Gramener EDA Case Study
SUBMISSION REPORT

- To use Concepts of EDA and decipher which types of Customers default on a loan.

Group Members : Rajarshi Palit | Amani Prasad | Vignesh R | Harsha Ravi

Business Objective

- The objective is to identify patterns which indicate if a borrower of a loan is likely to default, which may be used by a consumer finance company for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. The aim is to understand and highlight the **consumer attributes** and **loan attributes** that influence the tendency of default.

Problem Solving Methodology: The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals and give more meaning to the analysis. The checkpoints are represented in a sequential flow as below:

Importing the Data Set & Data Cleaning → Understanding the Data by performing Univariate Analysis for Numeric and Categorical Variables → Segmenting data from certain columns into slots for Segmented Univariate Analysis → Comparing sets of 2 numeric variables for Bivariate Analysis → Combining multiple categorical columns and performing Multivariate analysis to find a trend → Drawing conclusion regarding driver variables that lead to loans being Charged-off (defaulted).

Data Cleaning & Creating Derived Metrics

The following data cleansing processes were applied to make the data dependable so that it can provide significant business value by improving Decision Making Process:

Removing columns with unique values

- Some of the columns from the loan dataframe have only value. Deleting these columns with one unique value from the dataframe.

Removing 'Current Loans'

- For the current analysis, we are interested in only FULLY PAID and CHARGED OFF loan statuses. Removing all loans with statuses = "CURRENT" because we have no way of knowing if they will turn out to be fully_paid or charged_off loans.

Dropping Null values in Rows & Columns

- The **null values in some columns** that aren't required for analysis have been dropped from the loan_dataframe
- High Percentage of Null values in some columns doesnt make any sense or help in analysis. Hence **all rows containing null values** for these columns were dropped from master_frame.

Creating new Columns for Analysis

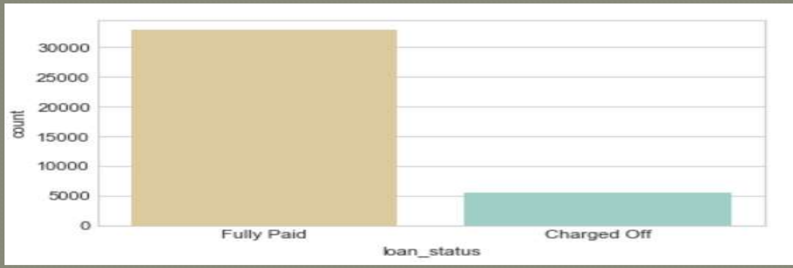
- Some columns like 'int_rate%' & 'revol_util%' have been broken into categorical slots for segmented analysis.
- Creating 2 new columns 'charged_off' and 'fully_paid' with values 1 for True and 0 for False. If Loan status = 'Charged Off', CHARGED_OFF column will have the value 1, else 0. Similarly, if Loan status = 'Fully Paid', FULLY_PAID column will have the value 1, else 0.

Univariate Analysis

The most important insights from Univariate Analysis are as follows:

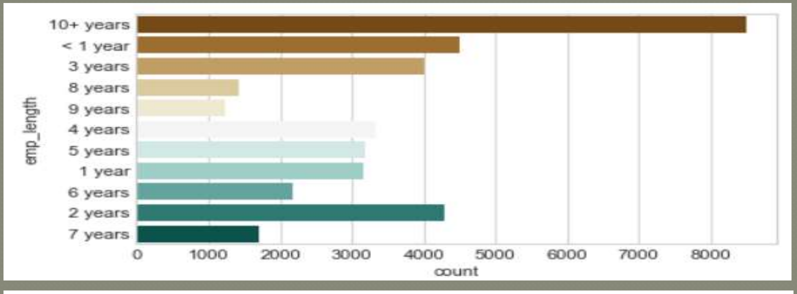
Loan Status

- About 1/7th of all accepted loans are 'Charged off'. The rest are fully paid. ('Current' Loans)



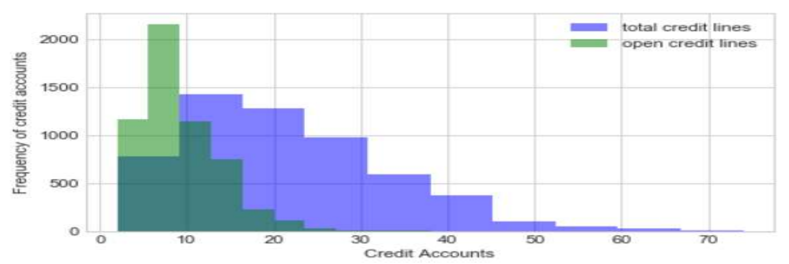
Employment Length

- Most of the accepted loans belong to borrowers having employment length of either more than 10 yrs or 0-3 years



Credit Lines

- Both the open credit lines and total credit lines have a decreasing trend in frequency as no of credit lines increase

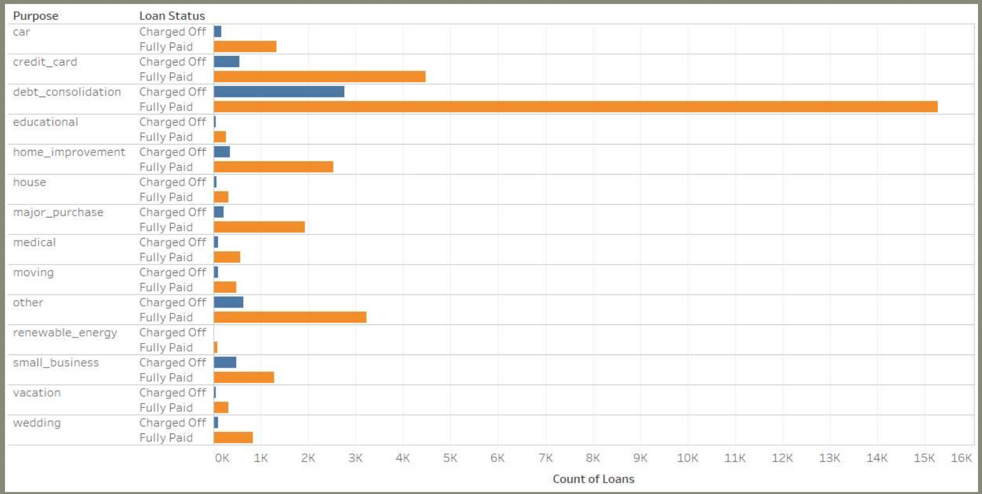


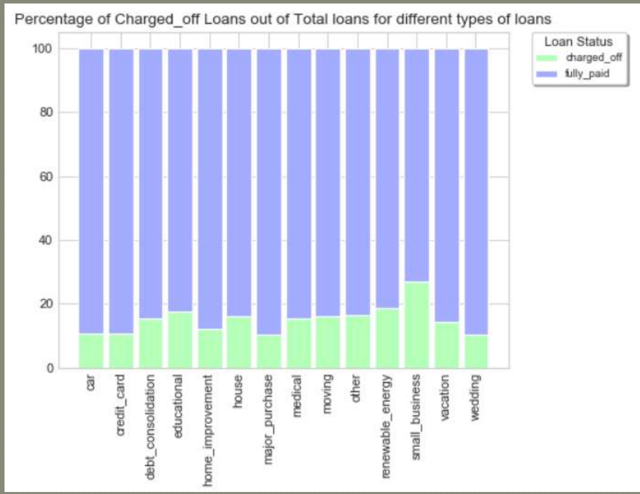
Univariate Segmented Analysis

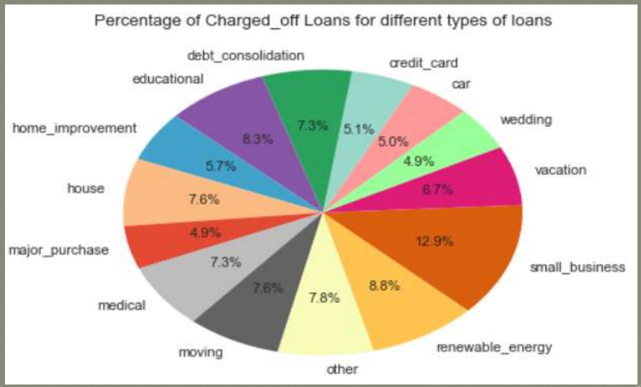
The most important insights from Univariate Analysis are as follows:

Loan Purpose

- Small business have highest % of charged off loans.
- Debt_consolidation has highest no of charged off loans.



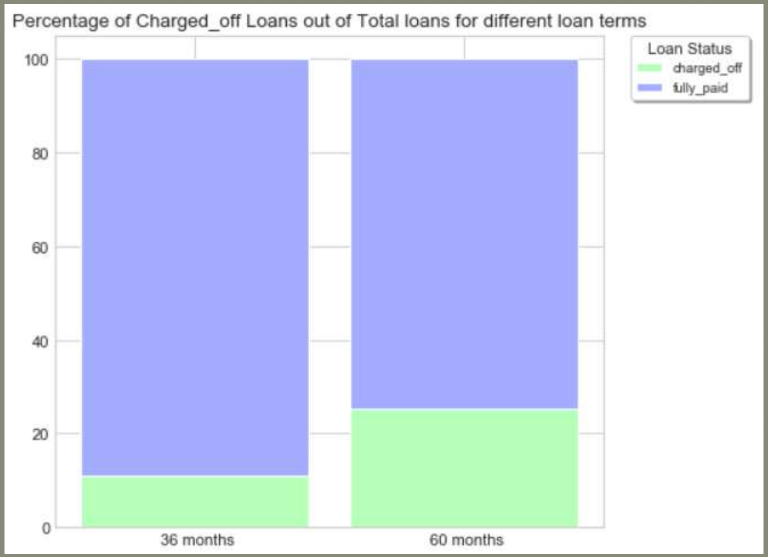


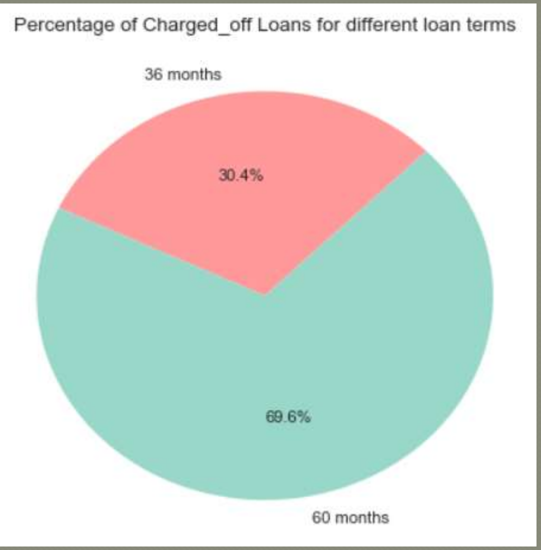


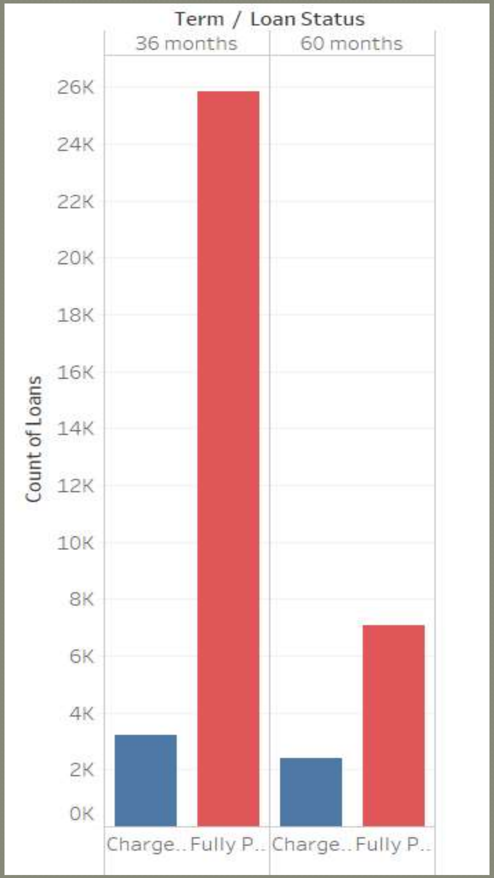
Univariate Segmented Analysis: contd…

Loan Term

- Loans with term=60 have more than double the the percentage of being charged off than those with term=36.
- There are more number of loans with term=36 that have become charged off than those with term=60.



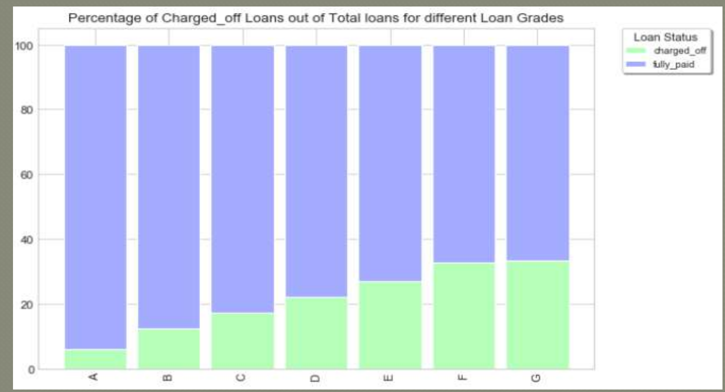


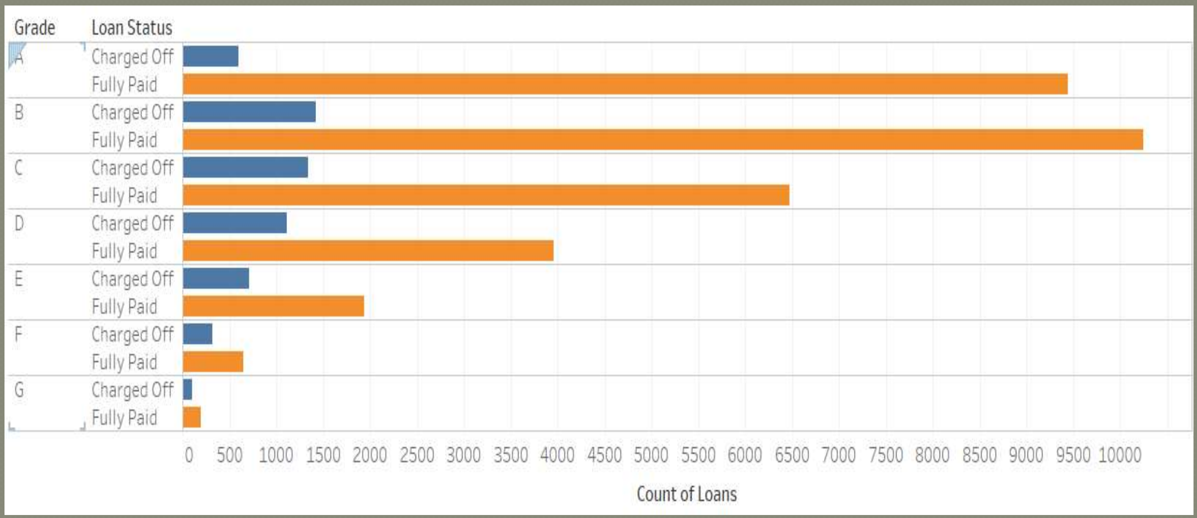


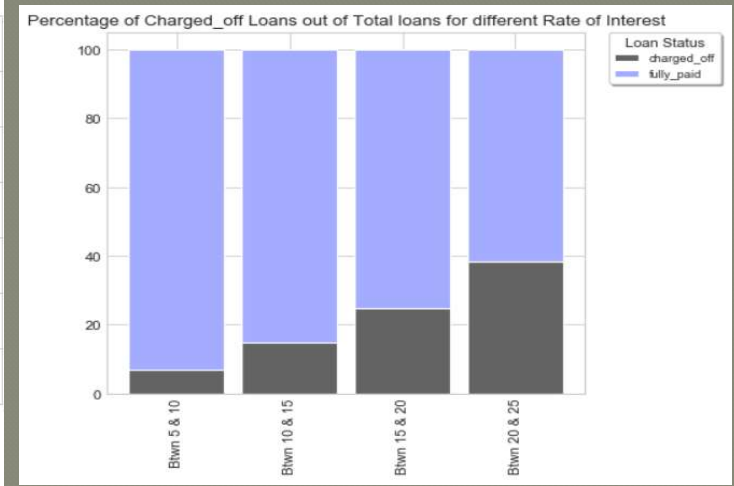
Univariate Segmented Analysis: contd…

Loan Grade & Rate of Interest (interdependent)

- Grade & Rate of Interest are interdependent on each other. Lower the Grade, higher is the Rate of interest.
- Percentage of Charged off loans have a more or less gradual increase with decrease in Grade & increase in Rate of Interest.
- Count of Charged off loans is highest for Grade B followed by A and C.



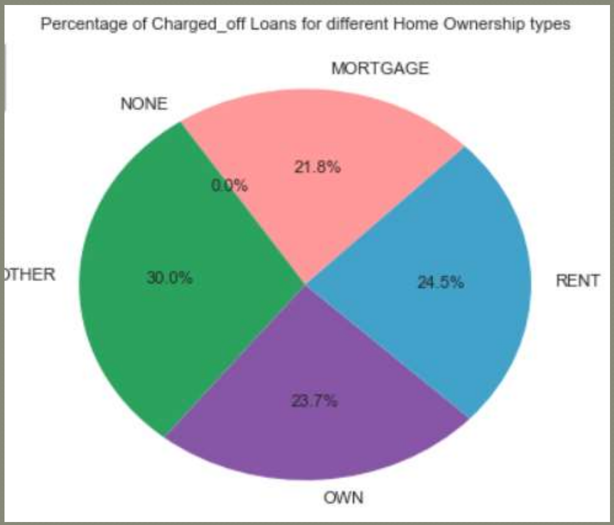


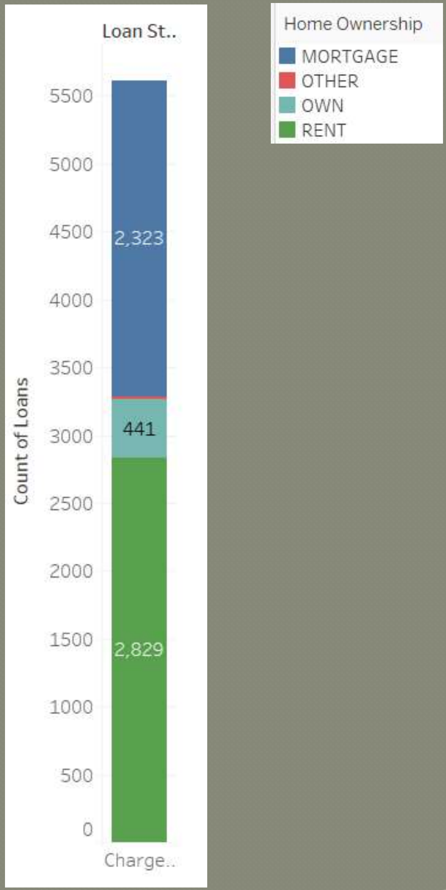


Univariate Segmented Analysis: contd…

Home Ownership

- Most number of Charged-off loans happen for borrowers living in Rented homes or homes on Mortgage
- Percentage of Charged off loans is however same more or less in all ownership types

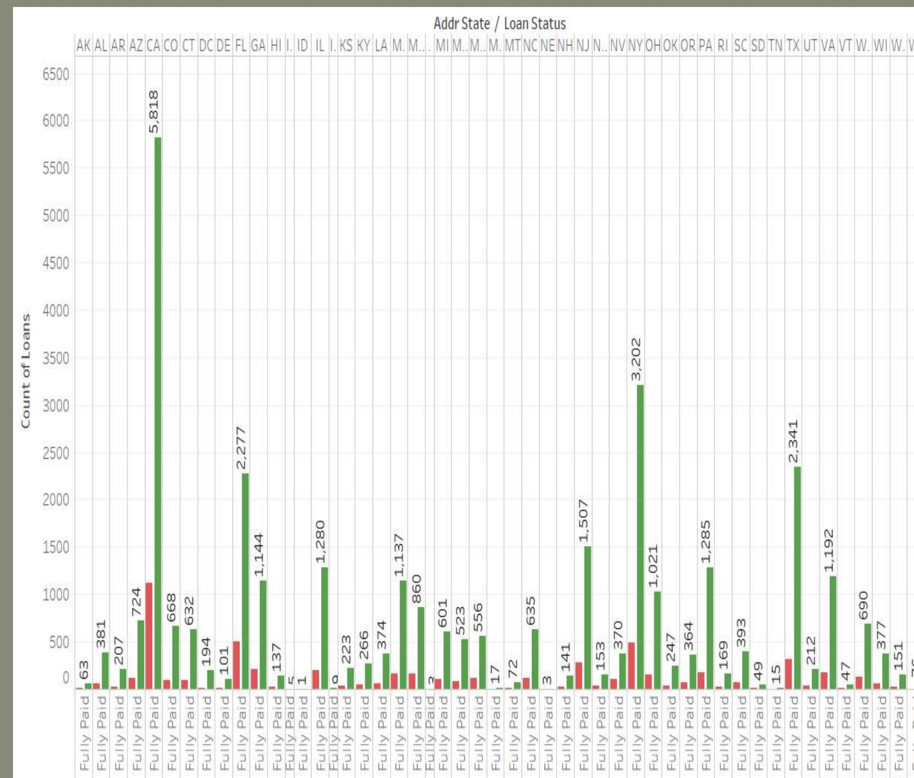


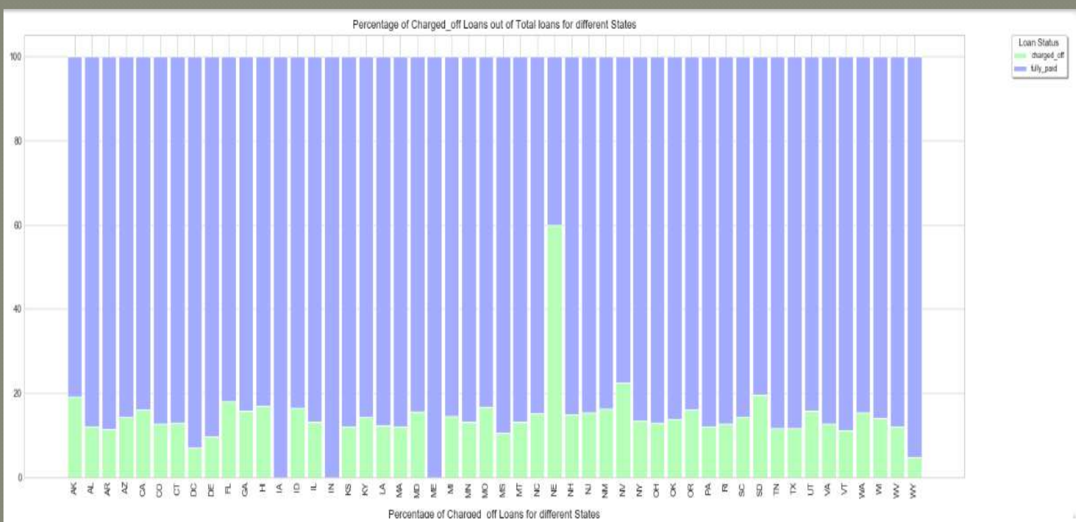


Univariate Segmented Analysis: contd…

Address State

- Percentage of charged off loans is highest in the state NE .
- The count of charged off loans is most in the state CA followed by NY, FL, TX and NJ.

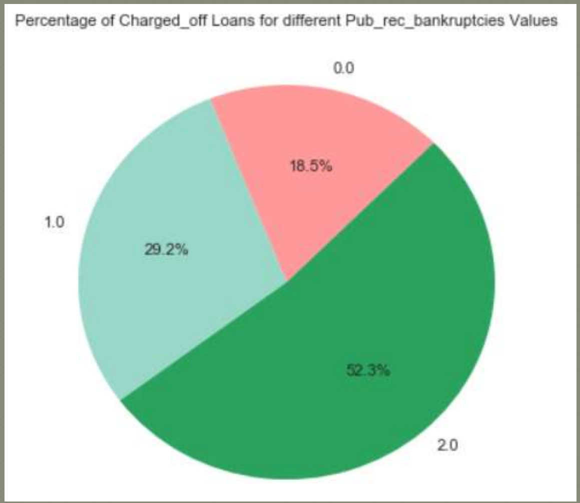


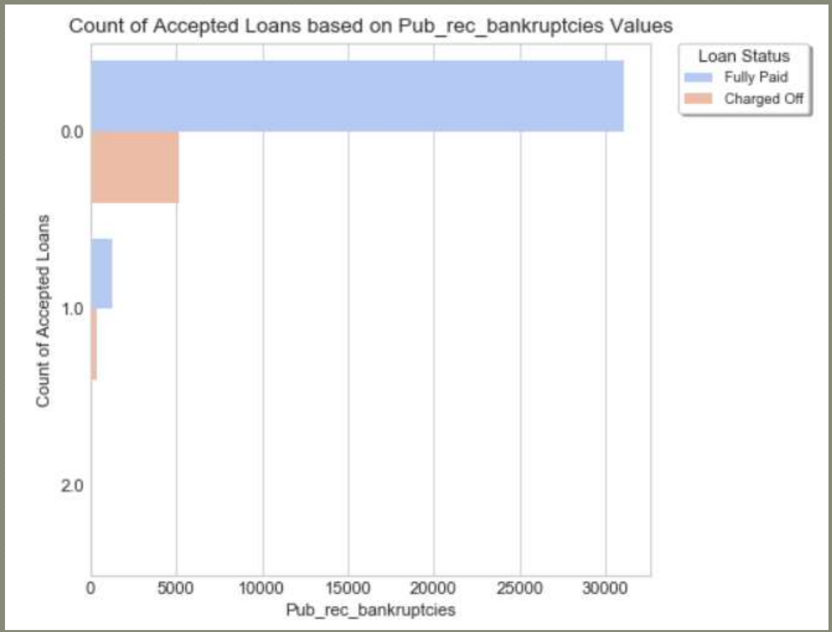


Univariate Segmented Analysis: contd…

Public Record Bankruptcies

- Percentage of charged off loans gradually increase with increase in the value of Pub_rec_bankruptcies.
- However, there are very few records with bankruptcy value = 2 or 1. Most of the charged off records have '0' bankruptcies.

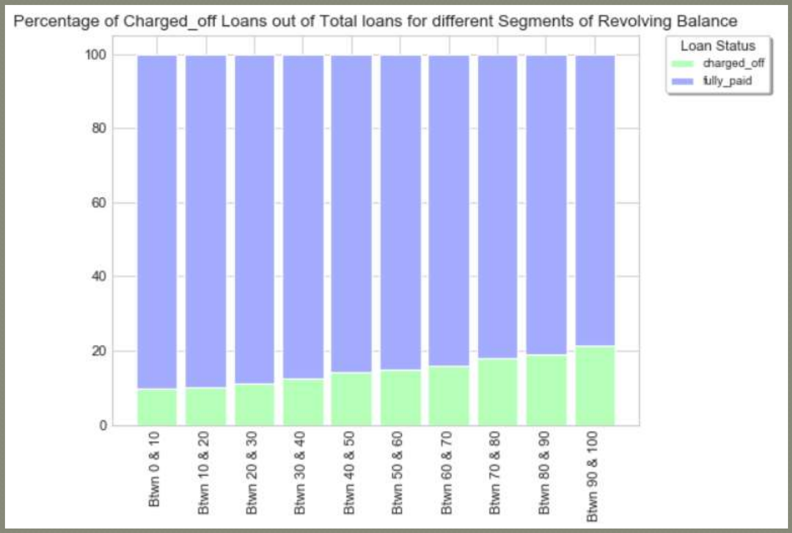


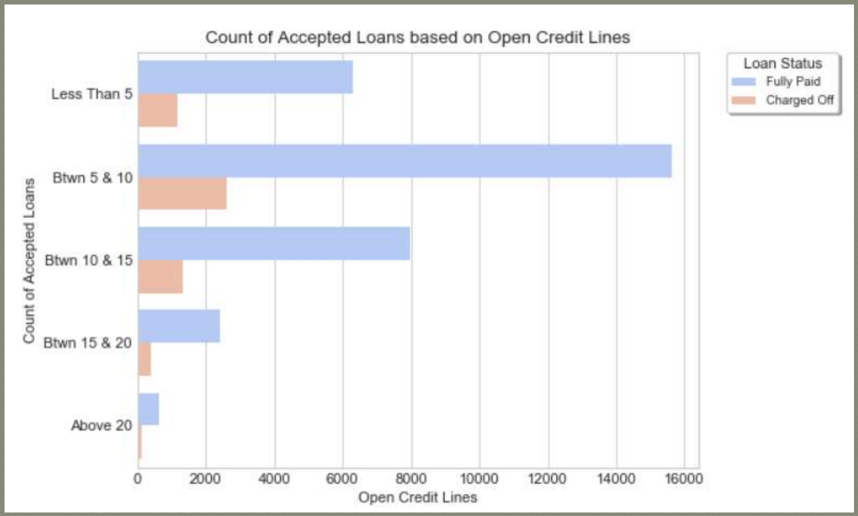


Univariate Segmented Analysis: contd...

Revolving Credit Utilization

Percentage of charged off loans gradually increase with increase in the value of revol_util.





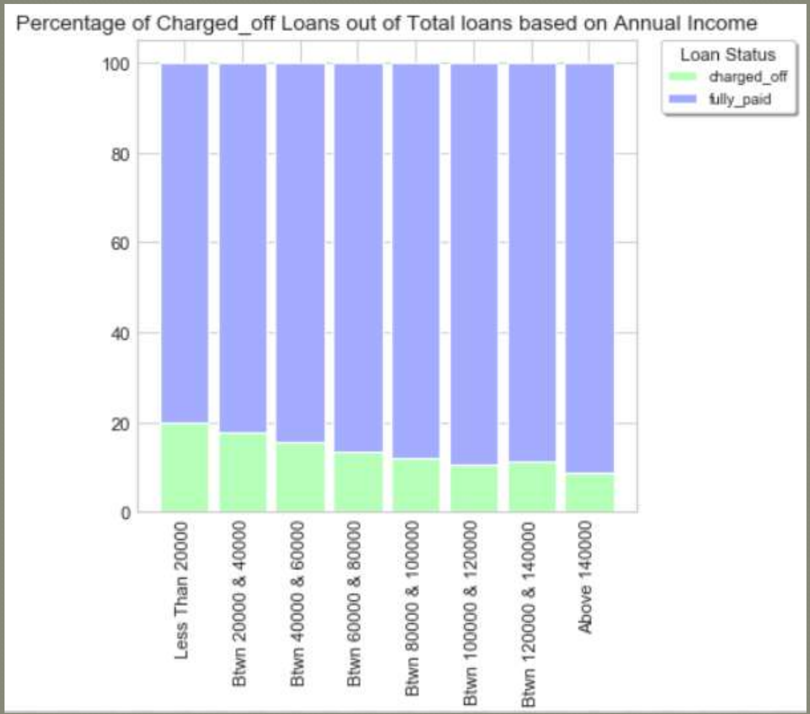
Count of charged off loans is highest when no of Open Credit Lines is between 5 & 10.

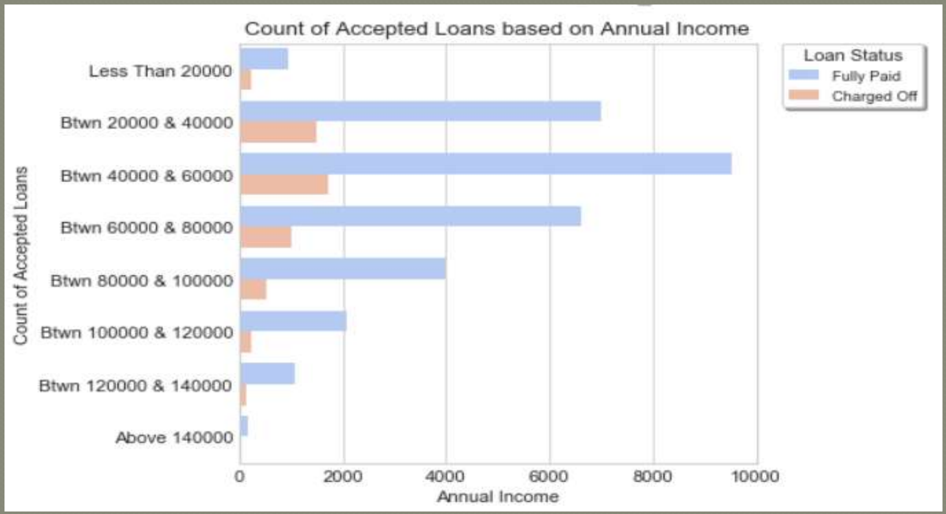
Open Credit Lines

Univariate Segmented Analysis: contd…

Annual Income

- The percentage of Charged off loans decreases more or less with the increase in annual income of the borrower.
- Most no of charged off loans can be seen for borrowers with annual income between 40K-60K followed by 20K-40K

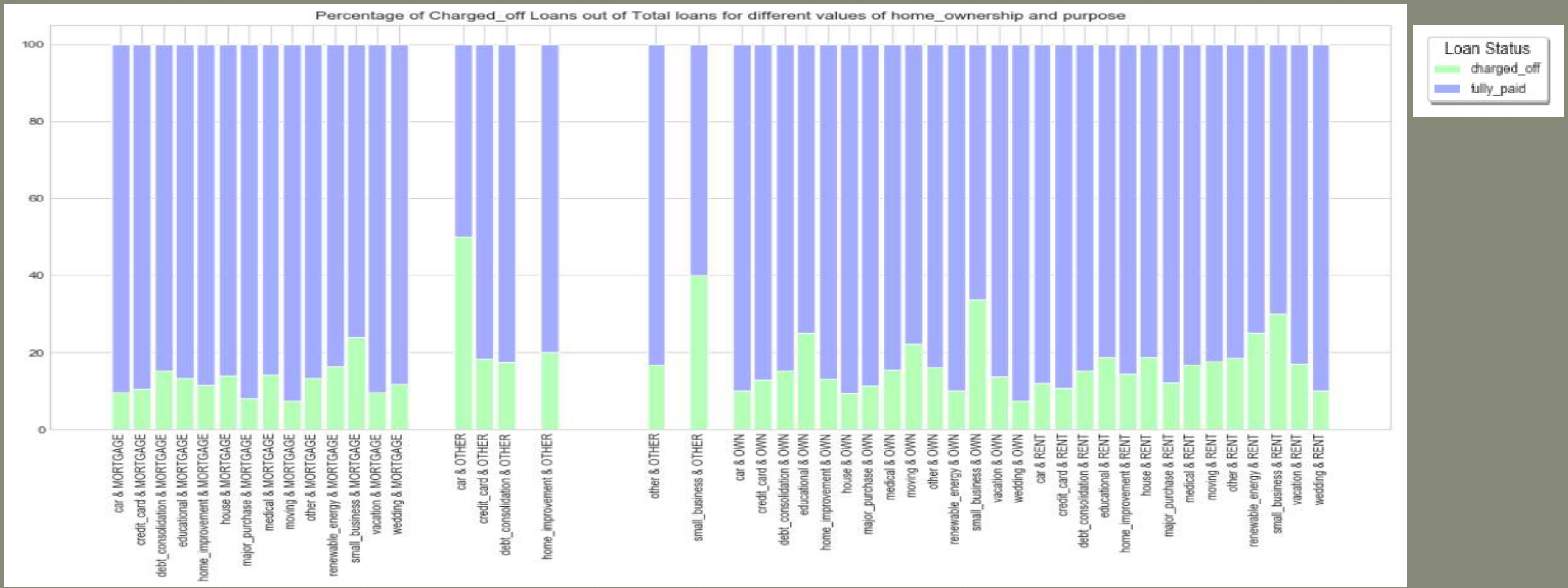




Bivariate & Multivariate Analysis:

Home Ownership & Loan Purpose

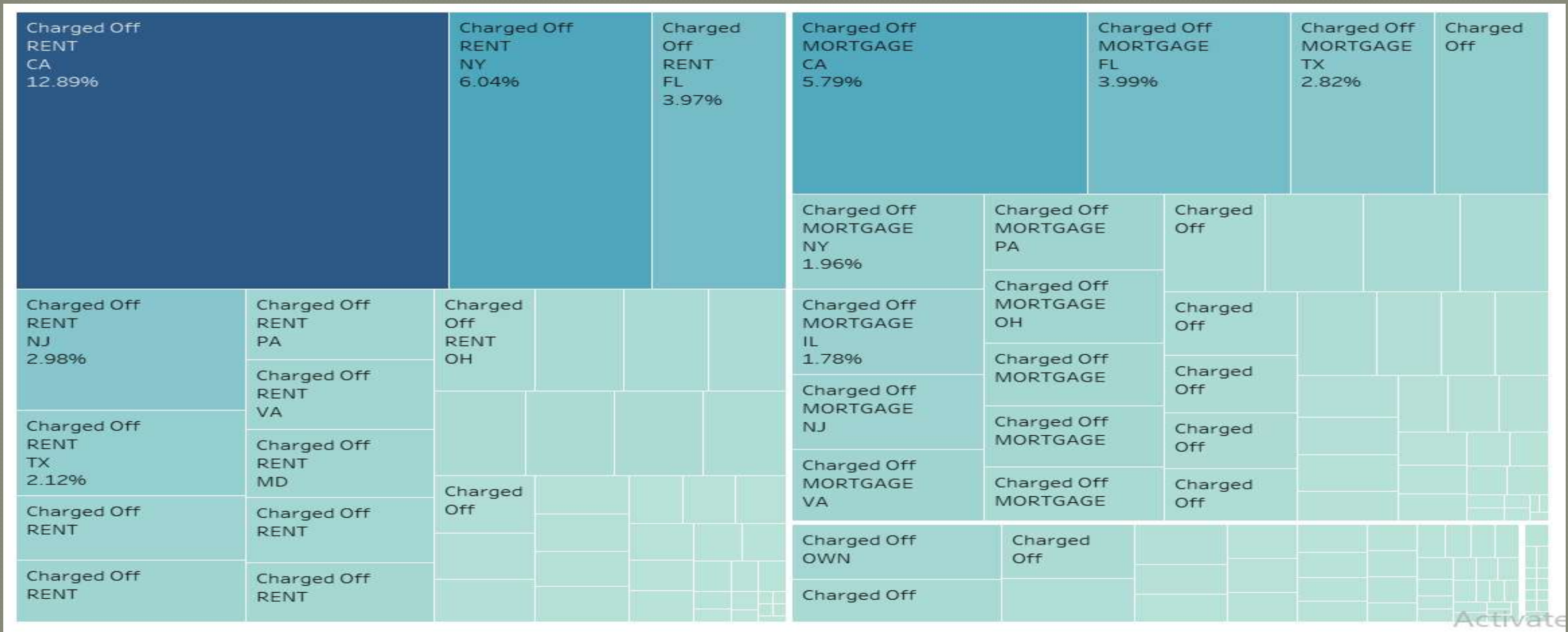
- The percentage of Charged off loans is highest for the below categories in descending order
  - loan_purpose='car' and home_ownership='Other'
  - loan_purpose='small_business' and home_ownership='Other'
  - loan_purpose='small_business' and home_ownership='Own'



Bivariate & Multivariate Analysis:

Home Ownership & Address State

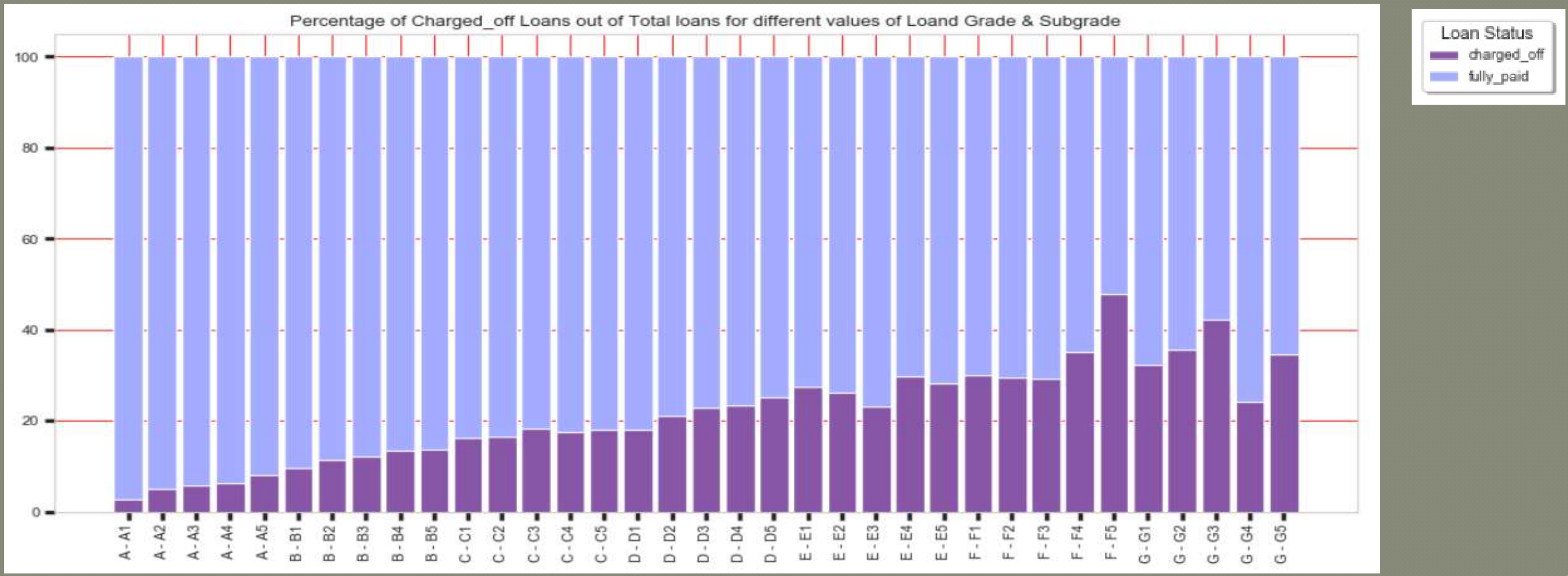
- The Count of Charged off loans is highest for the below categories in descending order:
  - Address State='CA' and home_ownership='Rent'
  - Address State='NY' and home_ownership='Rent'
  - Address State=CA' and home_ownership='Mortgage'



Bivariate & Multivariate Analysis:

Grade& Sub-grade

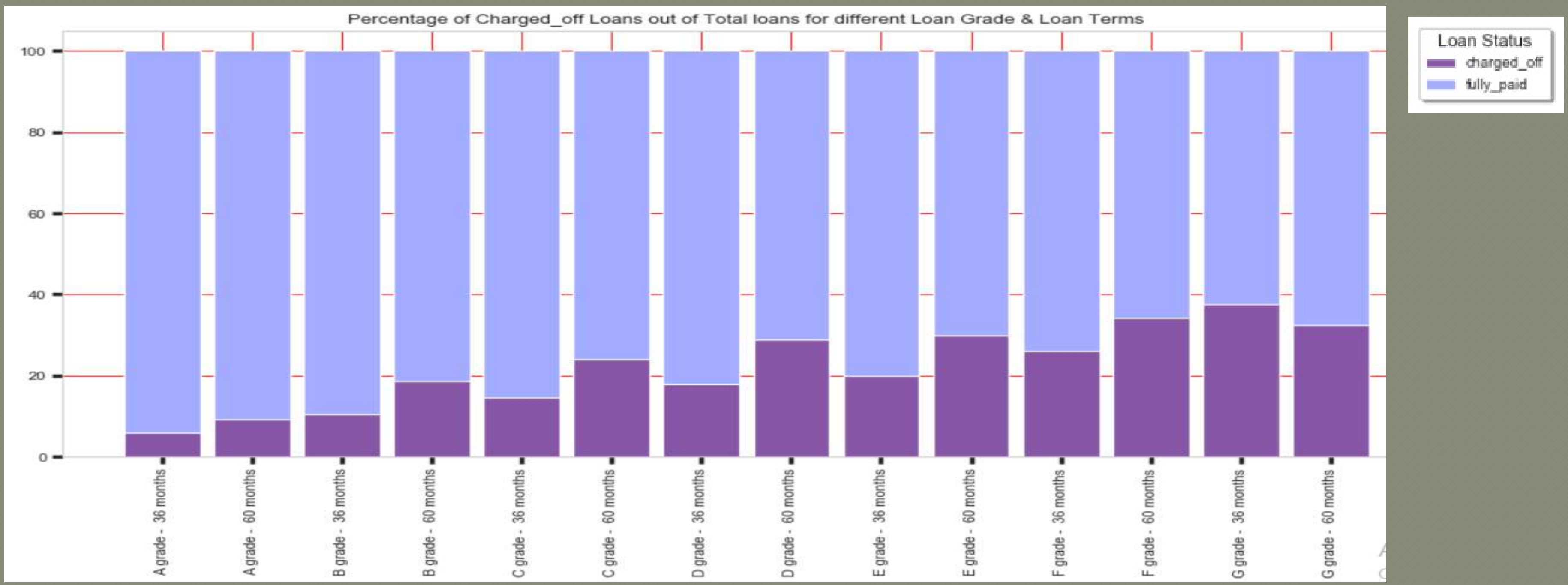
- The percentage of Charged off loans have a more or less gradual increase across Loan Grde and Sub-grade from A1 to G5.
- The percentage is highest for subgrade F5 followed by G3 and G5 and lowest for A1



Bivariate & Multivariate Analysis:

Grade & Loan Term

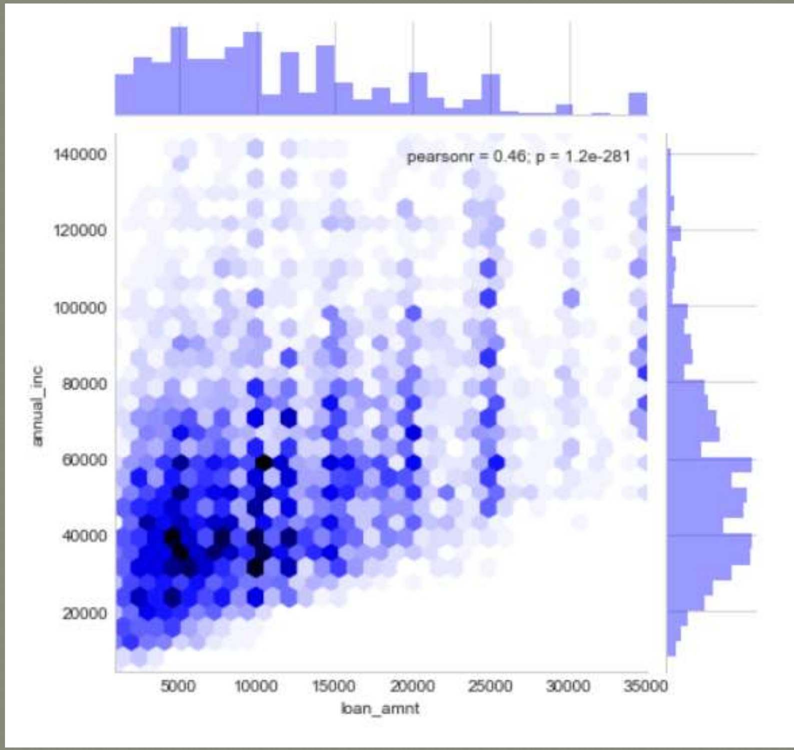
- The percentage of Charged off loans increase with decrease in loan grade and within each grade, it is usually more for term = 60 months and lesser for term = 36 months.



Bivariate & Multivariate Analysis:

Annual Income & Loan Amount

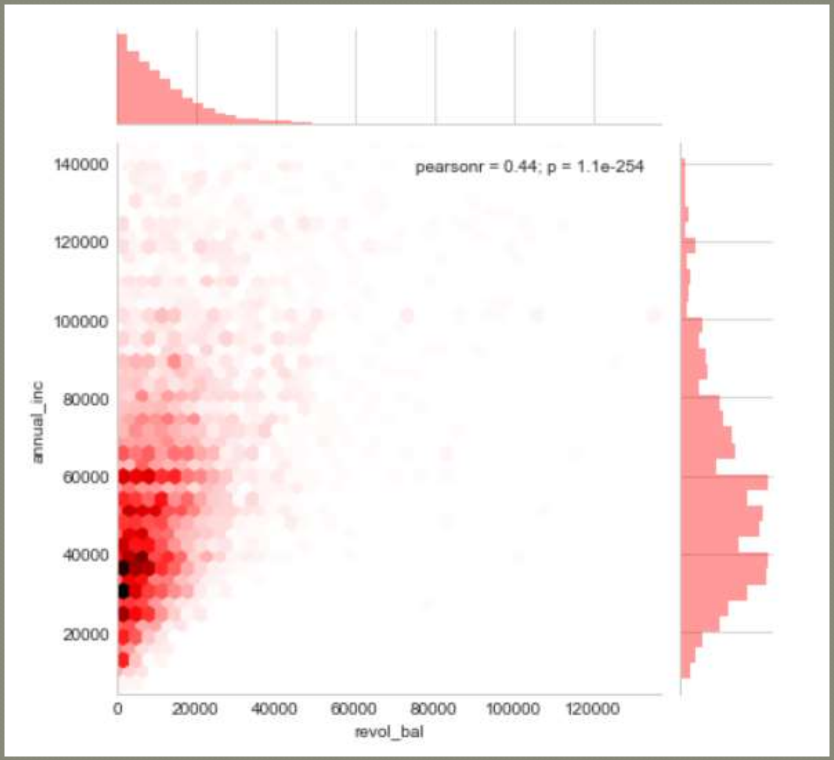
- For Charged off loans correlation coefficient between borrower's Annual Income and Loan amount is 0.46.
- As Annual income increase, loan amount also tends to increase whereas most of the observations are concentrated where annual income is between 30000 - 60000 and loan amount is below 13000



Bivariate & Multivariate Analysis:

Annual Income & Revolving Balance

- For Charged off loans correlation coefficient between borrower's Annual Income and Revolving balance is 0.44.
- The frequency of Defaulted loans is most below an annual income of 60K when revolving balance is under 20K.



Inferences

The driver variables and their combinations, which have been derived from the analysis, play a significant role in a loan being defaulted. The same has been illustrated below:

Five Most Important Driver Variables

- **Loan Purpose**
- **Loan Grade**
- **Loan Term**
- **Address State**
- **Revolving Utilization %**

Three Most Important Driver Variable Combinations

- **Annual Income & Loan Amount**
- **Home Ownership in Various States**
- **Loan Grade & Loan Term**

Defaulted Loan

Recommendations

The following points can help Lending Club, the consumer finance company to understand the driving factors (or driver variables) portfolio and risk assessment:

- The company should more concentrate on the category where defaulters are less: Major purchase, credit cards, car, home improvement and wedding loans and avoid categories like 'Small business' loans.
- Analysis shows that the percentage of charged off loans increase from Grades A to G. Hence loans of applicants falling into higher grades (towards A) should be considered for approval.
- Count of Loan defaults in home ownership category of RENT & MORTGAGE is the highest because high home rents may lead to borrowers being delinquent. This is even more observed in states with high cost of living like CA, NY, FL, etc.
- Study shows clearly that short term loans are much safer and hence the company should approve more loans of these kind.
- High utilization rate of the credit line indicates the extravagant spending and this directly impacts the repayment. Company should sanction loans to customers with lower rate of credit line utilization.
- Lower the income, higher the chances of default. Thus annual income is an important customer demographic that needs to be considered while sanctioning loans.