CompTIA

Exam Questions DA0-001

CompTIA Data+ Certification Exam

NEW QUESTION 1
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 2
An analysts building a monthly report for production and wants to ensure the audience is aware of its once-a-month cadence. Which of the following is the MOST important to convey that information?

A. The date of the dashboard build
B. The data refresh date
C. A report summary
D. Frequently asked questions

Answer: A

Explanation:
This is because the date of the dashboard build is the most important component to convey that information, which is the once-a-month cadence of the monthly report for production. The date of the dashboard build can convey that information by indicating when the dashboard was created or updated, as well as showing the frequency or interval of the dashboard creation or update. For example, the date of the dashboard build can convey that information by displaying a date format that includes the month and year, such as January 2020, February 2020, etc., or by displaying a text format that includes the word ??monthly??, such as Monthly Report for Production - January 2020, Monthly Report for Production - February 2020, etc. The other components are not the most important components to convey that information. Here is why:
? The data refresh date is a component that indicates when the data on the dashboard was refreshed or retrieved from the source or system, such as a database, a cloud service, or a web application. The data refresh date does not convey that information, but rather conveys how current or up-to-date the data on the dashboard is.
? A report summary is a component that provides an overview or a highlight of the main findings or insights from the dashboard, such as key metrics, indicators, or trends. A report summary does not convey that information, but rather conveys what the dashboard is about or what it shows.
? Frequently asked questions is a component that provides answers or explanations to common or expected questions from the audience or users of the dashboard, such as how to use or interpret the dashboard, what are the assumptions or limitations of the dashboard, etc. Frequently asked questions does not convey that information, but rather conveys how to understand or interact with the dashboard.

NEW QUESTION 3
Jhon is working on an ELT process that sources data from six different source systems.
Looking at the source data, he finds that data about the sample people exists in two of six systems.
What does he have to make sure he checks for in his ELT process? Choose the best answer.

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

Answer: C

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

NEW QUESTION 4
A data analyst received the information in the table below from a recently completed marketing campaign:

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

Which of the following is the total order conversion rate?

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

Answer: B

Explanation:
The correct answer is A. 13.2%.
The total order conversion rate is the ratio of the total number of orders to the total number of clicks, expressed as a percentage. To calculate the total order conversion rate, we need to sum up the clicks and orders from all the channels, and then divide the orders by the clicks and multiply by 100.
Using the data from the table, we can do the following:
? Total clicks = 580 + 800 + 1,200 + 300 + 620 = 3,500
? Total orders = 55 + 100 + 220 + 60 + 85 = 520
? Total order conversion rate = (520 / 3,500) x 100 = 14.857%
? Rounding to one decimal place, we get 14.9% Therefore, the total order conversion rate is 14.9%.

NEW QUESTION 5
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. PI I
B. PCI
C. PBI
D. PHI

Answer: B

NEW QUESTION 6
A data analyst must separate the column shown below into multiple columns for each component of the name:

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

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

A. Imputing
B. Transposing
C. Parsing
D. Concatenating

Answer: C

Explanation:
Parsing is the data manipulation technique that should be used to separate the column into multiple columns for each component of the name. Parsing is the process of breaking down a string of text into smaller units, such as words, symbols, or numbers. Parsing can be used to extract specific information from a text column, such as names, addresses, phone numbers, etc. Parsing can also be used to split a text column into multiple columns based on a delimiter, such as a comma, space, or dash1. In this case, the

analyst can use parsing to split the column by the comma delimiter and create three new columns: one for the last name, one for the first name, and one for the middle initial. This will make the data more organized and easier to analyze.

NEW QUESTION 7
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 8
A database administrator is required to mask certain table columns containing Pll in order to comply with the company privacy policy. Which of the following are the most likely types of information the administrator should mask? (Select two).

A. Government-issued ID
B. Address
C. Order ID
D. Order date
E. Customer ID
F. Referral number

Answer: AB

NEW QUESTION 9
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 10
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 10
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 12
A Chief Executive Officer (CEO) is requesting more up-to-date sales data for improved visibility prior to month-end. An analyst must determine the frequency of a sales report that was previously distributed on an as-needed basis. Which of the following would be the most appropriate frequency for this report?

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

Answer: C

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

NEW QUESTION 14
You should always choose the analytics tool that is most appropriate for any given situation, even if that means acquiring a new tool.

A. True.
B. False.

Answer: B

Explanation:
The statement is false. You should not always choose the analytics tool that is most appropriate for any given situation, even if that means acquiring a new tool. Acquiring a new tool can be costly, time-consuming, and risky, as it may not be compatible with your existing data sources, systems, or processes. It may also require additional training, maintenance, and support. Therefore, you should always consider the trade-offs between the benefits and drawbacks of acquiring a new tool versus using an existing one. You should also evaluate the feasibility, availability, and reliability of the new tool before making a decision. Reference: CompTIA Data+ (DA0-001) Practice Certification Exams | Udemy

NEW QUESTION 18
Which of the following can be used to translate data into another form so it can only be read by a user who has a key or a password?

A. Data encryption.
B. Data transmission.
C. Data protection.
D. Data masking.

Answer: A

Explanation:

Data encryption can be used to translate data into another form so it can only be read by a user who has a key or a password. Data encryption is a process of transforming data using an algorithm or a cipher to make it unreadable to anyone except those who have the key or the password to decrypt it. Data encryption is a common method of protecting data from unauthorized access, modification, or theft. Reference: Guide to CompTIA Data+ and Practice Questions - Pass Your Cert

NEW QUESTION 20
Joseph is interpreting a left skewed distribution of test scores. Joe scored at the mean, Alfonso scored at the median, and gaby scored and the end of the tail. Who had the highest score?

A. Joseph
B. Joe
C. Alfonso
D. Gaby

Answer: C

Explanation:

Alfonso had the highest score. A left skewed distribution is a distribution where the tail is longer on the left side than on the right side, meaning that most of the values are clustered on the right side and there are some outliers on the left side. In a left skewed distribution, the mean is less than the median, which is less than the mode. Therefore, Joseph, who scored at the mean, had the lowest score, Gaby, who scored at the end of the tail, had the second lowest score, and Alfonso, who scored at the median, had the highest score. Reference: Skewness - Statistics How To

NEW QUESTION 23
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 28
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 preserved1
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 31
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 35
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 39
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 41
What category of data stewardship work is focused on ensuring that the organization respects the wishes of data subjects?

A. Data quality.
B. Data privacy.
C. Data security.
D. Regulatory compliance.

Answer: B

Explanation:

Data privacy defines who has access to data, while data protection provides tools and policies to actually restrict access to the data. Compliance regulations help ensure that user's privacy requests are carried out by companies, and companies are responsible to take measures to protect private user data. Why is data privacy important?
When data that should be kept private gets in the wrong hands, bad things can happen. A data breach at a government agency can, for example, put top secret information in the hands of an enemy state. A breach at a corporation can put proprietary data in the hands of a competitor.

NEW QUESTION 42
Under which of the following circumstances should the null hypothesis be accepted when a
= 0.05?

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

Answer: C

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

NEW QUESTION 45
Which of the following tools would be best to use to calculate the interquartile range, median, mean, and standard deviation of a column in a table that has 5.000.000 rows?

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

Answer: B

NEW QUESTION 50
Which of the following is an example of structured data?

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

Answer: A

Explanation:
A credit card number is an example of structured data, which is a type of data that conforms to a data model, has a well-defined structure, follows a consistent order, and can be easily accessed and used by a person or a computer program. A credit card number consists of 16 digits that are divided into four groups of four digits each, separated by spaces or hyphens. The first six digits indicate the issuer identification number, the next nine digits indicate the account number, and the last digit is a check digit that validates the
number. A credit card number can be stored and processed in a structured format, such as a database or a spreadsheet1.

NEW QUESTION 54
Five dogs have the following heights in millimeters: 300, 430, 170, 470, 600
Which of the following is the mean height for the five dogs?

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

Answer: B

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

NEW QUESTION 56
You have two databases tables that you would like to join together using a foreign key relationship.
What term best describes this action?

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

Answer: D

Explanation:

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

NEW QUESTION 57
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 61
An analyst needs to provide a chart to identify the composition between the categories of the survey response data set:

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

Which of the following charts would be BEST to use?

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

Answer: B

Explanation:
A pie chart is the best choice to show the composition between the categories of the survey response data set. A pie chart represents the whole with a circle, divided by slices into parts. Each slice shows the relative size of each category as a percentage of the total. A pie chart is useful when the categories are mutually exclusive and add up to 100%. The table shows the favorite color and the number of responses for each color, which can be easily converted into percentages. A pie chart can show how each color contributes to the total number of responses.
Option A is incorrect because a histogram is used to show how data points are distributed along a numerical scale. The survey response data set is not numerical, but categorical. Option C is incorrect because a line chart is used to show trends or changes over time. The survey response data set does not have a time dimension.
Option D is incorrect because a scatter plot is used to show the relationship between two numerical variables. The survey response data set does not have two numerical variables. Option E is incorrect because a waterfall chart is used to show how an initial value is increased or decreased by a series of intermediate values. The survey response data set does not have an initial value or intermediate values.
References:
? How to Choose the Right Chart for Your Data - Infogram
? How to Choose the Right Data Visualization | Tutorial by Chartio
? Find the Best Visualizations for Your Metrics - The Data School
? How to choose the best chart or graph for your data

NEW QUESTION 63
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 Optimization1.
? MSSQLTips on SQL Query Performance2.
? Blog post on SQL Performance Optimization3.
? SQL Easy guide on improving SQL Query Performance4.
? LearnSQL.com on SQL for Data Analysis5.

NEW QUESTION 64
Andy is a pricing analyst for a retailer. Using a hypothesis test, he wants to assess whether people who receive electronic coupons spend more on average.
What should Andy's null hypothesis be?

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

Answer: C

Explanation:

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

NEW QUESTION 65
Given the following grocery store orders:

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

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

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

Answer: C

Explanation:

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

NEW QUESTION 66
An analyst needs to determine the appropriate data type for the following sample data: sample data collected:
Which of the following data types should be used for this data?

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

Answer: B

NEW QUESTION 68
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 analysis1.

NEW QUESTION 73
Which one of the following would not normally be considered a summary statistic?

A. z-score.
B. Mean.
C. Variance.
D. Standard deviation.

Answer: A

Explanation:

Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is. A z-score can be placed on a normal distribution curve.

NEW QUESTION 77
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 80
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 82
A data analyst is creating a report that will provide information about various regions, products, and time periods. Which of the following formats would be the MOST efficient way to deliver this report?

A. A workbook with multiple tabs for each region
B. A daily email with snapshots of regional summaries
C. A static report with a different page for every filtered view
D. A dashboard with filters at the top that the user can toggle

Answer: D

Explanation:
A dashboard with filters at the top that the user can toggle would be the most efficient way to deliver this report, because it allows the user to customize the view and explore different combinations of regions, products, and time periods. A workbook with multiple tabs for each region would be cumbersome and repetitive. A daily email with snapshots of regional summaries would not provide enough detail or interactivity. A static report with a different page for every filtered view would be too long and hard to navigate. References: CompTIA Data+ Certification Exam Objectives, page 14

NEW QUESTION 84
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 85
Which one of the following is a measure of dispersion?

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

Answer: A

NEW QUESTION 86
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 88
Samantha needs to share a list of her organization's top 50 customers with the VP of sales.
She would like to include the name of the customer, the business they represent, their contact information, and their total sales over the past year.
The VP does not have any specialized analytics skills or software but would like to make some personal notes on the dataset.
What would be the best tool for Samantha to use to share this information?

A. Power BI.

B. Microsoft Excel.
C. Minitab.
D. SAS.

Answer: B

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

NEW QUESTION 91
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 94
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) ?? 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) ?? 1.96 * sqrt(p * (1 - p) * (1/n1 + 1/n2)) = (0.084 - 0.072) ?? system The assistant??s response has exceeded the maximum character limit of [500]. Please shorten your response or split it into multiple messages.

NEW QUESTION 95
A military commander would like to see the health scorecards of the troops daily and filter them based on gender and rank. Considering this data is PHI, which of the following would be the best way for the commander to view the information?

A. An emailed report
B. A password-protected dashboard
C. A daily printout of a report
D. A cloud-hosted spreadsheet

Answer: B

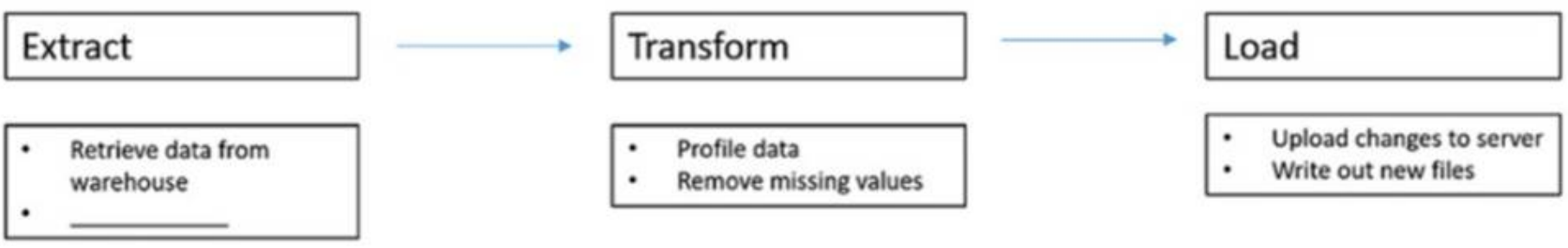
Explanation:
A password-protected dashboard is a type of web-based application that can display the health scorecards of the troops in a secure and interactive way. A password-protected dashboard can provide the following benefits for the commander:
? It can protect the PHI data from unauthorized access or disclosure by requiring a
valid username and password to log in. This can ensure that only the commander and other authorized personnel can view the information12
? It can allow the commander to filter the data based on gender and rank by using
drop-down menus, sliders, checkboxes, or other controls. This can enable the commander to customize the view and focus on the relevant data13
? It can update the data daily by connecting to a data source that refreshes
automatically or on demand. This can ensure that the commander always sees the latest and most accurate information14
? It can present the data in a visual and intuitive way by using charts, graphs, tables,
or other elements. This can help the commander to understand and analyze the data more easily and effectively1

NEW QUESTION 98
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 100
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 105
An analyst is creating a resource to improve users' experience when they select specific records based on particular dates. Which of the following should the analyst use to create a resource that best meets user needs?

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

Answer: B

Explanation:
A drop-down menu is a graphical user interface element that allows users to select one option from a list of options that are hidden until the user clicks on the menu. A drop-down menu can be used to create a resource that best meets user needs when they select specific records based on particular dates, because:
? A drop-down menu can provide a predefined list of dates or date ranges that are relevant and valid for the records, such as today, yesterday, last week, last month, custom range, etc. This can help users to avoid typing errors or invalid dates in a text field, and to save time and effort in entering the dates.
? A drop-down menu can also provide a calendar or a date picker that allows users
to select a specific date or a range of dates from a graphical representation of a calendar. This can help users to visualize and compare the dates, and to easily adjust or modify their selection.
? A drop-down menu can improve the user experience by making the interface more
compact and organized, as it only shows one option at a time and hides the rest of the options until the user clicks on the menu. This can help users to focus on their selection and to avoid clutter and distraction.

NEW QUESTION 110
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 113
A company's human resources department has asked a data analyst to categorize the income of all employees into five salary bands:

Employee_ID	Salary	Salary_band
003	$130,000	
014	$120,000	
004	$110,000	
013	$90,000	
002	$140,000	
012	$122,000	
016	$132,000	
006	$70,000	
017	$53,000	
009	$111,000	
019	$107,000	
008	$111,000	
018	$50,000	

Which of the following types of functions would be the most appropriate to use?

A. Statistical
B. Aggregate
C. Logical
D. Mathematical

Answer: C

Explanation:
Short Explanation: Logical functions are the most appropriate to use for categorizing data
into bands, because they allow the data analyst to apply conditional statements and criteria to the data values. For example, the IF function can be used to assign a band name based on whether a value meets a certain condition or not. Other logical functions that can be useful for categorizing data are AND, OR, NOT, and IFERROR12

NEW QUESTION 114
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 119
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 121
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 chart using the site and percentage of new custorners data.
D. Include a scatter chart using the site and the percent of new customers data.

Answer: A

Explanation:
The best type of chart to display the data is A. Include a bar chart using the site and the percentage of new customers data.
A bar chart is a good choice for comparing categorical data with numerical data, such as the site and the percentage of new customers. A bar chart can show the relative differences between the sites and highlight the site with the highest percentage of new customers. A bar chart can also be easily labeled and formatted to make the data clear and understandable.
A line chart is not suitable for this data, because it is used to show trends or changes over time, which is not relevant for the site and the percentage of new customers data. A line chart would also be confusing and misleading, as it would imply a connection or correlation between the sites that does not exist.
A pie chart is also not a good choice for this data, because it is used to show the proportion of a whole, not the comparison of different categories. A pie chart would also be difficult to read and interpret, as it would require labels or legends to identify the sites and their percentages. A pie chart would also not be able to show the exact values of the percentages, only their relative sizes.
A scatter chart is another inappropriate option for this data, because it is used to show the relationship or correlation between two numerical variables, not between a categorical and a numerical variable. A scatter chart would also be cluttered and unclear, as it would plot each site as a point on a coordinate plane, without any labels or axes. A scatter chart would also not be able to show the differences or rankings between the sites and their percentages.

NEW QUESTION 125
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 129
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 134
Which of the following is a best practice when updating a legacy data source?

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

Answer: C

Explanation:
When updating a legacy data source, it is a best practice to create a codebook to document field changes. A codebook serves as a detailed guide and record of the data structure, definitions, and any transformations or modifications made to the data fields. This documentation is crucial for maintaining data integrity, ensuring consistency, and facilitating future data use and understanding. It provides a reference that can be invaluable for data analysts, developers, and any stakeholders who need to work with the data.
Creating a codebook is preferred over placing old data in new fields, which can lead to confusion and data integrity issues. Keeping only the most recent data may result in the loss of valuable historical information. Removing the data source from production is not a practice related to updating data but rather to retiring a data source1234.
References:
? Legacy Data Migration: A Comprehensive Guide | OpenGeeksLab
? How to Successfully Complete Legacy Database Migration

? Methods for Saving and Integrating Legacy Data - DATAVERSITY
? Legacy Data Digitization - Learn The Best Practices

NEW QUESTION 138
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 B1
C. IBM SPSS
D. Python

Answer: D

Explanation:
The option that is a common data analytics tool that is also used as an interpreted, high-level, general-purpose programming language is Python. Python is a popular and versatile programming language that can be used for various purposes, such as web development, software development, automation, machine learning, and data analysis. Python has many features and libraries that make it suitable for data analytics, such as its simple syntax, dynamic typing, multiple paradigms, built-in data structures, NumPy, pandas, matplotlib, scikit-learn, etc. The other options are not programming languages, but software applications or platforms that are used for data analytics or related tasks. SAS is a software suite that provides advanced analytics, business intelligence, data management, and predictive analytics capabilities. Microsoft Power BI is a business analytics service that provides interactive visualizations and business intelligence capabilities. IBM SPSS is a software package that offers statistical analysis, data mining, text analytics, and predictive analytics capabilities. Reference: Python For Data Analysis - DataCamp

NEW QUESTION 140
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 142
During data profiling, an analyst decides to recode the status column in the following data set:

EMP ID	Date	Activity	Status
000352	1/2/2022	Course001	yes
000331	1/5/2022	Course001	completed
000347	1/10/2022	Course001	done
000364	1/12/2022	Course001	Y

Which of the following data concerns explains why the analyst wants to take this action?

A. Redundancy
B. Duplication
C. Invalidity
D. Inconsistency

Answer: D

Explanation:
The ??Status?? column in the dataset shows different terms such as ??yes??, ??completed??, ??done??, and ??Y?? that likely represent the same outcome - that a task has been completed. This variation in terms leads to inconsistency within the data. Data profiling aims to ensure that data is consistent, among other quality metrics, to facilitate accurate analysis and reporting. By recoding the ??Status?? column, the analyst seeks to address this inconsistency, ensuring that all entries indicating completion are represented uniformly. This enhances the data quality and usability for subsequent data analysis tasks.References:
The action of recoding is taken to standardize the data entries and eliminate inconsistencies, which is crucial for maintaining data integrity and ensuring reliable data analysis.

NEW QUESTION 146
Amanda needs to create a dashboard that will draw information from many other data sources and present it to business leaders.
Which one of the following tools is least likely to meet her needs?

A. QuickSight.
B. Tableau.
C. Power BI.
D. SPSS Modeler.

Answer: D

Explanation:
SPSS Modeler.
QuickSight, Tableau, and Power BI are all powerful analytics and reporting tools that can pull data from a variety of sources. SPSS Modeler is a powerful predictive analytics platform that is designed to bring predictive intelligence to decisions made by individuals, groups, systems and your enterprise.

NEW QUESTION 149
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 152
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 153
Which of the following technologies would be best suited for creating a multiple linear regression model?

A. Microsoft Power Bl
B. R
C. SQL
D. Tableau

Answer: B

Explanation:
R is a statistical programming language that is specifically designed for data analysis and statistical modeling, making it highly suitable for creating a multiple linear regression model. It has extensive libraries such as lm() for linear modeling, which simplifies the process of model creation, diagnostics, and interpretation. R also provides robust tools for data manipulation and visualization, which are essential for preparing data for regression analysis and understanding the results123.
While Microsoft Power BI, SQL, and Tableau have capabilities for regression analysis, they are more limited compared to R. Power BI and Tableau are primarily business intelligence tools that offer some built-in analytics capabilities, but they are not as comprehensive as
R. SQL is a database query language that can perform some statistical calculations, but it is not inherently designed for statistical modeling4567.
References:
? Multiple Linear Regression in R: Tutorial With Examples - DataCamp1.
? Implementing linear regression in Power BI - SQLBI5.
? Choosing a Predictive Model - Tableau6.
? How Predictive Modeling Functions Work in Tableau7.

NEW QUESTION 158
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 160
A recurring event is being stored in two databases that are housed in different geographical locations. A data analyst notices the event is being logged three hours earlier in one database than in the other database. Which of the following is the MOST likely cause of the issue?

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

Answer: C

Explanation:
The most likely cause of the issue is that the databases are recording the event in different time zones. For example, if one database is in New York and the other database is in Los Angeles, there is a three-hour difference between them. Therefore, an event that occurs at 12:00 PM in New York would be recorded as 9:00 AM in Los Angeles. To avoid this issue, the databases should either use a common time zone or convert the timestamps to a standard format. Therefore, option C is correct.
Option A is incorrect because the data analyst is not querying the databases incorrectly, but rather observing a discrepancy in the timestamps.
Option B is incorrect because the databases are recording the same event, but with different timestamps.
Option D is incorrect because the second database is not logging incorrectly, but rather using a different time zone.

NEW QUESTION 162
Given the following:

Candy	Has_nuts	Date_purchased	Cost	Quantity	Ext_cost
Snickers	Y	2021-08-24	$1.00	2	2.00
Starburst	N	8/24/2021	null	10	null
Snickers	Y	2020-11-13	$2.00	3	6.00

Which of the following is the most important thing for an analyst to do when transforming the table for a trend analysis?

A. Fill in the missing cost where it is null.
B. Separate the table into two tables and create a primary key
C. Replace the extended cost field with a calculated field.
D. Correct the dates so they have the same format.

Answer: D

Explanation:
Correcting the dates so they have the same format is the most important thing for an analyst to do when transforming the table for a trend analysis. Trend analysis is a method of analyzing data over time to identify patterns, changes, or relationships. To perform a trend analysis, the data needs to have a consistent and comparable format, especially for the date or time variables.
In the example, the date purchased column has two different formats: YYYY-MM-DD and MM/DD/YYYY. This could cause errors or confusion when sorting, filtering, or plotting the data over time. Therefore, the analyst should correct the dates so they have the same format, such as YYYY-MM-DD, which is a standard and unambiguous format.

NEW QUESTION 166
.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questons and Answers in PDF Format

DA0-001 Practice Exam Features:

* DA0-001 Questions and Answers Updated Frequently

* DA0-001 Practice Questions Verified by Expert Senior Certified Staff

* DA0-001 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* DA0-001 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year