GRAMENER CASE STUDY-
EDA CASE STUDY SUBMISSION

Group Name:
1. Pritha Banerjee
2. Priyanka Kapoor
3. Kanchan Kumari
4. Munish Bansal

GRAMENER CASE STUDY

- **Introduction :**

Risk analysis for a consumer Finance Company specialising in lending various types of loans to urban customers: Personal Loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through its fast online interface.

- **Objective :**

To understand how **consumer attributes** and **loan attributes** influence the tendency of default.

Risk assessment of new loan applicants

- **Strategy:**

- Identification of Loan Applicant traits that tend to 'default' paying back
- Understand the 'Driving Factors' or 'Driver Variables' behind Loan Default i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.
- There are two type of risks:

  If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company

  If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company.

GRAMENER CASE STUDY

- **Strategy: (Contd..)**

When a person applies for a loan, and the loan gets accepted by the company, there are 3 possible scenarios described below:

Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

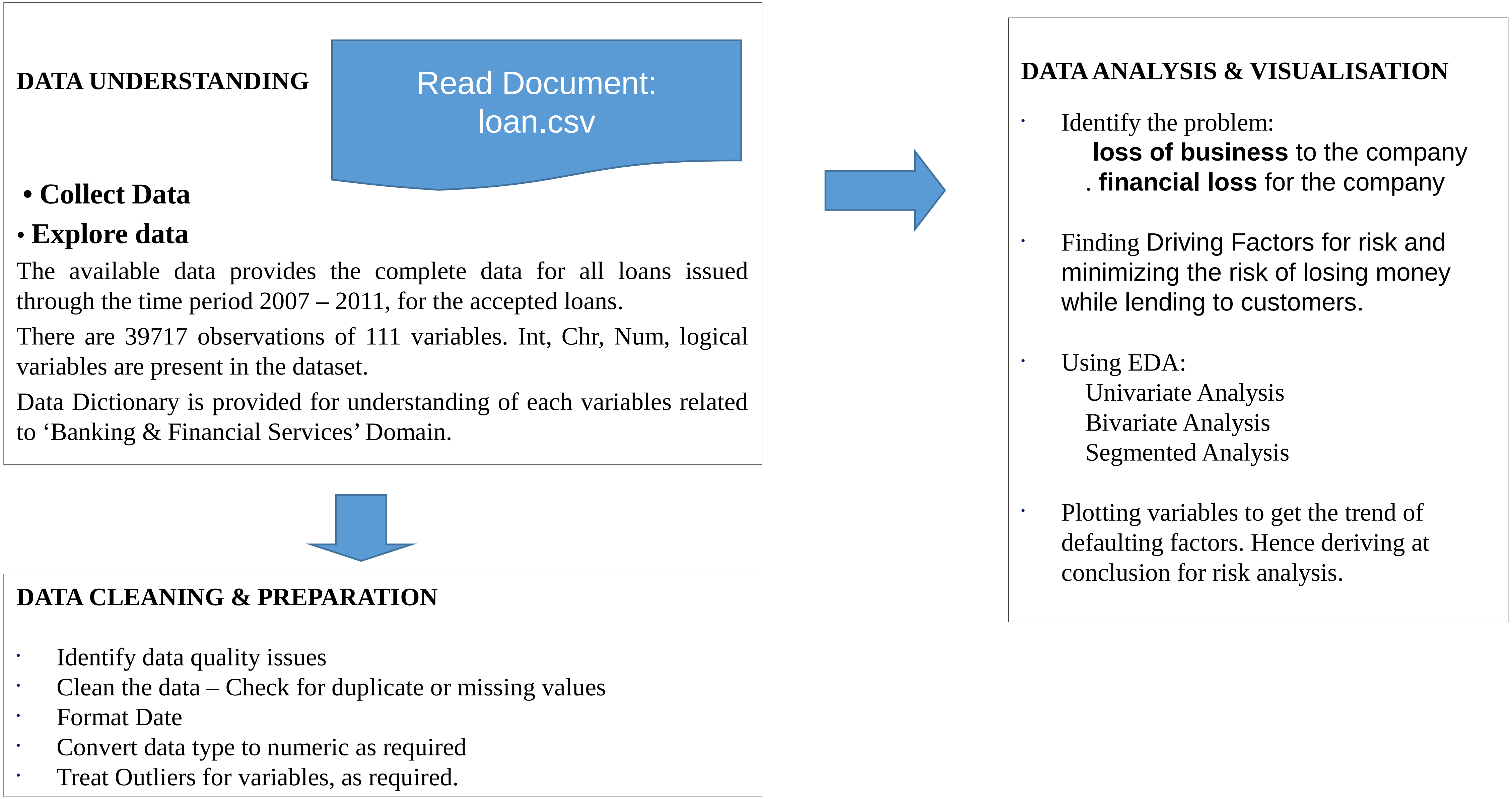
Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

Here, risk analysis has been done on Fully paid and Charged Off Loan status to derive insights.

- **Data Understanding:**

- The **loan** Dataset contains complete data for all loans issued through the time period 2007 – 2011, for the accepted loans.
- Regarding the rejected loans (because the candidate did not meet their requirements etc.), there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset).
- There are many missing and unnecessary data available in the dataset. Hence, data cleaning is required before performing data analysis.

GRAMENER CASE STUDY



DATA UNDERSTANDING

Read Document: loan.csv

- **Collect Data**
- **Explore data**

The available data provides the complete data for all loans issued through the time period 2007 – 2011, for the accepted loans.

There are 39717 observations of 111 variables. Int, Chr, Num, logical variables are present in the dataset.

Data Dictionary is provided for understanding of each variables related to ‘Banking & Financial Services’ Domain.

DATA CLEANING & PREPARATION

- Identify data quality issues
- Clean the data – Check for duplicate or missing values
- Format Date
- Convert data type to numeric as required
- Treat Outliers for variables, as required.

DATA ANALYSIS & VISUALISATION

- Identify the problem:
  loss of business to the company
  . **financial loss** for the company
- Finding Driving Factors for risk and minimizing the risk of losing money while lending to customers.
- Using EDA:
  Univariate Analysis
  Bivariate Analysis
  Segmented Analysis
- Plotting variables to get the trend of defaulting factors. Hence deriving at conclusion for risk analysis.

GRAMENER CASE STUDY

- **Data Cleaning and Manipulation**

1. Check for Duplicate values
2. Cleaning of Single Value Columns i.e. columns where count of unique values = 1: 60 Columns get deleted.
3. Deleting Further columns not requird for data analysis:

   -member_id, chargeoff_within_12_mths,desc,url,emp_title, tax_liens, title.

4. Checking For Missing Values: Deleting Columns with more than 50% missing Values
5. Formatting Data: Checking for Chr type of variables and converting them to Numbers as required for Analysis

   -term, int_rate and revol_util

   Formatting dates

6. Converting variables to factors:

   -term, Grade, sub grade, emp_length, home_ownership, verification_status, purpose, loan status

7. Considering loan_status which is not "CURRENT“: Keeping ‘Charged off’ & ‘Fully Paid’ loan status for further analysis

8. Treating Outliers in annual income of applicant.

Our Data is ready for analysis.

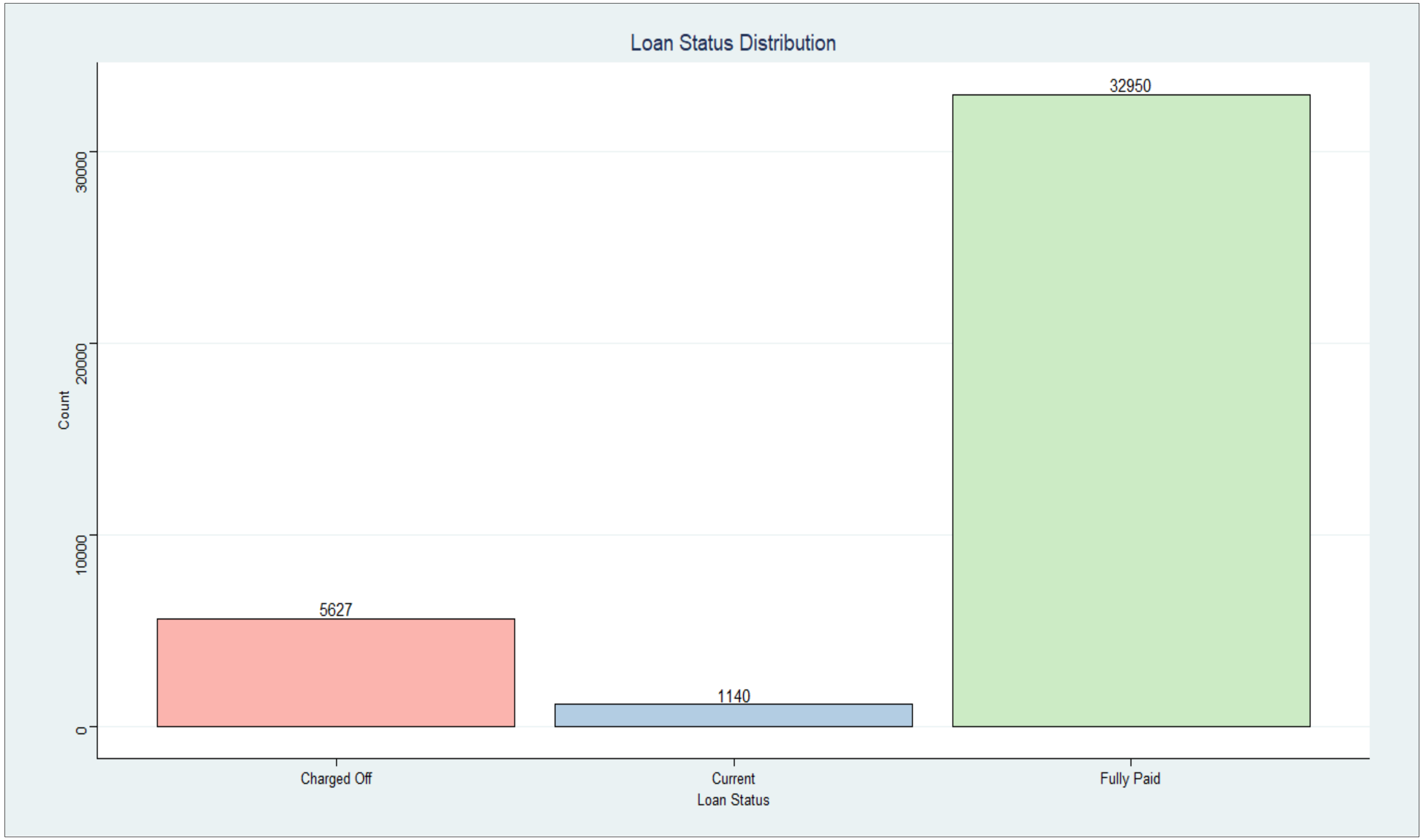
Data Analysis – Uni-variate Analysis

Plot -1 : Loan Status Distribution

There are 3 Status of Loans

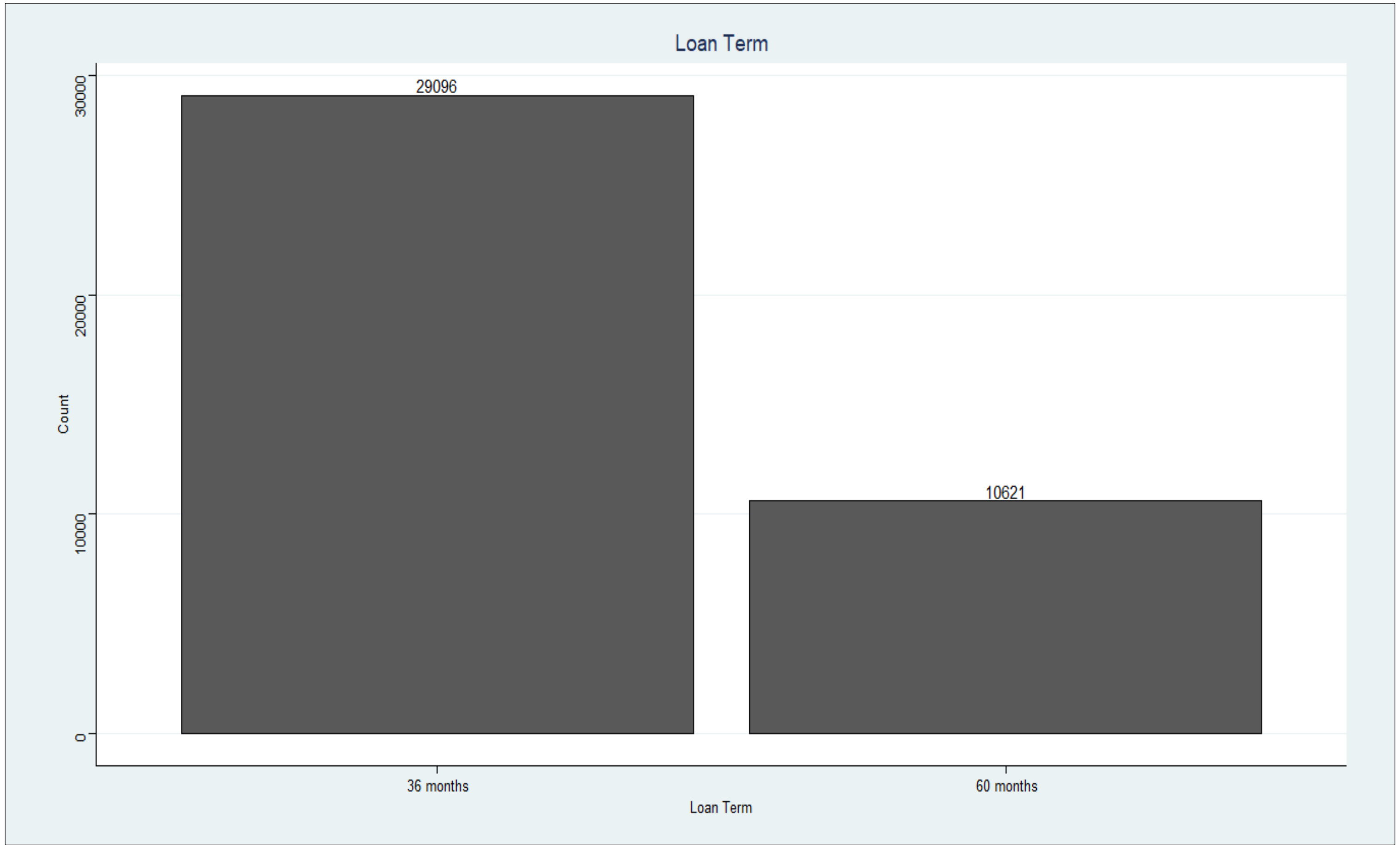
- Charged Off – These are the number of loans that were defaulted (5627)
- Current – These are the loans that are in progress (1140)
- Fully Paid – These are the loans that have been paid completely (32950)

- Since we need to find the factors that for risk analysis we will be considering only Charged Off and Fully Paid loans



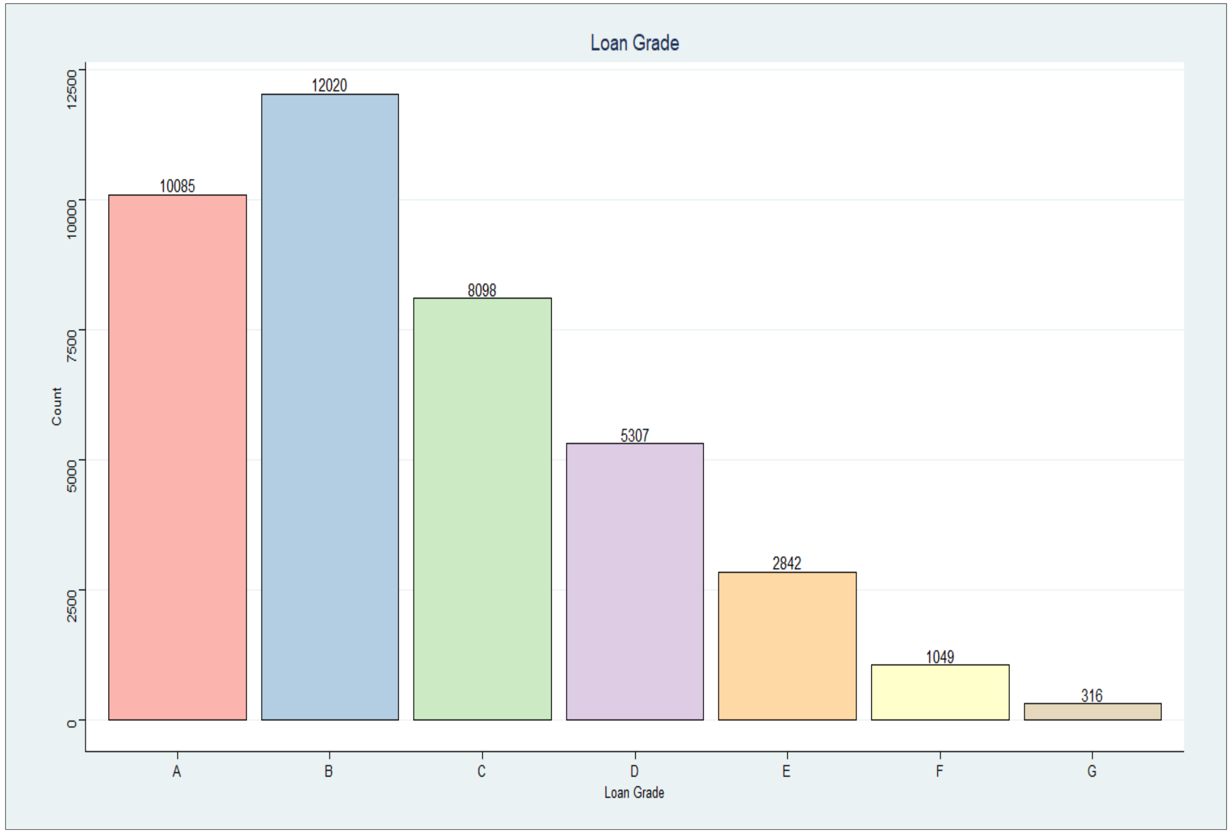
Plot -2 : Identifying the Terms for which loans are given

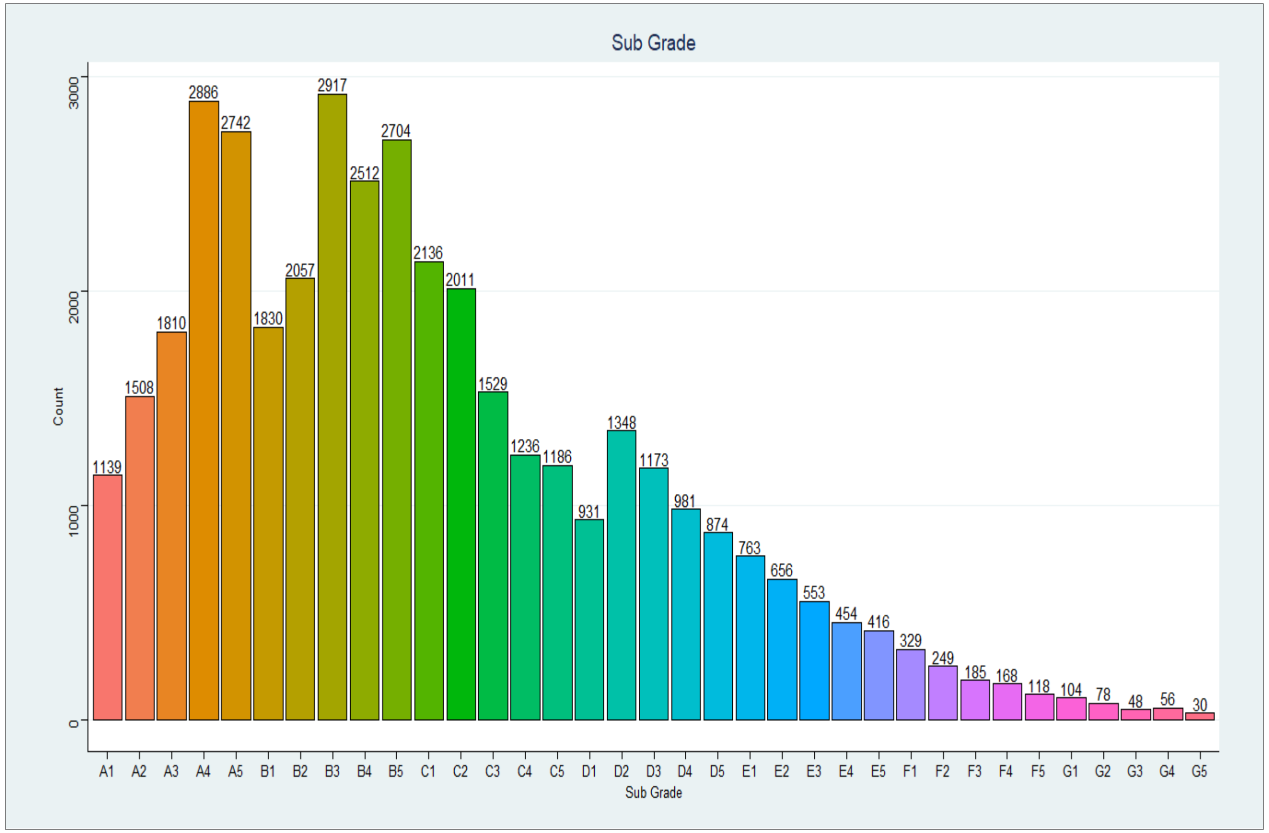
- Only 60 month & 36 month term loans are available, of which maximum is for 36 months term.
- 36 Months - 29096
- 60 Months – 10621
- Hence, it can be concluded that maximum applicants apply loans for short term repayment.



GRAMENER CASE STUDY

Plot -3,4 : Identifying different loan grades and sub-grades

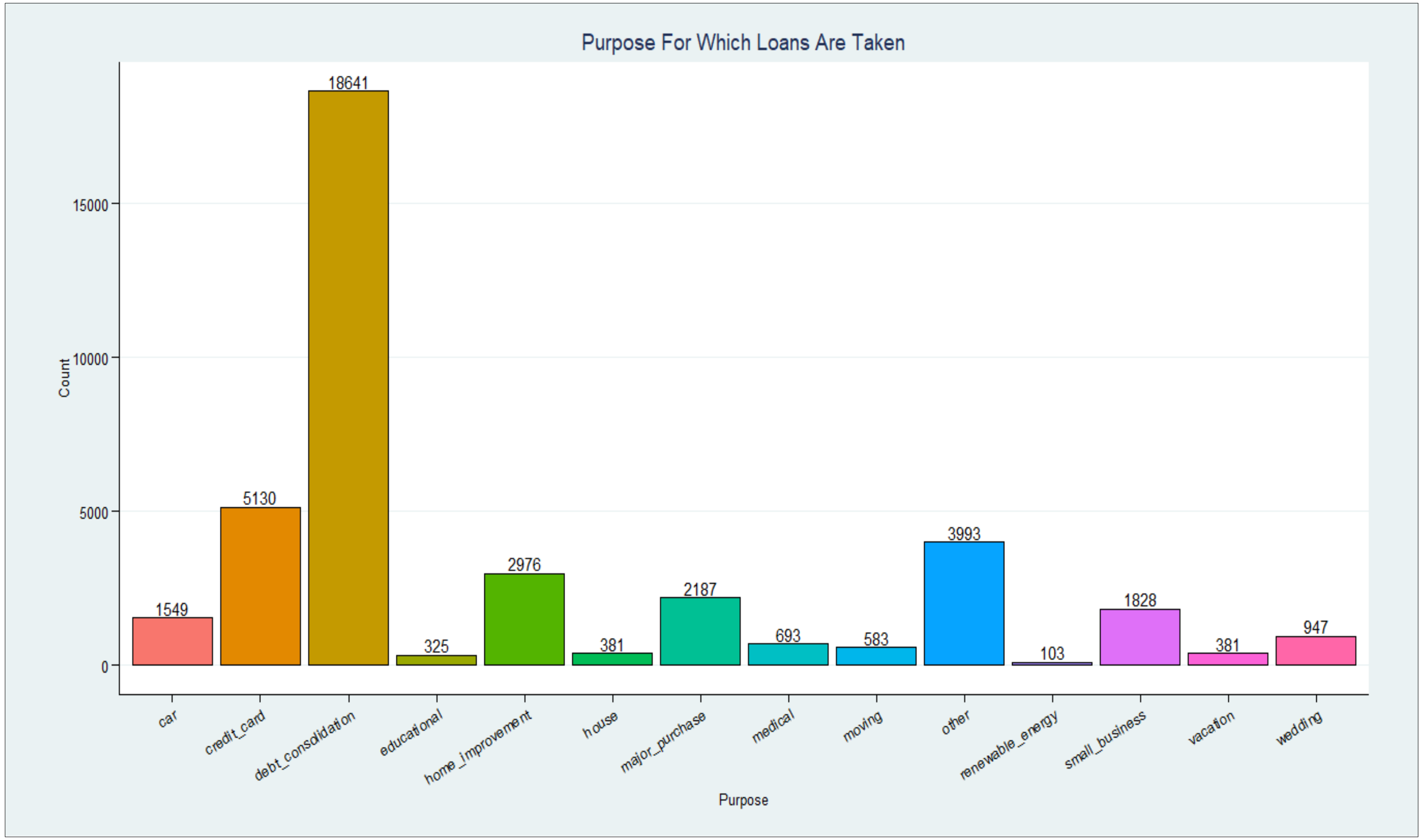




- 7 grades are there (A-G) for most loans, grades A & B are assigned whereas G grade is the minimum
- 35 sub-grades are there: 5 for each grades.
- Inference: Maximum loans are assigned to sub grades between A4-C1

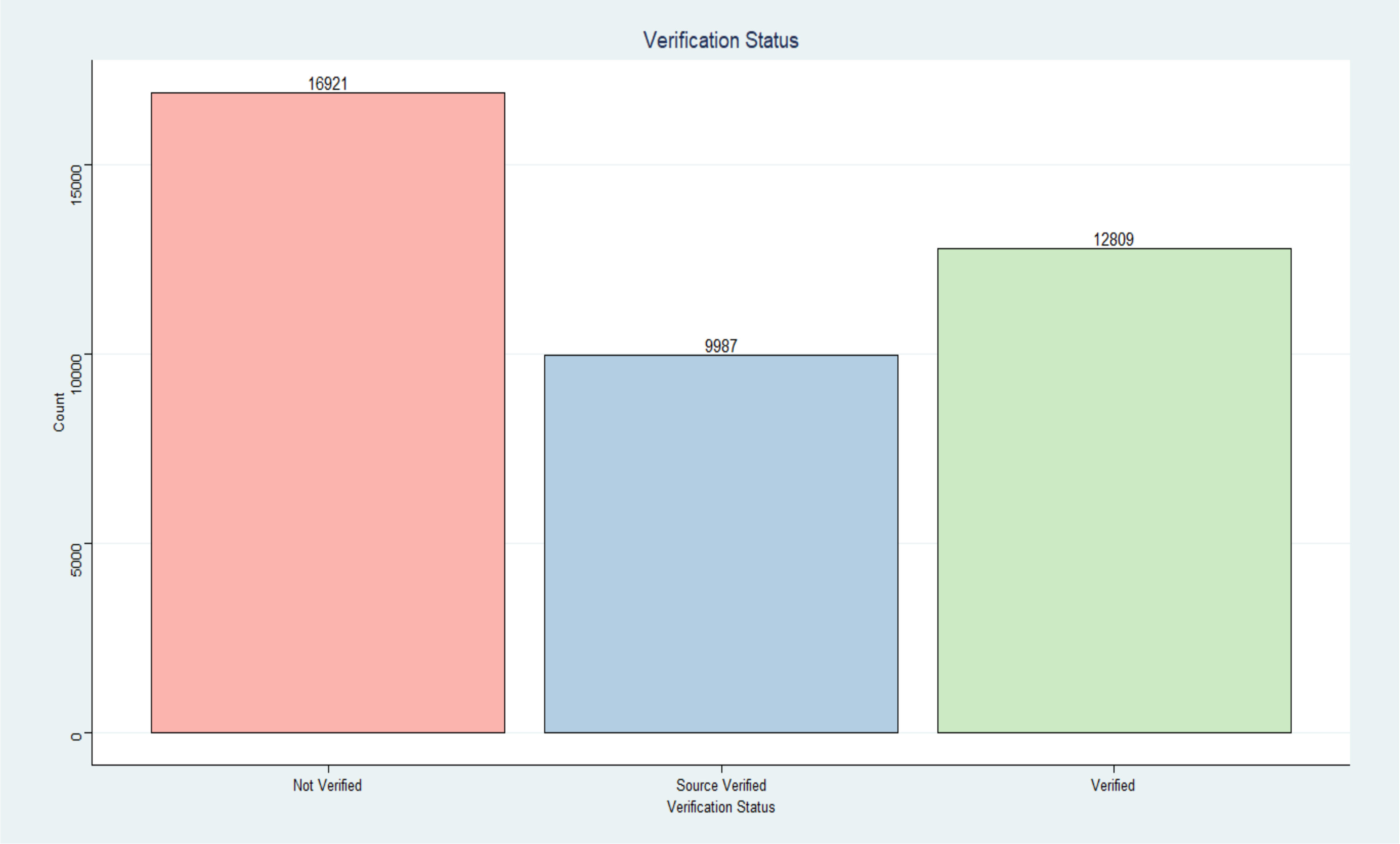
Plot -5 : Identifying the purpose for which most loans are taken

- The graph shows that there are a total of 14 different purposes for which loans are taken.
- However, for **debt consolidation** purpose most of the loans are taken i.e. 18055.
- This should be investigated further to understand if most customers default from this category or some other category.



Plot -6 : Different Verification Status amongst loans

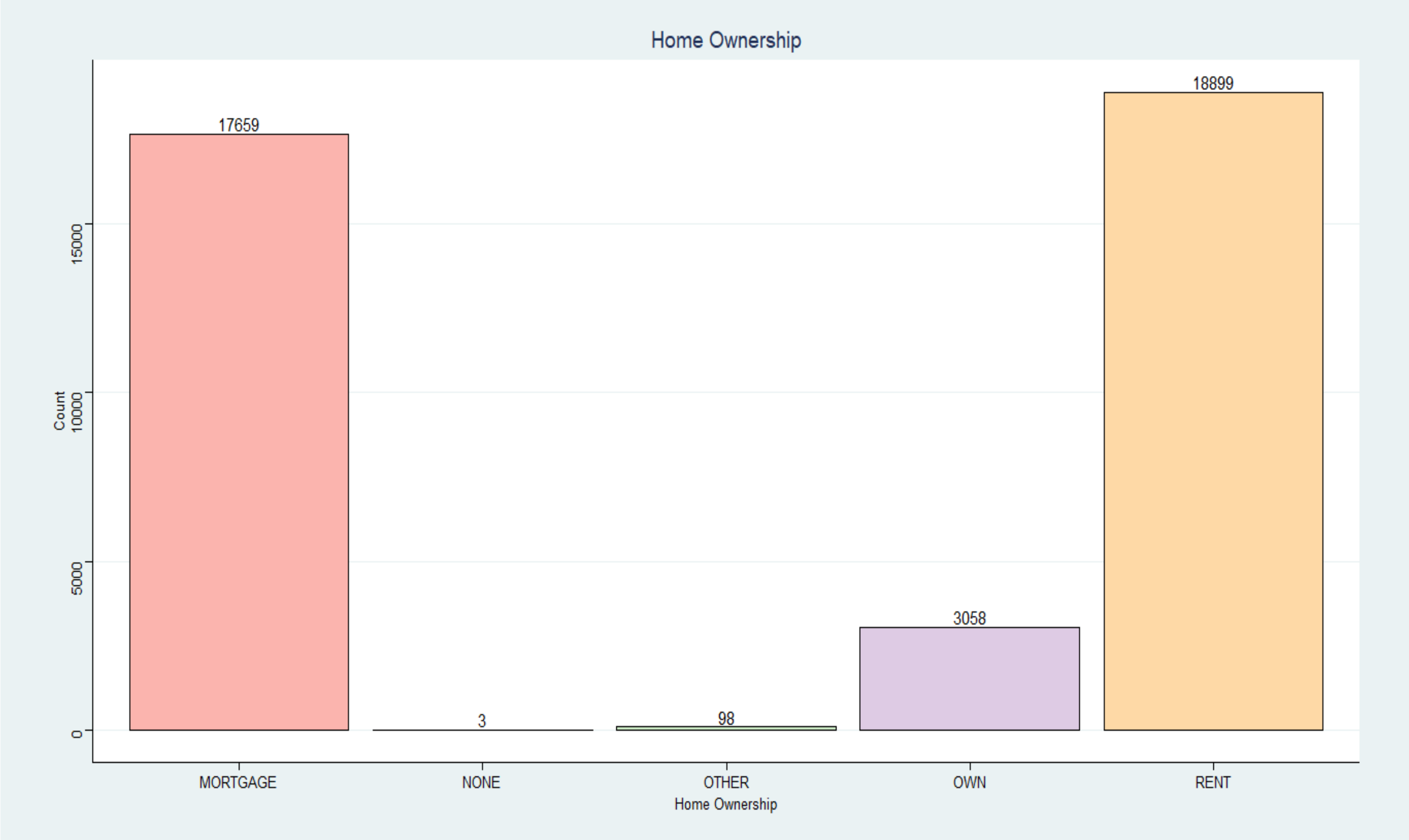
- We find that any loan will have one of the three statuses- verified, not verified and source verified
- Most of the loan are in “Not Verified” status.
- We will be inspecting them for further analysis.



Plot -7 : Identifying the home ownership status of borrowers

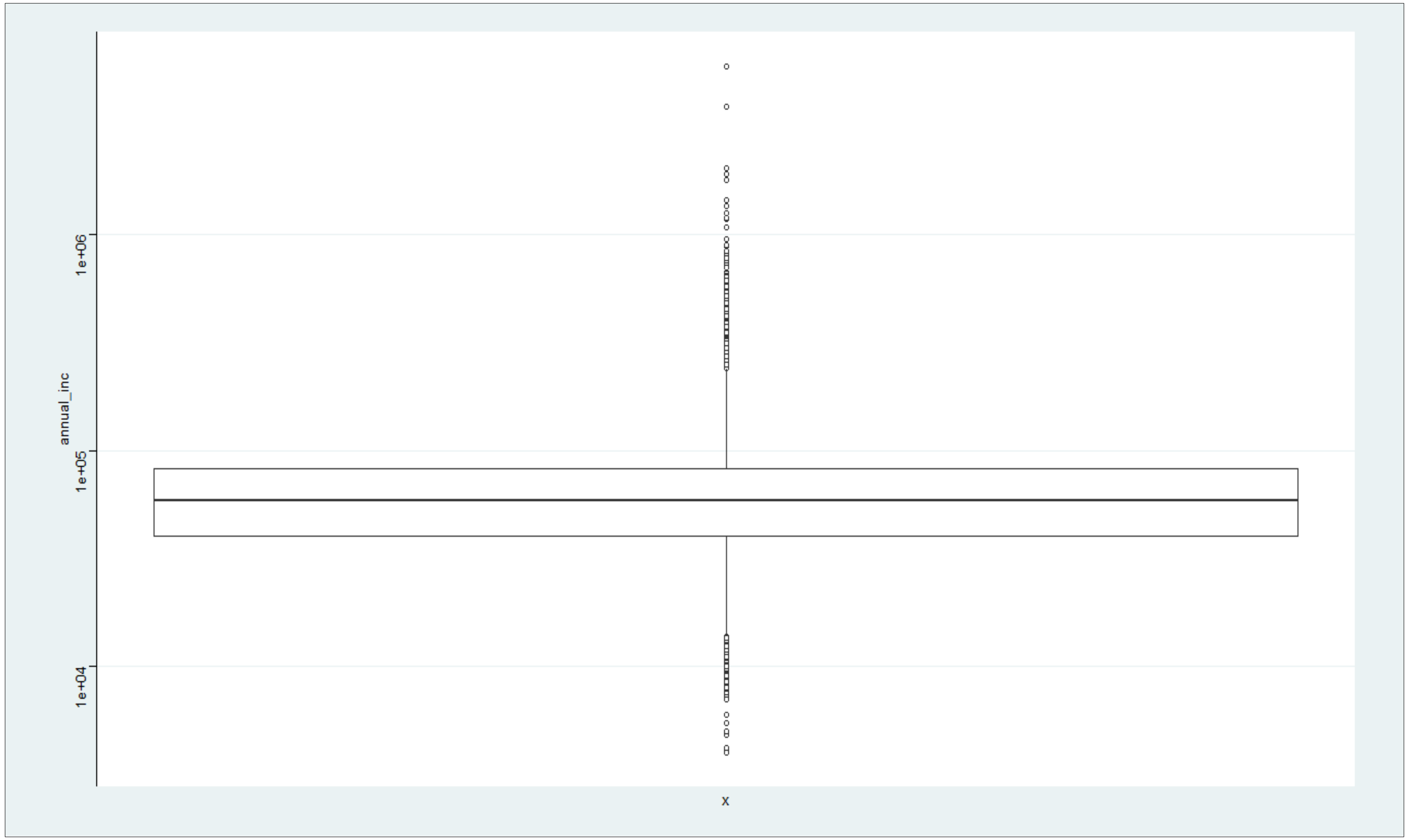
- The plot shows that every borrower has one of these 5 status while filing for the loan application

- Most of the borrowers live in rented houses. Also, borrowers living in mortgaged property is significantly higher.



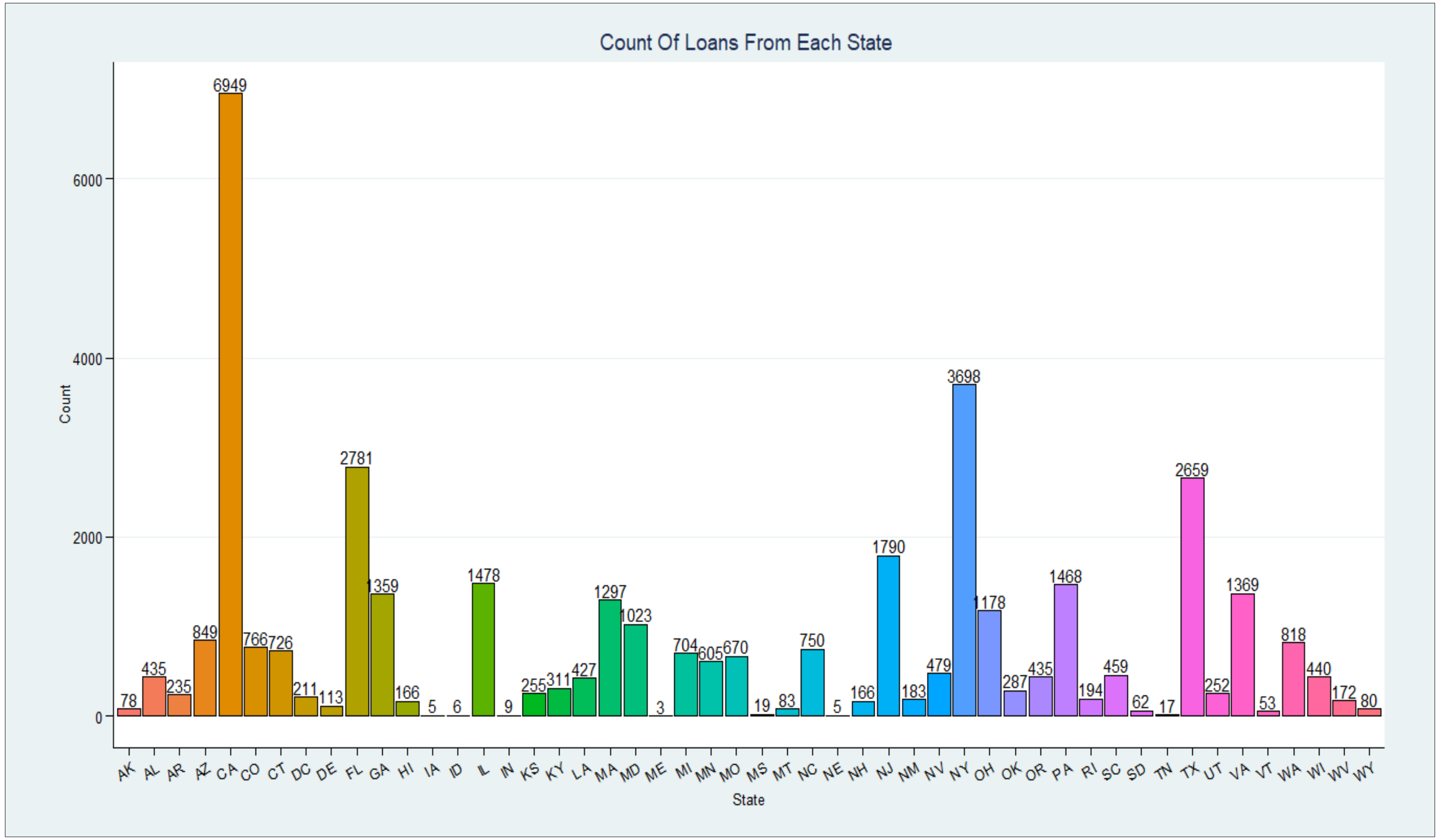
Plot -8 : Annual Income distribution

- The plot shows that there are a lot of outliers in the data. We will have to treat the outliers.
- So, removing the values that are greater than 1.5*IQR (Inter Quartile Range).



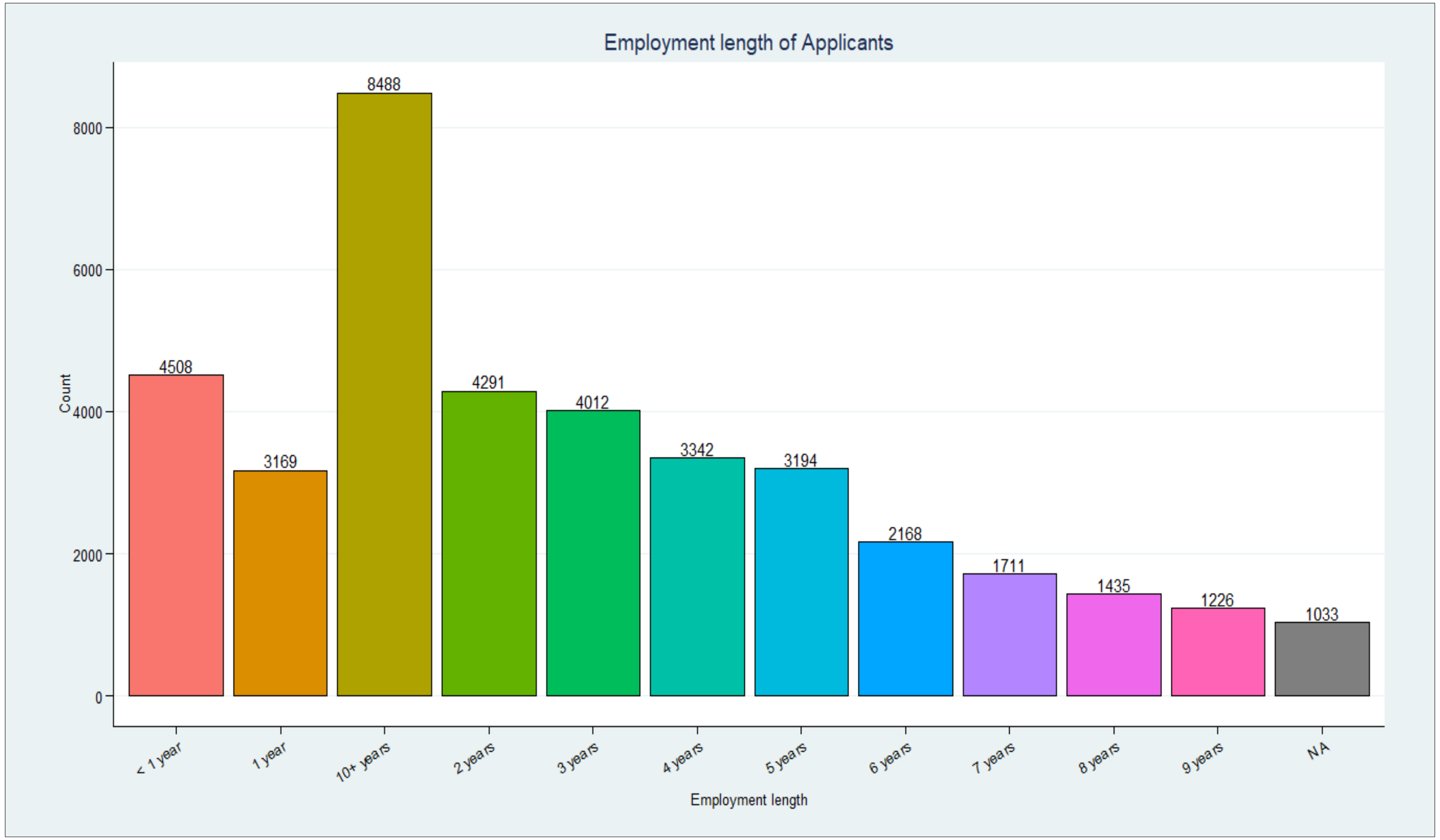
Plot -9 : Identifying the state in which most loans are taken

- The plot shows that the maximum loan applicants are from the state of California (CA).
- CA - 6949



Plot -10 : Employment Length of Applicants

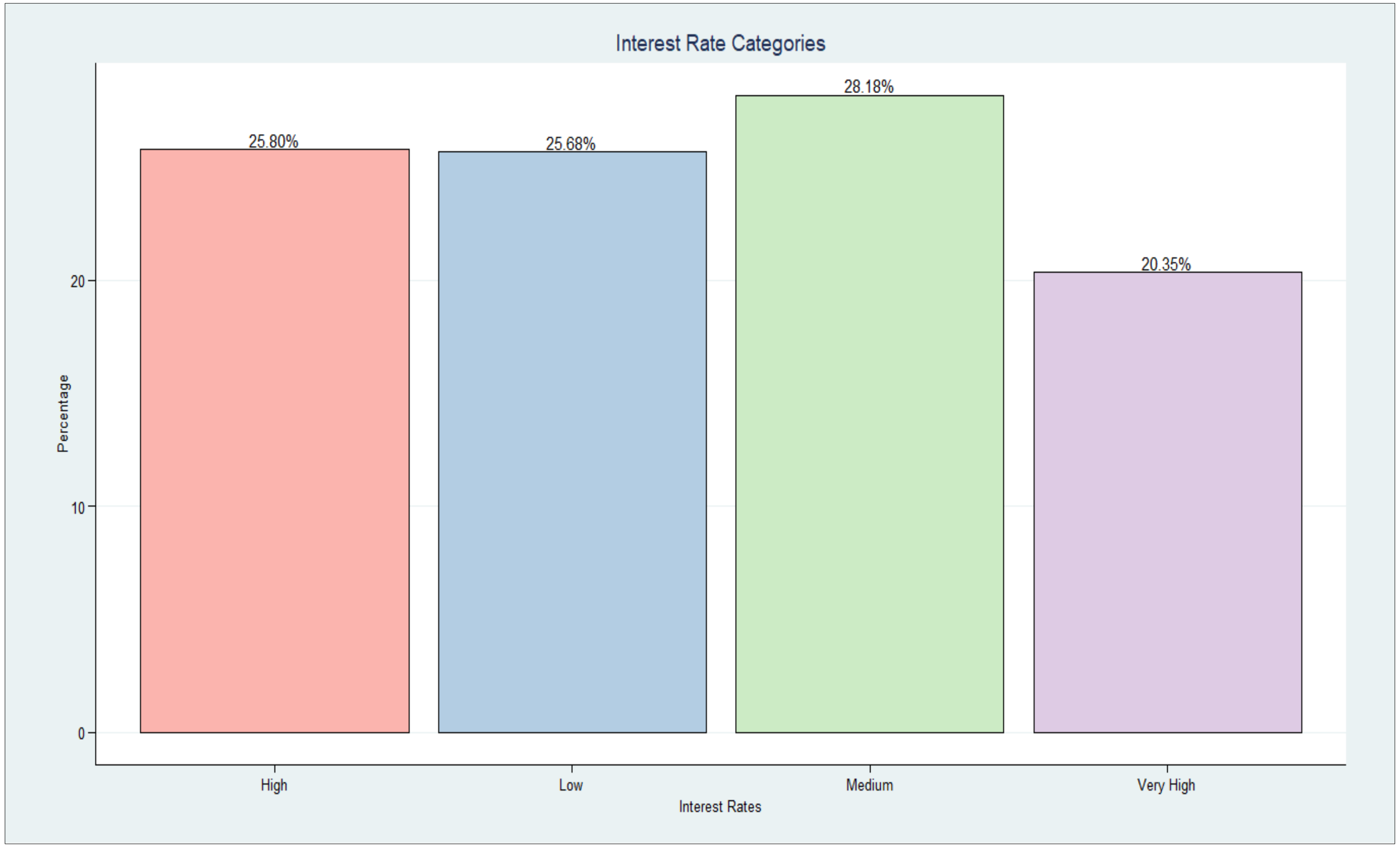
- The plot shows that the maximum loan applicants have experience of employment of 10+ years.
- Also, we find that for 1033 applicants this data is missing.



Data Analysis – Segmented Uni-variate Analysis

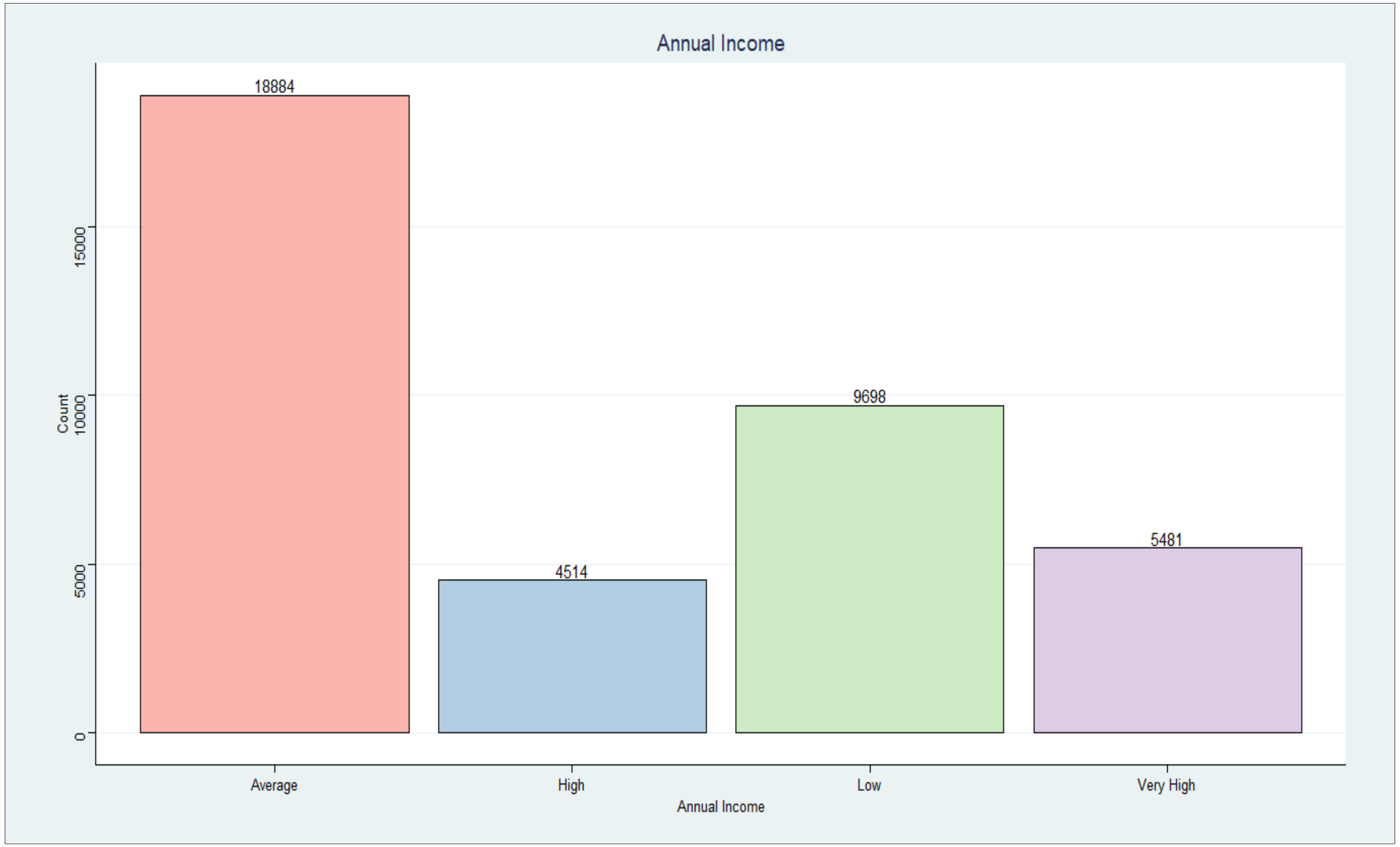
Plot -11 : Interest Rate Distribution – Category wise

- Since the interest rates lies between 5.42 to 24.59, segmenting it into bins in order to understand and analyze it clearly. (These bins are open to understanding of the analysts).
- Categorizing interest rate between 5.42 and 9.00 as low, 9.00 and 12.00 as medium, 12.00 and 15.00 as high and 15.00 and 24.59 as very high.
- So, we find that more loans are given at medium interest rate whereas very few loans are given at very high rates.



Plot -12 : Annual Income– Category wise

- Since the annual income lies between 4000 to 6000000, segmenting it into bins in order to understand and analyze it clearly. (These bins are open to understanding of the analysts).
- Categorizing annual income between 4000 and 40000 (min to 1st quartile) as low, 40000 and 80000(Aprox 3rd quartile start after this) as average, 80000 and 100000 as high and outliers as very high.
- The Plot shows that maximum loan applicants have average annual income.

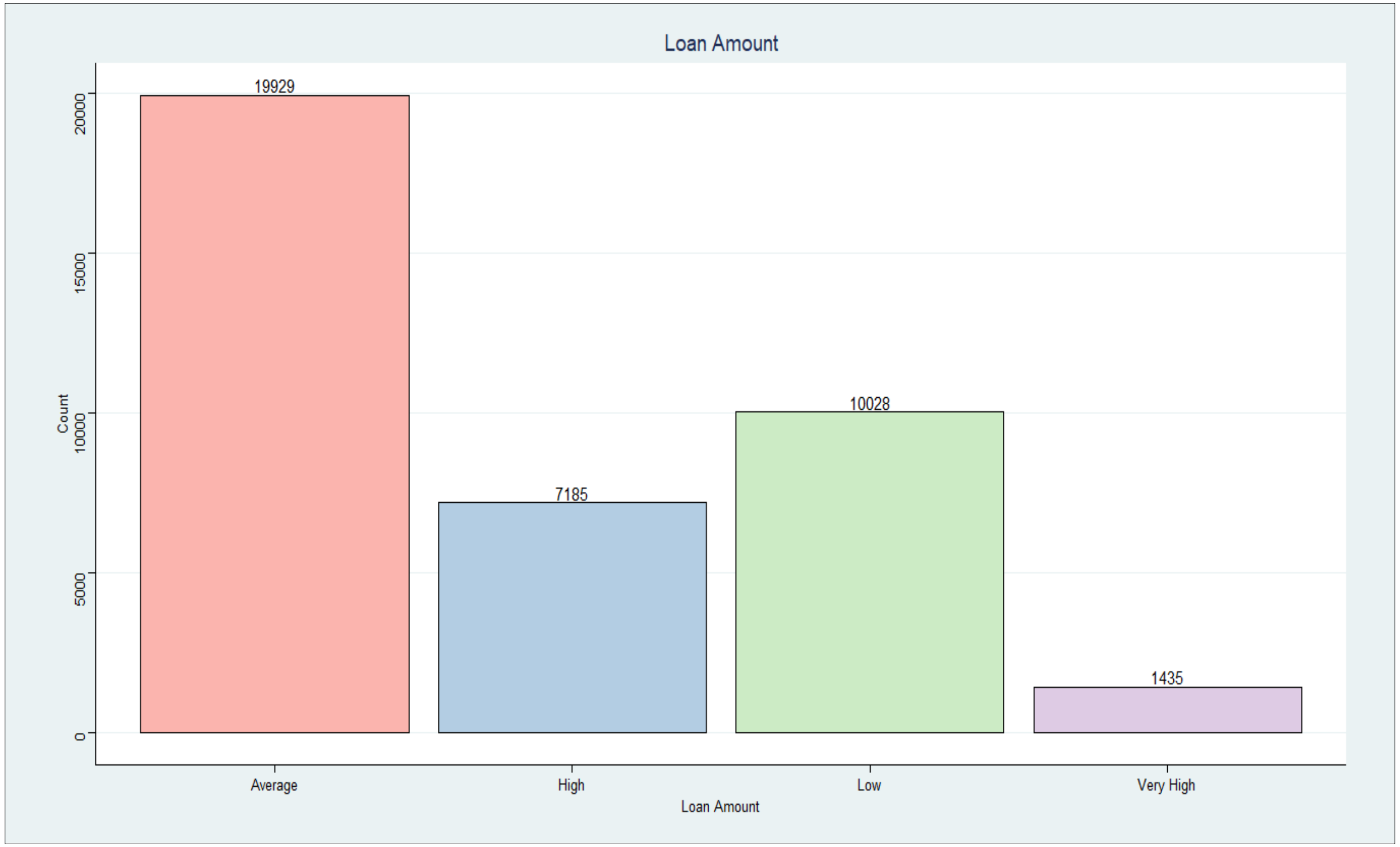


Plot -13 : Loan Amount– Category wise

- Similar approach was followed for segmenting loan amount

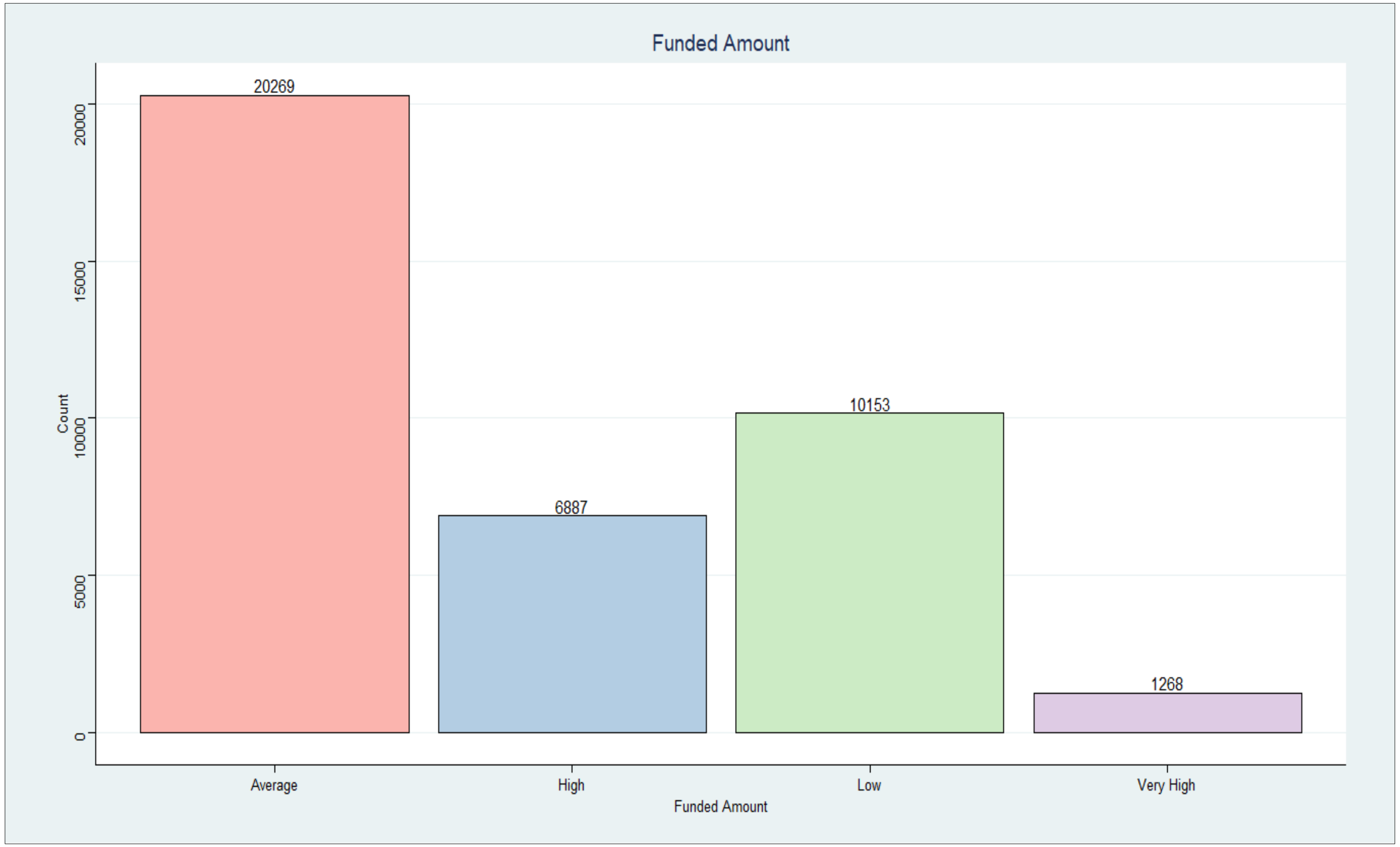
- Segmenting loan amount between 500and 5500(min to 1st quartile) as low, 5500and 15000(Aprox 3rd quartile start after this) as average, 15000and 25000 as high and outliers as very high.

- The Plot shows that maximum loan applicants are granted an average loan amount.



Plot -14 : Funded Amount– Category wise

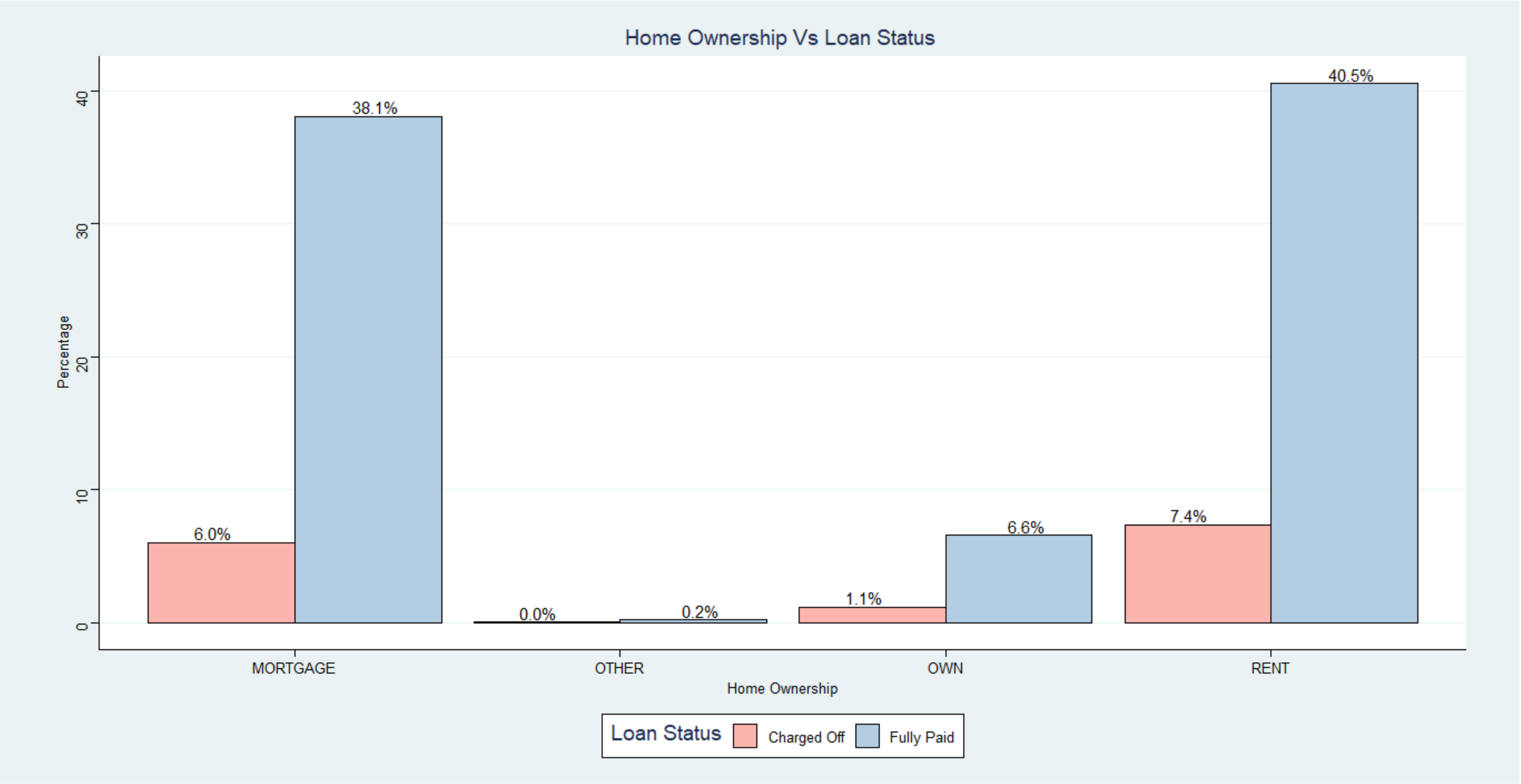
- Similar approach was followed for segmenting loan amount
- Segmenting loan amount between 500and 5500(min to 1st quartile) as low, 5500and 15000(Aprox 3rd quartile start after this) as average, 15000and 25000 as high and outliers as very high.
- The Plot shows that average funded amount is the highest.



Data Analysis – Bi-variate Analysis

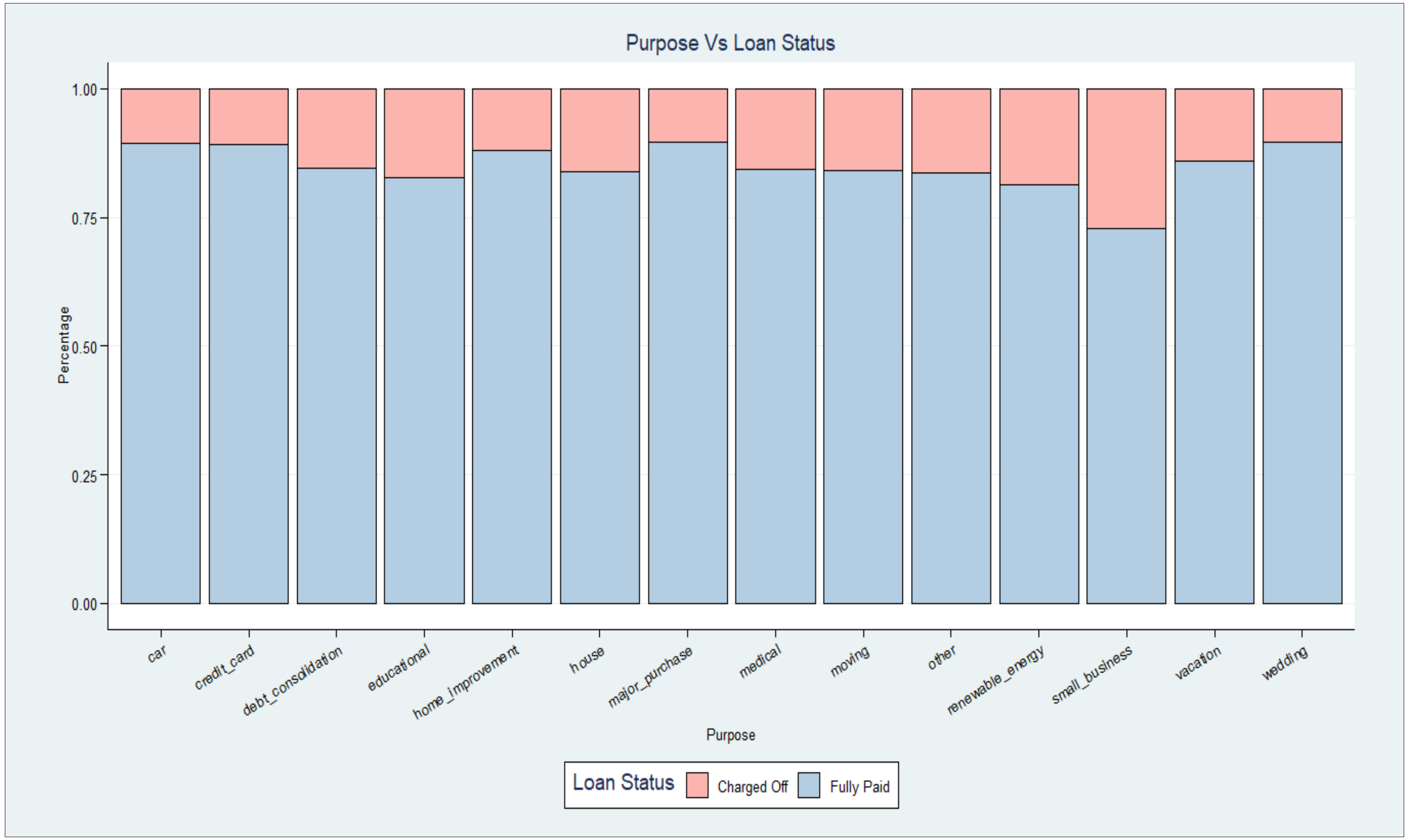
Plot -15 : Finding Relationship between Home Ownership and Loan Status

- This plot tells us that the borrowers with the home ownership status as “Other” have defaulted the most.
- This factor could be considered in future for giving the loans to the borrowers.



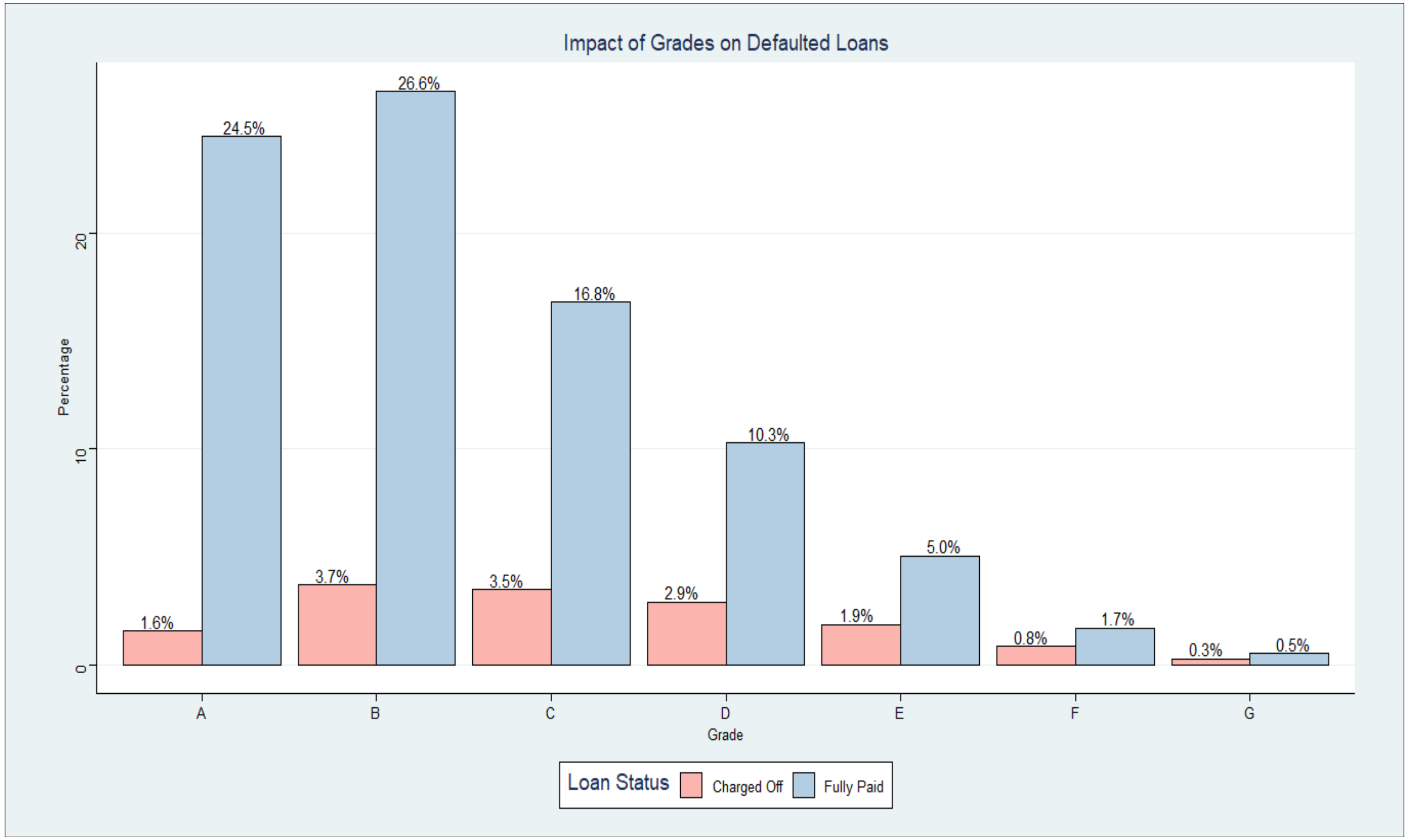
Plot -16 : Finding Relationship between Purpose and Loan Status

- The graph shows that loans taken for the purpose of small business have the most defaulters.
- However, most of the loans are taken for the purpose of debt consolidation. *[Plot -5]*.
- This factor can also be considered for our analysis as we see a significant defaulters for the purpose of small business.



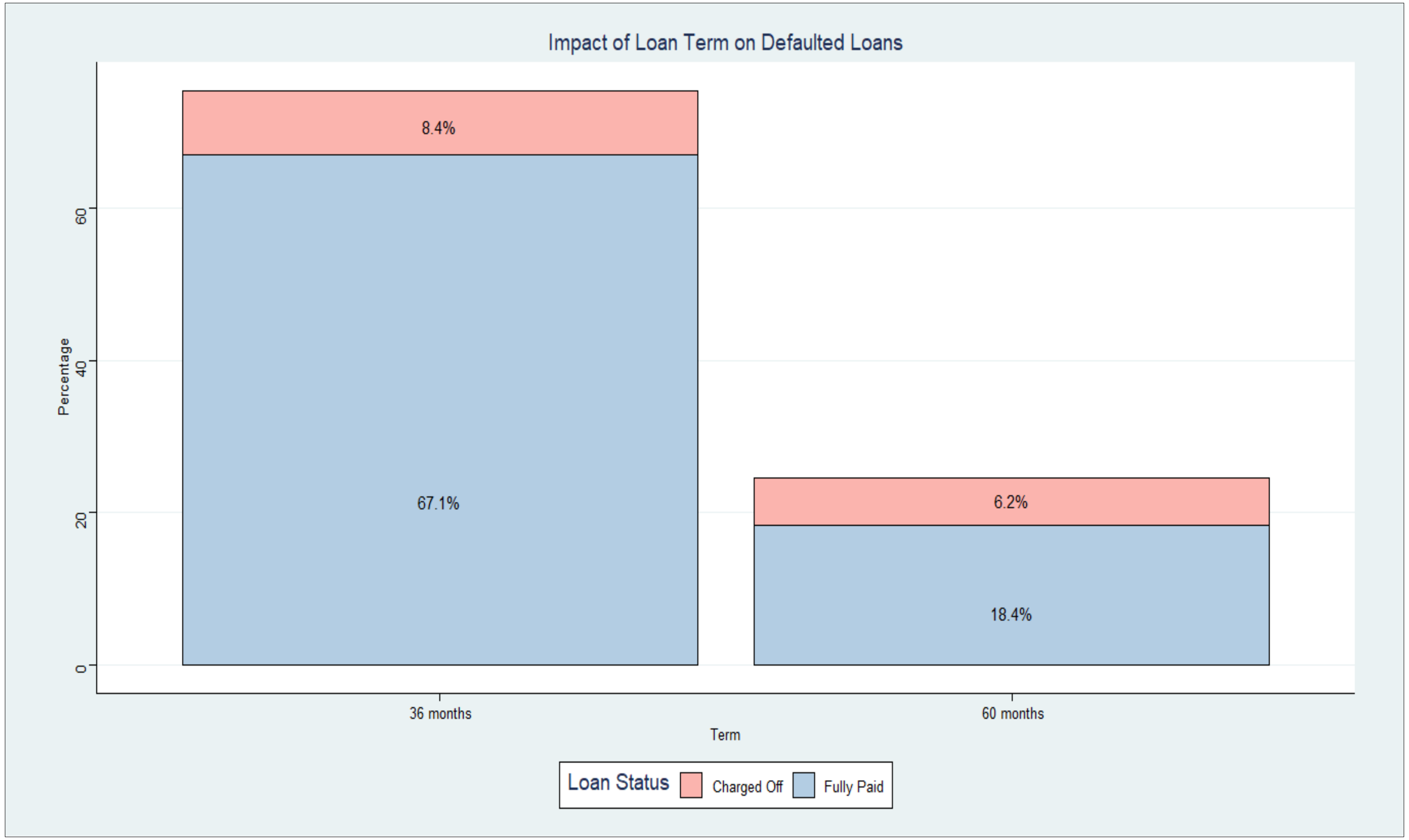
Plot -17 : Finding Relationship between Grade and Loan Status

- The plot tells that more defaulted customers are from Grade B and C



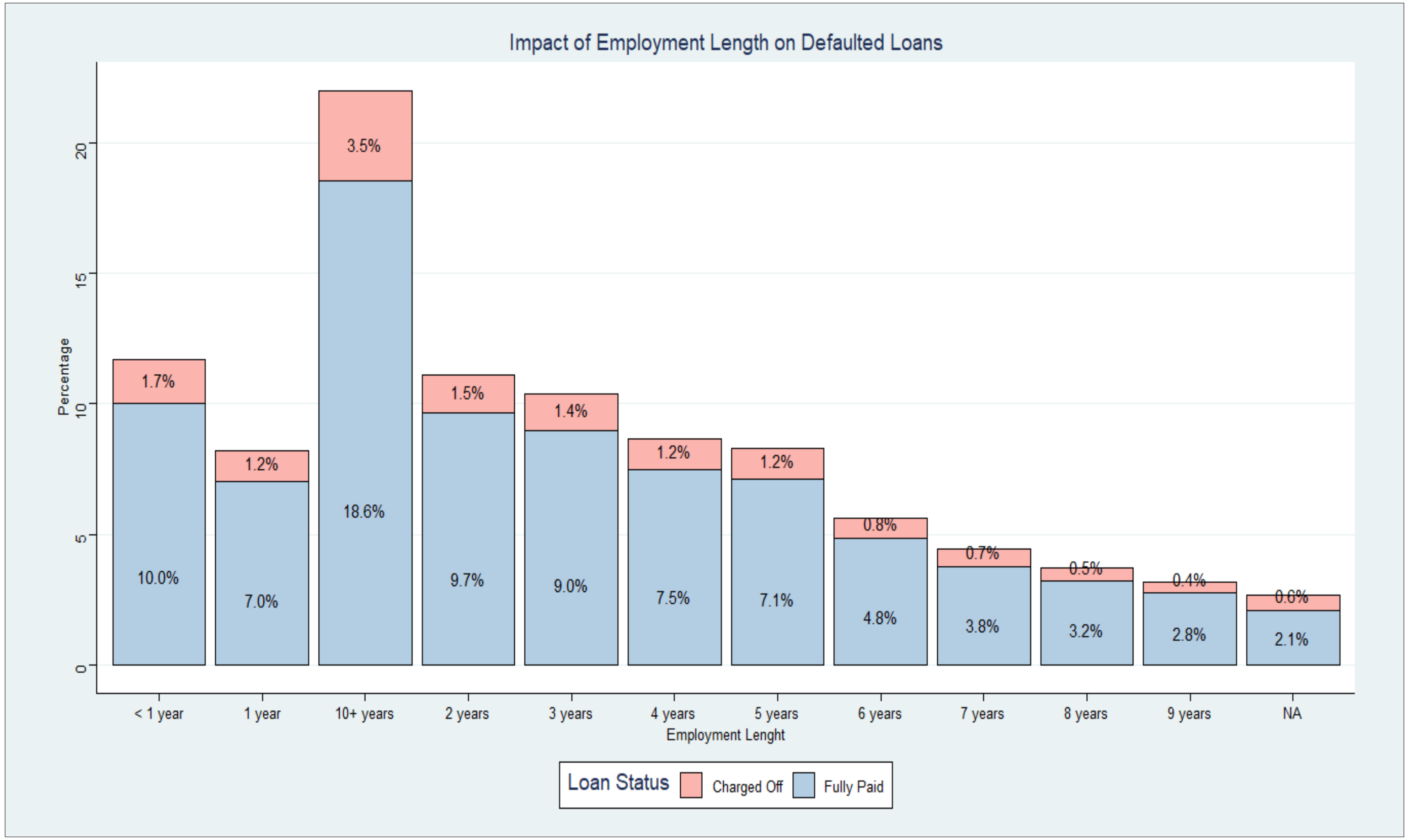
Plot -18 : Finding Relationship between Loan Term and Loan Status

- The plot tells that loan taken for the duration of 36 months is more likely to have defaulters than those taken for the period of 60 months.
- This can be considered a risk factor while disbursing loans in the future.



Plot -19 : Impact of Employment Length on Defaulted Loans

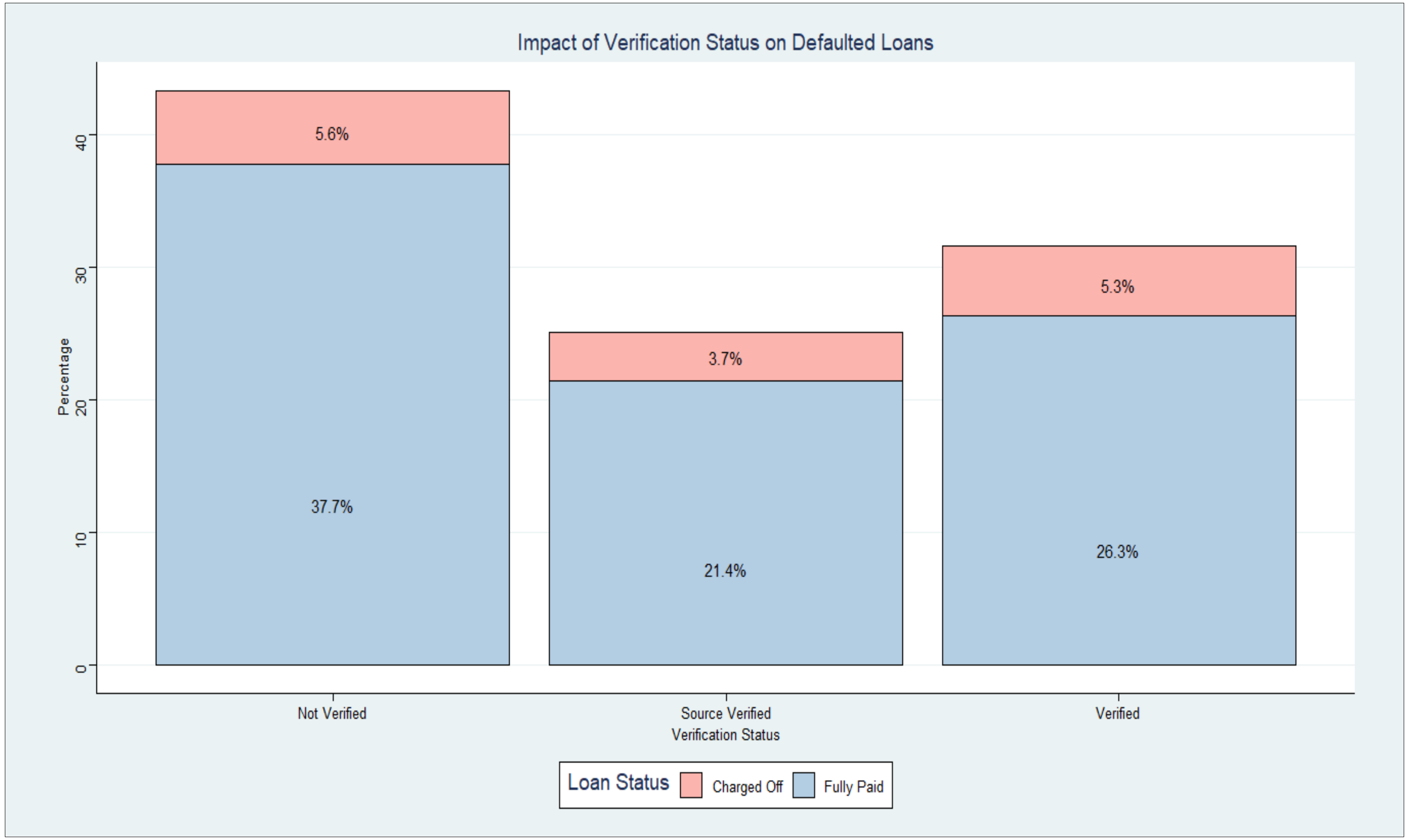
- We can see that there are the highest proportion of borrowers who have employment experience of more than 10 years as defaulters.
- But there seems to be not any specific correlation between the loan status and the employment term.



GRAMENER CASE STUDY

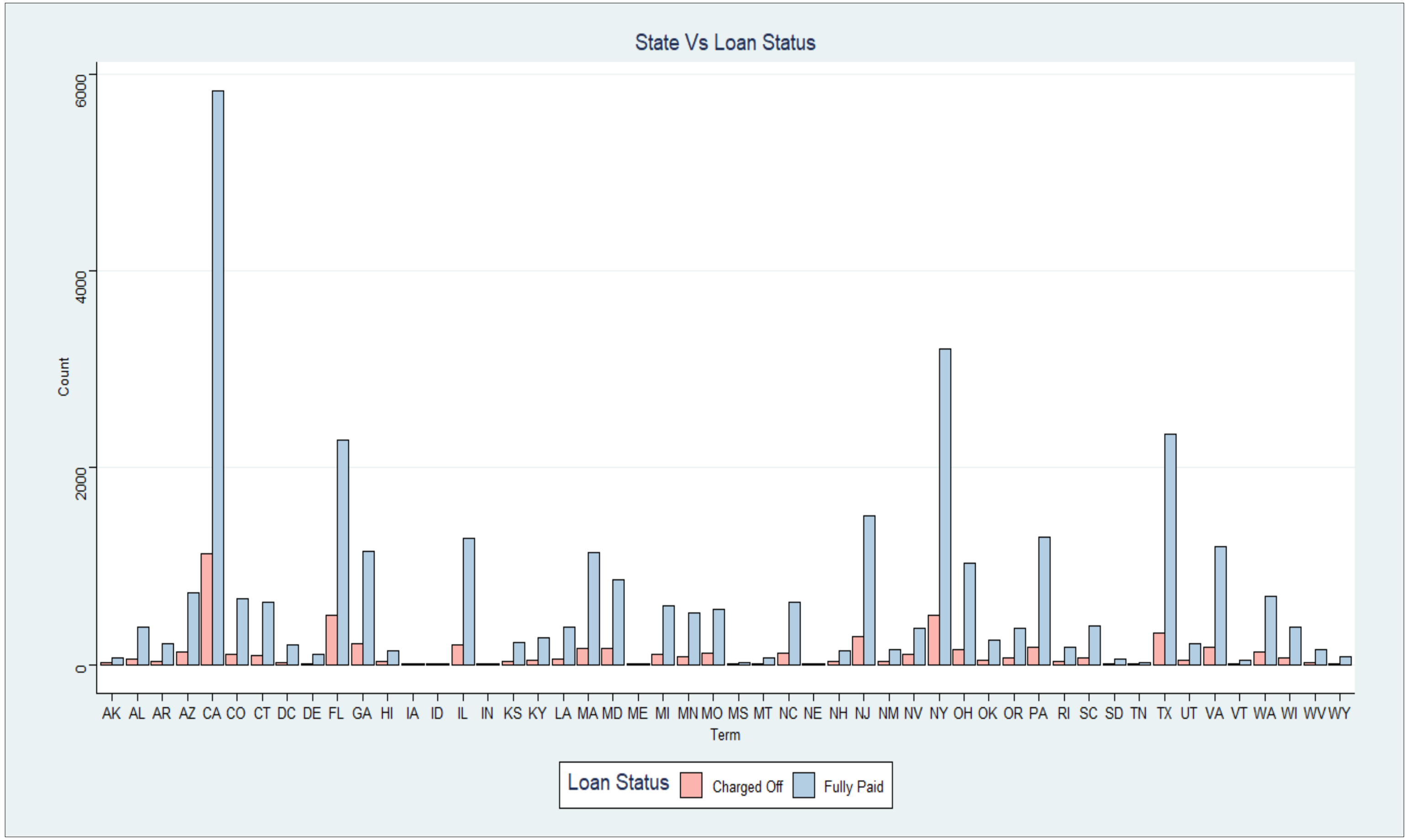
Plot -20 : Impact of Verification Status on Defaulted Loans

- From the plot it is surprising to see that verified loans have almost an equal amount of defaulters as not verified loans.
- This factor can not be considered for analyzing risky customers as there seem to be no clear correlation.



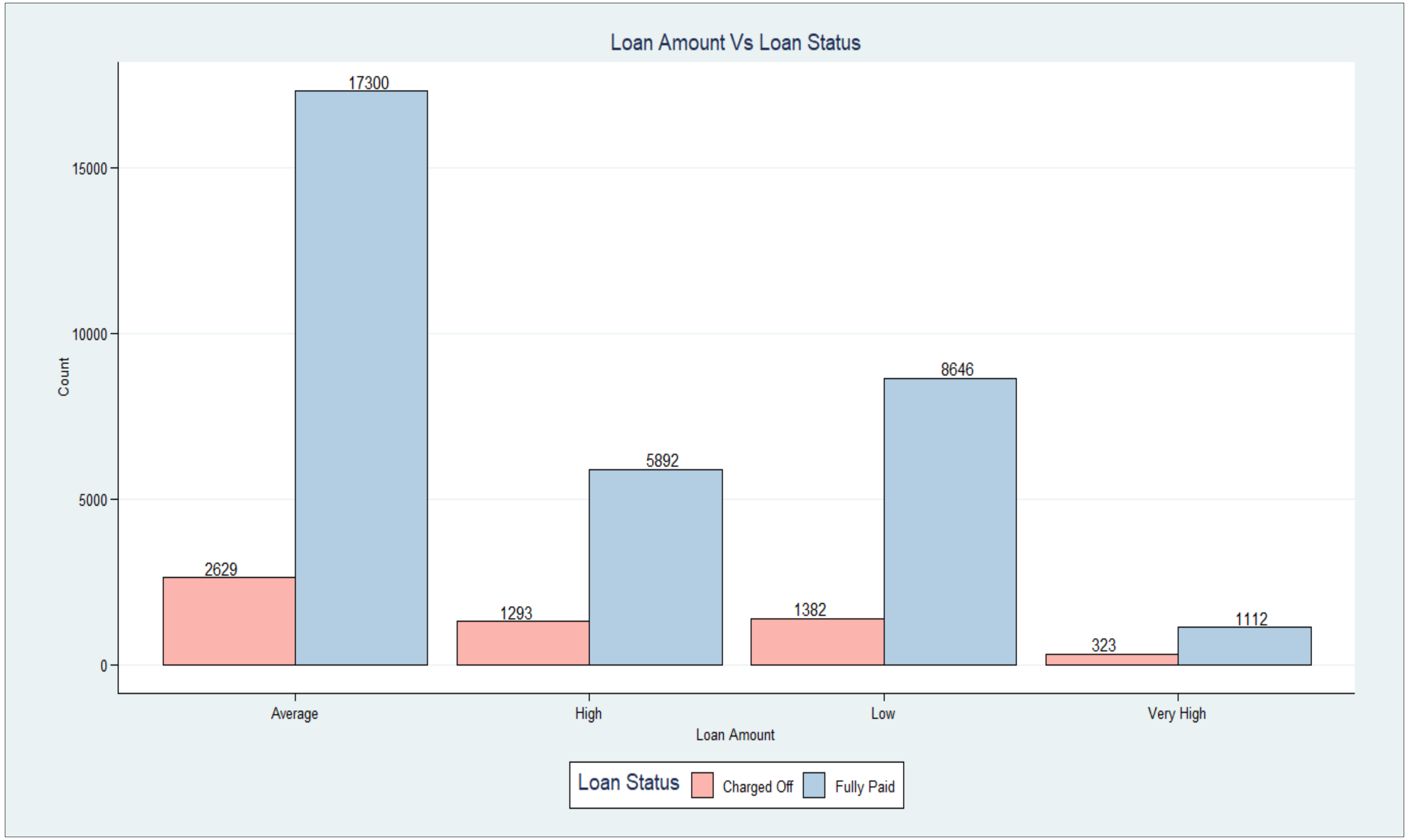
Plot -21 : Finding Relationship between State and Loan Status

- The plot shows that the maximum number of defaulters are from the state of California (CA).
- In future, the bank can suspect the borrowers from this state and go for a strict verification before lending to them.



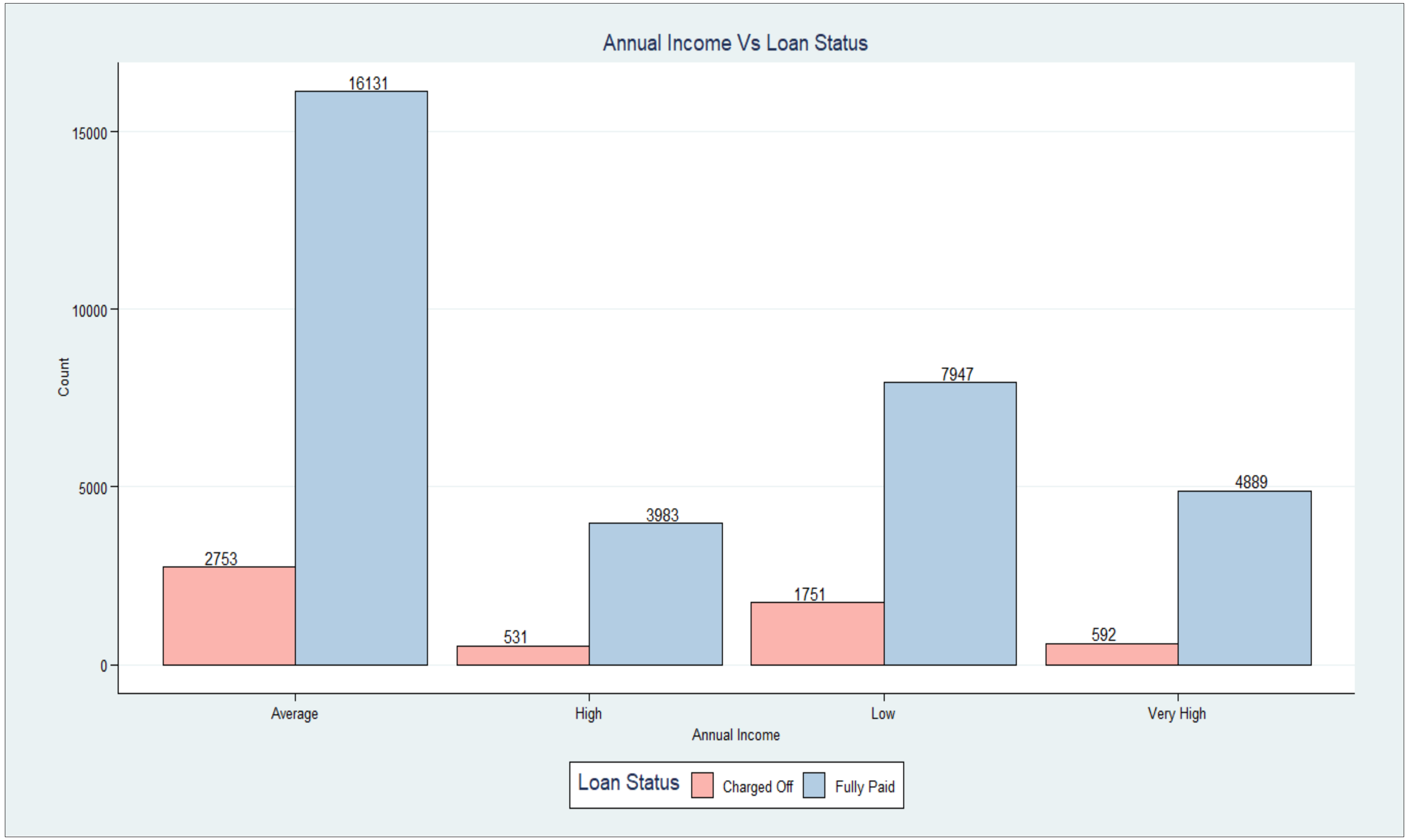
Plot -17 : Finding Relationship between Loan Amount and Loan Status

- The plot shows that the borrowers with average loan amount are more likely to default.
- One possibility for this could be that the interest on average loan amount is less.
- Another possibility could be that the borrowers who are given average loan amount have less or average annual income and they fail to fill the loan in due date.
- We need to analyze it further.



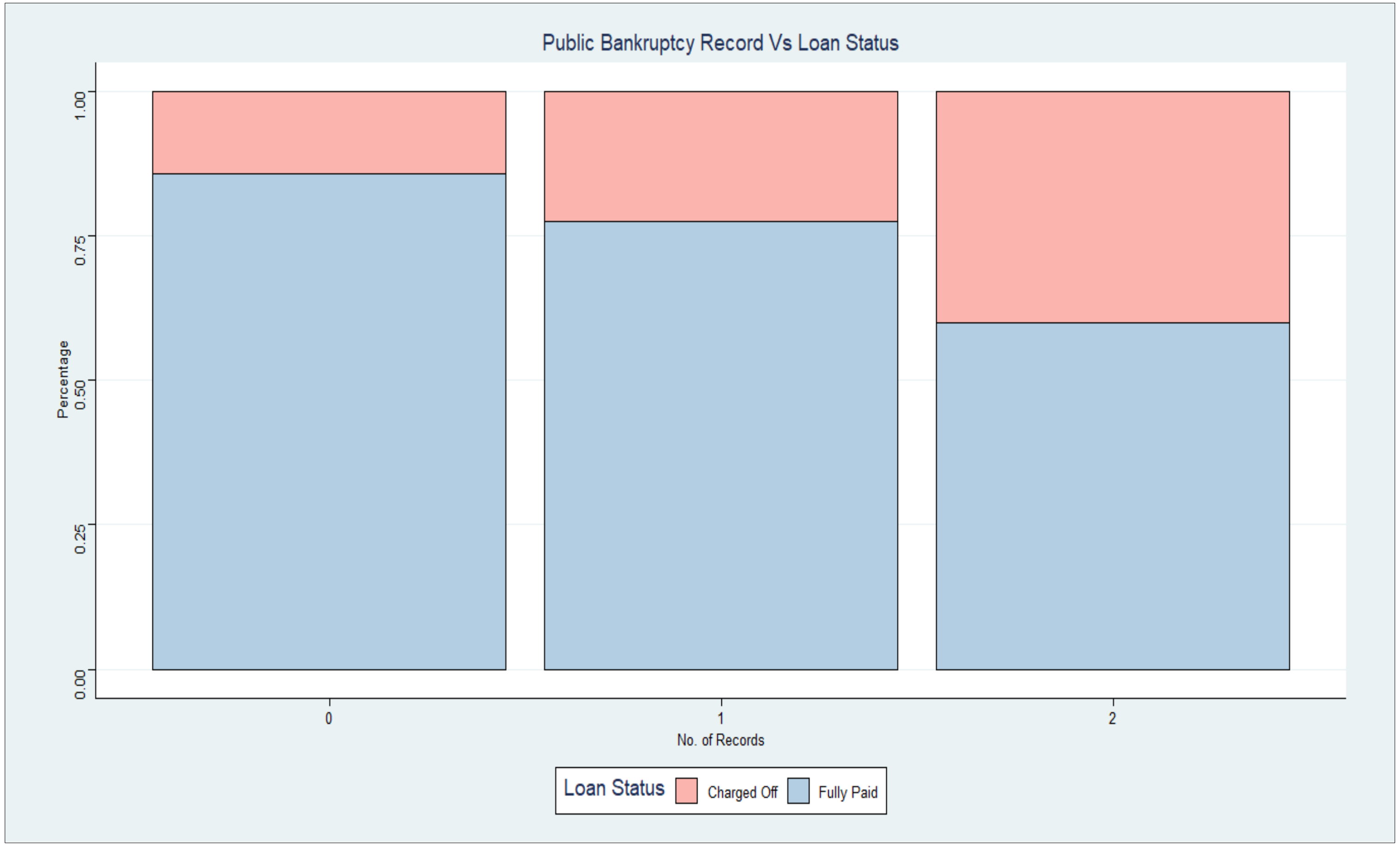
Plot -18 : Finding Relationship between Loan Status and Annual Income

- The plot helps prove our assumption from the previous plot that borrowers having average and low annual income are more likely to default than those having higher annual income.



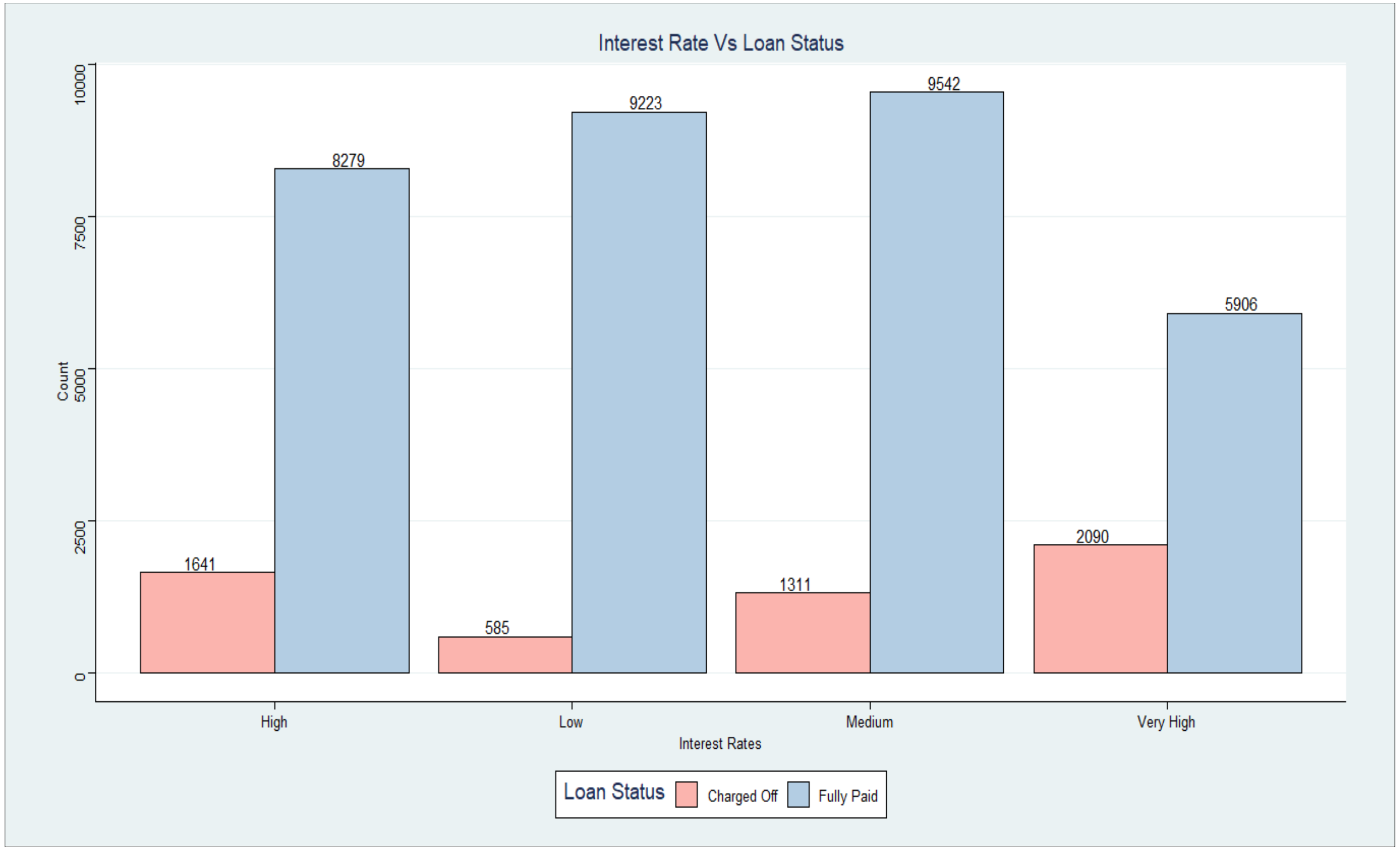
Plot -19 : Finding Relationship between Loan Status and Public Bankruptcy Record

- The plot shows that the borrowers having a higher number of public bankruptcy record default more than others.
- This should be considered as an important factor in providing loans to the prevent the company from loss in future.



Plot -20 : Finding Relationship between Loan Status and Interest Rate

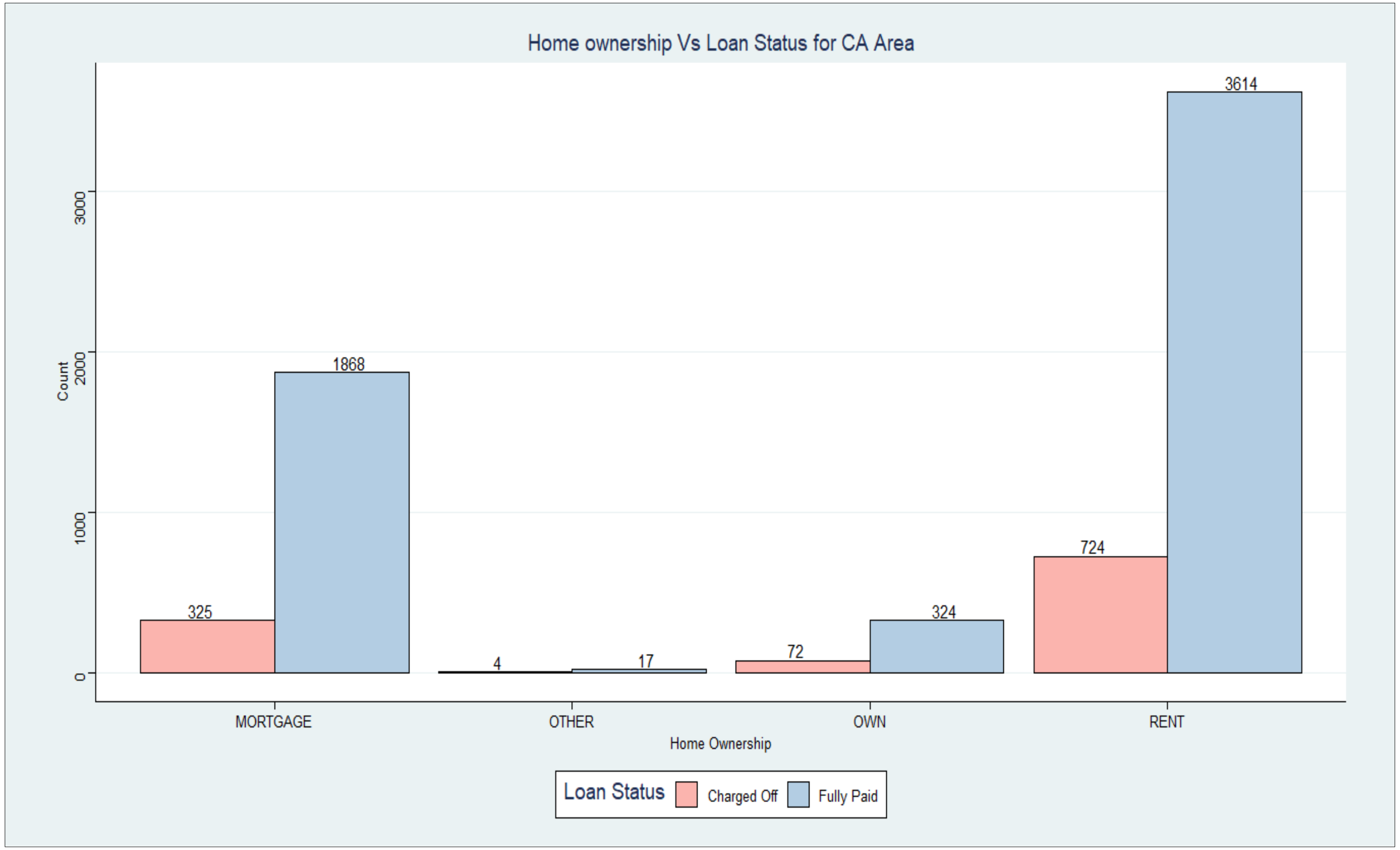
- The plot shows that borrowers who are given loan at high or very high interest rate default most.



Plot -21 : Finding the home ownership status of the borrowers from California

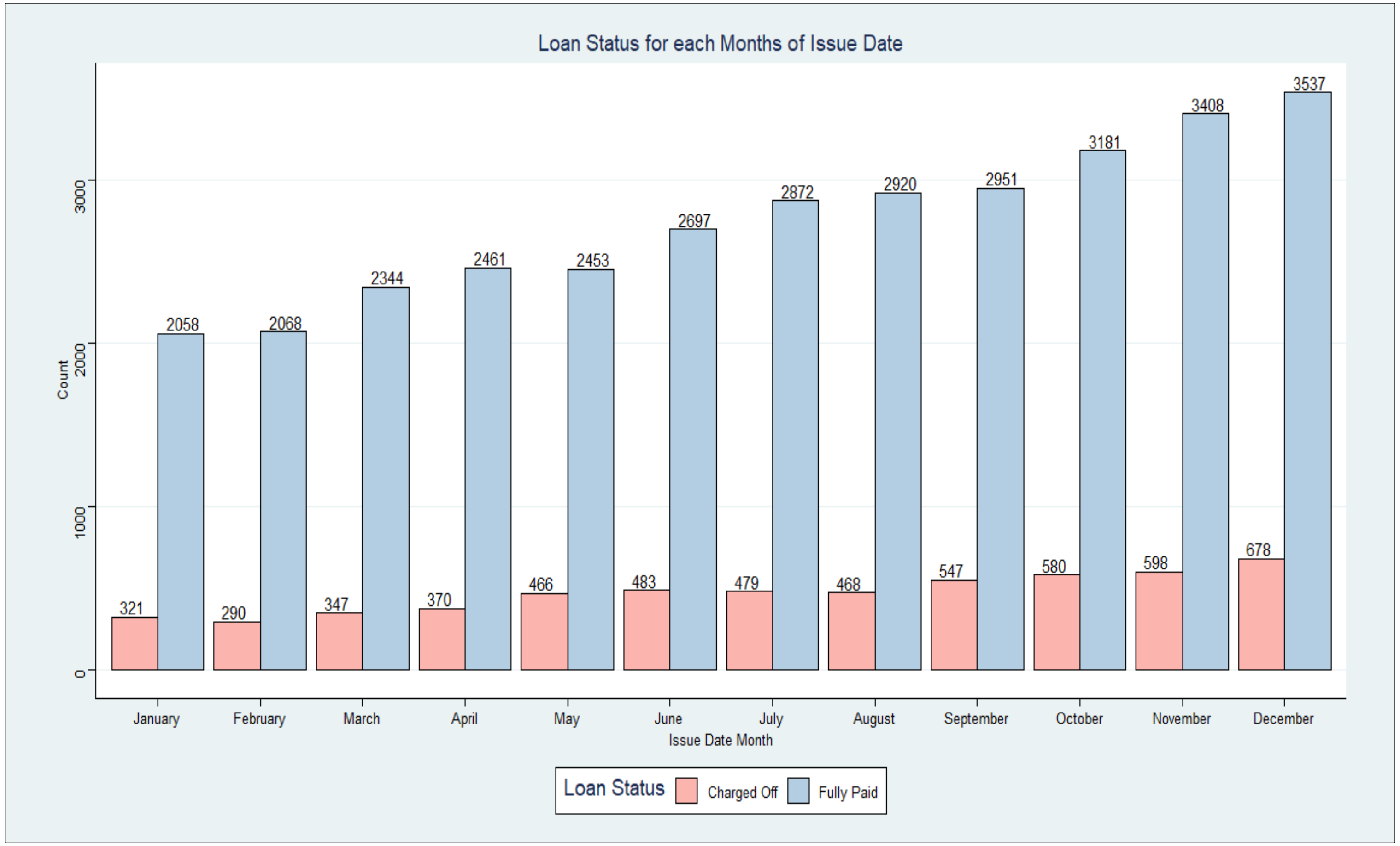
- The plot gives a very useful insight that the borrowers from the CA state defaulting most has the home ownership status of “RENT”.

- This can be considered for future risk analysis while lending loans to borrowers who are from the state of California and living in a rented home.



Plot -22 : Finding relationship between loan status and months of issue date : Derived Metrics

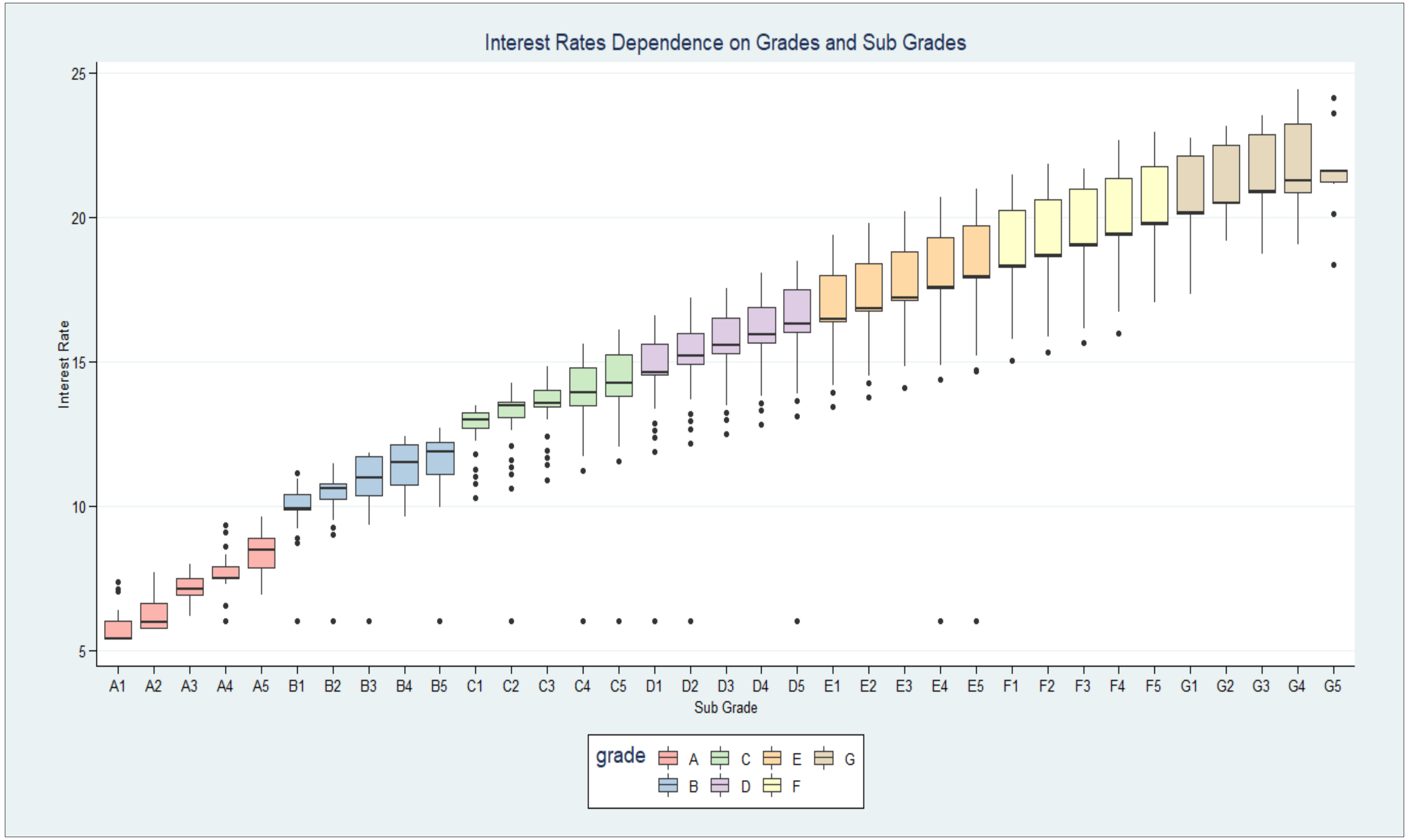
- The plot shows that the number of defaulters are relatively high in the months from September to December.
- There should be a strict verification specially during these months to ensure that the bank does not suffer losses from these borrowers.



Data Analysis – Multi-variate Analysis

Plot -23 : Grades, Sub-Grades and Interest Rates

- Interest rate shows a positive correlation with the sub-grades. Therefore giving the loan at a higher rate will depend on the sub-grade of the employee



Plot -24 : Corrplot of some varaibles

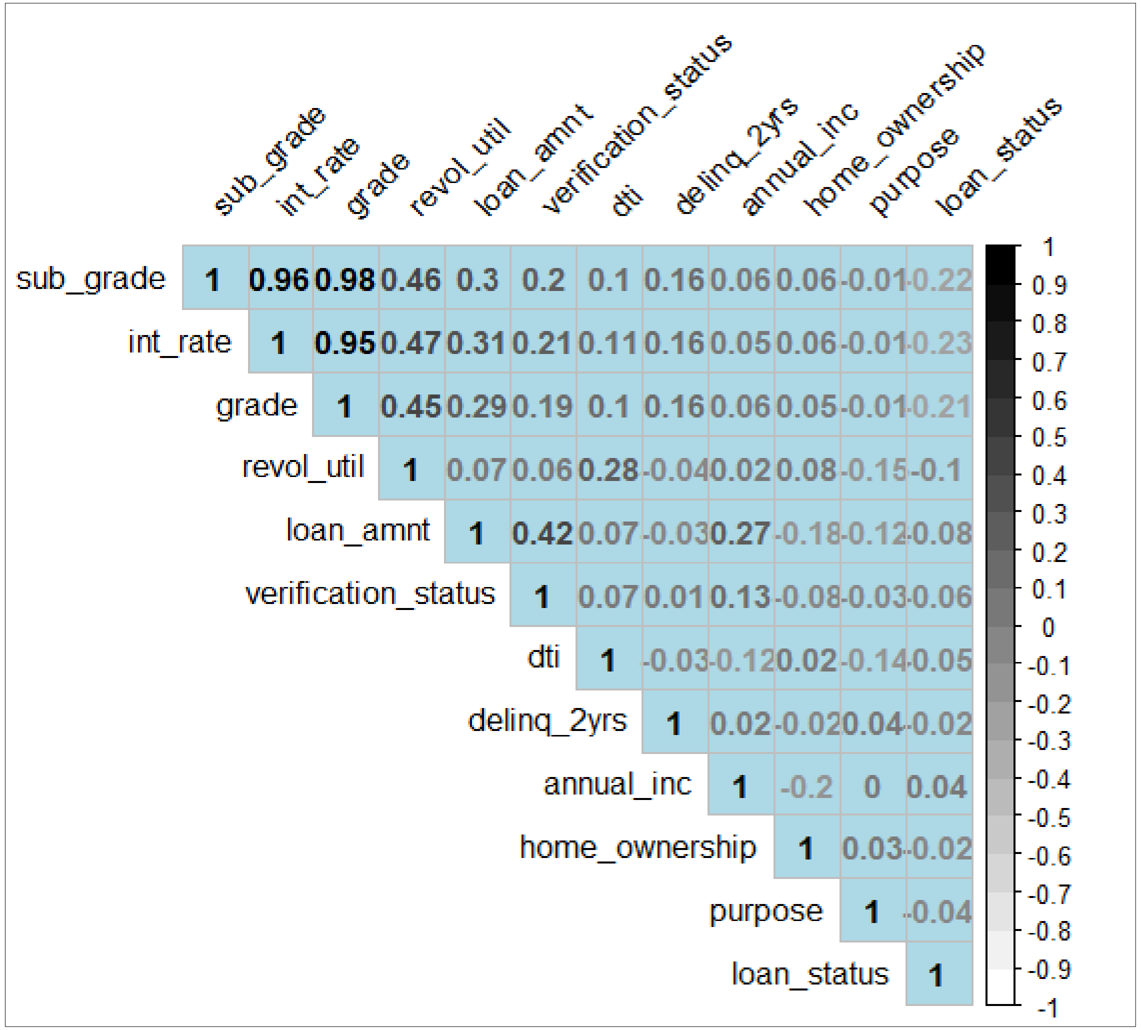
It is seen that

1. Sub grade has strong correlation with interest rate, revol_util and loan amount.

2. Interest rate and revol_util are moderately correlated

3. Revol_utli and dti are moderately correlated

4. Loan amount, verification status and annual income also have correlation which is obvious



	sub_grade	int_rate	grade	revol_util	loan_amnt	verification_status	dti	delinq_2yrs	annual_inc	home_ownership	purpose	loan_status
sub_grade	1	0.96	0.98	0.46	0.3	0.2	0.1	0.16	0.06	0.06	-0.01	-0.22
int_rate		1	0.95	0.47	0.31	0.21	0.11	0.16	0.05	0.06	-0.01	-0.23
grade			1	0.45	0.29	0.19	0.1	0.16	0.06	0.05	-0.01	-0.21
revol_util				1	0.07	0.06	0.28	-0.04	0.02	0.08	-0.15	-0.1
loan_amnt					1	0.42	0.07	-0.03	0.27	-0.18	-0.12	-0.08
verification_status						1	0.07	0.01	0.13	-0.08	-0.03	-0.06
dti							1	-0.03	-0.12	0.02	-0.14	-0.05
delinq_2yrs								1	0.02	-0.02	0.04	-0.02
annual_inc									1	-0.2	0	0.04
home_ownership										1	0.03	-0.02
purpose											1	-0.04
loan_status												1

Conclusions

From the above EDA analysis we can conclude following variables are strong indicators of default assumptions :

1. Purpose – *[Plot -5,16]*
2. Term –*[Plot -18]*
3. Home Ownership –*[Plot – 15]*
4. Public Bankruptcy Record –*[Plot – 19]*
5. State –*[Plot – 17,18 & 21]*

Recommendations

Some of the recommendations we would like to give to the bank are :

1. Restrict from giving loan to borrowers who take for the purpose of small business. It is seen that these are the ones who default the highest and bank can get in loss if it gives a lot of loans to these borrowers. If giving loans to them, bank can increase the interest rates so that the bank profits.
2. Bank should apply strict regulations and guidelines while lending to borrowers from California state living in rented homes as these are the ones who have been consistently defaulting. This can be reduced by either giving very less loan amounts or giving higher rates of interests.
3. The loan term should be changed from 36 months and increased to few more months. It is seen that most of the defaulters are those who take the loans for shorter term.
4. In order to reduce the losses bank should strictly avoid giving loans to people who have a higher public bankruptcy record. Since they already hold a record of being bankrupt, by lending to them bank is putting itself in a very high risk of loss.

Thank You !