EXPLORATORY DATA ANALYSIS

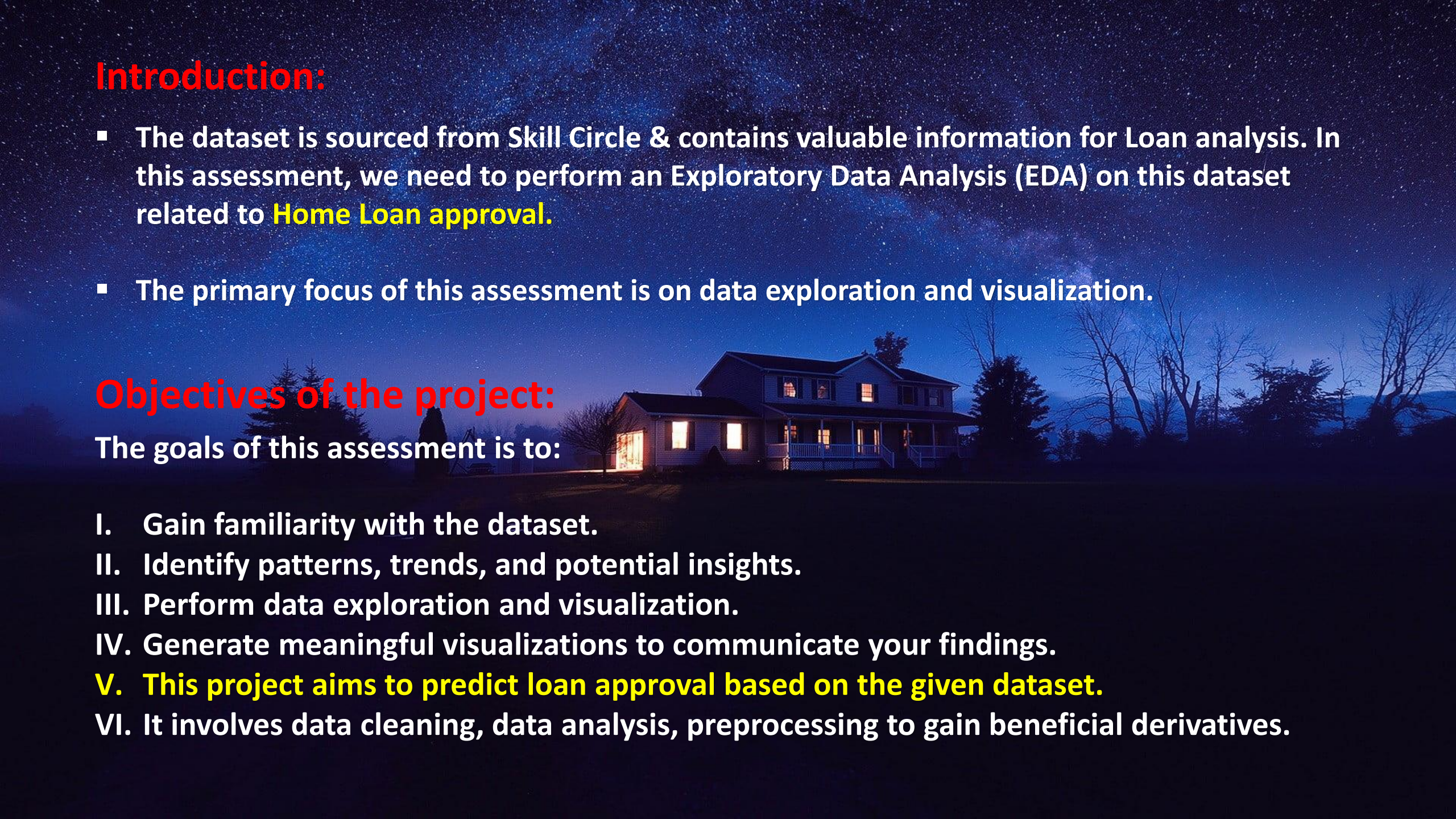
LOAN APPROVAL ANALYSIS

Project by
KUSHAGR SAXENA

Introduction:

- The dataset is sourced from Skill Circle & contains valuable information for Loan analysis. In this assessment, we need to perform an Exploratory Data Analysis (EDA) on this dataset related to Home Loan approval.

- 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, preprocessing to gain beneficial derivatives.

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 367 rows and 12 columns.
- For data cleaning, I have utilized libraries like Numpy, Pandas, Matplotlib, 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
```

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

```
Mounted at /content/drive
```

```
[3] data = '/content/drive/MyDrive/008 - My Projects/Loan Approval/Loan Sanction CSV.csv'
    df = pd.read_csv(data)
```

```
'''Let's drop any duplicate entries and check the shape of our dataset'''

df.drop_duplicates()
df.shape
```

```
(367, 12)
```

Description of Dataset:

The dataset includes information gathered from various borrowers, and upon initial examination, it appears to be related to home loans. We also notice that there are some Outliers and Missing values present in the data.

Key Features include:

- **Loan ID**: A unique identifier assigned to each loan application 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**: 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-employed or not (e.g., Yes, No).
- **Applicant Income**: The monthly income of the applicant.

```
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
```

Description of Dataset:

Key Features include:

- **Co-applicant Income**: The monthly income of the co-applicant, if any.
- **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, Semiurban, Rural).

```
'''Descriptive Statistics about our Dataset'''

df.describe()
```

	Dependents	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	357.000000	367.000000	367.000000	362.000000	361.000000	338.000000
mean	0.829132	4805.599455	1569.577657	136.132597	342.537396	0.825444
std	1.071302	4910.685399	2334.232099	61.366652	65.156643	0.380150
min	0.000000	0.000000	0.000000	28.000000	6.000000	0.000000
25%	0.000000	2864.000000	0.000000	100.250000	360.000000	1.000000
50%	0.000000	3786.000000	1025.000000	125.000000	360.000000	1.000000
75%	2.000000	5060.000000	2430.500000	158.000000	360.000000	1.000000
max	3.000000	72529.000000	24000.000000	550.000000	480.000000	1.000000

Data Cleaning & Pre-Processing:

There are total *84 Null values* present in our dataset. Out of which 34 values are from Categorical features and 50 values are from Numerical features.

Gender : As 11 Null values are present in the categorical 'Gender' attribute, Filling the missing values with 'Mode' can help maintain data completeness. Here, Mode is 'Male'.

```
df['Gender'] = df['Gender'].fillna(df['Gender'].mode()[0])
```

Self Employed : For another categorical 'Self Employed' attribute with 23 Null values, filling them with 'Mode' can be a feasible approach. Here, Mode is 'No'.

```
df['Self_Employed'] = df['Self_Employed'].fillna(df['Self_Employed'].mode()[0])
```

```
'''Let's find Null/Missing values

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

Data Cleaning & Pre-Processing:

Dependents : As this feature didn't contain any Outliers and is **Positively-skewed**, filling its 10 Null values with '**Median**' can be a cakewalk.

```
median_value = df['Dependents'].median()
median_value
```

```
0.0
```

```
df['Dependents'] = df['Dependents'].fillna(median_value).astype(float)
```

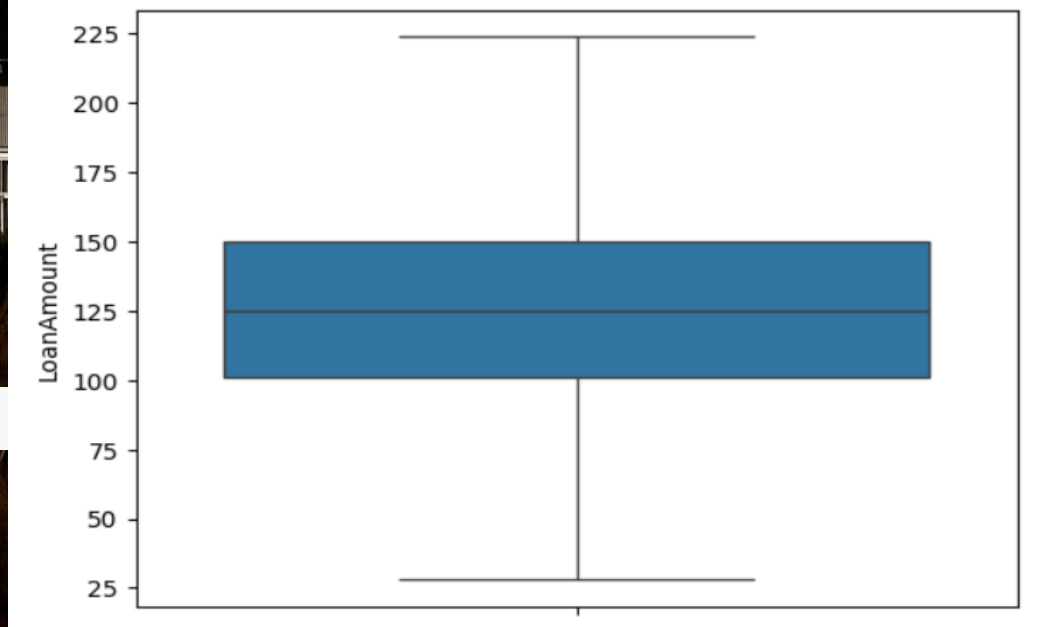
Loan Amount: This feature has both Outliers and Null values present in it. Firstly, 5 Null values has been filled with '**Median**' and then the Outliers has been handled by using **Inter-Quartile Range(IQR)** Method.

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

```
125.0
```

```
df['LoanAmount'] = df['LoanAmount'].fillna(median_loan_amount).astype(float)
```

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



Data Cleaning & Pre-Processing:

Loan Amount Term : Similarly, This feature also contains both Outliers & Null values. Initially, the 6 Null values were filled with the '**Median**', and subsequently, Outliers were addressed using the **Inter-Quartile Range (IQR) Method**.

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

```
360.0
```

```
df['Loan_Amount_Term'] = df['Loan_Amount_Term'].fillna(median_loan_term).astype(float)
```

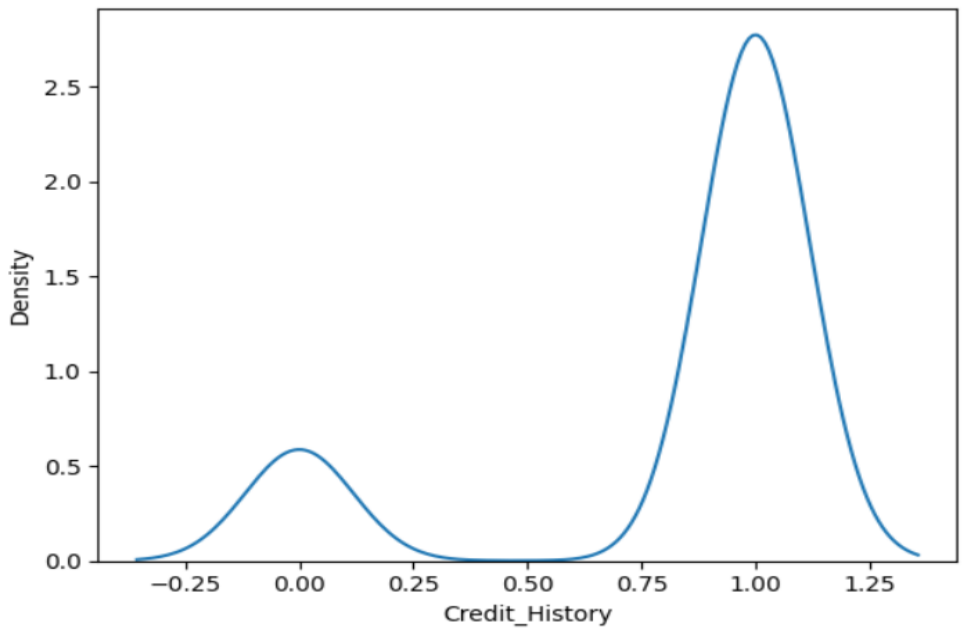
Credit History: For another Numerical feature with highest 29 Null values, filling values with '**Median**' is an easy approach because it didn't contain any Outliers & is **Negatively-skewed**.

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

```
1.0
```

```
df['Credit_History'] = df['Credit_History'].fillna(credit_median).astype(int)
```

```
sns.kdeplot(df['Credit_History'])
plt.show()
```



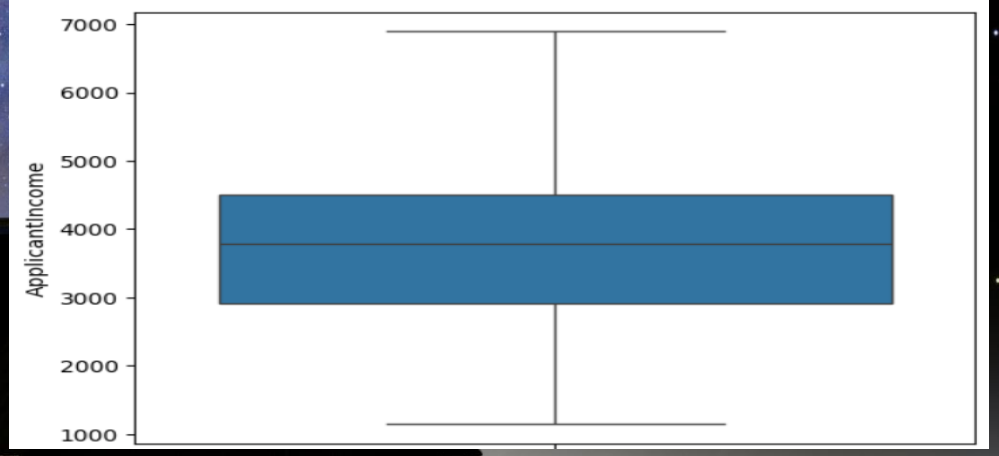
Data Cleaning & Pre-Processing:

Point to Ponder - As all the Null values has been handled, we still have two Numerical features (‘Applicant Income’, ‘Co-applicant Income’) left to check at least for Outliers to maintain data equilibrium for better insights.

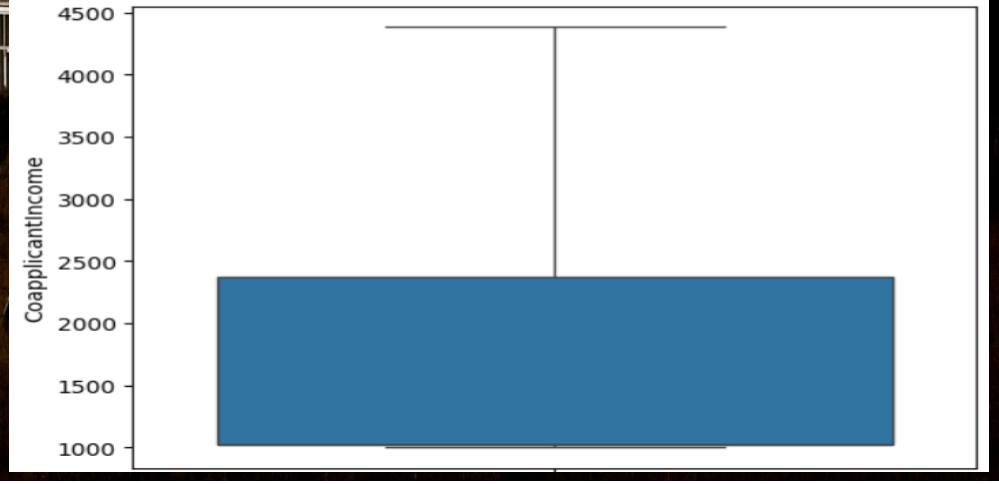
Applicant Income : This feature did contain a few Outliers which has been handled using the **Inter-Quartile Range (IQR) Method**. Note that a few applicant had 0 income, which is practically not possible in Home loan, so I have replaced that with ‘**Median**’ too.

Co-Applicant Income: Likewise, Outliers had been handled here too with **Inter-Quartile Range (IQR) Method**. Please note that some Co-applicants also reported an income of 0, which is again unrealistic for Home loan. I have replaced these values with the ‘**Median**’ as well.

```
sns.boxplot(df['ApplicantIncome'])
plt.show()
```




```
sns.boxplot(df['CoapplicantIncome'])
plt.show()
```

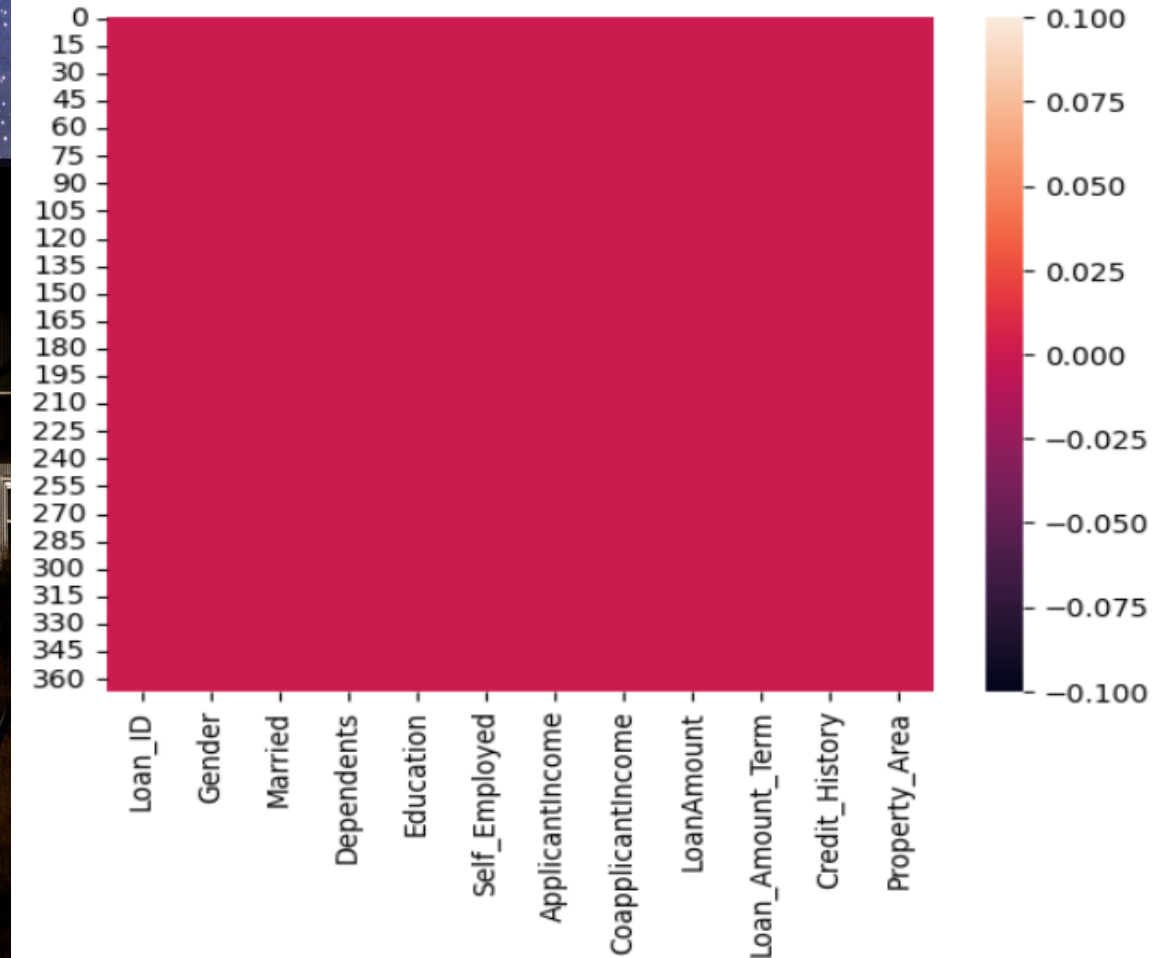


Data Cleaning & Pre-Processing:

Summary : To summarize, addressing Null values and Outliers necessitates a methodical approach tailored to the data's characteristics and specific attributes. By applying the outlined strategies, we can efficiently manage and fill in the missing values, thereby ensuring the dataset's completeness, integrity, and reliability for future analysis and insights.

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

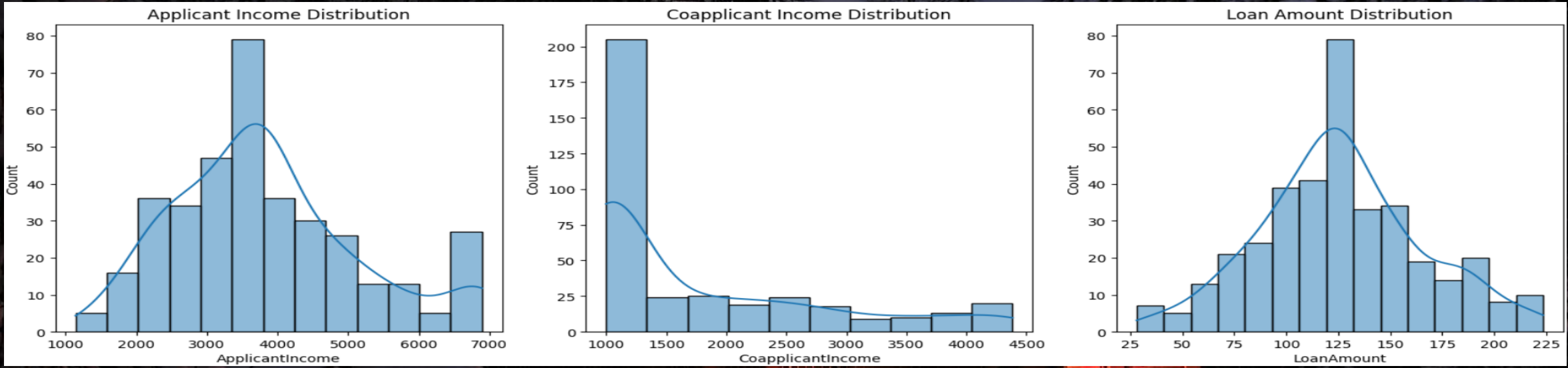
```
'''Let's draw a HEATMAP to ensure all Null values has been handled'''
sns.heatmap(df.isnull())
plt.show()
```

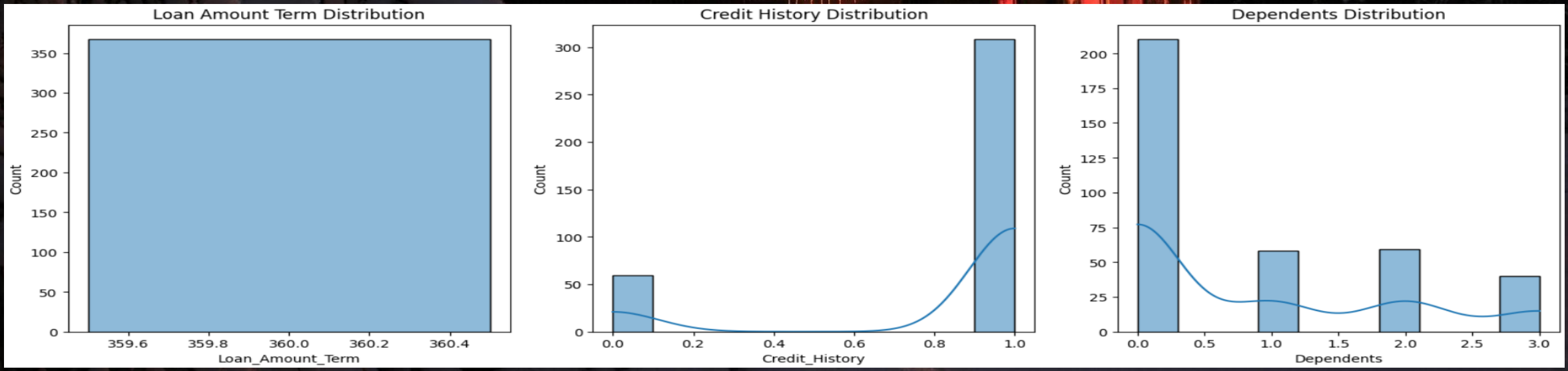


Data Visualization and Insights

Explore the distribution of numeric columns using the following visualizations:

Histograms: Plot the frequency distribution of key Numeric variables.





Data Visualization and Insights

Explore the distribution of numeric columns using the following visualizations:
Histograms: Plot the frequency distribution of key Numeric variables.

Key insights

- **Applicant Income:** The distribution is right-skewed, indicating a higher concentration of applicants with lower incomes. This suggests that the majority of loan applicants come from a specific income bracket.

- **Co-applicant Income:** Similar to applicant income, the co-applicant income distribution is also right-skewed, Implying that most co-applicants also have lower incomes.

- **Loan Amount:** The loan amount distribution shows a peak around a certain value, indicating a common loan amount requested by applicants. This could reflect the average affordability or typical loan requirements in the market.

- **Loan Amount Term:** The loan amount term distribution reveals the most frequent loan durations chosen by applicants. This information can help understand the preferred repayment timelines for loans.

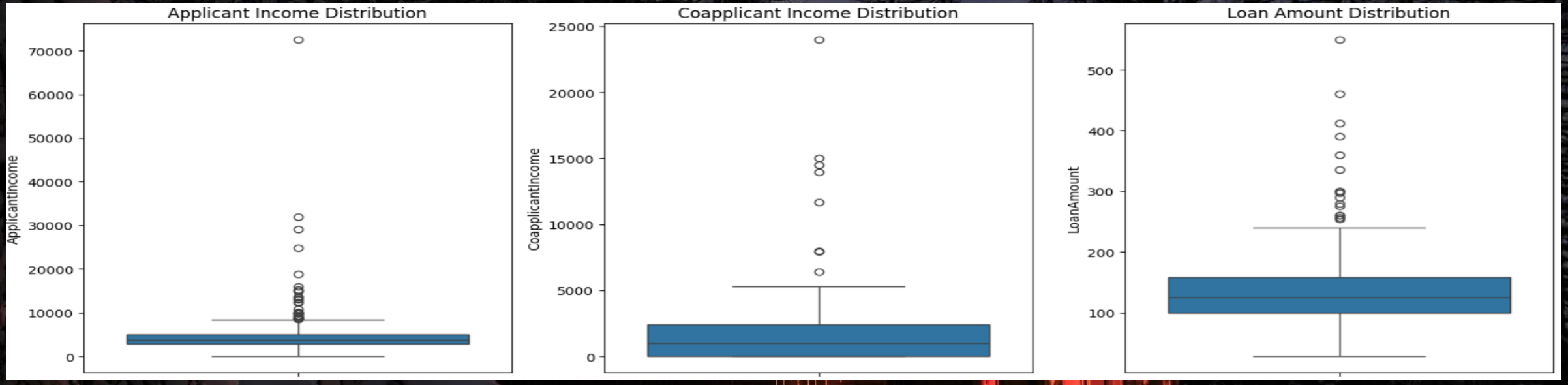
- **Credit History:** The credit history distribution shows a clear distinction between applicants with and without a credit history. This highlights the importance of credit history in loan approval decisions.

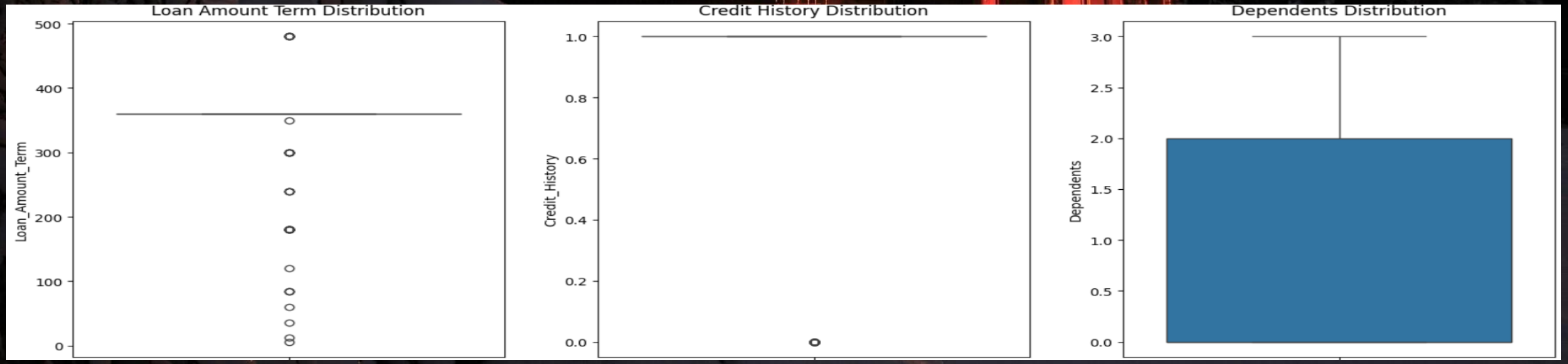
- **Dependents:** The dependents distribution indicates the number of dependents most commonly declared by applicants. This information can provide insights into the family structures of loan applicants.

Data Visualization and Insights

Explore the distribution of numeric columns using the following visualizations:

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





Data Visualization and Insights

Explore the distribution of numeric columns using the following visualizations:

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

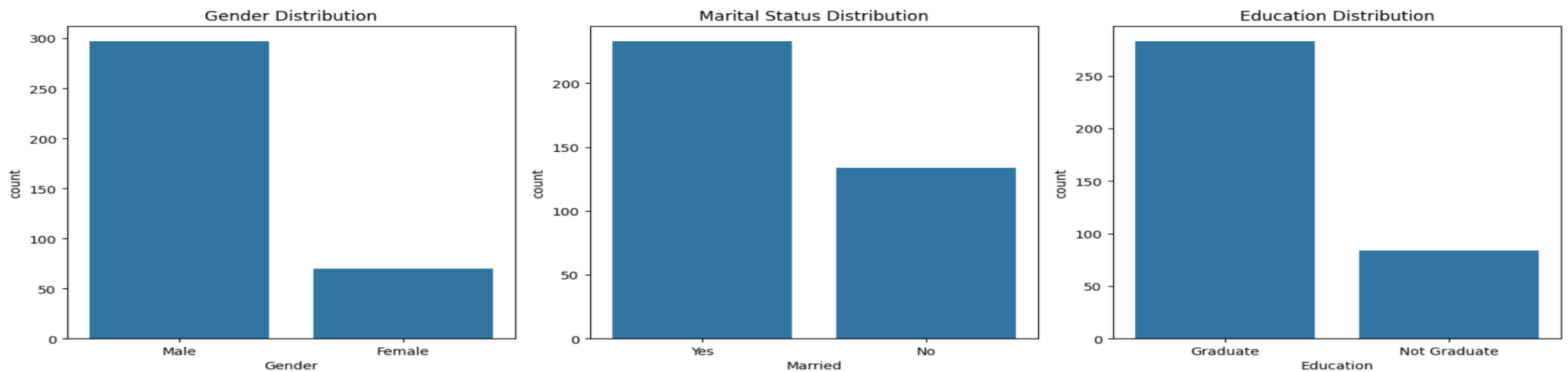
Key insights

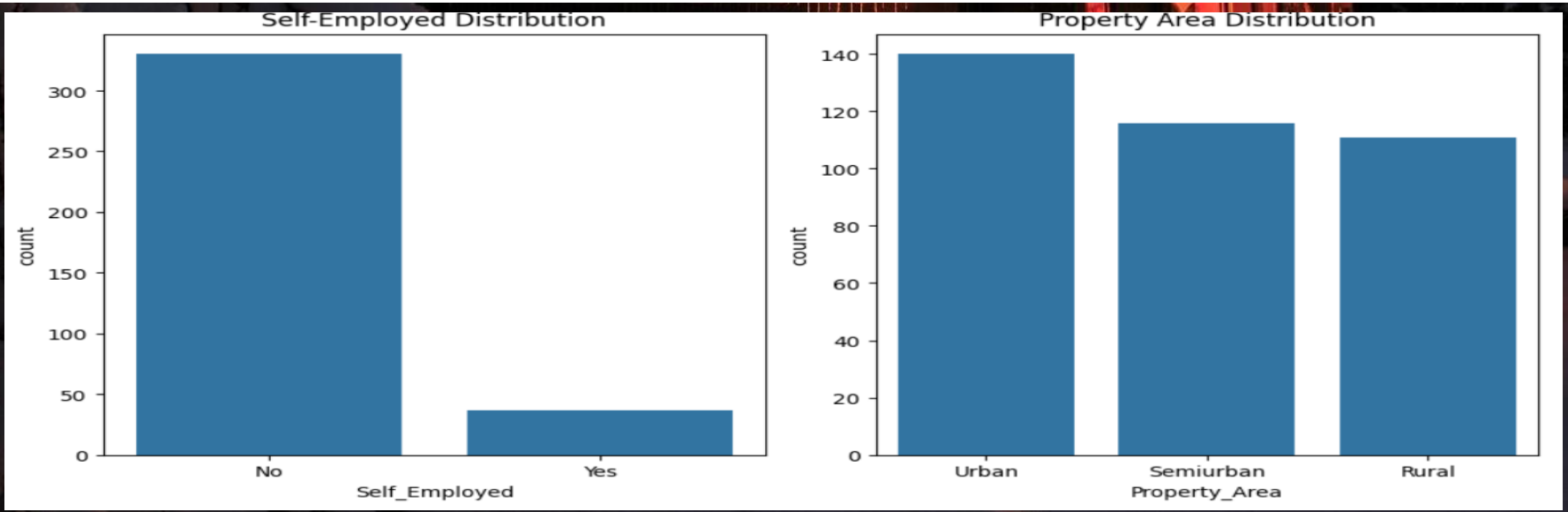
- **Applicant Income:** The box plot shows a number of outliers on the higher end of the income spectrum, suggesting there are some applicants with significantly higher incomes than the majority. The median income appears to be approximately around ₹3,800.
- **Co-applicant Income:** Similar to Applicant Income, there are outliers indicating some co-applicants have substantially higher incomes. The median co-applicant income seems to be lower than the applicant's, around ₹1,100.
- **Loan Amount:** The distribution of loan amounts is relatively symmetrical, with a median value around ₹125. There are a few outliers on the higher end, indicating some individuals are applying for larger loans.
- **Loan Amount Term:** Most loan terms are clustered around 360 months (30 years), with a few outliers representing shorter-term loans.
- **Credit History:** This plot clearly shows the majority of applicants have a credit history (value of 1). There's a smaller group with no credit history (value of 0).

Data Visualization and Insights

Analyze categorical variables by creating the following plots:

Bar Charts: Visualize the frequency distribution of categorical variables.





Data Visualization and Insights

Analyze categorical variables by creating the following plots:

Bar Charts: Visualize the frequency distribution of categorical variables.

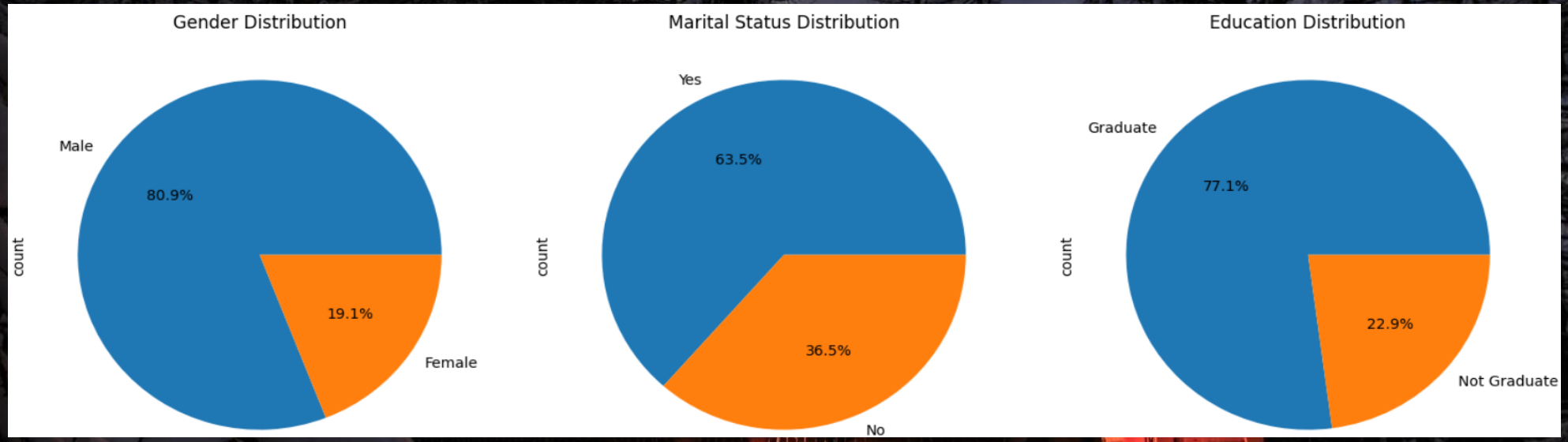
Key insights

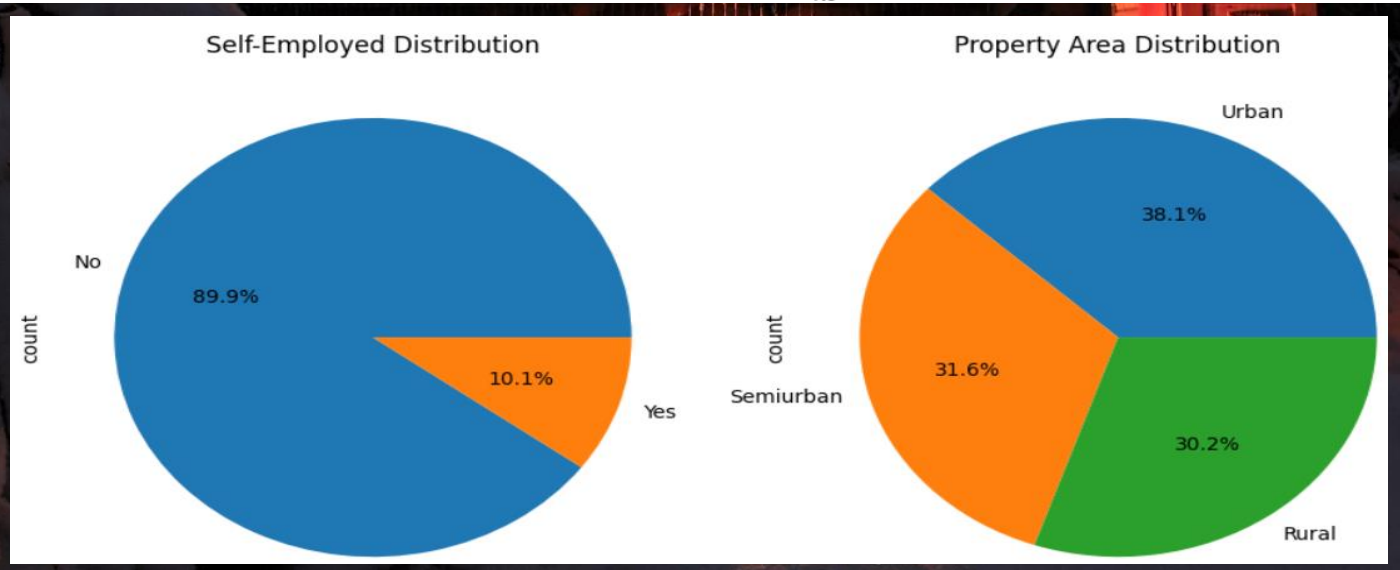
- **Gender:** The majority of loan applicants are male, indicating a potential gender bias in loan applications. This could be further investigated to understand the reasons behind this disparity.
- **Marital Status:** A significant portion of loan applicants are married, suggesting that married individuals may have a higher demand for loans. This could be attributed to factors like family responsibilities or joint financial planning.
- **Education:** Most loan applicants are graduates, implying that higher education levels might be associated with increased loan applications. This could reflect a greater awareness of financial products or a higher need for loans among educated individuals.
- **Self-Employed:** The majority of loan applicants are not self-employed, indicating that salaried individuals constitute a larger portion of loan seekers. This could be due to the perceived stability of salaried income compared to self-employment.
- **Property Area:** The distribution of property areas shows a relatively even distribution across urban, semi-urban, and rural areas. This suggests that loan applications are not significantly concentrated in any specific type of property area.

Data Visualization and Insights

Analyze categorical variables by creating the following plots:

Pie Charts: Represent the composition of categorical variables.





Data Visualization and Insights

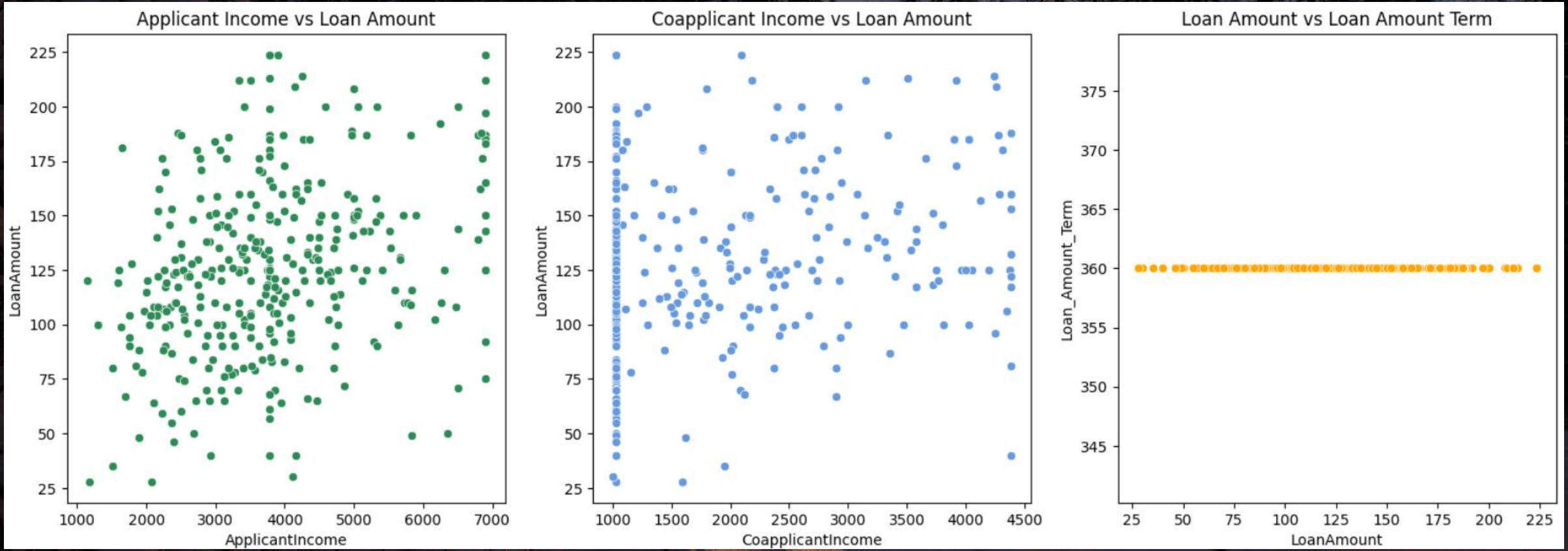
Analyze categorical variables by creating the following plots:
Pie Charts: Represent the composition of categorical variables.

Key insights

- **Gender Distribution:** Around 81% of the loan applicants are male, indicating a significant gender imbalance in loan applications.
- **Marital Status Distribution:** Approximately 65% of the applicants are married, suggesting that married individuals may be more likely to apply for loans.
- **Education Distribution:** A majority (around 78%) of the applicants are graduates, indicating a higher likelihood of loan applications from individuals with higher education levels.
- **Self-Employed Distribution:** Only about 14% of the applicants are self-employed, implying that a majority of loan applications come from salaried individuals.
- **Property Area Distribution:** The distribution of applicants across property areas (Semiurban, Urban, Rural) is relatively balanced, with semiurban areas having a slightly higher proportion (around 38%).

Data Visualization and Insights

Create scatter plots to explore relationships between pairs of numeric variables.



Data Visualization and Insights

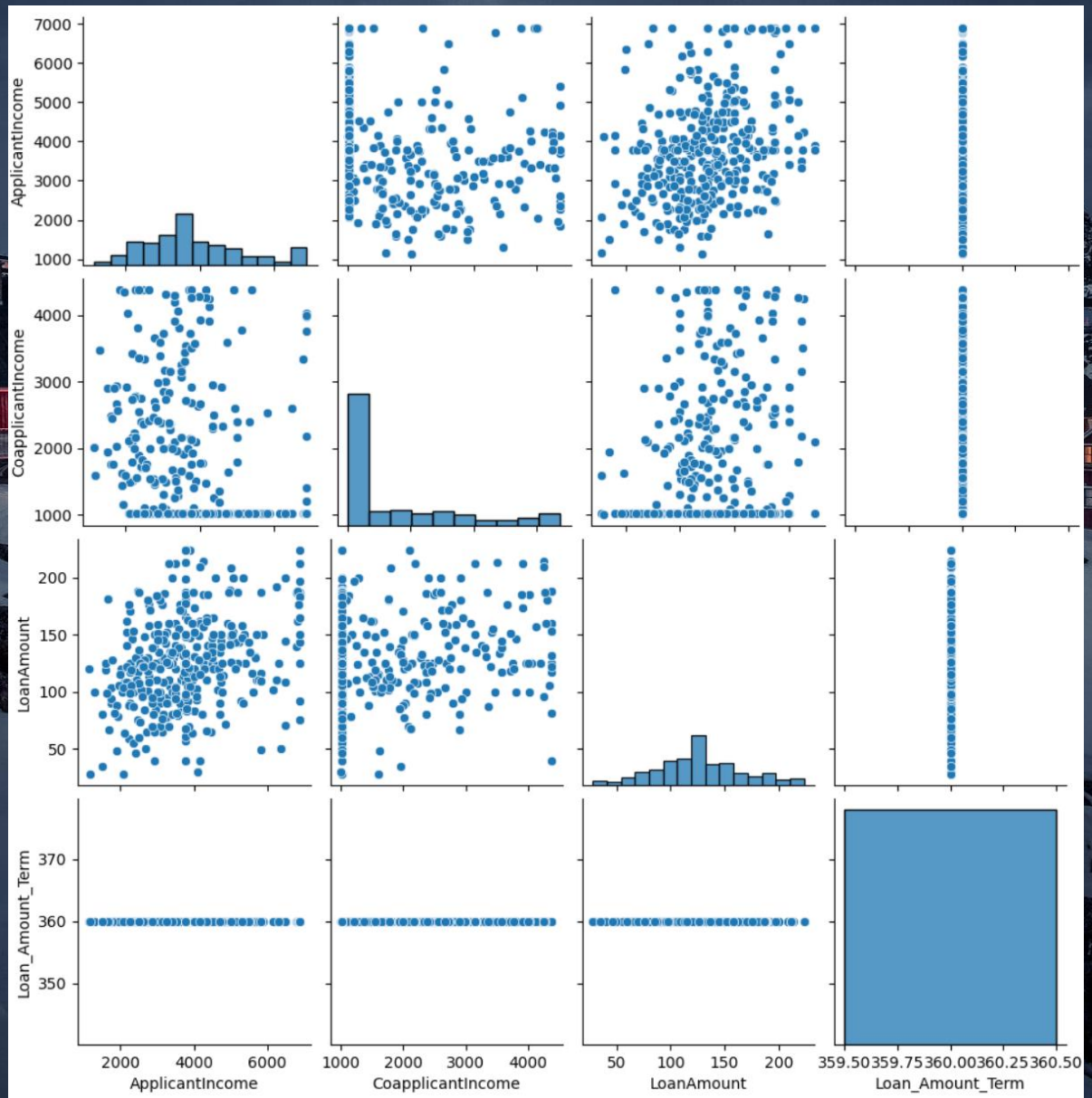
Create scatter plots to explore relationships between pairs of numeric variables.

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.
- **Loan Amount vs Loan Amount Term:** No clear correlation is observed, indicating loan amount and term are largely independent of each other. This suggests that loan term is determined based on factors other than the loan amount.

Data Visualization and Insights

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



Data Visualization and Insights

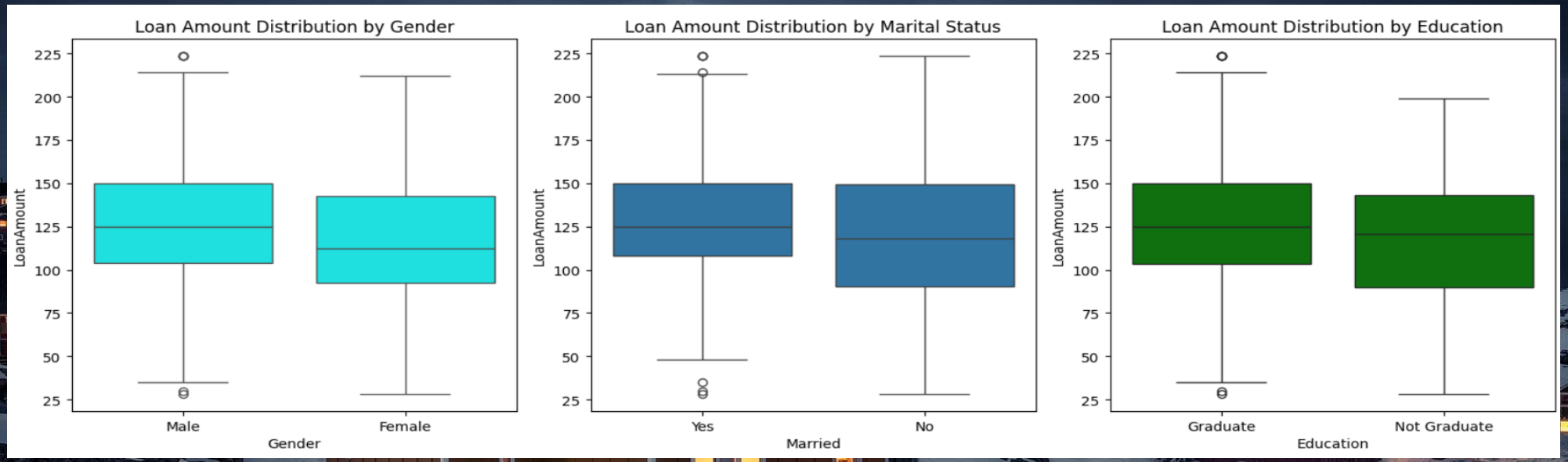
Use pair plots (scatter matrix) to visualize interactions between multiple numeric variables simultaneously.

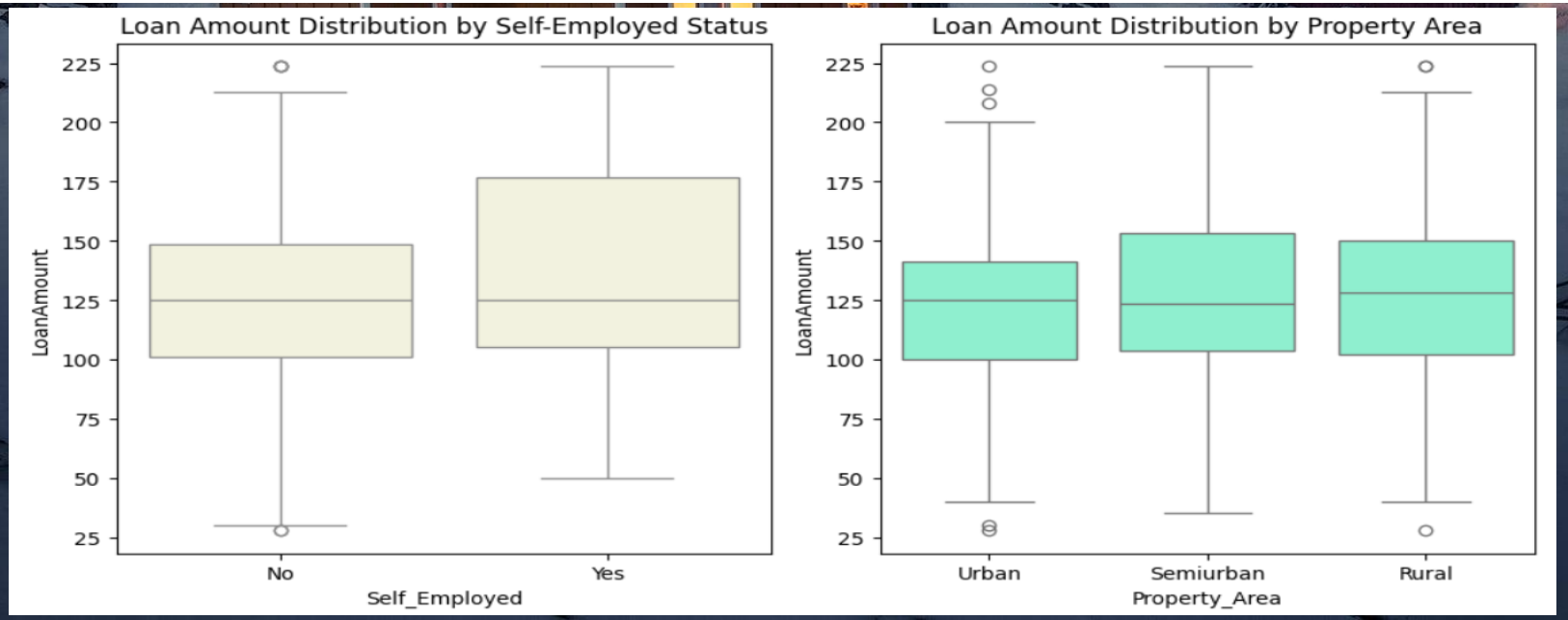
Key insights

- There's a positive correlation between 'Applicant Income' and 'Loan Amount'. As 'Applicant Income' increases, the 'Loan Amount' they are eligible for also tends to increase.
- There's no significant correlation between ‘Co-applicant Income' and 'Loan Amount'. The 'Loan Amount' doesn't seem to be strongly influenced by the ‘Co-applicant Income'.
- There's no clear correlation between 'Loan Amount' and 'Loan Amount Term'. The duration of the loan doesn't appear to have a strong relationship with the amount borrowed.

Data Visualization and Insights

Investigate the relationship between categorical and numeric variables using Box plots.





Data Visualization and Insights

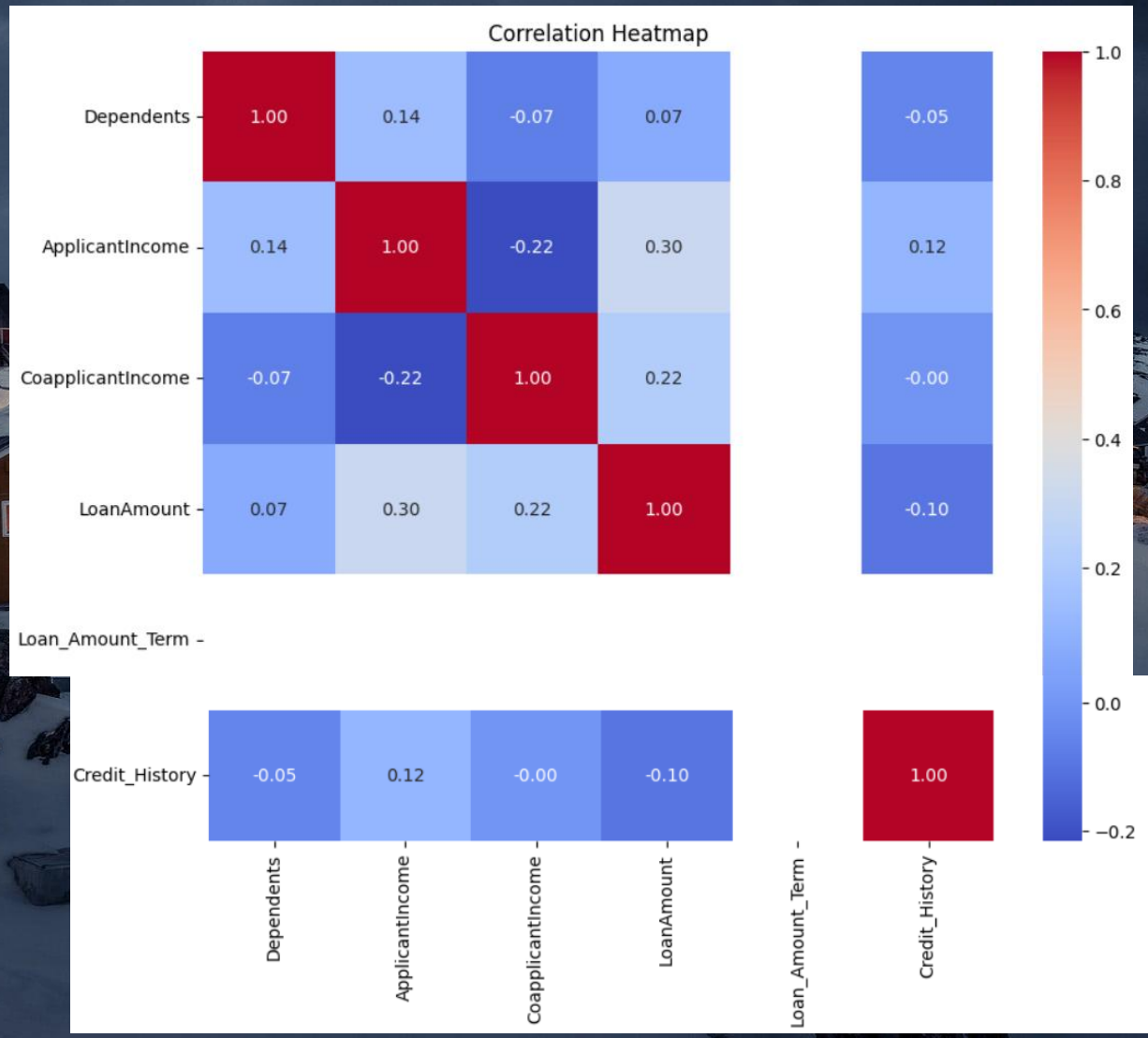
Investigate the relationship between categorical and numeric variables using Box plots.

Key insights

- There's no significant difference in 'Loan Amount' based on 'Gender'. Both males and females tend to apply for similar loan amounts.
- Married individuals tend to apply for slightly higher 'Loan Amount' compared to unmarried individuals.
- Graduates tend to apply for higher 'Loan Amount' compared to non-graduates. This suggests that higher education might be associated with greater financial needs or borrowing capacity.
- There's no substantial difference in 'Loan Amount' between self-employed and non-self-employed individuals.
- 'Loan Amount' distribution varies slightly across different 'Property Area'. Applicants from semiurban areas tend to apply for slightly higher loan amounts compared to those from urban or rural areas.

Data Visualization and Insights

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



Data Visualization and Insights

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

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.

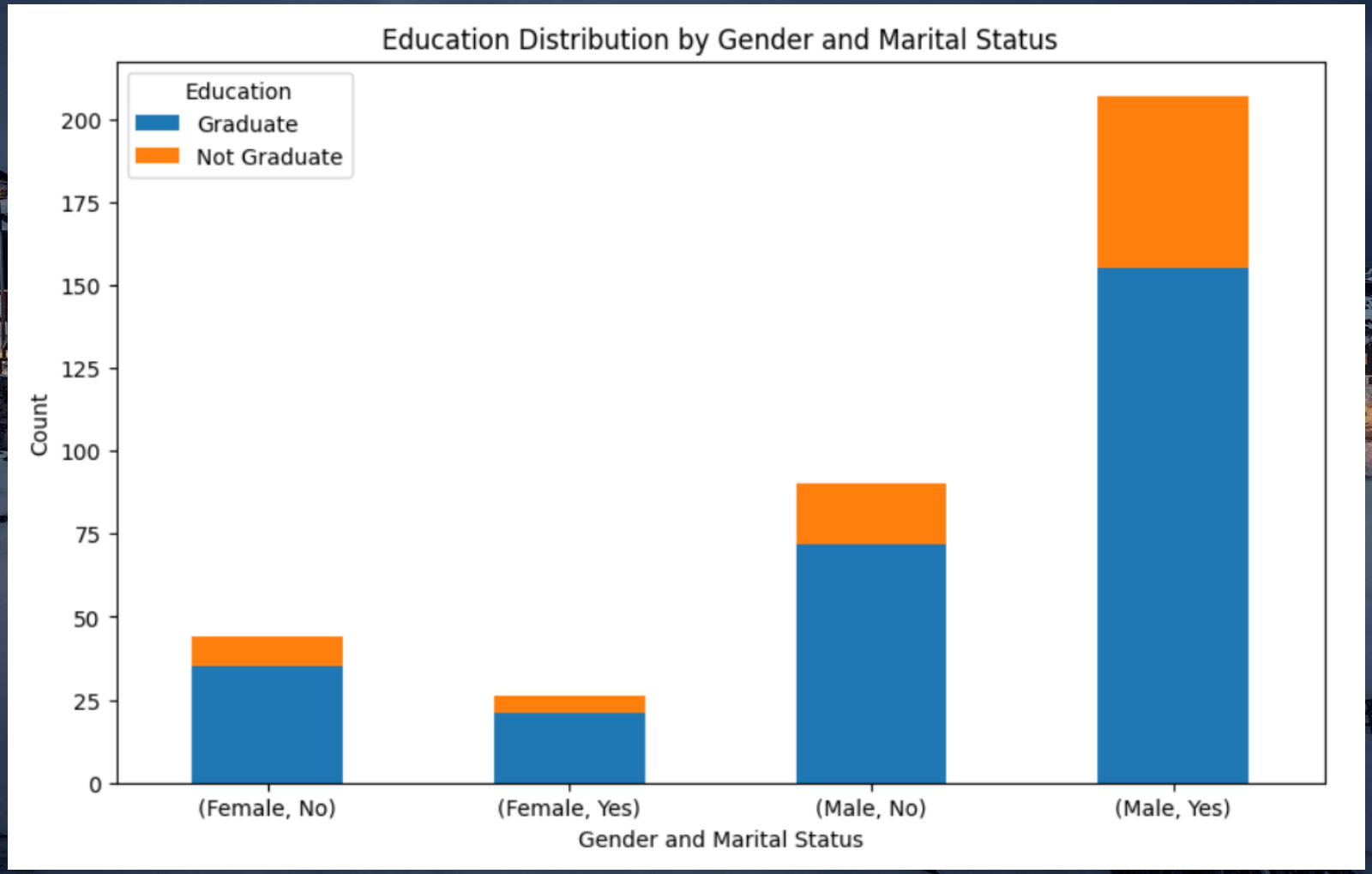
- **Weak Correlation between Loan Amount and Co-applicant Income:** The correlation between 'Loan Amount' and ‘Co-applicant Income' is relatively weak (0.19). This indicates that the co-applicant's income has a lesser impact on the loan amount compared to the applicant's income.

- **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 modeling, especially if using linear models sensitive to multicollinearity.

Data Visualization and Insights

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



Data Visualization and Insights

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

Key insights

- **Education and Gender:** A higher proportion of males across both married and unmarried categories have a graduate degree compared to females. The difference in graduate education between genders is more pronounced in the married category.

- **Education and Marital Status:** In both male and female categories, a higher proportion of married individuals have a graduate degree compared to unmarried individuals. This suggests a possible correlation between higher education and the likelihood of getting married.

- **Overall Trend:** The majority of loan applicants, regardless of gender or marital status, have a graduate degree. This indicates that higher education might be a common factor among those seeking loans.

- These insights can be further investigated to understand the impact of education on loan approval rates and other relevant factors.

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

KEY FINDINGS :

✓ Distribution of Numeric Variables :

- *Applicant Income and Co-applicant Income are right-skewed, indicating a concentration of lower incomes with a few high earners.*
- *Loan Amount shows a relatively normal distribution with some outliers.*
- *Loan Amount Term is predominantly concentrated at 360 months.*
- *Credit History is negatively skewed, suggesting most applicants have a credit history.*
- *Dependents distribution indicates a higher proportion of applicants with no dependents.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

KEY FINDINGS :

✓ Categorical Variable Analysis :

- *Majority of applicants are male and married.*
- *Most applicants are graduates and not self-employed.*
- *Property Area distribution is relatively balanced across urban, semiurban, and rural areas.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

KEY FINDINGS :

✓ Relationships between Variables:

- *Positive correlation between Applicant Income and Loan Amount, suggesting higher income applicants tend to request larger loans.*
- *Weak positive correlation between Co-applicant Income and Loan Amount.*
- *No strong linear relationship between Loan Amount and Loan Amount Term.*
- *Box plots reveal variations in Loan Amount distribution across different categories of Gender, Marital Status, Education, Self Employed, and Property Area.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

Conclusions :

- ***Income and Loan Amount***: *Applicant income plays a significant role in determining the loan amount.*
- ***Demographic Factors***: *Gender, marital status, education, and employment status influence loan application characteristics.*
- ***Credit History***: *A good credit history is prevalent among applicants.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

Recommendations :

- ***Target Marketing***: *Tailor loan products and marketing strategies based on income levels and demographic characteristics.*
- ***Risk Assessment***: *Consider income, credit history, and other factors for loan approval and risk assessment.*
- ***Product Diversification***: *Offer loan products with varying terms and amounts to cater to different customer needs.*
- ***Further Analysis***: *Explore additional factors and interactions to gain deeper insights and refine decision-making.*

THANK YOU FOR READING

For coding part, kindly refer to below link :-

https://colab.research.google.com/drive/1WhSYfizo6ITh262pg8Mfl7rqK2svsZFO#scrollTo=XzskCDMl-duH