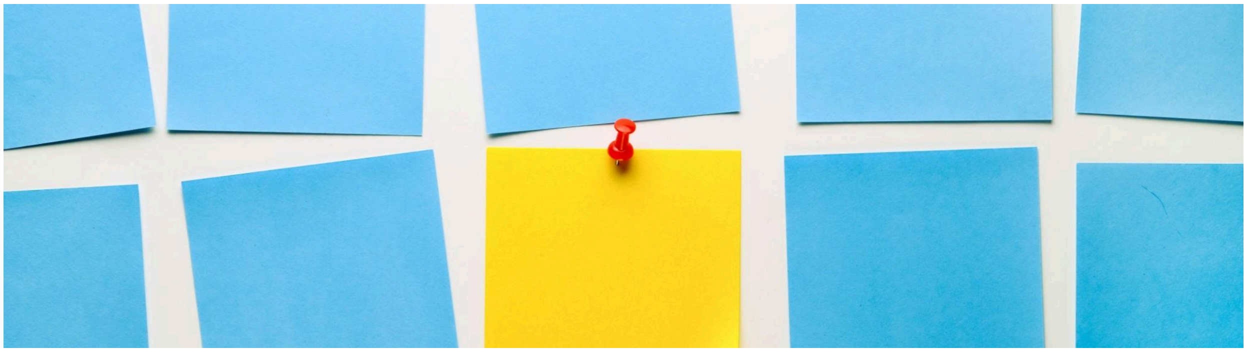

Exploratory Data Analysis

Goal

Get an initial overview of the dataset.

Techniques

Descriptive Statistics & Summary Measures

- **Task:** Calculate mean, median, standard deviation, quartiles, and range for numerical variables.
- **Purpose:** Understand the central tendencies and dispersion of the data, which helps in identifying the overall distribution and potential anomalies.

Missing Value Analysis

- **Task:** Identify missing values, analyze their patterns (random or systematic), and calculate the percentage of missing entries for each column.
- **Purpose:** Determine the impact of missing data on analysis and decide on strategies for imputation or exclusion.

Univariate Distribution Visualization

- **Task:** Create histograms, box plots, or density plots for individual variables.
- **Purpose:** Visualize the distribution of each feature, detect skewness, kurtosis, and outliers, and assess whether data transformations might be needed.

Correlation and Relationship Analysis

- **Task:** Compute and visualize the correlation matrix using heatmaps, scatter plots, or pair plots.
- **Purpose:** Identify relationships between features, detect multicollinearity, and discover potential predictors for modeling tasks.

Outlier Detection and Analysis

- **Task:** Use methods like box plots, Z-scores, or the IQR method to detect outliers in numerical features.
- **Purpose:** Investigate data points that deviate significantly from the norm, which might indicate data entry errors, anomalies, or interesting cases that warrant further investigation.

Frequency and Count Analysis

- **Task:** Calculate the frequency (counts and percentages) for each category in the feature.
- **Purpose:** Identify the most common and least common categories, understand the balance of categories, and detect any rare classes that might need special handling.

Visualization with Bar Charts and Pie Charts

- **Task:** Create bar charts or pie charts to visualize the distribution of categorical features.

- **Purpose:** Provide an intuitive visual overview of the data distribution, making it easier to spot imbalances or dominant categories.

Missing Value Analysis in Categorical Data

- **Task:** Check for missing or undefined entries in categorical features and determine the proportion of missing values.
- **Purpose:** Assess data quality and decide on an appropriate strategy (e.g., imputation, removal, or creating a separate "Unknown" category).

Cross-tabulation and Contingency Tables

- **Task:** Generate cross-tabulations (contingency tables) between categorical features or between a categorical feature and the target variable.
- **Purpose:** Explore relationships, dependencies, or interactions between categories, which can be particularly useful for feature engineering or understanding class associations.

Analyzing High Cardinality and Grouping Rare Categories

- **Task:** Identify categorical features with a large number of unique values (high cardinality) and decide if some categories should be grouped together (e.g., grouping low-frequency categories into an "Other" category).
- **Purpose:** Simplify the feature for modeling purposes and reduce noise, while preserving significant categorical patterns.

Methodologies

Explore these different strategies on the dataset, and finally pick 5 interesting observations you would like to dig deeper into. Present these in the class.