# APPROACH PAPER FOR BFS CAPSTONE PROJECT:


***PRESENTED BY:***

ALAY SHAH

RUTUJA MOWADE

PALLAVI KUMARI A

MEGHA GAWDE

## ❖ OVERALL APPROACH TO OUR CASE STUDY:

There are multiple ways or methods to approach any data analysis project. However, to avoid getting lost, we followed the robust process developed by data scientists to solve any analytics problem effectively in any industry–appropriately called the **Cross Industry Standard Process for Data Mining (CRISP–DM) framework.**

It involved a series of steps, as follows (not necessarily sequential in nature):

1. Business understanding, Data understanding
2. Data Cleaning/Preparation
3. Exploratory Data Analysis
4. Data Modelling
5. Model Evaluation
6. Model Deployment

Out of the above six steps, we are expected to complete the first three steps and present our work in the form of insights about data. And then we will be mentioning the approach for the remaining three steps i.e. Approach for Model Building, Evaluation and Score Card Preparation.

## ❖ 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. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

- ✓ 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 our project.

## ❖ DATA UNDERSTANDING

- ✓ There are two data sets in this project — **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.

    - The dimension of this dataset is: 71292 obs. of 12 variables.

- ✓ **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.

    - The dimension of this dataset is: 71292 obs. of 29 variables

- ✓ Both files contain a **performance tag** which represents whether the applicant has gone 90 days past due or worse in the past 12-months (i.e. defaulted) after getting a credit card.

## ❖ DATA CLEANING

- ✓ Both the dataset had 3 duplicated entries, which we dropped.

- ✓ We found the Application ID's or both data sets to be identical, hence, we merged both the datasets specifically for data cleaning operation, taking Application ID as our primary key column.

- ✓ Next step was to fix the NA's. Also, please note that we imported our data sets with a command to read blank values as NA's as well, so that all empty/blank/null/NA's can treated at once.

- ✓ So we checked for NA's, which amounted to 3180. However, out of 3180 NA's, when we looked at the breakdown of individual variables for NA's, we found out that the dependent variable i.e. Performace Tag column itself contributed to about half the number of total NA's i.e. 1425 NA's in Performance Tag variable.

- ✓ However, there is a method of interpreting NA's in Performance Tag column, which is termed as **"Reject Inference".**

- ✓ Usually, the outcome value (in our case – 'Performance Tag' is the outcome value) is only available for clients who were actually granted for credit or leases. This creates a selection bias, where the data set is not representative of the thought-the-door application population. Hence, Reject Inference attempts to mitigate this bias by estimating how rejected applicant would have performed if they had been accepted. So we will be applying the same concept as follows:

- ✓ In this case, NA's in the Performance Tag can be interpreted as customers whose Performance is not recorded because they did not have any i.e. They were just not granted credit card at the first place or their credit applications were rejected. And that is why their data of Performance is "Not Available". Now, getting a credit card application rejected itself signals that the loan granting organization thought this customers would default in the near future, maybe because they might be having some clear red signals in their profile. So, in a way, these customers having NA in their performance Tag column can be assumed no better than the customers who defaulted i.e. having values "1" in the Performance Tag Column (Since 1 is the value of Default in our column). Thus instead of dropping the data of these customers, it can be helpful if we take them into account by replacing their NA values with the Default value i.e. 1. By doing that, our model will now be prepared for such bad profiles as well, whose application were rejected because of some clear bad signs in their profiles.

- ✓ After fixing the NA values of Performance Tag Columns, there were only '1208' other rows which had one or more NA's. And, the total rows of our dataset =

71292 i.e. the original dimension of the dataset. So technically, the number of rows having one or more NA's accounted to just little above 1.5% of the total dataset. (1208 / 71292 = 1.68 %)

- ✓ Now, the thumb rule of data analysis is, if the values to be fixed are not significant enough in amount, then just drop it. And 1.69% is not a significant number. Hence, following that thumb rule, we decided to drop all the remaining rows having one or more NA's, instead of taking time out to interpret it through some other method.

- ✓ THE NEXT STEP WAS TO CHECK FOR **INVALID VALUES** IN THE DATASET, AS FOLLOWS:

- ✓ We found out around 80 negative values, with following individual breakdown:

    - Age – 1, Value = -3
    - Income – 79, Value = -0.5

- ✓ Since Age cannot be "-3" and Income cannot be "-0.5", Hence clearly that was a data quality issue. Also, again, the data was not significant enough to treat it by taking some mean or median. So we just dropped that bad entries.

- ✓ Now, looking from the business perspective, we know that loan cannot be granted to someone whose age is less than 18. However, there were 55 such entries with following unique values: 0,15,16,17. So all entries of customer, whose Age is less than 18 are considered as malicious entries, and we had to drop that.

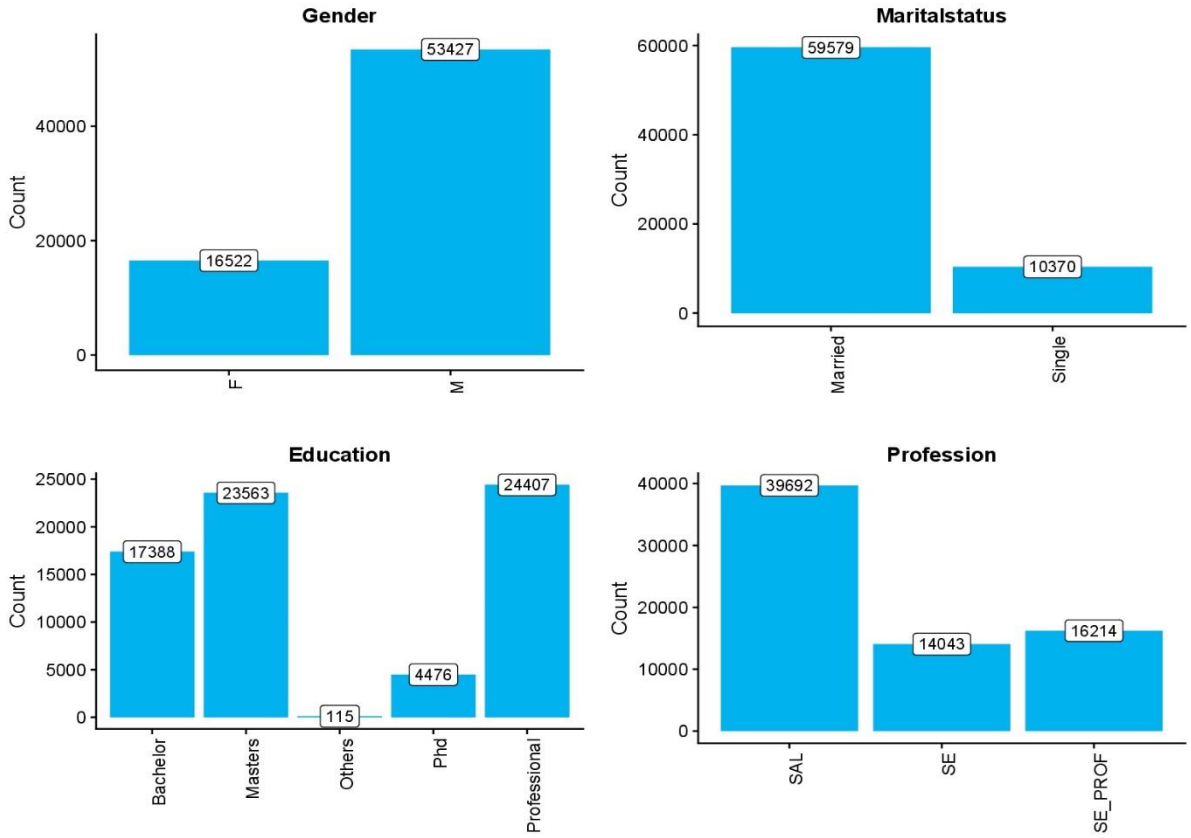
- ✓ Furthermore, after looking more closely at the data, we found out that the max value of the variable "No_of_times_30_DPD_or_worse_in_last_6_months" is 7, whereas it cannot exceed 6 i.e. [30,60,90,120,150,180]. Hence, that is clearly an error. However, instead of dropping that, this error can be tolerated by tweaking it a little i.e. by capping that variable to max possible value -- 6.

- ✓ And we are done here with basic data quality checks. Moving on to EDA now!

- ✓ However, for EDA, we again split the merged and cleaned data set into demographic and credit bureau data, to carry out individual exploratory data analysis of both the datasets.

# ❖ EXPLORATORY DATA ANALYSIS - DEMOGRAPHIC

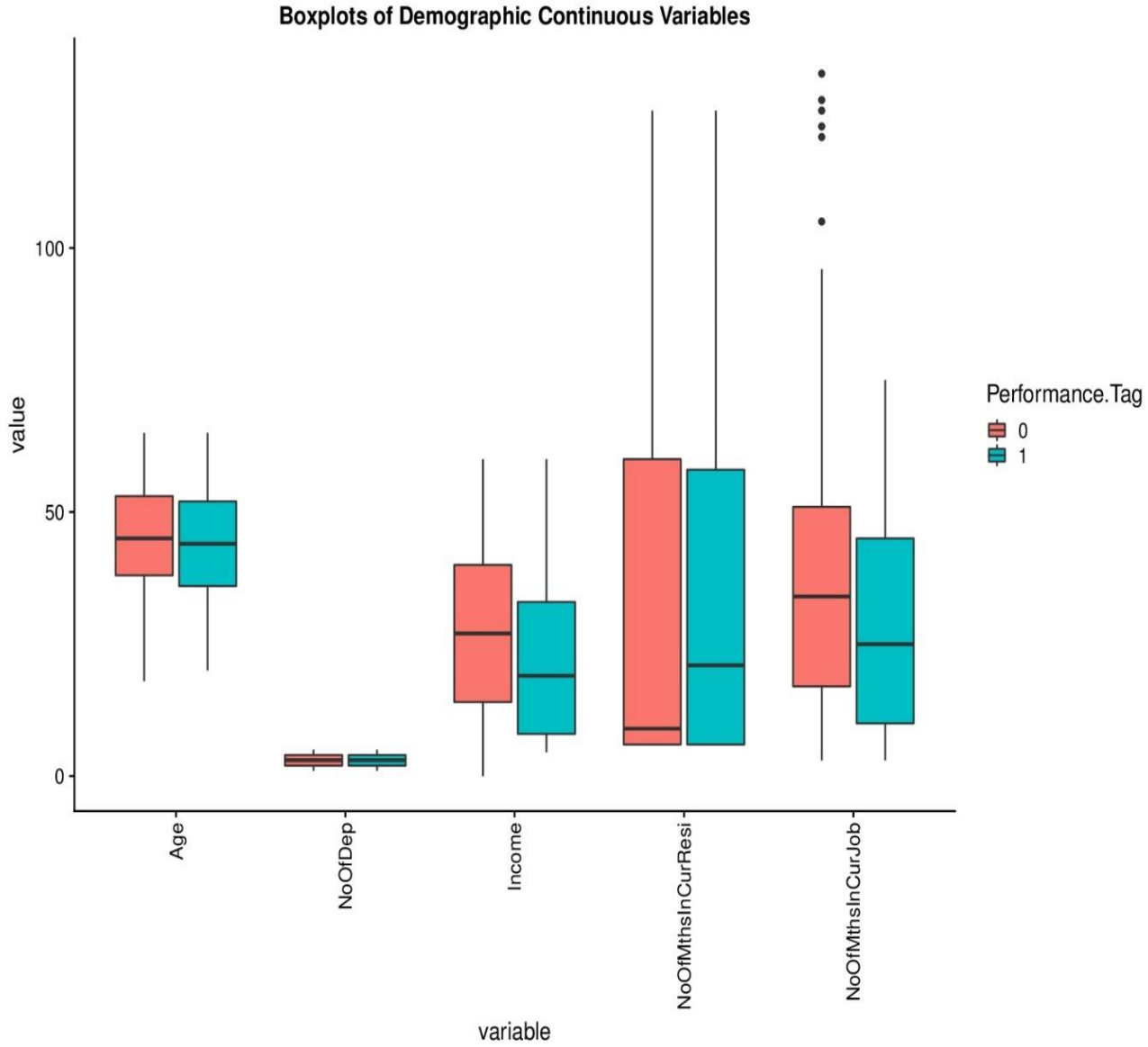
## ✓ *OVERALL DISTRIBUTION OF DATA FOR CATEGORICAL/DISCRETE VARIABLES:*




## ✓ *OBSERVATIONS:*

- For Gender, number of loan applications for Males is far larger than Females.
- Married People outnumber the Single People by a large factor in the dataset.
- The Others section in the Education sector had the lowest number of total credit card applicants amongst others in the same sector.
- Number of Salaried Applicants were the highest in the Profession Sector.

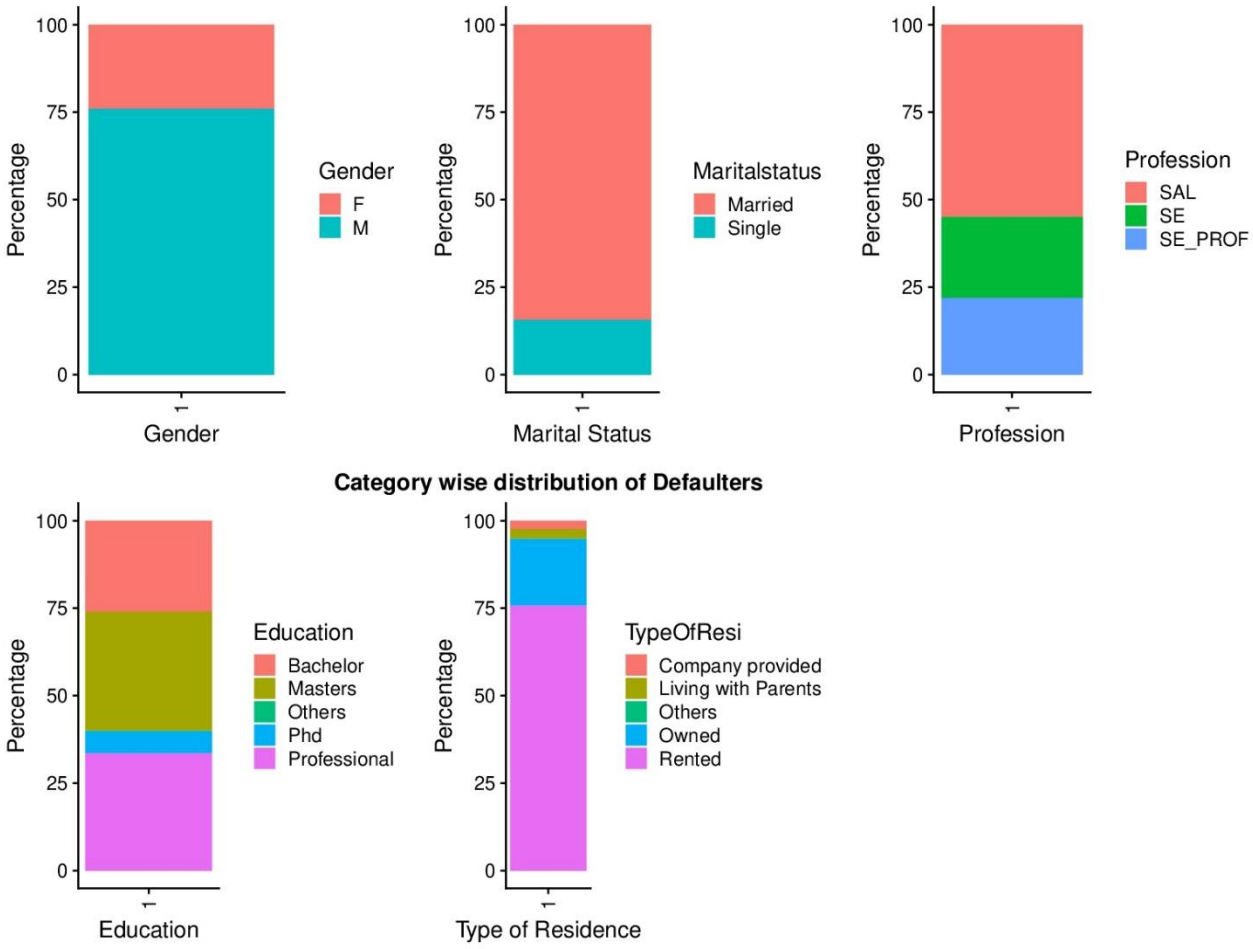
## ✓ *BOX PLOTS OF CONTINUOUS VARIABLES:*




## ✓ *OBSERVATIONS:*

- Some outliers exists in the variable “NoofMnthsInCurJob”, which should be treated before model building.

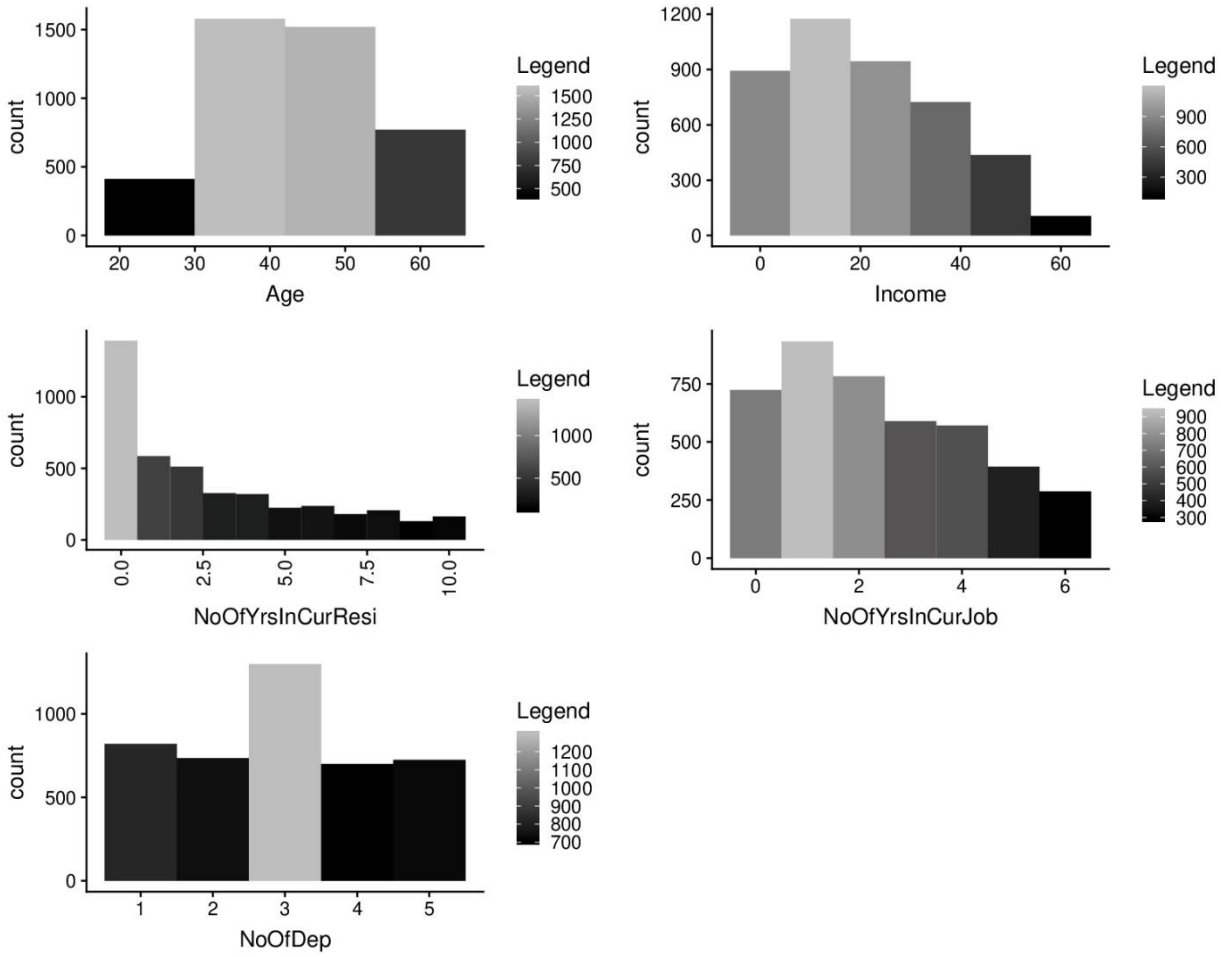
## ✓ DISTRIBUTION OF DEFAULTERS FOR CATEGORICAL VARIABLES:




## ✓ OBSERVATIONS:

- Male gender has more number of defaulters than Female Gender.
- Out of Total Defaulters, Married People accounts to 80% i.e. nearly 4 times as compared to Singles.
- The number of defaulters contributed by “SAL” profession is more compared to other two profession "SE" ,"SE_PROF".
- In the Education Sector, Others and PHD contributed to minimum number of defaulters, w.r.t to others in the same sector.
- 75% of total defaulters were living on rent whereas around 20% of the defaulters had their own houses. There were only a few number of defaulters who were “Living with Parents: or in the Company Provided Spaces, or some other other type of residence.

## ✓ HISTOGRAM DISTRIBUTION OF DEFAULTERS FOR CONTINUOUS VARIABLES:



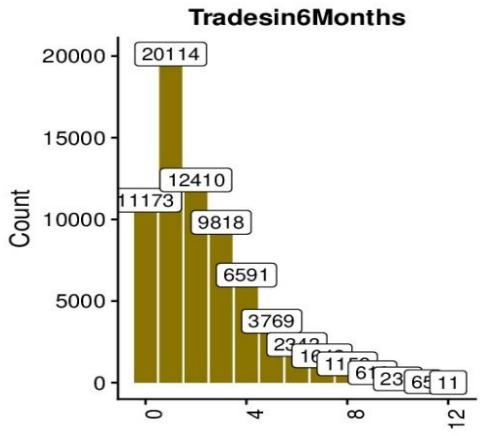

## ✓ OBSERVATIONS:

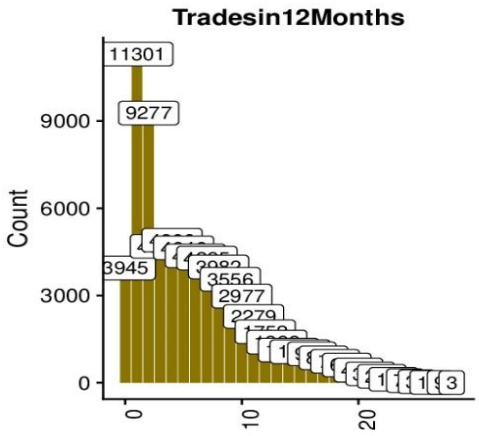
- Most of the defaulters were lying in the age ranging from 30 to 50.
- People having high income tended to default less; quite relatable fact.
- Usually, defaulters were the ones whose “Number of Years in Current Residence” or “Number of Years in Current Job” was less than 3.
- People having 3 dependents had the maximum count of defaults, nearly 1.5 times as compared to their counterparts.

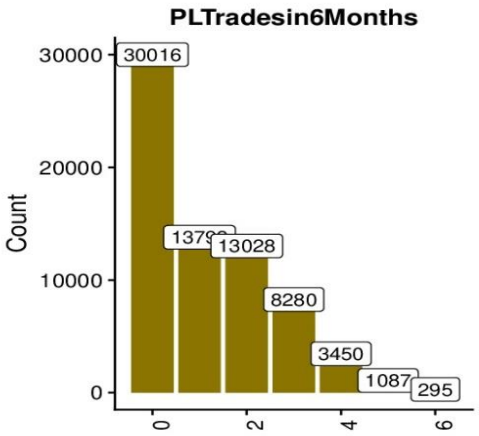
And that sums up some of our insights/observation of demographic data. Let proceed to the next analysis of Credit Bureau Data.

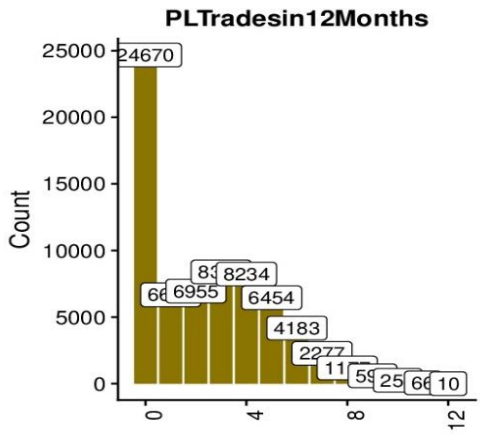
# ❖ EXPLORATORY DATA ANALYSIS – CREDIT BUREAU

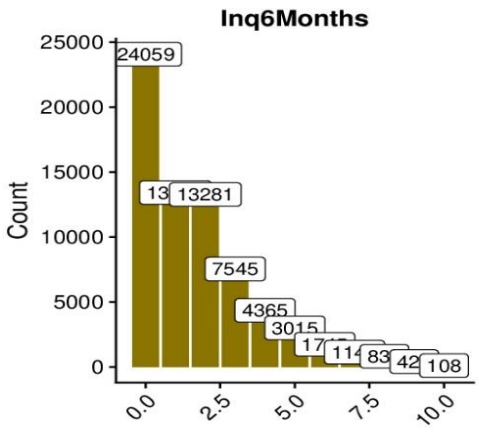
## ✓ *OVERALL DISTRIBUTION OF DATA*

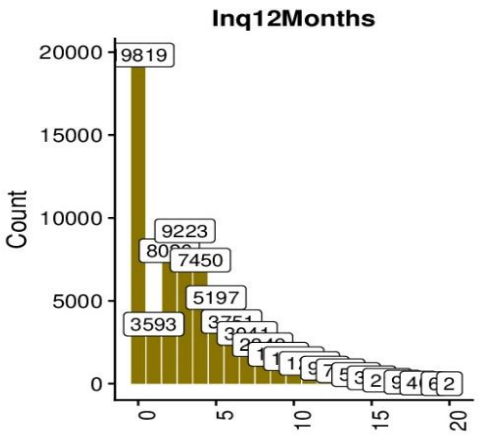


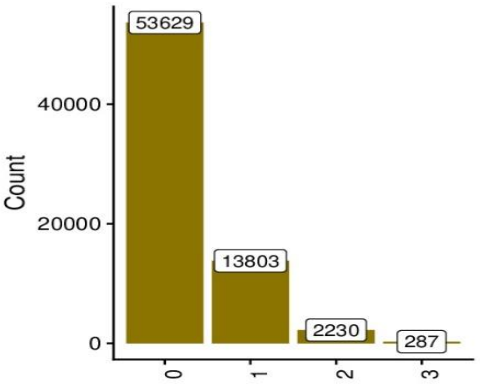



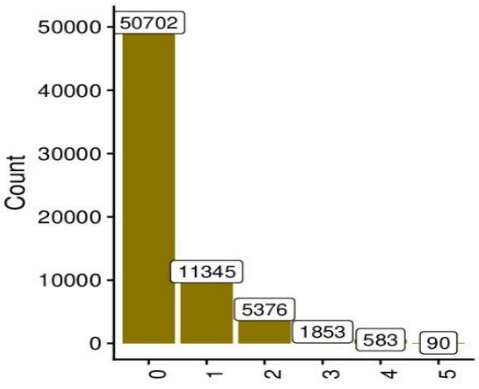



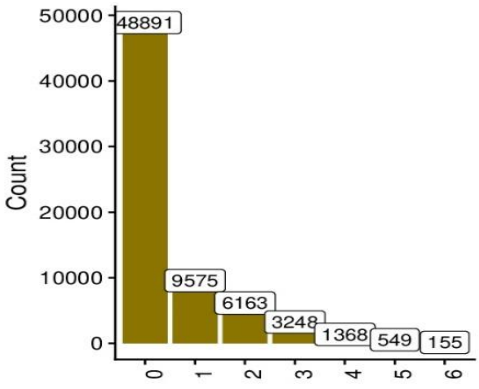



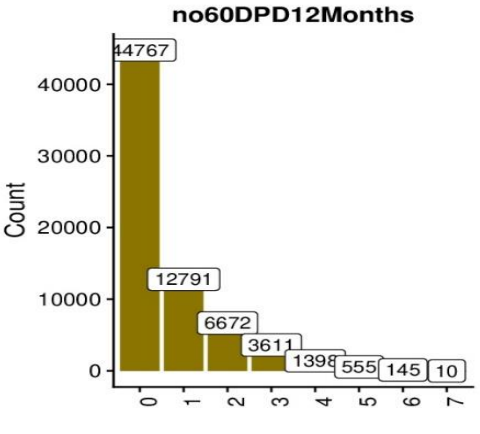



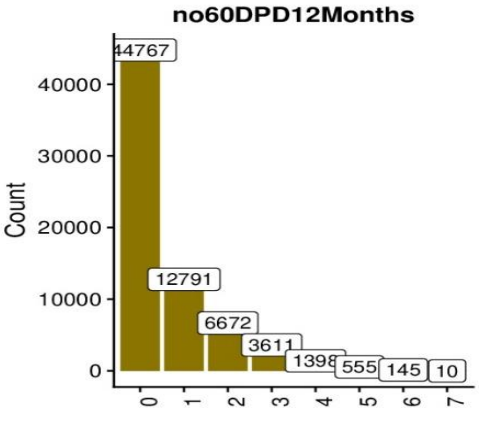



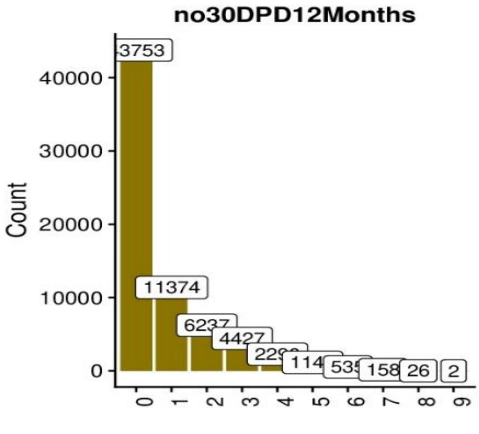



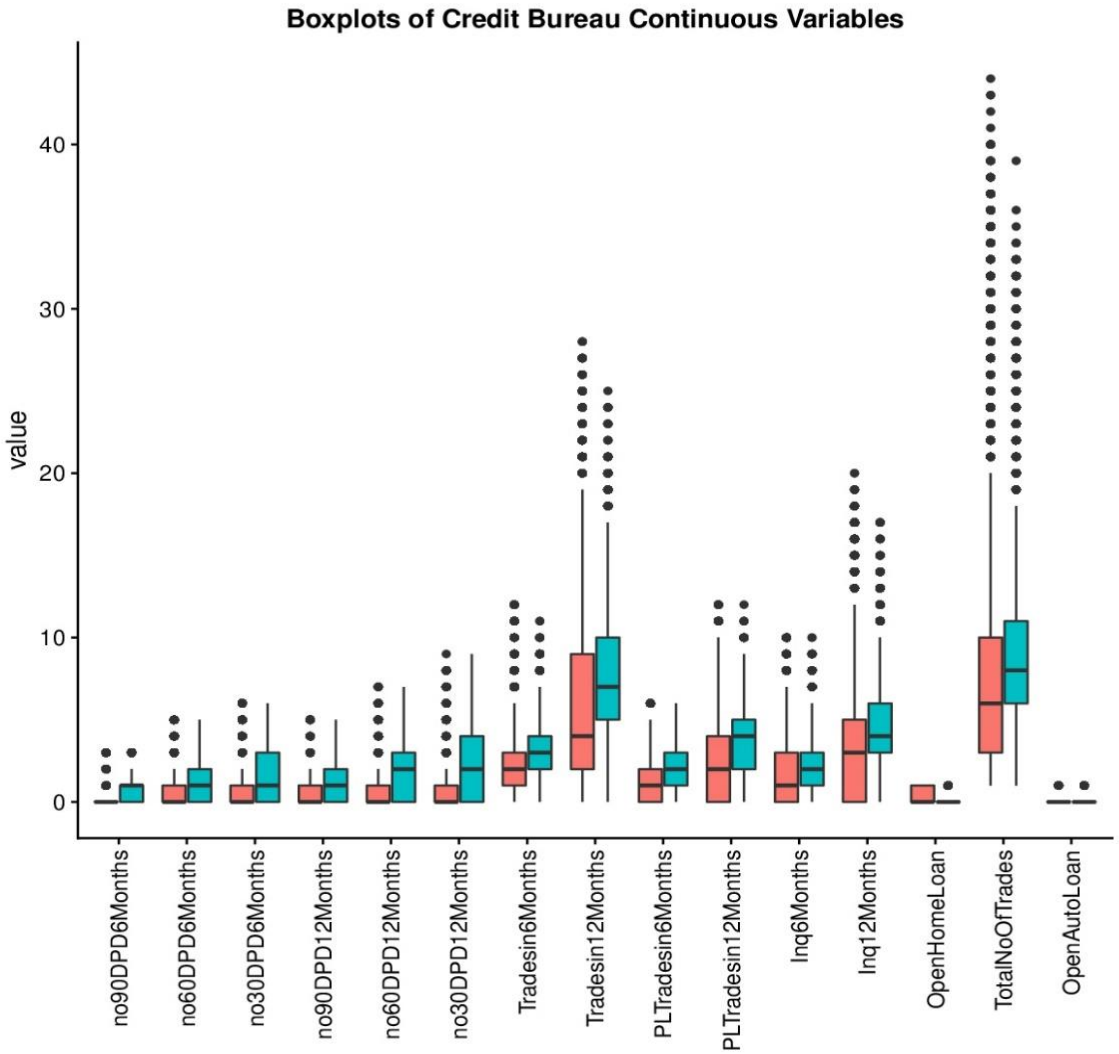



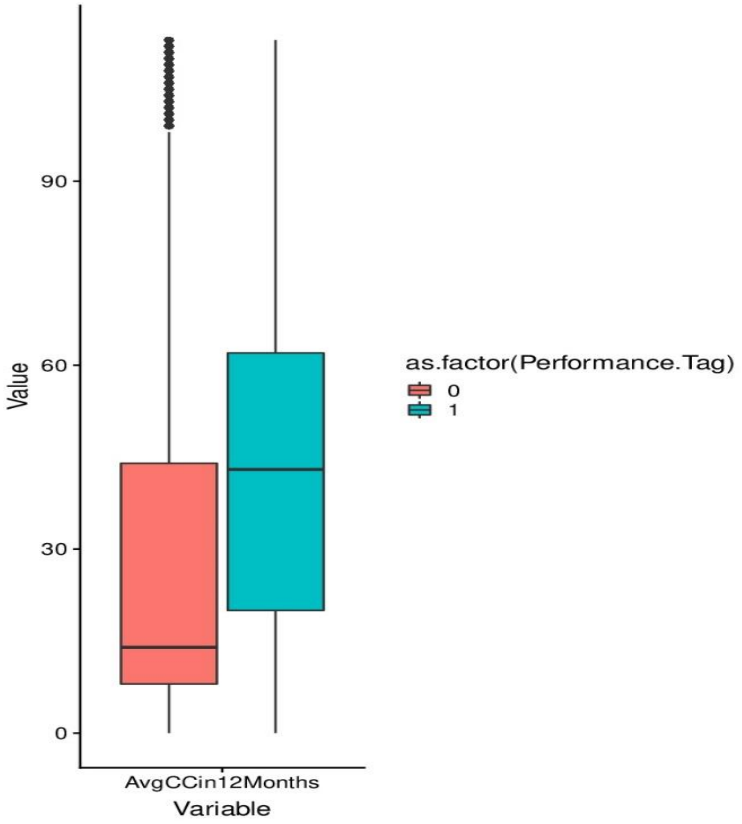



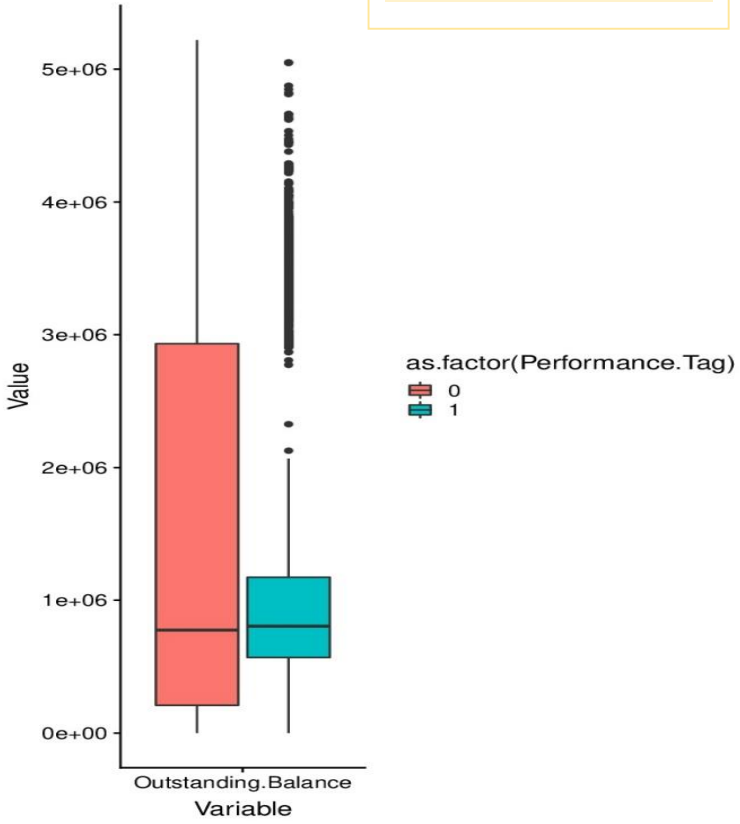
## ✓ *BOX PLOTS*



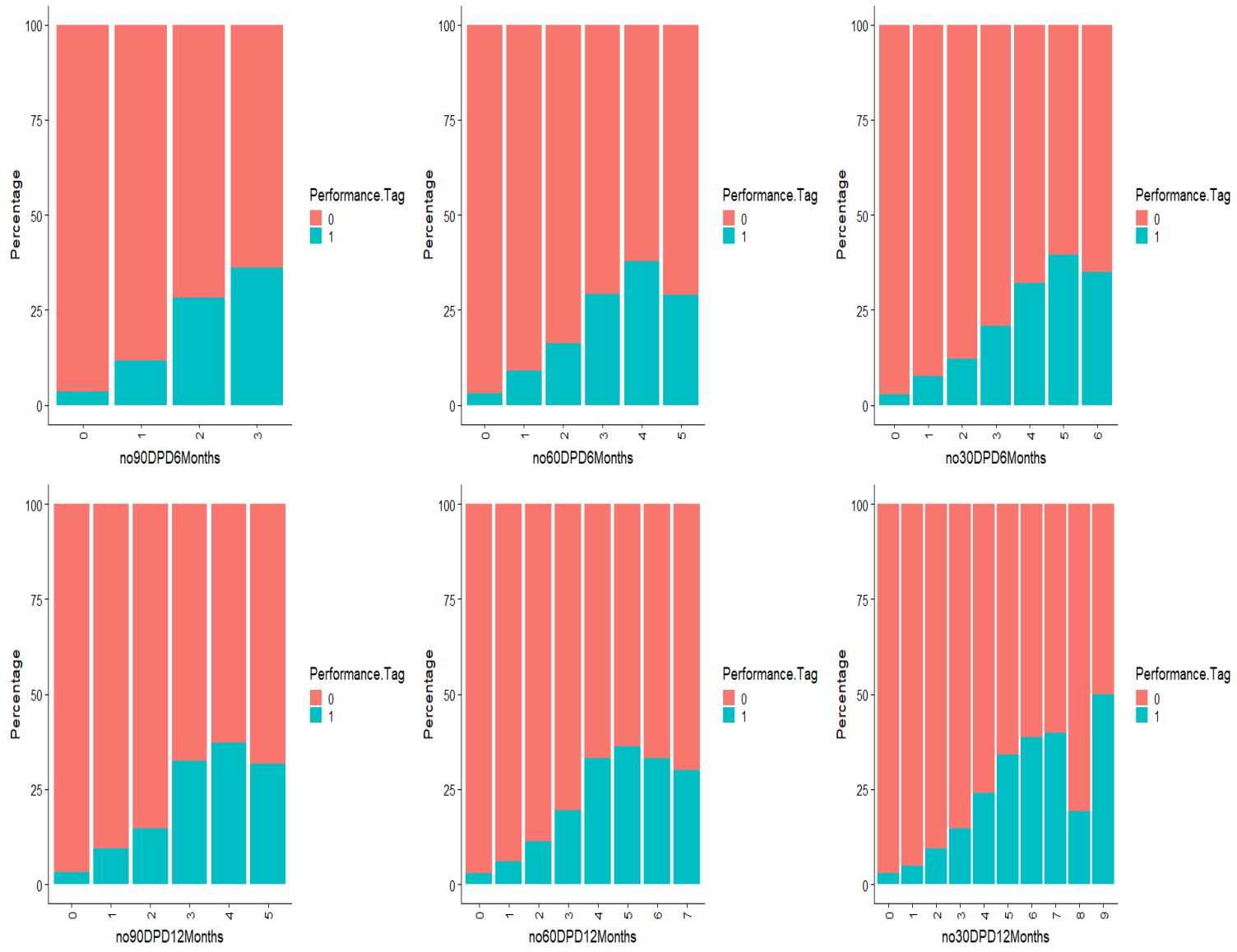

### Observations:

The box plots show us that considerable outliers exists in almost all the variables of this dataset, hence we will be treating them or capping them before model building.



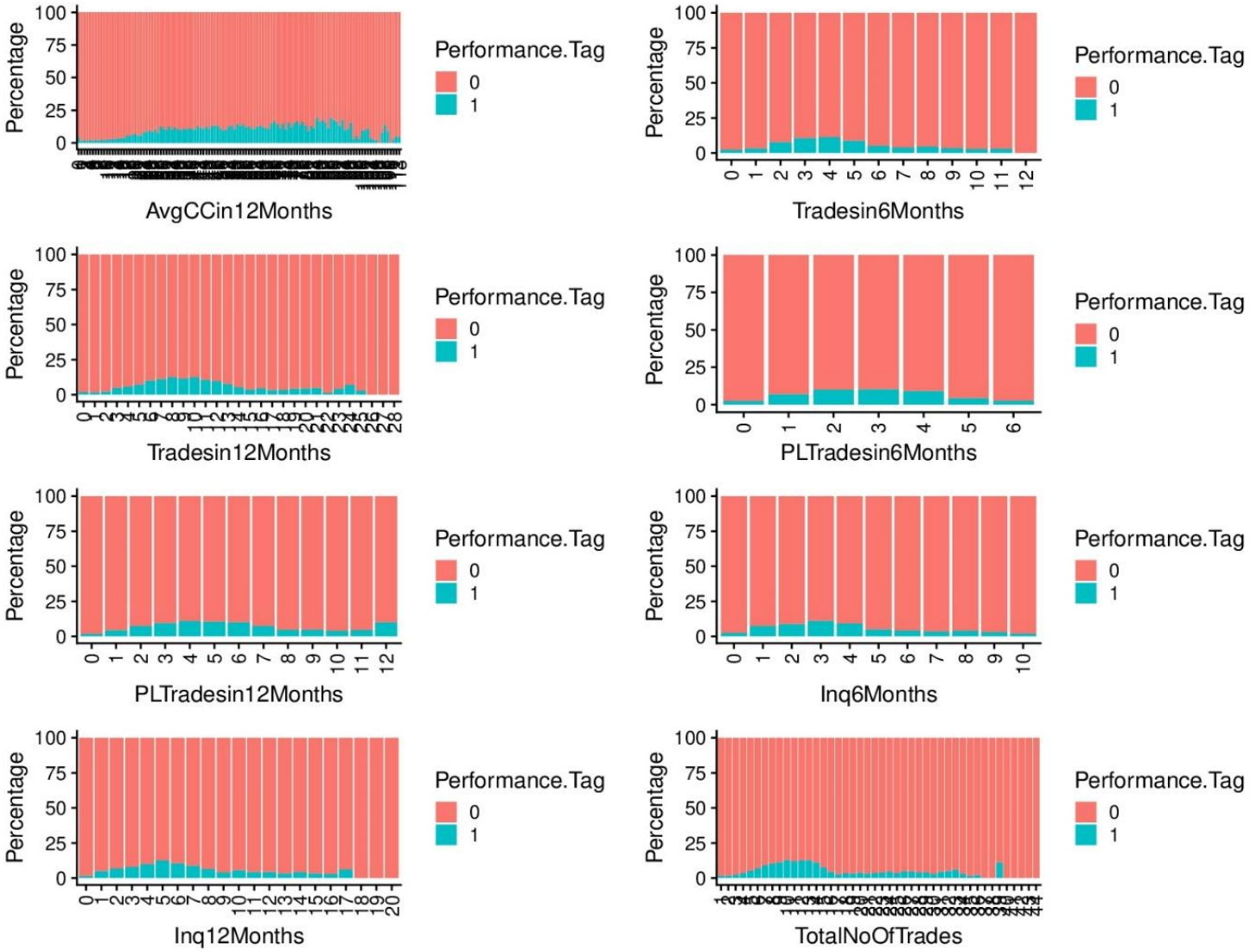
## ✓ DISTRIBUTION OF DEFAULTERS:




## ✓ OBSERVATIONS:

- It is clearly apparent from all the (DPD) graphs that as the number of DPD increases, the likelihood of default increases. So all the graphs follow almost a linear increase in default percentage. Hence all the six DPD variables can be of significant importance for predicting the defaulters.

## ✓ *DISTRIBUTION OF DEFAULTERS:*

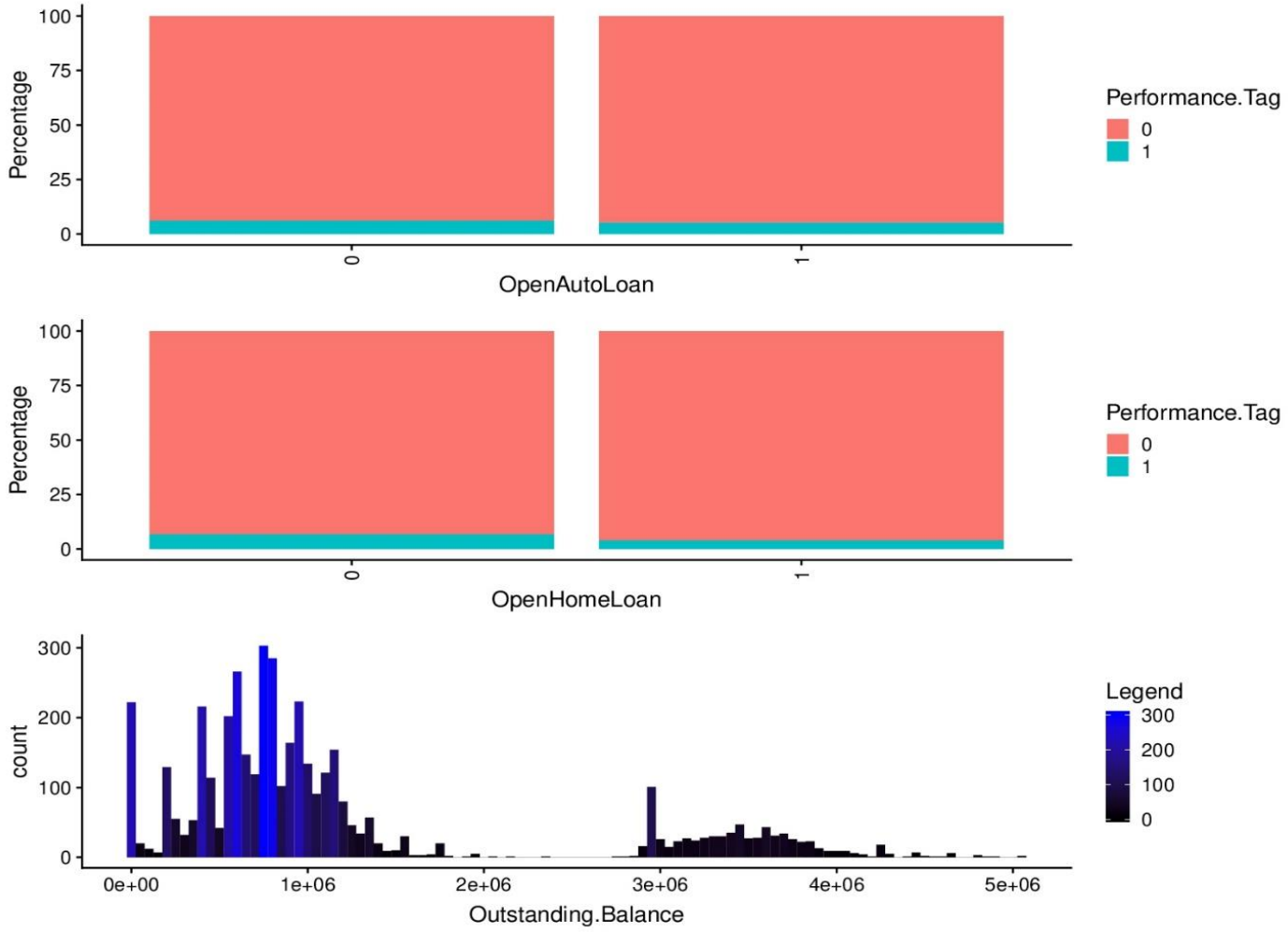



## ✓ *OBSERVATIONS:*

- Three defaulters of almost all the above six variables follow a nearly perfect bell curve i.e. lower percentage of defaulters towards both the ends, and higher percentage in the middle.

- However, the bell curve distribution of defaulters for the "TotalNoOfTrades" variable is slightly skewed towards the right. So we will be fixing them probably with log-transformation method before model building.

## ✓ DISTRIBUTION OF DEFAULTERS:

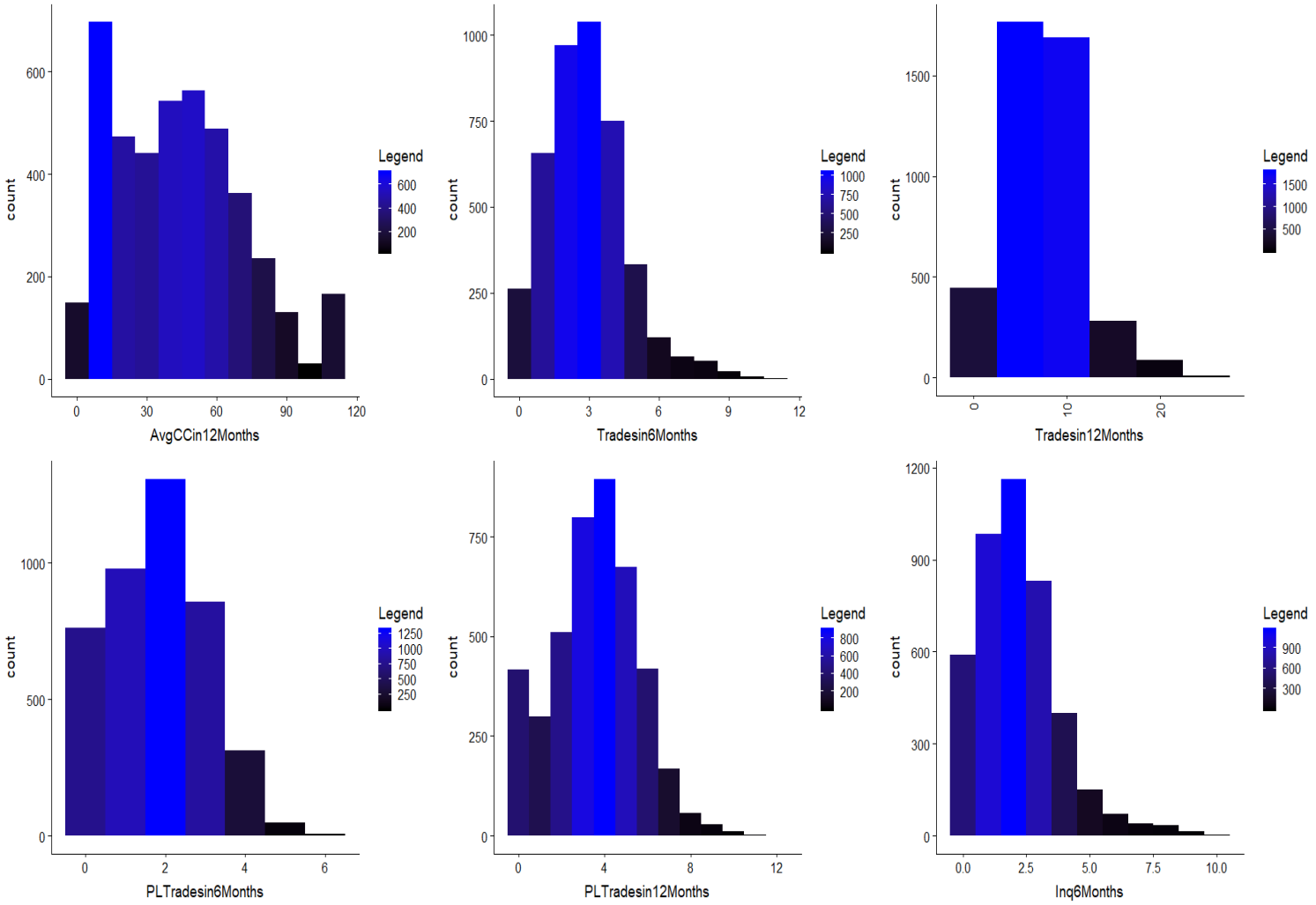



## ✓ OBSERVATIONS:

- Having an “Open Home Loan” or not having an “Open Home Loan” does not make much difference in predicting the defaulters. And same goes for the “Open Auto Loan” variable. So these variables does not appear of any significant importance for prediction.

- The Histogram Distribution of Outstanding Balance is almost random. However, the interesting insight is: The number of defaulters having less outstanding balance outnumber the number of defaulters having higher outstanding balance. Hence, this variable can be of moderate importance for us, later.

## ✓ DISTRIBUTION OF DEFAULTERS:



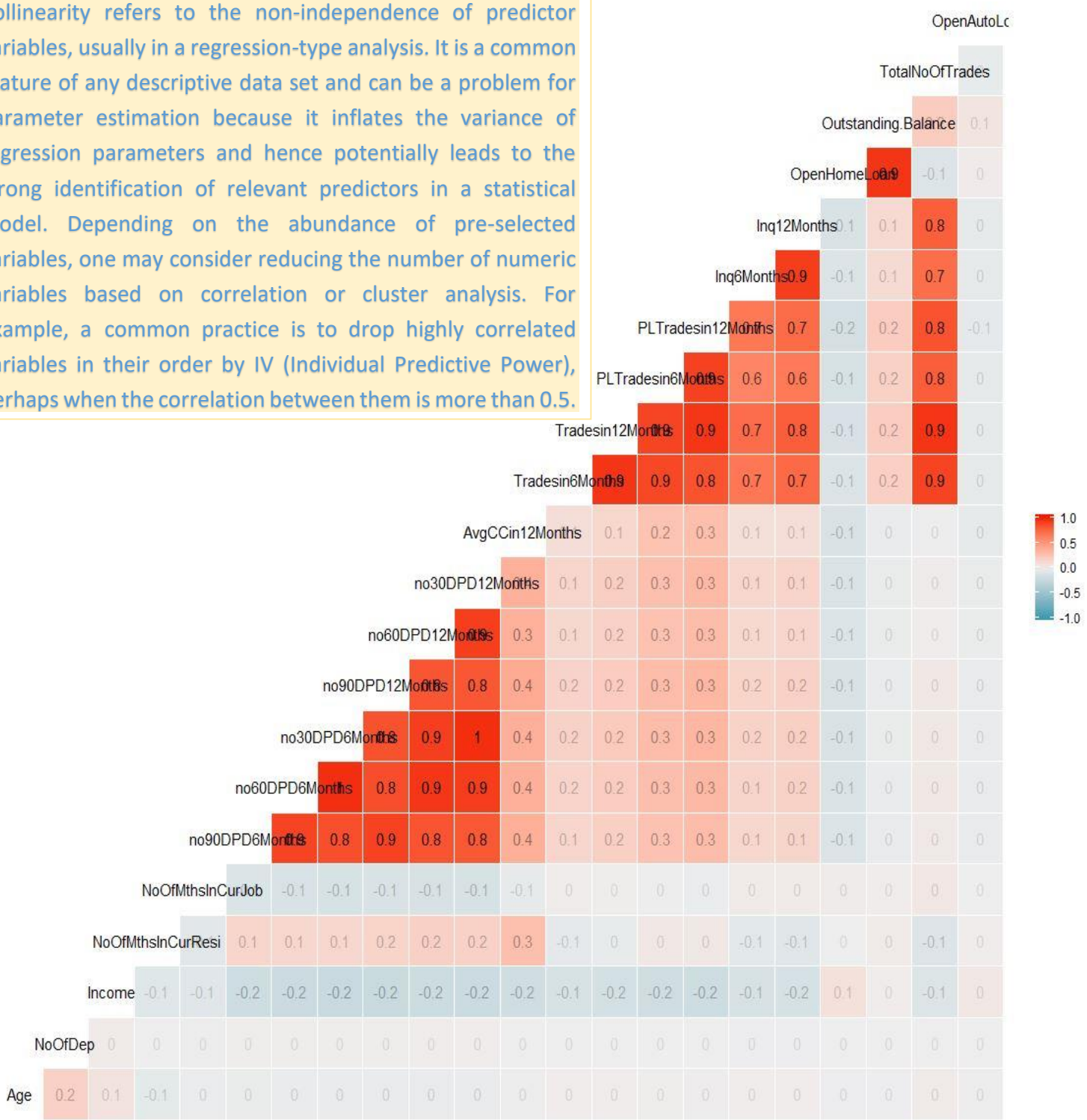

## ✓ OBSERVATIONS:

- Again, almost all the above six variables follow a bell curve i.e. lower percentage of defaulters towards both the ends, and higher percentage in the middle.

- However, the histograms of last three variables i.e. “PL Trades in 6 Months”, “PL Trades in 12 months” and “Inq 6 Months” is skewed towards right. So we will be making them more symmetric with log transformation method.

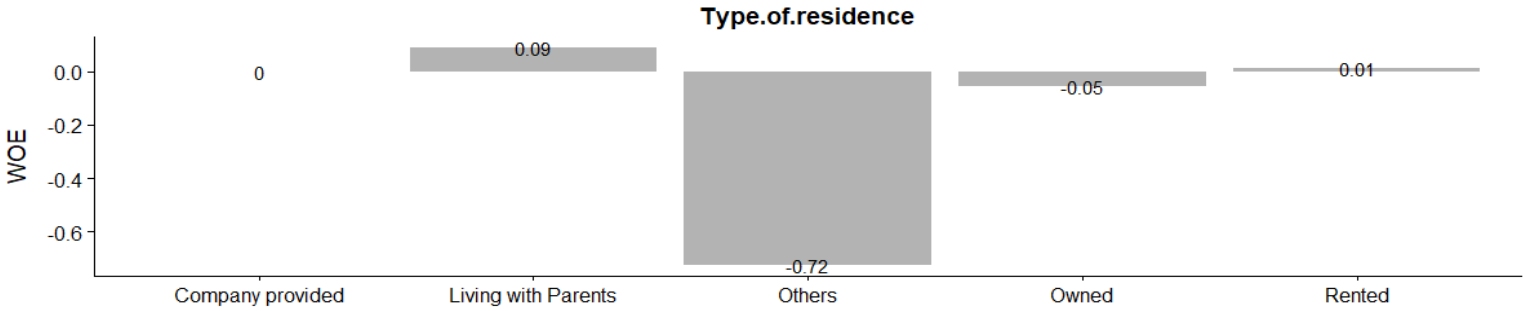
## ✓ CORRELATION MATRIX OF ALL THE VARIABLES [DEMOGRAPHIC + CREDIT BUREAU]:

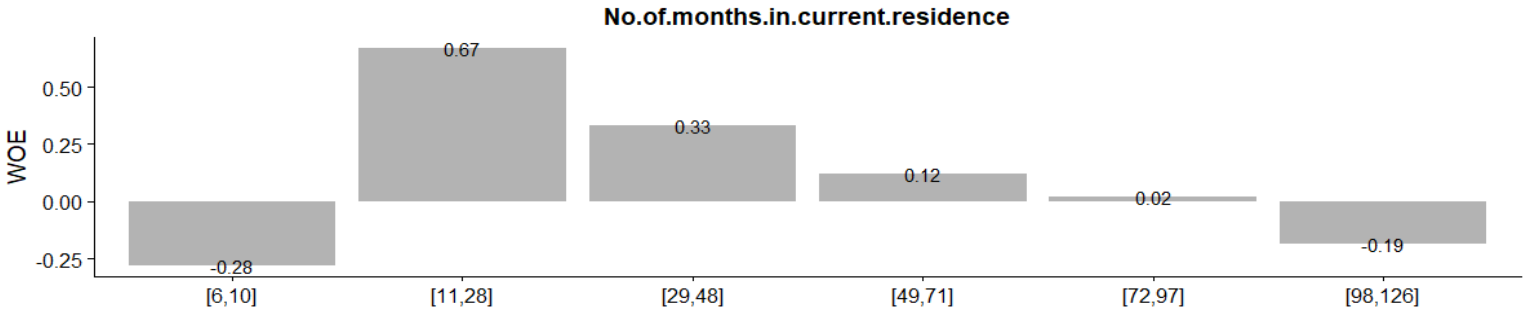
Collinearity refers to the non-independence of predictor variables, usually in a regression-type analysis. It is a common feature of any descriptive data set and can be a problem for parameter estimation because it inflates the variance of regression parameters and hence potentially leads to the wrong identification of relevant predictors in a statistical model. Depending on the abundance of pre-selected variables, one may consider reducing the number of numeric variables based on correlation or cluster analysis. For example, a common practice is to drop highly correlated variables in their order by IV (Individual Predictive Power), perhaps when the correlation between them is more than 0.5.

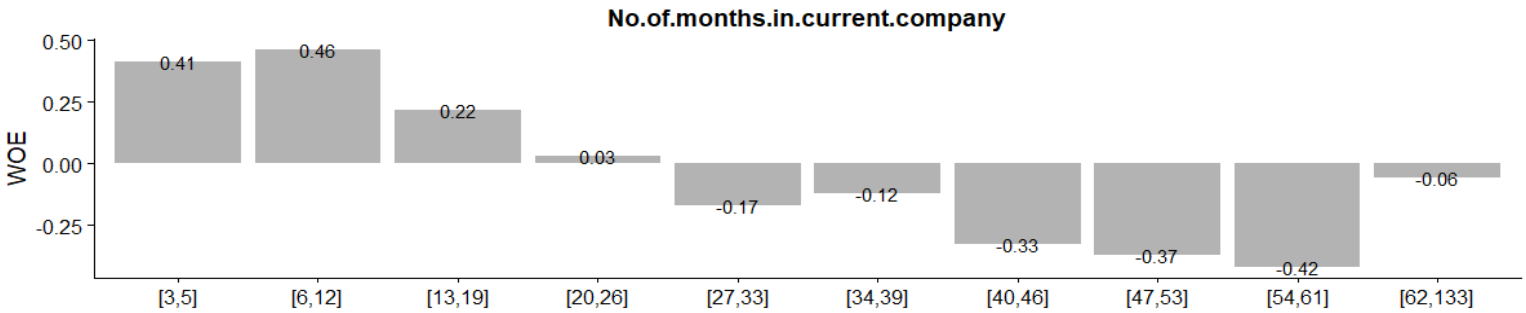


# ❖ WOE AND INFORMATION VALUE ANALYSIS:

## ✓ *WOE PLOTS:*




## ✓ *OBSERVATIONS:*

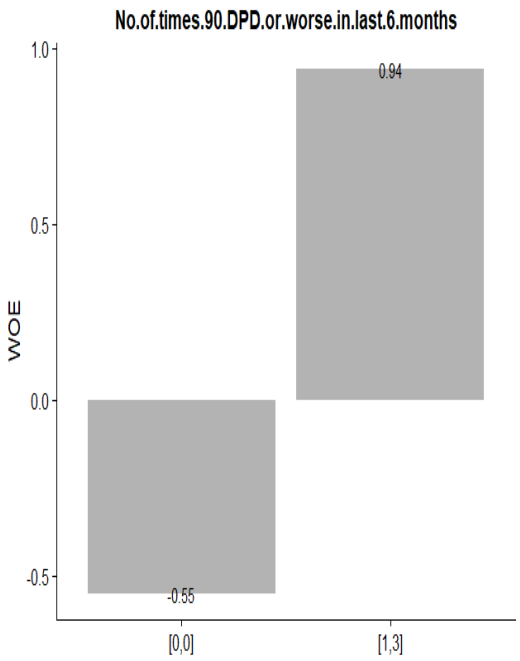
- Applicants who are living with parents tend to default more.
- Applicant whose “No. of months in current residence” is between 11 to 28 has the highest chances of default.
- Applicant whose “No. of months in current company” is between 6 to 12 has the highest chances of default.

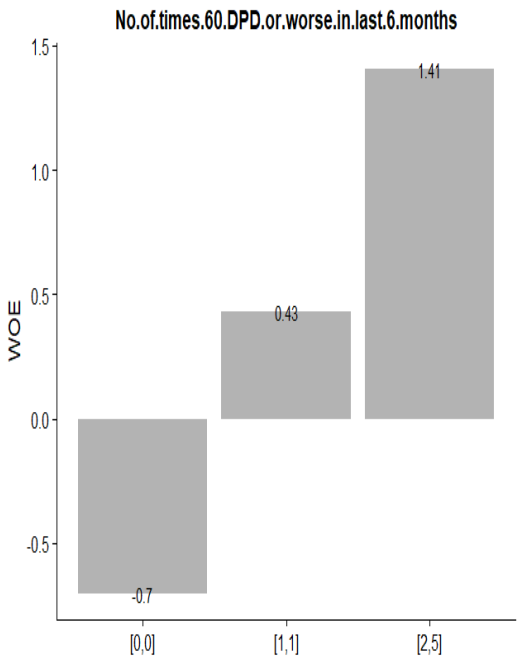
- ***PLEASE NOTE:***

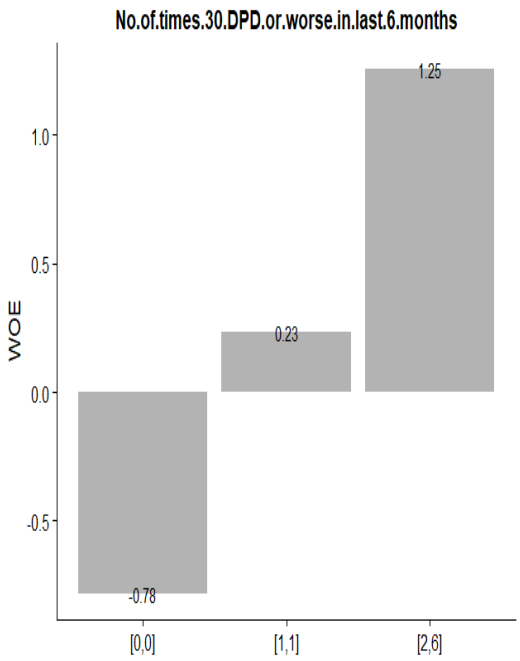
WOE describes the relationship between a predictive variable and a binary target variable Y. (Y = “Performance Tag” in our case). Now the greater the value of WOE, the higher the chance of observing Y=1.Hence in our dataset, since Y=1 is the indication of default, the interpretation for WOE specifically for our case will be:

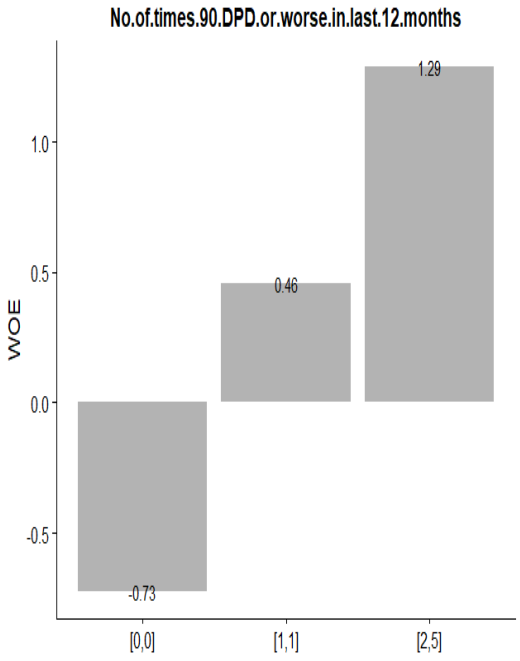
**Higher the WOE, higher the chances of default.**

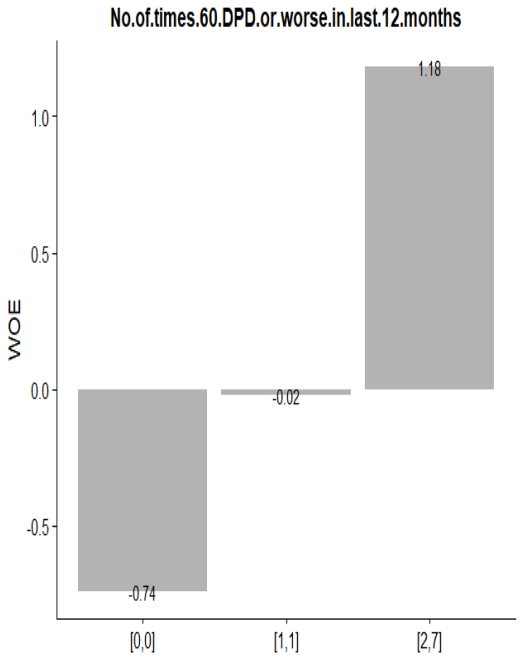
## ✓ *WOE PLOTS:*

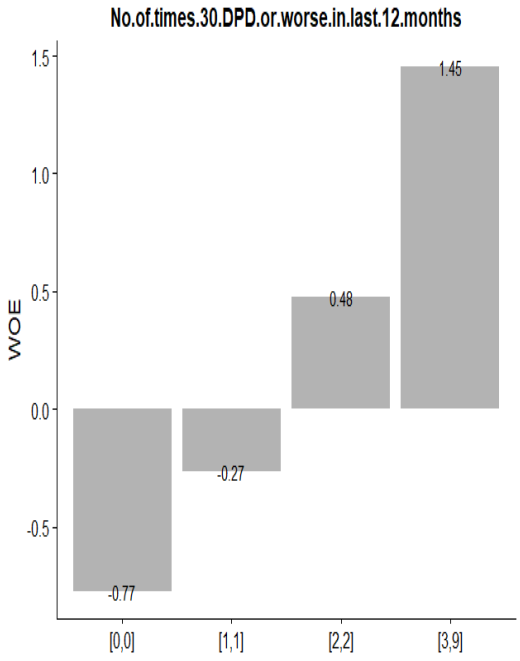


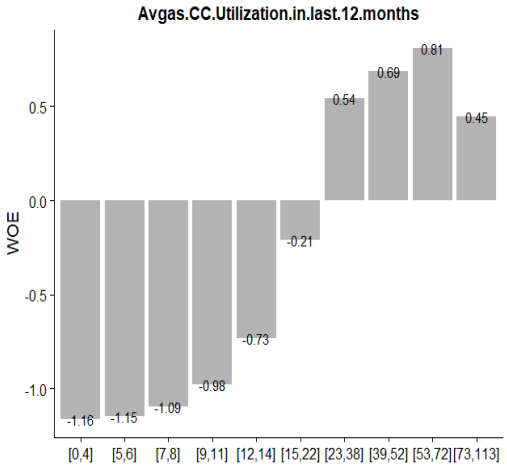



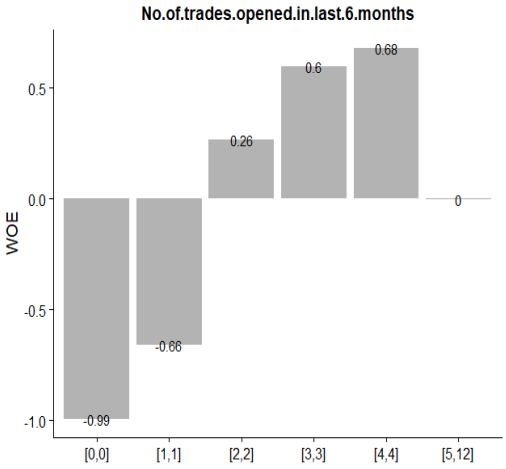



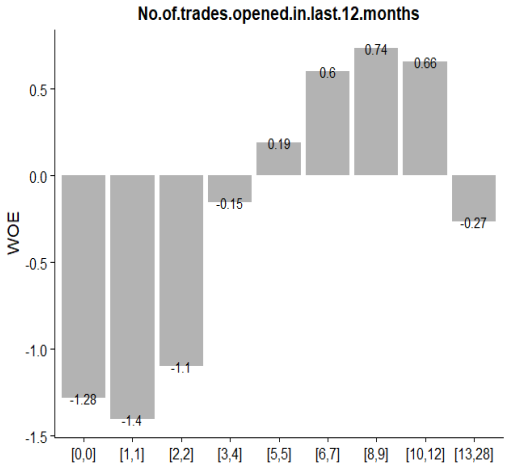

## ✓ *OBSERVATIONS:*

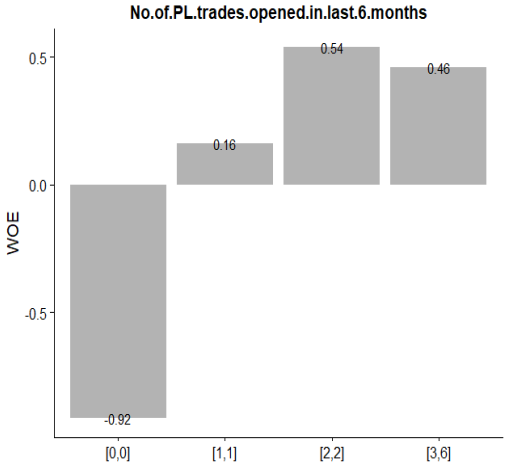
- All the six plots indicate the same insight i.e. as the “No of times of DPD of Applicant” exceeds 2, applicants become more likely to default.

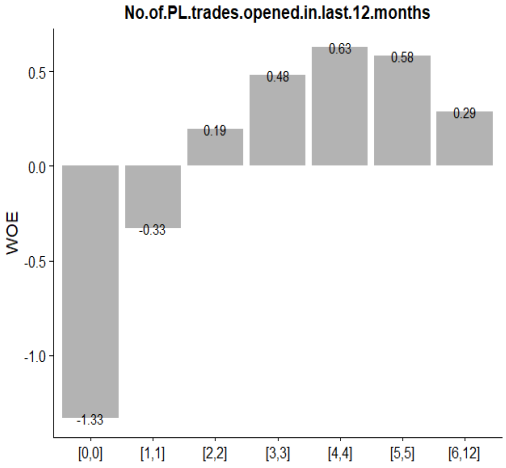
## ✓ WOE PLOTS:

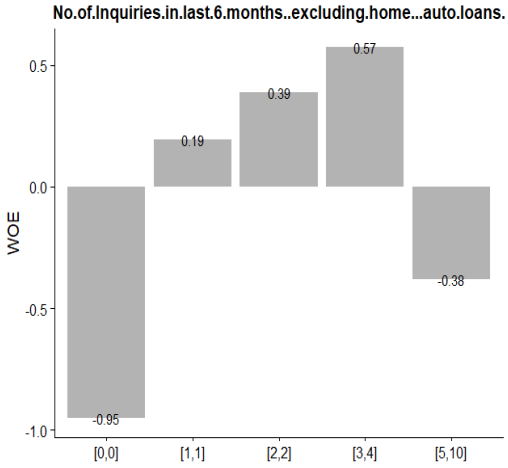


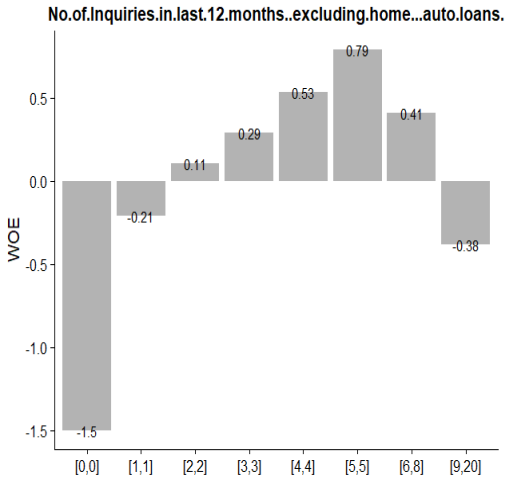



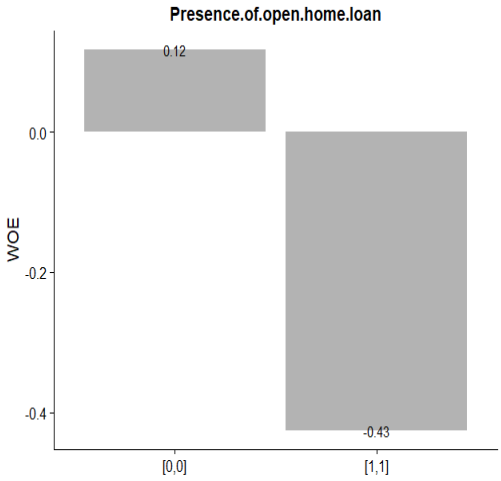



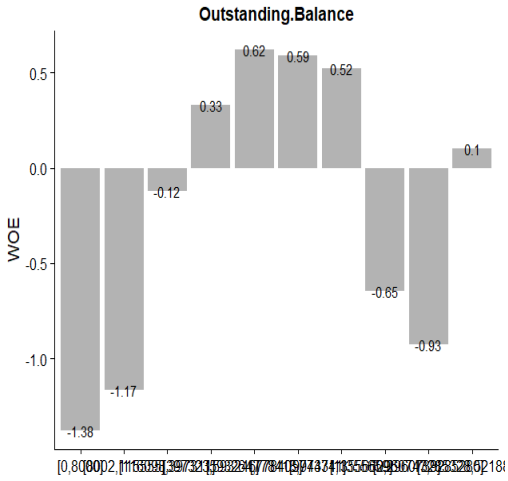

## ✓ OBSERVATIONS:

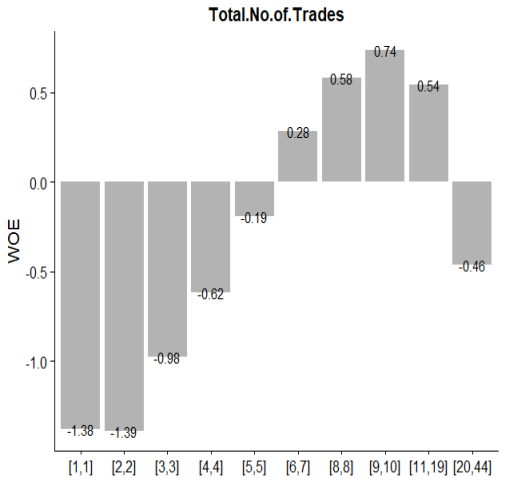
- Applicant whose “Average Credit Card Utilization in last 12 months” lies between 53 to 72 has the highest chances of default.
- Applicant whose “No of Trades opened in last 6 months” = 4 has the highest chances of default.
- Applicant whose “No of Trades opened in last 12 months” = 8 or 9 has the highest chances of default.
- Applicant whose “No of PL Trades opened in last 6 months” = 2 has the highest chances of default.
- Applicant whose “No of PL Trades opened in last 12 months” = 4 has the highest chances of default.
- Applicant whose “No of Inquires in last 6 months” = 3 or 4 has the highest chances of default.

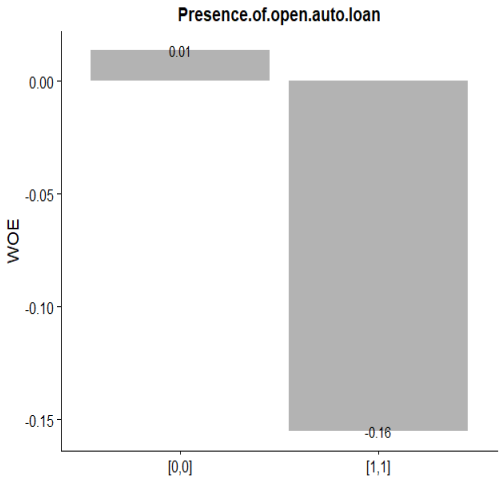
## ✓ WOE PLOTS:

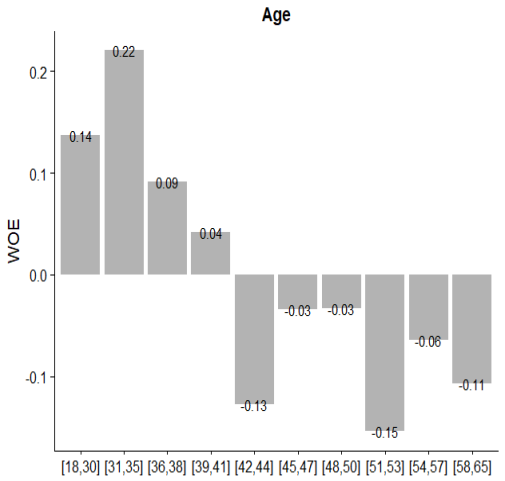


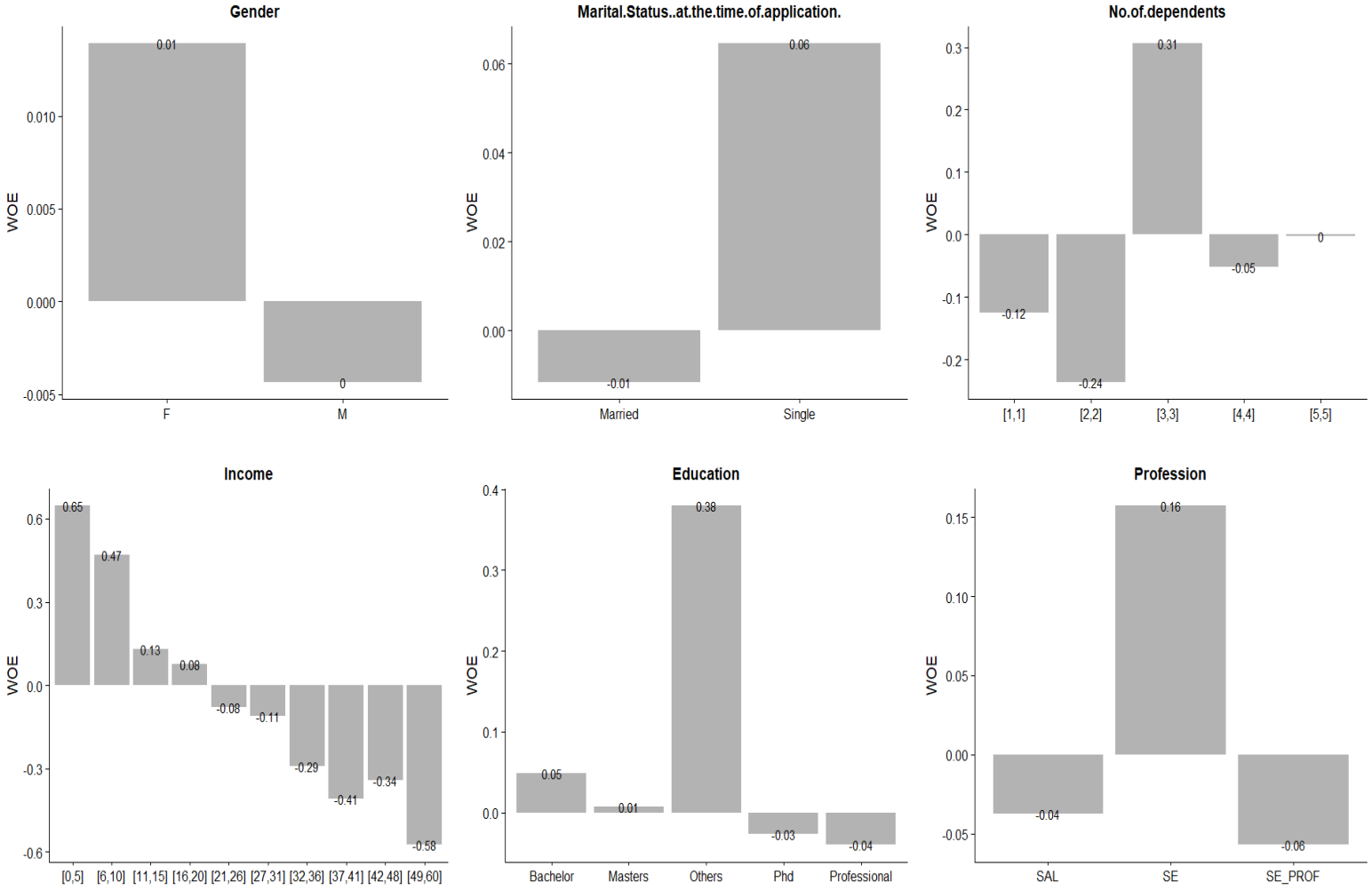




## ✓ OBSERVATIONS:

- Applicant whose “No of Inquires in last 12 months” = 5 has the highest chances of default.
- Applicant who “Do not have an open home loan” has higher chances of default compared to those who has an open home loan.
- Applicant whose “Outstanding Balance” is between 10 to 20 Lacs has higher chances of default.
- Applicant whose “Total no of Trades” is 9 or 10 has higher chances of default.
- Applicant who “Do not have an open auto loan” has higher chances of default compared to those who has an open auto loan.
- Applicant whose “Age” lies between 31 to 35 has the highest chances of default.

## ✓ *WOE PLOTS:*




## ✓ *OBSERVATIONS:*

- Females tend to default more than males.
- Applicant who are single tend to default more than the married ones.
- Applicant whose “No of dependents” = 3 has higher chances of default.
- Applicant whose “Income” is less than 5K tend to default more.
- Applicant who “Education” is Others has highest chances of default.
- Applicant who are Self Employed tend to default more.

# ❖ MODEL BUILDING APPROACH

- ✓ In this section, we provide an overview of popular predictive modelling techniques for credit scoring. We attempt to describe the both parametric and non-parametric modelling techniques. Each of the methods has its own strengths and weakness, which often vary according to the circumstances and depend on the data quality.

| Method | Main technique | Summary |
|---|---|---|
| Logistic regression | Maximum likelihood estimation | Determine formula to estimate binary response variable. |
| Decision trees | Recursive Partitioning Algorithms | Uses tree structure to maximize between group differences. |
| Neural Networks | Multilayer perceptron | Artificial Intelligence technique, whose results are difficult to interpret. |

## ✓ LOGISTIC REGRESSION MODEL:

Our Starting point will be the most commonly used parametric method for credit scoring. Logistic regression uses Maximum likelihood estimation process, which transforms the dependent variable into a log function and estimates the regression coefficients in a way it maximizes the log-likelihood.

Logistic regression requires the following assumption:

• The target variable is categorical (binary in the case of credit scoring)

• It has linear relationship with the log odds function

• It has independent error terms

• The predictor variables are uncorrelated

Logistic regression has been an optimal choice for developing credit scoring models as it's designed to handle binary outcome and it provides final probability that cannot fall outside of the range 0 to 1. Also, it provides fairly robust estimate of the true probability, given available Information.

## ✓ DECISION TREE MODELS : PRUNED & ENSEMBLE TREE MODELS:

Decision tree models utilize recursive partitioning algorithms to produce terminal nodes that are homogeneous.

The top of a tree is referred as the root node, each subsequent level as a daughter node, and the bottom ones are called the terminal nodes. The root node contains a sample of subjects from which the tree is grown in the learning/training sample. All nodes in the same layer constitute a partition of the root node. Moreover, every node in the tree is merely a subset of the learning sample

Recursive partitioning offers an alternative to parametric methods. Non-parametric approaches are appealing in a sense that it does not require, or few if any, assumptions about the underlying data, compared to parametric methods.

## ✓ NEURAL NETWORK MODELS:

In recent years, the progress on Artificial Intelligence (AI) has been made in attempt replace domain of thinking and decision making with computers. Neural Networks(NNs) are the core of AI for the use of predictive model as part of a decision process.

The end result of Neural Networks is something like a decision tree, however the details is much finer and the decision rules are much complex and difficult to interpret. The best suited Neural Network development technique for credit scoring is the Multilayer Perceptron (MLP) as it can deal with both non-linearity and interactions easily. The NNs have the advantages of being able process huge amount of data, discover patterns and relations in the data, and adaptable to changing circumstances.

On the other hand, they can overfit to the training sample and overly complex in a way that the relations detected by the models are very difficult to interpret. Nevertheless, one makes a use of such model as a benchmark comparison to other credit scoring models.

# ❖ MODEL EVALUATION APPROACH

Usually, several statistical tools are used to analyse the result of the scorecard development. However, we will mainly focus on the optimization through Gini coefficient and Area under the Curve(AUC) methods. Also we will use Receiver Operating Characteristic(ROC) curve as a visualization tool that illustrates the performance of a binary classifier system.

If a default event is correctly classified and predicted as default, we call it true positive (tp); and if a non-default event is wrongly predicted as a default, it's counted as a false positive (fp).

Accordingly, the following parameters are computed:

$$rp\ rate = TPR = \frac{Defaults\ correctly\ classified(tp)}{total\ defaults(p)}$$

$$fp\ rate = FPR = \frac{non\ defaults\ incorrectly\ classified(fp)}{total\ non\ defaults(n)}$$

Plotting the fraction of correctly classified defaults (tp rate) versus the incorrectly classified non-default (fp rate) give rise to the **ROC curve.**

Often simply referred as **AUC, or AUROC**, is equal to the probability that a classifier would rank a randomly chosen positive instance higher than a randomly chosen negative one, given that one assumes 'positive' ranks higher than 'negative'.

Once we obtain AUC, gini coefficient can be expressed as follows:

**Gini coefficient = $2 * AUC - 1$**

# ❖ APPLICATION SCORE CARD APPROACH

## ✓ BACKGROUND: CREDIT SCORING MODELS

Perhaps the simplest way to explain credit-scoring is to separate the word components into" **credit" and" scoring"**, where the word credit represents borrowers willingness and ability to pay back, as well as the potential financial impact due to observed or perceived changes in borrower's creditworthiness. Meanwhile, the word" scoring" implies ranking and ordering of cases or borrowers according to their perceived or observed quality which may distinguish into the separate groups of good and bad borrowers. Hence, credit scoring is the use of statistical models and techniques to translate relevant data of into numerical scores that guide credit decisions.

**Generally, higher the estimated score, less the riskiness of the client, and thus more reliable the client is perceived.**

## ✓ OUR APPROACH TO PREPARE CREDIT SCORING MODELS:

The above stated model building and evaluation approach will put us in good state to identify the customers as good or bad by analysing their profiles and predicting the odds of being good.

**(recall that the logit equation is log(odds) = ∑βixi ,where 'i' represents the ith variable.**

Note that we are using the odds of 'good' (rather than the odds of 'default' or 'bad'). Thus, the good applicants will have higher odds than the bad ones.

So if the probability of a certain applicant being 'good' is p = 70%, the odds of being good are: odds(good) = P(good) / P(bad) = 7/3 = 2.33

Once we have the odds, we can sort the applicants from high to low odds (i.e. good to bad). Now, the industry prefers to talk in terms of scores (e.g. between 200 to 900) rather than odds. Thus, we will simply calibrate the odds to a scale of scores, probably between some arbitrary range like 200 to 900.

Once the scores are calculated, we can decide a threshold score above which an applicant will be labelled 'good' or 'bad'. This is exactly equivalent to choosing a probability threshold in a logistic regression model.

## ❖ SCORE CARD ASSESSMENT APPROACH

Scorecard development is an iterative process and the resultant scorecard must satisfy a number of performance standards, including:

• **Stability** - The scorecard attributes should be estimated from a sufficiently sized dataset that covers appropriate historical period.

• **Discriminative** - The scorecard is expected to distinguish between the goods and bads.

• **Interpretable** - The output of the scorecard should be understandable and explainable to non-experts.

• **Not overly complex** - The scorecard should not-rely on any single feature. Also, it should not consist of too many features.

On the other hand, there is no quantitative definitive framework available to ensure that the criterion is satisfied. The decision making process varies for each organization, depending on a range of factors such as the available data and resources, as well as the corporate and regional cultures.

# THANK YOU!