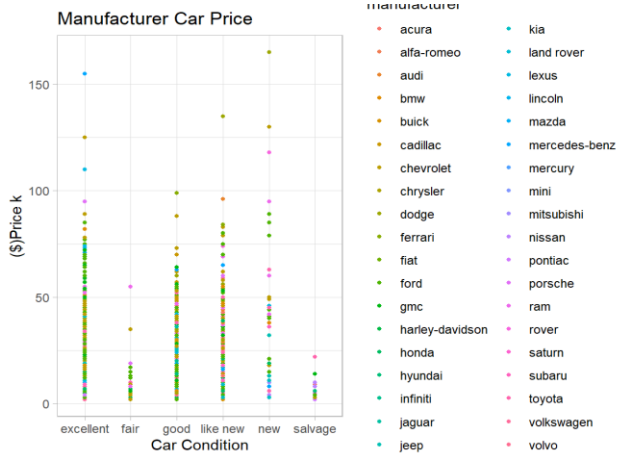
ID5059 Practical 1

Student ID: 210003557 (seed number 3557 in R markdown)

In this practical, I decided to build a random forest model using R because all the data types were strings and not number (even the price) hence I thought it would be better to use a classification method; Also, I printed the scatter plots in R using different variables and I could not find the trend of the data hence I thought linear regression might not work well in this scenario.



I started the process by cleaning the data using select function since only four variables were allowed and some of them were not relevant to my prediction of entry price. I then used filter function to filter out the weird values and low-end outliers because according to my research, the cheapest used cars I found online were around $2,000. Although I split the training data set and testing data set into an 80/20 ratio, the training data set was still too large for my laptop to run; Hence, I decided to make another subset for my training data set.

Next I built a random forest model of 500 trees and the error was massive, so I increased the number of trees to 1000 and found that the lowest error was somewhere around 900. I then tried to find the best mtry (covariates should be considered in each node) and the result was one.

To prove my selection was somewhat ideal, I compared the MSE between random forest with 500 trees (default number), 1000 trees (increased number for better overview), and 900 trees (best result according to the 1000-tree plot) using my test data set as model selection. The result showed that with a random forest with 900 trees using one covariate for each node would give me the best result. For more detailed number and data, please refer to the R markdown file.