EXPLORATORY DATA ANALYSIS
LOAN APPROVAL ANALYSIS

PROJECT BY :
NAINIKA THAPA

INTRODUCTION

A home loan, also known as a mortgage, is a financial product that allows individuals to purchase a property by borrowing money from a lender, housing finance companies , public banks , private banks . The loan amount, known as the principal, is repaid over a specified term, typically ranging from 15 to 30 years, through monthly installments. These installments include both the principal amount and interest, the latter being the cost of borrowing the money. Home loans come in various types, including fixed-rate mortgages, where the interest rate remains constant, and adjustable-rate mortgages, where the rate can change over time. Factors such as credit score, income stability, and the size of the down payment play crucial roles in securing a home loan and determining the interest rate offered. The process involves several stages, including pre-approval, property search, application submission, underwriting, and closing.

DATASET LOADING

```
# IMPORT THE REQUIRED LIBRARIES
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Importing required libraries for performing data analysis and creating visualization in Python . PANDAS is used for data manipulation , NUMPY is used for numerical operations , SEABORN and MATPLOTIB is used for creating static plots .

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive;
```

Then connecting the Google Drive to the google collab so you can access your files. It then sets the path to CSV file with loan information and reads this file into a dataframe.

DATASET LOADING

```
[ ] a = "/content/drive/MyDrive/loan_sanction_test.csv"
    df = pd.read_csv(a)
```

```
[ ] df.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
0	LP001015	Male	Yes	0	Graduate	No	5720	0	110.0	360.0	1.0	Urban
1	LP001022	Male	Yes	1	Graduate	No	3076	1500	126.0	360.0	1.0	Urban
2	LP001031	Male	Yes	2	Graduate	No	5000	1800	208.0	360.0	1.0	Urban
3	LP001035	Male	Yes	2	Graduate	No	2340	2546	100.0	360.0	NaN	Urban
4	LP001051	Male	No	0	Not Graduate	No	3276	0	78.0	360.0	1.0	Urban

I've named the dataset as 'df'. Then i use df.head() method . This is a method that displays the first few rows of the DataFrame. By default, it shows the first 5 rows. You can specify a different number of rows to display by passing an integer as an argument to the head() method

```
[6] df.shape

(367, 12)
```

There are total 367 rows and 12 columns present in the dataset.

DESCRIPTION

```
[5] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 367 entries, 0 to 366
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Loan_ID            367 non-null    object 
 1   Gender             356 non-null    object 
 2   Married            367 non-null    object 
 3   Dependents         357 non-null    object 
 4   Education          367 non-null    object 
 5   Self_Employed      344 non-null    object 
 6   ApplicantIncome    367 non-null    int64  
 7   CoapplicantIncome  367 non-null    int64  
 8   LoanAmount         362 non-null    float64
 9   Loan_Amount_Term   361 non-null    float64
 10  Credit_History     338 non-null    float64
 11  Property_Area      367 non-null    object 
dtypes: float64(3), int64(2), object(7)
memory usage: 34.5+ KB
```

•**ID:** Unique identifier for each loan application.

•**Gender:** Gender of the applicant (Male/Female).

•**Married:** Marital status of the applicant (Yes/No).

•**Dependents:** Number of dependents the applicant has.

•**Education:** Educational qualification of the applicant (Graduate/Not Graduate).

•**Self Employed:** Whether the applicant is self-employed (Yes/No).

•**Applicant Income:** Applicant's income.

•**Co applicant Income:** Co-applicant’s

•**Loan Amount:** Loan amount in thousands.

•**Loan Amount Term:** Term of the loan in months.

•**Credit History:** Credit history meets guidelines (0 or 1).

•**Property Area:** Urban/Semi-Urban/Rural .

DESCRIPTION

```
df.describe()
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	367.000000	367.000000	362.000000	361.000000	338.000000
mean	4805.599455	1569.577657	136.132597	342.537396	0.825444
std	4910.685399	2334.232099	61.366652	65.156643	0.380150
min	0.000000	0.000000	28.000000	6.000000	0.000000
25%	2864.000000	0.000000	100.250000	360.000000	1.000000
50%	3786.000000	1025.000000	125.000000	360.000000	1.000000
75%	5060.000000	2430.500000	158.000000	360.000000	1.000000
max	72529.000000	24000.000000	550.000000	480.000000	1.000000

There are 5 numerical columns present in the dataset i.e. ApplicantIncome ,LoanAmount , CoapplicantIncome , Loan_Amount_Term ,Credit_History .

CHECKING NULL VALUES

```
[ ] df.isnull().sum().sum()

84
```

There are 84 null values present in the dataset .

```
df.isnull().sum()
```

	0
Loan_ID	0
Gender	11
Married	0
Dependents	10
Education	0
Self_Employed	23
ApplicantIncome	0
CoapplicantIncome	0
LoanAmount	5
Loan_Amount_Term	6
Credit_History	29
Property_Area	0

Checked null values in the dataset using df.isnull().sum() . This provide column wise null values present in the dataset . As we can see there is 11 null values in Gender , 10 in Dependents , 23 in Self_Employed , 5 in LoanAmount , 6 in Loan_Amount_Term and 29 in Credit_History .

DATA CLEANING

❖ NOW , filling the null values present in the above mentioned columns :

➢ Filling the null values of Gender , Dependents and Self_Employed with mode because the columns are categorical.

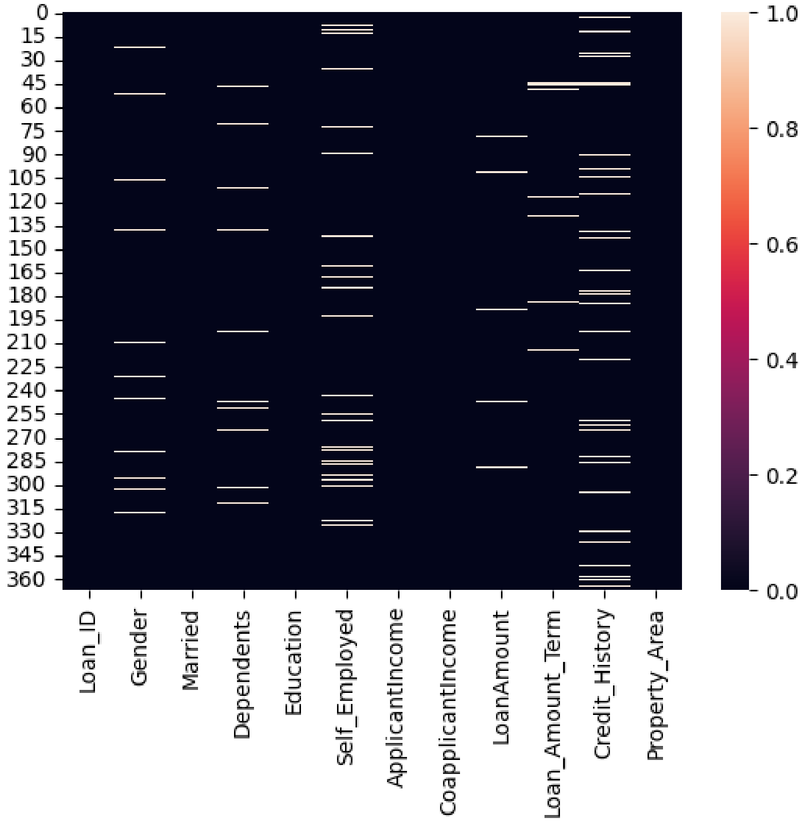
```
for column in ['Gender', 'Dependents', 'Self_Employed']:
    # Calculate the mode of the column
    mode_value = df[column].mode()[0]
    # Fill missing values with the mode
    df[column].fillna(mode_value, inplace=True)
```

➢ Filling the null values of LoanAmount , Credit_History , Loan_Amount_Term with median of the column because the median is less sensitive to extreme values .

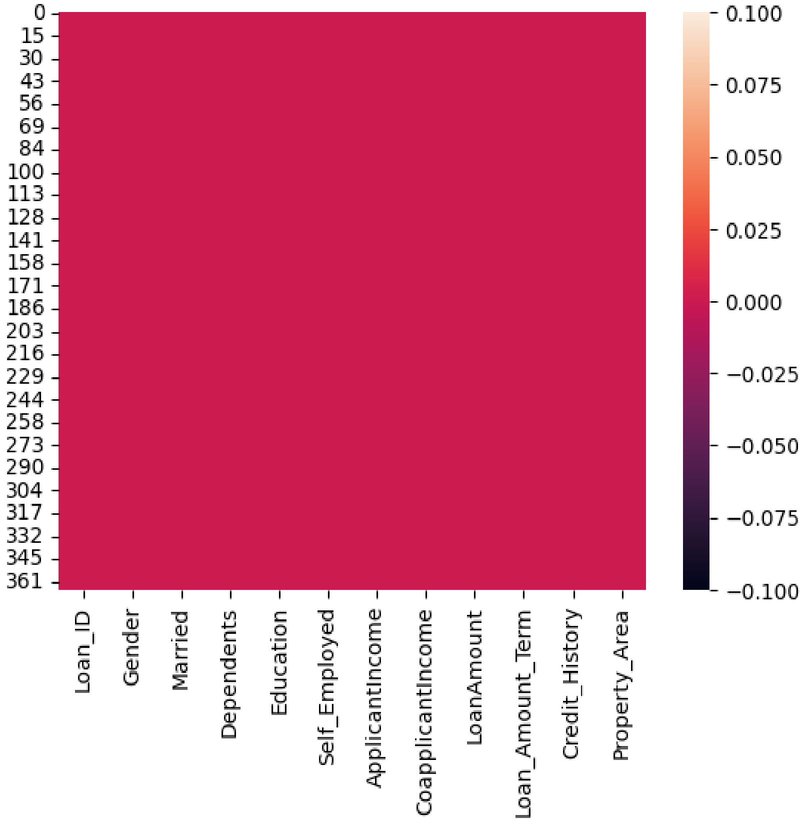
```
for column in ['LoanAmount','Credit_History','Loan_Amount_Term']:
    # Calculate the median of the column
    median_value = df[column].median()
    # Fill missing values with the median
    df[column].fillna(median_value, inplace=True)
```

DATA CLEANING

Heatmap before cleaning



Heatmap after cleaning



DATA CLEANING

- Handling Outliers : Here i used IQR method for outliers detection and box plot for visualization of outliers .
- Outliers in LoanAmount : There are 18 outliers present in the column and it was imputed by median

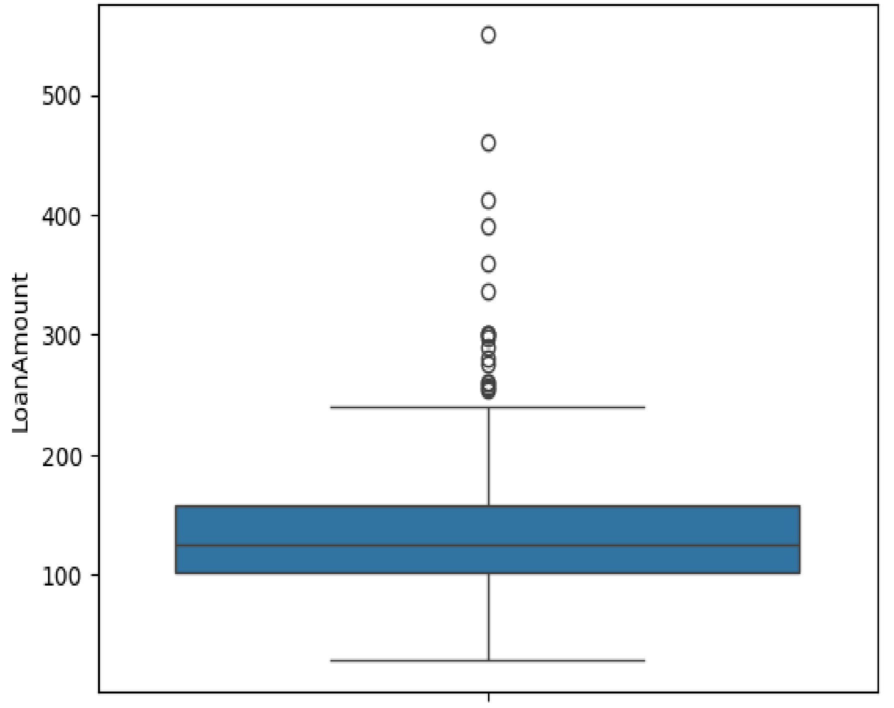
```
# Creating Boxplot before filling outliers
sns.boxplot(df['LoanAmount'])
plt.show()

#
Q1 = df['LoanAmount'].quantile(0.25)
print(f'Q1 is {Q1}')
Q3 = df['LoanAmount'].quantile(0.75)
print(f'Q3 is {Q3}')
IQR = Q3 - Q1
print(f'IQR is {IQR}')
# Define upper and lower bounds for outliers
upper_bound = Q3 + 1.5 * IQR
print(f'Upper bound is {upper_bound}')
lower_bound = Q1 - 1.5 * IQR
print(f'Lower bound is {lower_bound}')

# Identify outliers
outliers = df[(df['LoanAmount'] > upper_bound) | (df['LoanAmount'] < lower_bound)]
num_outliers = len(outliers)
print(f"Number of outliers in 'your_column': {num_outliers}")
```

```
Q1 is 101.0
Q3 is 157.5
IQR is 56.5
Upper bound is 242.25
Lower bound is 16.25
Number of outliers in 'your_column': 18
```

- Boxplot of LoanAmount before outlier handling.

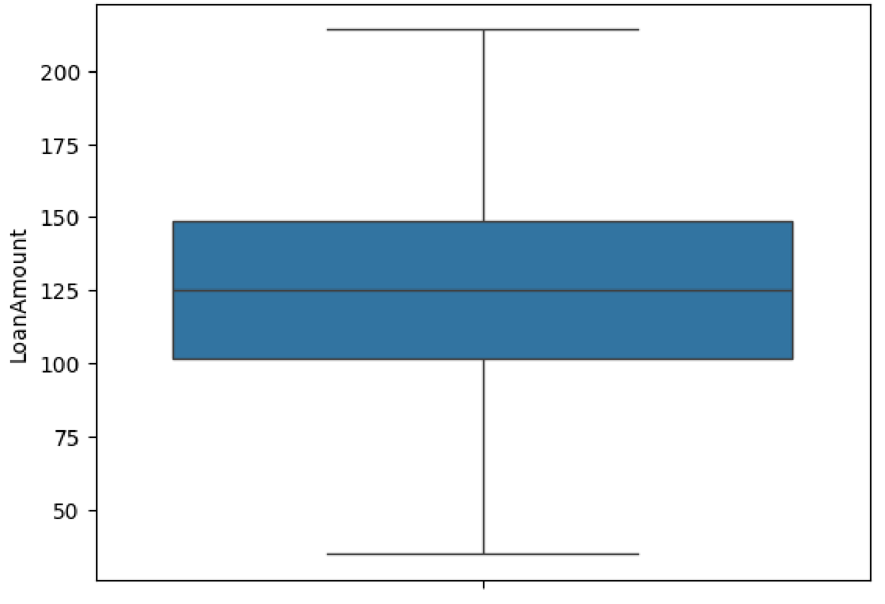


DATA CLEANING

```
# Filling the outliers with median
median_income = df['LoanAmount'].median()
df['LoanAmount'] = np.where(df['LoanAmount'] > upper_bound, median_income, df['LoanAmount'])
df['LoanAmount'] = np.where(df['LoanAmount'] < lower_bound, median_income, df['LoanAmount'])


# Creating Boxplot after filling outliers
sns.boxplot(df['LoanAmount'])
plt.show()
```

The outliers present in the LoanAmount attribute are imputed by median , the box plot shows the zero outliers present in LoanAmount term .

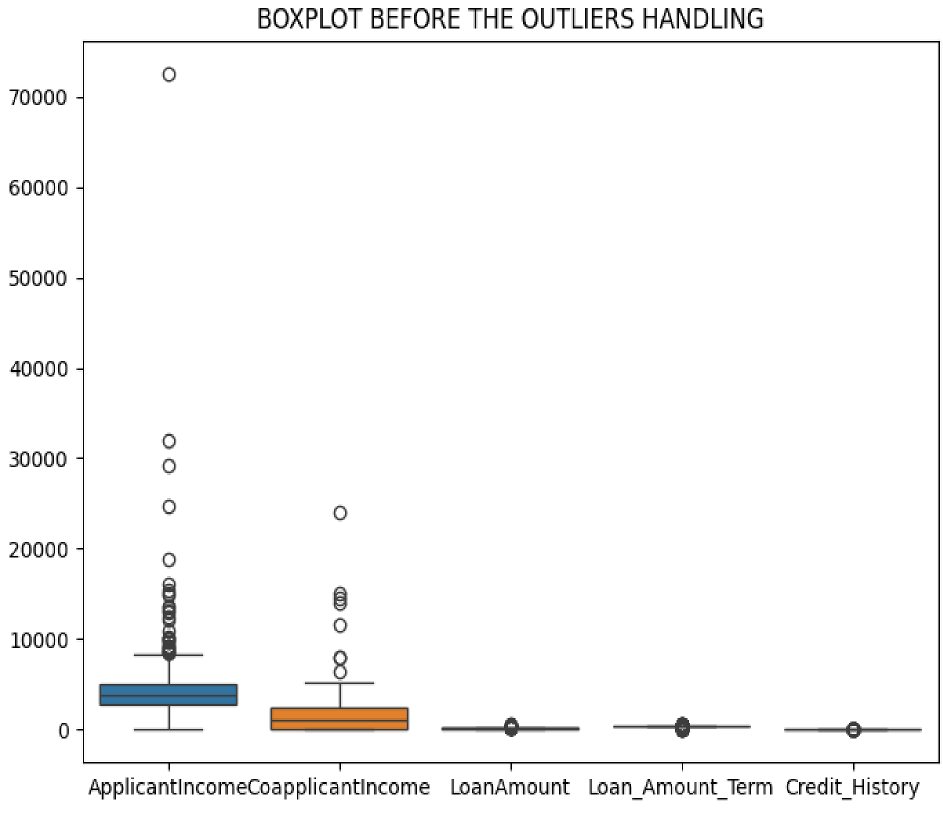


Boxplot of LoamAmount after outlier handling.

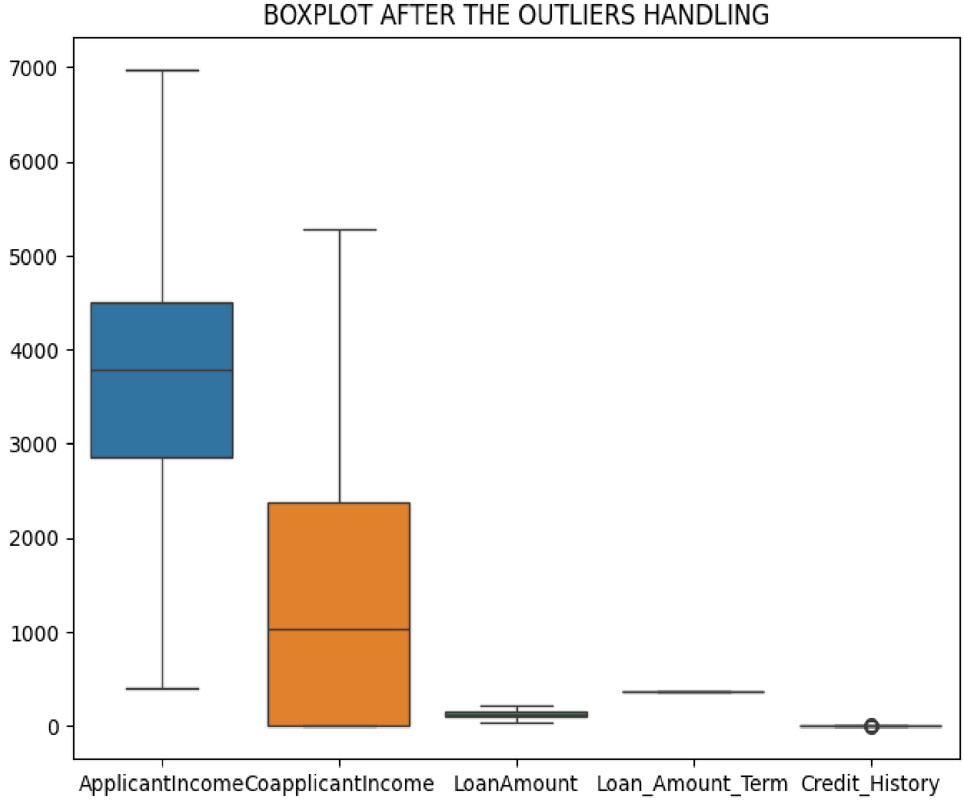
Outliers were present in the all numerical columns, other than LoanAmount were also addressed using the IQR method and capping to boundary values .

DATA CLEANING

Boxplots before Outliers handling :



Boxplots after Outliers handling :



SUMMARY :

- To summarize , addressing NULL VALUES & OUTLIERS necessitates a methodical approach tailored to the data's characteristics and specific attributes . Data cleaning and outliers handling are important steps for accurate analysis .

- The dataset contains missing values in 'Gender' , 'Dependents' , 'Self_employed' , 'LoanAmount' , 'Loan_Amount_Term' , 'Credit_History' attributes . These were handled by median/mode .

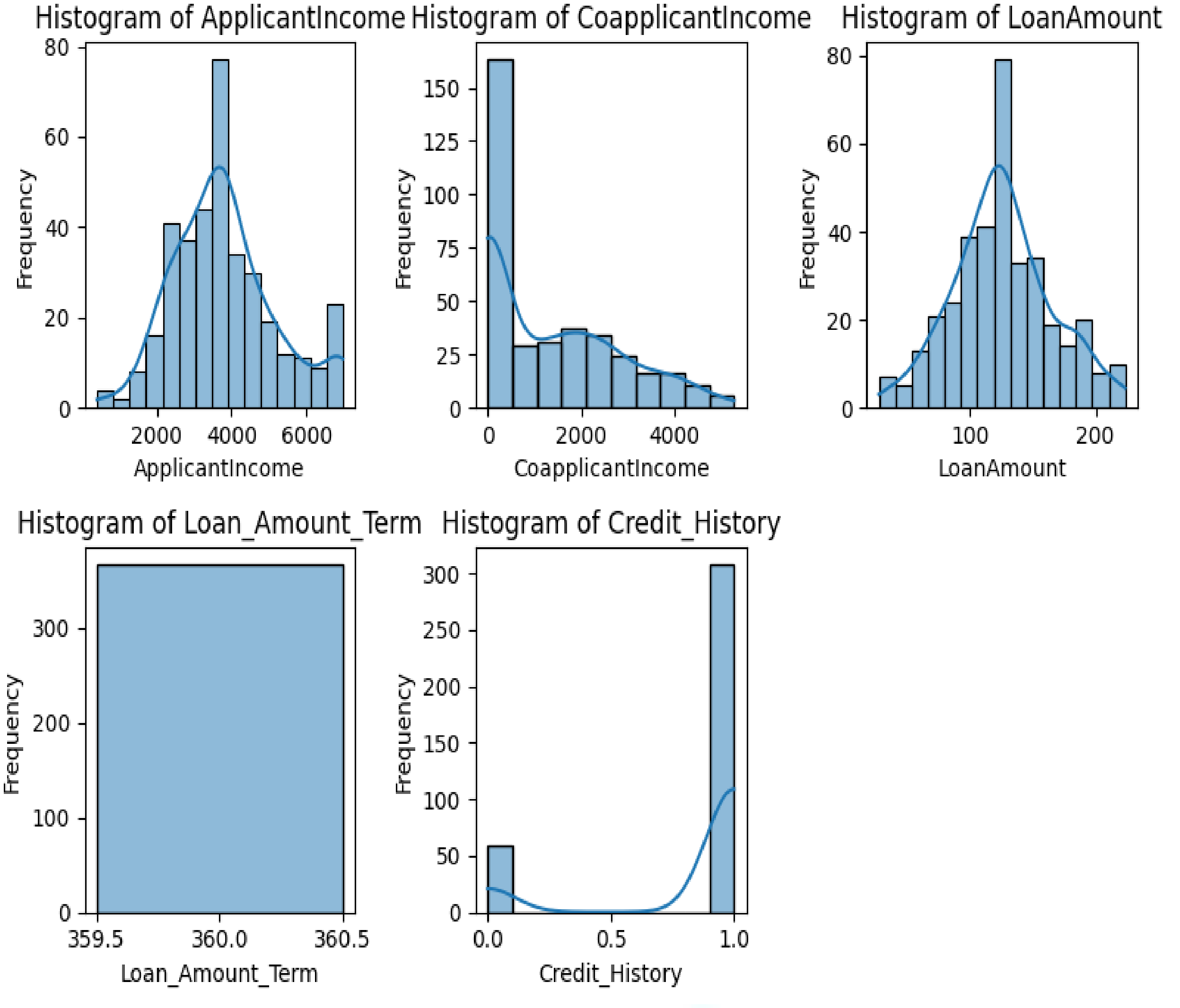
- Outliers present in all the numerical columns were addressed using IQR method and capping to boundary values .

- With the null , missing , and invalid values appropriately addressed , now we are ready to move forward with analyzing the dataset for visualization and insights .

DATA VISUALIZATION

Univariate Analysis

Histograms: Plot the frequency distribution of key numeric variables.



KEY INSIGHTS :

- ApplicantIncome: The distribution shows a peak around 4000, indicating that most applicants have an income in this range.

- CoapplicantIncome: The distribution peaks at 0, suggesting that many applicants not have a coapplicant or the coapplicant's income is not significant.

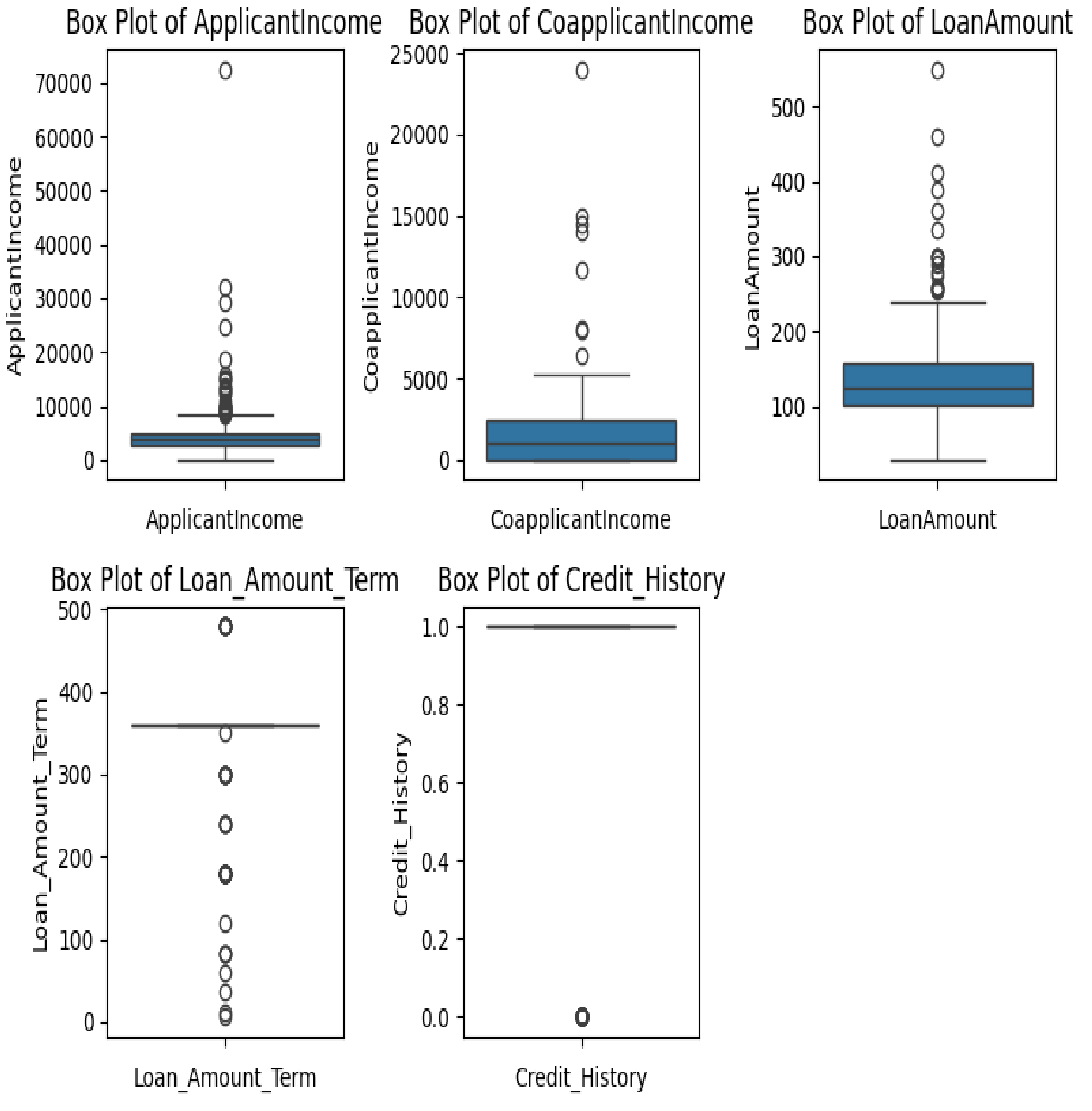
- LoanAmount: The peak around 100 suggests that most loan amounts are around this value.

- Loan_Amount_Term: The single bar at 360 indicates that the majority of loan terms are around 360 months (30 years).

- Credit_History: The peak at 1 shows that most applicants have a credit history terms are around 360 months (30 years).

DATA VISUALIZATION

Box Plots: Identify potential outliers and visualize the spread of data .

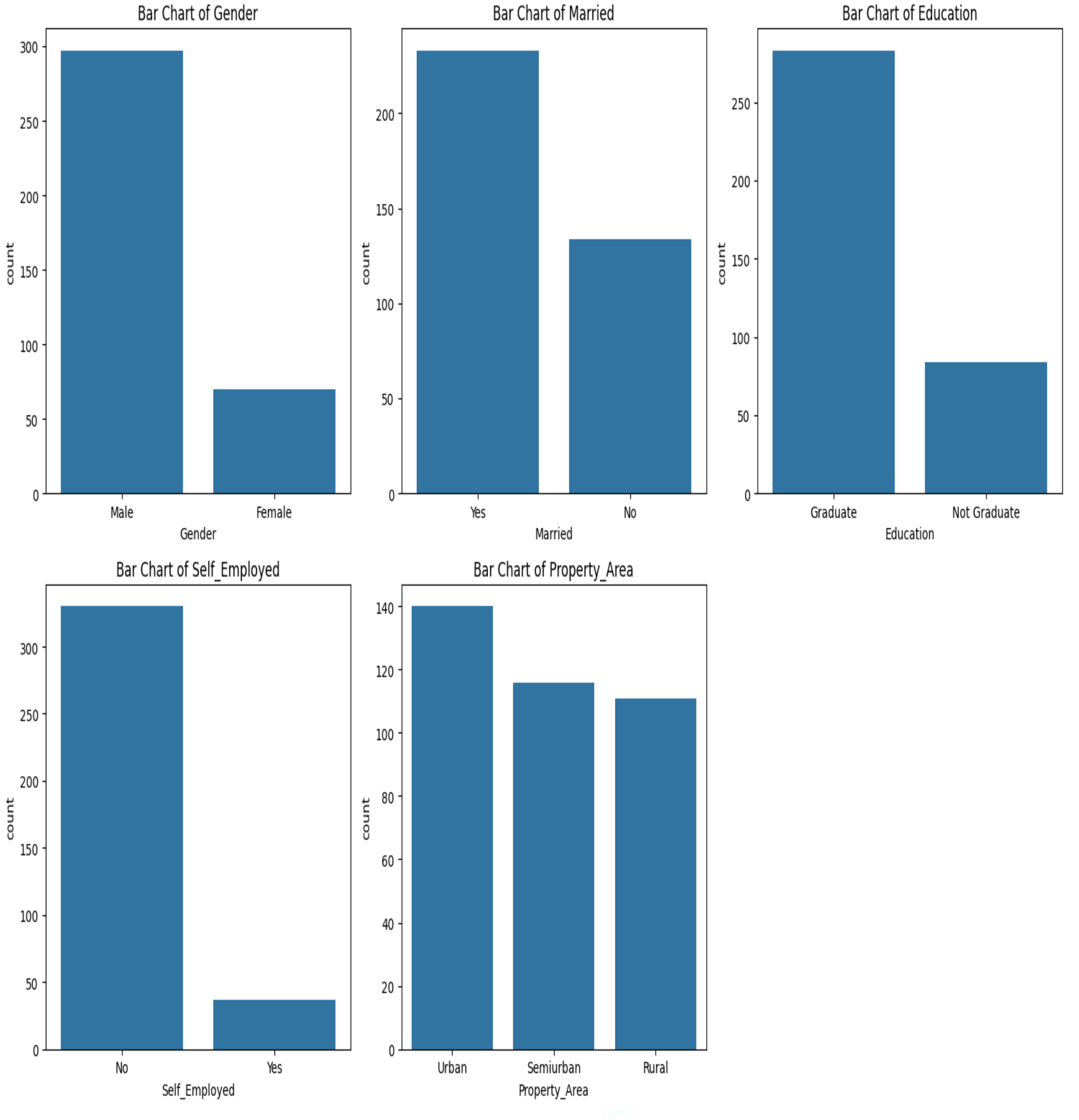


KEY INSIGHTS :

- ApplicantIncome: The majority of the data is concentrated below 10,000, with several outliers extending up to 70,000. This indicates that most applicants have a moderate income, but there are a few with significantly higher incomes.
- CoapplicantIncome: Most of the data is concentrated below 5,000, with outliers extending up to 20,000. This suggests that many applicants do not have a coapplicant or the coapplicant's income is relatively low.
- LoanAmount: The majority of the data is concentrated below 200, with outliers extending up to 500. This indicates that most loan amounts are moderate, but there are a few larger loan requests.
- Loan_Amount_Term: Most of the data is concentrated below 360, with outliers extending up to 480. This suggests that the majority of loan terms are around 30 years, with a few longer terms.
- Credit_History: Most of the data points are at 1, with a few outliers at 0. This indicates that most applicants have a good credit history, with a few exceptions

DATA VISUALIZATION

Bar Charts: Visualize the frequency distribution of categorical variables .

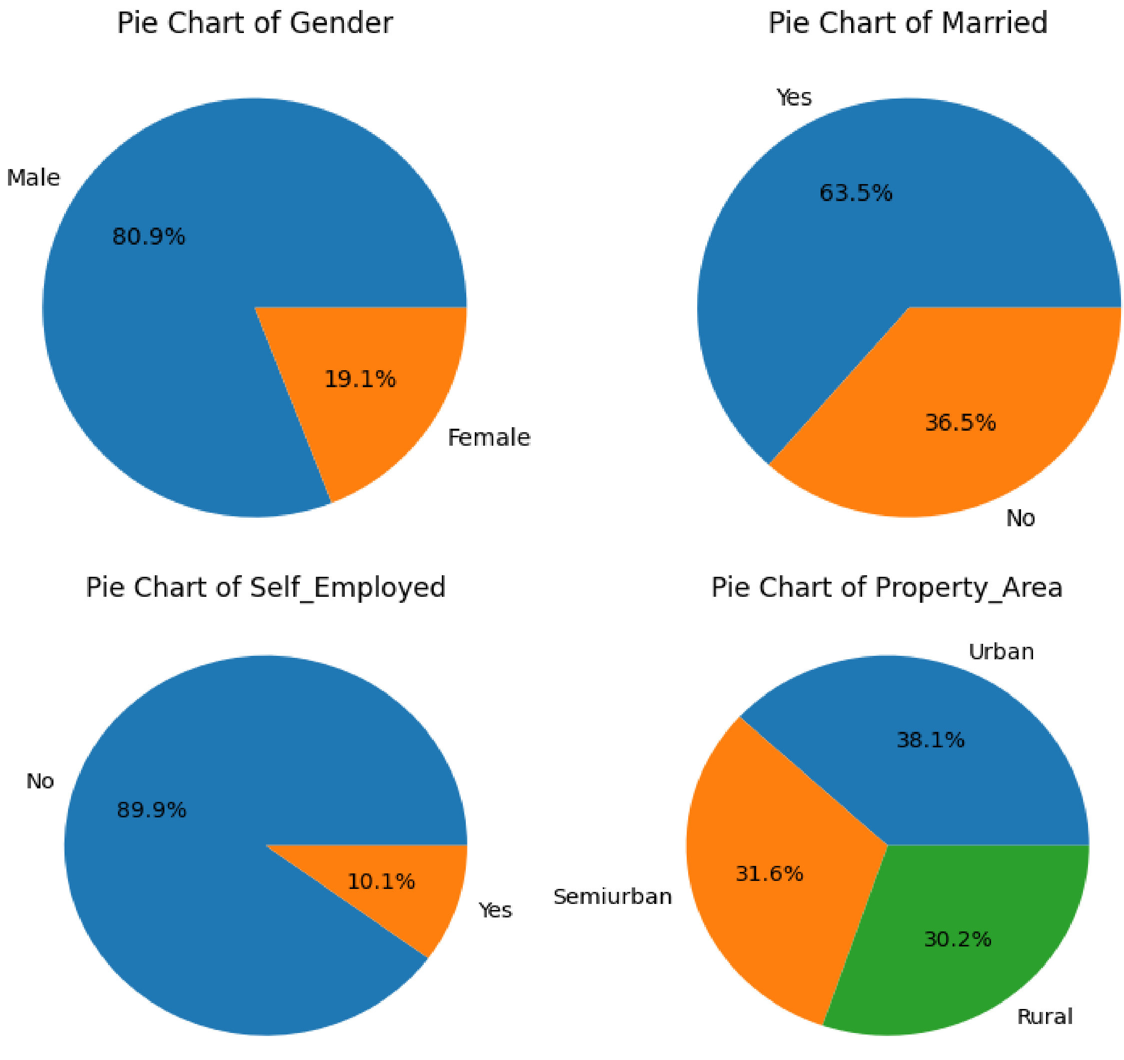


KEY INSIGHTS :

- Gender: There are significantly more males (around 300) compared to females (around 80). This indicates a higher number of male applicants.
- Marital Status: There are more married individuals (around 220) compared to unmarried individuals (around 140). This suggests that a majority of the applicants are married.
- Education: There are more graduates (around 270) compared to non-graduates (around 80). This indicates that most applicants have a higher level of education.
- Self-Employed: There are significantly more individuals who are not self-employed (around 320) compared to those who are self-employed (around 50). This suggests that most applicants are employed by others.
- Property Area: The counts are relatively close, with urban areas having the highest count (around 140), followed by semiurban (around 120), and rural areas (around 100). This indicates a diverse distribution of property areas among the applicants.

DATA VISUALIZATION

Pie Charts: Represent the composition of categorical variables .



KEY INSIGHTS :

- Gender: The majority of applicants are male (80.9%), while females make up 19.1%. This indicates a higher number of male applicants.

- Marital Status: Most applicants are married (63.5%), while 36.5% are unmarried. This suggests that a majority of the applicants are married.

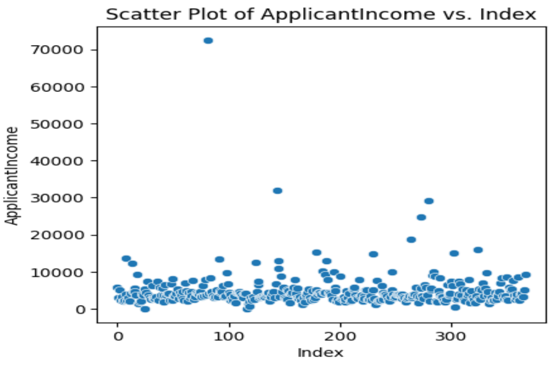
- Self-Employed: A significant majority of applicants are not self-employed (89.9%), while only 10.1% are self-employed. This indicates that most applicants are employed by others.

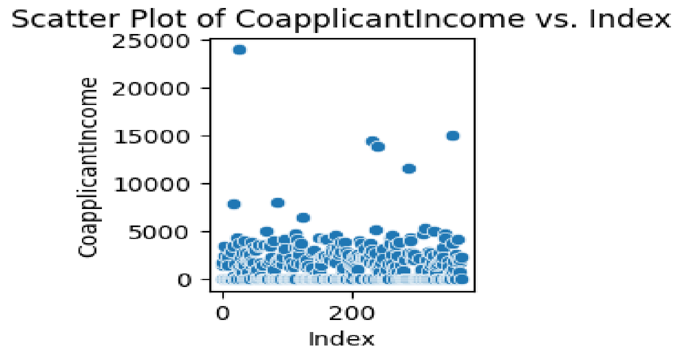
- Property Area: The distribution of property areas among the applicants is relatively balanced, with urban areas having the highest count (38.1%), followed by semiurban (31.6%), and rural areas (30.2%).

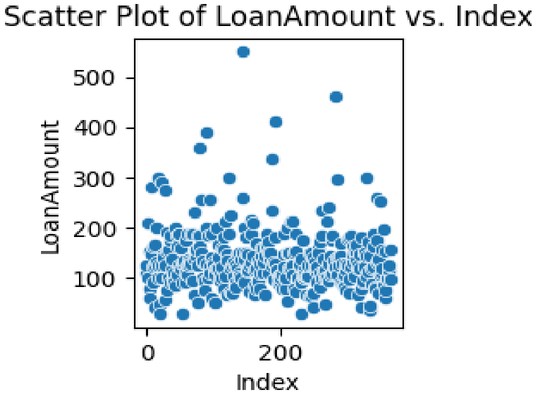
DATA VISUALIZATION

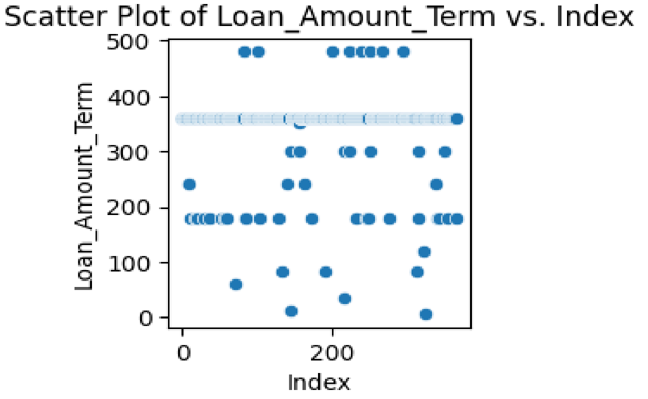
BIVARIATE ANALYSIS :

Create scatter plots to explore relationships between pairs of numeric variables







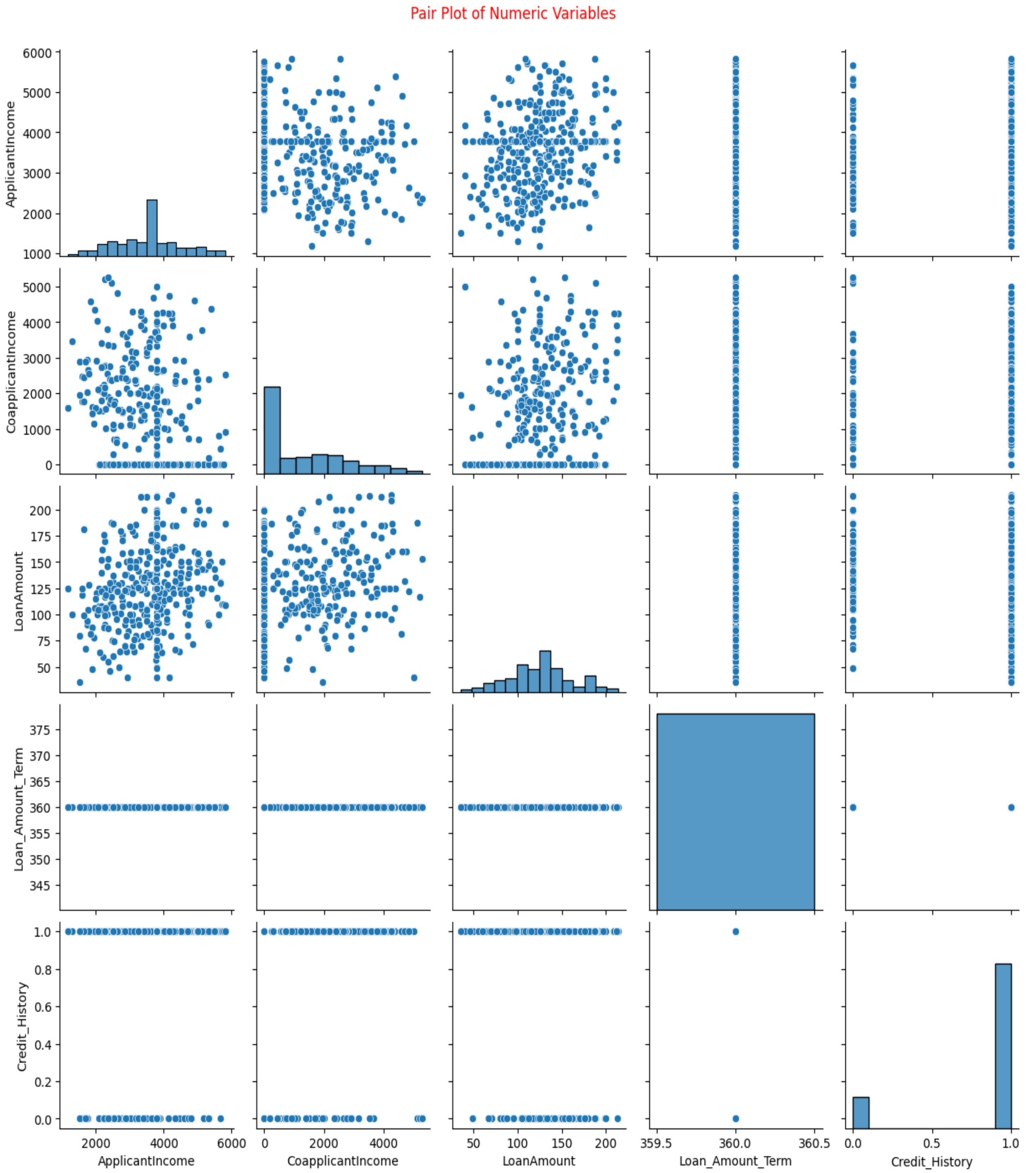


KEY INSIGHTS :

- ApplicantIncome: The majority of the data points are clustered between 0 and 10,000 on the y-axis, indicating that most applicants have a moderate income. However, there are a few outliers with incomes reaching up to 70,000, suggesting that some applicants have significantly higher incomes.
- CoapplicantIncome: The majority of the data points are clustered near the bottom of the plot, indicating lower coapplicant incomes.
- LoanAmount: The majority of the data points are clustered between 0 and 200 on the LoanAmount axis, indicating that most loan amounts are moderate.
- Loan_Amount_Term: The majority of the data points are concentrated around the 360 mark on the y-axis, indicating that most loan terms are around 360 months (30 years). There are a few data points scattered at different values, suggesting some variation in loan terms.
- Credit_History: The data points are clustered at two distinct y-values: 0.0 and 1.0. This indicates that the Credit_History variable is binary, with values either 0 or 1. Most data points are at 1.0, suggesting that the majority of applicants have a good credit history.

DATA VISUALIZATION

Use pair plots (scatter matrix) to visualize interactions between multiple numeric variables simultaneously.

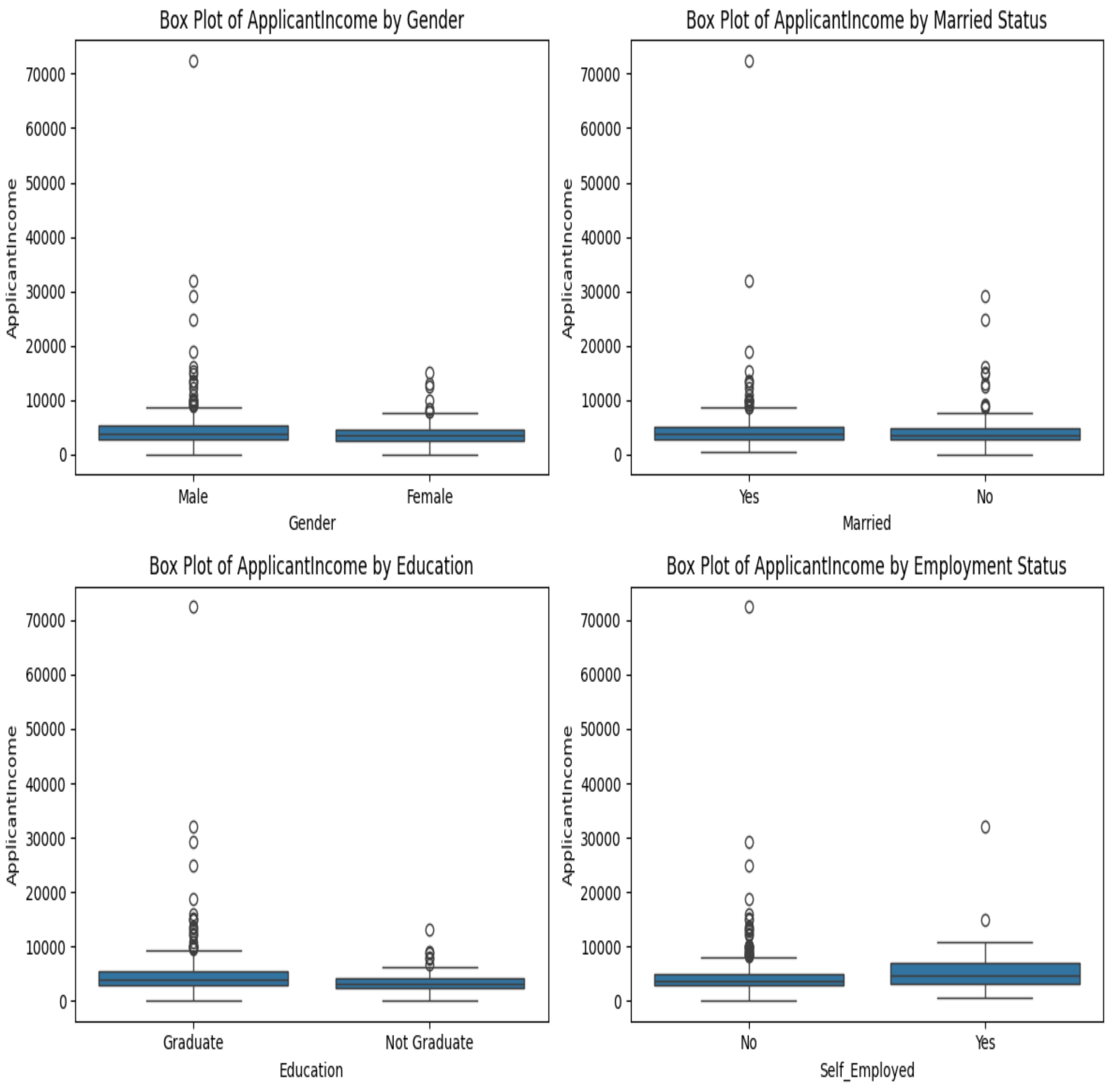


KEY INSIGHTS :

- ApplicantIncome: The histograms show that most applicants have an income below 10,000, with a few outliers having significantly higher incomes. The scatter plots indicate a positive correlation between ApplicantIncome and LoanAmount .
- CoapplicantIncome: The histograms reveal that many coapplicants have an income of 0, suggesting that they do not contribute financially. The scatter plots show a positive correlation between CoapplicantIncome and LoanAmount.
- LoanAmount: The histograms indicate that most loan amounts are below 200, with a few outliers requesting higher amounts. The scatter plots show positive correlations with both ApplicantIncome and CoapplicantIncome.
- Loan_Amount_Term: The histograms show that most loan terms are around 360 months (30 years). The scatter plots do not show strong correlations with other variables.
- Credit_History: The histograms reveal that most applicants have a credit history of 1, indicating a good credit record. The scatter plots show that applicants with a credit history of 1 are more likely to have higher loan amounts .

DATA VISUALIZATION

Create box plots for each combination of numerical and categorical columns



KEY INSIGHTS :

- Gender: The box plot shows that the median income for males is higher than that for females. There are also more outliers among males, indicating a wider range of incomes.

- Marital Status: Married applicants tend to have a higher median income compared to unmarried applicants. There are also more outliers among married applicants, suggesting a greater variation in incomes.

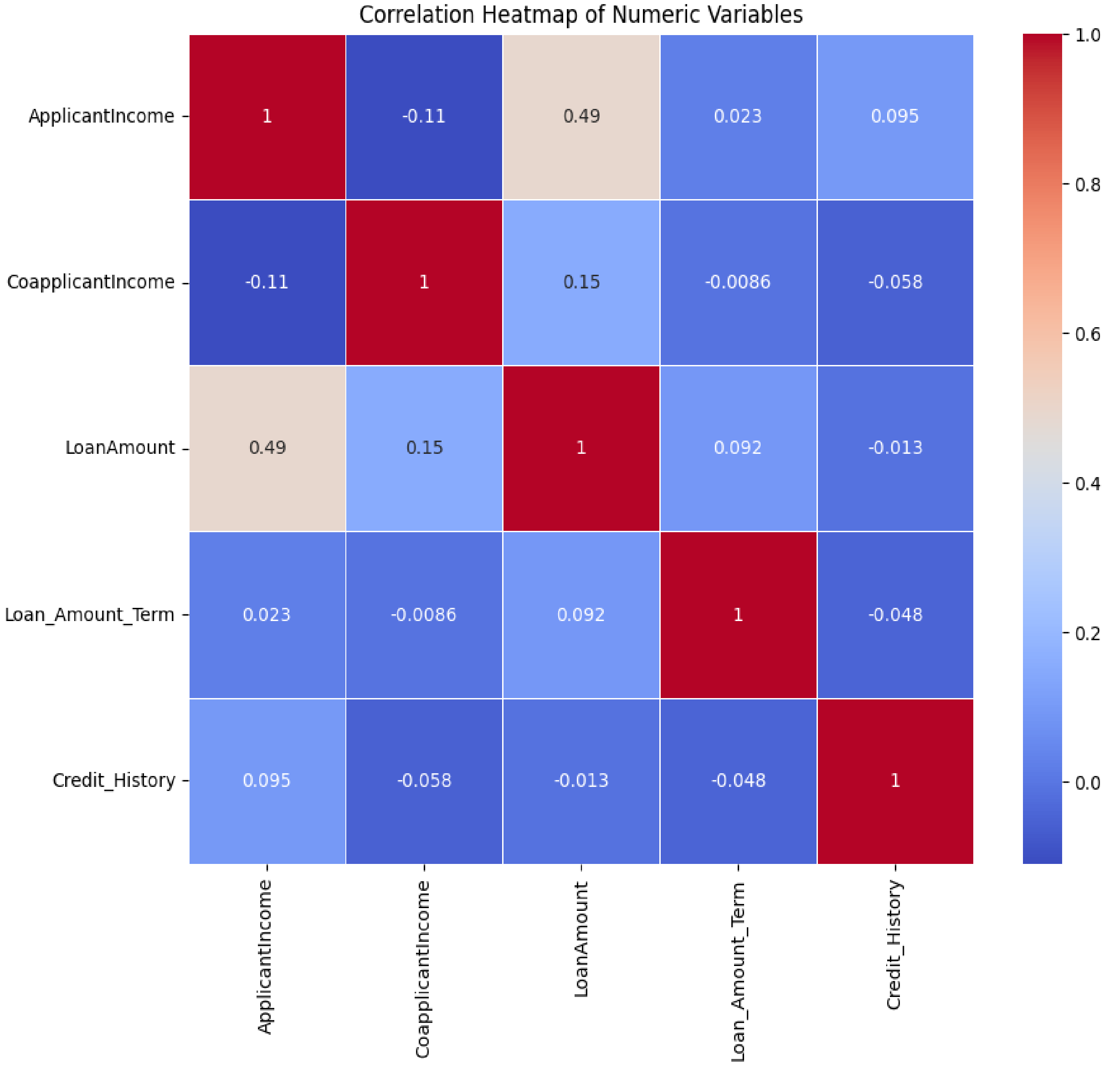
- Education: Graduates have a higher median income compared to non-graduates. The range of incomes is also wider for graduates, with more outliers present.

- Employment Status: Self-employed individuals have a higher median income compared to those who are not self-employed. There are also more outliers among self-employed individuals, indicating a greater variation in incomes.

DATA VISUALIZATION

Multivariate Analysis :

Perform a correlation analysis to identify relationships between numeric variables. Visualize correlations using a heatmap.



KEY INSIGHTS :

- ApplicantIncome:

Positively correlated with LoanAmount (0.49), indicating that higher applicant incomes are associated with higher loan amounts.

Weak positive correlation with Credit_History (0.095), suggesting that applicants with higher incomes tend to have a good credit history.

Weak negative correlation with CoapplicantIncome (-0.11), indicating that higher applicant incomes are slightly associated with lower coapplicant incomes.

- CoapplicantIncome:

Weak positive correlation with LoanAmount (0.15), suggesting that higher coapplicant incomes are slightly associated with higher loan amounts.

Weak negative correlation with Credit_History (-0.058), indicating that higher coapplicant incomes are slightly associated with poorer credit history.

- LoanAmount:

Weak positive correlation with Loan_Amount_Term (0.092), indicating that higher loan amounts are slightly associated with longer loan terms.

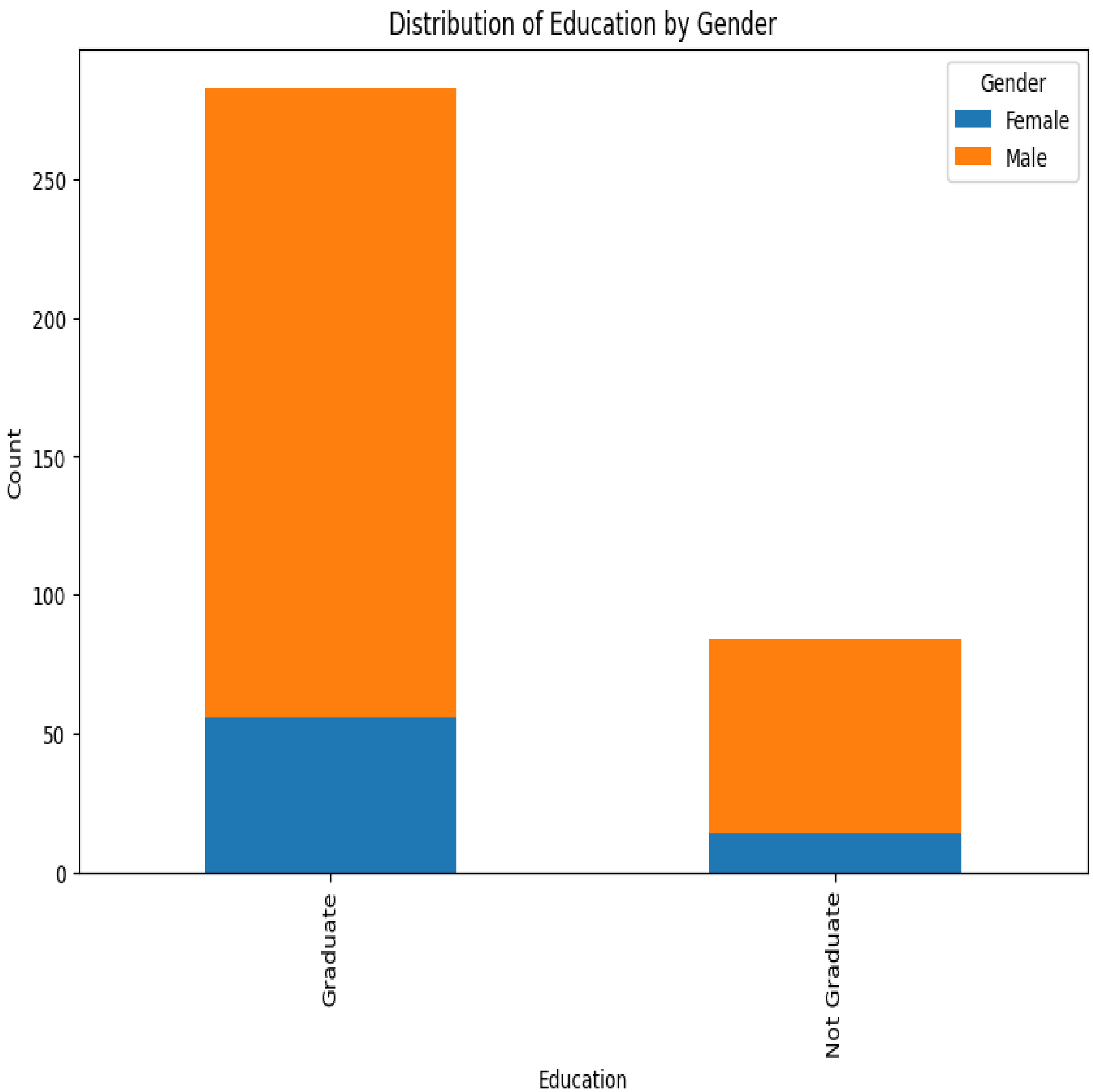
Weak negative correlation with Credit_History (-0.013) , suggesting that higher loan amounts are slightly associated with poorer credit history.

- Loan_Amount_Term:

Weak negative correlation with Credit_History (-0.048) , indicating that longer loan terms are slightly associated with poorer credit history.

DATA VISUALIZATION

Create a stacked bar chart to show the distribution of categorical variables across multiple categories



KEY INSIGHTS :

- Graduates: The number of graduates is significantly higher than the number of non-graduates. Among graduates, males outnumber females, with the male count being approximately 250 and the female count being around 50.

- Non-Graduates: Among non-graduates, males also outnumber females, with the male count being around 50 and the female count being slightly above 0.

FINAL REPORT

SUMMARIZING THE KEY FINDINGS , DRAWING CONCLUSIONS , AND PROVIDING RECOMMENDATIONS BASED ON THE INSIGHTS GAINED FROM THE ANALYSIS .

KEY FINDINGS :

- Demographics: Most applicants are male, married, and graduates. There is a higher representation of applicants from urban and semi-urban areas compared to rural areas.

- Income: The majority of applicants have a moderate income, with a few outliers having significantly higher incomes. Co-applicants often have lower or no income, suggesting they may not be primary contributors to loan repayment.

- Loan Characteristics: Most loan amounts are moderate, with a few outliers requesting larger loans. Loan terms are typically around 30 years, with a few longer terms. Most applicants have a good credit history.

FINAL REPORT

- **Relationships:** There is a positive correlation between ApplicantIncome and LoanAmount, indicating that higher applicant incomes are associated with higher loan amounts. A similar positive correlation exists between CoapplicantIncome and LoanAmount, though it is weaker.

- **Categorical Insights**: Males generally have higher incomes than females. Married applicants tend to have higher incomes than unmarried applicants. Graduates have higher incomes compared to non-graduates. Self-employed individuals have higher incomes than those who are not self-employed.

CONCLUSIONS :

- Loan applications are primarily driven by males with moderate to high incomes, who are more likely to be married and graduates.

- Co-applicants' income plays a less significant role in loan applications, suggesting their financial contribution may be secondary

- Applicants with higher incomes and good credit history are more likely to be approved for larger loan amounts.

- Loan approval may be influenced by demographic factors such as gender, marital status, education, and employment status.

RECOMMENDATIONS

- **Targeted Marketing:** Focus marketing efforts on males, married individuals, and graduates, as they constitute a larger portion of loan applicants.

- **Co-applicant Assessment:** Develop a more nuanced assessment of co-applicants' financial situation to better evaluate their contribution to loan repayment.

- **Income Verification:** Implement robust income verification processes to ensure accuracy and mitigate risks associated with outlier incomes.

- **Credit History:** Emphasize the importance of maintaining a good credit history to improve loan approval chances.

THANK YOU FOR READING

FOR CODING PART , KINDLY VISIT THE LINK BELOW

LOAN ANALYSIS.ipynb - Colab