KENTUCKY STATE UNIVERSITY, COMPUTER SCIENCE

Data Mining
HW1

Anil Varma B

September 15, 2019

1. Use plot to identify the strongest linear relationship among 4 attributes. Explain whether this is a positive relation or negative relation.

We have plotted a scatter plot to observe the linear relationship between attributes, the strongest relationship as observed from the below plot is between petal length and petal width. The next strongest relation is between petal length and sepal length.

```
plot(iris[-5])
```

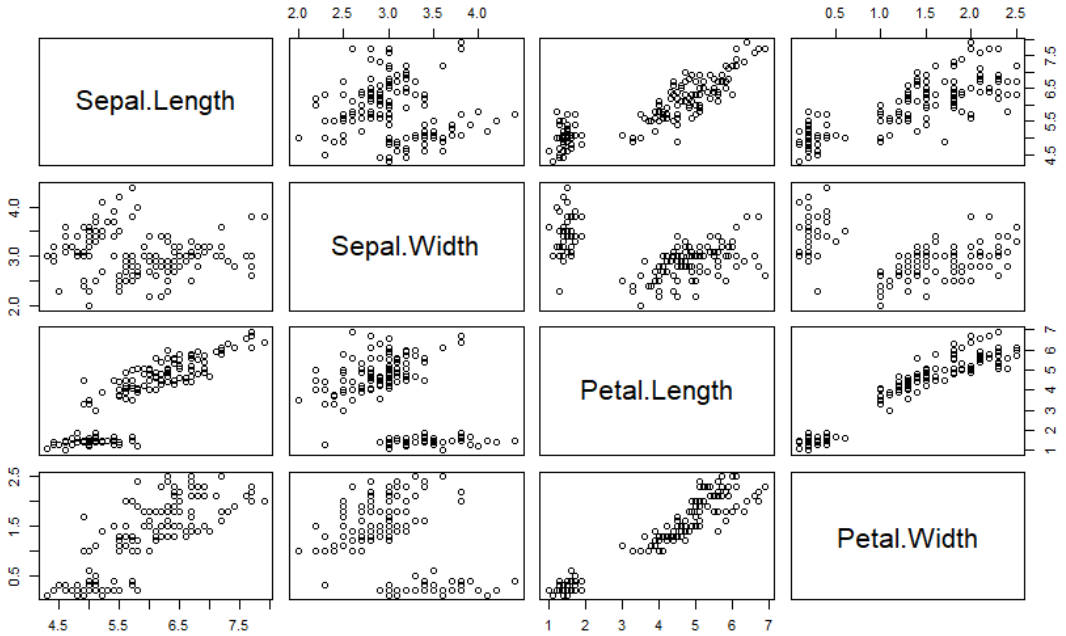


Figure 0.1: Matrix of scatter Plot.

Below we find the correlation values between attributes to find the positive or negative relation, and the strongest relationship is positive.

```
cor(iris[-5])
```

```
> cor(iris[-5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
> plot(iris[-5])
```

Figure 0.2: Correlation of iris data.

2. Draw a side by side boxplot for all 4 attributes and explain their possible outliers

```
boxplot(iris[-5])
```

An outlier is defined as a data point that is located outside the fences ("whiskers") of the boxplot (e.g: outside 1.5 times the interquartile range above the upper quartile and bellow the lower quartile). The same can be observed in Sepal Width.

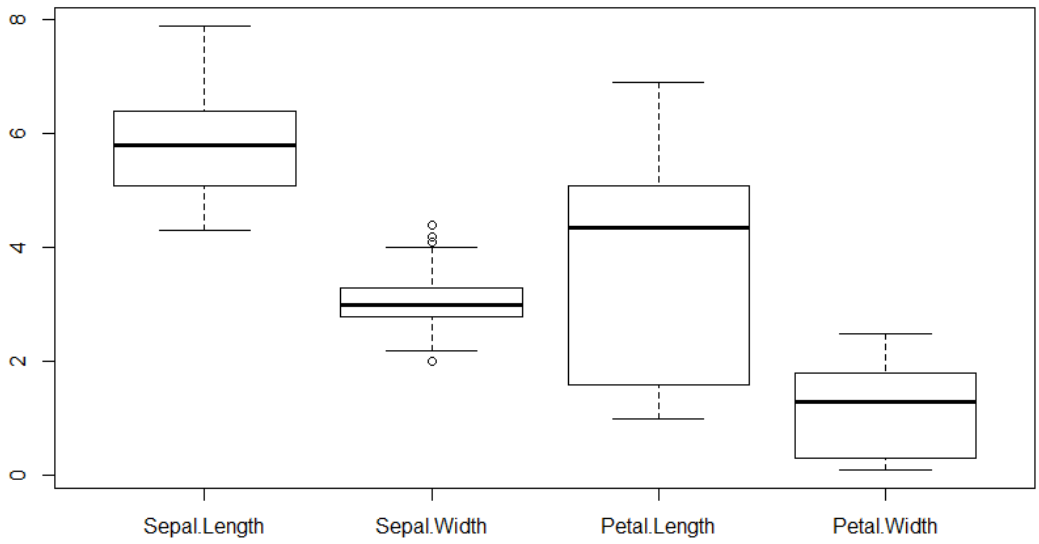


Figure 0.3: Box Plot Comparision of Iris data set.

3. Find mean values for all 4 attributes. Then compare their medians and explain if one of them is normally distributed.

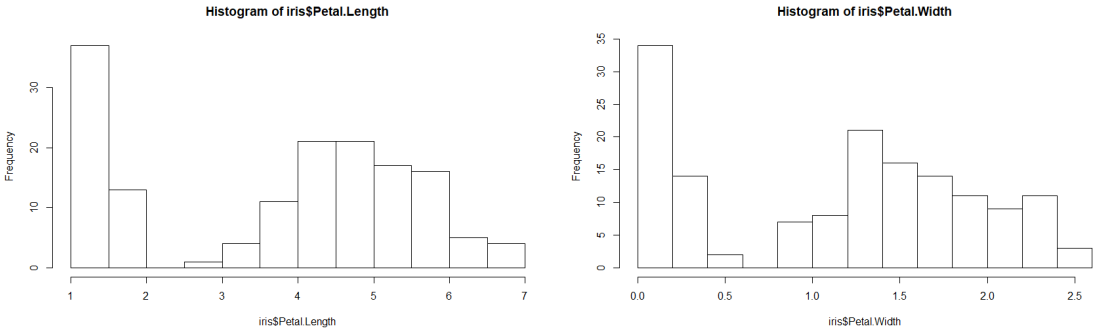
```
summary(iris)
```

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
```

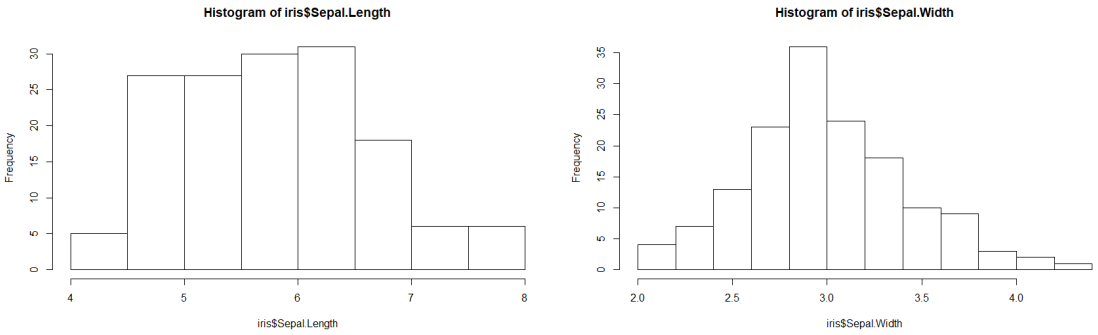
Figure 0.4: Mean and Median in Iris data set.

```
hist(iris$Petal.Length)
hist(iris$Petal.Width)
hist(iris$Sepal.Length)
hist(iris$Sepal.Width)
```

Plotting histogram to see the distribution of data, we can see that Sepal Width column is normally distributed.



(a) Petal Length and Width Histogram



(b) Sepal Length and Width Histogram

5. Use attached churn_1.txt a. Explore whether there are missing values for any of the variables. If they are, correct them by replacing with 0, mean and minimum values. There are more than 3 missing values. You need to replace one of them with 0, the other with mean, and the third one with minimum value of the column.

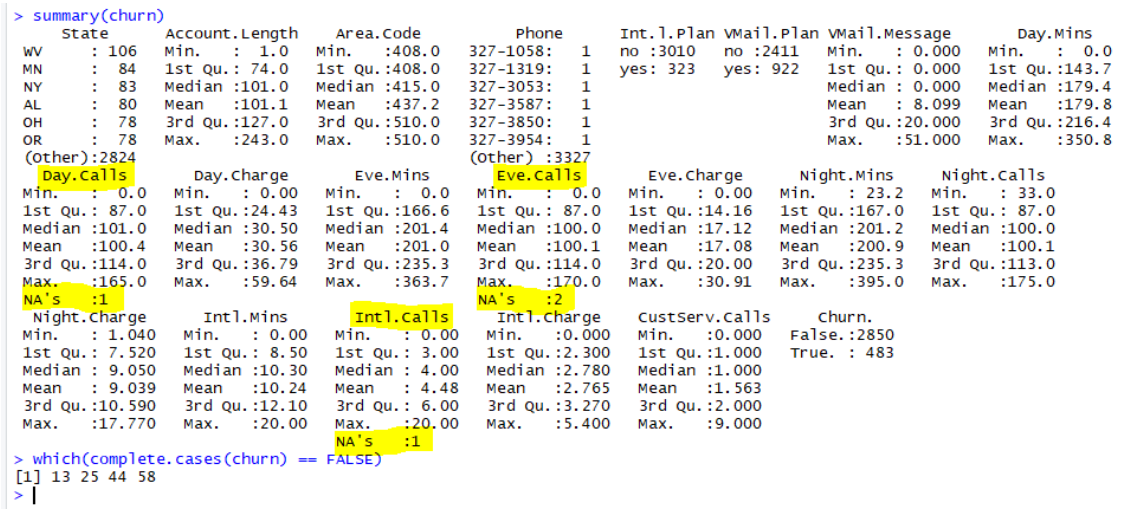


Figure 0.6: Missing Values in Churn data set.

We can see from the highlighted summary above that 3 columns have missing values Day.calls, Eve.Calls and Intl.Calls Below process shows the imputation of NA values.

```
> ave_DC <-mean(churn$Day.Calls, na.rm = TRUE)
> ave_EC <-mean(churn$Eve.Calls, na.rm = TRUE)
> ave_IC <-mean(churn$Intl.Calls, na.rm = TRUE)
> ave_DC
[1] 100.44
> ave_EC
[1] 100.1201
> ave_IC
[1] 4.479892
> churn$Day.Calls[is.na(churn$Day.Calls)==TRUE] <-ave_DC
> summary(churn$Day.Calls)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    87.0   101.0   100.4   114.0   165.0
> churn$Eve.Calls[is.na(churn$Eve.Calls)==TRUE] <-ave_EC
> churn$Intl.Calls[is.na(churn$Intl.Calls)==TRUE] <-ave_IC
> summary(churn$Eve.Calls)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    87.0   100.0   100.1   114.0   170.0
> summary(churn$Intl.Calls)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    3.00    4.00    4.48    6.00   20.00
> which(complete.cases(churn) == FALSE)
integer(0)
>
```

Figure 0.7: Missing Values imputation.

6. b. Use a boxplot to visually determine whether there are any outliers in the following columns day call, eve call .You may use either original data set or the new data set

```
boxplot(churn$Day.Calls,churn$Eve.Calls)
```

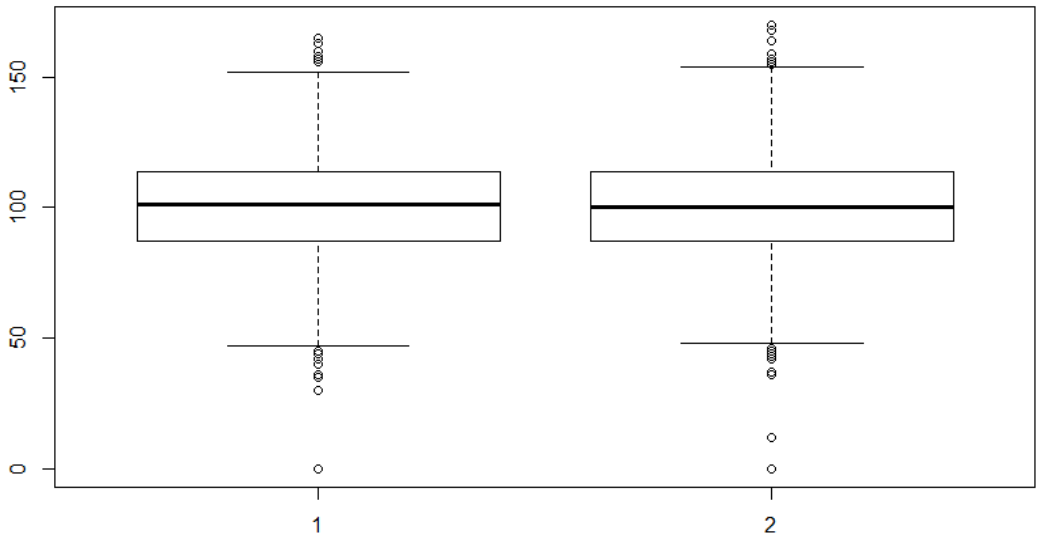


Figure 0.8: Box Plot Comparision of churn data set.

As shown above there are outliers present in both the columns which lie outside 1.5 times the interquartile range above the upper quartile and bellow the lower quartile.

7. c. Use plot to visually determine if there is linear relation among numerical columns. If so plot side by side these two columns.

Here we are checking the correlation values to see the strongest correlated values as highlighted in the below observation.

```
> churn_num <- Filter(is.numeric, churn)
> cor(churn_num)
                Account.Length    Area.Code VMail.Message      Day.Mins Day.Calls    Day.Charge      Eve.Mins Eve.Calls
Account.Length     1.000000000 -0.012463497 -0.0046278243  0.0062160205        NA  0.0062141347 -0.006757142        NA
Area.Code         -0.012463497  1.000000000 -0.0019943701 -0.0082643662        NA -0.0082644411  0.003580395        NA
VMail.Message     -0.004627824 -0.001994370  1.0000000000  0.0007782741        NA  0.0007755235  0.017562034        NA
Day.Mins           0.006216021 -0.008264366  0.0007782741  1.0000000000        NA  0.9999999522  0.007042511        NA
Day.Calls                   NA          NA            NA            NA         1            NA           NA        NA
Day.Charge         0.006214135 -0.008264441  0.0007755235  0.9999999522        NA  1.0000000000  0.007049607        NA
Eve.Mins          -0.006757142  0.003580395  0.0175620343  0.0070425110        NA  0.0070496072  1.000000000        NA
Eve.Calls                   NA          NA            NA            NA        NA            NA           NA         1
Eve.Charge        -0.006745302  0.003606690  0.0175777801  0.0070290353        NA  0.0070361315  0.999999776        NA
Night.Mins        -0.008955192 -0.005824660  0.0076811359  0.0043233666        NA  0.0043238794 -0.012583678        NA
Night.Calls       -0.013176275  0.016522317  0.0071230629  0.0229724555        NA  0.0229724195  0.007585643        NA
Night.Charge      -0.008959535 -0.005845376  0.0076632904  0.0043003570        NA  0.0043008608 -0.012592806        NA
Intl.Mins          0.009513902 -0.018288168  0.0028561959 -0.0101545856        NA -0.0101568616 -0.011034714        NA
Intl.Calls                  NA          NA            NA            NA        NA            NA           NA        NA
Intl.Charge        0.009545675 -0.018394696  0.0028836579 -0.0100919742        NA -0.0100942572 -0.011066621        NA
CustServ.Calls    -0.003795939  0.027572226 -0.0132625831 -0.0134231864        NA -0.0134269694 -0.012984553        NA
                  Eve.Charge   Night.Mins  Night.Calls Night.Charge    Intl.Mins Intl.Calls  Intl.Charge CustServ.Calls
Account.Length -0.006745302 -0.008955192 -0.013176275 -0.008959535  0.009513902         NA  0.009545675   -0.003795939
Area.Code       0.003606690 -0.005824660  0.016522317 -0.005845376 -0.018288168         NA -0.018394696    0.027572226
VMail.Message   0.017577780  0.007681136  0.007123063  0.007663290  0.002856196         NA  0.002883658   -0.013262583
Day.Mins        0.007029035  0.004323367  0.022972456  0.004300357 -0.010154586         NA -0.010091974   -0.013423186
Day.Calls                NA           NA           NA           NA           NA         NA           NA             NA
Day.Charge      0.007036131  0.004323879  0.022972420  0.004300861 -0.010156862         NA -0.010094257   -0.013426969
Eve.Mins        0.999999776 -0.012583678  0.007585643 -0.012592806 -0.011034714         NA -0.011066621   -0.012984553
Eve.Calls                NA           NA           NA           NA           NA         NA           NA             NA
Eve.Charge      1.000000000 -0.012592020  0.007595843 -0.012601142 -0.011042582         NA -0.011074499   -0.012987407
Night.Mins     -0.012592020  1.000000000  0.011203856  0.999999215 -0.015207297         NA -0.015179849   -0.009287613
Night.Calls     0.007595843  0.011203856  1.000000000  0.011187820 -0.013604996         NA -0.013630170   -0.012801927
Night.Charge   -0.012601142  0.999999215  0.011187820  1.000000000 -0.015213526         NA -0.015186139   -0.009276954
Intl.Mins      -0.011042582 -0.015207297 -0.013604996 -0.015213526  1.000000000         NA  0.999992742   -0.009639680
Intl.Calls               NA           NA           NA           NA           NA          1           NA             NA
Intl.Charge    -0.011074499 -0.015179849 -0.013630170 -0.015186139  0.999992742         NA  1.000000000   -0.009674732
CustServ.Calls -0.012987407 -0.009287613 -0.012801927 -0.009276954 -0.009639680         NA -0.009674732    1.000000000
> 
```

Figure 0.9: Correlation values Comparision of churn data set.

The strongest linear related columns are plotted below.

```
par(mfrow=c(3,1))
plot(churn$Day.Charge,churn$Day.Mins)
plot(churn$Eve.Charge,churn$Eve.Mins)
plot(churn$Intl.Charge,churn$Intl.Mins)
```

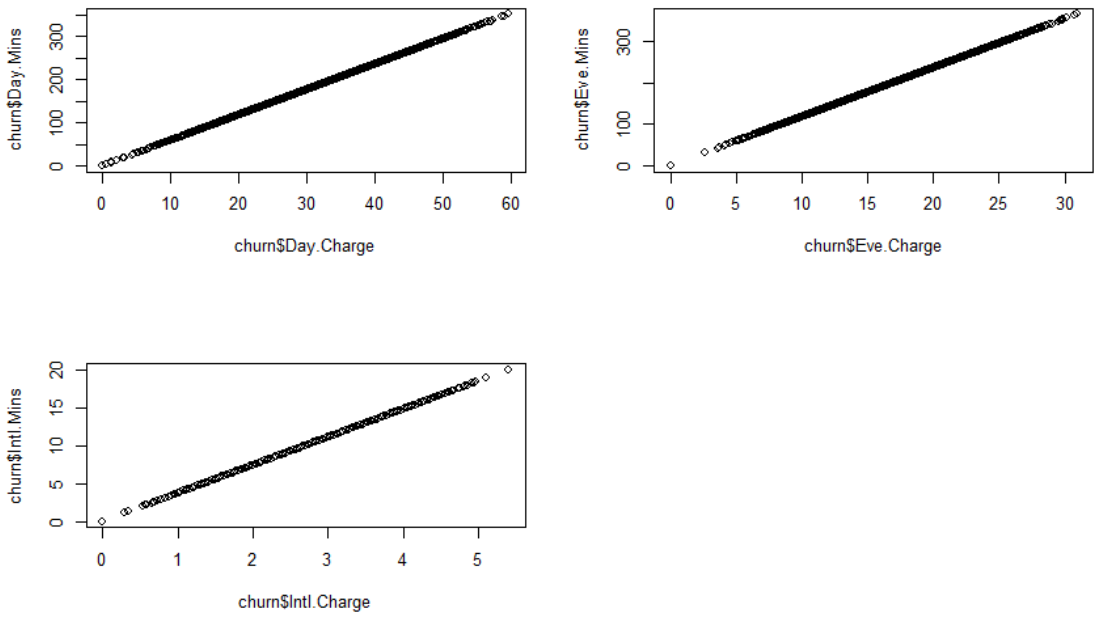


Figure 0.10: Plot Comparision of high correlated column