

# CompTIA

## Exam Questions DA0-001

CompTIA Data+ Certification Exam



#### NEW QUESTION 1

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

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

**Answer:** C

#### Explanation:

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

#### NEW QUESTION 2

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

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

**Answer:** B

#### NEW QUESTION 3

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 4

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:

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

Therefore, option A is correct.

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

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

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

#### NEW QUESTION 5

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

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

**Answer:** B

**Explanation:**

The correct answer is B. Transcription.

Transcription is a necessary step on the way to interpretation when a data set was recorded using multimedia technology. Multimedia technology refers to the use of various forms of media, such as audio, video, images, and text, to capture and present information<sup>1</sup> Transcription is the process of converting multimedia data into written or textual form, which can then be analyzed using various methods and tools<sup>2</sup> Transcription can help to make the data more accessible, searchable, and manageable, as well as to preserve the data for future use.

Structural equation modeling is not correct, because it is a statistical technique that tests the causal relationships between multiple variables using observed and latent variables. Structural equation modeling is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data.

Sequential analysis is not correct, because it is a method of analyzing the order and timing of events or behaviors in a data set. Sequential analysis is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data. Sampling is not correct, because it is the process of selecting a subset of data from a larger population for analysis. Sampling is not a necessary step on the way to interpretation, but rather a preliminary step that can be done before collecting or analyzing the data.

**NEW QUESTION 6**

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

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

**Answer:** A

**Explanation:**

A CSV file is a type of flat file that stores data as plain text in a table-like structure with rows and columns. Each row represents a single record, while columns represent fields or attributes of the data. A CSV file uses commas or other delimiters to separate the values in each row. A CSV file can be easily imported or exported by various applications and programs<sup>12</sup>

**NEW QUESTION 7**

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

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

**Answer:** D

**NEW QUESTION 8**

Which of the following concepts should be applied if a data set with 40 fields needs to be pared down to 20 fields and contains similar data across multiple fields?

- A. Duplication
- B. Consolidation
- C. Compliance
- D. Standardization

**Answer:** B

**Explanation:**

Consolidation is the process of combining multiple elements into a single, more effective or coherent whole. In the context of data analytics, consolidation would involve merging similar fields to reduce the overall number of fields in a dataset. This is particularly useful when a dataset contains redundant or similar data across multiple fields, as it helps to simplify the data structure and improve efficiency. Techniques such as dimensionality reduction are often applied to achieve this, where the goal is to retain the most informative and representative features of the data while reducing the number of total features. References:

? Applied Dimensionality Reduction — 3 Techniques using Python<sup>1</sup>.

? Seven Techniques for Data Dimensionality Reduction<sup>2</sup>.

? Best practices when working with datasets<sup>3</sup>.

? Effectively Handling Large Datasets<sup>4</sup>.

**NEW QUESTION 9**

A data analyst has a set of data that shows the number of gallons of oil produced each day. The company would like to know the standard deviation for the data set. The variance for the data is 36 gallons. Which of the following is the standard deviation for gallons produced?

- A. 1.16
- B. 6
- C. 36
- D. 72

**Answer:** B

**Explanation:**

The standard deviation is a measure of the amount of variation or dispersion in a set of values. It is calculated as the square root of the variance. Given that the variance for the data set is 36 gallons, the standard deviation can be found by taking the square root of 36, which is 6. Therefore, the standard deviation for the number of gallons of oil produced each day is 6 gallons.

References:

? The concept of standard deviation and its calculation is a fundamental aspect of statistics, which is well-documented in statistical textbooks and resources.

? The calculation performed to arrive at the answer is based on the mathematical operation of taking the square root of the variance value.

**NEW QUESTION 10**

You are working with a professional statistician to perform an analysis and would like to use a statistics package.

Which one of the following would be the most appropriate?

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 10**

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 11**

Which of the following is a domain-specific language used in programming that is designed for managing data that is held in a relational data stream management system?

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

**Answer:** B

**Explanation:**

SQL (Structured Query Language) is a domain-specific language used in programming, specifically designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS). It is the standard language for relational database management systems. SQL statements are used to perform tasks such as update data on a database, or retrieve data from a database. Unlike languages like Python or R, which are general-purpose programming languages, SQL is tailored specifically for database management and manipulation.

References:

? ResearchGate article on SQL1.

? SpringerLink chapter on Relational Databases and SQL Language2.

? DataCamp tutorial on SQL Server Installation3.

? Wikipedia page on SQL4.

**NEW QUESTION 14**

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 19**

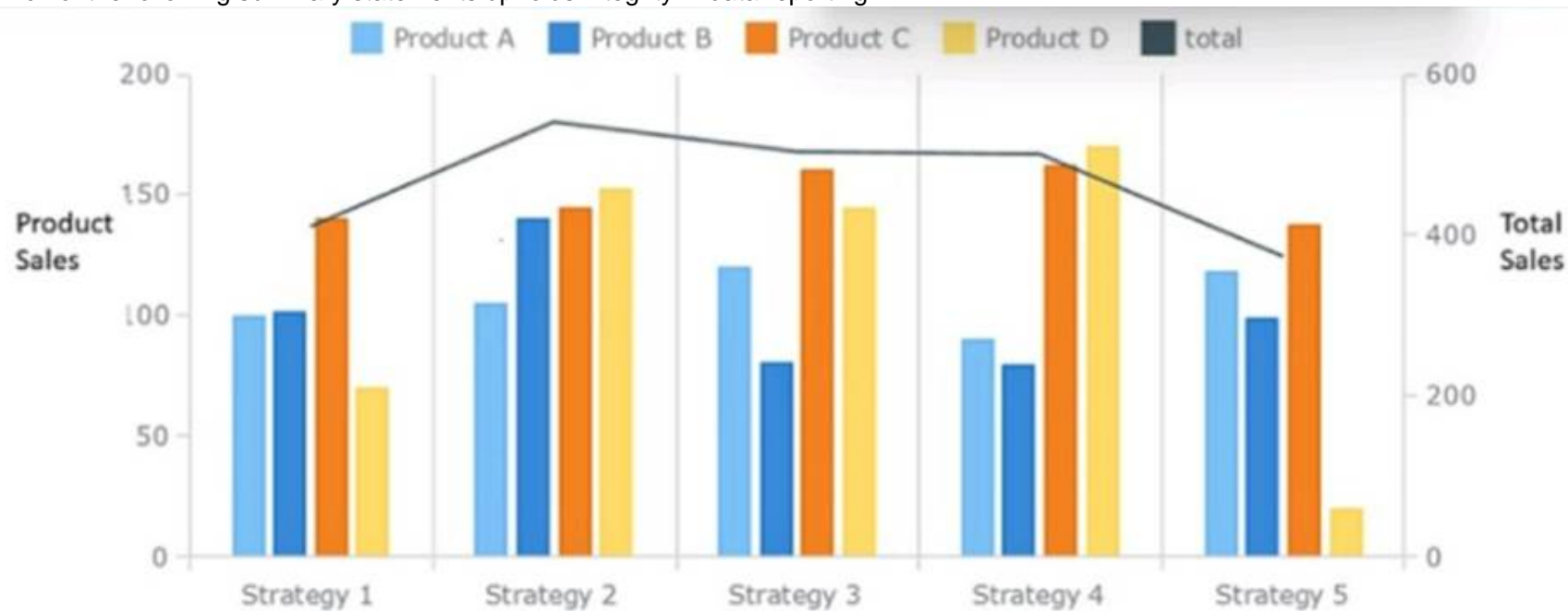
A sales director has requested a report for individual team members within the division be developed. The director would like the report to be shared with all team members, but individual team members should not be identifiable within the report Which of the following access requirements would support 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 20**

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 22**

A recurring event is being stored in two databases that are housed in different geographical locations. A data analyst notices the event is being logged three hours earlier in one database than in the other database. Which of the following is the MOST likely cause of the issue?

- A. The data analyst is not querying the databases correctly.
- B. The databases are recording different events.
- C. The databases are recording the event in different time zones.
- D. The second database is logging incorrectly.



**Answer:** C

**Explanation:**

The most likely cause of the issue is that the databases are recording the event in different time zones. A time zone is a region that observes a uniform standard time for legal, commercial, and social purposes. Different time zones have different offsets from Coordinated Universal Time (UTC), which is the primary time standard by which the world regulates clocks and time. For example, UTC-5 is five hours behind UTC, while UTC+3 is three hours ahead of UTC. If an event is being stored in two databases that are housed in different geographical locations with different time zones, it may appear that the event is being logged at different times, depending on how the databases handle the time zone conversion. For example, if one database records the event in UTC-5 and another database records the event in UTC+3, then an event that occurs at 12:00 PM in UTC-5 will appear as 9:00 AM in UTC+3. The other options are not likely causes of the issue, as they are either unrelated or implausible. The data analyst is not querying the databases incorrectly, as this would not affect the time stamps of the events. The databases are not recording different events, as they are supposed to record the same recurring event. The second database is not logging incorrectly, as there is no evidence or reason to assume that. Reference: [Time zone - Wikipedia]

**NEW QUESTION 24**

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

Online transactions:

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

Store transactions:

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

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

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

**Answer:** A

**Explanation:**

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

References:

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

**NEW QUESTION 26**

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 28

Given the following data:

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

Which of the following BEST describes the data set?

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

**Answer: C**

#### Explanation:

This is because inconsistency is a type of data quality issue that occurs when the data does not follow a common format, structure, or rule across different sources or systems, which can affect the efficiency and performance of the analysis or process. Inconsistency can be caused by having different spellings, punctuations, capitalizations, or abbreviations for the same or similar values in a data set, such as ??M??, ??m??, ??Male??, or ??male?? for gender in this case. Inconsistency can be eliminated or reduced by using data cleansing techniques, such as standardizing or normalizing the data values. The other options are not correct descriptions of the data set. Here is why:

? Data bias is a type of data quality issue that occurs when the data is not representative or proportional of the population or the parameter, which can affect the validity and reliability of the analysis or process. Data bias can be caused by having a sample that is too small, too large, or too skewed for the population or the parameter, such as having only male customers for a product that targets both genders in this case. Data bias can be eliminated or reduced by using sampling techniques, such as stratified or cluster sampling.

? The data is incomplete is a type of data quality issue that occurs when the data is absent or missing in a data set, which can affect the accuracy and reliability of the analysis or process. The data is incomplete can be caused by various factors, such as human error, system error, or non-response. The data is incomplete can be addressed by using various methods, such as replacing or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression.

? The data is outliers is a type of data quality issue that occurs when the data has values that are unusually high or low compared to the rest of the data set, which can affect the quality and validity of the analysis or process. The data is outliers can be caused by various factors, such as measurement error, natural variation, or extreme events. The data is outliers can be addressed by using various methods, such as removing or filtering out the outliers, or using robust statistics that are less sensitive to outliers, such as median, interquartile range, or box plot.

#### NEW QUESTION 30

Given the table below:

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

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

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

**Answer: C**

**Explanation:**

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

**NEW QUESTION 34**

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

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

**Answer: C**

**NEW QUESTION 39**

Analytics reports should follow corporate style guidelines.

- A. True.
- B. False.

**Answer: A**

**NEW QUESTION 42**

A data analyst received the information in the table below from a recently completed marketing campaign:

Channels	Clicks	Orders
Display	580	55
PPC	800	100
Social	1,200	220
Mobile	300	60
SEO	620	85

Which of the following is the total order conversion rate?

- A. 13.2%
- B. 14.8%
- C. 22.3%
- D. 85.2%

**Answer: B**

**Explanation:**

The correct answer is A. 13.2%.

The total order conversion rate is the ratio of the total number of orders to the total number of clicks, expressed as a percentage. To calculate the total order conversion rate, we need to sum up the clicks and orders from all the channels, and then divide the orders by the clicks and multiply by 100.

Using the data from the table, we can do the following:

? Total clicks = 580 + 800 + 1,200 + 300 + 620 = 3,500

? Total orders = 55 + 100 + 220 + 60 + 85 = 520

? Total order conversion rate = (520 / 3,500) x 100 = 14.857%

? Rounding to one decimal place, we get 14.9% Therefore, the total order conversion rate is 14.9%.

**NEW QUESTION 47**

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

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

**Answer: C**



**Explanation:**

Answer C. To confirm important details before dashboard development begins.

A dashboard mockup is a prototype of a finished dashboard directly in the product. It is a way to visualize the layout, design, and functionality of the dashboard before it is built with real data and code. A dashboard mockup can help the analyst to confirm important details with the user, such as the business objectives, the key performance indicators, the data sources, the filters, the charts, and the interactivity. By creating a dashboard mockup, the analyst can get immediate feedback and validation from the user, and avoid wasting time and resources on developing a dashboard that does not meet the user's expectations or needs<sup>1</sup>.

**NEW QUESTION 48**

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 51**

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

- A. PII
- B. PCI
- C. PBI
- D. PHI

**Answer:** B

**NEW QUESTION 56**

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<sup>12</sup>  
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<sup>12</sup>

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)<sup>3</sup>

? A primary key does not allow any NULL values, whereas a unique key allows one NULL value for the column(s)<sup>123</sup>

? A primary key can be a unique key, but a unique key cannot be a primary key<sup>12</sup>

**NEW QUESTION 59**

An analyst is reviewing the following data: Car IDSpeed

123155  
566436  
564418  
650567  
546436  
645638

Which of the following should the analyst include in the measures of central tendency for speed?

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

**Answer:** D

**Explanation:**

The measures of central tendency include the mode, median, and mean. The mode is the value that appears most frequently in a data set. In this case, the speed of 36 appears twice, making it the mode. The median is the middle value when a data set is ordered from least to greatest; for these speeds, when ordered (18, 36, 36, 38, 55, 67), the median is the average of the two middle numbers, which is ( $\frac{36 + 38}{2} = 37$ ). The mean is the average of all values, calculated as ( $\frac{18 + 36 + 36 + 38 + 55 + 67}{6} = 41.7$ ). References:

? The calculation of the mode, median, and mean is based on standard statistical formulas and definitions.

The measures of central tendency for speed include the mode, median, and mean. To calculate these, we first need to organize the data:

? Speeds in ascending order: 18, 36, 36, 38, 55, 67

? Mode is the value that appears most frequently, which is 36, as it appears twice.

? Median is the middle value when the data is ordered. Since we have an even number of observations, we take the average of the two middle values (36 and 38), resulting in 37.

? Mean is the sum of all values divided by the number of values.  $(18+36+36+38+55+67)/6=41.5$

Thus, the correct option is D, which includes Mode = 36, Median = 37, and Mean = 41.5. The range, maximum, and minimum values, although useful in understanding data dispersion, are not measures of central tendency and are therefore not relevant to this specific question.

#### NEW QUESTION 60

An analyst reviews the following data: 7

3  
5  
2  
3  
7  
7  
10

Which of the following is the value of the mode?

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

**Answer:** C

#### Explanation:

The mode is the value that appears most frequently in a data set. In the provided data set, the number 7 appears three times, which is more than any other number. Therefore, the mode of this data set is 7.

? 3 appears twice, but less frequently than 7.

? 5 and 10 each appear only once, so they cannot be the mode.

References:

? Mode in Statistics - Definition and Examples<sup>1</sup>

? Understanding Measures of Central Tendency<sup>2</sup>

? Mode (statistics) - Wikipedia<sup>3</sup>

#### NEW QUESTION 63

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

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

**Answer:** C

#### Explanation:

The best sampling method for the data analyst's need is C. Stratified sampling.

Stratified sampling is a type of probability sampling that involves dividing the population into homogeneous groups or strata based on some characteristic, such as state, and then randomly selecting a proportional number of individuals from each stratum. Stratified sampling ensures that every group is adequately represented in the sample, and reduces the sampling error and variability<sup>12</sup>

Systematic sampling is not correct, because it involves selecting every nth individual from the population, starting from a random point. Systematic sampling does not guarantee that every state will have a similar proportion of data in the sample, and may introduce bias or error if there is a hidden pattern or order in the population<sup>12</sup>

Convenience sampling is not correct, because it involves selecting individuals who are easily accessible or available to the researcher. Convenience sampling is a type of non-probability sampling that does not involve random selection, and may result in a biased or unrepresentative sample<sup>12</sup>

Random sampling is not correct, because it involves selecting individuals from the population at random, without any grouping or stratification. Random sampling may not produce a sample that has a similar proportion of data from every state, especially if the population is large or heterogeneous. Random sampling may also have a higher sampling error and variability than stratified sampling<sup>12</sup>

#### NEW QUESTION 64

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

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

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

- A. Imputing

- B. Transposing
- C. Parsing
- D. Concatenating

**Answer:** C

**Explanation:**

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

**NEW QUESTION 65**

Given the customer table below:

Customer_ID	Active_flag	Segment	Store_ID	Spend
004	N	Nursery	004C	\$7,000
009	Y	Prime	004A	\$2,000
008	N	Prime	004D	\$6,000
003	Y	Nursery	004U	\$1,000
002	Y	Prime	004S	\$2,000
001	N	Prime	004A	\$1,500
007	Y	Prime	004D	\$2,000

Which of the following chart types is the most appropriate to represent the average spending of active customers vs. inactive customers?

- A. Pie chart
- B. Heat graph
- C. Scatter plot
- D. Line chart

**Answer:** A

**Explanation:**

A Pie chart is the most suitable for representing the average spending of active customers versus inactive customers. Pie charts are effective for comparing parts of a whole, which makes them ideal for visually displaying the proportion of spend between two distinct groups. They are widely used to depict percentage distributions and are straightforward, allowing immediate analysis of the active vs. inactive customer spending distribution at a glance.

**NEW QUESTION 70**

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 74**

Which of the following differentiates a flat text file from other data types?

- A. Data is separated by a delimiter.
- B. Data is stored in defined rows.
- C. Data is defined with key-value pairs.

D. Data is housed in a markup language.

**Answer:** A

**Explanation:**

A flat text file is a type of data file that contains only plain text without any formatting or markup. Data in a flat text file is usually separated by a delimiter, which is a character that marks the boundary between different fields or values. For example, a comma-separated values (CSV) file is a flat text file that uses commas as delimiters. Other common delimiters are tabs, spaces, semicolons, and pipes. Therefore, the correct answer is A. References: Plain text - Wikipedia, Comparison of document markup languages - Wikipedia

**NEW QUESTION 78**

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

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

**Answer:** B

**NEW QUESTION 83**

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

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

**Answer:** D

**NEW QUESTION 87**

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 92**

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 94**

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 98**

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

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

**Answer: B**

**Explanation:**

The next step after understanding a business requirement for a data analysis report is to determine the data necessary for the analysis. This step involves identifying the data sources, variables, metrics, and dimensions that are relevant and sufficient to answer the business question or problem. This step also involves assessing the availability, quality, and accessibility of the data, and planning how to collect, clean, and prepare the data for analysis. The other options are not the next steps after understanding a business requirement, but rather subsequent steps in the data analysis process. Rephrasing the business requirement is a step that can help clarify and refine the business question or problem before determining the data necessary for the analysis. Building a mock dashboard/presentation layout is a step that can help design and visualize the report before performing the data analysis. Performing exploratory data analysis is a step that can help explore and summarize the data before drawing conclusions and recommendations from the data. Reference: Data Analysis Process - DataCamp

**NEW QUESTION 100**

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

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

**Answer: B**

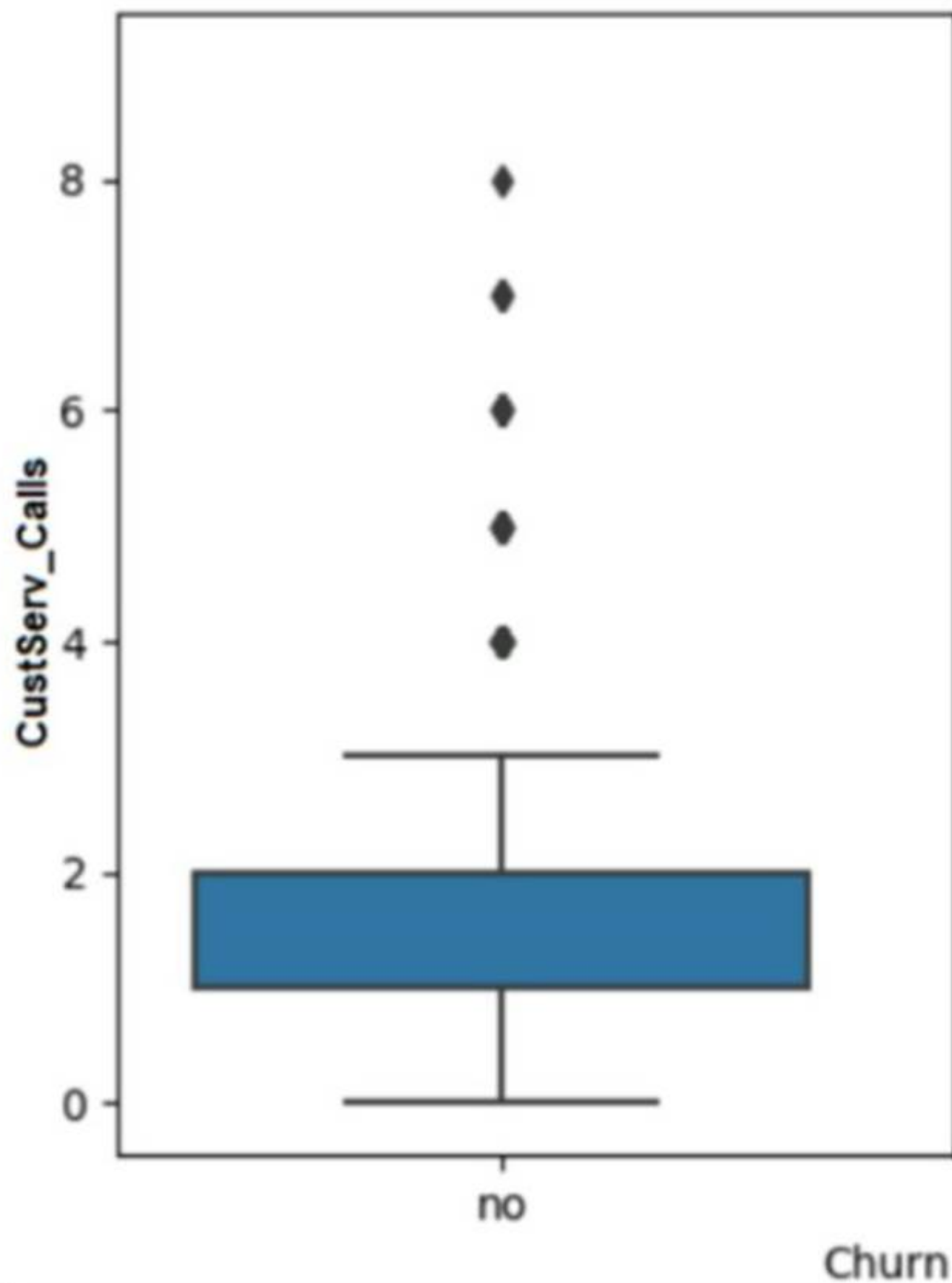
**Explanation:**

For a high-level, year-end report requested by a Chief Executive Officer (CEO), a recurring report type is most appropriate. Recurring reports are regular, scheduled reports that typically summarize information over a set period, such as a fiscal year. They provide a consistent format for executives to track performance over time, and their standardized nature makes them suitable for high-level analysis and decision-making. Since CEOs need to monitor performance and make strategic decisions, a recurring report that provides a comprehensive overview of the year's activities and outcomes would be valuable. This allows the CEO to evaluate the company's performance against its goals and objectives systematically.

Dynamic reports (A) are more interactive and typically used for in-depth analysis where users can drill down into the data. Ad hoc reports (C) are one-time, usually unscheduled reports tailored for specific questions, which may not be as comprehensive as a year-end report requires. Self-service reports (D) allow users to create their reports on demand, which may not be the formal, synthesized view a CEO would need for a year-end report.

**NEW QUESTION 104**

Given the image below:



The data should be cleaned because of the presence of:

- A. outlier
- B. non-parametric data.
- C. multicollinearity.
- D. invalid data.

**Answer: A**

**Explanation:**

The answer is A. Outlier.

Short Explanation: An outlier is a data point that differs significantly from the rest of the data in a dataset. An outlier can indicate an error, an anomaly, or a rare event in the data. An outlier can affect the statistical analysis and visualization of the data, such as skewing the mean, variance, or distribution of the data. Therefore, data should be cleaned to identify and remove or correct any outliers.

The image below shows a box plot graph with a vertical axis labeled "Customer Calls" and a horizontal axis labeled "Churn". The box plot is blue in color and the median value is around 2. There are 7 outliers above the box plot, ranging from 4 to 8. image)

A box plot is a type of graph that can show the distribution of data values using five summary statistics: minimum, maximum, median, first quartile, and third quartile. The box represents the interquartile range (IQR), which is the difference between the first and third quartiles. The median is shown as a line inside the box. The whiskers extend from the box to the minimum and maximum values, excluding any outliers. Outliers are shown as dots or circles outside the whiskers. In this graph, we can see that most of the customer calls are between 0 and 4, with a median of 2. However, there are 7 outliers that have more than 4 customer calls, up to 8. These outliers may indicate some customers who have more issues or complaints than others, or some errors or anomalies in the data collection or recording process. These outliers can affect the analysis and interpretation of the customer calls and churn relationship, such as making it seem that more customer calls lead to less churn, which may not be true for the majority of the customers. Therefore, data should be cleaned to investigate and handle these outliers appropriately.

**NEW QUESTION 108**

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

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

**Answer:** C

#### NEW QUESTION 111

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 114

Alex wants to use data from his corporate sale, CRM, and shipping systems to try and predict future sales. Which of the following systems is the most appropriate? Choose the best answer.

- A. Data mart.
- B. OLAP.
- C. Data Warehouse.
- D. OLTP.

**Answer:** C

#### Explanation:

Correct Answer: C. Data Warehouse.

Data warehouse bring together data from multiple systems used by an organization. A data mart is too narrow, as Alex needs data from across multiple divisions. OLAP is a broad term of analytical processing, and OLTP systems are transactional and not ideal for this task.

#### NEW QUESTION 116

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

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

**Answer:** A

#### Explanation:

This is because web scraping is a technique that allows the analyst to extract data from web pages, such as articles from a digital news outlet. Web scraping can be done using various tools and methods, such as Python libraries, browser extensions, or online services. The other techniques are not suitable for acquiring data from web pages. Here is why:

Sampling is a technique that involves selecting a subset of data from a larger population, usually for statistical analysis or testing purposes. Sampling does not help the analyst to acquire data from web pages, but rather to reduce the amount of data to be analyzed. Data wrangling is a technique that involves transforming and cleaning data to make it suitable for analysis or visualization. Data wrangling does not help the analyst to acquire data from web pages, but rather to improve the quality and usability of the data.

ETL stands for Extract, Transform, and Load, which is a process that involves moving data from one or more sources to a destination, such as a data warehouse or a database. ETL does not help the analyst to acquire data from web pages, but rather to store and organize the data.

#### NEW QUESTION 119

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

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

**Answer:** A

#### Explanation:

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



#### NEW QUESTION 121

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

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

**Answer:** BC

#### NEW QUESTION 126

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

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

**Answer:** C

#### Explanation:

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

#### NEW QUESTION 130

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 135

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 136

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 141

A data analyst is helping a retail store categorize its customers into five different groups based on the following information:

- How recently the customers made purchases
- How frequently the customers made purchases
- How much the customers spent Given the following information:

Customer_ID	Channel	Order_Date	Quantity	Territory	Amount (\$)
1001	Online	2/11/2020	12	North	1,250
2001	Store	2/10/2020	31	East	5,000
4001	Online	2/09/2020	24	West	2,500
3001	Online	2/11/2020	51	South	6,000
1001	Store	3/10/2020	22	North	2,000
1001	Online	1/09/2020	87	North	8,400
1001	Store	2/09/2020	23	North	2,000

Which of the following would be most important for the analysis?

- A. CustomerJ
- B. Channel, Order\_Date
- C. CustomerJD, Territor
- D. Amount
- E. CustomerJD, Order\_Dat
- F. Amount
- G. CustomerJ
- H. Quantity, Amount

**Answer: C**

#### NEW QUESTION 142

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

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

Which of the following functions did the analyst use to generate the data in the Sales\_indicator column?

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

**Answer: B**

#### Explanation:

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

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

The other functions are not suitable for generating the data in the Sales\_indicator column. Here is why:

Aggregate is a type of function that performs a calculation on a group of values, such as sum, average, count, etc. An aggregate function cannot generate the data in the Sales\_indicator column because it does not compare the values in the Sales column with a threshold or return a text value based on a condition.

Date is a type of function that manipulates or extracts information from dates, such as year, month, day, etc. A date function cannot generate the data in the Sales\_indicator column because it does not use the values in the Sales column or return a text value based on a condition.

Sort is a type of function that arranges the values in a column or a range in ascending or descending order. A sort function cannot generate the data in the Sales\_indicator column because it does not create a new column or return a text value based on a condition.

**NEW QUESTION 147**

A data analyst is compiling a report that a Chief Executive Officer needs for an impromptu meeting. The report should include information on the previous day's performance. Which of the following reports should the analyst provide?

- A. Tactical
- B. Ad hoc
- C. Dynamic
- D. Recurring

**Answer:** B

**NEW QUESTION 149**

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 153**

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

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

**Answer:** D

**Explanation:**

This is because stratified is a type of sampling in which elements of data are selected randomly from each of the small subgroups within a population, such as age



groups, gender groups, or income groups. Stratified sampling can be used to ensure that the sample is representative and proportional of the population, as well as reduce the sampling error or bias. For example, stratified sampling can be used to select a sample of voters from different political parties based on their proportion in the population. The other types of sampling are not the types of sampling in which elements of data are selected randomly from each of the small subgroups within a population. Here is why:

? Simple random is a type of sampling in which elements of data are selected randomly from the entire population, without dividing it into any subgroups. Simple random sampling can be used to ensure that every element in the population has an equal chance of being selected, as well as avoid any systematic error or bias. For example, simple random sampling can be used to select a sample of students from a school by using a lottery or a computer-generated number.

? Cluster is a type of sampling in which elements of data are selected randomly from a few large subgroups within a population, such as regions, districts, or schools. Cluster sampling can be used to reduce the cost and complexity of sampling, as well as increase the feasibility and convenience of sampling. For example, cluster sampling can be used to select a sample of households from a few neighborhoods by using a map or a list.

? Systematic is a type of sampling in which elements of data are selected at regular intervals from an ordered list or sequence within a population, such as every nth element or every kth element. Systematic sampling can be used to simplify and speed up the sampling process, as well as ensure that the sample covers the entire range or scope of the population. For example, systematic sampling can be used to select a sample of books from a library by using an alphabetical order or a numerical order.

#### NEW QUESTION 155

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 159

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

**Sales\_table**

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

**Order\_table**

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

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

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

**Answer:** D

#### Explanation:

The correct answer is D. OUTER JOIN, seven rows.

An OUTER JOIN is a type of SQL join that returns all the rows from both tables, regardless of whether there is a match or not. If there is no match, the missing side will have null values. An OUTER JOIN can be either a LEFT JOIN, a RIGHT JOIN, or a FULL JOIN, depending on which table's rows are preserved. Using the example tables, a FULL OUTER JOIN query would look like this:

SELECT Cust\_id, Order\_id, Order\_qty FROM Sales\_table FULL OUTER JOIN Order\_table ON Sales\_table.Order\_id = Order\_table.Order\_id;

The result of this query would be:

Cust\_id | Order\_id | Order\_qty  
 75 NULL | 5 | 10 NULL | 6 | 20 NULL | 7 | 15

As you can see, the query returns seven rows, one for each order in either table. The orders that are not in the Sales\_table have null values for the Cust\_id



column.  
To find the mean of the order quantity, we need to sum up the order quantities and divide by the number of rows. In this case, the mean is  $(100 + 50 + 25 + 75 + 10 + 20 + 15) / 7 = 42.14$ . Rounding to one decimal place, we get 42.1 as the mean of the order quantity.

**NEW QUESTION 161**

An analyst is reporting on the average income for a county and is reviewing the following data:

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

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

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

**Answer:** B

**NEW QUESTION 162**

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

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 164**

Which of the following describes the use of a representative amount of data from a main repository?

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

**Answer:** D

**Explanation:**

Sampling refers to the process of selecting a representative subset of data from a larger data set or repository. This technique is used when it is impractical or unnecessary to analyze the entire set of data. A representative sample should accurately reflect the characteristics of the larger population, allowing for analysis and inference about the population as a whole<sup>12</sup>.  
Observation (A) generally refers to the act of monitoring or recording data. Delta load (B) is a term used in data warehousing to describe the process of loading only the changes since the last data extraction, rather than the entire data set. Web scraping © is the process of extracting data from websites.  
References:  
? Understanding the importance of data sampling<sup>1</sup>.  
? The concept of a representative sample in statistics<sup>2</sup>.  
? Data repository management and usage<sup>3</sup>.  
? Benefits and methods of data sampling<sup>4</sup>.

**NEW QUESTION 169**

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

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

**Answer:** B

**Explanation:**

Creating a new report that automatically calculates weekly totals would streamline the process for the recipient. By setting this report to run at the end of business on Friday, it would provide the recipient with the necessary information for the entire week in one consolidated document. This eliminates the need for manual calculations and combines the previous week's data into one report, making it more efficient and less time-consuming.

References:

? Best practices in business analytics suggest automating repetitive tasks and consolidating reports where possible to improve efficiency and reduce the potential for human error.

**NEW QUESTION 171**

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 172**

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

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

**Answer:** C

**Explanation:**

A telephone number, despite being composed of digits, is not used for calculations and often includes formatting characters such as hyphens (-). Therefore, the most appropriate data type for a telephone number is Text (oVr ARCHAR in SQL databases), which can accommodate various formats and lengths, and preserve leading zeros that might be present in some phone numbers. Storing phone numbers as numeric data types would strip away any formatting and could lead to the loss of significant leading digits (like zeros in international numbers).

? Boolean is a binary data type and only represents true or false values.

? Date is a data type used for dates.

? Number could technically store phone numbers, but it is not suitable due to the reasons mentioned above.

References:

? Best Practices for Storing Phone Numbers<sup>1</sup>

? Data Types in SQL for Phone Numbers<sup>2</sup>

**NEW QUESTION 174**

Which of the following value is the measure of dispersion "range" between the scores of ten students in a test.

The scores of ten students in a test are 17, 23, 30, 36, 45, 51, 58, 66, 72, 77.

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

**Answer:** B

**Explanation:**

The correct answer is: 60

Range is the interval between the highest and the lowest score.

Range is a measure of variability or scatteredness of the varieties or observations among themselves and does not give an idea about the spread of the observations around some

central value. Symbolically  $R = H_s - L_s$ .

Where  $R$  = Range;  $H_s$  is the 'Highest score' and  $L_s$  is the Lowest Score.

The scores of ten students in a test are: 17, 23, 30, 36, 45, 51, 58, 66, 72, 77. The highest score is 77 and the lowest score is 17.

So the range is the difference between these two scores  $\text{Range} = 77 - 17 = 60$

**NEW QUESTION 176**

Given the information in the following tables:

## Online transactions:

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

## In-store transactions:

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

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

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

**Answer: D**

### Explanation:

Merging tables to create a master file that includes all transactions for both online and in-store sales is best described as data consolidation. This process involves combining data from various sources into a single, unified dataset. Data consolidation is essential for providing a comprehensive view of all transactions, which can be used for analysis, reporting, and decision-making purposes.

References: The answer is based on standard data management practices and the definition of data consolidation. No specific external documents were referenced for this response.

### NEW QUESTION 181

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 185

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

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

**Answer: C**

### NEW QUESTION 187

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 189

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

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

Which of the following is the mean?

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

Answer: C

Explanation:

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

NEW QUESTION 191

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 193

A data analyst reviews the following data set:



1
3
5
7
14
10
9
10
10

Which of the following is the range value?

- A. 9
- B. 10
- C. 12
- D. 13

**Answer:** D

#### NEW QUESTION 197

A web developer wants to ensure that malicious users can't type SQL statements when they asked for input, like their username/userid. Which of the following query optimization techniques would effectively prevent SQL Injection attacks?

- A. Indexing.
- B. Subset of records.
- C. Temporary table in the query set.
- D. Parametrization.

**Answer:** D

#### Explanation:

The correct answer is D: Parametrization. Parameterized SQL queries allow you to place parameters in an SQL query instead of a constant value. A parameter takes a value only when the query is executed, allowing the query to be reused with different values and purposes. Parameterized SQL statements are available in some analysis clients, and are also available through the Historian SDK. For example, you could create the following conditional SQL query, which contains a parameter for the collector's name: `SELECT* FROM ExamsDigest WHERE coursename=? ORDER BY tagname` SQL Injection is best prevented through the use of parameterized queries.

#### NEW QUESTION 198

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

- A. Data owner.
- B. Data steward.
- C. Data custodian.
- D. Data processor.

**Answer:** B

#### Explanation:

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

#### NEW QUESTION 200

Which of the following is a KPI metric for tracking sales performance?

- A. Order status percentage
- B. Customer acquisition percentage

- C. Gross profit percentage
- D. Click-through rate percentage

**Answer:** C

**Explanation:**

Gross profit percentage is a key performance indicator (KPI) that measures the profitability of a company's sales by showing the percentage of revenue that exceeds the cost of goods sold (COGS). It is a critical metric for tracking sales performance because it directly reflects the efficiency of a company in managing its production costs and the profitability of its products. This KPI is essential for understanding the financial health of a business and making informed decisions about pricing, cost control, and sales strategies.

References:

- ? Sales KPIs are essential for measuring the effectiveness of sales activities and the profitability of those efforts<sup>1</sup>.
- ? Gross profit percentage is highlighted as a crucial metric for assessing the financial success of sales initiatives<sup>2</sup>.
- ? Understanding the difference between sales metrics and KPIs, and the importance of gross profit percentage as a KPI<sup>1</sup>.
- ? The significance of gross profit percentage in evaluating sales team performance and guiding business decisions<sup>3</sup>.

**NEW QUESTION 203**

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 206

Given the following graph:



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

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

**Answer: B**

#### Explanation:

Strategy 4 provides the best sales in comparison to other strategies. This is because the total sales for Strategy 4 are the highest among all the strategies, as shown by the black line. The other statements are not accurate or do not uphold integrity in data reporting. Here is why:

Statement A is false because sales are not approximately equal for Product A and Product B across all strategies. For example, in Strategy 1, Product A has more sales than Product B, while in Strategy 3, Product B has more sales than Product A.

Statement C is misleading because it does not account for the difference in scale between the products. While Strategy 2 has the highest total sales among all products, it does not necessarily mean that it is the most effective for each product. For instance, Product D has very low sales in Strategy 2 compared to other strategies.

Statement D is biased because it does not provide any evidence or justification for why Product D should be promoted more than the other products in all strategies. It also ignores the fact that Product D has the lowest sales among all products in most of the strategies.

#### NEW QUESTION 209

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 213

An analyst wants to combine two data sets into a single spreadsheet. Column names from the first spreadsheet are listed in rows in the second spreadsheet. Which of the following is the first step the analyst should take to combine the data sets?

- A. Blend
- B. Merge
- C. Concatenate
- D. Transpose

**Answer: C**



#### NEW QUESTION 215

Five dogs have the following heights in millimeters: 300, 430, 170, 470, 600  
Which of the following is the mean height for the five dogs?

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

**Answer:** B

#### Explanation:

The mean height for the five dogs is 405mm. The mean, or average, is a measure of central tendency that represents the sum of all values divided by the number of values. To calculate the mean height for the five dogs, we can use the following formula:  $\text{Mean} = (300 + 430 + 170 + 470 + 600) / 5 = 2020 / 5 = 404$   
We can round up the result to the nearest millimeter, which is 405mm. The other options are not correct, as they are either too high or too low than the actual mean. Reference: [Mean - Math is Fun]

#### NEW QUESTION 220

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 221

You have two databases tables that you would like to join together using a foreign key relationship.  
What term best describes this action?

- A. Blending.
- B. Appending.
- C. Mixing.
- D. Merging.

**Answer:** D

#### Explanation:

Data merging is the process of combining two or more data sets into a single data set. Most often, this process is necessary when you have raw data stored in multiple files, worksheets, or data tables, that you want to analyze all in one go.

#### NEW QUESTION 226

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 227

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 228**

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 233**

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

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

Which of the following charts would be BEST to use?

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

Answer: B

Explanation:

A pie chart is the best choice to show the composition between the categories of the survey response data set. A pie chart represents the whole with a circle, divided by slices into parts. Each slice shows the relative size of each category as a percentage of the total. A pie chart is useful when the categories are mutually exclusive and add up to 100%. The table shows the favorite color and the number of responses for each color, which can be easily converted into percentages. A pie chart can show how each color contributes to the total number of responses.

Option A is incorrect because a histogram is used to show how data points are distributed along a numerical scale. The survey response data set is not numerical, but categorical. Option C is incorrect because a line chart is used to show trends or changes over time. The survey response data set does not have a time dimension.

Option D is incorrect because a scatter plot is used to show the relationship between two numerical variables. The survey response data set does not have two numerical variables. Option E is incorrect because a waterfall chart is used to show how an initial value is increased or decreased by a series of intermediate values. The survey response data set does not have an initial value or intermediate values.

References:

- ? How to Choose the Right Chart for Your Data - Infogram
- ? How to Choose the Right Data Visualization | Tutorial by Chartio
- ? Find the Best Visualizations for Your Metrics - The Data School
- ? How to choose the best chart or graph for your data

NEW QUESTION 234

A data analyst is working with a team to create a dashboard for a client who requires on- demand access. Which of the following is the best delivery method to support the clients?? requirement?

- A. Email
- B. Scheduled
- C. Subscription
- D. Static

Answer: C

Explanation:

The best delivery method to support the client??s requirement is C. Subscription.

Short Explanation: A subscription is a delivery method that allows the client to access the dashboard on-demand, whenever they need it. A subscription can be set up by the data analyst or the client themselves, and it can be configured to send an email notification when the dashboard is updated or refreshed. A subscription also allows the client to view the dashboard online or download it as a file format of their choice12

\* A. Email is not the best delivery method because it does not allow the client to access the dashboard on-demand. Email deliveries are sent at a fixed time or frequency, and they may not reflect the latest data or changes in the dashboard. Email deliveries also have limitations on the file size and format of the dashboard attachments1

\* B. Scheduled is not the best delivery method because it does not allow the client to access the dashboard on-demand. Scheduled deliveries are similar to email deliveries, except that they are triggered by a specific event or condition, such as a data update or a threshold value. Scheduled deliveries also have the same limitations as email deliveries on the file size and format of the dashboard attachments1

\* D. Static is not the best delivery method because it does not allow the client to access the dashboard on-demand. Static deliveries are one-time deliveries that are manually generated by the data analyst or the client. Static deliveries do not update or refresh automatically, and they may become outdated or irrelevant over time. Static deliveries also have limitations on the file size and format of the dashboard files3

#### NEW QUESTION 239

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

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

**Answer:** C

#### Explanation:

INSERT

The INSERT command is used to add new records to a database table.

The SELECT command is used to retrieve information from a database. It's the most commonly used command in SQL because it is used to pose queries to the database and retrieve the data that you're interested in working with.

The UPDATE command is used to modify rows in the database.

The CREATE command is used to create a new table within your database or a new database on your server.

#### NEW QUESTION 244

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

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

**Answer:** D

#### NEW QUESTION 248

Which of the following data governance concepts fits into the security requirements category?

- A. Data transmission
- B. Data deletion
- C. Data use agreements
- D. Personally identifiable information

**Answer:** D

#### NEW QUESTION 249

Which of the following is an example of PII?

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

**Answer:** B

#### Explanation:

A name is an example of personally identifiable information (PII), which is any data that can be used to identify someone, either on its own or with other relevant data. A name is a direct identifier, which means that it can uniquely identify a person without the need for any additional information. For example, a full name, such as John Smith, can be used to distinguish or trace an individual's identity<sup>1</sup>. Other examples of direct identifiers include:

- ? Social Security Number
- ? Passport number
- ? Driver's license number
- ? Email address
- ? Phone number

#### NEW QUESTION 254

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

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

**Answer:** D

#### Explanation:

The correct answer is D. Continuous data.

Continuous data is a type of quantitative data that can take any value within a range and can be measured with infinite precision. Continuous data can be expressed as fractions, decimals, or percentages. Examples of continuous data are height, weight, temperature, time, speed, etc<sup>12</sup>

The duration of a phone call in milliseconds is an example of continuous data, because it can take any value within a range (from zero to infinity) and can be measured with infinite precision (up to milliseconds or even smaller units). The duration of a phone call in milliseconds can also be expressed as fractions, decimals, or percentages of a larger unit (such as seconds, minutes, or hours).

Ordinal data is not correct, because ordinal data is a type of qualitative or categorical data that can be ordered or ranked according to some criterion. Ordinal data can have a logical order, but the intervals between the values are not equal or meaningful. Examples of ordinal data are grades, ratings, ranks, etc<sup>12</sup>



Nominal data is not correct, because nominal data is a type of qualitative or categorical data that can be labeled or named without any order or ranking. Nominal data can have a finite number of categories or classes, but the categories have no intrinsic value or hierarchy. Examples of nominal data are gender, color, nationality, etc<sup>12</sup>

Boolean data is not correct, because boolean data is a type of binary data that can have only two possible values: true or false. Boolean data can be used to represent logical statements, conditions, or outcomes. Examples of boolean data are yes/no, on/off, 1/0, etc.

#### NEW QUESTION 257

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 260

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

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

Customer Table -  
In-store Transactions –

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

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

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

**Answer: C**

**Explanation:**

An INNER JOIN returns only the rows that match the join condition in both tables. A LEFT JOIN returns all the rows from the left table, and the matched rows from the right table, or NULL if there is no match. In this case, the customer table is the left table and the in-store transactions table is the right table. The join condition is based on the customer\_id column, which is common in both tables.

To perform an INNER JOIN, we can use the following SQL query:

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

This query will return 9 rows of data, as shown below:

```
customer_id | name | lastname | gender | marital_status | transaction_id | amount | date
1 | MARC | TESCO | M | Y | 1 | 1000 | 2020-01-01
1 | MARC | TESCO | M | Y | 2 | 5000 | 2020-01-02
2 | ANNA | MARTIN | F | N | 3 | 2000 | 2020-01-03
2 | ANNA | MARTIN | F | N | 4 | 3000 | 2020-01-04
3 | EMMA | JOHNSON | F | Y | 5 | 4000 | 2020-01-05
4 | DARIO | PENTAL | M | N | 6 | 5000 | 2020-01-06
5 | ELENA | SIMSON | F | N | 7 | 6000 | 2020-01-07
6 | TIM | ROBITH | M | N | 8 | 7000 | 2020-01-08
7 | MILA | MORRIS | F | N | 9 | 8000 | 2020-01-09
```

To perform a LEFT JOIN, we can use the following SQL query:

```
SELECT * FROM customer LEFT JOIN in_store_transactions ON customer.customer_id = in_store_transactions.customer_id;
```

This query will return 15 rows of data, as shown below:

```
customer_id | name | lastname | gender | marital_status | transaction_id | amount | date
1 | MARC | TESCO | M | Y | 1 | 1000 | 2020-01-01
1 | MARC | TESCO | M | Y | 2 | 5000 | 2020-01-02
2 | ANNA | MARTIN | F | N | 3 | 2000 | 2020-01-03
2 | ANNA | MARTIN | F | N | 4 | 3000 | 2020-01-04
3 | EMMA | JOHNSON | F | Y | 5 | 4000 | 2020-01-05
4 | DARIO | PENTAL | M | N | 6 | 5000 | 2020-01-06
5 | ELENA | SIMSON | F | N | 7 | 6000 | 2020-01-07
6 | TIM | ROBITH | M | N | 8 | 7000 | 2020-01-08
7 | MILA | MORRIS | F | N | 9 | 8000 | 2020-01-09
8 | JENNY | DWARTH | F | Y | NULL | NULL | NULL
```

As you can see, the customers who do not have any transactions (customer\_id = 8) are still included in the result, but with NULL values for the transaction\_id, amount, and date columns.

Therefore, the correct answer is C: INNER: 9 rows; LEFT: 15 rows. Reference: SQL Joins - W3Schools

**NEW QUESTION 264**

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

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

Which of the following types of charts should be considered?

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

**Answer: C**

**Explanation:**

A scatter chart using sales volume and average sales per customer is the best type of chart to include in the dashboard. A scatter chart is a type of chart that displays the relationship between two numerical variables using dots or markers. A scatter chart can show how one variable affects another, how strong the correlation is between them, and how the data points are distributed. In this case, a scatter chart can show the story of sales and determine which site is providing the highest sales volume per customer by plotting the sales volume on the x-axis and the average sales per customer on the y-axis. Each dot on the chart will represent a site, and the analyst can easily compare the sites based on their position on the chart. A site with a high sales volume and a high average sales per customer will be in the upper right quadrant, indicating a high performance. A site with a low sales volume and a low average sales per customer will be in the lower left quadrant, indicating a low performance. A site with a high sales volume and a low average sales per customer will be in the lower right quadrant, indicating a high volume but low value. A site with a low sales volume and a high average sales per customer will be in the upper left quadrant, indicating a low volume but high value. A scatter chart can also show if there is a positive or negative correlation between the two variables, or if there is no correlation at all. A positive correlation means that as one variable increases, so does the other. A negative correlation means that as one variable increases, the other decreases. No correlation means that there is no relationship between the two variables.

The other types of charts are not as suitable for this purpose. A line chart is a type of chart that displays the change of one or more variables over time using lines. A line chart can show trends, patterns, and fluctuations in the data. However, in this case, there is no time variable involved, so a line chart would not be appropriate. A pie chart is a type of chart that displays the proportion of each category in a whole using slices of a circle. A pie chart can show how each category contributes to the total and compare the relative sizes of each category. However, in this case, there are two numerical variables involved, so a pie chart would not be able to show their relationship. A column chart is a type of chart that displays the comparison of one or more variables across categories using vertical bars. A column chart can show how each category differs from each other and rank them by size. However, in this case, a column chart would not be able to show the relationship between sales volume and average sales per customer, as it would only show one variable for each site.

**NEW QUESTION 267**

A user imports a data file into the accounts payable system each day. On a regular basis, the field input is not what the system is expecting, so it results in an error for the row and a broken import process. To resolve the issue, the user opens the file, finds the error in the row, and manually corrects it before attempting the import again. The import sometimes breaks on subsequent attempts, though. Which of the following changes should be made to this process to reduce the number of errors?

- A. Delete all incorrect inputs and upload the corrected file.
- B. Have the user manually review the file for data completeness before loading it
- C. Create a data field to data type validator to run the file through prior to import.
- D. Spot-check the file prior to import to catch and correct field errors.

**Answer: C**

**Explanation:**

A data field to data type validator is a tool or a process that checks if the data in each field of a file matches the expected data type, such as text, number, date, etc. A data field to data type validator can help to identify and correct any errors or inconsistencies in the data before importing it into the accounts payable system. This would reduce the number of errors and broken imports, as well as save time and effort for the user.

**NEW QUESTION 268**

An organization would like to add a secondary email field to its customer database in order to enrich the customer profiles. Which of the following data manipulation techniques should the analyst use to add this information?

- A. Blend
- B. Merge
- C. Append
- D. Aggregate

**Answer:** C

#### NEW QUESTION 272

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

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

**Answer:** B

#### NEW QUESTION 276

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

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

**Answer:** C

#### Explanation:

The correct answer is C. Indexing documents.

Indexing documents is a query optimization technique that involves creating a data structure that allows faster access to the data in the documents. Indexing documents can reduce the amount of data that needs to be scanned for a particular query, thus improving the performance and efficiency of the query. Indexing documents can also help with searching, sorting, filtering, and aggregating the data in the documents<sup>12</sup>

#### NEW QUESTION 281

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 283

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

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

**Answer:** C

#### Explanation:

The best reporting enhancement that the accounting department could implement to reduce the number of calls on current accounts is C. Increase the frequency of report generation.

By increasing the frequency of report generation, the accounting department could provide the collections manager with more up-to-date information on the customers who are past due on their accounts. This would help to avoid calling customers who have made payments in the time since the last report was printed, and thus reduce the number of calls on current accounts. Increasing the frequency of report generation would also improve the accuracy and timeliness of the data, and enhance the efficiency and effectiveness of the collections process.

Modifying the date range on the report, including a time stamp on the report, or adding a report run date to the report would not be sufficient to reduce the number of calls on current accounts. These enhancements would only provide information on when the report was generated or what period it covers, but they would not change the fact that the report could be outdated by the time it reaches the collections manager. Therefore, these enhancements would not solve the problem of calling customers who have already paid their accounts.

#### NEW QUESTION 288



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

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

**Answer:** BD

**Explanation:**

A data dictionary is a collection of metadata that describes the data elements in a database or dataset. It can help improve data acquisition by providing information about the data sources, formats, quality, and usage. It can also help remember specifics about data fields, such as their names, definitions, types, sizes, and relationships. Therefore, options B and D are correct.

Option A is incorrect because it is not a reason to create and maintain a data dictionary, but a benefit of doing so.

Option C is incorrect because specifying user groups for databases is not a function of a data dictionary, but a function of a database management system or a security policy.

Option E is incorrect because confining breaches of PHI data is not a function of a data dictionary, but a function of a data protection or encryption system.

Option F is incorrect because reducing processing power requirements is not a function of a data dictionary, but a function of a data compression or optimization system.

**NEW QUESTION 293**

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 295**

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 297**

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 302**

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 304**

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

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

**Answer:** A

**Explanation:**

The correct answer is A. UNION ALL.

UNION ALL is a SQL statement that appends two tables that have the same number of columns and compatible data types. UNION ALL preserves all the rows from both tables, including any duplicates<sup>12</sup>

\* B. MERGE is not correct, because MERGE is a SQL statement that combines the data of two tables based on a common column. MERGE can perform insert, update, or delete operations on the target table depending on the matching or non-matching rows from the source table<sup>34</sup>

\* C. GROUP BY is not correct, because GROUP BY is a SQL clause that groups the rows of a table based on one or more columns. GROUP BY is often used with aggregate functions, such as SUM, AVG, COUNT, etc., to calculate summary statistics for each group<sup>56</sup>

\* D. JOIN is not correct, because JOIN is a SQL clause that combines the data of two tables based on a common column or condition. JOIN can produce different results depending on the type of join, such as INNER JOIN, LEFT JOIN, RIGHT JOIN, etc.

**NEW QUESTION 307**

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

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 312**

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 315**

Which one of the following is a measure of dispersion?

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

D. Mean.

**Answer:** A

#### NEW QUESTION 320

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

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

**Answer:** A

#### Explanation:

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

#### NEW QUESTION 324

Which of the following best describes an exploratory analysis?

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

**Answer:** A

#### Explanation:

Answer A. Involves the use of descriptive statistics to understand observations. Exploratory data analysis (EDA) is a method of analyzing and investigating data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. EDA involves the use of descriptive statistics, such as mean, median, mode, standard deviation, frequency, or percentage, to understand the distribution, central tendency, variability, and relationship of the data. EDA helps to see what the data can reveal beyond the formal modeling or hypothesis testing, and provides a better understanding of data set variables and the interactions between them<sup>1</sup>.

#### NEW QUESTION 328

.....

## 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](#)**