|  |
| --- |
| Blackwell electronics |

### Kyle gardner

Utilization of R

## Review and Application of R for Data Analytics

|  |  |
| --- | --- |
|  | Introduction to R­ |

|  |  |
| --- | --- |
|  | Prediction: Distance Travelled |

|  |  |
| --- | --- |
|  | Prediction: Length of Petal |

# 

# Introduction

What is R:

R is an emerging programming languages software used in the industry for statistical and data visualization. The purpose of this report is to highlight experiences and best practices for using R and its applicability to be utilized by Blackwell Electronics to perform data analytics.

R is free software developed and maintained by plethora of user within the data community (open-source). One strength of R, is its user community, which allows users to share code, programs, and tactics for solving problems or determining the best analysis for a project. Another strength of R, is the ability to run a line of code and receive an instant output. R studio is a complement to and allows for greater functionality with R.

Unlike Rapidminer, where users drag and clicked to set up and/or run analysis, in R user must manually type out commands to set up and/or run analysis. This can be a bit intimidating for users whom are familiar with friendlier graphical user interfaces. Users have to become accustomed to writing code for variety of task such as, installing & reading packages, uploading and running statistical analysis on data. User should also be aware that in R is case-sensitive and to develop a consistent structure naming process or variables. Packages and extensions do not come with R and must be downloaded separately. However, R provides a general descriptions of all functions used and there are resources on the web to assist new user with R.

1

Tab 1

Tab 2

Predicting Distance Travelled of Cars:

R/RStudio was used to perform exploratory data analysis on a dataset, Cars, of interest to Blackwell Electronics. The car dataset had 50 observations and 3 variables. The datatype format for two of the variables was integers and the last variable was formatted as a factor data.

General statistic summary of car dataset

name.of.car speed.of.car distance.of.car

Dodge : 3 Min. : 4.0 Min. : 2.00

Honda : 3 1st Qu.:12.0 1st Qu.: 26.00

Jeep : 3 Median :15.0 Median : 36.00

KIA : 3 Mean :15.4 Mean : 42.98

Acura : 2 3rd Qu.:19.0 3rd Qu.: 56.00

Audi : 2 Max. :25.0 Max. :120.00

The mean speed of cars was 15.4 mph with ranges between 4 to 25 mph. The mean distance travelled was 42.9 ft, with ranges between 2 to 120 ft. The distribution for distance variable can be described as skewed right where as for speed the data is more normally distributed. Scatterplot appear to shows a relationship that when speed increases so does distance

# Prediction:

# Distance Travelled

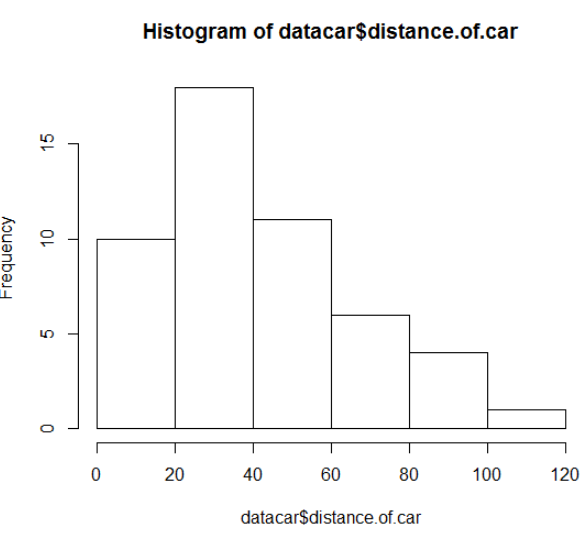
# 

# 

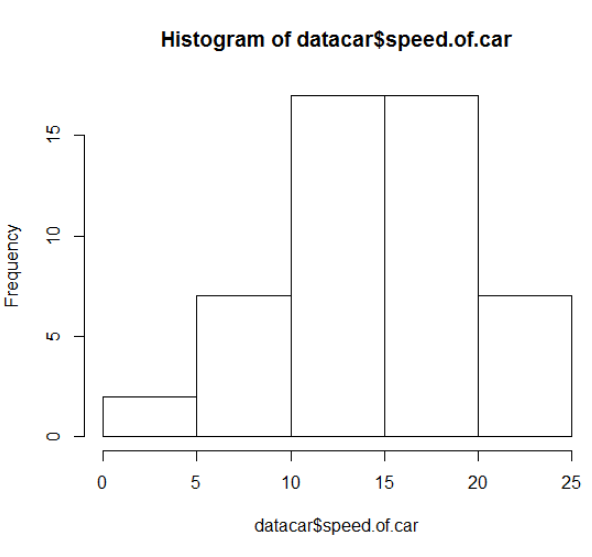
# Histograms Scatterplot

Tab 3

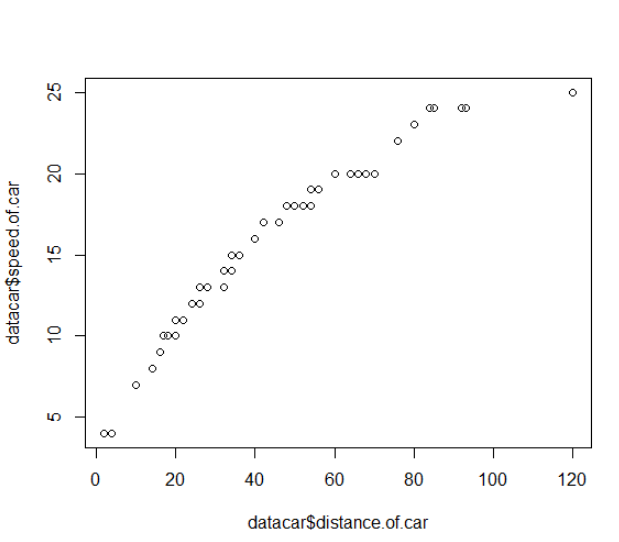
Histogram of distance travelled by car (ft)



Histogram speed of car (mph)



Scatterplot speed (mph) vs distance travelled (ft)



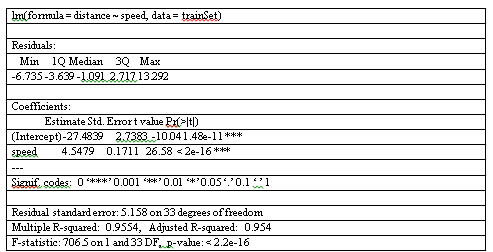
# 

# Prediction model

Tab 4

Regression Analysis:

Regression was recommended for analysis since the car dataset had 3 variables. Linear regression model was used to predict distance travelled (dependent variable) through speed of certain cars (independent variable). Before the model was applied, the dataset was split 70%/30% on size into a training (n=35) and testing (n=15) dataset. Splitting the data into separate datasets allows for assessing the quality and how effective the model is. The linear regression model created from training dataset, appears to be efficient as there is not a large difference between the actual and predicted values for distance (residuals). The R-squared value was 0.955, which shows the regression line fits the data well. The P –value was 2.2e-16 (statistically significant, below 0.05 confidence level), which shows distance is very much dependent on speed.



# 

# Prediction model

Regression Analysis Cont.:

Regression analysis predicted the distance travelled for 15 different cars. One of the Ford (#1) cars had a predicted distance travelled of -9.3 ft. One Buick (#46) and one Dodge (#50) had the highest predicted distance travelled, 86.2 ft and 81.7 ft, respectively. The predicted distance values calculated fall within a 95% confidence interval of lower and upper limit values.

**Predicted distanced travelled (ft) and 95% confidence intervals**

|  |  |  |  |
| --- | --- | --- | --- |
| Car ID | fit | Lower Limit | Upper Limit |
| 1 | -9.29229 | -13.5665 | -5.01804 |
| 5 | 8.899321 | 5.836855 | 11.96179 |
| 12 | 27.09094 | 25.00152 | 29.18035 |
| 13 | 27.09094 | 25.00152 | 29.18035 |
| 20 | 36.18674 | 34.36653 | 38.00696 |
| 21 | 36.18674 | 34.36653 | 38.00696 |
| 26 | 40.73465 | 38.95969 | 42.5096 |
| 28 | 45.28255 | 43.4853 | 47.0798 |
| 29 | 49.83045 | 47.94576 | 51.71514 |
| 33 | 54.37836 | 52.34947 | 56.40724 |
| 21 | 36.18674 | 34.36653 | 38.00696 |
| 26 | 40.73465 | 38.95969 | 42.5096 |
| 28 | 45.28255 | 43.4853 | 47.0798 |
| 29 | 49.83045 | 47.94576 | 51.71514 |
| 50 | 86.21368 | 82.35982 | 90.06754 |
| 35 | 54.37836 | 52.34947 | 56.40724 |

Tab 5

# Prediction:

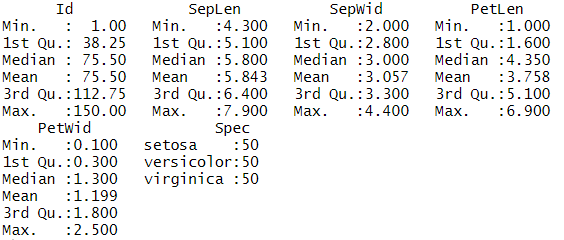
# Petal Length

Tab 6

Predicting growth of petal length:

The iris dataset had 150 observations and 6 variables. The datatype format for four of the variables was integers and the other variables were numeric and factor. There were a few spelling errors identified and fixed before preparing general statistics.

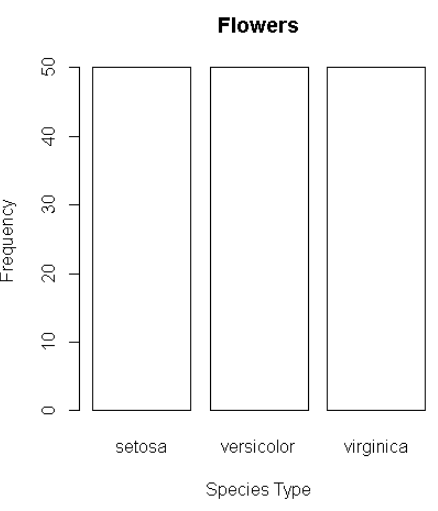
**General statistic summary of iris dataset**



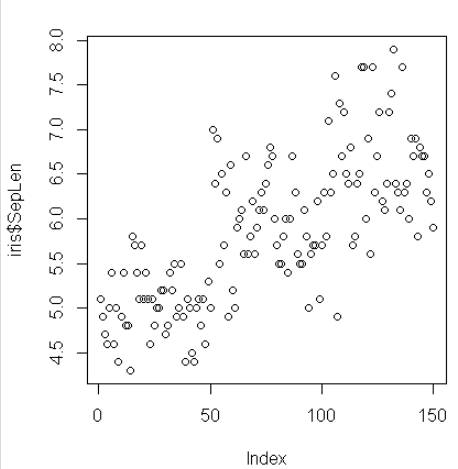
The mean petal length was 3.76 cm with ranges between 1 to 6.9 cm. The mean petal width was 1.20 cm with ranges between 0.1 to 2.5 cm. The flower species were evenly distributed within the dataset. Since species is a categorical variable, bar chart should be used instead of histogram chart.

# BarChart Scatterplot quantile plot

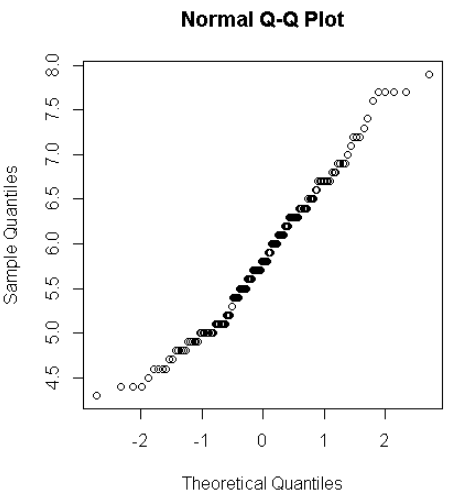
Barchart of flower species



Scatterplot of sepal length



Quantile plot of sepal length

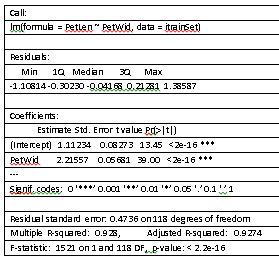


Tab 7

# Prediction model

Regression Analysis:

Linear regression model was used to predict petal length (dependent variable) from petal width (independent variable). Before the model was applied, the dataset was split 80%/20% on size into a training (n=120) and testing (n=20) dataset. Splitting the data into separate datasets allows for assessing the quality and how effective the model is. The linear regression model created from training dataset, appears to be efficient as there is not a large difference between the actual and predicted values for distance (residuals). The R-squared value was 0.928, which shows the regression line fits the data well. The P –value was 2.2e-16 (statistically significant, below 0.05 confidence level), which shows petal length is very much dependent on petal width.



Tab 8

# Prediction Model

Regression Analysis Cont.:

Regression analysis predicted the petal length for 20 different flowers. Virginica were predicted to have the longest petal length, followed by versicolor, than setosa. In the testing dataset, predicted petal length for virginica ranged from 5.1 to 6.4 cm, for versicolor ranged from 3.3 to 4.6 cm, and for setosa ranged for 1.5 to 1.9.

Tab 9