Exploratory Data Analysis (EDA)

1. **Data Preprocessing**
   1. Import datasets
   2. Check data types for all variables
   3. Check for missing/null/NaN values
      1. Dealing with missing values
         1. Categorical missing values
            1. Drop missing categorical data if < 10% of the overall dataset
            2. Bucket (create new category for all missing categorical data)
         2. Numerical missing values
            1. Drop missing numerical data if < 10% of the overall dataset
            2. Replace missing values by mean, median, mode of the dataset
            3. Imputation
2. **Descriptive Statistics**
   1. Distributions of all variables
      1. Numerical variables - check for normal distribution
         1. Histograms/density/boxplots
      2. Categorical variables
         1. Bar plots
         2. tables
      3. Deal with outliers
         1. qq-plots
         2. Remove an outlying observation if its influence is undue
         3. Keep an outlier if it’s a high leverage point
   2. Summary Statistics
      1. Tables
      2. Plots
3. **High Dimensionality Reduction**
   1. Correlation check
      1. Plots
      2. Tables
      3. Remove highly correlated variables
   2. Variable Interaction
      1. Understand relationships between variables
   3. Principal Component Analysis
      1. PCA - If dataset still contains extremely high number of predictors
         1. Pick the first principal components that explain at least 90% of the variance for model building