GRAMENER CASE STUDY

SUBMISSION

Risk Analysis of Loan Applicant Profiles

Group Name: **<u>Data Ninjas</u>**

1. Alvin Mark Windsor
2. Ravi Kiran Vissa
3. Abdi Adam
4. Alfred Mburugu

Case Study Overview

CONTEXT

The company is an online credit marketplace and acts as aggregator between the lenders [investors] & the borrowers of money. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Once the application is accepted and the loan is sanctioned then a borrower will re-pay the amount in monthly installments completely or can default leading to credit loss.

PROBLEM

The largest source of financial loss to the company is through credit loss resulting from lending loans to '**risky applicants**'. If a borrower fails to repay the loan then the loan is termed as '**charged off**' and the pending amount is the credit loss to the company.

INFORMATION AVAILABLE

The dataset provided in this assignment has consumer & loan attributes of previously approved loan applications. The attached data dictionary gives us a detailed summary of the significance of each attribute. Note- We do not have information regarding the previously rejected loan applications.

OBJECTIVE

The aim is to identify patterns and driving variables which indicate if a loan applicant is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

METHODOLOGY AND DELIVERABLES

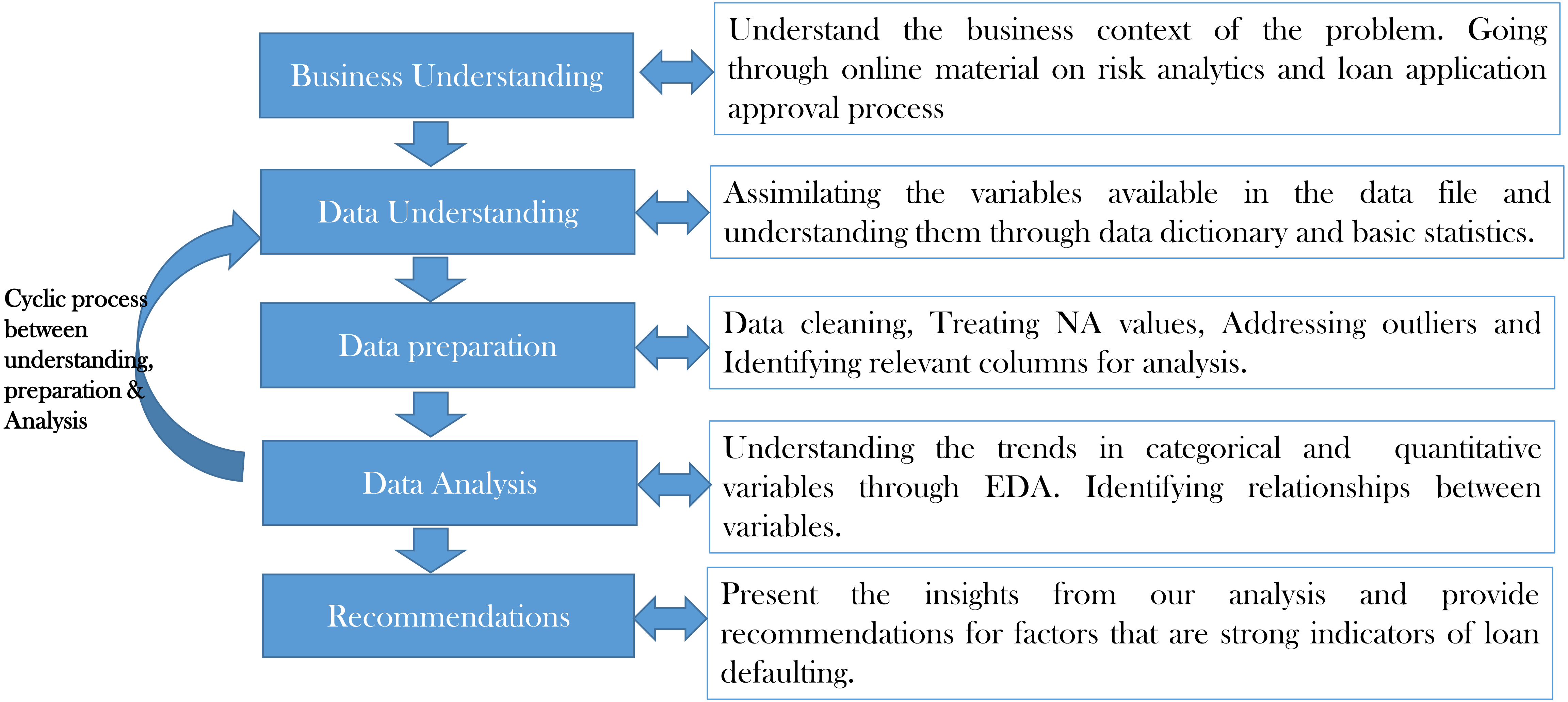
Implement exploratory data analysis on the input dataset to gain an understanding of risk analytics in a business environment.

[1] Perform Uni-variate and Multi-variate analysis of variables to identify latent patterns and inconsistencies in the dataset.

[2] Determine driving factors which are strong indicators of 'risky applicants'.

[3] Present findings through neat visualizations.

Problem Solving Methodology



Business Understanding

Understand the business context of the problem. Going through online material on risk analytics and loan application approval process

Data Understanding

Assimilating the variables available in the data file and understanding them through data dictionary and basic statistics.

Data preparation

Data cleaning, Treating NA values, Addressing outliers and Identifying relevant columns for analysis.

Data Analysis

Understanding the trends in categorical and quantitative variables through EDA. Identifying relationships between variables.

Recommendations

Present the insights from our analysis and provide recommendations for factors that are strong indicators of loan defaulting.

Cyclic process between understanding, preparation & Analysis

Assumptions and Data Handling

[1] Loan Status – The loan status has three levels or outcomes [current, fully paid and charged off]. In our analysis we will disregard all the records with loan status as ***Current***. The loans that are currently active have an uncertain outcome, they can either successfully result in fully paid or default and lead to charged off condition. Therefore we will only use the records with known outcomes [charged of or fully paid] to derive any insights.

[2] Data standardization-Following 5 fields are not in standard date format *[Issue_d, earliest_cr_line, last_pymnt_d, next_pymnt_d & last_credit_pull_d]*

Treatment <Using custom ***fundateconversion function*** to convert the non-standard date records to a standardized date object and including a dummy date as 01st of every month to represent it as yyyy-mm-dd format for the aforementioned five date attributes.>

[3] Data Cleansing- Removed the columns which are all NAs or with only one unique value as the lack of variability will not contribute to any useful insights. The ***datachop function*** checks all the records of each column for more than 1 unique value, if there is only one unique value or NA it deletes that column from the loan dataset.

[4] Loan Title is a drilldown of the loan purpose attribute and is specific to the applicants loan needs. Since it is a text heavy column with numerous entries branching from a main group it will not be essential for EDA. We will exclude this column.

[5] Loan desc is again a text based description of the purpose of the loan. Since text analysis isn't under the purview of this case study we will not consider this attribute as well

[6] URL leads to a web address specific to a particular loan application record. It is again a non-essential attribute in analyzing driver attributes to credit loss or fraudulent loan applicants. We will disregard this column as well.

[7] The attribute **collections_12_mths_ex_med** representing number of collections in the past year excluding medical collections contains records with 0 or Na values. Therefore we will disregard this attribute.

[8] The data structure of the **int_rate and revol_util** are represented as character type due to the presence of the % symbol. We will remove the symbol and represent it as a numerical value representative of the interest percentage. i.e. 10.75% will be converted into 10.75.

[9] The schema for this dataset reveals two primary keys [1] "id" denoting the unique ID assigned by LC to the loan application [2] member_id a unique code representing a loan applicant. Having both these primary keys are redundant we can eliminate the member_id and map all records to the unique LC "id". Note- no analysis will be done on the basis of the **LC id** either therefore it can technically be removed.

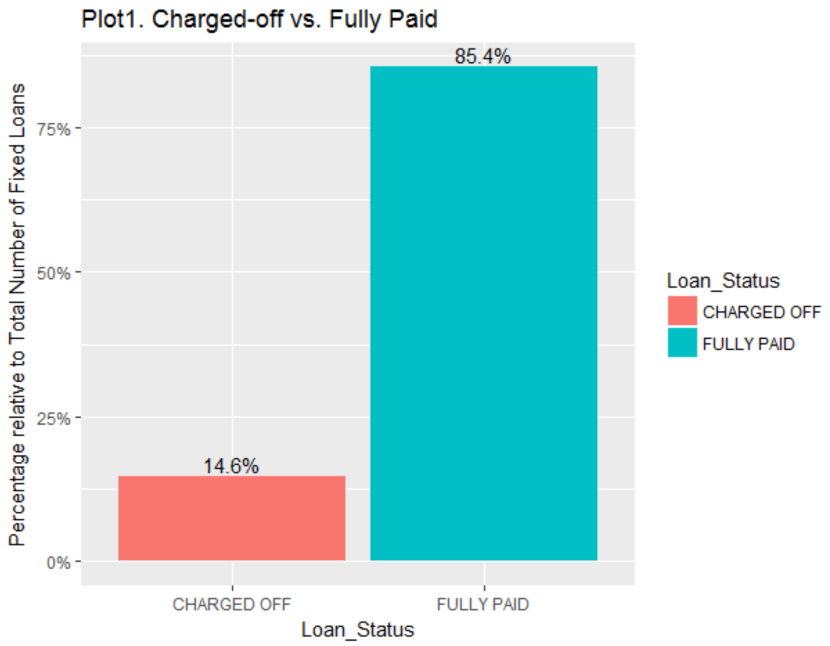
[10] The **emp_title** column representing the job title of the loan applicant contains records with numerous special characters and non-uniform text discrepancies. Therefore we will try to resolve some of these issues by manually sifting through the dataset and resolving the records with the highest frequency of occurrences.

[11] We will convert all character attribute records to upper case to avoid any case sensitive inconsistencies and data entry discrepancies. The case conversion will be done using a custom defined *function caseconversionfun.*

[12] During analysis of any variable the records with missing values or NA records will be handled appropriately through transformations but no direct data imputations will be made to the master dataset.

[13] Business driven metrics like FICO score the have not been computed due to the lack of required information in the dataset.

Overview of Business Status



- **Insight 1- From this graph it is evident that ≈15% of all the issued loans under consideration have resulted in credit loss. Therefore, it is critical to look into this issue.**

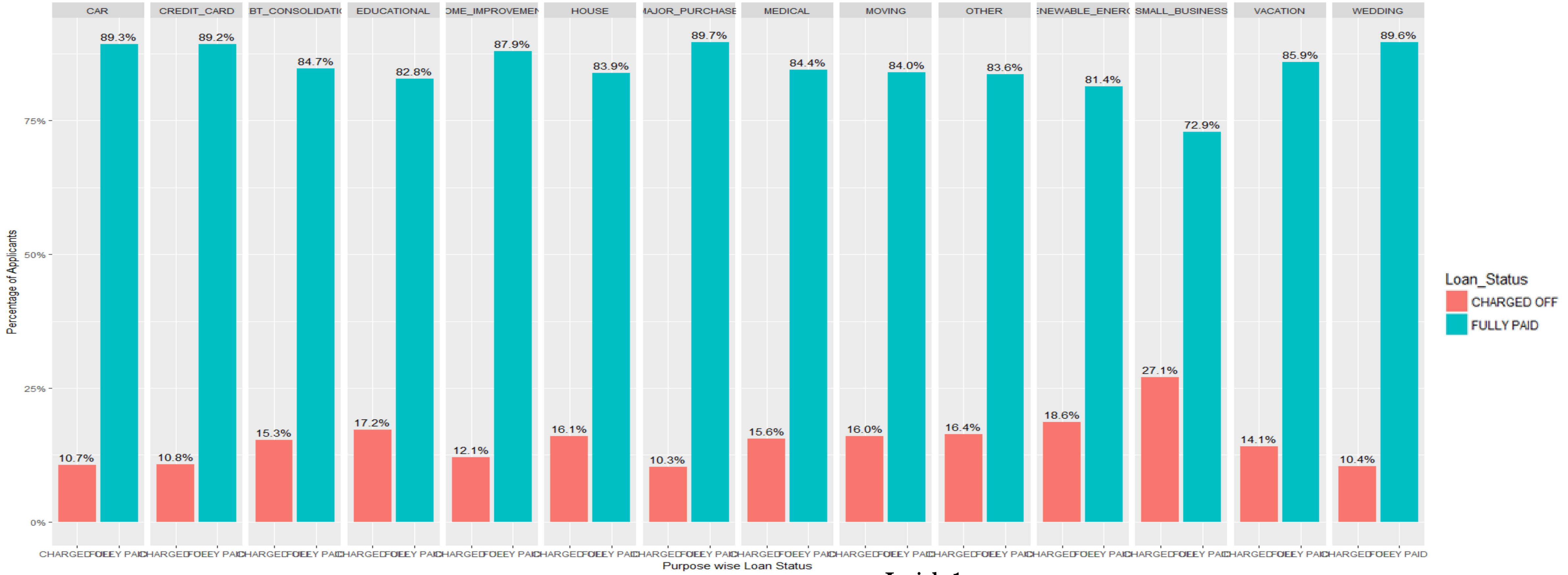
- The variables are classified as follows
- Input factors – variables which can be taken as input to analyze the applicant
- Customer Demographics – Variables related to customer demographics(these are also inputs of the applicant
- Customer Information – Variables which give more information about applicant.
- LC Loan Payment Variables – Variables which are related to behavior of payment for LC
- Output Factors – Variables related to Loan provided to applicant
- Others – all other variables which are not fitting in above categories

Data Understanding

Column Variable	Category	Unordered Categorical Variable(UOCV)/Ordered Categorical Variable(OCV)/Quantitative Variable(QV)
dti	Input Factors	QV
earliest_cr_line	Input Factors	OCV
inq_last_6mnths	Input Factors	QV
mnths_since_last_record	Input Factors	QV
open_acc	Input Factors	QV
revol_bal	Input Factors	QV
revol_util	Input Factors	QV
total_acc	Input Factors	QV
acc_now_delinq	Input Factors	QV
chargeoff_within_12_mths	Input Factors	QV
delinq_amnt	Input Factors	QV
pub_rec_bankruptcies	Input Factors	QV
Grade	Customer Demographics	OCV
Sub-Grade	Customer Demographics	OCV
home ownership	Customer Demographics	UOCV
annual_inc	Customer Demographics	QV
zip_code	Customer Demographics	UOCV
addr_state	Customer Demographics	UOCV
ID	Customer Information	UOCV
member_id	Customer Information	UOCV
verification_status	Customer Information	UOCV
issue_d	Customer Information	OCV
loan_status	Customer Information	UOCV
emp_title	Customer Information	UOCV
revol_bal	Input Factors	QV
revol_util	Input Factors	QV
total_acc	Input Factors	QV
acc_now_delinq	Input Factors	QV
chargeoff_within_12_mths	Input Factors	QV
delinq_amnt	Input Factors	QV
pub_rec_bankruptcies	Input Factors	QV
Grade	Customer Demographics	OCV
Sub-Grade	Customer Demographics	OCV
home ownership	Customer Demographics	UOCV
annual_inc	Customer Demographics	QV
zip_code	Customer Demographics	UOCV
addr_state	Customer Demographics	UOCV
ID	Customer Information	UOCV
member_id	Customer Information	UOCV
verification_status	Customer Information	UOCV
issue_d	Customer Information	OCV

Top Observation and Conclusions [1]

Plot38. Loan Purpose versus Loan Status Analysis



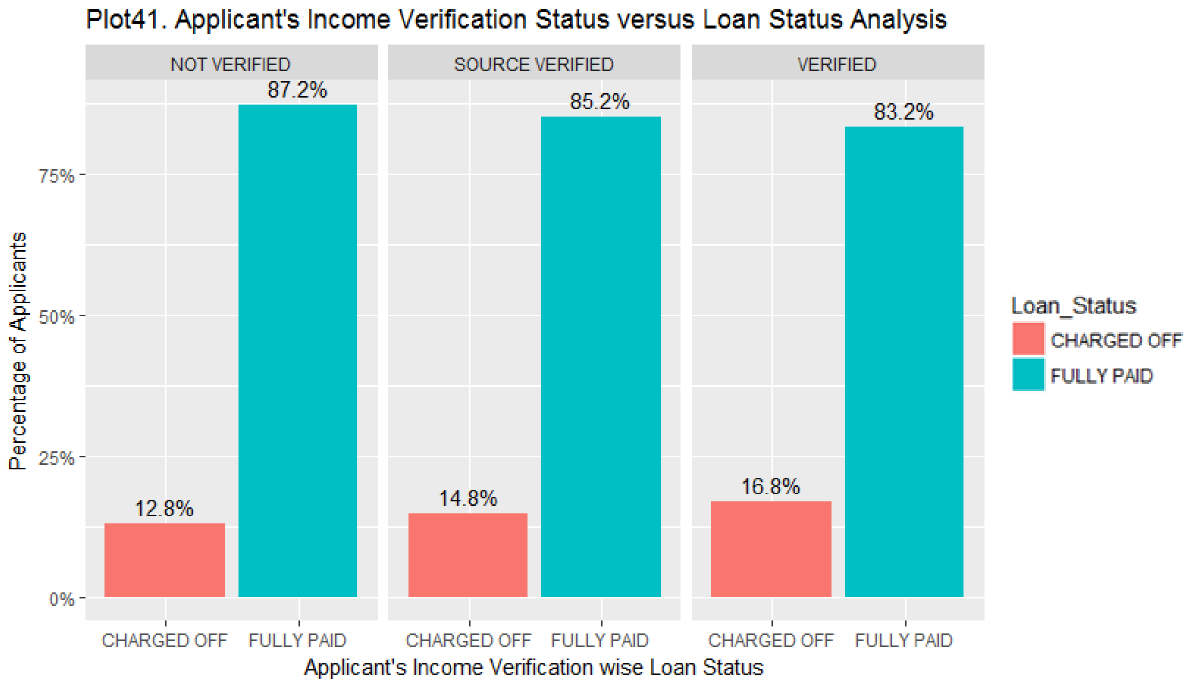
Loan Purpose	Charged Off	Fully Paid	Grand Total	Fraud Percentage [%]
small_business	475	1279	1754	27.08%
renewable_energy	19	83	102	18.63%
educational	56	269	325	17.23%
other	633	3232	3865	16.38%
house	59	308	367	16.08%
moving	92	484	576	15.97%
medical	106	575	681	15.57%
debt_consolidation	2767	15288	18055	15.33%
vacation	53	322	375	14.13%

Insight1

Although the highest number of loan applications are received for debt consolidation [Refer Table on Left] From the Plot it is clear that the highest percentage of defaulters state loan purpose as Small Business. **27% of all loans taken for small businesses result in credit loss to the company.**

As **'others'** contributes to the next highest default % we can look into providing more selection options for loan purpose in the loan application. This will lead to better analysis options.

Top Observation and Conclusions [2]

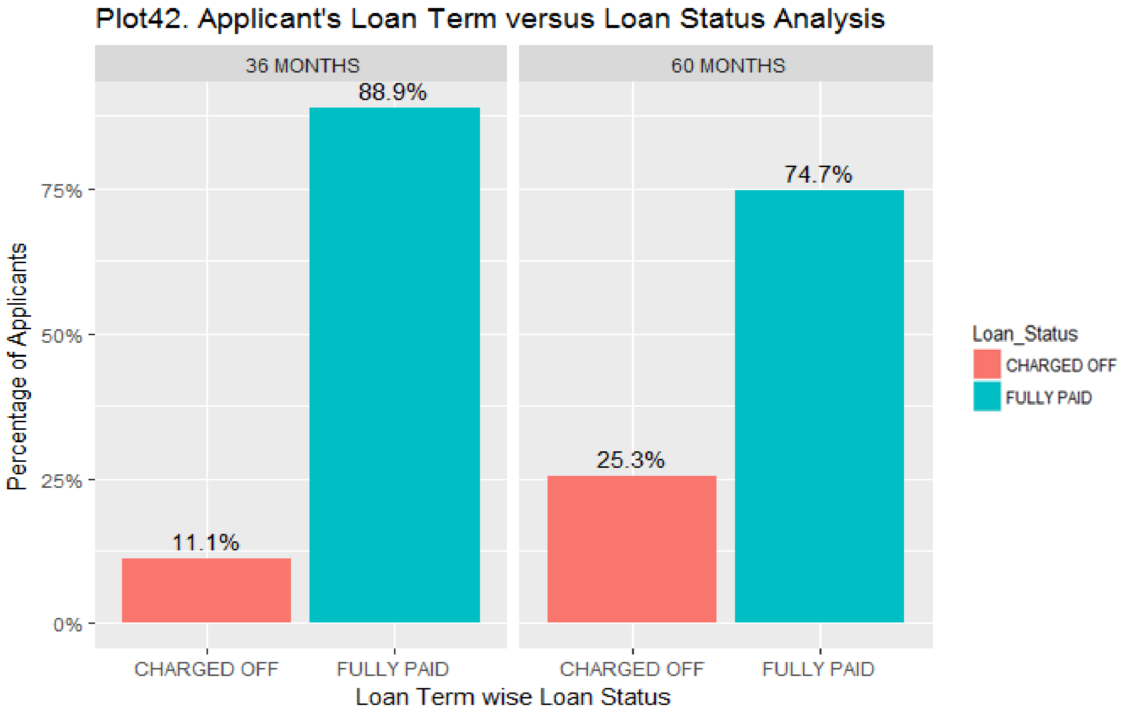


verification_status	Charged Off	Fully Paid	Grand Total	Fraud Percentage
Verified	2051	10155	12206	16.80%
Source Verified	1434	8243	9677	14.82%
Not Verified	2142	14552	16694	12.83%

Insight2

LC has incorporated a system to review the income source of the loan applicant. From the table it is clear that the highest number of applications received not verified. However, from the plot it is evident that <u>applications that have the income verified or source verified have a higher chance of leading to credit loss.</u>

LC must appraise and analyze the system implemented for verification.



Term	Charged Off	Fully Paid	Grand Total	Fraud Percentage
60 months	2400	7081	9481	25.31%
36 months	3227	25869	29096	11.09%

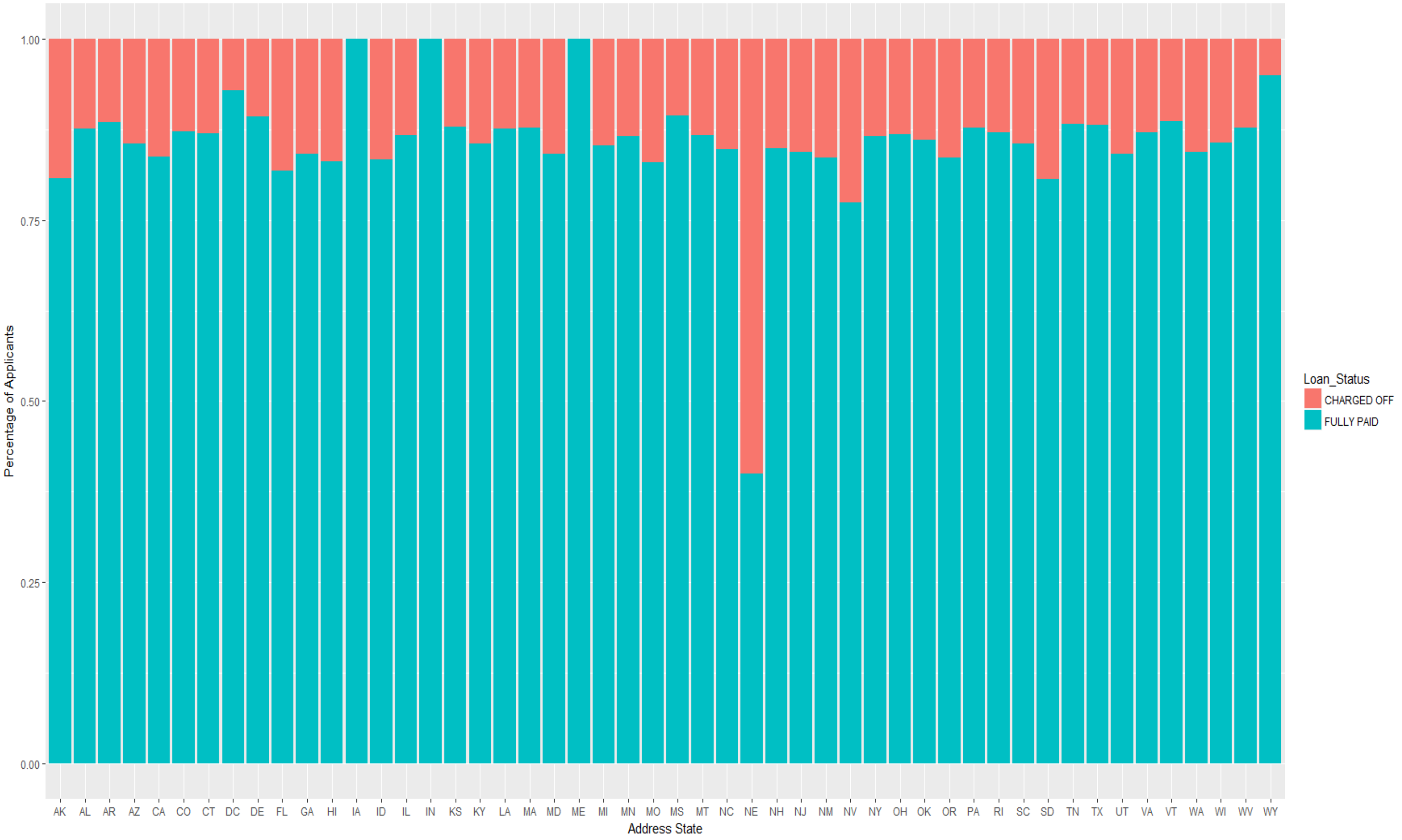
Insight3

Majority of loans are issue for a duration of 36 months. However, from the plot it is clear that a loans issued for a term of 60 months has a significantly higher chance of resulting in credit loss.

25% of all loans under consideration issued for a term of 60 months results in credit loss or loan default.

Top Observation and Conclusions [3]

Plot43. Applicant Residence State versus Loan Status Analysis

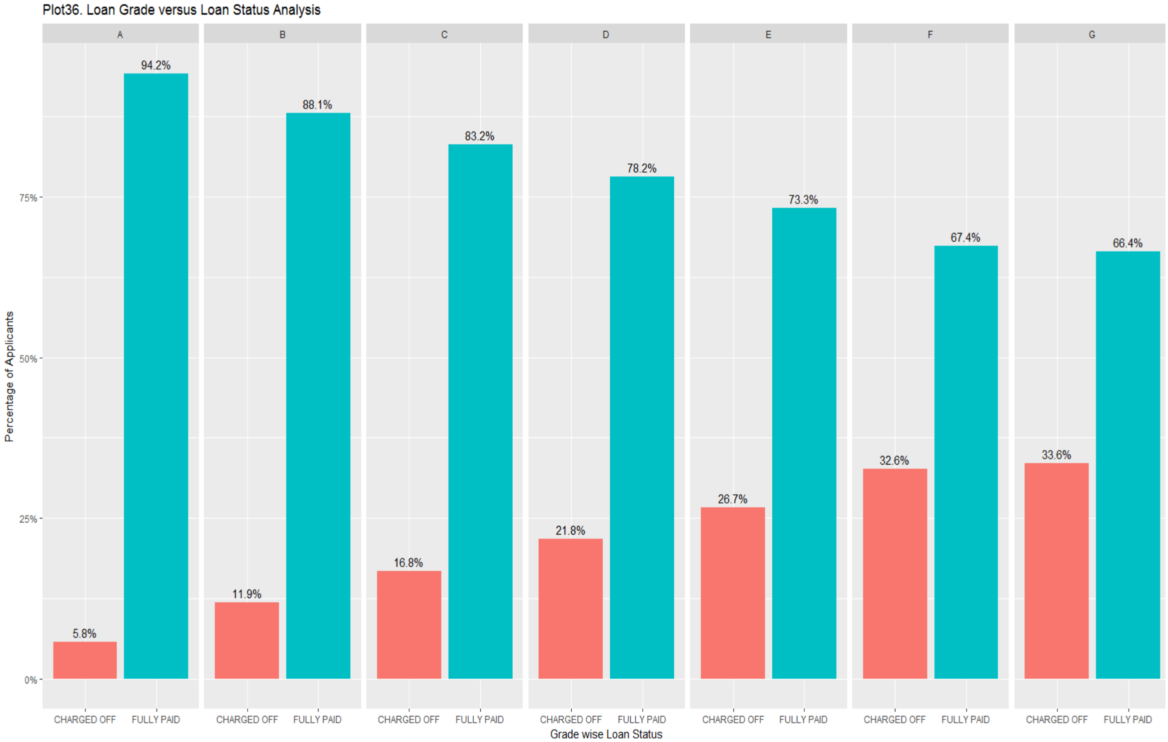


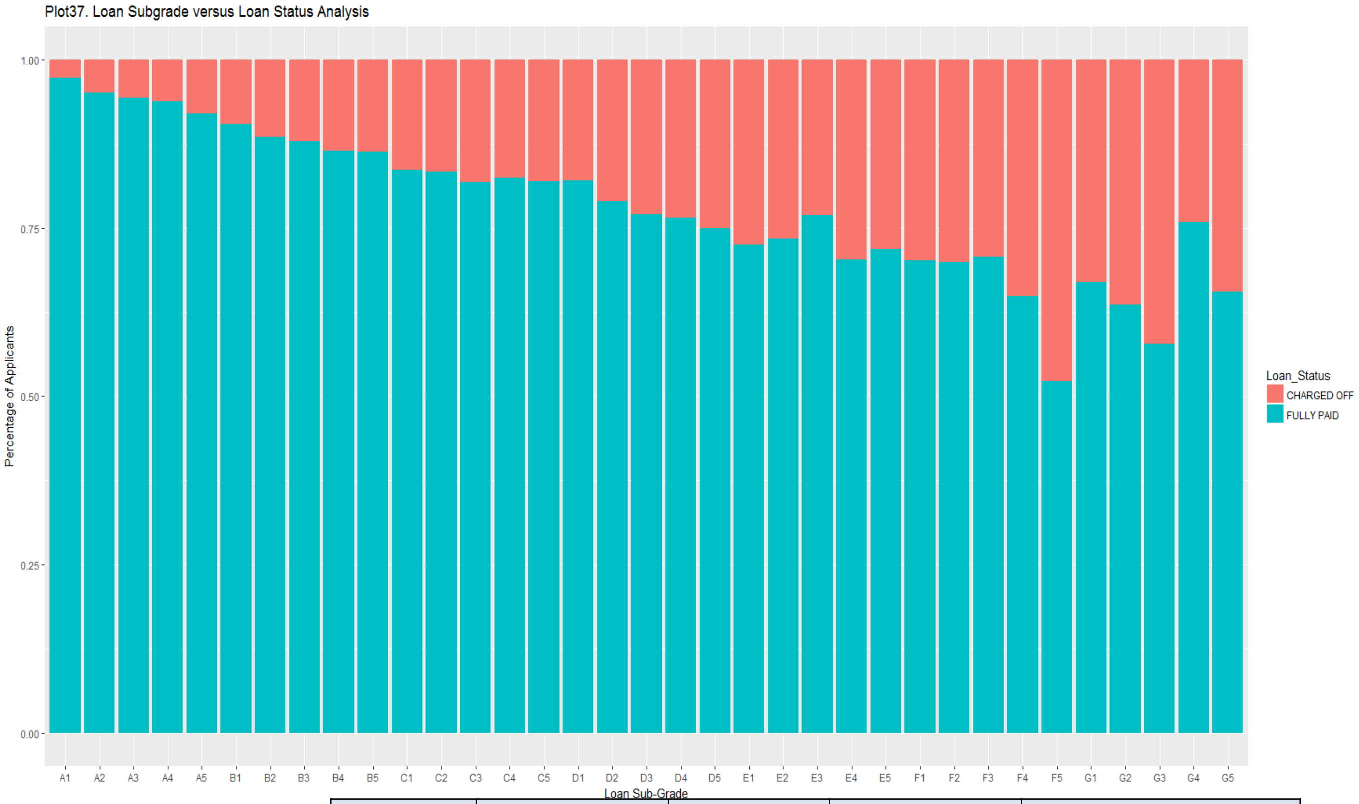
State	Charged Off	Fully Paid	Grand Total	Fraud Percentage
CA	1125	5824	6949	16.19%
NY	495	3203	3698	13.39%
FL	504	2277	2781	18.12%
TX	316	2343	2659	11.88%
NJ	278	1512	1790	15.53%
IL	197	1281	1478	13.33%
PA	180	1288	1468	12.26%
VA	177	1192	1369	12.93%
GA	215	1144	1359	15.82%
MA	159	1138	1297	12.26%
OH	155	1023	1178	13.16%
MD	162	861	1023	15.84%
AZ	123	726	849	14.49%
WA	127	691	818	15.53%
CO	98	668	766	12.79%
NC	114	636	750	15.20%
CT	94	632	726	12.95%
MI	103	601	704	14.63%
MO	114	556	670	17.01%
MN	81	524	605	13.39%
NV	108	371	479	22.55%

Insight4- From the proportionality plot it is clear that borrowers belonging to the state of **Nevada and Nebraska [with 60% and 22.55% default rate respectfully]** contribute to the highest default percentage. However, the sample size for the state of Nevada is only 5. Therefore we will disregard it and consider the next highest state, i.e Nebraska. The highest number of loan applications are received from **California, New York, Florida and Texas. Their respective default rates are 16.1%,13.4%,18.1% and 11.9%**

Top Observation and Conclusions [4]

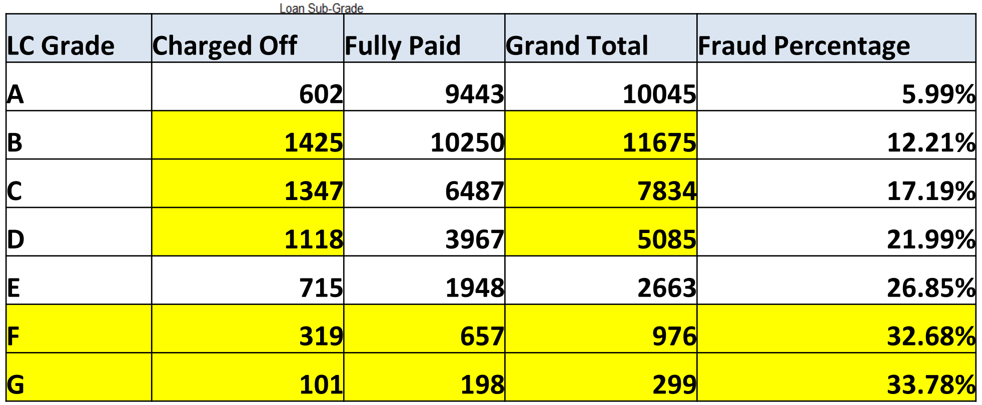
UpGrad


Plot36. Loan Grade versus Loan Status Analysis


Plot37. Loan Subgrade versus Loan Status Analysis

Insight5- From the table it is clear that a highest frequency of loans belong to grades A,B,C and D from which the highest number of defaulters belong to grades B, C and D. However, when we look at the proportion graph of Grade versus Percentage Default [Fraud Percentage] there is a clear trend of increase in the % of defaulters from the grade A to E. [With A having lowest percentage of default and **G having the highest 33.6%.**

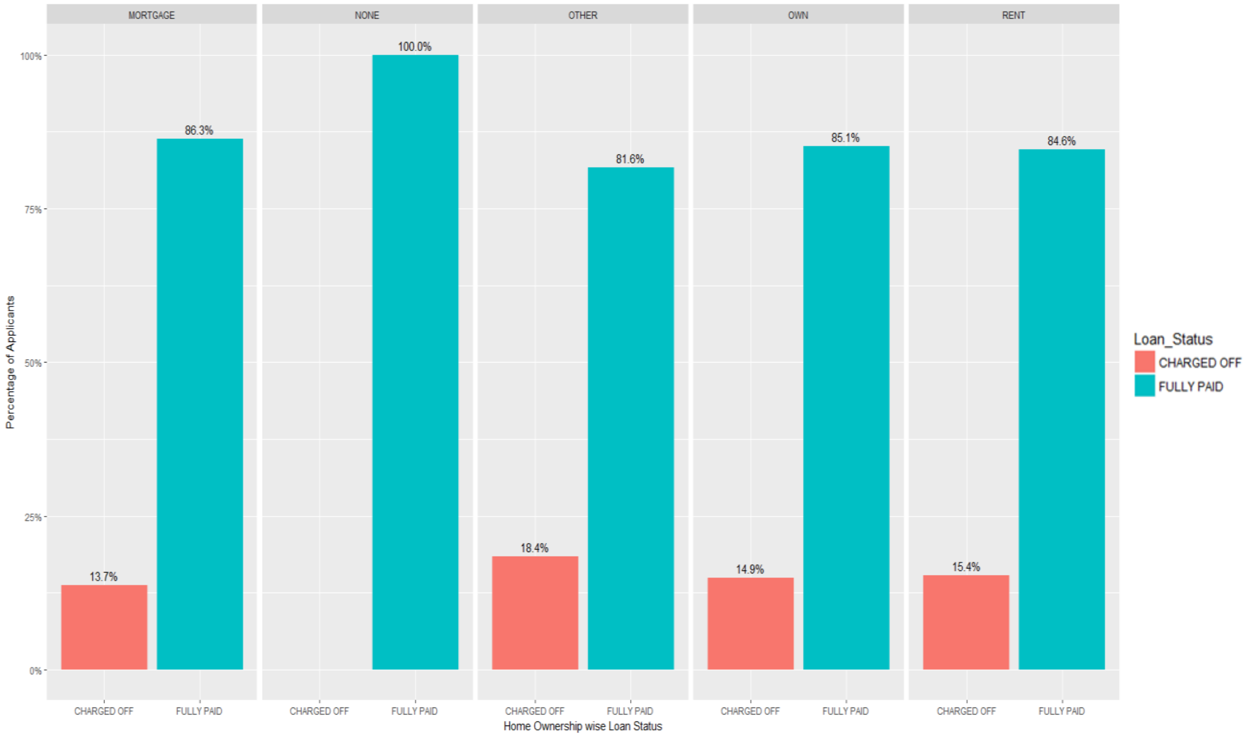
If we drill down to the subgrade and observe we can clearly see that borrowers with sub grades E1 to G5 have the highest chance of resulting in credit loss. **[With F5-48%, G3-44%, G2-41% representing the top 3 likely defaulters with their respective default percentage**]



LC Grade	Charged Off	Fully Paid	Grand Total	Fraud Percentage
A	602	9443	10045	5.99%
B	1425	10250	11675	12.21%
C	1347	6487	7834	17.19%
D	1118	3967	5085	21.99%
E	715	1948	2663	26.85%
F	319	657	976	32.68%
G	101	198	299	33.78%

Top Observation and Conclusions [5]

Plot40. Home Ownership versus Loan Status Analysis

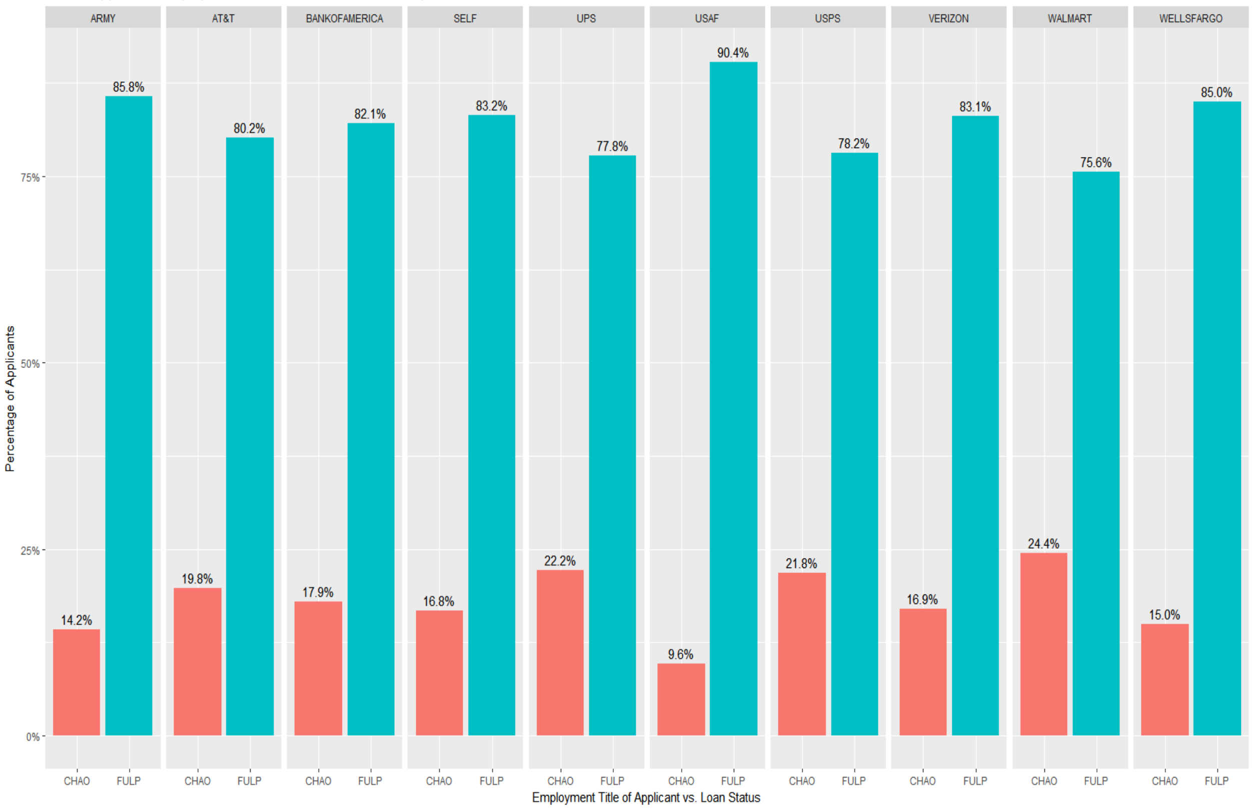


home_ownership	Charged Off	Fully Paid	Grand Total	Fraud Percentage
RENT	2839	15641	18480	15.36%
MORTGAGE	2327	14694	17021	13.67%
OWN	443	2532	2975	14.89%
OTHER	18	80	98	18.37%

Insight6- From the above plot it is clear that applicants who state home ownership as other have a higher chance of defaulting the loan payment.

From the table it is also clear that majority of loan applicants state rent or mortgage as home ownership. From which 15.4% and 13.7% result in credit loss respectively.

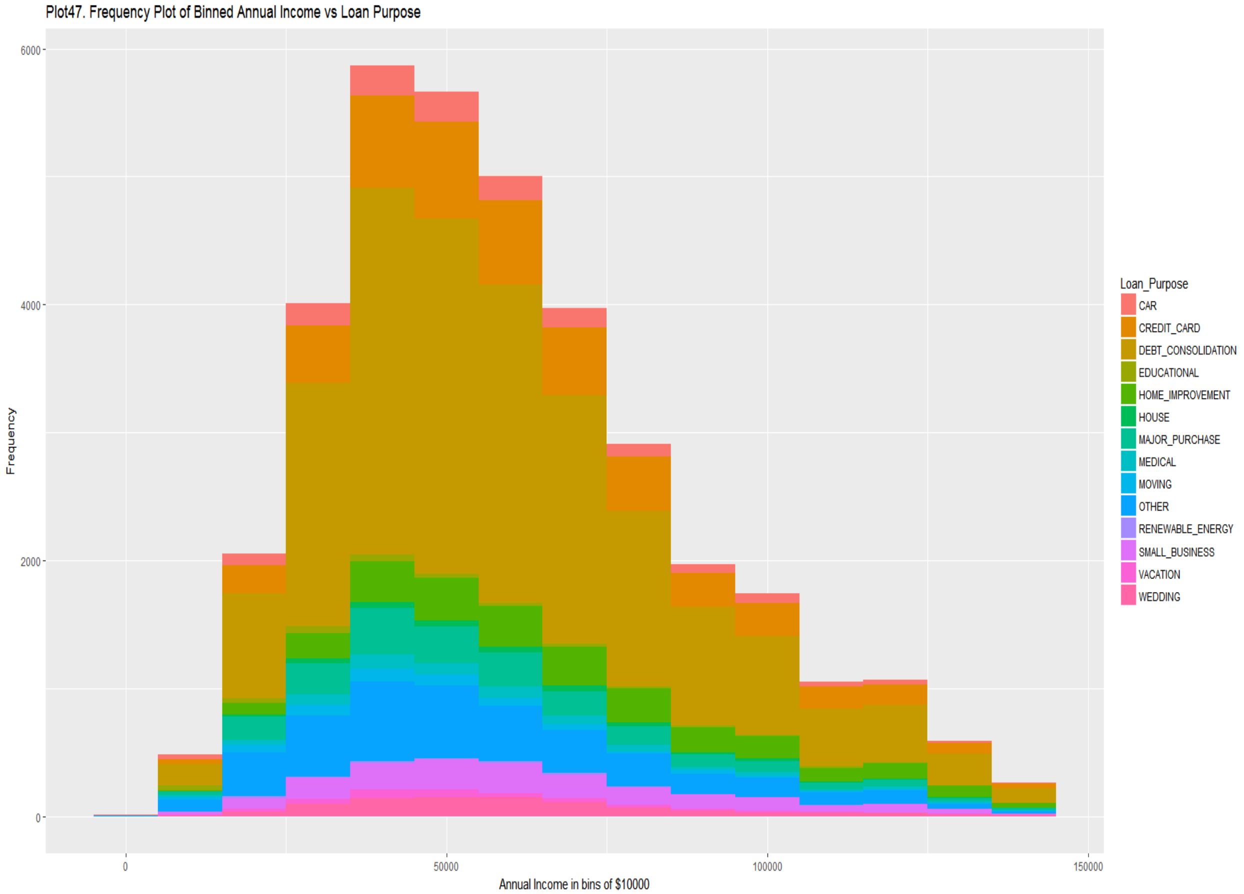
Plot45. Applicant's Employment Title versus Loan Status Analysis



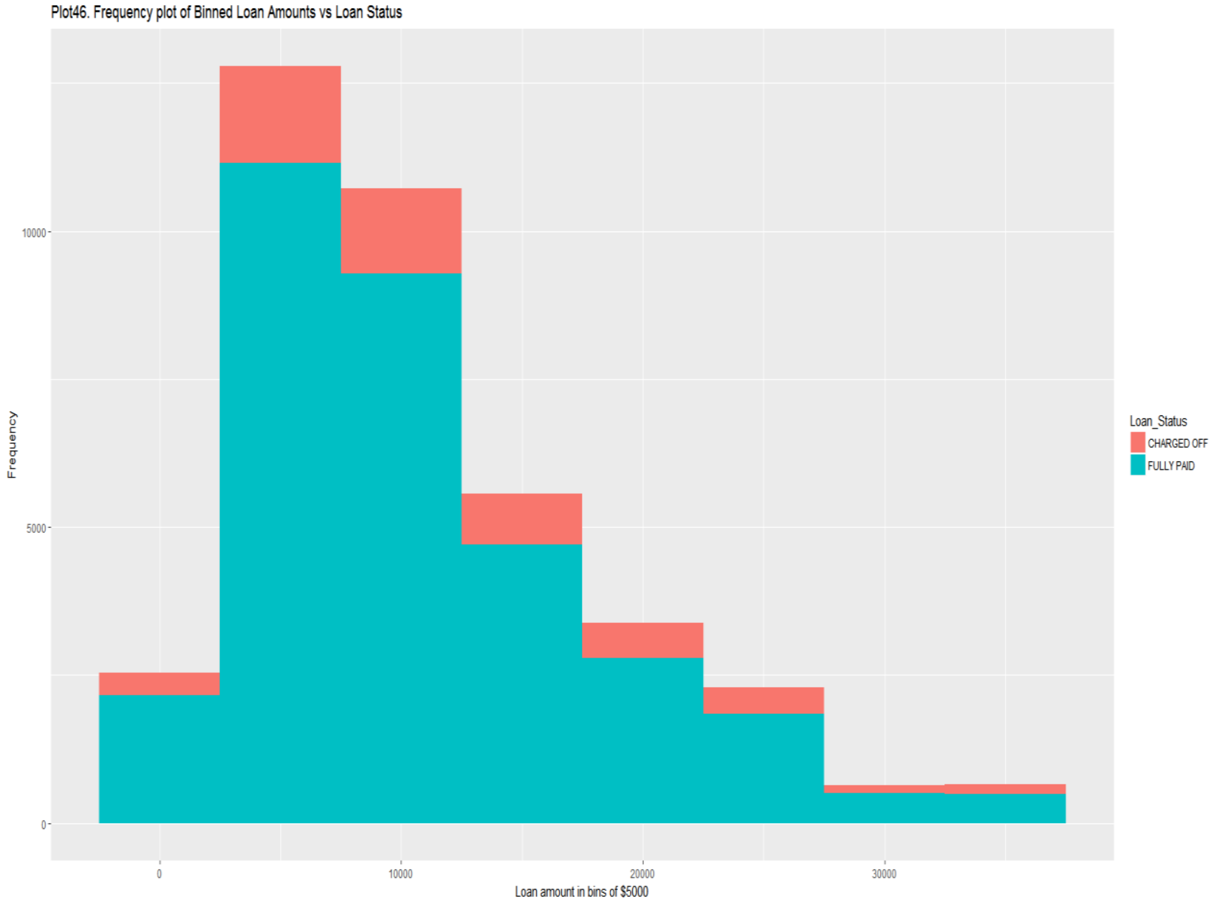
Insight7- From the above plot it is clear that applicants who state employment designation as Wallmart, United Parcel Service or US Postal Service have a 24.4%, 22.2% and 21.9% likelihood of defaulting.

Also, on observing the employment term versus loan status plot we see that majority of loan defaulters have an employment term of less than or equal to 1 year or beyond 10 years.

Top Observation and Conclusions [6]



Plot47. Frequency Plot of Binned Annual Income vs Loan Purpose



Plot46. Frequency plot of Binned Loan Amounts vs Loan Status

Insight8- On Analysis of the Annual Income after removal of outliers and loan purpose. It is clear that close to 68% of all loan applicants are have an annual income between $40,000-$70,000. In this segment Debt Consolidation and Small Businesses contribute to 53%.

Of the above 53% of loan applicants [27% of Small businesses loans and 16% of debt consolidation loans will lead to credit loss]

Insight9- On Analysis of Loan amount with respect to loan status it is clear that 58% of all borrowers apply for a loan less than $10,000.

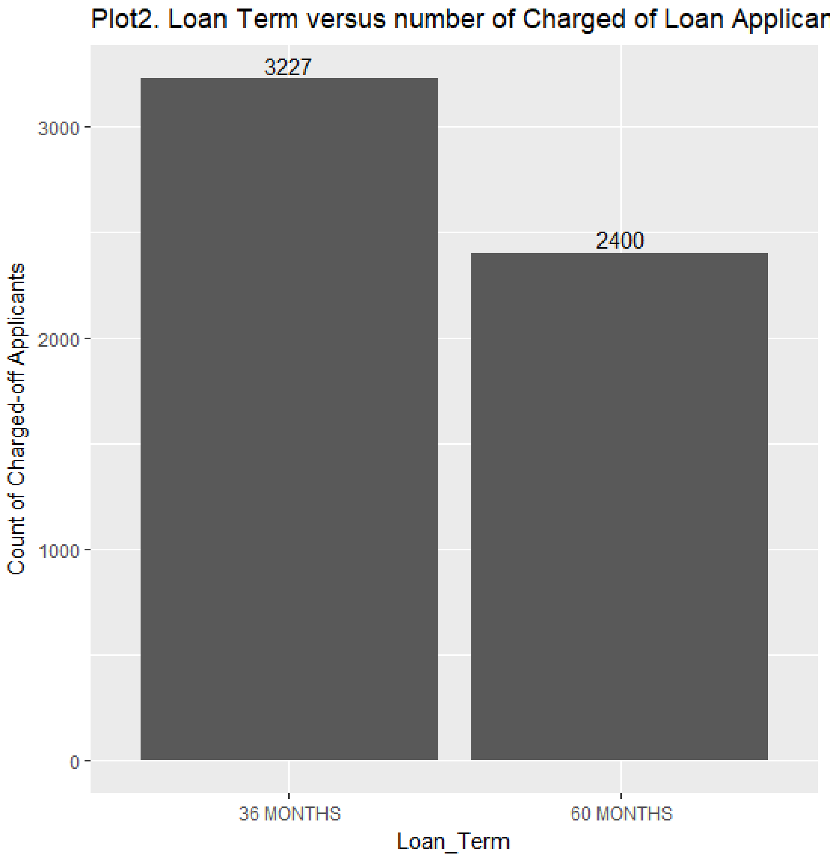
Of the above 58% of applicants 14% of applicants cause credit loss to the company.

Thank You

All Annexure Plots are available in the following slides

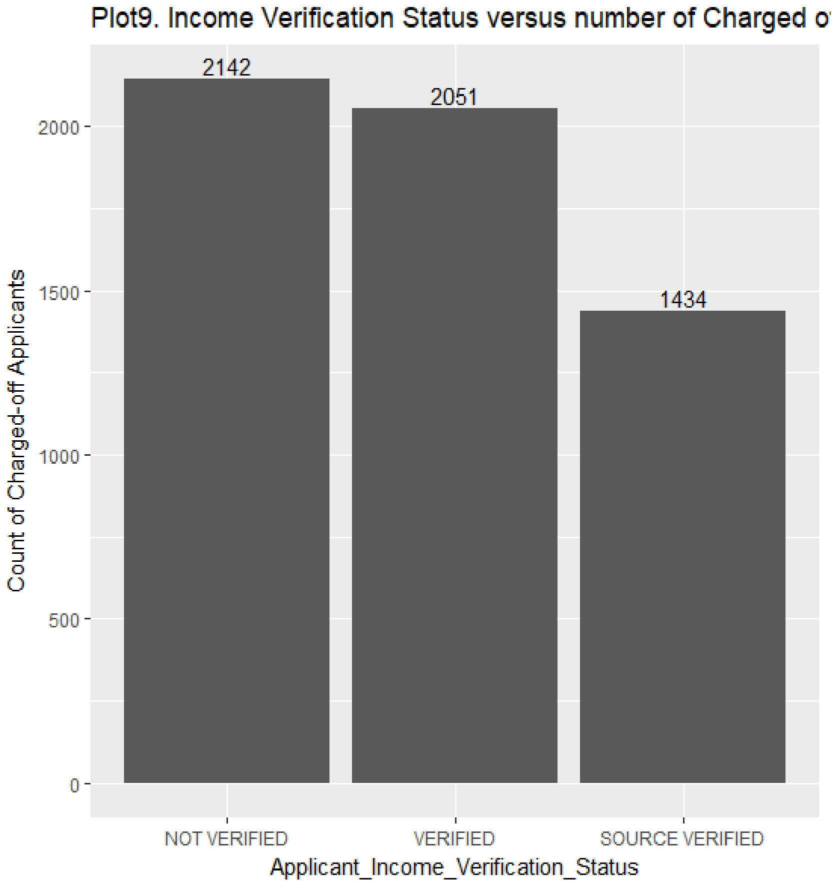
[A]Univariate Analysis of Categorical Variables

[A] 1. Analysis of Loan Term versus Count of Charged-Off Loans



- Observation [A]1. The number of charged-off loans for 3 years is higher than that of 5 years

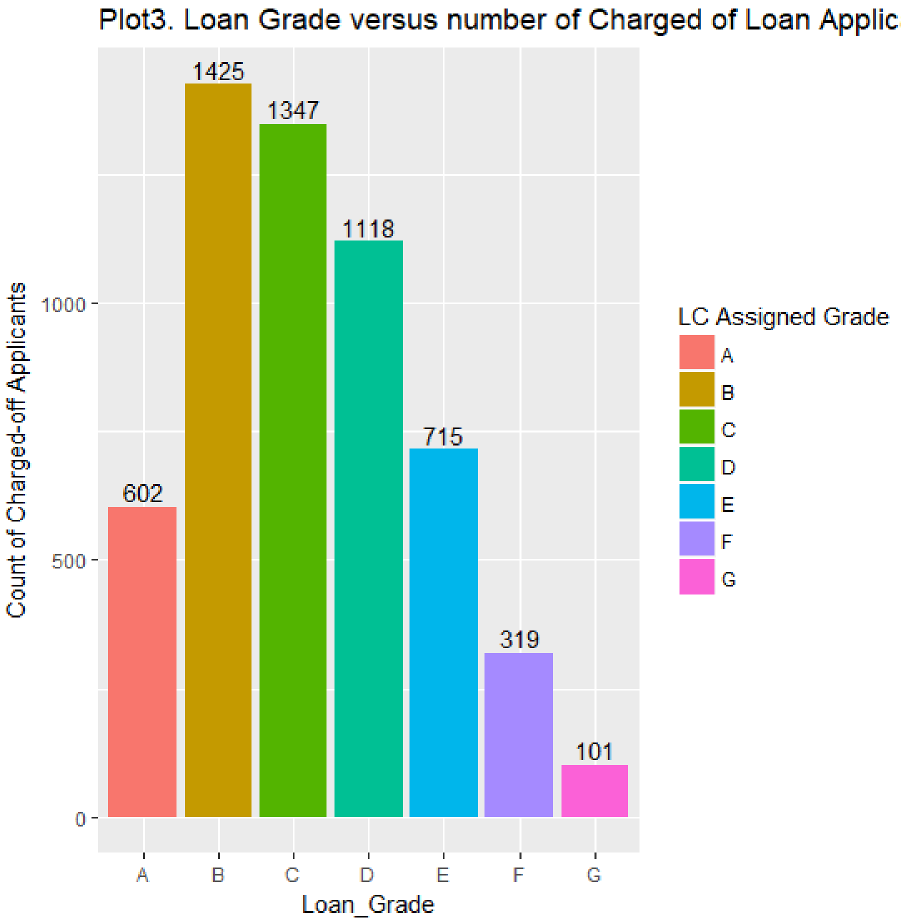
[A] 2. Analysis of Applicants Income Verification Status vs. Count of Charged-Off Loans.



- Observation [A]2. The number of charged-off loans for applicants with not verified income is higher than that of verified and source verified.

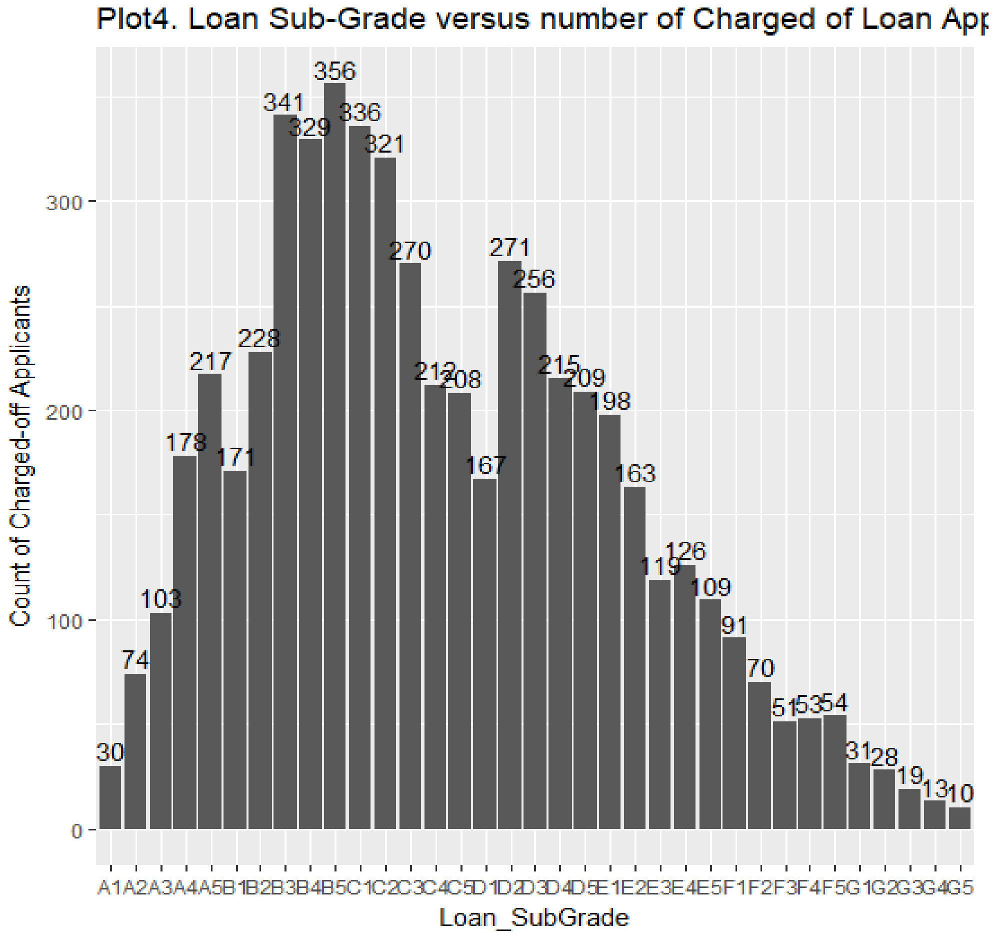
[A]Univariate Analysis of Categorical Variables

[A] 3. Analysis of LC Assigned grade versus Count of Charged-Off Loans



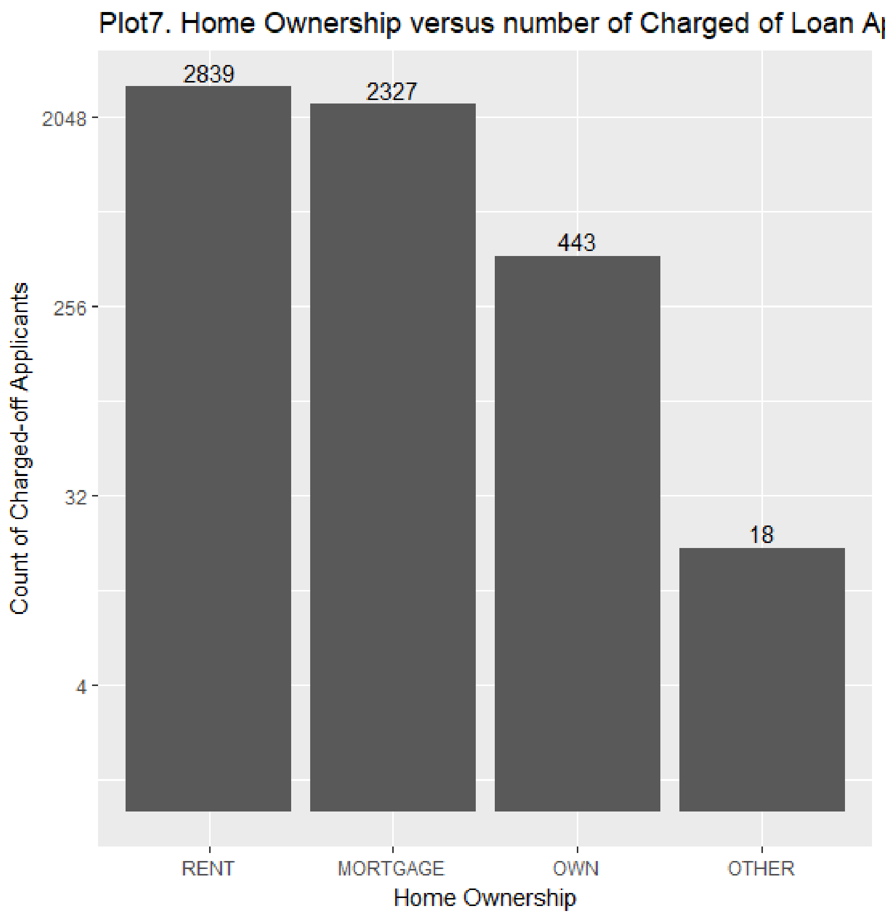
- Observation [A]3. LC Grades B,C and D contribute to the majority number of charged-off loans for

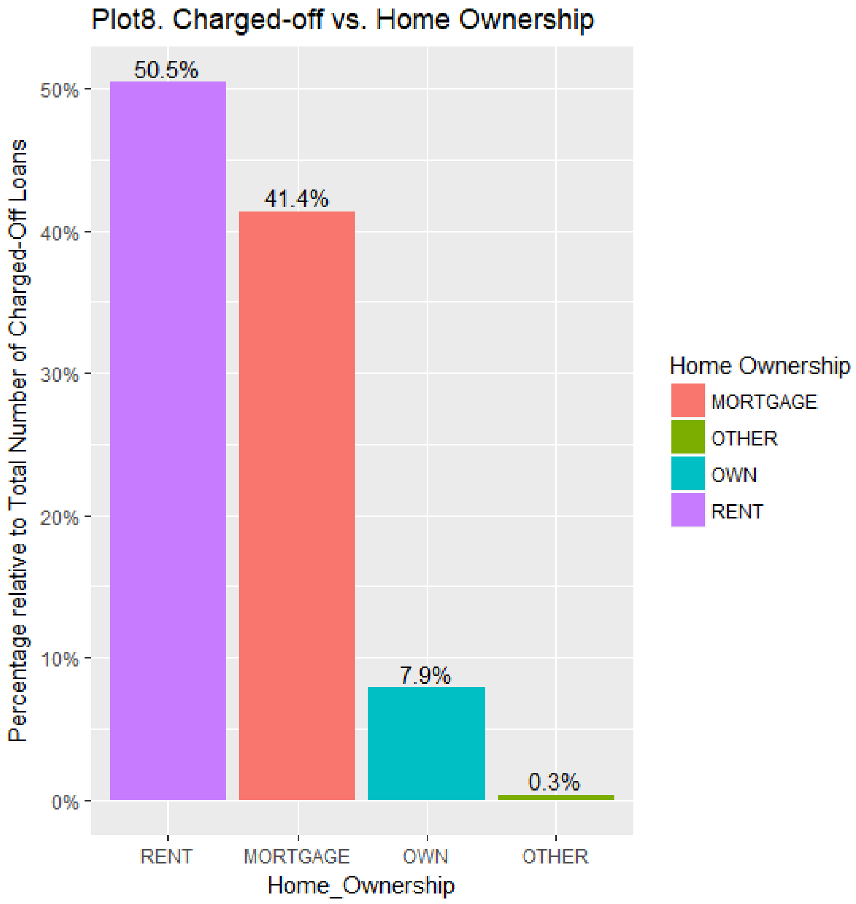
[A] 4. Analysis of LC Assigned sub-grade versus Count of Charged-Off Loans



Observation [A]4. The highest number of charged-off loans are in B3 ~ C3 and also D2~E1 sub grades.

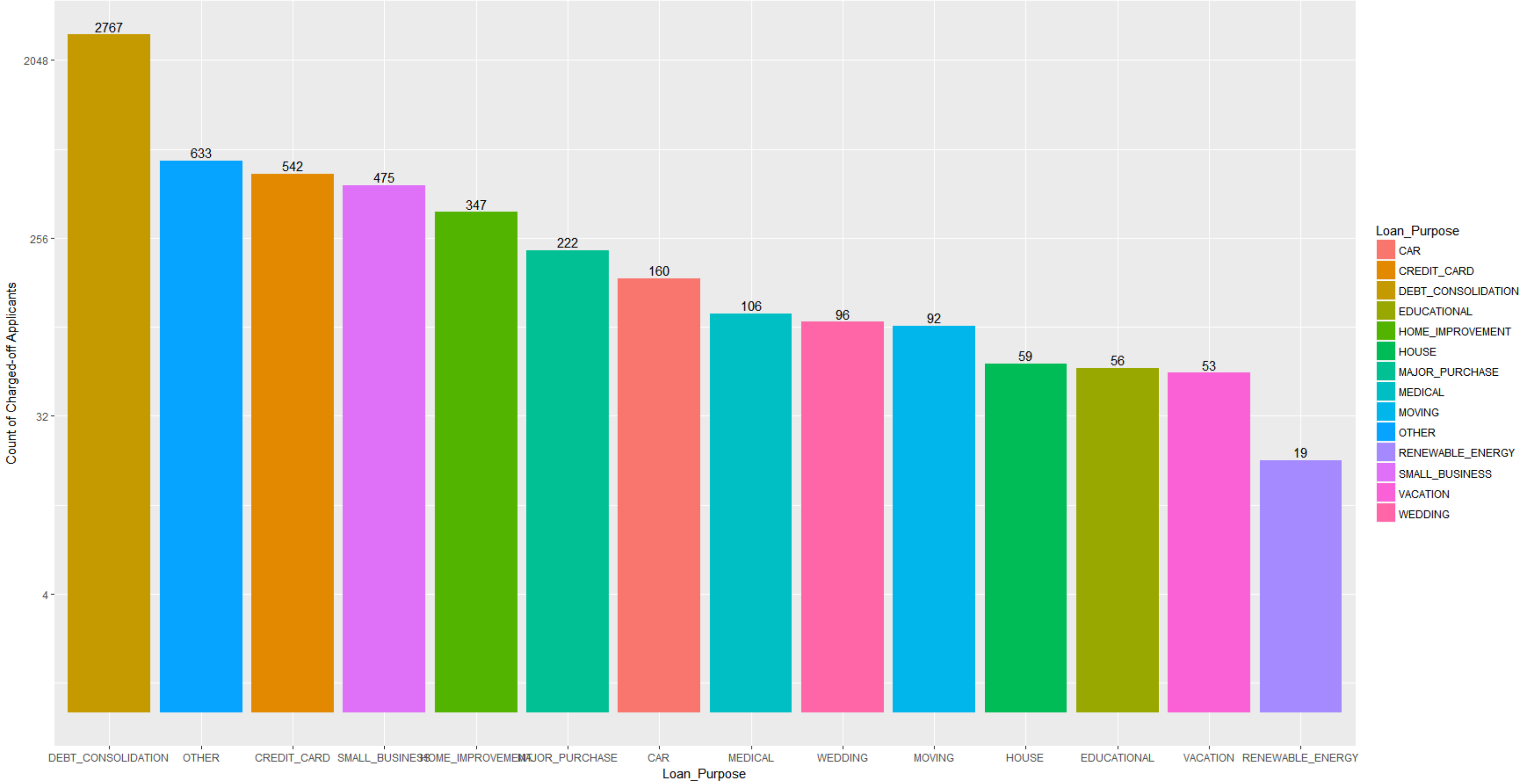
[A]5. Analysis of Home Ownership on Count of Charged-off Loans.





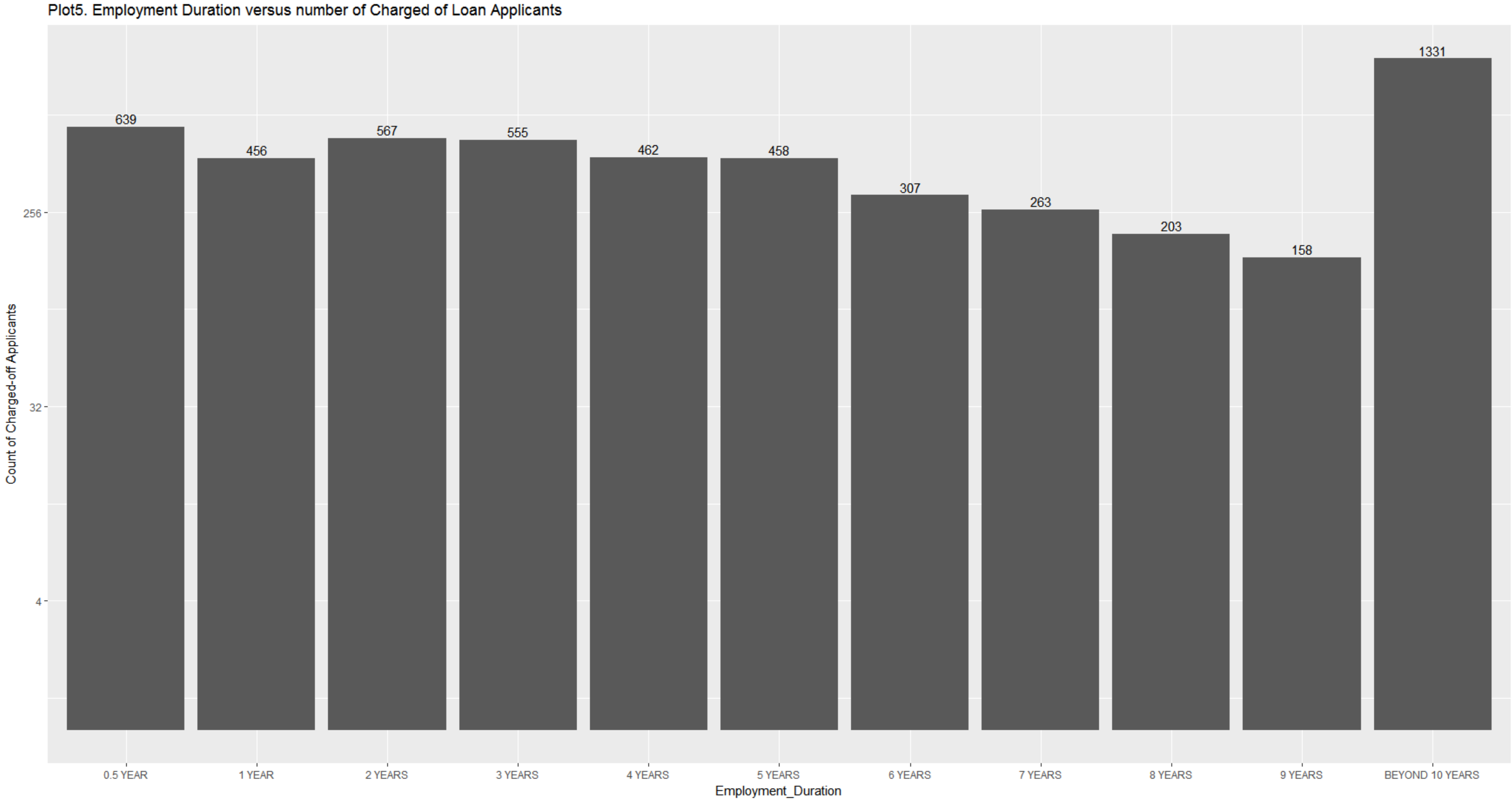
- Observation [A]5. People who live in rented accommodation constitute of 50.5% of the population

[A]6. Analysis of Loan_Purpose Vs. Count of Charged-Off Loans.


Plot10. Loan Purpose versus number of Charged of Loan Applicants

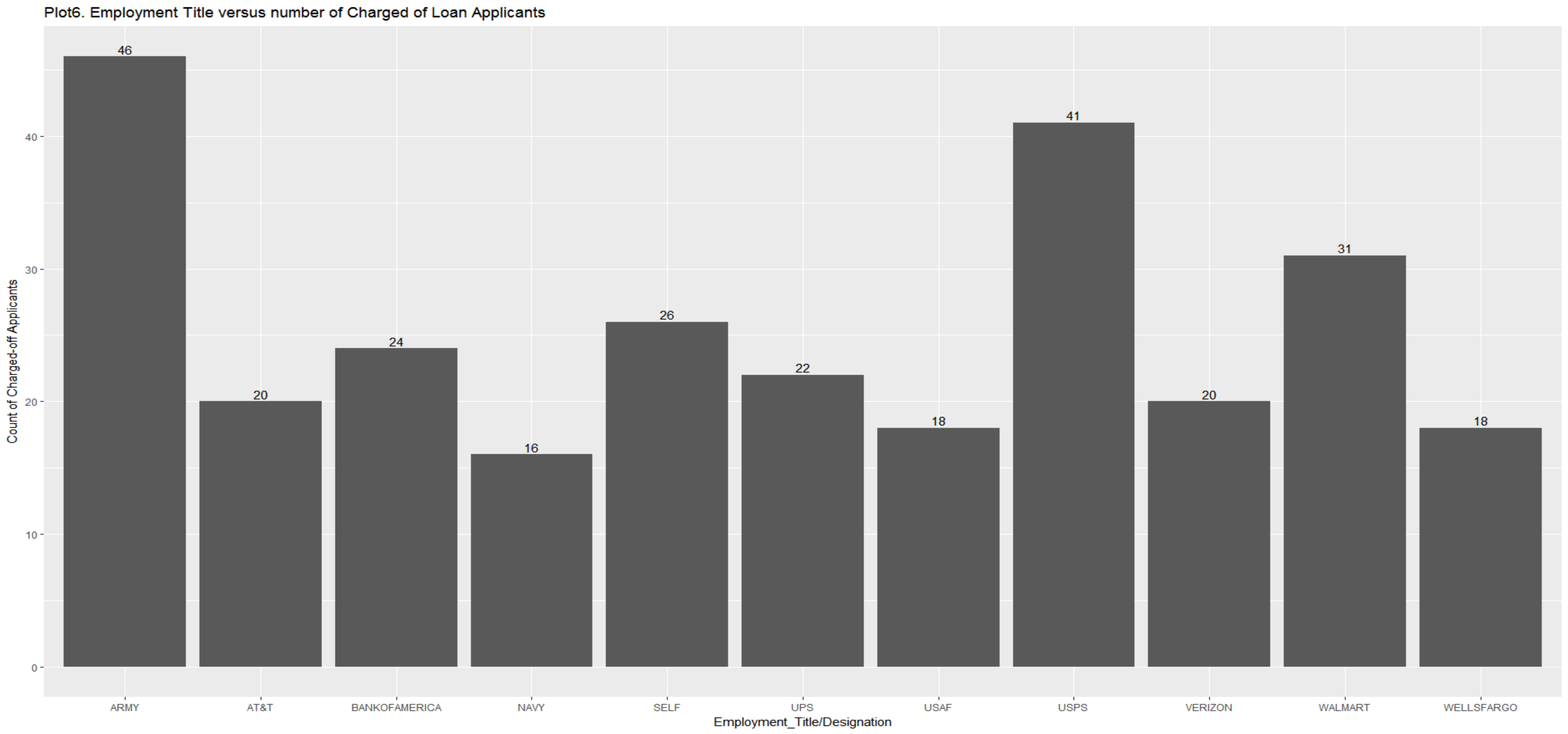
- Observation [A]6. Observations: Note that this is Log 2 Power plot, and yet the purpose of “Debt Consolidation” is prominently high. A high number of “Other” category also tells us that our data collection method is inadequate. We must add more categories to the selection drop down menu which is used at the time of applying for loans.

[A]7. Analysis of Employment Term versus Count of Charged-Off Loans



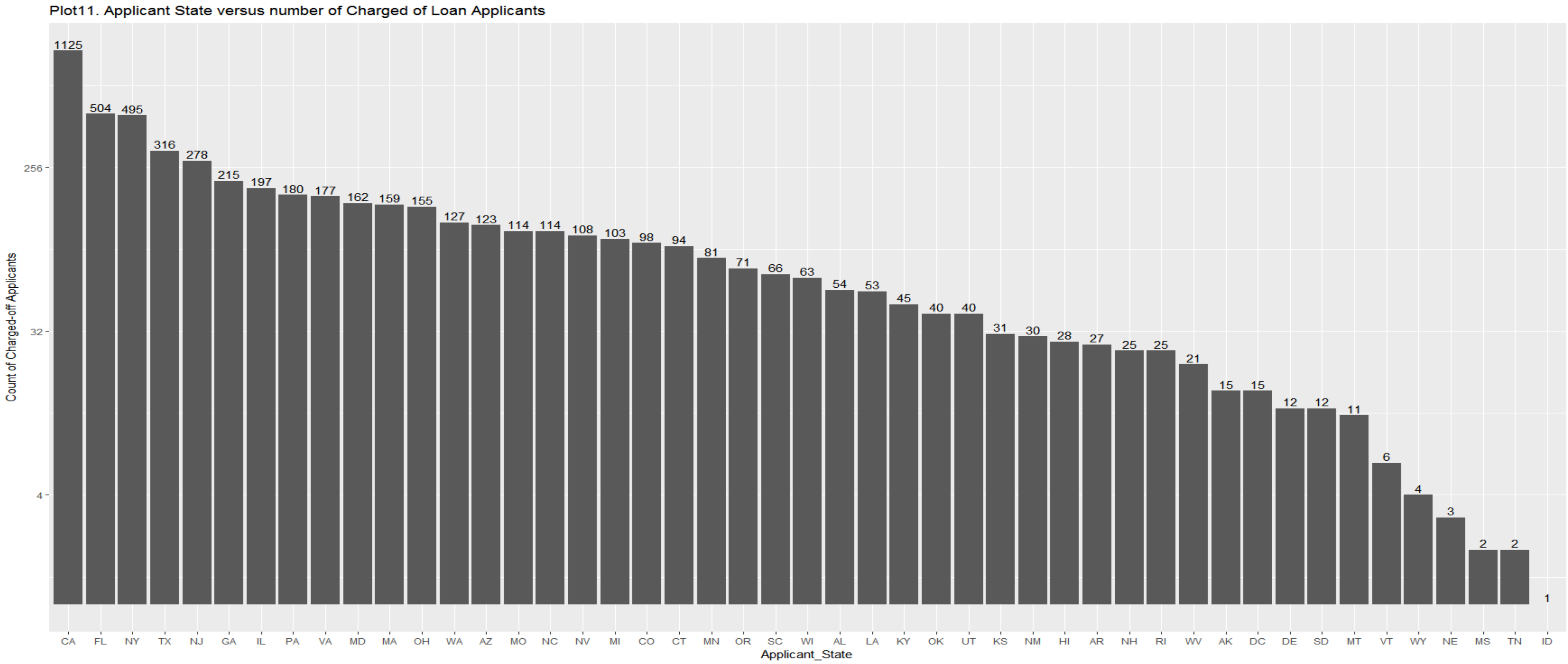
- Observation [A]7. Charged off loans has a decreasing trend with respect to tenure. However the number of Charged Off loans within the first and beyond 10 years is significantly high.

[A]8. Histogram of Employer Title vs. Count of Charged-Off Loans



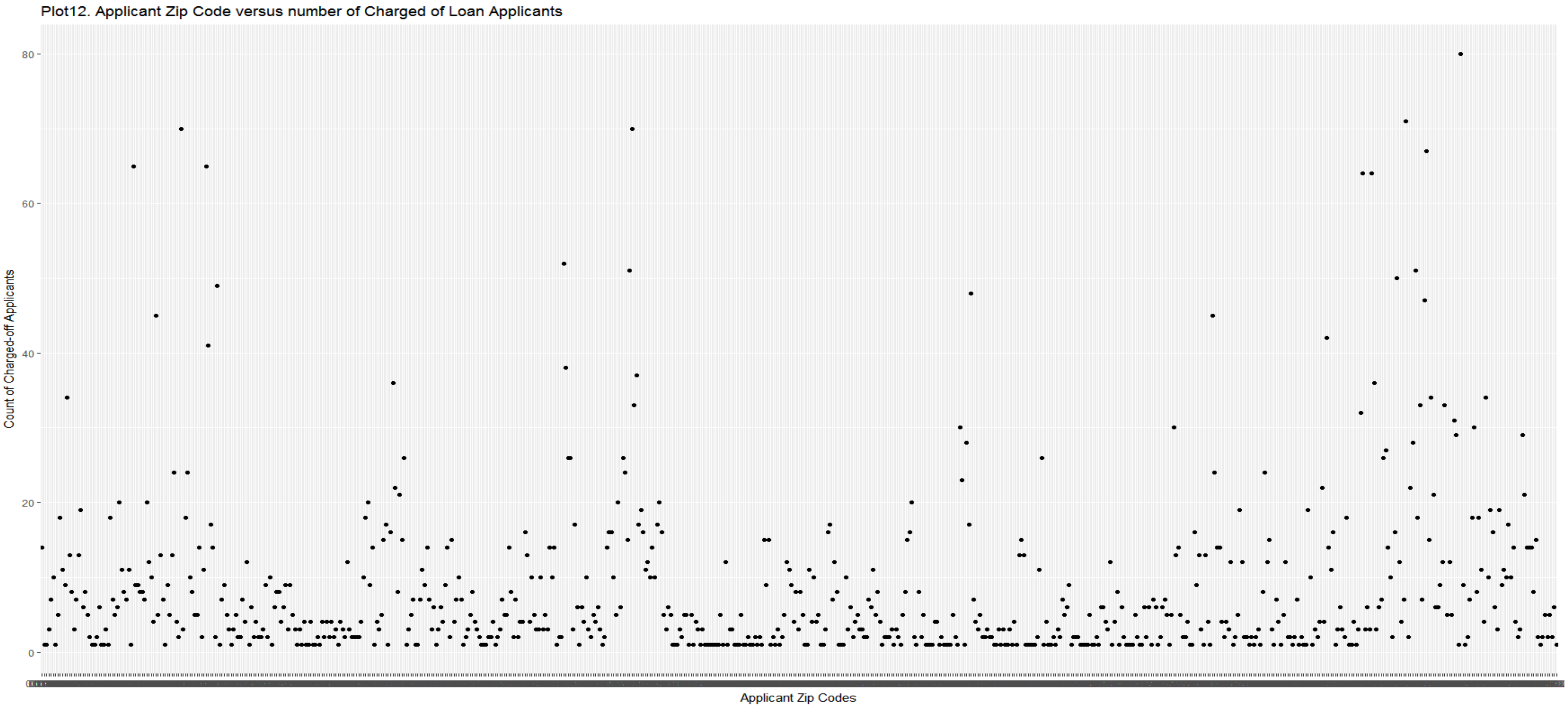
- Observation [A]8. The above plot shows the organizations along with the count of defaulting applicants from each organization.

[A]9. Analysis of Applicant Residence State Vs Count of Charged-Off Loans


Plot11. Applicant State versus number of Charged of Loan Applicants

- Observation [A]9. California, Florida, New York, Texas, New Jersey are top 5 affected states with loan status: “Charged Off”

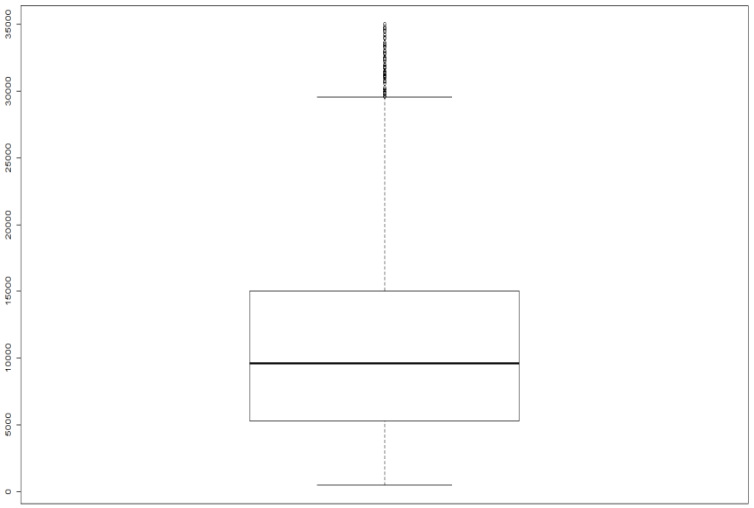
[A]10. Analysis of Applicant Zip-code Vs Count of Charged-Off Loans

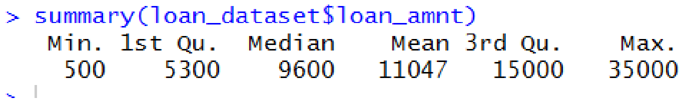


- Observation [A]10. Since zip codes are in ascending order in X axis we can see that no of “CHARGED OFF” loans are particularly high between 8xxxx and 9xxxx.

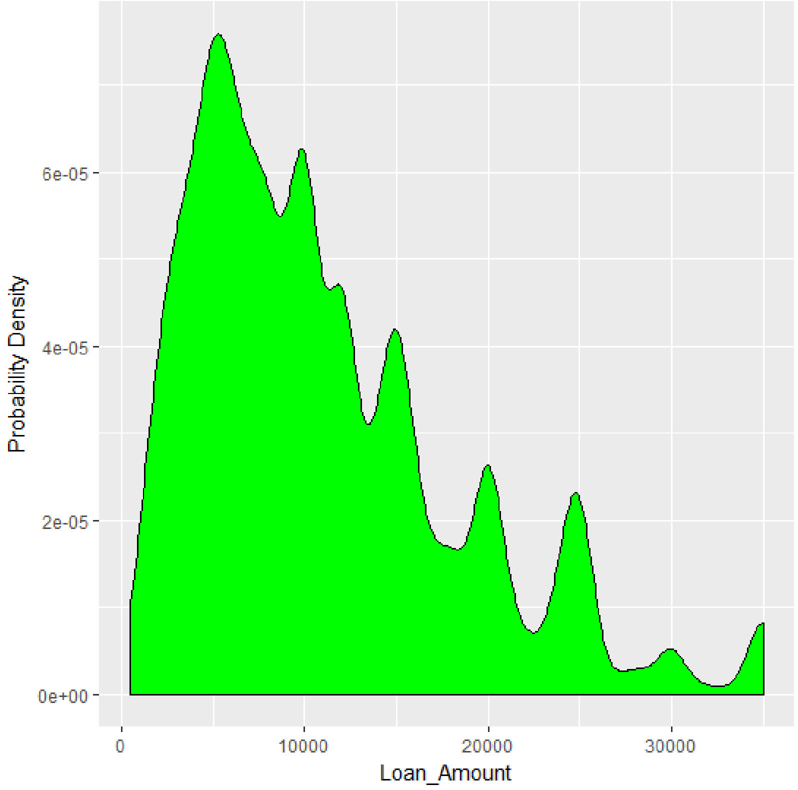
[B]Univariate Analysis of Quantitative Variables

[B] 1. Loan amount boxplot and summary and Analysis

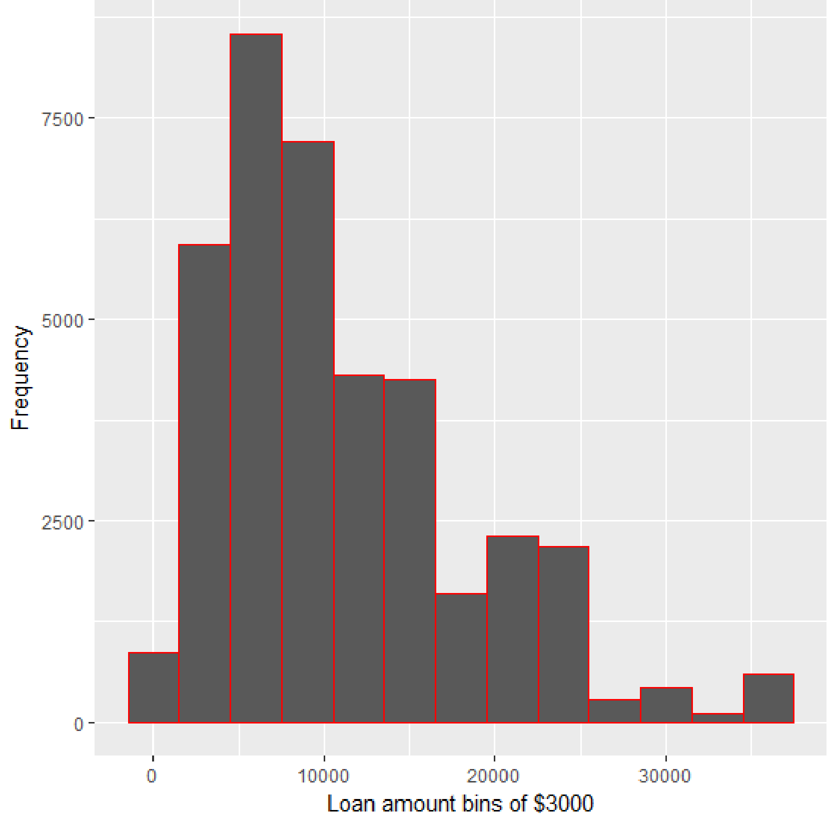




```
> summary(loan_dataset$loan_amnt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    500    5300    9600   11047   15000   35000 
```



Plot13. Density plot of Loan Amount

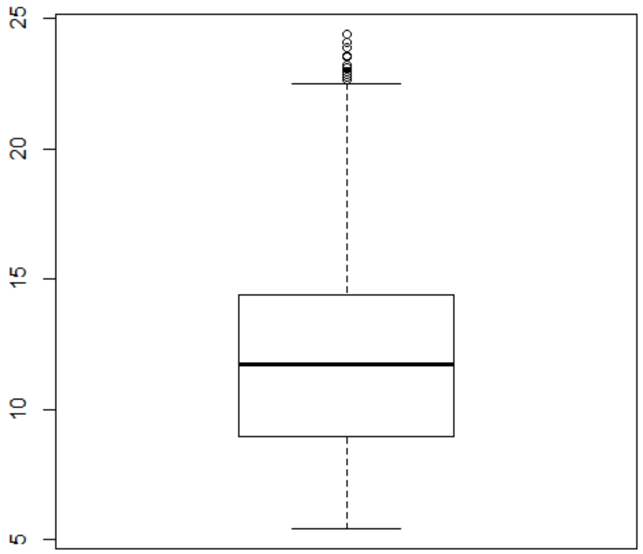


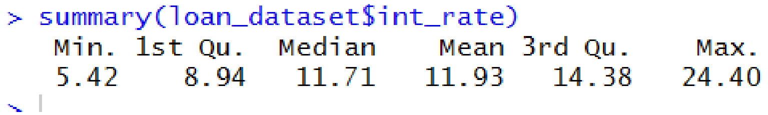
Plot14. Frequency Plot of Binned Loan Amounts

- Observation [B]1. From the box plot & the distribution, we can clearly see that it is approximately a Gaussian distribution but there are outliers at the far end.

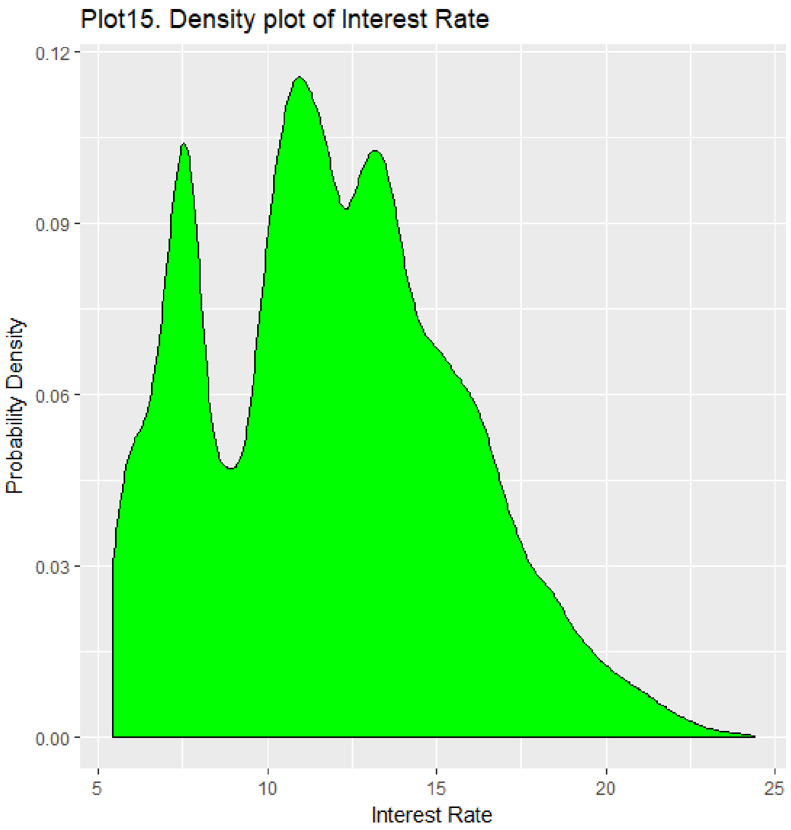
[B]Univariate Analysis of Quantitative Variables

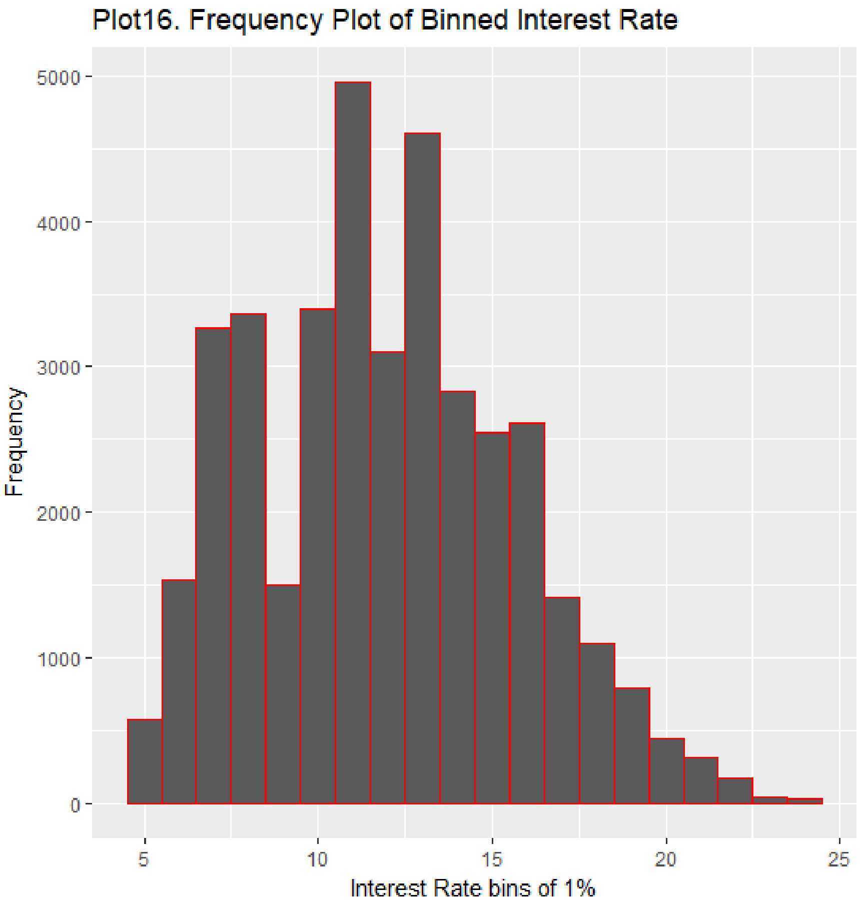
[B] 2. Interest Rate boxplot and summary and Analysis





```
> summary(loan_dataset$int_rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5.42    8.94   11.71   11.93   14.38   24.40 
```

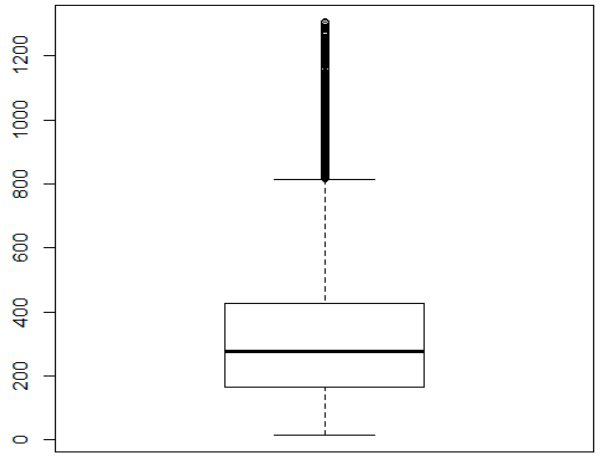


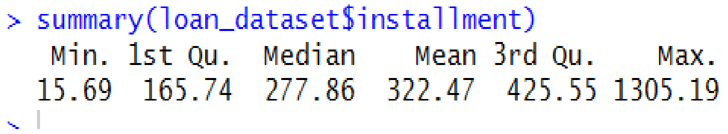


- Observation [B]2. From the box plot & the distribution, we can see that the count of loans with interest rate spikes between 7-8% and again between 11-13%.

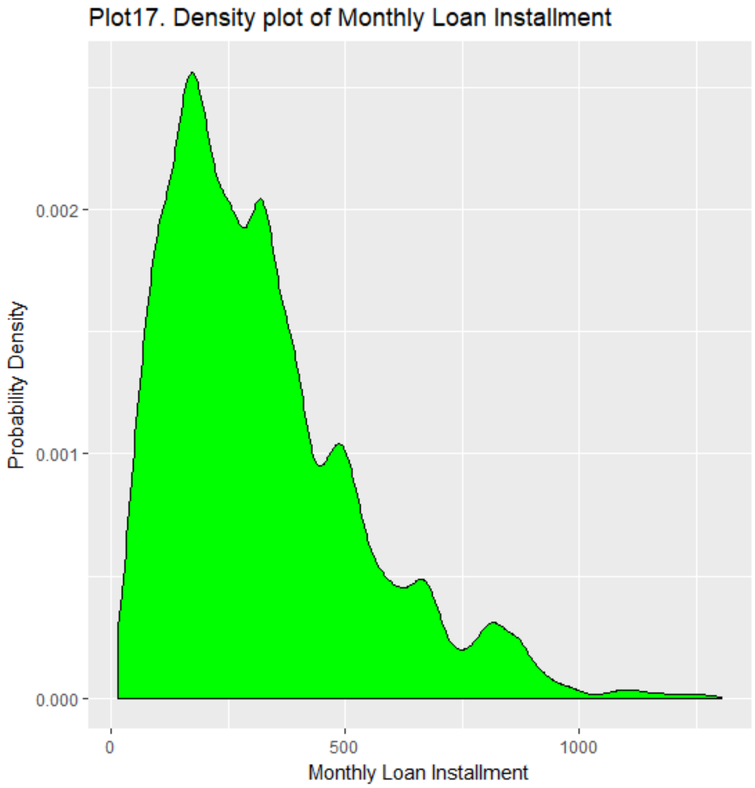
[B]Univariate Analysis of Quantitative Variables

[B] 3. Loan Installment boxplot and summary and Analysis

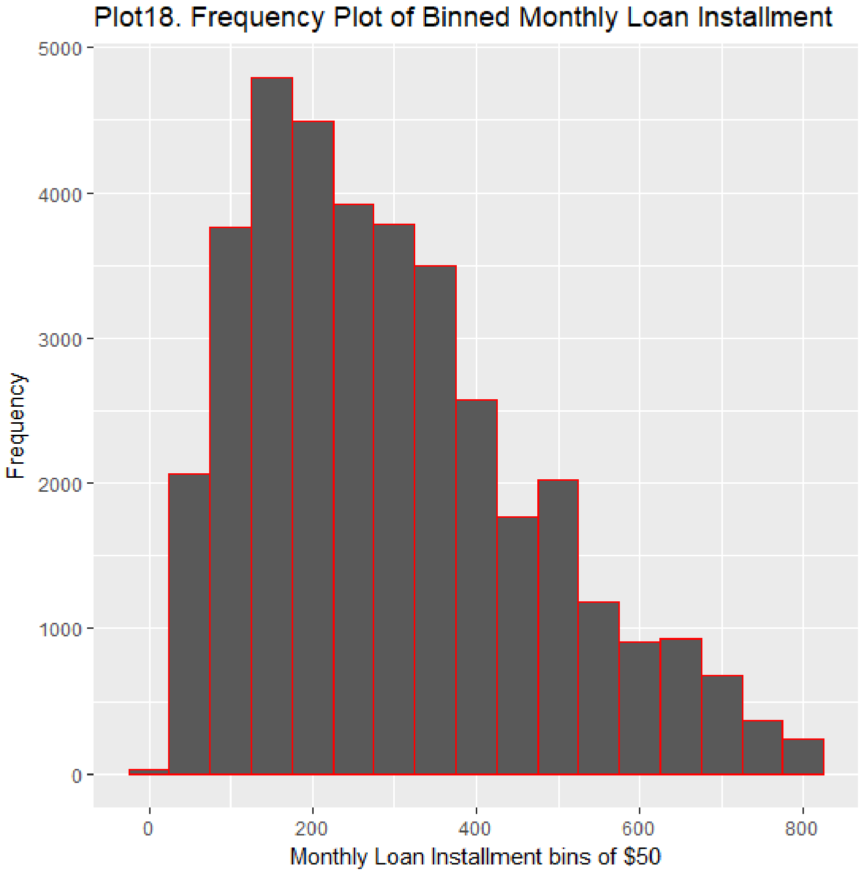




```
> summary(loan_dataset$installment)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.69  165.74  277.86  322.47  425.55 1305.19 
>
```



Plot17. Density plot of Monthly Loan Installment

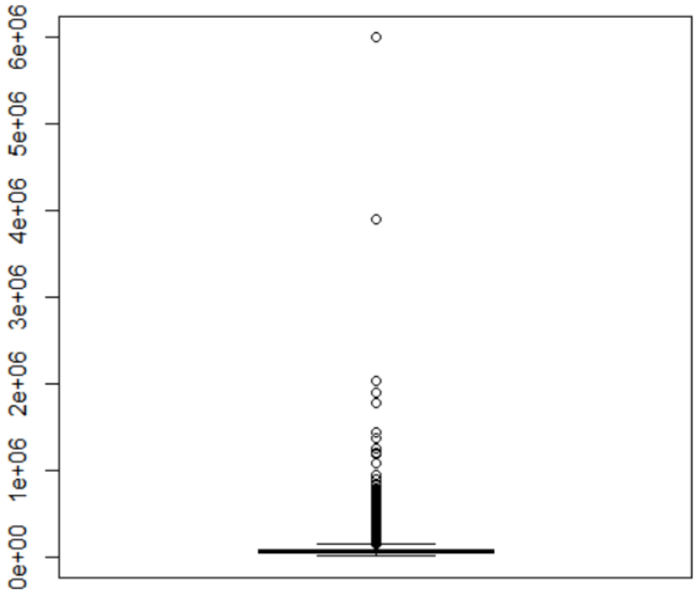


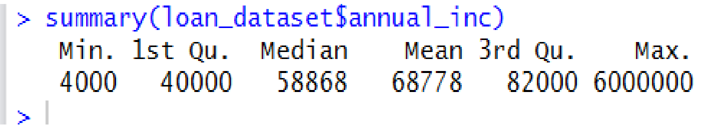
Plot18. Frequency Plot of Binned Monthly Loan Installment

- Observation [B]3. From the box plot & the distribution, we see that there are significant outliers on the higher side of monthly loan instalment. Using the 95 percentile rule we eliminate the outliers and set the upper bound to $800. On binned analysis we identify that the 68% of all monthly instalments are below $400.

[B]Univariate Analysis of Quantitative Variables

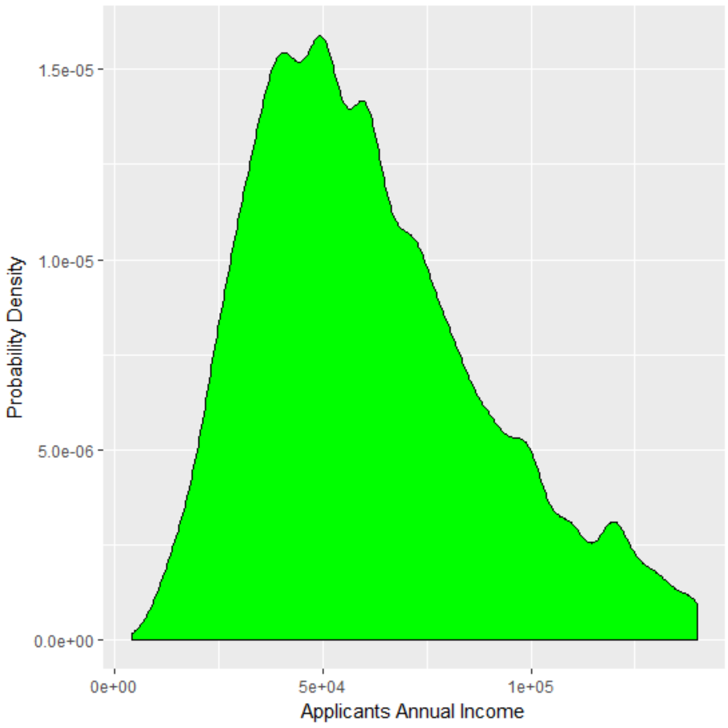
[B] 4. Annual Income boxplot and summary and Analysis



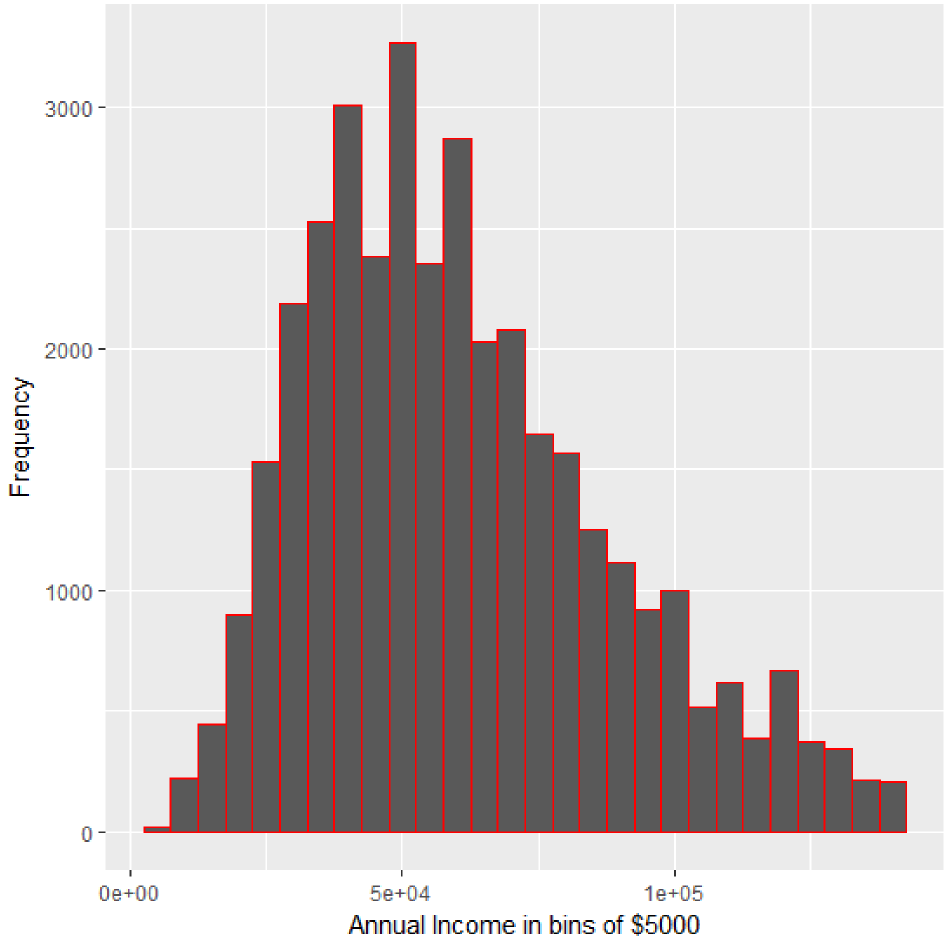


```
> summary(loan_dataset$annual_inc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   4000   40000   58868   68778   82000 6000000 
> 
```

Plot19. Density plot of Applicants Annual Income



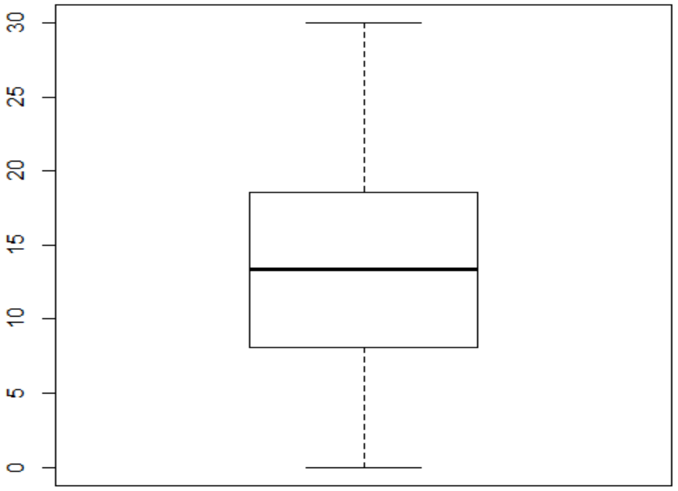
Plot20. Frequency Plot of Binned Annual Income

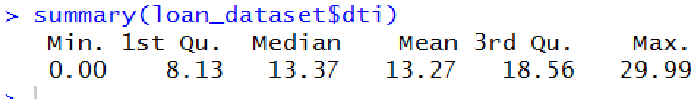


- Observation [B]4. From the box plot, we see that there are significant outliers on the higher side of annual income. Using the 95 percentile rule we eliminate the outliers and set the upper bound to $140000. On binned analysis we identify that majority are within 40-80 thousand $ annual Income.

[B]Univariate Analysis of Quantitative Variables

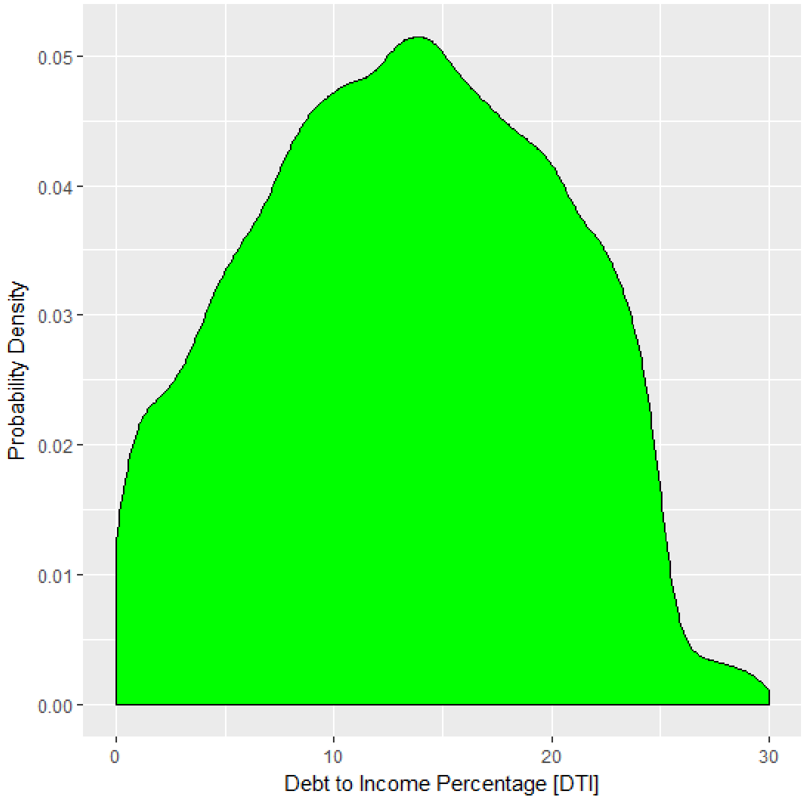
[B] 5. DTI boxplot and summary and Analysis



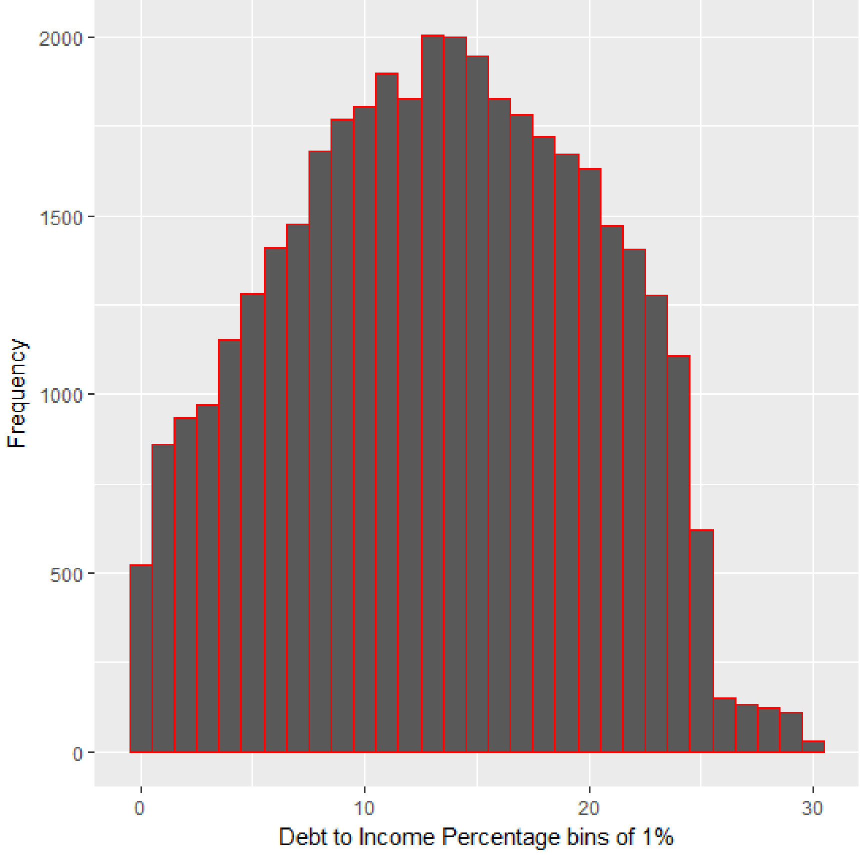


```
> summary(loan_dataset$dti)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    8.13   13.37   13.27   18.56   29.99 
```

Plot21. Density plot of Debt to Income Percentage [DTI]

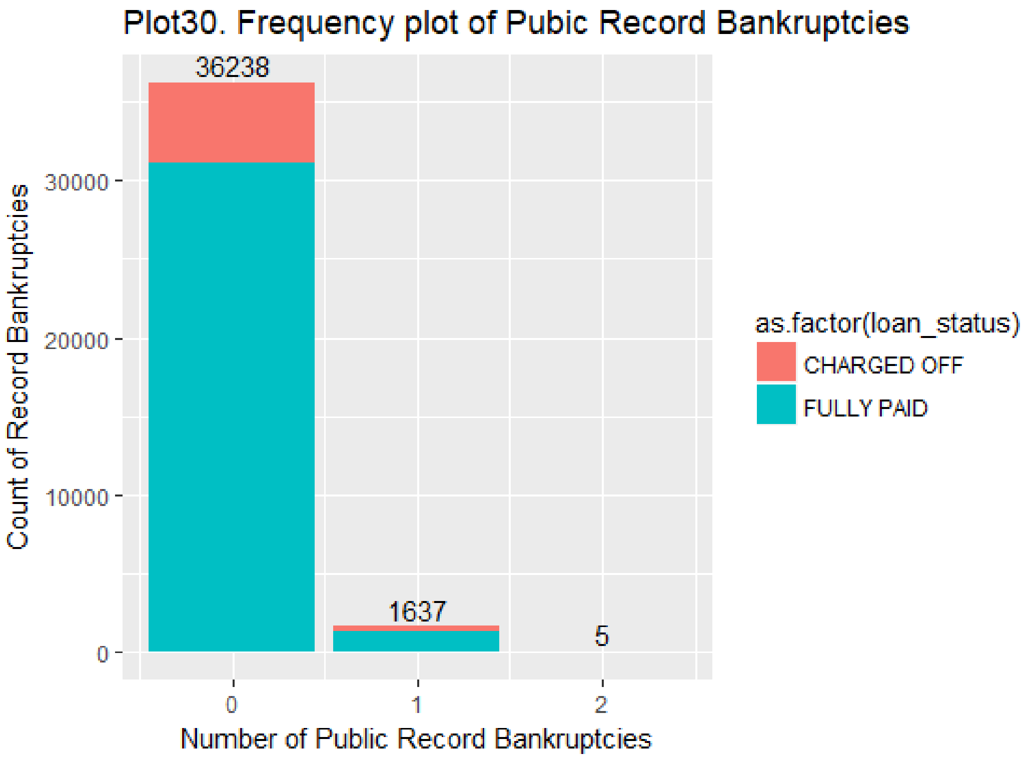


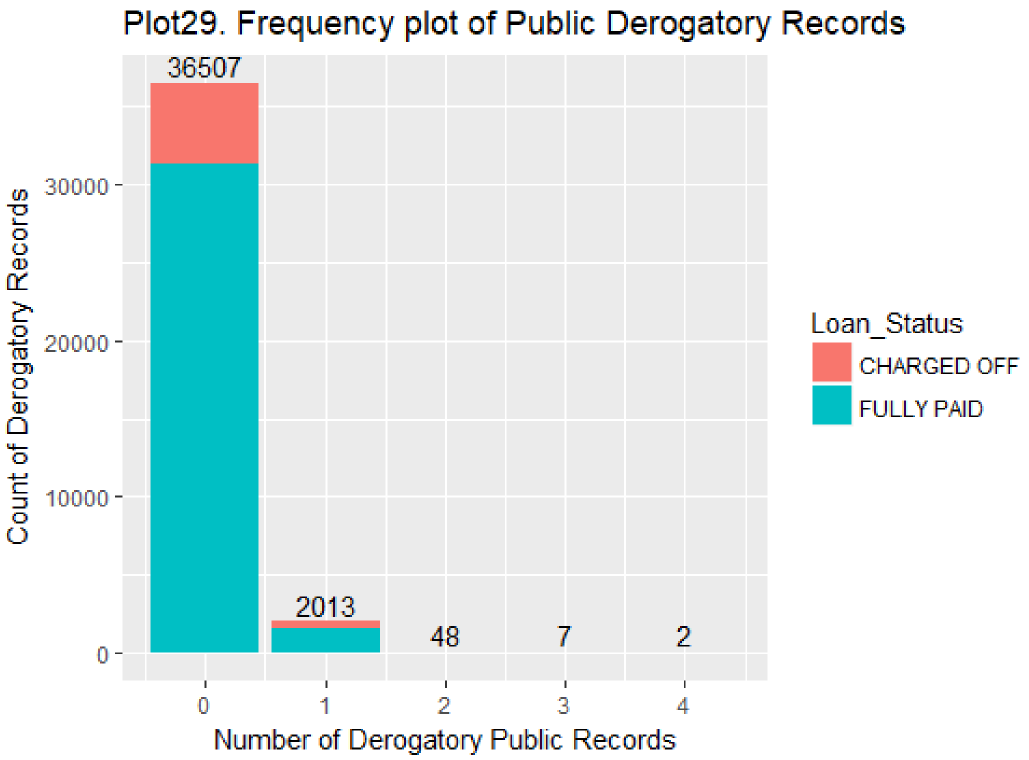
Plot22. Frequency Plot of Binned Debt to Interest Percentage



- Observation [B]5. From the box plot, and the distribution we can see that DTI is almost normally distributed with the between 13-14%

[B] 6. Public Record Bankruptcies and number of derogatory records


Plot30. Frequency plot of Pubic Record Bankruptcies

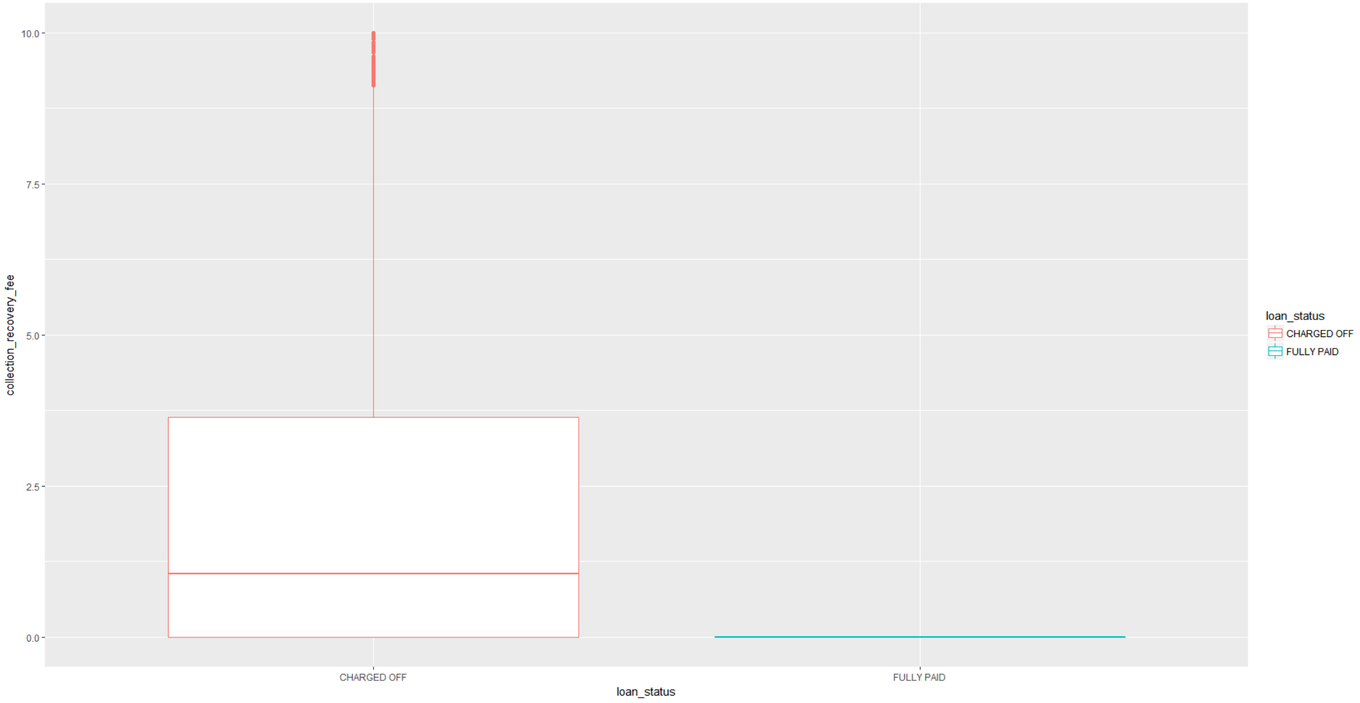

Plot29. Frequency plot of Public Derogatory Records

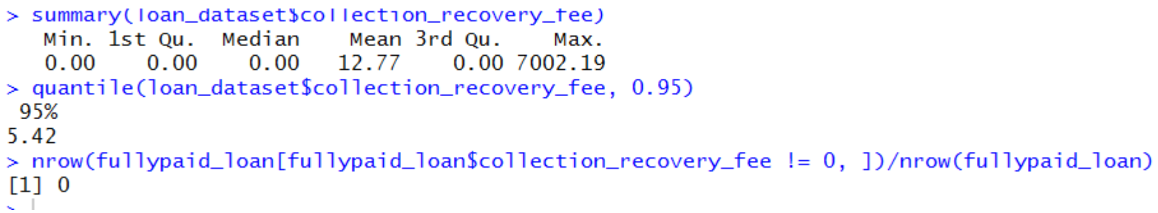
Observation [B] 6.

There are hardly any derogatory public records. Public Record of Bankruptcies column is sparsely populated too. We will not use these for Analysis.

[B] 7. Recovery Fees Analysis

Plot26. Box Plot Comparison of Collection Revovery Fees vs. Loan Status



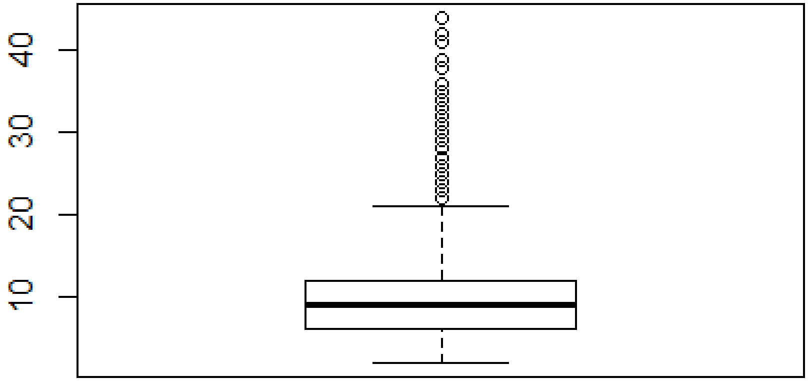


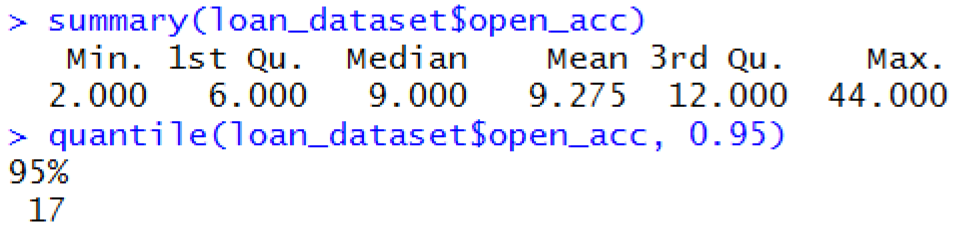
```
> summary(loan_dataset$collection_recovery_fee)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.00    0.00   12.77    0.00 7002.19 
> quantile(loan_dataset$collection_recovery_fee, 0.95)
 95% 
5.42 
> nrow(fullypaid_loan[fullypaid_loan$collection_recovery_fee != 0, ])/nrow(fullypaid_loan)
[1] 0
```

[B]7. Recovery Fees is not valid for Fully paid loans therefore we will not include this in out analysis.

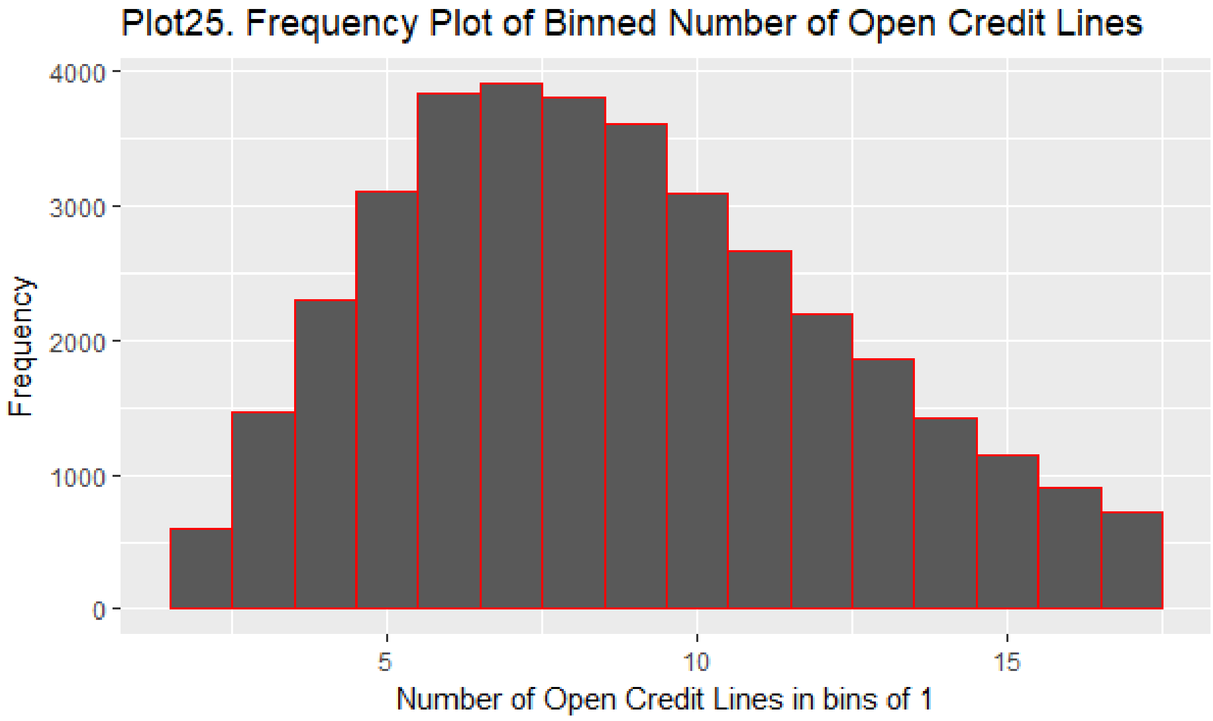
Revolving line utilization Rate is fairly normally distributed

[B] 8. OpenCredit Lines Analysis



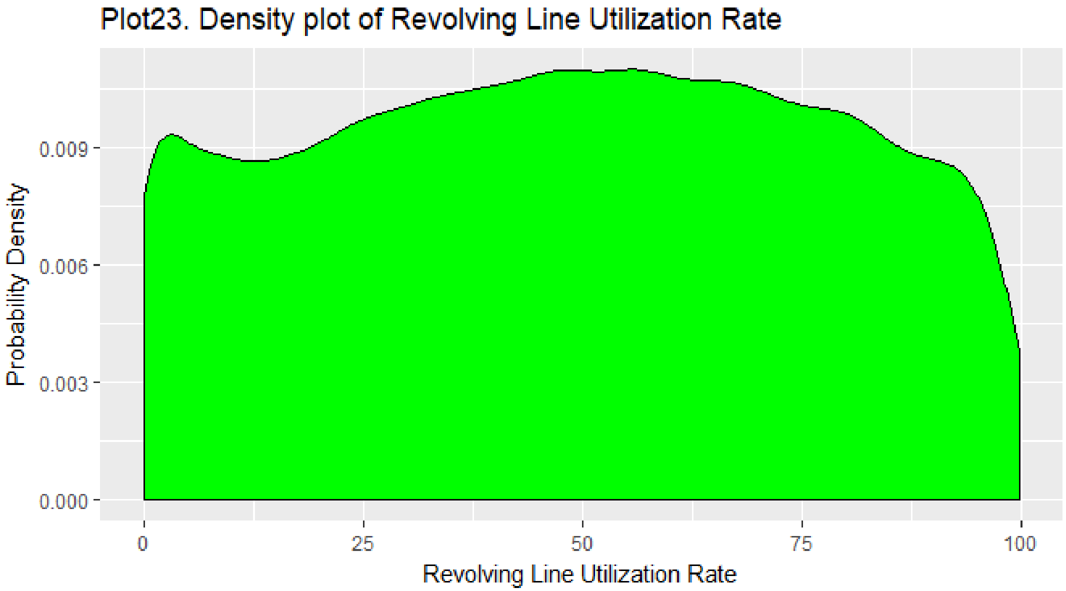


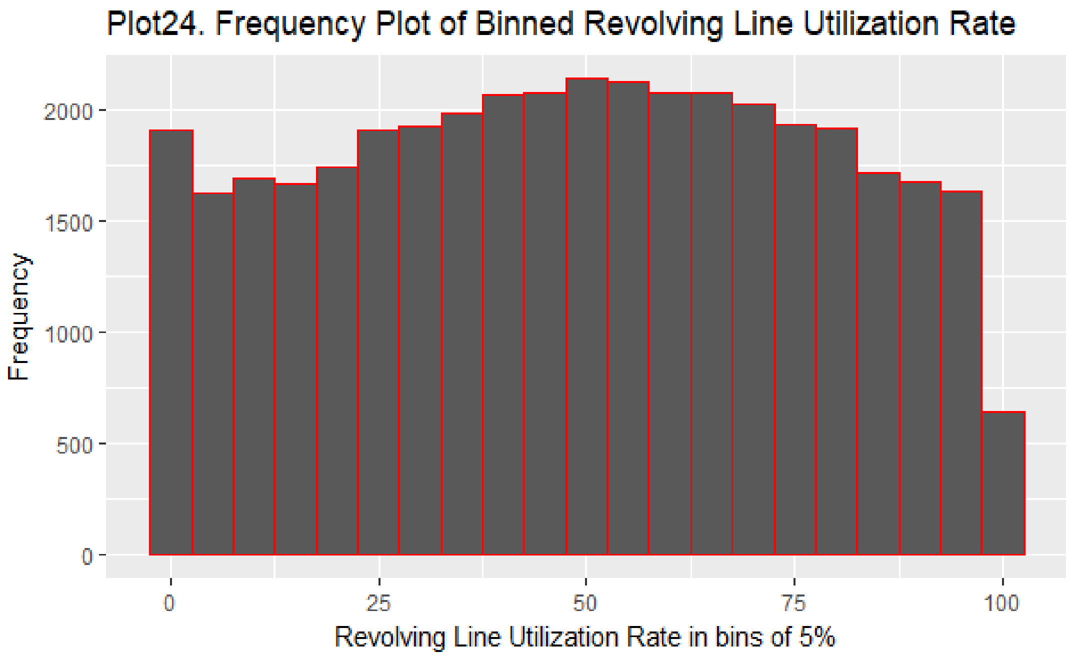
```
> summary(loan_dataset$open_acc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   6.000   9.000   9.275  12.000  44.000 
> quantile(loan_dataset$open_acc, 0.95)
95% 
 17 
```


Plot25. Frequency Plot of Binned Number of Open Credit Lines

[B] 8. After Addressing the outliers of number of open credit lines we can see that the majority number of loan applications is aggregated around 6-8 open accounts

[B] 9. Revolving Utilization Rate summary and Analysis

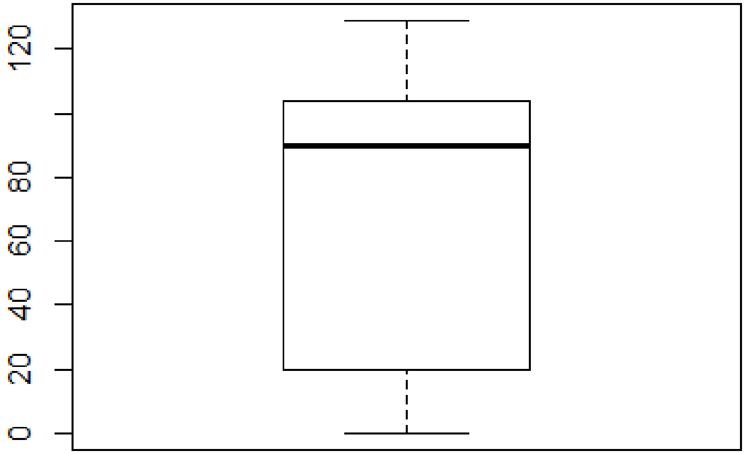


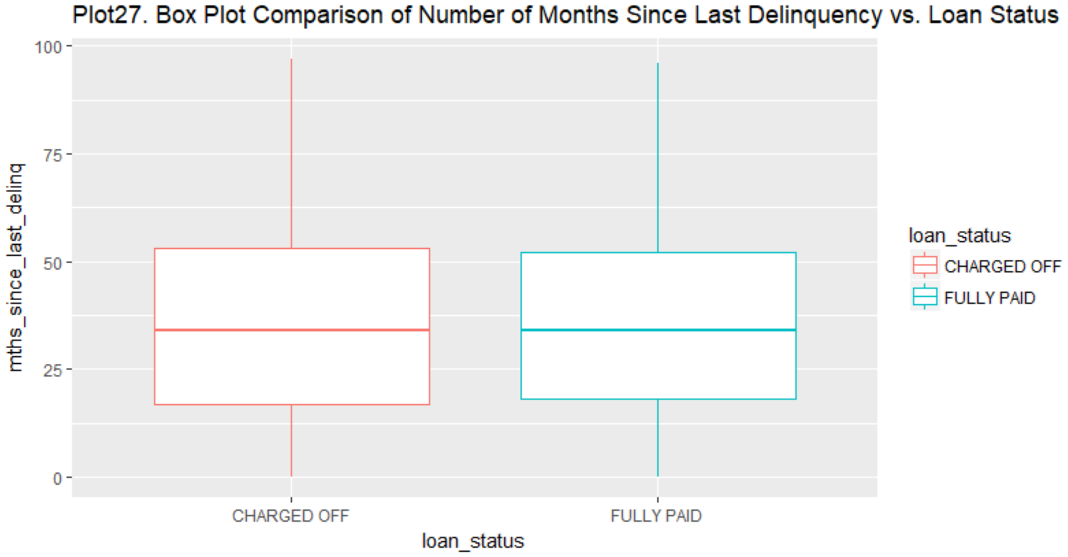


Observation [B]9. From the distribution we can see that revolving utilization rate is almost normally distributed with the between 40-60%

[C] Segmented analysis

boxplot(loan_dataset$mths_since_last_record)

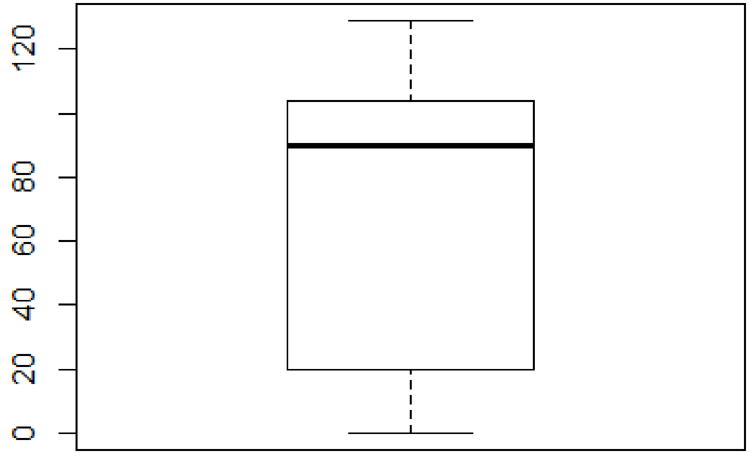



Plot27. Box Plot Comparison of Number of Months Since Last Delinquency vs. Loan Status

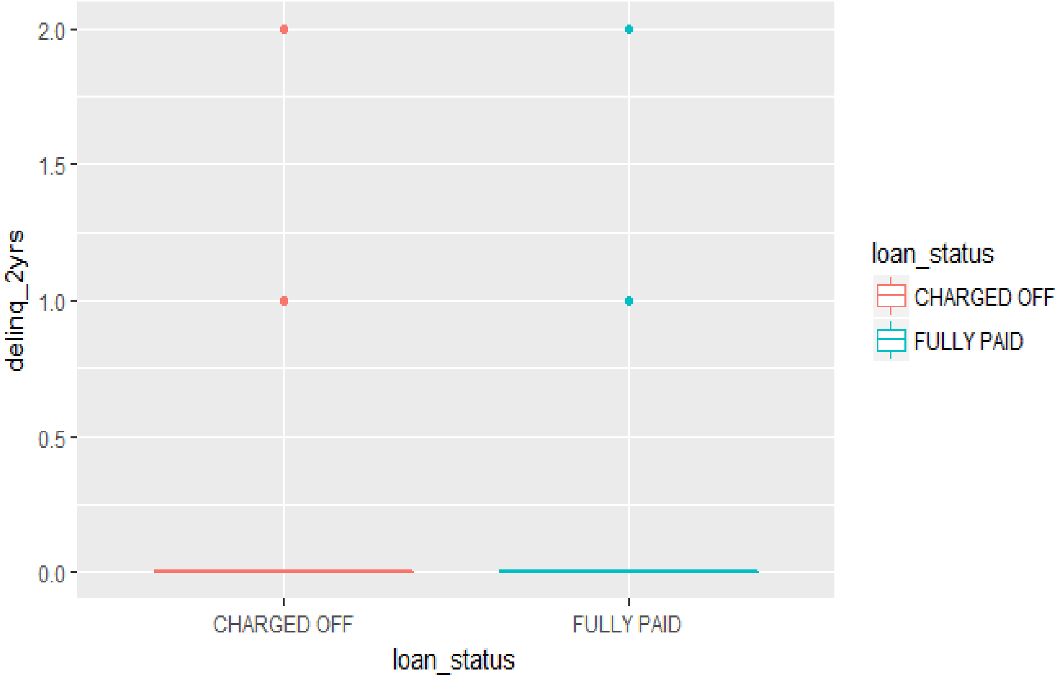
Very Similar and therefore not useful for analysis)

[C] Segmented analysis

Months since last delinquency Analysis



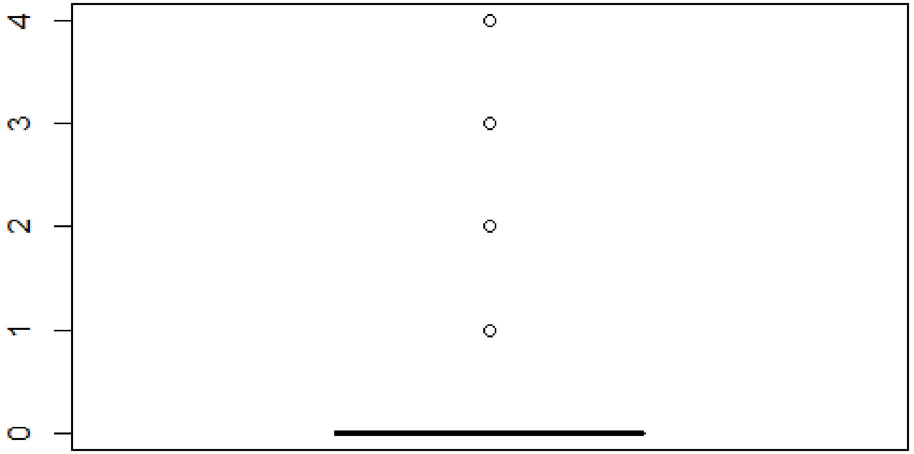
Plot27. Box Plot Comparison of Number of Delinquencies vs. Loan Status

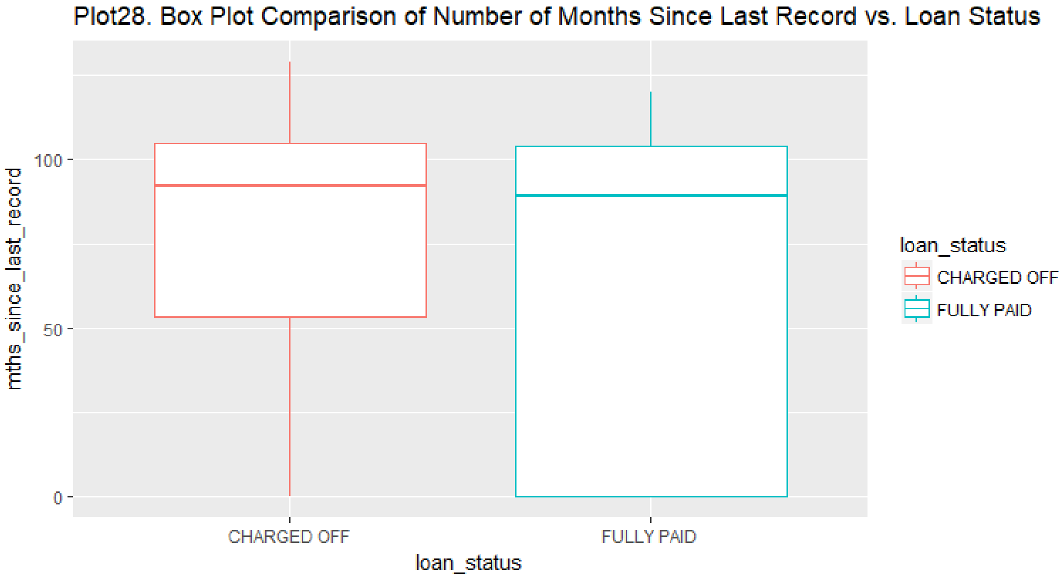


Too many 0 values therefore not useful for analysis.

[C] Segmented analysis

Months since last record Analysis





Too many 0 values therefore not useful for analysis.

- [D] For Bi-Variate Categorical Analysis please refer the slides 5-10 and the R-Code

- [E] For Bivariate Numerical analysis please refer the slide number 11 and the R-code.