CompTIA

Exam Questions DA0-001

CompTIA Data+ Certification Exam

NEW QUESTION 1
A table in a hospital database has a column for patient height in inches and a column for patient height in centimeters. This is an example of:

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

Answer: D

Explanation:
This is because redundant data is a type of data that is unnecessary or irrelevant for the analysis or purpose, which can affect the efficiency and performance of the analysis or process. Redundant data can be caused by having multiple data fields that store the same or similar information, such as patient height in inches and patient height in centimeters in this case. Redundant data can be eliminated or reduced by using data cleansing techniques, such as removing or merging the redundant data fields. The other types of data are not examples of data that is unnecessary or irrelevant for the analysis or purpose. Here is what they mean in terms of data quality:
? Dependent data is a type of data that relies on or is influenced by another data field or value, such as a formula or a calculation that uses other data fields or values as inputs or outputs. Dependent data can be useful or important for the analysis or purpose, as it can provide additional information or insights based on the existing data.
? Duplicate data is a type of data that is repeated or copied in a data set, which can
affect the quality and validity of the analysis or process. Duplicate data can be caused by having multiple records or rows that have the same or similar values for one or more data fields or columns, such as customer ID or order ID. Duplicate data can be eliminated or reduced by using data cleansing techniques, such as removing or filtering out the duplicate records or rows.
? Invalid data is a type of data that is incorrect or inaccurate in a data set, which can
affect the validity and reliability of the analysis or process. Invalid data can be caused by having values that do not match the expected format, type, range, or rule for a data field or column, such as an email address that does not have an @ symbol or a date that does not follow the YYYY-MM-DD format. Invalid data can be eliminated or reduced by using data cleansing techniques, such as validating or correcting the invalid values.

NEW QUESTION 2
A data set was recorded using multimedia technology. Which of the following is a necessary step on the way to interpretation?

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

Answer: B

Explanation:
The correct answer is B. Transcription.
Transcription is a necessary step on the way to interpretation when a data set was recorded using multimedia technology. Multimedia technology refers to the use of various forms of media, such as audio, video, images, and text, to capture and present information1 Transcription is the process of converting multimedia data into written or textual form, which can then be analyzed using various methods and tools2 Transcription can help to make the data more accessible, searchable, and manageable, as well as to preserve the data for future use.
Structural equation modeling is not correct, because it is a statistical technique that tests the causal relationships between multiple variables using observed and latent variables. Structural equation modeling is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data.
Sequential analysis is not correct, because it is a method of analyzing the order and timing of events or behaviors in a data set. Sequential analysis is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data. Sampling is not correct, because it is the process of selecting a subset of data from a larger population for analysis. Sampling is not a necessary step on the way to interpretation, but rather a preliminary step that can be done before collecting or analyzing the data.

NEW QUESTION 3
Which of the following is an example of a flat file?

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

Answer: A

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

NEW QUESTION 4
Which of the following is an example of a discrete data type?

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

Answer: B

Explanation:

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

NEW QUESTION 5
Jenny wants to study the academic performance of undergraduate sophomores and wants to determine the average grade point average at different points during an academic year.
What best describes the data set she needs?

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

Answer: A

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

NEW QUESTION 6
Given the table below:

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

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

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

Answer: C

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

NEW QUESTION 7
Analytics reports should follow corporate style guidelines.

A. True.
B. False.

Answer: A

NEW QUESTION 8
An analyst is building a new dashboard for a user. After an initial conversation with the user. the analyst created a mock-up of the dashboard. Which of the following best explains why the analyst created the mock-up?

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

Answer: C

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

NEW QUESTION 9
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 10
A data analyst needs to collect a similar proportion of data from every state. Which of the following sampling methods would be the most appropriate?

A. Systematic sampling
B. Convenience sampling
C. Stratified sampling
D. Random sampling

Answer: C

Explanation:
The best sampling method for the data analyst??s need is C. Stratified sampling.
Stratified sampling is a type of probability sampling that involves dividing the population into homogeneous groups or strata based on some characteristic, such as state, and then randomly selecting a proportional number of individuals from each stratum. Stratified sampling ensures that every group is adequately represented in the sample, and reduces the sampling error and variability12
Systematic sampling is not correct, because it involves selecting every nth individual from the population, starting from a random point. Systematic sampling does not guarantee that every state will have a similar proportion of data in the sample, and may introduce bias or error if there is a hidden pattern or order in the population12
Convenience sampling is not correct, because it involves selecting individuals who are easily accessible or available to the researcher. Convenience sampling is a type of non- probability sampling that does not involve random selection, and may result in a biased or unrepresentative sample12
Random sampling is not correct, because it involves selecting individuals from the population at random, without any grouping or stratification. Random sampling may not produce a sample that has a similar proportion of data from every state, especially if the population is large or heterogeneous. Random sampling may also have a higher sampling error and variability than stratified sampling12

NEW QUESTION 10
A database consists of one fact table that is composed of multiple dimensions. Each dimension is represented by a denormalized table. This structure is an example of a:

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

Answer: D

Explanation:
A star schema is a type of database schema that consists of one fact table and multiple dimension tables. The fact table contains the measures or metrics of the business process, such as sales, orders, or transactions. The dimension tables contain the attributes or characteristics of the business entities, such as products, customers, or locations. The fact table is connected to the dimension tables by foreign keys that reference the primary keys of the dimension tables. The fact table is located at the center of the schema, while the dimension tables are located at the edges, forming a star-like shape1.
A star schema is an example of a denormalized schema, which means that the dimension tables are not normalized and may contain redundant or repeated data. This is done to improve the performance and simplicity of queries, as there are fewer joins and tables involved. A star schema is suitable for data warehouses and business intelligence applications that require fast and efficient data retrieval2.

NEW QUESTION 14
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 19
Mario works with a group of R programmers tasked with copying data from an accounting system into a data warehouse.
In what phase are the group's R skills most relevant?

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

Answer: C

NEW QUESTION 22
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 27
An analyst develops an IT document and needs to describe the technical terms used in the document. Which of the following is where the analyst should include descriptions of the technical terms?

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

Answer: A

Explanation:
In technical documentation, a glossary is the designated section where definitions for technical terms are provided. It serves as a reference point for readers to understand specialized or uncommon words used within the document. Including descriptions of technical terms in a glossary ensures that readers have a consistent resource to refer to, which can improve comprehension and reduce misunderstandings12.
A system diagram (Option B) is a visual representation of the system??s components and their interactions, not a place for defining terms. User requirements (Option C) outline what end-users expect from the system, and an index (Option D) is an alphabetical list of topics covered in the document, usually with page numbers, but not definitions.
References:
? Creating effective technical documentation1.
? Best practices when writing technical descriptions3.

NEW QUESTION 30
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 32
Which of the following is a non-parametric test?

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

Answer: D

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

NEW QUESTION 33
Given the table below:

Transaction ID	Date	Year	Amount
XFW25091	10/1/2019	2019	$100.00
8741STKJG	5/3/2019	2019	$50.00
TIO335AL	8/15/2018	2018	$50.00
53KJNM1C	1/4/2020	2020	$250.00

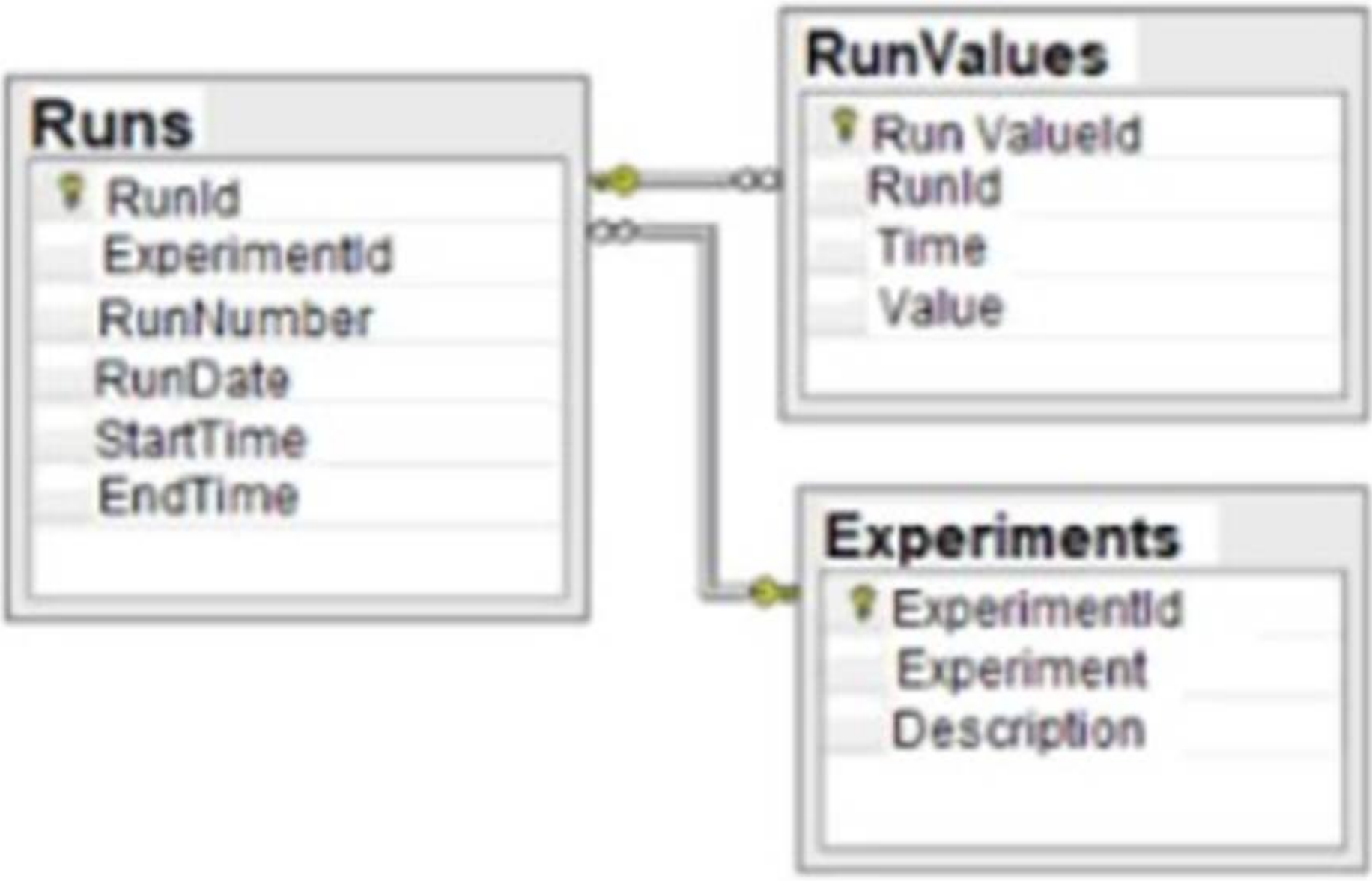
Which of the following variable types BEST describes the ??Year?? column?

A. Numeric
B. Date
C. Alphanumeric
D. Text

Answer: B

Explanation:
This is because date is a type of variable that represents a specific point or period in time, such as a day, a month, or a year. Date variables can be used to store, manipulate, or analyze temporal data, such as transaction dates, birth dates, or expiration dates. For example, date variables can be used to calculate the duration or the difference between two dates, or to filter or sort the data by date. The other variable types are not correct descriptions of the ??Year?? column. Here is why:
? Numeric is a type of variable that represents a numerical value, such as an integer, a decimal, or a fraction. Numeric variables can be used to store, manipulate, or analyze quantitative data, such as amounts, prices, or scores. For example, numeric variables can be used to perform arithmetic operations or calculations on the data, or to measure the central tendency or the dispersion of the data.
? Alphanumeric is a type of variable that represents a combination of alphabetic and
numeric characters, such as letters, numbers, symbols, or spaces. Alphanumeric variables can be used to store, manipulate, or analyze textual data, such as names, addresses, or codes. For example, alphanumeric variables can be used to concatenate or split the data, or to search or match the data using patterns or expressions.
? Text is a type of variable that represents a sequence of alphabetic characters,
such as letters or words. Text variables can be used to store, manipulate, or analyze textual data, such as names, categories, or labels. For example, text variables can be used to change the case or the length of the data, or to compare or classify the data using criteria or rules.

NEW QUESTION 34
Given the diagram below:



Which of the following data schemas shown?

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

Answer: D

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

NEW QUESTION 38
Which of the following report types is most appropriate for a high-level, year-end report requested by a Chief Executive Officer?

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

Answer: B

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

NEW QUESTION 42
Kelly wants to get feedback on the final draft of a strategic report that has taken her six months to develop.
What can she do to get prevent confusion as see seeks feedback before publishing the report?
Choose the best answer.

A. Distribute the report to the appropriate stakeholders via email.
B. Use a watermark to identify the report as a draft.
C. Show the report to her immediate supervisor.

D. Publish the report on an internally facing website.

Answer: B

Explanation:

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

NEW QUESTION 44
You should always choose the analytics tool that is most appropriate for any given situation, even if that means acquiring a new tool.

A. True.
B. False.

Answer: B

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

NEW QUESTION 48
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 51
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 53
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 56
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 60
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 65
Which of the following describes the use of a representative amount of data from a main repository?

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

Answer: D

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

NEW QUESTION 69
Which of the following are reasons to conduct data cleansing? (Select two).

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

Answer: CF

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

NEW QUESTION 74
Given the information in the following tables:

Online transactions:

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

In-store transactions:

Customer ID	Channel	Segment	Amount ($)
001	In-store	New	1,000
004	In-store	Existing	4,000
005	In-store	New	3,500

Which of the following describes merging these tables to create a master file that includes all transactions for both online and in-store sales?

A. Data audit
B. Data completeness
C. Data validation
D. Data consolidation

Answer: D

Explanation:
Merging tables to create a master file that includes all transactions for both online and in- store sales is best described as data consolidation. This process involves combining data from various sources into a single, unified dataset. Data consolidation is essential for providing a comprehensive view of all transactions, which can be used for analysis, reporting, and decision-making purposes.
References: The answer is based on standard data management practices and the definition of data consolidation. No specific external documents were referenced for this response.

NEW QUESTION 77
Which of the following is the best variable formal to store a customer's age using the least possible amount of storage data?

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

Answer: A

NEW QUESTION 79
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 83
Which one the following is not considered an aggregate function?

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

Answer: C

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

NEW QUESTION 85
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 90
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 93
Given the following graph:

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

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

Answer: B

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

NEW QUESTION 98
A data analyst needs to create a dashboard using the company's yearly revenue data sets. Which of the following would be the best way to plot the information to show the top- performing region?

A. A line chart
B. A waterfall chart
C. A heat map
D. A stacked bar chart

Answer: D

NEW QUESTION 103
An employer needs to maintain adequate office staffing during the winter and wants to track storm data. Which of the following data collection methods should the employer use?

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

Answer: B

Explanation:
For an employer looking to maintain adequate office staffing during winter while tracking storm data, the most effective method would be to use public databases. These databases often contain comprehensive records of weather patterns and storm data collected and verified by reputable meteorological organizations. Utilizing public databases allows for access to historical and real-time data that is crucial for making informed decisions about staffing during adverse weather conditions.
Web scraping (A) is not the most reliable method, as it may involve extracting data from various websites that might not always provide verified or consistent information. Observations © can be subjective and may not cover a wide enough area to be effective for decision-making on a larger scale. Weather surveys (D) could provide insights, but they are not as immediate or comprehensive as the data available in public databases. References:
? The systematic review on Big Data Analytics in Weather Forecasting suggests that
big data techniques and technologies can manage and analyze the huge volume of weather data from different resources, which supports the use of public databases1.
? NOAA??s approach to detecting severe weather events using instruments and
receiving information from storm spotters indicates the importance of reliable, collected data, which is typically stored in public databases2.
? The National Weather Service??s use of observational data collected by various
instruments, which are then fed into forecast models, further emphasizes the value of established data collection methods over individual observations or

surveys3.

NEW QUESTION 105
An analyst runs a report on a daily basis, and the number of datapoints must be validated before the data can be analyzed. The number of datapoints increases each day by approximately 20% of the total number from the day before. On a given day, the number of datapoints was 8,798. Which of the following should be the total number of datapoints on the next day?

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

Answer: C

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

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

Plugging in the given values, we get:

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

Next day = 8,798 * 1.2

Next day = 10,557.6
```

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

NEW QUESTION 109
Which of the following types of analysis is used when comparing last week's sales to the previous week's sales?

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

Answer: A

NEW QUESTION 111
An analyst needs to provide a chart to identify the composition between the categories of the survey response data set:

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

Which of the following charts would be BEST to use?

A. Histogram
B. Pie
C. Line
D. Scatter pot
E. Waterfall

Answer: B

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

NEW QUESTION 116
Each month an analyst needs to execute a data pull for the two prior months. Which of the following is the most efficient function for the analyst to use?

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

Answer: B

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

NEW QUESTION 118
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 119
A data analyst needs to write a SOL query measuring last month's website visits and distribute a summary report to the marketing team. Which of the following is the analyst creating?

A. Date range
B. Distribution list
C. Data content
D. Report view

Answer: D

NEW QUESTION 121
Given the following tables:

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

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

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

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

Answer: D

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

Name, Project_ID).
References:
? SQL documentation on FULL JOIN operations.

NEW QUESTION 126
Which of the following is an example of PII?

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

Answer: B

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

NEW QUESTION 131
Which of the following best describes a difference between JSON and XML?

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

Answer: A

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

NEW QUESTION 133
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 138
A data analyst is asked to create a sales report for the second-quarter 2020 board meeting, which will include a review of the business??s performance through the second quarter. The board meeting will be held on July 15, 2020, after the numbers are finalized. Which of the following report types should the data analyst create?

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

Answer: A

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

NEW QUESTION 139
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 142
An analyst is working with a data set that lists individuals' first and last names in separate columns. Which of the following processes should the analyst use to combine the first and last names into a single spreadsheet cell?

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

Answer: C

NEW QUESTION 147
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 150
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 152
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 156
Which of the following is a process that is used during data integration to collect, blend, and load data?

A. MDM
B. ETL
C. OLTP
D. BI

Answer: B

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

NEW QUESTION 159
Which of the following is the first step an analyst should perform upon receiving a business request for analysis?

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

Answer: C

Explanation:
Answer C. Review the business questions to understand the scope.
The first step an analyst should perform upon receiving a business request for analysis is to review the business questions to understand the scope of the problem, the objectives, and the expected outcomes. This will help the analyst to define the analytical approach, identify the data needs and sources, and plan the analysis process. Reviewing the business questions will also help the analyst to communicate with the stakeholders and clarify any assumptions or ambiguities1.
Option A is incorrect, as determining the data needs and sources for analysis is not the first step, but rather a subsequent step that depends on the business questions and the analytical approach.
Option B is incorrect, as initiating the analysis for exploratory data analysis is not the first step, but rather a part of the analysis process that involves examining and summarizing the data, identifying patterns and outliers, and testing hypotheses.
Option D is incorrect, as finalizing the methodology to solve the problem is not the first step, but rather a later step that involves selecting and applying the appropriate analytical techniques, tools, and models to answer the business questions.

NEW QUESTION 163
You are working with a dataset and need to swap the values in rows with those in columns. What action do you need to perform?

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

Answer: D

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

NEW QUESTION 164
Which one of the following is a measure of dispersion?

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

Answer: A

NEW QUESTION 165
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 166
Which of the following is a control measure for preventing a data breach?

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

Answer: D

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

NEW QUESTION 169
A salesperson who is prospecting potential clients collected the following data:

ID	Name	LName	Phone	Email
1	Jacob	Smith	(303)445-2323	jsmith@abc.com
2	Hans	Williams	(302)546-4588	hws@emc.com
3	Martha	Dion	(304)254-6575	dion@mail.com
4	Jules	Martin	(3OO)563-3435	jmartinxyz.com
5	Sabrina	Huggins	(323)655-3475	shug@emc.com

Which of the following is an issue with this data?

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

Answer: C

NEW QUESTION 170
A military commander would like to see the health scorecards of the troops daily and filter them based on gender and rank. Considering this data is PHI, which of the following would be the best way for the commander to view the information?

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

Answer: B

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

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

NEW QUESTION 172
A company notifies its employees that emails will be automatically moved to a cloud-based server in 180 days. Which of the following describes this concept?

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

Answer: C

NEW QUESTION 173
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 177
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 178
A company's human resources department has asked a data analyst to categorize the income of all employees into five salary bands:

Employee_ID	Salary	Salary_band
003	$130,000	
014	$120,000	
004	$110,000	
013	$90,000	
002	$140,000	
012	$122,000	
016	$132,000	
006	$70,000	
017	$53,000	
009	$111,000	
019	$107,000	
008	$111,000	
018	$50,000	

Which of the following types of functions would be the most appropriate to use?

A. Statistical
B. Aggregate
C. Logical
D. Mathematical

Answer: C

Explanation:
Short Explanation: Logical functions are the most appropriate to use for categorizing data
into bands, because they allow the data analyst to apply conditional statements and criteria to the data values. For example, the IF function can be used to assign a band name based on whether a value meets a certain condition or not. Other logical functions that can be useful for categorizing data are AND, OR, NOT, and IFERROR12

NEW QUESTION 179
Which of the following actions should be taken when transmitting data to mitigate the chance of a data leak occurring? (Choose two.)

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

Answer: DE

Explanation:

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

NEW QUESTION 182
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 185
Which of the ing is the correct ion for a tab-delimited spre file?

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

Answer: C

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

NEW QUESTION 189
A financial analyst is creating a daily billing report for a company. One night, the company's data warehouse did not update the data, which caused the data to be reported incorrectly the next day. Which of the following documentation elements should the analyst add to catch this error?

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

Answer: B

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

NEW QUESTION 194
Which of the following is the best technique for transferring data from one database to another with some data manipulation?

A. Application programming interfaces
B. Delta load
C. Extract, transform, load
D. Export/import

Answer: C

NEW QUESTION 197
Given the image below:

Which of the following file formats is depicted?

A. JSON
B. CSV
C. XML
D. HTML

Answer: A

Explanation:
The image depicts a snippet of code in the JSON format, which stands for JavaScript Object Notation. JSON is a lightweight data-interchange format that is easy for humans to read and write and easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language and is commonly used to transmit data in web
applications.
? CSV, or Comma-Separated Values, is a simple file format used to store tabular data, such as a spreadsheet or database. It uses commas to separate values.
? XML, or eXtensible Markup Language, is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.
? HTML, or HyperText Markup Language, is the standard markup language for documents designed to be displayed in a web browser.
References:
? JSON.org - Introducing JSON1
? W3Schools - JSON Introduction2
? Mozilla Developer Network - JSON3

NEW QUESTION 202
A data set for sales per month includes the following data:

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

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

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

Answer: B

NEW QUESTION 203
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 207
Amanda needs to create a dashboard that will draw information from many other data sources and present it to business leaders.
Which one of the following tools is least likely to meet her needs?

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

Answer: D

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

NEW QUESTION 212
Which of the following is the best description of the term "data governance"?

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

Answer: D

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

? Defining Data Governance and Its Business Value3.

NEW QUESTION 215
A development company is constructing a new unit in its apartment complex. The complex has the following floor plans:

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

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

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

Answer: C

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

```
Average cost per square foot = Total price / Total square feet
```

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

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

Average cost per square foot = $1,760,000 / 5,750

Average cost per square foot = $306
```

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

```
Price = Square feet * Average cost per square foot
```

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

```
Price = 2,300 * $306

Price = $705,200
```

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

NEW QUESTION 217
An e-commerce company recently tested a new website layout. The website was tested by a test group of customers, and an old website was presented to a control group. The table below shows the percentage of users in each group who made purchases on the websites:

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

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

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

Answer: A

Explanation:
The p-value is a measure of how likely it is to observe a difference in conversion rates as large or larger than the one observed, assuming that there is no difference between the groups. A common threshold for statistical significance is 0.05, meaning that there is a 5% or less chance of observing such a difference by chance alone. The table shows the p-values for each country, and we can see that only Germany has a p- value above 0.05 (0.13). This means that we cannot reject the null hypothesis that there is no difference in conversion rates between the test and control groups in Germany. Therefore, the increase in conversion from the new layout was not significant in Germany. For the other countries, the p-values are below 0.05, indicating that the increase in conversion from the new layout was statistically significant. Option A is correct.
Option B is incorrect because the increase in conversion from the new layout was
significant in France (p-value = 0.002).
Option C is incorrect because it does not account for the variation across countries. While the overall conversion rate for the test group (8.4%) is higher than the control group (6.8%), this difference may not be statistically significant when we consider the country-specific effects.
Option D is incorrect because the new layout has the highest conversion rate in the United Kingdom (9.6%), not the lowest.
References:
? P-value Calculator & Statistical Significance Calculator
? p-value Calculator | Formula | Interpretation
? How to obtain the P value from a confidence interval | The BMJ
? Confidence Intervals & P-values for Percent Change / Relative Difference

NEW QUESTION 221
Which of the following technologies would be best suited for creating a multiple linear regression model?

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

Answer: B

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

NEW QUESTION 223
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 224
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 226
A data analyst is asked on the morning of April 9, 2020, to create a sales report that identifies sales year to date. The daily sales data is current through the end of the day. Which of the following date ranges should be on the report?

A. January 1, 2020 to April 1, 2020
B. January 1, 2020 to April 7, 2020
C. January 1, 2020 to April 8, 2020
D. January 1, 2020 to April 9, 2020

Answer: D

Explanation:
This is because sales year to date refers to the sales that have occurred from the beginning of the current year until the current date. By creating a sales report that identifies sales year to date, the analyst can measure and compare the sales performance and progress of the current year. Since the analyst is asked to create the sales report on the morning of April 9, 2020, and the daily sales data is current through the end of the day, the date range that should be on the report is January 1, 2020 to April 9, 2020. The other date ranges are not correct for identifying sales year to date. Here is why:
? January 1, 2020 to April 1, 2020 would not include the sales that occurred in the first eight days of April, which would underestimate the sales year to date.
? January 1, 2020 to April 7, 2020 would not include the sales that occurred in the last two days of April, which would also underestimate the sales year to date.
? January 1, 2020 to April 8, 2020 would not include the sales that occurred on April 9, which would also underestimate the sales year to date.

NEW QUESTION 228
.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questons and Answers in PDF Format

DA0-001 Practice Exam Features:

* DA0-001 Questions and Answers Updated Frequently

* DA0-001 Practice Questions Verified by Expert Senior Certified Staff

* DA0-001 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* DA0-001 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year