Exam Questions DA0-001

CompTIA Data+ Certification Exam

https://www.2passeasy.com/dumps/DA0-001/

NEW QUESTION 1
Which of the following is an example of a flat file?

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

Answer: A

Explanation:
A CSV file is a type of flat file that stores data as plain text in a table-like structure with rows and columns. Each row represents a single record, while columns represent fields or attributes of the data. A CSV file uses commas or other delimiters to separate the values in each row. A CSV file can be easily imported or exported by various applications and programs12

NEW QUESTION 2
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 3
A sales team wants visibility of current sales numbers, pipeline, and team performance. The team would also like to see calculations of individuals?? earned commissions and projected commissions based on sales, but they want that information to be kept confidential. Which of the following would be the BEST way to provide this visibility?

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

Answer: D

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

NEW QUESTION 4
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 5
An analyst modified a data set that had a number of issues. Given the original and modified versions:

Original data:

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

Modified data:

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

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

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

Answer: B

Explanation:
The correct answer is B. Recoding.

Recoding is a data manipulation technique that involves changing the values or categories of a variable to make it more suitable for analysis. Recoding can be used to simplify or group the data, to correct errors or inconsistencies, or to create new variables from existing ones12

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

NEW QUESTION 6
Given the following data:

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

Which of the following BEST describes the data set?

A. There is data bias.
B. The data is incomplete.
C. The data is inconsistent.

D. The data is outliers.

Answer: C

Explanation:
 This is because inconsistency is a type of data quality issue that occurs when the data does not follow a common format, structure, or rule across different sources or systems, which can affect the efficiency and performance of the analysis or process. Inconsistency can be caused by having different spellings, punctuations, capitalizations, or abbreviations for the same or similar values in a data set, such as ??M??, ??m??, ??Male??, or ??male?? for gender in this case. Inconsistency can be eliminated or reduced by using data cleansing techniques, such as standardizing or normalizing the data values. The other options are not correct descriptions of the data set. Here is why:
? Data bias is a type of data quality issue that occurs when the data is not representative or proportional of the population or the parameter, which can affect the validity and reliability of the analysis or process. Data bias can be caused by having a sample that is too small, too large, or too skewed for the population or the parameter, such as having only male customers for a product that targets both genders in this case. Data bias can be eliminated or reduced by using sampling techniques, such as stratified or cluster sampling.
? The data is incomplete is a type of data quality issue that occurs when the data is
absent or missing in a data set, which can affect the accuracy and reliability of the analysis or process. The data is incomplete can be caused by various factors, such as human error, system error, or non-response. The data is incomplete can be addressed by using various methods, such as replacing or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression.
? The data is outliers is a type of data quality issue that occurs when the data has
values that are unusually high or low compared to the rest of the data set, which can affect the quality and validity of the analysis or process. The data is outliers can be caused by various factors, such as measurement error, natural variation, or extreme events. The data is outliers can be addressed by using various methods, such as removing or filtering out the outliers, or using robust statistics that are less sensitive to outliers, such as median, interquartile range, or box plot.

NEW QUESTION 7
Given the table below:

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

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

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

Answer: C

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

NEW QUESTION 8
Analytics reports should follow corporate style guidelines.

A. True.
B. False.

Answer: A

NEW QUESTION 9
A JSON file is an example of:

A. structured data.
B. web data.
C. machine data.
D. processed data.

Answer: A

Explanation:
A JSON (JavaScript Object Notation) file is a text-based format for representing structured data based on JavaScript object syntax. It is commonly used for transmitting data in web applications (e.g., sending some data from the server to the client, so it can be displayed on a web page, or vice versa). JSON files are human-readable and can be interpreted by various programming languages, making them ideal for data interchange123.
JSON files typically contain an array of objects, with each object representing a record with a series of name-value pairs. This structured format is both easy to understand and write by humans and easy for machines to parse and generate4.
References:
? JSON??s official definition and syntax rules1.
? A beginner??s guide to JSON and its data types2.
? Understanding the JSON file format3.
? Detailed explanation of JSON as a structured data format4.

NEW QUESTION 10
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 10
Which of the following is a difference between a primary key and a unique key?

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

Answer: B

Explanation:
The correct answer is B. There can be only one primary key in a data set, whereas there can be multiple unique keys.
A primary key is a column or a set of columns that uniquely identifies each row in a table. A table can have only one primary key, which also enforces the NOT NULL constraint on the column(s) involved. A primary key can also be referenced by a foreign key of another table to establish a relationship between the tables12
A unique key is a column or a set of columns that also uniquely identifies each row in a table, but it is not the primary key. A table can have more than one unique key, which also allows one NULL value for the column(s) involved. A unique key can also be referenced by a foreign key of another table to establish a relationship between the tables12
Some of the differences between a primary key and a unique key are:
? A primary key creates a clustered index on the column(s), whereas a unique key creates a non-clustered index on the column(s)3
? A primary key does not allow any NULL values, whereas a unique key allows one
NULL value for the column(s)123
? A primary key can be a unique key, but a unique key cannot be a primary key12

NEW QUESTION 12
Which of the following data types would a telephone number formatted as XXX-XXX-XXXX be considered?

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

Answer: D

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

NEW QUESTION 13
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 17
Given the customer table below:

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

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

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

Answer: A

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

NEW QUESTION 20
Given the following table:

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

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

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

Answer: B

NEW QUESTION 21
The process of performing initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization is called:

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

Answer: C

Explanation:
 This is because exploratory data analysis is a type of process that performs initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization, such as box plots, histograms, scatter plots, etc. Exploratory data analysis can be used to understand and summarize the data, as well as to generate hypotheses or questions for further analysis or research. For example, exploratory data analysis can be used to identify and visualize the characteristics, features, or behaviors of the data, as well as to measure their distribution, frequency, or correlation. The other options are not types of processes that perform initial investigations on data to spot outliers, discover patterns, and test assumptions with statistical insight and graphical visualization. Here is what they mean:
? A t-test is a type of statistical method that tests whether there is a significant difference between the means of two groups or samples, such as whether there is a difference between the average exam scores of two classes in this case. A t-test

can be used to test or verify a claim or an assumption about the data, as well as to measure the confidence or the error of the estimation.
? A performance analysis is a type of process that measures whether the data
meets certain goals or objectives, such as targets, benchmarks, or standards. A performance analysis can be used to identify and visualize the gaps, deviations, or variations in the data, as well as to measure the efficiency, effectiveness, or quality of the outcomes. For example, a performance analysis can be used to determine if there is a gap between a student??s test score and their expected score based on their previous performance.
? A link analysis is a type of process that determines whether the data is connected to other datapoints, such as entities, events, or relationships. A link analysis can be used to identify and visualize the patterns, networks, or associations among the datapoints, as well as to measure the strength, direction, or frequency of the connections. For example, a link analysis can be used to determine if there is a connection between a customer??s purchase history and their loyalty program status.

NEW QUESTION 24
Which of the following differentiates a flat text file from other data types?

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

Answer: A

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

NEW QUESTION 28
A county in Illinois is conducting a survey to determine the mean annual income per household. The county is 427sq mi (2.65q km). Which of the following sampling methods would MOST likely result in a representative sample?

A. A stratified phone survey of 100 people that is conducted between 2:00 p.
B. and 3:00 p.m.
C. A systematic survey that is sent to 100 single-family homes in the county
D. Surveys sent to ten randomly selected homes within 5mi (8km) of the county??s office
E. Surveys sent to 100 randomly selected homes that are reflective of the population

Answer: D

Explanation:
Surveys sent to 100 randomly selected homes that are reflective of the population. This is because a random sample is a type of sample that is selected by using a random method, such as a lottery or a computer-generated number, which ensures that every element in the population has an equal chance of being selected. A random sample can result in a representative sample, which means that the sample reflects the characteristics and diversity of the population. By sending surveys to 100 randomly selected homes that are reflective of the population, the analyst can ensure that the sample is representative of the county??s households and their income levels. The other sampling methods are not likely to result in a representative sample. Here is why:
A stratified phone survey of 100 people that is conducted between 2:00 p.m. and 3:00 p.m. would result in a biased sample, which means that the sample favors or excludes certain groups or elements in the population. By conducting the survey only between 2:00 p.m. and 3:00 p.m., the analyst would miss out on people who are not available or reachable at that time, such as those who are working or sleeping. This could affect the representativeness and generalizability of the sample.
A systematic survey that is sent to 100 single-family homes in the county would result in an unrepresentative sample, which means that the sample does not reflect the characteristics and diversity of the population. By sending surveys only to single-family homes, the analyst would ignore other types of households, such as apartments, condos, or mobile homes. This could affect the accuracy and reliability of the sample.
Surveys sent to ten randomly selected homes within 5mi (8km) of the county??s office would result in a small sample, which means that the sample size is too low to capture the variability and diversity of the population. By sending surveys only to ten homes within a limited area, the analyst would miss out on many households that are located in different parts of the county. This could affect the precision and confidence of the sample.

NEW QUESTION 33
Which of the following techniques is used to quantify data?

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

Answer: C

Explanation:
Answer C. Coding
Coding is a technique that is used to quantify data, especially qualitative data that are not expressed numerically. Coding involves assigning codes, such as numbers, letters, symbols, or colors, to different categories or themes that emerge from the data. For example, if you have a set of survey responses that ask about the satisfaction level of customers, you can code them as follows:
? Very satisfied = 5
? Satisfied = 4
? Neutral = 3
? Dissatisfied = 2
? Very dissatisfied = 1
By coding the data, you can convert them into quantitative data that can be analyzed using statistical methods, such as calculating the mean, median, mode, frequency, or percentage of each category12.
Option A is incorrect, as decoding is not a technique that is used to quantify data, but rather a process of interpreting or translating data from one form to another. For example, decoding can involve converting binary codes into text or images, or decrypting ciphertext into plaintext3.
Option B is incorrect, as enumeration is not a technique that is used to quantify data, but rather a process of listing or naming data in a specific order. For example, enumeration can involve listing the names of the states in alphabetical order, or naming the planets in order of their distance from the sun4.

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

NEW QUESTION 37
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 40
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 45
Which of the following best describes the law of large numbers?

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

Answer: B

Explanation:
The best answer is B. As a sample size grows, its mean gets closer to the average of the whole population.
The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population. This is due to the sample being more representative of the population as it increases in size. The law of large numbers guarantees stable long-term results for the averages of some random events1
* A. As a sample size decreases, its standard deviation gets closer to the average of the whole population is not correct, because it confuses the concepts of standard deviation and mean. Standard deviation is a measure of how much the values in a data set vary from the mean, not how close the mean is to the population average. Also, as a sample size decreases, its standard deviation tends to increase, not decrease, because the sample becomes less representative of the population.
* C. As a sample size decreases, its mean gets closer to the average of the whole population is not correct, because it contradicts the law of large numbers. As a sample size decreases, its mean tends to deviate from the average of the whole population, because the sample becomes less representative of the population.
* D. When a sample size doubles, the sample is indicative of the whole population is not correct, because it does not specify how close the sample mean is to the population average. Doubling the sample size does not necessarily make the sample indicative of the whole population, unless the sample size is large enough to begin with. The law of large numbers does not state a specific number or proportion of samples that are indicative of the whole population, but rather describes how the sample mean approaches the population average as the sample size increases indefinitely.

NEW QUESTION 48
Which of the following contains alphanumeric values?

A. 10.1²
B. 13.6
C. 1347
D. A3J7

Answer: D

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

NEW QUESTION 53
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 58
A customer list from a financial services company is shown below:

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

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

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

Answer: D

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

NEW QUESTION 60
A data scientist wants to see which products make the most money and which products attract the most customer purchasing interest in their company.
Which of the following data manipulation techniques would he use to obtain this information?

A. Data append
B. Data blending
C. Normalize data
D. Data merge

Answer: B

Explanation:

The correct answer is B: Data blending.
Data blending is combining multiple data sources to create a single, new dataset, which can be presented visually in a dashboard or other visualization and can then be processed or analyzed. Enterprises get their data from a variety of sources, and users may want to temporarily bring together different datasets to compare data relationships or answer a specific question. Data append is incorrect. Data append is a process that involves adding new data elements to an existing database. An example of a common data append would be the enhancement of a company's customer files. A data append takes the information they have, matches it against a larger database of business data, allowing the desired missing data fields to be added. Normalize data is incorrect.
Data normalization is the process of structuring your relational customer database, following a series of normal forms. This improves the accuracy and integrity of your data while ensuring that your database is easier to navigate. Data merge is incorrect. Data merging is the process of combining two or more data sets into a single data set.

NEW QUESTION 61
A research analyst collects ten data points from 1.000 specimens. The analyst will not need any additional data to complete the analysis and will not need to retrieve information by specifier. Which of the following is the best data structure for the analyst to use?

A. NoSQL
B. Flat file
C. JSON
D. Relational database

Answer: B

Explanation:
A flat file is a type of data structure that stores data in a plain text format, such as CSV, TSV, or TXT. A flat file consists of one or more records, each containing one or more fields, separated by a delimiter, such as a comma, tab, or space. A flat file does not have any hierarchical or relational structure, and does not support any complex queries or operations1.
A flat file may be the best data structure for the analyst to use in this scenario, because:
? The analyst collects ten data points from 1,000 specimens, which means the data is relatively small and simple, and can be easily stored and processed in a flat file.

? The analyst will not need any additional data to complete the analysis, which means the data is static and does not require any updates or modifications.
? The analyst will not need to retrieve information by specifier, which means the data
does not require any indexing or searching by key or value.

NEW QUESTION 62
Which of the following is an example of a data-mining ETL tool?

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

Answer: A

Explanation:
 A data-mining ETL tool is a software application that performs extract, transform, and load (ETL) operations on data for data mining purposes. Data mining is the process of discovering patterns, trends, and insights from large and complex data sets. ETL tools help to prepare the data for analysis by extracting data from various sources, transforming data into a consistent and suitable format, and loading data into a data warehouse or other destination. SSIS (SQL Server Integration Services) is an example of a data-mining ETL tool that is part of Microsoft SQL Server. SSIS provides graphical tools and wizards for building and debugging ETL packages that can work with various data
sources and destinations. Therefore, the correct answer is A. References: [Data Mining - SQL Server Integration Services (SSIS) | Microsoft Docs], [What Is Data Mining? | Oracle]

NEW QUESTION 65
A development company is constructing a new Init in its apartment complex. The complex has the following floor plans:

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

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

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

Answer: D

Explanation:
 The correct answer is D. $702,500.
To find the price of the Rose unit, we need to use the average cost per square foot of the original floor plans. The average cost per square foot is calculated by dividing the price by the square footage of each unit type. Using the data from the table, we can do the following:
? Jasmine: $345,000 / 1,000 = $345 per square foot
? Orchid: $525,000 / 2,000 = $262.5 per square foot
? Azalea: $375,000 / 1,500 = $250 per square foot
? Tulip: $450,000 / 1,800 = $250 per square foot
The average cost per square foot of the original floor plans is the mean of these four values, which is ($345 + $262.5 + $250 + $250) / 4 = $276.875 per square foot.
To find the price of the Rose unit, we need to multiply the average cost per square foot by the square footage of the Rose unit. The Rose unit has a square footage of 2,535, according to the table. Therefore, the price of the Rose unit is $276.875 x 2,535 = $702,421.875.
Rounding to the nearest whole number, we get $702,500 as the price of the Rose unit.

NEW QUESTION 66
While reviewing survey data, a research analyst notices data is missing from all the responses to a single question. Which of the following methods would BEST address this issue?

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

Answer: A

Explanation:
This is because missing data is a type of data quality issue that occurs when data is absent or incomplete in a data set, which can affect the accuracy and reliability of the analysis or process. Missing data can be caused by various factors, such as human error, system error, or non-response. Missing data can be addressed by using various methods, such as replacing missing data, which means filling in or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression. The other methods are not used to address missing data. Here is why:
? Remove duplicate data is a type of method that eliminates or reduces duplicate data, which is a type of data quality issue that occurs when data is repeated or copied in a data set. Removing duplicate data does not address missing data, but rather affects the quantity and validity of the data.
? Replace redundant data is a type of method that eliminates or reduces redundant data, which is a type of data quality issue that occurs when data is unnecessary or irrelevant for the analysis or purpose. Replacing redundant data does not address missing data, but rather affects the efficiency and performance of the analysis or process.
? Remove invalid data is a type of method that eliminates or reduces invalid data, which is a type of data quality issue that occurs when data is incorrect or inaccurate in a data set. Removing invalid data does not address missing data, but rather affects the validity and reliability of the analysis or process.

NEW QUESTION 69
An analyst is required to run a text analysis of data that is found in articles from a digital news outlet. Which of the following would be the BEST technique for the analyst to apply to acquire the data?

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

Answer: A

Explanation:
This is because web scraping is a technique that allows the analyst to extract data from web pages, such as articles from a digital news outlet. Web scraping can be done using various tools and methods, such as Python libraries, browser extensions, or online services. The other techniques are not suitable for acquiring data from web pages. Here is why:
Sampling is a technique that involves selecting a subset of data from a larger population, usually for statistical analysis or testing purposes. Sampling does not help the analyst to acquire data from web pages, but rather to reduce the amount of data to be analyzed. Data wrangling is a technique that involves transforming and cleaning data to make it suitable for analysis or visualization. Data wrangling does not help the analyst to acquire data from web pages, but rather to improve the quality and usability of the data.
ETL stands for Extract, Transform, and Load, which is a process that involves moving data from one or more sources to a destination, such as a data warehouse or a database. ETL does not help the analyst to acquire data from web pages, but rather to store and organize the data.

NEW QUESTION 73
Which one of the following in NOT a common data integration tool?

A. XSS
B. ELT
C. ETL
D. APIs

Answer: A

Explanation:

Cross-site Scripting (XSS) is a security vulnerability usually found in websites and/or web applications that accept user input.
XSS is a client-side vulnerability that targets other application users, while SQL injection is a server-side vulnerability that targets the application's database. How do I prevent XSS in PHP? Filter your inputs with a whitelist of allowed characters and use type hints or type casting.

NEW QUESTION 74
An analyst needs to summarize the number of people in Chicago in 2022 using the following set of data:

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

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

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

Answer: AC

NEW QUESTION 77

A data analyst is helping a retail store categorize its customers into five different groups based on the following information:
• How recently the customers made purchases
• How frequently the customers made purchases
• How much the customers spent Given the following information:

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

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

A. CustomerJ
B. Channel, Order_Date
C. CustomerJD, Territor
D. Amount
E. CustomerJD, Order_Dat
F. Amount
G. CustomerJ
H. Quantity, Amount

Answer: C

NEW QUESTION 78

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 82
An analyst is reporting on the average income for a county and is reviewing the following data:

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

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

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

Answer: B

NEW QUESTION 87
Angela is aggregating data from CRM system with data from an employee system.
While performing an initial quality check, she realizes that her employee ID is not associated with her identifier in the CRM system.
What kind of issues is Angela facing? Choose the best answer.

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

Answer: B

Explanation:

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

NEW QUESTION 89
A business intelligence team wants to create a new dashboard in order to solve a problem statement. Which of the following is the correct order of steps the team should take?

A. Determine business needs, find data sources, validate the data, create a mock-up, and analyze the information.
B. Find data sources, determine business needs, validate the data, create a mock-u
C. and analyze the information.
D. Create a mock-up, validate the data, analyze the information, determine business needs, and find data sources.
E. Validate the data, find data sources, analyze the information, and determine business needs.

Answer: C

NEW QUESTION 90
A data analyst must fulfill a request for information that is needed weekly and should be automatically emailed to a specific set of users. Which of the following types of reports should the analyst recommend?

A. A self-service report

B. A research report
C. An ad hoc report
D. An operational report

Answer: D

Explanation:
An operational report is the most suitable type of report for information that needs to be sent out on a regular, scheduled basis, such as weekly. Operational reports are designed to provide ongoing insights into the performance of an organization??s operations and are typically automated to be distributed at set intervals. This automation can include scheduling the reports to be emailed to a specific list of recipients, making it an efficient solution for the analyst??s requirement.
Operational reports are often generated from data that is continuously updated, ensuring that the recipients receive the most current information at the time of the report??s distribution. This contrasts with ad hoc reports ©, which are usually created as needed and are not scheduled. Self-service reports (A) require users to generate the report themselves, which is not the requirement here. Research reports (B) are generally more detailed and are not typically used for regular operational updates.
References:
? The guidelines on writing email reports suggest that for regular, scheduled information dissemination, structured reports like operational reports are preferred1.
? Best practices in reporting also recommend automated and scheduled reports for consistent and timely updates, which operational reports provide2.

NEW QUESTION 93
A data analyst is creating a dashboard and trying to identify the type of information that should be included. Which of the following should the analyst consider first?

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

Answer: D

Explanation:
The answer is D. Data sources and attributes.
Short Explanation: The data analyst should consider the data sources and attributes first when creating a dashboard, because they determine what kind of information can be
included and how it can be displayed. The data sources and attributes define the origin, quality, format, and structure of the data that will be used for the dashboard. They also affect the data refresh rate, the consumer types, and the access permissions of the dashboard12
* A. Data refresh rate is not the first thing to consider, because it depends on the data sources and attributes. The data refresh rate is how often the data in the dashboard is updated or refreshed to reflect the latest changes. The data refresh rate can vary depending on the type, frequency, and availability of the data sources1
* B. Consumer types are not the first thing to consider, because they depend on the data sources and attributes. The consumer types are the intended audiences or users of the dashboard, who may have different needs, preferences, and expectations for the dashboard. The consumer types can influence the design, layout, and functionality of the dashboard. However, the consumer types cannot be determined without knowing what kind of data is available and relevant for them1
* C. Access permissions are not the first thing to consider, because they depend on the data sources and attributes. The access permissions are the rules or policies that govern who can view, edit, or share the dashboard. The access permissions can protect the confidentiality, integrity, and availability of the data in the dashboard. However, the access permissions cannot be set without knowing what kind of data is involved and who needs to access it1

NEW QUESTION 95
While reviewing survey data, an analyst notices respondents entered ??Jan,?? ??January,?? and ??01?? as responses for the month of January. Which of the following steps should be taken to ensure data consistency?

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

Answer: C

Explanation:
Filter on any of the responses that do not say ??January?? and update them to ??January??. This is because filtering and updating are data cleansing techniques that can be used to ensure data consistency, which means that the data is uniform and follows a standard format. By filtering on any of the responses that do not say ??January?? and updating them to ??January??, the analyst can make sure that all the responses for the month of January are written in the same way. The other steps are not appropriate for ensuring data consistency. Here is why:
Deleting any of the responses that do not have ??January?? written out would result in data loss, which means that some information would be missing from the data set. This could affect the accuracy and reliability of the analysis.
Replacing any of the responses that have ??01?? would not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??Jan?? and ??January??. This could cause confusion and errors in the analysis. Sorting any of the responses that say ??Jan?? and updating them to ??01?? would also not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??01?? and ??January??. This could also cause confusion and errors in the analysis.

NEW QUESTION 98
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 100
A data analyst is developing a data dictionary that aligns with a company's data management processes and policies. Which of the following best describes what should be included in the data dictionary?

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

Answer: C

NEW QUESTION 105
Which of the following value is the measure of dispersion "range" between the scores of ten students in a test.
The scores of ten students in a test are 17, 23, 30, 36, 45, 51, 58, 66, 72, 77.

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

Answer: B

Explanation:

The correct answer is: 60
Range is the interval between the highest and the lowest score.
Range is a measure of variability or scatteredness of the varieties or observations among themselves and does not give an idea about the spread of the observations around some
central value. Symbolically R = Hs - Ls.
Where R = Range; Hs is the 'Highest score' and Ls is the Lowest Score.
The scores of ten students in a test are: 17, 23, 30, 36, 45, 51, 58, 66, 72, 77. The highest score is 77 and the lowest score is 17.
So the range is the difference between these two scores Range = 77 - 17 = 60

NEW QUESTION 108
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 113
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 117
Which of the following is a relational database?

A. SQL
B. Excel

C. JSON
D. NoSQL

Answer: A

NEW QUESTION 121
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 124
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 127
A data analyst reviews the following data set:

1
3
5
7
14
10
9
10
10

Which of the following is the range value?

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

Answer: D

NEW QUESTION 130
Which of the following roles is responsible for ensuring an organization's data quality, security, privacy, and regulatory compliance?

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

Answer: B

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

NEW QUESTION 132
A data analyst has been asked to create one table that has each employee's first name, last name, sales, and address. The sales and addresses are listed in the tables below:

Table 1

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

Table 2

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

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

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

Answer: D

NEW QUESTION 133
Given the following graph:
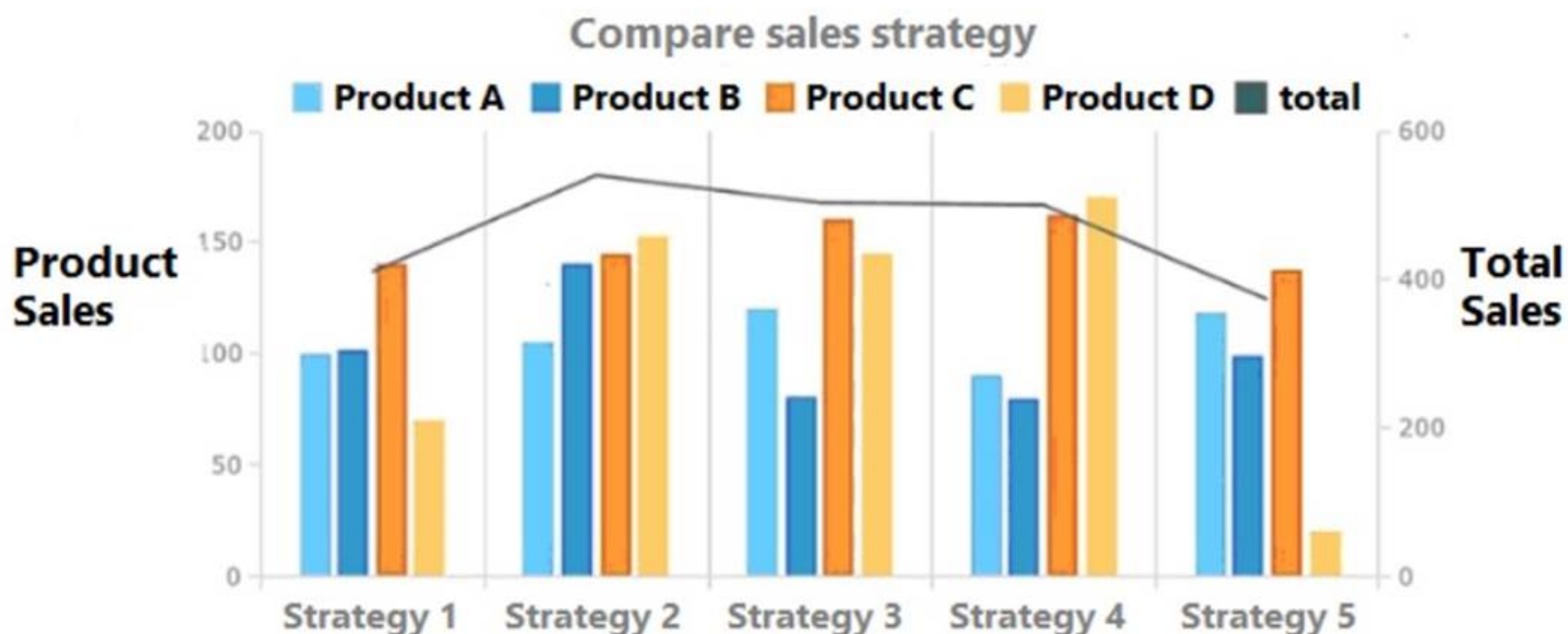
Which of the following summary statements upholds integrity in data reporting?

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

Answer: B

Explanation:
Strategy 4 provides the best sales in comparison to other strategies. This is because the total sales for Strategy 4 are the highest among all the strategies, as shown by the black line. The other statements are not accurate or do not uphold integrity in data reporting. Here is why:
Statement A is false because sales are not approximately equal for Product A and Product B across all strategies. For example, in Strategy 1, Product A has more sales than Product B, while in Strategy 3, Product B has more sales than Product A.
Statement C is misleading because it does not account for the difference in scale between the products. While Strategy 2 has the highest total sales among all products, it does not necessarily mean that it is the most effective for each product. For instance, Product D has very low sales in Strategy 2 compared to other strategies.
Statement D is biased because it does not provide any evidence or justification for why Product D should be promoted more than the other products in all strategies. It also ignores the fact that Product D has the lowest sales among all products in most of the strategies.

NEW QUESTION 138
An analyst must obtain the average daily sales for the following week:

Date	SalesTotal
2/10/2020	$36,986
2/11/2020	$37,981
2/12/2020	$40,551
2/13/2020	$42,442
2/14/2020	$56,216
2/15/2020	$81,117
2/16/2020	$63,815

Which of the following must the analyst perform to obtain this value?

A. Data normalization
B. Data append
C. Data aggregation
D. Data blending

Answer: C

Explanation:
Data aggregation is the process of compiling data from multiple sources and summarizing it into a single dataset. Data aggregation can be used to calculate statistics, such as averages, sums, counts, or percentages. In this case, the analyst must obtain the average daily sales for the following week, which is a statistic that can be calculated by aggregating the sales data from each day and dividing by the number of days. Data aggregation can be done using various tools and methods, such as spreadsheets, databases, or programming languages.

NEW QUESTION 139
Which of the following types of analysis is used when comparing last week's sales to the previous week's sales?

A. Trend analysis
B. Exploratory analysis
C. Prescriptive analysis
D. Link analysis

Answer: A

NEW QUESTION 140
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 141
A company wants to know how its customers interact with an e-commerce website based on clicks over items. Which of the following is the primary requirement for this report?

A. Data content

B. Frequency
C. Filtering
D. Views

Answer: B

NEW QUESTION 143
A data analyst has been asked to organize the table below in the following ways: By sales from high to low -
By state in alphabetic order -

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

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

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

Answer: D

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

NEW QUESTION 144
Which one of the following values will appear first if they are sorted in descending order?

A. Aaron.
B. Molly.
C. Xavier.
D. Adam.

Answer: C

Explanation:
The value that will appear first if they are sorted in descending order is Xavier. Descending order means arranging values from the largest to the smallest, or from the last to the first in alphabetical order. In this case, Xavier is the last name in alphabetical order, so it will appear first when sorted in descending order. The other names will appear in the following order: Molly, Adam, Aaron. Reference: Sorting Data - W3Schools

NEW QUESTION 149
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 151
A data analyst is designing a dashboard that will provide a story of sales and determine which site is providing the highest sales volume per customer. The analyst must choose an appropriate chart to include in the dashboard. The following data is available:

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

Which of the following types of charts should be considered?

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

Answer: C

Explanation:
 A scatter chart using sales volume and average sales per customer is the best type of chart to include in the dashboard. A scatter chart is a type of chart that displays the relationship between two numerical variables using dots or markers. A scatter chart can show how one variable affects another, how strong the correlation is between them, and how the data points are distributed. In this case, a scatter chart can show the story of sales and determine which site is providing the highest sales volume per customer by plotting the sales volume on the x-axis and the average sales per customer on the y-axis. Each dot on the chart will represent a site, and the analyst can easily compare the sites based on their position on the chart. A site with a high sales volume and a high average sales per customer will be in the upper right quadrant, indicating a high performance. A site with a low sales volume and a low average sales per customer will be in the lower left quadrant, indicating a low performance. A site with a high sales volume and a low average sales per customer will be in the lower right quadrant, indicating a high volume but low value. A site with a low sales volume and a high average sales per customer will be in the upper left quadrant, indicating a low volume but high value. A scatter chart can also show if there is a positive or negative correlation between the two variables, or if there is no correlation at all. A positive correlation means that as one variable increases, so does the other. A negative correlation means that as one variable increases, the other decreases. No correlation means that there is no relationship between the two variables.
The other types of charts are not as suitable for this purpose. A line chart is a type of chart that displays the change of one or more variables over time using lines. A line chart can show trends, patterns, and fluctuations in the data. However, in this case, there is no time variable involved, so a line chart would not be appropriate. A pie chart is a type of chart that displays the proportion of each category in a whole using slices of a circle. A pie chart can show how each category contributes to the total and compare the relative sizes of each category. However, in this case, there are two numerical variables involved, so a pie chart would not be able to show their relationship. A column chart is a type of chart that displays the comparison of one or more variables across categories using vertical bars. A column chart can show how each category differs from each other and rank them by size. However, in this case, a column chart would not be able to show the relationship between sales volume and average sales per customer, as it would only show one variable for each site.

NEW QUESTION 155
Which of the following best describes a difference between JSON and XML?

A. JSON is quicker to read and write.
B. JSON has to use an end tag.
C. JSON strings are longer
D. JSON is much more difficult to parse.

Answer: A

Explanation:
The best answer is A. JSON is quicker to read and write.
JSON (JavaScript Object Notation) is a lightweight data-interchange format that is based on the JavaScript programming language and easy to understand and generate. JSON uses a simple syntax that consists of name-value pairs and arrays, and does not require any end tags or attributes. JSON is quicker to read and write than XML (Extensible Markup Language), which is a markup language that uses a tag structure to represent data items. XML has a more complex and verbose syntax that requires end tags, attributes, and namespaces123

NEW QUESTION 158
An organization would like to add a secondary email field to its customer database in order
to enrich the customer profiles. Which of the following data manipulation techniques should the analyst use to add this information?

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

Answer: C

NEW QUESTION 159
Which of the following is the correct data type for text?

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

Answer: B

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

NEW QUESTION 160
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 161
A research analyst wants to determine whether the data being analyzed is connected to other datapoints. Which of the following is the BEST type of analysis to conduct?

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

Answer: C

Explanation:
This is because link analysis is a type of analysis that determines whether the data being analyzed is connected to other datapoints, such as entities, events, or relationships. Link analysis can be used to identify and visualize the patterns, networks, or associations among the datapoints, as well as measure the strength, direction, or frequency of the connections. For example, link analysis can be used to determine if there is a connection between a customer??s purchase history and their loyalty program status. The other types of analysis are not the best types of analysis to conduct to determine whether the data being analyzed is connected to other datapoints. Here is why:
? Trend analysis is a type of analysis that determines whether the data being analyzed is changing over time, such as increasing, decreasing, or fluctuating. Trend analysis can be used to identify and visualize the patterns, cycles, or movements in the data points, as well as measure the rate, direction, or magnitude of the changes. For example, trend analysis can be used to determine if there is a change in a company??s sales revenue over a period of time.
? Performance analysis is a type of analysis that determines whether the data being
analyzed is meeting certain goals or objectives, such as targets, benchmarks, or standards. Performance analysis can be used to identify and visualize the gaps, deviations, or variations in the data points, as well as measure the efficiency, effectiveness, or quality of the outcomes. For example, performance analysis can be used to determine if there is a gap between a student??s test score and their expected score based on their previous performance.
? Exploratory analysis is a type of analysis that determines whether there are any insights or discoveries in the data being analyzed, such as patterns, relationships, or anomalies. Exploratory analysis can be used to identify and visualize the characteristics, features, or behaviors of the data points, as well as measure their distribution, frequency, or correlation. For example, exploratory analysis can be used to determine if there are any outliers or unusual values in a dataset.

NEW QUESTION 164
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 169
Which of the following query optimization techniques involves examining only the data that is needed for a particular task?

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

Answer: C

Explanation:
The correct answer is C. Indexing documents.
Indexing documents is a query optimization technique that involves creating a data structure that allows faster access to the data in the documents. Indexing documents can reduce the amount of data that needs to be scanned for a particular query, thus improving the performance and efficiency of the query. Indexing documents can also help with searching, sorting, filtering, and aggregating the data in the documents12

NEW QUESTION 173
A database consists of one fact table that is composed of multiple dimensions. Depending on the dimension, each one can be represented by a denormalized table or multiple normalized tables. This structure is an example of a:

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

Answer: B

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

NEW QUESTION 177
A data analyst has received a data set that contains actual and projected sales for the fourth quarter of 2019. Which of the following statistical methods should the analyst use to find the measure of dispersion?

A. Mean

B. Variance
C. Correlation
D. Confidence interval

Answer: B

Explanation:
The measure of dispersion is used to describe the spread of data around a central value. In the context of a data set containing actual and projected sales, the measure of dispersion will help to understand the variability or consistency of sales figures. The variance is the most appropriate statistical method for finding the measure of dispersion because it calculates the average of the squared differences from the Mean, providing a clear picture of data spread. It is especially useful in comparing the spread between different data sets and understanding the distribution of data points.
? Mean is a measure of central tendency, not dispersion.
? Correlation measures the relationship between two variables, not the spread of a single variable.
? Confidence intervals are used to estimate the range within which a population parameter will fall, but they do not measure dispersion within the data set itself.
References:
? Measures of Dispersion in Statistics1
? Measures of Dispersion - Definition, Formulas, Examples2
? Statistical dispersion - Wikipedia3

NEW QUESTION 178
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 180
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 183
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 184
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 189
Encryption is a mechanism for protecting data. When should encryption be applied to data? Choose the best answer.

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

Answer: B

Explanation:

Correct answer B. When data is at rest or in transit.
To provide maximum protection, encrypt data both in transit and at rest.

NEW QUESTION 194
An e-commerce company recently tested a new website layout. The website was tested by a test group of customers, and an old website was presented to a control group. The table below shows the percentage of users in each group who made purchases on the websites:

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

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

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

Answer: C

Explanation:
The conclusion that is accurate at a 95% confidence interval is that in general, users who visit the new website are more likely to make a purchase. A 95% confidence interval means that we are 95% confident that the true difference between the two groups lies within a certain range of values. To calculate the 95% confidence interval, we can use the following formula:
$CI = (p1 - p2) \pm 1.96 * sqrt(p * (1 - p) * (1/n1 + 1/n2))$
where p1 and p2 are the conversion rates for the test and control groups, respectively, p is the pooled conversion rate, n1 and n2 are the sample sizes for the test and control groups, respectively, and 1.96 is the z-score for a 95% confidence level.
Using this formula, we can calculate the 95% confidence interval for each country as follows:
Country | p1 | p2 | n1 | n2 | p | CI United States | 0.12 | 0.11 | 2000 | 2000 | 0.115 | (-0.006, 0.026) Germany | 0.06 | 0.04 | 1000 | 1000 | 0.05 | (-0.002, 0.042)
United Kingdom | 0.09 | 0.07 | 1500 | 1500 | 0.08 | (-0.003, 0.053) France | 0.08 | 0.08 | 1200 | 1200 | 0.08 | (-0.024, 0.024) Canada | 0.05 | 0.03 | 800 | 800 | 0.04 | (-0.005, 0.045)
We can see that for all countries except France, the confidence interval does not include zero, which means that the difference between the test and control groups is statistically significant at a 95% confidence level. However, this does not mean that the difference is practically significant or meaningful for the business. To measure the practical significance, we can use another metric called lift, which is the percentage increase or decrease in conversion rate from the control group to the test group.
$Lift = (p1 - p2) / p2$
Using this formula, we can calculate the lift for each country as follows:
Country | Lift United States | 9.09% Germany | 50% United Kingdom |28.57% France|0% Canada|66.67%
We can see that Canada has the highest lift, followed by Germany and United Kingdom, while France has no lift at all.

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

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

Group|Weighted average Test|0.084 Control|0.072

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

CI = (p1 - p2) ?? 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 199

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

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

Answer: C

NEW QUESTION 200

A cereal manufacturer wants to determine whether the sugar content of its cereal has increased over the years. Which of the following is the appropriate descriptive statistic to use?

A. Frequency
B. Percent change
C. Variance
D. Mean

Answer: B

Explanation:

This is because percent change is a type of descriptive statistic that measures the relative change or difference of a variable over time, such as the sugar content of cereal over years in this case. Percent change can be used to determine whether the sugar content of cereal has increased over years by comparing the initial and final values of the sugar content, as well as calculating the ratio or proportion of the change. For example, percent change can be used to determine whether the sugar content of cereal has increased over years by finding out how much more (or less) sugar there is in cereal now than before, as well as expressing it as a fraction or a percentage of the original sugar content. The other descriptive statistics are not appropriate to use to determine whether the sugar content of cereal has increased over years. Here is why:

? Frequency is a type of descriptive statistic that measures how often or how likely a value or an event occurs in a data set, such as how many times a certain sugar content appears in cereal in this case. Frequency does not measure the relative change or difference of a variable over time, but rather measures the occurrence or chance of a variable at a given time.

? Variance is a type of descriptive statistic that measures how much the values in a data set vary or deviate from the mean or average of the data set, such as how much variation there is in sugar content among different cereals in this case. Variance does not measure the relative change or difference of a variable over time, but rather measures the dispersion or spread of a variable at a given time.

? Mean is a type of descriptive statistic that measures the average value or central tendency of a data set, such as what is the typical sugar content of cereal in this case. Mean does not measure the relative change or difference of a variable over time, but rather measures the summary or representation of a variable at a given time.

NEW QUESTION 204

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 206

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:
A summary with statistics, conclusions, and recommendations from the data analyst is the best way to communicate the results of an online marketing campaign to the marketing manager. A summary can provide a concise and clear overview of the most important KPIs and measure the return on marketing investment, as well as highlight the main findings and insights from the data analysis. A summary can also include actionable suggestions and best practices for improving the campaign performance and achieving the marketing objectives. A summary is different from other options, such as a real-time monitor, a self-service dashboard, or a spreadsheet of raw data, which may not provide enough context, interpretation, or guidance for the manager. Therefore, the correct answer is D. References: How to Write a Data Analysis Report: 6 Essential Tips, How to Write a Marketing Report (with Pictures) - wikiHow

NEW QUESTION 209
Which of the following is used for calculations and pivot tables?

A. IBM SPSS
B. SAS
C. Microsoft Excel
D. Domo

Answer: C

Explanation:
This is because Microsoft Excel is a type of software application that allows users to create, edit, and analyze data in spreadsheets, which are composed of rows and columns of cells that can store various types of data, such as numbers, text, or formulas. Microsoft Excel can be used for calculations and pivot tables, which are two common features or functions in data analysis. Calculations are mathematical operations or expressions that can be performed on the data in the cells, such as addition, subtraction, multiplication, division, average, sum, etc. Pivot tables are interactive tables that can summarize and display the data in different ways, such as by grouping, filtering, sorting, or aggregating the data based on various criteria or categories. The other software applications are not used for calculations and pivot tables. Here is why:
IBM SPSS is a type of software application that allows users to perform statistical analysis and modeling on data sets, such as regression, correlation, ANOVA, etc. IBM SPSS does not use spreadsheets or cells to store or manipulate data, but rather uses data views or variable views to display the data in rows and columns. IBM SPSS does not have pivot tables as a feature or function, but rather has output views or charts to display the results of the analysis.
SAS is a type of software application that allows users to perform data management and analysis using a programming language that consists of statements and commands. SAS does not use spreadsheets or cells to store or manipulate data, but rather uses data sets or tables that are stored in libraries or folders. SAS does not have pivot tables as a feature or function, but rather has procedures or macros that can produce summary tables or reports based on the data.
Domo is a type of software application that allows users to create and share dashboards and visualizations that display data from various sources and systems, such as databases, cloud services, or web applications. Domo does not use spreadsheets or cells to store or manipulate data, but rather uses connectors or APIs to access and integrate the data from different sources. Domo does not have pivot tables as a feature or function, but rather has cards or widgets that can show different aspects or metrics of the data.

NEW QUESTION 212
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 217
A data analyst needs to apply quality control concepts to a data set for accuracy. Which of the following is the best way to do this?

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

Answer: D

NEW QUESTION 219
Given the data below:

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

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

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

Answer: B

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

NEW QUESTION 220
A data analyst needs to create a weekly recurring report on sales performance and distribute it to all sales managers. Which of the following would be the BEST method to automate and ensure successful delivery for this task?

A. Use scheduled report delivery.
B. Implement subscription access delivery.
C. Print out a copy.
D. Upload the report to the server.

Answer: A

Explanation:
Scheduled report delivery is a feature that allows a data analyst to automate the generation and distribution of a report at a specified time and frequency. This would be the best method to ensure that the sales managers receive the weekly report on sales performance without manual intervention. Subscription access delivery is a feature that allows users to subscribe to a report and access it on demand, but it does not automate the delivery. Printing out a copy or uploading the report to the server are manual methods that require more time and effort from the data analyst. Reference: CertMaster Practice for Data+ Exam Prep - CompTIA

NEW QUESTION 225
An analyst is designing a dashboard that will provide a story of the sales and sales customer ratio. The following data is available:

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

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

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

Answer: A

Explanation:
For a dashboard that aims to tell a story about sales and the sales customer ratio, a column chart is an effective choice. Column charts are particularly useful for showing data changes over a period of time or for illustrating comparisons among items. In this case, a column chart can clearly display the sales figures for each site, allowing for easy comparison across different sites. Additionally, it can be used to represent the sales customer ratio by showing the proportion of sales per customer, which can provide insights into customer behavior and sales effectiveness.
? Line charts are best suited for displaying data trends over time, rather than for comparing individual categories.
? Pie charts could show the proportion of sales for each site, but they are not as effective as column charts for comparing multiple categories.
? Scatter charts are used to show the relationship between two variables, which is not the focus in this scenario.
References:
? Effective Use of Column Charts1
? Choosing the Right Chart for Your Data2
? Sales Dashboards: Examples & Templates3

NEW QUESTION 226
An analyst is updating a customer contacts database with information obtained from a survey of new customers. Which of the following data manipulation techniques should the analyst use?

A. Join
B. Append
C. Transform
D. Blend

Answer: B

NEW QUESTION 227
Which of the following actions should be taken when transmitting data to mitigate the chance of a data leak occurring? (Choose two.)

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

Answer: DE

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

NEW QUESTION 231
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 234
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 237
Which of the following would be the best way to identify multicollinear attributes in a data set?

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

Answer: A

Explanation:
Multicollinearity in a dataset refers to the situation where two or more predictor variables are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In such cases, the correlation coefficient is a key statistical measure used to identify the presence of multicollinearity. It quantifies the degree to which two variables are linearly related.
The Variance Inflation Factor (VIF) is another commonly used metric that is derived from the correlation coefficient. It assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be equal to 1.
While the other options listed—Chi-squared test, Two-sample f-test, and Two-way ANOVA—are valuable statistical tools, they serve different purposes and are not typically used to detect multicollinearity. The Chi-squared test is used for testing relationships between categorical variables, the Two-sample f-test compares variances across groups, and Two-way ANOVA is used to understand the interaction between two independent categorical variables on a continuous dependent variable.
References:
? Multicollinearity in Regression Analysis: Problems, Detection, and Solutions1.
? What is multicollinearity and how to remove it?2.
? Detect and Treat Multicollinearity in Regression with Python3.

NEW QUESTION 239
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 244

A data analyst needs to create a data visualization that aids in un the cumulative impact of sequentially introduced values that are positive or negative. Which of the following data visualization methods should the analyst use?

A. A bubble chart
B. A waterfall chart
C. A scatter plot
D. A line chart

Answer: B

Explanation:

A waterfall chart is a type of data visualization that shows the cumulative impact of sequentially introduced values that are positive or negative. A waterfall chart typically has an initial value and a final value, with intermediate values shown as floating columns that either add to or subtract from the initial value. A waterfall chart can help visualize how different factors contribute to a net change in a value over time. Therefore, the correct answer is B. References: [Waterfall Chart | Definition & Examples - Investopedia], [Waterfall Charts in Excel | How to Create Waterfall Chart in Excel?] 4of30

NEW QUESTION 247

A data analyst needs to create a dashboard to help identify trends in the data sets. Which of the following is an appropriate consideration for dashboard development?

A. Data sources and attributes
B. Frequently asked questions
C. A report from the data source
D. A comparison of data sets

Answer: A

Explanation:

When creating a dashboard to identify trends in data sets, the most appropriate consideration is the data sources and attributes. This is because the quality, reliability, and structure of the data sources directly influence the dashboard??s ability to accurately reflect trends. Attributes, such as the type of data and the time frame it covers, are crucial for trend analysis. A well-designed dashboard should provide a clear and intuitive representation of the data, allowing for easy identification of trends and patterns. Frequently asked questions (B) can inform the design of the dashboard but are not a direct consideration for the development process itself. A report from the data source © might be an output of the dashboard but does not guide its development. A comparison of data sets (D) could be a feature of the dashboard, but the underlying data sources and attributes must be considered first to ensure accurate and meaningful comparisons. References:

? Best practices in dashboard design emphasize the importance of understanding and consolidating different data sources and creating a mix of useful metrics, which aligns with the choice of data sources and attributes1.

? Fundamental dashboard design principles include the clear and efficient display of information, which is dependent on the proper selection and use of data sources and attributes2.

? Effective dashboard communication is achieved by using colors, shapes, sizes, labels, and legends meaningfully, all of which rely on the underlying data sources and attributes3.

NEW QUESTION 251

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 256

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

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

D. Summary

Answer: B

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

NEW QUESTION 257
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 259
A data analyst has been asked to create a daily manufacturing report for the floor manager Which of the following metrics should be included in the report?

A. Tons of steel produced per hour
B. Annual sales budget
C. End-of-day stock price
D. Daily corporate employee count

Answer: A

NEW QUESTION 262
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 267
An analyst needs to conduct a quick analysis. Which of the following is the FIRST step the analyst should perform with the data?

A. Conduct an exploratory analysis and use descriptive statistics.
B. Conduct a trend analysis and use a scatter chart.
C. Conduct a link analysis and illustrate the connection points.
D. Conduct an initial analysis and use a Pareto chart.

Answer: A

Explanation:

The first step the analyst should perform with the data is to conduct an exploratory analysis and use descriptive statistics. Exploratory analysis is a type of analysis that aims to summarize the main characteristics of the data, identify patterns, outliers, and relationships, and generate hypotheses for further investigation. Descriptive statistics are numerical measures that describe the central tendency, variability, and distribution of the data, such as mean, median, mode, standard deviation, range, quartiles, etc. Exploratory analysis and descriptive statistics can help the analyst gain a better understanding of the data and its quality, as well as prepare the data for further analysis.

NEW QUESTION 270
A development company is constructing a new unit in its apartment complex. The complex has the following floor plans:

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

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

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

Answer: C

Explanation:
This is because the price of the Rose unit can be estimated using the average cost per square foot of the original floor plans, which are Jasmine, Orchid, Azalea, and Tulip. To find the average cost per square foot of the original floor plans, we can use the following formula:

$$\text{Average cost per square foot} = \text{Total price} / \text{Total square feet}$$

Plugging in the values from the original floor plans, we get:

$$\text{Average cost per square foot} = (\$345,000 + \$425,000 + \$465,000 + \$525,000) / (1,000 + 1,250 + 1,500 + 2,000)$$

$$\text{Average cost per square foot} = \$1,760,000 / 5,750$$

$$\text{Average cost per square foot} = \$306$$

To find the price of the Rose unit, we can use the following formula:

$$\text{Price} = \text{Square feet} * \text{Average cost per square foot}$$

Plugging in the values from the Rose unit, we get:

$$\text{Price} = 2,300 * \$306$$

$$\text{Price} = \$705,200$$

Therefore, the price of the Rose unit should be $705,200, using the average cost per square foot of the original floor plans.

NEW QUESTION 272
After completing web scraping, which of the following file formats needs to be parsed?

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

Answer: A

Explanation:
The correct answer is .html.
Short Explanation: Web scraping is the process of extracting data from websites by parsing the HTML code of the web pages. HTML stands for HyperText Markup

Language and it is the standard markup language for creating web pages and web applications. HTML files have the extension .html and they contain tags, elements, attributes, and content that define the structure and appearance of a web page. Web scraping tools need to parse the HTML files to extract the relevant data from the web pages12

NEW QUESTION 273
Which of the following is a characteristic of a relational database?

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

Answer: C

Explanation:
It is structured in nature. This is because a relational database is a type of database that organizes data into tables, which consist of rows and columns. A relational database is structured in nature, which means that the data has a predefined schema or format, and follows certain rules and constraints, such as primary keys, foreign keys, or referential integrity. A relational database can be used to store, query, and manipulate data using a structured query language (SQL). The other characteristics are not true for a relational database. Here is why:
It utilizes key-value pairs. This is not true for a relational database, because key-value pairs are a way of storing data that associates each value with a unique key, such as an identifier or a name. Key-value pairs are typically used in non-relational databases, such as NoSQL databases, which do not have tables, rows, or columns, but rather store data in various formats, such as documents, graphs, or columns.
It has undefined fields. This is not true for a relational database, because fields are another name for columns in a table, which define the attributes or properties of each row or record in the table. Fields have defined names, types, and lengths in a relational database, which specify the format and size of the data that can be stored in each field.
It uses minimal memory. This is not true for a relational database, because memory is the amount of space or storage that is used by a database to store and process data. Memory usage depends on various factors, such as the size, complexity, and number of tables and queries in a relational database. A relational database can use a lot of memory if it has many tables with many rows and columns, or if it performs complex or frequent queries on the data.

NEW QUESTION 275
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 279
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 284

......

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual DA0-001 Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the DA0-001 Product From:

https://www.2passeasy.com/dumps/DA0-001/

Money Back Guarantee

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