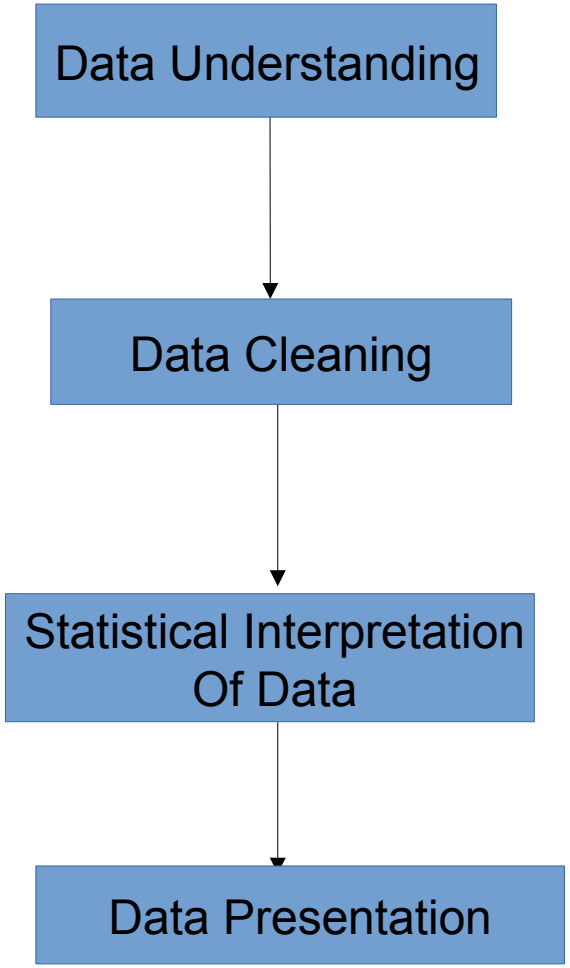
Gramener Case Study

SUBMISSION
Jayanth B
APFE187000049

Group Name:

1. Parthasarathi Panda
2. Sambunath Jena
3. Sumit Gupta
4. Jayanth B

WorK Flow


Data Understanding
Data Cleaning
Statistical Interpretation Of Data
Data Presentation

Case Study Objective

- Identification of Loan Applicant patterns that tend to ‘default' Loan amount
- Understand the ‘Driving Factors’ or ‘Driver Variables’ behind Loan Default phenomena
- Gramener may choose to utilize this knowledge for its portfolio and risk assessment of new loan applicants

Data Cleaning

- Identify all columns that don't provide any value. E.g. the url column: It provides us a link to source of the data.
- Remove all columns that need Text Processing.
- Identify all columns that don't have any other value other than NA and 0. Remove such columns.
- Identification of duplicate entries in columns with discrepency.
- Example: In emp_title column names. E.g. ARMY and US ARMY are same. Walmart, WALMART, Walmart and WAL-MART are same.
- The columns with more than 75% of the values were missing are discarded. The columns with outliers are identified through boxplots.
- Since we are dealing with the aggregate, we may not need the primary keys in this analysis such as id and member_id. We can remove these as well

- Top 5 Observations
- Uni-varate Analysis
- Bi-variate Analysis

Conclusion top 5 Observations

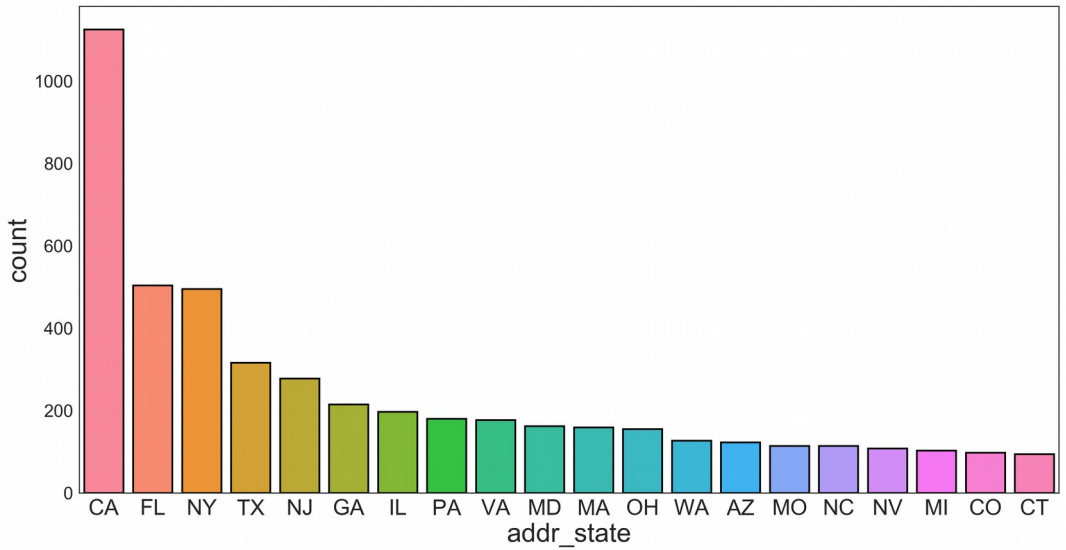


Fig1: Frequency distribution of Address of a state for charged of loans

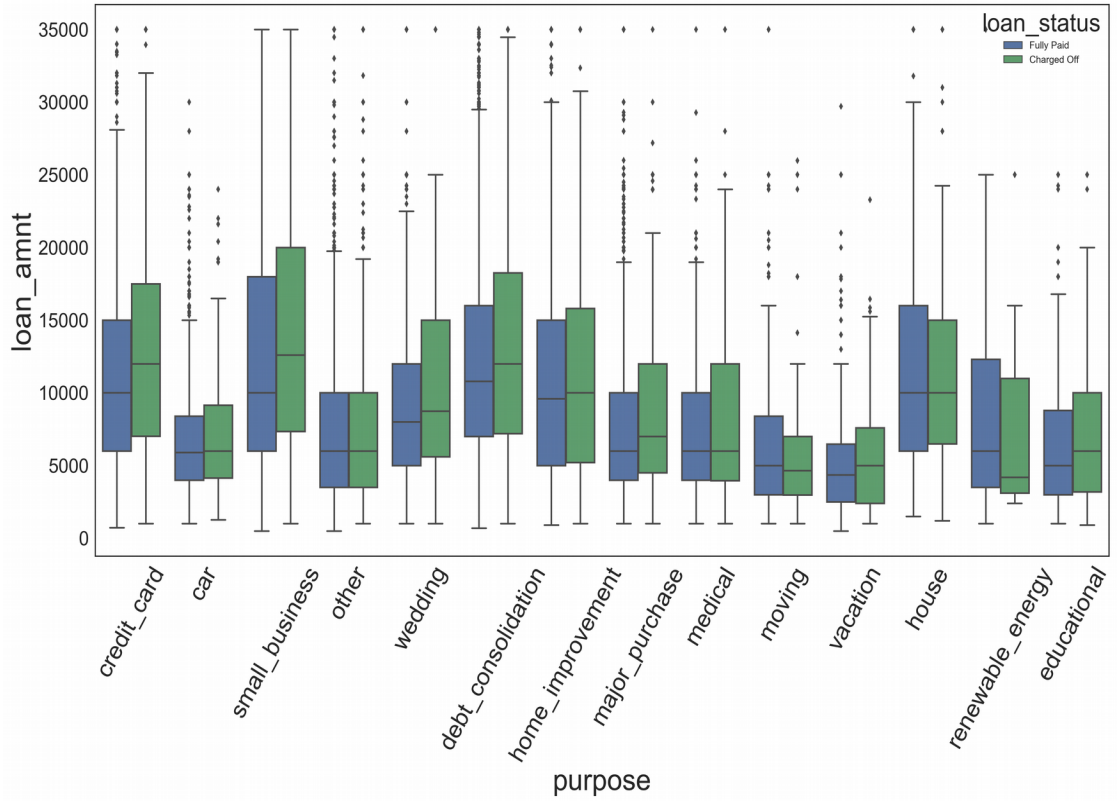
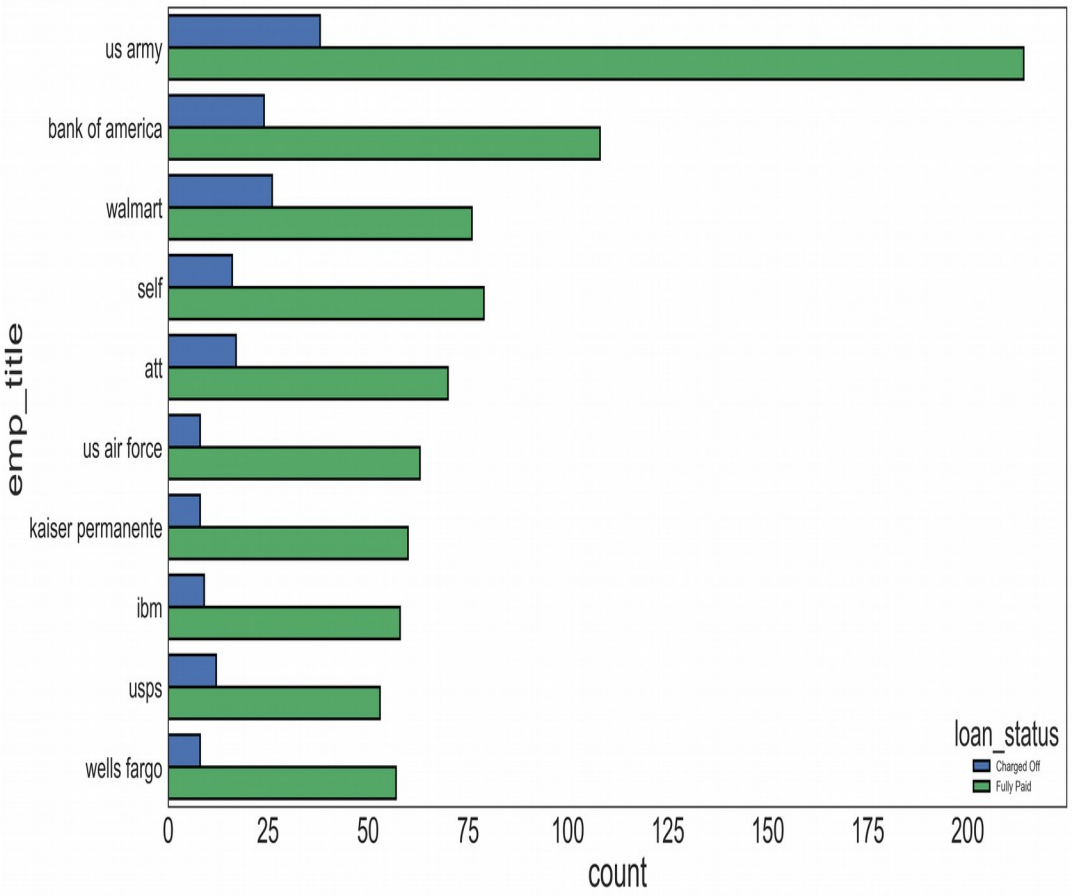


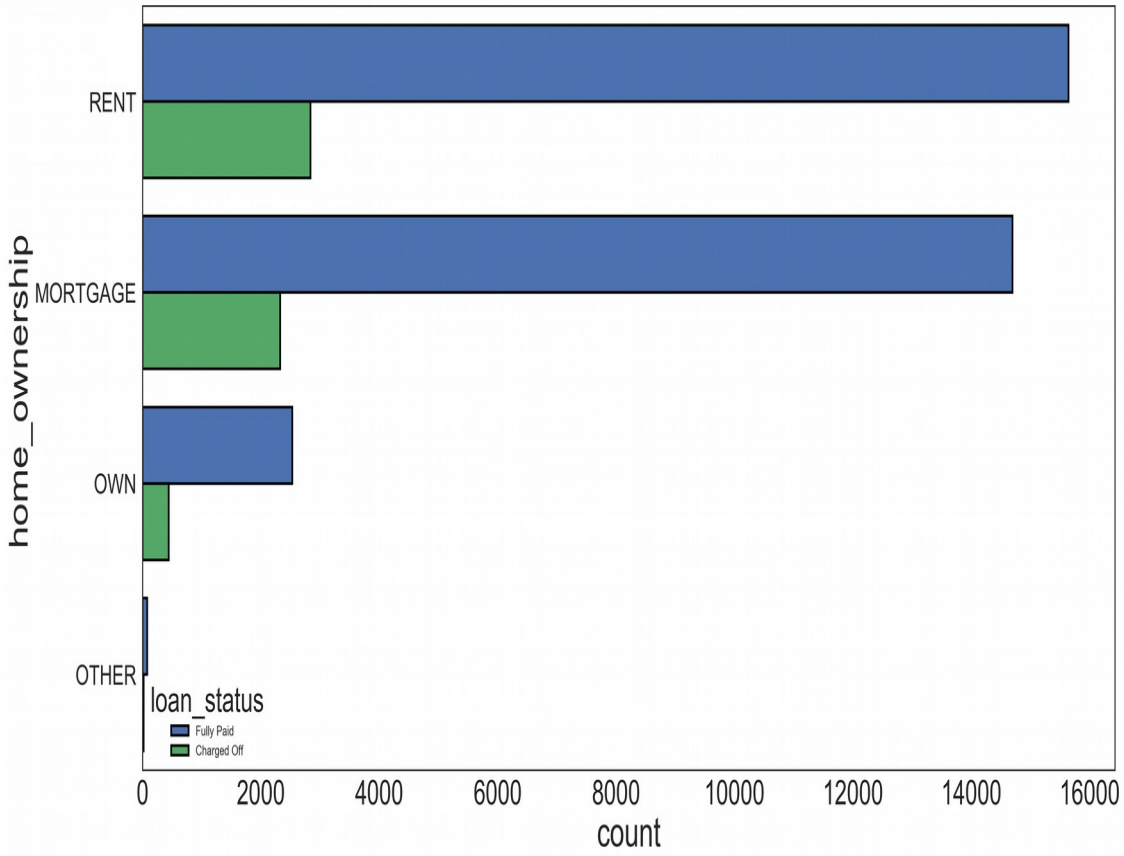
Fig2: Boxplot representation of loan purpose vs loan amount.

Inference 1: State CA(california) is the major defaulter.
Inference 2: Small business , credit card are major defaulters.

The Frequency of Employ title

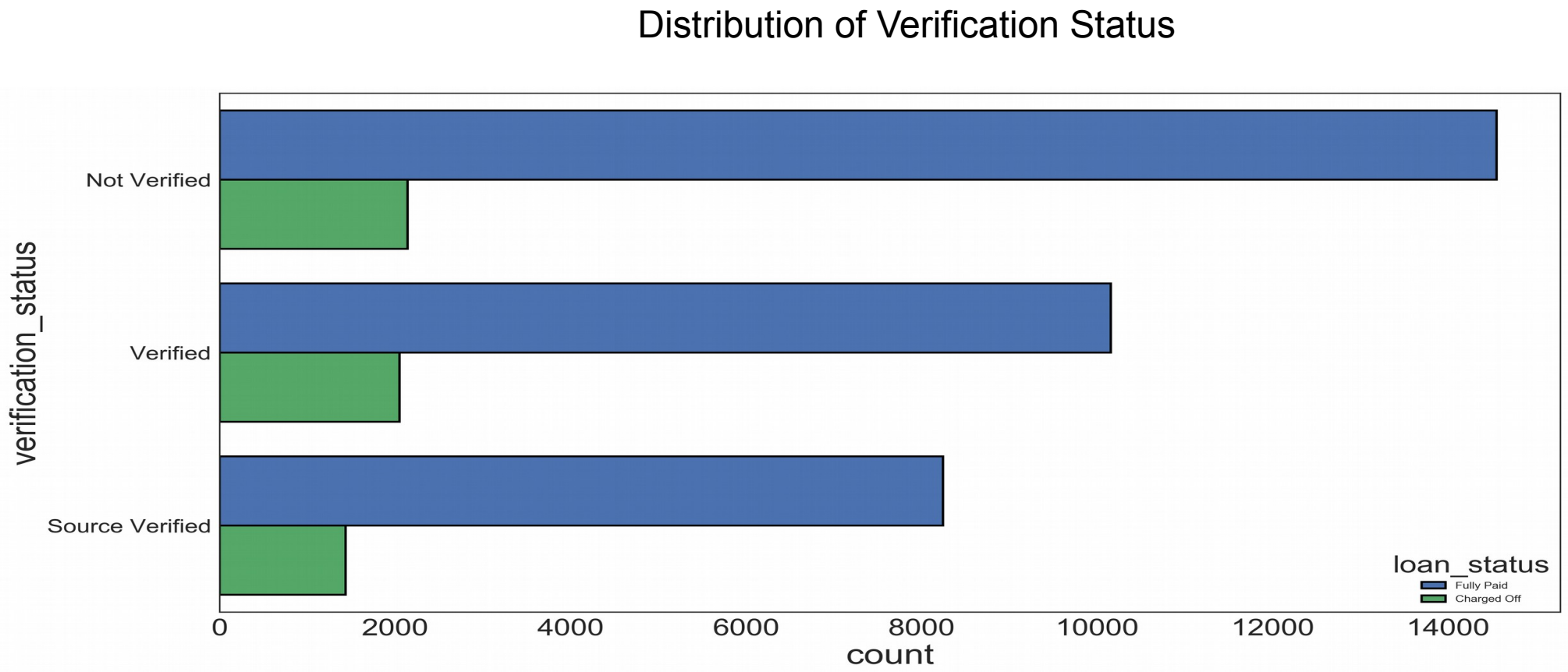


Frequency of Home ownership



Inference 3 : Most defaulters recide in rented homes. The person with own house is less likely to default, compared to person in rented house

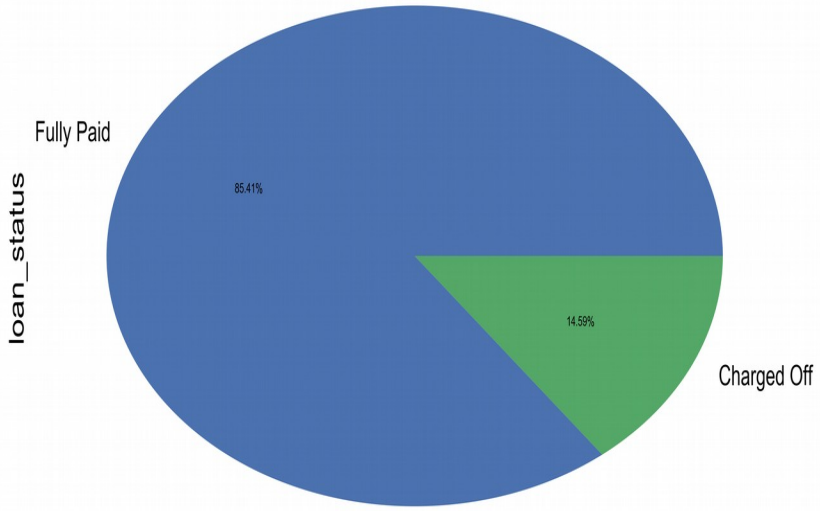
Inference 4 : Most defaulters are employed with US army.

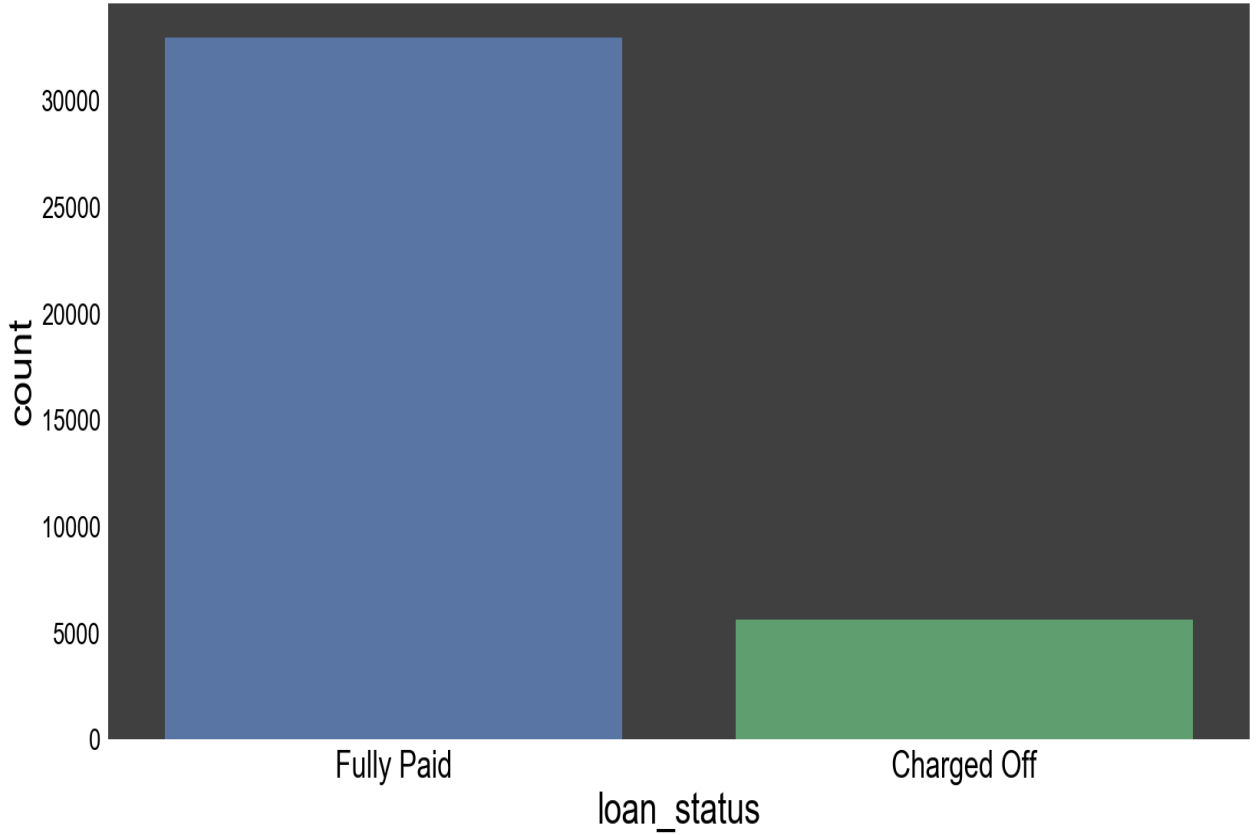


Inference 5:The percentage of defaulters is higher in “Not Verified” cases.
Suggestion 1: Verification Process needs to be reviewed and improved accordingly

UNIVARIATE ANALYSIS

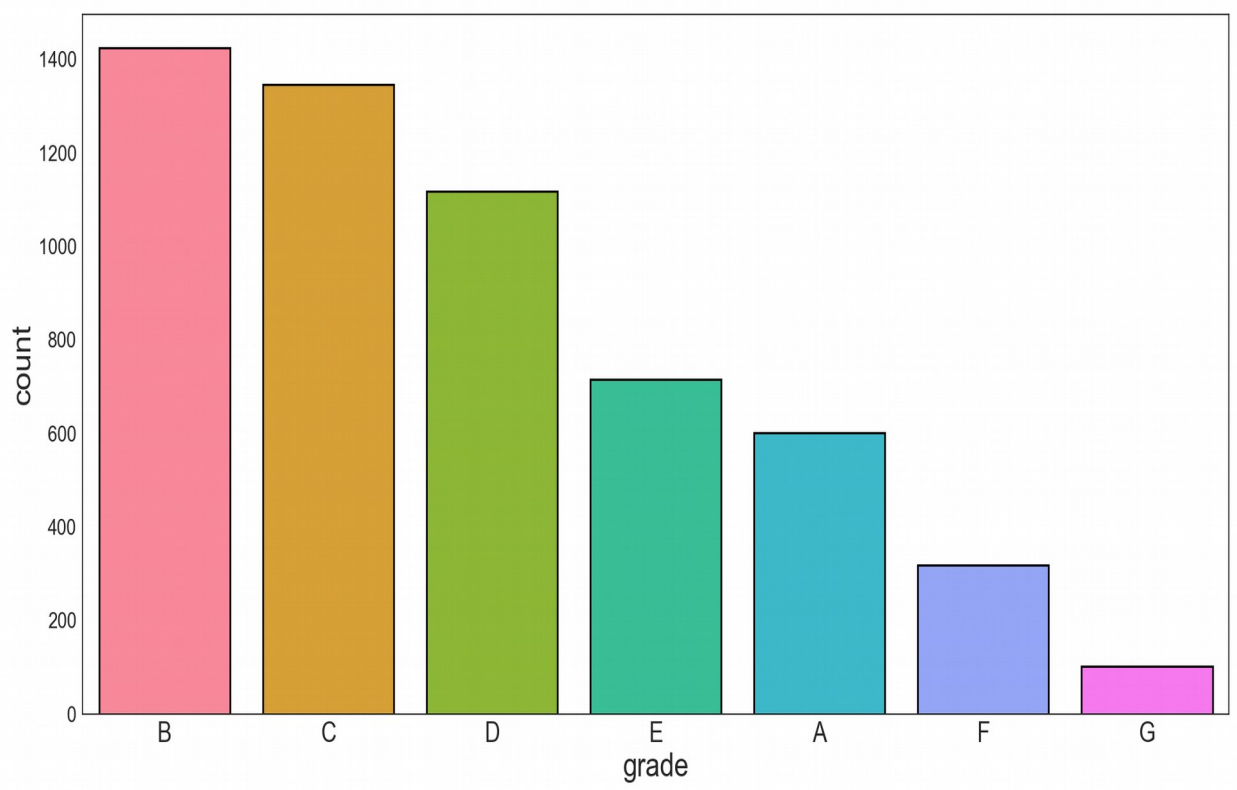
Distribution of Loan Status



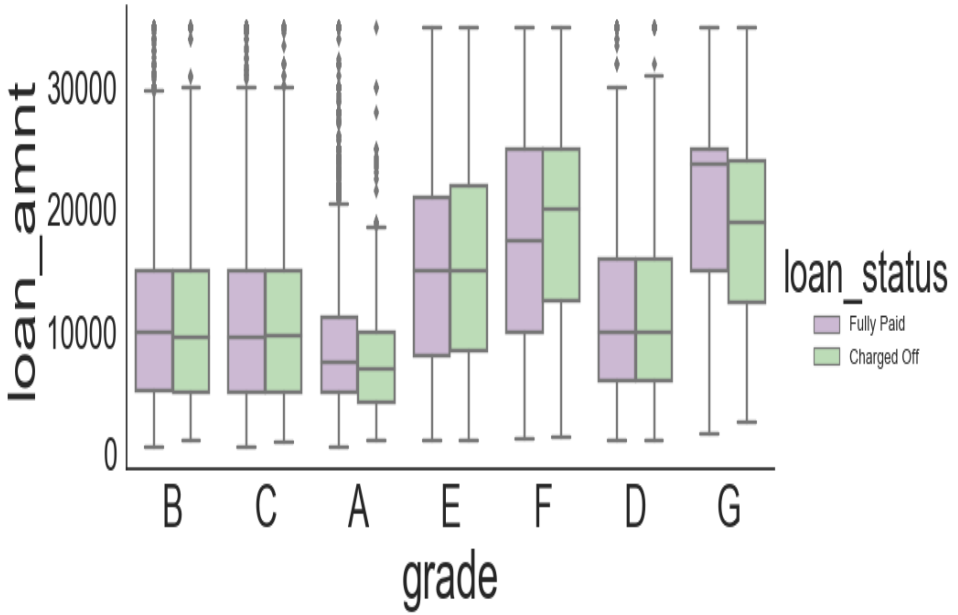


Observation: The percentage of charged off loans lower than Fully paid loans

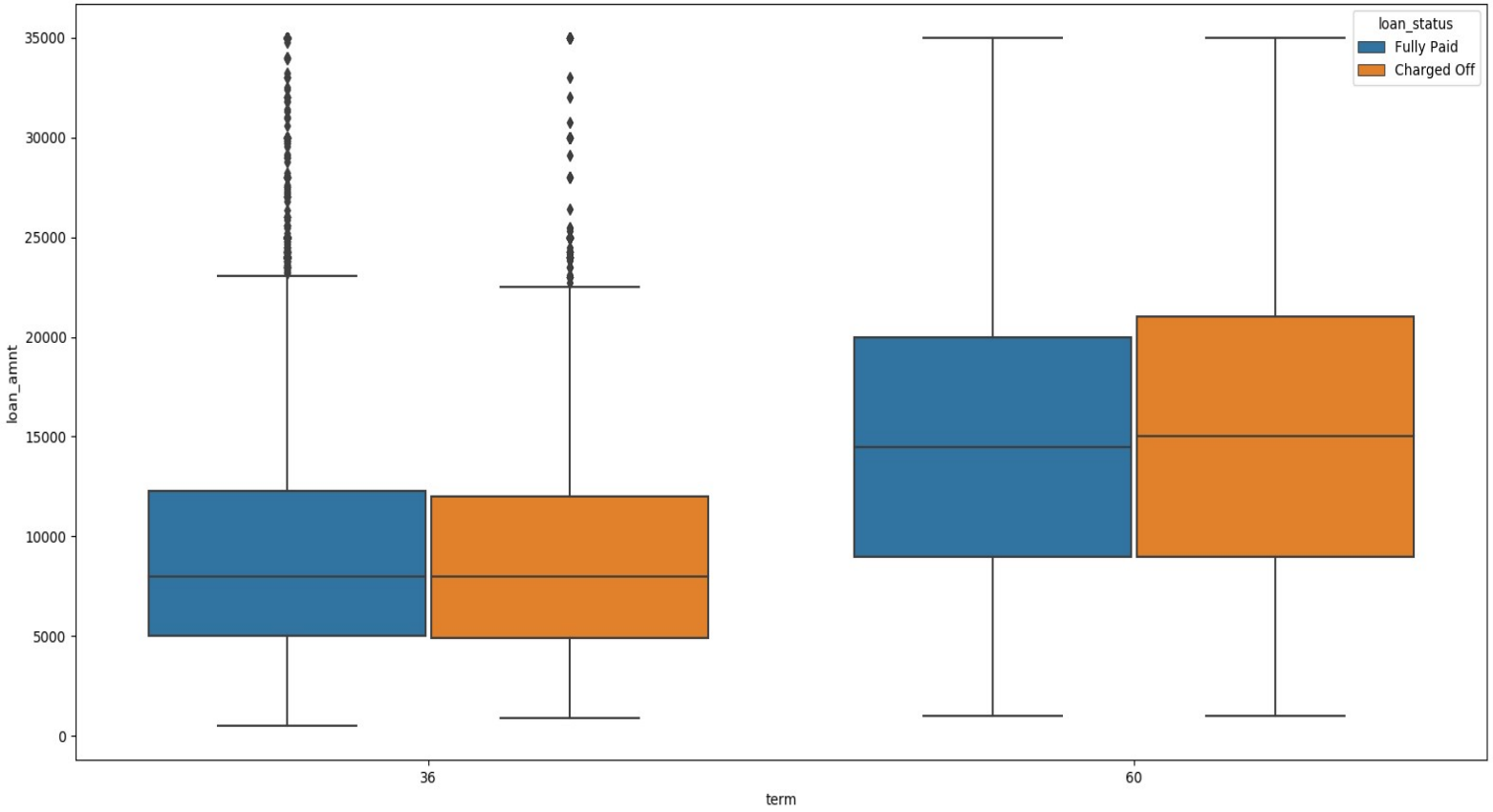
The Grade distribution in charged of loans



Comparison of grade , loan amount and Loan Status

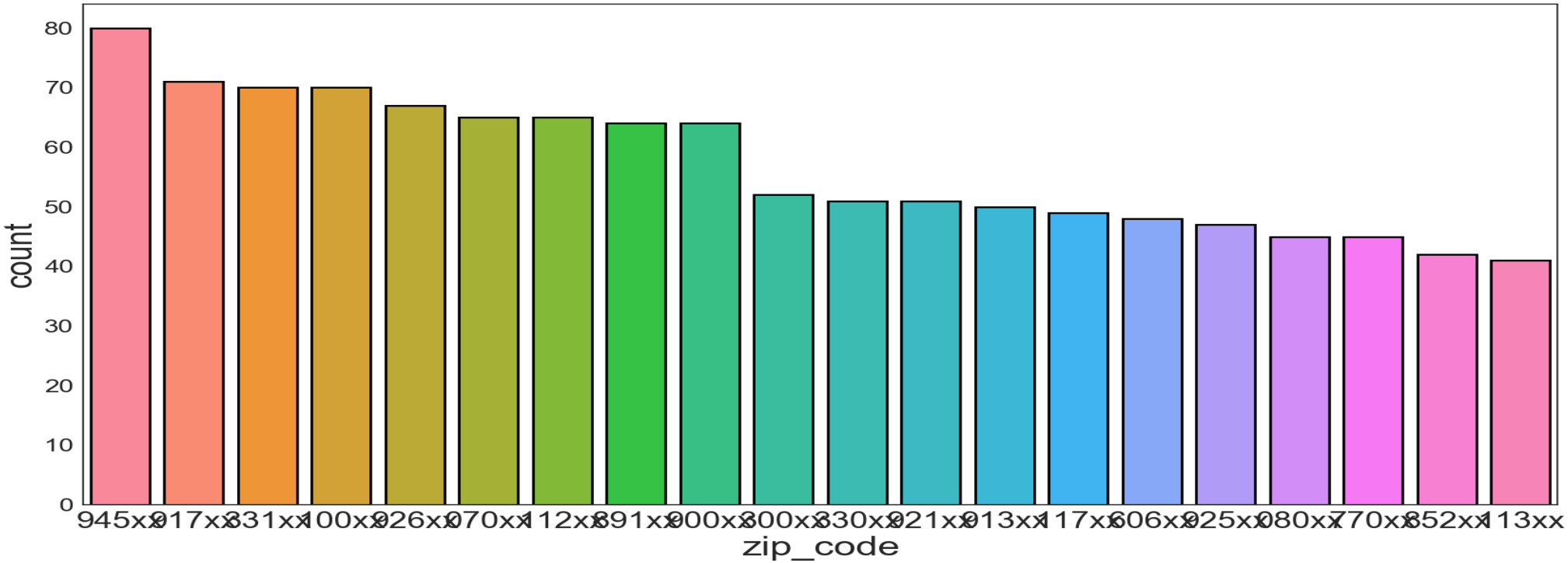


The Grade F has high default rate.



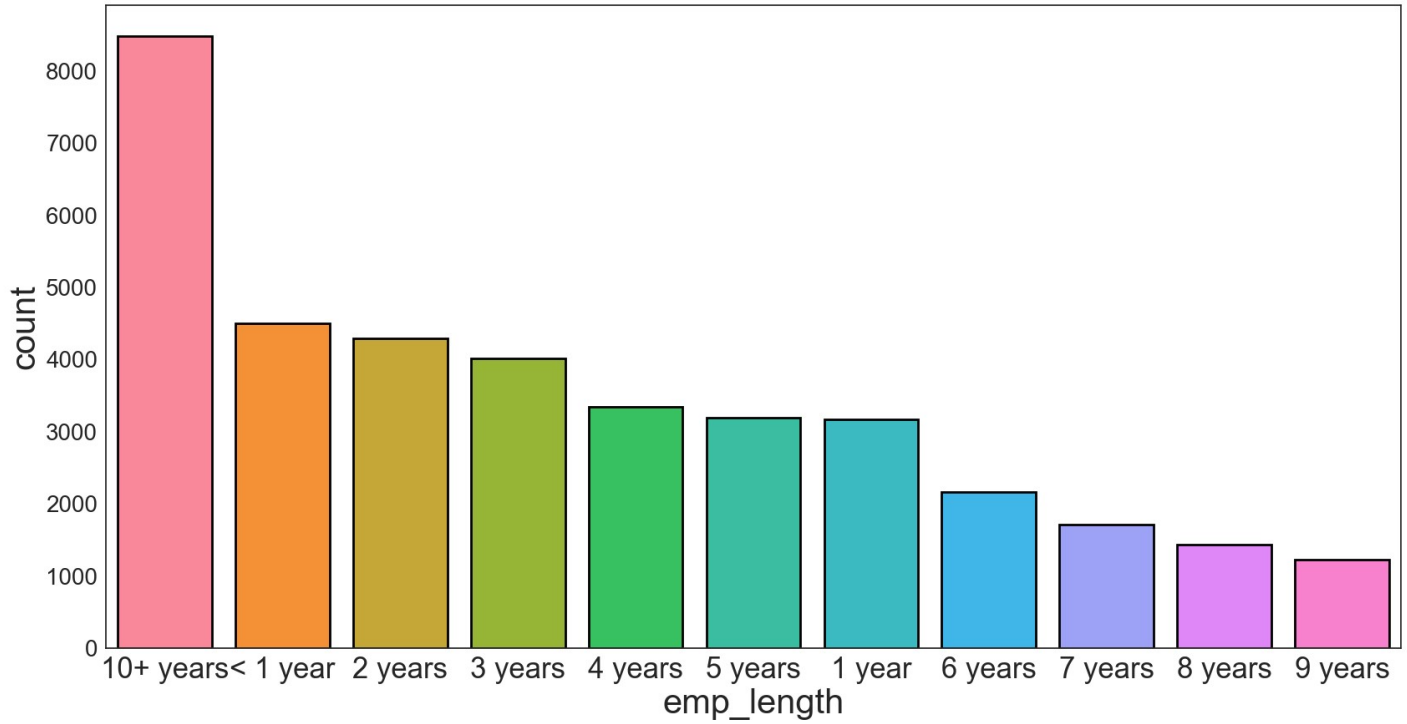
The default rate is little higher in term period of 60 months

Zip code vs charged off loans frequency



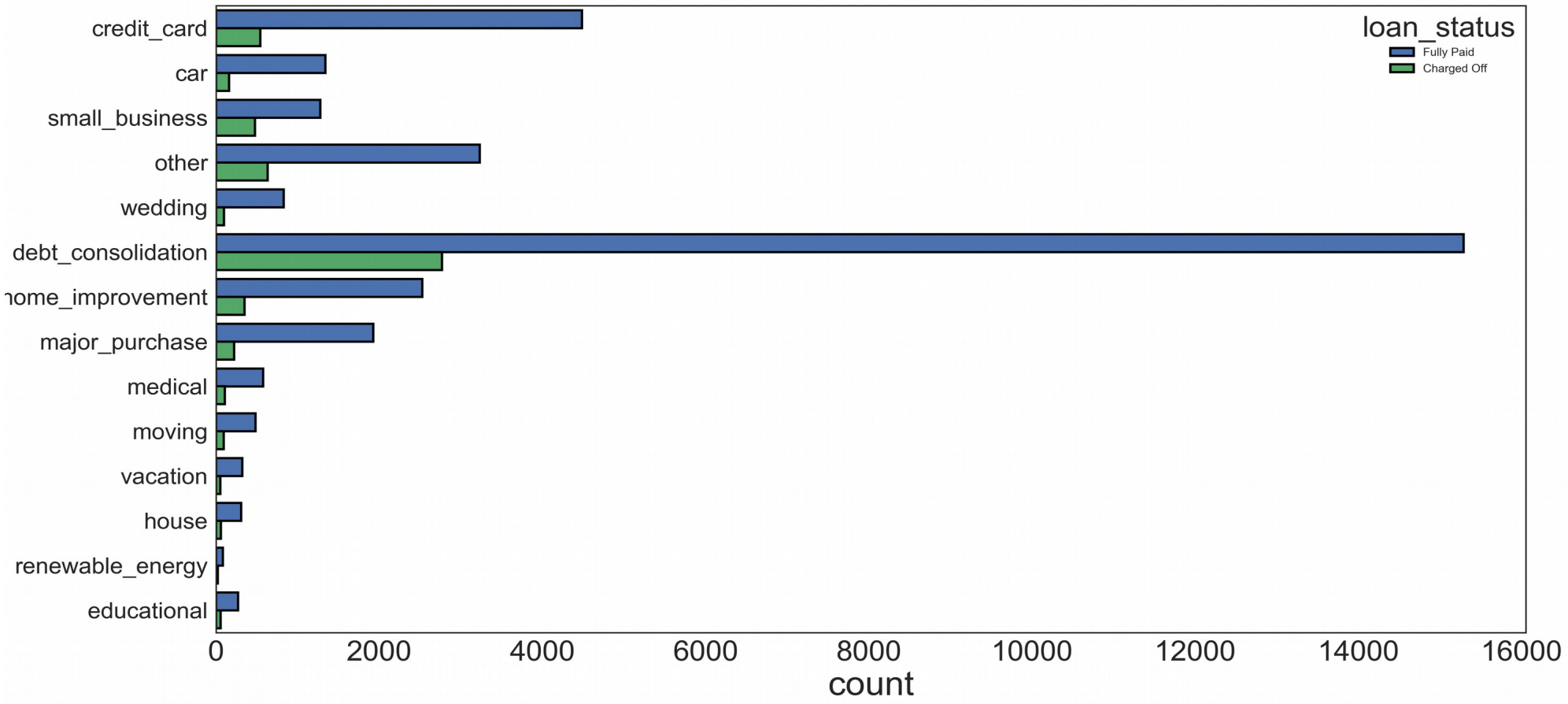
Observations: Since zip codes are in ascending order in X axis we can see that no of “CHARGED OFF” loans are particularly high in 9xxxx. This is the state of California.

Histogram of Employment length in Charged Off Loans



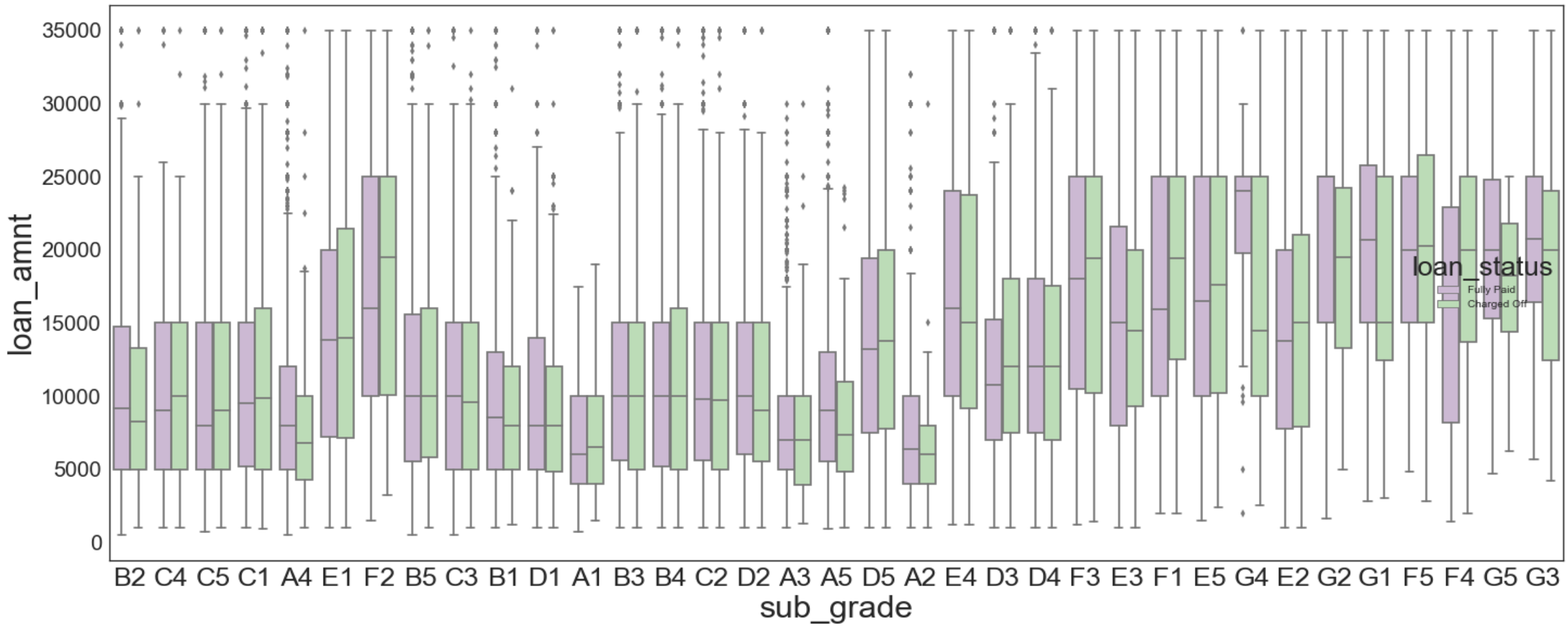
Observations: Charged off loans has a decreasing trend w.r.t tenure. However the number of Charged Off loans within the first year is high.

Distribution of purpose of Loan



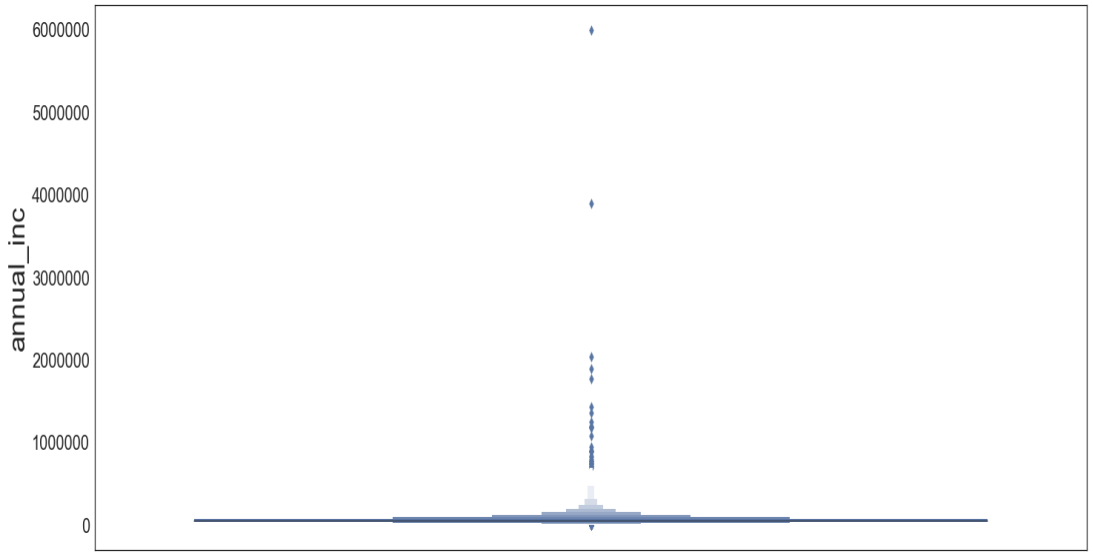
Observation: The loans from debt consolidation section is very high. Debt consolidation is much broad term. This suggests that data collection method is inadequate.

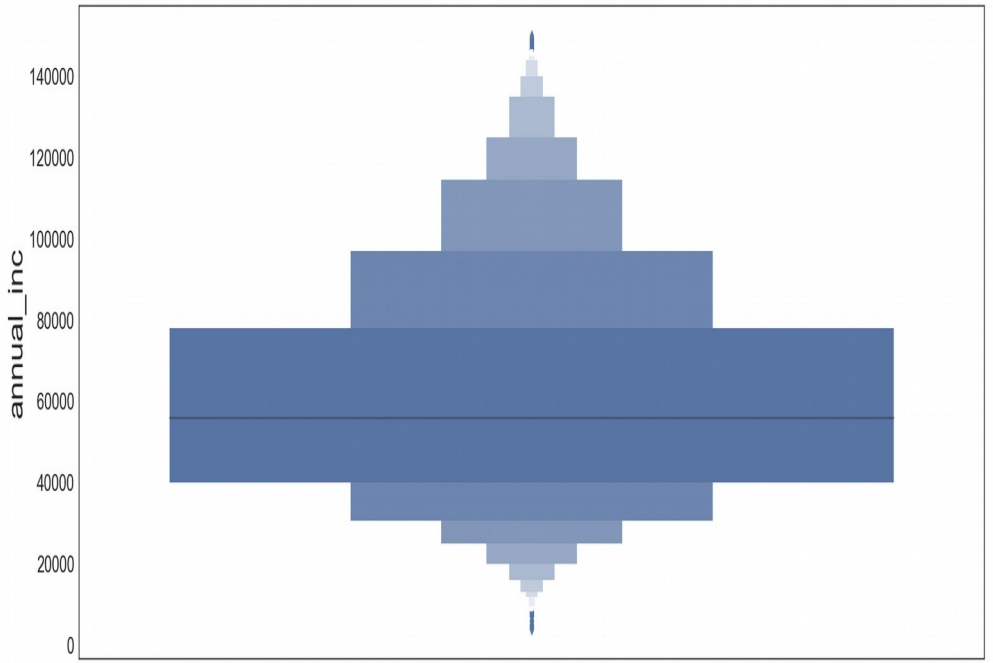
Analysis of sub grade Analysis



The Grade F2 and F4 has high default rate compared to other grades

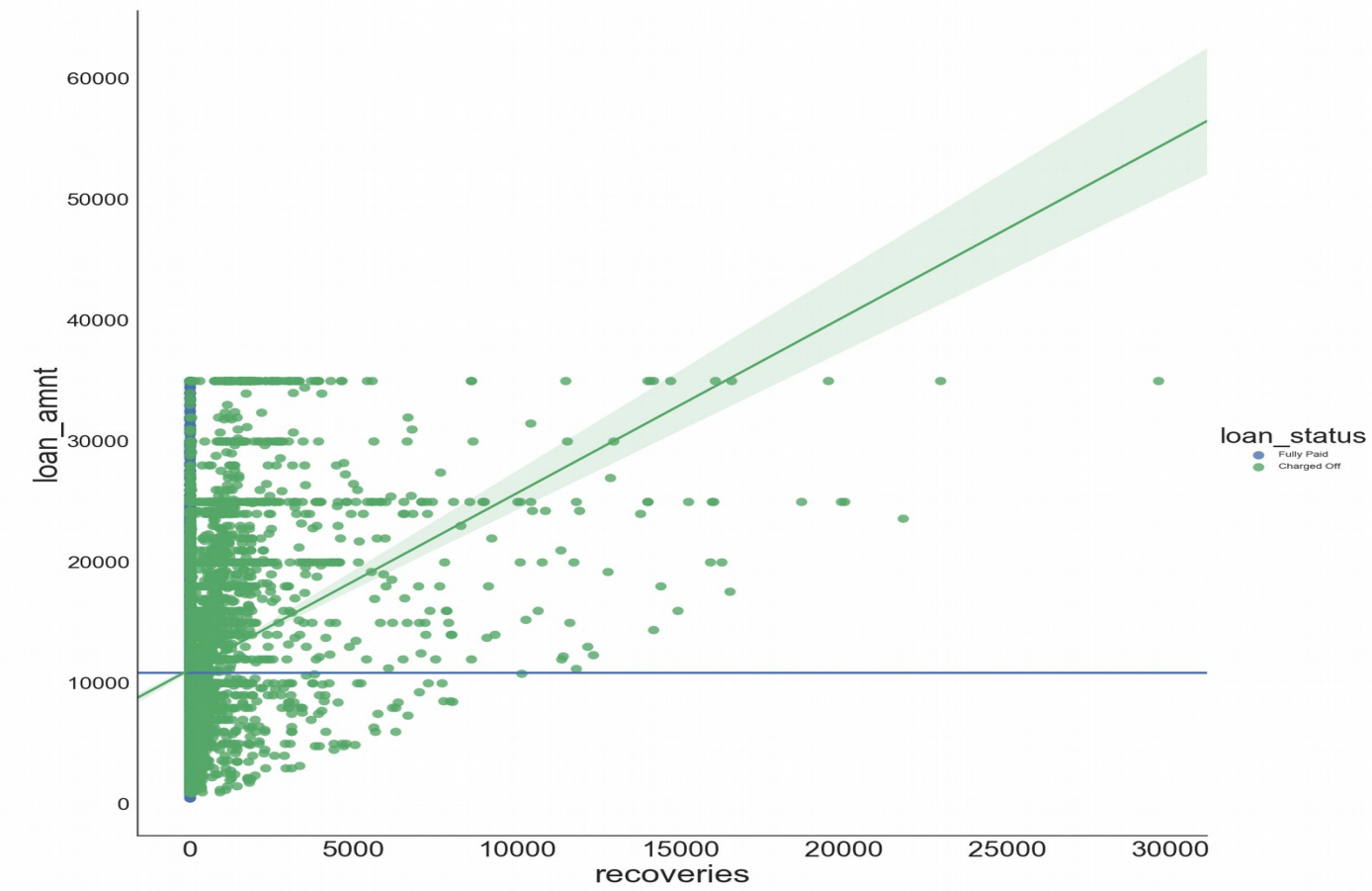
Analysis of Annual Income





Observation: The annual income has relatively high number of outliers

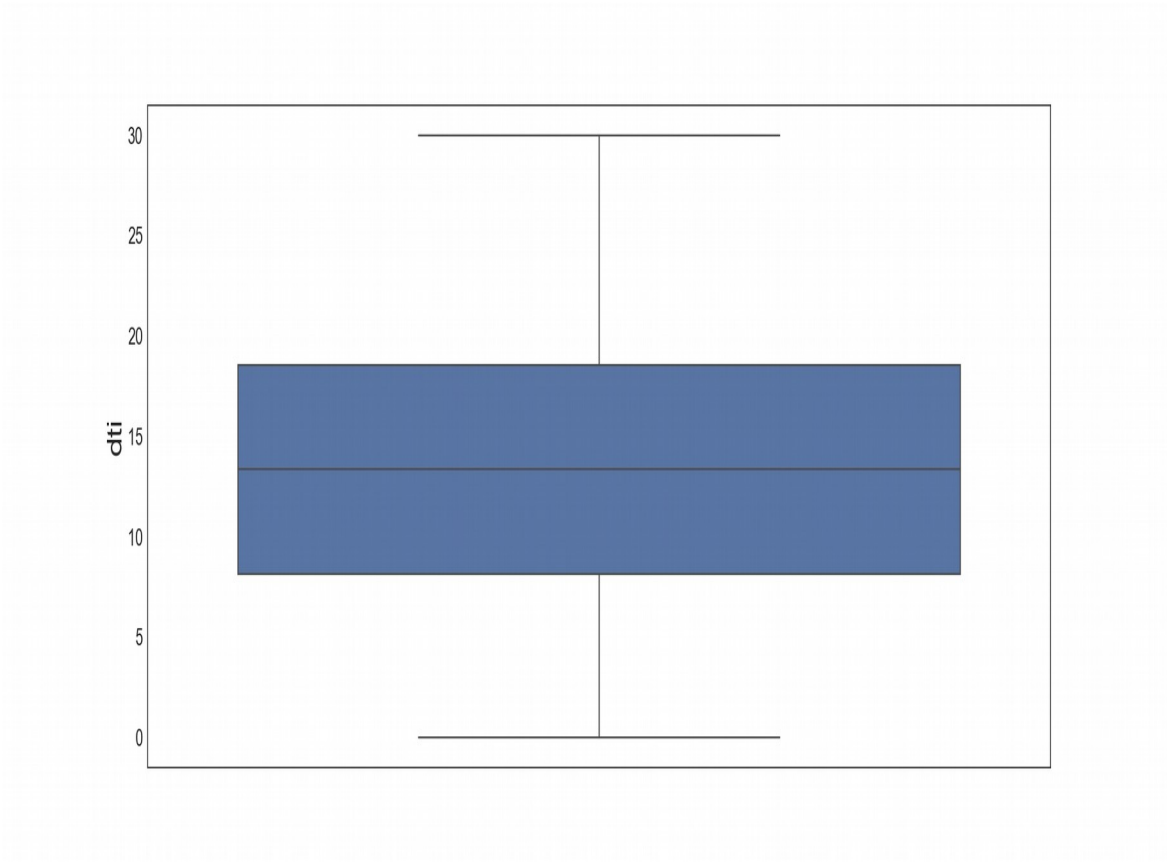
Analysis of Collection Recovery Fee



Observation: The number of recoveries in fully paid condition is almost zero.

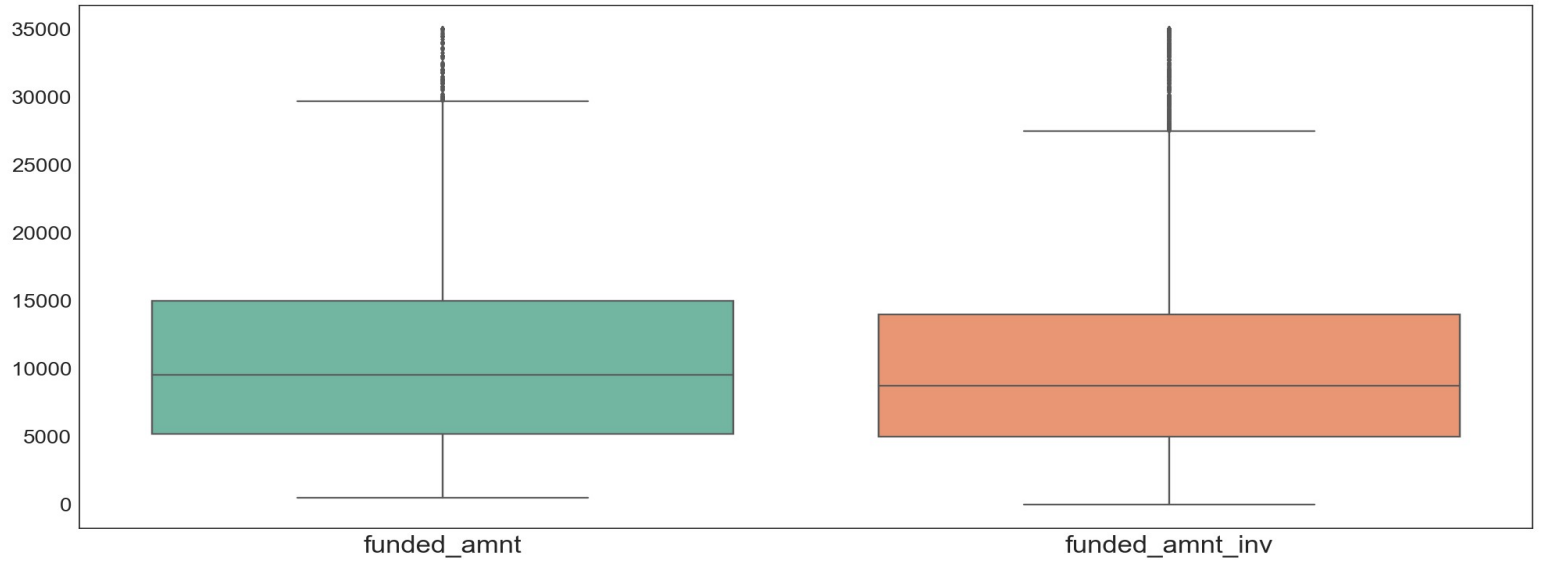
In charged of condition the number of recoveries Increases with loan amount

- Analysis of DTI column



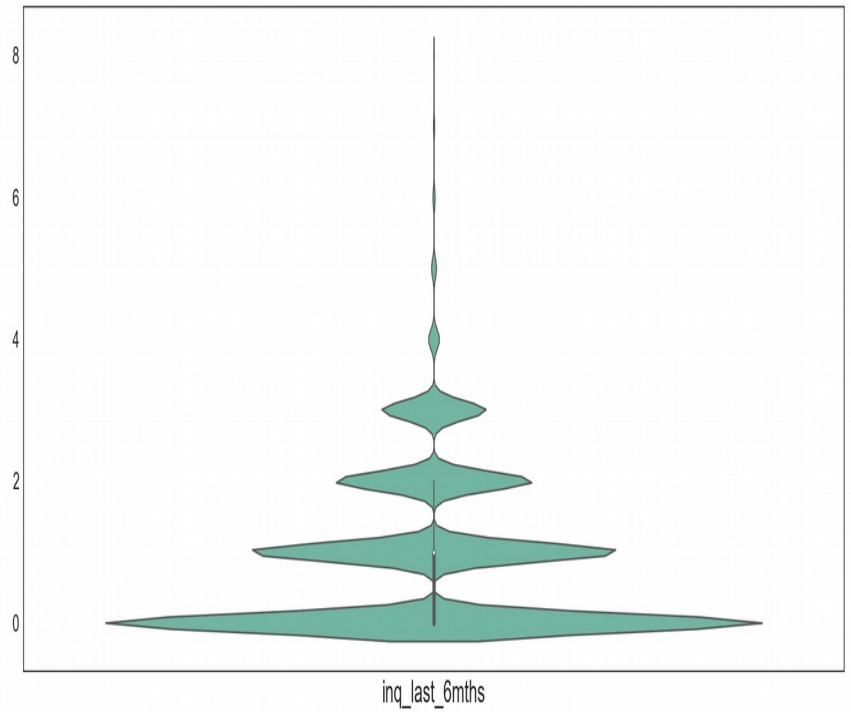
Observations: The DTI column seems to be well distributed with no outliers

- Analysis of Funded amount and Funded amount (investor)



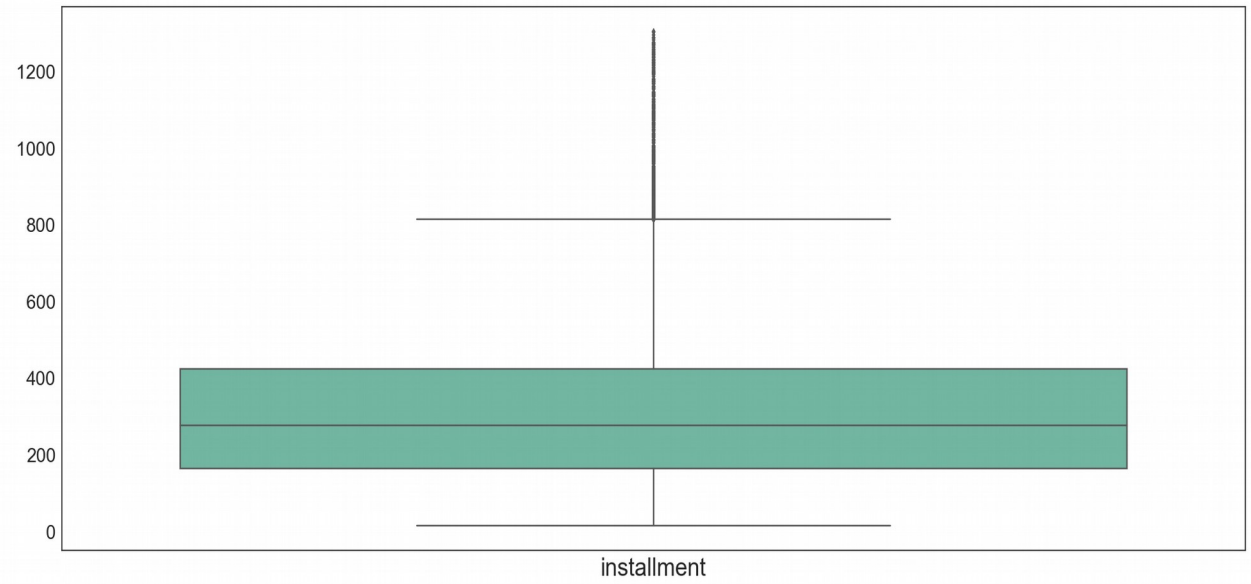
Observations: There are few outliers on top end of the distribution

Analysis of No of inquiries in last 6 months



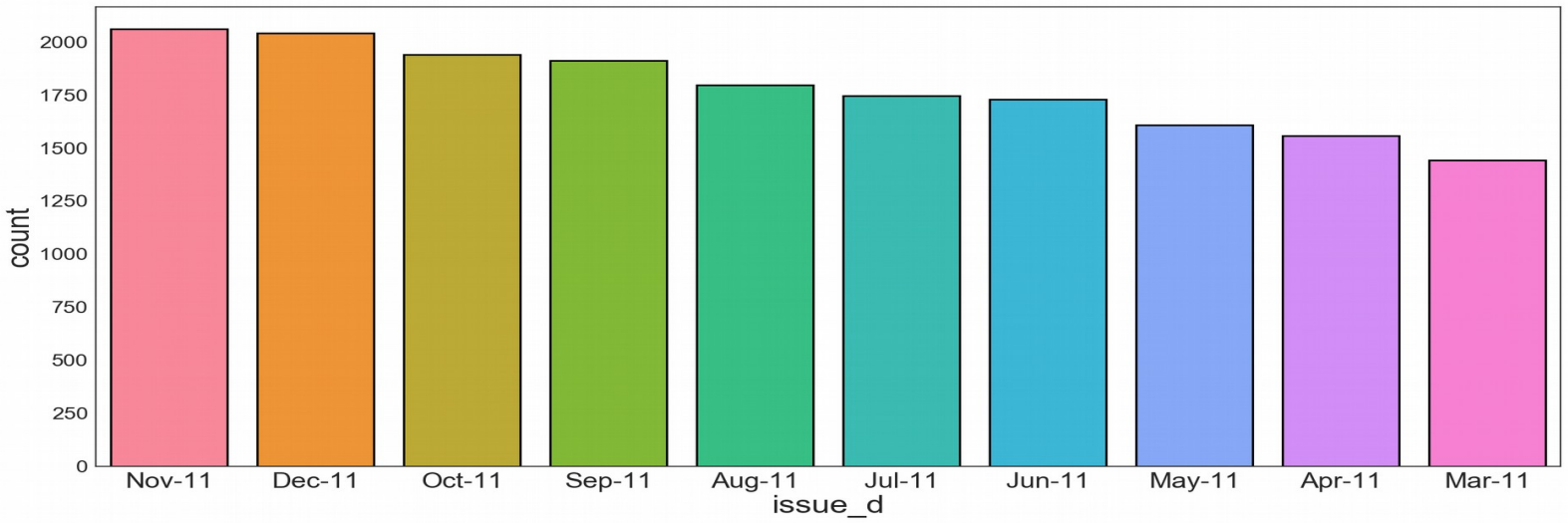
Observations: From the plots, it appears that there are too many 0 values (no inquiries in last 6 months)

Analysis of Installment



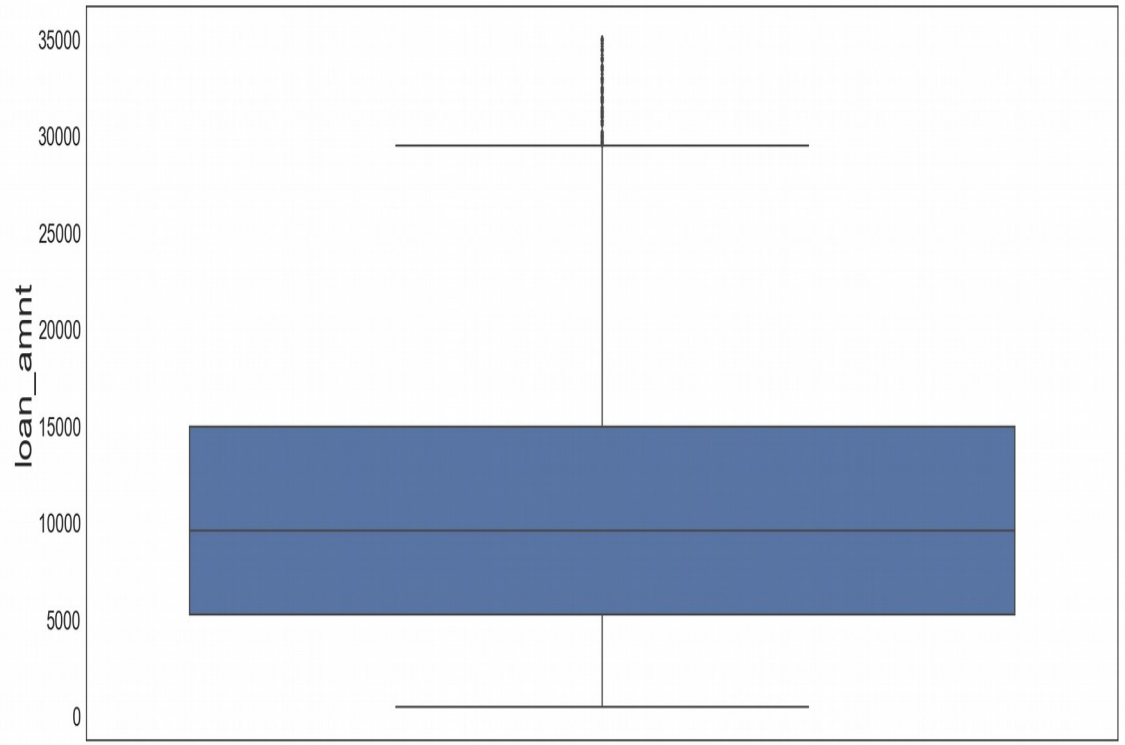
Observations: From the box plot, it appears that there are a considerable number of outliers.

Analysis of Issue Date



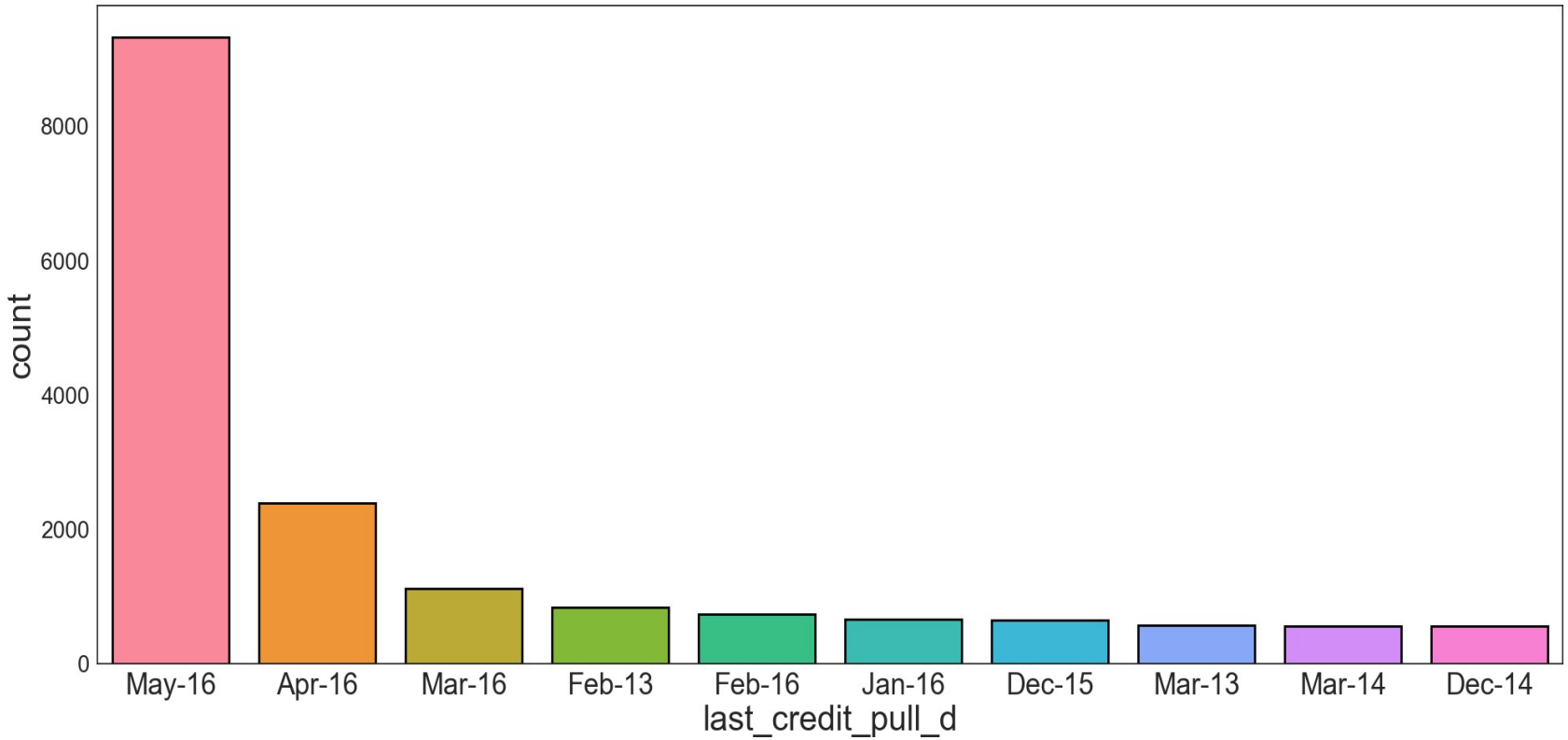
Observations: There is an increasing trend in the no. of issued loans every month

Analysis of Loan Amount



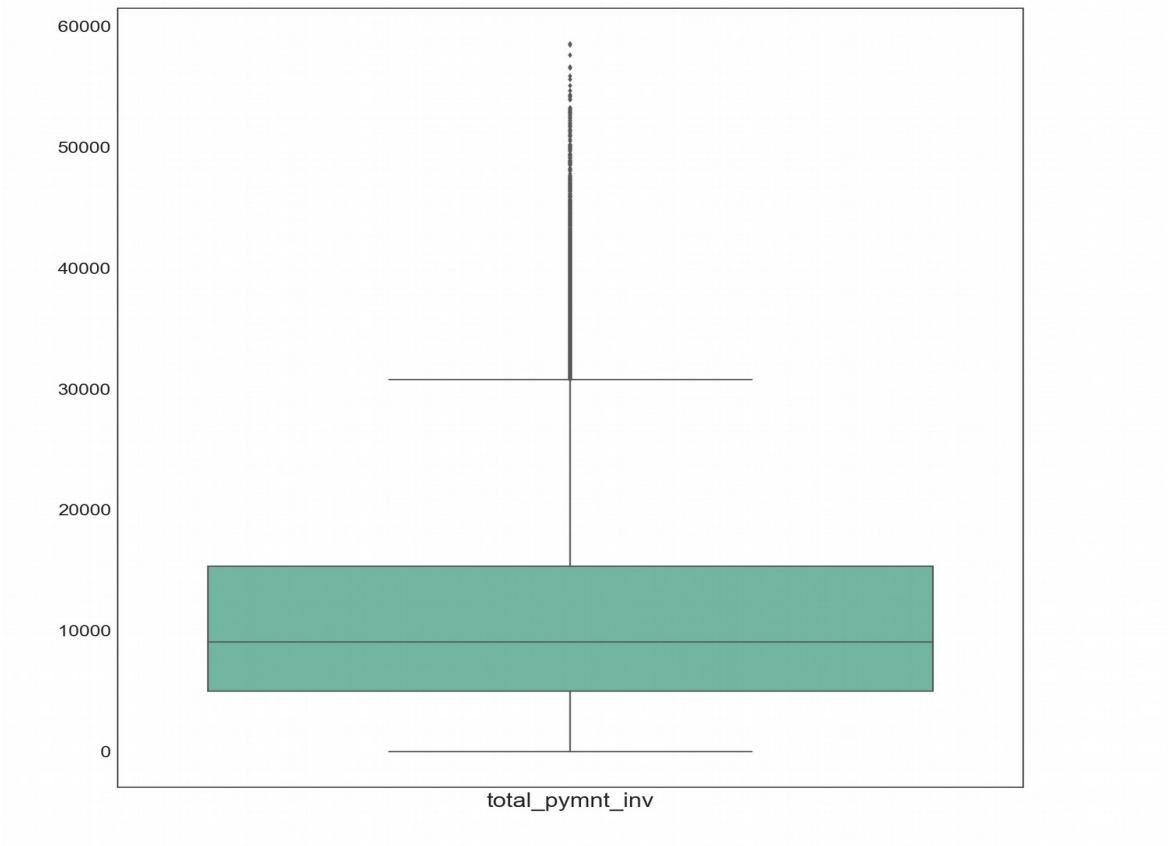
Observations: From the box plot outliers at the far end.

Analysis of Last Credit Pull Date



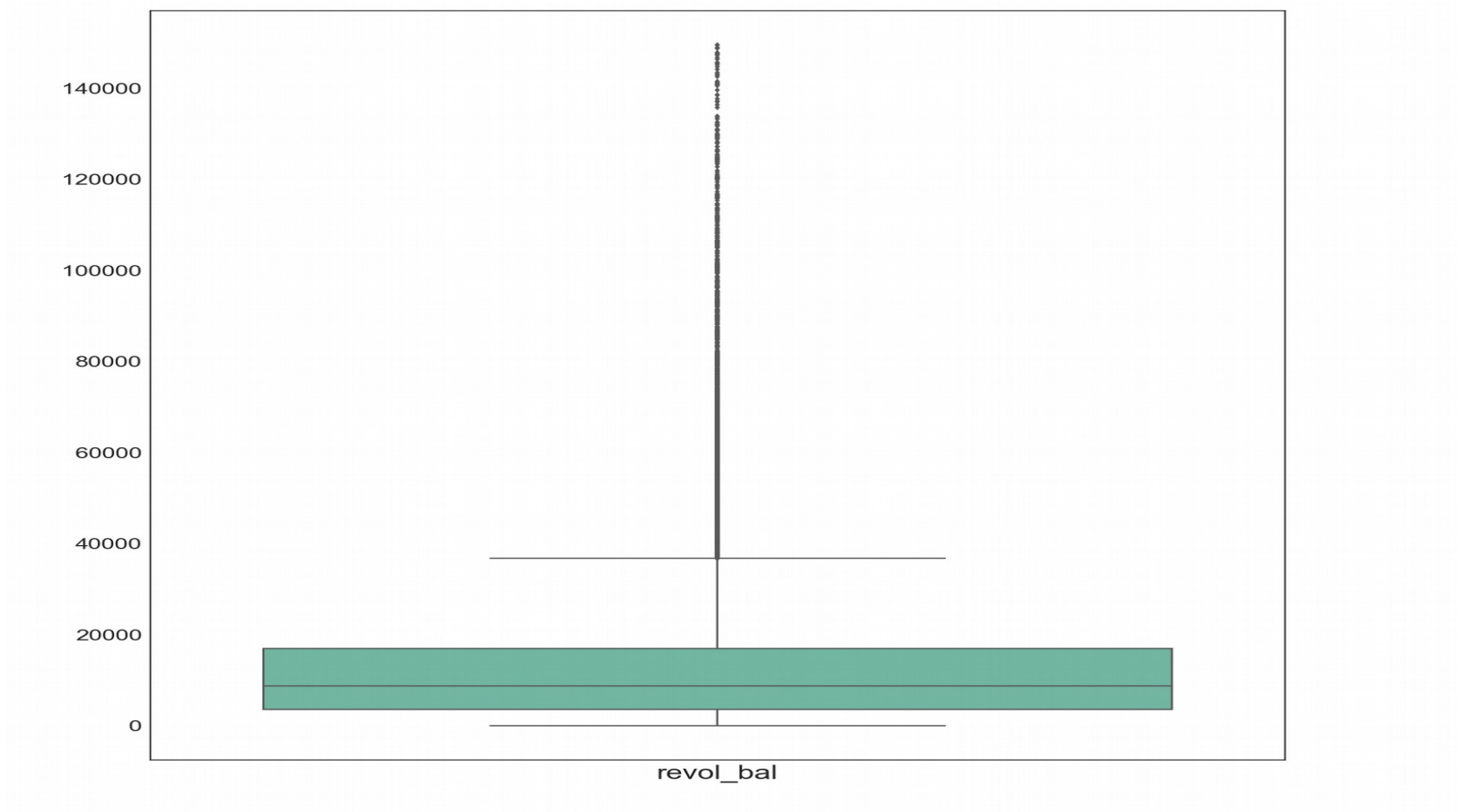
Observation: The credit request pull was very high in may 2016

Analysis of Total Payment



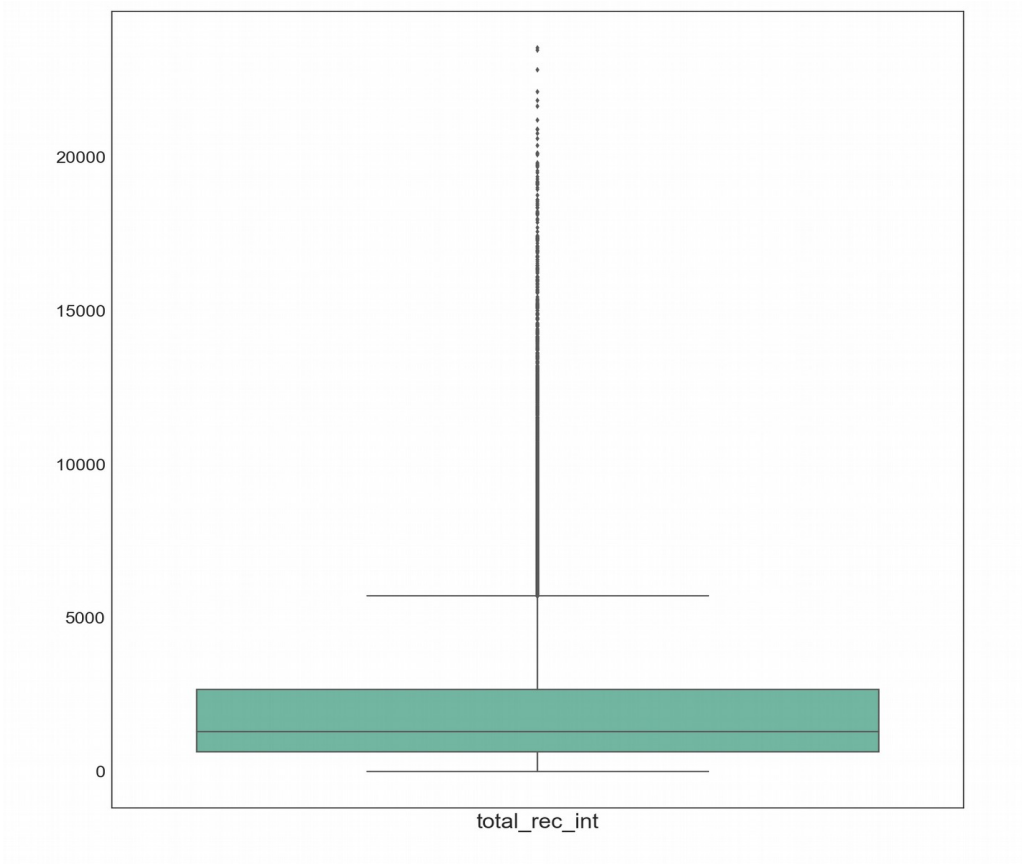
Observations: From the box plot it appears that there are a considerable number of outliers

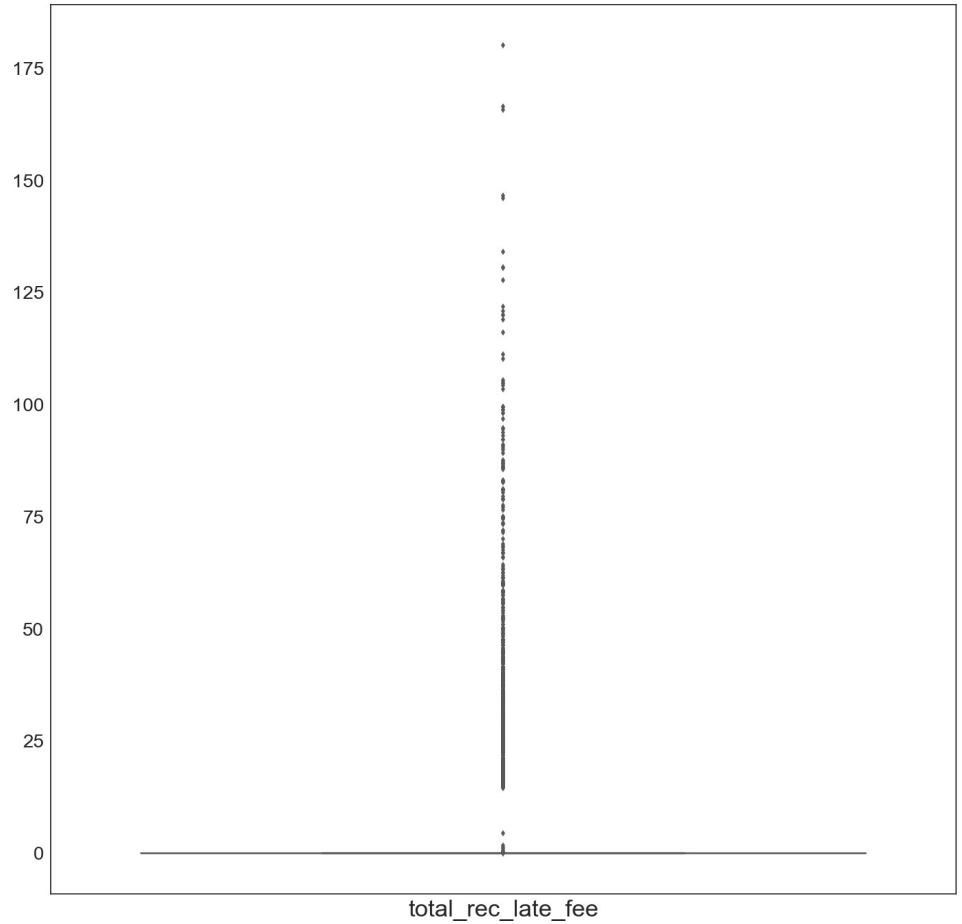
Analysis of Revolving Balance



Observations: From the box plot for Revolving Balance, it appears that there are a considerable number of outliers that we may need to remove

Analysis of Total Interest Recd., Total Late Fee Recd.,

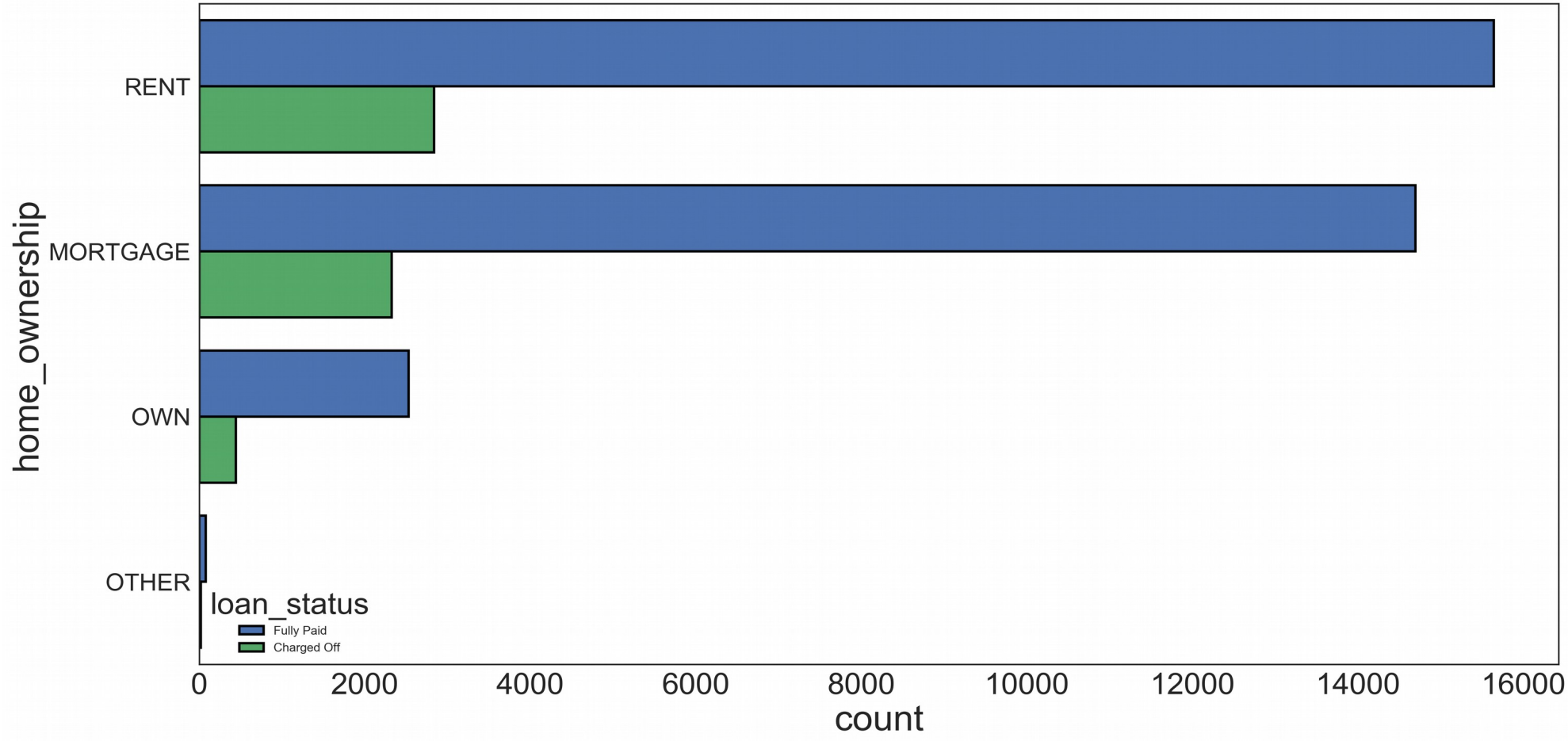




Observations: From the box plot for all 2 columns, it appears that there are a considerable number of outliers

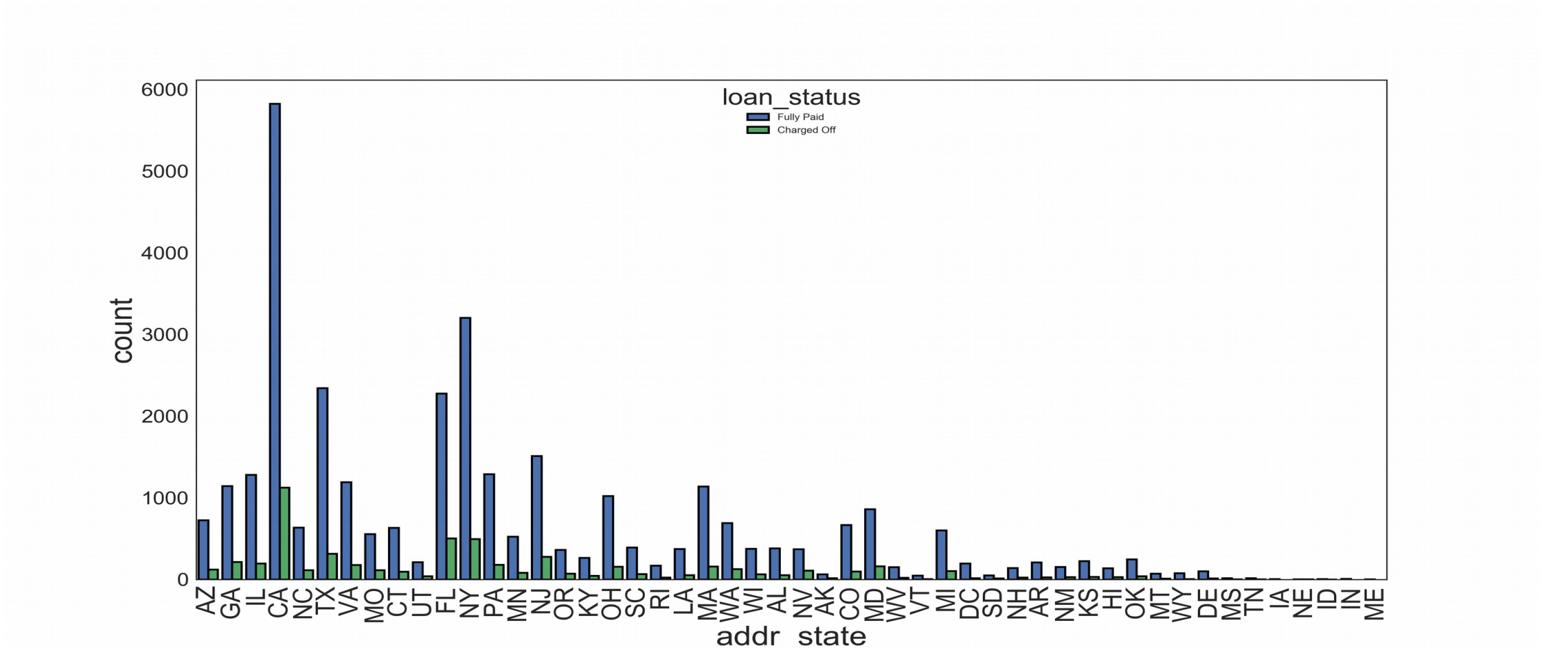
Bivariate analysis

Home owner Attribute vs. Loan Status



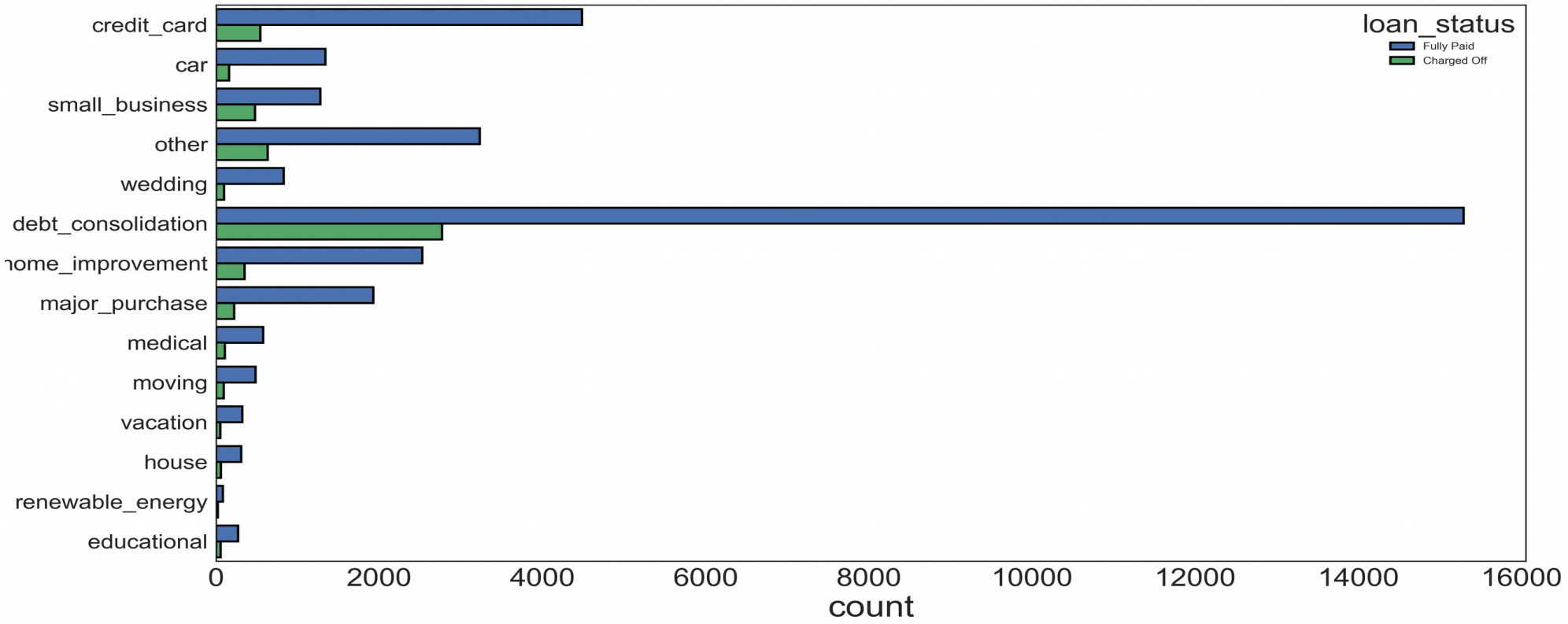
Observation: If the borrower's home ownership status is "OWN", he is less likely to be a defaulter compared to others.

State vs. Loan Status



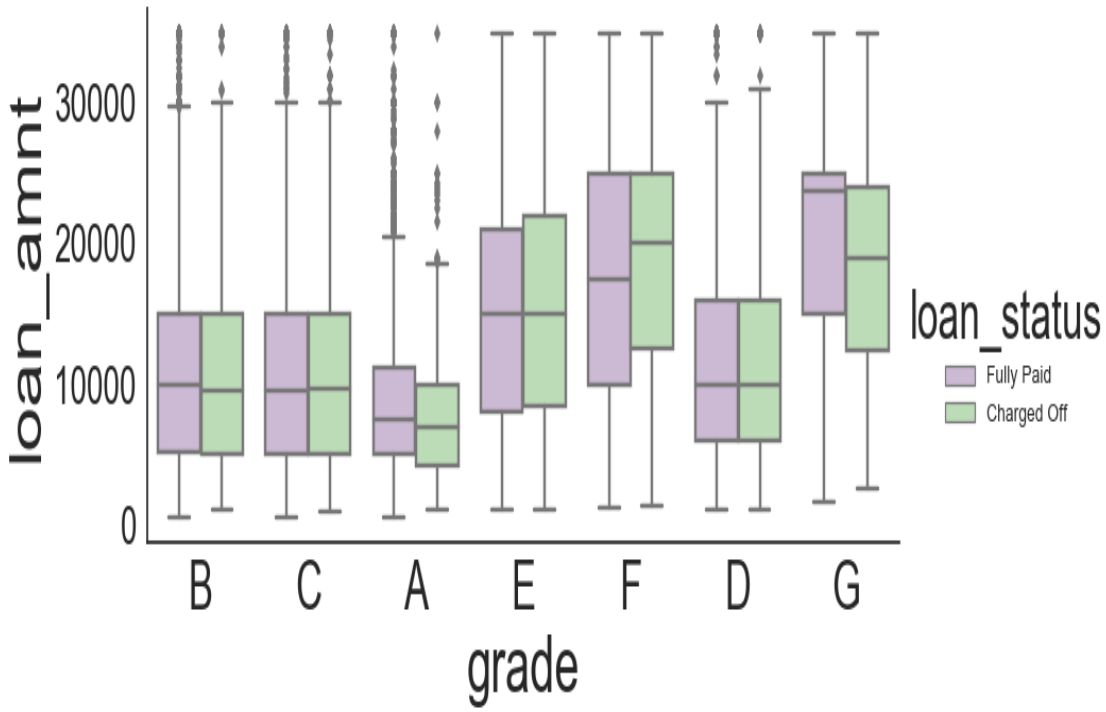
Observation: CA is the most risky state to lend a loan compared to other states

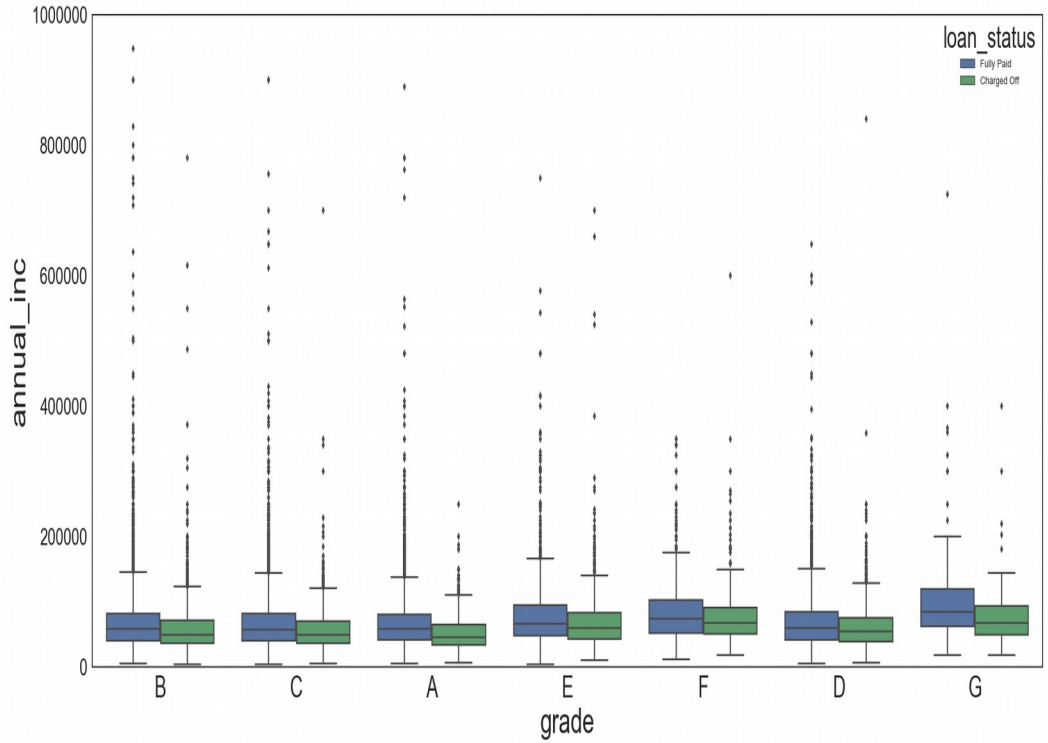
Purpose Attribute vs. Loan Status



Observation: Borrowers for ‘Small Business , credit card’ are most likely to default for the loan compared to other purposes

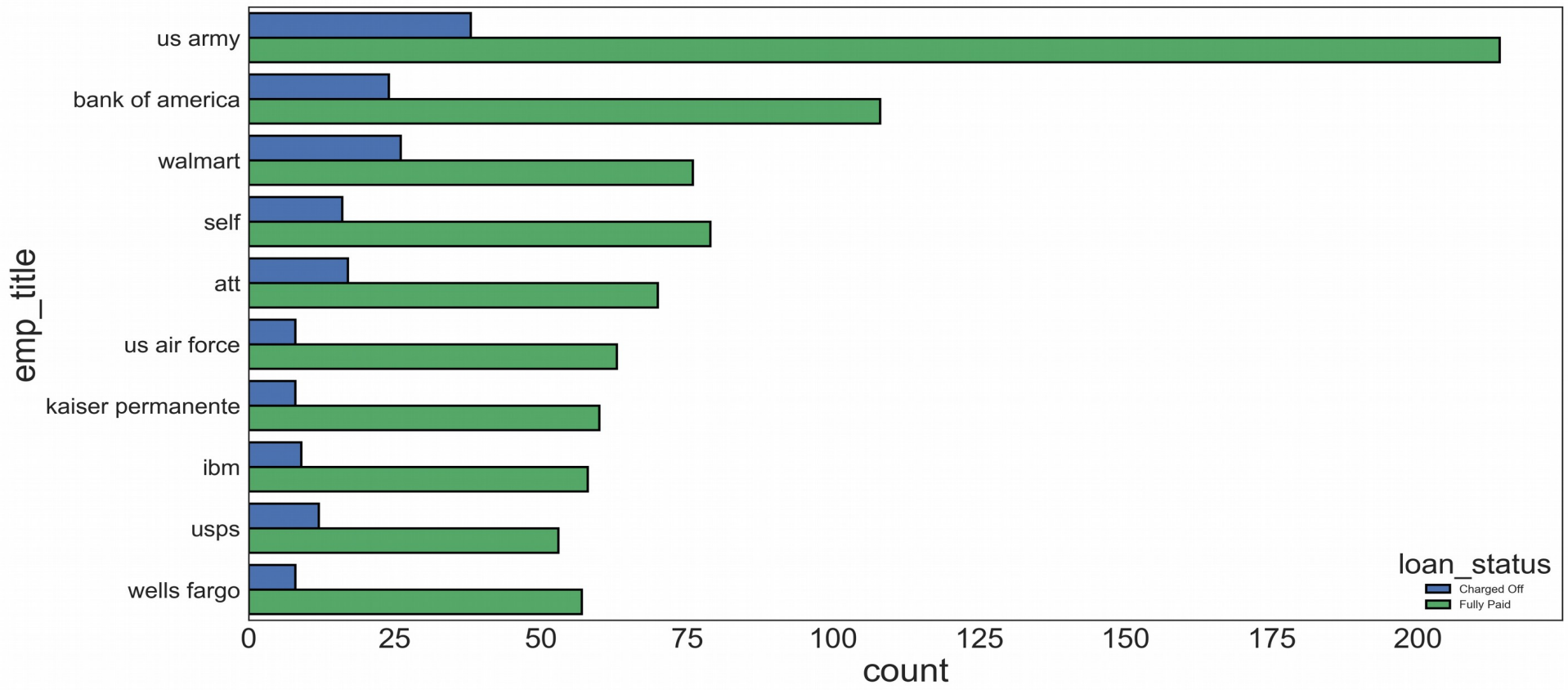
Grade vs. Loan Status





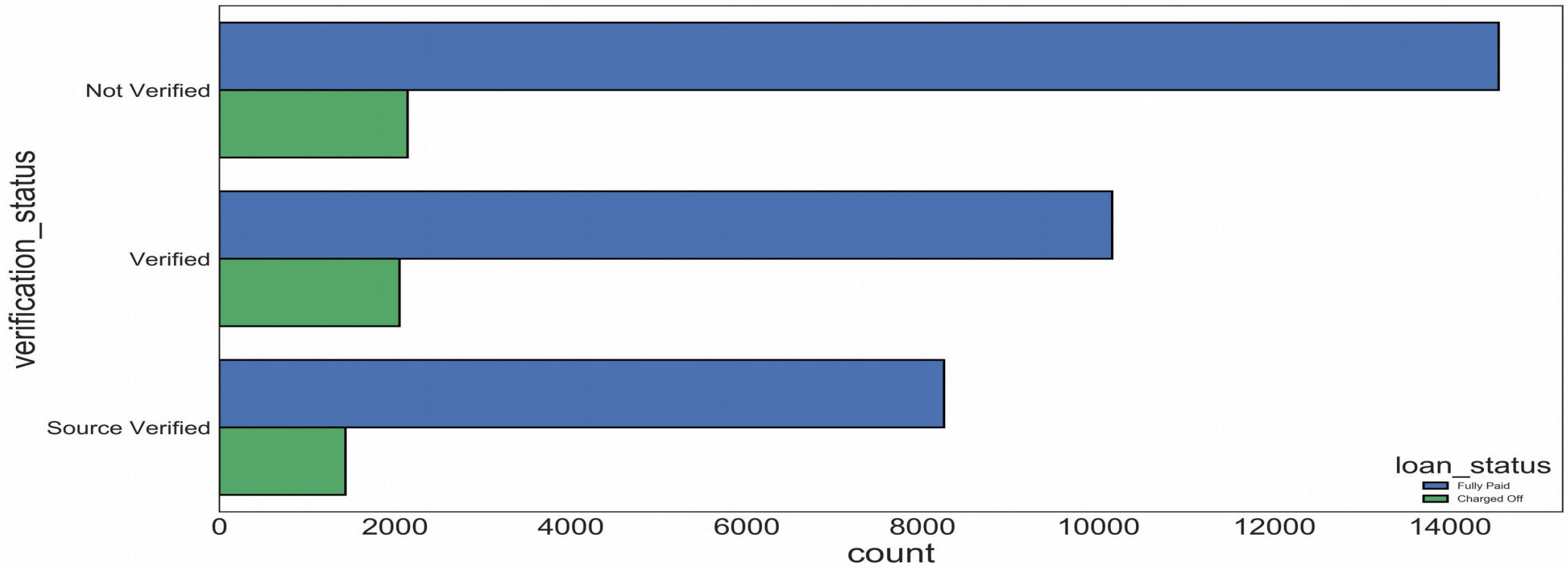
Observation: Grade F has higher loan amount

Employers vs. Loan Status



Observation: Employers working in US army are more likely to default

Verification Status vs. Loan Status



Observation: Not verified source are likely to default