# Project Report: Bank Loan Case Study

## Project Description

The Bank Loan Case Study aimed to provide a comprehensive analysis of a dataset containing information about loan applications. The project's primary objectives were to identify missing data, handle it appropriately, identify outliers, analyze data imbalance, and perform various exploratory data analysis (EDA) techniques, including univariate, segmented univariate, and bivariate analyses.

## Approach

### Data Loading and Exploration

The project began by loading three datasets: `application\_data`, `previous\_application`, and `columns\_description`. The datasets were examined to understand their structure, features, and contents.

### Task A: Identify Missing Data and Deal with it Appropriately

1. \*\*Identifying Missing Data:\*\*

- Utilized Pandas to identify missing values in the `application\_data` dataset.

- Generated a summary of missing values for each column.

2. \*\*Visualizing Missing Data:\*\*

- Created a bar plot using Matplotlib and Seaborn to visualize the proportion of missing values in the dataset.

3. \*\*Handling Missing Data:\*\*

- Imputed missing numerical values with the mean to preserve data integrity.

- Imputed missing categorical values with the mode to maintain the distribution of categorical variables.

### Task B: Identify Outliers in the Dataset

- Leveraged box plots to identify outliers in numerical columns of `application\_data`.

- Examined and visualized potential outliers for further investigation.

### Task C: Analyze Data Imbalance

- Utilized Matplotlib and Seaborn to visualize the distribution of the target variable (`TARGET`) using both a pie chart and a bar chart.

- Recognized the imbalance in the dataset, where one class significantly outnumbered the other.

### Task D: Perform Univariate, Segmented Univariate, and Bivariate Analysis

#### 1. Univariate Analysis:

- \*\*Histograms for Numerical Variables:\*\*

- Selected numerical columns and plotted histograms to visualize the distribution of each variable.

- \*\*Bar Charts for Categorical Variables:\*\*

- Selected categorical columns and created count plots to represent the distribution of each category.

#### 2. Segmented Univariate Analysis:

- \*\*Grouped Bar Charts for Categorical Variables Across Scenarios:\*\*

- Employed count plots to analyze how categorical variables varied across different scenarios.

#### 3. Bivariate Analysis:

- \*\*Scatter Plots for Numerical-Target Variable Relationships:\*\*

- Created scatter plots to explore the relationship between numerical variables and the target variable (`TARGET`).

- \*\*Box Plots for Categorical Variables:\*\*

- Utilized box plots to analyze the distribution of income based on education and target.

### Task E: Identify Top Correlations for Different Scenarios

- Developed a function (`analyze\_correlations`) to analyze and visualize correlation matrices for different segments based on the `TARGET` variable.

- Visualized top correlated features with the target variable using heatmaps.

## Tech-Stack Used

- Python 3.11

- Libraries: Pandas, Matplotlib, Seaborn, NumPy

## Insights

1. Handling Missing Data:

- Successfully addressed missing data, resulting in a dataset with minimal missing values.

- Ensured data completeness for subsequent analyses.

2. Outlier Identification:

- Identified potential outliers, providing insights into data quality issues and potential data cleansing needs.

3. Data Imbalance:

- Recognized a significant imbalance in the target variable, which may impact model performance.

4. Exploratory Data Analysis:

- Conducted a thorough EDA, revealing patterns and trends in the data.

- Utilized various visualization techniques for effective data exploration.

5. Correlation Analysis:

- Examined correlations between numerical variables in different scenarios.

- Identified features strongly correlated with the target variable (`TARGET`).

## Result

The project successfully achieved its objectives, providing a comprehensive analysis of the bank loan dataset. By addressing missing data, identifying outliers, and conducting thorough exploratory data analysis, the team gained valuable insights into the dataset's characteristics. The correlation analysis helped identify features strongly correlated with the target variable, laying the foundation for future predictive modeling.

## githuhb link  
https://github.com/mrunali25/ML/blob/main/bank\_loan\_case\_study