# EXPLORATORY DATA ANALYSIS

On Bank Loan Approval Dataset

# DATA OVERVIEW

- **Introduction**

- **Welcome to our journey through the data! In this exploratory analysis, we'll unravel the mysteries hidden within our dataset. Buckle up as we dive into the financial dimensions of our applicants.**

# The Dataset in a Nutshell

- **Our dataset comprises loan applicants, each represented by a unique identifier (**SK_ID_CURR**). These Candidates seek financial assistance for various purposes. Let's peek into the data provided by them:**
- **1. **Demographics and Basics**:**
- **- We have both male and female applicants.**
- **- Some own cars, while others hold real estate .**
- **- Families vary in size, with children  adding to the mix.**
- **2. **Financial Landscape**:**
- **- **Income**: Our applicants span a wide income range—some are swimming in gold coins, while others are counting every rupee.**
- **- **Credit Requests**: They seek different credit amounts (**AMT_CREDIT**), reflecting their aspirations or needs.**
- **- **Annuities**: Monthly payments are on the horizon—annuities (**AMT_ANNUITY**).**
- **- **Goods Price**: The price of their desired goods (**AMT_GOODS_PRICE**) fuels their aspirations.**

- **Employment and Time Metrics:**

- **Education level (NAME_EDUCATION_TYPE)**
- **Income type (NAME_INCOME_TYPE)**
- **Days since birth and employment (DAYS_BIRTH, DAYS_EMPLOYED)**
- **Ownership of car and realty (FLAG_OWN_CAR, FLAG_OWN_REALTY)**
- **Social circle observations (OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE etc.)**

- **Buying Habits:**

- **Days Since Last Phone change**
- **Average Credit Score from different sources**

# IMPORTANT DATA CLEANING STEPS

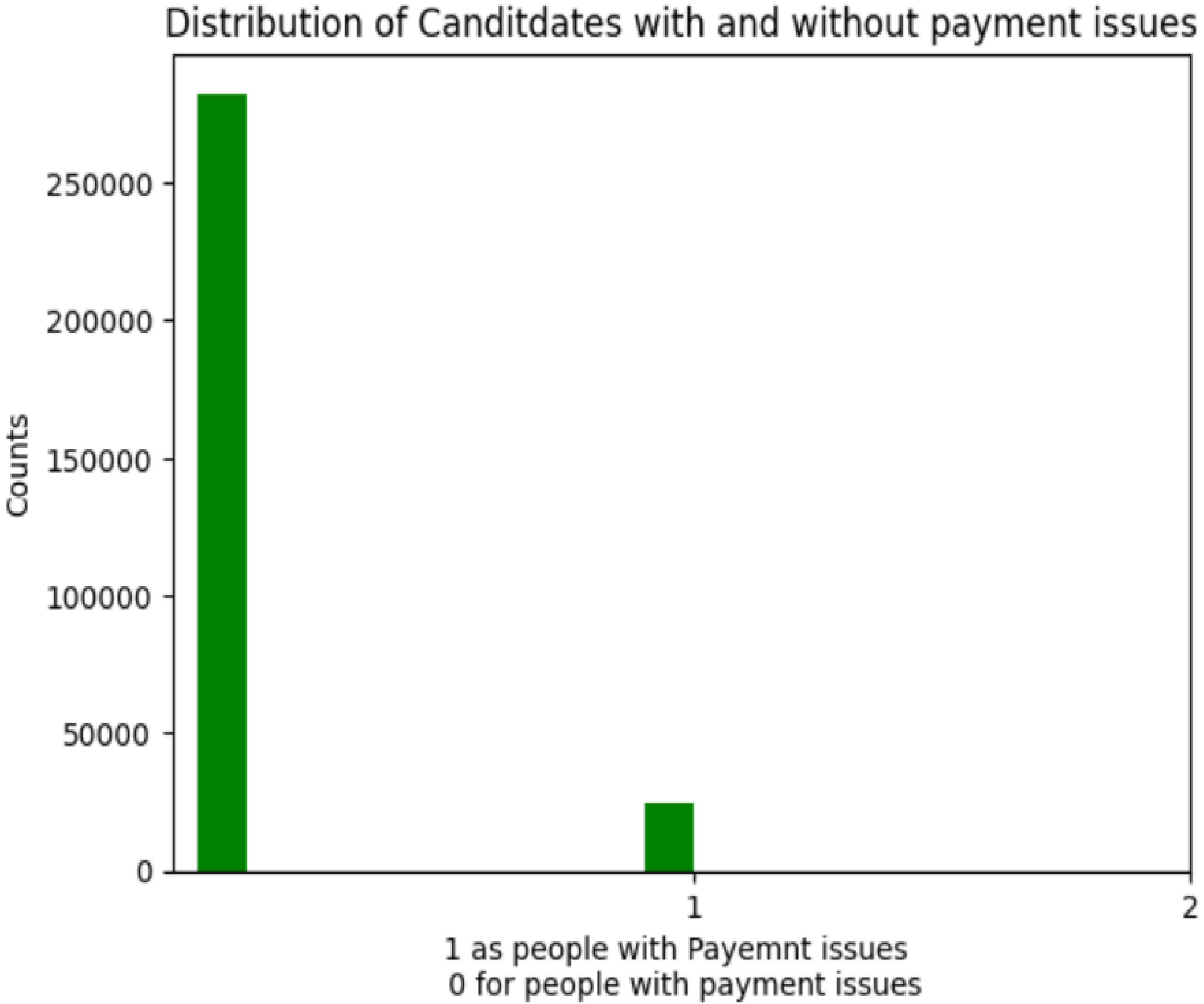
1. **<u>Dropping Irrelevant columns</u>: Columns related to user locations were not of use for our Analysis. So the columns like Apartments Avg etc which contained information about candidate's Apartment/Locations were removed.**
2. **Checking and Handling Missing values in each column: First we have checked for missing values in each column and noted down the number of missing values and data type of each column.**
3. **We broadly classified the columns into Numerical and Categorical.**
4. **For Numeric columns like "AMT_INCOME_TOTAL" , "CNT_CHILDREN" etc we have checked the skewness of each field and filled with mean or median depending upon each case.**
5. **For some numerical columns like EXT_Source_1 it didn't seem good to alter the information, so we have replaced its missing values with NAN/Null.**
6. **For Categorical columns like "NAME_TYPE_SUITE" we have replaced its missing values with Mode. For other categorical columns like "OCCUPATION_TYPE" we considered grouping missing values into different category named as "Unknown".**

# IMPORTANT DATA CLEANING STEPS

1. **Further we considered replacing 'Unknown' values in "NAME_FAMILY_STATUS" column with Mode i.e "Married" as this column contained relatively very less, missing values.**
2. **Columns like DAYS_BIRTH, DAYS_EMPLOYED contained values in form of negative days so we considered converting them into positive values and formatting them into year type and renaming columns into more readable "Age" and "Work_Experience".**
3. **Also Work_Experience had some outliers like 998 or value of 49 against one whose age was just 30 years. To handle this all at once we have used feature Engineering. In India people start working at the age of 18 so any value in work experience column could not be more than (Age – 18). Thus, we have then transformed the Wok Experience column Accordingly.**

# IMPORTANT DATA CLEANING STEPS

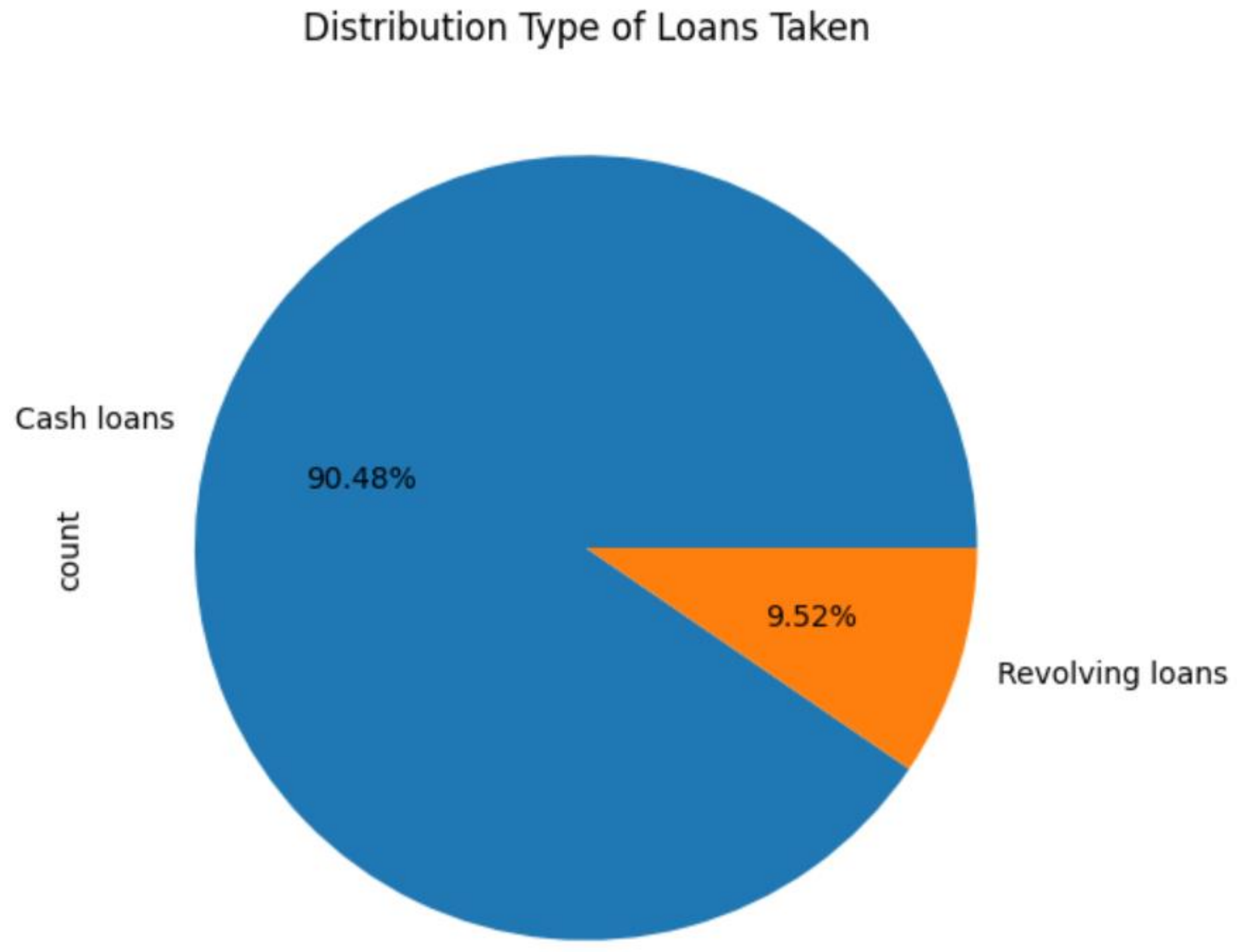
1. Then we considered merging/aggregating multiple similar columns into single column for better Analysis.
2. Like we had 3 sources for credit score of an Individual. We considered taking average of those for each individual and noting them down in one single column.
3. Also to remember we had some missing values in credit scores from first source so instead of dropping or imputing those records we chose to skip those and calculate average using only two sources against the individuals who had missing values.
4. Also, we had 21 different document submitted flag columns in our dataset. For our analysis purpose we have considered adding the number of documents provided by a single user and noting them into one column. This was much more informative and feasible then analysing each columns given we didn’t knew the purpose of any document.




## Checking Distribution of Target Variable

**As we see here:**

1. **Most of the customers didn't have difficulties in payments.**
2. **Also, according to our calculated ratio 1 out of 11 candidates had payment issues.**



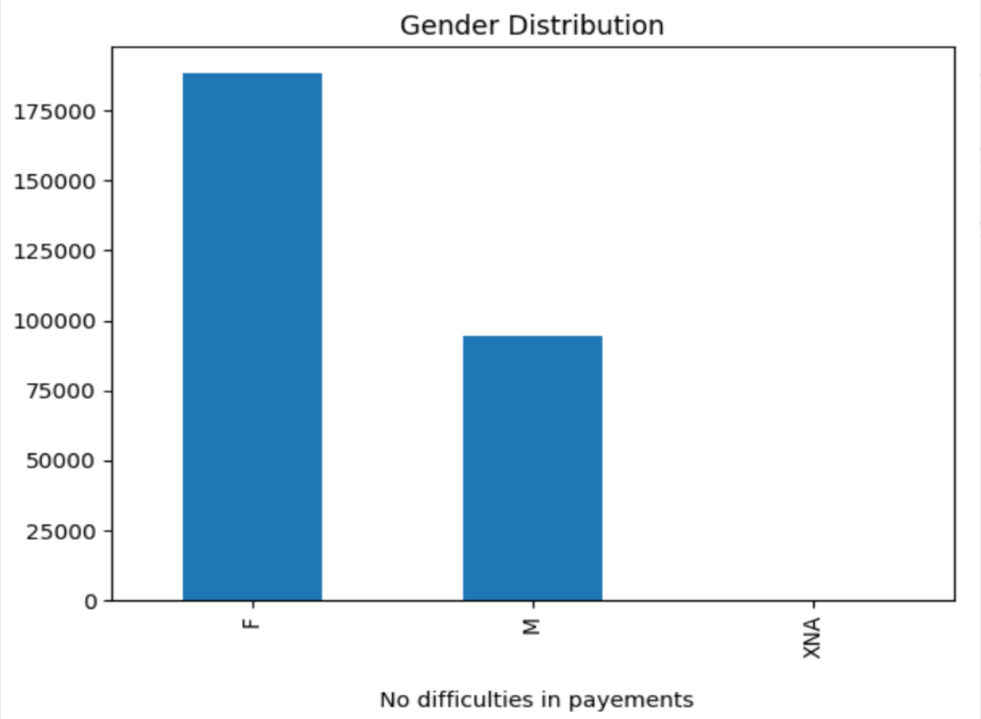

# Checking Distribution of Loan Type

**Most of candidates preferred taking Cash Loans over Revolving Loans**

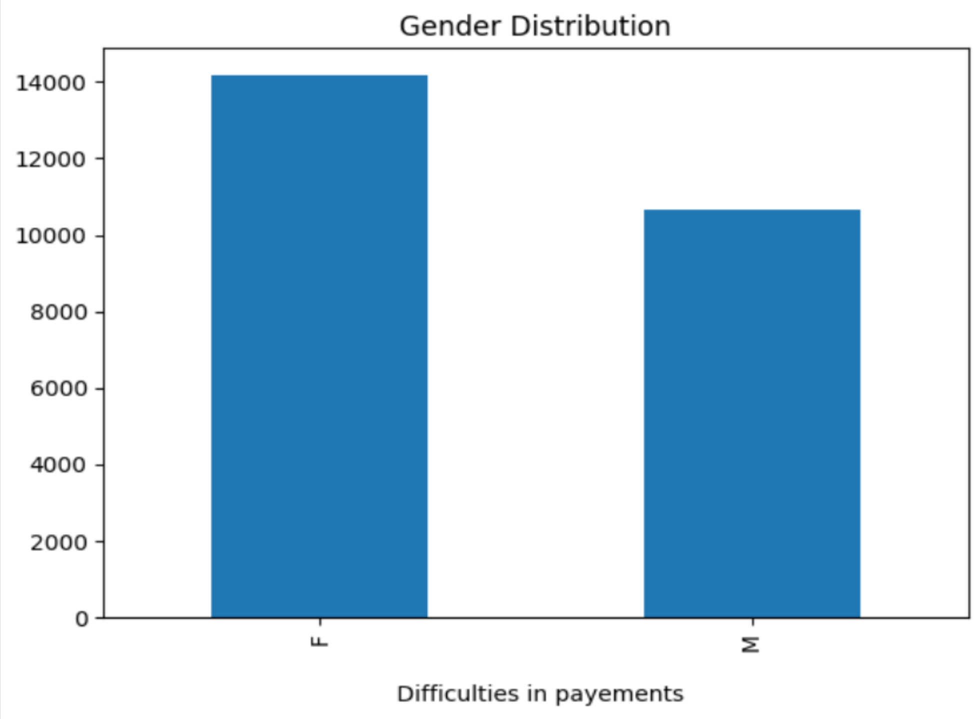
**From here we have divided our dataset into two parts with respect to target variable for separate analysis. Here Target 1 with contain data of candidates with payment issues, Whereas target 0 will contain data of candidates with no payment issues.**

# Distribution According to Gender

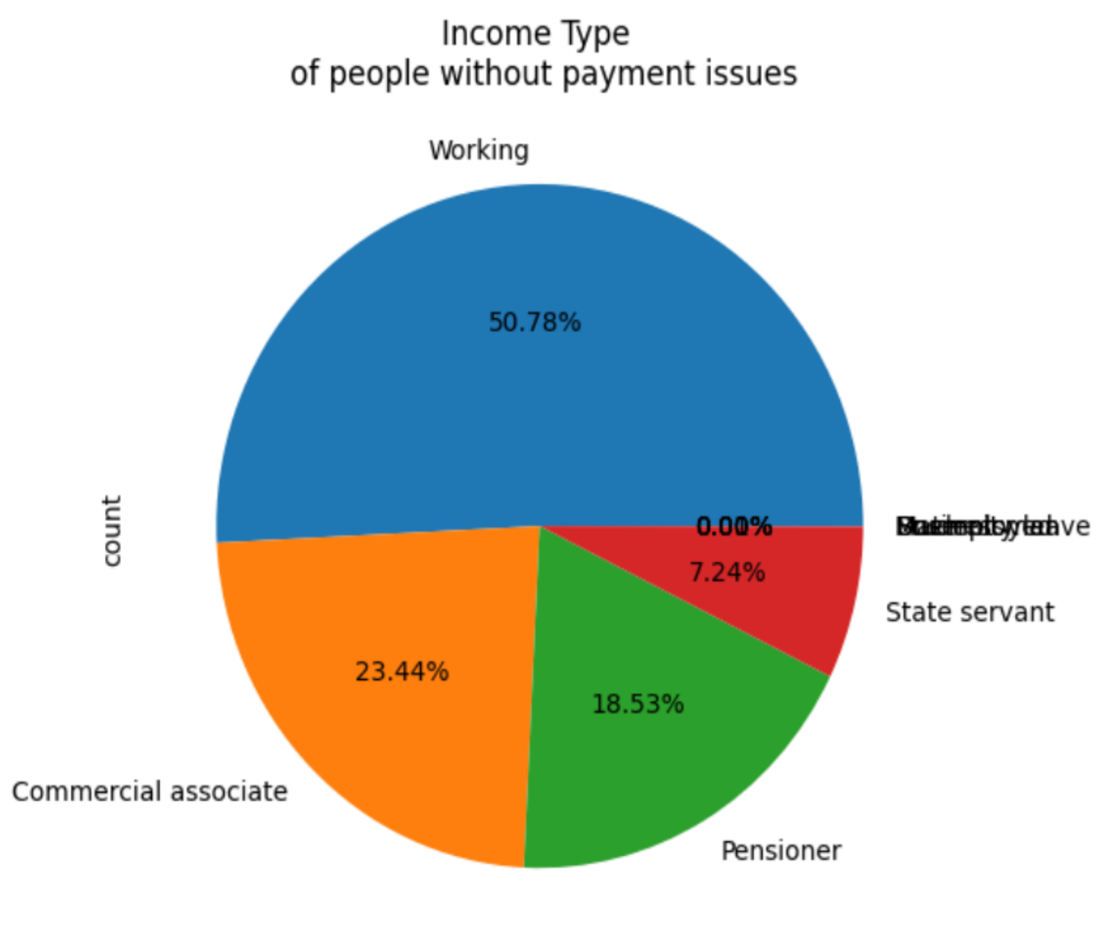
**Count of Males and Females who all had no issues with payments**

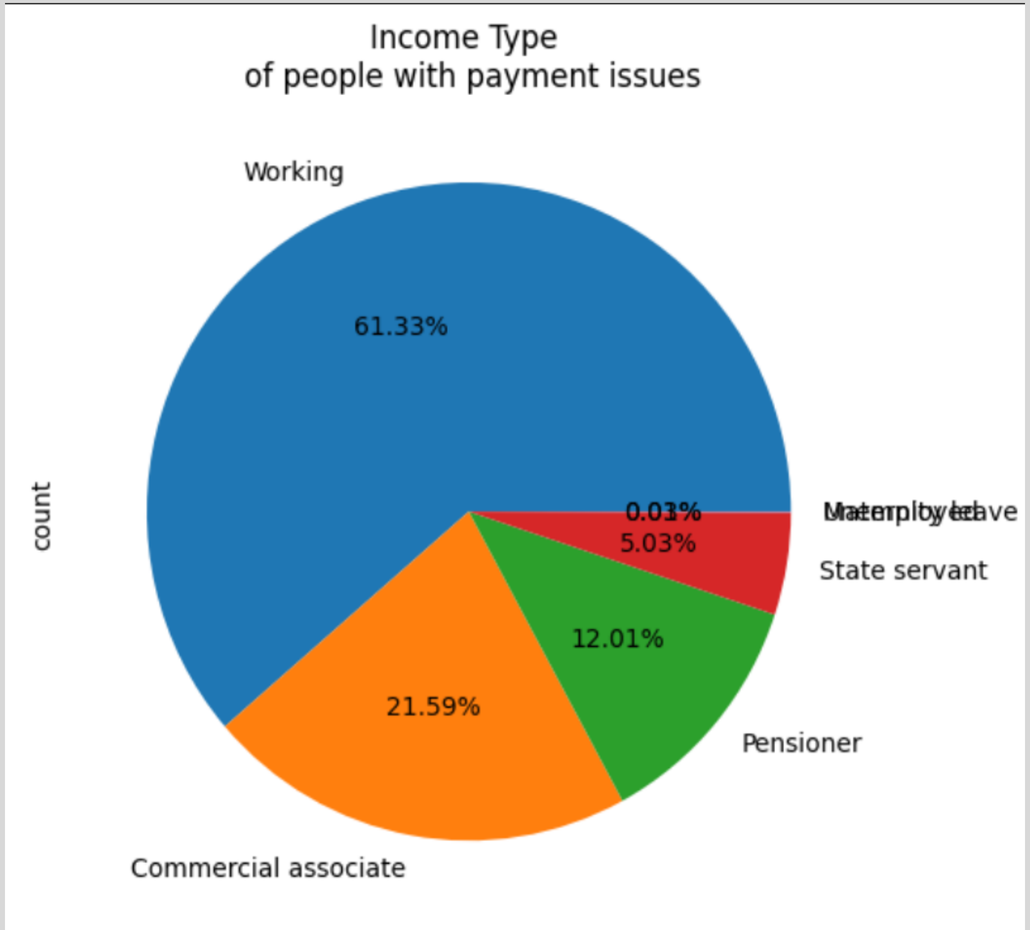



**Count of Males and Females who all had issues with payments**

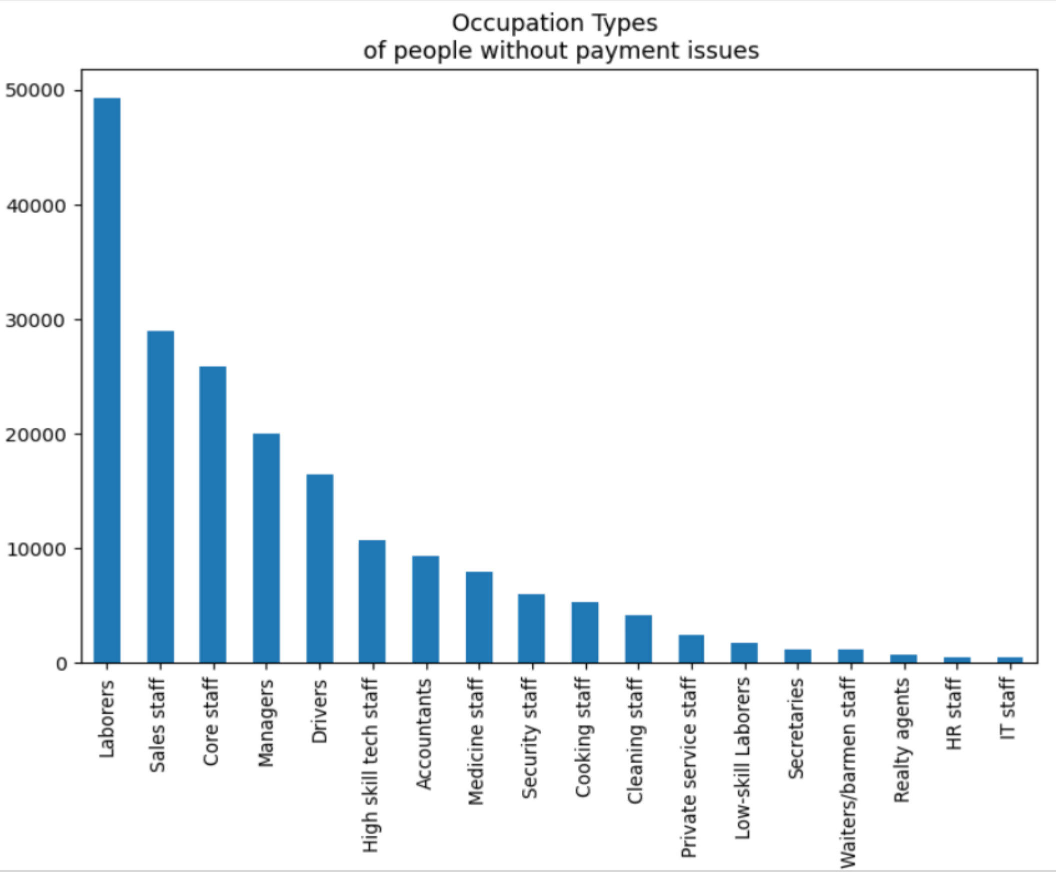


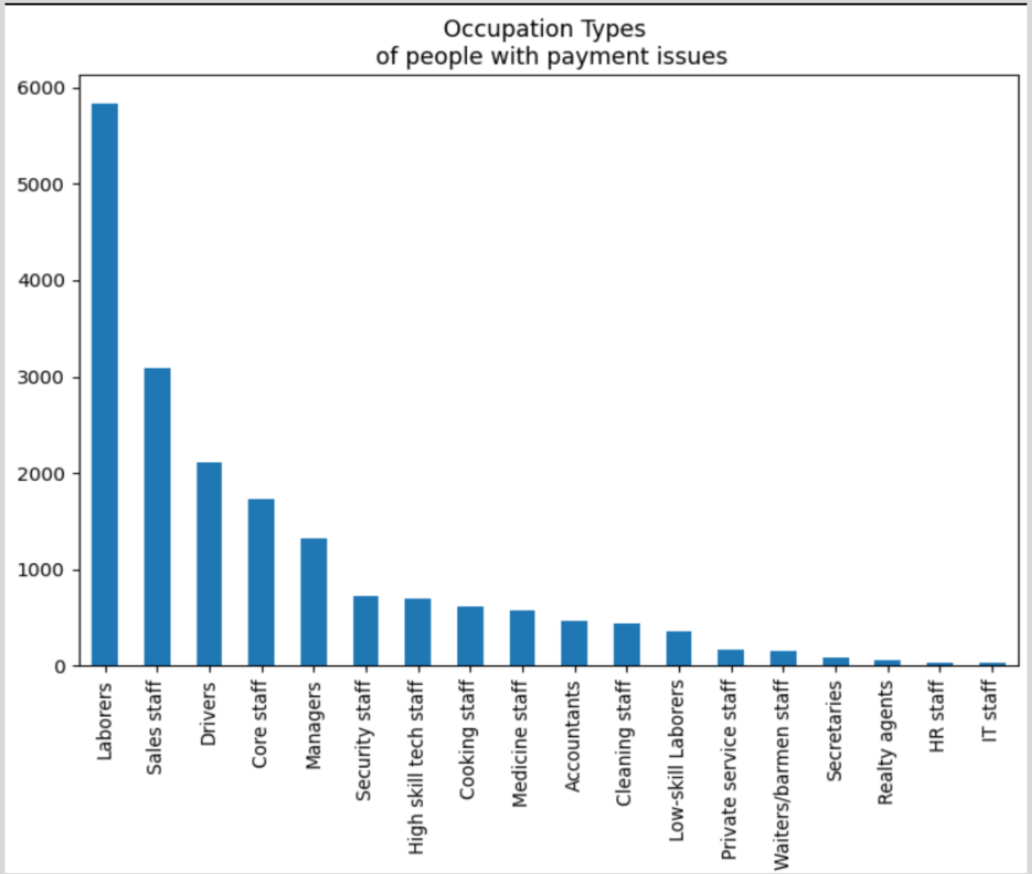
# Comparing Income Types With Respect to Payment issues.



# Occupation Types with Respect to Payment Issues



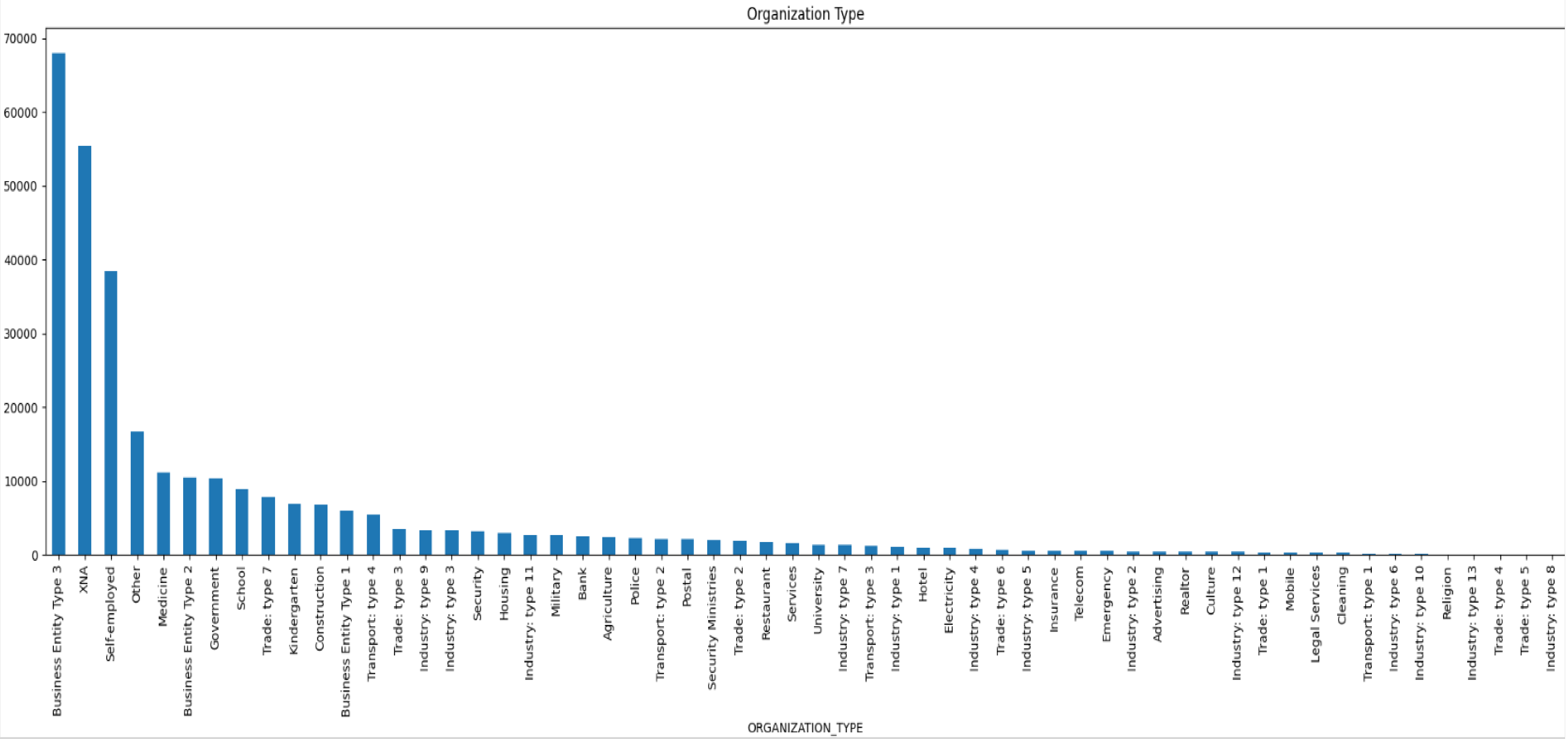
# Inferences

- **It has been noted that a significant majority of loan applicants with no payment issues are Working Professionals, followed by Commercial Associates and Pensioners. However, among those reporting payment issues, Working Professionals constitute over 60% of the observed cases.**

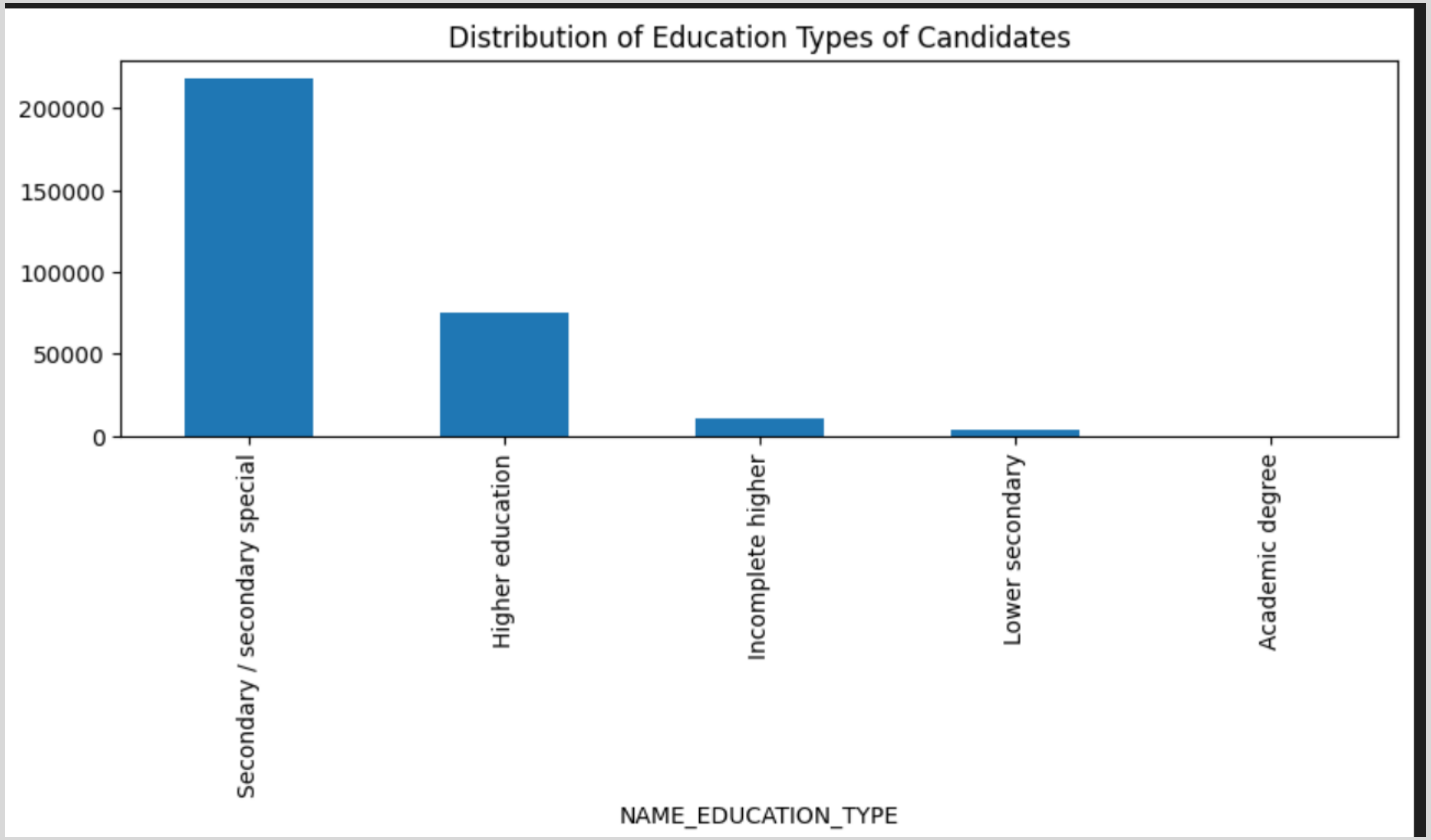
- **A notable trend reveals that the Occupation type 'Unknown' predominates among loan applicants. Labourers are the most frequent applicants in both categories—those with and without payment issues—followed closely by Sales Staff.**

- **Additionally, it is observed that Security and other staff, excluding Sales Staff, show fewer instances of payment issues compared to Managers.**

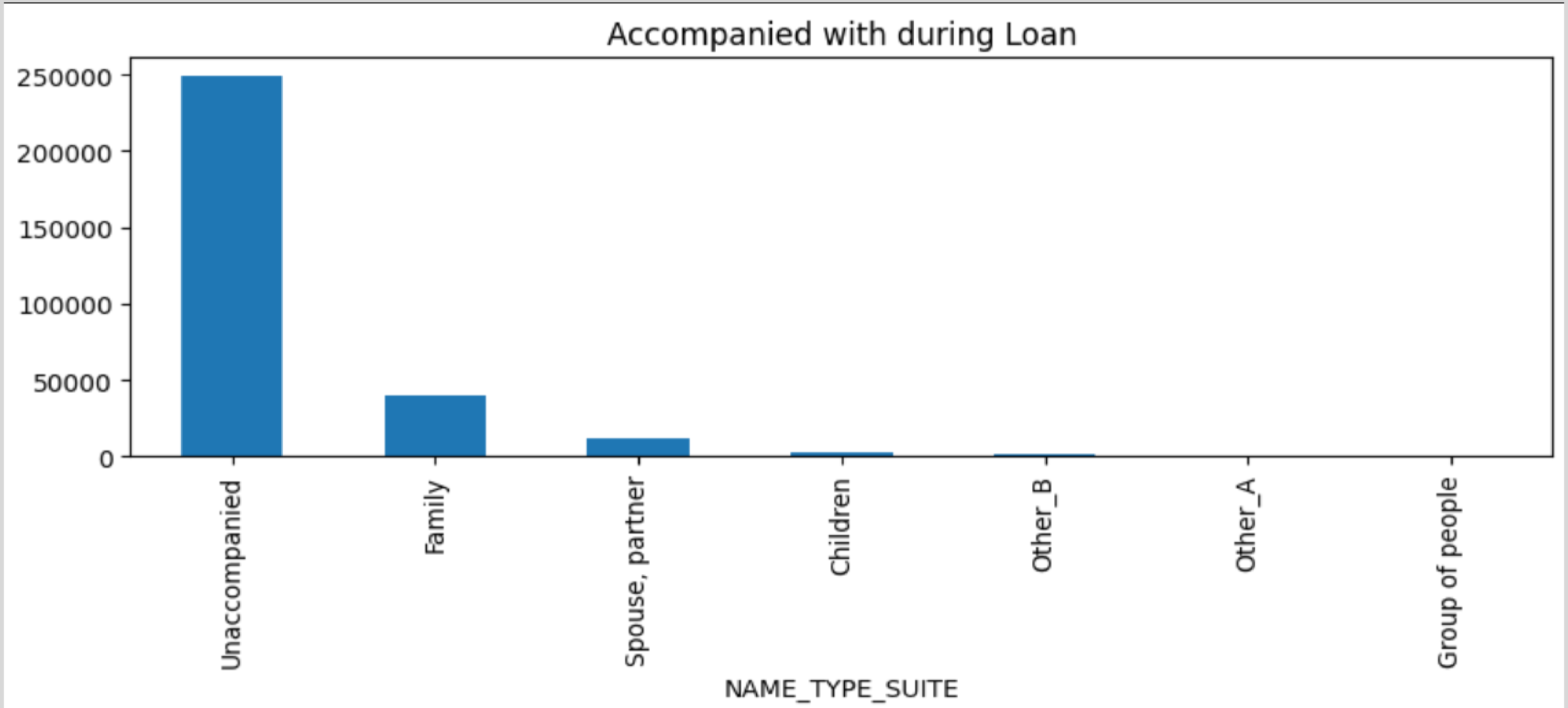
# Organization Type Distribution



# Education Types of Candidates



# Candidates Accompanied with During Loan Process



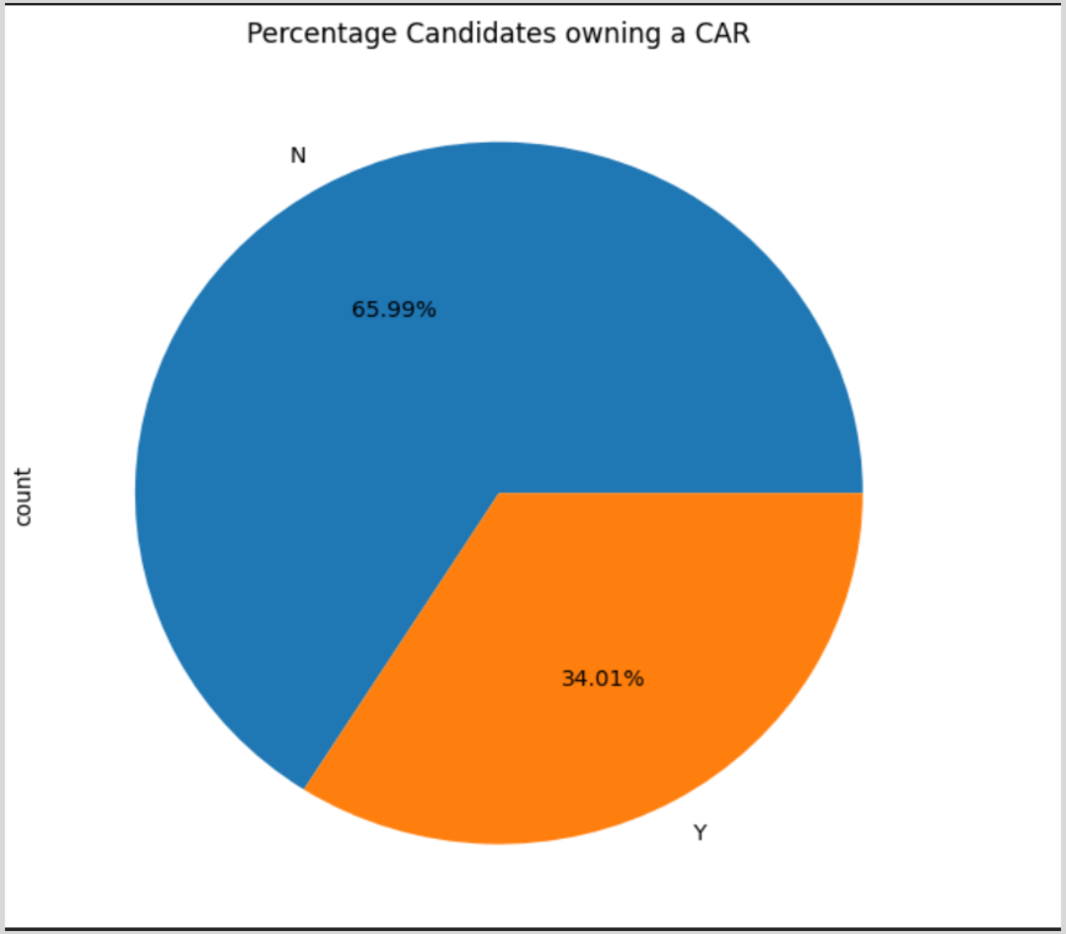
# Inferences

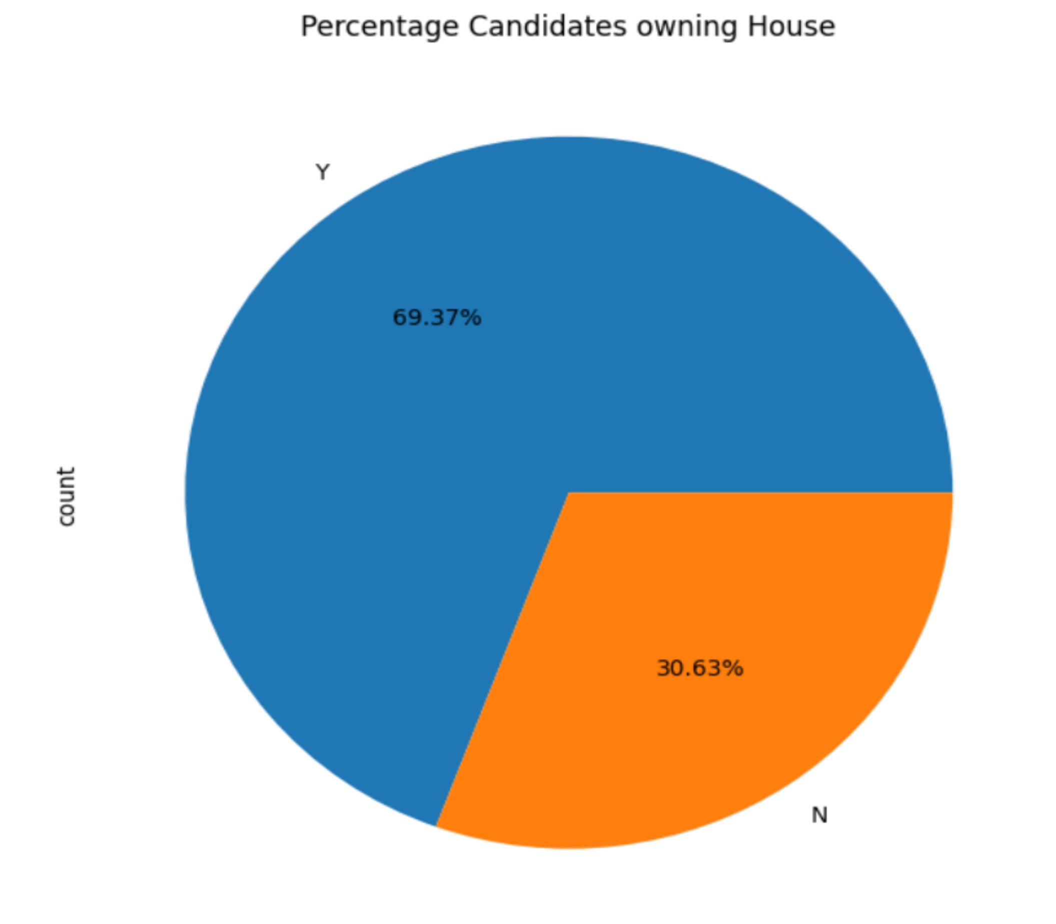
- **A significant majority of loan applicants were unaccompanied during the application process, suggesting they are financially independent decision-makers.**

- **A large percentage of professionals obtaining loans are either from the business sector or are self-employed. This trend aligns with the initial need for substantial investment in their businesses. It also explains why the majority of candidates choose to keep their Occupation Type as 'Unknown'; many individuals starting businesses prefer not to disclose their occupation until their business is established.**

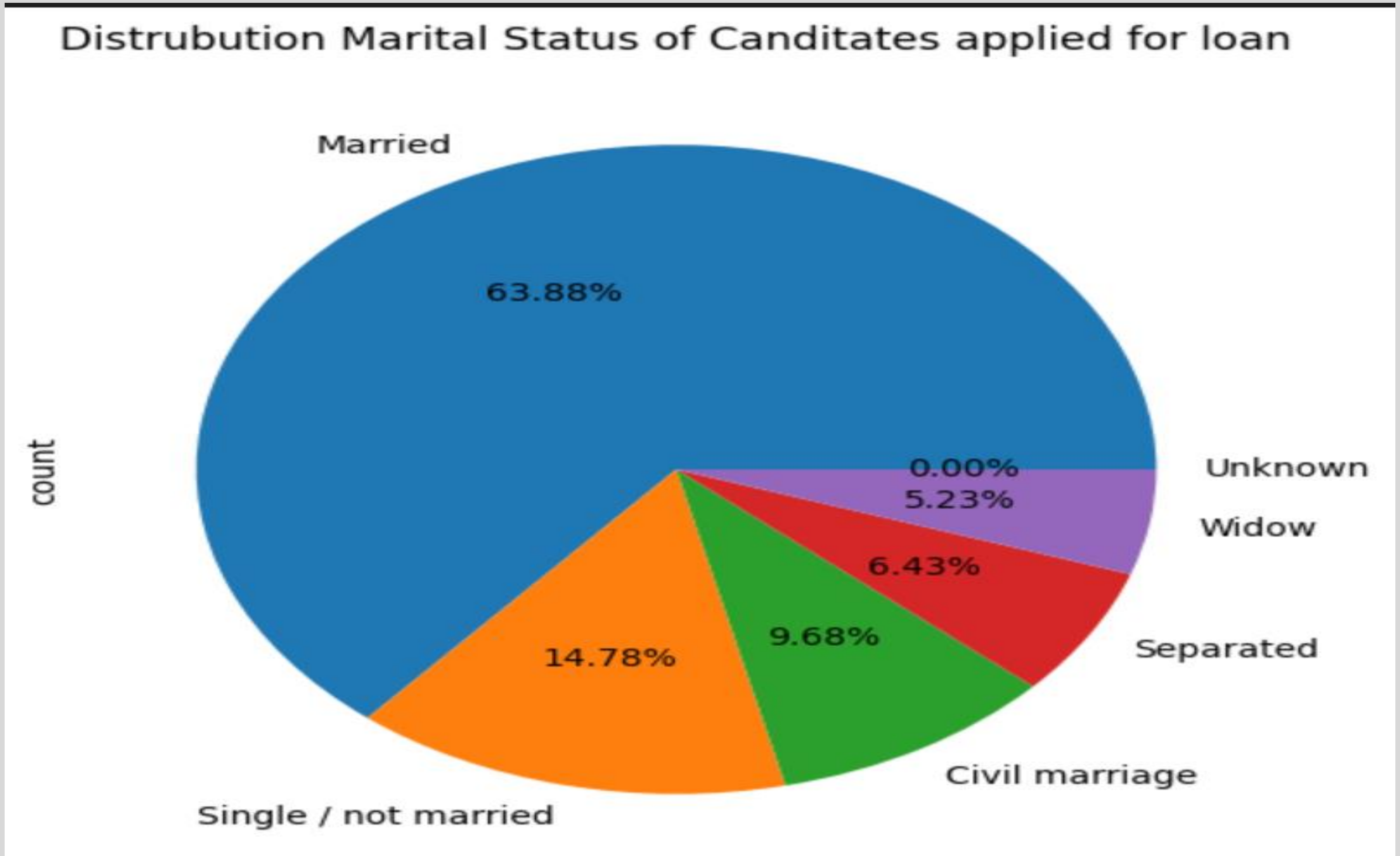
- **Candidates with secondary education backgrounds appear to take more loans compared to those with higher education qualifications. This trend suggests they may be applying for educational loans to finance further educational pursuits.**

# Taking a view on assets owned by our Candidates

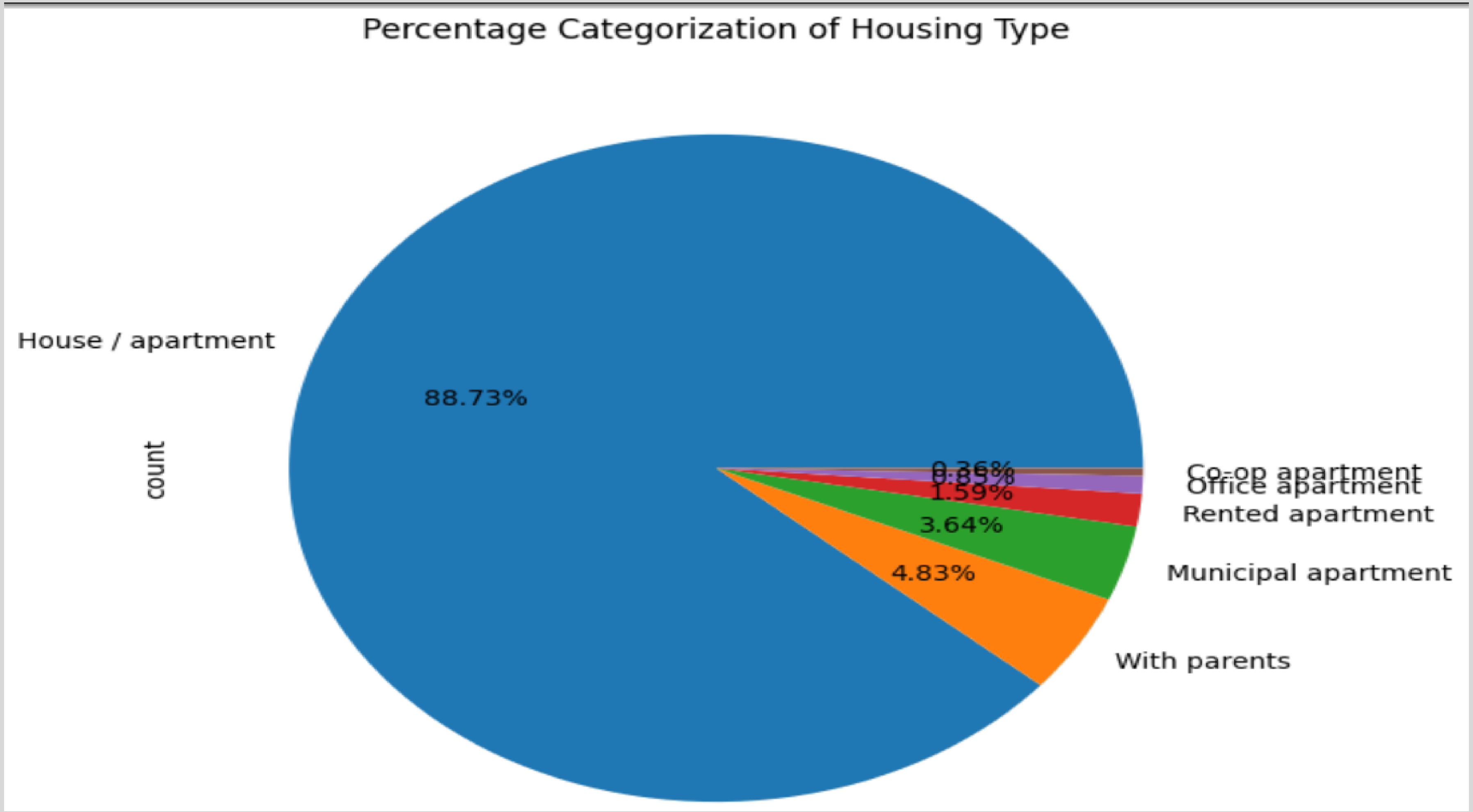


# Marital Status Distribution of Candidates



# Distribution of Housing Type

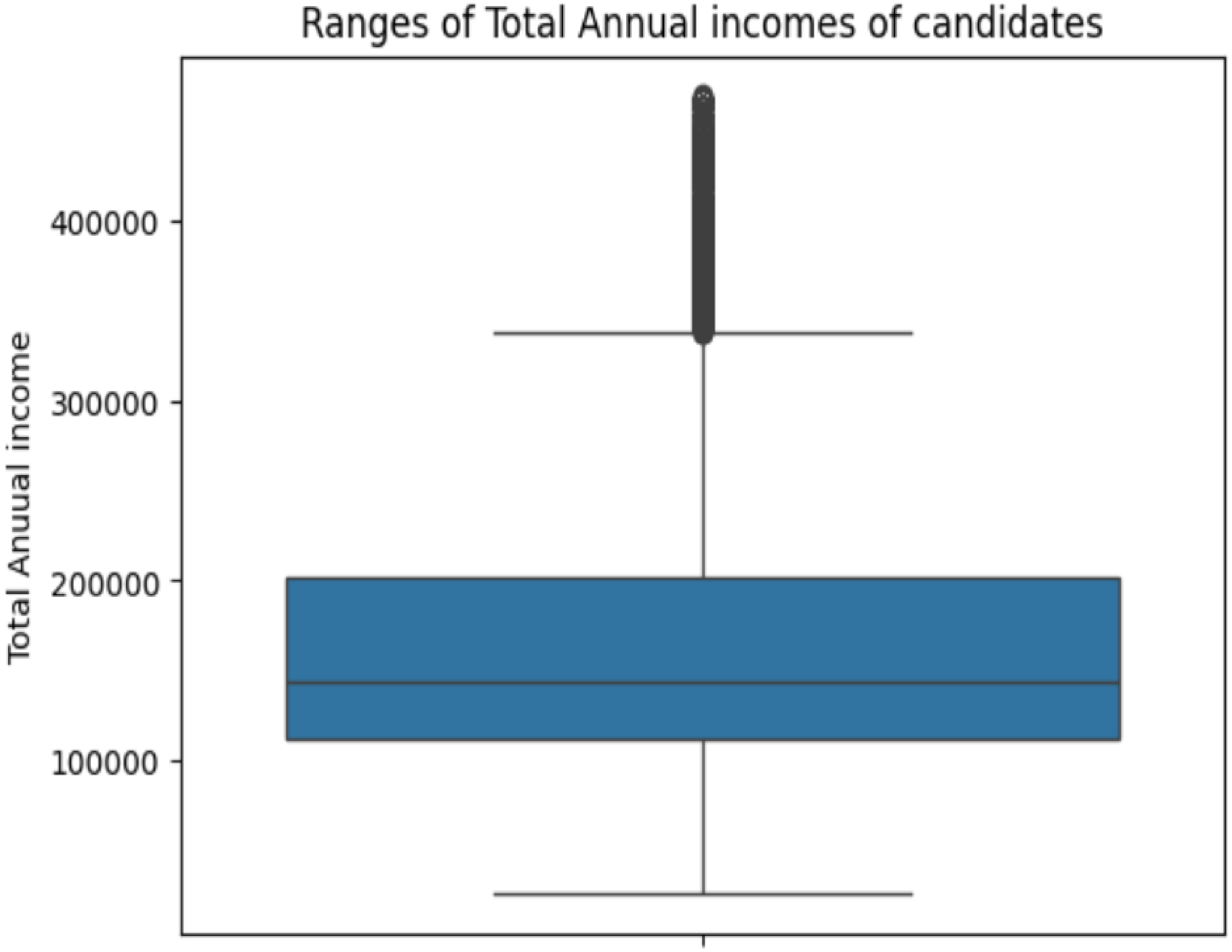


# Inferences

- **The majority of loan applicants are married, followed by single individuals. This suggests they may be taking loans for future planning purposes.**

- **Most loan applicants own a house but do not own a car. This is likely because candidates with an existing car loan are less inclined to take on another loan simultaneously.**

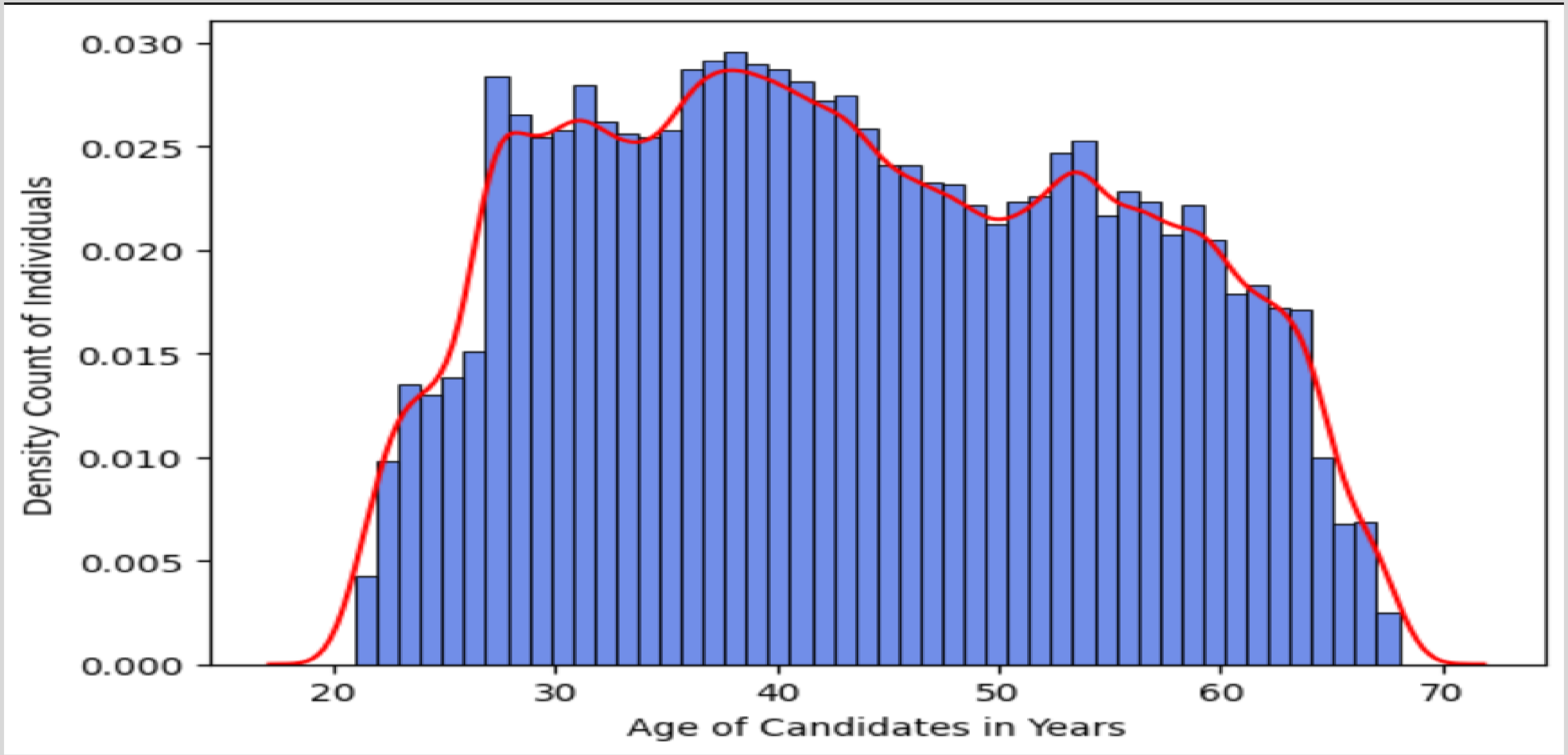
- **Candidates who do not own a house are less likely to apply for a loan. This could be attributed to their relatively lower financial stability.**




# Total Annual Incomes of Candidates

- **Most loan applicants have an annual salary below 2 lakhs, with fewer applicants earning above this threshold.**
- **However, there is a data point where an applicant with a salary of 12 lakhs per annum has applied for a loan, which warrants careful scrutiny and verification before approval.**

# Age Distribution of Candidates



# Inferences

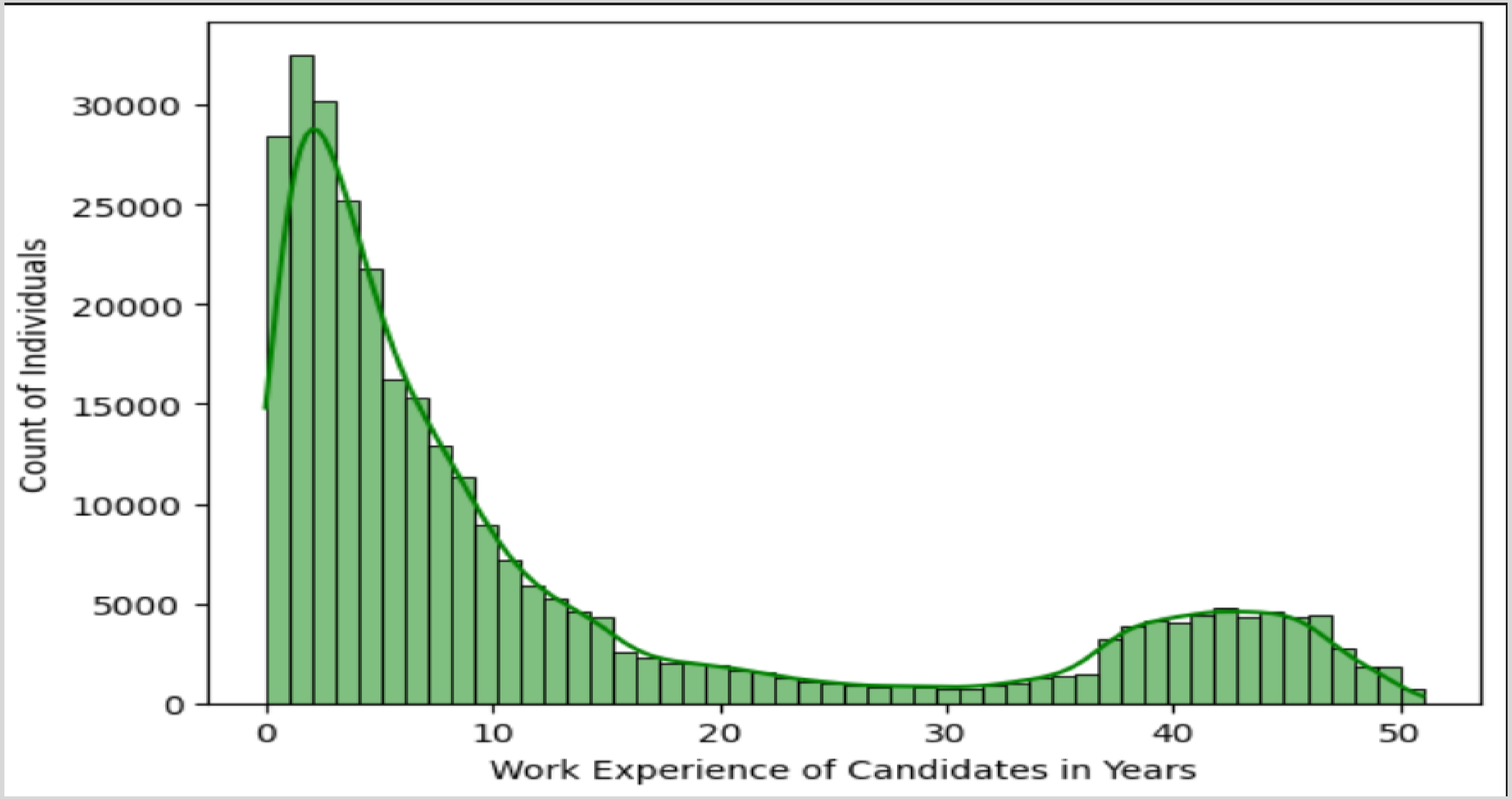
- **The age group from mid-30s to mid-40s shows the highest concentration of candidates applying for loans.**

- **Candidates in their 20s have relatively fewer loans, with a majority presumably applying for educational loans. The trend shifts in the late 20s, with an increasing number of applicants.**

- **There is a significant presence of candidates throughout the age range from 30s to 60s.**

- **Loan applicants decrease significantly after the age of 60, which correlates with retirement age in India. Many individuals in this age bracket tend to achieve financial stability and are less likely to seek loans.**

# Work Experience of Candidates



# Inferences

- **Similar to age, candidates in the early years of their careers tend to take loans, as they have fewer financial obligations and can manage repayments more easily."**

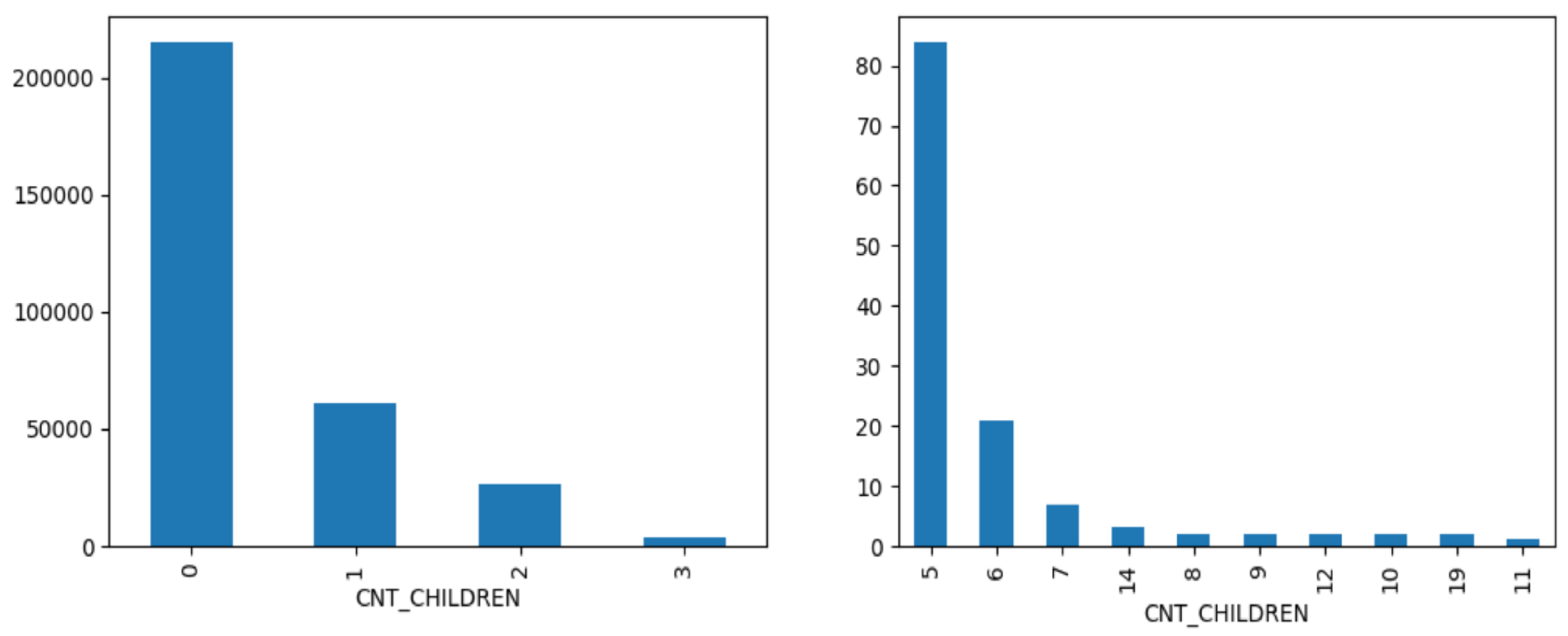
- **This trend diminishes significantly within 10-15 years, resulting in fewer candidates in this range.**

- **There is a slight increase again in loan applicants within the last 10 years (40-50 years) of work experience. These candidates may have chosen to invest in properties or start small-scale family businesses**

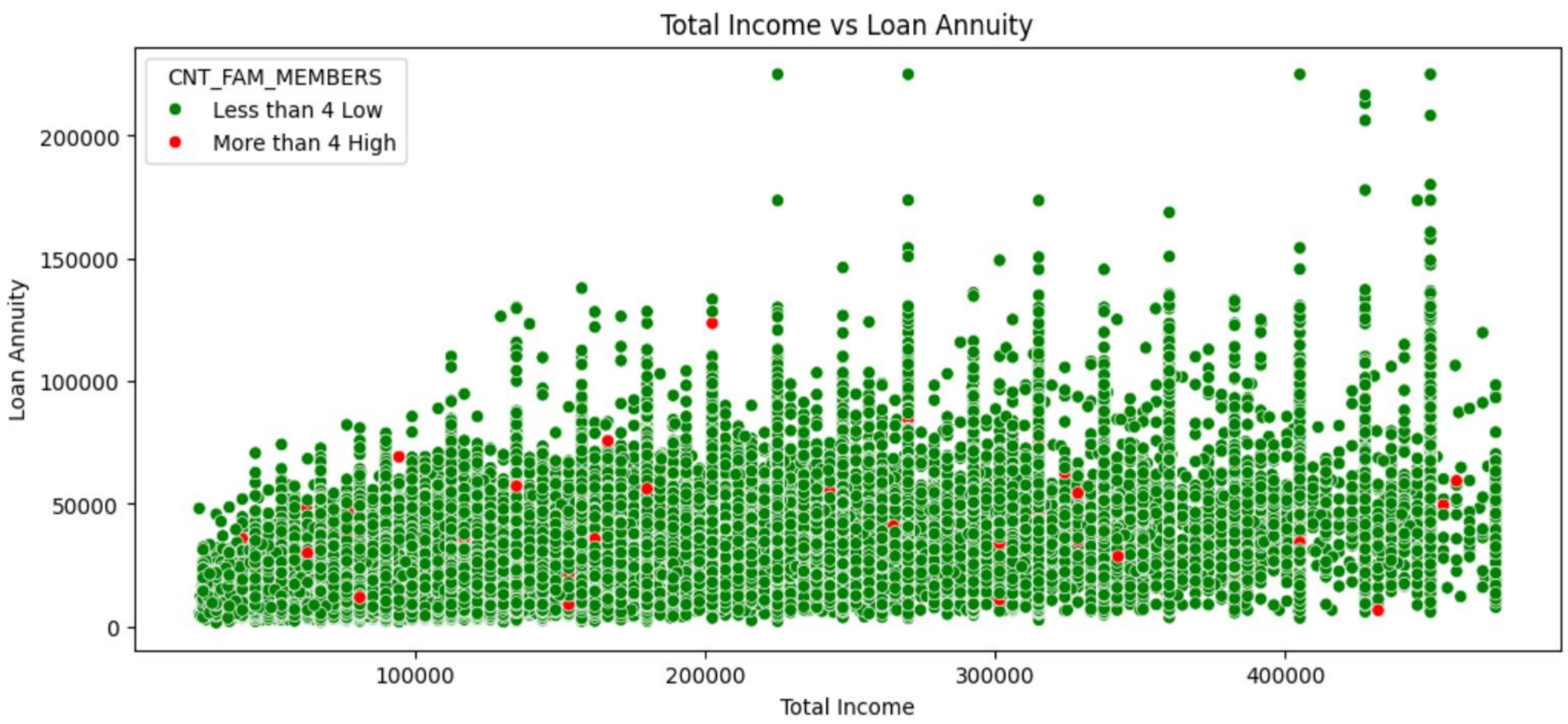
- Candidates with fewer children tend to take more loans, as expected.

- The trend in loan applications decreases as the number of children increases.

- Therefore, it can be inferred that the count of children has a negative linear relationship with the likelihood of a candidate taking a loan



# Visualizing how Loan Annuity varies with Total Income.

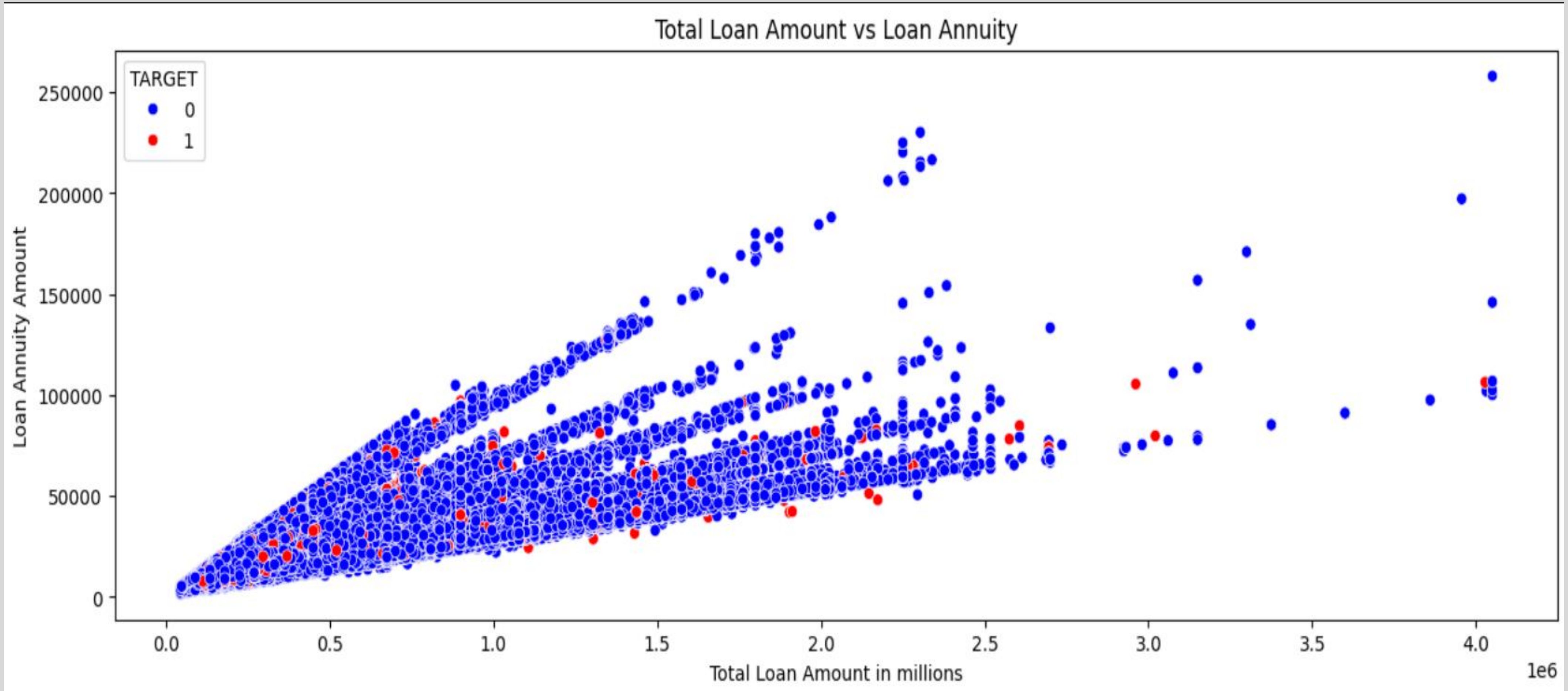


# Insights

- **The most common annuity amount appears to be under 1 lakh.**
- **With increasing income, a significant percentage of candidates seem to have annuity amounts between 1 lakh to 1.5 lakhs. Only a few candidates have annuity amounts exceeding 1.5 lakhs.**
- **Candidates with fewer family members appear more willing to pay higher annuity amounts.**
- **Conversely, candidates with larger families are likely to pay lower annuity amounts, even with higher salaries.**

# Visualizing how Total Loan Amount Approved to Candidates varies with Total Income

# Some more Insights on Total loan vs Loan Annuity with Respect to Payment issues of Candidates

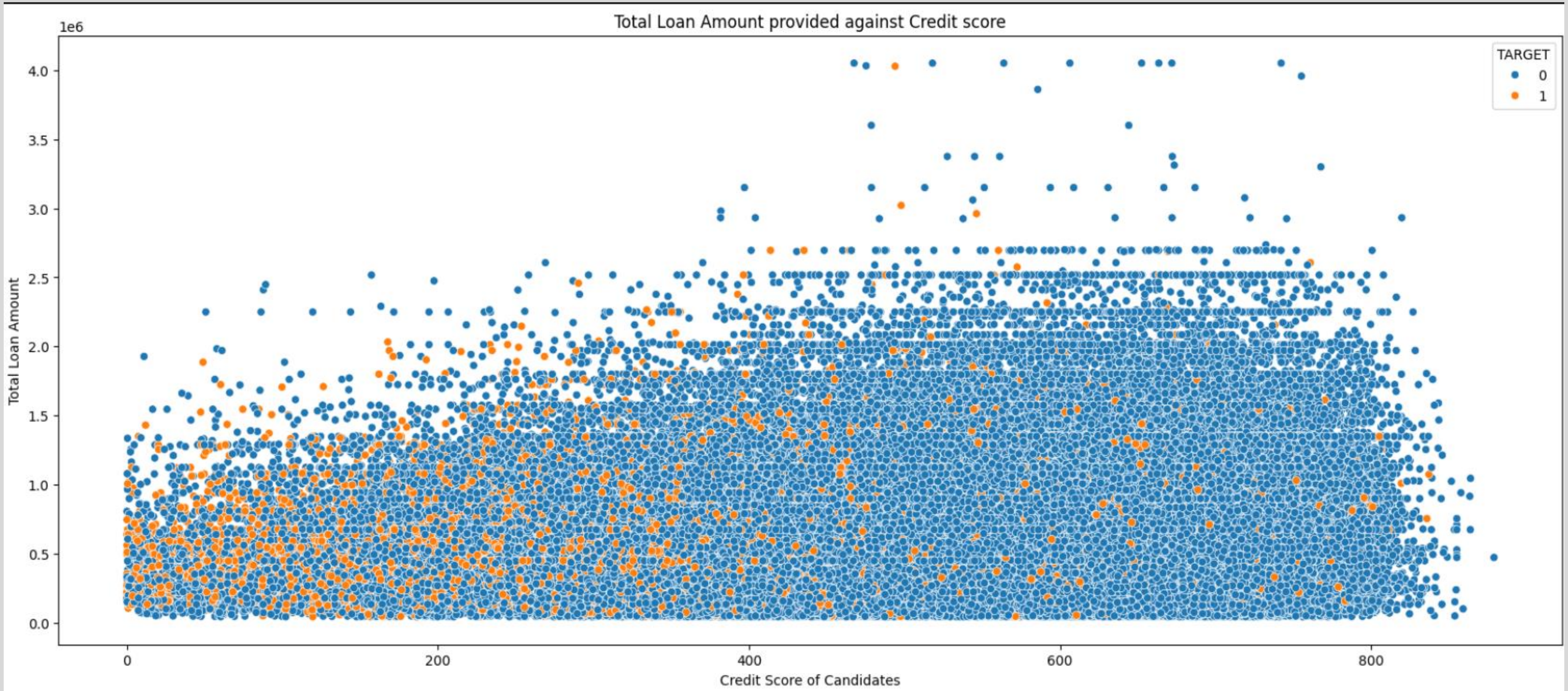


# Insights

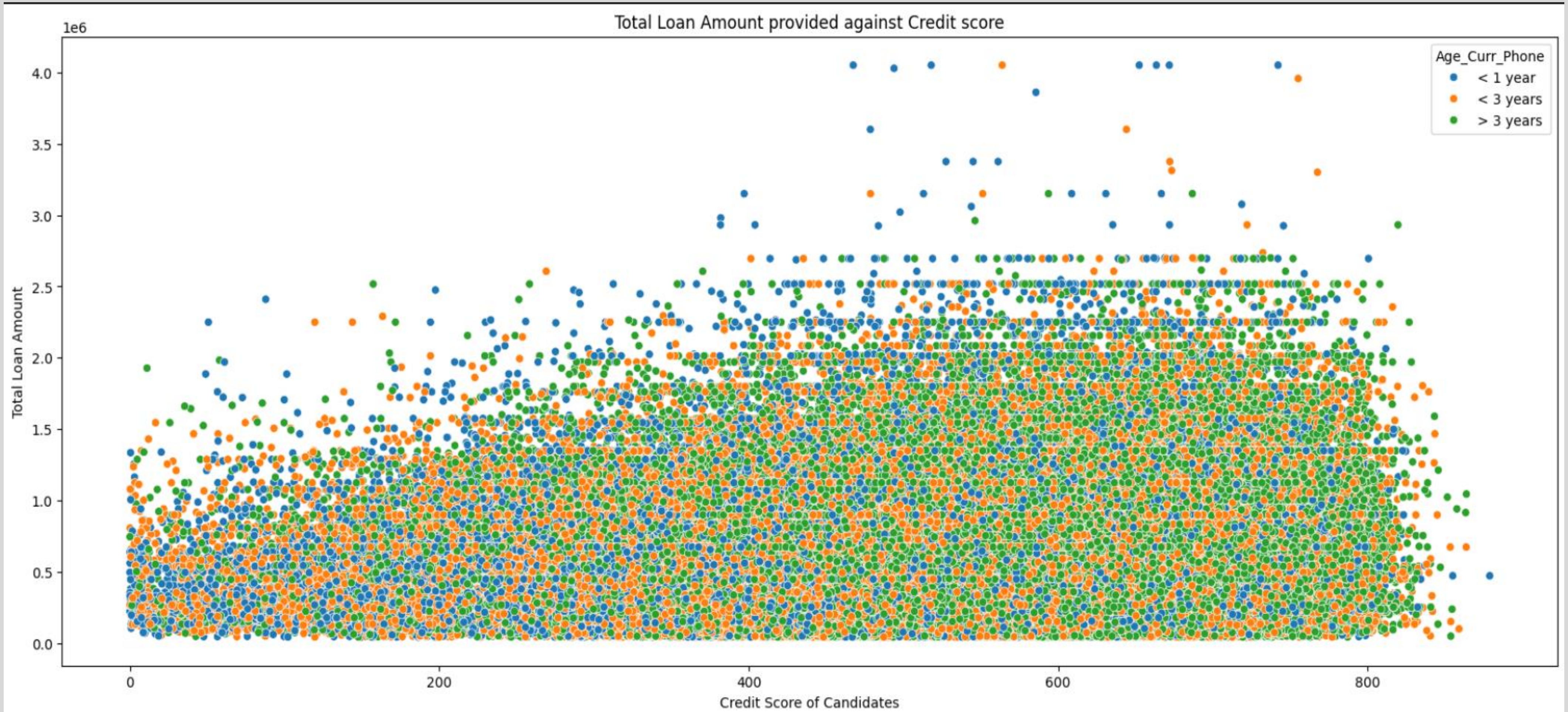
- **There is a strong positive linear relationship observed between the Amount Credit and the Amount Annuity.**

- **Interestingly, candidates with comparatively lower annuities are observed to have more payment issues.**

# Loan Amount Provided against Credit Score with respect to issues in Payments



# Total Loan provided against credit score with respect to how frequently candidates change their phone.

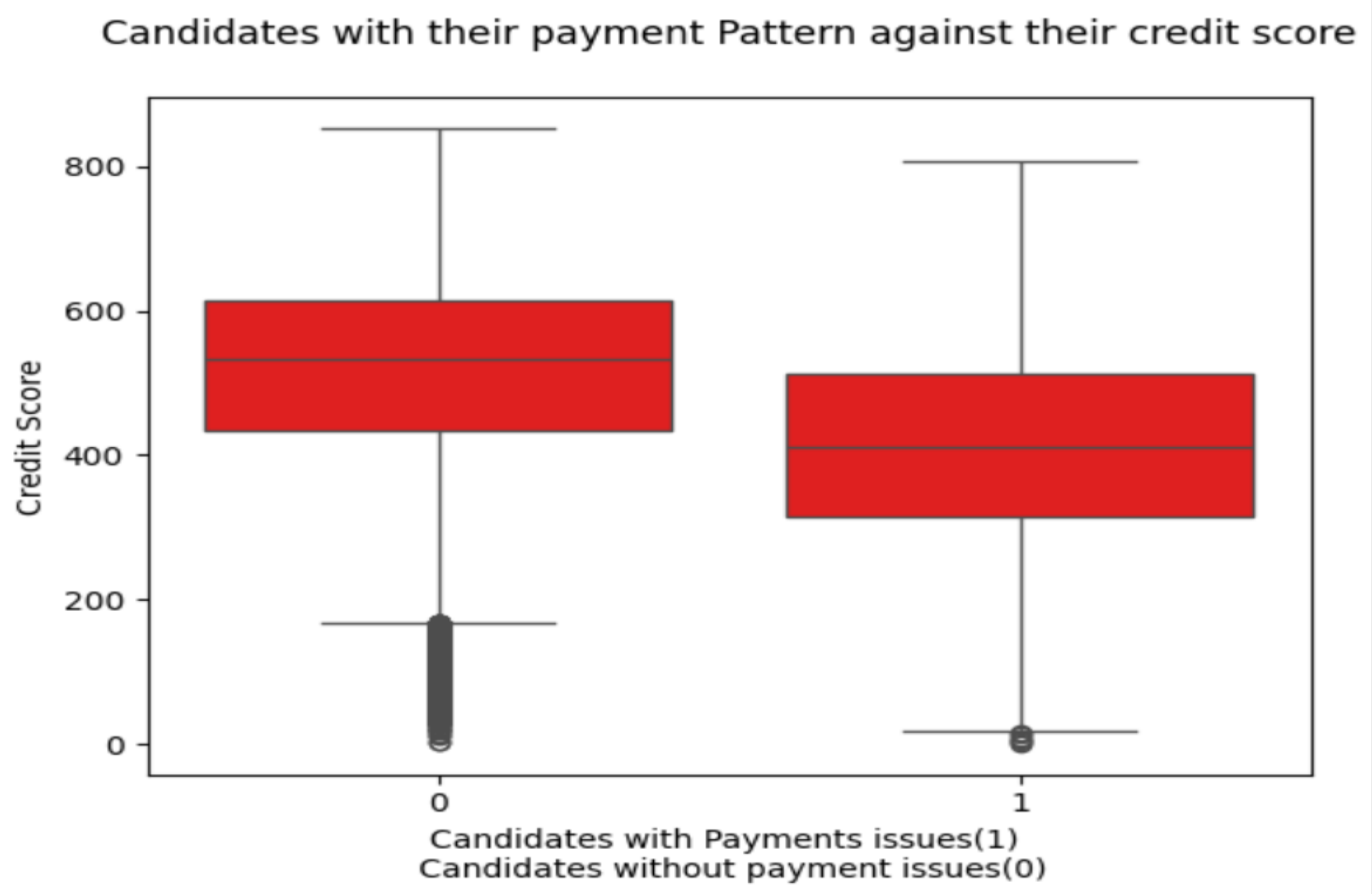


# Insights

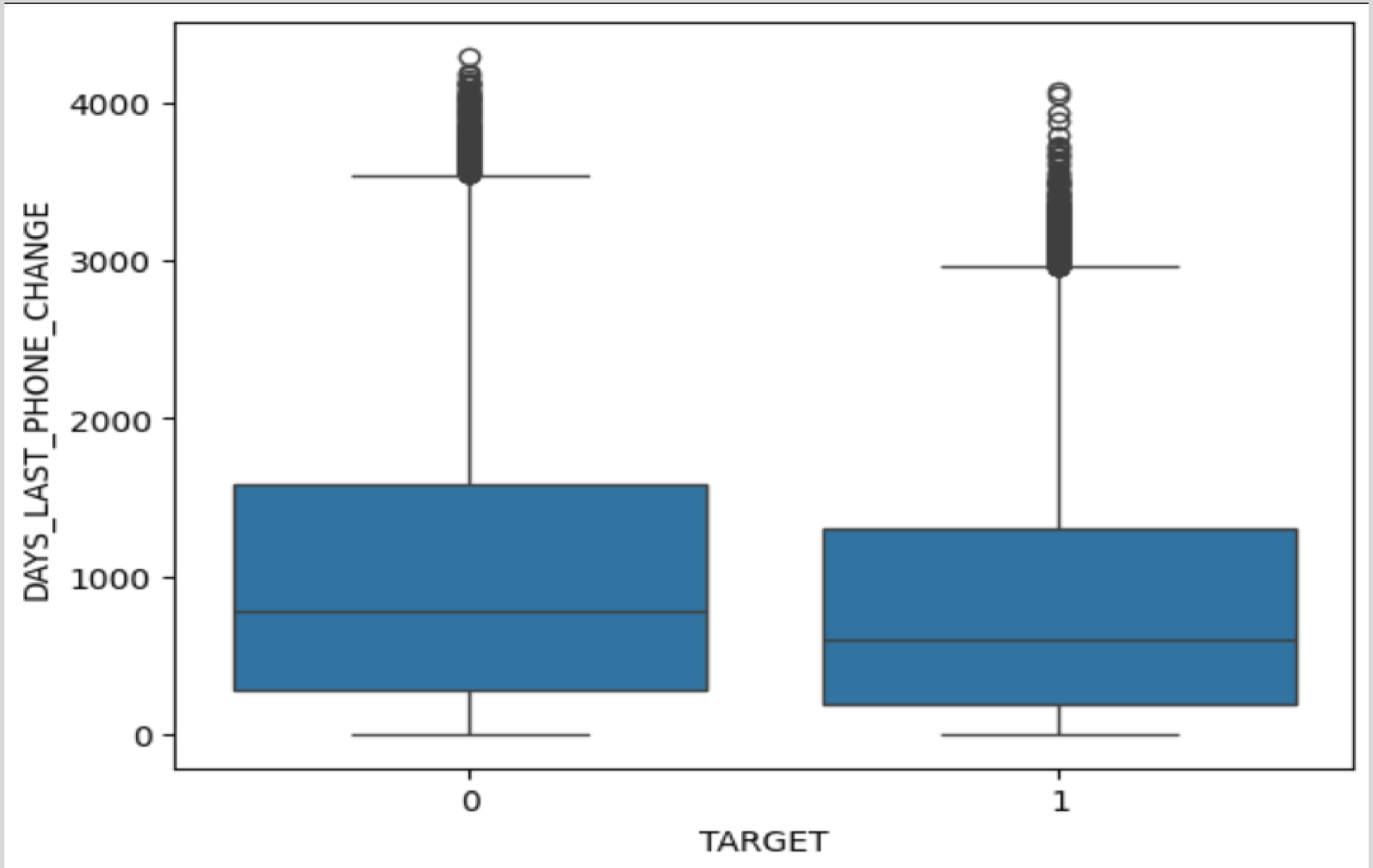
- **Most candidates with low credit scores experience payment issues. Conversely, candidates with medium to high credit scores have fewer instances of payment issues, which increases their likelihood of obtaining higher loan amounts. Further research will be conducted to substantiate this observation."**

- **Additionally, it is evident from the second Visual that candidates who changed their Mobile Phones recently, within 1-3 years, tend to have lower credit scores compared to those who have not changed their phones recently. It raises the question of whether there are valid reasons for these changes or if this was a first-time occurrence. We will analyze this pattern individually to gain deeper insights.**

To verify if payment issues are dependent on the credit scores of candidates, we will create a box plot of Payment issues against credit scores.



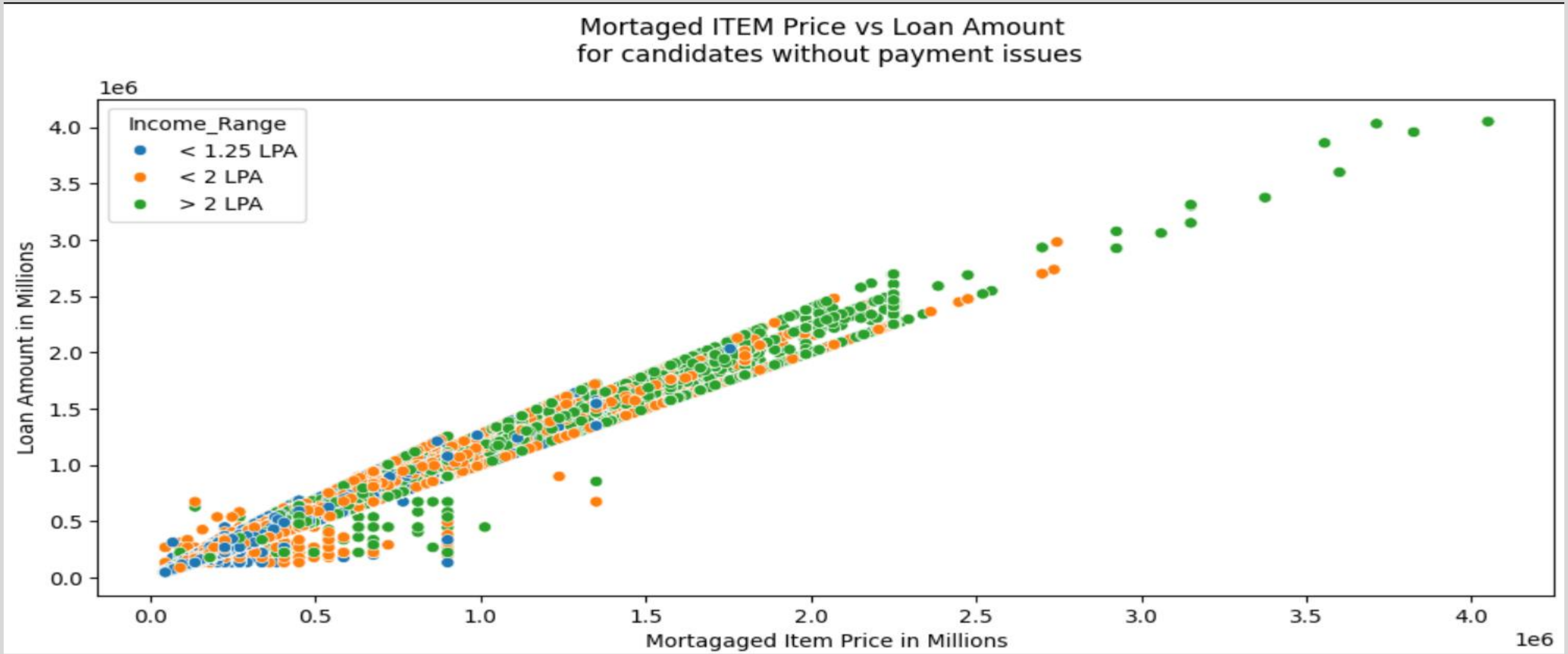
# Confirming if Frequently changing Mobile Phone Leads to Payment issues or not.



# Inferences

- We can clearly see that Candidates with Payment issues have a median credit score around 400 which is less that of candidates with no payment issues who have median credit score around 550.
- People with payment issues seems to have less intervals between current and last phone change.
- Possible reason explaining this slight variation can be that people who change their Mobile phones or other commodities frequently have more buying habits which leads to payment issues for some individuals.
- This confirms that there is dependency of payment issues with buying habits of candidates.

# Approval of Loan Amount against Price of Mortagaged Item. For Candidates without Payment Issues.



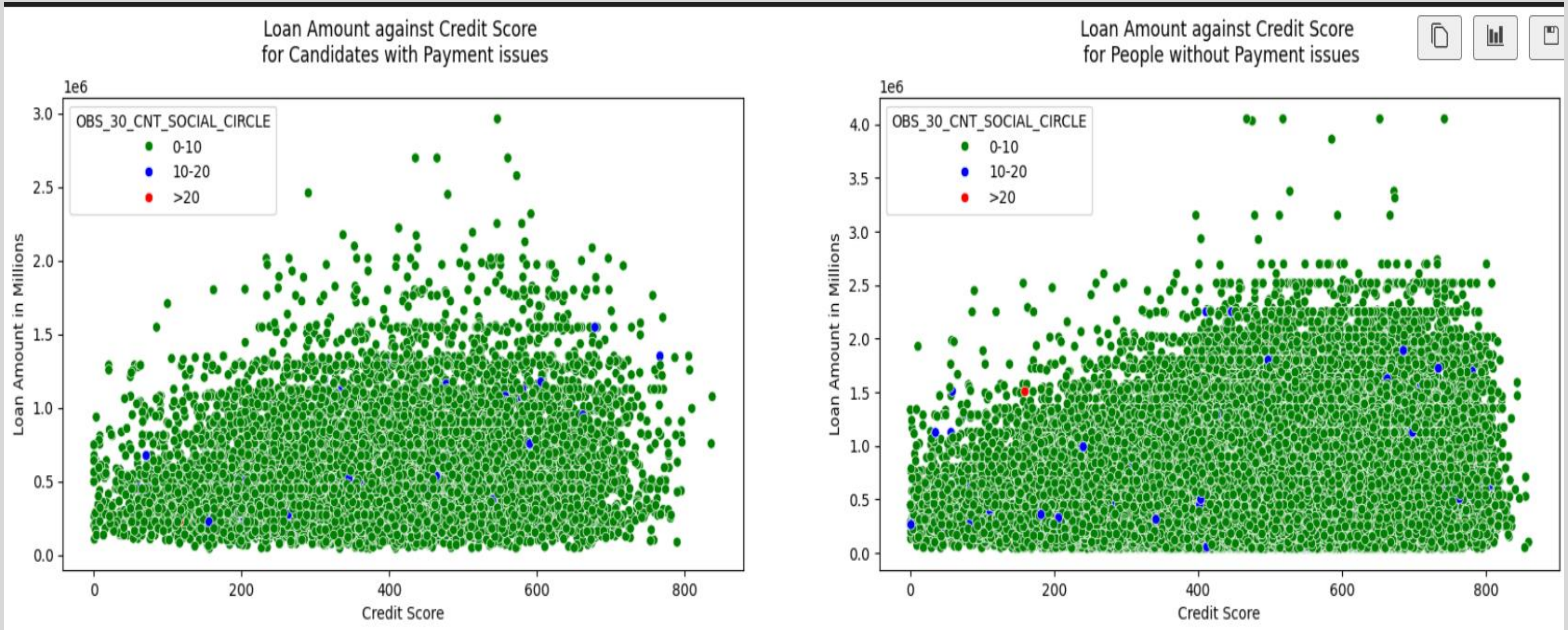
# For Candidates with Payment Issues

# Insights

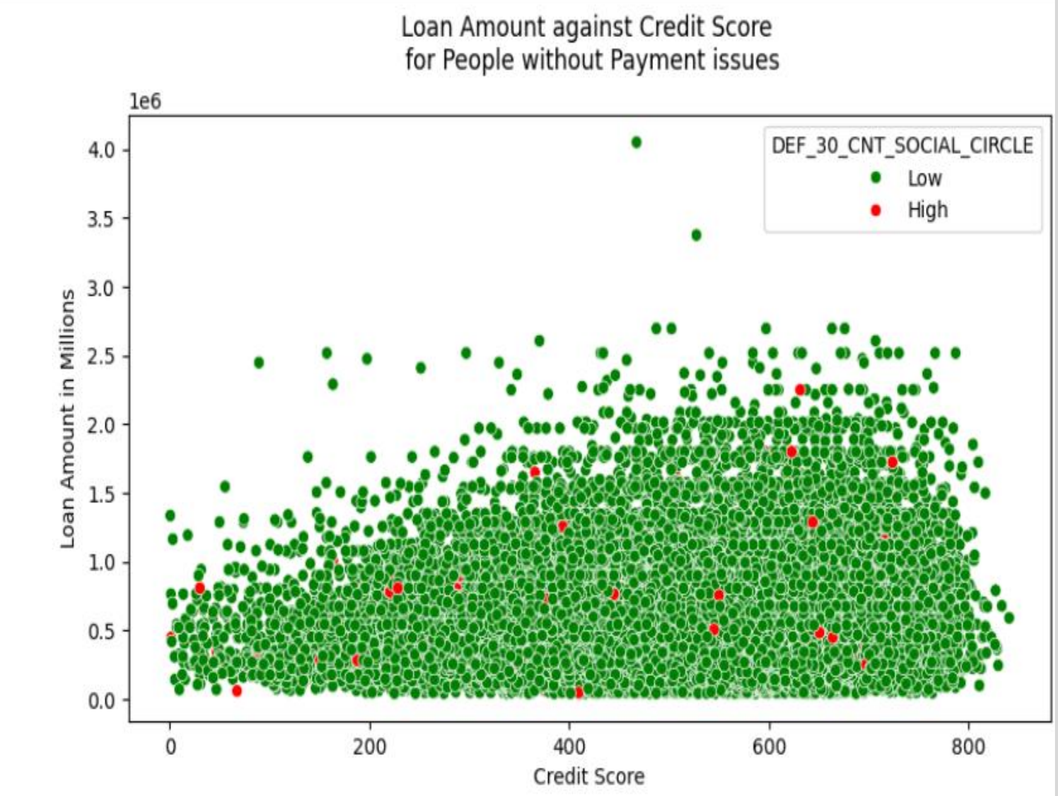
1. <u>Correlation between Asset Value and Loan Amounts:</u> There is a clear relationship where higher valued assets used as collateral tend to secure larger loan amounts. This is standard practice as lenders assess collateral value to determine borrowing limits.
2. <u>Income Influence on Asset Selection (No Payment Issues)</u>: Borrowers with higher incomes typically mortgage higher valued assets, reflecting their ability to secure larger loans. Conversely, those with lower incomes tend to mortgage assets of lower value.
3. <u>Risk Factors in Payment Issues:</u> A notable finding is that borrowers with lower incomes sometimes secure larger loans by mortgaging high valued assets. This practice may contribute to payment issues if borrowers are unable to sustain repayments commensurate with their income levels.
4. <u>Verification of Asset Value:</u> It is essential to rigorously verify the true value of mortgaged assets, particularly when borrowers with lower incomes present high valued items as collateral. This verification process ensures transparency and reduces the risk of misrepresentation or overvaluation.
5. <u>Policy Implications:</u> These observations underscore the importance of robust risk assessment frameworks and borrower education initiatives. Lenders should implement stringent policies to evaluate income-asset alignment and mitigate potential default risks associated with misaligned borrowing practices.

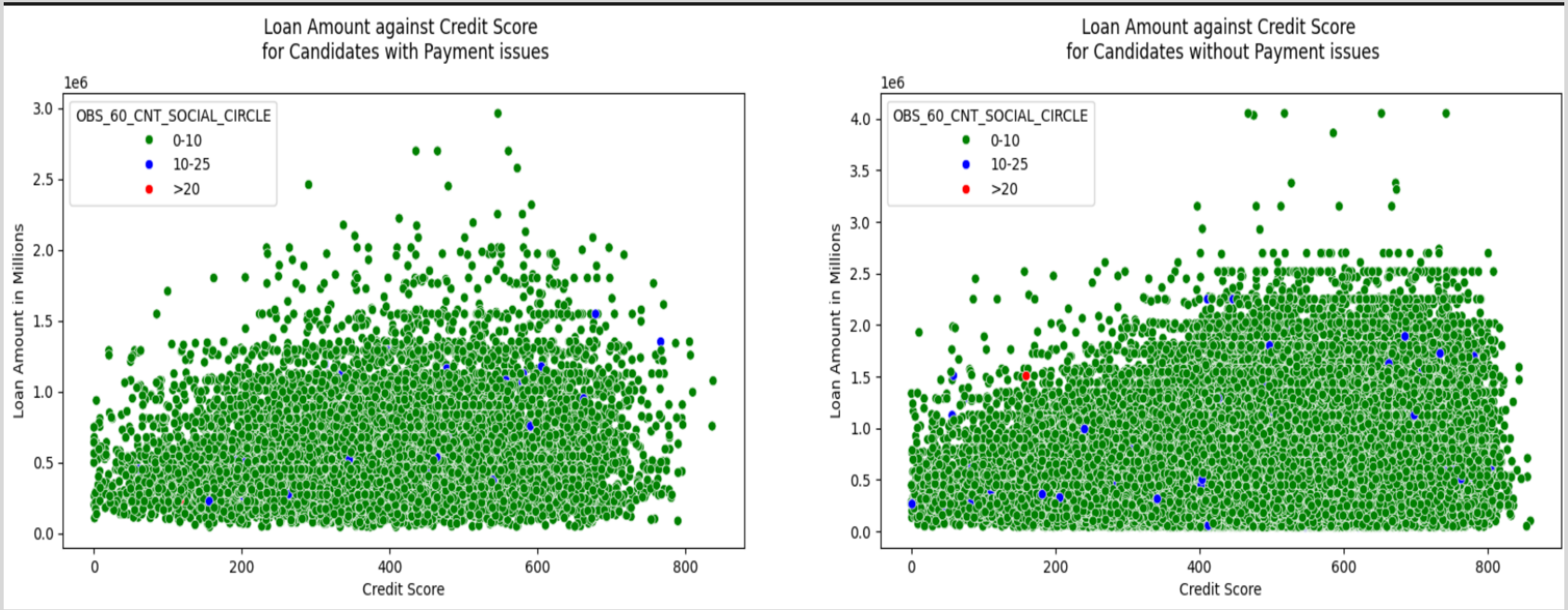
- In summary, these insights emphasize the need for thorough due diligence in asset valuation and income verification to uphold sound lending practices and mitigate financial risks for both lenders and borrowers.

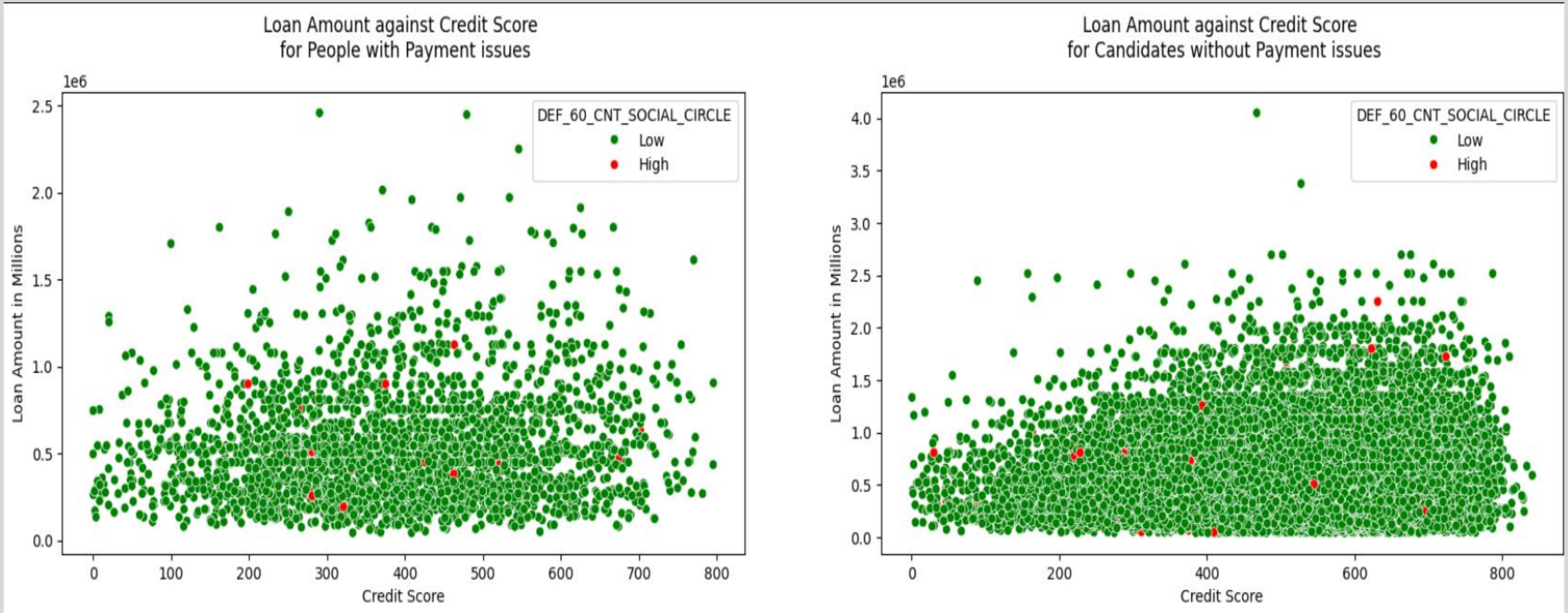
# Visualizing How Payment issues arise with respect to their social Circles and dependent on their credit scores.




Loan Amount against Credit Score
for Candidates with Payment issues
1e6
DEF_30_CNT_SOCIAL_CIRCLE
Low
High
Loan Amount in Millions
Credit Score
0.0
0.5
1.0
1.5
2.0
2.5
0
100
200
300
400
500
600
700
800


Loan Amount against Credit Score
for People without Payment issues
1e6
DEF_30_CNT_SOCIAL_CIRCLE
Low
High
Loan Amount in Millions
Credit Score
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
200
400
600
800


Loan Amount against Credit Score
for Candidates with Payment issues
1e6
OBS_60_CNT_SOCIAL_CIRCLE
0-10
10-25
>20
Loan Amount in Millions
Credit Score
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
200
400
600
800
Loan Amount against Credit Score
for Candidates without Payment issues
1e6
OBS_60_CNT_SOCIAL_CIRCLE
0-10
10-25
>20
Loan Amount in Millions
Credit Score
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
200
400
600
800


Loan Amount against Credit Score
for People with Payment issues
1e6
DEF_60_CNT_SOCIAL_CIRCLE
Low
High
Loan Amount in Millions
Credit Score
Loan Amount against Credit Score
for Candidates without Payment issues
1e6
DEF_60_CNT_SOCIAL_CIRCLE
Low
High
Loan Amount in Millions
Credit Score

# Insights

- **Impact of Social Circle on Payment Issues: Analysis reveals a notable trend among candidates experiencing payment issues: a higher prevalence of defaults within their social circles, particularly within 30-60 days. In contrast, candidates without payment issues exhibit lower default rates among their acquaintances.**

- **Credit Score and Default Patterns: Individuals who defaulted or exhibited default tendencies typically demonstrated lower credit scores compared to those who maintained consistent payments.**

- **Social Circle Influence on Credit Scores: There is a discernible correlation between the frequency of observed and actual defaults within a candidate's social circle and their credit scores. Candidates associated with social circles characterized by frequent defaults tend to have lower credit scores.**

- **Impact on Payment Patterns: These findings suggest that the social environment in which a candidate resides can significantly influence their creditworthiness and subsequent payment behaviors. Factors such as peer influence and shared financial practices within social circles may indirectly impact individual credit scores and payment reliability.**