HOME LOAN APPROVAL ANALYSIS

Prepared By :
Rishabh Bansal

Purpose of the Project :

The home loan approval dataset is a collection of financial records and associated information used to determine the eligibility of individuals or organizations for home loans. This dataset is used in various applications, such as predicting loan eligibility and analyzing factors affecting loan approval.

The dataset contains various features, including loan amount, gender, marital status, education, number of dependents, income, credit history, and loan status (approved or not). The goal is to predict loan eligibility based on these features and to analyze the factors that influence loan approval.

Exploratory Data Analysis (EDA) is an essential step in understanding the dataset, identifying trends, and preparing the data for modeling. EDA techniques include univariate analysis, bivariate analysis, and multivariate analysis. Univariate analysis involves analyzing individual features, such as histograms for numeric columns and bar charts for categorical columns. Bivariate analysis explores relationships between two features, such as scatter plots for numeric-numeric relationships and box plots for categorical-numeric relationships. Multivariate analysis involves analyzing relationships between multiple features simultaneously, such as pair plots and heatmaps for correlation analysis.

In summary, the home loan approval dataset is used to predict loan eligibility and analyze factors affecting loan approval. EDA, data preprocessing, and machine learning techniques are essential steps in preparing the dataset and building predictive models for loan eligibility.

Objectives:

- **Familiarity with the Dataset:**

This involves getting to know the data you are working with, understanding its structure, variables, and any potential issues or limitations. It's about gaining a comprehensive understanding of what information is available to you.

- **Data Exploration and Visualization:**

This step involves delving into the data to uncover patterns, relationships, and anomalies. By using various statistical and visualization techniques, you can explore the data in depth to gain insights and a better understanding of its characteristics.

- **Identification of Patterns, Trends, and Insights:**

Here, the focus is on analyzing the data to identify recurring patterns, trends over time, correlations between variables, and any other significant insights that can be derived from the data. This step is crucial for extracting valuable information from the dataset.

- **Generation of Meaningful Visualizations:**

The final goal is to create visual representations of the findings in a clear and concise manner. Visualizations such as charts, graphs, and plots can effectively communicate the patterns, trends, and insights discovered during the analysis process. These visualizations help in presenting the information in a way that is easily understandable to others.

Dataset Description:

The Dataset comprises information related to Home Loan Approval Analysis:

- Loan_ID: This is a unique identifier assigned to each loan application.
- Gender: Indicates the gender of the loan applicant (Male/Female).
- Married: Denotes the marital status of the applicant.
- Dependents: Represents the number of dependents of the customer (0, 1, 2, 3, or more).
- Education: Indicates the educational background of the applicant.
- Self_Employed: Specifies whether the applicant is self-employed or not.
- ApplicantIncome: Refers to the income of the primary applicant.
- CoapplicantIncome: Represents the income of the co-applicant, if applicable.
- LoanAmount: Denotes the amount of the loan applied for.
- Loan_Amount_Term: Indicates the term (in months) for which the loan is taken.
- Credit_History: Reflects the credit history of the applicant.
- Property_Area: Specifies the area where the property for which the loan is sought is located.

The dataset contains valuable information for analysis. The primary focus of this assessment is on data exploration and visualization.

Project Tasks:

Task 1: Data Exploration

- Load the dataset into a Python environment (e.g., Jupyter Notebook).
- Display the initial rows of the dataset to understand its structure.
- Check for missing values and handle them if necessary.
- Summarize basic statistics (mean, median, standard deviation) for the numeric columns.

Task 2: Data Visualization

2.1 Univariate Analysis

- Explore the distribution of numeric columns using histograms and box plots.
- Analyze categorical variables with bar charts and pie charts.

2.2 Bivariate Analysis

- Create scatter plots to explore relationships between numeric variables.
- Use pair plots to visualize interactions between multiple numeric variables.
- Investigate relationships between categorical and numeric variables using box plots or violin plots.

2.3 Multivariate Analysis

- Perform correlation analysis to identify relationships between numeric variables and visualize them using a heatmap.
- Create a stacked bar chart to show the distribution of categorical variables across multiple categories.

Task 1 : Data Exploration

Data Exploration

Data Loading:

To Begin data exploration ,we must first load the dataset in a Python Environment such as Jupyter Notebook. In order to do so we can use pd.read_csv() method to load the csv file.

Null Values:

The Dataset Contains following null values.

- Gender Column : 11 Null Values.
- Dependents Column : 50 Null Values.
- Self_Employed Column : 23 Null Values.
- LoanAmount Column : 5 Null Values.
- Loan_Amount_Term Column : 6 Null Values.
- Credit_History Column : 29 Null Values.

Basic Summarize of Dataset:

Basic Summarize of Dataset contains the Mean ,Standard Deviation ,

- Dependents
  - Dependents
  - Mean is 0.479564
  - Standard deviation of 0.756749
  - Minimum is 0.00
  - Maximum is 2.00.
- ApplicantIncome
  - Mean is 4238.96457
  - Median is 3786.00
  - Standard deviation of 1950.97635
  - Minimum is 0.0
  - Maximum is 8354.000000
- CoapplicantIncome
  - Mean is 1425.5013
  - Median is 1025.00
  - Standard deviation of 1600.08625
  - Minimum is 0.
  - Maximum is 6076.250000

	Dependents	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	367.000000	367.000000	367.000000	367.000000	367.000000	367.000000
mean	0.479564	4238.964578	1425.501362	135.980926	342.822888	0.839237
std	0.756749	1950.976357	1600.086250	60.959739	64.658402	0.367814
min	0.000000	0.000000	0.000000	28.000000	6.000000	0.000000
25%	0.000000	2864.000000	0.000000	101.000000	360.000000	1.000000
50%	0.000000	3786.000000	1025.000000	125.000000	360.000000	1.000000
75%	1.000000	5060.000000	2430.500000	157.500000	360.000000	1.000000
max	2.000000	8354.000000	6076.250000	550.000000	480.000000	1.000000

Task 1 : Data Visualizations

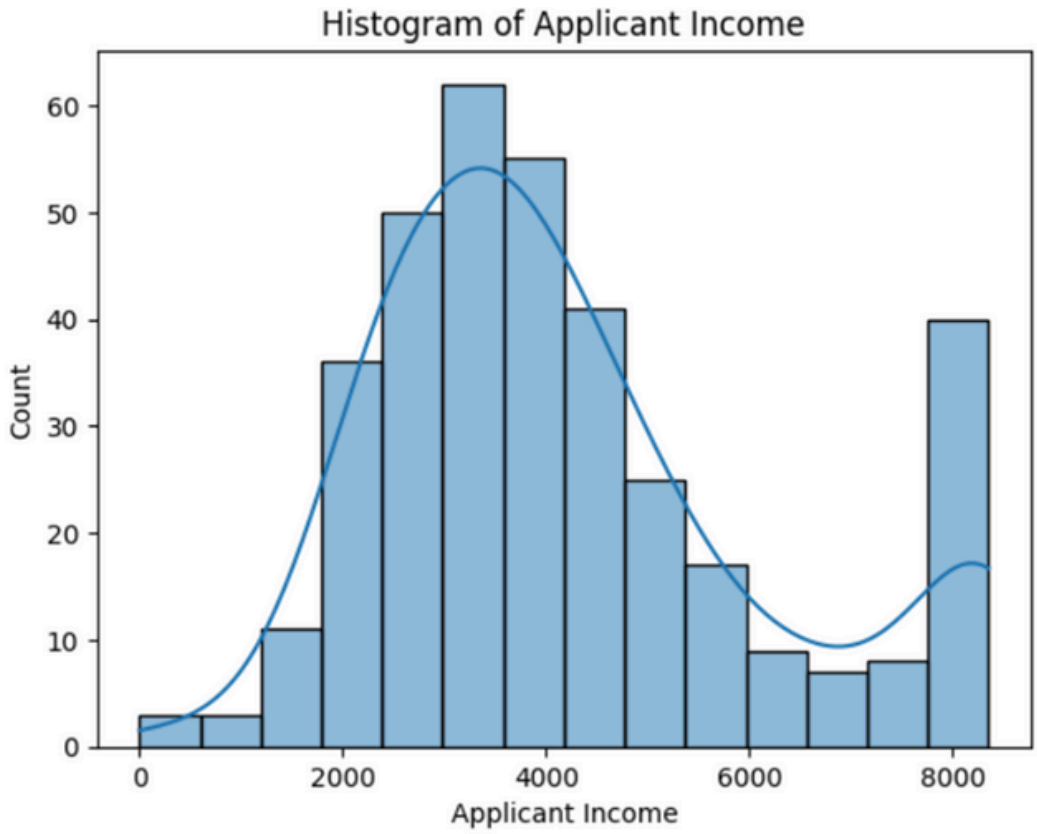
Data Visualizations:

Uni-Variate Analysis:

Univariate analysis is a fundamental concept in statistics and data analysis that deals with the examination of a single variable or attribute within a dataset. This type of analysis is often used as an initial step in understanding the data, as it provides a basic overview of the variable's distribution, central tendencies, and dispersion. The primary goal of univariate analysis is to describe the data and find patterns, without considering the relationship between the variable and other factors.

In univariate analysis, the data is typically represented in tables, graphs, or statistical measures. Frequency distribution tables, histograms, bar charts, and pie charts are common graphical representations used to visualize the distribution of a single variable. Statistical measures, such as mean, median, mode, standard deviation, and variance, are used to summarize the data's central tendencies and dispersion.

Histogram

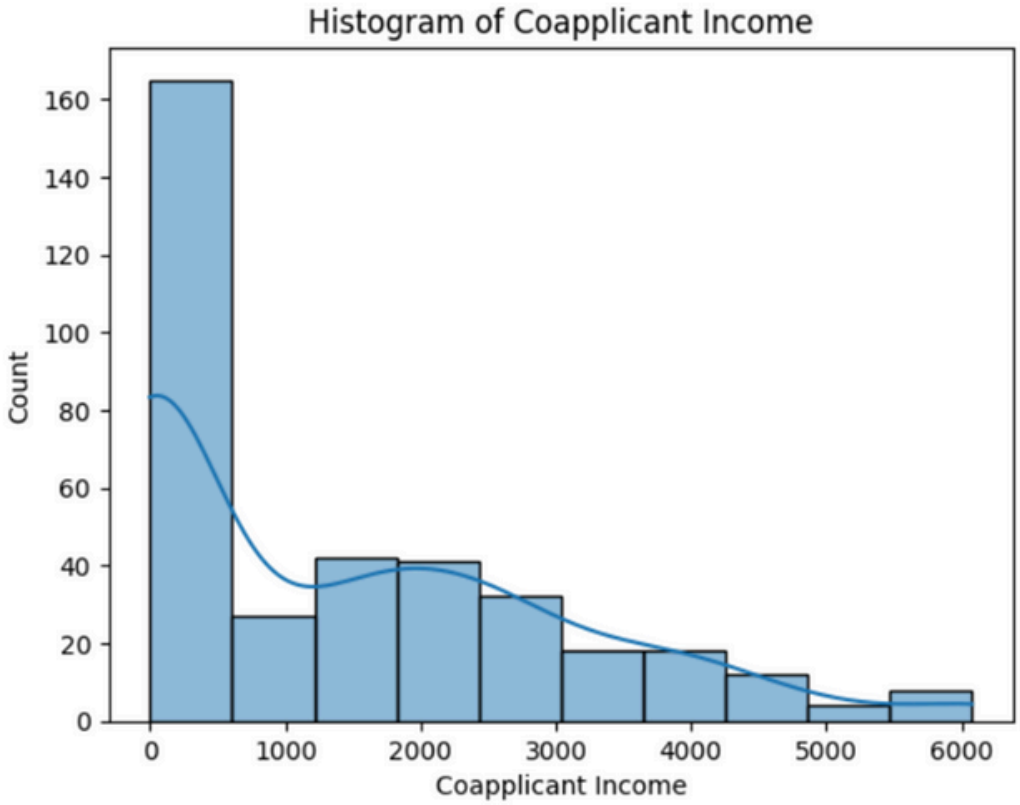


Observations:

- The histogram shows the distribution of ApplicantIncome, with a range of income values on the x-axis.
- The KDE line provides a smooth representation of the underlying distribution of the data.
- The histogram shows a peak around a certain income value, indicating a high concentration of applicants with that income.
- The histogram also shows a long tail on the right side, indicating the presence of applicants with high income values.

Insights:

- The distribution of ApplicantIncome is right-skewed, indicating that there are more applicants with lower income values than higher income values.
- The peak of the distribution suggests that there is a common income range for applicants.
- The long tail on the right side of the distribution indicates the presence of a small number of applicants with very high income values.
- The KDE line provides a more accurate representation of the underlying distribution of the data than a traditional histogram.
- The histogram can be used to identify any potential outliers or skewness in the data.
- The KDE line can help to smooth out the histogram and provide a more accurate representation of the underlying distribution of the data.
- The histogram can be used to inform decisions related to loan approvals, such as setting income thresholds for loan eligibility.

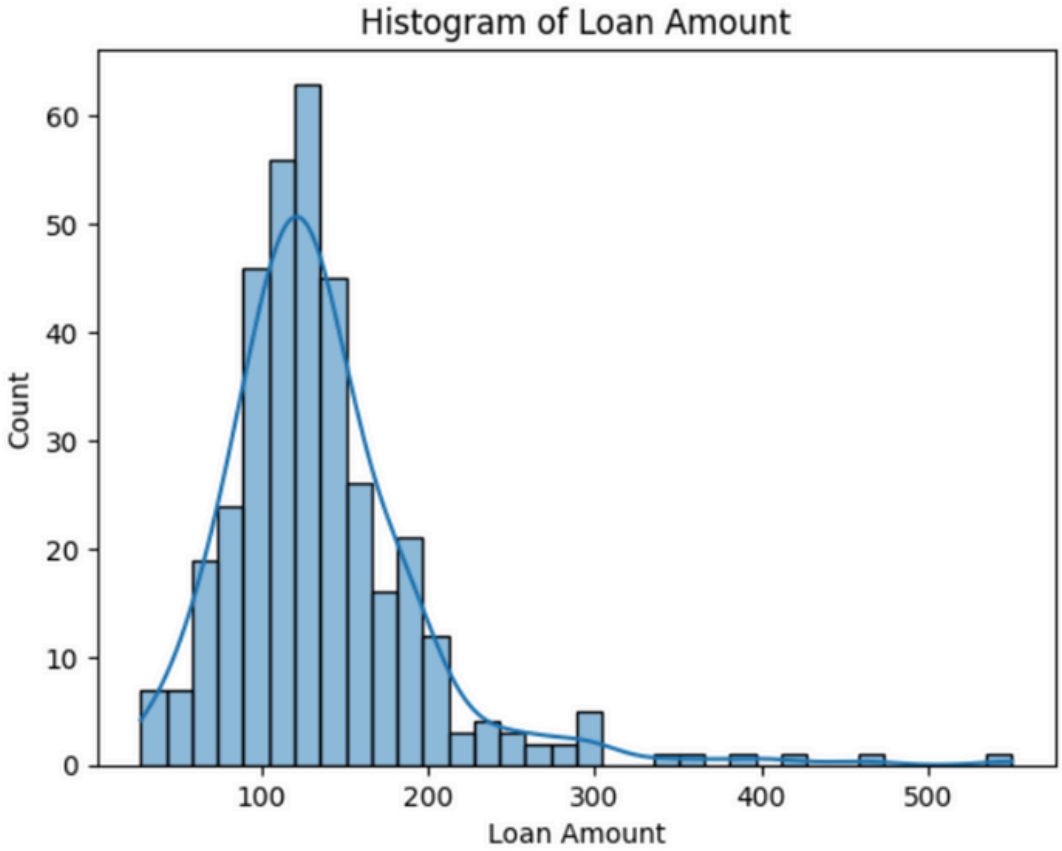


Observations:

- The histogram shows the distribution of CoapplicantIncome, with a range of income values on the x-axis.
- The KDE line provides a smooth representation of the underlying distribution of the data.
- The histogram shows a peak around a certain income value, indicating a high concentration of coapplicants with that income.
- The histogram also shows a long tail on the right side, indicating the presence of coapplicants with high income values.

Insights:

- The distribution of CoapplicantIncome is right-skewed, indicating that there are more coapplicants with lower income values than higher income values.
- The peak of the distribution suggests that there is a common income range for coapplicants.
- The long tail on the right side of the distribution indicates the presence of a small number of coapplicants with very high income values.
- The KDE line provides a more accurate representation of the underlying distribution of the data than a traditional histogram.
- The histogram can be used to identify any potential outliers or skewness in the data.
- The KDE line can help to smooth out the histogram and provide a more accurate representation of the underlying distribution of the data.
- The histogram can be used to inform decisions related to loan approvals, such as setting income thresholds for loan eligibility.



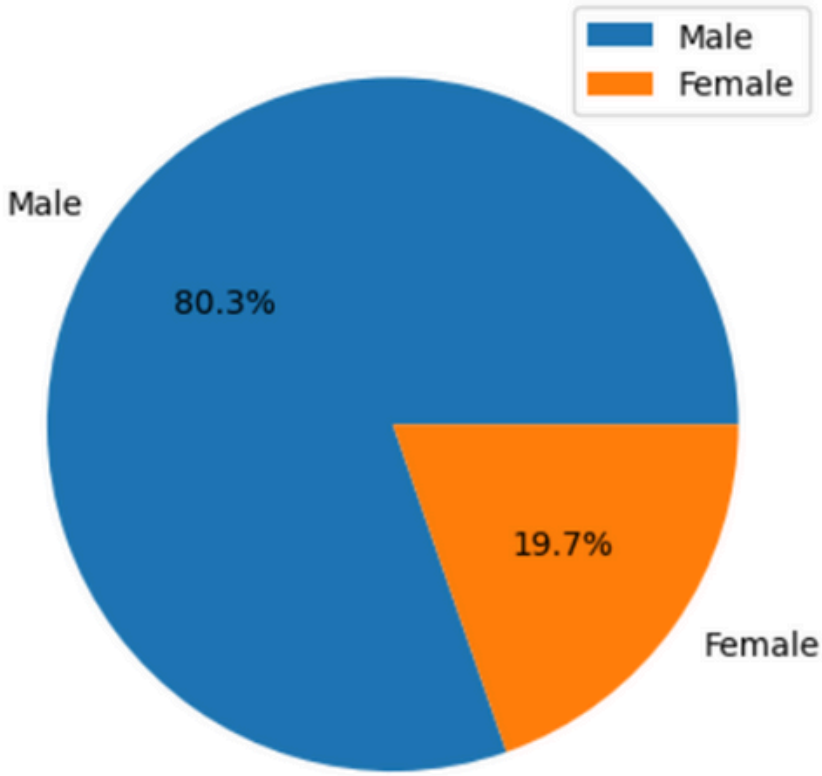
Observations:

- The histogram shows the distribution of LoanAmount, with a range of loan amounts on the x-axis.
- The KDE line provides a smooth representation of the underlying distribution of the data.
- The histogram shows a peak around a certain loan amount, indicating a high concentration of loans with that amount.
- The histogram also shows a long tail on the right side, indicating the presence of loans with high loan amounts.

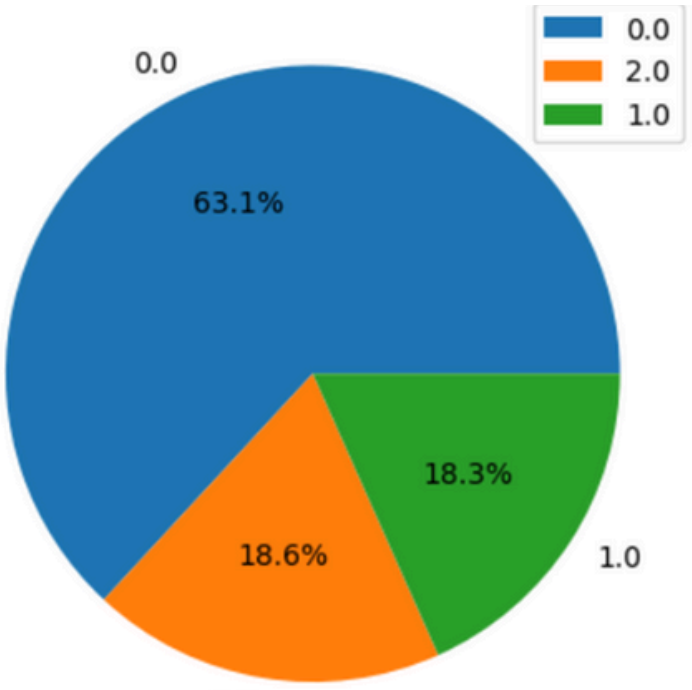
Insights:

- The distribution of LoanAmount is right-skewed, indicating that there are more loans with lower loan amounts than higher loan amounts.
- The peak of the distribution suggests that there is a common loan amount range.
- The long tail on the right side of the distribution indicates the presence of a small number of loans with very high loan amounts.
- The KDE line provides a more accurate representation of the underlying distribution of the data than a traditional histogram.
- The histogram can be used to identify any potential outliers or skewness in the data.
- The KDE line can help to smooth out the histogram and provide a more accurate representation of the underlying distribution of the data.
- The histogram can be used to inform decisions related to loan approvals, such as setting loan amount thresholds for loan eligibility.

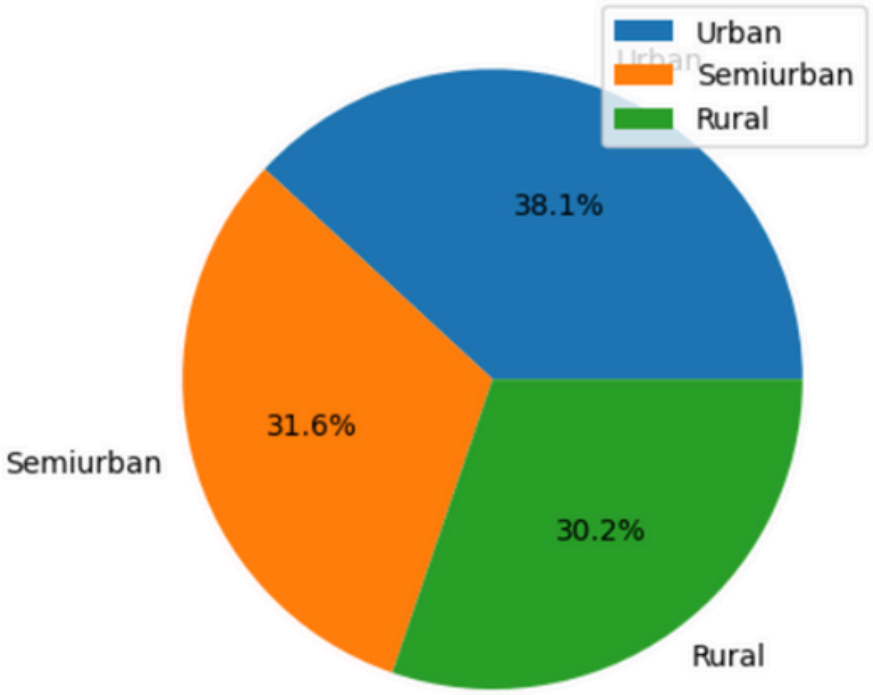
Pie Charts



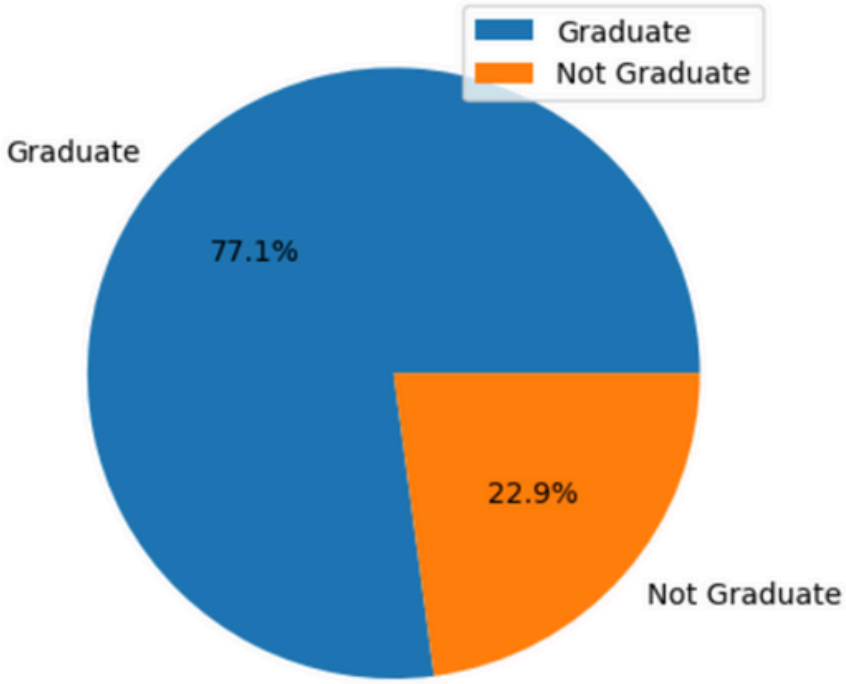
- The pie chart shows the distribution of gender in the dataset, with the number of occurrences represented by the size of each slice.
- The majority of the dataset consists of Male records, making up approximately 80.3% of the total & Female, accounting for around 19.7% of the total.
- The dataset is slightly skewed towards Male records, which could have implications for any analysis or modeling based on this data.
- The pie chart provides a clear visual representation of the gender distribution, making it easy to understand the relative proportions of each gender in the dataset.
- In summary, the pie chart effectively communicates the gender distribution in the dataset, highlighting the dominance of Male records. This information can be valuable for understanding the data's composition and for guiding any further analysis or modeling efforts.



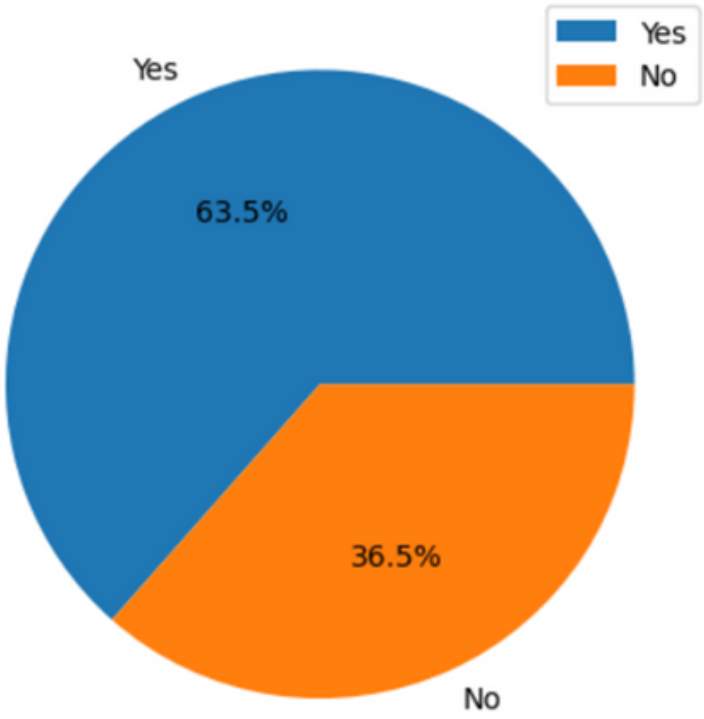
- The pie chart shows the distribution of the number of dependents in the dataset, with the number of occurrences represented by the size of each slice.
- The largest portion of the dataset consists of individuals with no dependents, making up approximately 63.1% of the total.
- The second-largest group is those with two dependents, accounting for around 18.6% of the total. The remaining portions of the dataset are individuals with one dependent around 18.3%.
- The dataset is skewed towards individuals with no or few dependents, which could have implications for any analysis or modeling based on this data.
- The pie chart provides a clear visual representation of the distribution of dependents, making it easy to understand the relative proportions of each category.
- In summary, the pie chart effectively communicates the distribution of dependents in the dataset, highlighting the dominance of individuals with no dependents. This information can be valuable for understanding the data's composition and for guiding any further analysis or modeling efforts.



- The pie chart illustrates the distribution of property areas in the dataset, with each slice representing a different property area category.
- The largest portion of the dataset corresponds to Semiurban properties, accounting for approximately 31.6% of the total.
- The Urban property area category follows closely behind, representing around 38.1% of the dataset.
- The Rural property area category is the smallest segment, making up approximately 30.2% of the total.
- The dataset shows a varied distribution across different property areas, with a significant presence in all three distinct property areas.
- The pie chart effectively visualizes the proportions of each property area category, making it easy to grasp the relative distribution.
- In summary, the pie chart provides a clear overview of the distribution of property areas in the dataset, highlighting the prevalence of Semiurban properties. This information can be valuable for understanding the composition of properties in different areas and may offer insights for further analysis or decision-making related to property trends or investments.



- The pie chart visualizes the distribution of education levels within the dataset, with each slice representing a different education category.
- The majority of individuals in the dataset are categorized as Graduates, constituting approximately 77.1% of the total.
- The remaining portion of the dataset consists of individuals classified as Not Graduates, making up around 22.9% of the total.
- The dataset is heavily skewed towards individuals with Graduate education, indicating a higher proportion of educated individuals in the dataset.
- The pie chart effectively communicates the distribution of education levels, making it easy to interpret the relative proportions of Graduates and Not Graduates.
- In summary, the pie chart offers a clear representation of the education distribution in the dataset, emphasizing the dominance of Graduate individuals. This insight can be valuable for understanding the educational background of the dataset and may influence decision-making processes or further analysis related to education levels and their impact on the data.



- The pie chart visualizes the distribution of marital status within the dataset, with each slice representing a different marital status category.
- The majority of individuals in the dataset are categorized as Married, constituting approximately 63.5% of the total.
- The remaining portion of the dataset consists of individuals classified as Not Married (single, divorced, or widowed), making up around 36.5% of the total.
- The dataset is skewed towards individuals who are Married, indicating a higher proportion of married individuals in the dataset.
- The pie chart effectively communicates the distribution of marital status, making it easy to interpret the relative proportions of Married and Not Married individuals.
- In summary, the pie chart offers a clear representation of the marital status distribution in the dataset, emphasizing the dominance of Married individuals. This insight can be valuable for understanding the marital background of the dataset and may influence decision-making processes or further analysis related to marital status and its impact on the data.

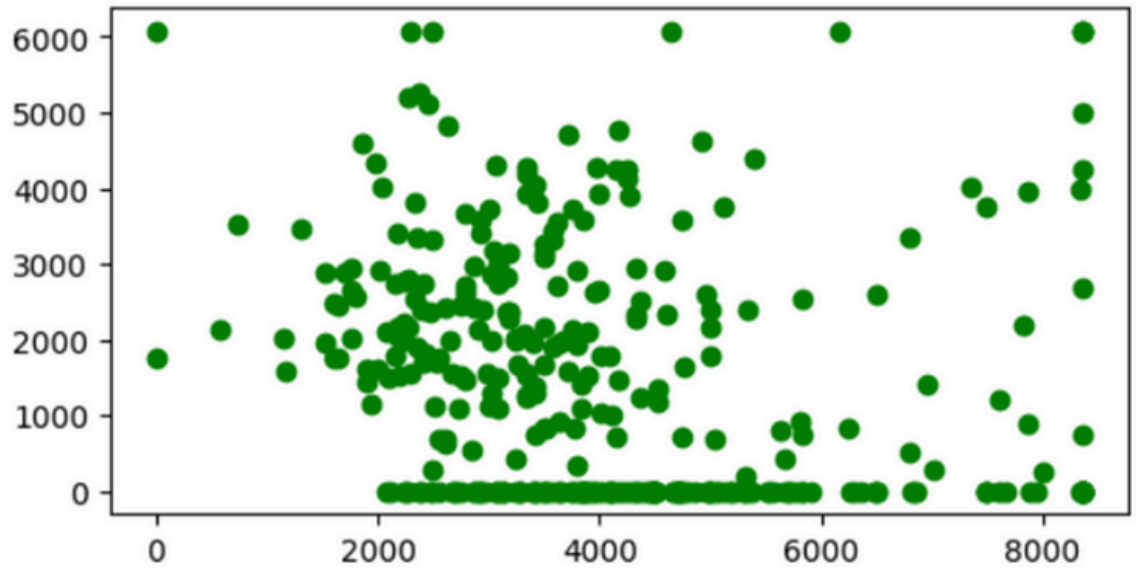
Data Visualizations:

Bi-Variate Analysis:

Bivariate analysis is a statistical method used to analyze the relationship between two variables. It involves the analysis of two variables to determine the empirical relationship between them, and can be descriptive or inferential. Types of bivariate analysis include scatter plots, correlation coefficients, and regression analysis.

The two variables can be dependent and independent variable. Bivariate analysis is used to determine the changes that occur between the two variables and to what extent. Examples of bivariate analysis include investigating the connection between education and income, where one variable could be the level of education and the other could be income, and determining if there is a significant relationship between these two variables.

Scatter Plot

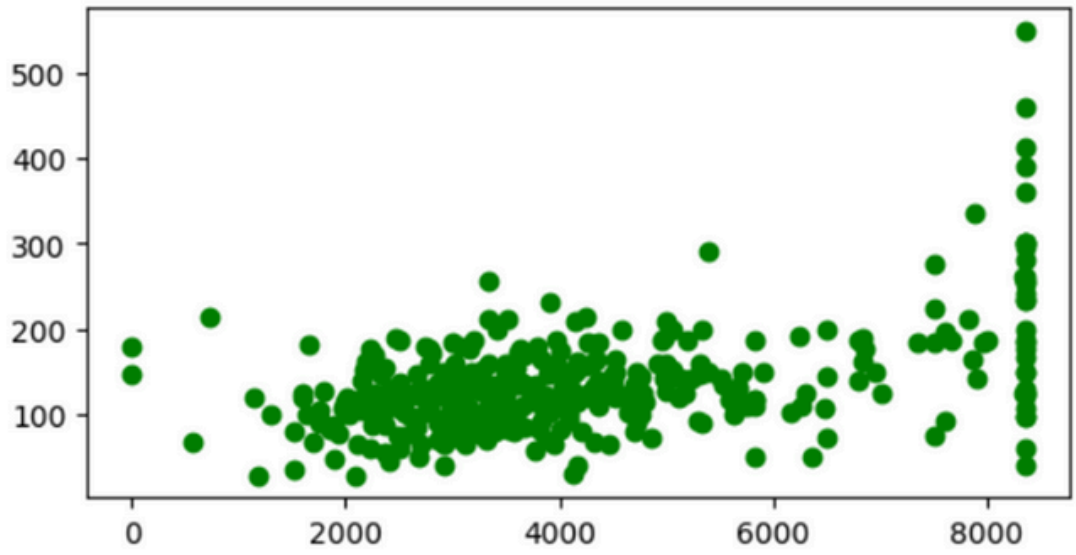


Observations:

- The scatter plot shows the relationship between the income of the applicant and the co-applicant.
- There are a large number of data points, indicating that the dataset is quite large.
- The data points are scattered throughout the plot area, suggesting that there is no strong correlation between the two variables.
- Some data points are clustered in certain areas, indicating that there may be certain income levels that are more common among applicants and co-applicants.

Insights:

- The lack of a strong correlation between ApplicantIncome and CoapplicantIncome suggests that the income of the co-applicant may not be a major factor in the decision-making process.
- The clustering of data points in certain areas may indicate that there are certain income levels that are more desirable or more common among applicants and co-applicants.
- Further analysis may be needed to determine if there are any other factors that are more strongly correlated with the decision-making process.
- The large number of data points suggests that there is a lot of variation in the data, which may make it difficult to identify any clear patterns or trends.

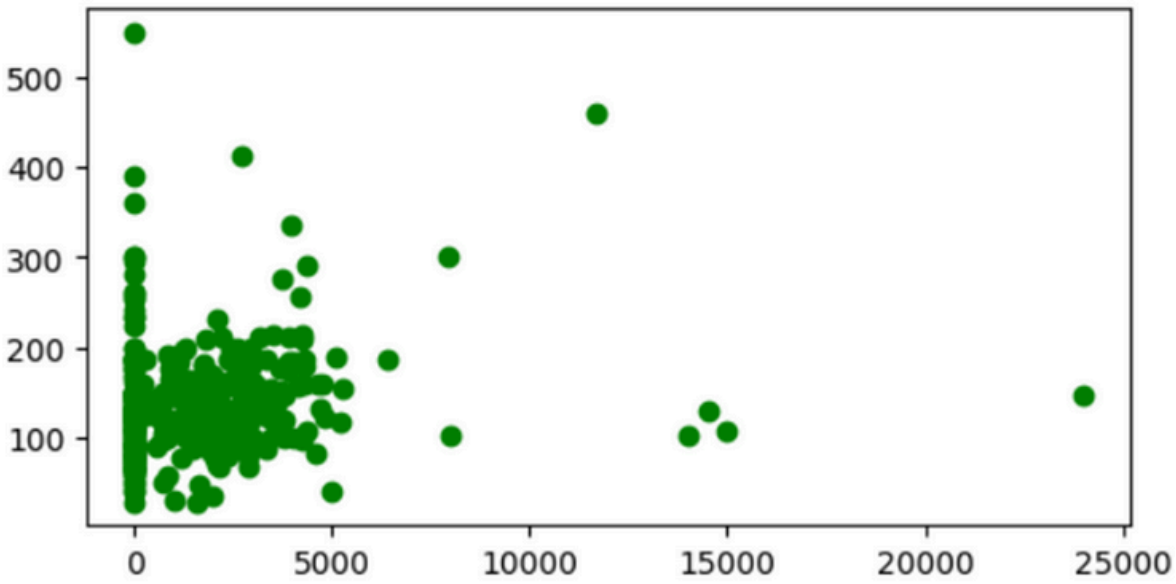


Observations:

- The scatter plot shows the relationship between the income of the applicant and the loan amount.
- There are a large number of data points, indicating that the dataset is quite large.
- The data points are scattered throughout the plot area, suggesting that there is no strong correlation between the two variables.
- Some data points are clustered in certain areas, indicating that there may be certain income levels that are more common among applicants with certain loan amounts.

Insights:

- The lack of a strong correlation between ApplicantIncome and LoanAmount suggests that the income of the applicant may not be a major factor in determining the loan amount.
- The clustering of data points in certain areas may indicate that there are certain income levels that are more common among applicants with certain loan amounts.
- Further analysis may be needed to determine if there are any other factors that are more strongly correlated with the loan amount.
- The large number of data points suggests that there is a lot of variation in the data, which may make it difficult to identify any clear patterns or trends.
- The shape of the data points may suggest that there are some outliers with very high loan amounts, which may be worth investigating further.
- It would be interesting to see if there is any correlation between the loan amount and other factors such as the loan term or the interest rate.



Observations:

- The scatter plot illustrates the relationship between the co-applicant's income and the loan amount.
- The data points are spread across the plot, indicating varying combinations of coapplicant income and loan amounts.
- There seems to be a moderate spread of data points, suggesting a potential correlation between coapplicant income and loan amount.
- Some clusters of data points may indicate common income levels among coapplicants with specific loan amounts.

Insights:

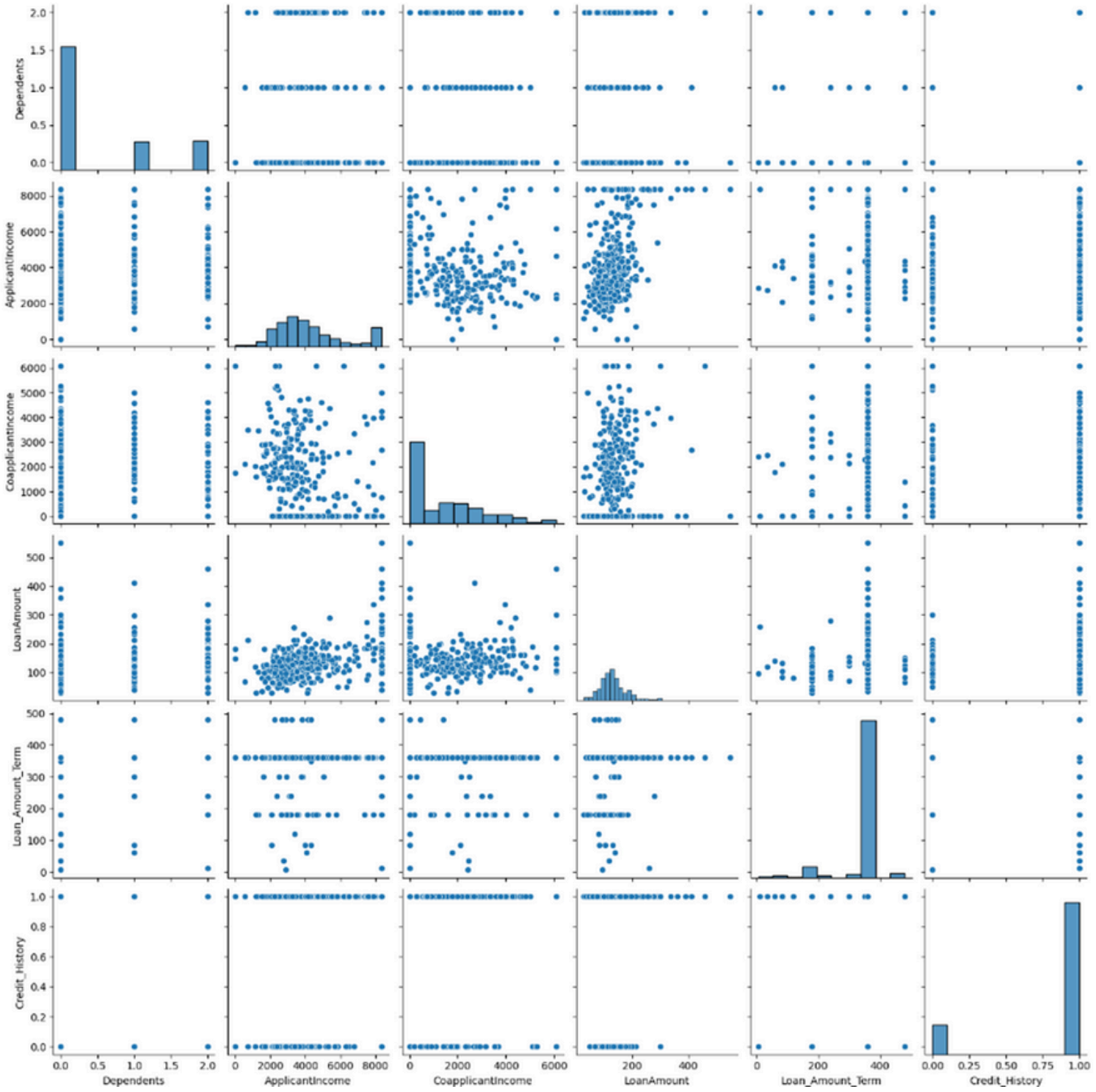
- The scatter plot hints at a potential relationship between CoapplicantIncome and LoanAmount, implying that the coapplicant's income may influence the loan amount.
- Further analysis could reveal if there is a significant correlation between the coapplicant's income and the loan amount, which could be valuable for loan approval decisions.
- Identifying any outliers in the data could provide insights into cases where the coapplicant's income significantly impacts the loan amount.
- Exploring additional factors such as credit scores or employment status alongside coapplicant income could enhance the understanding of loan approval criteria.
- It would be beneficial to investigate if there are specific income thresholds for coapplicants that affect the loan amount approval process.
- Conducting a deeper analysis to understand the distribution of loan amounts concerning different levels of coapplicant income could provide valuable insights for financial institutions.

PairPlot

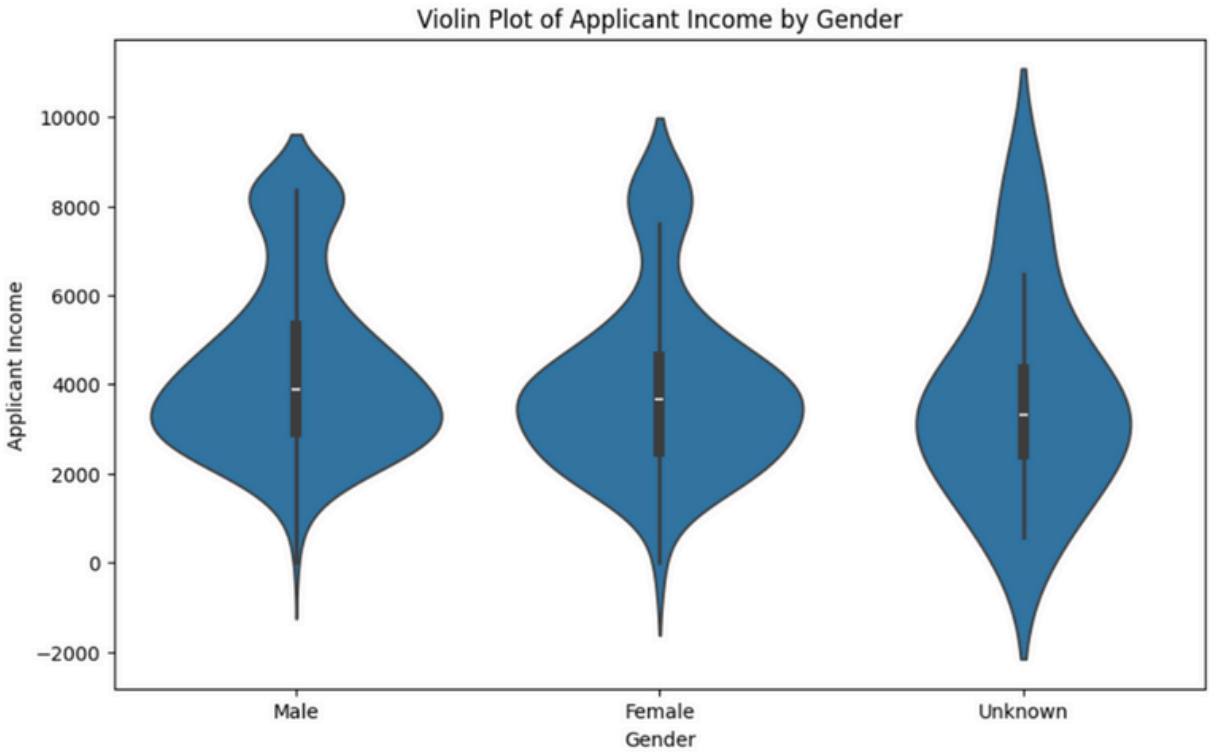
Seaborn pairplot is a visualization tool that allows for the exploration of pairwise relationships between variables in a dataset. It creates a grid of scatter plots where each variable is shared across the y-axes in a single row and the x-axes in a single column. The diagonal plots show the marginal distribution of the data in each column.

Pairplot can be used with categorical variables, and the markers parameter can be used to style the off-diagonal axes. The plot_kws and diag_kws parameters can be used to customize the off-diagonal and diagonal plots, respectively. The corner parameter can be used to plot only the lower triangle. The kind parameter can be used to select the kind of plot for the off-diagonal axes, and the diag_kind parameter can be used to select the kind of plot for the diagonal.

The hue parameter can be used to map plot aspects to different colors based on a variable in the dataset. The vars, x_vars, and y_vars parameters can be used to select the variables to plot. The size parameter can be used to control the size of the figure, and the aspect parameter can be used to control the aspect ratio of the subplots


Dependents
ApplicantIncome
CoapplicantIncome
LoanAmount
Loan_Amount_Term
Credit_History

Violin Plot



The violin plot shows the distribution of data points in a way that allows for easy comparison between different groups. It displays a box plot in the center, which shows the median, quartiles, and whiskers of the data. On either side of the box plot, the plot shows a kernel density estimation of the data, which is represented by a "violin" shape.

In this case, the x-axis represents the gender of the applicant, and the y-axis represents the applicant income. The plot will show the distribution of applicant income for each gender, allowing for easy comparison between the two groups.

This kind of plot can be useful for identifying differences in the distribution of data between different groups, and can help to identify any potential biases or inequalities in the data.

The violin plot shows the distribution of data points in a way that allows for easy comparison between different groups. It displays a box plot in the center, which shows the median, quartiles, and whiskers of the data. On either side of the box plot, the plot shows a kernel density estimation of the data, which is represented by a "violin" shape.

In this case, the x-axis represents the Property_Area of the applicant, and the y-axis represents the applicant income. The hue parameter is used to distinguish between different levels of education of the applicant. The plot will show the distribution of applicant income for each combination of Property_Area and Education, allowing for easy comparison between the different groups.

This kind of plot can be useful for identifying differences in the distribution of data between different groups, and can help to identify any potential biases or inequalities in the data. It can also help to identify any correlations between different variables, such as the relationship between the applicant's education level and their income.

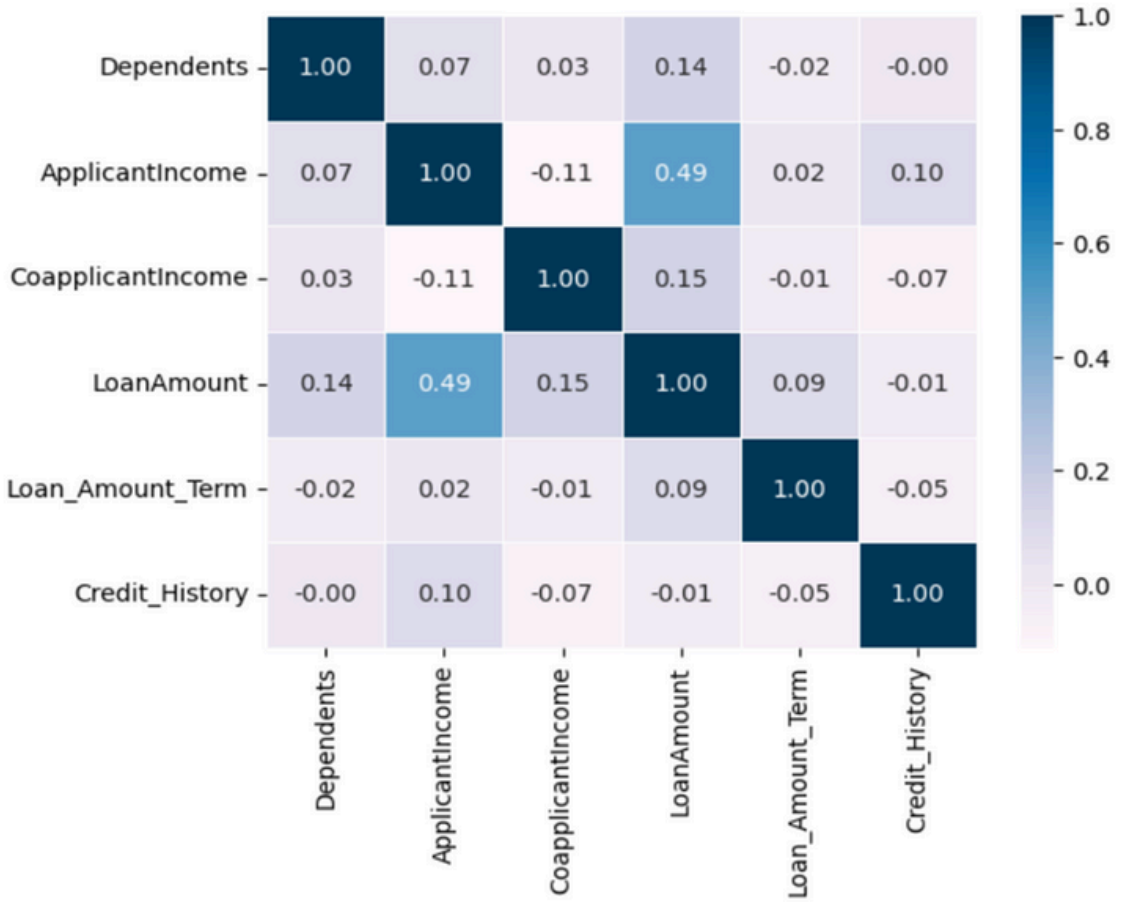
Data Visualizations:

Multi-Variate Analysis:

Multivariate analysis is a statistical method used to analyze the relationships between multiple variables in a dataset. It is a more complex form of analysis than univariate or bivariate analysis, which only consider one or two variables, respectively. Multivariate analysis allows researchers to identify patterns and correlations between several variables simultaneously, providing a more comprehensive understanding of the data.

Multivariate analysis can be used to analyze both dependent and independent variables. Dependent techniques, such as multiple linear regression and multiple logistic regression, are used to examine the relationship between one dependent variable and multiple independent variables. Independent techniques, such as factor analysis and cluster analysis, are used to explore the structure of a dataset without a specific dependent variable in mind.

Correlation



This is a Heatmap which is used to show the correlation between different factors that might be considered by a bank when deciding whether to approve a loan. Here are the factors listed on the heatmap and how they relate to each other according to the colors:

- Dependents: A positive correlation is shown with loan amount (0.14) and applicant income (0.07). A negative correlation is shown with credit history (-0.00). This could mean that people with more dependents are more likely to get larger loans, but may also have a lower credit score.

- Applicant Income: A positive correlation is shown with loan amount (0.49) and co-applicant income (0.49). This means that people with a higher income are more likely to get a larger loan. People with a higher income are also more likely to have a co-applicant with a higher income.

- Co-applicant Income: A positive correlation is shown with loan amount (0.15) and applicant income (0.49). This means that people with a co-applicant who has a higher income are more likely to get a larger loan.

- Loan Amount: A positive correlation is shown with loan amount term (0.09) and applicant income (0.49). This could mean that people who get larger loans are more likely to get a longer loan term. People with a higher income are also more likely to get a larger loan.

- Loan Amount Term: A weak positive correlation is shown with loan amount (0.09).

- Credit History: There is a weak positive correlation with applicant income (0.10). This could mean that people with a higher credit score are more likely to get a loan approval.

Conclusion

After conducting a comprehensive analysis of home loan approval dataset have several significant factors that influence the approval process. Mainly, Applicant Income level and Credit History score are crucial, with higher income and better Credit History have much better increasing chance of home loan approval.

Moreover Employment status also plays a vital role, with employed individuals more likely to be approved compared to unemployed or self-employed applicants.

Furthermore trends, such as differences in approval rates among various age groups and genders, are also noteworthy. As a person with relatively young age tends to not get home loan approval in comparison to a adults as adults are relatively much mature.

Gender differences do exist in home loan access, with female typically receiving smaller home loan amounts rather than male. These differences may be due to structural factors, such as the age, credit history, and Applicant Income.

By using the insights obtained from this analysis, financial institutions can make more informed decisions regarding home loan approvals.