Lending Club Case Study

VIBHOR SHRIVASTAVA

Problem Statement

We are given dataset from a large online loan marketplace where people can avail loans easily. We are given data about customers who are currently paying their loan, customers who already paid their loans and the ones who defaulted on their loan.

The company wants us to find the drivers to identify risky loans (loans which are likely to be defaulted)

Data Sourcing

We are provided with a dataset loan.csv

On analysing we find that it has 39717 rows and 111 columns

We are also provided a data dictionary to look into the meaning of columns

Data cleaning

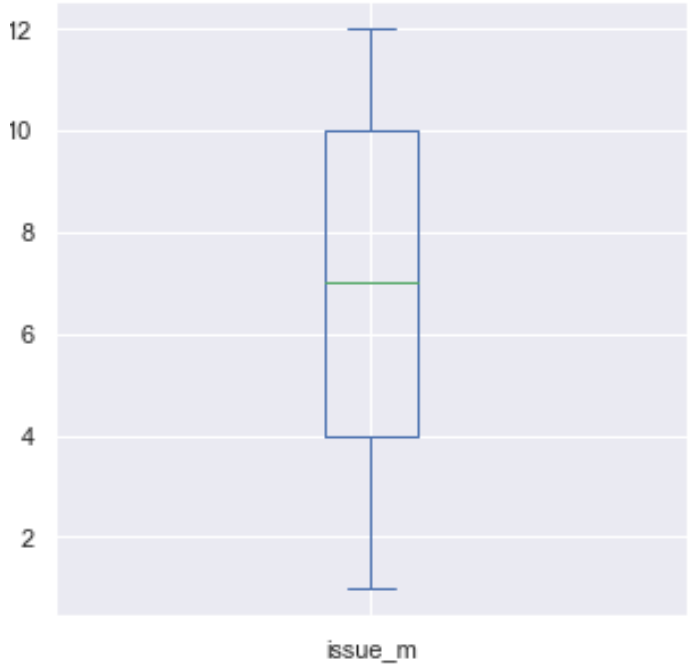
- Delete all the columns having all values as null.
- Delete all the columns having most of the values as null.
- Delete all the columns having one constant value through out
- Delete all the columns which doesn’t carry insightful information
- Since the dataset has a lot of columns, analyse chunks of columns and check which one of those seems relevant.
- If the values seem same for multiple columns, check if they are correlated. If yes, delete all except one column out of them

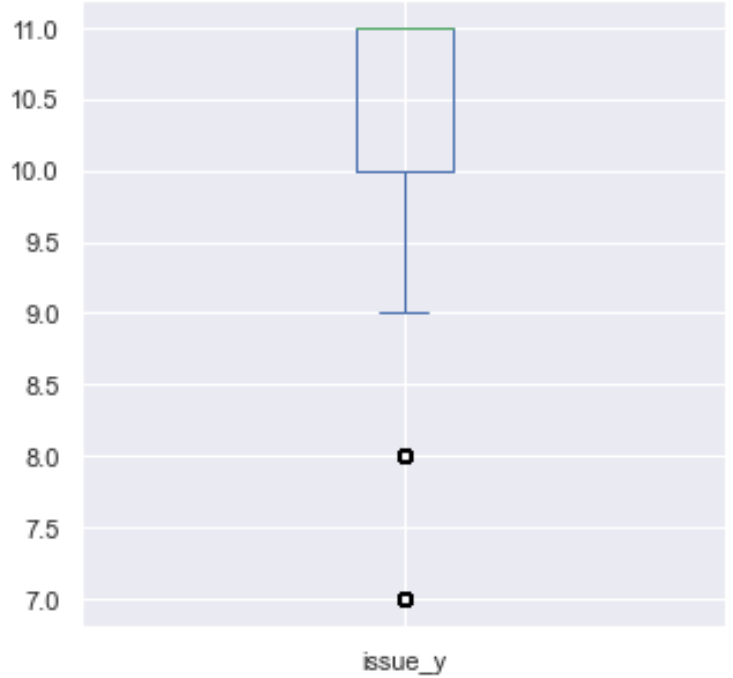
Data Cleaning Continued

- After performing steps mentioned earlier – we will be left with fewer columns.
- Our target variable is ‘loan_status’
- We don’t need data about customers with their loan currently running, so we will delete them
- Then we will check the types of columns
- If there are certain object type columns which could have been numerical, we will change their types
- Eg: emp_length, revol_util
- We will try to extract months and years from date time columns (last payment date, loan issued date, last credit pulled date)

Analysis

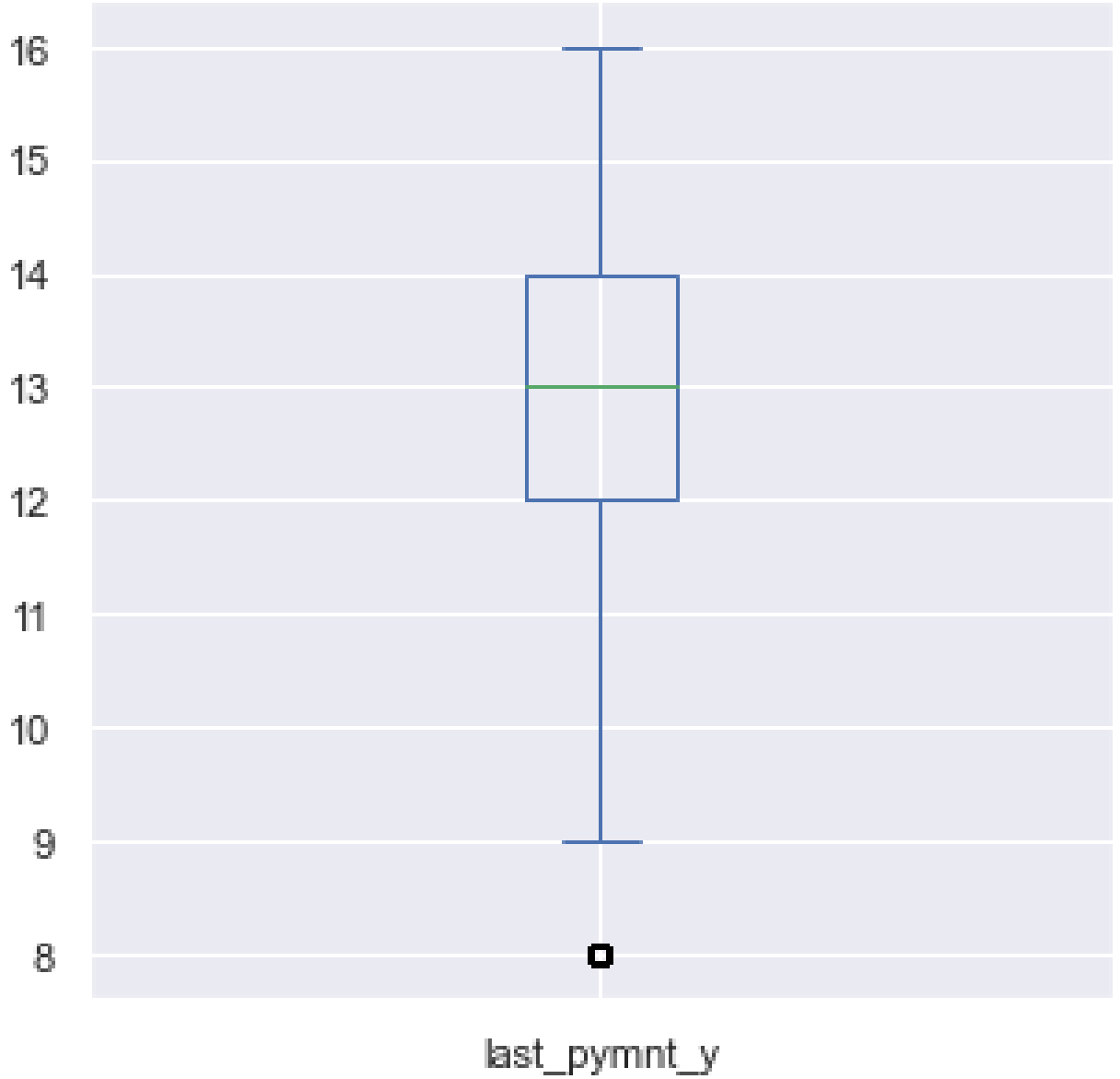
We will perform some univariate analysis on date columns (years and months) to check when the loans where issued. Left one is loan issued month and the right one is loan issued year





Analysis continued

We can clearly see that most loans were issued from 2009. We will remove the rows where loans were issued earlier than 2009 since they were outliers. 2008 was also the time of recession and hence economic parameters could be different at that time which could affect our analysis.We will perform the same analysis for last payment year. It doesn't give any insight though

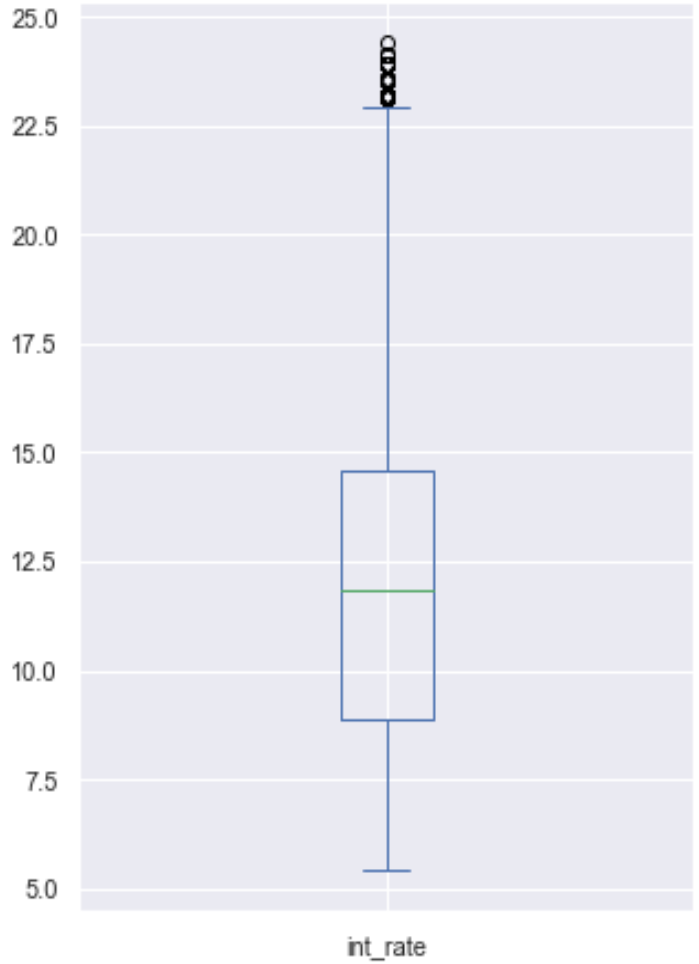


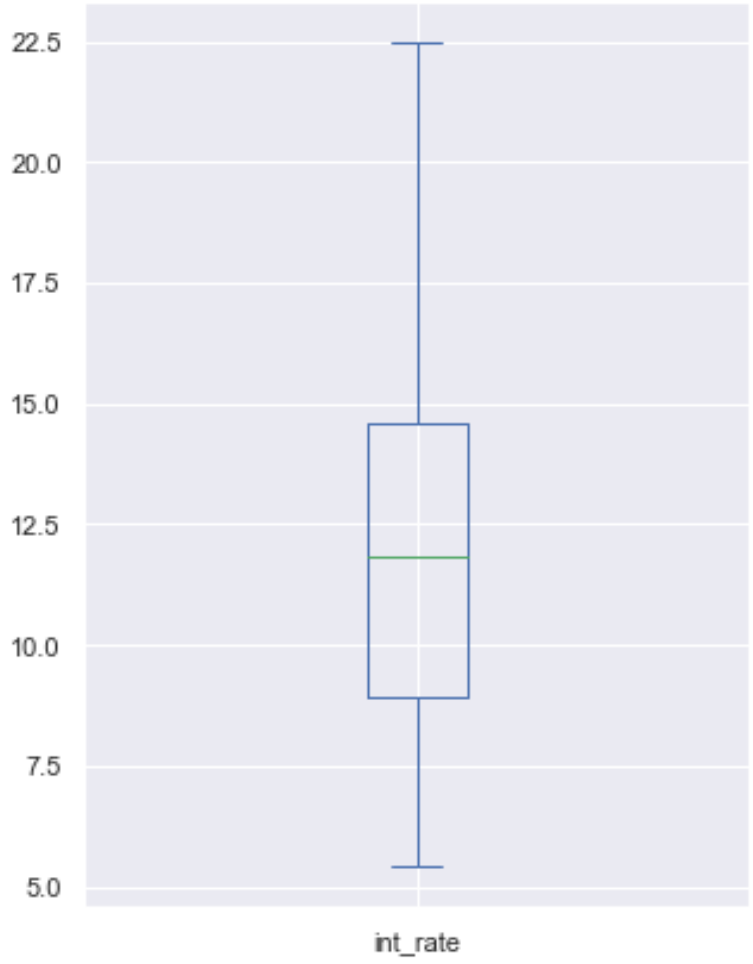
More univariate analysis

Column: interest rate

First image is with outliers

Second image is without outliers removed so that our analysis won't be skewed





More univariate analysis

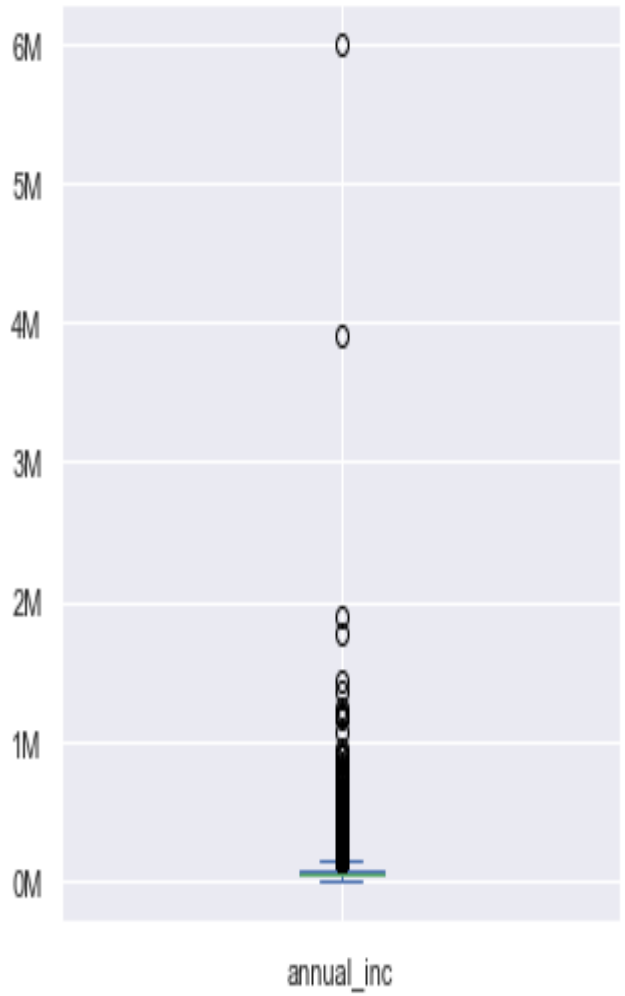
Column: annual income

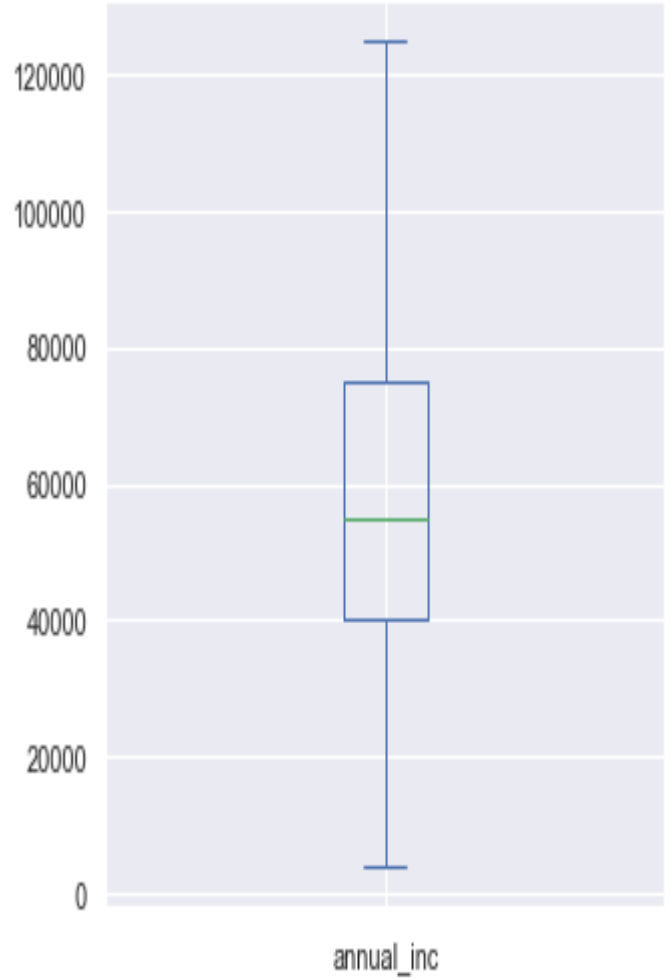
First image is with outliers

Second image is without outliers

Customers with higher annual income could have skewed the analysis since they have capability to pay off their loans easily

Please note that the scale for first analysis is in Millions





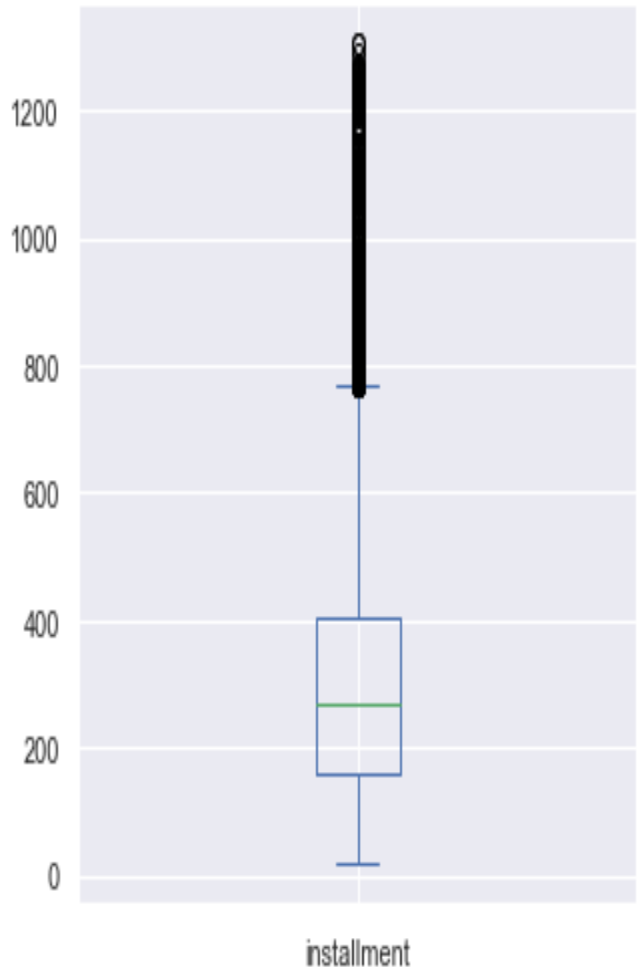
More univariate analysis

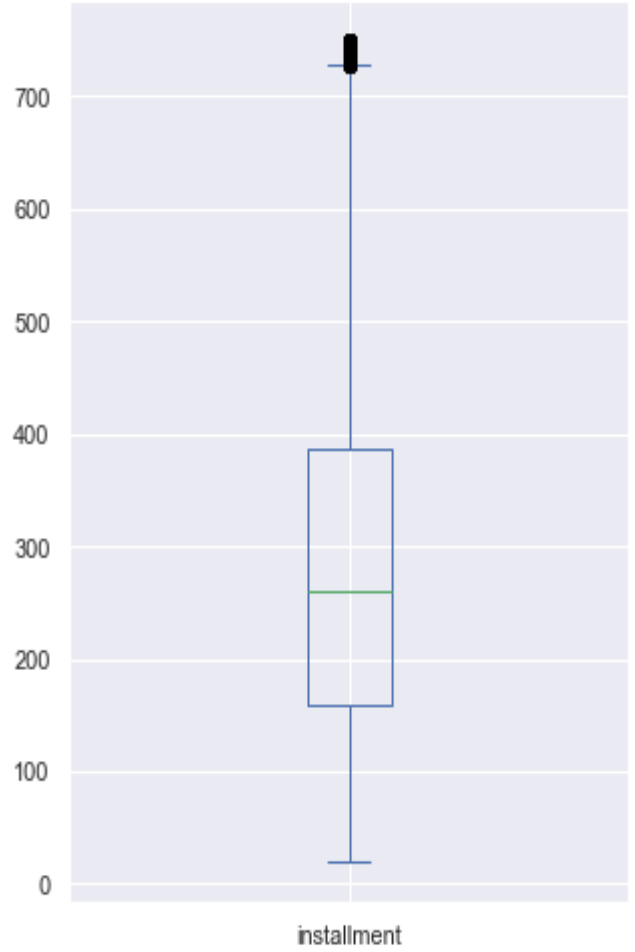
Column: installment

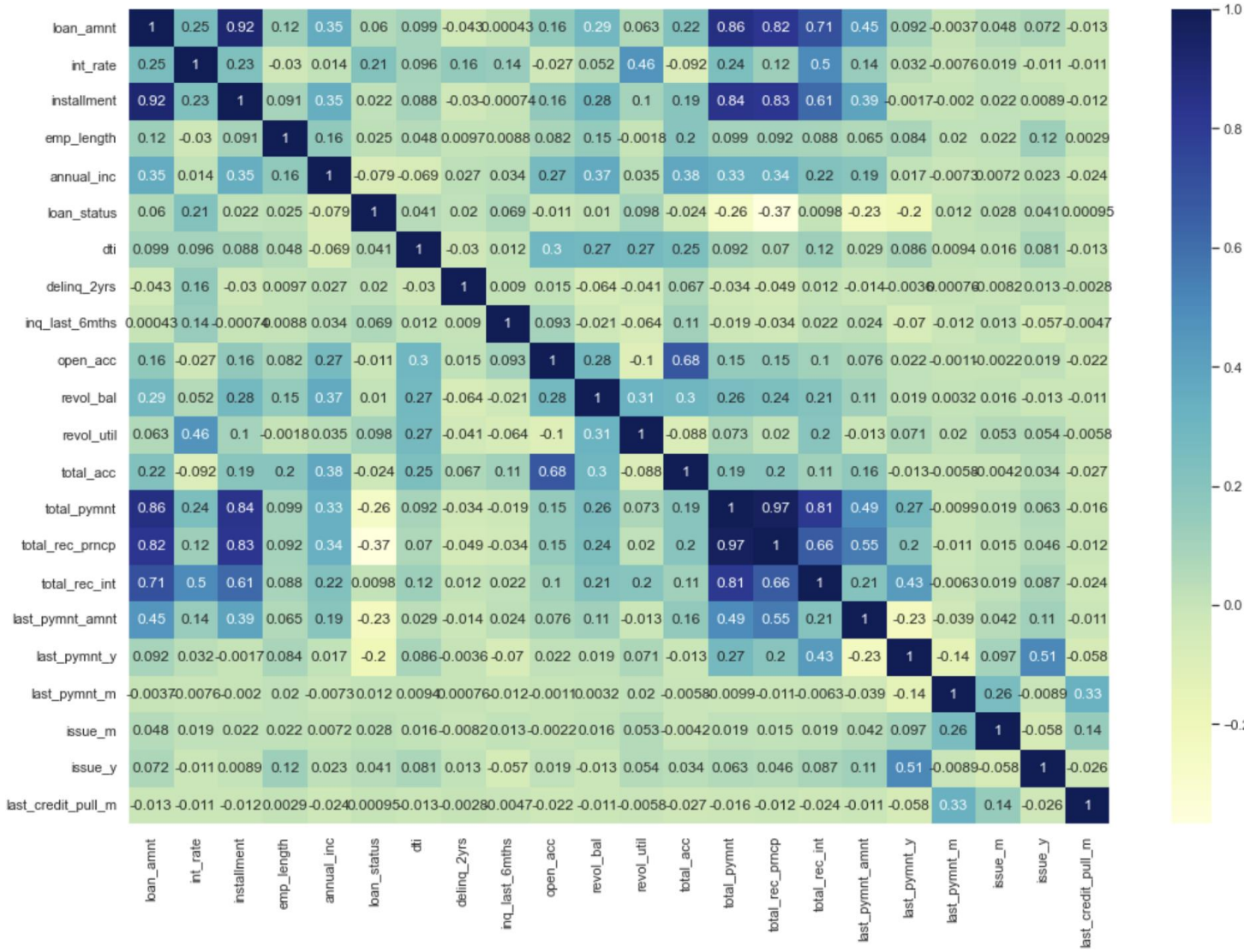
First image is with outliers

Second image is without outliers(or with fewer outliers)

Rows with installment greater than 750 were removed so that out analysis won't be distorted







Correlation Heat map(Derived)

Since there are many columns to analyze, it will make things easier since we could check which columns have higher correlations with the loan status

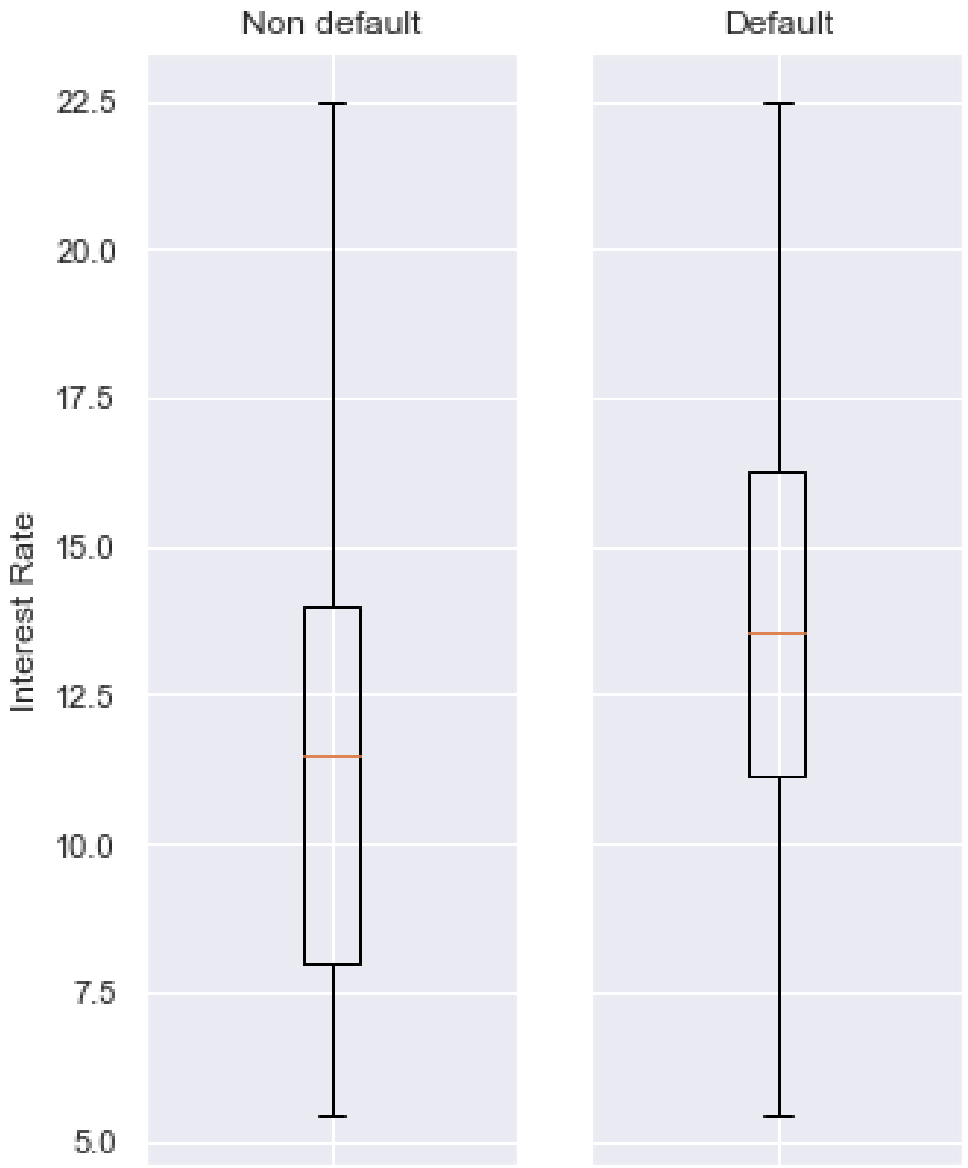
Learnings from heat map

- Total payments made till date, total recorded principal, total recorded interest, loan amount and instalment are highly correlated
- We will keep only total payments and delete the rest
- Interest rate is positively correlated with loan
- Total payments made till date, total recorded principal and last payment amount towards the loan are negatively correlated with loan status.
- Let’s do some bivariate analysis to confirm the observation

Interest Rate vs Loan Status

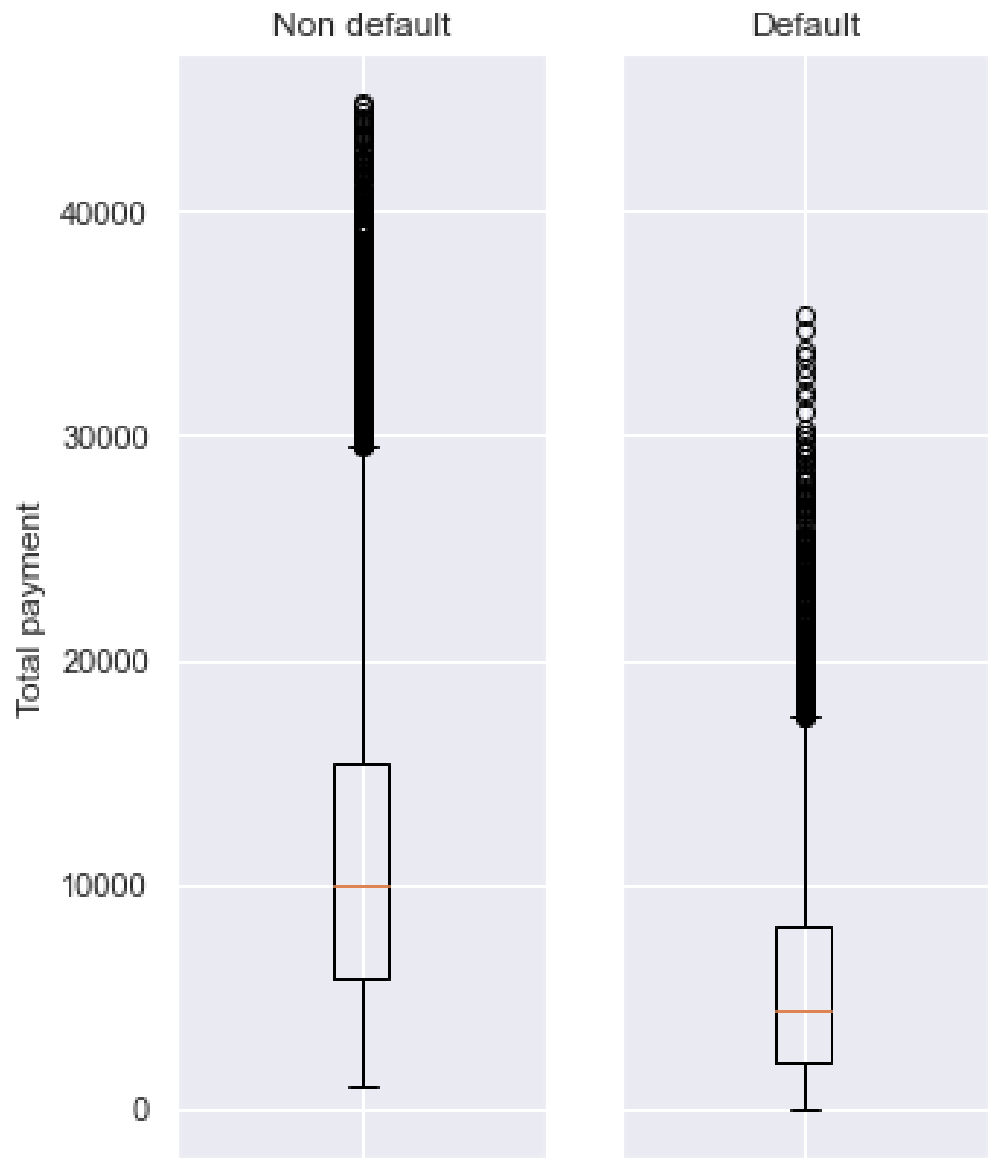
It’s been quite clear from the box plots that customers with higher interest rate on their loan tend to default more often.

They might feel they are paying a lot more money on top of the actual loan hence they tend to default.



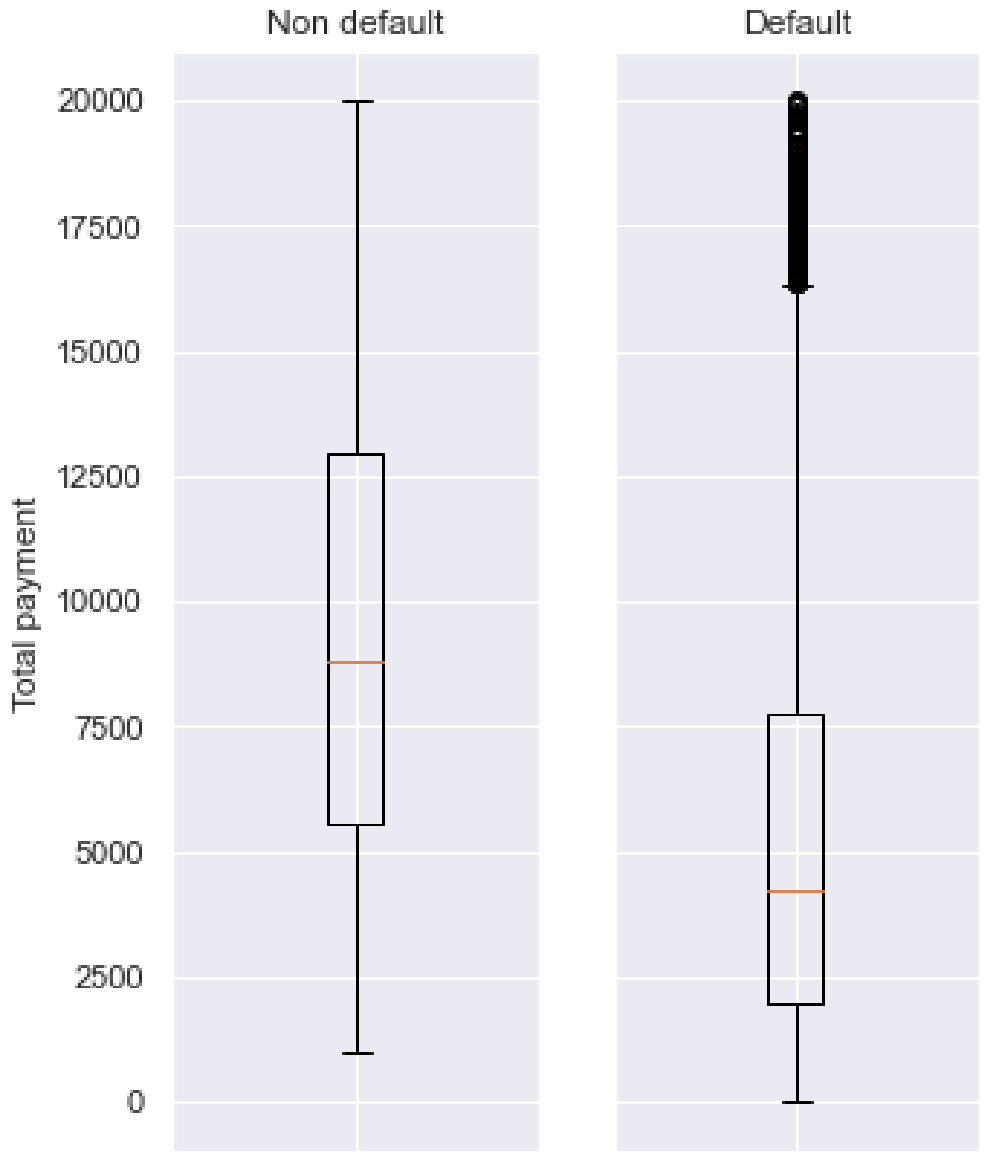
Total Payment vs Loan Status

It’s been quite clear from the box plots that customers who paid lower total payments towards their loan were the one who tends to default more.



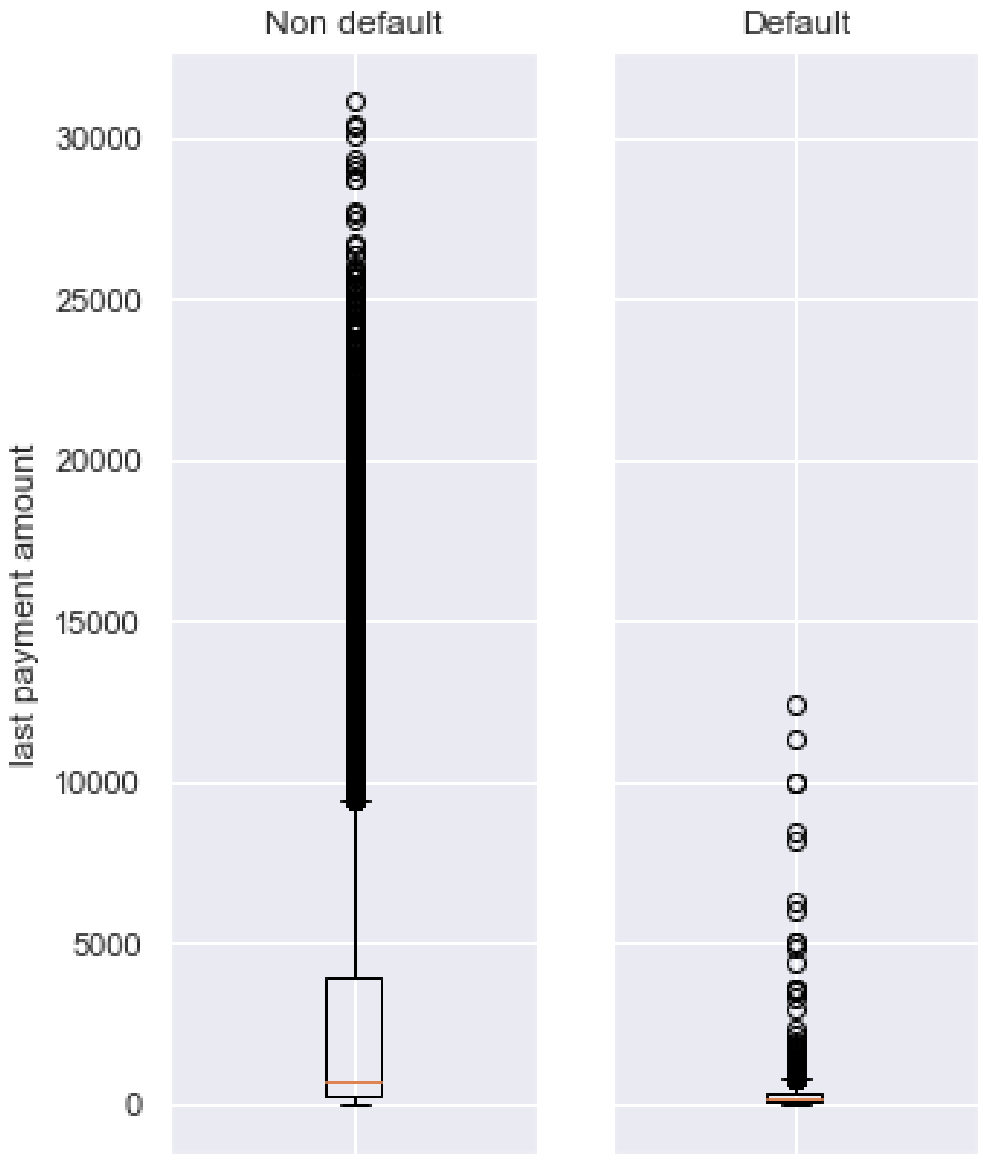
Total Payment vs Loan Status

We see the same pattern even after removing quite a lot of outliers.



Last Payment Amount vs Loan Status

The pattern is same with last payment amount towards the loan. Clearly, customers capable of paying off their loans, or the ones with the intention of paying off their loans – have big last payment amount and vice versa

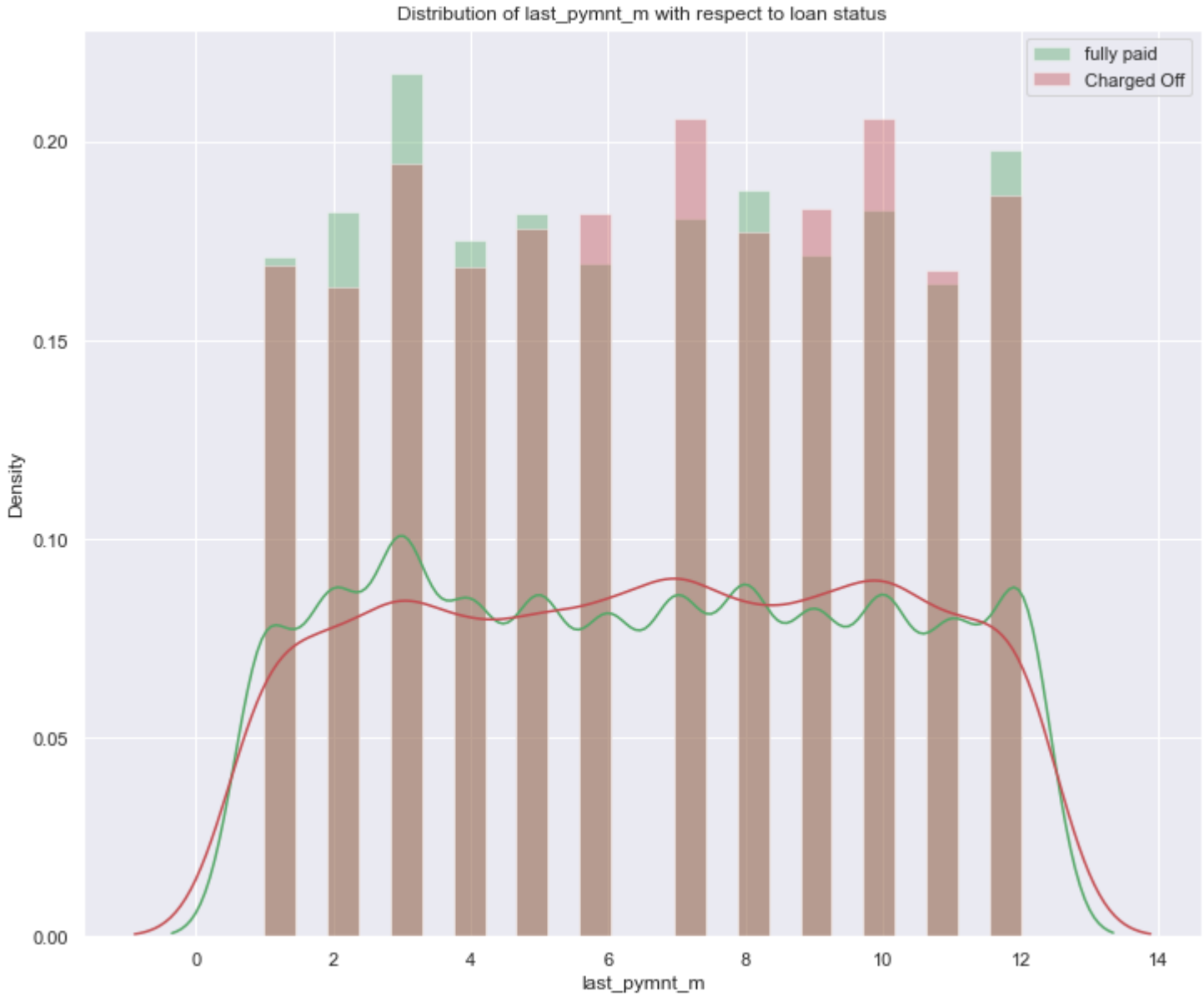


Date related columns

- Let’s play with date related columns and try to see if they give any drivers for defaults
- Last payment month, Last credit pull month, Loan issued year (Derived metrics*)
- We already know from heat map there shouldn’t be much of a correlation here

Last payment month vs Loan Status

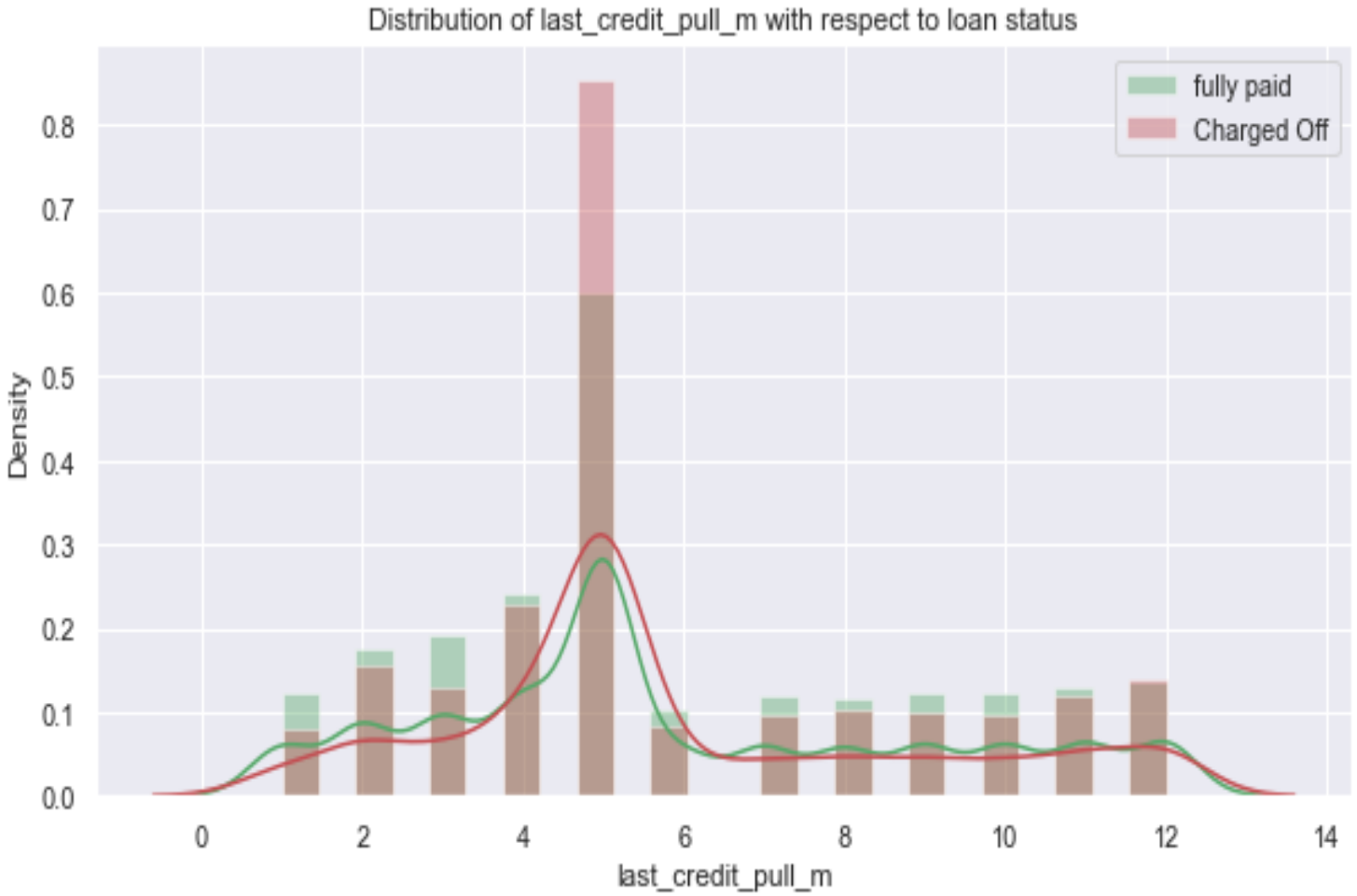
Not much of a correlation here as expected



Last Credit Pull Month vs Loan Status

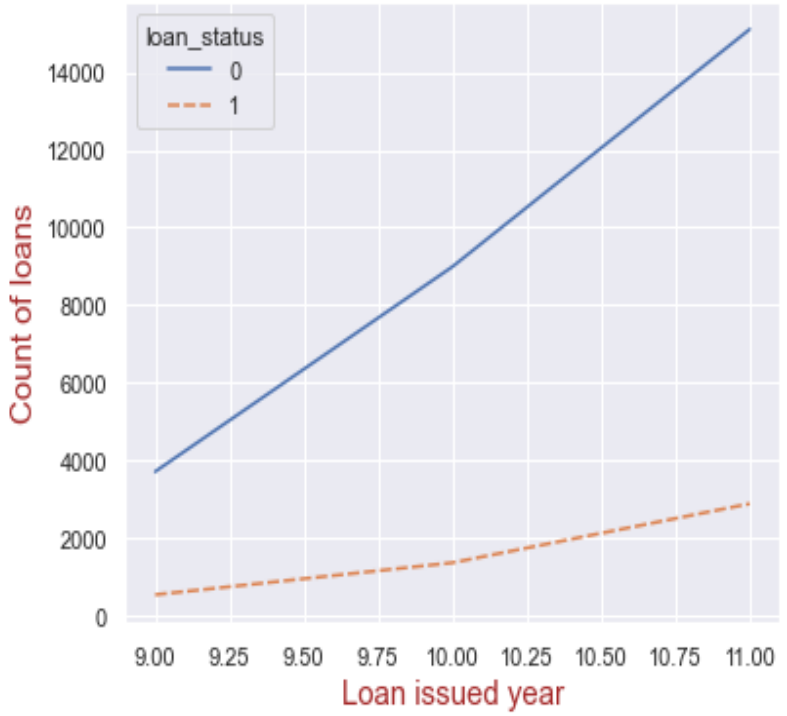
This also looks normal except one anomaly

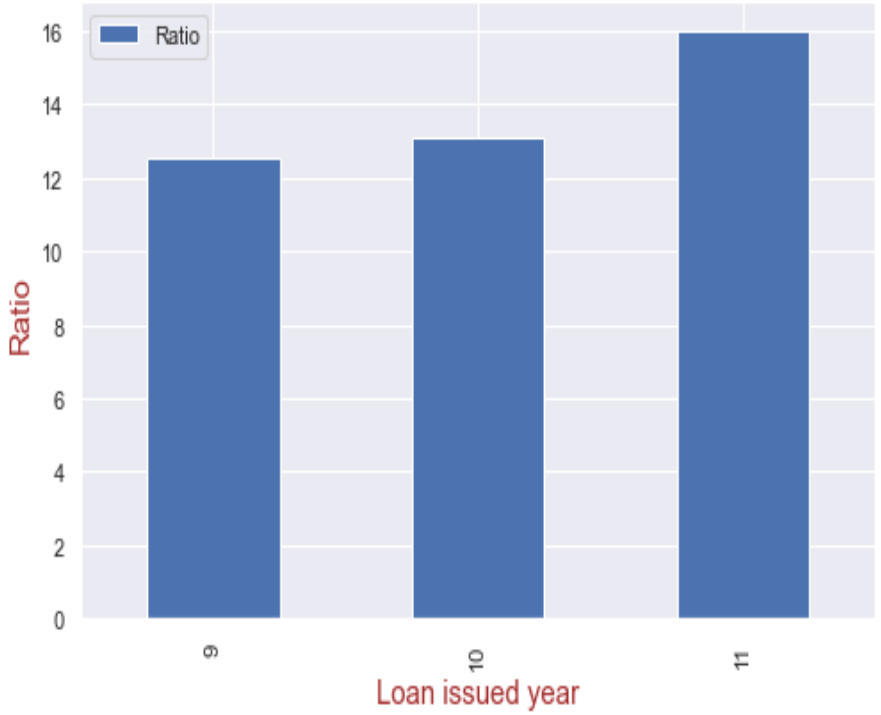
For most of the loans when LC last pulled credit in ‘May’, the loan tends to be default. Strange but this certainly is noticable



Loan issued year

There is something about loan issued in year 2011. It has the most percentage of loan defaults as compared to the previous years. But until and unless the correct reason is known for this behavior, it is difficult to use this learning to predict future set of default loans





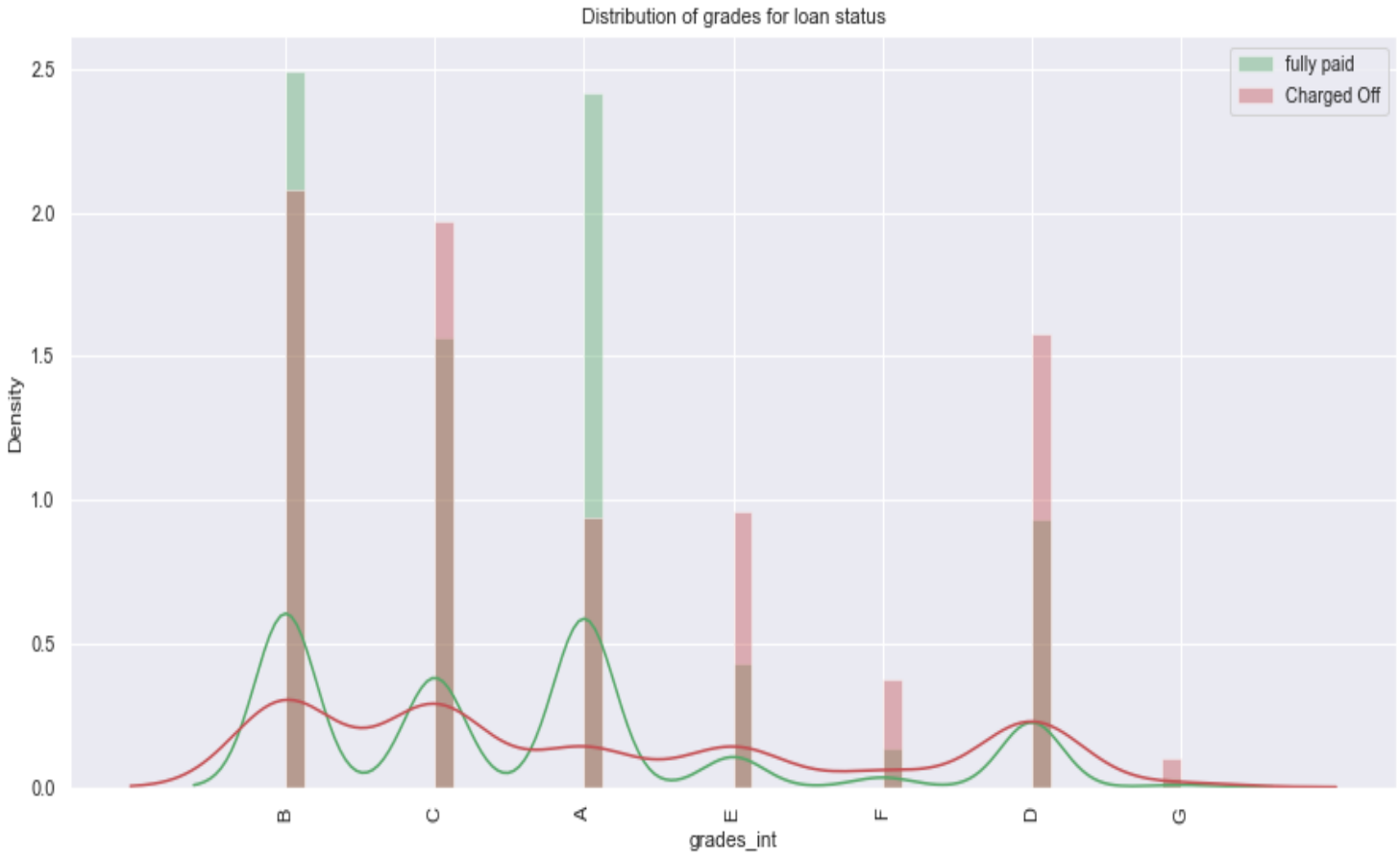
More analysis from categorical variables

- Grades
- Sub grades
- Purpose
- Address state
- Term type
- home ownership
- Verification status
- Bivariate analysis

Grades vs Loan Status

Loans assigned with A and B grades tend to be safe and are not likely to defaulted

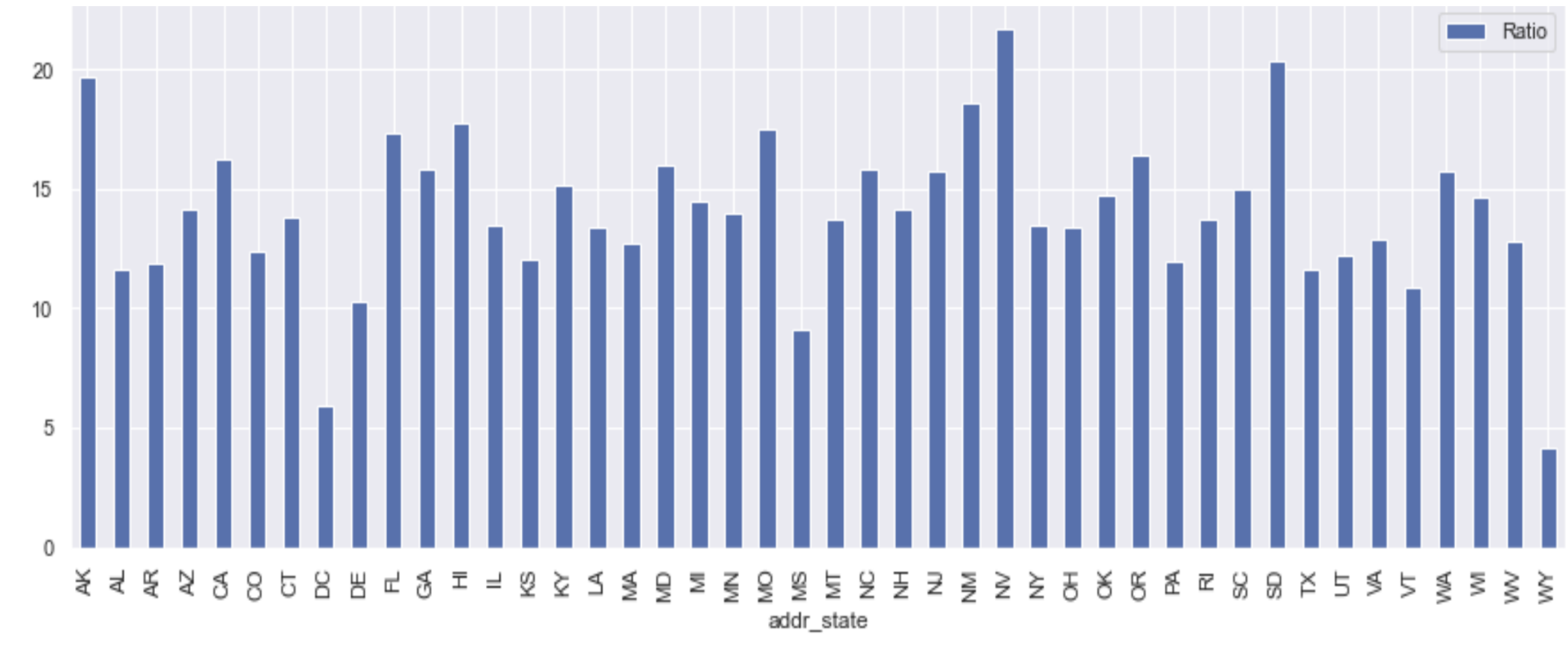
Loans assigned with E, F and G are more likely to default



Subgrades vs Loan Status

Loans with assigned sub grades under A are most likely to be safe

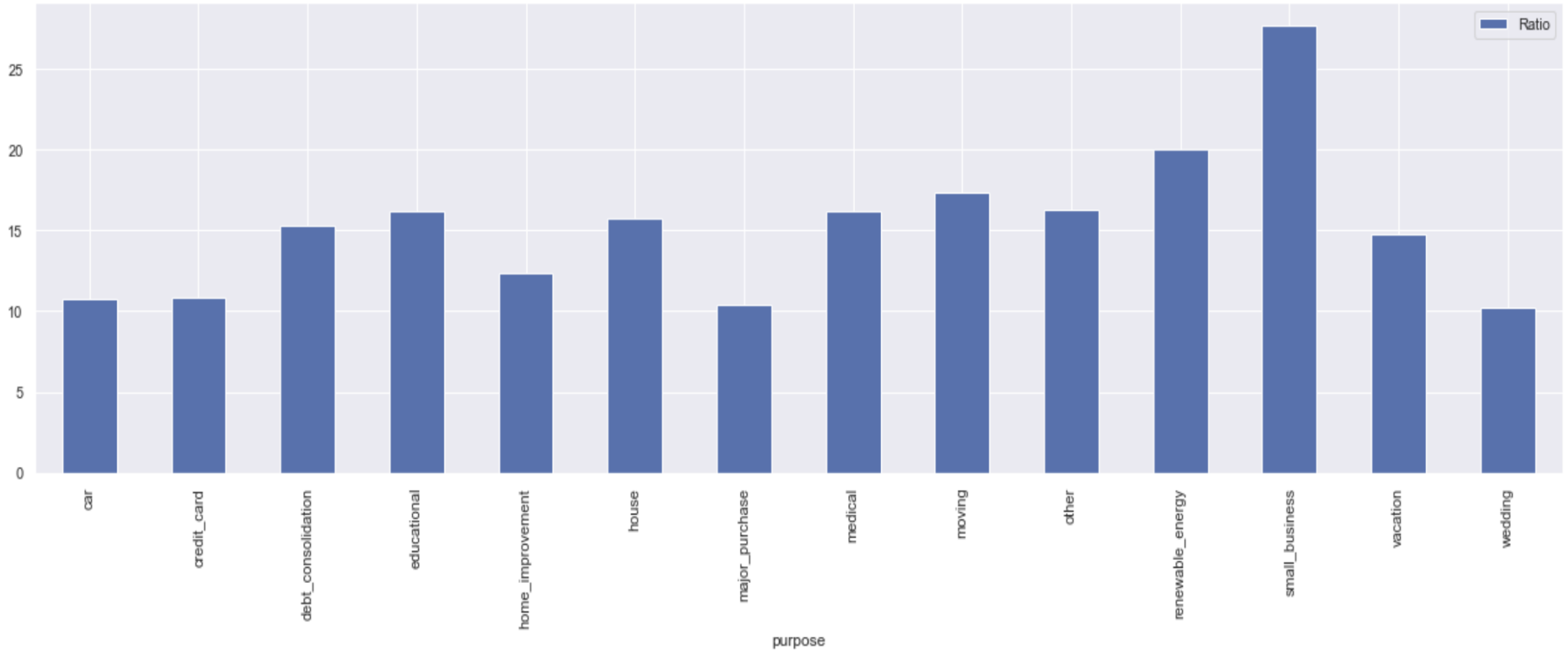
Loans under subcategories of B seems to be safe as well, but not as safe as A sub categories



Default loan ratio – state wise

Customers from NV, SD and AK have higher default ratio

Derived metric

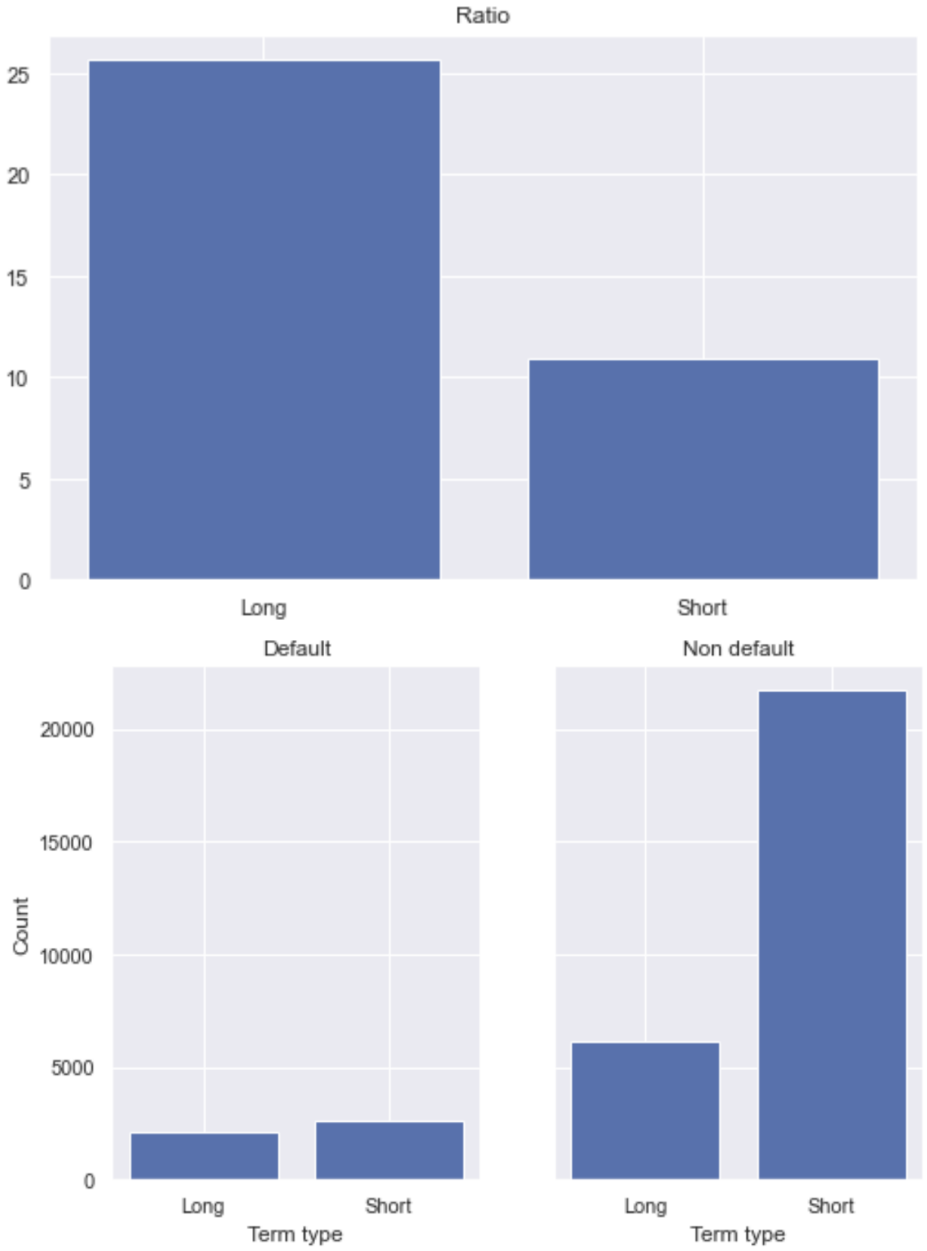


Default loan ratio – purpose wise

Loan purpose vs ratio of loan defaults (defaults / total no of loans for each purpose)

Loans taken for small business and renewable energy have higher ratio of defaulted loans

Derived metric

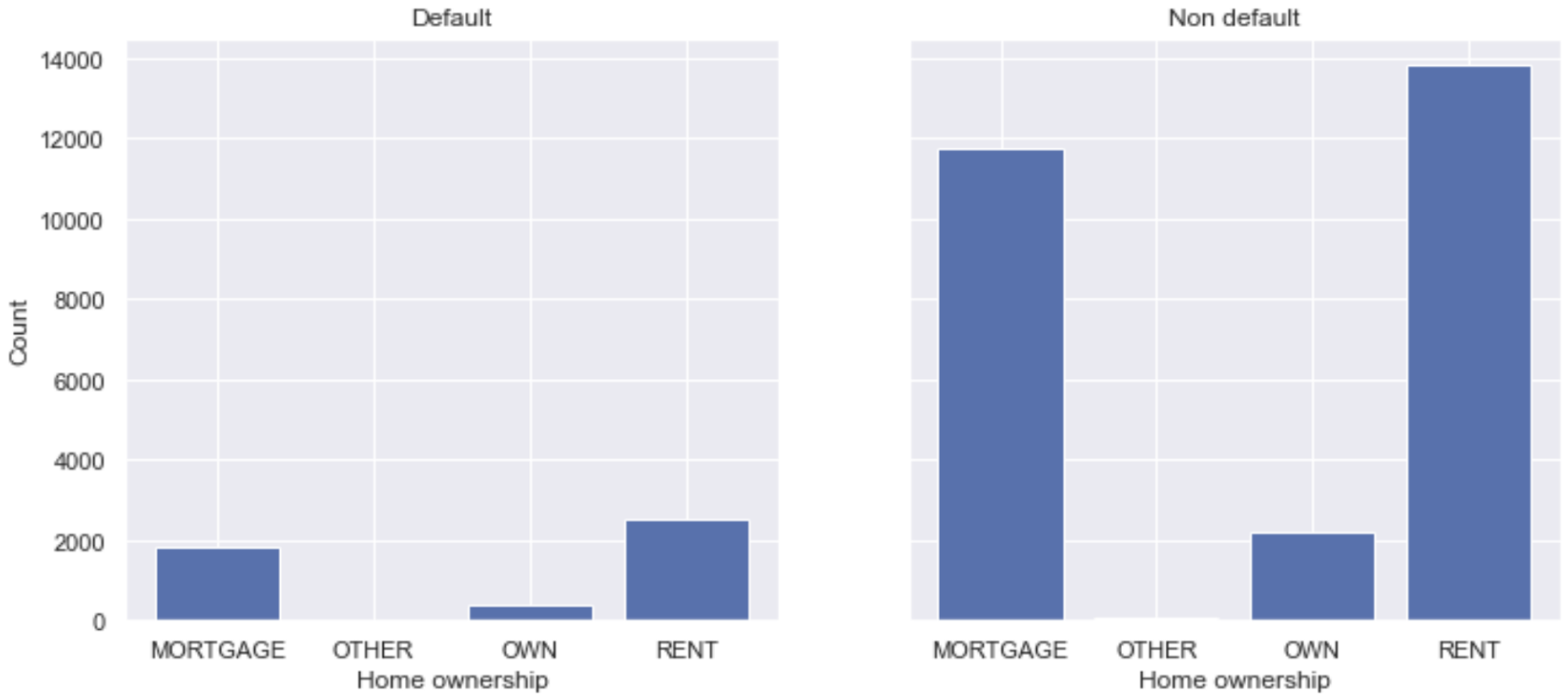


Default loan count and ratio – loan term wise

There is a certain strong correlation between loan term type and loan status. Loan with long term (60 months) has much higher default ratio as compared to short term loans. This is one of the strong driving factors for defaulted loans.

Graph on top displays ratio for long and short term defaulted loans

Graph on bottom displays count of loans broken into long and short term for defaulted and non defaulted loans

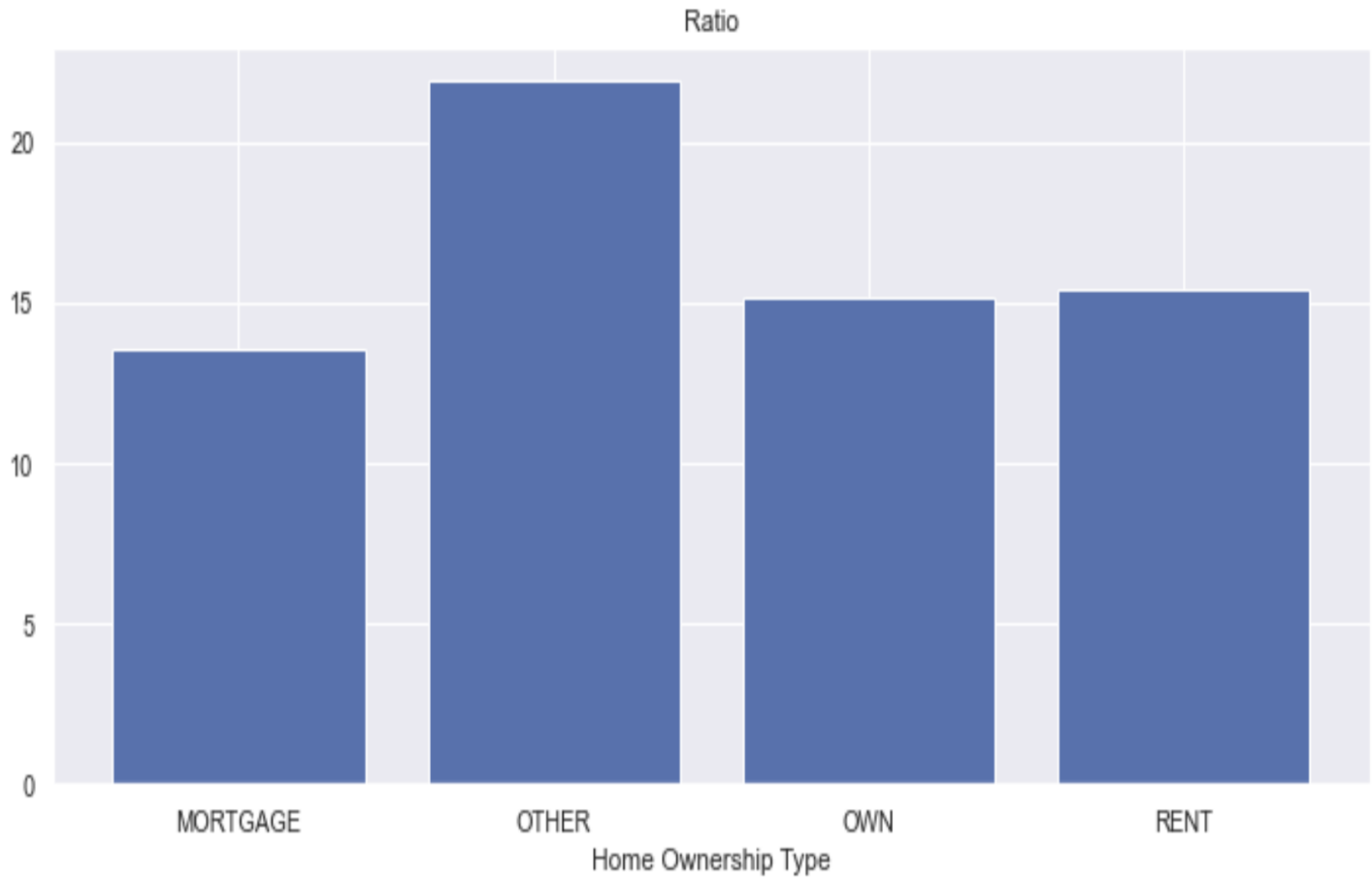


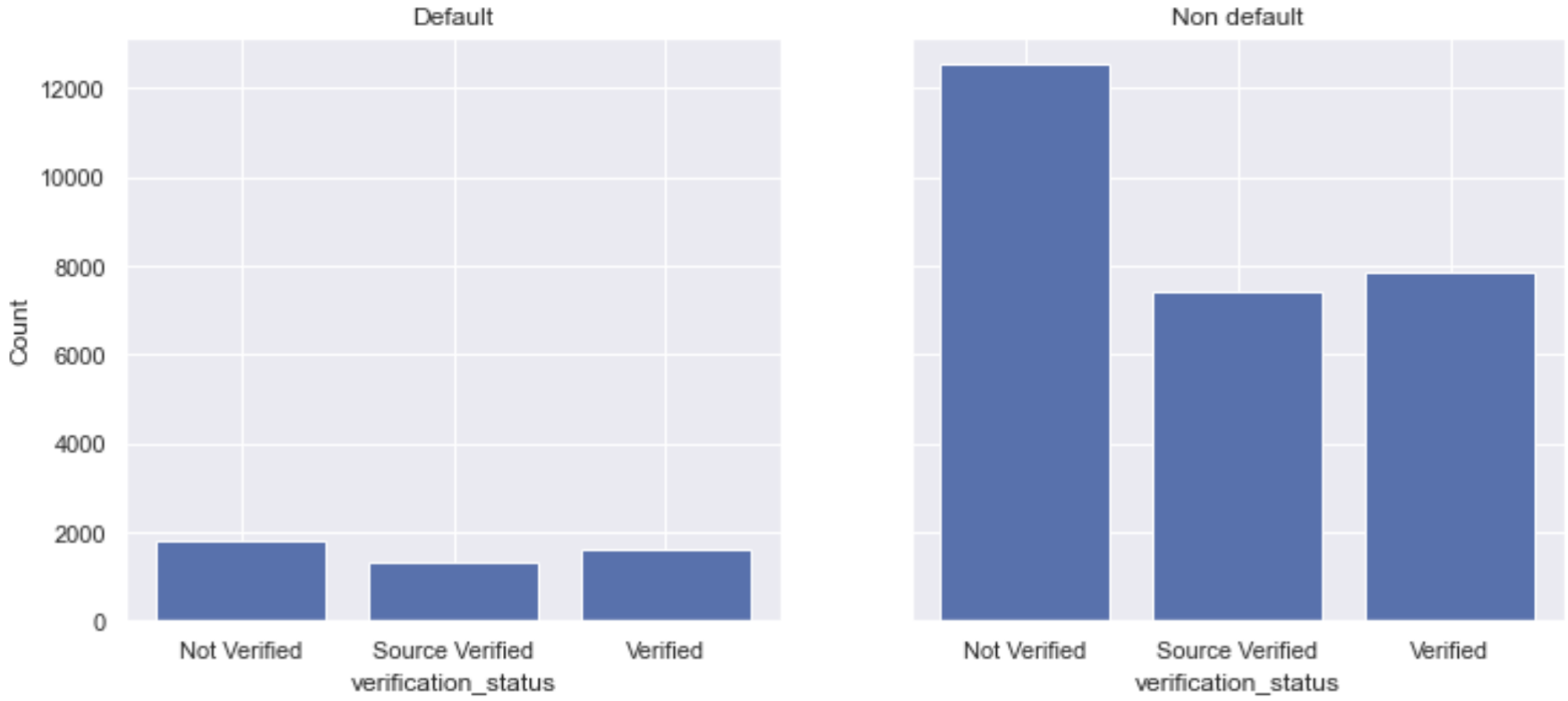
Default loan count categorised for home ownership

Default loan ratio for home ownership

‘Other’ category can be ignored since it does not have much data (as seen in the previous slide)

There is not much of any driving factor here for loan status

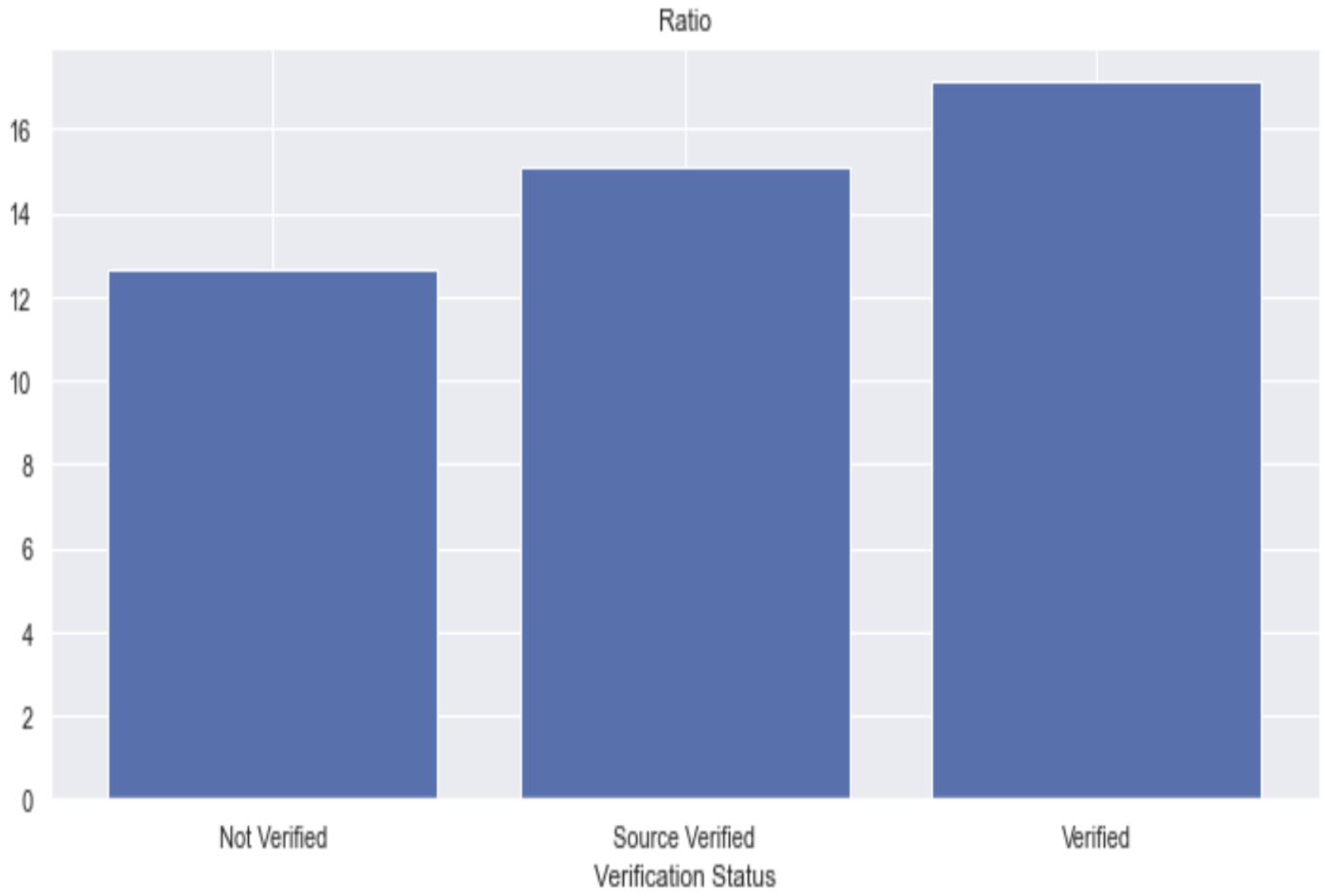




Default loan count categorised for Verification status type

Default loan ratio for verification status

There is definite difference in the ratio, but it is not much. We can say there is a weak relationship between verification status and loan status

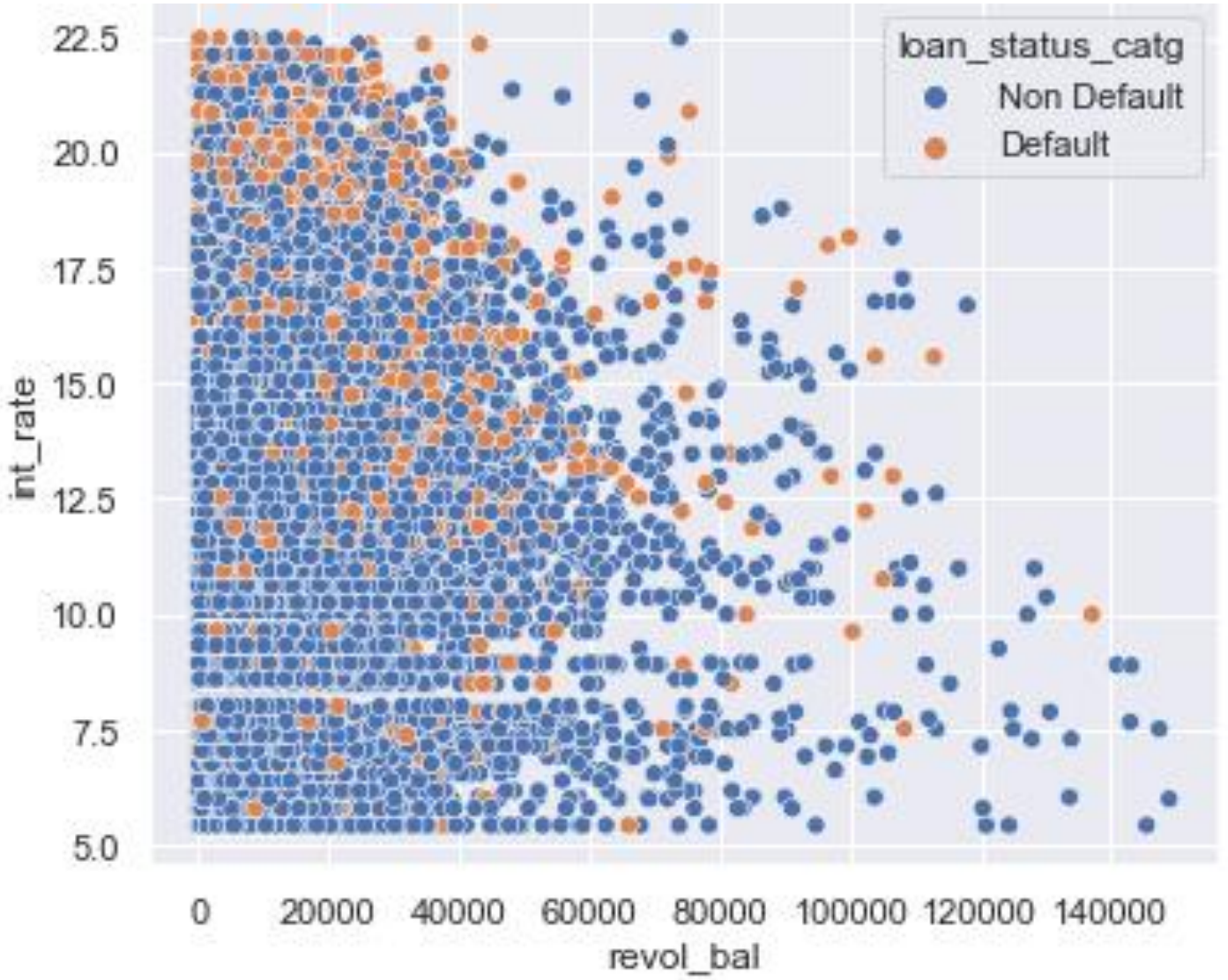


Are there more hidden patterns?

- Looks like we have covered most of the pattern to identify defaulters and default loans
- Let’s check few more columns and see if there is any pattern
- Credit revolving balance in conjunction with interest rate
- Annual income in conjunction with interest rate

Revolving balance and interest rate with respect to loan status

More defaults towards higher interest rate irrespective of total credit revolving balance



Annual income and interest rate with respect to loan status

Default loans are spread throughout the annual income range, but mostly on the higher interest rate, and we already saw this earlier. Basically annual income is not playing any role here

Summary

- Loan with high interest tend to be defaulted
- Loan term is correlated with loan status as well. Customers with long term loan tend to be defaulters as compared to customers with short term loan.
- Customers who have paid large amount of total payment of loan (total_pymnt) tends to be non defaulter and vice versa is also true
- Customers with lesser last payment paid(last_pymnt_amnt) towards their loan tend to be defaulters and vice versa
- Loans with LC assigned loan grade A and B tend to be non default
- Loan with LC assigned loan sub grades under A and B are likely to be non default (same conclusion as above point)
- Loans with last_credit_pull_m(month when LC pulled credit for loan) as May, tend to get defaulted
- Loans taken for small businesses tend to be defaulted. The reason could be small businesses are risky and they couldn't make profit and hence defaulted
- Loans taken from NV, AK and SD state has higher default ratio