Lending Club Case Study

Compiled by Achin K Das

Agenda

- Problem Statement
- Assumptions
- Approach
- Results
- Recommendations
- Conclusions

Problem Statement

Lending Club (LC) is a NBFC which is in the business of providing loans to consumers and gain from the interest earned.

- It competes with heavyweights like Bajaj Finserv, PaySense, Muthoot Finance etc.
- It needs to attract customers with low interest rates while ensuring minimal defaults
- It loses business if it rejects loan applications due to conservative credit assessments
- It makes losses if customer defaults
- LC is looking to understand the **driving factors** behind loan defaults so that they can effectively use those factors to determine risky customers

The **objective** of this exercise is to analyse the available data regarding past loans and identify those factors or variables which are strong indicators of default

Assumptions

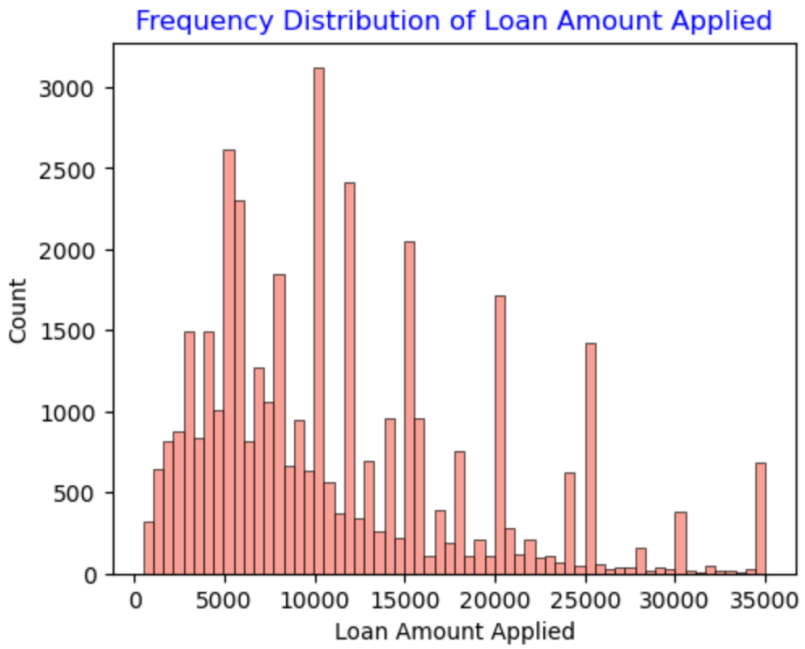
The assumptions for this exercise are as follows

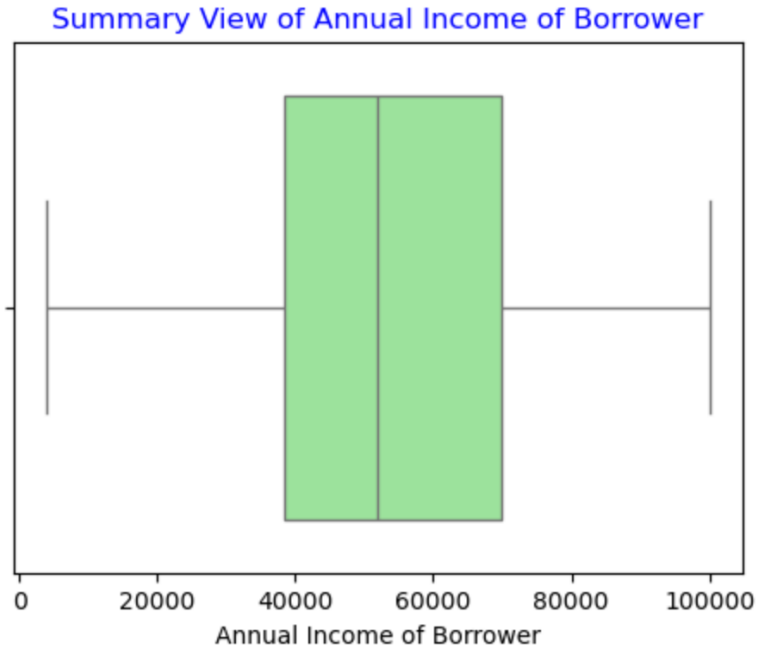
- Exploratory Data Analysis (EDA) technique is used for analysis of available data
- The analysis is carried out on the past loans data provided by LC
- The data is adequately anonymised to prevent revelation of personal sensitive information
- Loans with ”Current” status are ignored from the analysis process as those have neither defaulted nor paid up
- Outliers are not removed from the dataset to prevent loss of important data unless those may skew the results of analysis
- Python Notebook with Pandas, Matplotlib, Seaborn and other relevant libraries are used to prepare, analyse and visualise data

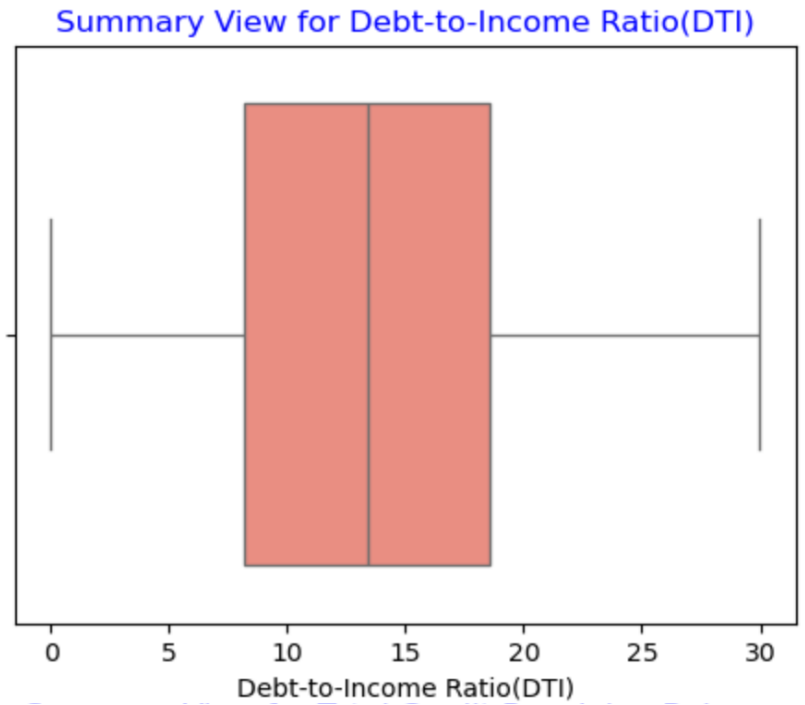
Approach for Exploratory Data Analysis

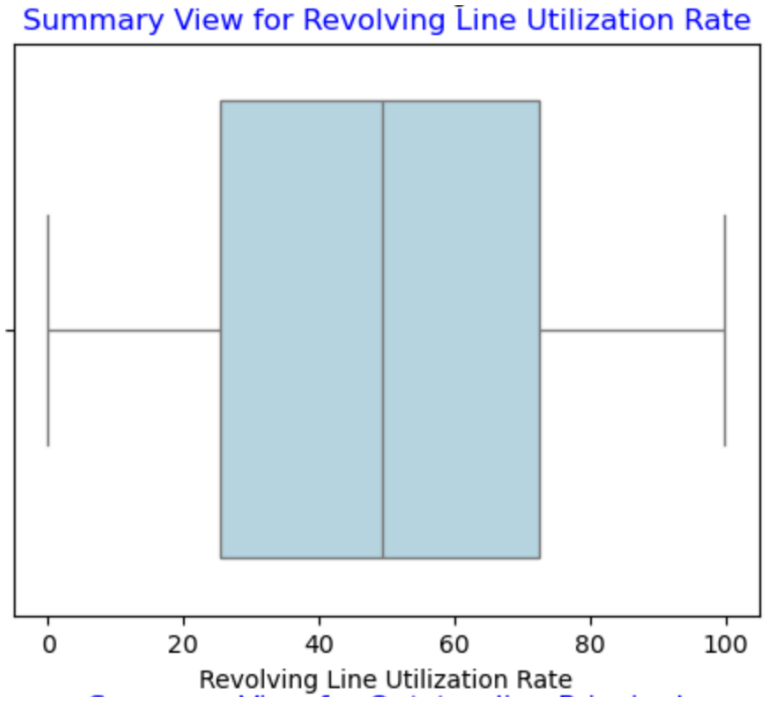
A six-step approach is used for to analyse the data for this exercise:

- Data in .csv-form is **imported** into Python Notebook
- Columns with missing (null) data are either **imputed or removed**
- Some of the columns (variables) are **converted** from object-type to float-type
- Variables are divided into buckets of **numerical** and **categorical** data types
- **Univariate** analysis of numerical (**histplot/ boxplot**) and categorical (**countplot**) variable is done
- **Bivariate** analysis of numerical (**boxplot**) and categorical (**barplot**) variables is done against target variable i.e. Loan Status
- **Multi-variate** analysis is done on numerical variables using **Heatmap**
- Finally, conclusions are drawn, and report is made



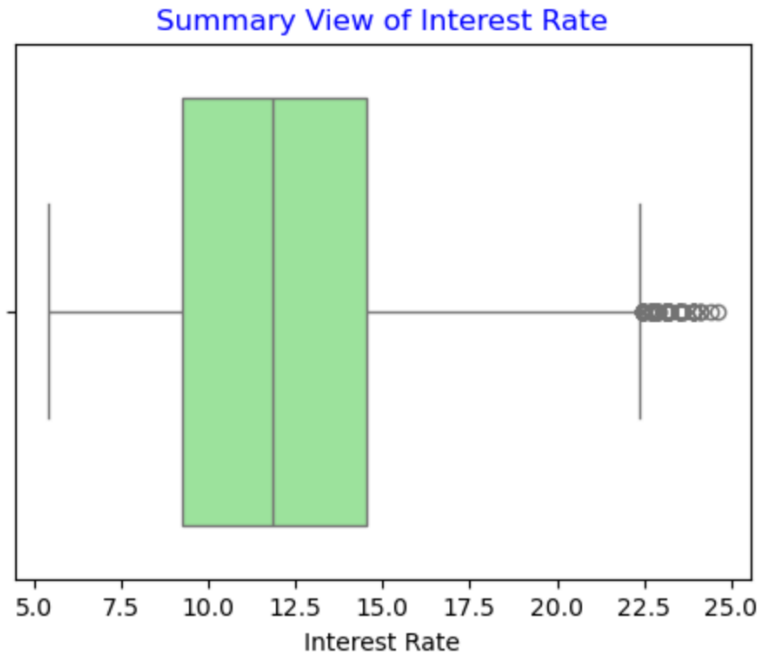


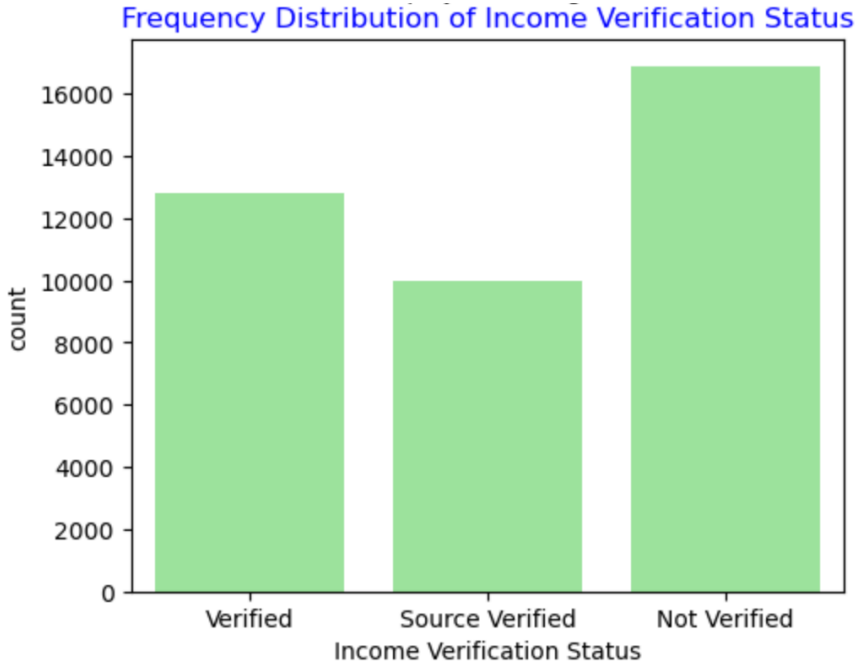


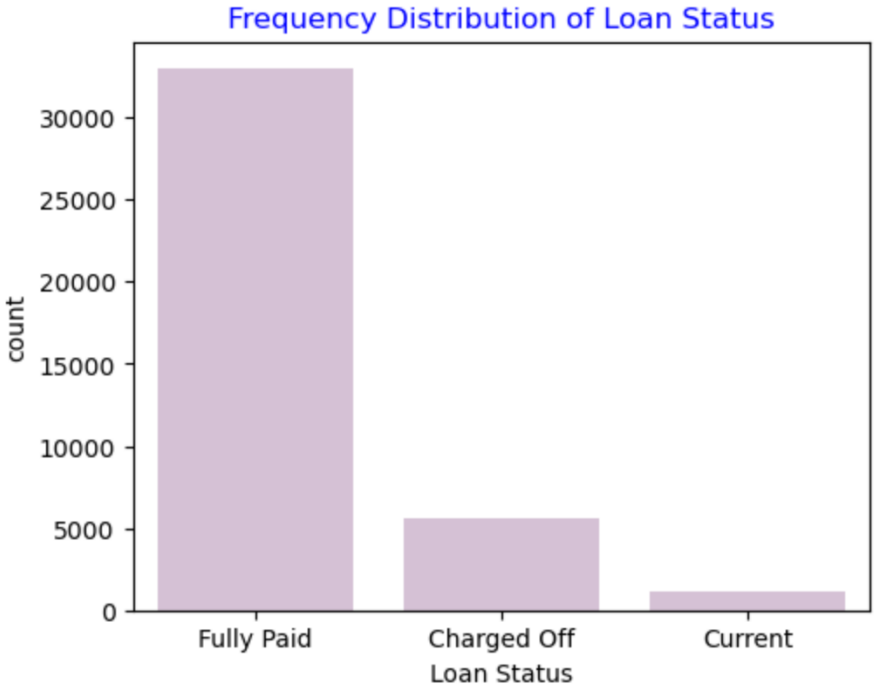


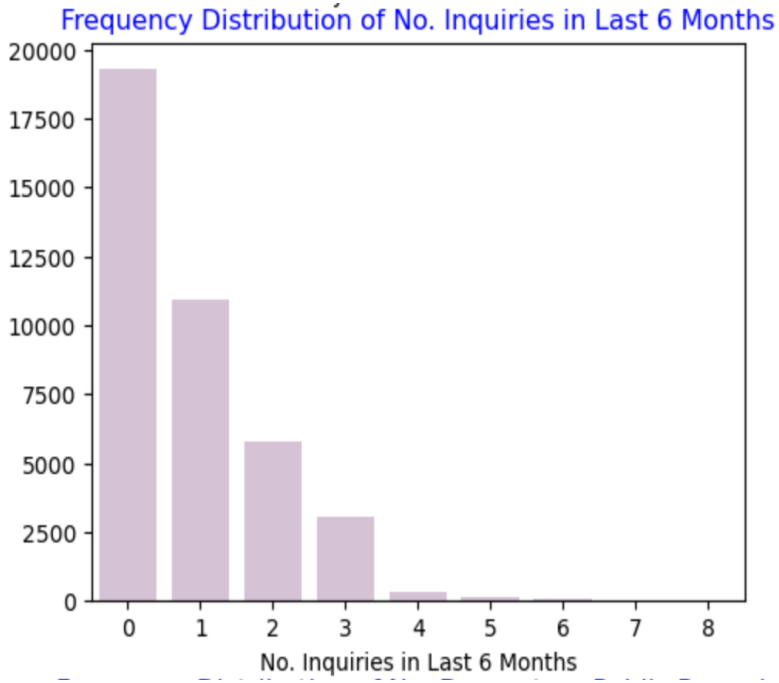
Univariate Analysis Observations – Numerical Variables

- Loans typically sought in multiples of $2500
- 75-percentile loans are for people earning ≦ 82.4K
- 3/4th borrowers maintain healthy debt-to-income ratio of ≦ 19%
- At 48.8%, revolving credit utilisation rate is high among borrowers



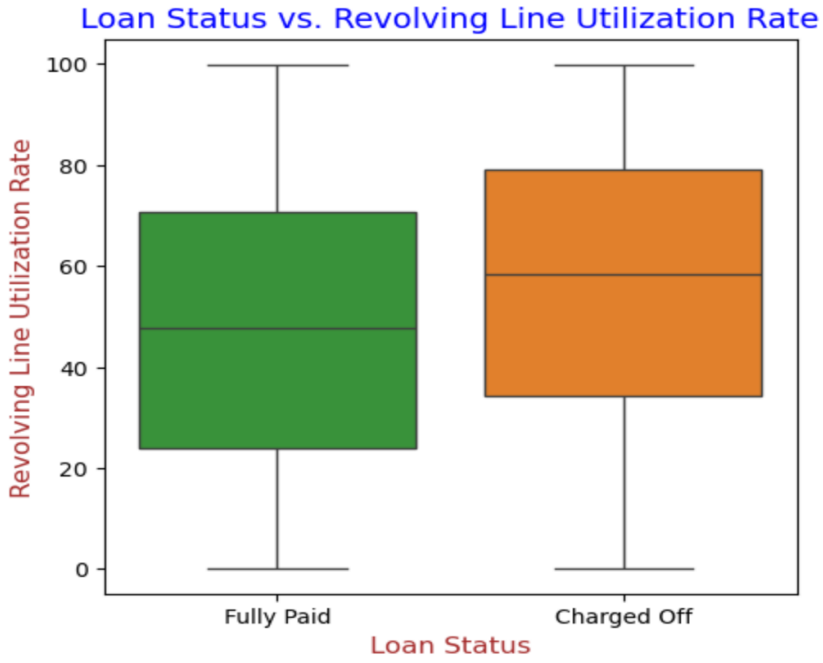


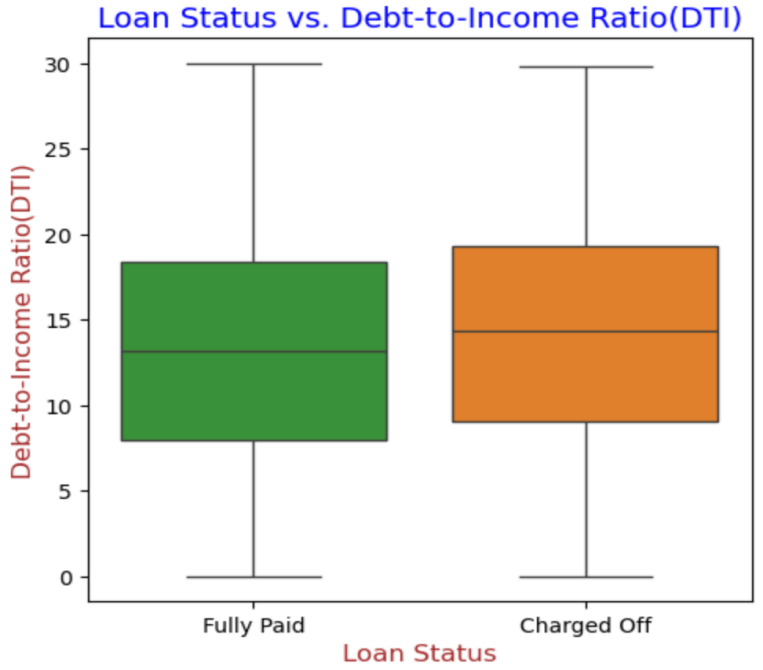


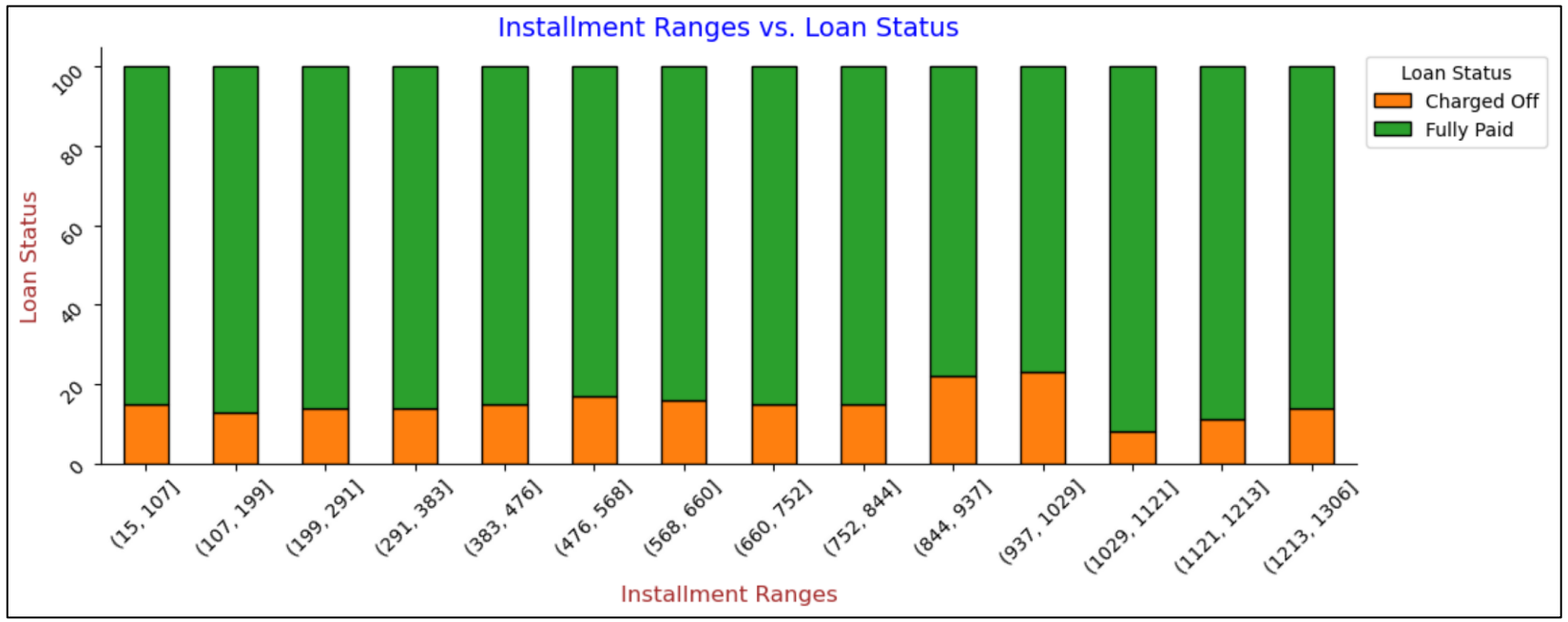


Univariate Analysis Observations – Categorical Variables

- 1/4th borrowers take loan at high interest ≧14.5%
- Only a third (32%) of loans get income verified
- 14% of the loans get charged off by providers
- About 24% borrowers get 2 or more enquiries in the last six months
- 50% borrowers have 9 or more open credit lines

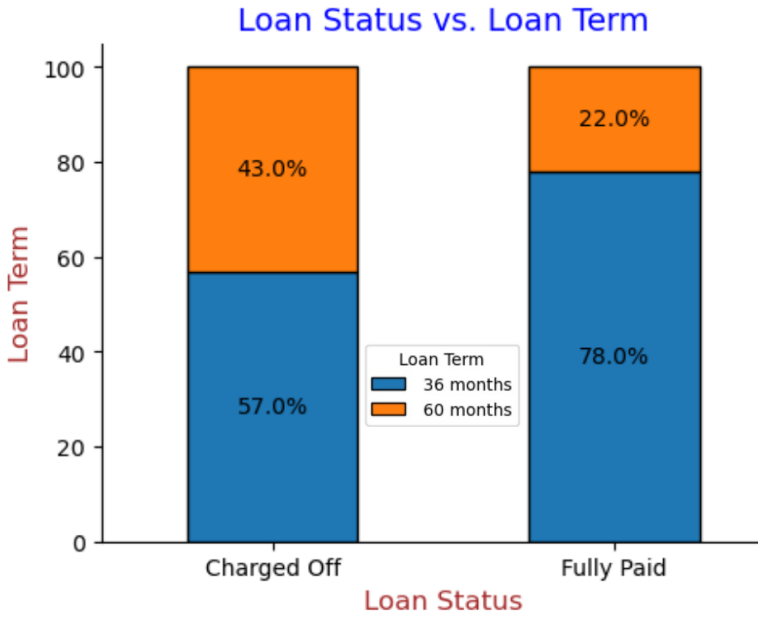


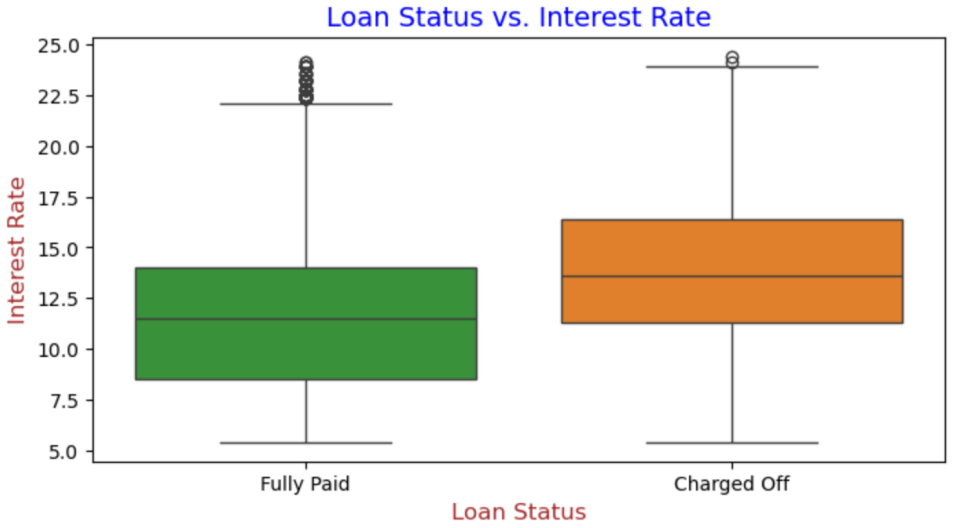


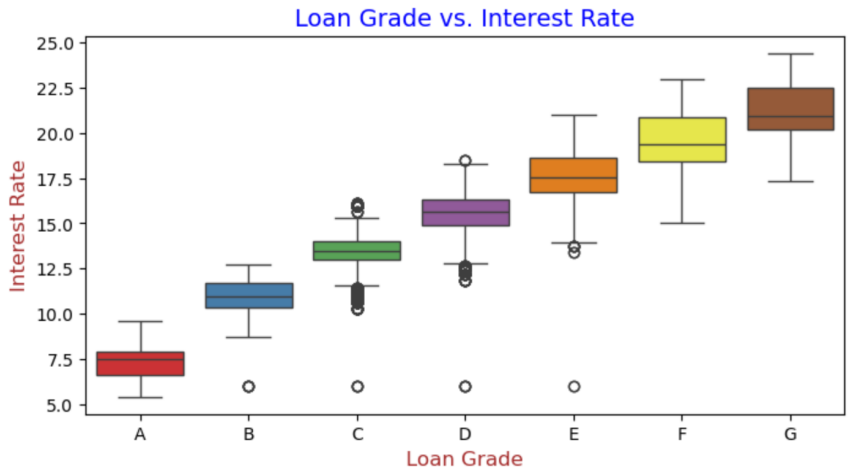


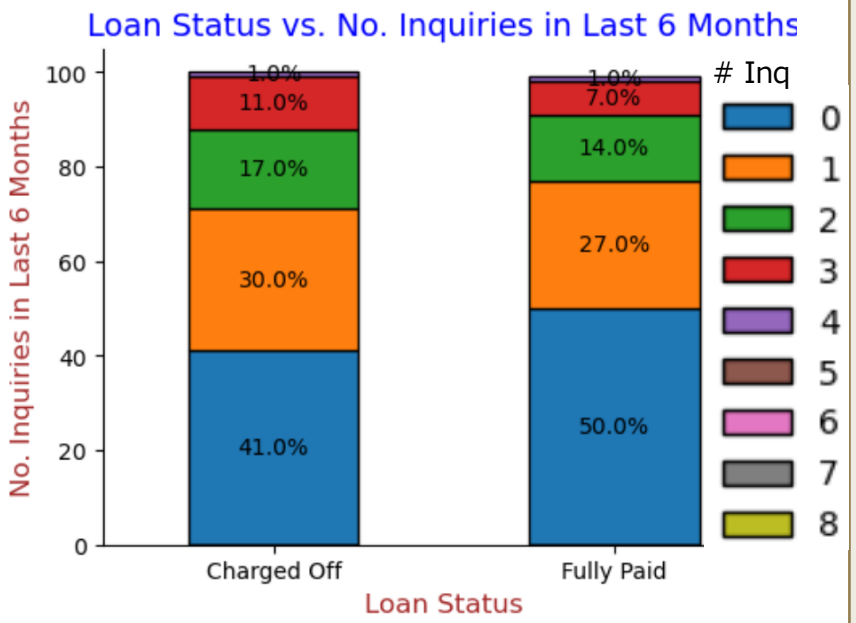
Bivariate Analysis Observations - Numerical Variables w.r.t. Loan Status

- Revolving Line Utilization Rate is 8-10% higher for Charged Off loans
- The Charged Off loans have only slightly higher DTI ratio than Fully Paid loans
- No significant variation in Charged Off percentages with different installment ranges



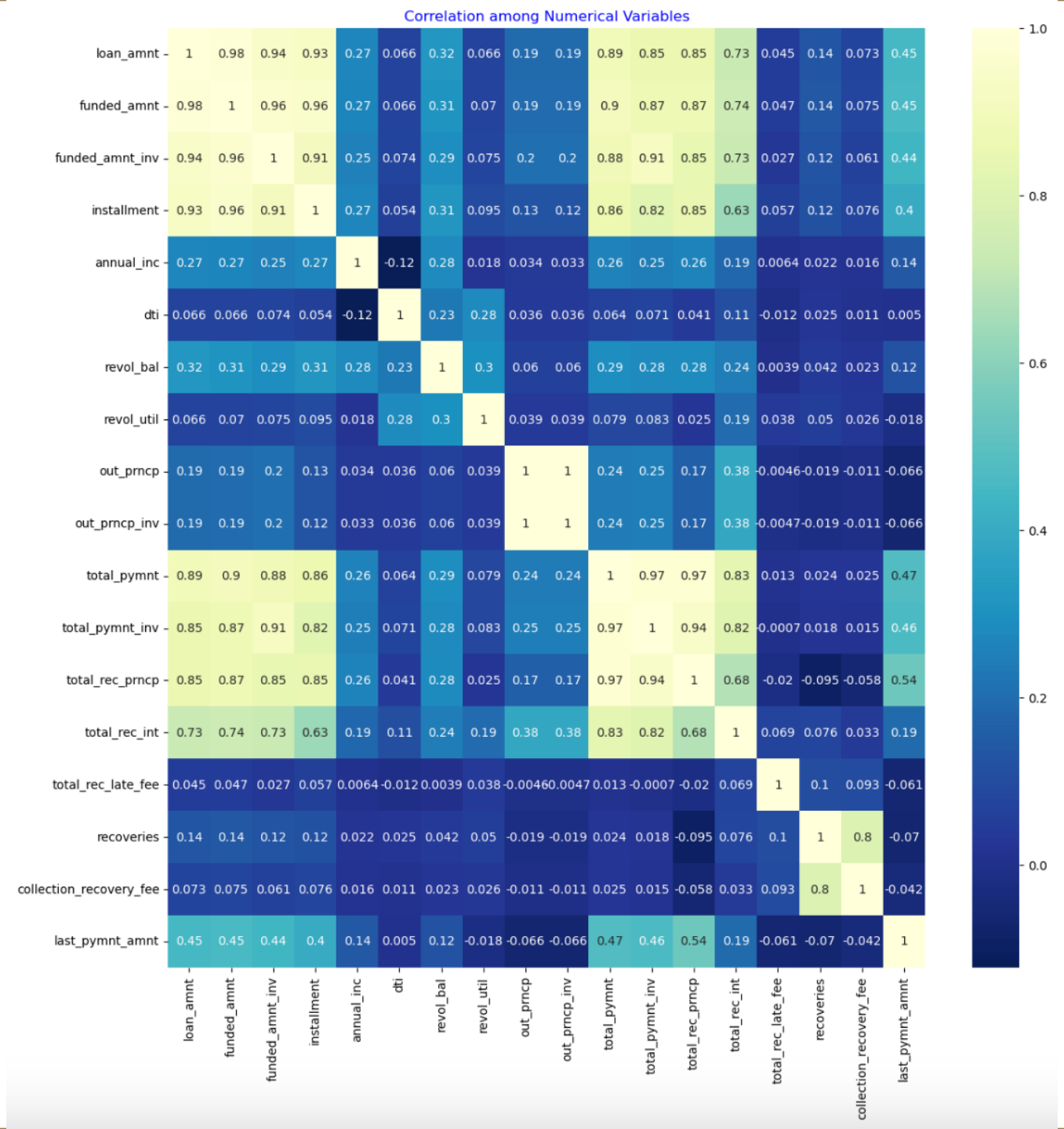






Bivariate Analysis Observations - Categorical Variables w.r.t. Loan Status

- Longer term loans (60 months) have higher(21%+) propensity for default.
- Charged Off loans have interest rate higher by 3-4%
- Lower Loan Grades (C and below) slap higher interest
- Loans taken for small business have two-times more default likelihood
- Charged Off borrowers have 9% more enquiries than Fully Paid



Multivariate Analysis Observations – Numerical Variables

- HeatMap was analysed among the numerical variables
- Recoveries has strong correlation with Post Charge-off Collection Fee
- This is because Collection Fees go up with increased recovery effort

Recommendations

Based on analysis of 53 variables (58 variables dropped due to null values), the following factors are identified to be of interest to Lending Club team.

Driving Factors with **high influence** on loan default:

- Revolving Line Utilization Rate – Avoid customers having more than 80% utilization rate
- No. of Enquiries – Avoid customers with more than 1 enquiries in past 6 months
- Longer Term Loan – Avoid giving 60-months loans as much as possible

Driving Factors with **moderate influence** on loan default:

- Loans for Small Business – Keep such loans to less than 5% of portfolio
- Lower Graded Loans – Such loans should be provided with caution
- Debt-to-Income Ratio– Keep DTI below 20% as much as possible

Additionally, verify income at the time of loan origination to check the authenticity of the income data

Conclusion

In conclusion,

- The exercise produced a set of variables which can help LC to determine potential defaulting customers in advance
- Exploratory Data Analysis technique was adequate to identify relationships within variables
- More than 50% columns had no data. More analysis can be possible if data are made available for those columns
- More advanced data models can be built as next steps that can be used by LC for risk assessment

Thank You