# Comptia

(DA0-001)

Data+



Total: **219 Questions**

Link:

**Question: 1**

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

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

Which of the following is the mean?

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

**Answer: A**

**Explanation:**

The mean can be calculated by adding up all the Q1 sales values and then dividing by the number of items in the data set. In this case, the total of all the Q1 sales is $12,330.88, and there are 5 items in the data set.So, the mean is calculated as follows:$12,330.88 ÷ 5 = $2,466.18Therefore, the mean is $2,466.18.Answer: A.
$2,466.18

**Question: 2**

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

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

**Answer: D**

**Explanation:**

The most efficient format for delivering a report providing information about various regions, products, and time periods is **D. A dashboard with filters at the top that the user can toggle.**

Here's why:

A dashboard with filters offers interactivity. Unlike a static report (C) which presents a pre-defined view, a dashboard allows users to dynamically explore the data according to their specific needs. They can instantly focus on a particular region, product, or time period by toggling the filters. This eliminates the need for creating and distributing numerous static reports or workbook tabs (A) each representing a different filtered view, significantly saving time and resources.

A daily email with snapshots (B) is less efficient because it presents a limited, pre-selected subset of the data. Users may need information beyond what's included in the snapshot, requiring them to request additional reports or perform their own analysis. This approach also lacks the real-time aspect of a dashboard.

Dashboards leverage data visualization principles, presenting complex information in a concise and easily digestible manner. Users can quickly identify trends, patterns, and outliers by interacting with charts and graphs, which leads to faster and better-informed decision-making. They also promote self-service analytics, empowering users to answer their own questions without relying on analysts to create custom reports. This improves overall organizational efficiency and reduces the burden on the analytics team.

Furthermore, dashboard platforms often support features like data drill-down, enabling users to explore underlying data points for deeper insights. This level of granularity is often unavailable in static reports or summary emails. Modern dashboard tools are designed for scalability and can handle large datasets from various sources.

In summary, a dashboard is superior because of its interactivity, efficiency in delivering multiple views, support for data visualization, promotion of self-service analytics, and potential for scalability. It's the most effective way to empower users to explore the data and gain valuable insights.

For further reading on dashboards and data visualization best practices, consider exploring these resources:

Tableau's guide to dashboard design: https://www.tableau.com/learn/articles/dashboard-design-best-practices
Perceptual Edge - Information Dashboard Design: https://www.perceptualedge.com/example1.php (although somewhat older, it provides solid foundation principles).

**Question: 3**

A customer list from a financial services company is shown below:

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

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

A.Recode the variables.

B.Calculate the percentiles of the variables.

C.Calculate the standard deviations of the variables.

D.Normalize the variables.

**Answer: D**

**Explanation:**

D. Normalize the variables. Normalization involves transforming the values of the variables so that they have the same scale and range, typically from 0 to 1. This allows the analyst to ensure that each variable has the same weight in the calculation, since all the values are on the same scale. Normalizing the variables will also make it easier to interpret the results and compare the scores of different customers. This is important in

creating a likely-to-buy score because it ensures that all three variables have an equal impact on the final score, rather than one variable having a larger weight due to its scale or range.

**Question: 4**

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

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

**Answer: DE**

**Explanation:**

Here's a detailed justification for why data encryption and data masking are the most effective actions to mitigate data leaks during transmission:

**Justification:**

Data leaks occur when sensitive information is exposed to unauthorized individuals. Encryption and data masking directly address this risk during the transmission phase of data handling.

**Data Encryption (D):** Encryption transforms data into an unreadable format (ciphertext) using an algorithm and a key. Even if an attacker intercepts the transmitted data, they cannot understand it without the correct decryption key. End-to-end encryption (E2EE) provides the strongest level of protection, as data is encrypted on the sender's end and can only be decrypted on the intended recipient's end. This mitigates the risk of data interception during transmission. Encryption employs robust cryptographic algorithms (e.g., AES, RSA) to secure data.

**Data Masking (E):** Data masking, also known as data obfuscation, involves replacing sensitive data elements with modified versions while preserving their format. This can be achieved through techniques like substitution, shuffling, or tokenization. While the data is still usable for testing or reporting (if designed that way), the original sensitive information is concealed during transmission. It's useful in situations where data is being transmitted to environments where it's processed, but the actual sensitive values aren't required (e.g., sending a masked customer ID to an analytics platform, where tracking general usage patterns is required but the specific IDs are irrelevant).

**Why the other options are less suitable in the context of data transmission:**

**Data Identification (A):** Data identification is important for classifying data sensitivity levels, but it doesn't directly prevent a data leak during transmission. It's a prerequisite for deciding what data needs encryption or masking.

**Data Processing (B):** Data processing involves transforming data and doesn't inherently mitigate data leaks during transmission. Data may still be vulnerable during the transfer process, regardless of how it's being processed.

**Data Reporting (C):** Data reporting involves generating summaries or insights from data. It has no direct effect on mitigating data leaks during the data transmission phase. If not handled with care, sending

unencrypted or unmasked reports over the internet would actually constitute a data leak.

**Data Removal (F):** While data removal, or more specifically, data retention policies, are important for data security, removing the data is not an option if the goal is to transmit the data. It doesn't solve the problem of secure data transmission.

**In summary:** Encryption and data masking are specifically designed to protect data during transmission, rendering it unintelligible or unusable to unauthorized parties even if it's intercepted. This makes them the most effective choices for mitigating data leaks in the context of the question.

**Authoritative Links for Further Research:**

**NIST Guidelines on Cryptographic Key Management:**https://csrc.nist.gov/publications/detail/sp/800-57/part-1/rev-5/final (For encryption standards and key management)
**NIST Special Publication 800-113: Guide to Database Security:**
https://csrc.nist.gov/publications/detail/sp/800-113/rev-1/final (Provides guidance on data masking techniques.)
**OWASP (Open Web Application Security Project):**https://owasp.org/ (Provides resources on application security, including data protection and encryption.)

## Question: 5

A data analyst has been asked to organize the table below in the following ways:

By sales from high to low -

By state in alphabetic order -

| First_name | Last_name | Address | City | State | Sales |
|---|---|---|---|---|---|
| Ed | Edens | 2851 N. Southport | Chicago | IL | $125,689 |
| Pat | Mudd | 710 Bridle Ridge Road | Eagan | MN | $101,259 |
| Katie | Hofstad | 2851 S. Windwood Lane | Rosemount | NY | $105,779 |
| Edward | Frank | 281 S. Northport | Chicago | IL | $456,231 |
| Rachel | Newman | 505 Big Timber Trail | Wheaton | CO | $99,876 |
| Karolyn | Keath | 332 Richfield Drive | Lakeview | MN | $166,874 |

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

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

**Answer: D**

**Explanation:**

D. Sorting is the function that will allow the data analyst to organize the table in the desired manner. Sorting allows the data analyst to arrange the data in a specific order, either in ascending or descending order, based on one or multiple columns. In this case, the analyst can sort the table by Age from high to low by clicking on the Age column and selecting "Sort Descending". They can then sort the table by Income in alphabetic order by clicking on the Income column and selecting "Sort A to Z". Sorting is a simple and efficient way to organize large data sets, making it easier for the analyst to analyze and present the data.

## Question: 6

Which of the following BEST describes the issue in which character values are mixed with integer values in a data set column?

A.Duplicate data
B.Missing data
C.Data outliers
D.Invalid data type

**Answer: D**

**Explanation:**

The correct answer is **D. Invalid data type**.

Here's why:

The scenario describes a situation where a column intended to hold only integers (numeric values) contains character values (strings or letters). This violates the expected data type constraint for that column. Data types are fundamental in data management and define the kind of values a field can store (e.g., integer, string, date, boolean). When a column is designated as an integer type but contains character values, a data type mismatch occurs, rendering the data invalid for calculations or comparisons expecting purely numeric input.

**A. Duplicate data:** Duplicate data refers to the presence of identical records or values within a dataset. This is not the primary issue described in the scenario. Duplicates can exist regardless of data type validity.

**B. Missing data:** Missing data signifies the absence of a value for a specific field in a record. While missing data is a common data quality issue, it's distinct from the problem of having incorrect data types.

**C. Data outliers:** Outliers are data points that significantly deviate from the other values in a dataset. While outliers can skew statistical analyses, the presence of character values mixed with integers is a more fundamental data integrity issue than simply having extreme values. Outliers pertain to the distribution of the data within the correct data type; invalid data type addresses the foundational issue of conformity to proper data type assignments.

Data validation is a crucial step in data processing, including data cleaning and preparation for analytics and machine learning. This step ensures that data conforms to predefined rules, including data type constraints. Failure to ensure data type validity can lead to errors during data processing, incorrect analysis results, and unreliable insights. Databases and data warehousing solutions rely heavily on defined schemas that specify datatypes for each column to ensure data integrity. When data violates these definitions it is typically rejected or results in unexpected errors.

For further research, you can refer to:

Data Types: https://dev.mysql.com/doc/refman/8.0/en/data-types.html
Data Validation: https://www.ibm.com/docs/en/i/7.3?topic=concepts-data-validation Data Quality: https://www.oracle.com/solutions/data-management/data-quality/

## Question: 7

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

A.MDM
B.ETL
C.OLTP

D.BI

**Answer: B**

**Explanation:**

The correct answer is B. ETL (Extract, Transform, Load) is the core process used in data integration for collecting, blending, and loading data. Let's break down why the other options are incorrect and further explain ETL.

**ETL (Extract, Transform, Load):** This is a three-stage process. First, data is extracted from various source systems (databases, files, APIs, etc.). Next, the data is transformed to clean, standardize, and conform it to a consistent schema for the target system. This transformation might include data cleansing, data enrichment, aggregations, and format conversions. Finally, the transformed data is loaded into the target data warehouse or other data repository. ETL pipelines are crucial for building data warehouses, data lakes, and other data repositories used for analytics and reporting.

**MDM (Master Data Management):** MDM focuses on creating a single, consistent, and authoritative source of master data (e.g., customer data, product data). While MDM can involve data integration, its primary goal is to maintain the quality and consistency of key business entities, not the general process of collecting, blending, and loading diverse data. MDM systems can use ETL to populate their master data stores, but MDM is not itself a process of collecting, blending, and loading.

**OLTP (Online Transaction Processing):** OLTP systems are designed for real-time transaction processing. They focus on handling many concurrent transactions, such as order entry, ATM transactions, and online shopping carts. OLTP systems are data sources for ETL processes but are not the process itself. They are concerned with rapid data insertion and retrieval, not data integration.

**BI (Business Intelligence):** BI encompasses tools and technologies used to analyze data and present actionable information to help executives, managers, and other corporate end users make informed business decisions. While BI relies on integrated data, it is a set of tools and practices built on top of data integration, not the integration process itself. BI uses the outputs of ETL processes.

In summary, ETL is the specific process that directly addresses the question's description of collecting, blending, and loading data for a consolidated view, making it the correct answer. The other options represent related but distinct concepts in the broader data management ecosystem.

For further research, you can explore these resources:

**AWS Data Integration:**https://aws.amazon.com/data-integration/ (Provides information on data integration tools and services in the cloud.)
**Microsoft Azure Data Factory (ETL Cloud Service):**https://azure.microsoft.com/en-us/products/data-factory/ (Describes a cloud-based ETL service.)
**Talend Data Integration:**https://www.talend.com/products/data-integration/ (Overview of an ETL tool.)

## Question: 8

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

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

**Answer: C**

**Explanation:**

The correct answer is C: Get stakeholder approval. Here's why: The dashboard creation process is iterative. After defining requirements, identifying data sources, and creating a wireframe, the next crucial step is to get stakeholder approval. The wireframe represents a visual blueprint of the dashboard. Sharing it with stakeholders ensures that their initial vision is being translated accurately and that the proposed layout, data visualization, and interactive elements meet their needs and expectations. Stakeholder approval at this stage prevents wasted effort and resources on building a dashboard that doesn't align with user requirements. It provides an opportunity to gather feedback, address concerns, and make necessary adjustments before investing significant time in development. Options A, B, and D are premature at this stage. Optimizing the dashboard (A) is done after the dashboard is functional and populated with data. Creating subscriptions (B) comes later in the process, once the dashboard is finalized and deployed. Deploying to production (D) is the final step, after thorough testing and stakeholder sign-off. Skipping stakeholder approval and directly moving to development could result in a final product that doesn't meet business needs, leading to rework, delays, and user dissatisfaction. Getting early buy-in from stakeholders is a key principle in data analytics projects to ensure the dashboard is valuable and effectively used. This aligns with general project management best practices which emphasize iterative development and continuous feedback.

Supporting Resources:

Dashboard Design Principles
Data Visualization Best Practices

**Question: 9**

A data analyst has been asked to derive a new variable labeled “Promotion_flag” based on the total quantity sold by each salesperson. Given the table below:

| Store_ID | Item | Salesperson | Quantity_sold | Promotion_flag |
|---|---|---|---|---|
| 104 | Pax-2 | James | 1,000,300 | |
| 204 | Pax-3 | Paul | 234,578 | |
| 304 | Pax-1 | Peter | 2,000,432 | |
| 404 | Pax-2 | Esther | 1,089,678 | |
| 204 | Pax-3 | May | 126,578 | |
| 304 | Pax-1 | Park | 200,432 | |
| 404 | Pax-2 | Mabel | 1,089,000 | |

Which of the following functions would the analyst consider appropriate to flag “Yes” for every salesperson who has a number above 1,000,000 in the Quantity_sold column?

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

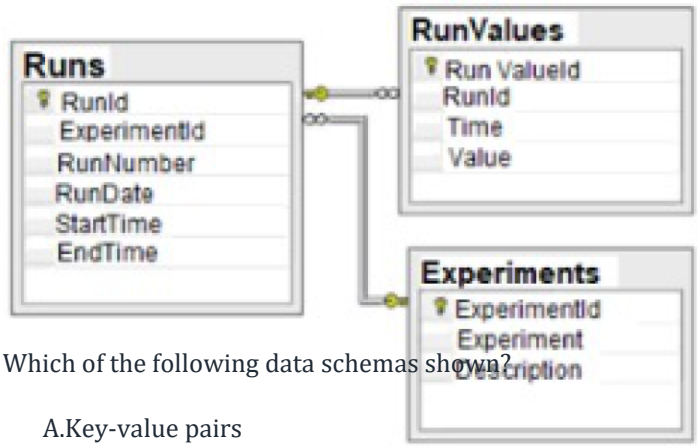
**Answer: C**

**Explanation:**

C. Logical functions would be appropriate to flag “Yes” for every salesperson who has a number above 1,000,000 in the Quantity_sold column. Logical functions in data analysis are used to make decisions based on a set of conditions. In this case, the analyst can use a logical function such as an IF statement, where the condition is that the salesperson's Quantity_sold is above 1,000,000. If the condition is met, then the Promotion_flag is set to “Yes,” otherwise it is set to “No.”For example, the formula could be something like:IF(Quantity_sold > 1000000, "Yes", "No")This formula would then be applied to the entire column of Quantity_sold values, and the resulting Promotion_flag column would indicate whether each salesperson is eligible for a promotion based on the number of items they have sold. Logical functions are an essential tool in data analysis, as they allow analysts to make decisions based on specific criteria and automate certain processes in the data analysis process.

**Question: 10**

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:**

D. The diagram shows a schema for a relational database, which is a type of database that stores data in tables that are related to each other through common columns or keys. In the diagram, the "Runs" table is connected to the "Athletes" and "Races" tables through the "Athlete_id" and "Race_id" columns, respectively.

These columns are foreign keys that reference the primary keys in the "Athletes" and "Races" tables. This type of structure allows data to be stored and accessed in a structured and organized way, and enables relationships between different data entities to be captured and queried. The other options (key-value pairs, online transactional processing, and data lake) do not describe a relational database schema.

**Question: 11**

A company’s marketing department wants to do a promotional campaign next month. A data analyst on the team has been asked to perform customer segmentation, looking at how recently a customer bought product, at what frequency, and at what value. Which of the following types of analysis would this practice be considered?

A.Prescriptive

B.Trend
C.Gap
D.Custer

**Answer: D**

**Explanation:**

The correct answer is **D. Cluster**. Here's why:

The scenario describes a data analyst segmenting customers based on recency, frequency, and monetary value (RFM). RFM analysis is a well-known technique for identifying groups of customers with similar behaviors. This inherently involves grouping or clustering customers based on these attributes.

**Clustering analysis** is a type of unsupervised machine learning technique that aims to group similar data points together. In this case, the data points are customers, and the similarity is based on their RFM values.
The marketing department can then tailor its campaigns to each cluster for better effectiveness.

Let's analyze why the other options are incorrect:

**Prescriptive analysis** focuses on recommending actions or strategies based on data analysis. While the results of the clustering might inform prescriptive actions (e.g., targeting a specific cluster with a promotion), the analysis itself isn't prescriptive. It's about understanding customer groups first.

**Trend analysis** involves identifying patterns and trends in data over time. While you could analyze RFM values over time, the immediate goal isn't to look at trends but to segment customers at a specific point in time for the upcoming campaign.

**Gap analysis** involves identifying the difference between a current state and a desired future state. This isn't what the data analyst is doing. They're not trying to identify gaps but rather trying to create homogenous groups of customers.

Therefore, the most fitting description of the data analyst's work is cluster analysis since they are grouping customers based on similar attributes for segmentation purposes.

For further research on cluster analysis, you can refer to:

**scikit-learn documentation on clustering:**https://scikit-learn.org/stable/modules/clustering.html **RFM Analysis:**https://www.putler.com/rfm-analysis/

## Question: 12

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:**

The best way to meet the stakeholder requirement of tracking all changes to documents before publication,

given multiple users are editing them, is to display the version number next to each submission on the dashboard (Option A). This is because version control is a fundamental aspect of managing documents that undergo frequent modifications, especially in collaborative environments.

Here's a detailed justification:

1. **Explicit Change Tracking:** Displaying version numbers directly associates each submission with a specific iteration of the document. This allows users to immediately see the sequence of changes and identify the latest version.

2. **Auditing and History:** Version numbers create an audit trail, making it easy to track the evolution of the document from initial submission to final publication. Users can revert to previous versions if needed or compare versions to understand what changes were made.

3. **Collaboration and Transparency:** In a multi-user environment, version numbers provide clarity and prevent confusion. Editors can be confident they are working on the correct version and understand the history of the document before making changes.

4. **Data Refresh Date (Option B) is Insufficient:** While a data refresh date indicates the last time the dashboard was updated, it doesn't provide granular information about individual submissions or the specific changes made to them.

5. **Corporate Style Guide (Option C) is Irrelevant:** Adhering to a corporate style guide is important for consistency and branding, but it doesn't address the core requirement of tracking changes to documents.

6. **Permissions (Option D) is Less Transparent:** While permissions are important for access control, restricting users to certain versions can hinder collaboration and transparency. It's better to allow users to see all versions but clearly identify the current version.

7. **Version Control Systems:** The concept of version control is widely used in software development and document management. Systems like Git (for software) and document management systems often incorporate robust versioning features.

8. **Benefits of Version Control:** Version control offers numerous benefits, including:

   **Collaboration:** Multiple users can work on the same document simultaneously without overwriting each other's changes.
   **Change tracking:** Every change is recorded, including who made the change, when, and why.
   **Reverting:** You can easily revert to previous versions of a document.
   **Branching:** You can create different versions of a document for different purposes.

Therefore, displaying the version number next to each submission on the dashboard directly addresses the stakeholder requirement of tracking all changes to documents before publication and facilitates collaboration in a transparent manner.

Authoritative links:

**Version Control:**https://www.atlassian.com/git/tutorials/what-is-version-control
**Document Management System:**https://en.wikipedia.org/wiki/Document_management_system

## Question: 13

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

A.continuous data.

B.categorical data.

C.ordinal data.

D.discrete data.

**Answer: D**

**Explanation:**

Here's a detailed justification for why the number of phone calls received by a call center in a day is considered discrete data:

Discrete data represents countable items. Think of individual, distinct units. You can have 1 call, 2 calls, 100 calls, but you can't have 1.5 calls. The values are whole numbers, and there are gaps between possible values. Essentially, discrete data can only take on specific, isolated values.

In the context of a call center, you count whole phone calls. A call center receives a specific number of complete calls during a day, such as 235 or 480. The number is always a whole number, reflecting the number of distinct phone call events that occurred. It's impossible to have a fraction of a phone call in this scenario; a call either happened or it didn't. Thus, the data is built from distinct, separate events.

Continuous data, on the other hand, can take on any value within a given range. Examples include temperature, height, or weight. Categorical data represents qualities or characteristics, such as eye color or type of service requested. Ordinal data is a type of categorical data with a meaningful order or ranking, like customer satisfaction ratings (e.g., "very dissatisfied," "dissatisfied," "neutral," "satisfied," "very satisfied").

Because phone calls are countable events that exist as separate numbers, the other options are inaccurate.Data classification is crucial in data analysis because it affects the methods used for calculation and representation. For example, the types of data are important when the IT team uses this information to provision server instances, network bandwidth, or storage capacity to handle the call volume.

Authoritative links:

**NIST - Engineering Statistics Handbook:**https://www.itl.nist.gov/div898/handbook/ (Look for sections on types of data)

**Statistics Canada:**https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch8/5214798-eng.htm (Provides definitions and examples of different data types)

## Question: 14

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:**

The correct answer is **A. Static**. Here's why:

A static report is a fixed snapshot of data at a specific point in time. In this scenario, the board meeting on July

15, 2020, requires a finalized view of the sales performance through the second quarter (ending June 30, 2020). A static report perfectly fulfills this need because it captures the data as it existed after the quarter ended and the numbers were finalized, preventing any changes afterward. The data analyst needs to present confirmed sales numbers for evaluation. Static reports are great when you want to use a snapshot of data at a particular point in time.

Real-time reports (Option B) would continuously update, which is undesirable for a finalized report presented to the board. The board needs stable, consistent figures, not a constantly evolving dataset.

Self-service reports (Option C) empower users to generate their own reports. While beneficial in general, they don't guarantee a standardized, prepared report ready for a formal board meeting. The objective is to provide a prepared report.

Dynamic reports (Option D) allow for some user interaction, such as filtering or drill-downs, but are still typically updated and not a fixed snapshot like the static report.

Therefore, because the board requires a fixed, finalized view of sales performance for a specific period, a static report is the most suitable option. The others are not a fit for the needs of the exam question, which is a snapshot of data at a particular point in time.

Further research on static reports:

**IBM - Static Report:**https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=reports-static-reports

## Question: 15

Which of the following would be considered non-personally identifiable information?

A.Cell phone device name
B.Customer’s name
C.Government ID number
D.Telephone number

**Answer: A**

**Explanation:**

The correct answer is A. Cell phone device name is considered non-personally identifiable information (non-PII) because it cannot be directly used, on its own, to identify a specific individual. It's a piece of information that describes the device itself, not its user. While a device name could potentially be correlated with other data to infer information about the user, it lacks the direct link necessary to be considered PII.

Customer’s name, Government ID number, and Telephone number are all PII. Customer's name directly identifies an individual. Government ID numbers (like Social Security numbers) are unique identifiers explicitly designed to link to a specific person and are highly sensitive. Telephone numbers are directly linked to a specific subscriber, effectively identifying them.

PII is any data that could potentially identify a specific individual. The key characteristic of PII is its ability to link back to a unique person either directly or when combined with other data points. Non-PII, on the other hand, is information that, on its own, cannot be used to determine the identity of an individual. This distinction is crucial for data privacy and security regulations, such as GDPR and CCPA.

The cell phone device name, while being a piece of data associated with a person's device, falls into the category of non-PII. A device name might be "John's iPhone" or "Samsung Galaxy S23," but this, by itself, doesn't reveal the owner's full legal name, address, or other identifying details. It's an attribute of the device,

rather than a direct identifier of the person. Context is essential, if the name has the persons full name it would become PII.

Understanding the difference between PII and non-PII is critical when working with data, especially in cloud computing environments where massive amounts of data are processed and stored. Cloud providers offer various services for identifying and classifying data, helping organizations comply with privacy regulations. Proper handling of PII and non-PII is essential for maintaining data privacy and avoiding legal liabilities. Many Cloud computing service providers such as AWS provide data masking services.

For further research on PII and Non-PII, refer to these authoritative links:

**NIST (National Institute of Standards and Technology):**https://www.nist.gov/
**HIPAA (Health Insurance Portability and Accountability Act):**https://www.hhs.gov/hipaa/index.html
**GDPR (General Data Protection Regulation):**https://gdpr-info.eu/

**Question: 16**

Which of the following is the correct data type for text?

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

**Answer: B**

**Explanation:**

The correct data type for storing textual information is a String. Let's examine why other options are incorrect and why a String is the appropriate choice.

A Boolean data type represents a logical value, which can be either true or false. It's used for representing binary states or conditions, not text.

An Integer data type represents whole numbers, both positive and negative, without any decimal points. It is suitable for numerical data such as counts or IDs, not text characters or words.

A Float data type represents numbers with decimal points. While it can store numerical values, it's designed for representing real numbers and doesn't have the capability to represent or store text characters or sequences of characters.

A String data type is designed explicitly to store sequences of characters. These characters can include letters, numbers, symbols, and spaces. Strings are fundamental to representing text in almost every programming language and database system. They are employed in data processing, data analytics, and data visualization to handle text-based information. For example, names, addresses, descriptions, and comments are all typically stored as Strings. Databases utilize String data types (often referred to as VARCHAR or TEXT) to handle textual information. Programming languages use String manipulation functions for tasks such as searching, replacing, and formatting text. Cloud-based data storage solutions also rely on String data types to store textual data effectively.

Therefore, given the options, the String data type is the only suitable and correct choice for storing text.

For more information, you can refer to these resources:

**Data Types on W3Schools:**https://www.w3schools.com/python/python_datatypes.asp

**Data Types on Oracle:**https://docs.oracle.com/cd/B19306_01/server.102/b14200/sql_elements001.htm

## Question: 17

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

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

**Answer: D**

**Explanation:**

The correct answer is D, performing exploratory data analysis (EDA). Here's why, and why the other options are less suitable immediately following the understanding of a business requirement for a data analysis report:

After clearly grasping the business objective, the next logical step is to dive into the available data and gain a preliminary understanding of its characteristics. EDA helps us determine if the available data is suitable and sufficient to answer the business question. This involves inspecting the data types, identifying missing values, detecting outliers, and exploring relationships between variables. By doing this, we can refine our
understanding of the data's potential and limitations. EDA techniques such as data visualization (histograms, scatter plots), summary statistics (mean, median, standard deviation), and correlation analysis are employed.

Option A, rephrasing the business requirement, is unnecessary at this point. We are already assuming we understand the requirement. Rephrasing might be needed later if the data analysis shows a misunderstanding, but not as the immediate next step.

Option B, determining the data necessary for the analysis, is partially correct, but it's intertwined with EDA. While you have a general idea of the required data, EDA helps you confirm that the data exists in the expected format and with sufficient quality. It might reveal unexpected data gaps or inconsistencies that necessitate adjusting the data requirements or refining the scope of the analysis. This information informs a more complete definition of the required data.

Option C, building a mock dashboard/presentation layout, is premature. Without understanding the data's content and quality through EDA, the mock dashboard would be based on assumptions rather than reality. EDA informs the selection of appropriate visualizations and ensures that the dashboard effectively presents the findings that emerge from the data.

In short, EDA provides the crucial bridge between the business requirement and the subsequent steps of data preparation, modeling, and visualization. It's the initial phase of investigation and discovery needed before more detailed analysis can begin.

Authoritative resources for further research:

**Towards Data Science - Exploratory Data Analysis**: https://towardsdatascience.com/exploratory-data-analysis-8e25eda1901f
**Wikipedia - Exploratory Data Analysis**: https://en.wikipedia.org/wiki/Exploratory_data_analysis **Analytics Vidhya - Exploratory Data Analysis**:
https://www.analyticsvidhya.com/blog/2020/03/understanding-exploratory-data-analysis/

**Question: 18**

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

A.SAS
B.Microsoft Power BI
C.IBM SPSS
D.Python

**Answer: D**

**Explanation:**

The correct answer is D. Python is indeed a popular choice in data analytics due to its versatility and extensive library ecosystem. Here's why:

**General-Purpose Programming Language:** Python is not solely a statistical or data analysis tool; it's a fully functional programming language capable of handling various tasks, from web development to automation. This broad applicability makes it a valuable asset for data professionals who need to perform tasks beyond data analysis itself.

**Interpreted Language:** Python is interpreted, meaning code is executed line by line, making it easier to debug and test, especially during the exploratory data analysis phase.

**Data Analysis Libraries:** Python boasts a rich collection of libraries specifically designed for data analysis. These libraries significantly simplify complex tasks:

**NumPy:** Provides powerful numerical computing capabilities, including array operations and linear algebra.
**Pandas:** Offers data structures (like DataFrames) and data analysis tools for manipulating and analyzing structured data.
**Scikit-learn:** Provides machine learning algorithms for classification, regression, clustering, and model selection.
**Matplotlib and Seaborn:** Enable data visualization, allowing analysts to create informative charts and graphs.

**Community and Resources:** Python has a large and active community, meaning ample online resources, tutorials, and support are available for data analysts.

**Integration with Cloud Platforms:** Python seamlessly integrates with cloud platforms like AWS, Azure, and Google Cloud. This is crucial for data analytics projects that leverage cloud storage, computing resources, and machine learning services. For example, Python can be used with AWS Sagemaker or Google Cloud AI Platform.

Now, let's examine why the other options are incorrect:

**A. SAS:** While SAS is a powerful statistical software suite, it is primarily a statistical programming language and environment rather than a general-purpose programming language.
**B. Microsoft Power BI:** Power BI is a business intelligence tool primarily used for data visualization and reporting. It's excellent for creating dashboards, but not a general-purpose programming language. **C. IBM SPSS:** SPSS is statistical software used for statistical analysis and data management, not a general-purpose programming language.

In conclusion, Python's combination of being a general-purpose programming language, having a comprehensive ecosystem of data analysis libraries, its ease of use due to it's interpreted nature, and it's integration with Cloud Computing makes it the correct answer.

**Authoritative Links:**

**Python's official website:**https://www.python.org/
**NumPy:**https://numpy.org/
**Pandas:**https://pandas.pydata.org/
**Scikit-learn:**https://scikit-learn.org/stable/

## Question: 19

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

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

**Answer: D**

**Explanation:**

The best way for a data analyst to communicate the online marketing campaign results and ROI to a marketing manager is through a summary report (Option D). Here's why:

Option D directly addresses the manager's needs. The manager wants to see key performance indicators (KPIs) and the return on marketing investment (ROI). A summary report can efficiently and clearly present these elements. The "statistics" can include metrics like click-through rates, conversion rates, cost per acquisition, and ROI. "Conclusions" interpret these statistics, explaining what they mean for the campaign's success. "Recommendations" offer actionable steps for future campaigns based on the findings. This provides the manager with a concise and valuable overview.

Option A (real-time monitor) might be useful for monitoring the campaign during its launch, but it's not the most effective way to communicate a summary of results and ROI. Real-time data can be overwhelming without proper context and analysis.

Option B (self-service dashboard) may be helpful for overall budget performance, but it doesn't specifically address the manager's request for the online marketing campaign's KPIs and ROI. While a self-service dashboard could be useful for deeper dives, a summary is better for initial presentation.

Option C (spreadsheet of raw data) is the least effective option. Raw data is difficult to interpret and requires significant processing. The manager likely lacks the time and expertise to analyze the data themselves. It fails to highlight the most important KPIs and calculate ROI in an accessible manner.

Therefore, the data analyst providing a tailored summary with statistics, conclusions, and recommendations is the most efficient and impactful way to communicate the campaign's performance and ROI to the marketing manager, allowing for informed decision-making.

To further research creating effective data summaries, reports, and dashboards, consider these resources:

**Tableau - Best Practices for Dashboard Design:**https://www.tableau.com/solutions/dashboard-design-best-practices
**HubSpot - Marketing Report Templates:**https://blog.hubspot.com/marketing/marketing-report-templates

## Question: 20

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

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

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

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

**Answer: D**

**Explanation:**

D. In order to determine the most popular movie genre, the data analyst must first delimit the Genre column. Delimiting means separating or breaking up the data in a column into separate values or categories. In this case, the Genre column contains multiple movie genres for each movie, which are separated by commas. In order to determine the most popular movie genre, the data analyst must first split the Genre column into separate rows or categories for each movie genre. This will allow the data analyst to count the number of movies in each genre and determine the most popular genre. The other options (append, merge, and concatenate) do not involve separating or breaking up data in a column.

**Question: 21**

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

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

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

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

D.The new layout has the lowest conversion rates in the United Kingdom.

**Answer: A**

**Explanation:**

A. In Germany, the increase in conversion from the new layout was not significant.This conclusion is accurate because the p-value for Germany is 0.13, which is higher than the commonly used significance level of 0.05. A p-value of 0.13 means that there is a 13% chance that the increase in conversion from the new layout was due to random chance, and therefore, the increase is not considered statistically significant. This means that we cannot reject the null hypothesis that the new layout does not have any effect on the conversion rate in Germany.In contrast, the p-values for the United States, France, and Canada are lower than 0.05, meaning that the increase in conversion from the new layout in these countries is considered statistically significant.

So, we can conclude that users who visit the new website are more likely to make a purchase in these countries

**Question: 22**

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:**

B. A pie chart would be the best chart to use to identify the composition between the categories of the survey response data set. A pie chart is a circular chart that shows the proportions or percentages of a whole. Each slice of the pie represents a category, and the size of the slice represents the proportion or percentage of the total that belongs to that category. Pie charts are useful for showing the relative sizes of categories or for comparing the proportions of different categories. The other chart types (histogram, line, scatter plot, and waterfall) are not appropriate for showing the composition between categories.

## Question: 23

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

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

**Answer: A**

**Explanation:**

The mean, also known as the average, is calculated by summing all the values in a dataset and then dividing by the number of values. In this scenario, the dataset consists of the heights of five dogs: 300mm, 430mm, 170mm, 470mm, and 600mm.

To find the mean height, we first sum these values: 300 + 430 + 170 + 470 + 600 = 1970.

Next, we divide the total sum by the number of dogs, which is 5: 1970 / 5 = 394.

Therefore, the mean height of the five dogs is 394mm. This result confirms that option A, 394mm, is the correct answer. The other options can be eliminated because they do not accurately represent the calculated average. The mean provides a central tendency measure, representing a typical value within the dataset.

While not directly related to cloud computing, the concept of calculating a mean is fundamental in data analysis, which is heavily utilized in cloud-based analytics platforms for processing large datasets. Cloud services like AWS, Azure, and Google Cloud provide tools for statistical calculations, including the computation of means for various metrics. If these were terabytes of pet heights, cloud solutions would be ideal for processing.

For more information on calculating the mean, you can refer to resources like:

**Khan Academy on Average/Mean:**https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/mean-median-basics/a/mean-median-and-mode-review
**Statistics How To on Mean:**https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/mean-median-mode/

## Question: 24

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

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

**Answer: AD**

**Explanation:**

The correct answers are **B. To remember specifics about data fields** and **D. To provide continuity through personnel turnover**.

Here's the justification:

A data dictionary, also known as a metadata repository, is a centralized repository of information about data. It documents the characteristics of data elements within a database system. One crucial reason to maintain a data dictionary is **to remember specifics about data fields (B)**. It provides a single source of truth for data definitions, including data types, formats, lengths, allowed values, and relationships with other data elements. This documentation ensures consistency and accuracy in data usage across different applications and teams.

Without a data dictionary, tribal knowledge becomes the primary means of understanding data, which is unreliable and inefficient.

Furthermore, data dictionaries **provide continuity through personnel turnover (D)**. When employees leave an organization, their knowledge of the data disappears with them. A comprehensive data dictionary mitigates this risk by codifying the understanding of data assets. New team members can quickly learn about the data landscape, fostering faster onboarding and reducing the risk of errors due to misunderstanding data structures. A well-maintained data dictionary acts as an institutional memory, preserving vital information and ensuring smooth operations even when staff changes occur.

Options A, C, E, and F are incorrect because they do not directly relate to the core purpose of a data dictionary. Data acquisition (A) is typically improved by data governance strategies and data integration tools.

User groups (C) are managed through database security features and access control mechanisms. PHI breaches (E) are prevented by data security policies, encryption, and access control. Processing power (F) is optimized through efficient database design, indexing, and query optimization. While a data dictionary can indirectly support some of these areas, they aren't the primary justifications for its existence.

Authoritative resources for further research:

**Database Design - 2nd Edition by Adrian Gulliver:** This book provides in-depth coverage of database design concepts, including the role of data dictionaries.

**Data Management Body of Knowledge (DMBOK) by DAMA International:** This comprehensive guide covers all aspects of data management, including metadata management and data dictionaries.

**Wikipedia - Data dictionary:**https://en.wikipedia.org/wiki/Data_dictionary

## Question: 25

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 discrepancy in the recorded event times across the two databases is that they are operating in different time zones. Data stored globally often needs to be normalized to a consistent time standard, like UTC (Coordinated Universal Time), but if the databases haven't been configured to do so, they

might be logging times based on their local geographical time zones.

Option A, that the data analyst is querying the databases incorrectly, is less likely because the time difference is consistently three hours, suggesting a systemic issue rather than a query error. Option B, that the databases are recording different events, is also less likely if the event is recurring and supposed to be the same across both locations. Option D, that the second database is logging incorrectly, could be a possibility, but attributing it solely to logging errors without considering the time zone difference is premature, especially with a consistent three-hour deviation.

Time zone differences are a common challenge in distributed systems and databases. Without proper handling, time-sensitive data can become skewed and lead to incorrect analysis and reporting. The databases likely record timestamps based on the server's operating system time zone setting. If one server is configured for Eastern Time (ET) and the other for Pacific Time (PT), the three-hour difference would be explained by the difference between these time zones.

Therefore, the most logical and directly relevant explanation is the difference in time zone configurations between the two databases. Properly configuring databases for consistent time zone handling is essential for data integrity and analysis in distributed environments.

Supporting Link:

PostgreSQL Time Zones (Example of time zone handling in a database system)
Understanding Time Zones in Databases

## Question: 26

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

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

**Answer: A**

**Explanation:**

The correct answer is A. CSV file. A flat file database stores data in a plain text format, where each row represents a record and columns are separated by a delimiter, most commonly a comma. This structure makes it simple to read and process using standard text editors and scripting languages.

A CSV (Comma Separated Values) file perfectly embodies this structure. Each line represents a record, and the values within each record are separated by commas. Therefore, it's a straightforward example of a flat file.

PDF files are designed for document presentation and are not structured for data analysis like flat files are. JSON files are hierarchical and use a key-value pair structure, representing more complex data structures than flat files can accommodate. JPEGs are image files that are binary and don't store data in a plain text, row-and-column format.

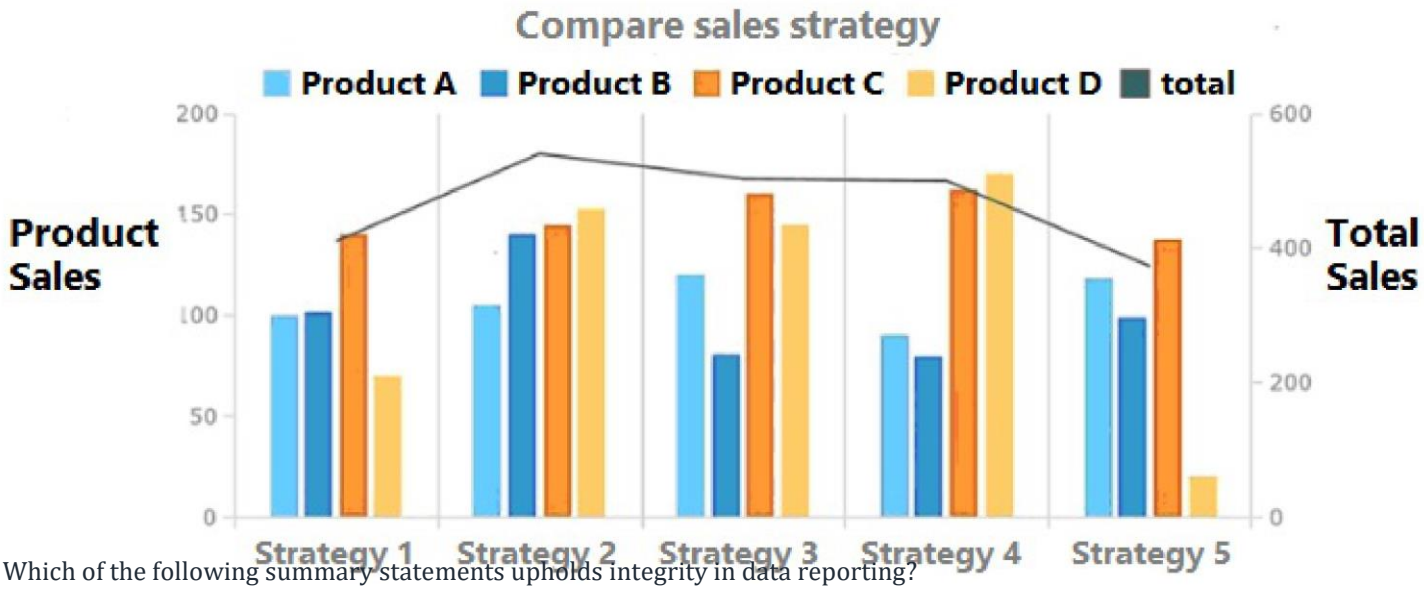
Flat files, and specifically CSV files, are often used for data import and export in data analysis tasks because of their simplicity and widespread compatibility with various data processing tools. They are easy to create, parse, and manipulate, making them a convenient choice for storing tabular data for analysis.

Further Reading:

Flat File Definition: https://www.techopedia.com/definition/2574/flat-file CSV
File Format: https://www.rfc-editor.org/rfc/rfc4180

**Question: 27**

Given the following graph:




Which of the following summary statements upholds integrity in data reporting?

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

**Answer: A**

**Explanation:**

A. Sales are approximately equal for Product A and Product B across all strategies.

This statement upholds integrity in data reporting because it is based on the data shown in the graph, which shows that the sales for Product A and Product B are approximately equal across all strategies. The statement is objective and accurately reflects the data, without drawing any conclusions or biases. By presenting the data in an impartial and unbiased manner, the statement upholds the integrity of the data reporting.

**Question: 28**

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

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

**Answer: A**

**Explanation:**

The correct answer is A, Web scraping, because it is the most suitable method for extracting data directly from website content, like news articles. Web scraping involves using automated tools or scripts to parse the HTML or XML structure of a webpage and extract the desired information, such as text, images, and links.

This is crucial when the data is embedded within the website's layout and not available through APIs or downloadable files. Sampling (B) focuses on selecting a representative subset of data, not acquiring it from a source. Data Wrangling (C) is about cleaning, transforming, and preparing data after it has been acquired, not the initial extraction process. ETL (D) (Extract, Transform, Load) is a broader data integration process, involving extracting data from various sources, transforming it, and loading it into a data warehouse or other data repository; while extraction is part of ETL, web scraping is the specific technique used to extract the data from the web in this scenario. Web scraping libraries and tools are specifically designed to automate the process of navigating websites and collecting data, making it ideal for the described scenario of acquiring data from digital news articles for text analysis.

Further reading on Web Scraping:

**Web Scraping with Python:**https://realpython.com/beautiful-soup-web-scraper-python/ (Real Python tutorial)
**Web Scraping - Wikipedia:**https://en.wikipedia.org/wiki/Web_scraping (General Overview)

## Question: 29

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:**

Here's a detailed justification for why the correct answer is C (10,600), given the provided information about datapoint growth:

The problem describes a scenario where the number of datapoints increases daily by approximately 20% of the previous day's total. We are given that on a certain day, the total number of datapoints is 8,798. To calculate the estimated number of datapoints for the next day, we need to determine 20% of 8,798 and add it to the original number.

First, we calculate 20% of 8,798:

0. 20 * 8798 = 1759.6

Next, we add this increase to the original number of datapoints:

8798 + 1759.6 = 10557.6

Now we must round this value to the nearest answer choice given.

Looking at the answer options:

A. 7,038 is significantly lower, indicating a decrease, not an increase.

B. 9,600 is lower than 10557.6 and does not accurately reflect the 20% increase. C. 10,600 is the closest value to our calculated value of 10557.6.

D. 10,800 is more of a departure from the calculated value of 10557.6 than option C is.

Therefore, the closest and most reasonable estimate for the total number of datapoints on the next day is 10,600. This solution directly applies the percentage increase provided in the problem statement and selects the option closest to the computed result.

**Question: 30**

An analyst has been tracking company intranet usage and has been asked to create a chat to show the most-used/most-clicked portions of a homepage that contains more than 30 links. Which of the following visualizations would BEST illustrate this information?

A.Scatter plot

B.Heat map

C.Pie chart

D.Infographic

**Answer: B**

**Explanation:**

A heat map is the best visualization for showcasing the most-used/most-clicked portions of a homepage with numerous links because it visually represents data through variations in color. In this scenario, a heat map could overlay the homepage design, with the most frequently clicked links appearing in warmer colors (e.g., red, orange) and less frequently clicked links in cooler colors (e.g., blue, green). This provides an immediate and intuitive understanding of user activity and areas of interest on the page.

A scatter plot primarily displays the relationship between two variables, making it unsuitable for highlighting the usage frequency of multiple links. A pie chart is useful for showing the proportion of each category to the whole but is less effective when dealing with a large number of categories (30+ links), as it becomes cluttered and difficult to interpret. An infographic can combine visuals and text to tell a story, but a heat map offers a more direct and concise visual representation of the data in question, specifically highlighting areas of concentration.

The strength of a heat map lies in its ability to quickly communicate patterns and trends within complex data sets. By using color intensity, it allows the analyst to easily identify the "hot spots" (most clicked areas) on the homepage, enabling data-driven decisions for website optimization and user experience improvements.

Furthermore, heatmaps can be dynamically generated, updating in real-time to reflect the most current data, allowing for the ongoing monitoring of user behavior and the adaptation of website layout or content as needed.

Authoritative Links:

Tableau - Heatmaps: https://help.tableau.com/current/pro/desktop/en-us/buildexamples_heatmap.htm
ResearchGate - A survey of heatmap visualization: https://www.researchgate.net/publication/311268033_A_survey_of_heatmap_visualization

**Question: 31**

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

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

Which of the following functions did the analyst use to generate the data in the Sales_indicator column?

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

**Answer: B**

**Explanation:**

The function that the analyst used to generate the data in the Sales_indicator column is likely a Logical function. The function likely uses a logical comparison between the sales data and a fixed value of $50,000 to determine whether each month's sales exceeded the threshold. For example, the formula may look like =IF(sales > 50000, "Exceeded", "Not Exceeded").

**Question: 32**

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:**

The correct answer is C: Filter on any of the responses that do not say “January” and update them to “January”. This is the most effective approach to ensure data consistency in this scenario. The goal is to standardize the representation of January across all survey responses.

Option A, deleting responses, is unacceptable as it leads to data loss and skews the survey results. All data points representing January are valid; only their format differs. Deletion is not a valid data cleaning technique unless the data is genuinely incorrect or irrelevant.

Option B, replacing "01," only addresses one specific inconsistency. It doesn't handle the "Jan" variation and leaves the dataset partially inconsistent. A complete solution should address all variants of January.

Option D, sorting and updating, is not practical and efficient. Sorting might help identify variations, but it's a manual process for a large dataset. Moreover, updating "Jan" to "01" doesn't improve readability and clarity compared to a full month name.

Option C provides a comprehensive solution. Filtering allows identifying all entries that don't conform to the desired "January" format. Then, a bulk update can be performed to change these filtered entries to "January." This results in a dataset where all instances of January are uniformly represented. Data cleaning tools often have filter-and-replace functionalities built-in to facilitate such transformations. This standardization enhances data analysis and reporting by eliminating ambiguity and simplifying aggregation and comparisons. Using a standardized format like "January" also improves data readability and interpretability. This is crucial for ensuring the validity of subsequent analysis and insights derived from the data.

For more information on data cleaning and standardization, you can refer to resources from organizations like the National Institute of Standards and Technology (NIST) and professional data management bodies. For example, the DAMA-DMBOK (Data Management Body of Knowledge) is a widely respected framework for data management principles and practices. While a specific URL for such a general concept isn't ideal, searching for "Data Quality Management best practices" will yield numerous resources. Specific software documentation for tools like Pandas (Python) or SQL databases will also contain relevant information.

**Question: 33**

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:**

The correct answer is B, duplicate data. Here's a detailed justification:

Data cleansing is the process of identifying and correcting errors and inconsistencies in a dataset to improve its quality. Several data quality issues can arise, including missing values, duplicate records, redundant information, and invalid entries. Each issue requires specific techniques to be addressed effectively.

The DISTINCT function, commonly found in SQL and data manipulation tools, focuses on identifying and removing duplicate rows from a dataset. Its primary purpose is to return only unique records, eliminating redundancies based on identical values across all specified columns or the entire row. When applied, the function compares each row in the dataset and removes any rows that are identical to another row, leaving only unique entries.

Missing data (option A) refers to situations where a data point is absent or unavailable. The DISTINCT function has no effect on missing values; other techniques like imputation or data deletion are required to handle this. Redundant data (option C), while related to duplication, generally refers to information that is unnecessarily repeated across different columns or tables, not identical rows. While removing duplicates can sometimes indirectly reduce redundancy, it's not the primary function or consistent outcome. Invalid data (option D) signifies data points that do not conform to predefined constraints, such as incorrect data types or values outside of an acceptable range. The DISTINCT function is incapable of detecting or correcting invalid data, as it only compares row equality.

Therefore, the DISTINCT function specifically tackles the problem of duplicate data by ensuring that only unique rows are retained in the final dataset. It's a fundamental tool for data deduplication, which is a crucial step in ensuring data integrity and accuracy.

For more information on data cleaning techniques and the DISTINCT function, refer to the following resources:

1. **Data Cleaning Overview**: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-396984b36c41
2. **SQL DISTINCT Clause**: https://www.w3schools.com/sql/sql_distinct.asp

These resources provide further context on data cleaning methodologies and the specific functionality of the DISTINCT function in database management systems.

**Question: 34**

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

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

**Answer: D**

**Explanation:**

Here's a detailed justification for why option D is the most likely to result in a representative sample for the income survey, along with why the other options are less suitable:

Option D, "Surveys sent to 100 randomly selected homes that are reflective of the population," is the best choice because it directly addresses the core requirement of representative sampling: ensuring that the sample accurately reflects the characteristics of the overall population. Random selection minimizes bias, giving each household an equal chance of being included. Furthermore, ensuring the selected homes are "reflective of the population" suggests a consideration of demographic factors like location (urban vs. rural), household size, and other relevant socio-economic indicators. This approach aims to create a miniature version of the county's population within the sample.

Option A, "A stratified phone survey of 100 people that is conducted between 2:00 p.m. and 3:00 p.m.," introduces several biases. Conducting the survey during a specific time of day (2:00-3:00 p.m.) excludes those working or unavailable at that time. A phone survey may exclude households without landlines or those who prefer not to answer unsolicited calls. Stratification could be beneficial if done correctly.

Option B, "A systematic survey that is sent to 100 single-family homes in the county," suffers from potential biases. Systematic sampling, while seemingly simple, can introduce bias if there's a pattern in the population data that aligns with the sampling interval. Focusing solely on single-family homes excludes apartment dwellers and those living in multi-family units, skewing the representation of household income.

Option C, "Surveys sent to ten randomly selected homes within 5mi (8km) of the county's office," creates significant geographic bias. Concentrating the sample in a small radius around the county office fails to capture the diversity of the entire county. This approach over-represents the income levels of individuals residing near the county office.

For a survey to accurately reflect the average income of a community, each household needs an equal opportunity to participate. This requires random sampling and adequate sample size. Stratification is also effective to ensure equal group representation.

Further research:

**Sampling Methods:**https://www.statisticshowto.com/probability-and-statistics/sampling-techniques/
**Random Sampling:**https://www.investopedia.com/terms/r/random_sampling.asp

**Question: 35**

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:**

The correct answer is B, the Chi-squared test. Here's why:

Chi-squared tests are specifically designed to analyze the relationship between two or more categorical variables. Categorical variables are those that represent groups or categories, such as eye color (blue, brown, green) or customer segment (enterprise, small business, individual). The Chi-squared test determines if there's a statistically significant association between these categories. It assesses whether the observed distribution of data across different categories deviates significantly from what would be expected if the variables were independent. In simpler terms, it checks if the categories of one variable influence the categories of another.

Simple linear regression, option A, is used to model the linear relationship between a single independent (predictor) variable and a continuous dependent variable. It aims to find the best-fitting line to predict the value of the dependent variable based on the independent variable. The Z-test (option C) and the Two-sample t-test (option D) are used to compare the means of populations or samples. They both typically involve numerical/continuous data rather than categorical data. The Z-test is generally used when the population standard deviation is known, or the sample size is large (typically n > 30). The Two-sample t-test is used to compare the means of two independent groups when the population standard deviation is unknown and the sample size is small. These tests analyze the difference between the means of these two groups. As such, they do not involve the relationship between two or more categorical variables.

Therefore, the Chi-squared test is the only method from the options provided that deals specifically with analyzing relationships between multiple categorical variables.

For further reading:

**Chi-Square Test:**https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/chi-square/
**Simple Linear Regression:**https://www.investopedia.com/terms/r/regression.asp
**Z-Test:**https://www.investopedia.com/terms/z/z-test.asp
**Two-Sample T-Test:**https://www.investopedia.com/terms/t/t-test.asp

**Question: 36**

Which of the following data manipulation techniques is an example of a logical function?

A.WHERE
B.AGGREGATE
C.BOOLEAN
D.IF

**Answer: D**

**Explanation:**

Here's a detailed justification for why the answer is D (IF) in the context of data manipulation and logical functions:

The question asks for an example of a data manipulation technique that exemplifies a logical function. Logical functions are fundamental in programming and data manipulation because they evaluate conditions and return a boolean value (TRUE or FALSE). These boolean values then determine subsequent actions or data transformations.

**A. WHERE:** The WHERE clause is commonly used in SQL to filter data based on specified conditions. While it uses conditions, it primarily serves to select rows that meet certain criteria, rather than directly manipulating the data based on a boolean result within a single field.

**B. AGGREGATE:** Aggregate functions (like SUM, AVG, COUNT, MIN, MAX) perform calculations across multiple rows of data to produce a single summary value. They don't fundamentally operate on boolean logic at the individual data point level.

**C. BOOLEAN:** While "BOOLEAN" describes a data type (TRUE or FALSE), it isn't a function that manipulates data. It's the result of a logical operation, not the operation itself.

**D. IF:** The IF function (or its equivalents like CASE statements in SQL or conditional statements in programming languages) directly embodies logical functionality. It evaluates a condition (which produces a boolean result) and then executes different actions or returns different values depending on whether the condition is TRUE or FALSE. This is a direct manipulation based on a logical evaluation.

Consider an example: IF(Sales > 1000, "High Sales", "Low Sales"). Here, the IF function evaluates the condition Sales > 1000. If the condition is TRUE, the function returns "High Sales"; otherwise, it returns "Low Sales". This directly alters or categorizes the data based on a logical test. This kind of functionality is crucial for data cleaning, transformation, and creating new derived columns based on conditions. Other variations of the IF function involve the "else if" where multiple condition can be verified.

**Authoritative Links:**

**SQL CASE Statement (equivalent to IF):**https://www.w3schools.com/sql/sql_case.asp
**Excel IF Function:**https://support.microsoft.com/en-us/office/if-function-69aed7c9-4e7a-4755-a9bc-aa8bbff73be2

**Question: 37**

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 management. Users can filter to see the data they want.

D.Create a dashboard with views for team, individuals, and management. Configure permissions to control access.

**Answer: A**

**Explanation:**

The best solution is to create a dashboard displaying sales numbers, pipeline, and team performance, with data refresh information. This addresses the need for current sales data visibility. Crucially, access controls and permissions should be configured to restrict access to individual commission data, fulfilling the confidentiality requirement. Options B, C, and D are less effective. Option B limits visibility to the management team, failing to address the entire sales team's needs. Option C, using filters, is insecure, as tech-savvy users might circumvent the intended restrictions and view confidential information. Option D, while incorporating role-based views and permissions, may not offer the necessary granularity in access control compared to individual data access restrictions. Displaying the data refresh date ensures transparency and allows users to understand the data's currency. Data security and access control are paramount, necessitating a solution that prioritizes permission management alongside data presentation.

Useful links:

**Data Visualization Best Practices:**https://www.tableau.com/learn/articles/data-visualization-best-practices **Role-Based Access Control (RBAC):**https://cloud.google.com/iam/docs/understanding-roles (Example using Google Cloud, but the principles are universal)

## Question: 38

Which of the following is a characteristic of a relational database?

A.It utilizes key-value pairs.

B.It has undefined fields.

C.It is structured in nature.

D.It uses minimal memory.

**Answer: C**

**Explanation:**

The correct answer is C. It is structured in nature. Here's why:

Relational databases, like those often deployed in cloud environments such as Amazon RDS, Azure SQL Database, or Google Cloud SQL, are fundamentally structured. This structure is defined by tables, where each table consists of rows (records) and columns (fields or attributes). Each column has a defined data type (e.g., integer, text, date), ensuring data consistency. The relationships between these tables are established using keys, primarily primary keys and foreign keys.

This structured approach contrasts sharply with other database types. Option A, "It utilizes key-value pairs," describes NoSQL databases like document stores (e.g., MongoDB) or key-value stores (e.g., Redis), not relational databases. Option B, "It has undefined fields," is the opposite of the rigid structure found in relational databases, where every field must adhere to a defined data type. Option D, "It uses minimal

memory," is not a defining characteristic; memory usage depends on the database size and workload, not the relational model itself. Relational databases can be very resource intensive.

The strict schema imposed by relational databases allows for efficient querying and data integrity, vital for transactional systems and reporting. The structured nature allows for the use of SQL (Structured Query Language) to interact with the data. This consistency and structure are key benefits of relational databases.

Further research:

**Relational Database Concepts:**https://www.essentialsql.com/what-is-a-relational-database/ **SQL Tutorial:**https://www.w3schools.com/sql/

## Question: 39

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

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

**Answer: C**

**Explanation:**

The correct date range for the sales report is January 1, 2020, to April 8, 2020. Here's why:

The request is for a "year-to-date" (YTD) sales report as of the morning of April 9, 2020. YTD means from the beginning of the current year up to the present. Because the prompt indicates that data is current "through the end of the day," implying the end of April 8th has already passed, the report should include all sales data from January 1, 2020, up to and including April 8, 2020. April 9th's sales data is still accumulating and shouldn't be included in a report generated on the morning of April 9th. Therefore, option C accurately captures the period for which sales data is available and relevant for an YTD report created at that specific time. Options A and B are both incorrect as they exclude valid sales data from the beginning of April, which should be included in a comprehensive report. Option D is also incorrect because a sales report created on the morning of April 9th would not yet incorporate data for the entirety of April 9th, as the sales day is still in progress.

Further research on calculating date ranges for reports and understanding the meaning of "year-to-date" can be found on resources that pertain to business analysis and financial reporting. Look for general resources rather than those tied to cloud computing.

A helpful definition of year-to-date can be found on Investopedia:
https://www.investopedia.com/terms/y/yeartodate.asp

## Question: 40

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: C**

**Explanation:**

C. Standardization of data field names.

In Master Data Management (MDM), the first step is to standardize data field names to ensure consistency across different systems and datasets. This process helps in creating a uniform structure, which is essential before performing other tasks like consolidation, compliance checks, or creating a data dictionary.

**Question: 41**

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

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

**Answer: C**

**Explanation:**

The correct answer is C. Microsoft Excel is widely recognized and extensively used for calculations and pivot tables due to its built-in features specifically designed for these tasks. Its spreadsheet format allows for easy organization and manipulation of data, facilitating both simple and complex calculations through formulas and functions. Pivot tables are a key component of Excel, enabling users to summarize and analyze large datasets by grouping and aggregating data based on different criteria.

While other options may have statistical or analytical capabilities, they are not as commonly used or

inherently designed for the broad range of calculations and pivot table functionalities offered by Excel. IBM SPSS is primarily a statistical software package used for advanced statistical analysis. SAS is also focused on statistical analysis, often used in enterprise environments for business intelligence and predictive analytics.
Domo is a cloud-based business intelligence platform focused on data visualization and dashboards.

Excel's intuitive interface and readily available functions make it a standard tool for data manipulation and reporting. Its capabilities for calculations range from basic arithmetic operations to advanced statistical and financial functions. Pivot tables allow users to quickly create summaries, cross-tabulations, and interactive reports to gain insights from their data. The user-friendliness and widespread accessibility of Excel contribute to its popularity for these specific tasks.

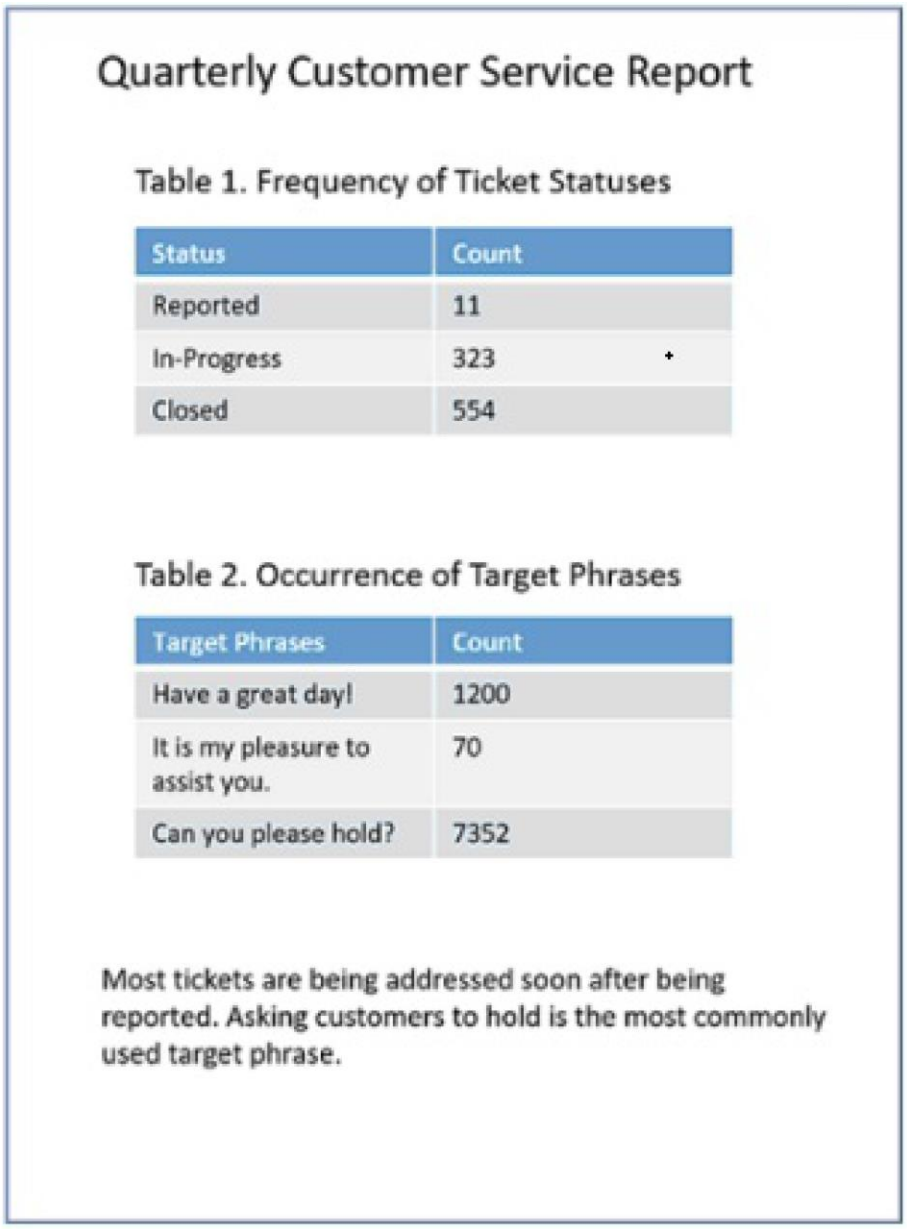
In conclusion, although the other software options have advanced capabilities, Microsoft Excel's design specifically caters to calculations and pivot tables in a way that is readily accessible and easily applicable to a broad range of data analysis needs.

References:

Microsoft Excel: https://www.microsoft.com/en-us/microsoft-365/excel
Pivot Tables: https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f991f7937291

**Question: 42**

Given the following report:




Which of the following components need to be added to ensure the report is point-in-time and static? (Choose two.)

A.A control group for the phrases
B.A summary of the KPIs
C.Filter buttons for the status
D.The date when the report was last accessed
E.The time period the report covers
F.The date on which the report was run

**Answer: EF**

**Explanation:**

E. The time period the report covers :This specifies the exact range of data included in the report, ensuring users understand the timeframe of the analysis.

F. The date on which the report was run :This captures when the report was generated, making it clear that the data reflects a particular snapshot in time.

**Question: 43**

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: BD**

**Explanation:**

Here's a detailed justification for why Integrity and Consistency are the best reasons to validate data quality, along with supporting explanations and resources:

**Justification:**

Data validation is critical for ensuring the usefulness and reliability of data for analysis and decision-making. Among the options, Integrity and Consistency are the most directly related to the concept of data quality in this context.

**Integrity:** Data integrity refers to the accuracy and completeness of data. Validating data for integrity checks if the data has been altered, corrupted, or lost during any stage of its lifecycle, from acquisition to storage. Without integrity, data can become unreliable, leading to incorrect analyses and flawed conclusions. Imagine a scenario where financial records are incomplete, leading to inaccurate profit calculations. Data validation ensures that the data adheres to predefined rules and constraints, minimizing errors and ensuring validity.

**Consistency:** Data consistency means that the same data should be represented in the same way across different datasets, systems, or locations. Validating data for consistency ensures that there are no contradictions or conflicts in the data. Inconsistent data can lead to confusion, errors, and difficulties in integrating data from different sources. For example, a customer address might be formatted differently in sales and marketing databases; validation would ensure a uniform format.

**Why the other options are less appropriate:**

**Retention:** While data retention is important for compliance and historical analysis, it doesn't directly validate the quality of the data itself. Retention policies define how long data is stored, not its accuracy or reliability.

**Transmission:** Data transmission concerns the reliable transfer of data between systems. While transmission errors can affect data quality, the validation process focuses on the data after it's been transmitted to ensure its integrity and consistency, regardless of potential transmission issues.

**Encryption:** Encryption protects data confidentiality, but it doesn't guarantee data quality. Encrypted data can still be inaccurate or inconsistent. Validation ensures the quality of the data irrespective of whether it's encrypted or not.

**Deletion:** Similar to retention, deletion pertains to data lifecycle management, not data quality. Deleting data, whether necessary or harmful to its quality, does not help in validation

**In summary:** Data integrity ensures data is accurate and complete, while data consistency ensures the data's uniformity across different systems. Validating data for these two characteristics guarantees the highest level of data quality.

**Supporting Resources:**

**Data Quality Dimensions:**
https://www.ibm.com/docs/en/SSEPGG_11.5.0/com.ibm.db2.luw.admin.dbconcepts.doc/doc/c0004756.html (IBM's definition of data quality dimensions including completeness, consistency, and accuracy)
**What is Data Integrity?:**https://www.oracle.com/database/what-is-data-integrity/ (Oracle's overview of data integrity principles)

## Question: 44

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

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

**Answer: C**

**Explanation:**

Here's a detailed justification for why link analysis is the best choice in this scenario:

The question highlights a need to understand connections or relationships between different data points. The analyst isn't necessarily looking at changes over time, efficiency, or overall data overview but specifically how data points relate to each other.

**Link analysis** directly addresses this requirement. It's a data analysis technique used to evaluate relationships between nodes (data points). This involves identifying patterns, associations, and dependencies that might not be immediately obvious through other analytical methods. Visualizing these connections using network graphs or similar diagrams is a common component of link analysis, making it easier to understand complex relationships.

**Trend analysis** focuses on identifying patterns and changes in data over time. While useful in many situations, it doesn't primarily concern itself with the relationships between individual data points at a specific moment.

**Performance analysis** is centered on evaluating the efficiency or effectiveness of a system, process, or entity. It's not directly aimed at uncovering connections between data points, even though such connections might indirectly affect performance.

**Exploratory analysis** is a broader initial approach to understand the main characteristics of a dataset, often involving summary statistics and visualizations. While it can reveal potential relationships, it doesn't have the specific focus on interconnection that link analysis provides.

Therefore, link analysis is the best choice because it is designed precisely to uncover and visualize connections between data points, which perfectly aligns with the analyst's goal of determining whether the data being analyzed is connected to other datapoints.

For further research:

**Wikipedia - Link Analysis:**https://en.wikipedia.org/wiki/Link_analysis
**Towards Data Science - Introduction to Link Analysis:**https://towardsdatascience.com/introduction-to-link-analysis-989216396acb

## Question: 45

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:**

The answer is D, "First Name", because many data software programs struggle with variable names containing spaces. Most programming languages and data analysis tools interpret spaces as delimiters, separating distinct elements or commands. Using a space within a variable name would cause the software to misinterpret "First" and "Name" as separate entities, leading to syntax errors or unexpected behavior.

Data analysis tools and scripting languages like Python (with libraries like Pandas), R, and SQL expect variable names to adhere to specific naming conventions. While some database systems might technically allow spaces if the names are enclosed in delimiters (e.g., backticks in MySQL or square brackets in SQL Server), this is generally discouraged due to increased complexity and potential inconsistencies across different platforms.

Options A, B, and C are less problematic because they use underscores or camel case to join the words "First" and "Name." These methods create single, unbroken variable names, easily interpretable by most data software. Underscores (as in "First_Name") are common in Python and SQL, while camel case (as in "FirstName") is often favored in languages like Java and JavaScript.

For consistent and robust data analysis, it is best to adhere to strict naming conventions across all aspects of the data lifecycle. This includes data ingestion, transformation, analysis, and reporting. Using alphanumeric characters and underscores provides cross-platform compatibility and avoids interpretation ambiguities.

By contrast, spaces in variable names necessitate special handling, which can be error-prone and make code less readable and maintainable. Data professionals are typically advised to avoid spaces and adopt more universally compatible conventions, ensuring smooth data workflows and reproducible results. Thus, the space in option D introduces an unnecessary complexity that is best avoided.

For more information on data naming conventions, consult resources on programming best practices or database design guidelines.

Authoritative links:

**Python Naming Conventions:**https://pep8.org/#naming-conventions
**SQL Naming Conventions:** (Search for "SQL naming conventions best practices" on reputable database vendor websites like Microsoft SQL Server documentation or Oracle documentation).

## Question: 46

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

A.Simple random
B.Cluster

C.Systematic
D.Stratified

**Answer: D**

**Explanation:**

The correct answer is D, Stratified sampling. Here's a detailed justification:

Stratified sampling is a statistical method used to divide a population into smaller subgroups, known as strata, based on shared characteristics. These characteristics could be anything relevant to the study, such as age, gender, income level, or, in a data context, specific datasets within a larger collection exhibiting similar properties. After creating these strata, a random sample is drawn from each stratum. The key differentiator is that every stratum is represented in the final sample, ensuring that no subgroup is overlooked. This
representation reflects the proportion of each stratum in the overall population, leading to a more accurate and representative sample than other methods.

In contrast, simple random sampling selects individuals or data points completely randomly from the entire population without any subgrouping. Cluster sampling divides the population into clusters and then randomly selects entire clusters to include in the sample. Systematic sampling selects individuals at regular intervals from an ordered list of the population.

Stratified sampling's benefit lies in reducing sampling error and improving the precision of estimates. By ensuring representation from all subgroups, it minimizes bias that might arise if a particular subgroup were over- or under-represented in the sample. In a data analytics context, imagine analyzing customer behavior.

Stratified sampling could involve creating strata based on customer demographics (age, location, spending habits) and then sampling customers from each stratum. This approach helps ensure that the analysis reflects the behavior of all customer segments, not just the largest or most easily accessible ones.

The other options do not fit the scenario as described in the question. Simple random sampling doesn't consider subgroups. Cluster sampling selects entire subgroups, not random elements within each subgroup. Systematic sampling selects elements at regular intervals, without regard for subgrouping.

Therefore, stratified sampling, with its focus on random selection within predefined subgroups (strata), aligns perfectly with the question's description.

For further research, consider these resources:

**National Institute of Standards and Technology (NIST):**https://www.nist.gov/ - Search for information on statistical sampling methods.

**Statistics How To:**https://www.statisticshowto.com/ - A comprehensive resource for understanding statistical concepts.

**Khan Academy Statistics and probability:**https://www.khanacademy.org/math/statistics-probability

## Question: 47

Given the following customer and order tables:

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

| Customer_ID | Order_date | Product | Discount_code |
|---|---|---|---|
| 004 | 02/02/2020 | PAX_1 | C |
| 004 | 02/01/2020 | PAX_2 | A |
| 008 | 01/02/2020 | PAX_1 | D |
| 003 | 12/02/2020 | PAX_2 | U |
| 001 | 11/01/2020 | PAX_1 | S |
| 001 | 11/02/2020 | PAX_3 | A |
| 002 | 01/02/2020 | PAX_2 | D |

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: A**

**Explanation:**

A. Five rows, eight columns.

Inner join returns only the rows where there is a match between the two tables in the specified join condition. In this case, the join condition would be based on the "Customer ID" column, which is present in both tables. So, the inner join of the two tables would result in 5 rows where there is a match in the "Customer ID" column, and each row would have 8 columns (all columns from both tables).

**Question: 48**

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

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

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

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

**Answer: D**

**Explanation:**

The Rose unit has 1,700 square feet. To calculate the price, we need to find the average cost per square foot of the original floor plans and then multiply that by the square footage of the Rose unit.To find the average cost per square foot of the original floor plans:(650,000 + 675,000 + 700,000 + 725,000 + 775,000) / 5 = 700,000So, the average cost per square foot is $700,000 / 5 = $140.The price of the Rose unit would be $140 x 1,700 = $238,000.Therefore, the answer is D. $702,500

**Question: 49**

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:**

Data encryption is the most effective control measure for preventing data breaches among the provided options. Encryption transforms data into an unreadable format, making it unusable to unauthorized individuals even if they gain access to it. This protects sensitive information both in transit and at rest.

Data transmission, while a necessary process, doesn't inherently prevent breaches; it's a pathway where data can be intercepted if not properly secured. Data attribution refers to identifying the source or origin of data, which is useful for tracing breaches but doesn't prevent them from happening in the first place. Data retention policies dictate how long data is stored, and while important for compliance and resource management, they don't directly block unauthorized access.

Encryption algorithms like Advanced Encryption Standard (AES) are widely used and considered secure, making them a vital component of a comprehensive data security strategy. Strong encryption ensures

confidentiality, a key principle of information security, by restricting access to the data based on possession of the correct decryption key. Without the key, the data remains unintelligible, neutralizing the impact of a potential breach. Other controls such as access controls, firewalls, and intrusion detection systems work in tandem with encryption to provide layered security. However, if all those measures fail, encryption stands as the last line of defense, preventing data exposure.

For further research, refer to the National Institute of Standards and Technology (NIST) guidelines on encryption and key management:

**NIST Special Publication 800-57, Recommendation for Key Management:**
https://csrc.nist.gov/publications/detail/sp/800-57/vol-1/rev-5/final
**NIST Special Publication 800-113, Guide to SSL VPNs:** (This is a general resource on secure VPNs which frequently employ encryption.) https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-113.pdf

## Question: 50

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

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

**Answer: D**

**Explanation:**

The most effective solution to provide a dynamic and seamless experience for the user analyzing company sales data across different date ranges is option D: Create a dashboard with a date range picker and calculations built in. Here's why:

Option A, creating multiple reports, is inefficient and leads to data redundancy. The user would have to navigate and manage multiple reports, increasing the risk of errors and making comparisons cumbersome.

Option B, building calculations directly into the report, is an improvement, but it lacks flexibility. If the user wants to analyze data for a date range not predefined in the report, they would still need to perform manual calculations.

Option C, adding macros to the report, automates some processes, but macros can be complex to create and maintain. Moreover, they don't offer the interactive data exploration capabilities a dashboard provides. Also, the question asks for a seamless experience which macros cannot fully provide.

Option D provides a consolidated, interactive environment. A date range picker allows the user to dynamically select the desired timeframe for analysis. Embedding calculations within the dashboard automates the data processing based on the selected date range. This minimizes manual effort, reduces the chance of errors, and offers a smooth user experience. Dashboards are also highly visual, which can lead to better insights.

Furthermore, dashboard technologies commonly leverage cloud computing principles such as scalability, accessibility, and data integration. Cloud-based dashboards can handle large datasets, be accessed from anywhere, and connect to various data sources seamlessly. Using tools like Tableau, Power BI, or Google Data Studio allows for easy deployment and management of such dashboards, especially in a cloud environment. Data refresh capabilities can also be automated.

Essentially, a dynamic dashboard allows the user to quickly get the answers needed in order to properly analyze information, which is not the case with the other options.

For further reading on dashboard design and best practices, refer to:

**Tableau's guide to dashboard design:**https://www.tableau.com/learn/articles/data-dashboard-design-best-practices
**Microsoft Power BI documentation:**https://docs.microsoft.com/en-us/power-bi/
**Google Data Studio help:**https://support.google.com/datastudio/?hl=en

**Question: 51**

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

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

**Answer: D**

**Explanation:**

The correct answer is **D. Redundant data**. Here's why:

The scenario describes two columns in a database table storing the same information (patient height) but in different units (inches and centimeters). This signifies redundancy. Redundant data exists when the same piece of information is stored in multiple places within a database or across different databases. While there might be a reason for having both, the data is inherently duplicate in its meaning. Having both columns means that updates must be performed to both, creating a risk of discrepancy.

**A. Dependent data** refers to data where one value is determined by another value within the same table or a related table. While you can convert from inches to centimeters, having both values explicitly stored does not make them strictly dependent data in the database structure.

**B. Duplicate data** is similar to redundant data, but it often refers to identical rows or records within a table. In this context, duplicate data would mean two rows containing the exact same patient information, including height in inches and centimeters. Here, the heights are stored in separate fields.

**C. Invalid data** refers to data that violates the defined constraints or data types of a database. For example, if the height in inches was stored as a negative number, or a non-numeric value, that would represent invalid data. The problem statement doesn't indicate any violation of data types or format; therefore, it's not invalid data.

The issue isn't that either column is inherently wrong, but that the second column is unnecessary and could be derived from the first through a simple conversion. Storing both values increases the chance of inconsistency.
If one changes, the other needs to be updated to reflect the change. Failure to do so would lead to inaccuracies.

In a cloud context, data redundancy can increase storage costs in cloud databases or data lakes. It also impacts data consistency across cloud services, especially when data is replicated across different regions for high availability or disaster recovery. Maintaining redundant data requires careful planning and management to avoid data inconsistencies and increased cloud resource consumption.

Further Reading:

**Data Redundancy:**https://www.techopedia.com/definition/26265/data-redundancy
**Database Normalization:**https://www.essentialsql.com/get-started-with-database-normalization/ (discusses eliminating data redundancy)

## Question: 52

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

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

**Answer: D**

**Explanation:**

Here's a detailed justification for why removing invalid data is the best approach in this scenario, and why the other options are less suitable, when a single question has no responses in a survey dataset:

The best course of action is to treat the missing responses to a single question as invalid data, warranting their removal, rather than attempting to replace them. The absence of any response for a specific question across the entire dataset indicates a systematic problem, not random occurrence. This could stem from a flaw in the survey design (e.g., a poorly worded question, technical issue preventing submission, or required field not properly configured).

Option A, replacing missing data (imputation), introduces bias. Filling in values when no one answered could skew the results. Imputation works best when dealing with randomly missing data from a portion of responses, not a complete lack of responses for one question. Since all the responses are missing, there is no logical or statistical basis for any form of replacement. It is critical to acknowledge that imputation techniques are powerful tools, but their validity rests on assumptions about the data and the missingness mechanism. In this case, the missingness is not random.

Option B, removing duplicate data, is irrelevant because missing data, in itself, isn't a duplicate. Duplicates would be identified by having identical responses across different participants or data points.

Option C, replacing redundant data, applies when information is unnecessarily repeated. While the problem of repeated data exists, it isn't directly related to the absence of responses for one particular question.

Therefore, since all responses are missing for a single question, this data is deemed invalid and should be removed or excluded from the analysis. This approach preserves the integrity and accuracy of the analysis by not introducing potentially erroneous or biased replacement values and by focusing on analyzing only the complete and valid responses. Removing the question entirely might be required for some forms of analyses. The question's reliability should be reviewed and corrected before using it again.

Here are a few links that might be helpful to review data management practices:

**Data Quality Assessment**: https://www.usability.gov/how-to-and-tools/methods/data-quality.html
**Handling Missing Data**: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3660819/

## Question: 53

Which of the following BEST describes standard deviation?

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

**Answer: C**

**Explanation:**

The correct answer is C, "A measure of the amount of dispersion of a set of values." Standard deviation fundamentally quantifies how spread out data points are within a dataset relative to the mean (average). A low standard deviation indicates that the data points tend to be clustered closely around the mean, implying less variability. Conversely, a high standard deviation suggests that the data points are more dispersed over a wider range, signifying greater variability.

Let's examine why the other options are incorrect. Option A, "A measure that is used to establish a relationship between two variables," describes correlation or regression analysis, not standard deviation. These techniques examine how changes in one variable relate to changes in another. Option B, "A measure of how data is distributed," is too general. While standard deviation contributes to understanding data distribution, it doesn't fully describe it. Distribution involves concepts like skewness, kurtosis, and shape, not just dispersion around the mean. Option D, "A measure that is used to find the significant difference between variables," is more akin to hypothesis testing (e.g., t-tests, ANOVA) which focuses on comparing the means of different groups to determine if the observed differences are statistically significant, and takes standard deviation into account, but standard deviation itself does not find the significant difference.

In a cloud context, understanding standard deviation is crucial for capacity planning and performance monitoring. For example, tracking the standard deviation of CPU utilization across a virtual machine fleet can reveal whether workloads are consistently demanding resources or if there are spikes causing instability. A high standard deviation in network latency might indicate intermittent network congestion issues requiring investigation. Monitoring services like AWS CloudWatch, Azure Monitor, and Google Cloud Monitoring extensively use standard deviation calculations to provide insights into the stability and reliability of cloud infrastructure. Recognizing patterns of increasing standard deviation allows for proactive scaling of resources or troubleshooting performance bottlenecks before they impact end-users. Understanding variance is also key for developing robust and resilient cloud solutions. Essentially, standard deviation helps assess the consistency and predictability of various cloud metrics, aiding in informed decision-making.

Here are some authoritative links for further research:

**Investopedia - Standard Deviation:**https://www.investopedia.com/terms/s/standarddeviation.asp
**Khan Academy - Standard Deviation:**https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/variance-standard-deviation-population/a/calculating-standard-deviation-step-by-step
**NIST - Standard Deviation:**https://www.nist.gov/itl/sed/products-services/reference-information/frequently-asked-questions/what-standard-deviation

## Question: 54

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:**

B. A scatter plot would BEST represent the relationship between the variables of study hours and exam scores for each student. A scatter plot shows the relationship between two variables by plotting their values in a two-dimensional graph. Each data point in the scatter plot represents a student's study hours and exam score. The scatter plot allows the analyst to see if there is a positive or negative relationship between the variables, and if there is a linear or nonlinear relationship. It provides a visual representation of the data, making it easier to understand and interpret.

**Question: 55**

Given the table below:

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

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

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

**Answer: A**

**Explanation:**

A. Numeric.

This is because the values in the "Year" column are all numbers representing the year, and they do not contain any alphabetic characters or special characters that would classify them as alphanumeric or text.
Additionally, they are not in a date format that would require them to be classified as a date variable type.

**Question: 56**

Given the following data:

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

Which of the following BEST describes the data set?

A.There is data bias.

B.The data is incomplete.
C.The data is inconsistent.
D.The data is outliers.

**Answer: D**

**Explanation:**

D. The data is outliers.

When a dataset is best described as containing "outliers," it means that there are values in the dataset that significantly deviate from the other observations. These extreme values can indicate errors in data collection, rare events, or other anomalies that might require further investigation or special handling in your analysis.

**Question: 57**

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

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

**Answer: B**

**Explanation:**

The correct answer is **B. The data refresh date**.

Here's why: The core objective is to inform the audience about the report's monthly cadence. The data refresh

date explicitly communicates when the data used to generate the report was last updated. This immediately conveys the monthly reporting frequency because the audience knows the report's contents are based on data current up to that date. Knowing this date, they understand a new report reflecting the following month's data will become available after the next refresh.

Option A (The date of the dashboard build) is less relevant because it only indicates when the report itself was created or last modified, not when the underlying data was last updated. The dashboard could be built once and the data refreshed monthly.

Option C (A report summary) is useful for understanding the report's contents but does not inherently communicate the reporting frequency. While the summary might mention a timeframe, it's not the primary way to communicate the data refresh cycle.

Option D (Frequently asked questions) could potentially include the refresh cycle, but relying on this would be less direct and potentially overlooked.

Therefore, explicitly displaying the **data refresh date** is the most direct and effective method to convey the monthly reporting cadence, ensuring users understand the currency of the data being presented. This clarity builds trust in the report's accuracy and usefulness for decision-making within the production environment.

Consistent data refresh dates allow users to plan their analyses and compare reports across months, facilitating trend identification and performance monitoring.

For further information on data reporting best practices, consider exploring resources like:

**Data Visualization Techniques: A Guide for Decision Making:**
https://www.coursera.org/lecture/datavisualization/data-visualization-techniques-a-guide-for-decision-making-GgU7t
**Information Dashboard Design** by Stephen Few

## Question: 58

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:**

The best measure to represent typical income in a dataset with significant outliers is the median. The median is the middle value in a sorted dataset, meaning half the values are above it, and half are below it. Unlike the mean (average), the median is resistant to extreme values or outliers. Outliers can disproportionately inflate or deflate the mean, misrepresenting the "typical" value. In the context of income data, a few very high incomes can significantly increase the mean income, making it seem like the typical family earns more than they actually do.

The mode, representing the most frequent value, might not be representative of the overall data, especially if the income distribution isn't normally distributed or has multiple modes. Standard deviation measures the spread or variability of the data around the mean; it doesn't provide a measure of typical value and is also influenced by outliers. Therefore, the median is the most robust measure of central tendency when dealing

with skewed data or data containing outliers because it focuses on the central position of the data rather than the actual values, thereby minimizing the distortion caused by extreme values. Using the median provides a more accurate and stable representation of the typical income in this scenario.https://www.mathsisfun.com/data/central-measures.htmlhttps://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/mean-median-basics/a/mean-median-and-mode-review

**Question: 59**

Which of the following would be used to store unstructured data from different sources?

A.A data lake
B.A database management system
C.A database
D.A data warehouse

**Answer: A**

**Explanation:**

Here's a detailed justification for why a data lake is the appropriate choice for storing unstructured data from various sources:

The core requirement is to store unstructured data from different sources. Traditional databases (relational or even NoSQL databases in many cases) and database management systems (DBMS) are primarily designed for structured or semi-structured data. They require predefined schemas and data models. A data warehouse, while capable of handling data from different sources, is also typically geared towards structured,
transformed, and cleaned data used for analytics. Data warehouses operate on the principle of ETL (Extract, Transform, Load).

Data lakes, on the other hand, are specifically built for storing vast amounts of raw data in its native format, regardless of structure (structured, semi-structured, or unstructured). This "schema-on-read" approach allows data scientists and analysts to explore and process data without requiring upfront data modeling or transformation. Data lakes are often built using distributed storage systems like Hadoop Distributed File System (HDFS) or object storage solutions like Amazon S3, Azure Data Lake Storage, or Google Cloud Storage. These platforms can handle various data types, including text files, images, audio, video, log files, and more, from diverse sources. The data is stored as-is, and transformation and analysis are performed only when needed. A data lake is essentially a centralized repository for data from various systems, including application logs, social media streams, sensor data, and legacy databases. The ability to store data without upfront transformation is key for exploratory data analysis, machine learning model training, and data discovery. This flexibility is crucial when dealing with diverse and evolving data sources, which are common in modern data environments. This makes a data lake the ideal solution for the scenario described in the question.

Authoritative Links:

**Microsoft Azure Data Lake Storage:**https://azure.microsoft.com/en-us/products/data-lake-storage/
**Amazon S3 Data Lake:**https://aws.amazon.com/solutions/data-lake-solution/
**Google Cloud Storage:**https://cloud.google.com/storage
**Data Lake vs. Data Warehouse:**https://www.talend.com/resources/data-lake-vs-data-warehouse/

**Question: 60**

An analyst is designing a dashboard to determine which site has the highest percentage of new customers. The analyst must choose an appropriate chart to include in the dashboard. The following data is available:

| Site | Customers | New customers | Percentage of new customers |
|---|---|---|---|
| A1 | 2236 | 277 | 12% |
| A2 | 885 | 300 | 34% |
| A3 | 333 | 200 | 60% |
| B1 | 483 | 167 | 35% |
| B2 | 2969 | 235 | 8% |
| B3 | 2357 | 153 | 6% |
| C1 | 1524 | 180 | 12% |
| C2 | 878 | 150 | 17% |
| C3 | 1925 | 142 | 7% |

Which of the following types of charts should be considered to BEST display the data?

A.Include a bar chart using the site and the percentage of new customers data. B.Include a line chart using the site and the percentage of new customers data.

C.Include a pie chat using the site and percentage of new customers data. D.Include a scatter chart using the site and the percent of new customers data.

**Answer: A**

**Explanation:**

A. Include a bar chart using the site and the percentage of new customers data.

Bar charts are ideal for comparing categorical data—in this case, different sites—and displaying a corresponding metric, such as the percentage of new customers.

Line charts are typically used for trends over time, which isn't the focus here.

Pie charts work well for showing proportions of a whole, but comparing percentages across multiple categories is clearer with a bar chart.

Scatter charts are best suited for exploring relationships between two continuous variables, not for comparing a categorical breakdown.

**Question: 61**

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

A.Frequency
B.Percent change
C.Variance

D.Mean

**Answer: D**

**Explanation:**

Here's a detailed justification for why the mean is the appropriate descriptive statistic to use in this scenario:

The goal is to determine if the sugar content of cereal has changed over the years. Descriptive statistics help summarize and describe the characteristics of a dataset. In this context, we have data on the sugar content of the cereal at different points in time (years).

**Mean (Average):** The mean represents the average sugar content for each year (or a group of years). By calculating the mean sugar content for different years and comparing them, the manufacturer can directly observe whether the average sugar level has increased, decreased, or remained the same over time. This provides a clear and easily understandable measure of central tendency.

**Frequency:** Frequency counts how often a particular sugar content value appears. While useful for understanding the distribution, it doesn't directly address the question of whether the sugar content has increased over time. It tells us how many times a specific value occurs, not the overall trend.

**Percent Change:** While percent change could be used to compare the sugar content of two specific years, using the mean and comparing those means allows for a comparison across multiple periods. In addition, the mean itself becomes a data point, allowing for trending over a longer period.

**Variance:** Variance measures the spread or dispersion of the data around the mean. While useful in understanding data variability, it does not directly measure the change in the sugar content over time. High variance might indicate inconsistent sugar levels within a given year, but it doesn't tell us if the average sugar content has increased over the years.

In summary, the mean provides the most direct and interpretable way to assess the trend in sugar content over time. By comparing the mean sugar content across different years, the manufacturer can effectively determine if there's a significant increase or decrease in sugar levels.

Authoritative links for further research:

**Descriptive Statistics:**https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/ **Mean (Average):**https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/mean-median-basics/a/mean-median-and-mode-review

## Question: 62

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

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

**Answer: C**

**Explanation:**

The correct answer is **C. an exploratory data analysis.**

Exploratory Data Analysis (EDA) is a crucial initial step in data analysis. It involves using a variety of techniques, primarily statistical summaries and visual representations, to gain a deeper understanding of a dataset. The key goals of EDA are to uncover underlying patterns, detect anomalies (outliers), test hypotheses, and assess assumptions about the data. This process helps refine subsequent analytical strategies and build more accurate models. It often uses graphical visualizations like histograms, scatter plots, box plots, and other techniques to reveal data distribution and relationships.

Option A, a t-test, is a specific statistical test used to compare the means of two groups. While it can be part of data analysis, it's a more targeted test applied after initial exploration to test specific hypotheses. Option B, a performance analysis, is focused on evaluating the efficiency and effectiveness of a system or process, not necessarily about understanding the inherent properties of the data itself. Option D, a link analysis, is used to discover relationships between entities within a network, rather than exploring the fundamental characteristics of the data values themselves. EDA is a much broader, more initial approach that lays the groundwork for more specific analyses like t-tests or link analysis. It uses statistical insight and visualization to understand the underlying data structure.

In a cloud computing context, where massive datasets are often stored and processed, EDA becomes exceptionally important. Before deploying complex machine learning models or drawing conclusions, understanding data quality, distribution, and potential biases is essential. Cloud platforms provide tools and services that make EDA scalable and efficient. For example, platforms like AWS, Azure, and Google Cloud offer managed services for data exploration and visualization that enable analysts to process large datasets effectively. Without EDA, potentially flawed data could be used in algorithms, leading to incorrect decisions.

Further Reading:

**NIST Engineering Statistics Handbook - Exploratory Data Analysis:**
https://www.itl.nist.gov/div898/handbook/eda/eda.htm
**Wikipedia - Exploratory Data Analysis:**https://en.wikipedia.org/wiki/Exploratory_data_analysis

## Question: 63

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:**

Here's a detailed justification for why options A (Data accuracy) and E (Data consistency) are the most likely issues to arise when different people manually type handwritten surveys into an online database:

**Data Accuracy:** Human error is inherent in manual data entry. Different individuals may misinterpret handwriting, leading to inaccuracies in the data. A number 7 could be mistaken for a 1 or a 9, or a 't' might be interpreted as an 'f'. This compromises the accuracy of the dataset, making it unreliable for analysis and decision-making. Poor handwriting and subjective interpretations contribute significantly to this problem.

**Data Consistency:** When multiple individuals are involved in data entry, they might interpret instructions or

formatting rules differently. One person might consistently abbreviate a certain field, while another spells it out in full. This results in inconsistency across the database, making it difficult to perform reliable searches, comparisons, and aggregations. The lack of standardized data entry protocols increases the risk of this issue occurring. Different interpretations of qualitative responses are also a source of inconsistency.

**Why other options are less likely (though not impossible):**

**B. Data constraints:** Data constraints are rules applied to data entry to ensure validity (e.g., a date field only accepts dates). While manually entered data can violate constraints, the constraints themselves are a property of the database, not a direct result of manual entry by different people.

**C. Data attribute limitations:** Attribute limitations refer to the predefined characteristics (name, data type, size) that are possible fields in a dataset. This is also less of a direct problem that is specific to multiple people manually entering data.

**D. Data bias:** While manual data entry can introduce bias (if, for example, data entry staff make assumptions based on the survey takers' demographics visible in the handwritten forms), it's less inherent than accuracy and consistency problems. Bias is more likely to arise in the survey design or analysis phases.

**F. Data manipulation:** Data manipulation implies intentional alteration. While theoretically possible with manual entry, it's less likely than unintentional errors (accuracy) and variations (consistency).

**In Summary:**

The key challenge with manual data entry by multiple people stems from the risk of human error (affecting accuracy) and differing interpretations and practices (affecting consistency). These are direct consequences of the manual process and the involvement of multiple individuals.

Relevant concepts: Data quality, data integrity, data governance, data validation.

For further reading on data quality and challenges in manual data entry:

**Data Quality Assessment** (MIT): https://dspace.mit.edu/handle/1721.1/4811
**Data Integrity** (Wikipedia): https://en.wikipedia.org/wiki/Data_integrity

## Question: 64

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

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

**Answer: D**

**Explanation:**

The correct answer is D, Stratified sampling. Here's a detailed justification:

Stratified sampling is a statistical technique where a population is divided into homogeneous subgroups, called strata, based on shared characteristics. These characteristics could be age, gender, income level, education, or any other relevant attribute. Once the population is divided into these strata, a random sample is then taken from each stratum, usually proportionally to the stratum's size within the overall population.

The primary benefit of stratified sampling is that it ensures representation from all relevant subgroups within the population. This leads to more accurate and reliable inferences about the population as a whole compared

to other sampling methods. For example, if you want to understand customer satisfaction with a new cloud service, you might stratify your customer base by their level of cloud usage (low, medium, high) to ensure you get feedback from users across all these usage levels.

Let's look at why the other options are incorrect:

**A. Systematic sampling:** This method selects samples at regular intervals (e.g., every 10th customer). It doesn't guarantee representation from all subgroups.

**B. Simple random sampling:** Every member of the population has an equal chance of being selected. While fair, it may not adequately represent smaller subgroups, potentially leading to skewed results.

**C. Convenience sampling:** This method selects samples based on ease of access. It's the least reliable as it's prone to bias and doesn't represent the population accurately. For instance, only surveying people who attend a specific cloud conference would be convenience sampling and would not be representative of all cloud users.

In the context of data analysis related to cloud computing, stratified sampling can be invaluable. Imagine analyzing the performance of a cloud application across different geographic regions. Stratifying your data by region would allow you to identify performance bottlenecks specific to certain areas, which a simple random sample might miss. Similarly, when analyzing the security vulnerabilities reported across a large cloud infrastructure, stratifying by the type of service (compute, storage, network) allows for targeted remediation efforts. Stratified sampling ensures that your analysis reflects the nuances of the population, leading to more meaningful and actionable insights.

Authoritative Links:

**Investopedia - Stratified Random Sampling:**
https://www.investopedia.com/terms/s/stratified_random_sampling.asp
**Statistics How To - Stratified Random Sample:**https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/stratified-random-sample/

## Question: 65

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

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

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

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

**Answer: C**

**Explanation:**

C. Parsing involves breaking down a single column that contains combined or composite data into multiple columns based on specific delimiters or patterns. In this case, the analyst would parse the name column to extract each component of the name (e.g., first name, last name) into separate columns. This allows for more

granular analysis and manipulation of the data based on its individual components.

**Question: 66**

Which of the following descriptive statistical methods are measures of central tendency? (Choose two.)

A.Mean
B.Minimum
C.Mode
D.Variance
E.Correlation
F.Maximum

**Answer: AC**

**Explanation:**

The correct answer identifies two key descriptive statistical methods that quantify the central tendency of a dataset: the mean and the mode. Central tendency measures aim to pinpoint a typical or representative value within a distribution.

A. **Mean:** The mean, often referred to as the average, is calculated by summing all values in a dataset and dividing by the total number of values. It provides a balanced representation of the dataset's center, sensitive to the magnitude of each data point. For instance, if analyzing the response times of a cloud application, the mean response time gives an overall indication of the application's typical performance. A high mean could signal performance bottlenecks requiring investigation. https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/mean-median-basics/a/mean-median-and-mode-review

C. **Mode:** The mode represents the value that appears most frequently within a dataset. It's particularly useful for identifying the most common occurrence or category. In cloud monitoring, if analyzing error logs, the mode might indicate the most frequent type of error, helping prioritize troubleshooting efforts. The mode is less sensitive to outliers than the mean. https://www.investopedia.com/terms/m/mode.asp

Why other options are incorrect:

B. **Minimum:** The minimum value represents the smallest value in the dataset. It describes the lower limit of the data range, not its center.

D. **Variance:** Variance measures the spread or dispersion of data points around the mean. It indicates how much the data deviates from the central tendency, not the central tendency itself. It is a measure of variability.

E. **Correlation:** Correlation describes the statistical relationship between two variables. While valuable for understanding dependencies, it doesn't provide information about the central tendency of either variable.

F. **Maximum:** The maximum value is the largest value in the dataset, indicating the upper limit of the data range, not its center.

In summary, mean and mode are fundamental measures of central tendency, providing insights into the typical values within a dataset, which is crucial for various data analysis tasks, including performance monitoring, anomaly detection, and capacity planning in cloud environments.