LENDING CLUB CASE STUDY

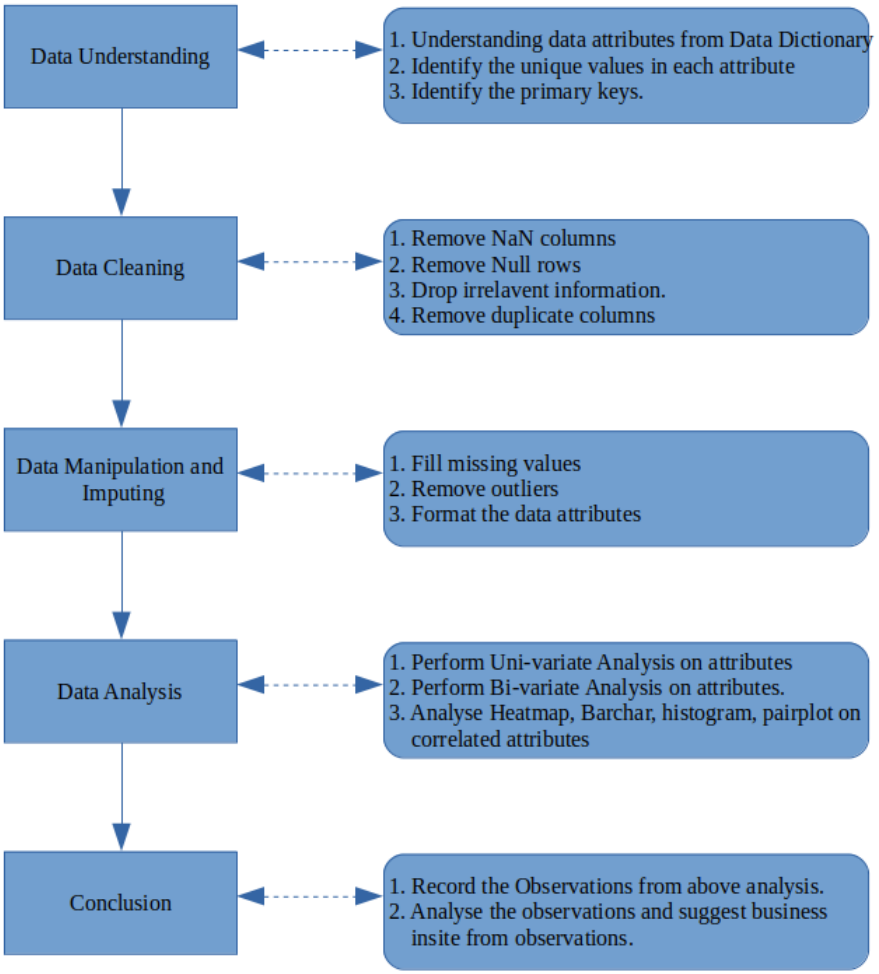
SUBMISSION

Team Member:
Kiran Sutar
Prateek Toshniwal

Objective

- Company is having good marketplace in financial loan sector and facilitats for personal loans, business loans, and financing of medical procedures. Borrowers can apply the loan by providing some basic personal informations.
- Company has to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default and which can be analysed by company for risk assessment based of borrower’s portfolio.

Problem Solving Strategy



Data Understanding

- Analyse the data and understand the attributes available in data dictionary.
- Identify the Primary keys – id, member_id
- Identify the continuous and categorical attributes.
  - Continuous – Interest rate, loan amount, annual income
  - Catagorical – Grade, emp_length, home ownership
- Look at the unique values in each attribute.
  - Home ownership – Mortgage, Own, Rent
- Rename the columns to have meaningfull name.

Data Cleaning

- Remove random values – Columns: emp_title, desc, title
- Remove column with same values – Columns: pymnt_plan, policy_code, initial_list_status,, application_type, acc_now_delinq, chargeoff_within_12_mths, delinq_amnt, tax_liens
- Remove columns with null values – Columns: mths_since_last_record, next_pymnt_d, mths_since_last_delinq
- Remove columns with mostly zero values – Columns: pub_rec, out_prncp, out_prncp_inv, collections_12_mths_ex_med
- Remove the empty rows.
- Drop duplicate values columns
  - url - duplicate with id column
  - addr_state - duplicate with zip code
  - total_pymnt_inv - duplicate for total_pymnt
- Drop unique key attributes as these will not contribute to pattern analysis

Data Manipulation and Imputing

- Fill missing values with mean/ Median/ Mode values of attribute: Columns: Grade, emp_length
- Remove percentage symbol and make values to numeric – Columns: Interest rate, revol_util
- Fix the datatype like date and numeric – Columns: issue_d, earliest_cr_line
- Remove post loan approved attributes (Domain Driven)
- Create Derived Columns (Domain Driven) – Columns: Annual Income and loan amount ratio

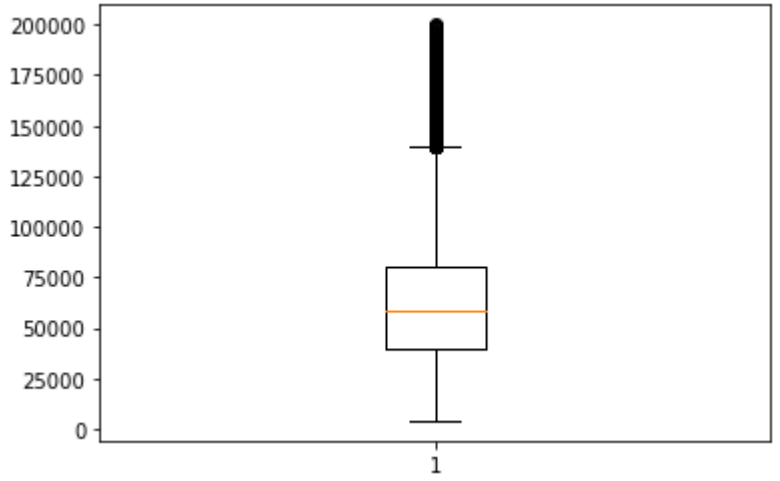
Analysis - Univariate Analysis

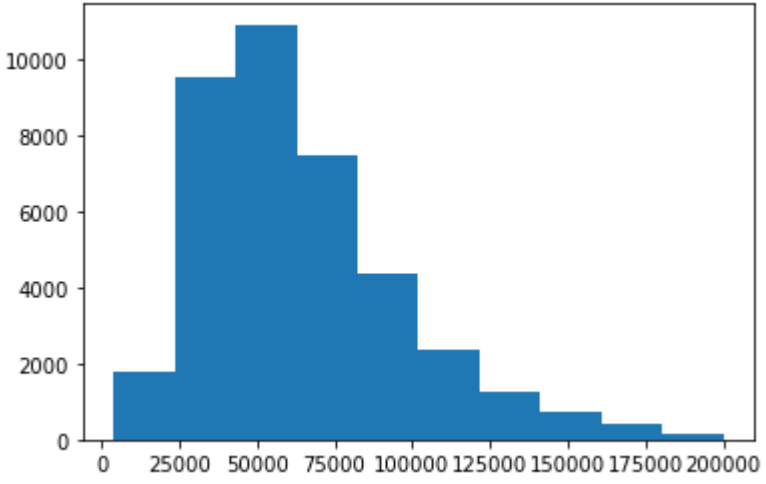
Analyse the following attibutes from categorical and continuous variables and indenfy the ouliners and perform the Univariate Analysis on the attributes with domain knowledge and correlation

- Annual income
- Interest Rate
- Loan Amount
- Grade
- Term
- Employee length
- Purpose
- Home ownership

Outliner handling

After examining the box plot and histogram for the annual income we found that there are 14 values above 10 Lacs and total 698 values above 2 Lac. We would consider the income below 2 lac for the data analysis as income above 2 lac is just 1.5% of whole data and can be treated as outlier.



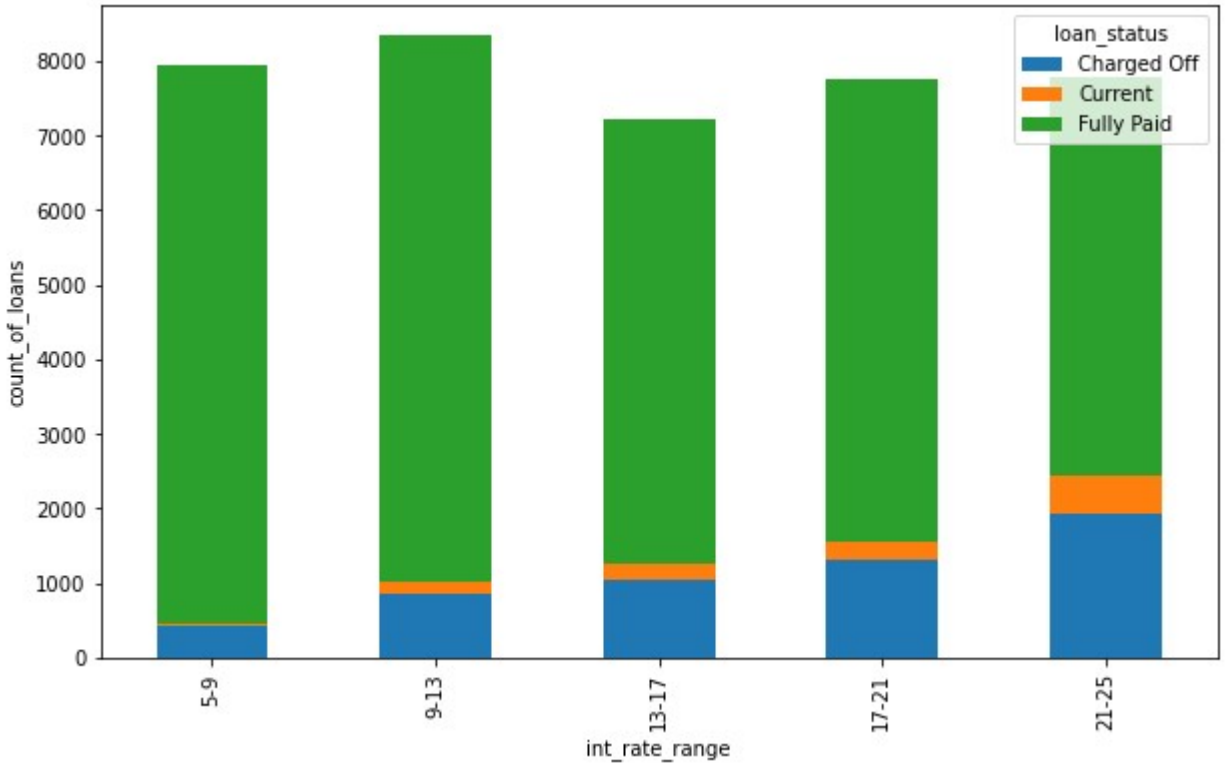


Analysis - Bivariate Analysis

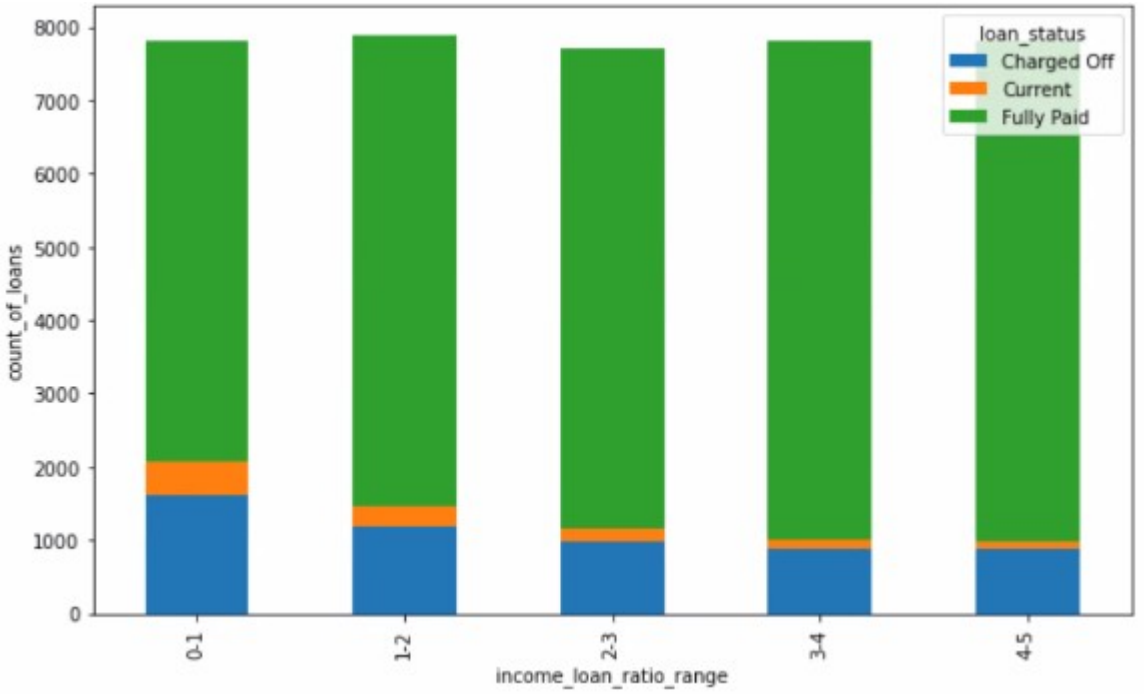
We found the following relationships between the attributes most relavent:

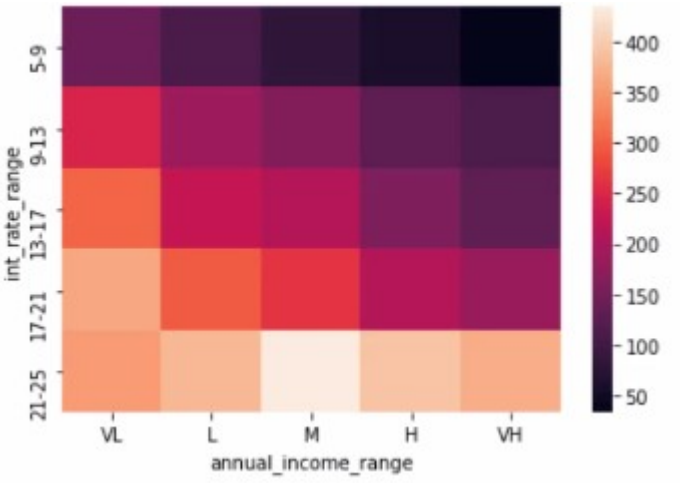
- Interest rate VS Loan status
- Annual income/Loan amount VS Loan status
- Purpose VS Loan status
- Employee length VS Loan status
- Annual income/Loan amount & Interest rate VS Loan Status

Note: For detailed explaination please refer Notebook.



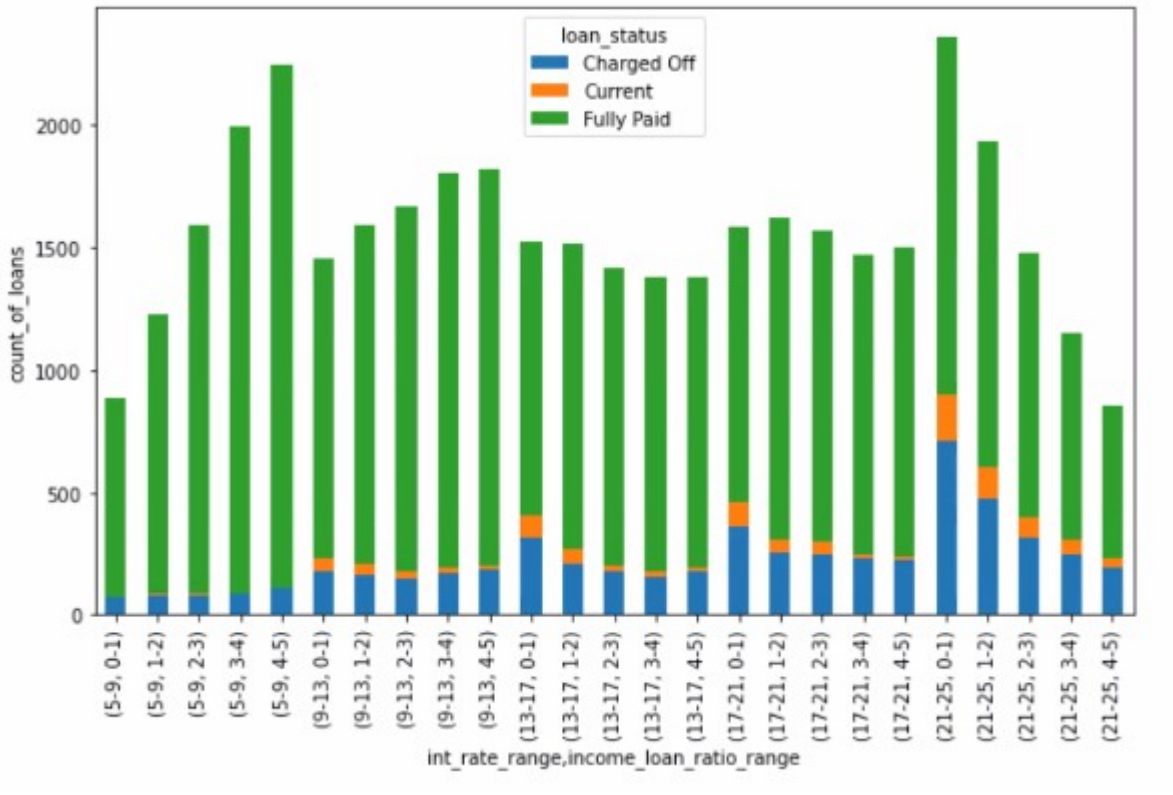
Inference -- The above chart clearly shows that the Interest rate has huge impact on the Charged off rate. Higher interest rate causes more probability of default.



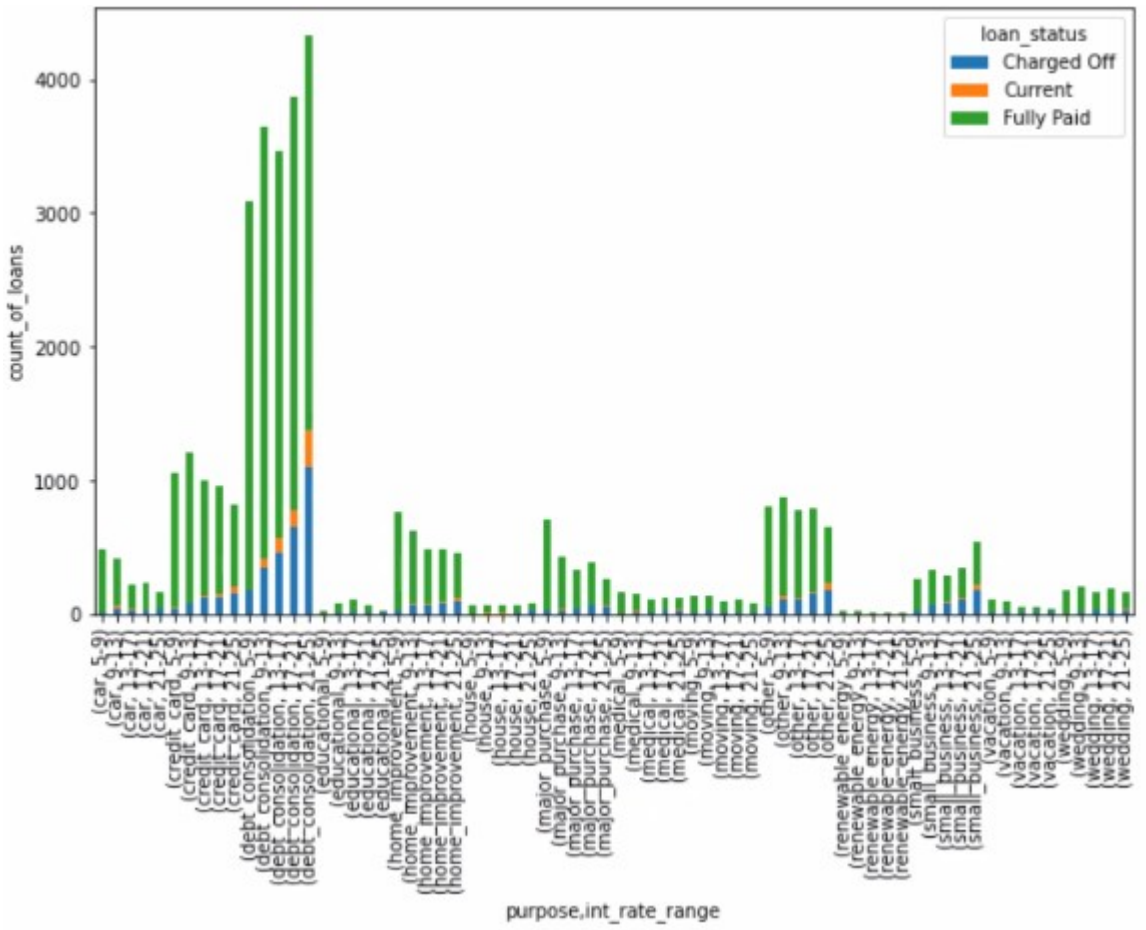


Interest rate vs Annual income for defaulters only

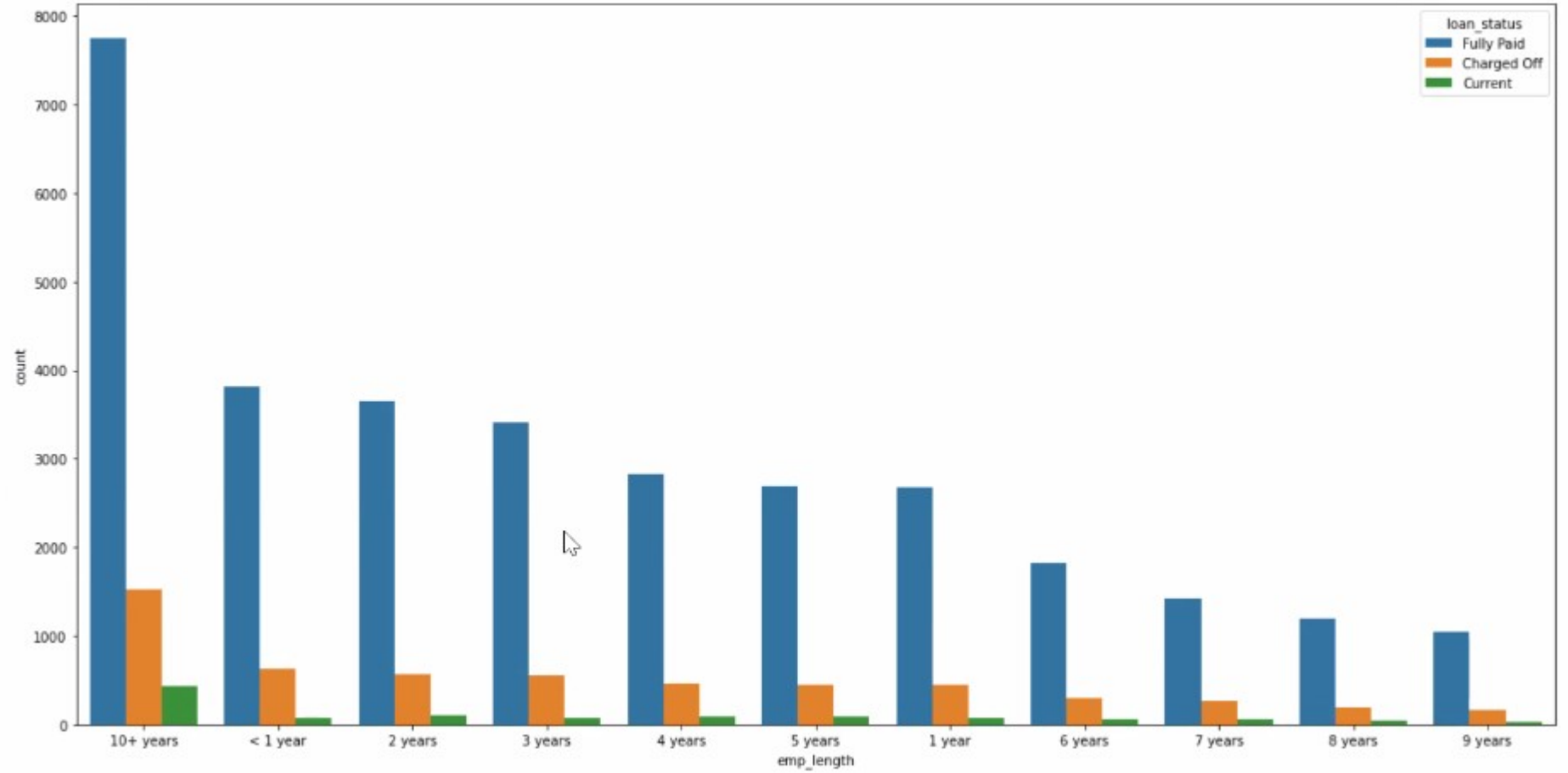
Inference -- These graphs clearly shows that the ratio of loan amount to the persons income have big impact on probability to default. And medium range income and higher interest rate have higher chance of defaults.



Inference -- Lower income loan ratio and high interest rate is the combination to avoid as it has the highest default probability - (21-25, 0-1)



Inference -- This graph shows is the loan is taken for debt consolidation and if the interest rate is high (21-25), the probability of default is high.



Inference -- These graphs does not show any specific patter but 1 year and <1 emplyee length have higher chance of defaults.

Conclusion

Following factors are highly impacting on probability of the defaulters and should be considered while processing the loan application:

- Higher interest rate
- Medium income range.
- Lower income-to-loan ratio and high interest rate is the combination to avoid as it has the highest default probability - (21-25, 0-1)
- If loan purpose is debt consolidation and if the interest rate is high (21-25).
- Employee’s experience - 1 year and <1 emplyee length have higher chance of defaults.