

**Exploratory Data Analysis:**

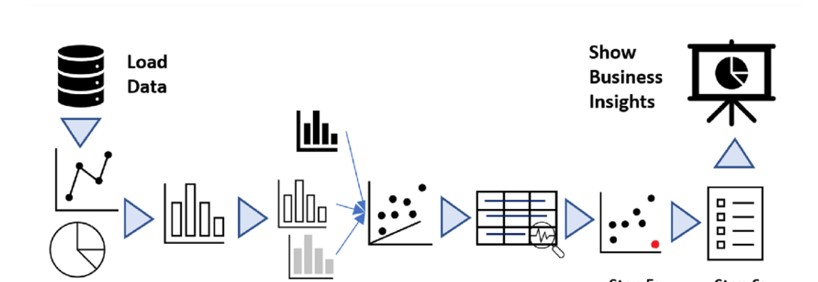
As a data analyst in a company, Exploratory Data Analysis (EDA) is one of the core tasks that Data analyst would typically perform. EDA is the process of exploring and analyzing data to gain insights and identify patterns, relationships, and anomalies in the data.

Exploratory data analysis (EDA) is an approach to analyzing and understanding data through visualizations and summary statistics. The goal of EDA is to gain insights into the data, identify patterns, and test hypotheses that can guide further analysis.

EDA involves several steps:

1. Data sourcing: Data can come from various sources, including databases, spreadsheets, text files, or web scraping tools. The data should be organized into a format that is easy to work with, such as a table or a data frame.
2. Data cleaning: Data cleaning involves identifying and correcting errors, missing values, and outliers. This step is critical to ensure that the analysis is based on accurate data.
3. Data visualization: Data visualization is a powerful tool for exploring patterns and relationships in the data. Graphs and charts can reveal trends, outliers, and correlations that may not be apparent from summary statistics alone.
4. Summary statistics: Summary statistics such as mean, median, standard deviation, and range provide an overview of the data distribution. They can help identify unusual values or patterns in the data.
5. Hypothesis testing: Hypothesis testing is used to determine whether a pattern or relationship in the data is statistically significant. This involves defining a null hypothesis (no relationship between variables) and an alternative hypothesis (a relationship exists), and then using statistical tests to determine the likelihood of each hypothesis.
6. Iterative process: EDA is an iterative process that involves refining hypotheses, exploring new variables, and testing different visualizations and statistical methods to gain insights into the data.

This data can be numbers, text, images, or even sounds. But for it to be a dataset, it usually needs to be organized in a specific way.

Ensuring clean data is key for reliable analysis. Data cleaning tackles two frequent issues: duplicates (data points appearing multiple times) and missing values (empty data points). Duplicates can be removed entirely or merged if slightly different. Missing values require strategies like deletion (if minimal), imputation (estimating values), or adding a new category (for informative missing reasons).

Data grouping by candidate name is a technique used to organize information based on individual candidates. It essentially sorts data entries where the "candidate name" field acts as the category. This allows you to analyze and compare information specific to each candidate.