Data Analytics: Predicting Customer Preferences

**COURSE 3 TASK 1 REPORT FOR BLACKWELL ELECTRONICS**

**R and R Studio :**

1. It was pretty straightforward to install and setup. Had to do some digging around to save files in R
2. Learnt R2 only cannot be a factor to determine if a model is fit or not but a combine output from residual plot and R2 to make the decision.

**Report for CARS dataset**

The cars data set had 3 columns . Car name , Speed and Distance. Did not find the necessity to change data type for car name field since we were doing a simple linear regression model to find the distance car travels depending on its speed.

Speed and Distance variables were linearly related. They had positive correlation of 0.9.

Chart, line chart, scatter chart

Description automatically generated

Both variables had normal distribution

Diagram

Description automatically generated with medium confidence

Used lm () model for this data . Find the summary below

Text

Description automatically generatedTable

Description automatically generated with low confidence

R2 was pretty high – 92%. But the Residual vs Fitted value plot had a curve and the points did not seem to be random, which could mean the model does not fit well for this data and we need to choose a different model.

**Report for IRIS dataset**

1. **Errors encountered with data and how it was rectified**
2. # install.packages(readr) Package already installed

# Error in install.packages : object 'readr' not found

library("readr")

1. IrisDataset <- read.csv('iris.csv') # Quotes were missing for file import
2. summary(IrisDataset) # data set name misspelled
3. str(IrisDataset) # data set name misspelled
4. hist(IrisDataset$Species) # Species must be numeric. Did not convert or use this since we were doing a simple linear reg model and this field was not used.
5. qqnorm(IrisDataset) # Missing field name in data set. Added field name
6. set.seed(405) This is not necessary, Since we already had another set.seed() in code
7. trainSetIris <- IrisDataset[training\_indices, ]

# testSetIris <- IrisDataset[-training\_indices, ]# this is not necessary since it was duplicated.

1. LinearModel<- lm(Petal.Width ~ Petal.Length, trainSetIris)

# LinearModel<- lm(trainSet$Petal.Length~trainSet$Petal.Width)

# Statement had testset.

# y~x variables were switched

1. **Model Analysis for IRIS data**

The cars data set had 6 columns . Out primary focus was on Peta length and petal width t build a linear model.

Text

Description automatically generated

Petal Length and Petal Width did not have a linear relation but had a correlation of 96%

Chart, scatter chart

Description automatically generated

Used lm () model for this data . Find the summary below **Table

Description automatically generated**

**A picture containing chart

Description automatically generated**

R2 value seems to be very low (0.15) indicating a very weak

Residual plot seems to have high variance between 0.2 ~ 0.35 .

there is not randomness in the residual plot which indicates a bad fit.