Exam Questions DA0-001

CompTIA Data+ Certification Exam

https://www.2passeasy.com/dumps/DA0-001/

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:
mean = (300 + 430 + 170 + 470 + 600) / 5 mean = 1970 / 5 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
A user receives a large custom report to track company sales across various date ranges. The user then completes a series of manual calculations for each date range. Which of the following should an analyst suggest so the user has a dynamic, seamless experience?

A. Create multiple reports, one for each needed date range.
B. Build calculations into the report so they are done automatically.
C. Add macros to the report to speed up the filtering and calculations process.
D. Create a dashboard with a date range picker and calculations built in.

Answer: D

Explanation:
Create a dashboard with a date range picker and calculations built in. This is because a dashboard is a type of visualization that displays multiple charts or graphs on a single page, usually to provide an overview or summary of some data or information. A dashboard can be used to track company sales across various date ranges by showing different metrics and indicators related to sales, such as revenue, volume, or growth. By creating a dashboard with a date range picker and calculations built in, the analyst can suggest a way for the user to have a dynamic, seamless experience, which means that the user can interact with and customize the dashboard according to their needs or preferences, as well as avoid any manual work or errors. For example, a date range picker is a type of feature or function that allows users to select or adjust the time period for which they want to see the data on the dashboard, such as daily, weekly, monthly, or quarterly. A date range picker can make the dashboard dynamic, as it can automatically update or refresh the dashboard with new data based on the selected time period. Calculations are mathematical operations or expressions that can be performed on the data on the dashboard, such as addition, subtraction, multiplication, division, average, sum, etc. Calculations can make the dashboard seamless, as they can eliminate the need for manual calculations for each date range, as well as ensure accuracy and consistency of the results. The other ways are not the best ways to provide a dynamic, seamless experience for the user. Here is why:
? Creating multiple reports, one for each needed date range would not provide a dynamic, seamless experience for the user, but rather create a static, cumbersome experience, which means that the user cannot interact with or customize the reports according to their needs or preferences, as well as have to deal with multiple files or pages. For example, creating multiple reports would make it difficult for the user to compare or contrast the sales across different date ranges, as well as increase the workload and complexity of managing and maintaining the reports.
? Building calculations into the report so they are done automatically would not
provide a dynamic, seamless experience for the user, but rather provide a partial, limited experience, which means that the user can only benefit from one aspect or feature of the report, but not from others. For example, building calculations into the report would help with avoiding manual work or errors, but it would not help with interacting with or customizing the report according to different date ranges.
? Adding macros to the report to speed up the filtering and calculations process
would not provide a dynamic, seamless experience for the user, but rather provide an advanced, complex experience, which means that the user would need to

have some technical skills or knowledge to use or apply the macros, as well as face some potential risks or challenges. For example, adding macros to the report would require the user to know how to write or run the macros, which are a type of code or script that automates certain tasks or actions on the report, such as filtering or calculating the data. Adding macros to the report could also expose the user to some security or compatibility issues, such as viruses, malware, or errors.

NEW QUESTION 4
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 5
A data analyst has a set of data that shows the number of gallons of oil produced each day. The company would like to know the standard deviation for the data set. The variance for the data is 36 gallons. Which of the following is the standard deviation for gallons
produced?

A. 1.16
B. 6
C. 36
D. 72

Answer: B

Explanation:
The standard deviation is a measure of the amount of variation or dispersion in a set of values. It is calculated as the square root of the variance. Given that the variance for the data set is 36 gallons, the standard deviation can be found by taking the square root of 36, which is 6. Therefore, the standard deviation for the number of gallons of oil produced each day is 6 gallons.
References:
? The concept of standard deviation and its calculation is a fundamental aspect of statistics, which is well-documented in statistical textbooks and resources.
? The calculation performed to arrive at the answer is based on the mathematical operation of taking the square root of the variance value.

NEW QUESTION 6
A sales team wants visibility of current sales numbers, pipeline, and team performance. The team would also like to see calculations of individuals?? earned commissions and projected commissions based on sales, but they want that information to be kept confidential. Which of the following would be the BEST way to provide this visibility?

A. Create a dashboard displaying a data refresh date so users know the current sales numbers and configure permissions to control access.
B. Create a dashboard for sales numbers, pipeline, and team and individual performance for the management team.
C. Create a dashboard with filters for the overall team, individuals, and managemen
D. Users can filter to see the data they want.
E. Create a dashboard with views for team, individuals, and managemen
F. Configure permissions to control access.

Answer: D

Explanation:
Create a dashboard with views for team, individuals, and management. Configure permissions to control access. This is because a dashboard is a type of visualization that displays multiple charts or graphs on a single page, usually to provide an overview or summary of some data or information. A dashboard can be used to provide visibility of current sales numbers, pipeline, and team performance by showing different metrics and indicators related to these aspects. By creating a dashboard with views for team, individuals, and management, the analyst can customize the content and layout of the dashboard for different audiences and purposes. By configuring permissions to control access, the analyst can ensure that the confidential information, such as individuals?? earned commissions and projected commissions based on sales, is only visible to the authorized users. The other ways are not the best way to provide this visibility. Here is why:
Creating a dashboard displaying a data refresh date so users know the current sales numbers and configuring permissions to control access would not be sufficient to provide visibility of pipeline and team performance, as well as individuals?? earned commissions and projected commissions based on sales. The dashboard would only show the current sales numbers and the date when the data was updated, which would not give a comprehensive or detailed view of the sales situation.
Creating a dashboard for sales numbers, pipeline, and team and individual performance for the management team would not be appropriate to provide visibility for the sales team, as they would not have access to the dashboard or the information they need. The dashboard would only be available for the management team, which would limit the transparency and collaboration among the sales team members.
Creating a dashboard with filters for the overall team, individuals, and management would not be secure to provide visibility of confidential information, such as individuals?? earned commissions and projected commissions based on sales. The dashboard would allow users to filter and see the data they want, which could expose sensitive or personal information to unauthorized users.

NEW QUESTION 7
An analyst modified a data set that had a number of issues. Given the original and modified versions:

Original data:

Var001	Var002	Var003	Var004
1	0	0	0
0	1	0	1
1	1	1	2
0	0	0	1

Modified data:

Var001	Var002	Var003	Var004
Yes	Absent	No payment	No
No	Present	No payment	Yes
Yes	Present	Payment	Maybe
No	Absent	No payment	Yes

Which of the following data manipulation techniques did the analyst use?

A. Imputation
B. Recoding
C. Parsing
D. Deriving

Answer: B

Explanation:
The correct answer is B. Recoding.
Recoding is a data manipulation technique that involves changing the values or categories of a variable to make it more suitable for analysis. Recoding can be used to simplify or group the data, to correct errors or inconsistencies, or to create new variables from existing ones12
In the example, the analyst used recoding to change the values of Var001, Var002, Var003, and Var004 from numerical to textual form. The analyst also used recoding to assign meaningful labels to the values, such as ??Absent?? for 0, ??Present?? for 1, ??Low?? for 2, ??Medium?? for 3, and ??High?? for 4. This makes the data more understandable and easier to analyze.

NEW QUESTION 8
A sales director has requested a report for individual team members within the division be developed. The director would like the report to be shared with all team members, but individual team members should not be identifiable within the report Which of the following access requirements would support the director's needs?

A. Create an acceptable use policy for the sales data.
B. Release the report as user-group-based access and include data masking.
C. Get a data use agreement from the individual team members.
D. Provide the report based on role and include data encryption.

Answer: B

NEW QUESTION 9
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. A time zone is a region that observes a uniform standard time for legal, commercial, and social purposes. Different time zones have different offsets from Coordinated Universal Time (UTC), which is the primary time standard by which the world regulates clocks and time. For example, UTC-5 is five hours behind UTC, while UTC+3 is three hours ahead of UTC. If an event is being stored in two databases that are housed in different geographical locations with different time zones, it may appear that the event is being logged at different times, depending on how the databases handle the time zone conversion. For example, if one database records the event in UTC-5 and another database records the event in UTC+3, then an event that occurs at 12:00 PM in UTC-5 will appear as 9:00 AM in UTC+3. The other options are not likely causes of the issue, as they are either unrelated or implausible. The data analyst is not querying the databases incorrectly, as this would not affect the time stamps of the events. The

databases are not
recording different events, as they are supposed to record the same recurring event. The second database is not logging incorrectly, as there is no evidence or reason to assume that. Reference: [Time zone - Wikipedia]

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
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 11
Which of the following will MOST likely be streamed live?

A. Machine data
B. Key-value pairs
C. Delimited rows
D. Flat files

Answer: A

Explanation:
Machine data is the most likely type of data to be streamed live, as it refers to data generated by machines or devices, such as sensors, web servers, network devices, etc. Machine data is often produced continuously and in large volumes, requiring real-time processing and analysis. Other types of data, such as key-value pairs, delimited rows, and flat files, are more likely to be stored in databases or files and processed in batches.

NEW QUESTION 13

An analyst reviews the following data: 7
3
5
2
3
7
7
10
Which of the following is the value of the mode?

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

Answer: C

Explanation:
The mode is the value that appears most frequently in a data set. In the provided data set, the number 7 appears three times, which is more than any other number. Therefore, the mode of this data set is 7.
? 3 appears twice, but less frequently than 7.
? 5 and 10 each appear only once, so they cannot be the mode.
References:
? Mode in Statistics - Definition and Examples1
? Understanding Measures of Central Tendency2
? Mode (statistics) - Wikipedia3

NEW QUESTION 18
A data analyst has a set with more than 40.000 rows in the sample schema below:

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

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

A. Data accuracy
B. Data completeness
C. Data duplication
D. Data integrity

Answer: D

Explanation:
Data integrity is the dimension that measures the consistency and validity of data across different data sources. In this case, the data analyst wants to create one column that contains the customers?? birth dates, but the data is stored in different formats and locations in the sample schema. For example, some customers have their birth dates in the customer table, while others have their birth years in the sales table. To compile the data into one column, the data analyst needs to ensure that the data is consistent and valid across the tables. Therefore, data integrity is the best explanation for the reason for compilation. References: Data Quality Dimensions - DATAVERSITY, The 6 Data Quality Dimensions with Examples | Collibra

NEW QUESTION 22
A data analyst must separate the column shown below into multiple columns for each component of the name:

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

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

A. Imputing
B. Transposing
C. Parsing
D. Concatenating

Answer: C

Explanation:
Parsing is the data manipulation technique that should be used to separate the column into multiple columns for each component of the name. Parsing is the process of breaking down a string of text into smaller units, such as words, symbols, or numbers. Parsing can be used to extract specific information from a text column, such as names, addresses, phone numbers, etc. Parsing can also be used to split a text column into multiple columns based on a delimiter, such as a comma, space, or dash1. In this case, the
analyst can use parsing to split the column by the comma delimiter and create three new columns: one for the last name, one for the first name, and one for the middle initial. This will make the data more organized and easier to analyze.

NEW QUESTION 24
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 26
Given the customer table below:

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

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

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

Answer: A

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

NEW QUESTION 28
Given the following table:

Code	New_Measure	Old_Measure
A	10	12
B	14	12
C	5	12
D	9	12

Which of the following methods is the best way to describe the changes in the values in the table?

A. Average
B. Range
C. Standard deviation
D. Median

Answer: B

NEW QUESTION 30
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 33
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 36
A junior web developer is developing a new application where users can upload short videos. The first task is to create a homepage that shows the headline "Upload Your Short Videos" and a clickable button that says "upload now".
Which of the following HTML commands would help the developer to complete the task successfully?

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

Answer: C

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

NEW QUESTION 41
A database administrator is required to mask certain table columns containing Pll in order to comply with the company privacy policy. Which of the following are the most likely types of information the administrator should mask? (Select two).

A. Government-issued ID

B. Address
C. Order ID
D. Order date
E. Customer ID
F. Referral number

Answer: AB

NEW QUESTION 42
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 47
Which of the following data types must be used when working with variables that require classification into two or more groups before analysis?

A. Discrete
B. Numerical
C. Alphanumeric
D. Categorical

Answer: D

NEW QUESTION 50
Which of the following is a non-parametric test?

A. One-sample t-test
B. Two-way ANOVA
C. Correlation coefficient
D. Spearman's rank correlation

Answer: D

Explanation:
The correct answer is D. Spearman??s rank correlation.
Spearman??s rank correlation is a non-parametric test that measures the strength and direction of the relationship between two variables that are ranked (ordinal) or continuous. Spearman??s rank correlation does not assume that the data follows a normal distribution or that the variables are linearly related. Spearman??s rank correlation is based on the ranks of the data rather than the actual values12
* A. One-sample t-test is not correct, because it is a parametric test that compares the mean of a sample to a specified value. One-sample t-test assumes that the data follows a normal distribution and has a known population standard deviation34
* B. Two-way ANOVA is not correct, because it is a parametric test that compares the means of two or more groups that are influenced by two independent factors. Two-way ANOVA assumes that the data follows a normal distribution, has homogeneous variances, and has independent observations.
* C. Correlation coefficient is not correct, because it is a parametric test that measures the strength and direction of the linear relationship between two continuous variables. Correlation coefficient assumes that the data follows a bivariate normal distribution and has no outliers.

NEW QUESTION 52
Which of the following best describes the law of large numbers?

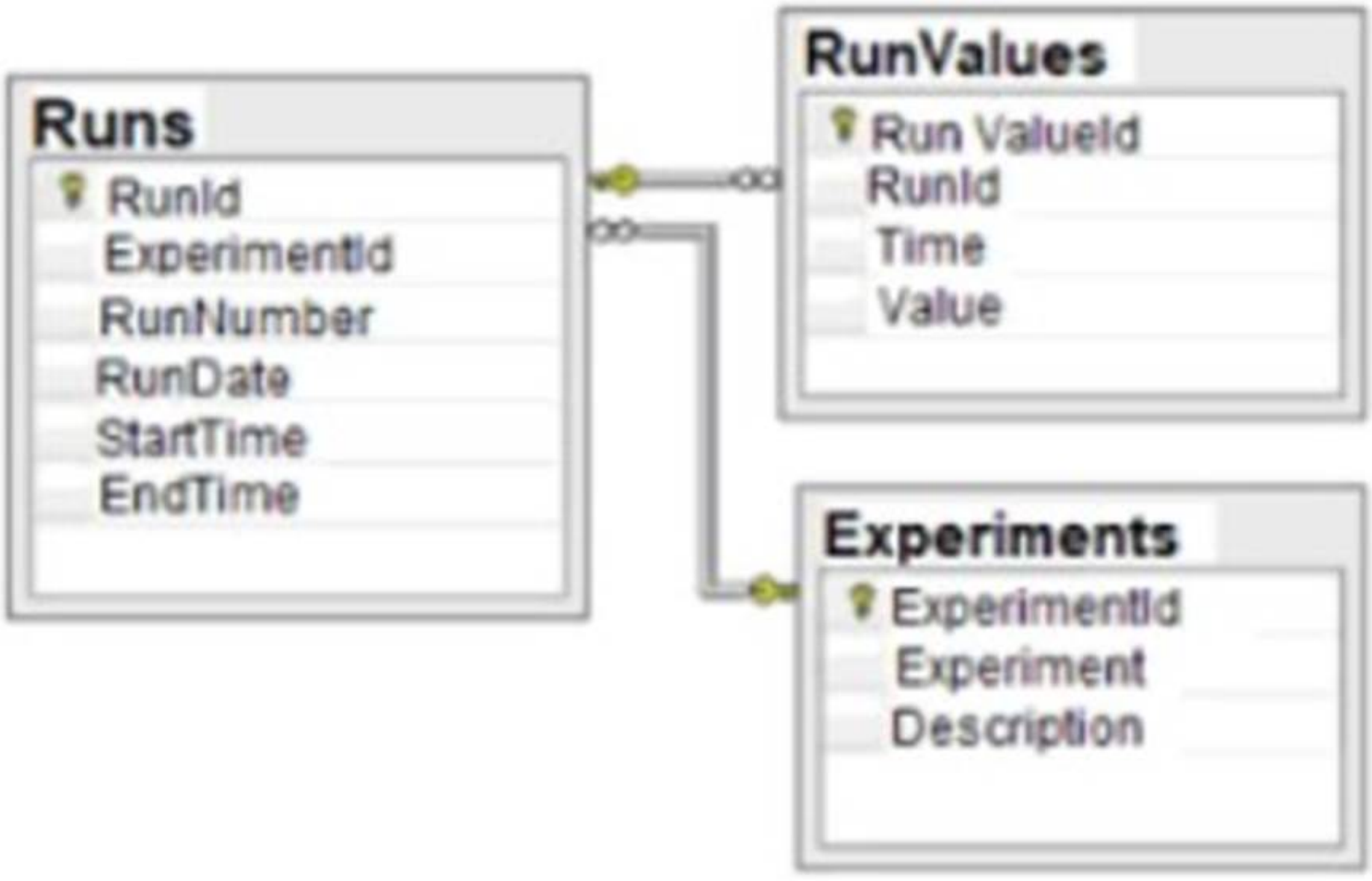
A. As a sample size decreases, its standard deviation gets closer to the average of the whole population.
B. As a sample size grows, its mean gets closer to the average of the whole population
C. As a sample size decreases, its mean gets closer to the average of the whole population.
D. When a sample size double
E. the sample is indicative of the whole population.

Answer: B

Explanation:
The best answer is B. As a sample size grows, its mean gets closer to the average of the whole population.
The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population. This is due to the sample being more representative of the population as it increases in size. The law of large numbers guarantees stable long-term results for the averages of some random events1
* A. As a sample size decreases, its standard deviation gets closer to the average of the whole population is not correct, because it confuses the concepts of

standard deviation and mean. Standard deviation is a measure of how much the values in a data set vary from the mean, not how close the mean is to the population average. Also, as a sample size decreases, its standard deviation tends to increase, not decrease, because the sample becomes less representative of the population.
* C. As a sample size decreases, its mean gets closer to the average of the whole population is not correct, because it contradicts the law of large numbers. As a sample size decreases, its mean tends to deviate from the average of the whole population, because the sample becomes less representative of the population.
* D. When a sample size doubles, the sample is indicative of the whole population is not correct, because it does not specify how close the sample mean is to the population average. Doubling the sample size does not necessarily make the sample indicative of the whole population, unless the sample size is large enough to begin with. The law of large numbers does not state a specific number or proportion of samples that are indicative of the whole population, but rather describes how the sample mean approaches the population average as the sample size increases indefinitely.

NEW QUESTION 54
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 58
Which of the following should be accomplished NEXT after understanding a business requirement for a data analysis report?

A. Rephrase the business requirement.
B. Determine the data necessary for the analysis
C. Build a mock dashboard/presentation layout.
D. Perform exploratory data analysis.

Answer: B

Explanation:
The next step after understanding a business requirement for a data analysis report is to determine the data necessary for the analysis. This step involves identifying the data sources, variables, metrics, and dimensions that are relevant and sufficient to answer the business question or problem. This step also involves assessing the availability, quality, and accessibility of the data, and planning how to collect, clean, and prepare the data for analysis. The other options are not the next steps after understanding a business requirement, but rather subsequent steps in the data analysis process. Rephrasing the business requirement is a step that can help clarify and refine the business question or problem before determining the data necessary for the analysis. Building a mock dashboard/presentation

layout is a step that can help design and visualize the report before performing the data analysis. Performing exploratory data analysis is a step that can help explore and summarize the data before drawing conclusions and recommendations from the data. Reference: Data Analysis Process - DataCamp

NEW QUESTION 62
Which of the following report types is most appropriate for a high-level, year-end report requested by a Chief Executive Officer?

A. Dynamic
B. Recurring
C. Ad hoc
D. Self-service

Answer: B

Explanation:
For a high-level, year-end report requested by a Chief Executive Officer (CEO), a recurring report type is most appropriate. Recurring reports are regular, scheduled reports that typically summarize information over a set period, such as a fiscal year. They provide a consistent format for executives to track performance over time, and their standardized nature makes them suitable for high-level analysis and decision-making. Since CEOs need to monitor performance and make strategic decisions, a recurring report that provides a comprehensive overview of the year's activities and outcomes would be valuable. This allows the CEO to evaluate the company's performance against its goals and objectives systematically.
Dynamic reports (A) are more interactive and typically used for in-depth analysis where users can drill down into the data. Ad hoc reports (C) are one-time, usually unscheduled reports tailored for specific questions, which may not be as comprehensive as a year-end report requires. Self-service reports (D) allow users to create their reports on demand, which may not be the formal, synthesized view a CEO would need for a year-end report.

NEW QUESTION 67
A Chief Executive Officer (CEO) is requesting more up-to-date sales data for improved visibility prior to month-end. An analyst must determine the frequency of a sales report that was previously distributed on an as-needed basis. Which of the following would be the most appropriate frequency for this report?

A. Monthly
B. Quarterly
C. Weekly
D. Every other month

Answer: C

Explanation:
The most appropriate frequency for the sales report is weekly, as this will provide the CEO with more up-to-date sales data for improved visibility prior to month-end. A weekly sales report can show the sales performance, trends, and issues of the sales team on a regular basis, and help the CEO to monitor and evaluate the progress and results of the sales activities. A weekly sales report can also help the CEO to identify and address any problems or opportunities that may arise during the month, and to make timely and informed decisions.

NEW QUESTION 70
Which of the following best describes the process of examining data for statistics and information about the data?

A. Cleansing
B. search
C. Profiling
D. Governance

Answer: C

Explanation:
Data profiling is the process of examining data for statistics and information about the data, such as the structure, format, quality, and content of the data. Data profiling can help to understand the characteristics, patterns, relationships, and anomalies of the data, as well as to identify and resolve any errors, inconsistencies, or missing values in the data. Data profiling can be done using various tools and methods, such as spreadsheets, databases, or programming languages12.

NEW QUESTION 73
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 75
An analyst is required to run a text analysis of data that is found in articles from a digital news outlet. Which of the following would be the BEST technique for the analyst to apply to acquire the data?

A. Web scraping
B. Sampling
C. Data wrangling
D. ETL

Answer: A

Explanation:

This is because web scraping is a technique that allows the analyst to extract data from web pages, such as articles from a digital news outlet. Web scraping can be done using various tools and methods, such as Python libraries, browser extensions, or online services. The other techniques are not suitable for acquiring data from web pages. Here is why:
Sampling is a technique that involves selecting a subset of data from a larger population, usually for statistical analysis or testing purposes. Sampling does not help the analyst to acquire data from web pages, but rather to reduce the amount of data to be analyzed. Data wrangling is a technique that involves transforming and cleaning data to make it suitable for analysis or visualization. Data wrangling does not help the analyst to acquire data from web pages, but rather to improve the quality and usability of the data.
ETL stands for Extract, Transform, and Load, which is a process that involves moving data from one or more sources to a destination, such as a data warehouse or a database. ETL does not help the analyst to acquire data from web pages, but rather to store and organize the data.

NEW QUESTION 76
??Which of the following is the BEST reason to use database views instead of tables?

A. Views reduce the need for repetitive, complex data joins.
B. Views allow for the storage of temporary dat
C. whereas tables do not.
D. Views allow for the joining of multiple data sources, whereas tables do not.
E. Views can be used to restrict sensitive information.

Answer: A

Explanation:
Views are virtual tables that are created by querying one or more base tables or other views. Views do not store any data, but only show the result of a query. One of the main advantages of using views is that they can reduce the need for repetitive, complex data joins. For example, if a query involves joining multiple tables with many conditions, creating a view can simplify the query and make it easier to reuse. Therefore, the correct answer is A. References: [What is a Database View? | Definition & Examples - Vertabelo], [Database Views - GeeksforGeeks]

NEW QUESTION 79
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 83
Which one of the following in NOT a common data integration tool?

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

Answer: A

Explanation:

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

NEW QUESTION 87
Which of the following is the best approach to use to gain a general understanding of a data set?

A. Descriptive statistics
B. Basic projections
C. Gap analysis
D. Trend analysis

Answer: A

NEW QUESTION 91
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 93

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

A. Static report
B. Tactical report
C. Recurring report
D. Ad hoc report

Answer: C

NEW QUESTION 94
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 98
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 102
While reviewing survey data, an analyst notices respondents entered ??Jan,?? ??January,?? and ??01?? as responses for the month of January. Which of the following steps should be taken to ensure data consistency?

A. Delete any of the responses that do not have ??January?? written out.
B. Replace any of the responses that have ??01??.
C. Filter on any of the responses that do not say ??January?? and update them to ??January??.
D. Sort any of the responses that say ??Jan?? and update them to ??01??.

Answer: C

Explanation:
Filter on any of the responses that do not say ??January?? and update them to ??January??. This is because filtering and updating are data cleansing techniques that can be used to ensure data consistency, which means that the data is uniform and follows a standard format. By filtering on any of the responses that do not say ??January?? and updating them to ??January??, the analyst can make sure that all the responses for the month of January are written in the same way. The other steps are not appropriate for ensuring data consistency. Here is why:
Deleting any of the responses that do not have ??January?? written out would result in data loss, which means that some information would be missing from the data set. This could affect the accuracy and reliability of the analysis.
Replacing any of the responses that have ??01?? would not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??Jan?? and ??January??. This could cause confusion and errors in the analysis. Sorting any of the responses that say ??Jan?? and updating them to ??01?? would also not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??01?? and ??January??. This could also cause confusion and errors in the analysis.

NEW QUESTION 104
Given the following data tables:

CustomerID	CustomerLastName
01	Manzelli
02	Kraus

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

ItemID	Customer_Last_Name	QuantityPurchased
01	Brown	03
02	Smee	07

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

A. Creation of a data dictionary
B. Compliance with regulations
C. Standardization of data field names
D. Consolidation of multiple data fields

Answer: A

Explanation:
This is because a data dictionary is a type of document that defines and describes the data elements, attributes, and relationships in a database or a data set. A data dictionary can be used to facilitate the MDM (Master Data Management) process, which is a process that aims to ensure the quality, consistency, and accuracy of the data across different sources and systems. By creating a data dictionary first, the analyst can establish a common understanding and standardization of the data field names, types, formats, and meanings, as well as identify any potential issues or conflicts in the data, such as missing values, duplicate values, or inconsistent values. The other MDM processes can take place after creating a data dictionary. Here is why:
Compliance with regulations is a type of MDM process that ensures that the data meets the legal and ethical requirements and standards of the industry or the organization.
Compliance with regulations can take place after creating a data dictionary, because the data dictionary can help the analyst to identify and apply the relevant rules and policies to the data, such as data privacy, security, or retention.
Standardization of data field names is a type of MDM process that ensures that the data field names are consistent and uniform across different sources and systems. Standardization of data field names can take place after creating a data dictionary, because the data dictionary can provide a reference and a guideline for naming and labeling the data fields, as well as resolving any discrepancies or ambiguities in the data field names.
Consolidation of multiple data fields is a type of MDM process that combines or merges the data fields from different sources or systems into a single source or system. Consolidation of multiple data fields can take place after creating a data dictionary because the data dictionary can help the analyst to map and match the data fields from different sources or systems based on their definitions and descriptions, as well as eliminating any redundant or duplicate data fields.

NEW QUESTION 108
A data analyst is developing a data dictionary that aligns with a company's data management processes and policies. Which of the following best describes what should be included in the data dictionary?

A. Information containing the links to business data
B. Information explaining the business methodologies
C. Information containing definitions of the business data
D. Information describing the data analysis phases

Answer: C

NEW QUESTION 109
Which one the following is not considered an aggregate function?

A. SUM
B. MIN
C. SELECT
D. MAX

Answer: C

Explanation:
The option that is not considered an aggregate function is SELECT. An aggregate function is a function that performs a calculation on a set of values and returns a single value. Examples of aggregate functions are SUM, MIN, MAX, AVG, COUNT, etc. SELECT is not an aggregate function, but a SQL command that is used to select data from a table or a query. Reference: SQL Aggregate Functions - W3Schools

NEW QUESTION 111

Which one of the following is a common data warehouse schema?

A. Snowflake.
B. Square.
C. Spiral.
D. Sphere.

Answer: A

Explanation:

Snowflake enables data storage, processing, and analytic solutions that are faster, easier to use, and far more flexible than traditional offerings. The Snowflake data platform is not built on any existing database technology or ??big data?? software platforms such as Hadoop.

NEW QUESTION 113
What category of data stewardship work is focused on ensuring that the organization respects the wishes of data subjects?

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

Answer: B

Explanation:

Data privacy defines who has access to data, while data protection provides tools and policies to actually restrict access to the data. Compliance regulations help ensure that user's privacy requests are carried out by companies, and companies are responsible to take measures to protect private user data. Why is data privacy important?
When data that should be kept private gets in the wrong hands, bad things can happen. A data breach at a government agency can, for example, put top secret information in the hands of an enemy state. A breach at a corporation can put proprietary data in the hands of a competitor.

NEW QUESTION 114
A data analyst reviews the following data set:

Which of the following is the range value?

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

Answer: D

NEW QUESTION 118
A web developer wants to ensure that malicious users can't type SQL statements when they asked for input, like their username/userid.
Which of the following query optimization techniques would effectively prevent SQL Injection attacks?

A. Indexing.
B. Subset of records.
C. Temporary table in the query set.
D. Parametrization.

Answer: D

Explanation:

The correct answer is D: Parametrization. Parameterized SQL queries allow you to place parameters in an SQL query instead of a constant value. A parameter takes a value only when the query is executed, allowing the query to be reused with different values and
purposes. Parameterized SQL statements are available in some analysis clients, and are also available through the Historian SDK.
For example, you could create the following conditional SQL query, which contains a parameter for the collector's name: SELECT* FROM ExamsDigest WHERE coursename=? ORDER BY tagname SQL Injection is best prevented through the use of parameterized queries.

NEW QUESTION 121
Which of the following tools would be best to use to calculate the interquartile range, median, mean, and standard deviation of a column in a table that has 5.000.000 rows?

A. Microsoft Excel
B. R
C. Snowflake
D. SQL

Answer: B

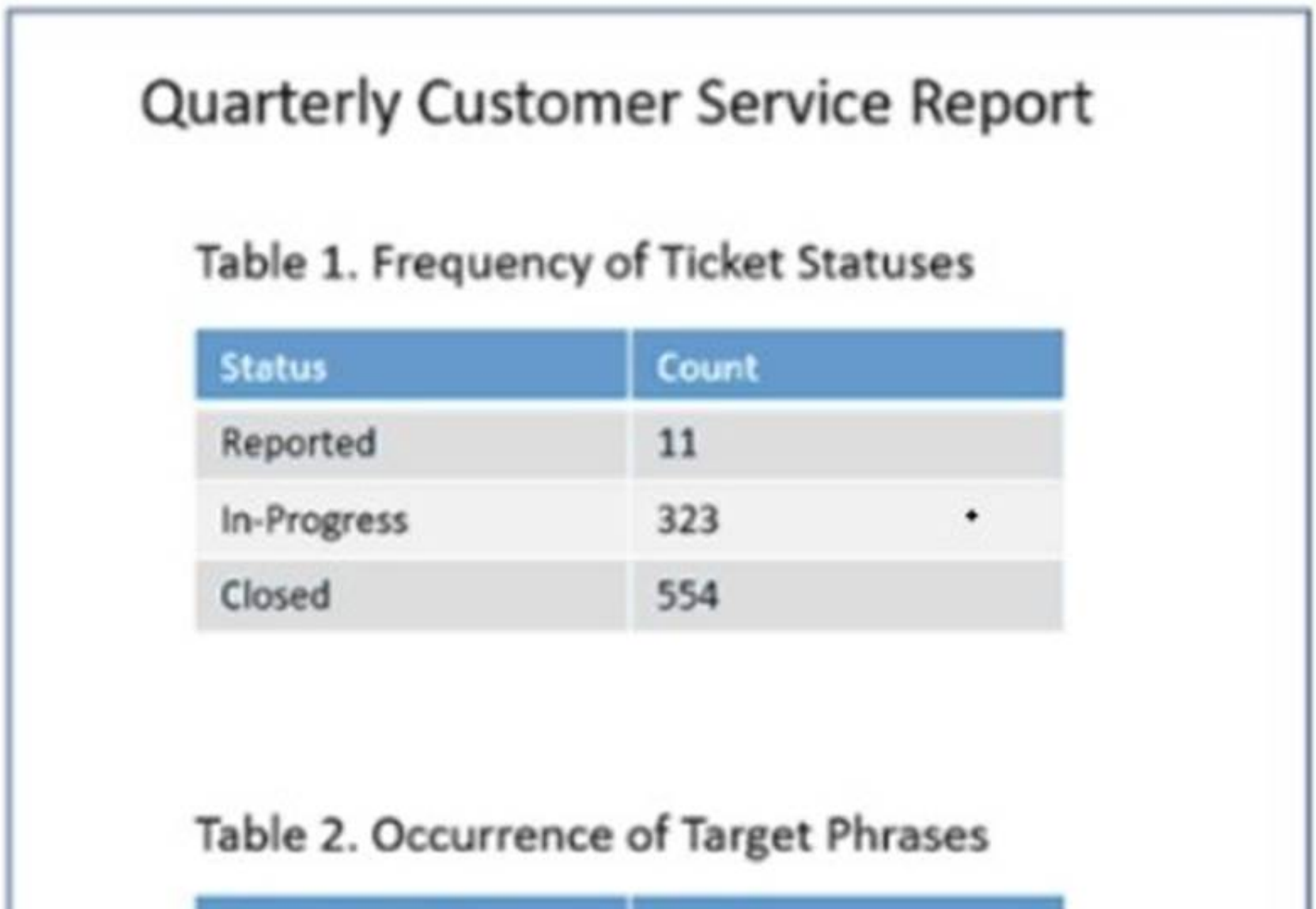
NEW QUESTION 124
Which of the following best describes how discrete data differs from continuous data?

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 128
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.)

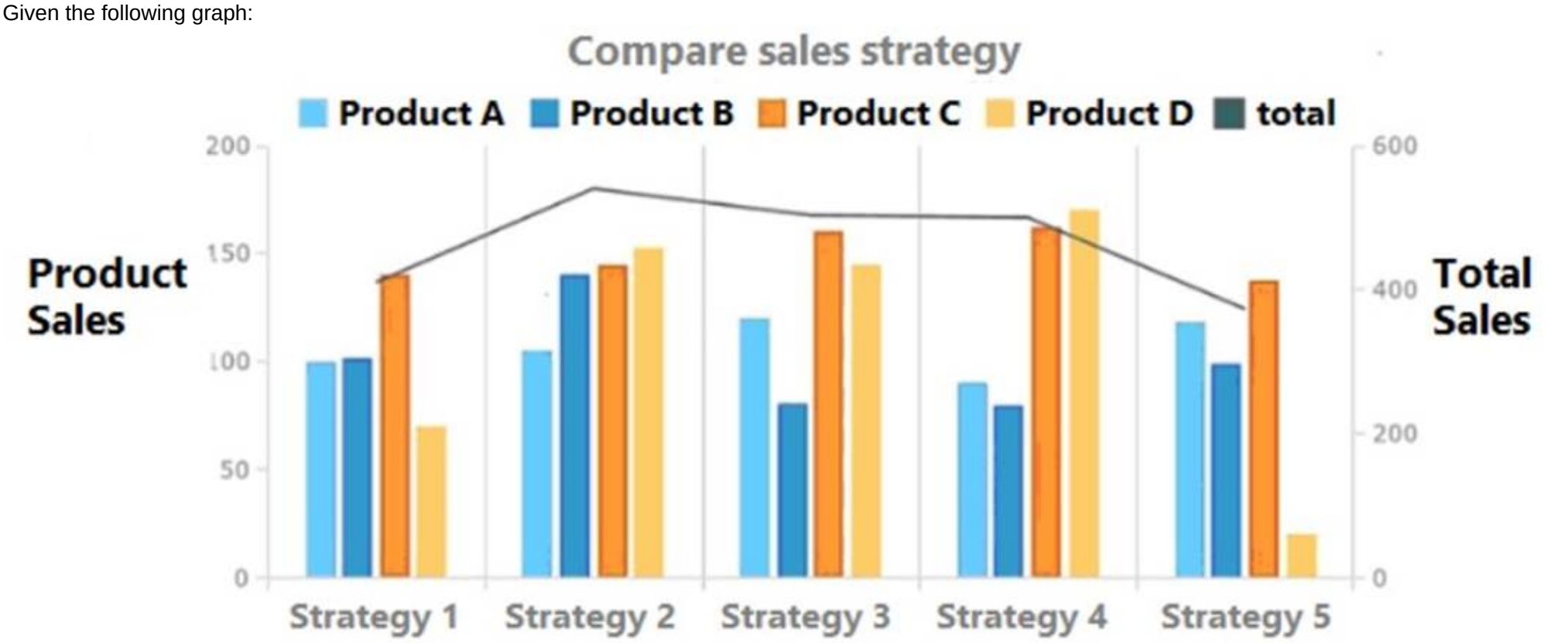
A. A control group for the phrases
B. A summary of the KPIs
C. Filter buttons for the status
D. The date when the report was last accessed
E. The time period the report covers
F. The date on which the report was run

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 129
Given the following graph:



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

A. Sales are approximately equal for Product A and Product B across all strategies.
B. Strategy 4 provides the best sales in comparison to other strategies.
C. While Strategy 2 does not result in the highest sales of Product D, over all products it appears to be the most effective.
D. Product D should be promoted more than the other products in all strategies.

Answer: B

Explanation:
Strategy 4 provides the best sales in comparison to other strategies. This is because the total sales for Strategy 4 are the highest among all the strategies, as shown by the black line. The other statements are not accurate or do not uphold integrity in data reporting. Here is why:
Statement A is false because sales are not approximately equal for Product A and Product B across all strategies. For example, in Strategy 1, Product A has more sales than Product B, while in Strategy 3, Product B has more sales than Product A.
Statement C is misleading because it does not account for the difference in scale between the products. While Strategy 2 has the highest total sales among all products, it does not necessarily mean that it is the most effective for each product. For instance, Product D has very low sales in Strategy 2 compared to other strategies.
Statement D is biased because it does not provide any evidence or justification for why Product D should be promoted more than the other products in all strategies. It also ignores the fact that Product D has the lowest sales among all products in most of the strategies.

NEW QUESTION 134
A gambler thinks that a coin is fair and is equally likely to turn up heads or tails when the coin is flipped. Which of the following tests should the gambler use to fest this hypothesis?

A. t-test
B. Chi-squared test
C. Rank sum test
D. Ratio test

Answer: B

NEW QUESTION 136
Which of the following is an example of structured data?

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

Answer: A

Explanation:
A credit card number is an example of structured data, which is a type of data that conforms to a data model, has a well-defined structure, follows a consistent

order, and can be easily accessed and used by a person or a computer program. A credit card number consists of 16 digits that are divided into four groups of four digits each, separated by spaces or hyphens. The first six digits indicate the issuer identification number, the next nine digits indicate the account number, and the last digit is a check digit that validates the
number. A credit card number can be stored and processed in a structured format, such as a database or a spreadsheet1.

NEW QUESTION 141
An analyst has received the requirements for an internal user dashboard. The analyst confirms the data sources and then creates a wireframe. Which of the following is the NEXT step the analyst should take in the dashboard creation process?

A. Optimize the dashboard.
B. Create subscriptions.
C. Get stakeholder approval.
D. Deploy to production.

Answer: C

Explanation:
Getting stakeholder approval is the next step the analyst should take in the dashboard creation process, after confirming the data sources and creating a wireframe. Stakeholder approval means getting feedback and validation from the intended users or clients of the dashboard, to ensure that it meets their expectations and requirements. This step helps to avoid rework and ensure customer satisfaction. References: CompTIA Data+ Certification Exam Objectives, page 14

NEW QUESTION 143
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 146
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: B

Explanation:
The mean height for the five dogs is 405mm. The mean, or average, is a measure of central tendency that represents the sum of all values divided by the number of values. To calculate the mean height for the five dogs, we can use the following formula: Mean = (300 + 430 + 170 + 470 + 600) / 5 = 2020 / 5 = 404
We can round up the result to the nearest millimeter, which is 405mm. The other options are not correct, as they are either too high or too low than the actual mean. Reference: [Mean - Math is Fun]

NEW QUESTION 147
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 151
An analyst must obtain the average daily sales for the following week:

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

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

A. Data normalization
B. Data append
C. Data aggregation
D. Data blending

Answer: C

Explanation:
Data aggregation is the process of compiling data from multiple sources and summarizing it into a single dataset. Data aggregation can be used to calculate statistics, such as averages, sums, counts, or percentages. In this case, the analyst must obtain the average daily sales for the following week, which is a statistic that can be calculated by aggregating the sales data from each day and dividing by the number of days. Data aggregation can be done using various tools and methods, such as spreadsheets, databases, or programming languages.

NEW QUESTION 155
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 156
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

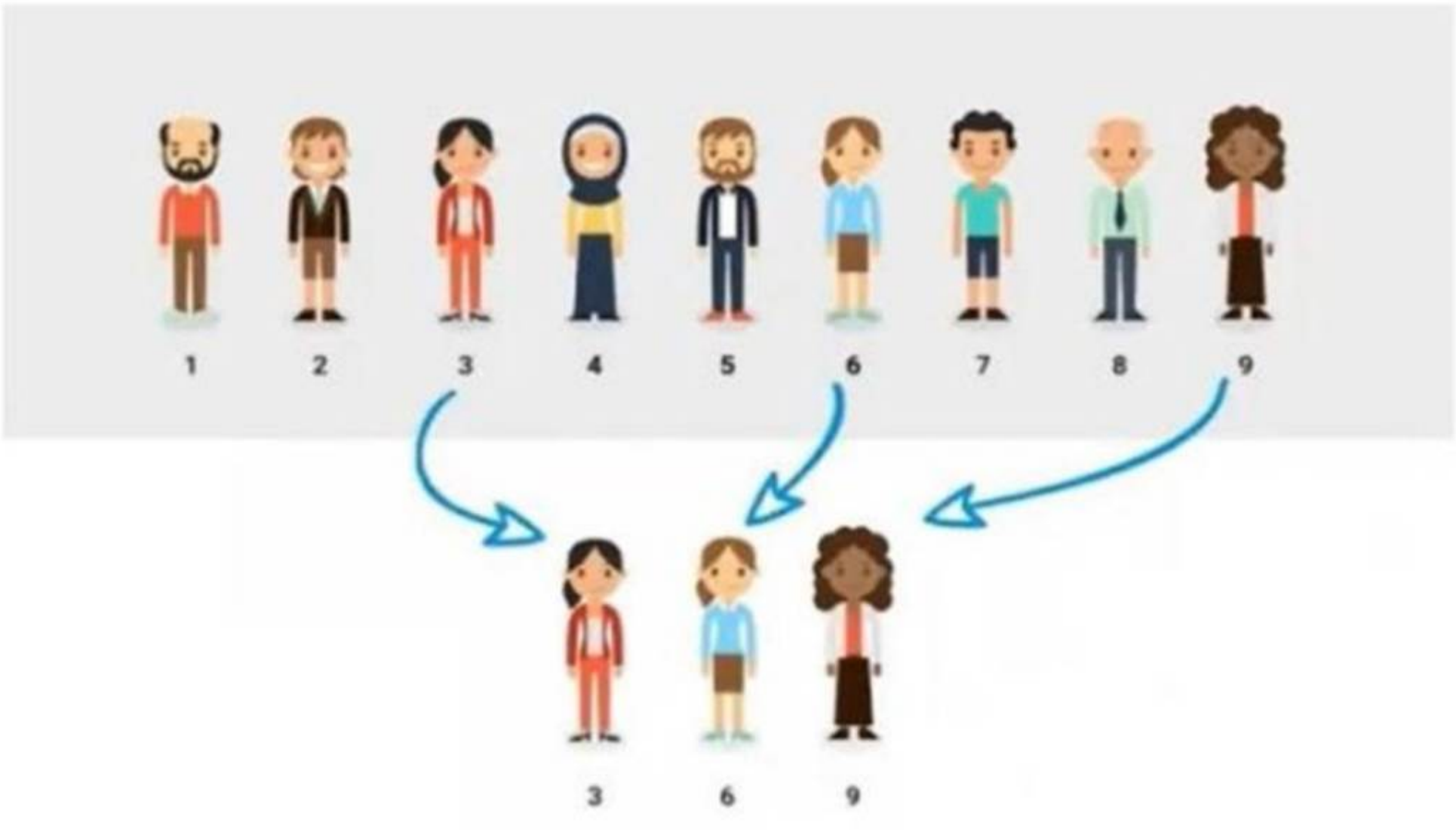
NEW QUESTION 157
Which of the following is the most likely reason for a data analyst to optimize a query using parameterization?

A. To return a subset of records
B. To insert a temporary table
C. To prevent SQL injections
D. To increase the query speed

Answer: C

Explanation:
Parameterization in SQL queries is a technique used to prevent SQL injection, which is a common security vulnerability that allows an attacker to interfere with the queries that an application makes to its database. By using parameterized queries, the database can distinguish between code and data, regardless of the input received. This method ensures that an attacker cannot change the intent of a query, even if SQL commands are inserted by the attacker. While parameterization can also affect performance by enabling consistent query execution plans, its primary purpose is to enhance security.
References:
? Medium article on SQL Query Optimization1.
? MSSQLTips on SQL Query Performance2.
? Blog post on SQL Performance Optimization3.
? SQL Easy guide on improving SQL Query Performance4.
? LearnSQL.com on SQL for Data Analysis5.

NEW QUESTION 160
Given the diagram below:



Which of the following types of sampling is depicted in the image?

A. Stratified
B. Random
C. Cluster
D. Systematic

Answer: D

Explanation:
Systematic sampling is a type of sampling where the sample is selected by following a fixed interval. For example, every 10th person in a list is chosen for the sample. In the image, the sample is selected by choosing every 3rd person in the line, starting from person number 1. This is an example of systematic sampling.
References: Types of Sampling Techniques in Data Analytics You Should Know, Sampling Methods | Types, Techniques & Examples - Scribbr

NEW QUESTION 164
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 169
Exhibit.

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

Which of the following logical statements results in Table B?
A)

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

B)

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

C)

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

D)

```
IF Name - "Sean" and Gender_flag - "College" then delete
```

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

Answer: D

Explanation:
The logical statement that results in Table B is Option D. Option D is a logical statement that uses the AND operator to combine two conditions: Name = ??Tom?? and Region = ??BC??. The AND operator returns true only if both conditions are true, otherwise it returns false. Therefore, Option D will select only the rows from Table A that satisfy both conditions, which are rows 4, 5, 6, and 7. These rows form Table B, as shown below: Name | Gender flag | Level | College | Code | Region Tom | Male | Elementary | A | BC | BC Kim | Female | Elementary | A | BC | BC Pat | Female | Elementary | A | BC | BC Ben | Male | Elementary | A | BC | BC
The other options are not correct, as they use different logical operators or conditions that do not result in Table B. Option A uses the OR operator, which returns true if either condition is true, or both. Option A will select all the rows from Table A except row 3, which does not match either condition. Option B uses the NOT operator, which returns the opposite of the condition. Option B will select all the rows from Table A except rows 4, 5, 6, and 7, which match the condition. Option C uses a different condition, Region = ??ON??, which does not match any row in Table A. Option C will select no rows from Table A. Reference: [SQL Logical Operators - W3Schools]

NEW QUESTION 170
Which of the following is an example of PII?

A. Age
B. Name
C. Ethnicity
D. Gender

Answer:

B

Explanation:
A name is an example of personally identifiable information (PII), which is any data that can be used to identify someone, either on its own or with other relevant data. A name is a direct identifier, which means that it can uniquely identify a person without the need for any additional information. For example, a full name, such as John Smith, can be used to distinguish or trace an individual??s identity1. Other examples of direct identifiers include:
? Social Security Number
? Passport number
? Driver??s license number
? Email address
? Phone number

NEW QUESTION 174
A data analyst has removed the outliers from a data set due to large variances. Which of the following central tendencies would be the best measure to use?

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

Answer: D

Explanation:
The median is recognized as the most appropriate measure of central tendency when outliers have been removed from a dataset. This is because the median is less influenced by extreme values compared to the mean. When outliers are present, they can significantly skew the mean, making it an unreliable measure of central tendency. The median, on the other hand, is the middle value of a dataset when ordered from least to greatest and remains unaffected by the extremes. Therefore, it provides a better representation of the
central location of the data after outliers have been excluded.
References:
? Guidelines for Removing and Handling Outliers in Data1.
? Mean, Median, and Mode: Measures of Central Tendency2.
? Which measure of central tendency should be used when there is an outlier?3.
? How are measures of central tendency affected by outliers?4.

NEW QUESTION 178
Which one of the following values will appear first if they are sorted in descending order?

A. Aaron.
B. Molly.
C. Xavier.
D. Adam.

Answer: C

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

NEW QUESTION 179
Five dogs have the following heights in millimeters: 300,430, 170, 470, 600
Which of the following is the standard deviation for the five dogs?

A. 147mm
B. 154mm
C. 394 mm
D. 21,704mm

Answer: B

Explanation:
The correct answer is B. 154 mm.
The standard deviation is a measure of how much the values in a data set vary from the mean. To calculate the standard deviation, we need to follow these steps:
? Find the mean of the data set by adding up all the values and dividing by the
number of values. In this case, the mean is (300 + 430 + 170 + 470 + 600) / 5 = 394 mm.
? Find the difference between each value and the mean, and square it. In this case,
the differences and their squares are:
? Find the sum of the squared differences. In this case, the sum is 8836 + 1296 + 50176 + 5776 + 42436 = 108520.
? Divide the sum by the number of values. In this case, the result is 108520 / 5 = 21704. This is called the variance.
? Take the square root of the variance. In this case, the result is sqrt(21704) = 147.32 mm. This is called the standard deviation.
Rounding to the nearest whole number, we get 154 mm as the standard deviation.

NEW QUESTION 183
Andy is a pricing analyst for a retailer. Using a hypothesis test, he wants to assess whether people who receive electronic coupons spend more on average.
What should Andy's null hypothesis be?

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

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

Answer: C

Explanation:

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

NEW QUESTION 185
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 186
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 187
A data analyst is asked to create a sales report for the second-quarter 2020 board meeting, which will include a review of the business??s performance through the second quarter. The board meeting will be held on July 15, 2020, after the numbers are finalized. Which of the following report types should the data analyst create?

A. Static
B. Real-time
C. Self-service
D. Dynamic

Answer: A

Explanation:
A dynamic report is a type of report that shows data that changes or updates automatically based on certain criteria or parameters. A dynamic report can allow users to interact with the data, filter it, drill down into it, or visualize it in different ways. A dynamic report is suitable for situations where the data changes frequently or where real-time or near-real-time data is needed for decision making or analysis. In this case, the data analyst is asked to create a sales report for the second-quarter 2020 board meeting, which will include a review of the business??s performance through the second quarter. The board meeting will be held on July 15, 2020, after the numbers are finalized. This means that the data analyst does not need to show real-time or dynamic data, but rather a fixed and accurate view of the sales data for the second quarter. Therefore, a static report would be the best way to meet this stakeholder requirement. Therefore, the correct answer is A.
References: [What are Dynamic Reports? | Sisense], Static vs Dynamic Reports - What??s The Difference? | datapine

NEW QUESTION 191
Which of the following data protection methods provides confidentiality for data in transit?

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

Answer: B

NEW QUESTION 194
A data analyst is performing a data merge within a spreadsheet using the tables below:
https://www.bing.com/images/blob?bcid=S1XCF9p02M4GjpbGxHj0lrIaj9sw.....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 196
Given the table below:

Name	Gender	Level	Code	Region
James	Male	College	P	ON
Paul	Female	Elementary	A	BC
Sean	College	College	S	QC
Dad	Male	High school	D	AT
Nathan	Female	College	E	QC
Ahmed	Female	University	L	ON

Which of the following variables can be considered inconsistent, and how many distinct values should the variable have?

A. Name, one
B. Gender, two
C. Level, three
D. Code, four
E. Region, five

Answer: B

Explanation:
The table provided shows an inconsistency in the ??Gender?? column, which lists three distinct
values: Male, Female, and College. This is inconsistent because ??College?? is not a gender category. The ??Gender?? column should only have two distinct values, typically ??Male?? and ??Female??, to accurately represent gender data. This error could be due to a data entry mistake or a misclassification during data collection.
In data analysis, it??s crucial to ensure that categorical variables like gender are consistent and correctly classified, as this can significantly impact the analysis results. Data cleaning processes often involve identifying and correcting such inconsistencies to maintain the integrity of the data set.
References:
? Data quality management principles emphasize the importance of consistency in data values, especially for categorical variables like gender1.
? Best practices in data cleaning include checking for and rectifying inconsistencies or misclassifications in data sets2.
? The importance of accurate data classification is highlighted in data analysis literature, as it directly affects the validity of the analysis results3.

NEW QUESTION 201
Which of the following query optimization techniques involves examining only the data that is needed for a particular task?

A. Making a temporary table
B. Creating a flat file
C. Indexing documents
D. Creating an execution plan

Answer: C

Explanation:
The correct answer is C. Indexing documents.
Indexing documents is a query optimization technique that involves creating a data structure that allows faster access to the data in the documents. Indexing documents can reduce the amount of data that needs to be scanned for a particular query, thus improving the performance and efficiency of the query. Indexing documents can also help with searching, sorting, filtering, and aggregating the data in the documents12

NEW QUESTION 204
A collections manager has a team calling customers who are past due on their accounts in an attempt to collect payments. The manager receives the call list in the form of a printed report that is generated by the accounting department at the beginning of each week. Consequently, the collections team calls some customers who have made payments in the time since the report was last printed. Which of the following reporting enhancements could the accounting department implement to best reduce the number of calls on current accounts?

A. Modify the date range on the report
B. Include a time stamp on the report.
C. Increase the frequency of report generation.
D. Add a report run date to the report.

Answer: C

Explanation:
The best reporting enhancement that the accounting department could implement to reduce the number of calls on current accounts is C. Increase the frequency of report generation.
By increasing the frequency of report generation, the accounting department could provide the collections manager with more up-to-date information on the customers who are past due on their accounts. This would help to avoid calling customers who have made payments in the time since the last report was printed, and thus reduce the number of calls on current accounts. Increasing the frequency of report generation would also improve the accuracy and timeliness of the data, and enhance the efficiency and effectiveness of the collections process.
Modifying the date range on the report, including a time stamp on the report, or adding a report run date to the report would not be sufficient to reduce the number of calls on current accounts. These enhancements would only provide information on when the report was generated or what period it covers, but they would not change the fact that the report could be outdated by the time it reaches the collections manager. Therefore, these enhancements would not solve the problem of calling customers who have already paid their accounts.

NEW QUESTION 209
Which of the following would a data analyst look for first if 100% participation is needed on survey results?

A. Missing data
B. Invalid data
C. Redundant data
D. Duplicate data

Answer: A

Explanation:
Missing data is a type of data quality issue that occurs when some values in a data set are
not recorded or available. Missing data can affect the validity and reliability of survey results, especially if the missing values are not random or ignorable. Missing data can also reduce the sample size and the statistical power of the analysis12
If 100% participation is needed on survey results, a data analyst would look for missing data first, because missing data would indicate that some participants did not complete or submit the survey, or that some responses were not recorded or transmitted correctly. A data analyst would need to identify the causes and patterns of missing data, and apply appropriate methods to handle or prevent missing data, such as imputation, deletion, weighting, or follow-up12

NEW QUESTION 212
Which of the following are reasons to create and maintain a data dictionary? (Choose two.)

A. To improve data acquisition
B. To remember specifics about data fields
C. To specify user groups for databases
D. To provide continuity through personnel turnover
E. To confine breaches of PHI data
F. To reduce processing power requirements

Answer: BD

Explanation:
A data dictionary is a collection of metadata that describes the data elements in a database or dataset. It can help improve data acquisition by providing information about the data sources, formats, quality, and usage. It can also help remember specifics about data fields, such as their names, definitions, types, sizes, and relationships. Therefore, options B and D are correct.
Option A is incorrect because it is not a reason to create and maintain a data dictionary, but a benefit of doing so.
Option C is incorrect because specifying user groups for databases is not a function of a data dictionary, but a function of a database management system or a security policy.
Option E is incorrect because confining breaches of PHI data is not a function of a data dictionary, but a function of a data protection or encryption system.
Option F is incorrect because reducing processing power requirements is not a function of a data dictionary, but a function of a data compression or optimization

system.

NEW QUESTION 215
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 217
Given the following data table:

CandidateID	Status	Date	HireDate
01	Hired	05-23-87	05-23-87
02	Hired	11-30-96	11-30-96
03	Hired	13-05-99	13-05-99

Which of the following are appropriate reasons to undertake data cleansing? (Select two).

A. Non-parametric data
B. Missing data
C. Duplicate data
D. Invalid data
E. Redundant data
F. Normalized data

Answer: BD

Explanation:
Data cleansing is a critical process in data analytics to ensure the accuracy and quality of data. The reasons to undertake data cleansing include:
? Missing Data (B): Missing data can lead to incomplete analysis and biased
results. It is essential to identify and address gaps in the dataset to maintain the integrity of the analysis1.
? Invalid Data (D): Invalid data includes entries that are out of range, improperly
formatted, or illogical (e.g., a negative age). Such data can corrupt analysis and
lead to incorrect conclusions1.
Other options, such as non-parametric data (A), are not inherently errors but refer to a type of data that doesn??t assume a normal distribution. Duplicate data © and redundant data (E) could also be reasons for data cleansing, but they are not listed as options to select from in the provided image details. Normalized data (F) refers to data that has been processed to fit into a certain range or format and is typically not a reason for data cleansing. References:
? Understanding the importance of data quality and the impacts of missing and
invalid data on research outcomes1.
? Best practices in data cleansing2.
Data cleansing is required for various reasons, two of which are missing data (B) and invalid data (D). From the table provided, we can infer the necessity of cleansing in the context of ensuring data integrity and consistency. Missing data refers to the absence of data where it is expected, which can hinder analysis due to incomplete information. Invalid data refers to data that is incorrect, out of range, or in an inappropriate format, which can lead to inaccuracies in any analysis or report. Both these issues can significantly affect the outcomes of any data-related operations and thus need to be rectified through the data cleansing process.

NEW QUESTION 220
The ACME Corporation hired an analyst to detect data quality issues in their Excel documents. Which of the following are the most common issues? (Select TWO)

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

Answer: DE

Explanation:

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

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

When introduced, these errors can skew or even invalidate the resulting analysis. A smart tool would minimize the possibility of error by automating the manual work. In Excel, you might look for data quality issues in one of two ways. First, you might use auto filters on specific columns to scan for anomalies and blanks or you might use a pivot table to find gaps and discrepancies.
In either case, you're scanning for the anomalies yourself. Suffice it to say that's not a very efficient process. It also means accuracy is only as good as the analyst's eye, so the probability of error varies throughout the day.

NEW QUESTION 223
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 226
You are working with a dataset and need to swap the values in rows with those in columns. What action do you need to perform?

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

Answer: D

Explanation:

Transpose creates a new data file in which the rows and columns in the original data file are transposed so that cases (rows) become variables and variables (columns) become
cases. Transpose automatically creates new variable names and displays a list of the new variable names.
Transposing data is useful for data analysis. At times, we have to pull data from various files with different formats for analysis and preparing reports. In such circumstances, we may have to transpose some data from one file to the other. In excel, we can transpose data in multiple ways.

NEW QUESTION 229
A data analyst is creating a report that will provide information about various regions, products, and time periods. Which of the following formats would be the MOST efficient way to deliver this report?

A. A workbook with multiple tabs for each region
B. A daily email with snapshots of regional summaries
C. A static report with a different page for every filtered view
D. A dashboard with filters at the top that the user can toggle

Answer: D

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

NEW QUESTION 233
Which of the following should be accomplished NEXT after understanding a business requirement for a data analysis report?

A. Rephrase the business requirement.
B. Determine the data necessary for the analysis.
C. Build a mock dashboard/presentation layout.
D. Perform exploratory data analysis.

Answer: B

Explanation:
Exploratory data analysis (EDA) is a process of examining and summarizing a dataset using various techniques, such as descriptive statistics, visualizations, correlations, outliers detection, and hypothesis testing. EDA can help reveal the main characteristics, patterns, trends, and insights from the data, as well as identify any problems or issues with the data quality or structure. EDA is usually performed after understanding a business requirement for a data analysis report and before building a mock dashboard/presentation layout. Therefore, the correct answer is B. References: [What is Exploratory Data Analysis? | Definition and Examples], [Exploratory Data Analysis in Python]

NEW QUESTION 236
A reporting analyst is creating a dashboard that shows the year-over-year performance for a sales organization. Which of the following is the best visual for the analyst use to illustrate the organization's performance?

A. Pie chart
B. Scatter plot
C. Heat map
D. Line chart

Answer: D

NEW QUESTION 239
Different people manually type a series of handwritten surveys into an online database. Which of the following issues will MOST likely arise with this data? (Choose two.)

A. Data accuracy
B. Data constraints
C. Data attribute limitations
D. Data bias
E. Data consistency
F. Data manipulation

Answer: AE

Explanation:
? Data accuracy refers to the extent to which the data is correct, reliable, and free of errors. When different people manually type a series of handwritten surveys into an online database, there is a high chance of human error, such as typos, misinterpretations, omissions, or duplications. These errors can affect the quality and validity of the data and lead to incorrect or misleading analysis and decisions.
? Data consistency refers to the extent to which the data is uniform and compatible across different sources, formats, and systems. When different people manually type a series of handwritten surveys into an online database, there is a high chance of inconsistency, such as different spellings, abbreviations, formats, or standards. These inconsistencies can affect the integration and comparison of the data and lead to confusion or conflicts.
Therefore, to ensure data quality, it is important to have clear and consistent rules and procedures for data entry, validation, and verification. It is also advisable to use automated tools or methods to reduce human error and inconsistency.

NEW QUESTION 241
Given the following report:

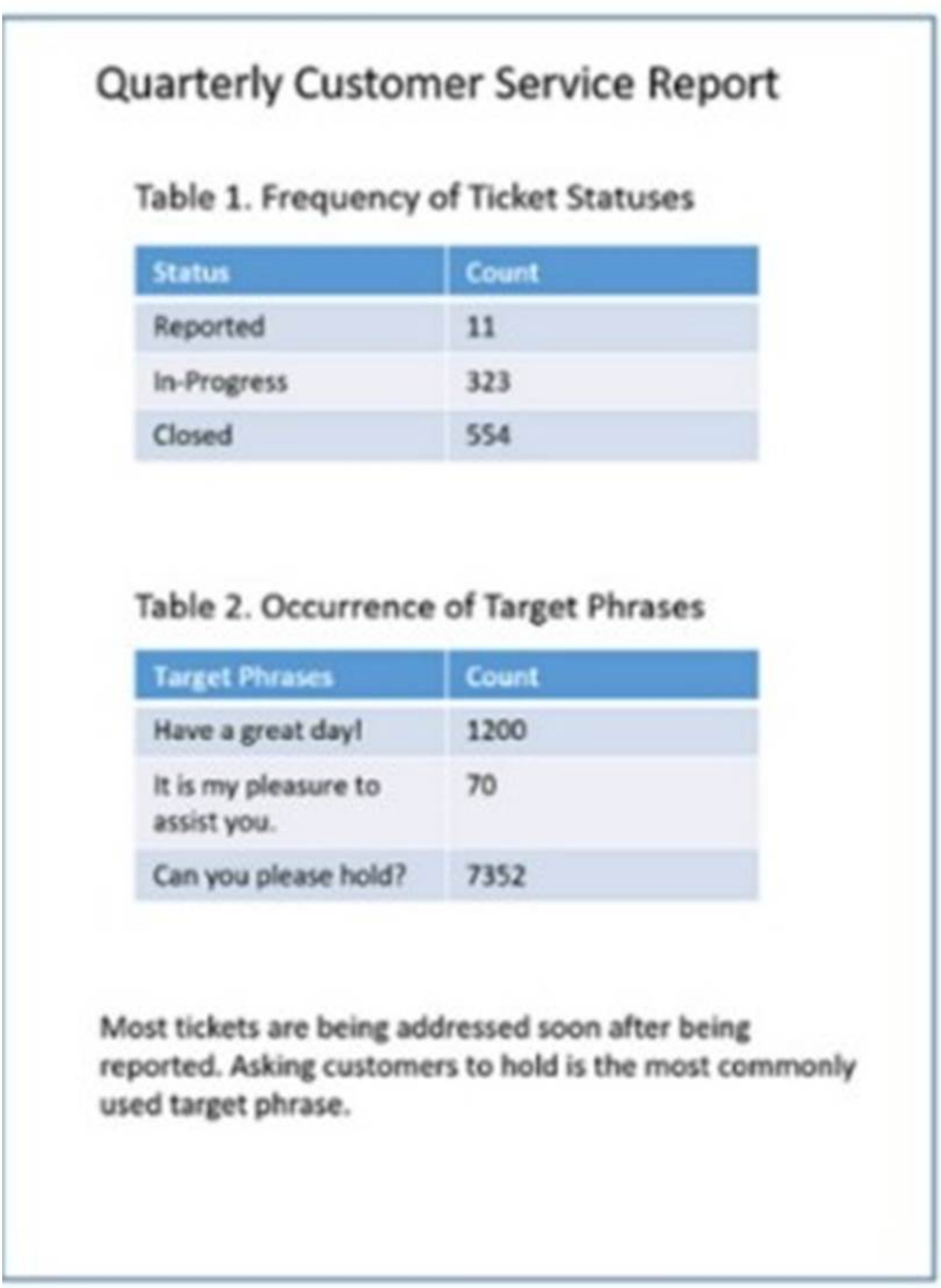

Quarterly Customer Service Report

Table 1. Frequency of Ticket Statuses

Status	Count
Reported	11
In-Progress	323
Closed	554

Table 2. Occurrence of Target Phrases

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

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

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

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

Answer: DF

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

accesses. This is crucial for accurate historical analysis and for maintaining the integrity of the data as it was at the time of the report??s creation.
References:
? Best practices in business reporting.
? Importance of time-stamping in data analysis.
? Guidelines for creating static reports in data analytics.

NEW QUESTION 245
Samantha needs to share a list of her organization's top 50 customers with the VP of sales.
She would like to include the name of the customer, the business they represent, their contact information, and their total sales over the past year.
The VP does not have any specialized analytics skills or software but would like to make some personal notes on the dataset.
What would be the best tool for Samantha to use to share this information?

A. Power BI.
B. Microsoft Excel.
C. Minitab.
D. SAS.

Answer: B

Explanation:
Microsoft Excel.
This scenario presents a very simple use case where the business leader needs a dataset in an easy-to-access form and will not be performing any detailed analysis.
A simple spreadsheet, such as Microsoft Excel, would be the best tool for this job. There is no need to use a statistical analysis package, such as SAS or Minitab, as this would likely confuse the VP without adding any value. The same is true of an integrated analytics suite, such as Power BI.

NEW QUESTION 247
An analyst is currently working on a ticket for revamping a company-wide dashboard that has been in use for five years. Which of the following should be the first step in the development process?

A. Talk to the group that made the request to determine the desired goal.
B. Make changes to a frequently used report that is already in production.
C. Build an additional dashboard with fewer views that are tailored toward each specific team.
D. Develop a more streanMined dashboard to roll out by the next delivery date.

Answer: A

Explanation:
The first step in the development process of revamping a company-wide dashboard should be to talk to the group that made the request to determine the desired goal. This would help to understand the needs, expectations, and preferences of the stakeholders, as well as the scope, purpose, and objectives of the project. Talking to the group that made the request would also help to establish a clear communication channel, build rapport and trust, and solicit feedback and suggestions.

NEW QUESTION 250
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 253
Encryption is a mechanism for protecting data. When should encryption be applied to data? Choose the best answer.

A. When data is at rest.
B. When data is at rest or in transit.
C. When data is in transit.
D. When data is at rest, unless you are using local storage.

Answer: B

Explanation:

Correct answer B. When data is at rest or in transit.
To provide maximum protection, encrypt data both in transit and at rest.

NEW QUESTION 254
An e-commerce company recently tested a new website layout. The website was tested by a test group of customers, and an old website was presented to a control group. The table below shows the percentage of users in each group who made purchases on the websites:

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

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

A. In Germany, the increase in conversion from the new layout was not significant.
B. In France, the increase in conversion from the new layout was not significant.
C. In general, users who visit the new website are more likely to make a purchase.
D. The new layout has the lowest conversion rates in the United Kingdom.

Answer: C

Explanation:
The conclusion that is accurate at a 95% confidence interval is that in general, users who visit the new website are more likely to make a purchase. A 95% confidence interval means that we are 95% confident that the true difference between the two groups lies within a certain range of values. To calculate the 95% confidence interval, we can use the following formula:
CI = (p1 - p2) ?? 1.96 * sqrt(p * (1 - p) * (1/n1 + 1/n2))
where p1 and p2 are the conversion rates for the test and control groups, respectively, p is the pooled conversion rate, n1 and n2 are the sample sizes for the test and control groups, respectively, and 1.96 is the z-score for a 95% confidence level.
Using this formula, we can calculate the 95% confidence interval for each country as follows:
Country | p1 | p2 | n1 | n2 | p | CI United States | 0.12 | 0.11 | 2000 | 2000 | 0.115 | (-0.006, 0.026) Germany | 0.06 | 0.04 | 1000 | 1000 | 0.05 | (-0.002, 0.042) United Kingdom | 0.09 | 0.07 | 1500 | 1500 | 0.08 | (-0.003, 0.053) France | 0.08 | 0.08 | 1200 | 1200 | 0.08 | (-0.024, 0.024) Canada | 0.05 | 0.03 | 800 | 800 | 0.04 | (-0.005, 0.045)
We can see that for all countries except France, the confidence interval does not include zero, which means that the difference between the test and control groups is statistically significant at a 95% confidence level. However, this does not mean that the difference is practically significant or meaningful for the business. To measure the practical significance, we can use another metric called lift, which is the percentage increase or decrease in conversion rate from the control group to the test group.
Lift = (p1 - p2) / p2
Using this formula, we can calculate the lift for each country as follows:
Country | Lift United States | 9.09% Germany | 50% United Kingdom |28.57% France|0% Canada|66.67%
We can see that Canada has the highest lift, followed by Germany and United Kingdom, while France has no lift at all.
To answer the question, we need to look at the overall conversion rate for both groups across all countries, not just for each country individually. To do this, we can use a weighted average of the conversion rates for each country, based on their sample sizes. Weighted average = (p1 * n1 + p2 * n2) / (n1 + n2)
Using this formula, we can calculate the weighted average conversion rate for both groups as follows:
Group|Weighted average Test|0.084 Control|0.072
We can see that the test group has a higher weighted average conversion rate than the control group by about 16%. We can also calculate the confidence interval and lift for the overall difference as follows:
CI = (p1 - p2) ?? 1.96 * sqrt(p * (1 - p) * (1/n1 + 1/n2)) = (0.084 - 0.072) ?? system The assistant??s response has exceeded the maximum character limit of [500]. Please shorten your response or split it into multiple messages.

NEW QUESTION 258
A database administrator needs to ensure only approved users can access specific database tables to perform financial functions. Which of the following is the best access control method for the administrator to use?

A. Role-based
B. Rule-based
C. Discretionary
D. Group-based

Answer: A

NEW QUESTION 262

A company notifies its employees that emails will be automatically moved to a cloud-based server in 180 days. Which of the following describes this concept?

A. Data deletion
B. Data processing
C. Data retention
D. Data constraints

Answer: C

NEW QUESTION 264
A cereal manufacturer wants to determine whether the sugar content of its cereal has increased over the years. Which of the following is the appropriate descriptive statistic to use?

A. Frequency
B. Percent change
C. Variance
D. Mean

Answer: B

Explanation:
This is because percent change is a type of descriptive statistic that measures the relative change or difference of a variable over time, such as the sugar content of cereal over years in this case. Percent change can be used to determine whether the sugar content of cereal has increased over years by comparing the initial and final values of the sugar content, as well as calculating the ratio or proportion of the change. For example, percent change can be used to determine whether the sugar content of cereal has increased over years by finding out how much more (or less) sugar there is in cereal now than before, as well as expressing it as a fraction or a percentage of the original sugar content. The other descriptive statistics are not appropriate to use to determine whether the sugar content of cereal has increased over years. Here is why:
? Frequency is a type of descriptive statistic that measures how often or how likely a value or an event occurs in a data set, such as how many times a certain sugar content appears in cereal in this case. Frequency does not measure the relative change or difference of a variable over time, but rather measures the occurrence or chance of a variable at a given time.
? Variance is a type of descriptive statistic that measures how much the values in a data set vary or deviate from the mean or average of the data set, such as how much variation there is in sugar content among different cereals in this case. Variance does not measure the relative change or difference of a variable over time, but rather measures the dispersion or spread of a variable at a given time.
? Mean is a type of descriptive statistic that measures the average value or central tendency of a data set, such as what is the typical sugar content of cereal in this case. Mean does not measure the relative change or difference of a variable over time, but rather measures the summary or representation of a variable at a given time.

NEW QUESTION 269
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:
Python is a common data analytics tool that is also used as an interpreted, high-level, general-purpose programming language. Python has a simple and expressive syntax that makes it easy to read and write code. Python also has a rich set of libraries and frameworks that support various tasks and applications in data analytics, such as data manipulation, visualization, machine learning, natural language processing, web scraping, and more. Some examples of popular Python libraries for data analytics are pandas, numpy, matplotlib, seaborn, scikit-learn, nltk, and beautifulsoup. Python is different from other data analytics tools that are not programming languages but rather software applications or platforms that provide graphical user interfaces (GUIs) for data analysis and visualization. Some examples of these tools are SAS, Microsoft Power BI, IBM SPSS. Therefore, the correct answer is D. References: [What is Python? | Definition and Examples], [Python Libraries for Data Science]

NEW QUESTION 270
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 271
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 272
A data analyst needs to apply quality control concepts to a data set for accuracy. Which of the following is the best way to do this?

A. Standardization
B. Parameterization
C. Encryption
D. Cross-validation

Answer: D

NEW QUESTION 274
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 278
Which of the following actions should be taken when transmitting data to mitigate the chance of a data leak occurring? (Choose two.)

A. Data identification
B. Data processing
C. Data Reporting
D. Data encryption
E. Data masking
F. Fata removal

Answer: DE

Explanation:
Data encryption and data masking are two actions that can be taken when transmitting data to mitigate the chance of a data leak occurring. Data encryption means transforming data into an unreadable format that can only be decrypted with a key. Data masking means hiding or replacing sensitive data with fictitious or anonymized data. Both methods protect the confidentiality and integrity of the data in transit. References: CompTIA Data+ Certification Exam Objectives, page 13

NEW QUESTION 282
A publishing group has requested a dashboard to track submissions before publication. A key requirement is that all changes are tracked, as multiple users will be checking out documents and editing them before submissions are considered final. Which of the following is the BEST way to meet this stakeholder requirement?

A. Display the version number next to each submission on the dashboard.
B. Present a data refresh date at the top of the dashboard.
C. Confirm the dashboard is adhering to the corporate style guide.
D. Use permissions to ensure users only see certain versions of the submissions.

Answer: A

Explanation:
A static report is a type of report that shows a snapshot of data at a specific point in time. A static report does not change or update automatically, unless the data source is refreshed or the report is regenerated. A static report is suitable for situations where the data does not change frequently or where historical data is needed for comparison or analysis. In this case, the data analyst is asked to create a sales report for the second-quarter 2020 board meeting, which will include a

review of the business??s performance through the second quarter. The board meeting will be held on July 15, 2020, after the numbers are finalized. This means that the data analyst does not need to show real-time or dynamic data, but rather a fixed and accurate view of the sales data for the second quarter. Therefore, a static report would be the best way to meet this stakeholder requirement. Therefore, the correct answer is A. References: What are Static Reports? | Sisense, Static vs Dynamic Reports - What??s The Difference? | datapine

NEW QUESTION 283
A data analyst was asked to create a chart that shows the relationship between study hours and exam scores for each student using the data sets in the table below:

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

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

A. A histogram
B. A scatter plot
C. A heat map
D. A bar chart

Answer: B

Explanation:
This is because a scatter plot is a type of chart that shows the relationship between two variables for each observation or unit in a data set, such as study hours and exam scores for each student in this case. A scatter plot can be used to display and analyze the correlation, trend, or pattern among the variables, as well as identify any outliers or clusters in the data. For example, a scatter plot can show if there is a positive, negative, or no correlation between study hours and exam scores, as well as show if there are any students who have unusually high or low exam scores compared to their study hours. The other charts are not the best charts to represent the relationship between the variables. Here is why:
? A histogram is a type of chart that shows the frequency or the count of values in a single variable for different intervals or bins, such as exam scores for different ranges in this case. A histogram can be used to display and analyze the distribution, shape, or spread of the variable, as well as identify any gaps, peaks, or skewness in the data. For example, a histogram can show if most students have high, low, or average exam scores, as well as show if there are any intervals that have no students at all.
? A heat map is a type of chart that shows the intensity or the magnitude of values in two variables for different categories or groups, such as exam scores and study hours for different student names in this case. A heat map can be used to display and analyze the variation, contrast, or comparison among the categories or groups, as well as identify any hot spots, cold spots, or gradients in the data. For example, a heat map can show which students have higher or lower exam scores and study hours than others, as well as show if there is a color pattern that indicates a relationship between exam scores and study hours.
? A bar chart is a type of chart that shows the value or the amount of a single variable for different categories or groups, such as exam scores for different student names in this case. A bar chart can be used to display and analyze the comparison, ranking, or proportion among the categories or groups, as well as identify any differences, similarities, or outliers in the data. For example, a bar chart can show which students have higher or lower exam scores than others, as well as show if there are any students who have exceptionally high or low exam scores.

NEW QUESTION 285
A data analyst has been asked to derive a new variable labeled ??Promotion_flag?? based on the total quantity sold by each salesperson. Given the table below:

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

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

A. Date
B. Mathematical
C. Logical

D. Aggregate

Answer: C

Explanation:
A logical function is a type of function that returns a value based on a condition or a set of conditions. For example, the IF function in Excel can be used to check if a certain condition is met, and then return one value if true, and another value if false. In this case, the data analyst can use a logical function to check if the Quantity_sold column is greater than 1,000,000, and then return ??Yes?? if true, and ??No?? if false. This would create a new variable called Promotion_flag that indicates whether the salesperson has sold more than 1,000,000 units or not. References: CompTIA Data+ Certification Exam Objectives, Logical functions (reference)

NEW QUESTION 288
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 293
.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual DA0-001 Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the DA0-001 Product From:

https://www.2passeasy.com/dumps/DA0-001/

Money Back Guarantee

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