GRAMENER CASE STUDY

Group Members:

1. Snigdha Prakash
2. Rahul Doshi
3. Amisha
4. Nitish Kanantha

Business Objective

The Gramener case study is for an online credit marketplace which facilitates personal loans, business loans and financing for medical procedures. Lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). Borrowers who **default** cause the largest amount of loss to the lenders.

Here, our aim is to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment so that such loans can be reduced thereby cutting down the amount of credit loss.

Business Constraints:

- Consider only consumers whose loan application is approved in analysis.

Goals:

- Understand the driving factors behind loan default
- Perform univariate and multivariate analysis of variables of interest
- Express the findings in terms of neat visualizations

Problem solving methodology

Data Collection – Loan dataset given by the company

Data Cleaning –

- Finding percentage of NA values in the columns like emp_length and revol_util and removing them
- Removing unnecessary non-numeric values from the numeric data column like removing % sign from int_rate and revol_util columns and the literal 'year' and 'month' from emp_length and term column respectively

Univariate Analysis – Analyzing the distribution of each variables individually

Multivariate Analysis – Finding the relationship between each variables

Conclusion – Finding which variables drives loan defaulting

Data Cleaning

- **We have identified 15 variable of interests which are listed down in the table –**

Target Variable	loan_status = “Charged Off”	
loan_amnt	funded_amnt	term
installment	int_rate	grade
emp_length	home_ownership	annual_inc
verification_status	purpose	dti
inq_last_6mths	open_acc	revol_util

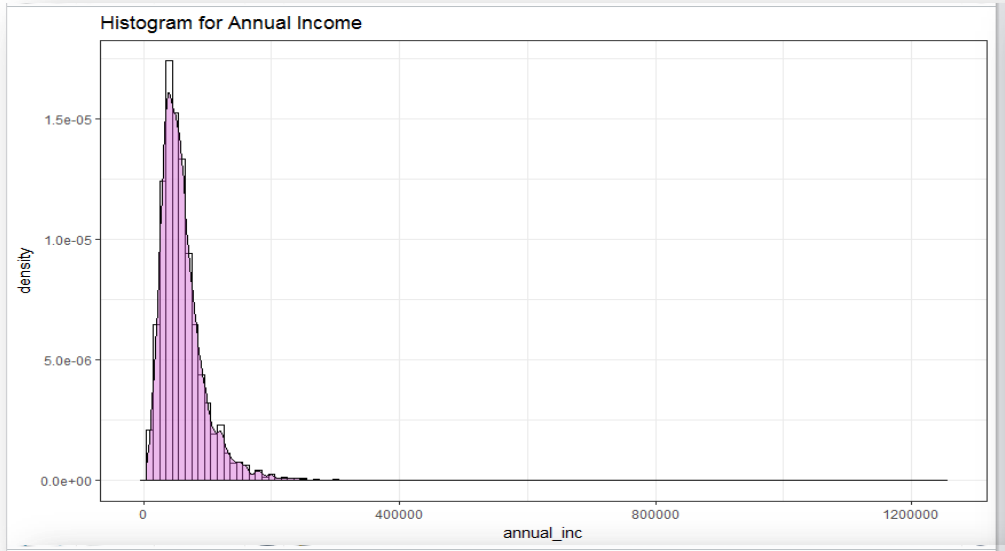
- Out of these variables, emp_length and revol_util contains NA values which constitutes 3% and 0.126% of total data respectively. We have removed these rows having NA values
- We aimed to keep Emp_length, int_rate, revol_util and term columns as numeric fields and hence removed the non-numeric values like % sign and other literals from the column data
- We have converted Grade, purpose, home_ownership and term into factor columns since they are categorical variables

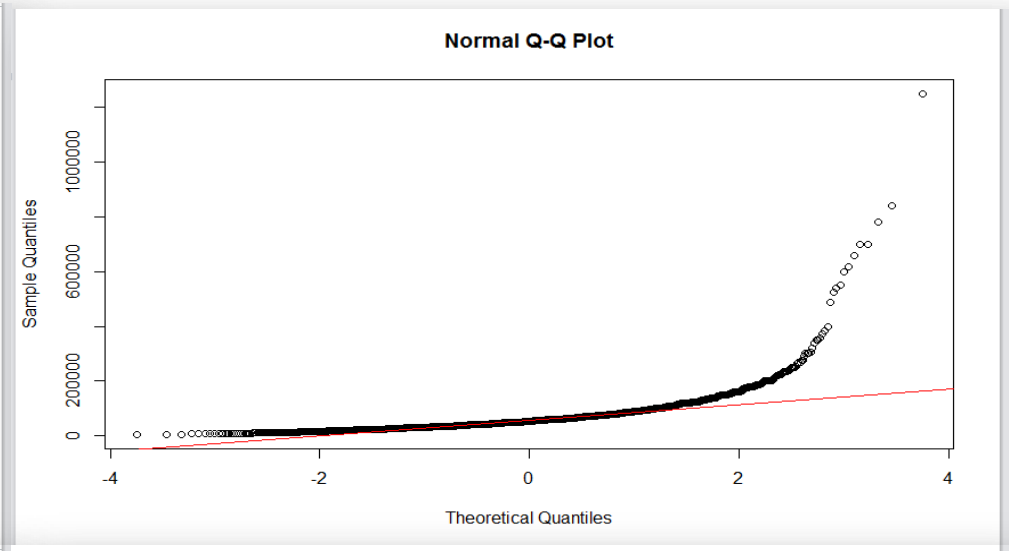
Performing univariate analysis

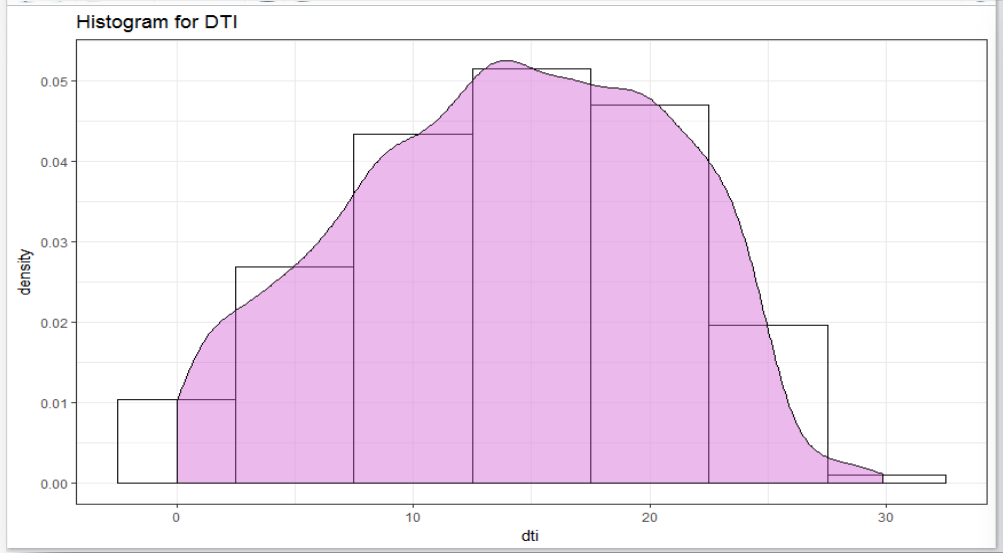
- **We have divided the univariate analysis into 2 parts –**
  - Analysing the behaviour of continuous variables using frequency and density plots
  - Analysing the behaviour of categorical variables using bar plots
- **Analyzing continuous variables –** We recorded the skewness of each variable to check the actual distribution. The graphs are shown in the next few slides

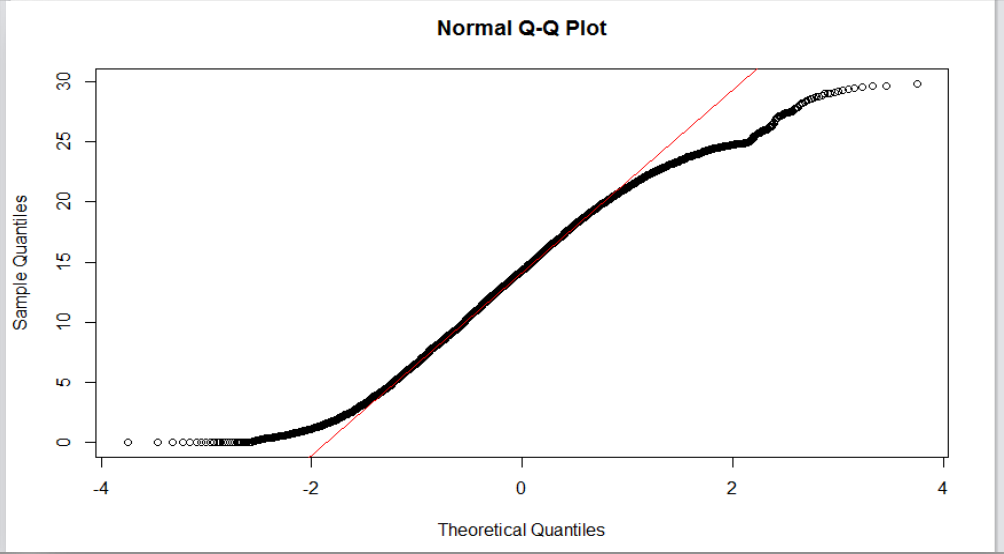
Variable	Skewness	Inference
loan_amnt	~0.89	The distribution is skewed towards the right
installment	~1.00	The distribution is skewed towards the right
int_rate	~0.1	Close to a normal distribution
annual_inc	~7.6	The distribution is strongly skewed towards the right
dti	~-0.17	The distribution is slightly skewed towards the left
open_acct	~1.07	The distribution is skewed towards the right
Revol_util	~0.324	The distribution is skewed towards the right

Univariate Analysis – Graphs for continuous variables

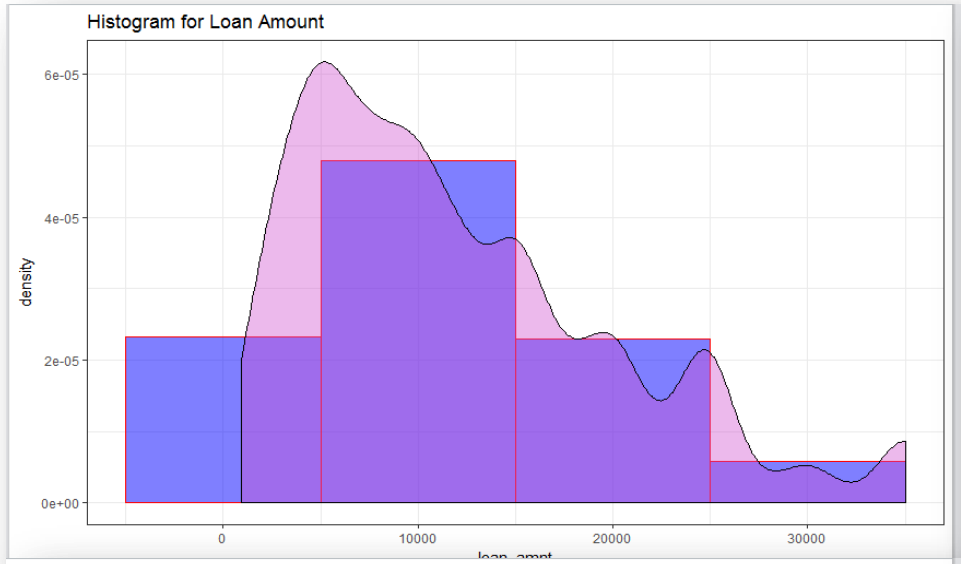


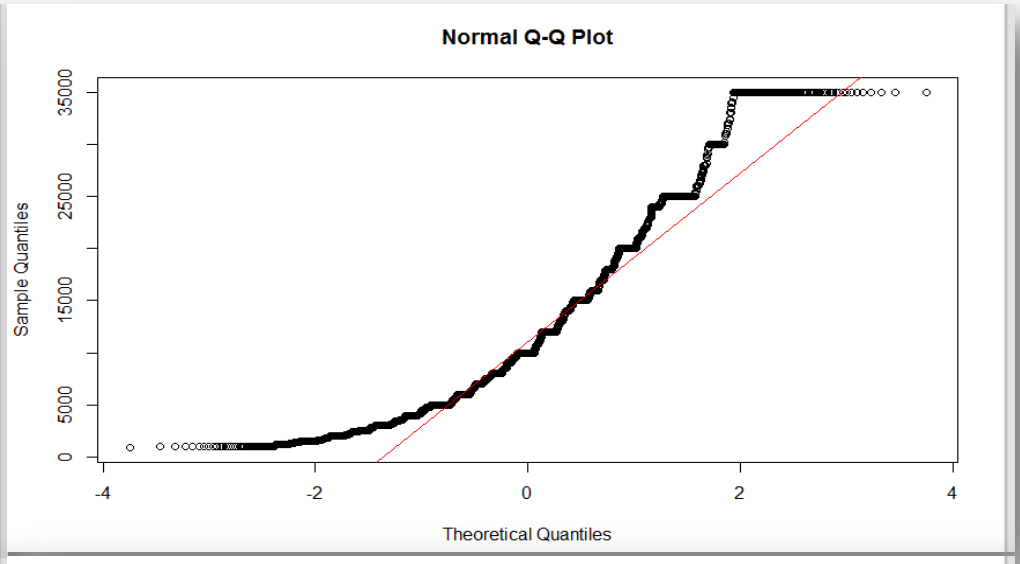


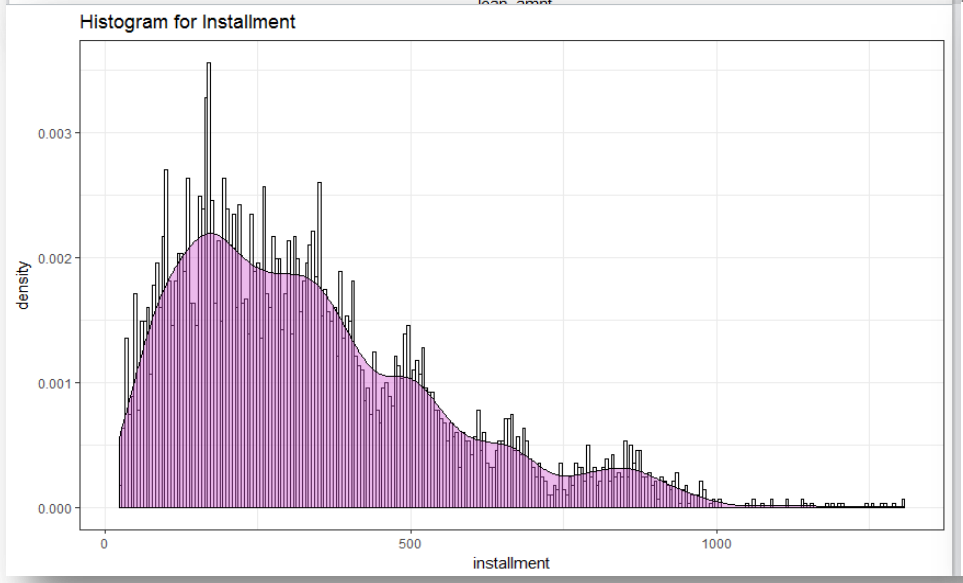


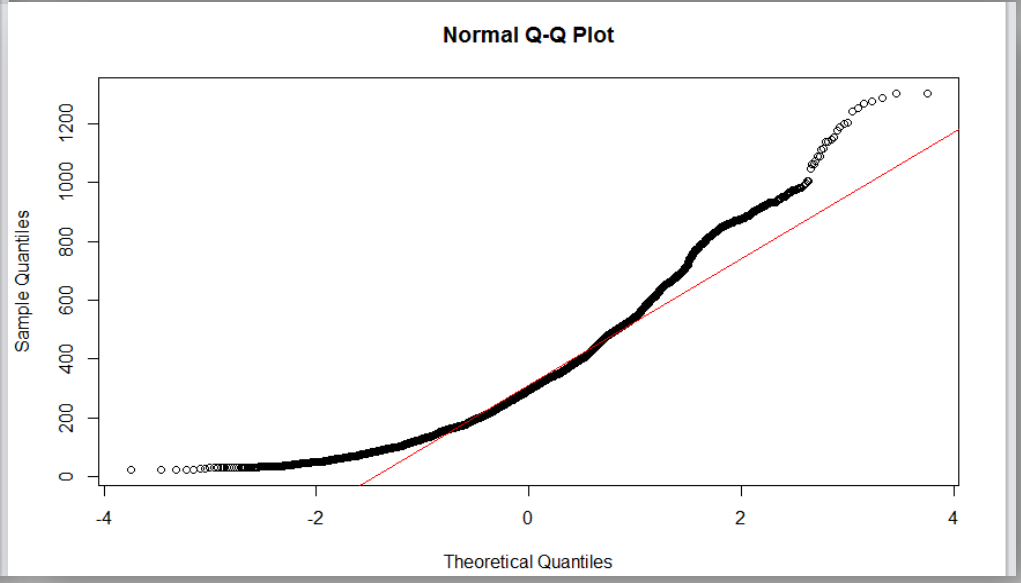


Univariate Analysis – Graphs for continuous variables

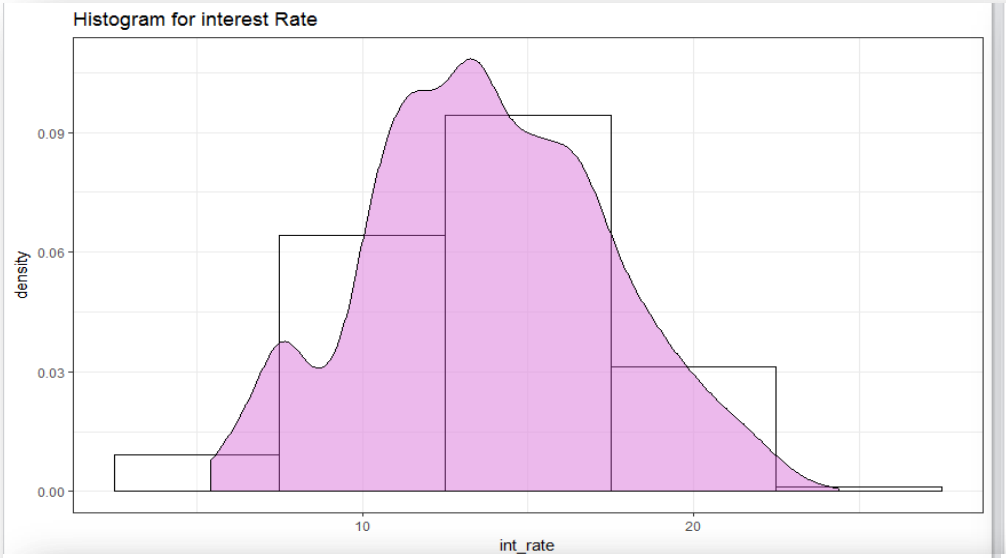


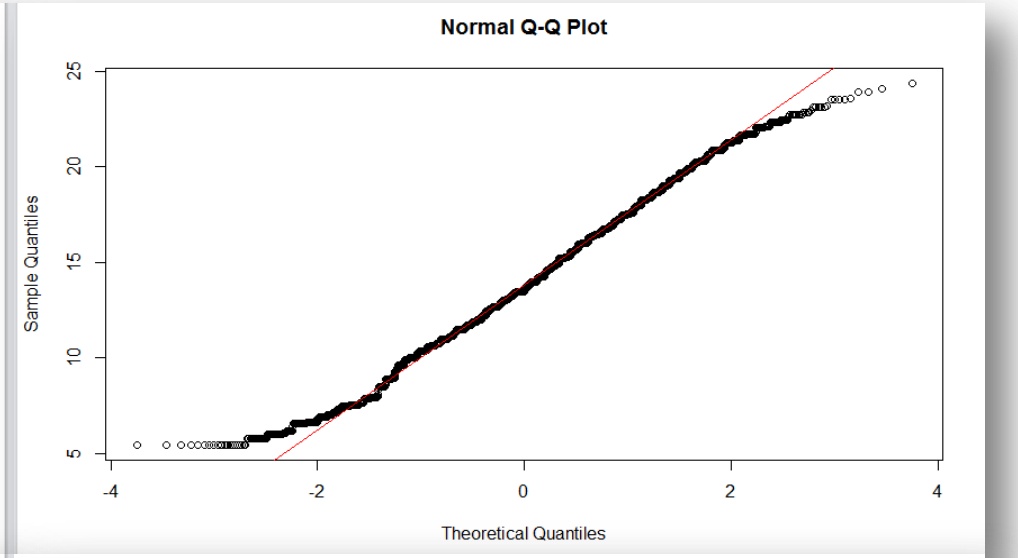


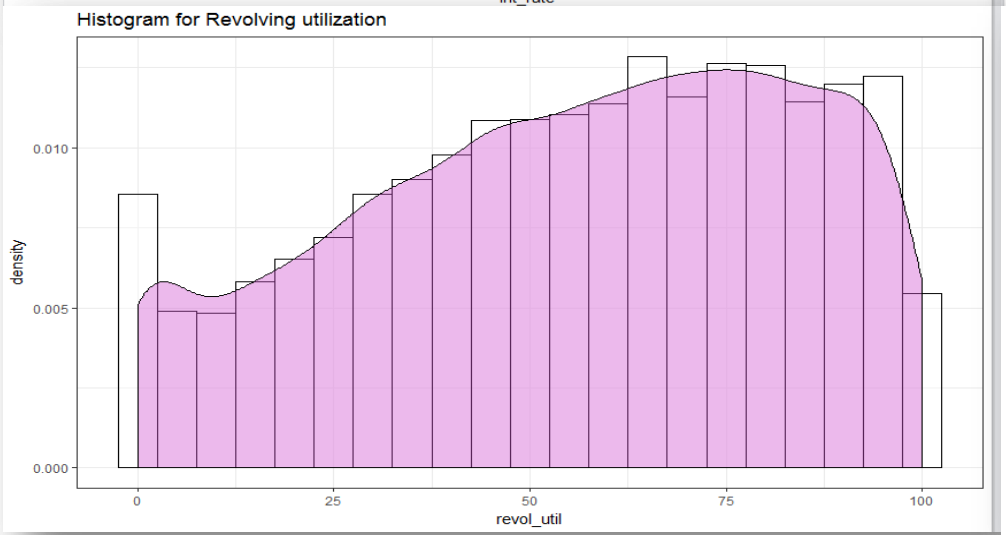


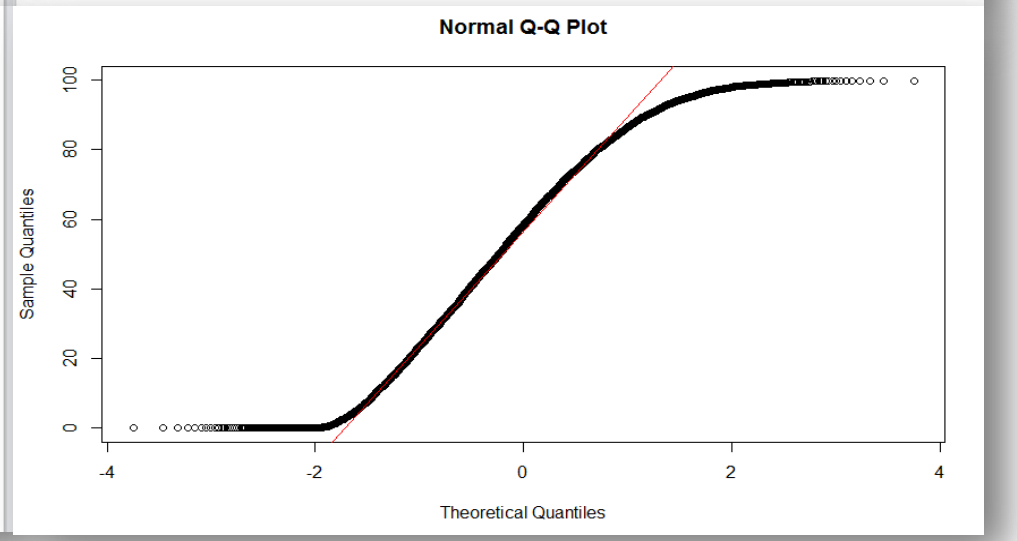


Univariate Analysis – Graphs for continuous variables

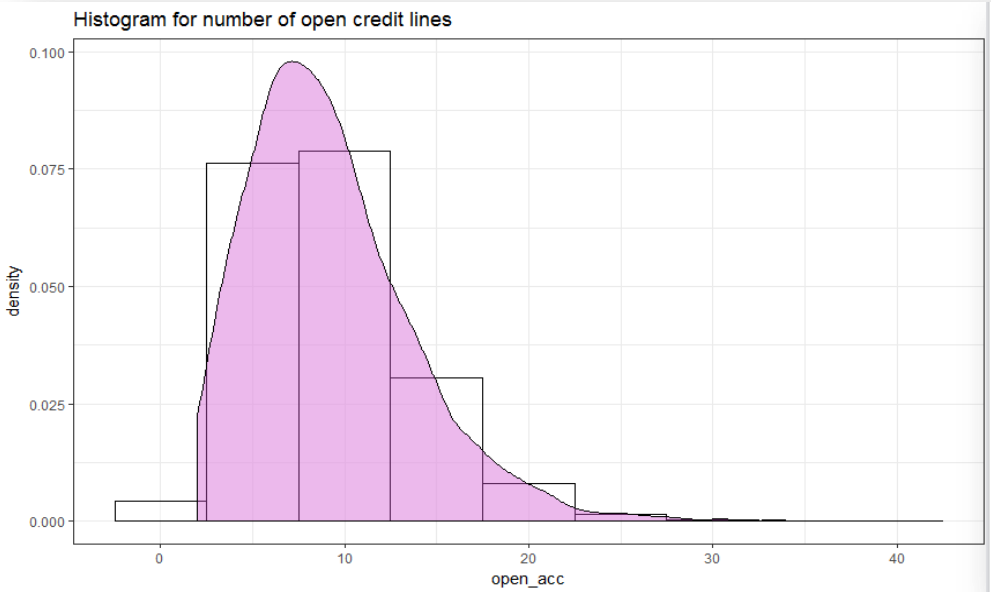


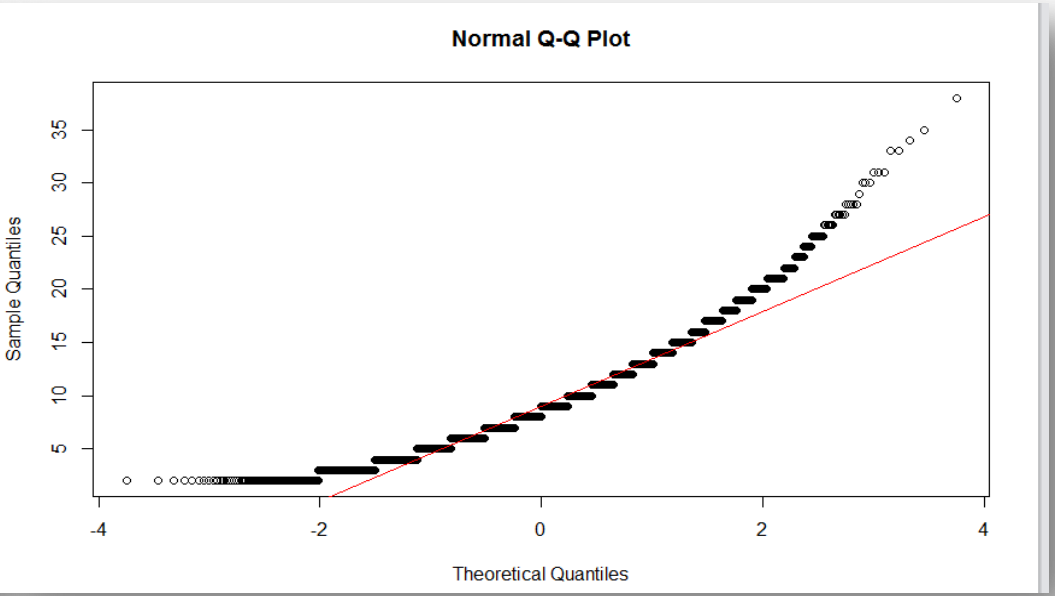




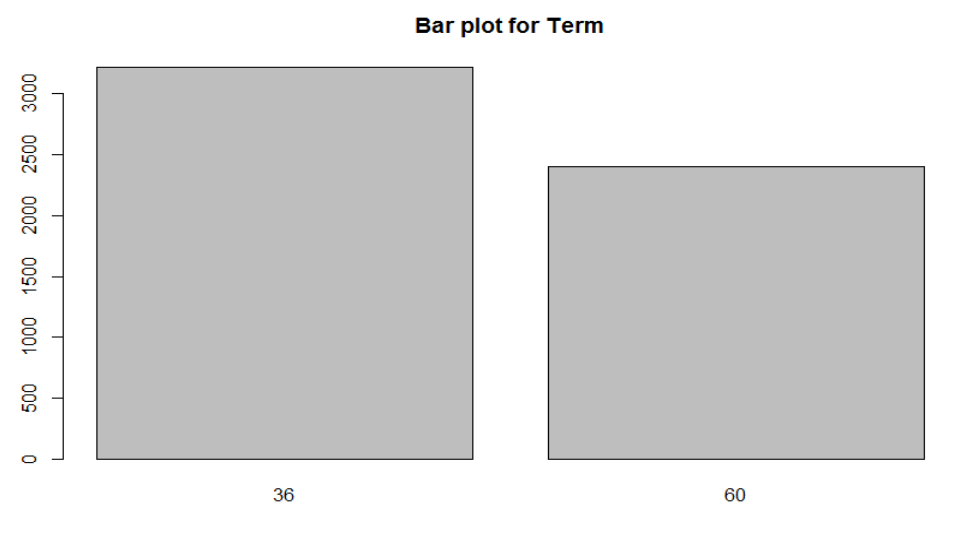


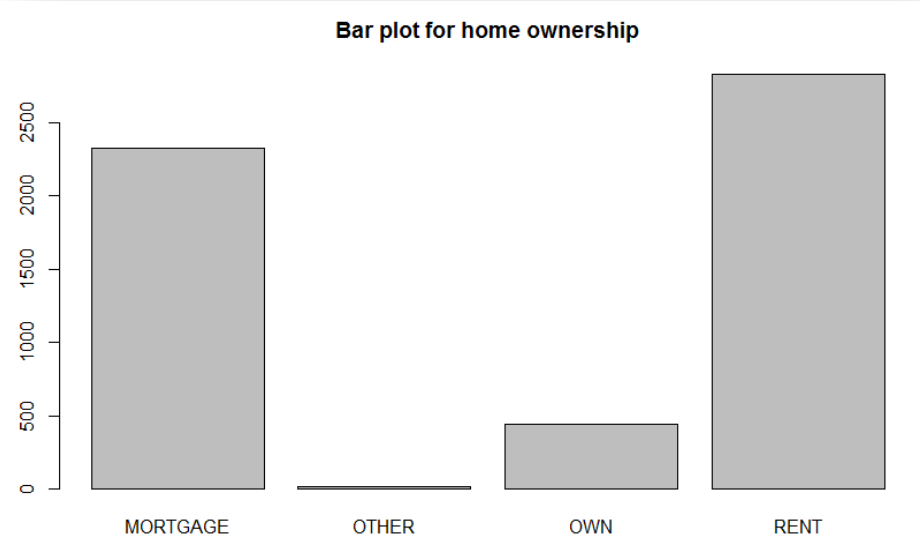
Univariate Analysis – Graphs for continuous variables

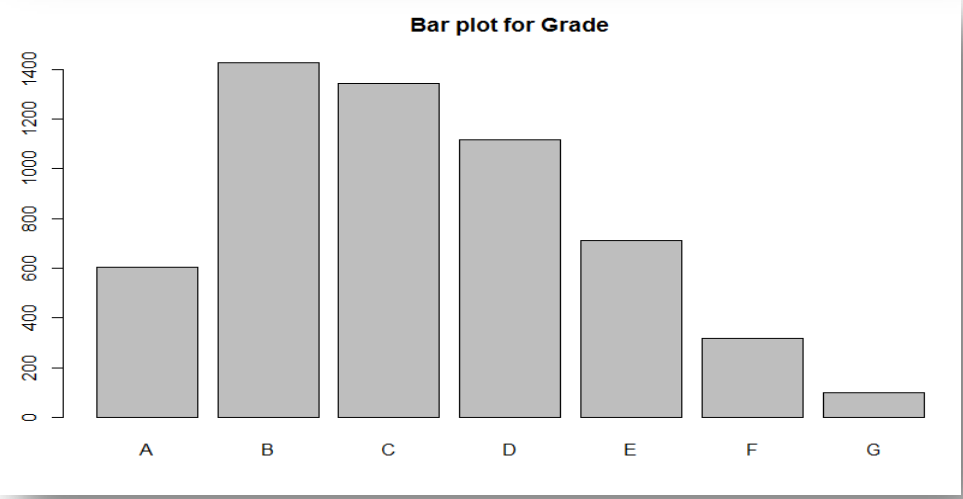




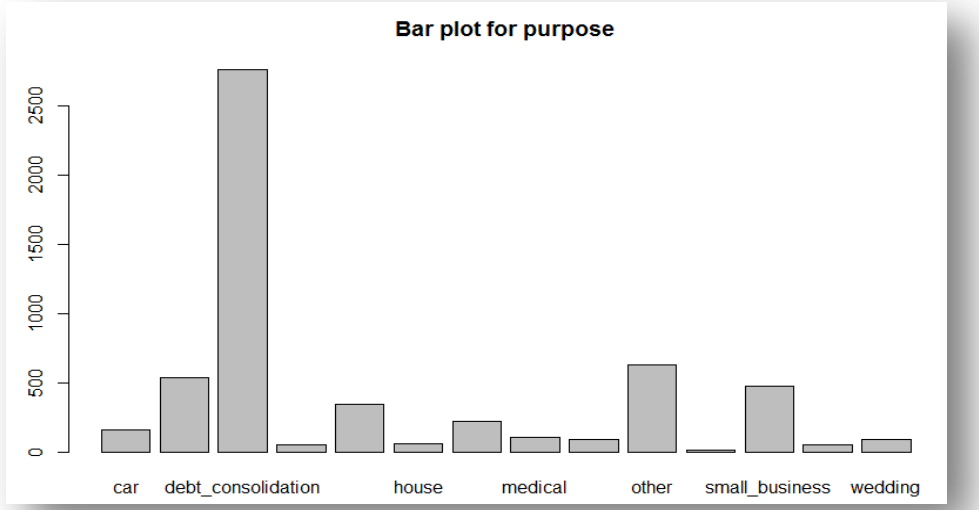
Univariate Analysis – Graphs for categorical variables

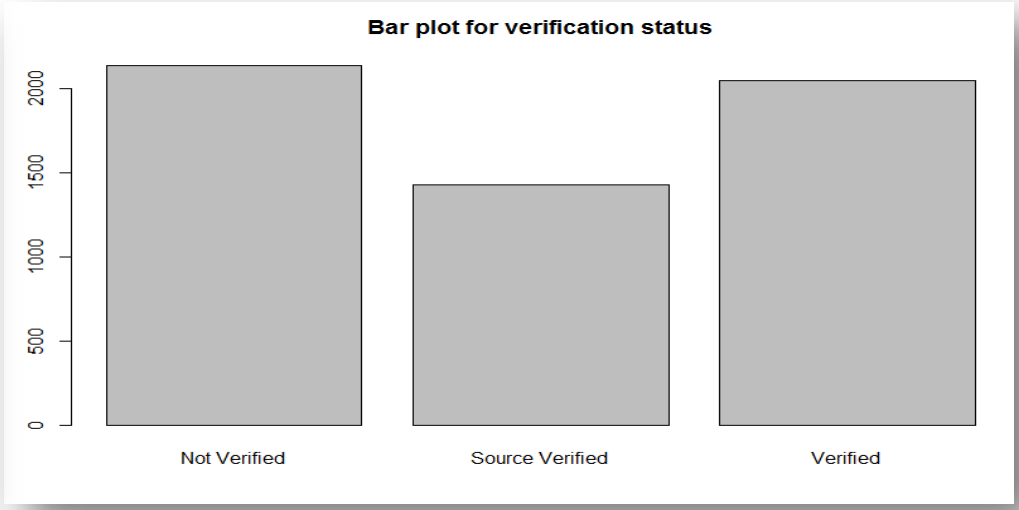


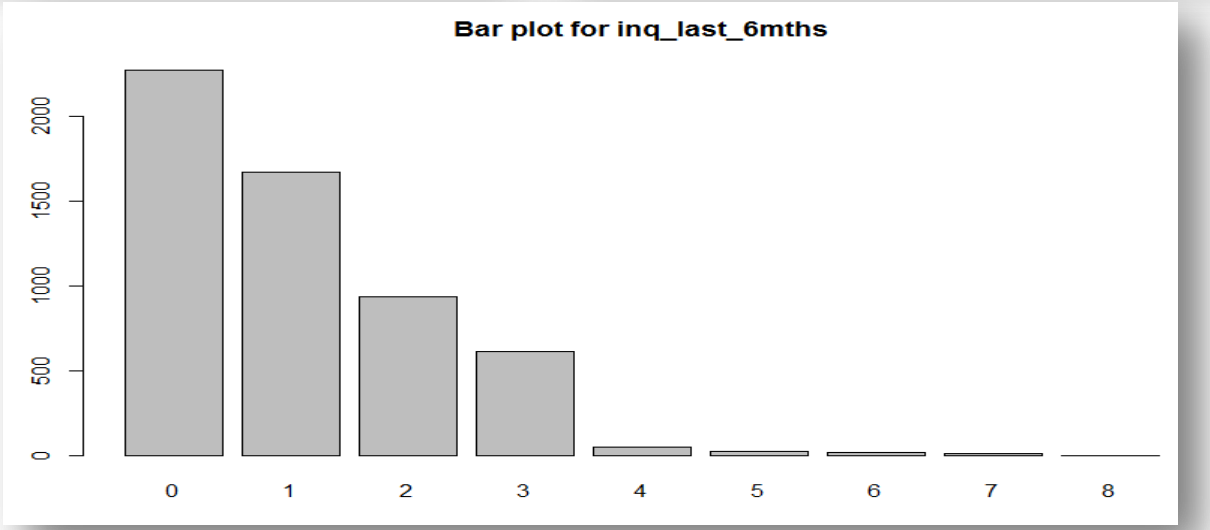




Univariate Analysis – Graphs for categorical variables







Performing multivariate analysis

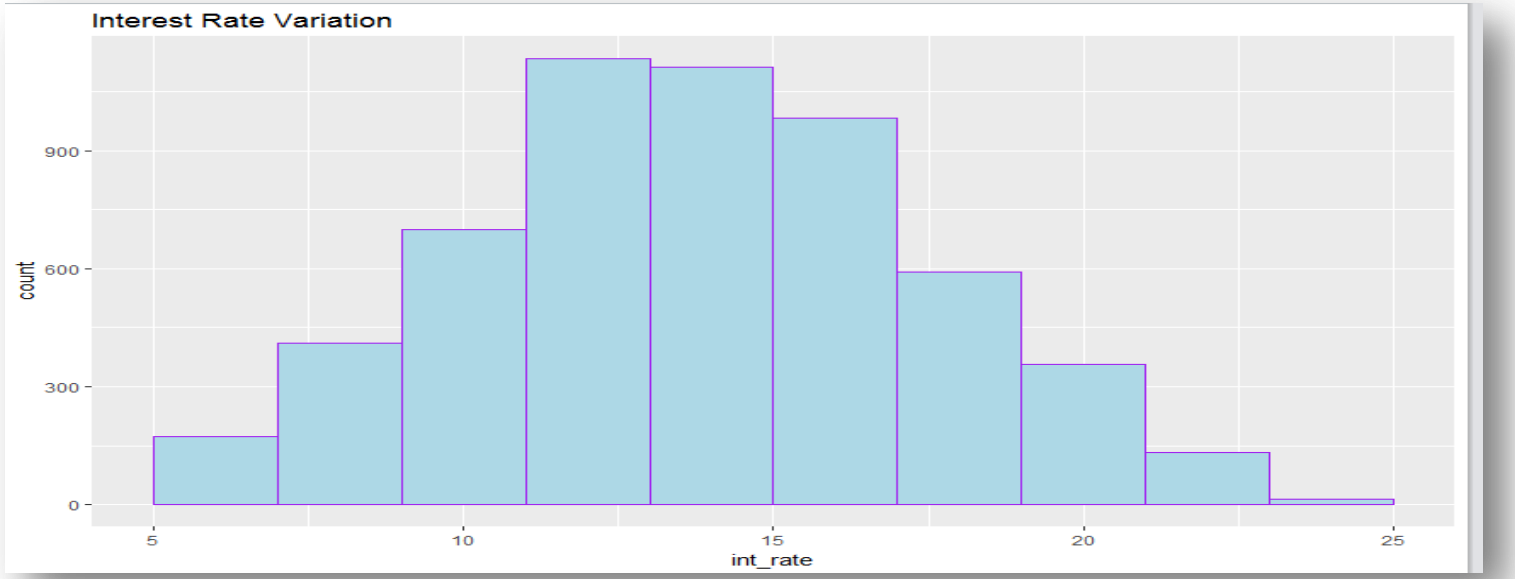
- **We have divided the multivariate analysis into 2 parts –**
  - Analysing how our variable of interest affects the interest rates
  - Analysing how variables of interest are related to each other

- **<u>Analyzing how the variable on interest affects the interest rates –</u>**
  - We have grouped the interest rate in 3 buckets – low, medium and high
  - For each category of interest rate, we have made a boxplot with the quantitative variables of interest and bar plot with the categorical ones

- **<u>Analyzing how the variable on interest are related to one another –</u>**
  - We have created histogram for the continuous variables and checked their relationship with the categorical variables using aesthetics

Analysis of interest Rates

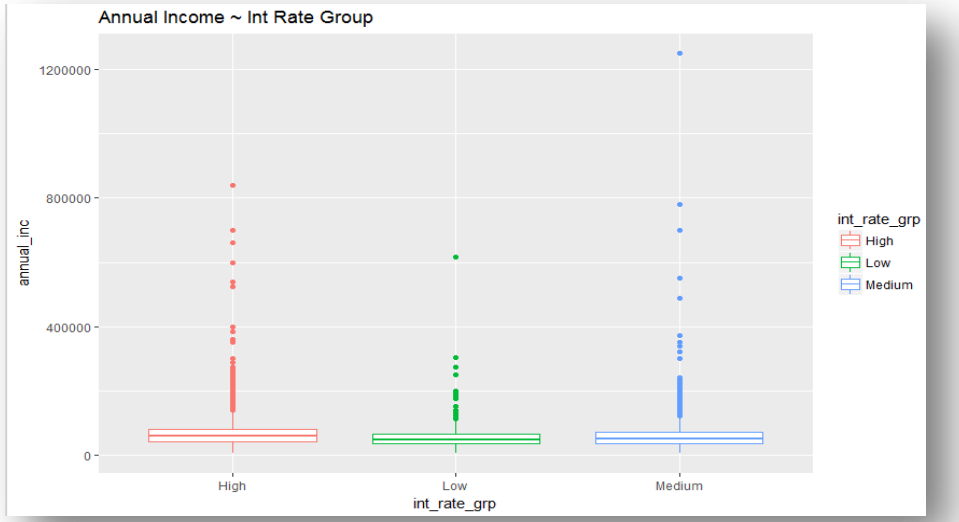
- The frequency plot of interest rate is shown below -

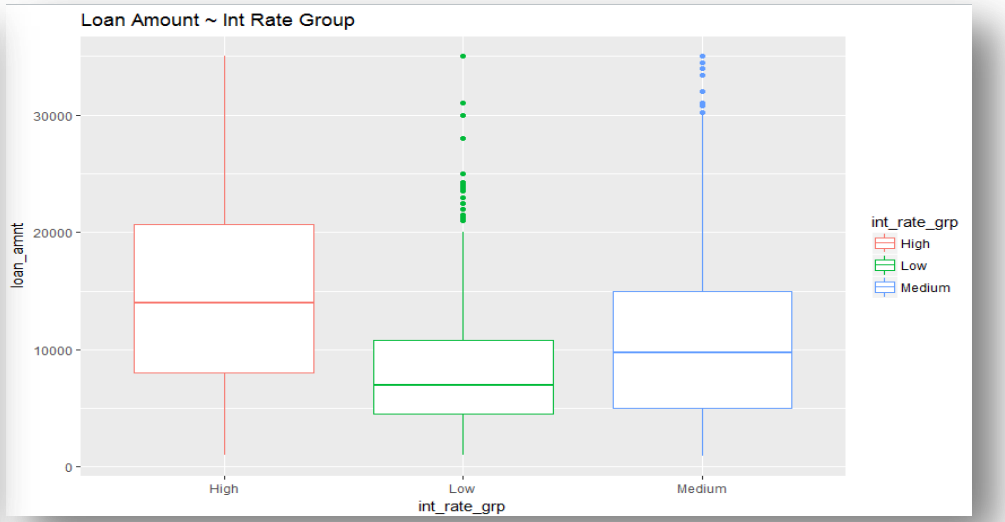


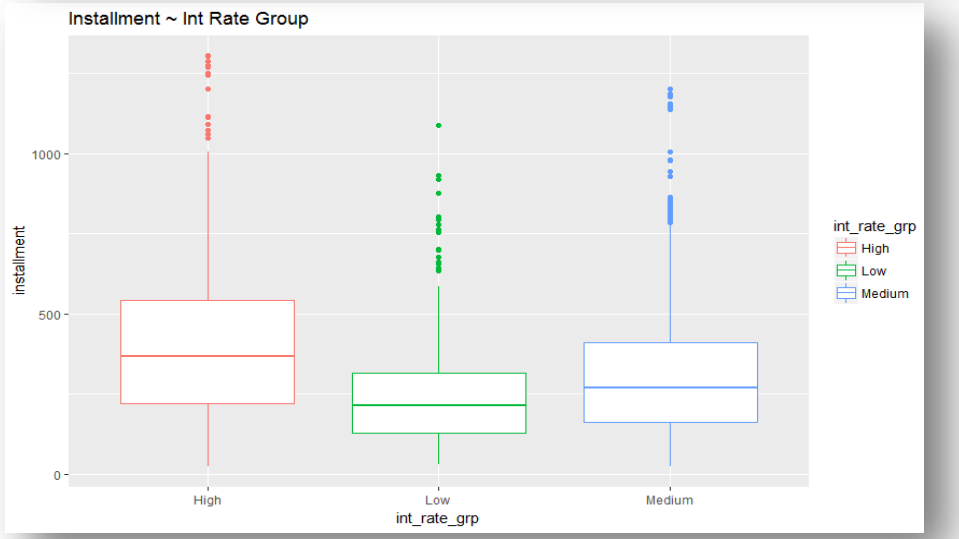
- We have categorized it into 3 groups -

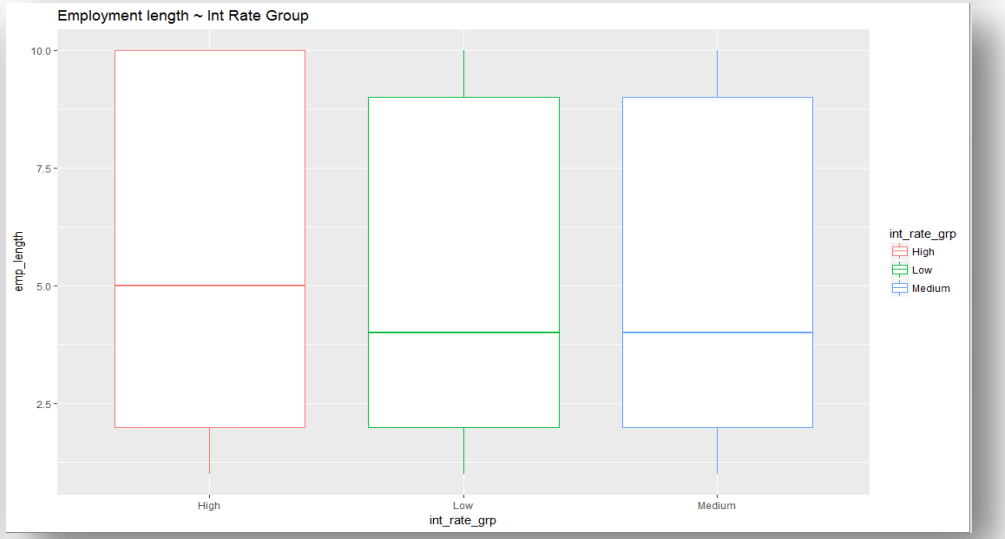
Data Range	Group
0 - <10	Low
10 – 15	Medium
>15	High

Multivariate Analysis – Variation of quantitative variables within interest rate groups

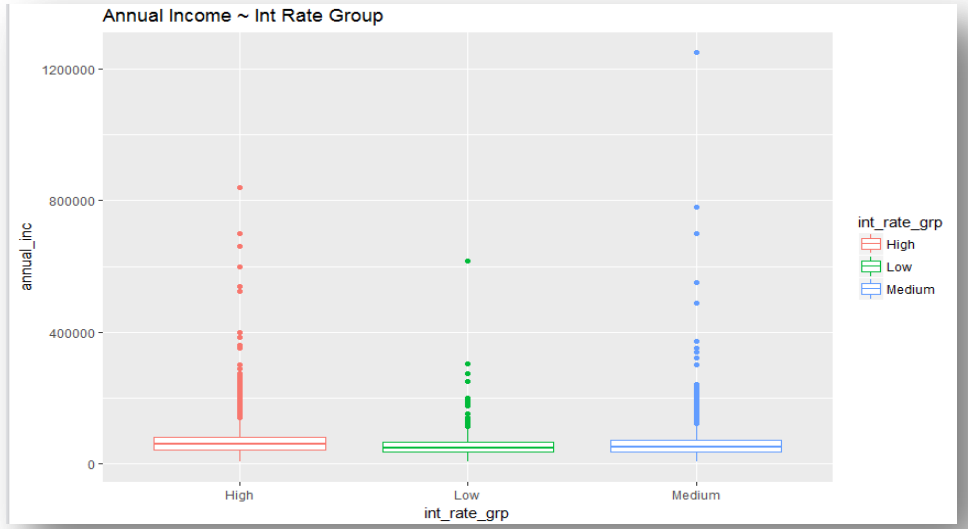


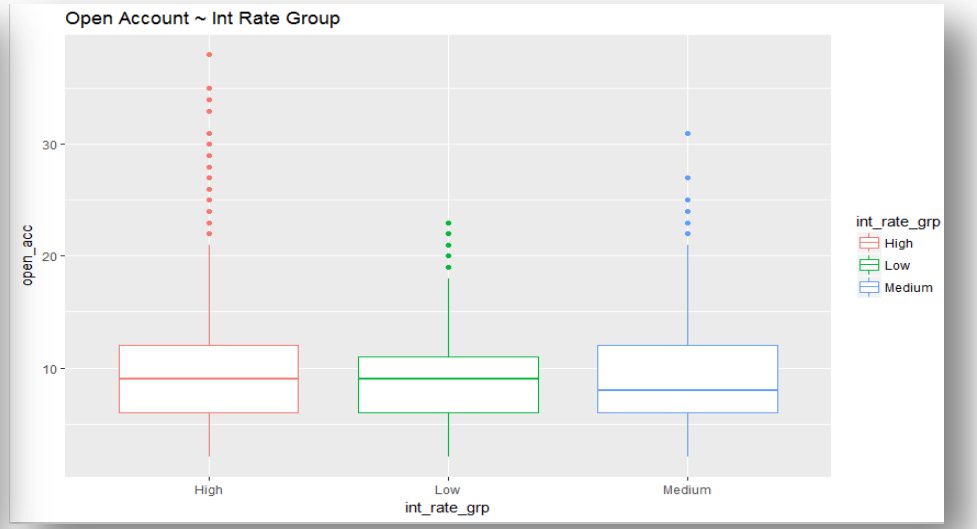


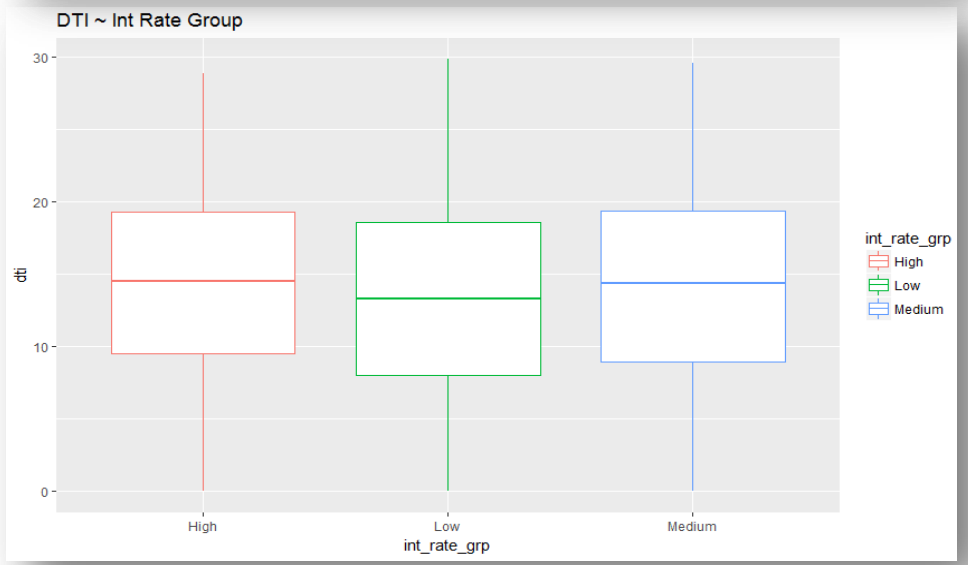


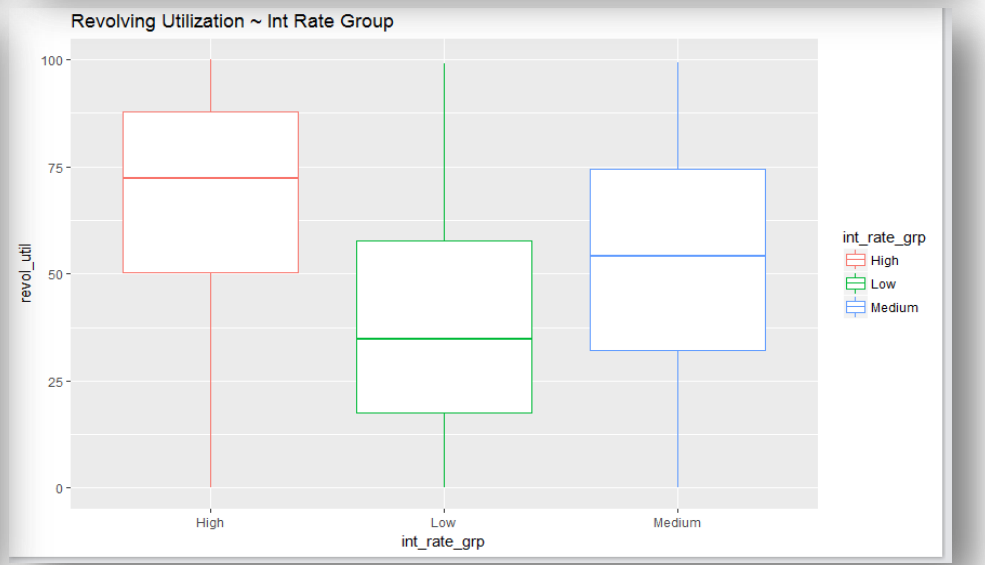


Multivariate Analysis – Variation of quantitative variables within interest rate groups

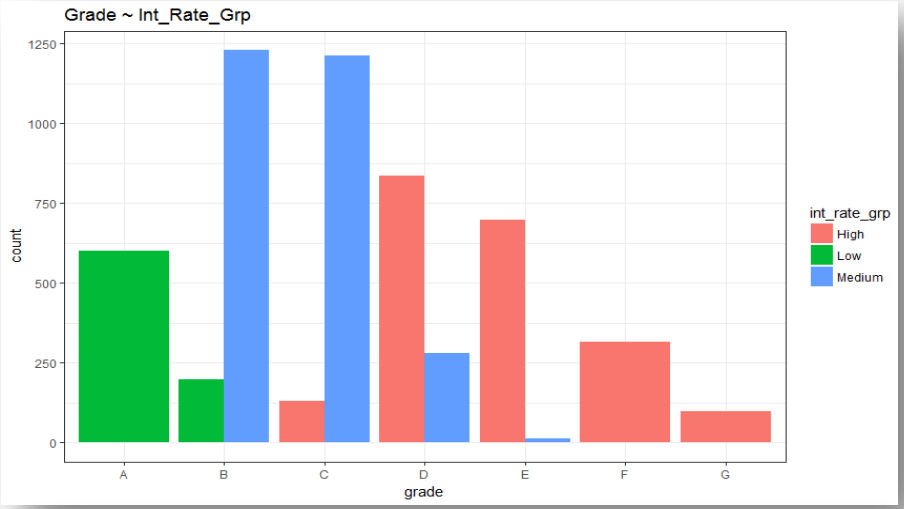


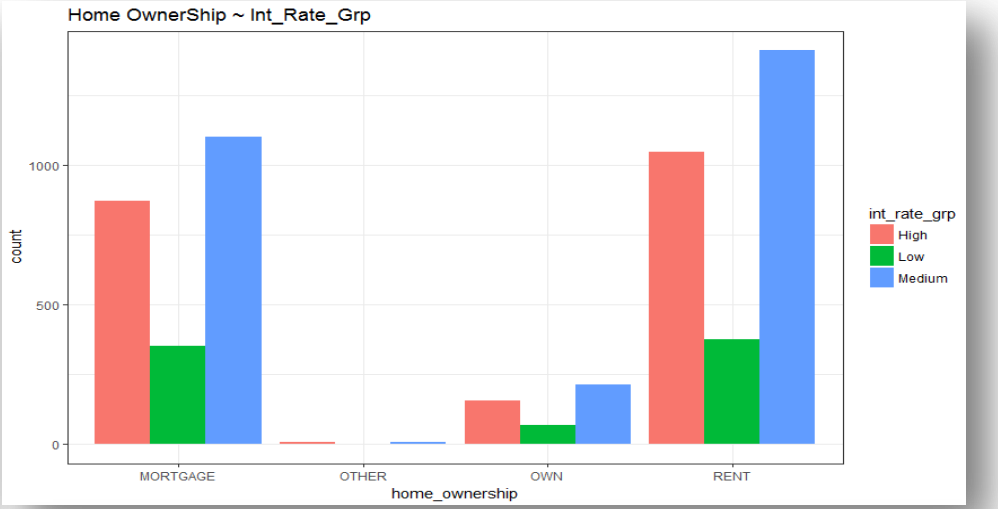


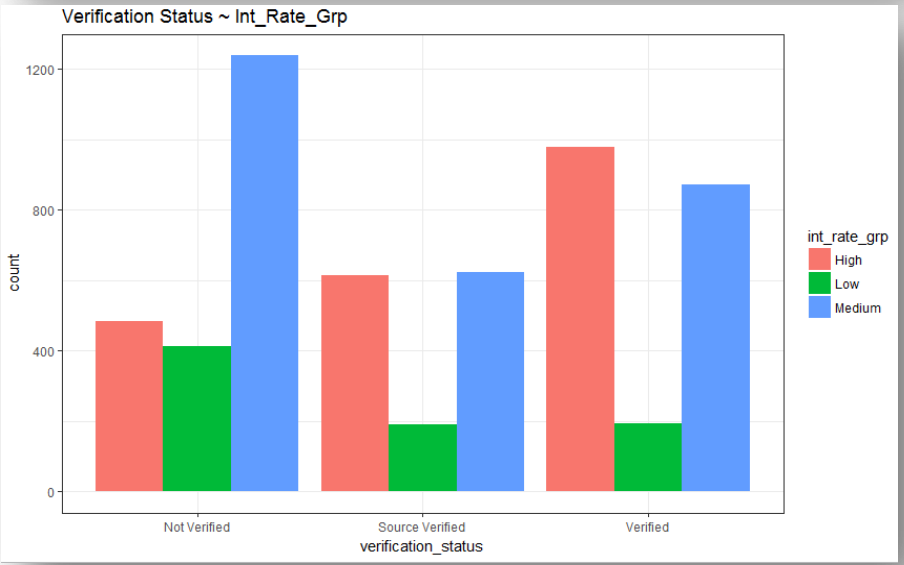


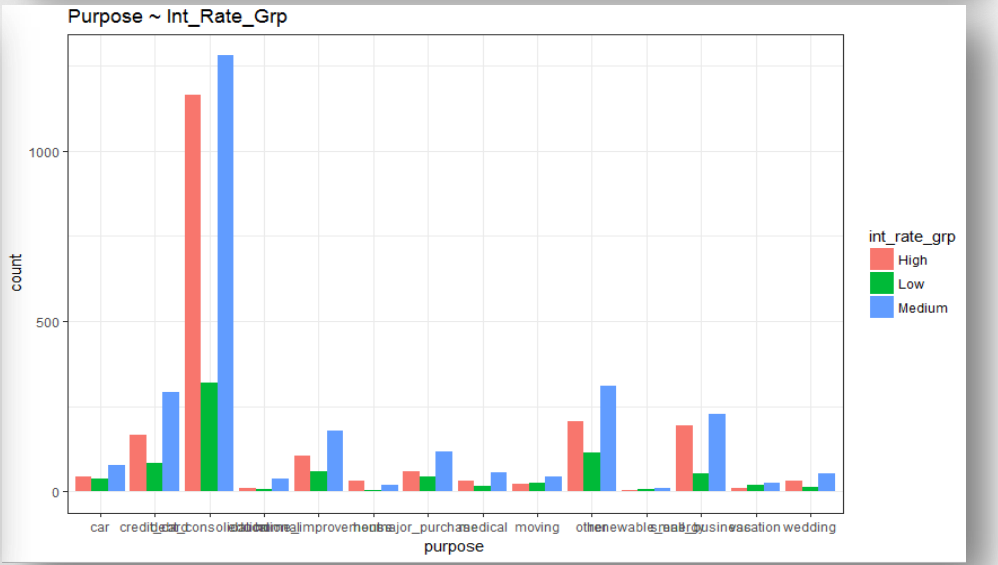


Multivariate Analysis – Variation of categorical variables within interest rate groups









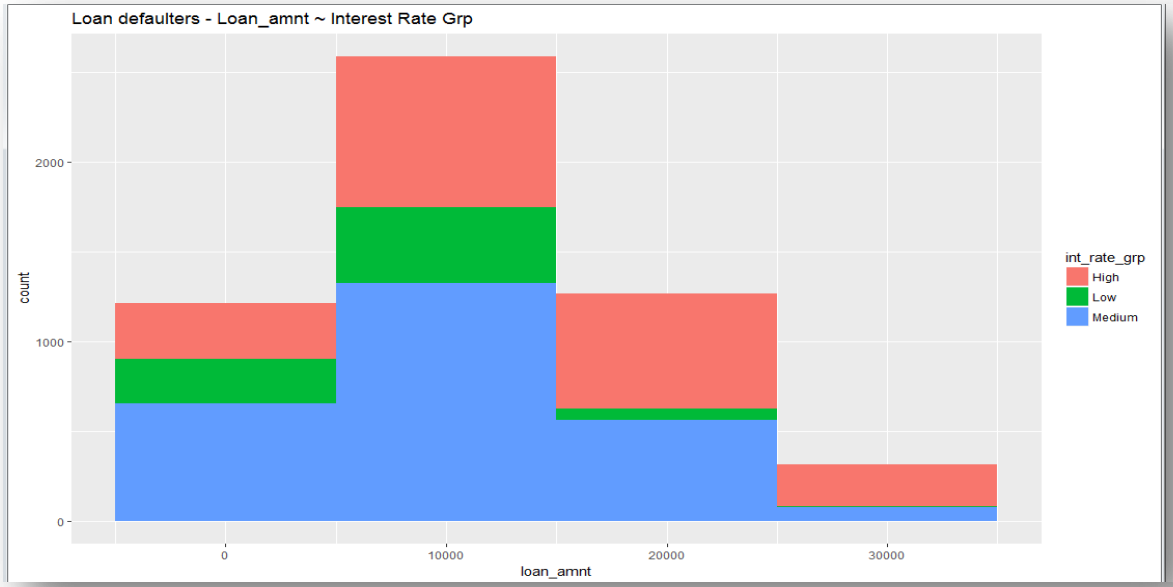
Relation of variables on interest with one another

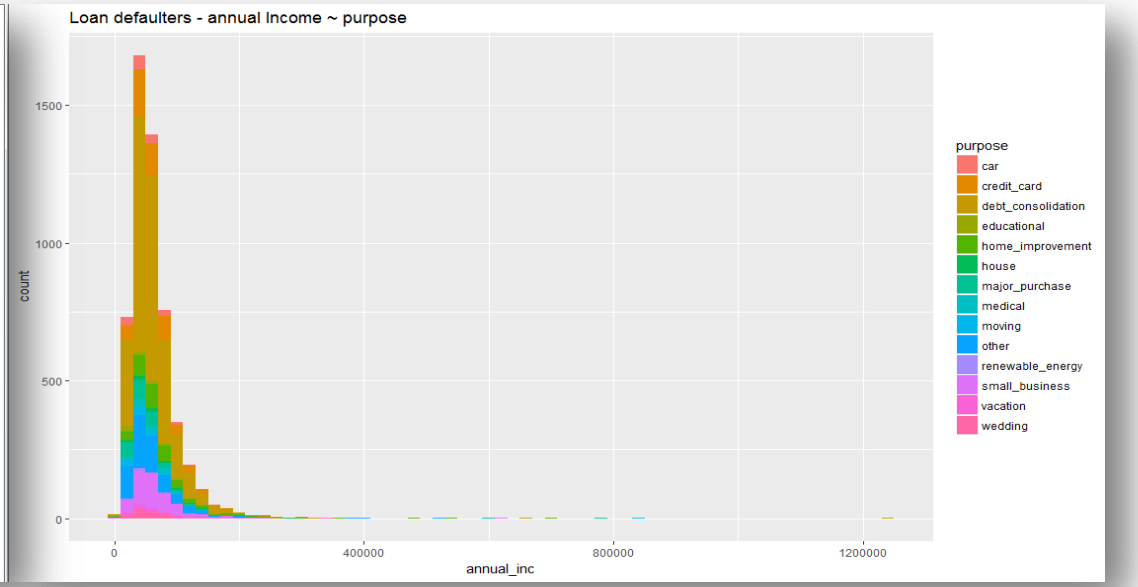
Variables	Correlation	Inference
loan_amnt ~ funded_amnt	~0.98	Positive and a strong correlation between the variables
Loan_amnt ~ installment	~0.92	Positive and a strong correlation between these variables
Loan_amnt ~ annual_inc	~0.35	Weak relation between these variables

- We have plotted the histogram for the continuous variables are observed their trends with each of the categorical variables using aesthetics. We have then found the percentage of the total amount and drawn relevant inferences

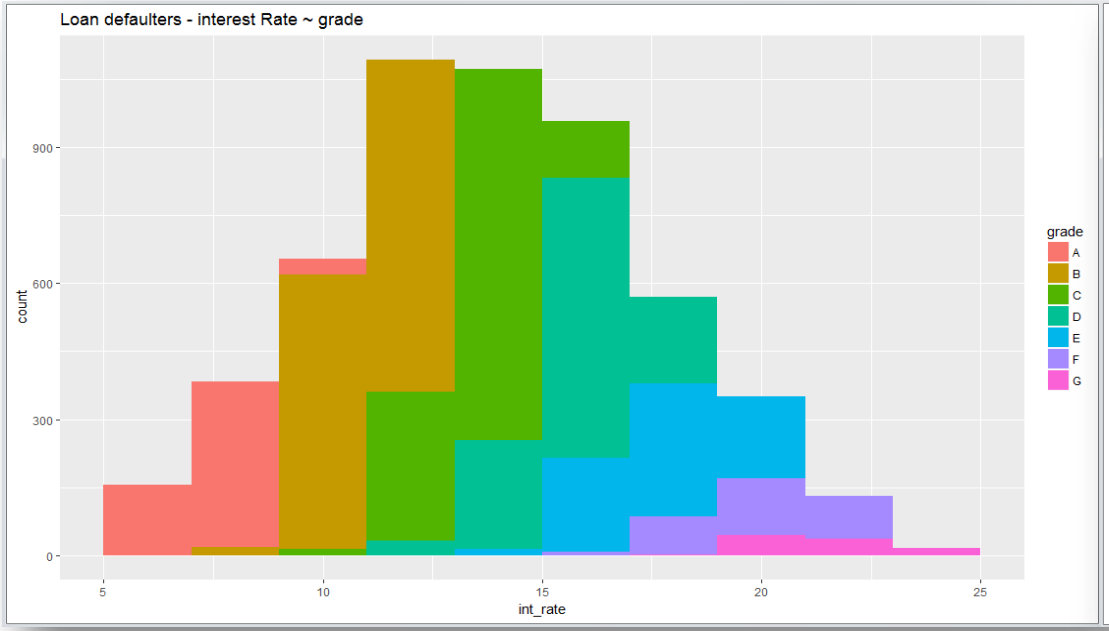
- The graphs are shown in subsequent slides

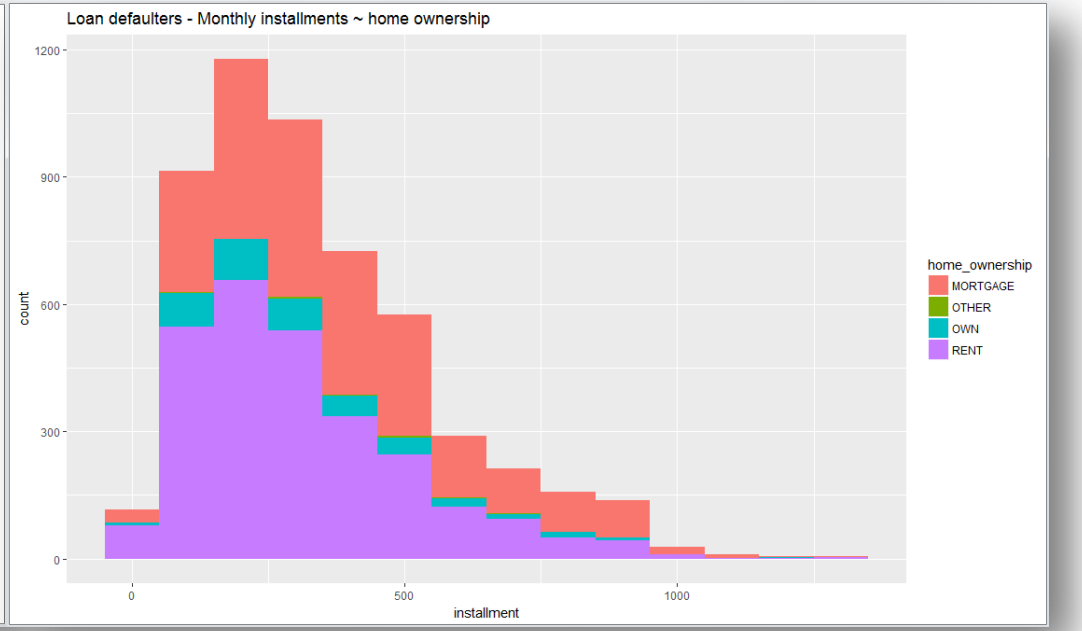
Conclusion



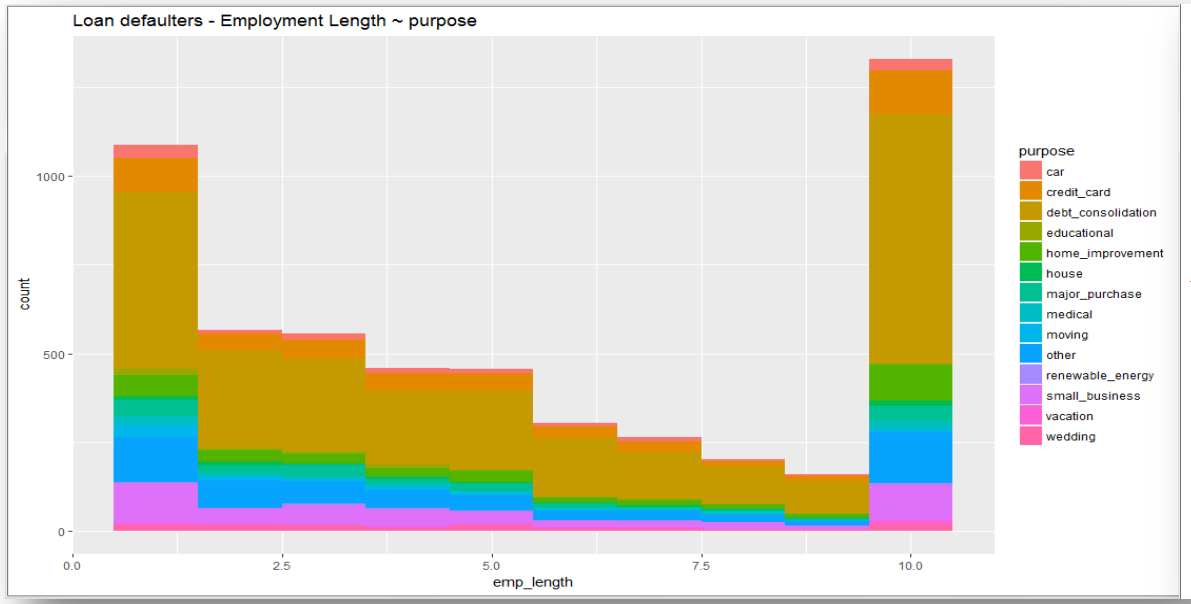


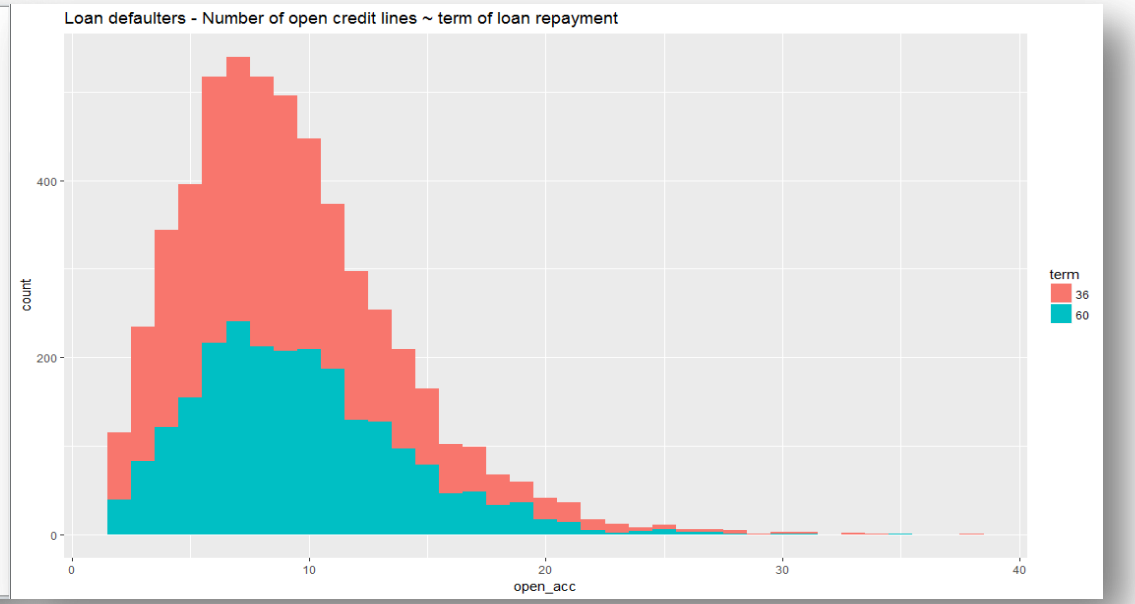
Inferences		
Loan amount ~ Interest rate group	51% of people take Loan Amount between 0 to 10000 and they are most likely to default	Among these people, ~54% of people are given medium rate of interest that is ranging from 5% - 20%
Annual Income ~ purpose	Close to 76% of people earning between 40K – 80 K are likely to default	Among them ~51% of people have taken loan for the purpose of debt consolidation





Inferences		
Interest rate ~ Grade	48% of people are given loan at interest rate between 10-15% and they are most likely to default	Among these people, ~45% of them belongs to Grade B
Monthly Installment ~ home ownership	Close to 58% of people who have monthly installments in the range of 200 – 400 are most likely to default	Among them ~41% have mortgaged their home and 51% lives in the rented flats





Inferences

Employment Length ~ purpose	Customers having employment length of less than 1 year and greater than 10 years together constitutes 45% of the total data who are most likely to default.	Among them, close to 50% takes loan for debt consolidation and they default the most
Number of open credit lines ~ term of loan repayment	Close to 47% of people who have 6 – 10 open credit lines are most likely to default	Among them ~57% of people have loan repayment term of 36 months are they are more likely to default