Data 606 Final Exam

Armenoush Aslanian-Persico

Part I.

a. Distribution 1 is right-skewed with a long tail, unimodal, nearly normal, but not completely symmetric. Distribution 2 is normal, unimodal, and symmetric.

b. Standard deviation is the square root of the variance, which is the average of the deviations of the observations. The sampling distribution has much less variance and therefore a smaller standard deviation.

c. The Central Limit Theorem states: "If a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by a normal model."

Part II.

```
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(data1$x)
```

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

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

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

```r
mean(data2$x)
```

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

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

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

```r
mean(data3$x)
```

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

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

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

```r
mean(data4$x)
```

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

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

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

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

```r
median(data1$x)
```

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

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

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

```r
median(data2$x)
```

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

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

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

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

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

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

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

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

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

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

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

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

```r
x <- data1$x
n <- length(x)
sd1x <- sd(x)
sd1x
```

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

```r
#Alternate method of calculation

sd1xm <- sqrt(sum((x - mean(x))^2) / (n - 1))
sd1xm
```

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

```r
y <- data1$y
sd1y <- sd(y)
sd1y
```

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

```r
x <- data2$x
sd2x <- sd(x)
sd2x
```

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

```
y <- data2$y
sd2y <- sd(y)
sd2y
```

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

```
x <- data3$x
sd3x <- sd(x)
sd3x
```

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

```
y <- data3$y
sd3y <- sd(y)
sd3y
```

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

```
x <- data4$x
n <-length(x)
sd4x <- sd(x)
sd4x
```

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

```
y <- data4$y
sd4y <- sd(y)
sd4y
```

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

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

d. The correlation (1 pt).

```
cor(data1$x, data1$y)
```

```
## [1] 0.82
```

```
cor(data2$x, data2$y)
```

```
## [1] 0.82
```

```
cor(data3$x, data3$y)
```

```
## [1] 0.82
```

```r
cor(data4$x, data4$y)
```

```
## [1] 0.82
```

e. Linear regression equation (2 pts).

$$y = 3.002 + 0.500x$$

f. R-Squared (2 pts).

```r
m1<- lm(y  ~ x, data = data1)
m2<- lm(y  ~ x, data = data2)
m3<- lm(y  ~ x, data = data3)
m4<- lm(y  ~ x, data = data4)

summary(m1)
```

```
##
## Call:
## lm(formula = y ~ x, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9213 -0.4558 -0.0414  0.7094  1.8388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.000      1.125    2.67   0.0257 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 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
```

```r
summary(m2)
```

```
##
## Call:
## lm(formula = y ~ x, data = data2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.901 -0.761  0.129  0.949  1.269
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125    2.67   0.0258 *
```

```
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 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
```

```
summary(m3)
```

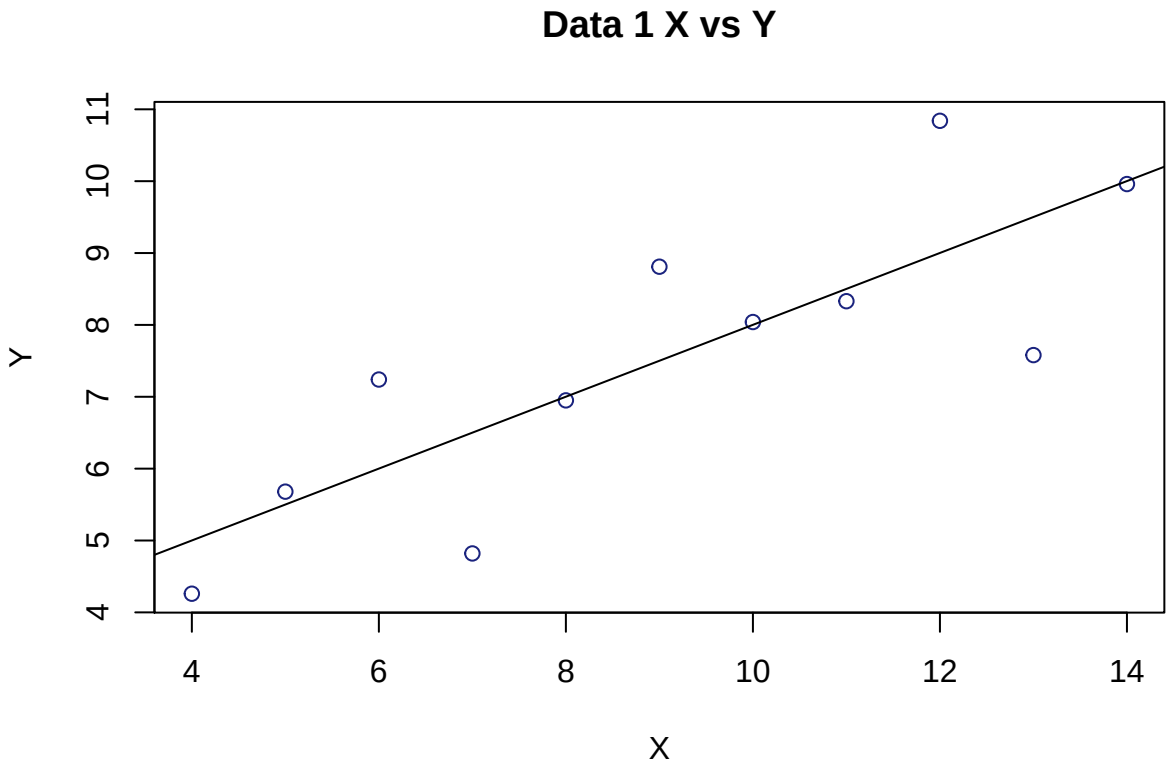
```
##
## Call:
## lm(formula = y ~ x, data = data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.159 -0.615 -0.230  0.154  3.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002      1.124    2.67   0.0256 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 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
```

```
summary(m4)
```

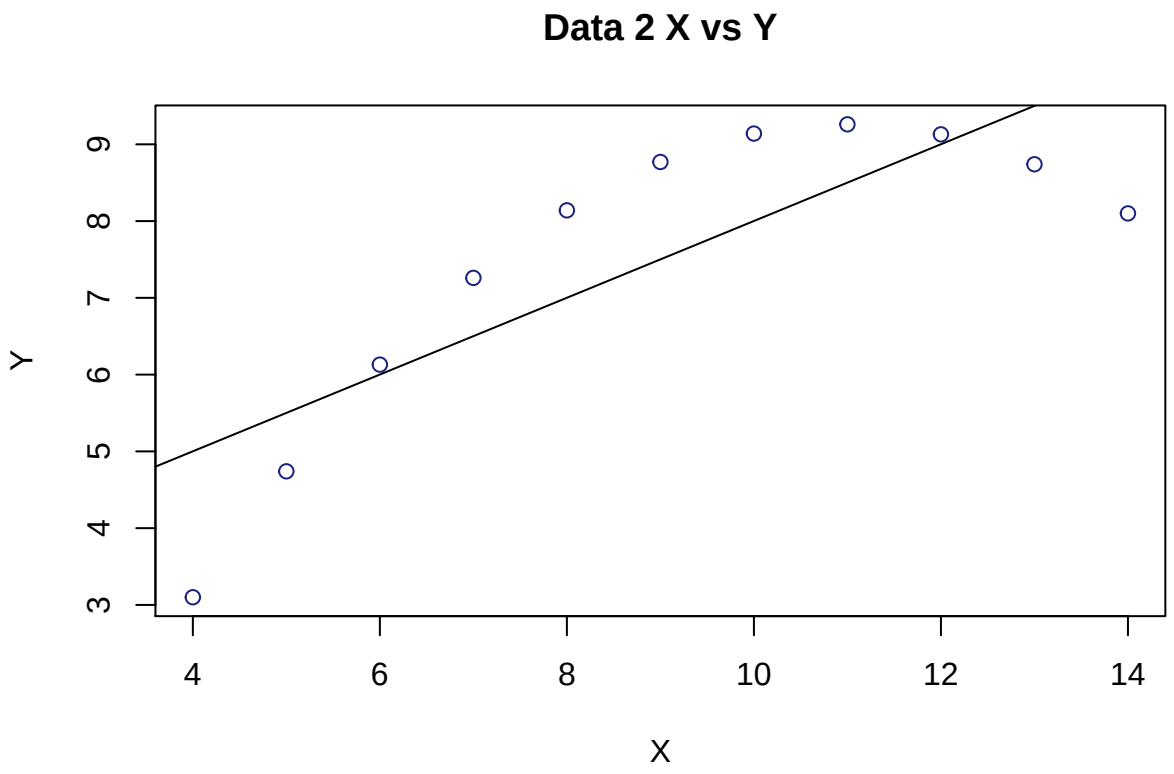
```
##
## Call:
## lm(formula = y ~ x, data = data4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002      1.124    2.67   0.0256 *
## x              0.500      0.118    4.24   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 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
```

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! (4 pts)

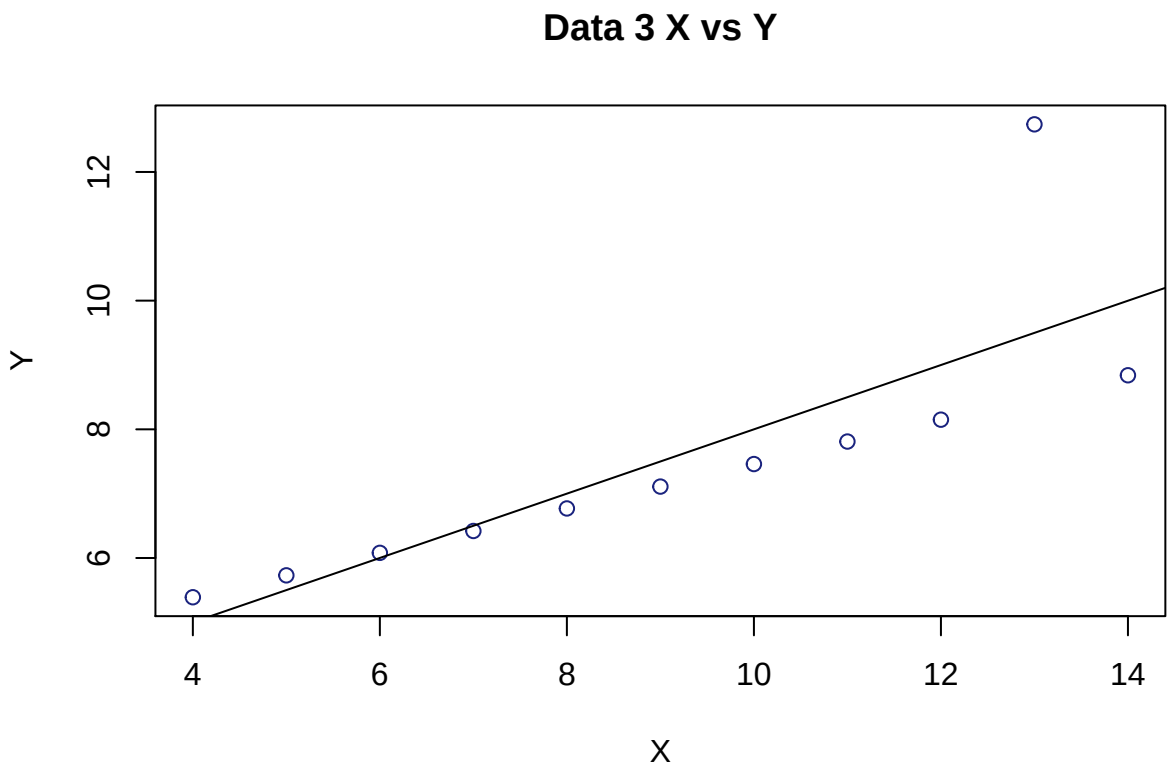
```
m1<- lm(y  ~ x, data = data1)
plot(data1$y ~ data1$x , col="#1a237e", main="Data 1 X vs Y", xlab="X", ylab="Y")
abline(m1)
```



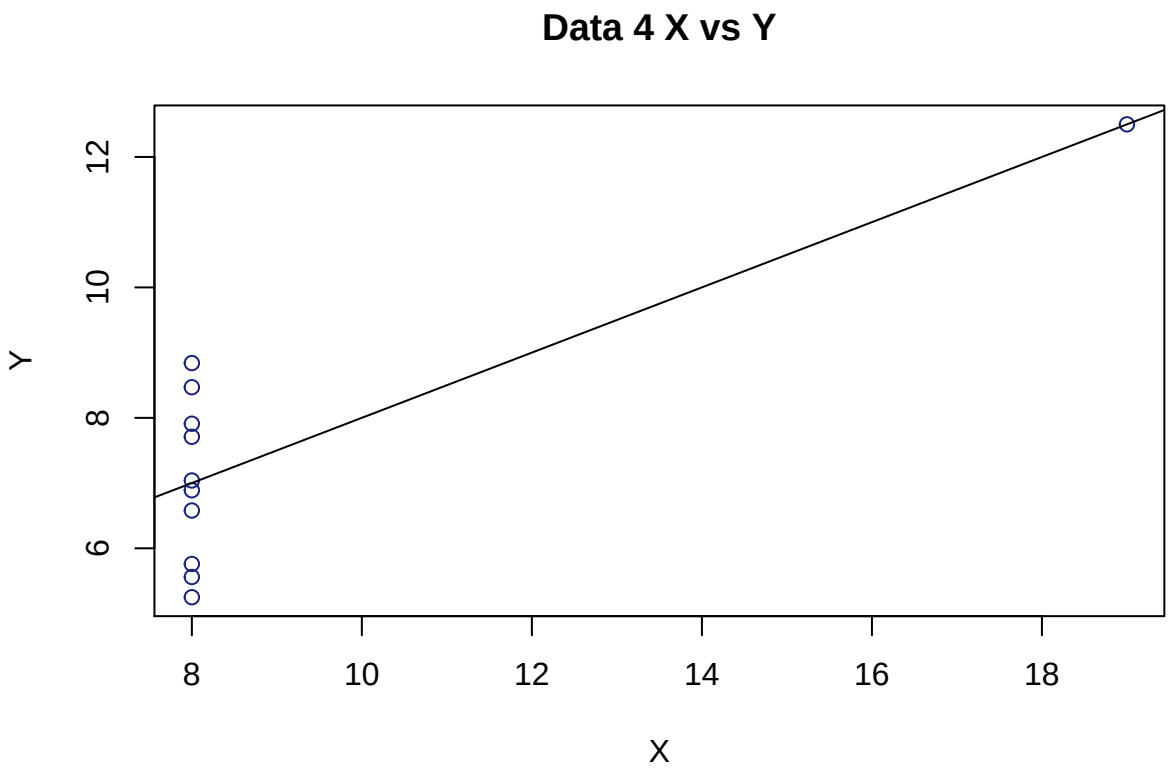

```
m2<- lm(y  ~ x, data = data2)
plot(data2$y ~ data2$x , col="#1a237e", main="Data 2 X vs Y", xlab="X", ylab="Y")
abline(m2)
```



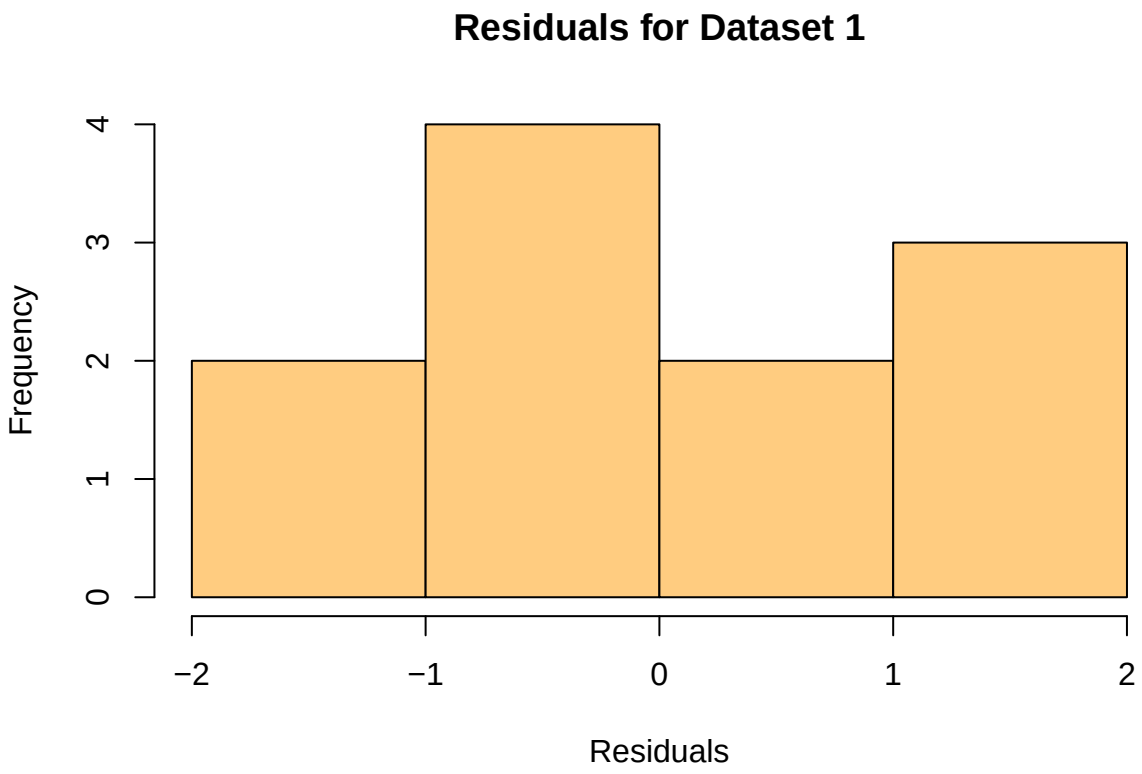

```
m3<- lm(y  ~ x, data = data3)
plot(data3$y ~ data3$x , col="#1a237e", main="Data 3 X vs Y", xlab="X", ylab="Y")
abline(m3)
```



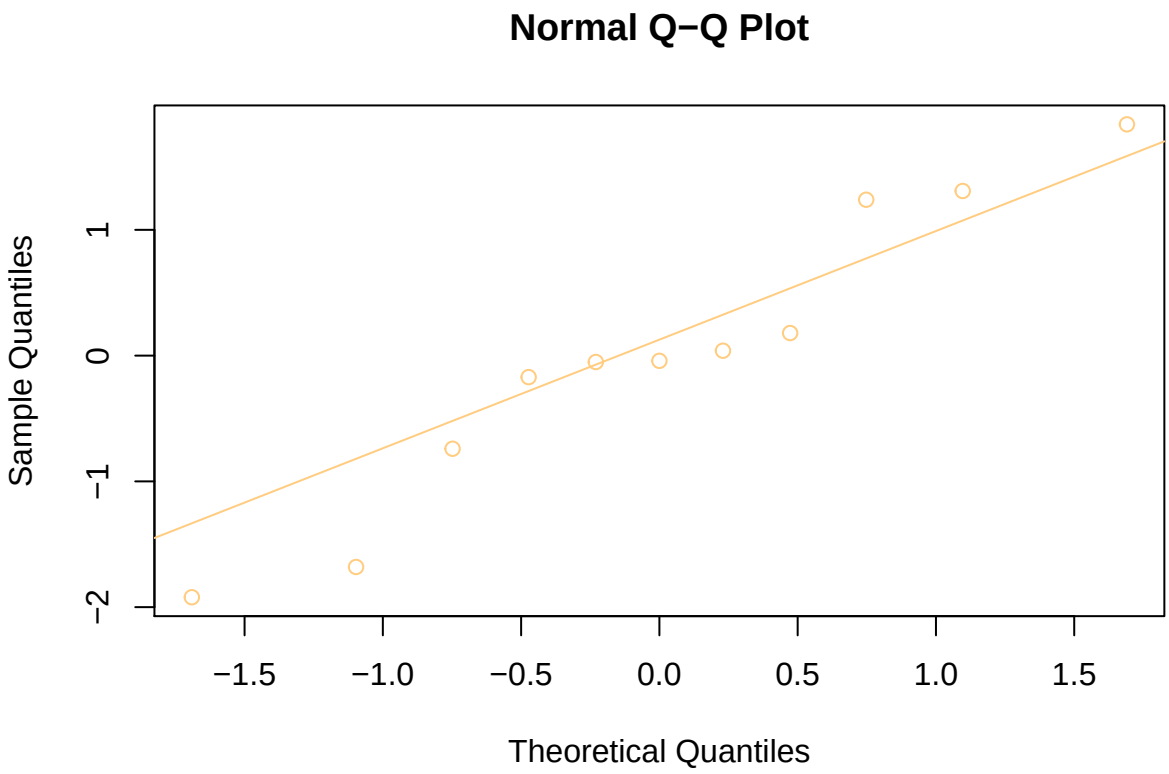

```
m4<- lm(y  ~ x, data = data4)
plot(data4$y ~ data4$x , col="#1a237e", main="Data 4 X vs Y", xlab="X", ylab="Y")
abline(m4)
```



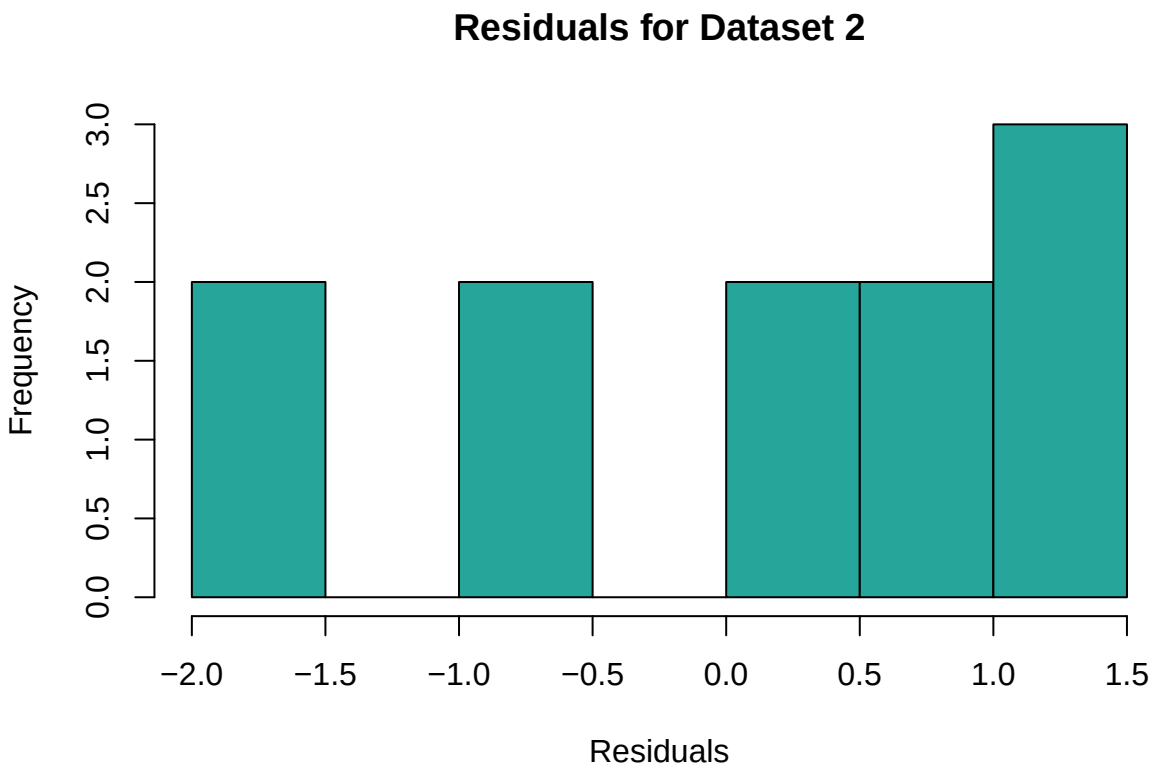

```r
hist(m1$residuals, main="Residuals for Dataset 1", xlab="Residuals" , col="#ffcc80")
```



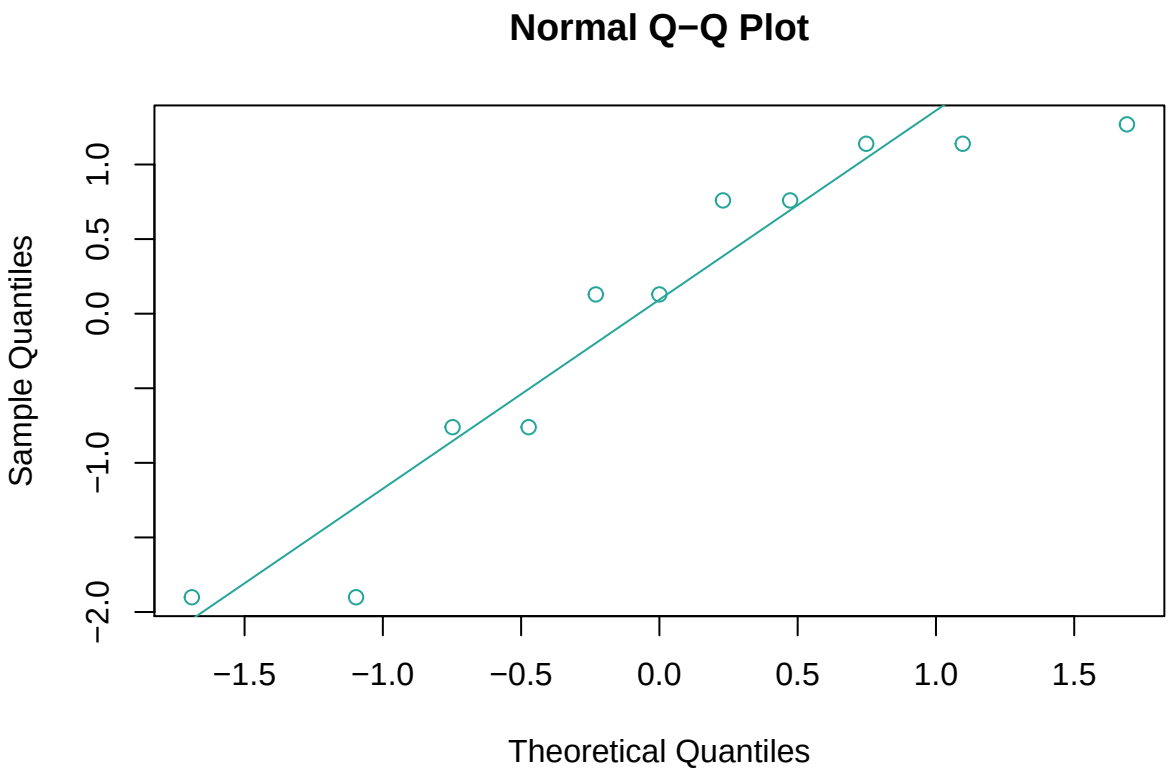

```
qqnorm(m1$residuals, col="#ffcc80")
qqline(m1$residuals, col="#ffcc80")
```



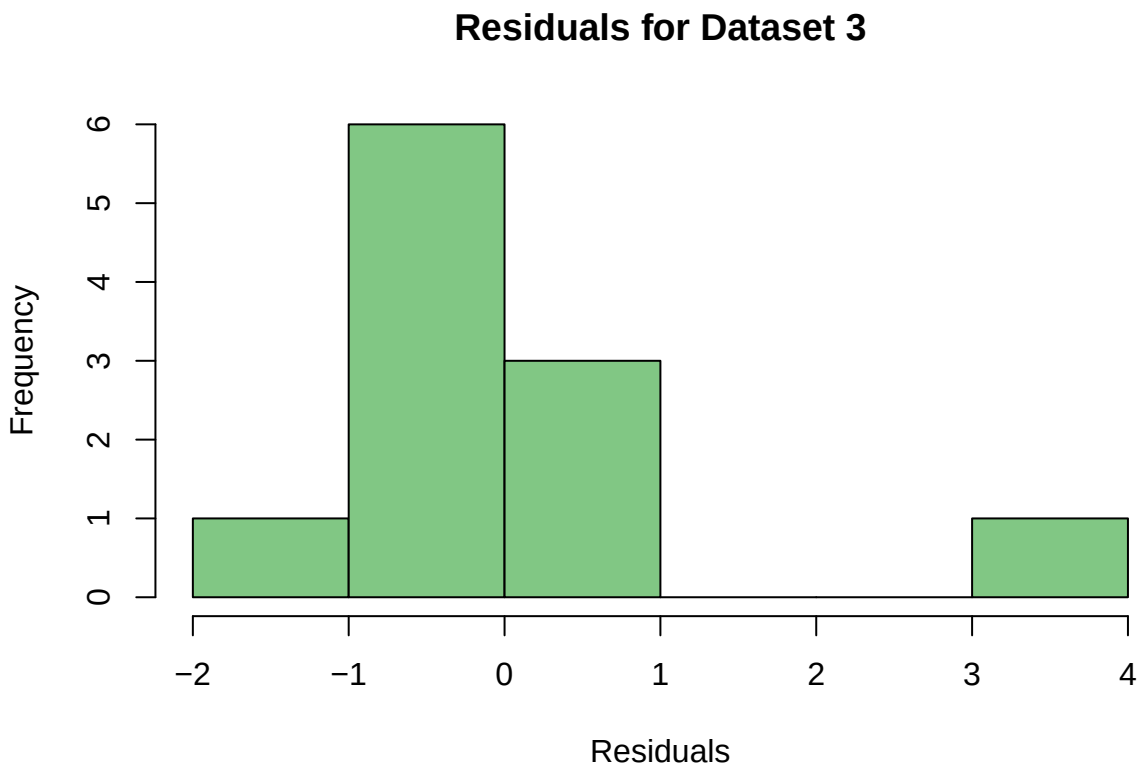

```r
hist(m2$residuals, main="Residuals for Dataset 2", xlab="Residuals" , col="#26a69a")
```



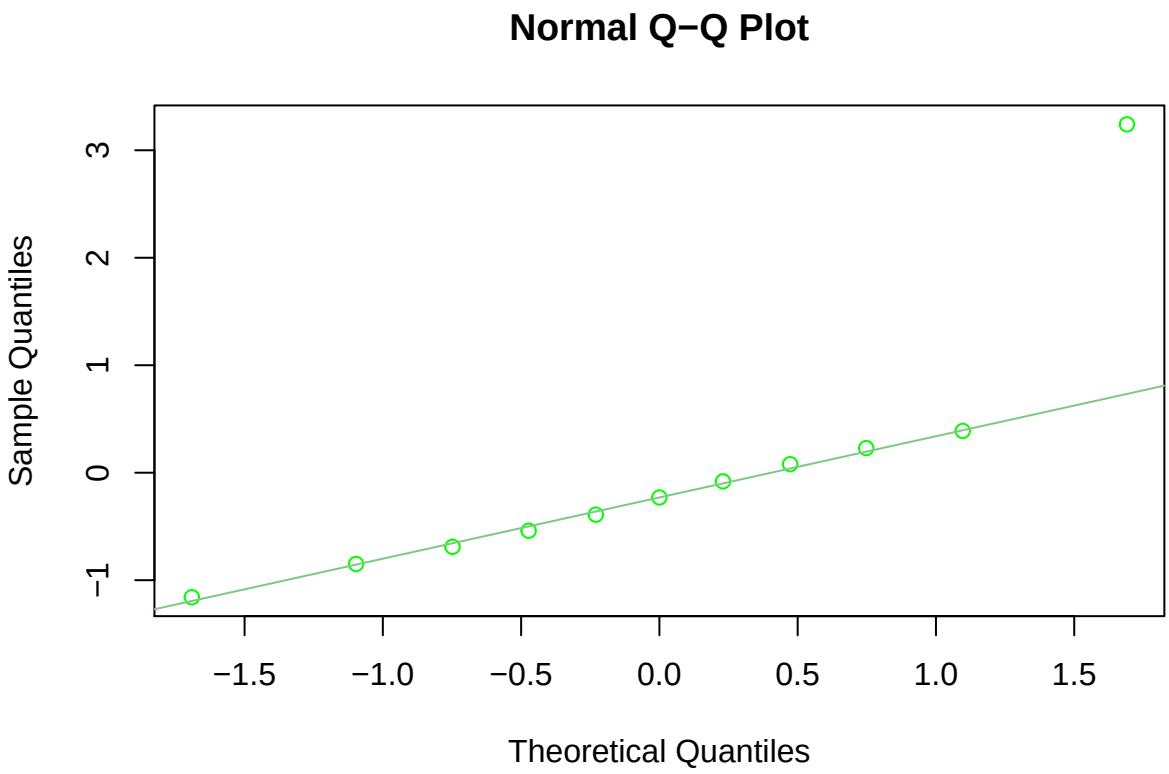

```
qqnorm(m2$residuals, col="#26a69a")
qqline(m2$residuals, col="#26a69a")
```



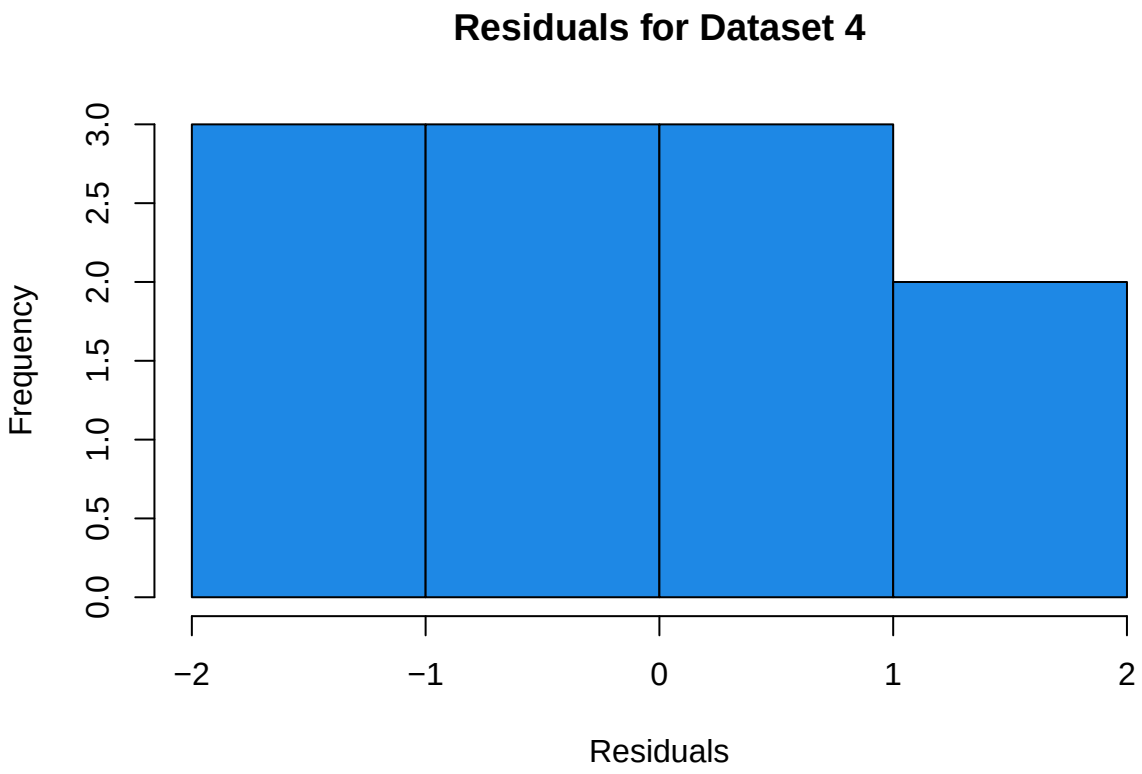

```r
hist(m3$residuals, main="Residuals for Dataset 3", xlab="Residuals" , col="#81c784")
```



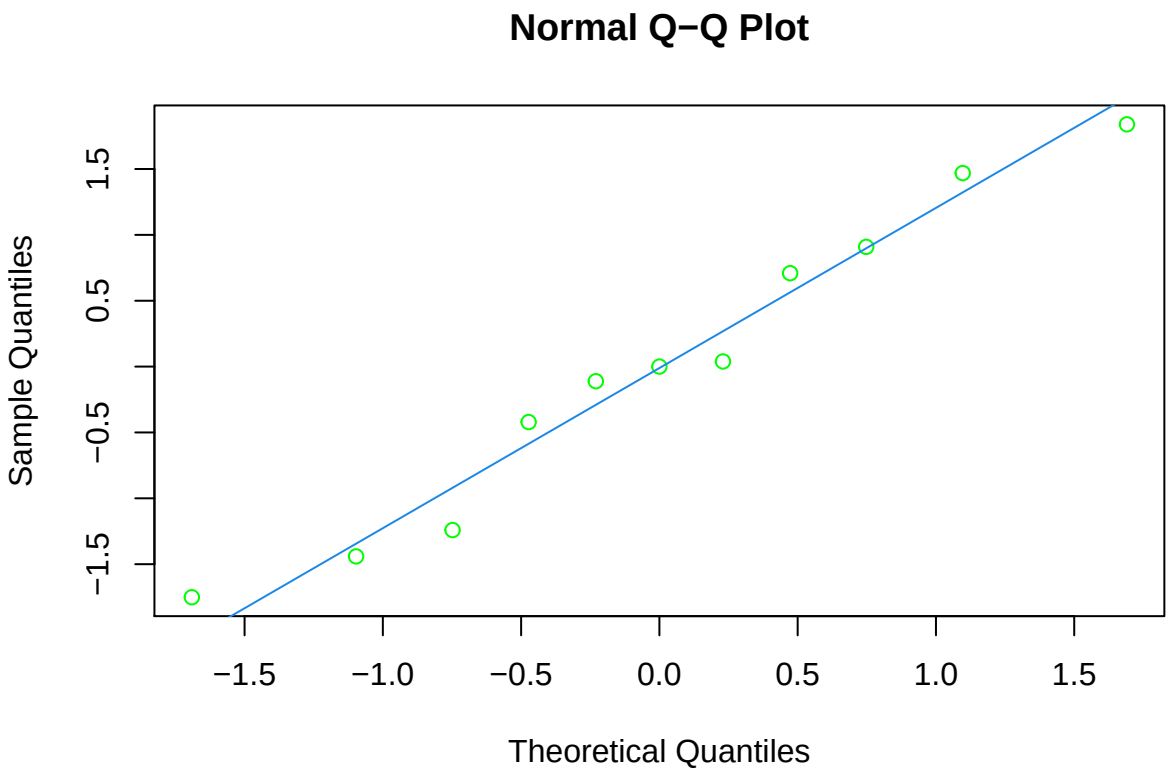

```r
qqnorm(m3$residuals, col="green")
qqline(m3$residuals, col="#81c784")
```




```r
hist(m4$residuals, main="Residuals for Dataset 4", xlab="Residuals" , col="#1e88e5")
```




```
qqnorm(m4$residuals, col="green")
qqline(m4$residuals, col="#1e88e5")
```



A linear regression model is not appropriate for datasets 2 and 4, as seen by the abline and explained below.

To determine if a least squares fit is appropriate for these data, we have to see if it meets the requirements:

1. Linearity: not met in dataset 2.
2. Nearly normal residuals: not met in datasets 2 and 4.
3. Constant variability: not met in dataset 4.
4. Independent observations: assumed for all sets.

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

Visualizations are important because the numbers alone do not show the full picture of trends. Visualizing a linear best fit line across a parabolic curve will show us that where a linear regression model is not appropriate. Visualizing residuals will show us their normality easily. Even data that looks similar in dataframes and statistical analysis may look very different when plotted.