**Name: Nikhil Sanjay Shinde**

**Roll No: TYITB120**

**Assignment No. 1**

**Aim:**

Perform Exploratory Data Analysis (EDA) on the loan dataset and draw insights with the help of Python.

**Objective:**

To perform Exploratory Data Analysis (EDA) on the loan dataset and derive meaningful insights that can be helpful for further modeling.

**Theory:**

Exploratory Data Analysis (EDA) is a critical step in any data science or machine learning workflow. It involves summarizing the main characteristics of a dataset, often using visual methods. Through EDA, one can:

* Identify missing values.
* Detect outliers.
* Understand data distributions.
* Find relationships and correlations.
* Make decisions about data cleaning and feature engineering.

**Measures of Central Tendency**

The central tendency of a dataset refers to its middle or typical values. The most commonly used measures of central tendency are:

* **Mean:** The arithmetic mean is calculated as the sum of all values divided by the number of observations.
* **Median:** The median is the middle value in an ordered dataset. If the number of observations is even, the median is the average of the two middle values.
* **Mode:** The mode is the most frequently occurring value in the dataset. A dataset may have one mode (unimodal), multiple modes (multimodal), or no mode at all.

In addition to these, specialized means like the geometric mean, harmonic mean, and trimmed mean are used in certain statistical applications.

**Measures of Variability**

The variability or dispersion of a dataset indicates how spread out the values are. Some common measures include:

* **Variance (s²):** Variance quantifies the average squared deviation from the mean: It provides insight into how much the data points deviate from the central value.
* **Standard Deviation (s):** The standard deviation is the square root of variance and retains the same unit as the original data, making it more interpretable: It helps in understanding the spread of data, particularly in normally distributed datasets.
* **Interquartile Range (IQR):** The IQR measures the spread of the middle 50% of the data and is calculated as: IQR=Q3−Q1 where Q1 (first quartile) is the 25th percentile and Q3(third quartile) is the 75th percentile of the dataset. IQR is useful for identifying outliers and assessing skewness in the data.

**Outlier Identification**

Outliers are extreme values that differ significantly from the rest of the data. They can be detected using:

* **Boxplots:** A boxplot visually represents the dataset’s distribution, showing the median, quartiles, and potential outliers. Outliers are typically identified as values lying beyond 1.5 times the IQR above Q3Q3 or below Q1Q1.
* **Z-Scores:** Standardizing data points using Z-scores helps in identifying values that deviate significantly (typically beyond ±3 standard deviations from the mean).

**Importance of EDA in Machine Learning:**

* **Understanding Structure**: Learn the structure, dimensions, and types of data.
* **Data Cleaning**: Spot and handle missing values or duplicate data.
* **Outlier Detection**: Use boxplots to identify unusual data points.
* **Correlation Analysis**: Understand how variables relate to each other.
* **Preprocessing Decisions**: Decide on normalization, encoding, or feature transformation.
* **Model Readiness**: Improve model accuracy with high-quality data inputs.

**Dataset:**

The dataset used is train.csv, which contains the following key features:

* **ApplicantIncome**: Income of the applicant.
* **CoapplicantIncome**: Income of the co-applicant.
* **LoanAmount**: Amount of the loan applied for.
* **Loan\_Amount\_Term**: Term of the loan in months.
* **Credit\_History**: Credit history (1 = good, 0 = bad/missing).
* **Property\_Area**: Urban, Semiurban, or Rural.
* **Loan\_Status**: Loan approved (Y) or not (N).

**Steps of Implementation:**

**1. Importing Libraries**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

**2. Loading and Exploring the Dataset**

df = pd.read\_csv("train.csv")

df.head()

df.info()

df.describe()

**3. Checking for Missing Values and Duplicates**

df.isnull().sum()

df.duplicated().sum()

Handling missing values:

df['LoanAmount'].fillna(df['LoanAmount'].mean(), inplace=True)

df['Credit\_History'].fillna(df['Credit\_History'].mode()[0], inplace=True)

df.dropna(inplace=True)

**4. Descriptive Statistics**

df[['ApplicantIncome', 'CoapplicantIncome', 'LoanAmount']].describe()

**5. Data Visualization**

**a. Histogram for numeric variables**

df['ApplicantIncome'].hist(bins=50)

plt.title("Distribution of Applicant Income")

**b. Boxplots to detect outliers**

sns.boxplot(df['LoanAmount'])

plt.title("Loan Amount Outliers")

**c. Bar chart for categorical features**

sns.countplot(x='Loan\_Status', data=df)

sns.countplot(x='Property\_Area', hue='Loan\_Status', data=df)

sns.countplot(x='Credit\_History', hue='Loan\_Status', data=df)

**d. Heatmap to visualize correlations**

corr\_matrix = df[['ApplicantIncome', 'CoapplicantIncome', 'LoanAmount', 'Credit\_History']].corr()

sns.heatmap(corr\_matrix, annot=True, cmap="YlGnBu")

**6. Feature Analysis**

**a. Loan Status vs Income**

sns.boxplot(x='Loan\_Status', y='ApplicantIncome', data=df)

**b. Credit History vs Loan Approval**

pd.crosstab(df['Credit\_History'], df['Loan\_Status'], normalize='index') \* 100

**c. Property Area Trends**

pd.crosstab(df['Property\_Area'], df['Loan\_Status'], normalize='index') \* 100

**Conclusion:**

The Exploratory Data Analysis (EDA) of the loan dataset revealed several key insights. It was observed that applicants with higher income levels generally tend to receive higher loan amounts, indicating a positive relationship between income and loan approval. Credit history emerged as a crucial factor influencing loan approval decisions — applicants with a good credit history (represented by the value 1) were significantly more likely to have their loans approved compared to those with poor or missing credit histories. In terms of geographical trends, loan approval rates were found to be higher in semiurban areas compared to rural areas, suggesting a regional influence on loan sanctioning. Additionally, the analysis of boxplots helped identify outliers, particularly among applicants with extremely high incomes or loan requests, which could potentially skew the dataset and affect model performance if not treated properly. Furthermore, missing values were found mainly in the 'LoanAmount' and 'Credit\_History' columns, and these were effectively handled through imputation techniques to ensure the completeness and reliability of the data**.**

**References:**

* [GeeksforGeeks - What is Exploratory Data Analysis?](https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/)
* <https://github.com/mrnik89/ML_Lab/blob/main/Assignment1_ML.ipynb>