EXPLORATORY DATA ANALYSIS

PROJECT BY
ADARSH SINGH

About Home Loan

- BUYING A HOUSE IS ONE OF THE BIGGEST DREAMS COME TRUE FOR MOST PEOPLE AND AN EXTRAVAGANT AFFAIR ALTOGETHER. IMPARTING LIFE TO SUCH A DREAM REQUIRES A LOT OF EFFORT FROM THE BUYERS' END, AND THE BEST ONE CAN DO TO ACCOMMODATE THE HOME IN THEIR BUDGET IS THROUGH A HOME LOAN.

- A HOME LOAN CAN BE OPTED TO BUY A NEW HOUSE/FLAT OR A PLOT OF LAND WHERE YOU CAN CONSTRUCT A HOUSE, AND EVEN FOR RENOVATION, EXTENSION, AND REPAIRS TO AN EXISTING HOUSE.

- THIS IS THE MOST COMMON TYPE OF HOME LOAN AVAILED TO PURCHASE A HOUSE. THERE ARE MANY HOUSING FINANCE COMPANIES, PUBLIC BANKS, AND PRIVATE BANKS THAT OFFER HOUSING LOANS WHERE YOU BORROW MONEY TO PURCHASE THE HOUSE OF YOUR CHOICE AND REPAY THE LOAN IN MONTHLY INSTALLMENTS.

INTRODUCTION

- THIS DATASET IS SOURCED FROM SKILL CIRCLE AND CONTAINS DATA COLLECTED FROM “LOAN APPROVAL ANALYSIS”. AFTER A QUICK VIEW OF THE DATASET, IT LOOKS LIKE A TYPICAL DATA FRAME. Home loan Approval.

- WE CAN SEE THAT THERE ARE NAN VALUES PRESENT IN SOME COLUMNS AS WELL. IT CONTAINS 8807 UNIQUE TV SHOWS AND MOVIES. THIS DATASET IS WIDELY USED BY BEGINNERS TO LEARN EDA

PURPOSE OF THE PROJECT:

- THE GOAL OF THE NETFLIX EDA PROJECT IS TO CONDUCT A COMPREHENSIVE EXPLORATION AND ANALYSIS OF NETFLIX'S CONTENT DATASET. THIS INCLUDES UNDERSTANDING THE DATA STRUCTURE, ENSURING DATA INTEGRITY BY HANDLING MISSING VALUES AND DUPLICATES, DERIVING DESCRIPTIVE STATISTICS, AND VISUALIZING

- THE GOALS OF THIS ASSESSMENT ARE TO:
- GAIN FAMILIARITY WITH THE DATASET.
- PERFORM DATA EXPLORATION AND VISUALIZATION.
- IDENTIFY PATTERNS, TRENDS, AND POTENTIAL INSIGHTS.
- GENERATE MEANINGFUL VISUALIZATIONS TO COMMUNICATE YOUR FINDINGS.

DESCRIPTION OF DATASET:

- I HAVE CONDUCTED MY WORK USING GOOGLE COLAB NOTEBOOK.
- THE DATASET HAS BEEN IMPORTED FROM GOOGLE DRIVE.
- AS WE BEGIN OUR EXPLORATORY DATA ANALYSIS (EDA),I'VE NAMED THE DATASET 'DF'.
- THE DATASET COMPRISES OF 8807 ROWS AND 12 COLUMNS.
- FOR DATA CLEANING/VISUALIZATION, I HAVE UTILIZED LIBRARIES LIKE NUMPY, PANDAS, SEABORN & PLOTLY.
- ANY DUPLICATE ENTRIES THAT WERE FOUND HAVE ALSO BEEN REMOVED

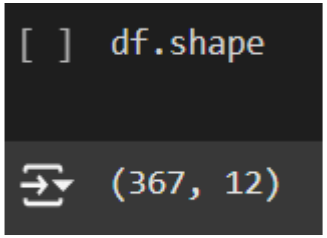
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
```

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

DESCRIPTION OF DATASET:

• The dataset under examination provides a comprehensive insight into home loan product offerings and sales dynamics. It encompasses 12 key attributes that shed light on various facets of the business:

• It looks like you have a Data Frame with 367 entries and 12 columns related to loan data. Here's a quick summary of the columns:



• **Loan_ID:**
• Type: Object
• Description: A unique identifier assigned to each loan application. This helps in tracking and referencing individual loans.
• **Gender:**
• Type: Object
• Description: The gender of the loan applicant (e.g., Male, Female). There are some missing values in this column

DESCRIPTION OF DATASET:

- **Married**
- Type: Object
- Description: Indicates whether the applicant is married or not (e.g., Yes, No). This can affect loan eligibility and risk assessment.
- **Dependents:**
- Type: Object
- Description: The number of dependents (e.g., 0, 1, 2+, or possibly "3+"). This may impact the applicant's financial responsibilities.
- **Education:**
- Type: Object
- Description: The educational background of the applicant (e.g., Graduate, Not Graduate). Higher education levels might correlate with higher income potential.
- **Self_Employed:**
- Type: Object
- Description: Indicates whether the applicant is self-employed (e.g., Yes, No). Self-employed individuals may have variable income, impacting loan approval.

DESCRIPTION OF DATASET:

- ApplicantIncome
- Type: Int64
- Description: The income of the applicant in a given time period (usually monthly). This is a crucial factor in determining loan eligibility and amount.
- CoapplicantIncome:
- Type: Int64
- Description: The income of the coapplicant (if any). Combining incomes can strengthen an application and affect the loan amount.
- LoanAmount:
- Type: Float64
- Description: The total amount of money requested for the loan. This can be influenced +by the applicant's income and creditworthiness.
- Loan_Amount_Term:
- Type: Float64
- Description: The duration (in months) over which the loan is to be repaid (e.g., 360 months for a 30-year mortgage). This affects the monthly payment and total interest paid.

DESCRIPTION OF DATASET:

```
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
```

- **Credit_History:**
- A binary indicator (usually 0 or 1) representing the applicant's credit history. A good credit history (1) suggests lower risk, while a poor credit history (0) indicates higher risk. There are some missing values here as well.

- **Property_Area**
- Description: The geographical area where the property is located (e.g., Urban, Semiurban, Rural). The area can impact property value and loan approval

DATA CLEANING & PRE-PROCESSING:

- The Dataset contains a total of 84 Null values. Of these, 44 are found in categorical features, while 40 are in numerical features.
- First handling missing value in numerical column
- ##Mean: Whenever your data is numeric and normally distributed, in this case you will impute missing values with Mean.
- ##Median: Whenever your data is numeric and skewed, in this case you will impute missing values with Median.

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

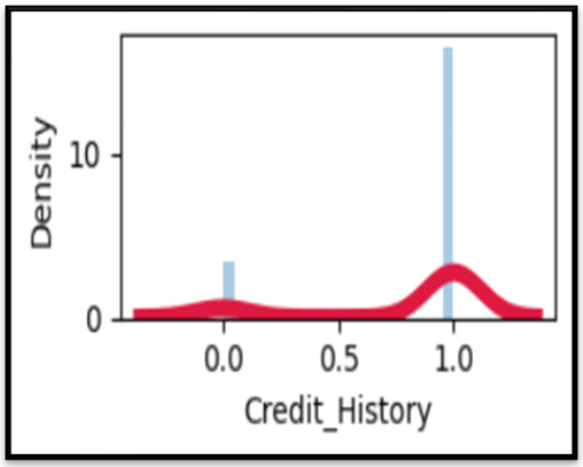
	0
Loan_ID	0
Gender	11
Married	0
Dependents	10
Education	0
Self_Employed	23
ApplicantIncome	0
CoapplicantIncome	0
LoanAmount	5
Loan_Amount_Term	6
Credit_History	29
Property_Area	0

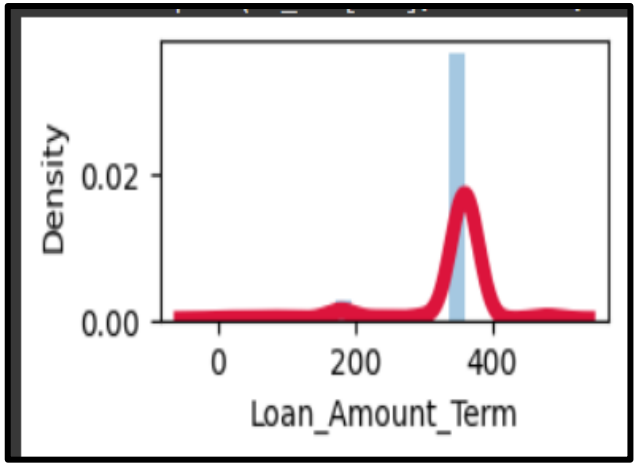
dtype: int64

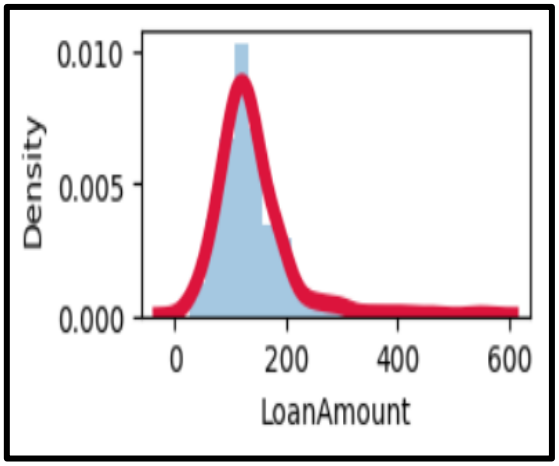
```
df.isnull().sum().sum()
```

```
84
```

DATA CLEANING & PRE-PROCESSING:







As you can see the column Credit_History , Loan_Amount_Term and LoanAmount are Skewed, in this case you will impute missing values with Median.

DATA CLEANING & PRE-PROCESSING:

```
[ ]  median_Credit_History = df['Credit_History'].median()
     median_Credit_History
```

```
1.0
```

```
[ ]  df['Credit_History'].fillna(median_Credit_History ,inplace=True)
```

```
Show hidden output
```

```
median_Loan_Amount_Term = df['Loan_Amount_Term'].median()
median_Loan_Amount_Term
```

```
360.0
```

```
[ ]  df['Loan_Amount_Term'].fillna(median_Loan_Amount_Term ,inplace=True)
```

```
Show hidden output
```

```
[ ]  median_LoanAmount = df['LoanAmount'].median()
     median_LoanAmount
```

```
125.0
```

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

```
Show hidden output
```

DATA CLEANING & PRE-PROCESSING:

- For the ‘Loan_Amount_Term’ attribute, which has 6 null value filling in the missing entries with the ‘median’ will help ensure data completeness. For the

- ‘Credit_History’ attribute, which has 29 null value filling in the missing entries with the ‘median’ will help ensure data completeness

- For the ‘LoanAmount’ attribute, which has 5null value filling in the missing entries with the ‘median’ will help ensure data completeness

DATA CLEANING & PRE-PROCESSING:

- You can fill the null values in the 'Dependents' and 'Self_Employed' columns with the mode:
- The mode is the most frequent value in a dataset. It's a suitable imputation method for categorical data, like 'Dependents' and 'Self_Employed', as it preserves the most common category.

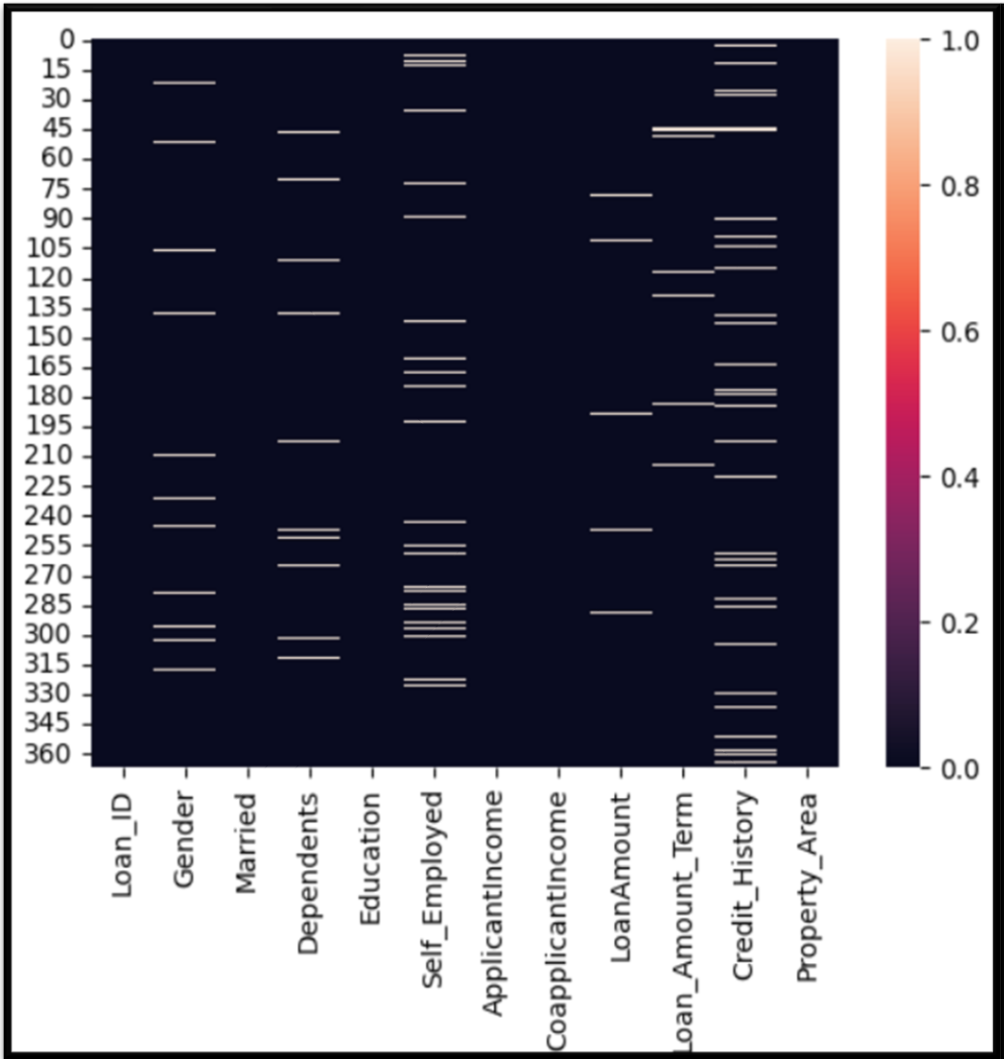
```
# Calculate the mode of the 'Dependents' column
dependents_mode = df['Dependents'].mode()[0]

# Fill null values in 'Dependents' with the mode
df['Dependents'].fillna(dependents_mode, inplace=True)

# Calculate the mode of the 'Self_Employed' column
self_employed_mode = df['Self_Employed'].mode()[0]

# Fill null values in 'Self_Employed' with the mode
df['Self_Employed'].fillna(self_employed_mode, inplace=True)
```

- For the 'Gender' attribute, which has 6 null value filling in the missing entries with the ' description are not given' will help ensure data completeness

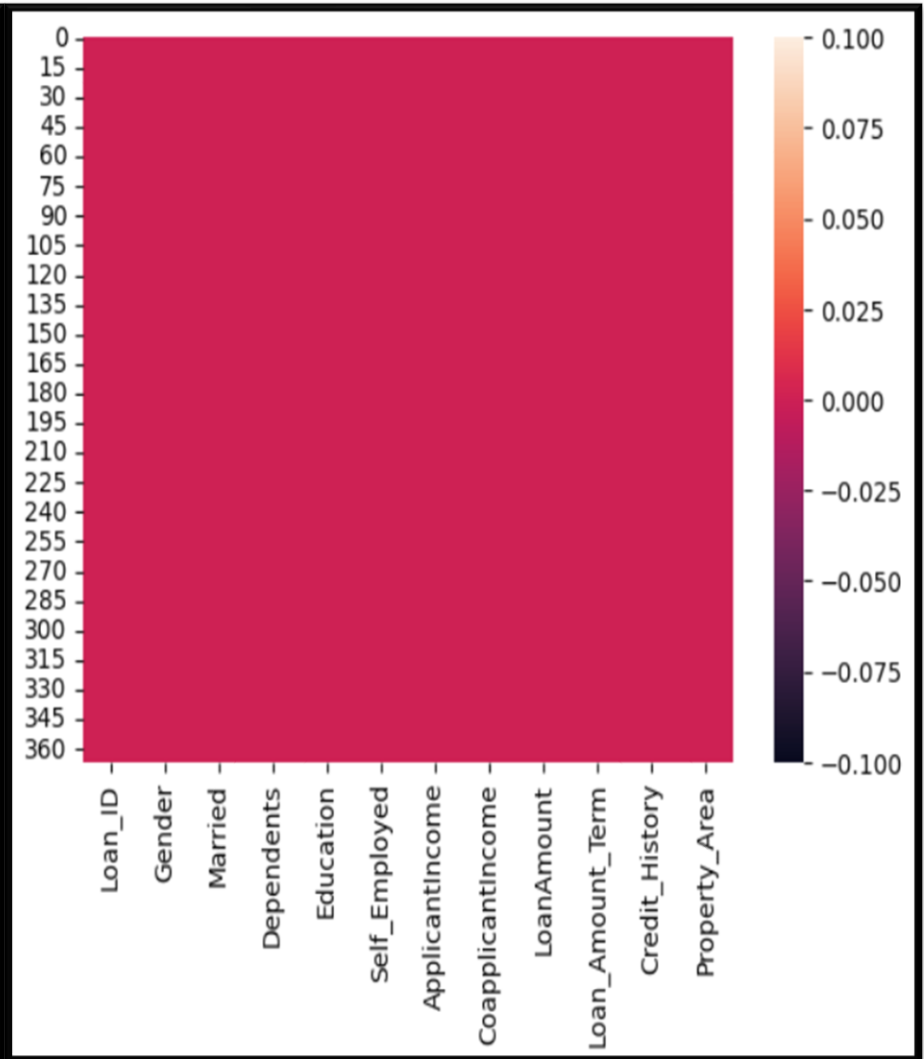
```
[18] df['Gender'].fillna('Gender are not given', inplace=True)
```

HEATMAPS

Before Cleaning



After Cleaning



REMOVING OUTLIERS

- After generating a box plot for the all numerical column we identified the presence of outliers in this column.

```
[ ]  numerical_cols = df.select_dtypes(include=['number']).columns
     numerical_cols
```

```
Index(['ApplicantIncome', 'CoapplicantIncome', 'LoanAmount',
       'Loan_Amount_Term', 'Credit_History'],
      dtype='object')
```

```
for col in numerical_cols:
        plt.figure(figsize=(8, 6))  # Adjust figure size as needed
        sns.boxplot(x=df[col])
        plt.title(f'Box Plot of {col}')
        plt.show()
```

REMOVING OUTLIERS

- To address these outliers, we will apply the IQR method.

- The code defines a function that uses the IQR method to identify and remove outliers from a specified column of a pandas DataFrame. It calculates the IQR, defines upper and lower bounds based on the IQR, and filters the DataFrame to keep only data points within those bounds

```python
def remove_outliers_iqr(df, column):
    """
    Removes outliers from a DataFrame column using the IQR method.

    Args:
        df: pandas DataFrame
        column: str, name of the column to remove outliers from

    Returns:
        pandas DataFrame with outliers removed
    """
    # Calculate quantiles
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1

    # Calculate upper and lower bounds
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Filter data
    df_filtered = df[(df[column] >= lower_bound) & (df[column] <= upper_bound)]

    return df_filtered
```

REMOVING OUTLIERS

This line of code in your notebook is using the function remove_outliers_iqr to remove outliers specifically from the all the numerical column of your Data Frame df.

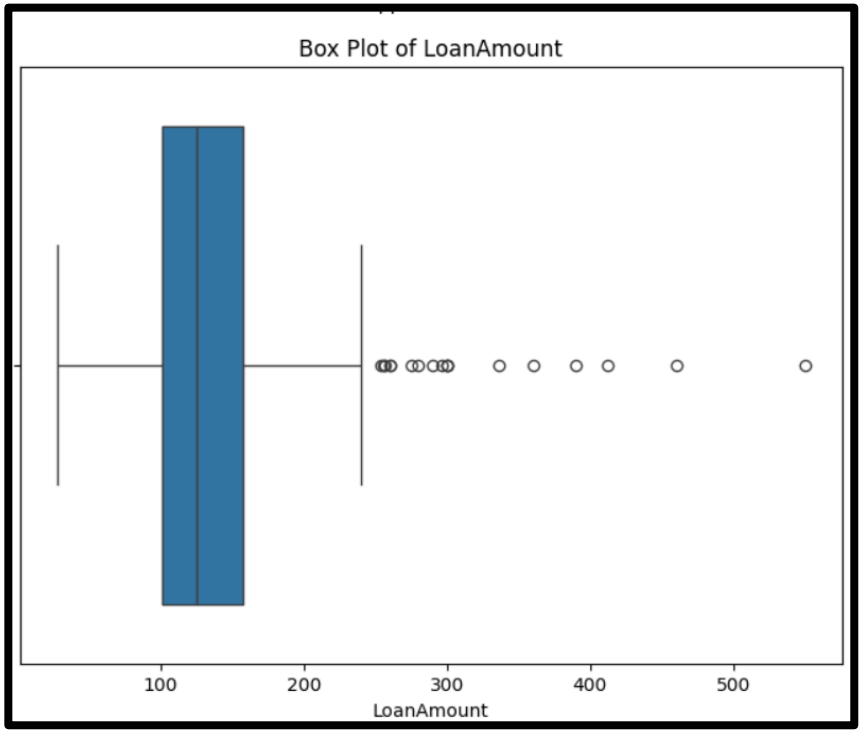
.

```
[ ]  df = remove_outliers_iqr(df, 'ApplicantIncome')

[ ]  df = remove_outliers_iqr(df, 'LoanAmount')

     df = remove_outliers_iqr(df, 'Loan_Amount_Term')

[ ]  df = remove_outliers_iqr(df, 'Credit_History')

[ ]  df = remove_outliers_iqr(df, 'CoapplicantIncome')
```

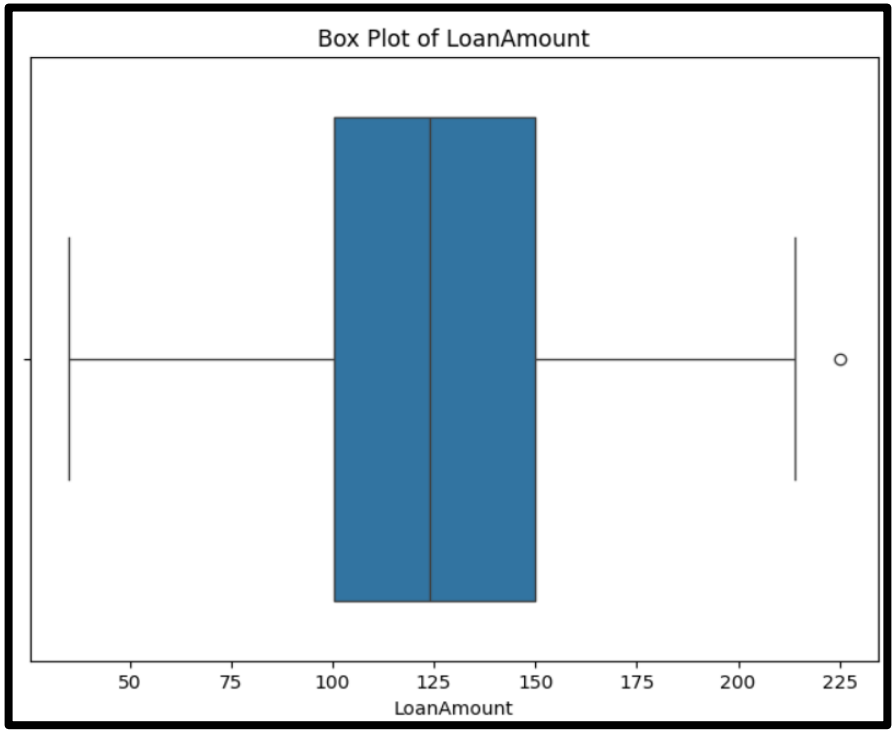
REMOVING OUTLIERS

- Take a example of Removing outliers by box plot the data frame

Before Removing



After Removing



DATA CLEANING & PRE-PROCESSING:

- **Summary –**

To summarize, addressing Null values and Outliers necessitates a methodical approach tailored to the data's characteristics and specific attributes. Data cleaning and Outlier handling are crucial steps for accurate analysis.

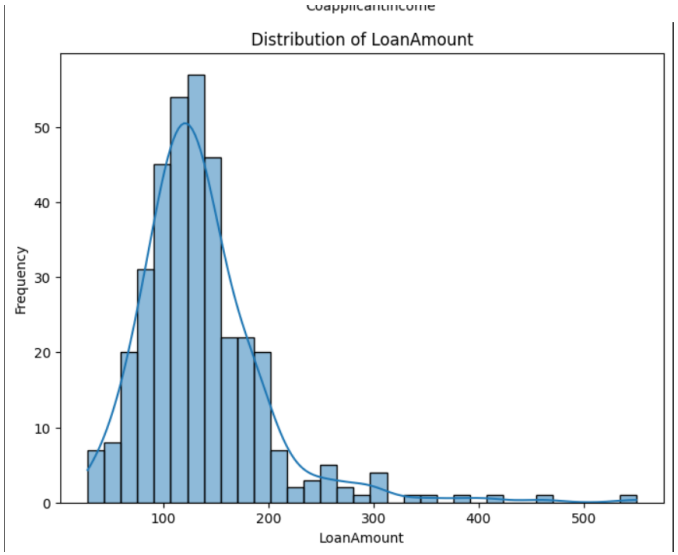
✓ The dataset contained Missing Values in “Gender”, “Dependents”, “Self_employed” , “LoanAmount”, “Loan_Amount_Term” , “Credit_History” features. These were handled by imputation (filling with median/mode) .

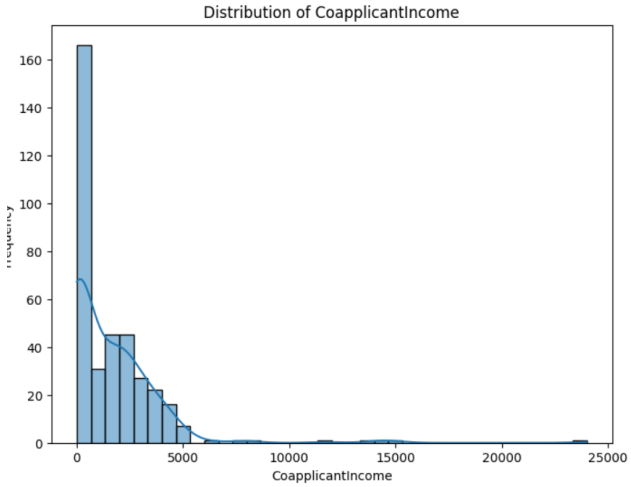
✓ Outliers were present in All the Numerical column . These were addressed using the IQR method and capping to boundary values.

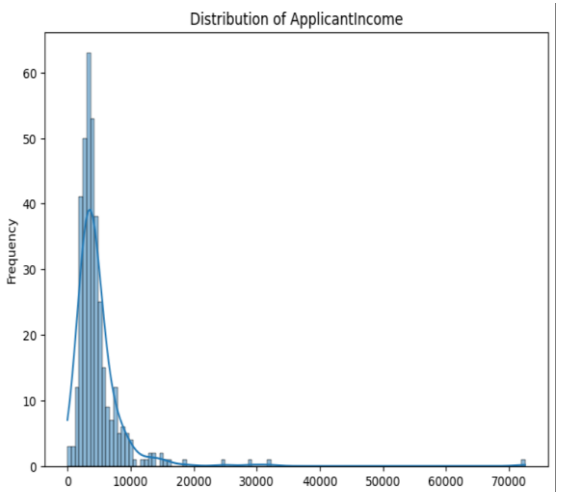
✓ With these Null, Missing, and Invalid values appropriately addressed, we are now ready to move forward with analyzing the dataset.

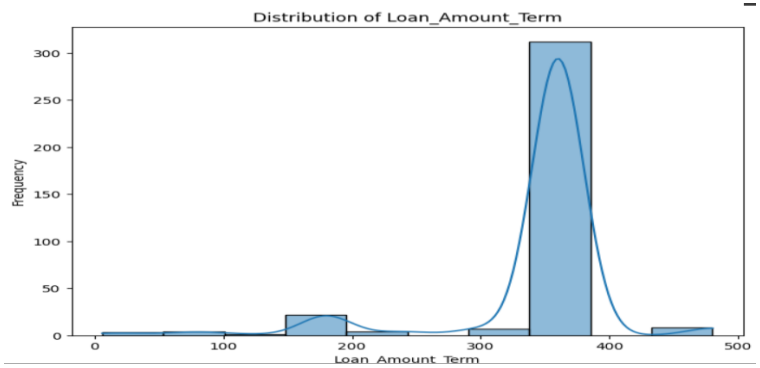
Data Visualization and Insights

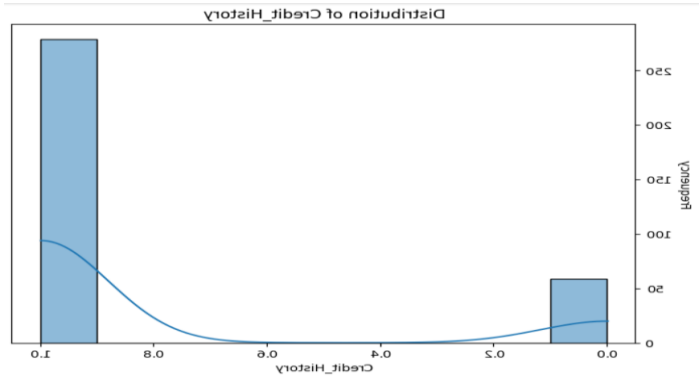
Histograms: Plot the frequency distribution of key numeric variables.











Data Visualization and Insights

- **Key Insights:**

- **Coapplicants :** The distribution is heavily right-skewed, meaning the majority of coapplicants have low incomes, clustered around the lower end of the range (between 0 and 5000). The number of coapplicants drastically drops as income increases.

- **Loan Amount:** The distribution is moderately right-skewed, with the majority of loans falling between 100 and 200 units (likely thousands of currency units), indicating this is a common loan size range. Most of the loans cluster around the 150-200 range, with the frequency peaking around these values. This suggests that loan requests in this range are the most typica

Data Visualization and Insights

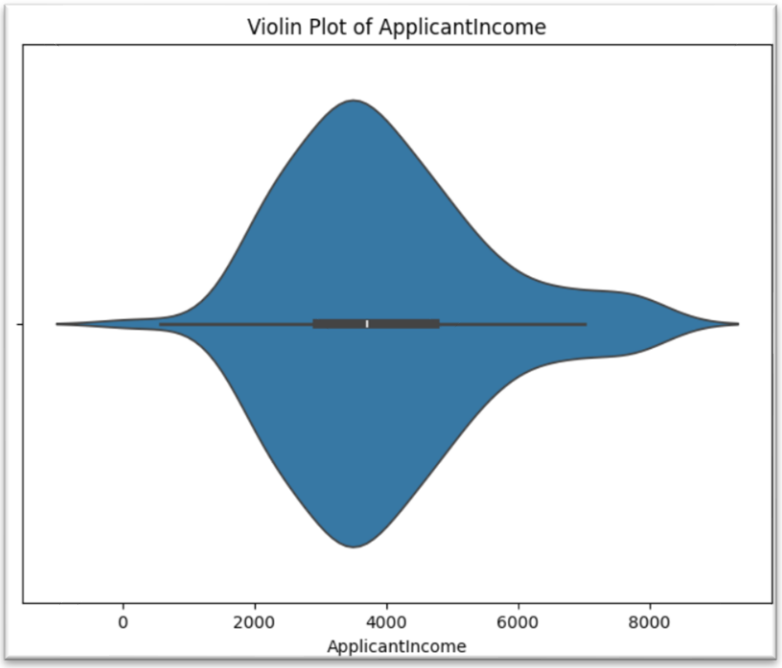
- **Loan Amount Term:** The majority of loan terms are concentrated at 360 months (or 30 years), with the highest frequency being over 300. This suggests that most loans are long-term, likely for mortgages or large financial commitments.

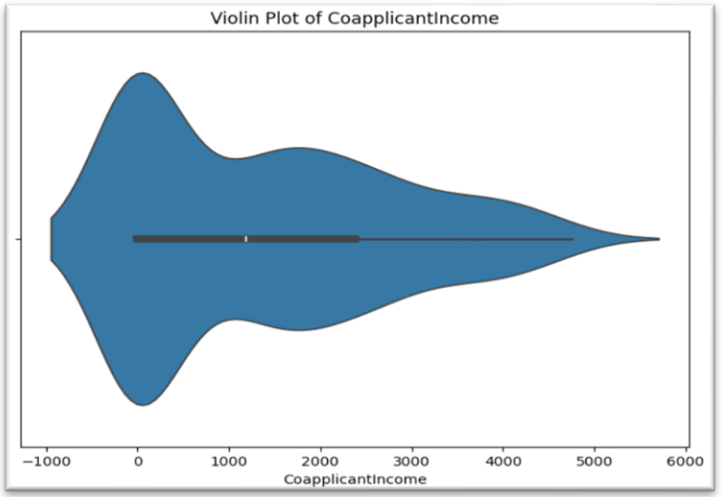
- **ApplicantIncome:** will reveal the distribution of income among loan applicants. You can observe whether the distribution is skewed, symmetric, or has any other distinct shape. This helps you understand the general income levels of the applicants.

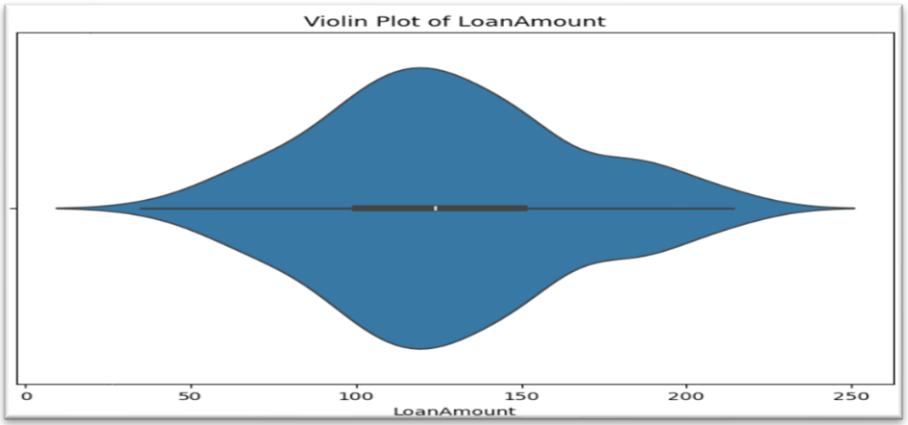
- **Credit_History :** histogram provides a crucial overview of the creditworthiness of loan applicants. It shows the distribution of applicants with different credit histories, typically represented as a binary value (0 or 1).

Data Visualization and Insights

Violin_Plot _Plots: Identify potential outliers and visualize the spread of data







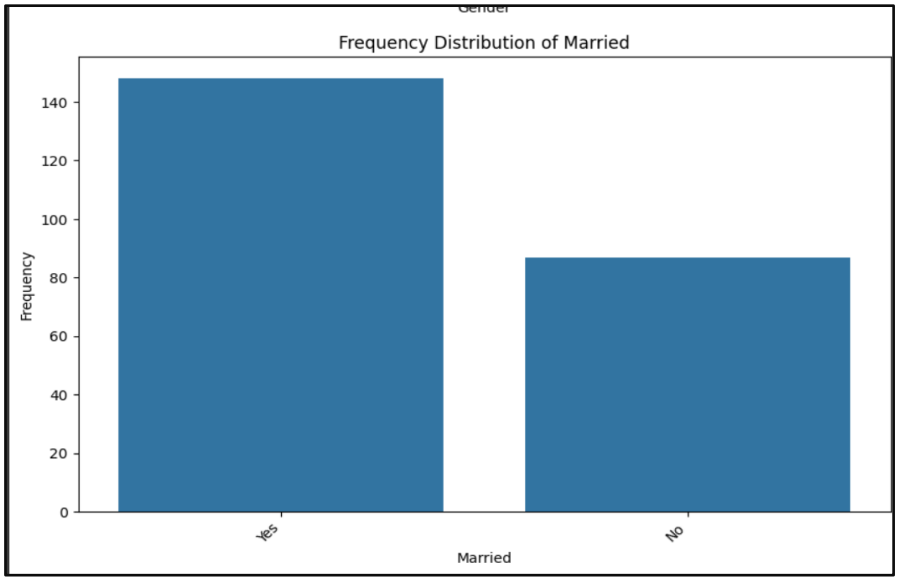
Data Visualization and Insights

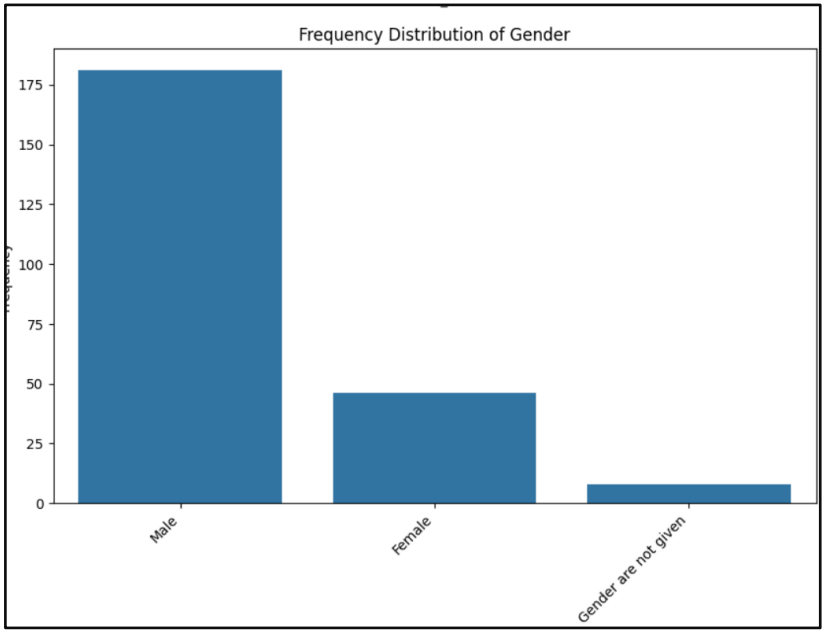
- **Key insights:**

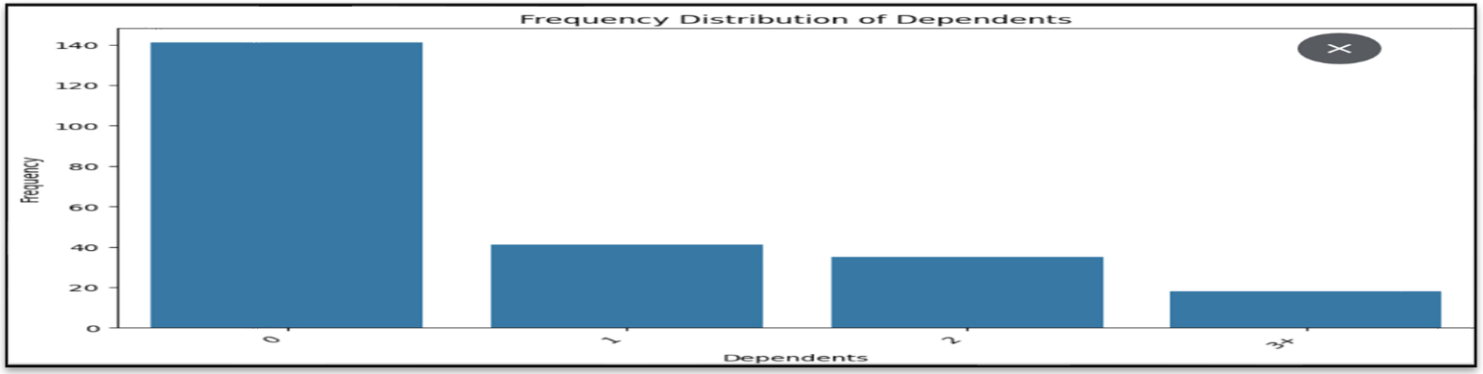
- **ApplicantIncome :** The violin plot for ApplicantIncome likely shows a right-skewed distribution. This means that the majority of applicants have lower incomes, while a smaller number of applicants have significantly higher incomes. The wider section of the violin plot at the lower income range indicates a higher density of data points in that region.
- **CoapplicantIncome :**The violin plot for CoapplicantIncome might display a distribution with a peak near zero and a long tail extending towards higher incomes. This suggests that a significant portion of loan applicants do not have a co-applicant, resulting in zero co-applicant income. The tail indicates the presence of some applicants with co-applicants who have substantial incomes.
- **LoanAmount :** The violin plot for LoanAmount might show a slightly right-skewed distribution, indicating that most loan amounts are concentrated towards the lower end, with a smaller number of loans having higher values. The wider section of the violin plot towards lower loan amounts signifies a higher density of data points in that range.

Data Visualization and Insights

Bar Charts: Visualize the frequency distribution of categorical variables.

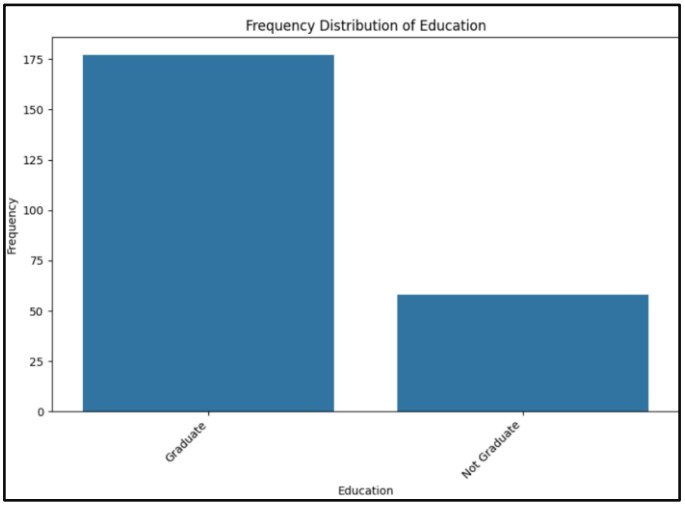


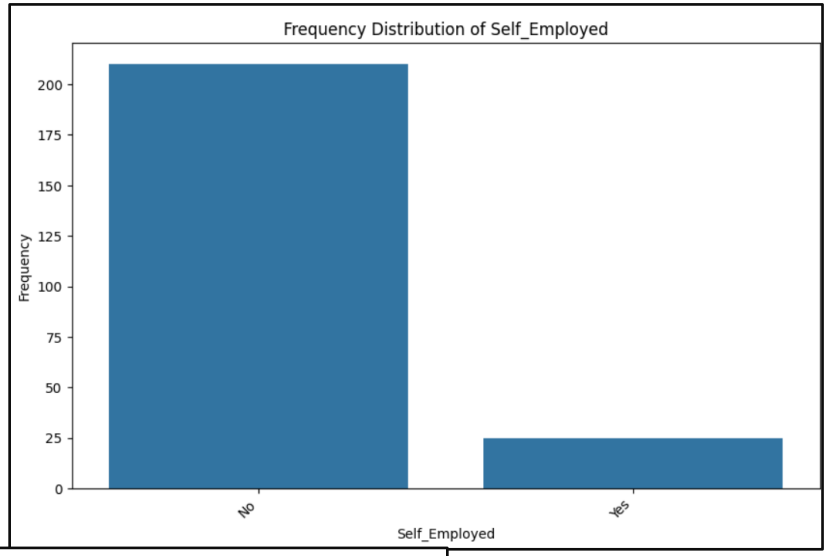


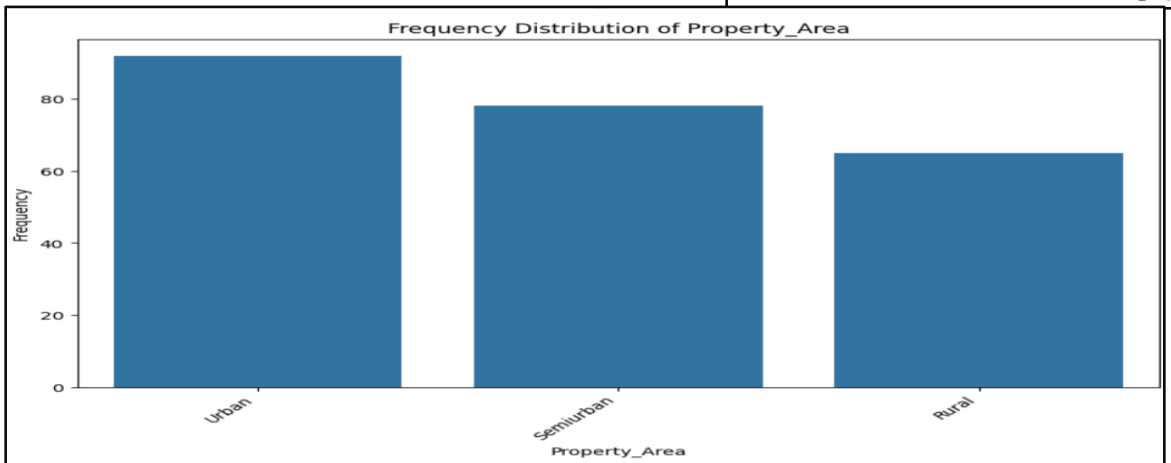


Data Visualization and Insights

Bar Charts: Visualize the frequency distribution of categorical variables.







Data Visualization and Insights

- **Key insights:**

- **Gender :** By comparing the heights of the bars, you can see which gender category has more loan applications. For example, if the 'Male' bar is significantly higher than the 'Female' bar, it indicates that there are more male applicants in the dataset.

- **Married :** By comparing the heights of the bars, you can see which marital status category has more loan applications. For example, if the 'Yes' bar is higher than the 'No' bar, it indicates that there are more married applicants in the dataset.

- **Distribution of Dependents :** By comparing the heights of the bars, you can see which category of dependents is most common among loan applicants. For example, if the bar for '0' dependents is the highest, it indicates that a large proportion of applicants do not have any dependents

Data Visualization and Insights

- **Key insights:**

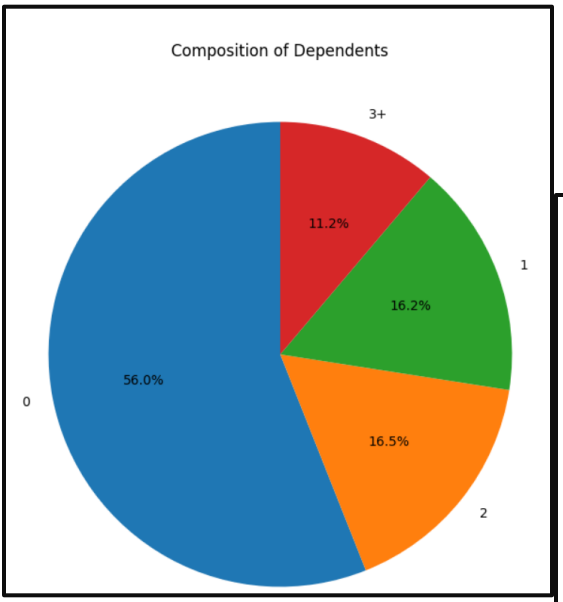
- **Education :** Comparing the heights of the bars, you can see which education level is more prevalent among loan applicants. For example, if the 'Graduate' bar is significantly higher than the 'Not Graduate' bar, it indicates that a larger proportion of applicants are graduates.

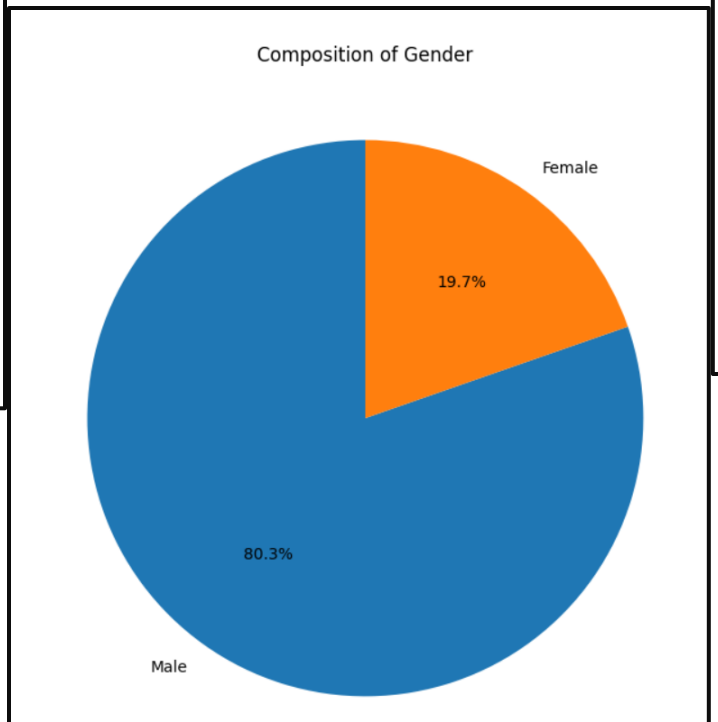
- **Self-Employed :** Comparing the heights of the bars, you can see which employment status is more prevalent among loan applicants. For example, if the 'No' bar is significantly higher than the 'Yes' bar, it indicates that a larger proportion of applicants are not self-employed.

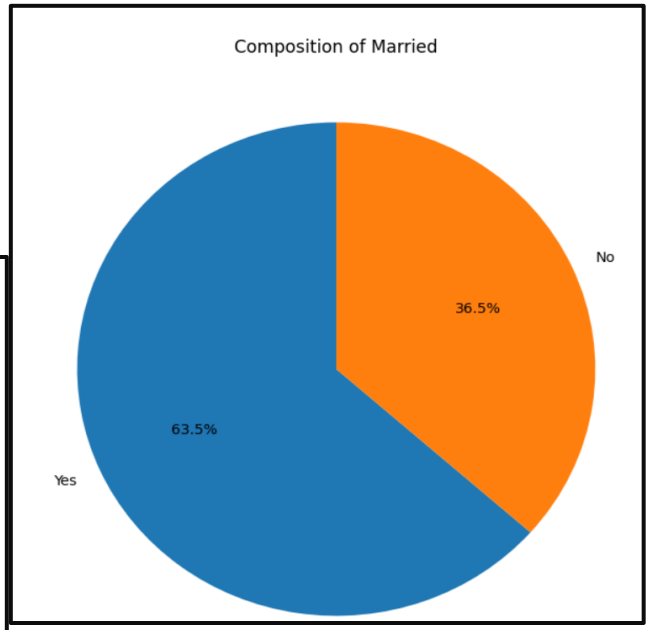
- **Property Area :** Comparing the heights of the bars, you can see which property area has the most loan applications. For example, if the 'Semiurban' bar is the highest, it indicates that a large proportion of applicants reside in semiurban areas.

Data Visualization and Insights

Pie Charts: Represent the composition of categorical variables.

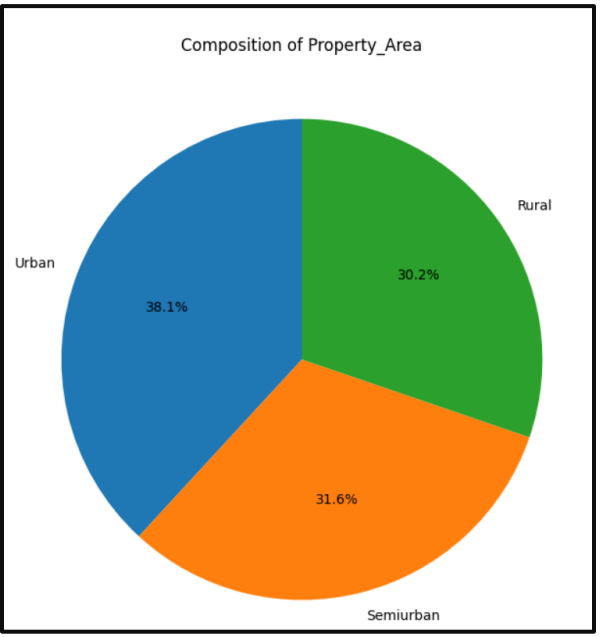


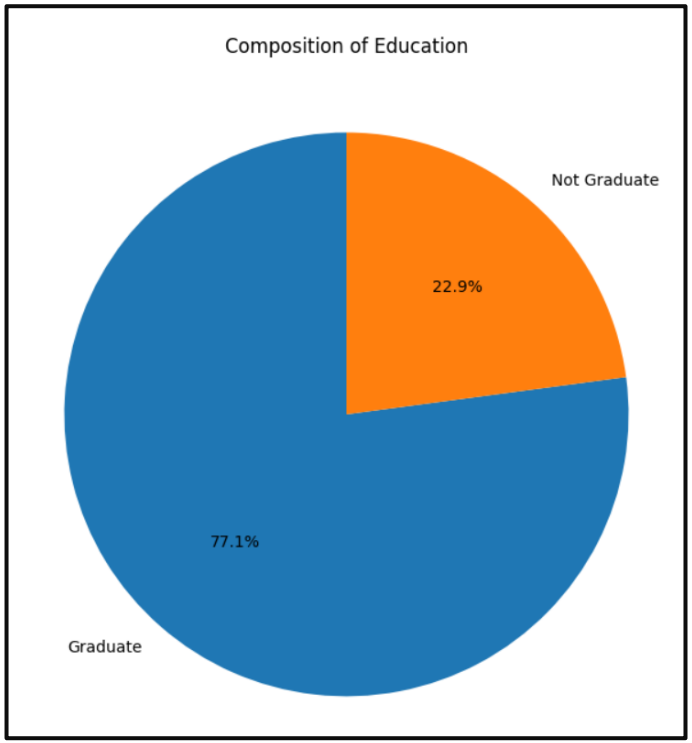


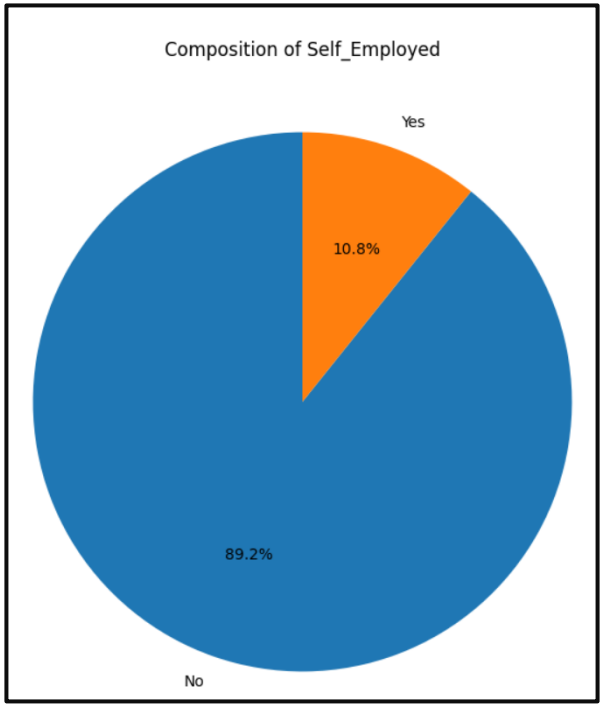


Data Visualization and Insights

Pie Charts: Represent the composition of categorical variables.







Data Visualization and Insights

- **Key insights:**
- **Gender :** The pie chart for the "Gender" column helps you understand the distribution of genders within the loan application data, identify the dominant gender, assess the proportion of missing gender information, explore potential relationships with other variables, and consider the impact of gender on loan sanction decisions.
- **Married :** The pie chart for the "Married" column provides a visual representation of the marital status distribution among loan applicants. It allows you to identify the dominant marital status, assess its potential impact on loan decisions, explore relationships with other variables, and consider any potential cultural or financial implications.
- **Dependents :** The pie chart for the "Dependents" column provides a visual representation of the distribution of dependents among loan applicants. It allows you to identify the dominant dependency category, assess its potential impact on loan decisions, explore relationships with other variables, and consider any potential cultural or financial implications.

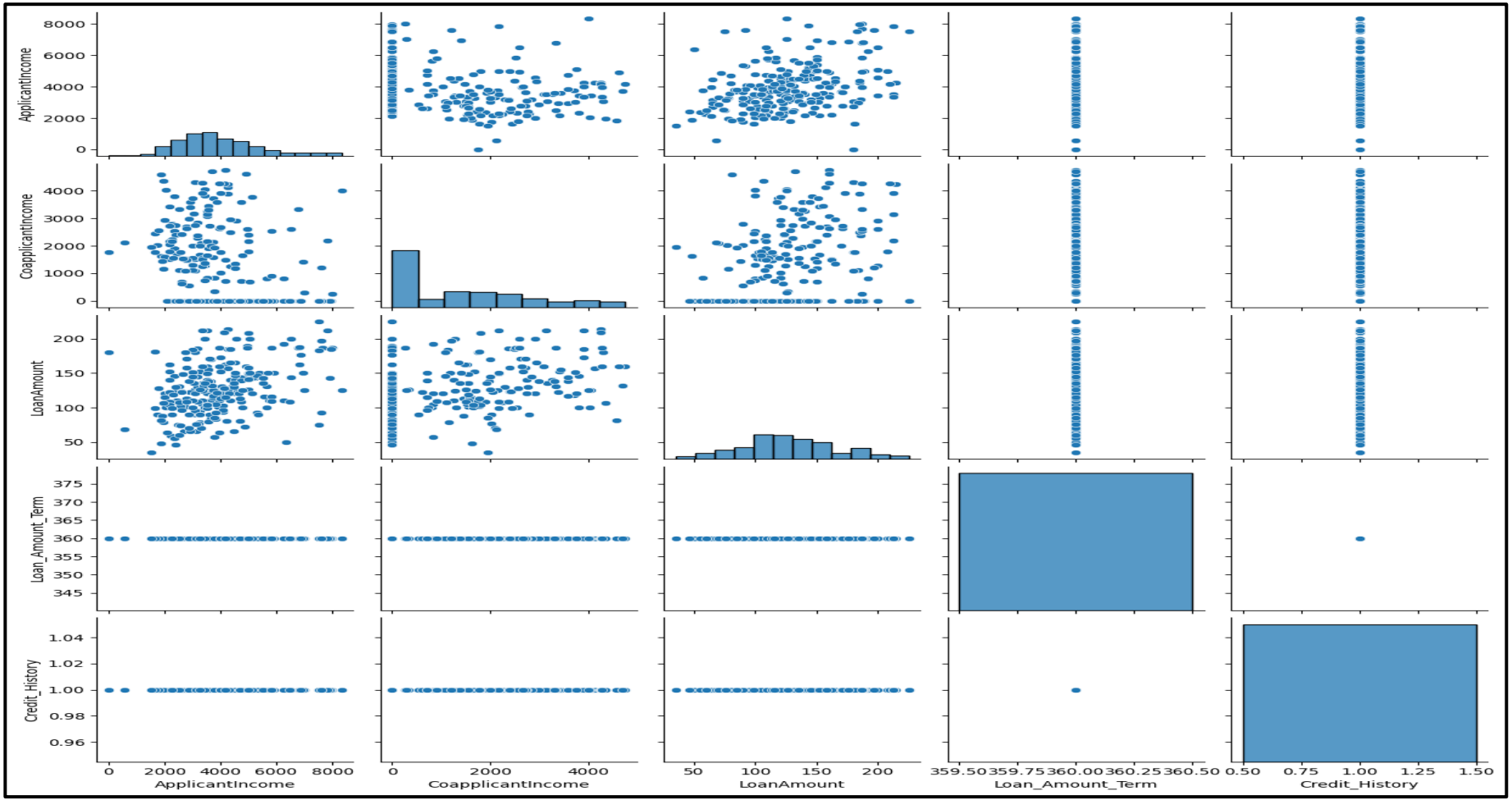
Data Visualization and Insights

- **Key insights:**

- **Education :** The pie chart for the "Education" column provides a visual representation of the education level distribution among loan applicants. It allows you to identify the dominant education level, assess its potential impact on loan decisions, explore relationships with other variables, and consider any potential socioeconomic implications.
- **Self-Employed :** The pie chart for the "Self-Employed" column provides a visual representation of the employment type distribution among loan applicants. It allows you to identify the dominant employment type, assess its potential impact on loan decisions, explore relationships with other variables, and consider any potential financial risk implications.
- **Property_Area :** The pie chart for the "Property_Area" column provides a visual representation of the geographical distribution of loan applicants. It allows you to identify the dominant property location, assess its potential impact on loan decisions, explore relationships with other variables, and consider any potential regional or socioeconomic implications.

Data Visualization and Insights

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



Data Visualization and Insights

Key insights:

- **ApplicantIncome vs. LoanAmount:**
- There is a positive trend between ApplicantIncome and LoanAmount, indicating that higher applicant incomes might be associated with higher loan amounts.
- **CoapplicantIncome vs. LoanAmount:**
- A weak relationship is visible. However, many coapplicants have income close to zero, suggesting single applicants or coapplicants with no reported income.
- **Distribution Patterns:**
- ApplicantIncome and LoanAmount have right-skewed distributions.
- CoapplicantIncome also shows skewness with many zeros.
- Loan_Amount_Term and Credit_History are categorical-like and have discrete values.

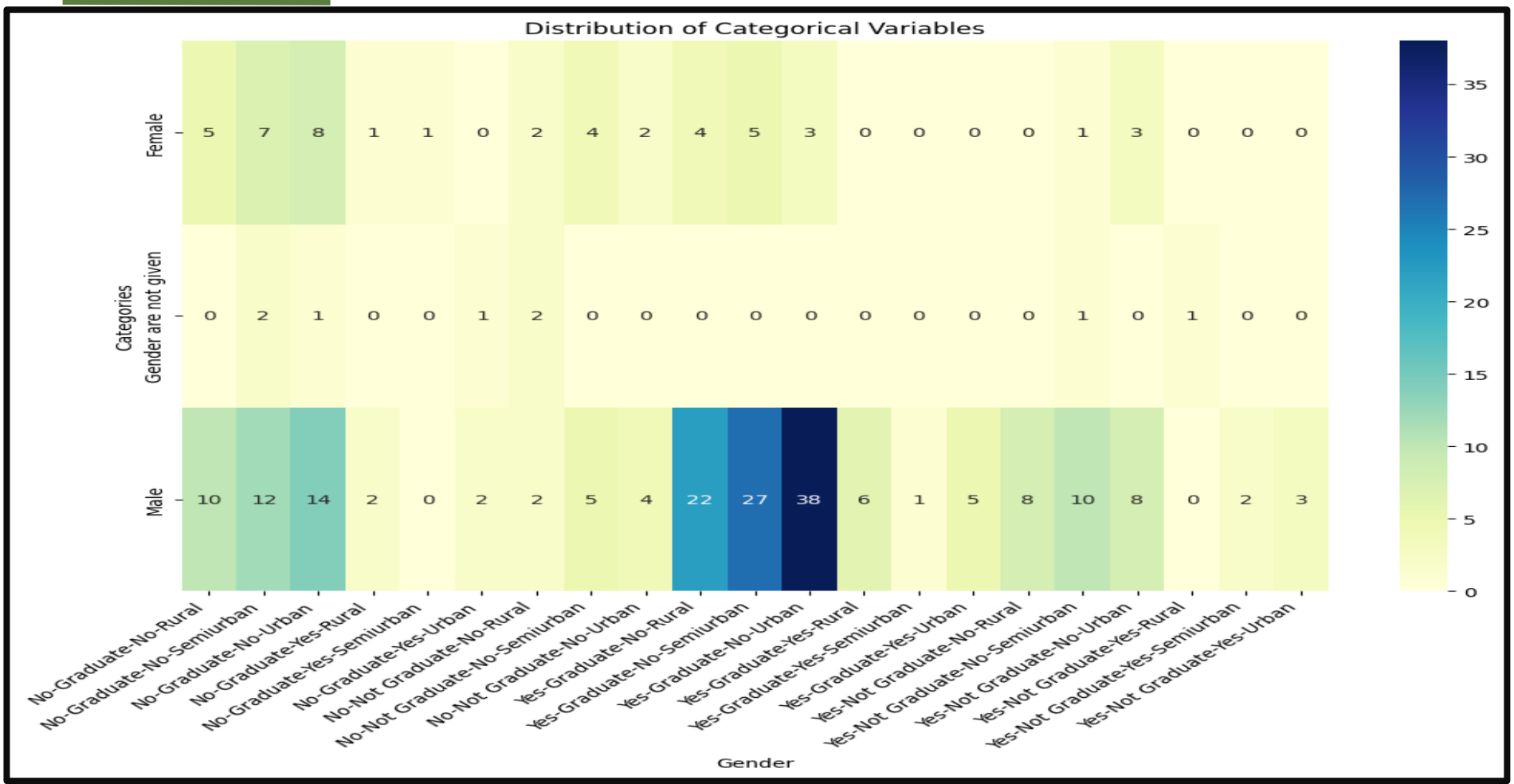
Data Visualization and Insights

- **Key insights:**

- **Credit_History:**
- Credit history appears to have two dominant values, likely representing binary categories (e.g., good vs. bad credit history).
- **Loan_Amount_Term:**
- Most loans have a similar term, with very little variation.
- **Outliers:**
- There may be outliers in ApplicantIncome and LoanAmount, which could affect analysis and model performance.
- **Key Patterns:**
- Credit history may influence loan approval but doesn't correlate significantly with other continuous variables.
- The relationship between income variables (ApplicantIncome and CoapplicantIncome) and LoanAmount may help predict loan amounts

Data Visualization and Insights

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



Data Visualization and Insights

- **Key insights:**

- **Gender and Education:**
- Males are more represented across all categories compared to females.
- Females have relatively fewer counts in all combinations of education and area categories, indicating possible gender disparity in data representation.
- **Graduation Status:**
- "Yes-Graduate" categories dominate for both males and females, especially in urban and semi-urban areas, which might indicate that graduates are more likely to apply for loans or be included in this dataset.
- **Urban vs. Rural:**
- Urban and semi-urban areas show higher counts for both males and females compared to rural areas.
- This trend is most pronounced in "Yes-Graduate" males in semi-urban and urban areas.

Data Visualization and Insights

•Key insights:

•Gender are not given:

•A small number of entries have missing or undefined gender information, which may need attention during data cleaning.

•Category Counts:

•The highest counts are observed for "Male-Yes-Graduate-Semiurban," suggesting that this demographic is most frequent in the dataset.

FINAL REPORT

Summarizing the key findings, drawing conclusions, and offering recommendations based on the insights obtained from the analysis:

❑Distribution of Numeric Variables :

Applicant Income and Co-applicant Income are right-skewed, which signifies a predominance of lower incomes with a few high earners. The Loan Amount exhibits a generally normal distribution, albeit with some outliers. The Loan Amount Term is mainly clustered around 360 months. Credit History is negatively skewed, implying that most applicants have a credit history. The distribution of Dependents shows that a larger proportion of applicants have no dependents.

FINAL REPORT

Summarizing the key findings, drawing conclusions, and offering recommendations based on the insights obtained from the analysis:

❑Categorical Variable Analysis :

The majority of applicants are male and married. Most applicants are graduates and not self-employed. The distribution of Property Area is fairly balanced across urban, semi-urban, and rural regions.

FINAL REPORT

Summarizing the key findings, drawing conclusions, and offering recommendations based on the insights obtained from the analysis:

❑Relationships between Variables:

There is a positive correlation between Applicant Income and Loan Amount, indicating that applicants with higher incomes tend to request larger loans. There is a weak positive correlation between Co-applicant Income and Loan Amount. There is no strong linear relationship between Loan Amount and Loan Amount Term. Box plots reveal variations in Loan Amount distribution across different categories such as Gender, Marital Status, Education, SelfEmployed status, and Property Area.

FINAL REPORT

Summarizing the key findings, drawing conclusions, and offering recommendations based on the insights obtained from the analysis:

❑CONCLISIONS:

➢ Income and Loan Amount: Applicant income significantly influences the amount of the loan they can receive.
➢ Demographic Factors: Factors such as gender, marital status, education, and employment status impact the characteristics of loan applications.
➢ Credit History: The credit history of applicants affects their loan application outcomes

FINAL REPORT

Summarizing the key findings, drawing conclusions, and offering recommendations based on the insights obtained from the analysis:

❑Recommendations :

➢ Target Marketing: Customize loan products and marketing strategies according to income levels and demographic characteristics.
➢ Risk Assessment: Evaluate income, credit history, and other factors to inform loan approval and risk assessment.
➢ Product Diversification: Provide loan products with different terms and amounts to meet diverse customer needs.
➢ Further Analysis: Investigate additional factors and their interactions to enhance insights and improve decision-making.


THANK YOU FOR READING
For coding part, kindly refer to below link :
https://colab.research.google.com/drive/1_eu97_furgwavLAL4u7EUcCneLdCsfZf?usp=sharing