|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DETAILS OF ASSIGNMENT | | | | | |  | | | | |
| **STUDENT NAME** | **Sanee Salim** | | | **SWINBURNE ID NUMBER** | | | | | | **101887181** |
| **EMAIL ADDRESS** | **101887181@student.swin.edu.au** | | | **PHONE CONTACT** | | | | | | **+6145966560** |
| **UNIT CODE & NAME** | **Data Mining – STA30004** | | | | | | | | | |
| **ASSESSMENT TITLE** | **Data Mining Assignment** | | | | | | | | | |
| **TUTOR’S NAME:** | **Pragalathan Apputhurai** | | **DATE OF SUBMISSION:** | | | | | | **20/10/2019** | |
|  | . | | | | | | | | | |
|  |  | | | | | | | | | |
| **DECLARATION** |  | | | | | | | | | |
| I declare that ( the first four boxes must be completed for the assignment to be accepted):  This assignment does not contain any material that has previously been submitted for assessment at this or any other university.  This is an original piece of work and no part has been completed by any other student than signed below.  I have read and understood the avoiding plagiarism guidelines at <http://www.swinburne.edu.au/ltas/plagiarism/students.htm> and no part of this work has been copied or paraphrased **from** **any other source** except where this has been clearly acknowledged in the body of the assignment and included in the reference list.  I have retained a copy of this assignment in the event of it becoming lost or damaged.  (optional) I agree to a copy of the assignment being retained as an exemplar for future students (subject to identifying details being removed). | | | | | | | | | | |
| **Student acknowledgement (by signing or typing your name you agree to the above):** | | Sanee Salim | | | | | **Date:** | 20/10/2019 | | |
|  | | | | | | | | | | |
| **DETAILS OF FEEDBACK** | | | | |  | | | | | |
|  | | | | | | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Office Use Only** | | | |
| Date Received |  | Received by |  |
| Total Mark / Grade |  | Marker |  |

# Introduction

This report is submitted as a project for the unit STA30004 – Data Mining. This report critically analyses motor insurance claim data. The dataset has more than 73000 policies. Some of the variables in the dataset are

CAR\_AGE measures the age of the insured car in years

DRIVERS measures the number of people who are specified as designated drivers

EXPOSURE measures the fraction of the year for which the policy was active

MILEAGE measures the expected mileage travelled in a single year

PRIMAGE gives the age of the primary driver in years

TOTAL gives the total amount claimed on the policy in the year

EXCESS = 0, 75 or 100 indicating the excess claim amount associated with each policy. The insurance company will not pay out claims below this excess amount.

USAGE specifies how the car is used (S=only social, SB=strictly business, SC=social and business, ST=social and taxi)

CLAIM=1 if there was at least one claim during the year, 0 otherwise.

The tools used for the analyses are RStudio (based on R programming) and the package consistently used is Rattle.

# Analysis

This section includes many subsections which analyse the data based on different levels as shown below:

* Descriptive Statistics
* Exploratory Analysis
* Association Analysis
* Classification Tree
* Classification with Random Forest and Boosting
* Classification with Regression and SVM
* Classification with Neural Networks
* Comparison of Classification Results

## Week 1 – R.

In the first week for this assignment, all the data was obtained from the specified source. The data file was downloaded, renamed and critically observed to understand more about the variables and the dataset. All the work of this week was performed on R studio, and use of rattle was not utilised yet.

First, we figured out how many values were present in the dataset. After that, a random sample of 5000 values with Claim =1 and 5000 values with Claim =0 was created. The seed for the creation of this random sample was the Swinbunre student ID. Two subsets with all false values and true values were created. Later they were compiled to make a table. Then, the first column’s name was changed to policy, and lastly, a new CSV file was created. The script and CSV file were also submitted with the assignment.

Many innovative questions arise as to the relationship between a claim are being registered or not based on different variables. Some of these are:

1. What is the relationship between the age of the primary variable and the Claim status?
2. How likely is a car with more than average mileage to be having more than average claim amount?

Many other questions were omitted at this point but were concluded throughout the semester and mentioned in the report.

## Week 2 – Descriptive Statistics

The aim of this week’s task was to perform descriptive statistics on this data set. The objective of this descriptive statistics was to understand the basic features of the data and get summary statistics for our dataset. From this week onwards, rattle was consistently used for feasibility and better analysation.

The first task of this week was to load the data into rattle and assign them roles accordingly. In our dataset the renamed column policy was the identifier, the claim was the target variable, and the total was the risk variable. Then, claim variable was transformed and recoded as categoric as the values were just one and zero and this can be deemed a qualitative measure for the summary statistics (FIGURE 1). For the total variable the first quartile is 0, that means 25% of the data is less than or equal to 0. It can be seen that 25% of the claim amount is zero.

For the total variable, the third quartile is 395, that means that 25% of the data is greater than or equal to 395. This shows that 25% of the claim amount is greater than or equal to 395. The minimum amount claimed is 0.The maximum amount claimed is 54271.3. On average a customer claims an amount of 611.3.

The boxplot for mileage (FIGURE 2)shows that the Inter Quartile Range (IQR) of the mileage for TFC Claim (0,1) is less than that of TFC Claim (0,0). The mileage for TFC Claim (0,0) and TFC Claim (0,1) is normal and not skewed. The medians reflect that TFC Claim (0,1) has more mileage than TFC Claim (0,0)

The boxplot for excess(FIGURE 3) shows that the IQR of the excess for TFC Claim (0,1) is like that of TFC Claim (0,0). The excess for TFC Claim (0,0) is positively skewed, and TFC Claim (0,0) is negatively skewed. The medians reflect that TFC Claim (0,1) has more excess than TFC Claim (0,0)

The boxplot for drivers (FIGURE 4)shows that the boxplot that the IQR of the Drivers for TFC Claim (0,1) is like that of TFC Claim (0,0). The Drivers for TFC Claim (0,0) is negatively skewed, and TFC Claim (0,0) is also negatively skewed. The medians reflect that TFC Claim (0,1) has similar drivers than TFC Claim (0,0)

The boxplot for car age (FIGURE 5)shows that the IQR of the Car\_AGE for TFC Claim (0,1) is smaller than that of TFC Claim (0,0). The Car\_age for TFC Claim (0,1) is slightly positively skewed and TFC Claim (0,0) is normal. The medians reflect that TFC Claim (0,1) has less car\_age than TFC Claim (0,0)

The boxplot for exposure (FIGURE 6)shows that the IQR of the Exposure for TFC Claim (0,1) is smaller than that of TFC Claim (0,0). The Exposure for TFC Claim (0,1) is negatively skewed, and TFC Claim (0,0) is also negatively skewed. The medians reflect that TFC Claim (0,1) has similar exposure as of TFC Claim (0,0)

The boxplot for primary age (FIGURE 7)shows that the IQR of the Primary Age for TFC Claim (0,1) is like that of TFC Claim (0,0). The Primary age for TFC Claim (0,1) is normal and TFC Claim (0,0) is also normal. The medians reflect that TFC Claim (0,1) has smaller primage compared to that of TFC Claim (0,0).

The pairs plot and correlation plot reflect (FIGURE 8) (FIGURE 9)

* A weak negative correlation between excess and mileage (-0.02)
* A weak negative correlation between excess and mileage (-0.02)
* A weak positive correlation between driver and mileage (0.10)
* A weak negative correlation between car\_age and mileage (-0.19)
* A weak negative correlation between exposure and mileage (-0.03)
* A weak negative correlation between primage and mileage (-0.27)
* A weak negative correlation between drivers and excess (-0.04)
* A weak negative correlation between car age and excess (-0.05)
* A weak negative correlation between exposure and excess (-0.02)
* A weak negative correlation between primage and excess (-0.18)
* A weak negative correlation between car age and drivers (-0.08)
* A weak negative correlation between exposure and drivers (-0.16)
* A weak negative correlation between primage and driver (-0.24)
* A weak positive correlation between exposure and car age (0.05)
* A weak positive correlation between primage and car age (0.10)
* A weak positive correlation between primage and exposure (0.12)

After that, the log of claim +1 was taken out as claim value was zero at times and that logging that would lead to 5000 indefinite amount.

Lastly, the data was partitioned as 40/30/30 for training validation, and testing, respectively and a tree was run in which total was an essential variable for the tree.

## Week 3 – Exploratory Data Analysis

The aim of this week’s task was to perform an exploratory on this data set. The objective of this exploratoty data analysis was to understand the underlying patterns of the data. For this week’s assignment, GGOBI was used to explore using brushing and other tools.

Firstly, automatic brushing was enabled for TFC\_Claim (Claim category) (FIGURE 10). In this analysis purple color was used for Claim = yes and yellow was used for Claim = no. (Figure 11). The scatterplot for all the variables can be used to check claim status against each variable and find patterns. One of the observed trends is that claim value is yes when primary age is less or when primary age is more. (Figure 12). Figure 13, 14, 15 and 16 reflect that younger drivers are more likely to get claims. After that, on rotating view the variables that give the best separation for TFC\_Claim are excess, car age and primary age. Lastly, a 1D of age of primary driver as the manipulation variable which provides evidence of a bimodal distribution. This bimodal distribution is perfectly bimodal which shows that the value of TFC\_Claim= yes and no both have highest frequencies.

## Week 4 – Association Analysis

This week’s task depended on the MBA Motor dataset in which the type of claim was reflected. Association analysis was performed to find relationships between these claims. The types of claims include:

WSCLMS=WS for windshield claims,

ADCLMS=AD for accidental damage,

FTCLMS =FT for fire or theft,

PDCLMS = PD for personal damage claims,

PICLMS = PI for personal injury claims,

After loading the data set in rattle and assigning appropriate rules, an association analysis was run. This showed us that there are 14701 datasets. The mean values for support, confidence and lift were 0.07, 0.56, and 2.42 respectively (Figure 21). After that, frequency plot for the probabilities of each of the type of claim was made (Figure 19) which shows the highest probability for AD claims. Next, an association analysis was conducted, and the rules in Figure 20 shows that the probability for both an AD and PI Claim is 0.04, it also shows that estimated probability of a PD claim for a policy claim that includes a PI claim is 0.89 and that PI type of claim increases the likelihood of a PD claim by 3.59. Lastly, someone with an AD and PD claim is very likely to have a PI Claim as well because of a larger lift value. We also discovered that as we decrease the minimum confidence, we see more rules because it finds even the slightest values of confidence and support and shows more rules.

## Week 5 – Classification Trees

For this week’s task, we go back to the old csv file. We run classification trees on our dataset which are used to predict categorical dependent variable based on predictor variables.

Upon running a classification tree the rules for claim = yes were: (Figure 22)

Rule number 20: Indicates that for 42% of the observations (probability = 0.42), when observed exposure is greater than or equal to 0.05886 and primage is greater than or equal to 34.5 and primage is less than 67.5 and excess is greater than or equal to 87.5 there is a claim recorded. The other information provided with the rule is that 2305 observations from the training dataset (i.e. 33 % of the observations in the training dataset) are covered by this rule.

Rule number 4: Indicates that for 11% of the observations (probability = 0.11), when observed exposure is greater than or equal to 0.05886 and primage is less than 34.5 there is a claim recorded. The other information provided with the rule is that 402 observations from the training dataset (i.e. 6 % of the observations in the training dataset) are covered by this rule.

On the other hand the rules for the nodes classifying claim = no were (Figure 23)

Rule number 3: Indicates that for 80% of the observations (probability = 0.89), when observed exposure is less than 0.05886 there is no claim recorded. The other information provided with the rule is that 479 observations from the training dataset (i.e. 7% of the observations in the training dataset) are covered by this rule

Rule number 11: Indicates that for 61% of the observation (probability = 0.61), when observed exposure is greater than or equal to 0.05886 and primage is greater than or equal to 34.5 and primage is greater than or equal to 67.5 there is no clam recorded. The other information provided with the rule is that 1389 observations from the training dataset (i.e. 20% of the observations in the training dataset) are covered by this rule.

Rule number 21: Indicates that for 51% of the observation (probability = 0.51), when observed exposure is greater than or equal to 0.05886 and primage is greater than or equal to 34.5 and primage is less than 67.5 and excess is greater than or equal to 87.5 there is no claim recorded. The other information provided with the rule is that 2425 observations from the training dataset (i.e. 35% of the observations in the training dataset) are covered by this rule.

It was observed that the optimal number of splits based on the cross validation result was three because as the cross validated error(xerror) and crossed validated standard deviation (xstd) values are the least when number of splits is 3 (Figure 24). Next, the same analysis was done assuming a loss matrix with losses half as big for a false positive (CatClaim=”Yes”) than a false negative (CatClaim=”No”). On rerunning the tree each of the nodes had one policy each (Figure 25,26,27). After that, we considered node two and tried to find the better decision based on the loss matrix and the results were Yes:1.06, No:0.47, yes is the better decision because lower loss for claim decision. Repeating the same process for node 3 gave us, Yes is he better decision to node 3. No:0.89, yes:0.22

## Week 6 - Classification with random forests and boosting

Random forests use the training data, but the runtimes are very fast. The random forest can also deal with unbalanced and missing data.

On running a random forest for 500 trees and randomly selecting four variables the OOB estimate of error rate was 40.56% (Figure 28). OOB means out of bag misclassification in the training data set. The error rate for TFC\_Claim yes was 42.7% and for no it was 38.4%. (Figure 29). According to the Gini measure of variable importance primary age, exposure and car age were the three most important variables. (Figure 30). The accuracy variable importance was defined by randomly permuting the values of a predictor variable across all observations, one at a time, and measuring the resulting decline in prediction accuracy for each tree. According to the error plot, the optimum number of trees was 300. (Figure 31)

On running a boosting ensemble, the OOB estimate of error rate was 36.7% (Figure 32). The error rate for TFC\_Claim yes was 34.2% and for no it was 39%.(Figure 33).According to the Gini measure of variable importance usage, drivers and primary age were the three most important variables. (Figure 34) According to the error plot, the optimum number of trees was 37 (Figure 35). The first five trees were weak learners because they were slightly correlated with the true classification. Amongst the single tree, random forest and a boosting ensemble a boosting ensemble would be the better tool for our data because of lower OOB error

## Week 7 – Classification with Regression & SVM

This week’s task is based on linear regression which is useful for finding relationships between continuous variables. We also use SVM (Support Vector Machine) which is a supervised machine learning algorithm which is used for classification problems

On running a linear regression model with TFC\_Claim as target variable on a logit link function, the model seemed overfitted as some of the variables are present in this model which are insignificant. The AIC value was 9141. (Figure 36)

Given a collection of models for the data, AIC estimates the quality of each model. When this model was refitted after omitting the usage variable, contrary to expectations, the AIC value increased to 9152.4. (Figure 37)

In my opinion, the driver’s variable should be omitted as well as it is also not significant. After that, the odds ratio for the predictors was calculated and interpreted below:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Coefficient (B) | Exp(B) | Interpret odds ratios |
| Mileage | -0.000032413 | 0.99996758752 | 0.1% reduction in the odds of the event on average of the event when mileage is increased by 1 when other predictors are controlled. |
| Excess | -0.008457325 | 0.99157833756 | 0.9% reduction in the odds of the event on average of the event when excess is increased by 1 when other predictors are controlled. |
| Drivers | -0.014587654 | 0.98551823033 | 1.5% reduction in the odds of the event on average of the event when drivers are increased by 1 when other predictors are controlled. |
| Car\_Age | 0.026428174 | 1.02678049508 | 2.6 % increase in the odds of the event on average of the event when the car age is increased by 1 when other predictors are controlled. |
| Exposure | -1.098742879 | 0.33328980606 | 66.6% reduction in the odds of the event on average of the event when exposure is increased by 1 when other predictors are controlled. |
| primage | 0.029613393 | 1.03005623002 | 3% increase in the odds of the event on average of the event when primage is increased by 1 when other predictors are controlled. |

Next, we used the same predictor variables to develop an SVM. In that case, the error rate was 0.3565 (35.7%). The number of support vectors used was 5556. (Figure 38. The comparison between linear model and SVM using the validation data showed us an error rate of 39.3% for regression and 37.9% for SVM. Similarly, the area under the ROC curve was 76% for regression and 78% for vector machine. (Figure 39,40). After that, evaluation of SVM and linear model were performed and SVM was deemed as the better model in terms of lift chart (Figure 41), precision chart (Figure 42) and the risk chart (Figure 43,44) In SVM the likelihood of seeing a claim was higher than that of linear for lift and precision charts. Similarly, the area under the ROC curve is more significant for SVM compared to linear regression model.

## Week 8 - Classification with neural networks

This week’s task required us to do classification based on neural networks and Self Organising Map.

Firstly, all the input variables were rescaled to a z score in the transformation tab. The hidden nodes were fixed to be two, and the neural network model was run on the validation data. The area under the ROC Curve is 0.6701 (Figure 45). This is also reflected in the Risk chart (Figure 46). When the model was refitted with three hidden nodes the ROC Curve value increased to 0.6795 (Figure 47), and it was also reflected in the risk chart (Figure 48). The model was again refitted with one hidden node. The area of ROC curve decreased this time to 0.6585 (Figure 49), and it was also reflected in the risk chart (Figure 50). Considering the area under the ROC curve values it can be concluded that the optimum number of hidden nodes for the neural network model would be three hidden nodes.

On considering a self-organising map the training progress showed that around 83 iterations were needed for the map to converge (Figure 51). The variable that seems to be most effective for differentiating policies with and without claim appears to be the driver variable (Figure 52). The quality plot is also very consistent. (Figure 53)

# Comparison of Classification Results

The table below shows the comparison between all the classification models above.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Simple Classification Tree | Boosting Ensemble | Random Forest | Simple Vector Machine | Linear Model | Neural Network |
| Error Rate | 40.1% | 36.8% | 38.8% | 38% | 39.3% | 37.2% |
| Area Under Recall | 76% | 79% | 77% | 78% | 76% | 79% |
| Area Under Risk | 56% | 58% | 58% | 59% | 63% | 59% |
| ROC Curve | 0.64 | 0.69 | 0.66 | 0.67 | 0.65 | 0.68 |

## Error Rate:

Error rate means the proportion of patterns that have been incorrectly classified by a decision model. Amongst all our models the model with the least error is the Boosting Ensemble, and the highest error rate is of that the Simple Classification tree

## The area under recall:

The area under recall means the proportion of actual positive cases which are correctly identified by the classification tree. In our case the most prominent area under recall is that of boosting ensemble and neural network that means these two classification methods do the best job at identifying positive cases.

## The area under risk:

The area under risk means the measure of risk associated with each observation. That means the adjustments that had to be made. Higher risk value means more adjustment on the observations. If the Risk values were very high, we might be inclined to not consider the classification tree, but in our case all the risk values are close to each other.

## ROC Curve:

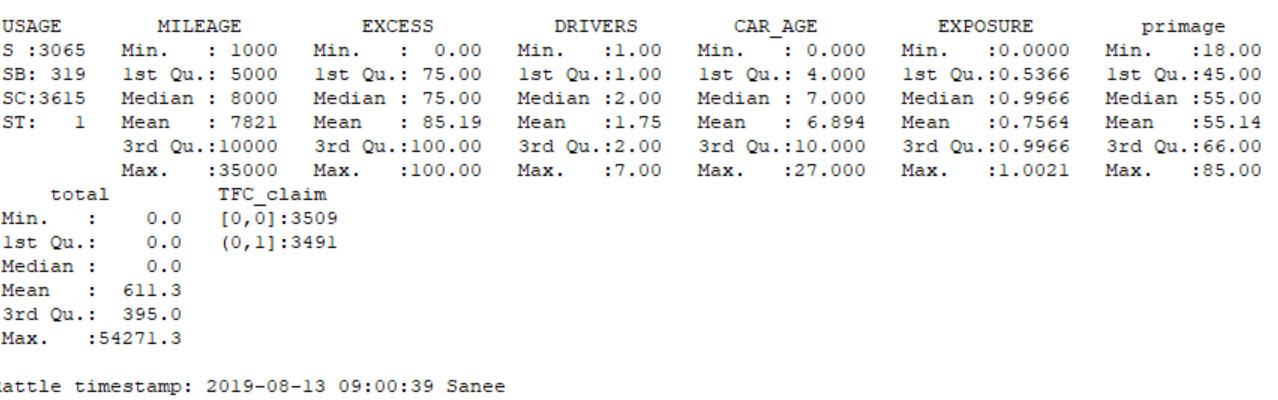
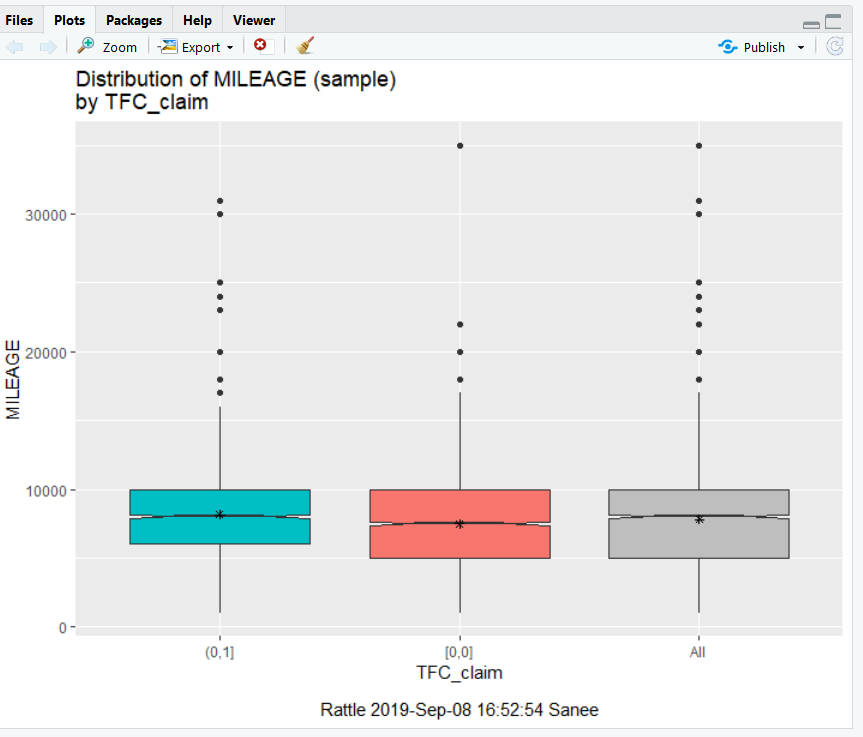
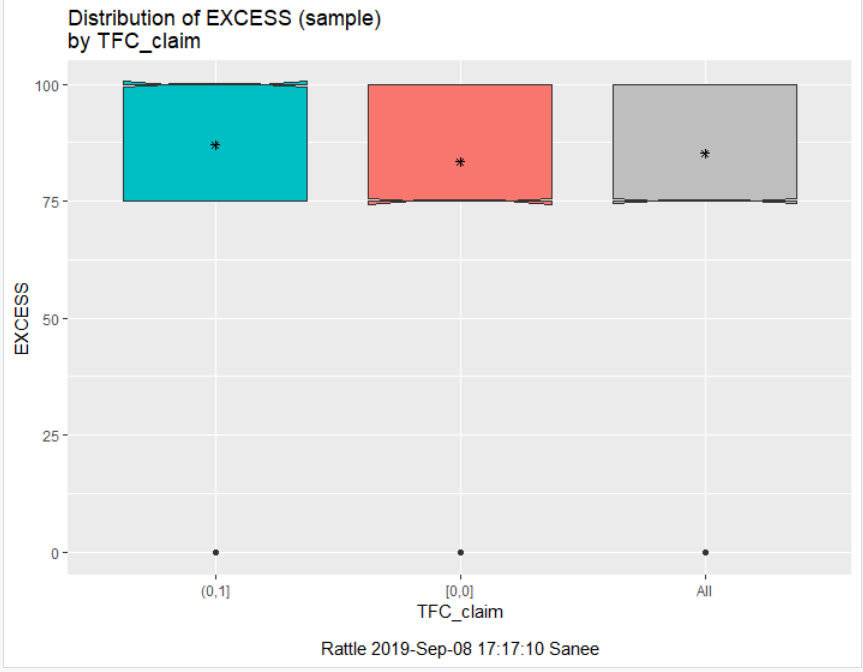
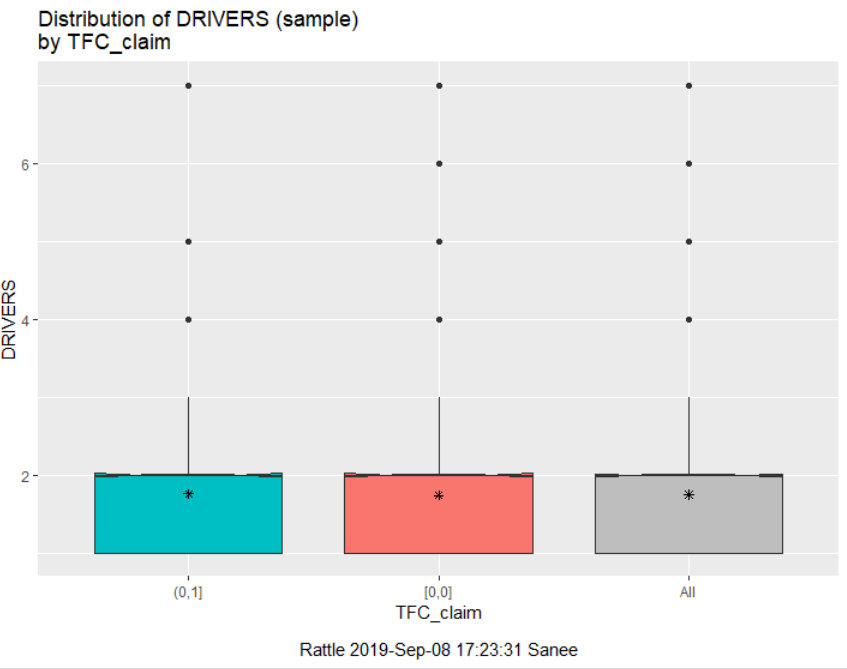
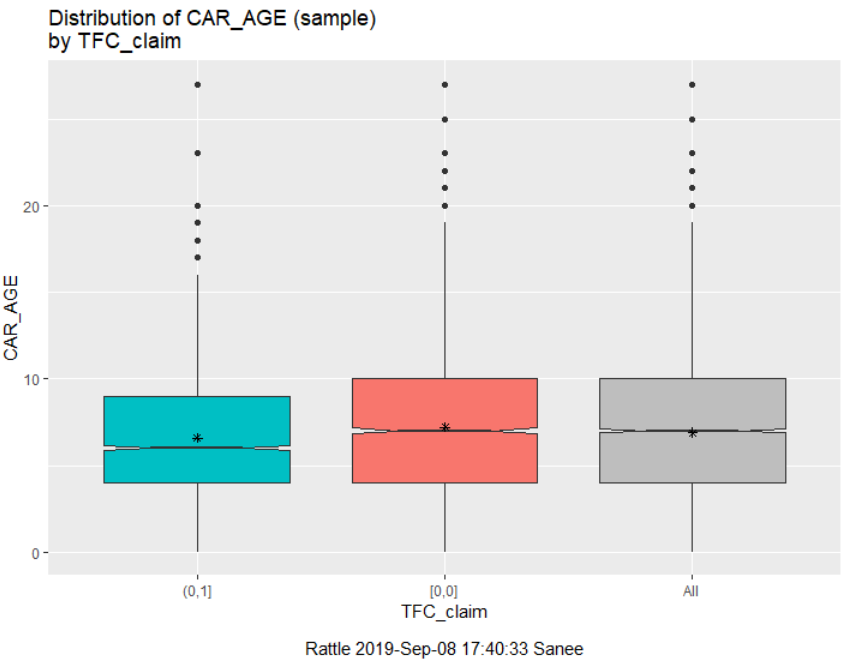
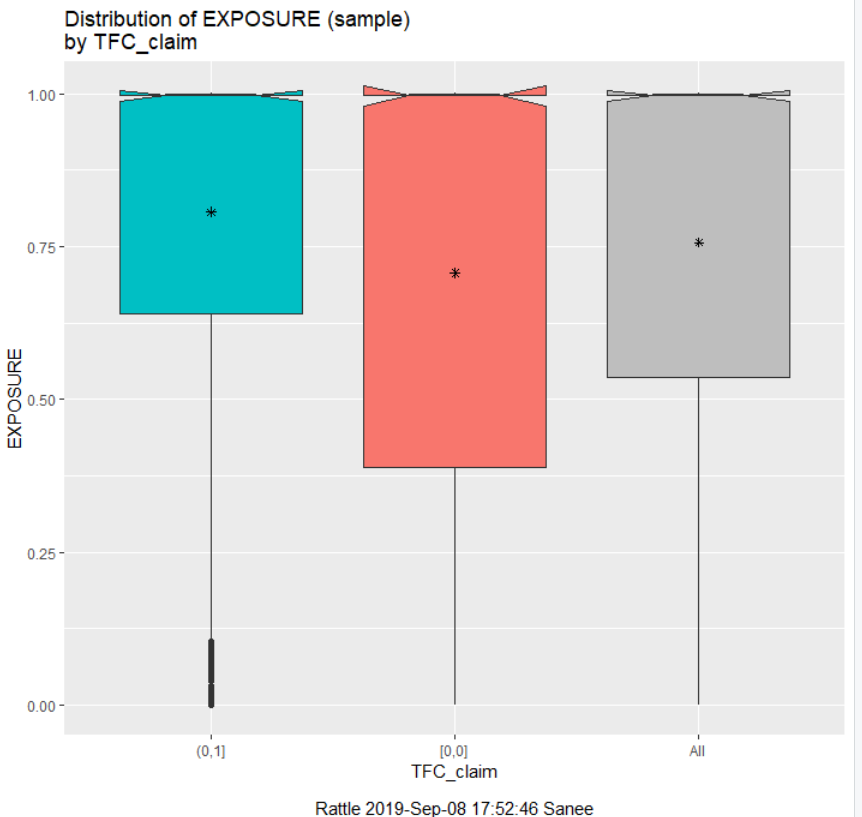
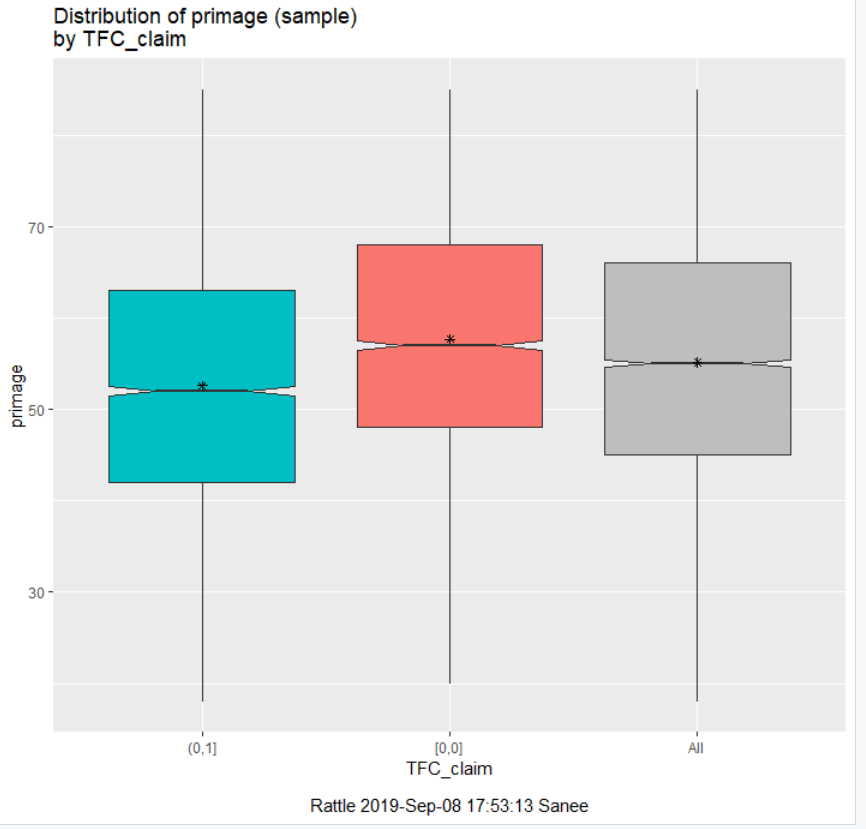
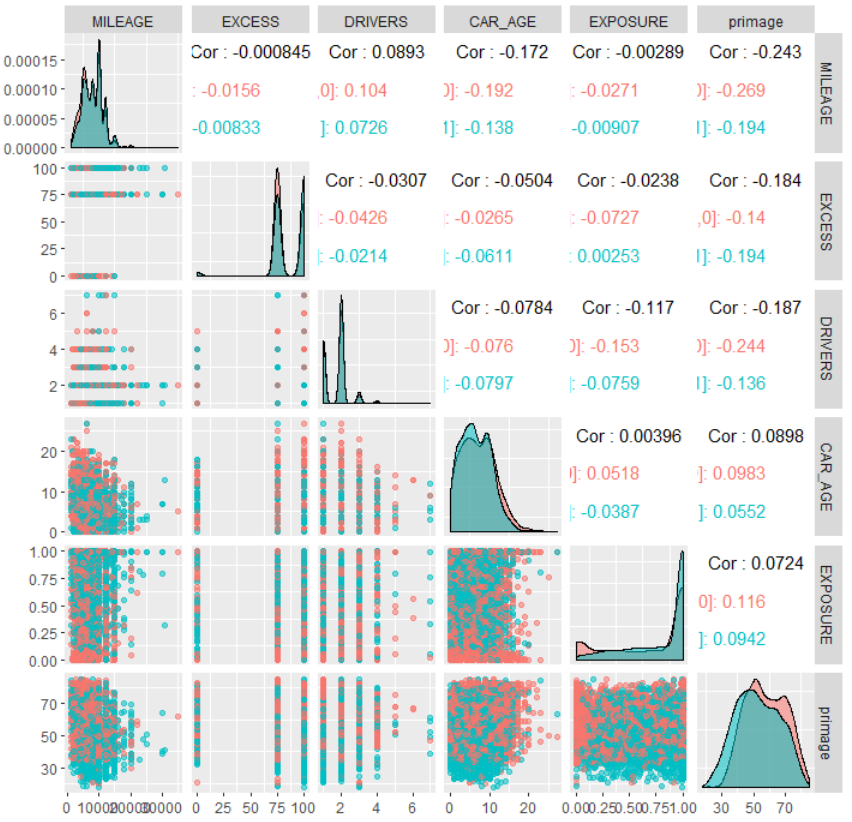
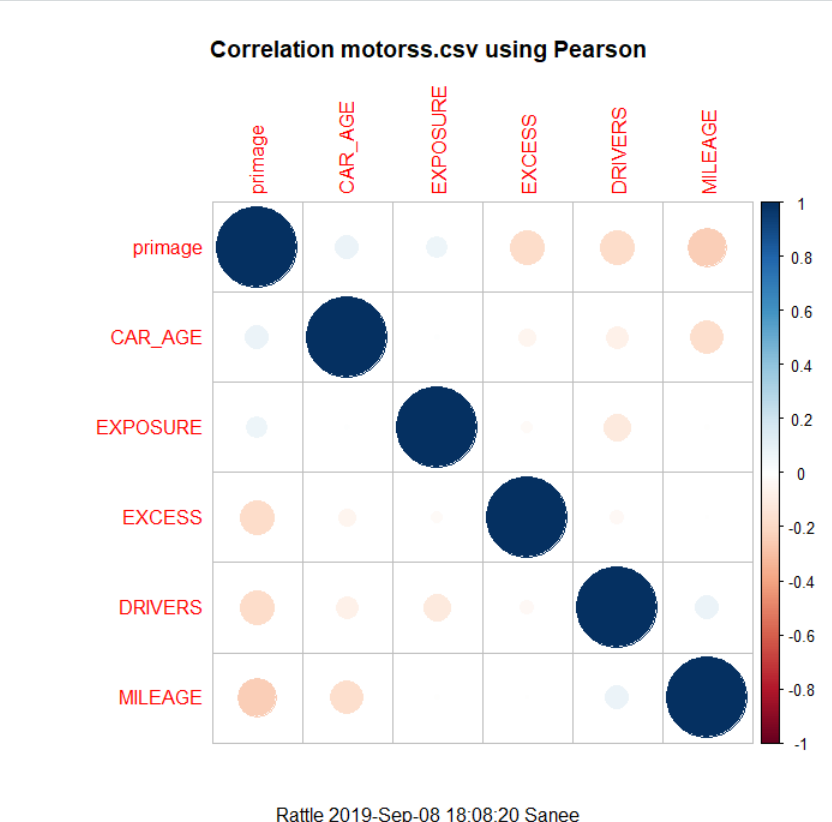
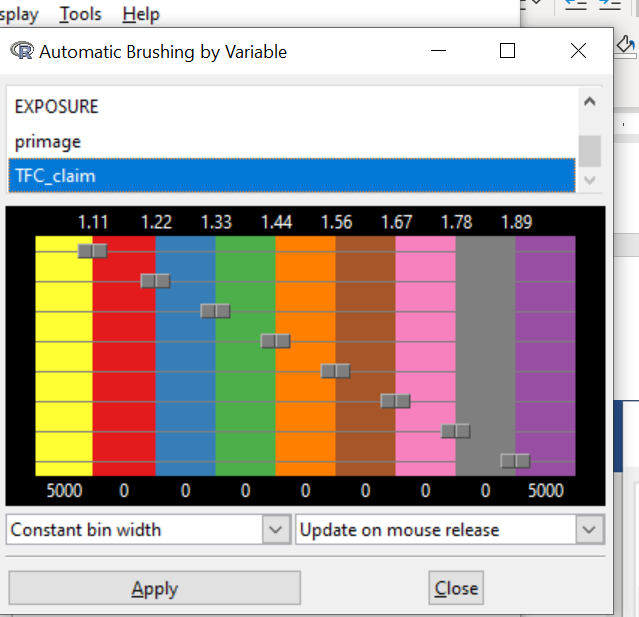
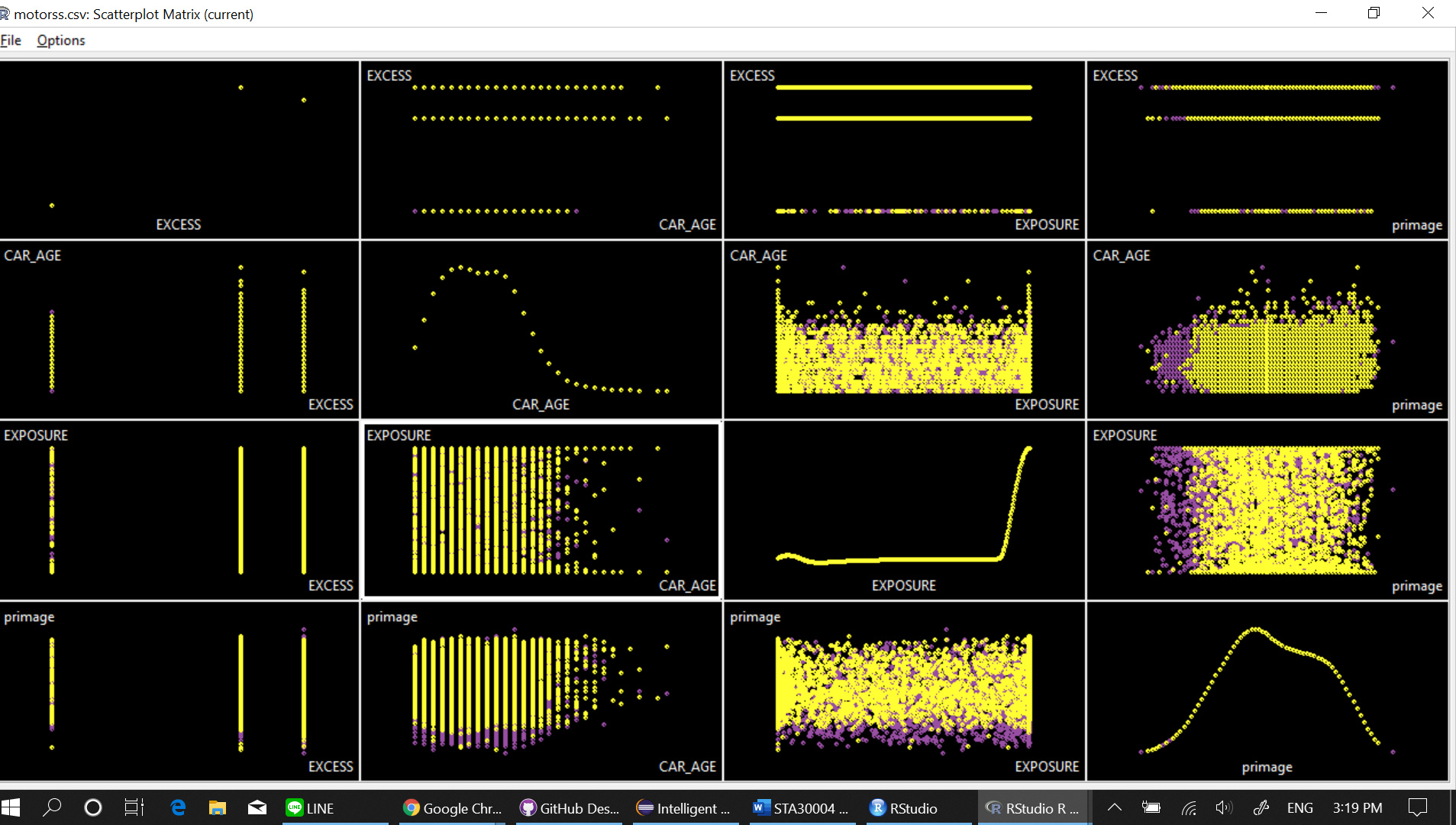
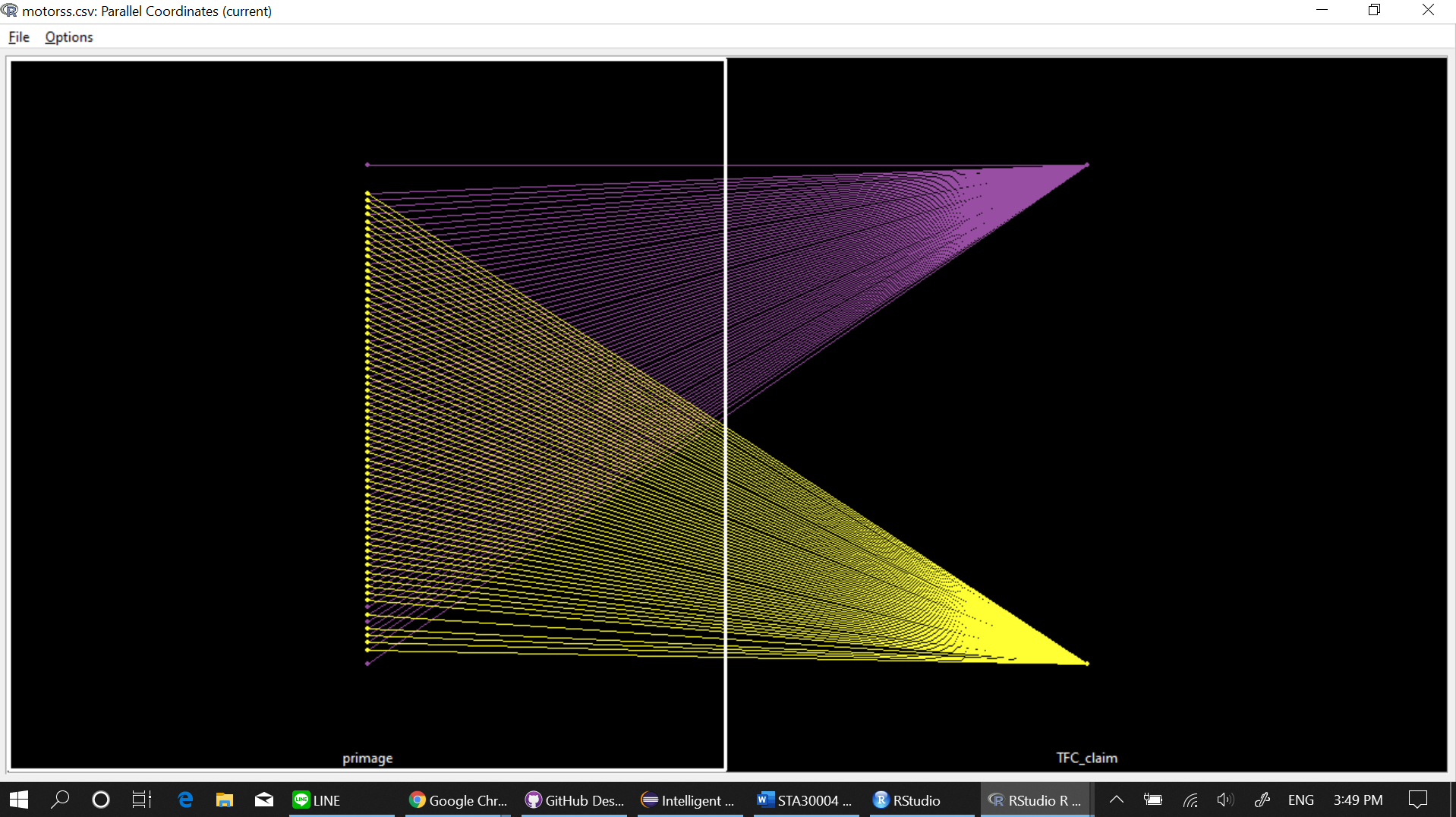
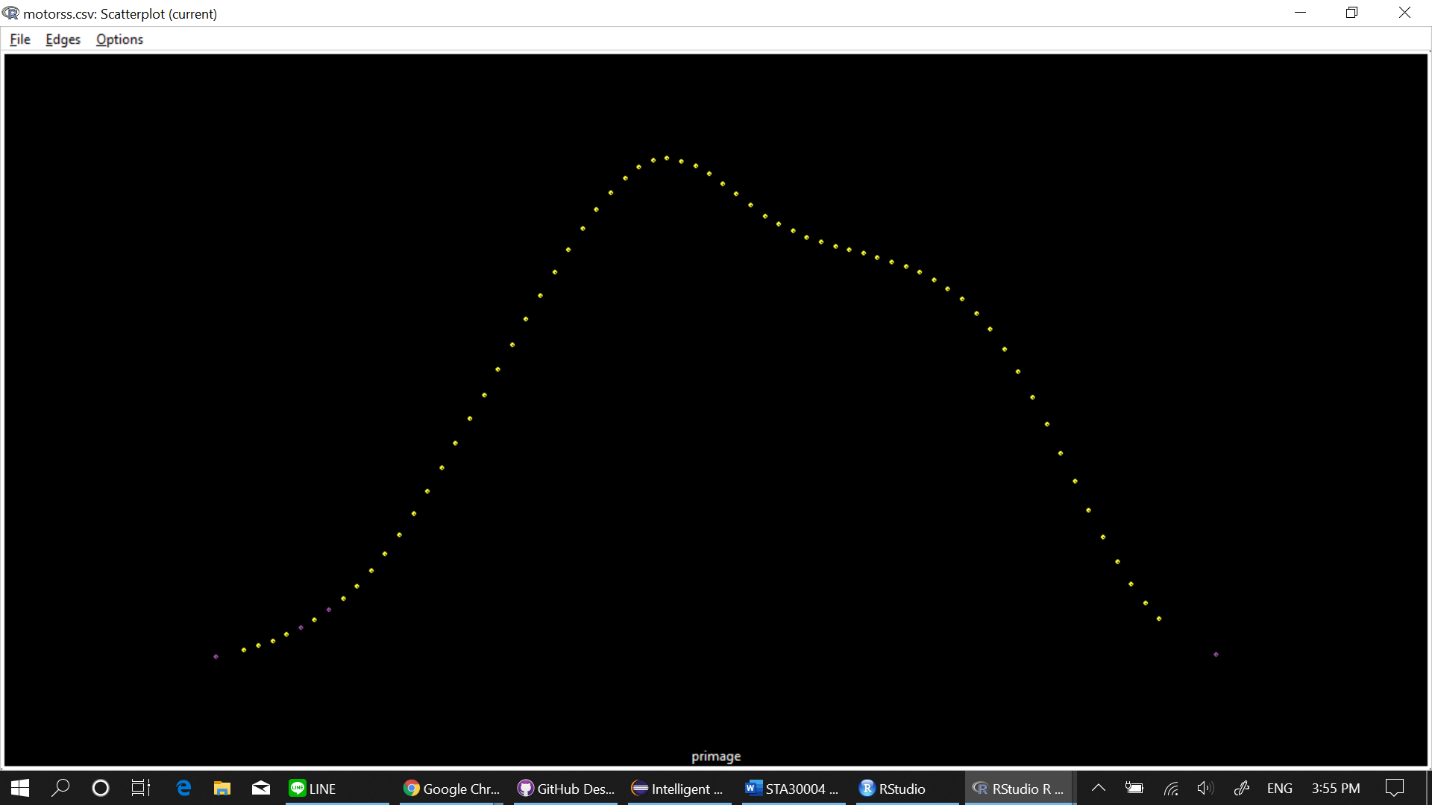
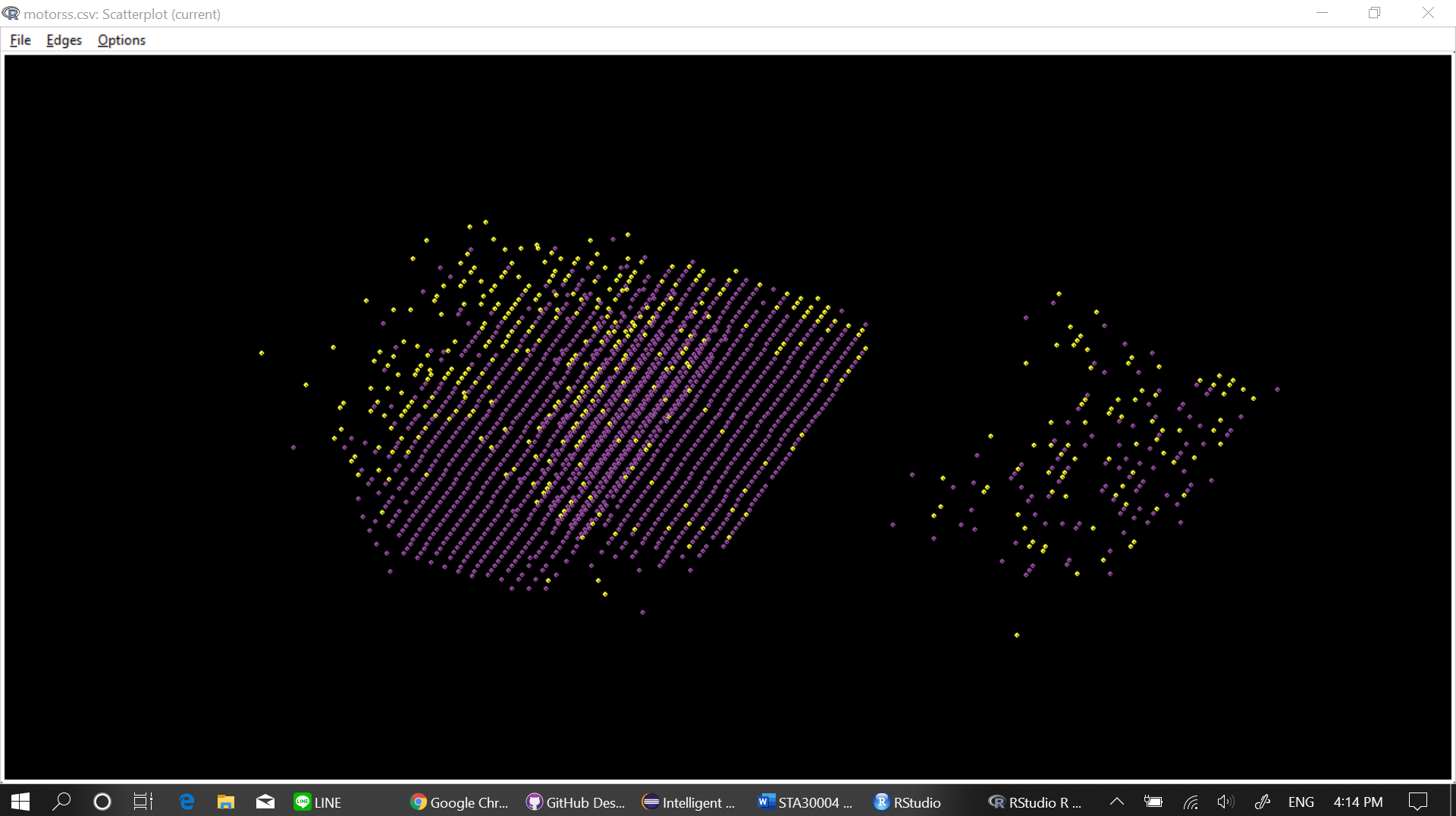
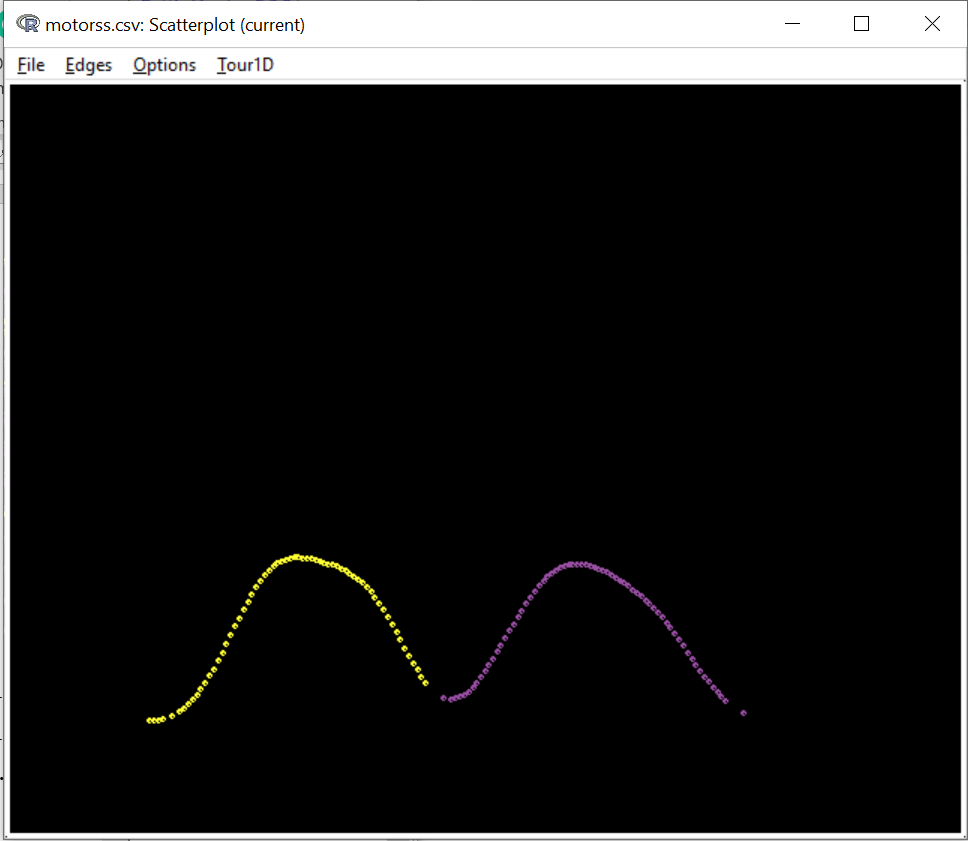
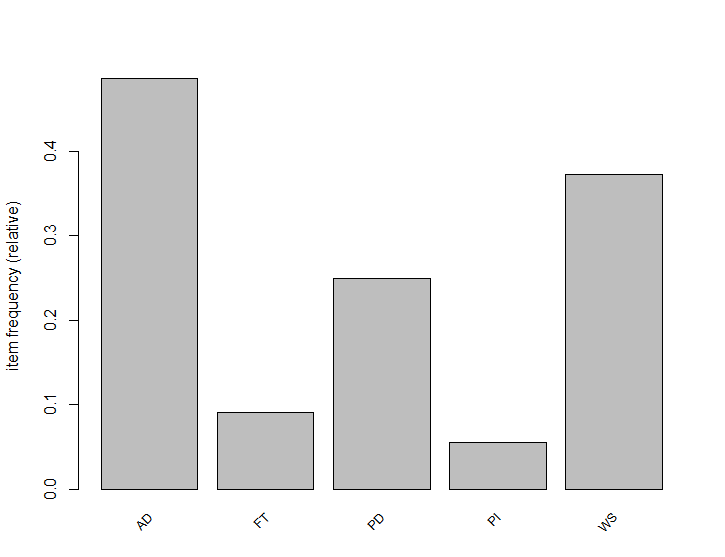
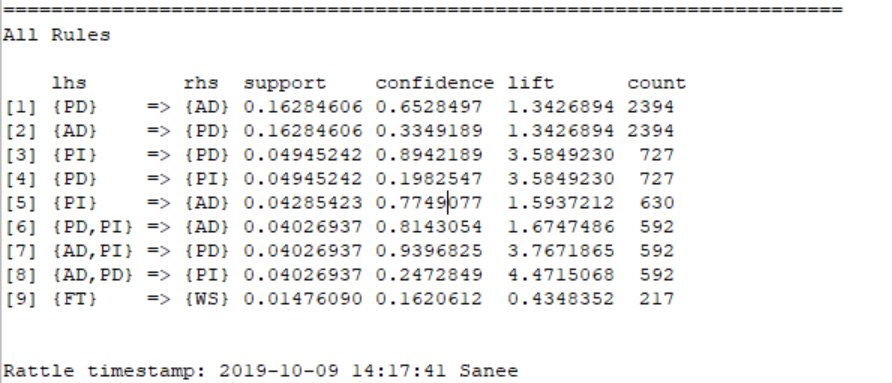
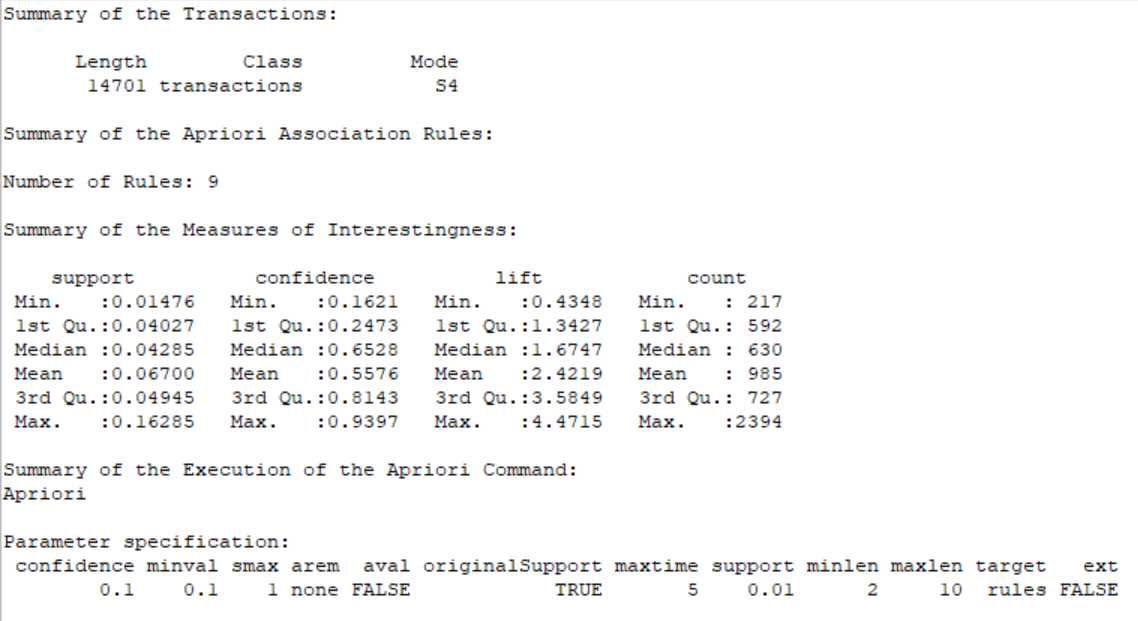
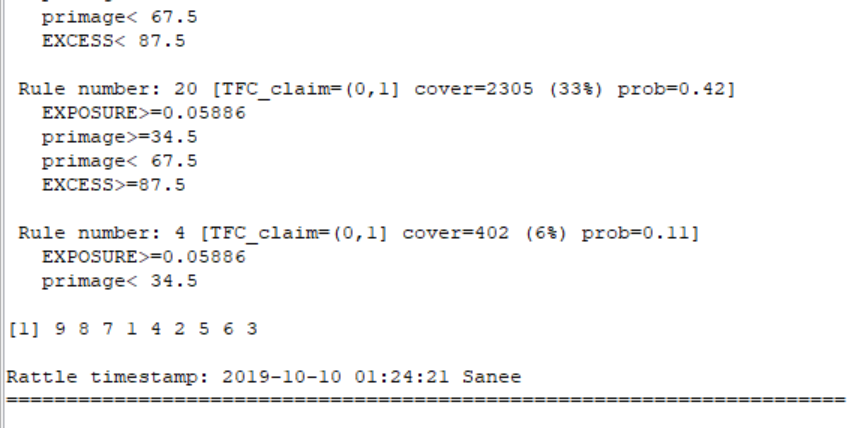
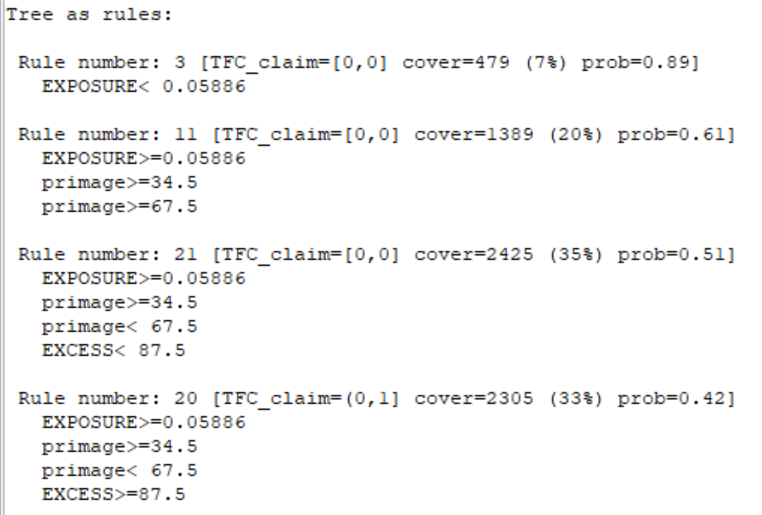
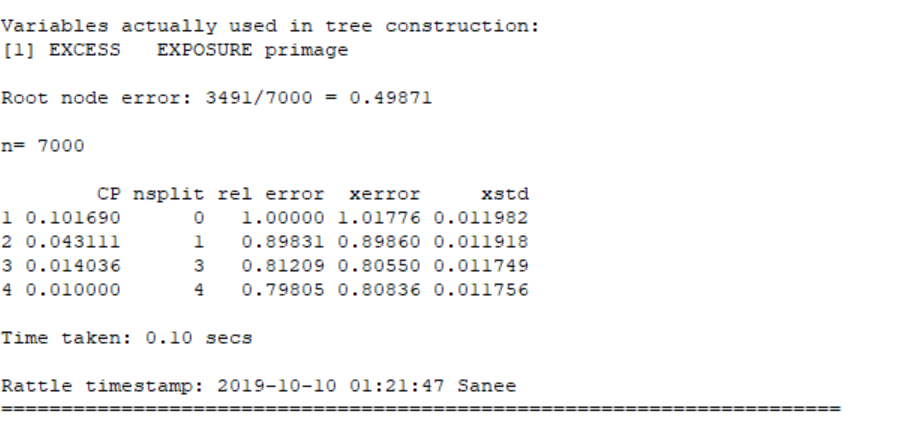
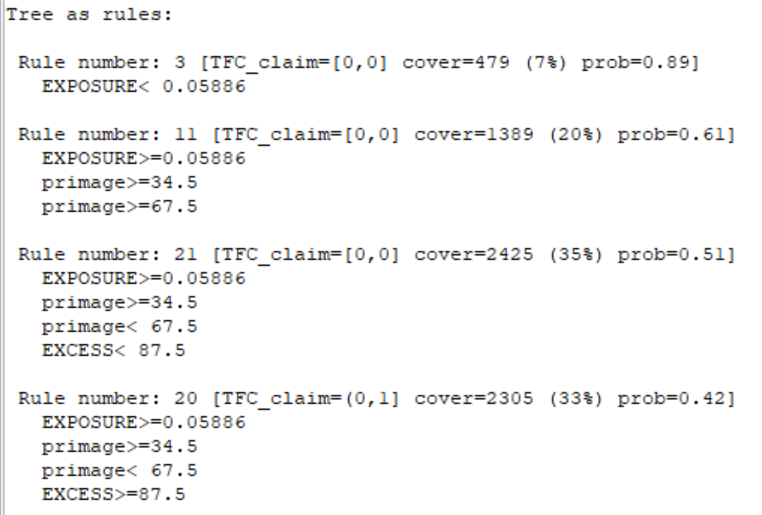
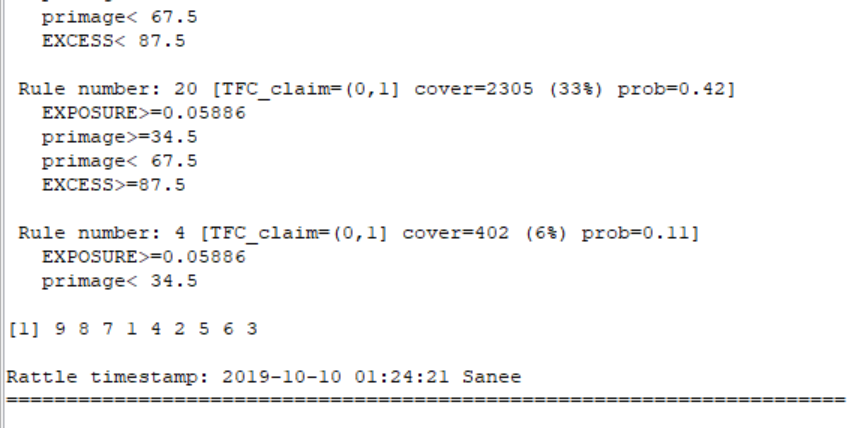
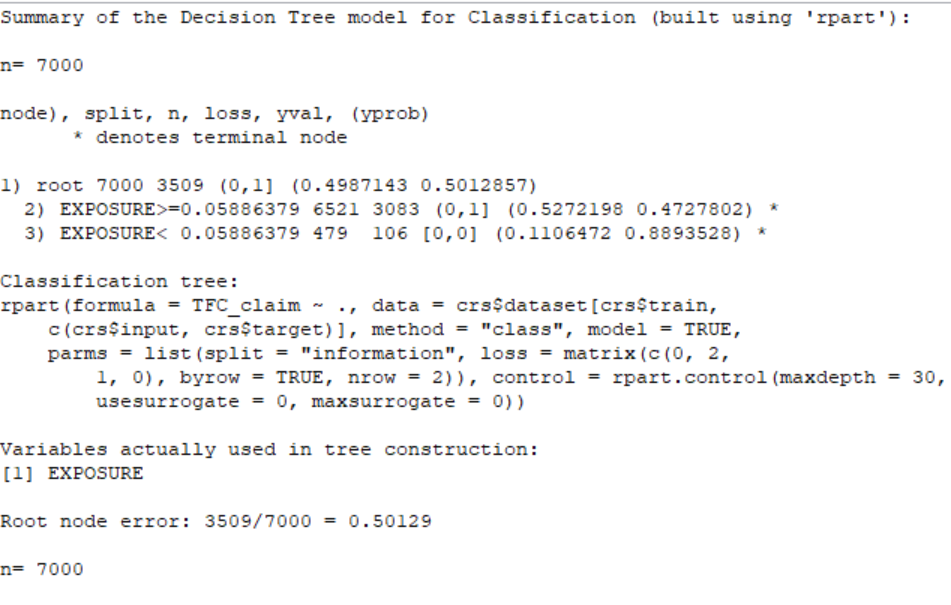
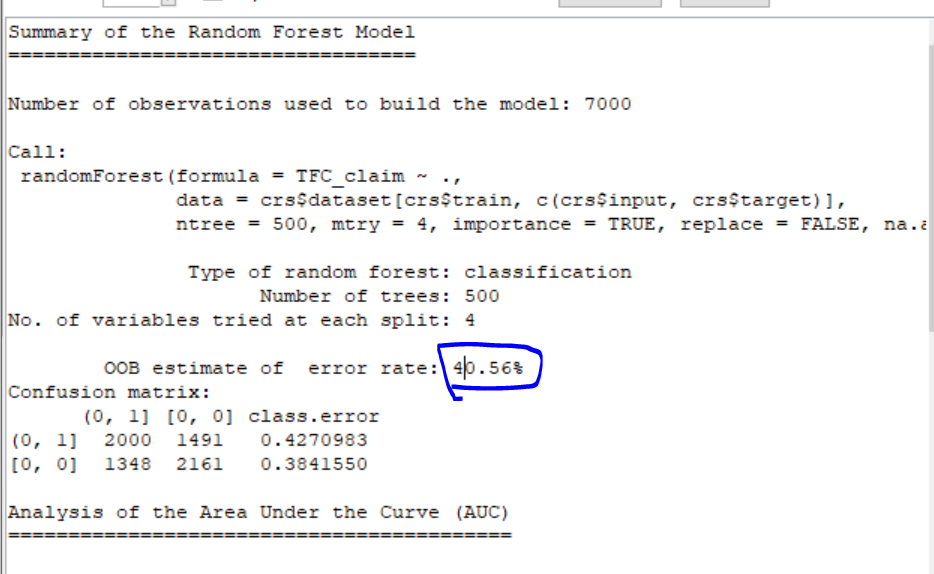
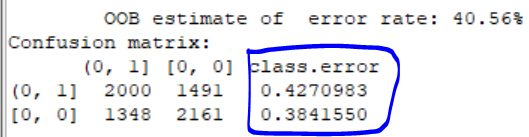
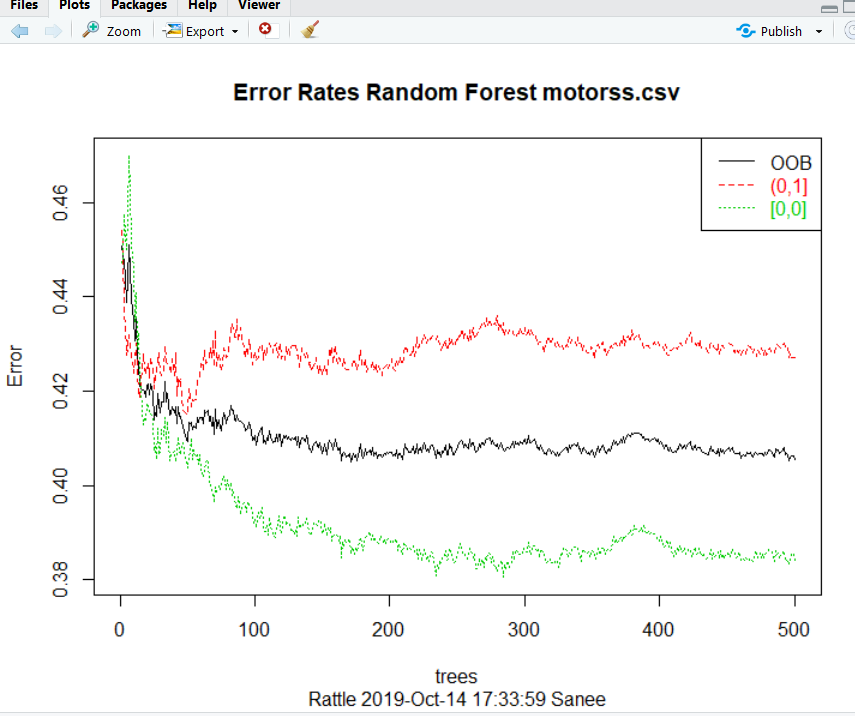
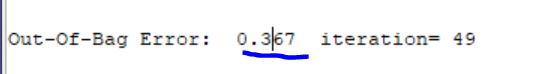
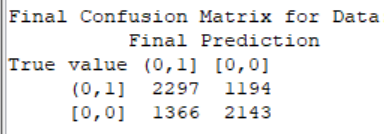
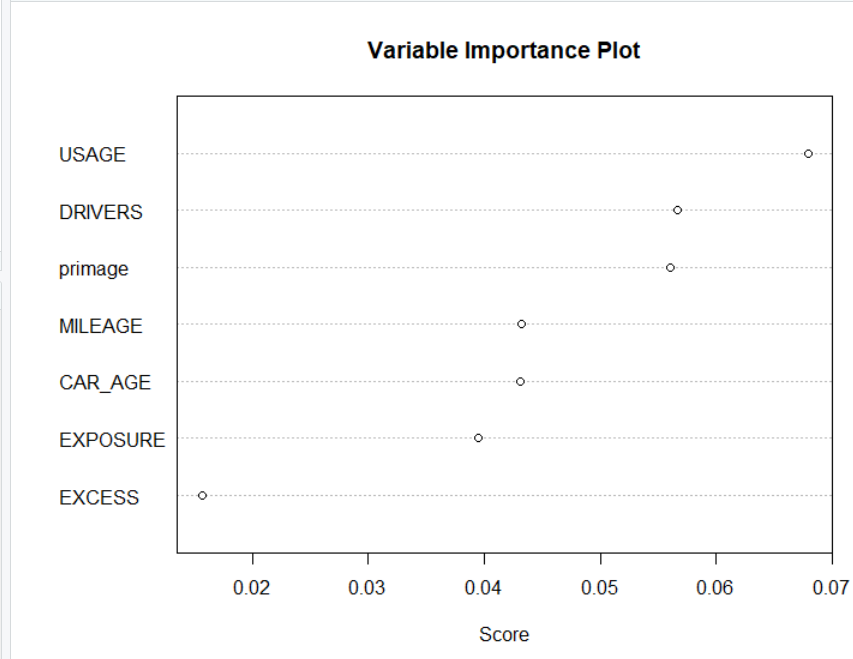
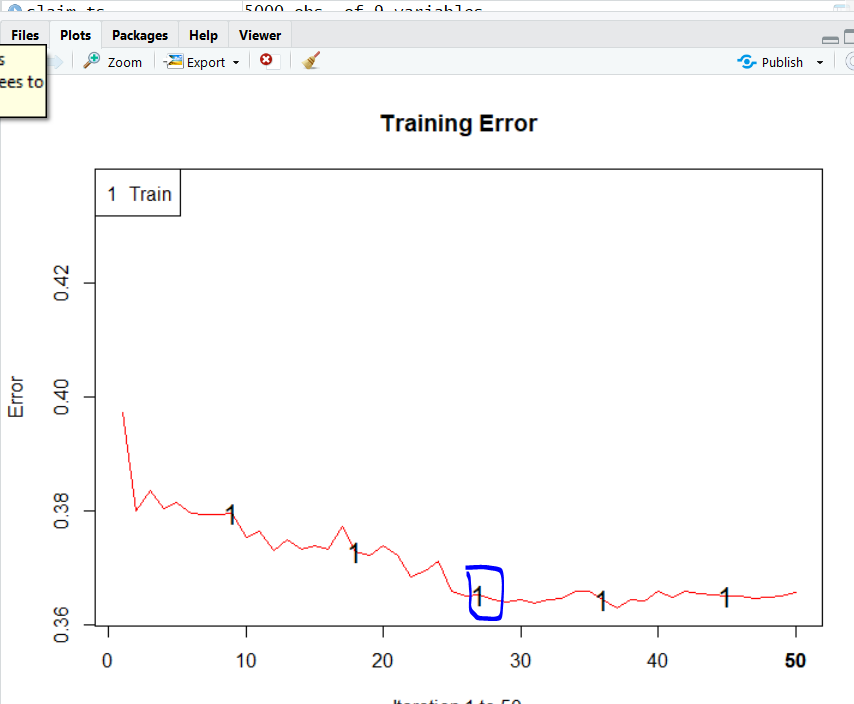
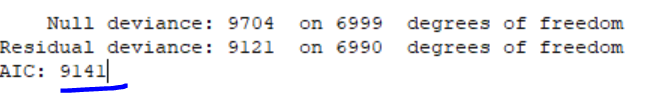
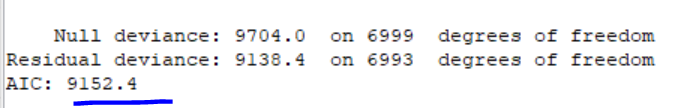
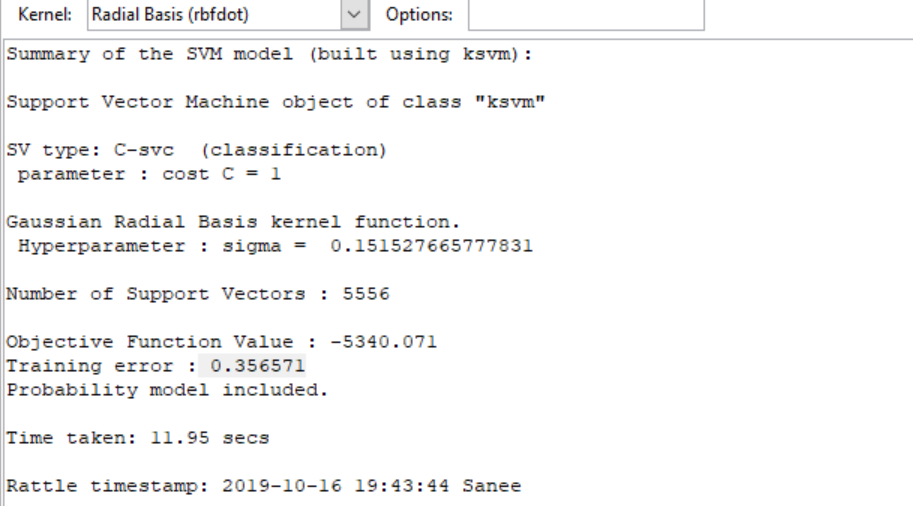
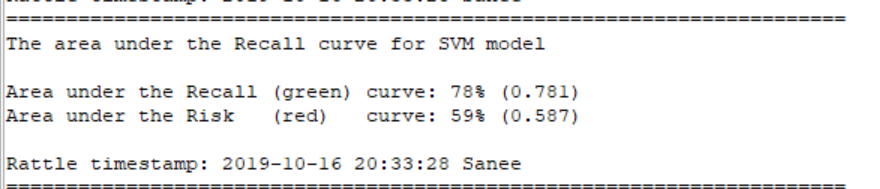
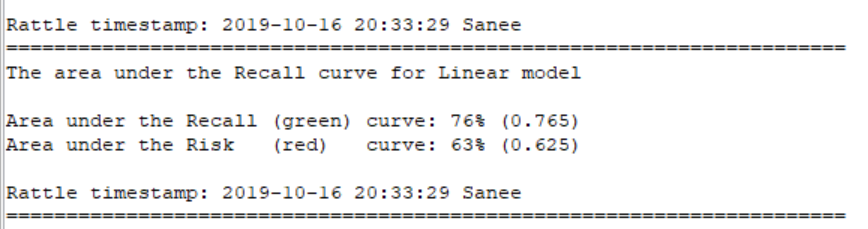
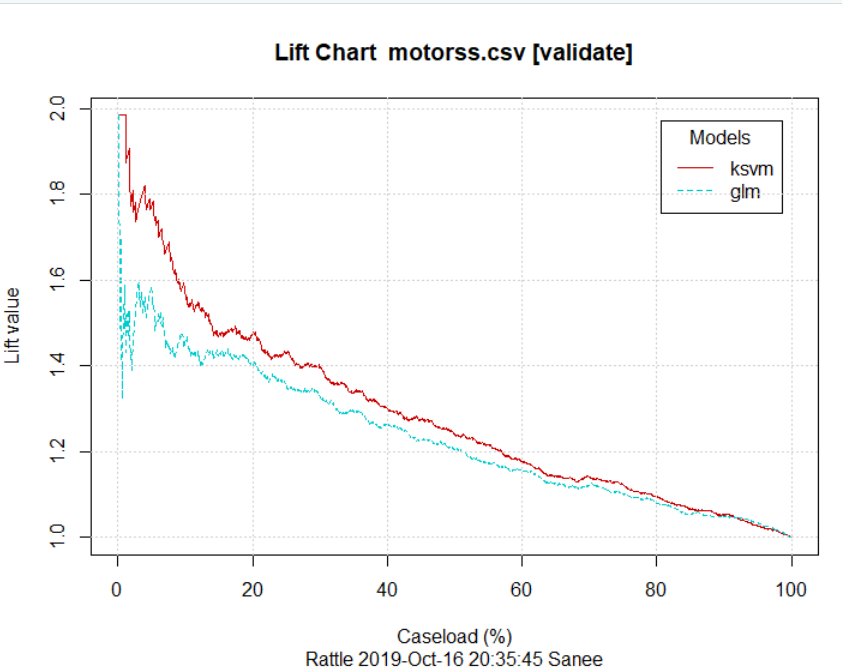
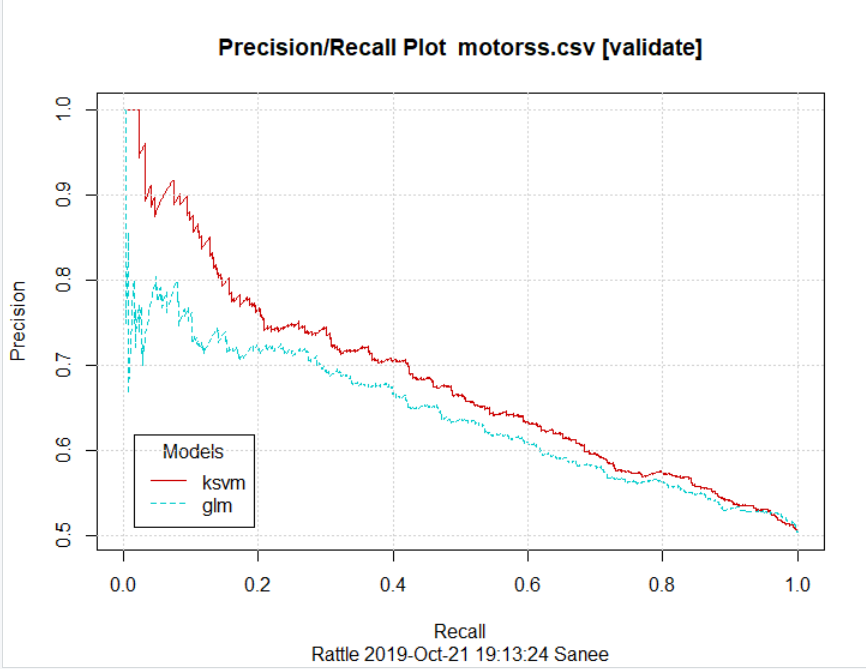
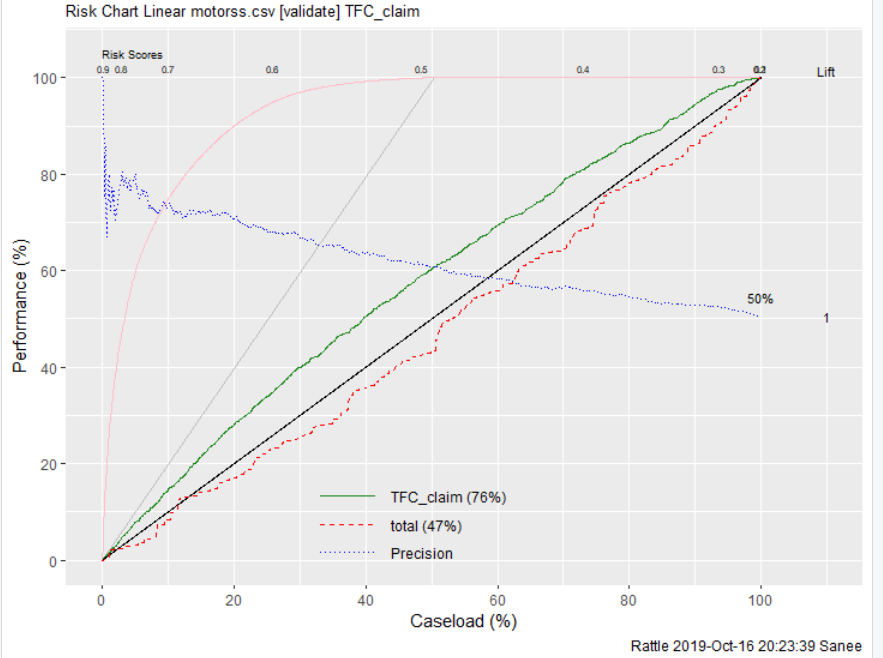
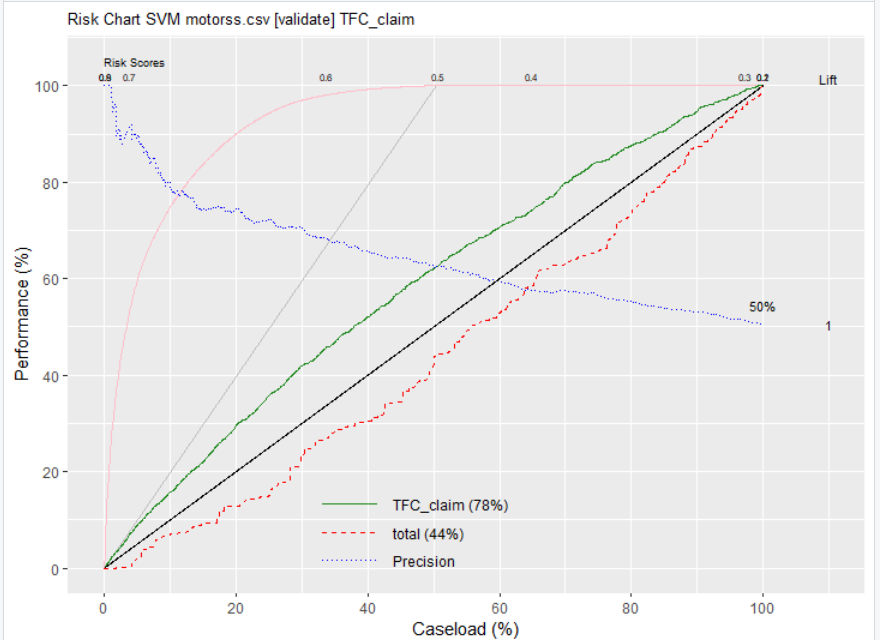
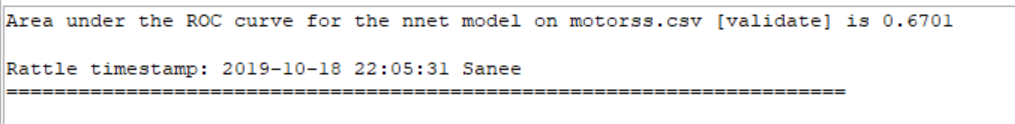
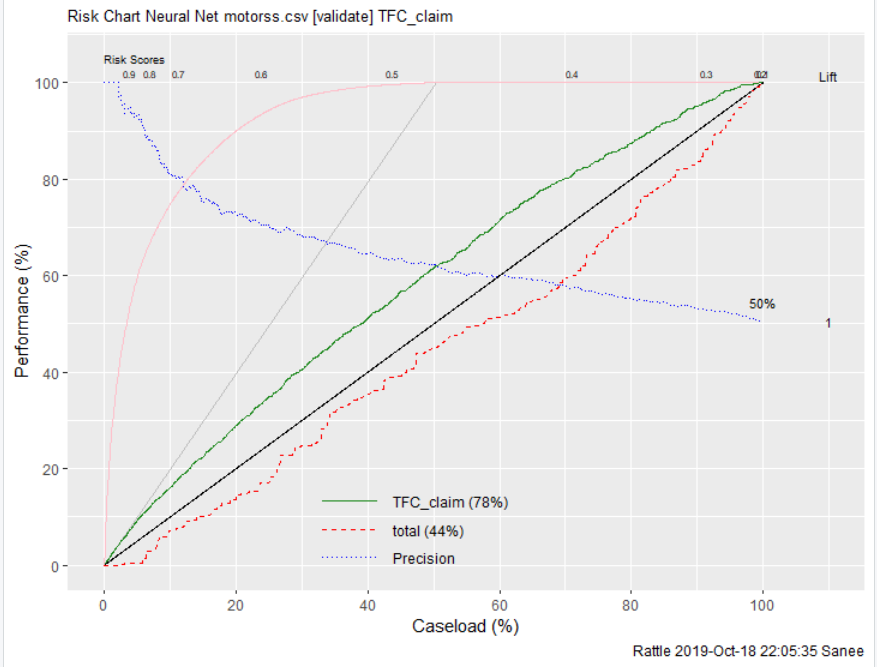
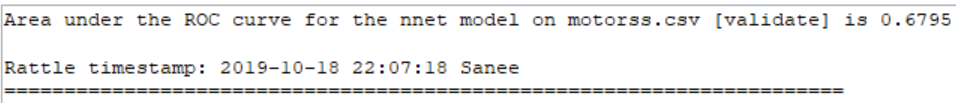
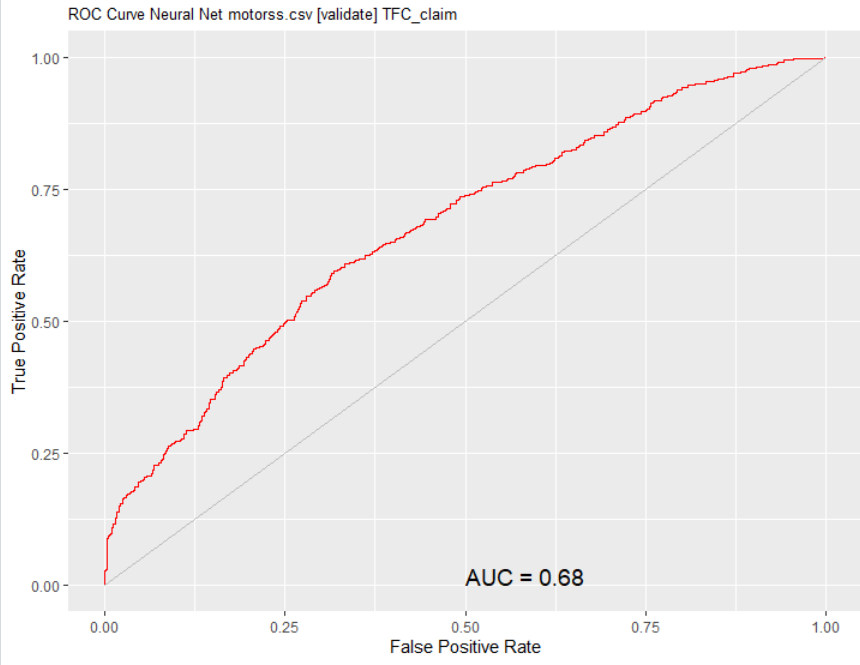
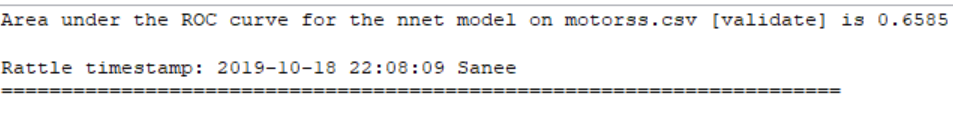
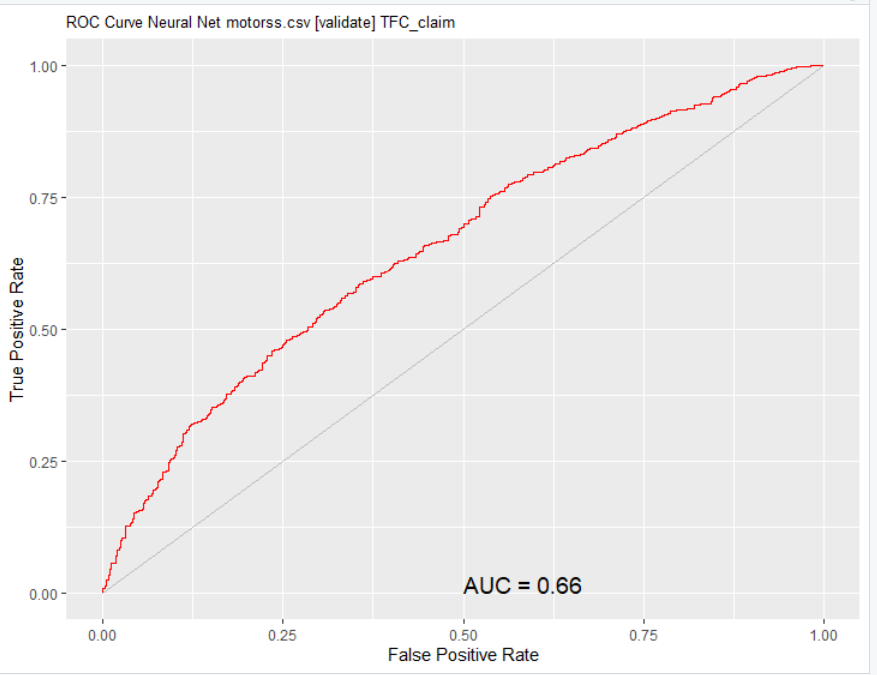
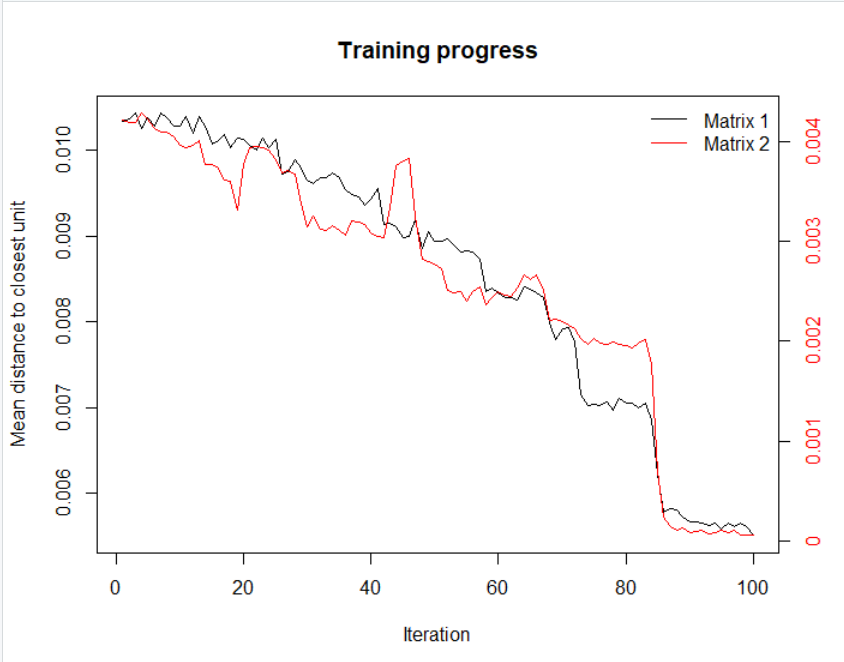
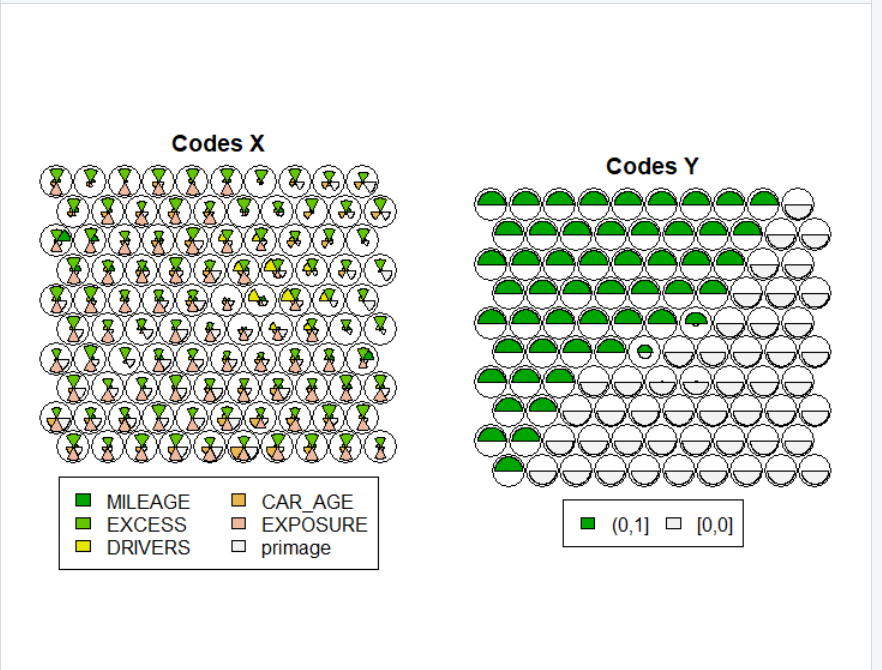
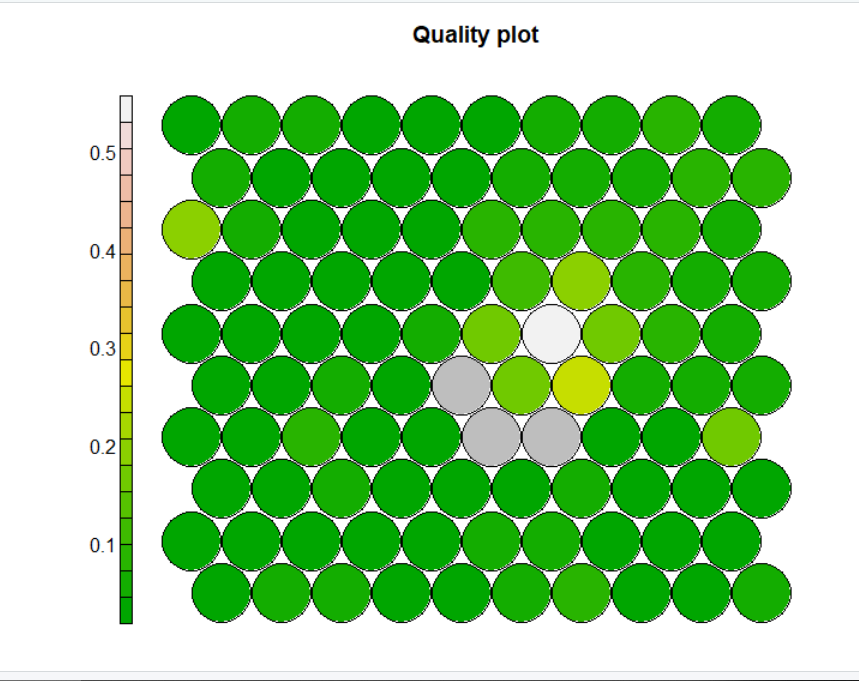
ROC curve is the plot between false positive and true positive rate. ROC itself is the ratio under the curve and the total area. The highest roc curve value in our case is for the boosting ensemble. And lowest is for simple classification tree

Besides these measures, lift, precision and sensitivity charts were also constructed, and all seem to point towards the fact that boosting ensemble is the best classification model for our datasets. (Figure 54,55,56,57,58,59)

# Conclusion:

The aim of this assignment to understand the dataset using descriptive and exploratory data analysis. Upon understanding the dataset, we further investigated the claims and type of claims made for association analysis. Our next objective was to find the best classification model for our dataset. For that reason, the first thing was to use all the types of classification models available in Rattle and conclude the dataset based on that. After that all the models were evaluated to find that the Boosting ensemble was indeed the best classification model for this dataset.

# Appendix

1. 
2. 
3. 
4. 
5. 
6. 
7. 
8. 
9. 
10. 
11. 
12. 
13. 
14. 
15. 
16. 
17. 
18. 
19. 
20. 
21. 
22. 
23. 
24. 
25. 
26. 
27. 
28. 
29. 
30. 
31. 
32. 
33. 
34. 
35. 
36. 
37. 
38. 
39. 
40. 
41. 
42. 
43. 
44. 
45. 
46. 
47. 
48. 
49. 
50. 
51. 
52. 
53. 
54. A close up of a map

    Description automatically generated
55. A close up of a map

    Description automatically generated
56. A close up of a map

    Description automatically generated
57. A close up of a map

    Description automatically generated
58. A close up of a map

    Description automatically generated
59. A close up of a map

    Description automatically generated