Problem statement

To mitigate credit risk to ‘acquire the right customers’.

To help CredX identify the right customers using predictive models. Using past data of the bank’s applicants

To determine the factors affecting credit risk

Create strategies to mitigate the acquisition risk and assess the financial benefit to CredX

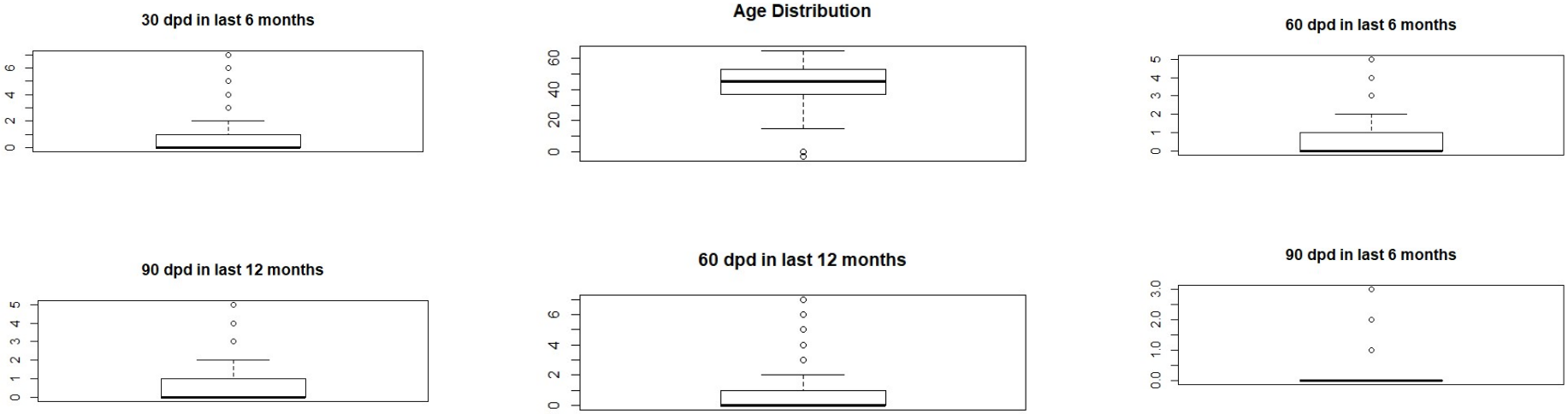
Data understanding

- Two data sets are present - demographic and credit bureau data.
- Demographic/application data: This is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
- Credit bureau: This is taken from the credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.

Data Preparation

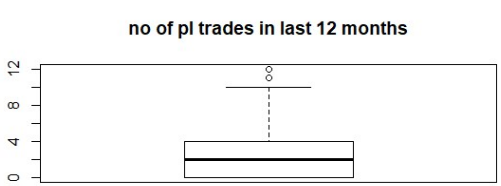
- Checking all the columns for missing values
- Checking all the columns for Nas
- Checking the necessary columns for duplicate values
- Outliers detection using the quartiles and boxplots
- Dummy variables creation for the necessary features

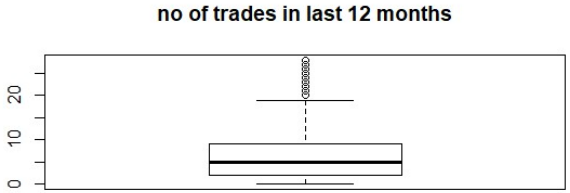
Outlier Detection Using Boxplots

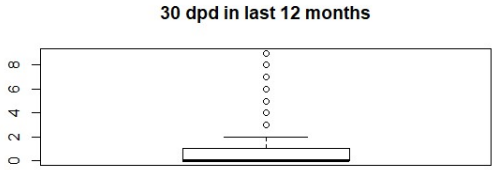


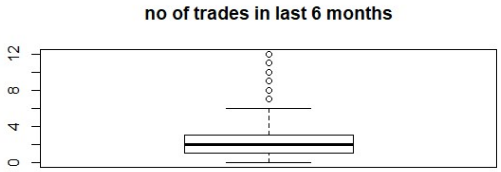
- There are few outliers in Age, having zero and negative values
- Other columns do not contain outliers, only the certain valid values lie outside the 3rd quartile range

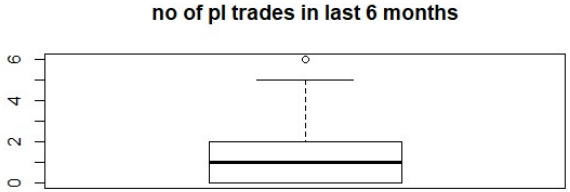
Outlier Detection Using Boxplots

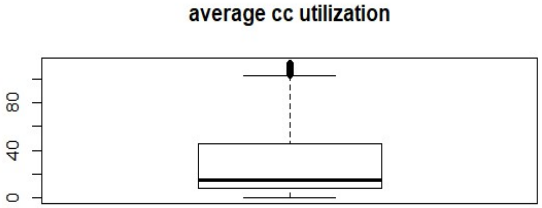






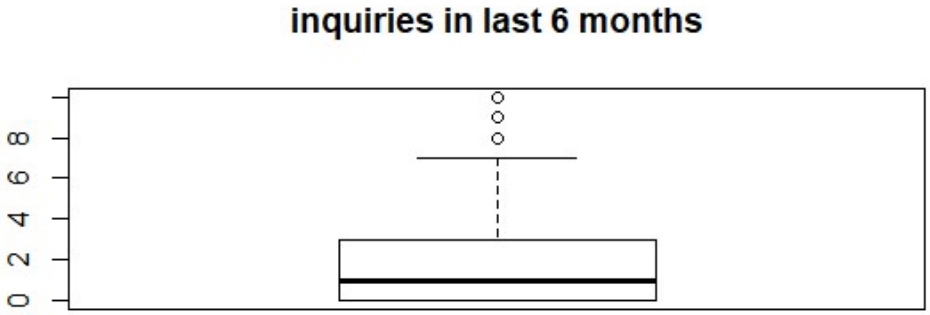


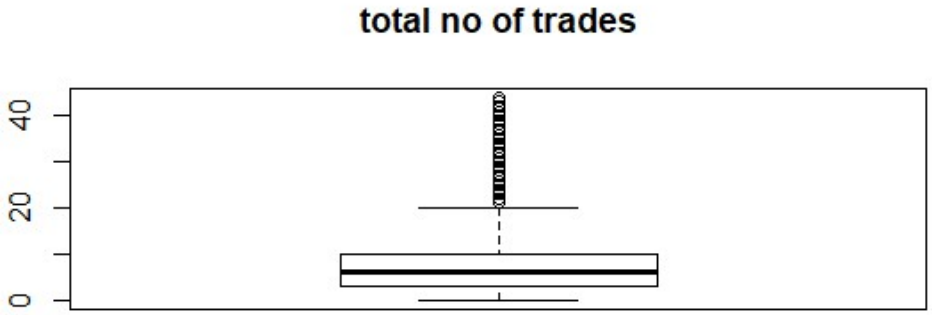


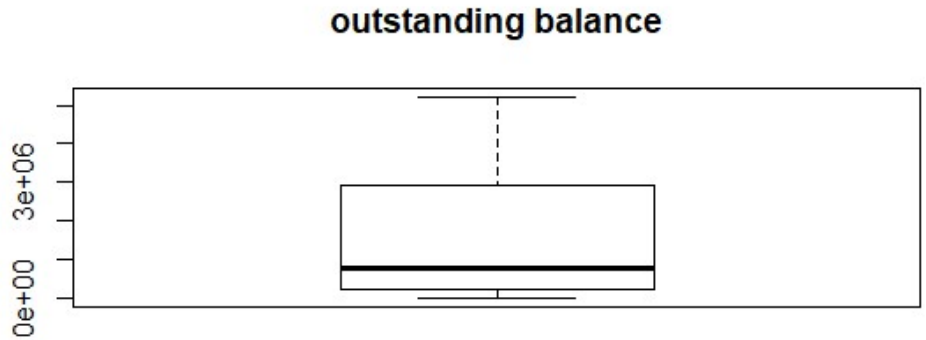


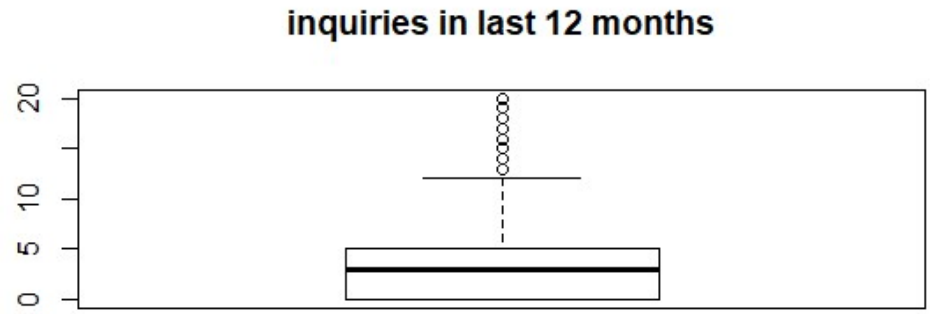
- Average Credit Card Utilization contains values more that 100, but those can account for over usage
- Other columns do not contain outliers, only the certain valid values lie outside the 3rd quartile range

Outlier Detection Using Boxplots







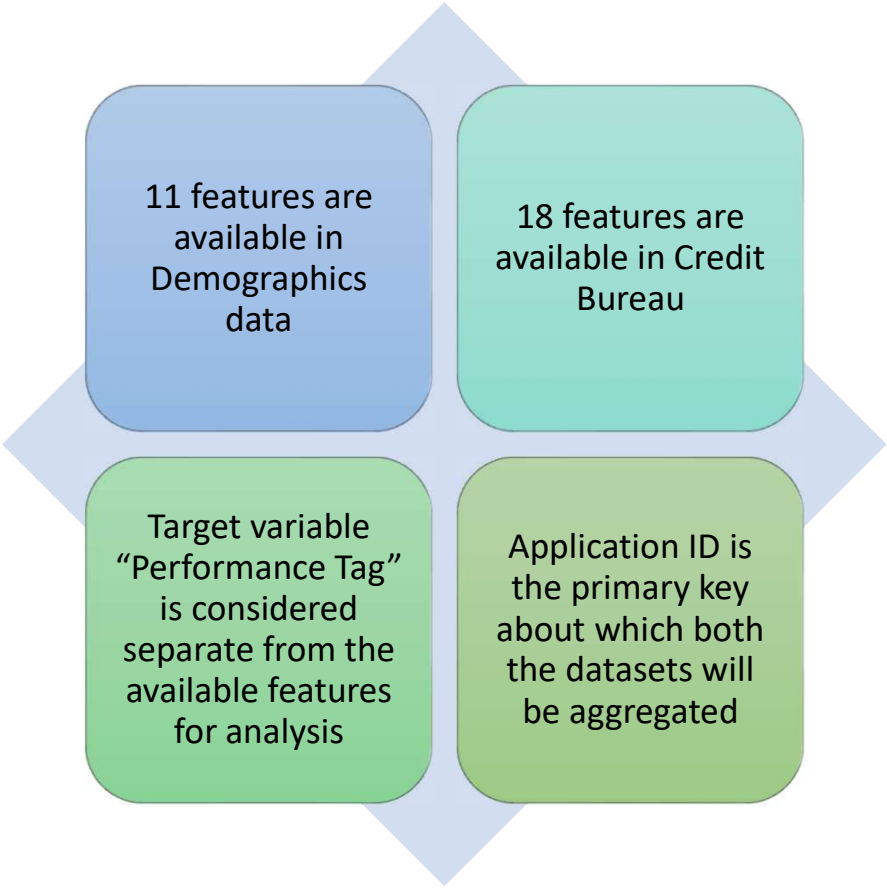


- Other columns do not contain outliers, only the certain valid values lie outside the 3rd quartile range

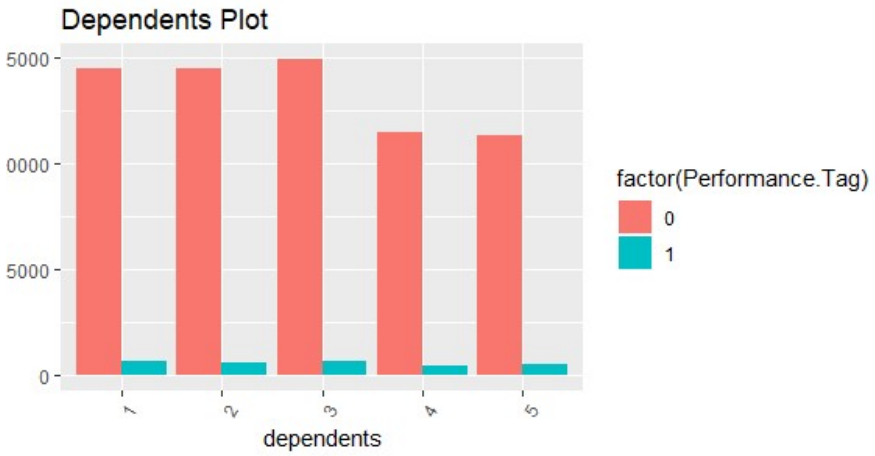
Data Cleaning

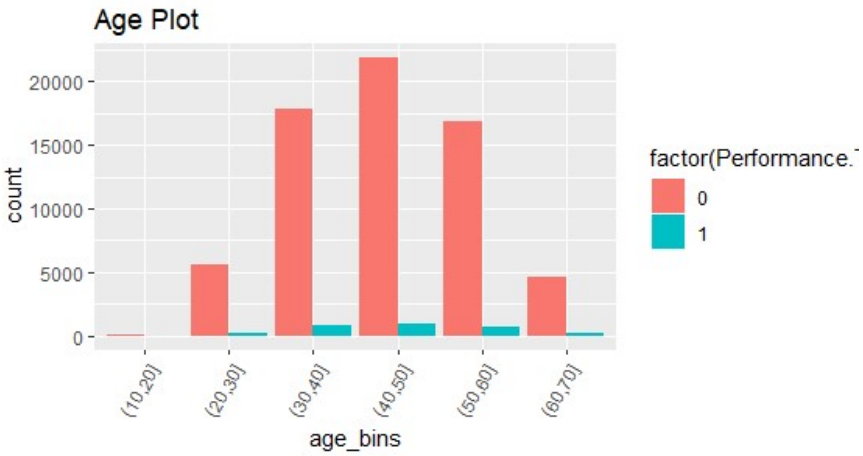
- Three duplicate application IDs were removed
- Negative and Zero values removed from Age, Income column
- Removed the rows where Gender, Marital Status, Profession was not mentioned
- Imputed the empty values in Education, Residence column with “others”
- Removed all the NAs from the average cc utilization columns, No.of.trades.opened.in.last.6.months, Presence.of.open.home.loan, Outstanding.Balance, Performance Tag column

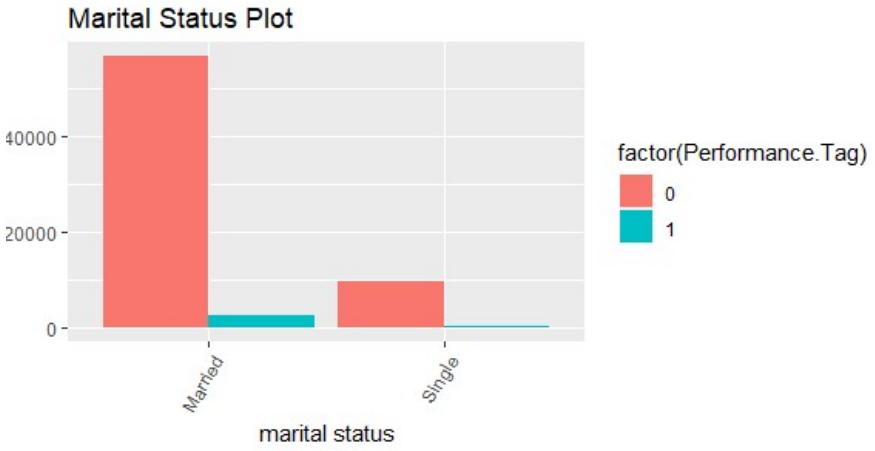
Data Aggregation

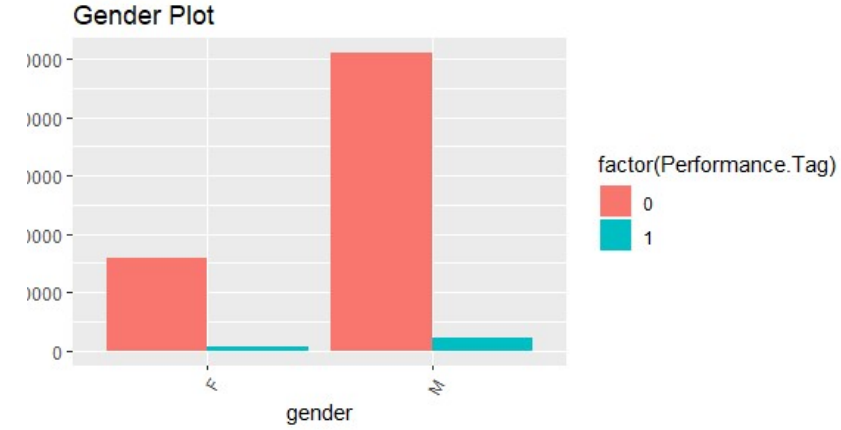


Exploratory Data Analytics

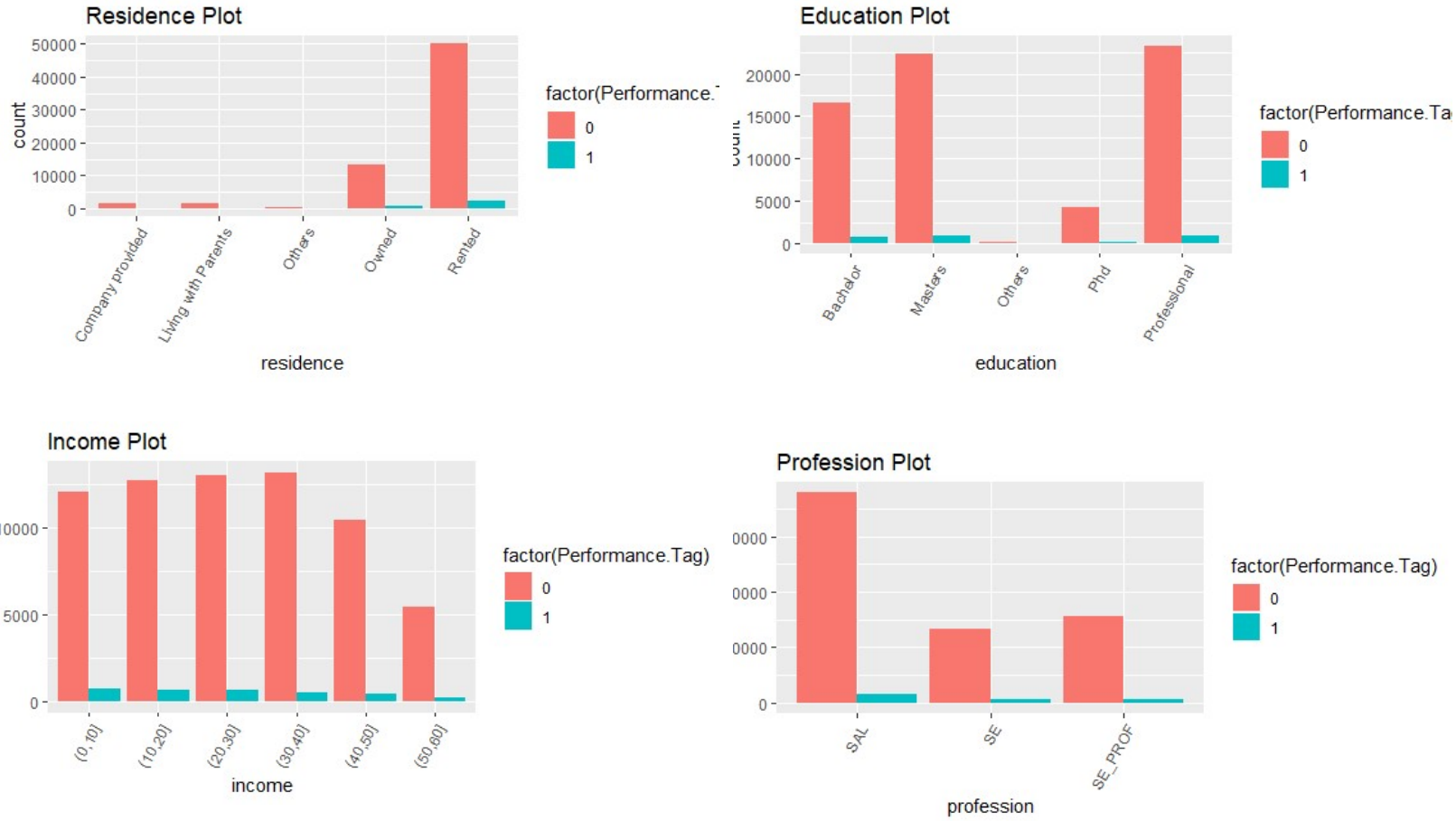




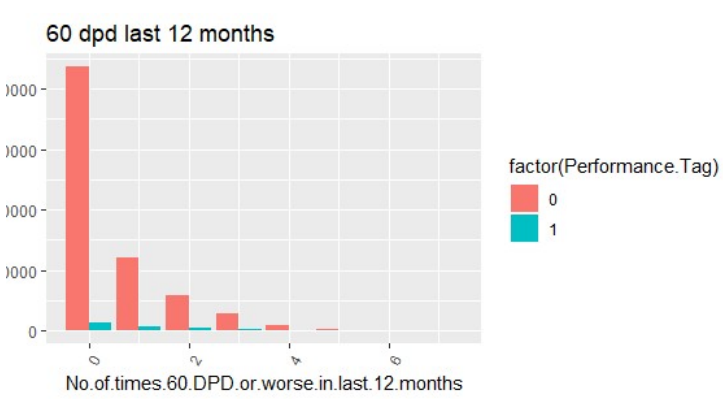


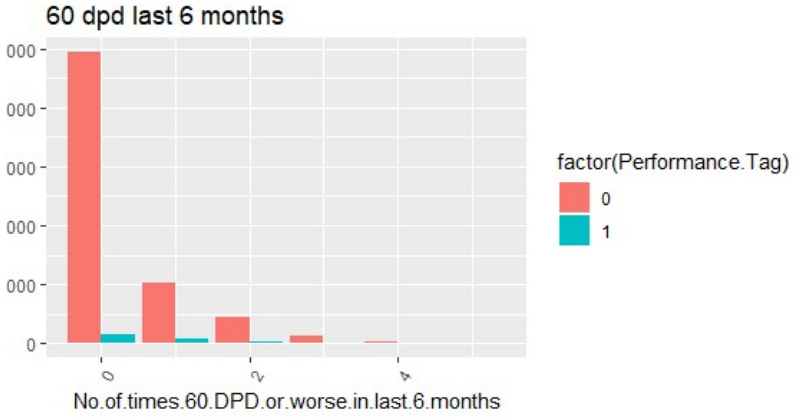


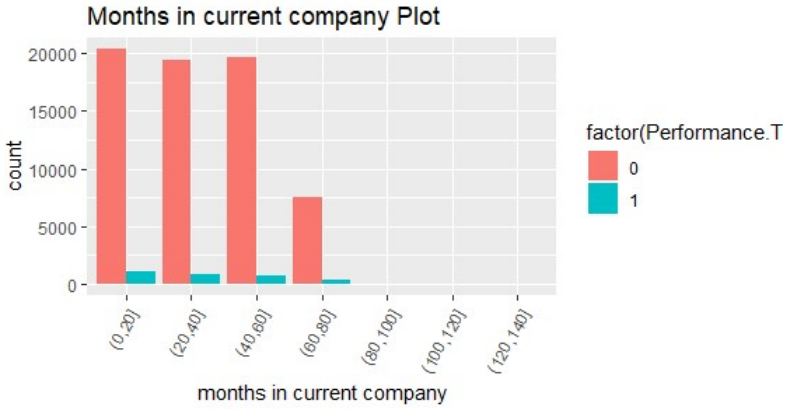
- No of dependents have equal distribution of defaulters, so may not be a significant contributor
- We can see maximum defaulters in the age group 30-60, so we nee to consider it
- Marital status and Gender also seem to affect the defaulter's behaviour

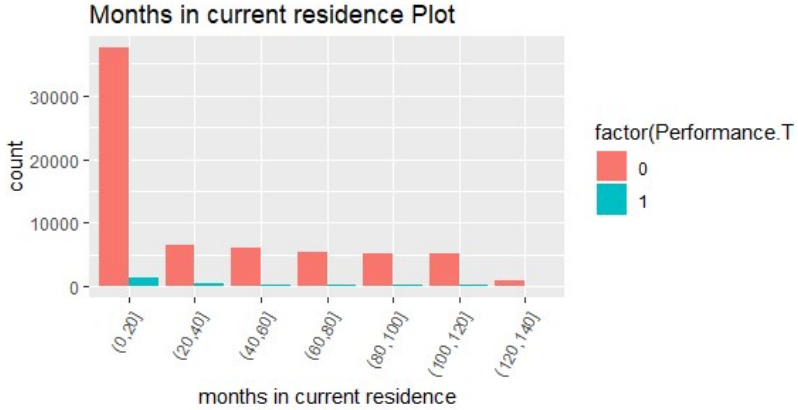


- Type of Residence, Education and Profession seems to be significant contributors
- Income seems to perform monotonically in all the groups

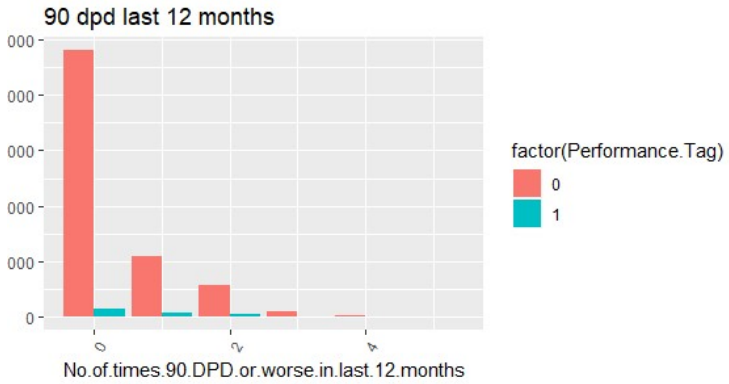


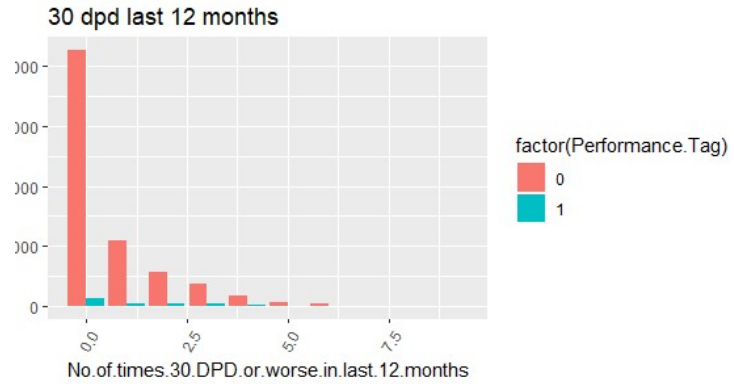


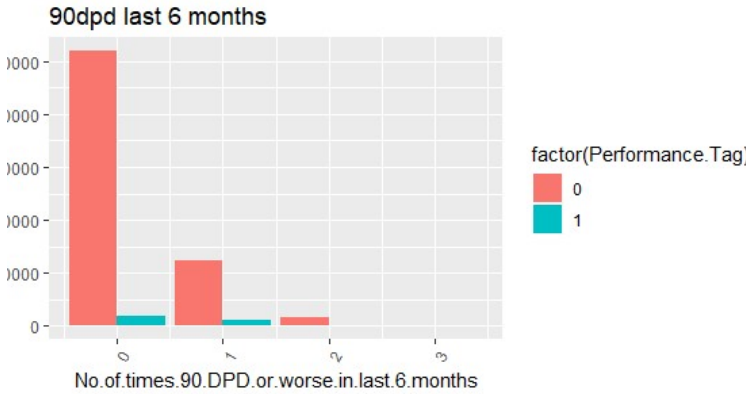


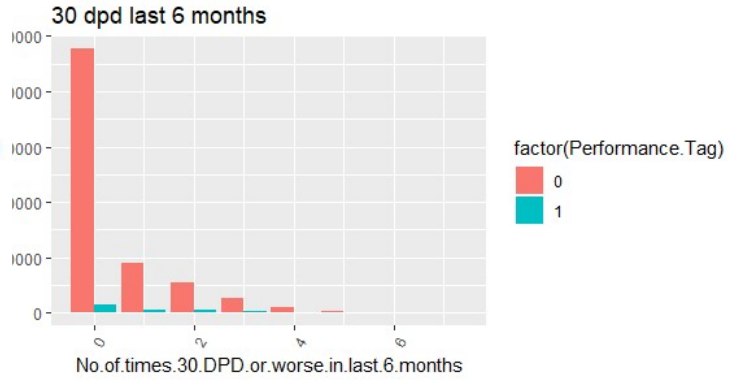


- All seems to be significant factors as there is present some trend in the defaulting behaviour
- 60 dpd in last 6 months and in last 12 months follow the same trend

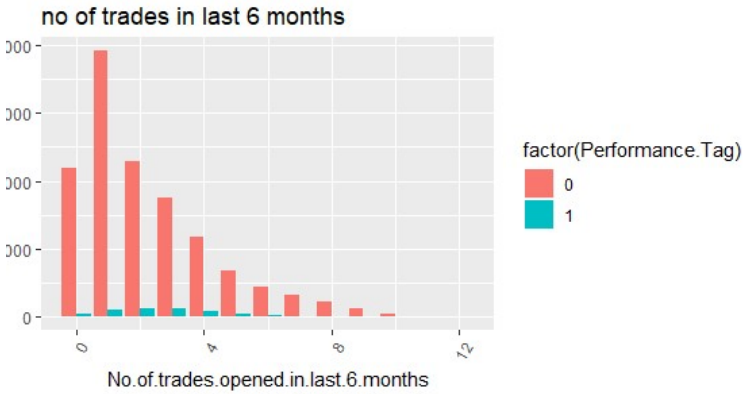


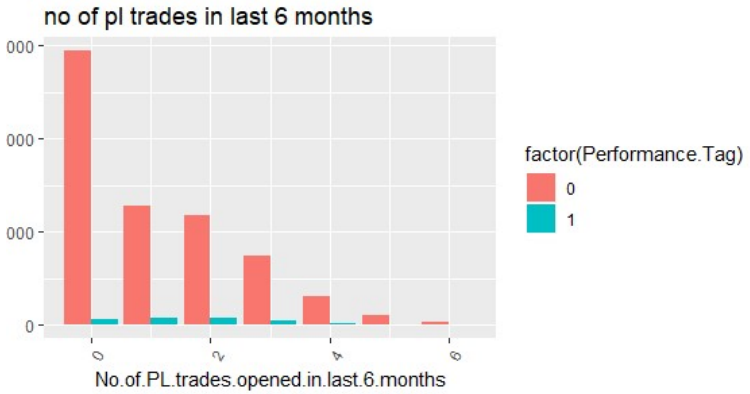


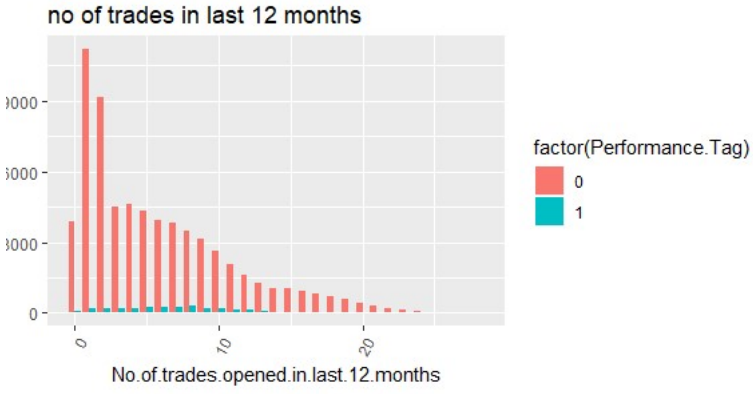


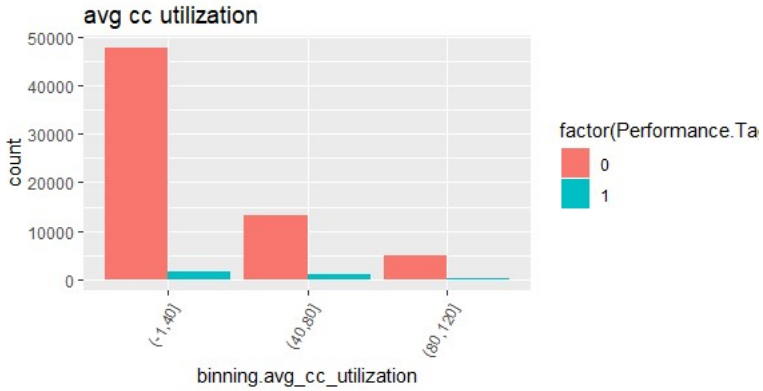


- All seem to be the significant contributing factors and all the four follow the same pattern

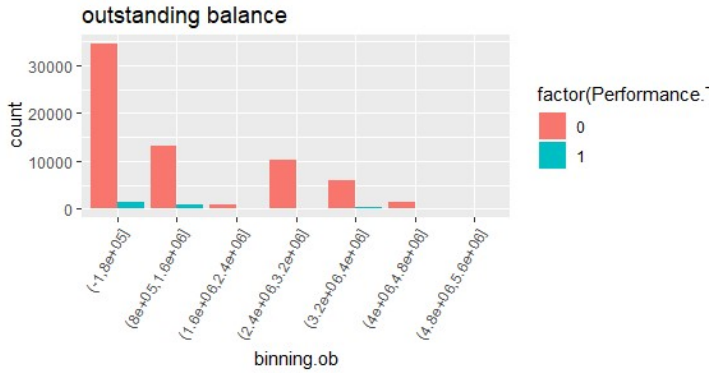


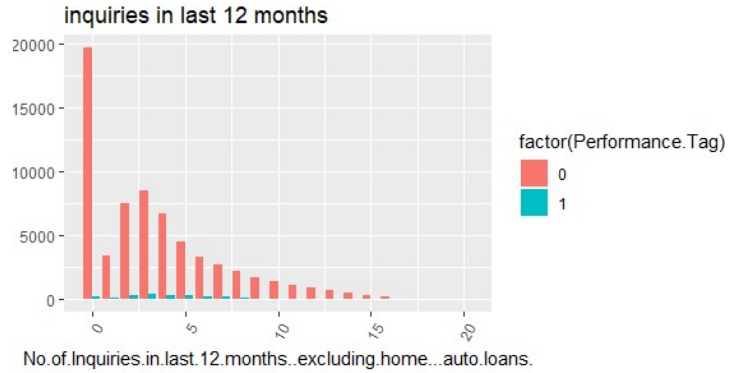


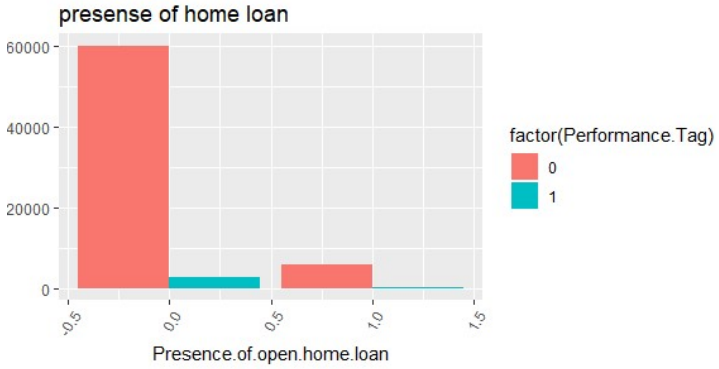


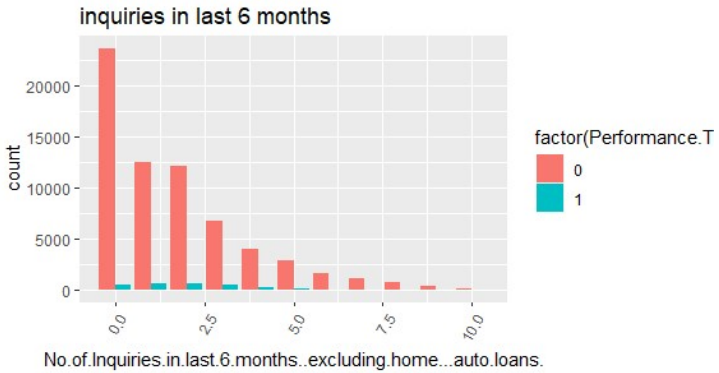


- All the factors seem to be significant in accounting for the default behaviour
- No of trades in last 6 months and in last 12 months seems to follow the same pattern

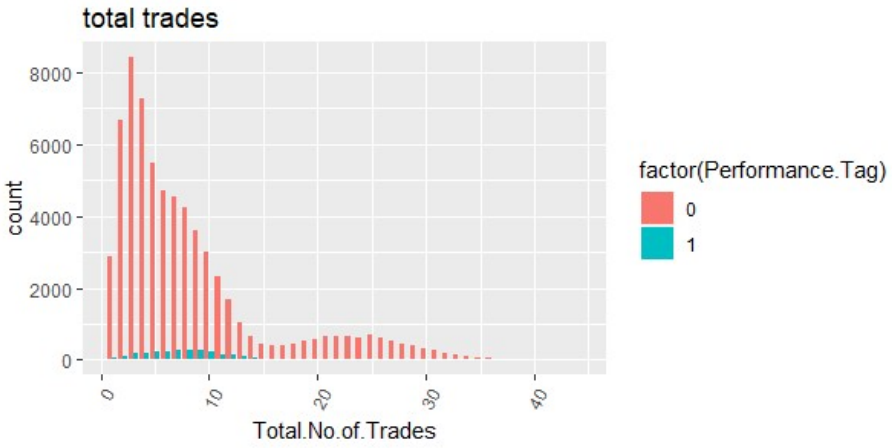


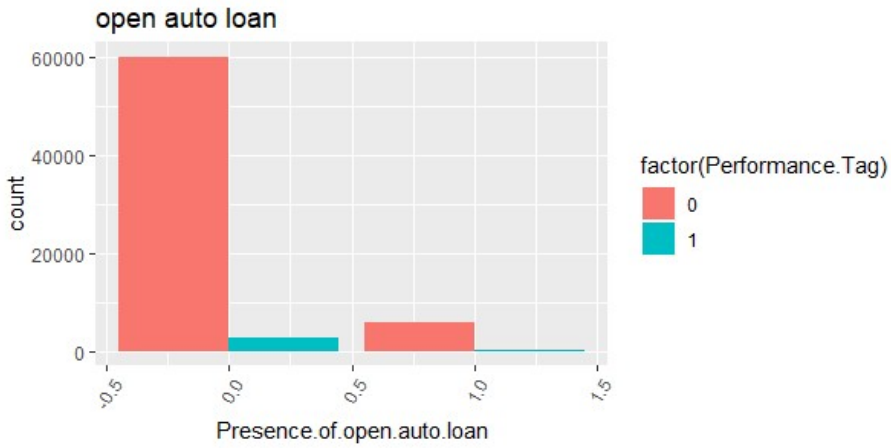






- All the factors seem to be significant in accounting for the default behaviour
- No of inquiries in last 6 months and in last 12 months seems to follow the same pattern





- Total trades seems to have more number of defaulters in the lower range
- Absence of auto loan seems to have a negative effect on the default behaviour

Conclusion Of EDA

- Following seems to be the contributing factors (after basic logistic model):
  - Marital.Status
  - No.of.times.30.DPD.or.worse.in.last.6.months
  - No.of.times.90.DPD.or.worse.in.last.12.months
  - Avgas.CC.Utilization.in.last.12.months
  - No.of.PL.trades.opened.in.last.12.months
  - No.of.Inquiries.in.last.6.months.excluding.home.auto.loans.
  - No.of.Inquiries.in.last.12.months.excluding.home.auto.loans.
  - Presence.of.open.home.loan
  - Total.No.of.Trades
  - No.of.dependents

Model Building

- Considering the classification problem of dividing the applicants in two categories based on the performance tag – Defaulters and Non Defaulters, we can use two different models.
  - Logistic Regression
  - Random Forest
- Not taking SVM into account as the amount of data is huge
- Segregating the data into test and train sets
- Will be using the drill down approach to remove the non significant variables on the basis of VIF and p-values.
- In random forest we need to vary the number of trees, min number of buckets and min number of leaves in a node.

Model Evaluation Techniques

- Plotting the sensitivity, specificity and accuracy at various cut-off values
- Choosing the best cut-off value where all the three parameters are very high
- Plotting the confusion matrix for the best cut-off value
- Using the KS-Statistics and Lift-Gain chart to check for better performing model out of the two
- Using the cross validation in case of logistics regression to fine tune the model

Creating the Application Score Card

We need to create the application scorecard with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.

We can use the above information to convert customer odds to customer scores

We can use the scorecard application package – “scorecard”