Gramener EDA Case Study Submission Report

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

Prepared By : Jyani AnkitKumar

Data Cleaning & Creating Derived Metrics

- 1) Removing Columns with Unique Value:

Some of the columns from the loan data frame have only value. Deleting these columns with one unique value from the data frame.

2) 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.

Data Cleaning & Creating Derived Metrics

- 3) Dropping Null Values in Rows & Columns:

The null values in some columns that aren't required for analysis have been dropped from the charged off

High Percentage of Null values in some columns doesn't make any sense or help in analysis. Hence all rows containing null values for these columns were dropped from master frame.

4) Creating New Columns for Analysis:

Some columns like ‘int_rate%’ & ‘revol_util%’ have been broken into categorical slots for segmented analysis.

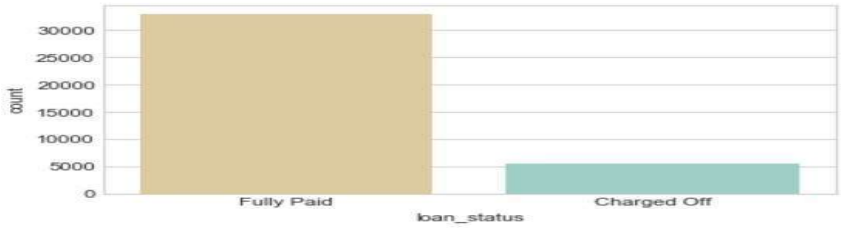
Creating 2 new columns 'charged_off' and 'fully_paid' with values 1 for True and 0 for False. If Loan status = 'Charged Off', CHARGED_OFF column will have the value 1, else 0. Similarly, if Loan status = 'Fully Paid', FULLY_PAID column will have the value 1, else 0.

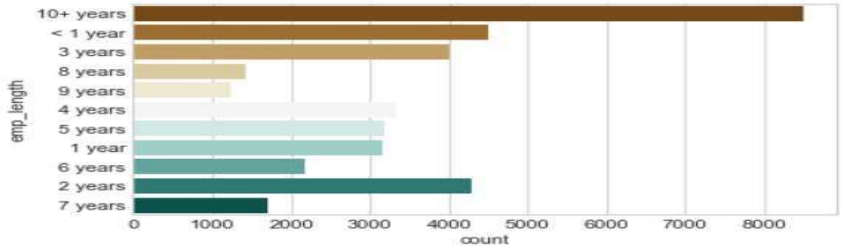
Univariate Analysis

- **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

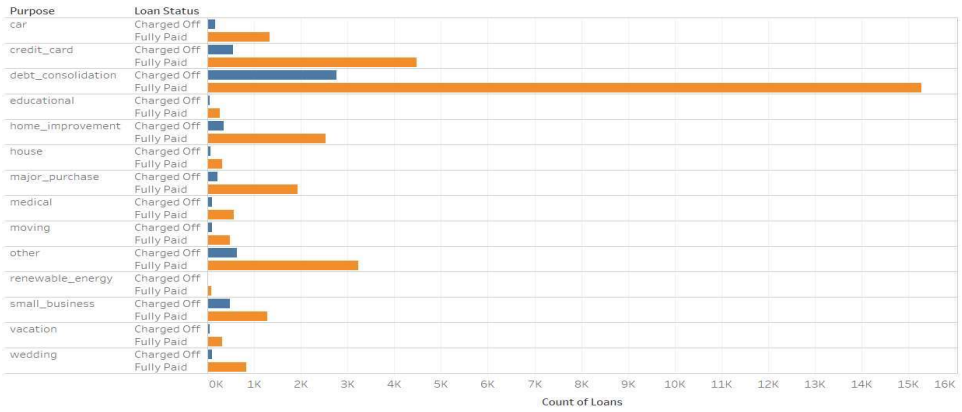


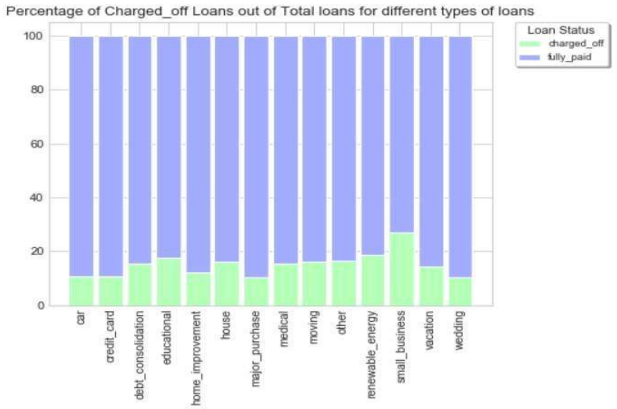


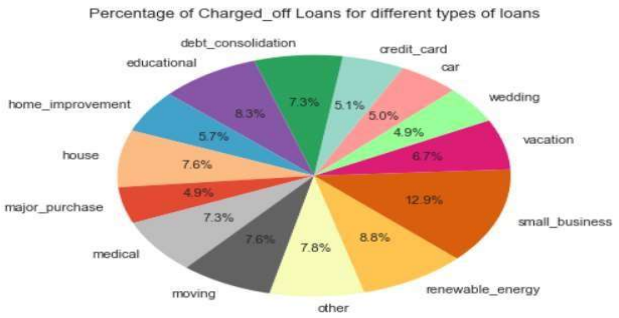
Univariate Segmented Analysis

Loan Purpose: Small business have highest % of charged off loans.

Debt_consolidation has highest no of charged off loans.

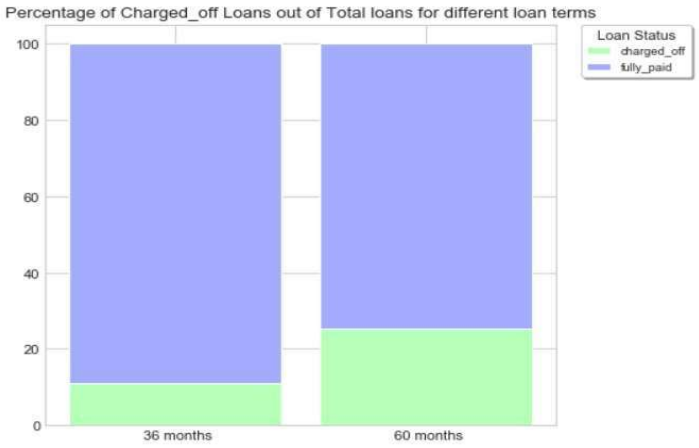


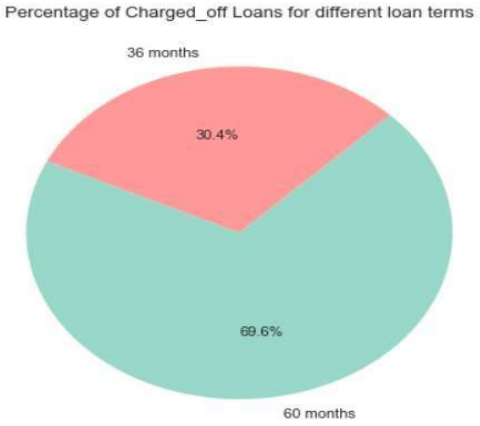


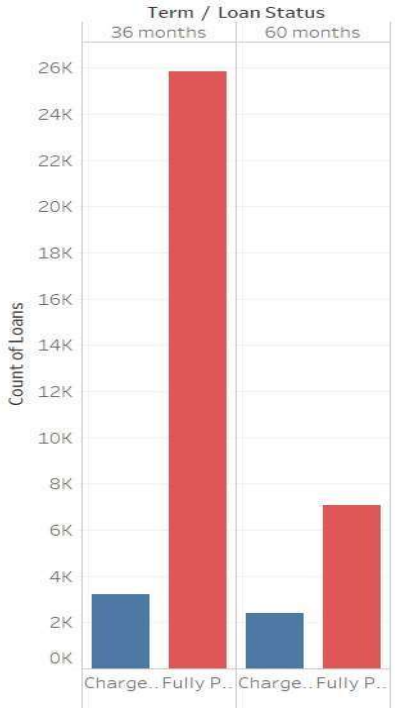


Univariate Segmented Analysis: Contd..

- 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.

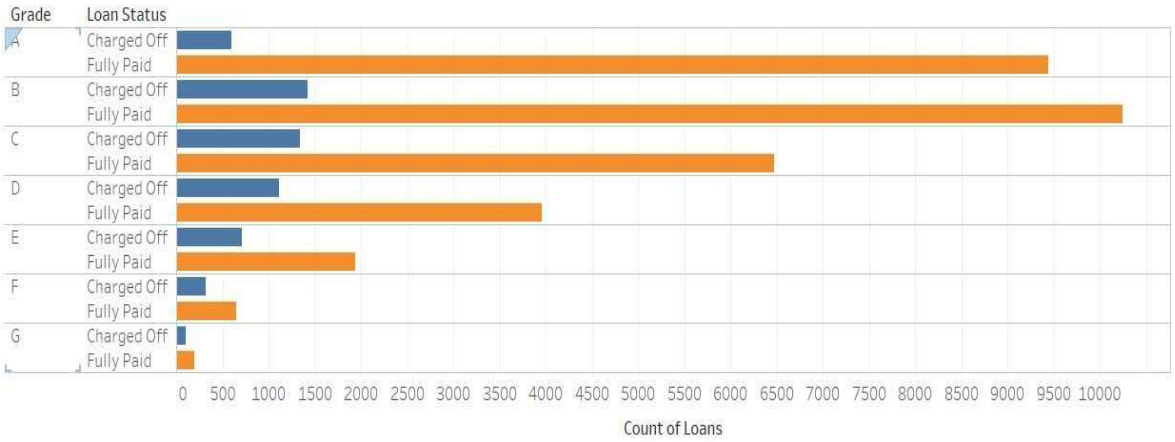


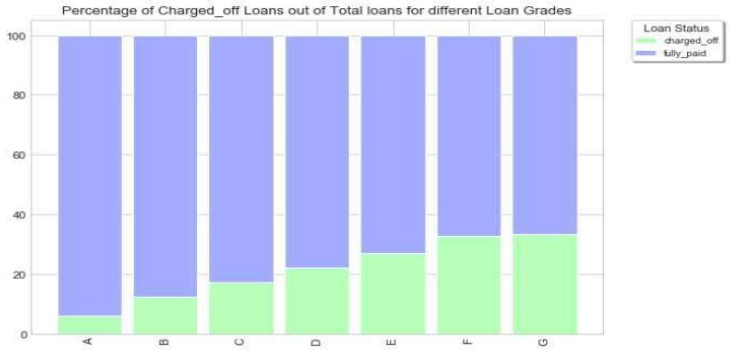


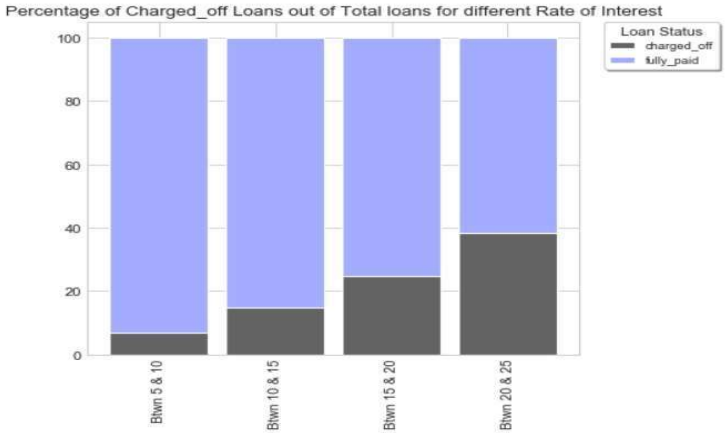


Univariate Segmented Analysis: Contd..

- 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.

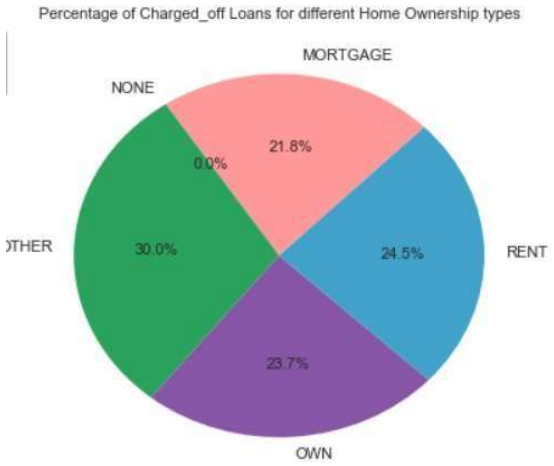






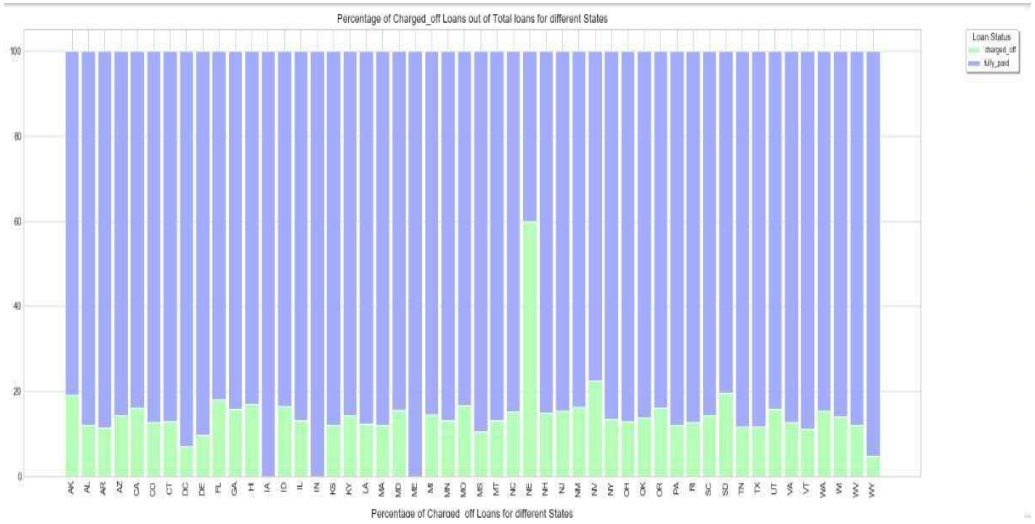
Univariate Segmented Analysis: Contd..

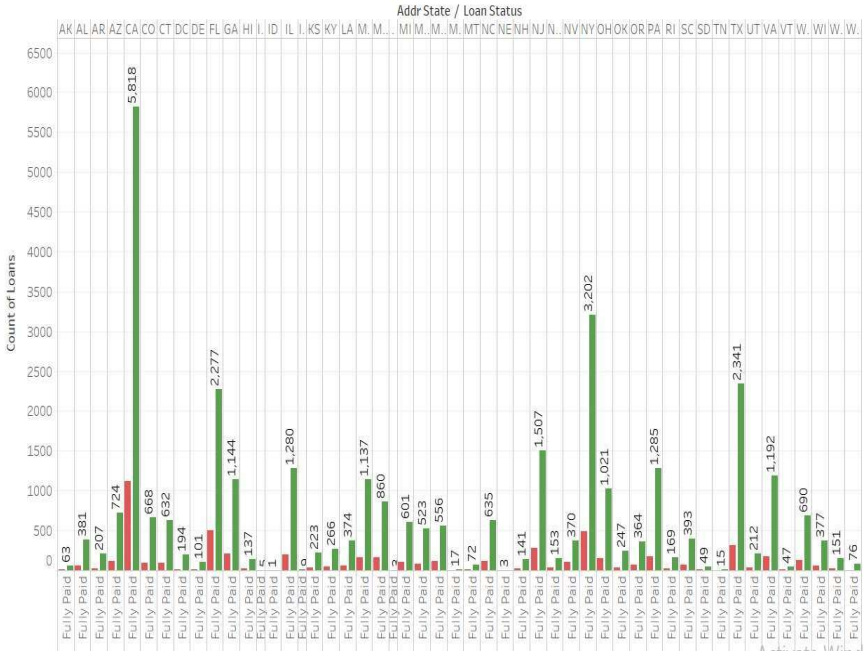
- Home Ownership: Most number of Charged-off loans happen for borrowers living in Rented homes or homes on Mortgage
- Percentage of Charged off loans is however same more or less in all ownership types



Univariate Segmented Analysis: Contd..

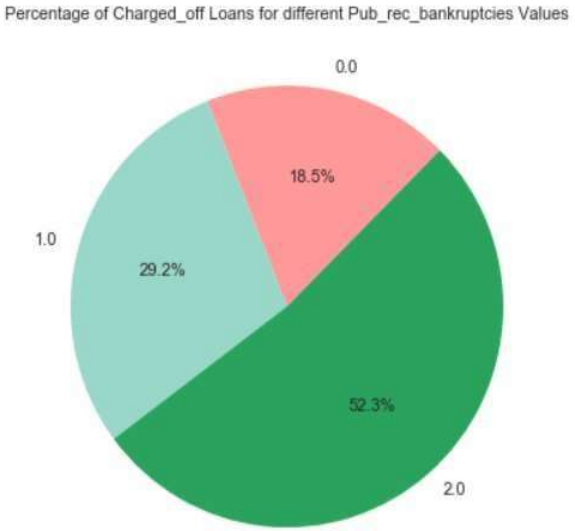
- Address State: Percentage of charged off loans is highest in the state NE .
- The count of charged off loans is most in the state CA followed by NY, FL, TX and NJ.

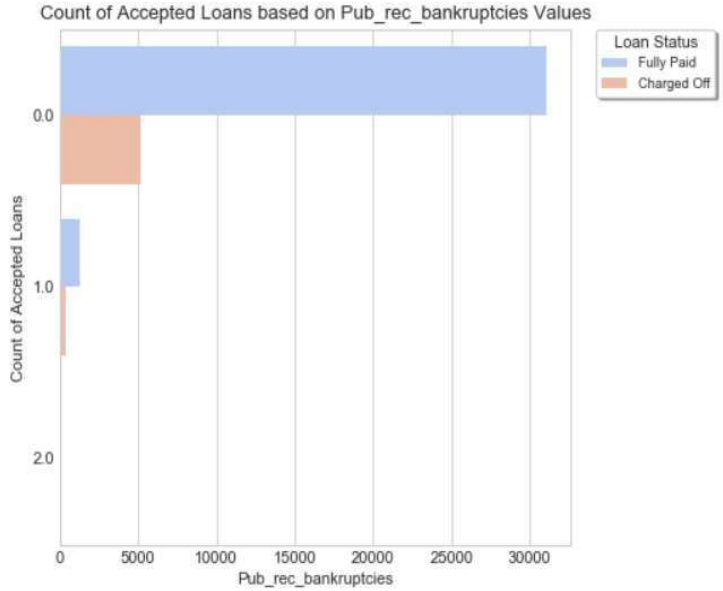




Univariate Segmented Analysis: Contd..

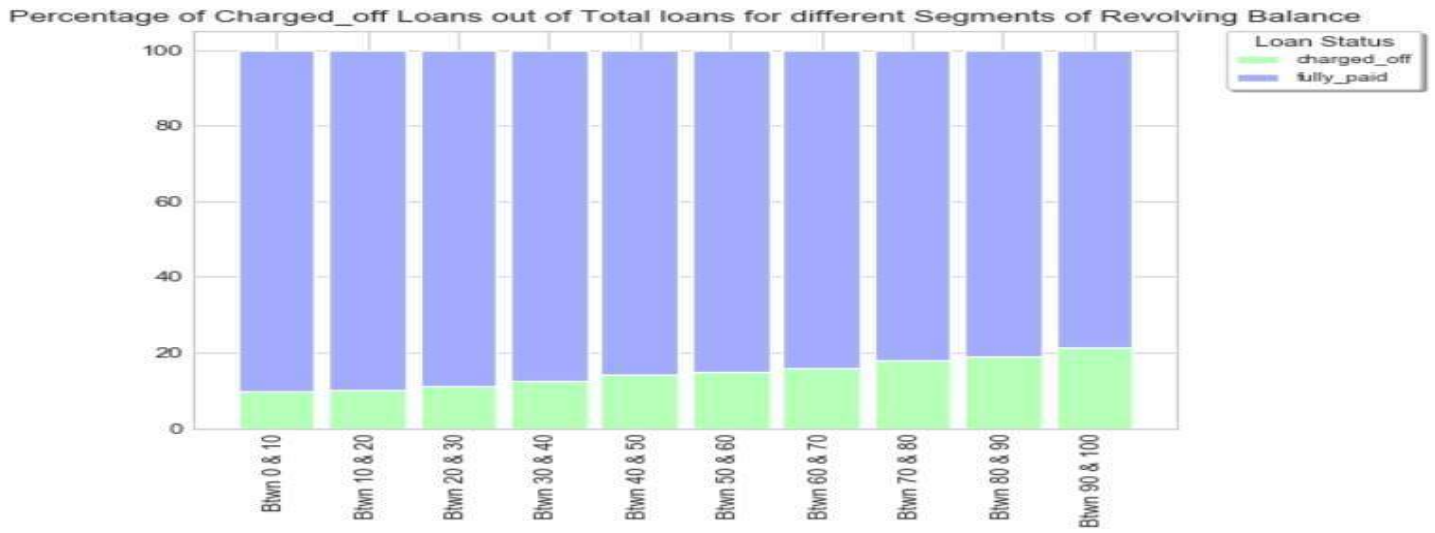
- Public Record Bankruptcies: Percentage of charged off loans gradually increase with increase in the value of Pub_rec_bankruptcies.
- However, there are very few records with bankruptcy value = 2 or 1. Most of the charged off records have '0' bankruptcies.





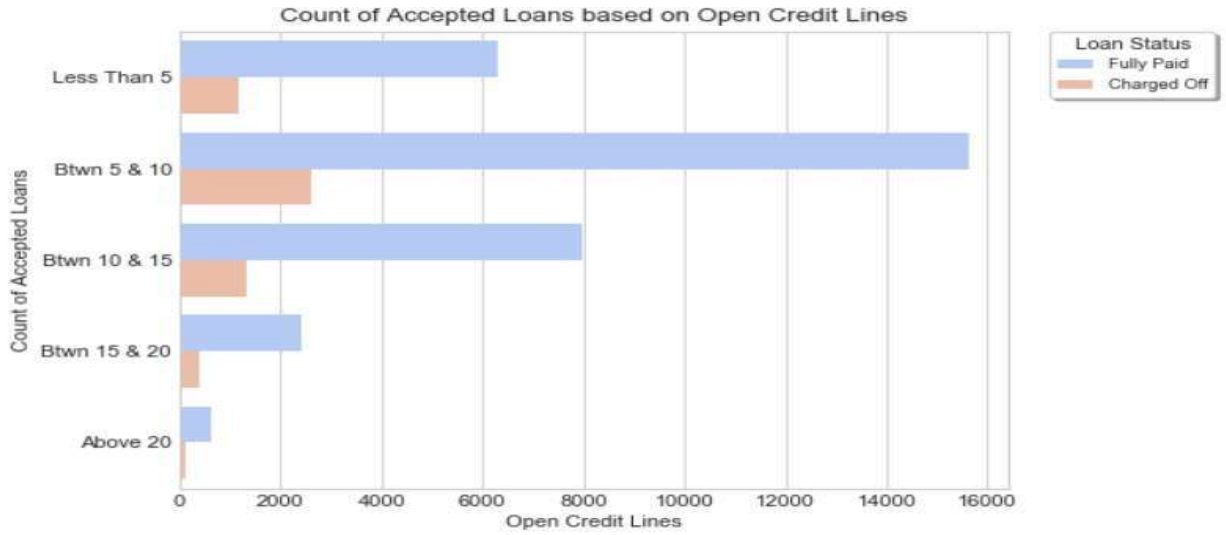
Univariate Segmented Analysis: Contd..

- Revolving Credit Utilization: Percentage of charged off loans gradually increase with increase in the value of revol_util.



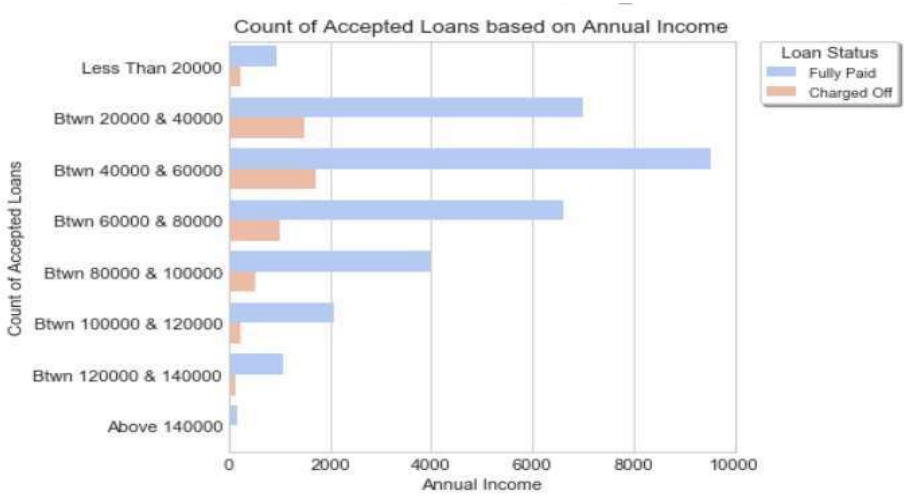
Univariate Segmented Analysis: Contd..

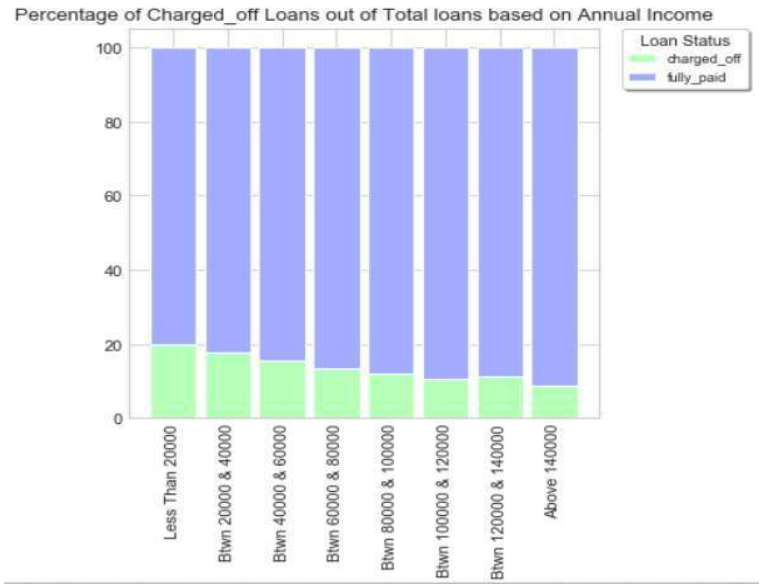
- Open Credit Line: Count of charged off loans is highest when no of Open Credit Lines is between 5 & 10.



Univariate Segmented Analysis: Contd..

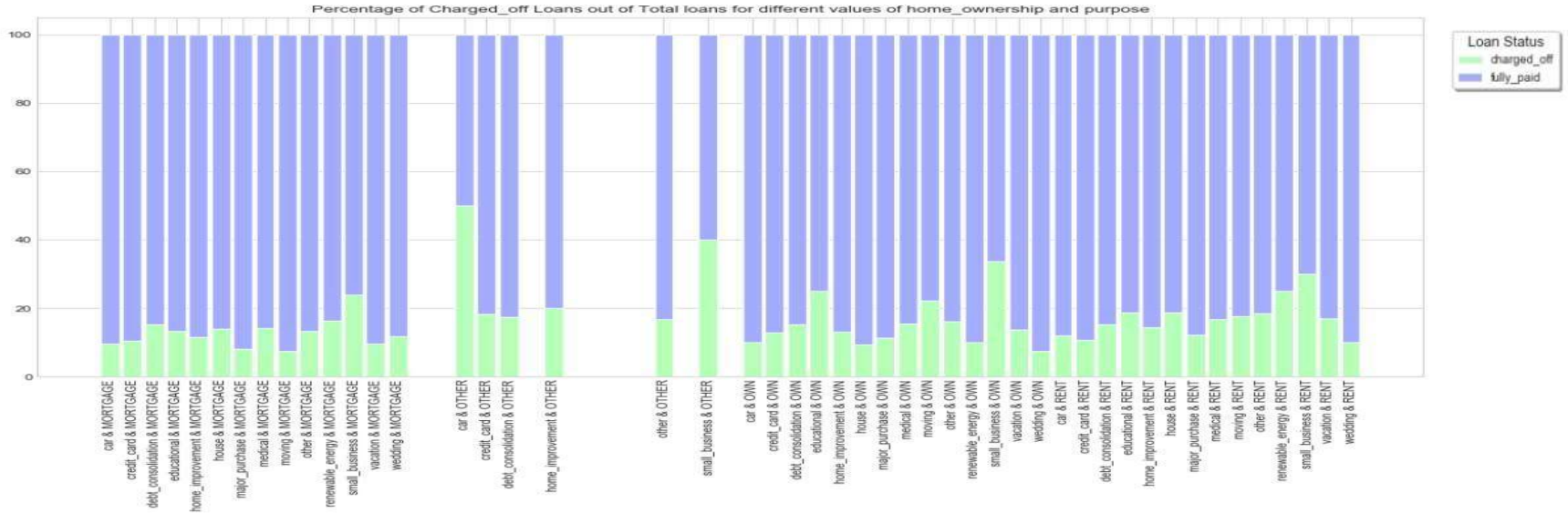
- Annual Income: The percentage of Charged off loans decreases more or less with the increase in annual income of the borrower.
- Most no of charged off loans can be seen for borrowers with annual income between 40K-60K followed by 20K-40K





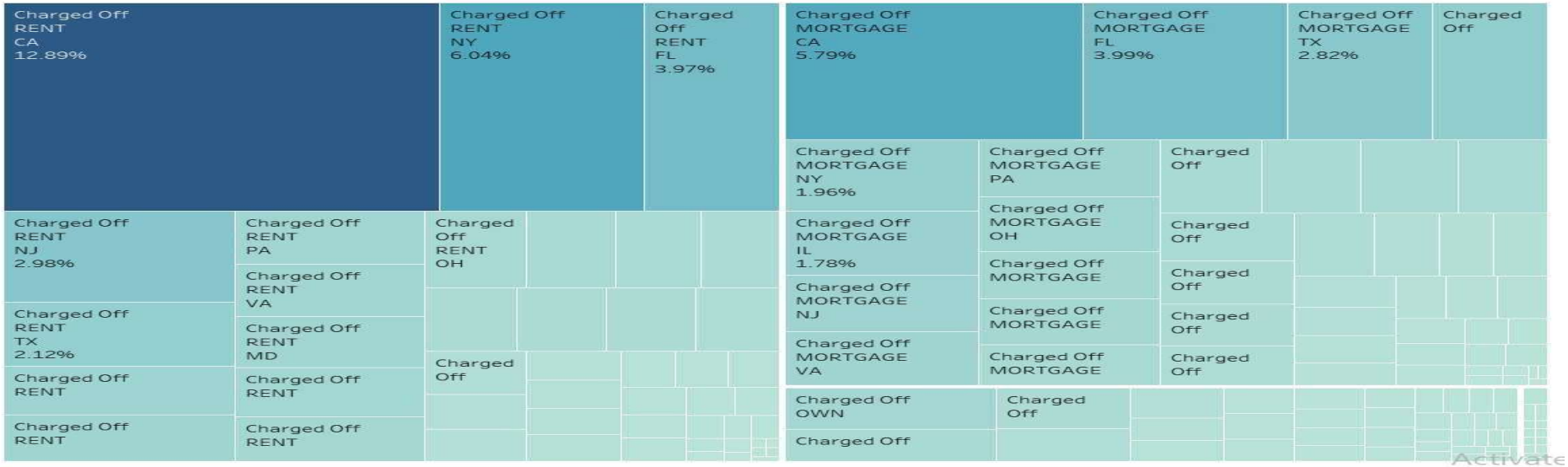
Bivariate & Multivariate Analysis:

- 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'



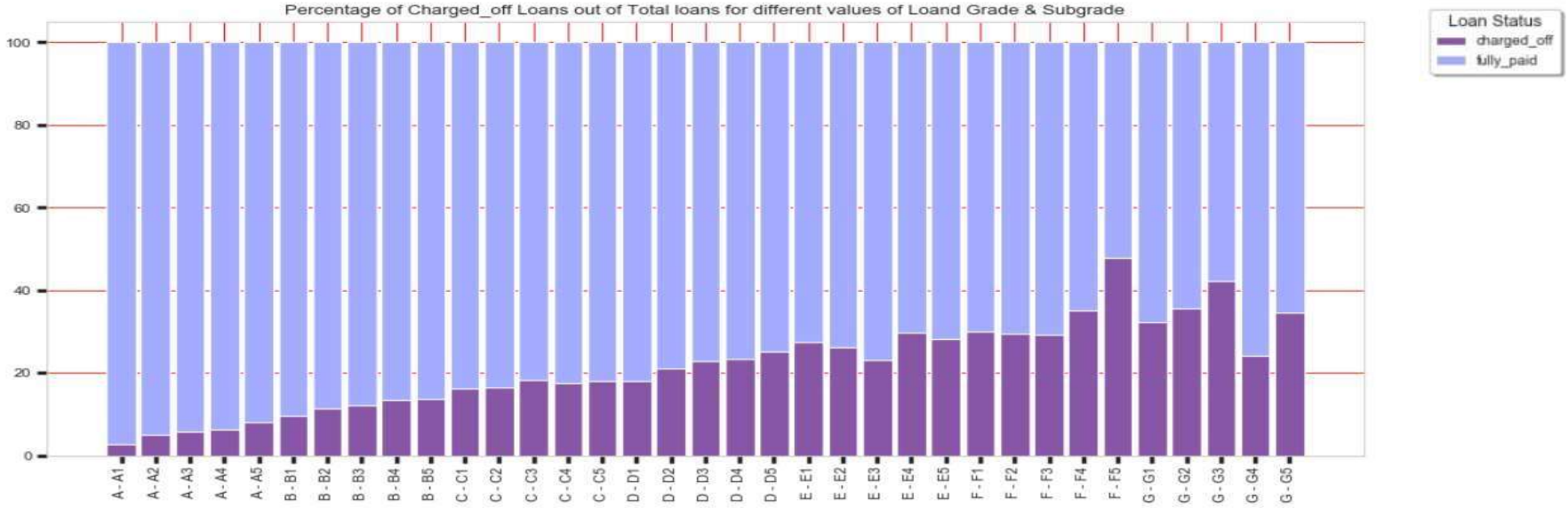
Bivariate & Multivariate Analysis:

- 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'



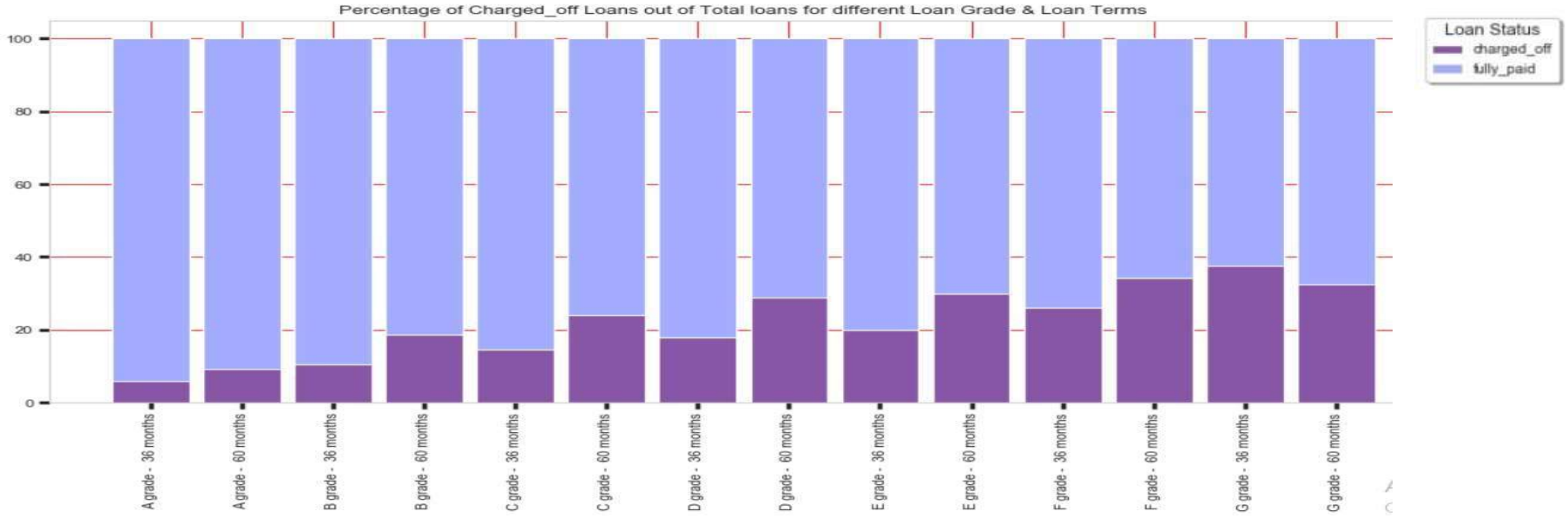
Bivariate & Multivariate Analysis:

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



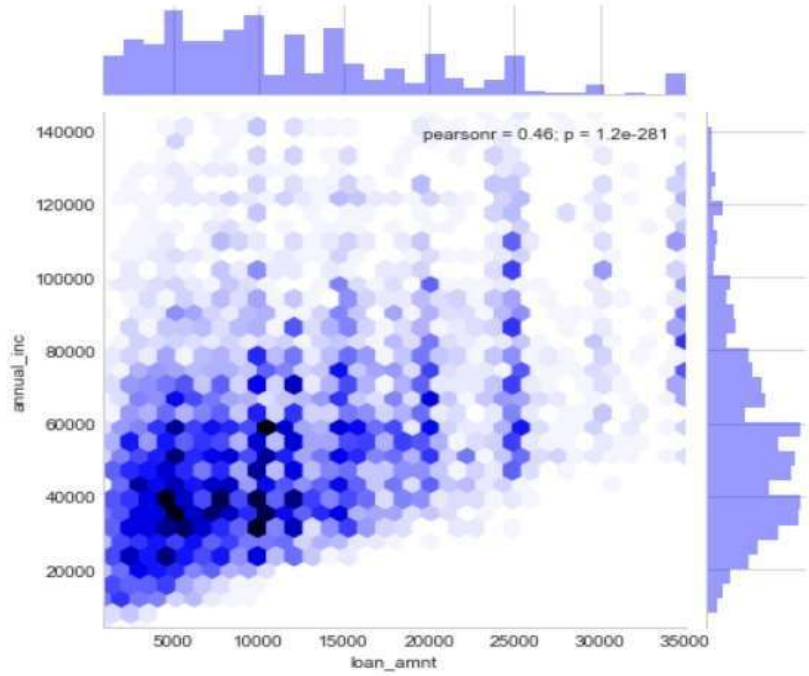
Bivariate & Multivariate Analysis:

- 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.



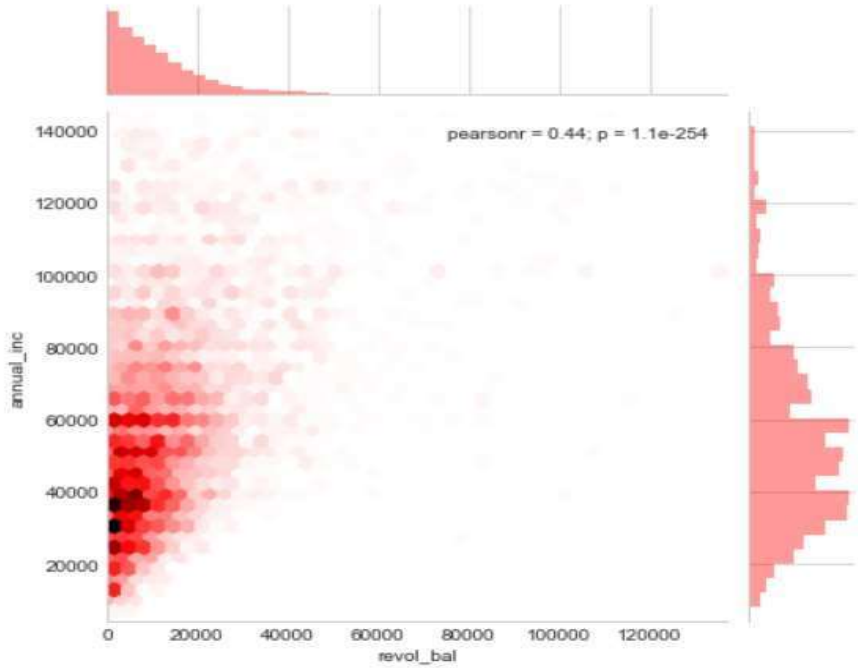
Bivariate & Multivariate Analysis:

- 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.
- Five Most Important driver variables.
- 1)Loan purpose.
- 2) Loan Grade.
- 3) Loan Term.
- 4)Address State
- 5)Revolving Utilization %.

Three Most Important Driver Variable Combination.

Annual Income & Loan Amount

Home Ownership in Various States

Loan Grade & Loan Term

IF all thing have problem then Loan Defaulted.

Recommendation

- 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.

Recommendation

- 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 sanctions 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.

Thank You