CompTIA

Exam Questions DA0-001

CompTIA Data+ Certification Exam

NEW QUESTION 1
What role in a data governance is typically responsible for day-to-day oversight of data use?

A. Data processors.
B. Data custodians
C. Data owners.
D. Data stewards.

Answer: D

NEW QUESTION 2
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 A portal that is refreshed daily would be too slow and outdated to monitor and diagnose issues when they occur.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 3
Which of the following is a domain-specific language used in programming that is designed for managing data that is held in a relational data stream management system?

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

Answer: B

Explanation:
SQL (Structured Query Language) is a domain-specific language used in programming, specifically designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS). It is the standard language for relational database management systems. SQL statements are used to perform tasks such as update data on a database, or retrieve data from a database. Unlike languages like Python or R, which are general-purpose programming languages, SQL is tailored specifically for database management and manipulation.
References:
? ResearchGate article on SQL1.
? SpringerLink chapter on Relational Databases and SQL Language2.
? DataCamp tutorial on SQL Server Installation3.
? Wikipedia page on SQL4.

NEW QUESTION 4
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.

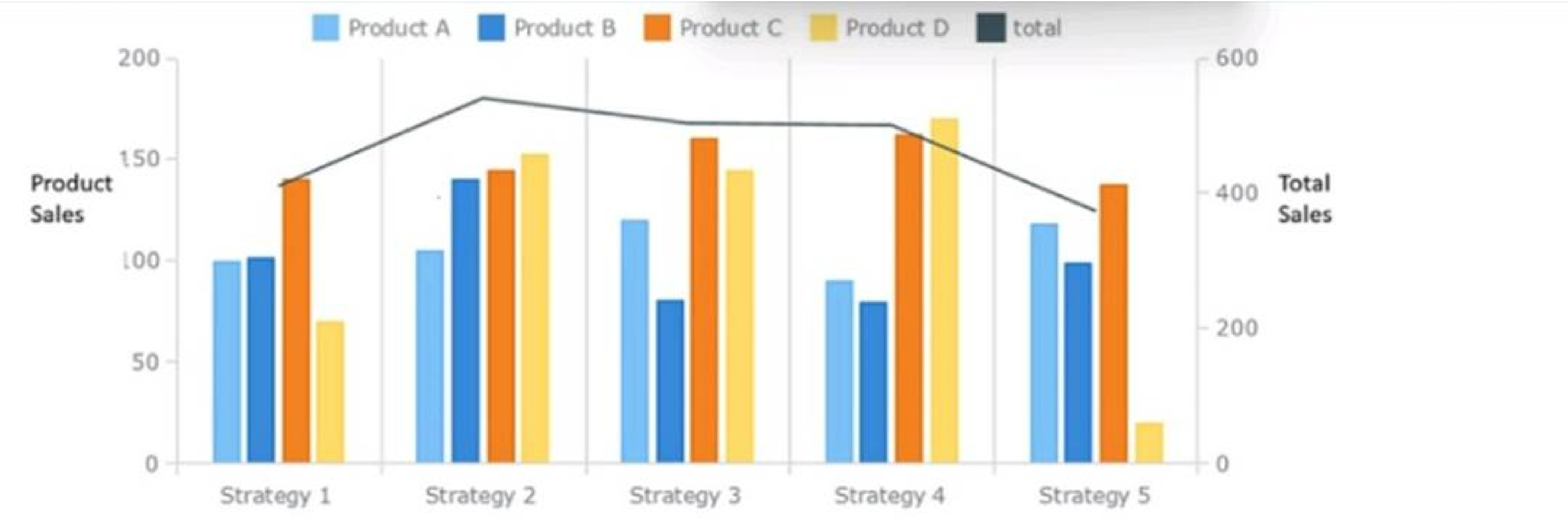
A. Duplicate Data.
B. Redundant Data.
C. Invalid Data.
D. Missing Data.

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 5
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 6
Analytics reports should follow corporate style guidelines.

A. True.
B. False.

Answer: A

NEW QUESTION 7
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 8
An analyst reviews the following data: 7
3
5
2

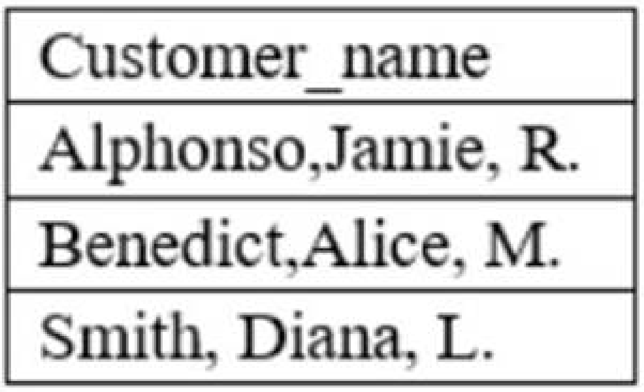
3
7
7
10
Which of the following is the value of the mode?

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

Answer: C

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

NEW QUESTION 9
A data analyst must separate the column shown below into multiple columns for each component of the name:



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

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

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

Answer: C

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

NEW QUESTION 10
A county in Illinois is conducting a survey to determine the mean annual income per household. The county is 427sq mi (2.65q km). Which of the following sampling methods would MOST likely result in a representative sample?

A. A stratified phone survey of 100 people that is conducted between 2:00 p.
B. and 3:00 p.m.
C. A systematic survey that is sent to 100 single-family homes in the county
D. Surveys sent to ten randomly selected homes within 5mi (8km) of the county??s office
E. Surveys sent to 100 randomly selected homes that are reflective of the population

Answer: D

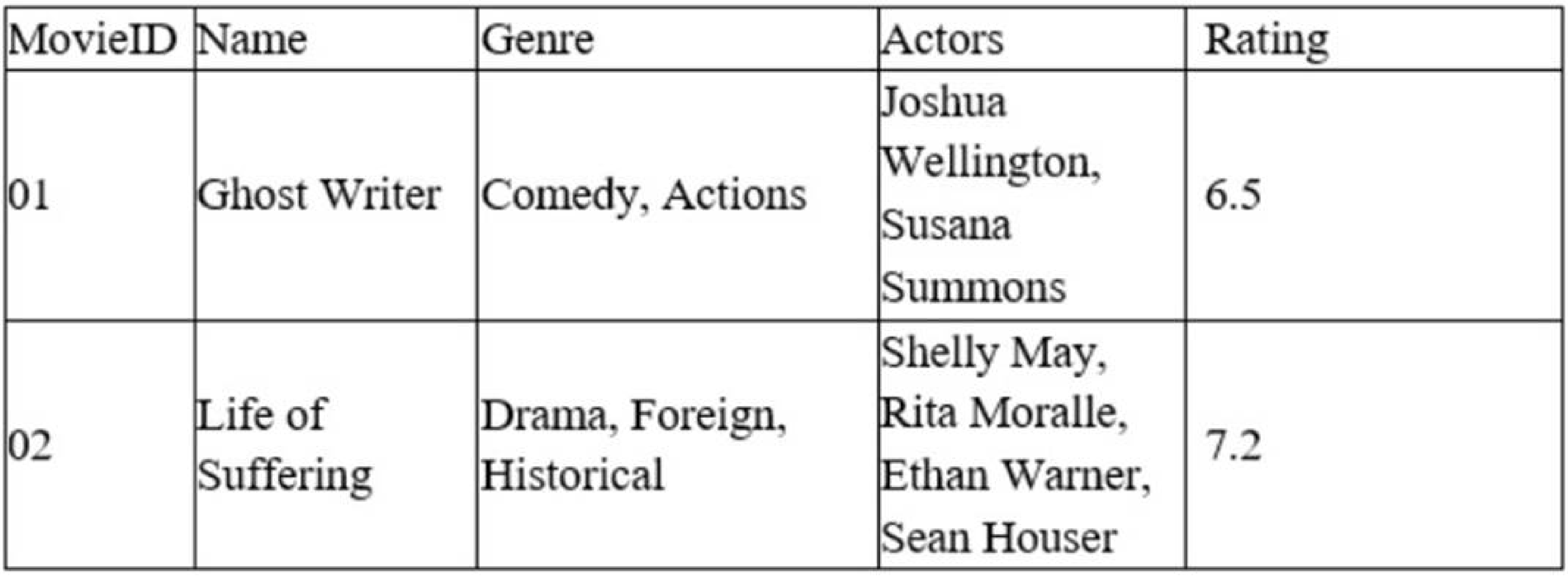
Explanation:
Surveys sent to 100 randomly selected homes that are reflective of the population. This is because a random sample is a type of sample that is selected by using a random method, such as a lottery or a computer-generated number, which ensures that every element in the population has an equal chance of being selected. A random sample can result in a representative sample, which means that the sample reflects the characteristics and diversity of the population. By sending surveys to 100 randomly selected homes that are reflective of the population, the analyst can ensure that the sample is representative of the county??s households and their income levels. The other sampling methods are not likely to result in a representative sample. Here is why:
A stratified phone survey of 100 people that is conducted between 2:00 p.m. and 3:00 p.m. would result in a biased sample, which means that the sample favors or excludes certain groups or elements in the population. By conducting the survey only between 2:00 p.m. and 3:00 p.m., the analyst would miss out on people who are not available or reachable at that time, such as those who are working or sleeping. This could affect the representativeness and generalizability of the sample.
A systematic survey that is sent to 100 single-family homes in the county would result in an unrepresentative sample, which means that the sample does not reflect the characteristics and diversity of the population. By sending surveys only to single-family homes, the analyst would ignore other types of households, such as apartments, condos, or mobile homes. This could affect the accuracy and reliability of the sample.
Surveys sent to ten randomly selected homes within 5mi (8km) of the county??s office would result in a small sample, which means that the sample size is too low to capture the variability and diversity of the population. By sending surveys only to ten homes within a limited area, the analyst would miss out on many households that are located in different parts of the county. This could affect the precision and confidence of the sample.

NEW QUESTION 10
A database administrator is required to mask certain table columns containing Pll in order to comply with the company privacy policy. Which of the following are the most likely types of information the administrator should mask? (Select two).

A. Government-issued ID
B. Address
C. Order ID
D. Order date
E. Customer ID
F. Referral number

Answer: AB

NEW QUESTION 13
A data analyst for a media company needs to determine the most popular movie genre. Given the table below:



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

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

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

Answer: D

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

NEW QUESTION 16
Which of the following techniques is used to quantify data?

A. Decoding
B. Enumeration
C. Coding
D. Structure

Answer: C

Explanation:
Answer C. Coding
Coding is a technique that is used to quantify data, especially qualitative data that are not expressed numerically. Coding involves assigning codes, such as numbers, letters, symbols, or colors, to different categories or themes that emerge from the data. For example, if you have a set of survey responses that ask about the satisfaction level of customers, you can code them as follows:
? Very satisfied = 5
? Satisfied = 4
? Neutral = 3
? Dissatisfied = 2
? Very dissatisfied = 1
By coding the data, you can convert them into quantitative data that can be analyzed using statistical methods, such as calculating the mean, median, mode, frequency, or percentage of each category12.
Option A is incorrect, as decoding is not a technique that is used to quantify data, but rather a process of interpreting or translating data from one form to another. For example, decoding can involve converting binary codes into text or images, or decrypting ciphertext into plaintext3.
Option B is incorrect, as enumeration is not a technique that is used to quantify data, but rather a process of listing or naming data in a specific order. For example,

enumeration can involve listing the names of the states in alphabetical order, or naming the planets in order of their distance from the sun4.
Option D is incorrect, as structure is not a technique that is used to quantify data, but rather a property or characteristic of data that describes how they are organized or arranged. For example, structure can refer to the format, type, or schema of data, such as structured, semi-structured, or unstructured data.

NEW QUESTION 20
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 25
Which of the following data types must be used when working with variables that require classification into two or more groups before analysis?

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

Answer: D

NEW QUESTION 28
Which of the following best describes the law of large numbers?

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

Answer: B

Explanation:
The best answer is B. As a sample size grows, its mean gets closer to the average of the whole population.
The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population. This is due to the sample being more representative of the population as it increases in size. The law of large numbers guarantees stable long-term results for the averages of some random events1
* A. As a sample size decreases, its standard deviation gets closer to the average of the whole population is not correct, because it confuses the concepts of standard deviation and mean. Standard deviation is a measure of how much the values in a data set vary from the mean, not how close the mean is to the population average. Also, as a sample size decreases, its standard deviation tends to increase, not decrease, because the sample becomes less representative of the population.
* C. As a sample size decreases, its mean gets closer to the average of the whole population is not correct, because it contradicts the law of large numbers. As a sample size decreases, its mean tends to deviate from the average of the whole population, because the sample becomes less representative of the population.
* D. When a sample size doubles, the sample is indicative of the whole population is not correct, because it does not specify how close the sample mean is to the population average. Doubling the sample size does not necessarily make the sample indicative of the whole population, unless the sample size is large enough to begin with. The law of large numbers does not state a specific number or proportion of samples that are indicative of the whole population, but rather describes how the sample mean approaches the population average as the sample size increases indefinitely.

NEW QUESTION 30
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 35
Which of the following types of analyses should be used to evaluate the connections and anomalies in a data set when either known patterns are being violated or new patterns are emerging?

A. Correlation
B. Descriptive
C. Graph
D. Regression

Answer: C

NEW QUESTION 36
A Chief Executive Officer (CEO) is requesting more up-to-date sales data for improved visibility prior to month-end. An analyst must determine the frequency of a sales report that was previously distributed on an as-needed basis. Which of the following would be the most appropriate frequency for this report?

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

Answer: C

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

NEW QUESTION 39
A research analyst collects ten data points from 1.000 specimens. The analyst will not need any additional data to complete the analysis and will not need to retrieve information by specifier. Which of the following is the best data structure for the analyst to use?

A. NoSQL
B. Flat file
C. JSON
D. Relational database

Answer: B

Explanation:
A flat file is a type of data structure that stores data in a plain text format, such as CSV, TSV, or TXT. A flat file consists of one or more records, each containing one or more fields, separated by a delimiter, such as a comma, tab, or space. A flat file does not have any hierarchical or relational structure, and does not support any complex queries or operations1.
A flat file may be the best data structure for the analyst to use in this scenario, because:
? The analyst collects ten data points from 1,000 specimens, which means the data is relatively small and simple, and can be easily stored and processed in a flat file.
? The analyst will not need any additional data to complete the analysis, which means the data is static and does not require any updates or modifications.
? The analyst will not need to retrieve information by specifier, which means the data
does not require any indexing or searching by key or value.

NEW QUESTION 41
Which of the following variable name formats would be problematic if used in the majority of data software programs?

A. First_Name_
B. FirstName
C. First_Name
D. First Name

Answer: D

Explanation:
This is because First Name is a variable name format that would be problematic if used in most of the data software programs, such as Excel, SQL, or Python. This is because First Name contains a space between two words, which could cause confusion or errors in the data software programs, as they might interpret the space as a separator or a delimiter between two different variables or values, rather than as part of a single variable name. For example, in SQL, a space is used to separate keywords, clauses,
or expressions in a statement, such as SELECT, FROM, WHERE, etc. Therefore, using First Name as a variable name in SQL could result in a syntax error or an unexpected result. The other variable name formats would not be problematic if used in most of the data software programs. Here is why:
? First_Name_ is a variable name format that uses an underscore (_) to separate two words, which is a common and acceptable practice in most of the data software programs, as it helps to improve the readability and clarity of the variable name. For example, in Python, an underscore is used to follow the PEP 8 style guide for naming variables, which recommends using lowercase letters and underscores for multi-word variable names.
? FirstName is a variable name format that uses camel case to separate two words,
which is another common and acceptable practice in most of the data software programs, as it helps to reduce the length and complexity of the variable name. For example, in Excel, camel case is used to follow the VBA naming conventions for naming variables, which recommends using mixed case letters for multi-word variable names.
? First_Name is a variable name format that also uses an underscore (_) to separate
two words, which is also a common and acceptable practice in most of the data software programs, as it helps to improve the readability and clarity of the variable name. For example, in SQL, an underscore is used to follow the ANSI SQL naming standards for naming variables, which recommends using lowercase letters and underscores for multi-word variable names.

NEW QUESTION 43
Which of the following query statements would be used when filtering data in a relational database management system? (Select two).

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

Answer: BC

NEW QUESTION 47
An analyst has written the following code: SELECT *
FROM Cust_table
WHERE age > 60 AND City = "New York"
Which of the following criteria is the analyst retrieving?

A. All customers older than age 60 in New York state
B. All customers aged 60 and older in New York state
C. All customers older than age 60 in New York City
D. All customers younger than age 60 in New York City

Answer: C

Explanation:
The SQL query provided is selecting all records from the Cust_table where the age column has values greater than 60 and the City column matches ??New York??. The > operator selects values that are strictly greater than the comparison value, so it does not include customers aged exactly 60. The term ??New York?? in the context of a city database typically refers to New York City, not the state of New York. Therefore, the correct answer is that the analyst is retrieving data for all customers older than age 60 in New York City.
References:
? The use of the > operator in SQL is to select values greater than the specified value1.
? Understanding the WHERE clause in SQL and its use in filtering records based on specified conditions2.
? Clarification on the distinction between city and state names in database records3.

NEW QUESTION 48
Joseph is interpreting a left skewed distribution of test scores. Joe scored at the mean, Alfonso scored at the median, and gaby scored and the end of the tail. Who had the highest score?

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

Answer: C

Explanation:

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

NEW QUESTION 49
The current date is July 14, 2020. A data analyst has been asked to create a report that shows the company??s year-over-year Q2 2020 sales. Which of the following reports should
the analyst compare?

A. A Q2 2020 and Q4 2019
B. YTD 2020 and YTD 2019
C. Q2 2020 and Q2 2019
D. Q2 2020 and Q2 2021

Answer: C

Explanation:
To create a report that shows the company??s year-over-year Q2 2020 sales, the analyst should compare the sales data from Q2 2020 and Q2 2019. Year-over-year (YoY) analysis is a method of comparing the performance of a business or a financial instrument over the same period in different years. It helps to identify trends, growth patterns, and seasonal fluctuations. Q2 refers to the second quarter of a year, which is usually from April to June. Therefore, the correct answer is C. References: YoY - Year over Year Analysis - Definition, Explanation & Examples, What is an Annual Sales Report: Definition, metrics, and tips - Snov.io

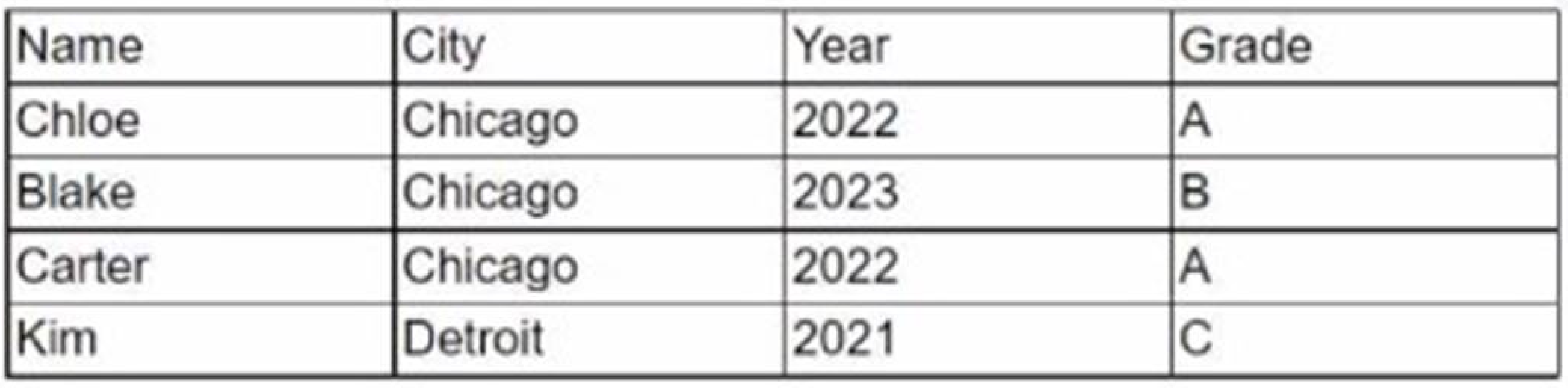
NEW QUESTION 53
Which of the following descriptive statistical methods are measures of central tendency? (Choose two.)

A. Mean
B. Minimum
C. Mode
D. Variance
E. Correlation
F. Maximum

Answer: AC

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

NEW QUESTION 58
An analyst needs to summarize the number of people in Chicago in 2022 using the following set of data:



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

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

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

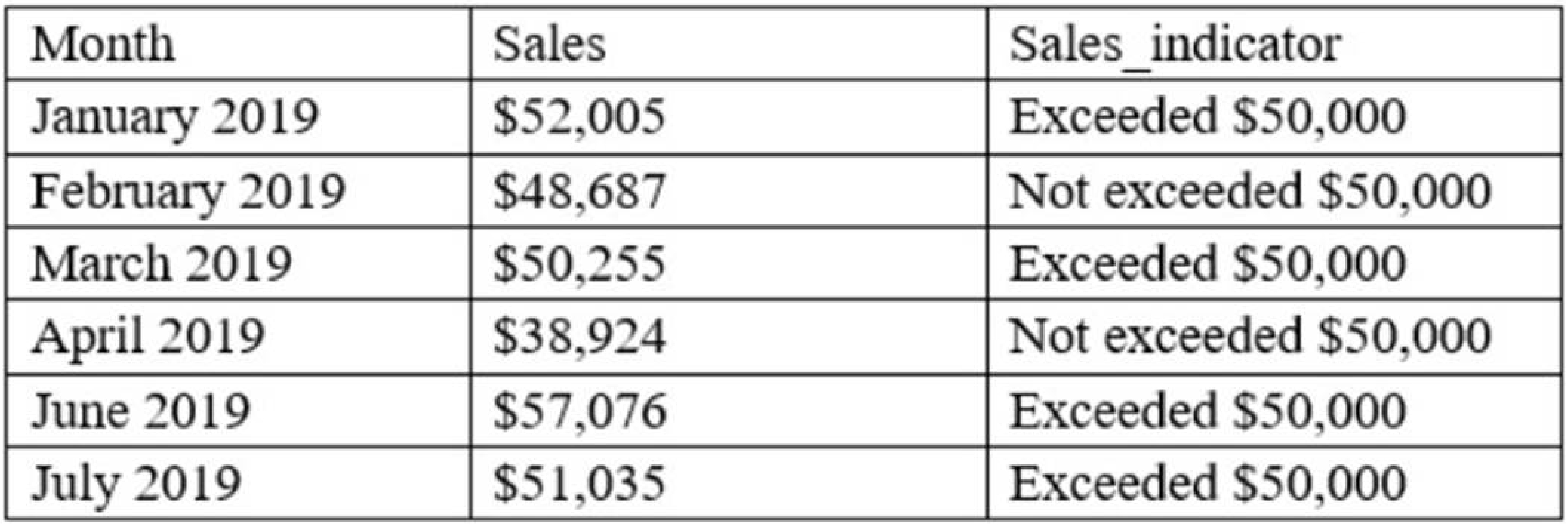
Answer: AC

NEW QUESTION 62
Which of the following is the best approach to use to gain a general understanding of a data set?

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

Answer: A

NEW QUESTION 65
An analyst has generated a report that includes the number of months in the first two quarters of 2019 when sales exceeded $50,000:



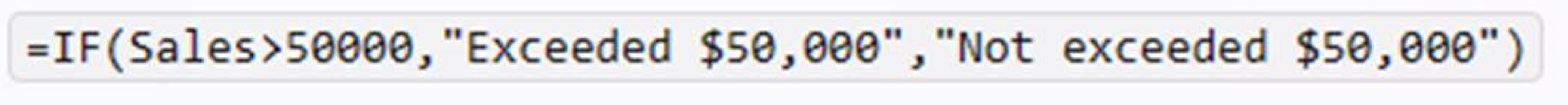
Month	Sales	Sales_indicator
January 2019	$52,005	Exceeded $50,000
February 2019	$48,687	Not exceeded $50,000
March 2019	$50,255	Exceeded $50,000
April 2019	$38,924	Not exceeded $50,000
June 2019	$57,076	Exceeded $50,000
July 2019	$51,035	Exceeded $50,000

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 67
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 72
Which of the following reports can be used when insight into operational performance is needed each Wednesday?

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

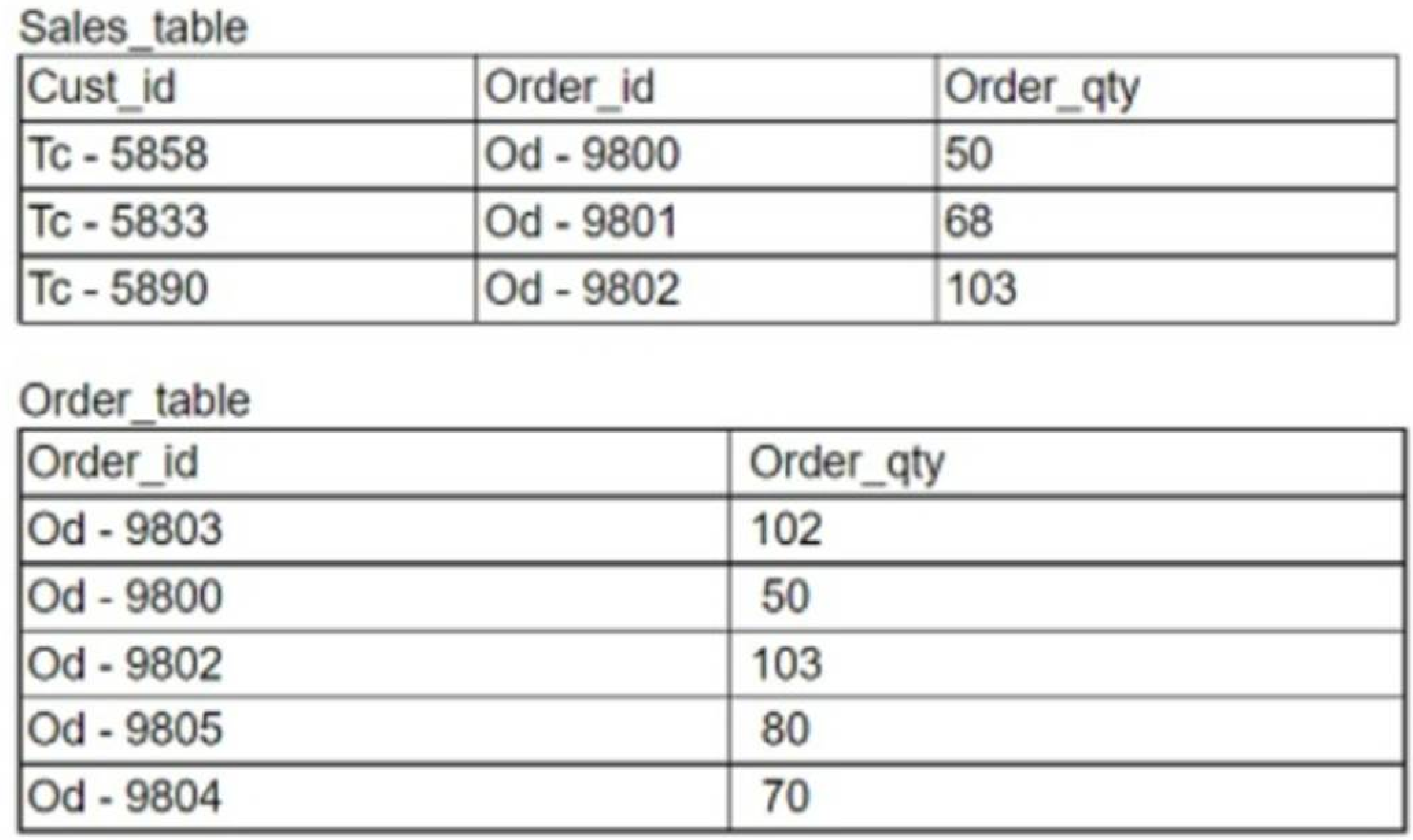
Answer: C

NEW QUESTION 73
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 75
A data analyst needs to perform a full outer join of a customer's orders using the tables below:



Sales_table

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

Order_table

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

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

A. 73.5
B. 76.5
C. 78.8
D. 81.5

Answer: D

Explanation:
The correct answer is D. OUTER JOIN, seven rows.
An OUTER JOIN is a type of SQL join that returns all the rows from both tables, regardless of whether there is a match or not. If there is no match, the missing side will have null values. An OUTER JOIN can be either a LEFT JOIN, a RIGHT JOIN, or a FULL JOIN, depending on which table??s rows are preserved1
Using the example tables, a FULL OUTER JOIN query would look like this:
SELECT Cust_id, Order_id, Order_qty FROM Sales_table FULL OUTER JOIN Order_table ON Sales_table.Order_id = Order_table.Order_id;
The result of this query would be:
Cust_id | Order_id | Order_qty --------??---------??--------- 1 | 1 | 100 2 | 2 | 50 3 | 3 | 25 4 | 4 |
75 NULL | 5 | 10 NULL | 6 | 20 NULL | 7 | 15
As you can see, the query returns seven rows, one for each order in either table. The orders that are not in the Sales_table have null values for the Cust_id column.
To find the mean of the order quantity, we need to sum up the order quantities and divide by the number of rows. In this case, the mean is (100 + 50 + 25 + 75 + 10 + 20 + 15) / 7 = 42.14. Rounding to one decimal place, we get 42.1 as the mean of the order quantity.

NEW QUESTION 79
Angela is aggregating data from CRM system with data from an employee system.
While performing an initial quality check, she realizes that her employee ID is not associated with her identifier in the CRM system.
What kind of issues is Angela facing? Choose the best answer.

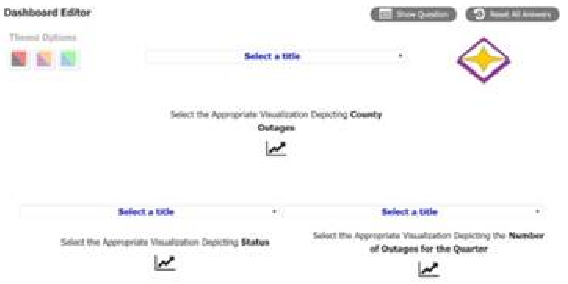
A. ETL process.
B. Record linkage.
C. ELT process.
D. System integration.

Answer: B

Explanation:

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

NEW QUESTION 80
SIMULATION
The director of operations at a power company needs data to help identify where company resources should be allocated in order to monitor activity for outages and restoration of power in the entire state. Specifically, the director wants to see the following:
* County outages
* Status
* Overall trend of outages INSTRUCTIONS:
Please, select each visualization to fit the appropriate space on the dashboard and choose an appropriate color scheme. Once you have selected all visualizations, please, select the appropriate titles and labels, if applicable. Titles and labels may be used more than once.
If at any time you would like to bring back the initial state of the simulation, please click the Reset All button.

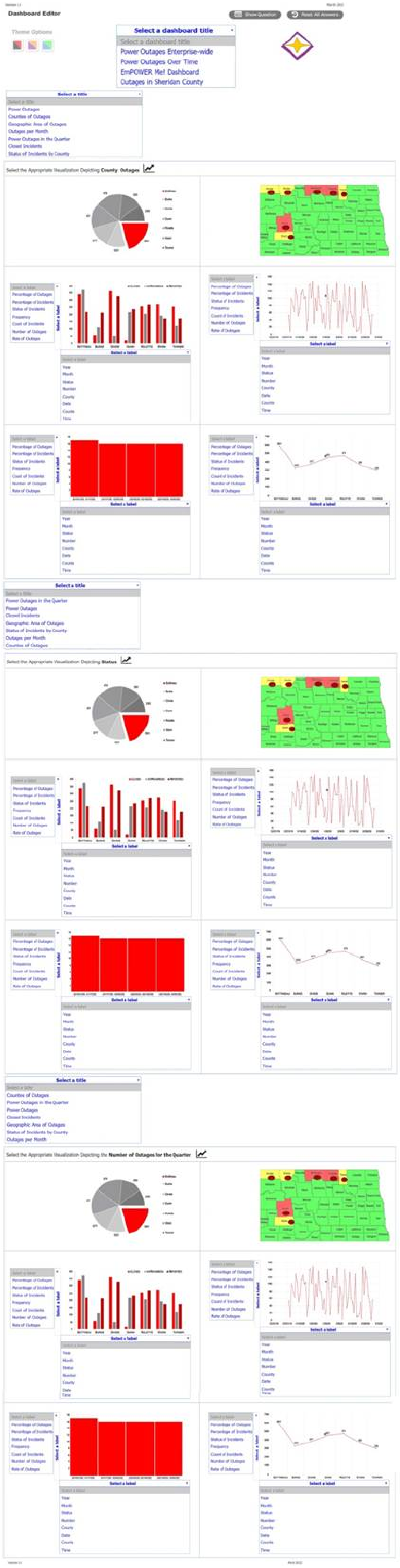


Dashboard Editor

Select the Appropriate Visualization Depicting **County Outages**

Select the Appropriate Visualization Depicting **Status**

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



Dashboard Editor

Select the Appropriate Visualization Depicting **County Outages**

Select the Appropriate Visualization Depicting **Status**

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
This is a simulation question that requires you to create a dashboard with visualizations that meet the director??s needs. Here are the steps to complete the task:
? Drag and drop the visualization that shows the county outages on the top left
space of the dashboard. This visualization is a map of the state with different colors indicating the number of outages in each county. You can choose any color scheme that suits your preference, but make sure that the colors are consistent and clear. For example, you can use a gradient of red to show the counties with more outages and green to show the counties with less outages.
? Drag and drop the visualization that shows the status of the outages on the top
right space of the dashboard. This visualization is a pie chart that shows the percentage of outages that are active, restored, or pending. You can choose any color scheme that suits your preference, but make sure that the colors are distinct and easy to identify. For example, you can use red for active, green for restored, and yellow for pending.
? Drag and drop the visualization that shows the overall trend of outages on the
bottom space of the dashboard. This visualization is a line graph that shows the number of outages over time. You can choose any color scheme that suits your preference, but make sure that the color is visible and contrasted with the background. For example, you can use blue for the line and white for the background.
? Select appropriate titles and labels for each visualization. Titles and labels may be
used more than once. For example, you can use ??County Outages?? as the title for the map, ??Status?? as the title for the pie chart, and ??Trend?? as the title for the line graph. You can also use ??County??, ??Number of Outages??, ??Active??, ??Restored??, ??Pending??, ??Time??, and ??Number of Outages?? as labels for the axes and legends of the visualizations.

NEW QUESTION 84
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

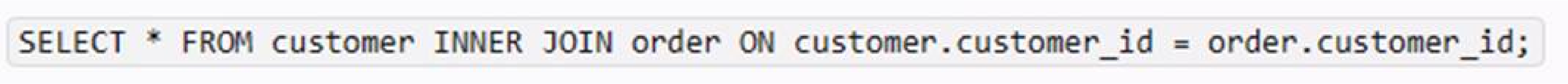
Explanation:
Creating a new report that automatically calculates weekly totals would streamline the process for the recipient. By setting this report to run at the end of business on Friday, it would provide the recipient with the necessary information for the entire week in one consolidated document. 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 86
Given the following customer and order tables:
Which of the following describes the number of rows and columns of data that would be present after performing an INNER JOIN of the tables?

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

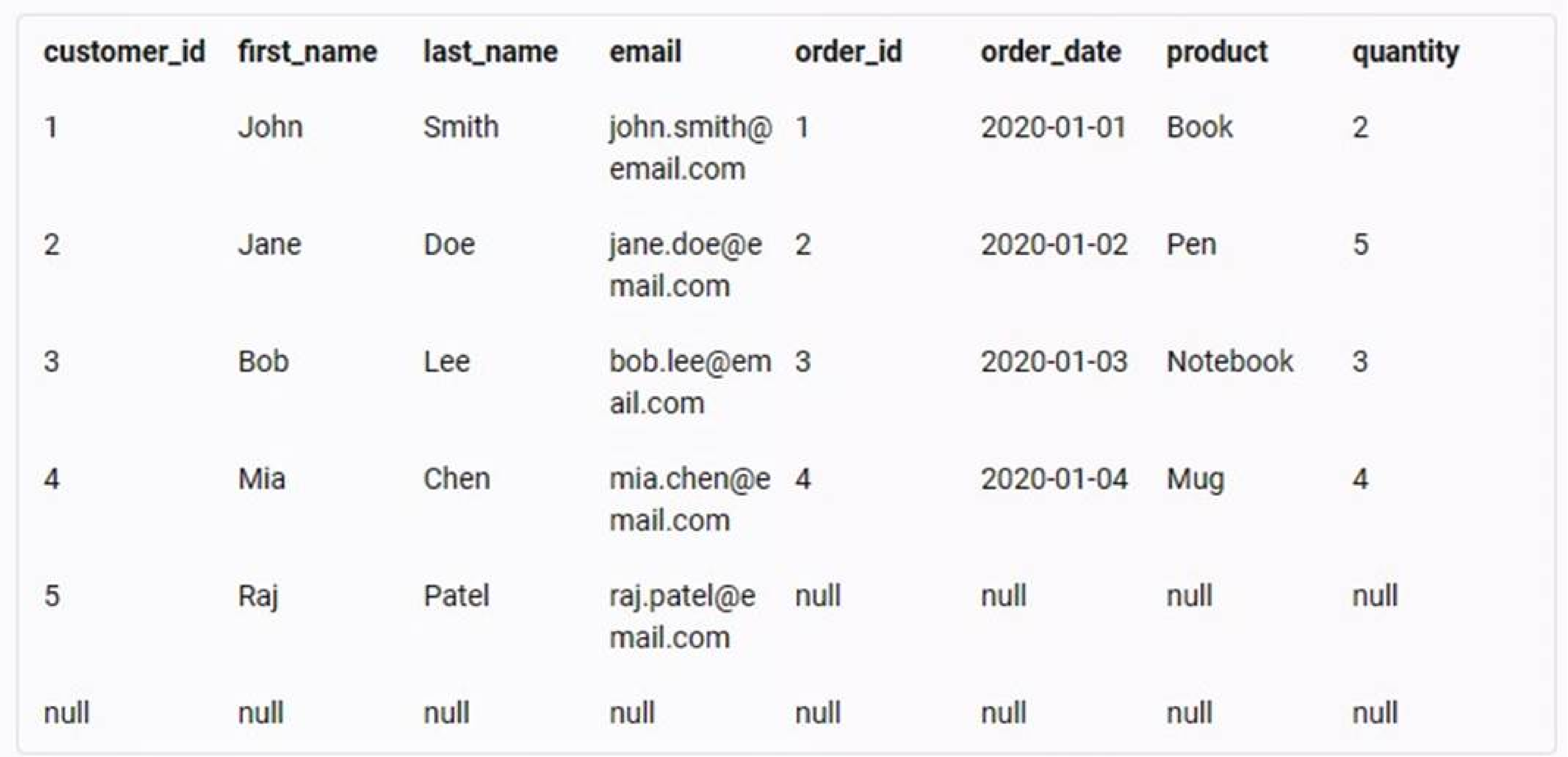
Answer: B

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



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

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



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

The reason why there are seven rows and eight columns in the result table is because:
? There are seven rows because there are six customers and six orders in the original tables, but only five customers have matching orders based on the customer ID column. Therefore, only five rows will have data from both tables, while one row will have data only from the customer table (customer 5), and one row will have no data at all (null values).
? There are eight columns because there are four columns in each of the original tables, and all of them are selected and joined in the result table. Therefore, the result table will have four columns from the customer table (customer ID, first name, last name, and email) and four columns from the order table (order ID, order date, product, and quantity).

NEW QUESTION 91
A data analyst is developing a data dictionary that aligns with a company's data management processes and policies. Which of the following best describes what should be included in the data dictionary?

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

Answer: C

NEW QUESTION 93
A customer's telephone number is in the format 123-456-7890. Which of the following data types is used for the phone number?

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

Answer: C

Explanation:
A telephone number, despite being composed of digits, is not used for calculations and often includes formatting characters such as hyphens (-). Therefore, the most appropriate data type for a telephone number is Text (oVr ARCHAR in SQL databases), which can accommodate various formats and lengths, and preserve leading zeros that might be present in some phone numbers. Storing phone numbers as numeric data types would strip away any formatting and could lead to the loss of significant leading digits (like zeros in international numbers).
? Boolean is a binary data type and only represents true or false values.
? Date is a data type used for dates.
? Number could technically store phone numbers, but it is not suitable due to the reasons mentioned above.
References:
? Best Practices for Storing Phone Numbers1
? Data Types in SQL for Phone Numbers2

NEW QUESTION 94
The senior management team at a company receives a detailed sales report at the end of each quarter. The report is several pages long and includes data from dozens of offices across the country. The team wants a better way to get a quick snapshot of what is included in the report. Which of the following modifications would best meet this requirement?

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

Answer: C

Explanation:
The purpose of an executive summary is to provide a concise and informative overview of a longer report, allowing busy stakeholders to quickly understand the key points and findings without reading the entire document. This summary should highlight the most important data, conclusions, and recommendations, and is typically placed at the beginning of the report for easy access12.
In the context of a detailed sales report for senior management, including a summary section with observations and insights would allow the team to quickly grasp the performance across various offices and identify any significant trends or issues that require attention. This approach aligns with best practices for executive reporting, which emphasize the importance of clear and concise summaries that focus on essential KPIs and actionable insights12.
References: 1: Databox - How to Write an Executive Summary for a Report: Step By Step Guide with Examples 2: LinkedIn - Best Practices for Writing Executive Summaries

NEW QUESTION 96
A stakeholder wants to see daily sales targets organized in a dashboard by country, state, city, and ZIP Code. Which of the following delivery considerations must a data analyst take into account when creating the dashboard?

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

Answer: B

NEW QUESTION 99
Which of the following data types best describe 4Ac1? (Select two).

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

Answer: AF

Explanation:
The term ??4Ac1?? is a combination of numbers and letters, which fits the definition of an alphanumeric string. Alphanumeric refers to a character set that contains both letters and numbers. In data analytics and programming, such a value is typically treated as a string, which is a sequence of characters. Strings can include letters, digits, and various other symbols.
A numeric data type would only include numbers, and a float is a specific kind of numeric data type that includes decimal points, neither of which applies to ??4Ac1??. A boolean data
type represents one of two values: true or false. Since ??4Ac1?? does not represent a true or false value, it cannot be classified as boolean. Lastly, symbolic is not a standard data type in the context of programming and data analytics.
References:
? Understanding Python 3 data types1.
? Basic Data Types in Python2.
? Java Data Types3.

NEW QUESTION 102
A data analyst is using a two-tailed, independent t-test to determine whether the type of stretching, dynamic or static, has any influence on a dancer's flexibility. Which of the following is the alternative hypothesis?

A. A dancer's flexibility is improved through static stretching.
B. The change in a dancer's flexibility is not equal to zero.
C. There is a difference in a dancer's flexibility between static and dynamic stretching.
D. The means of the static and dynamic stretching groups do not differ from each other.

Answer: C

NEW QUESTION 107
Which one of the following is a common data warehouse schema?

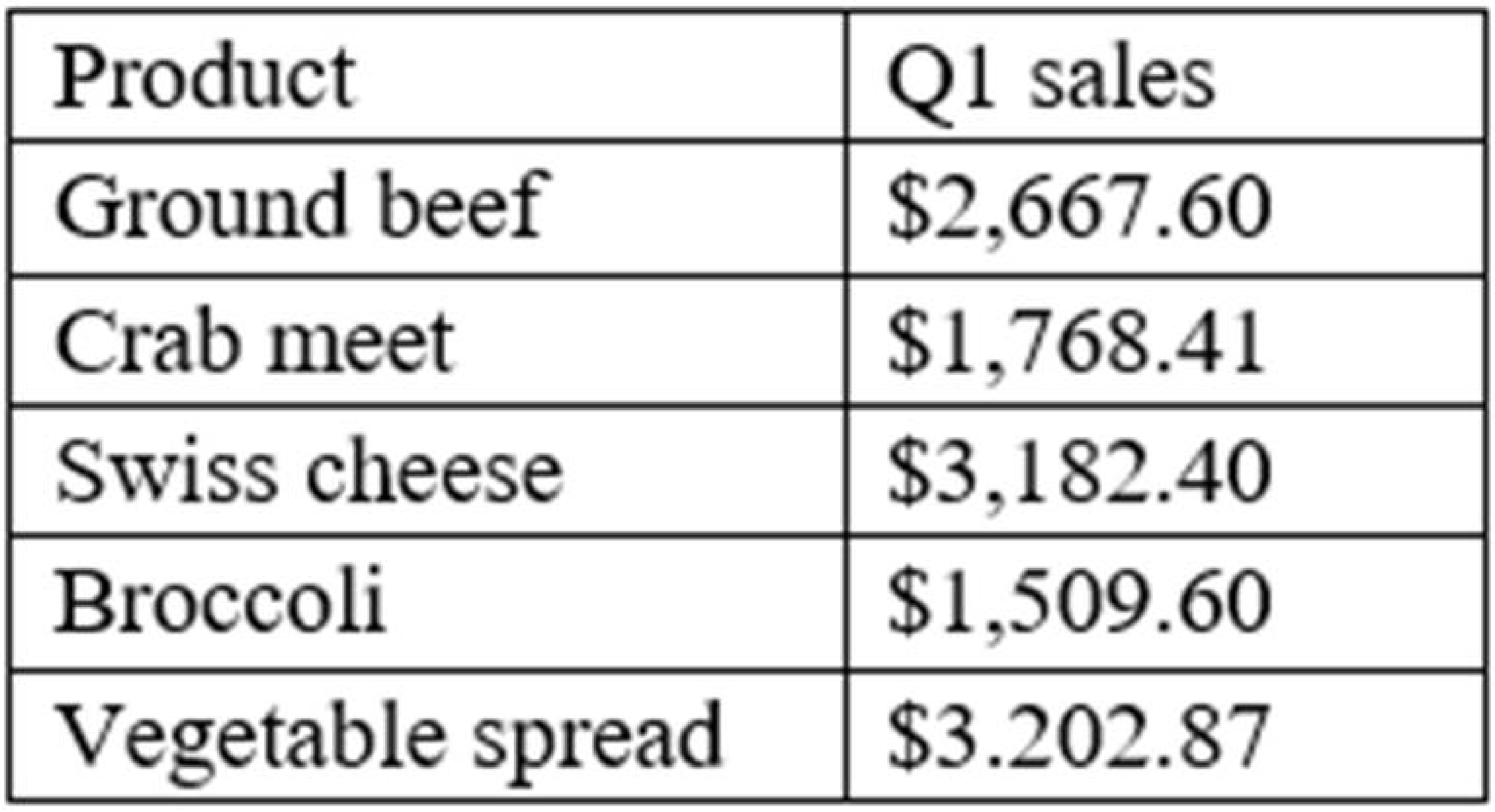
A. Snowflake.
B. Square.
C. Spiral.
D. Sphere.

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 109
A data analyst needs to calculate the mean for Q1 sales using the data set below:



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

Which of the following is the mean?

A. $2,466.18
B. $2,667.60
C. $3,082.72
D. $12,330.88

Answer: C

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

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

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

Answer: B

Explanation:

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

NEW QUESTION 118
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 121

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

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

Answer: B

NEW QUESTION 125
A data analyst has been asked to create one table that has each employee's first name, last name, sales, and address. The sales and addresses are listed in the tables below:

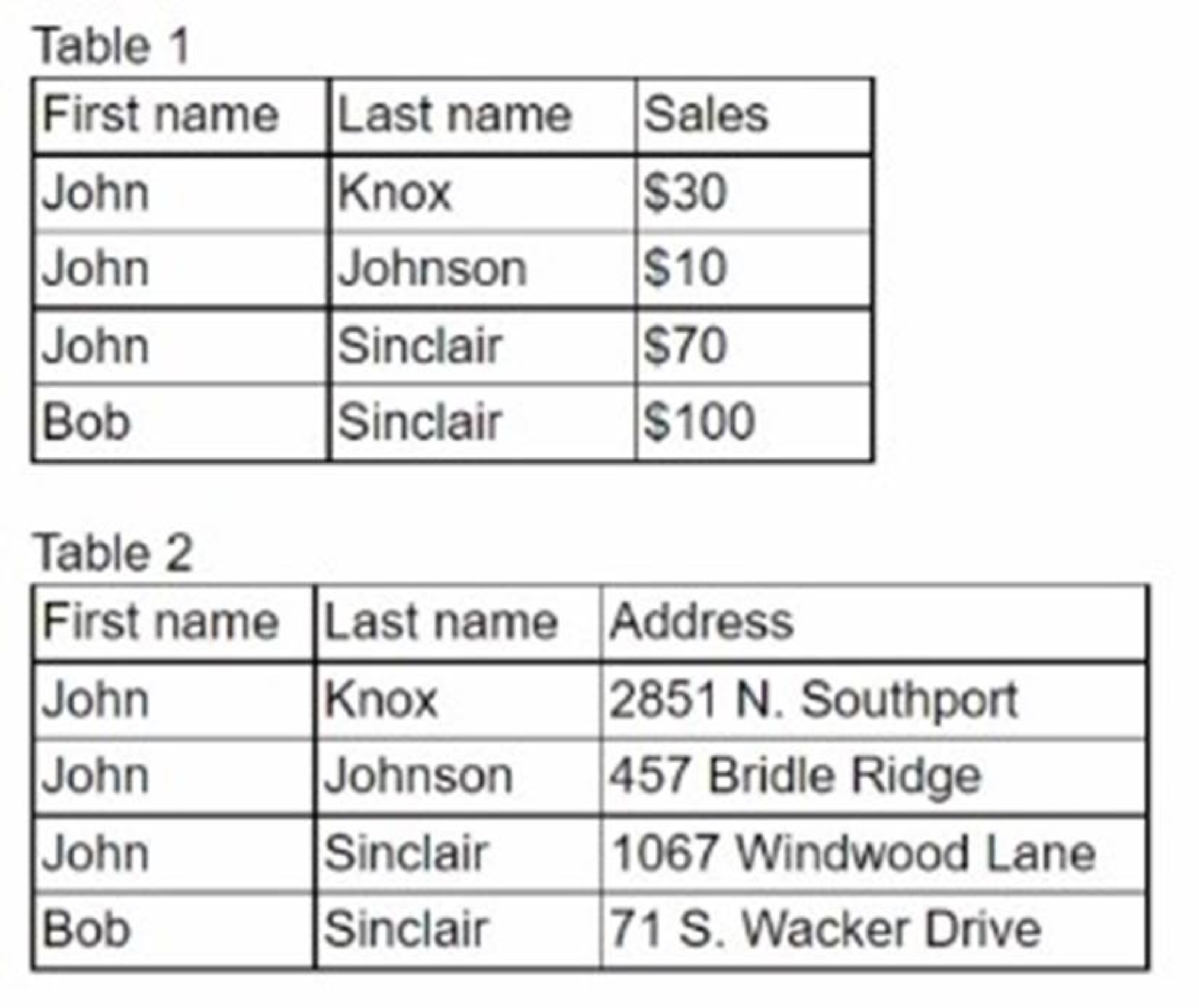


Table 1

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

Table 2

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

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

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

Answer: D

NEW QUESTION 129
Which of the following is an example of structured data?

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

Answer: A

Explanation:
A credit card number is an example of structured data, which is a type of data that conforms to a data model, has a well-defined structure, follows a consistent order, and can be easily accessed and used by a person or a computer program. A credit card number consists of 16 digits that are divided into four groups of four

digits each, separated by spaces or hyphens. The first six digits indicate the issuer identification number, the next nine digits indicate the account number, and the last digit is a check digit that validates the
number. A credit card number can be stored and processed in a structured format, such as a database or a spreadsheet1.

NEW QUESTION 130
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 134
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 136
Daniel is using the structured Query language to work with data stored in relational database.
He would like to add several new rows to a database table. What command should he use?

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

Answer: C

Explanation:
INSERT
The INSERT command is used to add new records to a database table.
The SELECT command is used to retrieve information from a database. It's the most commonly used command in SQL because it is used to pose queries to the database and retrieve the data that you're interested in working with.
The UPDATE command is used to modify rows in the database.
The CREATE command is used to create a new table within your database or a new database on your server.

NEW QUESTION 139
A company??s marketing department wants to do a promotional campaign next month. A data analyst on the team has been asked to perform customer segmentation, looking at how recently a customer bought the product, at what frequency, and at what value. Which of the following types of analysis would this practice be considered?

A. Prescriptive
B. Trend
C. Gap
D. Custer

Answer: D

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

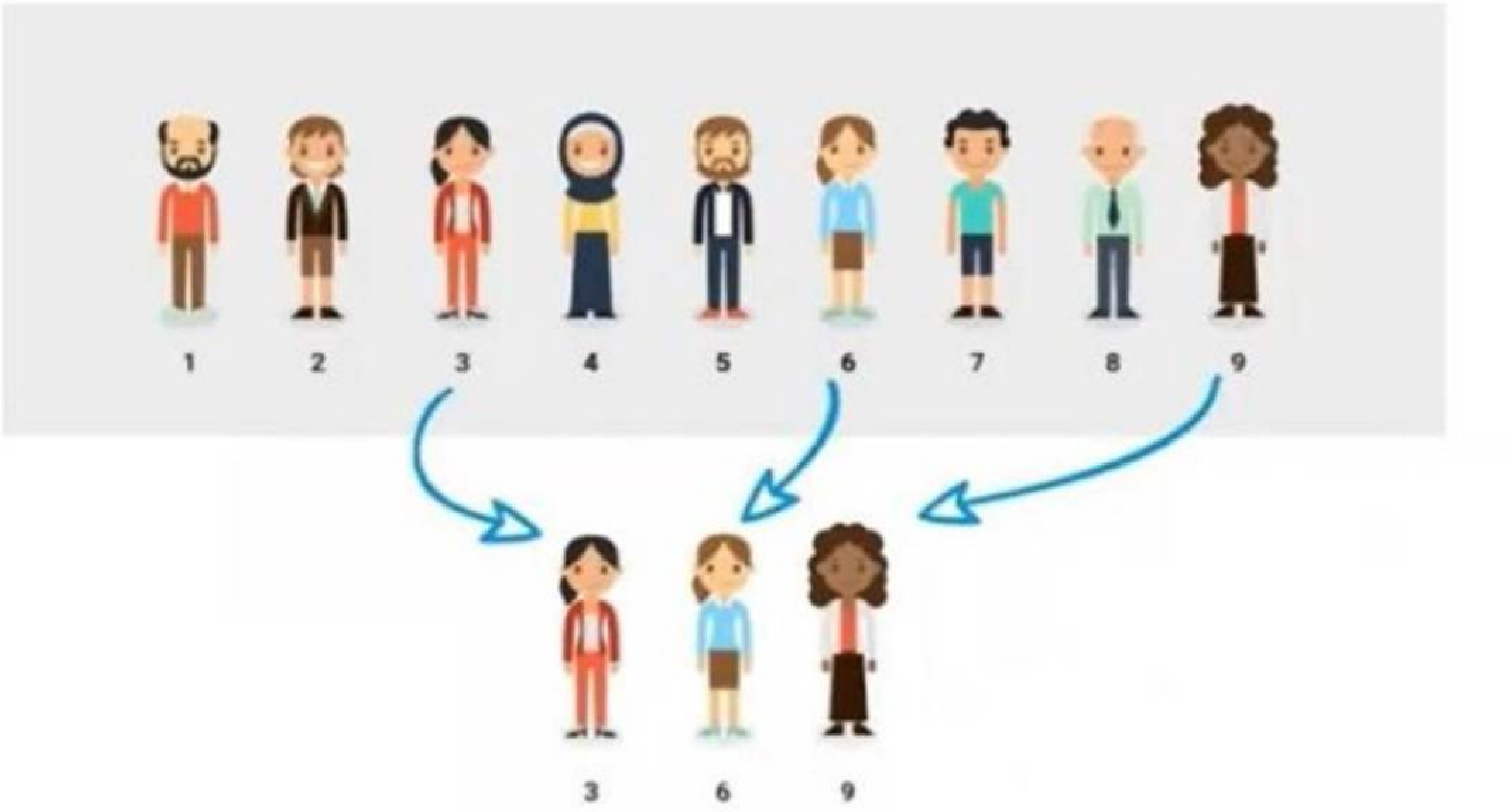
NEW QUESTION 144
Which of the following is the most likely reason for a data analyst to optimize a query using parameterization?

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

Answer: C

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

NEW QUESTION 148
Given the diagram below:



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

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

Answer: D

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

NEW QUESTION 152
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 157
A data engineer is creating a database field to capture whether a customer likes vanilla ice cream. Which of the following data types is the best to capture this information?

A. Integer
B. Boolean
C. Categorical

D. Numeric

Answer: B

NEW QUESTION 162
A data analyst has removed the outliers from a data set due to large variances. Which of the following central tendencies would be the best measure to use?

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

Answer: D

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

NEW QUESTION 166
Which one of the following values will appear first if they are sorted in descending order?

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

Answer: C

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

NEW QUESTION 171
The duration of a phone call in milliseconds is an example of:

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

Answer: D

Explanation:
The correct answer is D. Continuous data.
Continuous data is a type of quantitative data that can take any value within a range and can be measured with infinite precision. Continuous data can be expressed as fractions, decimals, or percentages. Examples of continuous data are height, weight, temperature, time, speed, etc12
The duration of a phone call in milliseconds is an example of continuous data, because it can take any value within a range (from zero to infinity) and can be measured with infinite precision (up to milliseconds or even smaller units). The duration of a phone call in milliseconds can also be expressed as fractions, decimals, or percentages of a larger unit (such as seconds, minutes, or hours).
Ordinal data is not correct, because ordinal data is a type of qualitative or categorical data that can be ordered or ranked according to some criterion. Ordinal data can have a logical order, but the intervals between the values are not equal or meaningful. Examples of ordinal data are grades, ratings, ranks, etc12
Nominal data is not correct, because nominal data is a type of qualitative or categorical data that can be labeled or named without any order or ranking. Nominal data can have a finite number of categories or classes, but the categories have no intrinsic value or hierarchy. Examples of nominal data are gender, color, nationality, etc12
Boolean data is not correct, because boolean data is a type of binary data that can have only two possible values: true or false. Boolean data can be used to represent logical statements, conditions, or outcomes. Examples of boolean data are yes/no, on/off, 1/0, etc.

NEW QUESTION 176
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 178
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?

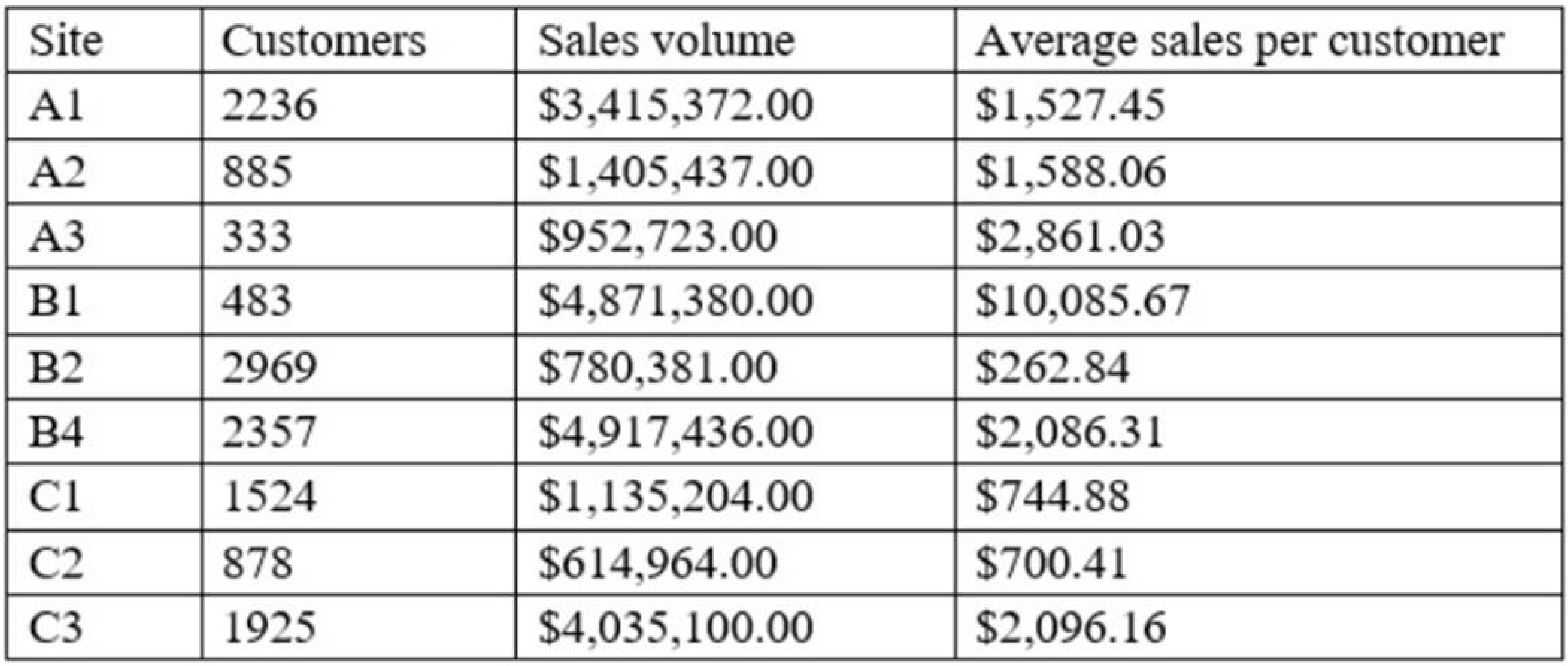
A. People who receive electronic coupons spend more on average.
B. People who receive electronic coupons spend less on average.
C. People who receive electronic coupons do not spend more on average.
D. People who do not receive electronic coupons spend more on average.

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 179
A data analyst is designing a dashboard that will provide a story of sales and determine which site is providing the highest sales volume per customer. The analyst must choose an appropriate chart to include in the dashboard. The following data is available:



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

Which of the following types of charts should be considered?

A. Include a line chart using the site and average sales per customer.
B. Include a pie chart using the site and sales to average sales per customer.
C. Include a scatter chart using sales volume and average sales per customer.
D. Include a column chart using the site and sales to average sales per customer.

Answer: C

Explanation:
A scatter chart using sales volume and average sales per customer is the best type of chart to include in the dashboard. A scatter chart is a type of chart that displays the relationship between two numerical variables using dots or markers. A scatter chart can show how one variable affects another, how strong the correlation is between them, and how the data points are distributed. In this case, a scatter chart can show the story of sales and determine which site is providing the highest sales volume per customer by plotting the sales volume on the x-axis and the average sales per customer on the y-axis. Each dot on the chart will represent a site, and the analyst can easily compare the sites based on their position on the chart. A site with a high sales volume and a high average sales per customer will be in the upper right quadrant, indicating a high performance. A site with a low sales volume and a low average sales per customer will be in the lower left quadrant, indicating a low performance. A site with a high sales volume and a low average sales per customer will be in the lower right quadrant, indicating a high volume but low value. A site with a low sales volume and a high average sales per customer will be in the upper left quadrant, indicating a low volume but high value. A scatter chart can also show if there is a positive or negative correlation between the two variables, or if there is no correlation at all. A positive correlation means that as one variable increases, so does the other. A negative correlation means that as one variable increases, the other decreases. No correlation means that there is no relationship between the two variables.
The other types of charts are not as suitable for this purpose. A line chart is a type of chart that displays the change of one or more variables over time using lines. A line chart can show trends, patterns, and fluctuations in the data. However, in this case, there is no time variable involved, so a line chart would not be appropriate. A pie chart is a type of chart that displays the proportion of each category in a whole using slices of a circle. A pie chart can show how each category contributes to the total and compare the relative sizes of each category. However, in this case, there are two numerical variables involved, so a pie chart would not be able to show their relationship. A column chart is a type of chart that displays the comparison of one or more variables across categories using vertical bars. A column chart can show how each category differs from each other and rank them by size. However, in this case, a column chart would not be able to show the

relationship between sales volume and average sales per customer, as it would only show one variable for each site.

NEW QUESTION 182
An analyst is preparing a report that contains weather data. The temperatures are shown in Fahrenheit. but they must be reported in Celsius. Which of the following should the analyst do to fix this issue?

A. Normalize the data.
B. Standardize the data.
C. Rescale the data.
D. Aggregate the data.

Answer: C

Explanation:
The analyst should rescale the data to fix this issue. Rescaling is a process of transforming data from one scale to another, such as changing the units of measurement. In this case, the analyst needs to rescale the temperatures from Fahrenheit to Celsius, which are two different scales for measuring temperature. To do this, the analyst can use the following formula:
Celsius = (Fahrenheit - 32) * 5/9
This formula converts each temperature value from Fahrenheit to Celsius by subtracting 32
and multiplying by 5/9. For example, if the temperature is 68??F, the rescaled value in Celsius is:
Celsius = (68 - 32) * 5/9 Celsius = 20??C
Rescaling the data can help the analyst to report the temperatures in a consistent and accurate way, and to avoid any confusion or errors that may arise from using different scales. Rescaling can also make the data more comparable and compatible with other data sources or standards that use the same scale12.

NEW QUESTION 187
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 188
Which of the following is the best description of discrete data types?

A. Non-numeric data used to describe attributes of a population sample
B. The frequency of the number of times each value occurs by using whole numbers
C. Numeric values that can be measured on a continuous scale
D. Non-numeric data used to describe attributes of a population sample ranked in a specific order

Answer: B

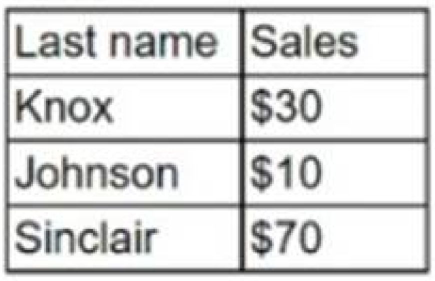
NEW QUESTION 192
Which of the following data protection methods provides confidentiality for data in transit?

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

Answer: B

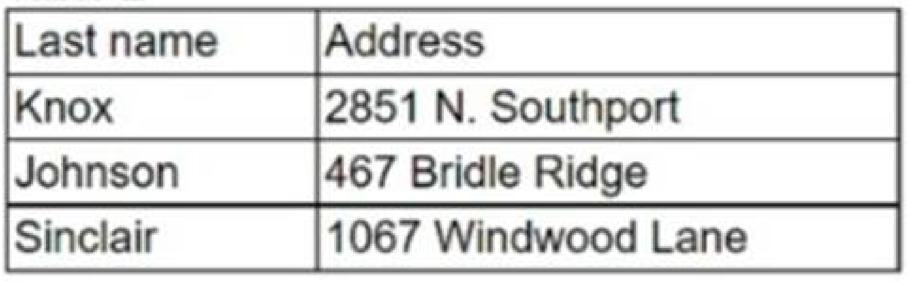
NEW QUESTION 193
A data analyst is performing a data merge within a spreadsheet using the tables below:
https://www.bing.com/images/blob?bcid=S1XCF9p02M4GjpbGxHj0lrIaj9sw.....4c

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 195
A research analyst wants to determine whether the data being analyzed is connected to other datapoints. Which of the following is the BEST type of analysis to conduct?

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

Answer: C

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

NEW QUESTION 199
Which of the following query optimization techniques involves examining only the data that is needed for a particular task?

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

Answer: C

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

documents can also help with searching, sorting, filtering, and aggregating the data in the documents12

NEW QUESTION 202
A collections manager has a team calling customers who are past due on their accounts in an attempt to collect payments. The manager receives the call list in the form of a printed report that is generated by the accounting department at the beginning of each week. Consequently, the collections team calls some customers who have made payments in the time since the report was last printed. Which of the following reporting enhancements could the accounting department implement to best reduce the number of calls on current accounts?

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

Answer: C

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

NEW QUESTION 203
Which of the following would a data analyst look for first if 100% participation is needed on survey results?

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

Answer: A

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

NEW QUESTION 208
An analyst has conducted a review of business questions. Which of the following should the analyst do next to conduct an analysis?

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

Answer: B

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

NEW QUESTION 209
Which one of the following would not normally be considered a summary statistic?

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

Answer: A

Explanation:

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

NEW QUESTION 210
The ACME Corporation hired an analyst to detect data quality issues in their Excel documents. Which of the following are the most common issues? (Select TWO)

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

Answer: DE

Explanation:

* 1. Duplicates
* 2. Misspellings
The most common data quality issues are difficult to resolve in Excel because of their rigidity. It forces analysts to do a ton of manual work, which results in a high probability of an error being introduced to the data set. Those common issues include:
- Blanks
- Nulls
- Outliers
- Duplicates
- Extra spaces
- Misspellings
- Abbreviations and domain-specific variations
- Formula error codes
When introduced, these errors can skew or even invalidate the resulting analysis. A smart tool would minimize the possibility of error by automating the manual work. In Excel, you might look for data quality issues in one of two ways. First, you might use auto filters on specific columns to scan for anomalies and blanks or you might use a pivot table to find gaps and discrepancies.
In either case, you're scanning for the anomalies yourself. Suffice it to say that's not a very efficient process. It also means accuracy is only as good as the analyst's eye, so the probability of error varies throughout the day.

NEW QUESTION 212
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 213
Which of the following statements would be used to append two tables that have the same number of columns?

A. UNION ALL
B. MERGE
C. GROUP BY
D. JOIN

Answer: A

Explanation:
The correct answer is A. UNION ALL.
UNION ALL is a SQL statement that appends two tables that have the same number of columns and compatible data types. UNION ALL preserves all the rows from both tables, including any duplicates12
* B. MERGE is not correct, because MERGE is a SQL statement that combines the data of two tables based on a common column. MERGE can perform insert, update, or delete operations on the target table depending on the matching or non-matching rows from the source table34
* C. GROUP BY is not correct, because GROUP BY is a SQL clause that groups the rows of a table based on one or more columns. GROUP BY is often used with aggregate functions, such as SUM, AVG, COUNT, etc., to calculate summary statistics for each group56
* D. JOIN is not correct, because JOIN is a SQL clause that combines the data of two tables based on a common column or condition. JOIN can produce different results depending on the type of join, such as INNER JOIN, LEFT JOIN, RIGHT JOIN, etc.

NEW QUESTION 215

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

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

Answer: B

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

NEW QUESTION 217
You are working with a dataset and want to change the names of categories that you used for different types of books.
What term best describes this action?

A. Recording.
B. Summarizing
C. Aggregating.
D. Filtering.

Answer: A

Explanation:
The term that best describes the action of changing the names of categories that you used for different types of books is recoding. Recoding is a process of transforming or modifying the values of a variable or a category to make them more meaningful, consistent, or accurate. For example, you can recode the names of book genres from ??Fiction??, ??Non-Fiction??, ??Biography??, etc. to ??FIC??, ??NF??, ??BIO??, etc. to make them shorter and easier to use. Reference:
Recoding Data - SPSS Tutorials - LibGuides at
Kent State University

NEW QUESTION 220
Which of the following best describes an exploratory analysis?

A. Involves the use of descriptive statistics to understand observations
B. Involves analysis of exploring data sets for performance tracking
C. Involves the testing of specific hypotheses
D. Involves the use of arithmetic algebra to determine the distribution

Answer: A

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

NEW QUESTION 224
A sales manager wants quarterly sales reports broken down by unit and week. Which of the following data output lists includes the most necessary information?

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

Answer: C

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

NEW QUESTION 228

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

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

Answer: B

NEW QUESTION 231
Which of the following is a common data analytics tool that is also used as an interpreted, high-level, general-purpose programming language?

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

Answer: D

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

NEW QUESTION 232
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 236
Which of the following database schemas features normalized dimension tables?

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

Answer: B

Explanation:
The correct answer is B. Snowflake.
A snowflake schema is a type of database schema that features normalized dimension tables. A database schema is a way of organizing and structuring the data in a database. A dimension table is a table that contains descriptive attributes or characteristics of the data, such as product name, category, color, etc. A normalized table is a table that follows the rules of normalization, which is a process of reducing data redundancy and improving data integrity by organizing the data into smaller and simpler tables12
A snowflake schema is a variation of the star schema, which is another type of database
schema that features denormalized dimension tables. A denormalized table is a table that does not follow the rules of normalization, and may contain redundant or duplicated data. A star schema consists of a central fact table that contains quantitative measures or facts, such as sales amount, order quantity, etc., and several dimension tables that are directly connected to the fact table. A snowflake schema differs from a star schema in that the dimension tables are further split into sub-dimension tables, creating a snowflake-like shape13
A snowflake schema has some advantages and disadvantages over a star schema. Some advantages are:
? It reduces the storage space required for the dimension tables, as it eliminates the
redundant data.
? It improves the data quality and consistency, as it avoids the update anomalies that may occur in denormalized tables.
? It allows more detailed analysis and queries, as it provides more levels of dimensions.
Some disadvantages are:

? It increases the complexity and number of joins required to retrieve the data from multiple tables, which may affect the query performance and speed.
? It reduces the readability and simplicity of the schema, as it has more tables and relationships to understand.
? It may require more maintenance and administration, as it has more tables to manage and update13

NEW QUESTION 237
A data analyst has been asked to create an ad-hoc sales report for the Chief Executive Officer (CEO).
Which of the following should be included in the report?

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

Answer: C

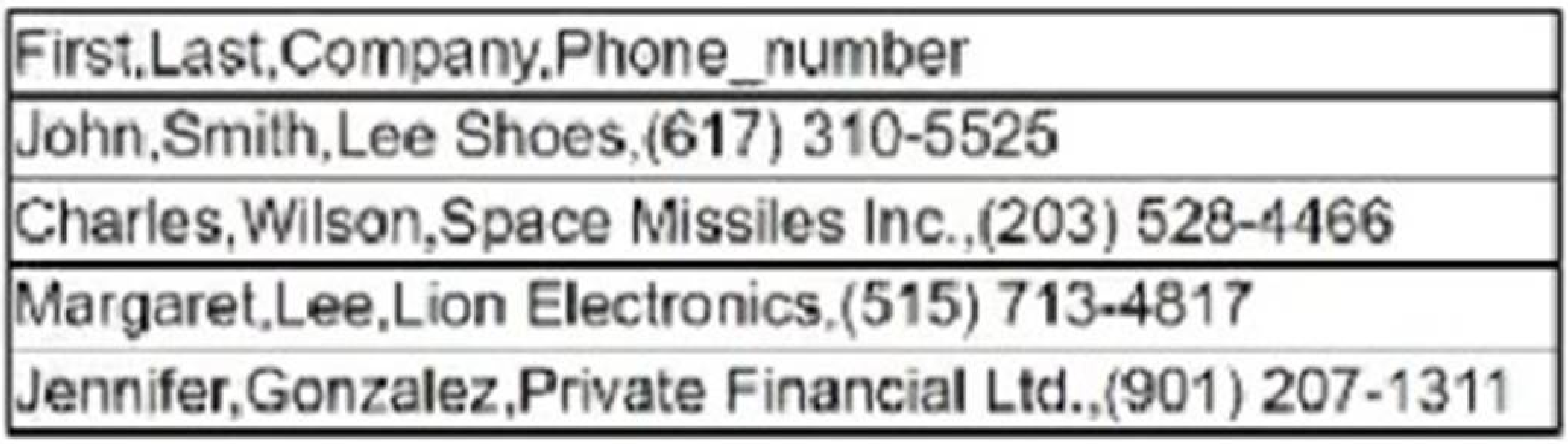
Explanation:
The report for the CEO should include YTD total sales, as this will provide a high-level overview of the sales performance of the company and show how it is meeting its annual goals. The other options are not appropriate for the CEO, as they are either too
detailed or irrelevant for the report. The sales representatives?? home addresses, line-item SKU numbers, and customers?? first and last names are not related to the sales performance and might compromise the privacy and security of the data.
Reference: CompTIA Data+ (DA0-001) Practice Certification Exams | Udemy

NEW QUESTION 238
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 240
Given the data below:



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

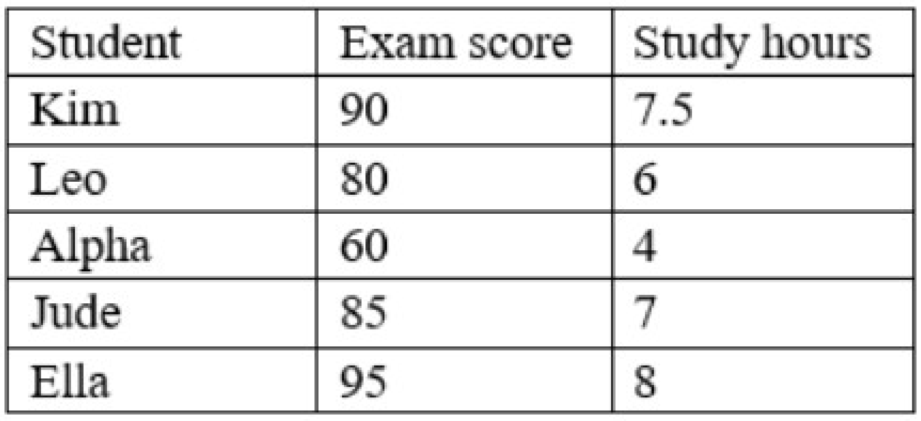
In which of the following file formats is the data presented?

A. Xs
B. CSV
C. RIF
D. XML

Answer: B

Explanation:
The data is presented in a CSV (comma-separated values) file format, which is a plain text format that stores tabular data. Each line of the file is a data record, and each record consists of one or more fields separated by commas. The first line of the file usually
contains the names of the fields, also known as the header. In this case, the data has four fields: Name, Age, Gender, and Occupation. Therefore, the correct answer is B. References: CSV File (What It Is & How to Open One), Comma-separated values - Wikipedia

NEW QUESTION 245
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 247
An analyst notices changes in sales ratios when analyzing a quarterly report. Which of the following is the analyst conducting?

A. A gap analysis
B. A link analysis
C. A trend analysis
D. A statistical analysis

Answer: C

Explanation:
When an analyst observes changes in sales ratios over a period, such as in a quarterly report, they are conducting a trend analysis. Trend analysis is a statistical method used to examine and evaluate the movement of data points over time to identify patterns or trends. This type of analysis is particularly useful for forecasting future events based on historical data. It differs from gap analysis, which assesses the difference between actual performance and potential or desired performance; link analysis, which is used to find associations among data; and statistical analysis, which is a broad term for all types of data analysis methods, including trend analysis.
References:
? Investopedia article on Ratio Analysis1.
? SpringerLink chapter on Financial Ratios Analysis2.
? ExamTopics page mentioning sales ratios in the context of analysis3.
? Investopedia definition of Ratio Analysis4.
? LiveWell article on Financial Ratio Analysis5.

NEW QUESTION 251
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.

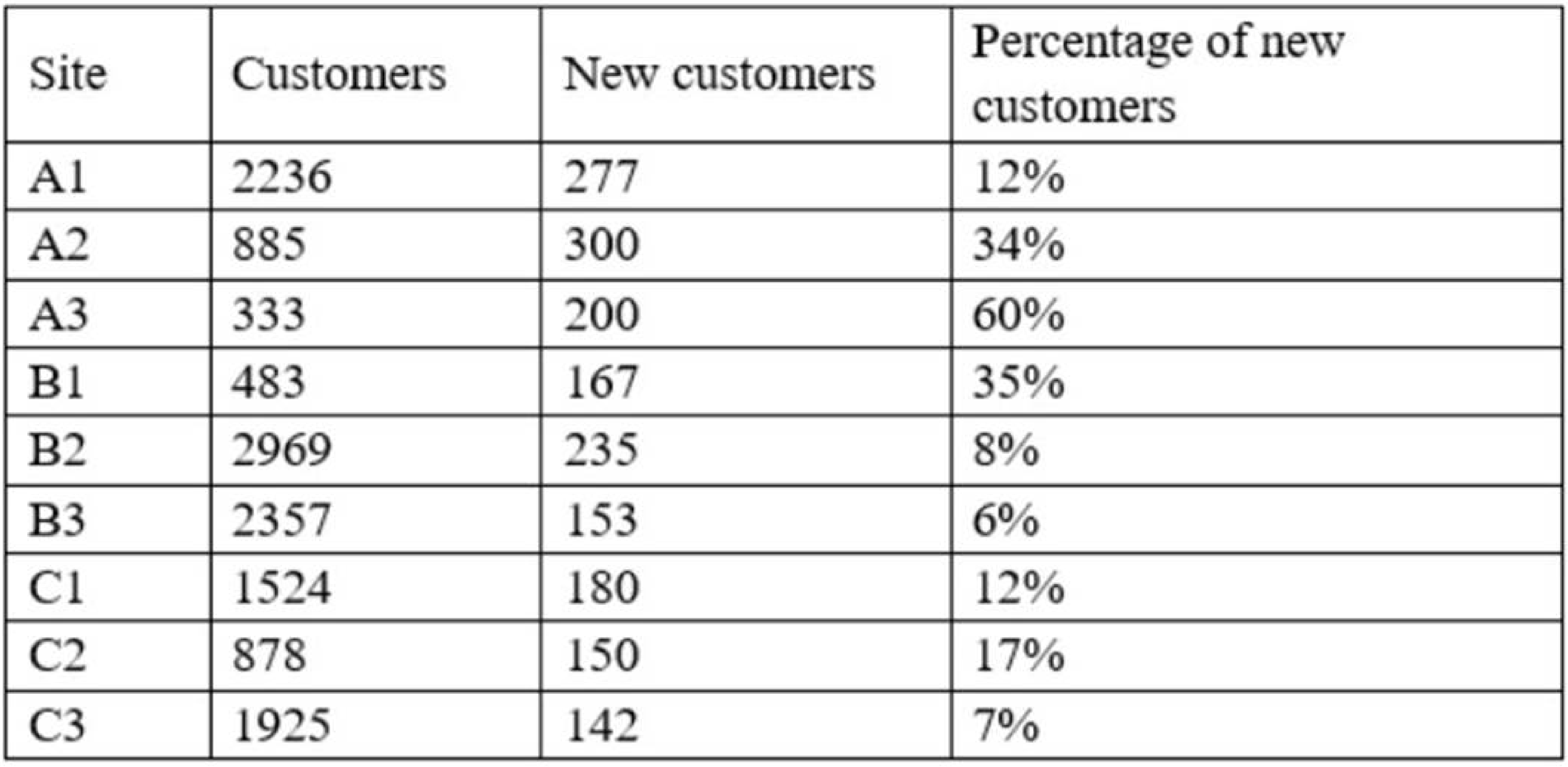
NEW QUESTION 254
A data analyst is developing a dashboard to track and monitor metrics. Which of the following best practices should be taken into during the FIRST pment process?

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 259
An analyst is designing a dashboard to determine which site has the highest percentage of new customers. The analyst must choose an appropriate chart to include in the dashboard. The following data is available:



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

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

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

Answer: A

Explanation:
This is because 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 the percentage of new customers for different sites 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 site has the highest or lowest percentage of new customers, as well as show how much each site contributes to the total percentage of new customers. The other types of charts are not the best charts to display the data. Here is why:
? A line chart is a type of chart that shows the change or the trend of a single variable over time, such as the percentage of new customers over months or years in this case. A line chart can be used to display and analyze the movement, cycle, or pattern of the variable, as well as identify any peaks, valleys, or fluctuations in the data. For example, a line chart can show how the percentage of new customers increases or decreases over time, as well as show if there are any seasonal or periodic variations in the data.
? A pie chart is a type of chart that shows the proportion or the percentage of a single variable for different categories or groups, such as the percentage of new customers for different sites in this case. A pie chart can be used to display and analyze the composition, distribution, or share of the variable, as well as identify any segments, slices, or fractions in the data. For example, a pie chart can show how much each site represents of the total percentage of new customers, as well

as show if there are any dominant or minor sites in the data.
? A scatter chart is a type of chart that shows the relationship between two variables for each observation or unit in a data set, such as the percentage of new customers and another variable for each site in this case. A scatter chart 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 chart can show if there is a positive, negative, or no correlation between the percentage of new customers and another variable, such as sales revenue or customer satisfaction.

NEW QUESTION 260
Q3 2020 has just ended, and now a data analyst needs to create an ad-hoc sales report that demonstrates how well the Q3 2020 promotion went versus last year's Q3 promotion.
Which of the following date parameters should the analyst use?

A. 2019 v
B. YTD 2020
C. Q3 2019 v
D. Q3 2020
E. YTD 2019 v
F. YTD 2020
G. Q4 2019 v
H. Q3 2020

Answer: B

Explanation:
The date parameters that the analyst should use are Q3 2019 vs. Q3 2020, as this will allow the analyst to compare the sales performance of the Q3 2020 promotion with the same period of last year. This will help to eliminate any seasonal or cyclical effects that might affect the sales data. The other options are not relevant for this purpose, as they either compare different quarters or different years. Reference: CertMaster Practice for Data+ Exam Prep - CompTIA

NEW QUESTION 262
Which of the following is a characteristic of a relational database?

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

Answer: C

Explanation:
It is structured in nature. This is because a relational database is a type of database that organizes data into tables, which consist of rows and columns. A relational database is structured in nature, which means that the data has a predefined schema or format, and follows certain rules and constraints, such as primary keys, foreign keys, or referential integrity. A relational database can be used to store, query, and manipulate data using a structured query language (SQL). The other characteristics are not true for a relational database. Here is why:
It utilizes key-value pairs. This is not true for a relational database, because key-value pairs are a way of storing data that associates each value with a unique key, such as an identifier or a name. Key-value pairs are typically used in non-relational databases, such as NoSQL databases, which do not have tables, rows, or columns, but rather store data in various formats, such as documents, graphs, or columns.
It has undefined fields. This is not true for a relational database, because fields are another name for columns in a table, which define the attributes or properties of each row or record in the table. Fields have defined names, types, and lengths in a relational database, which specify the format and size of the data that can be stored in each field.
It uses minimal memory. This is not true for a relational database, because memory is the amount of space or storage that is used by a database to store and process data. Memory usage depends on various factors, such as the size, complexity, and number of tables and queries in a relational database. A relational database can use a lot of memory if it has many tables with many rows and columns, or if it performs complex or frequent queries on the data.

NEW QUESTION 266
A data analyst needs to present the results of an online marketing campaign to the marketing manager. The manager wants to see the most important KPIs and measure the return on marketing investment. Which of the following should the data analyst use to BEST communicate this information to the manager?

A. A real-time monitor that allows the manager to view performance the day the campaign was launched
B. A sell-service dashboard that allows the manager to look at the company's annual budget performance
C. A spreadsheet of the raw data from all marketing campaigns and channels
D. A summary with statistics, conclusions, and recommendations from the data analyst

Answer: D

Explanation:
The option that the data analyst should use to best communicate the information to the manager is a summary with statistics, conclusions, and recommendations from the data analyst. A summary is a concise and clear way of presenting the main findings and insights from the data analysis report. A summary should include relevant statistics that support the conclusions and recommendations from the data analyst. A summary should also highlight the most important KPIs and measure the return on marketing investment in relation to the objectives of the online marketing campaign. The other options are not as effective as using a summary to communicate the information to the manager, as they either provide too much or too little information or do not address the manager??s needs or expectations. A real-time monitor may provide too much information that can be overwhelming or distracting for the manager who wants to see only the most important KPIs and measure the return on marketing investment. A self-service dashboard may provide too little information that can be insufficient or unclear for the manager who wants to see some guidance and interpretation from the data analyst. A spreadsheet of raw data may provide irrelevant or inaccurate information that can be confusing or misleading for the manager who wants to see some analysis and insights from the data analyst. Reference: [How to Write an Executive Summary for Your Data Analysis Report - Towards Data Science]

NEW QUESTION 271
An analyst has been tracking company intranet usage and has been asked to create a chat to show the most-used/most-clicked portions of a homepage that contains more than 30 links. Which of the following visualizations would BEST illustrate this information?

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

Answer: B

Explanation:
This is because a heat map is a visualization that uses colors to represent different values or intensities of a variable. A heat map can be used to show the most-used/most-clicked portions of a homepage that contains more than 30 links by assigning different colors to each link based on how frequently they are clicked by the users. For example, a link that is clicked very often can be colored red, while a link that is clicked rarely can be colored blue. A heat map can help the analyst to identify which links are more popular or important than others on the homepage. The other visualizations are not as effective as a heat map for this purpose. Here is why:
A scatter plot is a visualization that uses dots or points to represent the relationship between two variables. A scatter plot cannot show the most-used/most-clicked portions of a homepage that contain more than 30 links because it does not have a clear way of mapping each link to a point on the graph.
A pie chart is a visualization that uses slices or sectors to represent the proportion of each category in a whole. A pie chart cannot show the most-used/most-clicked portions of a homepage that contains more than 30 links because it does not have enough space to display all the categories clearly and accurately.
An infographic is a visualization that uses images, icons, charts, and text to convey information or tell a story. An infographic cannot show the most-used/most-clicked portions of a homepage that contain more than 30 links because it does not have a consistent or standardized way of representing each link and its click frequency.

NEW QUESTION 274
An analyst needs to create an analytics dashboard for an employee intranet site to improve the search functionality, display relevant information, and maintain an updated FAQ page. Which of the following visualizations would best represent what employees are searching for?

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

Answer: A

Explanation:
A word cloud is an ideal choice for visualizing what employees are searching for on an intranet site. It represents the frequency of word occurrence in a visually impactful way, with more commonly searched terms appearing larger in the cloud. This allows for quick identification of the most popular queries and topics of interest among employees. Unlike histograms, pie charts, or scatter plots, word clouds can effectively display textual data, which is the nature of search queries. They are particularly useful for analyzing text data from surveys or feedback forms, which can be similar to search query data in an intranet environment1234.
References: 1: ??What Are Word Clouds? Pros & Cons of Word Cloud Visualizations?? - Alida 2: ??Using Word Clouds for Powerful Data Visualization?? - WordCloud.app blog 3: ??Ultimate Google Data Studio Word Cloud Guide: Visualization 2024?? - AtOnce 4: ??How to Create Word Cloud in Power BI?? - Zebra BI

NEW QUESTION 279
Which of the following would be used to store unstructured data from different sources?

A. A data lake
B. A database management system
C. A database
D. A data warehouse

Answer: A

Explanation:
This is because a data lake is a type of storage system that stores unstructured data from different sources, such as text, images, audio, video, etc. A data lake can be used to store unstructured data from different sources by using a schema-on- read approach, which means that it does not impose any structure or format on the data when it is stored, but rather applies it when it is read or accessed. A data lake can also be used to store unstructured data from different sources by using a distributed file system, such as Hadoop, which means that it can store large volumes and varieties of data across multiple servers or nodes. The other storage systems are not used to store unstructured data from different sources. Here is why:
? A database management system is a type of software application that manages and controls databases, which are collections of structured or semi-structured data
that are organized into tables, rows, and columns. A database management system is not used to store unstructured data from different sources, but rather to store structured or semi-structured data from specific sources by using a schema- on-write approach, which means that it imposes a structure or format on the data when it is stored, and requires it to follow certain rules and constraints, such as primary keys, foreign keys, or referential integrity.
? A database is a type of storage system that stores structured or semi-structured
data that are organized into tables, rows, and columns. A database is not used to store unstructured data from different sources, but rather to store structured or semi-structured data from specific sources by using a relational model, which means that it establishes and maintains relationships between different tables based on common columns or keys. A database can also be used to store structured or semi-structured data from specific sources by using a query language, such as SQL, which means that it can access and manipulate the data using statements or commands.
? A data warehouse is a type of storage system that stores structured or semi-structured data that are integrated and aggregated from different sources or systems, such as databases, cloud services, or web applications. A data warehouse is not used to store unstructured data from different sources, but rather to store structured or semi-structured data from various sources by using an ETL process, which means that it extracts, transforms, and loads the data into a common format, structure, or schema. A data warehouse can also be used to store structured or semi-structured data from various sources by using an OLAP model, which means that it supports online analytical processing of the data using multidimensional cubes or queries.

NEW QUESTION 283
Which of the following is an example of a at flat file?

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

Answer: D

NEW QUESTION 286
What analytics suite is offered by Microsoft and directly integrates with SQL Server Databases?

A. Qlik.
B. Power BI.
C. Domo.
D. Dataroma.

Answer: B

Explanation:

Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data may be an Excel spreadsheet or a collection of cloud-based and on- premises hybrid data warehouses.

NEW QUESTION 287
A data analyst is creating a report that will provide information about various regions, products, and time periods. Which of the following formats would be the most efficient way to deliver this report?

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

Answer: D

Explanation:
The best format to deliver this report is D. A dashboard with filters at the top that the user can toggle.
A 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 glance1 A dashboard with filters at the top that the user can toggle would allow the user to easily and quickly access the information they need about various regions, products, and time periods, without having to navigate through multiple tabs, pages, or emails. A dashboard with filters would also enable the user to compare and contrast different views of the data and see how they change over time. A dashboard with filters would also be more interactive and engaging than a static or email report2
A workbook with multiple tabs for each region would not be an efficient way to deliver this report, because it would require the user to switch between different tabs to see the information they need. This would make it harder to compare and contrast different regions, products, and time periods, and also increase the risk of errors or confusion. A workbook with multiple tabs would also be less visually appealing and more cluttered than a dashboard3
A daily email with snapshots of regional summaries would not be an efficient way to deliver this report, because it would limit the user??s ability to explore the data in depth and customize their view. A daily email would also be dependent on the frequency and timing of the email delivery, which might not match the user??s needs or preferences. A daily email
would also be more likely to be ignored or deleted than a dashboard that is always accessible.
A static report with a different page for every filtered view would not be an efficient way to deliver this report, because it would create a very long and cumbersome report that would be difficult to read and understand. A static report would also not allow the user to change or update the filters as they wish, or see how the data changes over time. A static report would also be less interactive and engaging than a dashboard.

NEW QUESTION 291
......