DATA 606 Spring 2018 - Final Exam

Ritesh_Lohiya

5/17/2018

Part I

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

1. b. daysDrive
2. a. mean = 3.3, median = 3.5
3. d. Both studies (a) and (b) can be conducted in order to establish that the treatment does indeed cause improvement with regards to fever in Ebola patients.
4. d. eye color and natural hair color are independent
5. b. 17.8 and 69.0
6. d. median and interquartile range; mean and standard deviation

7a. Describe the two distributions (2pts).

A - The distribution is unimodal and right skewed.

B - The distribution is unimodal and normal.

7b. Explain why the means of these two distributions are similar but the standard deviations are not (2 pts).

For any population distribution with mean and standard deviation , the sampling distribution of the sample mean X¯ is approximately normal with mean and standard deviation /sqrt(n) , and the approximation improves as n increases. The standard deviation is standard error for the mean estimated from the data.

```
sd <- 3.22
n <-30
SE <- sd / sqrt(n)
SE
```

```
## [1] 0.5878889
```

So the standard deviations are not similar.

7c. What is the statistical principal that describes this phenomenon (2 pts)?

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).

```
#mean of data1
mean(data1$x)
```

```
## [1] 9
```

```
mean(data1$y)
```

```
## [1] 7.5
```

```
#mean of data2
mean(data2$x)
```

```
## [1] 9
```

```
mean(data2$y)
```

```
## [1] 7.5
```

```
#mean of data3
mean(data3$x)
```

```
## [1] 9
```

```
mean(data3$y)
```

```
## [1] 7.5
```

```
#mean of data4
mean(data4$x)
```

```
## [1] 9
```

```
mean(data4$y)
```

```
## [1] 7.5
```

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

```
#median of data1
median(data1$x)
```

```
## [1] 9
```

```
median(data1$y)
```

```
## [1] 7.6
```

```
#median of data2
median(data2$x)
```

```
## [1] 9
```

```
median(data2$y)
```

```
## [1] 8.1
```

```r
#median of data3
median(data3$x)
```

```
## [1] 9
```

```r
median(data3$y)
```

```
## [1] 7.1
```

```r
#median of data4
median(data4$x)
```

```
## [1] 8
```

```r
median(data4$y)
```

```
## [1] 7
```

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

```r
#Standard deviation of data1
sd(data1$x)
```

```
## [1] 3.3
```

```r
sd(data1$y)
```

```
## [1] 2
```

```r
#Standard deviation of data2
sd(data2$x)
```

```
## [1] 3.3
```

```r
sd(data2$y)
```

```
## [1] 2
```

```r
#Standard deviation of data3
sd(data3$x)
```

```
## [1] 3.3
```

```r
sd(data3$y)
```

```
## [1] 2
```

```r
#Standard deviation of data4
sd(data4$x)
```

```
## [1] 3.3
```

```r
sd(data4$y)
```

```
## [1] 2
```

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

d. The correlation (1 pt).

```
#Correlation deviation of data1
cor(data1)
```

```
##      x    y
## x 1.00 0.82
## y 0.82 1.00
```

```
#Correlation deviation of data2
cor(data2)
```

```
##      x    y
## x 1.00 0.82
## y 0.82 1.00
```

```
#Correlation deviation of data3
cor(data3)
```

```
##      x    y
## x 1.00 0.82
## y 0.82 1.00
```

```
#Correlation deviation of data4
cor(data4)
```

```
##      x    y
## x 1.00 0.82
## y 0.82 1.00
```

e. Linear regression equation (2 pts).

```
#Linear regression equation for data1
lm1 <- lm(x ~ y, data = data1)
summary(lm1)
```

```
##
## Call:
## lm(formula = x ~ y, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.652 -1.512 -0.266  1.234  3.895
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.998      2.434   -0.41   0.6916
## y              1.333      0.314    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2 on 9 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.629
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00217
```

```
#equation1: y = -0.998 + 1.333 * x

#Linear regression equation for data2
lm2 <- lm(x ~ y, data = data2)
```

```r
summary(lm2)
```

```
##
## Call:
## lm(formula = x ~ y, data = data2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.852 -1.432 -0.344  0.847  4.202
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.995      2.435   -0.41   0.6925
## y              1.332      0.314    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2 on 9 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.629
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.00218
```

```r
#equation2: y = -0.995 + 1.332 * x

#Linear regression equation for data3
lm3 <- lm(x ~ y, data = data3)
summary(lm3)
```

```
##
## Call:
## lm(formula = x ~ y, data = data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.987 -1.373 -0.027  1.320  3.213
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.000      2.436   -0.41   0.6910
## y              1.333      0.315    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2 on 9 degrees of freedom
## Multiple R-squared:  0.666,  Adjusted R-squared:  0.629
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.00218
```

```r
#equation3: y = -1.000 + 1.333 * x

#Linear regression equation for data4
lm4 <- lm(x ~ y, data = data4)
summary(lm4)
```

```
##
## Call:
## lm(formula = x ~ y, data = data4)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.786 -1.412 -0.185  1.455  3.333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.004      2.435   -0.41   0.6898
## y              1.334      0.314    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2 on 9 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.63
## F-statistic:   18 on 1 and 9 DF,  p-value: 0.00216
```

```
#equation4: y = -1.004 + 1.334 * x
```

Equation for data1: y = -0.998 + 1.333 * x

Equation for data2: y = -0.995 + 1.332 * x

Equation for data3: y = -1.000 + 1.333 * x

Equation for data4: y = -1.004 + 1.334 * x

f. R-Squared (2 pts).

```
#R-Squared for data1
summary(lm1)$r.squared
```

```
## [1] 0.67
```

```
#R-Squared for data2
summary(lm2)$r.squared
```

```
## [1] 0.67
```

```
#R-Squared for data3
summary(lm3)$r.squared
```

```
## [1] 0.67
```

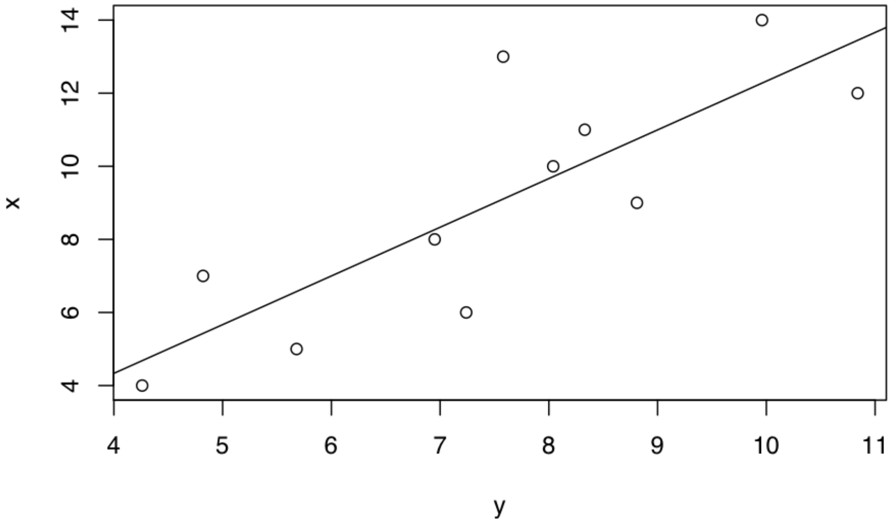
```
#R-Squared for data4
summary(lm4)$r.squared
```

```
## [1] 0.67
```

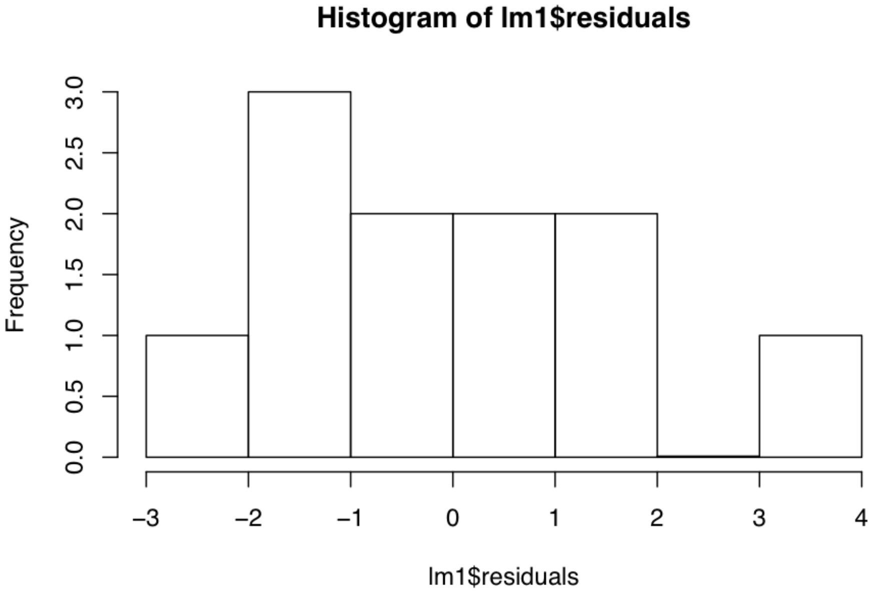
For each pair, is it appropriate to estimate a linear regression model? Why or why not? Be specific as to why for each pair and include appropriate plots! (4 pts)

#Plots for data1

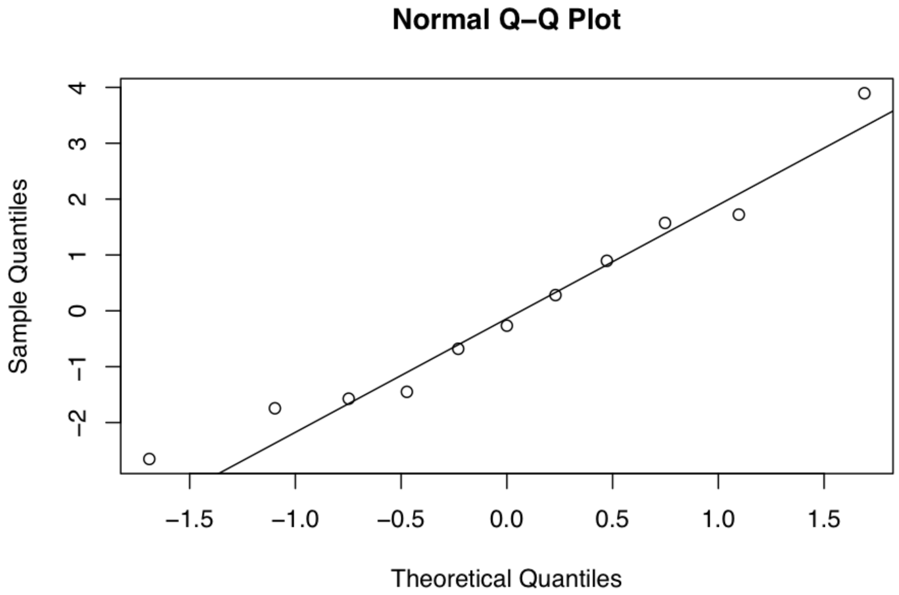
```
#data1 plots
plot(x ~ y, data1)
abline(lm1)
```



```r
hist(lm1$residuals)
```



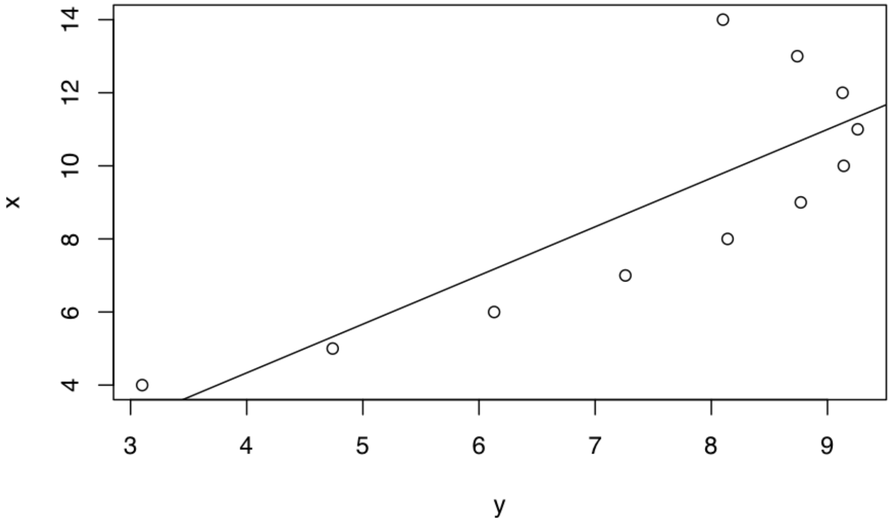
```r
qqnorm(lm1$residuals)
qqline(lm1$residuals)
```



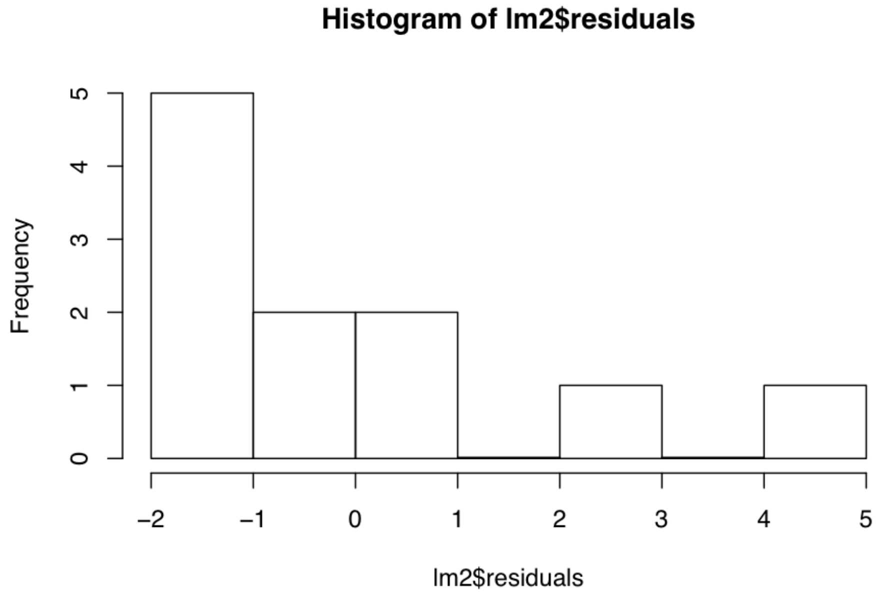
For data1: From the plots for data1 we can see that data plot suggests linearity but the residuals do not seems to follow a nearly normal distribution.

#Plots for data2

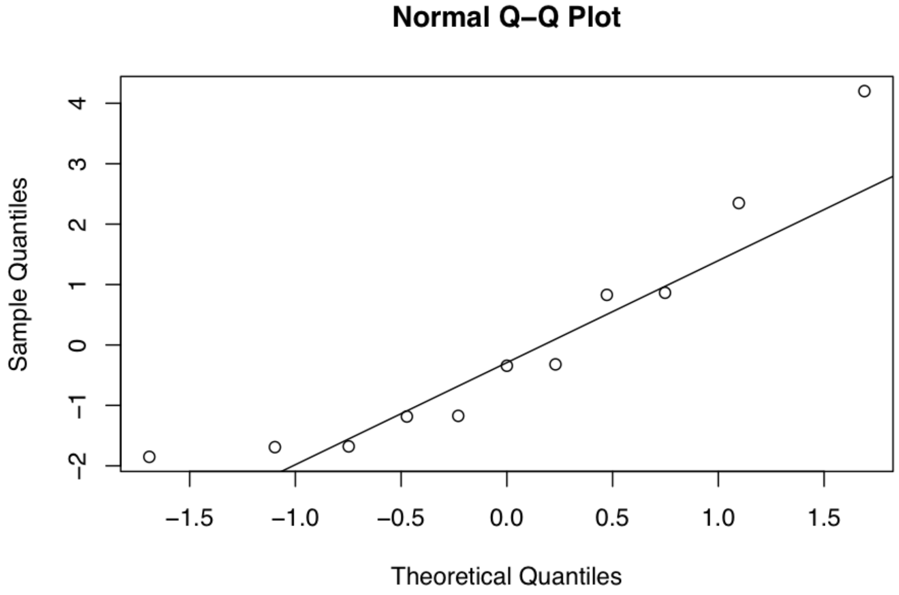
```
#data2 plots
plot(x ~ y, data2)
abline(lm2)
```



```r
hist(lm2$residuals)
```



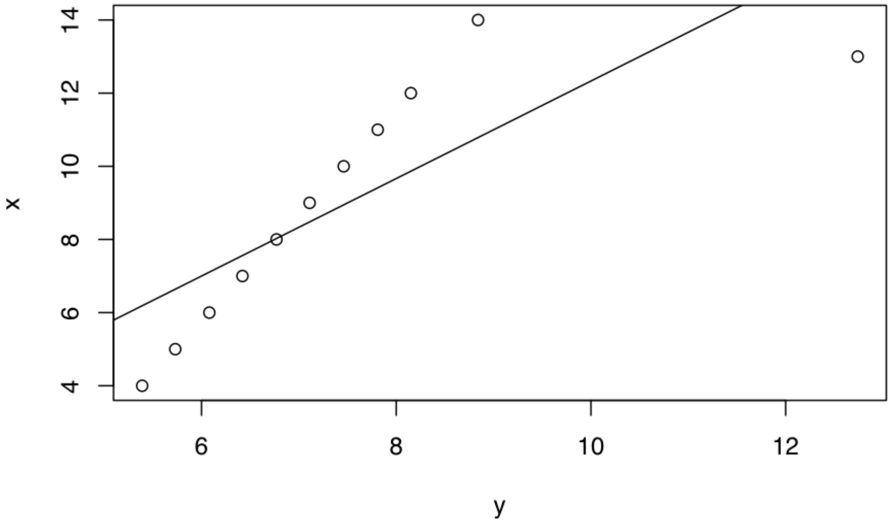
```r
qqnorm(lm2$residuals)
qqline(lm2$residuals)
```



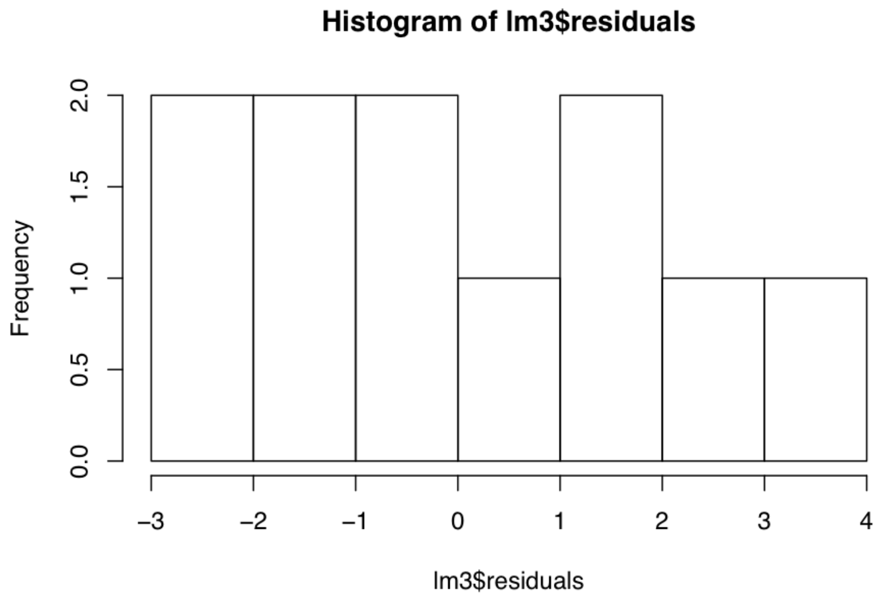
For data2: From the plots for data2 we can see that data plot does not suggests linearity. Also the residuals do not seems to follow a nearly normal distribution.

#Plots for data3

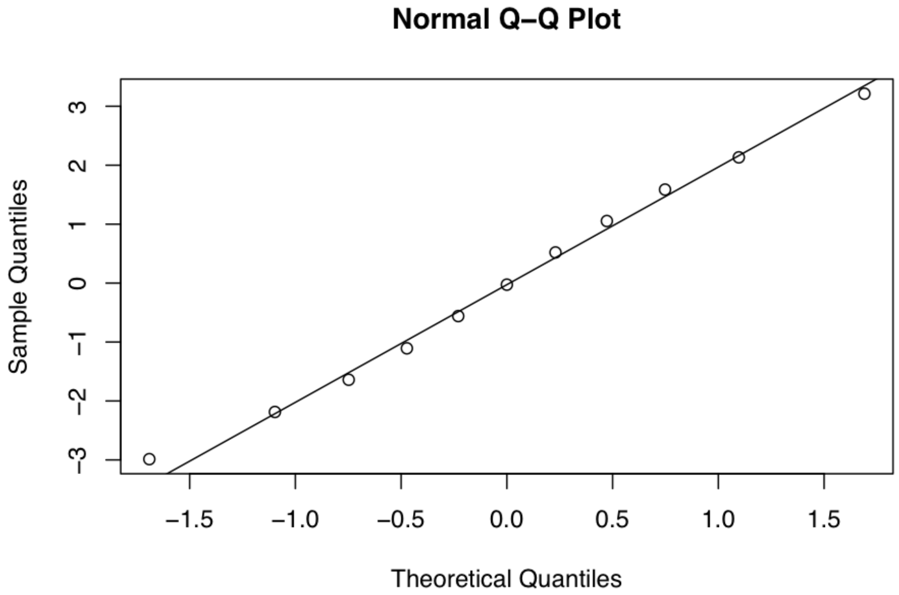
```r
#data3 plots
plot(x ~ y, data3)
abline(lm3)
```



```r
hist(lm3$residuals)
```



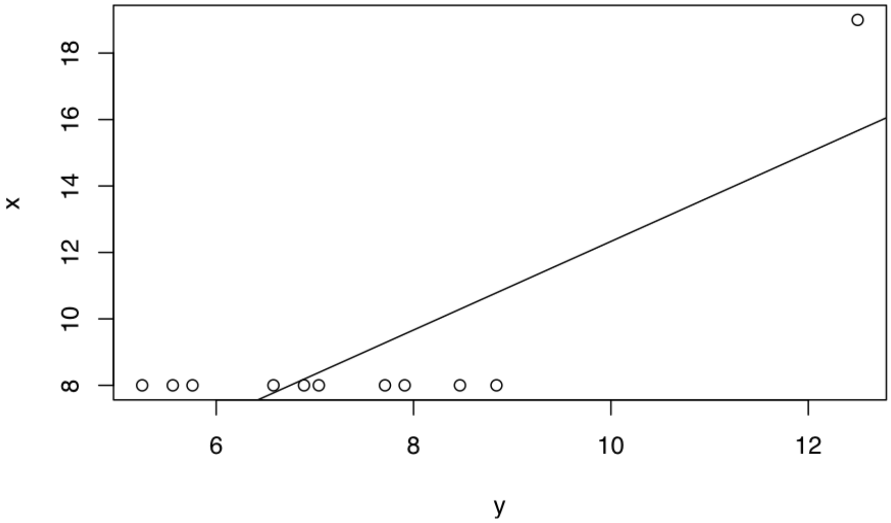
```
qqnorm(lm3$residuals)
qqline(lm3$residuals)
```



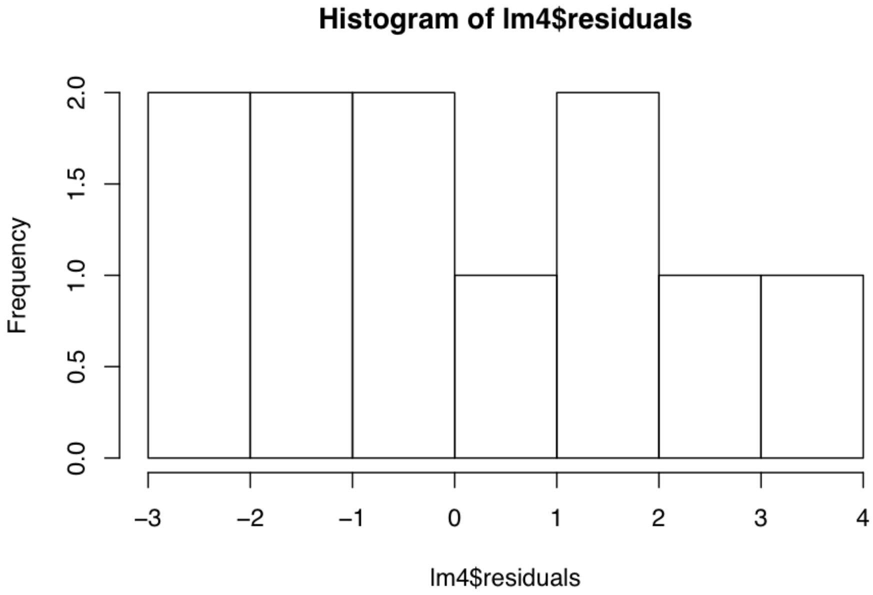
For data3: From the plots for data3 we can see that data plot suggests linearity although there is outlier. Also the residuals seems to follow somewhat normal distribution.

#Plots for data4

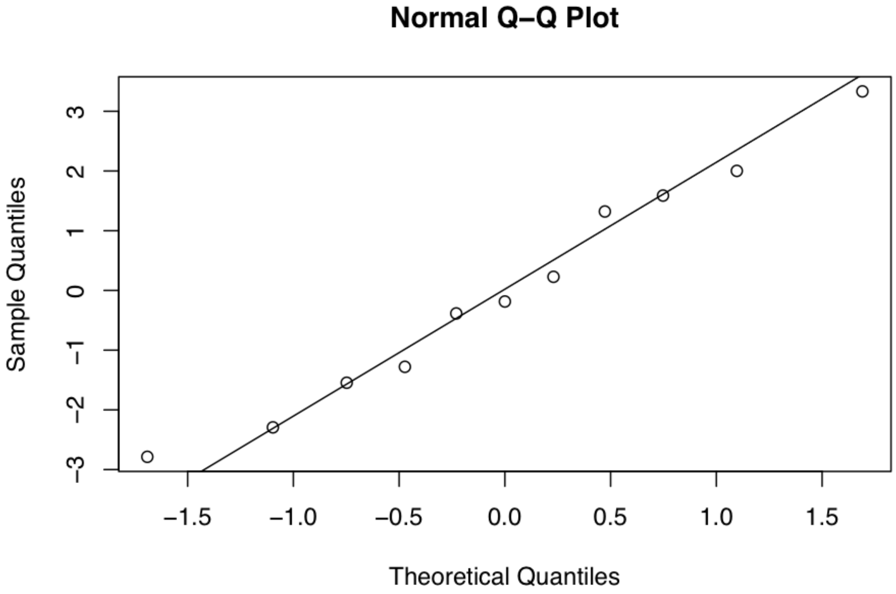
```
#data4 plots
plot(x ~ y, data4)
abline(lm4)
```



```
hist(lm4$residuals)
```



```
qqnorm(lm4$residuals)
qqline(lm4$residuals)
```



For data4: From the plots for data4 we can see that there is no linearity in the data and also the residuals does not follow normal distribution.

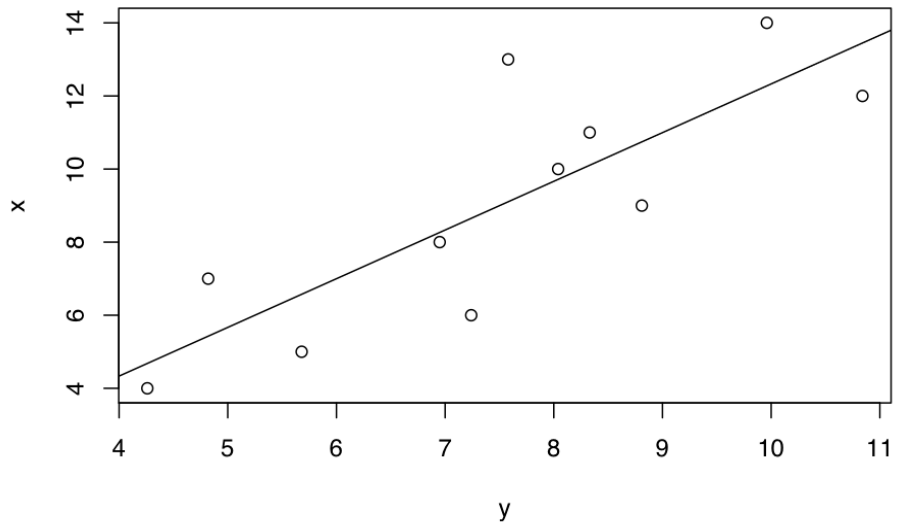
Explain why it is important to include appropriate visualizations when analyzing data. Include any visualization(s) you create. (2 pts)

Visualizations is the best way to see the trends in the data, residuals etc. It helps us know if there are any outliers in the data. It helps to visualize the linearity of the data. It helps in knowing if the residuals have variability or not. It helps in comparison of that different data(Which looks similar when we only see that data).

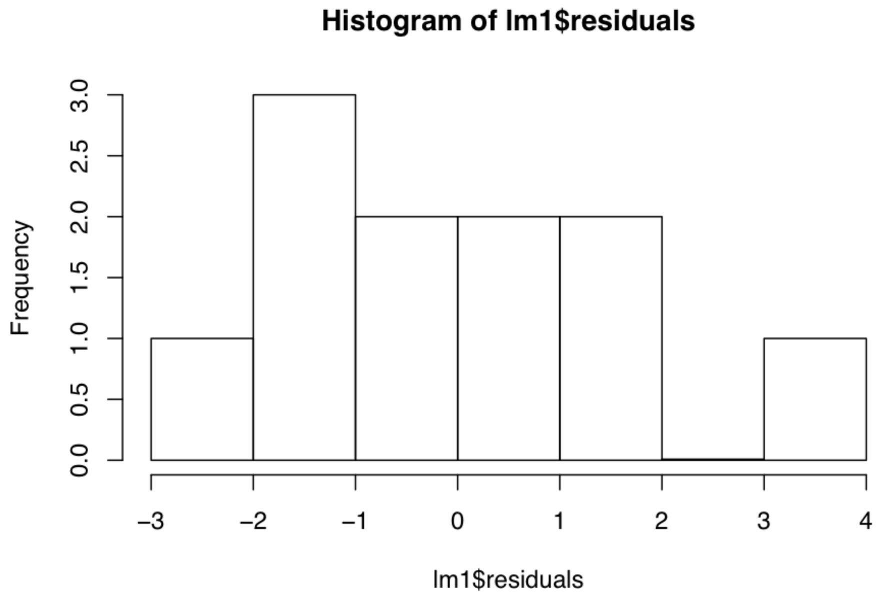
Example of Visualizations plots for data1:

#Plots for data1

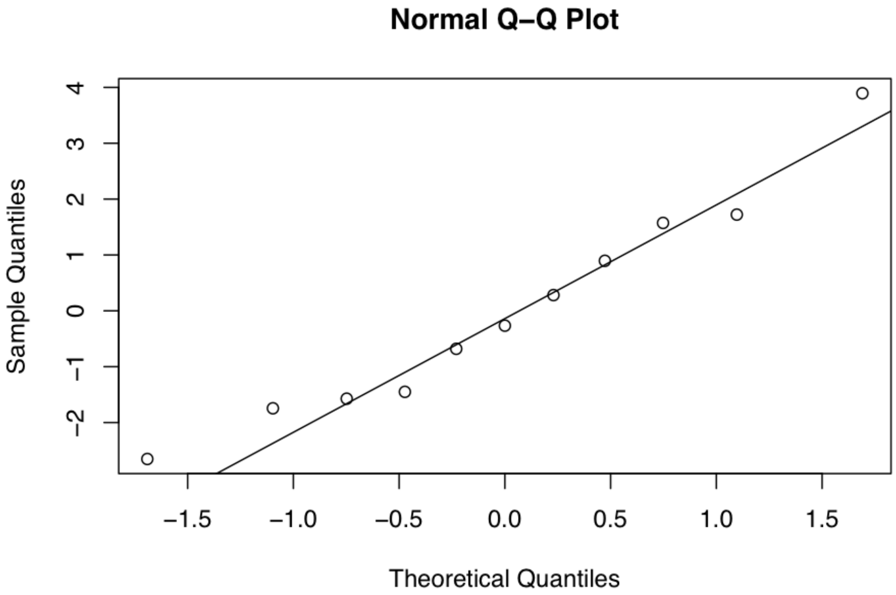
```
#data1 plots
plot(x ~ y, data1)
abline(lm1)
```



```
hist(lm1$residuals)
```



```
qqnorm(lm1$residuals)
qqline(lm1$residuals)
```


Normal Q−Q Plot
Sample Quantiles
−2
−1
0
1
2
3
4
Theoretical Quantiles
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5