**Name: Arpan Paul**

**GitHub Link: https://github.com/Hrick91/Bank-Loan-Case-Study**

**Technical Report: Data Analysis of Loan Application Dataset**

**Introduction**

This project focuses on analysing a loan application dataset to gain insights into various factors affecting loan approval and defaults. The dataset includes a range of attributes related to the applicant's financials, personal details, and loan characteristics. We aim to address key data analysis tasks such as handling missing data, detecting outliers, understanding data imbalance, performing univariate and bivariate analysis, and identifying correlations between key variables and the target variable, which indicates whether the loan application was approved (TARGET).

The analysis is conducted using Excel 2019, which provides robust functionality to process, clean, and visualize data. Below is a breakdown of the methodologies used to address each task, followed by the results obtained through various analyses.

**Task A: Identifying and Handling Missing Data**

Missing data is a common issue in datasets and needs to be addressed before performing further analysis. In the loan application dataset, we used Excel functions such as COUNT, ISBLANK, and IF to identify missing values across various columns.

**Steps:**

1. **Identifying Missing Data:**
   * We used the COUNTIF function to count the number of empty cells in each column.
   * The formula =COUNTIF(range, "") was applied to each column to determine the number of missing values.
2. **Dealing with Missing Data:**
   * For columns with numerical data such as AMT\_INCOME\_TOTAL, we replaced missing values with the **median** using the MEDIAN function, ensuring that the imputation did not skew the distribution of values.
   * For categorical variables like NAME\_EDUCATION\_TYPE, we replaced missing values with the most frequent value (mode) using the MODE function.
3. **Visualization:**
   * A bar chart was created to visualize the proportion of missing values for each variable. This helped in identifying the columns with significant missing data that required imputation.

**Task B: Identifying Outliers**

Outliers can significantly influence the results of statistical analysis. In this task, we used the **Interquartile Range (IQR)** method to detect outliers in numerical columns such as AMT\_INCOME\_TOTAL, AMT\_CREDIT, and AMT\_GOODS\_PRICE.

**Steps:**

1. **Calculating Quartiles:**
   * We used Excel's QUARTILE.EXC function to calculate the first and third quartiles (Q1 and Q3) for each numerical column.
2. **Calculating IQR:**
   * The IQR was calculated by subtracting Q1 from Q3: IQR = Q3 - Q1.
3. **Identifying Outliers:**
   * Any value greater than Q3 + 1.5 \* IQR or less than Q1 - 1.5 \* IQR was flagged as an outlier.
4. **Visualization:**
   * A **box plot** was used to visualize the distribution of numerical variables, highlighting potential outliers for further investigation.

**Task C: Analyzing Data Imbalance**

Data imbalance in binary classification problems, such as predicting loan defaults, can lead to biased results. To assess data imbalance, we analysed the distribution of the TARGET variable.

**Steps:**

1. **Calculating Class Proportions:**
   * We used the COUNTIF function to count the number of applicants with a loan default (TARGET = 1) and those without (TARGET = 0).
   * The class imbalance ratio was calculated as:  
     Imbalance Ratio = (Count of TARGET = 1) / (Count of TARGET = 0).
2. **Visualization:**
   * A **bar chart** was created to visualize the distribution of the target variable. The chart clearly highlighted the imbalance between the two classes, which is crucial for selecting appropriate machine learning models.

**Task D: Univariate, Segmented Univariate, and Bivariate Analysis**

In this task, we explored various aspects of the dataset through **univariate** and **bivariate** analysis to understand the relationship between individual variables and loan default status.

**Univariate Analysis:**

1. **Descriptive Statistics:**
   * We used Excel's AVERAGE, MEDIAN, and COUNTIF functions to summarize the distributions of numerical variables such as AMT\_INCOME\_TOTAL and AMT\_CREDIT.
   * The goal was to understand the central tendency and spread of each variable in isolation.
2. **Visualization:**
   * **Histograms** were used to visualize the distribution of income, credit amount, and other key financial variables.

**Segmented Univariate Analysis:**

1. **Comparing Groups:**
   * We segmented the data based on categories like NAME\_EDUCATION\_TYPE and compared the distributions of numerical variables (e.g., AMT\_INCOME\_TOTAL) across different education levels using pivot tables.
2. **Visualization:**
   * **Bar charts** were created to compare the average values of income and credit across education levels, allowing us to identify patterns or trends that might be indicative of loan approval likelihood.

**Bivariate Analysis:**

1. **Correlation Analysis:**
   * We calculated the correlation coefficients between the target variable TARGET and numerical variables such as AMT\_INCOME\_TOTAL, AMT\_CREDIT, and AMT\_ANNUITY using the CORREL function.
2. **Visualization:**
   * **Scatter plots** were used to visualize the relationships between key numerical variables and the target variable. These plots provided insights into how changes in income or credit amount might influence the probability of loan approval.

**Task E: Identifying Top Correlations for Different Scenarios**

Identifying the top correlations between variables and the target variable can provide valuable insights into the factors that most influence loan approval. In this task, we segmented the data based on different scenarios and identified the top correlations within each segment.

**Steps:**

1. **Calculating Correlations:**
   * For each segment (e.g., TARGET = 1 vs. TARGET = 0), we calculated the correlation coefficients between key variables (such as AMT\_INCOME\_TOTAL, AMT\_CREDIT, and DAYS\_BIRTH) and the target variable using the CORREL function.
2. **Visualization:**
   * A **heatmap** was created to visualize the correlation matrix, with colour intensity indicating the strength of the correlation. The correlations with the highest values were highlighted to identify the top indicators of loan default.

**Task F: Creating a Treemap to Visualize Correlation**

To create a **treemap** visualizing the correlation values between the TARGET variable and key features such as AMT\_INCOME\_TOTAL, AMT\_CREDIT, and DAYS\_BIRTH, we followed the steps below:

1. **Preparing the Data:**
   * We created a table with the correlation values between the TARGET and other variables.
2. **Creating the Treemap:**
   * We selected the table and inserted a **treemap** chart via the **Insert** tab. This allowed us to visually compare the relative importance of different features based on their correlation with the target variable.

**Conclusion**

This report outlines the detailed process of analysing a loan application dataset using Excel 2019. By effectively handling missing data, detecting outliers, understanding data imbalance, and conducting univariate and bivariate analysis, we were able to gain valuable insights into the factors affecting loan approvals. Identifying key correlations and visualizing these relationships provided a deeper understanding of the drivers behind loan default, which could inform further predictive modelling and decision-making processes.