

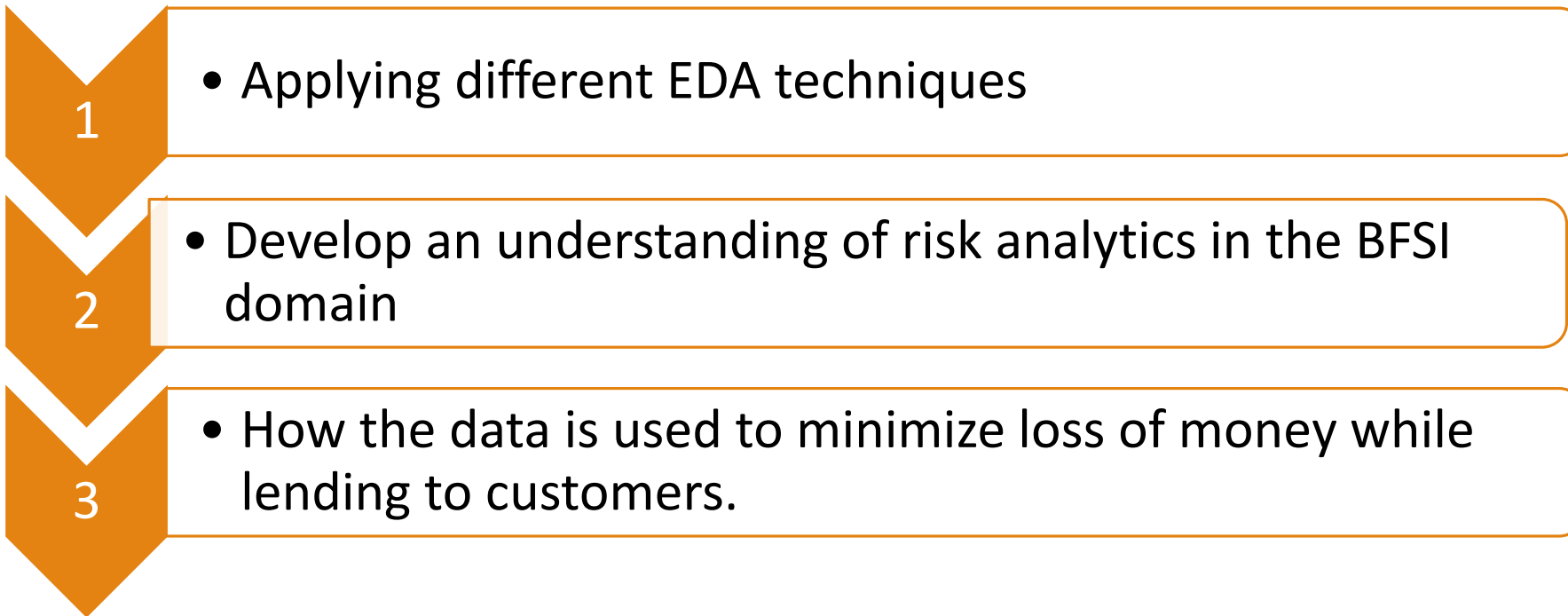
Lending Club Case Study

AVINASH KUMAR, KRISHNA MOJAMDAR



Objective

The aim of this case study is to get an idea of how real-world business problems are solved using EDA.

- 
- 1 • Applying different EDA techniques
 - 2 • Develop an understanding of risk analytics in the BFSI domain
 - 3 • How the data is used to minimize loss of money while lending to customers.

Business Understanding

When the company receives a loan application, the company must decide for loan approval based on the applicant's profile.

There are two types of risks associated with the bank's decision:

- 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

The data given contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Types of decisions

Loan
Accepted

- Fully Paid
- Charged Off
- Current

Loan
Rejected

- Not Considered and not part of the dataset

Loan Accepted

Fully Paid

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

Current

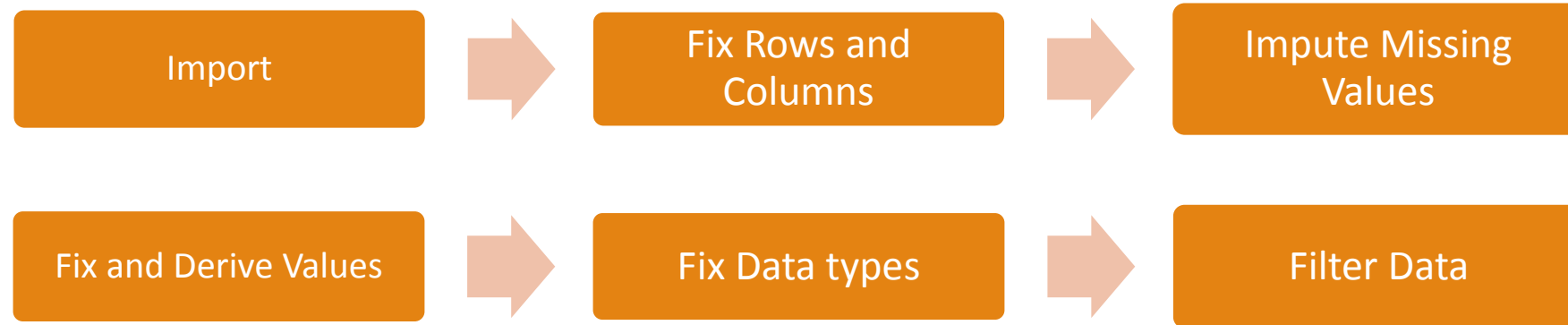
- Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

Charged-off

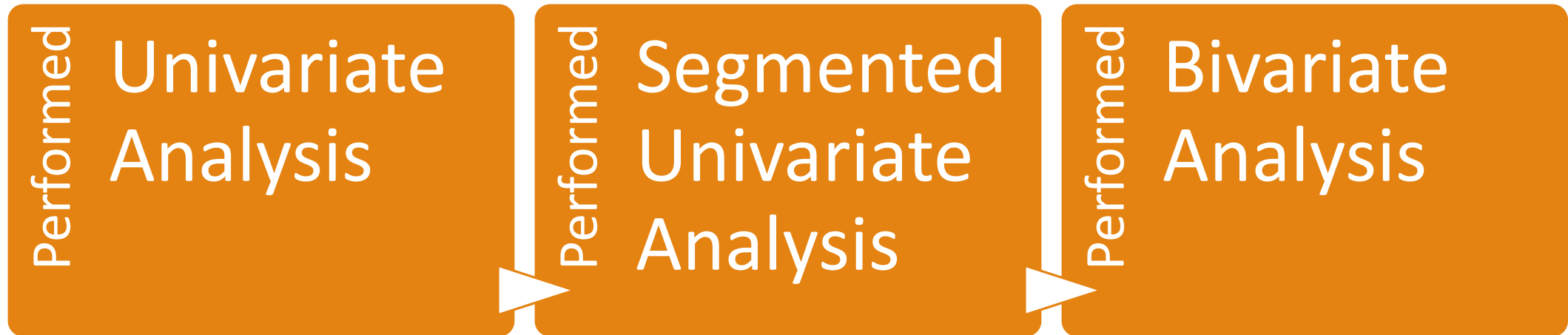
- Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

Data Cleanup

The following steps were taken to clean up the data and make it ready for EDA.



Exploratory Data Analysis



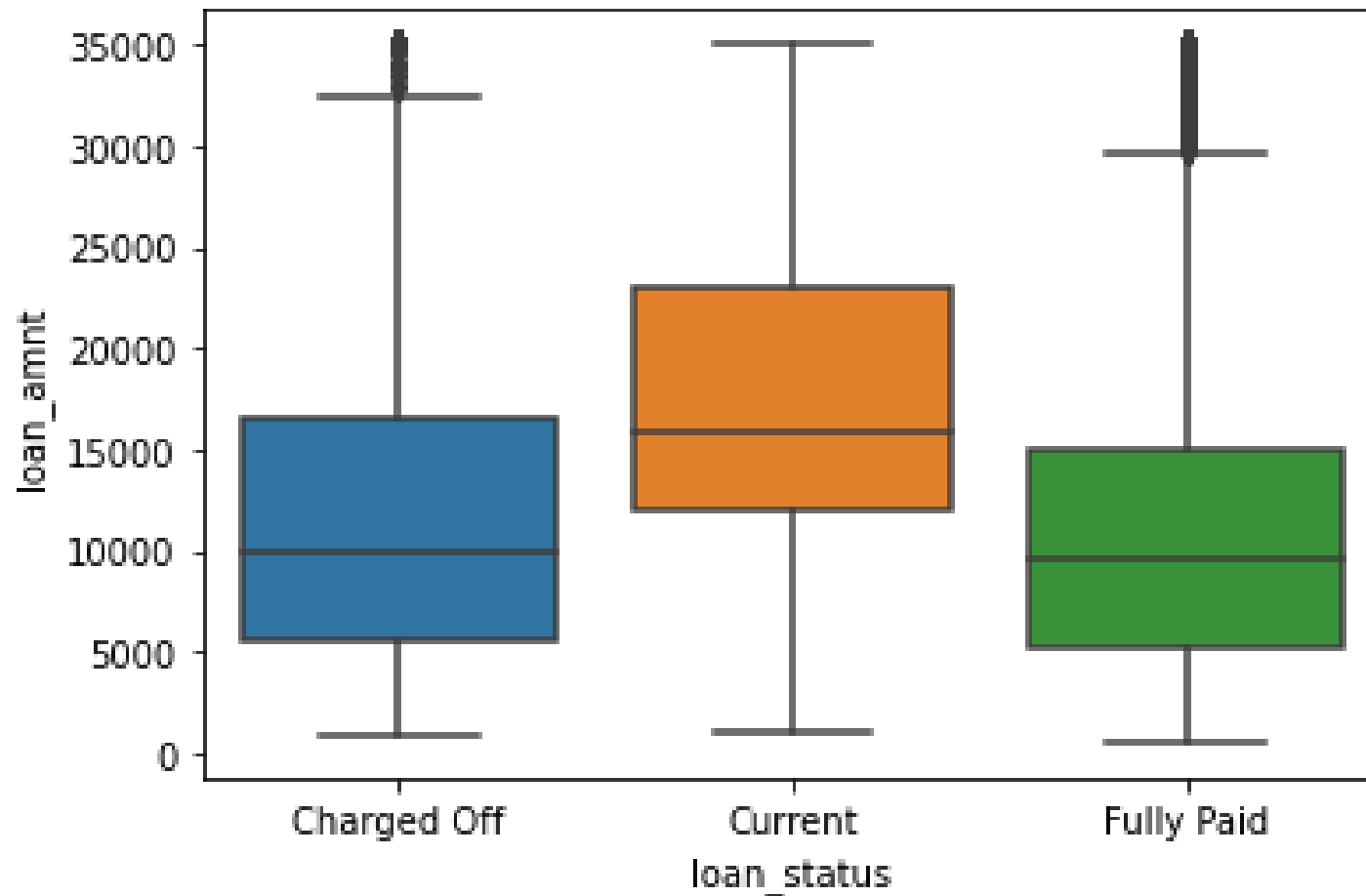


Univariate Analysis

Analysis of single variables.

Does not involve relationship with any other variable

Descriptive in nature

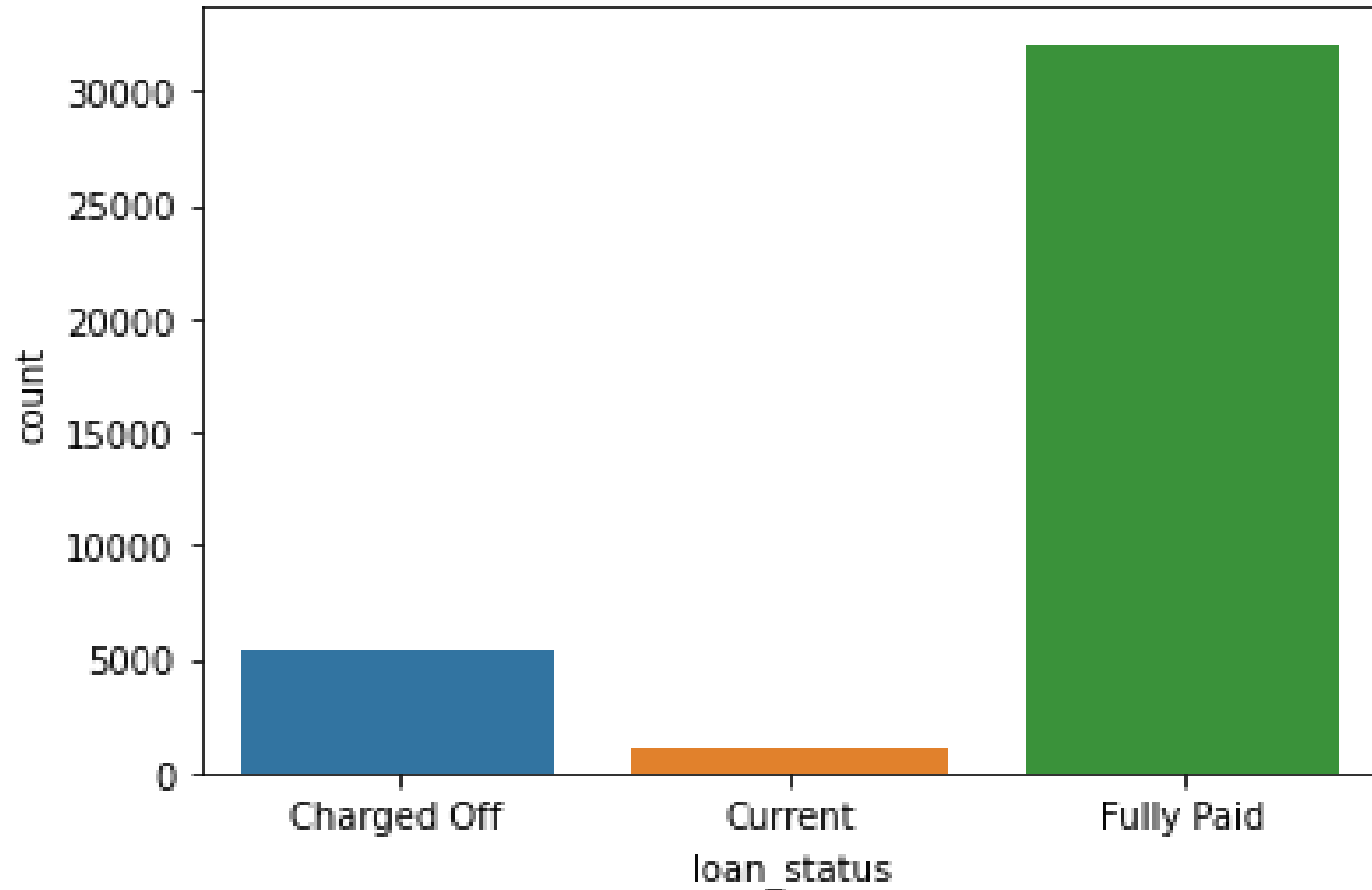


Loan Status

The loan amount ranges from 0 to 35000 with the mean being an amount of 10000.

There are a few outliers.

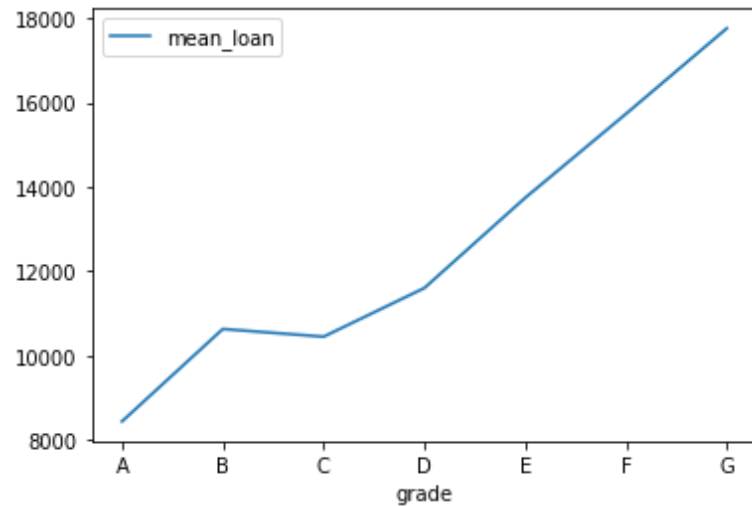
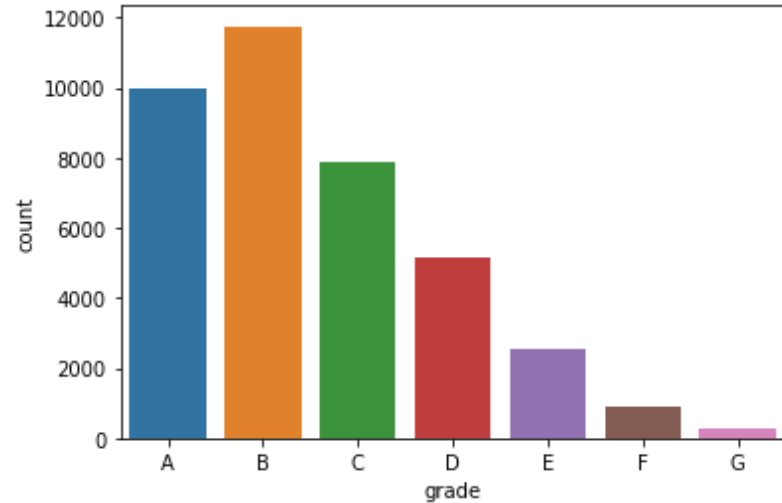
Outliers above 29250 were removed from the dataset



Loan Status

Majority of the loans have been fully paid.

There are a few which have been charged off.

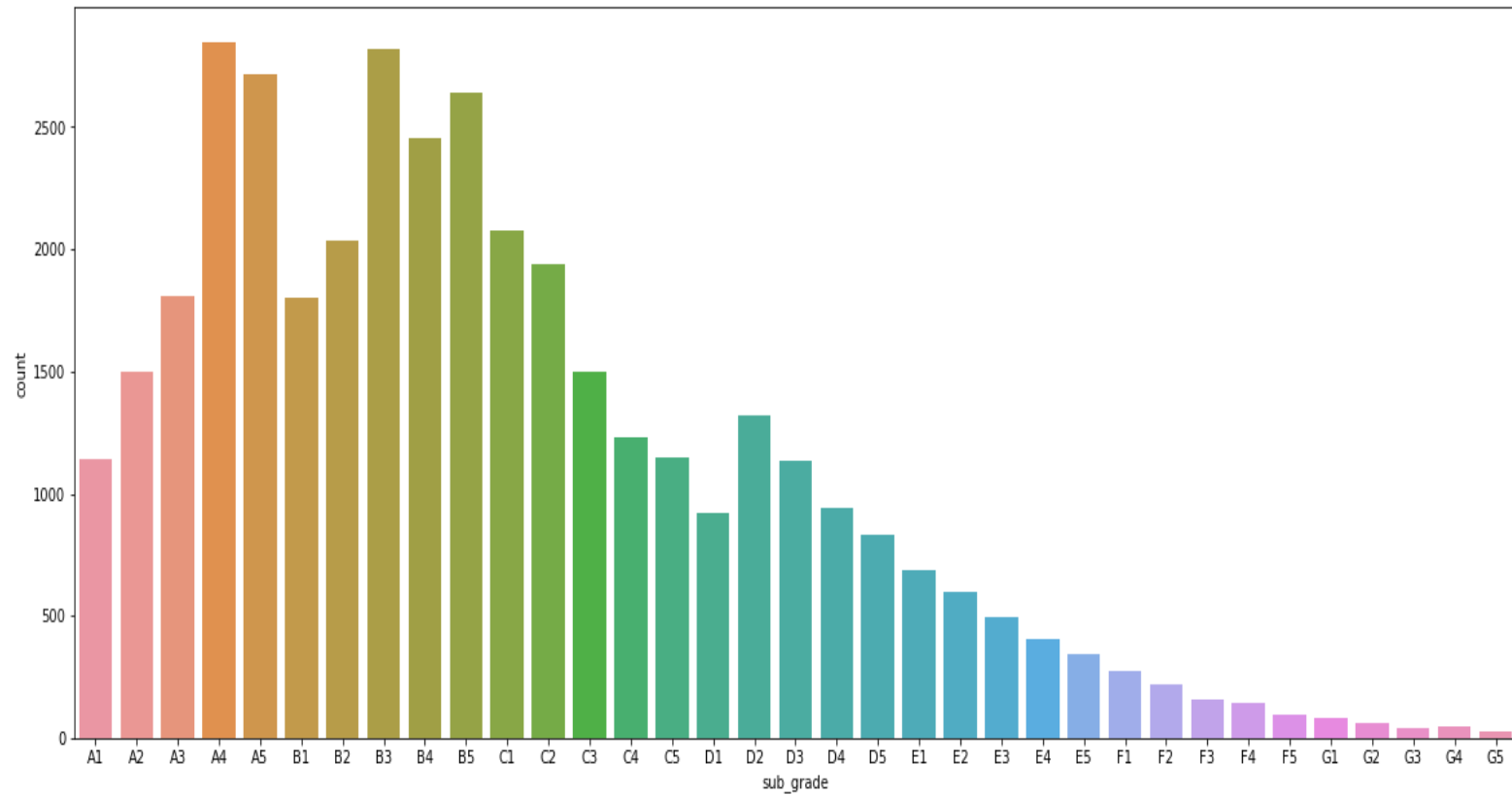


Loan Grade

Grade A is basic loan and Grade G is premium loan

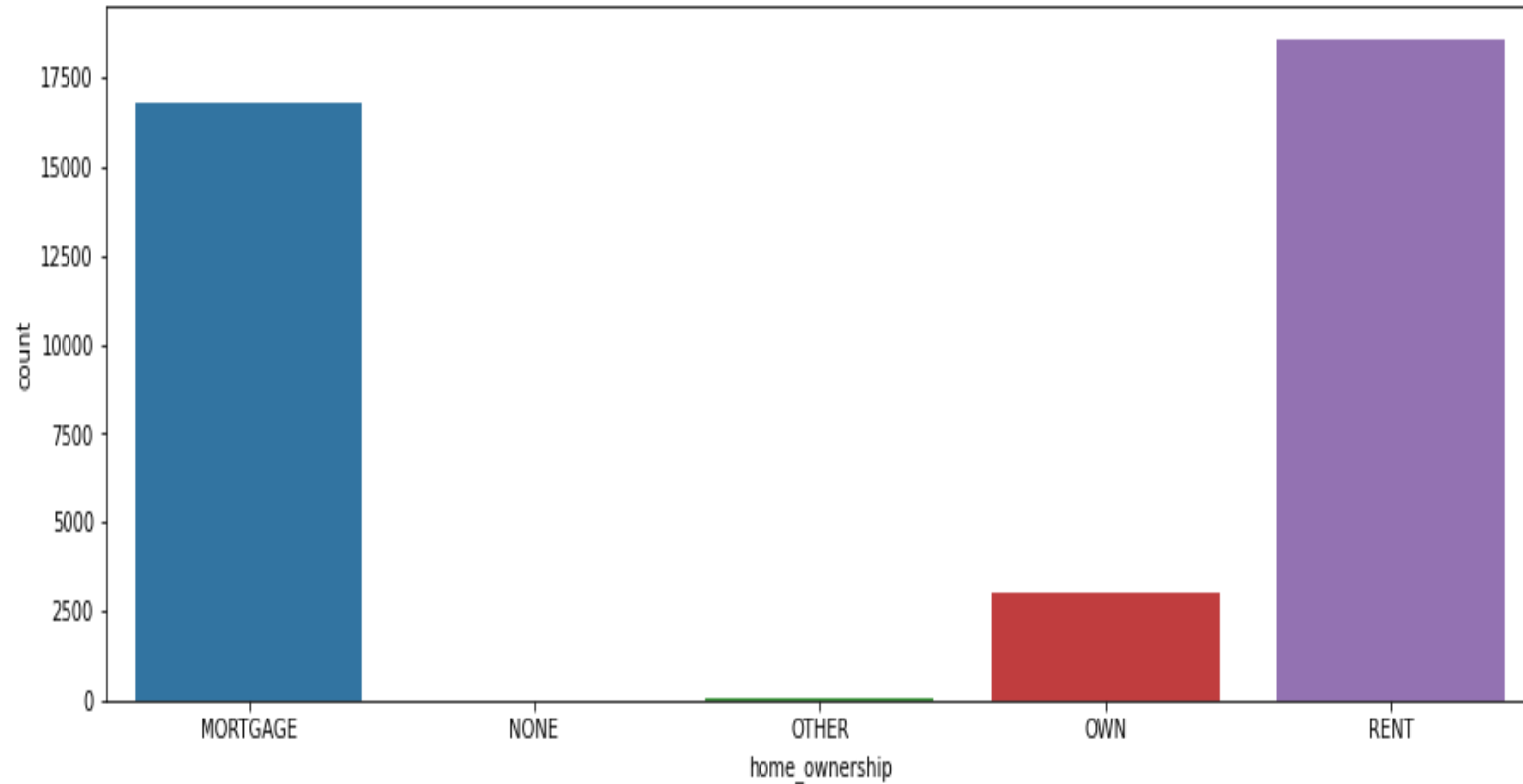
Most of the loans fall under grade A and B. Hence most of the loans are low graded loans.

Note: It's assumed that Grade G is the highest based on the mean of loans.



Loan Grade

Majority of the loan fall under sub grade A4 to B5

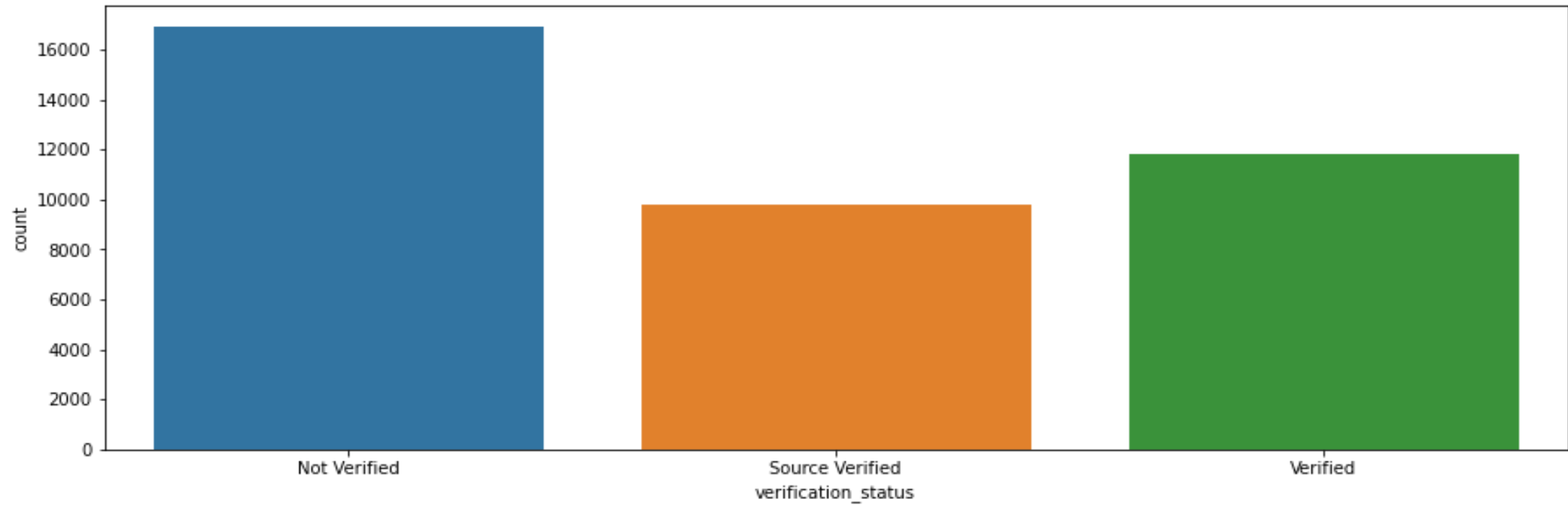


Home Ownership

A high number of the applicants live in rented and mortgaged houses.

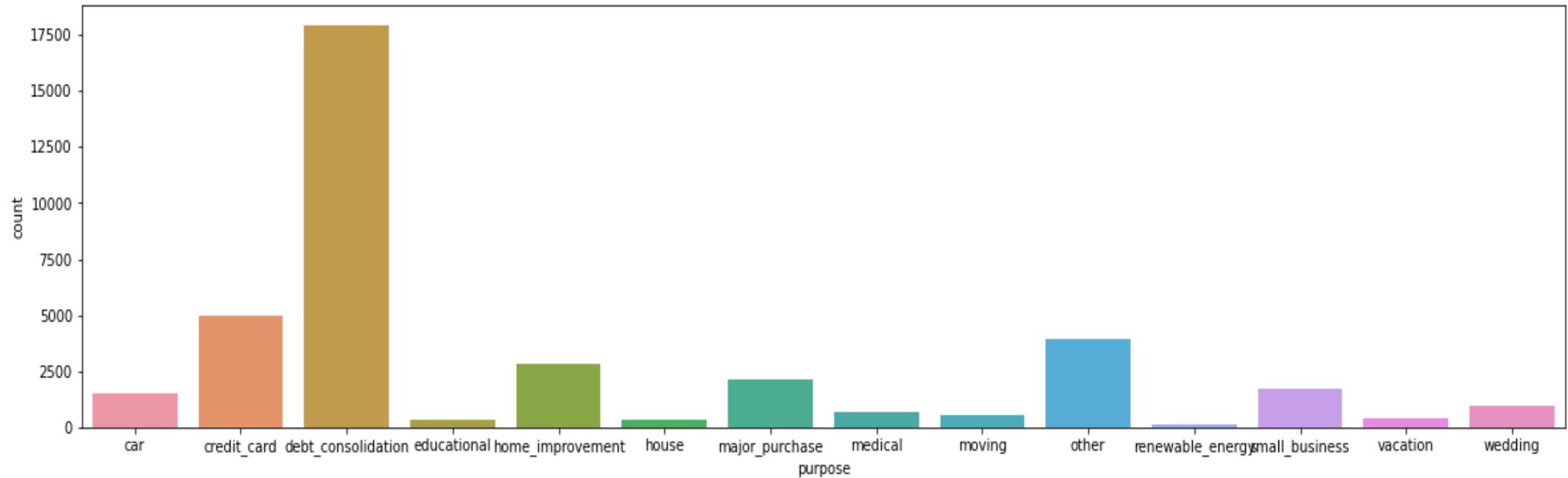
Very few of the customers own a house.

Verification Status



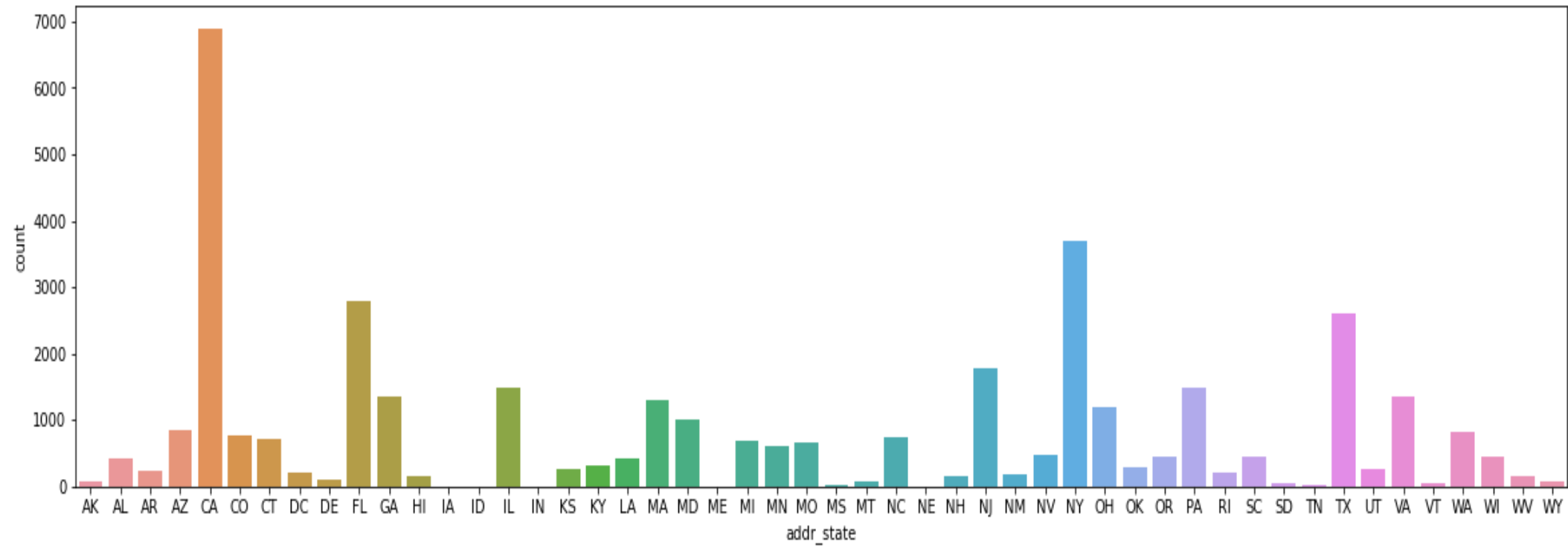
A large number of customers were totally not verified

Purpose of loan



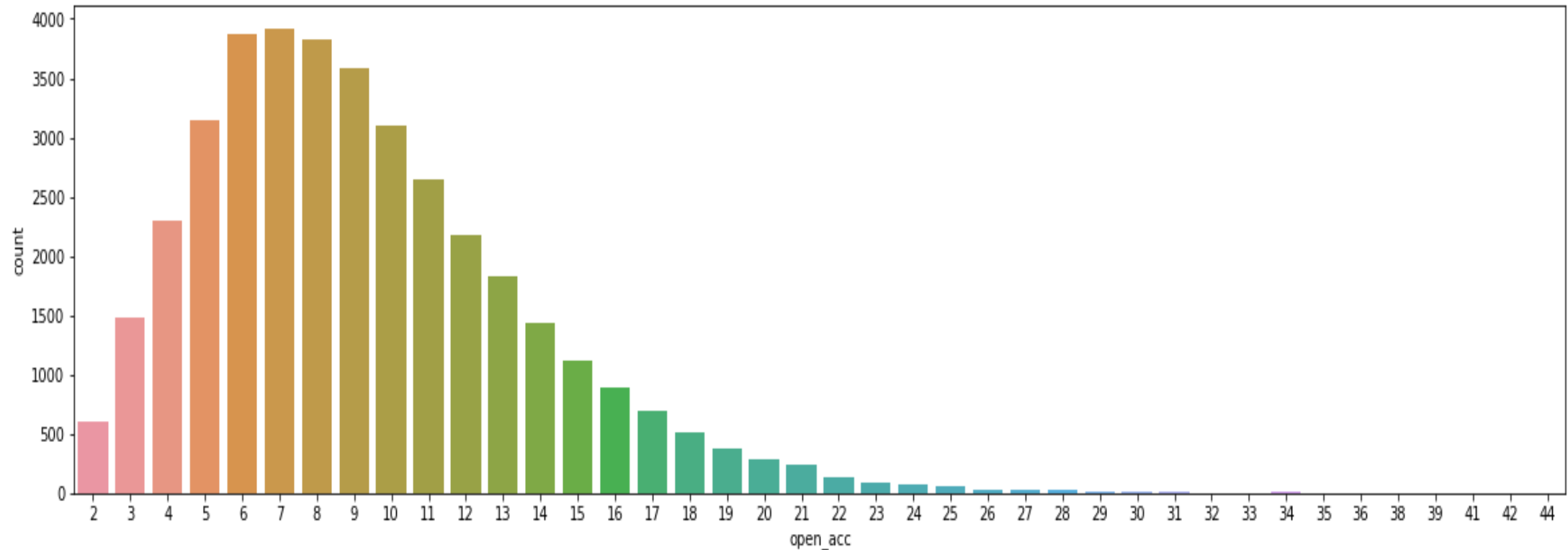
Debt consolidation seems to be the major purpose of taking a loan.

State Wise



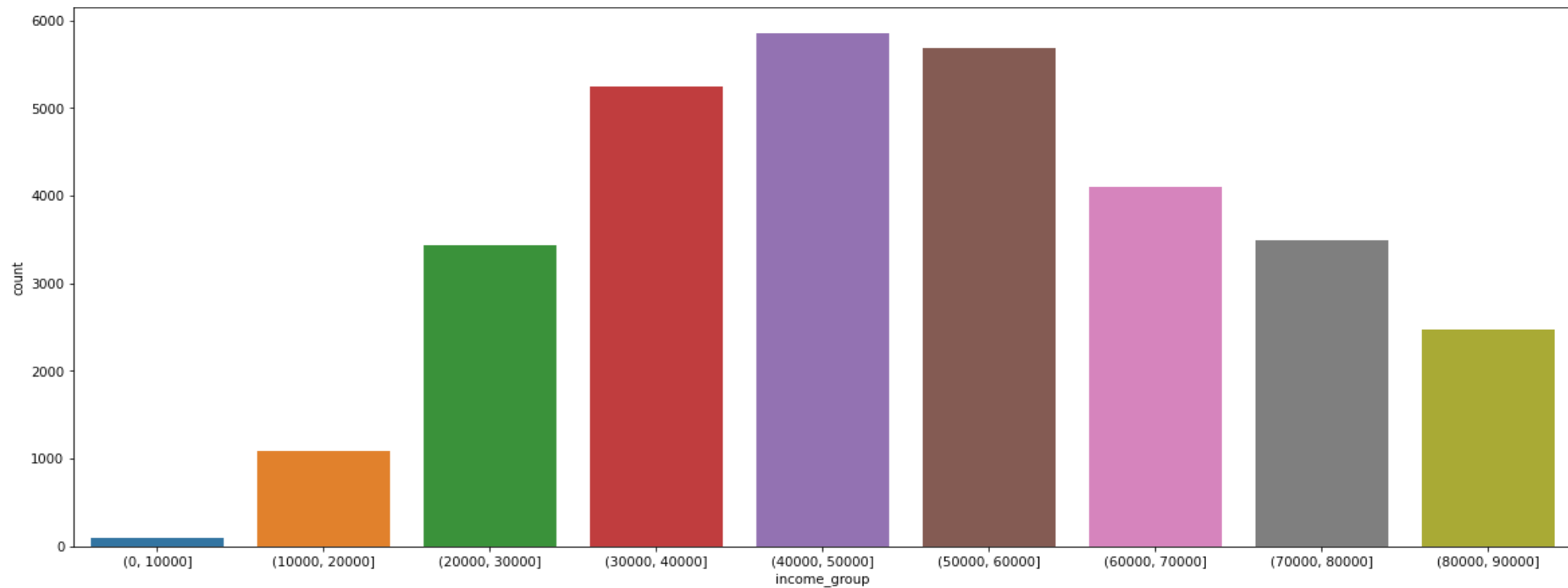
A large number of loans were granted in CA followed by NY and TX.

Credit Line



A majority of customers have more than 5 credit lines with the highest being in the range 5-11.

Income Groups



Income group 40000-60000 has the highest number of applicants in the dataset.

Loan Amount Category

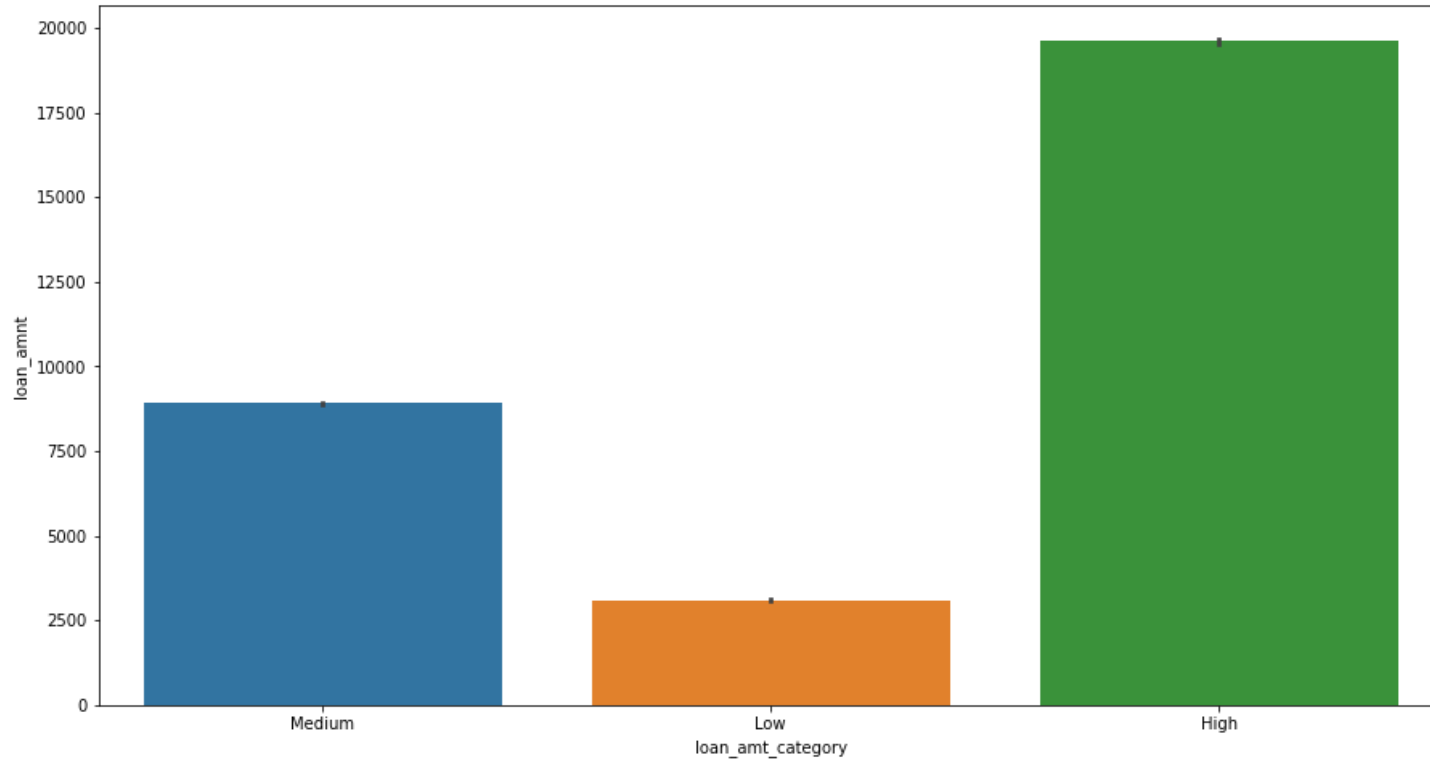
Majority of the loans are in the High Category where amount is greater than 15000.

Range of the Buckets

Low < 5000

Medium – 5000 to 15000

High > 15000



Segmented Univariate Analysis

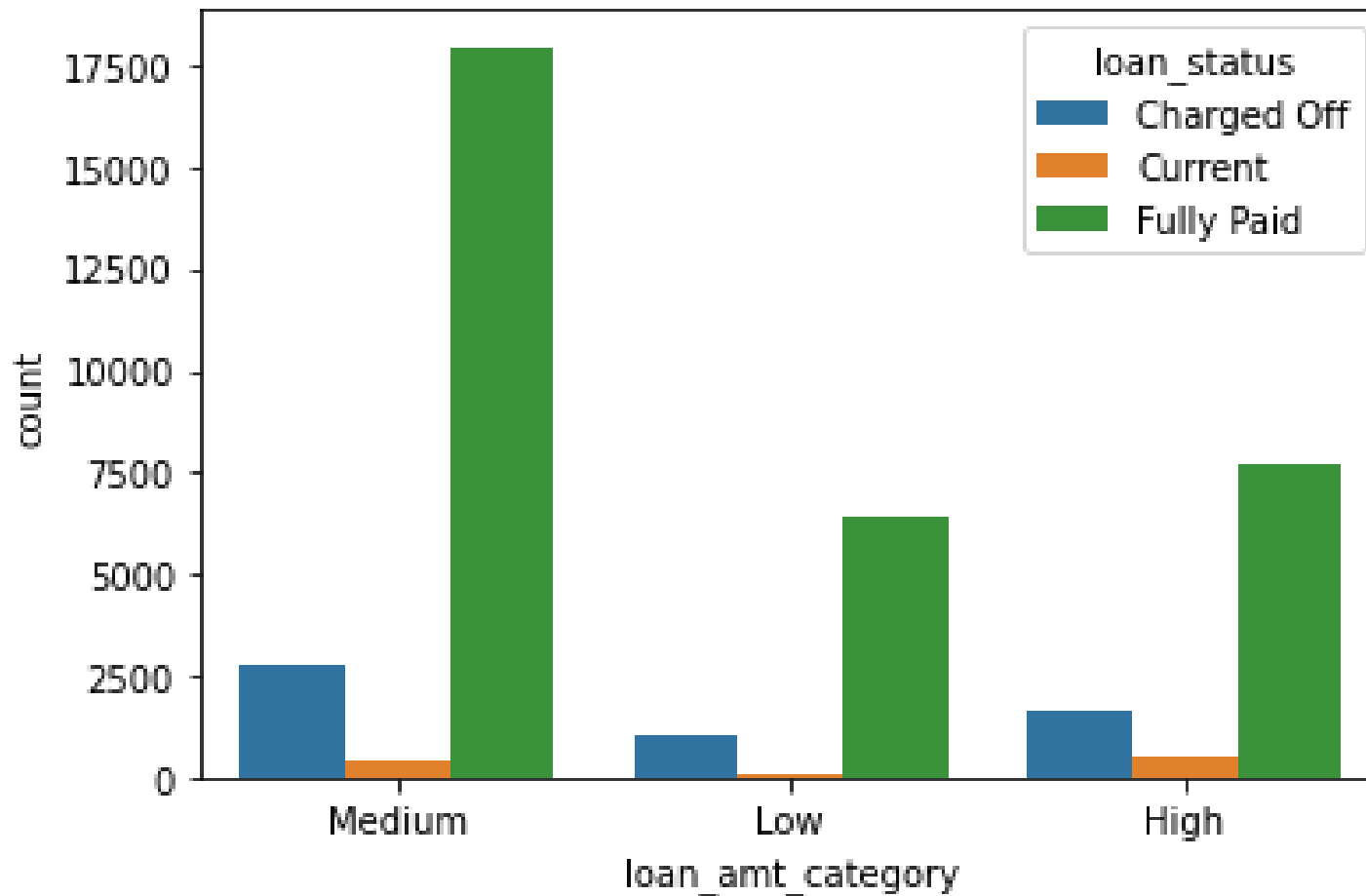
Analysis of single variables.

Does not involve relationship with any other variable

Descriptive in nature

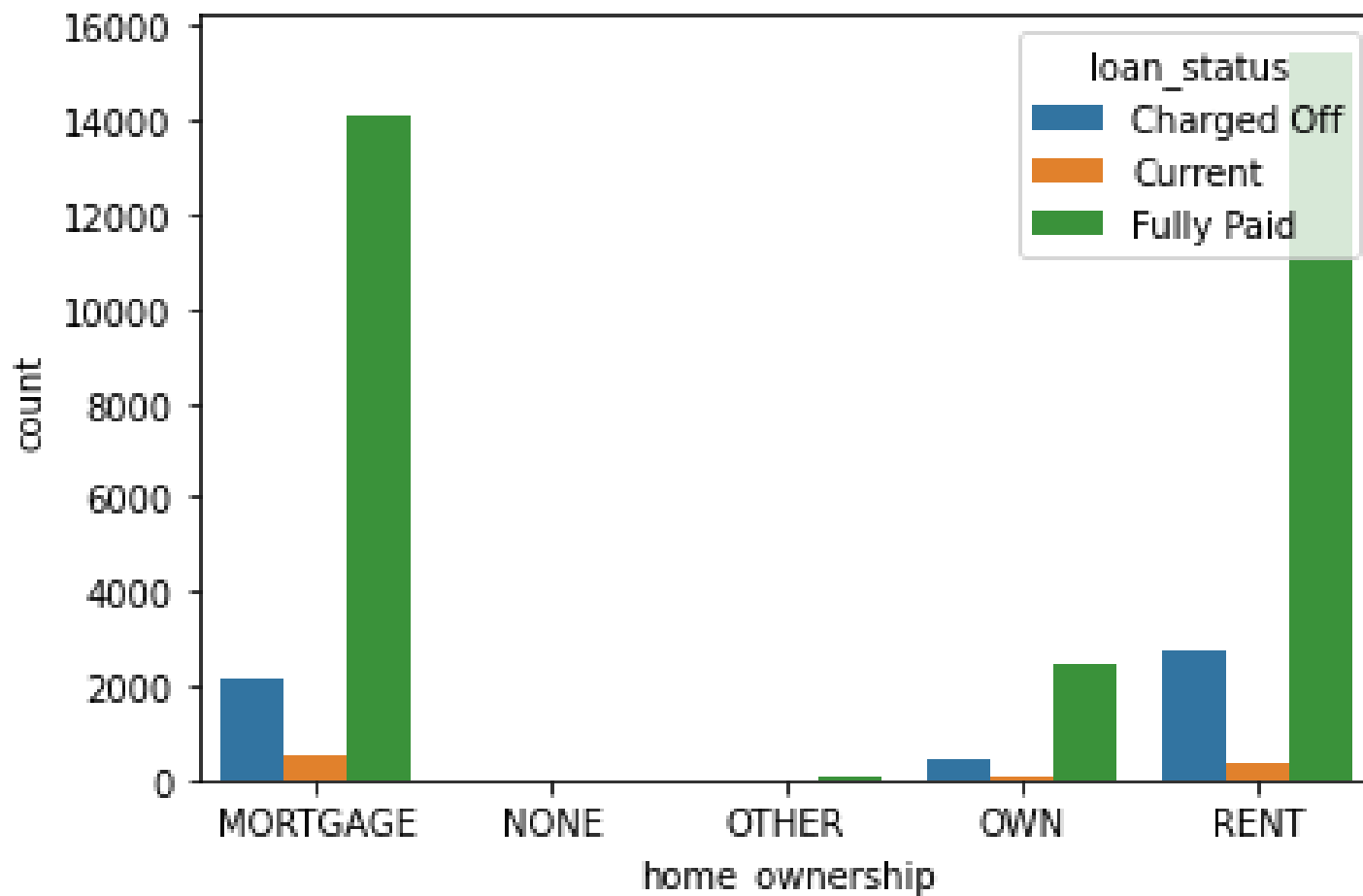
Dataset is analyzed in subsets.





Loan Category vs Loan Status

It's noted that there is a high Charge off for loans falling under Medium Loan Category i.e Above 5000 and Under 15000.

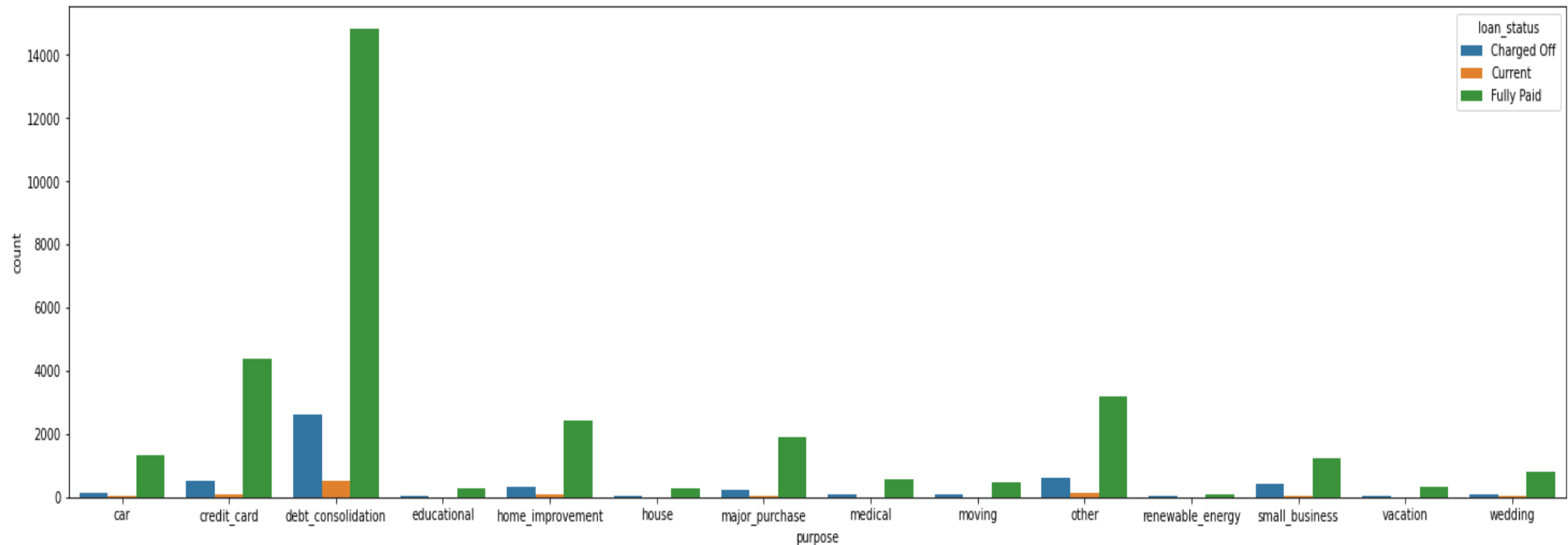


Home Ownership vs Loan Status

Applicants on Mortgage and Rent have the highest Charge off.

It appears that homeowners also default on loan repayment and are charged off however the number seems to be quite low.

Loan Status vs Purpose



It can be observed that even though debt consolidation has the highest fully paid applicants, it also has the highest charge offs.

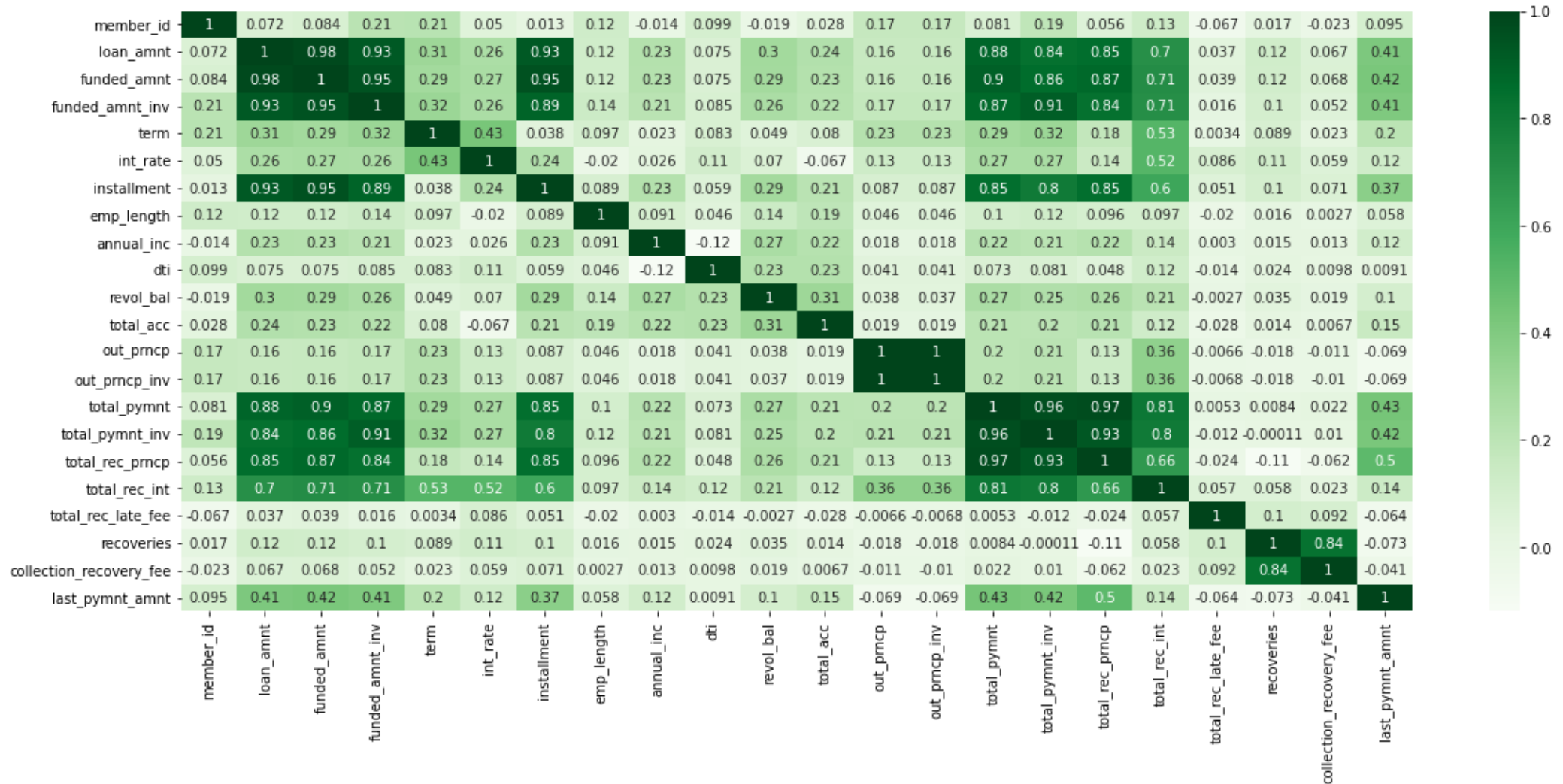
Bivariate Analysis

Analysis of two or more variables.

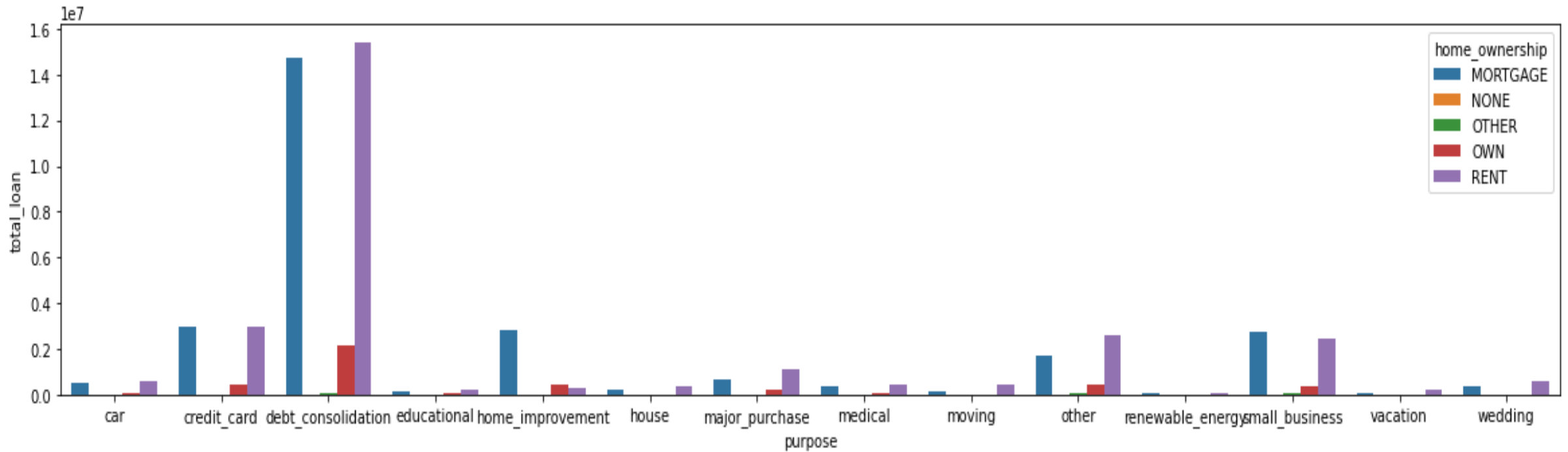
Includes relationship with other variable



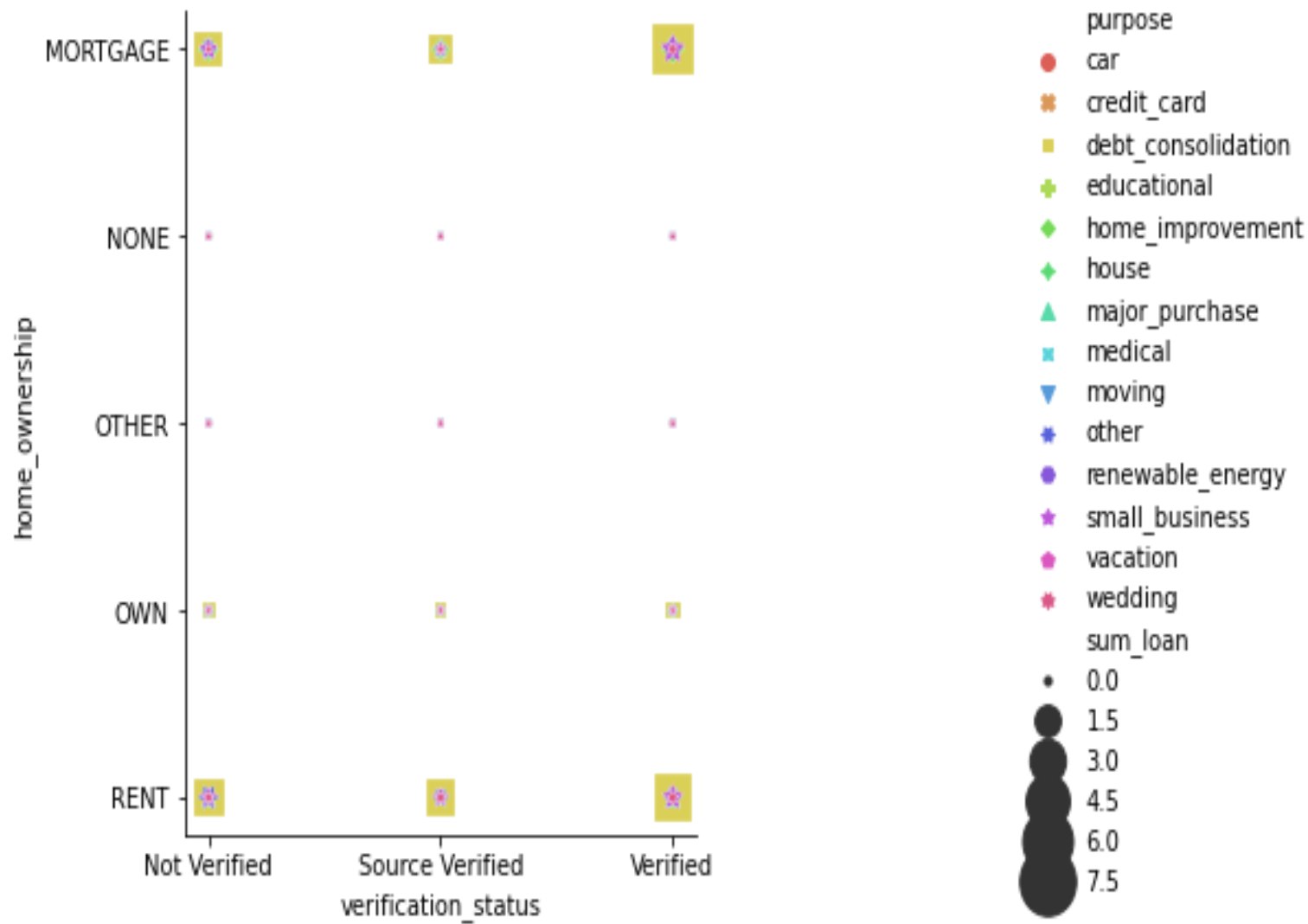
Correlation Heatmap



Purpose Vs Homeownership



A huge spike is observed in the total loans that have been waived is for people who do not own a house and borrowed money for debt consolidation

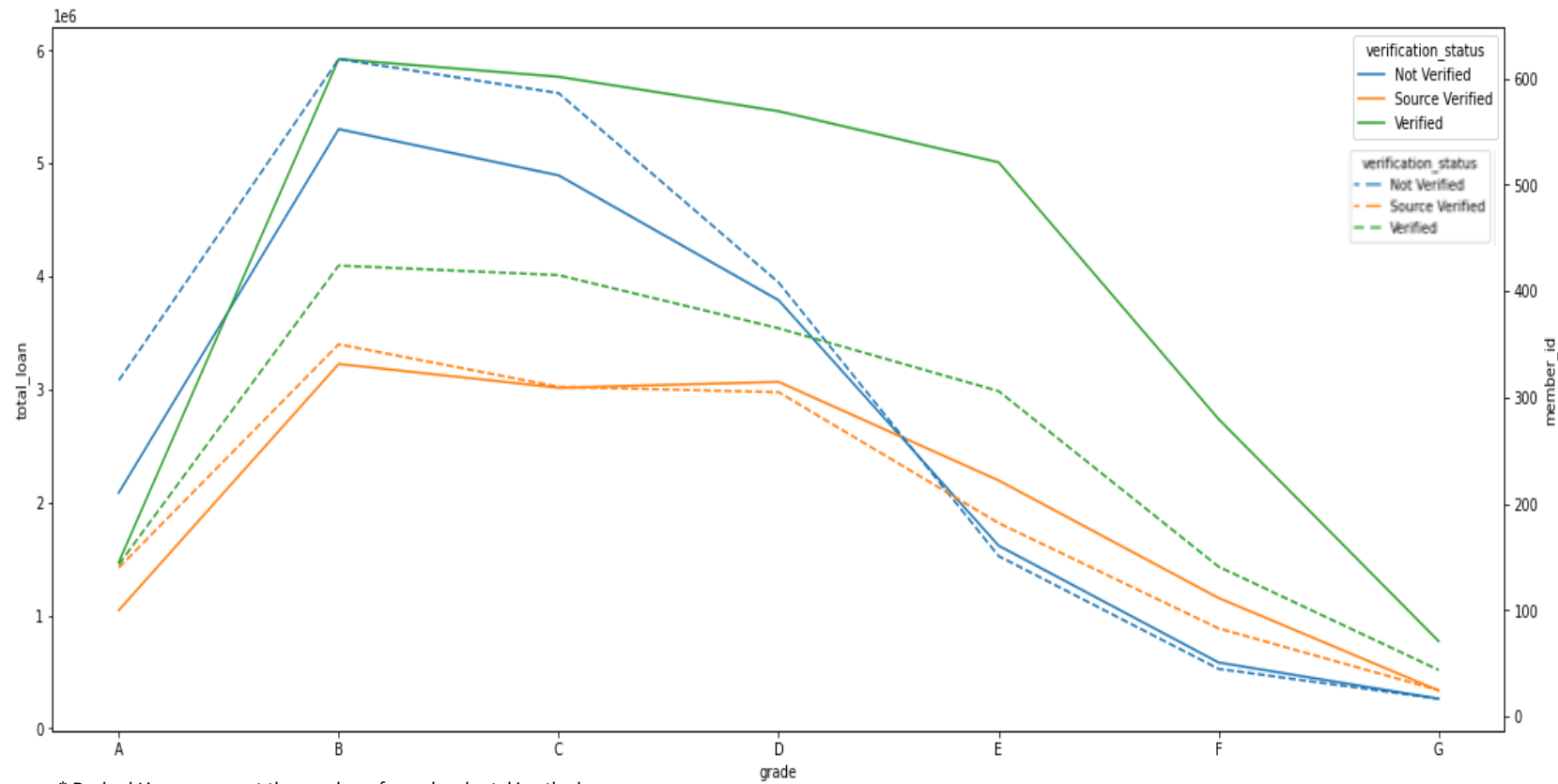


Homeownership Vs Verification and Purpose

A high amount of loan is being charged off from customers who are verified and have either mortgaged their home or customers who stay in rented homes. The same is true for customers who take loans for debt consolidation.

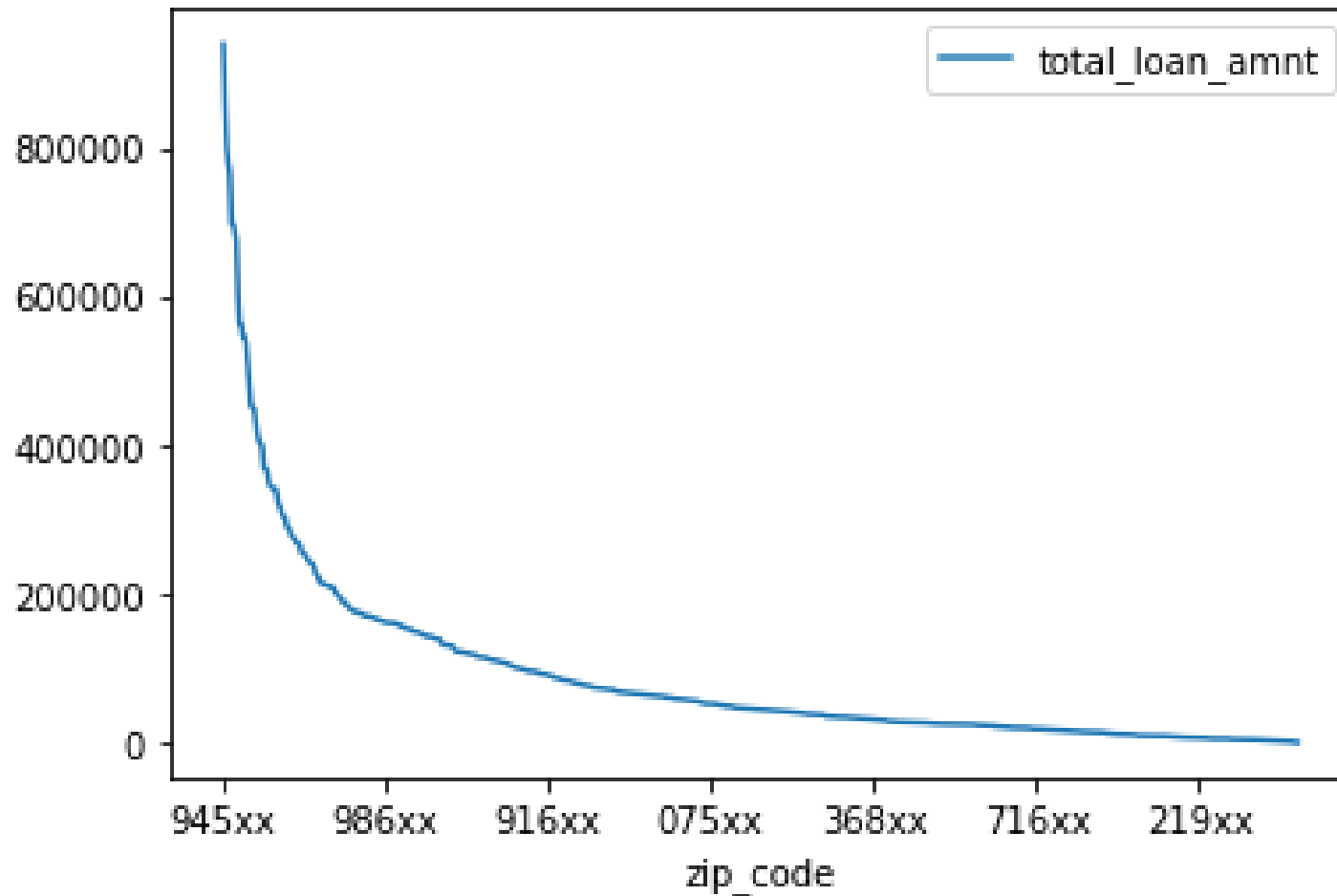
Note: The size of the points on the plot are proportional to the total sum of the loan and relative to the purpose for which the loan was taken. Each shape and color represent various purposes as can be seen in the legend.

Grade Vs Verification



* Dashed Lines represent the number of people who taking the loans
Solid lines is sum of the total loan amount

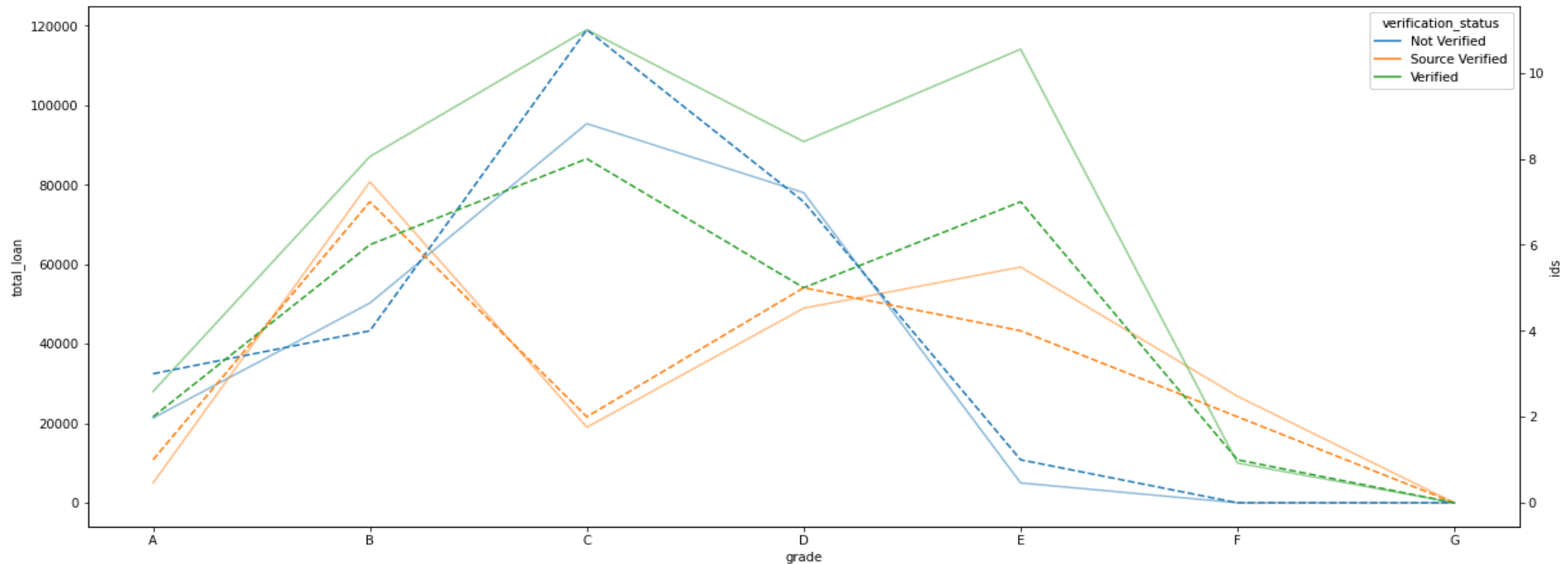
The total amount of charged off loans are for grade B and grade C given to Verified sources



Zip Code and Total Sum of Loan Amount

A huge spike is observed for people who reside in the zip code "945xx".

Loans in Zip 945xx vs Sum(Loan Amt) and Verification Status



Contrary to what was observed in the overall trend, the most charged off loans for zip code 945xx was for grade C and grade E. Grade G loans were not granted to people residing in this zip code

Summary (1/3)

The loan amount ranges from 0 to 35000 with the mean being an amount of 10000. The dataset had a few outliers and any amount above 29250 was considered to be the outlier and removed from the dataset under consideration to get an even distribution.

It was observed that a majority of the loans have been fully paid and a substantial number were charged off.

The dataset offered grades (A to G) for each loan with G being the premium loans (High loan amount).

Given the fact, it was observed that most of the loans fall under grade A, B and Sub Grade A4 to B5. Hence most of the loans are low graded loans i.e Low loan amount.

Summary (2/3)

A quick analysis of the identified data suggested the following:

1. Grade B has the highest number of applicants
2. A high number of the applicants live in rented and mortgaged houses. A few own a house.
3. Majority of the loans are fully paid.
4. Debt consolidation seems to be the major purpose of taking a loan.
5. More than 20000 of the applicants had credit inquires over the last 6 months.
6. A large population has more than 3 open credit lines
7. Income group 40000-60000 has the highest number of applicants in the dataset.
8. Majority of the loans are in the High category.
9. It was noted that there is a high Charge off for loans falling under Medium Loan Category
10. Applicants on Mortgage and Rent have the highest Charge off.

Summary (3/3)

- 11. B grade loans tend to be charged off for verified sources
- 12. California (CA) has the highest number of applicants.
- 13. A huge spike is observed for people who reside in the zip code "945xx"

Conclusion

A huge spike was observed in the total loans that have been waived is for people who have who do not own a house and borrowed money for debt consolidation and a high amount of loan is being charged off from customers who are verified and have either mortgaged their home or customers who stay in rented homes. The same is true for customers who take loans for debt consolidation.

Contrary to what was observed in the overall trend, the most charged off loans for zip code 945xx was for grade C and grade E. Grade G loans were not granted to people residing in this zip code