LENDING CLUB CASE STUDY

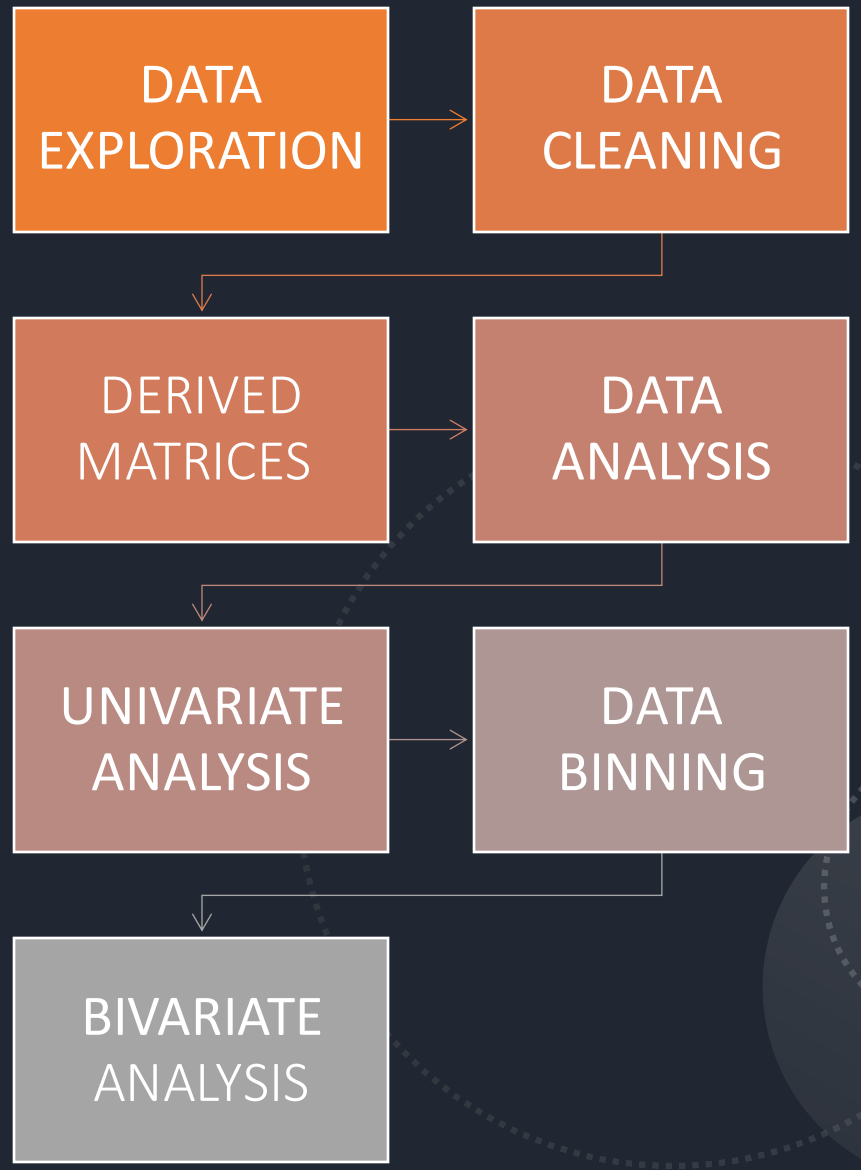
- Masters in Artificial intelligence and Machine learning – PGDP

- Ashwini Abhang
- Sailesh Bathala

Risk Analysis of Consumer Finance Company

Risk Analysis is important for any financial organisation that minimises the risk associated with the organisation decision. The exploratory data analysis helps in identifying the risk factors. The Lending club has given the loan dataset from 2007-2011 to analyse and to understand the various factors that causes loan to get defaulted. The dataset is loaded and various steps of EDA like data cleaning, missing value imputation and univariate, bivariate, multivariate analysis which are performed and proper visualisation is performed to identify the risky variables.

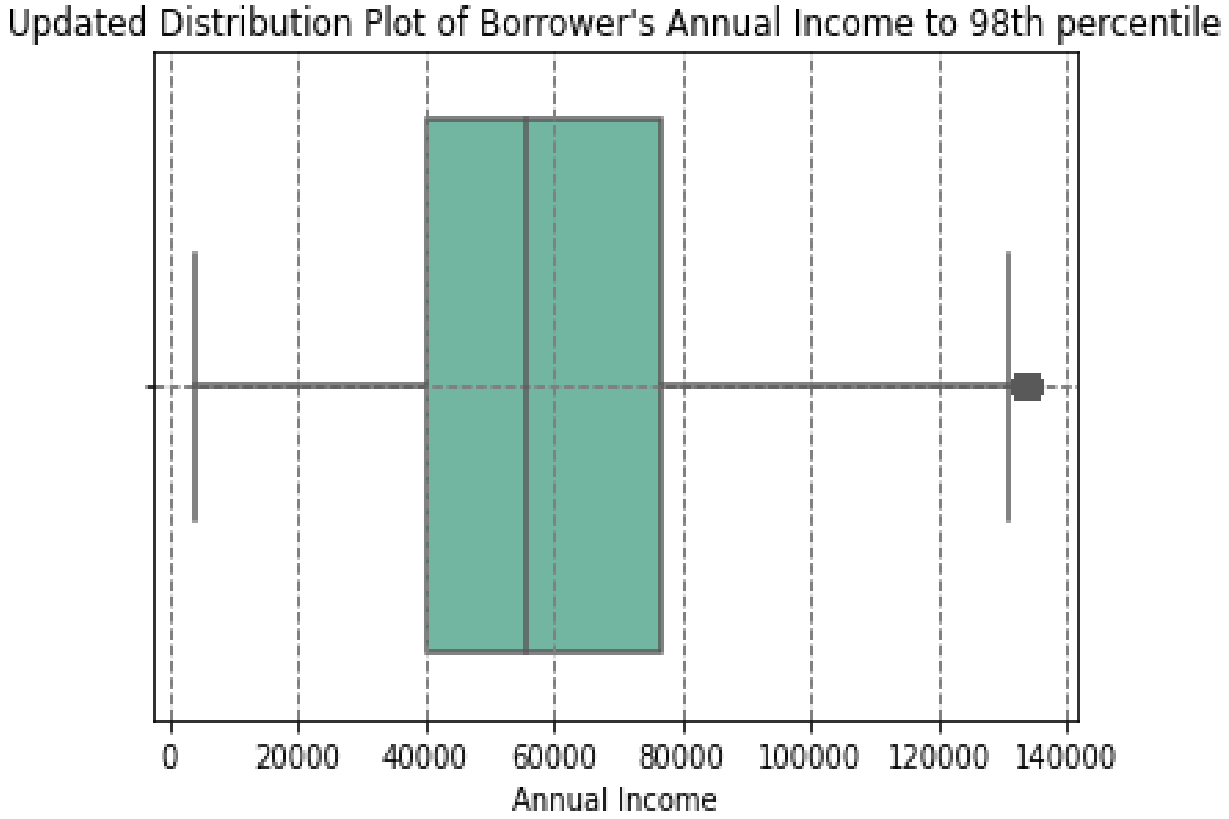
Problem solving methodology



DATA CLEANING

- We have loan.csv dataset on which we need to do data cleaning.
- Firstly, we'll fix the rows and columns.
- As we'll find the null values (NA) or zero (0) value in the dataset which are of no use to us.
- Then we'll find all the invalid values and missing values and fix them.
- After this we will get a much better dataset to do the further processes.
- Based on the definition of columns given in Data Dictionary the columns which are unlikely to impact the loan status are dropped.
- All the Missing values in the dataset are imputed.
- Data is formatted i.e., converting the data to appropriate format like int, float ,date time etc. e.g, loan["loan_amnt"] = loan["loan_amnt"].astype('float').

- This plot gives us a clear picture of outliers. By observation, we see that, after 2million, there is a huge discontinuity in the distribution. We can approximately consider 98% as the threshold. Let us get the quantile info to give us an idea of why 98% is chosen as the threshold.
- Now it is safe to say that after the 98th percentile, data seems to go off from the distribution, so let us now remove the threshold values.



DERIVED MATRICES

- int_rate is in object data type which is incorrect for calculation purpose. We will be changing it to float64
- loans_dataset.int_rate = pd.to_numeric(loans_dataset.int_rate)
- *#to_numeric by default converts to float64*
- loans_dataset.info()

- Lets now move to Derived Matrices
- We can pick the issue_d column for this.
- Derived matrices is important in data analysis.
- We can understand more about type of columns, we can extract its attributes.
- loans_dataset['issue_yr'] = pd.to_datetime(loans_dataset.issue_d, format = '%b-%y').dt.year

DATA ANALYSIS

- Basically after the data cleaning is performed, we will do the data analysis.
- There are few data analysis tech which are we are going to perform i.e.,

1. Univariate analysis
2. Bivariate analysis.

UNIVARIATE ANALYSIS

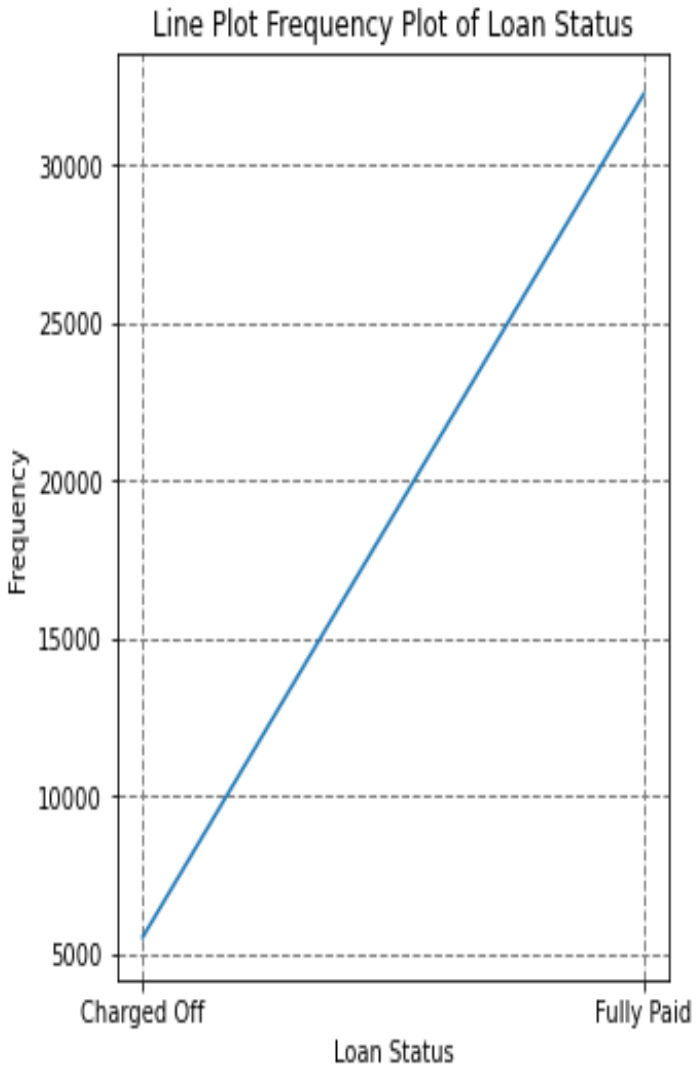
- Firstly, we will check all the ordered and unordered categorical variables.
- Then, we'll check all the Quantitative / Numeric variables in the dataset.
- Then we'll do data binning.

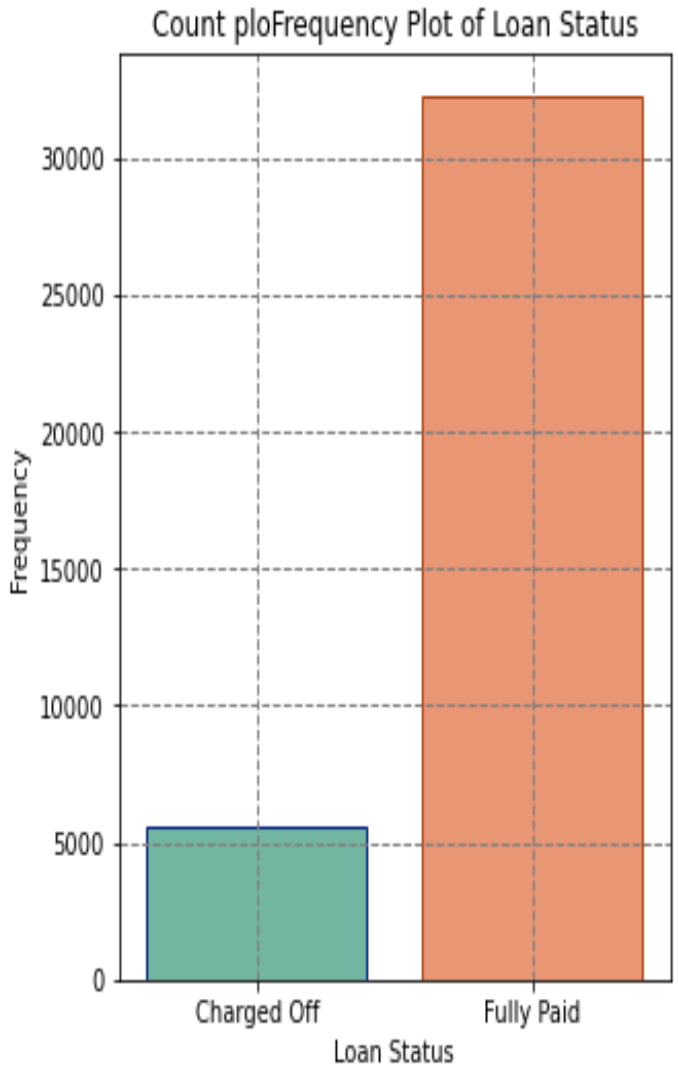
Univariate Analysis

Let us perform univariate analysis on the following columns:

1. LOAN STATUS

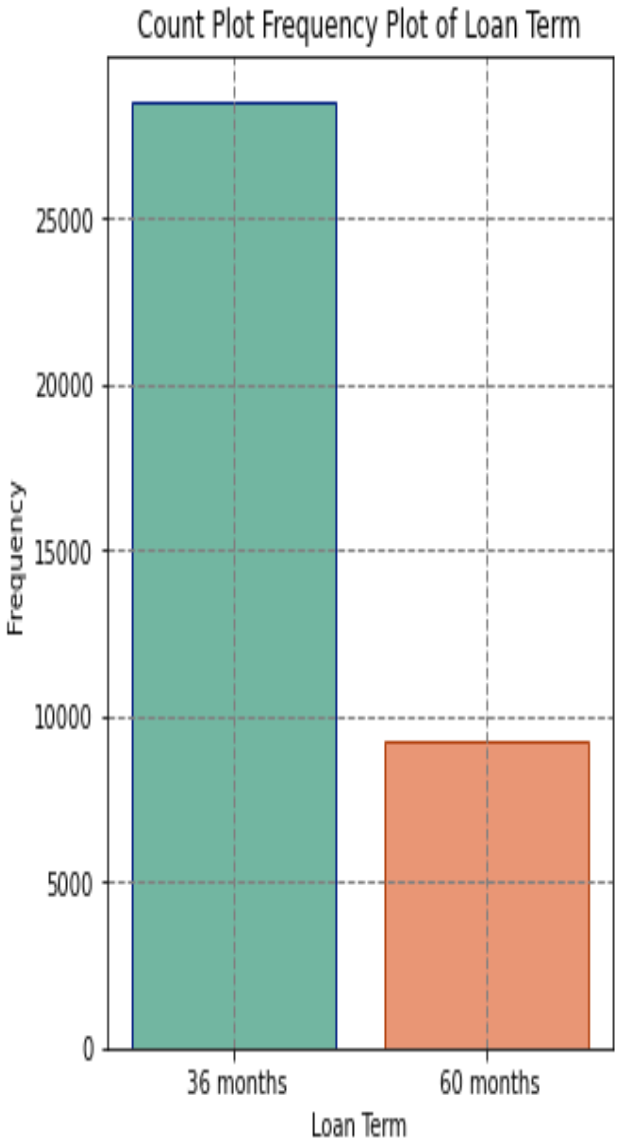
Out of 37042 values, 5448 are charged off and are our main area of analysis.

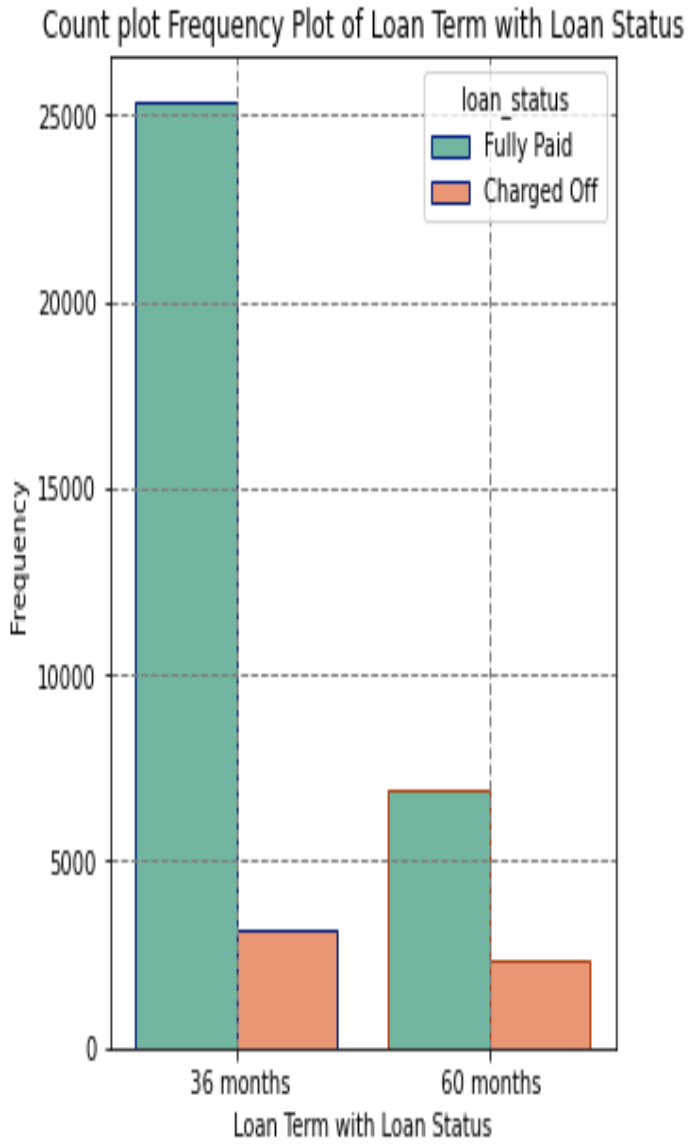




2. TERM

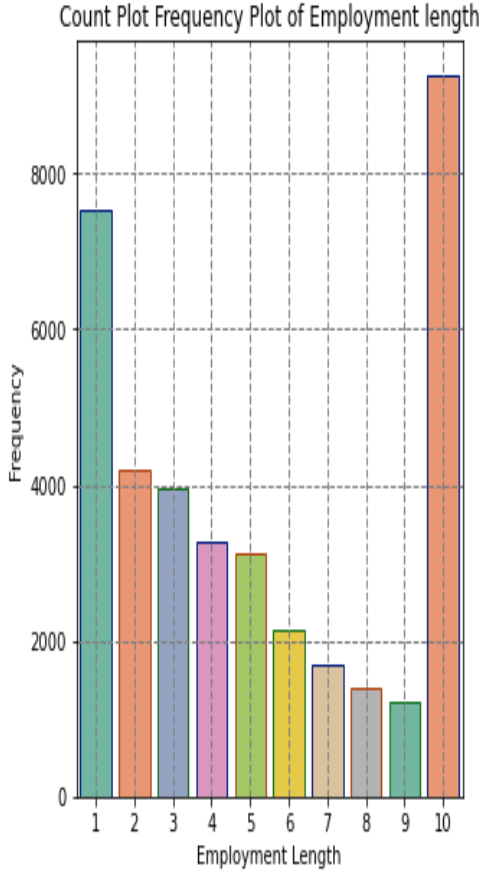
Most of the people who apply for the loan term of 36 months tend to be charged off or defaulted. 3171 people who opted 36 months defaulted/charged off against 2362 people who opted for 60months.

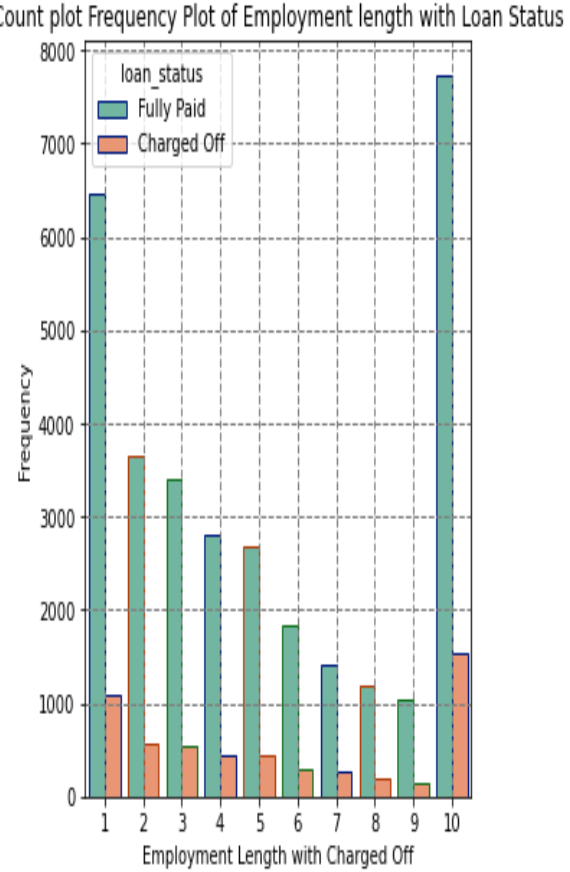




3. EMPLYOMENT LENGTH

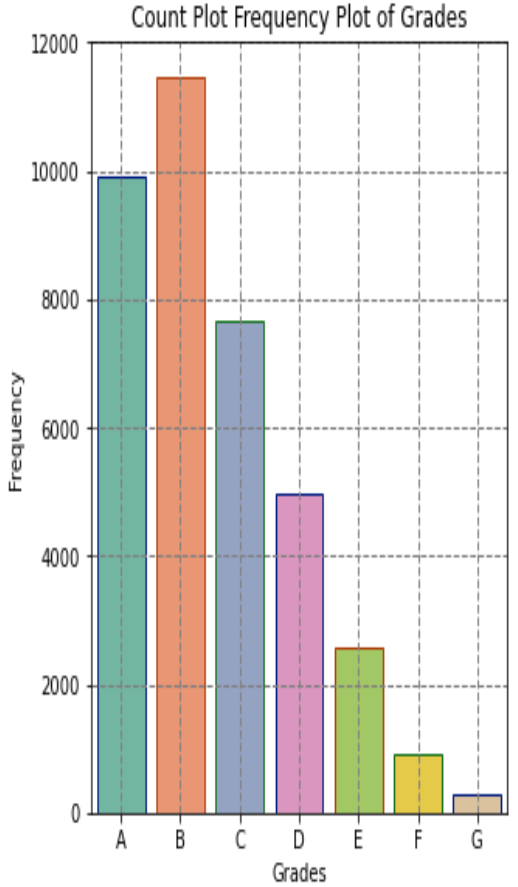
From the given plots we see that people who have the employment term as 10+ years and 1 year tend to be defaulted/charged off more than other employment terms.

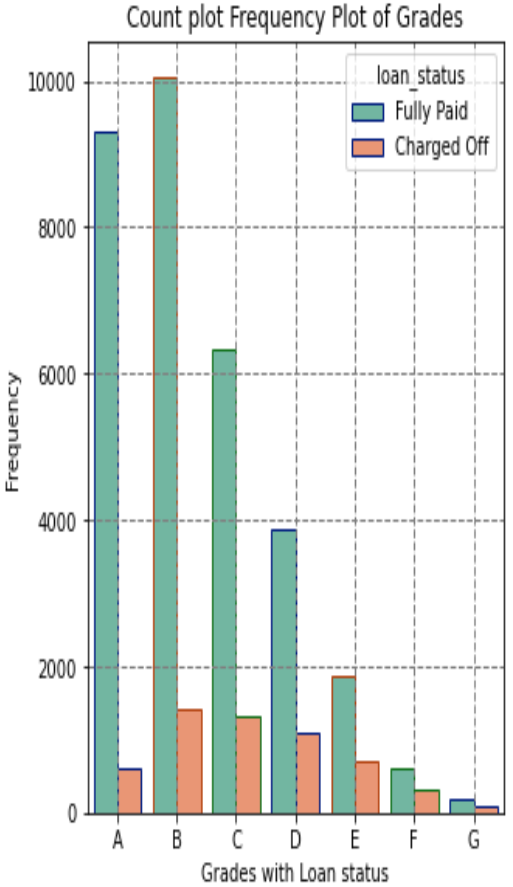




4. GRADES

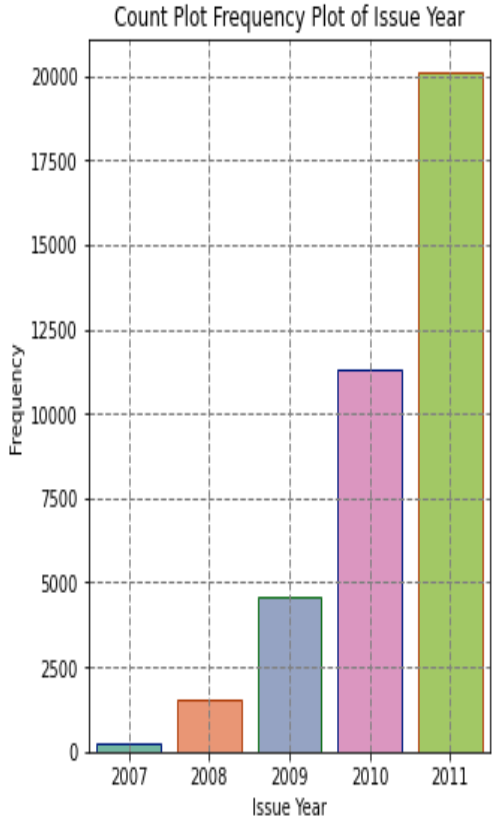
We can infer from the above plots that, grades B, C & D tend to be defaulted/charged off more than the other grades.

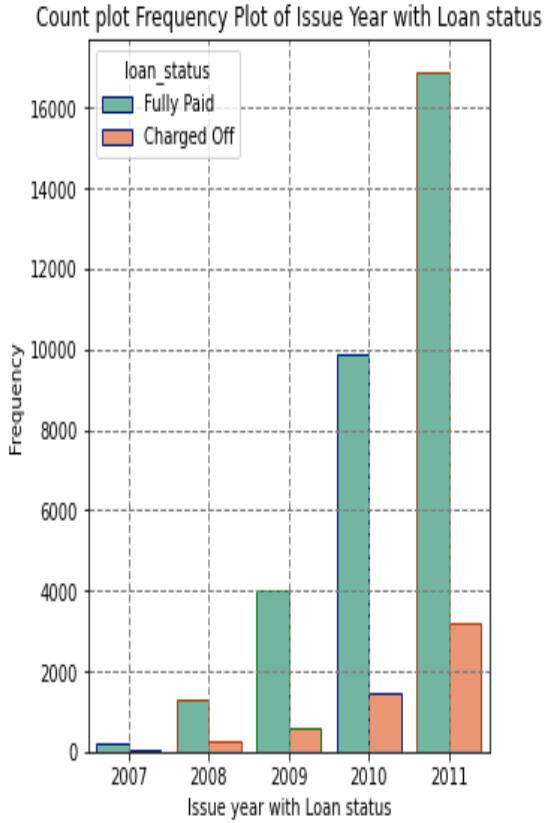




5. Issue Year

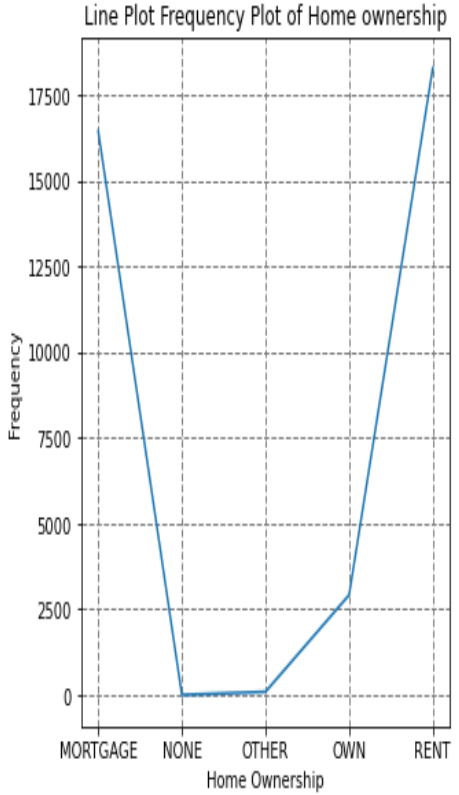
We can see that in the year 2011, the loans issued tend to be more defaulted than in the previous years followed by year 2010.

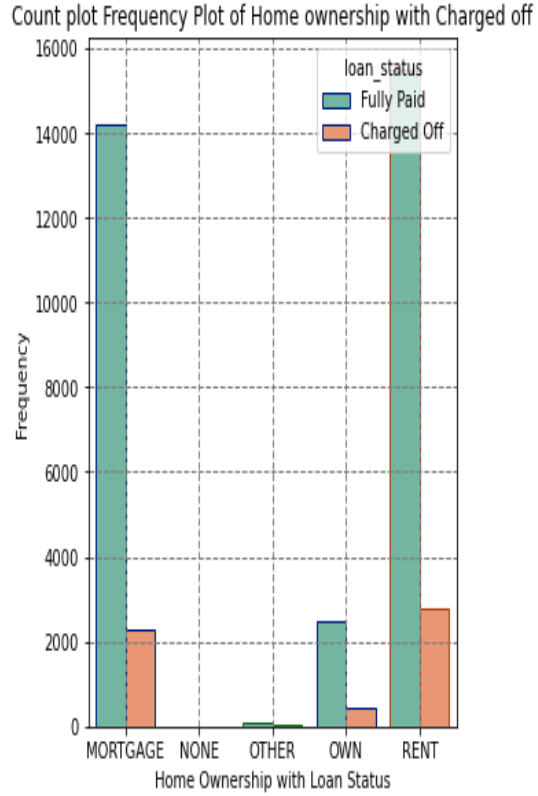




6. Home Ownership

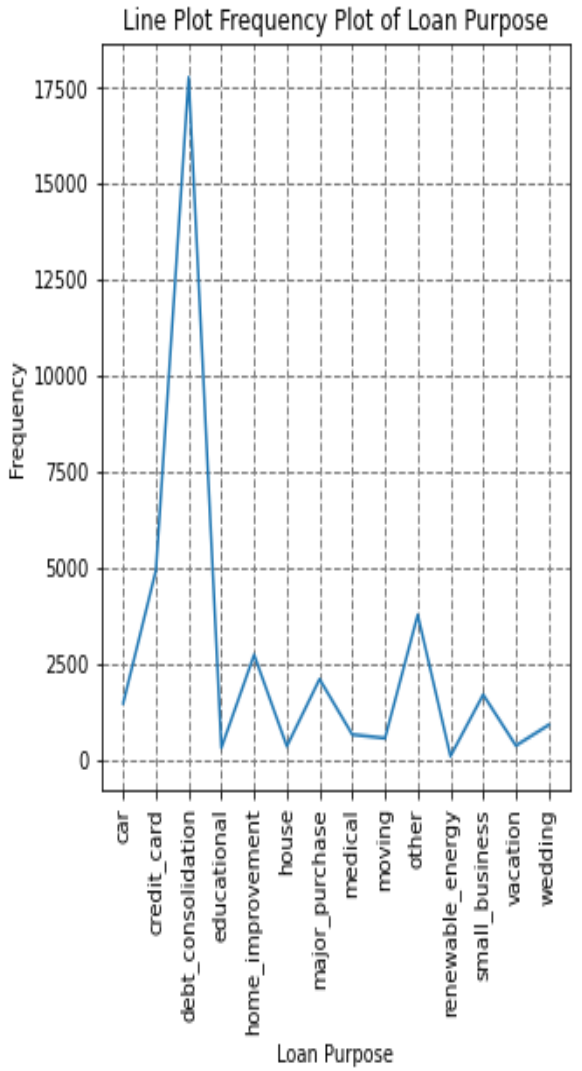
We can observe from the plots that there is a high occurrence of Defaulted/Charged off loan takers amongst Mortgage and Rented home ownership.

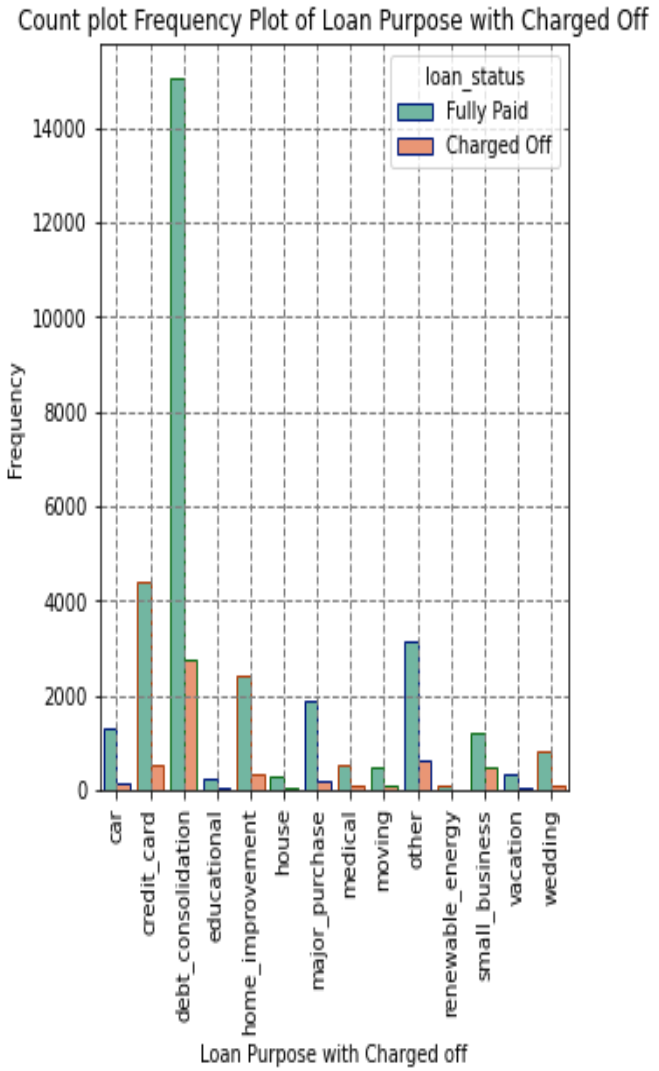




7. Purpose

- From the above plots we can conclude that, the main loan purposes that have been charged off are
- debt_consolidation
- credit_card
- other
- But in terms of higher percentage, the loan purpose that has higher charged off rate is small_business





BIVARIATE ANALYSIS

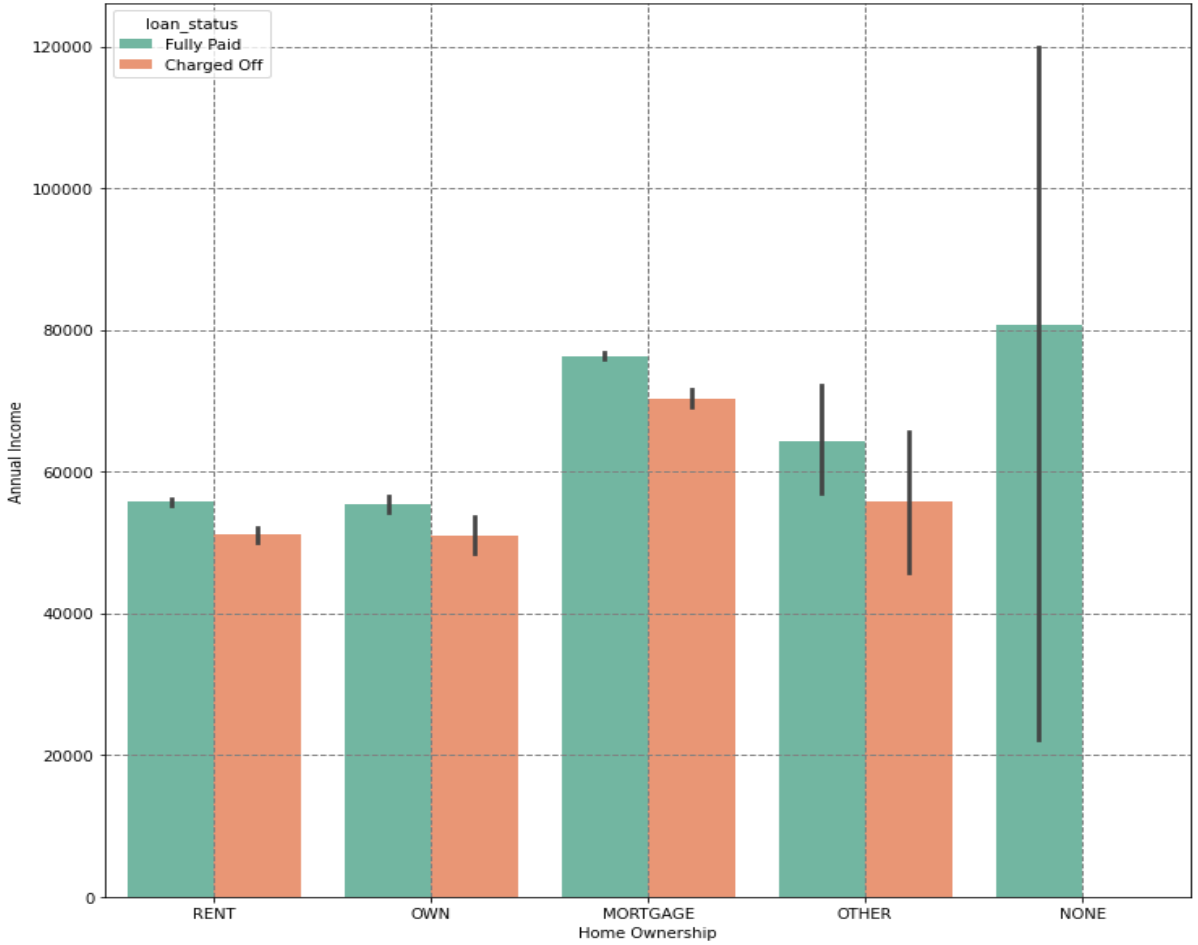
- Bivariate analysis on continuous variables.
- Bivariate analysis on categorical variables.
- Before we do the bivariate analysis, let us do the binning of certain columns, as we have a lot of data to plot and observe in a graphical representation.

DATA BINNING

- **A. Annual Income Binning**

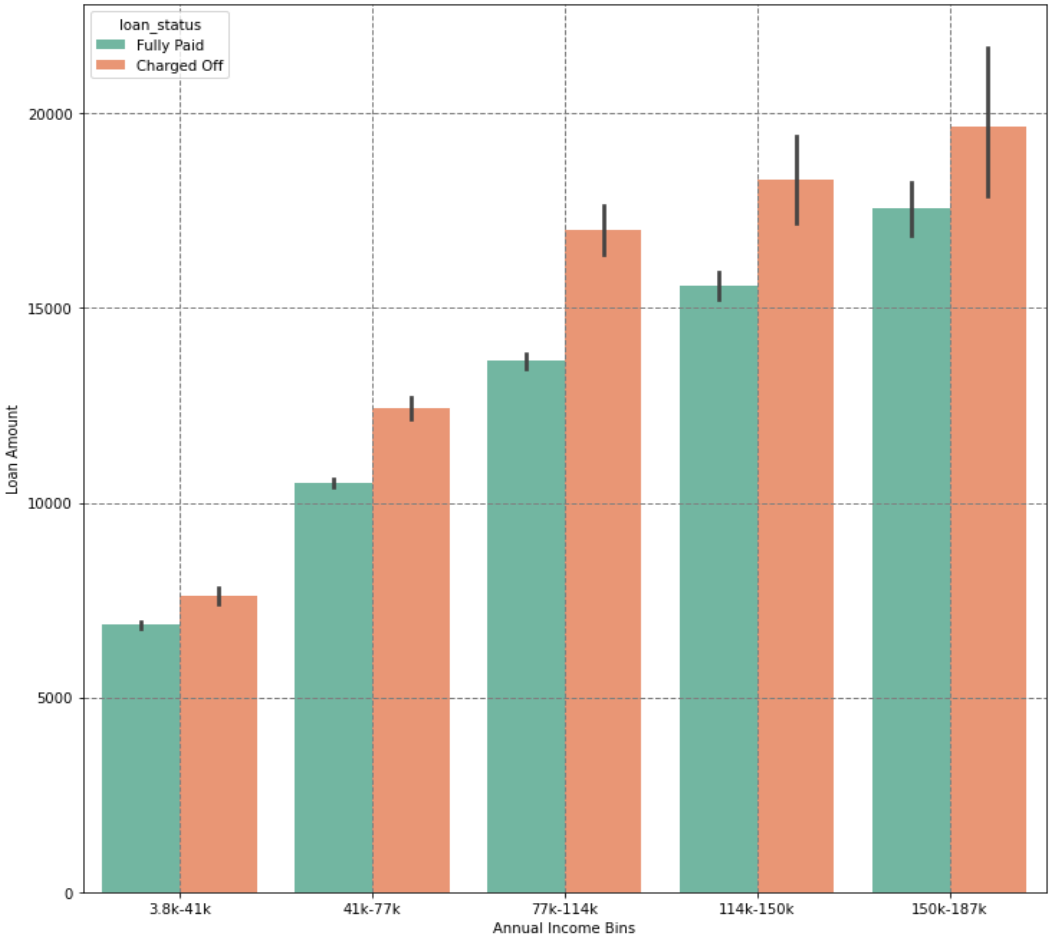
1. Annual salary with home ownership

- We observe that applicants who opt for Home ownership of 'Mortgage' and have an income range of 60,000 to 80,000 tend to default more than other combinations of Home ownership and Annual Income.



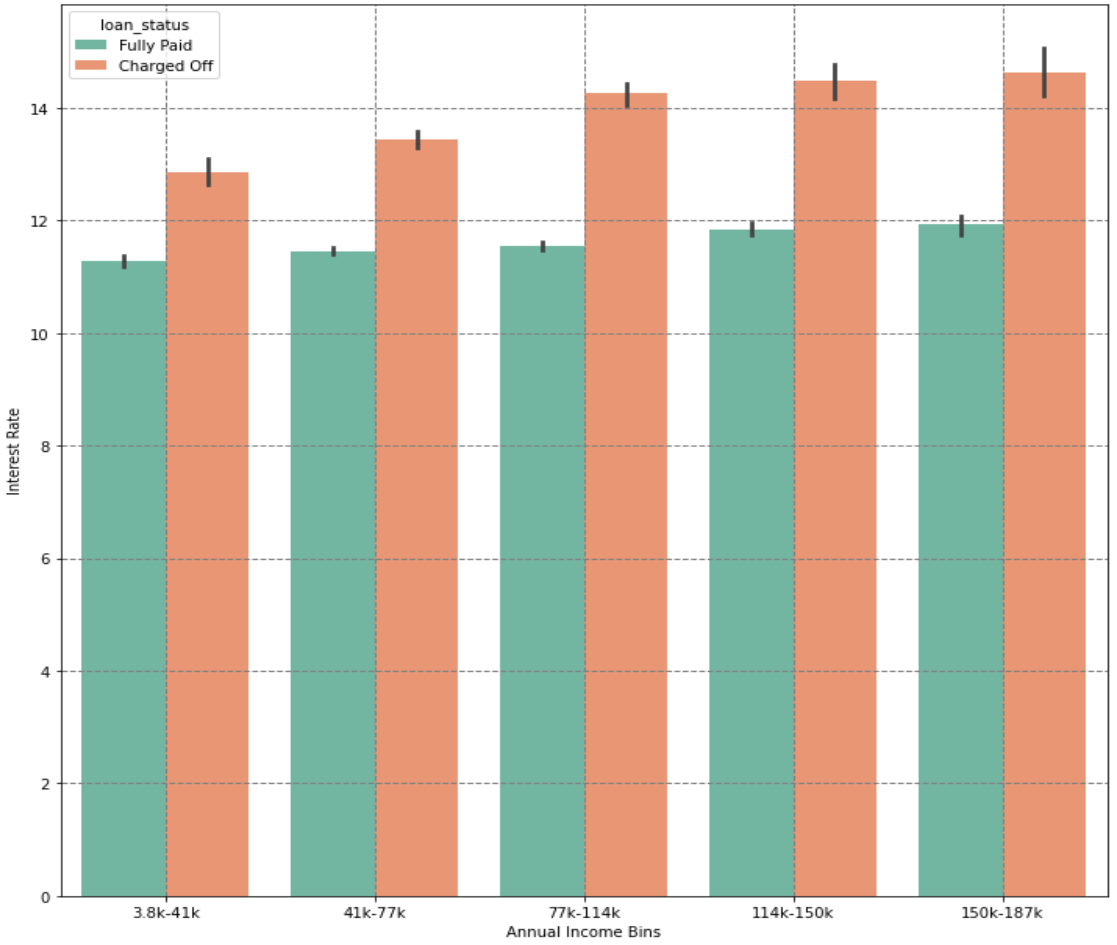
2. Annual Income with Loan amount.

We observe that applicants who opt for higher loan amount of have higher income range.



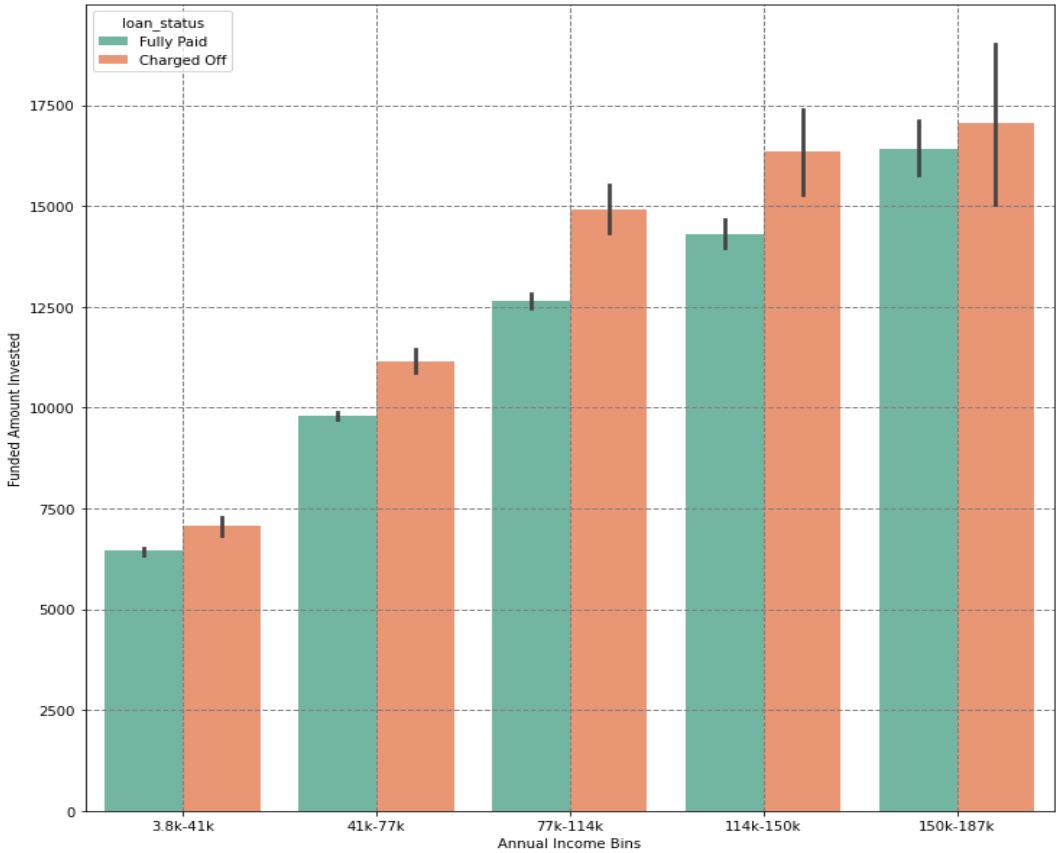
3. Annual Income with Interest Rates

We observe that applicants who have higher income range are charged higher interest rates.



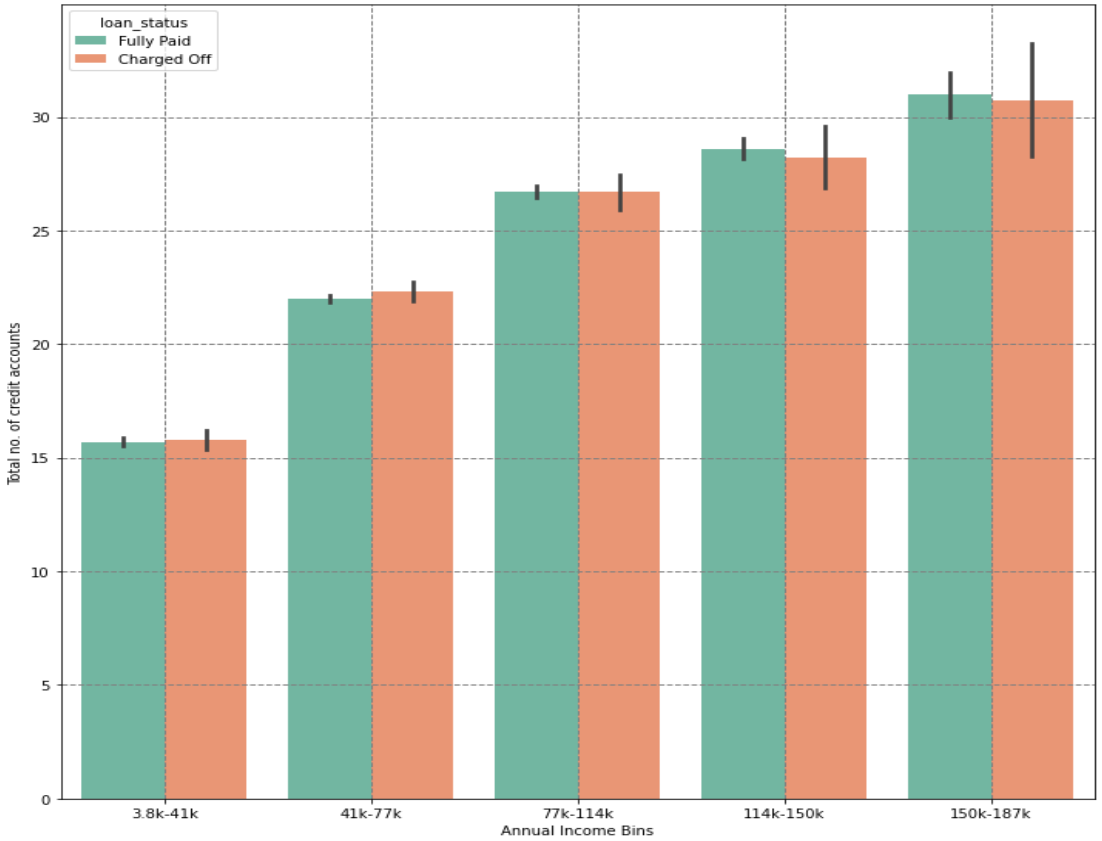
4. Annual Income with Funded Amount Invested

- When the Annual Income is in the range 77k to 114k and the funded amount invested is in range 12,500-15,000, the chances of a person being defaulted or charged off is relatively more.



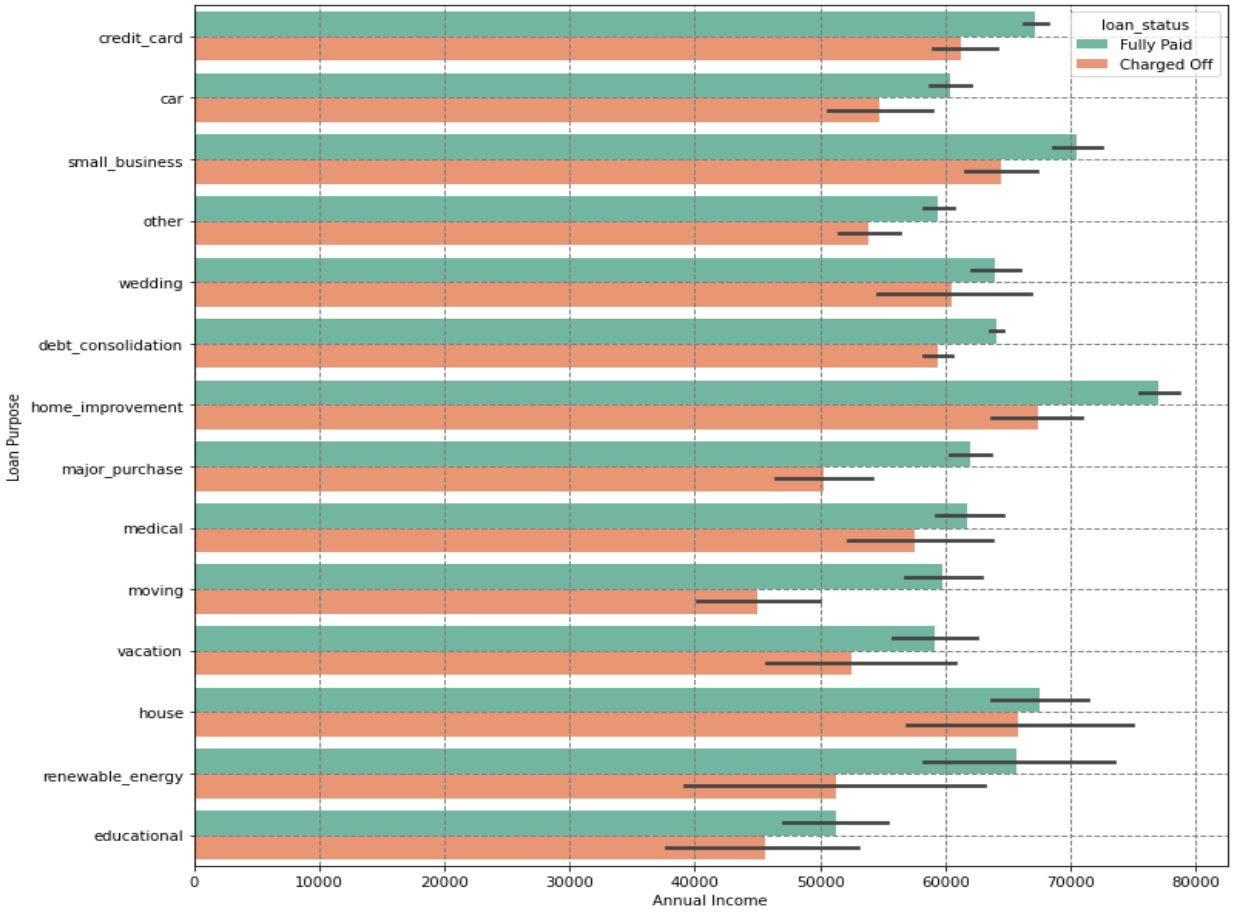
5. Annual Income with Total Credit accounts (total_acc)

- When the Annual Income is in the range 41k-77k and the total number of credit accounts is between 20 and 25, the chances of a person being defaulted or charged off is relatively more.



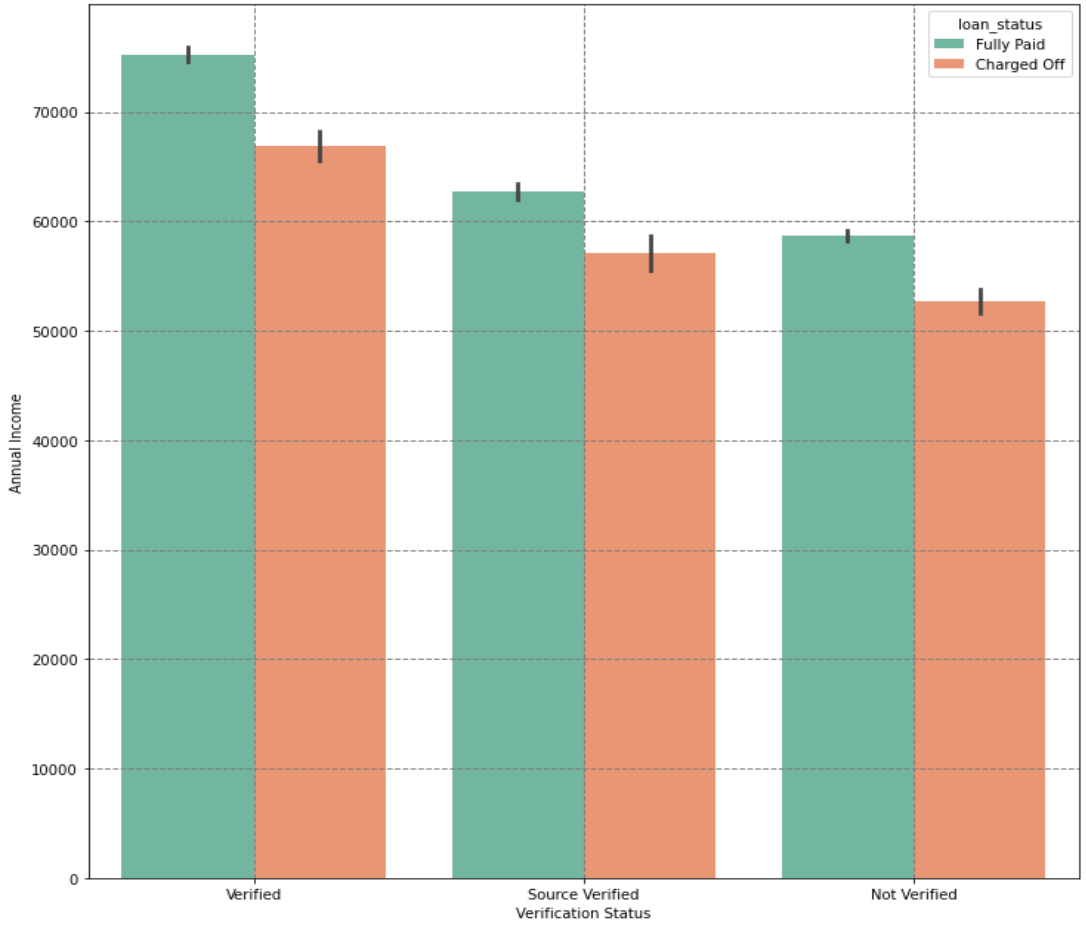
6. Annual Income with Purpose

- People who take loan with a purpose of 'Home Improvement' followed by 'House' and 'Small Business' have the highest income range of 60k to 70k



7. Annual Income with Verification status

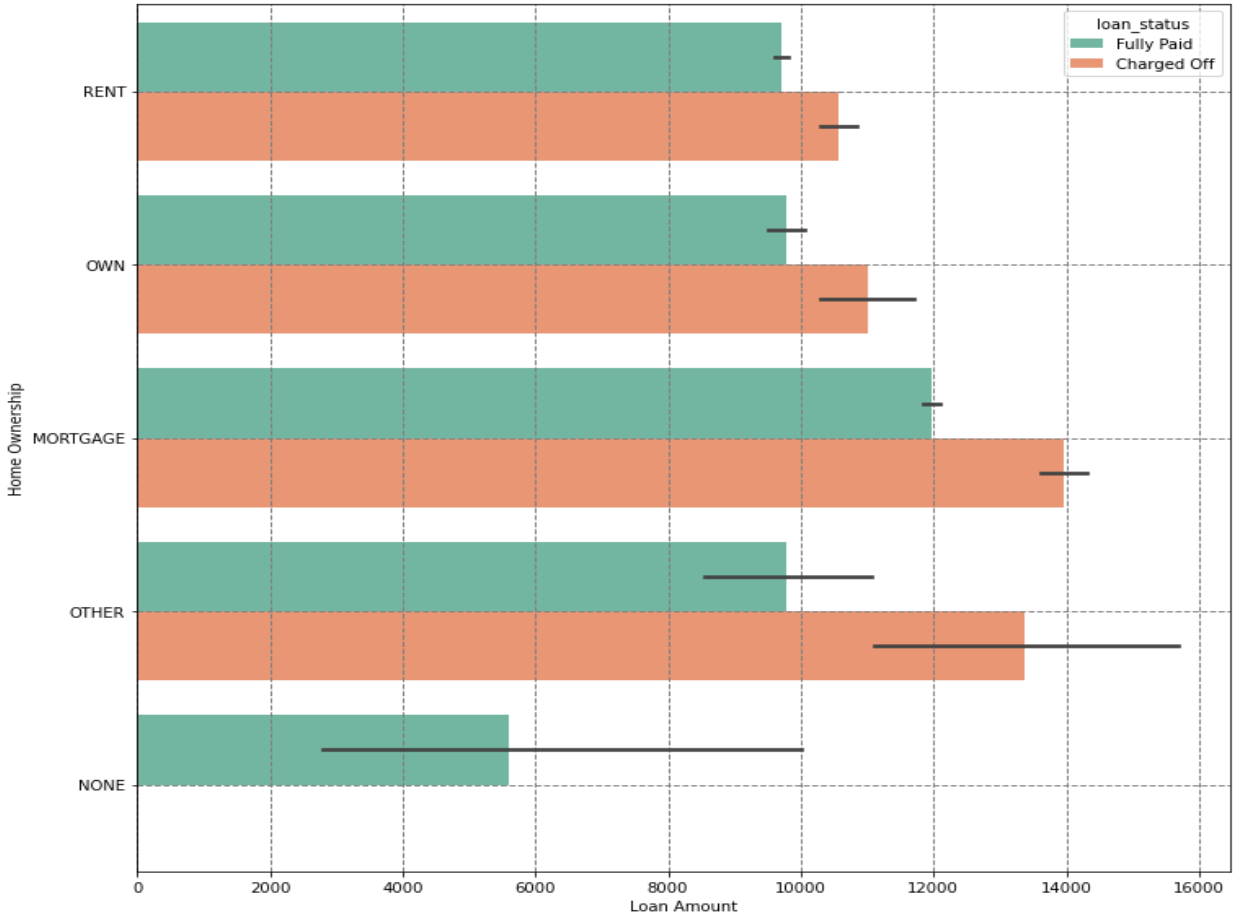
- Those with higher annual income are verified more compared to people with lower income range.



B. Loan Amount Binning

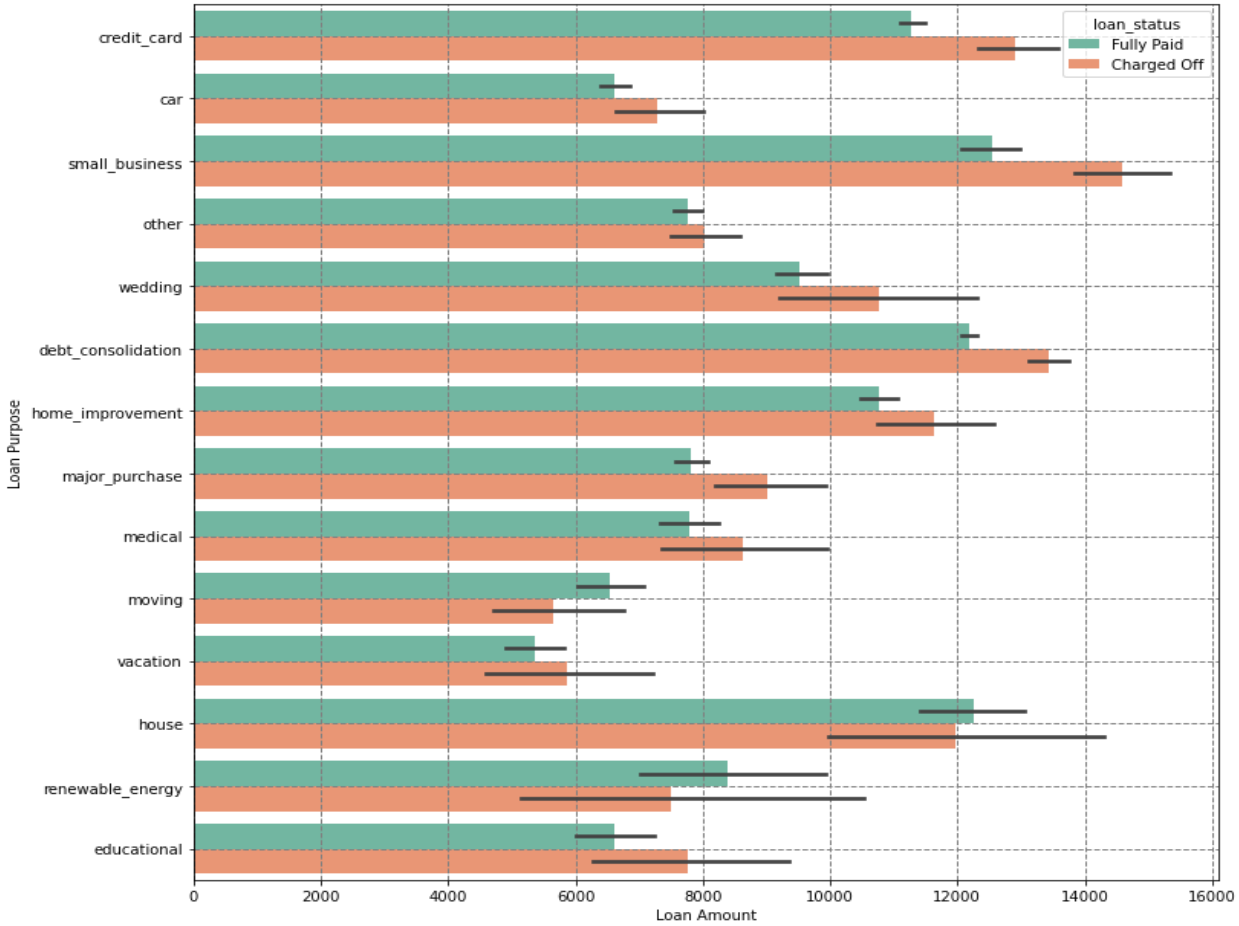
1. Loan Amount with home ownership

- Applicants whose Home ownership is Mortgage, have the highest Loan amount range of 12,000-14,000.



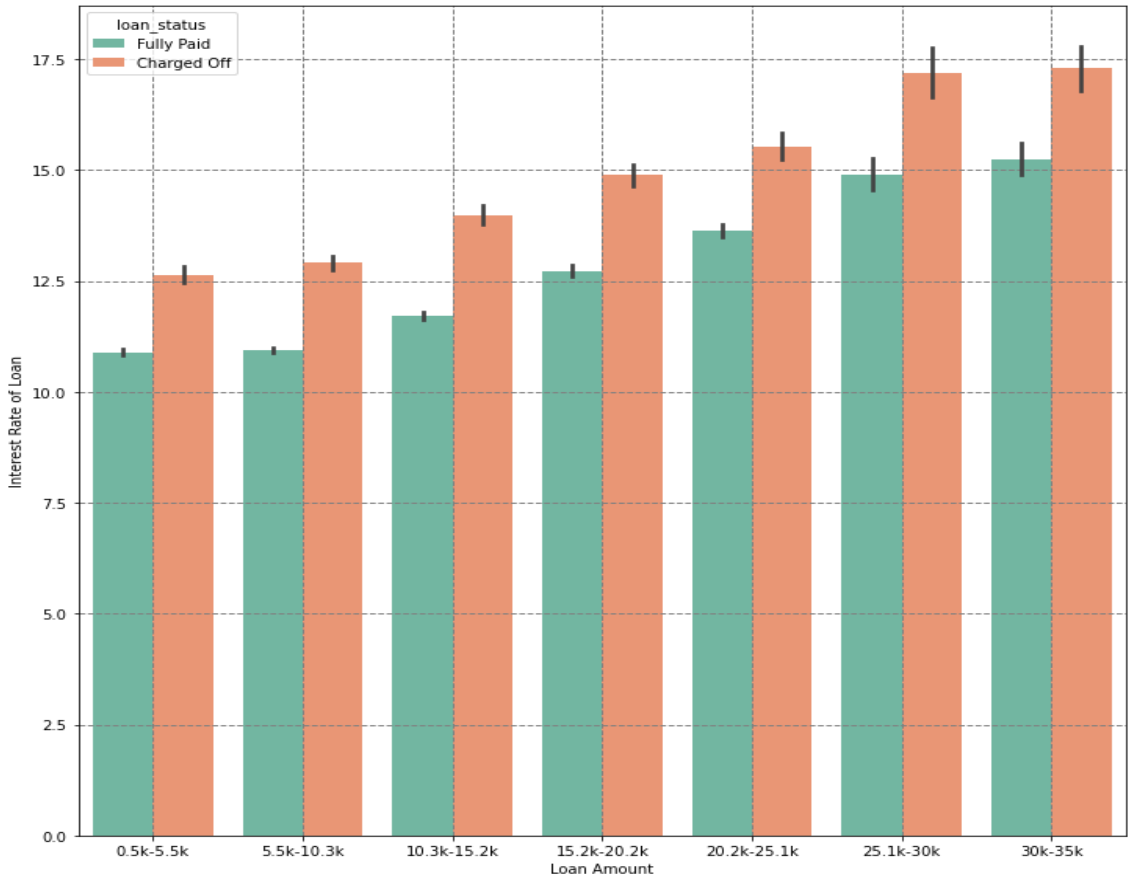
2. Loan Amount with Loan purpose

- Applicants whose loan Purpose is Small business, have the highest Loan amount range above 14,000 and also are more likely to be defaulted/Charged off.



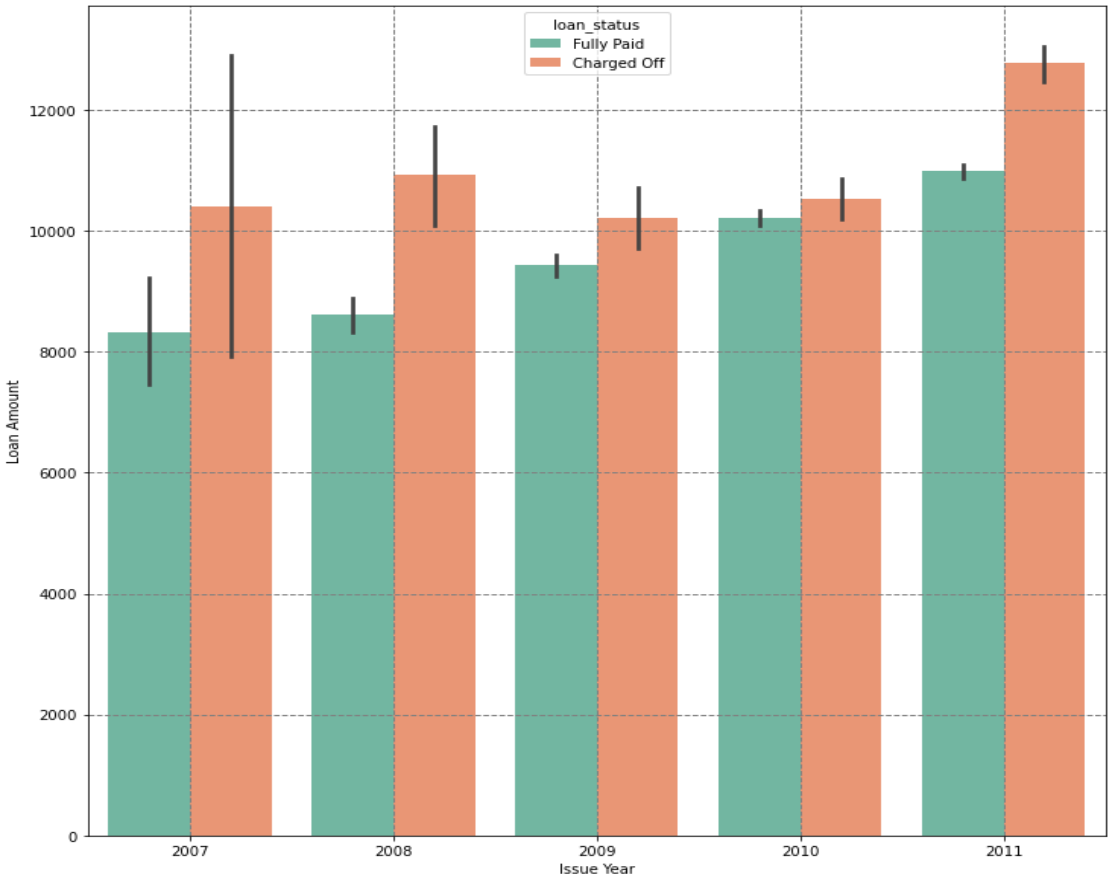
3. Loan amount with Interest rate

- Applicants whose loan Amount is higher, have the higher Interest Rate range. Here, we see that the Loan amount of 30k-35k has the highest Interest Rate of 15% to 17.5% and also are more likely to be defaulted/Charged off more as the Loan amount increases.



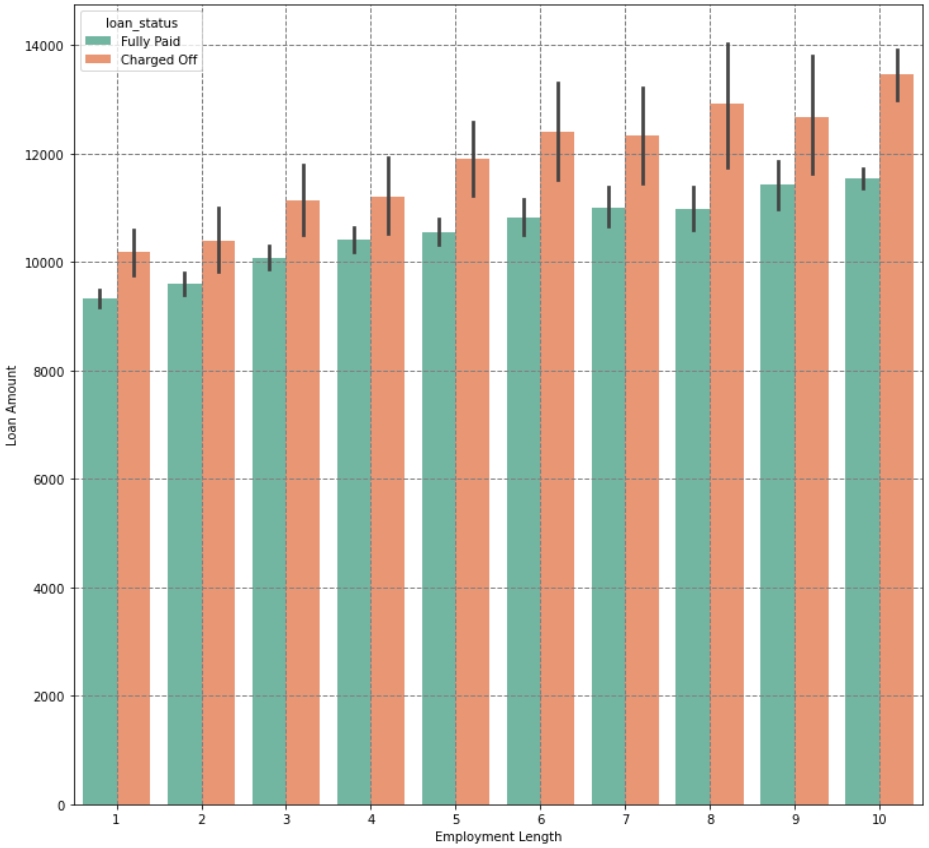
4. Loan amount with Issue Year

- Applicants who took the loan in the year 2011 have taken higher loan amount when compared to previous years. This also shows that the year 2011 has higher defaulters.

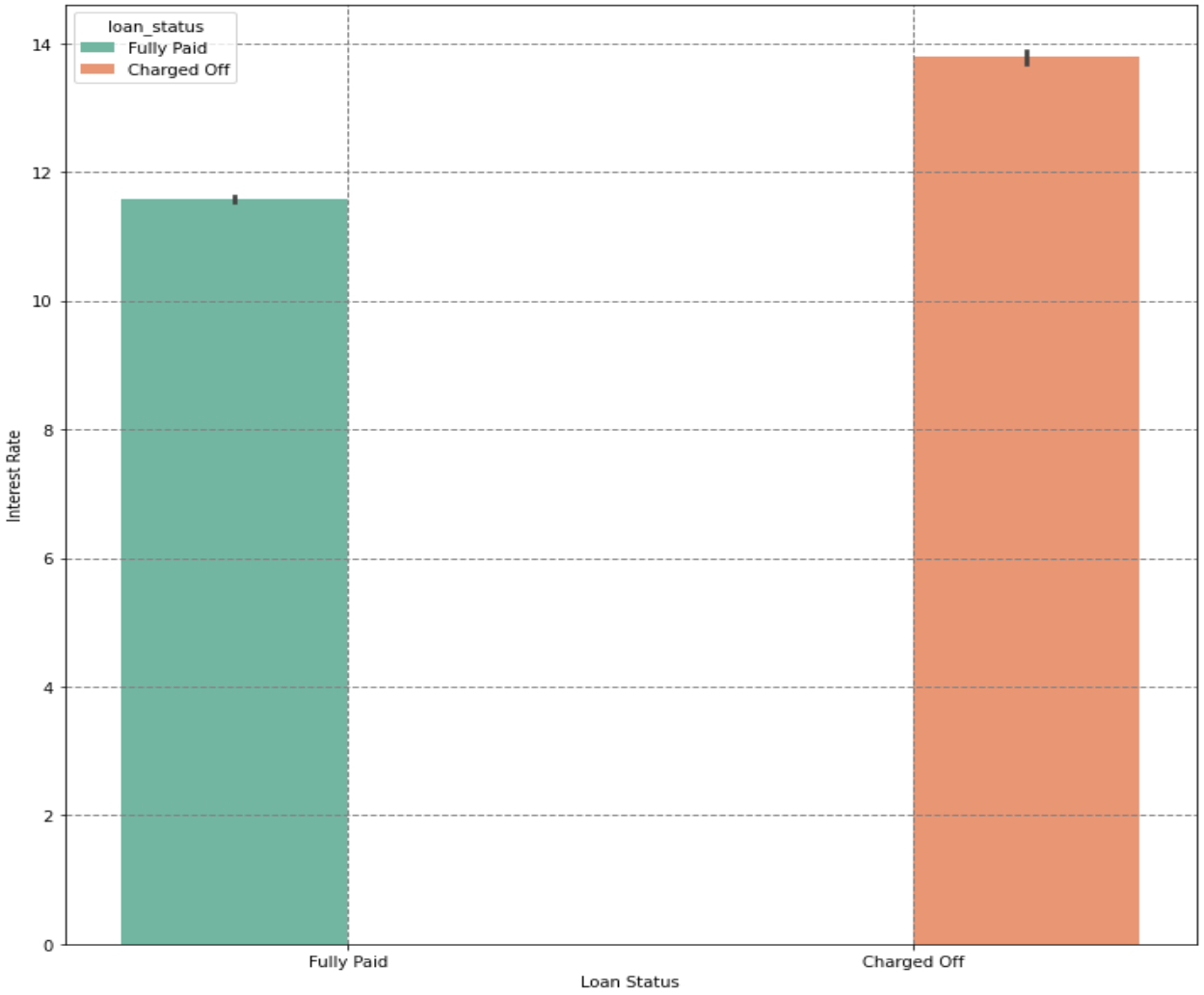


5. Loan Amount with Employment Length

Applicants whose Employment Length is 10 or greater than 10 years have taken higher loan amount of 12,000 to 14,000 when compared to Employment years less than 10 years. This also shows that the ones with Employment length of 10 years had has higher defaulters.



We can finally conclude that when the interest rates are high, we have higher chances of defaulting/charged off.



Conclusion

The Main risky factors are:

Loan taken for Dept_consolidation is more likely to get defaulted.

Borrowers who are working for more than 10 years are more unlikely to repay the loan.

Borrowers who are on rent or mortgage are are more unlikely to repay the loan.

60 Months term loans are better than 36months term loans.

Grades, interest rates are the other factors.