

CompTIA

Exam Questions DA0-001

CompTIA Data+ Certification Exam



NEW QUESTION 1

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

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

Answer: D

Explanation:

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

NEW QUESTION 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:

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

Therefore, option A is correct.

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

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

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

NEW QUESTION 3

A site reliability team wants to monitor the stability of their website. so they can proactively diagnose issues when they occur Which of the following deliverables would best suit their needs?

- A. A self-serve dashboard of website performance that updates in real time
- B. A weekly log report of site visits and user actions
- C. A portal that is refreshed daily and reports errors classified by type
- D. A daily summary email indicating website outages for the previous day

Answer: A

Explanation:

The best deliverable that would suit the site reliability team??s needs is A. A self-serve dashboard of website performance that updates in real time.

A self-serve dashboard is a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance. A self-serve dashboard of website performance that updates in real time would allow the site reliability team to easily and quickly access the information they need about the stability of their website, such as uptime, response time, error rate, traffic volume, etc. A self-serve dashboard would also enable the team to proactively diagnose issues when they occur, by providing alerts, notifications, or drill-down options. A self-serve dashboard would also be more interactive and engaging than a report or an email.

A weekly log report of site visits and user actions would not be a good deliverable for the site reliability team??s needs, because it would not provide timely or relevant information about the stability of their website. A weekly log report would be too infrequent and delayed to monitor and diagnose issues when they occur.

A weekly log report would also focus on the behavior and actions of the users, rather than the performance and functionality of the website.

A portal that is refreshed daily and reports errors classified by type would not be a good deliverable for the site reliability team??s needs, because it would not provide real-time or comprehensive information about the stability of their website. A portal that is refreshed daily would be too slow and outdated to monitor and diagnose issues when they occur. A portal that reports errors classified by type would be too narrow and limited to capture the full picture of the website performance.

A daily summary email indicating website outages for the previous day would not be a good deliverable for the site reliability team??s needs, because it would not provide real-time or actionable information about the stability of their website. A daily summary email would be too late and retrospective to monitor and diagnose issues when they occur. A daily summary email indicating website outages would also be too passive and generic to help the team resolve or prevent issues in the future.

NEW QUESTION 4

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

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

Answer: DF

NEW QUESTION 5

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

What best describes the data set she needs?

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

Answer: A

Explanation:

Correct answer A. Sample.

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

NEW QUESTION 6

Which of the following data cleansing issues will be fixed when a DISTINCT function is applied?

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

Answer: B

Explanation:

This is because duplicate data refers to data that is repeated or copied in a data set, which can affect the quality and validity of the analysis. A DISTINCT function is a type of function that removes duplicate values from a column or a table, leaving only unique values. For example, a DISTINCT function in SQL that can achieve this is:

```
SELECT DISTINCT column_name FROM table_name;
```

The other data cleansing issues will not be fixed by applying a DISTINCT function. Here is why:

Missing data refers to data that is absent or incomplete in a data set, which can affect the accuracy and reliability of the analysis. A DISTINCT function does not help with missing data, because it does not fill in or impute the missing values.

Redundant data refers to data that is unnecessary or irrelevant for the analysis, which can affect the efficiency and performance of the analysis. A DISTINCT function does not help with redundant data, because it does not remove or filter out the redundant values.

Invalid data refers to data that is incorrect or inaccurate in a data set, which can affect the validity and reliability of the analysis. A DISTINCT function does not help with invalid data, because it does not validate or correct the invalid values.

NEW QUESTION 7

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

Online transactions:

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

Store transactions:

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

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

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

Answer: A

Explanation:

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

field names in the actual datasets.

References:

? Best practices in data management.

? Principles of data integration and consolidation.

NEW QUESTION 8

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

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

Answer: C

NEW QUESTION 9

Which of following is a non-relational database?

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

Answer: A

Explanation:

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

NEW QUESTION 10

An analyst reviews the following data: 7

3
5
2
3
7
7
10

Which of the following is the value of the mode?

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

Answer: C

Explanation:

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

? 3 appears twice, but less frequently than 7.

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

References:

? Mode in Statistics - Definition and Examples¹

? Understanding Measures of Central Tendency²

? Mode (statistics) - Wikipedia³

NEW QUESTION 10

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

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

Answer: D

Explanation:

A star schema is a type of database schema that consists of one fact table and multiple dimension tables. The fact table contains the measures or metrics of the business process, such as sales, orders, or transactions. The dimension tables contain the attributes or characteristics of the business entities, such as products, customers, or locations. The fact table is connected to the dimension tables by foreign keys that reference the primary keys of the dimension tables. The fact table is located at the center of the schema, while the dimension tables are located at the edges, forming a star-like shape¹.

A star schema is an example of a denormalized schema, which means that the dimension tables are not normalized and may contain redundant or repeated data. This is done to improve the performance and simplicity of queries, as there are fewer joins and tables involved. A star schema is suitable for data warehouses and business intelligence applications that require fast and efficient data retrieval².

NEW QUESTION 14

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 19

A data analyst for a media company needs to determine the most popular movie genre. Given the table below:

MovieID	Name	Genre	Actors	Rating
01	Ghost Writer	Comedy, Actions	Joshua Wellington, Susana Summons	6.5
02	Life of Suffering	Drama, Foreign, Historical	Shelly May, Rita Moralle, Ethan Warner, Sean Houser	7.2

Which of the following must be done to the Genre column before this task can be completed?

- A. Append
- B. Merge
- C. Concatenate
- D. Delimit

Answer: D

Explanation:

Delimiting is the process of splitting a column of data into multiple columns based on a separator or delimiter character. Delimiting can help separate data that is combined or concatenated in one column into distinct values or categories. For example, if a column contains text values that are separated by commas, such as ??Comedy, Suspense??. Delimiting can split this column into two columns, one for ??Comedy?? and one for ??Suspense??. Delimiting is different from other options, such as appending, merging, or concatenating, which are methods of combining or joining data from multiple columns or sources. In this case, the data analyst needs to determine the most popular movie genre based on the Genre column in the table. However, this column contains multiple genres for each movie, separated by commas. Therefore, the data analyst must delimit this column before this task can be completed. Therefore, the correct answer is D. References: Split text into different columns with functions - Office Support, How to Split Text in Excel (Using Formulas & Split Function)

NEW QUESTION 23

Which of the following is a non-parametric test?

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

Answer: D

Explanation:

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

NEW QUESTION 27

Given the table below:

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

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

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

Answer: B

Explanation:

This is because date is a type of variable that represents a specific point or period in time, such as a day, a month, or a year. Date variables can be used to store, manipulate, or analyze temporal data, such as transaction dates, birth dates, or expiration dates. For example, date variables can be used to calculate the duration or the difference between two dates, or to filter or sort the data by date. The other variable types are not correct descriptions of the ??Year?? column. Here is why:

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

? Alphanumeric is a type of variable that represents a combination of alphabetic and numeric characters, such as letters, numbers, symbols, or spaces. Alphanumeric variables can be used to store, manipulate, or analyze textual data, such as names, addresses, or codes. For example, alphanumeric variables can be used to concatenate or split the data, or to search or match the data using patterns or expressions.

? Text is a type of variable that represents a sequence of alphabetic characters, such as letters or words. Text variables can be used to store, manipulate, or analyze textual data, such as names, categories, or labels. For example, text variables can be used to change the case or the length of the data, or to compare or classify the data using criteria or rules.

NEW QUESTION 31

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

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

Answer: B

Explanation:

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

NEW QUESTION 36

Kelly wants to get feedback on the final draft of a strategic report that has taken her six months to develop.

What can she do to get prevent confusion as see seeks feedback before publishing the report?

Choose the best answer.

- A. Distribute the report to the appropriate stakeholders via email.
- B. Use a watermark to identify the report as a draft.
- C. Show the report to her immediate supervisor.
- D. Publish the report on an internally facing website.

Answer: B

Explanation:

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

NEW QUESTION 40

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

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

Answer: C

Explanation:

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

NEW QUESTION 44

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

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

Answer: C

NEW QUESTION 48

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 52

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 54

SIMULATION

The director of operations at a power company needs data to help identify where company resources should be allocated in order to monitor activity for outages and restoration of power in the entire state. Specifically, the director wants to see the following:

* County outages

* Status

* Overall trend of outages **INSTRUCTIONS:**

Please, select each visualization to fit the appropriate space on the dashboard and choose an appropriate color scheme. Once you have selected all visualizations, please, select the appropriate titles and labels, if applicable. Titles and labels may be used more than once.

If at any time you would like to bring back the initial state of the simulation, please click the Reset All button.

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

This is a simulation question that requires you to create a dashboard with visualizations that meet the director's needs. Here are the steps to complete the task:

? Drag and drop the visualization that shows the county outages on the top left

space of the dashboard. This visualization is a map of the state with different colors indicating the number of outages in each county. You can choose any color scheme that suits your preference, but make sure that the colors are consistent and clear. For example, you can use a gradient of red to show the counties with more outages and green to show the counties with less outages.

? Drag and drop the visualization that shows the status of the outages on the top

right space of the dashboard. This visualization is a pie chart that shows the percentage of outages that are active, restored, or pending. You can choose any color scheme that suits your preference, but make sure that the colors are distinct and easy to identify. For example, you can use red for active, green for restored, and yellow for pending.

? Drag and drop the visualization that shows the overall trend of outages on the

bottom space of the dashboard. This visualization is a line graph that shows the number of outages over time. You can choose any color scheme that suits your preference, but make sure that the color is visible and contrasted with the background. For example, you can use blue for the line and white for the background.

? Select appropriate titles and labels for each visualization. Titles and labels may be

used more than once. For example, you can use "County Outages" as the title for the map, "Status" as the title for the pie chart, and "Trend" as the title for the line graph. You can also use "County", "Number of Outages", "Active", "Restored", "Pending", "Time", and "Number of Outages" as labels for the axes and legends of the visualizations.

NEW QUESTION 59

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

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

Answer: CF

Explanation:

Two reasons to conduct data cleansing are:

? To improve accuracy: Data cleansing helps to ensure that the data is correct, consistent, and reliable. This can improve the quality and validity of the analysis, as well as the decision-making and outcomes based on the data¹²

? To calculate trends: Data cleansing helps to remove or resolve any errors, outliers,

or missing values that could distort or skew the data. This can help to identify and measure the patterns, changes, or relationships in the data over time¹³

NEW QUESTION 60

Given the following data tables:

CustomerID	CustomerLastName
01	Manzelli
02	Kraus

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

ItemID	Customer_Last_Name	QuantityPurchased
01	Brown	03
02	Smee	07

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

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

Answer: A

Explanation:

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

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

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

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

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

NEW QUESTION 61

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 access¹².

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 insights¹².

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 66

A stakeholder wants to see daily sales targets organized in a dashboard by country, state, city, and ZIP Code. Which of the following delivery considerations must a data analyst take into account when creating the dashboard?

- A. Variable formatting
- B. Drill-down capability
- C. Saved searches
- D. Access permissions

Answer: B

NEW QUESTION 69

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

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

Answer: A

NEW QUESTION 71

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 73

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

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

Which of the following is the mean?

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

Answer: C

Explanation:

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

NEW QUESTION 76

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

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

Answer: C

Explanation:

The null hypothesis should be accepted when the p-value is greater than the alpha level, which is the significance level of the test. The p-value is the probability of obtaining a test statistic at least as extreme as the one observed in the sample, assuming that the null hypothesis is true. The alpha level is the probability of rejecting the null hypothesis when it is true, which is also known as a type I error¹².

In this case, the alpha level is 0.05, which means that there is a 5% chance of rejecting the null hypothesis when it is true. Therefore, to reject the null hypothesis, the p-value must be less than or equal to 0.05, which indicates that the test statistic is very unlikely to occur by chance under the null hypothesis. Conversely, to accept the null hypothesis, the p-value must be greater than 0.05, which indicates that the test statistic is not very unlikely to occur by chance under the null hypothesis.

Among the four options, only option D has a p-value that is greater than 0.05 ($p = 0.06$). Therefore, option D is the correct answer. When $p = 0.06$, it means that there is a 6% chance of obtaining a test statistic at least as extreme as the one observed in the sample, assuming that the null hypothesis is true. This probability is not very low, and therefore does not provide enough evidence to reject the null hypothesis.

NEW QUESTION 81

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

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

Answer: D

Explanation:

The correct answer is D: Parametrization. Parameterized SQL queries allow you to place parameters in an SQL query instead of a constant value. A parameter takes a value only when the query is executed, allowing the query to be reused with different values and purposes. Parameterized SQL statements are available in some analysis clients, and are also available through the Historian SDK.

For example, you could create the following conditional SQL query, which contains a parameter for the collector's name: `SELECT* FROM ExamsDigest WHERE coursename=? ORDER BY tagname` SQL Injection is best prevented through the use of parameterized queries.

NEW QUESTION 83

An analyst has received the requirements for an internal user dashboard. The analyst confirms the data sources and then creates a wireframe. Which of the following is the NEXT step the analyst should take in the dashboard creation process?

- A. Optimize the dashboard.
- B. Create subscriptions.
- C. Get stakeholder approval.
- D. Deploy to production.

Answer: C

Explanation:

Getting stakeholder approval is the next step the analyst should take in the dashboard creation process, after confirming the data sources and creating a wireframe. Stakeholder approval means getting feedback and validation from the intended users or clients of the dashboard, to ensure that it meets their expectations and requirements. This step helps to avoid rework and ensure customer satisfaction. References: CompTIA Data+ Certification Exam Objectives, page 14

NEW QUESTION 86

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 91

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

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

Answer: B

Explanation:

For an employer looking to maintain adequate office staffing during winter while tracking storm data, the most effective method would be to use public databases. These databases often contain comprehensive records of weather patterns and storm data collected and verified by reputable meteorological organizations. Utilizing public databases allows for access to historical and real-time data that is crucial for making informed decisions about staffing during adverse weather conditions.

Web scraping (A) is not the most reliable method, as it may involve extracting data from various websites that might not always provide verified or consistent information. Observations © can be subjective and may not cover a wide enough area to be effective for decision-making on a larger scale. Weather surveys (D) could provide insights, but they are not as immediate or comprehensive as the data available in public databases. References:

? The systematic review on Big Data Analytics in Weather Forecasting suggests that

big data techniques and technologies can manage and analyze the huge volume of weather data from different resources, which supports the use of public databases¹.

? NOAA??s approach to detecting severe weather events using instruments and receiving information from storm spotters indicates the importance of reliable, collected data, which is typically stored in public databases².

? The National Weather Service??s use of observational data collected by various instruments, which are then fed into forecast models, further emphasizes the value of established data collection methods over individual observations or surveys³.

NEW QUESTION 93

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 98

The number of phone calls that the call center receives in a day is an example of:

- A. continuous data.
- B. categorical data.
- C. ordinal data.
- D. discrete data.

Answer: D

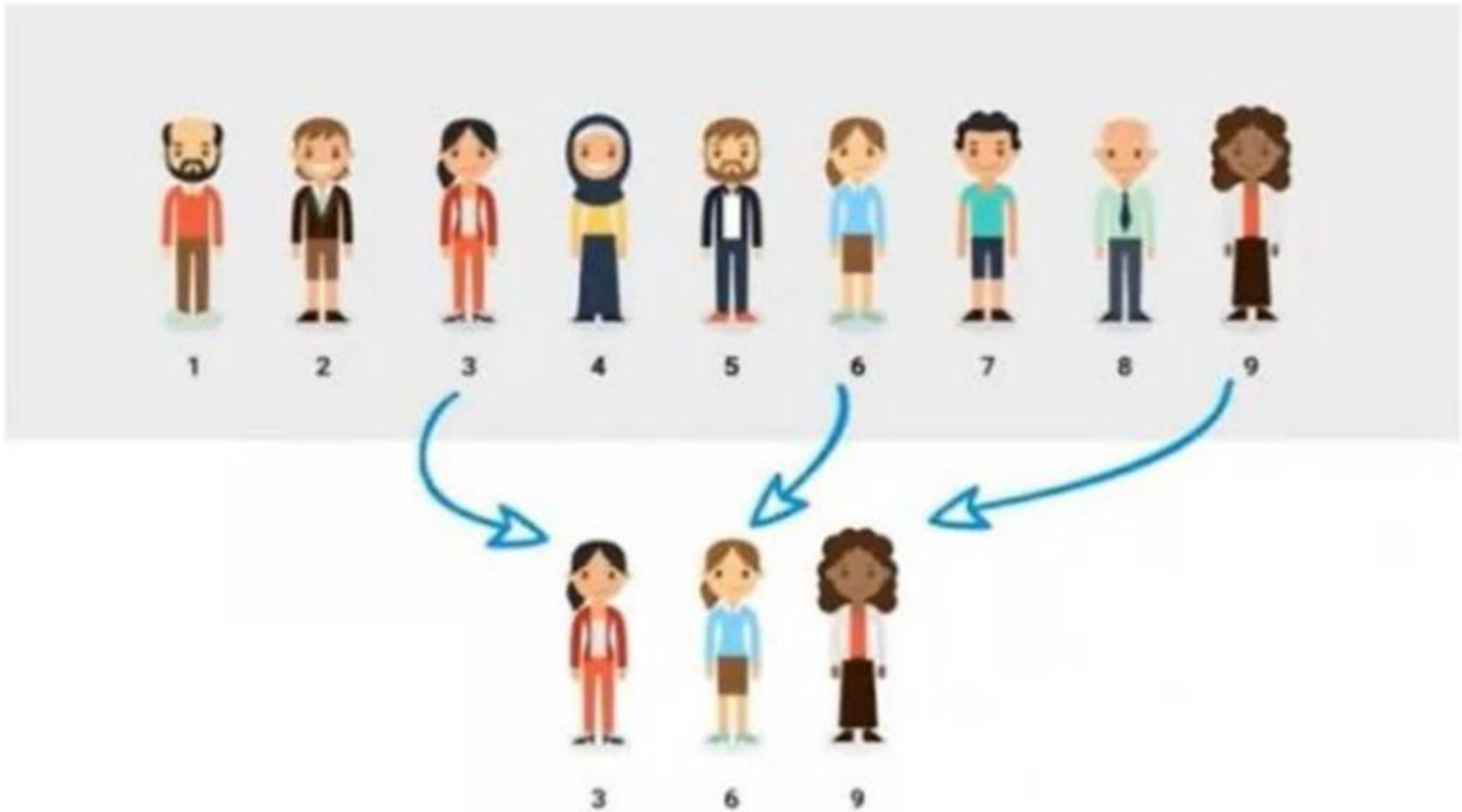
Explanation:

Discrete data is a type of data that can only take certain values, usually whole numbers or integers. Discrete data can be counted, but not measured. For

example, the number of students in a class, the number of books in a library, or the number of phone calls that a call center receives in a day are all examples of discrete data. Discrete data is different from continuous data, which can take any value within a range, and can be measured with precision. For example, the height of a person, the weight of a fruit, or the temperature of a room are all examples of continuous data. Therefore, the correct answer is D. References: [Discrete vs Continuous Data: Definition and Examples - Statistics How To], [Discrete Data - Definition and Examples | Math Goodies]

NEW QUESTION 100

Given the diagram below:



Which of the following types of sampling is depicted in the image?

- A. Stratified
- B. Random
- C. Cluster
- D. Systematic

Answer: D

Explanation:

Systematic sampling is a type of sampling where the sample is selected by following a fixed interval. For example, every 10th person in a list is chosen for the sample. In the image, the sample is selected by choosing every 3rd person in the line, starting from person number 1. This is an example of systematic sampling. References: Types of Sampling Techniques in Data Analytics You Should Know, Sampling Methods | Types, Techniques & Examples - Scribbr

NEW QUESTION 103

Which of the following is an example of PII?

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

Answer: B

Explanation:

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

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

NEW QUESTION 104

A data engineer is creating a database field to capture whether a customer likes vanilla ice cream. Which of the following data types is the best to capture this information?

- A. Integer
- B. Boolean
- C. Categorical

D. Numeric

Answer: B

NEW QUESTION 107

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

- A. 147mm
- B. 154mm
- C. 394 mm
- D. 21,704mm

Answer: B

Explanation:

The correct answer is B. 154 mm.

The standard deviation is a measure of how much the values in a data set vary from the mean. To calculate the standard deviation, we need to follow these steps:

? Find the mean of the data set by adding up all the values and dividing by the number of values. In this case, the mean is $(300 + 430 + 170 + 470 + 600) / 5 = 394$ mm.

? Find the difference between each value and the mean, and square it. In this case, the differences and their squares are:

? Find the sum of the squared differences. In this case, the sum is $8836 + 1296 + 50176 + 5776 + 42436 = 108520$.

? Divide the sum by the number of values. In this case, the result is $108520 / 5 = 21704$. This is called the variance.

? Take the square root of the variance. In this case, the result is $\sqrt{21704} = 147.32$ mm. This is called the standard deviation.

Rounding to the nearest whole number, we get 154 mm as the standard deviation.

NEW QUESTION 108

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

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

Answer: A

Explanation:

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

References: [What are Dynamic Reports? | Sisense], Static vs Dynamic Reports - What??s The Difference? | datapine

NEW QUESTION 111

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

<https://www.bing.com/images/blob?bcid=S1XCF9p02M4GjpbGxHj0lrlaj9sw.....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 115

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

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

Answer: C

NEW QUESTION 117

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

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

Answer: C

Explanation:

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

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

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

NEW QUESTION 120

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

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

Answer: A

Explanation:

Missing data is a type of data quality issue that occurs when some values in a data set are not recorded or available. Missing data can affect the validity and reliability of survey results, especially if the missing values are not random or ignorable. Missing data can also reduce the sample size and the statistical power of the analysis¹²

If 100% participation is needed on survey results, a data analyst would look for missing data first, because missing data would indicate that some participants did not complete or submit the survey, or that some responses were not recorded or transmitted correctly. A data analyst would need to identify the causes and patterns of missing data, and apply appropriate methods to handle or prevent missing data, such as imputation, deletion, weighting, or follow-up¹²

NEW QUESTION 123

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

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

Answer: BD

Explanation:

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

Option A is incorrect because it is not a reason to create and maintain a data dictionary, but a benefit of doing so.
 Option C is incorrect because specifying user groups for databases is not a function of a data dictionary, but a function of a database management system or a security policy.
 Option E is incorrect because confining breaches of PHI data is not a function of a data dictionary, but a function of a data protection or encryption system.
 Option F is incorrect because reducing processing power requirements is not a function of a data dictionary, but a function of a data compression or optimization system.

NEW QUESTION 125

Given the following data table:

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

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

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

Answer: BD

Explanation:

Data cleansing is a critical process in data analytics to ensure the accuracy and quality of data. The reasons to undertake data cleansing include:

? Missing Data (B): Missing data can lead to incomplete analysis and biased results. It is essential to identify and address gaps in the dataset to maintain the integrity of the analysis¹.

? Invalid Data (D): Invalid data includes entries that are out of range, improperly formatted, or illogical (e.g., a negative age). Such data can corrupt analysis and lead to incorrect conclusions¹.

Other options, such as non-parametric data (A), are not inherently errors but refer to a type of data that doesn't assume a normal distribution. Duplicate data (C) and redundant data (E) could also be reasons for data cleansing, but they are not listed as options to select from in the provided image details. Normalized data (F) refers to data that has been processed to fit into a certain range or format and is typically not a reason for data cleansing. References:

? Understanding the importance of data quality and the impacts of missing and invalid data on research outcomes¹.

? Best practices in data cleansing².

Data cleansing is required for various reasons, two of which are missing data (B) and invalid data (D). From the table provided, we can infer the necessity of cleansing in the context of ensuring data integrity and consistency. Missing data refers to the absence of data where it is expected, which can hinder analysis due to incomplete information. Invalid data refers to data that is incorrect, out of range, or in an inappropriate format, which can lead to inaccuracies in any analysis or report. Both these issues can significantly affect the outcomes of any data-related operations and thus need to be rectified through the data cleansing process.

NEW QUESTION 130

The ACME Corporation hired an analyst to detect data quality issues in their Excel documents. Which of the following are the most common issues? (Select TWO)

- A. Apostrophe.
- B. Commas.
- C. Symbols.
- D. Duplicates.
- E. Misspellings.

Answer: DE

Explanation:

* 1. Duplicates

* 2. Misspellings

The most common data quality issues are difficult to resolve in Excel because of their rigidity. It forces analysts to do a ton of manual work, which results in a high probability of an error being introduced to the data set. Those common issues include:

- Blanks
- Nulls
- Outliers
- Duplicates
- Extra spaces
- Misspellings
- Abbreviations and domain-specific variations
- Formula error codes

When introduced, these errors can skew or even invalidate the resulting analysis. A smart tool would minimize the possibility of error by automating the manual work. In Excel, you might look for data quality issues in one of two ways. First, you might use auto filters on specific columns to scan for anomalies and blanks or you might use a pivot table to find gaps and discrepancies.

In either case, you're scanning for the anomalies yourself. Suffice it to say that's not a very efficient process. It also means accuracy is only as good as the

analyst's eye, so the probability of error varies throughout the day.

NEW QUESTION 133

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

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

Answer: B

Explanation:

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

NEW QUESTION 135

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

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

Answer: A

Explanation:

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

NEW QUESTION 137

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 them¹.

NEW QUESTION 142

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

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

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

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

Answer: B

Explanation:

To compare the students?? scores, we need to standardize them by using the z-score formula, which is:

$$z = (x - \mu) / \sigma$$

where x is the raw score, μ is the mean, and σ is the standard deviation. The z-score tells us how many standard deviations a score is above or below the mean. A

higher z-score means a better score relative to the average.

Using the table, we can calculate the z-scores for each student as follows:

Randy: $z = (76 - 70) / 2 = 3$ Katie: $z = (86 - 80) / 3 = 2$ Ralph: $z = (80 - 75) / 2 = 2.5$ Jean: $z = (80 - 90) / 1 = -10$

The student with the highest z-score is Randy, with a z-score of 3. This means that Randy scored 3 standard deviations above the mean in math, which is the best performance among the four students. Therefore, the correct answer is A.

References: Comparing with z-scores (video) | Z-scores | Khan Academy, 17 Important Data Visualization Techniques | HBS Online

NEW QUESTION 143

A data analyst has been asked to create an ad-hoc sales report for the Chief Executive Officer (CEO).

Which of the following should be included in the report?

- A. The sales representatives' home addresses.
- B. Line-item SKU numbers.
- C. YTD total sales.
- D. The customers' first and last names.

Answer: C

Explanation:

The report for the CEO should include YTD total sales, as this will provide a high-level overview of the sales performance of the company and show how it is meeting its annual goals. The other options are not appropriate for the CEO, as they are either too detailed or irrelevant for the report. The sales representatives' home addresses, line-item SKU numbers, and customers' first and last names are not related to the sales performance and might compromise the privacy and security of the data.

Reference: CompTIA Data+ (DA0-001) Practice Certification Exams | Udemy

NEW QUESTION 144

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 147

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 148

A data analyst was asked to create a chart that shows the relationship between study hours and exam scores for each student using the data sets in the table below:

Student	Exam score	Study hours
Kim	90	7.5
Leo	80	6
Alpha	60	4
Jude	85	7
Ella	95	8

Which of the following charts would BEST represent the relationship between the variables?

- A. A histogram
- B. A scatter plot
- C. A heat map
- D. A bar chart

Answer: B

Explanation:

This is because a scatter plot is a type of chart that shows the relationship between two variables for each observation or unit in a data set, such as study hours and exam scores for each student in this case. A scatter plot can be used to display and analyze the correlation, trend, or pattern among the variables, as well as identify any outliers or clusters in the data. For example, a scatter plot can show if there is a positive, negative, or no correlation between study hours and exam scores, as well as show if there are any students who have unusually high or low exam scores compared to their study hours. The other charts are not the best charts to represent the relationship between the variables. Here is why:

? A histogram is a type of chart that shows the frequency or the count of values in a single variable for different intervals or bins, such as exam scores for different ranges in this case. A histogram can be used to display and analyze the distribution, shape, or spread of the variable, as well as identify any gaps, peaks, or skewness in the data. For example, a histogram can show if most students have high, low, or average exam scores, as well as show if there are any intervals that have no students at all.

? A heat map is a type of chart that shows the intensity or the magnitude of values in two variables for different categories or groups, such as exam scores and study hours for different student names in this case. A heat map can be used to display and analyze the variation, contrast, or comparison among the categories or groups, as well as identify any hot spots, cold spots, or gradients in the data. For example, a heat map can show which students have higher or lower exam scores and study hours than others, as well as show if there is a color pattern that indicates a relationship between exam scores and study hours.

? A bar chart is a type of chart that shows the value or the amount of a single variable for different categories or groups, such as exam scores for different student names in this case. A bar chart can be used to display and analyze the comparison, ranking, or proportion among the categories or groups, as well as identify any differences, similarities, or outliers in the data. For example, a bar chart can show which students have higher or lower exam scores than others, as well as show if there are any students who have exceptionally high or low exam scores.

NEW QUESTION 151

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 152

Which of the following data sampling methods involves dividing a population into subgroups by similar characteristics?

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

Answer: D

Explanation:

Stratified sampling is a data sampling method that involves dividing a population into subgroups by similar characteristics, such as age, gender, income, etc. Then, a simple random sample is drawn from each subgroup. This method ensures that each subgroup is adequately represented in the sample and reduces the sampling error. References: CompTIA Data+ Certification Exam Objectives, page 11.

NEW QUESTION 156

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 157

Q3 2020 has just ended, and now a data analyst needs to create an ad-hoc sales report that demonstrates how well the Q3 2020 promotion went versus last year's Q3 promotion.

Which of the following date parameters should the analyst use?

- A. 2019 v
- B. YTD 2020
- C. Q3 2019 v
- D. Q3 2020
- E. YTD 2019 v
- F. YTD 2020
- G. Q4 2019 v
- H. Q3 2020

Answer: B

Explanation:

The date parameters that the analyst should use are Q3 2019 vs. Q3 2020, as this will allow the analyst to compare the sales performance of the Q3 2020 promotion with the same period of last year. This will help to eliminate any seasonal or cyclical effects that might affect the sales data. The other options are not relevant for this purpose, as they either compare different quarters or different years. Reference: CertMaster Practice for Data+ Exam Prep - CompTIA

NEW QUESTION 160

Which of the following BEST describes standard deviation?

- A. A measure that is used to establish a relationship between two variables
- B. A measure of how data is distributed
- C. A measure of the amount of dispersion of a set of values
- D. A measure that is used to find the significant difference between variables

Answer: C

Explanation:

A measure of the amount of dispersion of a set of values. This is because standard deviation is a type of statistical measure that quantifies how much the values in a data set vary or deviate from the mean or the average of the data set. Standard deviation can be used to describe the spread or the distribution of the data, as well as to identify any outliers or extreme values in the data. For example, a low standard deviation indicates that the values are close to the mean, while a high standard deviation indicates that the values are far from the mean. The other options are not correct descriptions of standard deviation. Here is why:

? A measure that is used to establish a relationship between two variables is not a correct description of standard deviation, but rather a description of correlation or regression, which are types of statistical measures that quantify how two variables are related or associated with each other. Correlation or regression can be used to test or model the dependence or the influence of one variable on another variable, as well as to predict or estimate the value of one variable based on the value of another variable.

? A measure of how data is distributed is not a correct description of standard deviation, but rather a description of frequency or probability, which are types of statistical measures that quantify how often or how likely a value or an event occurs in a data set. Frequency or probability can be used to describe the occurrence or the chance of the data, as well as to compare or contrast different categories or groups of the data.

? A measure that is used to find the significant difference between variables is not a correct description of standard deviation, but rather a description of hypothesis testing or inferential statistics, which are types of statistical methods that use sample data to make generalizations or conclusions about a population or a parameter. Hypothesis testing or inferential statistics 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.

NEW QUESTION 161

A data analyst needs to present the results of an online marketing campaign to the marketing manager. The manager wants to see the most important KPIs and measure the return on marketing investment. Which of the following should the data analyst use to BEST communicate this information to the manager?

- A. A real-time monitor that allows the manager to view performance the day the campaign was launched
- B. A sell-service dashboard that allows the manager to look at the company's annual budget performance
- C. A spreadsheet of the raw data from all marketing campaigns and channels
- D. A summary with statistics, conclusions, and recommendations from the data analyst

Answer: D

Explanation:

The option that the data analyst should use to best communicate the information to the manager is a summary with statistics, conclusions, and recommendations from the data analyst. A summary is a concise and clear way of presenting the main findings and insights from the data analysis report. A summary should include relevant statistics that support the conclusions and recommendations from the data analyst. A summary should also highlight the most important KPIs and measure the return on marketing investment in relation to the objectives of the online marketing campaign. The other options are not as effective as using a summary to communicate the information to the manager, as they either provide too much or too little information or do not address the manager's needs or expectations. A real-time monitor may provide too much information that can be overwhelming or distracting for the manager who wants to see only the most important KPIs and measure the return on marketing investment. A self-service dashboard may provide too little information that can be insufficient or unclear for the manager who wants to see some guidance and interpretation from the data analyst. A spreadsheet of raw data may provide irrelevant or inaccurate information that can be confusing or misleading for the manager who wants to see some analysis and insights from the data analyst. Reference: [How to Write an Executive Summary for Your Data Analysis Report - Towards Data Science]

NEW QUESTION 163

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

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

Answer: AB

Explanation:

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

NEW QUESTION 165

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

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

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

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

Answer: A

Explanation:

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

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

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

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

References:

- ? P-value Calculator & Statistical Significance Calculator
- ? p-value Calculator | Formula | Interpretation
- ? How to obtain the P value from a confidence interval | The BMJ
- ? Confidence Intervals & P-values for Percent Change / Relative Difference

NEW QUESTION 166

An analyst needs to create an analytics dashboard for an employee intranet site to improve the search functionality, display relevant information, and maintain an

updated FAQ page. Which of the following visualizations would best represent what employees are searching for?

- A. A word cloud
- B. A histogram
- C. A pie chart
- D. A scatter plot

Answer: A

Explanation:

A word cloud is an ideal choice for visualizing what employees are searching for on an intranet site. It represents the frequency of word occurrence in a visually impactful way, with more commonly searched terms appearing larger in the cloud. This allows for quick identification of the most popular queries and topics of interest among employees. Unlike histograms, pie charts, or scatter plots, word clouds can effectively display textual data, which is the nature of search queries. They are particularly useful for analyzing text data from surveys or feedback forms, which can be similar to search query data in an intranet environment¹²³⁴.
References: 1: ??What Are Word Clouds? Pros & Cons of Word Cloud Visualizations?? - Alida 2: ??Using Word Clouds for Powerful Data Visualization?? - WordCloud.app blog 3: ??Ultimate Google Data Studio Word Cloud Guide: Visualization 2024?? - AtOnce 4: ??How to Create Word Cloud in Power BI?? - Zebra BI

NEW QUESTION 169

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

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

Answer: D

Explanation:

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

NEW QUESTION 173

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

- A. Qlik.
- B. Power BI.
- C. Domo.
- D. Dataroma.

Answer: B

Explanation:

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

NEW QUESTION 174

.....

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](#)