

# CompTIA

## Exam Questions DA0-001

CompTIA Data+ Certification Exam



**NEW QUESTION 1**

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

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

**Answer:** D

**Explanation:**

This is because redundant data is a type of data that is unnecessary or irrelevant for the analysis or purpose, which can affect the efficiency and performance of the analysis or process. Redundant data can be caused by having multiple data fields that store the same or similar information, such as patient height in inches and patient height in centimeters in this case. Redundant data can be eliminated or reduced by using data cleansing techniques, such as removing or merging the redundant data fields. The other types of data are not examples of data that is unnecessary or irrelevant for the analysis or purpose. Here is what they mean in terms of data quality:

? Dependent data is a type of data that relies on or is influenced by another data field or value, such as a formula or a calculation that uses other data fields or values as inputs or outputs. Dependent data can be useful or important for the analysis or purpose, as it can provide additional information or insights based on the existing data.

? Duplicate data is a type of data that is repeated or copied in a data set, which can affect the quality and validity of the analysis or process. Duplicate data can be caused by having multiple records or rows that have the same or similar values for one or more data fields or columns, such as customer ID or order ID. Duplicate data can be eliminated or reduced by using data cleansing techniques, such as removing or filtering out the duplicate records or rows.

? Invalid data is a type of data that is incorrect or inaccurate in a data set, which can affect the validity and reliability of the analysis or process. Invalid data can be caused by having values that do not match the expected format, type, range, or rule for a data field or column, such as an email address that does not have an @ symbol or a date that does not follow the YYYY-MM-DD format. Invalid data can be eliminated or reduced by using data cleansing techniques, such as validating or correcting the invalid values.

**NEW QUESTION 2**

A 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 3**

Five dogs have the following heights in millimeters: 300, 430, 170, 470, 600

Which of the following is the mean height for the five dogs?

- A. 394mm
- B. 405mm
- C. 493mm
- D. 504mm

**Answer:** A

**Explanation:**

The mean height for the five dogs is calculated by adding up all the heights and dividing by the number of dogs. The formula is:

$$\text{mean} = (300 + 430 + 170 + 470 + 600) / 5 \quad \text{mean} = 1970 / 5 \quad \text{mean} = 394$$

Therefore, option A is correct.

Option B is incorrect because it is the median height, which is the middle value when the heights are arranged in ascending order.

Option C is incorrect because it is the mean height multiplied by 1.25.

Option D is incorrect because it is the mean height multiplied by 1.28.

**NEW QUESTION 4**

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

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

**Answer:** D

**Explanation:**

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

Creating a dashboard for sales numbers, pipeline, and team and individual performance for the management team would not be appropriate to provide visibility for the sales team, as they would not have access to the dashboard or the information they need. The dashboard would only be available for the management team, which would limit the transparency and collaboration among the sales team members.

Creating a dashboard with filters for the overall team, individuals, and management would not be secure to provide visibility of confidential information, such as individuals' earned commissions and projected commissions based on sales. The dashboard would allow users to filter and see the data they want, which could expose sensitive or personal information to unauthorized users.

**NEW QUESTION 5**

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 6**

Which of the following programming languages are best suited for analysis and machine- learning applications? (Select two).

- A. Ruby
- B. Rust
- C. PHP
- D. Python
- E. Kotlin
- F. R

**Answer:** DF

**NEW QUESTION 7**

Jhon is working on an ELT process that sources data from six different source systems.

Looking at the source data, he finds that data about the sample people exists in two of six systems.

What does he have to make sure he checks for in his ELT process? Choose the best answer.

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

**Answer:** C

**Explanation:**

Duplicate Data.

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

**NEW QUESTION 8**

Jenny wants to study the academic performance of undergraduate sophomores and wants to determine the average grade point average at different points during

an academic year.  
What best describes the data set she needs?

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 9**

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

Original data:

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

Modified data:

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

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

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

**Answer:** B

**Explanation:**

The correct answer is B. Recoding.  
Recoding is a data manipulation technique that involves changing the values or categories of a variable to make it more suitable for analysis. Recoding can be used to simplify or group the data, to correct errors or inconsistencies, or to create new variables from existing ones<sup>12</sup>  
In the example, the analyst used recoding to change the values of Var001, Var002, Var003, and Var004 from numerical to textual form. The analyst also used recoding to assign meaningful labels to the values, such as ??Absent?? for 0, ??Present?? for 1, ??Low?? for 2, ??Medium?? for 3, and ??High?? for 4. This makes the data more understandable and easier to analyze.

**NEW QUESTION 10**

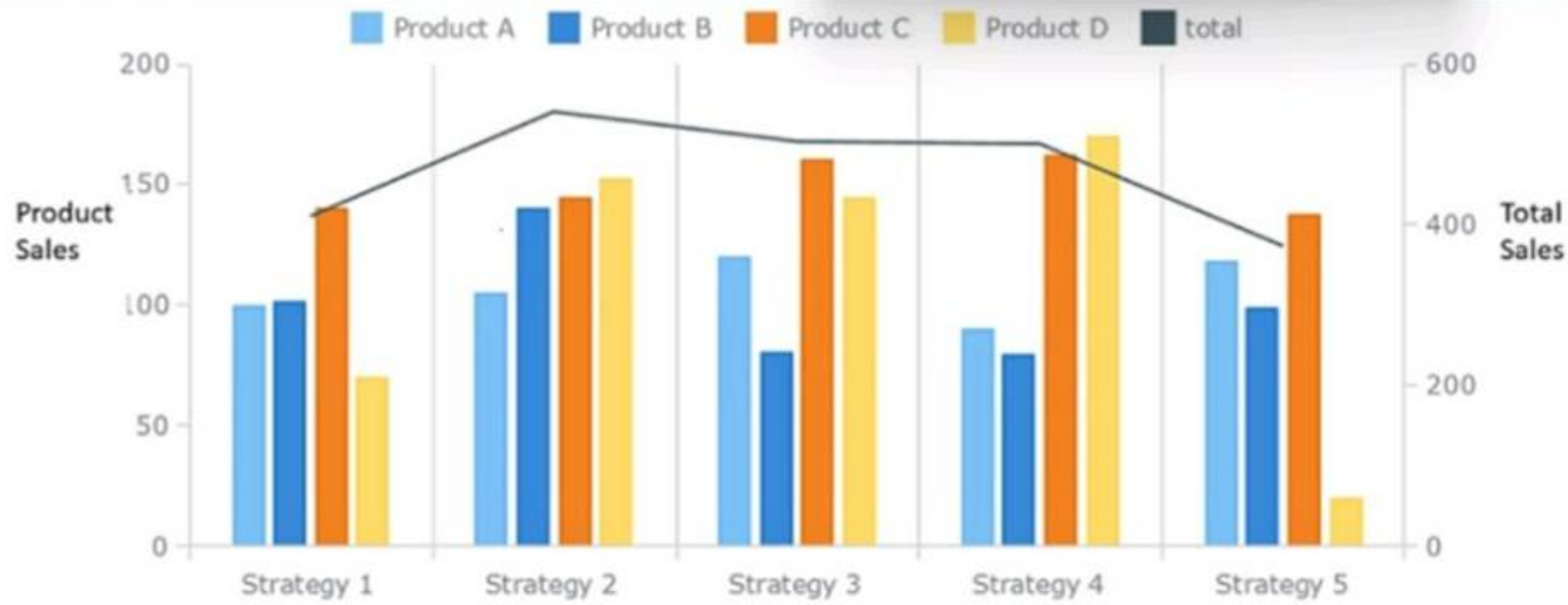
A sales director has requested a report for individual team members within the division be developed. The director would like the report to be shared with all team members, but individual team members should not be identifiable within the report Which of the following access requirements would support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

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

**Answer:** B

**NEW QUESTION 10**

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



- A. Sales are approximately equal for Product A and Product B across all strategies.
- B. Strategy 4 provides the best sales in comparison to other strategies.
- C. While Strategy 2 does not result in the highest sales of Product D. over all products it appears to be the most effective.
- 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 13**

A data analyst wants to create "Income Categories" that would be calculated based on the existing variable "Income". The "Income Categories" would be as follows:

Income category 1: less than \$1.

Income category 2: more than \$1 and less than \$20,000. Income category 3: more than \$20,001 and less than \$40,000. Income category 4: more than \$40,001.

Which of the following data manipulation techniques should the data analyst use to create "Income Categories"?

- A. Data merge
- B. Derived variables
- C. Data blending
- D. Data append

**Answer: B**

**Explanation:**

The correct answer is B: Derived variables. Derived variables are variables that you create by calculating or categorizing variables that already exist in your data set.

Data merge is incorrect. Data merging is the process of combining two or more data sets into a single data set. Data blending is incorrect.

Data blending involves pulling data from different sources and creating a single, unique, dataset for visualization and analysis.

Data append is incorrect. A data append is a process that involves adding new data elements to an existing database.

**NEW QUESTION 15**

Which of the following data manipulation techniques is an example of a logical function?

- A. WHERE
- B. AGGREGATE
- C. BOOLEAN
- D. IF

**Answer: D**

**Explanation:**

This is because an IF function is a type of logical function that returns a value based on a condition or a set of conditions. An IF function can be used to manipulate data by applying different actions or calculations depending on whether the condition is true or false. For example, an IF function in Excel that can achieve this is:



=IF (condition, value\_if\_true, value\_if\_false)

The other data manipulation techniques are not examples of logical functions. Here is why:

? WHERE is a type of clause that filters data based on a condition or a set of conditions. A WHERE clause can be used to manipulate data by selecting only the rows that satisfy the condition(s). For example, a WHERE clause in SQL that can achieve this is:

```
SELECT column_name FROM table_name WHERE condition;
```

? AGGREGATE is a type of function that performs a calculation on a group of values, such as sum, average, count, etc. An AGGREGATE function can be used to manipulate data by summarizing or aggregating the values in a column or a table. For example, an AGGREGATE function in SQL that can achieve this is:

```
SELECT AGGREGATE(column_name) FROM table_name;
```

? BOOLEAN is a type of data type that represents two possible values: true or false.

A BOOLEAN data type can be used to manipulate data by storing or returning logical values based on a condition or a set of conditions. For example, a BOOLEAN data type in Python that can achieve this is:

```
boolean_variable = condition
```

#### NEW QUESTION 16

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 17

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

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

**Answer:** B

#### Explanation:

The correct answer is B. There can be only one primary key in a data set, whereas there can be multiple unique keys.

A primary key is a column or a set of columns that uniquely identifies each row in a table. A table can have only one primary key, which also enforces the NOT NULL constraint on the column(s) involved. A primary key can also be referenced by a foreign key of another table to establish a relationship between the tables

A unique key is a column or a set of columns that also uniquely identifies each row in a table, but it is not the primary key. A table can have more than one unique key, which also allows one NULL value for the column(s) involved. A unique key can also be referenced by a foreign key of another table to establish a relationship between the tables

Some of the differences between a primary key and a unique key are:

? A primary key creates a clustered index on the column(s), whereas a unique key creates a non-clustered index on the column(s)

? A primary key does not allow any NULL values, whereas a unique key allows one

NULL value for the column(s)

? A primary key can be a unique key, but a unique key cannot be a primary key

#### NEW QUESTION 21

Which of the following will MOST likely be streamed live?

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

**Answer:** A

#### Explanation:

Machine data is the most likely type of data to be streamed live, as it refers to data generated by machines or devices, such as sensors, web servers, network

devices, etc. Machine data is often produced continuously and in large volumes, requiring real-time processing and analysis. Other types of data, such as key-value pairs, delimited rows, and flat files, are more likely to be stored in databases or files and processed in batches.

#### NEW QUESTION 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

Given the following table:

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

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

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

**Answer:** B

#### NEW QUESTION 30

The process of performing initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization is called:

- A. a t-test.
- B. a performance analysis.
- C. an exploratory data analysis.
- D. a link analysis.

**Answer:** C

#### Explanation:

This is because exploratory data analysis is a type of process that performs initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization, such as box plots, histograms, scatter plots, etc. Exploratory data analysis can be used to understand and summarize the data, as well as to generate hypotheses or questions for further analysis or research. For example, exploratory data analysis can be used to identify and visualize the characteristics, features, or behaviors of the data, as well as to measure their distribution, frequency, or correlation. The other options are not types of processes that perform initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization. Here is what they mean:

? A t-test is a type of statistical method that tests whether there is a significant difference between the means of two groups or samples, such as whether there is a difference between the average exam scores of two classes in this case. A t-test can be used to test or verify a claim or an assumption about the data, as well as to measure the confidence or the error of the estimation.

? A performance analysis is a type of process that measures whether the data meets certain goals or objectives, such as targets, benchmarks, or standards. A performance analysis can be used to identify and visualize the gaps, deviations, or variations in the data, as well as to measure the efficiency, effectiveness, or quality of the outcomes. For example, a performance analysis can be used to determine if there is a gap between a student's test score and their expected score based on their previous performance.

? A link analysis is a type of process that determines whether the data is connected to other datapoints, such as entities, events, or relationships. A link analysis can be used to identify and visualize the patterns, networks, or associations among the datapoints, as well as to measure the strength, direction, or frequency of the connections. For example, a link analysis can be used to determine if there is a connection between a customer's purchase history and their loyalty program status.

**NEW QUESTION 31**

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 category<sup>12</sup>.

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 plaintext<sup>3</sup>.

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 sun<sup>4</sup>.

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 32**

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

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

**Answer:** A

**Explanation:**

In technical documentation, a glossary is the designated section where definitions for technical terms are provided. It serves as a reference point for readers to understand specialized or uncommon words used within the document. Including descriptions of technical terms in a glossary ensures that readers have a consistent resource to refer to, which can improve comprehension and reduce misunderstandings<sup>12</sup>.

A system diagram (Option B) is a visual representation of the system's components and their interactions, not a place for defining terms. User requirements (Option C) outline what end-users expect from the system, and an index (Option D) is an alphabetical list of topics covered in the document, usually with page numbers, but not definitions.

References:

? Creating effective technical documentation<sup>1</sup>.

? Best practices when writing technical descriptions<sup>3</sup>.

**NEW QUESTION 33**

An analyst wants to create a historical data set for the past five years with each year in its own data set. Which of the following methods is the best way to create this historical data set?

- A. Data transpose
- B. Data concatenation
- C. Data append
- D. Data normalization

**Answer:** B

**NEW QUESTION 38**

A data analyst needs to create a master file that includes customer information from the tables below:



Table 1: Online Transactions

Order_ID	Customer_ID	Date	Amount	Quantity
002A	002	03/01/2020	\$800	109
001B	001	02/01/2020	\$400	14
001B	001	02/01/2020	\$400	14
001B	001	02/01/2020	\$400	14
004C	004	06/01/2020	\$700	52
003D	003	05/01/2020	\$900	20

Table 2: 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

Table 3: Customer Table

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

Given the three tables above, the analyst wants to filter down the information prior to joining it together. In which of the following orders should this data manipulation be approached for the most efficient result?

- A. Merge, append, deduplicate
- B. Merge, deduplicate, append
- C. Deduplicate, append, merge
- D. Append, deduplicate, merge

**Answer:** B

**Explanation:**

For efficient data manipulation, the ideal order would be to first merge related tables to create a comprehensive set of records, then deduplicate to remove any redundant information. Lastly, appending additional data, such as from another source or table, ensures that all relevant data is included without redundancy before the final analysis. This order prevents unnecessary duplication of effort, such as deduplicating both before and after appending, which would be less efficient.

In the context of the tables provided, merging would likely involve combining customer information from the online and in-store transaction tables with the customer table. Deduplication would remove any redundant customer records that may exist across these tables. Finally, appending would involve adding any additional transaction records to the master file, ensuring a complete dataset for analysis.

**NEW QUESTION 39**

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:

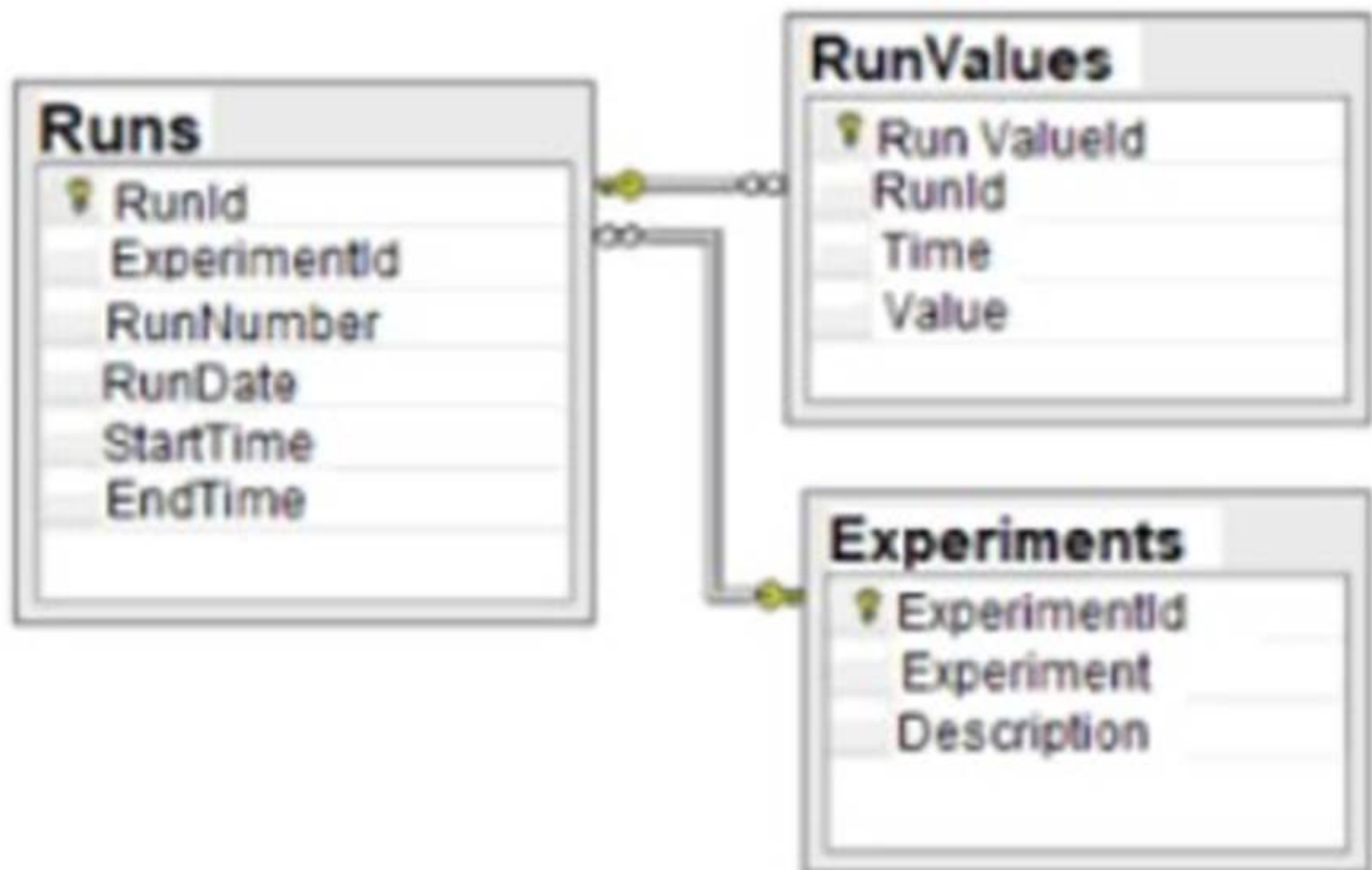
? Numeric is a type of variable that represents a numerical value, such as an integer, a decimal, or a fraction. Numeric variables can be used to store, manipulate, or analyze quantitative data, such as amounts, prices, or scores. For example, numeric variables can be used to perform arithmetic operations 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 44**

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 contains alphanumeric values?

- A. 10.1<sup>2</sup>
- B. 13.6
- C. 1347
- D. A3J7

**Answer:** D

**Explanation:**

Alphanumeric values are values that contain both letters and numbers, such as A3J7. The other options are numeric values, as they contain only numbers, such as 10.1E2, 13.6, and 1347. Reference: Guide to CompTIA Data+ and Practice Questions - Pass Your Cert

**NEW QUESTION 49**

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



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

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

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 53**

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

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

**Answer:** C

**Explanation:**

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

**NEW QUESTION 58**

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 operations<sup>1</sup>.

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 60**

Emma is working in a data warehouse and finds a finance fact table links to an organization dimension, which in turn links to a currency dimension that not linked to the fact table.

What type of design pattern is the data warehouse using?

- A. Star.
- B. Sun.
- C. Snowflake.
- D. Comet.



**Answer:** C

**Explanation:**

Correct answer C. Snowflake.  
 Since the dimension links to a dimension that isn't connected to the fact table, it must be a Snowflake, with a Star, all dimensions link directly to the fact table, Sun and Comet are not data warehouse design patterns.

**NEW QUESTION 64**

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

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

**Answer:** AC

**Explanation:**

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

**NEW QUESTION 65**

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

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

**Answer:** A

**NEW QUESTION 68**

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 69

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 70

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 72

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 preferred<sup>1</sup>.

? Best practices in reporting also recommend automated and scheduled reports for consistent and timely updates, which operational reports provide<sup>2</sup>.

#### NEW QUESTION 74

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

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

**Answer: C**

#### Explanation:

Filter on any of the responses that do not say ??January?? and update them to ??January??. This is because filtering and updating are data cleansing techniques that can be used to ensure data consistency, which means that the data is uniform and follows a standard format. By filtering on any of the responses that do not say ??January?? and updating them to ??January??., the analyst can make sure that all the responses for the month of January are written in the same way. The other steps are not appropriate for ensuring data consistency. Here is why:

Deleting any of the responses that do not have ??January?? written out would result in data loss, which means that some information would be missing from the data set. This could affect the accuracy and reliability of the analysis.

Replacing any of the responses that have ??01?? would not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??Jan?? and ??January??. This could cause confusion and errors in the analysis. Sorting any of the responses that say ??Jan?? and updating them to ??01?? would also not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??01?? and ??January??. This could also cause confusion and errors in the analysis.

#### NEW QUESTION 75

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

- A. To perform web scraping

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

**Answer:** CF

**Explanation:**

Two reasons to conduct data cleansing are:

? To improve accuracy: Data cleansing helps to ensure that the data is correct, consistent, and reliable. This can improve the quality and validity of the analysis, as well as the decision-making and outcomes based on the data<sup>12</sup>

? To calculate trends: Data cleansing helps to remove or resolve any errors, outliers, or missing values that could distort or skew the data. This can help to identify and measure the patterns, changes, or relationships in the data over time<sup>13</sup>

**NEW QUESTION 80**

Given the following data tables:

CustomerID	CustomerLastName
01	Manzelli
02	Kraus

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

ItemID	Customer_Last_Name	QuantityPurchased
01	Brown	03
02	Smee	07

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

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

**Answer:** A

**Explanation:**

This is because a data dictionary is a type of document that defines and describes the data elements, attributes, and relationships in a database or a data set. A data dictionary can be used to facilitate the MDM (Master Data Management) process, which is a process that aims to ensure the quality, consistency, and accuracy of the data across different sources and systems. By creating a data dictionary first, the analyst can establish a common understanding and standardization of the data field names, types, formats, and meanings, as well as identify any potential issues or conflicts in the data, such as missing values, duplicate values, or inconsistent values. The other MDM processes can take place after creating a data dictionary. Here is why:

Compliance with regulations is a type of MDM process that ensures that the data meets the legal and ethical requirements and standards of the industry or the organization.

Compliance with regulations can take place after creating a data dictionary, because the data dictionary can help the analyst to identify and apply the relevant rules and policies to the data, such as data privacy, security, or retention.

Standardization of data field names is a type of MDM process that ensures that the data field names are consistent and uniform across different sources and systems. Standardization of data field names can take place after creating a data dictionary, because the data dictionary can provide a reference and a guideline for naming and labeling the data fields, as well as resolving any discrepancies or ambiguities in the data field names.

Consolidation of multiple data fields is a type of MDM process that combines or merges the data fields from different sources or systems into a single source or system. Consolidation of multiple data fields can take place after creating a data dictionary because the data dictionary can help the analyst to map and match the data fields from different sources or systems based on their definitions and descriptions, as well as eliminating any redundant or duplicate data fields.

**NEW QUESTION 84**

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

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

**Answer:** C

**Explanation:**

The purpose of an executive summary is to provide a concise and informative overview of a longer report, allowing busy stakeholders to quickly understand the key points and findings without reading the entire document. This summary should highlight the most important data, conclusions, and recommendations, and is typically placed at the beginning of the report for easy access<sup>12</sup>.

In the context of a detailed sales report for senior management, including a summary section with observations and insights would allow the team to quickly grasp the performance across various offices and identify any significant trends or issues that require attention. This approach aligns with best practices for executive reporting, which emphasize the importance of clear and concise summaries that focus on essential KPIs and actionable insights<sup>12</sup>.

References: 1: Databox - How to Write an Executive Summary for a Report: Step By Step Guide with Examples 2: LinkedIn - Best Practices for Writing Executive Summaries

**NEW QUESTION 85**

Which of the following is a relational database?

- A. SQL
- B. Excel
- C. JSON
- D. NoSQL

**Answer:** A

**NEW QUESTION 86**

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 90**

Under which of the following circumstances should the null hypothesis be accepted when  $\alpha = 0.05$ ?

- 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 error<sup>12</sup>.

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 93**

Which of the following tools would be best to use to calculate the interquartile range, median, mean, and standard deviation of a column in a table that has 5,000,000 rows?

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

**Answer:** B

**NEW QUESTION 96**

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

- A. Discrete data cannot create a sloped line.
- B. Discrete data can only be a finite number of values.
- C. Discrete data can have decimal points.
- D. Discrete data applies only to numbers.

**Answer:** B



**Explanation:**

Discrete data are data that can only assume specific values that are countable and distinct. For example, the number of books, the number of heads in a coin toss, or the number of patients in a hospital are discrete data. Discrete data cannot have fractional or decimal values, and there are clear spaces between the possible values<sup>12</sup>. Continuous data are data that can assume any value within a range and can be meaningfully divided into smaller parts. For example, the weight, height, length, time, or temperature are continuous data. Continuous data can have fractional or decimal values, and there are infinite numbers of possible values between any two points<sup>12</sup>.

**NEW QUESTION 101**

Given the following report:

# Quarterly Customer Service Report

**Table 1. Frequency of Ticket Statuses**

Status	Count
Reported	11
In-Progress	323
Closed	554

**Table 2. Occurrence of Target Phrases**

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

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

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

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

**Answer:** E

**Explanation:**

The date on which the report was run. This is because the time period the report covers and the date on which the report was run are two components that need to be added to ensure the report is point-in-time and static, which means that the report shows the data as it was at a specific moment or interval in time, and does not change or update with new data. By adding the time period the report covers and the date on which the report was run, the analyst can indicate when and for how long the data was collected and analyzed, as well as avoid any confusion or ambiguity about the currency or validity of the data. The other components do not need to be added to ensure the report is point-in-time and static. Here is why:

A control group for the phrases is a type of group that serves as a baseline or a reference for comparison with another group that is exposed to some treatment or intervention, such as a target phrase in this case. A control group for the phrases does not need to be added to ensure the report is point-in-time and static, because it does not affect the time frame or the stability of the data. However, a control group for the phrases could be useful for evaluating the effectiveness or impact of the target phrases on customer satisfaction or retention.

A summary of the KPIs is a type of document that provides an overview or a highlight of the key performance indicators (KPIs), which are measurable values that indicate how well an organization or a process is achieving its goals or objectives. A summary of the KPIs does not need to be added to ensure the report is point-in-time and static, because it does not affect the time frame or the stability of the data. However, a summary of the KPIs could be useful for communicating or presenting the main findings or insights from the report.

Filter buttons for the status are a type of feature or function that allows users to select or deselect certain values or categories in a column or a table, such as ticket statuses in this case. Filter buttons for the status do not need to be added to ensure the report is point-in-time and static, because they do not affect the time frame or the stability of the data. However, filter buttons for the status could be useful for exploring or analyzing different aspects or segments of the data.

**NEW QUESTION 104**

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

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

**Answer:** B

**NEW QUESTION 107**

Which of the following is an example of structured data?

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

**Answer:** A

**Explanation:**

A credit card number is an example of structured data, which is a type of data that conforms to a data model, has a well-defined structure, follows a consistent order, and can be easily accessed and used by a person or a computer program. A credit card number consists of 16 digits that are divided into four groups of four digits each, separated by spaces or hyphens. The first six digits indicate the issuer identification number, the next nine digits indicate the account number, and the last digit is a check digit that validates the number. A credit card number can be stored and processed in a structured format, such as a database or a spreadsheet<sup>1</sup>.

**NEW QUESTION 109**

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 110**

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 databases<sup>1</sup>.

? 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 databases<sup>2</sup>.

? 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 surveys<sup>3</sup>.

#### NEW QUESTION 114

Given the below:

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

Which of the following numbers represents a Type I error?

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

**Answer:** C

#### NEW QUESTION 118

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

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

**Answer:** C

#### NEW QUESTION 119

An analyst is working with the income data of suburban families in the United States. The data set has a lot of outliers, and the analyst needs to provide a measure that represents the typical income. Which of the following would BEST fulfill the analyst's goal?

- A. Median
- B. Mean
- C. Mode
- D. Standard deviation

**Answer:** A

#### Explanation:

his is because median is a type of statistical measure that represents the typical value or central tendency of a data set, which means that it divides the data set into two equal halves, such that half of the values are above it and half are below it. Median can be used to provide a measure that represents the typical income of suburban families in the United States, especially when the data set has a lot of outliers, which means that it has values that are unusually high or low compared to the rest of the data set. Median can provide a measure that represents the typical income of suburban families in the United States, because it is not affected or skewed by the outliers, as it only depends on the middle value or the middle two values of the data set, regardless of how extreme or distant the outliers are. For example, median can provide a measure that represents the typical income of suburban families in the United States, by finding the income value that splits the data set into two equal groups of families, such that 50% of the families have higher incomes and 50% have lower incomes. The other statistical measures are not the best measures to represent the typical income of suburban families in the United States. Here is why:

? Mean is a type of statistical measure that represents the average value or central tendency of a data set, which means that it is the sum of all the values divided by the number of values. Mean is not a good measure to represent the typical income of suburban families in the United States, especially when the data set has a lot of outliers, because it is affected or skewed by the outliers, as it takes into account all the values in the data set, regardless of how extreme or distant they are. For example, mean can provide a measure that does not represent the typical income of suburban families in the United States, by finding the income value that is influenced by a few very high or very low incomes, which could make it higher or lower than most of the incomes in the data set.

? Mode is a type of statistical measure that represents the most frequent value or mode of a data set, which means that it is the value that occurs most often in the data set. Mode is not a good measure to represent the typical income of suburban families in the United States, especially when the data set has a lot of outliers, because it is not representative or indicative of the central tendency or distribution of the data set, as it only depends on the count or occurrence of a single value or a few values in the data set, regardless of how common or rare they are. For example, mode can provide a measure that does not represent the typical income of suburban families in the United States, by finding the income value that is repeated more often than others, which could be an outlier or an anomaly in the data set.

? Standard deviation is a type of statistical measure that represents the amount of dispersion or variation of a data set, which means that it quantifies how much the values in a data set vary or deviate from the mean or average of the data set. Standard deviation is not a measure that represents the typical income of suburban families in the United States, but rather a measure that describes the spread or distribution of their incomes, as well as identifies any outliers or extreme values in their incomes. For example, standard deviation can provide a measure that describes how diverse or homogeneous their incomes are, as well as how far their incomes are from their average income.



#### NEW QUESTION 122

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

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

**Answer: B**

#### NEW QUESTION 126

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

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

**Answer: D**

#### NEW QUESTION 131

Exhibit.

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

Which of the following logical statements results in Table B?

A)

IF Name = "James" and Gender\_flag = "College" then delete

B)

IF Name = "Sam" and Gender\_flag = "Male" then delete

C)

IF Name = "Pat" and Gender\_flag = "Female" then delete

D)

IF Name = "Sean" and Gender\_flag = "College" then delete

- A. Option A
- B. Option B



- C. Option C
- D. Option D

**Answer:** D

**Explanation:**

The logical statement that results in Table B is Option D. Option D is a logical statement that uses the AND operator to combine two conditions: Name = ??Tom?? and Region = ??BC??. The AND operator returns true only if both conditions are true, otherwise it returns false. Therefore, Option D will select only the rows from Table A that satisfy both conditions, which are rows 4, 5, 6, and 7. These rows form Table B, as shown below: Name | Gender flag | Level | College | Code |  
Region Tom | Male | Elementary | A | BC | BC Kim | Female | Elementary | A | BC | BC Pat | Female | Elementary | A | BC | BC Ben | Male | Elementary | A | BC | BC

The other options are not correct, as they use different logical operators or conditions that do not result in Table B. Option A uses the OR operator, which returns true if either condition is true, or both. Option A will select all the rows from Table A except row 3, which does not match either condition. Option B uses the NOT operator, which returns the opposite of the condition. Option B will select all the rows from Table A except rows 4, 5, 6, and 7, which match the condition. Option C uses a different condition, Region = ??ON??. which does not match any row in Table A. Option C will select no rows from Table A. Reference: [SQL Logical Operators - W3Schools]

**NEW QUESTION 136**

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

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

**Answer:** C

**Explanation:**

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

**NEW QUESTION 141**

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 144**

Which of the following is the correct data type for text?

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 149**

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

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

**Answer:** C

**Explanation:**

This is because link analysis is a type of analysis that determines whether the data being analyzed is connected to other datapoints, such as entities, events, or relationships. Link analysis can be used to identify and visualize the patterns, networks, or associations among the datapoints, as well as measure the strength, direction, or frequency of the connections. For example, link analysis can be used to determine if there is a connection between a customer's purchase history and their loyalty program status. The other types of analysis are not the best types of analysis to conduct to determine whether the data being analyzed is connected to other datapoints. Here is why:

? Trend analysis is a type of analysis that determines whether the data being analyzed is changing over time, such as increasing, decreasing, or fluctuating. Trend analysis can be used to identify and visualize the patterns, cycles, or movements in the data points, as well as measure the rate, direction, or magnitude of the changes. For example, trend analysis can be used to determine if there is a change in a company's sales revenue over a period of time.

? Performance analysis is a type of analysis that determines whether the data being analyzed is meeting certain goals or objectives, such as targets, benchmarks, or standards. Performance analysis can be used to identify and visualize the gaps, deviations, or variations in the data points, as well as measure the efficiency, effectiveness, or quality of the outcomes. For example, performance analysis can be used to determine if there is a gap between a student's test score and their expected score based on their previous performance.

? Exploratory analysis is a type of analysis that determines whether there are any insights or discoveries in the data being analyzed, such as patterns, relationships, or anomalies. Exploratory analysis can be used to identify and visualize the characteristics, features, or behaviors of the data points, as well as measure their distribution, frequency, or correlation. For example, exploratory analysis can be used to determine if there are any outliers or unusual values in a dataset.

#### NEW QUESTION 150

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 152

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

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

**Answer: B**

#### Explanation:

The measure of dispersion is used to describe the spread of data around a central value. In the context of a data set containing actual and projected sales, the measure of dispersion will help to understand the variability or consistency of sales figures. The variance is the most appropriate statistical method for finding the measure of dispersion because it calculates the average of the squared differences from the Mean, providing a clear picture of data spread. It is especially useful in comparing the spread between different data sets and understanding the distribution of data points.

? Mean is a measure of central tendency, not dispersion.

? Correlation measures the relationship between two variables, not the spread of a single variable.

? Confidence intervals are used to estimate the range within which a population parameter will fall, but they do not measure dispersion within the data set itself.

References:

? Measures of Dispersion in Statistics<sup>1</sup>

? Measures of Dispersion - Definition, Formulas, Examples<sup>2</sup>

? Statistical dispersion - Wikipedia<sup>3</sup>

#### NEW QUESTION 156

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 analysis<sup>12</sup>

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-up<sup>12</sup>

#### NEW QUESTION 161

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

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

**Answer: B**

#### Explanation:

After conducting a review of the business questions, the next step for the analyst is to determine the data needs and sources for analysis. This involves identifying the relevant data elements, variables, and metrics that are required to answer the business questions, as well as the data sources, formats, and quality that are available to access and use. This step will help the analyst to plan the data collection, preparation, and integration processes, as well as to assess the feasibility and limitations of the analysis<sup>1</sup>.

#### NEW QUESTION 165

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 BI
- 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 SharePoint<sup>12</sup>

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 169

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 ambiguities<sup>1</sup>.

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 174

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

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

**Answer: D**

#### Explanation:

Transpose creates a new data file in which the rows and columns in the original data file are transposed so that cases (rows) become variables and variables (columns) become

cases. Transpose automatically creates new variable names and displays a list of the new variable names.

Transposing data is useful for data analysis. At times, we have to pull data from various files with different formats for analysis and preparing reports. In such

circumstances, we may have to transpose some data from one file to the other. In excel, we can transpose data in multiple ways.

**NEW QUESTION 179**

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 181**

Which one of the following is a measure of dispersion?

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

**Answer: A**

**NEW QUESTION 185**

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 the 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 188

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

- A. Order number
- B. salesperson
- C. date shipped, recipient address, and price
- D. Item name, salesperson
- E. recipient address, shipping cost
- F. and date shipped
- G. Item number, item name, salesperson
- H. date sold
- I. and price
- J. Item name
- K. salesperson
- L. price
- M. shipping cost
- N. and date shipped

**Answer:** C

#### Explanation:

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

#### NEW QUESTION 192

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 193

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

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

**Answer:** B

#### Explanation:

Correct answer B. When data is at rest or in transit.

To provide maximum protection, encrypt data both in transit and at rest.

#### NEW QUESTION 197

An analyst is creating a resource to improve users' experience when they select specific records based on particular dates. Which of the following should the

analyst use to create a resource that best meets user needs?

- A. Drop-down menu
- B. Date range
- C. Text field
- D. Frequency

**Answer:** B

**Explanation:**

A drop-down menu is a graphical user interface element that allows users to select one option from a list of options that are hidden until the user clicks on the menu. A drop-down menu can be used to create a resource that best meets user needs when they select specific records based on particular dates, because:

? A drop-down menu can provide a predefined list of dates or date ranges that are relevant and valid for the records, such as today, yesterday, last week, last month, custom range, etc. This can help users to avoid typing errors or invalid dates in a text field, and to save time and effort in entering the dates.

? A drop-down menu can also provide a calendar or a date picker that allows users to select a specific date or a range of dates from a graphical representation of a calendar. This can help users to visualize and compare the dates, and to easily adjust or modify their selection.

? A drop-down menu can improve the user experience by making the interface more compact and organized, as it only shows one option at a time and hides the rest of the options until the user clicks on the menu. This can help users to focus on their selection and to avoid clutter and distraction.

**NEW QUESTION 198**

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 202**

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

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

**Answer:** A

**NEW QUESTION 207**

Given the data below:

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

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

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 210**

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 214**

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 JOIN, four rows
- C. four rows
- D. RIGHT JOIN, four rows
- 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 table<sup>1</sup>

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 216

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 220

A data analyst has been asked to derive a new variable labeled ??Promotion\_flag?? based on the total quantity sold by each salesperson. Given the table below:

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

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

- A. Date
- B. Mathematical
- C. Logical
- D. Aggregate

**Answer:** C

**Explanation:**

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

**NEW QUESTION 225**

A sales analyst needs to report how the sales team is performing to target. Which of the following files will be important in determining 2019 performance attainment?

- A. 2018 goal data
- B. 2018 actual revenue
- C. 2019 goal data
- D. 2019 commission plan

**Answer:** C

**Explanation:**

Answer: C. 2019 goal data

To report how the sales team is performing to target, the sales analyst needs to compare the actual sales revenue with the expected or planned sales revenue for the same period. The 2019 goal data is the file that contains the expected or planned sales revenue for the year 2019, which is the target that the sales team is aiming to achieve. By comparing the 2019 goal data with the 2019 actual revenue, the sales analyst can calculate the performance attainment, which is the percentage of the goal that was met by the sales team.

Option A is incorrect, as 2018 goal data is not relevant for determining 2019 performance attainment. The 2018 goal data contains the expected or planned sales revenue for the year 2018, which is not the target that the sales team is aiming to achieve in 2019.

Option B is incorrect, as 2018 actual revenue is not relevant for determining 2019 performance attainment. The 2018 actual revenue contains the actual sales revenue for the year 2018, which is not comparable with the 2019 goal data or the 2019 actual revenue. Option D is incorrect, as 2019 commission plan is not relevant for determining 2019 performance attainment. The 2019 commission plan contains the rules and rates for calculating and paying commissions to the sales team based on their performance attainment, but it does not contain the expected or planned sales revenue for the year 2019.

**NEW QUESTION 226**

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

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

**Answer:** C

**Explanation:**

When updating a legacy data source, it is a best practice to create a codebook to document field changes. A codebook serves as a detailed guide and record of the data structure, definitions, and any transformations or modifications made to the data fields. This documentation is crucial for maintaining data integrity, ensuring consistency, and facilitating future data use and understanding. It provides a reference that can be invaluable for data analysts, developers, and any stakeholders who need to work with the data.

Creating a codebook is preferred over placing old data in new fields, which can lead to confusion and data integrity issues. Keeping only the most recent data may result in the loss of valuable historical information. Removing the data source from production is not a practice related to updating data but rather to retiring a data source<sup>1234</sup>.

References:

- ? Legacy Data Migration: A Comprehensive Guide | OpenGeeksLab
- ? How to Successfully Complete Legacy Database Migration
- ? Methods for Saving and Integrating Legacy Data - DATAVERSITY
- ? Legacy Data Digitization - Learn The Best Practices

**NEW QUESTION 231**

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

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

**Answer:** A

**Explanation:**

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

#### NEW QUESTION 235

Which of the following is the correct extension for a tab-delimited spreadsheet file?

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

**Answer: C**

#### Explanation:

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

#### NEW QUESTION 236

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

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

**Answer: B**

#### Explanation:

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

#### NEW QUESTION 237

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

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

**Answer: C**

#### NEW QUESTION 242

After completing web scraping, which of the following file formats needs to be parsed?

- A. .html
- B. .txt
- C. .csv
- D. .tsv

**Answer: A**

#### Explanation:

The correct answer is .html.

Short Explanation: Web scraping is the process of extracting data from websites by parsing the HTML code of the web pages. HTML stands for HyperText Markup Language and it is the standard markup language for creating web pages and web applications. HTML files have the extension .html and they contain tags, elements, attributes, and content that define the structure and appearance of a web page. Web scraping tools need to parse the HTML files to extract the relevant data from the web pages.

#### NEW QUESTION 246

Which of the following is a characteristic of a relational database?

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

**Answer: C**

#### Explanation:

It is structured in nature. This is because a relational database is a type of database that organizes data into tables, which consist of rows and columns. A relational database is structured in nature, which means that the data has a predefined schema or format, and follows certain rules and constraints, such as primary keys, foreign keys, or referential integrity. A relational database can be used to store, query, and manipulate data using a structured query language (SQL). The other characteristics are not true for a relational database. Here is why:

It utilizes key-value pairs. This is not true for a relational database, because key-value pairs are a way of storing data that associates each value with a unique key, such as an identifier or a name. Key-value pairs are typically used in non-relational databases, such as NoSQL databases, which do not have tables, rows, or columns, but rather store data in various formats, such as documents, graphs, or columns.



It has undefined fields. This is not true for a relational database, because fields are another name for columns in a table, which define the attributes or properties of each row or record in the table. Fields have defined names, types, and lengths in a relational database, which specify the format and size of the data that can be stored in each field.

It uses minimal memory. This is not true for a relational database, because memory is the amount of space or storage that is used by a database to store and process data. Memory usage depends on various factors, such as the size, complexity, and number of tables and queries in a relational database. A relational database can use a lot of memory if it has many tables with many rows and columns, or if it performs complex or frequent queries on the data.

#### NEW QUESTION 248

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

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

**Answer:** AB

#### Explanation:

The reasons to create and maintain a data dictionary are to improve data acquisition and to remember specifics about data fields. A data dictionary is a document or a database that describes the structure, meaning, and usage of the data elements in a data source or a database. A data dictionary can help to improve data acquisition by providing clear and consistent definitions, rules, and standards for the data collection process. A data dictionary can also help to remember specifics about data fields by providing information such as data type, format, length, range, default value, constraints, relationships, etc. The other options are not reasons to create and maintain a data dictionary, as they are related to other aspects of data management or security. A data dictionary does not specify user groups for databases, as this is a function of access control or authorization. A data dictionary does not provide continuity through personnel turnover, as this is a function of documentation or knowledge transfer. A data dictionary does not confine breaches of PHI data, as this is a function of encryption or anonymization. A data dictionary does not reduce processing power requirements, as this is a function of optimization or compression. Reference: [What is a Data Dictionary? - DataCamp]

#### NEW QUESTION 253

A customer survey reveals 90% positive feedback. Which of the following statistical methods would be best to utilize to determine the reliability of a data set and predict how a larger sample of customers over the same time period might respond?

- A. Calculate a high variance on survey responses.
- B. Calculate the maximum range of the survey responses.
- C. Calculate a low standard deviation on survey responses.
- D. Remove any data more than 4 standard deviation from the mean.

**Answer:** C

#### Explanation:

A low standard deviation in survey responses indicates that the data points tend to be close to the mean, suggesting a high level of consistency among the responses. This consistency is crucial for determining the reliability of the data set and predicting future outcomes. If the standard deviation is low, it means that the positive feedback is not only high but also consistent, making it a reliable indicator of customer satisfaction and a good predictor of how a larger sample might respond.

References: The concept of using standard deviation to assess data reliability is a standard practice in statistics and data analysis<sup>123</sup>.

#### NEW QUESTION 257

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

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

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

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

**Answer:** A

#### Explanation:

The p-value is a measure of how likely it is to observe a difference in conversion rates as large or larger than the one observed, assuming that there is no difference between the groups. A common threshold for statistical significance is 0.05, meaning that there is a 5% or less chance of observing such a difference by chance alone. The table shows the p-values for each country, and we can see that only Germany has a p-value above 0.05 (0.13). This means that we cannot reject the null hypothesis that there is no difference in conversion rates between the test and control groups in Germany. Therefore, the increase in conversion from the new layout was not significant in Germany. For the other countries, the p-values are below 0.05, indicating that the increase in conversion from the new layout was statistically significant. Option A is correct.



Option B is incorrect because the increase in conversion from the new layout was significant in France (p-value = 0.002).

Option C is incorrect because it does not account for the variation across countries. While the overall conversion rate for the test group (8.4%) is higher than the control group (6.8%), this difference may not be statistically significant when we consider the country-specific effects.

Option D is incorrect because the new layout has the highest conversion rate in the United Kingdom (9.6%), not the lowest.

References:

? P-value Calculator & Statistical Significance Calculator

? p-value Calculator | Formula | Interpretation

? How to obtain the P value from a confidence interval | The BMJ

? Confidence Intervals & P-values for Percent Change / Relative Difference

#### NEW QUESTION 261

Which of the following file formats is best suited to start exploratory analysis within statistical software?

A. CSV

B. XLSM

C. XML

D. JSON

**Answer:** A

#### NEW QUESTION 262

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

A. January 1, 2020 to April 1, 2020

B. January 1, 2020 to April 7, 2020

C. January 1, 2020 to April 8, 2020

D. January 1, 2020 to April 9, 2020

**Answer:** D

#### Explanation:

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

? January 1, 2020 to April 1, 2020 would not include the sales that occurred in the first eight days of April, which would underestimate the sales year to date.

? January 1, 2020 to April 7, 2020 would not include the sales that occurred in the last two days of April, which would also underestimate the sales year to date.

? January 1, 2020 to April 8, 2020 would not include the sales that occurred on April 9, which would also underestimate the sales year to date.

#### NEW QUESTION 264

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

A. Qlik.

B. Power BI.

C. Domo.

D. Dataroma.

**Answer:** B

#### Explanation:

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

#### NEW QUESTION 266

A business unit made the following modification to the values in a table:

Previous value	New value
56.0	56.0456

Which of the following data quality dimensions was applied in this scenario?

A. Integrity

B. Consistency

C. Completeness

D. Accuracy

**Answer:** D

#### NEW QUESTION 268

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:

The best chart to use to identify the composition between the categories of the survey response data set is a pie chart. A pie chart is a circular chart that shows the relative proportions of different categories in a whole. A pie chart is divided into slices that represent the percentage or frequency of each category. A pie chart is suitable for displaying categorical data that has a few categories and does not have any hierarchical or temporal relationship. In this case, a pie chart can show the composition of the favorite colors among the survey respondents, as well as the percentage of each color. The other options are not as good as a pie chart for this purpose, as they are more suitable for displaying numerical data that has some kind of distribution, trend, correlation, or comparison. A histogram is a bar chart that shows the frequency distribution of a single numerical variable. A line chart is a chart that shows the change of one or more numerical variables over time or another continuous variable. A scatter plot is a chart that shows the relationship between two numerical variables by plotting them as points on a Cartesian plane. A waterfall chart is a chart that shows how an initial value is increased or decreased by a series of intermediate values, resulting in a final value. Reference: [Choosing the Right Chart Type - DataCamp]

#### NEW QUESTION 271

An analyst has been asked to validate data quality. Which of the following are the BEST reasons to validate data for quality control purposes? (Choose two.)

- A. Retention
- B. Integrity
- C. Transmission
- D. Consistency
- E. Encryption
- F. Deletion

**Answer:** B

#### Explanation:

Integrity and D. Consistency. This is because integrity and consistency are two of the best reasons to validate data for quality control purposes, which means to check and ensure that the data is accurate, complete, reliable, and usable for the intended analysis or purpose. By validating data for integrity and consistency, the analyst can prevent or correct any errors or issues in the data that could affect the validity or reliability of the analysis or the results. Here is what integrity and consistency mean in terms of data quality:

? Integrity refers to the completeness and validity of the data, which means that the data has no missing, incomplete, or invalid values that could compromise its meaning or usefulness. For example, validating data for integrity could involve checking for null values, outliers, or incorrect data types in the data set.

? Consistency refers to the uniformity and standardization of the data, which means that the data follows a common format, structure, or rule across different sources or systems. For example, validating data for consistency could involve checking for spelling, punctuation, or capitalization errors in the data set.

The other reasons are not the best reasons to validate data for quality control purposes. Here is why:

? Retention refers to the storage and preservation of the data, which means that the data is kept and maintained in a secure and accessible way for future use or reference. Retention does not need to be validated for quality control purposes, because it does not affect the accuracy or reliability of the data itself.

? Transmission refers to the transfer and exchange of the data, which means that the data is moved or shared between different sources or systems in a fast and efficient way. Transmission does not need to be validated for quality control purposes, because it does not affect the completeness or validity of the data itself.

? Encryption refers to the protection and security of the data, which means that the data is encoded or scrambled in a way that prevents unauthorized access or use. Encryption does not need to be validated for quality control purposes, because it does not affect the uniformity or standardization of the data itself.

? Deletion refers to the removal and disposal of the data, which means that the data is erased or destroyed in a way that prevents recovery or retrieval. Deletion does not need to be validated for quality control purposes, because it does not affect the meaning or usefulness of the data itself.

#### NEW QUESTION 275

.....

## Thank You for Trying Our Product

### We offer two products:

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

2nd - Questions 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

**100% Actual & Verified — Instant Download, Please Click**  
**[Order The DA0-001 Practice Test Here](#)**