Case Study: BFS Capstone Project

Business Understanding

CredX is a leading credit card provider that gets thousands of credit card applicants every year. But in the past few years, it has experienced an increase in credit loss.

In this project, we will help CredX identify the right customers using predictive models. Using past data of the bank's applicants, you need to determine the factors affecting credit risk, create strategies to mitigate the acquisition risk and assess the financial benefit of your project.

Data Understanding

The credx company has provided us two data sets, demographic/application data and credit bureau data.

The demographic data is obtained from the information provided by the applicants at the time of credit card application, which includes customer level information on age, gender, income, marital status etc.

The credit bureau data contains variables such as number of time 30 dpd or worse in last 3/6/12 months, outstanding balance, number of trades etc.

The demographic data consists of 71295 observations with 12 variables including 1425 na's in performance tag and 3 duplicates application id, the credit bureau data consists of 71295 observations with 19 variables including 1425 na's in performance tag and 3 duplicates application id.

Data Cleaning

We see that there are 272 NA's in Presence of open home loan , 272 NA's in Outstanding Balance and 1425 NA's in Performance Tag. Performance tag is our target variable. So, since the data has no information about default, this applicant are the one who have not been given credit card by the company, so we will keep this data in our validation sets.

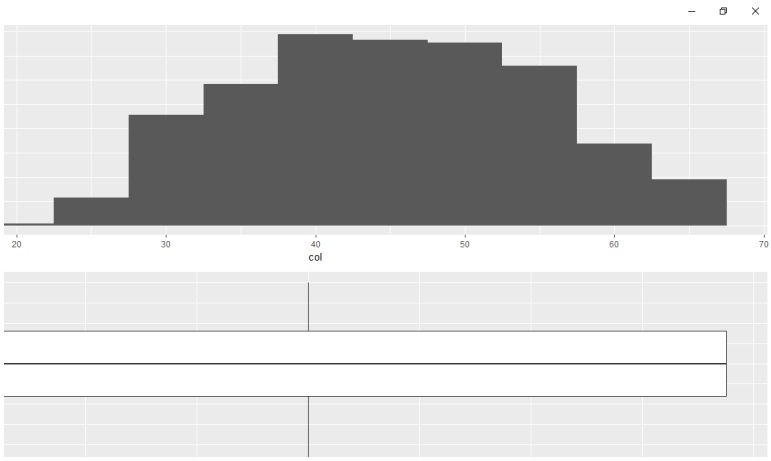
After removing validation data, we are left with 69867 obs.

We see Education has highest NA values 118, profession has 13 and type of residence has 8 , marital status 6, gender in 2 and 1425 NA's in Performance tag. We can remove the data with NA's in performance tag for validation set.

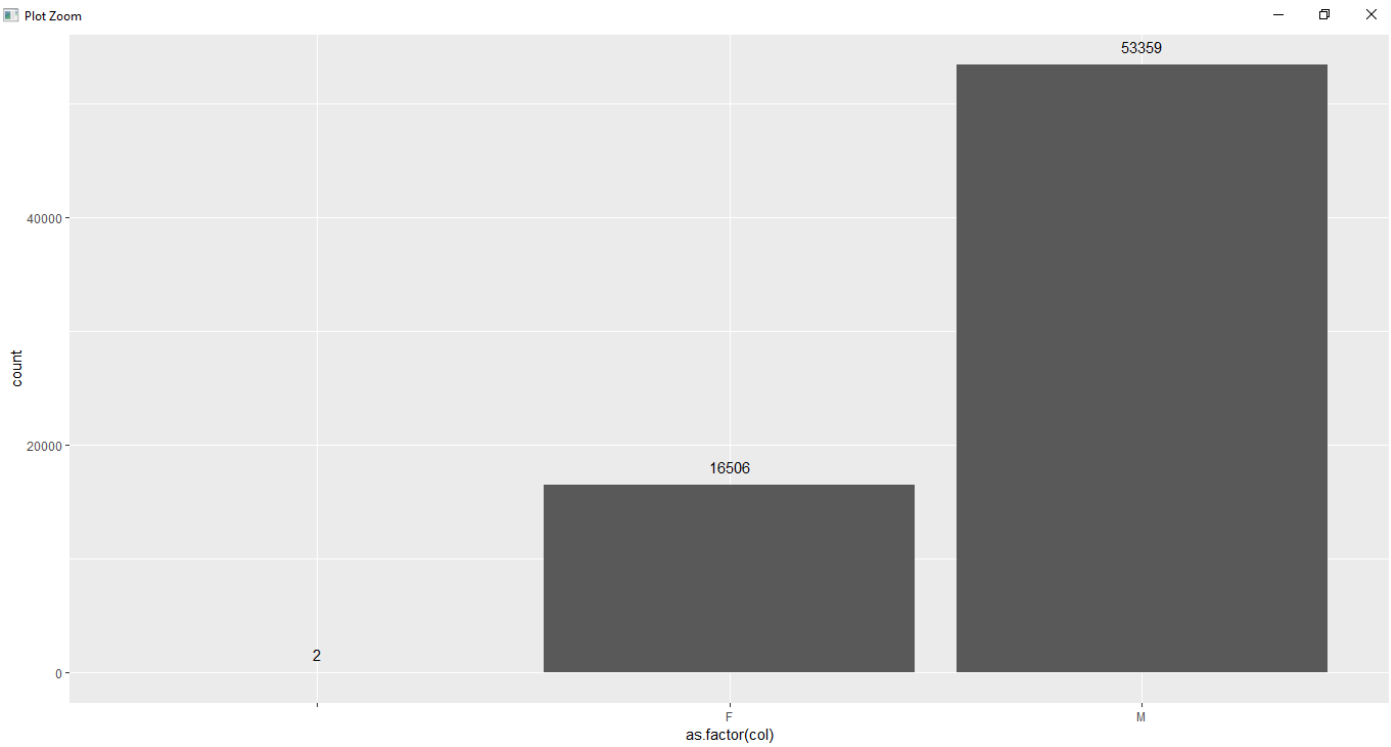
Univariate Analysis

We will analyse all the variables one by one. For EDA we will be using WOE and IV so before that we will start with univariate analysis.

1. *Age*



2. Gender

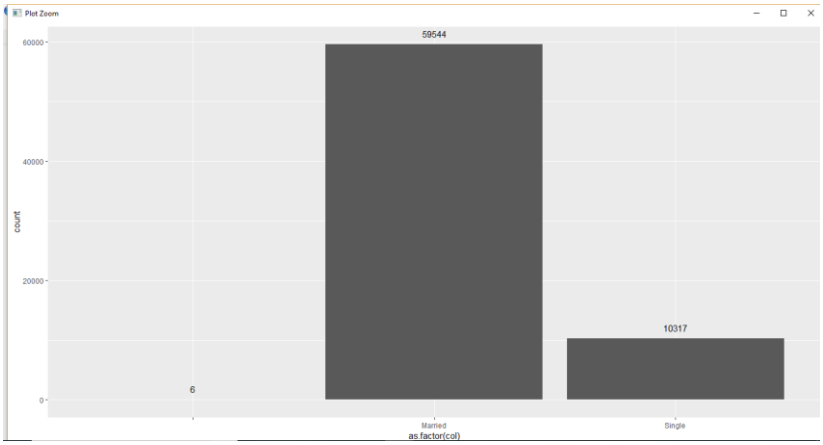


Total Application: 69870

Male Applicant: 53363 (76%)

Female Applicant: 16506 (24%)

3. Marital status



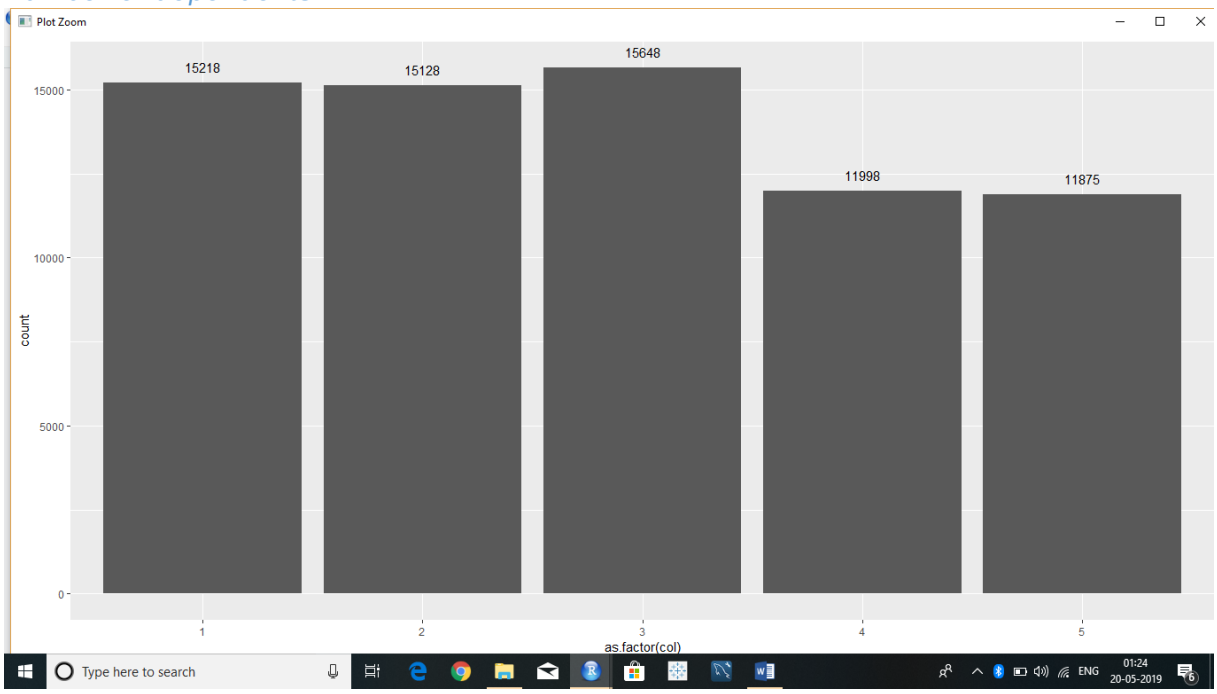
Total Application: 69870

Large Married people apply for credit card compare to single, might be due to their expenses.

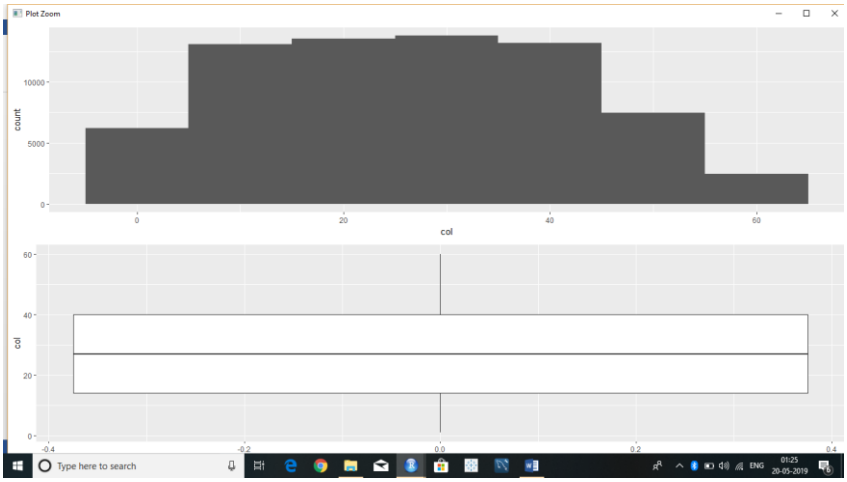
Married: 59547 (85%)

Single: 10317 (15%)

4. Number of dependents



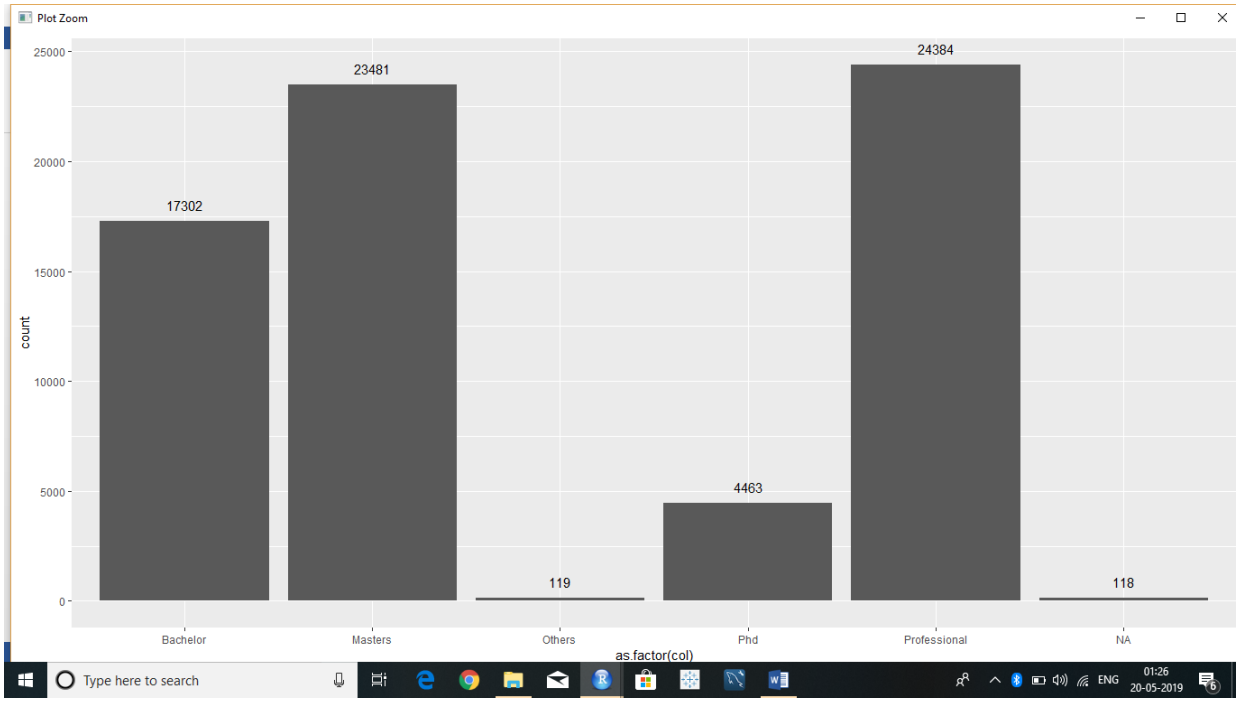
5. Income



Income is normally distributed and almost 80% among all applicant, earning between 10k to 50k.

No outlier in the data, just some invalid values.

6. Education

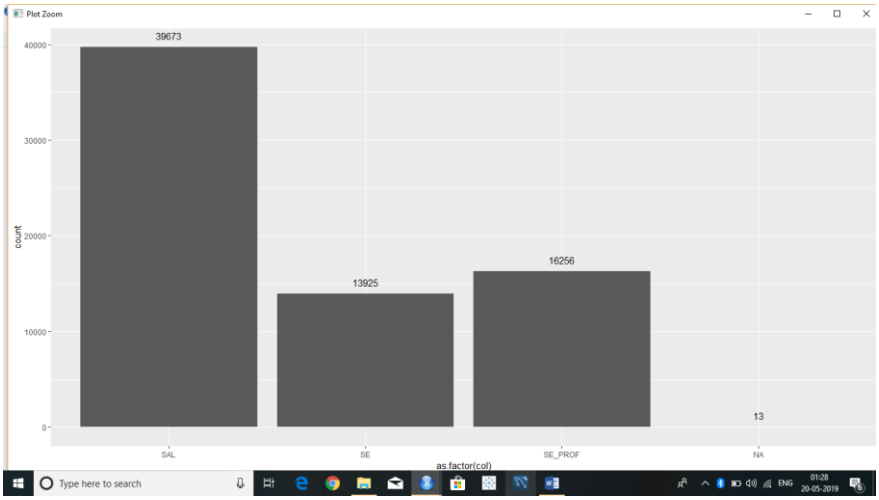


118 rows with blank. Replacing blank values to NA

47867 applicant are either Masters or Professional, which can be significant.

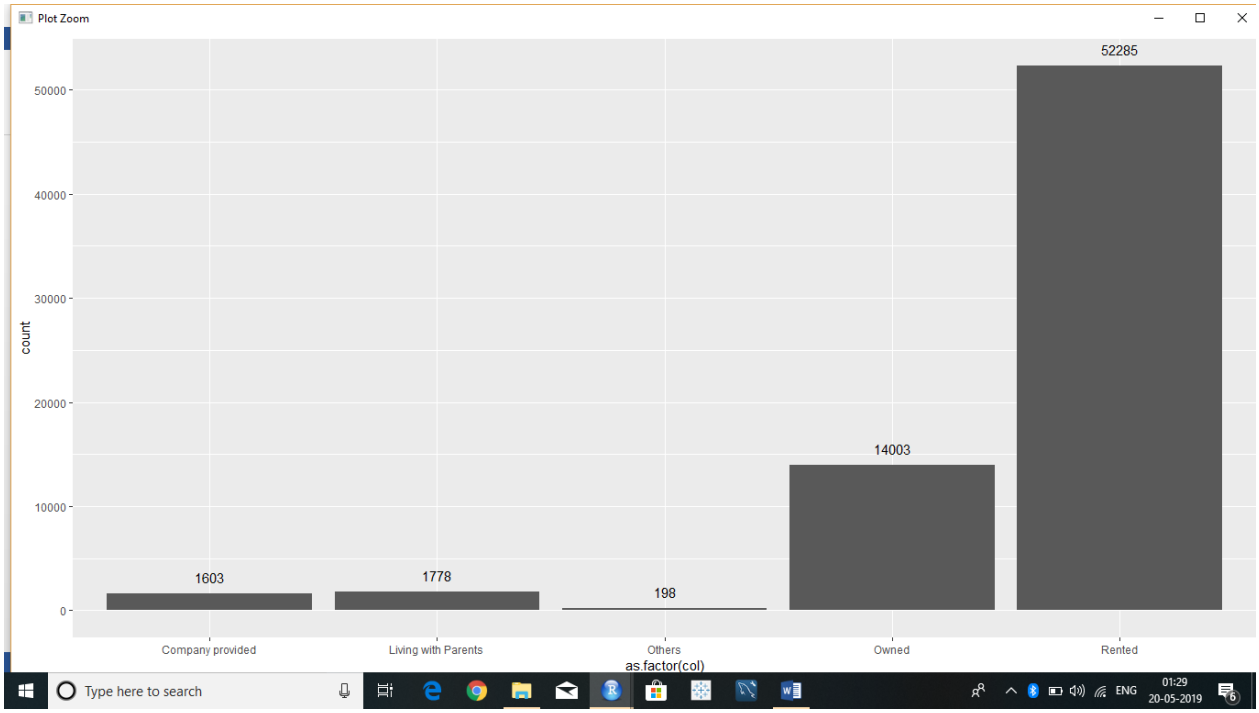
17302, applicant are Bachelor, which is bit less than Masters or professional.

7. Profession



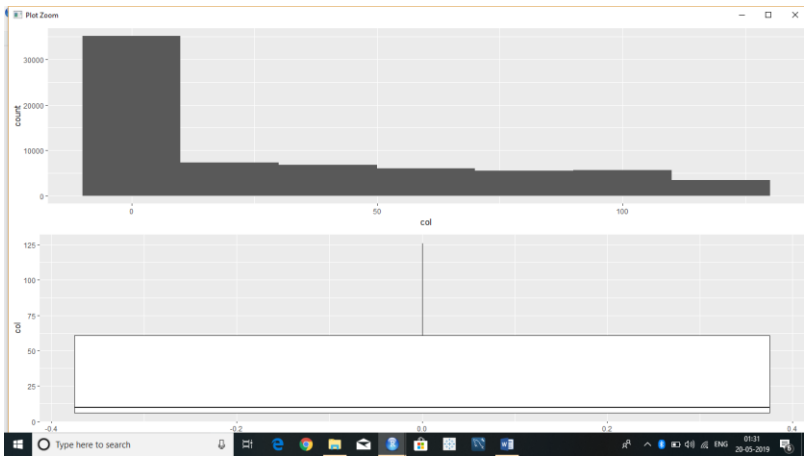
Almost 57% applicant are in "SAL" profession.
13 rows with blank.

8. Type of residence



Among all applicant 75% are rented.

9. Number if months in current residence



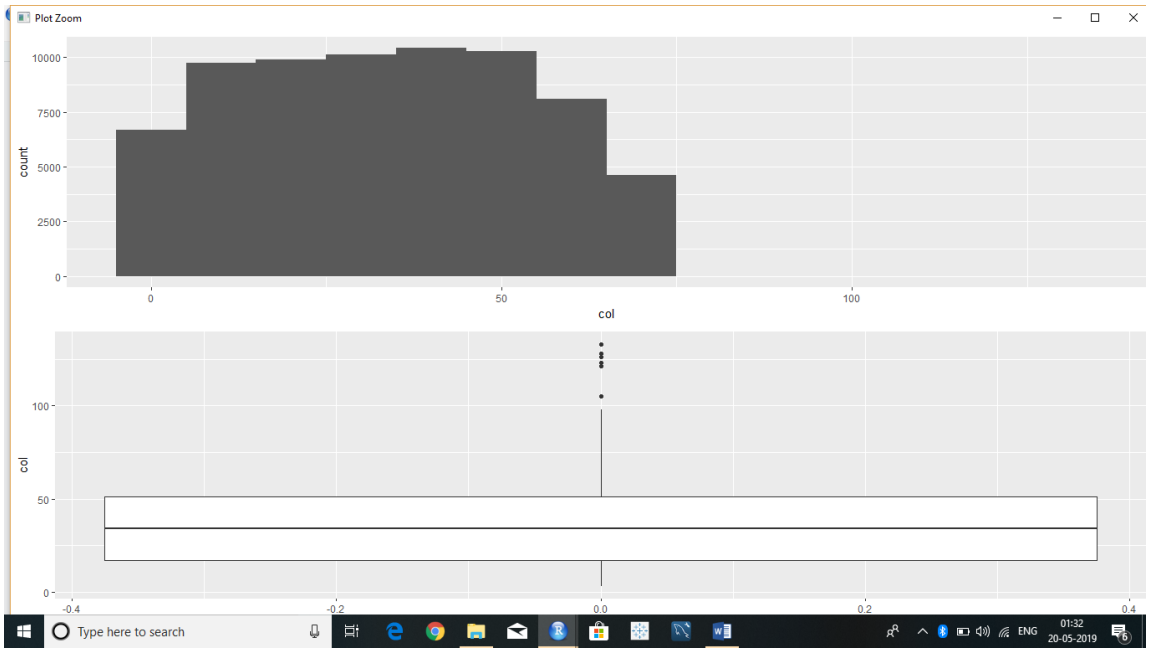
As per histogram, data is right skewed. Hence mean is higher than median.

Almost 50% applicant, spend between 6 to 10 months in current residence. Can be significant.

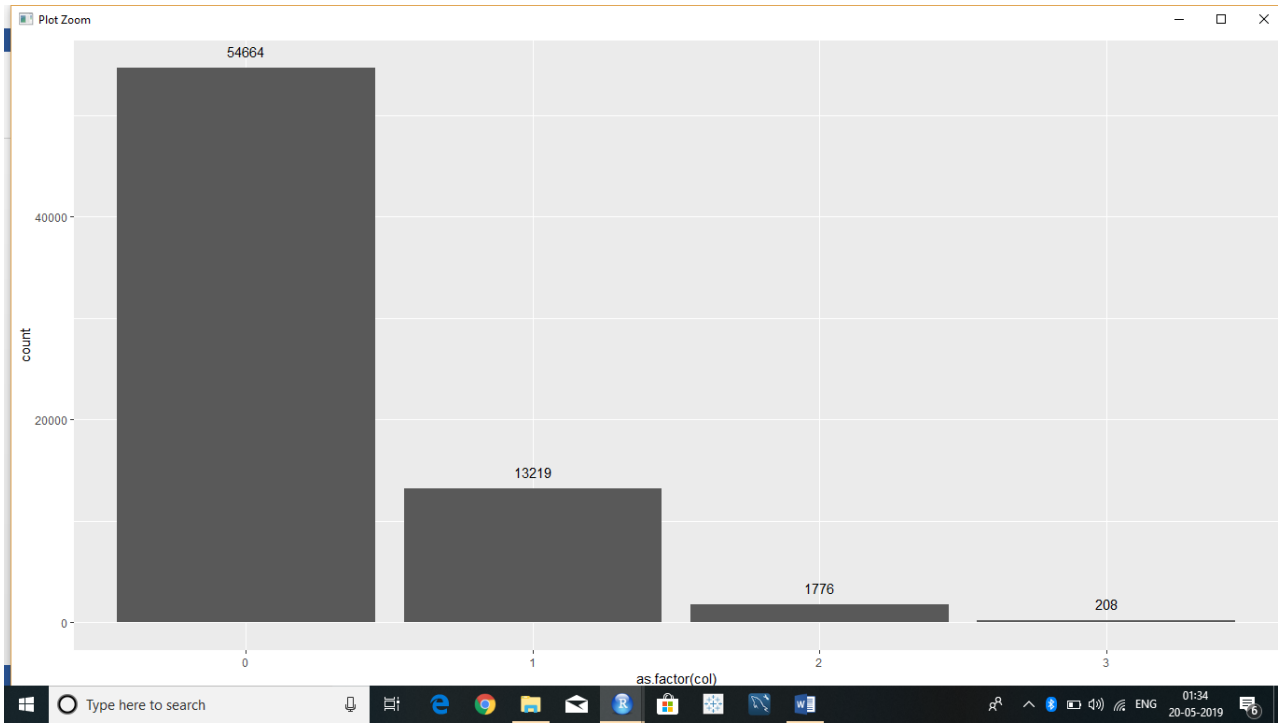
Because, most of them are rented. There are high changes that they will leave the residence frequently.

We can create bins for duration.

10. No.of.months.in.current.company



11. No.of.times.90.DPD.or.worse.in.last.6.months

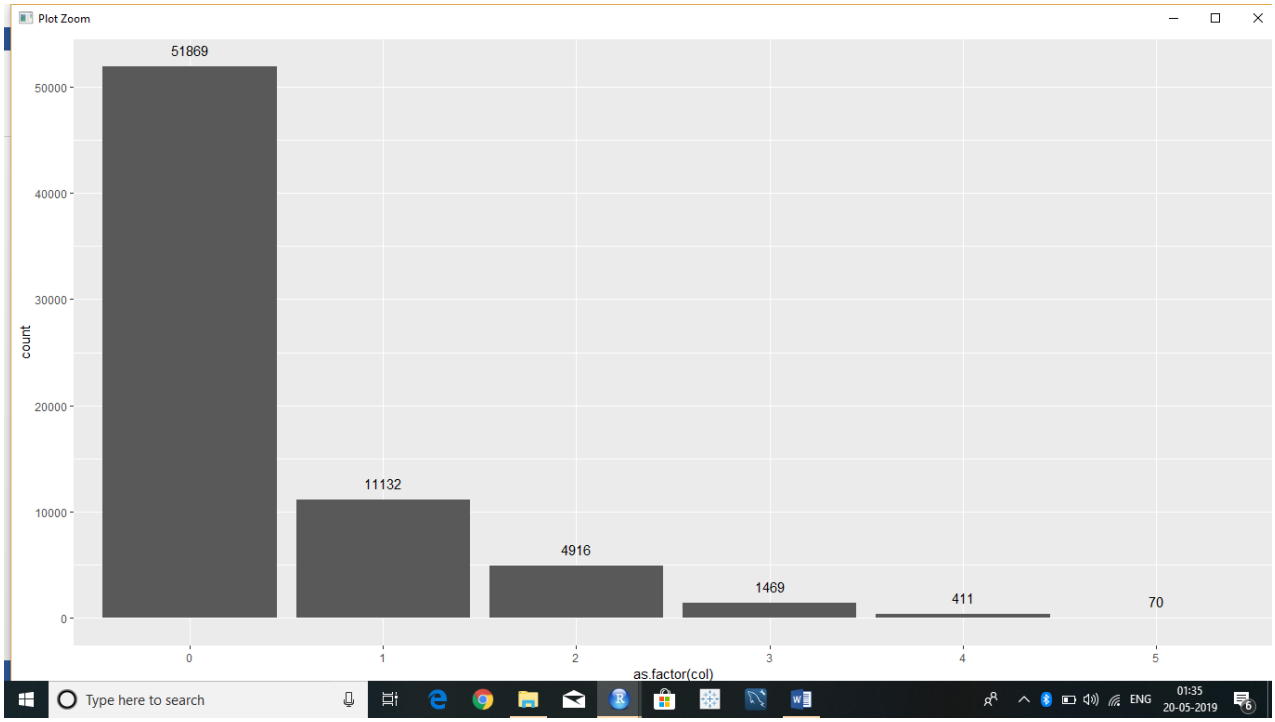


Mostly applicant having 0, 90 days dpd in last 6 months (78%).

Hence, 22% having at least 1, 90 days dpd in last 6 months.

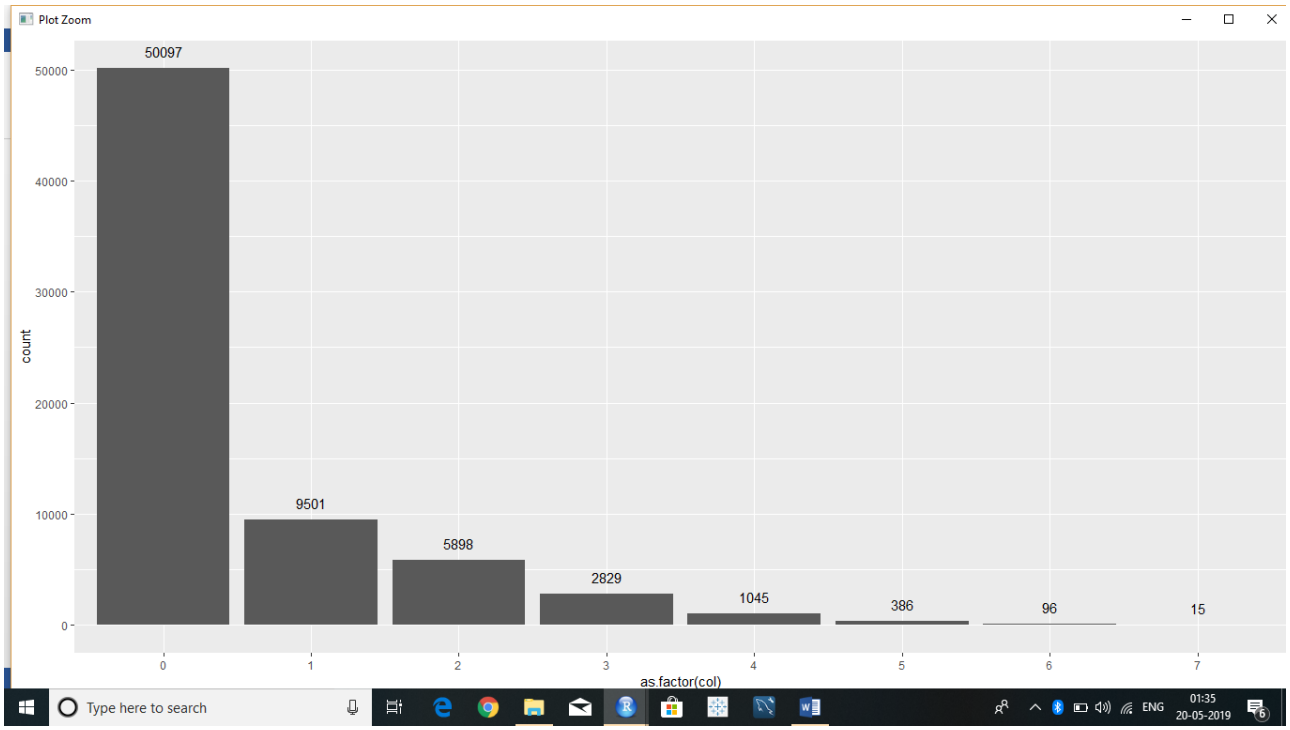
applicant having more or equal to 1, 90 days dpd past 6 months, almost 87% having 1 dpd.

12. No.of.times.60.DPD.or.worse.in.last.6.months



Mostly applicant having 0, 60 days dpd in last 6 months (74%).
Hence, 26% having at least 1, 60 days dpd in last 6 months. Which is 4% higher than 90 days dpd in 6 months.
applicant having more or equal to 1, 60 days dpd past 6 months, almost 62% having 1 time 60 days dpd and 28% having 2 time 60 days dpd.

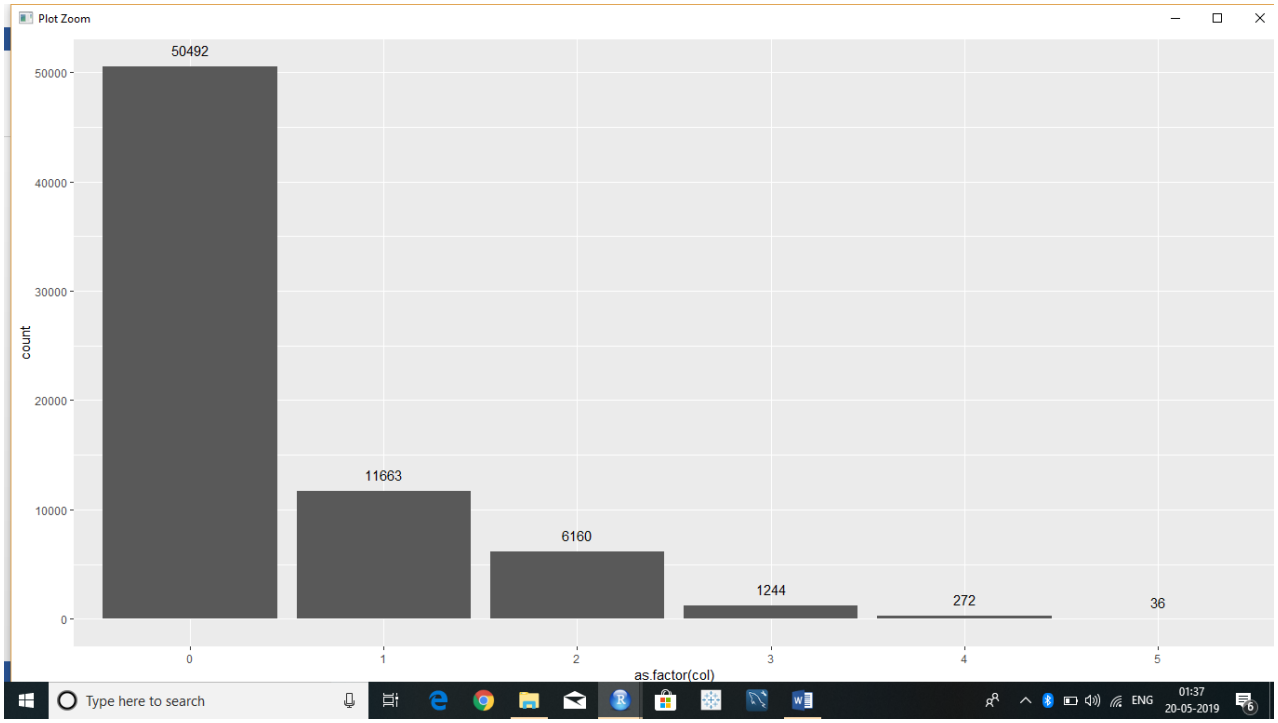
13. No.of.times.30.DPD.or.worse.in.last.6.months



Mostly applicant having 0, 30 days dpd in last 6 months (72%).
Hence, 28% having at least 1, 30 days dpd in last 6 months. Which is 2% higher than 90 days dpd in 6 months.

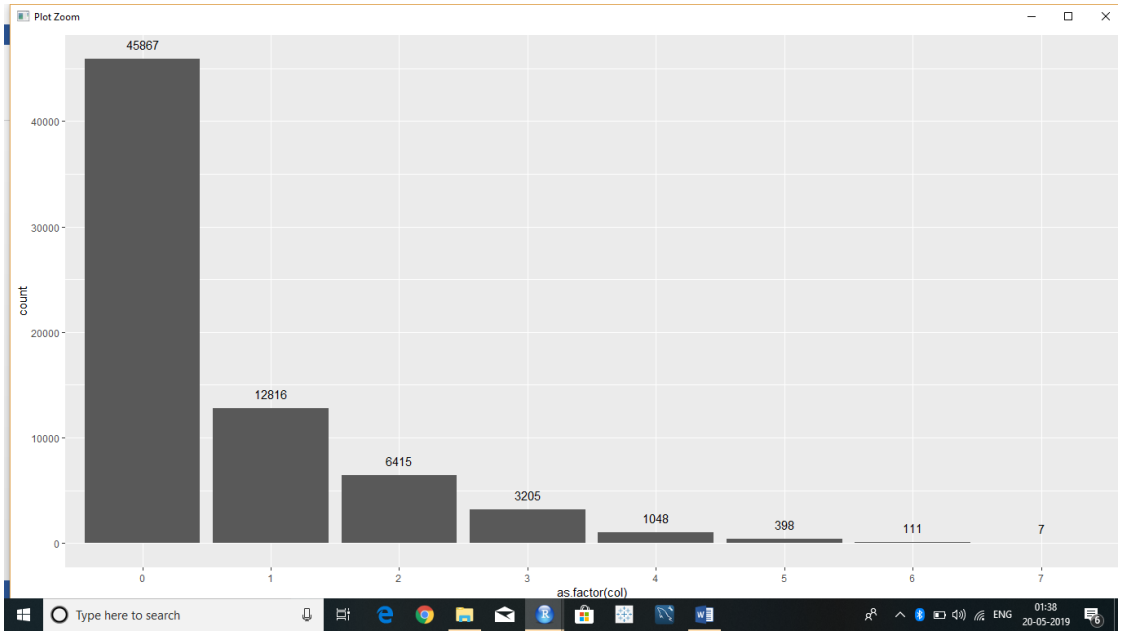
Applicant having more or equal to 1, 30 days dpd past 6 months, almost 78% having 1 or 2 time 60 days dpd.

14. No.of.times.90.DPD.or.worse.in.last.12.months



Mostly applicant having 0, 90 days dpd in last 12 months (72%).
Hence, 28% having at least 1, 90 days dpd in last 12 months. Which is 6% higher than 90 days dpd in 6 months.
applicant having more or equal to 1, 90 days dpd past 12 months, almost 92% having 1 or 2 time 90 days dpd.

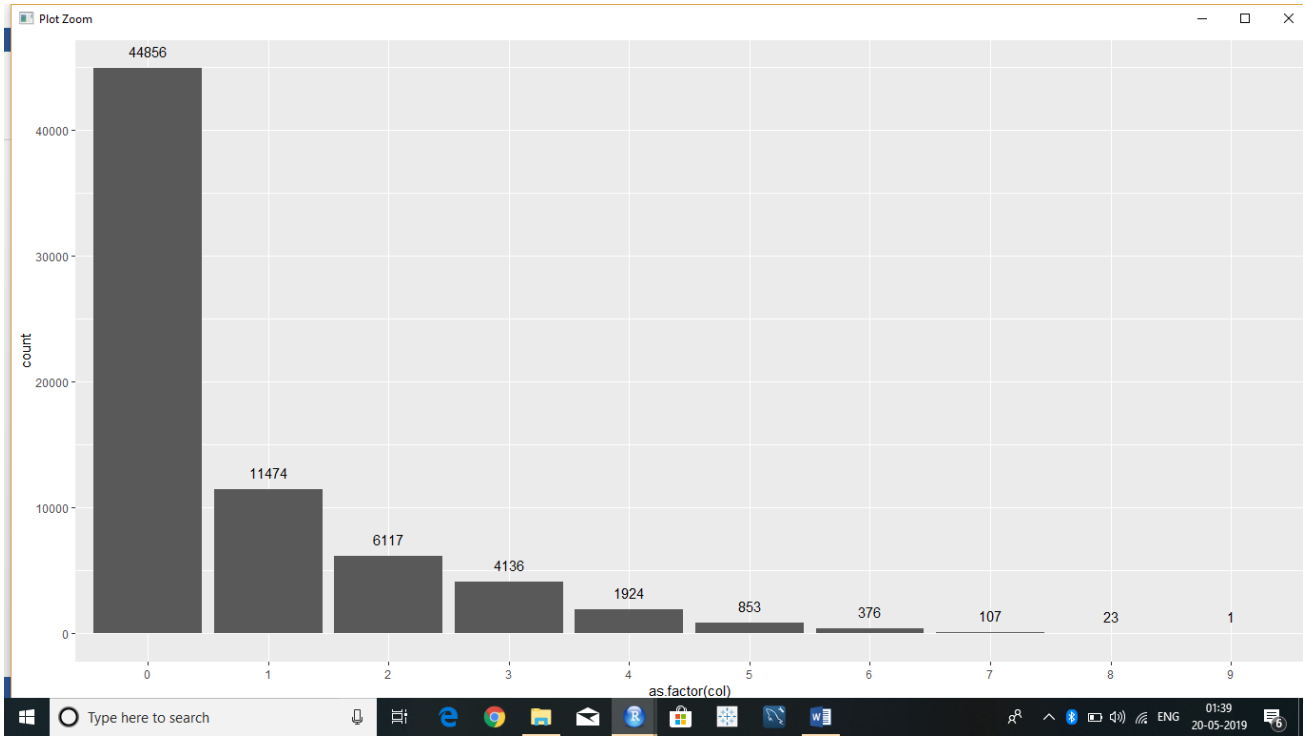
15. No.of.times.60.DPD.or.worse.in.last.12.months



66% applicant having 0, 60 days dpd in last 12 months.
Hence, 34% having at least 1, 60 days dpd in last 12 months.

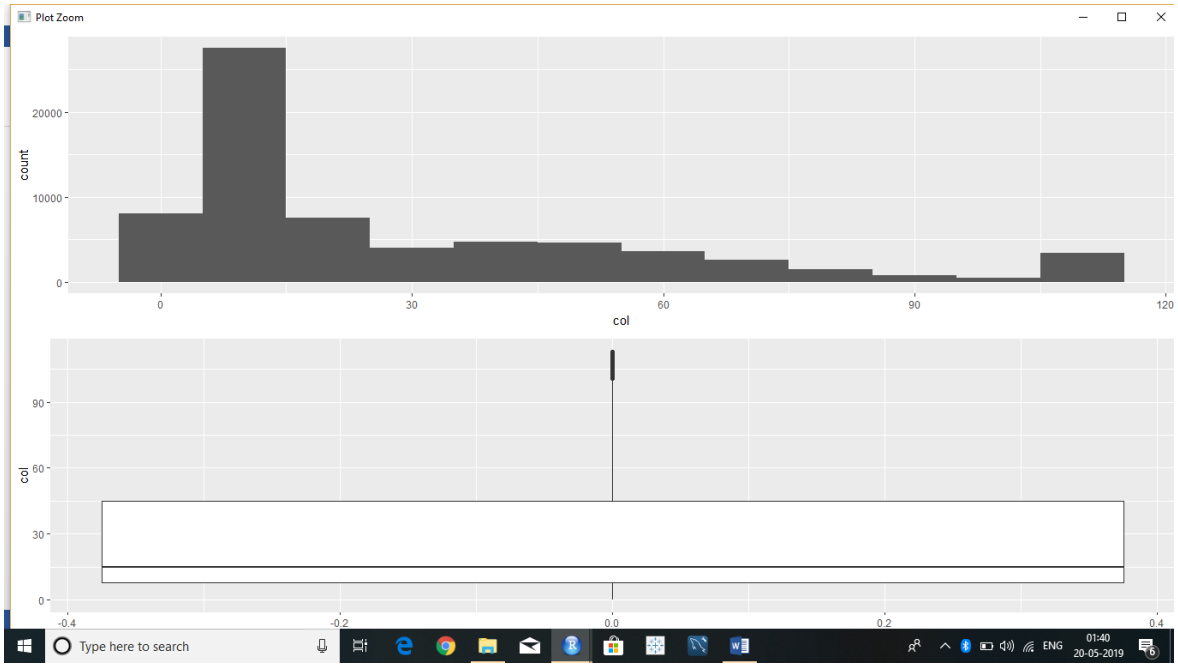
applicant having more or equal to 1, 60 days dpd past 12 months, almost 80% having 1 or 2 time 90 days dpd.

16. No.of.times.30.DPD.or.worse.in.last.12.months



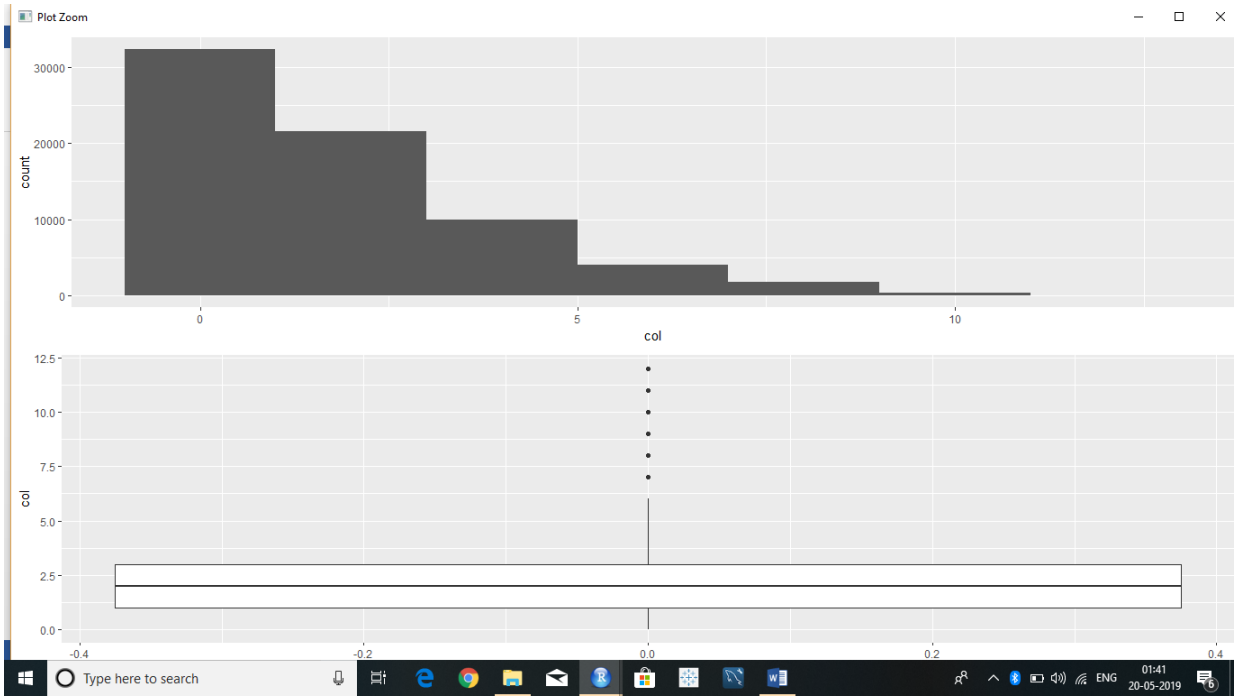
64% applicant (44858) having 0, 30 days dpd in last 12 months.
Hence, 36% having at least 1, 30 days dpd in last 12 months.
applicant having more or equal to 1, 30 days dpd past 12 months, almost 70% having 1 or 2 time 90 days dpd.

17. Avgas.CC.Utilization.in.last.12.months



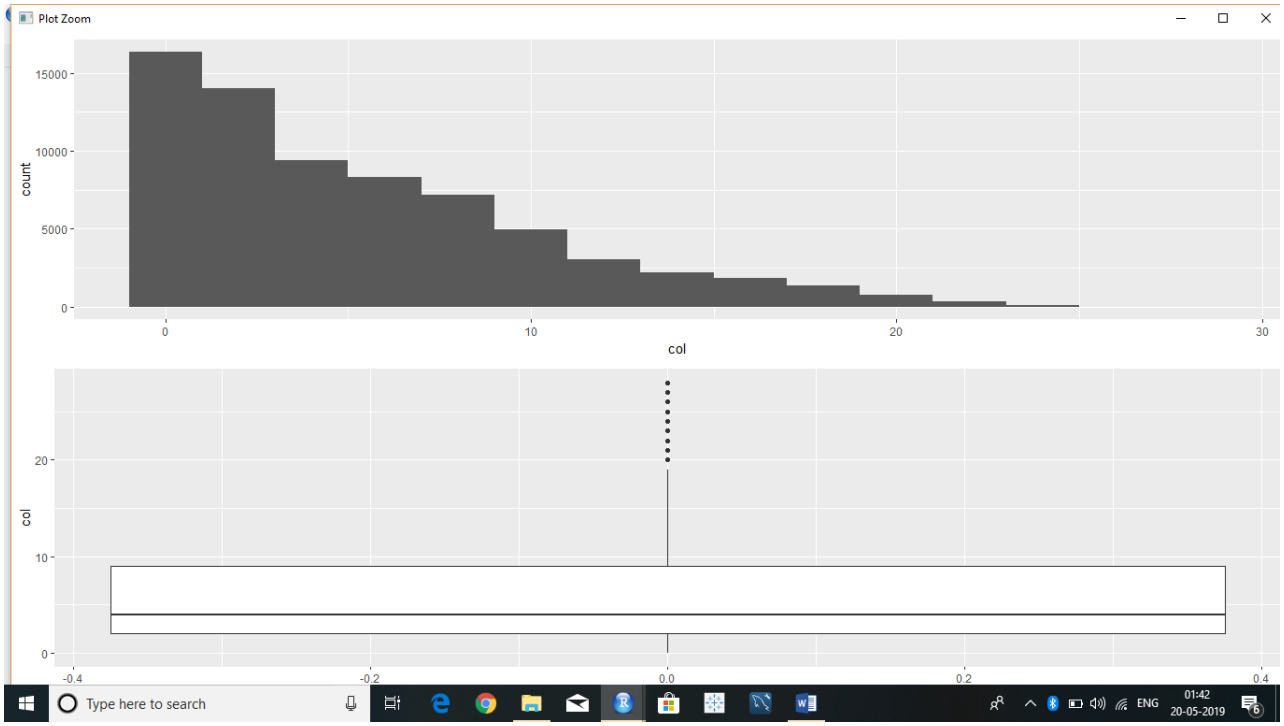
There are outliers after 94 percentile.
Data is right skewed and hence median is lesser than mean.
Mostly applicant are utilizing card 0 to 30 times.
Mostly they are utilizing between 10 to 20.

18. No.of.trades.opened.in.last.6.months



77% applicants have 3 or less open trades in last 6 months.

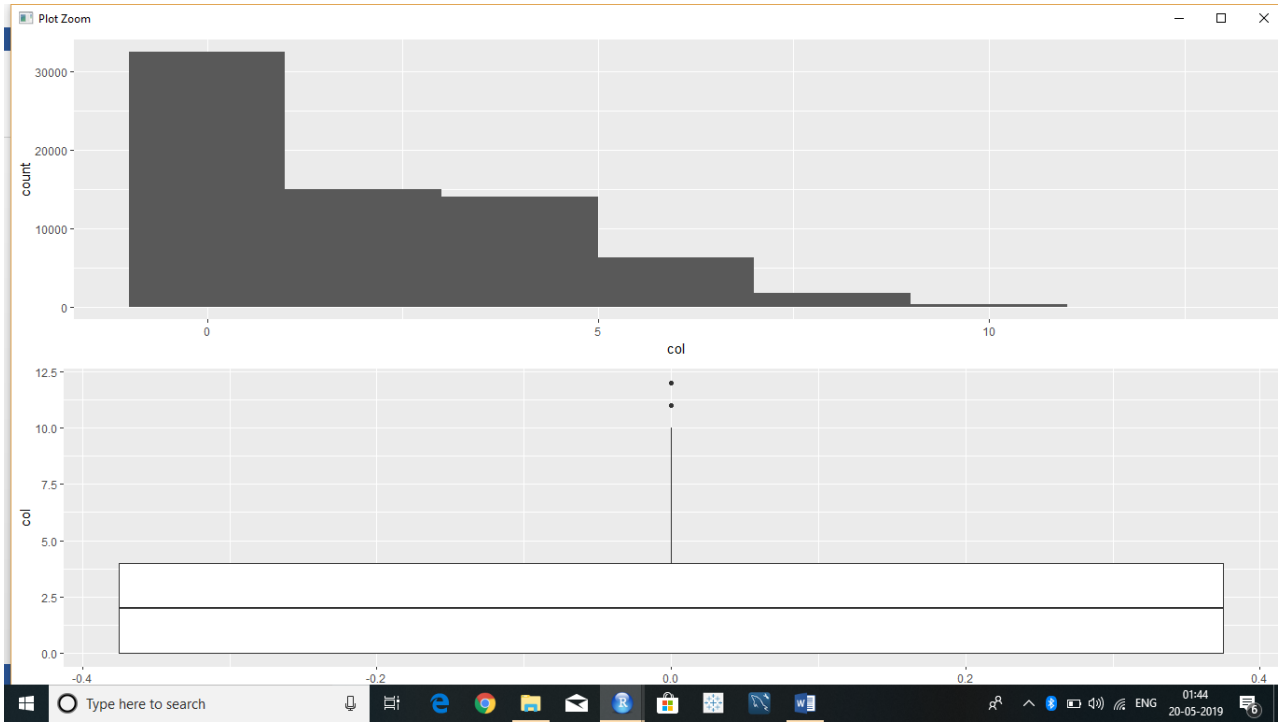
19. No.of.trades.opened.in.last.12.months



20. No.of.PL.trades.opened.in.last.6.months

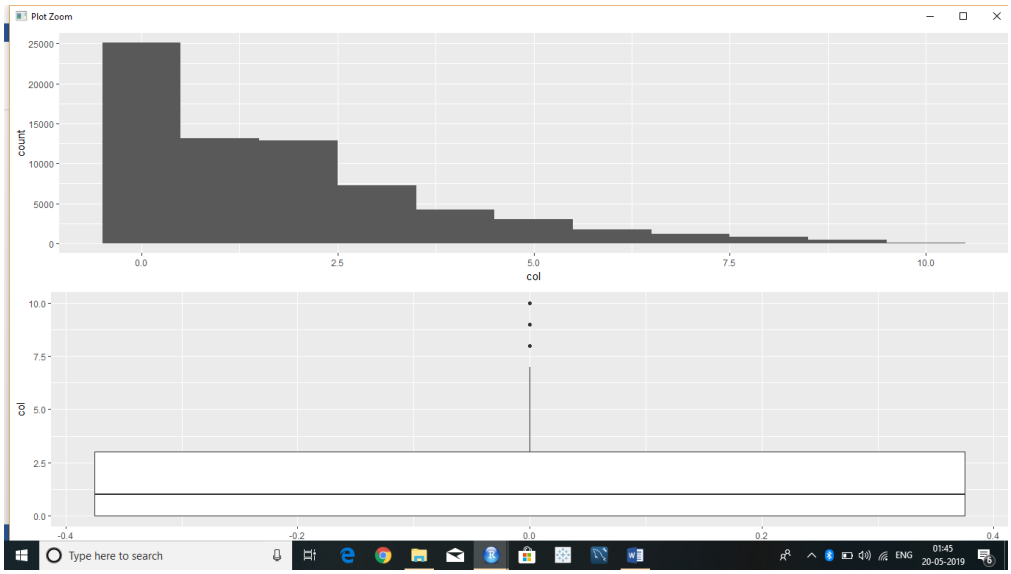
Almost 44% applicant having 0 open pl trade in last 6 months.

21. No.of.PL.trades.opened.in.last.12.months



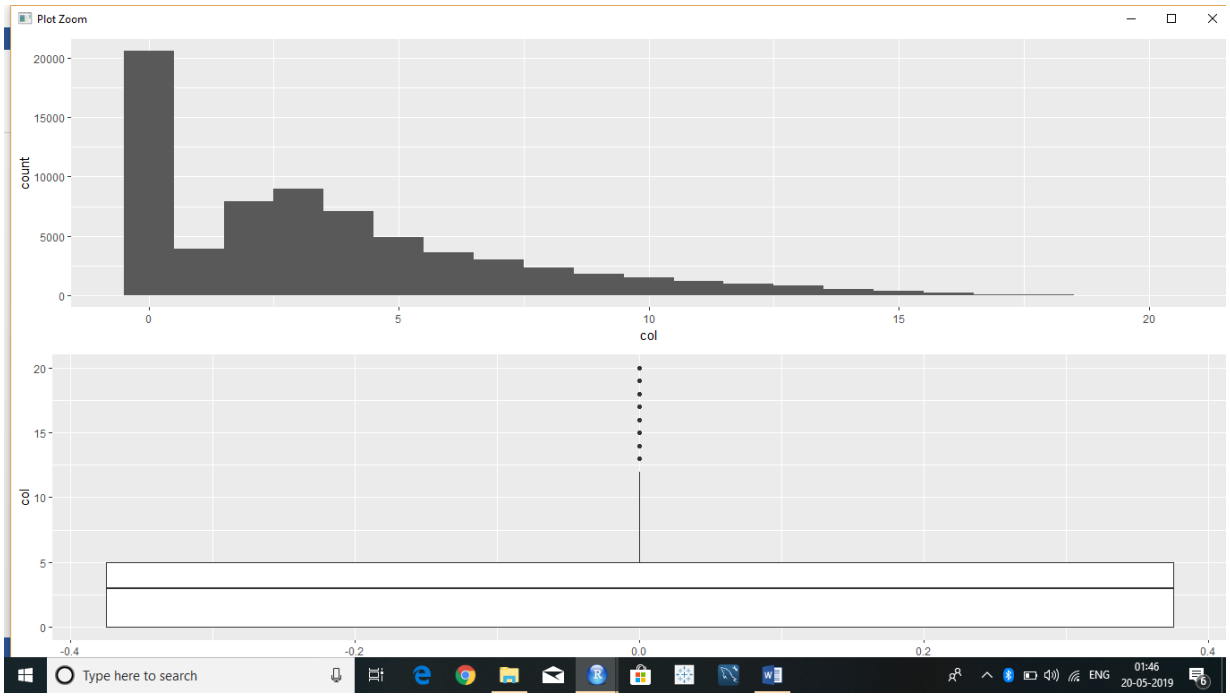
Almost 36% applicant having 0 open pl trade in last 12 months.
Means 8% more applicants have 1 or more, open pl trade compare to 6 months.

22. No.of.Inquiries.in.last.6.months..excluding.home...auto. Loans.



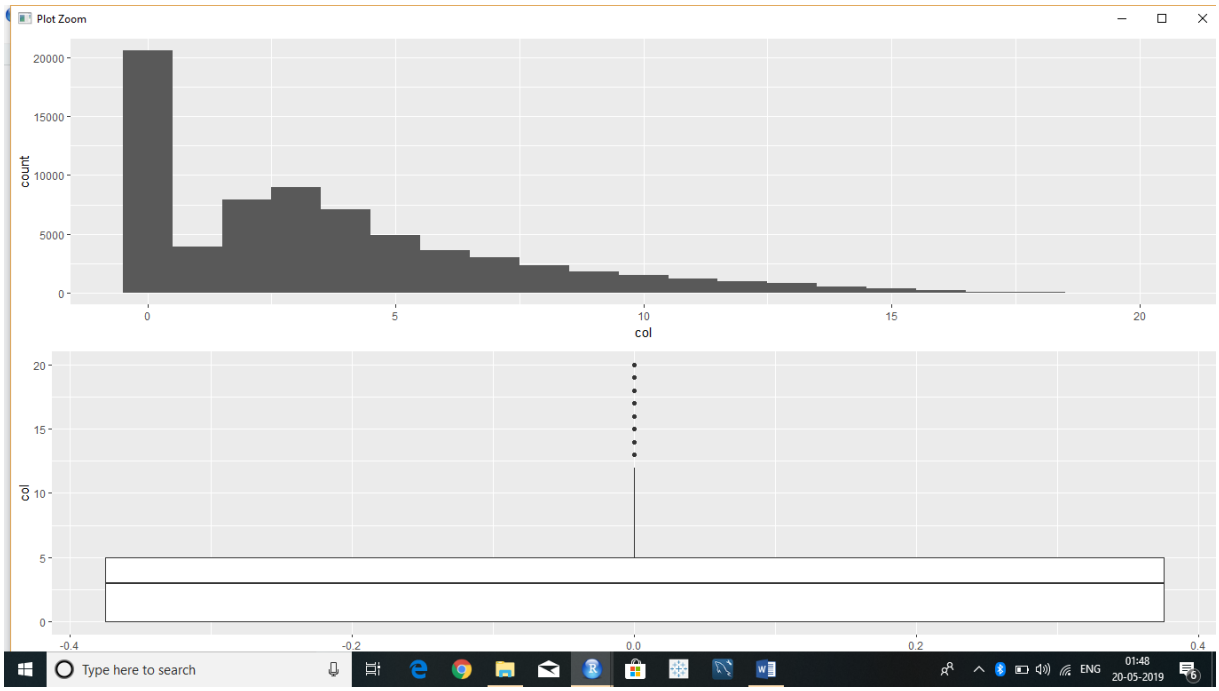
almost 35% applicants did 0 times cc inquiries.
47% applicant did cc inquiries 1 to 3 times.

23. No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.



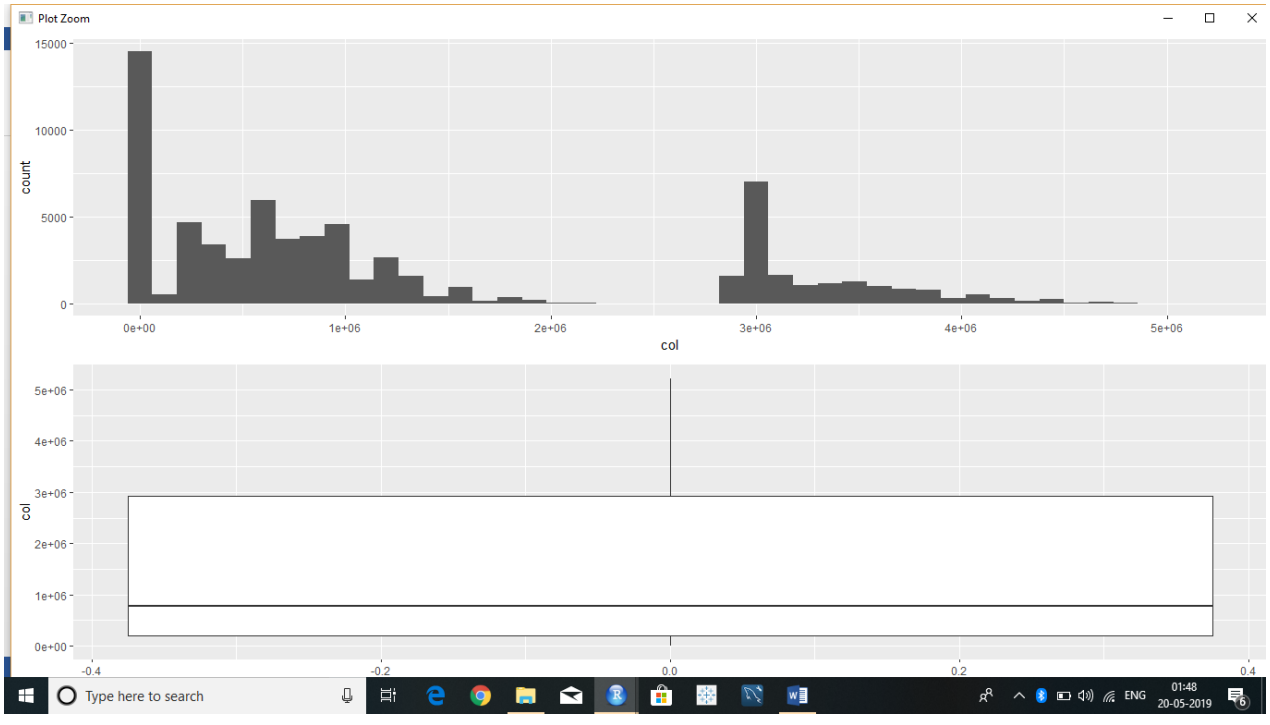
Almost 29% applicants did 0 times credit card inquiries. Which is less than credit card inquiries in 12 months.
Means over a period of time, applicant do more inquiry about credit card.
57% applicant did credit card inquiries 1 to 6 times.

24. Presence.of.open.home.loan

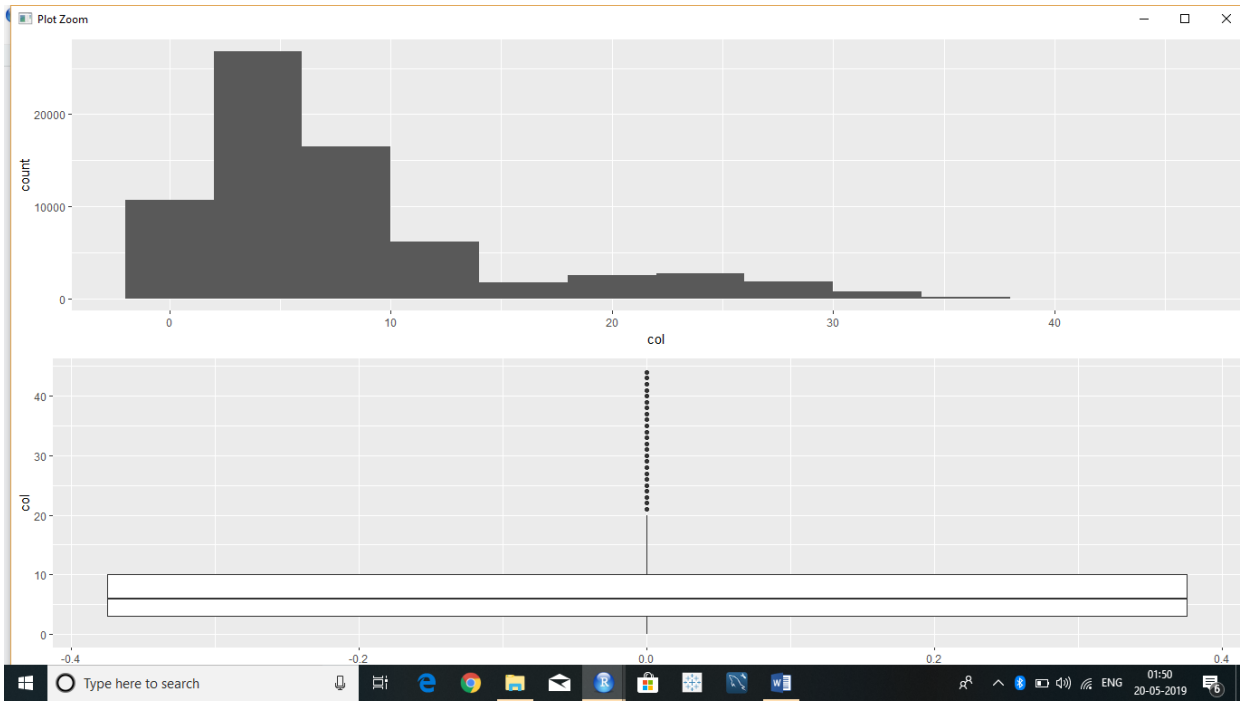


26% of applicants are having home loans

25. Outstanding Balance

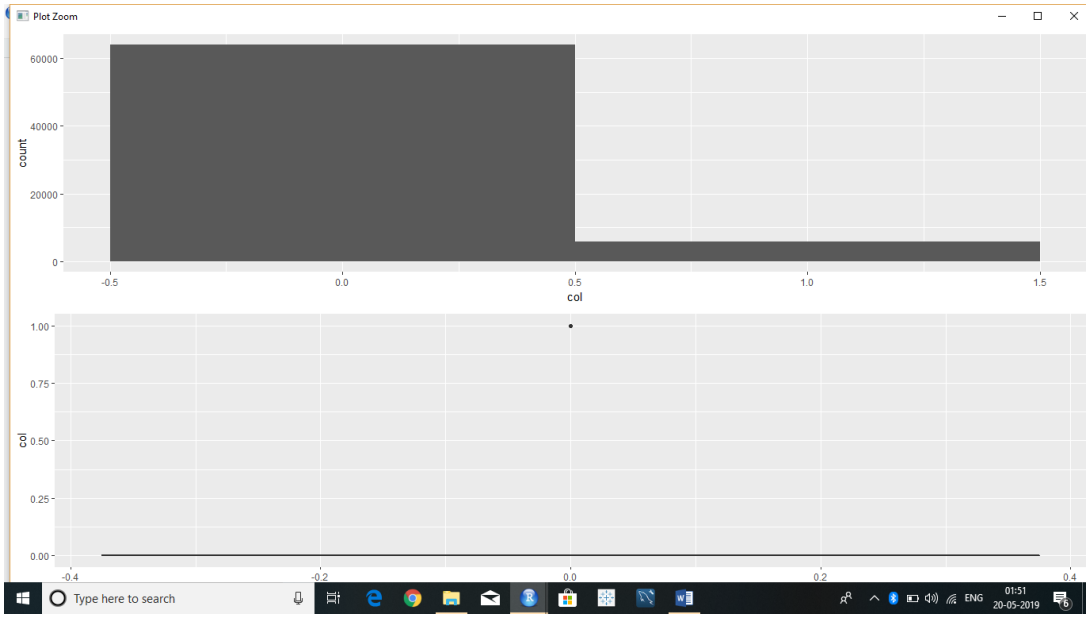


26. Total.No.of.Trades



Almost 88% applicant having total trades between 0 to 16.

27. *Presence.of.open.auto.loan*

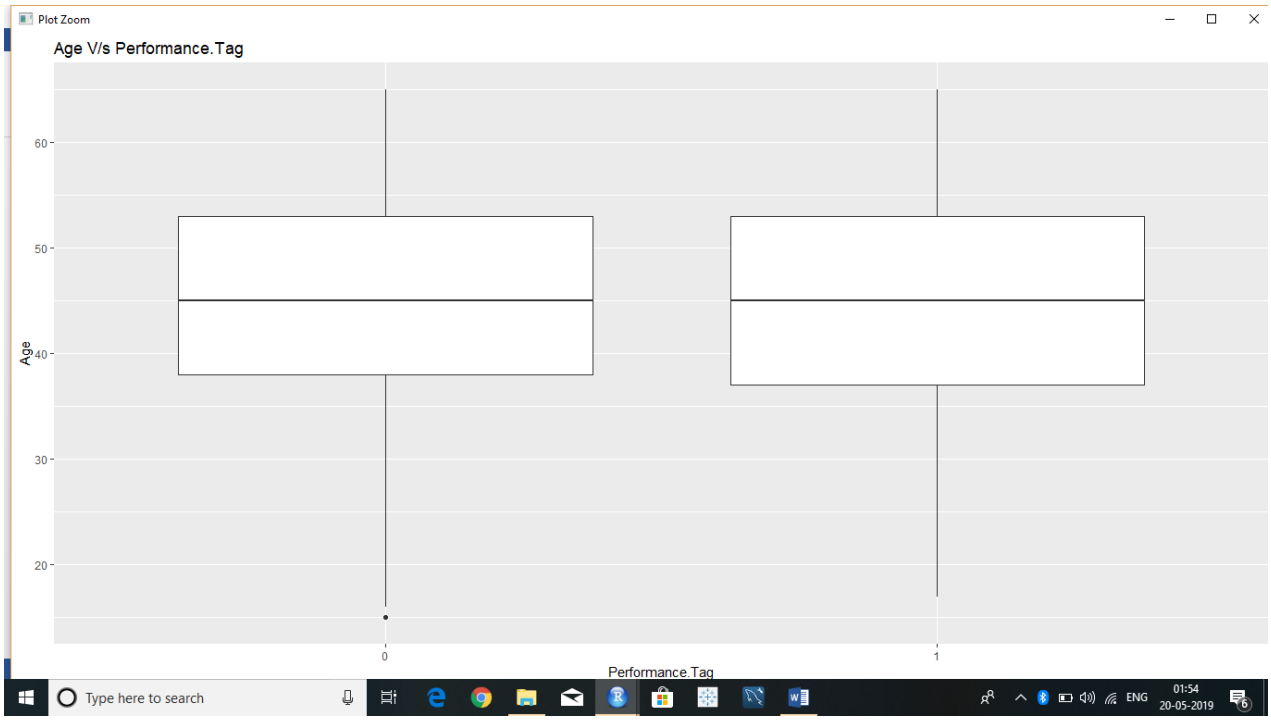


8% applicant has auto loan.

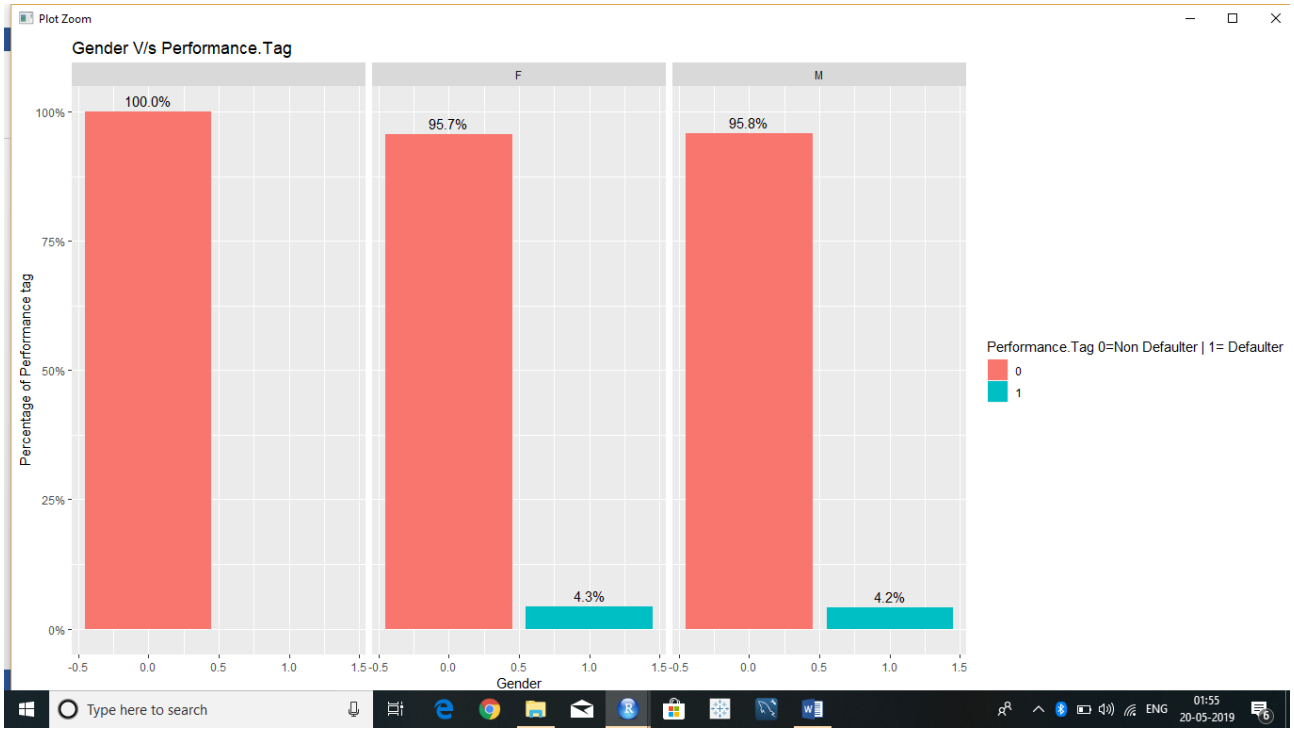
BIVARIATE ANALYSIS

Bivariate analysis is done analysing different variables against performance tag as that is out target variable.

1. *Age vs. Performance.Tag*

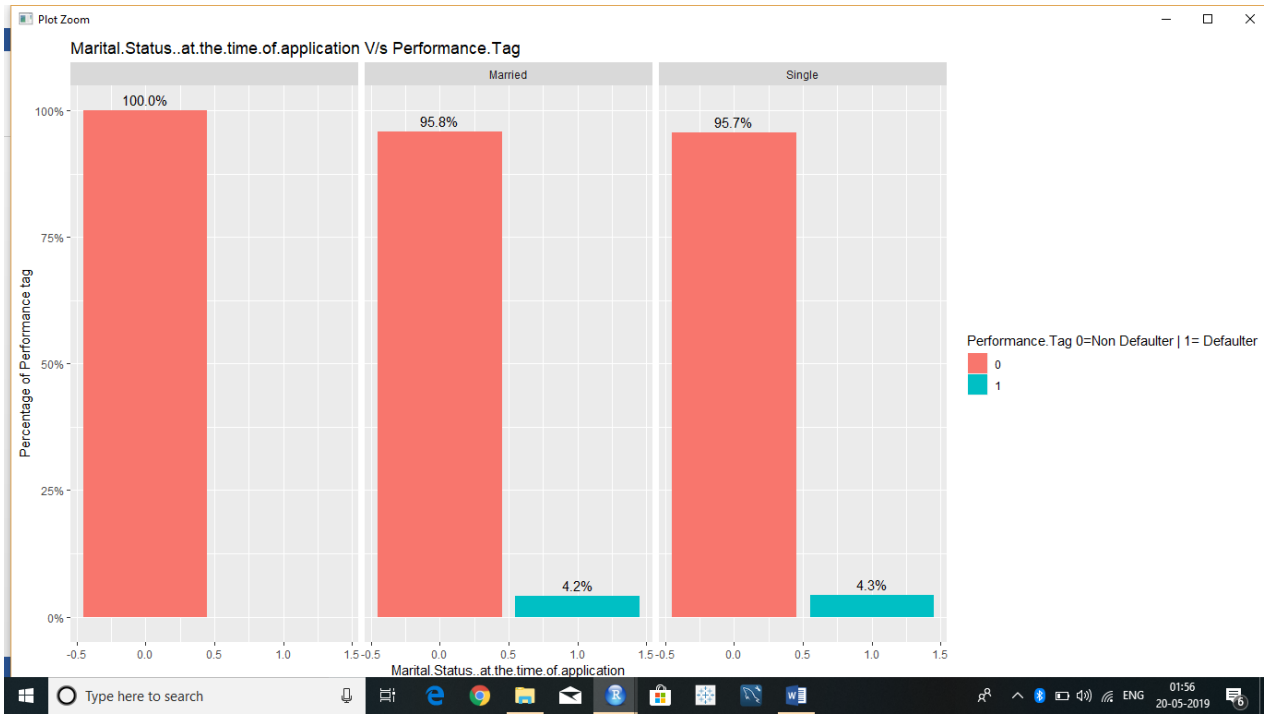


2. *Gender vs. Performance.Tag*

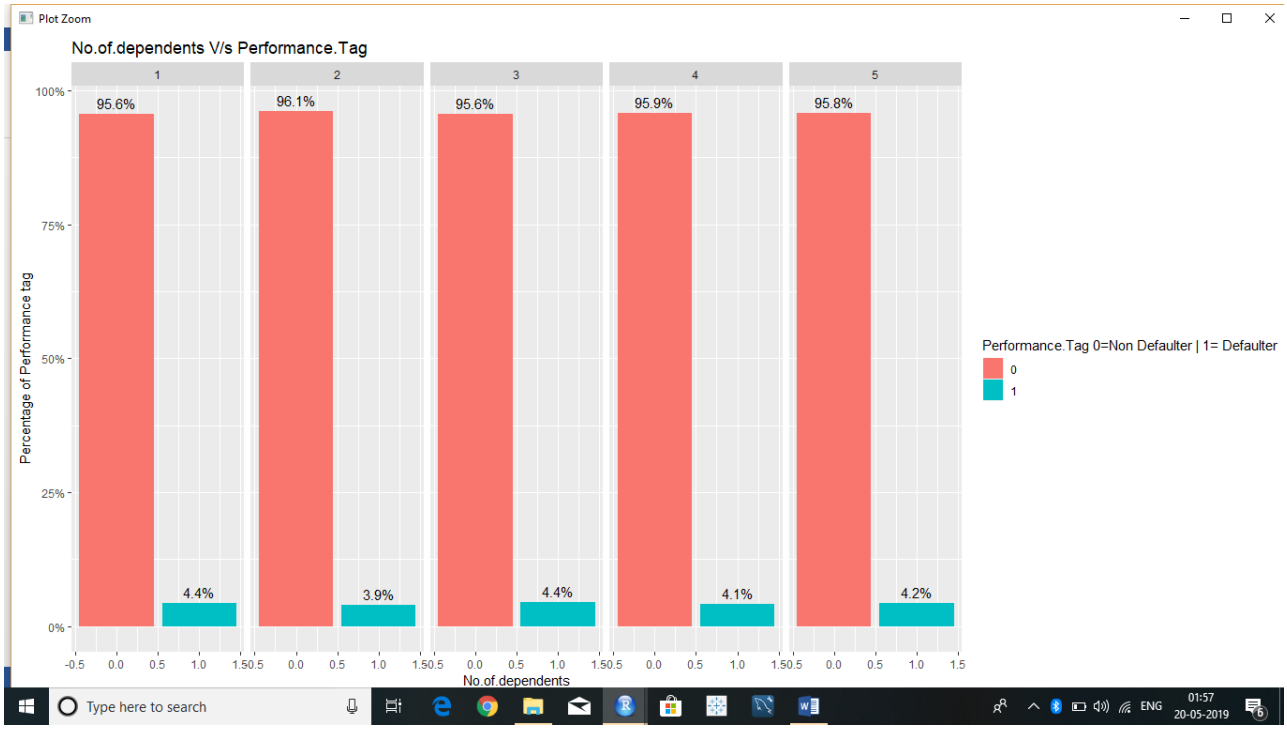


4% male and female have defaulted

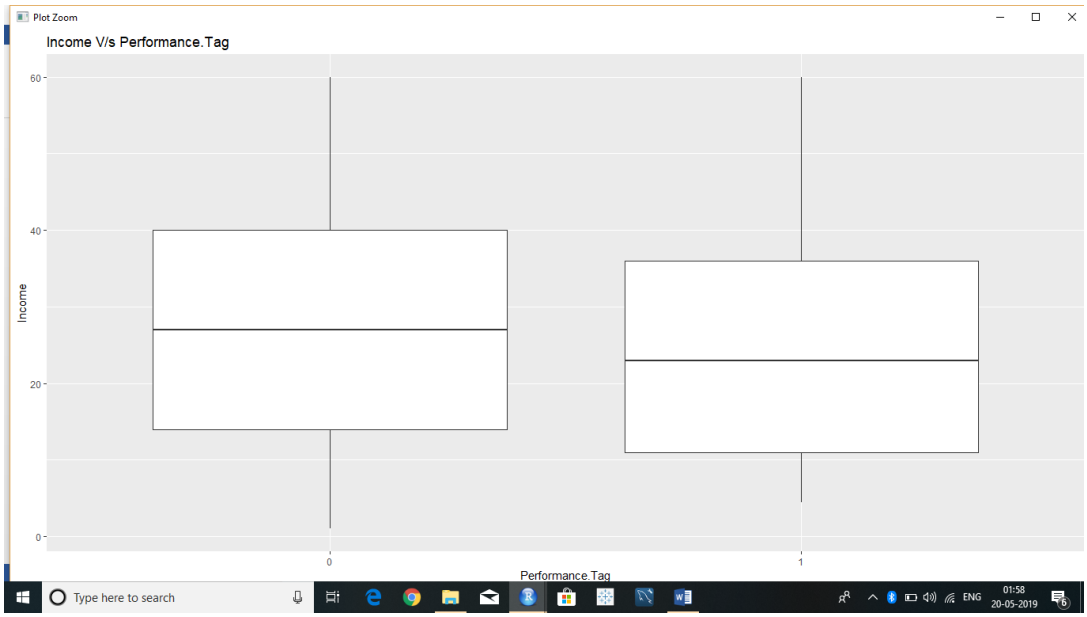
3. *Marital.Status..at.the.time.of.application vs. Performance.Tag*



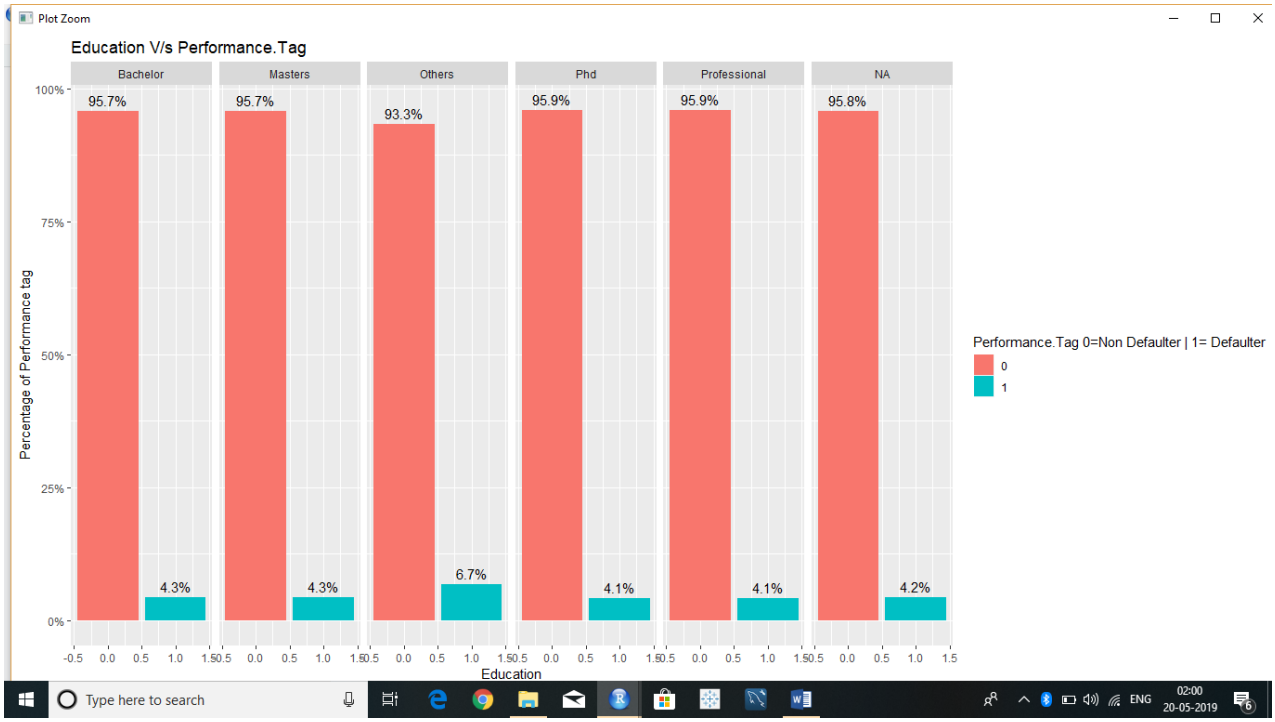
4. No.of.dependents vs. Performance.Tag



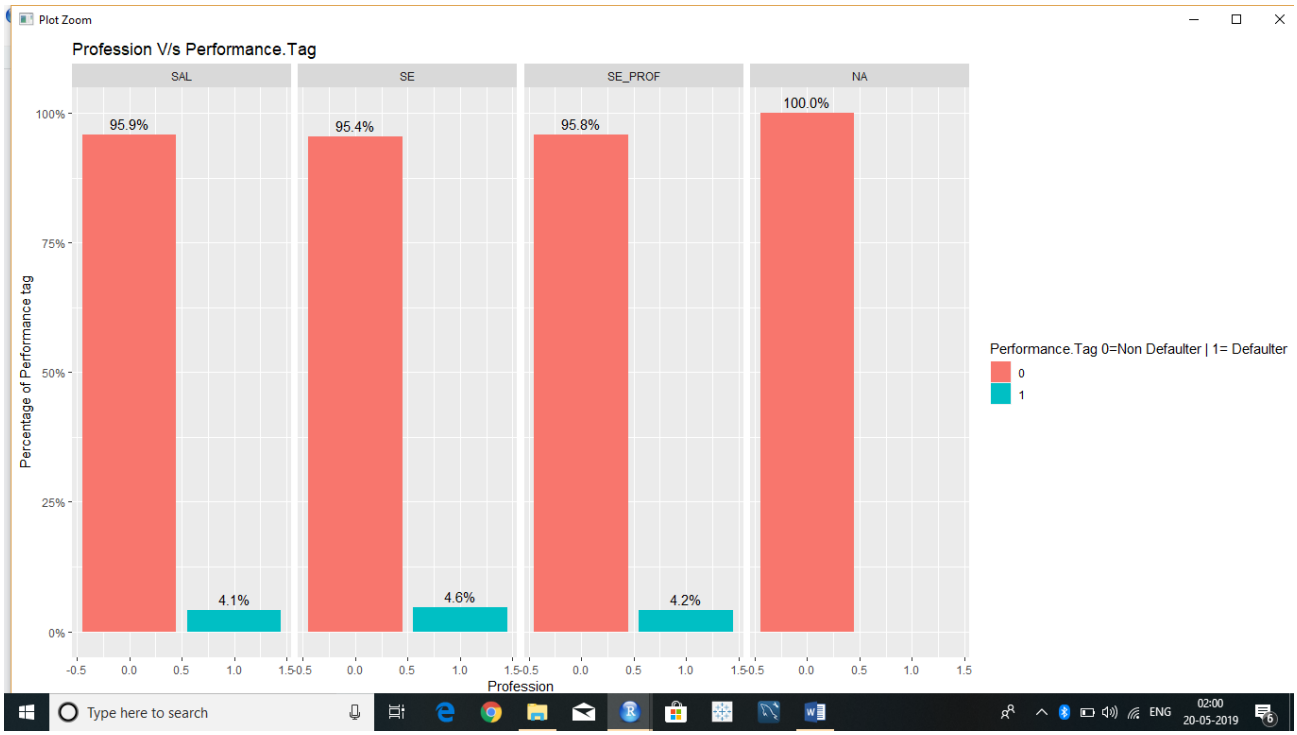
5. Income vs. Performance.Tag



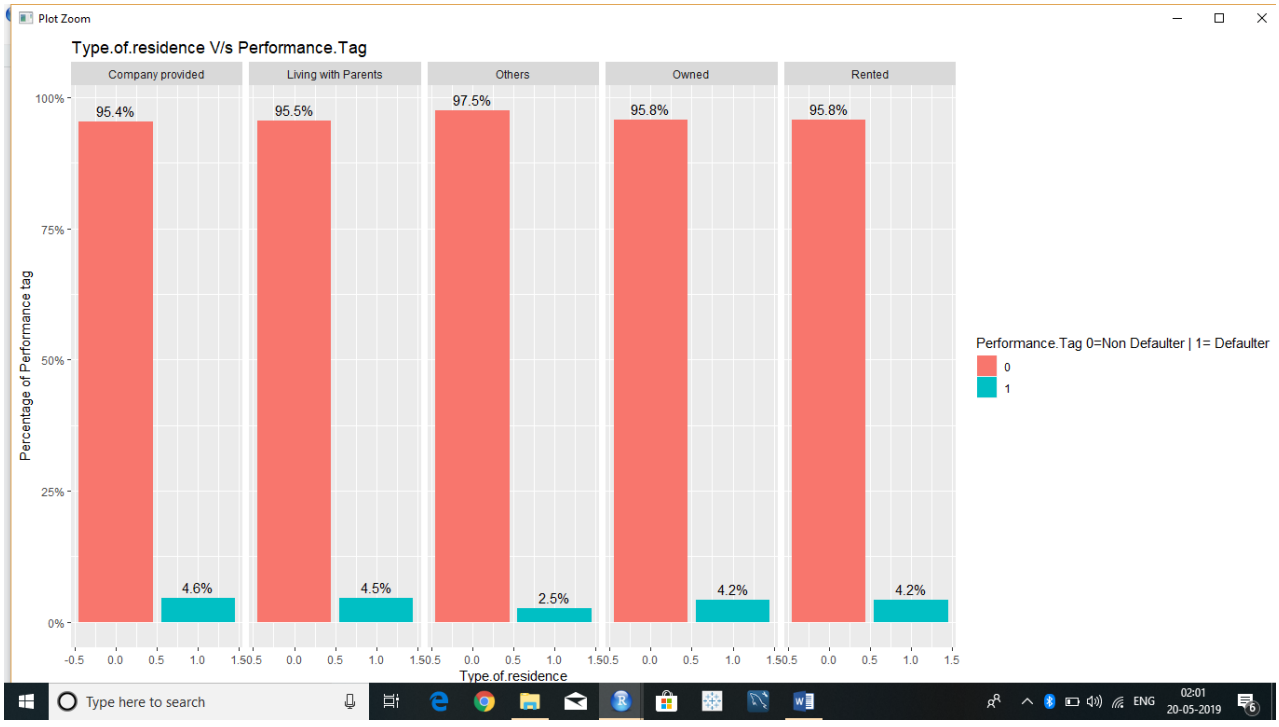
6. Education vs. Performance.Tag



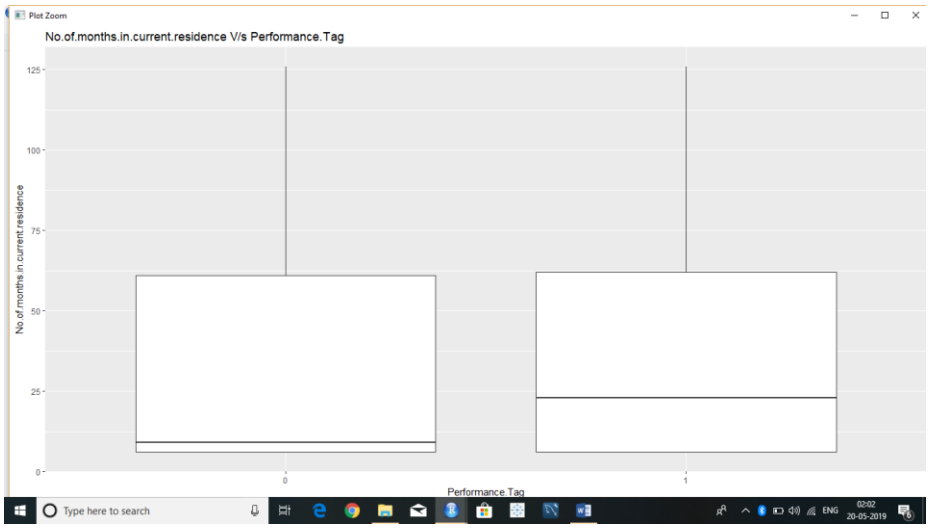
7. Profession vs. Performance.Tag



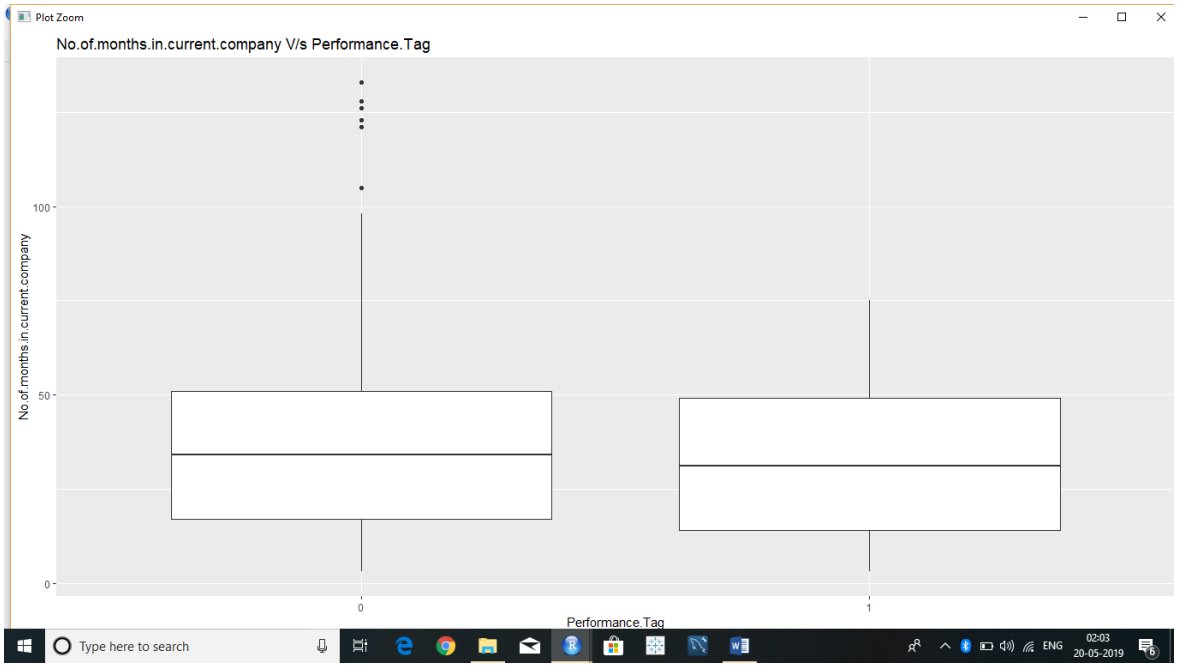
8. Type.of.residence vs. Performance.Tag



9. No.of.months.in.current.residence vs. Performance.Tag

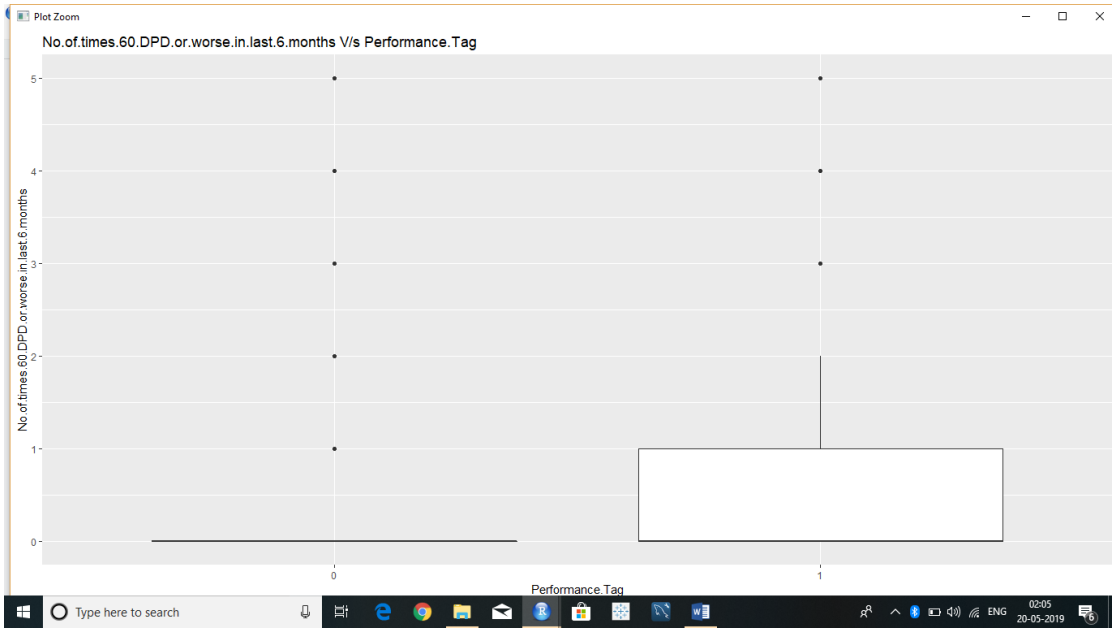


10. No.of.months.in.current.company vs. Performance.Tag

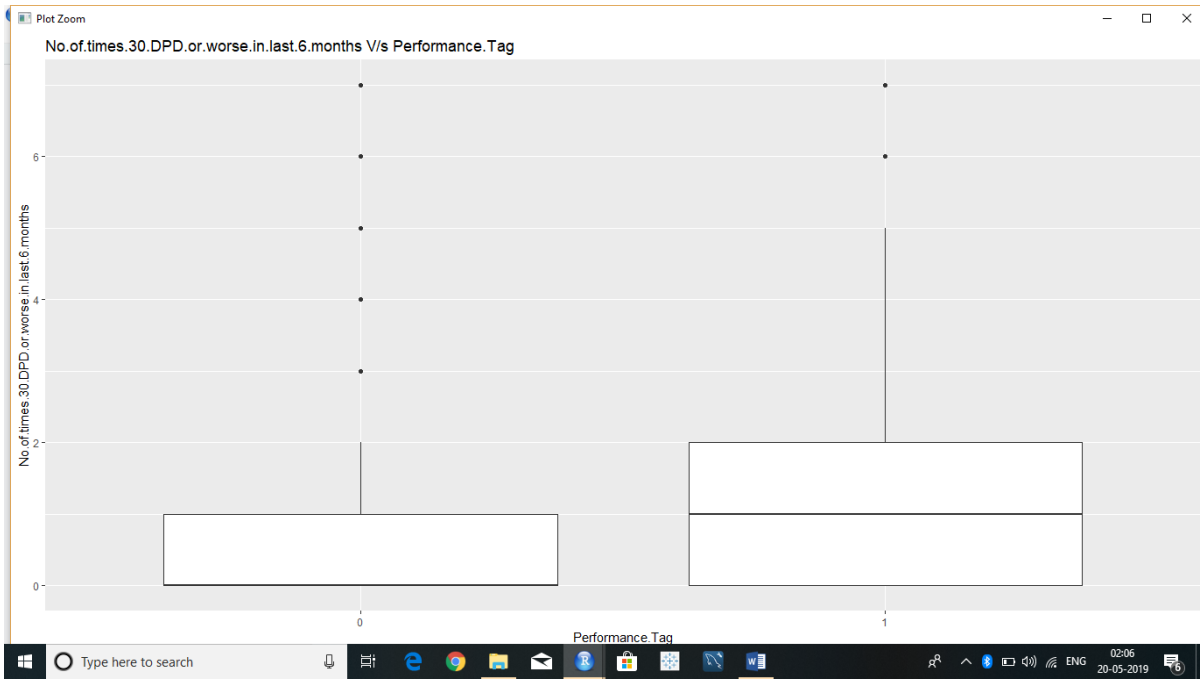


11. No.of.times.90.DPD.or.worse.in.last.6.months vs. Performance.Tag

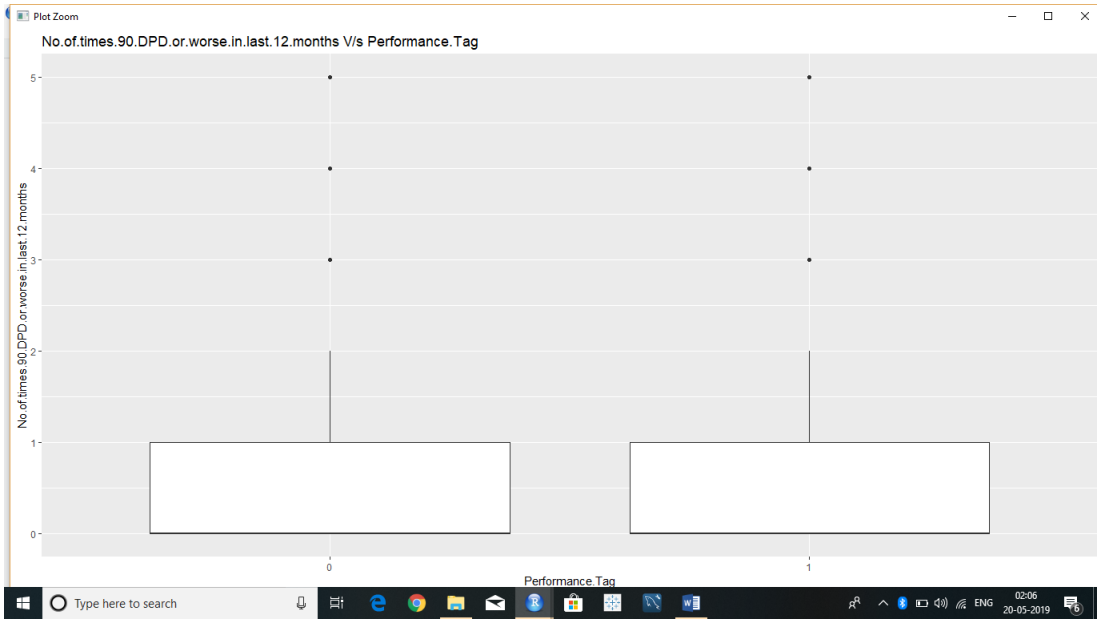
12. No.of.times.60.DPD.or.worse.in.last.6.months vs. Performance.Tag



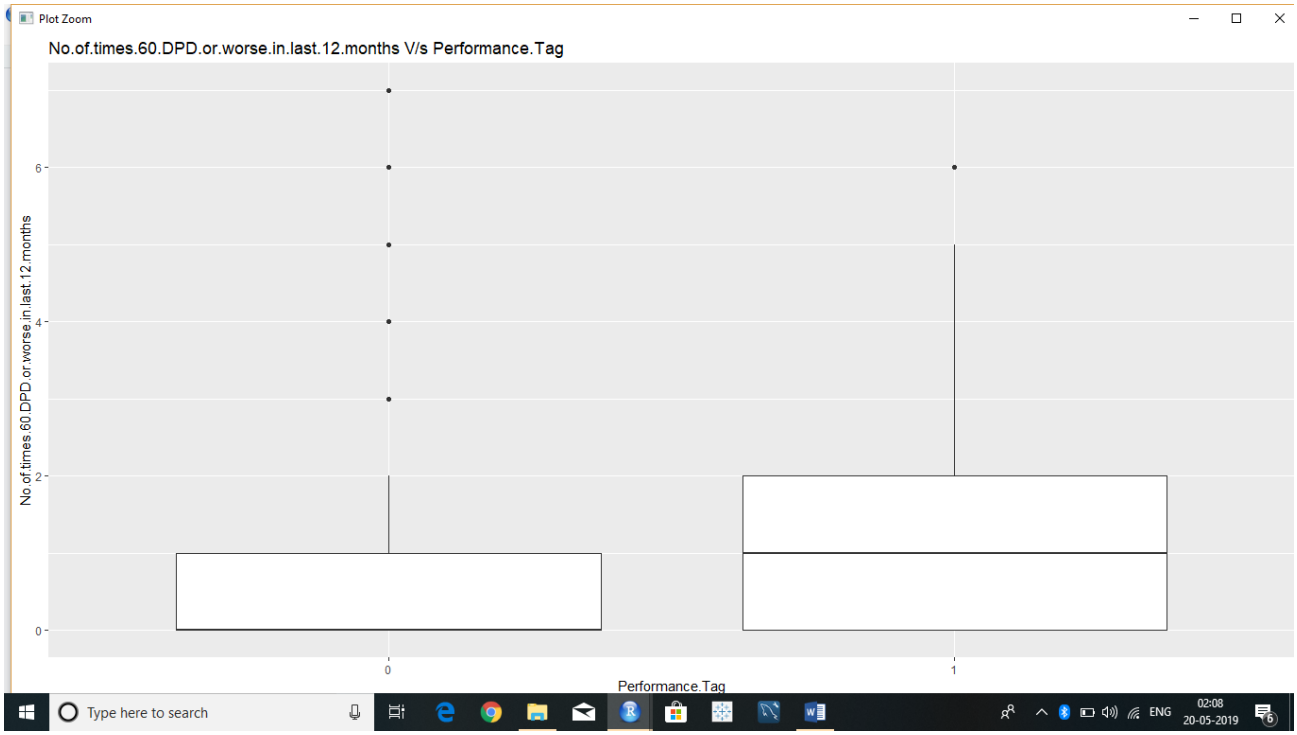
13. No.of.times.30.DPD.or.worse.in.last.6.months vs. Performance.Tag



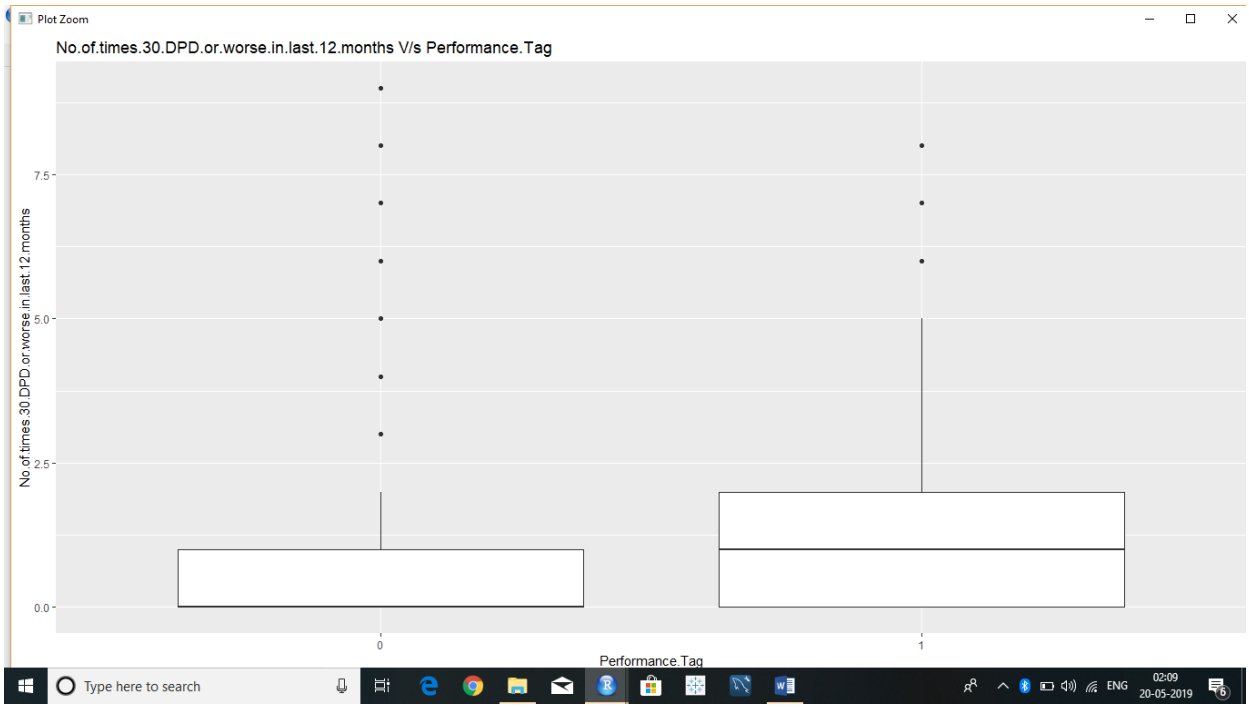
14. No.of.times.90.DPD.or.worse.in.last.12.months vs. Performance.Tag



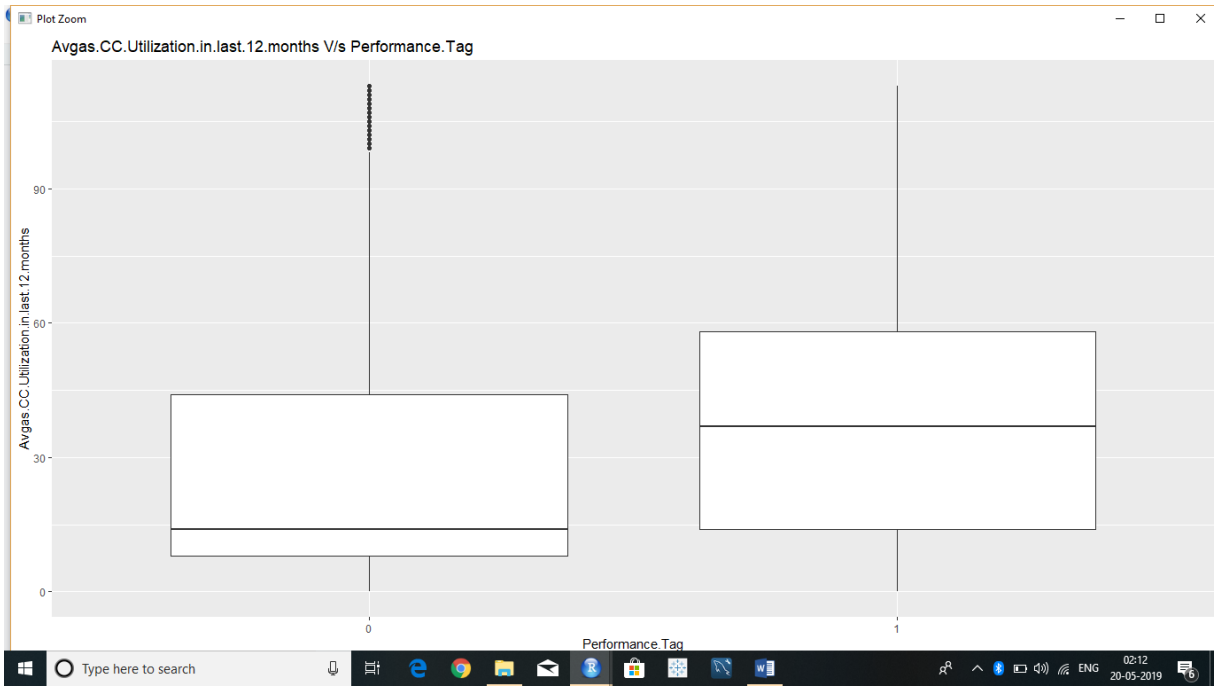
15. No.of.times.60.DPD.or.worse.in.last.12.months vs. Performance.Tag



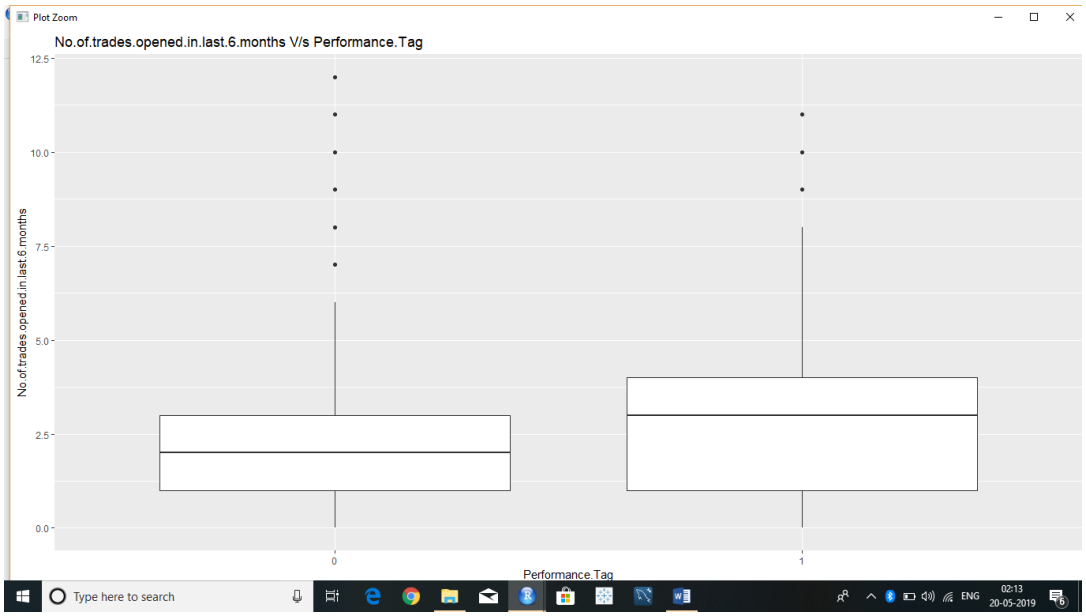
16. No.of.times.30.DPD.or.worse.in.last.12.months vs. Performance.Tag



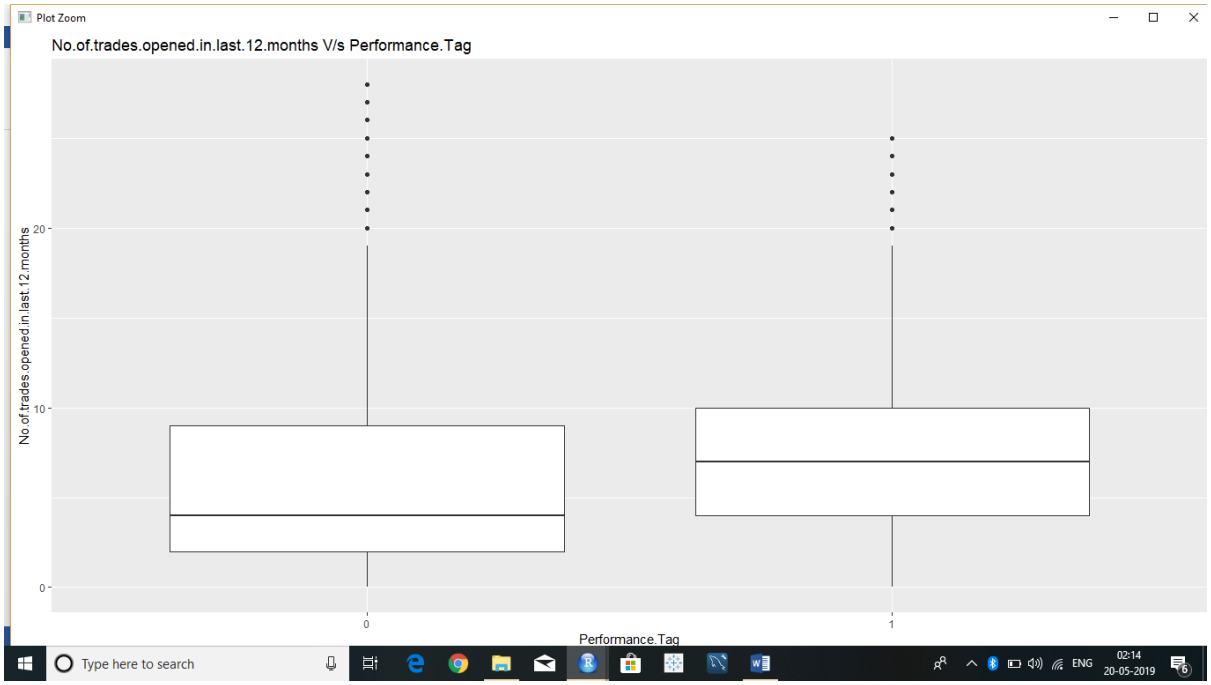
17. Avgas.CC.Utilization.in.last.12.months vs. Performance.Tag



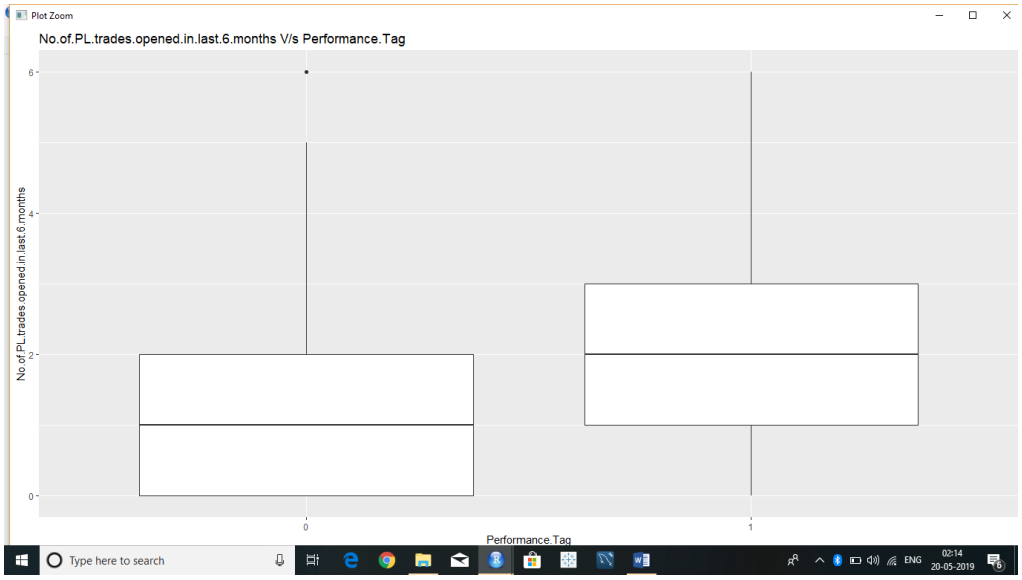
18. No.of.trades.opened.in.last.6.months vs. Performance.Tag



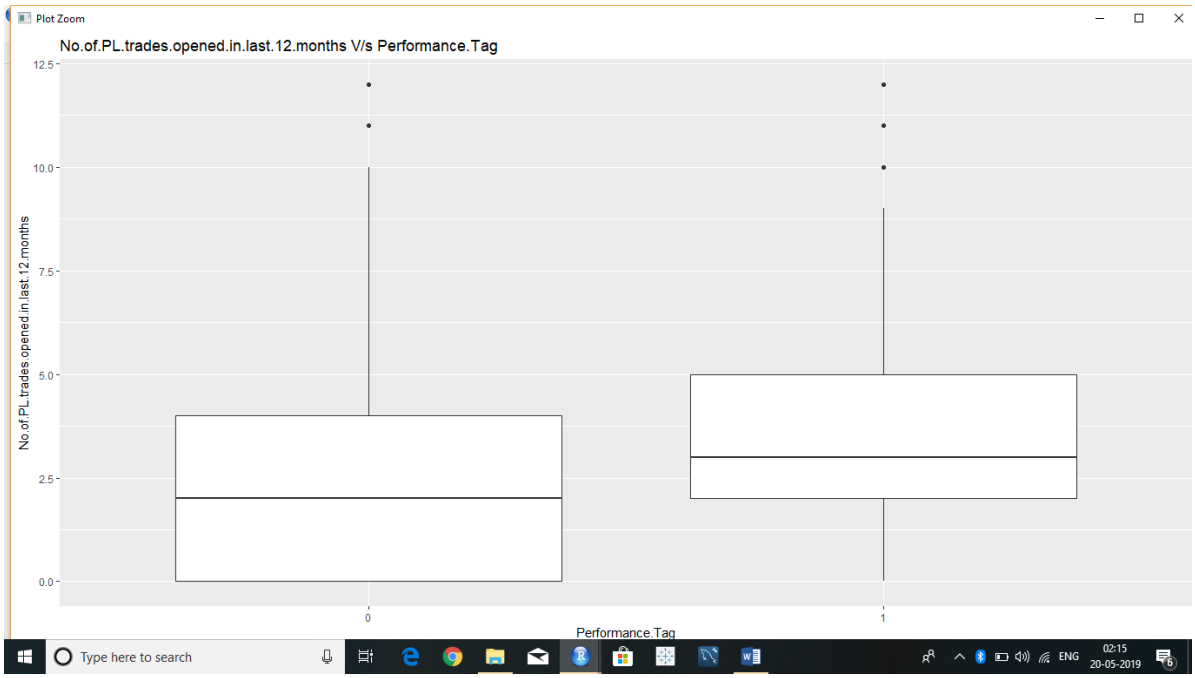
19. No.of.trades.opened.in.last.12.months vs. Performance.Tag



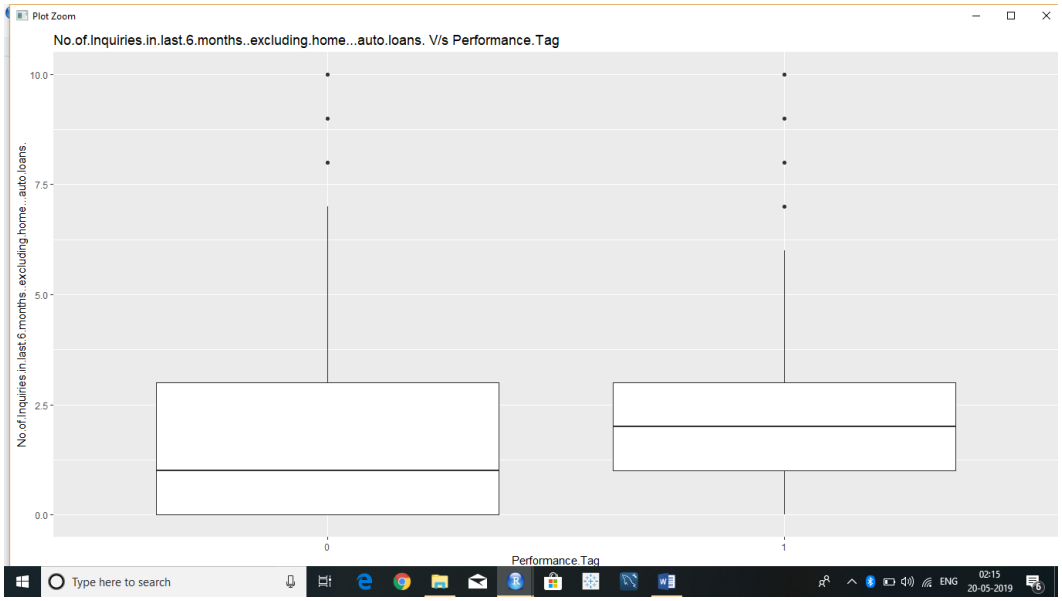
20. No.of.PL.trades.opened.in.last.6.months vs. Performance.Tag



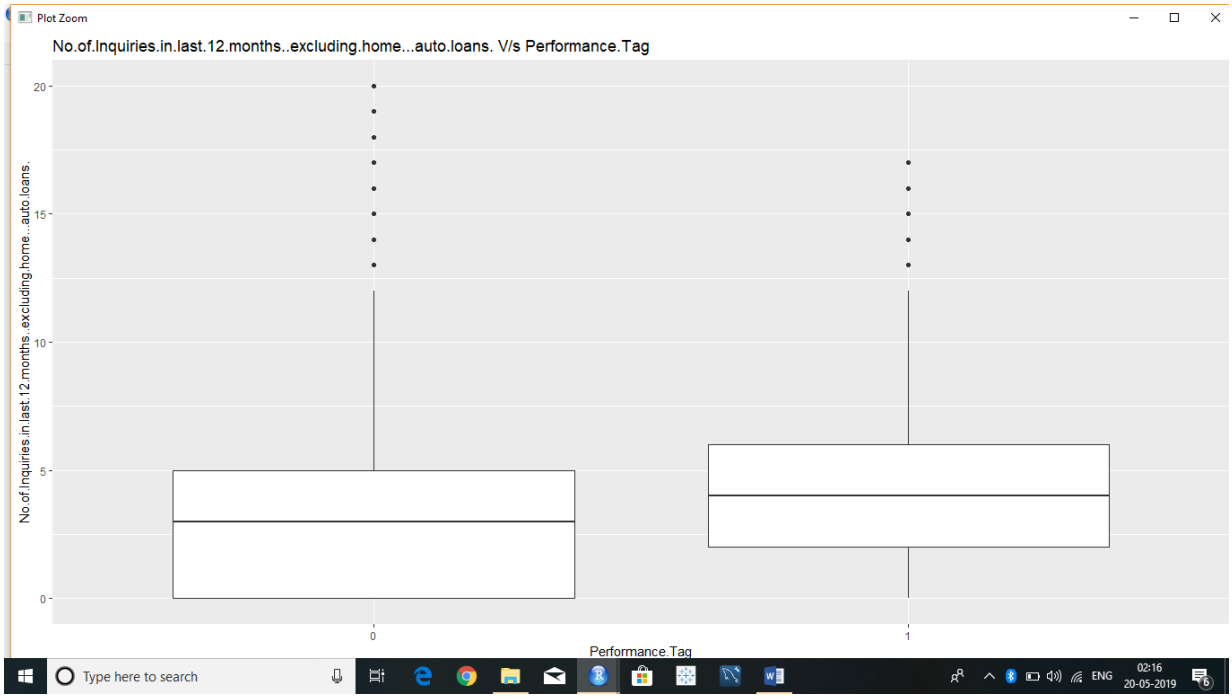
21. No.of.PL.trades.opened.in.last.12.months vs. Performance.Tag



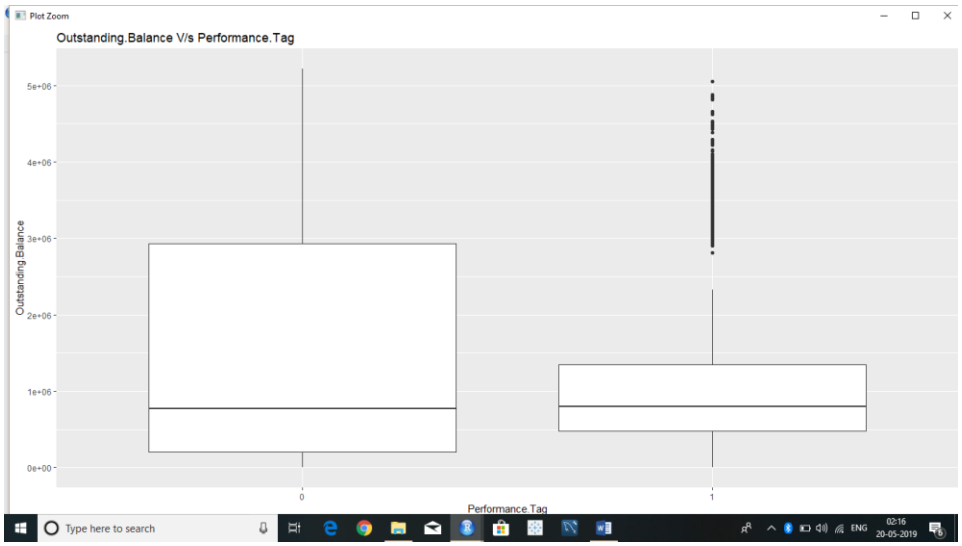
22. No.of.Inquiries.in.last.6.months..excluding.home...auto.loans. vs. Performance.Tag



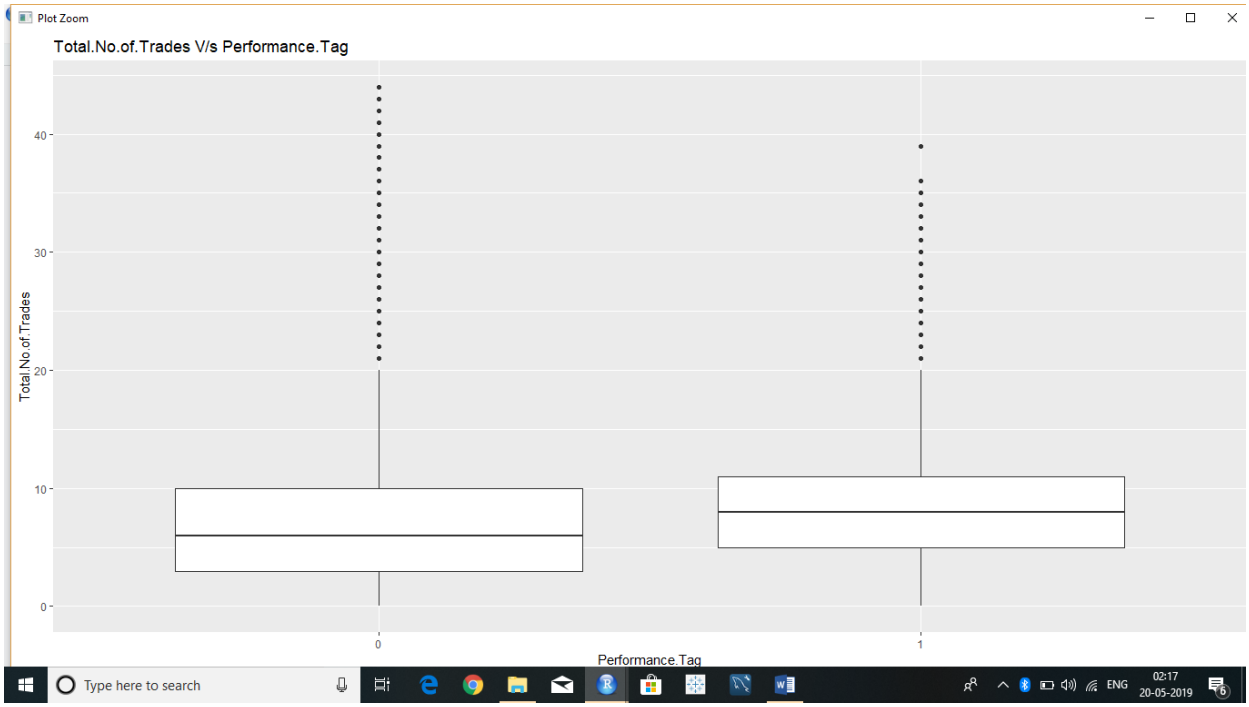
23. No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. vs. Performance.Tag



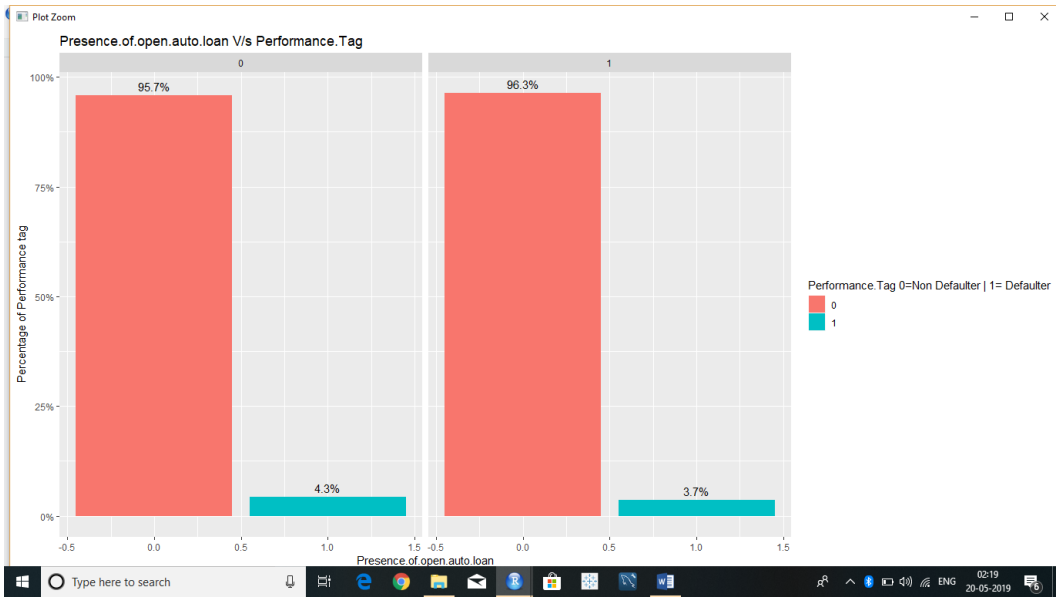
24. Outstanding.Balance vs. Performance.Tag



25. Total.No.of.Trades vs. Performance.Tag



26. Presence.of.open.auto.loan vs. Performance.Tag



EXPLORATORY DATA ANALYSIS

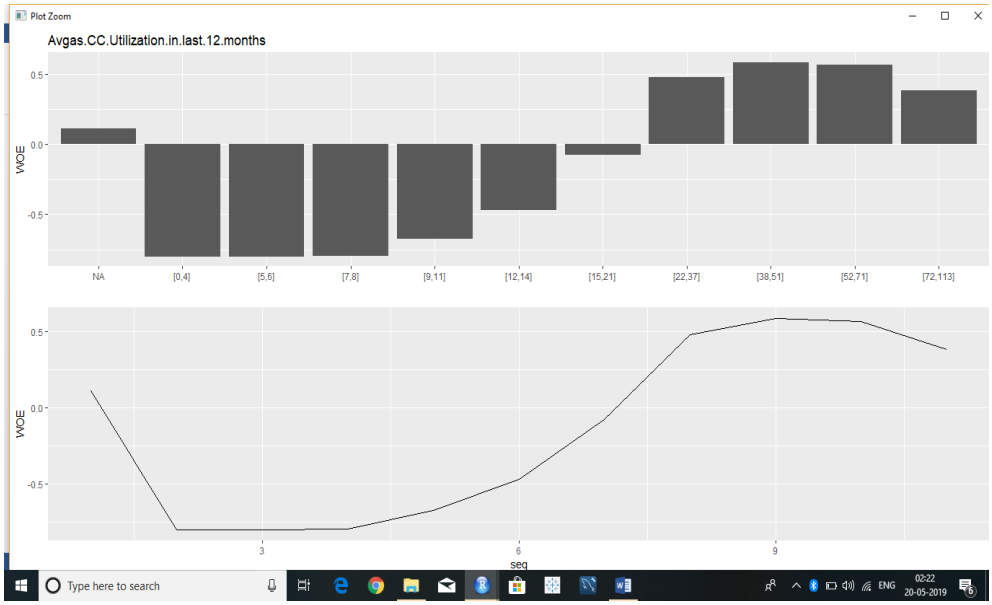
We will be using WOE and IV for EDA.

We see that "No of Inquiries in last 12 months excluding home auto loans", "Avgas CC Utilization in last 12 months", "No of PL trades opened in last 12 months", "No of trades opened in last 12 months" has IV more than 0.3 which indicates that these variable have

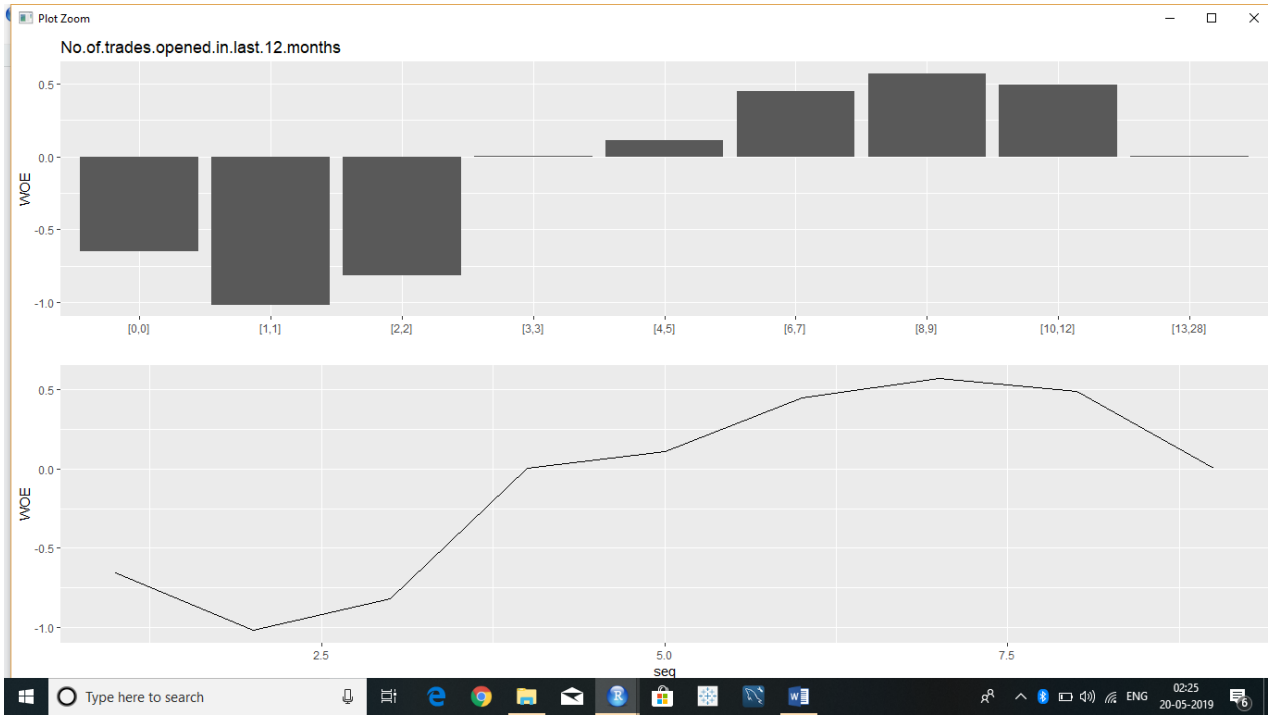
Strong predictive Power where as "Outstanding Balance" and "Total No of Trade" has Medium predictive Power.

Now let check the plot of these variables with respect to WOE.

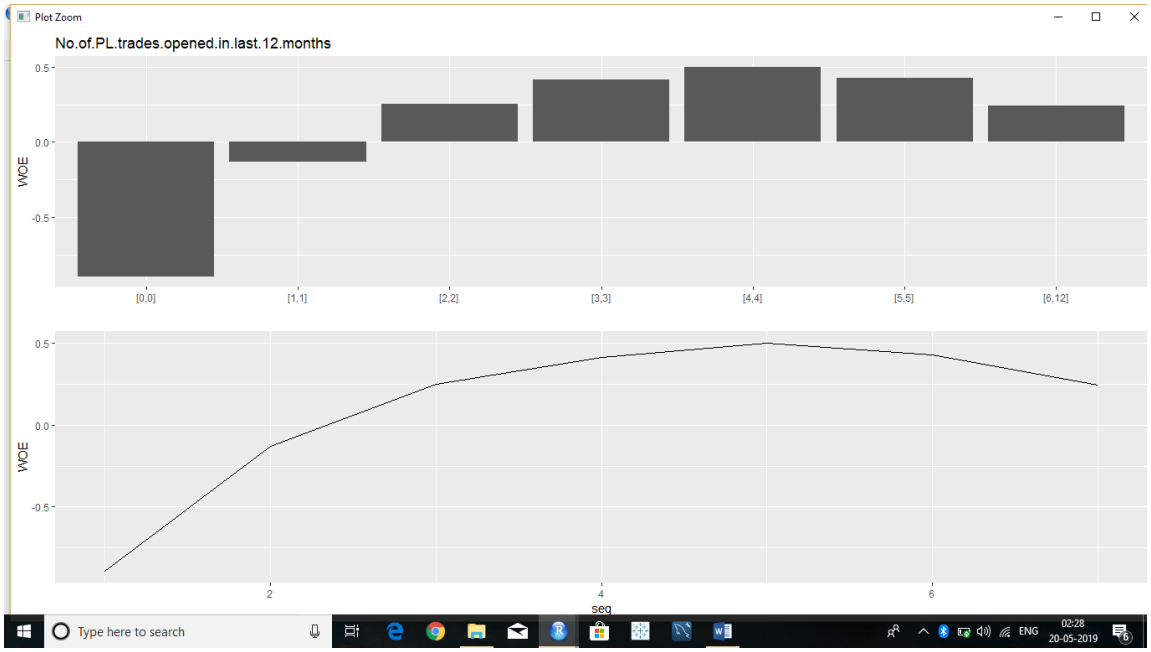
1. *Avgas.CC.Utilization.in.last.12.months: information value 0.3099*



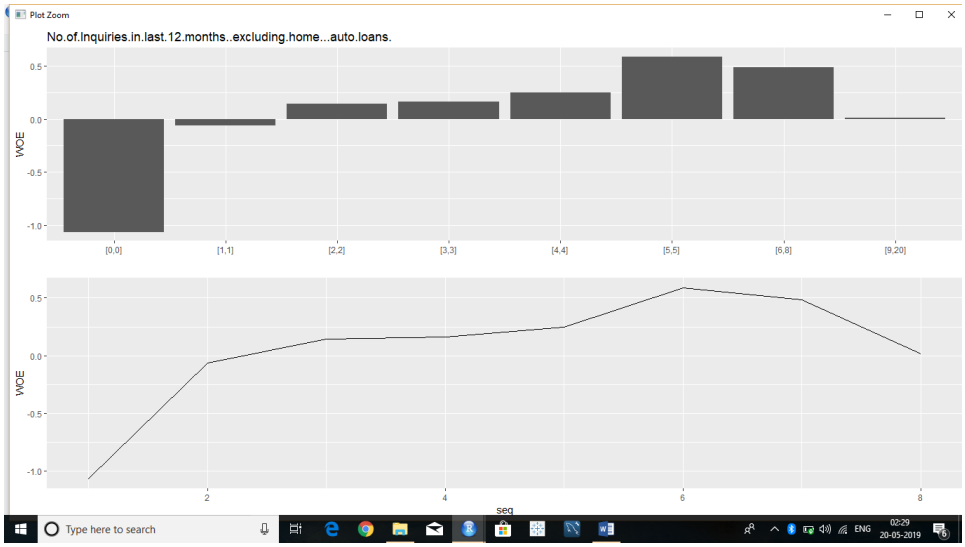
2. *No.of.trades.opened.in.last.12.months: information value 0.298 Approx. 0.3*



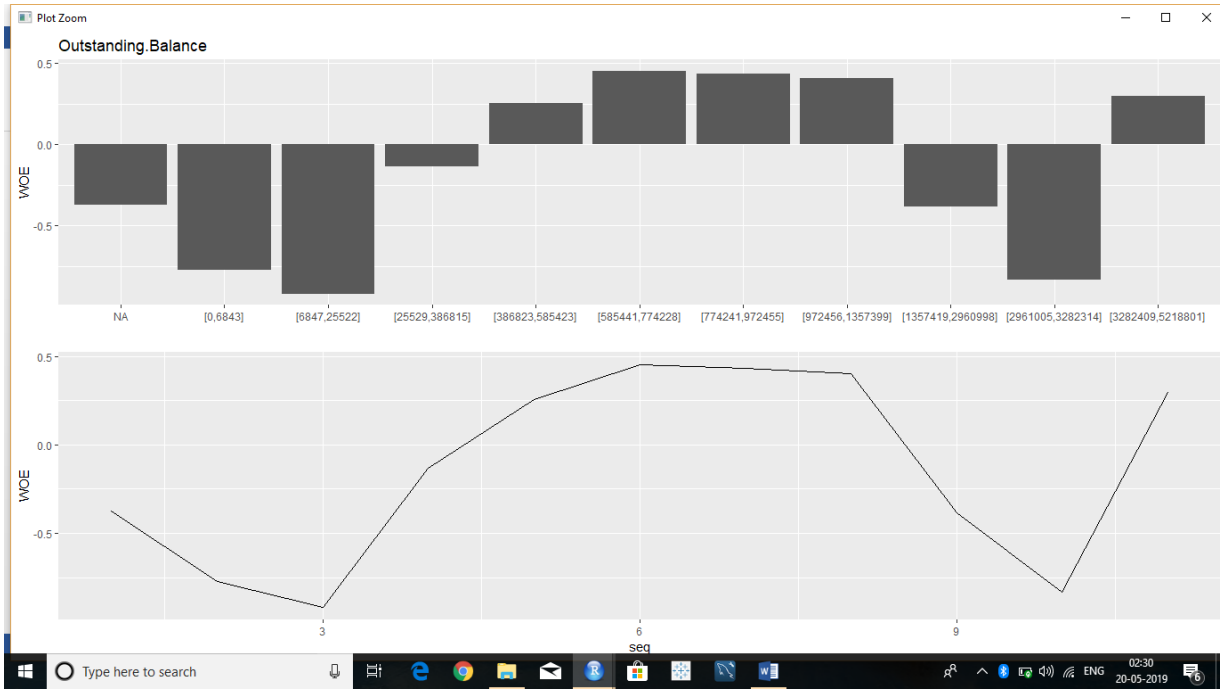
3. *No.of.PL.trades.opened.in.last.12.months: iv (0.296) approx. 0.3*



4. *No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.: iv (0.295) approx. 0.3*



5. *outstanding.balanace: iv (0.24)*



➢ Data Cleaning and Transformation

- **Outlier treatment** – We have use quartile and log transformation for outliers and skewed distribution.
- **Dummy Creation** – We have created k-1 dummy variable for all the categorical data using package AtConP form https://github.com/prk327/AtConP.git through function “df.matrix”.
- **Binning Variable** – We have created binning variable through “information package” and then impute the variable through “DF.Replace.Bin” of AtConP package.

- **WOE Variable** – We have created WOE values through "create_infotables" of "information package" and then use "DF.Replace.WOE" of "AtConP package" to impute the woe values into original data sets.

➢ Splitting the Dataset into train and test

- We have two data sets, one original raw data and another woe values data. We then have created a subset of demographic data from both raw and woe data sets.
- We have divided our four data into 70:30 ratio. The data sets is imbalance since 96% of the dependent variable are for good and only 4% repreent bad.
- We have used **ROSE** package for balancing our data sets. It helps to generate artificial data based on sampling methods and smoothed bootstrap approach.

➢ Model Building & Evaluation

First we will build Logistic regression only on demographic data and then same model is to be built on demographic + credit data. The accuracy can be low as there are less number of significant variables present in the data set. We will then use different variations of logistic regression like logistic regression-probit regression, cloglog regression. Then we will form decision tress and add weights to it to get better model.

- **Demographic data model –**

  We have use Logistic Regression to train our model and below are the metrics:

Significant variables in final model	Coefficients value (Numeric)
Age	0.0206
Noofmonthsincurrentresidence	2e-16
Noofmonthsincurrentcompany	9.78e-07

Final model metrics	Values (Numeric)
AIC value	16937
Null deviance	17100
Residual Deviance	16929

We have use StepAIC to select the model based on AIC value, then we have use VIF to check for any multicollinearity,

Then we have reduce the variable based on its P-Value, we have consider only those variables which have significant p-value. Below is our Model for demographic data

log(odds) = - 2e-16 + 0. 0206 (Age) + 2e-16 (Noofmonthsincurrentresidence) + 9.78e-07 (Noofmonthsincurrentcompany)

We have evaluated our model based on accuracy, Sensitivity, Specificity, C-statistic, KS-statistic and AUC:

Metrics	Values (Numeric)
KS-statistic	0.27
Overall Accuracy	75%
Sensitivity	38%
Specificity	73%
Area under the curve (AUC)	56%

➢ *Application Score Card*

We will use the above models mentioned to create an application score card, we will create a custom function to generate a score card.

Thank You