

# 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**

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:

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

Therefore, option A is correct.

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

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

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

**NEW QUESTION 3**

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 4**

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 5**

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 6**

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

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

**Answer: D**

**Explanation:**

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

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

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

**NEW QUESTION 7**

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

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

**Answer: B**

**Explanation:**

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

**NEW QUESTION 8**

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

Original data:

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

Modified data:

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

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

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

**Answer:** B

**Explanation:**

The correct answer is B. Recoding.

Recoding is a data manipulation technique that involves changing the values or categories of a variable to make it more suitable for analysis. Recoding can be used to simplify or group the data, to correct errors or inconsistencies, or to create new variables from existing ones<sup>12</sup>

In the example, the analyst used recoding to change the values of Var001, Var002, Var003, and Var004 from numerical to textual form. The analyst also used recoding to assign meaningful labels to the values, such as ??Absent?? for 0, ??Present?? for 1, ??Low?? for 2, ??Medium?? for 3, and ??High?? for 4. This makes the data more understandable and easier to analyze.

#### NEW QUESTION 9

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

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

**Answer:** B

#### NEW QUESTION 10

Which of the following 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 10

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 12

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 16**

Analytics reports should follow corporate style guidelines.

- A. True.
- B. False.

**Answer:** A

**NEW QUESTION 20**

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 =  $(520 / 3,500) \times 100 = 14.857\%$

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

**NEW QUESTION 21**

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 25**

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 28**

Which of the following will MOST likely be streamed live?

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

**Answer:** A

**Explanation:**

Machine data is the most likely type of data to be streamed live, as it refers to data generated by machines or devices, such as sensors, web servers, network devices, etc. Machine data is often produced continuously and in large volumes, requiring real-time processing and analysis. Other types of data, such as key-value pairs, delimited rows, and flat files, are more likely to be stored in databases or files and processed in batches.

**NEW QUESTION 29**

When analyzing the values of two variables, you decide to convert both variables so they are on a scale of 0 to 1. What term describes this action?

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

**Answer:** B

**Explanation:**

Normalization is the process of reorganizing data in a database so that it meets two basic requirements: There is no redundancy of data, all data is stored in only one place. Data dependencies are logical, all related data items are stored together. Put simply, data normalization ensures that your data looks, reads, and can be utilized the same way across all of the records in your customer database. This is done by standardizing the formats of specific fields and records within your customer database.

**NEW QUESTION 32**

Which of the following data types would a telephone number formatted as XXX-XXX-XXXX be considered?

- A. Numeric
- B. Date
- C. Float
- D. Text

**Answer:** D

**Explanation:**

A telephone number formatted as XXX-XXX-XXXX would be considered a text data type, as it is composed of alphanumeric characters and symbols. A numeric data type is composed of only numbers, such as integers or decimals. A date data type is composed of values that represent dates or times, such as YYYY-MM-DD or HH:MM:SS. A float data type is composed of numbers with fractional parts, such as 3.14 or 0.5. Reference: Guide to CompTIA Data+ and Practice Questions - Pass Your Cert

**NEW QUESTION 33**

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 38

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

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

**Answer:** C

#### Explanation:

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

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

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

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

### NEW QUESTION 43

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 47

What SQL command is used to delete an entire table from a database?

- A. DROP.
- B. MODIFY.
- C. DELETE.
- D. ALTER.

**Answer:** A

### NEW QUESTION 50

Which of the following techniques is used to quantify data?

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

**Answer:** C

#### Explanation:

Answer C. Coding



Coding is a technique that is used to quantify data, especially qualitative data that are not expressed numerically. Coding involves assigning codes, such as numbers, letters, symbols, or colors, to different categories or themes that emerge from the data. For example, if you have a set of survey responses that ask about the satisfaction level of customers, you can code them as follows:

- ? Very satisfied = 5
- ? Satisfied = 4
- ? Neutral = 3
- ? Dissatisfied = 2
- ? Very dissatisfied = 1

By coding the data, you can convert them into quantitative data that can be analyzed using statistical methods, such as calculating the mean, median, mode, frequency, or percentage of each category<sup>12</sup>.

Option A is incorrect, as decoding is not a technique that is used to quantify data, but rather a process of interpreting or translating data from one form to another. For example, decoding can involve converting binary codes into text or images, or decrypting ciphertext into plaintext<sup>3</sup>.

Option B is incorrect, as enumeration is not a technique that is used to quantify data, but rather a process of listing or naming data in a specific order. For example, enumeration can involve listing the names of the states in alphabetical order, or naming the planets in order of their distance from the sun<sup>4</sup>.

Option D is incorrect, as structure is not a technique that is used to quantify data, but rather a property or characteristic of data that describes how they are organized or arranged. For example, structure can refer to the format, type, or schema of data, such as structured, semi-structured, or unstructured data.

#### NEW QUESTION 52

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 56

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 59

Which of the following is a non-parametric test?

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

**Answer:** D

#### Explanation:

The correct answer is D. Spearman's rank correlation.

Spearman's rank correlation is a non-parametric test that measures the strength and direction of the relationship between two variables that are ranked (ordinal) or continuous. Spearman's rank correlation does not assume that the data follows a normal distribution or that the variables are linearly related. Spearman's rank correlation is based on the ranks of the data rather than the actual values<sup>12</sup>

\* A. One-sample t-test is not correct, because it is a parametric test that compares the mean of a sample to a specified value. One-sample t-test assumes that the data follows a normal distribution and has a known population standard deviation<sup>34</sup>

\* B. Two-way ANOVA is not correct, because it is a parametric test that compares the means of two or more groups that are influenced by two independent factors. Two-way ANOVA assumes that the data follows a normal distribution, has homogeneous variances, and has independent observations.

\* C. Correlation coefficient is not correct, because it is a parametric test that measures the strength and direction of the linear relationship between two continuous variables. Correlation coefficient assumes that the data follows a bivariate normal distribution and has no outliers.

#### NEW QUESTION 64

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 69**

Which of the following best describes the law of large numbers?

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

**Answer:** B

**Explanation:**

The best answer is B. As a sample size grows, its mean gets closer to the average of the whole population.

The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population. This is due to the sample being more representative of the population as it increases in size. The law of large numbers guarantees stable long-term results for the averages of some random events<sup>1</sup>

\* A. As a sample size decreases, its standard deviation gets closer to the average of the whole population is not correct, because it confuses the concepts of standard deviation and mean. Standard deviation is a measure of how much the values in a data set vary from the mean, not how close the mean is to the population average. Also, as a sample size decreases, its standard deviation tends to increase, not decrease, because the sample becomes less representative of the population.

\* C. As a sample size decreases, its mean gets closer to the average of the whole population is not correct, because it contradicts the law of large numbers. As a sample size decreases, its mean tends to deviate from the average of the whole population, because the sample becomes less representative of the population.

\* D. When a sample size doubles, the sample is indicative of the whole population is not correct, because it does not specify how close the sample mean is to the population average. Doubling the sample size does not necessarily make the sample indicative of the whole population, unless the sample size is large enough to begin with. The law of large numbers does not state a specific number or proportion of samples that are indicative of the whole population, but rather describes how the sample mean approaches the population average as the sample size increases indefinitely.

**NEW QUESTION 70**

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 75**

Which of the following would be considered non-personally identifiable information?

- A. Cell phone device name
- B. Customer's name
- C. Government ID number
- D. Telephone number

**Answer:** A

**Explanation:**

Non-personally identifiable information (non-PII) is any data that cannot be used to identify, contact, or locate a specific individual, either alone or combined with other sources. Non-PII can include aggregated statistics, anonymous data, device identifiers, IP addresses, cookies, and other types of information that do not reveal the identity or location of a person. Cell phone device name is an example of non-PII, as it does not reveal any personal information about the owner or user of the device. Therefore, the correct answer is A. References: What is Non-Personally Identifiable Information (Non-PII)? | Definition and Examples, What is Personally Identifiable Information (PII)? | Definition and Examples

**NEW QUESTION 78**

Which of the following best describes the process of examining data for statistics and information about the data?



- A. Cleansing
- B. search
- C. Profiling
- D. Governance

**Answer: C**

**Explanation:**

Data profiling is the process of examining data for statistics and information about the data, such as the structure, format, quality, and content of the data. Data profiling can help to understand the characteristics, patterns, relationships, and anomalies of the data, as well as to identify and resolve any errors, inconsistencies, or missing values in the data. Data profiling can be done using various tools and methods, such as spreadsheets, databases, or programming languages<sup>12</sup>.

**NEW QUESTION 80**

A development company is constructing a new Init in its apartment complex. The complex has the following floor plans:

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

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

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

**Answer: D**

**Explanation:**

The correct answer is D. \$702,500.

To find the price of the Rose unit, we need to use the average cost per square foot of the original floor plans. The average cost per square foot is calculated by dividing the price by the square footage of each unit type. Using the data from the table, we can do the following:

? Jasmine:  $\$345,000 / 1,000 = \$345$  per square foot

? Orchid:  $\$525,000 / 2,000 = \$262.5$  per square foot

? Azalea:  $\$375,000 / 1,500 = \$250$  per square foot

? Tulip:  $\$450,000 / 1,800 = \$250$  per square foot

The average cost per square foot of the original floor plans is the mean of these four values, which is  $(\$345 + \$262.5 + \$250 + \$250) / 4 = \$276.875$  per square foot.

To find the price of the Rose unit, we need to multiply the average cost per square foot by the square footage of the Rose unit. The Rose unit has a square footage of 2,535, according to the table. Therefore, the price of the Rose unit is  $\$276.875 \times 2,535 = \$702,421.875$ .

Rounding to the nearest whole number, we get \$702,500 as the price of the Rose unit.

**NEW QUESTION 83**

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

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

**Answer: A**

**Explanation:**

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

? Remove duplicate data is a type of method that eliminates or reduces duplicate data, which is a type of data quality issue that occurs when data is repeated or copied in a data set. Removing duplicate data does not address missing data, but rather affects the quantity and validity of the data.

? Replace redundant data is a type of method that eliminates or reduces redundant data, which is a type of data quality issue that occurs when data is unnecessary or irrelevant for the analysis or purpose. Replacing redundant data does not address missing data, but rather affects the efficiency and performance of the analysis or process.



? Remove invalid data is a type of method that eliminates or reduces invalid data, which is a type of data quality issue that occurs when data is incorrect or inaccurate in a data set. Removing invalid data does not address missing data, but rather affects the validity and reliability of the analysis or process.

#### NEW QUESTION 87

??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 92

An analyst has written the following code: SELECT \*  
FROM Cust\_table  
WHERE age > 60 AND City = "New York"  
Which of the following criteria is the analyst retrieving?

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

**Answer:** C

#### Explanation:

The SQL query provided is selecting all records from the Cust\_table where the age column has values greater than 60 and the City column matches ??New York??. The > operator selects values that are strictly greater than the comparison value, so it does not include customers aged exactly 60. The term ??New York?? in the context of a city database typically refers to New York City, not the state of New York. Therefore, the correct answer is that the analyst is retrieving data for all customers older than age 60 in New York City.

References:

- ? The use of the > operator in SQL is to select values greater than the specified value1.
- ? Understanding the WHERE clause in SQL and its use in filtering records based on specified conditions2.
- ? Clarification on the distinction between city and state names in database records3.

#### NEW QUESTION 94

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 98

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

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

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

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

**Answer:** AC

**NEW QUESTION 103**

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 108**

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 113**

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 118**

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

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

**Answer:** C

**NEW QUESTION 119**

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 -----??-----??----- 1 | 1 | 100 2 | 2 | 50 3 | 3 | 25 4 | 4 | 75 NULL | 5 | 10 NULL | 6 | 20 NULL | 7 | 15

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

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

#### NEW QUESTION 124

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 125

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 128

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 132

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 137**

A data analyst is creating a dashboard and trying to identify the type of information that should be included. Which of the following should the analyst consider first?

- A. Data refresh rate
- B. Consumer types
- C. Access permissions
- D. Data sources and attributes

**Answer:** D

**Explanation:**

The answer is D. Data sources and attributes.

Short Explanation: The data analyst should consider the data sources and attributes first when creating a dashboard, because they determine what kind of information can be

included and how it can be displayed. The data sources and attributes define the origin, quality, format, and structure of the data that will be used for the dashboard. They also affect the data refresh rate, the consumer types, and the access permissions of the dashboard<sup>12</sup>

\* A. Data refresh rate is not the first thing to consider, because it depends on the data sources and attributes. The data refresh rate is how often the data in the dashboard is updated or refreshed to reflect the latest changes. The data refresh rate can vary depending on the type, frequency, and availability of the data sources<sup>1</sup>

\* B. Consumer types are not the first thing to consider, because they depend on the data sources and attributes. The consumer types are the intended audiences or users of the dashboard, who may have different needs, preferences, and expectations for the dashboard. The consumer types can influence the design, layout, and functionality of the dashboard. However, the consumer types cannot be determined without knowing what kind of data is available and relevant for them<sup>1</sup>

\* C. Access permissions are not the first thing to consider, because they depend on the data sources and attributes. The access permissions are the rules or policies that govern who can view, edit, or share the dashboard. The access permissions can protect the confidentiality, integrity, and availability of the data in the dashboard. However, the access permissions cannot be set without knowing what kind of data is involved and who needs to access it<sup>1</sup>

**NEW QUESTION 141**

Given the following customer and order tables:

Which of the following describes the number of rows and columns of data that would be present after performing an INNER JOIN of the tables?

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

**Answer:** B

**Explanation:**

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

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

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

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

The reason why there are seven rows and eight columns in the result table is because:

? There are seven rows because there are six customers and six orders in the original tables, but only five customers have matching orders based on the customer ID column. Therefore, only five rows will have data from both tables, while one row will have data only from the customer table (customer 5), and one row will have no data at all (null values).

? There are eight columns because there are four columns in each of the original tables, and all of them are selected and joined in the result table. Therefore, the result table will have four columns from the customer table (customer ID, first name, last name, and email) and four columns from the order table (order ID, order date, product, and quantity).

#### NEW QUESTION 145

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 147

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

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

**Answer: C**

#### NEW QUESTION 149

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 150**

Which of the following is a relational database?

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

**Answer:** A

**NEW QUESTION 151**

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 155**

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

- 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 160**

Standardized tests are given to students in the middle of each month, and the results are ready by the end of the month. The superintendent needs a quick view of test performance. Which of the following would be the best recommendation to meet the superintendent's requirements?

- A. A dashboard with a continuous data stream and saved searches
- B. A report of test scores by classroom, emailed to the superintendent at the end of the month
- C. A report of test scores with pie charts showing student performance
- D. A dashboard with a scheduled delivery, the ability to filter scores by school, and bar charts for comparison

**Answer:** D

**Explanation:**

A dashboard with a scheduled delivery is an efficient way to provide a quick view of test performance. It allows for timely updates, which is crucial given that the superintendent needs the information promptly at the end of each month. The ability to filter scores by school enables the superintendent to easily segment and analyze the data as needed. Bar charts are effective for comparison and can visually communicate the performance across different schools or other categories, making it easier to identify trends and outliers at a glance.

References:

? Best practices in data visualization recommend using dashboards for real-time data monitoring and quick access to key metrics<sup>1</sup>.

? Guidelines for presenting performance data suggest that visual tools like bar charts are helpful in comparing and analyzing data effectively<sup>1</sup>.



? Educational performance data analysis often involves comparing scores across different schools or classrooms, which is facilitated by a well-designed dashboard2.

**NEW QUESTION 165**

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 170**

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

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

**Answer:** B

**Explanation:**

Discrete data are data that can only assume specific values that are countable and distinct. For example, the number of books, the number of heads in a coin toss, or the number of patients in a hospital are discrete data. Discrete data cannot have fractional or decimal values, and there are clear spaces between the possible values12. Continuous data are data that can assume any value within a range and can be meaningfully divided into smaller parts. For example, the weight, height, length, time, or temperature are continuous data. Continuous data can have fractional or decimal values, and there are infinite numbers of possible values between any two points12.

**NEW QUESTION 171**

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

Table 1

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

Table 2

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



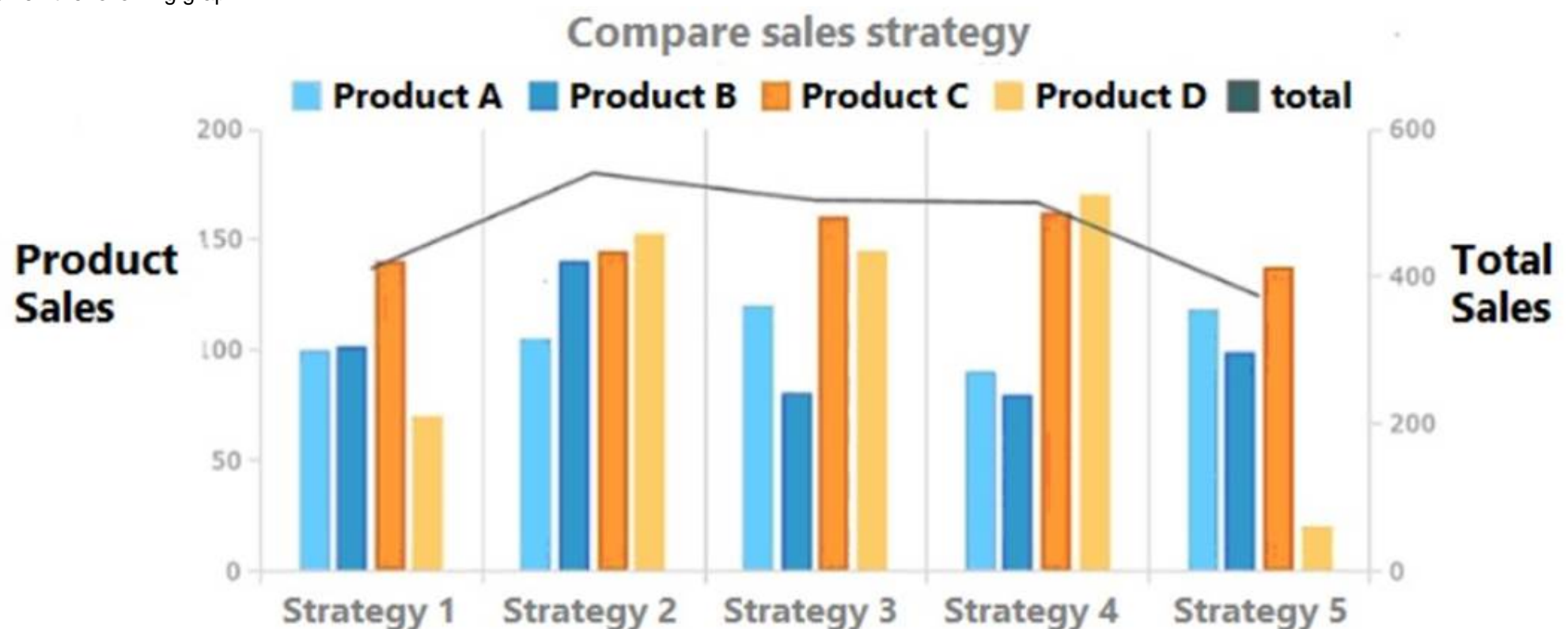
Which of the following steps should the analyst take to create the table?

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

**Answer: D**

#### NEW QUESTION 173

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 174

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

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

**Answer: D**

#### NEW QUESTION 176

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 178**

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 180**

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 184**

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

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

**Answer:** C

**NEW QUESTION 186**

A data analyst has been asked to create a sales report that calculates the rolling 12-month average for sales. If the report will be published on November 1, 2020, which of the following months should the report cover?

- A. October 1, 2019 to October 31, 2020
- B. October 31, 2020 to November 1, 2021
- C. November 1, 2019 to October 31, 2020
- D. October 31, 2019 to October 31, 2020

**Answer:** A

**Explanation:**

The report should cover the months from October 1, 2019 to October 31, 2020. A rolling 12-month average is a type of moving average that calculates the average of the last 12 months of data for each month. It is useful for smoothing out seasonal fluctuations and identifying long-term trends in the data. To calculate the rolling 12-month average for sales for November 1, 2020, the analyst needs to use the sales data from the previous 12 months, starting from November 1, 2019 and ending on October 31, 2020. The other options are either too short or too long to cover the required period.

**NEW QUESTION 188**

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 191**

A data analyst has been asked to organize the table below in the following ways: By sales from high to low -  
By state in alphabetic order -

First_name	Last_name	Address	City	State	Sales
Ed	Edens	2851 N. Southport	Chicago	IL	\$125,689
Pat	Mudd	710 Bridle Ridge Road	Eagan	MN	\$101,259
Katie	Hofstad	2851 S. Windwood Lane	Rosemount	NY	\$105,779
Edward	Frank	281 S. Northport	Chicago	IL	\$456,231
Rachel	Newman	305 Big Timber Trail	Wheaton	CO	\$99,876
Kaylyn	Korth	332 Richfield Drive	Lakeview	MN	\$166,874

Which of the following functions will allow the data analyst to organize the table in this manner?

- A. Conditional formatting
- B. Grouping
- C. Filtering
- D. Sorting

**Answer:** D

**Explanation:**

Sorting is the function that will allow the data analyst to organize the table in the desired manner. Sorting means arranging the data in a specific order, such as ascending or descending, based on one or more criteria. Sorting can be applied to any column in the table, such as sales or state. References: CompTIA Data+ Certification Exam Objectives, page 11

**NEW QUESTION 192**

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

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

**Answer:** C

**Explanation:**

Parameterization in SQL queries is a technique used to prevent SQL injection, which is a common security vulnerability that allows an attacker to interfere with the queries that an application makes to its database. By using parameterized queries, the database can distinguish between code and data, regardless of the input received. This method ensures that an attacker cannot change the intent of a query, even if SQL commands are inserted by the attacker. While parameterization can also affect performance by enabling consistent query execution plans, its primary purpose is to enhance security.

References:

- ? Medium article on SQL Query Optimization<sup>1</sup>.
- ? MSSQLTips on SQL Query Performance<sup>2</sup>.
- ? Blog post on SQL Performance Optimization<sup>3</sup>.
- ? SQL Easy guide on improving SQL Query Performance<sup>4</sup>.
- ? LearnSQL.com on SQL for Data Analysis<sup>5</sup>.

**NEW QUESTION 195**

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 199**

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 204

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

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

**Answer: D**

#### Explanation:

The median is recognized as the most appropriate measure of central tendency when outliers have been removed from a dataset. This is because the median is less influenced by extreme values compared to the mean. When outliers are present, they can significantly skew the mean, making it an unreliable measure of central tendency. The median, on the other hand, is the middle value of a dataset when ordered from least to greatest and remains unaffected by the extremes. Therefore, it provides a better representation of the central location of the data after outliers have been excluded.

References:

- ? Guidelines for Removing and Handling Outliers in Data<sup>1</sup>.
- ? Mean, Median, and Mode: Measures of Central Tendency<sup>2</sup>.
- ? Which measure of central tendency should be used when there is an outlier?<sup>3</sup>.
- ? How are measures of central tendency affected by outliers?<sup>4</sup>.

#### NEW QUESTION 205

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 not spend more on average.

#### NEW QUESTION 210

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 213

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 217

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 222

A data analyst is performing a data merge within a spreadsheet using the tables below:

<https://www.bing.com/images/blob?bcid=S1XCF9p02M4GjpbGxHj0Irlaj9sw.....4c>

Table 1

Last name	Sales
Knox	\$30
Johnson	\$10
Sinclair	\$70

Table 2

Last name	Address
Knox	2851 N. Southport
Johnson	467 Bridle Ridge
Sinclair	1067 Windwood Lane

The analyst is attempting to pull the addresses from Table 2 into Table 1 using the last names and is receiving an error message. Which of the following steps can the analyst perform to fix the error?

- A. Use concatenate to combine the tables.
- B. Ensure the formula is pulling from right to left.
- C. Sort the data by the last name field.
- D. Review the spelling and data type.

**Answer: D**

#### Explanation:

The error in merging data from Table 2 into Table 1 using last names could be due to discrepancies in spelling or data type between the two tables. It is essential to ensure that the last names are spelled consistently and that the data types are compatible for a successful merge. Option D suggests reviewing these aspects, which can potentially resolve the error, ensuring that each last name in Table 1 accurately corresponds to the same last name in Table 2, allowing for a successful data pull of addresses.

References: This answer is based on general data analytics practices and does not reference a specific document.

### NEW QUESTION 227

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

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

**Answer: C**

#### Explanation:

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

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

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

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

### NEW QUESTION 232

Given the table below:

Name	Gender	Level	Code	Region
James	Male	College	P	ON
Paul	Female	Elementary	A	BC
Sean	College	College	S	QC
Dad	Male	High school	D	AT
Nathan	Female	College	E	QC
Ahmed	Female	University	L	ON

Which of the following variables can be considered inconsistent, and how many distinct values should the variable have?

- A. Name, one
- B. Gender, two
- C. Level, three
- D. Code, four
- E. Region, five

**Answer: B**

#### Explanation:

The table provided shows an inconsistency in the ??Gender?? column, which lists three distinct values: Male, Female, and College. This is inconsistent because ??College?? is not a gender category. The ??Gender?? column should only have two distinct values, typically ??Male?? and ??Female??, to accurately represent gender data. This error could be due to a data entry mistake or a misclassification during data collection.

In data analysis, it's crucial to ensure that categorical variables like gender are consistent and correctly classified, as this can significantly impact the analysis results. Data cleaning processes often involve identifying and correcting such inconsistencies to maintain the integrity of the data set.

References:

- ? Data quality management principles emphasize the importance of consistency in data values, especially for categorical variables like gender<sup>1</sup>.
- ? Best practices in data cleaning include checking for and rectifying inconsistencies or misclassifications in data sets<sup>2</sup>.
- ? The importance of accurate data classification is highlighted in data analysis literature, as it directly affects the validity of the analysis results<sup>3</sup>.

### NEW QUESTION 236

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 240**

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

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 243**

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

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

**Answer:** A

**Explanation:**

Missing data is a type of data quality issue that occurs when some values in a data set are

not recorded or available. Missing data can affect the validity and reliability of survey results, especially if the missing values are not random or ignorable. Missing data can also reduce the sample size and the statistical power of the analysis<sup>12</sup>

If 100% participation is needed on survey results, a data analyst would look for missing data first, because missing data would indicate that some participants did not complete or submit the survey, or that some responses were not recorded or transmitted correctly. A data analyst would need to identify the causes and patterns of missing data, and apply appropriate methods to handle or prevent missing data, such as imputation, deletion, weighting, or follow-up<sup>12</sup>

**NEW QUESTION 247**

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 249**

An analyst in a consumer bank department wants to showcase the concentration of accounts opened in the United States by ZIP Code to describe the effectiveness of the bank's marketing campaigns. Which of the following would be the best way to visualize the data?

- A. A stacked chart
- B. A tree map
- C. A waterfall chart
- D. A geographic map

**Answer:** D

**NEW QUESTION 252**

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 256**

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 258**

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 262**

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

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

**Answer:** AE

**Explanation:**

? Data accuracy refers to the extent to which the data is correct, reliable, and free of errors. When different people manually type a series of handwritten surveys into an online database, there is a high chance of human error, such as typos, misinterpretations, omissions, or duplications. These errors can affect the quality and validity of the data and lead to incorrect or misleading analysis and decisions.

? Data consistency refers to the extent to which the data is uniform and compatible across different sources, formats, and systems. When different people manually type a series of handwritten surveys into an online database, there is a high chance of inconsistency, such as different spellings, abbreviations, formats, or standards. These inconsistencies can affect the integration and comparison of the data and lead to confusion or conflicts.

Therefore, to ensure data quality, it is important to have clear and consistent rules and procedures for data entry, validation, and verification. It is also advisable to use automated tools or methods to reduce human error and inconsistency.

**NEW QUESTION 267**

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 268**

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

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 273**

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

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 277**

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

ID	Name	LName	Phone	Email
1	Jacob	Smith	(303)445-2323	jsmith@abc.com
2	Hans	Williams	(302)546-4588	hws@emc.com
3	Martha	Dion	(304)254-6575	dion@mail.com
4	Jules	Martin	(300)563-3435	jmartinxyz.com
5	Sabrina	Huggins	(323)655-3475	shug@emc.com

Which of the following is an issue with this data?

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

**Answer:** C

**NEW QUESTION 282**

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

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

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

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

**Answer: C**

**Explanation:**

The conclusion that is accurate at a 95% confidence interval is that in general, users who visit the new website are more likely to make a purchase. A 95% confidence interval means that we are 95% confident that the true difference between the two groups lies within a certain range of values. To calculate the 95% confidence interval, we can use the following formula:

$$CI = (p1 - p2) \pm 1.96 * \sqrt{p * (1 - p) * (1/n1 + 1/n2)}$$

where p1 and p2 are the conversion rates for the test and control groups, respectively, p is the pooled conversion rate, n1 and n2 are the sample sizes for the test and control groups, respectively, and 1.96 is the z-score for a 95% confidence level.

Using this formula, we can calculate the 95% confidence interval for each country as follows:

Country | p1 | p2 | n1 | n2 | p | CI  
 United States | 0.12 | 0.11 | 2000 | 2000 | 0.115 | (-0.006, 0.026)  
 Germany | 0.06 | 0.04 | 1000 | 1000 | 0.05 | (-0.002, 0.042)  
 United Kingdom | 0.09 | 0.07 | 1500 | 1500 | 0.08 | (-0.003, 0.053)  
 France | 0.08 | 0.08 | 1200 | 1200 | 0.08 | (-0.024, 0.024)  
 Canada | 0.05 | 0.03 | 800 | 800 | 0.04 | (-0.005, 0.045)

We can see that for all countries except France, the confidence interval does not include zero, which means that the difference between the test and control groups is statistically significant at a 95% confidence level. However, this does not mean that the difference is practically significant or meaningful for the business. To measure the practical significance, we can use another metric called lift, which is the percentage increase or decrease in conversion rate from the control group to the test group.

$$Lift = (p1 - p2) / p2$$

Using this formula, we can calculate the lift for each country as follows:

Country | Lift  
 United States | 9.09%  
 Germany | 50%  
 United Kingdom | 28.57%  
 France | 0%  
 Canada | 66.67%

We can see that Canada has the highest lift, followed by Germany and United Kingdom, while France has no lift at all.

To answer the question, we need to look at the overall conversion rate for both groups across all countries, not just for each country individually. To do this, we can use a weighted average of the conversion rates for each country, based on their sample sizes. Weighted average = (p1 \* n1 + p2 \* n2) / (n1 + n2)

Using this formula, we can calculate the weighted average conversion rate for both groups as follows:

Group | Weighted average  
 Test | 0.084  
 Control | 0.072

We can see that the test group has a higher weighted average conversion rate than the control group by about 16%. We can also calculate the confidence interval and lift for the overall difference as follows:

$$CI = (p1 - p2) \pm 1.96 * \sqrt{p * (1 - p) * (1/n1 + 1/n2)} = (0.084 - 0.072) \pm 1.96 * \sqrt{0.078 * (1 - 0.078) * (1/2000 + 1/2000)}$$

**NEW QUESTION 283**

A database administrator needs to ensure only approved users can access specific database tables to perform financial functions. Which of the following is the best access control method for the administrator to use?

- A. Role-based
- B. Rule-based
- C. Discretionary
- D. Group-based

**Answer: A**

**NEW QUESTION 287**

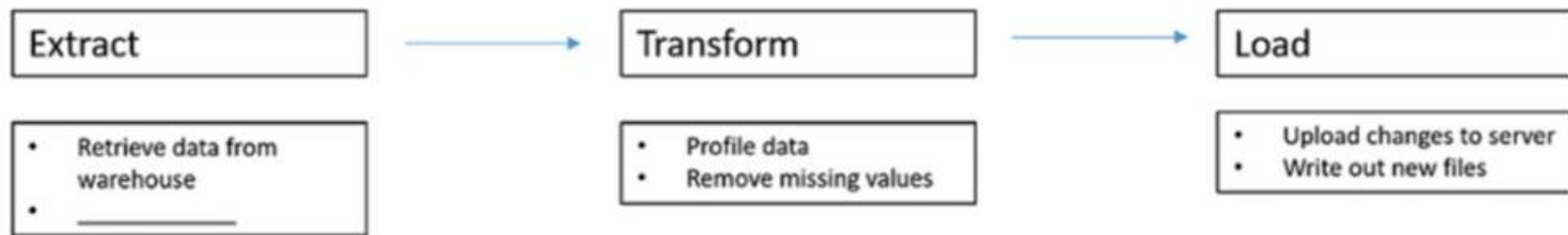
A company notifies its employees that emails will be automatically moved to a cloud-based server in 180 days. Which of the following describes this concept?

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

**Answer: C**

**NEW QUESTION 290**

Given the diagram below:



Which of the following steps is missing?

- A. Remove redundant data.
- B. Validate the data types.
- C. Connect to the data API.
- D. Normalize the data.

**Answer:** A

**Explanation:**

The missing step in the Extract, Transform, Load (ETL) process is typically the cleaning step, which involves removing redundant data or deduplication. This step is crucial in the ETL process to ensure that the data loaded into the destination is accurate and not inflated by duplicate records. The other options, like validating data types and connecting to the data API, are important but do not fit into the standard ETL process steps as a cleaning operation. Normalizing the data is part of the 'Transform' step, which was already listed.

**NEW QUESTION 291**

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

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 294**

Which of the following database schemas features normalized dimension tables?

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

**Answer:** B

**Explanation:**

The correct answer is B. Snowflake.

A snowflake schema is a type of database schema that features normalized dimension tables. A database schema is a way of organizing and structuring the data in a database. A dimension table is a table that contains descriptive attributes or characteristics of the data, such as product name, category, color, etc. A normalized table is a table that follows the rules of normalization, which is a process of reducing data redundancy and improving data integrity by organizing the data into smaller and simpler tables<sup>12</sup>

A snowflake schema is a variation of the star schema, which is another type of database

schema that features denormalized dimension tables. A denormalized table is a table that does not follow the rules of normalization, and may contain redundant or duplicated data. A star schema consists of a central fact table that contains quantitative measures or facts, such as sales amount, order quantity, etc., and several dimension tables that are directly connected to the fact table. A snowflake schema differs from a star schema in that the dimension tables are further split into sub-dimension tables, creating a snowflake-like shape<sup>13</sup>

A snowflake schema has some advantages and disadvantages over a star schema. Some advantages are:

? It reduces the storage space required for the dimension tables, as it eliminates the redundant data.

? It improves the data quality and consistency, as it avoids the update anomalies that may occur in denormalized tables.

? It allows more detailed analysis and queries, as it provides more levels of dimensions.

Some disadvantages are:

? It increases the complexity and number of joins required to retrieve the data from multiple tables, which may affect the query performance and speed.

? It reduces the readability and simplicity of the schema, as it has more tables and relationships to understand.

? It may require more maintenance and administration, as it has more tables to manage and update<sup>13</sup>

**NEW QUESTION 297**

Which of the following is used for calculations and pivot tables?

- A. IBM SPSS
- B. SAS



- C. Microsoft Excel
- D. Domo

**Answer:** C

**Explanation:**

This is because Microsoft Excel is a type of software application that allows users to create, edit, and analyze data in spreadsheets, which are composed of rows and columns of cells that can store various types of data, such as numbers, text, or formulas. Microsoft Excel can be used for calculations and pivot tables, which are two common features or functions in data analysis. Calculations are mathematical operations or expressions that can be performed on the data in the cells, such as addition, subtraction, multiplication, division, average, sum, etc. Pivot tables are interactive tables that can summarize and display the data in different ways, such as by grouping, filtering, sorting, or aggregating the data based on various criteria or categories. The other software applications are not used for calculations and pivot tables. Here is why:

IBM SPSS is a type of software application that allows users to perform statistical analysis and modeling on data sets, such as regression, correlation, ANOVA, etc. IBM SPSS does not use spreadsheets or cells to store or manipulate data, but rather uses data views or variable views to display the data in rows and columns. IBM SPSS does not have pivot tables as a feature or function, but rather has output views or charts to display the results of the analysis.

SAS is a type of software application that allows users to perform data management and analysis using a programming language that consists of statements and commands. SAS does not use spreadsheets or cells to store or manipulate data, but rather uses data sets or tables that are stored in libraries or folders. SAS does not have pivot tables as a feature or function, but rather has procedures or macros that can produce summary tables or reports based on the data.

Domo is a type of software application that allows users to create and share dashboards and visualizations that display data from various sources and systems, such as databases, cloud services, or web applications. Domo does not use spreadsheets or cells to store or manipulate data, but rather uses connectors or APIs to access and integrate the data from different sources. Domo does not have pivot tables as a feature or function, but rather has cards or widgets that can show different aspects or metrics of the data.

**NEW QUESTION 302**

A data analyst needs to apply quality control concepts to a data set for accuracy. Which of the following is the best way to do this?

- A. Standardization
- B. Parameterization
- C. Encryption
- D. Cross-validation

**Answer:** D

**NEW QUESTION 304**

Given the data below:

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

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

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 305**

What R package makes it easy to work with dates?

- A. Lubridate.
- B. Datemath.
- C. Stringr.
- D. ggplot.

**Answer:** A

**Explanation:**



Lubridate is an R package that makes it easier to work with dates and times.

#### NEW QUESTION 308

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

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

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

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

**Answer:** A

#### Explanation:

For a dashboard that aims to tell a story about sales and the sales customer ratio, a column chart is an effective choice. Column charts are particularly useful for showing data changes over a period of time or for illustrating comparisons among items. In this case, a column chart can clearly display the sales figures for each site, allowing for easy comparison across different sites. Additionally, it can be used to represent the sales customer ratio by showing the proportion of sales per customer, which can provide insights into customer behavior and sales effectiveness.

? Line charts are best suited for displaying data trends over time, rather than for comparing individual categories.

? Pie charts could show the proportion of sales for each site, but they are not as effective as column charts for comparing multiple categories.

? Scatter charts are used to show the relationship between two variables, which is not the focus in this scenario.

References:

? Effective Use of Column Charts1

? Choosing the Right Chart for Your Data2

? Sales Dashboards: Examples & Templates3

#### NEW QUESTION 310

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

Table 1

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

Table 2

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

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

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

**Answer:** B

**Explanation:**

The correct join the analyst should complete is B. LEFT JOIN, four rows.

A LEFT JOIN is a type of SQL join that returns all the rows from the left table, and the matched rows from the right table. If there is no match, the right table will have null values. A LEFT JOIN is useful when we want to preserve the data from the left table, even if there is no corresponding data in the right table.

Using the example tables, a LEFT JOIN query would look like this:

```
SELECT t1.Name, t1.City, t2.Age FROM Table1 t1 LEFT JOIN Table2 t2 ON t1.Name = t2.Name;
```

The result of this query would be:

Name City Age Jane Smith Detroit NULL John Smith Dallas 34 Candace Johnson Atlanta 45 Kyle Jacobs Chicago 39

As you can see, the query returns four rows, one for each name in Table1. The name John Smith appears twice in Table2, but only one of them is matched with the name in Table1. The name Jane Smith does not appear in Table2, so the age column has a null value for that row.

**NEW QUESTION 315**

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

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

**Answer:** DE

**Explanation:**

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

#### NEW QUESTION 318

A publishing group has requested a dashboard to track submissions before publication. A key requirement is that all changes are tracked, as multiple users will be checking out documents and editing them before submissions are considered final. Which of the following is the BEST way to meet this stakeholder requirement?

- A. Display the version number next to each submission on the dashboard.
- B. Present a data refresh date at the top of the dashboard.
- C. Confirm the dashboard is adhering to the corporate style guide.
- D. Use permissions to ensure users only see certain versions of the submissions.

**Answer:** A

#### Explanation:

A static report is a type of report that shows a snapshot of data at a specific point in time. A static report does not change or update automatically, unless the data source is refreshed or the report is regenerated. A static report is suitable for situations where the data does not change frequently or where historical data is needed for comparison or analysis. In this case, the data analyst is asked to create a sales report for the second-quarter 2020 board meeting, which will include a review of the business's performance through the second quarter. The board meeting will be held on July 15, 2020, after the numbers are finalized. This means that the data analyst does not need to show real-time or dynamic data, but rather a fixed and accurate view of the sales data for the second quarter. Therefore, a static report would be the best way to meet this stakeholder requirement. Therefore, the correct answer is A. References: What are Static Reports? | Sisense, Static vs Dynamic Reports - What's The Difference? | datapine

#### NEW QUESTION 319

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

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

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

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

**Answer:** C

#### Explanation:

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

#### NEW QUESTION 324

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

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

**Answer:** C

#### Explanation:

Answer: C. 2019 goal data

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



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

Option B is incorrect, as 2018 actual revenue is not relevant for determining 2019 performance attainment. The 2018 actual revenue contains the actual sales revenue for the year 2018, which is not comparable with the 2019 goal data or the 2019 actual revenue.

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

#### NEW QUESTION 327

Which of the following would be the best way to identify multicollinear attributes in a data set?

- A. Correlation coefficient
- B. Chi-squared test
- C. Two-sample f-test
- D. Two-way ANOVA

**Answer: A**

#### Explanation:

Multicollinearity in a dataset refers to the situation where two or more predictor variables are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In such cases, the correlation coefficient is a key statistical measure used to identify the presence of multicollinearity. It quantifies the degree to which two variables are linearly related.

The Variance Inflation Factor (VIF) is another commonly used metric that is derived from the correlation coefficient. It assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be equal to 1.

While the other options listed—Chi-squared test, Two-sample f-test, and Two-way ANOVA—are valuable statistical tools, they serve different purposes and are not typically used to detect multicollinearity. The Chi-squared test is used for testing relationships between categorical variables, the Two-sample f-test compares variances across groups, and Two-way ANOVA is used to understand the interaction between two independent categorical variables on a continuous dependent variable.

References:

? Multicollinearity in Regression Analysis: Problems, Detection, and Solutions<sup>1</sup>.

? What is multicollinearity and how to remove it?<sup>2</sup>.

? Detect and Treat Multicollinearity in Regression with Python<sup>3</sup>.

#### NEW QUESTION 329

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

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

**Answer: C**

#### Explanation:

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

| How to Open, Edit & Convert TSV Files]

#### NEW QUESTION 332

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

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



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

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

**Answer:** A

**Explanation:**

This is because a bar chart is a type of chart that shows the value or the amount of a single variable for different categories or groups, such as the percentage of new customers for different sites in this case. A bar chart can be used to display and analyze the comparison, ranking, or proportion among the categories or groups, as well as identify any differences, similarities, or outliers in the data. For example, a bar chart can show which site has the highest or lowest percentage of new customers, as well as show how much each site contributes to the total percentage of new customers. The other types of charts are not the best charts to display the data. Here is why:

? A line chart is a type of chart that shows the change or the trend of a single variable over time, such as the percentage of new customers over months or years in this case. A line chart can be used to display and analyze the movement, cycle, or pattern of the variable, as well as identify any peaks, valleys, or fluctuations in the data. For example, a line chart can show how the percentage of new customers increases or decreases over time, as well as show if there are any seasonal or periodic variations in the data.

? A pie chart is a type of chart that shows the proportion or the percentage of a single variable for different categories or groups, such as the percentage of new customers for different sites in this case. A pie chart can be used to display and analyze the composition, distribution, or share of the variable, as well as identify any segments, slices, or fractions in the data. For example, a pie chart can show how much each site represents of the total percentage of new customers, as well as show if there are any dominant or minor sites in the data.

? A scatter chart is a type of chart that shows the relationship between two variables for each observation or unit in a data set, such as the percentage of new customers and another variable for each site in this case. A scatter chart can be used to display and analyze the correlation, trend, or pattern among the variables, as well as identify any outliers or clusters in the data. For example, a scatter chart can show if there is a positive, negative, or no correlation between the percentage of new customers and another variable, such as sales revenue or customer satisfaction.

**NEW QUESTION 335**

.....

## 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](#)**