Lending Club Case Study

By

Vinoth Somu

Joseph Sanil

Objective

Main objective is to reduce the Credit Loss as much as possible. Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'. Aim of this case study is to identify the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.

In order to achieve this, we need to utilize various Exploratory Data Analysis (EDA) techniques.

Technique/Approach

To start with the EDA we were provided with the following

- Loan Data Set – This data set contains the complete data for all loans issued through time period 2007 to 2011
- Data Dictionary – Describes the meaning of variables

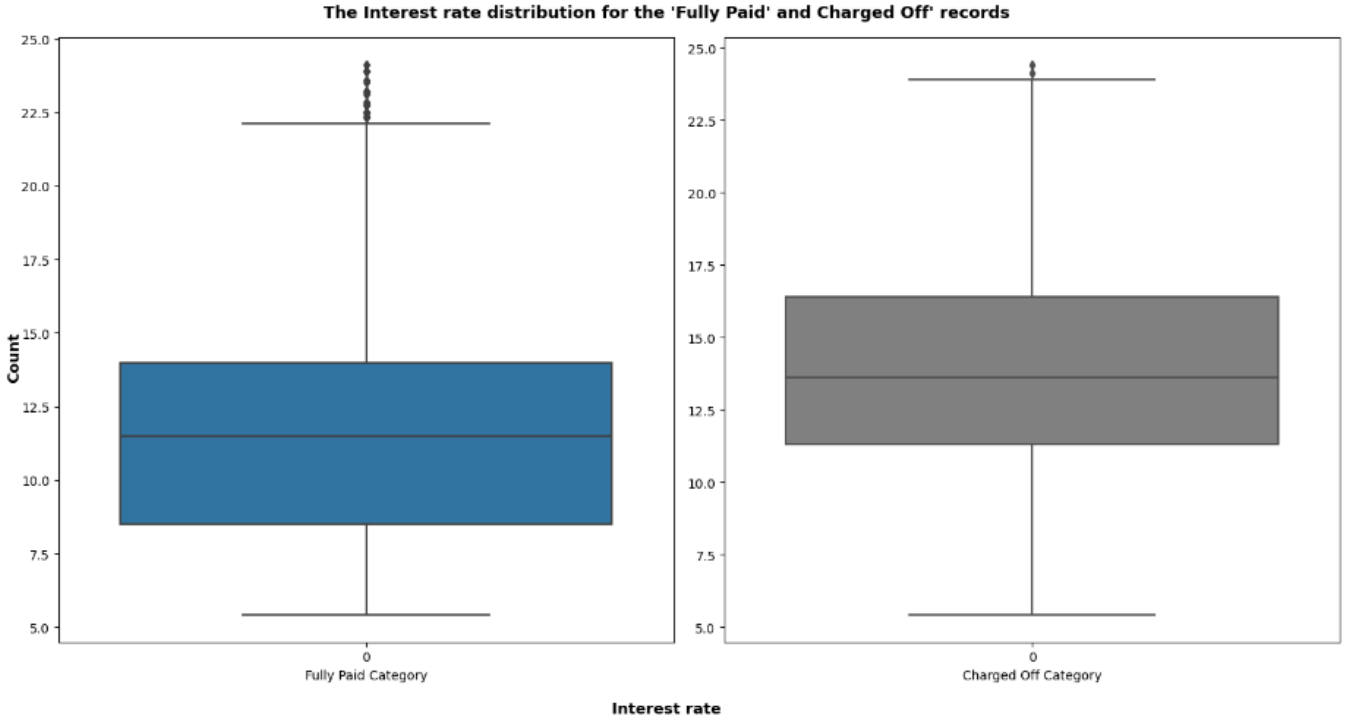
Approach

- After going through the data set its understood that set contains the details about all the loans which were categorized into 'Fully Paid', 'Current' and 'Charged Off'.
- As our main goal is to draft relation between the variables individually (or) in combination to help us identify the factors contributing to 'loan-default'.
- The 'Charged Off' category is where the defaulters fall into. Hence our focus was put mainly on 'Fully Paid' and 'Charged Off' categories.
- We utilized techniques like Univariate, Segmented Univariate and Bivariate analysis.
- Pre-requisite like below were followed before engaging on any kind of analysis
  - Import data set
  - Clean-up
    - Empty rows in the whole columns or above 50% of rows.
    - Same and single value columns.
    - Unique values columns.
    - Unwanted cols which does not add value to analysis
  - Standardization and Data type conversion.
  - Derived columns.
  - Performed basic sanity checks on the dataset after cleanups.

Summary/Conclusion

Based on the analysis, below are considered as the strong or good discriminators. These were identified as part of univariate or bivariate analysis.

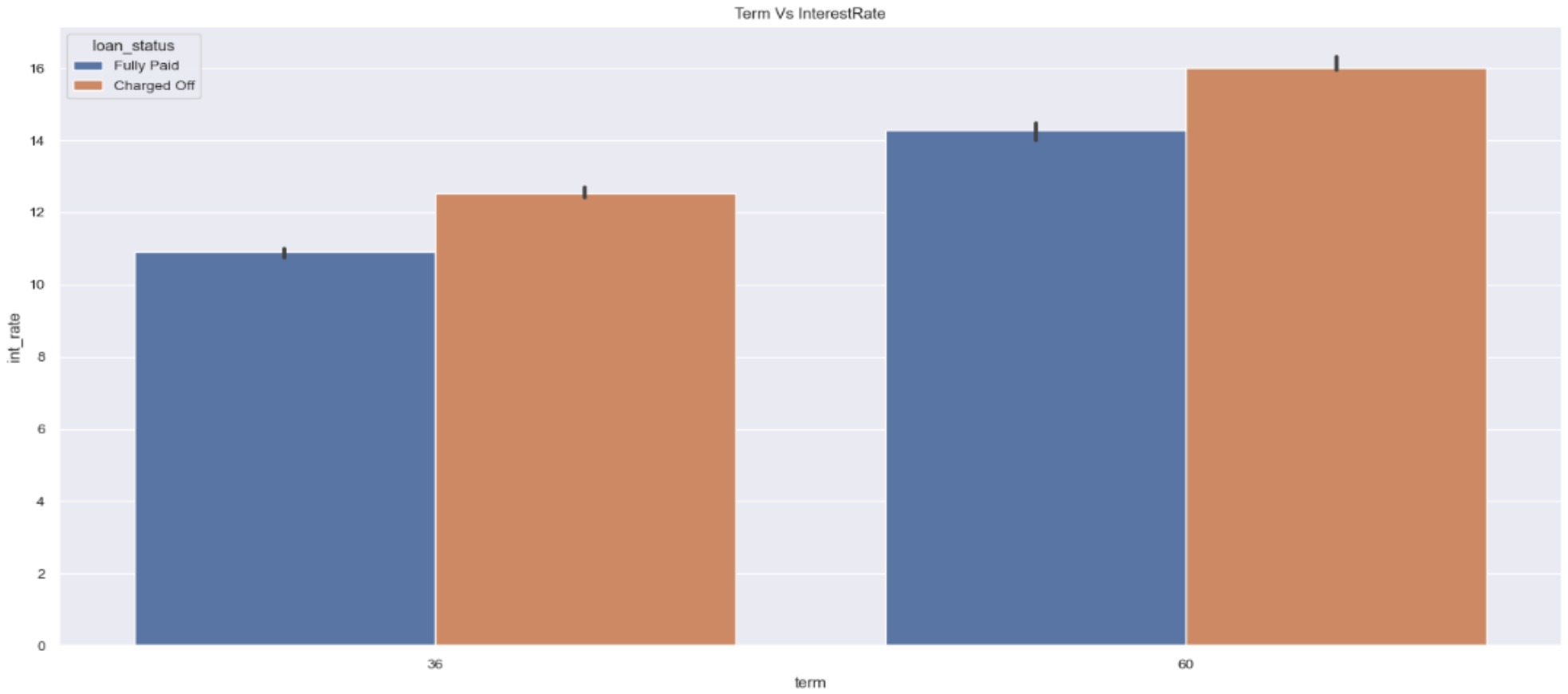
1. Interest Rate (int_rate)



Observations:

1. The average interest rates for the `Fully Paid` records are in the `range of 8 to 14.`
2. The average interest rates for the `Charged Off` records are in the `range of 11 to 17.`

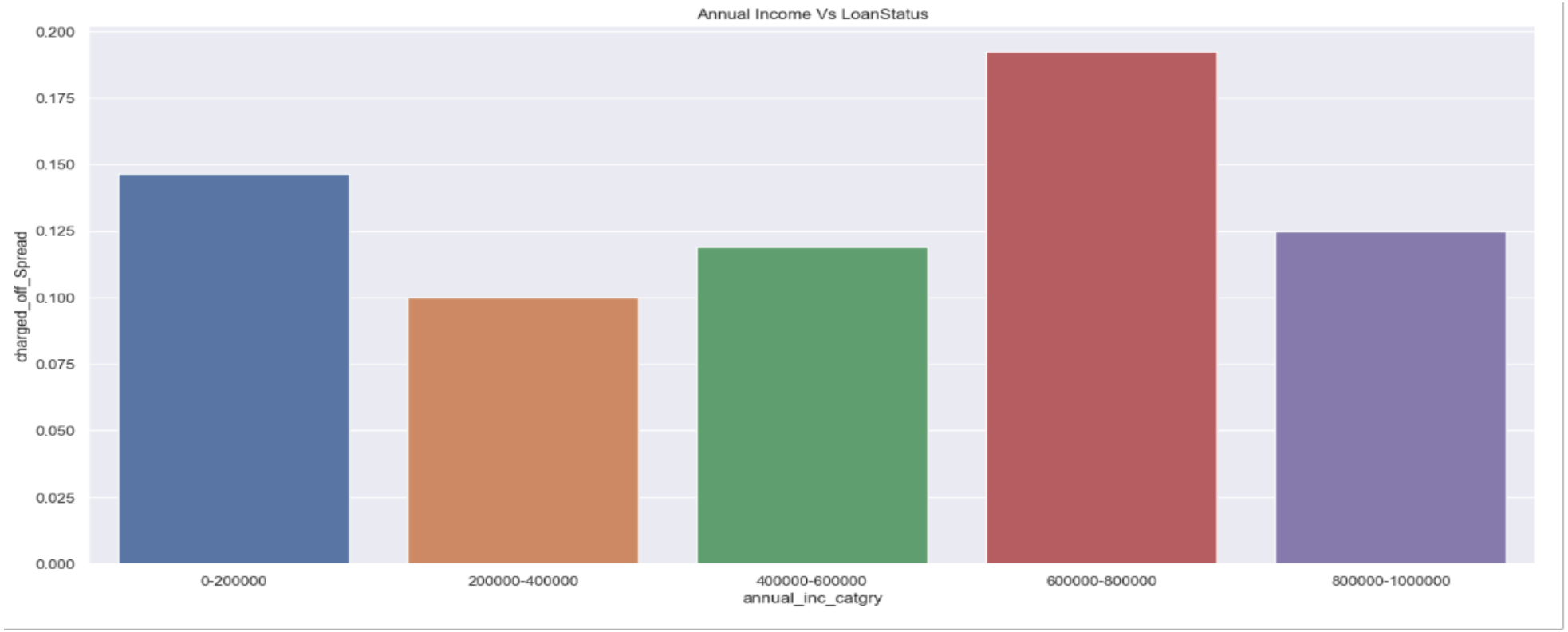
Good Discriminators....contd



Observation:

1. With higher interest rate, its observed the 'Charged Off' loans are also higher for both 36 and 60 month term.

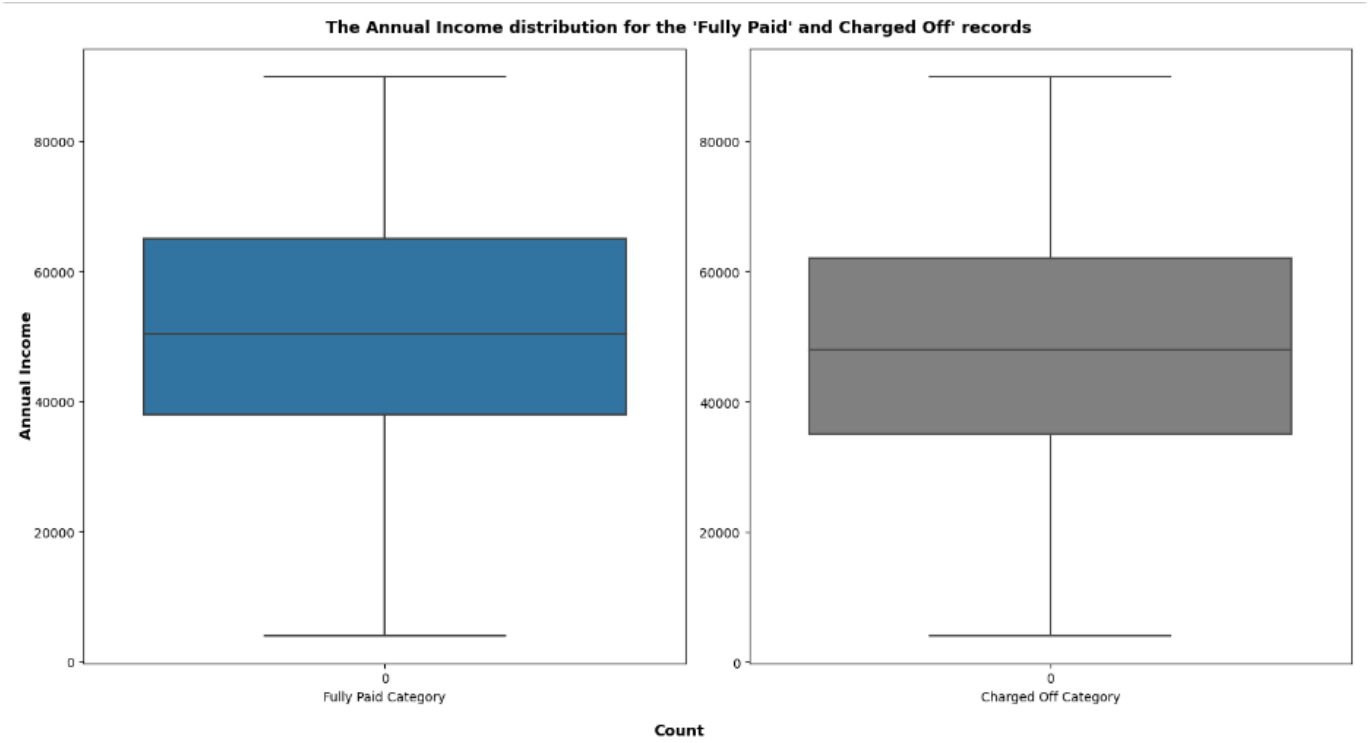
2. Annual Income (annual_inc)



Observations:

1. High chance of 'default' if 'annual_inc' is in 60K-80K category
2. The annual_inc_catgry 80K-100K has minimal in 'Charged Off'.
3. Lowest annual_inc_catgry of upto 20K has the highest in 'Charged Off'.
4. Inference as annual_inc increases Charged Off loan count decreases.

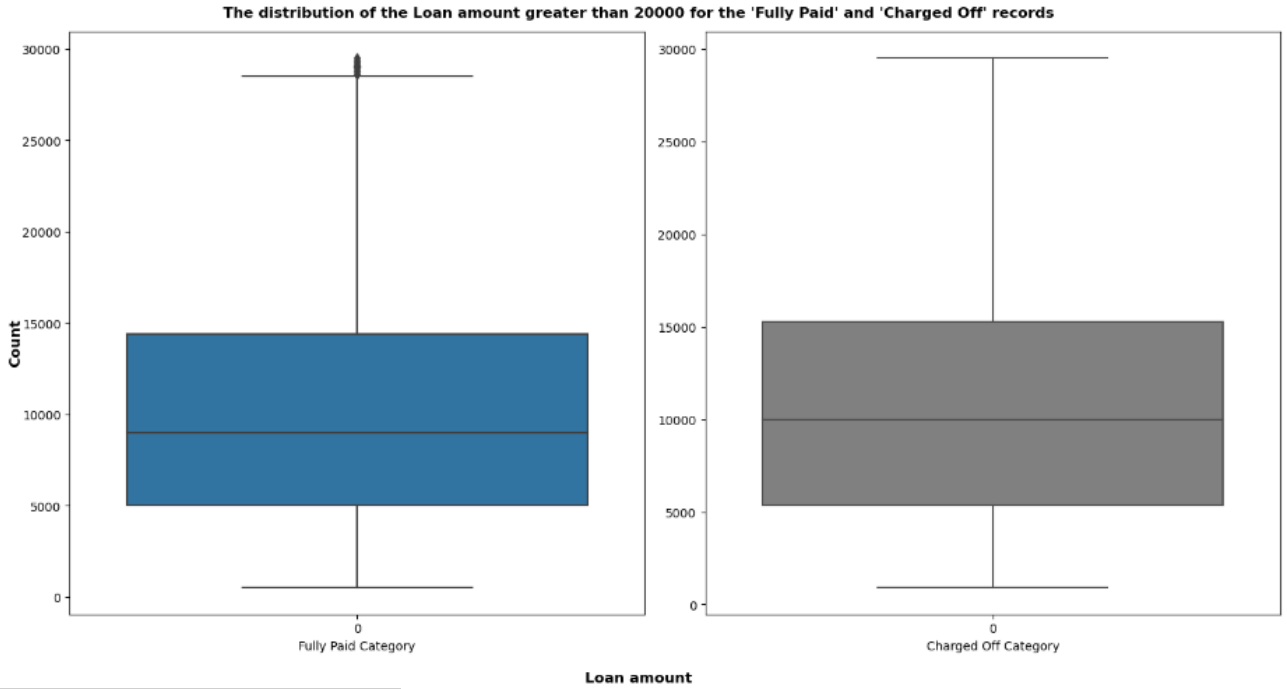
Good Discriminators....contd

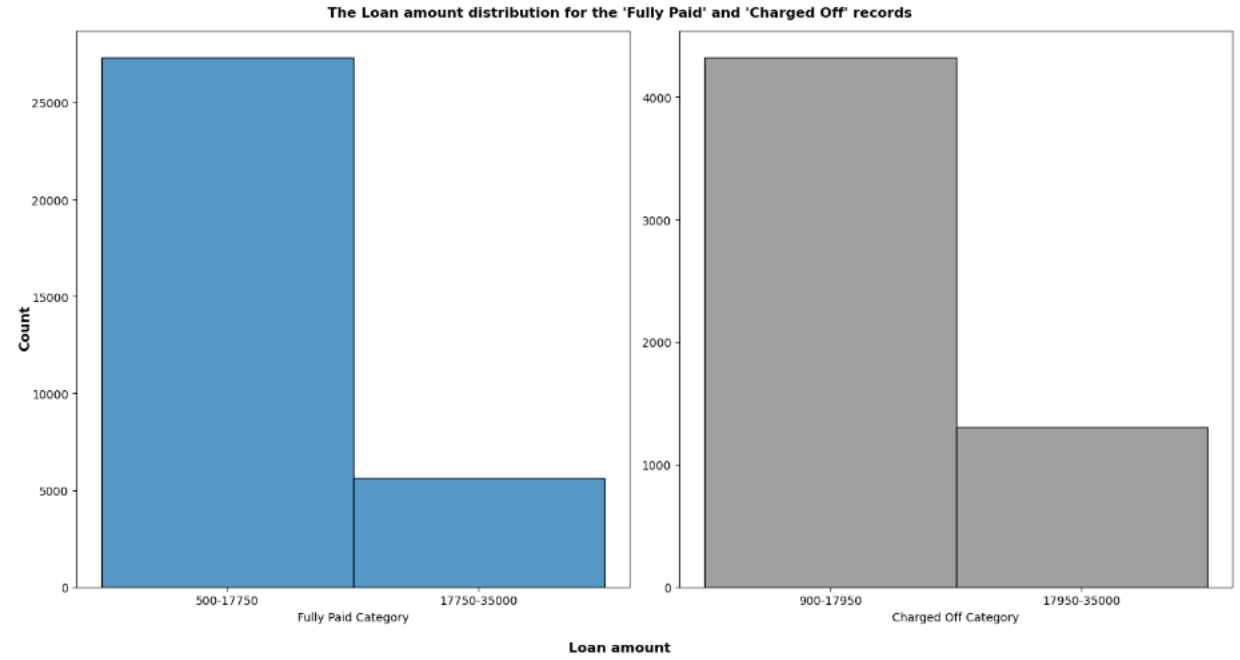


Observations:

1. The average annual income of 'Charged Off' category is lesser than 'Fully Paid' category.
2. `The Annual income consider as a discriminator variable.`

3. Loan Amount (loan_amnt)

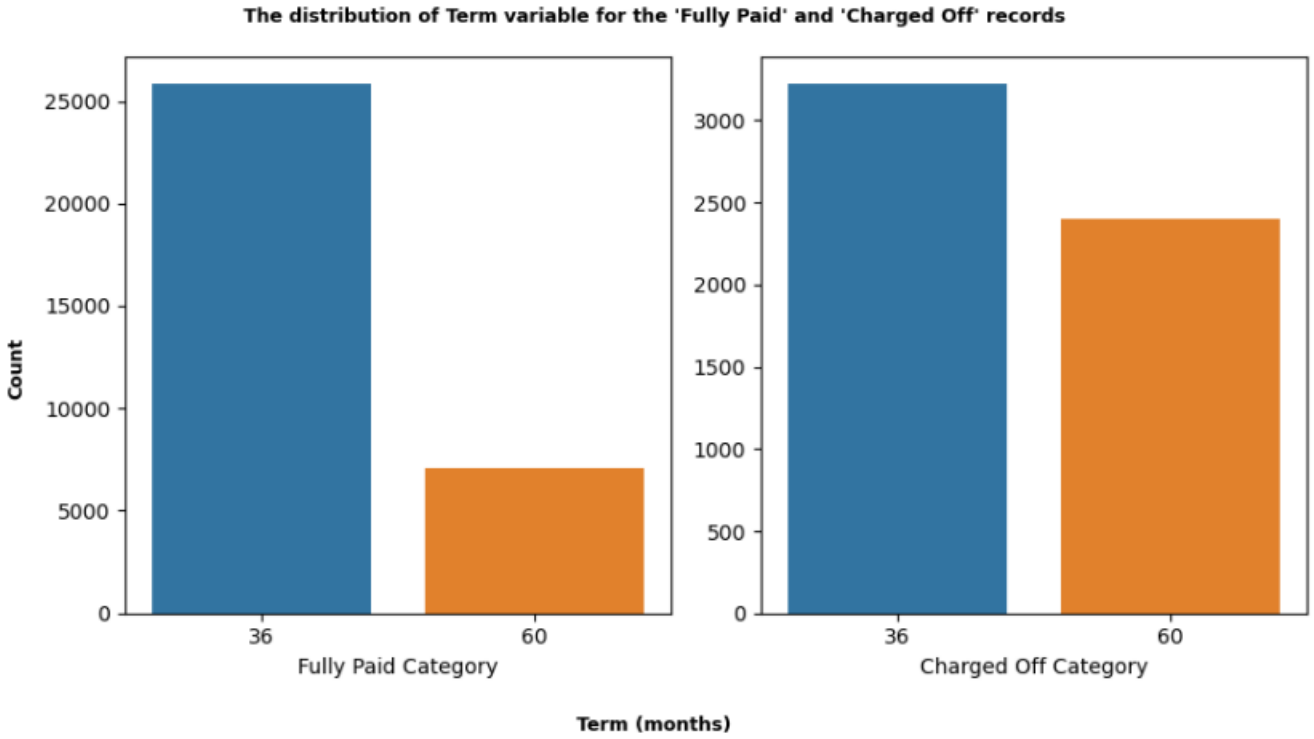




Observations:

1. Majority of cases had loan amount of 10000 for both Fully Paid and Charged Off categories.
2. The median distribution for the loan amount shows slight variation with both categories.
3. The **Loan amount greater than Q3 have variation between Fully Paid and Charged-Off.**
4. `The Loan Amount consider as a discriminator variable.`

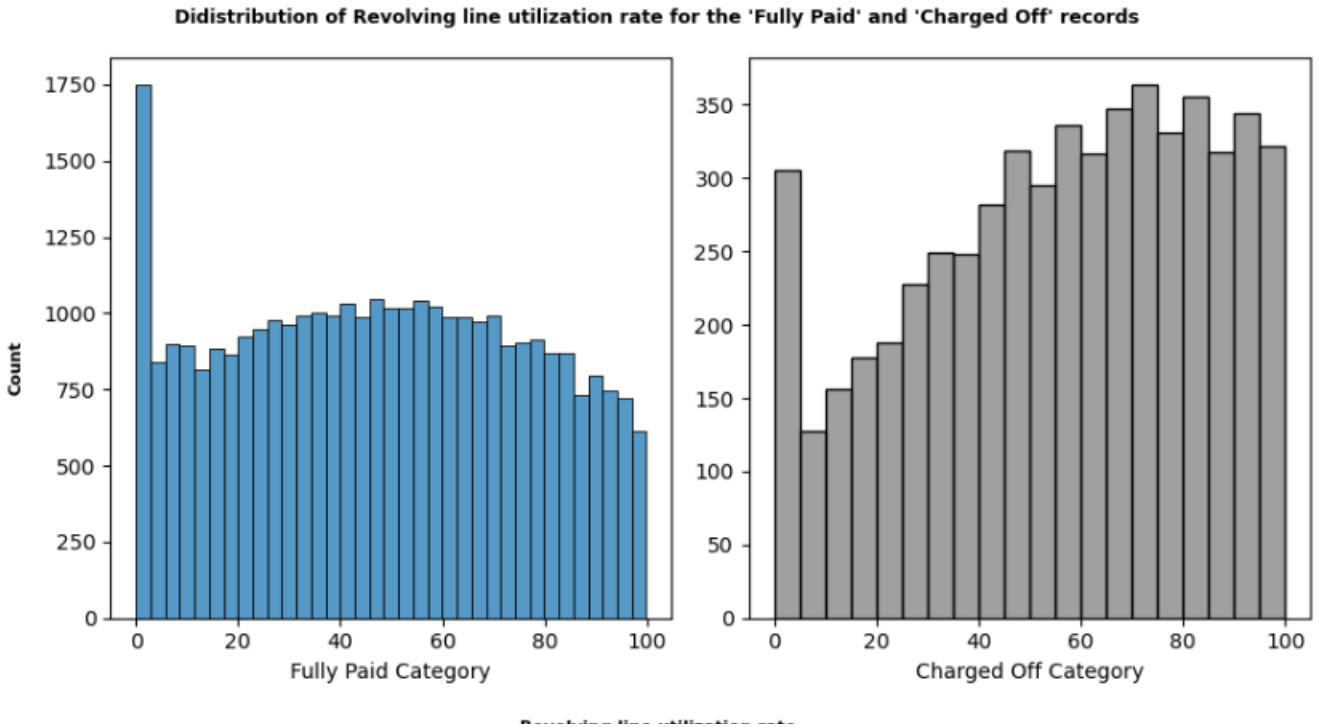
4. Term (term)



Observations:

1. This plot shows the records fall on 60 month is lesser in 'Fully Paid' category but much higher in 'Charged Off' category.
2. `This is a clear indication of the variable 'term' is a great discriminator.`

5. Revolving Line (revol_util)



Observations:

1. This plot shows the Revolving line utilization rate is almost even in 'Fully Paid' but in incremental fashion on 'Charged Off' category.
2. `This is a clear indication of the variable revol_util is a great discriminator.`

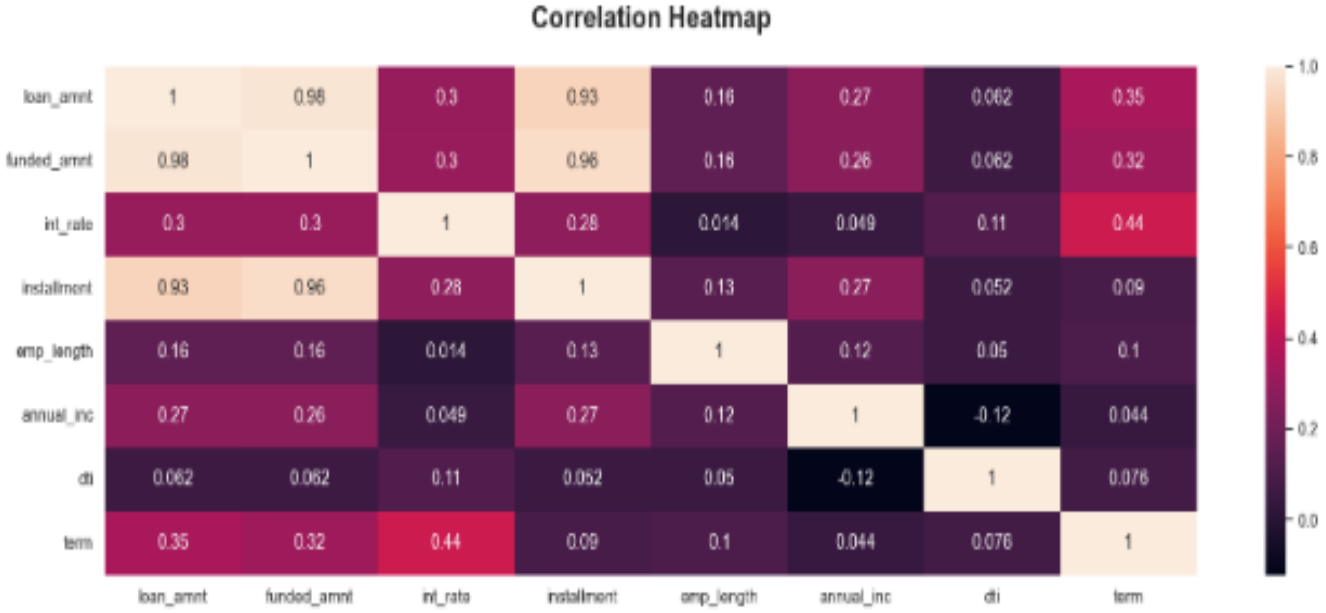
Other Observations

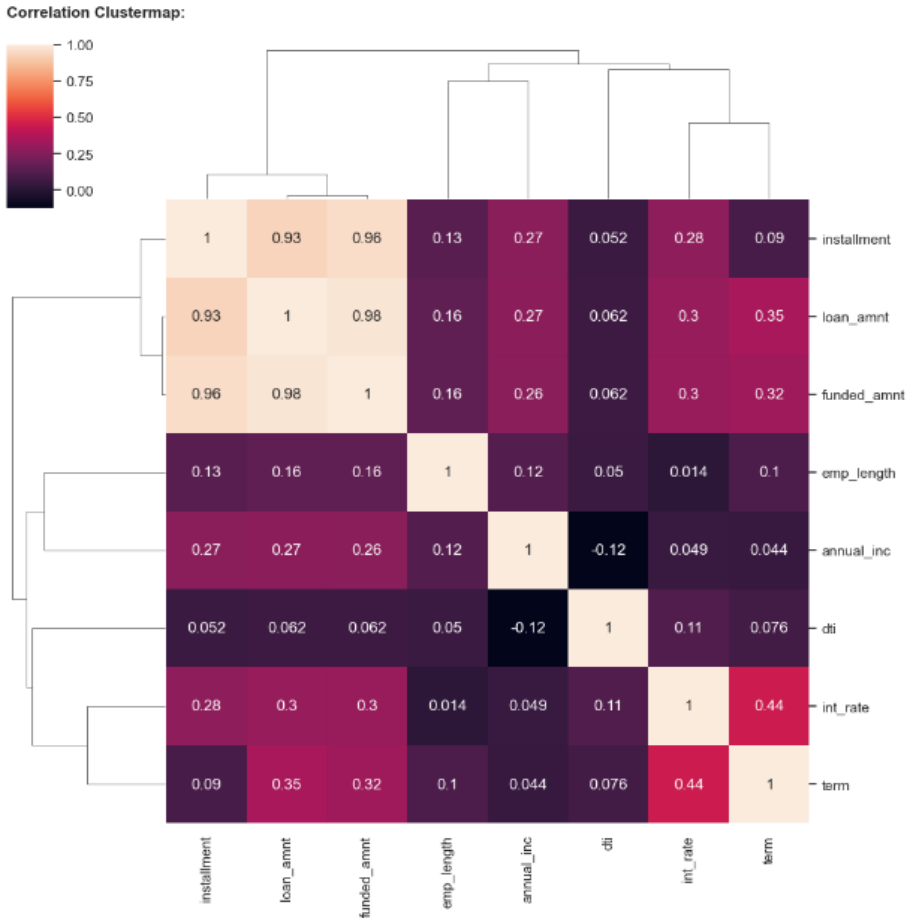
There were also multiple observations which were made, which would also add on to our analysis to identify the chances of 'Charged Off' loan.

All such observations are given below. Most of these were identified as part of bivariate analysis.

1. Loan Amount, Installment, Funded Amount - Strong correlation is visible between all these variables.

2. Employee Length and Annual Income are positively correlated.



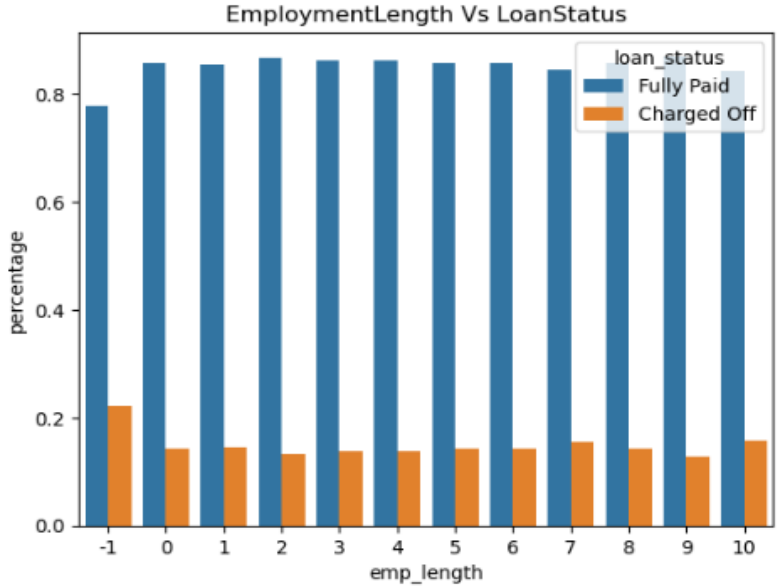


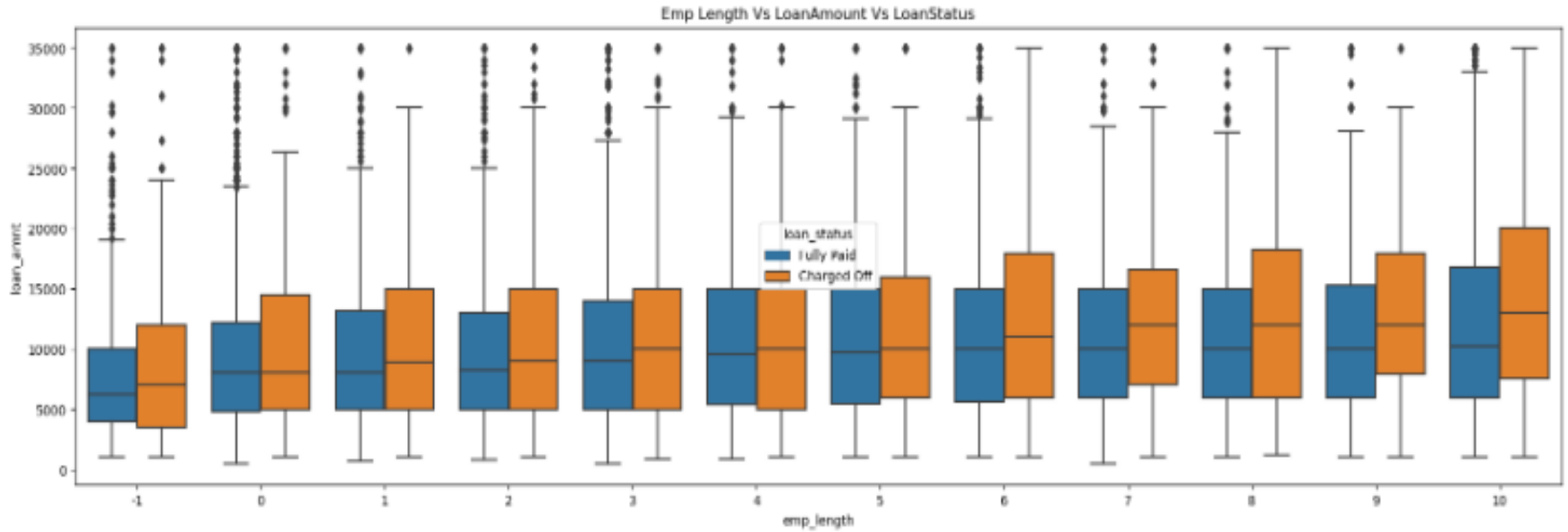
3. Employee Length

– As the employee experience increases,

chances of default also increases.

Highest is with employee in 10+ years.





4. Annual Income is negatively correlated with DTI (dti) which implies that 'dti' is LOW when 'annual_inc' id HIGH.