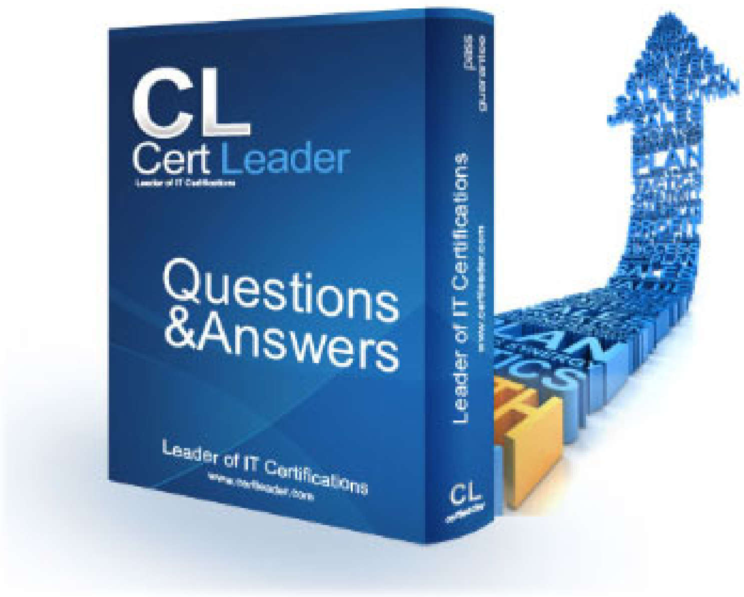
DA0-001 Dumps

CompTIA Data+ Certification Exam

https://www.certleader.com/DA0-001-dumps.html



NEW QUESTION 1
An analyst wants to check the progress and performance regarding the number of customers an organization served in the last six years. Which of the following represents the type of analysis the analyst should perform?

A. Correlation analysis
B. Trend analysis
C. Regression analysis
D. Descriptive analysis

Answer: B

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

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

Answer: A

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

NEW QUESTION 4
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 5
You are working with a professional statistician to perform an analysis and would like to use a statistics package.
Which one of the following would be the most appropriate?

A. Rapid Miner.
B. QLIK.
C. Power BI.
D. Minitab.

Answer: D

Explanation:

Minitab is statistical analysis software. It can be used for learning about statistics as well as statistical research. Statistical analysis computer applications have the advantage of being accurate, reliable, and generally faster than computing statistics and drawing graphs by hand.

NEW QUESTION 6
An analysts building a monthly report for production and wants to ensure the audience is aware of its once-a-month cadence. Which of the following is the MOST important to convey that information?

A. The date of the dashboard build
B. The data refresh date
C. A report summary
D. Frequently asked questions

Answer: A

Explanation:
This is because the date of the dashboard build is the most important component to convey that information, which is the once-a-month cadence of the monthly report for production. The date of the dashboard build can convey that information by indicating when the dashboard was created or updated, as well as showing the frequency or interval of the dashboard creation or update. For example, the date of the dashboard build can convey that information by displaying a date format that includes the month and year, such as January 2020, February 2020, etc., or by displaying a text format that includes the word ??monthly??, such as Monthly Report for Production - January 2020, Monthly Report for Production - February 2020, etc. The other components are not the most important components to convey that information. Here is why:
? The data refresh date is a component that indicates when the data on the dashboard was refreshed or retrieved from the source or system, such as a database, a cloud service, or a web application. The data refresh date does not convey that information, but rather conveys how current or up-to-date the data on the dashboard is.
? A report summary is a component that provides an overview or a highlight of the main findings or insights from the dashboard, such as key metrics, indicators, or trends. A report summary does not convey that information, but rather conveys what the dashboard is about or what it shows.
? Frequently asked questions is a component that provides answers or explanations to common or expected questions from the audience or users of the dashboard, such as how to use or interpret the dashboard, what are the assumptions or limitations of the dashboard, etc. Frequently asked questions does not convey that information, but rather conveys how to understand or interact with the dashboard.

NEW QUESTION 7
A site reliability team wants to monitor the stability of their website. so they can proactively diagnose issues when they occur Which of the following deliverables would best suit their needs?

A. A self-serve dashboard of website performance that updates in real time
B. A weekly log report of site visits and user actions
C. A portal that is refreshed daily and reports errors classified by type
D. A daily summary email indicating website outages for the previous day

Answer: A

Explanation:
The best deliverable that would suit the site reliability team??s needs is A. A self-serve dashboard of website performance that updates in real time.
A self-serve dashboard is a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance. A self-serve dashboard of website performance that updates in real time would allow the site reliability team to easily and quickly access the information they need about the stability of their website, such as uptime, response time, error rate, traffic volume, etc. A self-serve dashboard would also enable the team to proactively diagnose issues when they occur, by providing alerts, notifications, or drill-down options. A self-serve dashboard would also be more interactive and engaging than a report or an email.
A weekly log report of site visits and user actions would not be a good deliverable for the site reliability team??s needs, because it would not provide timely or relevant information about the stability of their website. A weekly log report would be too infrequent and delayed to monitor and diagnose issues when they occur. A weekly log report would also focus on the behavior and actions of the users, rather than the performance and functionality of the website.
A portal that is refreshed daily and reports errors classified by type would not be a good deliverable for the site reliability team??s needs, because it would not provide real-time or comprehensive information about the stability of their website. A portal that is refreshed daily would be too slow and outdated to monitor and diagnose issues when they occur. A portal that reports errors classified by type would be too narrow and limited to capture the full picture of the website performance.
A daily summary email indicating website outages for the previous day would not be a good deliverable for the site reliability team??s needs, because it would not provide real-time or actionable information about the stability of their website. A daily summary email would be too late and retrospective to monitor and diagnose issues when they occur. A daily summary email indicating website outages would also be too passive and generic to help the team resolve or prevent issues in the future.

NEW QUESTION 8
Jhon is working on an ELT process that sources data from six different source systems.
Looking at the source data, he finds that data about the sample people exists in two of six systems.
What does he have to make sure he checks for in his ELT process? Choose the best answer.

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

Answer: C

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

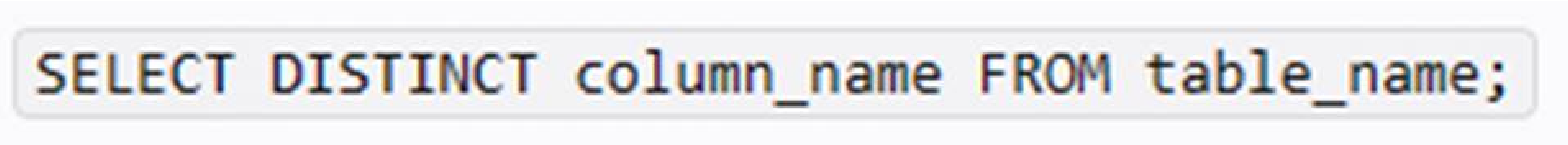
NEW QUESTION 9

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

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

Answer: B

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



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

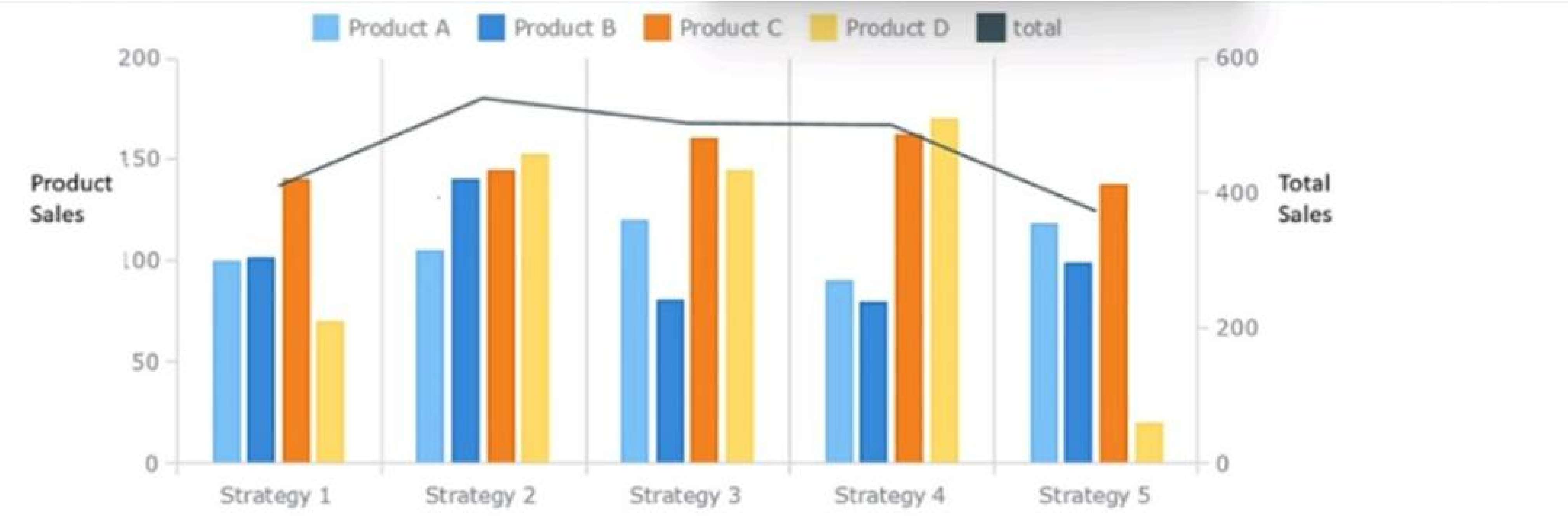
The other data cleansing issues will not be fixed by applying a DISTINCT function. Here is why:
Missing data refers to data that is absent or incomplete in a data set, which can affect the accuracy and reliability of the analysis. A DISTINCT function does not help with missing data, because it does not fill in or impute the missing values.
Redundant data refers to data that is unnecessary or irrelevant for the analysis, which can affect the efficiency and performance of the analysis. A DISTINCT function does not help with redundant data, because it does not remove or filter out the redundant values.
Invalid data refers to data that is incorrect or inaccurate in a data set, which can affect the validity and reliability of the analysis. A DISTINCT function does not help with invalid data, because it does not validate or correct the invalid values.

NEW QUESTION 10
A 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 10
Which of the following summary statements upholds integrity in data reporting?



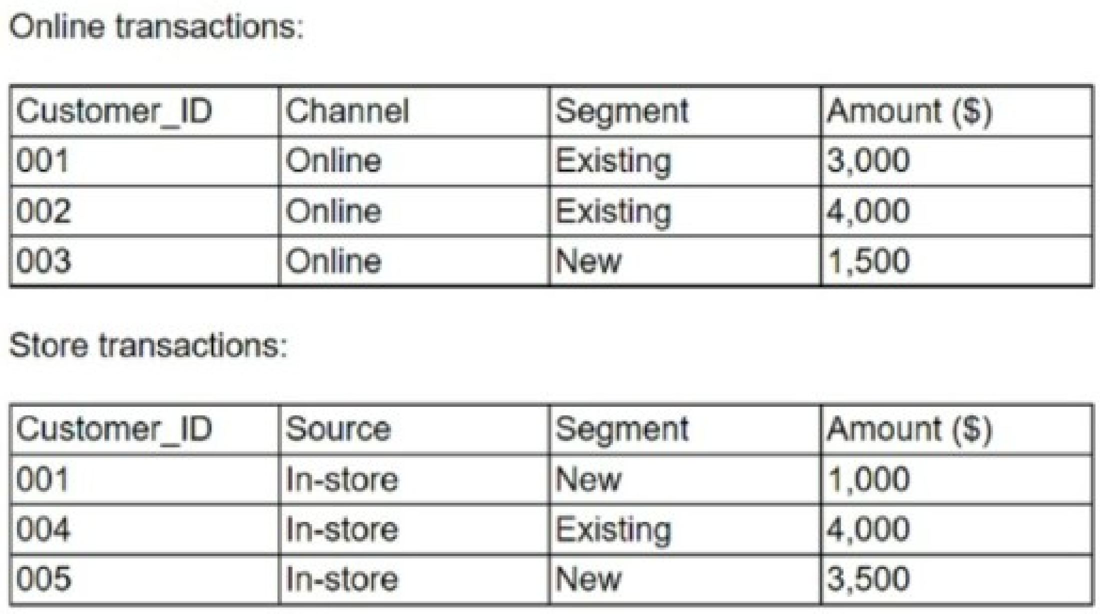
A. Sales are approximately equal for Product A and Product B across all strategies.
B. Strategy 4 provides the best sales in comparison to other strategies.
C. While Strategy 2 does not result in the highest sales of Product
D. over all products it appears to be the most effective.
E. Product D should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

Answer: C

Explanation:
Answer: C. While Strategy 2 does not result in the highest sales of Product D. over all products it appears to be the most effective.
A summary statement that upholds integrity in data reporting should be accurate, unbiased, and supported by evidence. Option C is the only statement that meets these criteria, as it reflects the data shown in the bar graph without exaggerating or distorting it. Option C also acknowledges the limitation of the statement by using the word ??appears??, which indicates that there may be other factors or variables that affect the sales performance.
Option A is inaccurate, as sales are not approximately equal for Product A and Product B across all strategies. Product A has higher sales than Product B in strategies 1, 3, and 5, while Product B has higher sales than Product A in strategies 2 and 4.
Option B is biased, as it does not consider the sales of different products in each strategy. Strategy 4 provides the best sales for Product B, but not for the other products. Strategy 5 has the highest total sales across all products, as shown by the black line graph.

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

NEW QUESTION 12
A marketing analytics team received customer transaction data from two different sources. The data is complete and accurate; however, the field names appear to be inconsistent. Given the following tables:



Online transactions:

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

Store transactions:

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

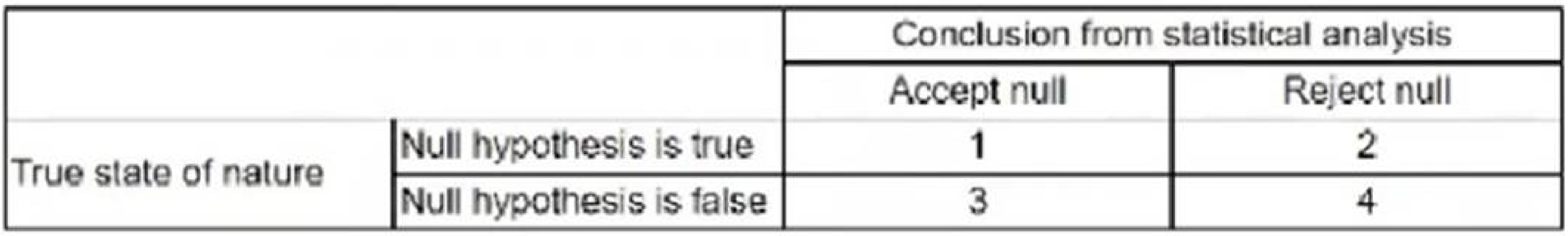
Which of the following is considered best practice if the team wants to consolidate the files and conduct further analysis?

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

Answer: A

Explanation:
When consolidating data from different sources, it is crucial to standardize field names to ensure consistency across datasets. This process involves aligning the field names so that they are the same in both tables, which simplifies the merging of data and subsequent analysis. Standardizing field names helps in maintaining data integrity and avoids confusion that may arise from having different names for the same data point. Recode the data values (B) would not be necessary unless the data values themselves are inconsistent or in different formats. Overwriting the field names in one of the tables © could
lead to loss of information or confusion. Editing the field names in the data dictionary (D) is helpful, but it does not address the immediate need to harmonize the field names in the actual datasets.
References:
? Best practices in data management.
? Principles of data integration and consolidation.

NEW QUESTION 17
Given the table below:



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

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

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

Answer: C

Explanation:
A Type II error is a false negative conclusion, which means failing to reject a null hypothesis that is actually false. In the table, box 3 indicates that a Type II error has occurred, because it shows that the null hypothesis is accepted when it is false in reality.
This means that the statistical test failed to detect a significant difference or relationship that actually exists. References: Type I & Type II Errors | Differences, Examples, Visualizations - Scribbr, Type I and type II errors - Wikipedia

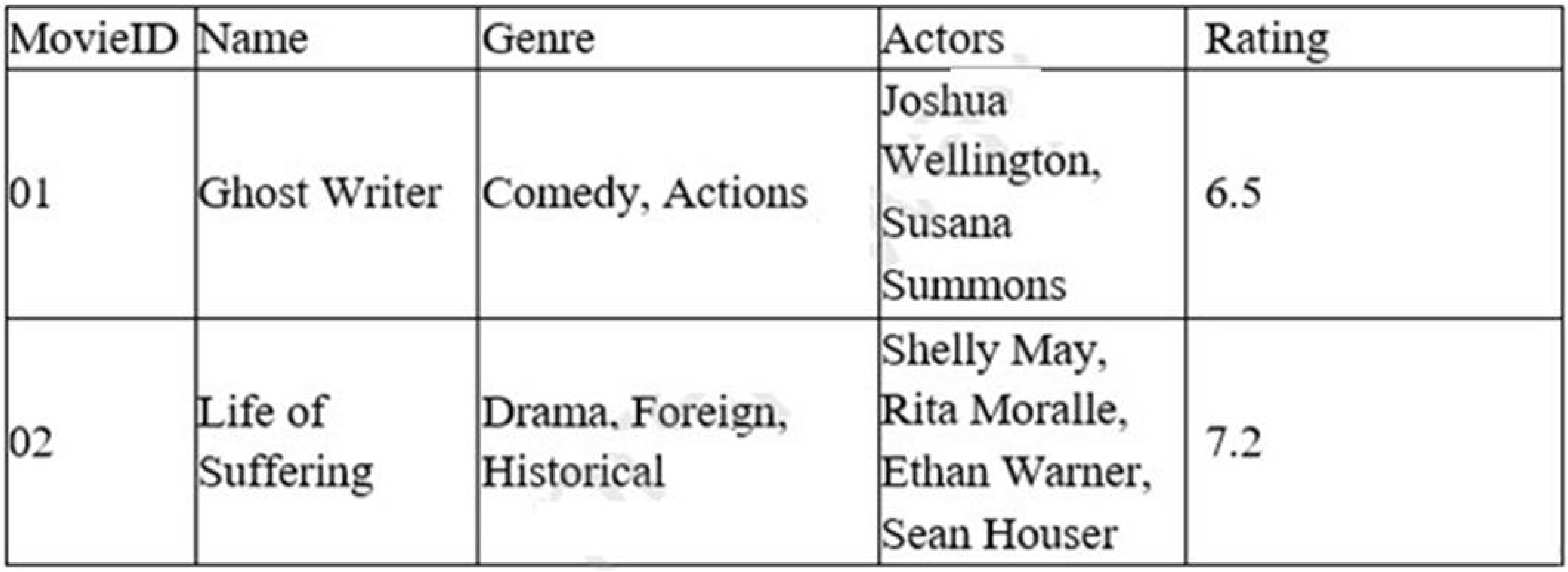
NEW QUESTION 20
A JSON file is an example of:

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

Answer: A

Explanation:
A JSON (JavaScript Object Notation) file is a text-based format for representing structured data based on JavaScript object syntax. It is commonly used for transmitting data in web applications (e.g., sending some data from the server to the client, so it can be displayed on a web page, or vice versa). JSON files are human-readable and can be interpreted by various programming languages, making them ideal for data interchange123.
JSON files typically contain an array of objects, with each object representing a record with a series of name-value pairs. This structured format is both easy to understand and write by humans and easy for machines to parse and generate4.
References:
? JSON??s official definition and syntax rules1.
? A beginner??s guide to JSON and its data types2.
? Understanding the JSON file format3.
? Detailed explanation of JSON as a structured data format4.

NEW QUESTION 22
A data analyst for a media company needs to determine the most popular movie genre. Given the table below:



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

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

A. Append
B. Merge
C. Concatenate
D. Delimit

Answer: D

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

NEW QUESTION 27
An analyst is building a new dashboard for a user. After an initial conversation with the user. the analyst created a mock-up of the dashboard. Which of the following best explains why the analyst created the mock-up?

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

Answer: C

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

NEW QUESTION 32
An analyst collected data that includes primary account numbers, expiration dates, and service codes. Which of the following data governance classifications is used to describe this data?

A. PI I
B. PCI
C. PBI

D. PHI

Answer: B

NEW QUESTION 35
Which of the following is a difference between a primary key and a unique key?

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

Answer: B

Explanation:
The correct answer is B. There can be only one primary key in a data set, whereas there can be multiple unique keys.
A primary key is a column or a set of columns that uniquely identifies each row in a table. A table can have only one primary key, which also enforces the NOT NULL constraint on the column(s) involved. A primary key can also be referenced by a foreign key of another table to establish a relationship between the tables12
A unique key is a column or a set of columns that also uniquely identifies each row in a table, but it is not the primary key. A table can have more than one unique key, which also allows one NULL value for the column(s) involved. A unique key can also be referenced by a foreign key of another table to establish a relationship between the tables12
Some of the differences between a primary key and a unique key are:
? A primary key creates a clustered index on the column(s), whereas a unique key creates a non-clustered index on the column(s)3
? A primary key does not allow any NULL values, whereas a unique key allows one
NULL value for the column(s)123
? A primary key can be a unique key, but a unique key cannot be a primary key12

NEW QUESTION 37
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 38
An analyst is reviewing the following data: Car IDSpeed
123155
566436
564418
650567
546436
645638
Which of the following should the analyst include in the measures of central tendency for speed?

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

Answer: D

Explanation:
The measures of central tendency include the mode, median, and mean. The mode is the value that appears most frequently in a data set. In this case, the speed of 36 appears twice, making it the mode. The median is the middle value when a data set is ordered from least to greatest; for these speeds, when ordered (18, 36, 36, 38, 55, 67), the median is the average of the two middle numbers, which is (\frac{36 + 38}{2} = 37). The mean is the average of all values, calculated as (\frac{55 + 36 + 18 + 67 + 36 + 38}{6} = 41.7). References:
? The calculation of the mode, median, and mean is based on standard statistical
formulas and definitions.
The measures of central tendency for speed include the mode, median, and mean. To calculate these, we first need to organize the data:
? Speeds in ascending order: 18, 36, 36, 38, 55, 67
? Mode is the value that appears most frequently, which is 36, as it appears twice.
? Median is the middle value when the data is ordered. Since we have an even number of observations, we take the average of the two middle values (36 and 38), resulting in 37.
? Mean is the sum of all values divided by the number of values. (18+36+36+38+55+67)/6=41.5(18+36+36+38+55+67)/6=41.5.
Thus, the correct option is D, which includes Mode = 36, Median = 37, and Mean = 41.5. The range, maximum, and minimum values, although useful in understanding data dispersion, are not measures of central tendency and are therefore not relevant to this specific question.

NEW QUESTION 41
When analyzing the values of two variables, you decide to convert both variables so they are on a scale of 0 to 1.
What term describes this action?

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

Answer: B

Explanation:

Normalization is the process of reorganizing data in a database so that it meets two basic requirements: There is no redundancy of data, all data is stored in only one place. Data dependencies are logical, all related data items are stored together.
Put simply, data normalization ensures that your data looks, reads, and can be utilized the same way across all of the records in your customer database. This is done by standardizing the formats of specific fields and records within your customer database.

NEW QUESTION 46
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 51
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 56
Mario works with a group of R programmers tasked with copying data from an accounting system into a data warehouse.
In what phase are the group's R skills most relevant?

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

Answer: C

NEW QUESTION 60
An analyst develops an IT document and needs to describe the technical terms used in the document. Which of the following is where the analyst should include descriptions of the technical terms?

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

Answer: A

Explanation:
In technical documentation, a glossary is the designated section where definitions for technical terms are provided. It serves as a reference point for readers to understand specialized or uncommon words used within the document. Including descriptions of technical terms in a glossary ensures that readers have a consistent resource to refer to, which can improve comprehension and reduce misunderstandings12.
A system diagram (Option B) is a visual representation of the system??s components and their interactions, not a place for defining terms. User requirements (Option C) outline what end-users expect from the system, and an index (Option D) is an alphabetical list of topics covered in the document, usually with page numbers, but not definitions.

References:
? Creating effective technical documentation1.
? Best practices when writing technical descriptions3.

NEW QUESTION 64
Which one of the following programming languages is specifically designed for use in analytics applications?

A. Python.
B. R
C. C++
D. Java.

Answer: B

NEW QUESTION 69
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.

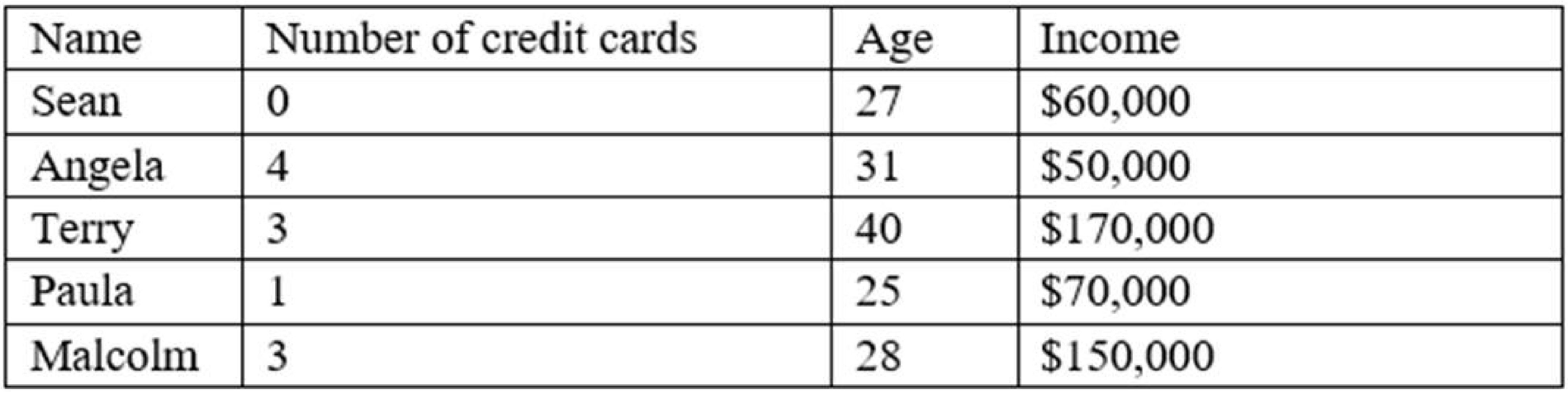
NEW QUESTION 74
Which of the following should be accomplished NEXT after understanding a business requirement for a data analysis report?

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 75
A customer list from a financial services company is shown below:



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

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

A. Recode the variables.
B. Calculate the percentiles of the variables.
C. Calculate the standard deviations of the variables.
D. Normalize the variables.

Answer: D

Explanation:
Normalizing the variables means scaling them to a common range, such as 0 to 1 or -1 to 1, so that they have the same weight in the score calculation. Recoding the variables means changing their values or categories, which would alter their meaning and distribution. Calculating the percentiles of the variables means ranking them relative to each other, which would not account for their actual magnitudes. Calculating the standard deviations of the variables means measuring their variability, which would not make them comparable. References: CompTIA Data+ Certification Exam Objectives, page 10

NEW QUESTION 77
A survey asks participants to rate a company on a scale of one to ten. Which of the following best describes the rating variable?

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

Answer: B

Explanation:
The rating variable in a survey where participants rate a company on a scale of one to ten is best described as ordinal. This is because the ratings are ranked in order, with each number representing a position on a scale of satisfaction or quality. The numbers are not just labels (which would be nominal), nor do they represent a continuous spectrum (which would be continuous). They also do not fit the definition of categorical, as that implies non- ordered groups or categories. In an ordinal scale, the order of the values is significant and meaningful12.
References:
? Qualtrics explains that ordinal scales have answer sets that occur in a logical and systematic order, providing qualitative data.
? Zonka Feedback describes a 1 to 10 rating scale survey, indicating that the numbers represent a ranking from most negative to most positive experience, which aligns with the characteristics of an ordinal scale.

NEW QUESTION 78
Which of the following is an example of a data-mining ETL tool?

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

Answer: A

Explanation:
A data-mining ETL tool is a software application that performs extract, transform, and load (ETL) operations on data for data mining purposes. Data mining is the process of discovering patterns, trends, and insights from large and complex data sets. ETL tools help to prepare the data for analysis by extracting data from various sources, transforming data into a consistent and suitable format, and loading data into a data warehouse or other destination. SSIS (SQL Server Integration Services) is an example of a data-mining ETL tool that is part of Microsoft SQL Server. SSIS provides graphical tools and wizards for building and debugging ETL packages that can work with various data
sources and destinations. Therefore, the correct answer is A. References: [Data Mining - SQL Server Integration Services (SSIS) | Microsoft Docs], [What Is Data Mining? | Oracle]

NEW QUESTION 79
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.

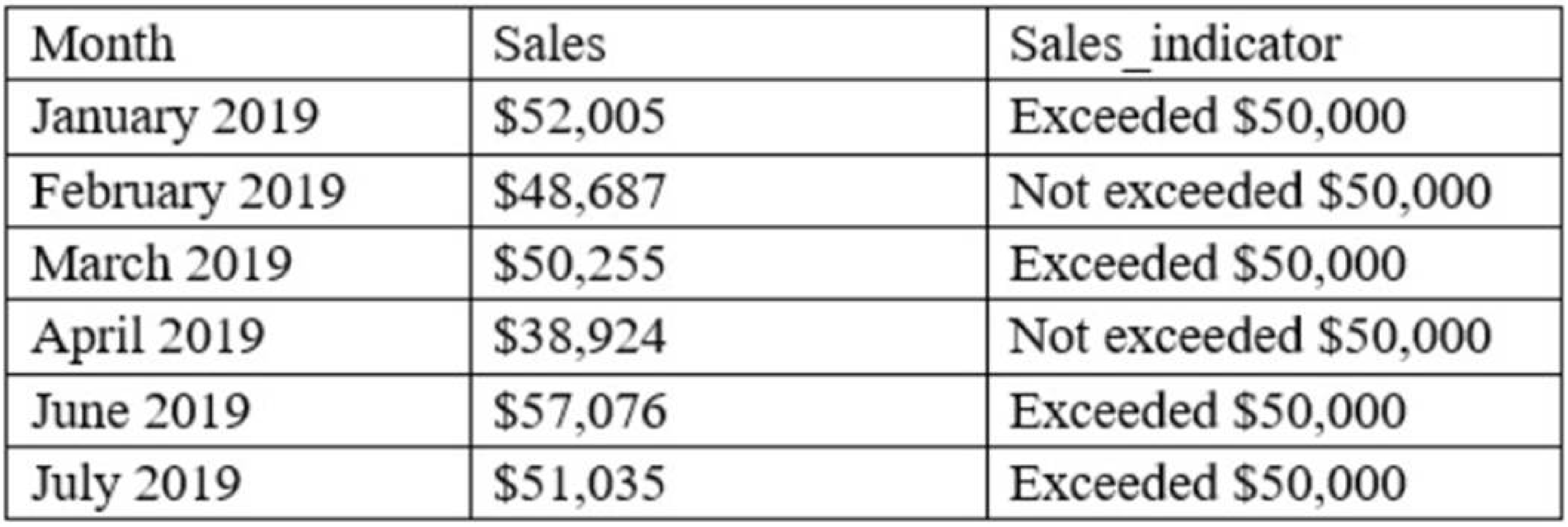
NEW QUESTION 82
??Which of the following is the BEST reason to use database views instead of tables?

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 83
An analyst has generated a report that includes the number of months in the first two quarters of 2019 when sales exceeded $50,000:



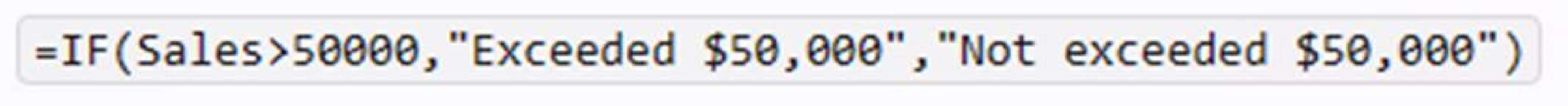
Month	Sales	Sales_indicator
January 2019	$52,005	Exceeded $50,000
February 2019	$48,687	Not exceeded $50,000
March 2019	$50,255	Exceeded $50,000
April 2019	$38,924	Not exceeded $50,000
June 2019	$57,076	Exceeded $50,000
July 2019	$51,035	Exceeded $50,000

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

A. Aggregate
B. Logical
C. Date
D. Sort

Answer: B

Explanation:
This is because a logical function is a type of function that returns a value based on a condition or a set of conditions. A logical function can be used to generate the data in the Sales_indicator column by comparing the values in the Sales column with a threshold of $50,000 and returning either ??Exceeded $50,000?? or ??Not exceeded $50,000?? accordingly. For example, a logical function in Excel that can achieve this is:



```
=IF(Sales>50000,"Exceeded $50,000","Not exceeded $50,000")
```

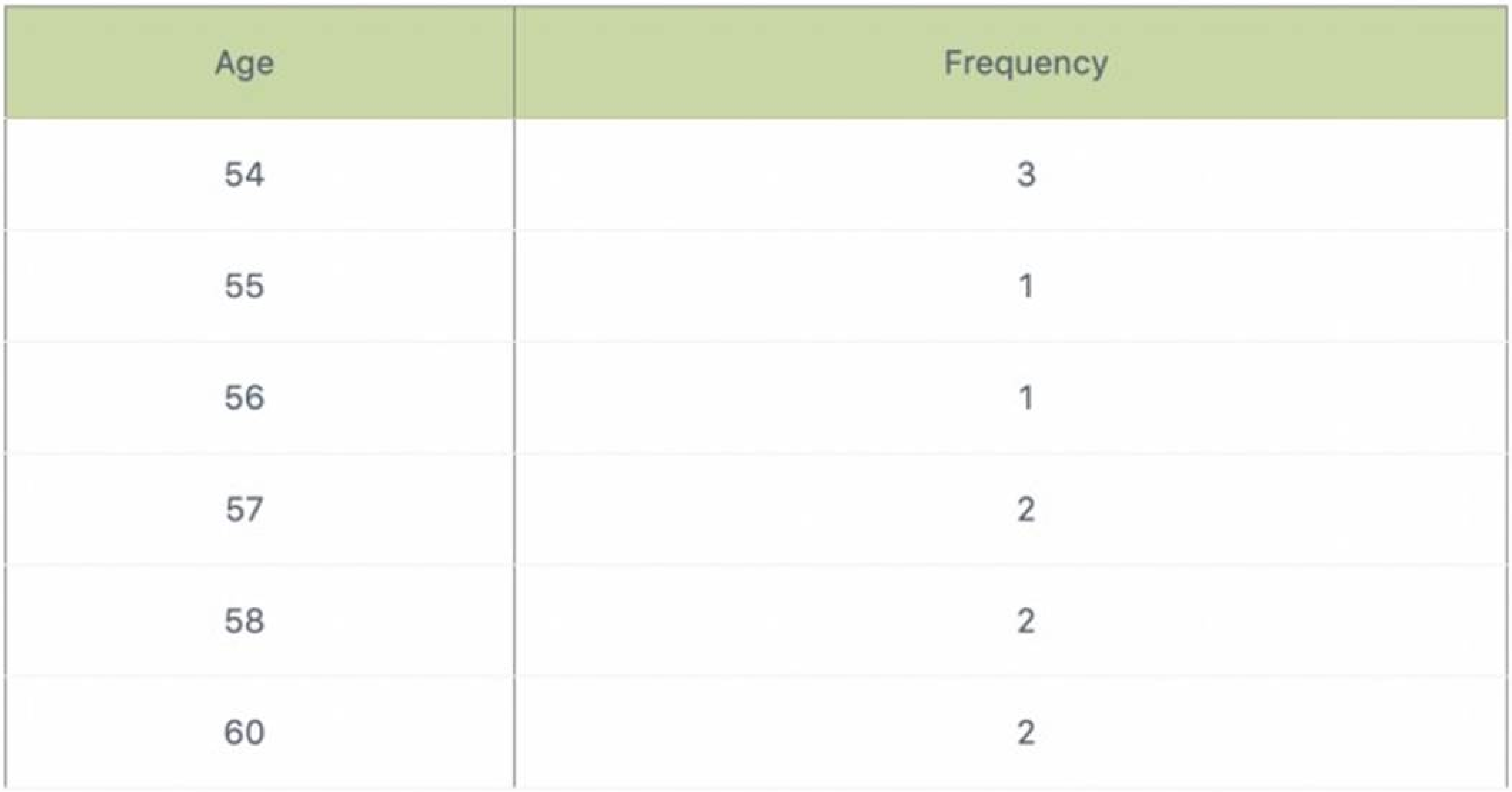
The other functions are not suitable for generating the data in the Sales_indicator column. Here is why:
Aggregate is a type of function that performs a calculation on a group of values, such as sum, average, count, etc. An aggregate function cannot generate the data in the Sales_indicator column because it does not compare the values in the Sales column with a threshold or return a text value based on a condition.
Date is a type of function that manipulates or extracts information from dates, such as year, month, day, etc. A date function cannot generate the data in the Sales_indicator column because it does not use the values in the Sales column or return a text value based on a condition.
Sort is a type of function that arranges the values in a column or a range in ascending or descending order. A sort function cannot generate the data in the Sales_indicator column because it does not create a new column or return a text value based on a condition.

NEW QUESTION 85
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 86
Consider this dataset showing the retirement age of 11 people, in whole years: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60
This tables show a simple frequency distribution of the retirement age data.



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

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

Answer: D

Explanation:

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

NEW QUESTION 90
Which of the following data manipulation techniques should an analyst use to hide unnecessary data during analysis?

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

Answer: A

NEW QUESTION 91
A report is scheduled to run and be distributed at the end of business each day. On Mondays, one of the recipients opens the previous week's reports and combines them to calculate the weekly totals and projections for the coming week. This is a tedious process, and the recipient asks an analyst for help. Which of the following should the analyst recommend?

A. Add calculation fields to the daily report so the totals are built in.
B. Create a new report with weekly totals set to run at the end of business on Friday.
C. Provide a daily summary to the report with totals to save the user the effort of manual calculations.
D. Reduce the frequency of the report to once a week and change the date range.

Answer: B

Explanation:
Creating a new report that automatically calculates weekly totals would streamline the process for the recipient. By setting this report to run at the end of business on Friday, it would provide the recipient with the necessary information for the entire week in one consolidated document. This eliminates the need for manual calculations and combines the previous week??s data into one report, making it more efficient and less time- consuming.
References:
? Best practices in business analytics suggest automating repetitive tasks and consolidating reports where possible to improve efficiency and reduce the potential for human error.

NEW QUESTION 96
A data analyst is creating a dashboard and trying to identify the type of information that should be included. Which of the following should the analyst consider first?

A. Data refresh rate
B. Consumer types
C. Access permissions
D. Data sources and attributes

Answer: D

Explanation:
The answer is D. Data sources and attributes.
Short Explanation: The data analyst should consider the data sources and attributes first when creating a dashboard, because they determine what kind of information can be
included and how it can be displayed. The data sources and attributes define the origin, quality, format, and structure of the data that will be used for the dashboard. They also affect the data refresh rate, the consumer types, and the access permissions of the dashboard12
* A. Data refresh rate is not the first thing to consider, because it depends on the data sources and attributes. The data refresh rate is how often the data in the dashboard is updated or refreshed to reflect the latest changes. The data refresh rate can vary depending on the type, frequency, and availability of the data sources1
* B. Consumer types are not the first thing to consider, because they depend on the data sources and attributes. The consumer types are the intended audiences or users of the dashboard, who may have different needs, preferences, and expectations for the dashboard. The consumer types can influence the design, layout, and functionality of the dashboard. However, the consumer types cannot be determined without knowing what kind of data is available and relevant for them1
* C. Access permissions are not the first thing to consider, because they depend on the data sources and attributes. The access permissions are the rules or policies that govern who can view, edit, or share the dashboard. The access permissions can protect the confidentiality, integrity, and availability of the data in the dashboard. However, the access permissions cannot be set without knowing what kind of data is involved and who needs to access it1

NEW QUESTION 98
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 99
Which of the following are reasons to conduct data cleansing? (Select two).

A. To perform web scraping
B. To track KPIs
C. To improve accuracy
D. To review data sets
E. To increase the sample size
F. To calculate trends

Answer: CF

Explanation:
Two reasons to conduct data cleansing are:
? To improve accuracy: Data cleansing helps to ensure that the data is correct, consistent, and reliable. This can improve the quality and validity of the analysis, as well as the decision-making and outcomes based on the data12
? To calculate trends: Data cleansing helps to remove or resolve any errors, outliers,
or missing values that could distort or skew the data. This can help to identify and measure the patterns, changes, or relationships in the data over time13

NEW QUESTION 101
Which of the following value is the measure of dispersion "range" between the scores of ten students in a test.
The scores of ten students in a test are 17, 23, 30, 36, 45, 51, 58, 66, 72, 77.

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

Answer: B

Explanation:

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

NEW QUESTION 106
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 107
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.

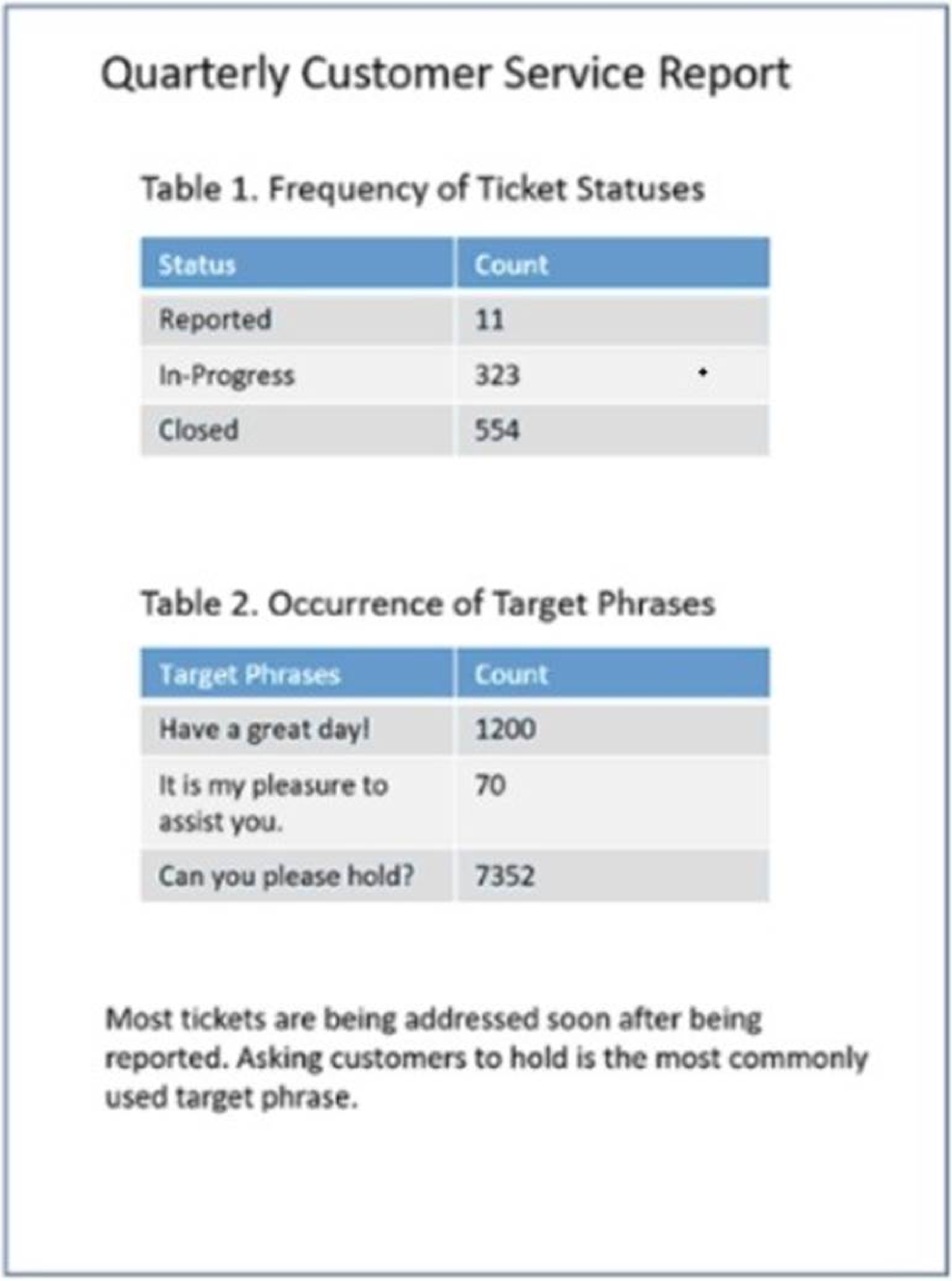
NEW QUESTION 108
Standardized tests are given to students in the middle of each month, and the results are ready by the end of the month. The superintendent needs a quick view of test performance. Which of the following would be the best recommendation to meet the superintendent's requirements?

A. A dashboard with a continuous data stream and saved searches
B. A report of test scores by classroom, emailed to the superintendent at the end of the month
C. A report of test scores with pie charts showing student performance
D. A dashboard with a scheduled delivery, the ability to filter scores by school, and bar charts for comparison

Answer: D

Explanation:
A dashboard with a scheduled delivery is an efficient way to provide a quick view of test performance. It allows for timely updates, which is crucial given that the superintendent needs the information promptly at the end of each month. The ability to filter scores by school enables the superintendent to easily segment and analyze the data as needed. Bar charts are effective for comparison and can visually communicate the performance across different schools or other categories, making it easier to identify trends and outliers at a glance.
References:
? Best practices in data visualization recommend using dashboards for real-time data monitoring and quick access to key metrics1.
? Guidelines for presenting performance data suggest that visual tools like bar charts are helpful in comparing and analyzing data effectively1.
? Educational performance data analysis often involves comparing scores across different schools or classrooms, which is facilitated by a well-designed dashboard2.

NEW QUESTION 111
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 115
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 117
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 121
You would like to measure how well an organization is achieving its goals. What type of analysis should you perform?

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

Answer: A

Explanation:

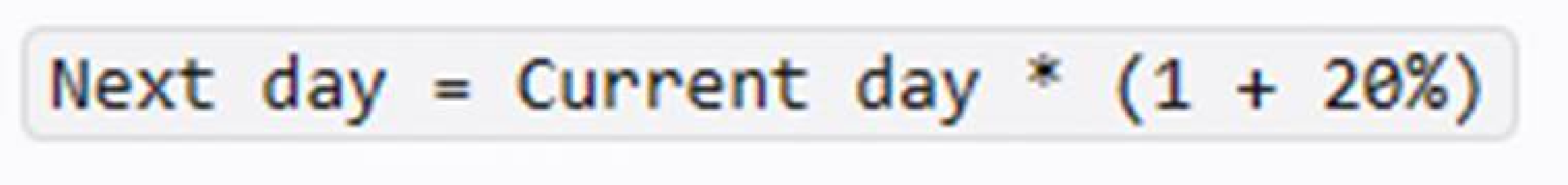
Performance analysis is the technique of studying or comparing the performance of a specific situation in contrast to the aim and yet executed. In Human Resources, performance analysis can help to review an employee's contribution towards a project or assignment, which they allotted him or her.

NEW QUESTION 124
An analyst runs a report on a daily basis, and the number of datapoints must be validated before the data can be analyzed. The number of datapoints increases each day by approximately 20% of the total number from the day before. On a given day, the number of datapoints was 8,798. Which of the following should be the total number of datapoints on the next day?

A. 7,038
B. 9,600
C. 10,600
D. 10,800

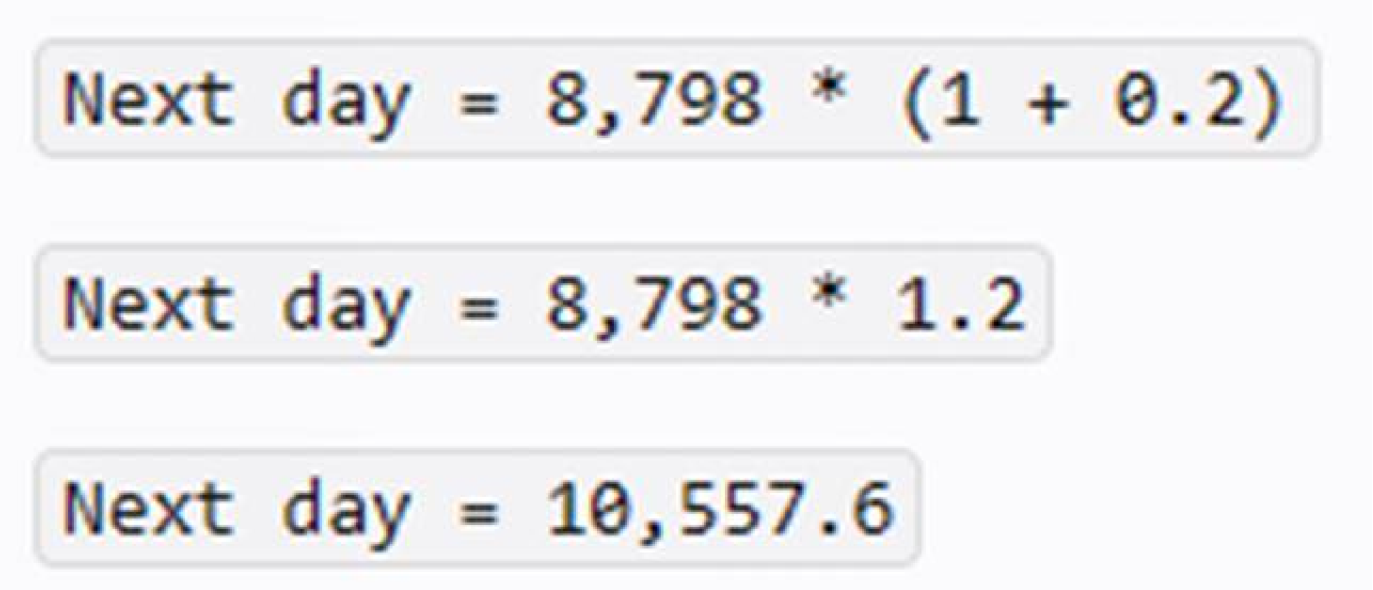
Answer: C

Explanation:
This is because the number of datapoints increases each day by approximately 20% of the total number from the day before. Therefore, to find the number of datapoints on the next day, we can use the formula:



```
Next day = Current day * (1 + 20%)
```

Plugging in the given values, we get:



```
Next day = 8,798 * (1 + 0.2)

Next day = 8,798 * 1.2

Next day = 10,557.6
```

Since we are dealing with whole numbers, we can round up the result to the nearest integer, which is 10,600.

NEW QUESTION 127
The total values in this month's revenue report are twice as much as last month's. Which of the following most likely occurred during the ETL process?

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

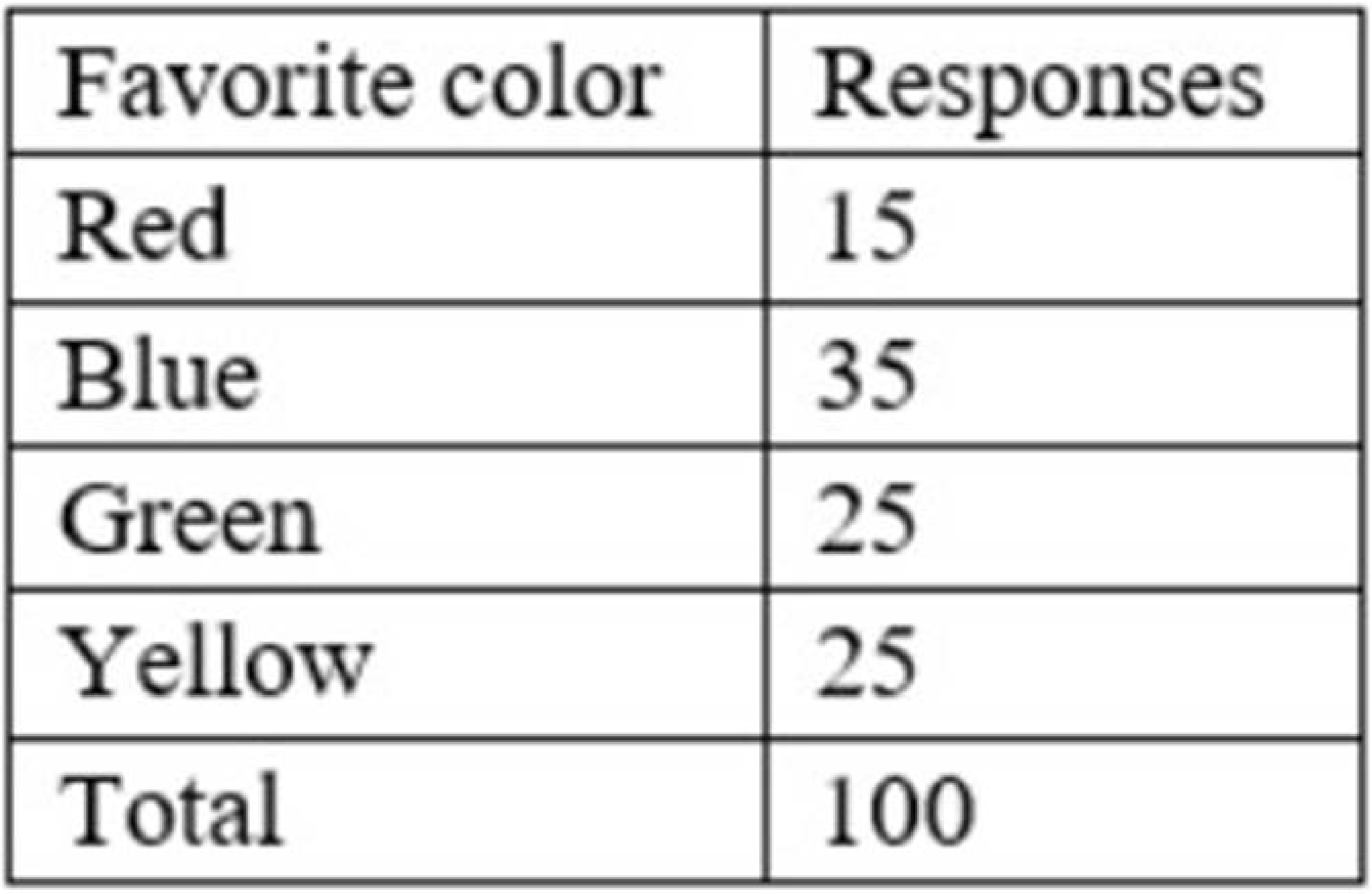
Answer: C

NEW QUESTION 131
Which of the following types of analysis is used when comparing last week's sales to the previous week's sales?

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

Answer: A

NEW QUESTION 134
An analyst needs to provide a chart to identify the composition between the categories of the survey response data set:



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

Which of the following charts would be BEST to use?

A. Histogram
B. Pie
C. Line

D. Scatter pot
E. Waterfall

Answer: B

Explanation:
A pie chart is the best choice to show the composition between the categories of the survey response data set. A pie chart represents the whole with a circle, divided by slices into parts. Each slice shows the relative size of each category as a percentage of the total. A pie chart is useful when the categories are mutually exclusive and add up to 100%. The table shows the favorite color and the number of responses for each color, which can be easily converted into percentages. A pie chart can show how each color contributes to the total number of responses.
Option A is incorrect because a histogram is used to show how data points are distributed along a numerical scale. The survey response data set is not numerical, but categorical. Option C is incorrect because a line chart is used to show trends or changes over time. The survey response data set does not have a time dimension.
Option D is incorrect because a scatter plot is used to show the relationship between two numerical variables. The survey response data set does not have two numerical variables. Option E is incorrect because a waterfall chart is used to show how an initial value is increased or decreased by a series of intermediate values. The survey response data set does not have an initial value or intermediate values.
References:
? How to Choose the Right Chart for Your Data - Infogram
? How to Choose the Right Data Visualization | Tutorial by Chartio
? Find the Best Visualizations for Your Metrics - The Data School
? How to choose the best chart or graph for your data

NEW QUESTION 139
A company wants to know how its customers interact with an e-commerce website based on clicks over items. Which of the following is the primary requirement for this report?

A. Data content
B. Frequency
C. Filtering
D. Views

Answer: B

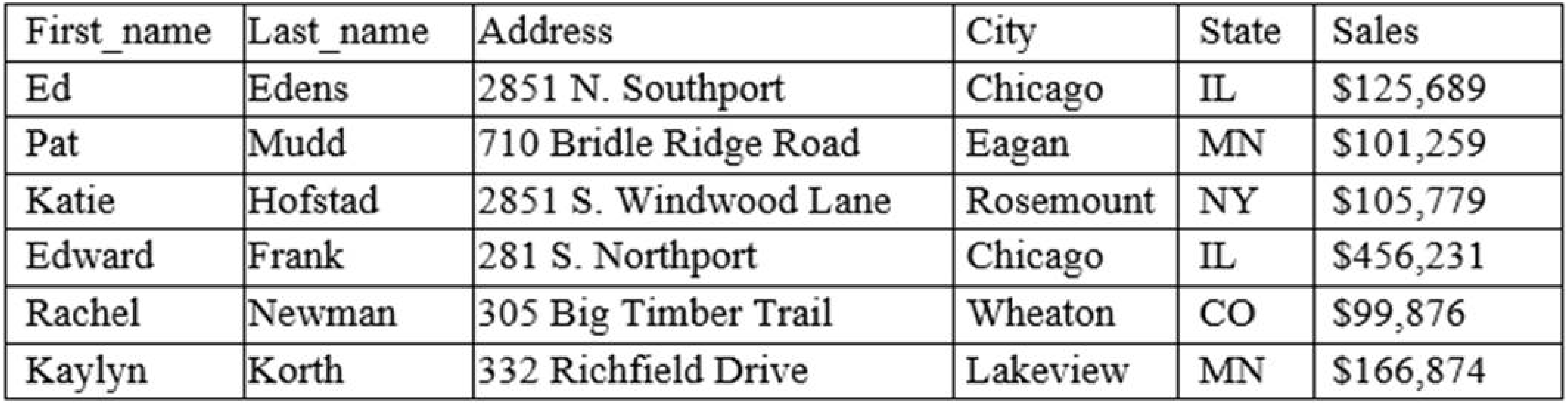
NEW QUESTION 144
Each month an analyst needs to execute a data pull for the two prior months. Which of the following is the most efficient function for the analyst to use?

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

Answer: B

Explanation:
The most efficient function for an analyst to execute a data pull for the two prior months would be the Date function. This function allows for the manipulation and formatting of date values within a database. Using Date functions, an analyst can dynamically calculate the start and end dates for the previous two months, ensuring that the data pull is accurate and automated without manual intervention.
For example, SQL functions like DATEADD and DATEDIFF can be used to determine the exact range of dates needed for the data pull. These functions can calculate the first and
last day of the previous months relative to the current date, which is essential for monthly reporting and analysis.
References:
? Discussions on Stack Overflow suggest using SQL date functions
like DATEADD and DATEDIFF to dynamically extract data for previous months, which supports the use of Date functions12.
? The use of Date functions is also recommended for ensuring that the data pull is
not only efficient but also accurate, as it avoids potential errors associated with manual date entry3.

NEW QUESTION 145
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

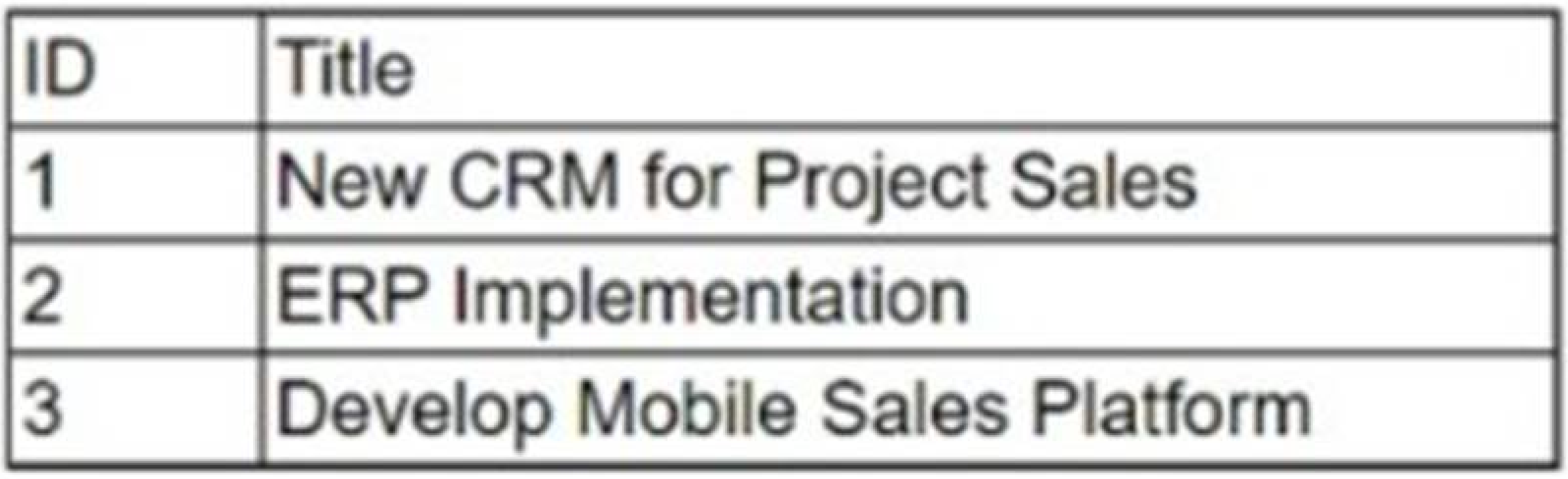
Explanation:
Sorting is the function that will allow the data analyst to organize the table in the desired manner. Sorting means arranging the data in a specific order, such as ascending or descending, based on one or more criteria. Sorting can be applied to any column in the table, such as sales or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

NEW QUESTION 149
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?

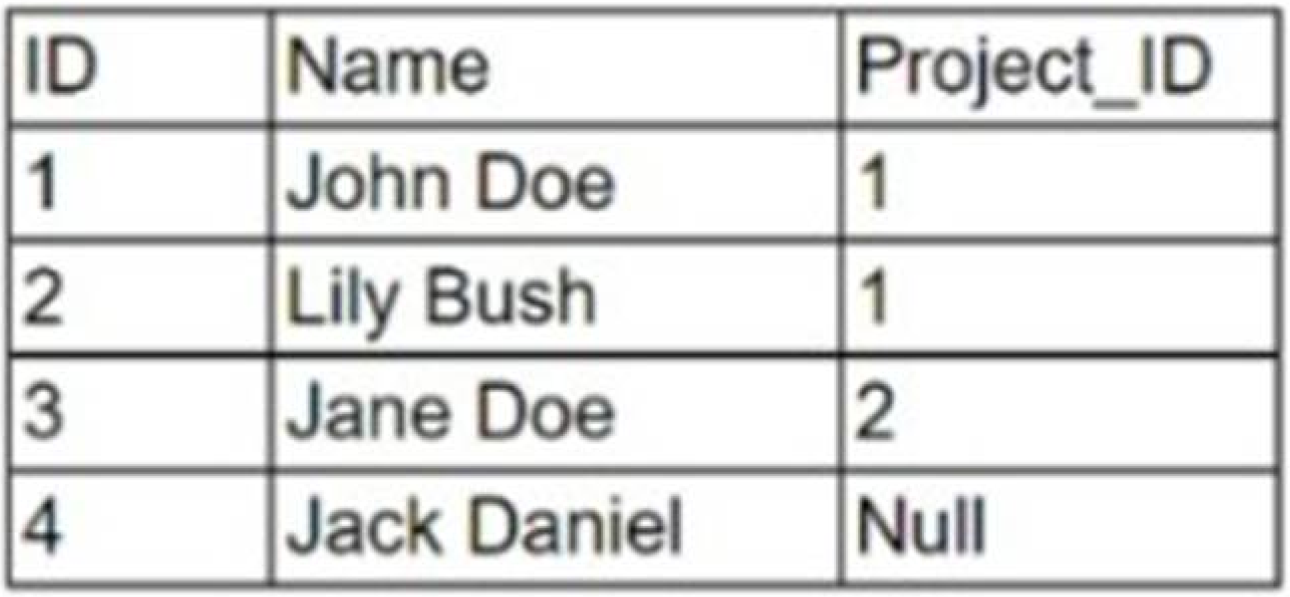
A. ETL
B. API
C. SQL
D. ELT

Answer: A

NEW QUESTION 150
Given the following tables:



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



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

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

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

Answer: D

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

NEW QUESTION 155
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 156
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?

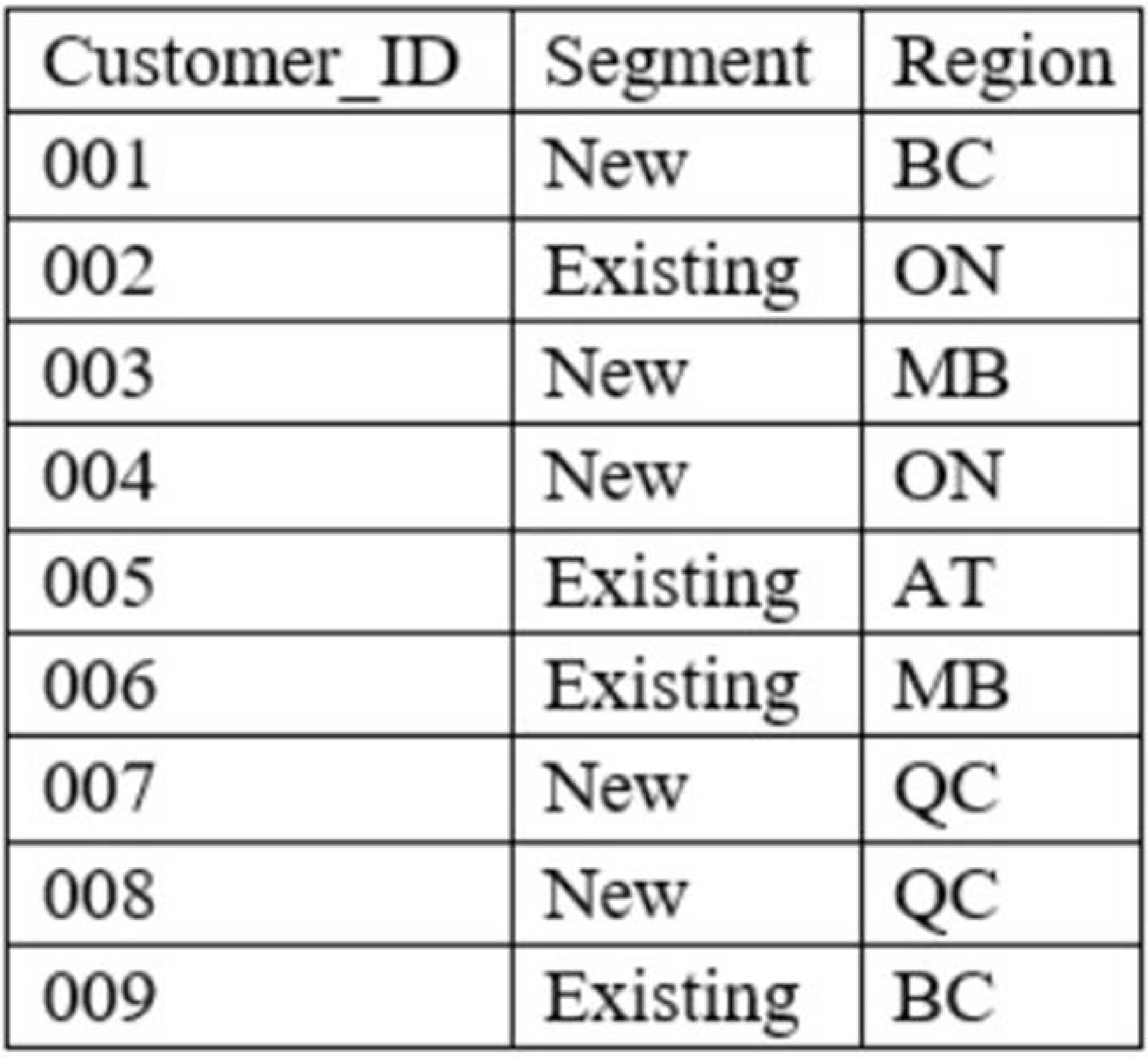
A. People who receive electronic coupons spend more on average.
B. People who receive electronic coupons spend less on average.
C. People who receive electronic coupons do not spend more on average.
D. People who do not receive electronic coupons spend more on average.

Answer: C

Explanation:

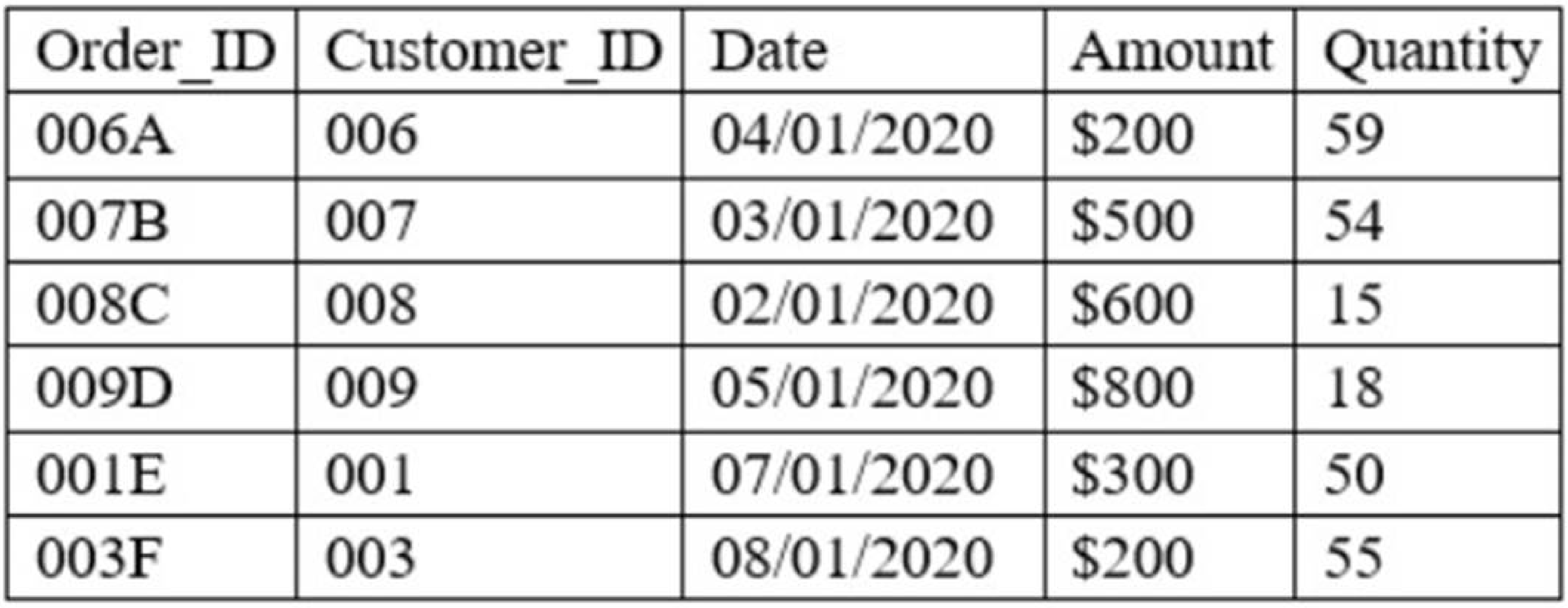
The null hypothesis presumes the status quo. Andy is testing whether or not people who receive an electronic coupon spend more on average, so, the null hypothesis states that people who receive the coupon do spend more on average.

NEW QUESTION 157
A data analyst has been asked to merge the tables below, first performing an INNER JOIN and then a LEFT JOIN:



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

Customer Table -
In-store Transactions –



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

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

A. INNER: 6 rows; LEFT: 9 rows
B. INNER: 9 rows; LEFT: 6 rows
C. INNER: 9 rows; LEFT: 15 rows
D. INNER: 15 rows; LEFT: 9 rows

Answer: C

Explanation:
An INNER JOIN returns only the rows that match the join condition in both tables. A LEFT JOIN returns all the rows from the left table, and the matched rows from the right table, or NULL if there is no match. In this case, the customer table is the left table and the in-store transactions table is the right table. The join condition is based on the customer_id column, which is common in both tables.
To perform an INNER JOIN, we can use the following SQL query:
SELECT * FROM customer INNER JOIN in_store_transactions ON customer.customer_id
= in_store_transactions.customer_id;
This query will return 9 rows of data, as shown below:
customer_id | name | lastname | gender | marital_status | transaction_id | amount | date 1 | MARC | TESCO | M | Y | 1 | 1000 | 2020-01-01 1 | MARC | TESCO | M | Y | 2 | 5000 | 2020-01-02 2 | ANNA | MARTIN | F | N | 3 | 2000 | 2020-01-03 2 | ANNA | MARTIN | F | N |
4 | 3000 | 2020-01-04 3 | EMMA | JOHNSON | F | Y | 5 | 4000 | 2020-01-05 4 | DARIO | PENTAL | M | N | 6 | 5000 | 2020-01-06 5 | ELENA | SIMSON| F|
N|7|6000|2020-01-07 6|TIM|ROBITH|M|N|8|7000|2020-01-08 7|MILA|MORRIS|F|N|9|8000|2020-01-09
To perform a LEFT JOIN, we can use the following SQL query:
SELECT * FROM customer LEFT JOIN in_store_transactions ON customer.customer_id = in_store_transactions.customer_id;
This query will return 15 rows of data, as shown below: customer_id|name|lastname|gender|marital_status|transaction_id|amount|date
1|MARC|TESCO|M|Y|1|1000|2020-01-01 1|MARC|TESCO|M|Y|2|5000|2020-01-02
2|ANNA|MARTIN|F|N|3|2000|2020-01-03 2|ANNA|MARTIN|F|N|4|3000|2020-01-04
3|EMMA|JOHNSON|F|Y|5|4000|2020-01-05 4|DARIO|PENTAL|M|N|6|5000|2020-01-06
5|ELENA|SIMSON||F||N||7||6000||2020-01-07 6||TIM||ROBITH||M||N||8||7000||2020-01-08
7||MILA||MORRIS||F||N||9||8000||2020-01-09
8||JENNY||DWARTH||F||Y||NULL||NULL||NULL
As you can see, the customers who do not have any transactions (customer_id = 8) are still included in the result, but with NULL values for the transaction_id, amount, and date columns.
Therefore, the correct answer is C: INNER: 9 rows; LEFT: 15 rows. Reference: SQL Joins - W3Schools

NEW QUESTION 161
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 163
Which of the following is the correct data type for text?

A. Boolean
B. String

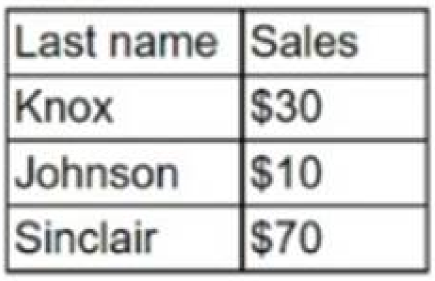
C. Integer
D. Float

Answer: B

Explanation:
The correct data type for text is string. A string is a data type that represents a sequence of characters, such as letters, numbers, symbols, or spaces. A string can be enclosed by single quotes (?? ') or double quotes (" ") in most programming languages. For example, ??Hello??, ??World??, and ??123?? are all strings. The other options are not data types for text, but for other kinds of values. A boolean is a data type that represents a logical value, either true or false. An integer is a data type that represents a whole number, such as 1, 0, or -5. A float is a data type that represents a number with a fractional part, such as 3.14, 0.5, or -2.7.
Reference: Data Types - W3Schools

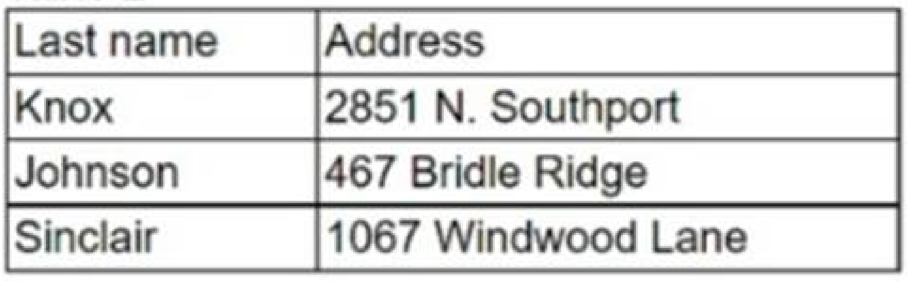
NEW QUESTION 167
A data analyst is performing a data merge within a spreadsheet using the tables below:
https://www.bing.com/images/blob?bcid=S1XCF9p02M4GjpbGxHj0lrIaj9sw.....4c

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 172
A research analyst wants to determine whether the data being analyzed is connected to other datapoints. Which of the following is the BEST type of analysis to conduct?

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

Answer: C

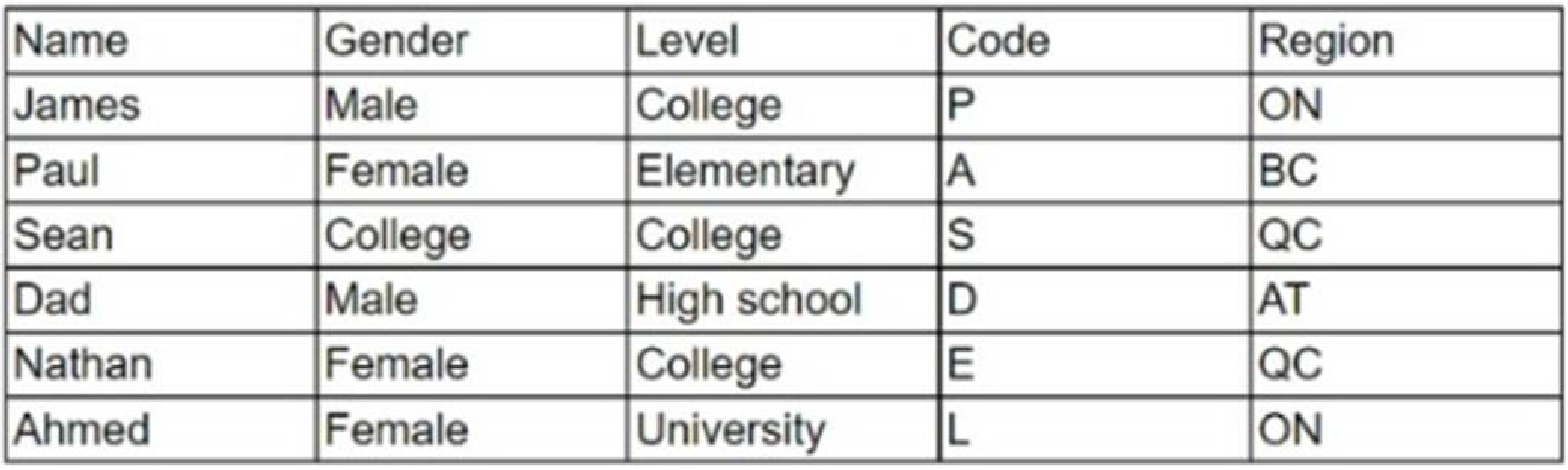
Explanation:
This is because link analysis is a type of analysis that determines whether the data being analyzed is connected to other datapoints, such as entities, events, or relationships. Link analysis can be used to identify and visualize the patterns, networks, or associations among the datapoints, as well as measure the strength, direction, or frequency of the connections. For example, link analysis can be used to determine if there is a connection between a customer??s purchase history and their loyalty program status. The other types of analysis are not the best types of analysis to conduct to determine whether the data being analyzed is connected to other datapoints. Here is why:
? Trend analysis is a type of analysis that determines whether the data being analyzed is changing over time, such as increasing, decreasing, or fluctuating. Trend analysis can be used to identify and visualize the patterns, cycles, or movements in the data points, as well as measure the rate, direction, or magnitude of the changes. For example, trend analysis can be used to determine if there is a change in a company??s sales revenue over a period of time.
? Performance analysis is a type of analysis that determines whether the data being
analyzed is meeting certain goals or objectives, such as targets, benchmarks, or standards. Performance analysis can be used to identify and visualize the gaps, deviations, or variations in the data points, as well as measure the efficiency, effectiveness, or quality of the outcomes. For example, performance analysis can be used to determine if there is a gap between a student??s test score and their expected score based on their previous performance.
? Exploratory analysis is a type of analysis that determines whether there are any insights or discoveries in the data being analyzed, such as patterns, relationships, or anomalies. Exploratory analysis can be used to identify and visualize the characteristics, features, or behaviors of the data points, as well as measure their distribution, frequency, or correlation. For example, exploratory analysis can be used to determine if there are any outliers or unusual values in a dataset.

NEW QUESTION 174
An analyst is working with a data set that lists individuals' first and last names in separate columns. Which of the following processes should the analyst use to combine the first and last names into a single spreadsheet cell?

A. Transpose
B. Blend
C. Concatenate
D. Merges

Answer: C

NEW QUESTION 175
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 179
A database consists of one fact table that is composed of multiple dimensions. Depending on the dimension, each one can be represented by a denormalized table or multiple normalized tables. This structure is an example of a:

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

Answer: B

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

NEW QUESTION 182
A data analyst has received a data set that contains actual and projected sales for the fourth quarter of 2019. Which of the following statistical methods should the analyst use to find the measure of dispersion?

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

Answer: B

Explanation:
The measure of dispersion is used to describe the spread of data around a central value. In the context of a data set containing actual and projected sales, the measure of dispersion will help to understand the variability or consistency of sales figures. The variance is the most appropriate statistical method for finding the measure of dispersion because it calculates the average of the squared differences from the Mean, providing a clear picture of data spread. It is especially useful in comparing the spread between different data sets and understanding the distribution of data points.
? Mean is a measure of central tendency, not dispersion.
? Correlation measures the relationship between two variables, not the spread of a single variable.
? Confidence intervals are used to estimate the range within which a population parameter will fall, but they do not measure dispersion within the data set itself.
References:
? Measures of Dispersion in Statistics1
? Measures of Dispersion - Definition, Formulas, Examples2
? Statistical dispersion - Wikipedia3

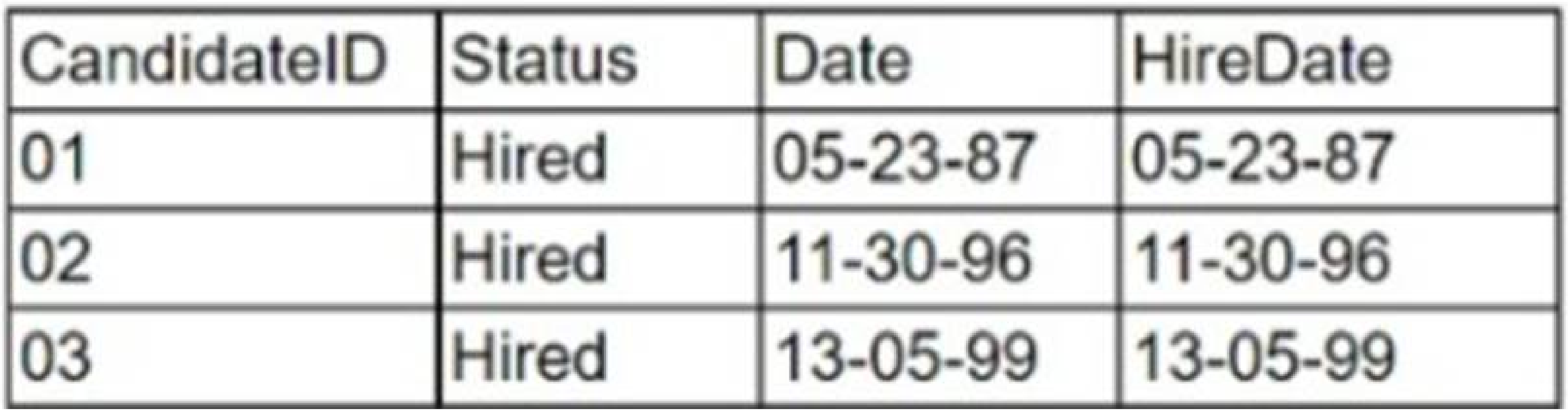
NEW QUESTION 184
Which of the following are reasons to create and maintain a data dictionary? (Choose two.)

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 189
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 194
A data analyst is attempting to understand how ice cream consumption is affected by different attributes. such as cost, temperature. and income level. Which of the following regression analyses should the data analyst perform to understand this relationship?

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

Answer: B

Explanation:
Answer: B. Ordinary least squares
Ordinary least squares (OLS) is a type of linear regression that is used to fit a regression model that describes the relationship between one or more predictor variables and a numeric response variable. Use when: The relationship between the predictor variable(s) and the response variable is reasonably linear. The response variable is a continuous numeric variable1.
In this case, the data analyst is interested in understanding how ice cream consumption (the response variable) is affected by different attributes, such as cost, temperature, and income level (the predictor variables). Assuming that these variables have a linear relationship, OLS can be used to estimate the coefficients of the regression equation that best fits the data. OLS can also provide measures of goodness-of-fit, such as R-squared and adjusted R-squared, and test the significance of the coefficients using t-tests and F- tests2.
Option A is incorrect, as logistic regression is used to fit a regression model that describes the relationship between one or more predictor variables and a binary response variable. Use when: The response variable is binary – it can only take on two values1. Ice cream consumption is not a binary variable, but rather a continuous numeric variable.
Option C is incorrect, as Cox regression is used to fit a regression model that describes the relationship between one or more predictor variables and a survival time response variable. Use when: The response variable is the time until an event of interest occurs, such as death, failure, or recovery3. Ice cream consumption is not a survival time variable, but rather a continuous numeric variable.
Option D is incorrect, as polynomial regression is used to fit a regression model that describes the relationship between one or more predictor variables and a numeric response variable. Use when: The relationship between the predictor variable(s) and the response variable is non-linear1. If there is no evidence of non-linearity in the data, polynomial regression may not be appropriate, as it may overfit the data and produce unreliable estimates.

NEW QUESTION 195
Which of the following is an object associated with a table that sorts and stores table row data in a key-value pair?

A. Foreign key
B. Function
C. Stored procedure
D. Clustered index

Answer: D

NEW QUESTION 198
Which of the following best describes a business analytics tool with interactive visualization and business capabilities and an interface that is simple enough for end users to create their own reports and dashboards?

A. Python
B. R
C. Microsoft Power Bl
D. SAS

Answer: C

Explanation:
The best answer is C. Microsoft Power BI.
Microsoft Power BI is a business analytics and business intelligence service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards. Power BI can connect to multiple data sources, clean and transform data, create custom calculations, and visualize data through charts, graphs, and tables. Power BI can be accessed through a web browser, mobile device, or desktop application and integrated with other Microsoft tools like Excel and SharePoint12
Python is not correct, because Python is a general-purpose programming language that can be used for various applications, including data analysis and visualization. However, Python is not a dedicated business analytics tool, and it requires coding or programming skills to create reports and dashboards.
R is not correct, because R is a programming language and software environment for statistical computing and graphics. R can be used for data analysis and visualization, but it is not a specialized business analytics tool, and it requires coding or programming skills to create reports and dashboards.
SAS is not correct, because SAS is a software suite for advanced analytics, business intelligence, data management, and predictive analytics. SAS can provide interactive visualizations and business capabilities, but it does not have an interface that is simple enough for end users to create their own reports and dashboards. SAS also requires coding or programming skills to use its features.

NEW QUESTION 202
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 204
Which one of the following is a measure of dispersion?

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

Answer: A

NEW QUESTION 205
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.

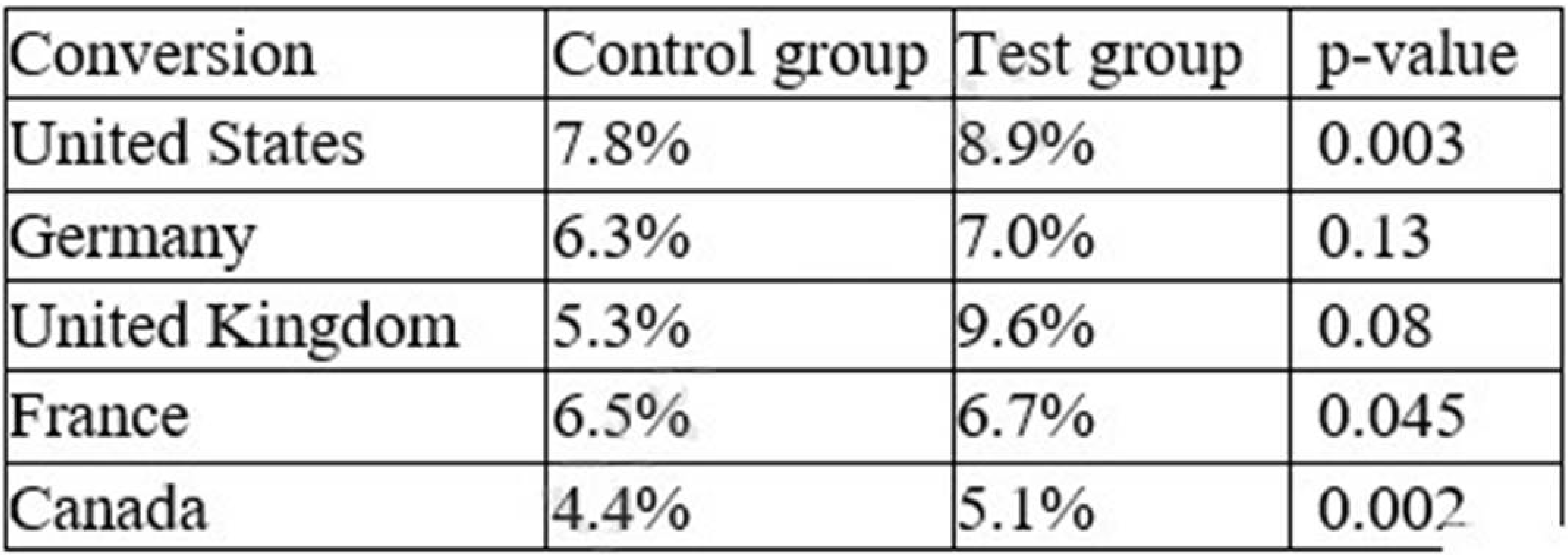
NEW QUESTION 210
Which of the following is a control measure for preventing a data breach?

A. Data transmission
B. Data attribution
C. Data retention
D. Data encryption

Answer: D

Explanation:
This is because data encryption is a type of control measure that prevents a data breach, which is an unauthorized or illegal access or use of data by an external or internal party. Data encryption can prevent a data breach by protecting and securing the data using a code or a key that scrambles or transforms the data into an unreadable or incomprehensible format, which can only be decoded or restored by authorized users who have the correct code or key. For example, data encryption can prevent a data breach by encrypting the data in transit or at rest, such as when the data is sent over a network or stored in a device. The other control measures are not used for preventing a data breach. Here is why:
? Data transmission is a type of process that transfers and exchanges data between different sources or systems, such as databases, cloud services, or web applications. Data transmission does not prevent a data breach, but rather exposes the data to potential risks or threats during the transfer or exchange. However, data transmission can be made more secure and less vulnerable to a data breach by using encryption or other methods, such as authentication or authorization.
? Data attribution is a type of feature or function that assigns and tracks the
ownership and origin of the data, such as the creator, modifier, or source of the data. Data attribution does not prevent a data breach but rather provides information and evidence about the data provenance and history. However, data attribution can be useful for detecting and responding to a data breach by using audit logs or metadata to identify and trace any unauthorized or illegal access or use of the data.
? Data retention is a type of policy or standard that specifies and regulates the
storage and preservation of the data, such as the duration, location, or format of the data. Data retention does not prevent a data breach, but rather affects the availability and accessibility of the data for future use or reference. However, data retention can be optimized and aligned with the legal and ethical requirements and standards of the industry or the organization to reduce the risk or impact of a data breach.

NEW QUESTION 214
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 215
A military commander would like to see the health scorecards of the troops daily and filter them based on gender and rank. Considering this data is PHI, which of the following would be the best way for the commander to view the information?

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

Answer: B

Explanation:
A password-protected dashboard is a type of web-based application that can display the health scorecards of the troops in a secure and interactive way. A password-protected dashboard can provide the following benefits for the commander:
? It can protect the PHI data from unauthorized access or disclosure by requiring a
valid username and password to log in. This can ensure that only the commander and other authorized personnel can view the information12
? It can allow the commander to filter the data based on gender and rank by using
drop-down menus, sliders, checkboxes, or other controls. This can enable the commander to customize the view and focus on the relevant data13
? It can update the data daily by connecting to a data source that refreshes
automatically or on demand. This can ensure that the commander always sees the latest and most accurate information14
? It can present the data in a visual and intuitive way by using charts, graphs, tables,
or other elements. This can help the commander to understand and analyze the data more easily and effectively1

NEW QUESTION 219
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 221
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 222
Which of the following statistical methods requires two or more categorical variables?

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

Answer: B

Explanation:
This is because a chi-squared test is a type of statistical method that tests the association or independence between two or more categorical variables, such as gender, race, or occupation. A chi-squared test can be used to compare the observed frequencies of the categories with the expected frequencies under the null hypothesis of no association or independence. For example, a chi-squared test can be used to determine if there is a relationship between smoking and lung cancer. The other statistical methods do not require two or more categorical variables. Here is why:
Simple linear regression is a type of statistical method that models the relationship between a continuous dependent variable and a continuous or categorical independent variable, such as height, weight, or education level. A simple linear regression can be used to estimate the slope and intercept of the best-fitting line that describes how the dependent variable changes with the independent variable. For example, a simple linear regression can be used to predict the weight of a person based on their height.
Z-test is a type of statistical method that tests the significance of the difference between a sample mean and a population mean, or between two sample means, when the population standard deviation or the sample sizes are large enough. A z-test can be used to compare the average scores of two groups of students on a standardized test.
Two-sample t-test is a type of statistical method that tests the significance of the difference between two sample means when the population standard deviation is unknown or the sample sizes are small. A two-sample t-test can be used to compare the average salaries of two groups of employees in different departments.

NEW QUESTION 227
Which of the following is most likely to be used as a data-mining ETL tool?

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

Answer: A

NEW QUESTION 232
Which of the following is used for calculations and pivot tables?

A. IBM SPSS
B. SAS

C. Microsoft Excel
D. Domo

Answer: C

Explanation:
This is because Microsoft Excel is a type of software application that allows users to create, edit, and analyze data in spreadsheets, which are composed of rows and columns of cells that can store various types of data, such as numbers, text, or formulas. Microsoft Excel can be used for calculations and pivot tables, which are two common features or functions in data analysis. Calculations are mathematical operations or expressions that can be performed on the data in the cells, such as addition, subtraction, multiplication, division, average, sum, etc. Pivot tables are interactive tables that can
summarize and display the data in different ways, such as by grouping, filtering, sorting, or aggregating the data based on various criteria or categories. The other software applications are not used for calculations and pivot tables. Here is why:
IBM SPSS is a type of software application that allows users to perform statistical analysis and modeling on data sets, such as regression, correlation, ANOVA, etc. IBM SPSS does not use spreadsheets or cells to store or manipulate data, but rather uses data views or variable views to display the data in rows and columns. IBM SPSS does not have pivot tables as a feature or function, but rather has output views or charts to display the results of the analysis.
SAS is a type of software application that allows users to perform data management and analysis using a programming language that consists of statements and commands. SAS does not use spreadsheets or cells to store or manipulate data, but rather uses data sets or tables that are stored in libraries or folders. SAS does not have pivot tables as a feature or function, but rather has procedures or macros that can produce summary tables or reports based on the data.
Domo is a type of software application that allows users to create and share dashboards and visualizations that display data from various sources and systems, such as databases, cloud services, or web applications. Domo does not use spreadsheets or cells to store or manipulate data, but rather uses connectors or APIs to access and integrate the data from different sources. Domo does not have pivot tables as a feature or function, but rather has cards or widgets that can show different aspects or metrics of the data.

NEW QUESTION 236
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 239
What R package makes it easy to work with dates?

A. Lubridate.
B. Datemath.
C. Stringr.
D. ggplot.

Answer: A

Explanation:

Lubridate is an R package that makes it easier to work with dates and times.

NEW QUESTION 242
A data analyst needs to create a weekly recurring report on sales performance and distribute it to all sales managers. Which of the following would be the BEST method to automate and ensure successful delivery for this task?

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

Answer: A

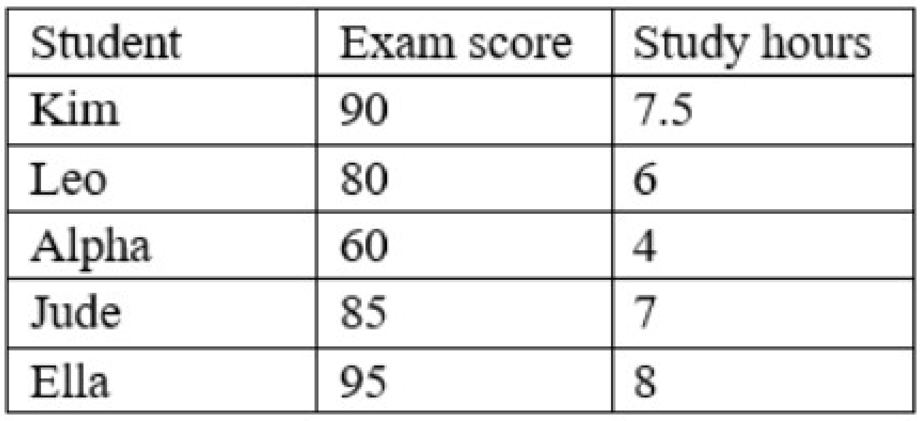
Explanation:
Scheduled report delivery is a feature that allows a data analyst to automate the generation and distribution of a report at a specified time and frequency. This would be the best method to ensure that the sales managers receive the weekly report on sales performance without manual intervention.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

NEW QUESTION 245
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 248
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

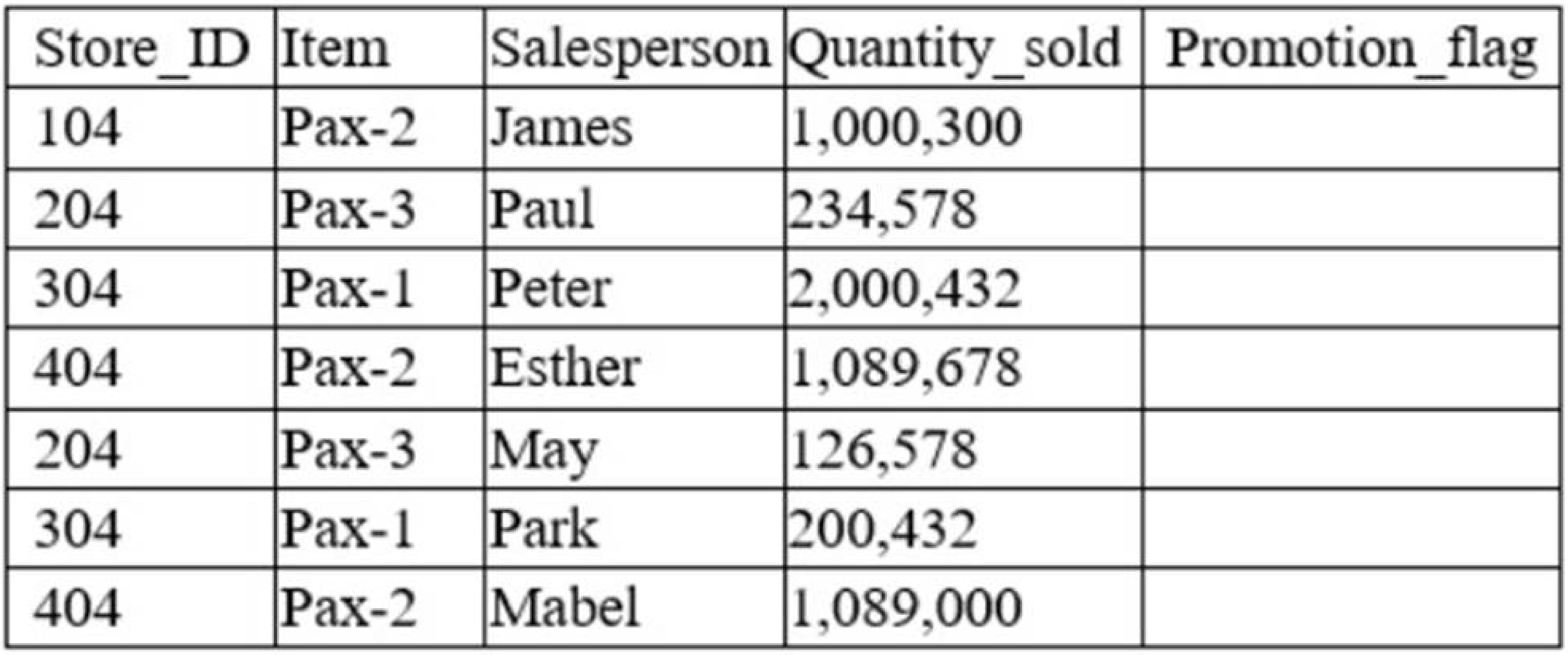
Which of the following charts would BEST represent the relationship between the variables?

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 253
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 255
Which of the following would be the best way to identify multicollinear attributes in a data set?

A. Correlation coefficient
B. Chi-squared test
C. Two-sample f-test
D. Two-way ANOVA

Answer: A

Explanation:
Multicollinearity in a dataset refers to the situation where two or more predictor variables are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In such cases, the correlation coefficient is a key statistical measure used to identify the presence of multicollinearity. It quantifies the degree to which two variables are linearly related.
The Variance Inflation Factor (VIF) is another commonly used metric that is derived from the correlation coefficient. It assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be equal to 1.
While the other options listed—Chi-squared test, Two-sample f-test, and Two-way ANOVA—are valuable statistical tools, they serve different purposes and are not typically used to detect multicollinearity. The Chi-squared test is used for testing relationships between categorical variables, the Two-sample f-test compares variances across groups, and Two-way ANOVA is used to understand the interaction between two independent categorical variables on a continuous dependent variable.
References:
? Multicollinearity in Regression Analysis: Problems, Detection, and Solutions1.
? What is multicollinearity and how to remove it?2.
? Detect and Treat Multicollinearity in Regression with Python3.

NEW QUESTION 257
Which of the following is a best practice when updating a legacy data source?

A. Placing old data in new fields
B. Keeping only the most recent data
C. Creating a codebook to document field changes
D. Removing the data source from production

Answer: C

Explanation:
When updating a legacy data source, it is a best practice to create a codebook to document field changes. A codebook serves as a detailed guide and record of the data structure, definitions, and any transformations or modifications made to the data fields. This documentation is crucial for maintaining data integrity, ensuring consistency, and facilitating future data use and understanding. It provides a reference that can be invaluable for data analysts, developers, and any stakeholders who need to work with the data.
Creating a codebook is preferred over placing old data in new fields, which can lead to confusion and data integrity issues. Keeping only the most recent data may result in the loss of valuable historical information. Removing the data source from production is not a practice related to updating data but rather to retiring a data source1234.
References:
? Legacy Data Migration: A Comprehensive Guide | OpenGeeksLab
? How to Successfully Complete Legacy Database Migration
? Methods for Saving and Integrating Legacy Data - DATAVERSITY
? Legacy Data Digitization - Learn The Best Practices

NEW QUESTION 262
A data analyst needs to create a data visualization that aids in un the cumulative impact of sequentially introduced values that are positive or negative. Which of the following data visualization methods should the analyst use?

A. A bubble chart
B. A waterfall chart
C. A scatter plot
D. A line chart

Answer: B

Explanation:
A waterfall chart is a type of data visualization that shows the cumulative impact of sequentially introduced values that are positive or negative. A waterfall chart typically has an initial value and a final value, with intermediate values shown as floating columns that either add to or subtract from the initial value. A waterfall chart can help visualize how different factors contribute to a net change in a value over time. Therefore, the correct answer is B. References: [Waterfall Chart | Definition & Examples - Investopedia], [Waterfall Charts in Excel | How to Create Waterfall Chart in Excel?] 4of30

NEW QUESTION 264
A data analyst needs to create a dashboard to help identify trends in the data sets. Which of the following is an appropriate consideration for dashboard development?

A. Data sources and attributes
B. Frequently asked questions
C. A report from the data source
D. A comparison of data sets

Answer: A

Explanation:
When creating a dashboard to identify trends in data sets, the most appropriate consideration is the data sources and attributes. This is because the quality, reliability, and structure of the data sources directly influence the dashboard??s ability to accurately reflect trends. Attributes, such as the type of data and the time frame it covers, are crucial for trend analysis. A well-designed dashboard should provide a clear and intuitive representation of the data, allowing for easy identification of trends and patterns. Frequently asked questions (B) can inform the design of the dashboard but are not a direct consideration for the development process itself. A report from the data source © might be an output of the dashboard but does not guide its development. A comparison of data sets (D) could be a feature of the dashboard, but the underlying data sources and attributes must be considered first to ensure accurate and meaningful comparisons. References:
? Best practices in dashboard design emphasize the importance of understanding and consolidating different data sources and creating a mix of useful metrics, which aligns with the choice of data sources and attributes1.
? Fundamental dashboard design principles include the clear and efficient display of information, which is dependent on the proper selection and use of data sources and attributes2.
? Effective dashboard communication is achieved by using colors, shapes, sizes, labels, and legends meaningfully, all of which rely on the underlying data sources and attributes3.

NEW QUESTION 268
A data analyst is developing a dashboard to track and monitor metrics. Which of the following best practices should be taken into during the FIRST pment process?

A. Create a A Aupirarrame:
B. Deploy to production.
C. Copy a dashboard design from the Internet.
D. Develop a dashboard.

Answer: A

Explanation:
A dashboard is a graphical display that summarizes and presents key performance indicators (KPIs) and metrics for a business or a project. A dashboard should be clear, concise, and easy to understand. To develop a dashboard, one of the best practices is to create a wireframe or a mockup first. A wireframe or a mockup is a low- fidelity sketch or prototype of the dashboard layout and design, which helps to define the scope, requirements, and functionality of the dashboard. Creating a wireframe or a mockup can help to save time and resources, as well as to get feedback from stakeholders and users before deploying the dashboard to production. Therefore, the correct answer is A. References: [Dashboard Design Best Practices: 4 Key Principles | Toptal], [How to Create an Effective Dashboard (with Examples) | Tableau]

NEW QUESTION 270
Which of the ing is the correct ion for a tab-delimited spre file?

A. tap
B. tar
C. sv
D. az

Answer: C

Explanation:
A tab-delimited spreadsheet file is a type of flat text file that uses tabs as delimiters to separate data values in a table. The file extension for a tab-delimited spreadsheet file is usually .tsv, which stands for tab-separated values. Therefore, the correct answer is C. References: [Tab-separated values - Wikipedia], [What is a TSV File?
| How to Open, Edit & Convert TSV Files]

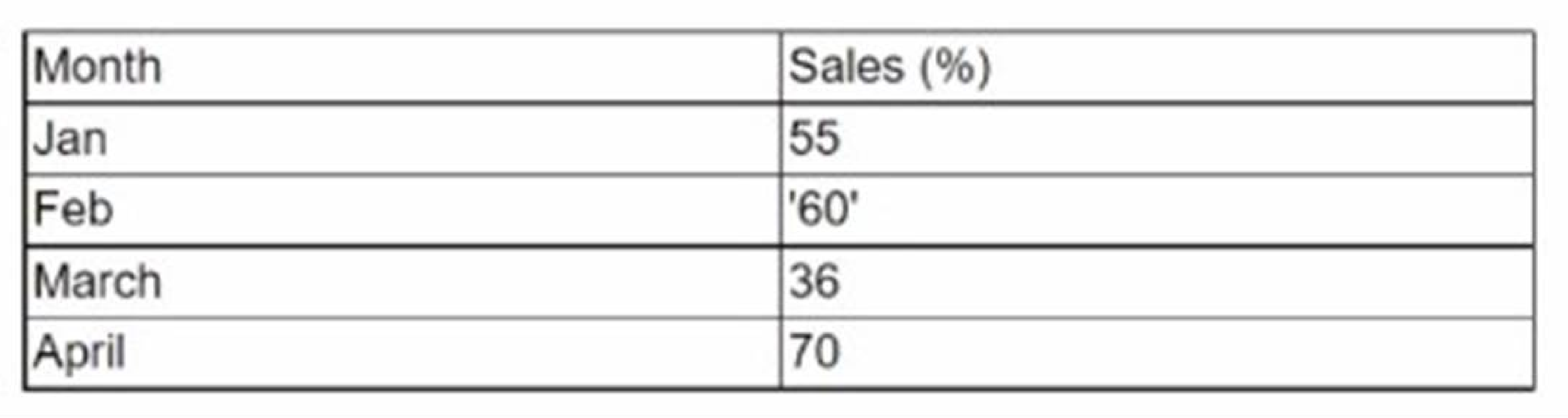
NEW QUESTION 275
A financial analyst is creating a daily billing report for a company. One night, the company's data warehouse did not update the data, which caused the data to be reported incorrectly the next day. Which of the following documentation elements should the analyst add to catch this error?

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

Answer: B

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

NEW QUESTION 276
A data set for sales per month includes the following data:



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

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

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

Answer: B

NEW QUESTION 277
A data analyst has been asked to create a daily manufacturing report for the floor manager Which of the following metrics should be included in the report?

A. Tons of steel produced per hour
B. Annual sales budget
C. End-of-day stock price
D. Daily corporate employee count

Answer: A

NEW QUESTION 282
Amanda needs to create a dashboard that will draw information from many other data sources and present it to business leaders.
Which one of the following tools is least likely to meet her needs?

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

Answer: D

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

NEW QUESTION 283
An analyst needs to conduct a quick analysis. Which of the following is the FIRST step the analyst should perform with the data?

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

Answer: A

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

NEW QUESTION 288
Which of the following is the best description of the term "data governance"?

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

Answer: D

Explanation:

Data governance refers to the overarching management of data??s availability, usability, integrity, and security within an organization. It involves setting policies and standards that govern data usage, determining data ownership, implementing data security measures, and ensuring that data is accessible for business insights while maintaining its quality. The goal of data governance is to ensure that data is consistent, trustworthy, and not misused, supporting compliance with data privacy regulations and enabling effective data analytics to optimize operations and drive business decision-making.
References:
? Understanding Data Governance and Its Importance1.
? The Role of Data Governance in Data Management2.
? Defining Data Governance and Its Business Value3.

NEW QUESTION 290
Which of the following technologies would be best suited for creating a multiple linear regression model?

A. Microsoft Power Bl
B. R
C. SQL
D. Tableau

Answer: B

Explanation:
R is a statistical programming language that is specifically designed for data analysis and statistical modeling, making it highly suitable for creating a multiple linear regression model. It has extensive libraries such as lm() for linear modeling, which simplifies the process of model creation, diagnostics, and interpretation. R also provides robust tools for data manipulation and visualization, which are essential for preparing data for regression analysis and understanding the results123.
While Microsoft Power BI, SQL, and Tableau have capabilities for regression analysis, they are more limited compared to R. Power BI and Tableau are primarily business intelligence tools that offer some built-in analytics capabilities, but they are not as comprehensive as
R. SQL is a database query language that can perform some statistical calculations, but it is not inherently designed for statistical modeling4567.
References:
? Multiple Linear Regression in R: Tutorial With Examples - DataCamp1.
? Implementing linear regression in Power BI - SQLBI5.
? Choosing a Predictive Model - Tableau6.
? How Predictive Modeling Functions Work in Tableau7.

NEW QUESTION 291
......