

CompTIA

Exam Questions DA0-001

CompTIA Data+ Certification Exam



NEW QUESTION 1

A financial institution is reporting on sales performance to a company at the account level. Due to the sensitive nature of the government the does il with, some account information is not shown. Which of the following fields should be masked?

- A. Sales volume
- B. Start date
- C. Product name
- D. Customer name

Answer: D

Explanation:

Customer name is the field that should be masked, because it contains sensitive information that could identify the government accounts that the financial institution deals with. Masking is a technique that replaces or obscures sensitive data with dummy or random data, such as asterisks or hashes. Masking can help protect the privacy and security of the data, while still allowing for some analysis and reporting. Therefore, the correct answer is D. References: [Data Masking | Definition, Techniques & Examples - Talend], [Data masking - Wikipedia]

NEW QUESTION 2

A data set was recorded using multimedia technology. Which of the following is a necessary step on the way to interpretation?

- A. Structural equation modeling
- B. Transcription
- C. Sequential analysis
- D. Sampling

Answer: B

Explanation:

The correct answer is B. Transcription.

Transcription is a necessary step on the way to interpretation when a data set was recorded using multimedia technology. Multimedia technology refers to the use of various forms of media, such as audio, video, images, and text, to capture and present information¹ Transcription is the process of converting multimedia data into written or textual form, which can then be analyzed using various methods and tools² Transcription can help to make the data more accessible, searchable, and manageable, as well as to preserve the data for future use.

Structural equation modeling is not correct, because it is a statistical technique that tests the causal relationships between multiple variables using observed and latent variables. Structural equation modeling is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data.

Sequential analysis is not correct, because it is a method of analyzing the order and timing of events or behaviors in a data set. Sequential analysis is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data. Sampling is not correct, because it is the process of selecting a subset of data from a larger population for analysis. Sampling is not a necessary step on the way to interpretation, but rather a preliminary step that can be done before collecting or analyzing the data.

NEW QUESTION 3

Which of the following concepts should be applied if a data set with 40 fields needs to be pared down to 20 fields and contains similar data across multiple fields?

- A. Duplication
- B. Consolidation
- C. Compliance
- D. Standardization

Answer: B

Explanation:

Consolidation is the process of combining multiple elements into a single, more effective or coherent whole. In the context of data analytics, consolidation would involve merging similar fields to reduce the overall number of fields in a dataset. This is particularly useful when a dataset contains redundant or similar data across multiple fields, as it helps to simplify the data structure and improve efficiency. Techniques such as dimensionality reduction are often applied to achieve this, where the goal is to retain the most informative and representative features of the data while reducing the number of total features. References:

? Applied Dimensionality Reduction — 3 Techniques using Python¹.

? Seven Techniques for Data Dimensionality Reduction².

? Best practices when working with datasets³.

? Effectively Handling Large Datasets⁴.

NEW QUESTION 4

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 managemen
- D. Users can filter to see the data they want.
- E. Create a dashboard with views for team, individuals, and managemen
- F. Configure permissions to control access.

Answer: D

Explanation:

Create a dashboard with views for team, individuals, and management. Configure permissions to control access. This is because a dashboard is a type of visualization that displays multiple charts or graphs on a single page, usually to provide an overview or summary of some data or information. A dashboard can be

used to provide visibility of current sales numbers, pipeline, and team performance by showing different metrics and indicators related to these aspects. By creating a dashboard with views for team, individuals, and management, the analyst can customize the content and layout of the dashboard for different audiences and purposes. By configuring permissions to control access, the analyst can ensure that the confidential information, such as individuals' earned commissions and projected commissions based on sales, is only visible to the authorized users. The other ways are not the best way to provide this visibility. Here is why: Creating a dashboard displaying a data refresh date so users know the current sales numbers and configuring permissions to control access would not be sufficient to provide visibility of pipeline and team performance, as well as individuals' earned commissions and projected commissions based on sales. The dashboard would only show the current sales numbers and the date when the data was updated, which would not give a comprehensive or detailed view of the sales situation.

Creating a dashboard for sales numbers, pipeline, and team and individual performance for the management team would not be appropriate to provide visibility for the sales team, as they would not have access to the dashboard or the information they need. The dashboard would only be available for the management team, which would limit the transparency and collaboration among the sales team members.

Creating a dashboard with filters for the overall team, individuals, and management would not be secure to provide visibility of confidential information, such as individuals' earned commissions and projected commissions based on sales. The dashboard would allow users to filter and see the data they want, which could expose sensitive or personal information to unauthorized users.

NEW QUESTION 5

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

Answer: A

Explanation:

The best deliverable that would suit the site reliability team's needs is A. A self-serve dashboard of website performance that updates in real time.

A self-serve dashboard is a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance. A self-serve dashboard of website performance that updates in real time would allow the site reliability team to easily and quickly access the information they need about the stability of their website, such as uptime, response time, error rate, traffic volume, etc. A self-serve dashboard would also enable the team to proactively diagnose issues when they occur, by providing alerts, notifications, or drill-down options. A self-serve dashboard would also be more interactive and engaging than a report or an email.

A weekly log report of site visits and user actions would not be a good deliverable for the site reliability team's needs, because it would not provide timely or relevant information about the stability of their website. A weekly log report would be too infrequent and delayed to monitor and diagnose issues when they occur.

A weekly log report would also focus on the behavior and actions of the users, rather than the performance and functionality of the website.

A portal that is refreshed daily and reports errors classified by type would not be a good deliverable for the site reliability team's needs, because it would not provide real-time or comprehensive information about the stability of their website. A portal that is refreshed daily would be too slow and outdated to monitor and diagnose issues when they occur. A portal that reports errors classified by type would be too narrow and limited to capture the full picture of the website performance.

A daily summary email indicating website outages for the previous day would not be a good deliverable for the site reliability team's needs, because it would not provide real-time or actionable information about the stability of their website. A daily summary email would be too late and retrospective to monitor and diagnose issues when they occur. A daily summary email indicating website outages would also be too passive and generic to help the team resolve or prevent issues in the future.

NEW QUESTION 6

Which of the following is a domain-specific language used in programming that is designed for managing data that is held in a relational data stream management system?

- A. SAS
- B. SQL
- C. Python
- D. R

Answer: B

Explanation:

SQL (Structured Query Language) is a domain-specific language used in programming, specifically designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS). It is the standard language for relational database management systems. SQL statements are used to perform tasks such as update data on a database, or retrieve data from a database. Unlike languages like Python or R, which are general-purpose programming languages, SQL is tailored specifically for database management and manipulation.

References:

? ResearchGate article on SQL1.

? SpringerLink chapter on Relational Databases and SQL Language2.

? DataCamp tutorial on SQL Server Installation3.

? Wikipedia page on SQL4.

NEW QUESTION 7

Jhon is working on an ELT process that sources data from six different source systems.

Looking at the source data, he finds that data about the sample people exists in two of six systems.

What does he have to make sure he checks for in his ELT process? Choose the best answer.

- A. Duplicate Data.
- B. Redundant Data.
- C. Invalid Data.
- D. Missing Data.

Answer: C

Explanation:

Duplicate Data.

While invalid, redundant, or missing data are all valid concerns, data about people exists in two of the six systems. As such, Jhon needs to account for duplicate data issues.

NEW QUESTION 8

Which of the following is an example of a discrete data type?

- A. 8in (20cm)
- B. 5 kids
- C. 2.5mi (4km)
- D. 10.7lbs (4.9kg)

Answer: B

Explanation:

A discrete data type is a data type that can only take on a finite number of values, such as integers or categories. An example of a discrete data type is the number of kids, as it can only be a whole number. The other options are examples of continuous data types, as they can take on any value within a range. The length in inches or centimeters, the distance in miles or kilometers, and the weight in pounds or kilograms are all continuous data types. Reference: CompTIA Data+ (DA0-001) Practice Certification Exams | Udemy

NEW QUESTION 9

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

Answer: B

Explanation:

This is because duplicate data refers to data that is repeated or copied in a data set, which can affect the quality and validity of the analysis. A DISTINCT function is a type of function that removes duplicate values from a column or a table, leaving only unique values. For example, a DISTINCT function in SQL that can achieve this is:

```
SELECT DISTINCT column_name FROM table_name;
```

The other data cleansing issues will not be fixed by applying a DISTINCT function. Here is why:

Missing data refers to data that is absent or incomplete in a data set, which can affect the accuracy and reliability of the analysis. A DISTINCT function does not help with missing data, because it does not fill in or impute the missing values.

Redundant data refers to data that is unnecessary or irrelevant for the analysis, which can affect the efficiency and performance of the analysis. A DISTINCT function does not help with redundant data, because it does not remove or filter out the redundant values.

Invalid data refers to data that is incorrect or inaccurate in a data set, which can affect the validity and reliability of the analysis. A DISTINCT function does not help with invalid data, because it does not validate or correct the invalid values.

NEW QUESTION 10

A data analyst received a large amount of third-party data that needs to be joined with in-house data files. After the data is joined, the analyst notices three columns all contain dates. Which of the following should the analyst do to maintain data consistency?

- A. Append all date columns and parse the strings.
- B. Impute all three date columns and then merge.
- C. Merge all date columns and unify the format.
- D. Separate the columns into a table and merge.

Answer: C

Explanation:

When dealing with multiple date columns from different data sources, it's crucial to ensure consistency and accuracy in the dataset. The best practice is to merge the date columns and standardize the date format across the entire dataset. This approach helps maintain data integrity, simplifies analysis, and avoids confusion that could arise from having multiple date formats. Unifying the date format is particularly important when the data will be used for time series analysis or when dates are key to joining with other datasets.

References:

? Best practices in data merging emphasize the importance of a single point of reference and the need to avoid data loss or damage to individual data structures¹.

? Power BI guides suggest that merging columns should be done carefully to maintain data integrity and avoid errors and inconsistencies².

? Oracle Blogs highlight the need for a consistent number of columns among data sources when combining data with unions³.

? Excel tutorials recommend organizing data before merging and using formulas for complex merges⁴.

? An Excel guide on merging date and time columns advises employing functions to ensure seamless handling of non-date values⁵.

NEW QUESTION 10

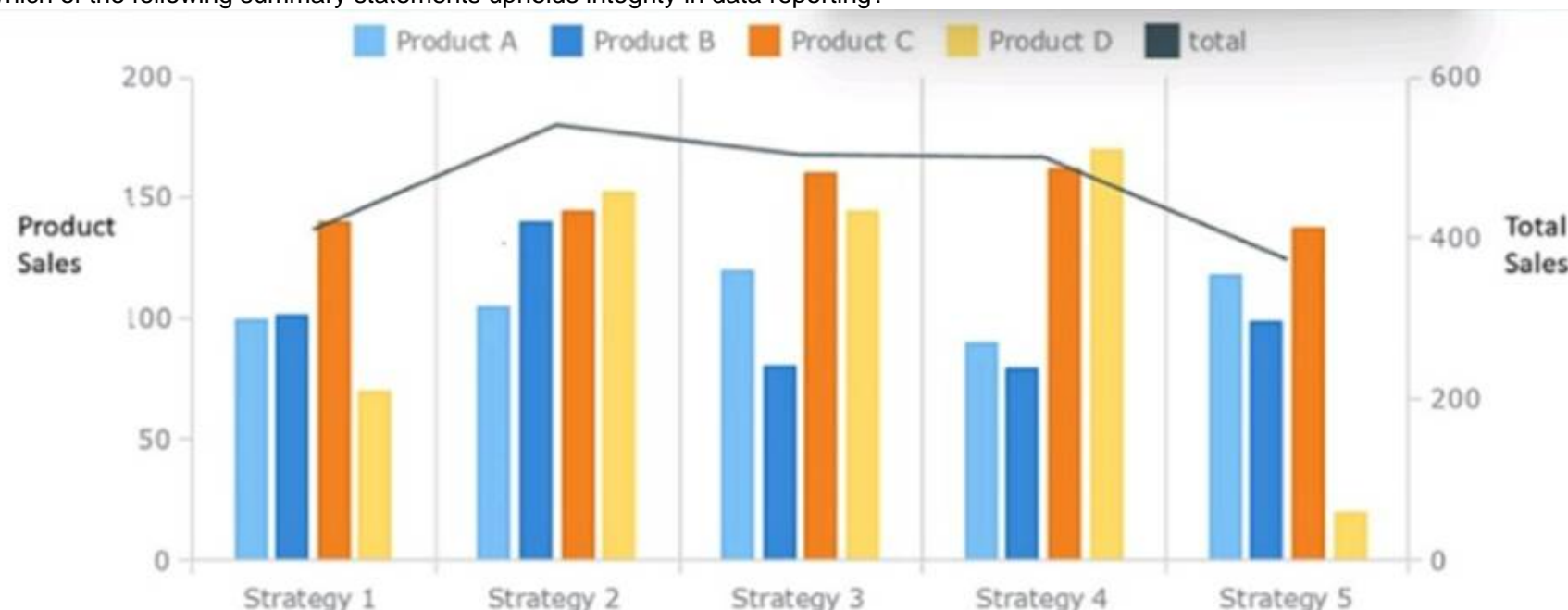
A sales director has requested a report for individual team members within the division be developed. The director would like the report to be shared with all team members, but individual team members should not be identifiable within the report Which of the following access requirements would support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

Answer: B

NEW QUESTION 12

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. over all products it appears to be the most effective.
- E. Product D should be promoted more than the other products in all strategies.

Answer: C

Explanation:

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

A summary statement that upholds integrity in data reporting should be accurate, unbiased, and supported by evidence. Option C is the only statement that meets these criteria, as it reflects the data shown in the bar graph without exaggerating or distorting it. Option C also acknowledges the limitation of the statement by using the word "appears", which indicates that there may be other factors or variables that affect the sales performance.

Option A is inaccurate, as sales are not approximately equal for Product A and Product B across all strategies. Product A has higher sales than Product B in strategies 1, 3, and 5, while Product B has higher sales than Product A in strategies 2 and 4.

Option B is biased, as it does not consider the sales of different products in each strategy. Strategy 4 provides the best sales for Product B, but not for the other products. Strategy 5 has the highest total sales across all products, as shown by the black line graph.

Option D is unsupported by evidence, as it does not explain why Product D should be promoted more than the other products in all strategies. Product D has the lowest sales among all products in strategies 1, 3, and 4, and only slightly higher sales than Product C in strategies 2 and 5.

NEW QUESTION 14

A marketing analytics team received customer transaction data from two different sources. The data is complete and accurate; however, the field names appear to be inconsistent. Given 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

Store transactions:

Customer_ID	Source	Segment	Amount (\$)
001	In-store	New	1,000
004	In-store	Existing	4,000
005	In-store	New	3,500

Which of the following is considered best practice if the team wants to consolidate the files and conduct further analysis?

- A. Standardize the field names.
- B. Recode the data values.
- C. Overwrite the field names in one of the tables.
- D. Edit the field names in the data dictionary.

Answer: A

Explanation:

When consolidating data from different sources, it is crucial to standardize field names to ensure consistency across datasets. This process involves aligning the

field names so that they are the same in both tables, which simplifies the merging of data and subsequent analysis. Standardizing field names helps in maintaining data integrity and avoids confusion that may arise from having different names for the same data point. Recode the data values (B) would not be necessary unless the data values themselves are inconsistent or in different formats. Overwriting the field names in one of the tables © could lead to loss of information or confusion. Editing the field names in the data dictionary (D) is helpful, but it does not address the immediate need to harmonize the field names in the actual datasets.

References:

? Best practices in data management.

? Principles of data integration and consolidation.

NEW QUESTION 18

A data analyst wants to create "Income Categories" that would be calculated based on the existing variable "Income". The "Income Categories" would be as follows:

Income category 1: less than \$1.

Income category 2: more than \$1 and less than \$20,000. Income category 3: more than \$20,001 and less than \$40,000. Income category 4: more than \$40,001.

Which of the following data manipulation techniques should the data analyst use to create "Income Categories"?

- A. Data merge
- B. Derived variables
- C. Data blending
- D. Data append

Answer: B

Explanation:

The correct answer is B: Derived variables Derived variables are variables that you create by calculating or categorizing variables that already exist in your data set.

Data merge is incorrect. Data merging is the process of combining two or more data sets into a single data set. Data blending is incorrect.

Data blending involves pulling data from different sources and creating a single, unique, dataset for visualization and analysis.

Data append is incorrect. A data append is a process that involves adding new data elements to an existing database.

NEW QUESTION 22

During data cleansing, an analyst conducts measures of central tendency on a data set. Which of the following data is the analyst attempting to identify?

- A. Duplicate
- B. Missing
- C. Outlying
- D. Invalid

Answer: C

NEW QUESTION 23

Analytics reports should follow corporate style guidelines.

- A. True.
- B. False.

Answer: A

NEW QUESTION 25

A JSON file is an example of:

- A. structured data.
- B. web data.
- C. machine data.
- D. processed data.

Answer: A

Explanation:

A JSON (JavaScript Object Notation) file is a text-based format for representing structured data based on JavaScript object syntax. It is commonly used for transmitting data in web applications (e.g., sending some data from the server to the client, so it can be displayed on a web page, or vice versa). JSON files are human-readable and can be interpreted by various programming languages, making them ideal for data interchange¹²³.

JSON files typically contain an array of objects, with each object representing a record with a series of name-value pairs. This structured format is both easy to understand and write by humans and easy for machines to parse and generate⁴.

References:

? JSON??s official definition and syntax rules¹.

? A beginner??s guide to JSON and its data types².

? Understanding the JSON file format³.

? Detailed explanation of JSON as a structured data format⁴.

NEW QUESTION 27

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%

Answer: B

Explanation:

The correct answer is A. 13.2%.

The total order conversion rate is the ratio of the total number of orders to the total number of clicks, expressed as a percentage. To calculate the total order conversion rate, we need to sum up the clicks and orders from all the channels, and then divide the orders by the clicks and multiply by 100.

Using the data from the table, we can do the following:

? Total clicks = 580 + 800 + 1,200 + 300 + 620 = 3,500

? Total orders = 55 + 100 + 220 + 60 + 85 = 520

? Total order conversion rate = $(520 / 3,500) \times 100 = 14.857\%$

? Rounding to one decimal place, we get 14.9% Therefore, the total order conversion rate is 14.9%.

NEW QUESTION 30

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

Answer: D

Explanation:

The action that must be done to the Genre column before this task can be completed is delimit. Delimit is a process of separating or splitting a string of text into multiple parts based on a delimiter, which is a character or a sequence of characters that marks the boundary between the parts. For example, a comma (,) or a semicolon (;) can be used as a delimiter. In this case, the Genre column contains multiple genres for each movie, separated by commas. To determine the most popular movie genre, the data analyst needs to delimit the Genre column by commas, so that each genre can be counted and compared separately. The other options are not relevant for this task, as they are related to combining or joining strings or tables, not separating them. Append is a process of adding or attaching one string or table to the end of another string or table. Merge is a process of combining or joining two or more tables into one table based on a common column or

key. Concatenate is a process of joining or linking two or more strings together into one string. Reference: [How to Split Text in Excel - Exceljet]

NEW QUESTION 31

An analyst is building a new dashboard for a user. After an initial conversation with the user, the analyst created a mock-up of the dashboard. Which of the following best explains why the analyst created the mock-up?

- A. To identify the dimensions and measures
- B. To send to the client after deploying the dashboard to production
- C. To confirm important details before dashboard development begins
- D. To receive client approval for the final dashboard design

Answer: C

Explanation:

Answer C. To confirm important details before dashboard development begins.

A dashboard mockup is a prototype of a finished dashboard directly in the product. It is a way to visualize the layout, design, and functionality of the dashboard before it is built with real data and code. A dashboard mockup can help the analyst to confirm important details with the user, such as the business objectives, the key performance indicators, the data sources, the filters, the charts, and the interactivity. By creating a dashboard mockup, the analyst can get immediate feedback and validation from the user, and avoid wasting time and resources on developing a dashboard that does not meet the user's expectations or needs¹.

NEW QUESTION 35

An analyst collected data that includes primary account numbers, expiration dates, and service codes. Which of the following data governance classifications is used to describe this data?

- A. PII
- B. PCI
- C. PBI
- D. PHI

Answer: B

NEW QUESTION 37

An analyst reviews the following data: 7

3
5
2
3
7
7
10

Which of the following is the value of the mode?

- A. 3
- B. 5
- C. 7
- D. 10

Answer: C

Explanation:

The mode is the value that appears most frequently in a data set. In the provided data set, the number 7 appears three times, which is more than any other number. Therefore, the mode of this data set is 7.

? 3 appears twice, but less frequently than 7.

? 5 and 10 each appear only once, so they cannot be the mode.

References:

? Mode in Statistics - Definition and Examples¹

? Understanding Measures of Central Tendency²

? Mode (statistics) - Wikipedia³

NEW QUESTION 38

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

Answer: C

Explanation:

The best sampling method for the data analyst's need is C. Stratified sampling.

Stratified sampling is a type of probability sampling that involves dividing the population into homogeneous groups or strata based on some characteristic, such as state, and then randomly selecting a proportional number of individuals from each stratum. Stratified sampling ensures that every group is adequately represented in the sample, and reduces the sampling error and variability¹²

Systematic sampling is not correct, because it involves selecting every nth individual from the population, starting from a random point. Systematic sampling does not guarantee that every state will have a similar proportion of data in the sample, and may introduce bias or error if there is a hidden pattern or order in the population¹²

Convenience sampling is not correct, because it involves selecting individuals who are easily accessible or available to the researcher. Convenience sampling is a

type of non- probability sampling that does not involve random selection, and may result in a biased or unrepresentative sample¹²
 Random sampling is not correct, because it involves selecting individuals from the population at random, without any grouping or stratification. Random sampling may not produce a sample that has a similar proportion of data from every state, especially if the population is large or heterogeneous. Random sampling may also have a higher sampling error and variability than stratified sampling¹²

NEW QUESTION 39

Given the customer table below:

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

Which of the following chart types is the most appropriate to represent the average spending of active customers vs. inactive customers?

- A. Pie chart
- B. Heat graph
- C. Scatter plot
- D. Line chart

Answer: A

Explanation:

A Pie chart is the most suitable for representing the average spending of active customers versus inactive customers. Pie charts are effective for comparing parts of a whole, which makes them ideal for visually displaying the proportion of spend between two distinct groups. They are widely used to depict percentage distributions and are straightforward, allowing immediate analysis of the active vs. inactive customer spending distribution at a glance.

NEW QUESTION 44

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.

Answer: A

Explanation:

A flat text file is a type of data file that contains only plain text without any formatting or markup. Data in a flat text file is usually separated by a delimiter, which is a character that marks the boundary between different fields or values. For example, a comma-separated values (CSV) file is a flat text file that uses commas as delimiters. Other common delimiters are tabs, spaces, semicolons, and pipes. Therefore, the correct answer is A. References: Plain text - Wikipedia, Comparison of document markup languages - Wikipedia

NEW QUESTION 49

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.

Answer: B

Explanation:

After writing the report, the most appropriate next step for the analyst is to complete a check for quality in the report. This involves reviewing the report for accuracy, clarity, completeness, consistency, and relevance. The analyst should ensure that the report addresses the director's business questions and objectives, that the data and analysis are correct and reliable, that the tables and graphs are well-designed and easy to understand, that the descriptions and summary are concise and informative, and that there are no errors or inconsistencies in the report. A quality check will help the analyst to improve the presentation and communication of the report, as well as to avoid any misunderstandings or misinterpretations by the director¹.

NEW QUESTION 51

What SQL command is used to delete an entire table from a database?

- A. DROP.
- B. MODIFY.
- C. DELETE.
- D. ALTER.

Answer: A

NEW QUESTION 53

A junior web developer is developing a new application where users can upload short videos. The first task is to create a homepage that shows the headline "Upload Your Short Videos" and a clickable button that says "upload now".

Which of the following HTML commands would help the developer to complete the task successfully?

- A. `< span >Upload Your Short Videos< /span >< button >upload now< /button >`
- B. `< p >Upload Your Short Videos< /p >< p >upload now< /p >`
- C. `< h1 >Upload Your Short Videos< /h1 >< button >upload now< /button >`
- D. `< h1 >Upload Your Short Videos< /h1 >< h1 >upload now< /h1 >`

Answer: C

Explanation:

The HTML commands that would help the developer to complete the task successfully are

`<h1>Upload Your Short Videos</h1>` and `<button>upload now</button>`. The `<h1>` tag defines a heading level 1, which is the largest and most important heading on a webpage. The `<button>` tag defines a clickable button that can perform some action when clicked. The other options are not suitable for the task, as they either use the wrong tags or do not create a clickable button. The `` tag defines a section of text with no specific meaning or formatting. The `<p>` tag defines a paragraph of text. The `<hl>` tag does not exist in HTML. Reference: HTML Tags - W3Schools

NEW QUESTION 57

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.
- B. and 3:00 p.m.
- C. A systematic survey that is sent to 100 single-family homes in the county
- D. Surveys sent to ten randomly selected homes within 5mi (8km) of the county's office
- E. Surveys sent to 100 randomly selected homes that are reflective of the population

Answer: D

Explanation:

Surveys sent to 100 randomly selected homes that are reflective of the population. This is because a random sample is a type of sample that is selected by using a random method, such as a lottery or a computer-generated number, which ensures that every element in the population has an equal chance of being selected. A random sample can result in a representative sample, which means that the sample reflects the characteristics and diversity of the population. By sending surveys to 100 randomly selected homes that are reflective of the population, the analyst can ensure that the sample is representative of the county's households and their income levels. The other sampling methods are not likely to result in a representative sample. Here is why:

A stratified phone survey of 100 people that is conducted between 2:00 p.m. and 3:00 p.m. would result in a biased sample, which means that the sample favors or excludes certain groups or elements in the population. By conducting the survey only between 2:00 p.m. and 3:00 p.m., the analyst would miss out on people who are not available or reachable at that time, such as those who are working or sleeping. This could affect the representativeness and generalizability of the sample. A systematic survey that is sent to 100 single-family homes in the county would result in an unrepresentative sample, which means that the sample does not reflect the characteristics and diversity of the population. By sending surveys only to single-family homes, the analyst would ignore other types of households, such as apartments, condos, or mobile homes. This could affect the accuracy and reliability of the sample.

Surveys sent to ten randomly selected homes within 5mi (8km) of the county's office would result in a small sample, which means that the sample size is too low to capture the variability and diversity of the population. By sending surveys only to ten homes within a limited area, the analyst would miss out on many households that are located in different parts of the county. This could affect the precision and confidence of the sample.

NEW QUESTION 59

An analyst develops an IT document and needs to describe the technical terms used in the document. Which of the following is where the analyst should include descriptions of the technical terms?

- A. Glossary
- B. System diagram
- C. User requirements
- D. Index

Answer: A

Explanation:

In technical documentation, a glossary is the designated section where definitions for technical terms are provided. It serves as a reference point for readers to understand specialized or uncommon words used within the document. Including descriptions of technical terms in a glossary ensures that readers have a consistent resource to refer to, which can improve comprehension and reduce misunderstandings¹².

A system diagram (Option B) is a visual representation of the system's components and their interactions, not a place for defining terms. User requirements (Option C) outline what end-users expect from the system, and an index (Option D) is an alphabetical list of topics covered in the document, usually with page numbers, but not definitions.

References:

? Creating effective technical documentation¹.

? Best practices when writing technical descriptions³.

NEW QUESTION 64

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

Answer: D

Explanation:

The correct answer is D. Spearman's rank correlation.

Spearman's rank correlation is a non-parametric test that measures the strength and direction of the relationship between two variables that are ranked (ordinal) or continuous. Spearman's rank correlation does not assume that the data follows a normal distribution or that the variables are linearly related. Spearman's rank correlation is based on the ranks of the data rather than the actual values¹²

* A. One-sample t-test is not correct, because it is a parametric test that compares the mean of a sample to a specified value. One-sample t-test assumes that the data follows a normal distribution and has a known population standard deviation³⁴

* B. Two-way ANOVA is not correct, because it is a parametric test that compares the means of two or more groups that are influenced by two independent factors. Two-way ANOVA assumes that the data follows a normal distribution, has homogeneous variances, and has independent observations.

* C. Correlation coefficient is not correct, because it is a parametric test that measures the strength and direction of the linear relationship between two continuous variables. Correlation coefficient assumes that the data follows a bivariate normal distribution and has no outliers.

NEW QUESTION 66

A data analyst needs to create a master file that includes customer information from the tables below:

Table 1: Online Transactions

Order_ID	Customer_ID	Date	Amount	Quantity
002A	002	03/01/2020	\$800	109
001B	001	02/01/2020	\$400	14
001B	001	02/01/2020	\$400	14
001B	001	02/01/2020	\$400	14
004C	004	06/01/2020	\$700	52
003D	003	05/01/2020	\$900	20

Table 2: 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

Table 3: 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

Given the three tables above, the analyst wants to filter down the information prior to joining it together. In which of the following orders should this data manipulation be approached for the most efficient result?

- A. Merge, append, deduplicate
- B. Merge, deduplicate, append
- C. Deduplicate, append, merge
- D. Append, deduplicate, merge

Answer: B

Explanation:

For efficient data manipulation, the ideal order would be to first merge related tables to create a comprehensive set of records, then deduplicate to remove any redundant information. Lastly, appending additional data, such as from another source or table, ensures that all relevant data is included without redundancy before the final analysis. This order prevents unnecessary duplication of effort, such as deduplicating both before and after appending, which would be less efficient.

In the context of the tables provided, merging would likely involve combining customer information from the online and in-store transaction tables with the customer table. Deduplication would remove any redundant customer records that may exist across these tables. Finally, appending would involve adding any additional transaction records to the master file, ensuring a complete dataset for analysis.

NEW QUESTION 71

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

Answer: B

Explanation:

For a high-level, year-end report requested by a Chief Executive Officer (CEO), a recurring report type is most appropriate. Recurring reports are regular, scheduled reports that typically summarize information over a set period, such as a fiscal year. They provide a consistent format for executives to track performance over time, and their standardized nature makes them suitable for high-level analysis and decision-making. Since CEOs need to monitor performance and make strategic decisions, a recurring report that provides a comprehensive overview of the year's activities and outcomes would be valuable. This allows the CEO to evaluate the company's performance against its goals and objectives systematically.

Dynamic reports (A) are more interactive and typically used for in-depth analysis where users can drill down into the data. Ad hoc reports (C) are one-time, usually unscheduled reports tailored for specific questions, which may not be as comprehensive as a year-end report requires. Self-service reports (D) allow users to create their reports on demand, which may not be the formal, synthesized view a CEO would need for a year-end report.

NEW QUESTION 73

Kelly wants to get feedback on the final draft of a strategic report that has taken her six months to develop.

What can she do to prevent confusion as she seeks feedback before publishing the report?

Choose the best answer.

- A. Distribute the report to the appropriate stakeholders via email.
- B. Use a watermark to identify the report as a draft.
- C. Show the report to her immediate supervisor.
- D. Publish the report on an internally facing website.

Answer: B

Explanation:

The best answer is to use a watermark to identify the report as a draft. A watermark is a faint image or text that appears behind the content of a document, indicating its status or ownership. By using a watermark, Kelly can clearly communicate that the report is not final and still subject to changes or feedback. This can prevent confusion among the readers and avoid any misuse or misinterpretation of the report. The other options are not as effective as using a watermark, as they either do not indicate the status of the report or do not reach the appropriate stakeholders. Distributing the report via email or publishing it on an internally facing website may not make it clear that the report is a draft and may cause confusion or errors. Showing the report to her immediate supervisor may not get enough feedback from other relevant stakeholders who may have different perspectives or insights. Reference: How to Add a Watermark in Microsoft Word - Lifewire

NEW QUESTION 77

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

Answer: C

Explanation:

Year-over-year (YoY) comparison is a method of evaluating two or more measured events to compare the results at one period with those from a comparable period on an annual basis. For a year-over-year comparison of Q2 2020 sales, the analyst should compare the sales figures from Q2 2020 with those from Q2 2019. This comparison will show the growth, stagnation, or decline in sales over the year and is a common practice in financial analysis to assess performance.

References:

- ? SlideTeam's article on sales comparison templates1.
- ? Salesforce help article on calculating YoY or Quarter-over-Quarter (QoQ) in reports2.
- ? Smartsheet's content on annual sales report templates3.
- ? TechRepublic article on creating a YoY comparison chart using a PivotChart in Excel4.

NEW QUESTION 78

A data scientist wants to see which products make the most money and which products attract the most customer purchasing interest in their company. Which of the following data manipulation techniques would he use to obtain this information?

- A. Data append
- B. Data blending
- C. Normalize data
- D. Data merge

Answer: B

Explanation:

The correct answer is B: Data blending.

Data blending is combining multiple data sources to create a single, new dataset, which can be presented visually in a dashboard or other visualization and can then be processed or analyzed. Enterprises get their data from a variety of sources, and users may want to temporarily bring together different datasets to compare data relationships or answer a specific question. Data append is incorrect. Data append is a process that involves adding new data elements to an existing database. An example of a common data append would be the enhancement of a company's customer files. A data append takes the information they have, matches it against a larger database of business data, allowing the desired missing data fields to be added. Normalize data is incorrect.

Data normalization is the process of structuring your relational customer database, following a series of normal forms. This improves the accuracy and integrity of your data while ensuring that your database is easier to navigate. Data merge is incorrect. Data merging is the process of combining two or more data sets into a single data set.

NEW QUESTION 81

You should always choose the analytics tool that is most appropriate for any given situation, even if that means acquiring a new tool.

- A. True.
- B. False.

Answer: B

Explanation:

The statement is false. You should not always choose the analytics tool that is most appropriate for any given situation, even if that means acquiring a new tool. Acquiring a new tool can be costly, time-consuming, and risky, as it may not be compatible with your existing data sources, systems, or processes. It may also require additional training, maintenance, and support. Therefore, you should always consider the trade-offs between the benefits and drawbacks of acquiring a new tool versus using an existing one. You should also evaluate the feasibility, availability, and reliability of the new tool before making a decision. Reference: CompTIA Data+ (DA0-001) Practice Certification Exams | Udemy

NEW QUESTION 82

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

Answer: D

Explanation:

This is because First Name is a variable name format that would be problematic if used in most of the data software programs, such as Excel, SQL, or Python. This is because First Name contains a space between two words, which could cause confusion or errors in the data software programs, as they might interpret the space as a separator or a delimiter between two different variables or values, rather than as part of a single variable name. For example, in SQL, a space is used to separate keywords, clauses,

or expressions in a statement, such as SELECT, FROM, WHERE, etc. Therefore, using First Name as a variable name in SQL could result in a syntax error or an unexpected result. The other variable name formats would not be problematic if used in most of the data software programs. Here is why:

? First_Name_ is a variable name format that uses an underscore (_) to separate two words, which is a common and acceptable practice in most of the data software programs, as it helps to improve the readability and clarity of the variable name. For example, in Python, an underscore is used to follow the PEP 8 style guide for naming variables, which recommends using lowercase letters and underscores for multi-word variable names.

? FirstName is a variable name format that uses camel case to separate two words, which is another common and acceptable practice in most of the data software programs, as it helps to reduce the length and complexity of the variable name. For example, in Excel, camel case is used to follow the VBA naming conventions for naming variables, which recommends using mixed case letters for multi-word variable names.

? First_Name is a variable name format that also uses an underscore (_) to separate two words, which is also a common and acceptable practice in most of the data software programs, as it helps to improve the readability and clarity of the variable name. For example, in SQL, an underscore is used to follow the ANSI SQL naming standards for naming variables, which recommends using lowercase letters and underscores for multi-word variable names.

NEW QUESTION 84

A development company is constructing a new Init 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 Init?

- A. \$640,900
- B. \$690,000
- C. \$705,200
- D. \$702,500

Answer: D

Explanation:

The correct answer is D. \$702,500.

To find the price of the Rose unit, we need to use the average cost per square foot of the original floor plans. The average cost per square foot is calculated by dividing the price by the square footage of each unit type. Using the data from the table, we can do the following:

? Jasmine: $\$345,000 / 1,000 = \345 per square foot

? Orchid: $\$525,000 / 2,000 = \262.5 per square foot

? Azalea: $\$375,000 / 1,500 = \250 per square foot

? Tulip: $\$450,000 / 1,800 = \250 per square foot

The average cost per square foot of the original floor plans is the mean of these four values, which is $(\$345 + \$262.5 + \$250 + \$250) / 4 = \$276.875$ per square foot.

To find the price of the Rose unit, we need to multiply the average cost per square foot by the square footage of the Rose unit. The Rose unit has a square footage of 2,535, according to the table. Therefore, the price of the Rose unit is $\$276.875 \times 2,535 = \$702,421.875$.

Rounding to the nearest whole number, we get \$702,500 as the price of the Rose unit.

NEW QUESTION 87

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

Answer: A

Explanation:

This is because web scraping is a technique that allows the analyst to extract data from web pages, such as articles from a digital news outlet. Web scraping can be done using various tools and methods, such as Python libraries, browser extensions, or online services. The other techniques are not suitable for acquiring data from web pages. Here is why:

Sampling is a technique that involves selecting a subset of data from a larger population, usually for statistical analysis or testing purposes. Sampling does not help the analyst to acquire data from web pages, but rather to reduce the amount of data to be analyzed. Data wrangling is a technique that involves transforming and cleaning data to make it suitable for analysis or visualization. Data wrangling does not help the analyst to acquire data from web pages, but rather to improve the quality and usability of the data.

ETL stands for Extract, Transform, and Load, which is a process that involves moving data from one or more sources to a destination, such as a data warehouse or a database. ETL does not help the analyst to acquire data from web pages, but rather to store and organize the data.

NEW QUESTION 91

??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 dat
- C. whereas tables do not.
- D. Views allow for the joining of multiple data sources, whereas tables do not.
- E. Views can be used to restrict sensitive information.

Answer: A

Explanation:

Views are virtual tables that are created by querying one or more base tables or other views. Views do not store any data, but only show the result of a query. One

of the main advantages of using views is that they can reduce the need for repetitive, complex data joins. For example, if a query involves joining multiple tables with many conditions, creating a view can simplify the query and make it easier to reuse. Therefore, the correct answer is A. References: [What is a Database View? | Definition & Examples - Vertabelo], [Database Views - GeeksforGeeks]

NEW QUESTION 92

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

Answer: AC

Explanation:

Mean and mode are measures of central tendency, which describe the typical or most common value in a distribution of data. Mean is the arithmetic average of all the values in a dataset, calculated by adding up all the values and dividing by the number of values. Mode is the most frequently occurring value in a dataset. Other measures of central tendency include median, which is the middle value when the data is sorted in ascending or descending order.

NEW QUESTION 96

Which one of the following is NOT a common data integration tool?

- A. XSS
- B. ELT
- C. ETL
- D. APIs

Answer: A

Explanation:

Cross-site Scripting (XSS) is a security vulnerability usually found in websites and/or web applications that accept user input. XSS is a client-side vulnerability that targets other application users, while SQL injection is a server-side vulnerability that targets the application's database. How do I prevent XSS in PHP? Filter your inputs with a whitelist of allowed characters and use type hints or type casting.

NEW QUESTION 100

An analyst needs to summarize the number of people in Chicago in 2022 using the following set of data:

Name	City	Year	Grade
Chloe	Chicago	2022	A
Blake	Chicago	2023	B
Carter	Chicago	2022	A
Kim	Detroit	2021	C

Which of the following steps should the analyst use to provide results? (Select two).

- A. Aggregation
- B. Sorting
- C. Filtering
- D. Indexing
- E. Cleaning
- F. Replacing

Answer: AC

NEW QUESTION 105

A data analyst is helping a retail store categorize its customers into five different groups based on the following information:

- How recently the customers made purchases
 - How frequently the customers made purchases
 - How much the customers spent
- Given the following information:

Customer_ID	Channel	Order_Date	Quantity	Territory	Amount (\$)
1001	Online	2/11/2020	12	North	1,250
2001	Store	2/10/2020	31	East	5,000
4001	Online	2/09/2020	24	West	2,500
3001	Online	2/11/2020	51	South	6,000
1001	Store	3/10/2020	22	North	2,000
1001	Online	1/09/2020	87	North	8,400
1001	Store	2/09/2020	23	North	2,000

Which of the following would be most important for the analysis?

- A. CustomerJ
- B. Channel, Order_Date
- C. CustomerJD, Territor
- D. Amount
- E. CustomerJD, Order_Dat
- F. Amount
- G. CustomerJ
- H. Quantity, Amount

Answer: C

NEW QUESTION 110

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

Answer: B

NEW QUESTION 115

Consider this dataset showing the retirement age of 11 people, in whole years: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60
This tables show a simple frequency distribution of the retirement age data.

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

- A. 56
- B. 55
- C. 57
- D. 54

Answer: D

Explanation:

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.
There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.
What is the mode?
The mode is the most commonly occurring value in a distribution.
The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

NEW QUESTION 118

Which of the following data manipulation techniques should an analyst use to hide unnecessary data during analysis?

- A. Filtering
- B. Parametrization
- C. Sorting
- D. Indexing

Answer: A

NEW QUESTION 121

An analyst is reporting on the average income for a county and is reviewing the following data:

Name	Address	Yearly income
Jessica Jones	145 Stonebridge Avenue	\$634,900
Spencer James	1567 Watercress	\$135,000
Olivia Baker	456 Harvard Road	\$95,000
Layla Harding	5674 Yarding Street	\$37,000

Which of the following is the reason the analyst would need to cleanse the data in this data set?

- A. Data completeness
- B. Data outliers
- C. Duplicate data
- D. Missing values

Answer: B

NEW QUESTION 124

Which of the following describes the use of a representative amount of data from a main repository?

- A. Observation
- B. Delta load
- C. Web scraping
- D. Sampling

Answer: D

Explanation:

Sampling refers to the process of selecting a representative subset of data from a larger data set or repository. This technique is used when it is impractical or unnecessary to analyze the entire set of data. A representative sample should accurately reflect the characteristics of the larger population, allowing for analysis and inference about the population as a whole¹².
Observation (A) generally refers to the act of monitoring or recording data. Delta load (B) is a term used in data warehousing to describe the process of loading only the changes since the last data extraction, rather than the entire data set. Web scraping © is the process of extracting data from websites.
References:
? Understanding the importance of data sampling¹.
? The concept of a representative sample in statistics².
? Data repository management and usage³.
? Benefits and methods of data sampling⁴.

NEW QUESTION 128

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.

TABLE 7

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

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.

? 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.

NEW QUESTION 131

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??.

Answer: C

Explanation:

Filter on any of the responses that do not say ??January?? and update them to ??January??. This is because filtering and updating are data cleansing techniques that can be used to ensure data consistency, which means that the data is uniform and follows a standard format. By filtering on any of the responses that do not say ??January?? and updating them to ??January??, the analyst can make sure that all the responses for the month of January are written in the same way. The other steps are not appropriate for ensuring data consistency. Here is why:

Deleting any of the responses that do not have ??January?? written out would result in data loss, which means that some information would be missing from the data set. This could affect the accuracy and reliability of the analysis.

Replacing any of the responses that have ??01?? would not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??Jan?? and ??January??. This could cause confusion and errors in the analysis. Sorting any of the responses that say ??Jan?? and updating them to ??01?? would also not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??01?? and ??January??. This could also cause confusion and errors in the analysis.

NEW QUESTION 133

Which of the following are reasons to conduct data cleansing? (Select 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

Answer: CF

Explanation:

Two reasons to conduct data cleansing are:

? To improve accuracy: Data cleansing helps to ensure that the data is correct, consistent, and reliable. This can improve the quality and validity of the analysis, as well as the decision-making and outcomes based on the data¹²

? To calculate trends: Data cleansing helps to remove or resolve any errors, outliers, or missing values that could distort or skew the data. This can help to identify and measure the patterns, changes, or relationships in the data over time¹³

NEW QUESTION 136

A customer's telephone number is in the format 123-456-7890. Which of the following data types is used for the phone number?

- A. Boolean
- B. Date
- C. Text
- D. Number

Answer: C

Explanation:

A telephone number, despite being composed of digits, is not used for calculations and often includes formatting characters such as hyphens (-). Therefore, the

most appropriate data type for a telephone number is Text (oVr ARCHAR in SQL databases), which can accommodate various formats and lengths, and preserve leading zeros that might be present in some phone numbers. Storing phone numbers as numeric data types would strip away any formatting and could lead to the loss of significant leading digits (like zeros in international numbers).

? Boolean is a binary data type and only represents true or false values.

? Date is a data type used for dates.

? Number could technically store phone numbers, but it is not suitable due to the reasons mentioned above.

References:

? Best Practices for Storing Phone Numbers¹

? Data Types in SQL for Phone Numbers²

NEW QUESTION 138

Which of the following value is the measure of dispersion "range" between the scores of ten students in a test.
 The scores of ten students in a test are 17, 23, 30, 36, 45, 51, 58, 66, 72, 77.

- A. 90
- B. 60
- C. 70
- D. 80

Answer: B

Explanation:

The correct answer is: 60
 Range is the interval between the highest and the lowest score.
 Range is a measure of variability or scatteredness of the varieties or observations among themselves and does not give an idea about the spread of the observations around some central value. Symbolically $R = H_s - L_s$.
 Where R = Range; H_s is the 'Highest score' and L_s is the Lowest Score.
 The scores of ten students in a test are: 17, 23, 30, 36, 45, 51, 58, 66, 72, 77. The highest score is 77 and the lowest score is 17.
 So the range is the difference between these two scores $\text{Range} = 77 - 17 = 60$

NEW QUESTION 141

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

Answer: D

Explanation:

Merging tables to create a master file that includes all transactions for both online and in- store sales is best described as data consolidation. This process involves combining data from various sources into a single, unified dataset. Data consolidation is essential for providing a comprehensive view of all transactions, which can be used for analysis, reporting, and decision-making purposes.

References: The answer is based on standard data management practices and the definition of data consolidation. No specific external documents were referenced for this response.

NEW QUESTION 143

The senior management team at a company receives a detailed sales report at the end of each quarter. The report is several pages long and includes data from dozens of offices across the country. The team wants a better way to get a quick snapshot of what is included in the report. Which of the following modifications would best meet this requirement?

- A. Modifying documentation elements to include reference data sources
- B. Modifying the font size and style so important data points are more visible
- C. Modifying the report to include a summary section with observations and insights
- D. Modifying the report layout so it is easier to follow and understand

Answer: C

Explanation:

The purpose of an executive summary is to provide a concise and informative overview of a longer report, allowing busy stakeholders to quickly understand the key points and findings without reading the entire document. This summary should highlight the most important data, conclusions, and recommendations, and is typically placed at the beginning of the report for easy access¹².

In the context of a detailed sales report for senior management, including a summary section with observations and insights would allow the team to quickly grasp the performance across various offices and identify any significant trends or issues that require attention. This approach aligns with best practices for executive reporting, which emphasize the importance of clear and concise summaries that focus on essential KPIs and actionable insights¹².

References: 1: Databox - How to Write an Executive Summary for a Report: Step By Step Guide with Examples 2: LinkedIn - Best Practices for Writing Executive Summaries

NEW QUESTION 144

A stakeholder wants to see daily sales targets organized in a dashboard by country, state, city, and ZIP Code. Which of the following delivery considerations must a data analyst take into account when creating the dashboard?

- A. Variable formatting
- B. Drill-down capability
- C. Saved searches
- D. Access permissions

Answer: B

NEW QUESTION 149

Which of the following data types best describe 4Ac1? (Select two).

- A. Alphanumeric
- B. Symbolic
- C. Numeric
- D. Float
- E. Boolean
- F. String

Answer: AF

Explanation:

The term 4Ac1?? is a combination of numbers and letters, which fits the definition of an alphanumeric string. Alphanumeric refers to a character set that contains both letters and numbers. In data analytics and programming, such a value is typically treated as a string, which is a sequence of characters. Strings can include letters, digits, and various other symbols.

A numeric data type would only include numbers, and a float is a specific kind of numeric data type that includes decimal points, neither of which applies to 4Ac1??. A boolean data

type represents one of two values: true or false. Since 4Ac1?? does not represent a true or false value, it cannot be classified as boolean. Lastly, symbolic is not a standard data type in the context of programming and data analytics.

References:

? Understanding Python 3 data types¹.

? Basic Data Types in Python².

? Java Data Types³.

NEW QUESTION 152

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.

Answer: C

NEW QUESTION 153

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

Answer: C

Explanation:

The mean is the average of all the values in a data set. To calculate the mean, we add up all the values and divide by the number of values. In this case, the mean for Q1 sales is $(\$2,000 + \$3,000 + \$4,000 + \$2,500 + \$3,500) / 5 = \$3,082.72$ References: CompTIA Data+ Certification Exam Objectives, page 9

NEW QUESTION 157

What category of data stewardship work is focused on ensuring that the organization respects the wishes of data subjects?

- A. Data quality.
- B. Data privacy.
- C. Data security.
- D. Regulatory compliance.

Answer: B

Explanation:

Data privacy defines who has access to data, while data protection provides tools and policies to actually restrict access to the data. Compliance regulations help ensure that user's privacy requests are carried out by companies, and companies are responsible to take measures to protect private user data. Why is data privacy important?

When data that should be kept private gets in the wrong hands, bad things can happen. A data breach at a government agency can, for example, put top secret information in the hands of an enemy state. A breach at a corporation can put proprietary data in the hands of a competitor.

NEW QUESTION 159

A data analyst reviews the following data set:

1
3
5
7
14
10
9
10
10

Which of the following is the range value?

- A. 9
- B. 10
- C. 12
- D. 13

Answer: D

NEW QUESTION 163

Which of the following roles is responsible for ensuring an organization's data quality, security, privacy, and regulatory compliance?

- A. Data owner.
- B. Data steward.
- C. Data custodian.
- D. Data processor.

Answer: B

Explanation:

Correct answer B. Data steward.
A data steward is responsible for leading an organization's data governance activities, which include data quality, security, privacy, and regulatory compliance.

NEW QUESTION 165

A data analyst has been asked to create one table that has each employee's first name, last name, sales, and address. The sales and addresses are listed in the tables below:

Table 1

First name	Last name	Sales
John	Knox	\$30
John	Johnson	\$10
John	Sinclair	\$70
Bob	Sinclair	\$100

Table 2

First name	Last name	Address
John	Knox	2851 N. Southport
John	Johnson	457 Bridle Ridge
John	Sinclair	1067 Windwood Lane
Bob	Sinclair	71 S. Wacker Drive

Which of the following steps should the analyst take to create the table?

- A. Transpose the first name and last name in both table
- B. Use lookup to pull the address field from Table 2 into Table 1.
- C. Use lookup with the first name or first name to pull the address field from Table 2 into Table 1.
- D. Use the append formula in both tables for the first name and last name
- E. Use lookup to pull the address field from Table 2 into Table 1.
- F. Create a column that concatenates the first name and last name in each table
- G. Use concatenate and lookup to bring the address field into Table 1.

Answer: D

NEW QUESTION 166

Which of the following is an example of structured data?

- A. A credit card number
- B. An email
- C. A photo
- D. Social media correspondence

Answer: A

Explanation:

A credit card number is an example of structured data, which is a type of data that conforms to a data model, has a well-defined structure, follows a consistent order, and can be easily accessed and used by a person or a computer program. A credit card number consists of 16 digits that are divided into four groups of four digits each, separated by spaces or hyphens. The first six digits indicate the issuer identification number, the next nine digits indicate the account number, and the last digit is a check digit that validates the number. A credit card number can be stored and processed in a structured format, such as a database or a spreadsheet.

NEW QUESTION 168

You would like to measure how well an organization is achieving its goals. What type of analysis should you perform?

- A. Performance analysis.
- B. Outlier analysis.
- C. Predictive analysis.
- D. Trend analysis.

Answer: A

Explanation:

Performance analysis is the technique of studying or comparing the performance of a specific situation in contrast to the aim and yet executed. In Human Resources, performance analysis can help to review an employee's contribution towards a project or assignment, which they allotted him or her.

NEW QUESTION 171

You have two databases tables that you would like to join together using a foreign key relationship. What term best describes this action?

- A. Blending.
- B. Appending.
- C. Mixing.
- D. Merging.

Answer: D

Explanation:

Data merging is the process of combining two or more data sets into a single data set. Most often, this process is necessary when you have raw data stored in multiple files, worksheets, or data tables, that you want to analyze all in one go.

NEW QUESTION 176

The total values in this month's revenue report are twice as much as last month's. Which of the following most likely occurred during the ETL process?

- A. The data cleansing processes failed to execute.
- B. The database connectivity failed.
- C. The report included the previous month's data.
- D. The data normalization processes failed.

Answer: C

NEW QUESTION 179

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

Answer: A

Explanation:

This is because median is a type of statistical measure that represents the typical value or central tendency of a data set, which means that it divides the data set into two equal halves, such that half of the values are above it and half are below it. Median can be used to provide a measure that represents the typical income of suburban families in the United States, especially when the data set has a lot of outliers, which means that it has values that are unusually high or low compared to the rest of the data set. Median can provide a measure that represents the typical income of suburban families in the United States, because it is not affected or skewed by the outliers, as it only depends on the middle value or the middle two values of the data set, regardless of how extreme or distant the outliers are. For example, median can provide a measure that represents the typical income of suburban families in the United States, by finding the income value that splits the data set into two equal groups of families, such that 50% of the families have higher incomes and 50% have lower incomes. The other statistical measures are not the best measures to represent the typical income of suburban families in the United States. Here is why:

? Mean is a type of statistical measure that represents the average value or central tendency of a data set, which means that it is the sum of all the values divided by the number of values. Mean is not a good measure to represent the typical income of suburban families in the United States, especially when the data set has a lot of outliers, because it is affected or skewed by the outliers, as it takes into account all the values in the data set, regardless of how extreme or distant they are. For example, mean can provide a measure that does not represent the typical income of suburban families in the United States, by finding the income value that is influenced by a few very high or very low incomes, which could make it higher or lower than most of the incomes in the data set.

? Mode is a type of statistical measure that represents the most frequent value or mode of a data set, which means that it is the value that occurs most often in the data set. Mode is not a good measure to represent the typical income of suburban families in the United States, especially when the data set has a lot of outliers, because it is not representative or indicative of the central tendency or distribution of the data set, as it only depends on the count or occurrence of a single value or a few values in the data set, regardless of how common or rare they are. For example, mode can provide a measure that does not represent the typical income of suburban families in the United States, by finding the income value that is repeated more often than others, which could be an outlier or an anomaly in the data set.

? Standard deviation is a type of statistical measure that represents the amount of dispersion or variation of a data set, which means that it quantifies how much the values in a data set vary or deviate from the mean or average of the data set. Standard deviation is not a measure that represents the typical income of suburban families in the United States, but rather a measure that describes the spread or distribution of their incomes, as well as identifies any outliers or extreme values in their incomes. For example, standard deviation can provide a measure that describes how diverse or homogeneous their incomes are, as well as how far their incomes are from their average income.

NEW QUESTION 183

Which of the following types of analysis is used when comparing last week's sales to the previous week's sales?

- A. Trend analysis
- B. Exploratory analysis
- C. Prescriptive analysis
- D. Link analysis

Answer: A

NEW QUESTION 187

A data analyst is working with a team to create a dashboard for a client who requires on- demand access. Which of the following is the best delivery method to support the clients?? requirement?

- A. Email
- B. Scheduled
- C. Subscription
- D. Static

Answer: C

Explanation:

The best delivery method to support the client??s requirement is C. Subscription.

Short Explanation: A subscription is a delivery method that allows the client to access the dashboard on-demand, whenever they need it. A subscription can be set up by the data analyst or the client themselves, and it can be configured to send an email notification when the dashboard is updated or refreshed. A subscription also allows the client to view the dashboard online or download it as a file format of their choice¹²

* A. Email is not the best delivery method because it does not allow the client to access the dashboard on-demand. Email deliveries are sent at a fixed time or frequency, and they may not reflect the latest data or changes in the dashboard. Email deliveries also have limitations on the file size and format of the dashboard attachments¹

* B. Scheduled is not the best delivery method because it does not allow the client to access the dashboard on-demand. Scheduled deliveries are similar to email deliveries, except that they are triggered by a specific event or condition, such as a data update or a threshold value. Scheduled deliveries also have the same limitations as email deliveries on the file size and format of the dashboard attachments¹

* D. Static is not the best delivery method because it does not allow the client to access the dashboard on-demand. Static deliveries are one-time deliveries that are manually generated by the data analyst or the client. Static deliveries do not update or refresh automatically, and they may become outdated or irrelevant over time. Static deliveries also have limitations on the file size and format of the dashboard files³

NEW QUESTION 190

Daniel is using the structured Query language to work with data stored in relational database. He would like to add several new rows to a database table. What command should he use?

- A. SELECT.
- B. ALTER.
- C. INSERT.
- D. UPDATE.

Answer: C

Explanation:

INSERT

The INSERT command is used to add new records to a database table.

The SELECT command is used to retrieve information from a database. It's the most commonly used command in SQL because it is used to pose queries to the database and retrieve the data that you're interested in working with.

The UPDATE command is used to modify rows in the database.

The CREATE command is used to create a new table within your database or a new database on your server.

NEW QUESTION 195

Each month an analyst needs to execute a data pull for the two prior months. Which of the following is the most efficient function for the analyst to use?

- A. Logical
- B. Date
- C. Aggregate
- D. System

Answer: B

Explanation:

The most efficient function for an analyst to execute a data pull for the two prior months would be the Date function. This function allows for the manipulation and formatting of date values within a database. Using Date functions, an analyst can dynamically calculate the start and end dates for the previous two months, ensuring that the data pull is accurate and automated without manual intervention.

For example, SQL functions like DATEADD and DATEDIFF can be used to determine the exact range of dates needed for the data pull. These functions can calculate the first and

last day of the previous months relative to the current date, which is essential for monthly reporting and analysis.

References:

? Discussions on Stack Overflow suggest using SQL date functions

like DATEADD and DATEDIFF to dynamically extract data for previous months, which supports the use of Date functions¹².

? The use of Date functions is also recommended for ensuring that the data pull is

not only efficient but also accurate, as it avoids potential errors associated with manual date entry³.

NEW QUESTION 197

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

- A. continuous data.
- B. categorical data.
- C. ordinal data.
- D. discrete data.

Answer: D

Explanation:

Discrete data is a type of data that can only take certain values, usually whole numbers or integers. Discrete data can be counted, but not measured. For

example, the number of students in a class, the number of books in a library, or the number of phone calls that a call center receives in a day are all examples of discrete data. Discrete data is different from continuous data, which can take any value within a range, and can be measured with precision. For example, the height of a person, the weight of a fruit, or the temperature of a room are all examples of continuous data. Therefore, the correct answer is D. References: [Discrete vs Continuous Data: Definition and Examples - Statistics How To], [Discrete Data - Definition and Examples | Math Goodies]

NEW QUESTION 202

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 the 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

Answer: D

Explanation:

Customer segmentation is a type of cluster analysis, which is a method of grouping data points based on their similarities or differences. Cluster analysis can help identify patterns and trends in the data, as well as target specific groups of customers for marketing purposes. One common technique for customer segmentation is RFM analysis, which stands for recency, frequency, and monetary value. This technique assigns a score to each customer based on how recently they bought the product, how often they buy the product, and how much they spend on the product. These scores can then be used to create clusters of customers with different characteristics and preferences. Therefore, the correct answer is D. References: Cluster Analysis - Statistics Solutions, RFM Analysis: The Ultimate Guide for Customer Segmentation

NEW QUESTION 207

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

Answer: C

Explanation:

Parameterization in SQL queries is a technique used to prevent SQL injection, which is a common security vulnerability that allows an attacker to interfere with the queries that an application makes to its database. By using parameterized queries, the database can distinguish between code and data, regardless of the input received. This method ensures that an attacker cannot change the intent of a query, even if SQL commands are inserted by the attacker. While parameterization can also affect performance by enabling consistent query execution plans, its primary purpose is to enhance security.

References:

- ? Medium article on SQL Query Optimization¹.
- ? MSSQLTips on SQL Query Performance².
- ? Blog post on SQL Performance Optimization³.
- ? SQL Easy guide on improving SQL Query Performance⁴.
- ? LearnSQL.com on SQL for Data Analysis⁵.

NEW QUESTION 211

An analyst wants to extract data from a variety of sources and store the data in a cloud-based environment prior to cleaning. Which of the following integration techniques should the analyst use?

- A. ETL
- B. API
- C. SQL
- D. ELT

Answer: A

NEW QUESTION 213

Given the following tables:

ID	Title
1	New CRM for Project Sales
2	ERP Implementation
3	Develop Mobile Sales Platform

ID	Name	Project_ID
1	John Doe	1
2	Lily Bush	1
3	Jane Doe	2
4	Jack Daniel	Null

Which of the following will be the dimensions from a FULL JOIN of the tables above?

- A. Two rows and three columns
- B. Three rows and four columns
- C. Four rows and two columns
- D. Four rows and four columns

Answer: D

Explanation:

A FULL JOIN in SQL combines all rows from two or more tables, regardless of whether a match exists. The result includes all records when there is a match in the joined tables and fills in NULLs for missing matches on either side. Given the two tables in the image, the first table has three rows, and the second table has four rows. The FULL JOIN of these tables will include all rows from both tables, resulting in four rows. Since there are three unique columns in the first table (ID, Title) and three unique columns in the second table (ID, Name, Project_ID), with the common column being ID, the resulting table will have four columns (ID, Title, Name, Project_ID).

References:

? SQL documentation on FULL JOIN operations.

NEW QUESTION 214

Which of the following data governance concepts fits into the security requirements category?

- A. Data transmission
- B. Data deletion
- C. Data use agreements
- D. Personally identifiable information

Answer: D

NEW QUESTION 216

Exhibit.

Name	Gender_flag	Level	Code	Region
James	Male	College	P	ON
Paul	Female	Elementary	A	BC
Sean	Male	College	S	QC
Dan	Female	Elementary	A	BC
Sam	Male	Elementary	A	BC
Ahmed	Male	University	L	ON
Tom	Male	Elementary	A	BC
Kim	Male	Elementary	A	BC
Pat	Female	Elementary	A	BC
Ben	Male	Elementary	A	BC
Ken	Male	High school	D	AT

Which of the following logical statements results in Table B?

A)

IF Name = "James" and Gender_flag = "College" then delete

B)

IF Name = "Sam" and Gender_flag = "Male" then delete

C)

IF Name = "Pat" and Gender_flag = "Female" then delete

D)

IF Name = "Sean" and Gender_flag = "College" then delete

- A. Option A
- B. Option B
- C. Option C
- D. Option D

Answer: D

Explanation:

The logical statement that results in Table B is Option D. Option D is a logical statement that uses the AND operator to combine two conditions: Name = ??Tom?? and Region = ??BC??. The AND operator returns true only if both conditions are true, otherwise it returns false. Therefore, Option D will select only the rows from Table A that satisfy both conditions, which are rows 4, 5, 6, and 7. These rows form Table B, as shown below: Name | Gender flag | Level | College | Code | Region Tom | Male | Elementary | A | BC | BC Kim | Female | Elementary | A | BC | BC Pat | Female | Elementary | A | BC | BC Ben | Male | Elementary | A | BC | BC

The other options are not correct, as they use different logical operators or conditions that do not result in Table B. Option A uses the OR operator, which returns true if either condition is true, or both. Option A will select all the rows from Table A except row 3, which does not match either condition. Option B uses the NOT operator, which returns the opposite of the condition. Option B will select all the rows from Table A except rows 4, 5, 6, and 7, which match the condition. Option C uses a different condition, Region = ??ON??. which does not match any row in Table A. Option C will select no rows from Table A. Reference: [SQL Logical Operators - W3Schools]

NEW QUESTION 218

A data analyst has removed the outliers from a data set due to large variances. Which of the following central tendencies would be the best measure to use?

- A. Range
- B. Mean
- C. Mode
- D. Median

Answer:

D

Explanation:

The median is recognized as the most appropriate measure of central tendency when outliers have been removed from a dataset. This is because the median is less influenced by extreme values compared to the mean. When outliers are present, they can significantly skew the mean, making it an unreliable measure of central tendency. The median, on the other hand, is the middle value of a dataset when ordered from least to greatest and remains unaffected by the extremes. Therefore, it provides a better representation of the central location of the data after outliers have been excluded.

References:

? Guidelines for Removing and Handling Outliers in Data¹.

? Mean, Median, and Mode: Measures of Central Tendency².

? Which measure of central tendency should be used when there is an outlier?³.

? How are measures of central tendency affected by outliers?⁴.

NEW QUESTION 223

Which one of the following values will appear first if they are sorted in descending order?

A. Aaron.

B. Molly.

C. Xavier.

D. Adam.

Answer: C

Explanation:

The value that will appear first if they are sorted in descending order is Xavier. Descending order means arranging values from the largest to the smallest, or from the last to the first in alphabetical order. In this case, Xavier is the last name in alphabetical order, so it will appear first when sorted in descending order. The other names will appear in the following order: Molly, Adam, Aaron. Reference: Sorting Data - W3Schools

NEW QUESTION 224

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. 394 mm

D. 21,704mm

Answer: B

Explanation:

The correct answer is B. 154 mm.

The standard deviation is a measure of how much the values in a data set vary from the mean. To calculate the standard deviation, we need to follow these steps:

? Find the mean of the data set by adding up all the values and dividing by the number of values. In this case, the mean is $(300 + 430 + 170 + 470 + 600) / 5 = 394$ mm.

? Find the difference between each value and the mean, and square it. In this case, the differences and their squares are:

? Find the sum of the squared differences. In this case, the sum is $8836 + 1296 + 50176 + 5776 + 42436 = 108520$.

? Divide the sum by the number of values. In this case, the result is $108520 / 5 = 21704$. This is called the variance.

? Take the square root of the variance. In this case, the result is $\sqrt{21704} = 147.32$ mm. This is called the standard deviation.

Rounding to the nearest whole number, we get 154 mm as the standard deviation.

NEW QUESTION 228

Andy is a pricing analyst for a retailer. Using a hypothesis test, he wants to assess whether people who receive electronic coupons spend more on average.

What should Andy's null hypothesis be?

A. People who receive electronic coupons spend more on average.

B. People who receive electronic coupons spend less on average.

C. People who receive electronic coupons do not spend more on average.

D. People who do not receive electronic coupons spend more on average.

Answer: C

Explanation:

The null hypothesis presumes the status quo. Andy is testing whether or not people who receive an electronic coupon spend more on average, so, the null hypothesis states that people who receive the coupon do spend more on average.

NEW QUESTION 229

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.

Answer: C

Explanation:

The analyst should rescale the data to fix this issue. Rescaling is a process of transforming data from one scale to another, such as changing the units of measurement. In this case, the analyst needs to rescale the temperatures from Fahrenheit to Celsius, which are two different scales for measuring temperature. To do this, the analyst can use the following formula:

$\text{Celsius} = (\text{Fahrenheit} - 32) * 5/9$

This formula converts each temperature value from Fahrenheit to Celsius by subtracting 32

and multiplying by 5/9. For example, if the temperature is 68°F, the rescaled value in Celsius is:

$\text{Celsius} = (68 - 32) * 5/9$ Celsius = 20°C

Rescaling the data can help the analyst to report the temperatures in a consistent and accurate way, and to avoid any confusion or errors that may arise from using different scales. Rescaling can also make the data more comparable and compatible with other data sources or standards that use the same scale¹².

NEW QUESTION 231

Which of the following best describes a difference between JSON and XML?

- A. JSON is quicker to read and write.
- B. JSON has to use an end tag.
- C. JSON strings are longer
- D. JSON is much more difficult to parse.

Answer: A

Explanation:

The best answer is A. JSON is quicker to read and write.

JSON (JavaScript Object Notation) is a lightweight data-interchange format that is based on the JavaScript programming language and easy to understand and generate. JSON uses a simple syntax that consists of name-value pairs and arrays, and does not require any end tags or attributes. JSON is quicker to read and write than XML (Extensible Markup Language), which is a markup language that uses a tag structure to represent data items. XML has a more complex and verbose syntax that requires end tags, attributes, and namespaces¹²³

NEW QUESTION 233

An organization would like to add a secondary email field to its customer database in order to enrich the customer profiles. Which of the following data manipulation techniques should the analyst use to add this information?

- A. Blend
- B. Merge
- C. Append
- D. Aggregate

Answer: C

NEW QUESTION 236

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

- A. Boolean
- B. String
- C. Integer
- D. Float

Answer: B

Explanation:

The correct data type for text is string. A string is a data type that represents a sequence of characters, such as letters, numbers, symbols, or spaces. A string can be enclosed by single quotes (' ') or double quotes (" ") in most programming languages. For example, "Hello", "World", and "123" are all strings. The other options are not data types for text, but for other kinds of values. A boolean is a data type that represents a logical value, either true or false. An integer is a data type that represents a whole number, such as 1, 0, or -5. A float is a data type that represents a number with a fractional part, such as 3.14, 0.5, or -2.7.

Reference: Data Types - W3Schools

NEW QUESTION 241

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

Answer: A

Explanation:

A dynamic report is a type of report that shows data that changes or updates automatically based on certain criteria or parameters. A dynamic report can allow users to interact with the data, filter it, drill down into it, or visualize it in different ways. A dynamic report is suitable for situations where the data changes frequently or where real-time or near-real-time data is needed for decision making or analysis. In this case, the 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. This means that the data analyst does not need to show real-time or dynamic data, but rather a fixed and accurate view of the sales data for the second quarter. Therefore, a static report would be the best way to meet this stakeholder requirement. Therefore, the correct answer is A.

References: [What are Dynamic Reports? | Sisense], Static vs Dynamic Reports - What's The Difference? | datapine

NEW QUESTION 244

Given the following grocery store orders:

Order_ID	Order_total
85495	\$132.49
28597	\$108.99
57490	\$96.19
35806	\$74.49
18014	\$178.59
39725	\$41.99
20935	\$136.99
25402	\$31.29
85023	\$24.49
27933	\$76.99

If a query is made to the table with the following logic: Order_Total > 132 OR (Order Total >= 25 AND Order_Total < 74)
Which of the following is the number of orders that will be returned by the query?

- A. Four
- B. Five
- C. Six
- D. Seven

Answer: C

Explanation:

Based on the query logic provided: Order_Total > 132 OR (Order Total >= 25 AND Order_Total < 74), we can manually determine which order totals fit this criteria. By examining the image, these are the Order_Total values that match:
? 132.49 (greater than 132)
? 108.99 (greater than or equal to 25 and less than 74)
? 96.19 (greater than or equal to 25 and less than 74)
? 74.49 (greater than or equal to 25 and less than 74)
? 41.99 (greater than or equal to 25 and less than 74)
? 31.29 (greater than or equal to 25 and less than 74) Thus, six orders satisfy the given conditions.

NEW QUESTION 249

Which of the following data protection methods provides confidentiality for data in transit?

- A. De-identification
- B. Encryption
- C. Masking
- D. Anonymization

Answer: B

NEW QUESTION 254

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

Answer: C

Explanation:

This is because link analysis is a type of analysis that determines whether the data being analyzed is connected to other datapoints, such as entities, events, or relationships. Link analysis can be used to identify and visualize the patterns, networks, or associations among the datapoints, as well as measure the strength, direction, or frequency of the connections. For example, link analysis can be used to determine if there is a connection between a customer's purchase history and their loyalty program status. The other types of analysis are not the best types of analysis to conduct to determine whether the data being analyzed is connected to other datapoints. Here is why:

? Trend analysis is a type of analysis that determines whether the data being analyzed is changing over time, such as increasing, decreasing, or fluctuating. Trend analysis can be used to identify and visualize the patterns, cycles, or movements in the data points, as well as measure the rate, direction, or magnitude of the changes. For example, trend analysis can be used to determine if there is a change in a company's sales revenue over a period of time.

? Performance analysis is a type of analysis that determines whether the data being analyzed is meeting certain goals or objectives, such as targets, benchmarks, or standards. Performance analysis can be used to identify and visualize the gaps, deviations, or variations in the data points, as well as measure the efficiency, effectiveness, or quality of the outcomes. For example, performance analysis can be used to determine if there is a gap between a student's test score and their expected score based on their previous performance.

? Exploratory analysis is a type of analysis that determines whether there are any insights or discoveries in the data being analyzed, such as patterns, relationships, or anomalies. Exploratory analysis can be used to identify and visualize the characteristics, features, or behaviors of the data points, as well as measure their distribution, frequency, or correlation. For example, exploratory analysis can be used to determine if there are any outliers or unusual values in a dataset.

NEW QUESTION 257

An analyst needs to determine the appropriate data type for the following sample data: sample data collected:
Which of the following data types should be used for this data?

- A. Text
- B. Float
- C. Alphanumeric
- D. Numeric

Answer: B

NEW QUESTION 260

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.

Answer: B

Explanation:

star schema is a type of database schema that consists of one fact table that is composed of multiple dimensions. A fact table contains quantitative measures or facts that are related to a specific event or transaction. A dimension table contains descriptive attributes or dimensions that provide context for the facts. A star schema is called so because it resembles a star, with the fact table at the center and the dimension tables radiating from it. A star schema is a type of dimensional schema, which is designed for data warehousing and analytical purposes. Other types of dimensional schemas include snowflake schema and galaxy schema. A snowflake schema is similar to a star schema, except that some or all of the dimension tables are normalized into multiple tables. A galaxy schema consists of multiple fact tables that share some common dimension tables. A transactional schema is a type of database schema that is designed for operational purposes, such as recording day-to-day transactions and activities. A transactional schema is usually normalized to reduce data redundancy and improve data integrity. A non-relational schema is a type of database schema that does not follow the relational model, which organizes data into tables with rows and columns. A non-relational schema can store data in various formats, such as documents, graphs, key-value pairs, etc.

NEW QUESTION 261

A data analyst has received a data set that contains actual and projected sales for the fourth quarter of 2019. Which of the following statistical methods should the analyst use to find the measure of dispersion?

- A. Mean
- B. Variance
- C. Correlation
- D. Confidence interval

Answer: B

Explanation:

The measure of dispersion is used to describe the spread of data around a central value. In the context of a data set containing actual and projected sales, the measure of dispersion will help to understand the variability or consistency of sales figures. The variance is the most appropriate statistical method for finding the measure of dispersion because it calculates the average of the squared differences from the Mean, providing a clear picture of data spread. It is especially useful in comparing the spread between different data sets and understanding the distribution of data points.

? Mean is a measure of central tendency, not dispersion.

? Correlation measures the relationship between two variables, not the spread of a single variable.

? Confidence intervals are used to estimate the range within which a population parameter will fall, but they do not measure dispersion within the data set itself.

References:

? Measures of Dispersion in Statistics¹

? Measures of Dispersion - Definition, Formulas, Examples²

? Statistical dispersion - Wikipedia³

NEW QUESTION 264

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.

Answer: C

Explanation:

The best reporting enhancement that the accounting department could implement to reduce the number of calls on current accounts is C. Increase the frequency of report generation.

By increasing the frequency of report generation, the accounting department could provide the collections manager with more up-to-date information on the customers who are past due on their accounts. This would help to avoid calling customers who have made payments in the time since the last report was printed, and thus reduce the number of calls on current accounts. Increasing the frequency of report generation would also improve the accuracy and timeliness of the data, and enhance the efficiency and effectiveness of the collections process.

Modifying the date range on the report, including a time stamp on the report, or adding a report run date to the report would not be sufficient to reduce the number of calls on current accounts. These enhancements would only provide information on when the report was generated or what period it covers, but they would not change the fact that the report could be outdated by the time it reaches the collections manager. Therefore, these enhancements would not solve the problem of calling customers who have already paid their accounts.

NEW QUESTION 265

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

Answer: BD

Explanation:

A data dictionary is a collection of metadata that describes the data elements in a database or dataset. It can help improve data acquisition by providing information about the data sources, formats, quality, and usage. It can also help remember specifics about data fields, such as their names, definitions, types, sizes, and relationships. Therefore, options B and D are correct.

Option A is incorrect because it is not a reason to create and maintain a data dictionary, but a benefit of doing so.

Option C is incorrect because specifying user groups for databases is not a function of a data dictionary, but a function of a database management system or a security policy.

Option E is incorrect because confining breaches of PHI data is not a function of a data dictionary, but a function of a data protection or encryption system.

Option F is incorrect because reducing processing power requirements is not a function of a data dictionary, but a function of a data compression or optimization system.

NEW QUESTION 266

Which one of the following would not normally be considered a summary statistic?

- A. z-score.
- B. Mean.
- C. Variance.
- D. Standard deviation.

Answer: A

Explanation:

Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is. A z-score can be placed on a normal distribution curve.

NEW QUESTION 267

An analyst in a consumer bank department wants to showcase the concentration of accounts opened in the United States by ZIP Code to describe the effectiveness of the bank's marketing campaigns. Which of the following would be the best way to visualize the data?

- A. A stacked chart
- B. A tree map
- C. A waterfall chart
- D. A geographic map

Answer: D

NEW QUESTION 271

The ACME Corporation hired an analyst to detect data quality issues in their Excel documents. Which of the following are the most common issues? (Select TWO)

- A. Apostrophe.
- B. Commas.
- C. Symbols.
- D. Duplicates.
- E. Misspellings.

Answer: DE

Explanation:

- * 1. Duplicates
- * 2. Misspellings

The most common data quality issues are difficult to resolve in Excel because of their rigidity. It forces analysts to do a ton of manual work, which results in a high probability of an error being introduced to the data set. Those common issues include:

- Blanks
- Nulls
- Outliers
- Duplicates
- Extra spaces
- Misspellings
- Abbreviations and domain-specific variations
- Formula error codes

When introduced, these errors can skew or even invalidate the resulting analysis. A smart tool would minimize the possibility of error by automating the manual work. In Excel, you might look for data quality issues in one of two ways. First, you might use auto filters on specific columns to scan for anomalies and blanks or you might use a pivot table to find gaps and discrepancies.

In either case, you're scanning for the anomalies yourself. Suffice it to say that's not a very efficient process. It also means accuracy is only as good as the analyst's eye, so the probability of error varies throughout the day.

NEW QUESTION 276

A data analyst is attempting to understand how ice cream consumption is affected by different attributes. such as cost, temperature. and income level. Which of the following regression analyses should the data analyst perform to understand this relationship?

- A. Logistic
- B. Ordinary least squares
- C. Cox
- D. Polynomial

Answer: B

Explanation:

Answer: B. Ordinary least squares

Ordinary least squares (OLS) is a type of linear regression that is used to fit a regression model that describes the relationship between one or more predictor variables and a numeric response variable. Use when: The relationship between the predictor variable(s) and the response variable is reasonably linear. The response variable is a continuous numeric variable¹.

In this case, the data analyst is interested in understanding how ice cream consumption (the response variable) is affected by different attributes, such as cost, temperature, and income level (the predictor variables). Assuming that these variables have a linear relationship, OLS can be used to estimate the coefficients of the regression equation that best fits the data. OLS can also provide measures of goodness-of-fit, such as R-squared and adjusted R-squared, and test the significance of the coefficients using t-tests and F- tests².

Option A is incorrect, as logistic regression is used to fit a regression model that describes the relationship between one or more predictor variables and a binary response variable. Use when: The response variable is binary – it can only take on two values¹. Ice cream consumption is not a binary variable, but rather a continuous numeric variable.

Option C is incorrect, as Cox regression is used to fit a regression model that describes the relationship between one or more predictor variables and a survival time response variable. Use when: The response variable is the time until an event of interest occurs, such as death, failure, or recovery³. Ice cream consumption is not a survival time variable, but rather a continuous numeric variable.

Option D is incorrect, as polynomial regression is used to fit a regression model that describes the relationship between one or more predictor variables and a numeric response variable. Use when: The relationship between the predictor variable(s) and the response variable is non-linear¹. If there is no evidence of non-linearity in the data, polynomial regression may not be appropriate, as it may overfit the data and produce unreliable estimates.

NEW QUESTION 281

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

Answer: B

Explanation:

ETL is a process that is used during data integration to collect, blend, and load data. ETL stands for extract, transform, and load, which are the three main steps involved in moving data from different sources to a common destination, such as a data warehouse or a data lake. ETL helps to consolidate and standardize data for analysis and reporting purposes. References: CompTIA Data+ Certification Exam Objectives, page 12

NEW QUESTION 286

You are working with a dataset and need to swap the values in rows with those in columns. What action do you need to perform?

- A. Recording
- B. Filtering.
- C. Aggregation.
- D. Transposition.

Answer: D

Explanation:

Transpose creates a new data file in which the rows and columns in the original data file are transposed so that cases (rows) become variables and variables (columns) become

cases. Transpose automatically creates new variable names and displays a list of the new variable names.

Transposing data is useful for data analysis. At times, we have to pull data from various files with different formats for analysis and preparing reports. In such circumstances, we may have to transpose some data from one file to the other. In excel, we can transpose data in multiple ways.

NEW QUESTION 287

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

Answer: A

Explanation:

The correct answer is A. UNION ALL.

UNION ALL is a SQL statement that appends two tables that have the same number of columns and compatible data types. UNION ALL preserves all the rows from both tables, including any duplicates¹²

* B. MERGE is not correct, because MERGE is a SQL statement that combines the data of two tables based on a common column. MERGE can perform insert, update, or delete operations on the target table depending on the matching or non-matching rows from the source table³⁴

* C. GROUP BY is not correct, because GROUP BY is a SQL clause that groups the rows of a table based on one or more columns. GROUP BY is often used with aggregate functions, such as SUM, AVG, COUNT, etc., to calculate summary statistics for each group⁵⁶

* D. JOIN is not correct, because JOIN is a SQL clause that combines the data of two tables based on a common column or condition. JOIN can produce different results depending on the type of join, such as INNER JOIN, LEFT JOIN, RIGHT JOIN, etc.

NEW QUESTION 292

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

Answer: D

Explanation:

A dashboard with filters at the top that the user can toggle would be the most efficient way to deliver this report, because it allows the user to customize the view and explore different combinations of regions, products, and time periods. A workbook with multiple tabs for each region would be cumbersome and repetitive. A daily email with snapshots of regional summaries would not provide enough detail or interactivity. A static report with a different page for every filtered view would be too long and hard to navigate. References: CompTIA Data+ Certification Exam Objectives, page 14

NEW QUESTION 296

Which one of the following is a measure of dispersion?

- A. Variance.
- B. Mode.
- C. Median.
- D. Mean.

Answer: A

NEW QUESTION 299

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

Answer: A

Explanation:

Answer A. Involves the use of descriptive statistics to understand observations. Exploratory data analysis (EDA) is a method of analyzing and investigating data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. EDA involves the use of descriptive statistics, such as mean, median, mode, standard deviation, frequency, or percentage, to understand the distribution, central tendency, variability, and relationship of the data. EDA helps to see what the data can reveal beyond the formal modeling or hypothesis testing, and provides a better understanding of data set variables and the interactions between them¹.

NEW QUESTION 301

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? (Select 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

Answer: DF

Explanation:

To ensure that a report is point-in-time and static, it should include the date when the report was last accessed and the date on which the report was run. These components confirm the specific time frame the data represents, making the report a fixed reference that does not change with subsequent data updates or

accesses. This is crucial for accurate historical analysis and for maintaining the integrity of the data as it was at the time of the report's creation.

References:

- ? Best practices in business reporting.
- ? Importance of time-stamping in data analysis.
- ? Guidelines for creating static reports in data analytics.

NEW QUESTION 306

Which of the following should an analyst do to best summarize the data on a data set?

- A. Filtering
- B. Aggregation
- C. Sorting
- D. Concatenation

Answer: B

NEW QUESTION 311

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

DA0-001 Practice Exam Features:

- * DA0-001 Questions and Answers Updated Frequently
- * DA0-001 Practice Questions Verified by Expert Senior Certified Staff
- * DA0-001 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * DA0-001 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
[Order The DA0-001 Practice Test Here](#)