Data Mining and Discovery

Assignment 1: Data Preprocessing

Student ID: 20070328

Introduction:

The aim of the report is to check outliers and check normalization. If outliers exist perform different task in the built-in R dataset.

Dataset Summarization:

The dataset used in this assignment is iris as it has continuous values which was the requirements for this assignment. It has 4 columns which represents the length and width of sepal and petal of different species.

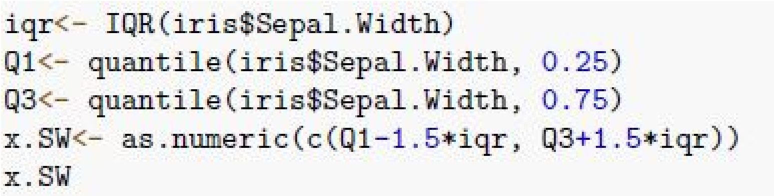
Using summary

summary(iris)

summary is used to describe brief information for all the columns in the dataset. Here, the columns are sepal length, sepal width, petal length, petal width. Summary () gives the minimum, maximum, mean, Q1, Q3, median of those columns.

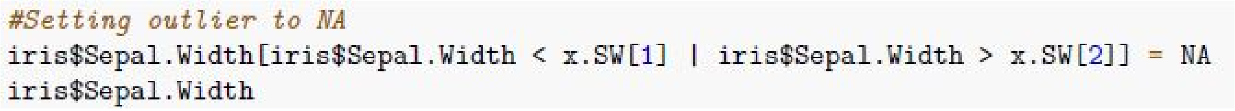
Detecting Outliers and Converting to NA:

The IQR rule is applied after finding the first interquartile range (IQR) and then calculating the limits (Q1 – 1.5IQR, Q3 + 1.5IQR) to each column. The R code for the following is,



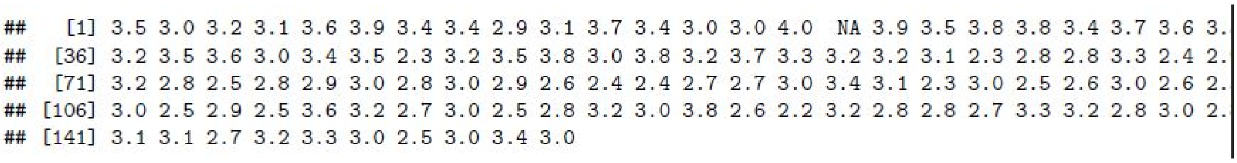
```
iqr<- IQR(iris$Sepal.Width)
Q1<- quantile(iris$Sepal.Width, 0.25)
Q3<- quantile(iris$Sepal.Width, 0.75)
x.SW<- as.numeric(c(Q1-1.5*iqr, Q3+1.5*iqr))
x.SW
```

After observing sepal width column has outliers and the rest of continuous columns did not have any outliers. The outliers in Sepal.Width column were changed to NA. Below is the following snippet.



```
#Setting outlier to NA
iris$Sepal.Width[iris$Sepal.Width < x.SW[1] | iris$Sepal.Width > x.SW[2]] = NA
iris$Sepal.Width
```

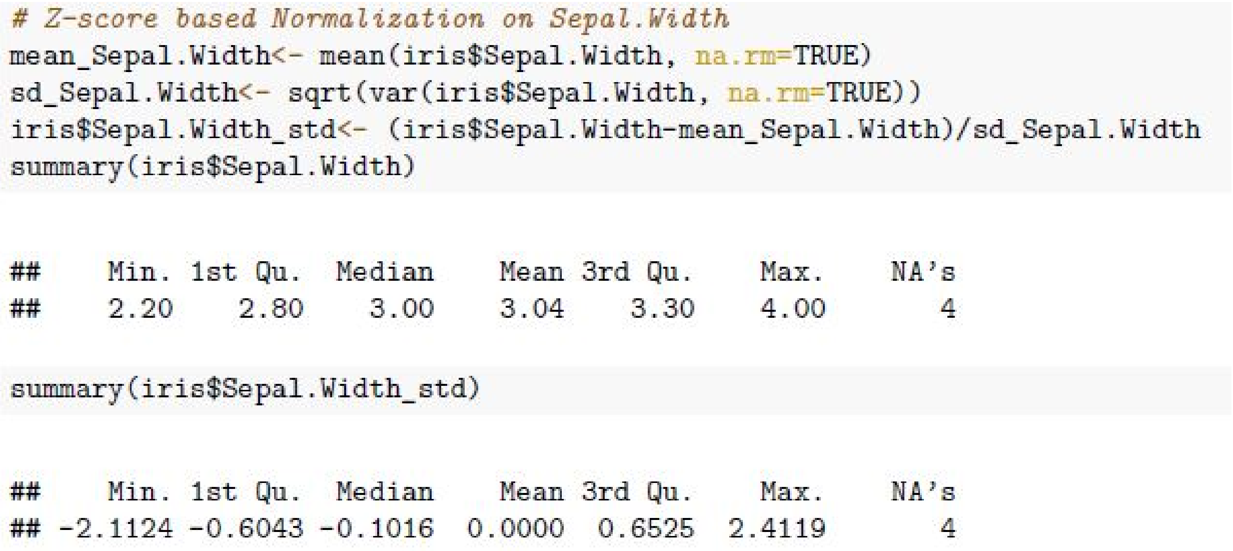
Below is the output for values that have been changed to NA



```
##   [1] 3.5 3.0 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 3.7 3.4 3.0 3.0 4.0  NA 3.9 3.5 3.8 3.8 3.4 3.7 3.6 3.
##  [36] 3.2 3.5 3.6 3.0 3.4 3.5 2.3 3.2 3.5 3.8 3.0 3.8 3.2 3.7 3.3 3.2 3.2 3.1 2.3 2.8 2.8 3.3 2.4 2.
##  [71] 3.2 2.8 2.5 2.8 2.9 3.0 2.8 3.0 2.9 2.6 2.4 2.4 2.7 2.7 3.0 3.4 3.1 2.3 3.0 2.5 2.6 3.0 2.6 2.
## [106] 3.0 2.5 2.9 2.5 3.6 3.2 2.7 3.0 2.5 2.8 3.2 3.0 3.8 2.6 2.2 3.2 2.8 2.8 2.7 3.3 3.2 2.8 3.0 2.
## [141] 3.1 3.1 2.7 3.2 3.3 3.0 2.5 3.0 3.4 3.0
```

Normalizing the data.

After setting the value to NA we used Z-Score to normalize the data for the continuous columns. After applying summary function to the normalization, the mean becomes 0, the variance was 1 and 1st and 3rd quartile were approximately -z to 1. Below is the following code snippet,



```
# Z-score based Normalization on Sepal.Width
mean_Sepal.Width<- mean(iris$Sepal.Width, na.rm=TRUE)
sd_Sepal.Width<- sqrt(var(iris$Sepal.Width, na.rm=TRUE))
iris$Sepal.Width_std<- (iris$Sepal.Width-mean_Sepal.Width)/sd_Sepal.Width
summary(iris$Sepal.Width)
```

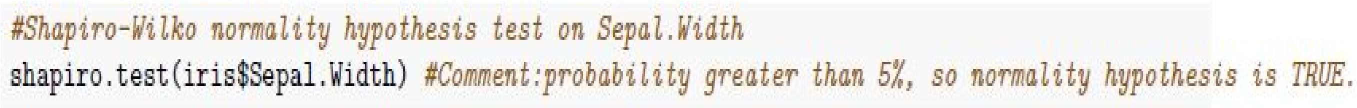
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.20    2.80    3.00    3.04    3.30    4.00       4
```

```
summary(iris$Sepal.Width_std)
```

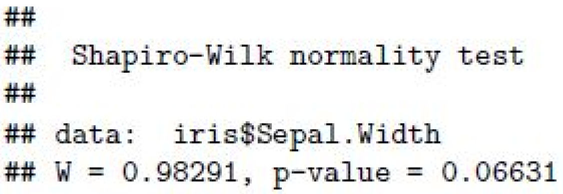
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -2.1124 -0.6043 -0.1016  0.0000  0.6525  2.4119       4
```

Using Shapiro-wilk test

Applying the Shapiro-Wilk test, the null hypothesis (if P < 0.05, then reject the hypothesis) was rejected for each column except the sepal Width whose P value was 0.06631. Below is the following code snippet.



```
#Shapiro-Wilko normality hypothesis test on Sepal.Width
shapiro.test(iris$Sepal.Width) #Comment:probability greater than 5%, so normality hypothesis is TRUE.
```

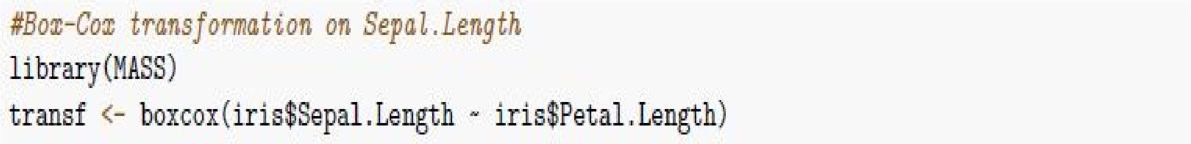


```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Sepal.Width
## W = 0.98291, p-value = 0.06631
```

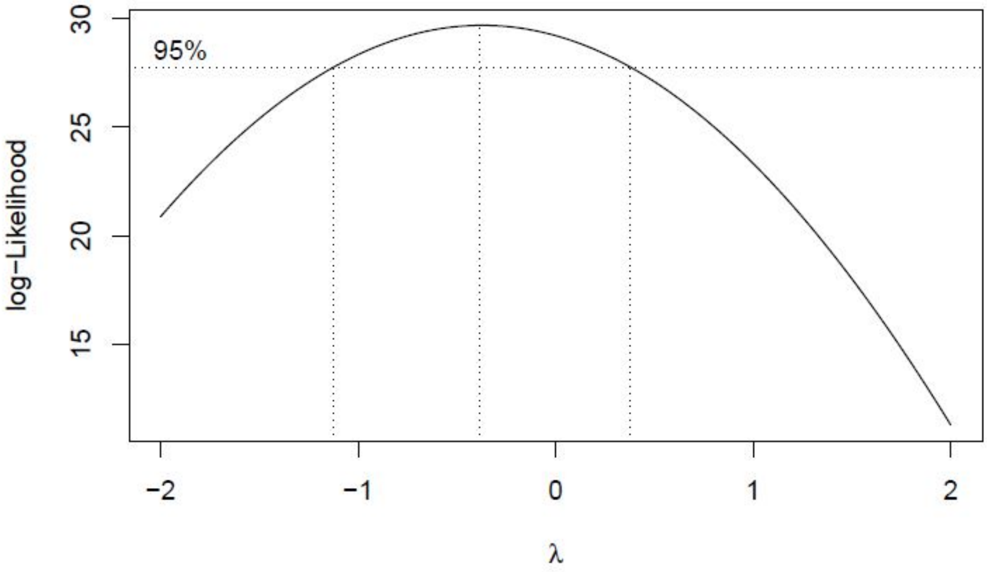
As the rest of the column has p-value is less than 0.05 we have to apply for Box-Cox transformation as per the assignment requirements on the speal.length, petal_length, and petal.width.

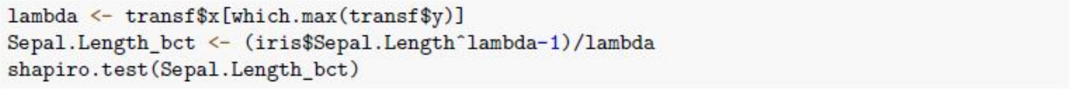
The Box-Cox transformation is applied by calculating **lambda value** and transforming the column using the formula (y^lambda – 1/lambda).

This transformation was applied to the **sepal length** column of the IRIS dataset and QQ plot was plotted for original and transformed data. Below is the following code snippet for the box-cox,



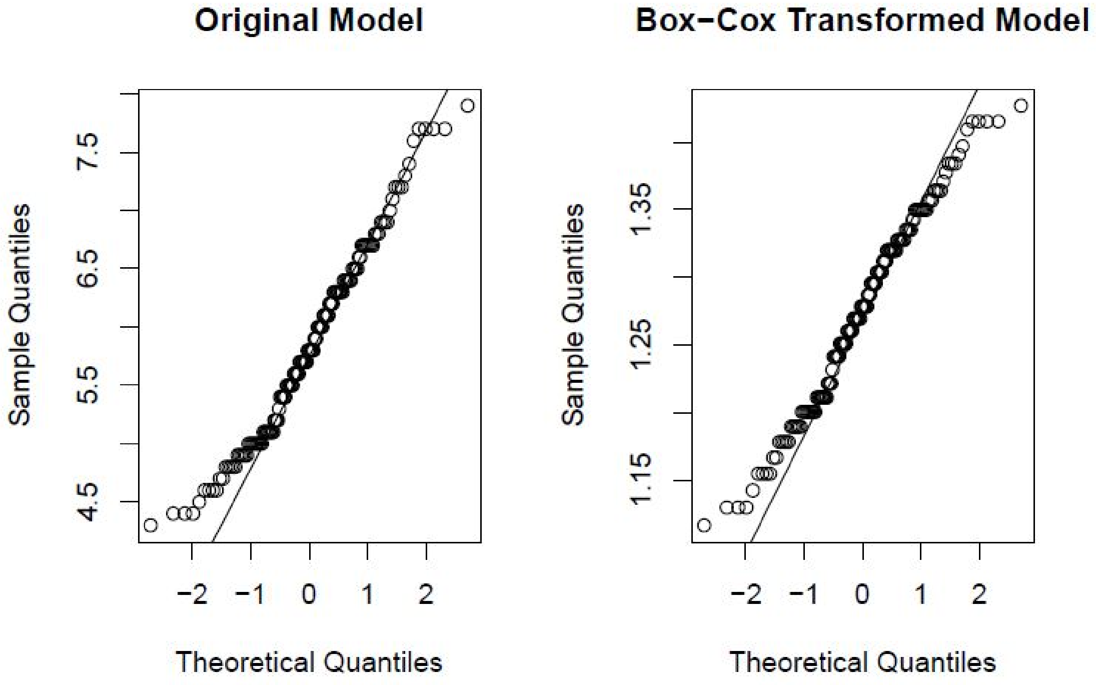
```
#Box-Cox transformation on Sepal.Length
library(MASS)
transf <- boxcox(iris$Sepal.Length ~ iris$Petal.Length)
```





```
lambda <- transf$x[which.max(transf$y)]
Sepal.Length_bct <- (iris$Sepal.Length^lambda-1)/lambda
shapiro.test(Sepal.Length_bct)
```

Below is the original mode and the box-cox transformed model of the iris dataset.


Original Model
Box−Cox Transformed Model
Sample Quantiles
Theoretical Quantiles