Home Task SVM

Coding Exercise 1

Short Answers

## 1. Data Quality

## - Common data quality problems

## What common data problems can be observed in the dataset

## (Answer in 5 Lines)

1. Missing values (? or NaN) observed in key features.
2. Duplicate columns or features in the dataset.
3. Presence of irrelevant or redundant variables.
4. Outliers that may distort analysis or model performance.
5. Skewed distributions or imbalanced target variable.

## - Exploratory data analysis (EDA)

## What are the common observations in the dataset

## (Answer in 5 Lines)

1. Data is numerical with several highly correlated features.
2. Target variable is imbalanced, creating potential bias.
3. Features have varying scales, requiring normalization.
4. Presence of extreme values in some features.
5. Some features lack sufficient variability, limiting their impact.

## - Anomaly detection

## What Anomalies can be seen in the dataset

## (Answer in 5 Lines)

1. Outliers detected in features like population, householdsize.
2. Z-scores highlight anomalies beyond 3 standard deviations.
3. Target variable extremes suggest potential data recording errors.
4. Unusual correlations between unrelated variables.
5. Rare combinations of feature values stand out.

## - Summary statistics

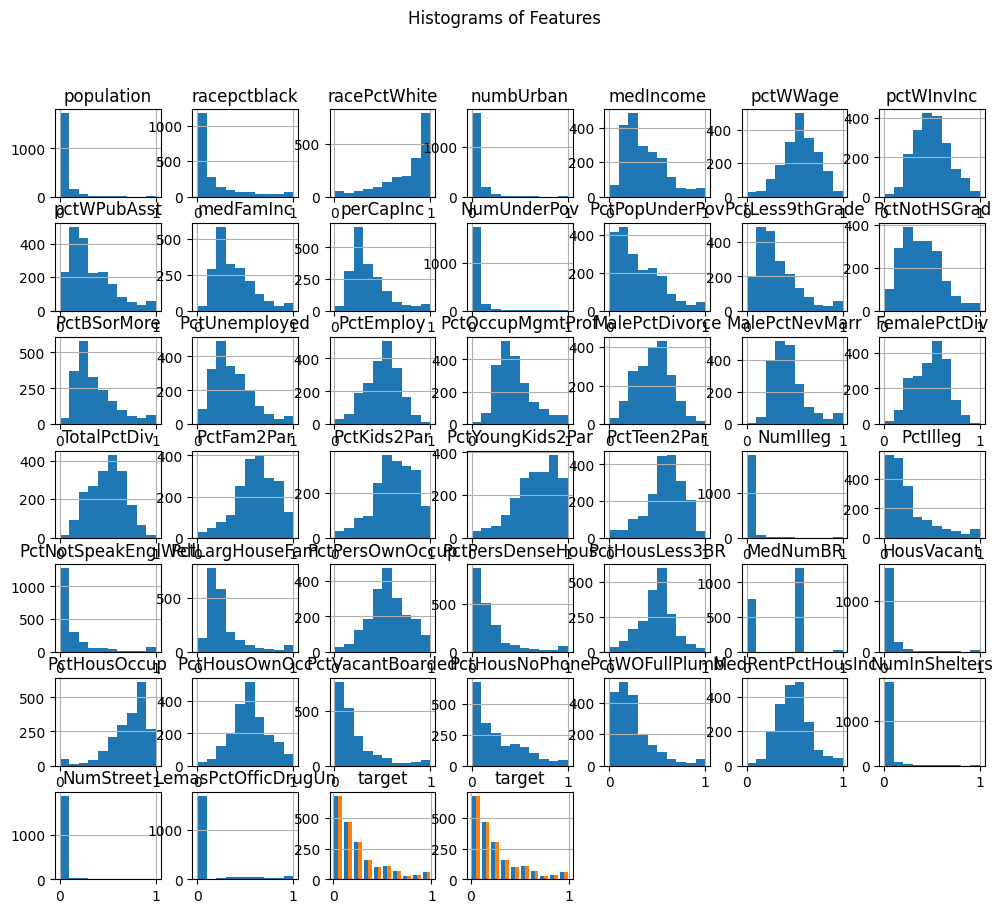
## Run code and explain the summary in 5 lines

1. Mean and median differ for several features, indicating skewness.
2. High standard deviation suggests significant variability.
3. Minimum and maximum values highlight potential outliers.
4. Many features are normalized, with values between 0 and 1.
5. Some features show limited ranges, indicating low variability.

## 2. Data Visualization

Run the code to see the below visualizations of the dataset’s Features:

- Histograms



- Scatter plots

A group of purple and white dots

Description automatically generated

- Contour plots

A purple square with green and yellow dots

Description automatically generated

- Matrix plots

A close-up of a data visualization

Description automatically generated

Now explain the observations and two pro and two cons of each of the above visualization methods in respect to the given dataset

* **Histograms**:
  + **Observations**: Distribution of features; skewness in some variables.
  + **Pros**:
    - Simple to interpret
    - Highlights skewness.
  + **Cons**:
    - Inefficient for high-dimensional data
    - Misses correlations.
* **Scatter Plots**
  + **Observations**: Relationships between features and target; clusters.
  + **Pros**:
    - Clear visualization of relationships
    - Identifies outliers.
  + **Cons**:
    - Limited to low-dimensional data
    - Cluttered with many points.
* **Contour Plots**:
  + **Observations**: Density regions for target and population.
  + **Pros**:
    - Good for 2D relationships
    - Identifies clusters.
  + **Cons**:
    - Hard to interpret for complex data
    - Prone to overlap issues.
* **Matrix Plots**:
  + **Observations**: Correlation heatmap of features.
  + **Pros**:
    - Identifies correlations
    - Effective for feature selection.
  + **Cons**:
    - Lacks direct interpretability
    - Poor with categorical data.