

**DESCRIPTIVE STATISTICS**

**AND**

**EXPLORATORY DATA ANALYSIS**

**ON**

**CREDIT DATA**

**Submitted to: Submitted By:**

**Dr. Bhajneet Kaur Parisha Bansal**

**(23PGDM225)**

**EXECUTIVE SUMMARY**

This report showcases the in-depth analysis of the loan application data and reveals significant insights into the factors influencing loan amounts and the characteristics of loan applicants.

The analysis reveals that loan amounts are generally influenced more by applicant income than co-applicant income. Additionally, individuals with good credit histories tend to receive larger loans compared to those with bad credit histories. The distribution of loan amounts is right-skewed, indicating that there are a few larger loans that pull the average higher. While property area has some influence on loan amounts, it is not the primary factor. Furthermore, outliers are present in several variables, suggesting that there may be exceptional cases that deviate from the general trends.

The analysis also highlights the relationships between these variables and provides valuable insights into the factors driving loan amounts.

.

**TABLE OF CONTENTS**

[INTRODUCTION 4](#_Toc175949549)

[METHODOLOGY 5](#_Toc175949550)

[CODE, VISUALIZATIONS & INSIGHTS 7](#_Toc175949551)

[DESCRIPTIVE STATISTICS 10](#_Toc175949553)

[EXPLORATORY DATA ANALYSIS 12](#_Toc175949585)

[1) UNIVARIATE ANALYSIS 12](#_Toc175949586)

[2) BIVARIATE ANALYSIS 16](#_Toc175949590)

[3) MULTIVARIATE ANALYSIS 20](#_Toc175949591)

[CONCLUSION 25](#_Toc175949593)

[REFERENCES 25](#_Toc175949594)

# INTRODUCTION

This comprehensive report presents an in-depth analysis of the "Credit Data" dataset, which appears to be related to customer credit information. By conducting exploratory data analysis (EDA), I aim to uncover valuable insights into the factors influencing loan amounts, the characteristics of loan applicants, and the potential risks associated with credit lending. This analysis will provide a solid foundation for further research and decision-making within the lending industry.

The dataset, likely containing variables such as credit scores, loan amounts, customer income, age, and payment history, offers a rich source of information for analyzing credit risk, predicting customer defaults, and understanding customer financial behavior. Through EDA, we can identify patterns, relationships, and trends that could inform business decisions, such as improving credit scoring models, risk management, and customer segmentation.

# METHODOLOGY

**Step 1 - Data Collection:**

1. The dataset was obtained from a reliable source, likely a financial data provider or exchange.
2. The dataset included key variables such as current price, P/E ratio, dividend yield, market capitalization, return on equity (ROE), industry, and company name.

**Step 2 - Data Cleaning and Preparation:**

The dataset was cleaned to remove any missing or inconsistent values, ensuring data accuracy and reliability.

**Step 3 - Descriptive Statistics:**

Summary statistics were calculated for each variable, including count, mean, standard deviation, minimum, 25th percentile (Q1), median (Q2), 75th percentile (Q3), and maximum. This provided a basic understanding of the data distribution and central tendency.

**Step 4 - Exploratory Data Analysis (EDA):**

* **Univariate Analysis:**
  + Histograms, density plots, and box plots were used to visualize the distribution of individual variables, such as current price, P/E ratio, and dividend yield.
  + This helped identify the shape of the distributions, identify outliers, and assess the range of values for each variable.
* **Bivariate Analysis:**
  + Scatter plots were used to visualize the relationships between pairs of variables, such as the relationship between current price and P/E ratio.
  + Correlation coefficients were calculated to quantify the strength and direction of these relationships.
* **Multivariate Analysis:**
  + Pair plots were used to visualize the relationships between multiple variables simultaneously, providing a comprehensive overview of the interdependencies.
  + Treemaps were used to visualize the distribution of market capitalization across industries and companies, highlighting the relative sizes of different industries and companies.
  + Correlation heatmaps were used to visualize the correlations between multiple variables, identifying groups of variables that are highly correlated or uncorrelated.

**Step 5 - Data Interpretation and Insights:**

1. Key findings from the EDA were identified and interpreted to gain a deeper understanding of the data.
2. The relationships between variables were analyzed to identify potential investment opportunities and risks.
3. Insights gained from the EDA were discussed in investment decision-making, considering factors such as valuation metrics, industry trends, and company performance.

# CODE, VISUALIZATIONS & INSIGHTS

**Importing the Dataset**

**Code**

import pandas as pd

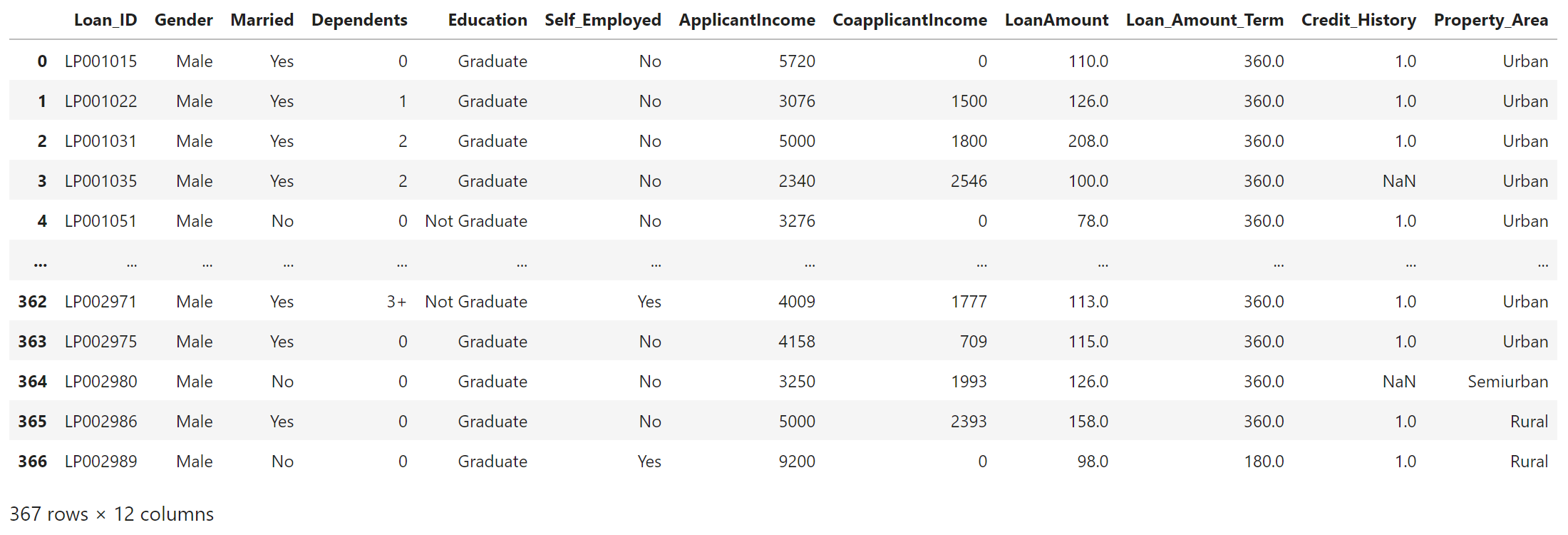
import seaborn as sns

import matplotlib.pyplot as plt

credit\_data = pd.read\_csv("C:\Credit\_Data.csv")

credit\_data

**Output**



**Data Cleaning**

**Code**

missing\_values = credit\_data.isnull().sum()

print("Missing values in each column:\n", missing\_values)

credit\_data\_cleaned = credit\_data.dropna()

for column in credit\_data.select\_dtypes(include=['object']).columns:

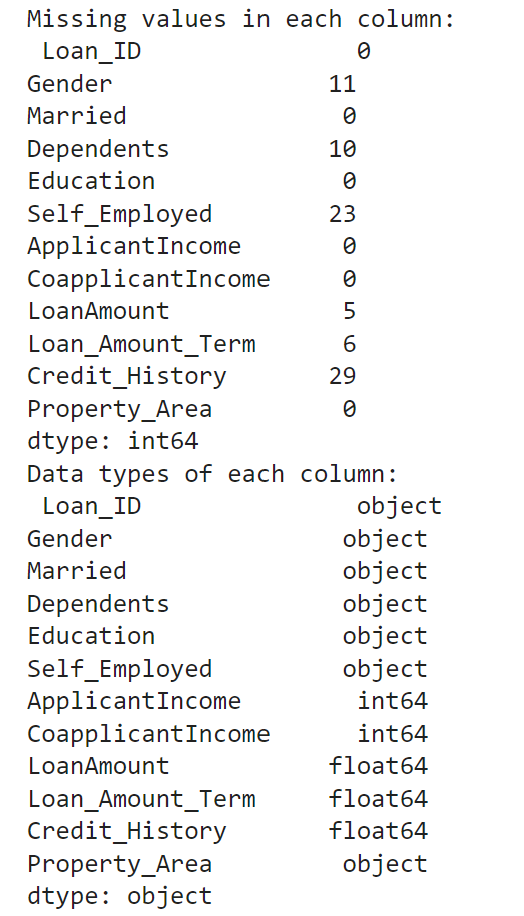
credit\_data[column].fillna(credit\_data[column].mode()[0], inplace=True)

credit\_data\_cleaned = credit\_data.drop\_duplicates()

print("Data types of each column:\n", credit\_data\_cleaned.dtypes)

print("Information of each column:\n", credit\_data\_cleaned.info)

**Output**



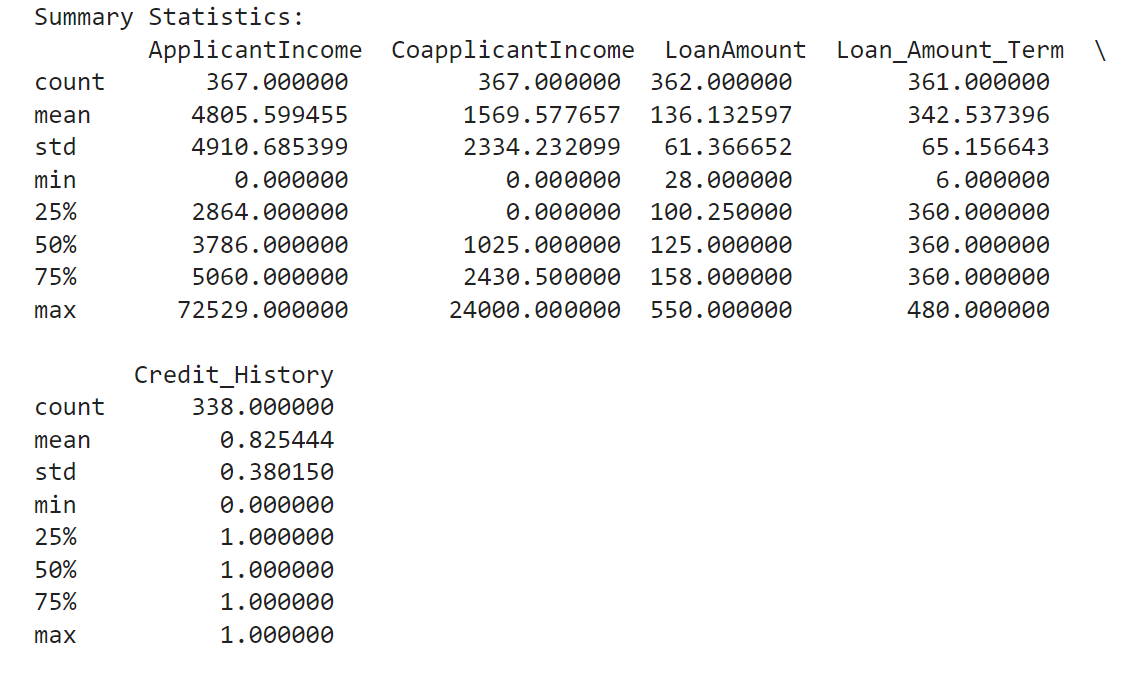
# 

# DESCRIPTIVE STATISTICS

**Code**

print("Summary Statistics:\n", credit\_data\_cleaned.describe())

**Output**



**Insights from the Descriptive Statistics**

# Applicant Income:

# The average applicant income is 4805.599455 units (assuming currency).

# There is a significant standard deviation (4910.685399), indicating a wide range of income levels among applicants.

# The minimum income is 0, suggesting some applicants might not have a primary source of income.

# The median income is 3786, indicating that half of the applicants earn less than this amount.

# The maximum income is very high at 72529, suggesting there are some outliers or high-income individuals in the dataset.

# Co-applicant Income:

# The average co-applicant income is 1569.577657 units.

# The standard deviation is 2334.232099, indicating a wider range of income levels among co-applicants compared to applicants.

# A significant number of co-applicants have 0 income, suggesting they might not be employed or might not contribute to the loan repayment.

# The median income is 1025, indicating that half of the co-applicants earn less than this amount.

# The maximum income is 24000, suggesting there are some high-income co-applicants.

# Loan Amount:

# The average loan amount is 136.132597 units.

# The standard deviation is 61.366652, indicating a moderate range of loan amounts.

# The minimum loan amount is 28, while the maximum is 550.

# The median loan amount is 125, suggesting that half of the loans are below this amount.

# Loan Amount Term:

# The average loan term is 342.537396 months.

# The standard deviation is 65.156643, indicating a moderate range of loan terms.

# The minimum term is 6 months, while the maximum is 480 months.

# The median term is 360 months, suggesting that most loans have a term of 30 years.

# Credit History:

# The average credit history is 0.825444. Since this is a binary variable (likely 1 for good credit history and 0 for bad credit history), it suggests that a majority of applicants have a good credit history.

# The standard deviation is 0.380150, indicating some variation in credit history.

# The minimum and maximum values are 0 and 1, respectively, confirming the binary nature of the variable.

# Overall Observations:

# The dataset appears to contain a mix of applicants with varying income levels and loan requirements.

# A significant number of coapplicants have no income.

# Most applicants have a good credit history.

# The loan amounts and terms vary, but the majority of loans are for 30 years.

# EXPLORATORY DATA ANALYSIS

# 1) UNIVARIATE ANALYSIS

**a) Histogram for Loan Amount**

**Code**

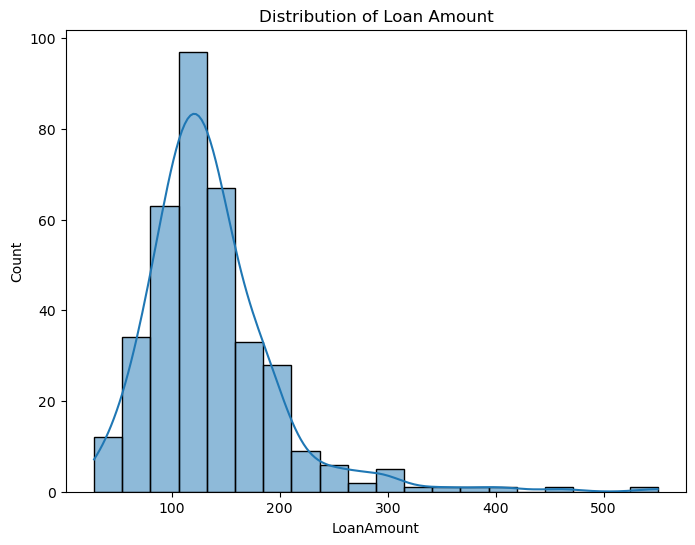
plt.figure(figsize=(8, 6))

sns.histplot(credit\_data\_cleaned['LoanAmount'], kde=True, bins=20)

plt.title('Distribution of Loan Amount')

plt.show()

**Output**



**Insights from the Loan Amount**

# The majority of loans are concentrated in the lower to mid-range.

# The presence of outliers indicates that a small number of loans are significantly larger than the typical loan amount.

# This skewness might impact statistical measures like the mean, which can be influenced by outliers.

**b) Box Plot to show distribution of Loan Amount by Property**

**Code**

plt**.**figure(figsize**=**(10, 6))

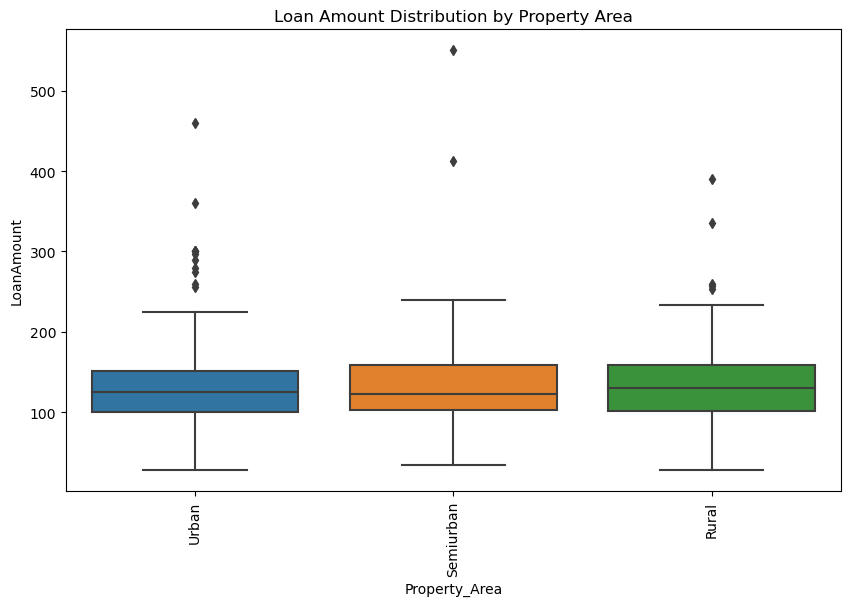
sns**.**boxplot(x**=**'Property\_Area', y**=**'LoanAmount', data**=**credit\_data\_cleaned)

plt**.**title('Loan Amount Distribution by Property Area')

plt**.**xticks(rotation**=**90)

plt**.**show()

**Output**



**Insights from the Loan Amount Distribution by Property Area**

1. **Property Area Impact:** While there are some minor differences in the distribution of loan amounts across property areas, the overall median loan amount is relatively consistent.
2. **Outliers:** The presence of outliers, especially in Urban and semi-urban areas, suggests that there are some larger loans in these areas.
3. **IQR:** The consistent IQR indicates that the variability of loan amounts within the middle 50% is similar across property areas.

**c) Pie Chart for Property Area Distribution**

**Code**

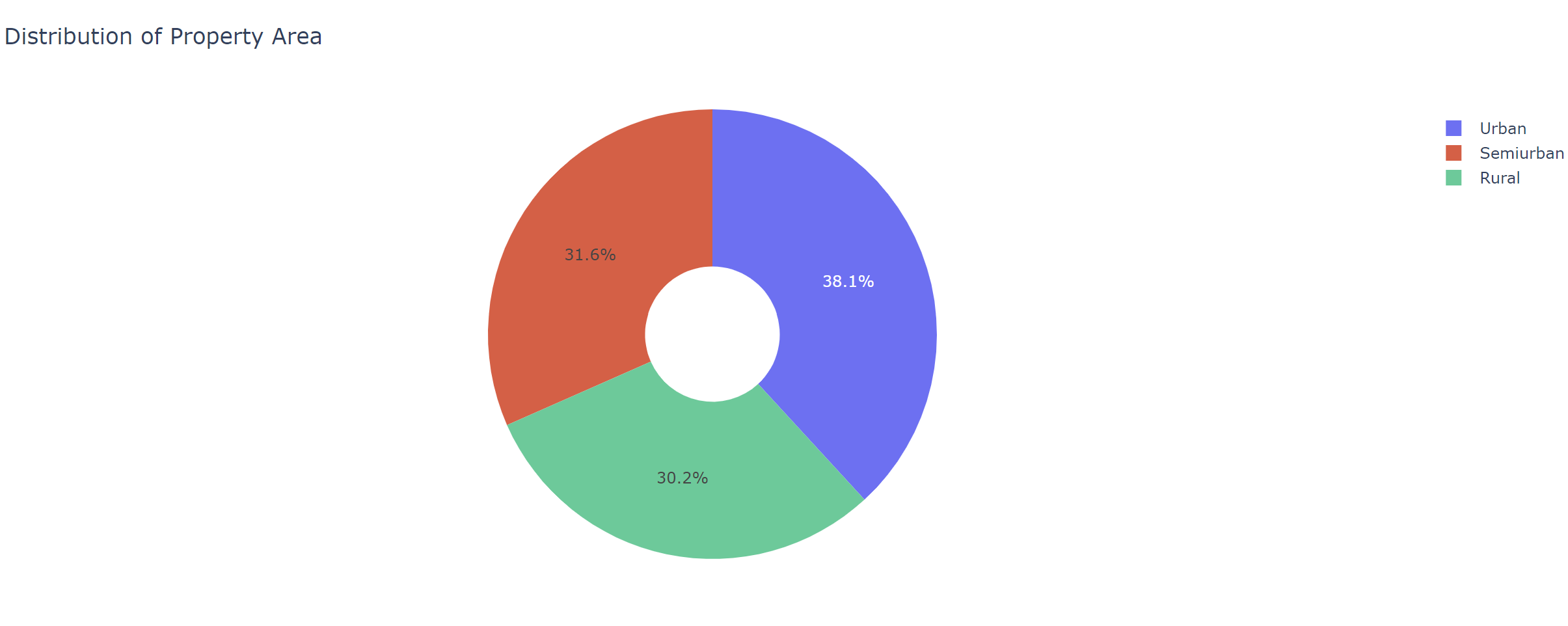
**import** plotly.express **as** px

fig **=** px**.**pie(credit\_data\_cleaned, names**=**'Property\_Area',

title**=**'Distribution of Property Area', hole**=**0.3) *# hole=0.3 creates a donut-shaped pie chart*

fig**.**show()

**Output**



**Insights from the Property Area Distribution**

1. **Urban Areas:** The most common property area is Urban, accounting for 38.1% of the data.
2. **Semiurban Areas:** The second most common property area is Semiurban, accounting for 31.6% of the data.
3. **Rural Areas:** The least common property area is Rural, accounting for 30.2% of the data.
4. The majority of the data is concentrated in Urban and semi-urban areas, indicating a higher proportion of loans in these regions.
5. Rural areas represent a slightly smaller portion of the dataset.

**d) Density Plot for Loan Amount**

**Code**

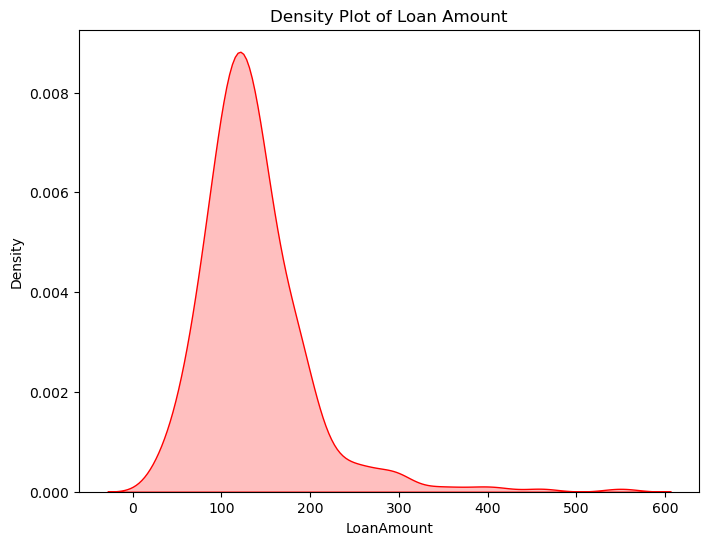
plt**.**figure(figsize**=**(8, 6))

sns**.**kdeplot(credit\_data\_cleaned['LoanAmount'], shade**=True**, color**=**'r')

plt**.**title('Density Plot of Loan Amount')

plt**.**show()

**Output**



**Insights from Density Plot of Loan Amount**

1. **Distribution:** The distribution of loan amounts is **right-skewed**, confirming the findings from the histogram. This is evident from the longer tail on the right side of the plot.
2. **Peak:** The peak of the density plot is around **100-150**, indicating the most frequent loan amount range.
3. **Skewness:** The right-skewness suggests that there are a few larger loans that pull the average higher.
4. The majority of loans are concentrated in the lower to mid-range.
5. The skewness highlights the presence of outliers, which can impact statistical measures like the mean.

# 2) BIVARIATE ANALYSIS

**a) Scatter Plot Between Applicant Income and Loan Amount**

**Code**

plt**.**figure(figsize**=**(12, 8))

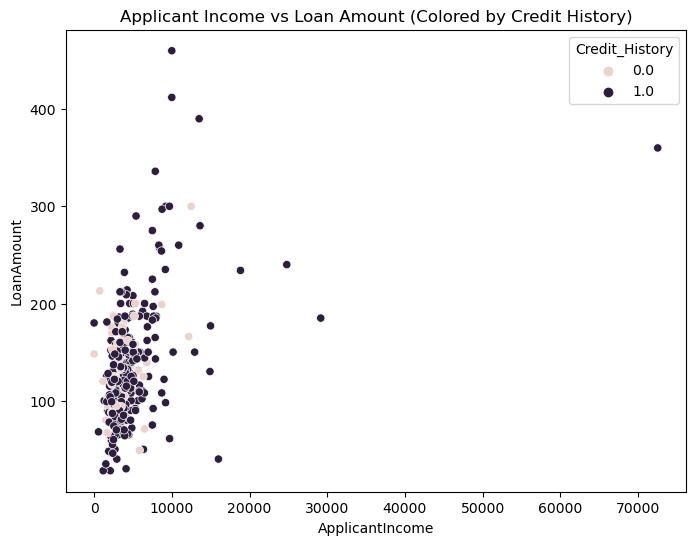
sns**.**violinplot(x**=**'Current Price (Rs.)', y**=**'Industry', data**=**nifty50\_data, scale**=**'width', inner**=None**, orient**=**"h")

sns**.**kdeplot(x**=**'Current Price (Rs.)', hue**=**'Industry', data**=**nifty50\_data, fill**=True**, common\_norm**=False**, alpha**=**0.2, linewidth**=**1.5)

plt**.**title('Ridge Plot of Current Price across Industries')

plt**.**show()

**Output**



**Insights from the Applicant Income vs Loan Amount**

1. **Income and Loan Amount:** The weak positive correlation suggests that lenders may consider applicant income when determining loan amounts. However, other factors, such as credit history, likely play a more significant role.
2. **Credit History Influence:** The clear separation between individuals with good and bad credit histories indicates that credit history is a strong predictor of loan amount. Lenders may be more willing to offer larger loans to applicants with good credit.
3. **Outliers:** The presence of outliers suggests that there may be specific cases where income or credit history does not follow the general trend. These cases could be due to other factors or potential errors in the data.

**b) Pair plot for Co-applicant Income and Credit History**

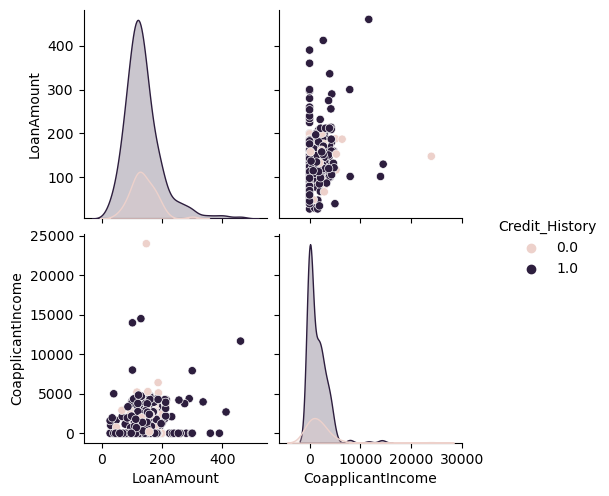
**Code**

selected\_columns **=** ['LoanAmount', 'CoapplicantIncome', 'Credit\_History']

sns**.**pairplot(credit\_data\_cleaned[selected\_columns], hue**=**'Credit\_History')

plt**.**show()

**Output**



**Insights from the Co-applicant Income and Credit History**

1. **Loan Amount and Co-applicant Income:** The weak positive correlation between Loan Amount and Co-applicant Income suggests that co-applicant income is a less significant factor in determining loan amounts than applicant income (as observed in the previous analysis).
2. **Credit History Influence:** The clear separation between individuals with good and bad credit histories across different co-applicant income levels reinforces the importance of credit history in loan approval and the determination of loan amounts.
3. **Distributions:** The right-skewness of both Loan Amount and co-applicant income indicates that there are a few outliers in both variables, which might be worth investigating further.

**c) Violin plot to compare distributions of Loan Amount by Credit History**

**Code**

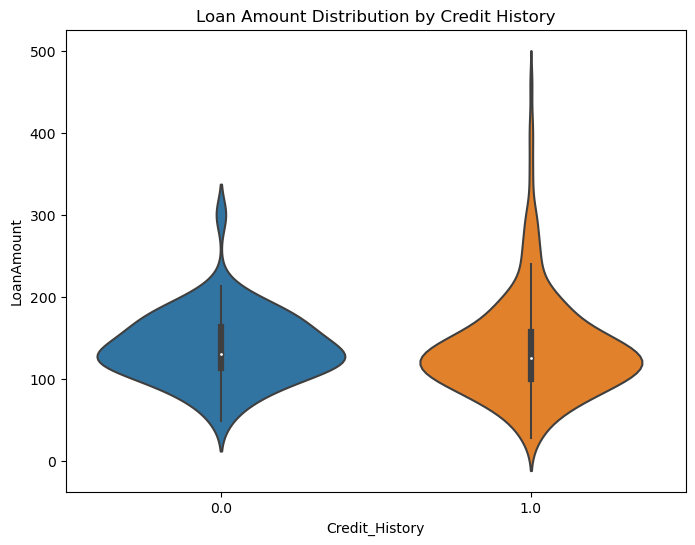
plt**.**figure(figsize**=**(8, 6))

sns**.**violinplot(x**=**'Credit\_History', y**=**'LoanAmount', data**=**credit\_data\_cleaned)

plt**.**title('Loan Amount Distribution by Credit History)

plt**.**show()

**Output**



**Insights from the Distributions of Loan Amount by Credit History**

1. **Credit History Impact:** Individuals with a good credit history tend to have higher loan amounts on average compared to those with a bad credit history.
2. **Distribution:** The right-skewness of the distribution suggests that there are a few larger loans in both groups, but the skewness is more pronounced for individuals with a bad credit history.
3. **Density:** The concentration of loan amounts in the middle range for both groups indicates that most loans are within a similar range, regardless of credit history.
4. **Outliers:** The presence of outliers, especially in the group with a bad credit history, suggests that there may be some exceptional cases where credit history does not fully explain the loan amount.

# 3) MULTIVARIATE ANALYSIS

**a) Heat Map for Correlation**

# Code

corr\_matrix **=** credit\_data\_cleaned[['LoanAmount', 'ApplicantIncome', 'CoapplicantIncome']]**.**corr()

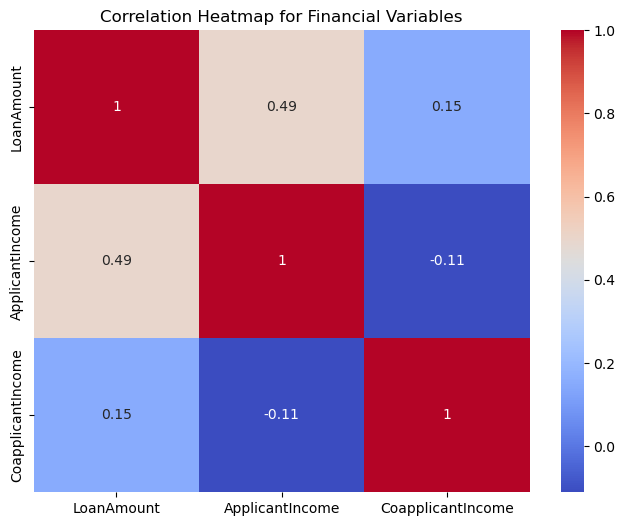
plt**.**figure(figsize**=**(8, 6))

sns**.**heatmap(corr\_matrix, annot**=True**, cmap**=**'coolwarm', cbar**=True**)

plt**.**title('Correlation Heatmap for Financial Variables')

plt**.**show()

**Output**



**Insights from the Heat Map for Correlation**

1. **Applicant Income:** Applicant income is a stronger predictor of loan amount compared to co-applicant income.
2. **Co-applicant Income:** While there is a slight positive relationship between co-applicant income and loan amount, it is not as strong as the relationship with applicant income.
3. **Negative Correlation:** The negative correlation between applicant income and co-applicant income might suggest that in some cases, individuals with higher incomes may have lower-earning co-applicants or may not have a co-applicant at all.

**b) 3D Scatter Plot of Income and Loan Amount**

**Code**

**import** plotly.express **as** px

credit\_data\_cleaned\_for\_plot **=** credit\_data\_cleaned[['ApplicantIncome', 'LoanAmount', 'CoapplicantIncome', 'Credit\_History']]**.**dropna()

fig **=** px**.**scatter\_3d(credit\_data\_cleaned\_for\_plot,

x**=**'ApplicantIncome',

y**=**'LoanAmount',

z**=**'CoapplicantIncome',

color**=**'Credit\_History',

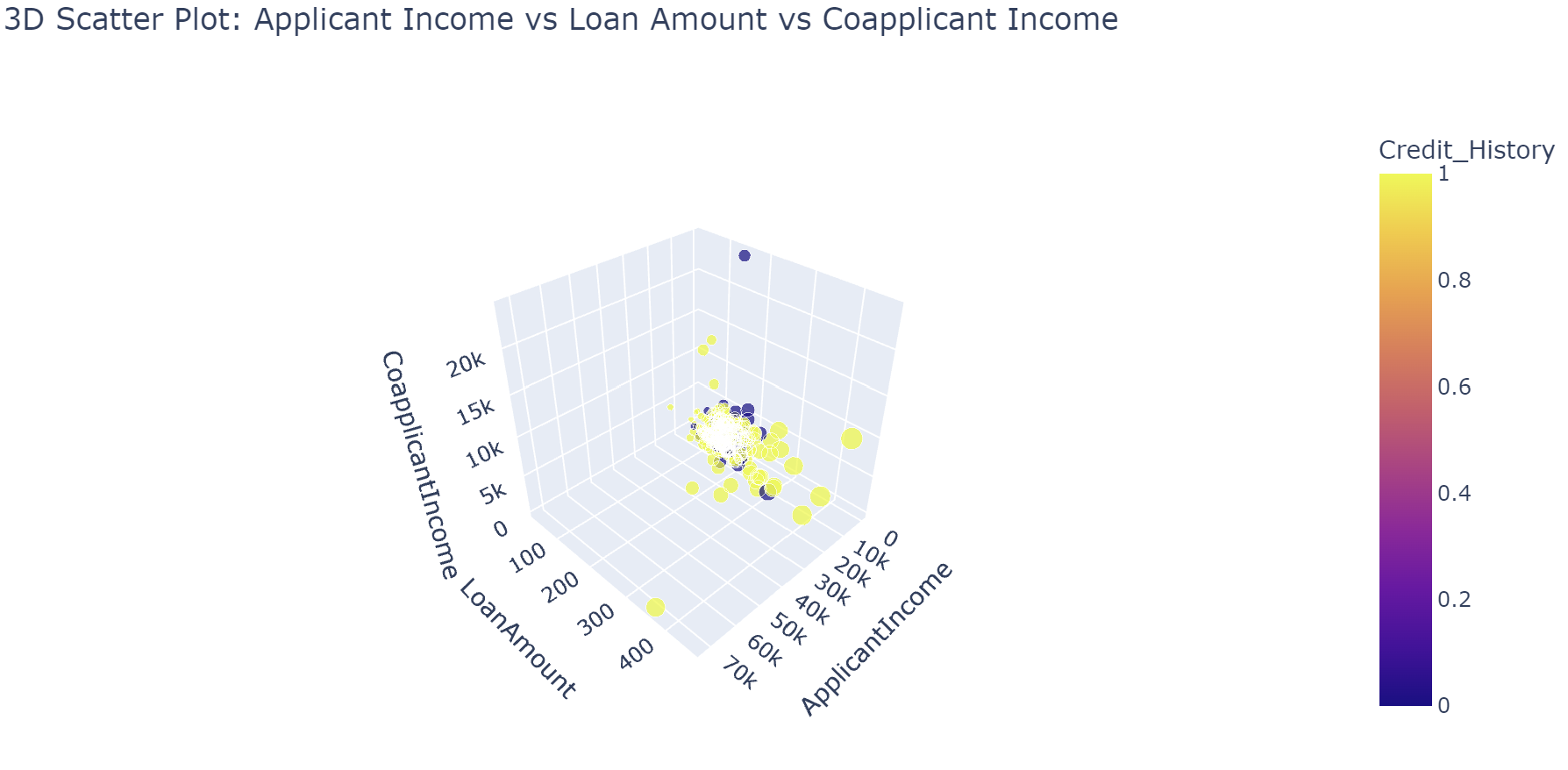
size**=**'LoanAmount',

opacity**=**0.7,

title**=**'3D Scatter Plot: Applicant Income vs Loan Amount vs Coapplicant Income')

fig**.**show()

**Output**



**Insights from the 3D Scatter Plot**

1. **Applicant Income and Loan Amount:** The plot confirms the positive relationship between Applicant Income and Loan Amount observed in previous analyses. As applicant income increases, loan amounts tend to increase as well.
2. **Coapplicant Income:** The plot suggests that coapplicant income plays a less significant role in determining loan amounts compared to applicant income. While there is a slight upward trend in loan amounts with increasing coapplicant income, the effect is not as pronounced.
3. **Credit History Influence:** The color coding by Credit History reveals that individuals with a good credit history (Credit\_History = 1) tend to have higher loan amounts across all combinations of Applicant Income and Coapplicant Income compared to those with a bad credit history (Credit\_History = 0).

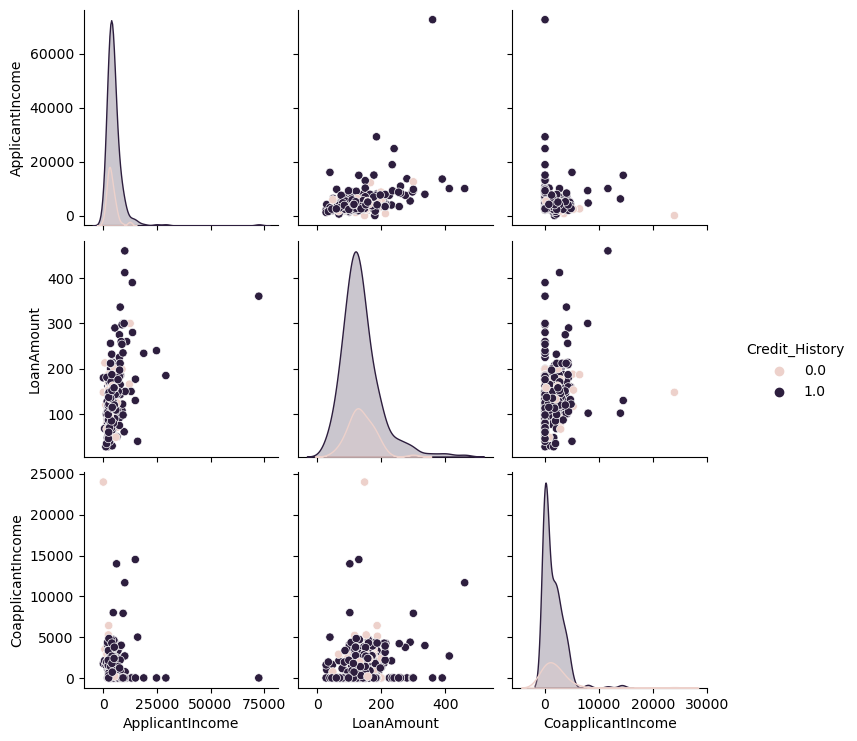
**c) Pairplot to visualize relationships between variables**

**Code**

sns**.**pairplot(credit\_data\_cleaned[['ApplicantIncome', 'LoanAmount', 'CoapplicantIncome', 'Credit\_History']], hue**=**'Credit\_History')

plt**.**show()

**Output**



**Insights from the Pair Plot from Different Variables**

1. **Applicant Income:** Applicant income is a stronger predictor of the loan amount compared to co-applicant income.
2. **Co-applicant Income:** While there is a slight positive relationship between co-applicant income and loan amount, it is not as strong as the relationship with applicant income.
3. **Credit History Influence:** The clear separation between individuals with good and bad credit histories across different combinations of Applicant Income and Co-applicant Income reinforces the importance of credit history in loan approval and the determination of loan amounts.
4. **Outliers:** The right-skewness of the distributions indicates the presence of outliers in all three variables, which might be worth investigating further.

**Overall Exploratory Data Analysis (EDA)**

**Key Findings:**

* **Loan Amount Distribution:** The distribution of loan amounts is right-skewed, with a few larger loans pulling the average higher.
* **Property Area:** While there are minor differences in the distribution of loan amounts across property areas, the overall median loan amount is relatively consistent.
* **Credit History:** Individuals with a good credit history tend to have higher loan amounts compared to those with a bad credit history.
* **Income:** Applicant income is a stronger predictor of the loan amount compared to co-applicant income. Co-applicant income has a weaker positive relationship with the loan amount.
* **Outliers:** Outliers are present in several variables, including loan amount, applicant income, and co-applicant income. These outliers might be worth investigating further.

**Relationships Between Variables:**

* **Applicant Income and Loan Amount:** There is a moderate positive correlation between Applicant Income and Loan Amount, suggesting that as applicant income increases, loan amounts tend to increase as well.
* **Co-applicant Income and Loan Amount:** There is a weak positive correlation between Co-applicant Income and Loan Amount, indicating a less significant relationship compared to the correlation with Applicant Income.
* **Applicant Income and Co-applicant Income:** There is a weak negative correlation between Applicant Income and Co-applicant Income, suggesting that as applicant income increases, co-applicant income tends to decrease slightly.

# CONCLUSION

The exploratory data analysis conducted in this report has provided a comprehensive understanding of the factors influencing loan amounts and the characteristics of loan applicants. The findings suggest that applicant income and credit history are key determinants of loan amounts, while coapplicant income plays a less significant role. The presence of outliers indicates the need for further investigation into exceptional cases.

Future research could delve deeper into the factors driving loan amounts within different credit history groups and property areas. Additionally, analyzing the impact of outliers on the overall analysis would provide a more comprehensive understanding of the data.

# REFERENCES

* Golisaivamsireddy. (2024, April 16). *EDA\_Financials\_Dataset*. Kaggle. <https://www.kaggle.com/code/golisaivamsireddy/eda-financials-dataset>
* Tiwari, M. (2023, December 3). Complete Exploratory Data Analysis (EDA)of loan data and visualization: understanding loan characteristics and default patterns. *Medium*. <https://medium.com/nerd-for-tech/complete-exploratory-data-analysis-eda-of-loan-data-and-visualization-understanding-loan-120651cfefc8>
* Descriptive statistics with Python. (n.d.). <https://www.pythonfordatascience.org/descriptive-statistics-python/>
* Dhaduk, H. (2022, September 27). *EDA: Exploratory Data Analysis with Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/eda-exploratory-data-analysis-with-python/