# LOAN APPROVAL ANALYSIS

**Presented by Fardeen Ansari**

LinkedIn

fardeenansari203@gmail.com

# INDEX

# About This Project

This project focuses on understanding a Loan Approval Dataset. The goal is to explore the data and find clear insights about what factors influence loan approval. Basic analysis and charts.

The dataset includes details like gender, education, income, loan amount, property area, and loan approval status. We first check and clean the data by handling missing values.

Then, we perform Exploratory Data Analysis (EDA) using charts such as histograms, bar graphs, pie charts, box plots, and scatter plots. These visuals help us see how the data is distributed, how variables are related, and whether there are any unusual values.

Overall, this project helps us understand the key patterns in the loan data and identify which factors play an important role in loan approval.

# Tools Used





# Aim

This project used a Loan Approval Dataset and followed a clear and organized data analysis process. It started with collecting the dataset and cleaning it by fixing missing values and correcting any wrong or inconsistent data. After cleaning, Exploratory Data Analysis (EDA) was performed to find patterns and understand the relationships between different features. Data visualization tools like Seaborn and Matplotlib were used to create clear and meaningful charts. In the final step, important insights and conclusions were drawn to identify the key factors that affect whether a loan application gets approved.

DATA COLLECTOIN

DATA CLEANING

EDA

DATA VISUALIZATION

# Objectives

1. To understand the loan approval data
2. See what information is available about the people who applied for loans.
3. To clean the data
4. Fix missing or wrong values so the data becomes easy to work with.
5. To find out what factors matter
6. Check which details—like income, loan amount, education, or credit history—affect loan approval.
7. To use simple charts and graphs
8. Show the data in an easy way so patterns are clear.
9. To understand the relationships
10. See how different factors are connected to each other and to loan approval.
11. To get clear insights
12. Understand why some people get their loan approved and why others do not.
13. To explain everything clearly
14. Present the findings in simple language so anyone can understand the results.

# Inside the Dataset

This project uses a Loan Approval Dataset that has basic information about people who applied for a loan. It includes their personal details, income, loan amount, credit history, and where their property is located. Here is a very simple meaning of each column:

Loan_ID: A special number given to each loan application.

1. Gender: If the person is Male or Female.
2. Married: If the person is married or not.
3. Dependents: How many people depend on the person (like children or family members).
4. Education: If the person is a Graduate or Not Graduate.
5. Self_Employed: If the person works for themselves (Yes/No).
6. ApplicantIncome: The monthly income of the main person applying for the loan.
7. CoapplicantIncome: Monthly income of the second person applying, if there is one.

# Inside the Dataset

8. LoanAmount: How much loan the person wants ( in thousands).
9. Loan_Amount_Term: How many months the person will take to repay the loan.
10. Credit_History: If the person has a good credit record (1 = good, 0 = bad/no record)
11. Property_Area: The type of area where the property is located — Urban, Semiurban, or Rural.

# Data Overviwe

```
# Load the Dataset
df = pd.read_csv('loan_sanction_test.csv')
# Display the first few row of the dataset
df.head()
```

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

```
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
```

```
# Check how many rows and columns in the dataset
df.shape
```

```
(367, 12)
```

# Handle Missing Values

```
# Check the missing values in the dataset
df.isnull().sum()
```

```
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
```

- To check how many missing values are present in the dataset, use df.isnull().sum()

```
# Fill Missing Values with Mode and Median
df['Self_Employed'] = df['Self_Employed'].fillna(df['Self_Employed'].mode()[0])
df['Gender'] = df['Gender'].fillna(df['Gender'].mode()[0])
df['LoanAmount'] = df['LoanAmount'].fillna(df['LoanAmount'].median())
df['Credit_History'] = df['Credit_History'].fillna(df['Credit_History'].median())
df['Dependents'] = df['Dependents'].fillna(df['Dependents'].mode()[0])
df['Loan_Amount_Term'] = df['Loan_Amount_Term'].fillna(df['Loan_Amount_Term'].median())
```

- Fill the missing values using mode for categorical columns and median for numerical columns

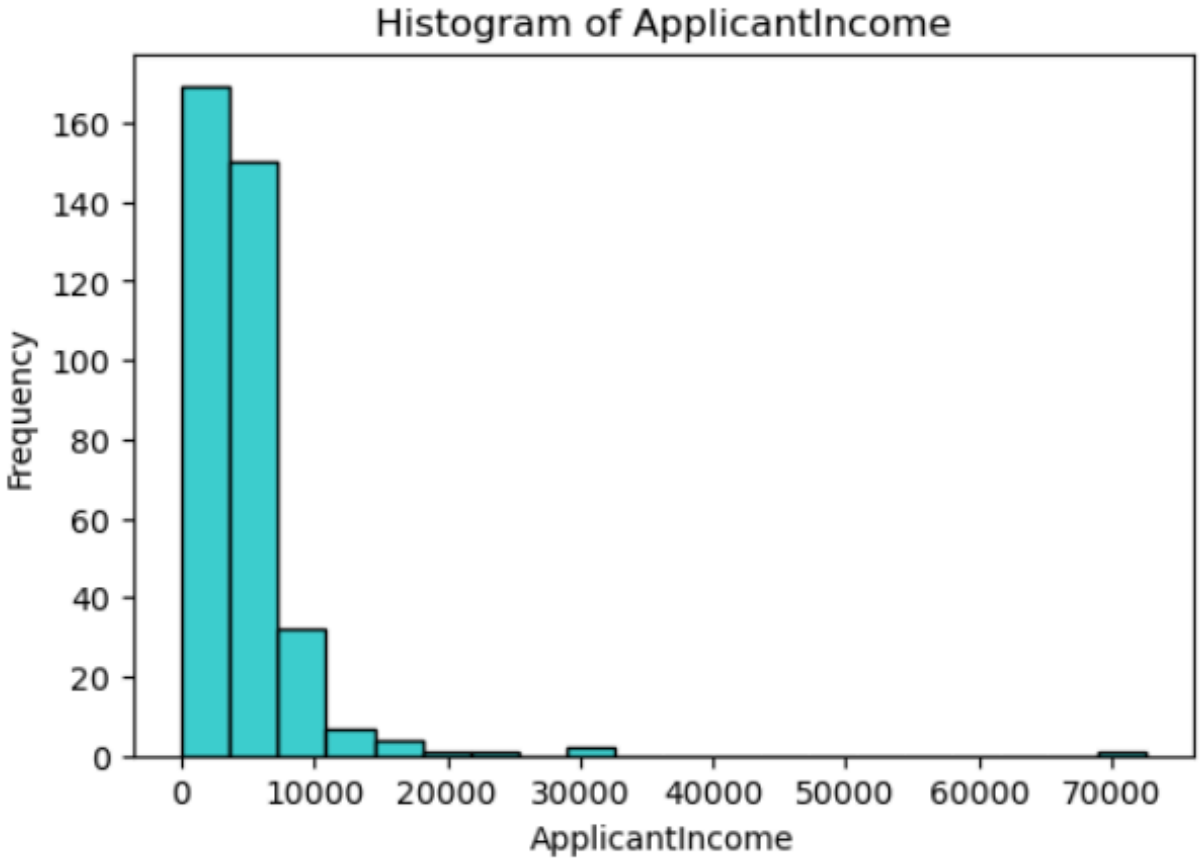
# Basic Statistics

```
df.describe()
```

| | ApplicantIncome | CoapplicantIncome | LoanAmount | Loan_Amount_Term | Credit_History |
|---|---|---|---|---|---|
| count | 367.000000 | 367.000000 | 367.000000 | 367.000000 | 367.000000 |
| mean | 4805.599455 | 1569.577657 | 135.980926 | 342.822888 | 0.839237 |
| std | 4910.685399 | 2334.232099 | 60.959739 | 64.658402 | 0.367814 |
| min | 0.000000 | 0.000000 | 28.000000 | 6.000000 | 0.000000 |
| 25% | 2864.000000 | 0.000000 | 101.000000 | 360.000000 | 1.000000 |
| 50% | 3786.000000 | 1025.000000 | 125.000000 | 360.000000 | 1.000000 |
| 75% | 5060.000000 | 2430.500000 | 157.500000 | 360.000000 | 1.000000 |
| max | 72529.000000 | 24000.000000 | 550.000000 | 480.000000 | 1.000000 |

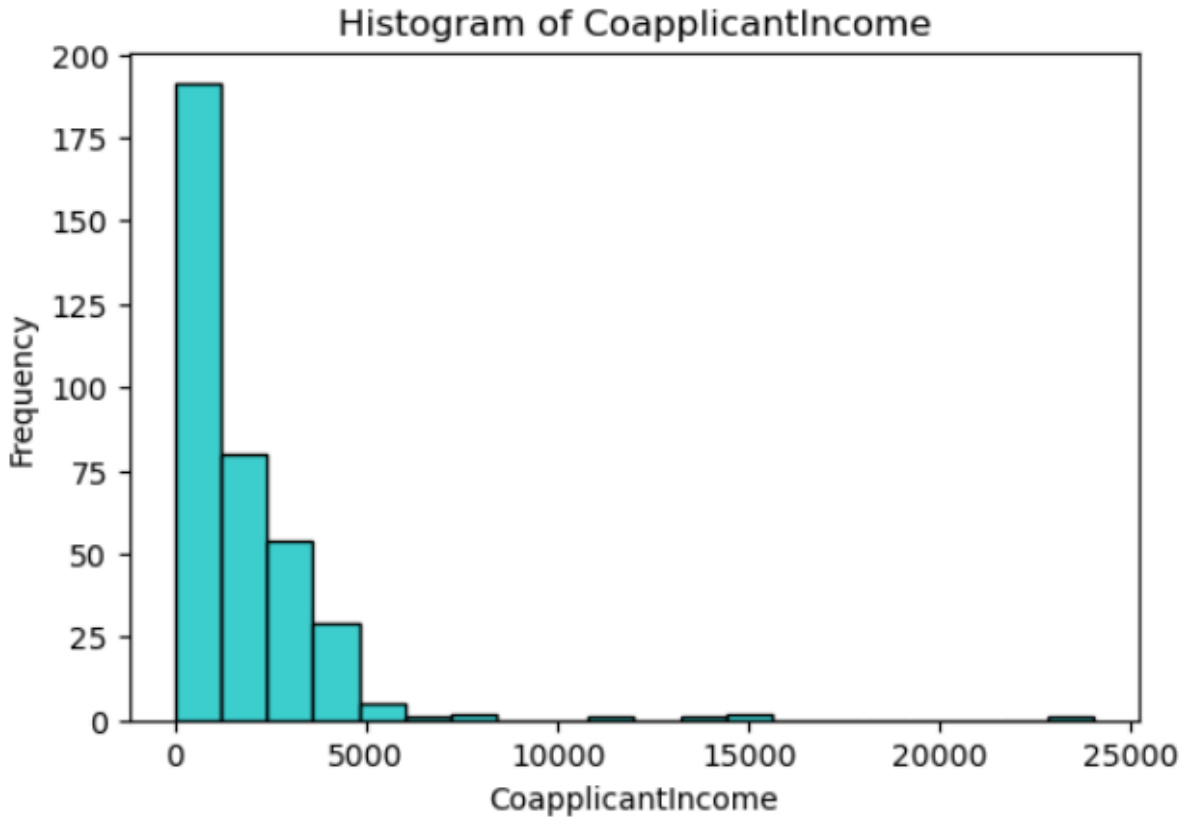
The numerical summary of the Loan Approval Dataset shows that applicant incomes vary widely, with an average of around ₹4,805. Co-applicant income is much lower on average, suggesting many applicants applied alone or with low-earning co-applicants. The average loan amount is about ₹136,000, and most loan terms are 360 months. The credit history mean of 0.83 indicates that most applicants had a positive credit record. Overall, the dataset shows high income diversity, a common long loan term, and generally good credit reliability.

# Frequency Distribution of Key Variables



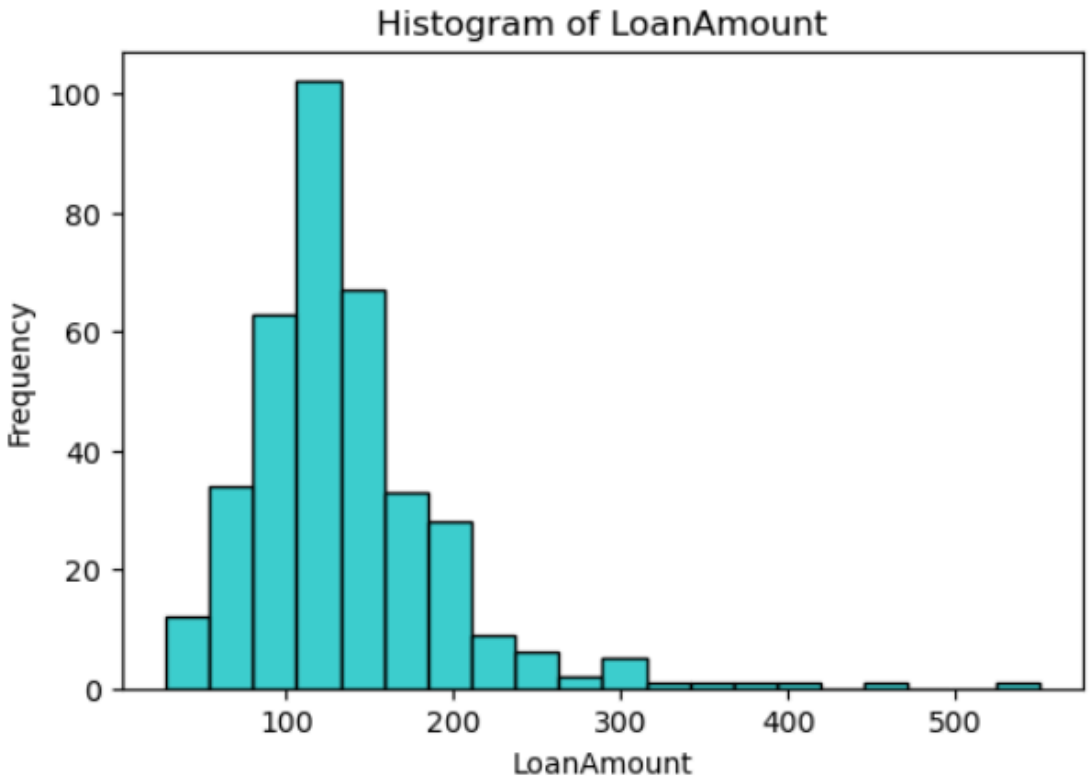

The histogram of ApplicantIncome shows that most applicants have low incomes, mainly below ₹10,000. Only a few people have very high incomes, which creates a long tail on the right side. The distribution is right-skewed, meaning the income values are uneven and mostly concentrated among lower-income applicants.

# Frequency Distribution of Key Variables



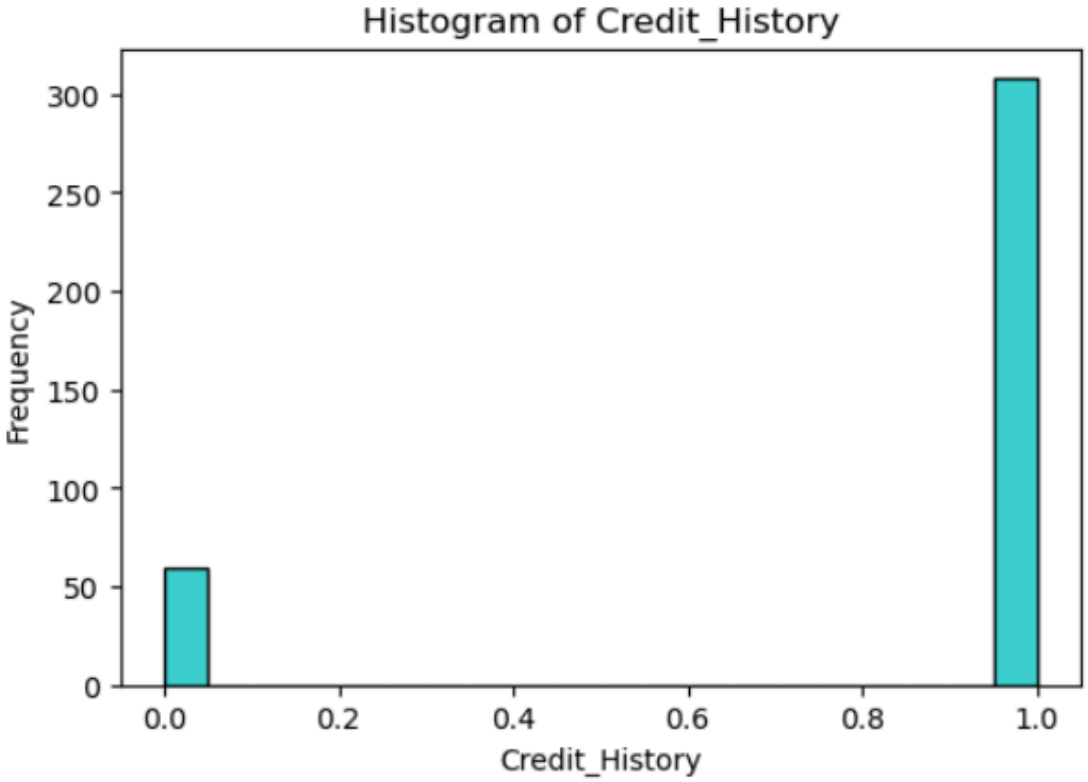

The histogram of CoapplicantIncome shows that most co-applicants have very low income, with the majority earning below ₹2,000. As income increases, the number of co-applicants decreases sharply, and only a few have high incomes above ₹10,000. The distribution is right-skewed, meaning most values are low while a small number of high-income outliers stretch the graph to the right

# Frequency Distribution of Key Variables



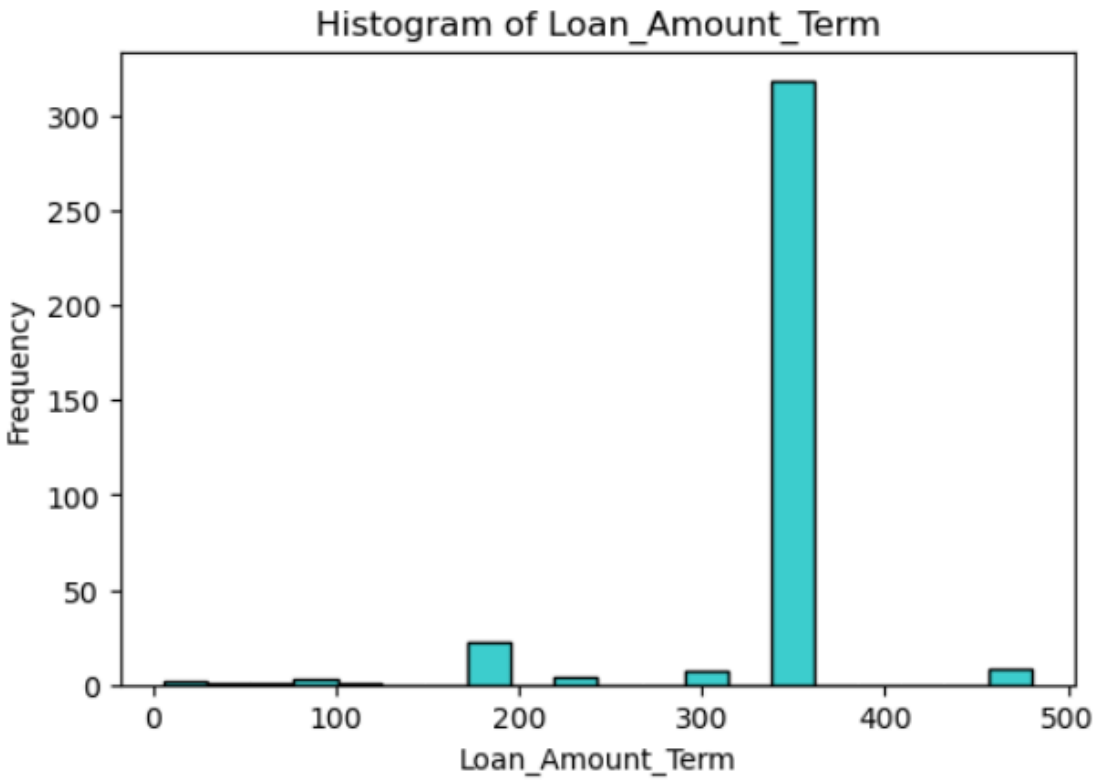

The LoanAmount histogram shows that most loan amounts are between 100 and 200, with a peak around 120–150. Very few applicants take large loans above 300, making the distribution right-skewed.

# Frequency Distribution of Key Variables



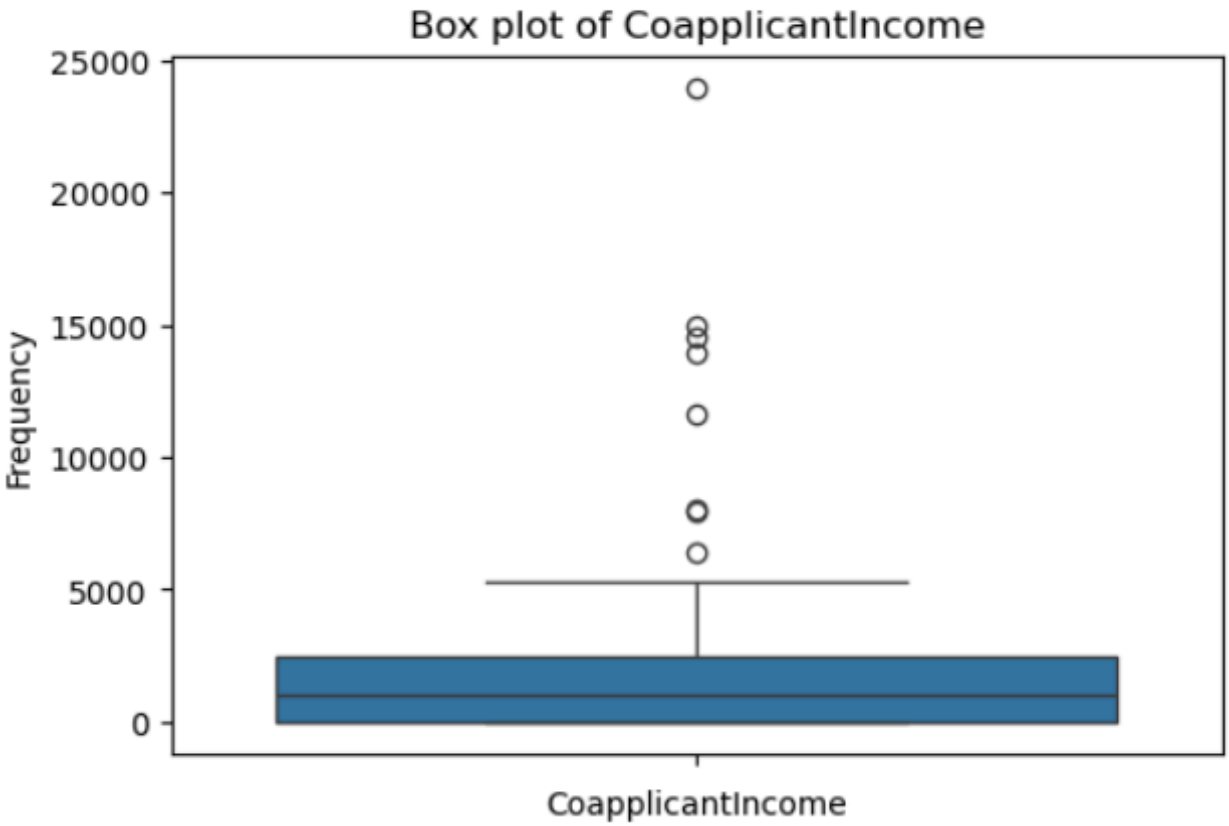

The histogram of Credit_History shows that most applicants have a credit history value of 1, meaning they have a good credit record. Only a small number have a value of 0. This creates a highly right-skewed distribution, with most values concentrated at 1.

# Frequency Distribution of Key Variables



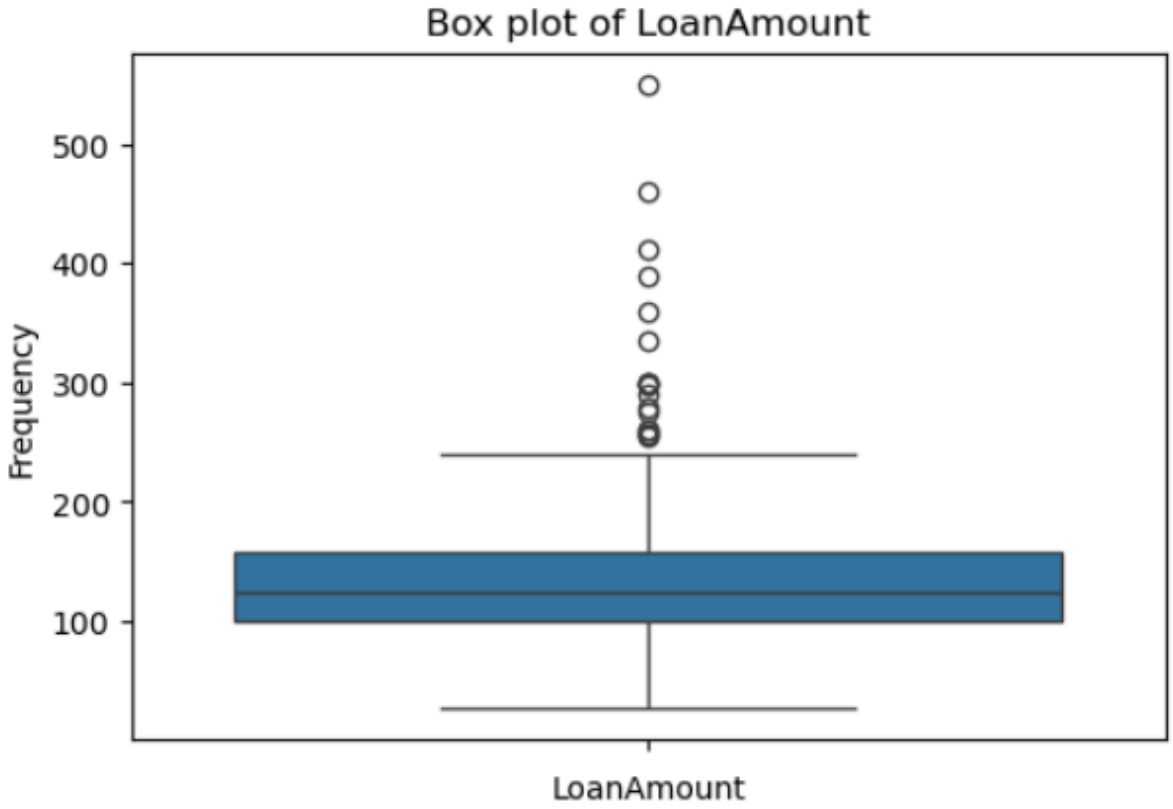

The histogram shows the distribution of Loan_Amount_Term values. Most of the loan terms are grouped around 360 months, meaning most applicants took long-term home loans. Very few applicants have shorter or very long terms, which is why the bars on the left and right are very small.

# Potential Outliers



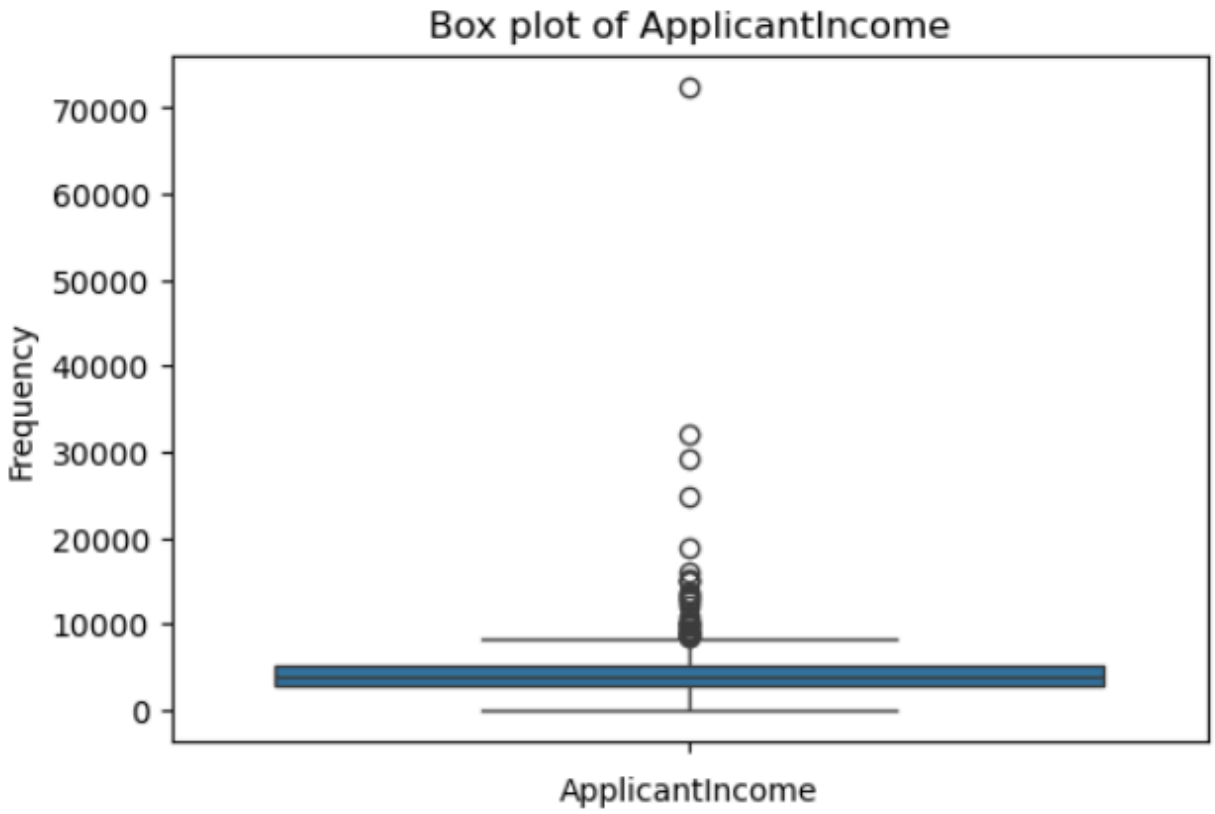

The box plot for Coapplicant Income shows that most coapplicants earn on the lower side, with most values grouped near the bottom range. There are several high-value outliers, including one around ₹25,000, which shows that a few coapplicants have much higher incomes than the rest.

# Potential Outliers



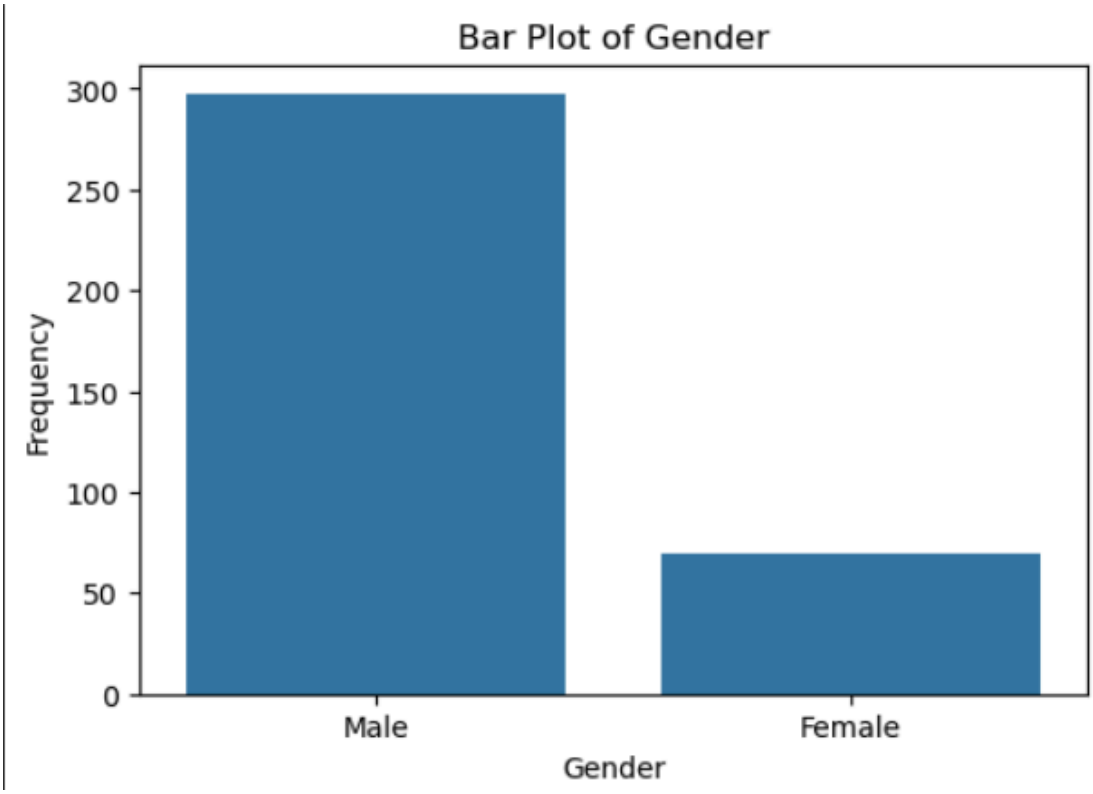

The box plot for LoanAmount shows that most loan amounts fall within a moderate range, with the majority of values between roughly 100 and 150. However, there are several high outliers above 250, including one close to 550, indicating that a few applicants requested much higher loan amounts than the rest.

# Potential Outliers




The box plot for Applicant Income shows that most applicants have incomes in the lower to moderate range, with values closely grouped near the bottom. However, there are many high-income outliers, including one above ₹70,000, indicating that a few applicants earn significantly more than the rest.

# Frequency Distribution (Categorical Variables)

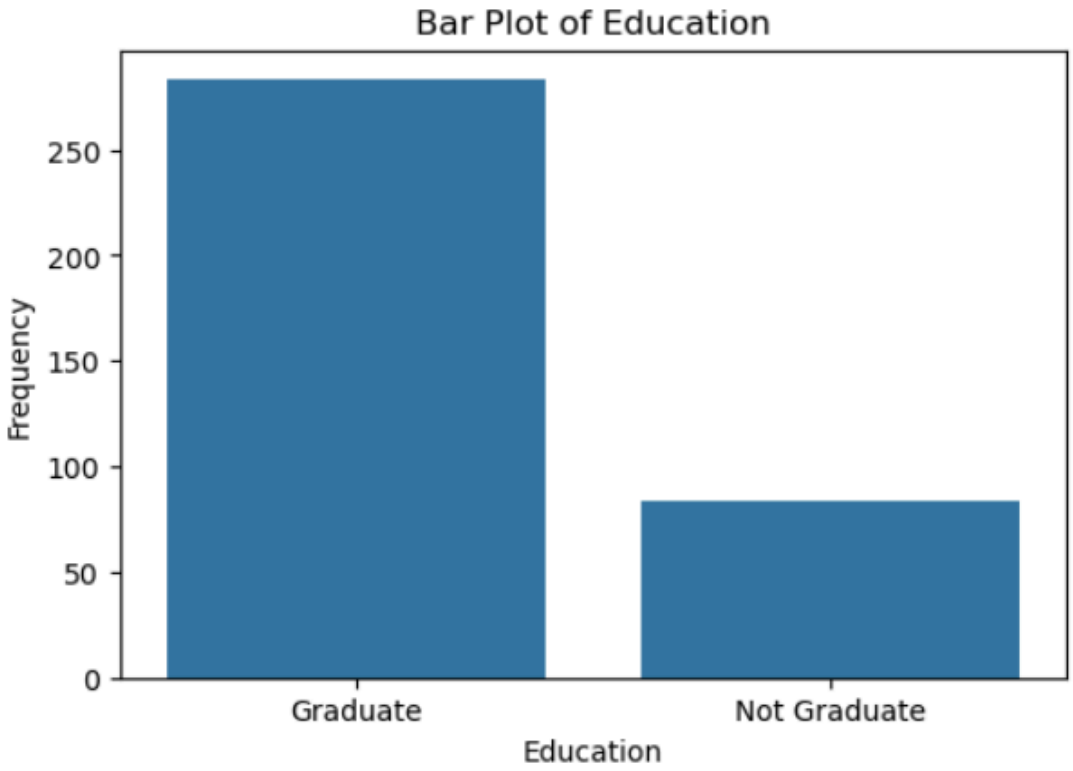



The bar plot shows the distribution of applicants based on gender.

There are many more male applicants compared to female applicants.

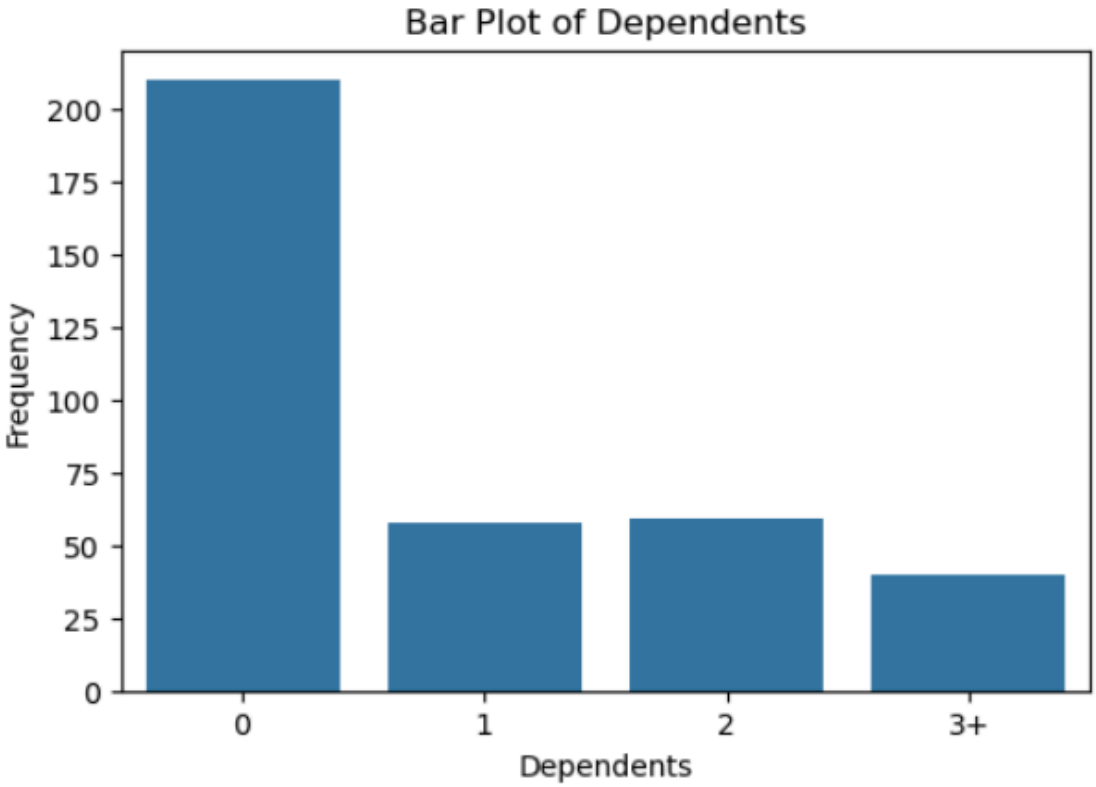
The count of male applicants is significantly higher, while the number of female applicants is much lowe

# Frequency Distribution (Categorical Variables)



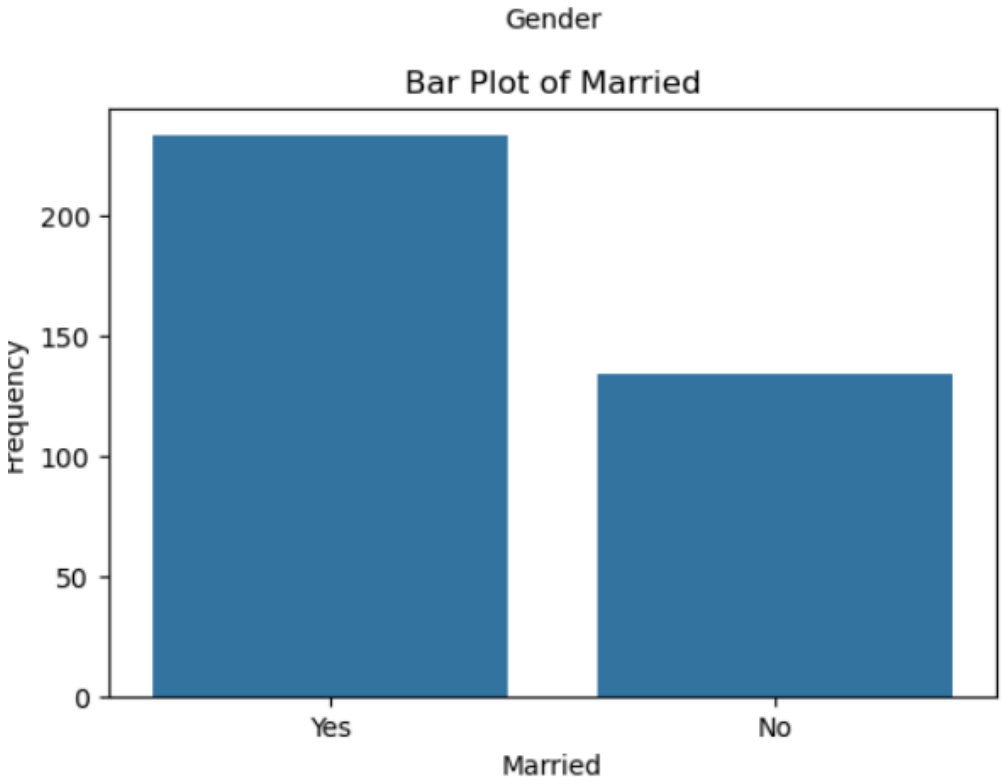

The bar plot shows the distribution of applicants based on their education levels.
Most applicants are Graduates, with their count being significantly higher.
In comparison, the number of Not Graduate applicants is much lower.

# Frequency Distribution (Categorical Variables)



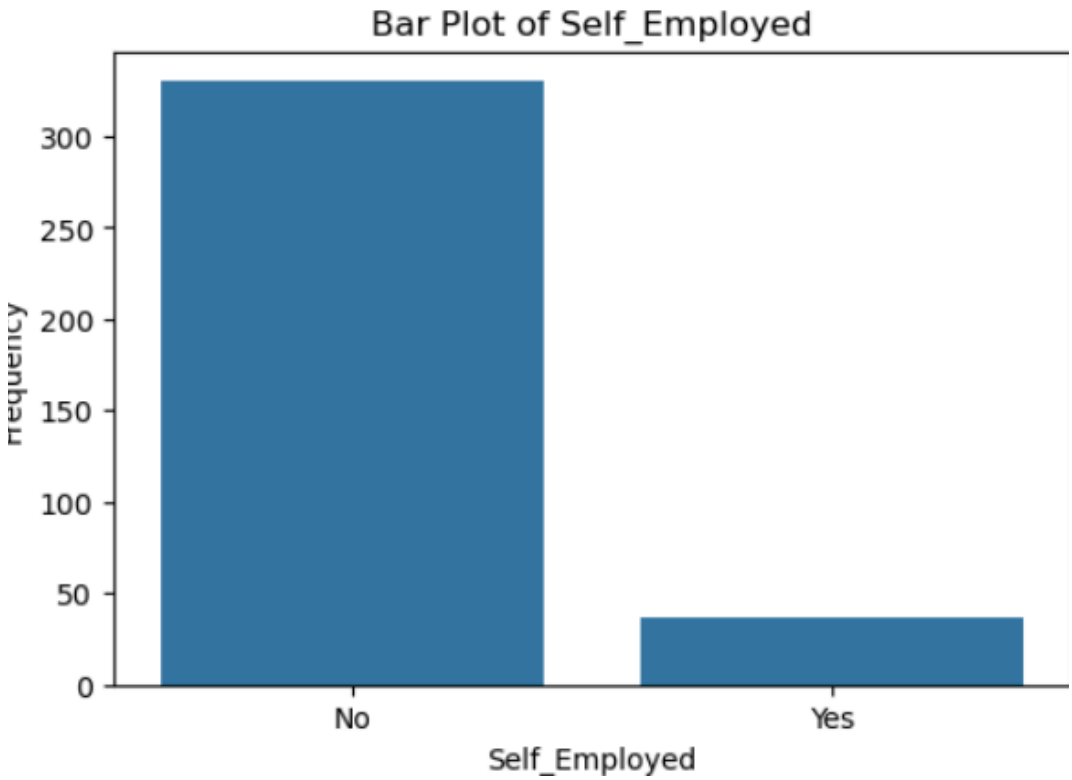

The bar plot shows how many dependents the loan applicants have.
Most applicants have 0 dependents, which is the highest group.
The number of applicants decreases as the number of dependents increases.
Applicants with 1 or 2 dependents are fewer and almost similar in count, while those with 3 or more dependents are the least.

# Frequency Distribution (Categorical Variables)



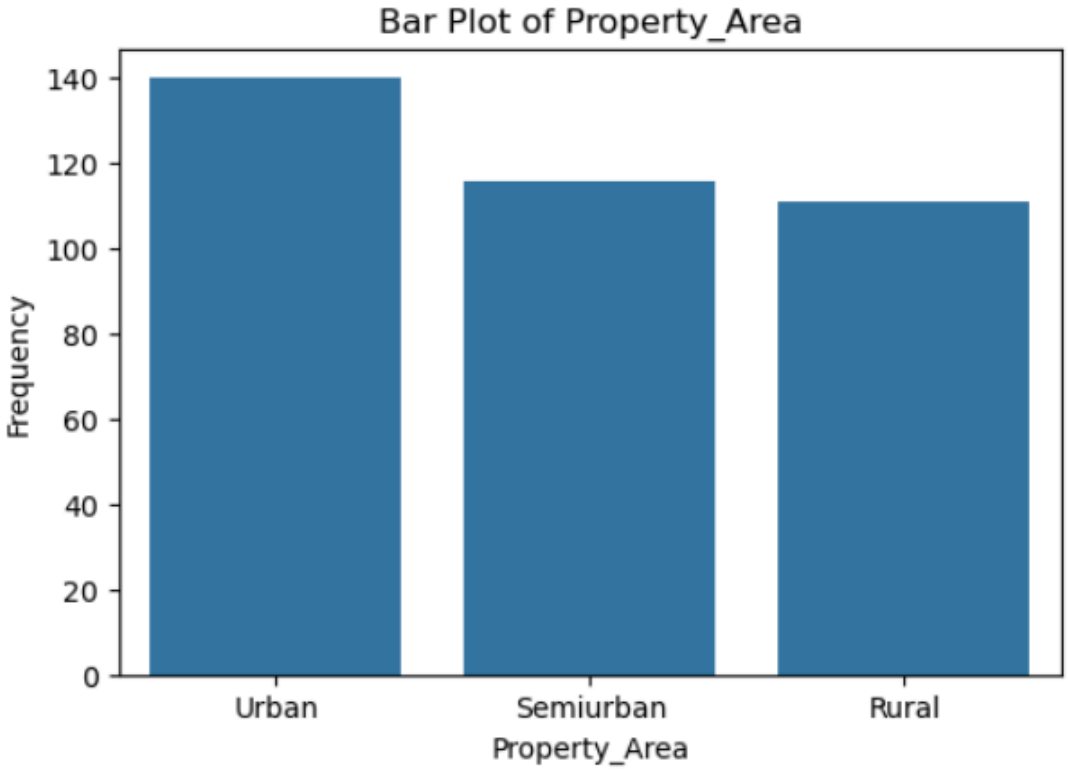

The bar plot shows the distribution of applicants based on their marital status.
Most applicants are married, forming the larger portion of the dataset.
A smaller number of applicants are not married.
Overall, the dataset contains more married applicants compared to unmarried ones.

# Frequency Distribution (Categorical Variables)



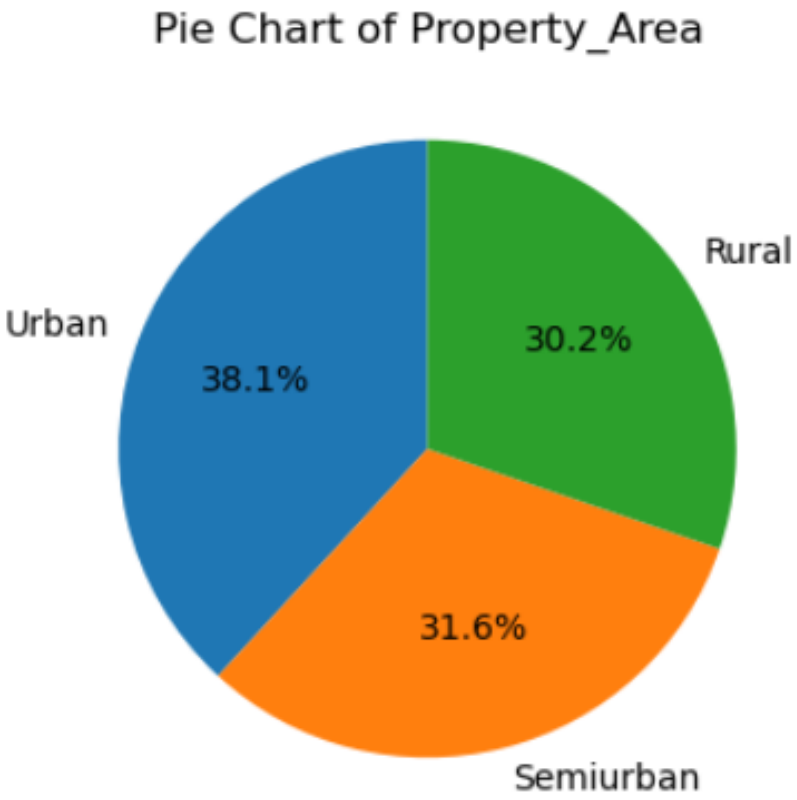

The bar plot shows the distribution of applicants based on whether they are self-employed.
Most applicants are not self-employed, which forms the largest group.
Only a small number of applicants fall under the self-employed category.

# Frequency Distribution (Categorical Variables)



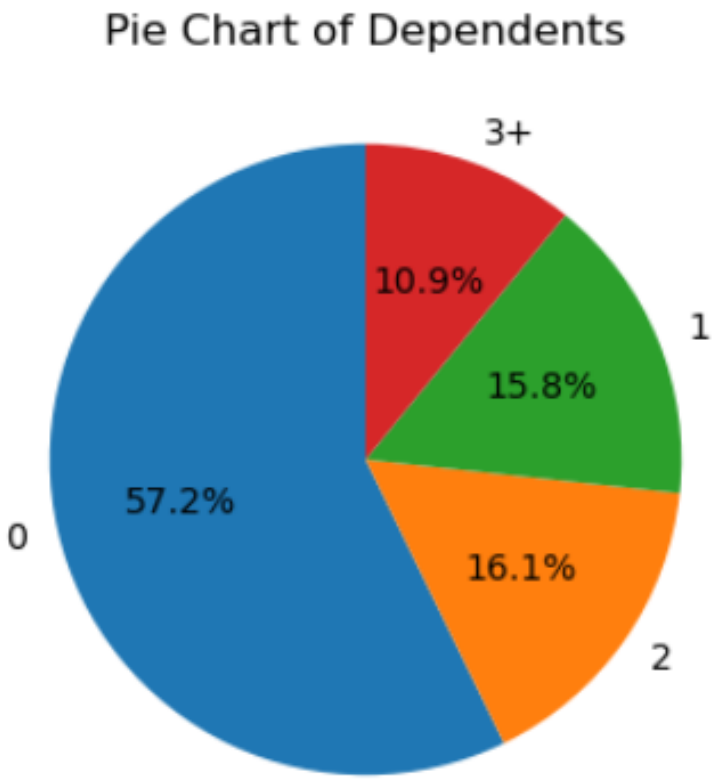

The bar plot shows the distribution of applicants based on the type of property area they live in.
The highest number of applicants are from Urban areas.
This is followed closely by Semiurban applicants.
The Rural category has slightly fewer applicants compared to the other two, but the difference is not very large.

# Composition of Categorical Variables



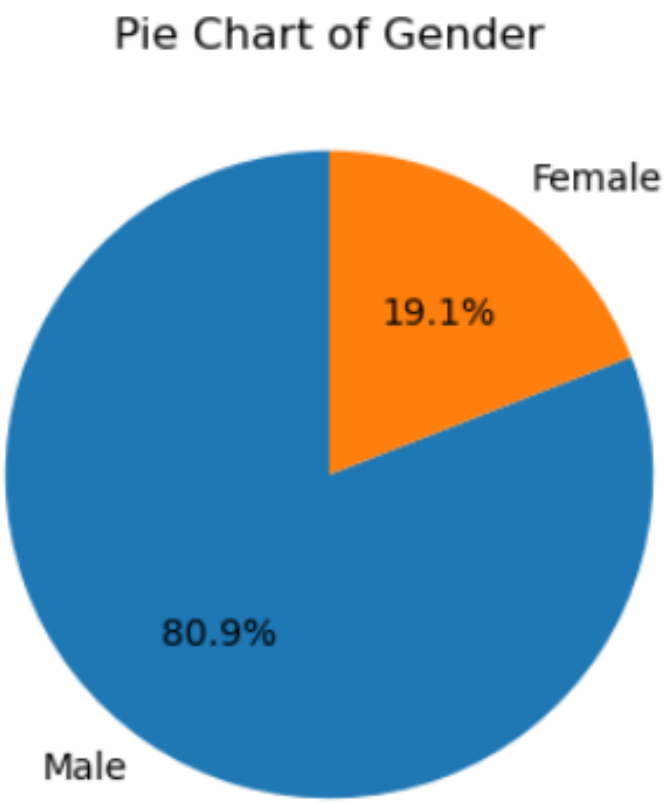

Urban areas have the most loan applicants. Semiurban and Rural areas have slightly fewer applicants, but the difference is small. This means people from all three areas apply for loans in almost equal numbers. Overall, Urban areas still have a little more applicants than the other two.

# Composition of Categorical Variables



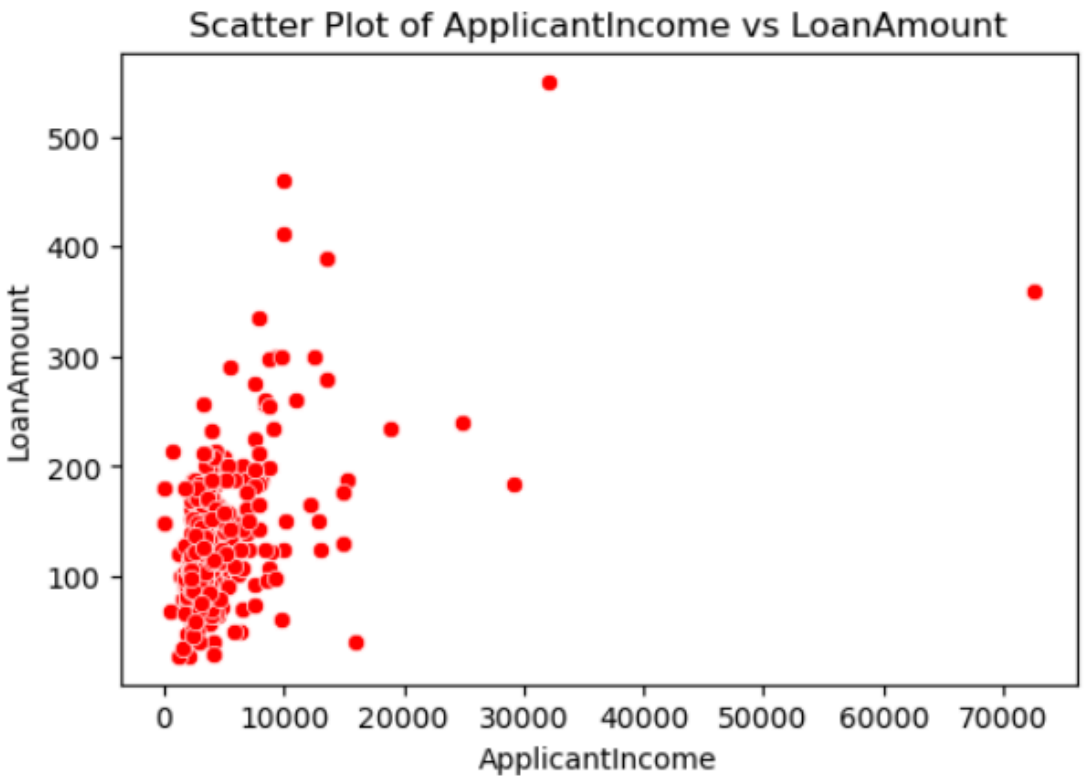

Most applicants have 0 dependents, making up more than half of the total. Applicants with 1 or 2 dependents form smaller but similar groups. The smallest group is applicants with 3 or more dependents. Overall, loan applications decrease as the number of dependents increases.

# Composition of Categorical Variables



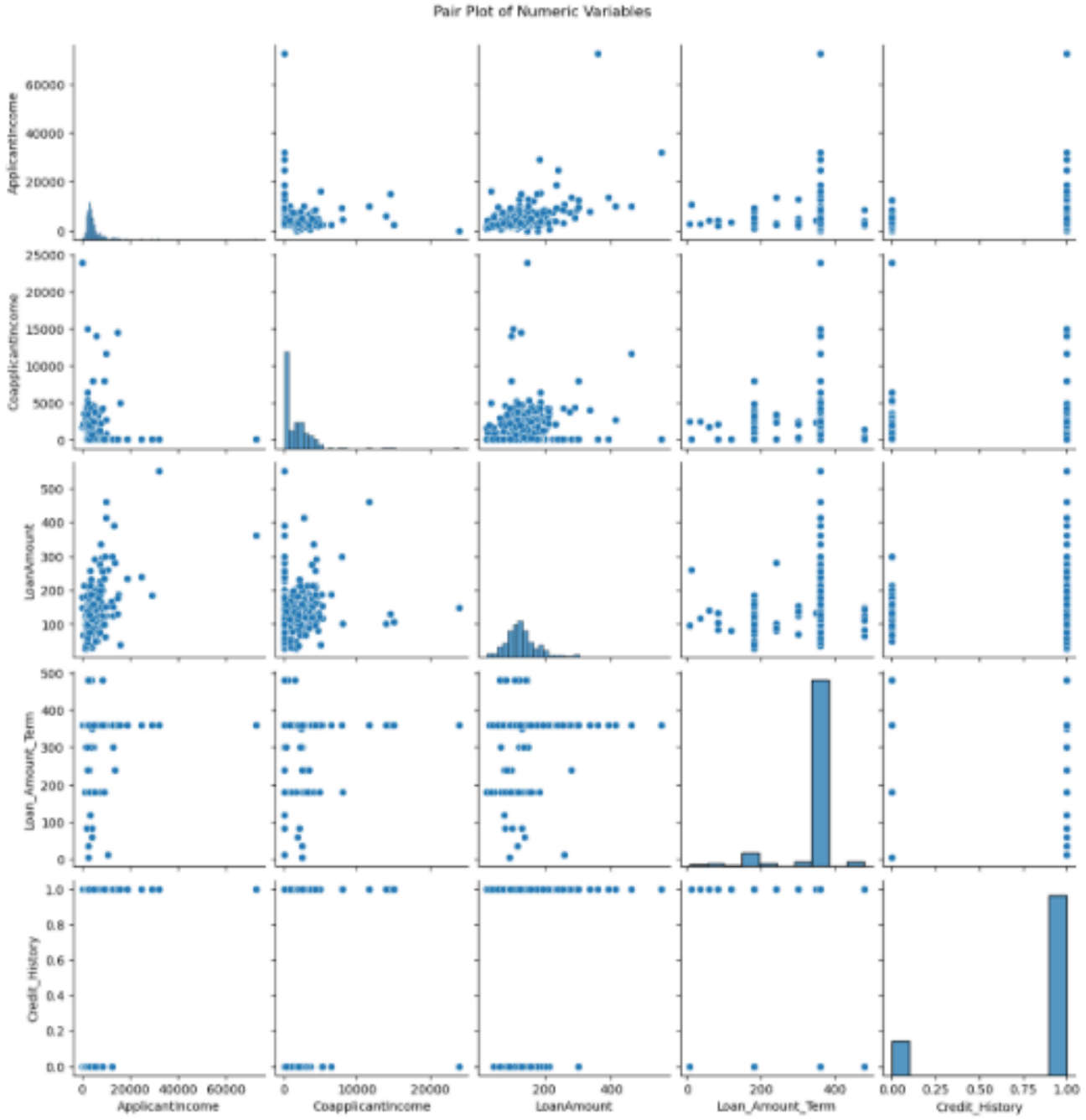

Most loan applicants are male, making up the largest share at 80.9%. Only 19.1% of the applicants are female. This shows a big difference between male and female participation. Overall, the dataset is mainly male-dominated, with far fewer female applicants applying for loans.

# Relationship Between Numerical Variables




The scatter plot shows the relationship between applicant income and loan amount.
Most points are clustered at lower income levels with moderate loan amounts.
As income increases, loan amounts also tend to increase, but the pattern is not very strong.
A few high-income applicants took much larger loans, showing some outliers.

# Relationship Between Numerical Variables



# Relationship Between Numerical Variables

**Pair Plot Summary of Numeric Variables**

The pair plot shows how different numeric variables—ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, and Credit_History—interact with each other.

Each subplot shows a scatter plot between two variables, while the diagonal plots display the individual distribution of each variable.

**Observation:**

A slight positive pattern is visible between ApplicantIncome and LoanAmount, meaning applicants with higher incomes often request higher loan amounts.

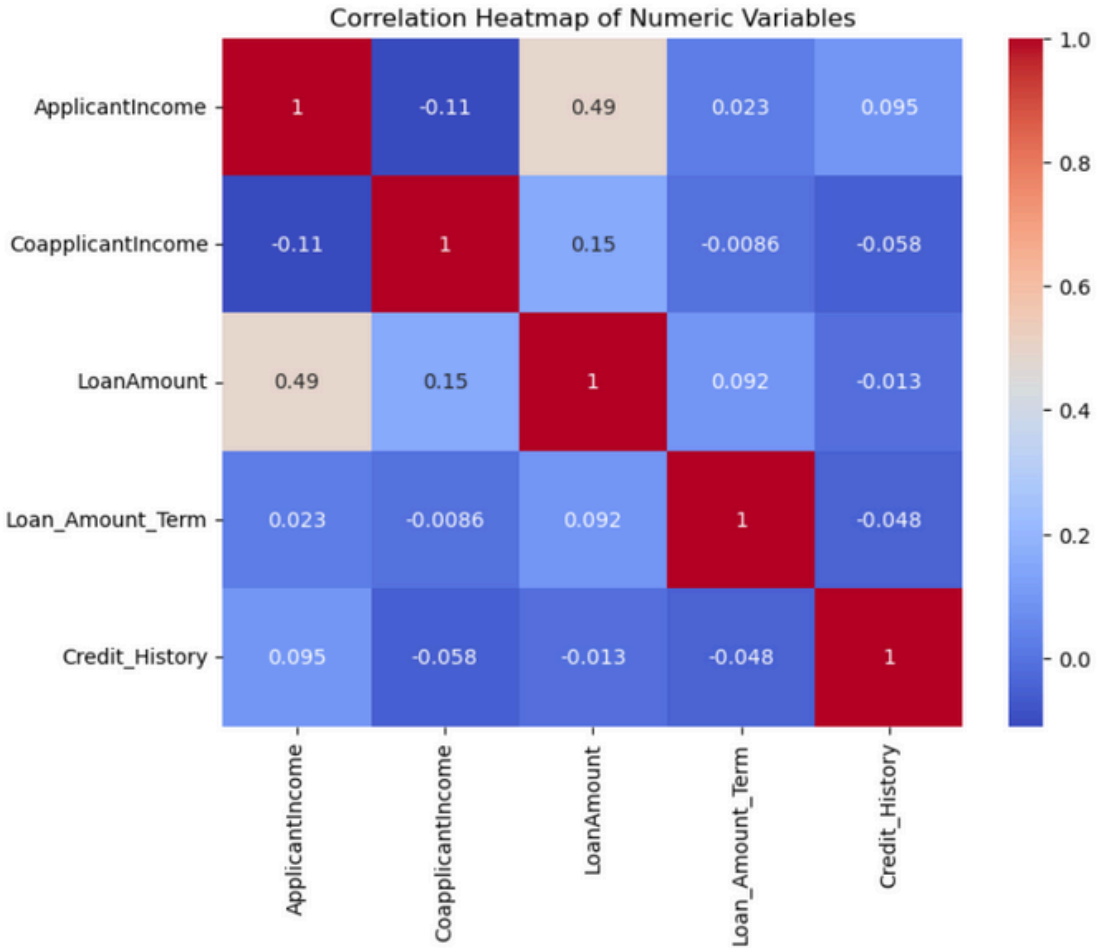
Other pairs, like CoapplicantIncome and LoanAmount, show weaker or no clear correlation.

The plot also helps identify outliers, such as very high incomes or unusually large loan amounts, and shows how the data is spread across variables.

**Insight:**

The pair plot gives a clear overview of the relationships between the numerical features. It helps identify which variables may influence loan amounts more strongly and provides a helpful visual understanding of data patterns and variability.

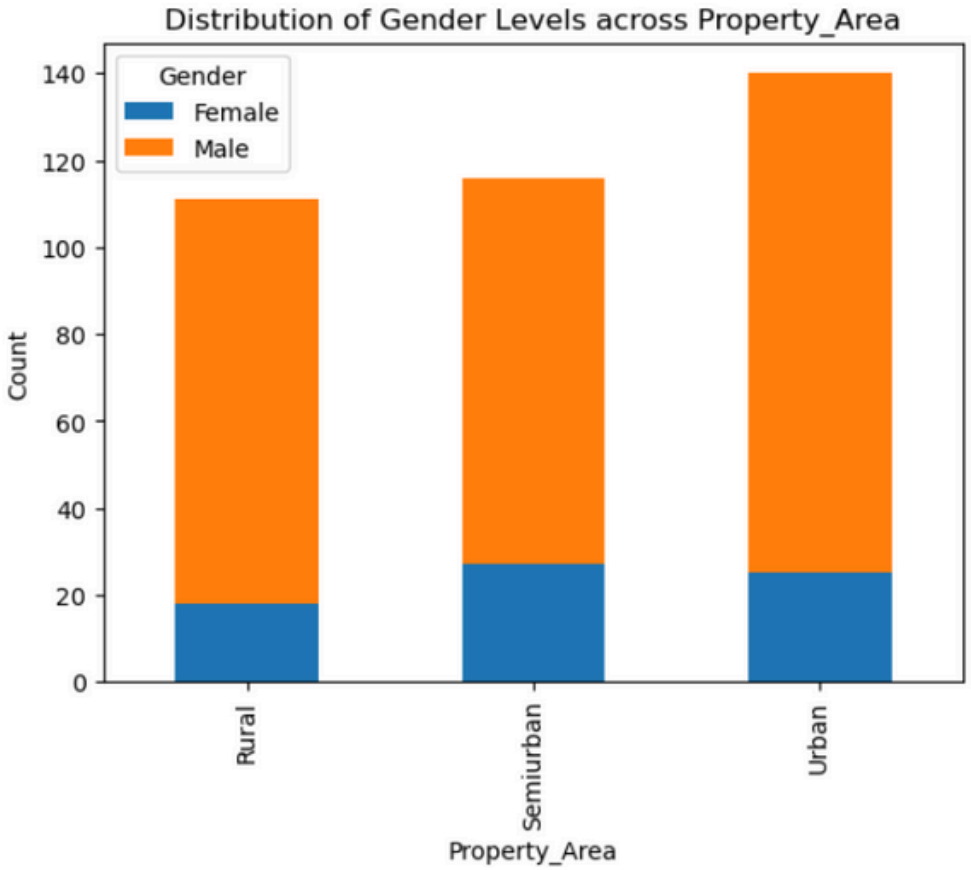
# Relationship Between Numerical Variables




The heatmap shows the correlation values between numeric variables such as ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, and Credit_History.

Positive values indicate that two variables increase together, while negative values show an opposite relationship. Values close to zero suggest little or no correlation.

# Relationship Between Numerical Variables




Across all property areas—Rural, Semiurban, and Urban—male applicants are much higher than female applicants.

Urban areas have the highest number of male applicants, while Rural areas have the lowest.

Female counts are lower in every area but remain fairly similar across regions.

# Insights and Findings

The dataset shows that most loan applicants are male (80%) and married (63%), indicating that family responsibilities and joint financial planning commonly drive loan applications.
A large proportion of applicants are graduates (77%) and not self-employed (89%), suggesting that individuals with stable, salaried jobs are more likely to apply for home loans.
The majority of applications come from Urban and Semi-Urban areas, highlighting greater loan demand in more developed and populated regions.
There is a positive correlation between ApplicantIncome and LoanAmount, meaning applicants with higher income generally qualify for larger loan amounts.
In contrast, CoapplicantIncome shows only a weak correlation, implying that secondary income plays a smaller role in determining loan size.
A few outliers in both ApplicantIncome and LoanAmount indicate applicants with unusually high earnings or significantly large loan requests.
Overall, the bar charts, pie charts, and distribution plots show that socio-economic and demographic factors—including gender, education, marital status, employment, and property area—have a significant influence on loan patterns and applicant behavior.

# Conclusion

This Loan Approval Data Analysis project helped uncover how different factors affect loan applications. After cleaning, exploring, and visualizing the data, it became clear that applicant income, education level, marital status, and property area strongly influence loan eligibility. Most loan seekers are graduates, salaried employees, and married individuals living in urban or semi-urban areas. The study also showed that higher applicant income results in higher loan amounts, while coapplicant income plays a smaller role.

Overall, the project highlights the value of data analysis in understanding financial patterns and supports banks in making smarter, data-driven loan decisions.

# THANK YOU