BFS Capstone Project

Credit Risk Analysis - Part 1(EDA,Model Building & Score Card preparation)

Group Name: Deep learners

1. Sivaiah Aelasrolu
2. Hariharan Subramaniam
3. Thripthi Raj
4. Suraj Kumar Talreja

About the Document

This Document intensively focuses on the following aspects of the project and termed as ***Part1***

1.Business Understanding
2.Project Framework
3.Data Cleaning and Preparation
4.Explorartory analysis
5.Model building and Evaluation
6.Preparation of shoe card

Part 2 which is another document focuses on Financial benefit.The intention to create a separate document to detail the financial benefits of the project helps maintain a document relevant for business stakeholders.This document will be termed as Part 2 of the 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. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

Objective:

To help CredX identify the right customers using predictive models.

Scope of the project:

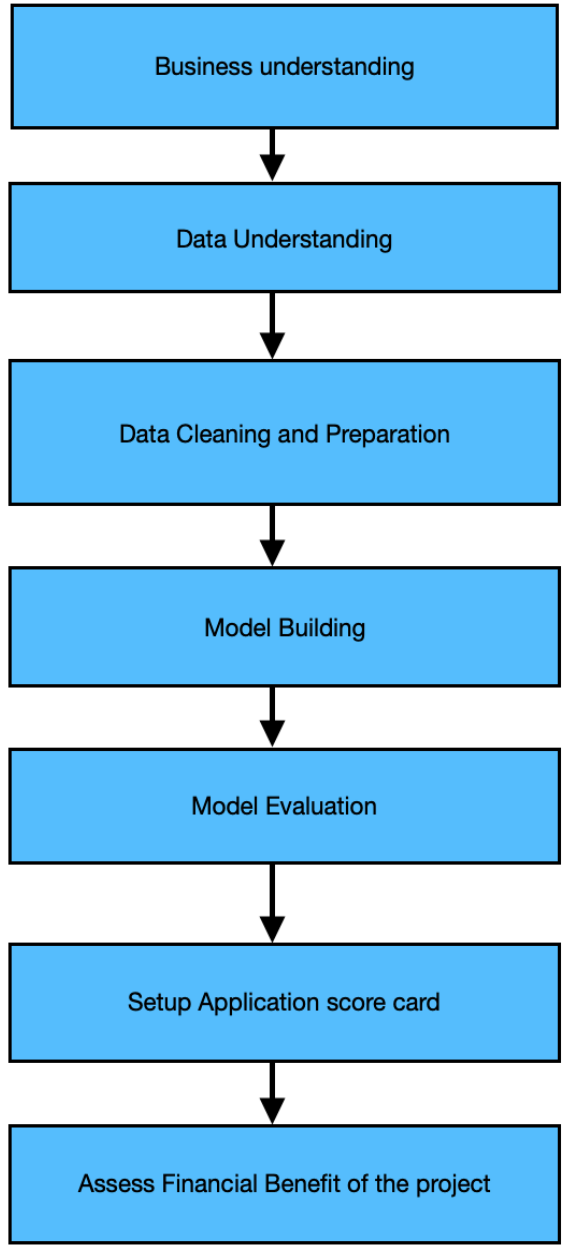
1.To determine the factors affecting credit risk.
2.Create strategies to mitigate the acquisition risk.
3.Assess the financial benefit of the project.

Project Framework

Frame work

CRISP DM framework will be used to execute the project. This framework has been practiced widely in the data science projects and hence been chosen for this project.

The overall steps that will be followed along with projects key objectives are laid out in the flow chart.



Data Cleaning and Preparation

- The datasets ‘Demographic’ and ‘Credit bureau’ were imported into R
- Its Dimensions, Structure and Summary statics were checked
- The datasets were checked for NAs columns wise
- The datasets ‘Demographic’ and ‘Credit bureau’ were merged using the ‘Application ID’ column
- Performed EDA by plotting each variable against the target variable (Performance tag) and observed distribution of data and its significance against the target variable, Outliers
- Treat outliers such as replacing the –ve income values using boundary values
- Remove unwanted data such as values below 18 in the ‘age’ variable as credit cards are provided to individuals who are 18 and above.
- Created separate data frame for rejected records i.e. rows that have ‘Performance tag’ as NAs
- Create a dataset with the predictor variables transformed into WOE, the NAs were replaced by WOE values
- Determined IV (Information value) to identify important predictors for modelling
- Created different subsets for datasets for the purpose of modelling and learning

Dataset Summary

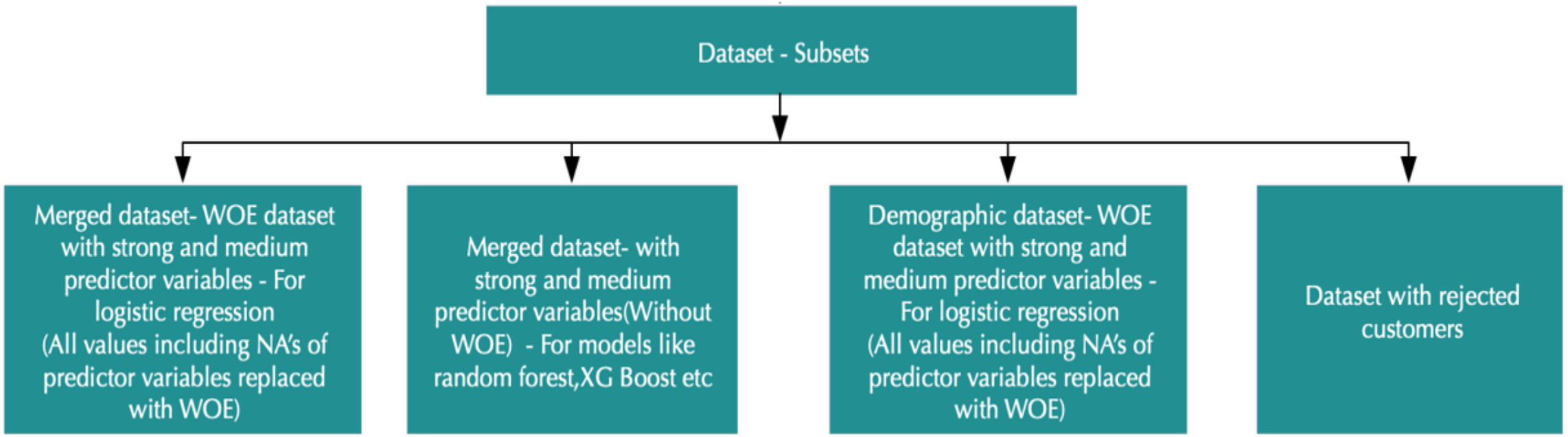
	Demographic Dataset	Credit Bureau Dataset
No of Columns	19	12
No of Rows	71295	71295
Target Variable	Performance Tag	Performance Tag
Duplicates	3	3
NAs	1577	3028
NAs -Columns wise(highest to lowest)	Performance Tag -1425,Education-119,Profession-14,Type of residence-8,Marital Status -6,No of dependents -3,Gender-2	Performance Tag-1425,Avgas CC Utilization in last 12 months -1058,Presence of open home loan-272,Outstanding Balance-272,No of trades opened in last 6 months -1
Common & Identical fields	Application_ID,Performance_Tag	Application_ID,Performance_Tag
Event Rate	4.13%	4.13%

Note :

- It is observed that the Target variable i.e. 'Performance Tag' is imbalanced as the defaulted customer's data is only 4.13 %.The dataset needs to be balanced prior to Model building.
- As the objective of the project is to find the good customers the values(0's,1's) of the target variables in the dataset will be swapped.0's in the dataset corresponds to good customers which will be converted into 1's likewise 1's in the dataset that corresponds to default customers will be converted into 0's
- It is observed that Demographics and Credit Bureau Dataset have 1425 NA values in 'Performance Tag' column, these are rejected applicants. A separate dataframe of Rejected Applicants will be created to compare the Score with Accepted applicants.

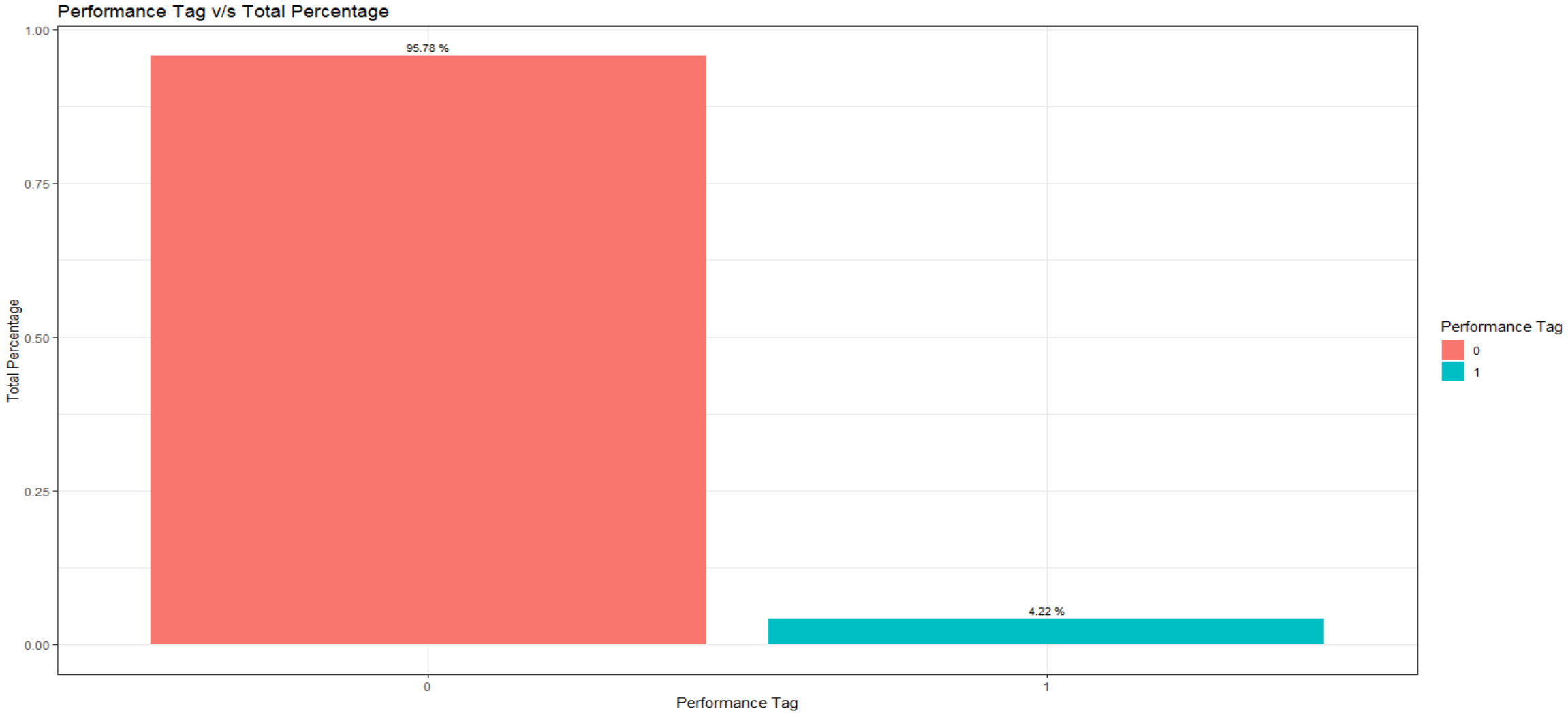
Datasets created

- Around 4 datasets were extracted from the merged and demographic datasets
- Each of these datasets are meant for different purposes such as for different modelling techniques and to meet other project objectives



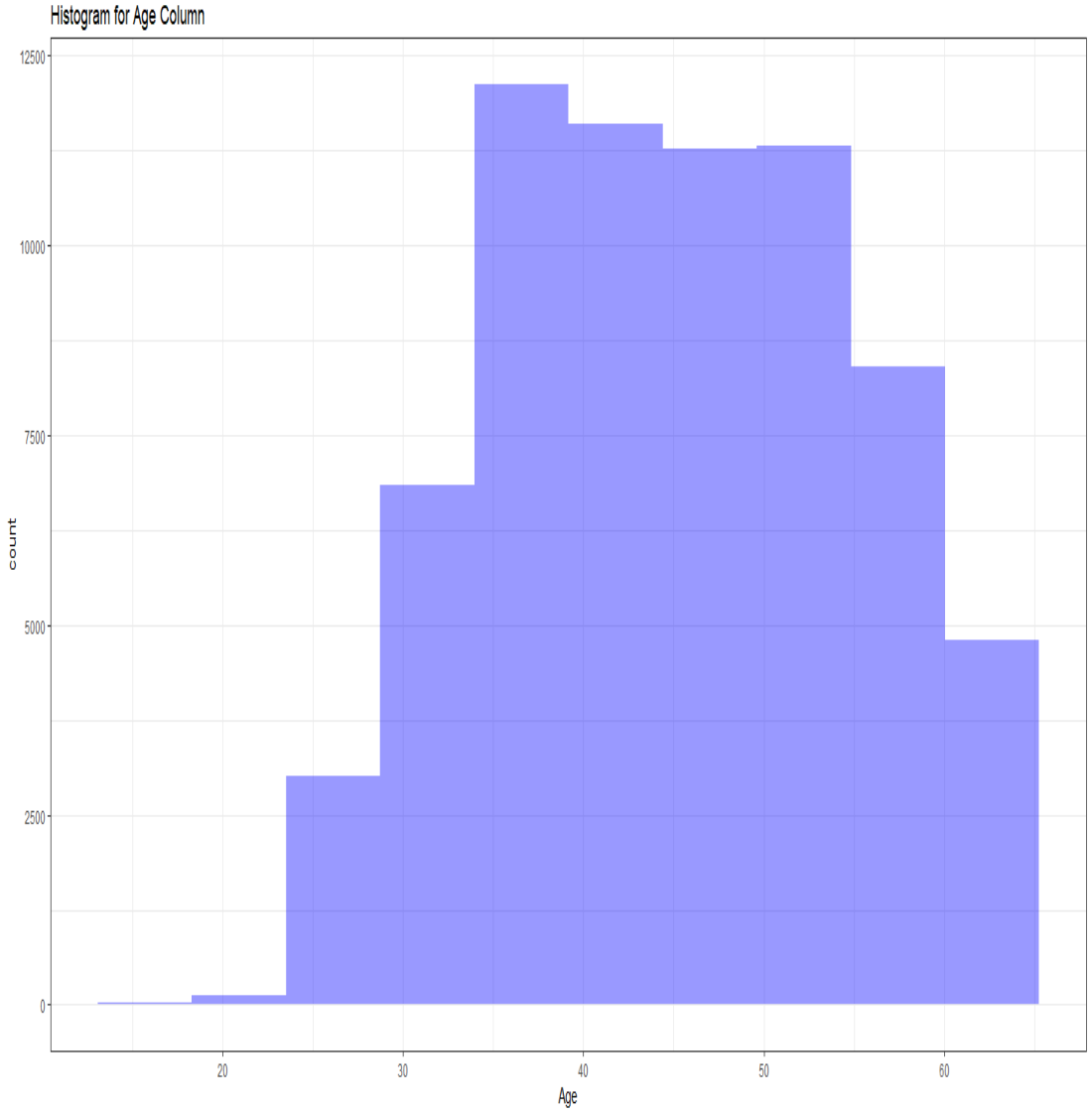
Exploratory Data Analysis(EDA)

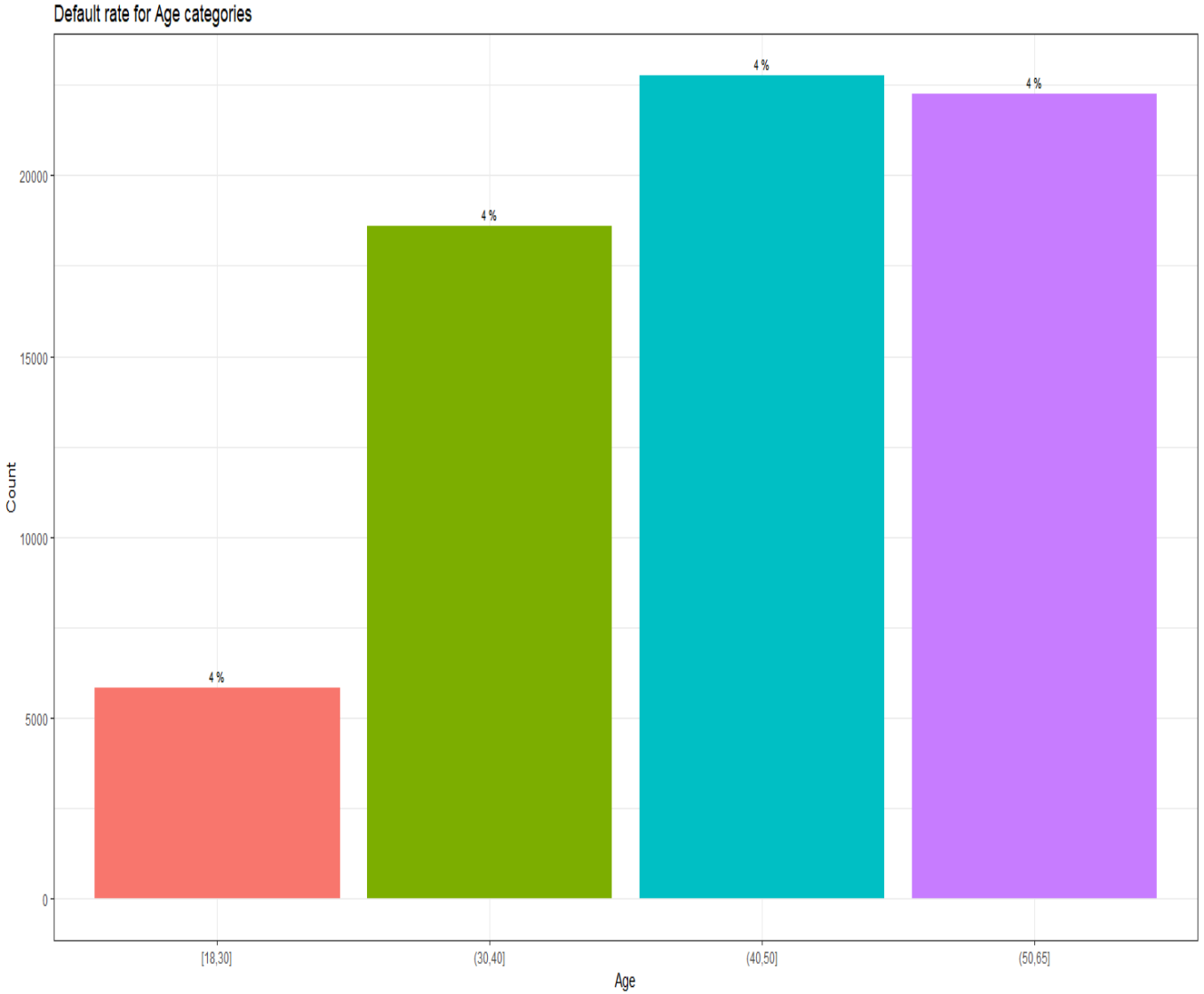
Analysis on the Target variable (Performance Tag)



- It reveals that the event rate is only 4.13% which indicates that the dataset is imbalanced

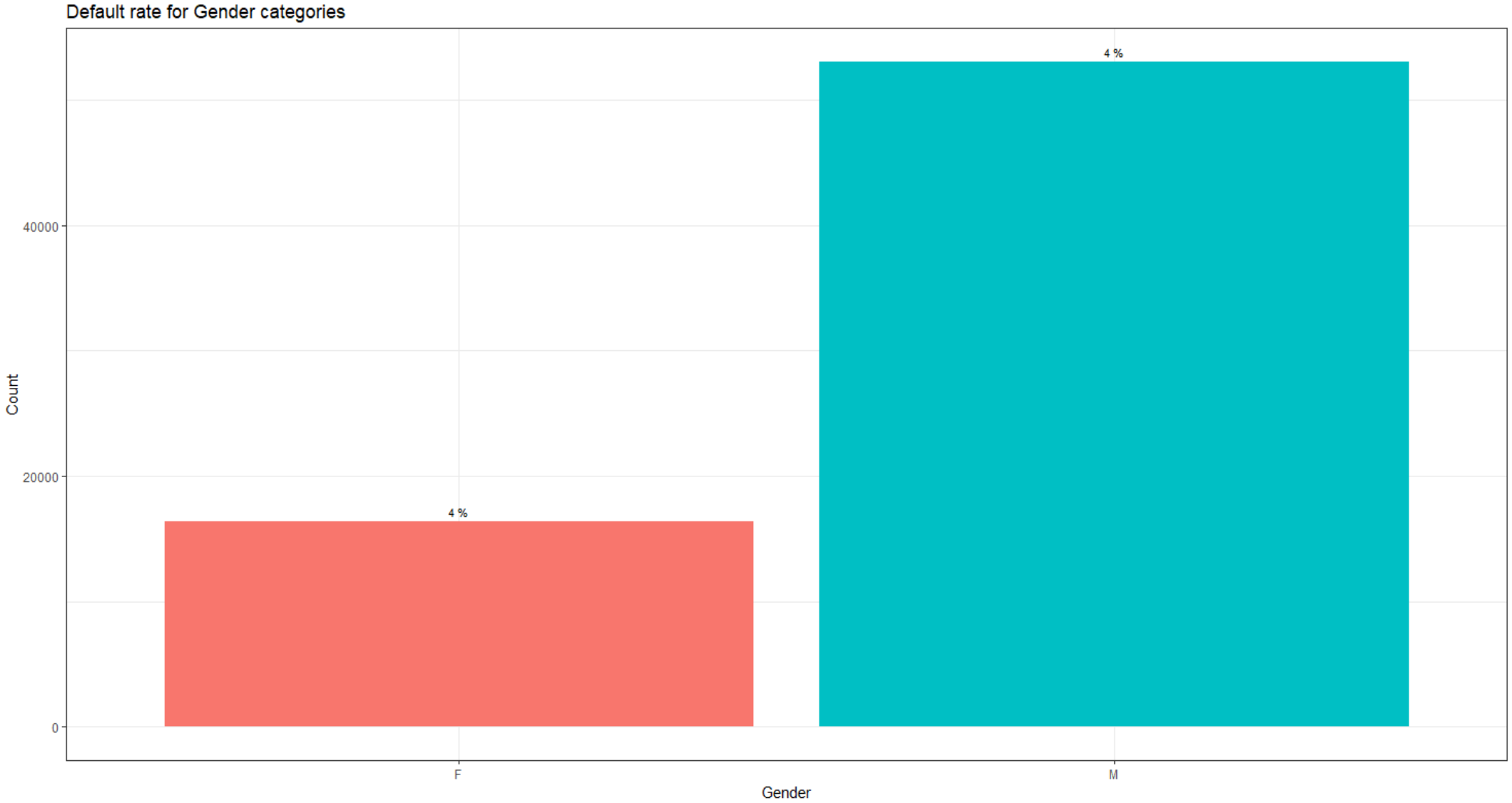
EDA - Variable ‘Age’





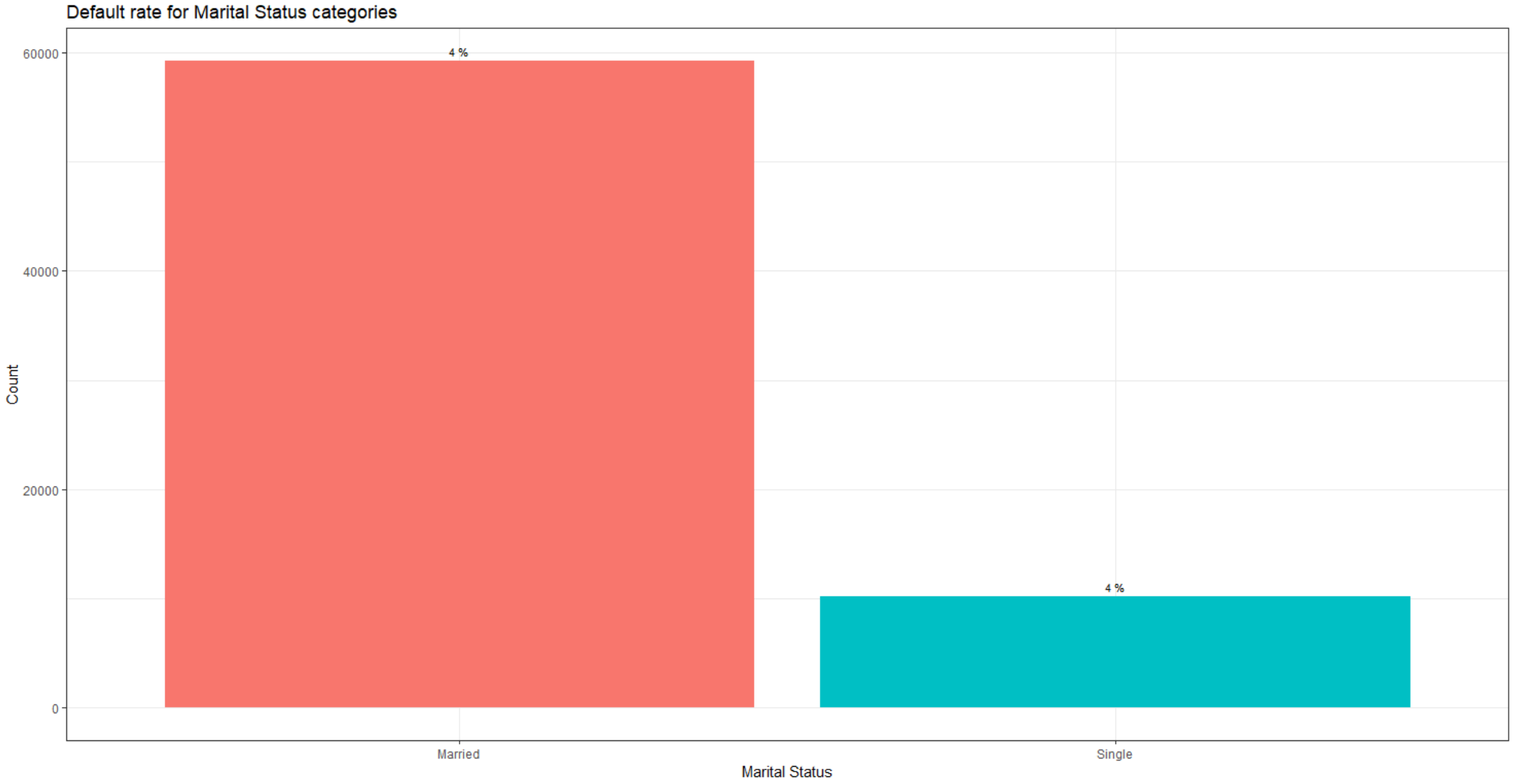
Age is not relevant variable for default and it is normally distributed

EDA - Variable ‘Gender’



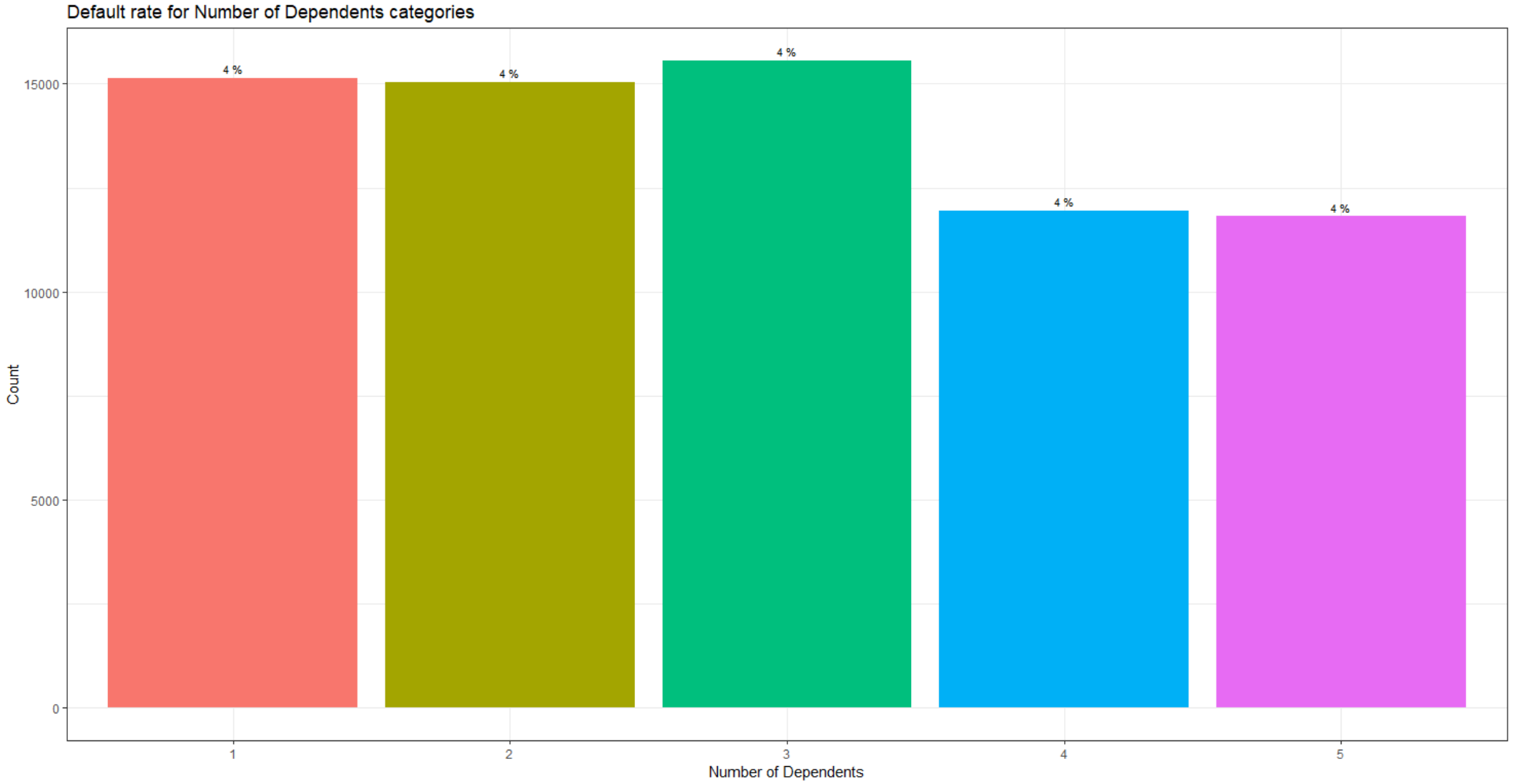
Gender is not relevant variable for default

EDA - Variable ‘Marital Status’



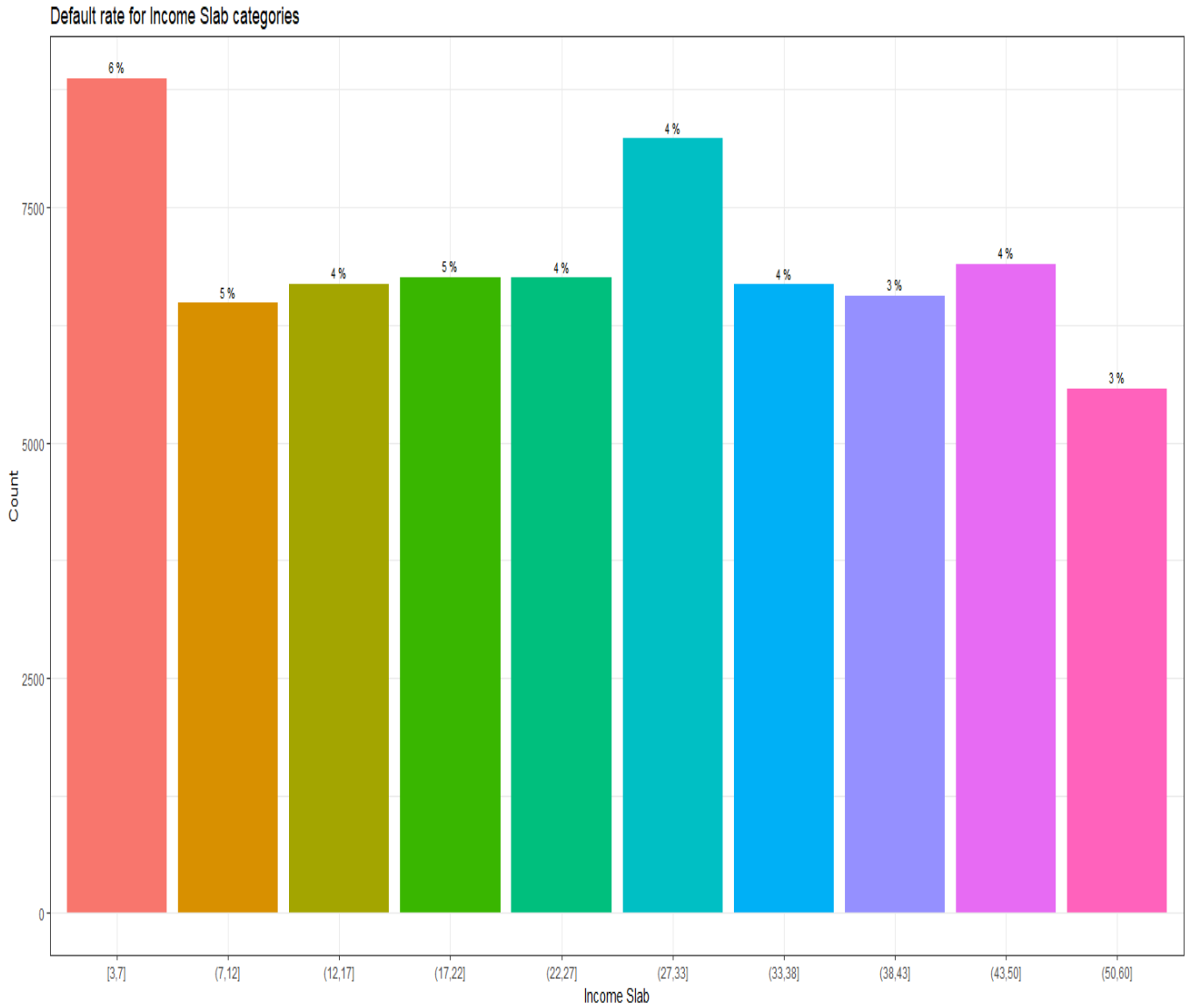
Marital Status is not relevant variable for default

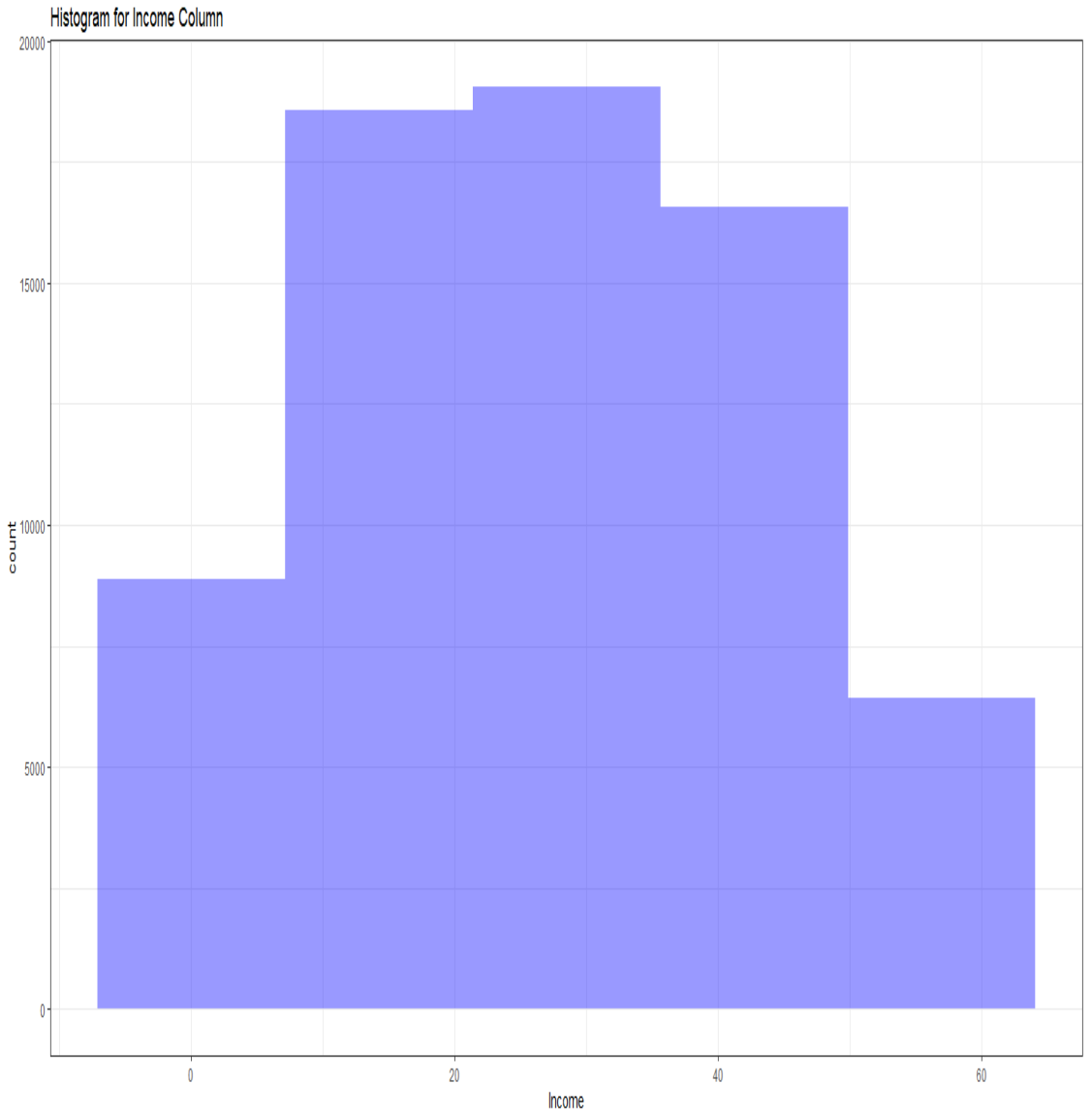
EDA - Variable ‘No. Of Dependents’



No. of Dependents is not relevant variable for default

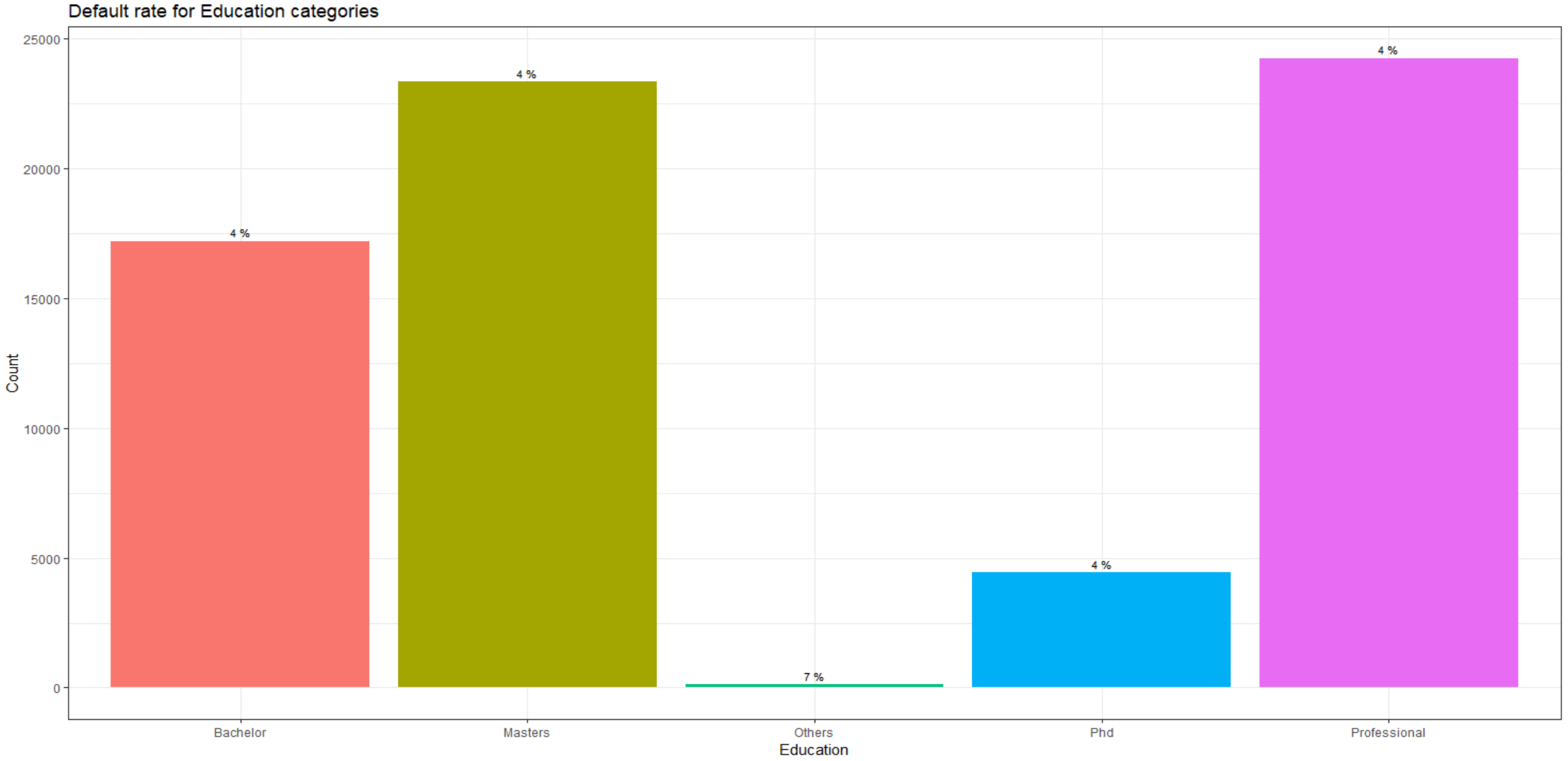
EDA - Variable ‘Income’





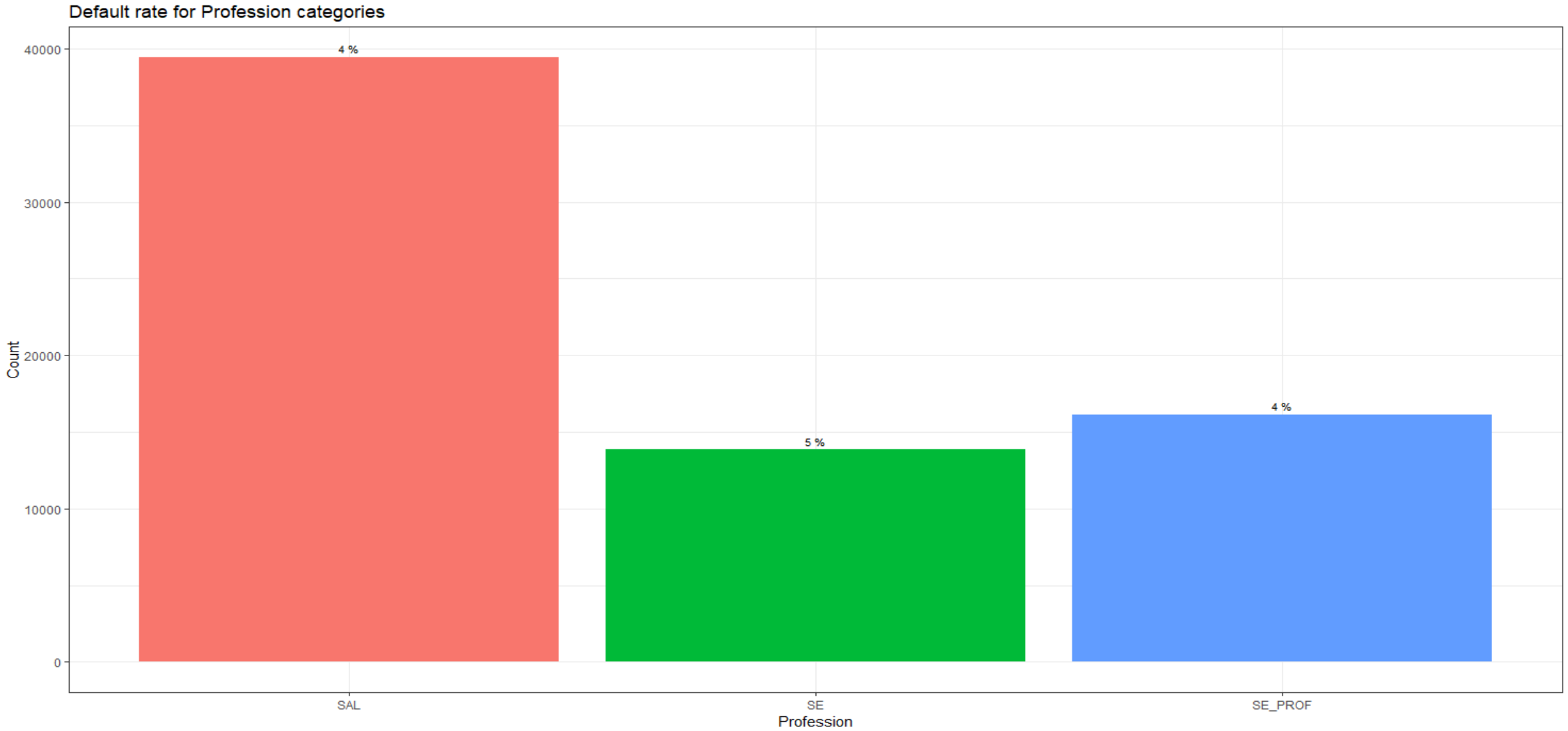
Income is not relevant variable for default and it is normally distributed

EDA - Variable ‘Education’



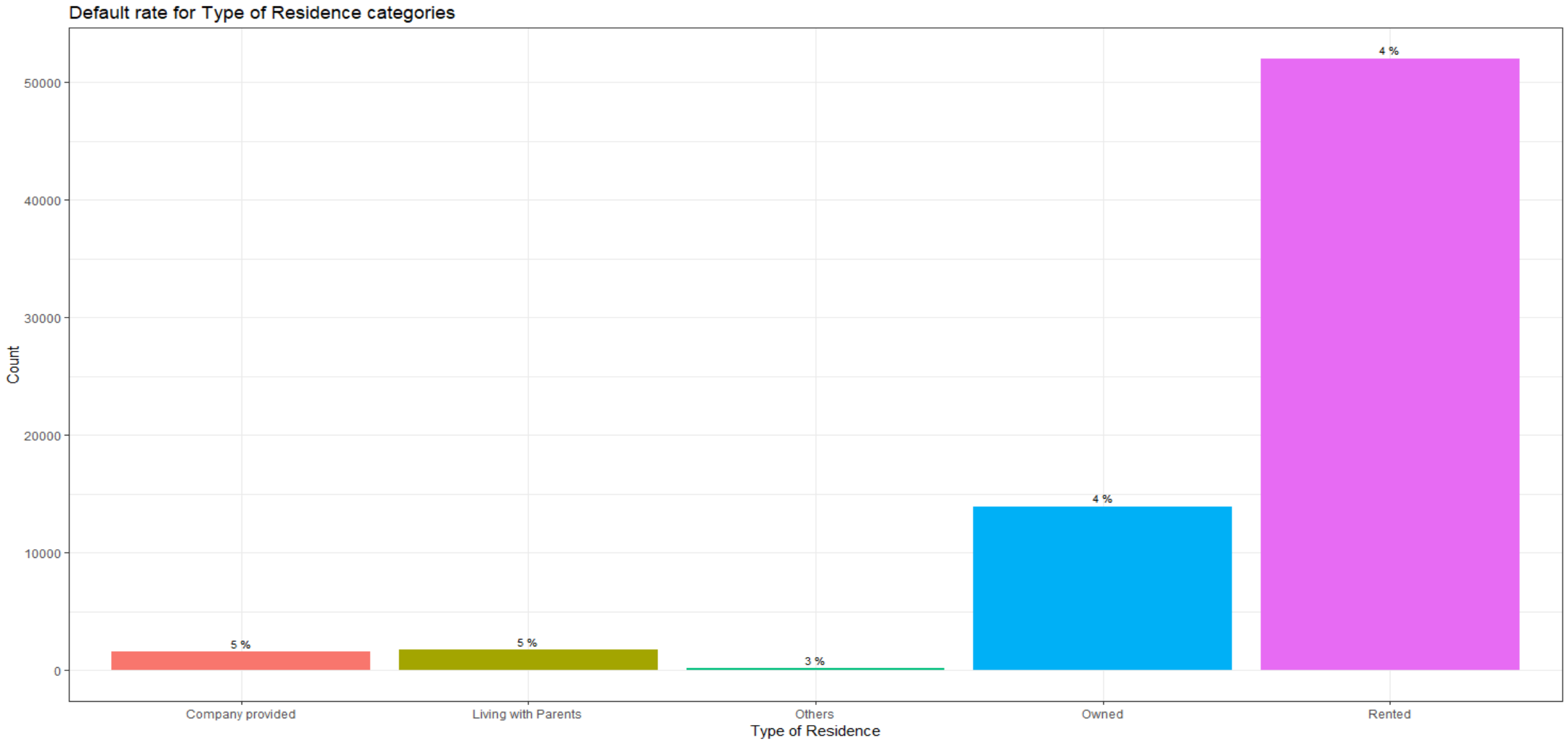
Education is not relevant variable for default

EDA - Variable ‘Profession’



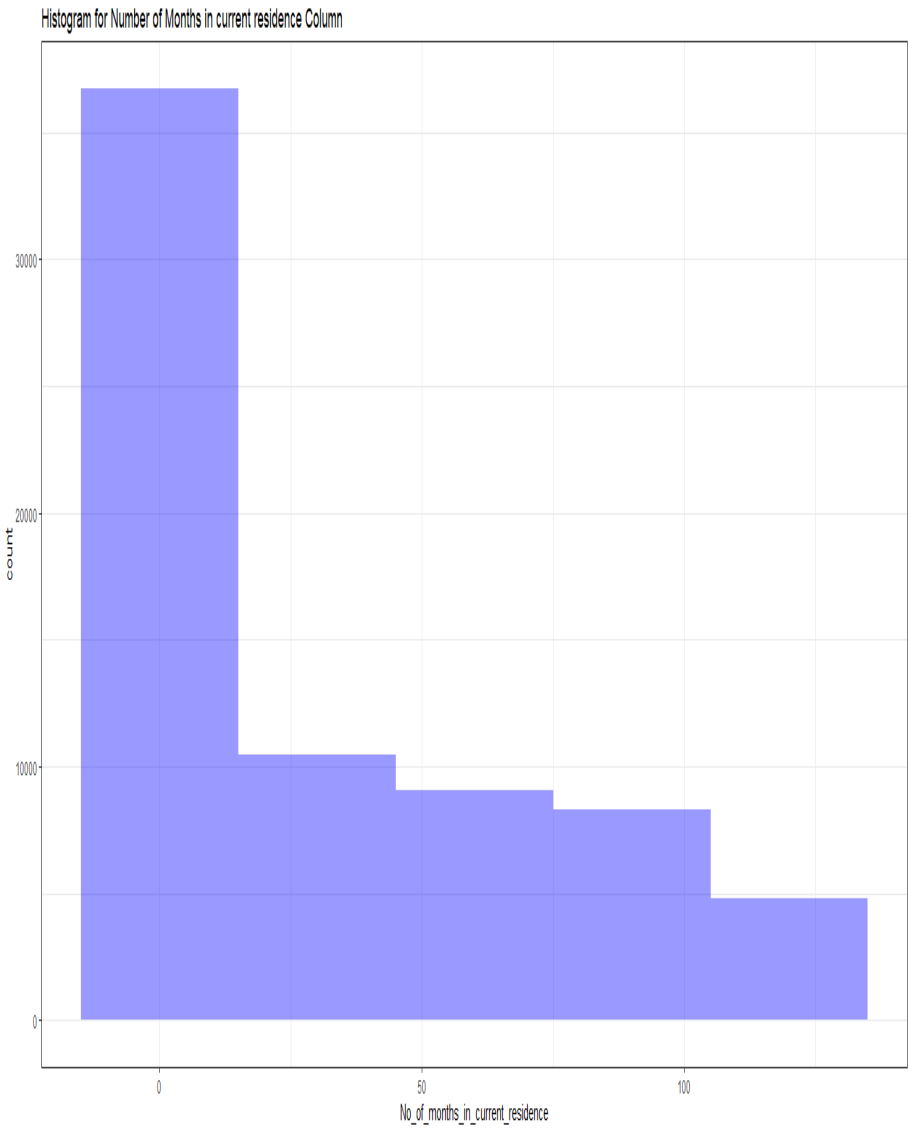
Profession is not relevant variable for default

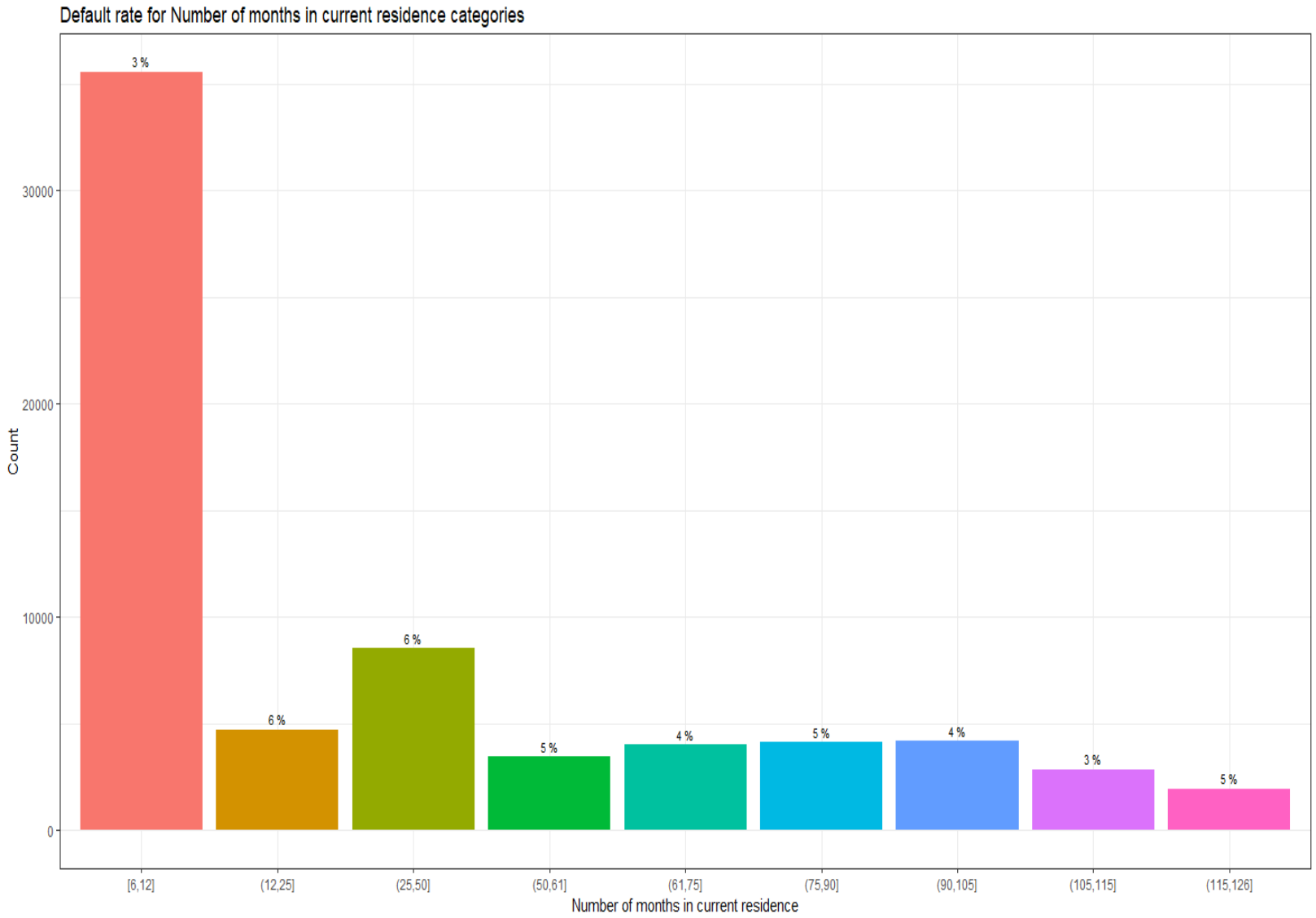
EDA - Variable ‘Type of Residence’



Type of Residence is not relevant variable for default

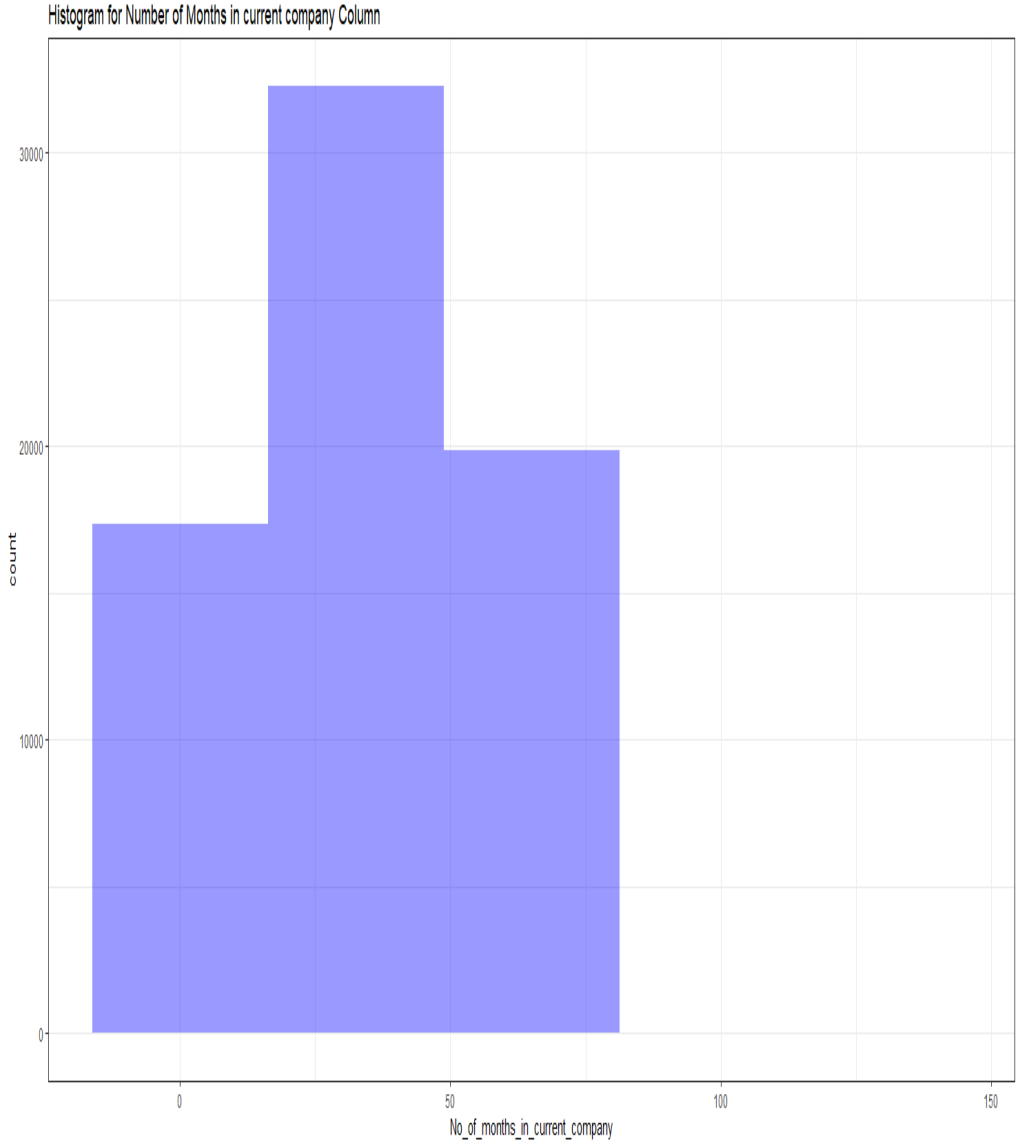
EDA - Variable ‘No of months in current residence’

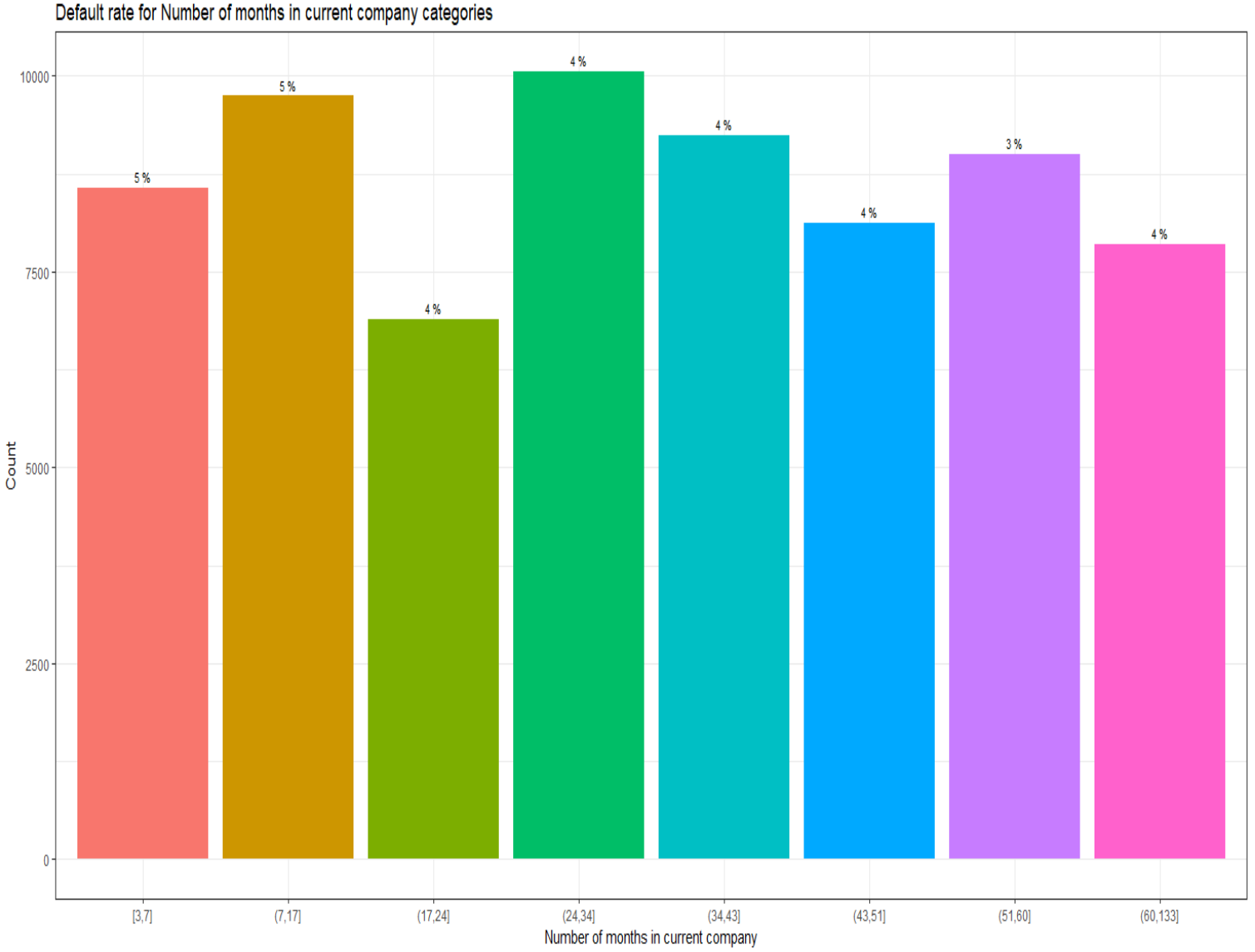




No. of months in current residence is not relevant variable for default and it is not normally distributed

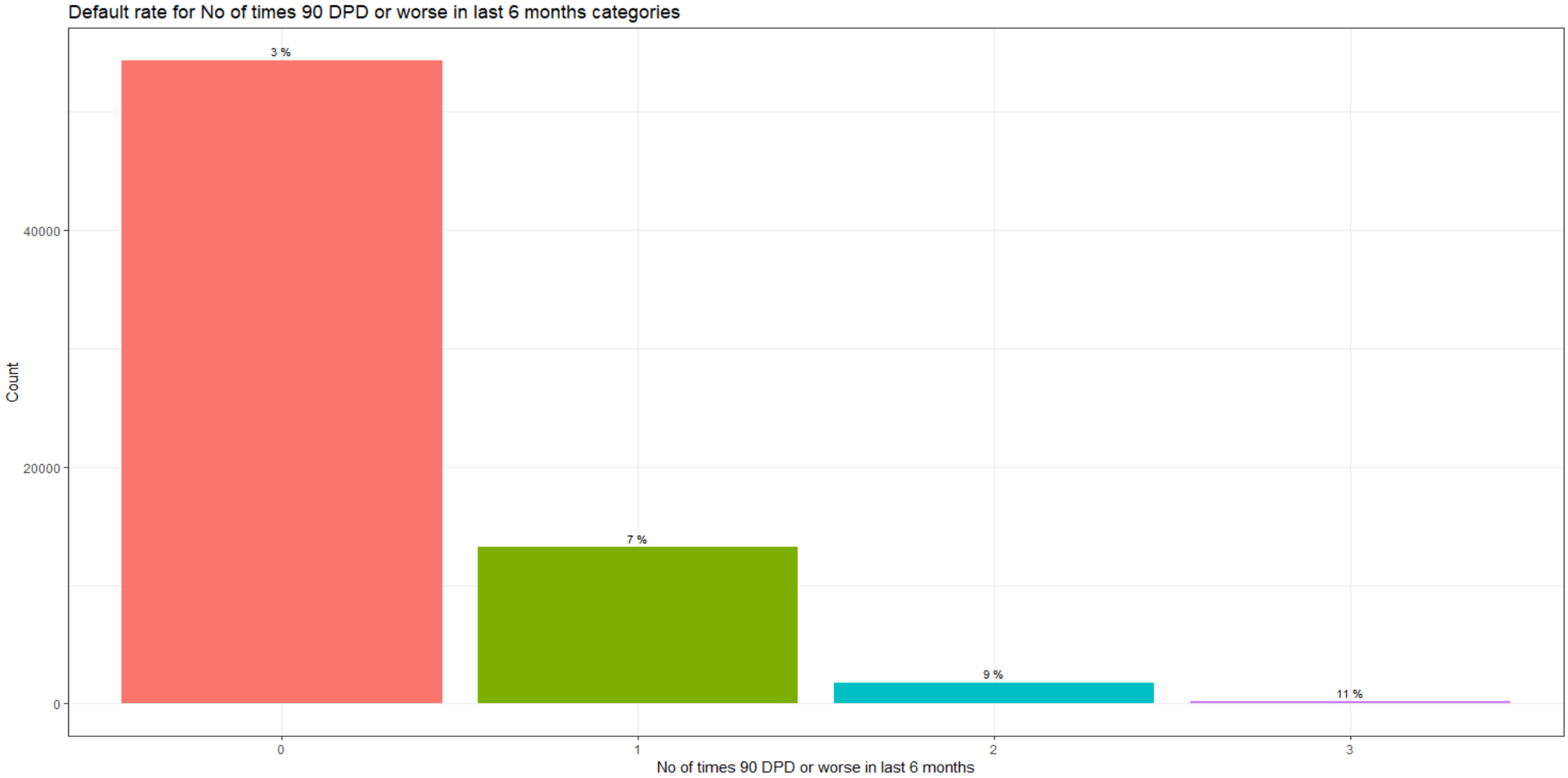
EDA - Variable ‘No. Of months in current company’





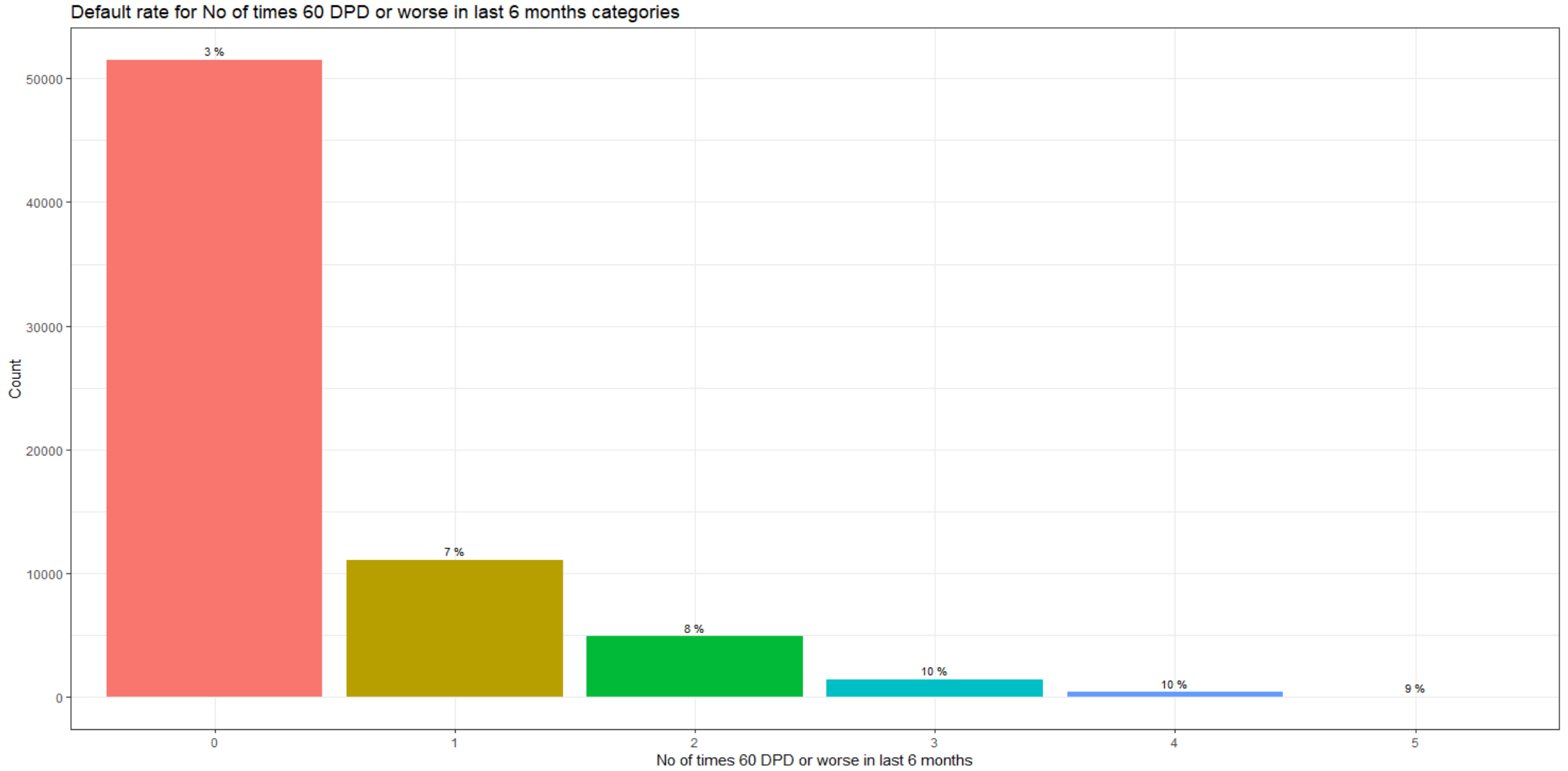
No. of months in current company is not relevant variable for default and it is normally distributed

EDA - Variable 'No_of_times_90_DPD_or_worse_in_last_6_months'



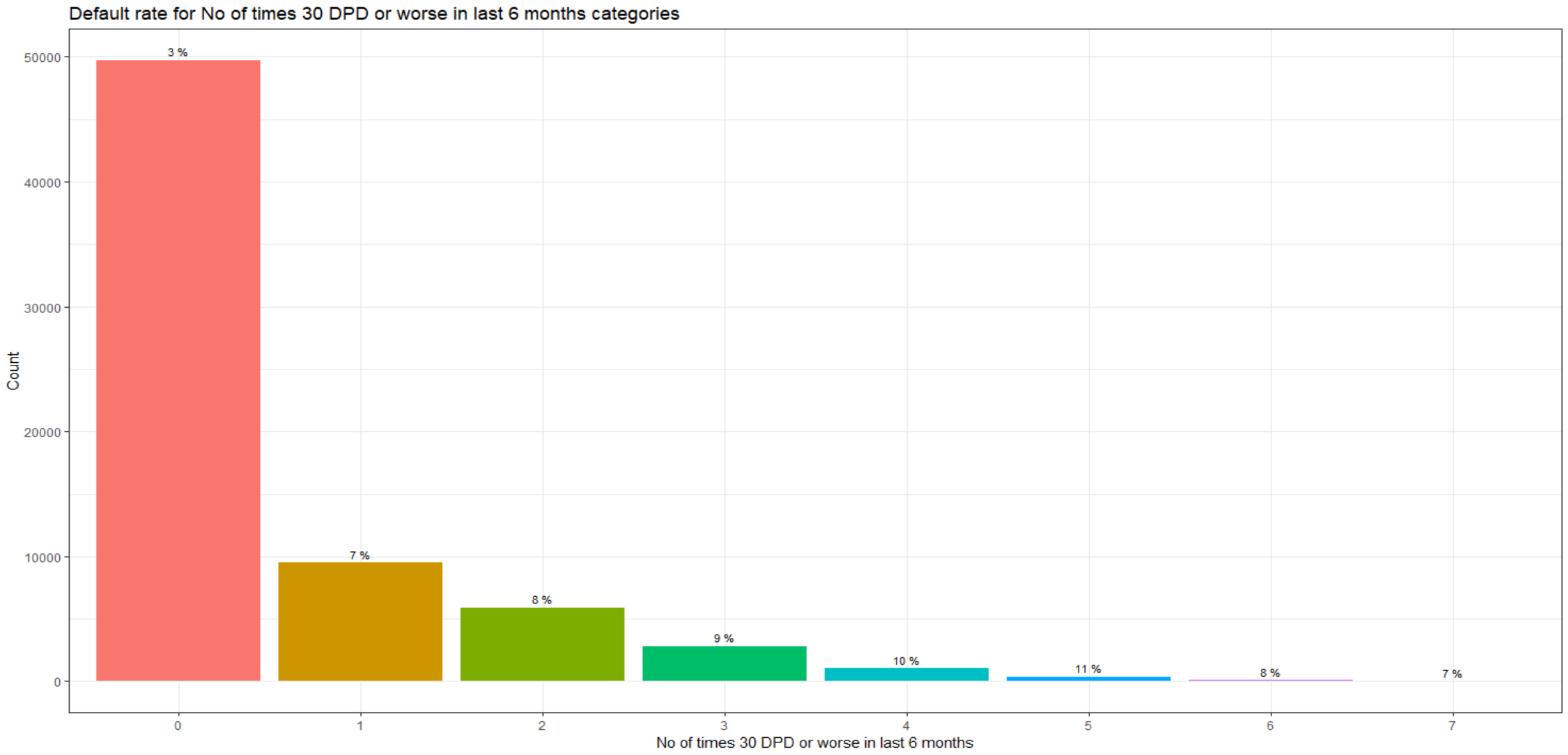
No. of times 90 DPD or worse in last 6 months is relevant variable for default

EDA - Variable 'No_of_times_60_DPD_or_worse_in_last_6_months'



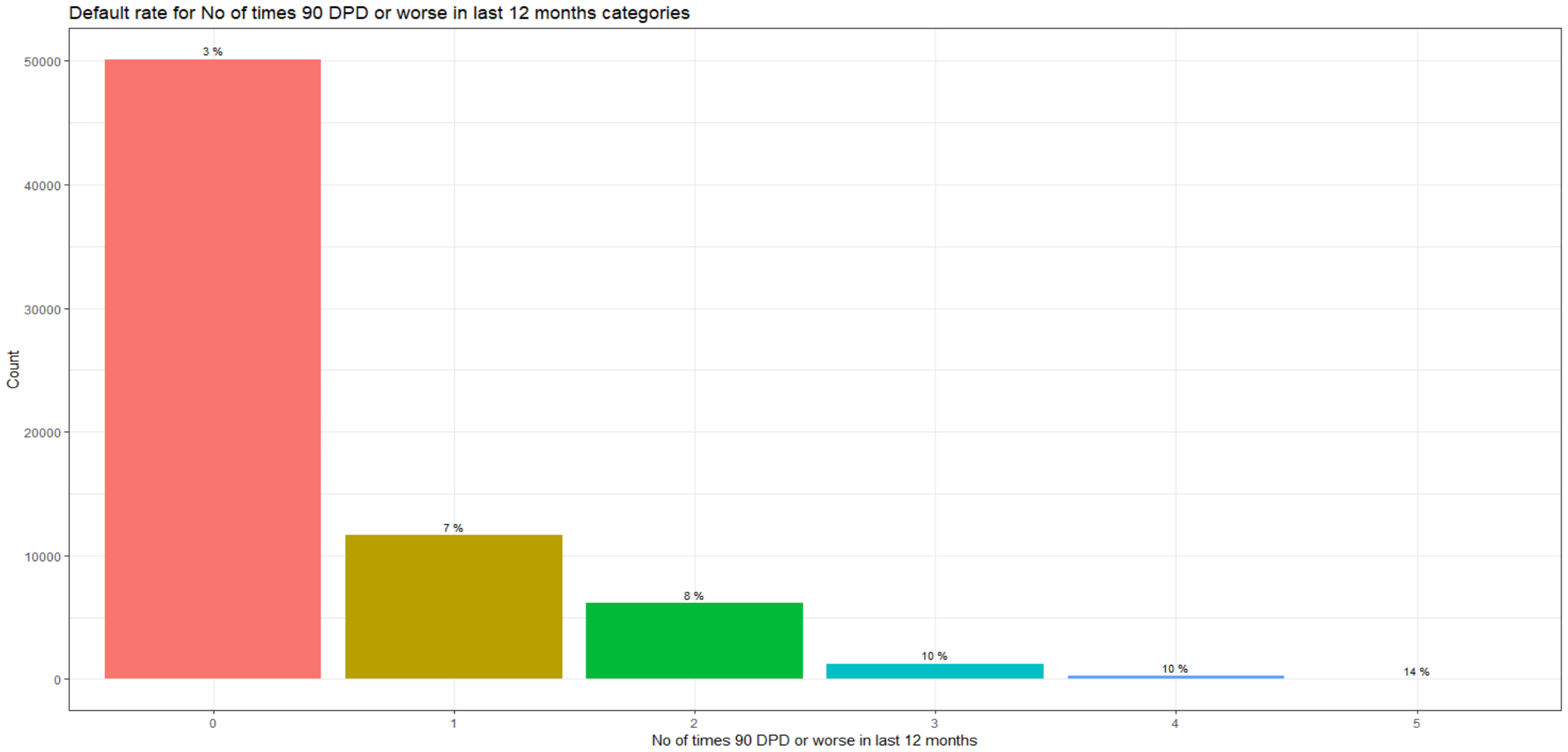
No. of times 60 DPD or worse in last 6 months is relevant variable for default

EDA - Variable 'No_of_times_30_DPD_or_worse_in_last_6_months'



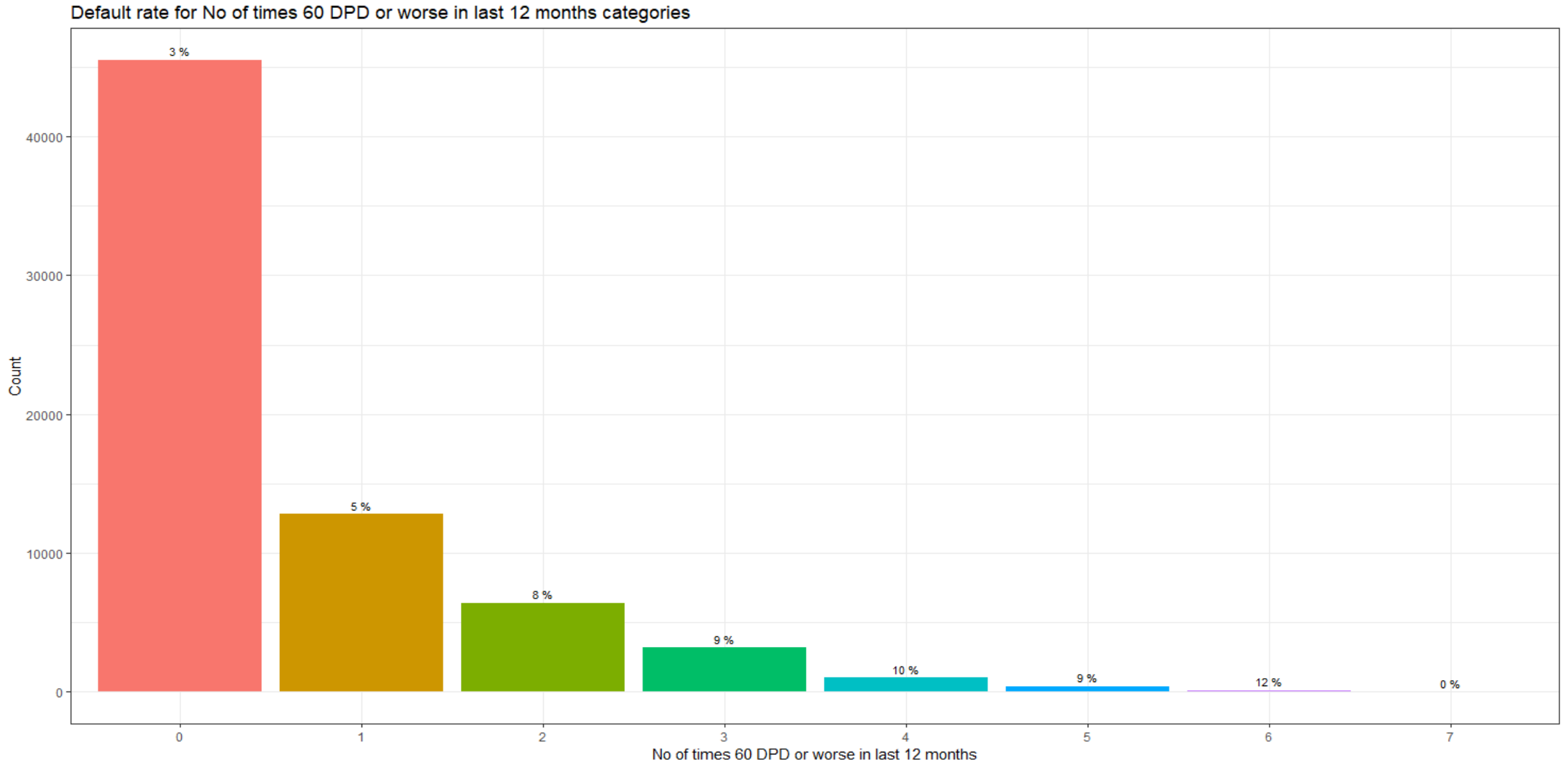
No. of times 30 DPD or worse in last 6 months is relevant variable for default

EDA - Variable 'No_of_times_90_DPD_or_worse_in_last_12_months'



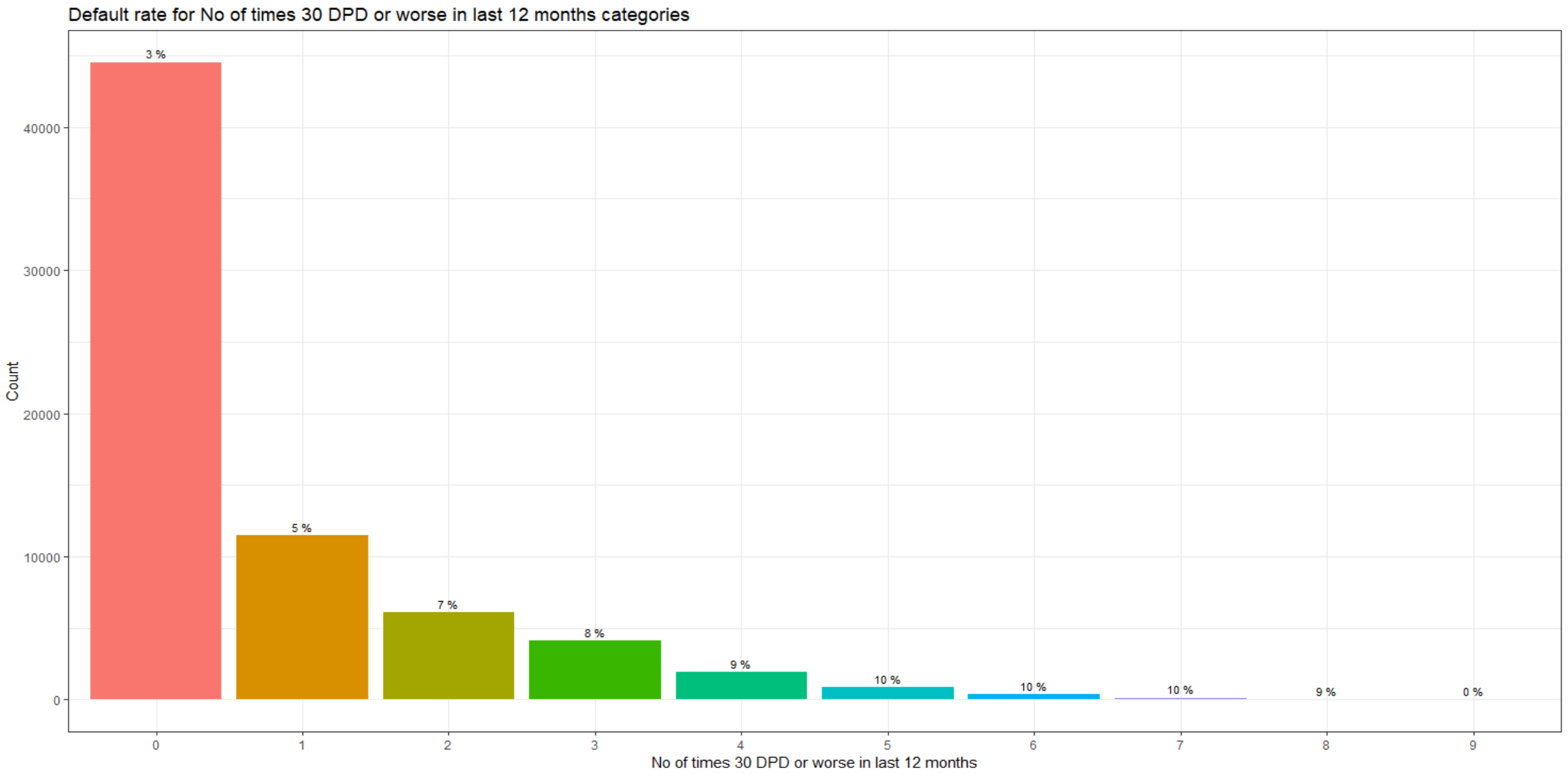
No. of times 90 DPD or worse in last 12 months is relevant variable for default

EDA - Variable 'No_of_times_60_DPD_or_worse_in_last_12_months'



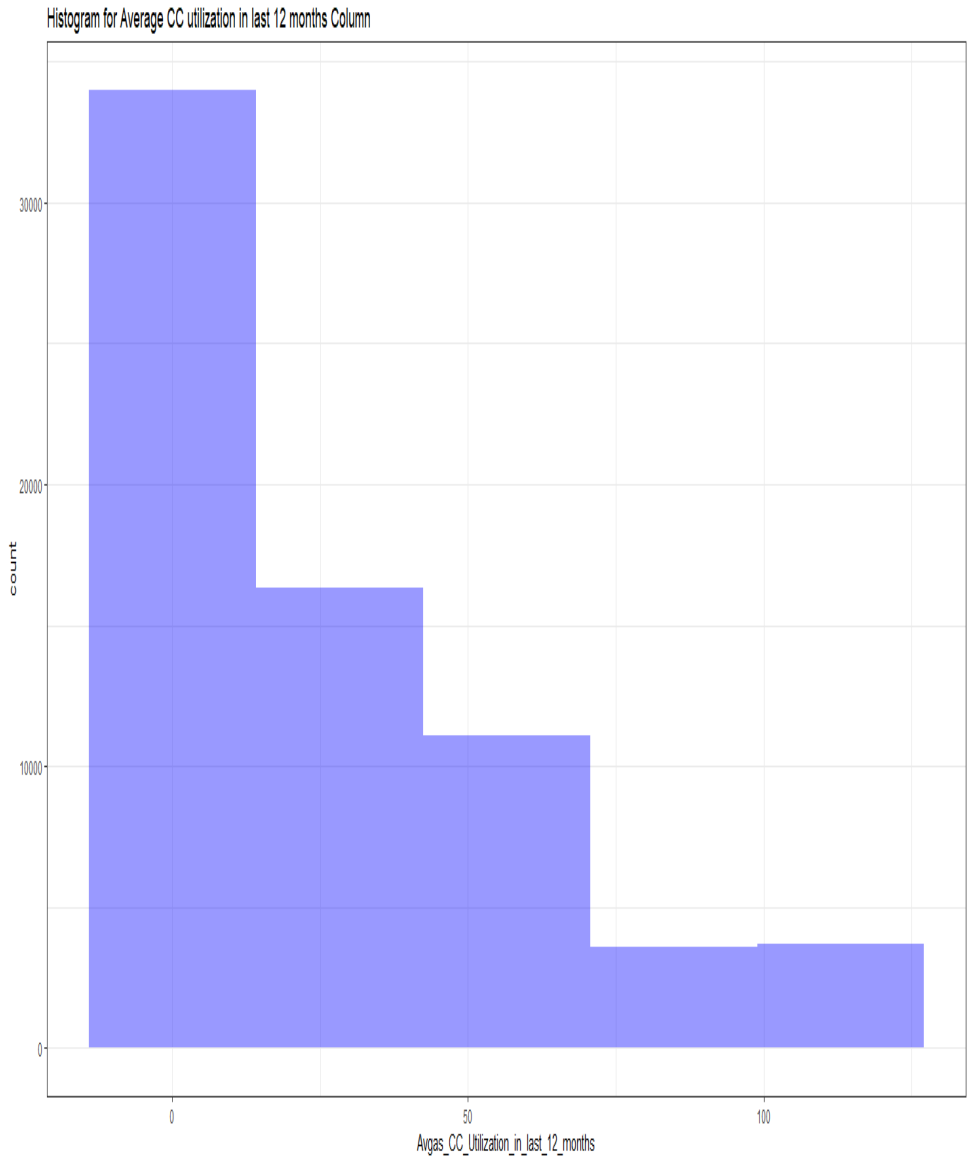
No. of times 60 DPD or worse in last 12 months is relevant variable for default

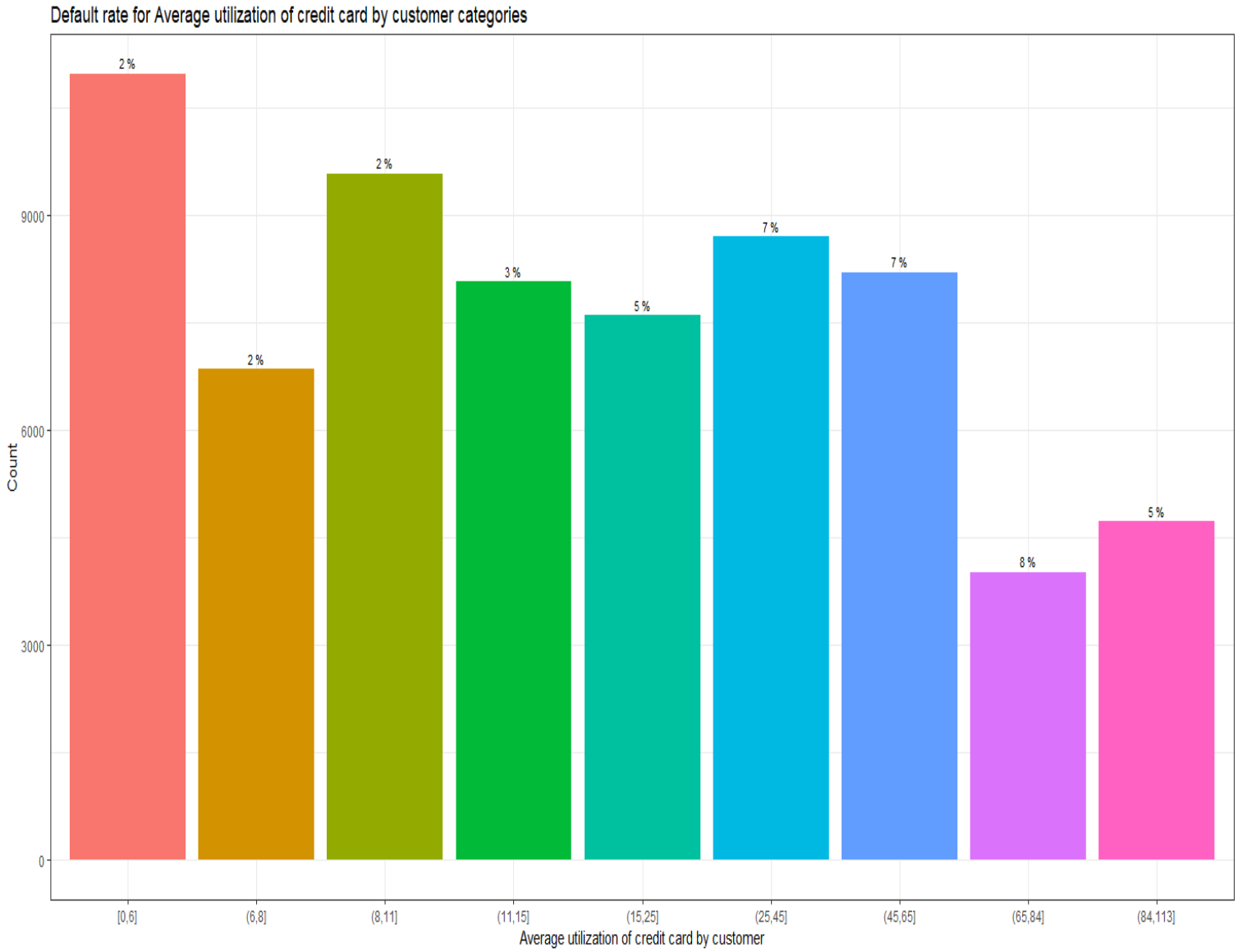
EDA - Variable 'No_of_times_30_DPD_or_worse_in_last_12_months'



No. of times 30 DPD or worse in last 12 months is relevant variable for default

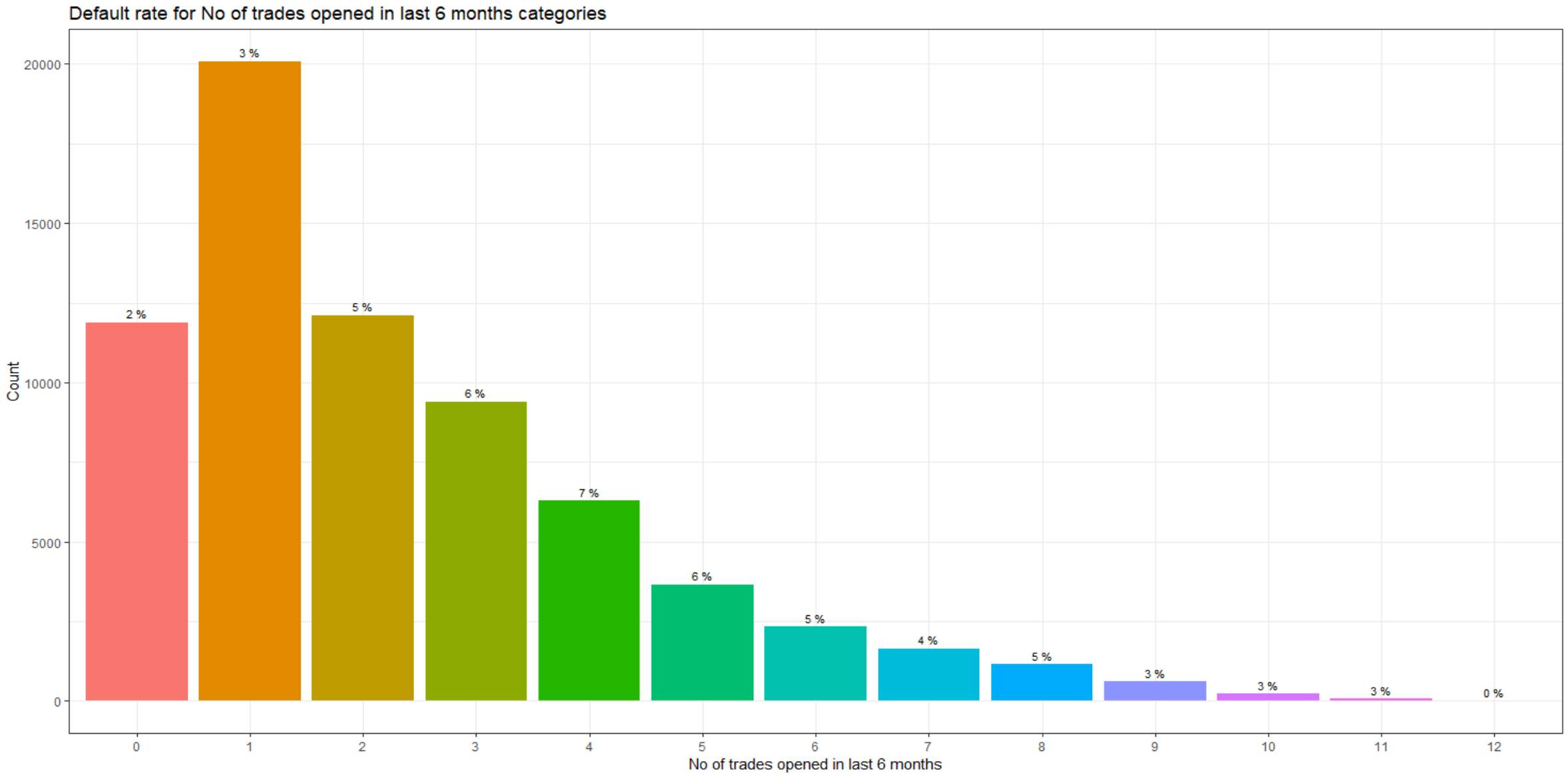
EDA - Variable 'Average CC Utilization in last 12 months'





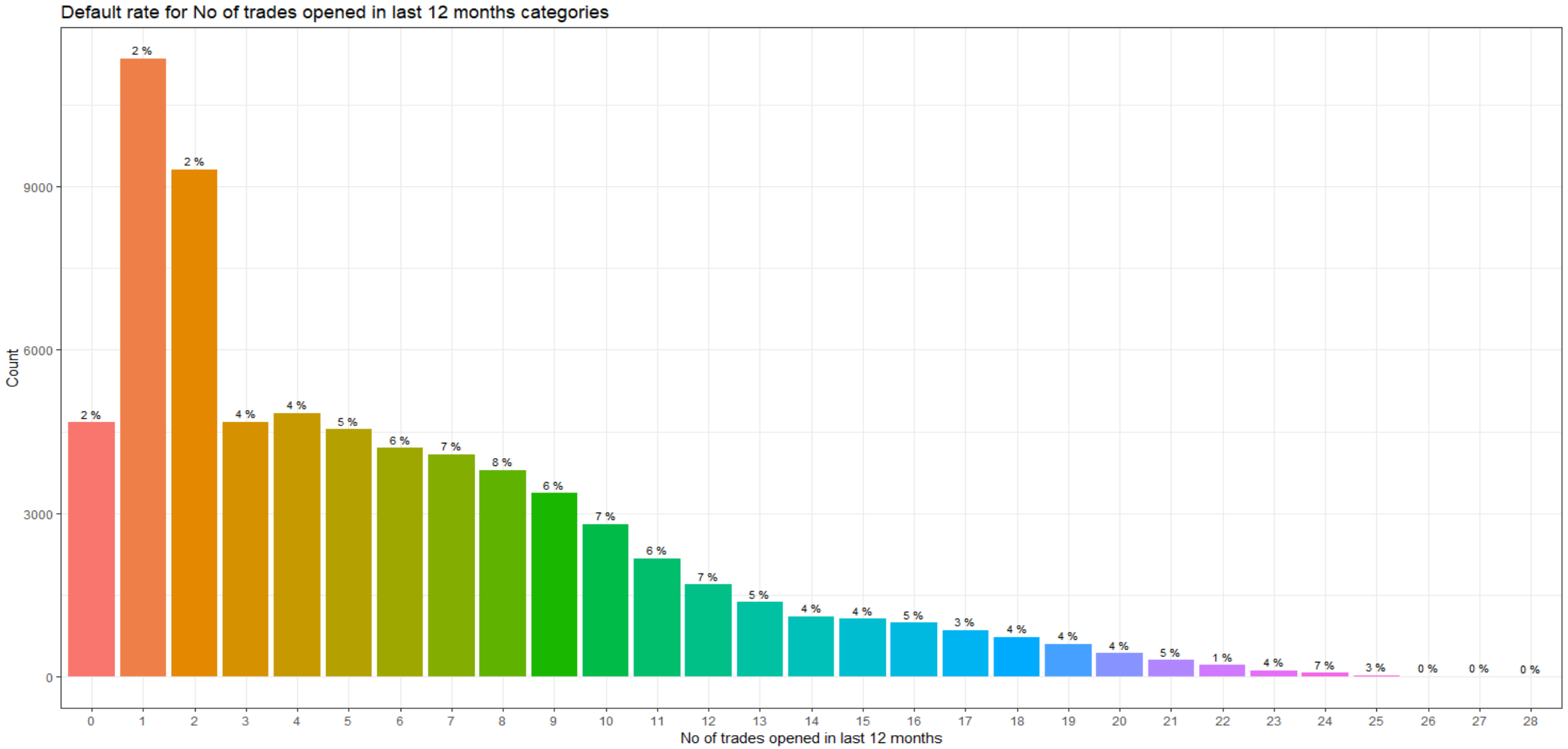
Average CC utilization in last 12 months is relevant variable for default and it is not normally distributed

EDA - Variable ‘No_of_trades_opened_in_last_6_months’



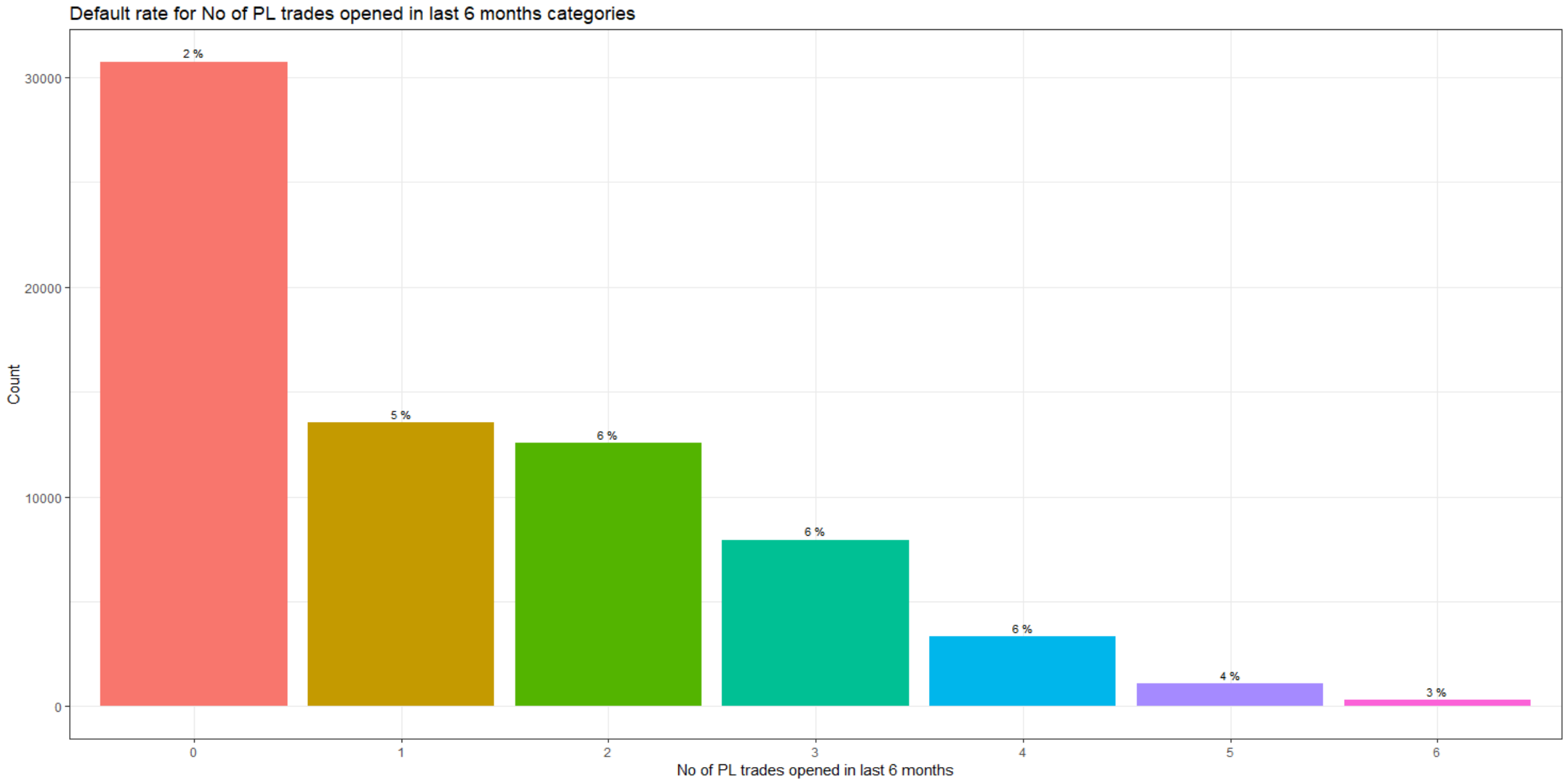
Number of trades opened in last 6 months is relevant variable for default

EDA - Variable 'No_of_trades_opened_in_last_12_months'



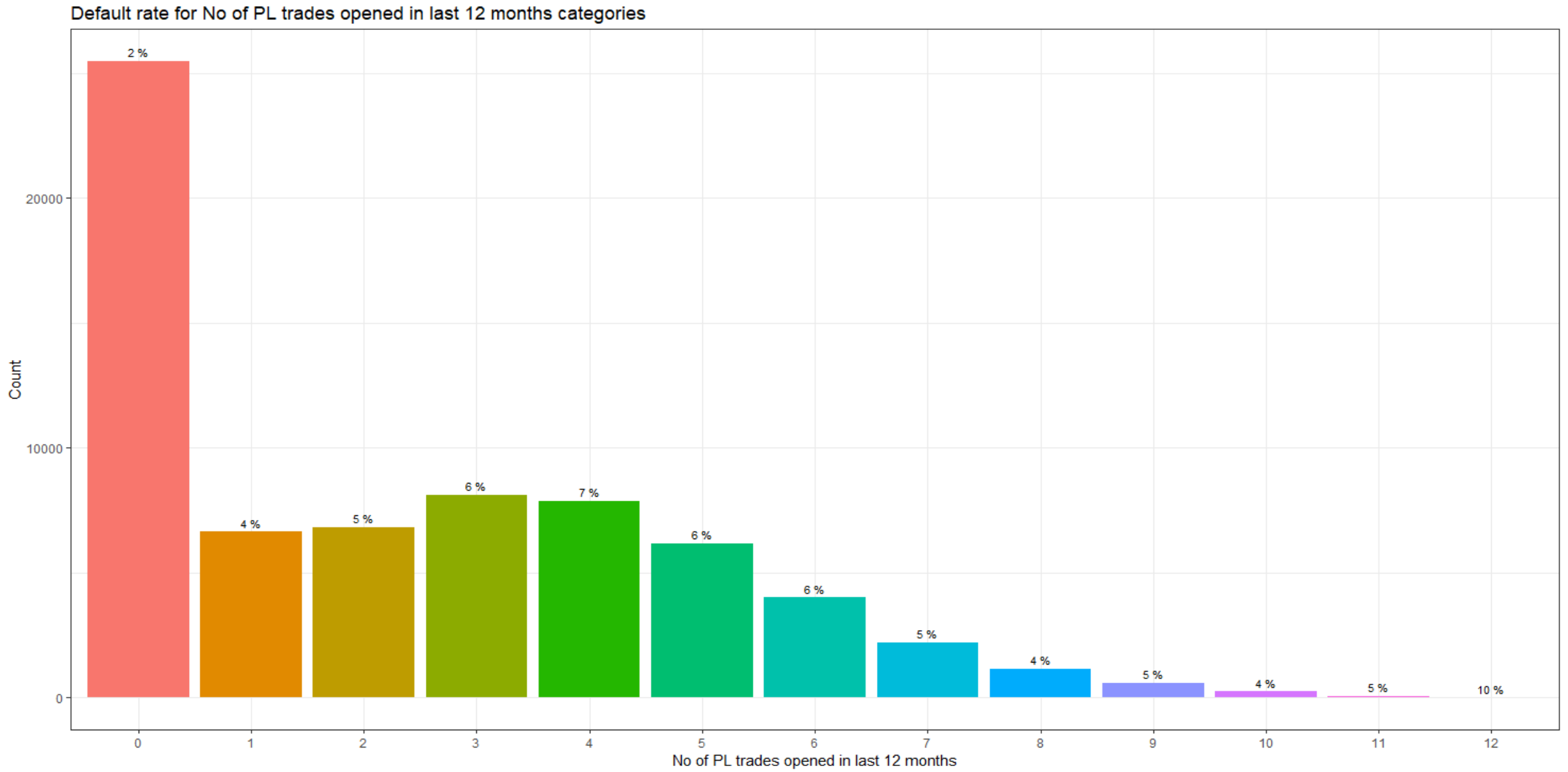
Number of trades opened in last 12 months is relevant variable for default

EDA - Variable 'No_of_PL_trades_opened_in_last_6_months'



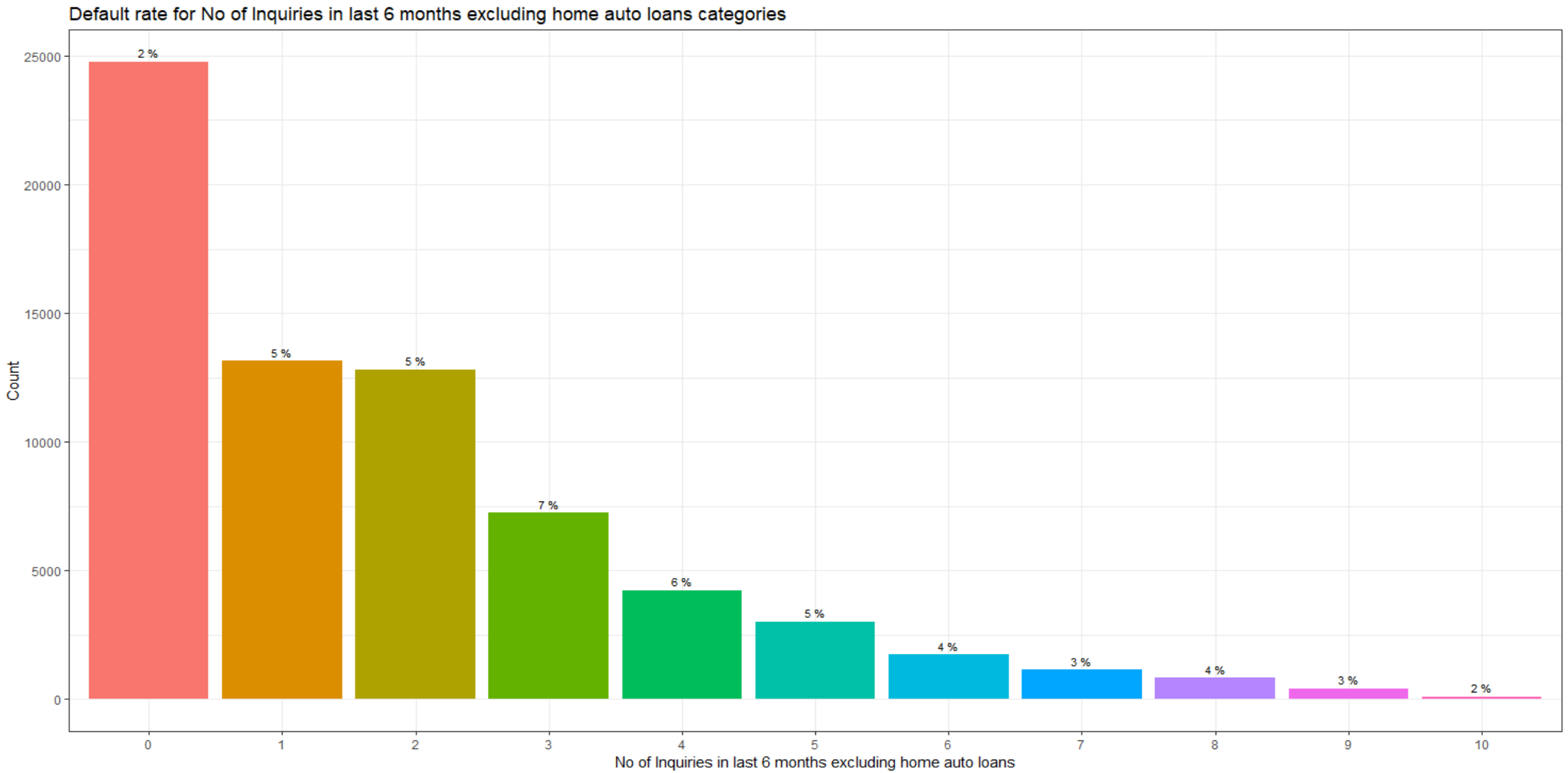
Number of PL trades opened in last 6 months is relevant variable for default

EDA - Variable 'No_of_PL_trades_opened_in_last_12_months'



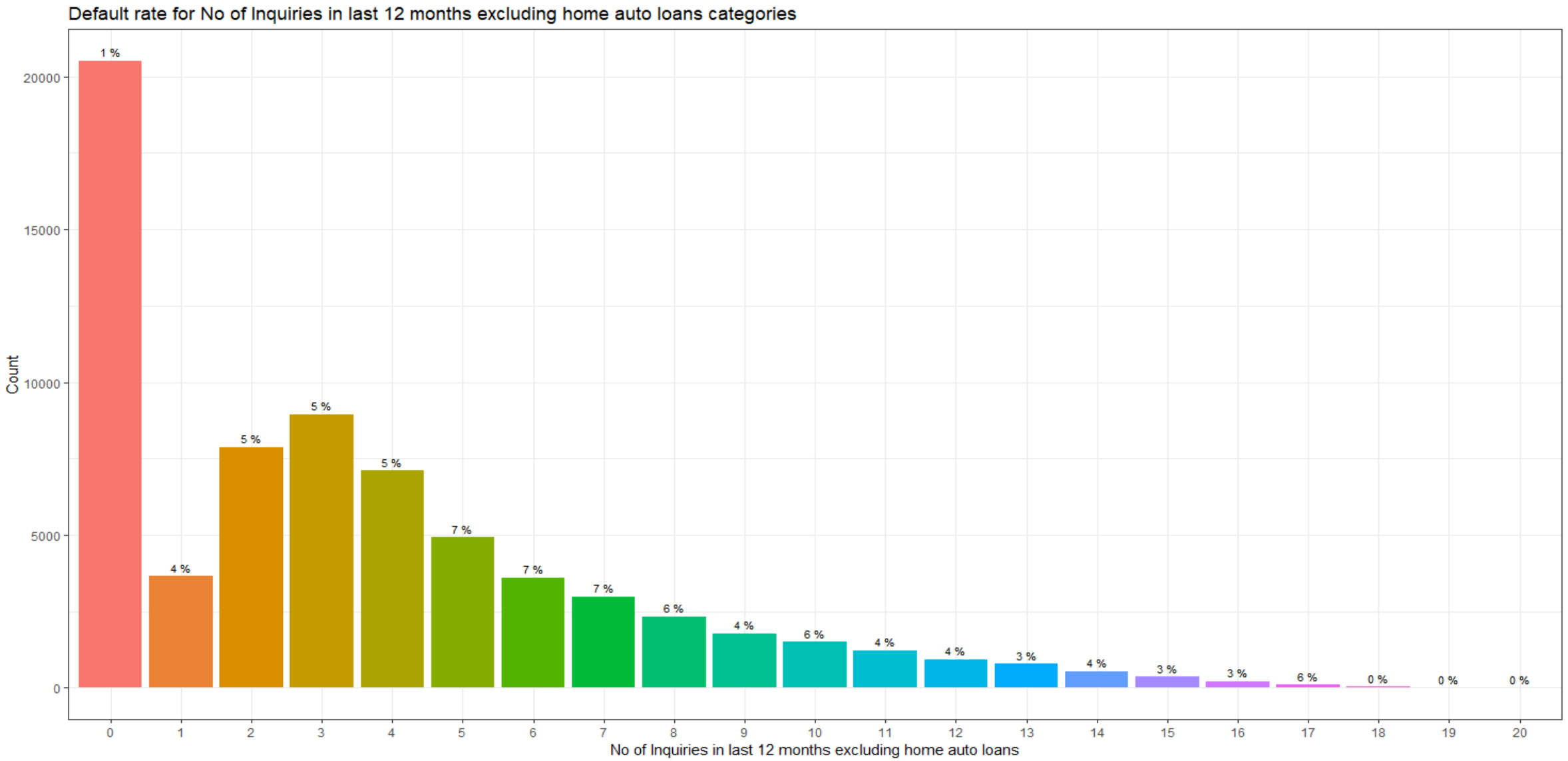
Number of PL trades opened in last 12 months is relevant variable for default

EDA – Variable 'No_of_Inquiries_in_last_6_months__excluding_home___auto_loans'



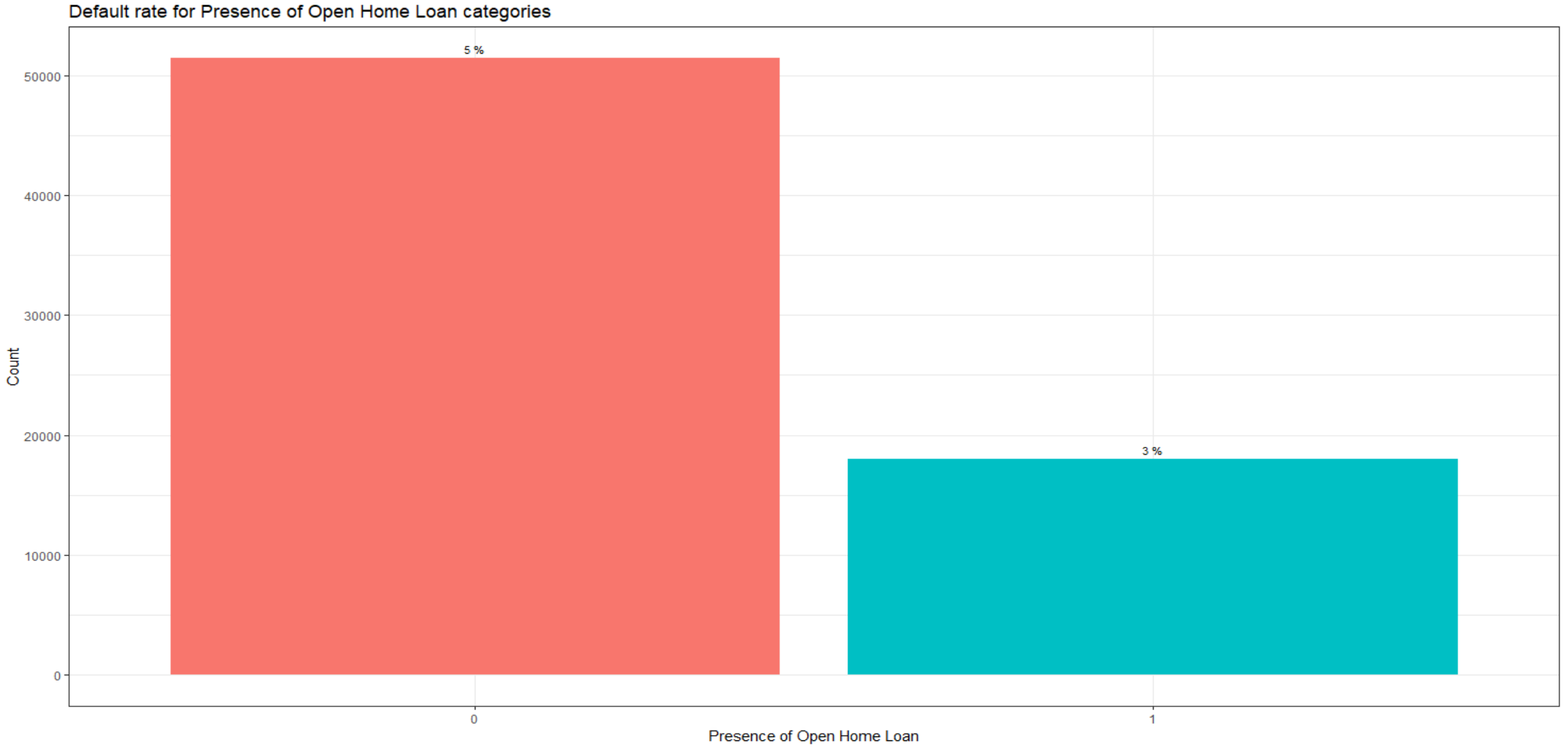
"Number of Inquiries in last 6 months excluding home auto loans" is relevant variable for default

EDA - Variable 'No_of_Inquiries_in_last_12_months__excluding_home___auto_loans'



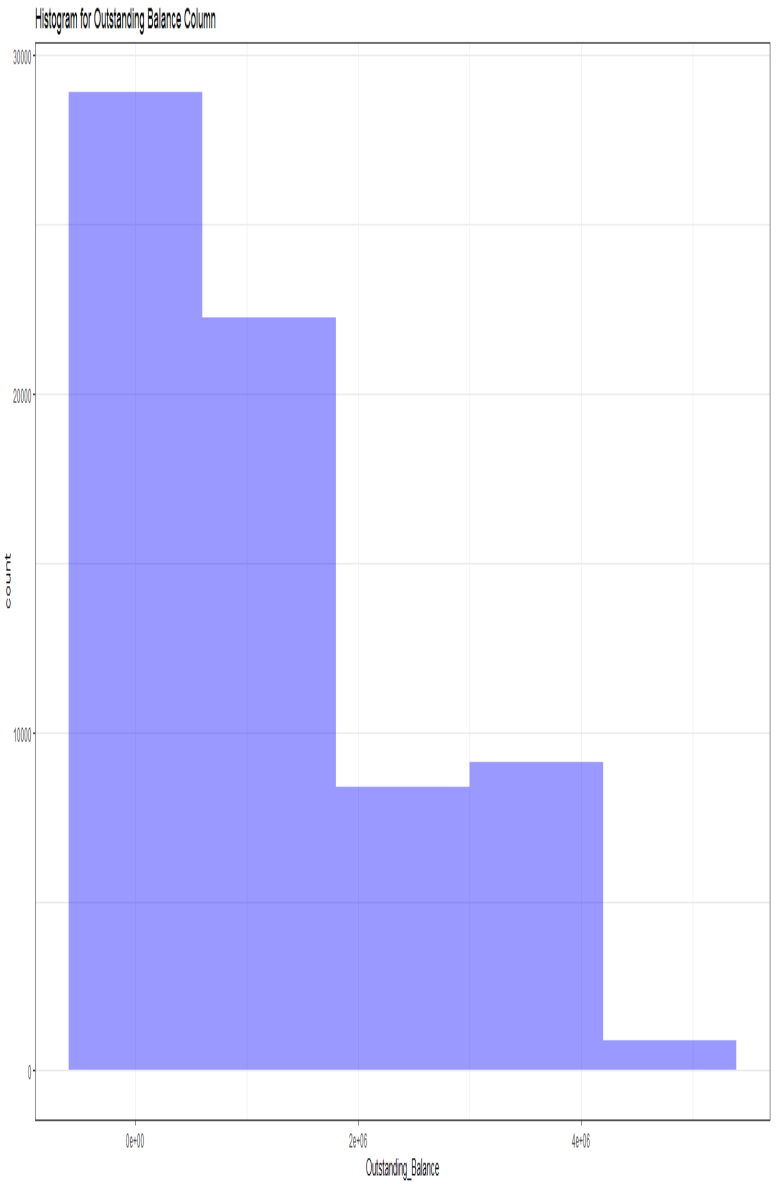
"Number of Inquiries in last 12 months excluding home auto loans" is relevant variable for default

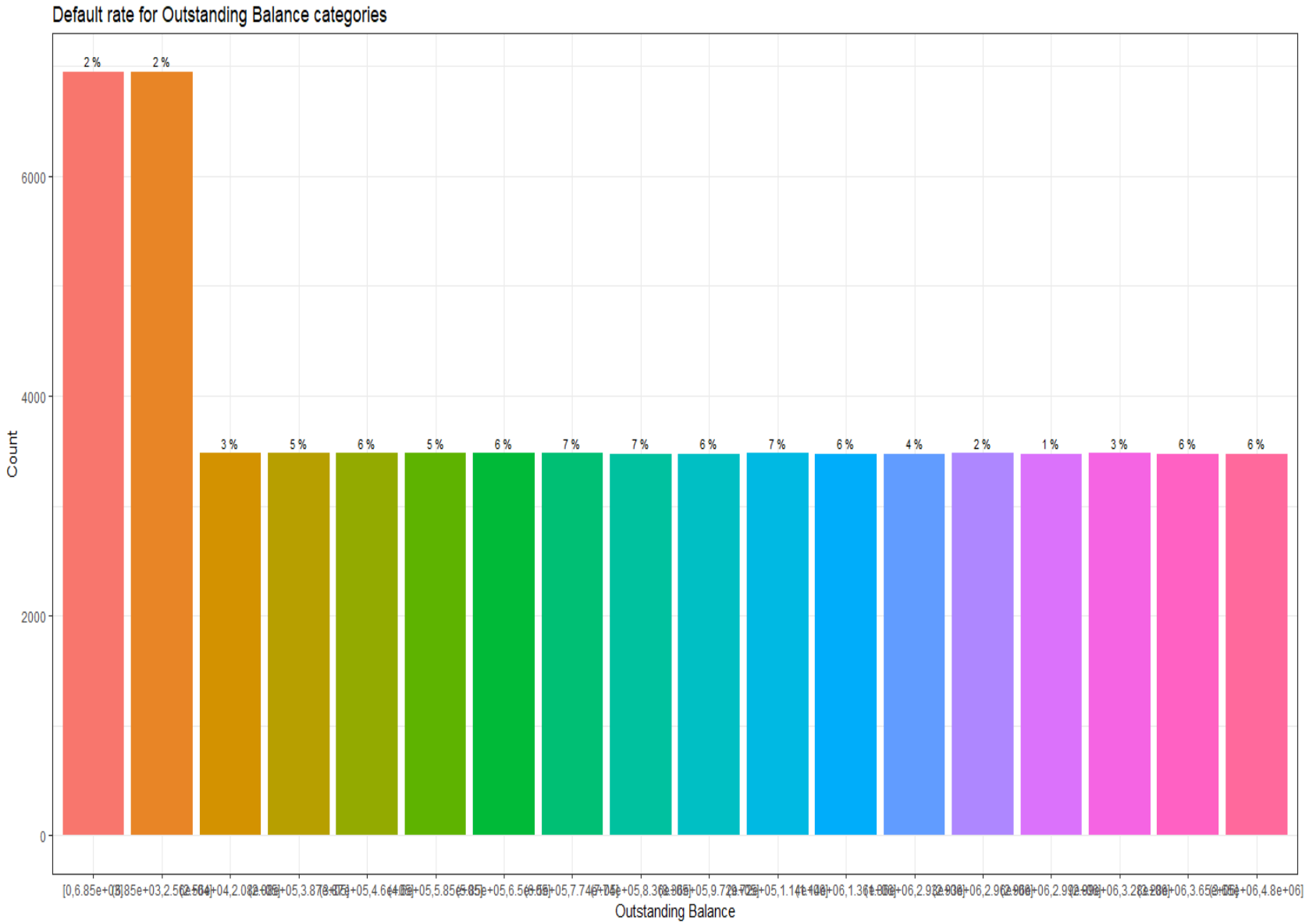
EDA - Variable ‘Presence_of_open_home_loan’



“Presence of open home loan” is not relevant variable for default

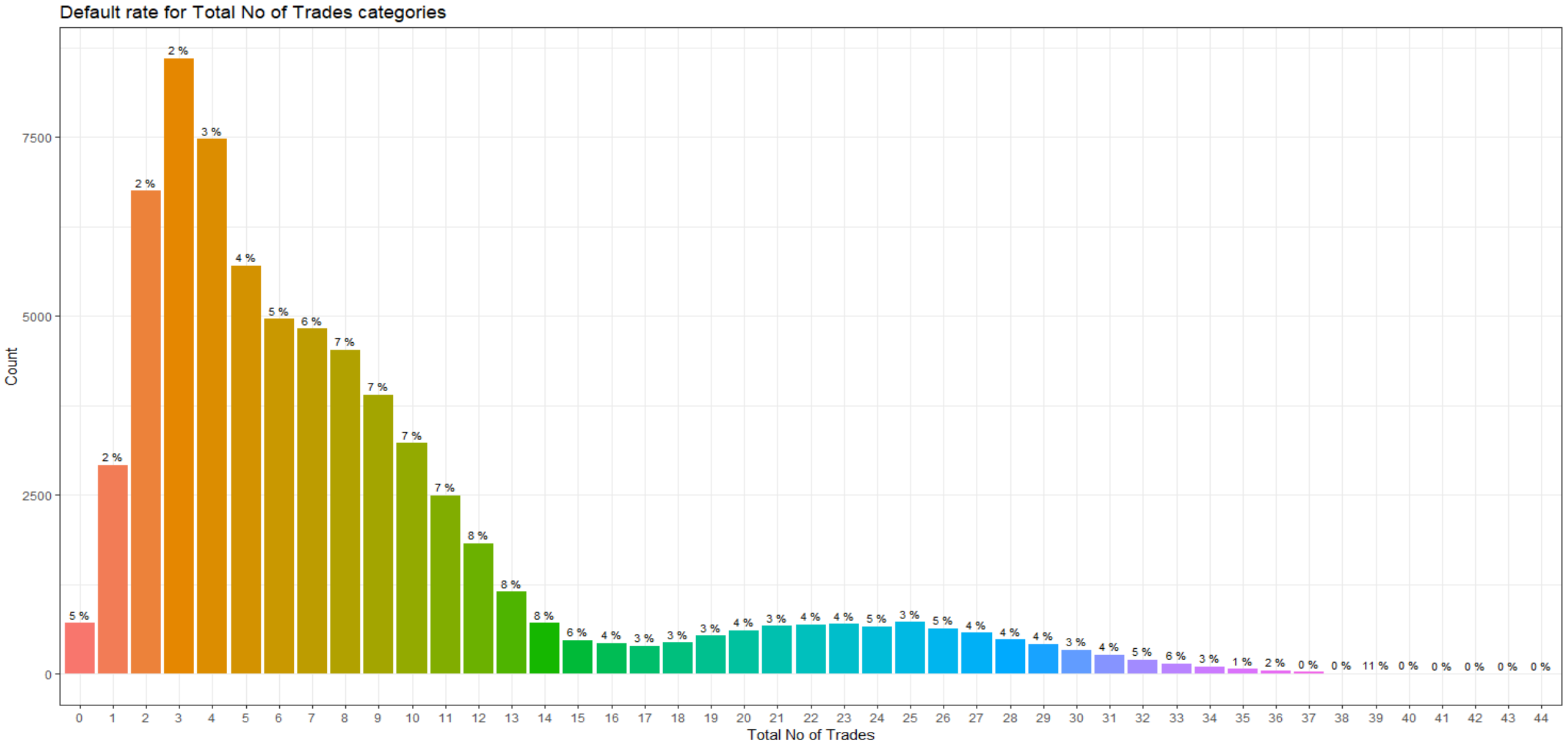
EDA - Variable ‘Outstanding Balance’





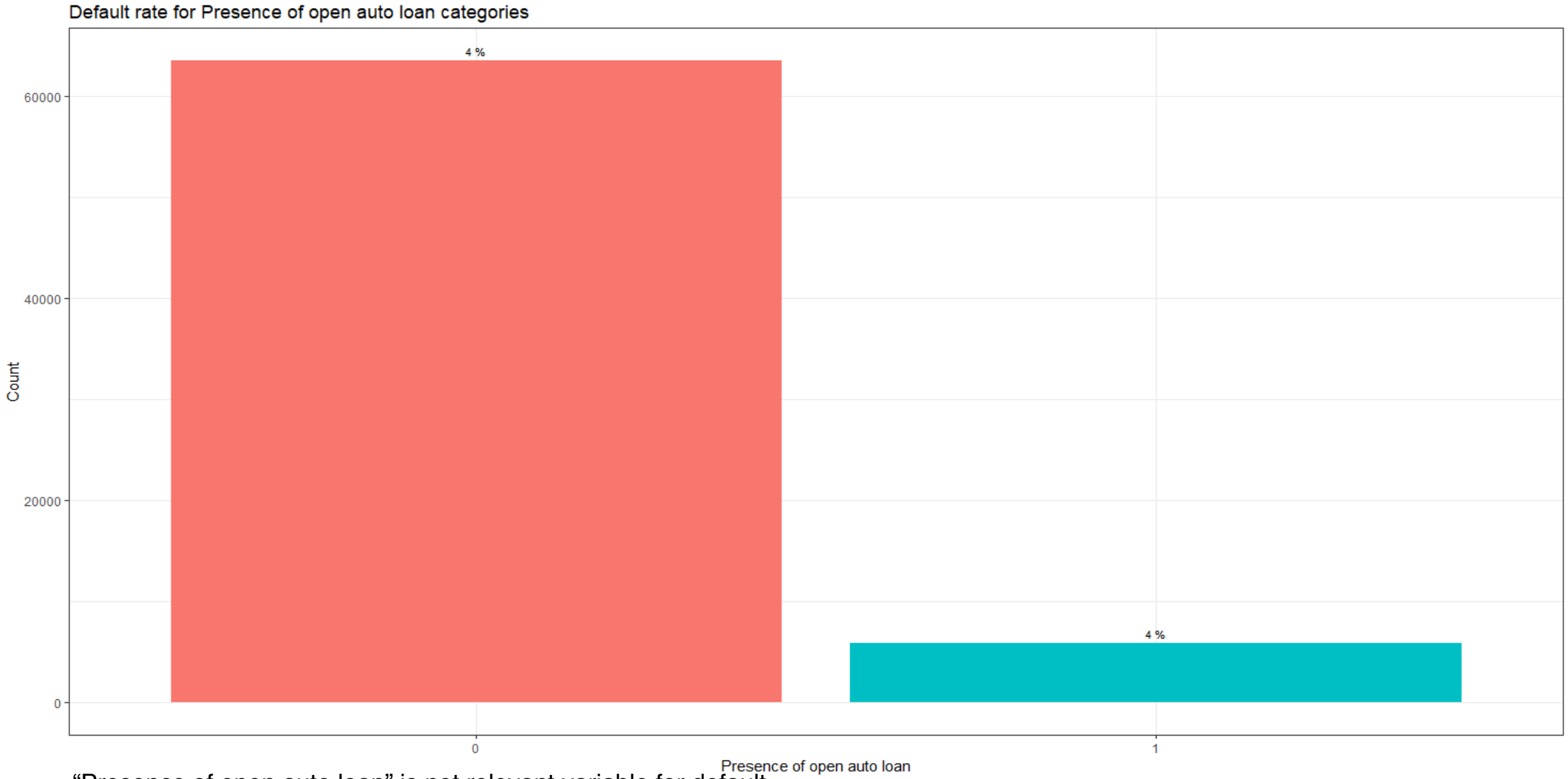
“Outstanding Balance” is relevant variable for default

EDA - Variable ‘Total No. Of Trades’



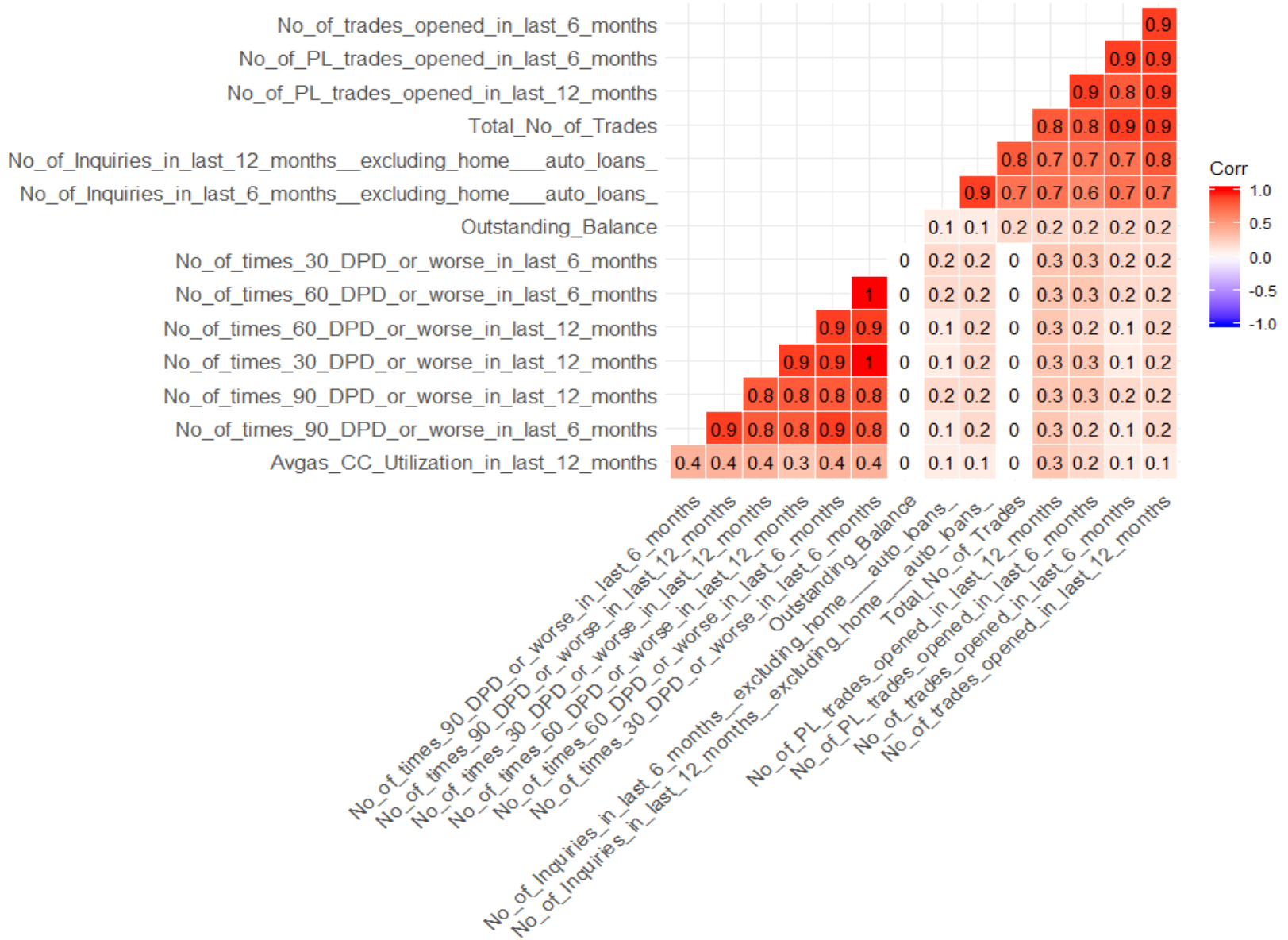
“Total No of Trades” is relevant variable for default

EDA - Variable ‘Presence of open auto loan’



“Presence of open auto loan” is not relevant variable for default

EDA – Correlation among numeric variables



EDA – Correlation among numeric variables

- All the DPD (Day Past Due) variables are highly correlated.
- Number of Trades, Total number of trades in last 6 & 12 months, Number of PL trades in last 6 and 12 months are highly correlated.
- Number of Inquiries in last 6 and 12 months excluding home & auto loans are highly correlated.
- Number of Inquiries in last 12 months excluding home & auto loans are highly correlated with Number of trades open in last 12 months and Total number of trades.

Note - Highly correlated here refers to Pearson correlation coefficient more than 0.70

EDA - Summary

Below table states the significant predictors identified during EDA

Variable
Avgas_CC_Utilization_in_last_12_months
No_of_trades_opened_in_last_12_months
No_of_PL_trades_opened_in_last_12_months
No_of_Inquiries_in_last_12_months__excluding_home___auto_loans_
Outstanding_Balance
No_of_times_30_DPD_or_worse_in_last_6_months
Total_No_of_Trades
No_of_PL_trades_opened_in_last_6_months
No_of_times_90_DPD_or_worse_in_last_12_months
No_of_times_60_DPD_or_worse_in_last_6_months
No_of_Inquiries_in_last_6_months__excluding_home___auto_loans_
No_of_times_30_DPD_or_worse_in_last_12_months
No_of_times_60_DPD_or_worse_in_last_12_months
No_of_trades_opened_in_last_6_months
No_of_times_90_DPD_or_worse_in_last_6_months
No_of_months_in_current_residence
Income
No_of_months_in_current_company
Presence_of_open_home_loan
Age
No_of_dependents
Profession
Presence_of_open_auto_loan
Type_of_residence
Education
Gender
Marital_Status__at_the_time_of_application_

	Significant
	Non Significant

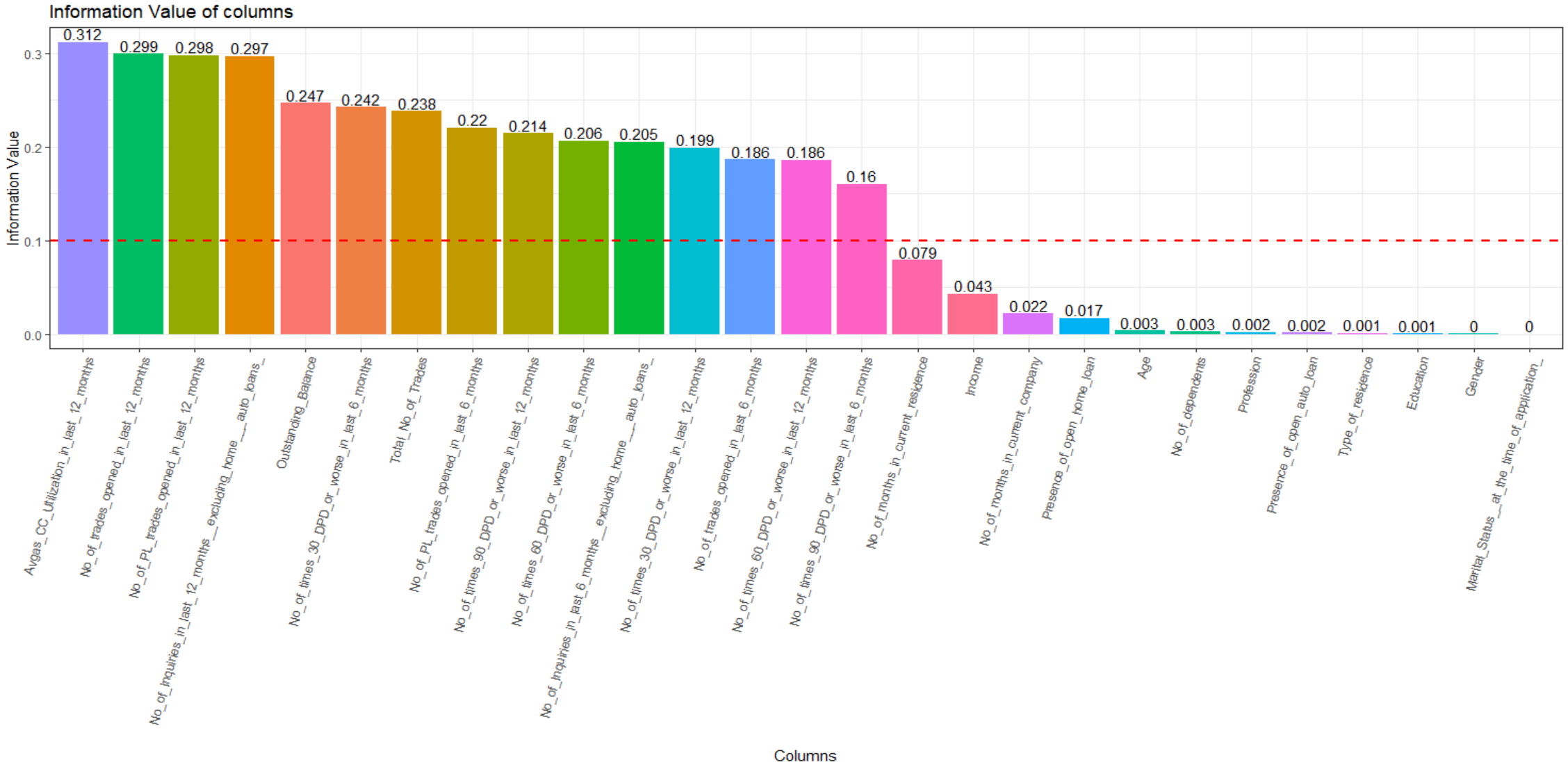
IV - Predictors and its strength determined using IV values

Variable	IV	Importance
Avgas_CC_Utilization_in_last_12_months	0.311135259	Strong
No_of_trades_opened_in_last_12_months	0.298386285	Medium
No_of_PL_trades_opened_in_last_12_months	0.296673173	Medium
No_of_Inquiries_in_last_12_months__excluding_home___auto_loans_	0.295887029	Medium
Outstanding_Balance	0.245971282	Medium
No_of_times_30_DPD_or_worse_in_last_6_months	0.242357073	Medium
Total_No_of_Trades	0.237044464	Medium
No_of_PL_trades_opened_in_last_6_months	0.219615138	Medium
No_of_times_90_DPD_or_worse_in_last_12_months	0.214460147	Medium
No_of_times_60_DPD_or_worse_in_last_6_months	0.206533623	Medium
No_of_Inquiries_in_last_6_months__excluding_home___auto_loans_	0.204693314	Medium
No_of_times_30_DPD_or_worse_in_last_12_months	0.198348292	Medium
No_of_times_60_DPD_or_worse_in_last_12_months	0.185614492	Medium
No_of_trades_opened_in_last_6_months	0.185353757	Medium
No_of_times_90_DPD_or_worse_in_last_6_months	0.160377161	Medium
No_of_months_in_current_residence	0.078384732	Weak
Income	0.042630075	Weak
No_of_months_in_current_company	0.022214018	Weak
Presence_of_open_home_loan	0.017106303	Not Useful
Age	0.003379143	Not Useful
No_of_dependents	0.002720527	Not Useful
Profession	0.002079843	Not Useful
Presence_of_open_auto_loan	0.001635871	Not Useful
Type_of_residence	0.000930864	Not Useful
Education	0.000759243	Not Useful
Gender	0.000356542	Not Useful
Marital_Status__at_the_time_of_application_	9.13033E-05	Not Useful

	Significant
	Non Significant

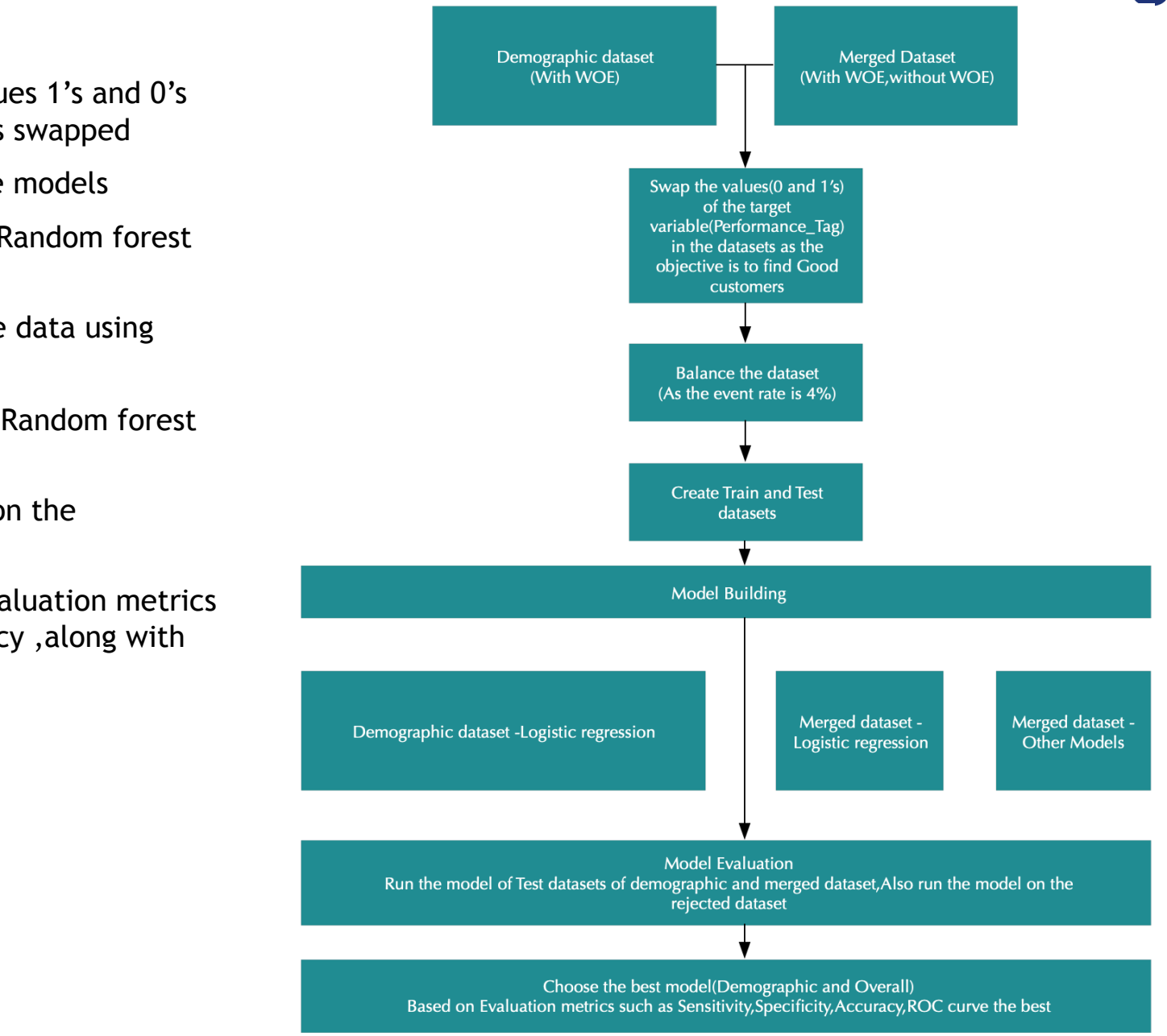
Plot IV - Predictors and its strength determined using IV values

Note : Variable that have "Strong" and "Medium" importance variable will only be used for model b



Model Building and Evaluation

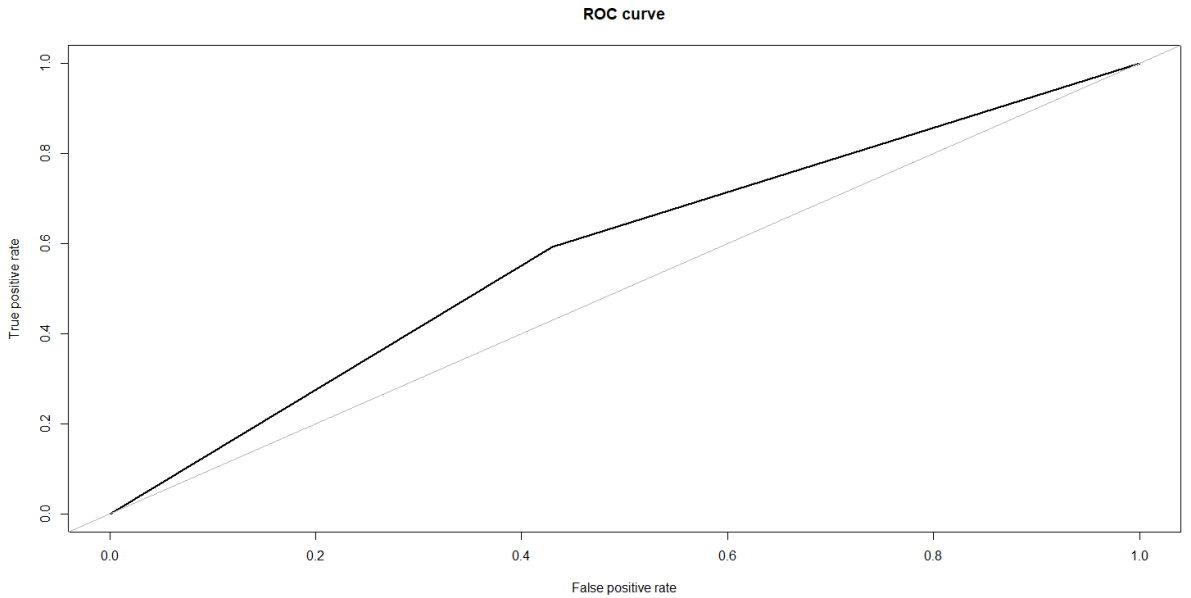
- As stated in the dataset summary the values 1's and 0's of the target variable(Performance Tag) is swapped
- Created train and test datasets for all the models
- Built the model using logistic regression, Random forest on unbalanced merged dataset
- Since the event rate is 4.13%, balance the data using ROSE method
- Build the model using logistic regression, Random forest etc. on the balanced Merged dataset
- Build the model using logistic regression on the demographic dataset
- Evaluate the models created using the evaluation metrics such as Sensitivity, specificity and Accuracy ,along with ROC curve on test datasets
- Choose the best model



‘Logistic Regression’ on Demographic data(Balanced data)

- As observed previously the data is imbalanced.In order to obtain efficient model ,the data is balanced using an R-package called ROSE (Random Over Sampling Example) prior to applying logistic regression method.
- After the model was built ,it was evaluated using confusion matrix .Metrics such as accuracy, sensitivity and specificity was captured for the test data. Following which an ROC curve was plotted.
- Below are the parameters obtained from the model built on balanced data.
- From the below metrics it can be understood that a model built on demographic data is not enough to predict the applicants who will default.
- KS Statistics obtained was **0.2590.**

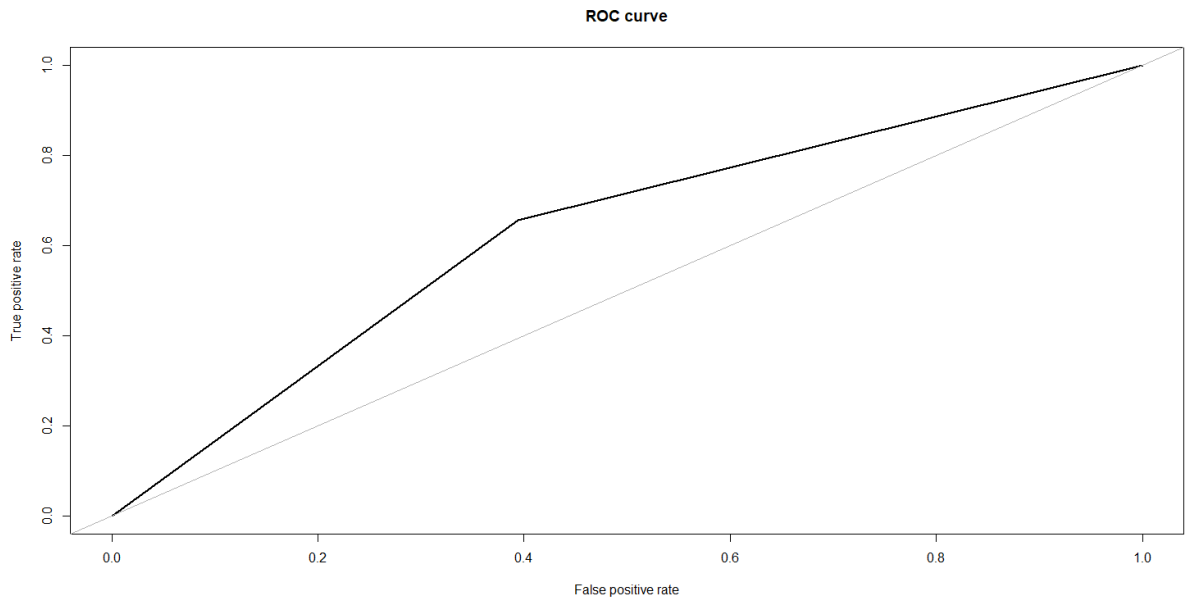
Evaluation Metrics	Values
Accuracy	42.89%
Sensitivity	42.98%
Specificity	40.74%
AUC (Area under the curve)	58.4%



‘Logistic Regression’ on Merged data(Balanced data)

- As previously stated Important variable for model building where chosen using Information value(IV).
- The values of Identified variables are converted into WOE values.
- The data in the variables are then balanced using the R-Package ROSE (Random Over Sampling Example) and following which the Logistic regression model was applied to the data.
- After the model was built ,it was evaluated using confusion matrix .Metrics such as accuracy, sensitivity and specificity was captured for the test data. Following which an ROC curve was plotted.
- Below are the parameters obtained from the model built on balanced data.

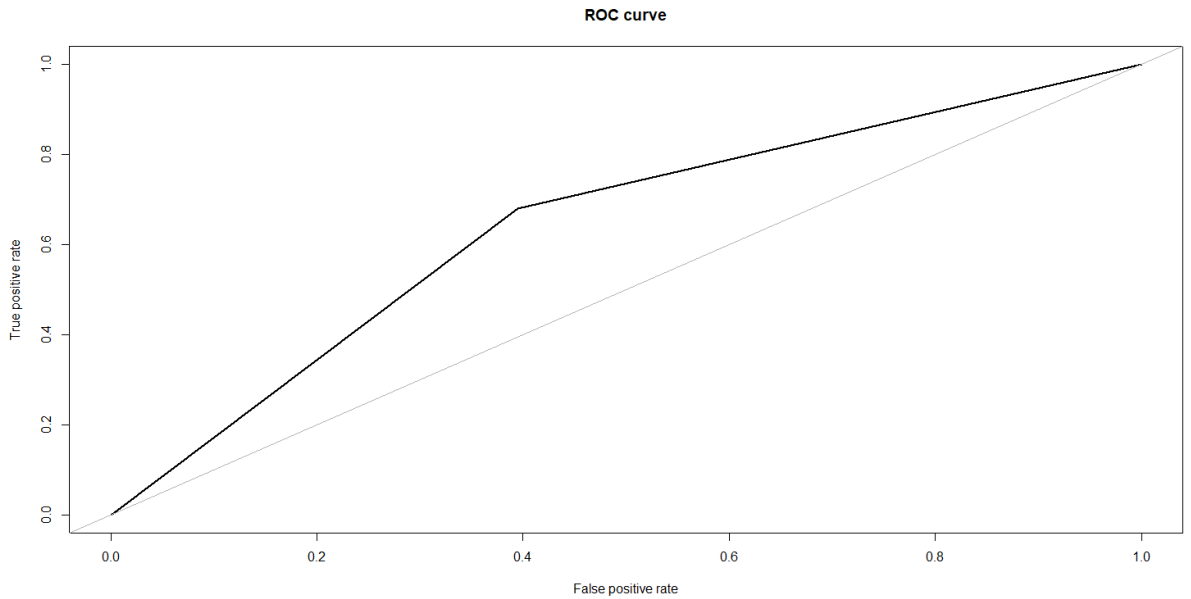
Evaluation Metrics	Values
Accuracy	65.19%
Sensitivity	65.40%
Specificity	60.49%
AUC (Area under the curve)	63.00%



‘Random Forest’ on Merged data(Balanced data)

- The data is setup with important variables identified through information value.
- Initially the model is being built on unbalanced data, later data is balanced using ROSE (Random Over Sampling Example) method and again the model is built.
- Model is built on the above data by cross validation using the metric as “ROC”.
- After the model was built ,it was evaluated using confusion matrix .Metrics such as accuracy, sensitivity and specificity was captured for the test data. Following which an ROC curve was plotted.
- Below are the evaluation parameters for model building on balanced data.

Evaluation Metrics	Values
Accuracy	67.62%
Sensitivity	67.79%
Specificity	60.49%
AUC (Area under the curve)	64.20%



Note : We selected Random Forest model on balanced data as the final model because of maximum evaluation metrics than other models.

Comparison of the models Built

	Accuracy %	Sensitivity %	Specificity %	ROC %
Logistic Regression - Demographic data	42.89	42.98	40.74	58.4
Logistic Regression - Merged data	65.19	65.40	60.49	63.00
Random Forest - Merged data	67.62	67.79	60.49	64.20

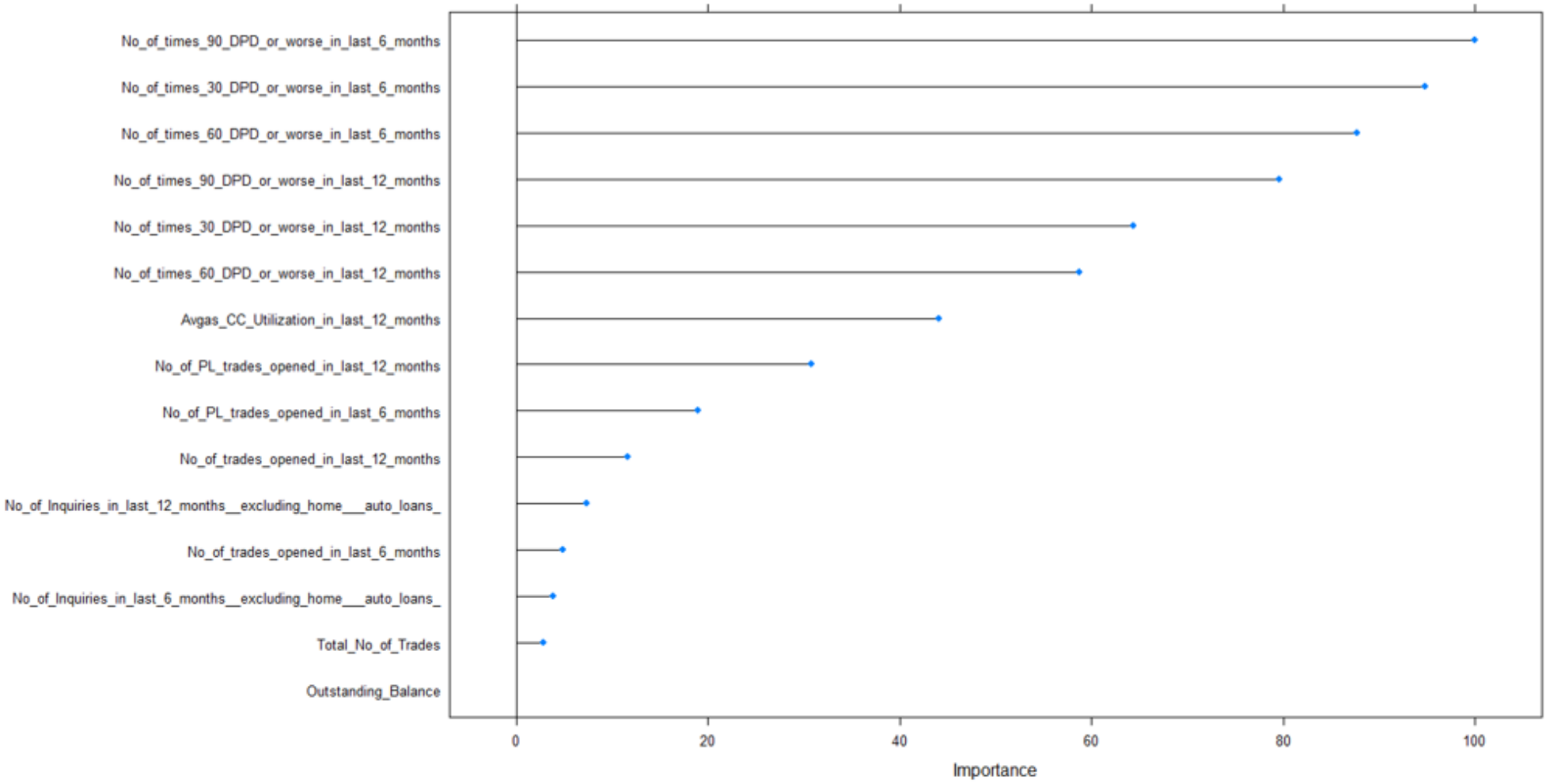
- The above table shows a comparison of the different models run on demographic and merged dataset.

Note *: The models shown are only run on balanced data.(Unbalanced data is highly unreliable ,hence it is not shown/neither used for prediction purpose*

- The table shows parameters pertaining to the model run on demographic dataset as well and is shown for the comparison purpose only however it is disqualified from being used for comparison with the likes of the models run on merged dataset.
- Of all,Random forest model run on the merged dataset seems to provide a better accuracy and other parameters compared to the logistic regression model.Hence Random forest is chosen as the final model to predict the defaulters of the loan.
- Model is able to correctly predict 57% of the applicants who have defaulted.
- Model is able to correctly predict 67.95% of the population who have defaulted.
- Model is able to correctly predict 99.85% of the ***rejected population*** as expected to default.

Important variables from the Final model(Built using Random Forest)

Below is graphical representation of the important variables from the final model which was built using using Random Forest.The longer the horizontal line the more degree of important the variable is in the model for prediction



Application Score card

- A scorecard was built on merged dataset that consists of accepted applicants.

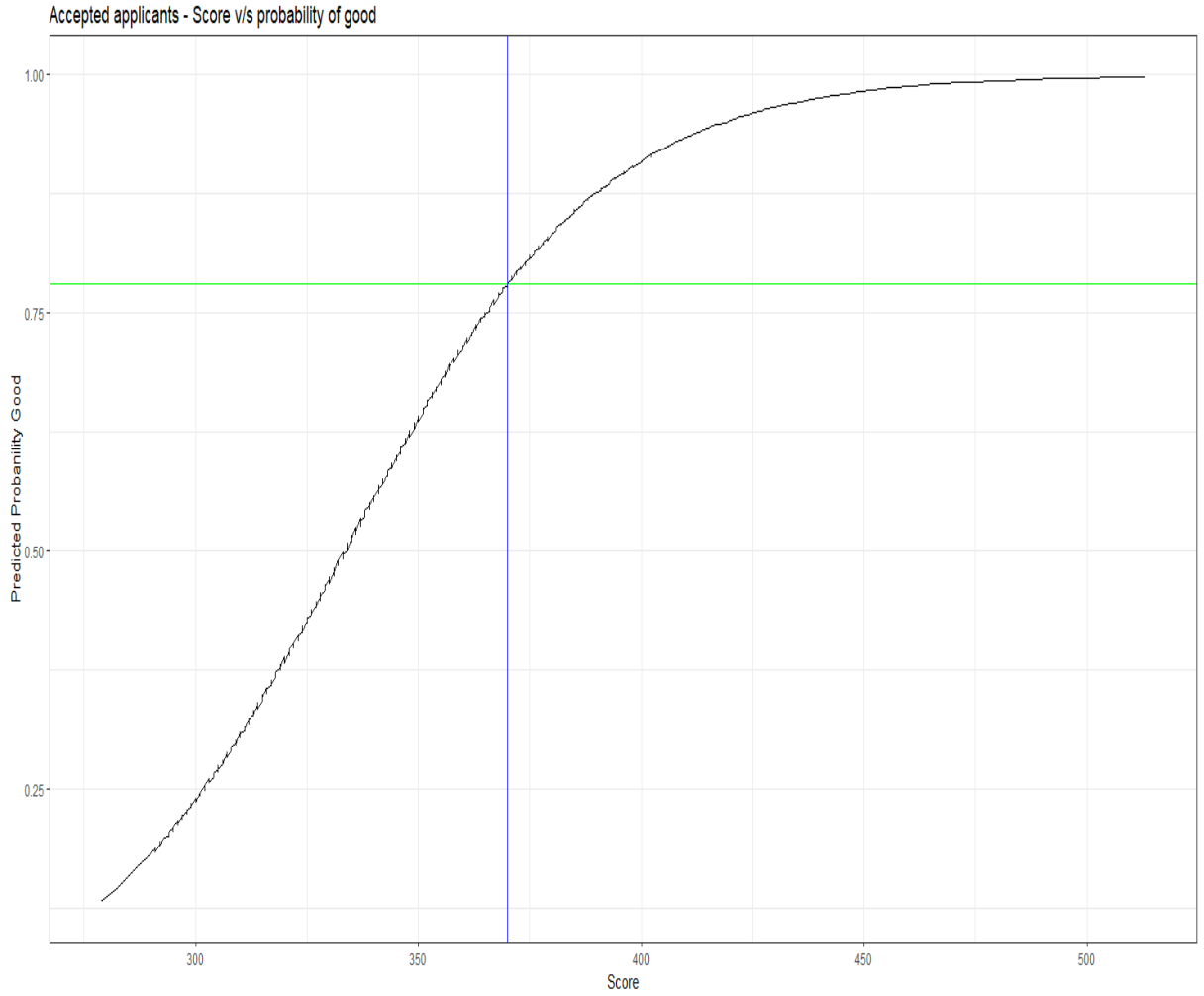
***Note** : This scorecard can help organization to categorise customers into Good Vs Bad and allows to select the customers that they want to provide loans*

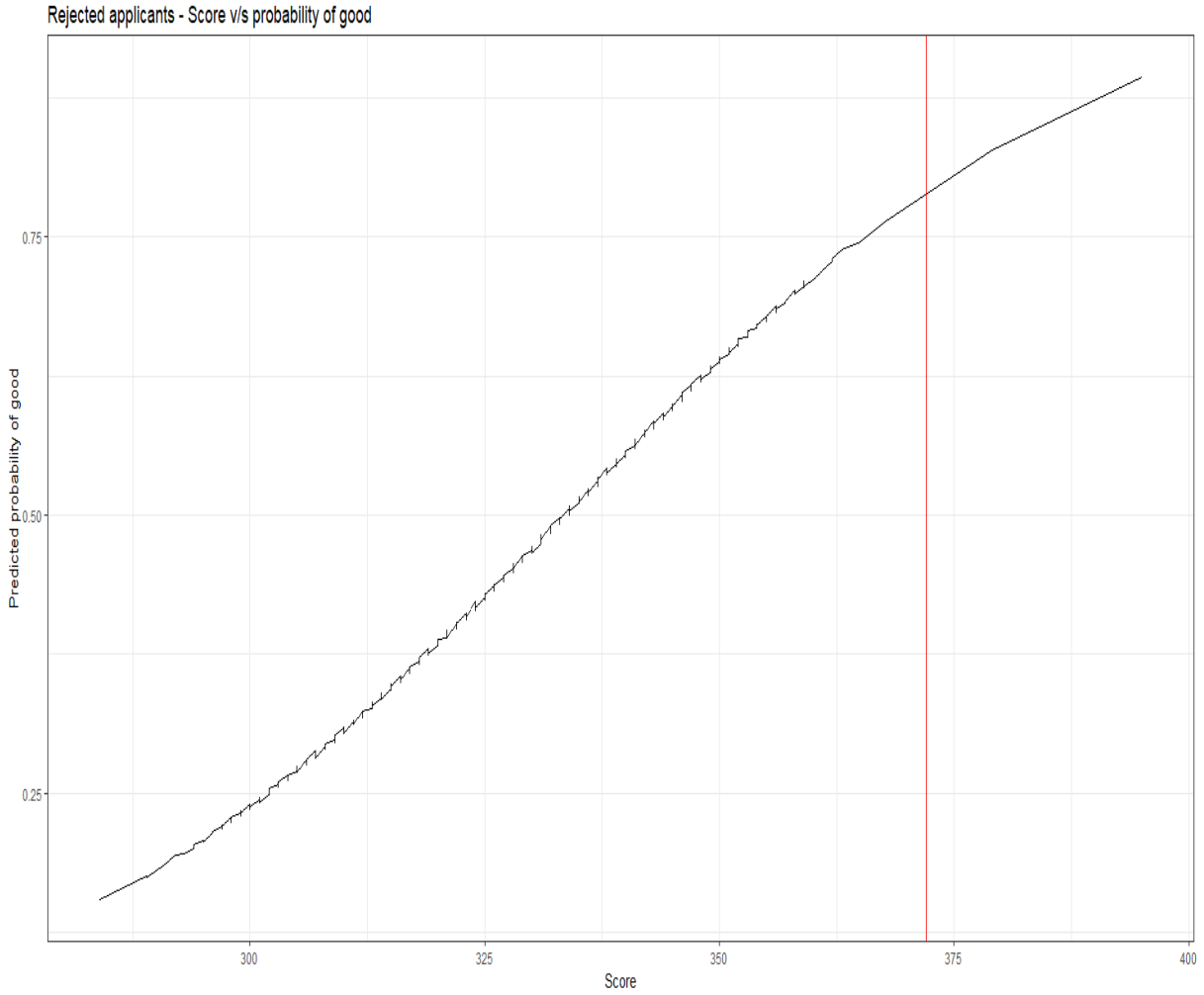
- The following were considered to build the score card
    - Point to Double,Base Score & good to bad odds was taken as 20, 400 & 10 respectively. After which the offset and scores were calculated for each applicant.
- Cut-off was identified as 370.Any applicants whose score falls below 370 can be expected to default and so Credex can choose to reject those applicants who are below the cutoff.
- Applicants whose score is above 370 can be chosen to be selected as customers.

Summary of the score card pertaining to the dataset

- Scores range from 279 to 513 for applicants with median score being 399
- There are 46486 applicants whose scores are above or equal 370 and thus will be accepted as customers
- There are 23015 applicants whose scores are below 370 and thus will be rejected being Credex customers

A Graphical Representation of the Application Score card





Score vs Probability of good for accepted and rejected applicants

END of Part1 of the project

Thank You

Note : For *financial benefits* of the project ,Please read the Financial benefit document I.e Part2 of the project