

Exploratory Data Analysis

LOAN APPROVAL



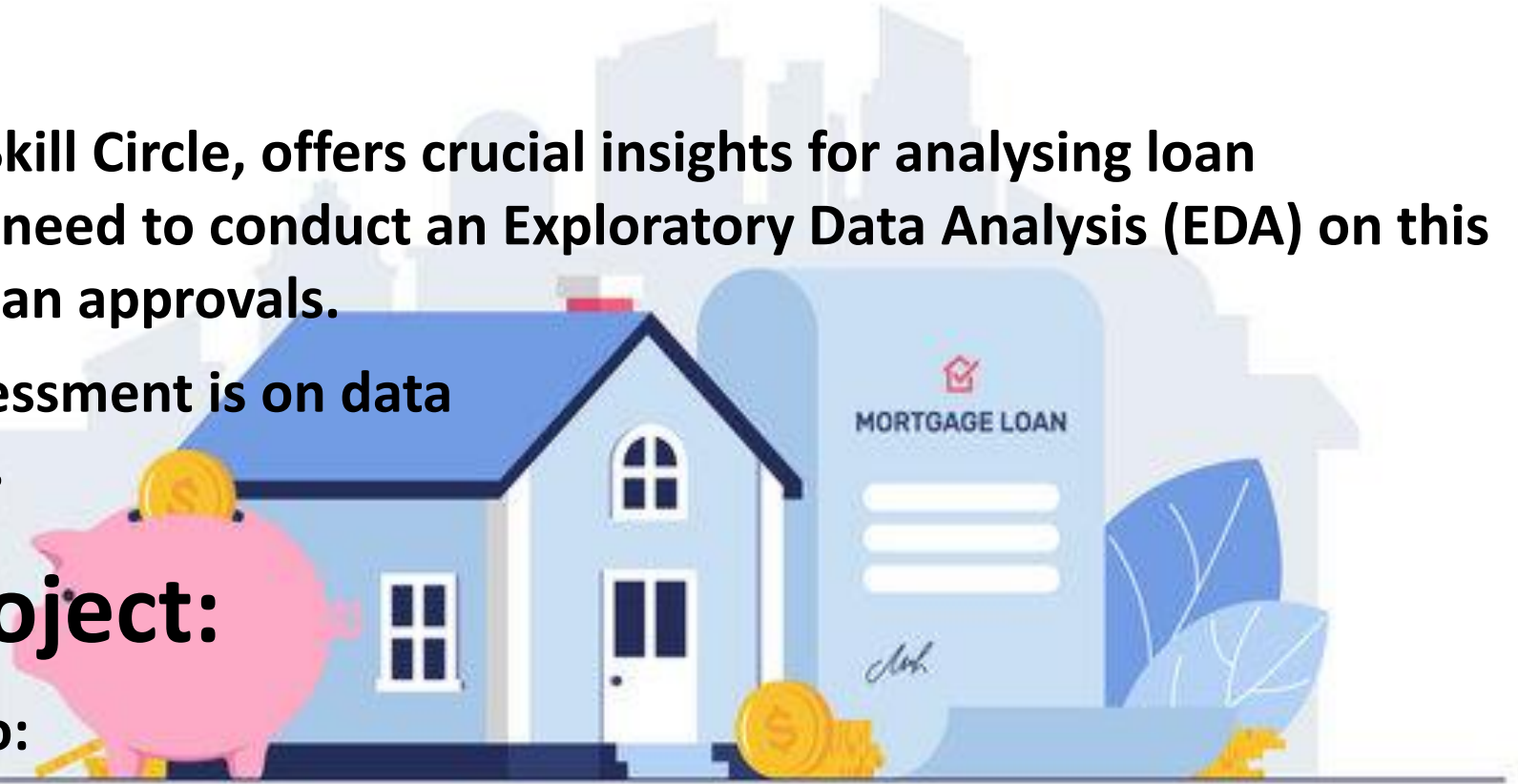
INTRODUCTION

- The dataset, obtained from Skill Circle, offers crucial insights for analysing loan applications. In this task, we need to conduct an Exploratory Data Analysis (EDA) on this dataset, focusing on home loan approvals.
- The primary focus of this assessment is on data exploration and visualization.

Objectives of the project:

The goals of this assessment is to:

- i. Gain familiarity with the dataset.
- ii. Identify patterns, trends, and potential insights.
- iii. Perform data exploration and visualization.
- iv. Generate meaningful visualizations to communicate your findings.
- v. This project aims to predict loan approval based on the given dataset.
- vi. It involves data cleaning, data analysis, pre-processing to gain beneficial derivatives.



Description of Dataset:

- I have conducted my work using Google Colab Notebook.
- The dataset has been imported from Files.
- As we begin our Exploratory Data Analysis (EDA), I've named the dataset 'loan'.
- The dataset comprises of 367 rows and 12 columns.
- For data cleaning, I have utilized libraries like Numpy , Pandas , Matplotlib , Plotly and Seaborn .
- Any duplicate entries that were found have also been removed.

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

```
loan = pd.read_csv('/content/loan_sanction_test.csv')
```

```
[ ] loan.drop_duplicates()
```

```
[ ] loan.shape
```

```
[ ] (367, 12)
```

Description of Dataset:

- The dataset contains information collected from multiple borrowers and, at first glance, seems to pertain to home loans. Additionally, there are some outliers and missing values evident in the data.

Key Features include:


- ❖ Loan ID : A unique identifier assigned to each loan application Gender for tracking and reference purposes.
- ❖ Gender : The gender of the applicant (e.g., Male, Female).
- ❖ Married : Marital status of the applicant (e.g., Yes, No).
- ❖ Dependents Education : The number of dependents or family members who rely on the applicant for financial support.
- ❖ Education : The educational qualification of the applicant (e.g., Graduate, Not Graduate).
- ❖ Self Employed : Indicates whether the applicant is self or not (e.g., Yes, No).
- ❖ Applicant Income : The monthly income of the applicant.

```
[ ] loan.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
```


Description of Dataset:

- **Co-applicant Income** : The monthly income of the co-applicant.
- **Loan Amount** : The total amount of the loan requested by the applicant.
- **Loan Amount Term** : The tenure or duration of the loan in months.
- **Credit History** : A numerical indicator of the applicant's credit history, reflecting their ability to repay the loan. (e.g., 1, 0).
- **Property Area**: The geographical area or type of area where the property is located (e.g., Urban, Semi-urban, Rural).



```
loan.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

Data Cleaning & Pre-Processing:

Our dataset has a total of **84 null values**. Of these, **34 are found in categorical features**, while **50 are in numerical features**.

- ❑ For the 'Gender' attribute, which has 11 null values, filling in the missing entries with the mode value—'male'—will help ensure data completeness.

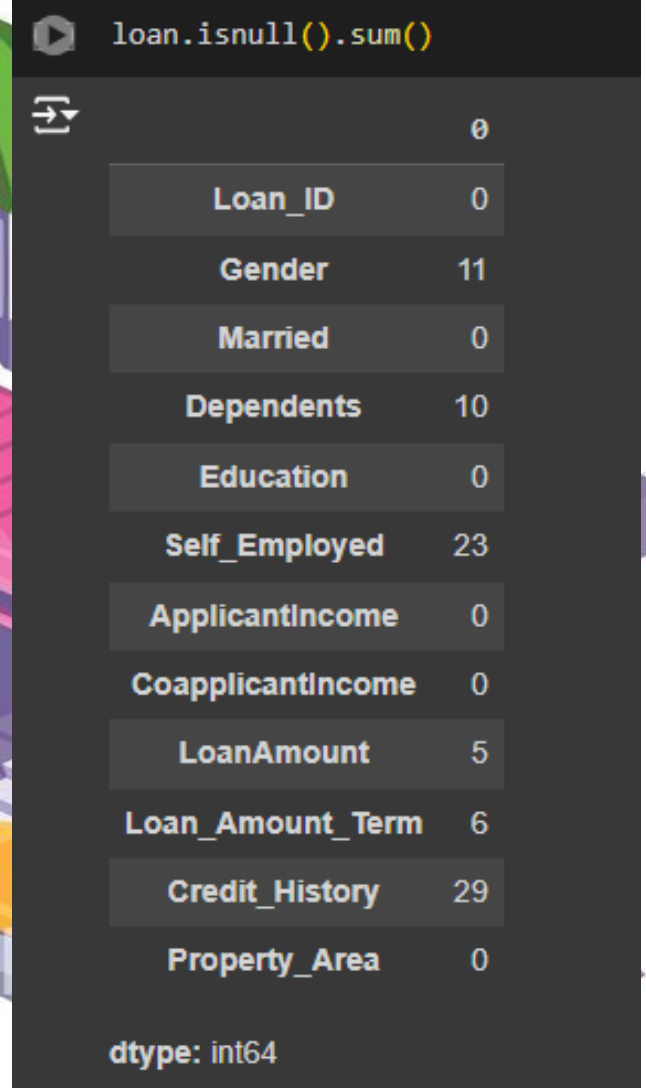
```
loan['Gender'].fillna(loan['Gender'].mode()[0],inplace=True)
```

- ❑ For the 'Dependents' attribute, which has 10 null values, filling in the missing entries with the mode value—'0'—will help ensure data completeness.

```
loan['Dependents'].fillna(loan['Dependents'].mode()[0],inplace=True)
```

- ❑ For the 'Self_Employed' attribute, which has 23 null values, filling in the missing entries with the mode value—'No'—will help ensure data completeness.

```
loan['Self_Employed'].fillna(loan['Self_Employed'].mode()[0],inplace=True)
```



```
loan.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

Data Cleaning & Pre-Processing:

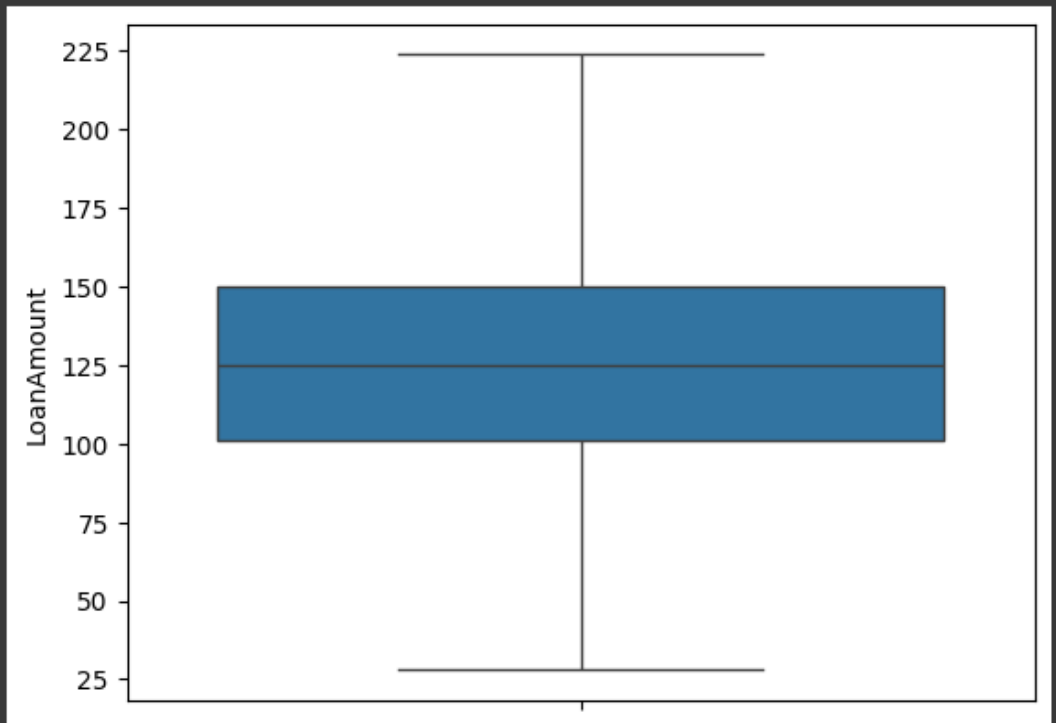
For the 'Loan Amount' feature, which has both outliers and null values, the process involved first filling the 5 null values with the median. Subsequently, the outliers were managed using the Interquartile Range (IQR) method.

```
[ ] median_loan_amount = loan['LoanAmount'].median()
median_loan_amount
```

125.0

```
[ ] loan['LoanAmount'].fillna(loan['LoanAmount'].median(),inplace=True)
```

```
[ ] sns.boxplot(loan['LoanAmount'])
plt.show()
```



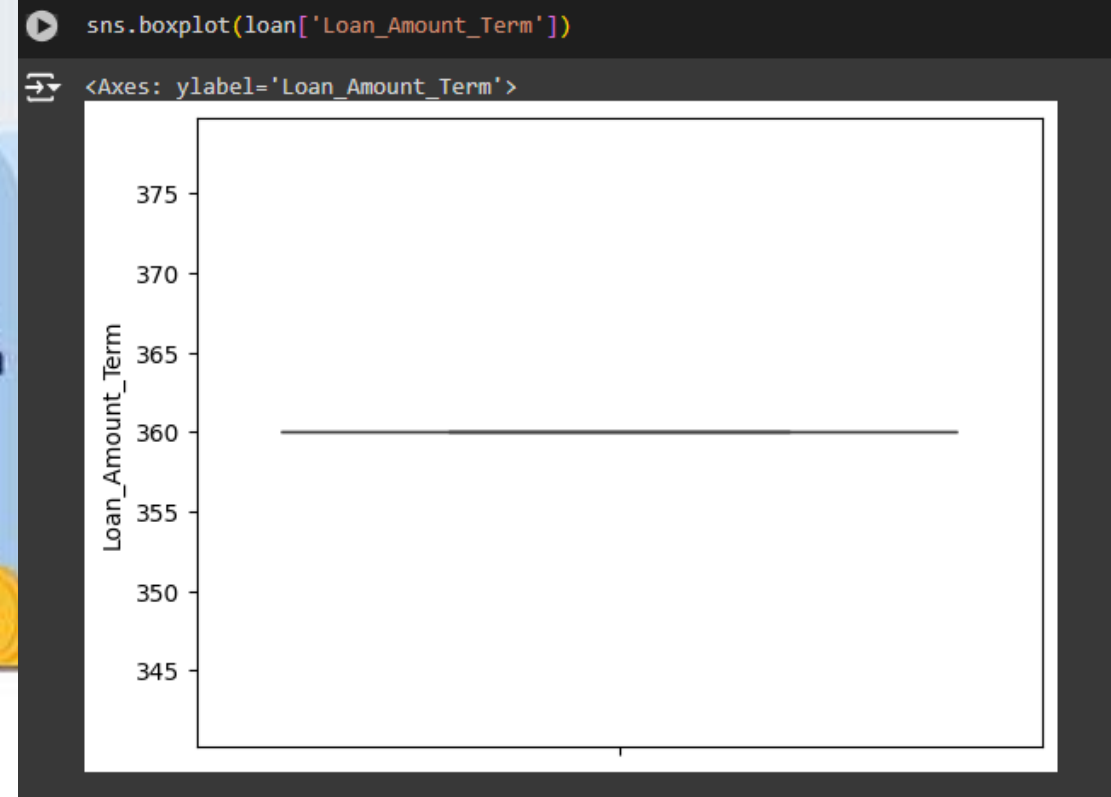
Data Cleaning & Pre-Processing:

For the 'Loan Amount Term' feature, which has both outliers and null values, the process involved first filling the 6 null values with the median. Subsequently, the outliers were managed using the Interquartile Range (IQR) method.

```
[ ] median = loan['Loan_Amount_Term'].median()  
      median
```

```
⇒ 360.0
```

```
[ ] loan['Loan_Amount_Term'] = loan['Loan_Amount_Term'].fillna(median).astype(float)
```



Data Cleaning & Pre-Processing:

For the 'Credit History' feature, which has both outliers and null values, the process involved first filling the 29 null values with the median. Subsequently, the outliers were managed using the Interquartile Range (IQR) method.

```
[16] credit_median = loan['Credit_History'].median()  
      credit_median
```

```
1.0
```

```
loan['Credit_History'] = loan['Credit_History'].fillna(loan['Credit_History'].median())
```

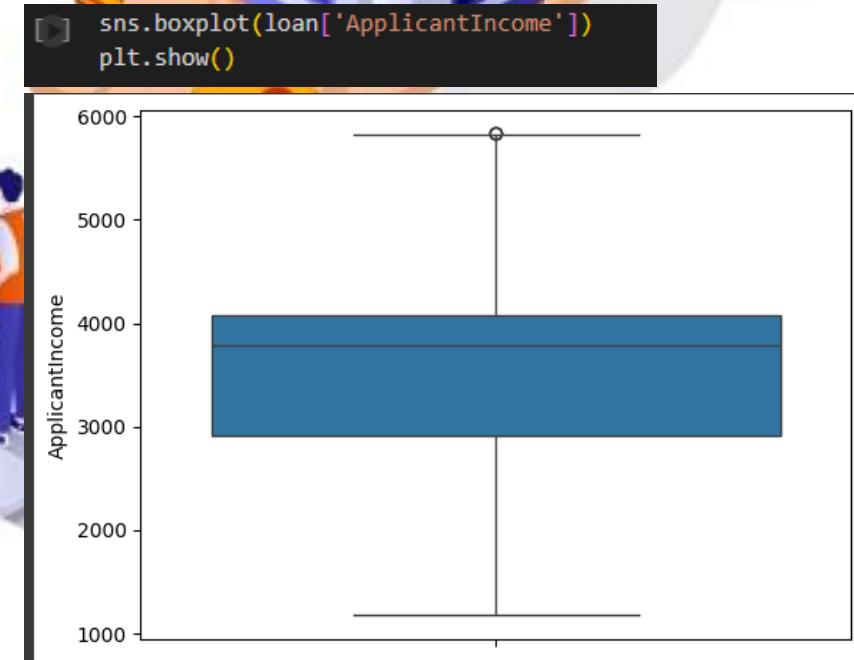
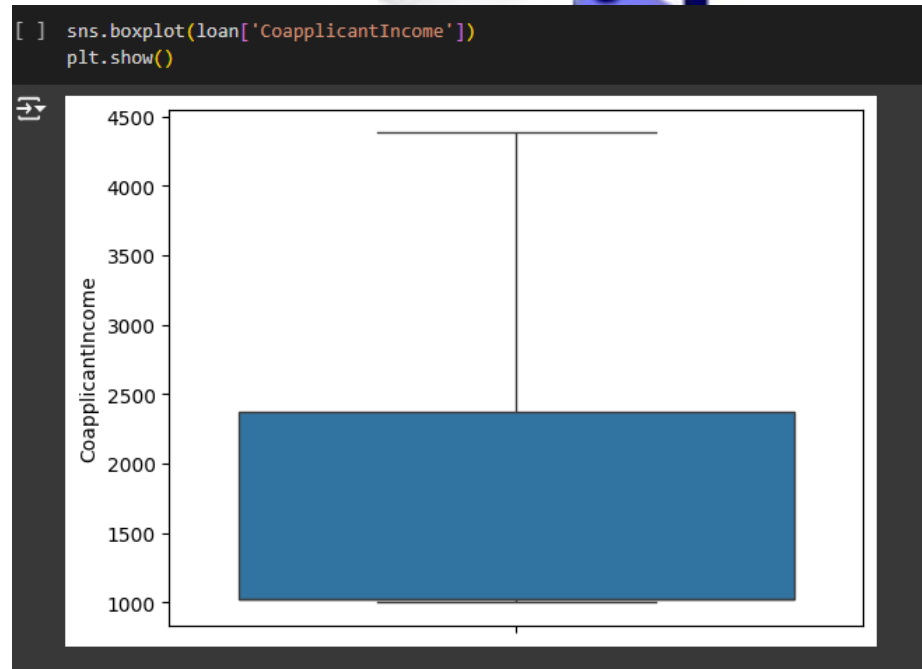
```
sns.boxplot(loan['Credit_History'])
```

```
<Axes: ylabel='Credit_History'>
```

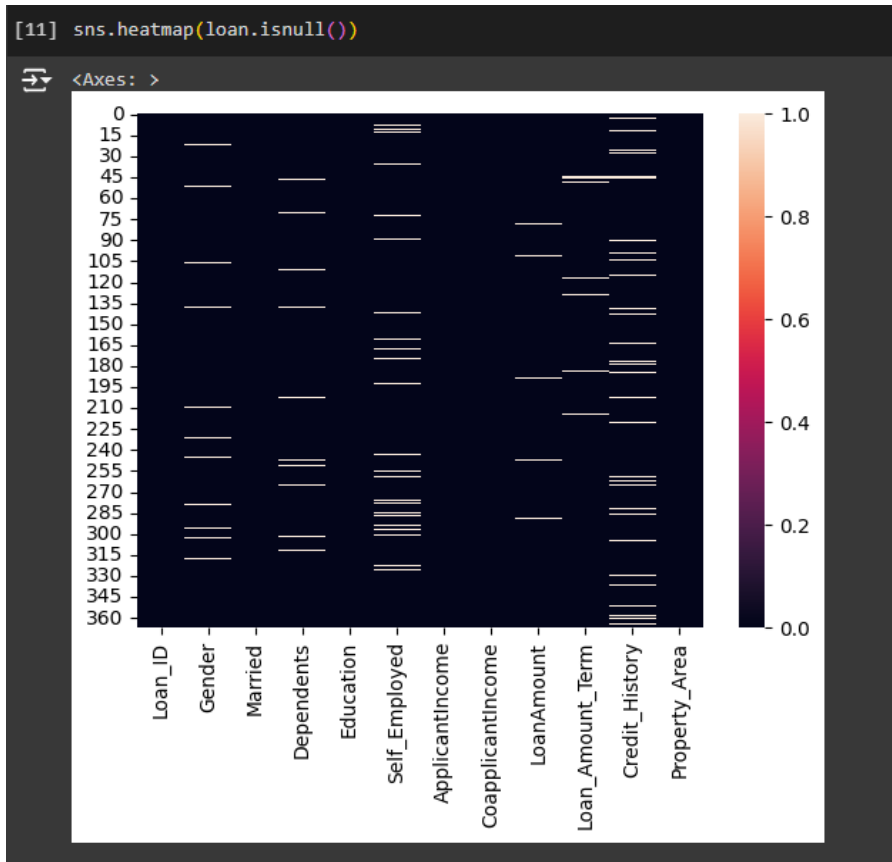


Data Cleaning & Pre-Processing:

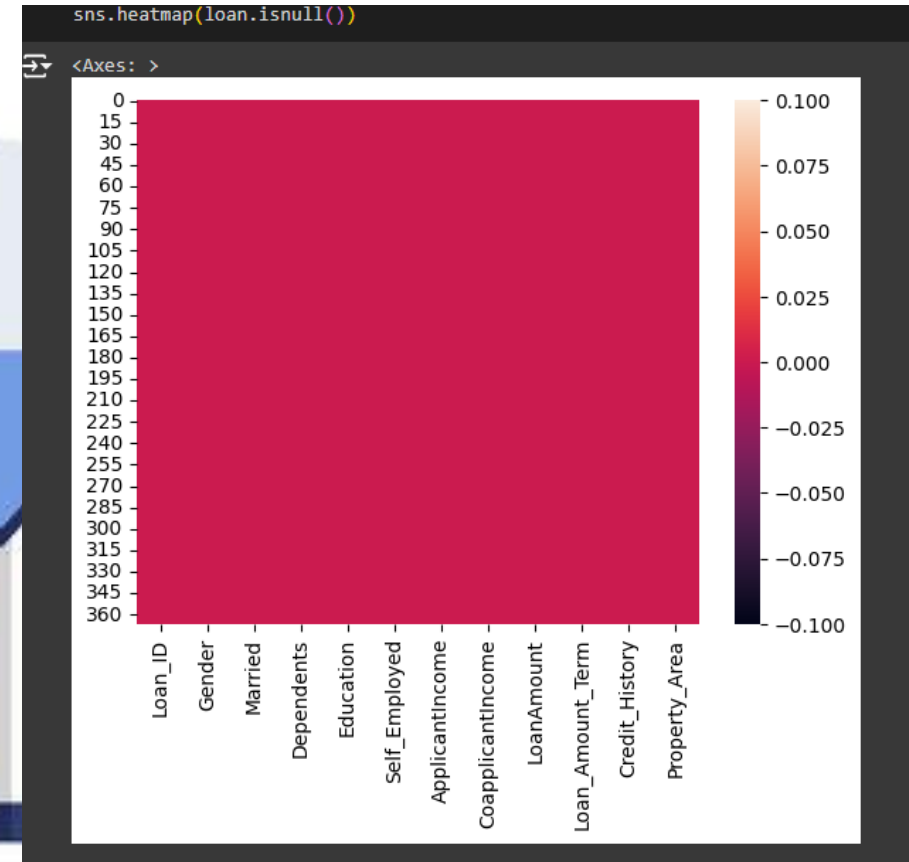
- Key consideration: Although we have addressed all the null values, we still need to examine two numerical features—'applicant income' and 'co-applicant income'—for outliers to ensure data balance and improve the quality of insights.
- For the 'Applicant Income' feature, outliers were addressed using the IQR method. Additionally, some applicants had an income of 0, which is unrealistic for a home loan scenario, so these values were also replaced with the median.
- For the 'Co-Applicant Income' feature, outliers were addressed using the IQR method. Additionally, some applicants had an income of 0, which is unrealistic for a home loan scenario, so these values were also replaced with the median.



BEFORE CLEANING



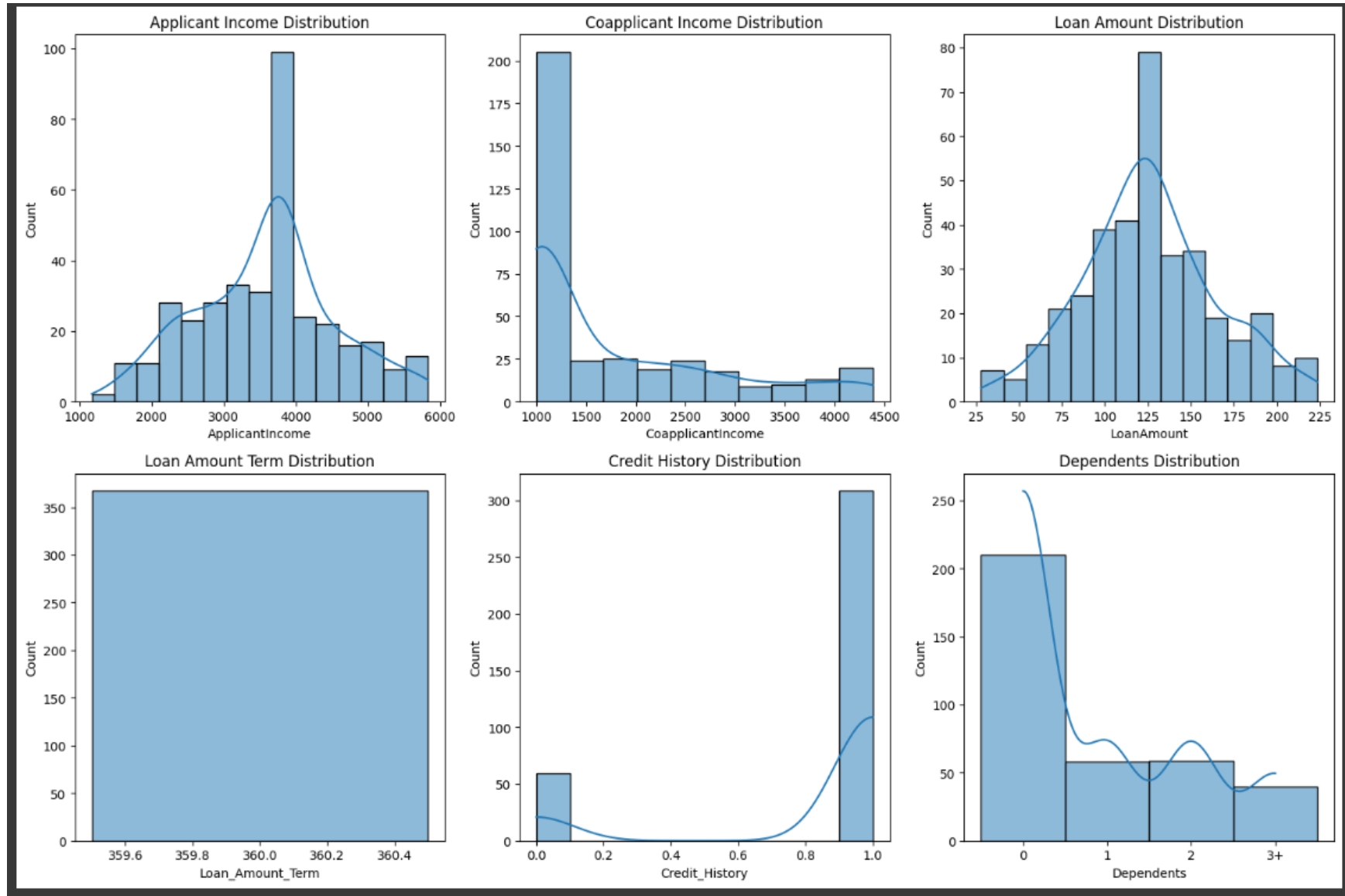
AFTER CLEANING



- **SUMMARY** : Handling null values and outliers requires a systematic approach that suits the specific characteristics and attributes of the data. By implementing the described strategies, we can effectively manage and fill in missing values, ensuring the dataset's completeness, integrity, and reliability for subsequent analysis and insights.
- With these null, missing, and invalid values properly addressed, we are now prepared to proceed with analyzing the dataset.

Data Visualization and Insights

Histograms: Plot the frequency distribution of key Numeric variables.

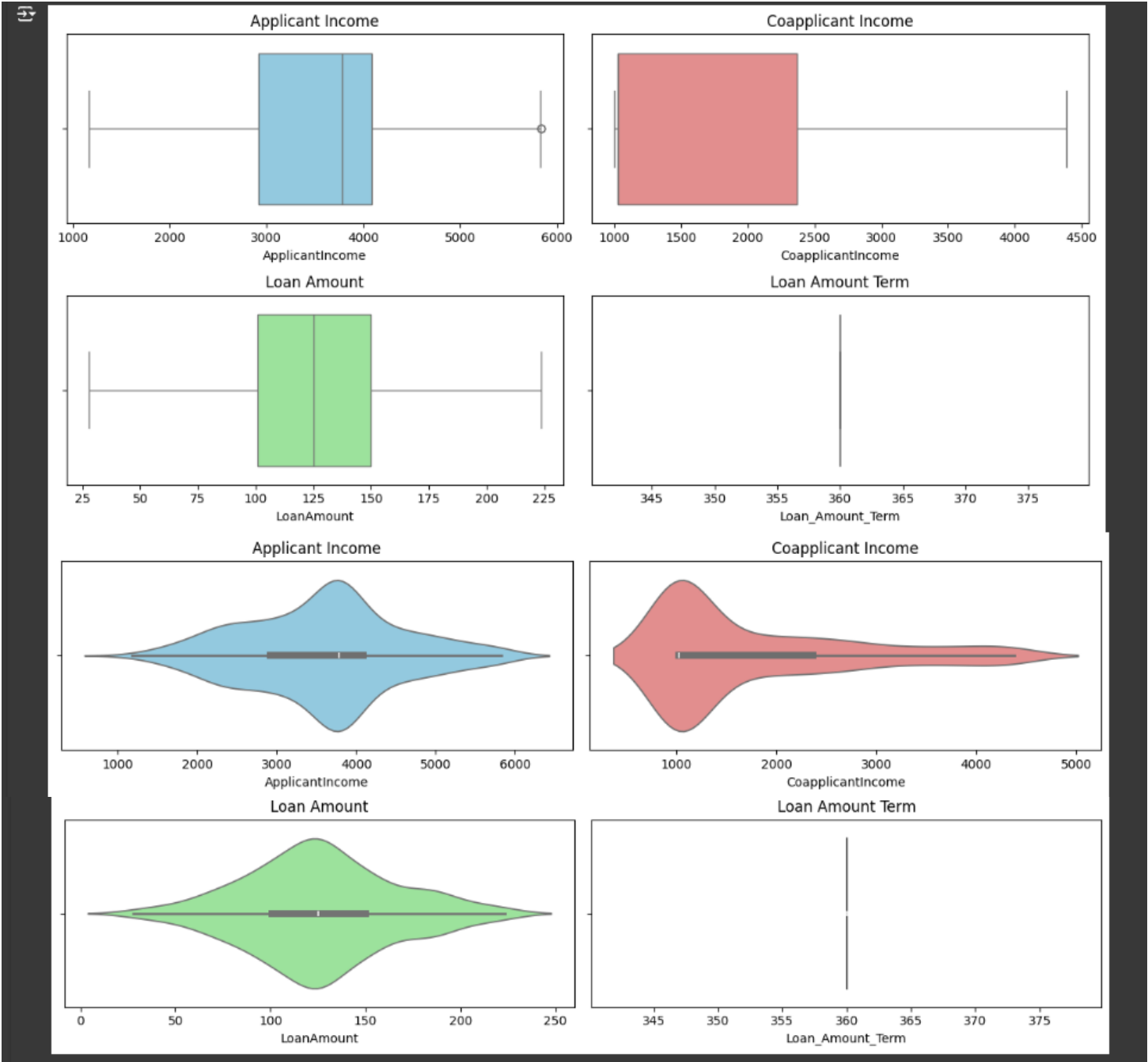


Key insights

- **Applicant Income Distribution:** The distribution is right-skewed, indicating a higher concentration of applicants with lower incomes . There might be a few high-income outliers influencing the distribution.
- **Co-applicant Income Distribution:** A significant portion of co-applicants have zero income (likely indicating single applicants).The distribution is also right-skewed, similar to applicant income.
- **Loan Amount Distribution:** The distribution appears to be relatively normal, with a peak around the average loan amount. This suggests that most loans fall within a typical range.
- **Loan Amount Term Distribution:** The majority of loans have a term of around 360 months (30 years). There are smaller peaks for other common loan terms (e.g., 180 months, 480 months).
- **Credit History Distribution:** Most applicants have a credit history (value of 1), which is a positive sign for loan approval. A smaller portion of applicants have no credit history (value of 0).
- **Dependents Distribution:** The majority of applicants have zero or one dependent. There's a decreasing trend in the number of applicants with more dependents.

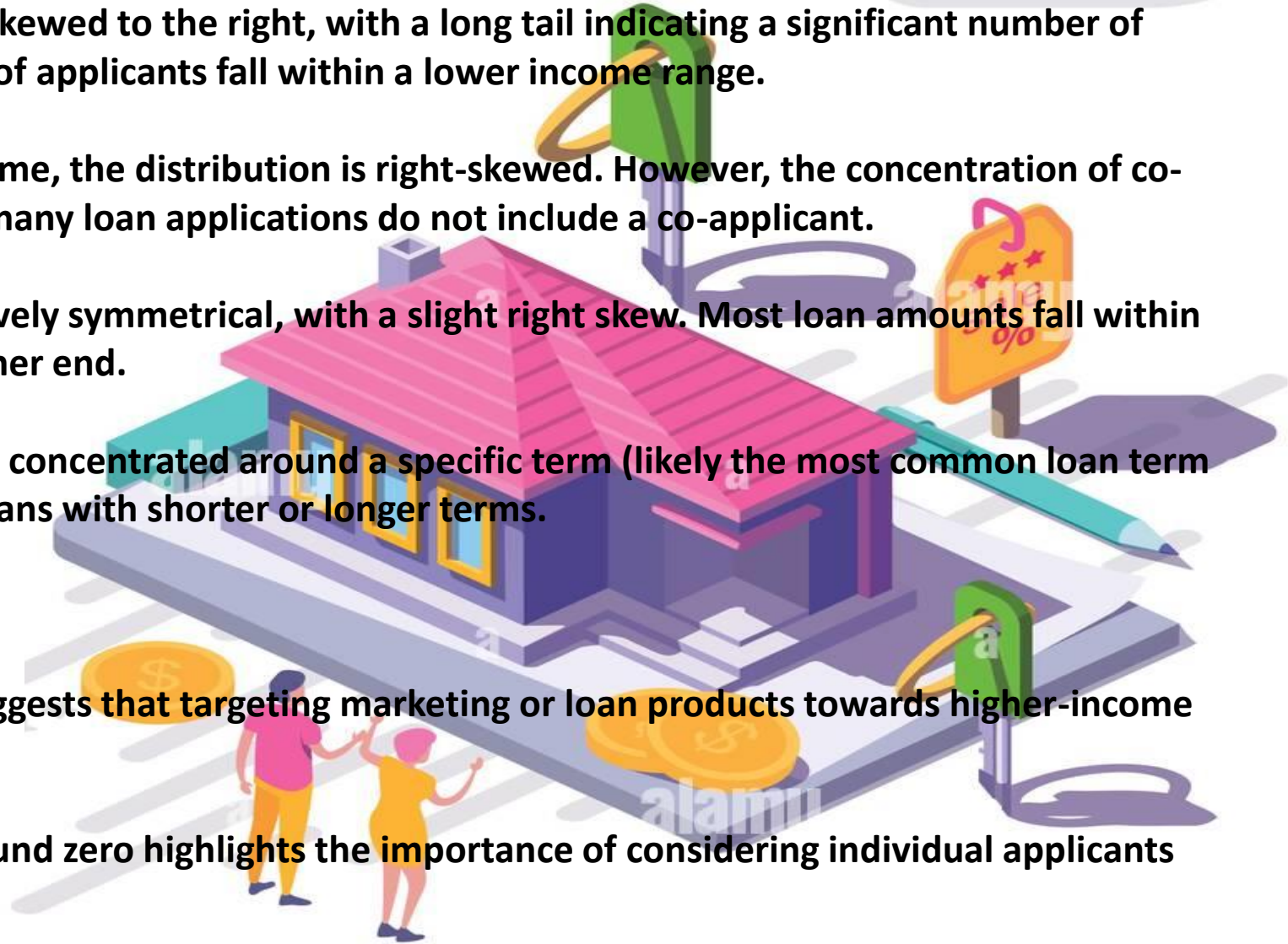


Identify potential outliers and visualize the spread of data.

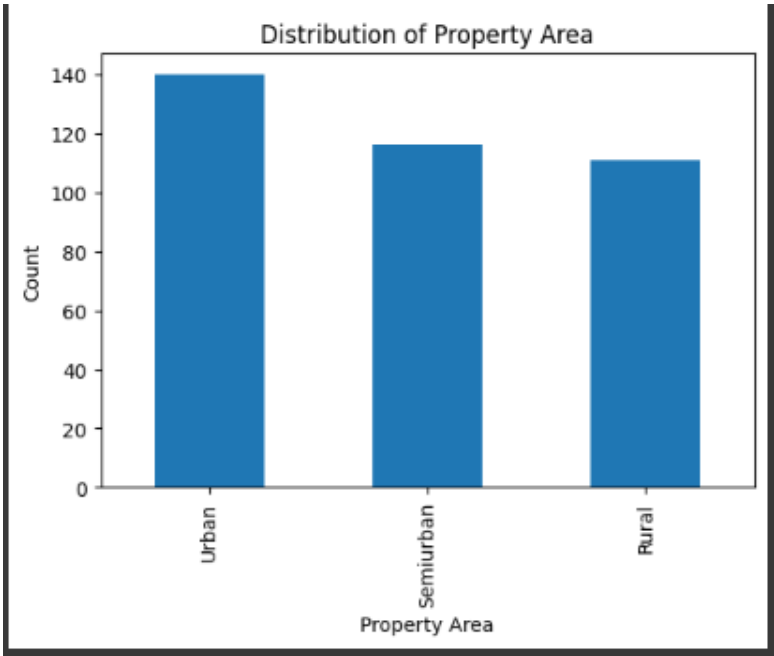
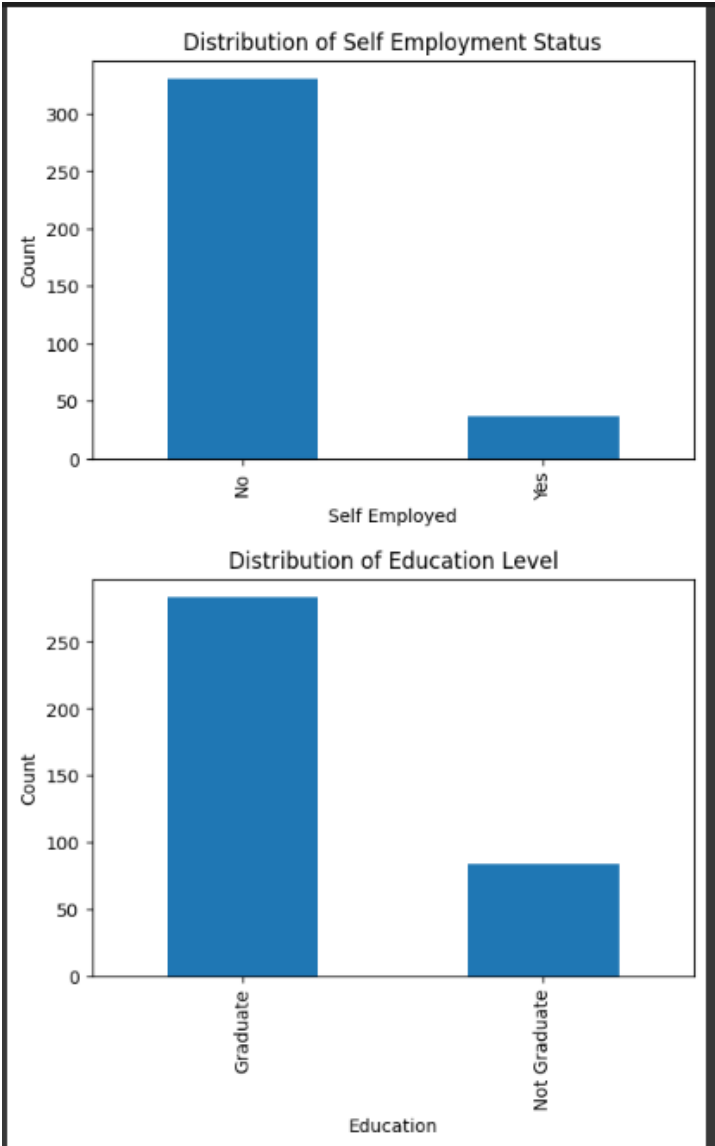
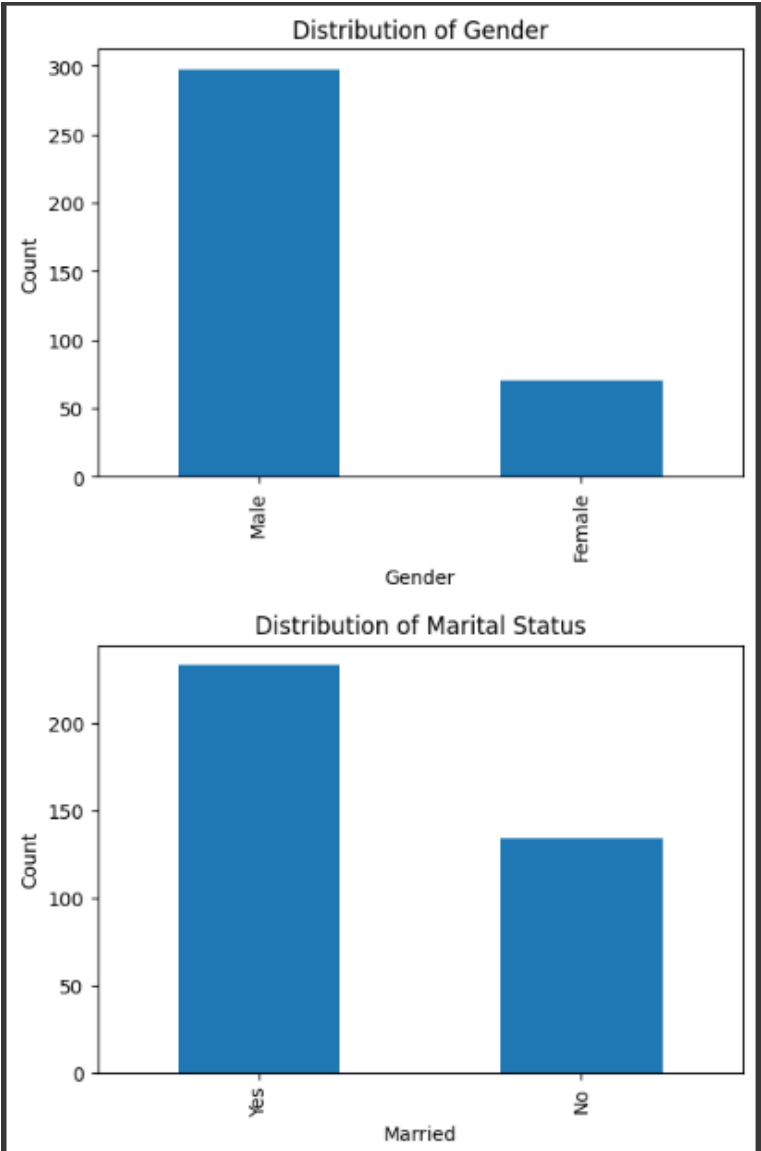


Key insights

- **Applicant Income:** The distribution is heavily skewed to the right, with a long tail indicating a significant number of applicants with higher incomes. The majority of applicants fall within a lower income range.
 - **Co-applicant Income:** Similar to applicant income, the distribution is right-skewed. However, the concentration of co-applicant incomes around zero suggests that many loan applications do not include a co-applicant.
 - **Loan Amount :** The distribution appears relatively symmetrical, with a slight right skew. Most loan amounts fall within a central range, with some outliers on the higher end.
 - **Loan Amount Term:** The distribution is heavily concentrated around a specific term (likely the most common loan term offered). There are a few outliers indicating loans with shorter or longer terms.
- ## ❖ Potential Implications
- The right skewness in income distributions suggests that targeting marketing or loan products towards higher-income individuals could be beneficial.
 - The concentration of co-applicant income around zero highlights the importance of considering individual applicants without co-applicants.
 - The outliers in loan amount and loan amount term warrant further investigation to understand the specific circumstances surrounding those loans.



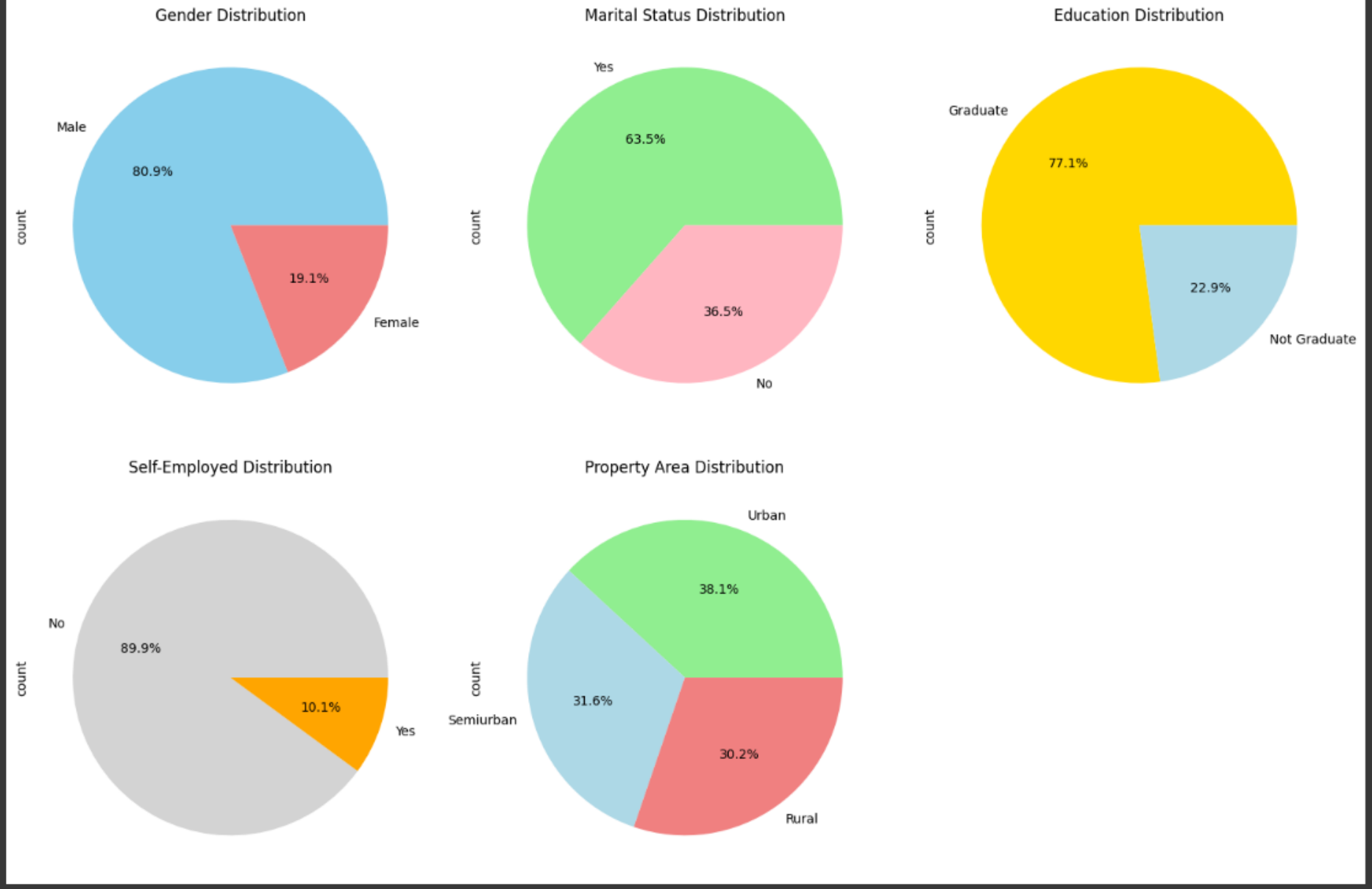
Visualize the frequency distribution of categorical variables.



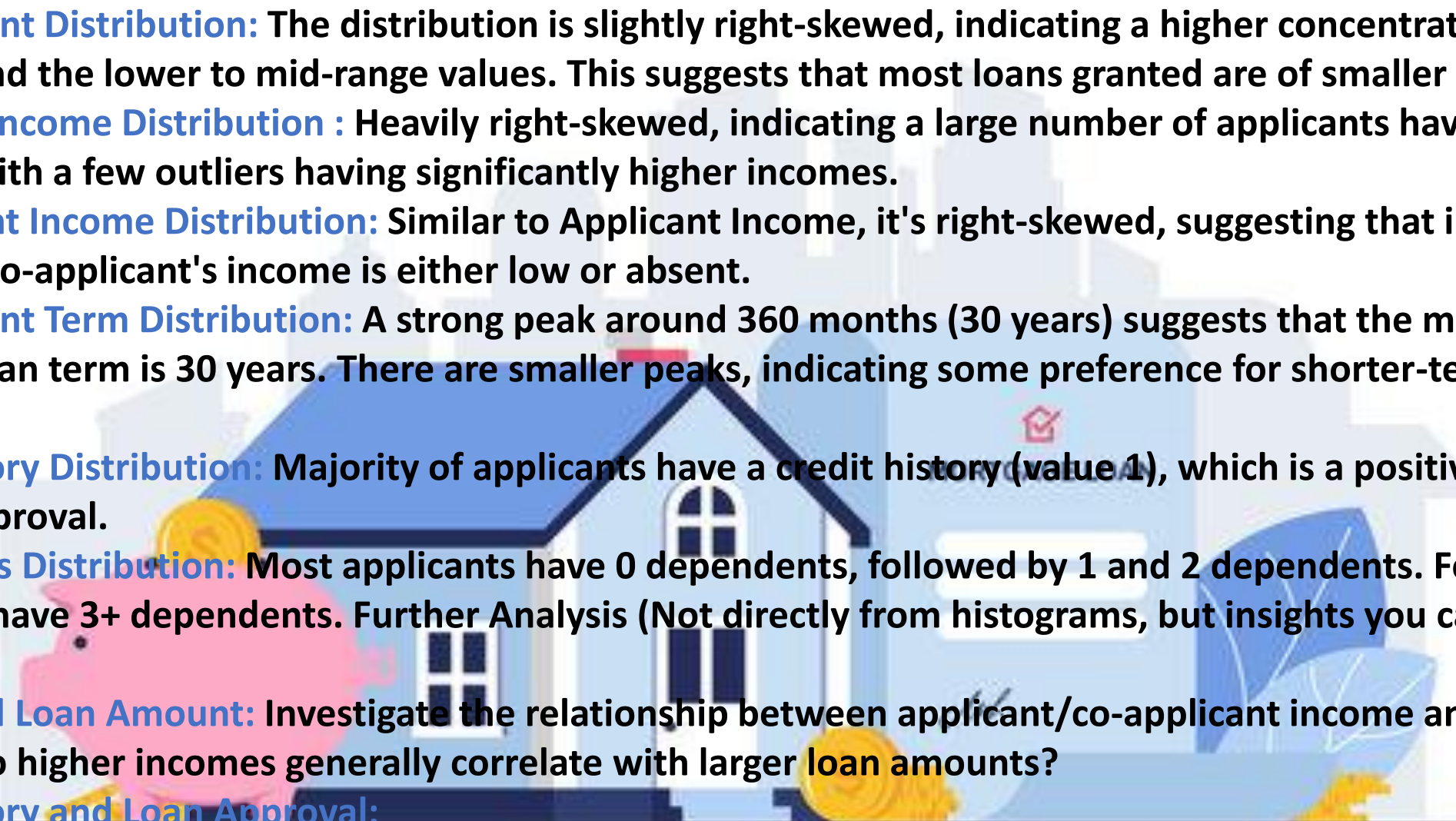
Key insights

- **Loan Amount Distribution:** The distribution of loan amounts is right-skewed, indicating a higher frequency of smaller loan amounts. The peak of the distribution suggests a common loan amount range. The long tail to the right indicates the presence of some larger loan amounts, potentially outliers.
- **Applicant Income Distribution:** The distribution of applicant incomes is also right-skewed, suggesting a higher concentration of lower-income applicants. The presence of outliers in the higher income range might influence loan approval decisions.
- **Co-applicant Income Distribution:** A significant portion of co-applicants have zero income, indicating a reliance on a single primary applicant's income in many cases. The right skew suggests that when co-applicants do contribute income, it's often lower than the primary applicant's.
- **Loan Amount Term Distribution:** The majority of loans have a term of 360 months (30 years), indicating a preference for long-term loans. Shorter-term loans are less common, suggesting they might be associated with specific loan types or applicant profiles.
- **Credit History Distribution:** A large proportion of applicants have a credit history (value of 1), which is a positive indicator for loan approval. The presence of applicants without credit history (value of 0) suggests a need for alternative assessment criteria for these cases.
- **Dependents Distribution:** The most common number of dependents is 0, followed by 1 and 2. This distribution reflects the family structures of loan applicants and might influence loan affordability assessments.

Represent the composition of categorical variables.

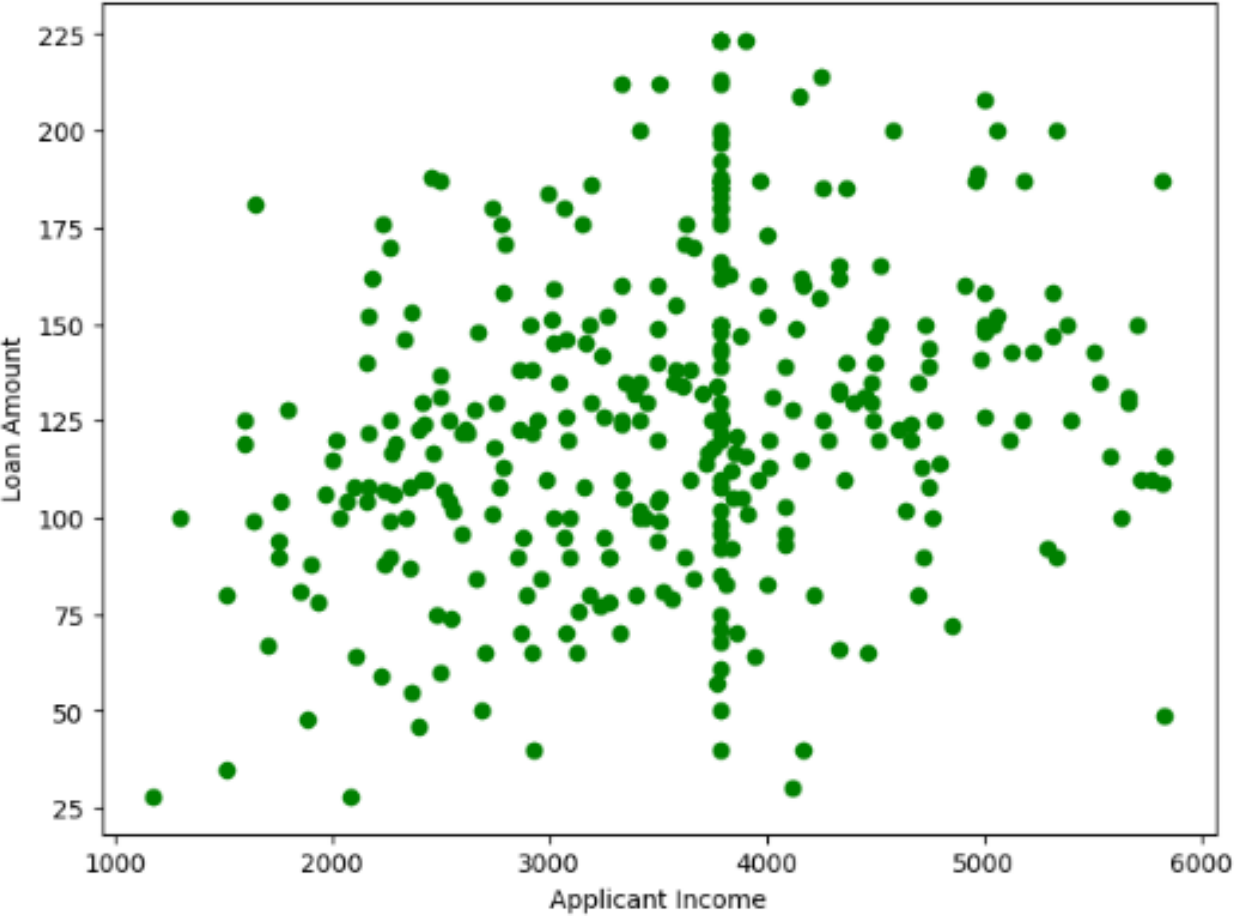


Key insights

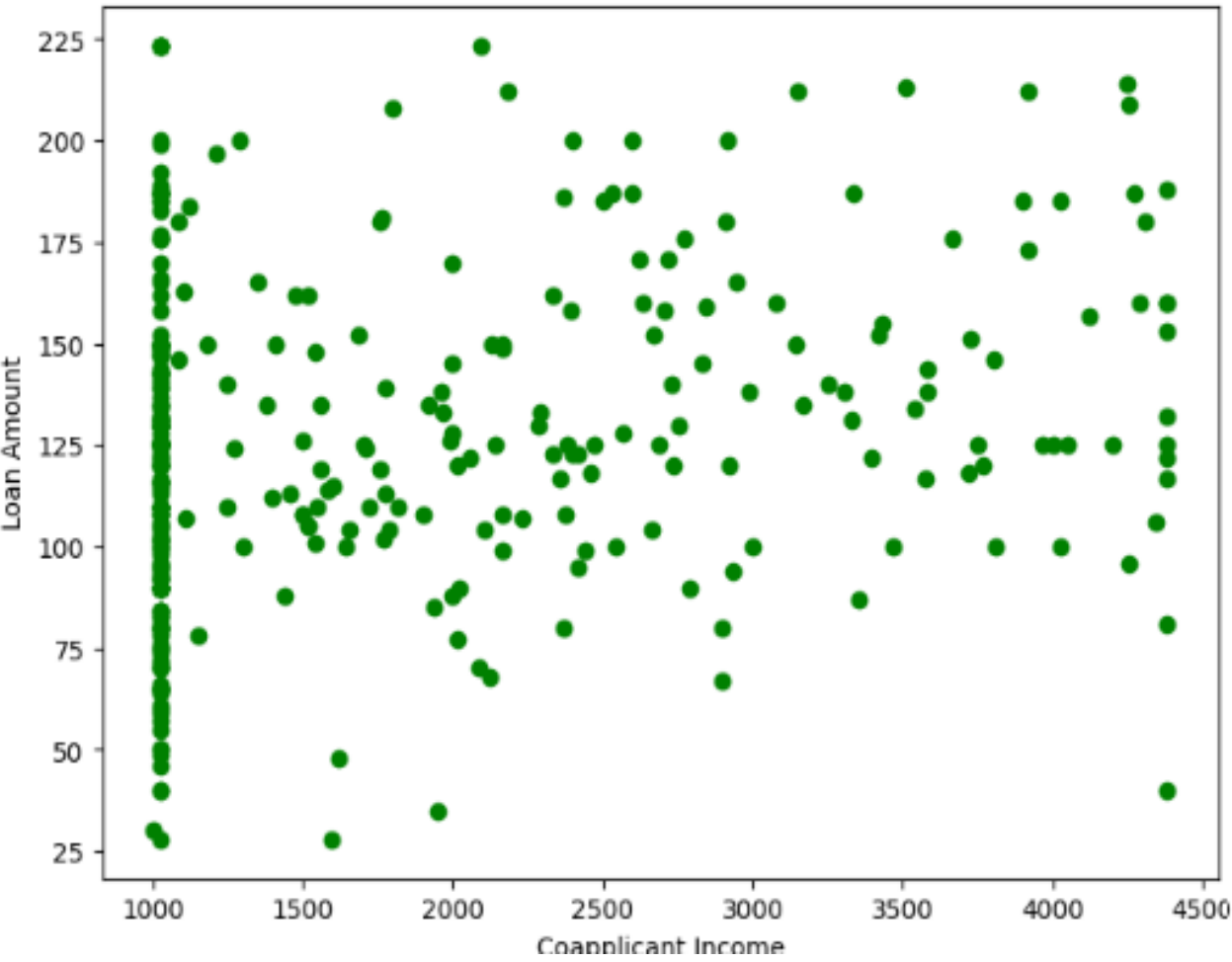
- 
- **Loan Amount Distribution:** The distribution is slightly right-skewed, indicating a higher concentration of loans around the lower to mid-range values. This suggests that most loans granted are of smaller amounts.
 - **Applicant Income Distribution :** Heavily right-skewed, indicating a large number of applicants have lower incomes, with a few outliers having significantly higher incomes.
 - **Co-applicant Income Distribution:** Similar to Applicant Income, it's right-skewed, suggesting that in many cases, the co-applicant's income is either low or absent.
 - **Loan Amount Term Distribution:** A strong peak around 360 months (30 years) suggests that the most common loan term is 30 years. There are smaller peaks, indicating some preference for shorter-term loans as well.
 - **Credit History Distribution:** Majority of applicants have a credit history (value 1), which is a positive sign for loan approval.
 - **Dependents Distribution:** Most applicants have 0 dependents, followed by 1 and 2 dependents. Fewer applicants have 3+ dependents. Further Analysis (Not directly from histograms, but insights you can derive)
 - **Income and Loan Amount:** Investigate the relationship between applicant/co-applicant income and loan amount. Do higher incomes generally correlate with larger loan amounts?
 - **Credit History and Loan Approval:**
 - Analyse the impact of credit history on loan approval rates
 - **Demographics and Loan Characteristics:** Explore how factors like gender, marital status, education, self-employment, and property area might influence loan amounts, terms, or approval rates.

Scatter plots to explore relationships between pairs of numeric variables.

Applicant Income vs Loan Amount



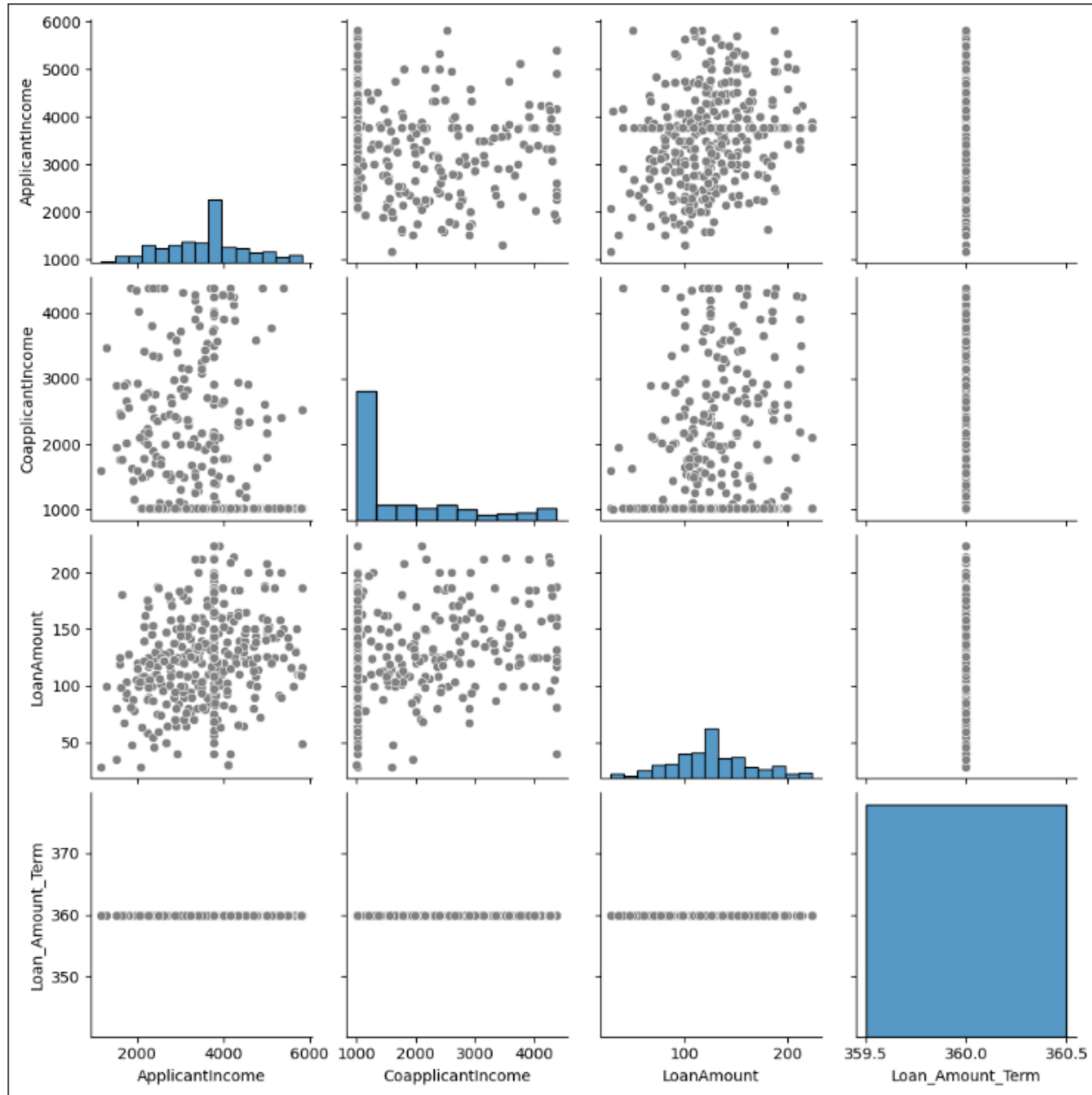
Coapplicant Income vs Loan Amount



Key insights

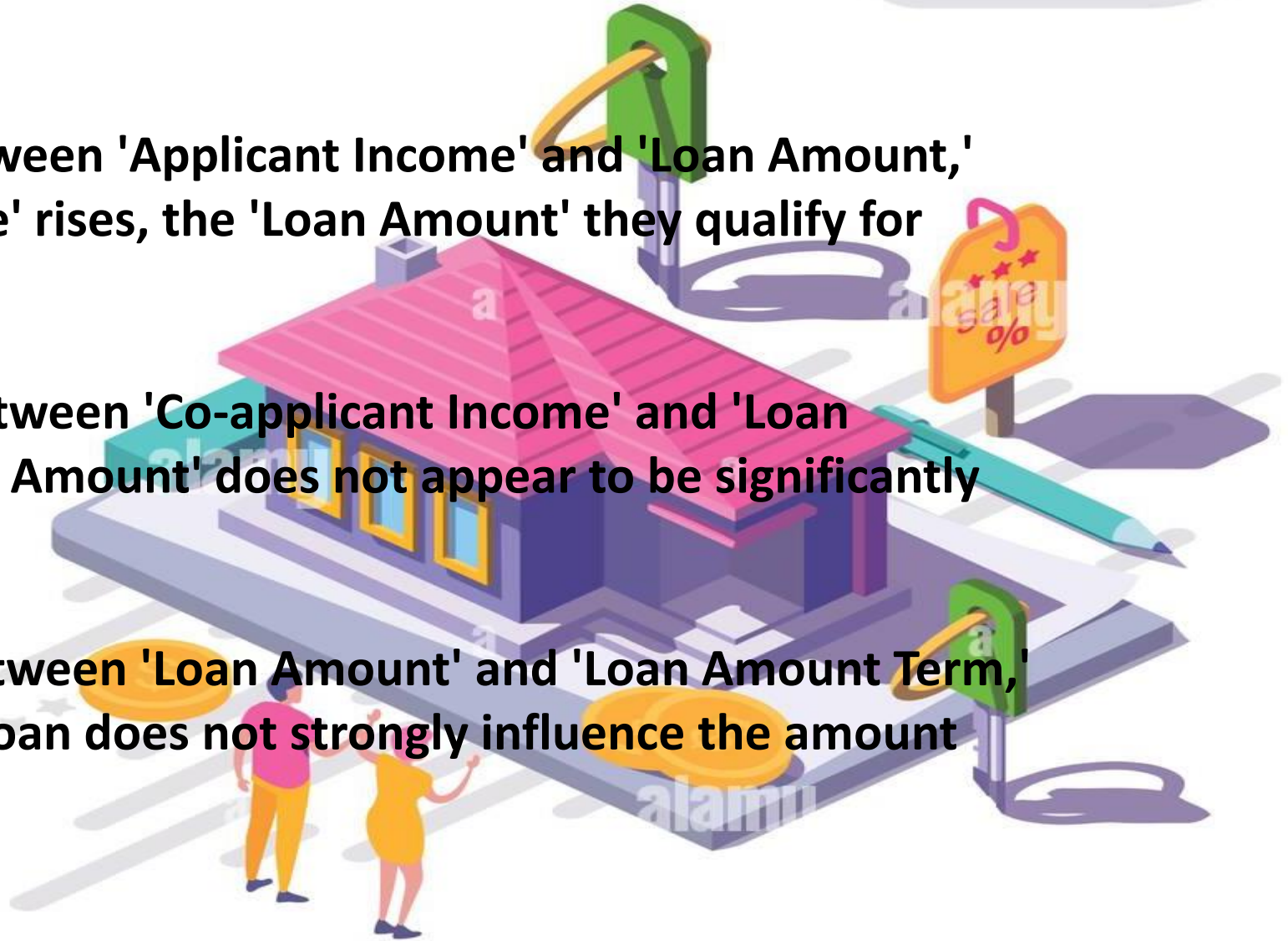
- **Applicant Income vs Loan Amount:** There's a slight positive correlation, indicating higher-income applicants tend to request larger loans. However, the correlation isn't very strong, suggesting other factors influence loan amount.
- **Co-applicant Income vs Loan Amount:** A weaker positive correlation than with Applicant Income, suggesting co applicant income plays a less significant role in loan amount determination.

Using scatter plot to visualize interactions between multiple numeric variables.

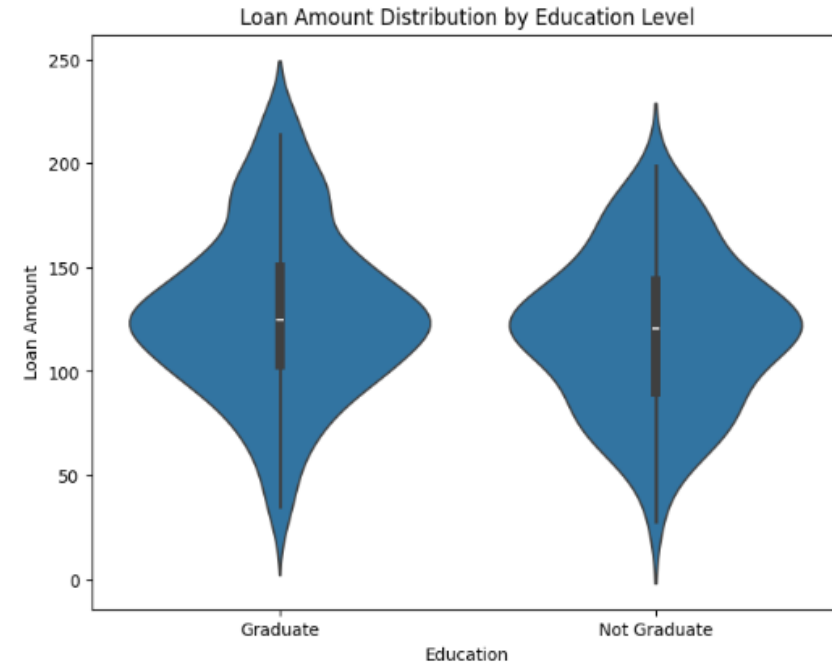
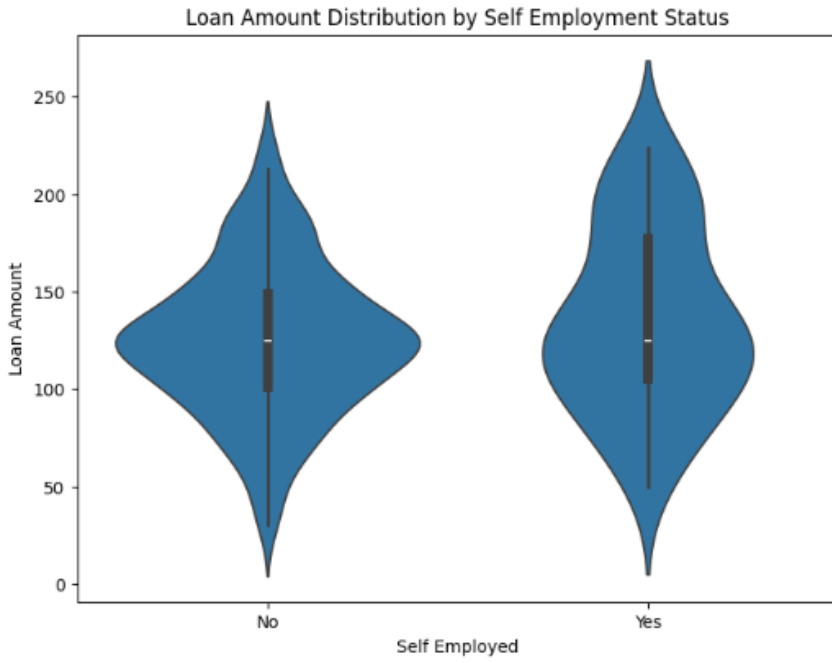
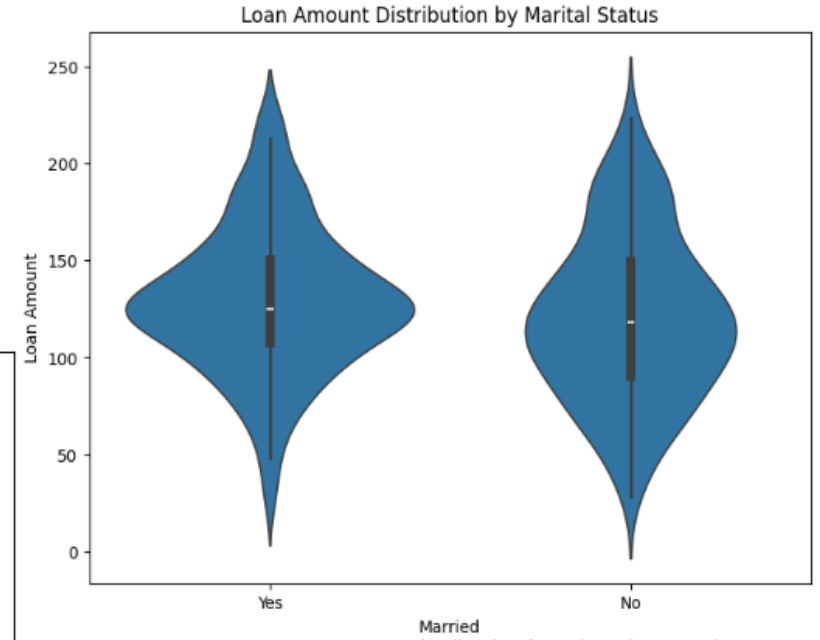
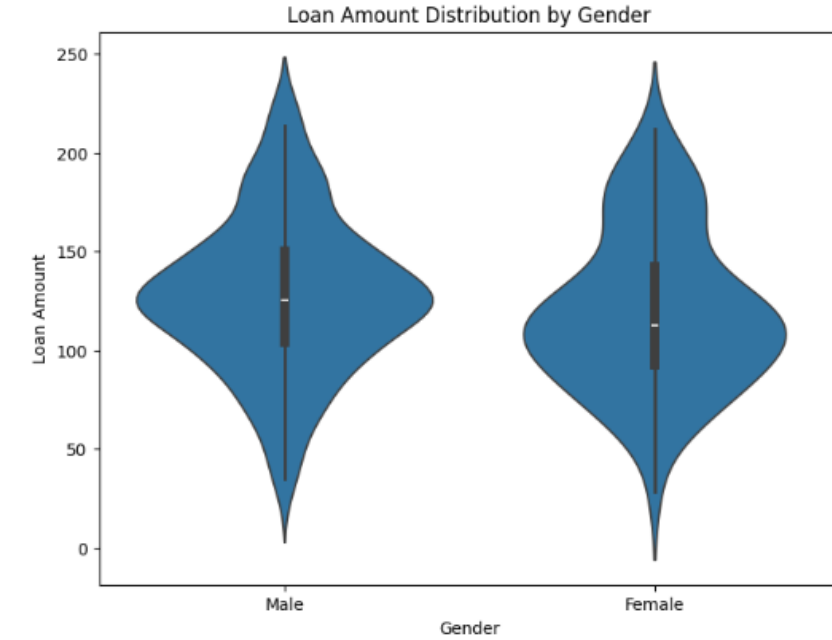


Key insights

- There is a positive correlation between 'Applicant Income' and 'Loan Amount,' meaning that as 'Applicant Income' rises, the 'Loan Amount' they qualify for generally increases as well.
- There is no notable correlation between 'Co-applicant Income' and 'Loan Amount,' indicating that the 'Loan Amount' does not appear to be significantly affected by 'Co-applicant Income.'
- There is no evident correlation between 'Loan Amount' and 'Loan Amount Term,' suggesting that the length of the loan does not strongly influence the amount borrowed.



Investigate the relationship between categorical and numeric variables violin plots.

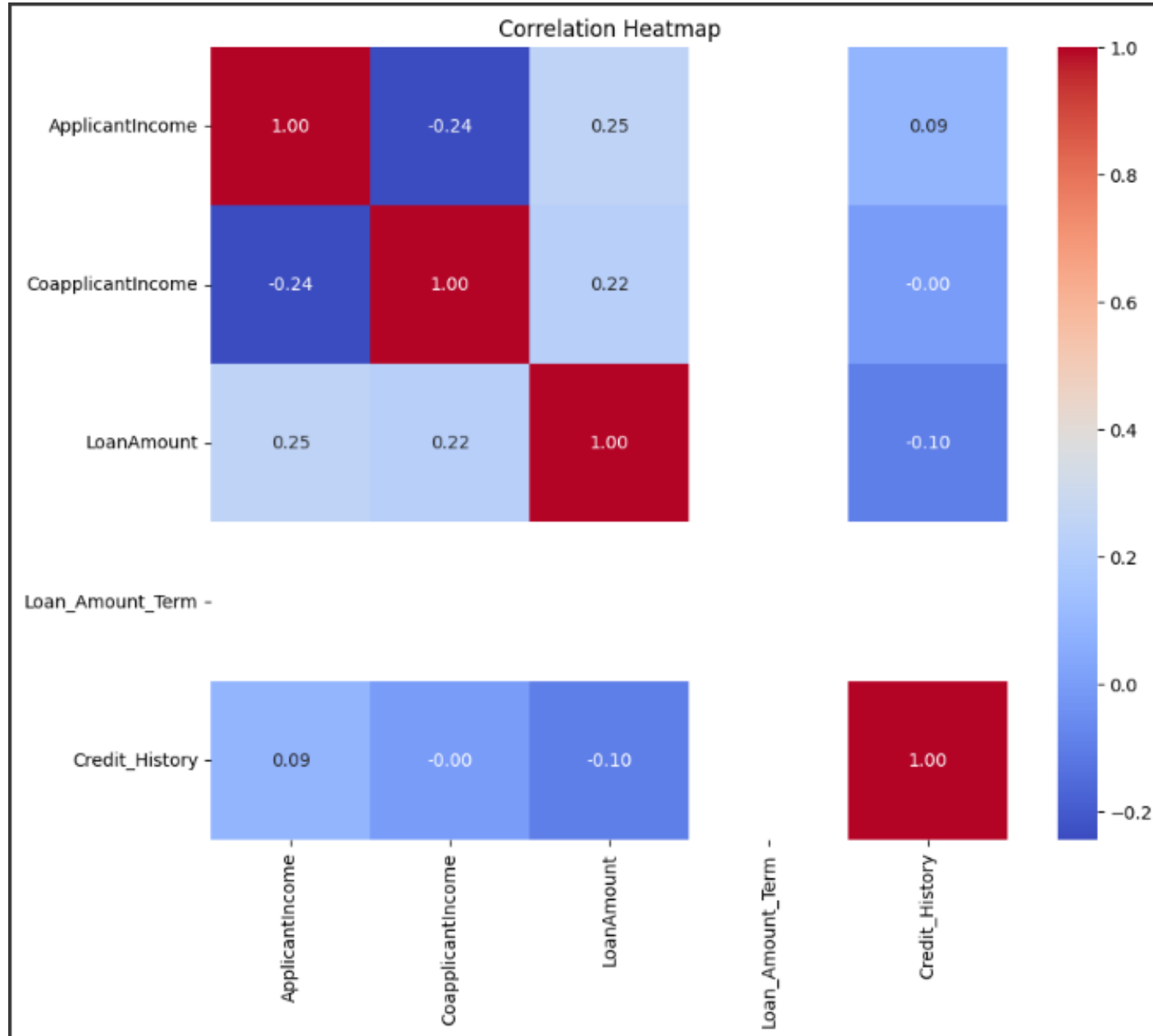


Key insights

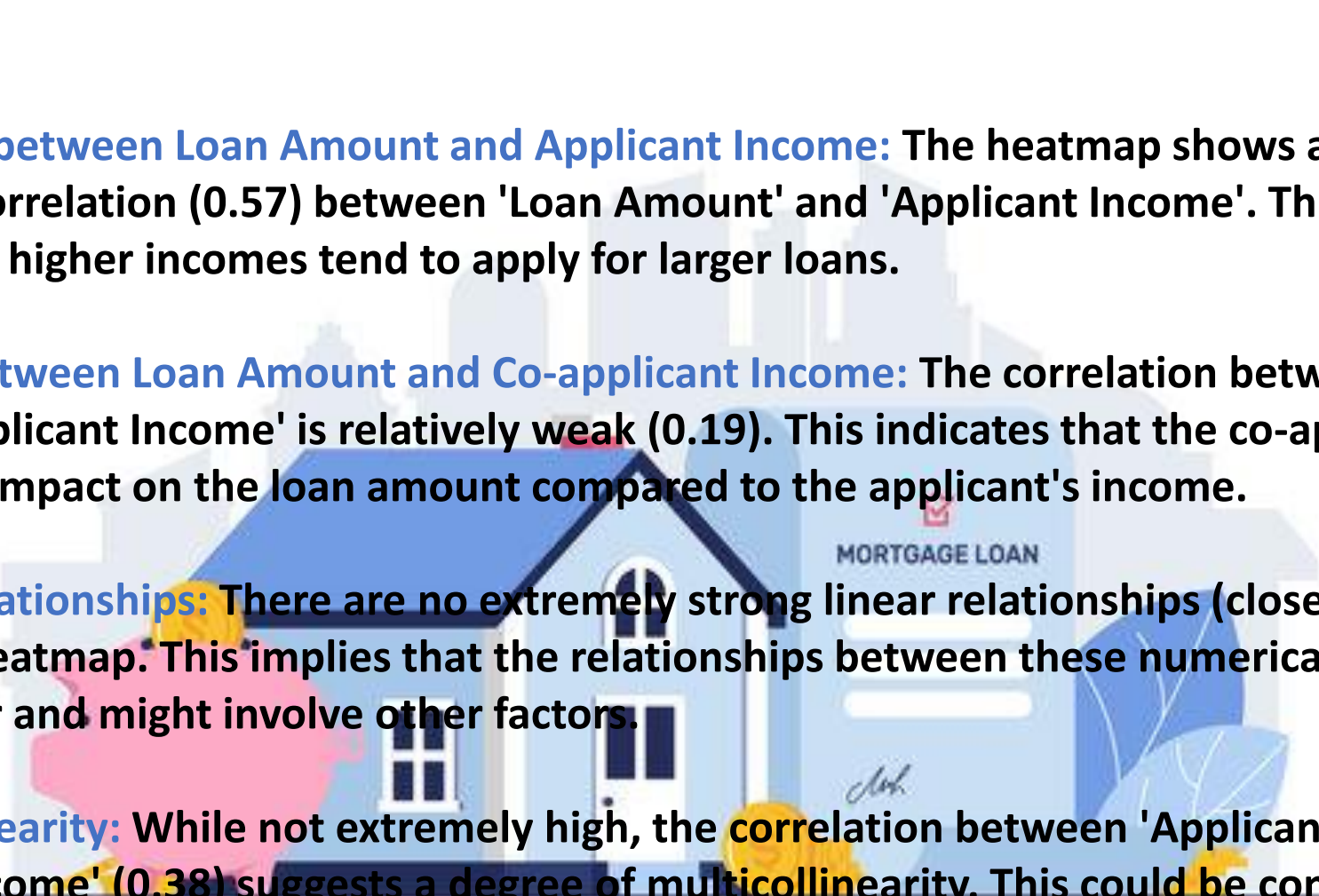
- **Loan Amount Variation Across Property Areas:** The plot shows how the distribution of loan amounts differs across property areas (Urban, Semi-urban, Rural). Observe the median, quartiles, and spread of loan amounts for each category. For example, you might find that median loan amounts are generally higher in urban areas compared to semi-urban or rural areas.
- **Outlier Analysis:** Look for potential outliers (data points far from the main distribution) in each property area. These outliers could indicate unusual loan applications that might require further investigation.
- **Property Area and Loan Amount Relationship:** Consider whether there's a clear pattern or trend in loan amounts based on property area.
- **Distribution Shape:** Analyse the shape of the violin plots for each property area. This can provide insights into the underlying factors influencing loan amounts in different areas.



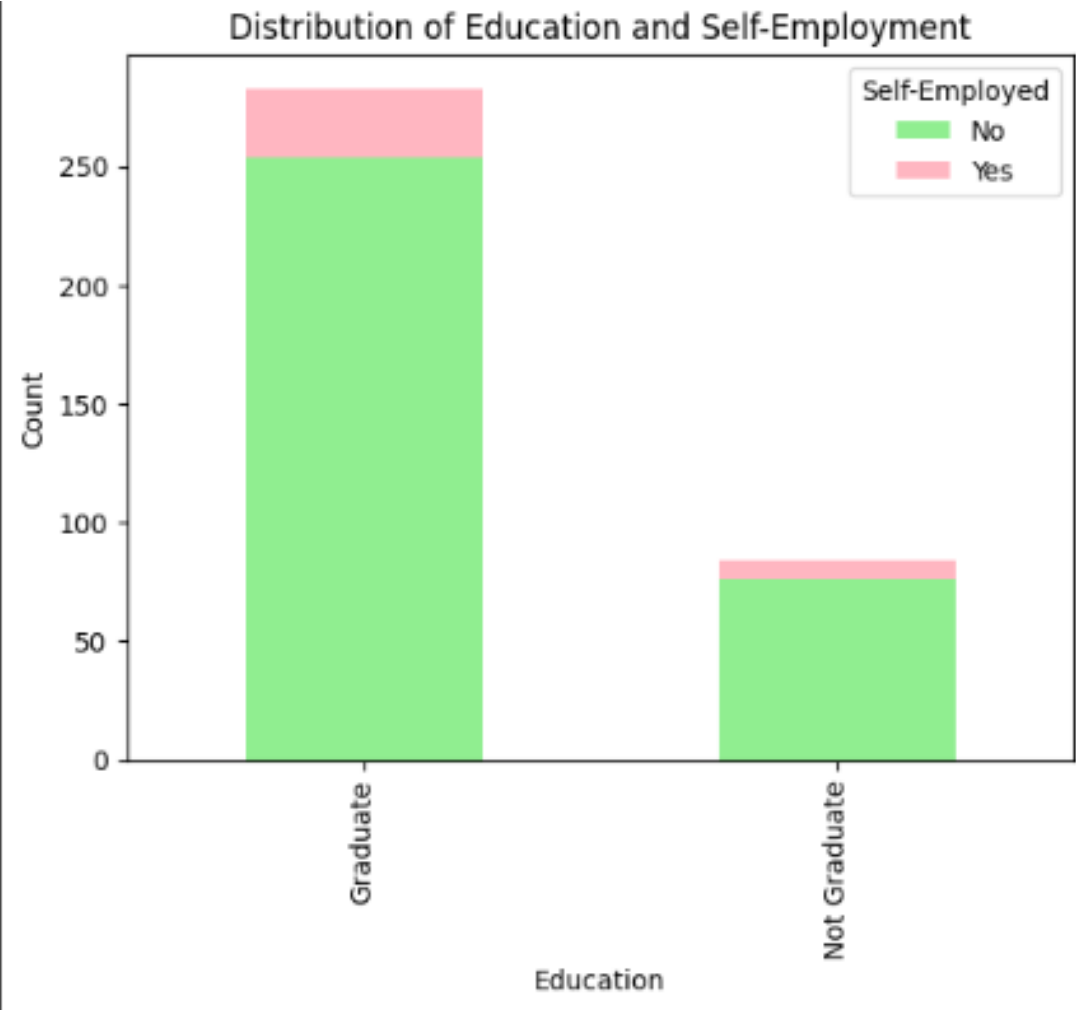
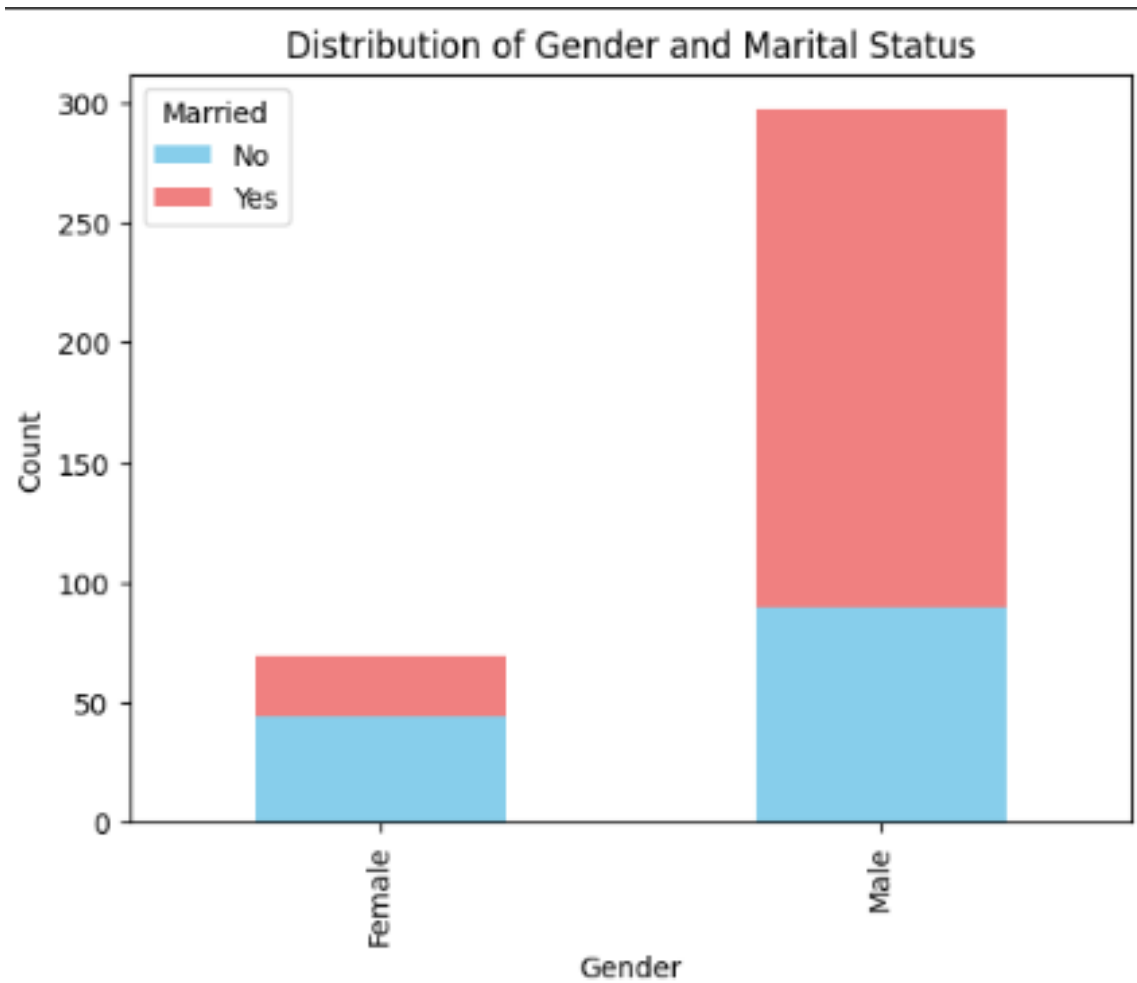
Correlation analysis to identify relationships between numeric variables.



Key insights

- 
- **Positive Correlation between Loan Amount and Applicant Income:** The heatmap shows a moderate positive correlation (0.57) between 'Loan Amount' and 'Applicant Income'. This suggests that individuals with higher incomes tend to apply for larger loans.
 - **Weak Correlation between Loan Amount and Co-applicant Income:** The correlation between 'Loan Amount' and 'Co-applicant Income' is relatively weak (0.19). This indicates that the co-applicant's income has a lesser impact on the loan amount compared to the applicant's income.
 - **No Strong Linear Relationships:** There are no extremely strong linear relationships (close to 1 or -1) observed in the heatmap. This implies that the relationships between these numerical variables are not strictly linear and might involve other factors.
 - **Potential Multicollinearity:** While not extremely high, the correlation between 'Applicant Income' and 'Co applicant Income' (0.38) suggests a degree of multicollinearity. This could be considered during feature selection for modelling, especially if using linear models sensitive to multicollinearity.

The distribution of categorical variables across multiple categories.



Key insights

- **Education and Self-Employment Relationship** : Most graduates are not self-employed, suggesting a preference for traditional employment. A significant portion of non-graduates are self-employed, indicating entrepreneurship in this group.
- **Potential Loan Targeting** : Financial institutions could tailor loan products for self-employed non-graduates, addressing their specific needs.



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.

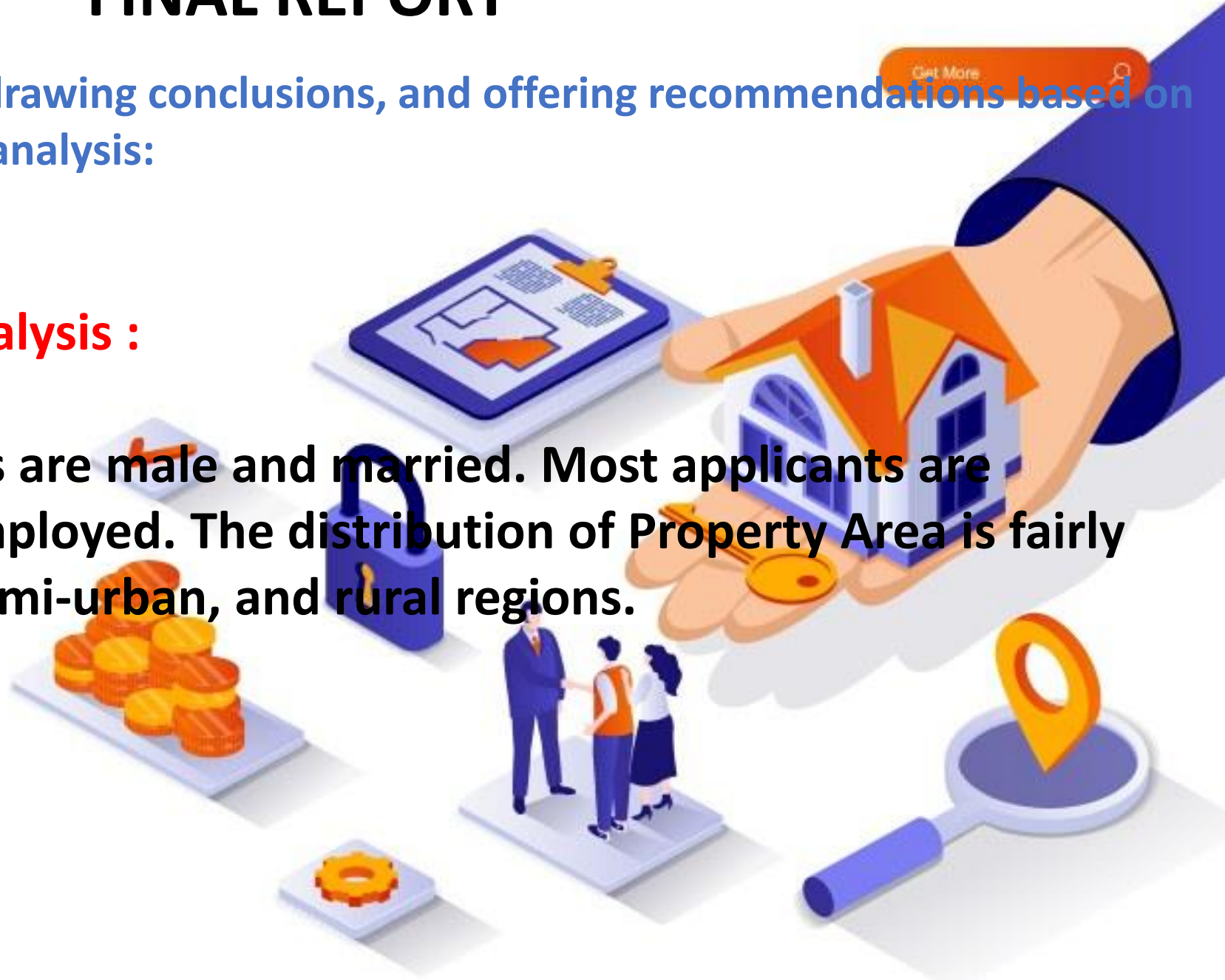


FINAL REPORT

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.

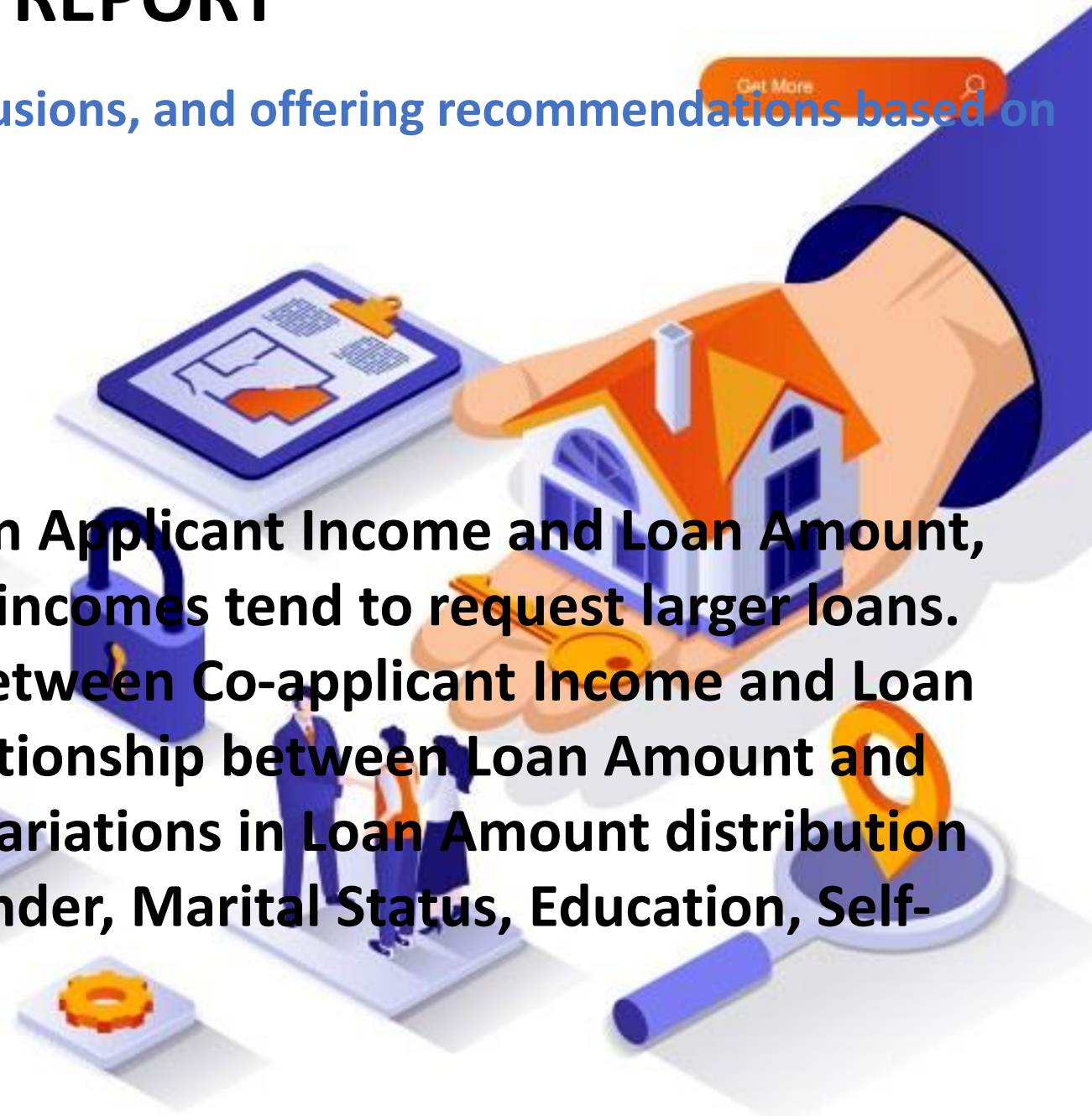


FINAL REPORT

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, Self-Employed status, and Property Area.

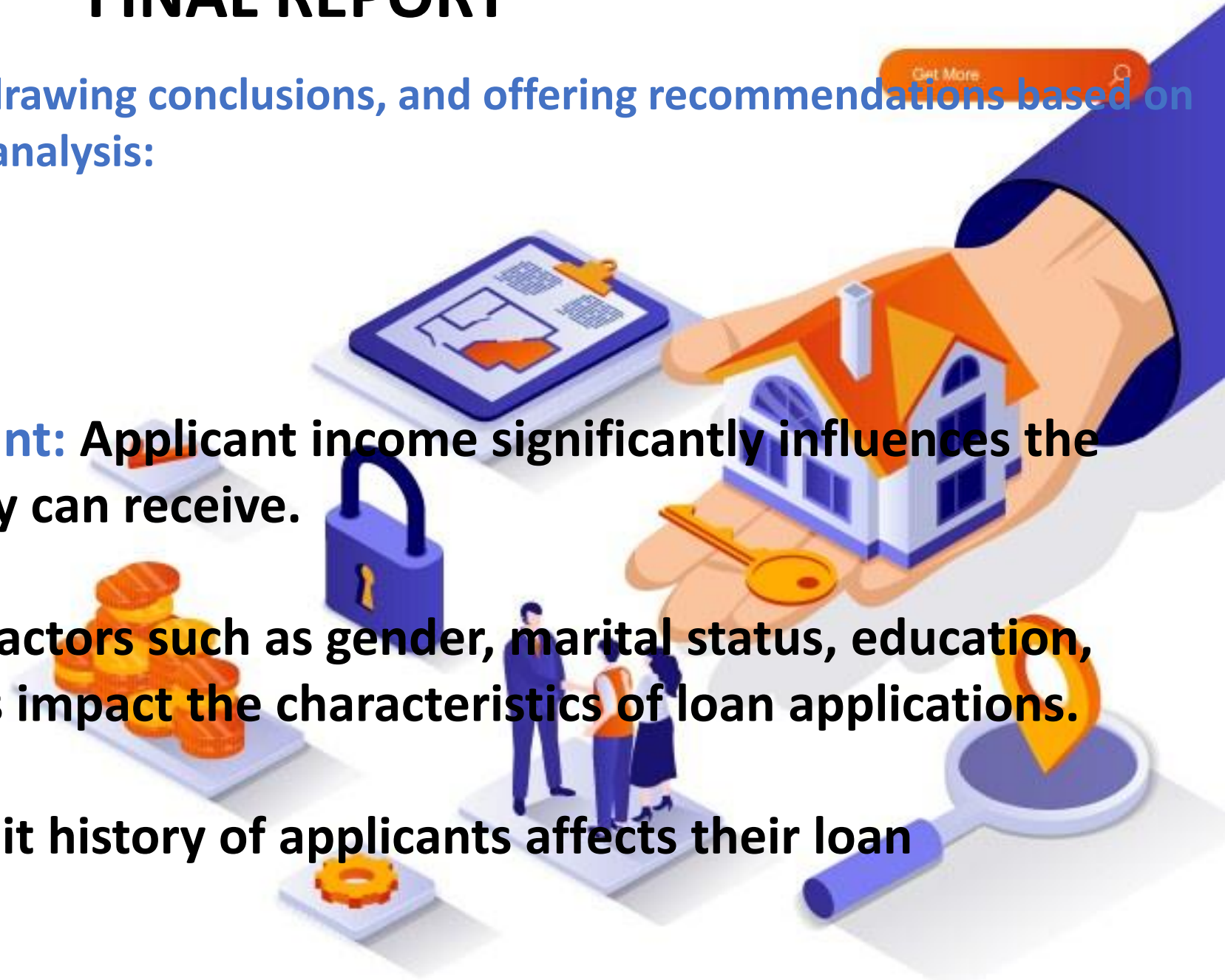


FINAL REPORT

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

❑ CONCLUSIONS:

- **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 section:-

<https://colab.research.google.com/drive/15YM1DIGKs9WAvYZJSp3OeaeZuEIBV157?usp=sharing>