**Getting Started With R**

**Introduction**

Transitioning from RapidMiner program into R and RStudio has been a very learned journey till now. Being a new bee to coding in R it has been a little overwhelming at times when errors are encountered in the script/codes. But till now I have been able to successfully overcome these errors with the help of online resources, especially the tutorial provided by the course.

One of the most important lessons that I have learned is the importance of writing codes properly in readable and manageable way. Additionally, I have also gained knowledge about the best practice of managing datasets in respective folders so that the program can by default save data analysis files like R.script, Rhistory, etc., in the respective data folder. This is an important practice to follow or else it will be difficult to reproduce correct project when required. The major difference between R/Rstudio and RapidMiner programs that I learned is that R/RStudio requires “coding/scripts” while RapidMiner does not such specific coding. Its always a good idea to

The first step in learning R is downloading and installing R and RStudio correctly. It was pretty straightforward and easy to install this program on Windows platform. Learning R has been a great learning journey till now with lots of ups and downs. As a novice and beginner in R the only advice I would like my fellow colleagues who are going to learn R in near future is “be patient and all will be good”. Running into errors can lead to frustration but if one can be patient; these errors can be easily eliminated. Also, if one practices variety of datasets in R, more exposure is guaranteed leading to better knowledge. I recommend my fellow peers to start leaning R using the tutorial provided. I have found this tutorial to be highly informative and beneficial in leaning R/RStudio. This tutorial goes over two datasets Cars and Iris that have been analyzed using linear regression model. The main purpose of this tutorial is to learn the basics of analytics and visualization using R.

Figure 1 is a bar graph showing number of products in each product type category. A total of 80 existing products have been analyzed in this study.

1. ***Cars Data Set - Predicting Distance (how far a certain car can travel based on speed)***

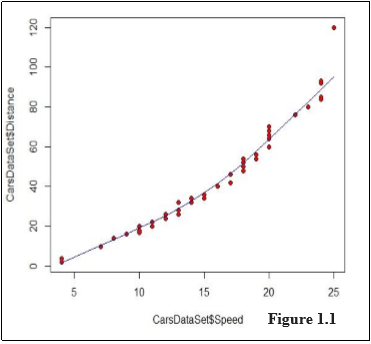
In this dataset the purpose is to predicting distances through the speed of certain cars using simple liner regression model. Therefore, is in this dataset “Speed” is dependent variable and “Distance” is dependent variable.

Figure 1.1 the correlation coefficient is r = 0.807 which for a sample size of 50 indicates a strong positive linear correlation between the two variables.

Table 1.1 shows the distance travelled by cars in dataset based on speed of respective cars.

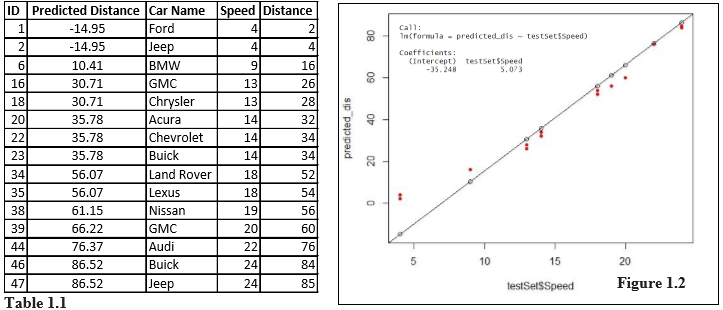
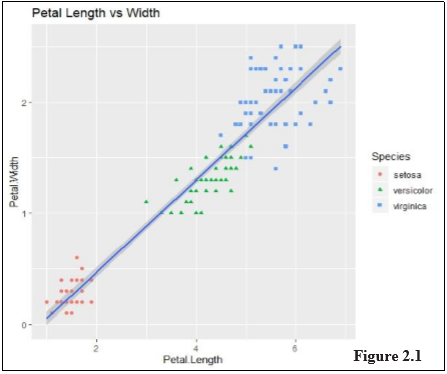


Figure 1.2: the regression equation of the linear model with distance as a response variable (y) and speed as the explanatory variable (x) is y=5.073x – 35.248. This relationship has an r2=0.9225. That means approximately 92.25% of the variation in distance is explained by the speed of the car. ID numbers 1and 2 show negative predicted distances and therefore, are outliers.

**Conclusions**

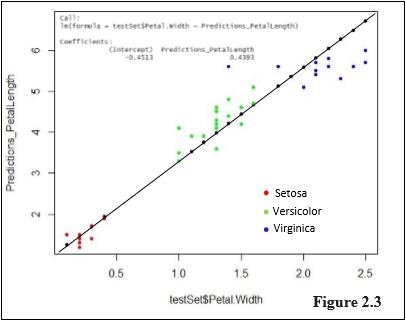
Based on the data used in this analysis, it can be concluded that the distance of a car does depend on the speed the car was traveling. It cannot be ignored that there are few limitations in this conclusion since the distance travelled by a car can also be influenced by weight, height, and length of the respective car.

1. ***Iris Data Set - Predicting Petal Length using Petal Width***

Since Blackwell believes that that best way to learn through trial and error. Therefore, script of code was run for Iris dataset to test the knowledge gained from Cars dataset. The main purpose of this task was to run the script of codes as provided in the tutorial and then record errors in script followed by using correct codes to analyze this Iris dataset and predict Petal Length using Petal Width using simple liner regression model.

Figure 2.1: Slope of line goes upwards implying positive relationship between petal length and petal width. This figure also shows that petal length of the Setosa is clearly a differentiated cluster so it will be a good predictor.

***Results and Conclusions***

Figure 2.3: the regression equation of the linear model with Petal Length as a response variable (y) and Petal Width as the explanatory variable (x) is y=0.4393x – 0.4513. This relationship has an r2=0.927. That means approximately 92.7% of the variation in Petal Length is explained by Petal Width. This implies that model fits the data.

Although it is observed that there is a strong correlation between Petal Length and Petal Width, there can be other factors/variables that can be considered in analysis like age, types of environment, weather conditions, etc.