DATA 606 Spring 2019 - Final Exam

Santosh Cheruku

Part I

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

1. **b**
2. **a**
   In a left skewed distribution, median is always higher than mean
3. **a**
   Observational studies cannot prove causality, it has to be randomized.
4. **c**
   An association can be derived if the test statistic is higher.
5. **b**
   Outliers will be Q1 - 1.5 X IQR and Q3 + 1.5 X IQR, which will be 17.8 and 69 respectively.
6. **d**

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

Answer: **Figure A represents a highly right skewed model**
Figure B represents a normal distribution model

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

Answer: The means are similar since the 2nd distribution is a sample derived from the first.

7c. What is the statistical principal that describes this phenomenon (2 pts)? Answer: **Central Limit Theorem describes the phenomenon which says that "given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population"**

Part II

Consider the four 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.

```
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).

```
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)
data4.x.mean <- mean(data4$x)
data4.y.mean <- mean(data4$y)
```

b. The median (for x and y separately; 1 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)
data4.x.median <- median(data4$x)
data4.y.median <- median(data4$y)
```

c. The standard deviation (for x and y separately; 1 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)
data4.x.sd <- sd(data4$x)
data4.y.sd <- sd(data4$y)
```

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

d. The correlation (1 pt).

```
data1.correlation <- cor(data1$x, data1$y)
data2.correlation <- cor(data2$x, data2$y)
data3.correlation <- cor(data3$x, data3$y)
data4.correlation <- cor(data4$x, data4$y)
```

e. Linear regression equation (2 pts).

```
data1.slope <- coef(lm(data = data1, y~x))["x"]
data2.slope <- coef(lm(data = data2, y~x))["x"]
data3.slope <- coef(lm(data = data3, y~x))["x"]
data4.slope <- coef(lm(data = data4, y~x))["x"]

data1.intercept <- coef(lm(data = data1, y~x))[1]
data2.intercept <- coef(lm(data = data2, y~x))[1]
data3.intercept <- coef(lm(data = data4, y~x))[1]
data4.intercept <- coef(lm(data = data1, y~x))[1]
```

f. R-Squared (2 pts).

```
data1.rsquared <- summary(lm(data = data1, y~x))$r.squared
data2.rsquared <- summary(lm(data = data2, y~x))$r.squared
data3.rsquared <- summary(lm(data = data3, y~x))$r.squared
data4.rsquared <- summary(lm(data = data4, y~x))$r.squared
```

```
## Warning: package 'knitr' was built under R version 3.5.2
```

```
## Warning: package 'kableExtra' was built under R version 3.5.3
```

	Data 1		Data 2		Data 3		Data 4	
	x	**y**	**x**	**y**	**x**	**y**	**x**	**y**
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Median	9.00	7.58	9.00	8.14	9.00	7.11	8.00	7.04
SD	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
r	0.82		0.82		0.82		0.82	
Intercept	3.00		3.00		3.00		3.00	
Slope	0.50		0.50		0.50		0.50	
R-Squared	0.67		0.67		0.67		0.67	

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)

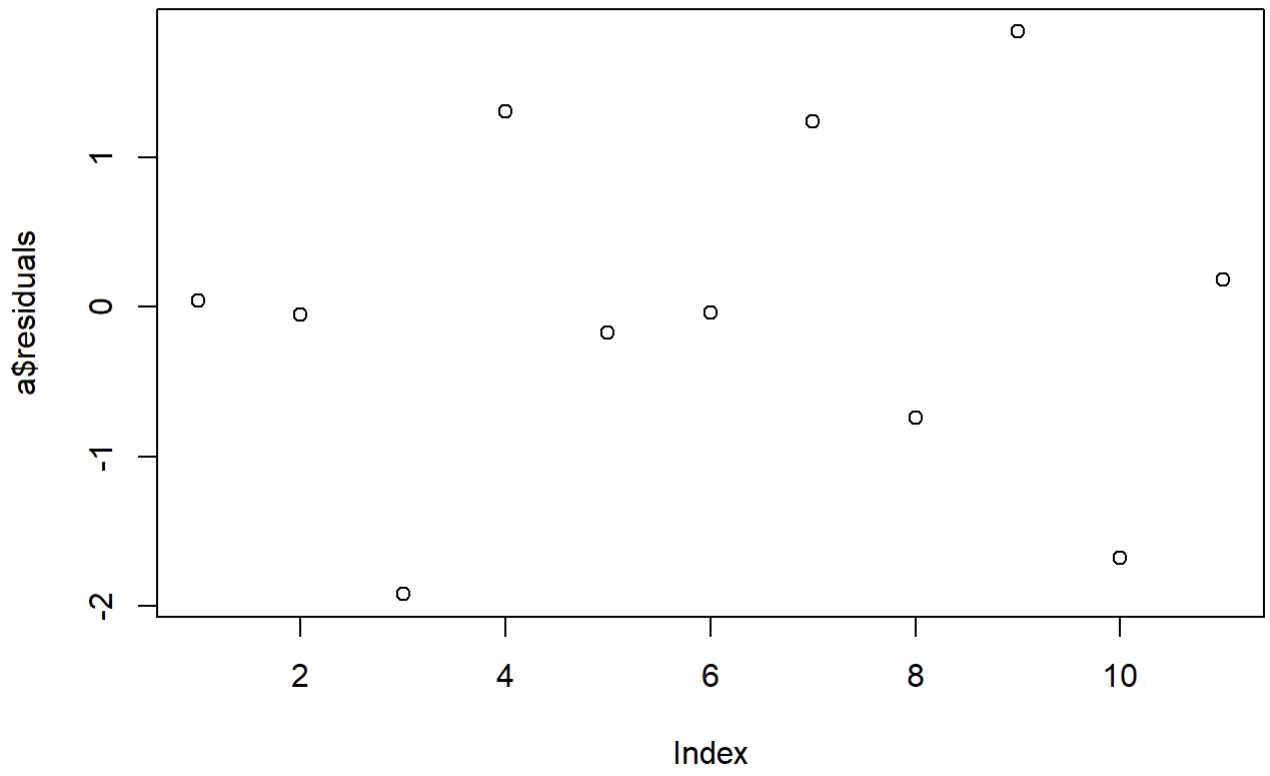
```
Answer: The linear regression model fits only if they appear to follow few conditions such as ha
ving a correlation, near normal residuals, and a linear or atleast near linear pattern. Here in
certain cases it doesn't seem to be the case.
```

Data1: Suitable

```
library(ggplot2)
cor(data1$x, data1$y)^2
```

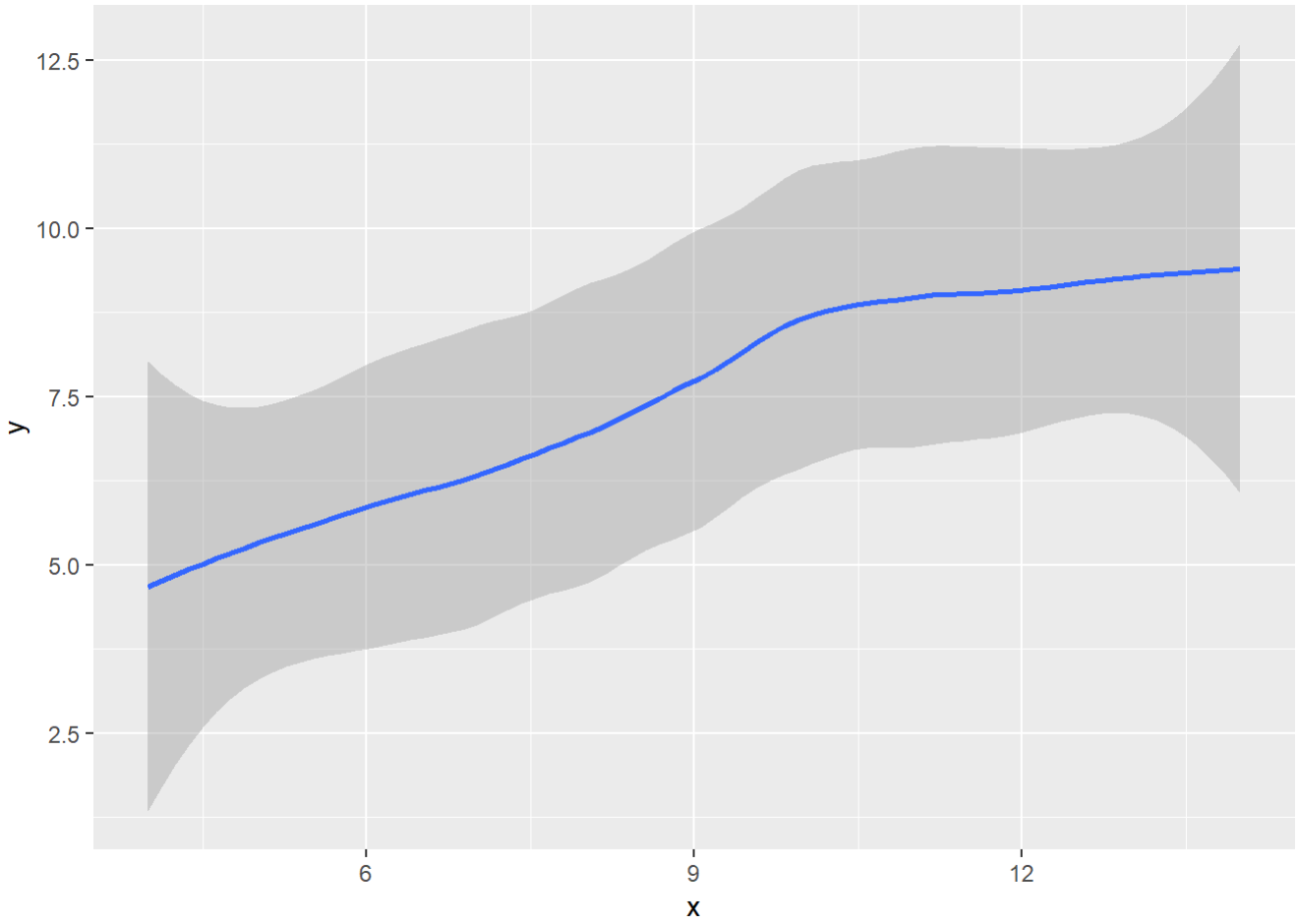
```
## [1] 0.67
```

```
a <- lm(formula = y ~ x, data = data1)
plot(a$residuals)
```




```
ggplot(data1, aes(x,y)) + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

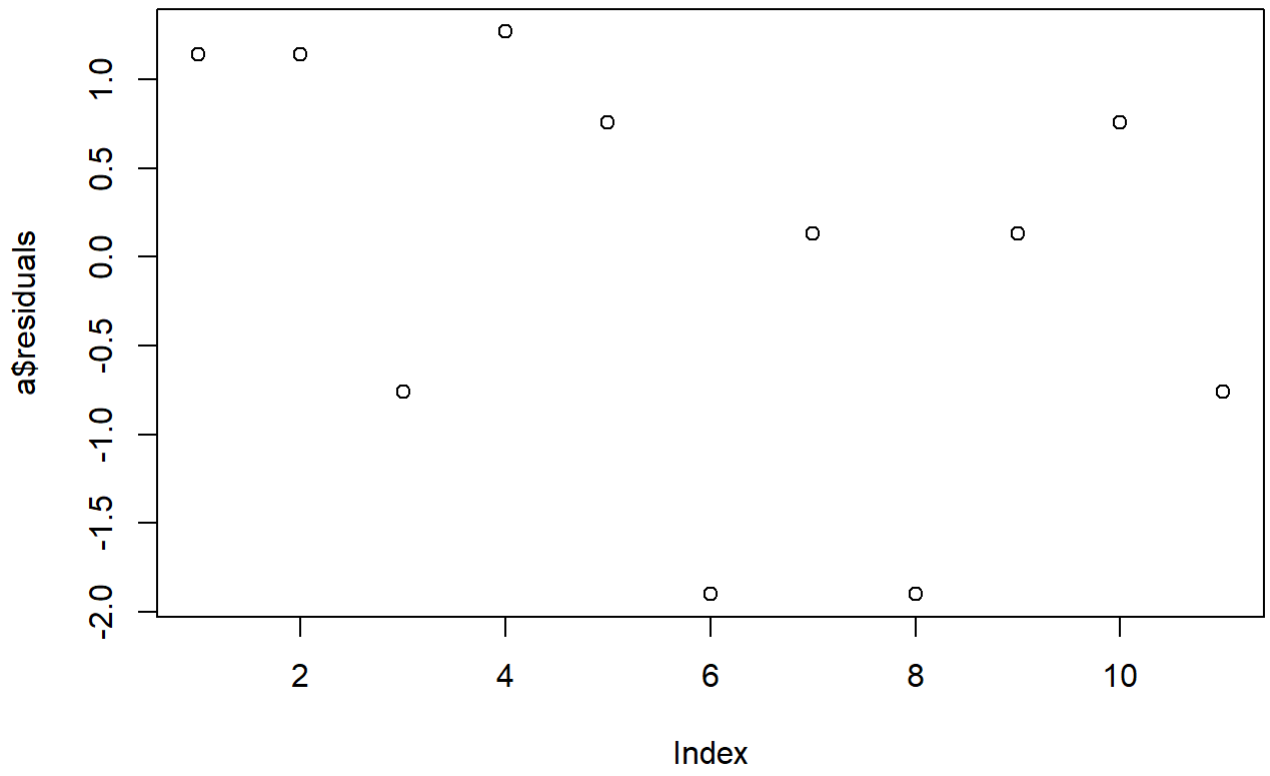


Data2: Not suitable

```
cor(data2$x, data2$y)^2
```

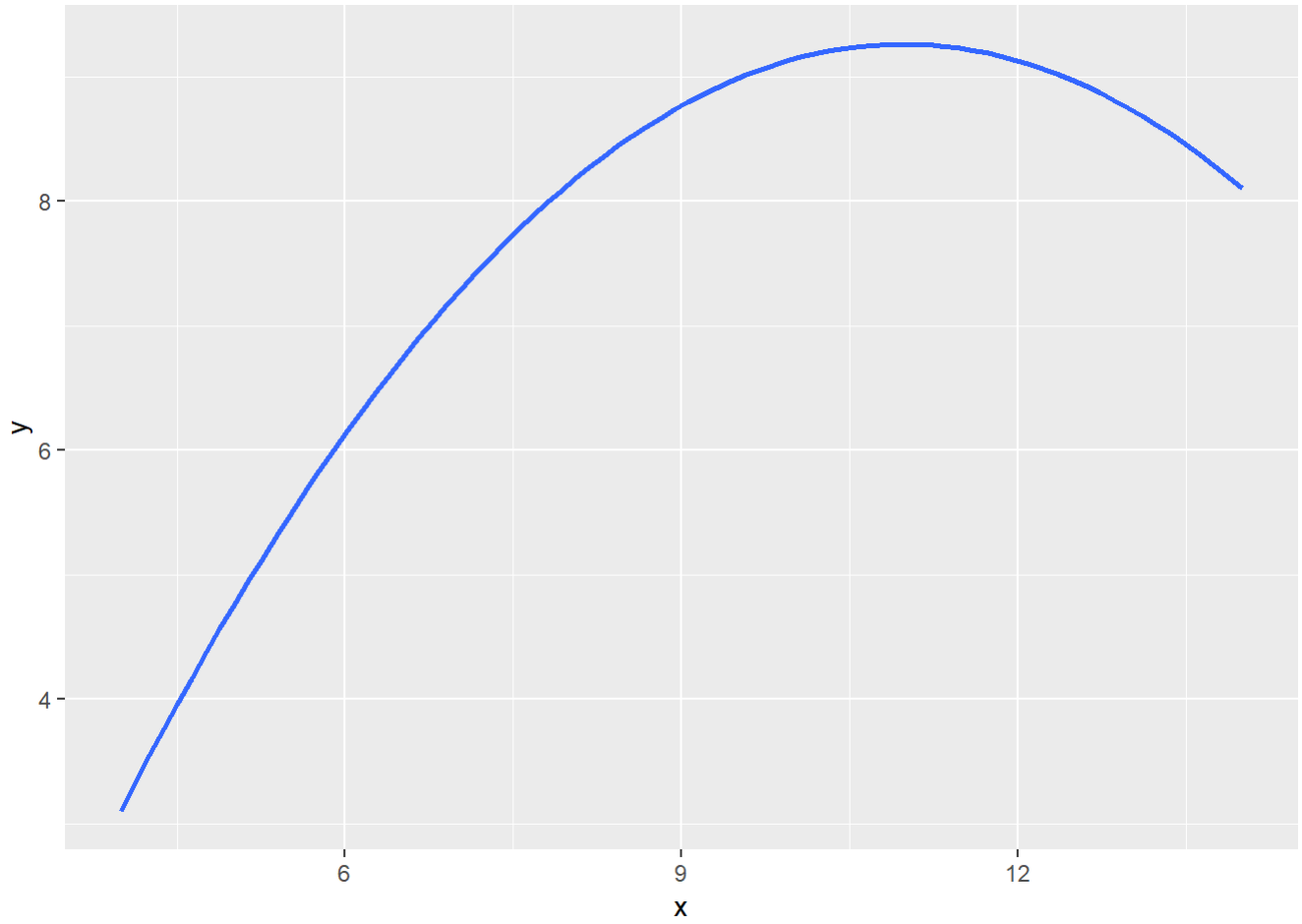
```
## [1] 0.67
```

```
a <- lm(formula = y ~ x, data = data2)
plot(a$residuals)
```




```
ggplot(data2, aes(x, y)) + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

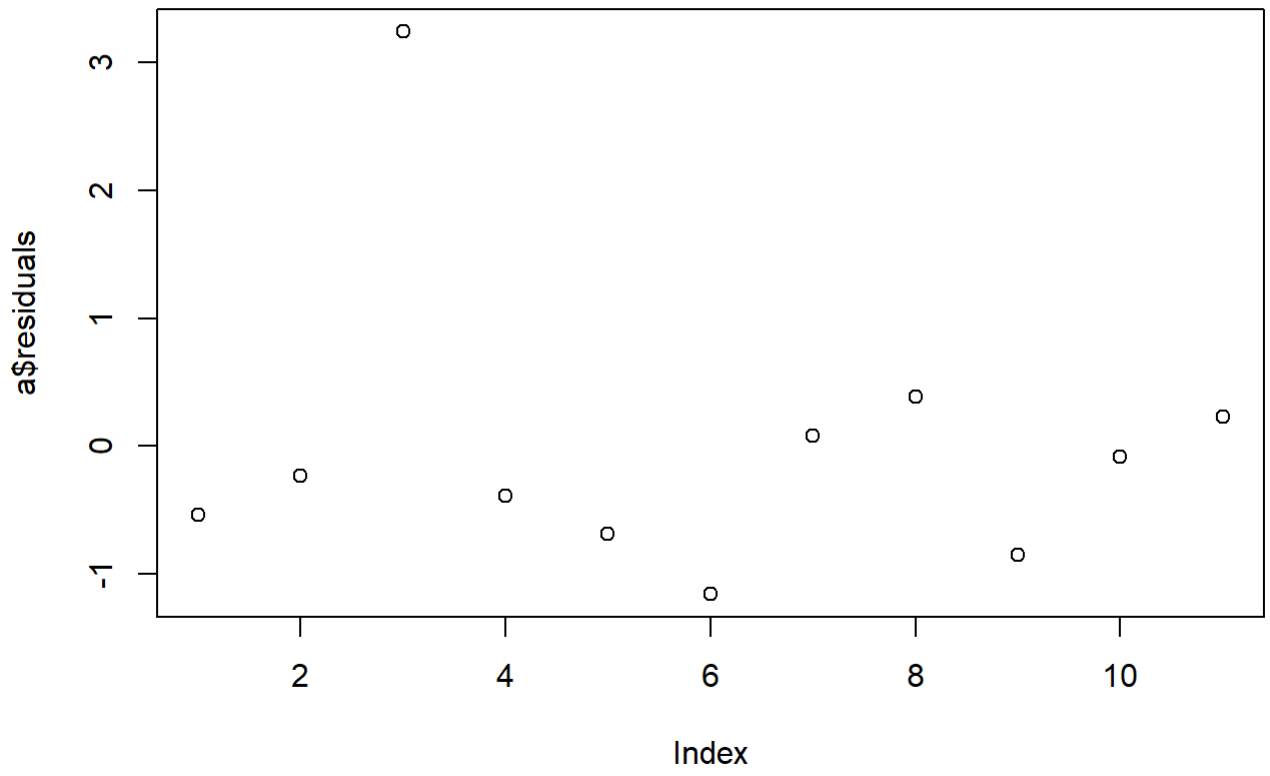


Data3: Suitable (Excluding single outlier residual)

```
cor(data3$x, data3$y)^2
```

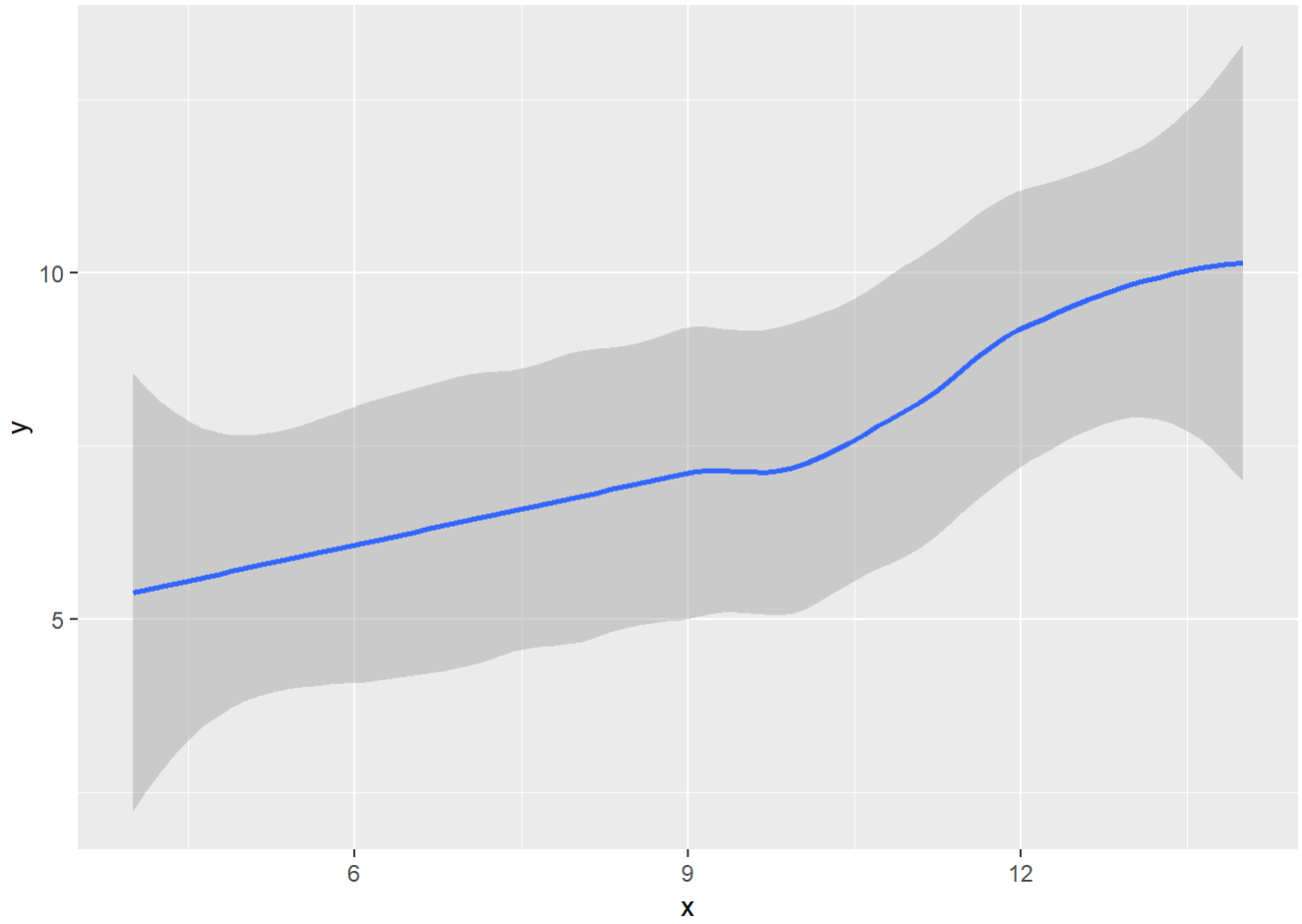
```
## [1] 0.67
```

```
a <- lm(formula = y ~ x, data = data3)
plot(a$residuals)
```




```
ggplot(data3, aes(x, y)) + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

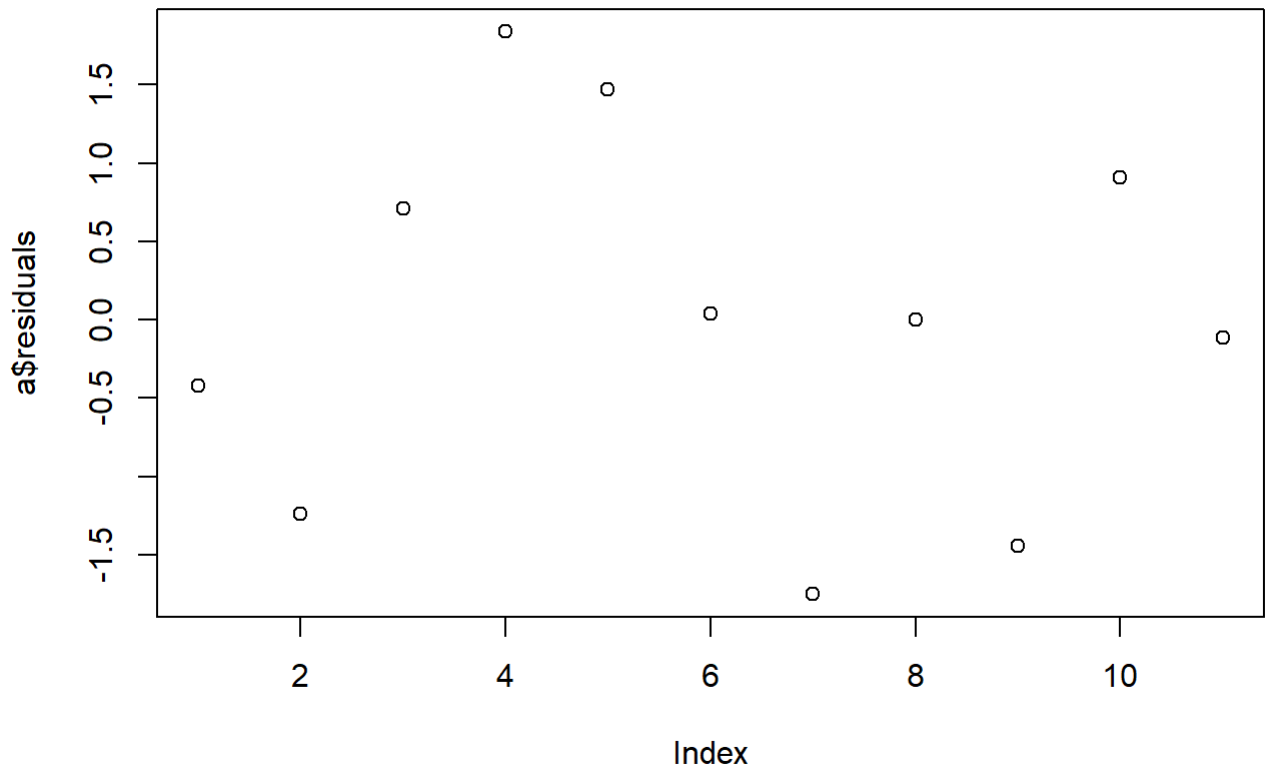


Data4: Not suitable

```
cor(data4$x, data4$y)^2
```

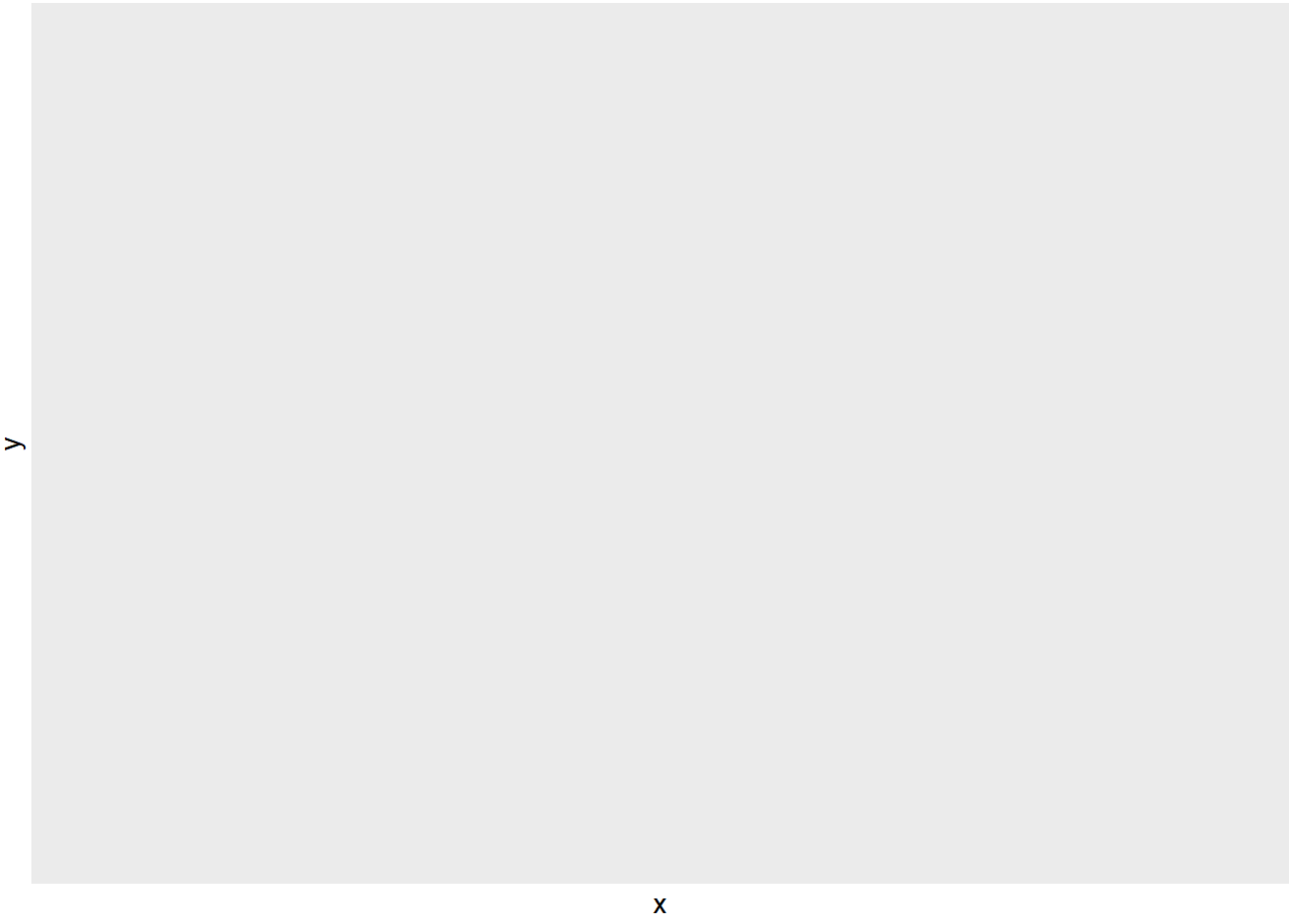
```
## [1] 0.67
```

```
a <- lm(formula = y ~ x, data = data4)
plot(a$residuals)
```




```
ggplot(data4, aes(x, y)) + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



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

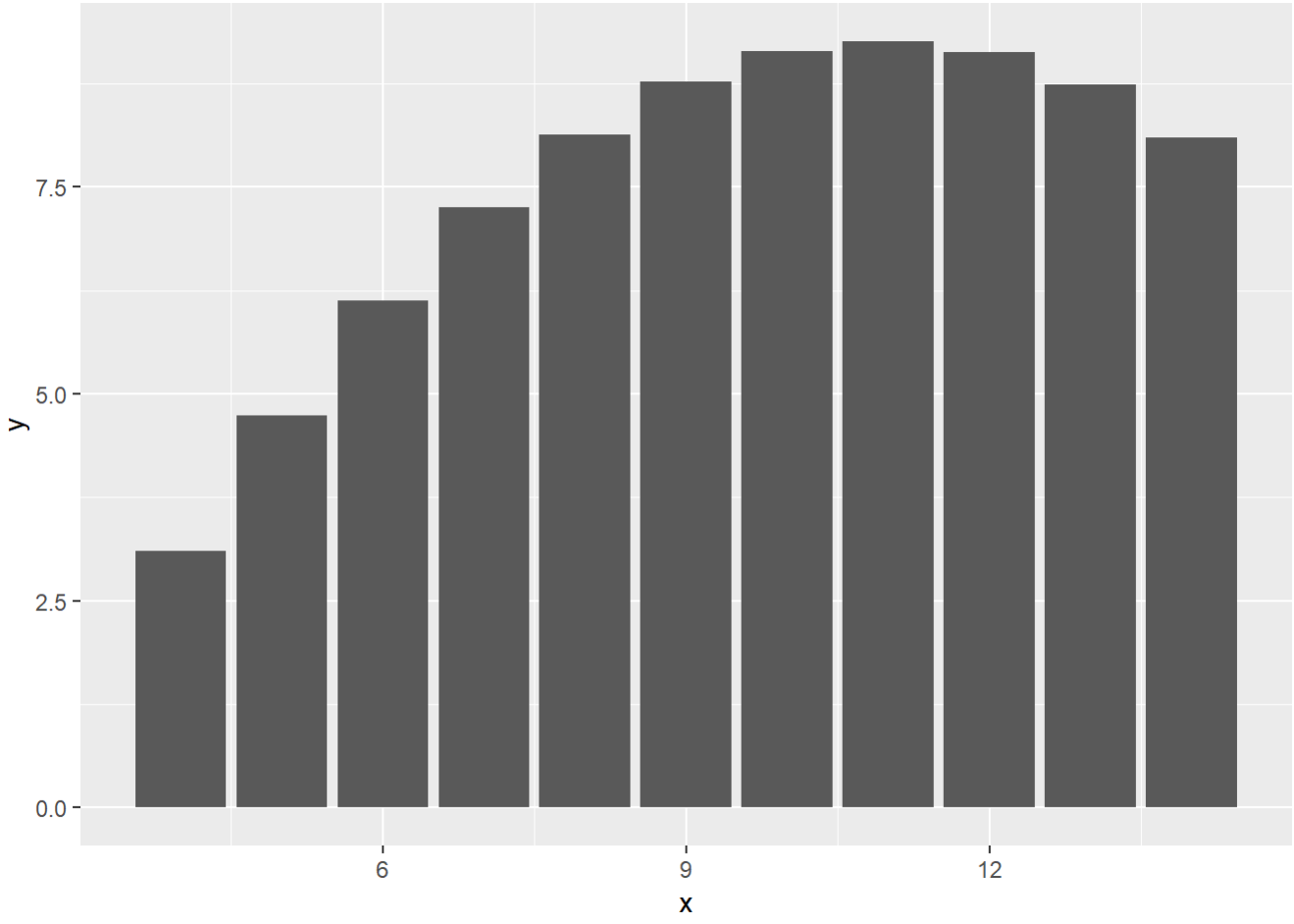
```
Answer:
    1) Including appropriate visualizations makes lot of difference analyzing the data because t
he underlying patterns cannot be identified by looking at large amount of raw data lying in numb
ers, but if they are converted into a graph or a plot they make more sense.
    2) The audience always need not be technically proficient in numbers and also may not be wil
ling to spend time to understand them. Therefore, it makes sense to project the same data visual
ly since it's universally well understood and can be easily grasped.
```

Below two graphs visually represent the raw numbers and make more sense.

```
data2
```

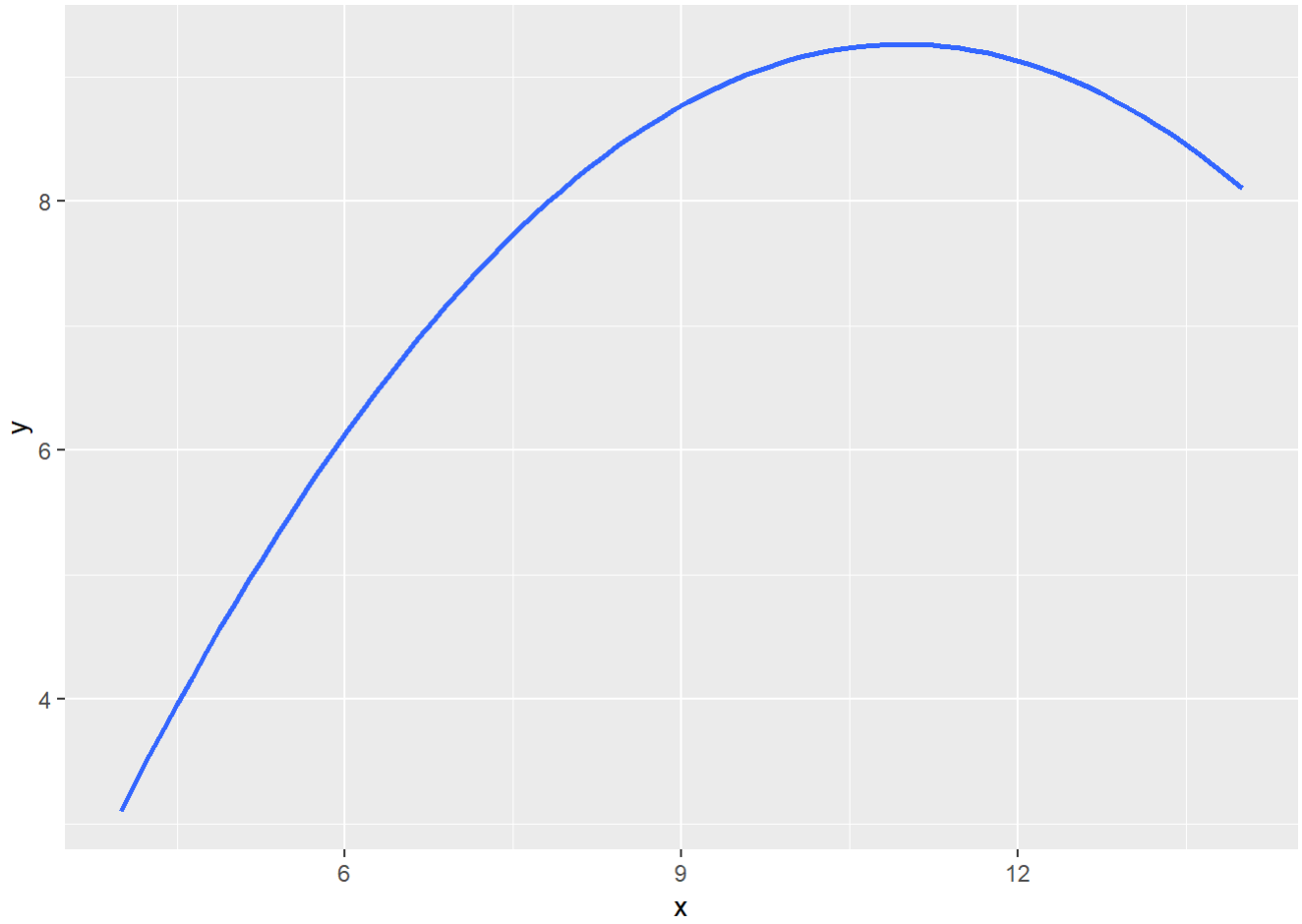
```
##     x   y
## 1  10 9.1
## 2   8 8.1
## 3  13 8.7
## 4   9 8.8
## 5  11 9.3
## 6  14 8.1
## 7   6 6.1
## 8   4 3.1
## 9  12 9.1
## 10  7 7.3
## 11  5 4.7
```

```
ggplot(data2, aes(x, y)) + geom_col()
```




```
ggplot(data2, aes(x, y)) + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


8
6
4
y
6
9
12
x