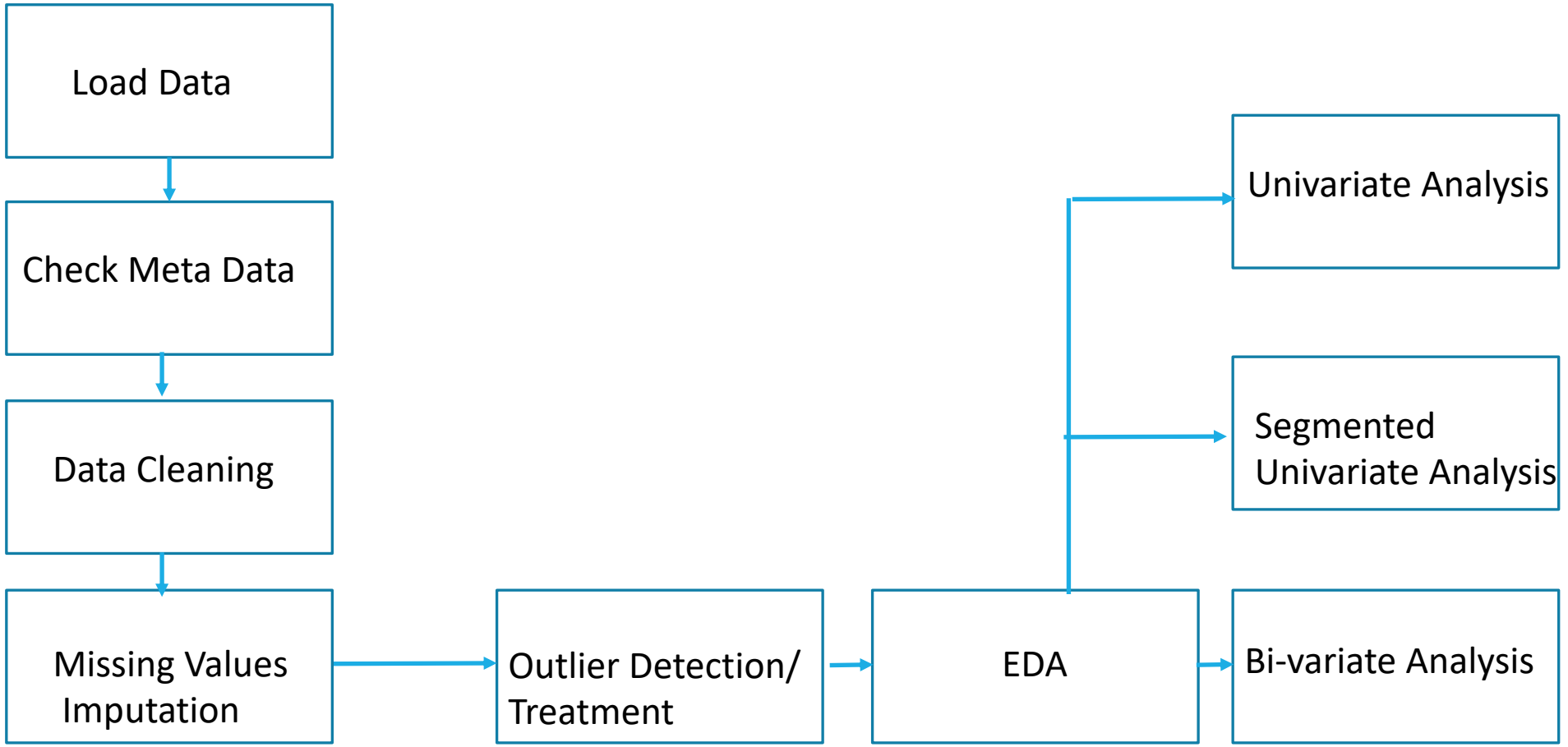
Lending Club Case Study Presentation

GEETA DESAI

Lending Club Case Study –Problem Statement

- Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss).
- In other words, borrowers who **default** cause the largest amount of loss to the lenders.
- We need to figure out the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Analytics Approach

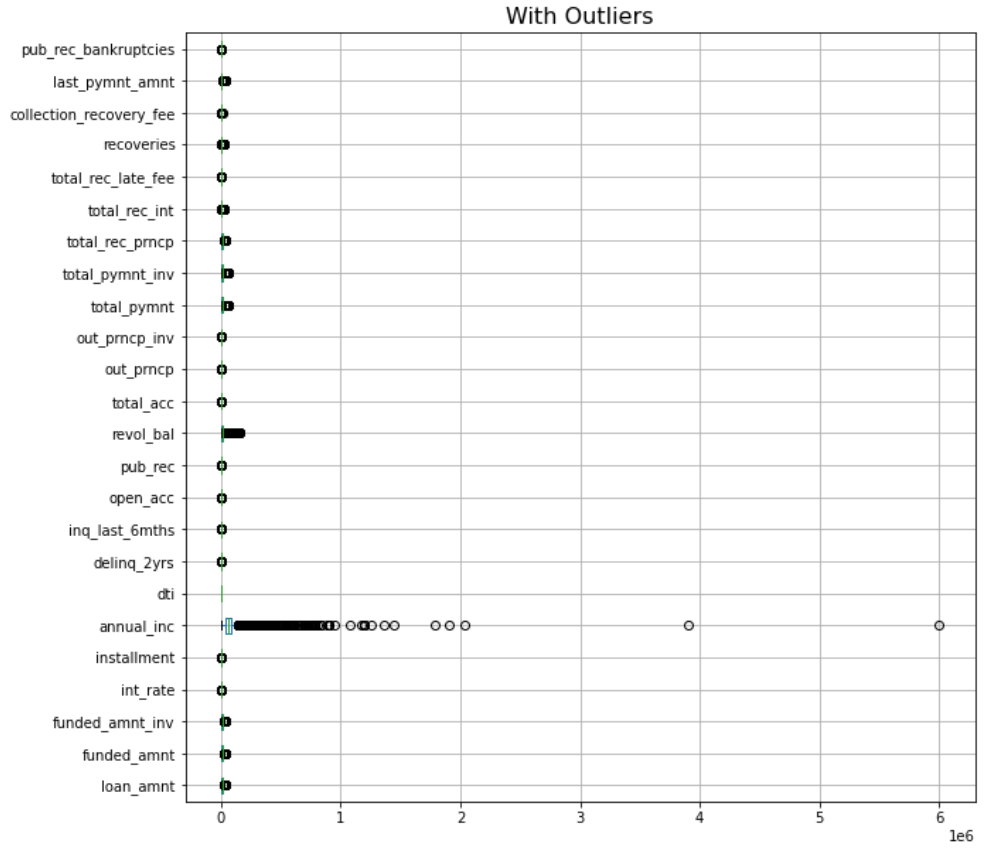


Data Cleaning

- Remove columns with more than 50% null.
- Check and remove the columns which do not have an impact on EDA. The columns that do not have duplicates and are totally unique or all values are same these columns will not have impact on EDA and hence can be removed. This reduces columns from 111 to 36
- Checking for rows with null. Only two rows have null columns and hence can be ignored
- Checking for duplicate rows. There are no duplicate rows.
- Standardize the columns like round off, conversion, etc
- Impute columns where necessary like employee length column

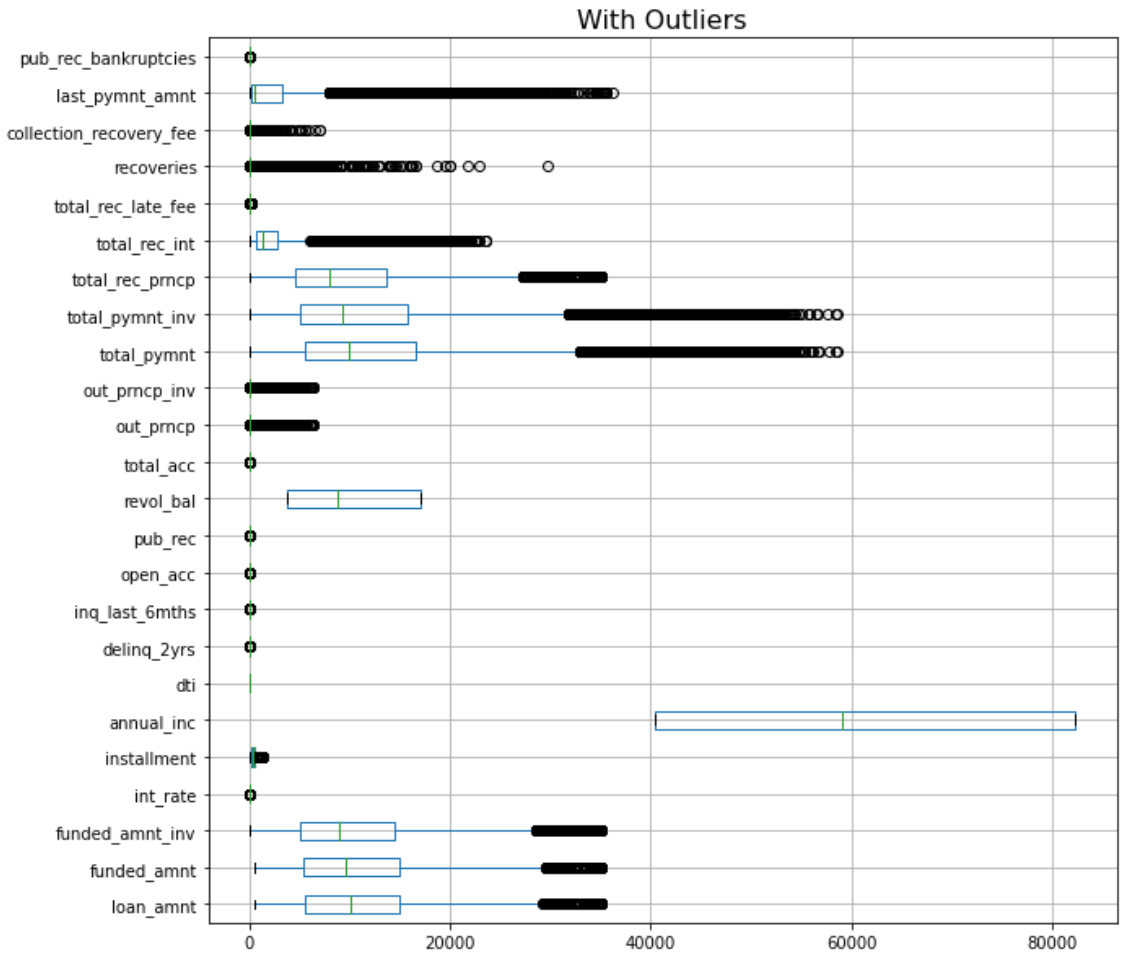
Outliers in Data

Box plot of continuous variables shows there are many columns with outliers

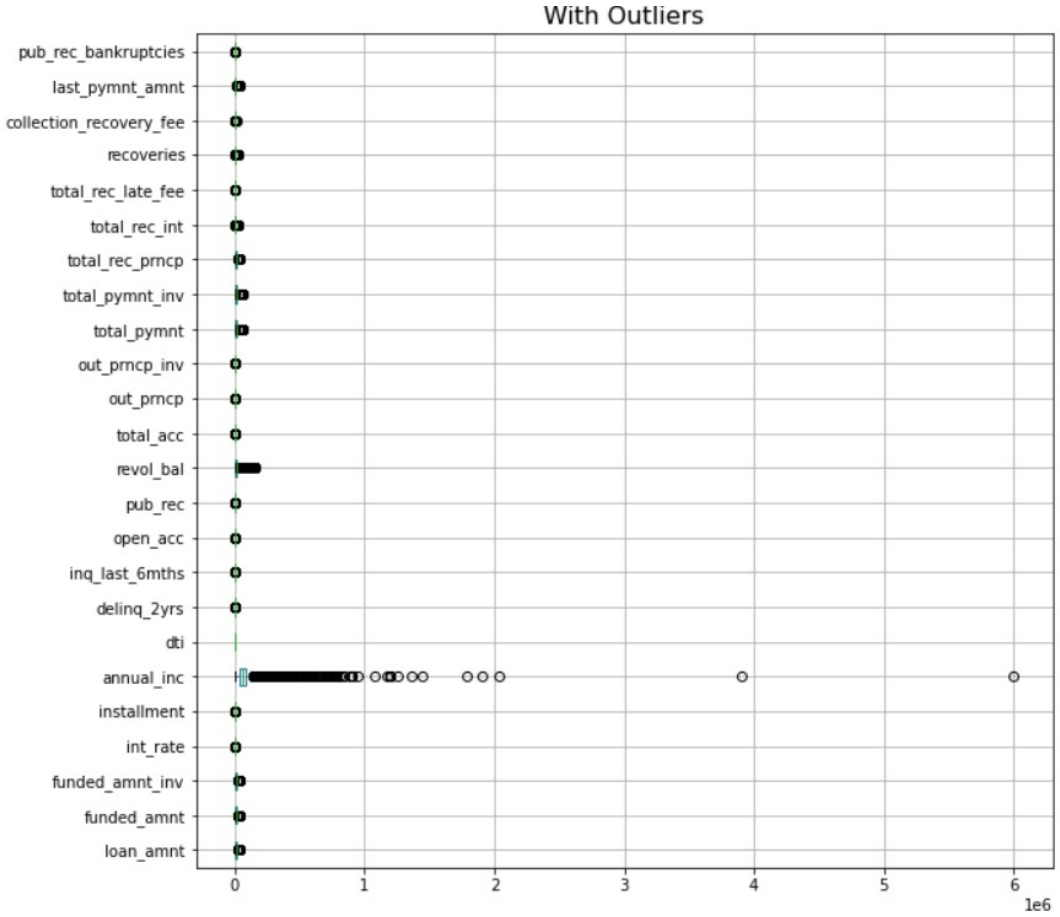


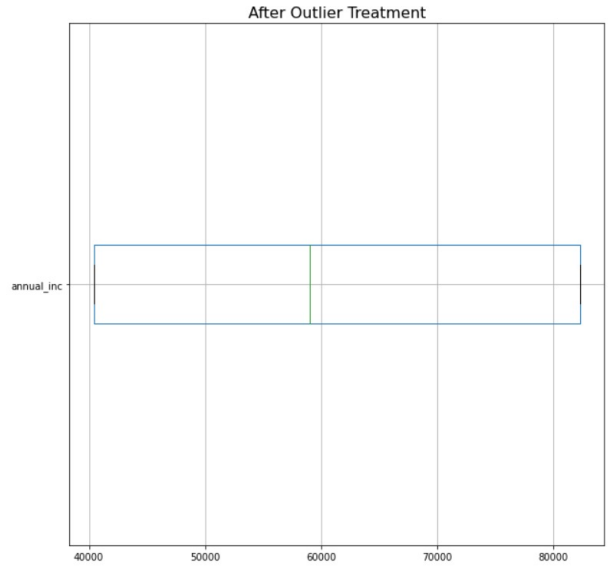
Outlier Handling

- **Check outliers with 0.25 and 0.75 quantile.**
- **Remove outlier values if they are less than 50 %**
- **If outlier values are more than 50%, treat them like imputing with lower range and upper range and box plot again. Now all required columns outliers are treated.**



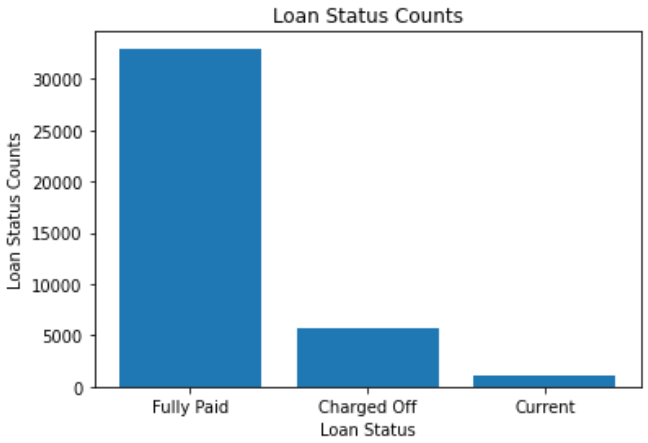
Outlier detection and treatment



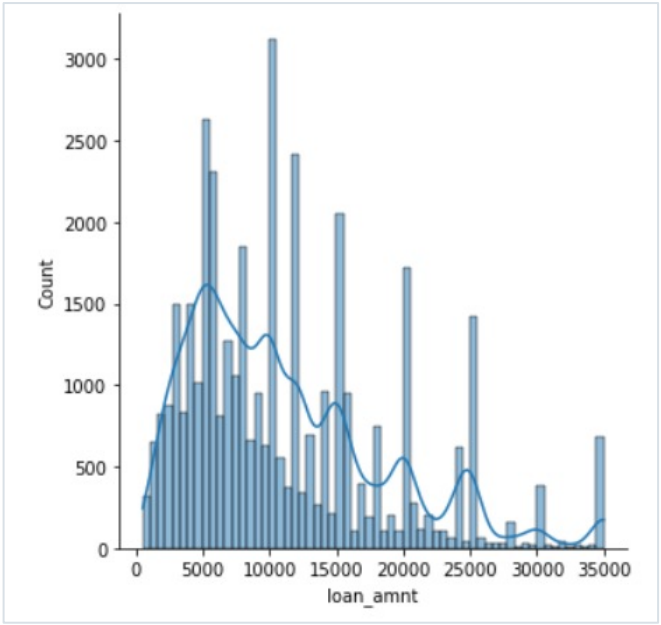


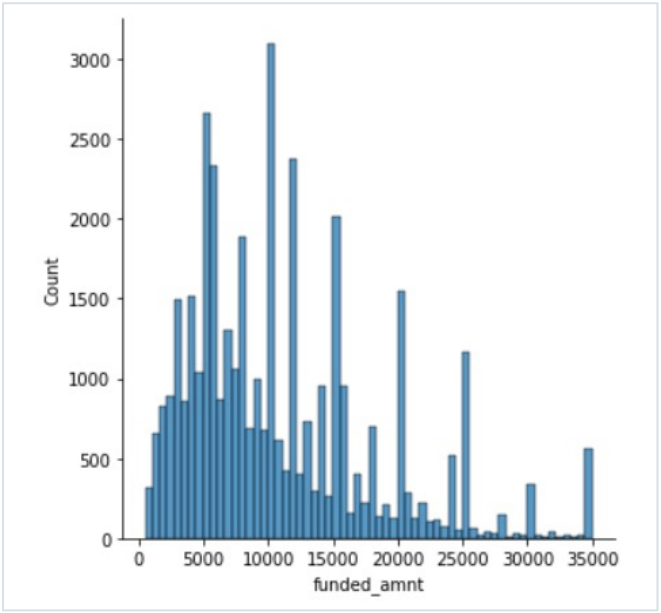
Outliers are detected in multiple numerical variables- For example Annual_inc, these outliers were removed by replaced by restricting the value to the Upper range

Univariate Analysis



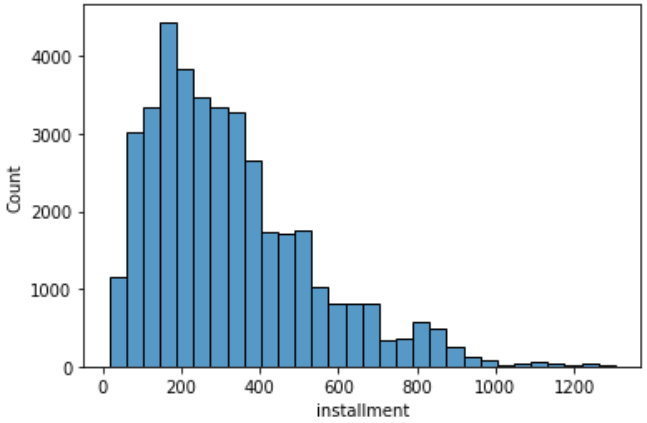
Fully paid customers are approx. 6 times greater than Charged of customers

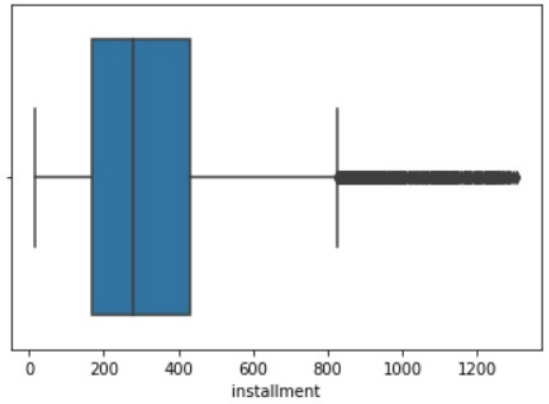




Histogram plots for loan amount and funded amount show that there are peaks at an interval of 5K . The patterns of count of loan amount and fund amount are exactly matching

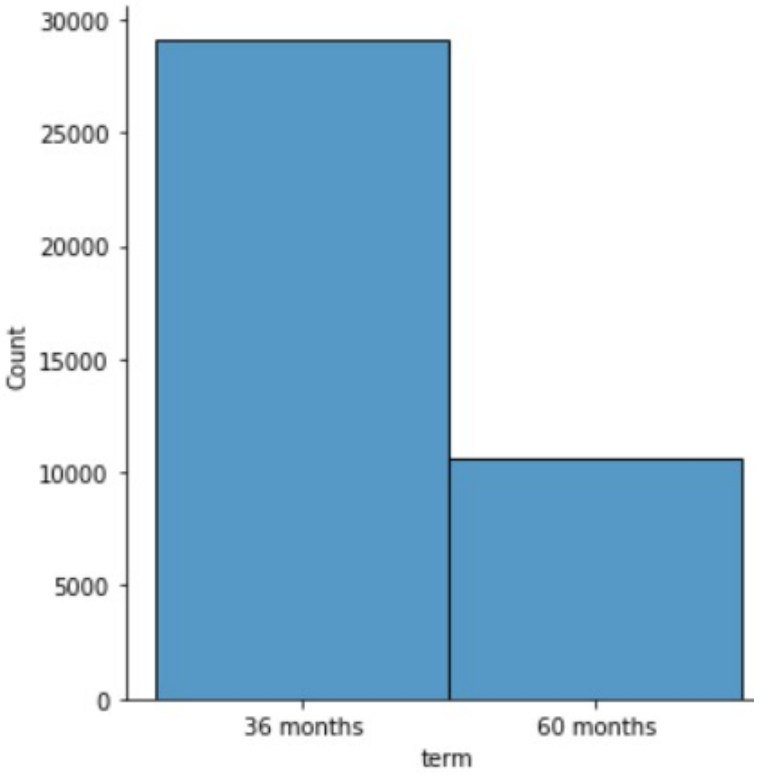
Univariate Analysis



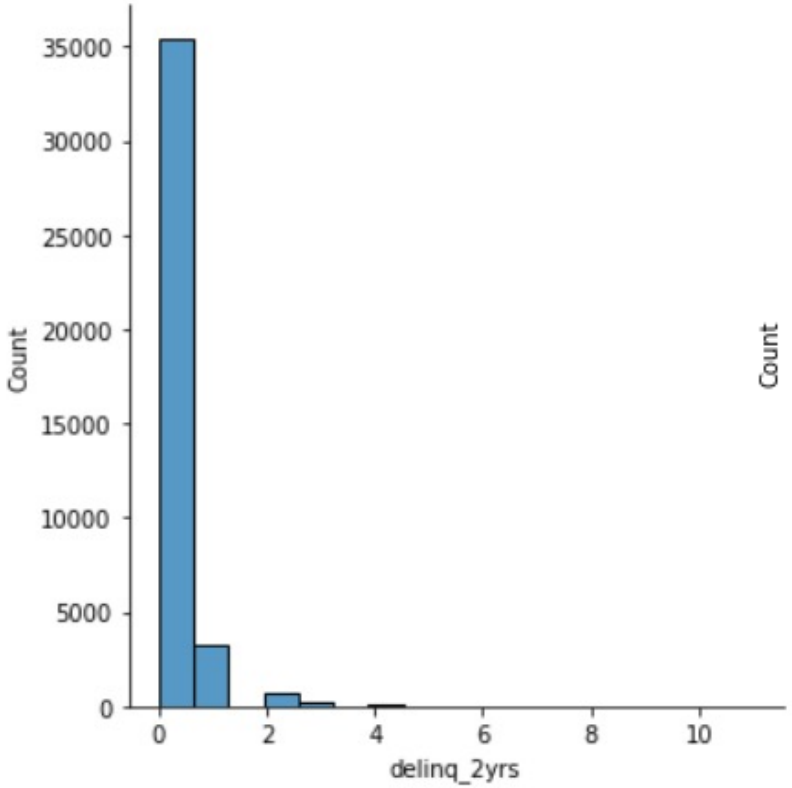


Histogram plot for monthly payment owed by the borrowers (installment) is right-skewed with the most common installment amount of Approx 200.
The box plot shows that there are outliers on the higher range of the installment amount

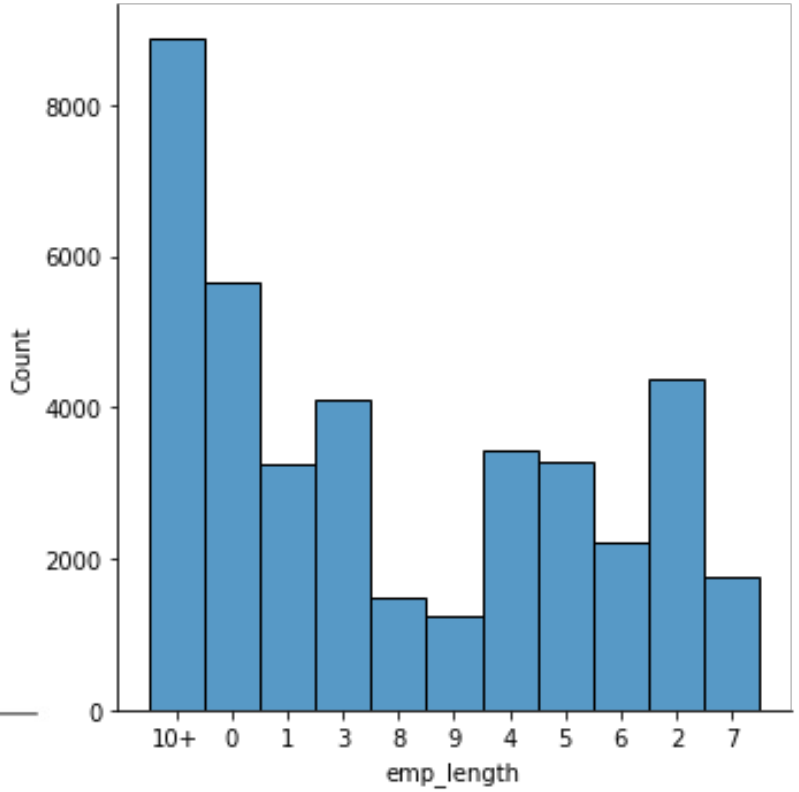
Univariate Analysis



No of customers who have taken a loan for 36 months duration are almost thrice those who have taken a loan for 60 months

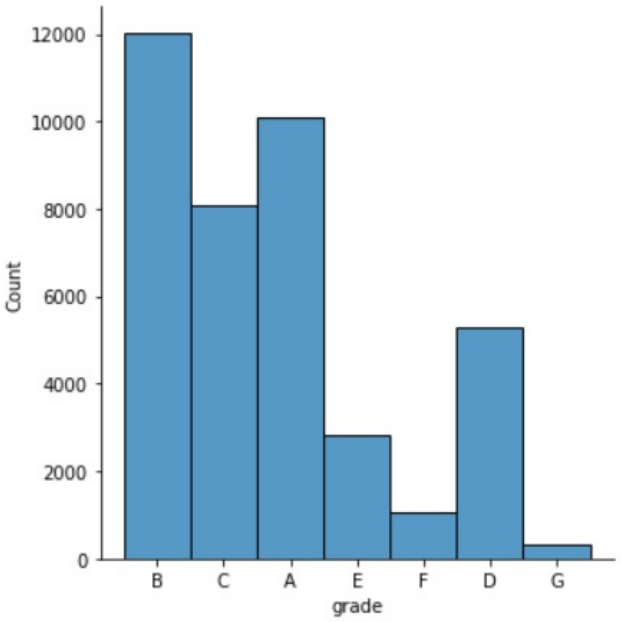


Most of the customers are non-delinquent

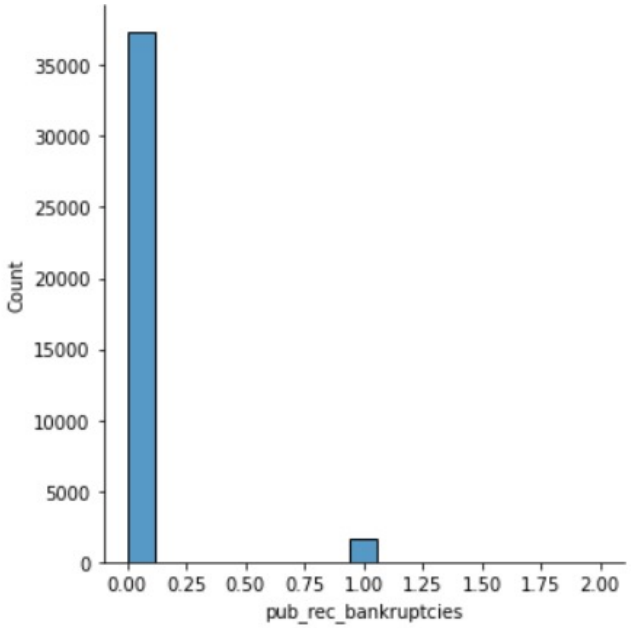


Most of the customers have employee length greater than 10 years

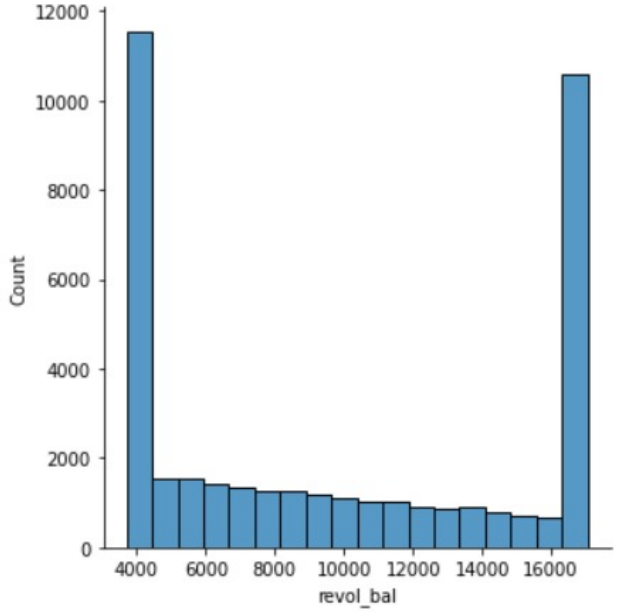
Univariate Analysis



Most of the customers are A or B graded

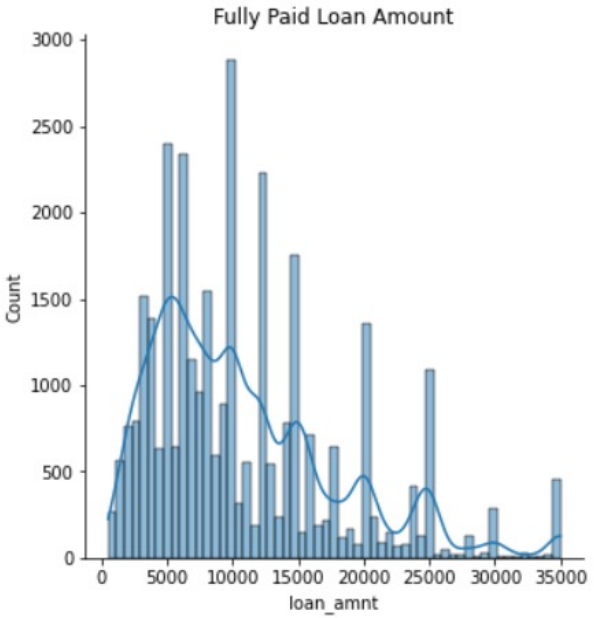


Most of the customers are non-delinquent

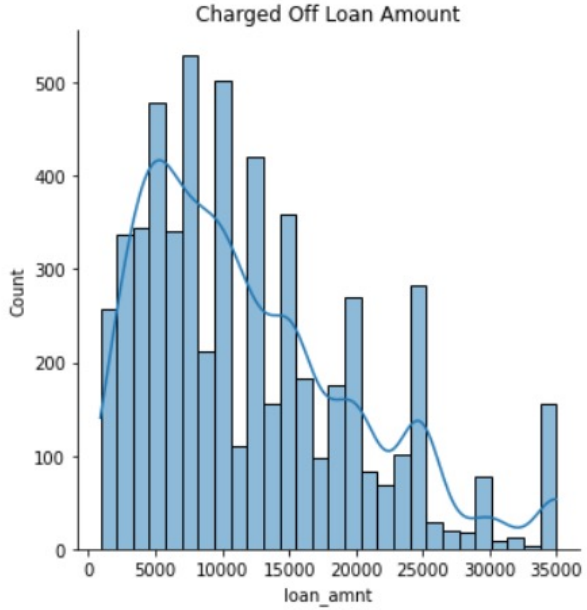


Most of the customers have either higher or lower revolving balance

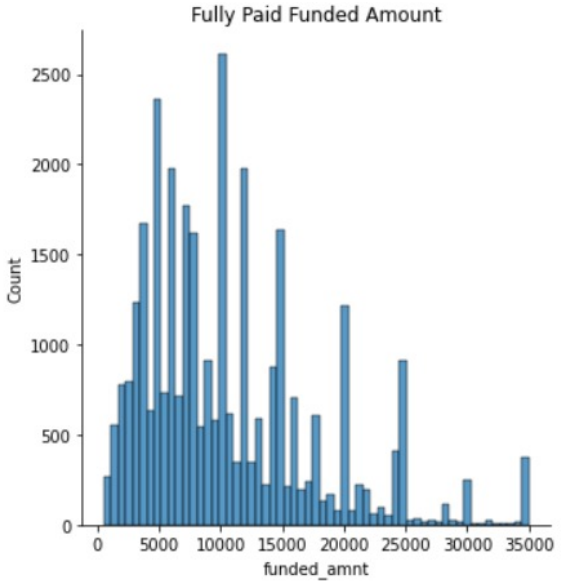
Segmented Univariate Analysis



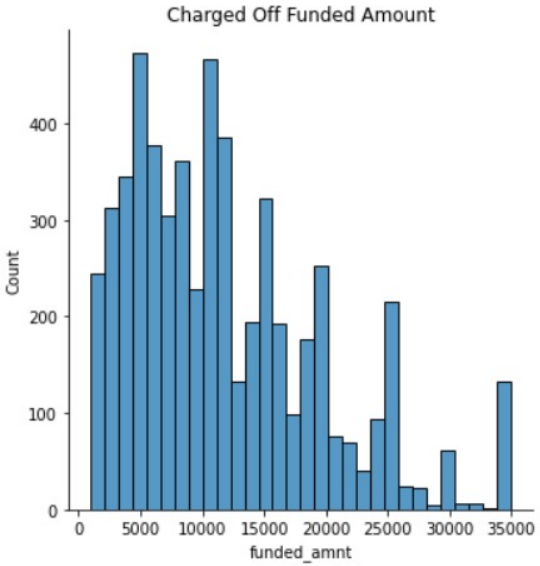
Most of the customers are A or B graded



Most of the customers are A or B graded

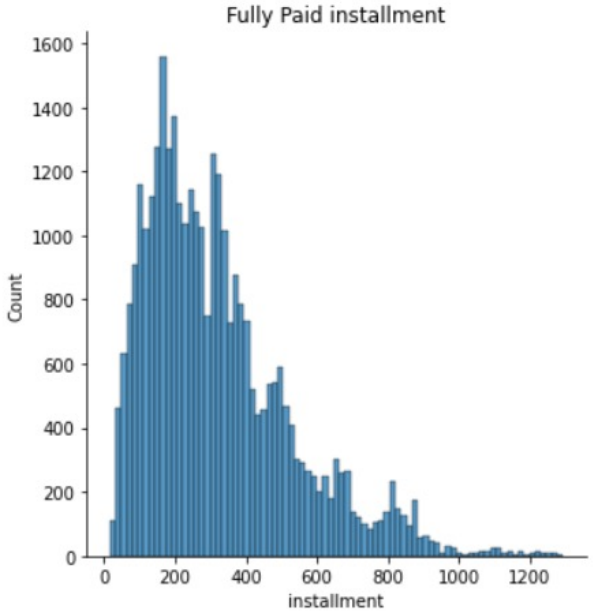


Most of the customers are A or B graded

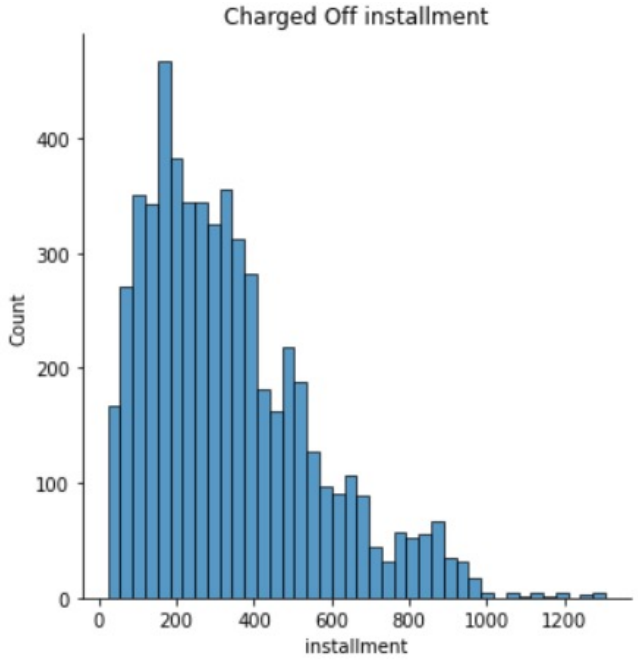


Most of the customers are A or B graded

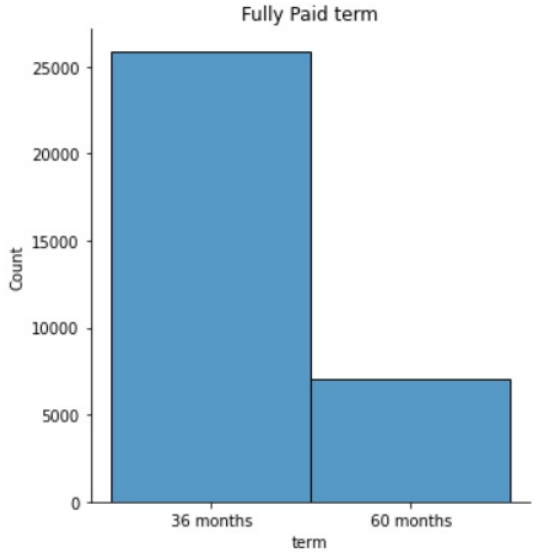
Segmented Univariate Analysis



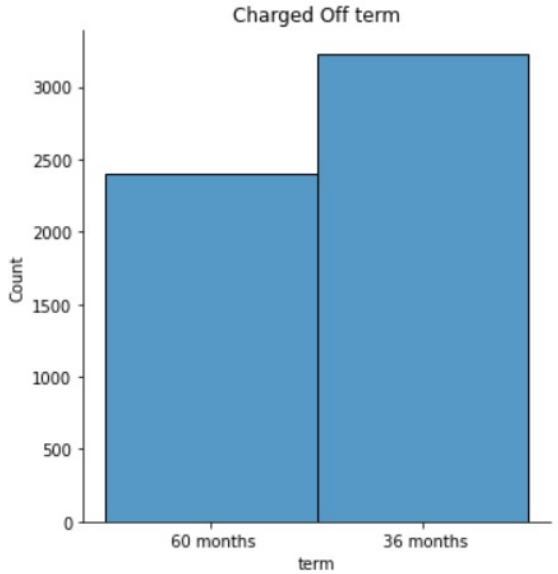
Most of the customers are A or B graded



Most of the customers are A or B graded

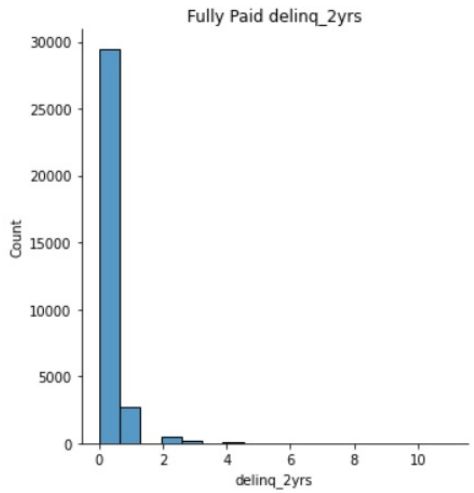


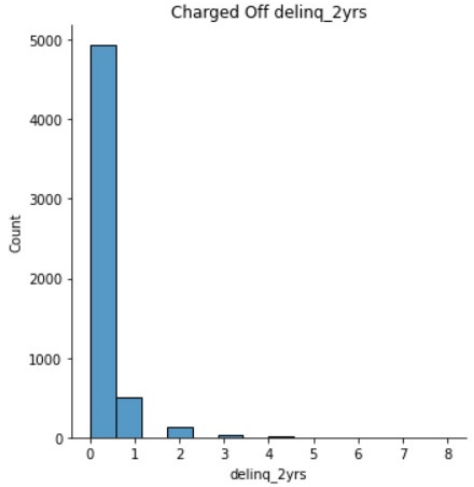
Most of the Fully Paid loaners are pf 36 months tenure



May defaulters have 60 months tenure

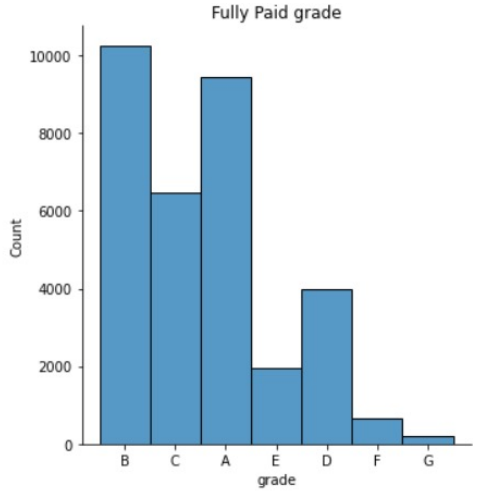
Segmented Univariate Analysis



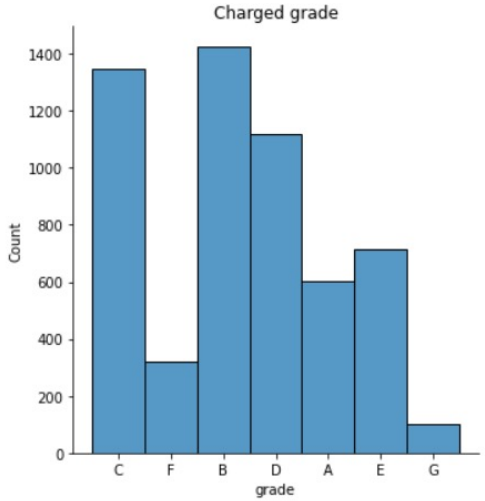


Most of the customers have delinquency as 0

Most of the customers have delinquency as 0

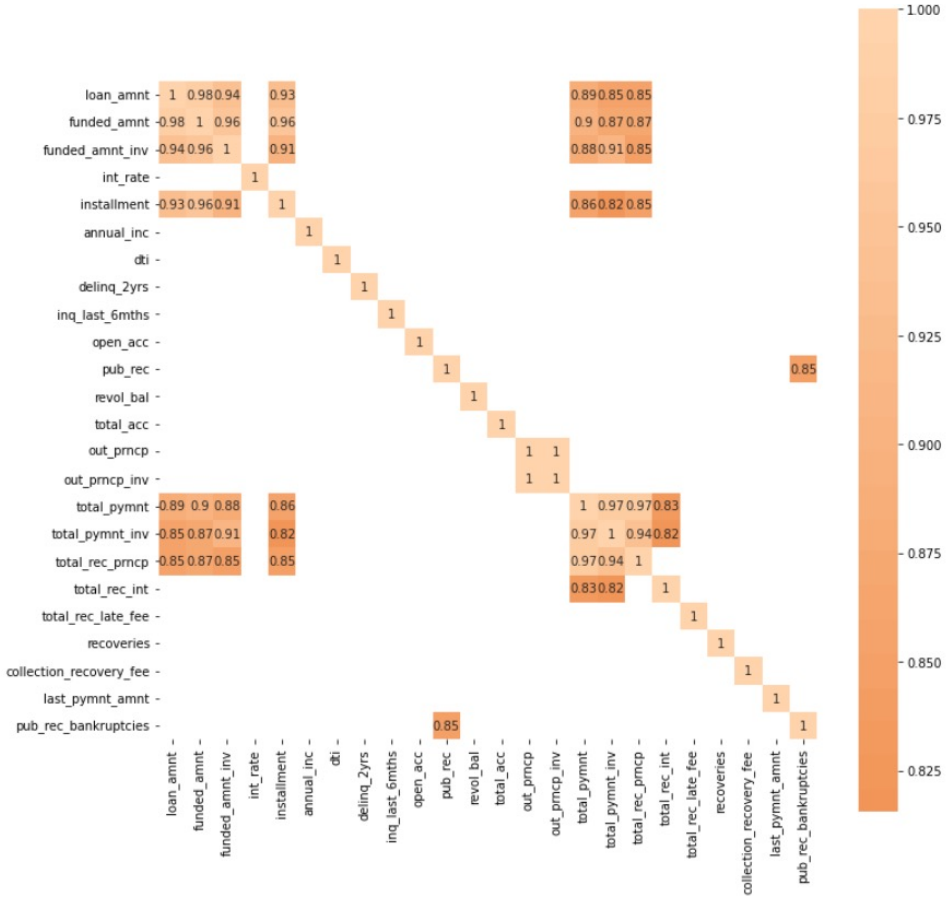


Most of the customers are A or B graded



Most of the customers are A or B or D graded

Bivariate / Multivariate Analysis

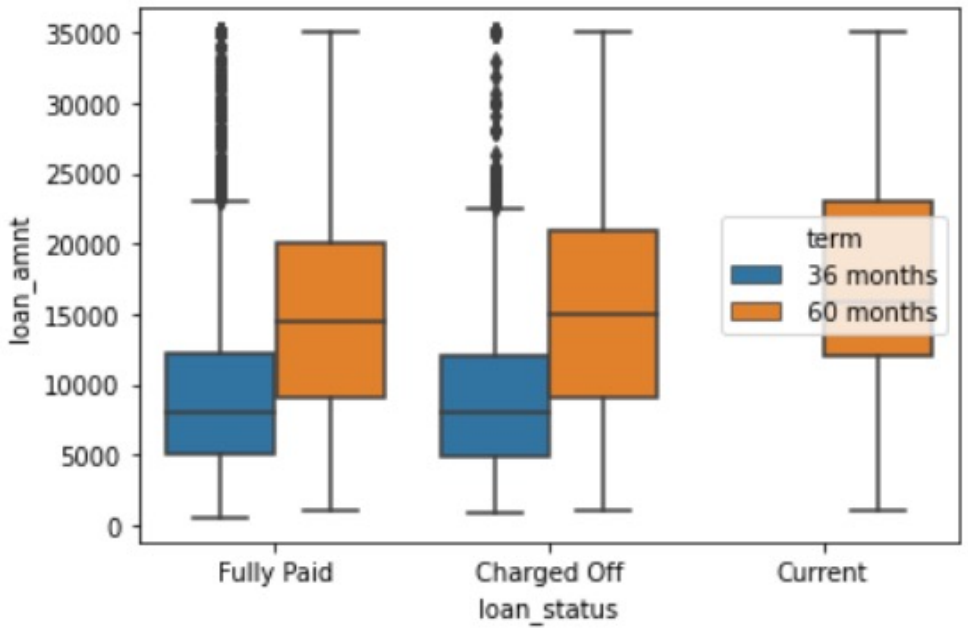


Variables that are highly correlated are …

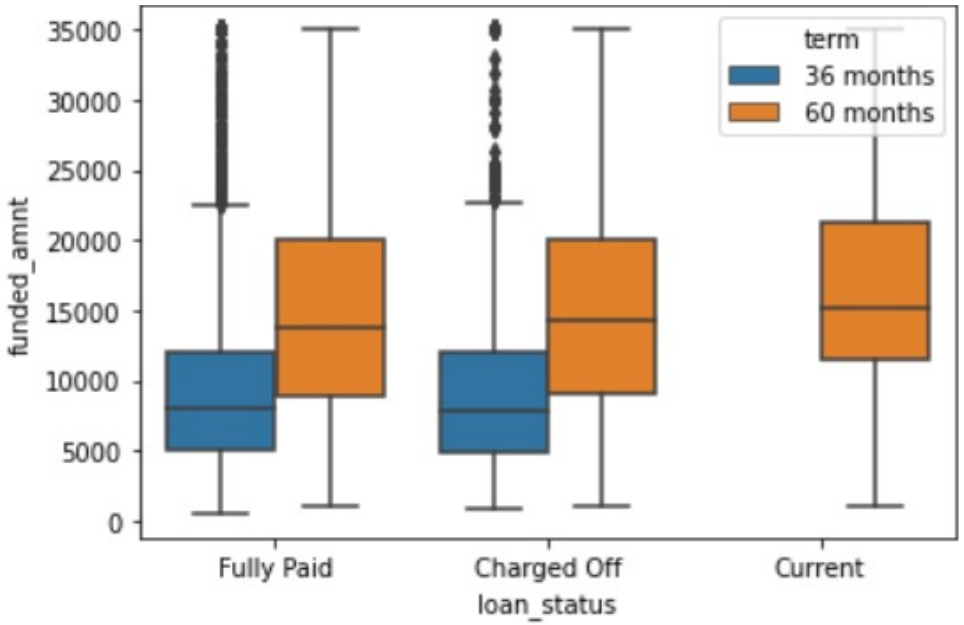
Variables that are moderately correlated are…

Variables that are poorly correlated are…

Bivariate analysis

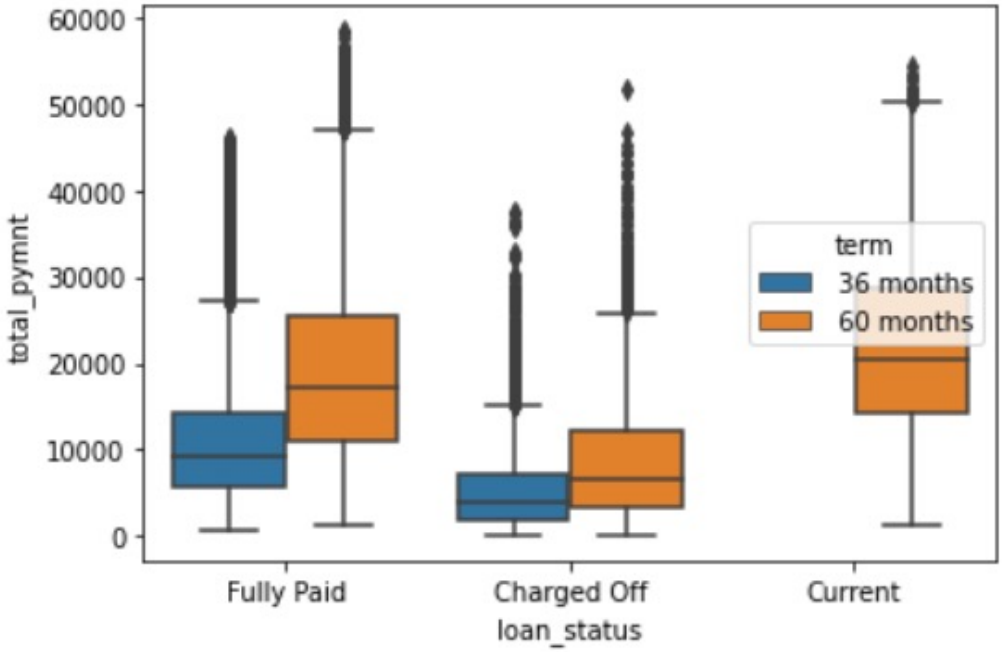


Median for 60 months tenure is higher

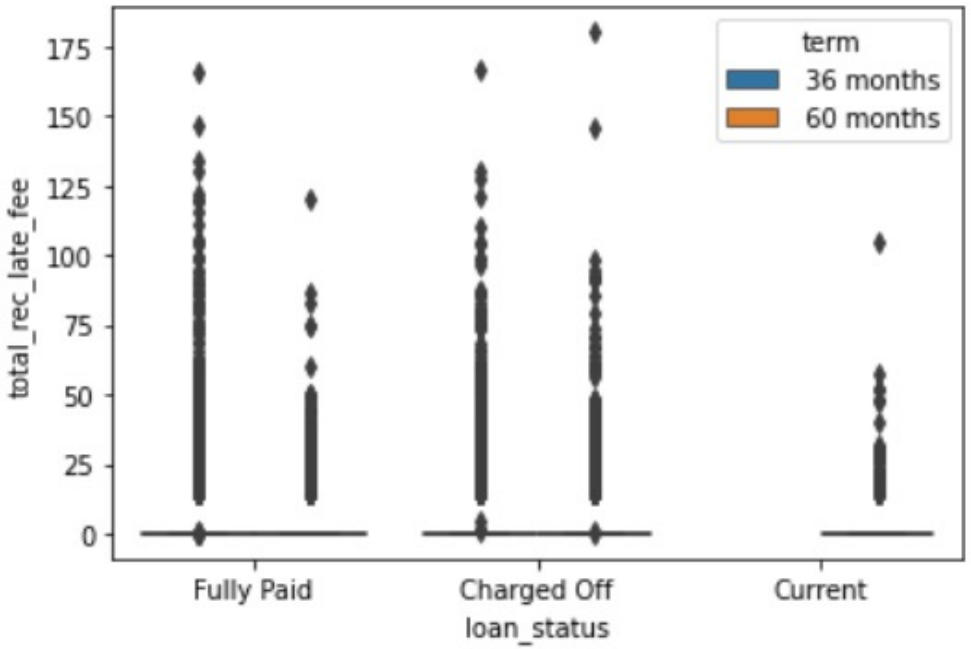


Median for 60 months tenure is higher

Bivariate Analysis

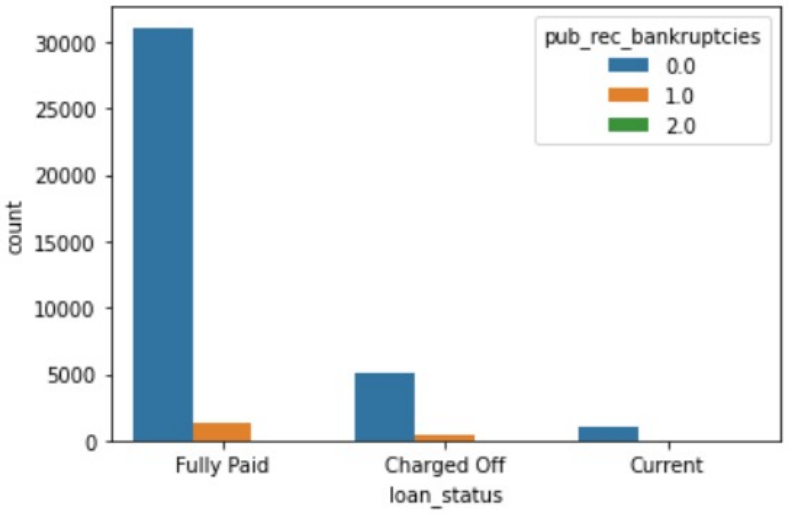


Charged off Customers have lower total payment median

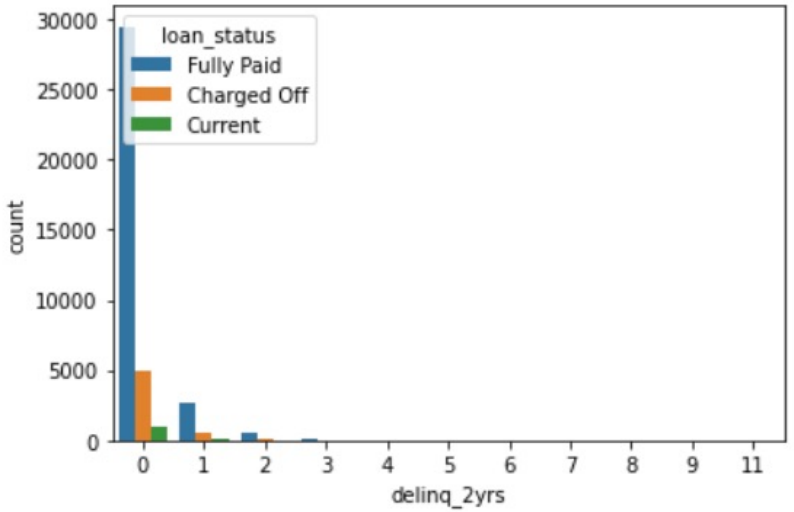


Charged off customers have paid higher late fee for 60 months term

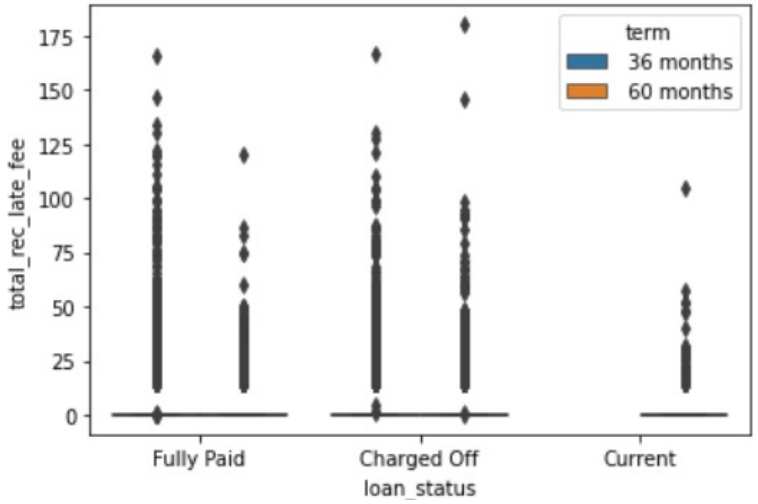
Bivariate Analysis



Charged off customers have higher bankruptcies

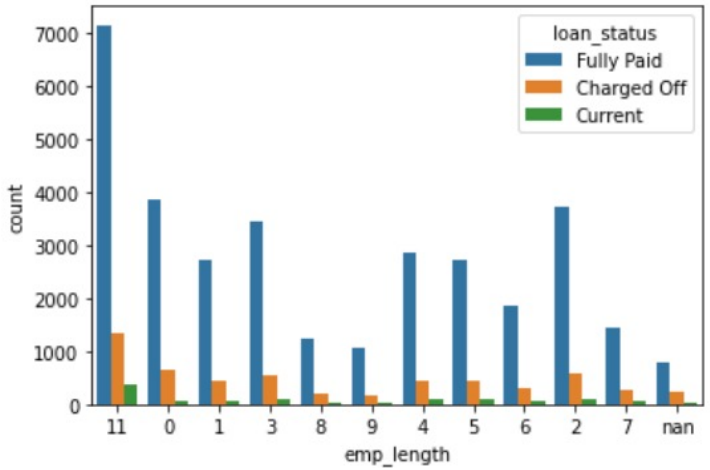


Less customers with delinquency

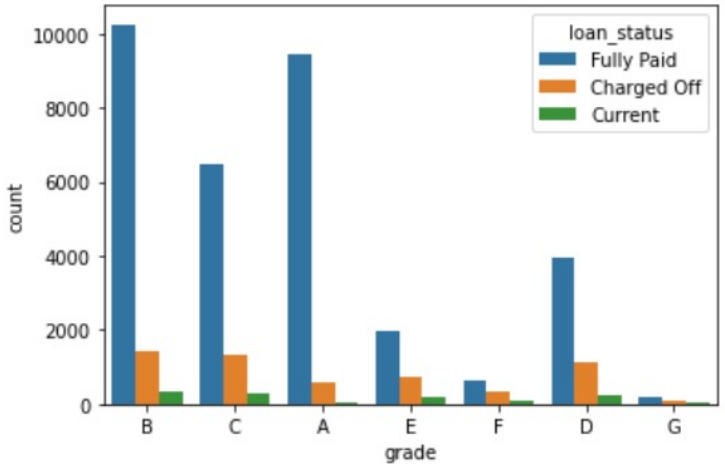


Charged off customers have more outliers in 60 months period

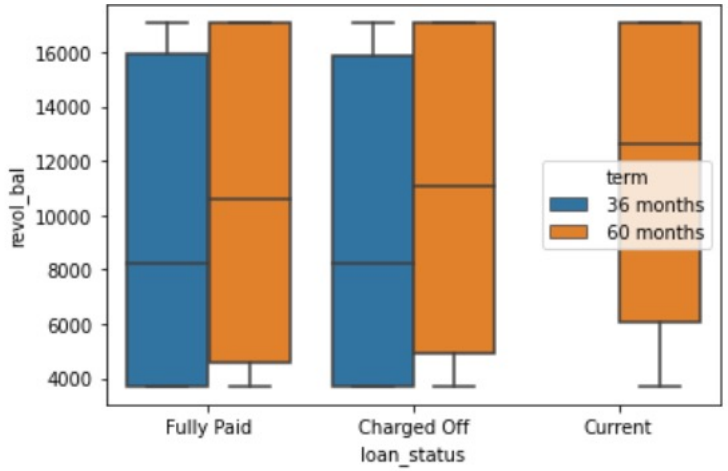
Bivariate Analysis



Employee Length is higher for Fully Paid customers

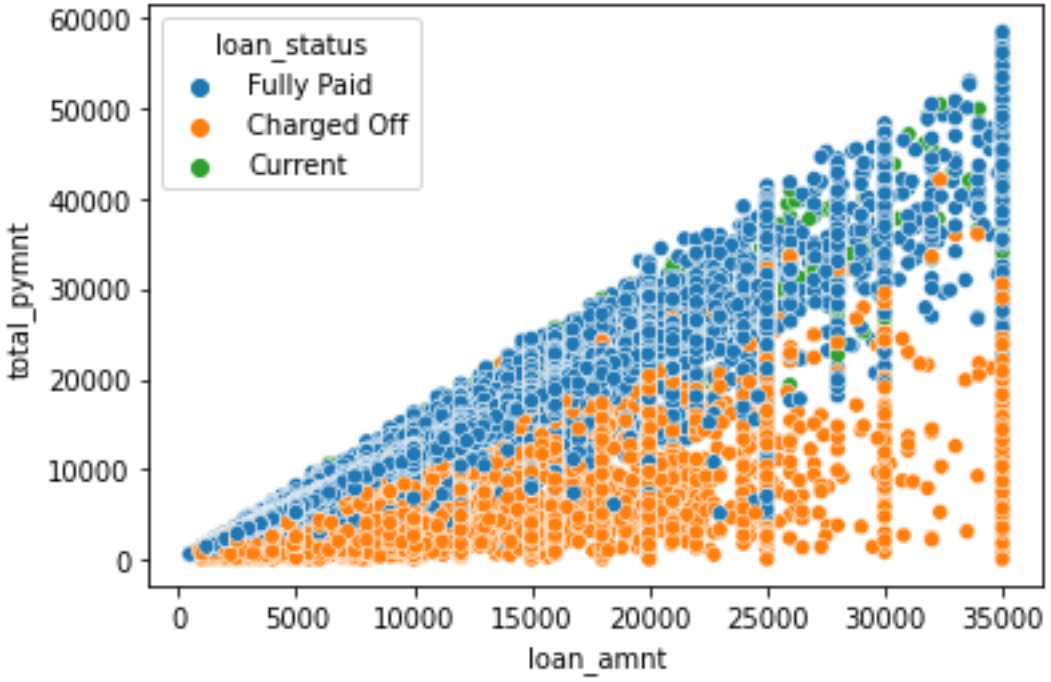


Most Fully Paid customers are having grade A or B Or D

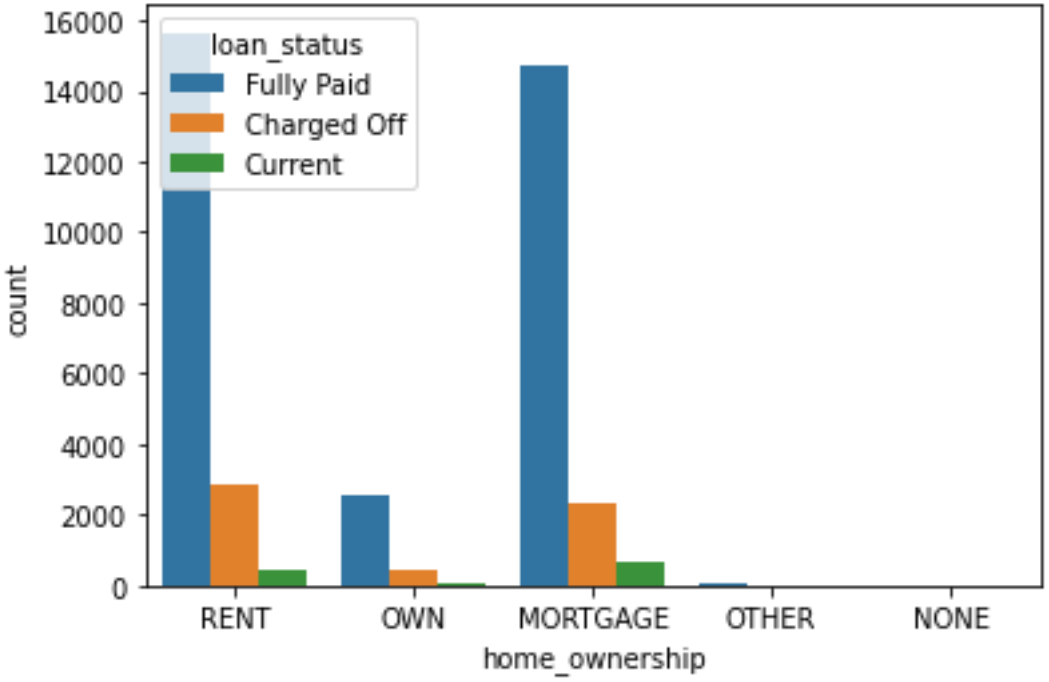


Revolving Balance has similar median for Fully Paid and Charged off across 36/60 months tenure

Bivariate Analysis



Defaulting Customers are having less total payments



Defaulting Customers have rented or mortgage home ownership

Final Analysis

- Based on the above EDA performed the **driving factors (or driver variables)** behind loan default can be
  - Employee length
  - Installment
  - Total received late fee
  - Total payment
  - Home ownership