

CompTIA

Exam Questions DA0-001

CompTIA Data+ Certification Exam



NEW QUESTION 1

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

- A. dependent data.
- B. duplicate data.
- C. invalid data
- D. redundant data

Answer: D

Explanation:

This is because redundant data is a type of data that is unnecessary or irrelevant for the analysis or purpose, which can affect the efficiency and performance of the analysis or process. Redundant data can be caused by having multiple data fields that store the same or similar information, such as patient height in inches and patient height in centimeters in this case. Redundant data can be eliminated or reduced by using data cleansing techniques, such as removing or merging the redundant data fields. The other types of data are not examples of data that is unnecessary or irrelevant for the analysis or purpose. Here is what they mean in terms of data quality:

? Dependent data is a type of data that relies on or is influenced by another data field or value, such as a formula or a calculation that uses other data fields or values as inputs or outputs. Dependent data can be useful or important for the analysis or purpose, as it can provide additional information or insights based on the existing data.

? Duplicate data is a type of data that is repeated or copied in a data set, which can affect the quality and validity of the analysis or process. Duplicate data can be caused by having multiple records or rows that have the same or similar values for one or more data fields or columns, such as customer ID or order ID. Duplicate data can be eliminated or reduced by using data cleansing techniques, such as removing or filtering out the duplicate records or rows.

? Invalid data is a type of data that is incorrect or inaccurate in a data set, which can affect the validity and reliability of the analysis or process. Invalid data can be caused by having values that do not match the expected format, type, range, or rule for a data field or column, such as an email address that does not have an @ symbol or a date that does not follow the YYYY-MM-DD format. Invalid data can be eliminated or reduced by using data cleansing techniques, such as validating or correcting the invalid values.

NEW QUESTION 2

Five dogs have the following heights in millimeters: 300, 430, 170, 470, 600

Which of the following is the mean height for the five dogs?

- A. 394mm
- B. 405mm
- C. 493mm
- D. 504mm

Answer: A

Explanation:

The mean height for the five dogs is calculated by adding up all the heights and dividing by the number of dogs. The formula is:

$\text{mean} = (300 + 430 + 170 + 470 + 600) / 5$ $\text{mean} = 1970 / 5$ $\text{mean} = 394$

Therefore, option A is correct.

Option B is incorrect because it is the median height, which is the middle value when the heights are arranged in ascending order.

Option C is incorrect because it is the mean height multiplied by 1.25.

Option D is incorrect because it is the mean height multiplied by 1.28.

NEW QUESTION 3

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

- A. CSV file
- B. PDF file
- C. JSON file
- D. JPEG file

Answer: A

Explanation:

A CSV file is a type of flat file that stores data as plain text in a table-like structure with rows and columns. Each row represents a single record, while columns represent fields or attributes of the data. A CSV file uses commas or other delimiters to separate the values in each row. A CSV file can be easily imported or exported by various applications and programs¹²

NEW QUESTION 4

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 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 programming languages are best suited for analysis and machine- learning applications? (Select two).

- A. Ruby
- B. Rust
- C. PHP
- D. Python
- E. Kotlin
- F. R

Answer: DF

NEW QUESTION 7

Jenny wants to study the academic performance of undergraduate sophomores and wants to determine the average grade point average at different points during an academic year.

What best describes the data set she needs?

- A. Sample.
- B. Observation.
- C. Variable.
- D. Population.

Answer: A

Explanation:

Correct answer A. Sample.

Jenny does not have data for the entire population of all undergraduate sophomores. While a specific grade point average is an observation of variable, jenny needs sample data.

NEW QUESTION 8

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 9

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

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

- A. WHERE
- B. AGGREGATE
- C. BOOLEAN
- D. IF

Answer: D

Explanation:

This is because an IF function is a type of logical function that returns a value based on a condition or a set of conditions. An IF function can be used to manipulate data by applying different actions or calculations depending on whether the condition is true or false. For example, an IF function in Excel that can achieve this is:

=IF (condition, value_if_true, value_if_false)

The other data manipulation techniques are not examples of logical functions. Here is why:

? WHERE is a type of clause that filters data based on a condition or a set of conditions. A WHERE clause can be used to manipulate data by selecting only the rows that satisfy the condition(s). For example, a WHERE clause in SQL that can achieve this is:

```
SELECT column_name FROM table_name WHERE condition;
```

? AGGREGATE is a type of function that performs a calculation on a group of values, such as sum, average, count, etc. An AGGREGATE function can be used to manipulate data by summarizing or aggregating the values in a column or a table. For example, an AGGREGATE function in SQL that can achieve this is:

```
SELECT AGGREGATE(column_name) FROM table_name;
```

? BOOLEAN is a type of data type that represents two possible values: true or false.

A BOOLEAN data type can be used to manipulate data by storing or returning logical values based on a condition or a set of conditions. For example, a BOOLEAN data type in Python that can achieve this is:

```
boolean_variable = condition
```


NEW QUESTION 10

Given the following data:

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

Which of the following BEST describes the data set?

- A. There is data bias.
- B. The data is incomplete.
- C. The data is inconsistent.
- D. The data is outliers.

Answer: C

Explanation:

This is because inconsistency is a type of data quality issue that occurs when the data does not follow a common format, structure, or rule across different sources or systems, which can affect the efficiency and performance of the analysis or process. Inconsistency can be caused by having different spellings, punctuations, capitalizations, or abbreviations for the same or similar values in a data set, such as ??M??. ??m??. ??Male??. or ??male?? for gender in this case. Inconsistency can be eliminated or reduced by using data cleansing techniques, such as standardizing or normalizing the data values. The other options are not correct descriptions of the data set. Here is why:

? Data bias is a type of data quality issue that occurs when the data is not representative or proportional of the population or the parameter, which can affect the validity and reliability of the analysis or process. Data bias can be caused by having a sample that is too small, too large, or too skewed for the population or the parameter, such as having only male customers for a product that targets both genders in this case. Data bias can be eliminated or reduced by using sampling techniques, such as stratified or cluster sampling.

? The data is incomplete is a type of data quality issue that occurs when the data is absent or missing in a data set, which can affect the accuracy and reliability of the analysis or process. The data is incomplete can be caused by various factors, such as human error, system error, or non-response. The data is incomplete can be addressed by using various methods, such as replacing or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression.

? The data is outliers is a type of data quality issue that occurs when the data has values that are unusually high or low compared to the rest of the data set, which can affect the quality and validity of the analysis or process. The data is outliers can be caused by various factors, such as measurement error, natural variation, or extreme events. The data is outliers can be addressed by using various methods, such as removing or filtering out the outliers, or using robust statistics that are less sensitive to outliers, such as median, interquartile range, or box plot.

NEW QUESTION 11

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 14

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 19

Which of following is a non-relational database?

- A. Neo4j
- B. SQLite
- C. MySQL
- D. PostgreSQL

Answer: A

Explanation:

Neo4j is a type of non-relational database that uses a graph model to store data. A graph database is a database that represents data as nodes and edges, where nodes are entities and edges are relationships between them. A graph database can store complex and diverse data that is not easily structured in tables. A graph database can also perform fast and efficient queries on the data by traversing the connections between the nodes

NEW QUESTION 22

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

- A. A unique key cannot take null values, whereas a primary key can take null values.
- B. There can be only one primary key in a data set, whereas there can be multiple unique keys.
- C. A primary key can take a value more than once, whereas a unique key cannot take a value more than once.
- D. A primary key cannot be a date variable, whereas a unique key can be.

Answer: B

Explanation:

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

A primary key is a column or a set of columns that uniquely identifies each row in a table. A table can have only one primary key, which also enforces the NOT NULL constraint on the column(s) involved. A primary key can also be referenced by a foreign key of another table to establish a relationship between the tables¹²
A unique key is a column or a set of columns that also uniquely identifies each row in a table, but it is not the primary key. A table can have more than one unique key, which also allows one NULL value for the column(s) involved. A unique key can also be referenced by a foreign key of another table to establish a relationship between the tables¹²

Some of the differences between a primary key and a unique key are:

? A primary key creates a clustered index on the column(s), whereas a unique key creates a non-clustered index on the column(s)³

? A primary key does not allow any NULL values, whereas a unique key allows one NULL value for the column(s)¹²³

? A primary key can be a unique key, but a unique key cannot be a primary key¹²

NEW QUESTION 23

An analyst is reviewing the following data: Car IDSpeed

123155
566436
564418
650567
546436
645638

Which of the following should the analyst include in the measures of central tendency for speed?

- A. Mode = 38 Range = 31 Mean = 42.5
- B. Range = 49 Max = 67 Min = 18
- C. Mode = 36 Max = 67 Min = 18
- D. Mode = 36 Median = 37 Mean = 41.5

Answer: D

Explanation:

The measures of central tendency include the mode, median, and mean. The mode is the value that appears most frequently in a data set. In this case, the speed of 36 appears twice, making it the mode. The median is the middle value when a data set is ordered from least to greatest; for these speeds, when ordered (18, 36, 36, 38, 55, 67), the median is the average of the two middle numbers, which is $(\frac{36 + 38}{2} = 37)$. The mean is the average of all values, calculated as $(\frac{18 + 36 + 36 + 38 + 55 + 67}{6} = 41.7)$. References:

? The calculation of the mode, median, and mean is based on standard statistical formulas and definitions.

The measures of central tendency for speed include the mode, median, and mean. To calculate these, we first need to organize the data:

? Speeds in ascending order: 18, 36, 36, 38, 55, 67

? Mode is the value that appears most frequently, which is 36, as it appears twice.

? Median is the middle value when the data is ordered. Since we have an even number of observations, we take the average of the two middle values (36 and 38), resulting in 37.

? Mean is the sum of all values divided by the number of values. $(18+36+36+38+55+67)/6=41.5$

Thus, the correct option is D, which includes Mode = 36, Median = 37, and Mean = 41.5. The range, maximum, and minimum values, although useful in understanding data dispersion, are not measures of central tendency and are therefore not relevant to this specific question.

NEW QUESTION 25

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 29

A database consists of one fact table that is composed of multiple dimensions. Each dimension is represented by a denormalized table. This structure is an example of a:

- A. non-relational schema.
- B. galaxy schema.
- C. snowflake schema.
- D. star schema.

Answer: D

Explanation:

A star schema is a type of database schema that consists of one fact table and multiple dimension tables. The fact table contains the measures or metrics of the business process, such as sales, orders, or transactions. The dimension tables contain the attributes or characteristics of the business entities, such as products, customers, or locations. The fact table is connected to the dimension tables by foreign keys that reference the primary keys of the dimension tables. The fact table is located at the center of the schema, while the dimension tables are located at the edges, forming a star-like shape¹.

A star schema is an example of a denormalized schema, which means that the dimension tables are not normalized and may contain redundant or repeated data. This is done to improve the performance and simplicity of queries, as there are fewer joins and tables involved. A star schema is suitable for data warehouses and business intelligence applications that require fast and efficient data retrieval².

NEW QUESTION 33

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

- A. a t-test.

- B. a performance analysis.
- C. an exploratory data analysis.
- D. a link analysis.

Answer: C

Explanation:

This is because exploratory data analysis is a type of process that performs initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization, such as box plots, histograms, scatter plots, etc. Exploratory data analysis can be used to understand and summarize the data, as well as to generate hypotheses or questions for further analysis or research. For example, exploratory data analysis can be used to identify and visualize the characteristics, features, or behaviors of the data, as well as to measure their distribution, frequency, or correlation. The other options are not types of processes that perform initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization. Here is what they mean:

? A t-test is a type of statistical method that tests whether there is a significant difference between the means of two groups or samples, such as whether there is a difference between the average exam scores of two classes in this case. A t-test

can be used to test or verify a claim or an assumption about the data, as well as to measure the confidence or the error of the estimation.

? A performance analysis is a type of process that measures whether the data

meets certain goals or objectives, such as targets, benchmarks, or standards. A performance analysis can be used to identify and visualize the gaps, deviations, or variations in the data, as well as to measure the efficiency, effectiveness, or quality of the outcomes. For example, a 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.

? A link analysis is a type of process that determines whether the data is connected to other datapoints, such as entities, events, or relationships. A link analysis can be used to identify and visualize the patterns, networks, or associations among the datapoints, as well as to measure the strength, direction, or frequency of the connections. For example, a link analysis can be used to determine if there is a connection between a customer's purchase history and their loyalty program status.

NEW QUESTION 36

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 41

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 director.

NEW QUESTION 44

Mario works with a group of R programmers tasked with copying data from an accounting system into a data warehouse. In what phase are the group's R skills most relevant?

- A. Extract.
- B. Load.
- C. Transform.
- D. Purge.

Answer: C

NEW QUESTION 47

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:

Delimiting is the process of splitting a column of data into multiple columns based on a separator or delimiter character. Delimiting can help separate data that is combined or concatenated in one column into distinct values or categories. For example, if a column contains text values that are separated by commas, such as ??Comedy, Suspense??. delimiting can split this column into two columns, one for ??Comedy?? and one for ??Suspense??. Delimiting is different from other options, such as appending, merging, or concatenating, which are methods of combining or joining data from multiple columns or sources. In this case, the data analyst needs to determine the most popular movie genre based on the Genre column in the table. However, this column contains multiple genres for each movie, separated by commas. Therefore, the data analyst must delimit this column before this task can be completed. Therefore, the correct answer is D. References: Split text into different columns with functions - Office Support, How to Split Text in Excel (Using Formulas & Split Function)

NEW QUESTION 48

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 49

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

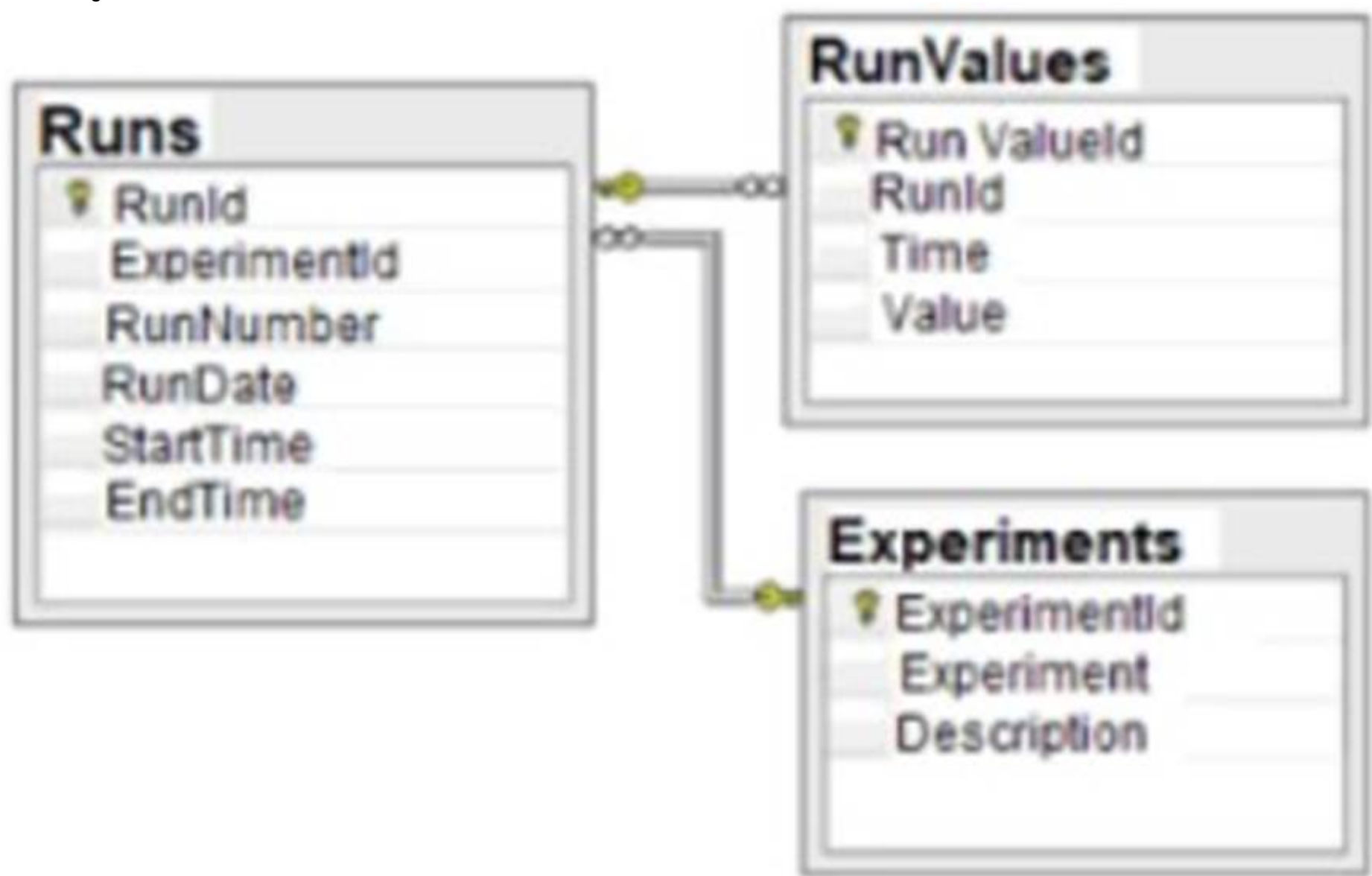
Explanation:

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 53

Given the diagram below:



Which of the following data schemas shown?

- A. Key-value pairs
- B. Online transactional processing
- C. Data Lake
- D. Relational database

Answer: D

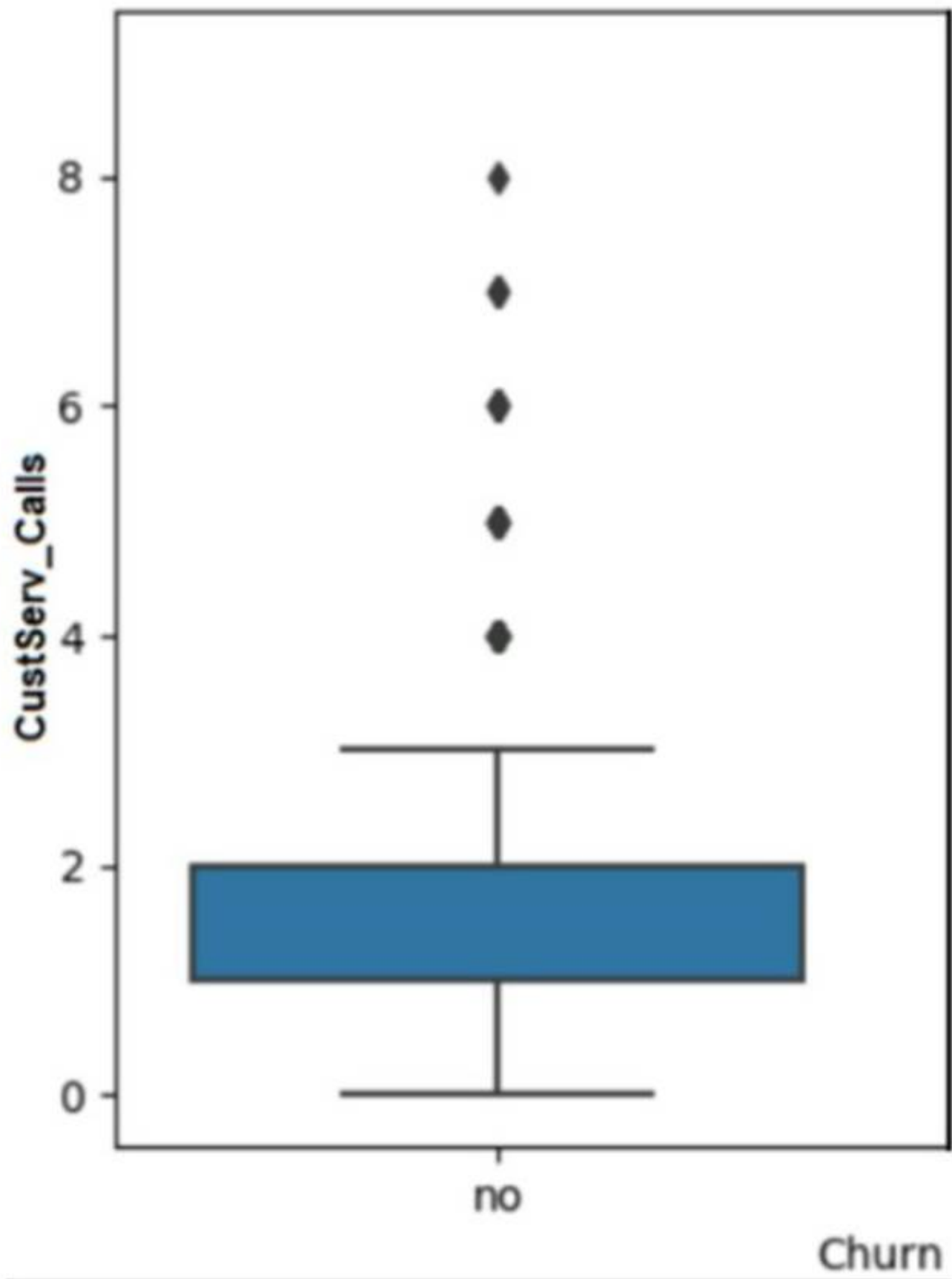
Explanation:

A relational database is a type of database that organizes data into tables, where each table has a fixed number of columns and a variable number of rows. Each row in a table represents a record or an entity, and each column represents an attribute or a property of that entity. The tables are linked by common fields, called keys, which enable the database to establish relationships between the data. A relational database schema is a diagram that shows the structure and organization of the tables, columns, keys, and constraints in a relational database. The diagram given in the question is an example of a relational database schema, as it shows two tables: ??Runs?? and ??Experiments??, with their respective columns, data types, and primary keys. The ??Runs?? table also has a foreign key that references the ??ExperimentId?? column in the ??Experiments?? table, indicating a relationship between the two tables. Therefore, the correct answer is D.

References: What is a database schema? | IBM, Database Schema - Javatpoint

NEW QUESTION 54

Given the image below:



The data should be cleaned because of the presence of:

- A. outlier
- B. non-parametric data.
- C. multicollinearity.
- D. invalid data.

Answer: A

Explanation:

The answer is A. Outlier.

Short Explanation: An outlier is a data point that differs significantly from the rest of the data in a dataset. An outlier can indicate an error, an anomaly, or a rare event in the data. An outlier can affect the statistical analysis and visualization of the data, such as skewing the mean, variance, or distribution of the data. Therefore, data should be cleaned to identify and remove or correct any outliers.

The image below shows a box plot graph with a vertical axis labeled "Customer Calls" and a horizontal axis labeled "Churn". The box plot is blue in color and the median value is around 2. There are 7 outliers above the box plot, ranging from 4 to 8. image)

A box plot is a type of graph that can show the distribution of data values using five summary statistics: minimum, maximum, median, first quartile, and third quartile. The box represents the interquartile range (IQR), which is the difference between the first and third quartiles. The median is shown as a line inside the box. The whiskers extend from the box to the minimum and maximum values, excluding any outliers. Outliers are shown as dots or circles outside the whiskers. In this graph, we can see that most of the customer calls are between 0 and 4, with a median of 2. However, there are 7 outliers that have more than 4 customer calls, up to 8. These outliers may indicate some customers who have more issues or complaints than others, or some errors or anomalies in the data collection or recording process. These outliers can affect the analysis and interpretation of the customer calls and churn relationship, such as making it seem that more customer calls lead to less churn, which may not be true for the majority of the customers. Therefore, data should be cleaned to investigate and handle these outliers appropriately.

NEW QUESTION 56

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

- A. Cell phone device name

- B. Customer??s name
- C. Government ID number
- D. Telephone number

Answer: A

Explanation:

Non-personally identifiable information (non-PII) is any data that cannot be used to identify, contact, or locate a specific individual, either alone or combined with other sources. Non-PII can include aggregated statistics, anonymous data, device identifiers, IP addresses, cookies, and other types of information that do not reveal the identity or location of a person. Cell phone device name is an example of non-PII, as it does not reveal any personal information about the owner or user of the device. Therefore, the correct answer is A. References: What is Non-Personally Identifiable Information (Non-PII)? | Definition and Examples, What is Personally Identifiable Information (PII)? | Definition and Examples

NEW QUESTION 57

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 get prevent confusion as see 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 58

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 59

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

- A. Continuous
- B. Ordinal
- C. Categorical
- D. Nominal

Answer: B

Explanation:

The rating variable in a survey where participants rate a company on a scale of one to ten is best described as ordinal. This is because the ratings are ranked in order, with each number representing a position on a scale of satisfaction or quality. The numbers are not just labels (which would be nominal), nor do they represent a continuous spectrum (which would be continuous). They also do not fit the definition of categorical, as that implies non- ordered groups or categories. In an ordinal scale, the order of the values is significant and meaningful¹².

References:

? Qualtrics explains that ordinal scales have answer sets that occur in a logical and systematic order, providing qualitative data.

? Zonka Feedback describes a 1 to 10 rating scale survey, indicating that the numbers represent a ranking from most negative to most positive experience, which aligns with the characteristics of an ordinal scale.

NEW QUESTION 60

A client has requested an analysis of all pet care items purchased by current customers and their social media connections in the past 12 months. Which of the

following data analysis techniques would be the best choice given these requirements?

- A. Trend analysis
- B. Performance analysis
- C. Link analysis
- D. Exploratory data analysis

Answer: C

NEW QUESTION 63

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

- A. Replace missing data.
- B. Remove duplicate data.
- C. Replace redundant data.
- D. Remove invalid data.

Answer: A

Explanation:

This is because missing data is a type of data quality issue that occurs when data is absent or incomplete in a data set, which can affect the accuracy and reliability of the analysis or process. Missing data can be caused by various factors, such as human error, system error, or non-response. Missing data can be addressed by using various methods, such as replacing missing data, which means filling in or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression. The other methods are not used to address missing data. Here is why:

? Remove duplicate data is a type of method that eliminates or reduces duplicate data, which is a type of data quality issue that occurs when data is repeated or copied in a data set. Removing duplicate data does not address missing data, but rather affects the quantity and validity of the data.

? Replace redundant data is a type of method that eliminates or reduces redundant data, which is a type of data quality issue that occurs when data is unnecessary or irrelevant for the analysis or purpose. Replacing redundant data does not address missing data, but rather affects the efficiency and performance of the analysis or process.

? Remove invalid data is a type of method that eliminates or reduces invalid data, which is a type of data quality issue that occurs when data is incorrect or inaccurate in a data set. Removing invalid data does not address missing data, but rather affects the validity and reliability of the analysis or process.

NEW QUESTION 64

Which of the following query statements would be used when filtering data in a relational database management system? (Select two).

- A. ORDER BY
- B. HAVING
- C. WHERE
- D. SELECT
- E. INSERT
- F. GROUP BY

Answer: BC

NEW QUESTION 66

Which of the following can be used to translate data into another form so it can only be read by a user who has a key or a password?

- A. Data encryption.
- B. Data transmission.
- C. Data protection.
- D. Data masking.

Answer: A

Explanation:

Data encryption can be used to translate data into another form so it can only be read by a user who has a key or a password. Data encryption is a process of transforming data using an algorithm or a cipher to make it unreadable to anyone except those who have the key or the password to decrypt it. Data encryption is a common method of protecting data from unauthorized access, modification, or theft. Reference: Guide to CompTIA Data+ and Practice Questions - Pass Your Cert

NEW QUESTION 68

Joseph is interpreting a left skewed distribution of test scores. Joe scored at the mean, Alfonso scored at the median, and gaby scored and the end of the tail. Who had the highest score?

- A. Joseph
- B. Joe
- C. Alfonso
- D. Gaby

Answer: C

Explanation:

Alfonso had the highest score. A left skewed distribution is a distribution where the tail is longer on the left side than on the right side, meaning that most of the values are clustered on the right side and there are some outliers on the left side. In a left skewed distribution, the mean is less than the median, which is less than the mode. Therefore, Joseph, who scored at the mean, had the lowest score, Gaby, who scored at the end of the tail, had the second lowest score, and Alfonso, who scored at the median, had the highest score. Reference: Skewness - Statistics How To

NEW QUESTION 70

Emma is working in a data warehouse and finds a finance fact table links to an organization dimension, which in turn links to a currency dimension that not linked to the fact table.
What type of design pattern is the data warehouse using?

- A. Star.
- B. Sun.
- C. Snowflake.
- D. Comet.

Answer: C

Explanation:

Correct answer C. Snowflake.
Since the dimension links to a dimension that isn't connected to the fact table, it must be a Snowflake, with a Star, all dimensions link directly to the fact table, Sun and Comet are not data warehouse design patterns.

NEW QUESTION 75

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 80

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 85

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 87

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 89

Angela is aggregating data from CRM system with data from an employee system.
 While performing an initial quality check, she realizes that her employee ID is not associated with her identifier in the CRM system.

What kind of issues is Angela facing? Choose the best answer.

- A. ETL process.
- B. Record linkage.
- C. ELT process.
- D. System integration.

Answer: B

Explanation:

While this scenario describes a system integration challenge that can be solved with ETL or ELT, Angela is facing a Record linkage issue.

NEW QUESTION 91

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 94

Given the following customer and order tables:

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

- A. Five rows, eight columns
- B. Seven rows, eight columns
- C. Eight rows, seven columns
- D. Nine rows, five columns

Answer: B

Explanation:

This is because an INNER JOIN is a type of join that combines two tables based on a matching condition and returns only the rows that satisfy the condition. An INNER JOIN can be used to merge data from different tables that have a common column or a key, such as customer ID or order ID. To perform an INNER JOIN of the customer and order tables, we can use the following SQL statement:

```
SELECT * FROM customer INNER JOIN order ON customer.customer_id = order.customer_id;
```

This statement will select all the columns (*) from both tables and join them on the customer ID column, which is the common column between them. The result of this statement will be a new table that has seven rows and eight columns, as shown below:

customer_id	first_name	last_name	email	order_id	order_date	product	quantity
1	John	Smith	john.smith@email.com	1	2020-01-01	Book	2
2	Jane	Doe	jane.doe@email.com	2	2020-01-02	Pen	5
3	Bob	Lee	bob.lee@email.com	3	2020-01-03	Notebook	3
4	Mia	Chen	mia.chen@email.com	4	2020-01-04	Mug	4
5	Raj	Patel	raj.patel@email.com	null	null	null	null
null	null	null	null	null	null	null	null

The reason why there are seven rows and eight columns in the result table is because:

? There are seven rows because there are six customers and six orders in the original tables, but only five customers have matching orders based on the customer ID column. Therefore, only five rows will have data from both tables, while one row will have data only from the customer table (customer 5), and one row will have no data at all (null values).

? There are eight columns because there are four columns in each of the original tables, and all of them are selected and joined in the result table. Therefore, the result table will have four columns from the customer table (customer ID, first name, last name, and email) and four columns from the order table (order ID, order date, product, and quantity).

NEW QUESTION 97

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 101

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 105

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 109

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 112

Which of the following is the best variable format to store a customer's age using the least possible amount of storage data?

- A. Int
- B. Float
- C. Char
- D. Double

Answer: A

NEW QUESTION 116

Which of the following is a relational database?

- A. SQL
- B. Excel
- C. JSON
- D. NoSQL

Answer: A

NEW QUESTION 119

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

- A. Discrete data cannot create a sloped line.
- B. Discrete data can only be a finite number of values.
- C. Discrete data can have decimal points.
- D. Discrete data applies only to numbers.

Answer: B

Explanation:

Discrete data are data that can only assume specific values that are countable and distinct. For example, the number of books, the number of heads in a coin toss, or the number of patients in a hospital are discrete data. Discrete data cannot have fractional or decimal values, and there are clear spaces between the possible values¹². Continuous data are data that can assume any value within a range and can be meaningfully divided into smaller parts. For example, the weight, height, length, time, or temperature are continuous data. Continuous data can have fractional or decimal values, and there are infinite numbers of possible values between any two points¹².

NEW QUESTION 122

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 127

Given the following report:

Quarterly Customer Service Report

Table 1. Frequency of Ticket Statuses

Status	Count
Reported	11
In-Progress	323
Closed	554

Table 2. Occurrence of Target Phrases

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

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

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

- A. A control group for the phrases

- B. A summary of the KPIs
- C. Filter buttons for the status
- D. The date when the report was last accessed
- E. The time period the report covers
- F. The date on which the report was run

Answer: E

Explanation:

The date on which the report was run. This is because the time period the report covers and the date on which the report was run are two components that need to be added to ensure the report is point-in-time and static, which means that the report shows the data as it was at a specific moment or interval in time, and does not change or update with new data. By adding the time period the report covers and the date on which the report was run, the analyst can indicate when and for how long the data was collected and analyzed, as well as avoid any confusion or ambiguity about the currency or validity of the data. The other components do not need to be added to ensure the report is point-in-time and static. Here is why:

A control group for the phrases is a type of group that serves as a baseline or a reference for comparison with another group that is exposed to some treatment or intervention, such as a target phrase in this case. A control group for the phrases does not need to be added to ensure the report is point-in-time and static, because it does not affect the time frame or the stability of the data. However, a control group for the phrases could be useful for evaluating the effectiveness or impact of the target phrases on customer satisfaction or retention.

A summary of the KPIs is a type of document that provides an overview or a highlight of the key performance indicators (KPIs), which are measurable values that indicate how well an organization or a process is achieving its goals or objectives. A summary of the KPIs does not need to be added to ensure the report is point-in-time and static, because it does not affect the time frame or the stability of the data. However, a summary of the KPIs could be useful for communicating or presenting the main findings or insights from the report.

Filter buttons for the status are a type of feature or function that allows users to select or deselect certain values or categories in a column or a table, such as ticket statuses in this case. Filter buttons for the status do not need to be added to ensure the report is point-in-time and static, because they do not affect the time frame or the stability of the data. However, filter buttons for the status could be useful for exploring or analyzing different aspects or segments of the data.

NEW QUESTION 132

An analyst wants to combine two data sets into a single spreadsheet. Column names from the first spreadsheet are listed in rows in the second spreadsheet. Which of the following is the first step the analyst should take to combine the data sets?

- A. Blend
- B. Merge
- C. Concatenate
- D. Transpose

Answer: C

NEW QUESTION 137

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 142

An employer needs to maintain adequate office staffing during the winter and wants to track storm data. Which of the following data collection methods should the employer use?

- A. Web scraping
- B. Public databases
- C. Observations
- D. Weather surveys

Answer: B

Explanation:

For an employer looking to maintain adequate office staffing during winter while tracking storm data, the most effective method would be to use public databases. These databases often contain comprehensive records of weather patterns and storm data collected and verified by reputable meteorological organizations. Utilizing public databases allows for access to historical and real-time data that is crucial for making informed decisions about staffing during adverse weather conditions.

Web scraping (A) is not the most reliable method, as it may involve extracting data from various websites that might not always provide verified or consistent information. Observations © can be subjective and may not cover a wide enough area to be effective for decision-making on a larger scale. Weather surveys (D) could provide insights, but they are not as immediate or comprehensive as the data available in public databases. References:

? The systematic review on Big Data Analytics in Weather Forecasting suggests that

big data techniques and technologies can manage and analyze the huge volume of weather data from different resources, which supports the use of public databases¹.

? NOAA??s approach to detecting severe weather events using instruments and receiving information from storm spotters indicates the importance of reliable, collected data, which is typically stored in public databases².

? The National Weather Service??s use of observational data collected by various instruments, which are then fed into forecast models, further emphasizes the value of established data collection methods over individual observations or surveys³.

NEW QUESTION 146

Given the below:

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

Which of the following numbers represents a Type I error?

- A. 1
- B. 2
- C. 3
- D. 4

Answer: C

NEW QUESTION 147

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

- A. October 1, 2019 to October 31, 2020
- B. October 31, 2020 to November 1, 2021
- C. November 1, 2019 to October 31, 2020
- D. October 31, 2019 to October 31, 2020

Answer: A

Explanation:

The report should cover the months from October 1, 2019 to October 31, 2020. A rolling 12-month average is a type of moving average that calculates the average of the last 12 months of data for each month. It is useful for smoothing out seasonal fluctuations and identifying long-term trends in the data. To calculate the rolling 12-month average for sales for November 1, 2020, the analyst needs to use the sales data from the previous 12 months, starting from November 1, 2019 and ending on October 31, 2020. The other options are either too short or too long to cover the required period.

NEW QUESTION 152

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 153

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 157

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 159

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 162

A data analyst has been asked to organize the table below in the following ways: By sales from high to low -

By state in alphabetic order -

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

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

- A. Conditional formatting
- B. Grouping
- C. Filtering
- D. Sorting

Answer: D

Explanation:

Sorting is the function that will allow the data analyst to organize the table in the desired manner. Sorting means arranging the data in a specific order, such as ascending or descending, based on one or more criteria. Sorting can be applied to any column in the table, such as sales or state. References: CompTIA Data+ Certification Exam Objectives, page 11

NEW QUESTION 166

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 170

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

- A. ordinal data.
- B. nominal data.
- C. boolean data.
- D. continuous data.

Answer: D

Explanation:

The correct answer is D. Continuous data.

Continuous data is a type of quantitative data that can take any value within a range and can be measured with infinite precision. Continuous data can be expressed as fractions, decimals, or percentages. Examples of continuous data are height, weight, temperature, time, speed, etc¹²

The duration of a phone call in milliseconds is an example of continuous data, because it can take any value within a range (from zero to infinity) and can be measured with infinite precision (up to milliseconds or even smaller units). The duration of a phone call in milliseconds can also be expressed as fractions, decimals, or percentages of a larger unit (such as seconds, minutes, or hours).

Ordinal data is not correct, because ordinal data is a type of qualitative or categorical data that can be ordered or ranked according to some criterion. Ordinal data can have a logical order, but the intervals between the values are not equal or meaningful. Examples of ordinal data are grades, ratings, ranks, etc¹²

Nominal data is not correct, because nominal data is a type of qualitative or categorical data that can be labeled or named without any order or ranking. Nominal data can have a finite number of categories or classes, but the categories have no intrinsic value or hierarchy. Examples of nominal data are gender, color, nationality, etc¹²

Boolean data is not correct, because boolean data is a type of binary data that can have only two possible values: true or false. Boolean data can be used to represent logical statements, conditions, or outcomes. Examples of boolean data are yes/no, on/off, 1/0, etc.

NEW QUESTION 175

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

- A. Delete all incorrect inputs and upload the corrected file.
- B. Have the user manually review the file for data completeness before loading it
- C. Create a data field to data type validator to run the file through prior to import.
- D. Spot-check the file prior to import to catch and correct field errors.

Answer: C

Explanation:

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

NEW QUESTION 180

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 184

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 189

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 194

A data analyst is performing a data merge within a spreadsheet using the tables below:
<https://www.bing.com/images/blob?bcid=S1XCF9p02M4GjpbGxHj0lrlaj9sw.....4c>

Table 1

Last name	Sales
Knox	\$30
Johnson	\$10
Sinclair	\$70

Table 2

Last name	Address
Knox	2851 N. Southport
Johnson	467 Bridle Ridge
Sinclair	1067 Windwood Lane

The analyst is attempting to pull the addresses from Table 2 into Table 1 using the last names and is receiving an error message. Which of the following steps can the analyst perform to fix the error?

- A. Use concatenate to combine the tables.
- B. Ensure the formula is pulling from right to left.
- C. Sort the data by the last name field.
- D. Review the spelling and data type.

Answer: D

Explanation:

The error in merging data from Table 2 into Table 1 using last names could be due to discrepancies in spelling or data type between the two tables. It is essential to ensure that the last names are spelled consistently and that the data types are compatible for a successful merge. Option D suggests reviewing these aspects, which can potentially resolve the error, ensuring that each last name in Table 1 accurately corresponds to the same last name in Table 2, allowing for a successful data pull of addresses.

References: This answer is based on general data analytics practices and does not reference a specific document.

NEW QUESTION 195

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 199

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 203

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

- A. Determine the data needs and review the observations.
- B. Determine the data needs and sources for analysis.
- C. Determine the data needs and schedule interviews.
- D. Determine the data needs and begin the analysis.

Answer: B

Explanation:

After conducting a review of the business questions, the next step for the analyst is to determine the data needs and sources for analysis. This involves identifying the relevant data elements, variables, and metrics that are required to answer the business questions, as well as the data sources, formats, and quality that are available to access and use. This step will help the analyst to plan the data collection, preparation, and integration processes, as well as to assess the feasibility and limitations of the analysis¹.

NEW QUESTION 208

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 213

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 217

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

- A. Determine the data needs and sources for analysis.
- B. Initiate the analysis for exploratory data analysis.
- C. Review the business questions to understand the scope.
- D. Finalize the methodology to solve the problem.

Answer: C

Explanation:

Answer C. Review the business questions to understand the scope.

The first step an analyst should perform upon receiving a business request for analysis is to review the business questions to understand the scope of the problem, the objectives, and the expected outcomes. This will help the analyst to define the analytical approach, identify the data needs and sources, and plan the analysis process. Reviewing the business questions will also help the analyst to communicate with the stakeholders and clarify any assumptions or ambiguities¹.

Option A is incorrect, as determining the data needs and sources for analysis is not the first step, but rather a subsequent step that depends on the business questions and the analytical approach.

Option B is incorrect, as initiating the analysis for exploratory data analysis is not the first step, but rather a part of the analysis process that involves examining and summarizing the data, identifying patterns and outliers, and testing hypotheses.

Option D is incorrect, as finalizing the methodology to solve the problem is not the first step, but rather a later step that involves selecting and applying the appropriate analytical techniques, tools, and models to answer the business questions.

NEW QUESTION 220

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 221

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

- A. Order numbe
- B. salesperso
- C. date shipped, recipient address, and price
- D. Item name, salesperso
- E. recipient address, shipping cos
- F. and date shipped
- G. Item number, item name, salesperso
- H. date sol
- I. and price
- J. Item nam
- K. salesperso
- L. pric
- M. shipping cos
- N. and date shipped

Answer: C

Explanation:

To create a quarterly sales report broken down by unit and week, the most necessary information is the item number, item name, salesperson, date sold, and price. These data elements can help the sales manager to track the sales volume, revenue, and performance of each unit and each week within a quarter. The item number and item name can identify the products or services sold by each unit. The salesperson can indicate the individual or team responsible for each sale. The date sold can show when each sale occurred and how it relates to the weekly and quarterly goals. The price can show how much revenue each sale generated and how it contributes to the unit and quarterly totals.

NEW QUESTION 222

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 226

A military commander would like to see the health scorecards of the troops daily and filter them based on gender and rank. Considering this data is PHI, which of the following would be the best way for the commander to view the information?

- A. An emailed report
- B. A password-protected dashboard
- C. A daily printout of a report
- D. A cloud-hosted spreadsheet

Answer: B

Explanation:

A password-protected dashboard is a type of web-based application that can display the health scorecards of the troops in a secure and interactive way. A password-protected dashboard can provide the following benefits for the commander:

? It can protect the PHI data from unauthorized access or disclosure by requiring a

valid username and password to log in. This can ensure that only the commander and other authorized personnel can view the information¹²

? It can allow the commander to filter the data based on gender and rank by using

drop-down menus, sliders, checkboxes, or other controls. This can enable the commander to customize the view and focus on the relevant data¹³

? It can update the data daily by connecting to a data source that refreshes

automatically or on demand. This can ensure that the commander always sees the latest and most accurate information¹⁴

? It can present the data in a visual and intuitive way by using charts, graphs, tables,

or other elements. This can help the commander to understand and analyze the data more easily and effectively¹

NEW QUESTION 229

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

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

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

- A. Randy
- B. Katie
- C. Ralph
- D. Jean

Answer: B

Explanation:

To compare the students' scores, we need to standardize them by using the z-score formula, which is:

$$z = (x - \mu) / \sigma$$

where x is the raw score, μ is the mean, and σ is the standard deviation. The z-score tells us how many standard deviations a score is above or below the mean. A higher z-score means a better score relative to the average.

Using the table, we can calculate the z-scores for each student as follows:

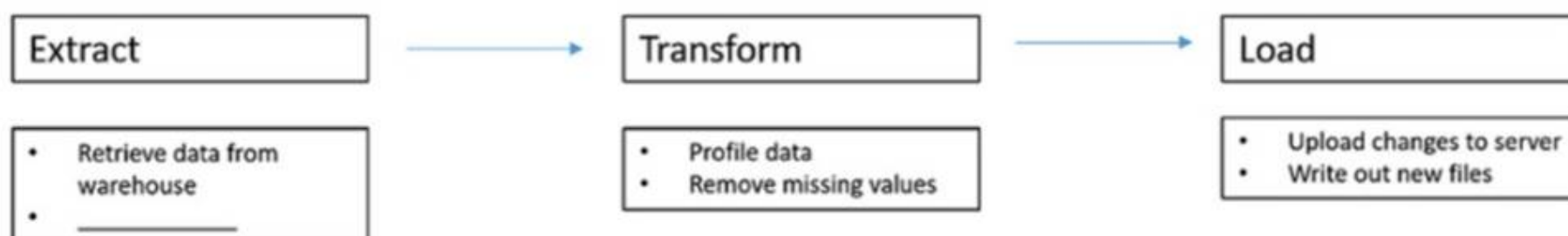
Randy: $z = (76 - 70) / 2 = 3$ Katie: $z = (86 - 80) / 3 = 2$ Ralph: $z = (80 - 75) / 2 = 2.5$ Jean: $z = (80 - 90) / 1 = -10$

The student with the highest z-score is Randy, with a z-score of 3. This means that Randy scored 3 standard deviations above the mean in math, which is the best performance among the four students. Therefore, the correct answer is A.

References: Comparing with z-scores (video) | Z-scores | Khan Academy, 17 Important Data Visualization Techniques | HBS Online

NEW QUESTION 234

Given the diagram below:



Which of the following steps is missing?

- A. Remove redundant data.
- B. Validate the data types.
- C. Connect to the data API.
- D. Normalize the data.

Answer: A

Explanation:

The missing step in the Extract, Transform, Load (ETL) process is typically the cleaning step, which involves removing redundant data or deduplication. This step is crucial in the ETL process to ensure that the data loaded into the destination is accurate and not inflated by duplicate records. The other options, like validating data types and connecting to the data API, are important but do not fit into the standard ETL process steps as a cleaning operation. Normalizing the data is part of the 'Transform' step, which was already listed.

NEW QUESTION 239

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

- A. Simple linear regression
- B. Chi-squared test
- C. Z-test
- D. Two-sample t-test

Answer: B

Explanation:

This is because a chi-squared test is a type of statistical method that tests the association or independence between two or more categorical variables, such as gender, race, or occupation. A chi-squared test can be used to compare the observed frequencies of the categories with the expected frequencies under the null hypothesis of no association or independence. For example, a chi-squared test can be used to determine if there is a relationship between smoking and lung cancer. The other statistical methods do not require two or more categorical variables. Here is why:

Simple linear regression is a type of statistical method that models the relationship between a continuous dependent variable and a continuous or categorical independent variable, such as height, weight, or education level. A simple linear regression can be used to estimate the slope and intercept of the best-fitting line that describes how the dependent variable changes with the independent variable. For example, a simple linear regression can be used to predict the weight of a person based on their height.

Z-test is a type of statistical method that tests the significance of the difference between a sample mean and a population mean, or between two sample means, when the population standard deviation or the sample sizes are large enough. A z-test can be used to compare the average scores of two groups of students on a standardized test.

Two-sample t-test is a type of statistical method that tests the significance of the difference between two sample means when the population standard deviation is unknown or the sample sizes are small. A two-sample t-test can be used to compare the average salaries of two groups of employees in different departments.

NEW QUESTION 241

Which of the following is most likely to be used as a data-mining ETL tool?

- A. SSIS
- B. Stata
- C. SPSS
- D. Cognos

Answer: A

NEW QUESTION 244

Which of the following database schemas features normalized dimension tables?

- A. Flat
- B. Snowflake
- C. Hierarchical
- D. Star

Answer: B

Explanation:

The correct answer is B. Snowflake.

A snowflake schema is a type of database schema that features normalized dimension tables. A database schema is a way of organizing and structuring the data in a database. A dimension table is a table that contains descriptive attributes or characteristics of the data, such as product name, category, color, etc. A normalized table is a table that follows the rules of normalization, which is a process of reducing data redundancy and improving data integrity by organizing the data into smaller and simpler tables¹²

A snowflake schema is a variation of the star schema, which is another type of database schema that features denormalized dimension tables. A denormalized table is a table that does not follow the rules of normalization, and may contain redundant or duplicated data. A star schema consists of a central fact table that contains quantitative measures or facts, such as sales amount, order quantity, etc., and several dimension tables that are directly connected to the fact table. A snowflake schema differs from a star schema in that the dimension tables are further split into sub-dimension tables, creating a snowflake-like shape¹³

A snowflake schema has some advantages and disadvantages over a star schema. Some advantages are:

? It reduces the storage space required for the dimension tables, as it eliminates the redundant data.

? It improves the data quality and consistency, as it avoids the update anomalies that may occur in denormalized tables.

? It allows more detailed analysis and queries, as it provides more levels of dimensions.

Some disadvantages are:

? It increases the complexity and number of joins required to retrieve the data from multiple tables, which may affect the query performance and speed.

? It reduces the readability and simplicity of the schema, as it has more tables and relationships to understand.

? It may require more maintenance and administration, as it has more tables to manage and update¹³

NEW QUESTION 247

A data analyst has been asked to create an ad-hoc sales report for the Chief Executive Officer (CEO).

Which of the following should be included in the report?

- A. The sales representatives' home addresses.
- B. Line-item SKU numbers.
- C. YTD total sales.
- D. The customers' first and last names.

Answer: C

Explanation:

The report for the CEO should include YTD total sales, as this will provide a high-level overview of the sales performance of the company and show how it is meeting its annual goals. The other options are not appropriate for the CEO, as they are either too detailed or irrelevant for the report. The sales representatives' home addresses, line-item SKU numbers, and customers' first and last names are not related to the sales performance and might compromise the privacy and security of the data.

Reference: CompTIA Data+ (DA0-001) Practice Certification Exams | Udemy

NEW QUESTION 248

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

- A. Use scheduled report delivery.
- B. Implement subscription access delivery.
- C. Print out a copy.
- D. Upload the report to the server.

Answer: A

Explanation:

Scheduled report delivery is a feature that allows a data analyst to automate the generation and distribution of a report at a specified time and frequency. This would be the best method to ensure that the sales managers receive the weekly report on sales performance without manual intervention. Subscription access delivery is a feature that allows users to subscribe to a report and access it on demand, but it does not automate the delivery. Printing out a copy or uploading the report to the server are manual methods that require more time and effort from the data analyst. Reference: CertMaster Practice for Data+ Exam Prep - CompTIA

NEW QUESTION 250

An analyst is updating a customer contacts database with information obtained from a survey of new customers. Which of the following data manipulation techniques should the analyst use?

- A. Join
- B. Append
- C. Transform
- D. Blend

Answer: B

NEW QUESTION 255

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

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

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

- A. Include a bar chart using the site and the percentage of new customers data.
- B. Include a line chart using the site and the percentage of new customers data.
- C. Include a pie chart using the site and percentage of new customers data.
- D. Include a scatter chart using the site and the percent of new customers data.

Answer: A

Explanation:

The best type of chart to display the data is A. Include a bar chart using the site and the percentage of new customers data.

A bar chart is a good choice for comparing categorical data with numerical data, such as the site and the percentage of new customers. A bar chart can show the relative differences between the sites and highlight the site with the highest percentage of new customers. A bar chart can also be easily labeled and formatted to make the data clear and understandable.

A line chart is not suitable for this data, because it is used to show trends or changes over time, which is not relevant for the site and the percentage of new customers data. A line chart would also be confusing and misleading, as it would imply a connection or correlation between the sites that does not exist.

A pie chart is also not a good choice for this data, because it is used to show the proportion of a whole, not the comparison of different categories. A pie chart would also be difficult to read and interpret, as it would require labels or legends to identify the sites and their percentages. A pie chart would also not be able to show the exact values of the percentages, only their relative sizes.

A scatter chart is another inappropriate option for this data, because it is used to show the relationship or correlation between two numerical variables, not between a categorical and a numerical variable. A scatter chart would also be cluttered and unclear, as it would plot each site as a point on a coordinate plane, without any labels or axes. A scatter chart would also not be able to show the differences or rankings between the sites and their percentages.

NEW QUESTION 256

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

- A. The temperature of a hot tub
- B. The height of a horse
- C. The time to complete a task
- D. The number of people in an office

Answer: D

Explanation:

A discrete variable is a variable that can only take on a finite number of values, such as integers or categories. The number of people in an office is an example of a discrete variable, as it can only be a whole number. The temperature of a hot tub, the height of a horse, and the time to complete a task are examples of continuous variables, as they can take on any value within a range. Reference: CompTIA Data+ (DAO-001) Practice Certification Exams | Udemy

NEW QUESTION 258

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

- A. 2018 goal data
- B. 2018 actual revenue
- C. 2019 goal data
- D. 2019 commission plan

Answer: C

Explanation:

Answer: C. 2019 goal data

To report how the sales team is performing to target, the sales analyst needs to compare the actual sales revenue with the expected or planned sales revenue for the same period. The 2019 goal data is the file that contains the expected or planned sales revenue for the year 2019, which is the target that the sales team is aiming to achieve. By comparing the 2019 goal data with the 2019 actual revenue, the sales analyst can calculate the performance attainment, which is the percentage of the goal that was met by the sales team.

Option A is incorrect, as 2018 goal data is not relevant for determining 2019 performance attainment. The 2018 goal data contains the expected or planned sales revenue for the year 2018, which is not the target that the sales team is aiming to achieve in 2019.

Option B is incorrect, as 2018 actual revenue is not relevant for determining 2019 performance attainment. The 2018 actual revenue contains the actual sales revenue for the year 2018, which is not comparable with the 2019 goal data or the 2019 actual revenue. Option D is incorrect, as 2019 commission plan is not relevant for determining 2019 performance attainment. The 2019 commission plan contains the rules and rates for calculating and paying commissions to the sales team based on their performance attainment, but it does not contain the expected or planned sales revenue for the year 2019.

NEW QUESTION 259

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

- A. SAS
- B. Microsoft Power BI
- C. IBM SPSS
- D. Python

Answer: D

Explanation:

The option that is a common data analytics tool that is also used as an interpreted, high-level, general-purpose programming language is Python. Python is a popular and versatile programming language that can be used for various purposes, such as web development, software development, automation, machine learning, and data analysis. Python has many features and libraries that make it suitable for data analytics, such as its simple syntax, dynamic typing, multiple paradigms, built-in data structures, NumPy, pandas, matplotlib, scikit-learn, etc. The other options are not programming languages, but software applications or platforms that are used for data analytics or related tasks. SAS is a software suite that provides advanced analytics, business intelligence, data management, and predictive analytics capabilities. Microsoft Power BI is a business analytics service that provides interactive visualizations and business intelligence capabilities. IBM SPSS is a software package that offers statistical analysis, data mining, text analytics, and predictive analytics capabilities. Reference: Python For Data Analysis - DataCamp

NEW QUESTION 260

A financial analyst is creating a daily billing report for a company. One night, the company's data warehouse did not update the data, which caused the data to be reported incorrectly the next day. Which of the following documentation elements should the analyst add to catch this error?

- A. Version number
- B. Data refresh
- C. Frequently asked questions tab
- D. Summary

Answer: B

Explanation:

A data refresh is a documentation element that indicates when the data was last updated or refreshed from the source. A data refresh can help the analyst to catch the error of the data warehouse not updating the data, as it will show a discrepancy between the expected and actual date of the data update. A data refresh can also help the users of the report to verify the timeliness and accuracy of the data, and to avoid making decisions based on outdated or incorrect data

NEW QUESTION 261

A data set for sales per month includes the following data:

Month	Sales (%)
Jan	55
Feb	'60'
March	36
April	70

Which of the following cleaning and profiling methods should be applied to the data set?

- A. Data outliers
- B. Invalid data
- C. Duplicate data
- D. Data type validation

Answer: B

NEW QUESTION 263

Amanda needs to create a dashboard that will draw information from many other data sources and present it to business leaders. Which one of the following tools is least likely to meet her needs?

- A. QuickSight.
- B. Tableau.
- C. Power BI.
- D. SPSS Modeler.

Answer: D

Explanation:

SPSS Modeler.
QuickSight, Tableau, and Power BI are all powerful analytics and reporting tools that can pull data from a variety of sources. SPSS Modeler is a powerful predictive analytics platform that is designed to bring predictive intelligence to decisions made by individuals, groups, systems and your enterprise.

NEW QUESTION 265

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

- A. Conduct an exploratory analysis and use descriptive statistics.
- B. Conduct a trend analysis and use a scatter chart.
- C. Conduct a link analysis and illustrate the connection points.
- D. Conduct an initial analysis and use a Pareto chart.

Answer: A

Explanation:

The first step the analyst should perform with the data is to conduct an exploratory analysis and use descriptive statistics. Exploratory analysis is a type of analysis that aims to summarize the main characteristics of the data, identify patterns, outliers, and relationships, and generate hypotheses for further investigation. Descriptive statistics are numerical measures that describe the central tendency, variability, and distribution of the data, such as mean, median, mode, standard deviation, range, quartiles, etc. Exploratory analysis and descriptive statistics can help the analyst gain a better understanding of the data and its quality, as well as prepare the data for further analysis.

NEW QUESTION 268

Which of the following is the best description of the term "data governance"?

- A. Data governance governs the development of a data visualization dashboard in an organization.
- B. Data governance is the policy that protects against data breaches by cybercriminals.
- C. Data governance is the process of analyzing, manipulating, and reporting data in an organization.
- D. Data governance is the availability, usability, integrity, and security of data in an enterprise.

Answer: D

Explanation:

Data governance refers to the overarching management of data??s availability, usability, integrity, and security within an organization. It involves setting policies and standards that govern data usage, determining data ownership, implementing data security measures, and ensuring that data is accessible for business insights while maintaining its quality. The goal of data governance is to ensure that data is consistent, trustworthy, and not misused, supporting compliance with data privacy regulations and enabling effective data analytics to optimize operations and drive business decision-making.

References:

- ? Understanding Data Governance and Its Importance1.
- ? The Role of Data Governance in Data Management2.
- ? Defining Data Governance and Its Business Value3.

NEW QUESTION 272

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

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

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

- A. \$640,900
- B. \$690,000

- C. \$705,200
D. \$702,500

Answer: C

Explanation:

This is because the price of the Rose unit can be estimated using the average cost per square foot of the original floor plans, which are Jasmine, Orchid, Azalea, and Tulip. To find the average cost per square foot of the original floor plans, we can use the following formula:

$$\text{Average cost per square foot} = \text{Total price} / \text{Total square feet}$$

Plugging in the values from the original floor plans, we get:

$$\text{Average cost per square foot} = (\$345,000 + \$425,000 + \$465,000 + \$525,000) / (1,000 + 1,250 + 1,500 + 2,000)$$

$$\text{Average cost per square foot} = \$1,760,000 / 5,750$$

$$\text{Average cost per square foot} = \$306$$

To find the price of the Rose unit, we can use the following formula:

$$\text{Price} = \text{Square feet} * \text{Average cost per square foot}$$

Plugging in the values from the Rose unit, we get:

$$\text{Price} = 2,300 * \$306$$

$$\text{Price} = \$705,200$$

Therefore, the price of the Rose unit should be \$705,200, using the average cost per square foot of the original floor plans.

NEW QUESTION 273

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

- A. It utilizes key-value pairs.
B. It has undefined fields.
C. It is structured in nature.
D. It uses minimal memory.

Answer: C

Explanation:

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

It utilizes key-value pairs. This is not true for a relational database, because key-value pairs are a way of storing data that associates each value with a unique key, such as an identifier or a name. Key-value pairs are typically used in non-relational databases, such as NoSQL databases, which do not have tables, rows, or columns, but rather store data in various formats, such as documents, graphs, or columns.

It has undefined fields. This is not true for a relational database, because fields are another name for columns in a table, which define the attributes or properties of each row or record in the table. Fields have defined names, types, and lengths in a relational database, which specify the format and size of the data that can be stored in each field.

It uses minimal memory. This is not true for a relational database, because memory is the amount of space or storage that is used by a database to store and process data. Memory usage depends on various factors, such as the size, complexity, and number of tables and queries in a relational database. A relational database can use a lot of memory if it has many tables with many rows and columns, or if it performs complex or frequent queries on the data.

NEW QUESTION 276

A customer survey reveals 90% positive feedback. Which of the following statistical methods would be best to utilize to determine the reliability of a data set and predict how a larger sample of customers over the same time period might respond?

- A. Calculate a high variance on survey responses.
B. Calculate the maximum range of the survey responses.
C. Calculate a low standard deviation on survey responses.
D. Remove any data more than 4 standard deviation from the mean.

Answer: C

Explanation:

A low standard deviation in survey responses indicates that the data points tend to be close to the mean, suggesting a high level of consistency among the responses. This consistency is crucial for determining the reliability of the data set and predicting future outcomes. If the standard deviation is low, it means that the positive feedback is not only high but also consistent, making it a reliable indicator of customer satisfaction and a good predictor of how a larger sample might respond.

References: The concept of using standard deviation to assess data reliability is a standard practice in statistics and data analysis¹²³.

NEW QUESTION 277

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

- A. Scatter plot
- B. Heat map
- C. Pie chart
- D. Infographic

Answer: B

Explanation:

This is because a heat map is a visualization that uses colors to represent different values or intensities of a variable. A heat map can be used to show the most-used/most-clicked portions of a homepage that contains more than 30 links by assigning different colors to each link based on how frequently they are clicked by the users. For example, a link that is clicked very often can be colored red, while a link that is clicked rarely can be colored blue. A heat map can help the analyst to identify which links are more popular or important than others on the homepage. The other visualizations are not as effective as a heat map for this purpose.

Here is why:

A scatter plot is a visualization that uses dots or points to represent the relationship between two variables. A scatter plot cannot show the most-used/most-clicked portions of a homepage that contain more than 30 links because it does not have a clear way of mapping each link to a point on the graph.

A pie chart is a visualization that uses slices or sectors to represent the proportion of each category in a whole. A pie chart cannot show the most-used/most-clicked portions of a homepage that contains more than 30 links because it does not have enough space to display all the categories clearly and accurately.

An infographic is a visualization that uses images, icons, charts, and text to convey information or tell a story. An infographic cannot show the most-used/most-clicked portions of a homepage that contain more than 30 links because it does not have a consistent or standardized way of representing each link and its click frequency.

NEW QUESTION 281

An analyst needs to create an analytics dashboard for an employee intranet site to improve the search functionality, display relevant information, and maintain an updated FAQ page. Which of the following visualizations would best represent what employees are searching for?

- A. A word cloud
- B. A histogram
- C. A pie chart
- D. A scatter plot

Answer: A

Explanation:

A word cloud is an ideal choice for visualizing what employees are searching for on an intranet site. It represents the frequency of word occurrence in a visually impactful way, with more commonly searched terms appearing larger in the cloud. This allows for quick identification of the most popular queries and topics of interest among employees. Unlike histograms, pie charts, or scatter plots, word clouds can effectively display textual data, which is the nature of search queries. They are particularly useful for analyzing text data from surveys or feedback forms, which can be similar to search query data in an intranet environment¹²³⁴.

References: 1: ??What Are Word Clouds? Pros & Cons of Word Cloud Visualizations?? - Alida 2: ??Using Word Clouds for Powerful Data Visualization?? - WordCloud.app blog 3: ??Ultimate Google Data Studio Word Cloud Guide: Visualization 2024?? - AtOnce 4: ??How to Create Word Cloud in Power BI?? - Zebra BI

NEW QUESTION 282

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

- A. The data analyst is not querying the databases correctly.
- B. The databases are recording different events.
- C. The databases are recording the event in different time zones.
- D. The second database is logging incorrectly.

Answer: C

Explanation:

The most likely cause of the issue is that the databases are recording the event in different time zones. For example, if one database is in New York and the other database is in Los Angeles, there is a three-hour difference between them. Therefore, an event that occurs at 12:00 PM in New York would be recorded as 9:00 AM in Los Angeles. To avoid this issue, the databases should either use a common time zone or convert the timestamps to a standard format. Therefore, option C is correct.

Option A is incorrect because the data analyst is not querying the databases incorrectly, but rather observing a discrepancy in the timestamps.

Option B is incorrect because the databases are recording the same event, but with different timestamps.

Option D is incorrect because the second database is not logging incorrectly, but rather using a different time zone.

NEW QUESTION 285

.....

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](#)