Exploratory Data Analysis

Part-1: EDA: Univariate and Bivariate

- **Taxonomy of variable types:**
  - continuous vs. discrete
  - categorical: binary vs. polytomous
  - nominal, ordinal, interval, ratio
- **Visual statistical description: Overview**
  - Univariate: central tendency and variability
  - Bivariate: correlation, covariance, association, by-group analysis
  - Multivariate: conditioning, facets
  - Distributions, hypothesis testing about the means
  - Visual comprehension: box plots, violin plots, density plots, histograms, bar charts, pie charts, contour plots, hexagonal binning, parallel coordinate plots, cluster dendrogram
- **Univariate and Bivariate EDA**
  - *Central tendency* (location measures):
    - Means: arithmetic/harmonic/geometric; trimmed, weighted
    - Medians: median, weighted median
    - Mode, expected value for categorical variables
  - *Variability* (spread): variance, standard deviation, mean absolute deviation, range, order statistics (ranks), percentile (quantile), interquartile range (IQR)
  - *Correlation and covariance*
  - Group-by analysis
  - *Distributions:* normal, bi-modal,
  - Statistical tests to compare central tendency and variability
- **Proximity metrics:**
  - Similarity vs. dissimilarity vs. distance
  - Similarity for binary data
  - Similarity for continuous data
  - Distance for continuous data
  - Distance for categorical data
  - Proximity for mixed attribute types
  - Entropy measures

Part-2: EDA for Data Preparation, Validation and Cleaning

- **Data characteristics:**
  - dimensionality, sparcity, resolution
  - missingness
  - noise, outliers
  - inconsistencies and duplicates
- **Sampling and data reduction**
  - Sample vs. population
  - Sample statistic vs. population parameters
  - Resampling: bootstrap sample vs. permutation

        - Sampling strategies: Simple Random Sample, Stratified Random Sample, Cluster Sample, Systematic Sample
        - Sample characteristics and design
- **Linear Transformations:**
    - Centering
    - Standardizing
    - Z-scores
    - Normalizing
    - Rescaling
- **Non-linear Transformations:**
    - Ladder of Roots of Power
    - Log transformation
    - Box-Cox transformation
    - Rank transformation
    - Change the shape of distribution
    - Transforming for linearity, constant spread, and normality
    - Logit & probit for skewed proportions
- **Dimensionality reduction vs. feature selection**
    - Principal Component Analysis (PCA)
    - Univariate feature selection (e.g., t-test)

Part-3: Hypothesis Testing

- **A/B Testing**
    - Treatment and control
    - Randomization
    - Test statistic
    - Null and Alternative Hypothesis:
        - One-tailed and Two-Tailed Test
        - p-value, alpha
    - Bootstrap sampling:
        - Statistical Simulation
        - Independent and paired samples
    - Comparison Tests:
        - Comparing means (t-test)
        - Comparing proportions (Z-test, chi-squared test, McNamir test)
- **Distributions and Confidence Intervals**
    - Descriptive vs. Inferential Statistics
    - Confidence interval: Bootstrap Procedure
    - Statistical Distributions for different types of variables
        - Continuous (Gaussian), binary (binomial), counts (Poisson)
    - Statistical Distributions and tests:
        - Normal, t-, F-, χ^2-distributions and tests

- **Design and Analysis of Statistical Tests**
    - Effective size
    - Power
    - Sample size

    - Significance level, alpha
    - Type I and II errors
    - Sampling procedure and its effect on test interpretability
- **Multiple Testing**
    - Adjusted p-values
    - False Discovery Rate (FDR)
    - Overfitting and mitigation strategies
        - Bonferroni adjustment.
        - Cross-validation