DATA 606 Fall 2017 - Final Exam

Anjal Husssan

Part I

Please put the answers for Part I next to the question number (2pts each):

Question:1

Ans: Variable daysDrive is quantitative and discrete since the value is counted in whole days. Variables car and color are qualitative and gasMonth is not discrete.

Question:2

Ans: The histogram is left skewed, therefore there is a greater chance of having median greater than mean. It also can be assumed that 3.8 as a Median is too high. So the Answer is a. mean = 3.3, median = 3.5

Question:3

Ans: Random selection for testing and observing the data points on how new testing affects a group will help to see if the treatment causes improvement in Ebola patients. So the Answer is d. Both a) and c). Although A or C should mention to observe the new testing affects on Male and Female cause new testing may work differently depending on gender.

Question:4

Ans: a) There's a difference between average eye color and average hair color

Since we are taking a large chi square means that we will reject the null hypothesis that the there is no difference between average eye color and average hair color.

Question:5

Ans: b. 17.8 and 69.0

lower: $Q1-1.5\times IQR = 17.8$ Upper: $Q3+1.5\times IQR=69$

Question:6

Ans: d. median and interquartile range; mean and standard deviation

Question:7

Ans: a)

Distribution A is unimodal and Right to the right with a mean of 5 and spread is small. Distribution B is unimodal, symmetrical.

Ans: b)

The means of the two distributions are similar because distribution B is a sample distribution of the population A. The standard deviations are different because distribution B has wider spread with a smaller population compare to distribution A

Ans: c)

This phenomenon is described by the Central Limit Theorem.

Part II

Consider the four datasets, each with two columns (x and y), provided below.

```
options(digits=2)
data1 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68))
data2 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74))
data3 <- data.frame(x=c(10,8,13,9,11,14,6,4,12,7,5),
                    y=c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73))
data4 <- data.frame(x=c(8,8,8,8,8,8,8,19,8,8,8),
                    y=c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.5,5.56,7.91,6.89))
```

For each column, calculate (to two decimal places):

a. The mean (for x and y separately; 1 pt).

```
meanx1 <- mean(data1$x)
cat("Mean of Data1 X is: ", meanx1, "\n")
```

```
## Mean of Data1 X is:  9
```

```
meanx2 <- mean(data2$x)
cat("Mean of Data2 X is: ", meanx2, "\n")
```

```
## Mean of Data2 X is:  9
```

```
meanx3 <- mean(data3$x)
cat("Mean of Data3 X is: ", meanx3, "\n")
```

```
## Mean of Data3 X is:  9
```

```
meanx4 <- mean(data4$x)
cat("Mean of Data4 X is: ", meanx4, "\n")
```

```
## Mean of Data4 X is:  9
```

```
meany1 <- mean(data1$y)
cat("Mean of Data1 Y is: ", meany1, "\n")
```

```
## Mean of Data1 Y is:  7.5
```

```
meany2 <- mean(data2$y)
cat("Mean of Data2 Y is: ", meany2, "\n")
```

```
## Mean of Data2 Y is:  7.5
```

```
meany3 <- mean(data3$y)
cat("Mean of Data3 Y is: ", meany3, "\n")
```

```
## Mean of Data3 Y is:  7.5
```

```
meany4 <- mean(data4$y)
cat("Mean of Data4 Y is: ", meany4, "\n")
```

```
## Mean of Data4 Y is:  7.5
```

b. The median (for x and y separately; 1 pt).

```
medianx= median(data1$x)
cat("The Median of data1_x is: ", medianx, "\n")
```

```
## The Median of data1_x is:  9
```

```
mediany=median(data1$y)
cat("The Median of data1_y is: ", mediany, "\n")
```

```
## The Median of data1_y is:  7.6
```

```
medianx= median(data2$x)
cat("The Median of data2_x is: ", medianx, "\n")
```

```
## The Median of data2_x is:  9
```

```
mediany=median(data2$y)
cat("The Median of data2_y is: ", mediany, "\n")
```

```
## The Median of data2_y is:  8.1
```

```
medianx= median(data3$x)
cat("The Median of data3_x is: ", medianx, "\n")
```

```
## The Median of data3_x is:  9
```

```
mediany=median(data3$y)
cat("The Median of data3_y is: ", mediany, "\n")
```

```
## The Median of data3_y is:  7.1
```

```
medianx= median(data4$x)
cat("The Median of data4_x is: ", medianx, "\n")
```

```
## The Median of data4_x is:  8
```

```
mediany=median(data4$y)
cat("The Median of data4_y is: ", mediany, "\n")
```

```
## The Median of data4_y is:  7
```

c. The standard deviation (for x and y separately; 1 pt).

```
stdx <- sd(data1$x)
cat("The standard Deviation of data1_x is: ", stdx, "\n")
```

```
## The standard Deviation of data1_x is:  3.3
```

```
stdx <- sd(data2$x)
cat("The standard Deviation of data2_x is: ", stdx, "\n")
```

```
## The standard Deviation of data2_x is:  3.3
```

```
stdx <- sd(data3$x)
cat("The standard Deviation of data3_x is: ", stdx, "\n")
```

```
## The standard Deviation of data3_x is:  3.3
```

```
stdx <- sd(data4$x)
cat("The standard Deviation of data4_x is: ", stdx, "\n")
```

```
## The standard Deviation of data4_x is:  3.3
```

```
stdy <- sd(data1$y)
cat("The standard Deviation of data1_Y is: ", stdy, "\n")
```

```
## The standard Deviation of data1_Y is:  2
```

```
stdy <- sd(data2$y)
cat("The standard Deviation of data2_Y is: ", stdy, "\n")
```

```
## The standard Deviation of data2_Y is:  2
```

```
stdy <- sd(data3$y)
cat("The standard Deviation of data3_Y is: ", stdy, "\n")
```

```
## The standard Deviation of data3_Y is:  2
```

```
stdy <- sd(data4$y)
cat("The standard Deviation of data4_Y is: ", stdy, "\n")
```

```
## The standard Deviation of data4_Y is:  2
```

For each x and y pair, calculate (also to two decimal places; 1 pt):

d. The correlation (1 pt).

```
cat("The Correlation for data1 is: ", cor(data1$x, data1$y), "\n")
```

```
## The Correlation for data1 is:  0.82
```

```
cat("The Correlation for data2 is: ", cor(data2$x, data2$y), "\n")
```

```
## The Correlation for data2 is:  0.82
```

```
cat("The Correlation for data3 is: ", cor(data3$x, data3$y), "\n")
```

```
## The Correlation for data3 is:  0.82
```

```
cat("The Correlation for data4 is: ", cor(data4$x, data4$y), "\n")
```

```
## The Correlation for data4 is:  0.82
```

e. Linear regression equation (2 pts).

```
# Build linear regression model for each data set
equation1 <- lm(data1$y ~ data1$x)
equation2 <- lm(data2$y ~ data2$x)
equation3 <- lm(data3$y ~ data3$x)
equation4 <- lm(data4$y ~ data4$x)

coefficients <- rbind(equation1$coefficients, equation2$coefficients,
                      equation3$coefficients, equation4$coefficients)
rownames(coefficients) <- c("data1", "data2", "data3", "data4")
coefficients
```

```
##       (Intercept) data1$x
## data1           3     0.5
## data2           3     0.5
## data3           3     0.5
## data4           3     0.5
```

Linear regression equations is same for all data set: y=3+0.5x

f. R-Squared (2 pts).

```
cat("R^2 for DataSet1: ",summary(equation1)$r.squared, "\n")
```

```
## R^2 for DataSet1:  0.67
```

```
cat("R^2 for DataSet2: ",summary(equation2)$r.squared, "\n")
```

```
## R^2 for DataSet2:  0.67
```

```
cat("R^2 for DataSet3: ",summary(equation3)$r.squared, "\n")
```

```
## R^2 for DataSet3:  0.67
```

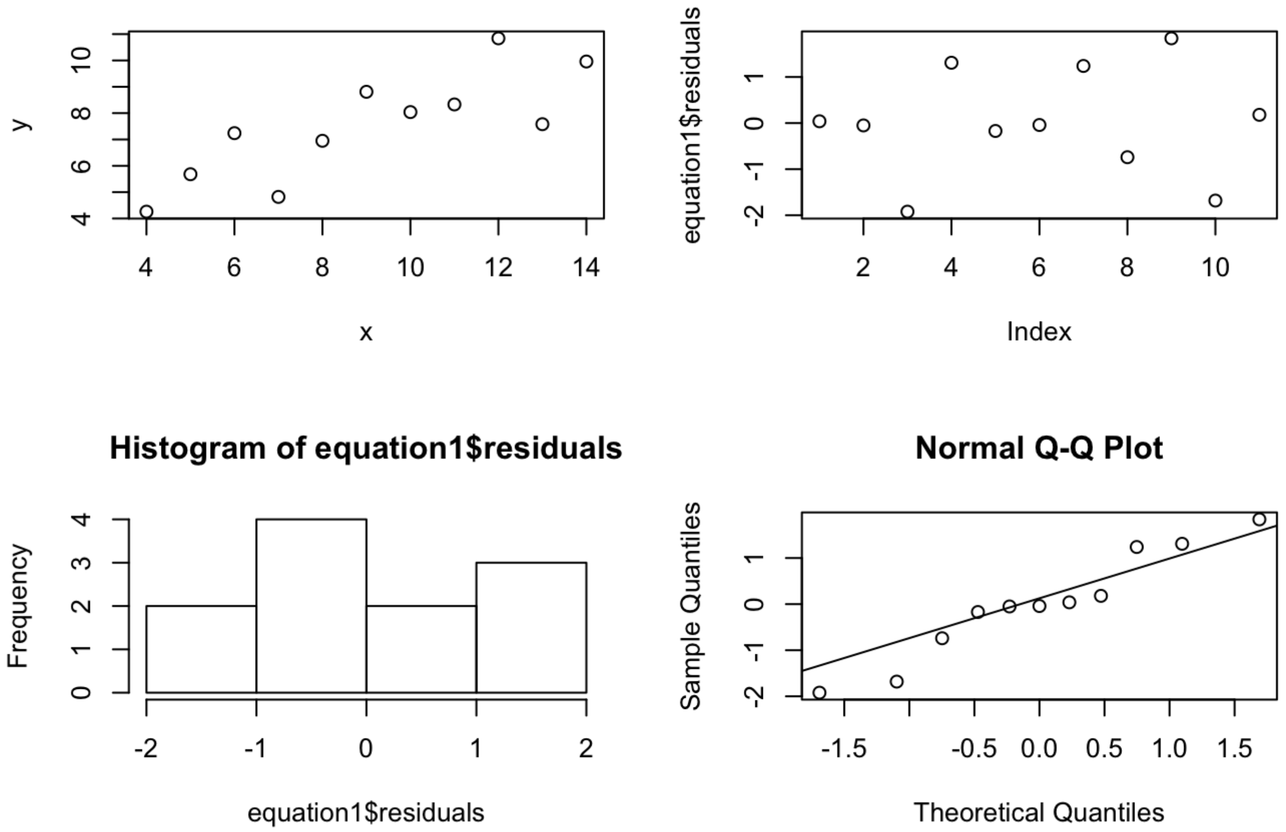
```
cat("R^2 for DataSet4: ",summary(equation4)$r.squared, "\n")
```

```
## R^2 for DataSet4:  0.67
```

Therefore, R^2 = 0.67

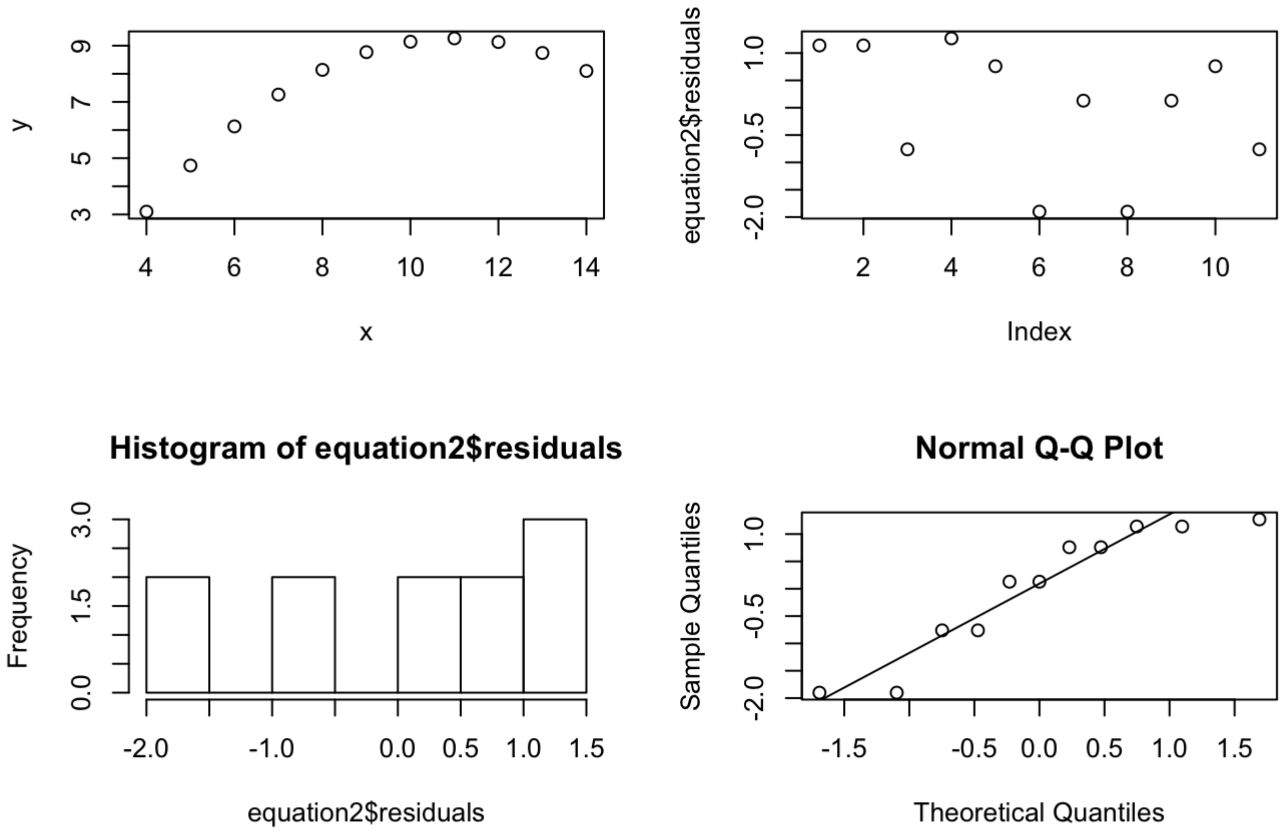
For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

```
par(mfrow=c(2,2))
plot(data1)
plot(equation1$residuals)
hist(equation1$residuals)
qqnorm(equation1$residuals)
qqline(equation1$residuals)
```



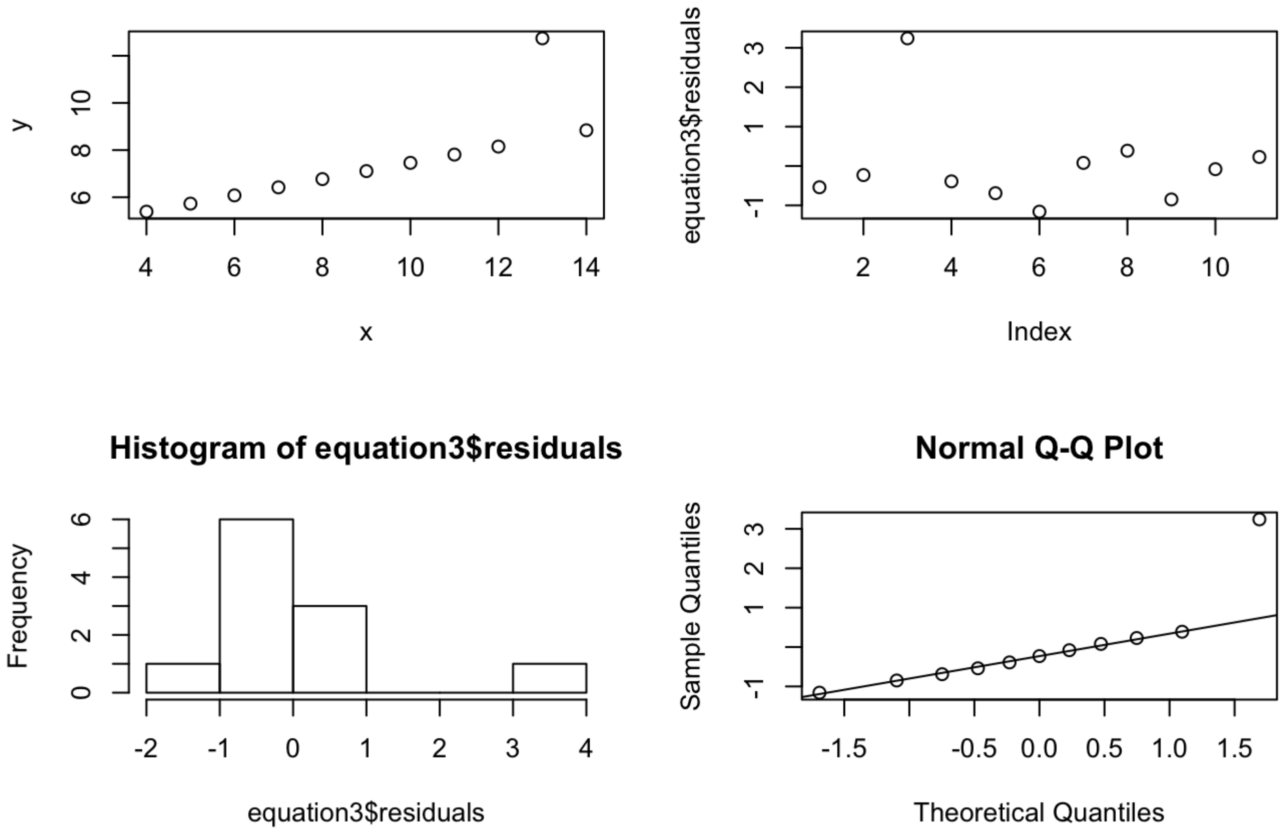
plot seems to have linearity but Data1 does not have residuals that imply a normal distribution.

```
par(mfrow=c(2,2))
plot(data2)
plot(equation2$residuals)
hist(equation2$residuals)
qqnorm(equation2$residuals)
qqline(equation2$residuals)
```



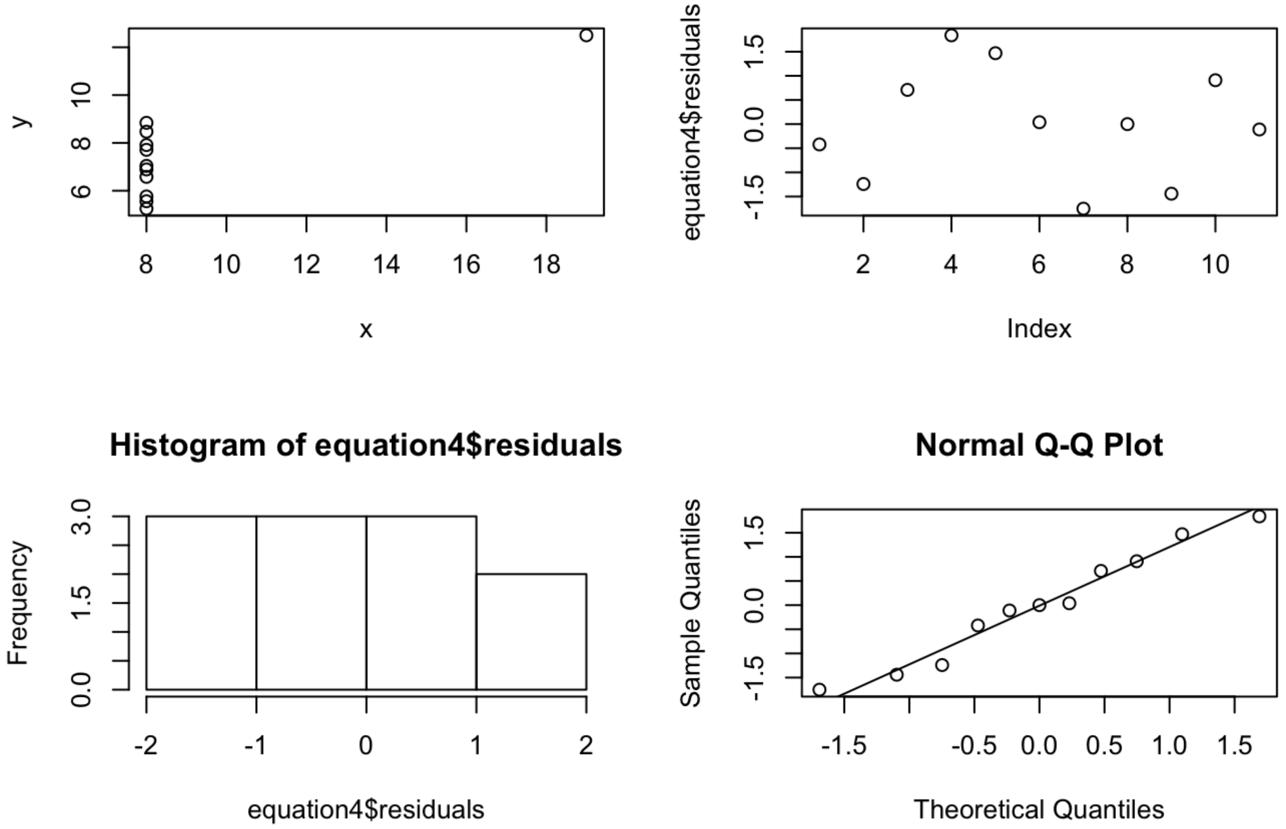
Plots for Data 2 do not show linearity or have residuals that follow a normal distribution.

```
par(mfrow=c(2,2))
plot(data3)
plot(equation3$residuals)
hist(equation3$residuals)
qqnorm(equation3$residuals)
qqline(equation3$residuals)
```



Plots for Data3 show linearity and the residuals seems to follow a normal distribution.

```
par(mfrow=c(2,2))
plot(data4)
plot(equation4$residuals)
hist(equation4$residuals)
qqnorm(equation4$residuals)
qqline(equation4$residuals)
```



Plots for Data 4 indicates that there is no linearity and the residuals don’t follow any normal distribution.

Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (2 pts)

Ans: Visualizations indicates and supports any statements that we make when we analyze data. Visulizations are also important to indicate trends and insights that cannot be observed by just matrices.

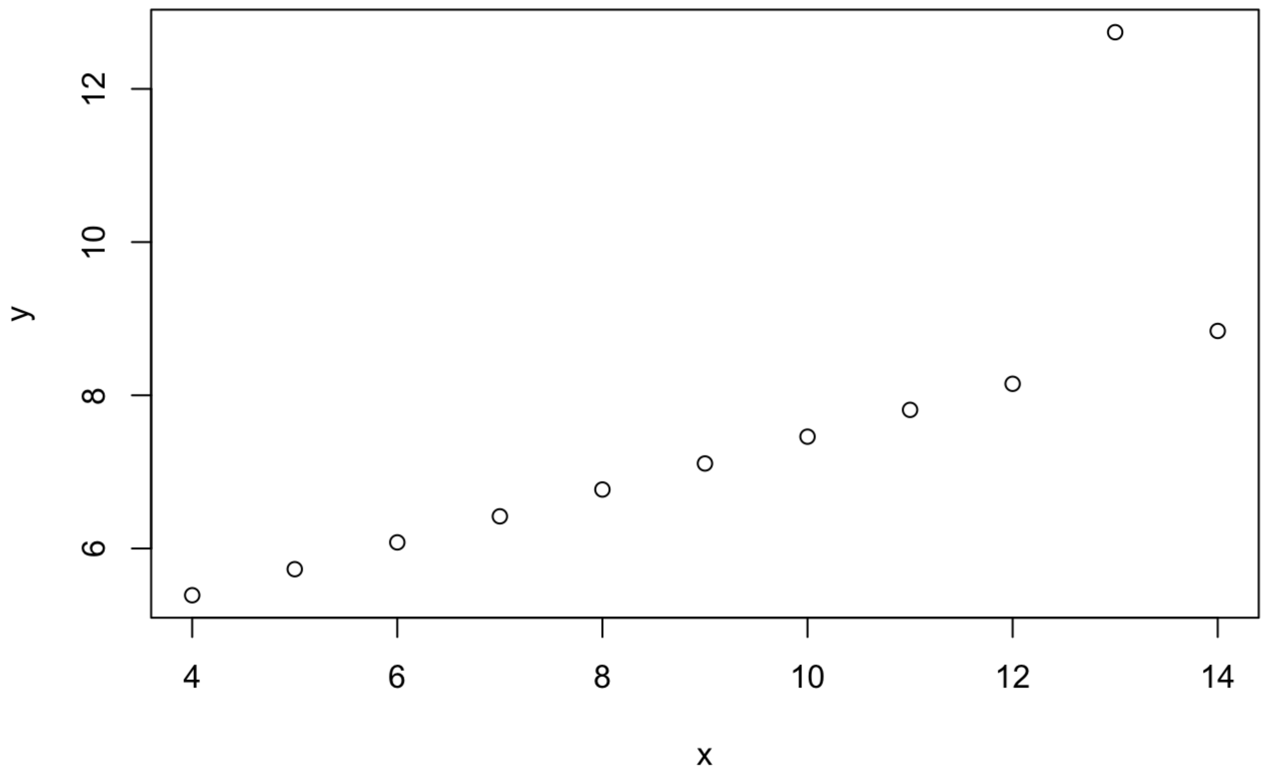
To elaborate, when we look at the Data3 table, we don’t see any trend. If the data set is extremely large, it’s hard to find any pattern.

```
head(data3)
```

```
##    x    y
## 1 10  7.5
## 2  8  6.8
## 3 13 12.7
## 4  9  7.1
## 5 11  7.8
## 6 14  8.8
```

But when we generate a plot it is very obvious that it’s following a pattern.

```
plot(data3)
```



End of this file