

## DA0-001 Dumps

### CompTIA Data+ Certification Exam

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



**NEW QUESTION 1**

A data set was recorded using multimedia technology. Which of the following is a necessary step on the way to interpretation?

- A. Structural equation modeling
- B. Transcription
- C. Sequential analysis
- D. Sampling

**Answer:** B

**Explanation:**

The correct answer is B. Transcription.

Transcription is a necessary step on the way to interpretation when a data set was recorded using multimedia technology. Multimedia technology refers to the use of various forms of media, such as audio, video, images, and text, to capture and present information<sup>1</sup> Transcription is the process of converting multimedia data into written or textual form, which can then be analyzed using various methods and tools<sup>2</sup> Transcription can help to make the data more accessible, searchable, and manageable, as well as to preserve the data for future use.

Structural equation modeling is not correct, because it is a statistical technique that tests the causal relationships between multiple variables using observed and latent variables. Structural equation modeling is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data.

Sequential analysis is not correct, because it is a method of analyzing the order and timing of events or behaviors in a data set. Sequential analysis is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data. Sampling is not correct, because it is the process of selecting a subset of data from a larger population for analysis. Sampling is not a necessary step on the way to interpretation, but rather a preliminary step that can be done before collecting or analyzing the data.

**NEW QUESTION 2**

Which of the following concepts should be applied if a data set with 40 fields needs to be pared down to 20 fields and contains similar data across multiple fields?

- A. Duplication
- B. Consolidation
- C. Compliance
- D. Standardization

**Answer:** B

**Explanation:**

Consolidation is the process of combining multiple elements into a single, more effective or coherent whole. In the context of data analytics, consolidation would involve merging similar fields to reduce the overall number of fields in a dataset. This is particularly useful when a dataset contains redundant or similar data across multiple fields, as it helps to simplify the data structure and improve efficiency. Techniques such as dimensionality reduction are often applied to achieve this, where the goal is to retain the most informative and representative features of the data while reducing the number of total features. References:

? Applied Dimensionality Reduction — 3 Techniques using Python<sup>1</sup>.

? Seven Techniques for Data Dimensionality Reduction<sup>2</sup>.

? Best practices when working with datasets<sup>3</sup>.

? Effectively Handling Large Datasets<sup>4</sup>.

**NEW QUESTION 3**

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

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

**Answer:** D

**Explanation:**

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

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

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

**NEW QUESTION 4**

Which of the following is a domain-specific language used in programming that is designed for managing data that is held in a relational data stream management

system?

- A. SAS
- B. SQL
- C. Python
- D. R

**Answer:** B

**Explanation:**

SQL (Structured Query Language) is a domain-specific language used in programming, specifically designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS). It is the standard language for relational database management systems. SQL statements are used to perform tasks such as update data on a database, or retrieve data from a database. Unlike languages like Python or R, which are general-purpose programming languages, SQL is tailored specifically for database management and manipulation.

References:

- ? ResearchGate article on SQL1.
- ? SpringerLink chapter on Relational Databases and SQL Language2.
- ? DataCamp tutorial on SQL Server Installation3.
- ? Wikipedia page on SQL4.

**NEW QUESTION 5**

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

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

**Answer:** DF

**NEW QUESTION 6**

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

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

**Answer:** C

**Explanation:**

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

**NEW QUESTION 7**

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

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

**Answer:** B

**Explanation:**

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

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

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

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

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

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

**NEW QUESTION 8**

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

Original data:

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

Modified data:

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

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

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

**Answer:** B

**Explanation:**

The correct answer is B. Recoding.

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

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

#### NEW QUESTION 9

A data analyst wants to create "Income Categories" that would be calculated based on the existing variable "Income". The "Income Categories" would be as follows:

Income category 1: less than \$1.

Income category 2: more than \$1 and less than \$20,000. Income category 3: more than \$20,001 and less than \$40,000. Income category 4: more than \$40,001.

Which of the following data manipulation techniques should the data analyst use to create "Income Categories"?

- A. Data merge
- B. Derived variables
- C. Data blending
- D. Data append

**Answer:** B

**Explanation:**

The correct answer is B: Derived variables Derived variables are variables that you create by calculating or categorizing variables that already exist in your data set.

Data merge is incorrect. Data merging is the process of combining two or more data sets into a single data set. Data blending is incorrect.

Data blending involves pulling data from different sources and creating a single, unique, dataset for visualization and analysis.

Data append is incorrect. A data append is a process that involves adding new data elements to an existing database.

#### NEW QUESTION 10

Analytics reports should follow corporate style guidelines.

- A. True.
- B. False.

**Answer:** A

#### NEW QUESTION 10

An analyst collected data that includes primary account numbers, expiration dates, and service codes. Which of the following data governance classifications is used to describe this data?

- A. PII
- B. PCI
- C. PBI
- D. PHI

**Answer:** B

#### NEW QUESTION 14

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

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

**Answer:** B

#### Explanation:

Normalization is the process of reorganizing data in a database so that it meets two basic requirements: There is no redundancy of data, all data is stored in only one place. Data dependencies are logical, all related data items are stored together.

Put simply, data normalization ensures that your data looks, reads, and can be utilized the same way across all of the records in your customer database. This is done by standardizing the formats of specific fields and records within your customer database.

#### NEW QUESTION 18

A data analyst needs to collect a similar proportion of data from every state. Which of the following sampling methods would be the most appropriate?

- A. Systematic sampling
- B. Convenience sampling
- C. Stratified sampling
- D. Random sampling

**Answer:** C

#### Explanation:

The best sampling method for the data analyst's need is C. Stratified sampling.

Stratified sampling is a type of probability sampling that involves dividing the population into homogeneous groups or strata based on some characteristic, such as state, and then randomly selecting a proportional number of individuals from each stratum. Stratified sampling ensures that every group is adequately represented in the sample, and reduces the sampling error and variability<sup>12</sup>

Systematic sampling is not correct, because it involves selecting every nth individual from the population, starting from a random point. Systematic sampling does not guarantee that every state will have a similar proportion of data in the sample, and may introduce bias or error if there is a hidden pattern or order in the population<sup>12</sup>

Convenience sampling is not correct, because it involves selecting individuals who are easily accessible or available to the researcher. Convenience sampling is a type of non-probability sampling that does not involve random selection, and may result in a biased or unrepresentative sample<sup>12</sup>

Random sampling is not correct, because it involves selecting individuals from the population at random, without any grouping or stratification. Random sampling may not produce a sample that has a similar proportion of data from every state, especially if the population is large or heterogeneous. Random sampling may also have a higher sampling error and variability than stratified sampling<sup>12</sup>

#### NEW QUESTION 19

A data analyst has a set with more than 40,000 rows in the sample schema below:

Name	Birth date - sales system	Birth date - marketing system	Birth date - accounting system
Tom	1/4/1989		
Frank		7/5/1994	
Carrie		8/3/1973	
Joe			3/2/2001

The analyst would like to create one column that contains the customers' birth dates. Which of the following data quality dimensions would BEST explain the reason for compilation?

- A. Data accuracy
- B. Data completeness
- C. Data duplication
- D. Data integrity

**Answer:** D

**Explanation:**

Data integrity is the dimension that measures the consistency and validity of data across different data sources. In this case, the data analyst wants to create one column that contains the customers' birth dates, but the data is stored in different formats and locations in the sample schema. For example, some customers have their birth dates in the customer table, while others have their birth years in the sales table. To compile the data into one column, the data analyst needs to ensure that the data is consistent and valid across the tables. Therefore, data integrity is the best explanation for the reason for compilation. References: Data Quality Dimensions - DATAVERSITY, The 6 Data Quality Dimensions with Examples | Collibra

**NEW QUESTION 20**

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

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

**Answer:** D

**Explanation:**

A star schema is a type of database schema that consists of one fact table and multiple dimension tables. The fact table contains the measures or metrics of the business process, such as sales, orders, or transactions. The dimension tables contain the attributes or characteristics of the business entities, such as products, customers, or locations. The fact table is connected to the dimension tables by foreign keys that reference the primary keys of the dimension tables. The fact table is located at the center of the schema, while the dimension tables are located at the edges, forming a star-like shape<sup>1</sup>.

A star schema is an example of a denormalized schema, which means that the dimension tables are not normalized and may contain redundant or repeated data. This is done to improve the performance and simplicity of queries, as there are fewer joins and tables involved. A star schema is suitable for data warehouses and business intelligence applications that require fast and efficient data retrieval<sup>2</sup>.

**NEW QUESTION 25**

An analyst is working on a project for a director. During this process, the analyst pulled the data, created summarized tables and graphs with descriptions, created a report summary, and inserted all items into a report. After writing the report, which of the following would be the most appropriate next step?

- A. Complete an audit on the data pulled for the report.
- B. Complete a check for quality in the report.
- C. Complete a review of the data and a check for consistency
- D. Complete a trend analysis to be included in the report.

**Answer:** B

**Explanation:**

After writing the report, the most appropriate next step for the analyst is to complete a check for quality in the report. This involves reviewing the report for accuracy, clarity, completeness, consistency, and relevance. The analyst should ensure that the report addresses the director's business questions and objectives, that the data and analysis are correct and reliable, that the tables and graphs are well-designed and easy to understand, that the descriptions and summary are concise and informative, and that there are no errors or inconsistencies in the report. A quality check will help the analyst to improve the presentation and communication of the report, as well as to avoid any misunderstandings or misinterpretations by the director<sup>1</sup>.

**NEW QUESTION 29**

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 34**

Which of the following techniques is used to quantify data?

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

**Answer:** C

**Explanation:**

Answer C. Coding

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

? Very satisfied = 5

? Satisfied = 4

? Neutral = 3

? Dissatisfied = 2

? Very dissatisfied = 1

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

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

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

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

#### NEW QUESTION 38

Which of the following is a non-parametric test?

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

**Answer: D**

#### Explanation:

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

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

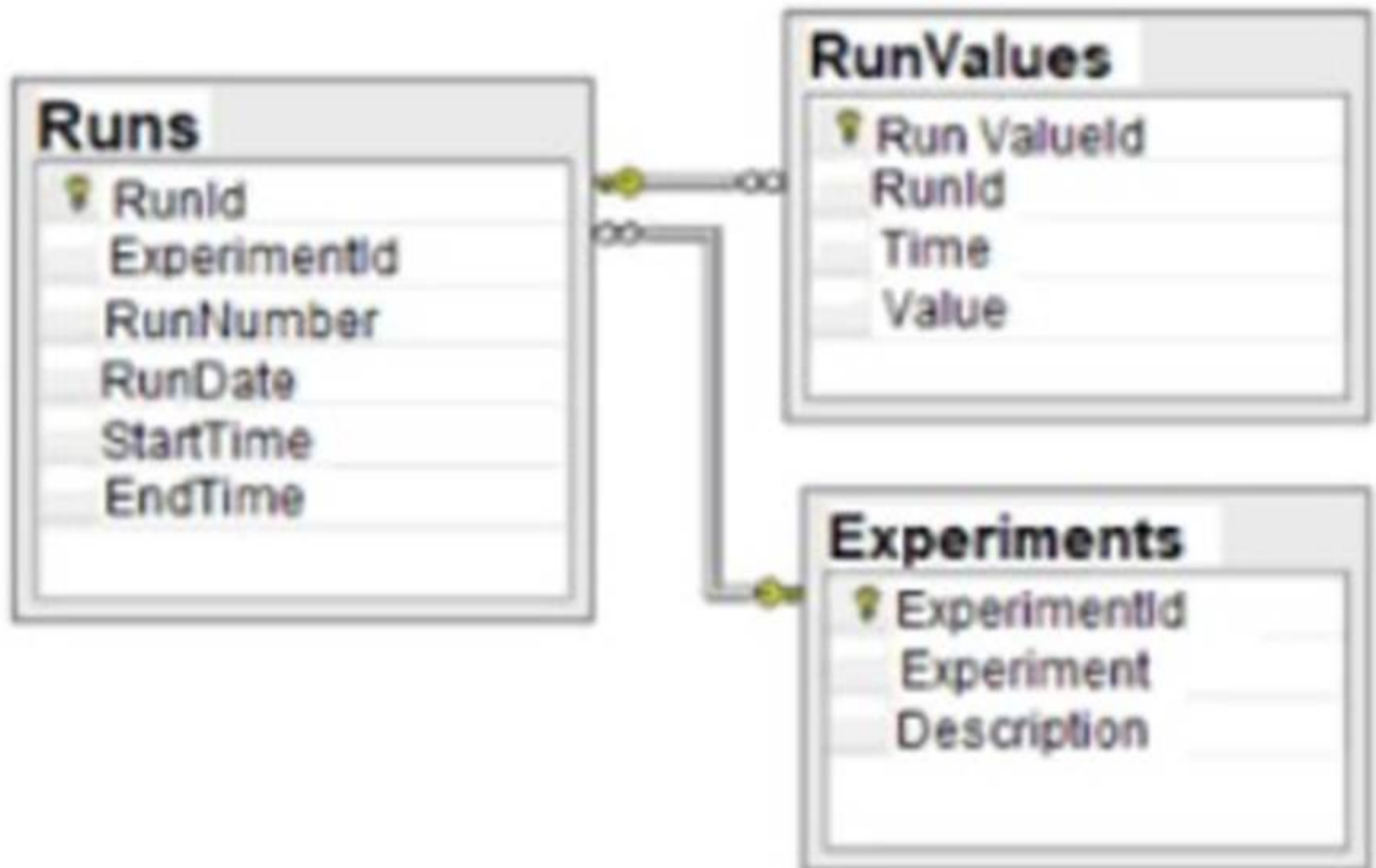
\* A. One-sample t-test is not correct, because it is a parametric test that compares the mean of a sample to a specified value. One-sample t-test assumes that the data follows a normal distribution and has a known population standard deviation.

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

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

#### NEW QUESTION 42

Given the diagram below:



Which of the following data schemas shown?

- A. Key-value pairs
- B. Online transactional processing
- C. Data Lake
- D. Relational database

**Answer: D**

#### Explanation:

A relational database is a type of database that organizes data into tables, where each table has a fixed number of columns and a variable number of rows. Each row in a table represents a record or an entity, and each column represents an attribute or a property of that entity. The tables are linked by common fields, called keys, which enable the database to establish relationships between the data. A relational database schema is a diagram that shows the structure and organization of the tables, columns, keys, and constraints in a relational database. The diagram given in the question is an example of a relational database schema, as it shows two tables: Runs and Experiments, with their respective columns, data types, and primary keys. The Runs table also has a foreign key that references the ExperimentId column in the Experiments table, indicating a relationship between the two tables. Therefore, the correct answer is D.

References: What is a database schema? | IBM, Database Schema - Javatpoint

**NEW QUESTION 45**

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 46**

Which of the following variable name formats would be problematic if used in the majority of data software programs?

- A. First\_Name\_
- B. FirstName
- C. First\_Name
- D. First Name

**Answer:** D

**Explanation:**

This is because First Name is a variable name format that would be problematic if used in most of the data software programs, such as Excel, SQL, or Python.

This is because First Name contains a space between two words, which could cause confusion or errors in the data software programs, as they might interpret the space as a separator or a delimiter between two different variables or values, rather than as part of a single variable name. For example, in SQL, a space is used to separate keywords, clauses,

or expressions in a statement, such as SELECT, FROM, WHERE, etc. Therefore, using First Name as a variable name in SQL could result in a syntax error or an unexpected result. The other variable name formats would not be problematic if used in most of the data software programs. Here is why:

? First\_Name\_ is a variable name format that uses an underscore (\_) to separate two words, which is a common and acceptable practice in most of the data software programs, as it helps to improve the readability and clarity of the variable name. For example, in Python, an underscore is used to follow the PEP 8 style guide for naming variables, which recommends using lowercase letters and underscores for multi-word variable names.

? FirstName is a variable name format that uses camel case to separate two words,

which is another common and acceptable practice in most of the data software programs, as it helps to reduce the length and complexity of the variable name. For example, in Excel, camel case is used to follow the VBA naming conventions for naming variables, which recommends using mixed case letters for multi-word variable names.

? First\_Name is a variable name format that also uses an underscore (\_) to separate

two words, which is also a common and acceptable practice in most of the data software programs, as it helps to improve the readability and clarity of the variable name. For example, in SQL, an underscore is used to follow the ANSI SQL naming standards for naming variables, which recommends using lowercase letters and underscores for multi-word variable names.

**NEW QUESTION 49**

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 51**

A data analyst is helping a retail store categorize its customers into five different groups based on the following information:

- How recently the customers made purchases
  - How frequently the customers made purchases
  - How much the customers spent
- Given the following information:

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

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

- A. CustomerJ
- B. Channel, Order\_Date
- C. CustomerJD, Territor
- D. Amount
- E. CustomerJD, Order\_Dat
- F. Amount
- G. CustomerJ
- H. Quantity, Amount

**Answer: C**

#### NEW QUESTION 55

A data analyst is compiling a report that a Chief Executive Officer needs for an impromptu meeting. The report should include information on the previous day's performance. Which of the following reports should the analyst provide?

- A. Tactical
- B. Ad hoc
- C. Dynamic
- D. Recurring

**Answer: B**

#### NEW QUESTION 59

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 61

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 66

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 67

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

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

**Answer:** C

#### Explanation:

Filter on any of the responses that do not say ??January?? and update them to ??January??. This is because filtering and updating are data cleansing techniques that can be used to ensure data consistency, which means that the data is uniform and follows a standard format. By filtering on any of the responses that do not say ??January?? and updating them to ??January??. the analyst can make sure that all the responses for the month of January are written in the same way. The other steps are not appropriate for ensuring data consistency. Here is why:

Deleting any of the responses that do not have ??January?? written out would result in data loss, which means that some information would be missing from the data set. This could affect the accuracy and reliability of the analysis.

Replacing any of the responses that have ??01?? would not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??Jan?? and ??January??. This could cause confusion and errors in the analysis. Sorting any of the responses that say ??Jan?? and updating them to ??01?? would also not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??01?? and ??January??. This could also cause confusion and errors in the analysis.

#### NEW QUESTION 72

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

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

**Answer:** CF

#### Explanation:

Two reasons to conduct data cleansing are:

? To improve accuracy: Data cleansing helps to ensure that the data is correct, consistent, and reliable. This can improve the quality and validity of the analysis, as well as the decision-making and outcomes based on the data<sup>12</sup>

? To calculate trends: Data cleansing helps to remove or resolve any errors, outliers, or missing values that could distort or skew the data. This can help to identify and measure the patterns, changes, or relationships in the data over time<sup>13</sup>

#### NEW QUESTION 75

Given the following data tables:

CustomerID	CustomerLastName
01	Manzelli
02	Kraus

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

ItemID	Customer_Last_Name	QuantityPurchased
01	Brown	03
02	Smee	07

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

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

**Answer:** A

#### Explanation:

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

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

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

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

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

#### NEW QUESTION 77

The senior management team at a company receives a detailed sales report at the end of each quarter. The report is several pages long and includes data from dozens of offices across the country. The team wants a better way to get a quick snapshot of what is included in the report. Which of the following modifications would best meet this requirement?

- A. Modifying documentation elements to include reference data sources

- B. Modifying the font size and style so important data points are more visible
- C. Modifying the report to include a summary section with observations and insights
- D. Modifying the report layout so it is easier to follow and understand

**Answer:** C

**Explanation:**

The purpose of an executive summary is to provide a concise and informative overview of a longer report, allowing busy stakeholders to quickly understand the key points and findings without reading the entire document. This summary should highlight the most important data, conclusions, and recommendations, and is typically placed at the beginning of the report for easy access<sup>12</sup>.

In the context of a detailed sales report for senior management, including a summary section with observations and insights would allow the team to quickly grasp the performance across various offices and identify any significant trends or issues that require attention. This approach aligns with best practices for executive reporting, which emphasize the importance of clear and concise summaries that focus on essential KPIs and actionable insights<sup>12</sup>.

References: 1: Databox - How to Write an Executive Summary for a Report: Step By Step Guide with Examples 2: LinkedIn - Best Practices for Writing Executive Summaries

**NEW QUESTION 78**

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

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

**Answer:** B

**NEW QUESTION 81**

Which of the following is a relational database?

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

**Answer:** A

**NEW QUESTION 83**

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 86**

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

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

**Answer:** D

**Explanation:**

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

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

**NEW QUESTION 91**

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

- A. Discrete data cannot create a sloped line.

- B. Discrete data can only be a finite number of values.
- C. Discrete data can have decimal points.
- D. Discrete data applies only to numbers.

**Answer:** B

**Explanation:**

Discrete data are data that can only assume specific values that are countable and distinct. For example, the number of books, the number of heads in a coin toss, or the number of patients in a hospital are discrete data. Discrete data cannot have fractional or decimal values, and there are clear spaces between the possible values<sup>12</sup>. Continuous data are data that can assume any value within a range and can be meaningfully divided into smaller parts. For example, the weight, height, length, time, or temperature are continuous data. Continuous data can have fractional or decimal values, and there are infinite numbers of possible values between any two points<sup>12</sup>.

**NEW QUESTION 92**

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 96**

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

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

**Answer:** B

**Explanation:**

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

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

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

big data techniques and technologies can manage and analyze the huge volume of weather data from different resources, which supports the use of public 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 98**

An analyst runs a report on a daily basis, and the number of datapoints must be validated before the data can be analyzed. The number of datapoints increases each day by approximately 20% of the total number from the day before. On a given day, the number of datapoints was 8,798. Which of the following should be the total number of datapoints on the next day?

- A. 7,038
- B. 9,600
- C. 10,600
- D. 10,800

**Answer:** C

**Explanation:**

This is because the number of datapoints increases each day by approximately 20% of the total number from the day before. Therefore, to find the number of datapoints on the next day, we can use the formula:

$$\text{Next day} = \text{Current day} * (1 + 20\%)$$

Plugging in the given values, we get:

$$\text{Next day} = 8,798 * (1 + 0.2)$$

$$\text{Next day} = 8,798 * 1.2$$

$$\text{Next day} = 10,557.6$$

Since we are dealing with whole numbers, we can round up the result to the nearest integer, which is 10,600.

#### NEW QUESTION 101

Given the below:

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

Which of the following numbers represents a Type I error?

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

**Answer:** C

#### NEW QUESTION 105

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

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

**Answer:** C

#### NEW QUESTION 106

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

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

**Answer:** A

#### Explanation:

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

#### NEW QUESTION 107

An analyst is working with the income data of suburban families in the United States. The data set has a lot of outliers, and the analyst needs to provide a measure that represents the typical income. Which of the following would BEST fulfill the analyst's goal?

- A. Median
- B. Mean
- C. Mode
- D. Standard deviation

**Answer:** A

#### Explanation:

This is because the median is a type of statistical measure that represents the typical value or central tendency of a data set, which means that it divides the data set

into two equal halves, such that half of the values are above it and half are below it. Median can be used to provide a measure that represents the typical income of suburban families in the United States, especially when the data set has a lot of outliers, which means that it has values that are unusually high or low compared to the rest of the data set. Median can provide a measure that represents the typical income of suburban families in the United States, because it is not affected or skewed by the outliers, as it only depends on the middle value or the middle two values of the data set, regardless of how extreme or distant the outliers are. For example, median can provide a measure that represents the typical income of suburban families in the United States, by finding the income value that splits the data set into two equal groups of families, such that 50% of the families have higher incomes and 50% have lower incomes. The other statistical measures are not the best measures to represent the typical income of suburban families in the United States. Here is why:

? Mean is a type of statistical measure that represents the average value or central tendency of a data set, which means that it is the sum of all the values divided by the number of values. Mean is not a good measure to represent the typical income of suburban families in the United States, especially when the data set has a lot of outliers, because it is affected or skewed by the outliers, as it takes into account all the values in the data set, regardless of how extreme or distant they are. For example, mean can provide a measure that does not represent the typical income of suburban families in the United States, by finding the income value that is influenced by a few very high or very low incomes, which could make it higher or lower than most of the incomes in the data set.

? Mode is a type of statistical measure that represents the most frequent value or mode of a data set, which means that it is the value that occurs most often in the data set. Mode is not a good measure to represent the typical income of suburban families in the United States, especially when the data set has a lot of outliers, because it is not representative or indicative of the central tendency or distribution of the data set, as it only depends on the count or occurrence of a single value or a few values in the data set, regardless of how common or rare they are. For example, mode can provide a measure that does not represent the typical income of suburban families in the United States, by finding the income value that is repeated more often than others, which could be an outlier or an anomaly in the data set.

? Standard deviation is a type of statistical measure that represents the amount of dispersion or variation of a data set, which means that it quantifies how much the values in a data set vary or deviate from the mean or average of the data set. Standard deviation is not a measure that represents the typical income of suburban families in the United States, but rather a measure that describes the spread or distribution of their incomes, as well as identifies any outliers or extreme values in their incomes. For example, standard deviation can provide a measure that describes how diverse or homogeneous their incomes are, as well as how far their incomes are from their average income.

#### NEW QUESTION 109

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 111

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

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

Which of the following charts would be BEST to use?

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

**Answer:** B

**Explanation:**

A pie chart is the best choice to show the composition between the categories of the survey response data set. A pie chart represents the whole with a circle, divided by slices into parts. Each slice shows the relative size of each category as a percentage of the total. A pie chart is useful when the categories are mutually exclusive and add up to 100%. The table shows the favorite color and the number of responses for each color, which can be easily converted into percentages. A pie chart can show how each color contributes to the total number of responses.

Option A is incorrect because a histogram is used to show how data points are distributed along a numerical scale. The survey response data set is not numerical, but categorical. Option C is incorrect because a line chart is used to show trends or changes over time. The survey response data set does not have a time dimension.

Option D is incorrect because a scatter plot is used to show the relationship between two numerical variables. The survey response data set does not have two numerical variables. Option E is incorrect because a waterfall chart is used to show how an initial value is increased or decreased by a series of intermediate values. The survey response data set does not have an initial value or intermediate values.

References:

- ? How to Choose the Right Chart for Your Data - Infogram
- ? How to Choose the Right Data Visualization | Tutorial by Chartio
- ? Find the Best Visualizations for Your Metrics - The Data School
- ? How to choose the best chart or graph for your data

**NEW QUESTION 112**

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 115**

A data analyst needs to write a SOL query measuring last month's website visits and distribute a summary report to the marketing team. Which of the following is the analyst creating?

- A. Date range
- B. Distribution list
- C. Data content

D. Report view

**Answer:** D

#### NEW QUESTION 117

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

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

**Answer:** D

#### Explanation:

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

References:

? Guidelines for Removing and Handling Outliers in Data<sup>1</sup>.

? Mean, Median, and Mode: Measures of Central Tendency<sup>2</sup>.

? Which measure of central tendency should be used when there is an outlier?<sup>3</sup>.

? How are measures of central tendency affected by outliers?<sup>4</sup>.

#### NEW QUESTION 120

The duration of a phone call in milliseconds is an example of:

- A. ordinal data.
- B. nominal data.
- C. boolean data.
- D. continuous data.

**Answer:** D

#### Explanation:

The correct answer is D. Continuous data.

Continuous data is a type of quantitative data that can take any value within a range and can be measured with infinite precision. Continuous data can be expressed as fractions, decimals, or percentages. Examples of continuous data are height, weight, temperature, time, speed, etc<sup>12</sup>

The duration of a phone call in milliseconds is an example of continuous data, because it can take any value within a range (from zero to infinity) and can be measured with infinite precision (up to milliseconds or even smaller units). The duration of a phone call in milliseconds can also be expressed as fractions, decimals, or percentages of a larger unit (such as seconds, minutes, or hours).

Ordinal data is not correct, because ordinal data is a type of qualitative or categorical data that can be ordered or ranked according to some criterion. Ordinal data can have a logical order, but the intervals between the values are not equal or meaningful. Examples of ordinal data are grades, ratings, ranks, etc<sup>12</sup>

Nominal data is not correct, because nominal data is a type of qualitative or categorical data that can be labeled or named without any order or ranking. Nominal data can have a finite number of categories or classes, but the categories have no intrinsic value or hierarchy. Examples of nominal data are gender, color, nationality, etc<sup>12</sup>

Boolean data is not correct, because boolean data is a type of binary data that can have only two possible values: true or false. Boolean data can be used to represent logical statements, conditions, or outcomes. Examples of boolean data are yes/no, on/off, 1/0, etc.

#### NEW QUESTION 122

Five dogs have the following heights in millimeters: 300,430, 170, 470, 600

Which of the following is the standard deviation for the five dogs?

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

**Answer:** B

#### Explanation:

The correct answer is B. 154 mm.

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

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

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

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

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

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

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

#### NEW QUESTION 123

An analyst is preparing a report that contains weather data. The temperatures are shown in Fahrenheit. but they must be reported in Celsius. Which of the following should the analyst do to fix this issue?

- A. Normalize the data.

- B. Standardize the data.
- C. Rescale the data.
- D. Aggregate the data.

**Answer: C**

**Explanation:**

The analyst should rescale the data to fix this issue. Rescaling is a process of transforming data from one scale to another, such as changing the units of measurement. In this case, the analyst needs to rescale the temperatures from Fahrenheit to Celsius, which are two different scales for measuring temperature. To do this, the analyst can use the following formula:

$\text{Celsius} = (\text{Fahrenheit} - 32) \times \frac{5}{9}$

This formula converts each temperature value from Fahrenheit to Celsius by subtracting 32

and multiplying by 5/9. For example, if the temperature is 68°F, the rescaled value in Celsius is:

$\text{Celsius} = (68 - 32) \times \frac{5}{9}$  Celsius = 20°C

Rescaling the data can help the analyst to report the temperatures in a consistent and accurate way, and to avoid any confusion or errors that may arise from using different scales. Rescaling can also make the data more comparable and compatible with other data sources or standards that use the same scale<sup>12</sup>.

**NEW QUESTION 124**

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 127**

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 132**

Given the table below:

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

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

- A. Name, one

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

**Answer:** B

**Explanation:**

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

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

References:

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

**NEW QUESTION 135**

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

- A. Mean
- B. Variance
- C. Correlation
- D. Confidence interval

**Answer:** B

**Explanation:**

The measure of dispersion is used to describe the spread of data around a central value. In the context of a data set containing actual and projected sales, the measure of dispersion will help to understand the variability or consistency of sales figures. The variance is the most appropriate statistical method for finding the measure of dispersion because it calculates the average of the squared differences from the Mean, providing a clear picture of data spread. It is especially useful in comparing the spread between different data sets and understanding the distribution of data points.

? Mean is a measure of central tendency, not dispersion.

? Correlation measures the relationship between two variables, not the spread of a single variable.

? Confidence intervals are used to estimate the range within which a population parameter will fall, but they do not measure dispersion within the data set itself.

References:

- ? Measures of Dispersion in Statistics<sup>1</sup>
- ? Measures of Dispersion - Definition, Formulas, Examples<sup>2</sup>
- ? Statistical dispersion - Wikipedia<sup>3</sup>

**NEW QUESTION 136**

A data analyst is attempting to understand how ice cream consumption is affected by different attributes. such as cost, temperature. and income level. Which of the following regression analyses should the data analyst perform to understand this relationship?

- A. Logistic
- B. Ordinary least squares
- C. Cox
- D. Polynomial

**Answer:** B

**Explanation:**

Answer: B. Ordinary least squares

Ordinary least squares (OLS) is a type of linear regression that is used to fit a regression model that describes the relationship between one or more predictor variables and a numeric response variable. Use when: The relationship between the predictor variable(s) and the response variable is reasonably linear. The response variable is a continuous numeric variable<sup>1</sup>.

In this case, the data analyst is interested in understanding how ice cream consumption (the response variable) is affected by different attributes, such as cost, temperature, and income level (the predictor variables). Assuming that these variables have a linear relationship, OLS can be used to estimate the coefficients of the regression equation that best fits the data. OLS can also provide measures of goodness-of-fit, such as R-squared and adjusted R-squared, and test the significance of the coefficients using t-tests and F- tests<sup>2</sup>.

Option A is incorrect, as logistic regression is used to fit a regression model that describes the relationship between one or more predictor variables and a binary response variable. Use when: The response variable is binary – it can only take on two values<sup>1</sup>. Ice cream consumption is not a binary variable, but rather a continuous numeric variable.

Option C is incorrect, as Cox regression is used to fit a regression model that describes the relationship between one or more predictor variables and a survival time response variable. Use when: The response variable is the time until an event of interest occurs, such as death, failure, or recovery<sup>3</sup>. Ice cream consumption is not a survival time variable, but rather a continuous numeric variable.

Option D is incorrect, as polynomial regression is used to fit a regression model that describes the relationship between one or more predictor variables and a numeric response variable. Use when: The relationship between the predictor variable(s) and the response variable is non-linear<sup>1</sup>. If there is no evidence of non-linearity in the data, polynomial regression may not be appropriate, as it may overfit the data and produce unreliable estimates.

**NEW QUESTION 139**

You are working with a dataset and need to swap the values in rows with those in columns. What action do you need to perform?

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

**Answer:** D

**Explanation:**

Transpose creates a new data file in which the rows and columns in the original data file are transposed so that cases (rows) become variables and variables (columns) become cases. Transpose automatically creates new variable names and displays a list of the new variable names. Transposing data is useful for data analysis. At times, we have to pull data from various files with different formats for analysis and preparing reports. In such circumstances, we may have to transpose some data from one file to the other. In excel, we can transpose data in multiple ways.

**NEW QUESTION 143**

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 148**

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 150**

Which of the following best describes an exploratory analysis?

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

**Answer:** A

**Explanation:**

Answer A. Involves the use of descriptive statistics to understand observations. Exploratory data analysis (EDA) is a method of analyzing and investigating data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. EDA involves the use of descriptive statistics, such as mean, median, mode, standard deviation, frequency, or percentage, to understand the distribution, central tendency, variability, and relationship of the data. EDA helps to see what the data can reveal beyond the formal modeling or hypothesis testing, and provides a better understanding of data set variables and the interactions between them1.

**NEW QUESTION 151**

Samantha needs to share a list of her organization's top 50 customers with the VP of sales.

She would like to include the name of the customer, the business they represent, their contact information, and their total sales over the past year.

The VP does not have any specialized analytics skills or software but would like to make some personal notes on the dataset.

What would be the best tool for Samantha to use to share this information?

- A. Power BI.
- B. Microsoft Excel.
- C. Minitab.
- D. SAS.

**Answer:** B

**Explanation:**

Microsoft Excel.

This scenario presents a very simple use case where the business leader needs a dataset in an easy-to-access form and will not be performing any detailed analysis.

A simple spreadsheet, such as Microsoft Excel, would be the best tool for this job. There is no need to use a statistical analysis package, such as SAS or Minitab, as this would likely confuse the VP without adding any value. The same is true of an integrated analytics suite, such as Power BI.

**NEW QUESTION 156**

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

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

**Answer:** A

**Explanation:**

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

**NEW QUESTION 161**

A sales manager wants quarterly sales reports broken down by unit and week. Which of the following data output lists includes the most necessary information?

- A. Order number
- B. salesperson
- C. date shipped, recipient address, and price
- D. Item name, salesperson
- E. recipient address, shipping cost
- F. and date shipped
- G. Item number, item name, salesperson
- H. date sold
- I. and price
- J. Item name
- K. salesperson
- L. price
- M. shipping cost
- N. and date shipped

**Answer:** C

**Explanation:**

To create a quarterly sales report broken down by unit and week, the most necessary information is the item number, item name, salesperson, date sold, and price. These data elements can help the sales manager to track the sales volume, revenue, and performance of each unit and each week within a quarter. The item number and item name can identify the products or services sold by each unit. The salesperson can indicate the individual or team responsible for each sale. The date sold can show when each sale occurred and how it relates to the weekly and quarterly goals. The price can show how much revenue each sale generated and how it contributes to the unit and quarterly totals.

**NEW QUESTION 162**

Which of the following is a control measure for preventing a data breach?

- A. Data transmission
- B. Data attribution
- C. Data retention
- D. Data encryption

**Answer:** D

**Explanation:**

This is because data encryption is a type of control measure that prevents a data breach, which is an unauthorized or illegal access or use of data by an external or internal party. Data encryption can prevent a data breach by protecting and securing the data using a code or a key that scrambles or transforms the data into an unreadable or incomprehensible format, which can only be decoded or restored by authorized users who have the correct code or key. For example, data encryption can prevent a data breach by encrypting the data in transit or at rest, such as when the data is sent over a network or stored in a device. The other control measures are not used for preventing a data breach. Here is why:

? Data transmission is a type of process that transfers and exchanges data between different sources or systems, such as databases, cloud services, or web applications. Data transmission does not prevent a data breach, but rather exposes the data to potential risks or threats during the transfer or exchange. However, data transmission can be made more secure and less vulnerable to a data breach by using encryption or other methods, such as authentication or authorization.

? Data attribution is a type of feature or function that assigns and tracks the ownership and origin of the data, such as the creator, modifier, or source of the data. Data attribution does not prevent a data breach but rather provides information and evidence about the data provenance and history. However, data attribution can be useful for detecting and responding to a data breach by using audit logs or metadata to identify and trace any unauthorized or illegal access or use of the data.

? Data retention is a type of policy or standard that specifies and regulates the storage and preservation of the data, such as the duration, location, or format of the data. Data retention does not prevent a data breach, but rather affects the availability and accessibility of the data for future use or reference. However, data retention can be optimized and aligned with the legal and ethical requirements and standards of the industry or the organization to reduce the risk or impact of a data breach.

#### NEW QUESTION 163

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

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

**Answer: B**

#### Explanation:

A password-protected dashboard is a type of web-based application that can display the health scorecards of the troops in a secure and interactive way. A password-protected dashboard can provide the following benefits for the commander:

- ? It can protect the PHI data from unauthorized access or disclosure by requiring a valid username and password to log in. This can ensure that only the commander and other authorized personnel can view the information<sup>12</sup>
- ? It can allow the commander to filter the data based on gender and rank by using drop-down menus, sliders, checkboxes, or other controls. This can enable the commander to customize the view and focus on the relevant data<sup>13</sup>
- ? It can update the data daily by connecting to a data source that refreshes automatically or on demand. This can ensure that the commander always sees the latest and most accurate information<sup>14</sup>
- ? It can present the data in a visual and intuitive way by using charts, graphs, tables, or other elements. This can help the commander to understand and analyze the data more easily and effectively<sup>1</sup>

#### NEW QUESTION 167

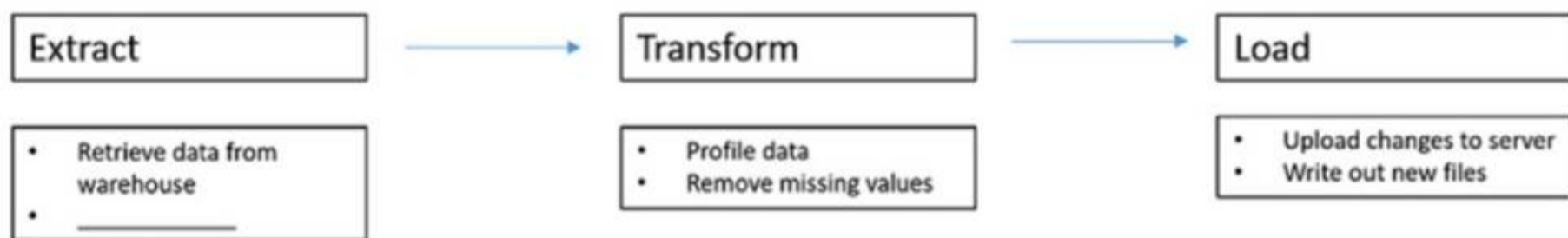
A database administrator needs to ensure only approved users can access specific database tables to perform financial functions. Which of the following is the best access control method for the administrator to use?

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

**Answer: A**

#### NEW QUESTION 172

Given the diagram below:



Which of the following steps is missing?

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

**Answer: A**

#### Explanation:

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

#### NEW QUESTION 177

Which of the following statistical methods requires two or more categorical variables?

- A. Simple linear regression
- B. Chi-squared test
- C. Z-test
- D. Two-sample t-test

**Answer: B**

#### Explanation:

This is because a chi-squared test is a type of statistical method that tests the association or independence between two or more categorical variables, such as gender, race, or occupation. A chi-squared test can be used to compare the observed frequencies of the categories with the expected frequencies under the null hypothesis of no association or independence. For example, a chi-squared test can be used to determine if there is a relationship between smoking and lung cancer. The other statistical methods do not require two or more categorical variables. Here is why:

Simple linear regression is a type of statistical method that models the relationship between a continuous dependent variable and a continuous or categorical independent variable, such as height, weight, or education level. A simple linear regression can be used to estimate the slope and intercept of the best-fitting line that describes how the dependent variable changes with the independent variable. For example, a simple linear regression can be used to predict the weight of a person based on their height.

Z-test is a type of statistical method that tests the significance of the difference between a sample mean and a population mean, or between two sample means,

when the population standard deviation or the sample sizes are large enough. A z-test can be used to compare the average scores of two groups of students on a standardized test.

Two-sample t-test is a type of statistical method that tests the significance of the difference between two sample means when the population standard deviation is unknown or the sample sizes are small. A two-sample t-test can be used to compare the average salaries of two groups of employees in different departments.

**NEW QUESTION 180**

Which of the following database schemas features normalized dimension tables?

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

**Answer:** B

**Explanation:**

The correct answer is B. Snowflake.

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

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

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

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

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

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

Some disadvantages are:

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

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

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

**NEW QUESTION 185**

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 186**

Given the data below:

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

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

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

**Answer: B**

**Explanation:**

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

**NEW QUESTION 189**

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

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

**Answer: A**

**Explanation:**

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

**NEW QUESTION 194**

Joe, an analyst, tests the loading time on a dashboard he is preparing to go live and finds it is slower than he would like. Which of the following must occur to decrease the loading time?

- A. Deploy the dashboard to production.
- B. Change the field definitions.
- C. Update the dashboard subscribers.
- D. Optimize the dashboard.

**Answer: D**

**Explanation:**

Optimizing the dashboard is the process of improving its performance and reducing its loading time by applying various techniques and best practices. Some of the common ways to optimize a dashboard are:

- ? Reducing the size and complexity of the data model, such as removing unnecessary columns, aggregating data at the source, or using data compression techniques<sup>12</sup>
- ? Leveraging caching strategies, such as setting appropriate cache refresh intervals or utilizing Power BI's built-in caching mechanisms, to minimize data retrieval delays<sup>2</sup>
- ? Utilizing query folding, direct query, or live connection to enhance data processing efficiency and enable real-time data updates<sup>23</sup>
- ? Optimizing DAX queries, such as avoiding nested calculations, using variables, or simplifying measures, to improve data calculation speed<sup>23</sup>
- ? Reducing visualizations and calculations, such as using fewer or simpler charts, filters, or parameters, to speed up dashboard rendering<sup>12</sup>
- ? Evaluating the impact of custom visuals on dashboard load time and avoiding or replacing those that are slow or inefficient<sup>2</sup>
- ? Applying aggregation and summarization techniques, such as using extract filters, context filters, or level of detail expressions, to reduce the amount of data displayed on the dashboard<sup>1</sup>
- ? Troubleshooting and resolving any issues that may cause slow dashboard load, such as network latency, server overload, or hardware limitations<sup>24</sup>

**NEW QUESTION 195**

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

Table 1

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

Table 2

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

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

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

**Answer:** B

**Explanation:**

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

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

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

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

The result of this query would be:

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

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

**NEW QUESTION 199**

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

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

**Answer:** DE

**Explanation:**

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

### NEW QUESTION 201

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

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

**Answer:** A

#### Explanation:

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

### NEW QUESTION 202

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

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

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

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

**Answer:** B

#### Explanation:

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

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

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

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

### NEW QUESTION 207

Which of the following is a best practice when updating a legacy data source?

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

**Answer:** C

#### Explanation:

When updating a legacy data source, it is a best practice to create a codebook to document field changes. A codebook serves as a detailed guide and record of the data structure, definitions, and any transformations or modifications made to the data fields. This documentation is crucial for maintaining data integrity, ensuring consistency, and facilitating future data use and understanding. It provides a reference that can be invaluable for data analysts, developers, and any stakeholders who need to work with the data.

Creating a codebook is preferred over placing old data in new fields, which can lead to confusion and data integrity issues. Keeping only the most recent data may result in the loss of valuable historical information. Removing the data source from production is not a practice related to updating data but rather to retiring a data

source1234.

References:

- ? Legacy Data Migration: A Comprehensive Guide | OpenGeeksLab
- ? How to Successfully Complete Legacy Database Migration
- ? Methods for Saving and Integrating Legacy Data - DATAVERSITY
- ? Legacy Data Digitization - Learn The Best Practices

#### NEW QUESTION 210

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 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 213

Which of the following is the best technique for transferring data from one database to another with some data manipulation?

- A. Application programming interfaces
- B. Delta load
- C. Extract, transform, load
- D. Export/import

**Answer:** C

#### NEW QUESTION 217

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 218

Which of the following is the best description of the term "data governance"?

- A. Data governance governs the development of a data visualization dashboard in an organization.
- B. Data governance is the policy that protects against data breaches by cybercriminals.
- C. Data governance is the process of analyzing, manipulating, and reporting data in an organization.
- D. Data governance is the availability, usability, integrity, and security of data in an enterprise.

**Answer:** D

#### Explanation:

Data governance refers to the overarching management of data??s availability, usability, integrity, and security within an organization. It involves setting policies and standards that govern data usage, determining data ownership, implementing data security measures, and ensuring that data is accessible for business insights while maintaining its quality. The goal of data governance is to ensure that data is consistent, trustworthy, and not misused, supporting compliance with data privacy regulations and enabling effective data analytics to optimize operations and drive business decision-making.

References:

- ? Understanding Data Governance and Its Importance1.
- ? The Role of Data Governance in Data Management2.
- ? Defining Data Governance and Its Business Value3.

#### NEW QUESTION 220

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

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

**Answer:** A

#### Explanation:

The correct answer is .html.

Short Explanation: Web scraping is the process of extracting data from websites by parsing the HTML code of the web pages. HTML stands for HyperText Markup Language and it is the standard markup language for creating web pages and web applications. HTML files have the extension .html and they contain tags, elements, attributes, and content that define the structure and appearance of a web page. Web scraping tools need to parse the HTML files to extract the relevant data from the web pages12

**NEW QUESTION 225**

A customer survey reveals 90% positive feedback. Which of the following statistical methods would be best to utilize to determine the reliability of a data set and predict how a larger sample of customers over the same time period might respond?

- A. Calculate a high variance on survey responses.
- B. Calculate the maximum range of the survey responses.
- C. Calculate a low standard deviation on survey responses.
- D. Remove any data more than 4 standard deviation from the mean.

**Answer:** C

**Explanation:**

A low standard deviation in survey responses indicates that the data points tend to be close to the mean, suggesting a high level of consistency among the responses. This consistency is crucial for determining the reliability of the data set and predicting future outcomes. If the standard deviation is low, it means that the positive feedback is not only high but also consistent, making it a reliable indicator of customer satisfaction and a good predictor of how a larger sample might respond.

References: The concept of using standard deviation to assess data reliability is a standard practice in statistics and data analysis123.

**NEW QUESTION 228**

An analyst needs to create an analytics dashboard for an employee intranet site to improve the search functionality, display relevant information, and maintain an updated FAQ page. Which of the following visualizations would best represent what employees are searching for?

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

**Answer:** A

**Explanation:**

A word cloud is an ideal choice for visualizing what employees are searching for on an intranet site. It represents the frequency of word occurrence in a visually impactful way, with more commonly searched terms appearing larger in the cloud. This allows for quick identification of the most popular queries and topics of interest among employees. Unlike histograms, pie charts, or scatter plots, word clouds can effectively display textual data, which is the nature of search queries. They are particularly useful for analyzing text data from surveys or feedback forms, which can be similar to search query data in an intranet environment1234.

References: 1: ??What Are Word Clouds? Pros & Cons of Word Cloud Visualizations?? - Alida 2: ??Using Word Clouds for Powerful Data Visualization?? - WordCloud.app blog 3: ??Ultimate Google Data Studio Word Cloud Guide: Visualization 2024?? - AtOnce 4: ??How to Create Word Cloud in Power BI?? - Zebra BI

**NEW QUESTION 231**

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

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

**Answer:** C

**Explanation:**

The most likely cause of the issue is that the databases are recording the event in different time zones. For example, if one database is in New York and the other database is in Los Angeles, there is a three-hour difference between them. Therefore, an event that occurs at 12:00 PM in New York would be recorded as 9:00 AM in Los Angeles. To avoid this issue, the databases should either use a common time zone or convert the timestamps to a standard format. Therefore, option C is correct.

Option A is incorrect because the data analyst is not querying the databases incorrectly, but rather observing a discrepancy in the timestamps.

Option B is incorrect because the databases are recording the same event, but with different timestamps.

Option D is incorrect because the second database is not logging incorrectly, but rather using a different time zone.

**NEW QUESTION 232**

Which of the following is an example of a flat file?

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

**Answer:** D

**NEW QUESTION 237**

Consider the following dataset which contains information about houses that are for sale:

```
sonery=# select * from melb limit 5;
 houseid | address | regionname | type | rooms | date | distance | price
-----+-----+-----+-----+-----+-----+-----+-----
      1 | 85 Turner St | Northern Metropolitan | h | 2 | 2016-03-12 | 2.5 | 1.48e+06
      2 | 25 Bloomberg St | Northern Metropolitan | h | 2 | 2016-04-02 | 2.5 | 1.035e+06
      3 | 5 Charles St | Northern Metropolitan | h | 3 | 2017-04-03 | 2.5 | 1.465e+06
      4 | 40 Federation La | Northern Metropolitan | h | 3 | 2017-04-03 | 2.5 | 850000
      5 | 55a Park St | Northern Metropolitan | h | 4 | 2016-04-06 | 2.5 | 1.6e+06
(5 rows)
```

Which of the following string manipulation commands will combine the address and region name columns to create a full address?

full\_address----- 85 Turner St, Northern Metropolitan 25 Bloomberg St, Northern Metropolitan 5 Charles St, Northern Metropolitan 40 Federation La, Northern Metropolitan 55a Park St, Northern Metropolitan

- A. SELECT CONCAT(address, ' ', regionname) AS full\_address FROM melb LIMIT 5;
- B. SELECT CONCAT(address, '-', regionname) AS full\_address FROM melb LIMIT 5;
- C. SELECT CONCAT(regionname, ' ', address) AS full\_address FROM melb LIMIT 5
- D. SELECT CONCAT(regionname, '-', address) AS full\_address FROM melb LIMIT 5;

**Answer:** A

**Explanation:**

The correct answer is A: SELECT CONCAT(address, ' ', regionname) AS full\_address FROM melb LIMIT 5; String manipulation (or string handling) is the process of changing, parsing, splicing, pasting, or analyzing strings. SQL is used for managing data in a relational database. The CONCAT () function adds two or more strings together. Syntax CONCAT(string1, string2,... string\_n) Parameter Values Parameter Description string1, string2, string\_n Required. The strings to add together.

#### NEW QUESTION 242

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 246

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 251

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

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

Which of the following charts would be BEST to use?

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

**Answer:** B

**Explanation:**

The best chart to use to identify the composition between the categories of the survey response data set is a pie chart. A pie chart is a circular chart that shows the relative proportions of different categories in a whole. A pie chart is divided into slices that represent the percentage or frequency of each category. A pie chart is suitable for displaying categorical data that has a few categories and does not have any hierarchical or temporal relationship. In this case, a pie chart can show the composition of the favorite colors among the survey respondents, as well as the percentage of each color. The other options are not as good as a pie chart for this purpose, as they are more suitable for displaying numerical data that has some kind of distribution, trend, correlation, or comparison. A histogram is a bar chart that shows the frequency distribution of a single numerical variable. A line chart is a chart that shows the change of one or more numerical variables over time or another continuous variable. A scatter plot is a chart that shows the relationship between two numerical variables by plotting them as points on a Cartesian plane. A waterfall chart is a chart that shows how an initial value is increased or decreased by a series of intermediate values, resulting in a final value. Reference: [Choosing the Right Chart Type - DataCamp]

**NEW QUESTION 255**

An analyst has been asked to validate data quality. Which of the following are the BEST reasons to validate data for quality control purposes? (Choose two.)

- A. Retention
- B. Integrity
- C. Transmission
- D. Consistency
- E. Encryption
- F. Deletion

**Answer:** B

**Explanation:**

Integrity and D. Consistency. This is because integrity and consistency are two of the best reasons to validate data for quality control purposes, which means to check and ensure that the data is accurate, complete, reliable, and usable for the intended analysis or purpose. By validating data for integrity and consistency, the analyst can prevent or correct any errors or issues in the data that could affect the validity or reliability of the analysis or the results. Here is what integrity and consistency mean in terms of data quality:

? Integrity refers to the completeness and validity of the data, which means that the data has no missing, incomplete, or invalid values that could compromise its meaning or usefulness. For example, validating data for integrity could involve checking for null values, outliers, or incorrect data types in the data set.

? Consistency refers to the uniformity and standardization of the data, which means that the data follows a common format, structure, or rule across different sources or systems. For example, validating data for consistency could involve checking for spelling, punctuation, or capitalization errors in the data set.

The other reasons are not the best reasons to validate data for quality control purposes. Here is why:

? Retention refers to the storage and preservation of the data, which means that the data is kept and maintained in a secure and accessible way for future use or reference. Retention does not need to be validated for quality control purposes, because it does not affect the accuracy or reliability of the data itself.

? Transmission refers to the transfer and exchange of the data, which means that the data is moved or shared between different sources or systems in a fast and efficient way. Transmission does not need to be validated for quality control purposes, because it does not affect the completeness or validity of the data itself.

? Encryption refers to the protection and security of the data, which means that the data is encoded or scrambled in a way that prevents unauthorized access or use. Encryption does not need to be validated for quality control purposes, because it does not affect the uniformity or standardization of the data itself.

? Deletion refers to the removal and disposal of the data, which means that the data is erased or destroyed in a way that prevents recovery or retrieval. Deletion does not need to be validated for quality control purposes, because it does not affect the meaning or usefulness of the data itself.

**NEW QUESTION 257**

Given the following:

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

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

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

**Answer:** D

**Explanation:**

Correcting the dates so they have the same format is the most important thing for an analyst to do when transforming the table for a trend analysis. Trend analysis is a method of analyzing data over time to identify patterns, changes, or relationships. To perform a trend analysis, the data needs to have a consistent and comparable format, especially for the date or time variables.

In the example, the date purchased column has two different formats: YYYY-MM-DD and MM/DD/YYYY. This could cause errors or confusion when sorting, filtering, or plotting the data over time. Therefore, the analyst should correct the dates so they have the same format, such as YYYY-MM-DD, which is a standard and unambiguous format.

**NEW QUESTION 258**

.....

## 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>