DATA 606 Fall 2023 - Final Exam

Victor H Torres

Part I

Please put the answers for Part I next to the question number (please enter only the letter options; 4 points each):

1. C
2. A
3. D
4. B
5. B
6. E
7. D
8. E
9. B
10. C

Part II

Consider the three datasets, each with two columns (x and y), provided below. Be sure to replace the `NA` with your answer for each part (e.g. assign the mean of `x` for `data1` to the `data1.x.mean` variable). When you Knit your answer document, a table will be generated with all the answers.

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

```
data1.x.mean <- mean(data1$x)
data1.y.mean <- mean(data1$y)
data2.x.mean <- mean(data2$x)
data2.y.mean <- mean(data2$y)
data3.x.mean <- mean(data3$x)
data3.y.mean <- mean(data3$y)
```

a. The mean (for x and y separately; 5 pt).

```
data1.x.median <- median(data1$x)
data1.y.median <- median(data1$y)
data2.x.median <- median(data2$x)
```

```
data2.y.median <- median(data2$y)
data3.x.median <- median(data3$x)
data3.y.median <- median(data3$y)
```

b. The median (for x and y separately; 5 pt).

```
data1.x.sd <- sd(data1$x)
data1.y.sd <- sd(data1$y)
data2.x.sd <- sd(data2$x)
data2.y.sd <- sd(data2$y)
data3.x.sd <- sd(data3$x)
data3.y.sd <- sd(data3$y)
```

c. The standard deviation (for x and y separately; 5 pt).

For each x and y pair, calculate (also to four decimal places):

```
round(cor(data1),2)
```

d. The correlation (5 pt).

```
##       x     y
## x  1.00 -0.06
## y -0.06  1.00
```

```
round(cor(data2),2)
```

```
##       x     y
## x  1.00 -0.07
## y -0.07  1.00
```

```
round(cor(data3),2)
```

```
##       x     y
## x  1.00 -0.06
## y -0.06  1.00
```

```
data1.correlation <- -0.06
data2.correlation <- -0.07
data3.correlation <- -0.06
```

```r
model1 <- lm(x ~ y, data = data1)
summary(model1)
```

e. Linear regression equation (5 points).

```
##
## Call:
## lm(formula = x ~ y, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.58 -10.56  -0.98  10.29  43.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.1827     2.8792   19.51   <2e-16 ***
## y            -0.0401     0.0525   -0.76     0.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.8 on 140 degrees of freedom
## Multiple R-squared:  0.00416,    Adjusted R-squared:  -0.00296
## F-statistic: 0.584 on 1 and 140 DF,  p-value: 0.446
```

```r
model2 <- lm(x ~ y, data = data2)
summary(model2)
```

Formula for equation 1: y=-0.0401*x+56.1827

```
##
## Call:
## lm(formula = x ~ y, data = data2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35.91 -11.20  -0.02  10.33  40.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.3218     2.8788   19.56   <2e-16 ***
## y            -0.0429     0.0525   -0.82     0.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.8 on 140 degrees of freedom
## Multiple R-squared:  0.00476,    Adjusted R-squared:  -0.00235
## F-statistic: 0.669 on 1 and 140 DF,  p-value: 0.415
```

```r
model3 <- lm(x ~ y, data = data3)
summary(model3)
```

Formula for equation 2: y=-0.0429*x+56.3218

```
##
## Call:
## lm(formula = x ~ y, data = data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37.42 -13.76  -0.69  15.03  38.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.1756     2.8799   19.51   <2e-16 ***
## y            -0.0399     0.0525   -0.76     0.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.8 on 140 degrees of freedom
## Multiple R-squared:  0.00411,    Adjusted R-squared:  -0.003
## F-statistic: 0.578 on 1 and 140 DF,  p-value: 0.448
```

```r
data1.slope <- -0.0401
data2.slope <- -0.0429
data3.slope <- -0.0399

data1.intercept <- 56.1827
data2.intercept <- 56.3218
data3.intercept <- 56.1756
```

Formula for equation 3: y=-0.0399*x+56.1756

```r
data1.rsquared <- summary(model1)$r.squared
data2.rsquared <- summary(model2)$r.squared
data3.rsquared <- summary(model3)$r.squared
```

f. R-Squared (5 points). Summary Table

	Data 1		Data 2		Data 3	
	x	y	x	y	x	y
Mean	54.2633	47.8323	54.2678	47.8359	54.2661	47.8347

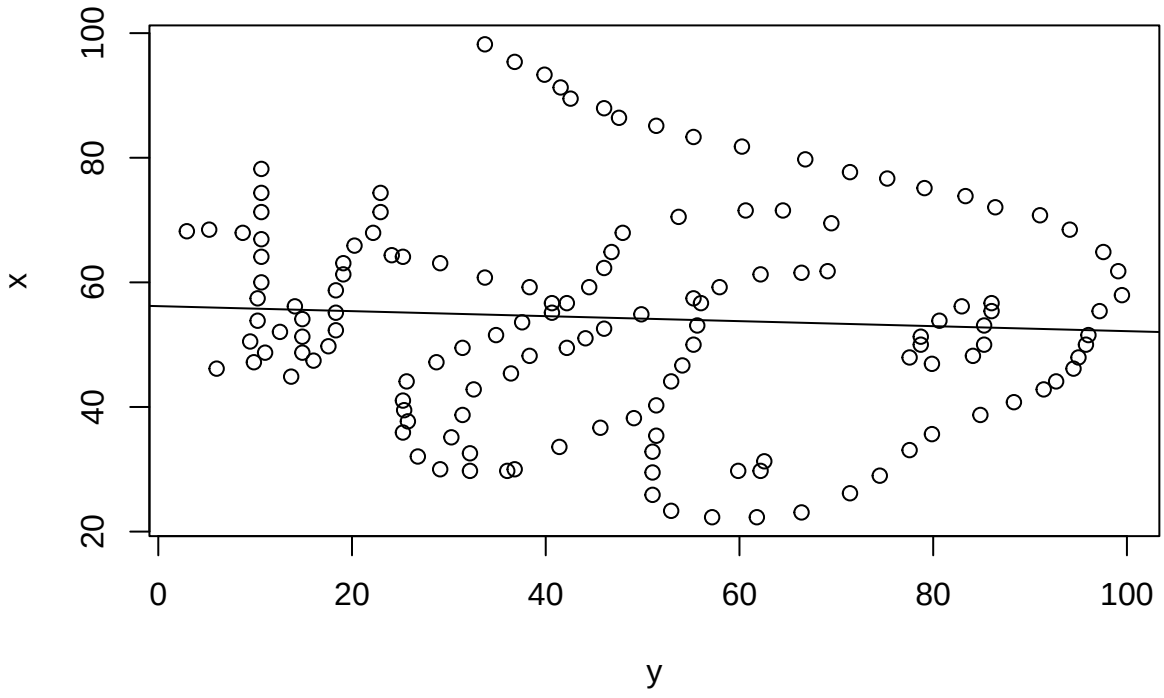
Median	53.3333	46.0256	53.1352	46.4013	53.3403	47.5353
SD	16.7651	26.9354	16.7668	26.9361	16.7698	26.9397
r	-0.0600		-0.0700		-0.0600	
Intercept	56.1827		56.3218		56.1756	
Slope	-0.0401		-0.0429		-0.0399	
R-Squared	0.0042		0.0048		0.0041	

g. 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! (15 points)

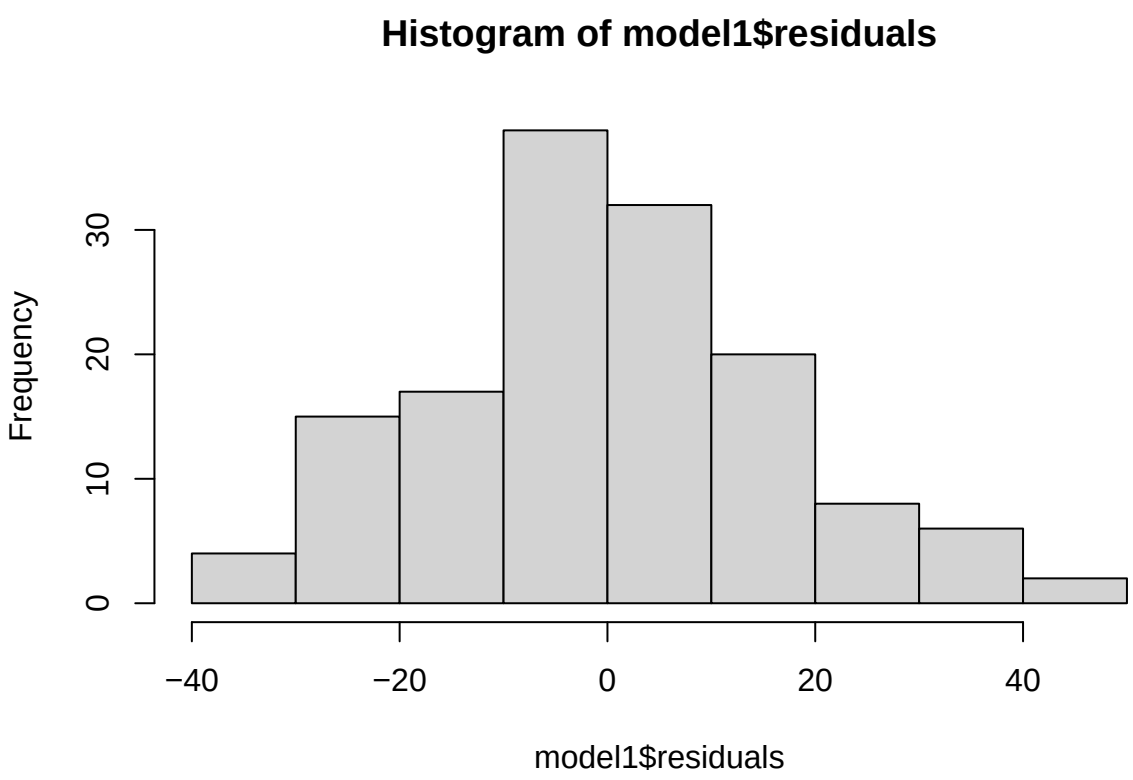
Data set 1 Yes

Based on the plots for data1 below, data1 seems to have linearity although there are outlier. In the residuals plot. we can see that the data seems to have normal distribution.

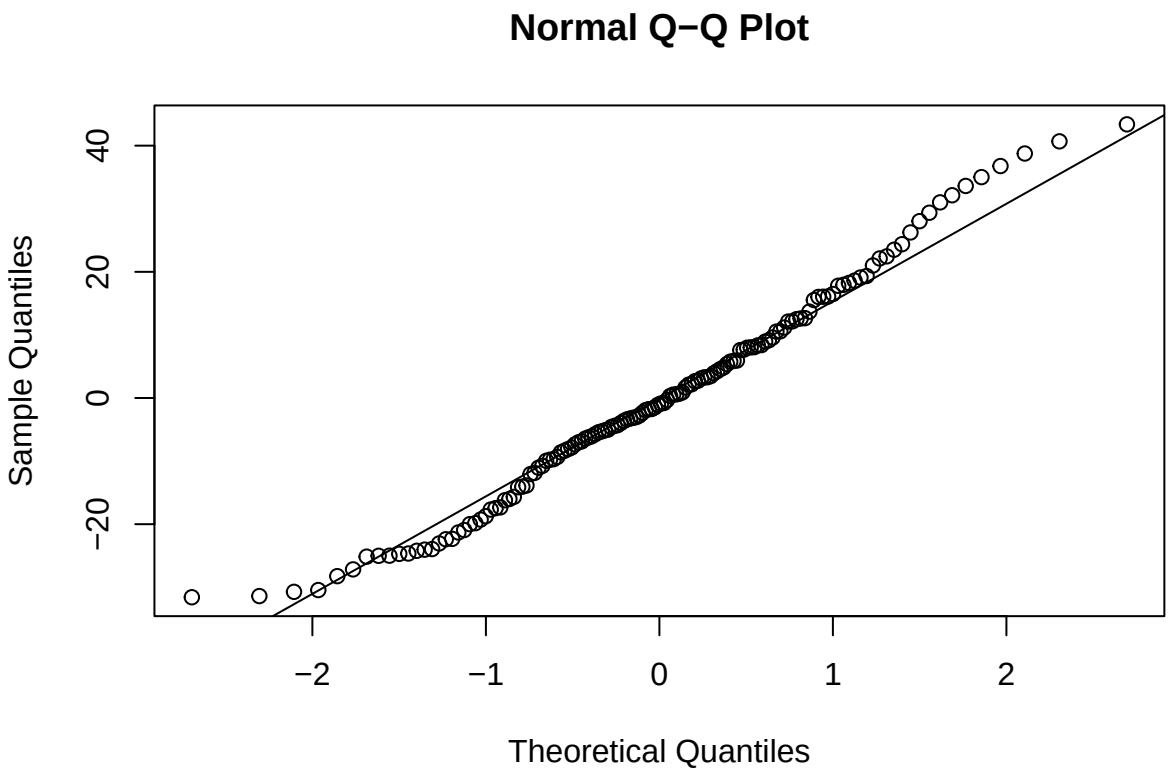
```
plot(x ~ y, data1)
abline(model1)
```




```
hist(model1$residuals)
```



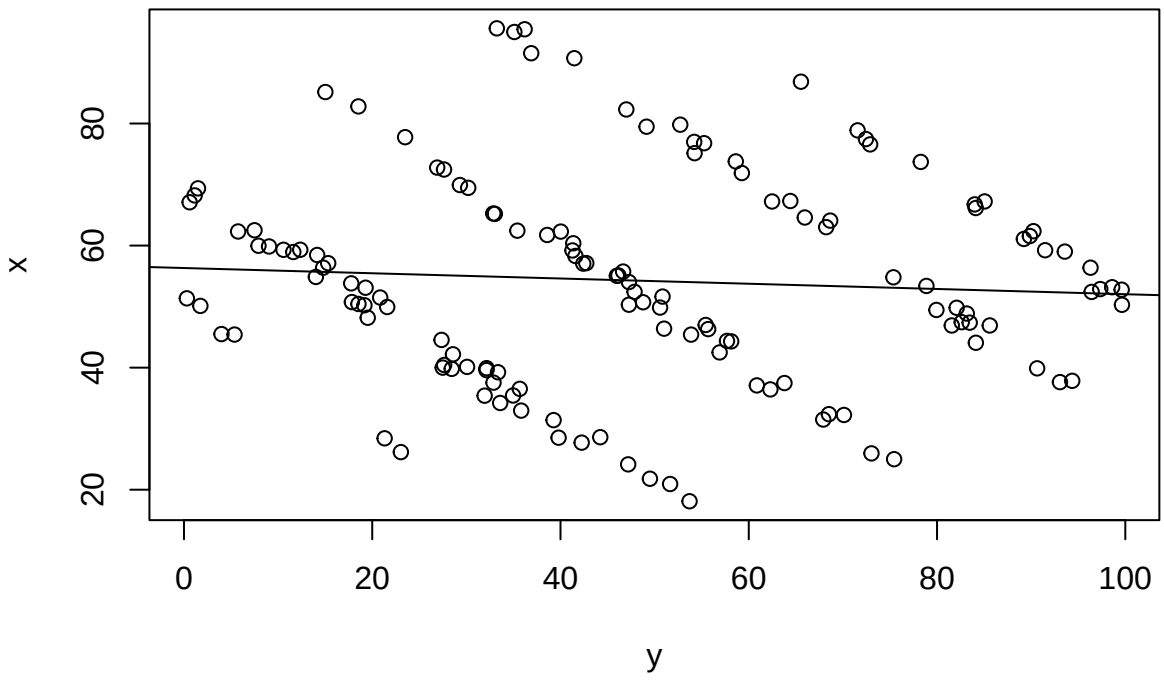

```
qqnorm(model1$residuals)
qqline(model1$residuals)
```



Data set 2 No

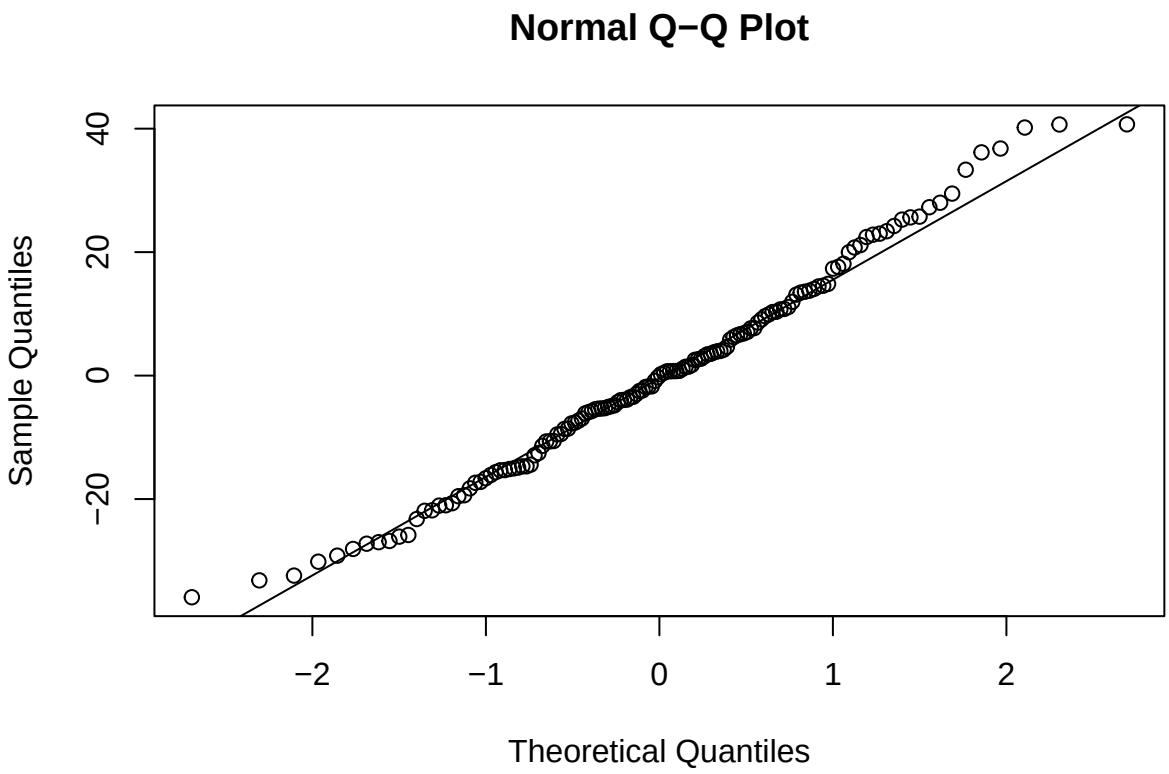
Based on plots of data2 below, the is not linearity between columns.The residuals plot doesn't show a nearly normal distribution

```
plot(x ~ y, data2)
abline(model2)
```




```
hist(model2$residuals)
```

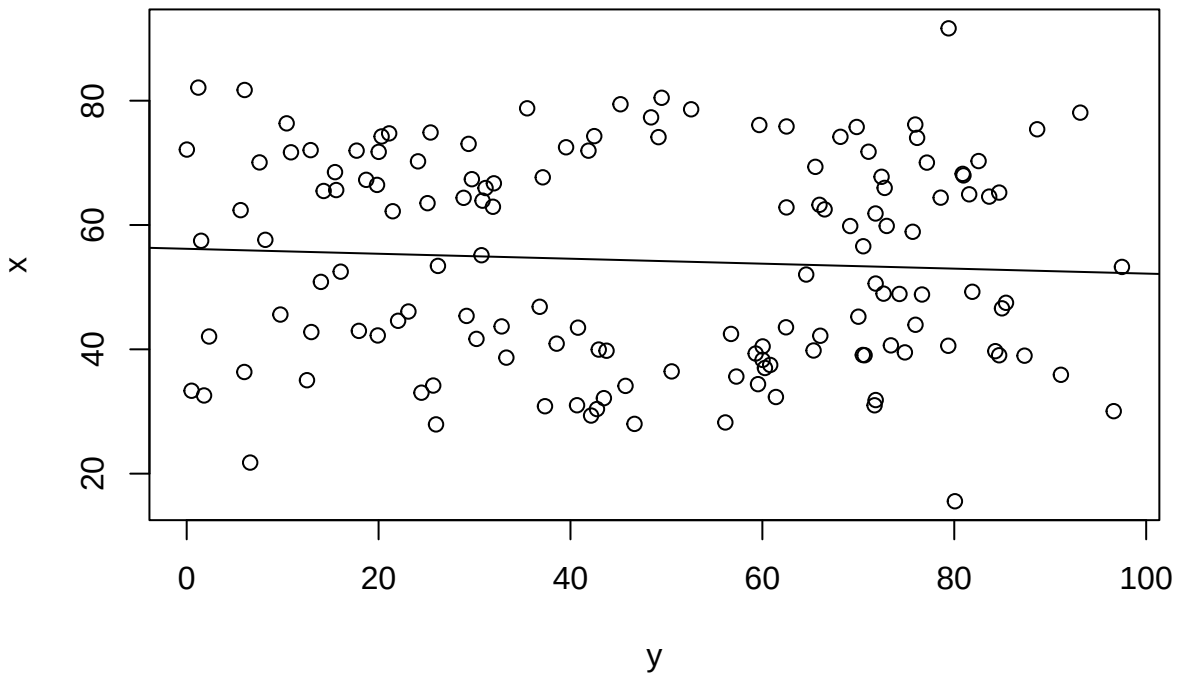

```
qqnorm(model2$residuals)
qqline(model2$residuals)
```



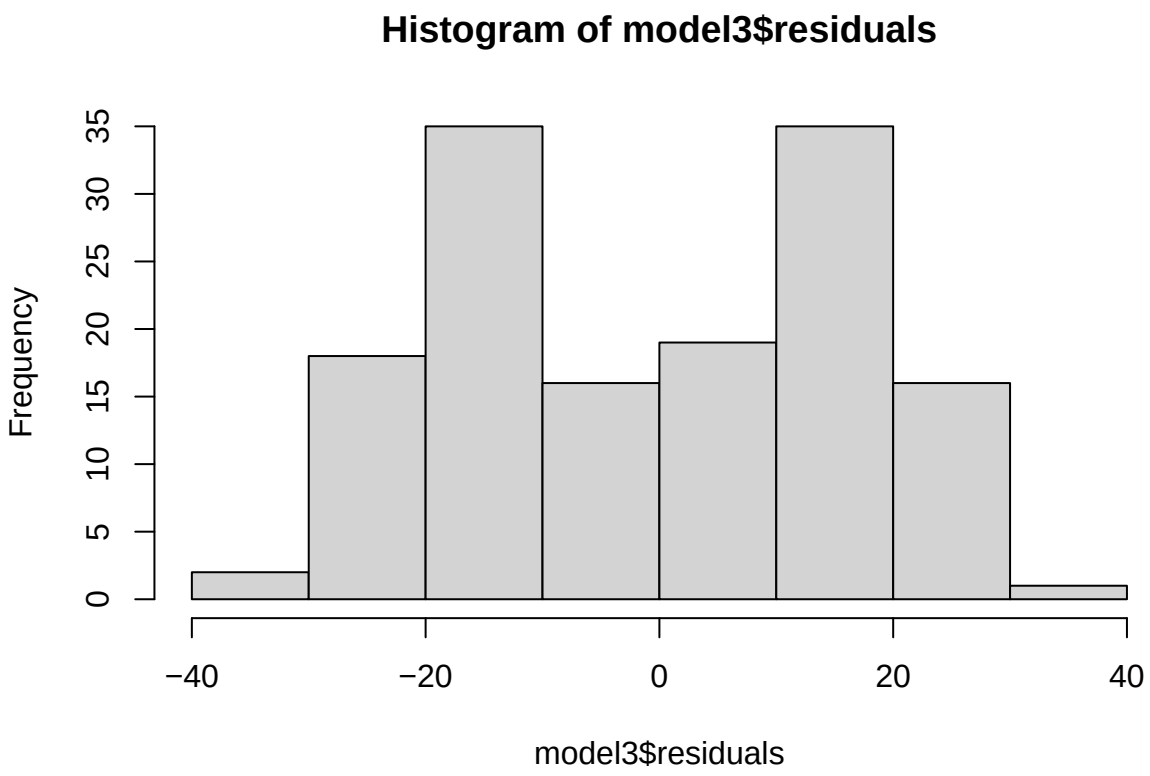
Data set 3 No

Based on plots of data3, there is no linearity in the data, the residuals plot does not show a normal distribution.

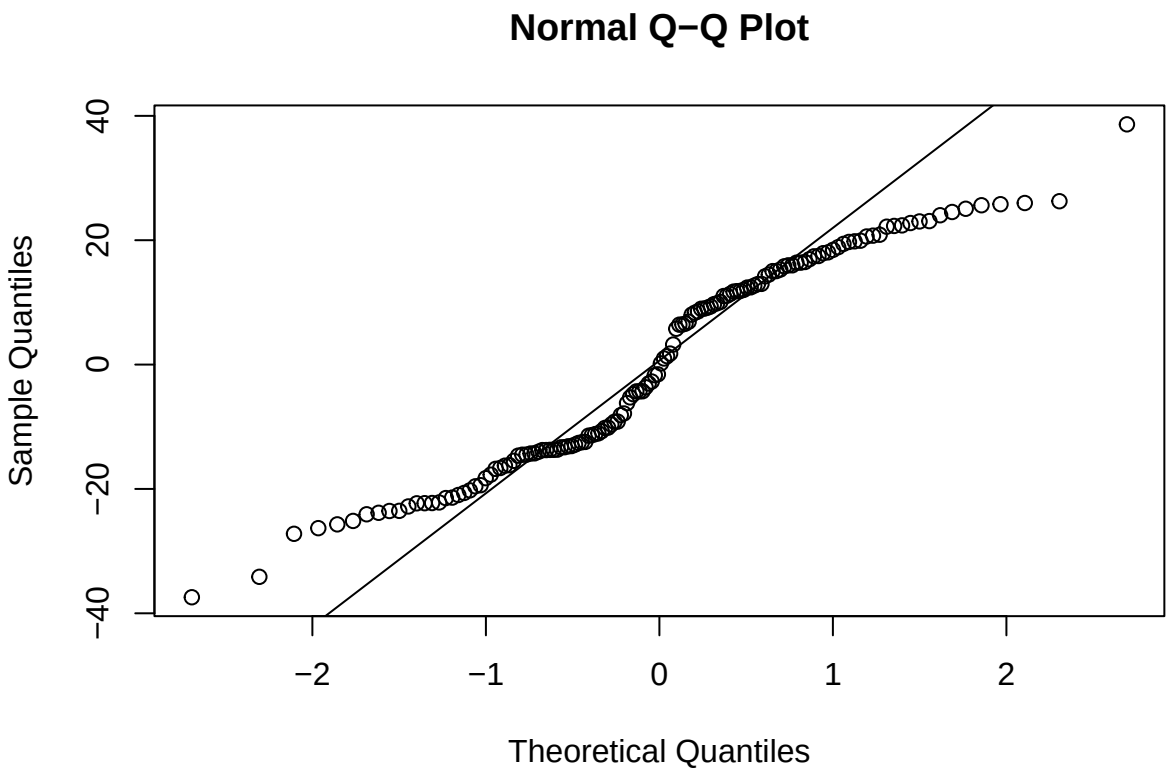
```
plot(x ~ y, data3)
abline(model3)
```




```
hist(model3$residuals)
```




```
qqnorm(model3$residuals)
qqline(model3$residuals)
```



h. Why it is important to include appropriate visualizations when analyzing data? Be sure to ground your reasoning in the context of the analyses completed above. Include any visualization(s) you create. (15 points) Data visualizations helps the researcher display data findings into a form easier to understand by highlighting the trends and the outliers. The graphs that I used for this test, helped me find the necessary conditions need it to create a linear regression model. I used an scatterplot using to "abline" function to display a line in the graph to find out the linearity between X and Y in the data. Also I used a histogram and a QQ plot of the residuals to find if the data is distributed normally or not.