GRAMENER CASE STUDY

SUBMISSION

Names:

1. Yuvraj Shinde
2. Summit Sethi
3. Mahesh Sawant
4. Darshan Vora

Consumer Finance Case Study

Introduction

- A consumer finance company which specializes in lending various types of loans to urban customers like personal loans, business loans, and financing of medical procedures.
- The lending company is the largest online loan marketplace.

Business Goals

- The company receives a loan application and has to make a decision for loan approval or disapproval based on the applicant's profile.
- Borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.
- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Risks

- The company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the company's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Data Analysis

- Analysis to identify different variables for loan default applicants.
- Analyze using univariate , bivariate techniques using visualizations to identify patterns of loan default.
- Data is available in loan file, for all loans issued through the time period 2007 to 2011.

Data Analysis Flowchart

Part 1: Collect the data and create the dataset for loan, ensure correct encoding. Inspect the size of dataset. Refer data dictionary and identify loan status and its value as the Target variable.

Part 2: Data cleaning– Remove unnecessary columns, handle missing value in rows and columns, filter duplicate data.

Part 3: Derived metrics – Create bins for numerical values like annual income, installment , dti, interest rate, loan amount .

Part 4: Univariate analysis on categorical data. Look at the count frequency to see number of times the event occurs for variables . Plot the count plot using library seaborn.

Part 5: Univariate analysis on quantitative variables. Look at the distribution skewness of the numbers. Plot the box plot to know the spread of the data and outliers.

Part 6: Bivariate analysis on loan status variable. Eg. Loan Status by purpose to know the count of Charged Off by purpose and so on .
Part 7: Bivariate analysis on continous variables. Correlation of numeric columns using Heat map.

Part 8: Multivariate analysis on categorical vs. continous variables. Eg. Installment groups vs. Loan Amount by Loan status

Part 9: Data driven metrics Charged Off % . To know the charged off% trend and strong indicators within each variable to help loan company to reduce credit loss.

Part 1 & 2: Data Collection and Cleaning

- **Data Collection and Inspection:**
  - Import files loan and create the loan dataset in python using the pandas library.
  - We see the encoding is utf-8 using the chardet library so we will continue to use it for our analysis.
  - In the loan dataset we see there are 39717 observations and 111 variables.
- **Data Cleaning:**
- **1. Fixing and filtering columns:**
  - Removing columns with missing values: After removing columns with all missing values we are left with 54 columns.
  - Removing columns with single/constant value: There are 9 such columns. We remove them.
  - Filter columns that are not required for the analysis:
    - If we inspect the data closely columns: ID, member_id, url, desc are of not much use. So we will remove all these columns.
    - recoveries and collection_recovery_fee are post charge off gross recovery and collection fee respectively. Hence , we can keep recoveries as its gross amount and remove collection_recovery_fee column.
    - total_pymnt_inv and total_pymnt is the payments received in portion and total payment received respectively. So we can keep the total_pymnt column for our analysis further and remove the column total_pymnt_inv.
    - Also total_pymnt is sum of total_rec_prncp, total_rec_int, total_rec_late_fee and recoveries. So we can remove the 4 columns and keep total_pymt.
  - **After fixing and filtering columns ,we are remaining with 38 columns.**
- **2. Fixing and filtering rows:**
  - There are , about 4% rows have more than 3 missing values. Let's remove these rows and count the number of missing values remaining.
  - When last_pymnt_d is null which means the loan status was charged off(defaulted) and hence last payment date is missing vlaues. We will not remove the missing values for last_pymnt_d else we will also loose the Charged Off values in loan_status.
  - **After fixing and filtering rows, we have lost about 4% observations in cleaning the missing values.**

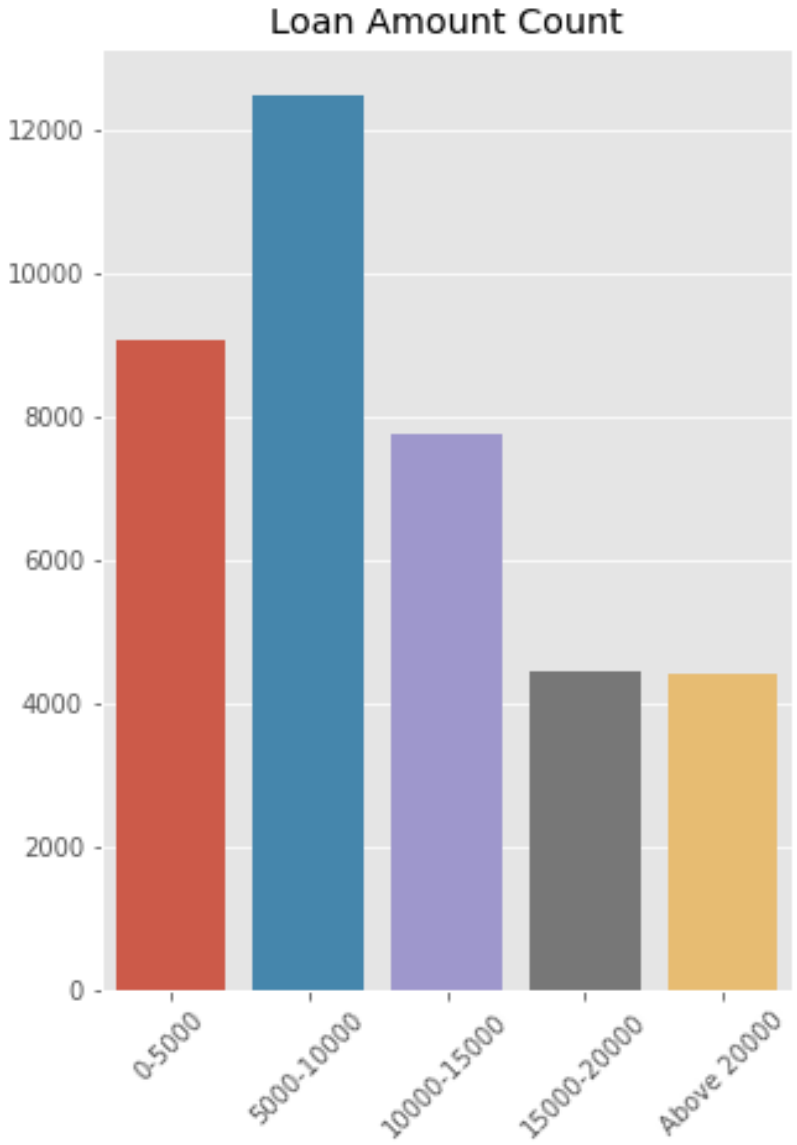
- **After data cleaning, we are left with 39717 rows and 38 columns.**

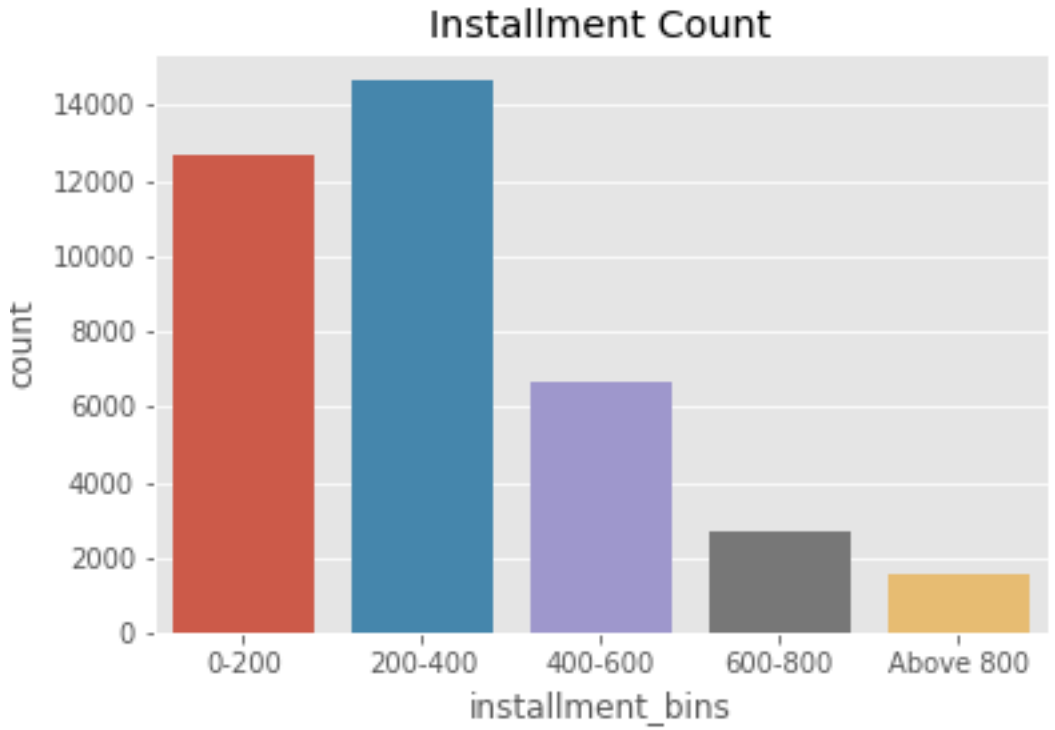
Part 3 : Type Derived Metrics

- **We created following Type Driven Derived metrics:**
  - For analyzing the numerical values created variable with bins/groups/categories for the numerical values
  - Extracted month from issue date to create a variable with issue month

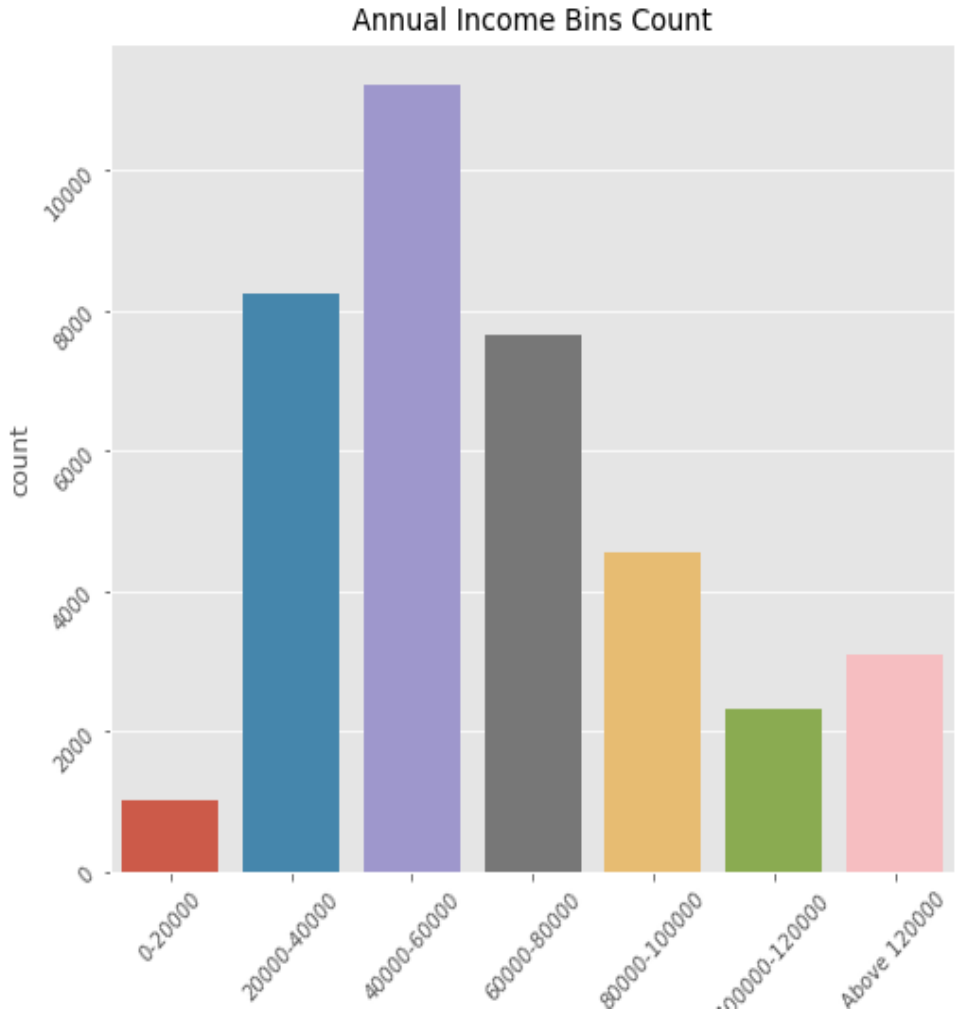
Derived Metric/Column name	Derived from Metric/Column Name	Metric description
annual_inc_bins	annual_inc	Annual Income
loan_amnt_bins	loan_amnt	Loan Amount
installment_bins	loan_amnt	Installment
mths_since_last_delinq_bins	mths_since_last_delinq	number of months since the borrower's last delinquency
mths_since_last_record_bins	mths_since_last_record	number of months since the last public record
open_acc_bins	open_acc	open_account
revol_bal_bins	revol_bal	revolving balance
revol_util_bins	revol_util	revolving util
dti_bins	dti	Dti
issue_d_month	issue_d	issue date

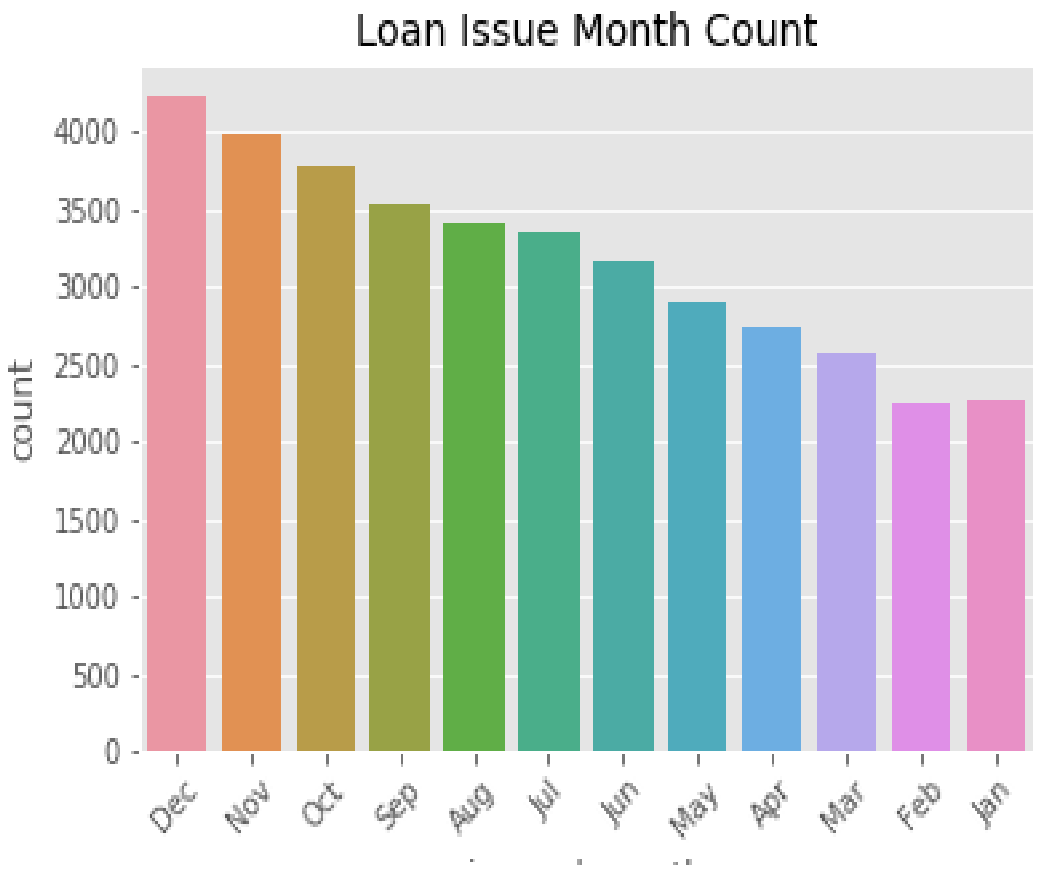
Part4: Univariate analysis on Categorical variables



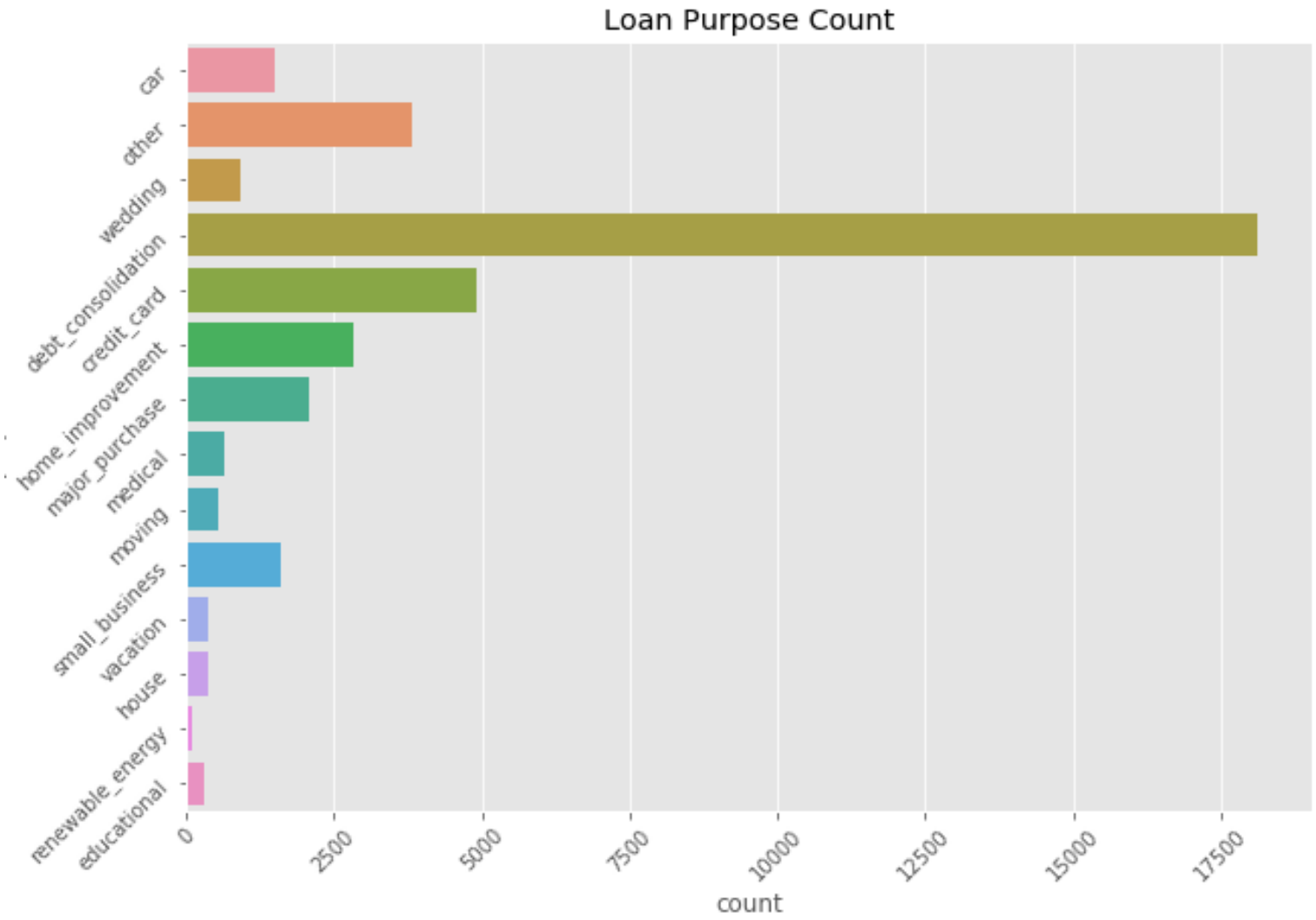


Univariate analysis on Categorical variables

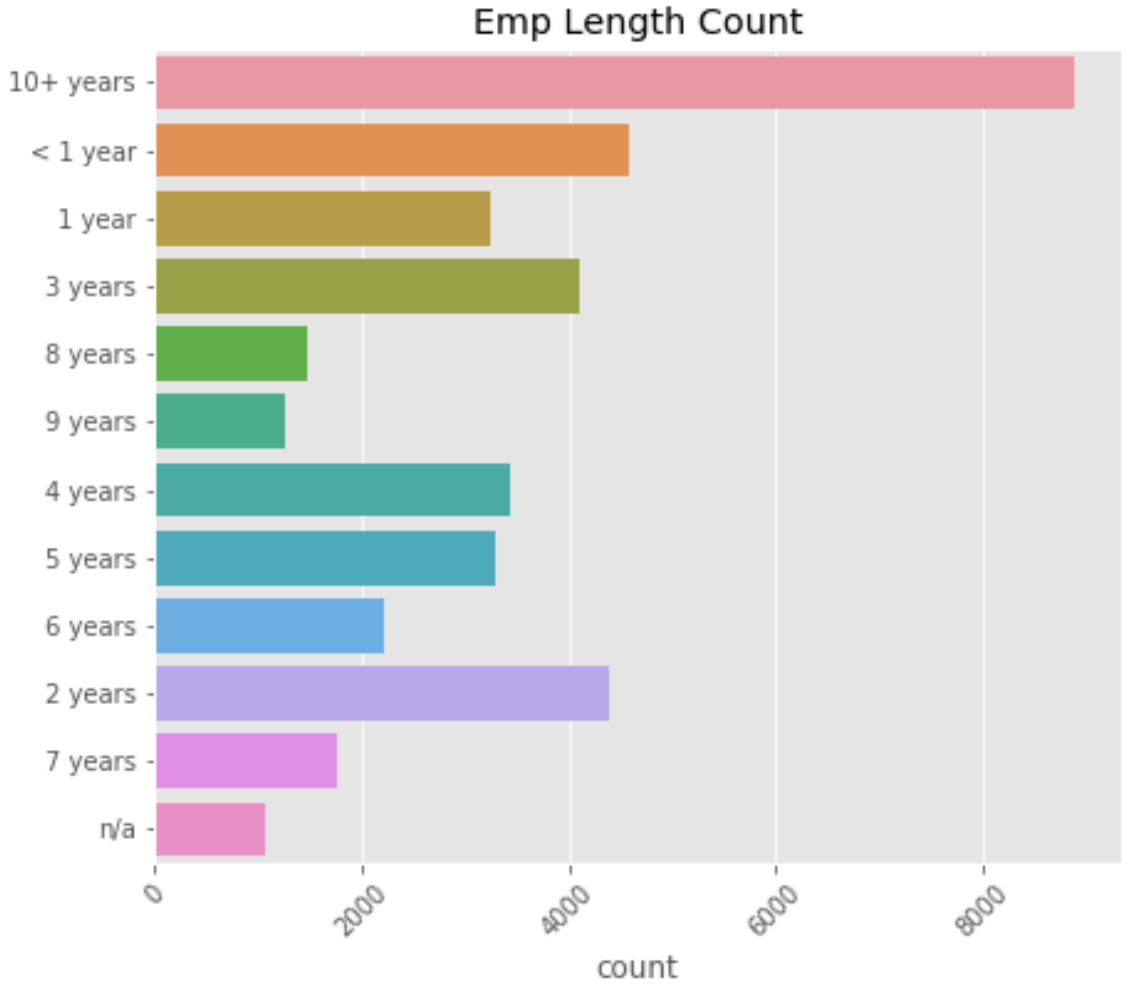


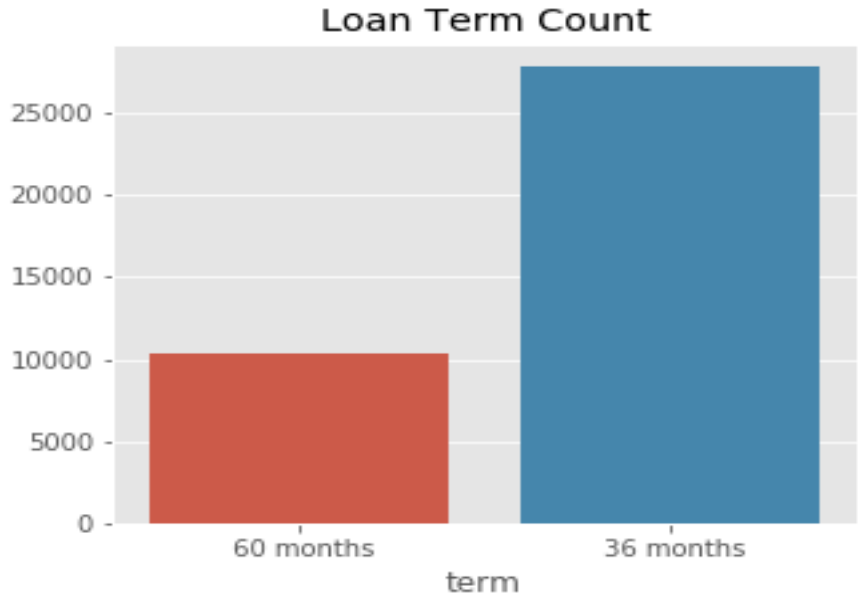


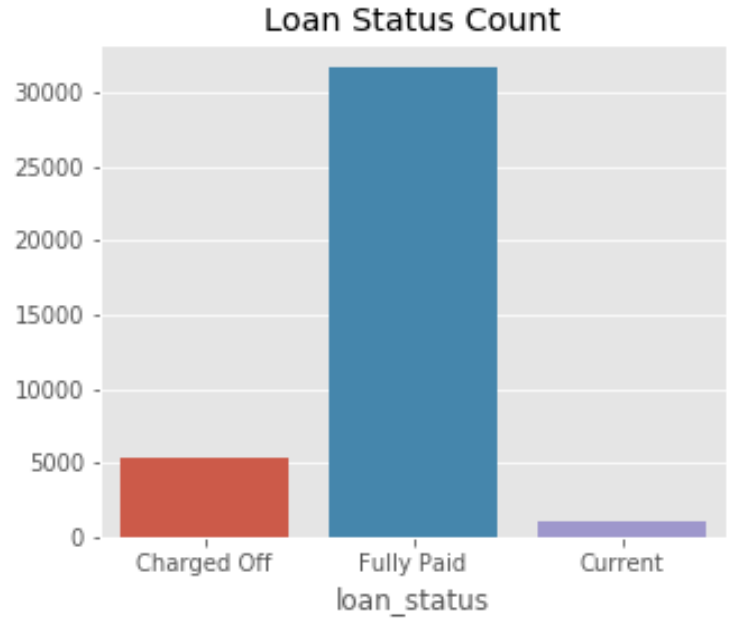
Univariate analysis on Categorical variables



Univariate analysis on Categorical variables



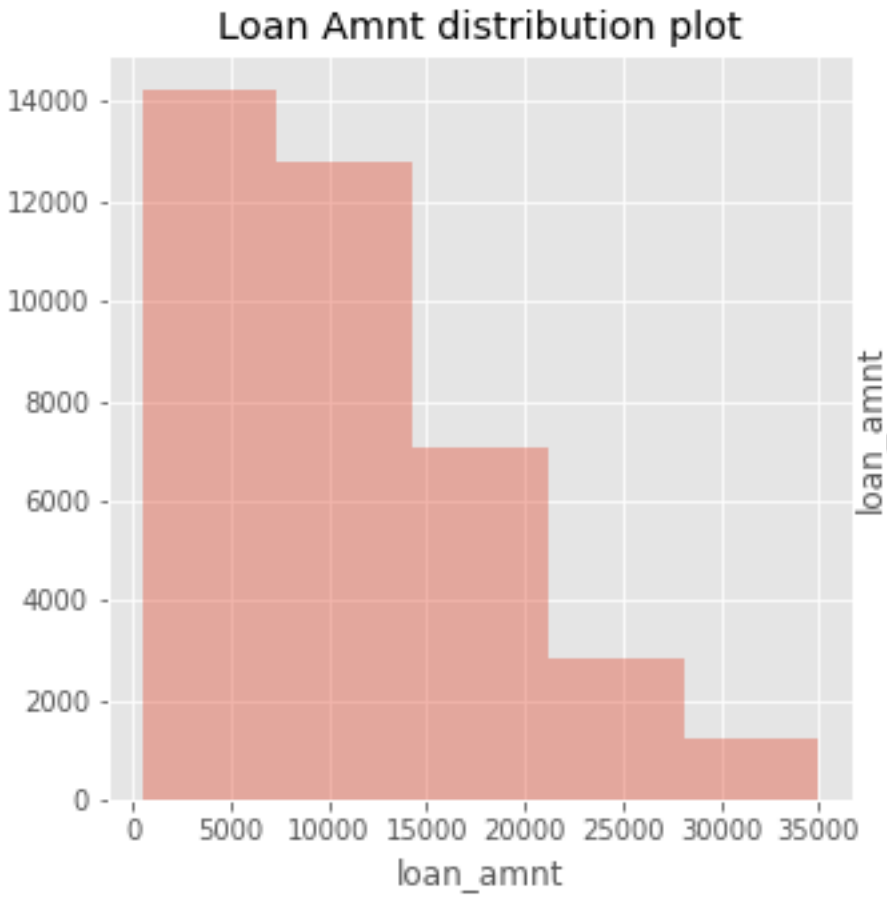


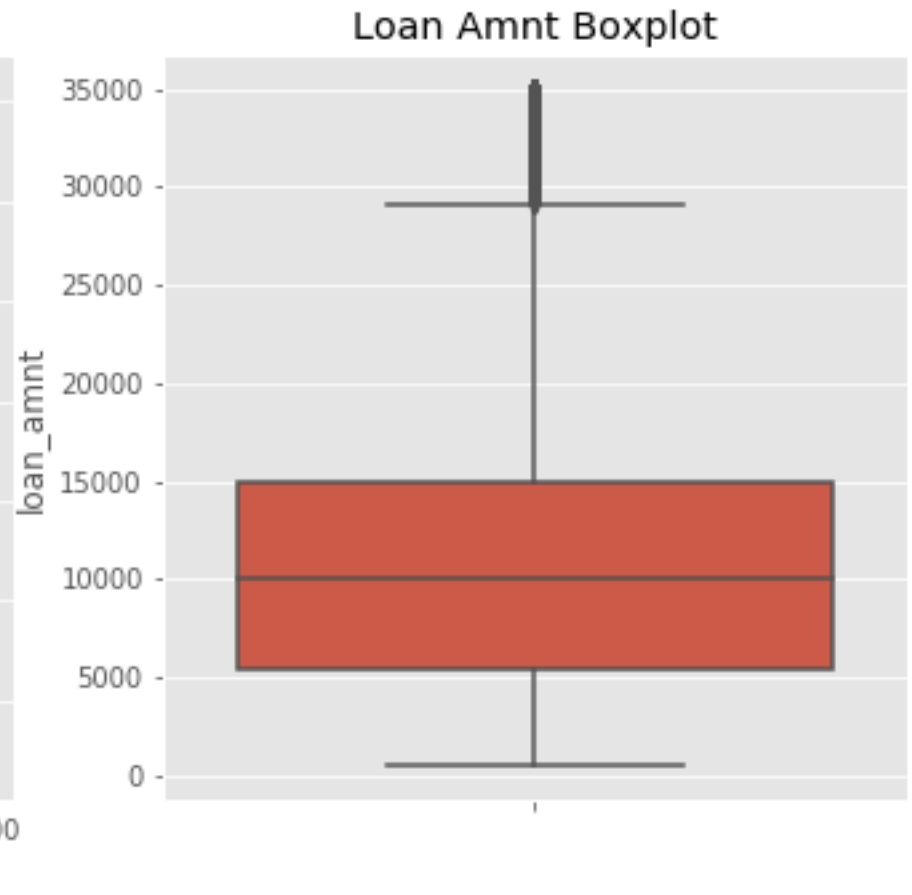


Univariate analysis on Categorical variables

Variable	Observation
Term	There are 29096 borrowers with 36 months term which is higher than 60 months.
Loan Status	There number of borrowers who defaulted are around 5627 .
Purpose	There number of borrowers are highest for debt consolidation around 18641. The number of borrowers who defaulted is highest for purpose debt_consolidation around 2767.
Emp_Length	The number of borrowers with greater than 10+ years of experience are the highest around 8879. The number of borrowers who defaulted is highest for 10+ years of experience around 1331.
Loan Amount	The number of borrowers with loan amount range 5000-10000 are the highest around 32%
Installment	The number of borrowers with Installment range 200-400 are the highest around 38%
Issue Month	Most no of loan issued towards end of year with increasing number
Annual Income	The number of borrowers with annual income range 40000-60000 are the highest around 11500.

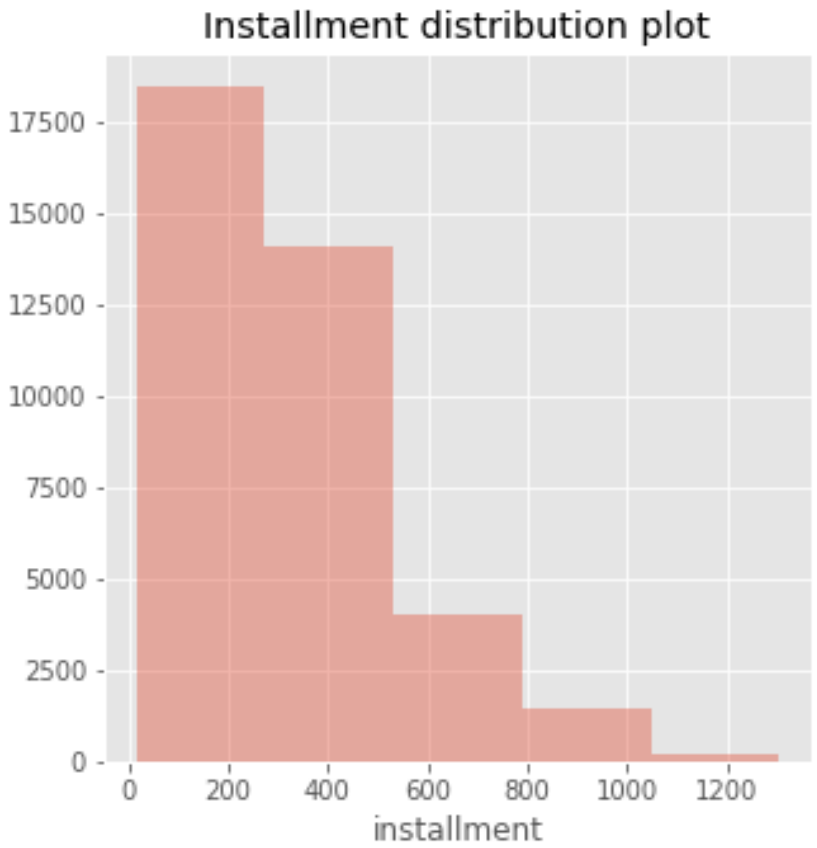
Part 5: Univariate Analysis Quantitative variables

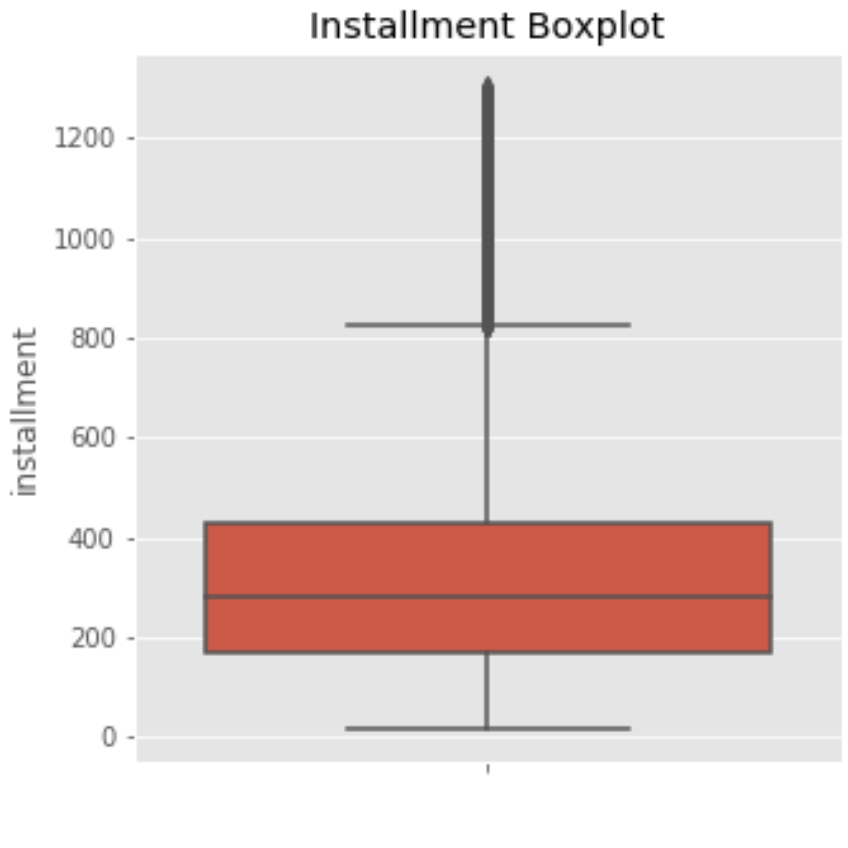




- The distribution plot shows us the skewness towards the right.
- The Loan amount spread is around 9500 (i.e. IQR, interquartile difference).
- Outliers are 3.5 * IQR.

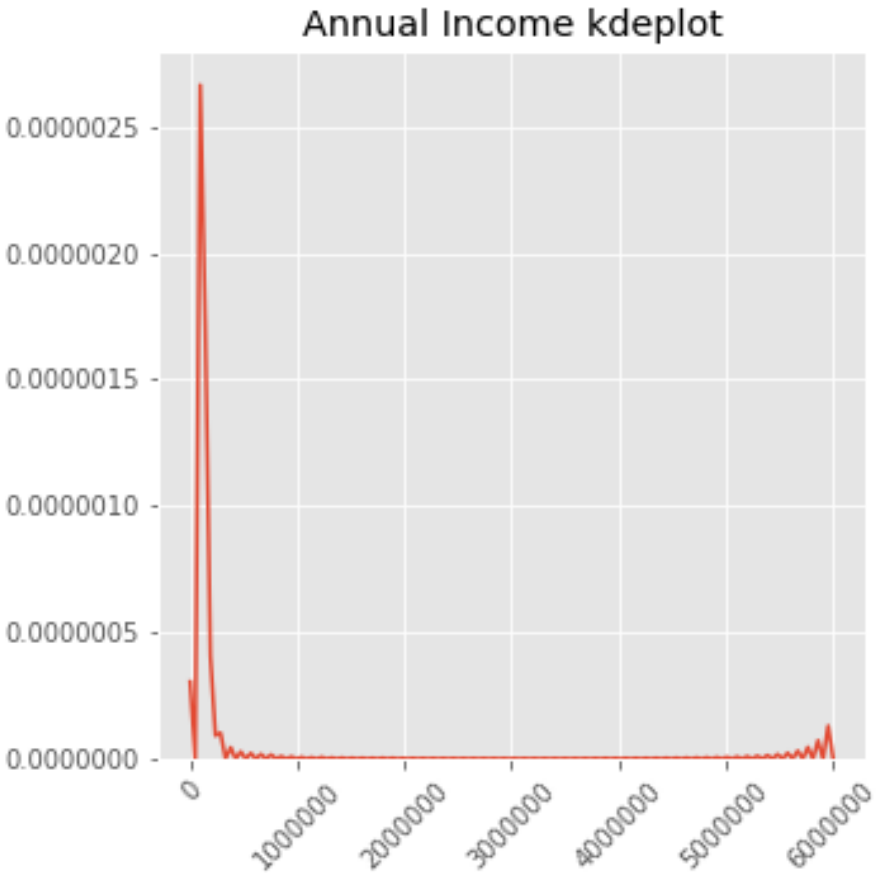
Part 5: Univariate Analysis Quantitative variables

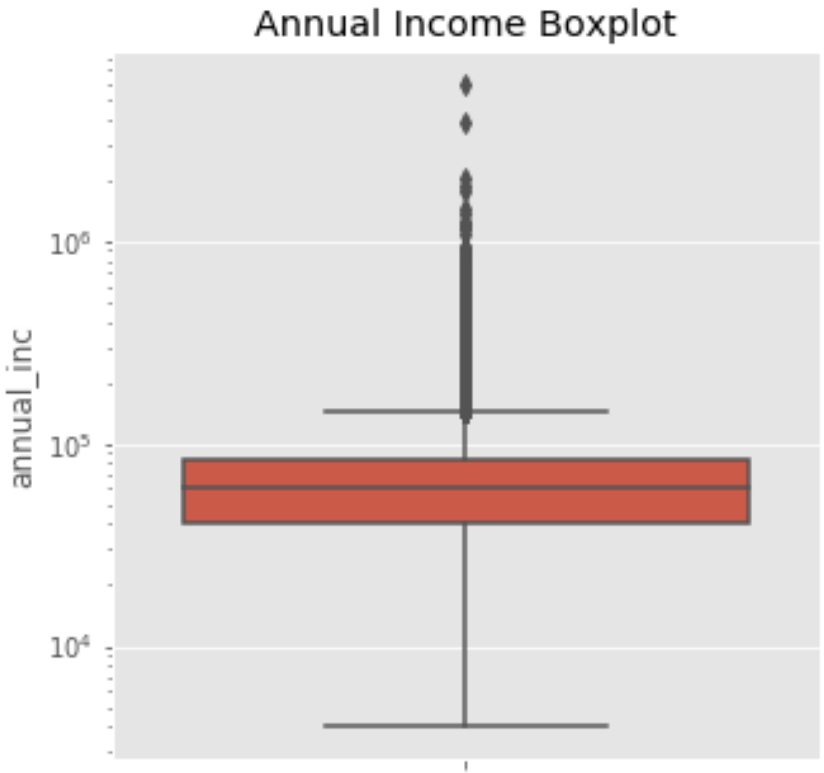




- The distribution plot shows us the skewness towards the right.
- The Installment amount spread is around 263 (i.e. IQR, interquartile difference).
- Outliers are 3.5 * IQR.

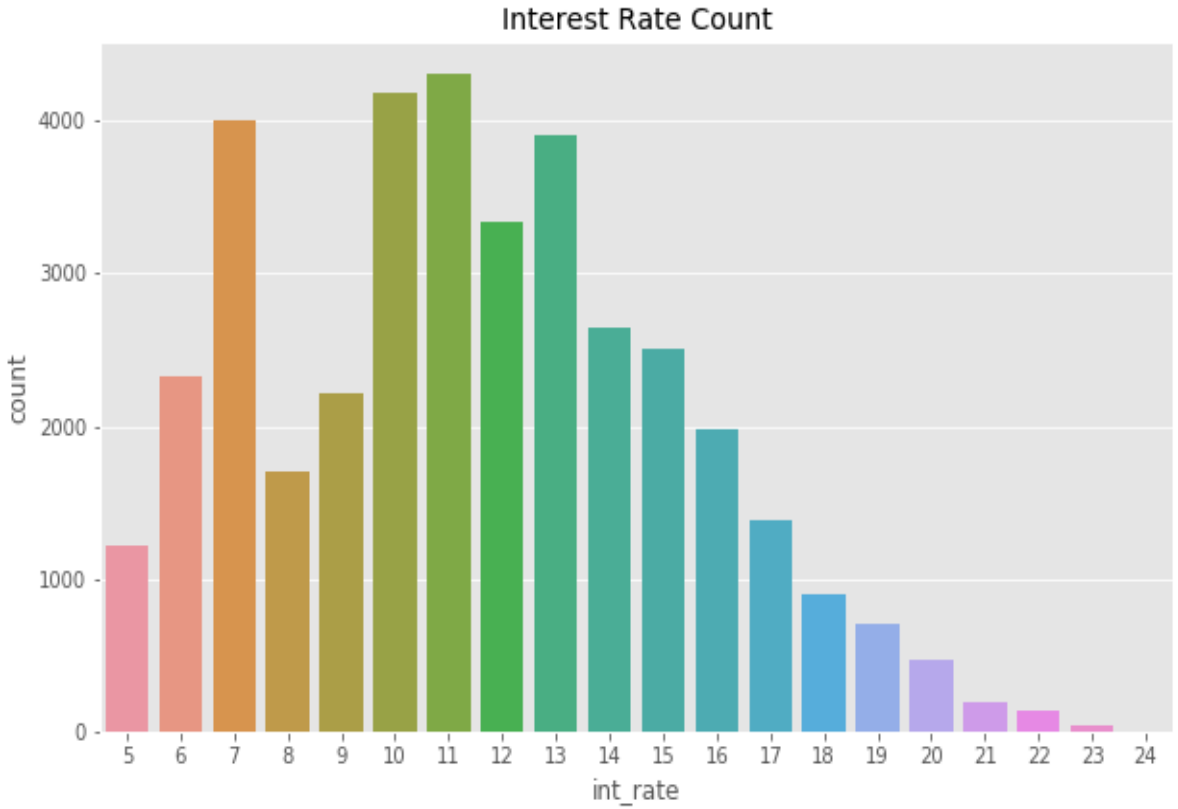
Part 5: Univariate Analysis Quantitative variables



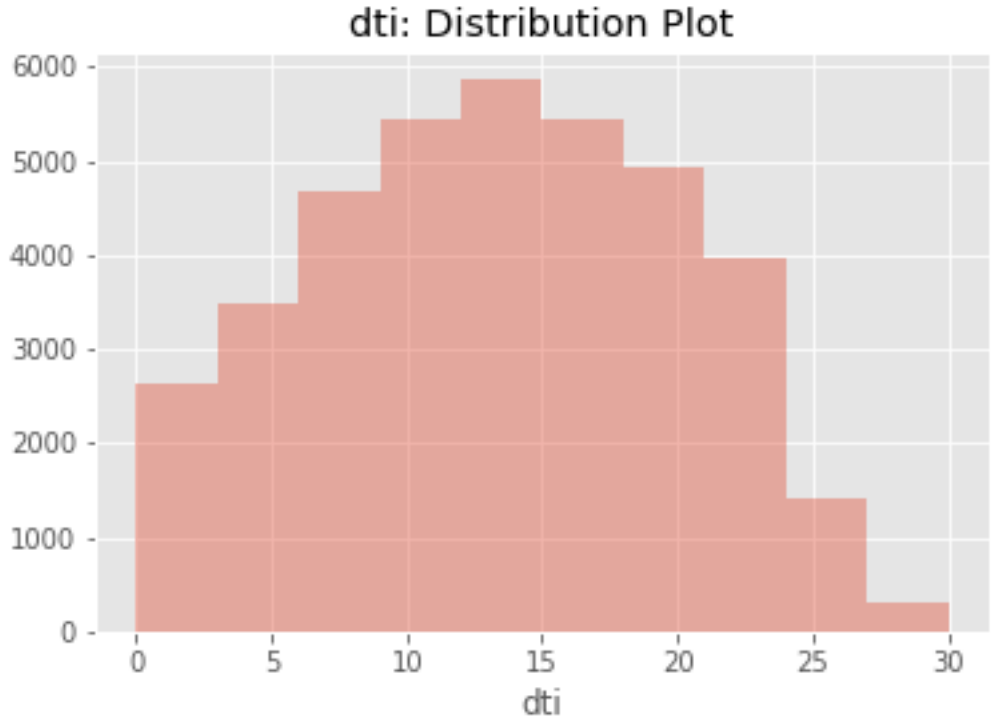


- The Annual Income spread is around 41900 (i.e. IQR, interquartile difference).
- Outliers are 6 * IQR.

Part 5: Univariate Analysis Quantitative variables

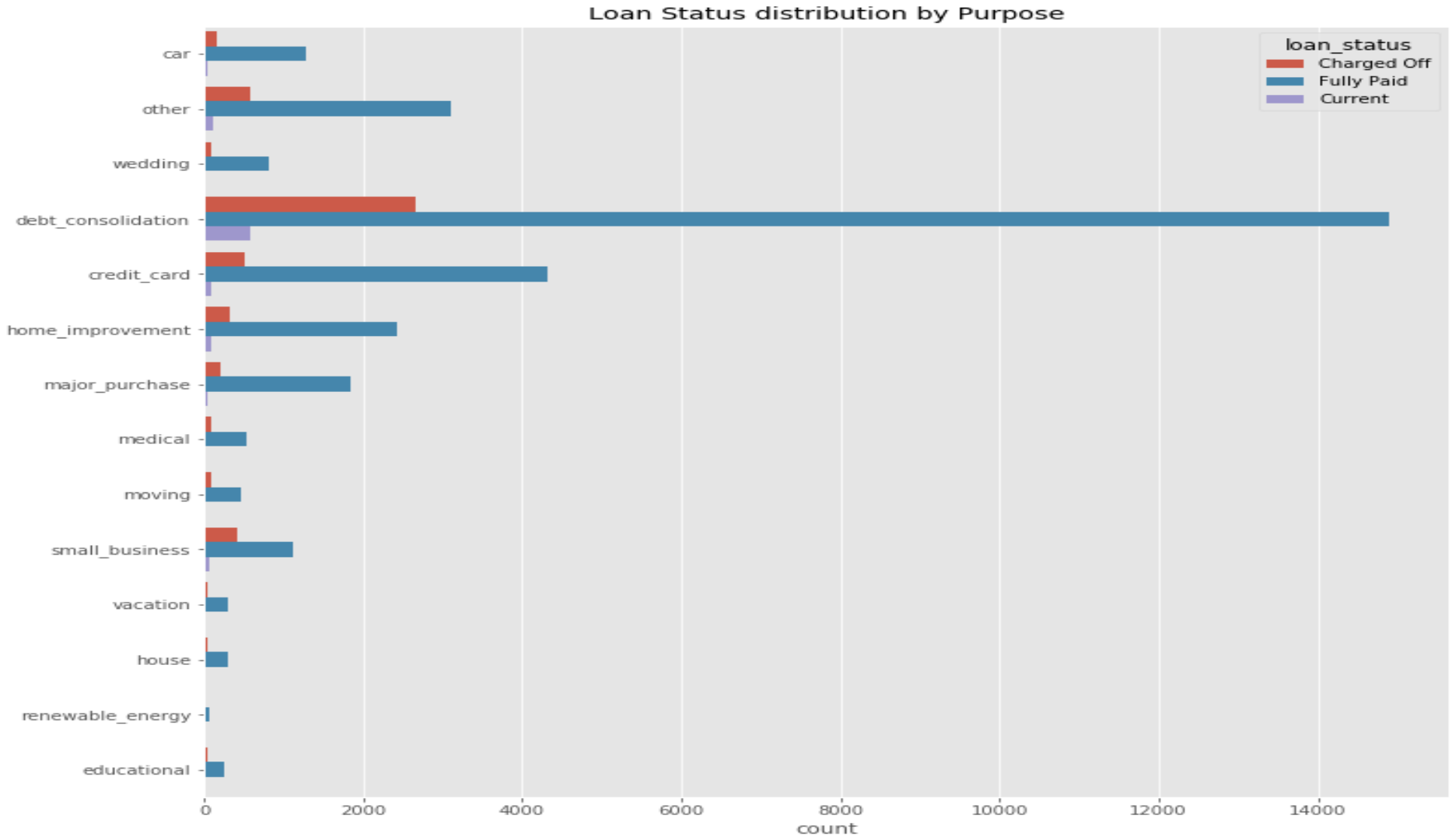


Most number of borrowers are charged 11% interest rate.



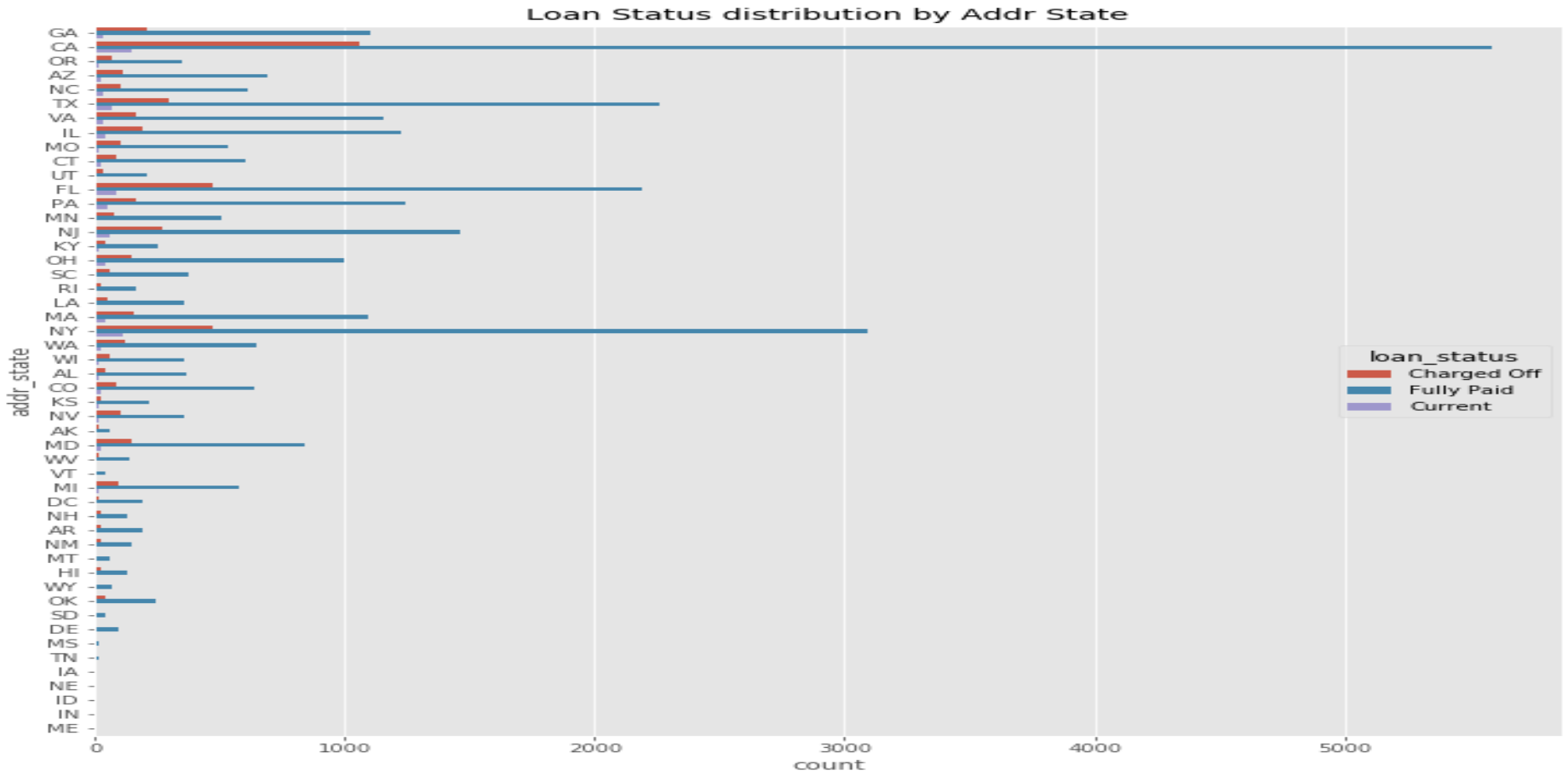
- The Mean value for dti is almost equal to the Median value around 13.
- A closer look at the distribution plot shows its symmertic in nature.

Part 6 :Bivariate Analysis on Loan Status



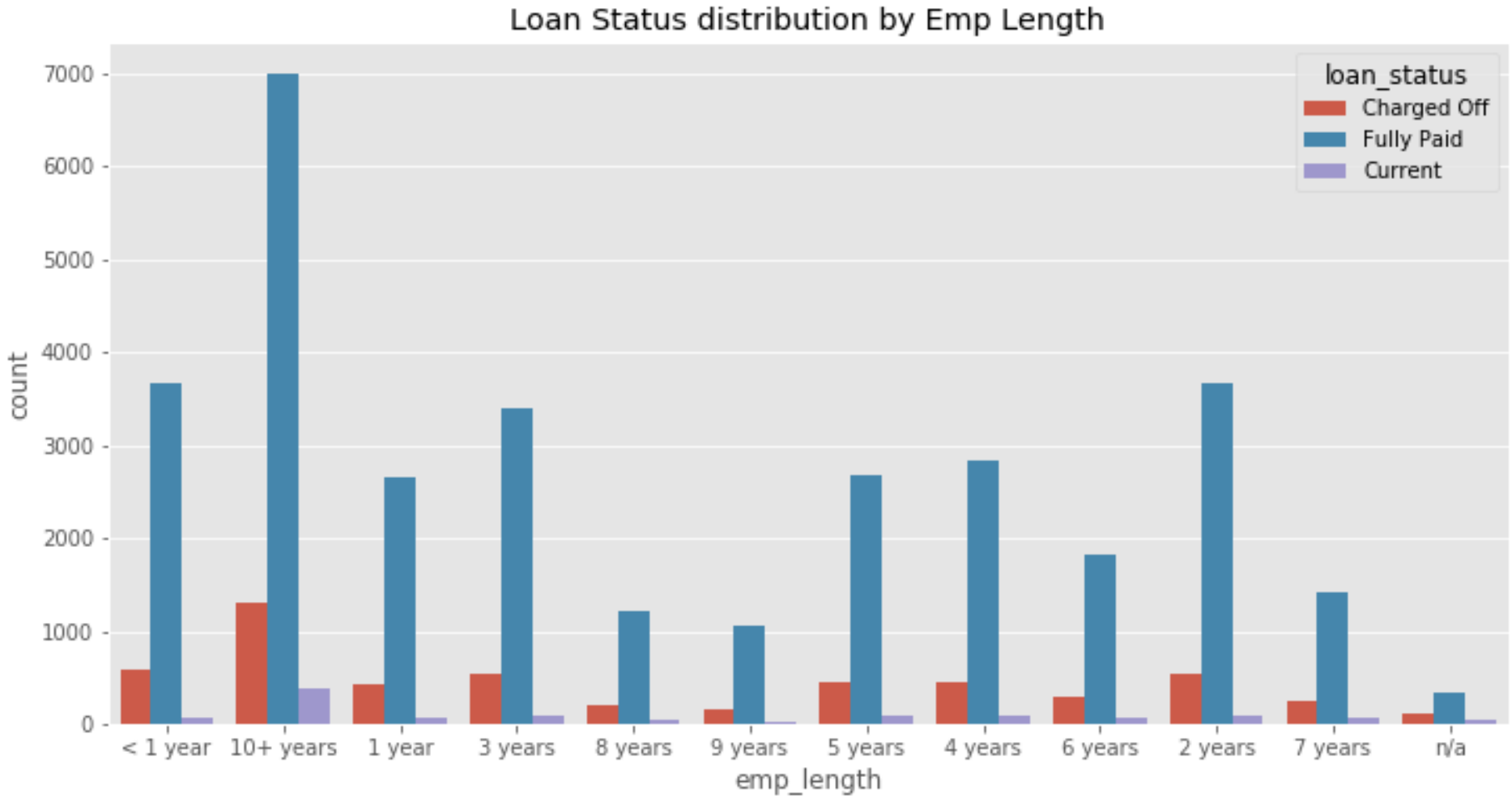
The highest count for Charged Off is for purpose debt_consolidation around 2660.

Part 6 :Bivariate Analysis on Loan Status



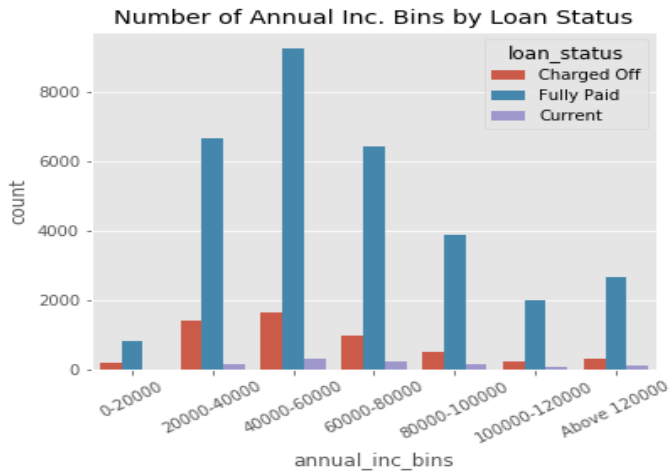
- Most borrowers are from the State CA.
- The Charged Off count in the State CA is 1058.

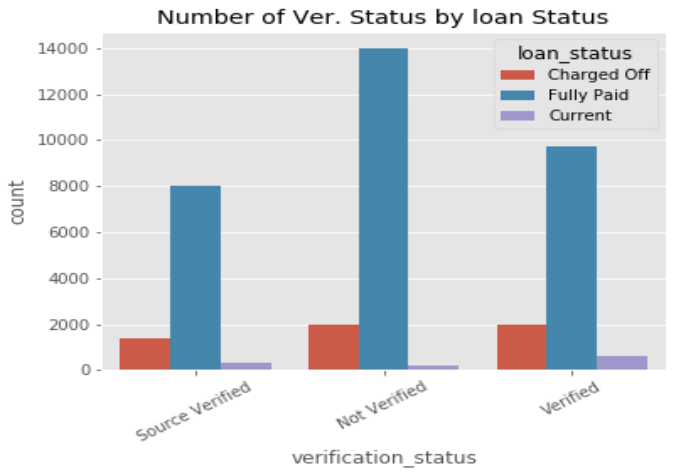
Part 6 :Bivariate Analysis on Loan Status



The number of borrowers who defaulted is highest for 10+ years of experience around 1297.

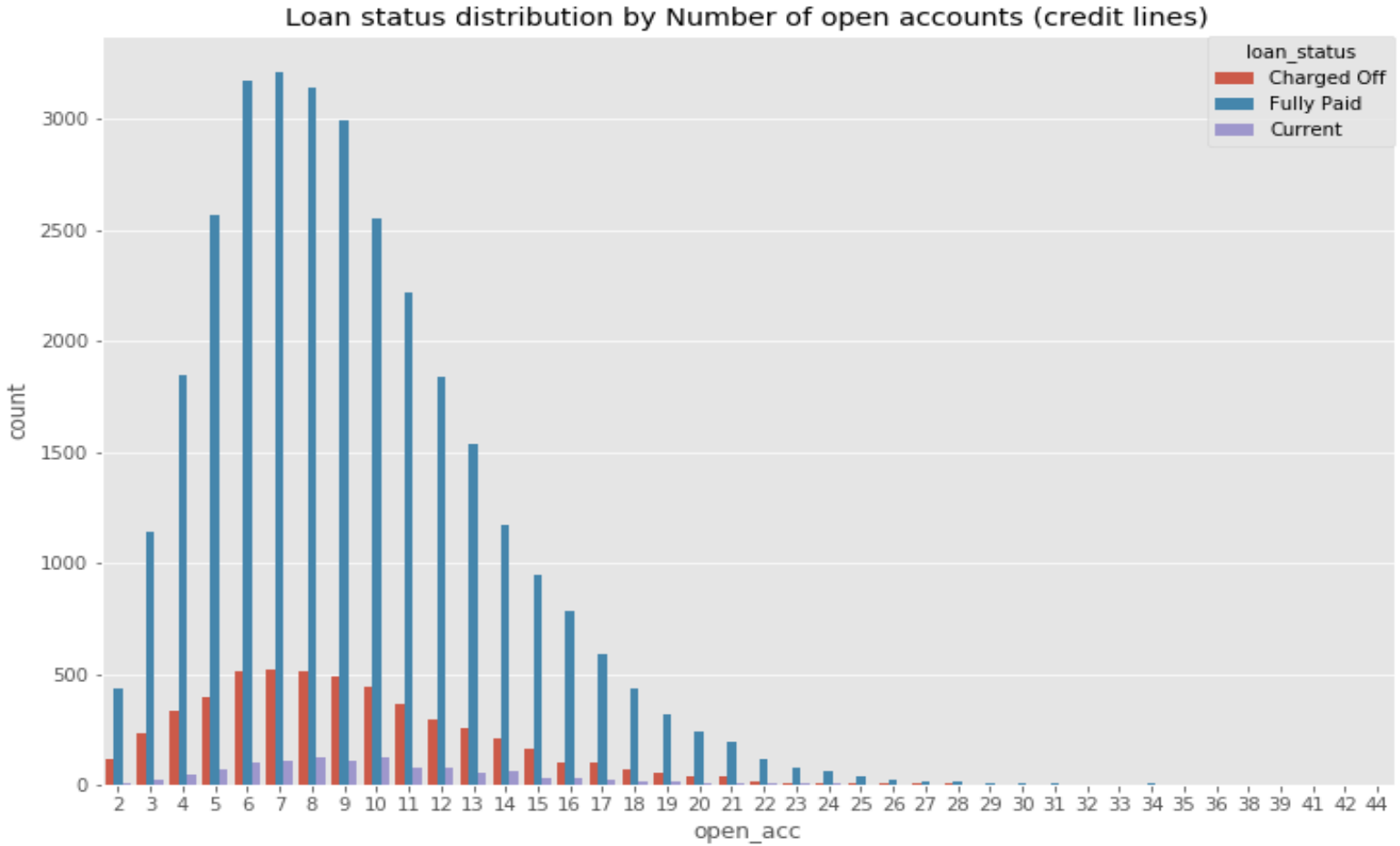
Part 6 :Bivariate Analysis on Loan Status





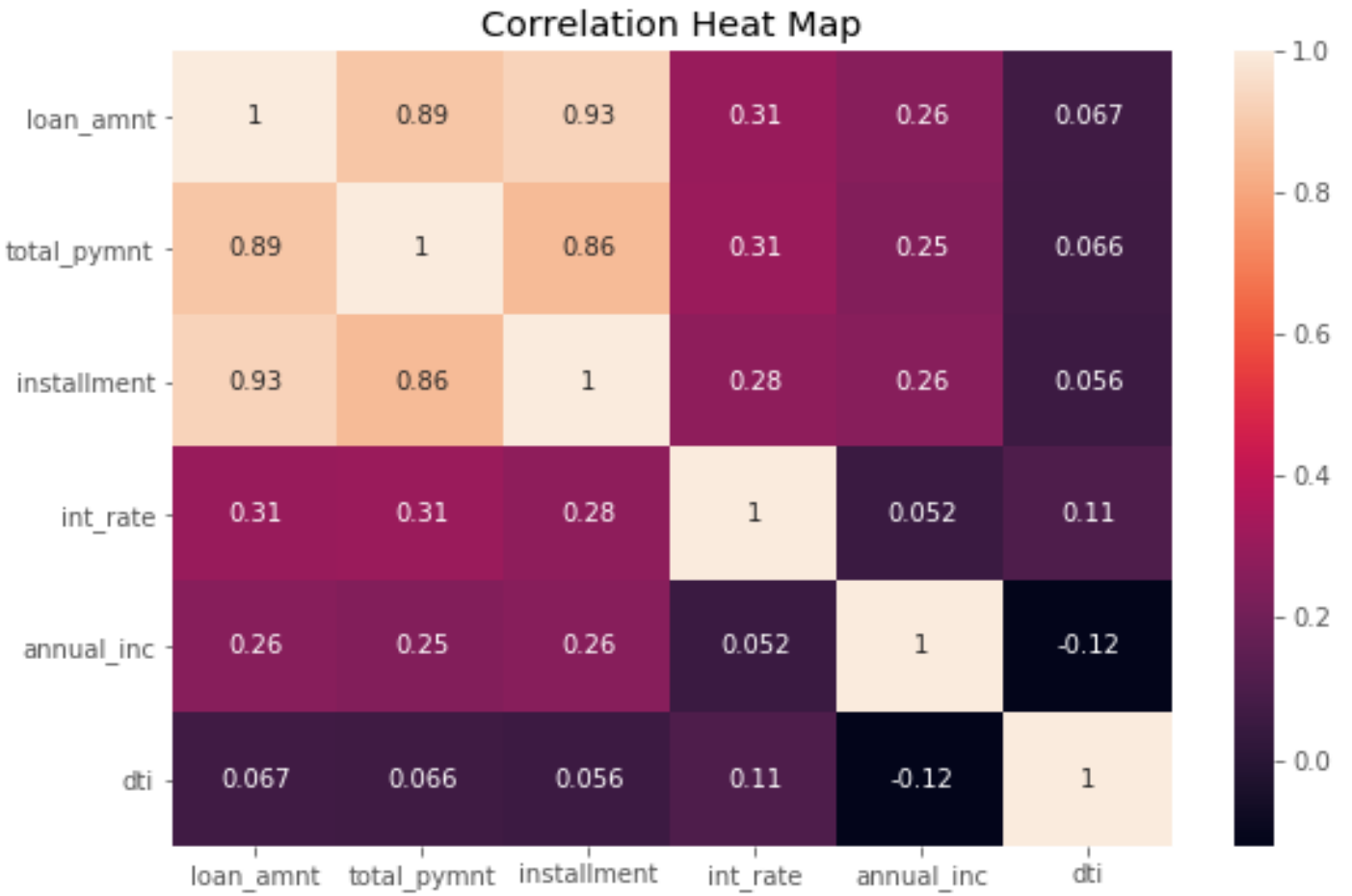
- Verification_status indicates if Income was verified or not.
- The number of borrowers who defaulted is highest for annual income range 40000-60000 around 1648.
- The number of borrowers who Defaulted is highest for Not Verifed around 1992

Part 6 :Bivariate Analysis on Loan Status



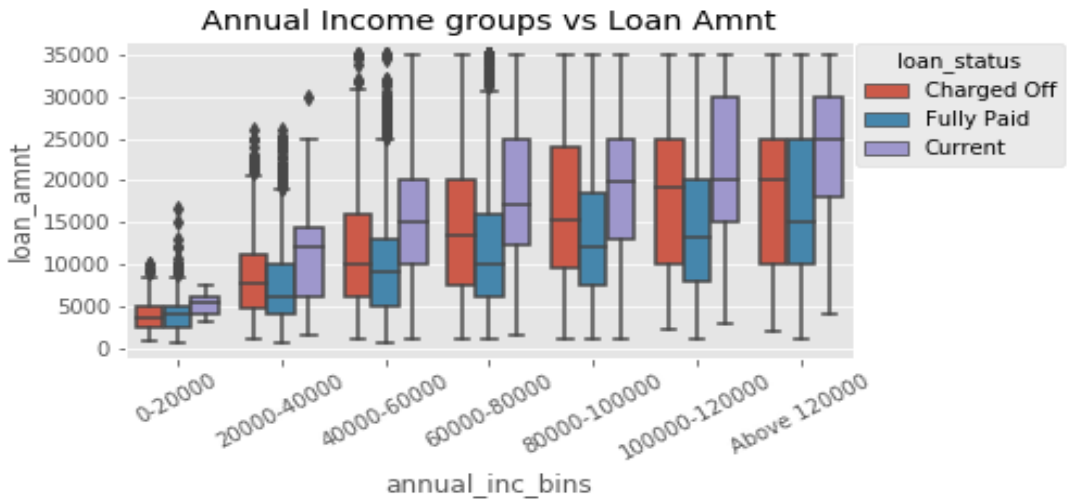
The number of borrowers and number of defaulters rise for number of open accounts from 0 to 7 with highest number of defaulters at open accounts of 7 and it reduces from number of open accounts of 8 and onwards and we observe skewness to the right of the distribution

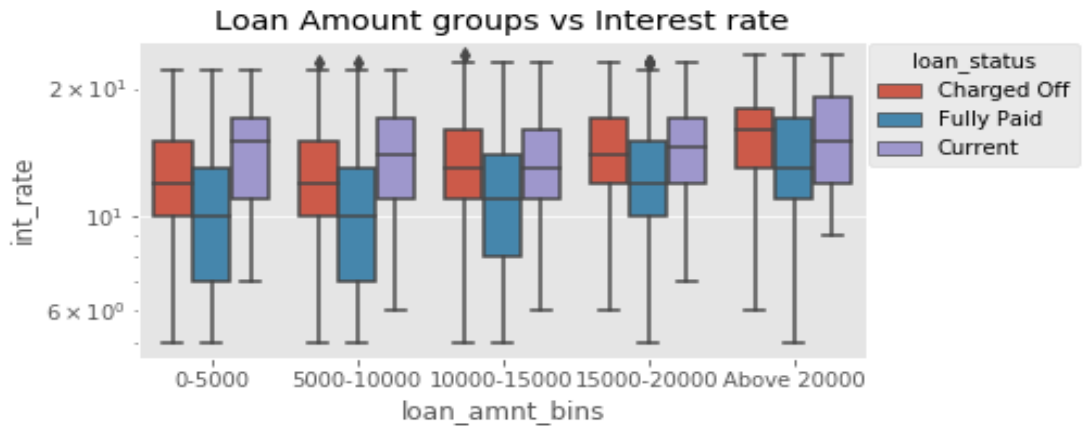
Part 7: Bivariate Analysis on Continuous Variables

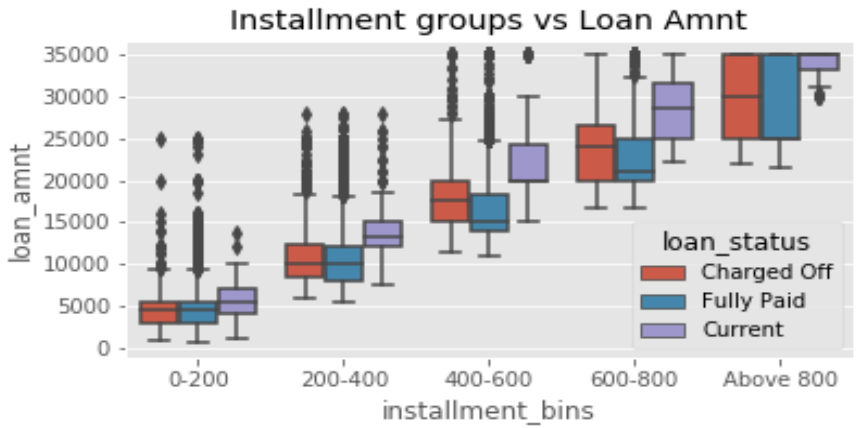


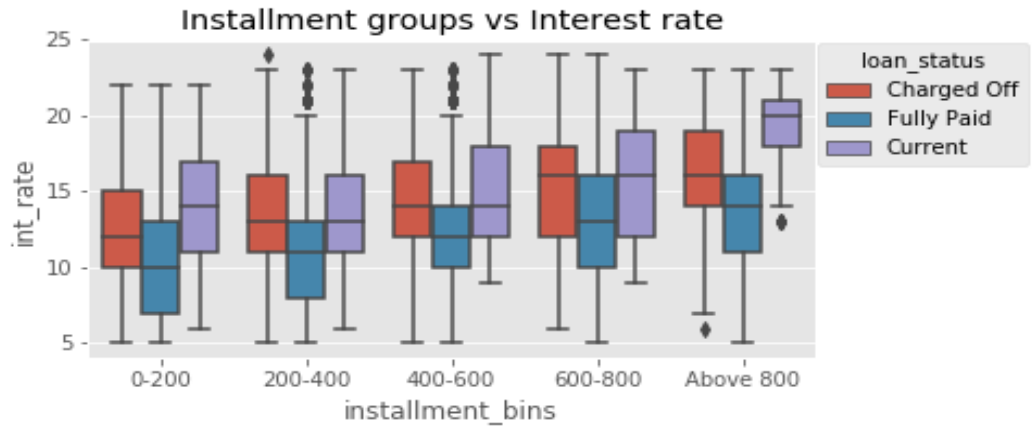
- Loan Amount and Installment have strong positive correlation.
- Loan Amount and total_pymnt have strong positive correlation.
- total_pymnt have positive correlation with installment.

Part 8: Multivariate Analysis on Continuous vs Categorical Variables









- Annual Income groups vs Loan Amnt: The variation in loan amount increases as Income groups increase for Charged Off.
- Installment groups vs Loan Amnt: As installment increases we see the loan amount also increases, though the variation is high for installments Above 800 for the Charged Off.

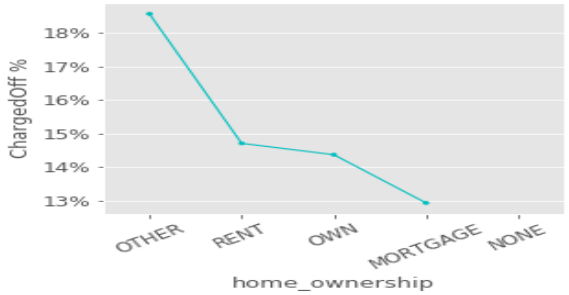
Part 9 : Data and Business Driven Derived Metrics

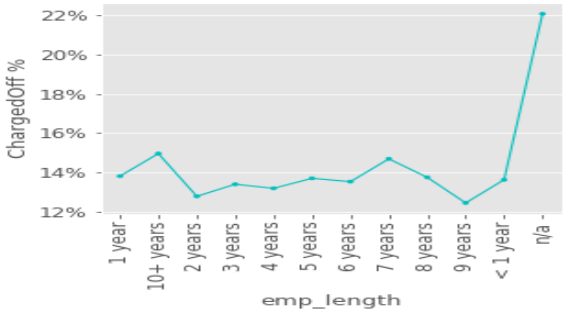
- We created following Data and Business Driven Derived metrics:

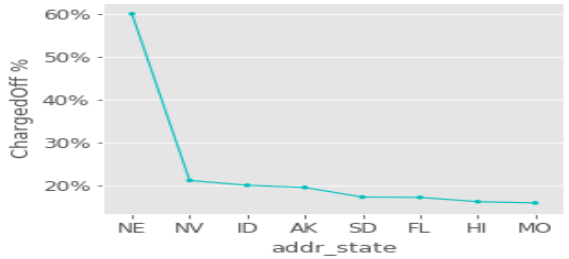
Data Driven Derived Metric/Column name	Derived Metric Formula
Charge Off %	Charge Off % = Number of Loan Status Charged Off by Category/Number of Loan Status(all status) by category
Non Verification Rate	Non Verification Rate = Number of Verification status Not Verified by Category/Number of Verification status(all status) by category
Business Driven Derived Metric/Column name	**Derived Metric Formula**
Monthly InHand Savings	Monthly InHand Savings Available to applicant after deductions = dti * monthly Salary /100

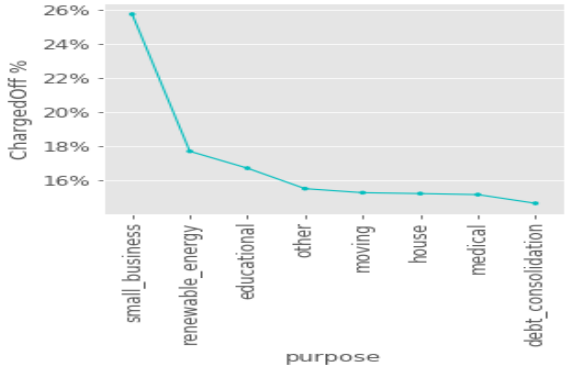
Part 9 : Data and Business Driven Derived Metrics

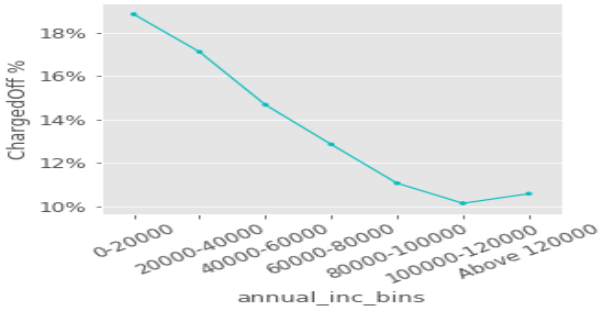
Charge Off % for basic borrowers characteristics:







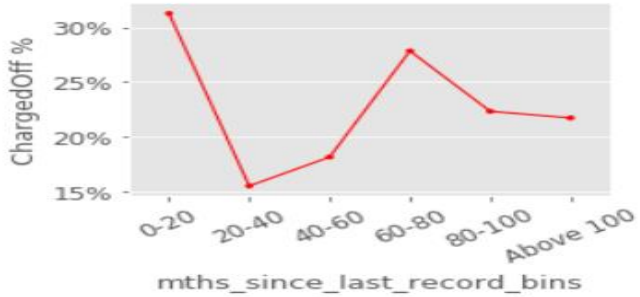


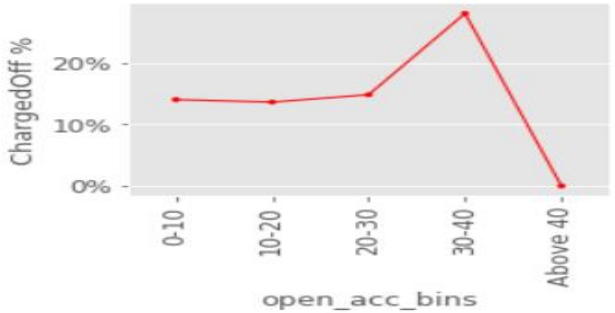


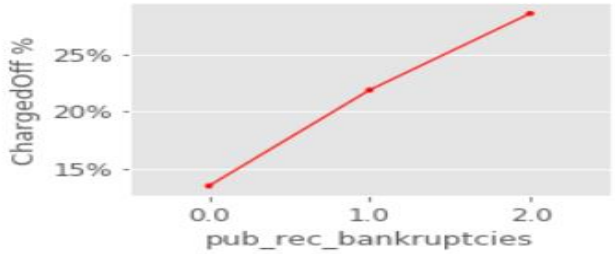
- **Charged Off % for borrower's characteristics is as seen below:-**
  - **It is highest for Home ownership 'RENT'(ignoring OTHER) around 15%**
  - **It is highest for emp length of '10+ yrs' around 15%**
  - **It is highest for Addr state as 'NE' around 60%**
  - **It is highest for purpose of 'small business' around 25.5%**
  - **It is highest for income range '0-20000' around 18.5%**

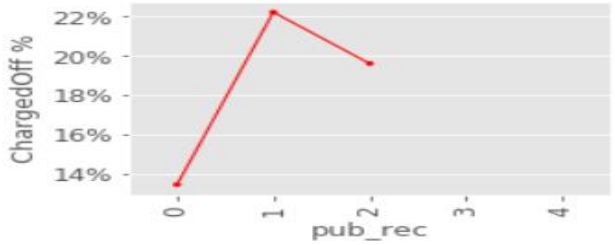
Part 9 : Data and Business Driven Derived Metrics

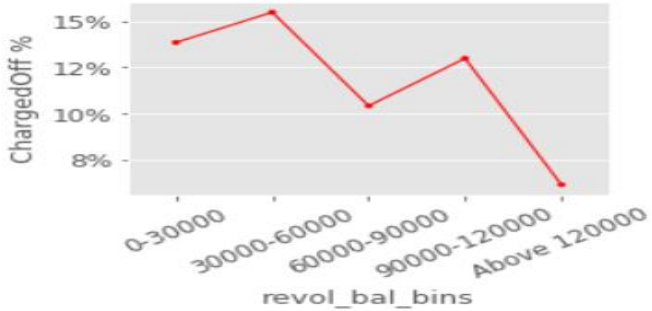
Charge Off % for additional borrowers characteristics:

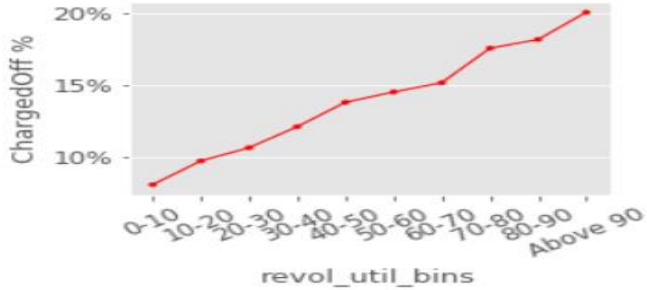








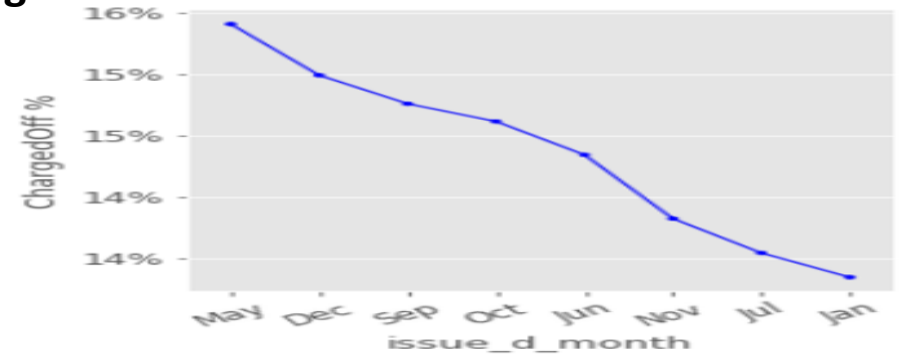


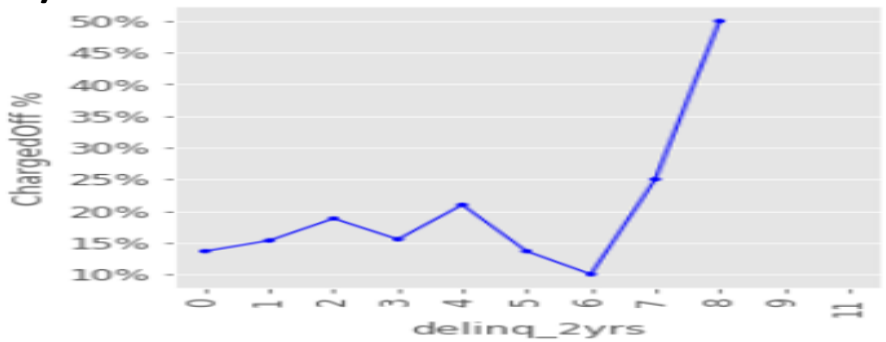


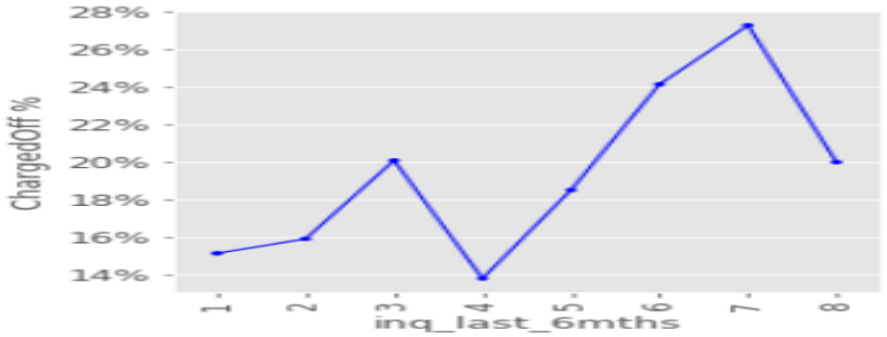
- Charged Off % for borrower's characteristics is as seen below:-
  - Month since last record is highest for range 0-20 around 32%
  - Open account is highest for range 30-40 around 28%
  - Public record bankruptcies is highest for value 2 around 27%
  - Public record is highest for any value above 0 around 20-22%
  - Revolving balance is highest for range 0-60000 around 13-16%
  - Revolving balance utilization shows continuous increment and highest for range above 90% at 20%

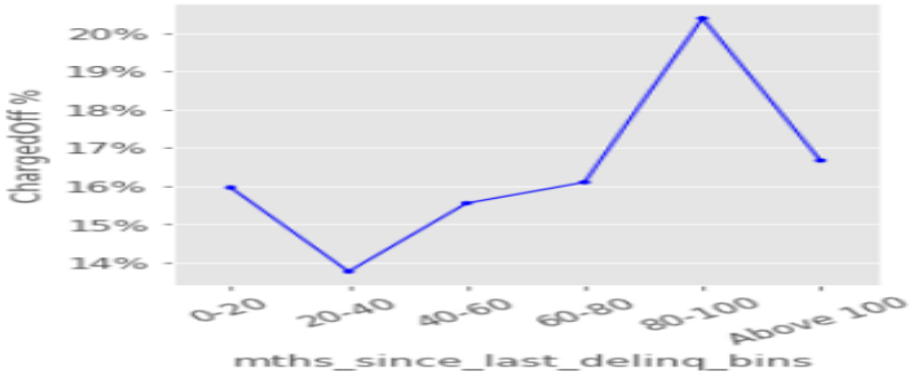
Part 9 : Data and Business Driven Derived Metrics

Charge Off % for additional borrowers characteristics (contd.):





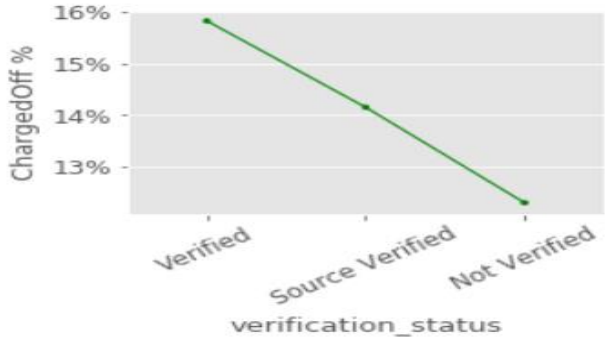


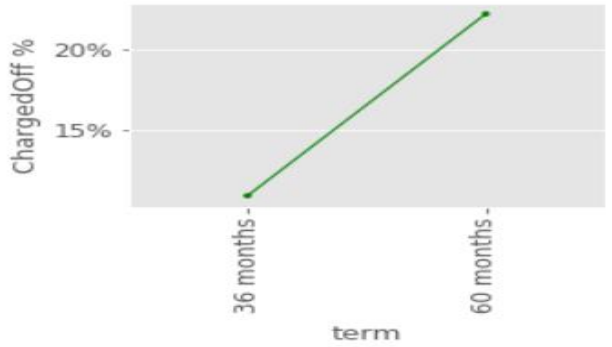


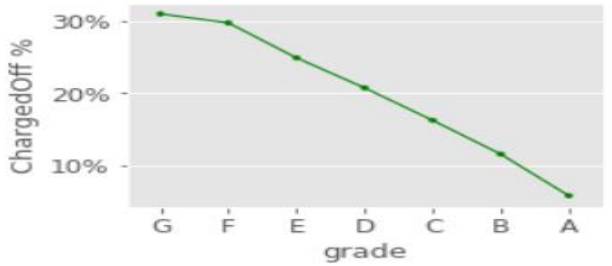
- **Charged Off % for each variable is as seen below:-**
  - **It is highest for Loan issued in month of "May & Dec" with charge off rate: around 15-16%**
  - **It is highest for no of delinquency in last 2 years with value 7 seen with highest charge off rate i.e. 50%**
  - **No of inquiries with value 6 & 7 seen with highest charge off rate around 24-28%**
  - **No of month since last delinquency with range 80-100 seen with highest charge off rate around 21% i.e. Applicant may not pay loan after showing decent behavior for few months after failing to pay EMI**

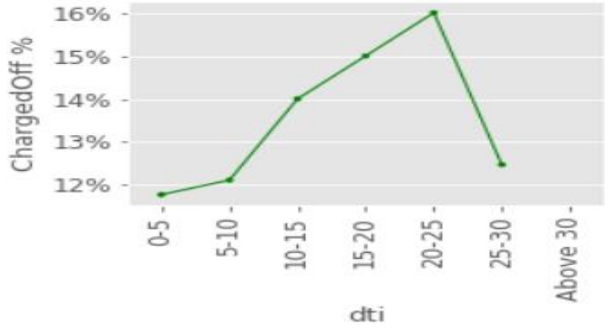
Part 9 : Data and Business Driven Derived Metrics

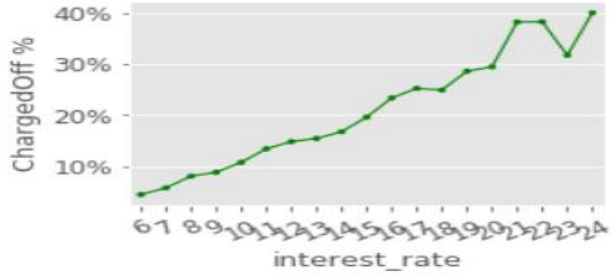
Charge Off % for Lenders parameters:

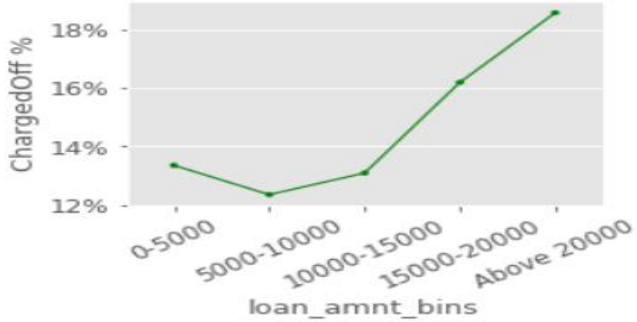








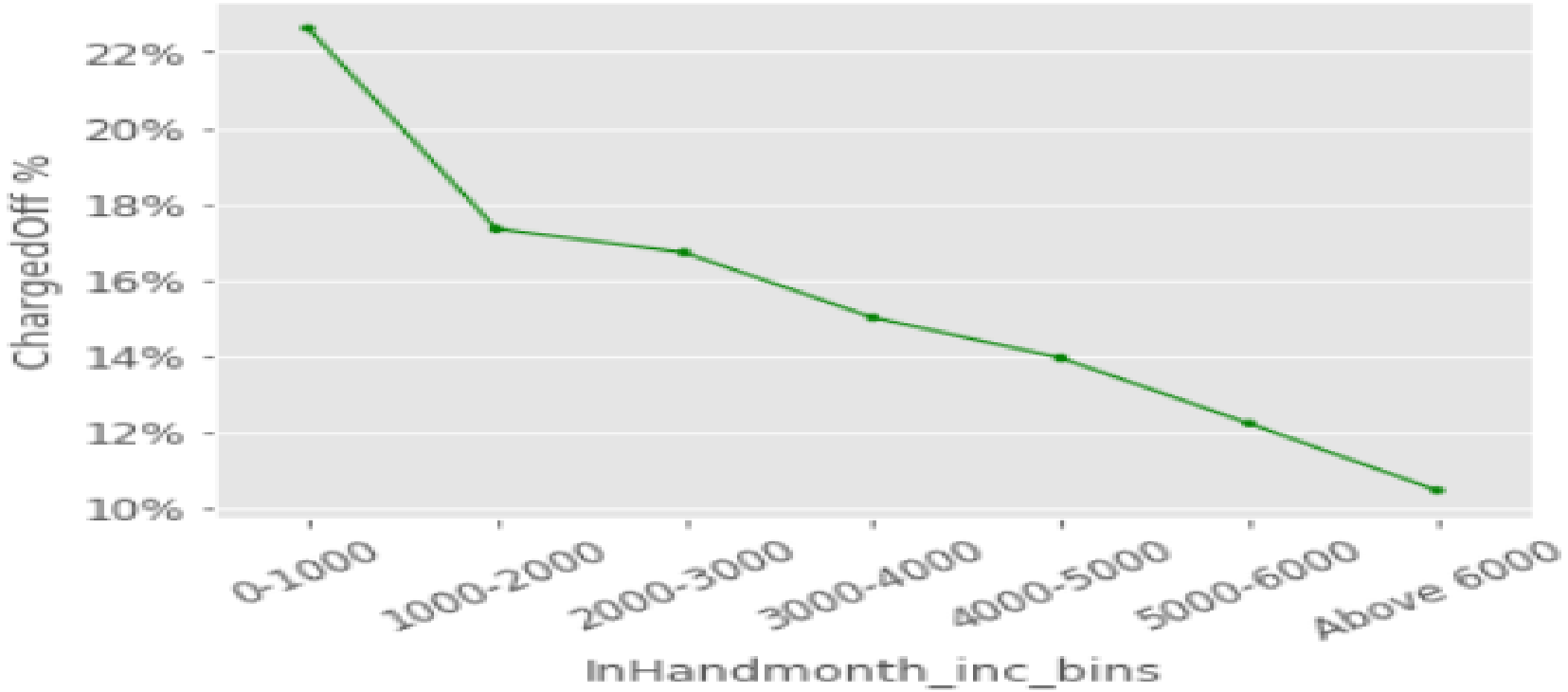




- **Charged Off % for lenders parameters is as seen below:-**
  - **It is highest for Verification Status as 'Verified' around 16%**
  - **It is highest for term of '60 months' around 23%**
  - **It is highest for grade of 'G' around 30% and there is increase from A to G**
  - **It is highest for dti range of '20-25' around 16% and shows increase from range of '5-10' to '20-25'**
  - **It shows increasing trend for Interest rate with steep increase between 20 to 21%**
  - **It shows an increasing trend for Loan Amounts greater than 10,000 with sharp increase after 15000**

Part 9 : Data and Business Driven Derived Metrics

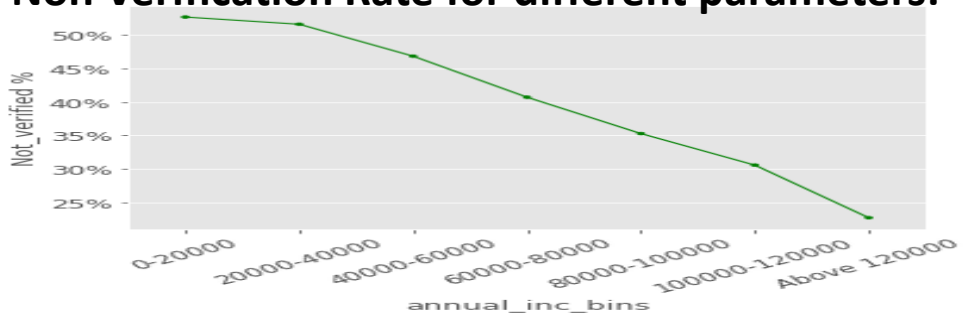
Charge Off % against Monthly InHand Savings:

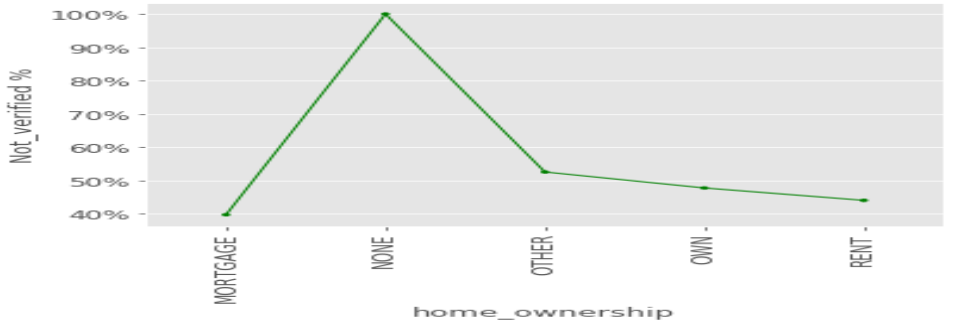


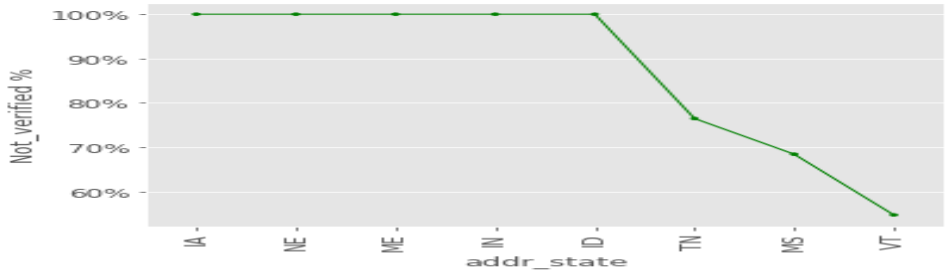
- **This shows that borrower having Monthly InHand Savings in range "0 to 1000" available after deducting all EMIs from monthly salary has highest charge of rate 22%**

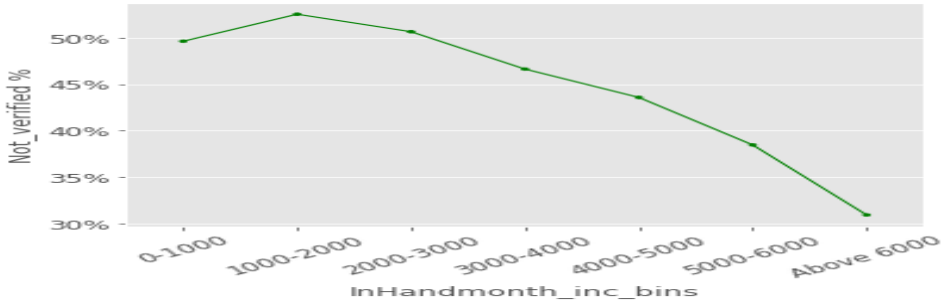
Part 9 : Data and Business Driven Derived Metrics

Non Verification Rate for different parameters:









Non verification rate is high among following categories which also has very charge off rate:

1) "Annual Income" for low salary bins has more than "50%" non verified rate same low salary bins also has highest charge off rate i.e. approx. "20%"
2) Home ownership with value as "None" for low salary bins has 100% non verified rate same option i.e. 'None' also has highest charge off rate i.e. approx. 20%
3) Address state, e.g. "NE" has 100% non verified rate same state "NE" also has highest charge off rate i.e. "60%"
4) Monthly InHand Savings ("Monthly Income left after paying all EMIs“) has more than 50% non verified rate for values 0 to 3000 same values also has highest charge off rate i.e. 15 to 25%

To summarize, if we check the charge off% plot with the parameters 'Annual Income', 'Home ownership', 'Address state' and 'Monthly InHand Savings’ (Monthly Income left after paying all EMIs) it is clear that the charge off% is higher for sections/values/ranges where non-verification rate is high
We have also seen in previous plots that Charge off rate is highest among verified borrower
Looking at above plot, Finance company should improve borrower verification process

Summary of Analysis

- Thus, the loan company should look at the below **Driver variables** based on the Charge Off %:
  Home Ownership for RENT
  Emp Length for 10+ yrs
  Addr State for NE
  Purpose for small business.
  Annual Income for income range 0-20000

  Verification Status for Verified
  Term for loan in 60 months
  Grade especially G
  dti range from 5-10 till 20-25
  Interest rate greater than 20%
  Loan Amounts greater than 10,000

  Month since last record in range 0-20 which has highest charge off around 32%
  Open accounts in range 30-40 which has highest charge off around 28%
  Public record bankruptcies for value 2 which has highest charge off around 27%
  Public record for any value above 0 which has highest charge off around 20-22%
  Revolving balance in range 0-60000 which has highest charge off around 13-16%
  Revolving balance utilization for range above 90% as it shows continuous increment for charge off and is highest at 20%

  Loan issued in month of "May & Dec" seen with highest charge off rate: around 15-16%
  No of delinquency in last 2 years with value 7 seen with highest charge off rate i.e. 50%
  No of inquiries in last 6 months with value 6 & 7 seen with highest charge off rate around 24-28%
  No of month since last delinquency with range 80-100 seen with highest charge off rate around 21% i.e. Applicant may not pay loan after showing decent behavior for few months after failing to pay EMI
  Borrower having **Monthly InHand Savings** range "0 to 1000" available after all EMIs deducted from monthly salary has highest charge of rate 22%

- **Finance company should improve borrower verification process to minimize loss.**

Final Conclusion/Recommendation

The consumer finance company should look into the below mentioned variables and their critical values related to loan ChargedOff % to help it identify the loan defaulters in the past.

Variables	Critical Values
Home Ownership	RENT
Emp Length for	10+ yrs.
Addr State	NE
Purpose	small_business
Annual Income	range 0-20000
Verification Status	Verified
Term	60 months
Grade	G
dti	range from 5-10 till 20-25
Interest rate	Greater than 20%
Loan Amounts	greater than 10,000

Variables	Critical Values
Month since last record	range 0-20
Open accounts	range 30-40
Public record bankruptcies	Value 2
Public record	Values above 0
Revolving balance	range 0-60000
Revolving balance utilization	range above 90%
No of inquiries in last 6 months	Values 6 & 7
No. of delinquency in last 2 years	Value 7
No of month since last delinquency	range 80-100
Monthly InHand Savings	Range "0 to 1000"
Loan issued in month	May & Dec