Interactive activity _NOTUPLOADED

What is the primary goal of Exploratory Data Analysis (EDA)?

The primary goal of EDA is to gain an initial understanding of the data, summarize its main characteristics, and 'listen' to what the data has to tell us without imposing preconceived notions.

How does understanding the data or EDA help in improving the outcome of a data analysis or data science project?

EDA leads to a more robust, accurate, and insightful understanding of the data, which in turn enhances the quality and effectiveness of the subsequent analyses and decision-making processes.

What are some common outputs or results of conducting EDA on a dataset?

Common outputs include a better understanding of the data's structure, a refined list of potential variables for modeling, identified data quality issues, insights into the dataset, and further questions or hypotheses for investigation.

How can EDA facilitate the communication of findings to stakeholders without a technical background?

EDA facilitates communication through the use of visualizations and summary statistics that make the findings accessible and understandable to stakeholders who may not have a technical background.

What are the key tasks involved in the Data Profiling step of EDA?

Key tasks in Data Profiling include checking data types, ranges of values, unique counts, presence of null values, and overall data quality checks.

What steps might be taken during the Data Cleaning phase as a result of findings from data profiling?

Steps in the Data Cleaning phase can include handling missing values, correcting errors, and dealing with outliers based on the preliminary findings from data profiling.

How do Univariate and Bivariate/Multivariate Analysis differ in their approach and purpose within EDA?

Univariate Analysis focuses on examining single variables to understand their distribution, central tendency, and dispersion. Bivariate/Multivariate Analysis, on the other hand, examines relationships between two or more variables to identify correlations, patterns, and trends that suggest relationships or associations.

What are the two main types of data?

The two main types of data mentioned are quantitative (numerical) data and qualitative (categorical) data.

What is another term used for quantitative data as mentioned in the paragraph?

Another term used for quantitative data is numerical data.

What is another term used for qualitative data according to the paragraph?

Another term used for qualitative data is categorical data.

What are the 7 attributes of data that data quality checks aim to ensure?

The five attributes are accuracy, completeness, consistency, relevance, and reliability.

How is the concept of 'Timeliness' important in data quality checks?

Timeliness is important as it ensures that the data is up-to-date and reflects the current situation or the relevant period of interest for the analysis, which is crucial for its applicability and usefulness in decision-making processes.

What does evaluating the reliability of data involve according to the text?

Evaluating the reliability of data involves assessing the trustworthiness of the data's source and the process used to collect the data.

What are the two stages where missing values might occur in data?

Missing values may occur during the data extraction and data collection stages.

What is typically easier to correct, errors from data extraction or data collection?

Errors at the data extraction stage are typically easier to find and correct.

Why is it crucial to handle missing data?

Handling missing data is crucial to avoid biasing the conclusions and leading the business to make wrong decisions.

What are two common methods mentioned for treating missing values?

The two methods mentioned are deletion and imputation.

What is the major advantage of the deletion method for handling missing values?

The major advantage of the deletion method is its simplicity.

What is a significant drawback of using the deletion method to handle missing values?

A significant drawback is that it reduces the power of the model because it decreases the sample size.

What does the imputation method involve?

The imputation method involves filling in the missing values with estimated ones, making it one of the most frequently used methods.

What specific types of imputation are mentioned in the text?

The types of imputation mentioned are mean, mode, and median imputation.

How does mean/mode/median imputation work?

This method consists of replacing missing data for a given attribute with the mean or median (for quantitative attributes) or mode (for qualitative attributes) of all known values for that attribute.

Why might data collection errors be harder to correct than extraction errors?

Data collection errors are harder to correct because they occur during the gathering of data, often from external sources or processes that might not be easily adjusted or replicated after the fact.