CompTIA

Exam Questions DA0-001

CompTIA Data+ Certification Exam

NEW QUESTION 1
Taylor wants to investigate how manufacturing, marketing, and sales expenditures impact overall profitability for her company.
Which of the following systems is the most appropriate?

A. OLTP.
B. OLAP.
C. Data warehouse.
D. Data mart.

Answer: C

Explanation:

A Data mart is too narrow, because Taylor needs data from across multiple divisions. OLAP is a broad term for analytical processing, and OLTP systems are transactional and not ideal for the task. Since Taylor is working with data across multiple different divisions, she will work with a Data warehouse.

NEW QUESTION 2
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 3
A financial institution is reporting on sales performance to a company at the account level. Due to the sensitive nature of the government the does il with, some account information is not shown. Which of the following fields should be masked?

A. Sales volume
B. Start date
C. Product name
D. Customer name

Answer: D

Explanation:
Customer name is the field that should be masked, because it contains sensitive information that could identify the government accounts that the financial institution deals with. Masking is a technique that replaces or obscures sensitive data with dummy or random data, such as asterisks or hashes. Masking can help protect the privacy and security of the data, while still allowing for some analysis and reporting. Therefore, the correct answer is D. References: [Data Masking | Definition, Techniques & Examples - Talend], [Data masking - Wikipedia]

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

A. Rapid Miner.

B. QLIK.
C. Power BI.
D. Minitab.

Answer: D

Explanation:

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

NEW QUESTION 6
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 7
A data analyst received a large amount of third-party data that needs to be joined with in- house data files. After the data is joined, the analyst notices three columns all contain dates. Which of the following should the analyst do to maintain data consistency?

A. Append all date columns and parse the strings.
B. Impute all three date columns and then merge.
C. Merge all date columns and unify the format.
D. Separate the columns into a table and merge.

Answer: C

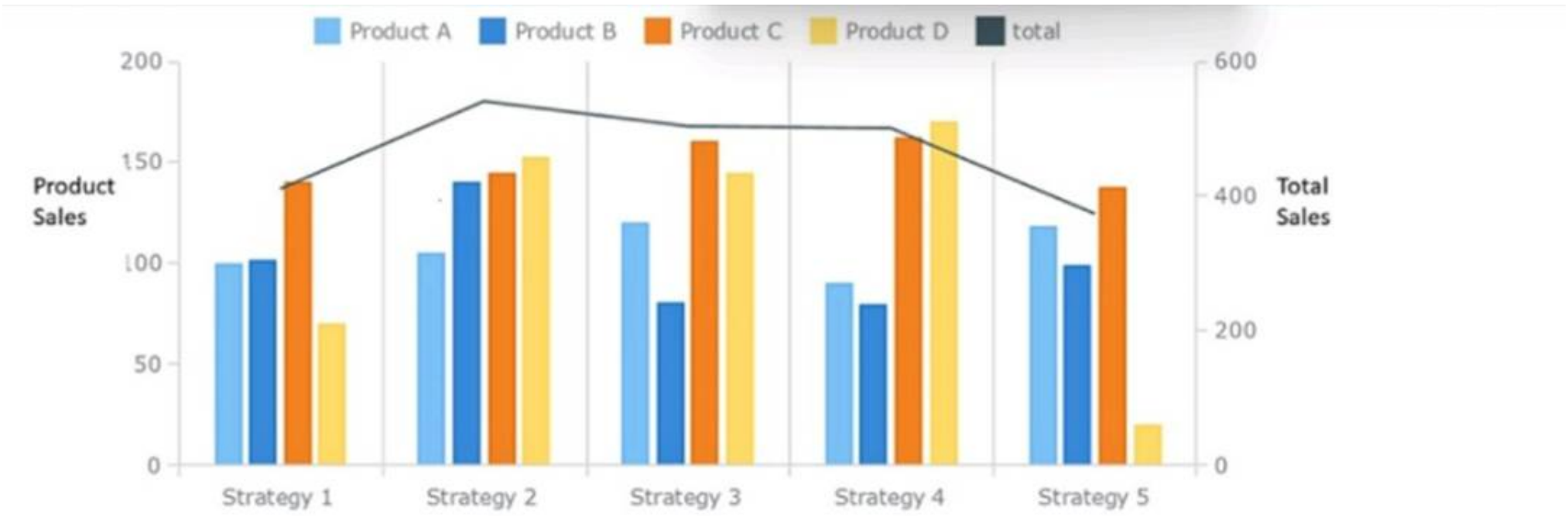
Explanation:
When dealing with multiple date columns from different data sources, it??s crucial to ensure consistency and accuracy in the dataset. The best practice is to merge the date columns and standardize the date format across the entire dataset. This approach helps maintain data integrity, simplifies analysis, and avoids confusion that could arise from having multiple date formats. Unifying the date format is particularly important when the data will be used for time series analysis or when dates are key to joining with other datasets.
References:
? Best practices in data merging emphasize the importance of a single point of reference and the need to avoid data loss or damage to individual data structures1.
? Power BI guides suggest that merging columns should be done carefully to maintain data integrity and avoid errors and inconsistencies2.
? Oracle Blogs highlight the need for a consistent number of columns among data sources when combining data with unions3.
? Excel tutorials recommend organizing data before merging and using formulas for complex merges4.
? An Excel guide on merging date and time columns advises employing functions to ensure seamless handling of non-date values5.

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

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

Answer: B

NEW QUESTION 9
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 10
A marketing analytics team received customer transaction data from two different sources. The data is complete and accurate; however, the field names appear to be inconsistent. Given the following tables:

Online transactions:

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

Store transactions:

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

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

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

Answer: A

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

References:
? Best practices in data management.
? Principles of data integration and consolidation.

NEW QUESTION 10
Which of the following is the correct data type for text?

A. Boolean
B. String
C. Integer
D. Float

Answer: B

Explanation:
A string is a data type that represents a sequence of characters, such as text, symbols, numbers, or punctuation marks. Strings are enclosed in quotation marks, such as ??Hello??, ??123??, or ??!@#??. Strings can be manipulated, concatenated, sliced, indexed, formatted, and searched using various methods and functions. A string is different from other data types, such as boolean, integer, or float, which represent logical values (true or false), whole numbers, or decimal numbers respectively. Therefore, the correct answer is B. References: What is a String? | Definition and Examples, Python String Methods

NEW QUESTION 12
Given the following data:

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

Which of the following BEST describes the data set?

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

Answer: C

Explanation:
This is because inconsistency is a type of data quality issue that occurs when the data does not follow a common format, structure, or rule across different sources or systems, which can affect the efficiency and performance of the analysis or process. Inconsistency can be caused by having different spellings, punctuations, capitalizations, or abbreviations for the same or similar values in a data set, such as ??M??, ??m??, ??Male??, or ??male?? for gender in this case. Inconsistency can be eliminated or reduced by using data cleansing techniques, such as standardizing or normalizing the data values. The other options are not correct descriptions of the data set. Here is why:
? Data bias is a type of data quality issue that occurs when the data is not representative or proportional of the population or the parameter, which can affect the validity and reliability of the analysis or process. Data bias can be caused by having a sample that is too small, too large, or too skewed for the population or the parameter, such as having only male customers for a product that targets both genders in this case. Data bias can be eliminated or reduced by using sampling techniques, such as stratified or cluster sampling.
? The data is incomplete is a type of data quality issue that occurs when the data is
absent or missing in a data set, which can affect the accuracy and reliability of the analysis or process. The data is incomplete can be caused by various factors, such as human error, system error, or non-response. The data is incomplete can be addressed by using various methods, such as replacing or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression.
? The data is outliers is a type of data quality issue that occurs when the data has
values that are unusually high or low compared to the rest of the data set, which can affect the quality and validity of the analysis or process. The data is outliers can be caused by various factors, such as measurement error, natural variation, or extreme events. The data is outliers can be addressed by using various methods, such as removing or filtering out the outliers, or using robust statistics that are less sensitive to outliers, such as median, interquartile range, or box plot.

NEW QUESTION 14
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 15
During data cleansing, an analyst conducts measures of central tendency on a data set. Which of the following data is the analyst attempting to identify?

A. Duplicate
B. Missing
C. Outlying
D. Invalid

Answer: C

NEW QUESTION 17
A JSON file is an example of:

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

Answer: A

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

NEW QUESTION 19
Which of following is a non-relational database?

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

Answer: A

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

NEW QUESTION 24
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 29
A data analyst has a set with more than 40.000 rows in the sample schema below:

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

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

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

Answer: D

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

NEW QUESTION 34
An analyst is working on a project for a director. During this process. the analyst pulled the data. created summarized tables and graphs with descriptions, created a report summary, and inserted all items into a report. After writing the report, which of the following would be the most appropriate next step?

A. Complete an audit on the data pulled for the report.
B. Complete a check for quality in the report.
C. Complete a review of the data and a check for consistency
D. Complete a trend analysis to be included in the report.

Answer: B

Explanation:
After writing the report, the most appropriate next step for the analyst is to complete a check for quality in the report. This involves reviewing the report for accuracy, clarity, completeness, consistency, and relevance. The analyst should ensure that the report addresses the director??s business questions and objectives, that the data and analysis are correct and reliable, that the tables and graphs are well-designed and easy to understand, that the descriptions and summary are concise and informative, and that there are no errors or inconsistencies in the report. A quality check will help the analyst to improve the presentation and communication of the report, as well as to avoid any misunderstandings or misinterpretations by the director1.

NEW QUESTION 39
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 40
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 44
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

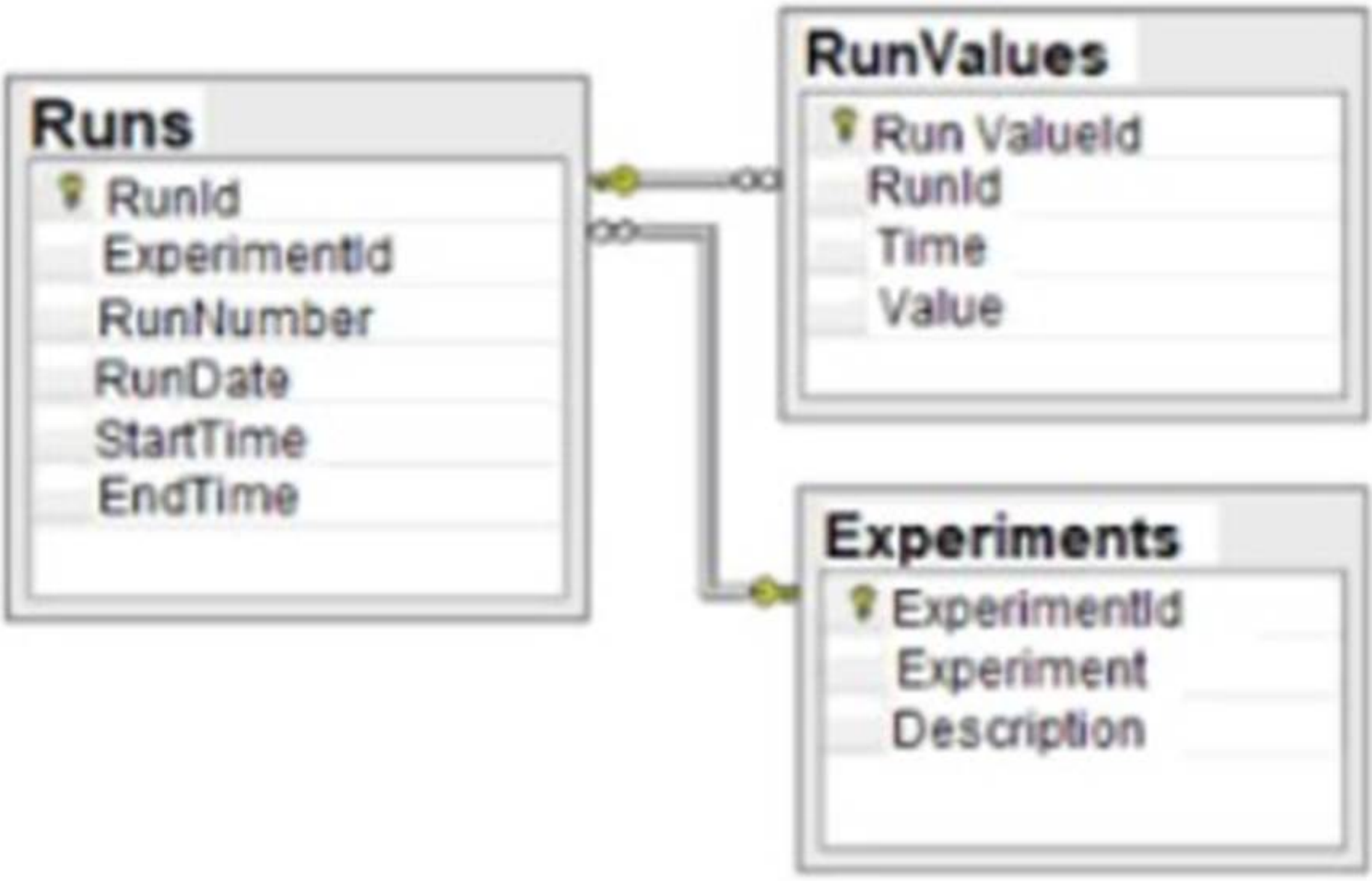
Which of the following variable types BEST describes the ??Year?? column?

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 46
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 48
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 50
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 55
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 56
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 60
A data scientist wants to see which products make the most money and which products attract the most customer purchasing interest in their company.
Which of the following data manipulation techniques would he use to obtain this information?

A. Data append
B. Data blending
C. Normalize data
D. Data merge

Answer: B

Explanation:

The correct answer is B: Data blending.
Data blending is combining multiple data sources to create a single, new dataset, which can be presented visually in a dashboard or other visualization and can then be processed or analyzed. Enterprises get their data from a variety of sources, and users may want to temporarily bring together different datasets to compare data relationships or answer a specific question. Data append is incorrect. Data append is a process that involves adding new data elements to an existing database. An example of a common data append would be the enhancement of a company's customer files. A data append takes the information they have, matches it against a larger database of business data, allowing the desired missing data fields to be added. Normalize data is incorrect.
Data normalization is the process of structuring your relational customer database, following a series of normal forms. This improves the accuracy and integrity of your data while ensuring that your database is easier to navigate. Data merge is incorrect. Data merging is the process of combining two or more data sets into a single data set.

NEW QUESTION 65
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 67
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 69
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 74
A client has requested an analysis of all pet care items purchased by current customers and their social media connections in the past 12 months. Which of the following data analysis techniques would be the best choice given these requirements?

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

Answer: C

NEW QUESTION 78
While reviewing survey data, a research analyst notices data is missing from all the responses to a single question. Which of the following methods would BEST address this issue?

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

Answer: A

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

? Remove duplicate data is a type of method that eliminates or reduces duplicate data, which is a type of data quality issue that occurs when data is repeated or copied in a data set. Removing duplicate data does not address missing data, but rather affects the quantity and validity of the data.
? Replace redundant data is a type of method that eliminates or reduces redundant data, which is a type of data quality issue that occurs when data is unnecessary or irrelevant for the analysis or purpose. Replacing redundant data does not address missing data, but rather affects the efficiency and performance of the analysis or process.
? Remove invalid data is a type of method that eliminates or reduces invalid data, which is a type of data quality issue that occurs when data is incorrect or inaccurate in a data set. Removing invalid data does not address missing data, but rather affects the validity and reliability of the analysis or process.

NEW QUESTION 79
An analyst is required to run a text analysis of data that is found in articles from a digital news outlet. Which of the following would be the BEST technique for the analyst to apply to acquire the data?

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

Answer: A

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

NEW QUESTION 80
??Which of the following is the BEST reason to use database views instead of tables?

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

Answer: A

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

NEW QUESTION 81
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 82
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 85
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 86
Which one of the following in NOT a common data integration tool?

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

Answer: A

Explanation:

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

NEW QUESTION 87
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 90
Consider this dataset showing the retirement age of 11 people, in whole years: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60
This tables show a simple frequency distribution of the retirement age data.

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

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

Answer: D

Explanation:

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

NEW QUESTION 95
Which of the following describes the method of sampling in which elements of data are selected randomly from each of the small subgroups within a population?

A. Simple random
B. Cluster
C. Systematic
D. Stratified

Answer: D

Explanation:
This is because stratified is a type of sampling in which elements of data are selected randomly from each of the small subgroups within a population, such as age groups, gender groups, or income groups. Stratified sampling can be used to ensure that the sample is representative and proportional of the population, as well as reduce the sampling error or bias. For example, stratified sampling can be used to select a sample of voters from different political parties based on their proportion in the population. The other types of sampling are not the types of sampling in which elements of data are selected randomly from each of the small subgroups within a population. Here is why:
? Simple random is a type of sampling in which elements of data are selected randomly from the entire population, without dividing it into any subgroups. Simple random sampling can be used to ensure that every element in the population has an equal chance of being selected, as well as avoid any systematic error or bias. For example, simple random sampling can be used to select a sample of students from a school by using a lottery or a computer-generated number.
? Cluster is a type of sampling in which elements of data are selected randomly from
a few large subgroups within a population, such as regions, districts, or schools. Cluster sampling can be used to reduce the cost and complexity of sampling, as well as increase the feasibility and convenience of sampling. For example, cluster sampling can be used to select a sample of households from a few neighborhoods by using a map or a list.
? Systematic is a type of sampling in which elements of data are selected at regular
intervals from an ordered list or sequence within a population, such as every nth element or every kth element. Systematic sampling can be used to simplify and speed up the sampling process, as well as ensure that the sample covers the entire range or scope of the population. For example, systematic sampling can be used to select a sample of books from a library by using an alphabetical order or a numerical order.

NEW QUESTION 99
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 102
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 106
An analyst is reporting on the average income for a county and is reviewing the following data:

Name	Address	Yearly income
Jessica Jones	145 Stonebridge Avenue	$634,900
Spencer James	1567 Watercress	$135,000
Olivia Baker	456 Harvard Road	$95,000
Layla Harding	5674 Yarding Street	$37,000

Which of the following is the reason the analyst would need to cleanse the data in this data set?

A. Data completeness
B. Data outliers
C. Duplicate data
D. Missing values

Answer: B

NEW QUESTION 107
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 109
A data analyst must fulfill a request for information that is needed weekly and should be automatically emailed to a specific set of users. Which of the following types of reports should the analyst recommend?

A. A self-service report
B. A research report
C. An ad hoc report
D. An operational report

Answer: D

Explanation:
An operational report is the most suitable type of report for information that needs to be sent out on a regular, scheduled basis, such as weekly. Operational reports are designed to provide ongoing insights into the performance of an organization??s operations and are typically automated to be distributed at set

intervals. This automation can include scheduling the reports to be emailed to a specific list of recipients, making it an efficient solution for the analyst??s requirement.
Operational reports are often generated from data that is continuously updated, ensuring that the recipients receive the most current information at the time of the report??s distribution. This contrasts with ad hoc reports ©, which are usually created as needed and are not scheduled. Self-service reports (A) require users to generate the report themselves, which is not the requirement here. Research reports (B) are generally more detailed and are not typically used for regular operational updates.
References:
? The guidelines on writing email reports suggest that for regular, scheduled information dissemination, structured reports like operational reports are preferred1.
? Best practices in reporting also recommend automated and scheduled reports for consistent and timely updates, which operational reports provide2.

NEW QUESTION 111
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 114
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 116
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 118
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 122
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 127
Under which of the following circumstances should the null hypothesis be accepted when a
= 0.05?

A. When p is 0.00003
B. When p is 0.001
C. When p is 0.04
D. When p is 0.06

Answer: C

Explanation:
The null hypothesis should be accepted when the p-value is greater than the alpha level, which is the significance level of the test. The p-value is the probability of obtaining a test statistic at least as extreme as the one observed in the sample, assuming that the null hypothesis is true. The alpha level is the probability of rejecting the null hypothesis when it is true, which is also known as a type I error12.
In this case, the alpha level is 0.05, which means that there is a 5% chance of rejecting the null hypothesis when it is true. Therefore, to reject the null hypothesis, the p-value must be less than or equal to 0.05, which indicates that the test statistic is very unlikely to occur by chance under the null hypothesis. Conversely, to accept the null hypothesis, the p-value must be greater than 0.05, which indicates that the test statistic is not very unlikely to occur by chance under the null hypothesis.
Among the four options, only option D has a p-value that is greater than 0.05 (p = 0.06). Therefore, option D is the correct answer. When p = 0.06, it means that there is a 6% chance of obtaining a test statistic at least as extreme as the one observed in the sample, assuming that the null hypothesis is true. This probability is not very low, and therefore does not provide enough evidence to reject the null hypothesis.

NEW QUESTION 128
Which of the following roles is responsible for ensuring an organization's data quality, security, privacy, and regulatory compliance?

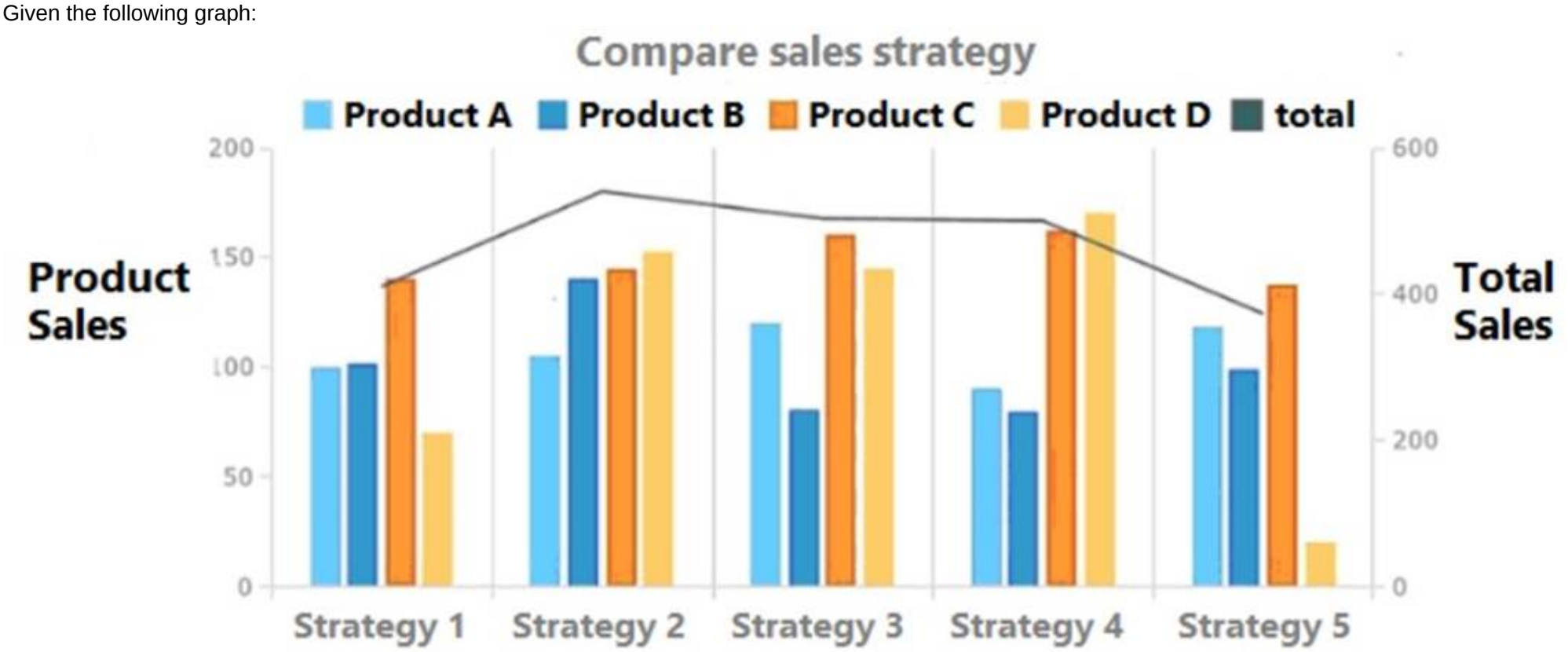
A. Data owner.

B. Data steward.
C. Data custodian.
D. Data processor.

Answer: B

Explanation:
Correct answer B. Data steward.
A data steward is responsible for leading an organization's data governance activities, which include data quality, security, privacy, and regulatory compliance.

NEW QUESTION 133
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 138
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 139
You would like to measure how well an organization is achieving its goals. What type of analysis should you perform?

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

Answer: A

Explanation:

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

NEW QUESTION 143
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 148
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 149
An analyst must obtain the average daily sales for the following week:

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

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

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

Answer: C

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

NEW QUESTION 152
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

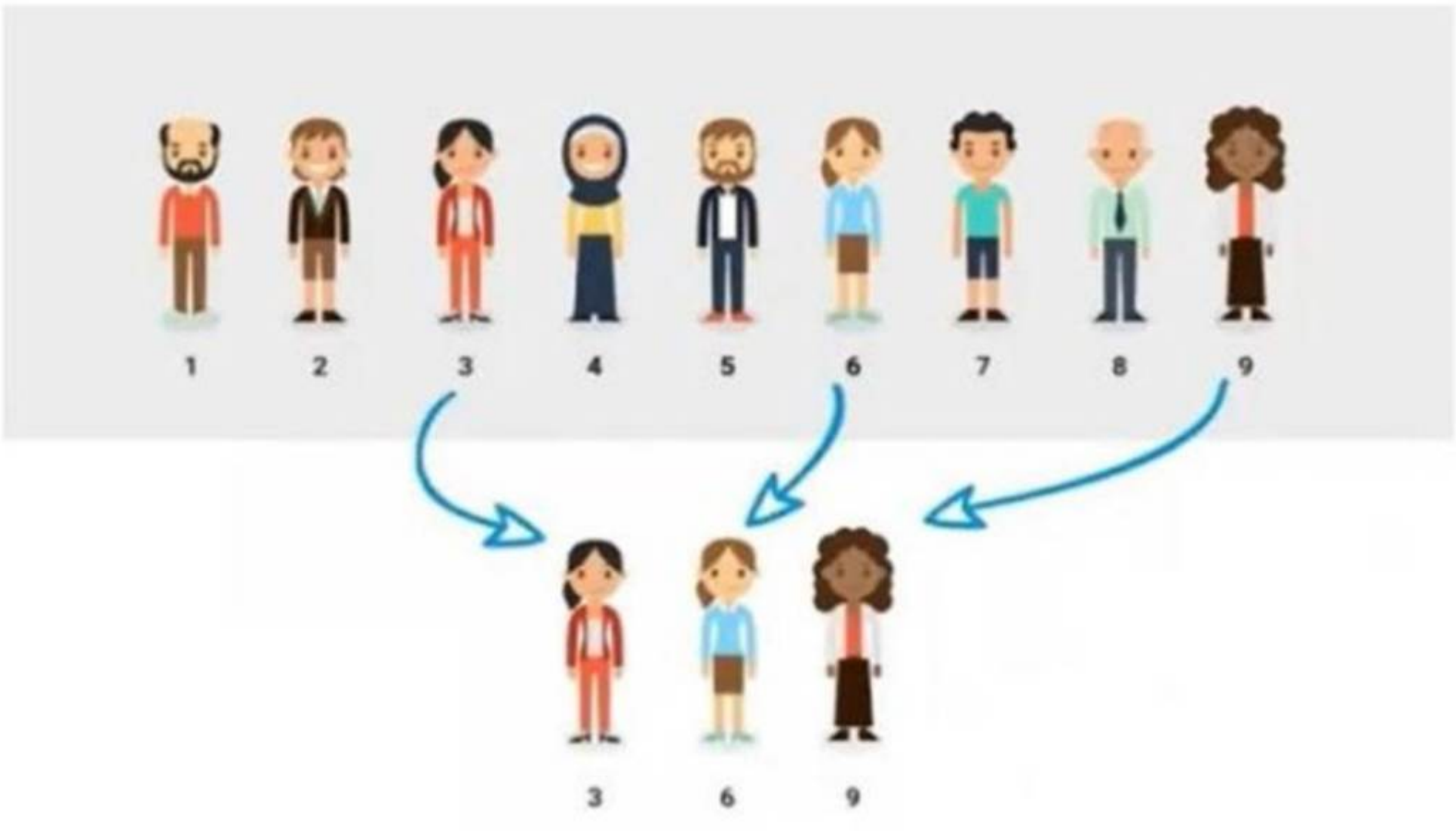
Which of the following charts would be BEST to use?

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 153
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 155
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 156
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 159
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 162
A data analyst has been asked to merge the tables below, first performing an INNER JOIN and then a LEFT JOIN:

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

Customer Table -
In-store Transactions –

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

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

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

Answer: C

Explanation:
An INNER JOIN returns only the rows that match the join condition in both tables. A LEFT JOIN returns all the rows from the left table, and the matched rows from the right table, or NULL if there is no match. In this case, the customer table is the left table and the in-store transactions table is the right table. The join condition is based on the customer_id column, which is common in both tables.
To perform an INNER JOIN, we can use the following SQL query:
SELECT * FROM customer INNER JOIN in_store_transactions ON customer.customer_id
= in_store_transactions.customer_id;
This query will return 9 rows of data, as shown below:
customer_id | name | lastname | gender | marital_status | transaction_id | amount | date 1 | MARC | TESCO | M | Y | 1 | 1000 | 2020-01-01 1 | MARC | TESCO | M | Y | 2 | 5000 | 2020-01-02 2 | ANNA | MARTIN | F | N | 3 | 2000 | 2020-01-03 2 | ANNA | MARTIN | F | N |

4 | 3000 | 2020-01-04 3 | EMMA | JOHNSON | F | Y | 5 | 4000 | 2020-01-05 4 | DARIO | PENTAL | M | N | 6 | 5000 | 2020-01-06 5 | ELENA | SIMSON| F| N|7|6000|2020-01-07 6|TIM|ROBITH|M|N|8|7000|2020-01-08 7|MILA|MORRIS|F|N|9|8000|2020-01-09
To perform a LEFT JOIN, we can use the following SQL query:
SELECT * FROM customer LEFT JOIN in_store_transactions ON customer.customer_id = in_store_transactions.customer_id;
This query will return 15 rows of data, as shown below: customer_id|name|lastname|gender|marital_status|transaction_id|amount|date
1|MARC|TESCO|M|Y|1|1000|2020-01-01 1|MARC|TESCO|M|Y|2|5000|2020-01-02
2|ANNA|MARTIN|F|N|3|2000|2020-01-03 2|ANNA|MARTIN|F|N|4|3000|2020-01-04
3|EMMA|JOHNSON|F|Y|5|4000|2020-01-05 4|DARIO|PENTAL|M|N|6|5000|2020-01-06
5|ELENA|SIMSON||F||N||7||6000||2020-01-07 6||TIM||ROBITH||M||N||8||7000||2020-01-08
7||MILA||MORRIS||F||N||9||8000||2020-01-09
8||JENNY||DWARTH||F||Y||NULL||NULL||NULL
As you can see, the customers who do not have any transactions (customer_id = 8) are still included in the result, but with NULL values for the transaction_id, amount, and date columns.
Therefore, the correct answer is C: INNER: 9 rows; LEFT: 15 rows. Reference: SQL Joins - W3Schools

NEW QUESTION 166
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 168
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 169
Given the following grocery store orders:

Order_ID	Order_total
85495	$132.49
28597	$108.99
57490	$96.19
35806	$74.49
18014	$178.59
39725	$41.99
20935	$136.99
25402	$31.29
85023	$24.49
27933	$76.99

If a query is made to the table with the following logic: Order_Total > 132 OR (Order Total >= 25 AND Order_Total < 74)
Which of the following is the number of orders that will be returned by the query?

A. Four
B. Five
C. Six
D. Seven

Answer: C

Explanation:
Based on the query logic provided: Order_Total > 132 OR (Order Total >= 25 AND Order_Total < 74), we can manually determine which order totals fit this criteria.
By examining the image, these are the Order_Total values that match:
? 132.49 (greater than 132)
? 108.99 (greater than or equal to 25 and less than 74)
? 96.19 (greater than or equal to 25 and less than 74)
? 74.49 (greater than or equal to 25 and less than 74)
? 41.99 (greater than or equal to 25 and less than 74)
? 31.29 (greater than or equal to 25 and less than 74) Thus, six orders satisfy the given conditions.

NEW QUESTION 172
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 177
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 180
An analyst needs to determine the appropriate data type for the following sample data: sample data collected:
Which of the following data types should be used for this data?

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

Answer: B

NEW QUESTION 182
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 183
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 186
A database consists of one fact table that is composed of multiple dimensions. Depending on the dimension, each one can be represented by a denormalized table or multiple normalized tables. This structure is an example of a:

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

Answer: B

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

NEW QUESTION 188
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 189
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 194
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 199
Given the following data table:

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

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

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

Answer: BD

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

NEW QUESTION 201
Which of the following best describes a business analytics tool with interactive visualization and business capabilities and an interface that is simple enough for end users to create their own reports and dashboards?

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

Answer: C

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

NEW QUESTION 206
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 209
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 211
A reporting analyst is creating a dashboard that shows the year-over-year performance for a sales organization. Which of the following is the best visual for the analyst use to illustrate the organization's performance?

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

Answer: D

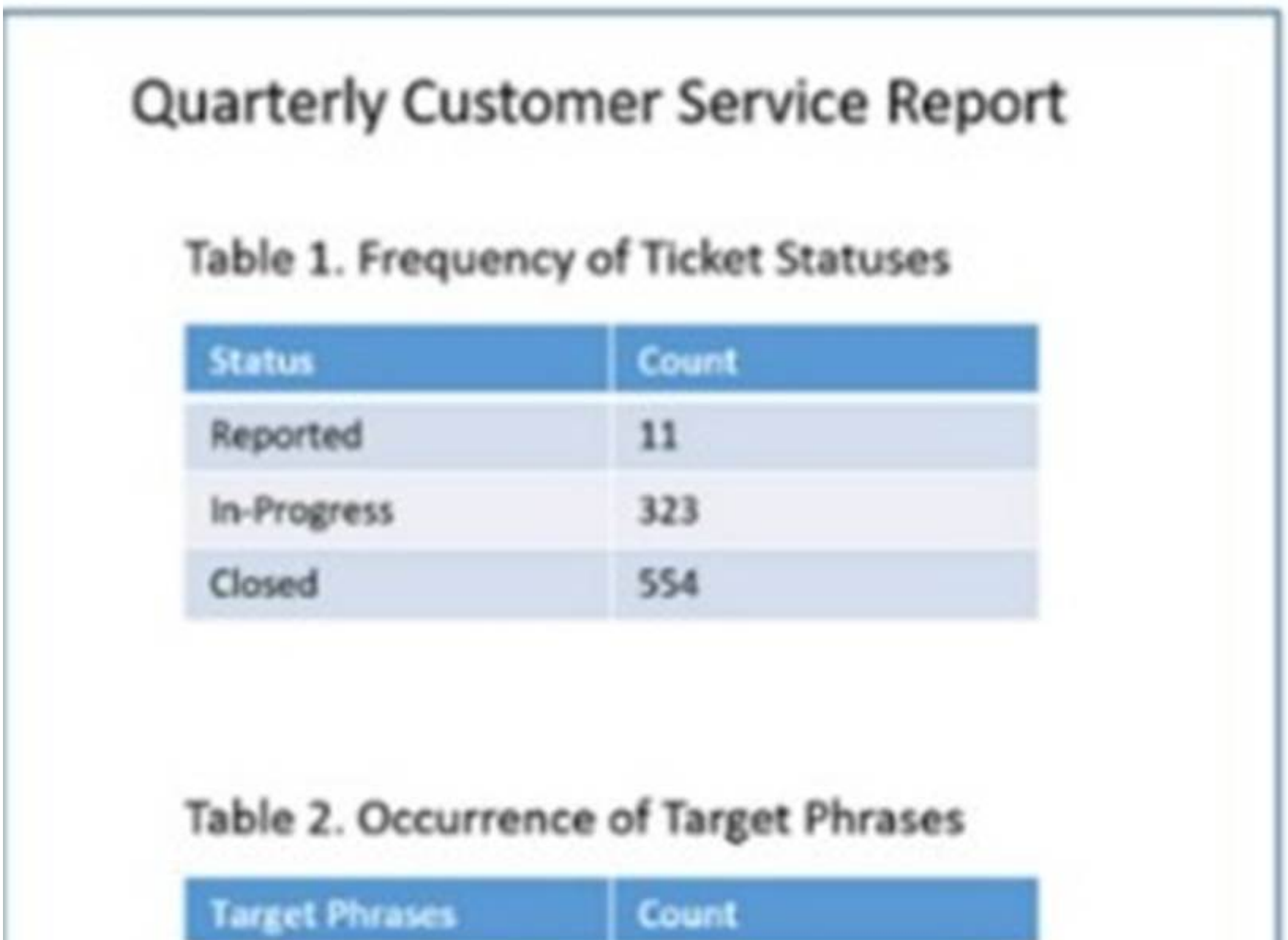
NEW QUESTION 214
Different people manually type a series of handwritten surveys into an online database. Which of the following issues will MOST likely arise with this data? (Choose two.)

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

Answer: AE

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

NEW QUESTION 215
Given the following report:



Quarterly Customer Service Report

Table 1. Frequency of Ticket Statuses

Status	Count
Reported	11
In-Progress	323
Closed	554

Table 2. Occurrence of Target Phrases

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

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

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

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

Answer: DF

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

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

NEW QUESTION 216
Samantha needs to share a list of her organization's top 50 customers with the VP of sales.
She would like to include the name of the customer, the business they represent, their contact information, and their total sales over the past year.
The VP does not have any specialized analytics skills or software but would like to make some personal notes on the dataset.
What would be the best tool for Samantha to use to share this information?

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

Answer: B

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

NEW QUESTION 217
An analyst is currently working on a ticket for revamping a company-wide dashboard that has been in use for five years. Which of the following should be the first step in the development process?

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

Answer: A

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

NEW QUESTION 222
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 226
Which of the following BEST describes the issue in which character values are mixed with integer values in a data set column?

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

Answer: D

Explanation:
The invalid data type is the best description for the issue in which character values are mixed with integer values in a data set column. Invalid data type means that the data does not match the expected or required format or structure for a given variable or attribute. For example, if a column is supposed to store numerical values, but some rows contain text values, then those rows have an invalid data type. References: CompTIA Data+ Certification Exam Objectives, page 10

NEW QUESTION 228
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 231
Randy scored 76 on a math test, Katie scored 86 on a science test, Ralph scored 80 on a history test, and Jean scored 80 on an English test. The table below contains the mean and standard deviation of the scores for each of the courses:

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

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

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

Answer: B

Explanation:
To compare the students?? scores, we need to standardize them by using the z-score formula, which is:
z = (x -) /
where x is the raw score, is the mean, and is the standard deviation. The z-score tells us how many standard deviations a score is above or below the mean. A higher z-score means a better score relative to the average.
Using the table, we can calculate the z-scores for each student as follows:
Randy: z = (76 - 70) / 2 = 3 Katie: z = (86 - 80) / 3 = 2 Ralph: z = (80 - 75) / 2 = 2.5 Jean: z
= (80 - 90) / 1 = -10
The student with the highest z-score is Randy, with a z-score of 3. This means that Randy scored 3 standard deviations above the mean in math, which is the best performance among the four students. Therefore, the correct answer is A.
References: Comparing with z-scores (video) | Z-scores | Khan Academy, 17 Important Data Visualization Techniques | HBS Online

NEW QUESTION 236
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 241
A cereal manufacturer wants to determine whether the sugar content of its cereal has increased over the years. Which of the following is the appropriate descriptive statistic to use?

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

Answer: B

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

NEW QUESTION 246
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:
A summary with statistics, conclusions, and recommendations from the data analyst is the best way to communicate the results of an online marketing campaign to the marketing manager. A summary can provide a concise and clear overview of the most important KPIs and measure the return on marketing investment, as well as highlight the main findings and insights from the data analysis. A summary can also include actionable suggestions and best practices for improving the campaign performance and achieving the marketing objectives. A summary is different from other options, such as a real-time monitor, a self-service dashboard, or a spreadsheet of raw data, which may not provide enough context, interpretation, or guidance for the manager. Therefore, the correct answer is D. References: How to Write a Data Analysis Report: 6 Essential Tips, How to Write a Marketing Report (with Pictures) - wikiHow

NEW QUESTION 251
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 256
A data analyst needs to create a weekly recurring report on sales performance and distribute it to all sales managers. Which of the following would be the BEST method to automate and ensure successful delivery for this task?

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

Answer: A

Explanation:
Scheduled report delivery is a feature that allows a data analyst to automate the generation and distribution of a report at a specified time and frequency. This would be the best method to ensure that the sales managers receive the weekly report on sales performance without manual intervention. Subscription access

delivery is a feature that allows users to subscribe to a report and access it on demand, but it does not automate the delivery. Printing out a copy or uploading the report to the server are manual methods that require more time and effort from the data analyst. Reference: CertMaster Practice for Data+ Exam Prep - CompTIA

NEW QUESTION 257
An analyst is designing a dashboard that will provide a story of the sales and sales customer ratio. The following data is available:

Site	Customers	New customers	Percentage of new customers	Sales volume	Average sales per customer
A1	2236	277	12%	$3,415,372.00	$1,527.45
A2	885	300	34%	$1,405,437.00	$1,588.06
A3	333	200	60%	$952,723.00	$2,861.03
B1	483	167	35%	$4,871,380.00	$10,085.67
B2	2969	235	8%	$780,381.00	$262.84
B3	2357	153	6%	$4,917,436.00	$2,086.31
C1	1524	180	12%	$1,135,204.00	$744.88
C2	878	150	17%	$614,964.00	$700.41
C2	1925	142	7%	$4,035,100.00	$2,096.16

Which of the following charts should the analyst consider including in the dashboard?

A. A column chart with site and sales
B. A line chart with site and sales
C. A pie chart with site and sales
D. A scatter chart with site and sales

Answer: A

Explanation:
For a dashboard that aims to tell a story about sales and the sales customer ratio, a column chart is an effective choice. Column charts are particularly useful for showing data changes over a period of time or for illustrating comparisons among items. In this case, a column chart can clearly display the sales figures for each site, allowing for easy comparison across different sites. Additionally, it can be used to represent the sales customer ratio by showing the proportion of sales per customer, which can provide insights into customer behavior and sales effectiveness.
? Line charts are best suited for displaying data trends over time, rather than for comparing individual categories.
? Pie charts could show the proportion of sales for each site, but they are not as effective as column charts for comparing multiple categories.
? Scatter charts are used to show the relationship between two variables, which is not the focus in this scenario.
References:
? Effective Use of Column Charts1
? Choosing the Right Chart for Your Data2
? Sales Dashboards: Examples & Templates3

NEW QUESTION 262
An analyst is updating a customer contacts database with information obtained from a survey of new customers. Which of the following data manipulation techniques should the analyst use?

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

Answer: B

NEW QUESTION 264
Joe. an analyst. tests the loading time on a dashboard he is preparing to go live and finds it is slower than he would like. Which of the following must occur to decrease the loading time?

A. Deploy the dashboard to production.
B. Change the field definitions.
C. Update the dashboard subscribers.
D. Optimize the dashboard.

Answer: D

Explanation:
Optimizing the dashboard is the process of improving its performance and reducing its loading time by applying various techniques and best practices. Some of the common ways to optimize a dashboard are:
? Reducing the size and complexity of the data model, such as removing
unnecessary columns, aggregating data at the source, or using data compression techniques12
? Leveraging caching strategies, such as setting appropriate cache refresh intervals
or utilizing Power BI??s built-in caching mechanisms, to minimize data retrieval delays2
? Utilizing query folding, direct query, or live connection to enhance data processing

efficiency and enable real-time data updates23
? Optimizing DAX queries, such as avoiding nested calculations, using variables, or simplifying measures, to improve data calculation speed23
? Reducing visualizations and calculations, such as using fewer or simpler charts, filters, or parameters, to speed up dashboard rendering12
? Evaluating the impact of custom visuals on dashboard load time and avoiding or replacing those that are slow or inefficient2
? Applying aggregation and summarization techniques, such as using extract filters, context filters, or level of detail expressions, to reduce the amount of data displayed on the dashboard1
? Troubleshooting and resolving any issues that may cause slow dashboard load, such as network latency, server overload, or hardware limitations24

NEW QUESTION 265
An analyst needs to join two tables of data together for analysis. All the names and cities in the first table should be joined with the corresponding ages in the second table, if applicable.

Table 1

Name	City
Jane Smith	Detroit
John Smith	Dallas
Candace Johnson	Atlanta
Kyle Jacobs	Chicago

Table 2

Name	Age
John Smith	34
John Smith	56
Candace Johnson	45
Kyle Jacobs	39

Which of the following is the correct join the analyst should complete. and how many total rows will be in one table?

A. INNER JOIN, two rows
B. LEFT JOI
C. four rows
D. RIGHT JOI
E. five rows
F. OUTER JOIN, seven rows

Answer: B

Explanation:
The correct join the analyst should complete is B. LEFT JOIN, four rows.
A LEFT JOIN is a type of SQL join that returns all the rows from the left table, and the matched rows from the right table. If there is no match, the right table will have null values. A LEFT JOIN is useful when we want to preserve the data from the left table, even if there is no corresponding data in the right table1
Using the example tables, a LEFT JOIN query would look like this:
SELECT t1.Name, t1.City, t2.Age FROM Table1 t1 LEFT JOIN Table2 t2 ON t1.Name = t2.Name;
The result of this query would be:
Name City Age Jane Smith Detroit NULL John Smith Dallas 34 Candace Johnson Atlanta 45 Kyle Jacobs Chicago 39
As you can see, the query returns four rows, one for each name in Table1. The name John Smith appears twice in Table2, but only one of them is matched with the name in Table1. The name Jane Smith does not appear in Table2, so the age column has a null value for that row.

NEW QUESTION 266
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 271
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 273
.....

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