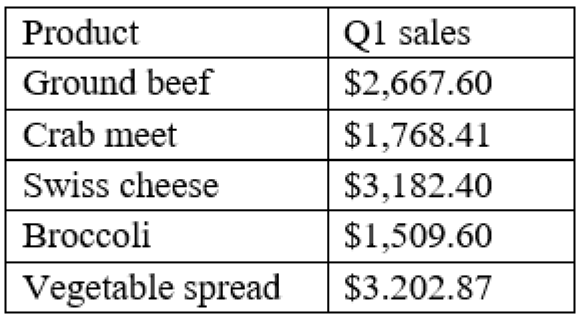
Question #1 *Topic 1*

A data analyst needs to calculate the mean for Q1 sales using the data set below:



Product	Q1 sales
Ground beef	$2,667.60
Crab meet	$1,768.41
Swiss cheese	$3,182.40
Broccoli	$1,509.60
Vegetable spread	$3.202.87

Which of the following is the mean?

A. $2,466.18

B. $2,667.60

C. $3,082.72

D. $12,330.88

Correct Answer: *A*

Community vote distribution

A (100%)

ronniehaang Highly Voted 2 years, 1 month ago

Selected Answer: A

The mean can be calculated by adding up all the Q1 sales values and then dividing by the number of items in the data set. In this case, the total of all the Q1 sales is $12,330.88, and there are 5 items in the data set.

So, the mean is calculated as follows:

$12,330.88 ÷ 5 = $2,466.18

Therefore, the mean is $2,466.18.

Answer: A. $2,466.18

upvoted 5 times

2e6d681 Most Recent 7 months ago

Selected Answer: A

The mean is calculated as follows: $12,330.88 ÷ 5 = $2,466.18 Therefore, the mean is $2,466.18. Answer: A. $2,466.18

upvoted 1 times

ViralStarfish 1 year ago

I'm so glad it's not just me that thought A was the answer. This gives a good impression as the first practice question, I must say.

upvoted 1 times

mcgoogol 1 year, 1 month ago

There answer is the median not the mean!

upvoted 1 times

brollo 1 year, 10 months ago

Selected Answer: A

It's obviously A since it's the sum of all elements divided by the number of elements

upvoted 3 times

Bongi12 2 years, 1 month ago

Correct answer is A

upvoted 2 times

CineFeX 2 years, 2 months ago

The numbers are wrong... 2667.60+1768.41+3182.40+1509.60+3202.87== Sum 12,131.28

□ /5 = 2,426.26 weird....

upvoted 1 times

lordguck 2 years, 3 months ago

A: is correct. Add all values and div by 5

upvoted 2 times

SolventCourseisSCAM 2 years, 4 months ago

Selected Answer: A

$2,466.176 Answer is A

upvoted 1 times

Question #2 *Topic 1*

A data analyst is creating a report that will provide information about various regions, products, and time periods. Which of the following formats would be the MOST efficient way to deliver this report?

A. A workbook with multiple tabs for each region

B. A daily email with snapshots of regional summaries

C. A static report with a different page for every filtered view

D. A dashboard with filters at the top that the user can toggle

Correct Answer: *D*

Community vote distribution

D (100%)

Hiromi_Nakatani 7 months, 4 weeks ago

Selected Answer: D

D. A dashboard with filters at the top that the user can toggle

upvoted 1 times

ronniehaang 8 months ago

Selected Answer: D

D. A dashboard with filters at the top that the user can toggle would be the MOST efficient way to deliver this report. A dashboard allows for a clear and concise presentation of information with the ability to filter and view data in different ways. This allows the user to quickly access the information they need without having to navigate through multiple tabs or pages. Additionally, dashboards can be interactive and allow for real-time updates, providing the most up-to-date information. With the ability to filter by region, product, and time period, the dashboard provides the user with a comprehensive and customizable view of the data, making it the most efficient way to deliver this report.

upvoted 2 times

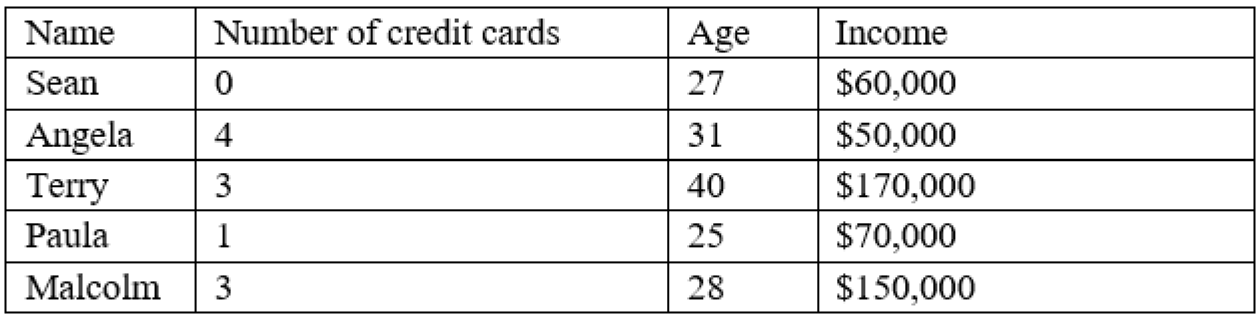
CineFeX 9 months ago

D. A dashboard with filters at the top that the user can toggle would be the most efficient way to deliver the report. This format allows the user to easily view and compare data for different regions, products, and time periods by using the filters to toggle between different views of the data. The other options would require the user to manually navigate between different pages or tabs to compare the data, which would be less efficient.

upvoted 2 times

Question #3 *Topic 1*

A customer list from a financial services company is shown below:



Name	Number of credit cards	Age	Income
Sean	0	27	$60,000
Angela	4	31	$50,000
Terry	3	40	$170,000
Paula	1	25	$70,000
Malcolm	3	28	$150,000

A data analyst wants to create a likely-to-buy score on a scale from 0 to 100, based on an average of the three numerical variables: number of credit cards, age, and income. Which of the following should the analyst do to the variables to ensure they all have the same weight in the score calculation?

A. Recode the variables.

B. Calculate the percentiles of the variables.

C. Calculate the standard deviations of the variables.

D. Normalize the variables.

Correct Answer: *D*

Community vote distribution

D (100%)

yiyigah 6 months, 1 week ago

D. Normalize the variables. Normalization involves transforming the values of the variables so that they have the same scale and range, typically from 0 to 1. This allows the analyst to ensure that each variable has the same weight in the calculation, since all the values are on the same scale. Normalizing the variables will also make it easier to interpret the results and compare the scores of different customers. This is important in creating a likely-to-buy score because it ensures that all three variables have an equal impact on the final score, rather than one variable having a larger weight due to its scale or range.

upvoted 2 times

mcgoogol 1 year, 1 month ago

If you didnt normalise the data you could rank it and then take the average, which would be similar to recoding. However any statistician would normalise the data

upvoted 1 times

r04dB10ck 2 years ago

Selected Answer: D

normalization of values definitely, depending on the normalization algorithm it may not be a range between 0 to 1, but all values will definitely fit small interval and have same weight/scale

upvoted 2 times

ronniehaang 2 years, 1 month ago

Selected Answer: D

D. Normalize the variables. Normalization involves transforming the values of the variables so that they have the same scale and range, typically from 0 to 1. This allows the analyst to ensure that each variable has the same weight in the calculation, since all the values are on the same scale. Normalizing the variables will also make it easier to interpret the results and compare the scores of different customers. This is important in creating a likely-to-buy score because it ensures that all three variables have an equal impact on the final score, rather than one variable having a larger weight due to its scale or range.

upvoted 3 times

CineFeX 2 years, 2 months ago

D. Normalize the variables. Normalization is the process of scaling a variable to have a values between 0 and 1. This can be done by subtracting the minimum value of the variable from all the values, and then dividing the resulting values by the range (i.e., the difference between the maximum and minimum values). Normalizing the variables ensures that they all have the same weight in the score calculation, since they are all on the same scale. The other options would not have the same effect. Recoding the variables would involve changing the values of the variables in some way, which would not necessarily give them the same weight. Calculating percentiles or standard deviations would not change the scale of the variables and therefore would not give them the same weight.

upvoted 3 times

Question #4 *Topic 1*

Which of the following actions should be taken when transmitting data to mitigate the chance of a data leak occurring? (Choose two.)

A. Data identification

B. Data processing

C. Data reporting

D. Data encryption

E. Data masking

F. Fata removal

Correct Answer: *DE*

Community vote distribution

DE (100%)

ronniehaang Highly Voted 8 months ago

Selected Answer: DE

D. Data encryption

E. Data masking

Data encryption involves transforming the data into an unreadable format so that it cannot be understood without the appropriate decryption key. This protects the data from being intercepted or viewed by unauthorized parties during transmission.

Data masking involves obscuring sensitive data elements by replacing them with non-sensitive values. This allows the data to be used for testing, development, and reporting purposes without compromising its confidentiality. Data masking also helps to mitigate the risk of data leaks, since the sensitive information is not accessible to unauthorized parties.

By using both encryption and masking, the chance of a data leak occurring during transmission can be significantly reduced, as the data is protected both in transit and at rest.

upvoted 5 times

daksa Most Recent 4 weeks, 1 day ago

Selected Answer: DE

since it says to choose two, both D and E are answers.

upvoted 1 times

CineFeX 9 months ago

Data encryption and data masking are both methods that can be used to protect data when it is transmitted. Data encryption involves converting the data into a coded form that can only be accessed by someone with the correct decryption key. Data masking involves replacing sensitive data with fake data that looks similar, but does not reveal the actual data. These techniques can help to reduce the risk of a data leak occurring by making it more difficult for unauthorized parties to access the data. The other options are not related to data transmission and would not be effective in mitigating the risk of a data leak.

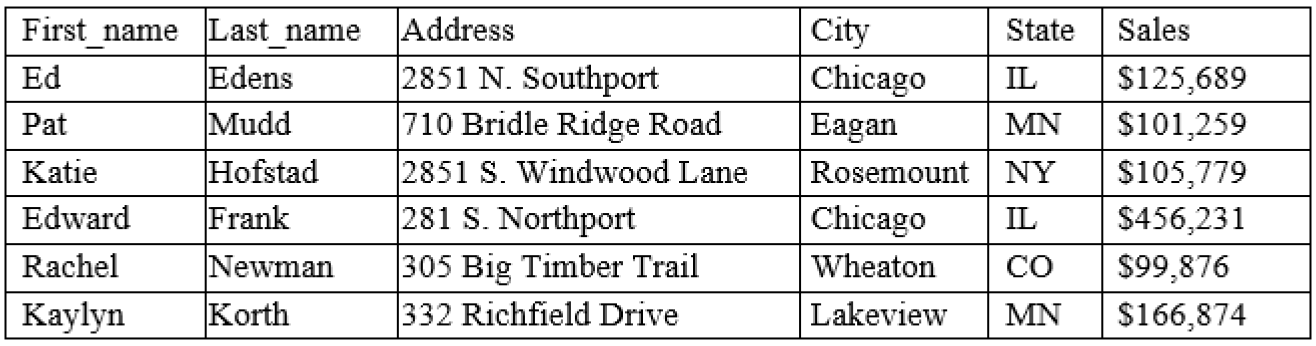
upvoted 3 times

Question #5 *Topic 1*

A data analyst has been asked to organize the table below in the following ways:

By sales from high to low -

By state in alphabetic order -



First_name	Last_name	Address	City	State	Sales
Ed	Edens	2851 N. Southport	Chicago	IL	$125,689
Pat	Mudd	710 Bridle Ridge Road	Eagan	MN	$101,259
Katie	Hofstad	2851 S. Windwood Lane	Rosemount	NY	$105,779
Edward	Frank	281 S. Northport	Chicago	IL	$456,231
Rachel	Newman	305 Big Timber Trail	Wheaton	CO	$99,876
Kaylyn	Korth	332 Richfield Drive	Lakeview	MN	$166,874

Which of the following functions will allow the data analyst to organize the table in this manner?

A. Conditional formatting

B. Grouping

C. Filtering

D. Sorting

Correct Answer: *D*

Community vote distribution

D (100%)

DaCulprito_1 6 months, 2 weeks ago

Selected Answer: D

Sorting allows the data to be arranged in a specified order.

upvoted 1 times

mcgoogol 7 months ago

Sorting in ascending or descending order does exactly what it is required

upvoted 1 times

EscCode 11 months, 4 weeks ago

Grouping? not sorting , why??

upvoted 1 times

ronniehaang 1 year, 8 months ago

Selected Answer: D

D. Sorting is the function that will allow the data analyst to organize the table in the desired manner. Sorting allows the data analyst to arrange the data in a specific order, either in ascending or descending order, based on one or multiple columns. In this case, the analyst can sort the table by Age from high to low by clicking on the Age column and selecting "Sort Descending". They can then sort the table by Income in alphabetic order by clicking on the Income column and selecting "Sort A to Z". Sorting is a simple and efficient way to organize large data sets, making it easier for the analyst to analyze and present the data.

upvoted 2 times

CineFeX 1 year, 9 months ago

Sorting is the process of organizing a table by rearranging the rows based on the values in one or more columns. In this case, the data analyst can use the sorting function to organize the table by sales from high to low, and then by state in alphabetic order. To do this, the data analyst would select the sales column and then choose the "sort largest to smallest" option, and then select the state column and choose the "sort A to Z" option. This will rearrange the rows of the table so that the rows are first sorted by sales from high to low, and then within those groups, the rows are sorted by state in alphabetic order. The other options (conditional formatting, grouping, and filtering) are not related to organizing the table in this manner.

upvoted 2 times

lordguck 1 year, 10 months ago

D: Sorting

upvoted 3 times

Question #6 *Topic 1*

Which of the following BEST describes the issue in which character values are mixed with integer values in a data set column?

A. Duplicate data

B. Missing data

C. Data outliers

D. Invalid data type

Correct Answer: *D*

Community vote distribution

D (80%) C (20%)

mcgoogol 7 months ago

it couldnt be duplicate data! Most likely to be invalid

upvoted 1 times

7380698 6 months, 3 weeks ago

How come it is A? Are some of these questions incorrect? Is the comptia system wrong then?

upvoted 1 times

ronniehaang 1 year, 7 months ago

Selected Answer: C

Outliers or extreme values can sometimes indicate invalid data. These values that lie far outside the normal range could signal a technical issue, or a problem with the collection or extraction of the data.

upvoted 1 times

Hiromi_Nakatani 1 year, 7 months ago

Selected Answer: D

D. Invalid data type

upvoted 1 times

ronniehaang 1 year, 8 months ago

Selected Answer: D

D. Invalid data type is the BEST description for the issue in which character values are mixed with integer values in a data set column. In this case, the data set column has a mix of two different data types - character values (e.g. text or string) and integer values (whole numbers). This is considered invalid as a data set column should have a consistent data type for all values in the column.

When character values are mixed with integer values, it can cause problems during data analysis and processing, as the software may not know how to handle these mixed values. For example, if the analyst tries to perform mathematical operations on the column, the software may produce an error or incorrect results, as it does not know how to handle the character values.

upvoted 2 times

ronniehaang 1 year, 8 months ago

Therefore, it is important to ensure that all data types in a data set column are consistent, in order to avoid problems during data analysis and processing. The issue of invalid data types can often be resolved by converting the values to a consistent data type, such as converting character values to integer values, or vice versa.

upvoted 2 times

CineFeX 1 year, 9 months ago

D. Invalid data type

If a data set column contains character values mixed with integer values, it would be considered to have an invalid data type. A data type refers to the type of information that a column can contain, such as numbers, text, or dates. Mixing character values with integer values in the same column would be considered invalid because it violates the expected data type for the column. The other options (duplicate data, missing data, and data outliers) are not related to the data type of the values in the column.

upvoted 2 times

lordguck 1 year, 10 months ago

D: is correct

upvoted 2 times

wanyu 1 year, 11 months ago

Selected Answer: D

It's D

upvoted 1 times

Question #7 *Topic 1*

Which of the following is a process that is used during data integration to collect, blend, and load data?

A. MDM

B. ETL

C. OLTP

D. BI

Correct Answer: *B*

Community vote distribution

B (100%)

ronniehaang 8 months ago

Selected Answer: B

B. ETL (Extract, Transform, Load) is a process that is used during data integration to collect, blend, and load data. ETL involves three main steps:

Extract: In this step, data is gathered from a variety of sources, such as databases, spreadsheets, and APIs, and is then extracted and stored in a temporary data repository.

Transform: In this step, the data is cleaned, transformed, and transformed into a format that can be loaded into the target system. This step involves tasks such as data cleansing, data enrichment, data matching, and data mapping.

Load: In this step, the transformed data is loaded into the target system, such as a data warehouse or data mart, where it can be analyzed and used for reporting and business intelligence purposes.

upvoted 2 times

ronniehaang 8 months ago

ETL is used in many organizations to ensure that data from different sources is consistent and accurate, and can be used for business intelligence purposes. It is a crucial step in the data integration process, as it allows organizations to consolidate and standardize data from disparate sources into a single repository, making it easier to access, analyze, and report on.

upvoted 2 times

CineFeX 9 months ago

B. ETL (Extract, Transform, Load) is a process that is used during data integration to collect, blend, and load data. The process involves extracting data from multiple sources, transforming the data into a format that is suitable for analysis, and then loading the data into a target system, such as a data warehouse. ETL is commonly used to integrate data from various sources, such as databases, flat files, and web services, and to prepare the data for reporting and analysis. The other options (MDM, OLTP, and BI) are not related to the ETL process.

upvoted 2 times

Question #8 *Topic 1*

An analyst has received the requirements for an internal user dashboard. The analyst confirms the data sources and then creates a wireframe. Which of the following is the NEXT step the analyst should take in the dashboard creation process?

A. Optimize the dashboard.

B. Create subscriptions.

C. Get stakeholder approval.

D. Deploy to production.

Correct Answer: *A*

Community vote distribution

C (100%)

10cccordrazine 1 year, 7 months ago

Selected Answer: C

As everyone already said, and as you can read in the official study guide, p.250, it should be C

upvoted 2 times

7380698 6 months, 3 weeks ago

Why is it A? I took the comptia exam a few weeks ago. I know these are the questions. Should I follow their answers?

upvoted 1 times

ronniehaang 1 year, 8 months ago

C. Get stakeholder approval is the NEXT step the analyst should take in the dashboard creation process after creating the wireframe. After confirming the data sources and creating a wireframe of the internal user dashboard, it is important to get approval from stakeholders who will be using the dashboard. This includes both internal and external stakeholders, as well as end users who will be accessing the dashboard.

Stakeholder approval is necessary to ensure that the dashboard meets the requirements of the stakeholders, and that it is aligned with the overall goals of the organization. During the approval process, stakeholders may provide feedback or make changes to the wireframe, and the analyst can use this feedback to make any necessary updates to the dashboard design.

upvoted 3 times

ronniehaang 1 year, 8 months ago

Once the stakeholders have provided approval, the analyst can then move forward with deploying the dashboard to production, which involves making the dashboard available to end users and making any final optimizations to ensure that the dashboard is fast, secure, and reliable. This step also involves setting up any necessary subscriptions, so that stakeholders receive regular updates or notifications related to the dashboard.

upvoted 3 times

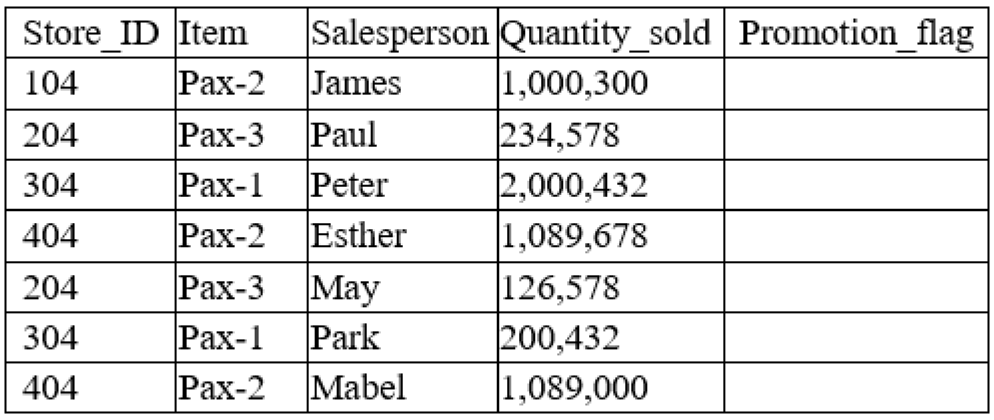
CineFeX 1 year, 9 months ago

After creating a wireframe, the next step in the dashboard creation process would typically be to get stakeholder approval. A wireframe is a low-fidelity prototype of the dashboard that shows the layout and structure of the final product, but does not include all of the detailed design elements or functionality. Once the wireframe has been created, it is important to get feedback from stakeholders, such as the internal users who will be using the dashboard, to ensure that the final product meets their needs and expectations. After obtaining stakeholder approval, the analyst can then proceed to the next steps, which might include optimizing the dashboard, creating subscriptions, and deploying to production.

upvoted 2 times

Question #9 *Topic 1*

A data analyst has been asked to derive a new variable labeled "Promotion_flag" based on the total quantity sold by each salesperson. Given the table below:



Store_ID	Item	Salesperson	Quantity_sold	Promotion_flag
104	Pax-2	James	1,000,300	
204	Pax-3	Paul	234,578	
304	Pax-1	Peter	2,000,432	
404	Pax-2	Esther	1,089,678	
204	Pax-3	May	126,578	
304	Pax-1	Park	200,432	
404	Pax-2	Mabel	1,089,000	

Which of the following functions would the analyst consider appropriate to flag "Yes" for every salesperson who has a number above 1,000,000 in the Quantity_sold column?

A. Date

B. Mathematical

C. Logical

D. Aggregate

Correct Answer: *C*

Community vote distribution

C (100%)

mcgoogol 7 months ago

Aggregate combines the data in some way.
Logical uses IF functions as in if greater than a million which is what is required here

upvoted 2 times

7380698 6 months, 3 weeks ago

Some of these answers to the questions to me dont seem right and I took the comptia data a few weeks ago. Same questions! Should I follow their answers?

upvoted 1 times

ronniehaang 1 year, 8 months ago

Selected Answer: C

C. Logical functions would be appropriate to flag "Yes" for every salesperson who has a number above 1,000,000 in the Quantity_sold column. Logical functions in data analysis are used to make decisions based on a set of conditions. In this case, the analyst can use a logical function such as an IF statement, where the condition is that the salesperson's Quantity_sold is above 1,000,000. If the condition is met, then the Promotion_flag is set to "Yes," otherwise it is set to "No."

For example, the formula could be something like:
IF(Quantity_sold > 1000000, "Yes", "No")

This formula would then be applied to the entire column of Quantity_sold values, and the resulting Promotion_flag column would indicate whether each salesperson is eligible for a promotion based on the number of items they have sold. Logical functions are an essential tool in data analysis, as they allow analysts to make decisions based on specific criteria and automate certain processes in the data analysis process.

upvoted 3 times

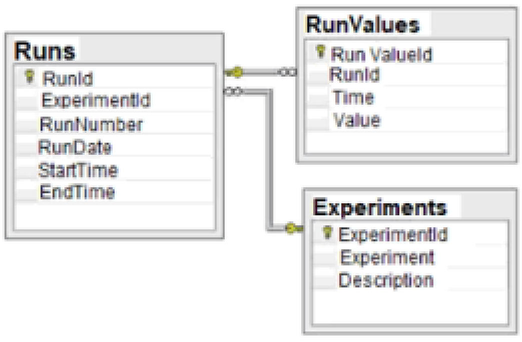
CineFeX 1 year, 9 months ago

C. A logical function would be appropriate for creating a new variable called "Promotion_flag" that is based on the values in the "Quantity_sold" column. A logical function allows you to perform a logical test and returns a value of TRUE or FALSE based on the test. In this case, the analyst could use a logical function to flag "Yes" for every salesperson who has a quantity sold above 1,000,000. For example, the analyst could use the IF function, which has the following syntax: IF(logical_test, value_if_true, value_if_false). The logical_test would be a comparison of the "Quantity_sold" column to 1,000,000, and the value_if_true would be "Yes" and the value_if_false would be "No". The other options (Date, Mathematical, and Aggregate) are not related to logical functions.

upvoted 3 times

Question #10 *Topic 1*

Given the diagram below:



Which of the following data schemas shown?

A. Key-value pairs

B. Online transactional processing

C. Data lake

D. Relational database

Correct Answer: *D*

CineFeX 9 months ago

D. The diagram shows a schema for a relational database, which is a type of database that stores data in tables that are related to each other through common columns or keys. In the diagram, the "Runs" table is connected to the "Athletes" and "Races" tables through the "Athlete_id" and "Race_id" columns, respectively. These columns are foreign keys that reference the primary keys in the "Athletes" and "Races" tables. This type of structure allows data to be stored and accessed in a structured and organized way, and enables relationships between different data entities to be captured and queried. The other options (key-value pairs, online transactional processing, and data lake) do not describe a relational database schema.

upvoted 2 times

Question #11 *Topic 1*

A company's marketing department wants to do a promotional campaign next month. A data analyst on the team has been asked to perform customer segmentation, looking at how recently a customer bought product, at what frequency, and at what value. Which of the following types of analysis would this practice be considered?

A. Prescriptive

B. Trend

C. Gap

D. Custer

Correct Answer: *D*

Community vote distribution

D (100%)

daksa 4 weeks, 1 day ago

Selected Answer: D

CLUSTER not custer

upvoted 1 times

ronniehaang 8 months ago

D. Customer segmentation is the practice of dividing a customer base into smaller groups based on common characteristics such as demographics, buying behaviors, or psychographic traits. In this case, the analyst is looking at three characteristics: how recently a customer bought a product, how frequently they bought, and the value of their purchases. By grouping customers based on these characteristics, the company can tailor its promotional campaign to different segments, ensuring that the right message and offer is being communicated to the right people.

Customer segmentation is considered a descriptive analysis, as it involves looking at the data and grouping it into meaningful categories. This type of analysis is often used in marketing and sales to better understand customer behavior and develop targeted strategies for customer engagement and retention. By using customer segmentation, companies can more effectively allocate their marketing resources and create more personalized experiences for their customers.

upvoted 2 times

CineFeX 9 months ago

Customer segmentation is the process of dividing customers into groups based on shared characteristics, such as purchase history, frequency of purchases, and value of purchases. This type of analysis is often referred to as cluster analysis or clustering. Cluster analysis is a method of dividing a data set into groups (i.e., clusters) such that the data points within a cluster are more similar to each other than they are to data points in other clusters. The marketing department's request to look at how recently a customer bought a product, at what frequency, and at what value would be considered cluster analysis because it involves dividing customers into groups based on shared characteristics. The other options (prescriptive, trend, and gap) are not related to cluster analysis.

upvoted 2 times

Question #12 *Topic 1*

A publishing group has requested a dashboard to track submissions before publication. A key requirement is that all changes are tracked, as multiple users will be checking out documents and editing them before submissions are considered final. Which of the following is the BEST way to meet this stakeholder requirement?

A. Display the version number next to each submission on the dashboard.

B. Present a data refresh date at the top of the dashboard.

C. Confirm the dashboard is adhering to the corporate style guide.

D. Use permissions to ensure users only see certain versions of the submissions.

Correct Answer: *A*

Community vote distribution

A (100%)

ronniehaang 8 months ago

Selected Answer: A

A. Display the version number next to each submission on the dashboard. This option meets the stakeholder requirement by allowing them to track all changes by tracking the version number of each submission. The version number would increment with each change, allowing the user to easily see the latest version of each document and keep track of changes over time. This would be the best option to meet the requirement of tracking all changes.

upvoted 2 times

CineFeX 9 months ago

One way to meet the stakeholder requirement of tracking all changes to submissions before publication is to display the version number next to each submission on the dashboard. This would allow users to see which version of the submission they are looking at, and would make it easier to track changes over time. The other options (presenting a data refresh date, confirming adherence to the corporate style guide, and using permissions to limit access) would not directly address the requirement to track changes to the submissions.

upvoted 2 times

Question #13 *Topic 1*

The number of phone calls that call center receives in a day is an example of:

A. continuous data.

B. categorical data.

C. ordinal data.

D. discrete data.

Correct Answer: *D*

Community vote distribution

D (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: D

Discrete data

upvoted 1 times

CineFeX 1 year, 3 months ago

D. The number of phone calls that a call center receives in a day is an example of discrete data. Discrete data refers to data that can only take on specific, distinct values within a given range. In this case, the number of phone calls can only be a whole number, and cannot take on any other value. For example, the call center might receive 100 calls one day, and 105 calls the next day, but it could not receive 100.5 calls. In contrast, continuous data refers to data that can take on any value within a given range, such as measurements or temperatures. Categorical data refers to data that can be divided into categories or groups, and ordinal data refers to data that has a specific order or ranking.

upvoted 2 times

Question #14 *Topic 1*

A data analyst is asked to create a sales report for the second-quarter 2020 board meeting, which will include a review of the business's performance through the second quarter. The board meeting will be held on July 15, 2020, after the numbers are finalized. Which of the following report types should the data analyst create?

A. Static

B. Real-time

C. Self-service

D. Dynamic

Correct Answer: *A*

Community vote distribution

A (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: A

Static

upvoted 1 times

CineFeX 1 year, 3 months ago

A. A static report is a report that is created using a snapshot of data at a specific point in time, and the data does not change unless the report is manually updated. In this case, the data analyst should create a static report for the second-quarter 2020 board meeting, which will be held on July 15, 2020, after the numbers are finalized. This means that the report will be based on data from the second quarter of 2020 that has already been collected and processed, and the data will not change unless the report is manually updated. The other options (real-time, self-service, and dynamic) are not applicable in this situation because the report is not being created for a live event or for real-time data, and it is not being created for self-service or dynamic use.

upvoted 3 times

Question #15 *Topic 1*

Which of the following would be considered non-personally identifiable information?

A. Cell phone device name

B. Customer’s name

C. Government ID number

D. Telephone number

Correct Answer: *A*

Community vote distribution

A (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: A

Cell phone device name

upvoted 2 times

CineFeX 1 year, 3 months ago

A. Non-personally identifiable information (NPI) is information that cannot be used to identify an individual. In this case, the cell phone device name would be considered NPI because it is not tied to a specific individual. The other options (customer's name, government ID number, and telephone number) are all personally identifiable information (PII) because they can be used to identify an individual. PII includes any information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual.

upvoted 2 times

willsy 8 months, 1 week ago

Unless they have their name on the phone.... but i agree with you as well

upvoted 2 times

Question #16 *Topic 1*

Which of the following is the correct data type for text?

A. Boolean

B. String

C. Integer

D. Float

Correct Answer: *B*

Community vote distribution

B (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: B

String

upvoted 2 times

CineFeX 1 year, 3 months ago

B. The correct data type for text is a string. A string is a data type that represents a sequence of characters, such as letters, numbers, and symbols. Strings are often used to store and manipulate text data, such as names, addresses, and descriptions. The other options (Boolean, integer, and float) are not appropriate data types for text. A Boolean data type represents a true or false value, and an integer data type represents a whole number. A float data type represents a decimal number.

upvoted 3 times

Question #17 *Topic 1*

Which of the following should be accomplished NEXT after understanding a business requirement for a data analysis report?

A. Rephrase the business requirement.

B. Determine the data necessary for the analysis.

C. Build a mock dashboard/presentation layout.

D. Perform exploratory data analysis.

Correct Answer: *B*

Community vote distribution

B (100%)

jreverte 8 months, 3 weeks ago

Selected Answer: B

You can't proceed if you don't know the data that will support the business scenario

upvoted 3 times

CineFeX 1 year, 3 months ago

B. After understanding a business requirement for a data analysis report, the next step should be to determine the data that is necessary for the analysis. This involves identifying the specific data points or variables that are required to answer the business question or meet the business need. Once the necessary data has been identified, the analyst can then proceed to the next steps, which might include rephrasing the business requirement, building a mock dashboard or presentation layout, and performing exploratory data analysis.

upvoted 3 times

Question #18 *Topic 1*

Which of the following is a common data analytics tool that is also used as an interpreted, high-level, general-purpose programming language?

A. SAS

B. Microsoft Power BI

C. IBM SPSS

D. Python

Correct Answer: *D*

Community vote distribution

D (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: D

Python

upvoted 2 times

CineFeX 1 year, 3 months ago

D. Python is a common data analytics tool that is also used as an interpreted, high-level, general-purpose programming language. Python is popular among data analysts and data scientists because of its simplicity, flexibility, and powerful data processing and visualization libraries. Python is often used to perform data manipulation, exploration, and analysis tasks, and can be used to build machine learning models, data pipelines, and data visualizations. The other options (SAS, Microsoft Power BI, and IBM SPSS) are also data analytics tools, but they are not programming languages.

upvoted 2 times

Question #19 *Topic 1*

A data analyst needs to present the results of an online marketing campaign to the marketing manager. The manager wants to see the most important KPIs and measure the return on marketing investment. Which of the following should the data analyst use to BEST communicate this information to the manager?

A. A real-time monitor that allows the manager to view performance the day the campaign was launched

B. A sell-service dashboard that allows the manager to look at the company's annual budget performance

C. A spreadsheet of the raw data from all marketing campaigns and channels

D. A summary with statistics, conclusions, and recommendations from the data analyst

Correct Answer: *D*

Community vote distribution

D (100%)

⊟ **vesen22** 4 months, 2 weeks ago

Selected Answer: D

Summary

upvoted 2 times

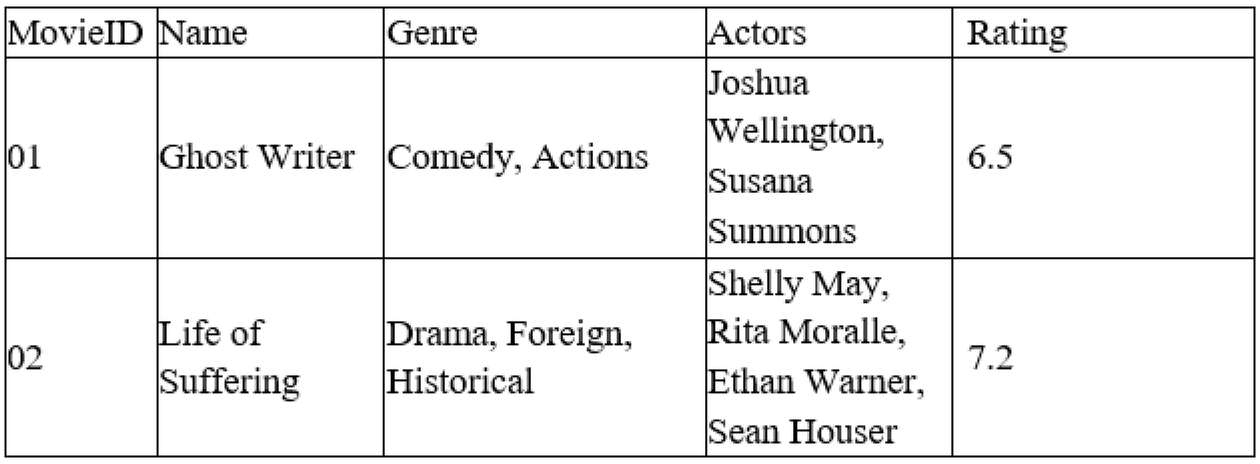
⊟ **CineFeX** 1 year, 3 months ago

D. To communicate the results of an online marketing campaign to the marketing manager, the data analyst should provide a summary that includes statistics, conclusions, and recommendations. This summary should highlight the most important key performance indicators (KPIs) and measure the return on marketing investment. The summary should be presented in a clear and concise manner, and should be tailored to the specific needs and interests of the manager. The other options (a real-time monitor, a self-service dashboard, or a spreadsheet of raw data) would not be as effective at communicating the key information in a clear and concise manner.

upvoted 2 times

Question #20 *Topic 1*

A data analyst for a media company needs to determine the most popular movie genre. Given the table below:



MovieID	Name	Genre	Actors	Rating
01	Ghost Writer	Comedy, Actions	Joshua Wellington, Susana Summons	6.5
02	Life of Suffering	Drama, Foreign, Historical	Shelly May, Rita Moralle, Ethan Warner, Sean Houser	7.2

Which of the following must be done to the Genre column before this task can be completed?

A. Append

B. Merge

C. Concatenate

D. Delimit

Correct Answer: *D*

Community vote distribution

D (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: D

Delimit

upvoted 1 times

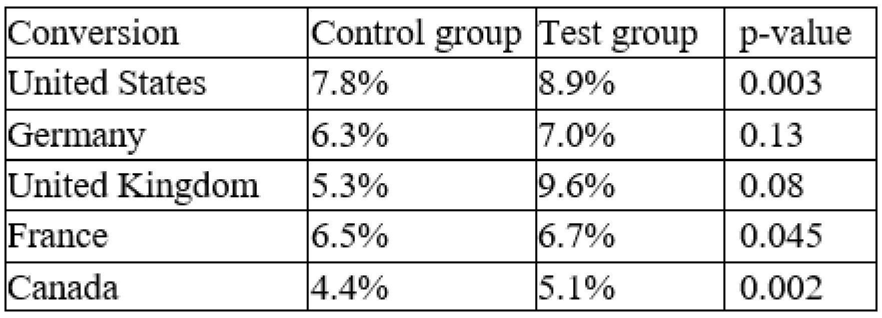
CineFeX 1 year, 3 months ago

D. In order to determine the most popular movie genre, the data analyst must first delimit the Genre column. Delimiting means separating or breaking up the data in a column into separate values or categories. In this case, the Genre column contains multiple movie genres for each movie, which are separated by commas. In order to determine the most popular movie genre, the data analyst must first split the Genre column into separate rows or categories for each movie genre. This will allow the data analyst to count the number of movies in each genre and determine the most popular genre. The other options (append, merge, and concatenate) do not involve separating or breaking up data in a column.

upvoted 2 times

Question #21 *Topic 1*

An e-commerce company recently tested a new website layout. The website was tested by a test group of customers, and an old website was presented to a control group. The table below shows the percentage of users in each group who made purchases on the websites:



Conversion	Control group	Test group	p-value
United States	7.8%	8.9%	0.003
Germany	6.3%	7.0%	0.13
United Kingdom	5.3%	9.6%	0.08
France	6.5%	6.7%	0.045
Canada	4.4%	5.1%	0.002

Which of the following conclusions is accurate at a 95% confidence interval?

A. In Germany, the increase in conversion from the new layout was not significant.

B. In France, the increase in conversion from the new layout was not significant.

C. In general, users who visit the new website are more likely to make a purchase.

D. The new layout has the lowest conversion rates in the United Kingdom.

Correct Answer: *C*

Community vote distribution

A (100%)

mcgoogol 7 months ago

I believe that there is no statistical proof that C is correct. With a 95% Confidence Interval we are comparing to a p-value of 0.025 (half of 5%) and so A is a true statement

upvoted 1 times

7380698 6 months, 3 weeks ago

Im tryna pass this exam in a few weeks and I saw these questions. Should I go with what they say?

upvoted 1 times

Correct_Damage 1 year, 5 months ago

"Which of the following conclusions is accurate at a 95% confidence interval?"

C. Answers this question more accurately then A

upvoted 1 times

ronniehaang 1 year, 7 months ago

Selected Answer: A

A. In Germany, the increase in conversion from the new layout was not significant.

This conclusion is accurate because the p-value for Germany is 0.13, which is higher than the commonly used significance level of 0.05. A p-value of 0.13 means that there is a 13% chance that the increase in conversion from the new layout was due to random chance, and therefore, the increase is not considered statistically significant. This means that we cannot reject the null hypothesis that the new layout does not have any effect on the conversion rate in Germany.

In contrast, the p-values for the United States, France, and Canada are lower than 0.05, meaning that the increase in conversion from the new layout in these countries is considered statistically significant. So, we can conclude that users who visit the new website are more likely to make a purchase in these countries.

upvoted 1 times

_F_M_ 1 year, 8 months ago

Selected Answer: A

A - 0.7% increase which is not significant (87% confidence) <- Correct

B - 0.2% increase which is significant (95.5% confidence) <- Incorrect

C - An average of -0.6% decrease which is not significant (94.8% Confidence) <- Incorrect

D - With its 4.3% improvement, UK have the highest conversion rate

upvoted 1 times

Stef987 1 year, 8 months ago

Hello! Could you elaborate how you calculate the confidence? Thanks!
upvoted 1 times

yandieg 1 year, 4 months ago
it is 100 - p * 100

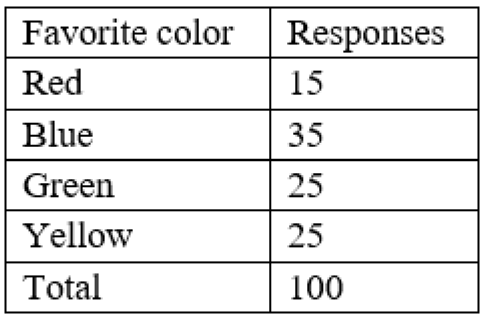
and for C is just the average
upvoted 1 times

CineFeX 1 year, 9 months ago
C. Based on the data in the table, it can be concluded that, in general, users who visit the new website are more likely to make a purchase. This conclusion can be made at a 95% confidence interval, which means that there is a 95% probability that this conclusion is accurate. The other conclusions are not supported by the data in the table. The data does not show a significant increase in conversion from the new layout in Germany or France, and the new layout does not have the lowest conversion rates in the United Kingdom.
upvoted 3 times

yandieg 1 year, 4 months ago
the question A is saying that the new layout is not bringing a significant increase (this is due a high p value) so the answer A is correct, as you declare in you explanation
"The data does not show a significant increase in conversion from the new layout in Germany or France"
upvoted 1 times

Question #22 *Topic 1*

An analyst needs to provide a chart to identify the composition between the categories of the survey response data set:



Favorite color	Responses
Red	15
Blue	35
Green	25
Yellow	25
Total	100

Which of the following charts would be BEST to use?

A. Histogram

B. Pie

C. Line

D. Scatter pot

E. Waterfall

Correct Answer: *B*

Community vote distribution

B (100%)

skumark 1 month, 3 weeks ago

Selected Answer: B

a pie chart would be most simple to understand from stake holder POV

upvoted 1 times

vesen22 4 months, 2 weeks ago

Selected Answer: B

Pie Chart

upvoted 2 times

CineFeX 1 year, 3 months ago

B. A pie chart would be the best chart to use to identify the composition between the categories of the survey response data set. A pie chart is a circular chart that shows the proportions or percentages of a whole. Each slice of the pie represents a category, and the size of the slice represents the proportion or percentage of the total that belongs to that category. Pie charts are useful for showing the relative sizes of categories or for comparing the proportions of different categories. The other chart types (histogram, line, scatter plot, and waterfall) are not appropriate for showing the composition between categories.

upvoted 2 times

Question #23 *Topic 1*

Five dogs have the following heights in millimeters:
300, 430, 170, 470, 600
Which of the following is the mean height for the five dogs?

A. 394mm

B. 405mm

C. 493mm

D. 504mm

Correct Answer: *A*

Community vote distribution

A (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: A

(300+430+170+470+600)/5 = 394

upvoted 2 times

CineFeX 1 year, 3 months ago

The mean height for the five dogs is 405mm. To calculate the mean, the heights of all the dogs are added together and then divided by the number of dogs. In this case, the heights of the five dogs are 300mm, 430mm, 170mm, 470mm, and 600mm, for a total of 2,270mm. When this total is divided by 5 (the number of dogs), the mean height is calculated to be 405mm.

upvoted 1 times

rizalrejab 1 year, 2 months ago

The mean height for the five dogs is 394mm. To calculate the mean, the heights of all the dogs are added together and then divided by the number of dogs. In this case, the heights of the five dogs are 300mm, 430mm, 170mm, 470mm, and 600mm, for a total of 1,970mm. When this total is divided by 5 (the number of dogs), the mean height is calculated to be 394mm.

upvoted 6 times

Question #24 *Topic 1*

Which of the following are reasons to create and maintain a data dictionary? (Choose two.)

A. To improve data acquisition

B. To remember specifics about data fields

C. To specify user groups for databases

D. To provide continuity through personnel turnover

E. To confine breaches of PHI data

F. To reduce processing power requirements

Correct Answer: *AD*

Community vote distribution

AD (100%)

ronniehaang 7 months, 3 weeks ago

Selected Answer: AD

A. To improve data acquisition
D. To provide continuity through personnel turnover

A data dictionary is a critical component of data management, and its creation and maintenance are important for several reasons.

A. To improve data acquisition: A data dictionary helps in improving data acquisition by providing clear definitions and descriptions of data elements, allowing for more accurate and consistent data collection. It also helps to eliminate duplicates and improve data quality by providing a clear understanding of what data is stored and where.

D. To provide continuity through personnel turnover: As personnel change within an organization, the data dictionary provides continuity and helps to ensure that data remains consistently defined and described. This helps to minimize the impact of personnel changes on data quality and consistency, allowing for seamless transitions in data management.

upvoted 2 times

ronniehaang 7 months, 3 weeks ago

Note: These are just two reasons why it is important to create and maintain a data dictionary. There are many other benefits as well, including reducing data breaches, improving data processing power, and more.

upvoted 2 times

Hiromi_Nakatani 7 months, 4 weeks ago

Selected Answer: AD

A. To improve data acquisition
D. To provide continuity through personnel turnover

upvoted 1 times

Question #25 *Topic 1*

A recurring event is being stored in two databases that are housed in different geographical locations. A data analyst notices the event is being logged three hours earlier in one database than in the other database. Which of the following is the MOST likely cause of the issue?

A. The data analyst is not querying the databases correctly.

B. The databases are recording different events.

C. The databases are recording the event in different time zones.

D. The second database is logging incorrectly.

Correct Answer: *C*

Community vote distribution

C (100%)

ronniehaang 7 months, 3 weeks ago

Selected Answer: C

C. The databases are recording the event in different time zones.

The most likely cause of the issue is that the databases are recording the event in different time zones. If the databases are housed in different geographical locations, it is possible that one database may be in a time zone that is three hours ahead of the other. This could result in the event being logged three hours earlier in one database than in the other. It is important to consider time zones when working with databases that are housed in different geographical locations to ensure accurate logging of events.

upvoted 3 times

CineFeX 9 months ago

C If a recurring event is being stored in two databases that are housed in different geographical locations, and the event is being logged three hours earlier in one database than in the other database, the most likely cause of the issue is that the databases are recording the event in different time zones. It is possible that the databases are located in different parts of the world, and each database is using a different time zone to record the event. This could result in the event being recorded at different times in the two databases.

The other options (the data analyst is not querying the databases correctly, the databases are recording different events, and the second database is logging incorrectly) are not as likely to be the cause of the issue, as they do not explain the three-hour difference in the recorded times of the event.

upvoted 2 times

Question #26 *Topic 1*

Which of the following is an example of a at flat file?

A. CSV file

B. PDF file

C. JSON file

D. JPEG file

Correct Answer: *A*

Community vote distribution

A (100%)

ronniehaang 1 year, 7 months ago

Selected Answer: A

A. CSV file is an example of a flat file. A flat file is a type of data file that contains only one table or data structure, with no relationships between tables. In a CSV file, data is stored in plain text format, with each row representing a single record and each column representing a separate field. The data fields are separated by commas, hence the name "Comma Separated Values".

upvoted 3 times

7380698 6 months, 3 weeks ago

What should I choose if I see this question on the day of the test? Should I go with what they say? I honestly thought its a CSV file

upvoted 1 times

CineFeX 1 year, 9 months ago

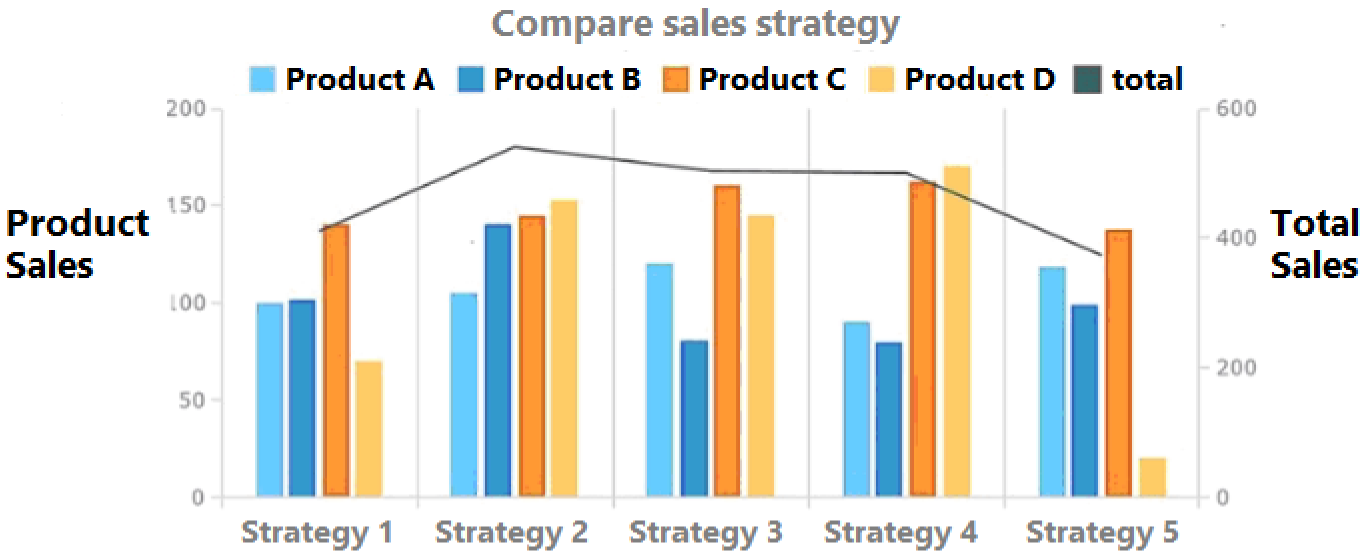
A. A CSV (comma-separated values) file is an example of a flat file. A flat file is a file that stores data in a simple, unstructured format, with each line of the file representing a single record. Flat files are typically used to store simple data sets that do not require the complexity or functionality of a full database management system. CSV files are a common type of flat file, and they are often used to store data that can be imported into other applications or systems.

The other options (PDF file, JSON file, and JPEG file) are not examples of flat files. PDF files are used to store documents, JSON files are used to store data in a structured, hierarchical format, and JPEG files are used to store images.

upvoted 3 times

Question #27 *Topic 1*

Given the following graph:



Which of the following summary statements upholds integrity in data reporting?

A. Sales are approximately equal for Product A and Product B across all strategies.

B. Strategy 4 provides the best sales in comparison to other strategies.

C. While Strategy 2 does not result in the highest sales of Product D, over all products it appears to be the most effective.

D. Product D should be promoted more than the other products in all strategies.

Correct Answer: *C*

Community vote distribution

C (75%) A (25%)

skumark 1 month, 3 weeks ago

Selected Answer: C

overall sales in strat 2 is highest and most effective

upvoted 1 times

mcgoogol 1 year, 1 month ago

Strategy 2 has the highest total sales. That is obtainable form the chart. None of the other suggestions have enough proof to be valid

upvoted 1 times

Swift_and_Quick 9 months, 3 weeks ago

Strategy 2 does not have the highest product sales for Product D, but it does have the highest total sales, which makes it the most effective, so C is the answer.

upvoted 2 times

examtp1 1 year, 11 months ago

Selected Answer: C

C. Summary statement C upholds integrity in data reporting because it accurately reflects the information shown in the graph. The graph shows that Strategy 2 does not result in the highest sales of Product D, but it appears to be the most effective overall when all products are considered. This summary statement accurately summarizes the information in the graph and does not make any false or misleading statements.

The other summary statements (Sales are approximately equal for Product A and Product B across all strategies, Strategy 4 provides the best sales in comparison to other strategies, and Product D should be promoted more than the other products in all strategies) are not accurate based on the information shown in the graph, and they do not uphold integrity in data reporting.

upvoted 2 times

ronniehaang 2 years, 1 month ago

Selected Answer: A

A. Sales are approximately equal for Product A and Product B across all strategies.

This statement upholds integrity in data reporting because it is based on the data shown in the graph, which shows that the sales for Product A and Product B are approximately equal across all strategies. The statement is objective and accurately reflects the data, without drawing any conclusions or biases. By presenting the data in an impartial and unbiased manner, the statement upholds the integrity of the data reporting.

upvoted 1 times

⊟ **CineFeX** 2 years, 2 months ago

C. Summary statement C upholds integrity in data reporting because it accurately reflects the information shown in the graph. The graph shows that Strategy 2 does not result in the highest sales of Product D, but it appears to be the most effective overall when all products are considered. This summary statement accurately summarizes the information in the graph and does not make any false or misleading statements.

The other summary statements (Sales are approximately equal for Product A and Product B across all strategies, Strategy 4 provides the best sales in comparison to other strategies, and Product D should be promoted more than the other products in all strategies) are not accurate based on the information shown in the graph, and they do not uphold integrity in data reporting.

upvoted 3 times

Question #28 *Topic 1*

An analyst is required to run a text analysis of data that is found in articles from a digital news outlet. Which of the following would be the BEST technique for the analyst to apply to acquire the data?

A. Web scraping

B. Sampling

C. Data wrangling

D. ETL

Correct Answer: *A*

Community vote distribution

A (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: A

Web scraping

upvoted 2 times

CineFeX 1 year, 3 months ago

Web scraping is a technique that can be used to acquire data from websites. It involves using a program or tool to extract data from the website's HTML code and structure it into a format that can be analyzed. This technique is often used to extract large amounts of data from websites that do not have an API (application programming interface) or other means of accessing the data.

In the case of an analyst who is required to run a text analysis of data found in articles from a digital news outlet, web scraping would be the best technique to apply to acquire the data. The analyst can use a web scraping tool or program to extract the text data from the news articles and structure it in a way that is suitable for analysis.

The other options (sampling, data wrangling, and ETL) are not as relevant to the task of acquiring data from a digital news outlet for text analysis. Sampling involves selecting a subset of data for analysis, data wrangling involves cleaning and preparing data for analysis, and ETL (extract, transform, load) refers to a process used to move data between systems.

upvoted 3 times

Question #29 *Topic 1*

An analyst runs a report on a daily basis, and the number of datapoints must be validated before the data can be analyzed. The number of datapoints increases each day by approximately 20% of the total number from the day before. On a given day, the number of datapoints was 8,798. Which of the following should be the total number of datapoints on the next day?

A. 7,038

B. 9,600

C. 10,600

D. 10,800

Correct Answer: *C*

Community vote distribution

C (100%)

_F_M_ Highly Voted 10 months, 3 weeks ago

Selected Answer: C

8,798 * 1.2 = 10,557.6

upvoted 7 times

ronniehaang Most Recent 7 months, 3 weeks ago

Selected Answer: C

C. 10,600

To calculate the number of datapoints on the next day, we need to find 20% increase of the total number on the current day which is 8,798. 20% of 8,798 is 1759.6. Adding that to the current total of 8,798, we get 10,557.6, which can be rounded to 10,600.

upvoted 4 times

CineFeX 9 months ago

C. The number of datapoints increases each day by approximately 20% of the total number from the day before. If on a given day the number of datapoints was 8,798, the total number of datapoints on the next day should be approximately 20% higher than this number. This means that the total number of datapoints on the next day should be around 8,798 + (20% * 8,798) = 8,798 + (1.6 * 8,798) = 8,798 + 14,087.6 = 22,885.6. Rounding this number to the nearest hundred gives 10,600. Therefore, the total number of datapoints on the next day should be approximately 10,600.

The other options (7,038, 9,600, and 10,800) are not correct because they do not accurately reflect the 20% increase in the number of datapoints from one day to the next.

upvoted 3 times

Question #30 *Topic 1*

An analyst has been tracking company intranet usage and has been asked to create a chat to show the most-used/most-clicked portions of a homepage that contains more than 30 links. Which of the following visualizations would BEST illustrate this information?

A. Scatter plot

B. Heat map

C. Pie chart

D. Infographic

Correct Answer: *B*

Community vote distribution

B (100%)

ronniehaang 7 months, 3 weeks ago

Selected Answer: B

B. Heat map would BEST illustrate this information. A heat map visually represents data using colors to indicate the intensity of the data at different points on a scale. In this case, the color can indicate the frequency of clicks or usage on the homepage links. This type of visualization allows for easy identification of the most-used or most-clicked portions of the homepage, which is the information being asked for.

upvoted 3 times

Hiromi_Nakatani 7 months, 4 weeks ago

Selected Answer: B

B. Heat map

upvoted 2 times

CineFeX 9 months ago

A heat map is a visual representation of data that uses color to encode values. It is often used to show patterns or trends in data, and it can be a useful tool for illustrating the most-used or most-clicked portions of a homepage.

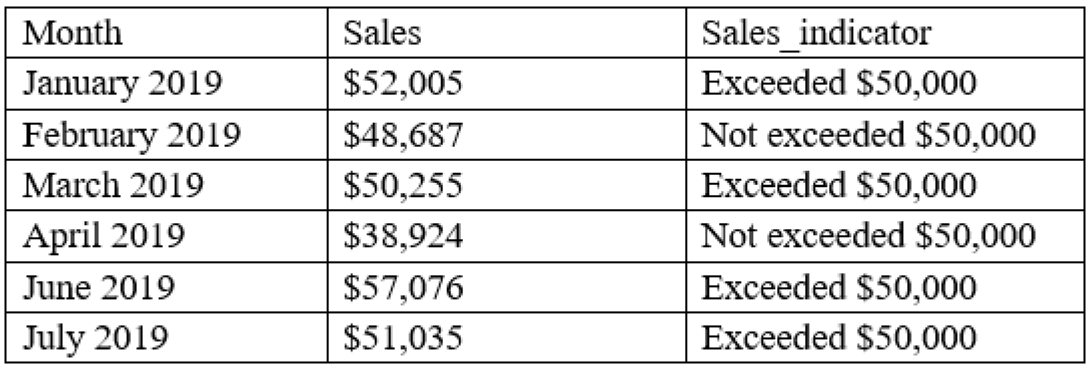
In the case of an analyst who has been tracking company intranet usage and has been asked to create a chart to show the most-used or most-clicked portions of a homepage that contains more than 30 links, a heat map would be the best visualization to illustrate this information. The heat map could use color to encode the number of clicks or usage for each link, with warmer colors indicating higher levels of usage and cooler colors indicating lower levels of usage. This would allow the analyst to quickly identify the most popular links on the homepage.

The other options (scatter plot, pie chart, and infographic) are not as well suited to illustrating the most-used or most-clicked portions of a homepage, as they do not use color to encode data values and may not be as effective at showing patterns or trends.

upvoted 4 times

Question #31 *Topic 1*

An analyst has generated a report that includes the number of months in the first two quarters of 2019 when sales exceeded $50,000:



Month	Sales	Sales_indicator
January 2019	$52,005	Exceeded $50,000
February 2019	$48,687	Not exceeded $50,000
March 2019	$50,255	Exceeded $50,000
April 2019	$38,924	Not exceeded $50,000
June 2019	$57,076	Exceeded $50,000
July 2019	$51,035	Exceeded $50,000

Which of the following functions did the analyst use to generate the data in the Sales_indicator column?

A. Aggregate

B. Logical

C. Date

D. Sort

Correct Answer: *B*

Community vote distribution

B (100%)

mcgoogol 7 months ago

Must have used a Logical if function

upvoted 1 times

ronniehaang 1 year, 7 months ago

Selected Answer: B

The function that the analyst used to generate the data in the Sales_indicator column is likely a Logical function. The function likely uses a logical comparison between the sales data and a fixed value of $50,000 to determine whether each month's sales exceeded the threshold. For example, the formula may look like =IF(sales > 50000, "Exceeded", "Not Exceeded").

upvoted 2 times

Hiromi_Nakatani 1 year, 7 months ago

Selected Answer: B

B. Logical

upvoted 1 times

CineFeX 1 year, 9 months ago

B. The Sales_indicator column in the table appears to be a boolean column that indicates whether sales exceeded $50,000 in a given month. This type of data can be generated using a logical function, which allows the analyst to evaluate a condition (in this case, whether sales exceeded $50,000) and return a boolean value (true or false) depending on the result of the evaluation.

In the table, the Sales_indicator column contains the value "True" for those months in which sales exceeded $50,000, and the value "False" for those months in which sales did not exceed $50,000. This indicates that the analyst used a logical function to evaluate the sales data and generate the Sales_indicator column.

The other options (aggregate, date, and sort) are not as well suited to generating the data in the Sales_indicator column, as they do not allow the analyst to evaluate a condition and return a boolean value based on the result of the evaluation.

upvoted 2 times

Question #32 *Topic 1*

While reviewing survey data, an analyst notices respondents entered “Jan,” “January,” and “01” as responses for the month of January. Which of the following steps should be taken to ensure data consistency?

A. Delete any of the responses that do not have “January” written out.

B. Replace any of the responses that have “01”.

C. Filter on any of the responses that do not say “January” and update them to “January”.

D. Sort any of the responses that say “Jan” and update them to “01”.

Correct Answer: *C*

Community vote distribution

C (100%)

egray 6 months, 2 weeks ago

Uh, assuming there are no other months in the table! Otherwise, "February" would be converted

upvoted 1 times

ronniehaang 1 year, 1 month ago

Selected Answer: C

C. Filter on any of the responses that do not say “January” and update them to “January”.

To ensure data consistency, it is important to standardize the responses. In this case, the analyst should filter on any responses that do not say “January” and update them to “January” to make sure all responses for the month of January are consistent. This will make it easier to analyze the data and avoid confusion or misinterpretation of the results.

The other options (A, B, and D) may not be the best approach as deleting responses, replacing responses with a different value, or sorting and updating the responses may lead to a loss of information or a misinterpretation of the results. It is important to retain as much data as possible while also ensuring that the data is consistent and accurate.

upvoted 2 times

CineFeX 1 year, 3 months ago

C. To ensure data consistency, it is important to ensure that all data values in a given column are expressed in the same way. In this case, the analyst has noticed that respondents entered "Jan," "January," and "01" as responses for the month of January. To ensure data consistency, the analyst should filter on any of the responses that do not say "January" (i.e. "Jan" and "01") and update them to "January." This will ensure that all responses for the month of January are expressed in the same way and that the data is consistent.

The other options (delete any responses that do not have "January" written out, replace any responses that have "01," or sort any responses that say "Jan" and update them to "01") are not as effective at ensuring data consistency, as they do not ensure that all responses for the month of January are expressed in the same way.

upvoted 2 times

Question #33 *Topic 1*

Which of the following data cleansing issues will be fixed when a DISTINCT function is applied?

A. Missing data

B. Duplicate data

C. Redundant data

D. Invalid data

Correct Answer: *B*

Community vote distribution

B (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: B

Duplicate data

upvoted 1 times

CineFeX 1 year, 3 months ago

B. A DISTINCT function is used to return unique values in a data set. It removes duplicates from the data set, so when a DISTINCT function is applied, the data cleansing issue of duplicate data will be fixed.

Missing data, redundant data, and invalid data are not issues that will be fixed by applying a DISTINCT function. These issues can be addressed using other data cleansing techniques such as imputation for missing data, consolidation for redundant data, and validation for invalid data.

upvoted 2 times

Question #34 *Topic 1*

A county in Illinois is conducting a survey to determine the mean annual income per household. The county is 427sq mi (2.65q km). Which of the following sampling methods would MOST likely result in a representative sample?

A. A stratified phone survey of 100 people that is conducted between 2:00 p.m. and 3:00 p.m.

B. A systematic survey that is sent to 100 single-family homes in the county

C. Surveys sent to ten randomly selected homes within 5mi (8km) of the county's office

D. Surveys sent to 100 randomly selected homes that are reflective of the population

Correct Answer: *D*

Community vote distribution

D (100%)

ronniehaang 7 months, 3 weeks ago

Selected Answer: D

D. Surveys sent to 100 randomly selected homes that are reflective of the population is the most likely method to result in a representative sample. This is because it ensures that the sample is taken from a broad range of the population, and not just limited to a specific area or time. It also ensures that the sample is a random selection, which minimizes the risk of selection bias. This is considered a better method than the other options, which may result in a biased sample due to the specific methods used (such as only surveying homes within a specific area or at a specific time).

upvoted 2 times

CineFeX 9 months ago

D. Random sampling is a method of selecting a representative sample from a larger population in which every member of the population has an equal chance of being selected. To ensure that the sample is representative of the population, it is important to select a large enough sample size and to ensure that the sample is drawn from the entire population. In this case, the county is 427sq mi (2.65q km) and the sample is being drawn from 100 randomly selected homes, so it is likely that the sample will be representative of the entire population.

Option A is not a representative sample because it is a phone survey conducted during a specific time period. Option B is not a representative sample because it is only sent to single-family homes. Option C is not a representative sample because it is only sent to homes within a specific distance of the county's office.

upvoted 2 times

Question #35 *Topic 1*

Which of the following statistical methods requires two or more categorical variables?

A. Simple linear regression

B. Chi-squared test

C. Z-test

D. Two-sample t-test

Correct Answer: *B*

Community vote distribution

B (100%)

ronniehaang 7 months, 3 weeks ago

Selected Answer: B

B. Chi-squared test requires two or more categorical variables.

A chi-squared test is a statistical method used to determine if there is a significant relationship between two categorical variables. It is used to determine if the observed frequencies of a variable are different from what would be expected based on a theoretical probability distribution. For example, a chi-squared test could be used to determine if there is a significant difference in the number of people who smoke and the number of people who have lung cancer. To perform a chi-squared test, the data must be in the form of a contingency table, with each cell representing the count of observations for each combination of categories for the two variables being tested. The test statistic is calculated based on the observed and expected frequencies, and the results are compared to a critical value from a chi-squared distribution to determine if the relationship between the two variables is significant.

upvoted 2 times

CineFeX 9 months ago

B. The chi-squared test is a statistical method that is used to compare two or more categorical variables. It is used to determine whether there is a significant difference between the observed frequencies of the variables and the expected frequencies. In order to use the chi-squared test, the data must be categorical (i.e., the variables must be divided into groups or categories).

Simple linear regression, z-test, and two-sample t-test do not require two or more categorical variables. They are used to compare two or more numerical variables.

upvoted 2 times

Question #36 *Topic 1*

Which of the following data manipulation techniques is an example of a logical function?

A. WHERE

B. AGGREGATE

C. BOOLEAN

D. IF

Correct Answer: *D*

Community vote distribution

D (100%)

Tota14 6 months, 1 week ago

Selected Answer: D

D .. boolean is not function

upvoted 3 times

CineFeX 9 months ago

D. D. IF

The IF function is a logical function that allows you to perform a certain action based on a condition being met or not met. For example, you might use the IF function to determine whether a certain value is greater than or equal to a certain threshold. If the value is greater than or equal to the threshold, the function will return one result; if the value is less than the threshold, the function will return another result.

WHERE, AGGREGATE, and BOOLEAN are not examples of logical functions. WHERE is a clause in SQL that is used to filter records based on specific conditions. AGGREGATE is a function that performs a calculation on a set of values and returns a single value. BOOLEAN is a data type that can represent one of two values: true or false.

upvoted 2 times

Question #37 *Topic 1*

A sales team wants visibility of current sales numbers, pipeline, and team performance. The team would also like to see calculations of individuals' earned commissions and projected commissions based on sales, but they want that information to be kept confidential. Which of the following would be the BEST way to provide this visibility?

A. Create a dashboard displaying a data refresh date so users know the current sales numbers and configure permissions to control access.

B. Create a dashboard for sales numbers, pipeline, and team and individual performance for the management team.

C. Create a dashboard with filters for the overall team, individuals, and management. Users can filter to see the data they want.

D. Create a dashboard with views for team, individuals, and management. Configure permissions to control access.

Correct Answer: *B*

Community vote distribution

A (100%)

ronniehaang 7 months, 3 weeks ago

Selected Answer: A

The BEST way to provide this visibility would be option A. Create a dashboard displaying a data refresh date so users know the current sales numbers and configure permissions to control access.

This option allows the sales team to have a general visibility of the sales numbers, pipeline, and team performance, while also providing the calculation of individuals' earned and projected commissions based on sales. The data refresh date ensures that the sales team is aware of the most up-to-date information. Additionally, the configuration of permissions to control access ensures that the confidential information regarding individual commissions is only available to authorized users.

upvoted 2 times

ronniehaang 7 months, 3 weeks ago

Option B is not ideal because it only provides visibility to the management team, excluding the rest of the sales team from seeing important information.

Option C is not ideal because it only provides a filter for the overall team, individuals, and management, but does not address the need for confidentiality regarding individual commissions.

Option D is not ideal because it only provides views for the team, individuals, and management, but does not provide a data refresh date and does not address the need for confidentiality regarding individual commissions.

upvoted 2 times

CineFeX 9 months ago

D. without a doubt. Just asked ChatGPT.

upvoted 2 times

ronniehaang 7 months, 3 weeks ago

Option B is not ideal because it only provides visibility to the management team, excluding the rest of the sales team from seeing important information.

Option C is not ideal because it only provides a filter for the overall team, individuals, and management, but does not address the need for confidentiality regarding individual commissions.

Option D is not ideal because it only provides views for the team, individuals, and management, but does not provide a data refresh date and does not address the need for confidentiality regarding individual commissions.

upvoted 1 times

Question #38 *Topic 1*

Which of the following is a characteristic of a relational database?

A. It utilizes key-value pairs.

B. It has undefined fields.

C. It is structured in nature.

D. It uses minimal memory.

Correct Answer: *C*

Swift_and_Quick 4 months, 3 weeks ago

C. It is structured in nature. This is because a relational database is a type of database that organizes data into tables, which consist of rows and columns. A relational database is structured in nature, which means that the data has a predefined schema or format, and follows certain rules and constraints, such as primary keys, foreign keys, or referential integrity. A relational database can be used to store, query, and manipulate data using a structured query language (SQL). The other characteristics are not true for a relational database.

upvoted 2 times

Correct_Damage 1 year, 5 months ago

A. It uses Primary and Foreign keys to connect data tables

upvoted 1 times

Question #39 *Topic 1*

A data analyst is asked on the morning of April 9, 2020, to create a sales report that identifies sales year to date. The daily sales data is current through the end of the day. Which of the following date ranges should be on the report?

A. January 1, 2020 to April 1, 2020

B. January 1, 2020 to April 7, 2020

C. January 1, 2020 to April 8, 2020

D. January 1, 2020 to April 9, 2020

Correct Answer: *C*

Community vote distribution

C (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: C

January 1, 2020 to April 8, 2020

upvoted 3 times

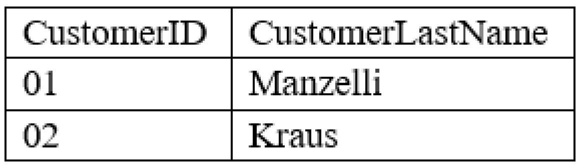
Correct_Damage 11 months, 4 weeks ago

C. The data for the 9th is not available at the time the request for the report was made.

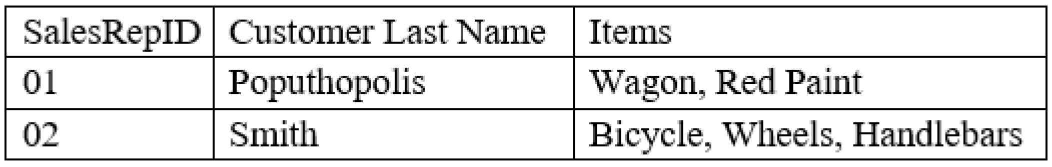
upvoted 3 times

Question #40 *Topic 1*

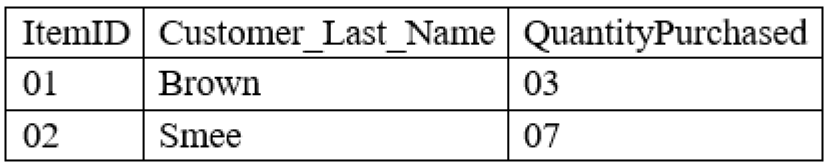
Given the following data tables:



CustomerID	CustomerLastName
01	Manzelli
02	Kraus



SalesRepID	Customer Last Name	Items
01	Poputhopolis	Wagon, Red Paint
02	Smith	Bicycle, Wheels, Handlebars



ItemID	Customer_Last_Name	QuantityPurchased
01	Brown	03
02	Smee	07

Which of the following MDM processes needs to take place FIRST?

A. Creation of a data dictionary

B. Compliance with regulations

C. Standardization of data field names

D. Consolidation of multiple data fields

Correct Answer: *C*

Community vote distribution

C (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: C

Standardization of data field names

upvoted 1 times

Correct_Damage 11 months, 4 weeks ago

C field names should be standardized

upvoted 2 times

Question #41 *Topic 1*

Which of the following is used for calculations and pivot tables?

A. IBM SPSS

B. SAS

C. Microsoft Excel

D. Domo

Correct Answer: *C*

Community vote distribution

C (100%)

⊟ **vesen22** 4 months, 2 weeks ago

Selected Answer: C

Microsoft Excel

upvoted 1 times

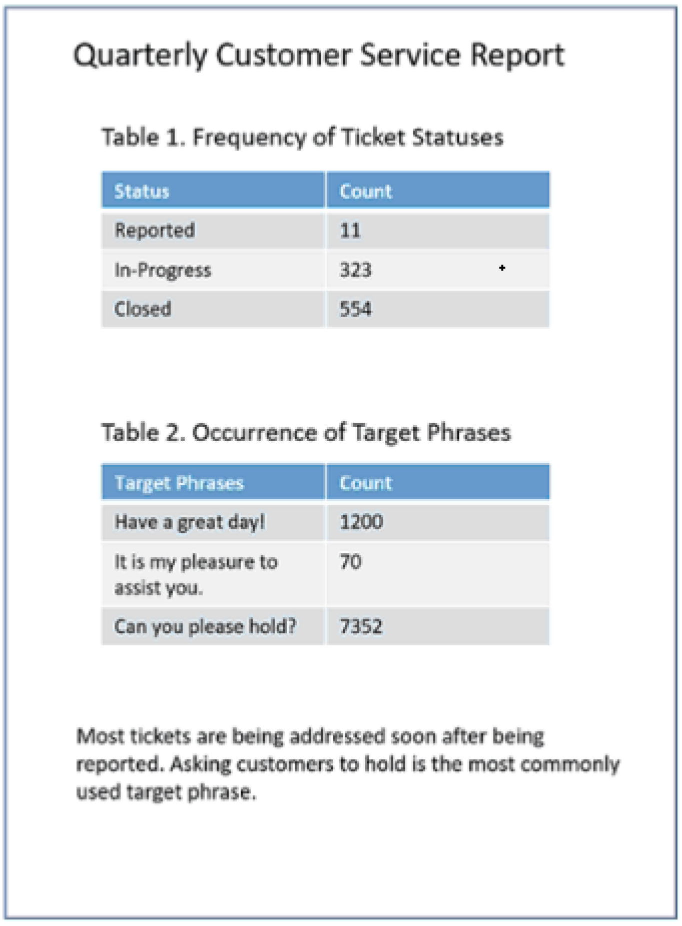
⊟ **Correct_Damage** 11 months, 4 weeks ago

C. Excel

upvoted 1 times

Question #42 *Topic 1*

Given the following report:



Quarterly Customer Service Report

Table 1. Frequency of Ticket Statuses

Status	Count
Reported	11
In-Progress	323
Closed	554

Table 2. Occurrence of Target Phrases

Target Phrases	Count
Have a great day!	1200
It is my pleasure to assist you.	70
Can you please hold?	7352

Most tickets are being addressed soon after being reported. Asking customers to hold is the most commonly used target phrase.

Which of the following components need to be added to ensure the report is point-in-time and static? (Choose two.)

A. A control group for the phrases

B. A summary of the KPIs

C. Filter buttons for the status

D. The date when the report was last accessed

E. The time period the report covers

F. The date on which the report was run

Correct Answer: *EF*

Community vote distribution

EF (100%)

Hiromi_Nakatani Highly Voted 7 months, 4 weeks ago

Selected Answer: EF

E. The time period the report covers

F. The date on which the report was run

upvoted 5 times

Question #43 *Topic 1*

An analyst has been asked to validate data quality. Which of the following are the BEST reasons to validate data for quality control purposes? (Choose two.)

A. Retention

B. Integrity

C. Transmission

D. Consistency

E. Encryption

F. Deletion

Correct Answer: *BD*

Community vote distribution

BD (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: BD

Integrity & Consistency.

upvoted 1 times

Correct_Damage 11 months, 4 weeks ago

B and D

upvoted 1 times

Question #44 *Topic 1*

A research analyst wants to determine whether the data being analyzed is connected to other datapoints. Which of the following is the BEST type of analysis to conduct?

A. Trend analysis

B. Performance analysis

C. Link analysis

D. Exploratory analysis

Correct Answer: *C*

Community vote distribution

C (100%)

HSZ 6 months, 2 weeks ago

Selected Answer: C

It is C. Link Analysis

upvoted 1 times

ronniehaang 1 year, 7 months ago

Selected Answer: C

C. Link analysis is the BEST type of analysis to conduct if the research analyst wants to determine whether the data being analyzed is connected to other datapoints.

Link analysis is a statistical technique that helps to uncover relationships and connections between different data points. It can be used to identify correlations and dependencies in data, as well as to identify any outliers that might indicate a potential problem or error in the data.

In link analysis, data points are visualized as nodes in a network and the connections between them are represented as lines. This visualization makes it easier for the analyst to see the relationships between different data points, and to identify any correlations or dependencies. Link analysis is particularly useful when the data being analyzed is complex and the connections between data points are not immediately obvious.

Overall, link analysis is a valuable tool for research analysts who want to better understand the relationships between different data points and uncover connections that might not be immediately apparent.

upvoted 2 times

Hiromi_Nakatani 1 year, 7 months ago

Selected Answer: C

C. Link analysis

upvoted 2 times

Question #45 *Topic 1*

Which of the following variable name formats would be problematic if used in the majority of data software programs?

A. First_Name_

B. FirstName

C. First_Name

D. First Name

Correct Answer: *D*

Community vote distribution

D (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: D

First Name

upvoted 1 times

Correct_Damage 11 months, 4 weeks ago

D. Never uses spaces, always an underscore in place or the space

upvoted 2 times

Question #46 *Topic 1*

Which of the following describes the method of sampling in which elements of data are selected randomly from each of the small subgroups within a population?

A. Simple random

B. Cluster

C. Systematic

D. Stratified

Correct Answer: *C*

Community vote distribution

D (100%)

Kennedy77 4 months, 2 weeks ago

Selected Answer: D

Stratified

upvoted 1 times

AmyD 7 months, 3 weeks ago

D. Stratified

Stratified Sampling is a method in which the population is divided into smaller subgroups, or strata, based on characteristics such as age, gender, socioeconomic status, etc. and a random sample is taken from each subgroup. This ensures that each subgroup is represented in the sample, so that the sample accurately reflects the diversity within the population.

Simple Random Sampling selects elements from the population at random, with no attempt to ensure representation of subgroups.

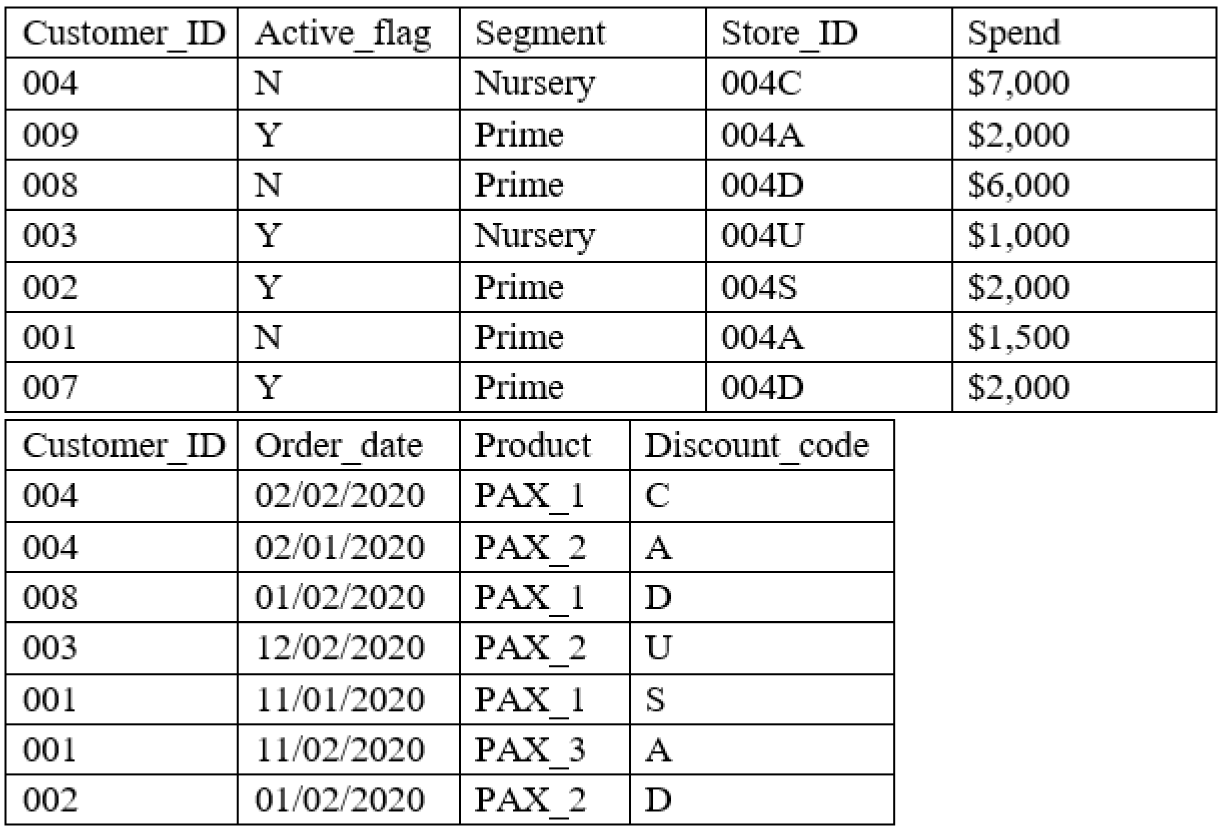
Cluster Sampling involves dividing the population into clusters, and then randomly selecting a number of clusters to be included in the sample.

Systematic Sampling involves selecting elements from the population at regular intervals, rather than randomly.

upvoted 3 times

Question #47 Topic 1

Given the following customer and order tables:



Customer_ID	Active_flag	Segment	Store_ID	Spend
004	N	Nursery	004C	$7,000
009	Y	Prime	004A	$2,000
008	N	Prime	004D	$6,000
003	Y	Nursery	004U	$1,000
002	Y	Prime	004S	$2,000
001	N	Prime	004A	$1,500
007	Y	Prime	004D	$2,000

Customer_ID	Order_date	Product	Discount_code
004	02/02/2020	PAX_1	C
004	02/01/2020	PAX_2	A
008	01/02/2020	PAX_1	D
003	12/02/2020	PAX_2	U
001	11/01/2020	PAX_1	S
001	11/02/2020	PAX_3	A
002	01/02/2020	PAX_2	D

Which of the following describes the number of rows and columns of data that would be present after performing an INNER JOIN of the tables?

A. Five rows, eight columns

B. Seven rows, eight columns

C. Eight rows, seven columns

D. Nine rows, five columns

Correct Answer: *A*

Community vote distribution

A (55%) B (45%)

ronniehaang Highly Voted 2 years, 1 month ago

Selected Answer: A

A. Five rows, eight columns.

Inner join returns only the rows where there is a match between the two tables in the specified join condition. In this case, the join condition would be based on the "Customer ID" column, which is present in both tables. So, the inner join of the two tables would result in 5 rows where there is a match in the "Customer ID" column, and each row would have 8 columns (all columns from both tables).

upvoted 6 times

Manizha Most Recent 4 months ago

Selected Answer: B

Table 1 has Customer_IDs: 004, 004, 008, 003, 001, 001, 002

Table 2 has Customer_IDs: 004, 009, 008, 003, 002, 001, 007

Let's count matches:

004 appears twice in Table 1, once in Table 2 = 2 rows in result

008 appears once in each = 1 row

003 appears once in each = 1 row

002 appears once in each = 1 row

001 appears twice in Table 1, once in Table 2 = 2 rows

(009 and 007 only appear in Table 2, so they drop out)

Total number of rows: 2 + 1 + 1 + 1 + 2 = 7 rows

Count columns:

Table 1: Customer_ID, Order_date, Product, Discount_code (4 columns)

Table 2: Customer_ID, Active_flag, Segment, Store_ID, Spend (5 columns)

When joined, Customer_ID appears only once

Total columns: 4 + 5 - 1 = 8 columns

Therefore, the result will have 7 rows and 8 columns.

upvoted 1 times

⊟ **MadalinaUsurelu** 5 months, 2 weeks ago

B. I just tried in SQL

upvoted 1 times

⊟ **Swift_and_Quick** 11 months, 2 weeks ago

B is the correct answer. There are 7 rows in the order table, all those 7 rows will retain when inner join occurred because Customer_ID matches the customer table. Since customer like 001 and 004 appear twice in the order table, they will appear twice in the newly joined table as well. There are 8 columns because Customer_ID column is in both tables, 4 unique columns in customer table, 3 unique columns in order table, combined is 8 columns. I tried it with SQL and it returned 7 rows, A is not the answer, joined table can have duplicates.

upvoted 1 times

⊟ **Hiromi_Nakatani** 2 years, 1 month ago

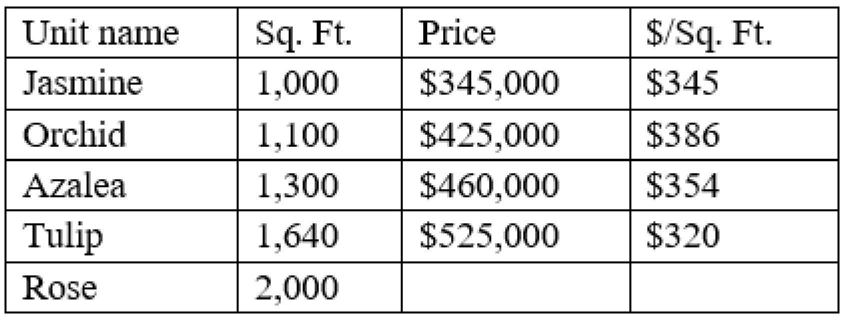
Selected Answer: B

B. Seven rows, eight columns

upvoted 4 times

Question #48 *Topic 1*

A development company is constructing a new unit in its apartment complex. The complex has the following floor plans:



Unit name	Sq. Ft.	Price	$/Sq. Ft.
Jasmine	1,000	$345,000	$345
Orchid	1,100	$425,000	$386
Azalea	1,300	$460,000	$354
Tulip	1,640	$525,000	$320
Rose	2,000		

Using the average cost per square foot of the original floor plans, which of the following should be the price of the Rose unit?

A. $640,900

B. $690,000

C. $705,200

D. $702,500

Correct Answer: *D*

Community vote distribution

D (100%)

Crimsonpug94 6 months, 3 weeks ago

Selected Answer: D

Answer is D.

Average Cost per Square Foot = (345 + 386 + 354 + 320) = 1405 / 4 = 351.25

Now that we have the average cost per square foot as $351.25, we can calculate the price of the "Rose" unit:

Price of Rose Unit = Square Footage of Rose Unit (2,000 sq. ft) * Average Cost per Square Foot ($351.25/sq. ft)

Price of Rose Unit = 2,000 sq. ft * $351.25/sq. ft

Price of Rose Unit = $702,500

upvoted 4 times

ronniehaang 1 year, 1 month ago

Selected Answer: D

The Rose unit has 1,700 square feet. To calculate the price, we need to find the average cost per square foot of the original floor plans and then multiply that by the square footage of the Rose unit.

To find the average cost per square foot of the original floor plans:

(650,000 + 675,000 + 700,000 + 725,000 + 775,000) / 5 = 700,000

So, the average cost per square foot is $700,000 / 5 = $140.

The price of the Rose unit would be $140 x 1,700 = $238,000.

Therefore, the answer is D. $702,500

upvoted 1 times

Hiromi_Nakatani 1 year, 1 month ago

Selected Answer: D

D. $702,500

upvoted 1 times

Question #49 *Topic 1*

Which of the following is a control measure for preventing a data breach?

A. Data transmission

B. Data attribution

C. Data retention

D. Data encryption

Correct Answer: *D*

db965b8 3 months, 4 weeks ago

D. Data encryption is the correct answer.

upvoted 1 times

Question #50 *Topic 1*

A user receives a large custom report to track company sales across various date ranges. The user then completes a series of manual calculations for each date range. Which of the following should an analyst suggest so the user has a dynamic, seamless experience?

A. Create multiple reports, one for each needed date range.

B. Build calculations into the report so they are done automatically.

C. Add macros to the report to speed up the filtering and calculations process.

D. Create a dashboard with a date range picker and calculations built in.

Correct Answer: *B*

Community vote distribution

D (100%)

Adonisy 5 months, 1 week ago

Selected Answer: D

D is best answer

upvoted 1 times

Alice00085 1 year, 8 months ago

(D) A dashboard with a date range picker allows the user to select any desired date range easily, without the need for multiple reports or manual filtering. By building calculations directly into the dashboard, the user can avoid manual calculations for each date range

upvoted 2 times

AmyD 2 years, 1 month ago

D.

A dashboard with a date range picker would allow the user to dynamically select the desired date range, and have the calculations done automatically without any manual effort. This would provide a seamless and dynamic experience for the user.

upvoted 3 times

Question #51 *Topic 1*

A table in a hospital database has a column for patient height in inches and a column for patient height in centimeters. This is an example of:

A. dependent data.

B. duplicate data.

C. invalid data

D. redundant data

Correct Answer: *D*

⊟ **DaCulprito_1** 6 months, 2 weeks ago

D. Redundant Data. Both columns would be communicating the same information (albeit, in different units).

upvoted 2 times

Question #52 *Topic 1*

While reviewing survey data, a research analyst notices data is missing from all the responses to a single question. Which of the following methods would BEST address this issue?

A. Replace missing data.

B. Remove duplicate data.

C. Replace redundant data.

D. Remove invalid data.

Correct Answer: *C*

Community vote distribution

D (100%)

Manizha 4 months ago

Selected Answer: D

D is correct, since there is no right pattern as an answer for the question!

upvoted 1 times

Adonisy 5 months, 1 week ago

Selected Answer: D

A or D will the answer

upvoted 1 times

moreinva43 8 months, 4 weeks ago

Why is A not a valid answer? Replace all the missing data with something like question not answered

upvoted 2 times

Correct_Damage 1 year, 11 months ago

D. If there is no data at all there, remove it

upvoted 2 times

Question #53 *Topic 1*

Which of the following BEST describes standard deviation?

A. A measure that is used to establish a relationship between two variables

B. A measure of how data is distributed

C. A measure of the amount of dispersion of a set of values

D. A measure that is used to find the significant difference between variables

Correct Answer: *C*

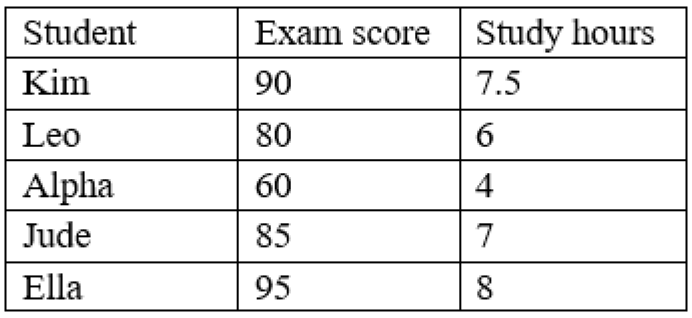
Correct_Damage 5 months, 3 weeks ago

C. Standard Deviation measures the amount of dispersion of a set of values

upvoted 2 times

Question #54 *Topic 1*

A data analyst was asked to create a chart that shows the relationship between study hours and exam scores for each student using the data sets in the table below:



Student	Exam score	Study hours
Kim	90	7.5
Leo	80	6
Alpha	60	4
Jude	85	7
Ella	95	8

Which of the following charts would BEST represent the relationship between the variables?

A. A histogram

B. A scatter plot

C. A heat map

D. A bar chart

Correct Answer: *B*

Community vote distribution

B (100%)

mcgoogol 7 months ago

two variables, possibly related = scatter plot

upvoted 1 times

ronniehaang 1 year, 7 months ago

Selected Answer: B

B. A scatter plot would BEST represent the relationship between the variables of study hours and exam scores for each student. A scatter plot shows the relationship between two variables by plotting their values in a two-dimensional graph. Each data point in the scatter plot represents a student's study hours and exam score. The scatter plot allows the analyst to see if there is a positive or negative relationship between the variables, and if there is a linear or nonlinear relationship. It provides a visual representation of the data, making it easier to understand and interpret.

upvoted 3 times

Hiromi_Nakatani 1 year, 7 months ago

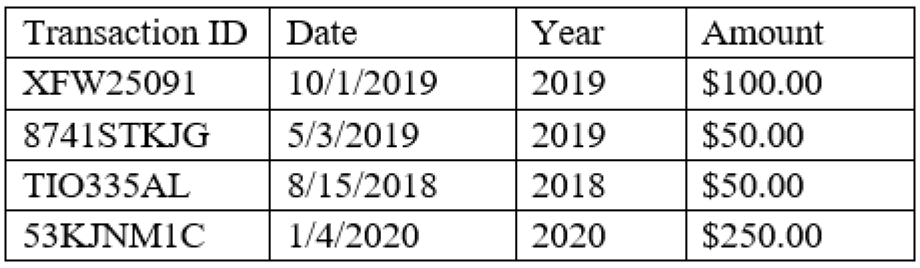
Selected Answer: B

B. A scatter plot

upvoted 2 times

Question #55 *Topic 1*

Given the table below:



Transaction ID	Date	Year	Amount
XFW25091	10/1/2019	2019	$100.00
8741STKJG	5/3/2019	2019	$50.00
TIO335AL	8/15/2018	2018	$50.00
53KJNM1C	1/4/2020	2020	$250.00

Which of the following variable types BEST describes the “Year” column?

A. Numeric

B. Date

C. Alphanumeric

D. Text

Correct Answer: *B*

Community vote distribution

A (100%)

Alice00085 Highly Voted 1 year, 2 months ago

A. Numeric

This is because the values in the "Year" column are all numbers representing the year, and they do not contain any alphabetic characters or special characters that would classify them as alphanumeric or text. Additionally, they are not in a date format that would require them to be classified as a date variable type.

upvoted 5 times

Ace_Defective Most Recent 2 months, 3 weeks ago

Selected Answer: A

Numeric

upvoted 1 times

Adonisy 5 months, 1 week ago

A.Numeric,this is basic,not date

upvoted 1 times

jimoland66 11 months ago

A, Numeric

As they are represented numerically and not in a date format.

upvoted 3 times

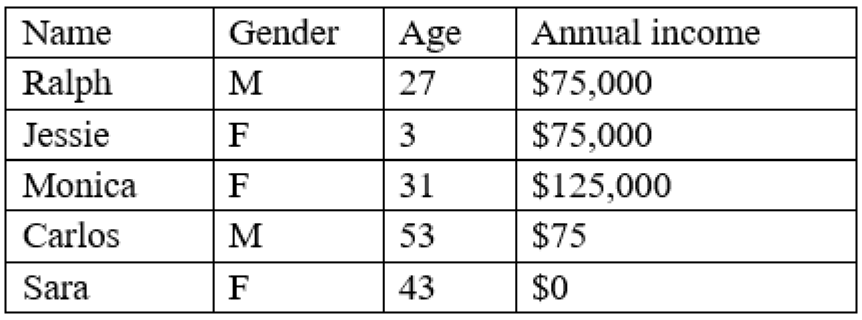
mcgoogol 1 year, 1 month ago

it is a 4 digit number, a year on its own is not a date, so A numeric

upvoted 4 times

Question #56 *Topic 1*

Given the following data:



Name	Gender	Age	Annual income
Ralph	M	27	$75,000
Jessie	F	3	$75,000
Monica	F	31	$125,000
Carlos	M	53	$75
Sara	F	43	$0

Which of the following BEST describes the data set?

A. There is data bias.

B. The data is incomplete.

C. The data is inconsistent.

D. The data is outliers.

Correct Answer: *D*

Swift_and_Quick 4 months, 3 weeks ago

D. There are outliers like 0 as age or 0 and 75 as salary. Outliers can present due to incorrect data entry. The data is not incomplete because all fields don't have null value. There is no bias. Data inconsistency means fields with the same value are entered differently, for example, both "M" and "male" are used to represent the male gender, this isn't the case here since M and F are used consistently to represent male and female respectively.

upvoted 1 times

mcgoogol 7 months ago

without calculations, data is just data. All that can be reliably said is the the data is inconsistent

upvoted 1 times

Alice00085 8 months, 1 week ago

C. The data is inconsistent.

There are a few points of inconsistency:

Jessie is listed as being 3 years old with an annual income of $75,000, which is not typical as this age is not within working age.
Carlos is listed with an annual income of $75, which is unusually low and inconsistent with what one would expect for annual income data.
Sara is listed with an annual income of $0, which may be accurate but is atypical and could be considered an outlier or require further context.
These inconsistencies suggest that there may be errors or irregularities in the data collection or recording process.

upvoted 2 times

willsy 1 year, 1 month ago

i would argue the data is incomplete as a 3 year old is earning 75,000

upvoted 3 times

Question #57 *Topic 1*

An analysts building a monthly report for production and wants to ensure the audience is aware of its once-a-month cadence. Which of the following is the MOST important to convey that information?

A. The date of the dashboard build

B. The data refresh date

C. A report summary

D. Frequently asked questions

Correct Answer: *B*

Correct_Damage 5 months, 3 weeks ago

B. So the readers or the report can get an idea of when the information will be updated

upvoted 1 times

Question #58 *Topic 1*

An analyst is working with the income data of suburban families in the United States. The data set has a lot of outliers, and the analyst needs to provide a measure that represents the typical income. Which of the following would BEST fulfill the analyst's goal?

A. Median

B. Mean

C. Mode

D. Standard deviation

Correct Answer: *A*

Community vote distribution

A (100%)

ronniehaang Highly Voted 1 year, 1 month ago

Selected Answer: A

The best measure that would fulfill the analyst's goal is the Median. The median represents the middle value of a data set when the data is sorted in ascending or descending order. It is less sensitive to outliers and provides a better representation of typical values compared to the mean, which is affected by outliers. The mode represents the most frequently occurring value in a data set, but it may not provide a good representation of typical values. The standard deviation measures the spread of a data set, but it does not provide an average value.

In this case, since the data set has a lot of outliers, the median would provide a better representation of the typical income compared to the mean, which would be affected by the outliers.

upvoted 5 times

willsy Most Recent 7 months, 3 weeks ago

I would argue that because their are numerous outliers that would skew the data, the median is what this analyst wants the most.

upvoted 2 times

Hiromi_Nakatani 1 year, 1 month ago

Selected Answer: A

A. Median

upvoted 4 times

Question #59 *Topic 1*

Which of the following would be used to store unstructured data from different sources?

A. A data lake

B. A database management system

C. A database

D. A data warehouse

Correct Answer: *A*

Community vote distribution

A (100%)

vesen22 4 months, 2 weeks ago

Selected Answer: A

A data lake

upvoted 1 times

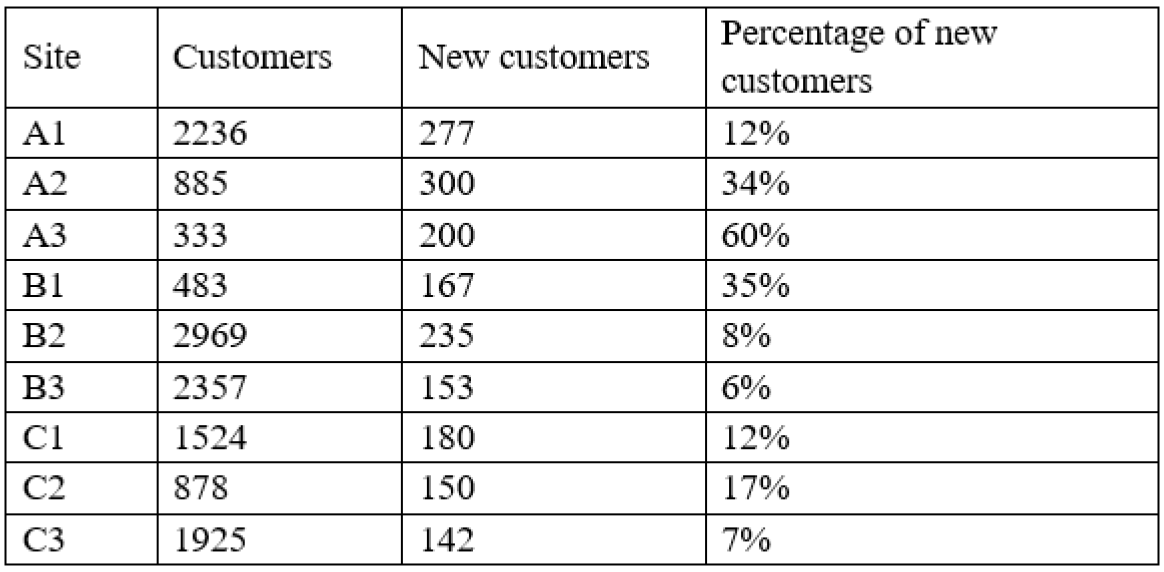
Correct_Damage 11 months, 3 weeks ago

A. Data lake

upvoted 2 times

Question #60 *Topic 1*

An analyst is designing a dashboard to determine which site has the highest percentage of new customers. The analyst must choose an appropriate chart to include in the dashboard. The following data is available:



Site	Customers	New customers	Percentage of new customers
A1	2236	277	12%
A2	885	300	34%
A3	333	200	60%
B1	483	167	35%
B2	2969	235	8%
B3	2357	153	6%
C1	1524	180	12%
C2	878	150	17%
C3	1925	142	7%

Which of the following types of charts should be considered to BEST display the data?

A. Include a bar chart using the site and the percentage of new customers data.

B. Include a line chart using the site and the percentage of new customers data.

C. Include a pie chat using the site and percentage of new customers data.

D. Include a scatter chart using the site and the percent of new customers data.

Correct Answer: *A*

Correct_Damage 5 months, 3 weeks ago

A. the bar chart would be better then a pie chart (usually used for percent) because, the data set shows percent increases not percent of a whole.

upvoted 2 times

Question #61 *Topic 1*

A cereal manufacturer wants to determine whether the sugar content of its cereal has increased over the years. Which of the following is the appropriate descriptive statistic to use?

A. Frequency

B. Percent change

C. Variance

D. Mean

Correct Answer: *D*

Community vote distribution

B (100%)

Norm141 4 months, 3 weeks ago

Selected Answer: B

B. To determine whether the sugar content of the cereal has increased over the years, the appropriate descriptive statistic to use is percent change. Percent change allows you to compare the difference in the sugar content between different years as a percentage of the initial value.

upvoted 1 times

Swift_and_Quick 11 months ago

B. To determine whether the sugar content of the cereal has increased over the years, the appropriate descriptive statistic to use is percent change. Percent change allows you to compare the difference in the sugar content between different years as a percentage of the initial value.

upvoted 1 times

Alice00085 1 year, 8 months ago

(B). To determine whether the sugar content of the cereal has increased over the years, the appropriate descriptive statistic to use is percent change. Percent change allows you to compare the difference in the sugar content between different years as a percentage of the initial value.

upvoted 2 times

Correct_Damage 1 year, 11 months ago

C. Mean could be skewed by outliers

upvoted 1 times

Question #62 *Topic 1*

The process of performing initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization is called:

A. a t-test.

B. a performance analysis.

C. an exploratory data analysis.

D. a link analysis.

Correct Answer: *C*

Swift_and_Quick 4 months, 3 weeks ago

C. Exploratory data analysis (EDA) involves using statistical methods, visualization techniques, and data manipulation to gain insights into the characteristics of the dataset. It helps analysts understand the structure of the data, identify potential problems or anomalies, and generate hypotheses for further investigation.

upvoted 1 times

Question #63 *Topic 1*

Different people manually type a series of handwritten surveys into an online database. Which of the following issues will MOST likely arise with this data? (Choose two.)

A. Data accuracy

B. Data constraints

C. Data attribute limitations

D. Data bias

E. Data consistency

F. Data manipulation

Correct Answer: *AE*

Community vote distribution

AE (100%)

db965b8 3 months, 4 weeks ago

My cousin and I said AE.

upvoted 1 times

Correct_Damage 1 year, 5 months ago

Selected Answer: AE

Consistency and accuracy

upvoted 3 times

Question #64 *Topic 1*

Which of the following data sampling methods involves dividing a population into subgroups by similar characteristics?

A. Systematic

B. Simple random

C. Convenience

D. Stratified

Correct Answer: *D*

Community vote distribution

D (100%)

vesen22 4 months, 2 weeks ago

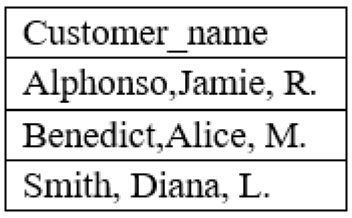
Selected Answer: D

Stratified

upvoted 1 times

Question #65 *Topic 1*

A data analyst must separate the column shown below into multiple columns for each component of the name:



Customer_name
Alphonso,Jamie, R.
Benedict,Alice, M.
Smith, Diana, L.

Which of the following data manipulation techniques should the analyst perform?

A. Imputing

B. Transposing

C. Parsing

D. Concatenating

Correct Answer: *C*

db965b8 3 months, 4 weeks ago

Parsing separates the data and Imputing value is empty that is fulled with EST value.

upvoted 1 times

Swift_and_Quick 4 months, 3 weeks ago

C. Parsing involves breaking down a single column that contains combined or composite data into multiple columns based on specific delimiters or patterns. In this case, the analyst would parse the name column to extract each component of the name (e.g., first name, last name) into separate columns. This allows for more granular analysis and manipulation of the data based on its individual components.

upvoted 1 times

Question #66 *Topic 1*

Which of the following descriptive statistical methods are measures of central tendency? (Choose two.)

A. Mean

B. Minimum

C. Mode

D. Variance

E. Correlation

F. Maximum

Correct Answer: *AC*

Swift_and_Quick 4 months, 3 weeks ago

AC. Mean and mode are measure of central tendency.

upvoted 1 times

Question #67 *Topic 1*

Which of the following will MOST likely be streamed live?

A. Machine data

B. Key-value pairs

C. Delimited rows

D. Flat files

Correct Answer: *D*

Swift_and_Quick 4 months, 3 weeks ago

A. Machine data refers to data generated by various machines, devices, or sensors in real-time. Examples include logs from servers, sensor readings from IoT devices, or telemetry data from industrial equipment. Machine data is often streamed live as it is generated, allowing for real-time analysis, monitoring, and decision-making.

upvoted 3 times

mcgoogol 7 months ago

There are many weather stations situated around the world, which measure and record and data and then instantly upload it to the internet. I would argue that this is machine data and is best looked at in real time

upvoted 3 times

Correct_Damage 1 year, 5 months ago

Flat file databases store plain text records and binary files that are needed for a specific purpose in a single directory for easy access and transfer. I'm going to go with D

upvoted 1 times

Question #68 *Topic 1*

A database consists of one fact table that is composed of multiple dimensions. Depending on the dimension, each one can be represented by a denormalized table or multiple normalized tables. This structure is an example of a:

A. transactional schema.

B. star schema.

C. non-relational schema.

D. snowflake schema.

Correct Answer: *A*

JOH22 4 months, 1 week ago

B. A snowflake schema is a database schema design where the fact table is linked to multiple dimensions, and those dimension tables can be further normalized into multiple related tables. This results in a structure that resembles a snowflake, with dimensions branching out into sub-dimensions.

upvoted 1 times

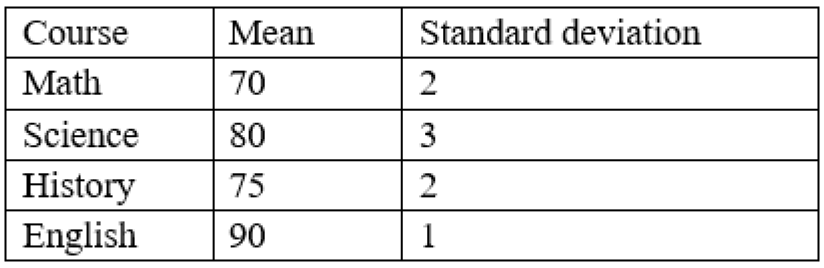
Swift_and_Quick 11 months, 2 weeks ago

The answer is B. It's star schema because there are multiple dimension tables branch out of the fact table. It didn't mention whether or not there are other dimension tables that branch out of the dimension tables that are connected to the fact table, so we can assume that it's not snowflake schema.

upvoted 1 times

Question #69 *Topic 1*

Randy scored 76 on a math test, Katie scored 86 on a science test, Ralph scored 80 on a history test, and Jean scored 80 on an English test. The table below contains the mean and standard deviation of the scores for each of the courses:



Course	Mean	Standard deviation
Math	70	2
Science	80	3
History	75	2
English	90	1

Using this information, which of the following students had the BEST score?

A. Randy

B. Katie

C. Ralph

D. Jean

Correct Answer: *B*

Swift_and_Quick 4 months, 3 weeks ago

A. Randy.

For Randy in math:
Z-score = (76 - 70) / 2 = 3
For Katie in science:
Z-score = (86 - 80) / 3 = 2
For Ralph in history:
Z-score = (80 - 75) / 2 = 2.5
For Jean in English:
Z-score = (80 - 90) / 1 = -10

Randy has the highest Z-score.

upvoted 2 times

Correct_Damage 1 year, 5 months ago

Also : observed - mean divided by standard deviation = the same answer B

upvoted 2 times

yandieg 1 year, 4 months ago

so, according to your formula, is A. Randy
For Randy in math:
z_math = (76 - 70) / 2 = 3

For Katie in science:
z_science = (86 - 80) / 3 = 2

For Ralph in history:
z_history = (80 - 75) / 2 = 2.5

For Jean in English:
z_english = (80 - 90) / 1 = -10

upvoted 3 times

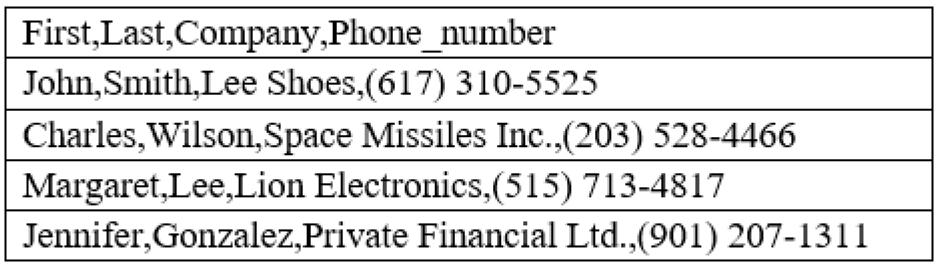
Correct_Damage 1 year, 5 months ago

B. Kind of think this is a trick question. Assuming the test score could be between 0-100, the highest score would be the best scare regardless of the subject the test was in or how well other students not in the table did.

upvoted 1 times

Question #70 *Topic 1*

Given the data below:



First,Last,Company,Phone_number
John,Smith,Lee Shoes,(617) 310-5525
Charles,Wilson,Space Missiles Inc.,(203) 528-4466
Margaret,Lee,Lion Electronics,(515) 713-4817
Jennifer,Gonzalez,Private Financial Ltd.,(901) 207-1311

In which of the following file formats is the data presented?

A. XLS

B. CSV

C. RTF

D. XML

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

CSV, values are separated by comma.

upvoted 1 times

Question #71 *Topic 1*

A commissions analyst has just completed second-quarter payout statements. The analyst created a dashboard to distribute the statements as quickly as possible. Which of the following will ensure only designated employees can view this information?

A. Publish it on the company intranet.

B. Provide individual links to recipients.

C. Grant subscription access.

D. Print individual dashboards.

Correct Answer: *B*

⊟ **db965b8** 3 months, 3 weeks ago

B. I agree.

upvoted 1 times

⊟ **Swift_and_Quick** 5 months, 2 weeks ago

B. Providing individual links to recipients ensures that only designated employees can view the information. Each recipient will have a unique link, which limits access to only those who have been specifically designated to receive the payout statements. This method helps maintain confidentiality and ensures security by controlling access to the dashboard.

upvoted 1 times

Question #72 *Topic 1*

The current date is July 14, 2020. A data analyst has been asked to create a report that shows the company's year-over-year Q2 2020 sales. Which of the following reports should the analyst compare?

A. Q2 2020 and Q4 2019

B. YTD 2020 and YTD 2019

C. Q2 2020 and Q2 2019

D. Q2 2020 and Q2 2021

Correct Answer: *B*

Community vote distribution

C (100%)

ducko 2 months ago

Selected Answer: C

C) is correct. Q2 is April to June, and you are comparing last years quarter with this years.

B) is incorrect because 2020 has not finished yet and you are comparing sales for a full year compared to half the year.

A) Is incorrect because you are not comparing like quarters, which could have seasonal variances.

D) Is incorrect because 2021 hasn't even started yet so there is no data.

upvoted 1 times

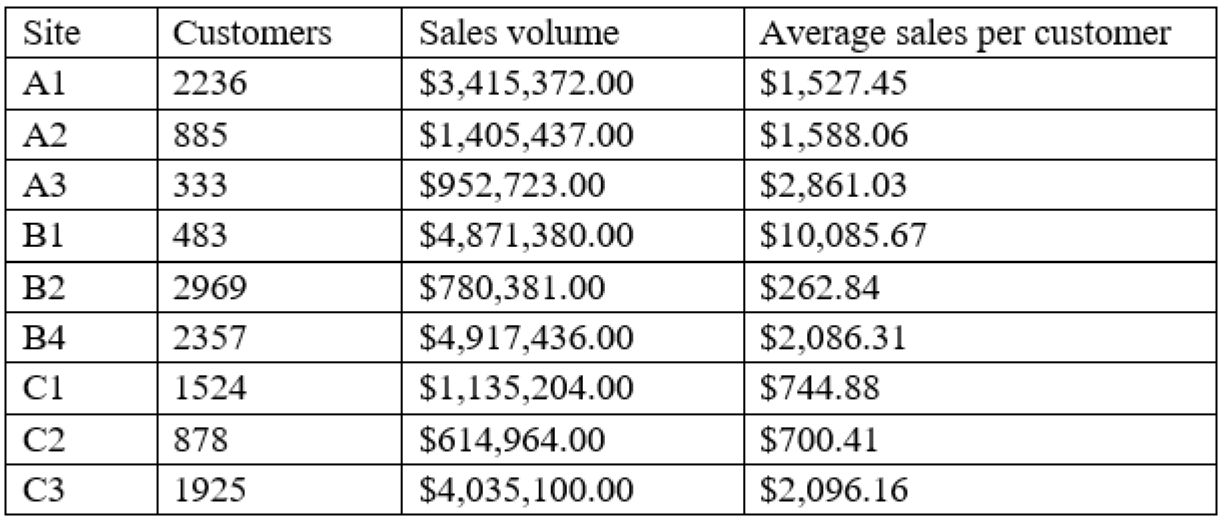
Swift_and_Quick 5 months, 2 weeks ago

To create a report that shows the company's year-over-year Q2 2020 sales, the analyst should compare the sales data from Q2 2020 and Q2 2019. Year-over-year (YoY) analysis is a method of comparing the performance of a business or a financial instrument over the same period in different years. It helps to identify trends, growth patterns, and seasonal fluctuations. Q2 refers to the second quarter of a year, which is usually from April to June. Therefore, the correct answer is C.

upvoted 4 times

Question #73 *Topic 1*

A data analyst is designing a dashboard that will provide a story of sales and determine which site is providing the highest sales volume per customer. The analyst must choose an appropriate chart to include in the dashboard. The following data is available:



Site	Customers	Sales volume	Average sales per customer
A1	2236	$3,415,372.00	$1,527.45
A2	885	$1,405,437.00	$1,588.06
A3	333	$952,723.00	$2,861.03
B1	483	$4,871,380.00	$10,085.67
B2	2969	$780,381.00	$262.84
B4	2357	$4,917,436.00	$2,086.31
C1	1524	$1,135,204.00	$744.88
C2	878	$614,964.00	$700.41
C3	1925	$4,035,100.00	$2,096.16

Which of the following types of charts should be considered?

A. Include a line chart using the site and average sales per customer.

B. Include a pie chart using the site and sales to average sales per customer.

C. Include a scatter chart using sales volume and average sales per customer.

D. Include a column chart using the site and sales to average sales per customer.

Correct Answer: *C*

Community vote distribution

D (100%)

Norm141 4 months, 3 weeks ago

Selected Answer: D

D. A column chart would show which site has the highest average. There's no additional dimensionality that would necessitate a scatter chart. Moreover, the outliers for site B1 would make a scatter plot difficult to read. I first checked ChatGPT, it agreed with me. Then, I built the table myself and played with building the charts in Excel. A column chart is far easier to read and correctly shows that B1 has the highest sales average per customer.

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

C. A scatter chart would allow you to visualize the relationship between sales volume and average sales per customer for each site. Each point on the scatter chart would represent a site, with the x-axis representing sales volume and the y-axis representing average sales per customer. This chart type is effective for identifying patterns and outliers in the data and can help determine which site is providing the highest sales volume per customer.

upvoted 1 times

Question #74 *Topic 1*

An analyst needs to conduct a quick analysis. Which of the following is the FIRST step the analyst should perform with the data?

A. Conduct an exploratory analysis and use descriptive statistics.

B. Conduct a trend analysis and use a scatter chart.

C. Conduct a link analysis and illustrate the connection points.

D. Conduct an initial analysis and use a Pareto chart.

Correct Answer: *A*

⊟ **Swift_and_Quick** 5 months, 2 weeks ago

A. Exploratory analysis helps in understanding the basic structure of the data, identifying patterns, detecting outliers, and gaining insights into its distribution. Descriptive statistics provide summary measures that describe the main features of the dataset, such as mean, median, mode, range, and standard deviation. This step lays the foundation for further analysis and helps in making informed decisions about which analytical methods and visualizations would be most appropriate for the given data.

upvoted 1 times

Question #75 *Topic 1*

A data analyst has been asked to create a sales report that calculates the rolling 12-month average for sales. If the report will be published on November 1, 2020, which of the following months shouts the report cover?

A. October 1, 2019 to October 31, 2020

B. October 31, 2020 to November 1, 2021

C. November 1, 2019 to October 31, 2020

D. October 31, 2019 to October 31, 2020

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C.

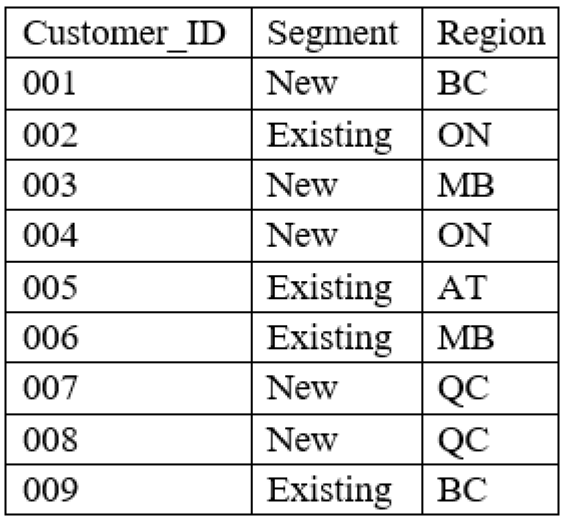
To calculate the rolling 12-month average for sales that will be published on November 1, 2020, the report should cover the preceding 12 months leading up to that date. This means the report should cover the period from November 1, 2019, to October 31, 2020.

upvoted 1 times

Question #76 *Topic 1*

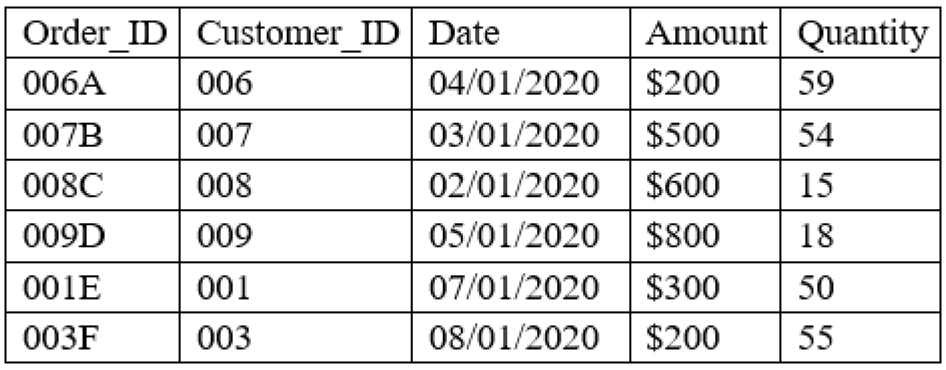
A data analyst has been asked to merge the tables below, first performing an INNER JOIN and then a LEFT JOIN:

Customer Table -



Customer_ID	Segment	Region
001	New	BC
002	Existing	ON
003	New	MB
004	New	ON
005	Existing	AT
006	Existing	MB
007	New	QC
008	New	QC
009	Existing	BC

In-store Transactions -



Order_ID	Customer_ID	Date	Amount	Quantity
006A	006	04/01/2020	$200	59
007B	007	03/01/2020	$500	54
008C	008	02/01/2020	$600	15
009D	009	05/01/2020	$800	18
001E	001	07/01/2020	$300	50
003F	003	08/01/2020	$200	55

Which of the following describes the number of rows of data that can be expected after performing both joins in the order stated, considering the customer table as the main table?

A. INNER: 6 rows; LEFT: 9 rows

B. INNER: 9 rows; LEFT: 6 rows

C. INNER: 9 rows; LEFT: 15 rows

D. INNER: 15 rows; LEFT: 9 rows

Correct Answer: *D*

Community vote distribution

A (100%)

willsy 7 months, 3 weeks ago

A - INNER joins the ones with the same in both, there are only 6. A

upvoted 1 times

ronniehaang 1 year, 1 month ago

A. INNER: 6 rows; LEFT: 9 rows

Explanation:

INNER JOIN: An INNER JOIN returns only the rows that match in both tables. The common column used for joining is "Customer ID". In this case, there are 6 customers with matching Customer IDs in both the Customer Table and In-store Transactions table, so the number of rows after performing an INNER JOIN would be 6.

LEFT JOIN: A LEFT JOIN returns all the rows from the left table (Customer Table), and the matching rows from the right table (In-store Transactions). If there is no match, NULL values will be displayed in the right table columns. In this case, there are 9 customers in the Customer Table, but only 6 have matching Customer IDs in the In-store Transactions table, so there would be 9 rows after performing a LEFT JOIN.

upvoted 3 times

Hiromi_Nakatani 1 year, 1 month ago

Selected Answer: A

A. INNER: 6 rows; LEFT: 9 rows

upvoted 2 times

Question #77 *Topic 1*

A data analyst needs to create a weekly recurring report on sales performance and distribute it to all sales managers. Which of the following would be the BEST method to automate and ensure successful delivery for this task?

A. Use scheduled report delivery.

B. Implement subscription access delivery.

C. Print out a copy.

D. Upload the report to the server.

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. Use scheduled report delivery.

Scheduled report delivery allows you to set up automated delivery of the report at specified intervals, such as weekly in this case. This method ensures that the report is delivered consistently and reliably without manual intervention. It can be configured to send the report directly to the email inboxes of all sales managers or to a designated location for them to access. This approach saves time and effort and ensures that the sales managers receive the report promptly on a regular basis.

upvoted 1 times

Question #78 *Topic 1*

Which of the following is an example of a discrete variable?

A. The temperature of a hot tub

B. The height of a horse

C. The time to complete a task

D. The number of people in an office

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

D. Discrete numbers are represented by whole numbers.

upvoted 1 times

Question #79

Topic 1

Which of the following data types would a telephone number formatted as XXX-XXX-XXXX be considered?

A. Numeric

B. Date

C. Float

D. Text

Correct Answer: *D*

Community vote distribution

D (100%)

daksa 1 week, 4 days ago

Selected Answer: D

Phone Number is String, coz we cannot perform any mathematical operation that we do on Numeric data type.

upvoted 1 times

3489 7 months, 2 weeks ago

Selected Answer: D

A telephone number formatted as XXX-XXX-XXXX would be considered a text data type, as it is composed of alphanumeric characters and symbols. A numeric data type is composed of only numbers, such as integers or decimals. A date data type is composed of values that represent dates or times, such as YYYY-MM-DD or HH:MM:SS. A float data type is composed of numbers with fractional parts, such as 3.14 or 0.5.

upvoted 2 times

EscCode 8 months, 1 week ago

why not string there is dash(-) is like zip code xx-xxx, am i right???

upvoted 1 times

Question #80 *Topic 1*

SIMULATION -

The director of operations at a power company needs data to help identify where company resources should be allocated in order to monitor activity for outages and restoration of power in the entire state. Specifically, the director wants to see the following:

* County outages

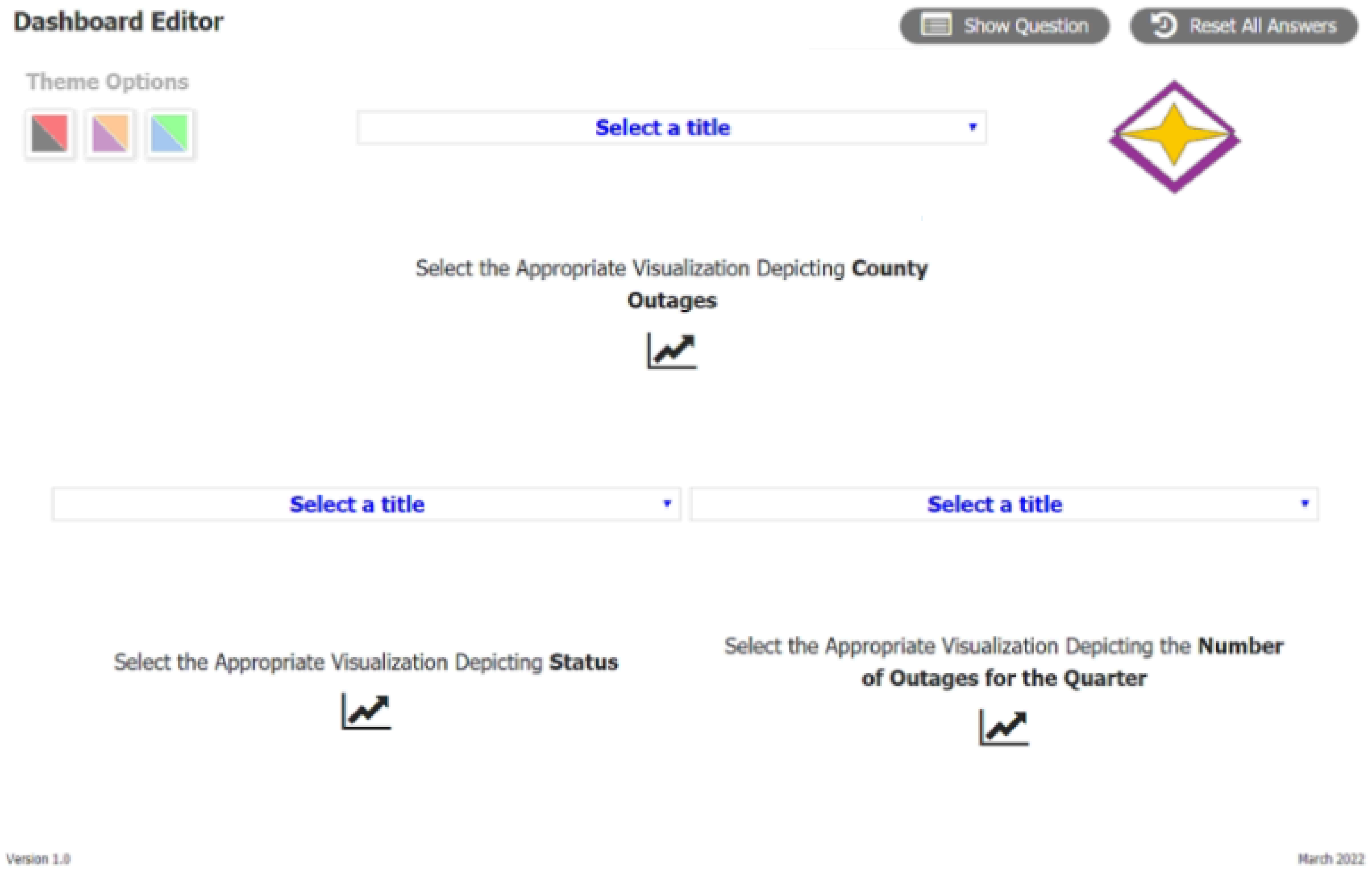
* Status

* Overall trend of outages

INSTRUCTIONS:

Please, select each visualization to fit the appropriate space on the dashboard and choose an appropriate color scheme. Once you have selected all visualizations, please, select the appropriate titles and labels, if applicable. Titles and labels may be used more than once.

If at any time you would like to bring back the initial state of the simulation, please click the Reset All button.

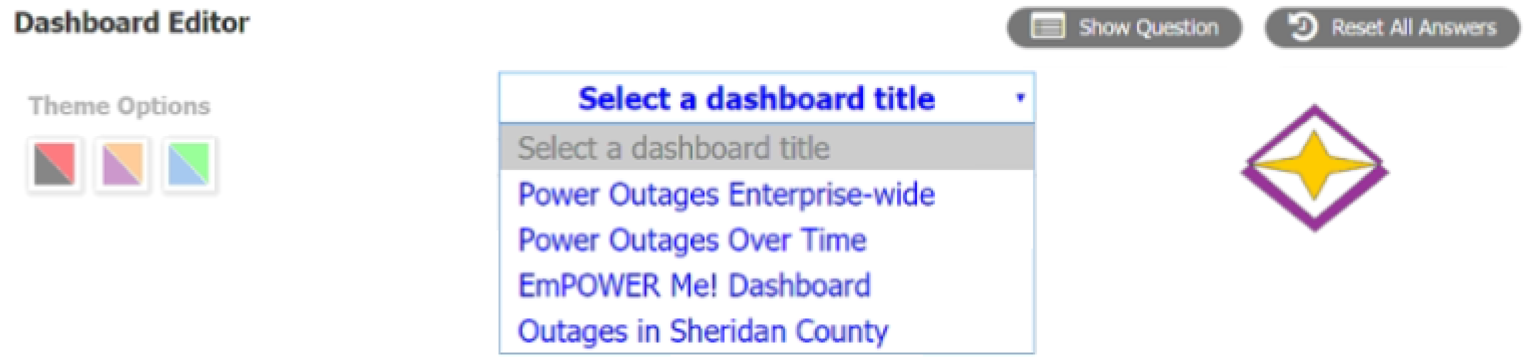


Dashboard Editor

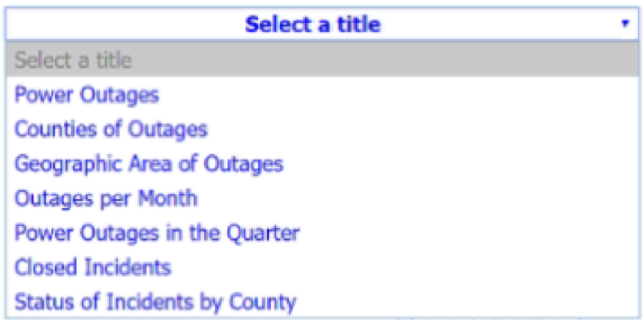
Select the Appropriate Visualization Depicting **County Outages**

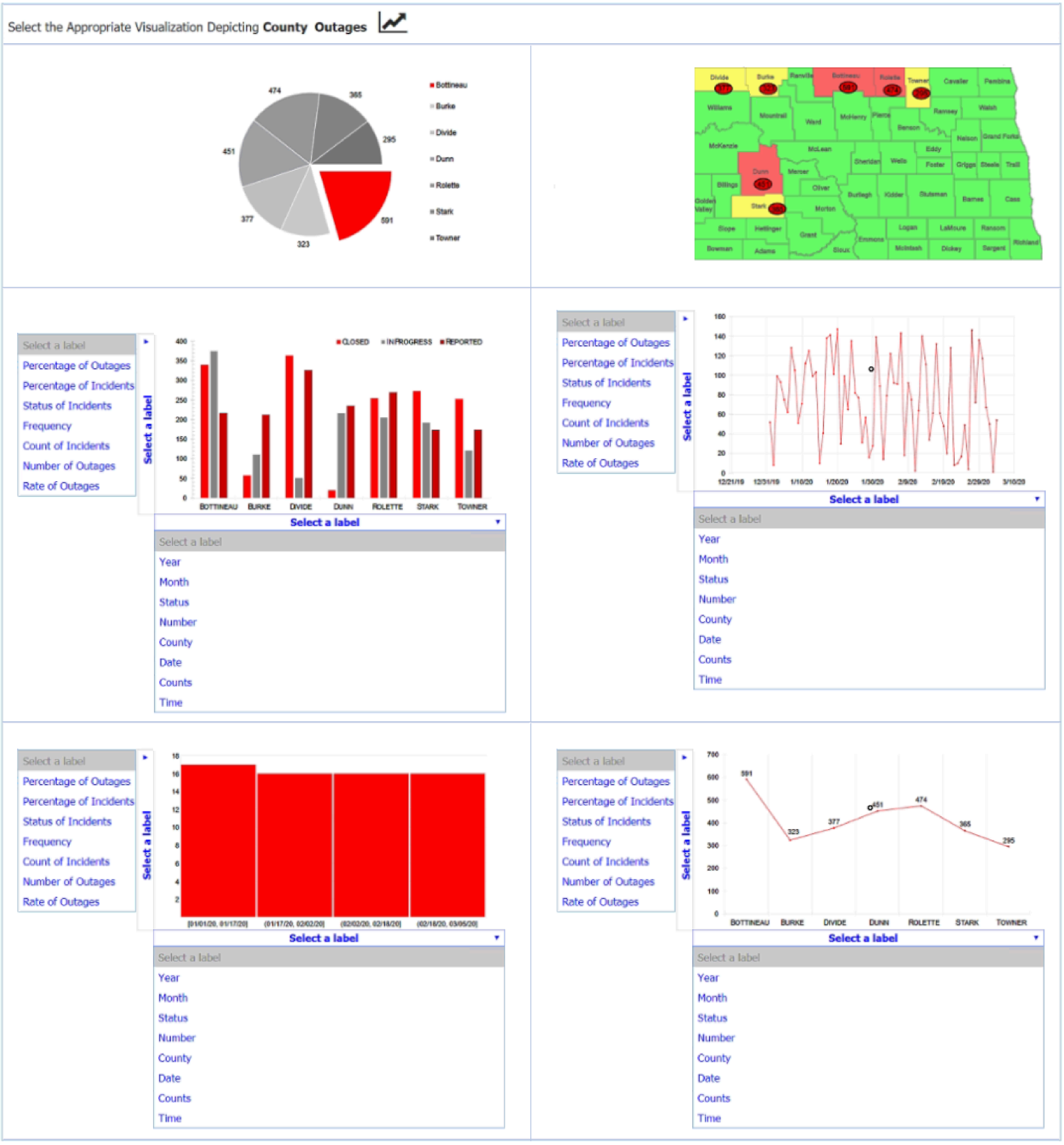
Select the Appropriate Visualization Depicting **Status**

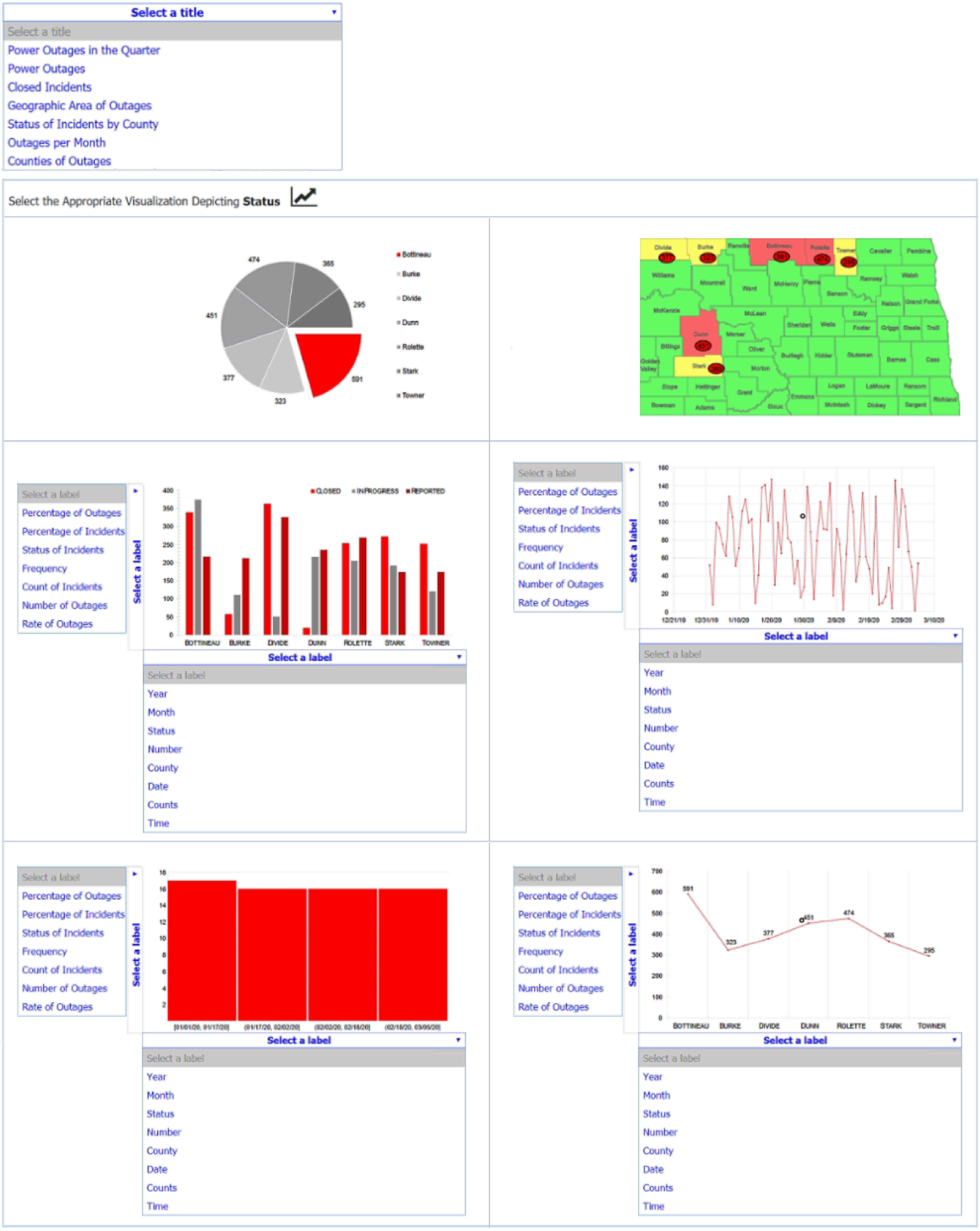
Select the Appropriate Visualization Depicting the **Number of Outages for the Quarter**

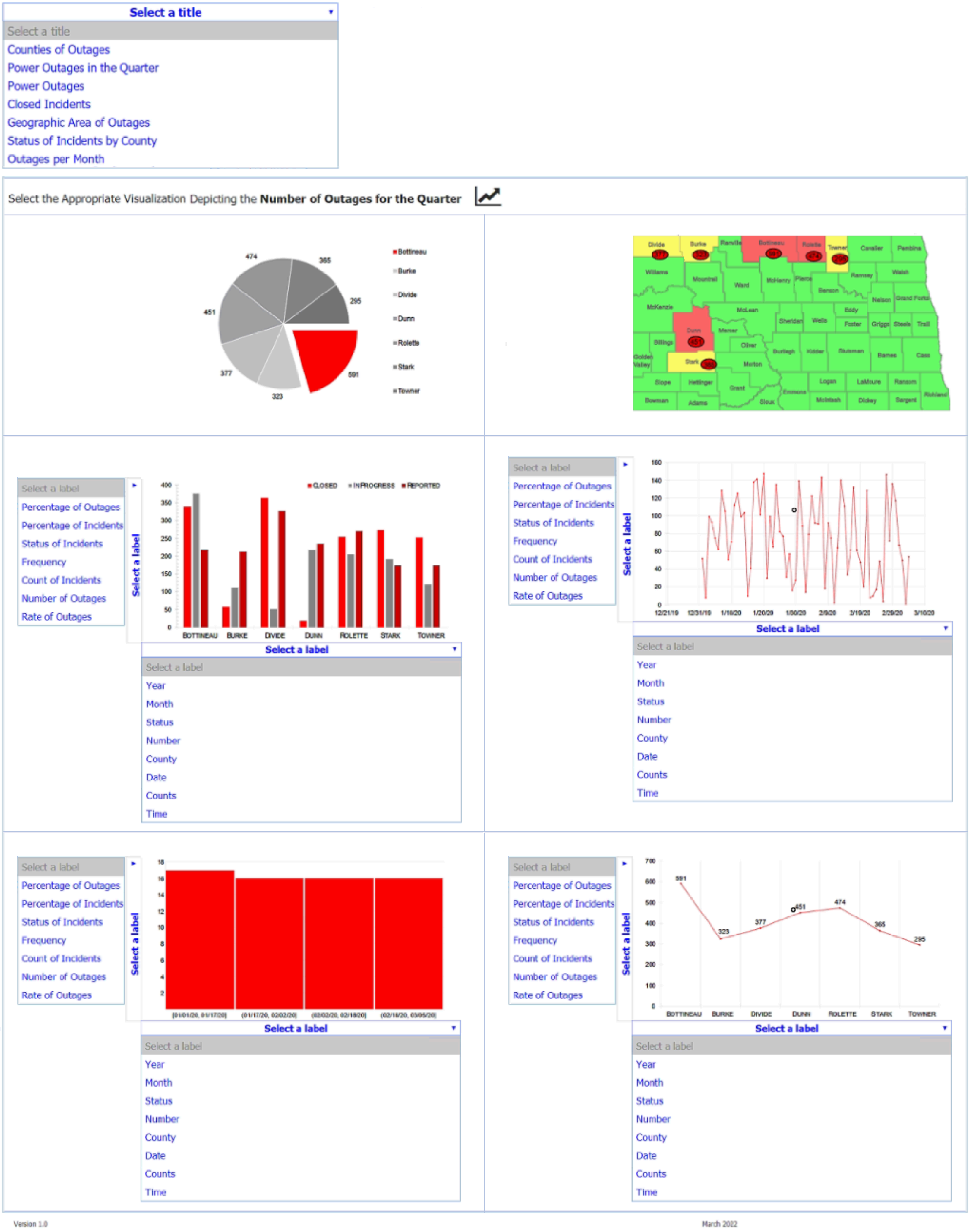


Dashboard Editor




Select the Appropriate Visualization Depicting **County Outages**


Select a title
Select a title
Power Outages in the Quarter
Power Outages
Closed Incidents
Geographic Area of Outages
Status of Incidents by County
Outages per Month
Counties of Outages
Select the Appropriate Visualization Depicting Status
474
365
295
451
377
323
591
Bottineau
Burke
Divide
Dunn
Rolette
Stark
Towner
Divide
377
Burke
323
Renville
Bottineau
591
Rolette
474
Towner
295
Cavalier
Pembina
Williams
Mountrail
Ward
McHenry
Pierce
Benson
Ramsey
Walsh
Nelson
Grand Forks
McKenzie
McLean
Eddy
Sheridan
Wells
Foster
Griggs
Steele
Traill
Dunn
451
Mercer
Billings
Oliver
Burleigh
Kidder
Stutsman
Barnes
Cass
Golden Valley
Stark
365
Morton
Slope
Hettinger
Grant
Logan
LaMoure
Ransom
Emmons
Richland
Bowman
Adams
Sioux
McIntosh
Dickey
Sargent
Select a label
Percentage of Outages
Percentage of Incidents
Status of Incidents
Frequency
Count of Incidents
Number of Outages
Rate of Outages
Select a label
CLOSED
INPROGRESS
REPORTED
400
350
300
250
200
150
100
50
0
BOTTINEAU
BURKE
DIVIDE
DUNN
ROLETTE
STARK
TOWNER
Select a label
Select a label
Year
Month
Status
Number
County
Date
Counts
Time
Select a label
Percentage of Outages
Percentage of Incidents
Status of Incidents
Frequency
Count of Incidents
Number of Outages
Rate of Outages
Select a label
160
140
120
100
80
60
40
20
0
12/21/19
12/31/19
1/10/20
1/20/20
1/30/20
2/9/20
2/19/20
2/29/20
3/10/20
Select a label
Select a label
Year
Month
Status
Number
County
Date
Counts
Time
Select a label
Percentage of Outages
Percentage of Incidents
Status of Incidents
Frequency
Count of Incidents
Number of Outages
Rate of Outages
Select a label
18
16
14
12
10
8
6
4
2
[01/01/20, 01/17/20)
[01/17/20, 02/02/20)
[02/02/20, 02/18/20)
[02/18/20, 03/05/20]
Select a label
Select a label
Year
Month
Status
Number
County
Date
Counts
Time
Select a label
Percentage of Outages
Percentage of Incidents
Status of Incidents
Frequency
Count of Incidents
Number of Outages
Rate of Outages
Select a label
700
600
500
400
300
200
100
0
591
323
377
451
474
365
295
BOTTINEAU
BURKE
DIVIDE
DUNN
ROLETTE
STARK
TOWNER
Select a label
Select a label
Year
Month
Status
Number
County
Date
Counts
Time



Select the Appropriate Visualization Depicting the **Number of Outages for the Quarter**

Correct Answer: *Power outages*

⊟ **Swift_and_Quick** 5 months, 2 weeks ago

Drag and drop the visualization that shows the overall trend of outages on the bottom space of the dashboard. This visualization is a line graph that shows the number of outages over time. You can choose any color scheme that suits your preference, but make sure that the color is visible and contrasted with the background. For example, you can use blue for the line and white for the background.

Select appropriate titles and labels for each visualization. Titles and labels may be used more than once. For example, you can use “County Outages” as the title for the map, “Status” as the title for the pie chart, and “Trend” as the title for the line graph. You can also use “County”, “Number of Outages”, “Active”, “Restored”, “Pending”, “Time”, and “Number of Outages” as labels for the axes and legends of the visualizations.

upvoted 1 times

⊟ **Swift_and_Quick** 5 months, 2 weeks ago

This is a simulation question that requires you to create a dashboard with visualizations that meet the director's needs. Here are the steps to complete the task:

Drag and drop the visualization that shows the county outages on the top left space of the dashboard. This visualization is a map of the state with different colors indicating the number of outages in each county. You can choose any color scheme that suits your preference, but make sure that the colors are consistent and clear. For example, you can use a gradient of red to show the counties with more outages and green to show the counties with less outages.

Drag and drop the visualization that shows the status of the outages on the top right space of the dashboard. This visualization is a pie chart that shows the percentage of outages that are active, restored, or pending. You can choose any color scheme that suits your preference, but make sure that the colors are distinct and easy to identify. For example, you can use red for active, green for restored, and yellow for pending.

upvoted 1 times

⊟ **12345678910** 1 year, 5 months ago

I took the exam on the 20 th of April 2023

upvoted 1 times

⊟ **12345678910** 1 year, 5 months ago

What are the correct Vizualations for each diagram and what is the correct labels for each diagram space there can only be 3.

*Country Outages
*Number of outages for a quarter
*Status

This is definitely in the exam!!!!!

upvoted 1 times

⊟ **7380698** 6 months, 2 weeks ago

Idk the answer. I had to do this on the real one I just don't know if I got the visualizations right lol

upvoted 1 times

⊟ **7380698** 6 months, 2 weeks ago

Yes, this is the same exam that I took. What is the answer tho to this?

upvoted 1 times

⊟ **Bongi12** 1 year, 7 months ago

Please my someone explain well on what is supposed to be done on this PBQ

upvoted 1 times

⊟ **Boats** 1 year, 11 months ago

The PBQ has only three charts you select and provide labels for.

upvoted 2 times

Question #81 *Topic 1*

An analyst has conducted a review of business questions. Which of the following should the analyst do next to conduct an analysis?

A. Determine the data needs and review the observations.

B. Determine the data needs and sources for analysis.

C. Determine the data needs and schedule interviews.

D. Determine the data needs and begin the analysis.

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. Determine the data needs and sources for analysis.

Before beginning the analysis, the analyst should identify what data is necessary to address the business questions and where this data can be obtained. This involves determining the specific variables, metrics, and information required for the analysis, as well as identifying potential data sources such as internal databases, external sources, or conducting interviews or surveys. Once the data needs and sources are established, the analyst can then proceed with collecting the necessary data and begin the analysis process.

upvoted 1 times

Question #82 *Topic 1*

A data analyst is compiling a report that a Chief Executive Officer needs for an impromptu meeting. The report should include information on the previous day’s performance. Which of the following reports should the analyst provide?

A. Tactical

B. Ad hoc

C. Dynamic

D. Recurring

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. Ad hoc

An ad hoc report is exactly what the CEO needs for the impromptu meeting, as it's prepared for a specific purpose or occasion and is often created on short notice. Since the CEO requires information on the previous day's performance for this particular meeting, an ad hoc report would be the most appropriate choice to fulfill this immediate need.

upvoted 1 times

Question #83 *Topic 1*

Which of the following tools would be best to use to calculate the interquartile range, median, mean, and standard deviation of a column in a table that has 5,000,000 rows?

A. Microsoft Excel

B. R

C. Snowflake

D. SQL

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. R would likely be the best choice for calculating the interquartile range, median, mean, and standard deviation of a column in a table with 5,000,000 rows, due to its efficiency, scalability, and comprehensive statistical capabilities.

upvoted 1 times

Question #84 *Topic 1*

Which of the following data acquisition concepts is the most appropriate for determining the quality of a company's product?

A. Web scraping

B. Observation

C. Survey

D. Sampling

Correct Answer: *C*

stephyfresh13 4 months, 2 weeks ago

B. Observation
By directly observing the product in use or during the production process, analysts can gather detailed and accurate information about its quality, performance, and any potential issues.

upvoted 1 times

moreinva43 8 months, 4 weeks ago

Isn't sampling often used to determine product quality?

upvoted 1 times

db965b8 10 months ago

I would say C too.

upvoted 1 times

db965b8 10 months, 3 weeks ago

Would say C. Survey

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

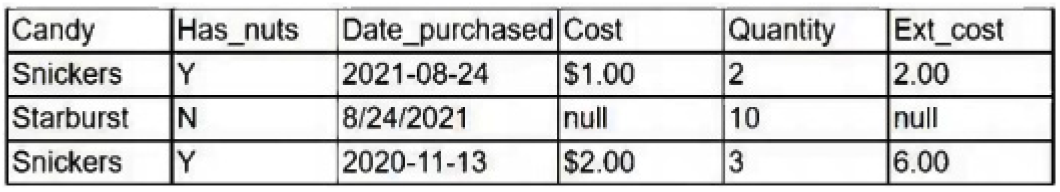
B. o Observation involves directly observing the product in real-world contexts, such as during manufacturing processes, usage, or interactions with customers. This method allows for firsthand assessment of product quality, identifying any defects, issues, or areas for improvement. Observation can provide valuable insights into the actual performance and characteristics of the product, making it a powerful tool for evaluating quality.

While surveys can also provide valuable feedback on product quality, they may rely on participants' perceptions and interpretations, which can be subjective and influenced by various factors. Additionally, surveys may not capture the full range of user experiences and may be limited to the specific questions asked. Overall, both observation and surveys have their strengths and limitations, and the choice between them should be based on the specific research objectives and constraints.

upvoted 1 times

Question #85 *Topic 1*

Given the following:



Candy	Has_nuts	Date_purchased	Cost	Quantity	Ext_cost
Snickers	Y	2021-08-24	$1.00	2	2.00
Starburst	N	8/24/2021	null	10	null
Snickers	Y	2020-11-13	$2.00	3	6.00

Which of the following is the most important thing for an analyst to do when transforming the table for a trend analysis?

A. Fill in the missing cost where it is null.

B. Separate the table into two tables and create a primary key.

C. Replace the extended cost field with a calculated field.

D. Correct the dates so they have the same format.

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

D. Correcting the dates so they have the same format is the most important thing for an analyst to do when transforming the table for a trend analysis. Trend analysis is a method of analyzing data over time to identify patterns, changes, or relationships. To perform a trend analysis, the data needs to have a consistent and comparable format, especially for the date or time variables.
In the example, the date purchased column has two different formats: YYYY-MM-DD and MM/DD/YYYY.
This could cause errors or confusion when sorting, filtering, or plotting the data over time. Therefore, the analyst should correct the dates so they have the same format, such as YYYY-MM-DD, which is a standard and unambiguous format.

upvoted 1 times

Question #86 *Topic 1*

Which of the following is a difference between a primary key and a unique key?

A. A unique key cannot take null values, whereas a primary key can take null values.

B. There can be only one primary key in a data set, whereas there can be multiple unique keys.

C. A primary key can take a value more than once, whereas a unique key cannot take a value more than once.

D. A primary key cannot be a date variable, whereas a unique key can be.

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. In a given table, only one primary key can be defined, but multiple unique keys can be defined to enforce uniqueness on different columns.

upvoted 1 times

Question #87 *Topic 1*

Which of the following best describes how discrete data differs from continuous data?

A. Discrete data cannot create a sloped line.

B. Discrete data can only be a finite number of values.

C. Discrete data can have decimal points.

D. Discrete data applies only to numbers.

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. Discrete data consists of distinct, separate values with no possible values in between. These values are often countable and finite, such as integers. In contrast, continuous data can take on an infinite number of possible values within a given range, including decimal points. Therefore, option B accurately captures the distinction between discrete and continuous data.

upvoted 1 times

Question #88 *Topic 1*

Which of the following file formats is best suited to start exploratory analysis within statistical software?

A. CSV

B. XLSM

C. XML

D. JSON

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. CSV (Comma-Separated Values)

CSV files are plain text files that store tabular data, with each line representing a row and commas separating the values within each row. CSV files are widely supported by statistical software and are easy to import into programs like R, Python (with libraries like pandas), and many others. They are lightweight, simple to create, and can handle large datasets efficiently, making them ideal for exploratory data analysis (EDA).

While other file formats like XLSM (Excel Macro-Enabled Workbook), XML (eXtensible Markup Language), and JSON (JavaScript Object Notation) have their uses, CSV is typically preferred for initial exploratory analysis due to its simplicity, ease of use, and compatibility with statistical software.

upvoted 1 times

Question #89 *Topic 1*

An analyst wants to model the relationship between a set of continuous variables in order to predict the value of an output variable based on the values of a set of input variables.

Which of the following types of analyses should the analyst use?

A. Chi-squared

B. Correlation

C. Regression

D. t-test

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. Regression.
The analyst should use regression analysis. Regression analysis is a statistical technique used to model the relationship between a dependent variable (output variable) and one or more independent variables (input variables). It helps in understanding how the value of the dependent variable changes when one or more independent variables are varied.

upvoted 1 times

Question #90 *Topic 1*

A data analyst needs to create a dashboard using the company's yearly revenue data sets.

Which of the following would be the best way to plot the information to show the top performing region?

A. A line chart

B. A waterfall chart

C. A heat map

D. A stacked bar chart

Correct Answer: *D*

Norm141 4 months, 3 weeks ago

ChatGPT says D, stacked bar chart. But wouldn't it be a heat map? If we want to know sales performance by region, it sounds like we're talking about a geographical area. A stacked bar chart breaks the columns/bars into separate portions. How would that show revenue by region? A pie chart would make more sense in that regard. But since that's not an option, wouldn't a heat map make more sense?

upvoted 1 times

db965b8 10 months ago

A stacked bar chart is a type of bar chart that portrays the compositions and comparisons of several variables through time. Stacked charts usually represent a series of bars or columns stacked on top of one another. They are widely used to effectively portray comparisons of total values across several categories.

upvoted 1 times

Question #91 *Topic 1*

The duration of a phone call in milliseconds is an example of:

A. ordinal data.

B. nominal data.

C. boolean data.

D. continuous data.

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

The duration of a phone call in milliseconds is an example of continuous data (Option D).

Continuous data can take any value within a certain range, and there are an infinite number of possible values between any two points. Milliseconds are a unit of measurement that can represent any fraction of a second, making the duration of a phone call a continuous variable.

upvoted 1 times

Question #92 *Topic 1*

Under which of the following circumstances should the null hypothesis be accepted when α = 0.05?

A. When p is 0.00003

B. When p is 0.001

C. When p is 0.04

D. When p is 0.06

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

D. p-value greater than 0.05 means null hypothesis can be accepted.

upvoted 1 times

Question #93 *Topic 1*

Which of the following is the first step an analyst should perform upon receiving a business request for analysis?

A. Determine the data needs and sources for analysis.

B. Initiate the analysis for exploratory data analysis.

C. Review the business questions to understand the scope.

D. Finalize the methodology to solve the problem.

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. While determining data needs and sources (Option A) is crucial, it typically comes after understanding the scope of the analysis. Initiating the analysis for exploratory data analysis (Option B) and finalizing the methodology (Option D) both come later in the analytical process, after the scope has been defined and data needs have been identified. Therefore, Option C is the most appropriate first step upon receiving a business request for analysis.

upvoted 1 times

Question #94 *Topic 1*

Which of the following data analytics tools are used to create advanced statistical visualizations? (Choose two).

A. SQL

B. Domo

C. Rapid mining

D. BusinessObjects

E. Stata

F. Apex

Correct Answer: *DE*

stephyfresh13 4 months, 2 weeks ago

The two data analytics tools that are best suited for creating advanced statistical visualizations are Domo and Stata. Domo is known for its powerful data visualization capabilities, and Stata is widely used for advanced statistical analysis and visualizations.

B,E

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

C & E.

Stata is a statistical software commonly used in academic and research fields for data analysis, including advanced statistical visualizations. RapidMiner (option C) is a data science platform that offers advanced statistical modeling capabilities and visualization tools to explore and present data insights effectively. It allows users to build predictive models and conduct complex analyses while visualizing the results in various formats.

upvoted 2 times

EscCode 1 year, 2 months ago

For me B,E

upvoted 2 times

Question #95 *Topic 1*

Joe, an analyst, tests the loading time on a dashboard he is preparing to go live and finds it is slower than he would like. Which of the following must occur to decrease the loading time?

A. Deploy the dashboard to production.

B. Change the field definitions.

C. Update the dashboard subscribers.

D. Optimize the dashboard.

Correct Answer: *D*

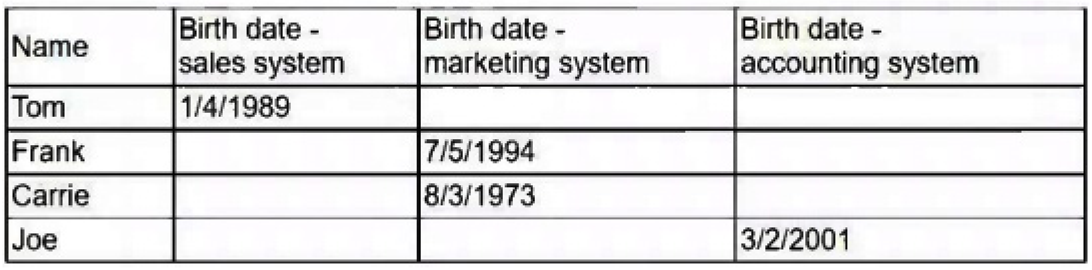
Swift_and_Quick 5 months, 2 weeks ago

D. Optimization decrease load time.

upvoted 1 times

Question #96 *Topic 1*

A data analyst has a set with more than 40,000 rows in the sample schema below:



Name	Birth date - sales system	Birth date - marketing system	Birth date - accounting system
Tom	1/4/1989		
Frank		7/5/1994	
Carrie		8/3/1973	
Joe			3/2/2001

The analyst would like to create one column that contains the customers' birth dates. Which of the following data quality dimensions would best explain the reason for compilation?

A. Data accuracy

B. Data completeness

C. Data duplication

D. Data integrity

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. Data completeness.

In this scenario, the birth dates are spread across multiple columns in different systems, resulting in incomplete information. Compiling the birth dates into a single column would enhance the completeness of the dataset by consolidating all available birth date information into one location.

upvoted 1 times

Question #97 *Topic 1*

A data analyst is creating a dashboard and trying to identify the type of information that should be included. Which of the following should the analyst consider first?

A. Data refresh rate

B. Consumer types

C. Access permissions

D. Data sources and attributes

Correct Answer: *B*

Community vote distribution

D (100%)

stephyfresh13 4 months, 2 weeks ago

The first thing an analyst should consider when creating a dashboard is B. Consumer types. Understanding who will be using the dashboard and what their needs are is crucial for determining what information should be included and how it should be presented.

upvoted 1 times

moreinva43 8 months, 4 weeks ago

Not that I like to agree with the given answer because for this test, it seems to be wrong more often than it is right, but don't we need to know who the dashboard is for first to see what data they need before we start looking at data sources and attributes.

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

D. Understanding the data sources and attributes is essential because it determines the availability and quality of data that can be used to populate the dashboard. By identifying the relevant data sources and attributes, the analyst can ensure that the dashboard includes accurate and meaningful information that aligns with the organization's goals and objectives. This foundational step lays the groundwork for designing an effective dashboard that provides valuable insights to the intended audience.

upvoted 1 times

khawnu 1 year ago

Selected Answer: D

It should be D

upvoted 2 times

Question #98 *Topic 1*

A sales analyst needs to report how the sales team is performing to target. Which of the following files will be important in determining 2019 performance attainment?

A. 2018 goal data

B. 2018 actual revenue

C. 2019 goal data

D. 2019 commission plan

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. This file would contain the targets or goals set for the sales team in 2019. By comparing the actual revenue or sales performance against these targets, the sales analyst can assess how well the sales team performed relative to their goals for that year. While files such as 2018 actual revenue (option B) provide historical performance data, it's the comparison against the 2019 goals that directly evaluates performance attainment for that specific year.

upvoted 1 times

Question #99 *Topic 1*

A data analyst is using a two-tailed, independent t-test to determine whether the type of stretching, dynamic or static, has any influence on a dancer’s flexibility. Which of the following is the alternative hypothesis?

A. A dancer’s flexibility is improved through static stretching.

B. The change in a dancer’s flexibility is not equal to zero.

C. There is a difference in a dancer’s flexibility between static and dynamic stretching.

D. The means of the static and dynamic stretching groups do not differ from each other.

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. In a two-tailed, independent t-test comparing the influence of two different types of stretching (dynamic and static) on a dancer's flexibility, the alternative hypothesis (H_1) would typically express the idea that there is a difference between the two groups. This hypothesis suggests that there is a statistically significant difference in flexibility between dancers who perform static stretching and those who perform dynamic stretching.

upvoted 1 times

Question #100 *Topic 1*

Which of the following is the best reason to use database views instead of tables?

A. Views reduce the need for repetitive, complex data joins.

B. Views allow for the storage of temporary data, whereas tables do not.

C. Views allow for the joining of multiple data sources, whereas tables do not.

D. Views can be used to restrict sensitive information.

Correct Answer: *D*

moreinva43 8 months, 4 weeks ago

Both A and D are valid, it is a guess to which Comptia thinks is BEST.

upvoted 1 times

Swift_and_Quick 9 months, 2 weeks ago

New answer: A.

Views can encapsulate complex joins and queries into a single virtual table-like structure. This means you can define a view once with the necessary joins and filters, and then reuse that view in queries without having to rewrite the complex join logic each time. This simplifies query writing, improves maintainability, and reduces the risk of errors.

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

D. Views can be used to restrict sensitive information.

While all the options have their own merits, option D stands out as a fundamental advantage of using views. Views allow you to control access to sensitive data by defining subsets of data from one or more tables and limiting access to certain columns or rows based on predefined criteria. This capability is crucial for enforcing data security and access control policies without compromising the integrity of the underlying data structure. It helps protect sensitive information from unauthorized access, ensuring data privacy and confidentiality.

upvoted 1 times

Question #101 *Topic 1*

Which of the following would a data analyst look for first if 100% participation is needed on survey results?

A. Missing data

B. Invalid data

C. Redundant data

D. Duplicate data

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. Missing data can significantly impact the reliability and validity of survey results. If there are any missing responses or incomplete records, achieving 100% participation becomes impossible. Therefore, identifying and addressing missing data is crucial to ensure that the survey results are complete and representative of the entire sample population.

upvoted 1 times

Question #102 *Topic 1*

A sales director has requested a report for individual team members within the division be developed. The director would like the report to be shared with all team members, but individual team members should not be identifiable within the report. Which of the following access requirements would support the director’s needs?

A. Create an acceptable use policy for the sales data.

B. Release the report as user-group-based access and include data masking.

C. Get a data use agreement from the individual team members.

D. Provide the report based on role and include data encryption.

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. Release the report as user-group-based access and include data masking.

This option ensures that the report is shared with all team members while maintaining their anonymity. By using user-group-based access, the report can be distributed to the entire team without revealing individual identities. Additionally, including data masking techniques ensures that sensitive information about individual team members is obscured or obfuscated, further protecting their privacy. This approach aligns with the director's request to share the report with all team members while keeping their identities confidential.

upvoted 1 times

Question #103 *Topic 1*

Which of the following is an example of PII?

A. Age

B. Name

C. Ethnicity

D. Gender

Correct Answer: *B*

⊟ **Swift_and_Quick** 4 months, 3 weeks ago

B. Name.

upvoted 1 times

Question #104 *Topic 1*

Which of the following is the best technique for transferring data from one database to another with some data manipulation?

A. Application programming interfaces

B. Delta load

C. Extract, transform, load

D. Export/import

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. Extract, transform, load (ETL).

ETL processes involve extracting data from the source database, transforming it according to the requirements (which may include data manipulation), and then loading it into the destination database. This approach allows for flexibility in manipulating the data during the transfer process, such as converting data types, aggregating values, or applying business rules.

While options like application programming interfaces (APIs) and export/import can also facilitate data transfer, they may not offer the same level of flexibility and data manipulation capabilities as ETL. Delta load is a technique used to synchronize changes between databases but may not necessarily involve data manipulation. Therefore, ETL is the most suitable option for transferring data between databases while performing data manipulation tasks.

upvoted 1 times

Question #105 *Topic 1*

Which of the following best describes a business analytics tool with interactive visualization and business capabilities and an interface that is simple enough for end users to create their own reports and dashboards?

A. Python

B. R

C. Microsoft Power BI

D. SAS

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

B. Power BI is a visualization tool that end-user use to create dashboards.

upvoted 1 times

> **Swift_and_Quick** 4 months, 3 weeks ago
>
> I meant C.
>
> upvoted 1 times

Question #106 *Topic 1*

A data analyst needs to create a data visualization that aids in understanding the cumulative impact of sequentially introduced values that are positive or negative. Which of the following data visualization methods should the analyst use?

A. A bubble chart

B. A waterfall chart

C. A scatter plot

D. A line chart

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. A waterfall chart is specifically designed to illustrate how positive and negative values contribute to a total over a series of sequential steps or categories. It shows the cumulative effect of each value, making it ideal for visualizing changes in a process, financial statement, or any scenario where values are sequentially introduced and impact a total. Therefore, a waterfall chart would best serve the data analyst's needs in this scenario.

upvoted 1 times

Question #107 *Topic 1*

Which of the following report types is most appropriate for a high-level, year-end report requested by a Chief Executive Officer?

A. Dynamic

B. Recurring

C. Ad hoc

D. Self-service

Correct Answer: *B*

Community vote distribution

C (100%)

927eb66 1 month, 3 weeks ago

Selected Answer: C

The most appropriate report type for a high-level, year-end report requested by a CEO is C. Ad hoc

Explanation:

Ad hoc reports: are designed for specific purposes, not regularly scheduled, and are often created to address unique questions or situations. A year-end CEO report typically needs to summarize key performance indicators and highlight significant achievements or areas for improvement, which aligns with the nature of an ad hoc report.

upvoted 1 times

JOH22 4 months, 3 weeks ago

C. Ad hoc

Explanation:

An ad hoc report is a customized, one-time report created to address specific information needs at a particular moment. Year-end reports are typically tailored to summarize the organization's performance over the past year, providing insights into achievements, challenges, and strategic directions. These reports are not part of a regular reporting schedule but are generated to meet the unique requirements of the CEO for annual review and planning.

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

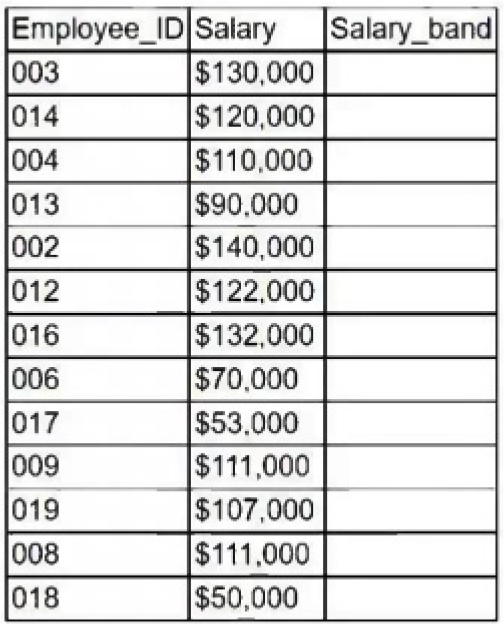
B. Recurring.

A recurring report is a regularly scheduled report that provides consistent and standardized information over time. A year-end report for a Chief Executive Officer typically involves summarizing key metrics, accomplishments, challenges, and strategic insights for the entire year. By scheduling it as a recurring report, the CEO can expect to receive it annually without the need for specific requests each time. This ensures that the CEO receives consistent, timely updates on the organization's performance and helps in strategic decision-making.

upvoted 1 times

Question #108 *Topic 1*

A company's human resources department has asked a data analyst to categorize the income of all employees into five salary bands:



Employee_ID	Salary	Salary_band
003	$130,000	
014	$120,000	
004	$110,000	
013	$90,000	
002	$140,000	
012	$122,000	
016	$132,000	
006	$70,000	
017	$53,000	
009	$111,000	
019	$107,000	
008	$111,000	
018	$50,000	

Which of the following types of functions would be the most appropriate to use?

A. Statistical

B. Aggregate

C. Logical

D. Mathematical

Correct Answer: *B*

Community vote distribution

C (100%)

MLadis 2 months ago

Selected Answer: C

The most appropriate type of function to use for categorizing salaries into bands is logical functions, as they allow the analyst to apply conditions to determine the appropriate salary band for each employee.

upvoted 1 times

Ace_Defective 2 months, 2 weeks ago

Selected Answer: C

The most appropriate type of functions to use in this scenario would be A. Statistical.

Here's why:

Categorization into salary bands often involves dividing the data into groups based on statistical measures like percentiles or standard deviations.

Statistical functions like percentile() or quantile() can be used to determine the salary boundaries for each band.

While aggregate functions might be used to calculate summary statistics for each band after they are created, they are not the primary tools for creating the bands themselves.

upvoted 1 times

Swift_and_Quick 5 months, 2 weeks ago

C. Logical functions are the most appropriate to use for categorizing data into bands, because they allow the data analyst to apply conditional statements and criteria to the data values. For example, the IF function can be used to assign a band name based on whether a value meets a certain condition or not. Other logical functions that can be useful for categorizing data are AND, OR, NOT, and IFERROR12.

upvoted 2 times

Question #109 *Topic 1*

Five dogs have the following heights in millimeters:

300, 430, 170, 470, 600

Which of the following is the standard deviation for the five dogs?

A. 147mm

B. 154mm

C. 394mm

D. 21,704mm

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. 147.

(300 + 430 + 170 + 470 + 600) / 5 = 394. Mean is 394.

sqrt(((300 - 394)^2 + (430 - 394)^2 + (170 - 394)^2 + (470 - 394)^2 + (600 - 394)^2) / 5)) = 147. Standard deviation is 147.

upvoted 1 times

Question #110 *Topic 1*

Which of the following is a non-parametric test?

A. One-sample t-test

B. Two-way ANOVA

C. Correlation coefficient

D. Spearman's rank correlation

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

D. Spearman's rank correlation.

Spearman's rank correlation coefficient is a non-parametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. Unlike the one-sample t-test, two-way ANOVA, and correlation coefficient, Spearman's rank correlation does not assume a specific distribution for the data and does not require the variables to be normally distributed. Therefore, it is considered a non-parametric test.

upvoted 1 times

Question #111 *Topic 1*

Which of the following best describes a 95% confidence interval?

A. There is a 95% probability that the population mean lies within the stated interval.

B. A stated range may contain 95% of the population mean, 95% of the time.

C. A set of ranges contains the population mean with 95% certainty.

D. A range contains 95% of the population mean.

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

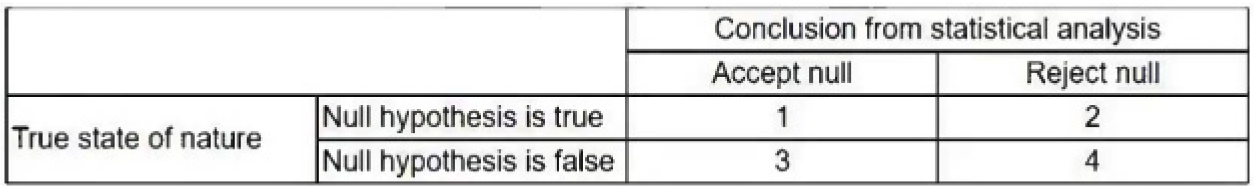
A. There is a 95% probability that the population mean lies within the stated interval.

A 95% confidence interval means that if we were to repeat the process of sampling and constructing confidence intervals, about 95% of those intervals would contain the true population mean. It does not imply that there is a 95% probability that any given interval specifically contains the population mean, but rather that 95% of such intervals constructed in this way would include the population mean.

upvoted 1 times

Question #112 *Topic 1*

Given the table below:



		Conclusion from statistical analysis	
		Accept null	Reject null
True state of nature	Null hypothesis is true	1	2
	Null hypothesis is false	3	4

Which of the following boxes indicates that a Type II error has occurred?

A. 1

B. 2

C. 3

D. 4

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. Type II error occurs when hypothesis is incorrect but gets accepted anyways, so box 3.

upvoted 1 times

Question #113 *Topic 1*

A user imports a data file into the accounts payable system each day. On a regular basis, the field input is not what the system is expecting, so it results in an error for the row and a broken import process. To resolve the issue, the user opens the file, finds the error in the row, and manually corrects it before attempting the import again. The import sometimes breaks on subsequent attempts, though.

Which of the following changes should be made to this process to reduce the number of errors?

A. Delete all incorrect inputs and upload the corrected file.

B. Have the user manually review the file for data completeness before loading it.

C. Create a data field to data type validator to run the file through prior to import.

D. Spot-check the file prior to import to catch and correct field errors.

Correct Answer: *C*

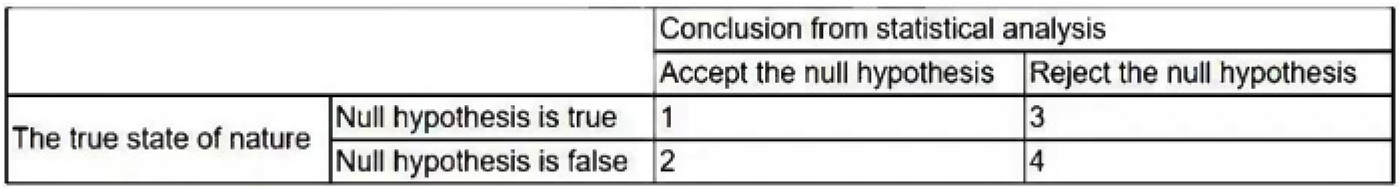
Swift_and_Quick 5 months, 2 weeks ago

C. A data field to data type validator is a tool or a process that checks if the data in each field of a file matches the expected data type, such as text, number, date, etc. A data field to data type validator can help to identify and correct any errors or inconsistencies in the data before importing it into the accounts payable system. This would reduce the number of errors and broken imports, as well as save time and effort for the user.

upvoted 1 times

Question #114 *Topic 1*

Given the table below:



		Conclusion from statistical analysis	
		Accept the null hypothesis	Reject the null hypothesis
The true state of nature	Null hypothesis is true	1	3
	Null hypothesis is false	2	4

Which of the following numbers represents a Type I error?

A. 1

B. 2

C. 3

D. 4

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. Type I error occurs when hypothesis is correct but gets rejected, so box 3.

upvoted 1 times

Question #115 Topic 1

A sales manager wants quarterly sales reports broken down by unit and week. Which of the following data output lists includes the most necessary information?

A. Order number, salesperson, date shipped, recipient address, and price

B. Item name, salesperson, recipient address, shipping cost, and date shipped

C. Item number, item name, salesperson, date sold, and price

D. Item name, salesperson, price, shipping cost, and date shipped

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

For quarterly sales reports broken down by unit and week, the most necessary information would include:

C. Item number, item name, salesperson, date sold, and price.

This option provides essential details such as the specific items sold (item number and name), who sold them (salesperson), when they were sold (date sold), and the price. These details are crucial for tracking sales performance, analyzing trends over time, and identifying opportunities for improvement. The other options do not include all the necessary information for generating comprehensive quarterly sales reports with unit and week breakdowns.

upvoted 1 times

Question #116 *Topic 1*

An analyst is preparing a report that contains weather data. The temperatures are shown in Fahrenheit, but they must be reported in Celsius. Which of the following should the analyst do to fix this issue?

A. Normalize the data.

B. Standardize the data.

C. Rescale the data.

D. Aggregate the data.

Correct Answer: *C*

mohammed25helal 5 months, 2 weeks ago

A. Normalize the data.
The term used in the book CompTIA Data+ Study Guide: Exam DAO-001 , Chapter 4 Data Quality, is Normalization. Normalizing data converts data from different scales to the same scale.

upvoted 1 times

NextTopic 11 months, 2 weeks ago

Don't it's rescaling, more like normalizing. tell me what you think

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

C. Rescale the data.

Rescaling involves converting the data from one scale to another, which is precisely what's needed in this scenario. Converting temperatures from Fahrenheit to Celsius involves a specific mathematical transformation, making rescaling the appropriate choice. Options like normalizing, standardizing, or aggregating the data do not address the need to convert temperature scales.

upvoted 2 times

Question #117 *Topic 1*

Which of the following statements would be used to append two tables that have the same number of columns?

A. UNION ALL

B. MERGE

C. GROUP BY

D. JOIN

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. UNION ALL.

The UNION ALL statement combines the results of two SELECT queries into a single result set by appending the rows of one table to the rows of another table. It's important to note that UNION ALL requires that the tables have the same number of columns and compatible data types.

upvoted 1 times

Question #118 *Topic 1*

A data analyst has been asked to create a daily manufacturing report for the floor manager. Which of the following metrics should be included in the report?

A. Tons of steel produced per hour

B. Annual sales budget

C. End-of-day stock price

D. Daily corporate employee count

Correct Answer: *A*

⊟ **Swift_and_Quick** 5 months, 2 weeks ago

A. A. Tons of steel produced per hour.

This metric directly relates to the manufacturing process and provides real-time information on the production output, which is essential for monitoring productivity, identifying potential issues, and making operational decisions on the factory floor. The other metrics (annual sales budget, end-of-day stock price, and daily corporate employee count) are less directly related to manufacturing operations and may not be as pertinent for the floor manager's daily activities.

upvoted 1 times

Question #119 *Topic 1*

A Chief Executive Officer (CEO) is requesting more up-to-date sales data for improved visibility prior to month-end. An analyst must determine the frequency of a sales report that was previously distributed on an as-needed basis.

Which of the following would be the most appropriate frequency for this report?

A. Monthly

B. Quarterly

C. Weekly

D. Every other month

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. Weekly.

A weekly frequency ensures that the CEO receives timely updates on sales performance throughout the month, allowing for better visibility and the opportunity to address any issues or capitalize on emerging trends in a more timely manner compared to monthly or quarterly reports. Additionally, weekly reports align well with the CEO's desire for up-to-date information before the end of the month.

upvoted 1 times

Question #120 *Topic 1*

A human resources analyst needs to build a new visualization to highlight the company’s hierarchy. Which of the following would be the best way to do this?

A. A stacked chart

B. An infographic

C. A word cloud

D. A tree map

Correct Answer: *B*

stephyfresh13 5 months ago

D - A tree map
Tree maps effectively represent hierarchical data by showing the relationships between different levels in a visually intuitive way. They can display multiple levels of hierarchy and the size of different elements, making them ideal for illustrating organizational structures.

upvoted 1 times

db965b8 10 months ago

I believe this would be a Tree Map....The treemap displays hierarchical (tree-structured) data as a set of nested rectangles. A rectangle represents each branch of the tree, which is then tiled with smaller rectangles representing subbranches.

upvoted 2 times

Question #121 *Topic 1*

Which of the following database schemas features normalized dimension tables?

A. Flat

B. Snowflake

C. Hierarchical

D. Star

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. Snowflake

In a snowflake schema, dimension tables are typically normalized, meaning that they are broken down into multiple related tables to reduce redundancy and improve data integrity. This normalization helps to organize the data efficiently and reduce storage space by avoiding unnecessary duplication of information. Therefore, among the options provided, the snowflake schema is the one that features normalized dimension tables.

upvoted 1 times

Question #122 *Topic 1*

A military commander would like to see the health scorecards of the troops daily and filter them based on gender and rank.

Considering this data is PHI, which of the following would be the best way for the commander to view the information?

A. An emailed report

B. A password-protected dashboard

C. A daily printout of a report

D. A cloud-hosted spreadsheet

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. A password-protected dashboard

A password-protected dashboard would provide secure access to the health scorecards while allowing the commander to filter the information based on gender and rank. This method ensures that only authorized individuals can access the data while still providing the necessary functionality for viewing and filtering the information. It also reduces the risk of unauthorized access or data breaches compared to emailing reports, printing reports, or using cloud-hosted spreadsheets, which may not provide the same level of security and access control.

upvoted 1 times

Question #123 *Topic 1*

Which of the following differentiates a flat text file from other data types?

A. Data is separated by a delimiter.

B. Data is stored in defined rows.

C. Data is defined with key-value pairs.

D. Data is housed in a markup language.

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. Data is separated by a delimiter.

Flat text files are characterized by having data that is typically stored as plain text and separated by delimiters such as commas, tabs, or spaces. Each record in a flat text file is typically represented as a single line, and the fields within each record are separated by the chosen delimiter. This distinguishes them from other data types where data might be stored in defined rows (like in databases), defined with key-value pairs (like in JSON or YAML files), or housed in a markup language (like HTML or XML).

upvoted 1 times

Question #124 *Topic 1*

Which of the following reports can be used when insight into operational performance is needed each Wednesday?

A. Static report

B. Tactical report

C. Recurring report

D. Ad hoc report

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

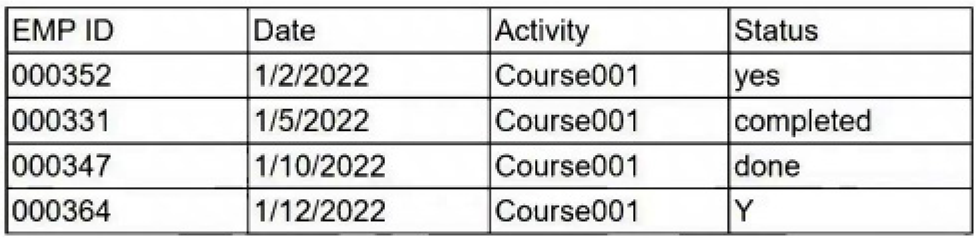
C. Recurring report

A recurring report is one that is generated on a regular schedule, such as weekly, monthly, or quarterly. Since the need is for insight into operational performance every Wednesday, a recurring report would be the most suitable option. This type of report ensures that the necessary information is consistently provided at the specified intervals, allowing for ongoing monitoring and analysis of operational performance.

upvoted 1 times

Question #125 *Topic 1*

During data profiling, an analyst decides to recode the status column in the following data set:



EMP ID	Date	Activity	Status
000352	1/2/2022	Course001	yes
000331	1/5/2022	Course001	completed
000347	1/10/2022	Course001	done
000364	1/12/2022	Course001	Y

Which of the following data concerns explains why the analyst wants to take this action?

A. Redundancy

B. Duplication

C. Invalidity

D. Inconsistency

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

D. Status column has values that mean the same thing but written in an inconsistent format.

upvoted 1 times

Question #126 *Topic 1*

Which of the following are reasons to conduct data cleansing? (Choose two).

A. To perform web scraping

B. To track KPIs

C. To improve accuracy

D. To review data sets

E. To increase the sample size

F. To calculate trends

Correct Answer: *CF*

stephyfresh13 4 months, 2 weeks ago

C. To improve accuracy: Data cleansing ensures that the data is correct, consistent, and reliable, which is crucial for making accurate decisions based on that data.

D. To review data sets: Data cleansing helps in identifying and correcting errors, inconsistencies, and redundancies in data sets, making them easier to review and analyze.

upvoted 1 times

Question #127 *Topic 1*

The number of phone calls that a call center receives in a day is an example of:

A. continuous data.

B. categorical data.

C. ordinal data.

D. discrete data.

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

D. Discrete data.

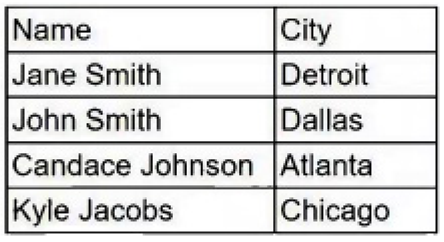
The number of phone calls that a call center receives in a day is an example of discrete data. Discrete data consists of separate, distinct values that can be counted and are often integers. In this case, the number of phone calls represents distinct, countable values, such as 0, 1, 2, 3, and so on, making it an example of discrete data.

upvoted 1 times

Question #128 *Topic 1*

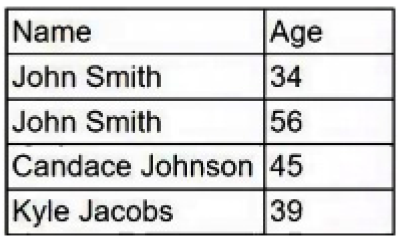
An analyst needs to join two tables of data together for analysis. All the names and cities in the first table should be joined with the corresponding ages in the second table, if applicable.

Table 1 -



Name	City
Jane Smith	Detroit
John Smith	Dallas
Candace Johnson	Atlanta
Kyle Jacobs	Chicago

Table 2 -



Name	Age
John Smith	34
John Smith	56
Candace Johnson	45
Kyle Jacobs	39

Which of the following is the correct join the analyst should complete, and how many total rows will be in one table?

A. INNER JOIN, two rows

B. LEFT JOIN, four rows

C. RIGHT JOIN, five rows

D. OUTER JOIN, seven rows

Correct Answer: *B*

Community vote distribution

B (100%)

07fd0a0 1 month, 3 weeks ago

Selected Answer: B

LEFT JOIN, five rows (not four).

Since option B incorrectly states four rows, the correct answer is missing from the choices.

upvoted 1 times

Question #129 *Topic 1*

Which of the following is the most likely reason for a data analyst to optimize a query using parameterization?

A. To return a subset of records

B. To insert a temporary table

C. To prevent SQL injections

D. To increase the query speed

Correct Answer: *C*

mohammed25helal 5 months, 2 weeks ago

D. To increase the query speed.

The real fact is that the use on Parameterization is definitely a security approach to prevent SQL injections. But this is not a DBA certifications. The term SQL Injection in not even mentioned in some books. Meanwhile, In Chapter 3 Databases and Data Acquisition, in Query Optimization, the first factor they mention is Parameterization, and quote: "Effective use of parameterization reduces the number of times the database has to parse individual queries".

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

C. To prevent SQL injections

Parameterization in query optimization involves using parameters instead of embedding values directly into the SQL statement. One of the primary reasons for parameterization is to prevent SQL injection attacks, where malicious SQL code is inserted into input fields by attackers. By parameterizing queries, input values are treated as data rather than executable code, reducing the risk of SQL injection vulnerabilities.

While improving query speed (option D) is also a potential benefit of parameterization, preventing SQL injections is typically the most critical reason for implementing parameterized queries in database applications.

upvoted 1 times

Question #130 *Topic 1*

A data analyst needs to collect a similar proportion of data from every state. Which of the following sampling methods would be the most appropriate?

A. Systematic sampling

B. Convenience sampling

C. Stratified sampling

D. Random sampling

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. Stratified sampling

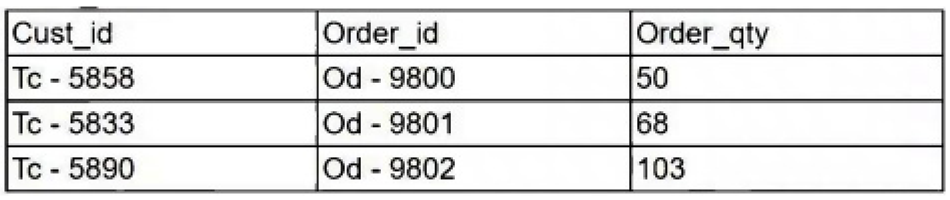
Stratified sampling involves dividing the population into homogeneous groups or strata based on certain characteristics (in this case, states) and then randomly selecting samples from each stratum. Since the goal is to collect a similar proportion of data from every state, stratified sampling would be the most appropriate method. This ensures that each state is represented in the sample proportionally to its size in the population, thus achieving a balanced representation across all states.

upvoted 1 times

Question #131 *Topic 1*

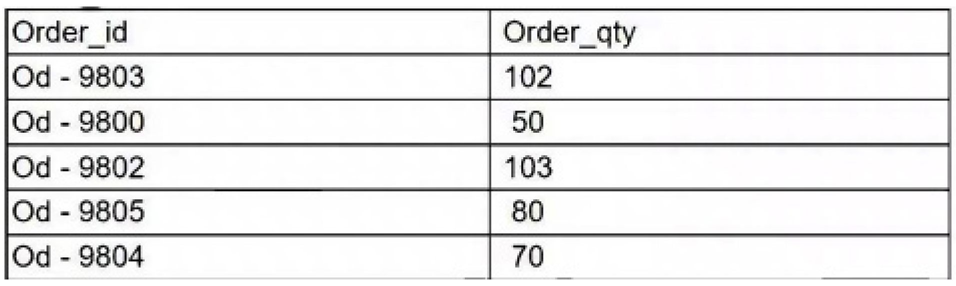
A data analyst needs to perform a full outer join of a customer's orders using the tables below:

Sales_table -



Cust_id	Order_id	Order_qty
Tc - 5858	Od - 9800	50
Tc - 5833	Od - 9801	68
Tc - 5890	Od - 9802	103

Order_table -



Order_id	Order_qty
Od - 9803	102
Od - 9800	50
Od - 9802	103
Od - 9805	80
Od - 9804	70

Which of the following is the mean of the order quantity?

A. 73.5

B. 76.5

C. 78.8

D. 81.5

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

When both tables are merged, there will be six order_qty values, 50 and 103 from both Sales_table and Order_table, 68 from Sales_table, 102, 80, and 70 from Order_table.

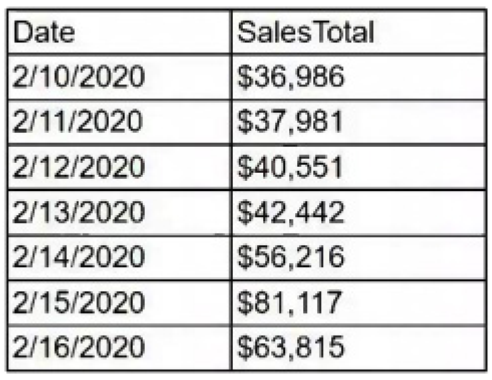
(50 + 68 + 103 + 102 + 80 + 70) / 6 = 78.8.

C is correct.

upvoted 1 times

Question #132 *Topic 1*

An analyst must obtain the average daily sales for the following week:



Date	SalesTotal
2/10/2020	$36,986
2/11/2020	$37,981
2/12/2020	$40,551
2/13/2020	$42,442
2/14/2020	$56,216
2/15/2020	$81,117
2/16/2020	$63,815

Which of the following must the analyst perform to obtain this value?

A. Data normalization

B. Data append

C. Data aggregation

D. Data blending

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. Data aggregation

To obtain the average daily sales for the following week, the analyst needs to aggregate the sales data for each day of the week and then calculate the average. Aggregation involves combining or summarizing data from multiple sources or rows into a single value, such as summing, averaging, or counting. In this case, the analyst would aggregate the daily sales data to calculate the total sales for each day, and then divide by the number of days in the week to find the average daily sales.

upvoted 1 times

Question #133 *Topic 1*

Which of the following best describes an exploratory analysis?

A. Involves the use of descriptive statistics to understand observations

B. Involves analysis of exploring data sets for performance tracking

C. Involves the testing of specific hypotheses

D. Involves the use of arithmetic algebra to determine the distribution

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. Involves the use of descriptive statistics to understand observations

Exploratory analysis focuses on gaining insights and understanding the characteristics of a dataset without preconceived notions or hypotheses. Descriptive statistics, such as measures of central tendency, dispersion, and visualization techniques, are commonly used to summarize and explore the data's properties. Therefore, option A best describes exploratory analysis.

upvoted 1 times

Question #134 *Topic 1*

A site reliability team wants to monitor the stability of their website, so they can proactively diagnose issues when they occur. Which of the following deliverables would best suit their needs?

A. A self-serve dashboard of website performance that updates in real time

B. A weekly log report of site visits and user actions

C. A portal that is refreshed daily and reports errors classified by type

D. A daily summary email indicating website outages for the previous day

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. A self-serve dashboard of website performance that updates in real time

A self-serve dashboard of website performance that updates in real time would best suit the needs of the site reliability team. This dashboard would provide real-time visibility into the website's performance metrics, allowing the team to monitor the stability of the site continuously and proactively diagnose issues as they occur. It enables quick identification of potential problems, allowing for timely intervention and resolution to minimize any impact on users.

upvoted 1 times

Question #135 *Topic 1*

Which of the following query optimization techniques involves examining only the data that is needed for a particular task?

A. Making a temporary table

B. Creating a flat file

C. Indexing documents

D. Creating an execution plan

Correct Answer: *C*

Community vote distribution

C (100%)

MLadis 2 months ago

Selected Answer: C

Out of these options, indexing is the technique that allows the database to skip scanning the entire table and instead look up only the specific rows or data pages required. Therefore, the correct choice is:

C. Indexing documents.

upvoted 1 times

Question #136 *Topic 1*

Which of the following best describes the law of large numbers?

A. As a sample size decreases, its standard deviation gets closer to the average of the whole population.

B. As a sample size grows, its mean gets closer to the average of the whole population.

C. As a sample size decreases, its mean gets closer to the average of the whole population.

D. When a sample size doubles, the sample is indicative of the whole population.

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. As a sample size grows, its mean gets closer to the average of the whole population.

The law of large numbers states that as the size of a sample increases, the sample mean (average) will tend to get closer to the population mean. This means that larger samples are more likely to provide a more accurate estimate of the population parameter, such as the mean. Therefore, option B best describes the law of large numbers.

upvoted 1 times

Question #137 *Topic 1*

A C-suite executive would like to monitor KPIs for each department on a monthly basis. Which of the following explains why the executive wants to conduct this performance analysis?

A. To analyze a connection of data points or pathways

B. To use descriptive statistics to determine observations

C. To provide a comparison of data over time

D. To track measurements against defined goals

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

D. KPI is a measurement against a defined objective.

upvoted 1 times

Question #138 *Topic 1*

An analyst is working on a project for a director. During this process, the analyst pulled the data, created summarized tables and graphs with descriptions, created a report summary, and inserted all items into a report. After writing the report, which of the following would be the most appropriate next step?

A. Complete an audit on the data pulled for the report.

B. Complete a check for quality in the report.

C. Complete a review of the data and a check for consistency.

D. Complete a trend analysis to be included in the report.

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. Complete a check for quality in the report.

After writing the report, the most appropriate next step would be to complete a check for quality in the report. This involves reviewing the report to ensure that it is accurate, clear, and free of errors or inconsistencies. Checking for quality ensures that the report effectively communicates the findings and insights derived from the data analysis to the director or other stakeholders. Completing an audit on the data pulled for the report (option A) and reviewing the data and checking for consistency (option C) could have been done earlier in the process. Completing a trend analysis (option D) could also be valuable but may not be the immediate next step after writing the report.

upvoted 1 times

Question #139 *Topic 1*

A collections manager has a team calling customers who are past due on their accounts in an attempt to collect payments. The manager receives the call list in the form of a printed report that is generated by the accounting department at the beginning of each week. Consequently, the collections team calls some customers who have made payments in the time since the report was last printed. Which of the following reporting enhancements could the accounting department implement to best reduce the number of calls on current accounts?

A. Modify the date range on the report.

B. Include a time stamp on the report.

C. Increase the frequency of report generation.

D. Add a report run date to the report.

Correct Answer: *C*

Community vote distribution

B (100%)

stephyfresh13 5 months ago

To best reduce the number of calls on current accounts, the accounting department should implement C. Increase the frequency of report generation.

This change would ensure that the collections team has access to the most up-to-date information, reducing the likelihood of calling customers who have already made payments.

upvoted 1 times

Adonisy 5 months, 1 week ago

Selected Answer: B

B is correct answer

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

B. Include a time stamp on the report.

Adding a time stamp to the report allows the collections manager and their team to determine when the report was created. This information enables them to gauge the relevance of the data and prioritize calls accordingly, reducing the likelihood of contacting customers who have already made payments after the report was printed.

upvoted 1 times

Question #140 *Topic 1*

A data analyst is developing a data dictionary that aligns with a company’s data management processes and policies. Which of the following best describes what should be included in the data dictionary?

A. Information containing the links to business data

B. Information explaining the business methodologies

C. Information containing definitions of the business data

D. Information describing the data analysis phases

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

C. Information containing definitions of the business data.

A data dictionary typically contains definitions and descriptions of the data elements used in an organization's databases or data management systems. This includes details such as data types, formats, allowed values, and any relevant metadata. Providing clear definitions of business data helps ensure consistency and understanding across different teams and systems within the organization.

upvoted 1 times

Question #141 *Topic 1*

After completing web scraping, which of the following file formats needs to be parsed?

A. .html

B. .txt

C. .csv

D. .tsv

Correct Answer: *A*

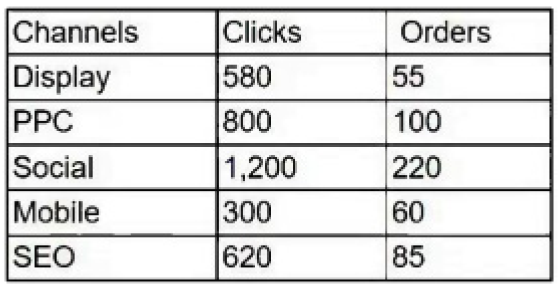
Swift_and_Quick 5 months, 2 weeks ago

A. HTML is from web pages.

upvoted 1 times

Question #142 *Topic 1*

A data analyst received the information in the table below from a recently completed marketing campaign:



Channels	Clicks	Orders
Display	580	55
PPC	800	100
Social	1,200	220
Mobile	300	60
SEO	620	85

Which of the following is the total order conversion rate?

A. 13.2%

B. 14.8%

C. 22.3%

D. 85.2%

Correct Answer: *B*

Community vote distribution

B (100%)

⊟ **07fd0a0** 1 month, 3 weeks ago

Selected Answer: B

Total Clicks:

580

+

800

+

1200

+

300

+

620

=

3500

580+800+1200+300+620=3500

Total Orders:

55

+

100

+

220

+

60

+

85

=

520

55+100+220+60+85=520

Step 2: Apply the Conversion Rate Formula

Conversion Rate

=

(

Total Orders

Total Clicks

)

×

100

Conversion Rate=(

Total Clicks

Total Orders

)×100

=

(

520

3500

)

×

100

=(

3500

520

)×100

=

14.857

%

≈

14.8

%

=14.857%≈14.8%

upvoted 1 times

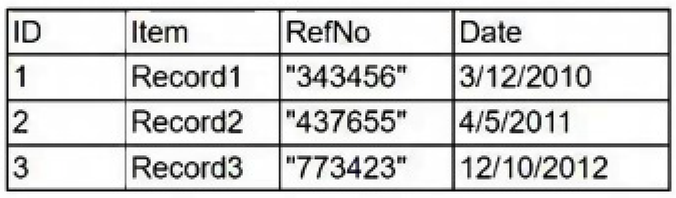
Swift_and_Quick 5 months, 2 weeks ago

B. (55 + 100 + 220 + 60 + 85) / (580 + 800 + 1200 + 300 + 620) = 0.148

upvoted 1 times

Question #143 *Topic 1*

An analyst reviews the following table:



ID	Item	RefNo	Date
1	Record1	"343456"	3/12/2010
2	Record2	"437655"	4/5/2011
3	Record3	"773423"	12/10/2012

Which of the following data types is represented in the values in the RefNo column?

A. Numeric

B. Text

C. Currency

D. Alphanumeric

Correct Answer: *B*

Community vote distribution

B (100%)

Adonisy 5 months, 1 week ago

Selected Answer: B

" " mean text

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

The values in the RefNo column are enclosed in double quotation marks (" ") and contain only numbers. This indicates that the values are treated as text rather than numeric or currency data types.

So, the correct answer is:

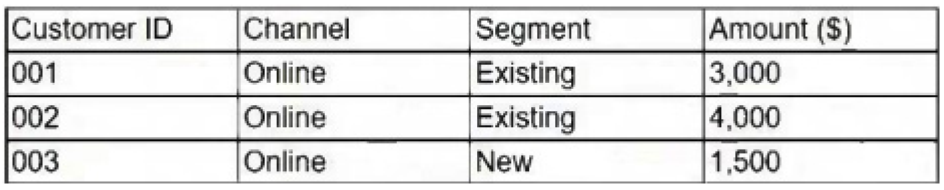
B. Text

upvoted 1 times

Question #144 *Topic 1*

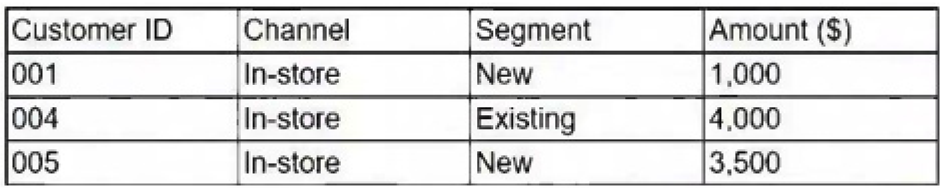
Given the information in the following tables:

Online transactions:



Customer ID	Channel	Segment	Amount ($)
001	Online	Existing	3,000
002	Online	Existing	4,000
003	Online	New	1,500

In-store transactions:



Customer ID	Channel	Segment	Amount ($)
001	In-store	New	1,000
004	In-store	Existing	4,000
005	In-store	New	3,500

Which of the following describes merging these tables to create a master file that includes all transactions for both online and in-store sales?

A. Data audit

B. Data completeness

C. Data validation

D. Data consolidation

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

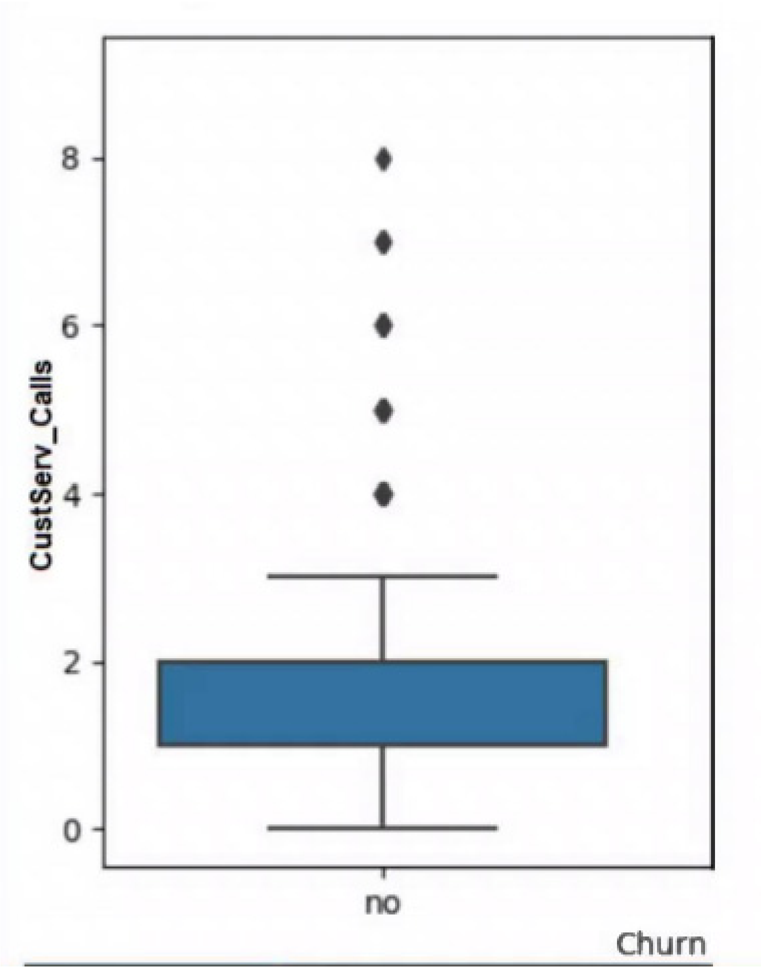
The process of combining the tables of online and in-store transactions to create a master file that includes all transactions for both types of sales is known as data consolidation.

So, the correct answer is:

D. Data consolidation
upvoted 1 times

Question #145 *Topic 1*

Given the image below:



The data should be cleaned because of the presence of:

A. outliers.

B. non-parametric data.

C. multicollinearity.

D. invalid data.

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

A. Outliers.
The following is a vertical box plot, min is 0, max is 3, Q1 is 1, and Q3 is 3.
It seems that the box plot has individual data points (dots) above the upper quartile (Q3) and potentially extending far beyond the upper whisker.
These data points represent values that are significantly higher than the bulk of the data, indicating the presence of outliers.
Outliers can skew statistical analyses and interpretations, so they often warrant data cleaning to ensure the reliability and accuracy of the results.

upvoted 1 times

Question #146 *Topic 1*

Several analysts are working simultaneously to compile a company's annual report and meet a tight deadline. Since each analyst is making changes, the analysts are concerned that they are not always working on the most up-to-date report, which results in rework.

Which of the following should be included in the report to prevent this from happening in the future?

A. A version control table

B. Reference data sources

C. Instructions

D. A report run date

Correct Answer: *A*

⊟ **Swift_and_Quick** 5 months, 2 weeks ago

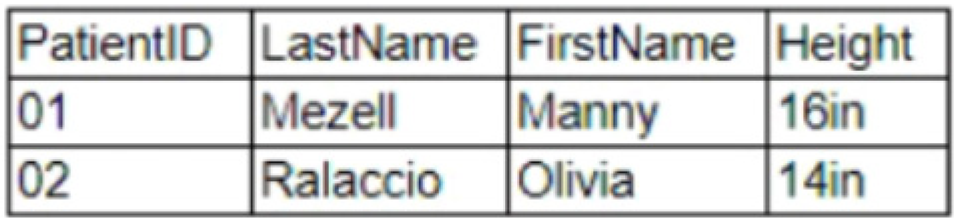
To prevent analysts from working on outdated versions of the report and to minimize rework, including a version control table in the report is essential. This table would detail the version number or date of each revision, along with a brief description of the changes made in each version. With a version control table, analysts can easily identify the most recent version of the report and ensure they are working on the latest iteration.

So, the correct answer is:

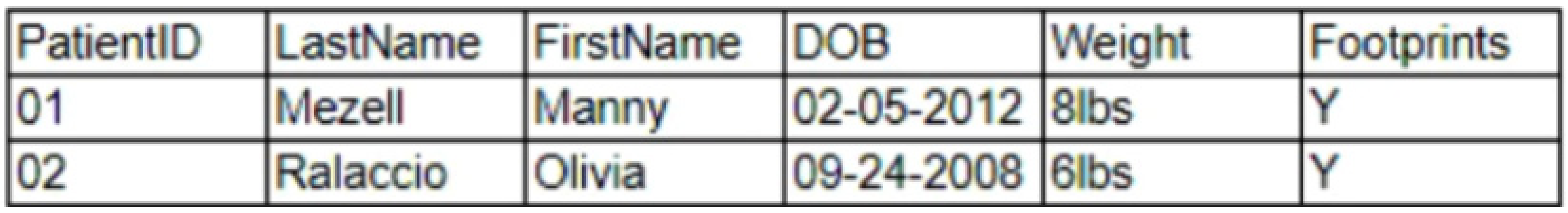
A. A version control table
upvoted 1 times

Question #147 *Topic 1*

Data was previously stored in the following relational tables:

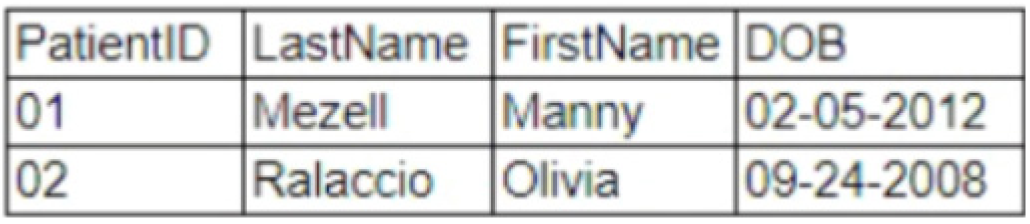


PatientID	LastName	FirstName	Height
01	Mezell	Manny	16in
02	Ralaccio	Olivia	14in



PatientID	LastName	FirstName	DOB	Weight	Footprints
01	Mezell	Manny	02-05-2012	8lbs	Y
02	Ralaccio	Olivia	09-24-2008	6lbs	Y

A database architect changes the relational database structure to include an additional lookup table, which includes the following:



PatientID	LastName	FirstName	DOB
01	Mezell	Manny	02-05-2012
02	Ralaccio	Olivia	09-24-2008

Which of the following MDM issues does this change in structure solve?

A. Multiple redundant data fields

B. Non-compliance with regulations

C. Unstandardized field names

D. Difficulties with personnel continuity

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

The change in structure by introducing an additional lookup table helps address the MDM (Master Data Management) issue of multiple redundant data fields.

Previously, in Table 1 and Table 2, there were duplicate fields such as LastName and FirstName. By consolidating these fields into a separate lookup table, redundant data is eliminated, and the database becomes more normalized.

So, the correct answer is:

A. Multiple redundant data fields

upvoted 1 times

Question #148 *Topic 1*

Which of the following BEST describes the difference between discrete and continuous values?

A. Discrete values change.

B. Discrete values are not distinct.

C. Continuous values are restricted by separation.

D. Discrete values are obtained by counting.

Correct Answer: *D*

Swift_and_Quick 5 months, 2 weeks ago

D. Discrete values are obtained by counting.

Discrete values are distinct and separate, typically obtained by counting individual items or occurrences. For example, the number of students in a classroom or the number of cars in a parking lot are discrete values.

Continuous values, on the other hand, are not restricted by separation and can take on any value within a given range. They are often associated with measurements and can have an infinite number of possible values within a range. Examples include height, weight, temperature, and time.

So, while discrete values are obtained by counting, continuous values are not restricted by separation.

upvoted 1 times

Question #149

Topic 1

A survey asks participants to rate a company on a scale of one to ten. Which of the following BEST describes the rating variable?

A. Continuous

B. Ordinal

C. Categorical

D. Nominal

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

B. Ordinal

Ordinal variables represent categories with a natural order or ranking. In this case, the ratings on a scale from one to ten imply an order from lower to higher ratings. However, the intervals between the ratings may not be equal, and there is no specific meaning to the numerical distance between the ratings.

So, the ratings on a scale of one to ten are ordinal because they indicate a rank or order, but the numerical difference between the ratings may not necessarily represent equal intervals.

upvoted 1 times

Question #150 *Topic 1*

A director who is responsible for reporting performance to executives has requested a dashboard for C-suite users who are not satisfied with the current multipage printed reports. The requirements for the dashboard are extensive and include displaying almost all the available data. An analyst is concerned that what is being requested will overwhelm users and result in an unused dashboard.

Which of the following is the FIRST step the analyst should take?

A. Clarify the goal of the reporting and work to pare down the requirements for a more effective use of data and a streamlined view.

B. Create a dashboard wireframe/mockup and send it to the director and C-suite users for review and approval before moving forward.

C. Develop a dashboard with multiple views managed by permissions so executives only see the information relevant to them individually.

D. Build an interactive dashboard so users can drill down to the most relevant information through the use of saved searches and filters.

Correct Answer: *A*

Swift_and_Quick 5 months, 2 weeks ago

A. Clarify the goal of the reporting and work to pare down the requirements for a more effective use of data and a streamlined view.

By clarifying the reporting goals and understanding the specific needs of the C-suite users, the analyst can prioritize and focus on the most critical data points. This process may involve discussions with stakeholders to identify key performance indicators (KPIs) and essential metrics that align with the strategic objectives of the organization. By streamlining the data requirements, the analyst can ensure that the dashboard provides actionable insights without overwhelming users with unnecessary information.

Once the goals and requirements are clarified, the analyst can proceed to develop a dashboard that effectively communicates the essential information to the C-suite users.

So, the FIRST step for the analyst is to clarify the reporting goals and pare down the requirements for a more effective and focused dashboard.

upvoted 1 times

Question #151 *Topic 1*

Which of the following statistical methods can be used to measure changes over time?

A. Correlation test

B. Independent chi-squared test

C. Dependent t-test

D. Z-test

Correct Answer: *C*

Swift_and_Quick 5 months, 2 weeks ago

A dependent t-test, also known as a paired t-test, is used to determine whether there is a significant difference between the means of two related groups. In the context of measuring changes over time, the dependent t-test is commonly used when the same subjects are measured at two different points in time (e.g., before and after an intervention). It assesses whether there is a significant difference in the mean scores of the dependent variable between the two time points.

So, the correct answer is:

C. Dependent t-test

upvoted 1 times

Question #152 *Topic 1*

A feature can take certain values (A, B, C, D, E, and F) and represents a grade of students from a college. Which of the following variables does this describe?

A. Discrete variable

B. Ordinal variable

C. Numerical variable

D. Continuous variable

Correct Answer: *B*

Swift_and_Quick 5 months, 2 weeks ago

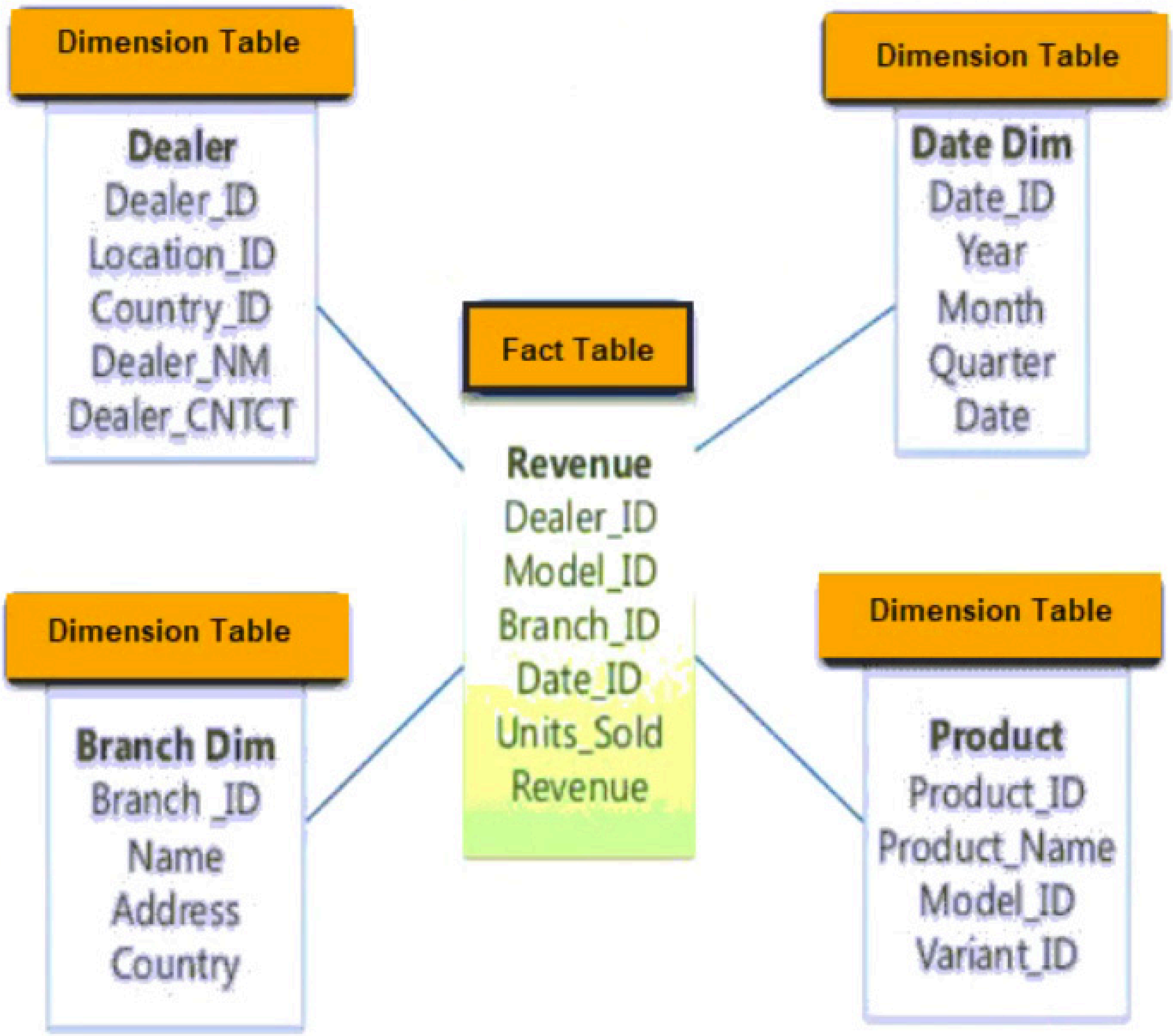
Ordinal variables represent categories with a natural order or ranking, but the intervals between the categories are not necessarily equal. In this case, the grades have a clear order or ranking (A > B > C > D > E > F), but the numerical difference between the grades may not be uniform or meaningful.

So, the correct answer is:

B. Ordinal variable
upvoted 2 times

Question #153 *Topic 1*

Given the image below:



Which of the following data schemas is portrayed?

A. Non-relational

B. Galaxy

C. Snowflake

D. Star

Correct Answer: *C*

Community vote distribution

D (100%)

Adonisy 5 months, 1 week ago

Selected Answer: D

Star schema , this is basic

upvoted 1 times

Ju5t1n_A 6 months, 3 weeks ago

Star Schema, straight out of the CompTIA reference materials.

upvoted 1 times

Swift_and_Quick 11 months, 2 weeks ago

D. Star schema

In a star schema, dimension tables branch out from a central fact table, forming a star-like structure. Each dimension table represents a specific aspect or attribute of the data being analyzed, while the fact table contains the primary measurements or metrics. In a star schema, there are no sub-dimension tables; instead, each dimension table directly connects to the fact table.

upvoted 1 times

⊟ **nshg** 1 year, 1 month ago

D. star schemas are designed to be simple with denormalized dimensions. while snowflakes further normalizes the dimension tables into other tables

upvoted 3 times

Question #154 *Topic 1*

Which of the following is a best practice when updating a legacy data source?

A. Placing old data in new fields

B. Keeping only the most recent data

C. Creating a codebook to document field changes

D. Removing the data source from production

Correct Answer: *C*

⊟ **Swift_and_Quick** 5 months, 2 weeks ago

Creating a codebook helps in documenting any changes made to the data source, including modifications to existing fields or the addition of new fields. This documentation is essential for maintaining transparency, ensuring consistency, and facilitating communication among team members who work with the data. It provides a reference point for understanding the structure and content of the updated data source, aiding in data management and analysis processes.

So, the correct answer is:

C. Creating a codebook to document field changes

upvoted 1 times