|  |  |
| --- | --- |
| To: | Danielle; CTO; Blackwell |
| From: | John Enrietto |
| Date: | 5/25/22 |
| Subject: | Request for analysis and prediction of cars and Iris database |

Per Danielle’s request, two different data files were analyzed. These files were supplied in a zip file titled *R Tutorial Data Sets*. Two files were downloaded: Cars.cvs and Iris.cvs. These files were analyzed in R, using R Studio. Two regression analysis were developed and data listed below.

An R LinearModel (LM) was developed to analysis Cars.csv. The LM output is listed below. A graph is also copied to show actual data vs predict data vs the linear regression line from lm. The lm resulted in an Adjusted R-Squared value of 0.9276. This is reasonable, and as shown in the chart, the predicted values match what looks like the average of the actual values, except at very high and very low values. The low p-value (<.05) this analysis shows good correlation.

|  |
| --- |
| Call:  lm(formula = distance ~ speed, data = car\_df)  Residuals:  Min 1Q Median 3Q Max  -8.493 -4.472 -1.102 1.929 31.939  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -29.3371 3.0606 -9.586 9.94e-13 \*\*\*  speed 4.6959 0.1882 24.956 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 6.965 on 48 degrees of freedom  Multiple R-squared: 0.9284, Adjusted R-squared: 0.927  F-statistic: 622.8 on 1 and 48 DF, p-value: < 2.2e-16 |

The same can be seen in the table listed below, which show testSet data, then a 4th calculated column (predict) using the Y int and slope of the lm analysis. You can see data does not match very tightly at the two ends of the data.

|  |  |  |  |
| --- | --- | --- | --- |
| testSet data | | | Y=4.66 X - 29.33 |
| Index | speed | distance | Distance (predict) |
| 1 | 4 | 2 | -10.6 |
| 3 | 7 | 10 | 3.5 |
| 8 | 10 | 18 | 17.6 |
| 14 | 12 | 26 | 27.0 |
| 16 | 13 | 26 | 31.7 |
| 17 | 13 | 28 | 31.7 |
| 22 | 14 | 34 | 36.4 |
| 33 | 18 | 50 | 55.2 |
| 38 | 19 | 56 | 59.9 |
| 49 | 24 | 93 | 83.4 |

Iris was evaluated next. Supplied code was run and analyzed. The table at the end of this report shows initial code, and corrections made to produce workable code in R. Initial Analysis was to determine if there was any difference in the average size of each of the flower sizes, based on the different species. Table below lists the average measurements of the different varieties.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Average Flower Size** | | | | |
|  | Species | Sep\_Len\_avg | Sep\_Wid\_avg | Pet\_Len\_avg | Pet\_Wid\_avg |
|  | setosa | 5.006 | 3.428 | 1.462 | 0.246 |
|  | versicolor | 5.936 | 2.77 | 4.26 | 1.326 |
|  | virginica | 6.588 | 2.974 | 5.552 | 2.026 |

Like the cars file, Iris data was also analyzed using an R – LM in an attempt to develop a model to predict a petal's length using the petal’s width as an independent variable. Although the three different species show some differences, all three do follow the same trend line as shown below in the visualization. The LM output is listed below and a graph is shown to give a visualization of the predicted data fit.

With the Linear Model, our objective is to get a linear fit of the data using a y-mx + b line. The coefficients give us this best fit model, with the intercept of our model being 1.08246, and the sloe of the line is 2.22203. With a large t value, and extremely low Pr, we can be confident this data will accurately represent the actual plant data. The residual standard error of .5064 says that although the line may fit well on average, any individual point may be off by 0.5064 on average, With an average petal length of 2.2, being off by 0.5, may be considered larger than desired. Overall though, with the .92 R-Squared value, the linear projection does a decent job of fitting our projection to the data. This can be seen in the curve below.

|  |
| --- |
| Call:  lm(formula = Petal.Length ~ Petal.Width, data = trainSet)  Residuals:  Min 1Q Median 3Q Max  -1.31533 -0.32661 -0.02686 0.27611 1.40670  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.08246 0.08689 12.46 <2e-16 \*\*\*  Petal.Width 2.22203 0.05994 37.07 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.5064 on 118 degrees of freedom  Multiple R-squared: 0.9209, Adjusted R-squared: 0.9203  F-statistic: 1374 on 1 and 118 DF, p-value: < 2.2e-16 |

Questions were raised about the effectiveness of R and R studio. Overall the editing capabilities were very good, but with caution. Multiple script windows open at the same time can cause errors and confusion. If variables are shared between two script windows, and not refreshed when switching between windows, it can cause confusion and inaccurate results. This needs careful attention by the user. There were some questions that needed raised during the programming, and it showed confusion on my part between different information sites which needed raised to other group members to resolve. This was confusing at times, but typical with open source type code instructions.

|  |  |
| --- | --- |
| **Supplied code analysis and corrections**. | |
| Original Code | Corrected Code |
| install.packages(readr) |  |
| library("readr") |  |
| IrisDataset <- read.csv("C:\\Users\\johne\\Documents\\  Purdue DA\\C3 rstudio\\RTutorialDataSets\\iris.csv") | Updated for local drive location |
| attributes(IrisDataset) | OK |
| summary(risDataset) | Spelling(I) |
| str(IrisDatasets) | Spelling (s) |
| names(IrisDataset) | OK |
| hist(IrisDataset$Species) | Variable (species) must be numeric. Commented out |
| plot(IrisDataset$Sepal.Length | Missing parentheses |
| qqnorm(IrisDataset) | Must be numeric variable(s). … qqnorm(IrisDataset$Sepal.Length) |
| IrisDataset$Species<- as.numeric(IrisDataset$Species) | Warning message:  NAs introduced by coercion  “Species” is string, cannot be converted to numeric. Add categorical variable  IrisDataset$Species <- as.factor(ifelse(IrisDataset$Species == "setosa", 1,  ifelse(IrisDataset$Species == "versicolor", 2,  ifelse(IrisDataset$Species == "virginica", 3, 0)))) |
| set.seed(123) | OK |
| trainSize <- round(nrow(IrisDataset) \* 0.2) | Train should be ~.7 - .8 |
| testSize <- nrow(IrisDataset) - trainSet | Trainset is not correct variable: Train Size |
| trainSizes | Spelling (s) |
| testSize | OK |
| trainSet <- IrisDataset[training\_indices, ] | Missing line to assign training\_indices, inserted. TrainSet OK |
| testSet <- IrisDataset[-training\_indices, ] | OK |
| set.seed(405) | Commented out. Duplicate analysis, run second set before reassigning |
| trainSet <- IrisDataset[training\_indices, ] | Commented out. Duplicate analysis, run second set before reassigning |
| testSet <- IrisDataset[-training\_indices, ] | Commented out. Duplicate analysis, run second set before reassigning |
| LinearModel<- lm(trainSet$Petal.Width ~ testingSet$Petal.Length) | Wrong variable: testingSet$... changed to trainSet$... |
| summary(LinearModel) | OK |
| prediction<-predict(LinearModeltestSet) | Missing comma |
| predictions | Spelling (s) |