

## DA0-001 Dumps

### CompTIA Data+ Certification Exam

<https://www.certleader.com/DA0-001-dumps.html>



**NEW QUESTION 1**

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

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

**Answer: B**

**NEW QUESTION 2**

A table in a hospital database has a column for patient height in inches and a column for patient height in centimeters. This is an example of:

- A. dependent data.
- B. duplicate data.
- C. invalid data
- D. redundant data

**Answer: D**

**Explanation:**

This is because redundant data is a type of data that is unnecessary or irrelevant for the analysis or purpose, which can affect the efficiency and performance of the analysis or process. Redundant data can be caused by having multiple data fields that store the same or similar information, such as patient height in inches and patient height in centimeters in this case. Redundant data can be eliminated or reduced by using data cleansing techniques, such as removing or merging the redundant data fields. The other types of data are not examples of data that is unnecessary or irrelevant for the analysis or purpose. Here is what they mean in terms of data quality:

? Dependent data is a type of data that relies on or is influenced by another data field or value, such as a formula or a calculation that uses other data fields or values as inputs or outputs. Dependent data can be useful or important for the analysis or purpose, as it can provide additional information or insights based on the existing data.

? Duplicate data is a type of data that is repeated or copied in a data set, which can affect the quality and validity of the analysis or process. Duplicate data can be caused by having multiple records or rows that have the same or similar values for one or more data fields or columns, such as customer ID or order ID. Duplicate data can be eliminated or reduced by using data cleansing techniques, such as removing or filtering out the duplicate records or rows.

? Invalid data is a type of data that is incorrect or inaccurate in a data set, which can affect the validity and reliability of the analysis or process. Invalid data can be caused by having values that do not match the expected format, type, range, or rule for a data field or column, such as an email address that does not have an @ symbol or a date that does not follow the YYYY-MM-DD format. Invalid data can be eliminated or reduced by using data cleansing techniques, such as validating or correcting the invalid values.

**NEW QUESTION 3**

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

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

**Answer: D**

**Explanation:**

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

**NEW QUESTION 4**

A user receives a large custom report to track company sales across various date ranges. The user then completes a series of manual calculations for each date range. Which of the following should an analyst suggest so the user has a dynamic, seamless experience?

- A. Create multiple reports, one for each needed date range.
- B. Build calculations into the report so they are done automatically.
- C. Add macros to the report to speed up the filtering and calculations process.
- D. Create a dashboard with a date range picker and calculations built in.

**Answer: D**

**Explanation:**

Create a dashboard with a date range picker and calculations built in. 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 track company sales across various date ranges by showing different metrics and indicators related to sales, such as revenue, volume, or growth. By creating a dashboard with a date range picker and calculations built in, the analyst can suggest a way for the user to have a dynamic, seamless experience, which means that the user can interact with and customize the dashboard according to their needs or preferences, as well as avoid any manual work or errors. For example, a date range picker is a type of feature or function that allows users to select or adjust the time period for which they want to see the data on the dashboard, such as daily, weekly, monthly, or quarterly. A date range picker can make the dashboard dynamic, as it can automatically update or refresh the dashboard with new data based on the selected time period. Calculations are mathematical operations or expressions that can be performed on the data on the dashboard, such as addition, subtraction, multiplication, division, average, sum, etc. Calculations can make the dashboard seamless, as they can eliminate the need for manual calculations for each date range, as well as ensure accuracy and consistency of the results. The other ways are not the best ways to provide a dynamic, seamless experience for the user. Here is why:

? Creating multiple reports, one for each needed date range would not provide a dynamic, seamless experience for the user, but rather create a static, cumbersome experience, which means that the user cannot interact with or customize the reports according to their needs or preferences, as well as have to deal with multiple files or pages. For example, creating multiple reports would make it difficult for the user to compare or contrast the sales across different date ranges, as well as increase the workload and complexity of managing and maintaining the reports.

? Building calculations into the report so they are done automatically would not provide a dynamic, seamless experience for the user, but rather provide a partial, limited experience, which means that the user can only benefit from one aspect or feature of the report, but not from others. For example, building calculations into the report would help with avoiding manual work or errors, but it would not help with interacting with or customizing the report according to different date ranges.

? Adding macros to the report to speed up the filtering and calculations process would not provide a dynamic, seamless experience for the user, but rather provide an advanced, complex experience, which means that the user would need to have some technical skills or knowledge to use or apply the macros, as well as face some potential risks or challenges. For example, adding macros to the report would require the user to know how to write or run the macros, which are a type of code or script that automates certain tasks or actions on the report, such as filtering or calculating the data. Adding macros to the report could also expose the user to some security or compatibility issues, such as viruses, malware, or errors.

**NEW QUESTION 5**

You are working with a professional statistician to perform an analysis and would like to use a statistics package. Which one of the following would be the most appropriate?

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 6**

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

**Original data:**

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

**Modified data:**

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

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

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

**Answer:** B

**Explanation:**

The correct answer is B. Recoding.

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

In the example, the analyst used recoding to change the values of Var001, Var002, Var003, and Var004 from numerical to textual form. The analyst also used

recoding to assign meaningful labels to the values, such as ??Absent?? for 0, ??Present?? for 1, ??Low?? for 2, ??Medium?? for 3, and ??High?? for 4. This makes the data more understandable and easier to analyze.

**NEW QUESTION 7**

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

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

**Answer:** B

**NEW QUESTION 8**

Which of the following will MOST likely be streamed live?

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 9**

An analyst is reviewing the following data: Car IDSpeed

123155  
566436  
564418  
650567  
546436  
645638

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

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

**Answer:** D

**Explanation:**

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

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

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

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

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

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

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

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

**NEW QUESTION 10**

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 dash. 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 10**

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 14**

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 15**

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

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

**Answer:** A

**NEW QUESTION 20**

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 24**

Mario works with a group of R programmers tasked with copying data from an accounting system into a data warehouse. In what phase are the group's R skills most relevant?

- A. Extract.
- B. Load.
- C. Transform.
- D. Purge.

**Answer:** C

**NEW QUESTION 29**

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

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

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

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 30**

An analyst develops an IT document and needs to describe the technical terms used in the document. Which of the following is where the analyst should include descriptions of the technical terms?

- A. Glossary
- B. System diagram
- C. User requirements
- D. Index

**Answer:** A

**Explanation:**

In technical documentation, a glossary is the designated section where definitions for technical terms are provided. It serves as a reference point for readers to understand specialized or uncommon words used within the document. Including descriptions of technical terms in a glossary ensures that readers have a consistent resource to refer to, which can improve comprehension and reduce misunderstandings<sup>12</sup>.

A system diagram (Option B) is a visual representation of the system's components and their interactions, not a place for defining terms. User requirements (Option C) outline what end-users expect from the system, and an index (Option D) is an alphabetical list of topics covered in the document, usually with page numbers, but not definitions.

References:

? Creating effective technical documentation<sup>1</sup>.

? Best practices when writing technical descriptions<sup>3</sup>.

**NEW QUESTION 31**

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 36**

Which of the following contains alphanumeric values?

- A. 10.1<sup>2</sup>
- B. 13.6
- C. 1347
- D. A3J7

**Answer:** D

**Explanation:**

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

**NEW QUESTION 37**

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 42**

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

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

**Answer:** B

**Explanation:**

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

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

**NEW QUESTION 46**

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

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

**Answer: C**

**NEW QUESTION 49**

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

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

**Answer: C**

**Explanation:**

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

**NEW QUESTION 50**

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 53**

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

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

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

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

D. \$702,500

**Answer:** D

**Explanation:**

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

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

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

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

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

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

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

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

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

**NEW QUESTION 57**

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

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 62**

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

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 63**

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

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

**Answer:** C

**Explanation:**

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

**NEW QUESTION 67**

Which one of the following is 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 71**

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

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

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

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

**Answer: B**

**Explanation:**

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

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

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

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

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

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

**NEW QUESTION 76**

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

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

**Answer: D**

**Explanation:**

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

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

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

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

**NEW QUESTION 77**

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

**Sales\_table**

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

**Order\_table**

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

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

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

**Answer:** D

**Explanation:**

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

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

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

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

The result of this query would be:

Cust\_id | Order\_id | Order\_qty | 1 | 1 | 100 | 2 | 2 | 50 | 3 | 3 | 25 | 4 | 4 |

75 NULL | 5 | 10 NULL | 6 | 20 NULL | 7 | 15

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

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

**NEW QUESTION 80**

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 84**

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**

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

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

**Answer:** B

**Explanation:**

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

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

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

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

The reason why there are seven rows and eight columns in the result table is because:  
 ? There are seven rows because there are six customers and six orders in the original tables, but only five customers have matching orders based on the customer ID column. Therefore, only five rows will have data from both tables, while one row will have data only from the customer table (customer 5), and one row will have no data at all (null values).  
 ? There are eight columns because there are four columns in each of the original tables, and all of them are selected and joined in the result table. Therefore, the result table will have four columns from the customer table (customer ID, first name, last name, and email) and four columns from the order table (order ID, order date, product, and quantity).

**NEW QUESTION 93**

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 98**

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

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

**Answer:** A

**NEW QUESTION 100**

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

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

**Answer:** C

**NEW QUESTION 102**

Which one the following is not considered an aggregate function?

- A. SUM
- B. MIN
- C. SELECT
- D. MAX

**Answer:** C

**Explanation:**

The option that is not considered an aggregate function is SELECT. An aggregate function is a function that performs a calculation on a set of values and returns a single value. Examples of aggregate functions are SUM, MIN, MAX, AVG, COUNT, etc. SELECT is not an aggregate function, but a SQL command that is used to select data from a table or a query. Reference: SQL Aggregate Functions - W3Schools

**NEW QUESTION 106**

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

Product	Q1 sales
Ground beef	\$2,667.60
Crab meet	\$1,768.41
Swiss cheese	\$3,182.40
Broccoli	\$1,509.60
Vegetable spread	\$3,202.87

Which of the following is the mean?

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

D. \$12,330.88

**Answer:** C

**Explanation:**

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

**NEW QUESTION 111**

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 115**

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

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

**Answer:** D

**Explanation:**

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

References:

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

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

? Educational performance data analysis often involves comparing scores across different schools or classrooms, which is facilitated by a well-designed dashboard<sup>2</sup>.

**NEW QUESTION 116**

Which of the following tools would be best to use to calculate the interquartile range, median, mean, and standard deviation of a column in a table that has 5,000,000 rows?

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

**Answer:** B

**NEW QUESTION 121**

A gambler thinks that a coin is fair and is equally likely to turn up heads or tails when the coin is flipped. Which of the following tests should the gambler use to test this hypothesis?

- A. t-test
- B. Chi-squared test
- C. Rank sum test
- D. Ratio test

**Answer:** B

**NEW QUESTION 125**

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

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

**Answer:** D

**NEW QUESTION 130**

Which of the following is an example of structured data?

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

**Answer:** A

**Explanation:**

A credit card number is an example of structured data, which is a type of data that conforms to a data model, has a well-defined structure, follows a consistent order, and can be easily accessed and used by a person or a computer program. A credit card number consists of 16 digits that are divided into four groups of four digits each, separated by spaces or hyphens. The first six digits indicate the issuer identification number, the next nine digits indicate the account number, and the last digit is a check digit that validates the number. A credit card number can be stored and processed in a structured format, such as a database or a spreadsheet<sup>1</sup>.

**NEW QUESTION 135**

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

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

**Answer:** C

**NEW QUESTION 138**

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

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

**Answer:** B

**Explanation:**

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

**NEW QUESTION 139**

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

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

**Answer:** B

**Explanation:**

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

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

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

big data techniques and technologies can manage and analyze the huge volume of weather data from different resources, which supports the use of public databases<sup>1</sup>.

? NOAA's approach to detecting severe weather events using instruments and receiving information from storm spotters indicates the importance of reliable, collected data, which is typically stored in public databases<sup>2</sup>.

? The National Weather Service's use of observational data collected by various instruments, which are then fed into forecast models, further emphasizes the value of established data collection methods over individual observations or surveys<sup>3</sup>.

**NEW QUESTION 144**

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

Date	Sales Total
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 145**

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

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

**Answer:** C

**Explanation:**

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

Short Explanation: A subscription is a delivery method that allows the client to access the dashboard on-demand, whenever they need it. A subscription can be set up by the data analyst or the client themselves, and it can be configured to send an email notification when the dashboard is updated or refreshed. A subscription also allows the client to view the dashboard online or download it as a file format of their choice<sup>12</sup>

\* A. Email is not the best delivery method because it does not allow the client to access the dashboard on-demand. Email deliveries are sent at a fixed time or frequency, and they may not reflect the latest data or changes in the dashboard. Email deliveries also have limitations on the file size and format of the dashboard attachments<sup>1</sup>

\* B. Scheduled is not the best delivery method because it does not allow the client to access the dashboard on-demand. Scheduled deliveries are similar to email deliveries, except that they are triggered by a specific event or condition, such as a data update or a threshold value. Scheduled deliveries also have the same limitations as email deliveries on the file size and format of the dashboard attachments<sup>1</sup>

\* D. Static is not the best delivery method because it does not allow the client to access the dashboard on-demand. Static deliveries are one-time deliveries that are manually generated by the data analyst or the client. Static deliveries do not update or refresh automatically, and they may become outdated or irrelevant over time. Static deliveries also have limitations on the file size and format of the dashboard files<sup>3</sup>

**NEW QUESTION 149**

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

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

**Answer:** C

**Explanation:**

INSERT

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

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

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

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

**NEW QUESTION 152**

Each month an analyst needs to execute a data pull for the two prior months. Which of the following is the most efficient function for the analyst to use?

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

**Answer:** B

**Explanation:**

The most efficient function for an analyst to execute a data pull for the two prior months would be the Date function. This function allows for the manipulation and formatting of date values within a database. Using Date functions, an analyst can dynamically calculate the start and end dates for the previous two months, ensuring that the data pull is accurate and automated without manual intervention.

For example, SQL functions like DATEADD and DATEDIFF can be used to determine the exact range of dates needed for the data pull. These functions can calculate the first and

last day of the previous months relative to the current date, which is essential for monthly reporting and analysis.

References:

? Discussions on Stack Overflow suggest using SQL date functions

like DATEADD and DATEDIFF to dynamically extract data for previous months, which supports the use of Date functions<sup>12</sup>.

? The use of Date functions is also recommended for ensuring that the data pull is

not only efficient but also accurate, as it avoids potential errors associated with manual date entry<sup>3</sup>.

**NEW QUESTION 153**

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

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

**Answer:** C

**Explanation:**

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

References:

? Medium article on SQL Query Optimization<sup>1</sup>.

? MSSQLTips on SQL Query Performance<sup>2</sup>.

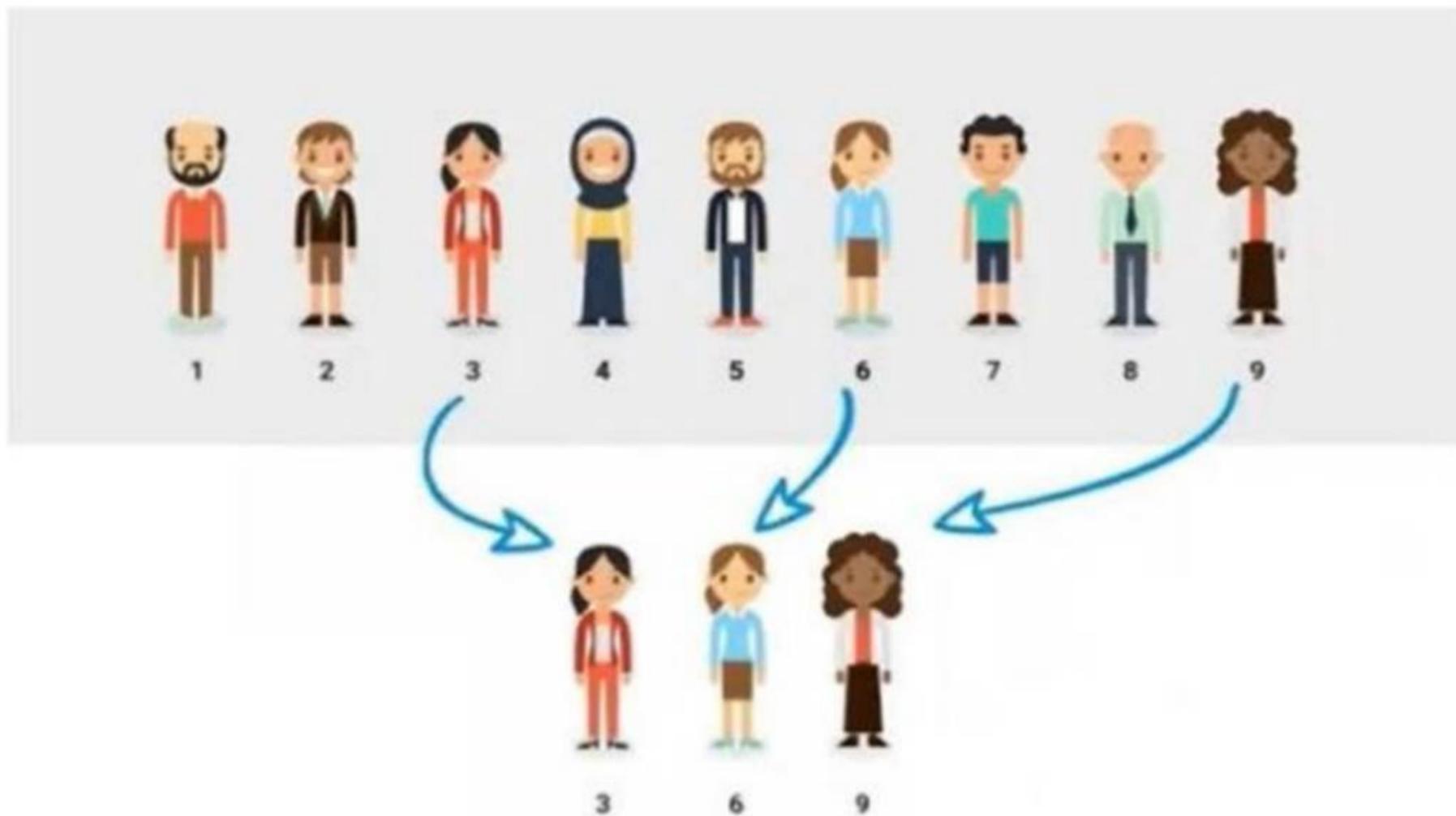
? Blog post on SQL Performance Optimization<sup>3</sup>.

? SQL Easy guide on improving SQL Query Performance<sup>4</sup>.

? LearnSQL.com on SQL for Data Analysis<sup>5</sup>.

**NEW QUESTION 156**

Given the diagram below:



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

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

**Answer:** D

**Explanation:**

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

**NEW QUESTION 160**

An analyst wants to extract data from a variety of sources and store the data in a cloud-based environment prior to cleaning. Which of the following integration techniques should the analyst use?

- A. ETL
- B. API
- C. SQL
- D. ELT

**Answer:** A

**NEW QUESTION 161**

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

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

**Answer:** B

**NEW QUESTION 164**

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 169**

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 174**

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

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

Customer Table -  
In-store Transactions –

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

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

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

**Answer: C**

**Explanation:**

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

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

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

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

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

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

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

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

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

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

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

**NEW QUESTION 178**

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

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

**Answer: C**

**Explanation:**

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

**NEW QUESTION 181**

Which of the following is the best description of discrete data types?

- A. Non-numeric data used to describe attributes of a population sample
- B. The frequency of the number of times each value occurs by using whole numbers

- C. Numeric values that can be measured on a continuous scale
- D. Non-numeric data used to describe attributes of a population sample ranked in a specific order

**Answer:** B

**NEW QUESTION 183**

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

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 185**

A data analyst is performing a data merge within a spreadsheet using the tables below:  
<https://www.bing.com/images/blob?bcid=S1XCF9p02M4GjpbGxHj0Irlaj9sw.....4c>

Table 1

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

Table 2

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

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

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

**Answer:** D

**Explanation:**

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

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

**NEW QUESTION 186**

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 189**

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 191**

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 195**

An analyst has conducted a review of business questions. Which of the following should the analyst do next to conduct an analysis?

- A. Determine the data needs and review the observations.
- B. Determine the data needs and sources for analysis.
- C. Determine the data needs and schedule interviews.
- D. Determine the data needs and begin the analysis.

**Answer: B**

**Explanation:**

After conducting a review of the business questions, the next step for the analyst is to determine the data needs and sources for analysis. This involves identifying the relevant data elements, variables, and metrics that are required to answer the business questions, as well as the data sources, formats, and quality that are available to access and use. This step will help the analyst to plan the data collection, preparation, and integration processes, as well as to assess the feasibility and limitations of the analysis.

**NEW QUESTION 197**

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

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 199**

Given the following data table:

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

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

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

**Answer:** BD

**Explanation:**

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

? Missing Data (B): Missing data can lead to incomplete analysis and biased results. It is essential to identify and address gaps in the dataset to maintain the integrity of the analysis<sup>1</sup>.

? Invalid Data (D): Invalid data includes entries that are out of range, improperly formatted, or illogical (e.g., a negative age). Such data can corrupt analysis and lead to incorrect conclusions<sup>1</sup>.

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

? Understanding the importance of data quality and the impacts of missing and invalid data on research outcomes<sup>1</sup>.

? Best practices in data cleansing<sup>2</sup>.

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

**NEW QUESTION 201**

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

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

**Answer:** DE

**Explanation:**

\* 1. Duplicates

\* 2. Misspellings

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

- Blanks
- Nulls
- Outliers
- Duplicates

- Extra spaces
- Misspellings
- Abbreviations and domain-specific variations
- Formula error codes

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

In either case, you're scanning for the anomalies yourself. Suffice it to say that's not a very efficient process. It also means accuracy is only as good as the analyst's eye, so the probability of error varies throughout the day.

**NEW QUESTION 205**

Which of the following is a process that is used during data integration to collect, blend, and load data?

- A. MDM
- B. ETL
- C. OLTP
- D. BI

**Answer: B**

**Explanation:**

ETL is a process that is used during data integration to collect, blend, and load data. ETL stands for extract, transform, and load, which are the three main steps involved in moving data from different sources to a common destination, such as a data warehouse or a data lake. ETL helps to consolidate and standardize data for analysis and reporting purposes. References: CompTIA Data+ Certification Exam Objectives, page 12

**NEW QUESTION 207**

Which of the following is an object associated with a table that sorts and stores table row data in a key-value pair?

- A. Foreign key
- B. Function
- C. Stored procedure
- D. Clustered index

**Answer: D**

**NEW QUESTION 209**

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

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

**Answer: A**

**Explanation:**

The correct answer is A. UNION ALL.

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

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

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

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

**NEW QUESTION 211**

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

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

**Answer: D**

**Explanation:**

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

**NEW QUESTION 215**

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

- A. Rephrase the business requirement.
- B. Determine the data necessary for the analysis.

- C. Build a mock dashboard/presentation layout.
- D. Perform exploratory data analysis.

**Answer:** B

**Explanation:**

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

**NEW QUESTION 220**

A reporting analyst is creating a dashboard that shows the year-over-year performance for a sales organization. Which of the following is the best visual for the analyst use to illustrate the organization's performance?

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

**Answer:** D

**NEW QUESTION 222**

Which one of the following is a measure of dispersion?

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

**Answer:** A

**NEW QUESTION 227**

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

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

**Answer:** AE

**Explanation:**

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

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

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

**NEW QUESTION 229**

Which of the following best describes an exploratory analysis?

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 234**

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 236**

Which of the following should an analyst do to best summarize the data on a data set?

- A. Filtering
- B. Aggregation
- C. Sorting
- D. Concatenation

**Answer:** B

**NEW QUESTION 241**

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 242**

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 244**

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 247**

What R package makes it easy to work with dates?

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 252**

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 254**

Which of the following is an example of a discrete variable?

- A. The temperature of a hot tub
- B. The height of a horse
- C. The time to complete a task
- D. The number of people in an office

**Answer:** D

**Explanation:**

A discrete variable is a variable that can only take on a finite number of values, such as integers or categories. The number of people in an office is an example of a discrete variable, as it can only be a whole number. The temperature of a hot tub, the height of a horse, and the time to complete a task are examples of continuous variables, as they can take on any value within a range. Reference: CompTIA Data+ (DA0-001) Practice Certification Exams | Udemy

**NEW QUESTION 259**

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

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

**Answer:** C

**Explanation:**

Answer: C. 2019 goal data

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

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

Option B is incorrect, as 2018 actual revenue is not relevant for determining 2019 performance attainment. The 2018 actual revenue contains the actual sales revenue for the year 2018, which is not comparable with the 2019 goal data or the 2019 actual revenue. Option D is incorrect, as 2019 commission plan is not relevant for determining 2019 performance attainment. The 2019 commission plan contains the rules and rates for calculating and paying commissions to the sales team based on their performance attainment, but it does not contain the expected or planned sales revenue for the year 2019.

**NEW QUESTION 262**

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

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

D. Stratified

**Answer:** D

**Explanation:**

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

**NEW QUESTION 266**

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 source<sup>1234</sup>.

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 269**

An analyst notices changes in sales ratios when analyzing a quarterly report. Which of the following is the analyst conducting?

- A. A gap analysis
- B. A link analysis
- C. A trend analysis
- D. A statistical analysis

**Answer:** C

**Explanation:**

When an analyst observes changes in sales ratios over a period, such as in a quarterly report, they are conducting a trend analysis. Trend analysis is a statistical method used to examine and evaluate the movement of data points over time to identify patterns or trends. This type of analysis is particularly useful for forecasting future events based on historical data. It differs from gap analysis, which assesses the difference between actual performance and potential or desired performance; link analysis, which is used to find associations among data; and statistical analysis, which is a broad term for all types of data analysis methods, including trend analysis.

References:

- ? Investopedia article on Ratio Analysis<sup>1</sup>.
- ? SpringerLink chapter on Financial Ratios Analysis<sup>2</sup>.
- ? ExamTopics page mentioning sales ratios in the context of analysis<sup>3</sup>.
- ? Investopedia definition of Ratio Analysis<sup>4</sup>.
- ? LiveWell article on Financial Ratio Analysis<sup>5</sup>.

**NEW QUESTION 271**

A data analyst is developing a dashboard to track and monitor metrics. Which of the following best practices should be taken into during the FIRST pment process?

- A. Create a Aupirarrame:
- B. Deploy to production.
- C. Copy a dashboard design from the Internet.
- D. Develop a dashboard.

**Answer:** A

**Explanation:**

A dashboard is a graphical display that summarizes and presents key performance indicators (KPIs) and metrics for a business or a project. A dashboard should be clear, concise, and easy to understand. To develop a dashboard, one of the best practices is to create a wireframe or a mockup first. A wireframe or a mockup is a low-fidelity sketch or prototype of the dashboard layout and design, which helps to define the scope, requirements, and functionality of the dashboard. Creating a wireframe or a mockup can help to save time and resources, as well as to get feedback from stakeholders and users before deploying the dashboard to production. Therefore, the correct answer is A. References: [Dashboard Design Best Practices: 4 Key Principles | Toptal], [How to Create an Effective Dashboard (with Examples) | Tableau]

**NEW QUESTION 272**

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

- A. SAS

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

**Answer:** D

**Explanation:**

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 274**

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 278**

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 281**

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 282**

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 286**

What subset of Structured Query Language (SQL) is used to add, remove, modify, or retrieve the information stored within a relational database?

- A. DDL.
- B. DSL.
- C. DQL.
- D. DML.

**Answer:** D

**Explanation:**

Correct answer D. DML.

The Data Manipulation Language (DML) is used to work with the data stored in a database.

DML includes the SELECT, INSERT, UPDATE, and DELETE commands.

The Data Definition Language (DDL) contains the commands used to create and structure a relational database. It includes the CREATE, ALTER, and DROP commands.

DDL and DML are the only two sublanguages of SQL.

**NEW QUESTION 289**

Which of the following file formats is best suited to start exploratory analysis within statistical software?

- A. CSV
- B. XLSM
- C. XML
- D. JSON

**Answer:** A

**NEW QUESTION 293**

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: D**

**Explanation:**

The best type of chart to display the data is D. Include a column chart using the site and sales to average sales per customer.

A column chart is a good choice for comparing categorical data with numerical data, such as the site and sales to average sales per customer. A column chart can show the relative differences between the sites and highlight the site with the highest sales volume per customer. A column chart can also be easily labeled and formatted to make the data clear and understandable.

A line chart is not suitable for this data, because it is used to show trends or changes over time, which is not relevant for the site and sales to average sales per customer data. A line chart would also be confusing and misleading, as it would imply a connection or correlation between the sites that does not exist.

A pie chart is also not a good choice for this data, because it is used to show the proportion of a whole, not the comparison of different categories. A pie chart would also be difficult to read and interpret, as it would require labels or legends to identify the sites and their sales to average sales per customer. A pie chart would also not be able to show the exact values of the sales to average sales per customer, only their relative sizes.

A scatter chart is another inappropriate option for this data, because it is used to show the relationship or correlation between two numerical variables, not between a categorical and a numerical variable. A scatter chart would also be cluttered and unclear, as it would plot each site as a point on a coordinate plane, without any labels or axes. A scatter chart would also not be able to show the differences or rankings between the sites and their sales to average sales per customer.

**NEW QUESTION 295**

What would be an example of an acceptable form of primary identification for the Data+ exam?

- A. Passport.
- B. School ID card.
- C. Employee ID card.
- D. Credit card with photo and signature.

**Answer: A**

**NEW QUESTION 299**

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

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

**Answer: D**

**Explanation:**

The best format to deliver this report is D. A dashboard with filters at the top that the user can toggle.

A dashboard is a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance. A dashboard with filters at the top that the user can toggle would allow the user to easily and quickly access the information they need about various regions, products, and time periods, without having to navigate through multiple tabs, pages, or emails. A dashboard with filters would also enable the user to compare and contrast different views of the data and see how they change over time. A dashboard with filters would also be more interactive and engaging than a static or email report.

A workbook with multiple tabs for each region would not be an efficient way to deliver this report, because it would require the user to switch between different tabs to see the information they need. This would make it harder to compare and contrast different regions, products, and time periods, and also increase the risk of

errors or confusion. A workbook with multiple tabs would also be less visually appealing and more cluttered than a dashboard3

A daily email with snapshots of regional summaries would not be an efficient way to deliver this report, because it would limit the user's ability to explore the data in depth and customize their view. A daily email would also be dependent on the frequency and timing of the email delivery, which might not match the user's needs or preferences. A daily email

would also be more likely to be ignored or deleted than a dashboard that is always accessible.

A static report with a different page for every filtered view would not be an efficient way to deliver this report, because it would create a very long and cumbersome report that would be difficult to read and understand. A static report would also not allow the user to change or update the filters as they wish, or see how the data changes over time. A static report would also be less interactive and engaging than a dashboard.

#### **NEW QUESTION 300**

.....

## Thank You for Trying Our Product

\* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

\* One year free update

You can enjoy free update one year. 24x7 online support.

\* Trusted by Millions

We currently serve more than 30,000,000 customers.

\* Shop Securely

All transactions are protected by VeriSign!

**100% Pass Your DA0-001 Exam with Our Prep Materials Via below:**

<https://www.certleader.com/DA0-001-dumps.html>