Assignment 1

Pre-coding work:

Since the datasets were provided within a Powerpoint slide, the first step was to clean the data and store it in separate csv files for easier reading. Also, judging by the convention followed in other datasets and for clarity of working, the second column in Dataset1 was renamed to Y1 instead of Y2.

Assignment Steps: ***Click Here for comprehensive step by step execution as mentioned in the assignment***

R Code: ***Click Here to go straight to the R Code***

Scatter Plots: ***Click here to go straight to the scatter plots***

Assignment Steps

Step 1: For each of the dataset given above, please find the summary statistics.

The datasets have already been loaded into memory in the form of Data Frames by using the read.csv() function.

Now we are using the summary() function to simply calculating the 5 number summary for each column of the 4 datasets respectively.

Code snippet:

```
> summary(Dataset1)
      X1              Y1
 Min.   : 4.0   Min.   : 4.260
 1st Qu.: 6.5   1st Qu.: 6.315
 Median : 9.0   Median : 7.580
 Mean   : 9.0   Mean   : 7.501
 3rd Qu.:11.5   3rd Qu.: 8.570
 Max.   :14.0   Max.   :10.840

> summary(Dataset2)
      X2              Y2
 Min.   : 4.0   Min.   :3.100
 1st Qu.: 6.5   1st Qu.:6.695
 Median : 9.0   Median :8.140
 Mean   : 9.0   Mean   :7.501
 3rd Qu.:11.5   3rd Qu.:8.950
 Max.   :14.0   Max.   :9.260

> summary(Dataset3)
      X3              Y3
 Min.   : 4.0   Min.   : 5.39
 1st Qu.: 6.5   1st Qu.: 6.25
 Median : 9.0   Median : 7.11
 Mean   : 9.0   Mean   : 7.50
 3rd Qu.:11.5   3rd Qu.: 7.98
 Max.   :14.0   Max.   :12.74

> summary(Dataset4)
      X4              Y4
 Min.   : 8   Min.   : 5.250
 1st Qu.: 8   1st Qu.: 6.170
 Median : 8   Median : 7.040
 Mean   : 9   Mean   : 7.501
 3rd Qu.: 8   3rd Qu.: 8.190
 Max.   :19   Max.   :12.500
```

Step 2: Compare the summary statistics of all the datasets and provide insights

- A cursory glance on the summary stats for the four datasets reveals that all five-point summary values for the first columns for Datasets 1, 2 and 3 are same. In fact all values in these columns are same which indicate that these three datasets might represent the variation of 3 different parameters (the respective second columns) against the same variable (first column).
- The summary stats of the second column for these 3 datasets show that the mean values are the same/nearly the same. However, other values differ significantly which indicate different distribution patterns for these variables.
- The summary of dataset4 shows that the minimum value, first quartile, median and third quartile values are all the same, i.e. 8. The max value however, happens to be an outlier – 19 and is throwing off the mean value to 9.

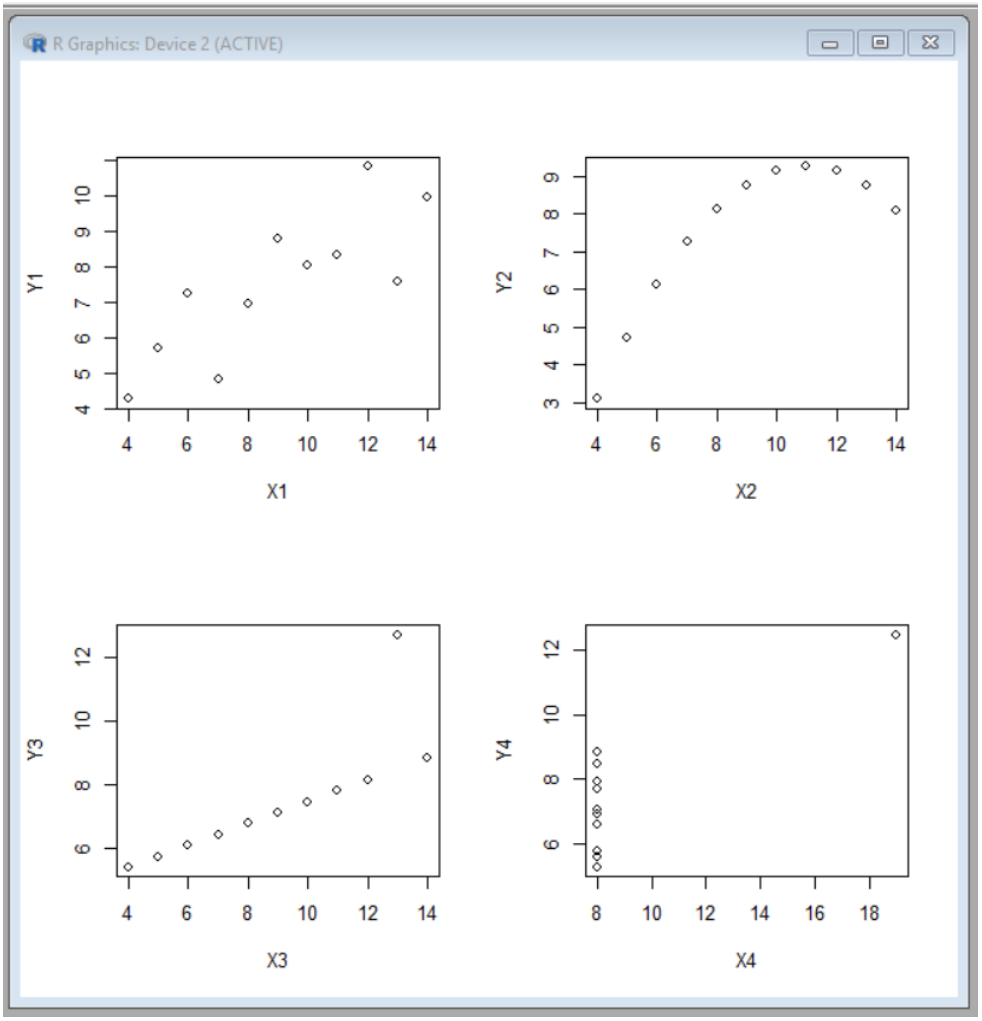
Step 3: Find the scatter plot of all the datasets.

- The scatter plots are plotted using the simple plot() function and are arranged in a single view by making a 2X2 grid using par() function.

R Code:

```
> par(mfrow=c(2,2))
> plot(Dataset1)
> plot(Dataset2)
> plot(Dataset3)
> plot(Dataset4)
```

- The scatter plots thus obtained:



It is possible to colorize the plots and to add titles to add clarity of representation. However, given that this is very small data and we don't specifically know what the data represents, it would be futile to do so.

Step 4: Provide Interpretation of different Scatter Plots.

Dataset1-

- There seems to be a positive correlation between the two variables. The Correlation coefficient between X1 and Y1 is 0.8164 (calculated using the cor() function). We may calculate the slope of the closest fitting straight line for this plot.

Dataset2 –

- The values clearly show a parabolic distribution with the Y2 variable peaking for X2 = 11.
- All data points follow the pattern with no outliers.
- Correlation coefficient between X2 and Y2 is 0.8162 which is almost the same as for the variables in dataset1. However, since dataset2 shows a parabolic plot, it would not be correct to use a correlation coefficient to make sense out of this data. For further analysis, we could calculate the equation between X2 and Y2.

Dataset3 –

- 10 out of the 11 data points follow the same pattern of a straight line with one very significant exception. This indicates that either this data point corresponds to an outlier or is a misreading/ wrong data.
- It would be appropriate to ignore this outlier data point while analyzing the dataset.
- Here when we first calculate the correlation coefficient using all datapoints, the value is 0.8162 which is close to what we got for the first 2 datasets.

```
> cor(Dataset3)
       X3       Y3
X3 1.0000000 0.8162867
Y3 0.8162867 1.0000000
```

- However, as seen from the plot, we should ignore the datapoint (13,12.74) which is the 3rd value. As seen below, after cleaning the data, the correlation coefficient is 0.99999 which is basically 1 and shows a perfect linear & positive correlation between the variables.

```
> cor(Dataset3[-3,])
       X3       Y3
X3 1.0000000 0.9999966
Y3 0.9999966 1.0000000
```

Dataset4 –

- Similar to the previous dataset, the plot shows an outlier. Rest of all values are the same indicating some kind of constant.

- Calculating the correlation coefficient without cleaning the data gives a similar value as previous datasets. However, ignoring the one outlier value (which is the 8th datapoint) , the picture is very clear:

```
> cor(Dataset4[-8,])
   X4 Y4
X4  1 NA
Y4 NA  1
Warning message:
In cor(Dataset4[-8, ]) : the standard deviation is zero
```

- The function cor() now shows that the correlation coefficient between X4 and Y4 is NA and also warns that the std. dev is zero. This is because the variable X4 is not varying at all and is a constant.
- The five point summary after cleaning the data would also give a good idea:

```
> summary(Dataset4[-8,])
      X4          Y4
 Min.   :8   Min.   :5.250
 1st Qu.:8   1st Qu.:5.965
 Median :8   Median :6.965
 Mean   :8   Mean   :7.001
 3rd Qu.:8   3rd Qu.:7.860
 Max.   :8   Max.   :8.840
```

- In this case, we might infer that the value of Y4 is being determined by something other than the variable X4. Further analysis of the case and more data will give clearer information.

Step 5: Summarize your insights

- Looking at the scatter plots, we can see that the distribution patterns of all four datasets are different. However, if we only look at the summary statistics and correlation coefficients, it would not give us a complete understanding. Looking at the graphic representation, i.e. the plots gives us a better idea about the data.
- As seen in the cases of Datasets 3 and 4, it is very important to cleanup the data before analysing it to reach to a conclusion.

R Code

```
> getwd()
[1] "C:/Users/Madhulica Singh/Documents"
> setwd("M:/Work/Residency1/coursework")

>#  Read the data

> Dataset1 = read.csv("Dataset1.csv")
> Dataset2 = read.csv("Dataset2.csv")
> Dataset3 = read.csv("Dataset3.csv")
> Dataset4 = read.csv("Dataset4.csv")

> summary(Dataset1)
       X1              Y1
 Min.   : 4.0   Min.   : 4.260
 1st Qu.: 6.5   1st Qu.: 6.315
 Median : 9.0   Median : 7.580
 Mean   : 9.0   Mean   : 7.501
 3rd Qu.:11.5   3rd Qu.: 8.570
 Max.   :14.0   Max.   :10.840

> summary(Dataset2)
       X2              Y2
 Min.   : 4.0   Min.   :3.100
 1st Qu.: 6.5   1st Qu.:6.695
 Median : 9.0   Median :8.140
 Mean   : 9.0   Mean   :7.501
 3rd Qu.:11.5   3rd Qu.:8.950
 Max.   :14.0   Max.   :9.260
> summary(Dataset3)
       X3              Y3
 Min.   : 4.0   Min.   : 5.39
 1st Qu.: 6.5   1st Qu.: 6.25
 Median : 9.0   Median : 7.11
 Mean   : 9.0   Mean   : 7.50
 3rd Qu.:11.5   3rd Qu.: 7.98
 Max.   :14.0   Max.   :12.74
> summary(Dataset4)
       X4          Y4
 Min.   : 8   Min.   : 5.250
 1st Qu.: 8   1st Qu.: 6.170
 Median : 8   Median : 7.040
 Mean   : 9   Mean   : 7.501
 3rd Qu.: 8   3rd Qu.: 8.190
 Max.   :19   Max.   :12.500

> par(mfrow=c(2,2))
> plot(Dataset1)
> plot(Dataset2)
```

```
> plot(Dataset3)
> plot(Dataset4)


# Calculating the correlation coefficients between the variables in all four datasets
> cor(Dataset1)
          X1        Y1
X1 1.0000000 0.8164205
Y1 0.8164205 1.0000000

> cor(Dataset3)
          X3        Y3
X3 1.0000000 0.8162867
Y3 0.8162867 1.0000000

> cor(Dataset2)
          X2        Y2
X2 1.0000000 0.8162365
Y2 0.8162365 1.0000000

> cor(Dataset4)
          X4        Y4
X4 1.0000000 0.8165214
Y4 0.8165214 1.0000000

> cor(Dataset4[-8,])
   X4 Y4
X4  1 NA
Y4 NA  1
Warning message:
In cor(Dataset4[-8, ]) : the standard deviation is zero

> summary(Dataset4[-8,])
       X4          Y4
 Min.   :8   Min.   :5.250
 1st Qu.:8   1st Qu.:5.965
 Median :8   Median :6.965
 Mean   :8   Mean   :7.001
 3rd Qu.:8   3rd Qu.:7.860
 Max.   :8   Max.   :8.840
```

Scatter Plots of all four datasets