Problem Set 1

ENGRD 2700

Cort Breuer

09/08/19

Question 1

```
quartet <- read.csv("Data/Quartet.csv")
```

Part A

x_1 Sample Mean $= \frac{10+8+13+9+11+14+6+4+12+7+5}{11} = 9$

x_1 Sample Median $= 4, 5, 6, 7, 8, \mathbf{9}, 10, 11, 12, 13, 14 = 9$

x_1 Sample Standard Deviation $= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{10}[(4-9)^2 + ...(14-9)^2]} = \sqrt{11} = 3.32$

```
colNames <- c("X1", "Y1", "X2", "Y2", "X3", "Y3", "X4", "Y4")

quartetMean <- apply(quartet, 2, mean)
quartetMedian <- apply(quartet, 2, median)
quartetSD <- apply(quartet, 2, sd)

summaryTable <- tibble(colNames, quartetMean, quartetMedian, quartetSD) %>% rename(Column = colNames, "

kable(summaryTable) %>% kable_styling(bootstrap_options = c("striped", "hover"))
```

Column	Sample Mean	Sample Median	Sample Standard Deviation
X1	9.0	9.00	3.32
Y1	7.5	7.58	2.03
X2	9.0	9.00	3.32
Y2	7.5	8.14	2.03
X3	9.0	9.00	3.32
Y3	7.5	7.11	2.03
X4	9.0	8.00	3.32
Y4	7.5	7.04	2.03

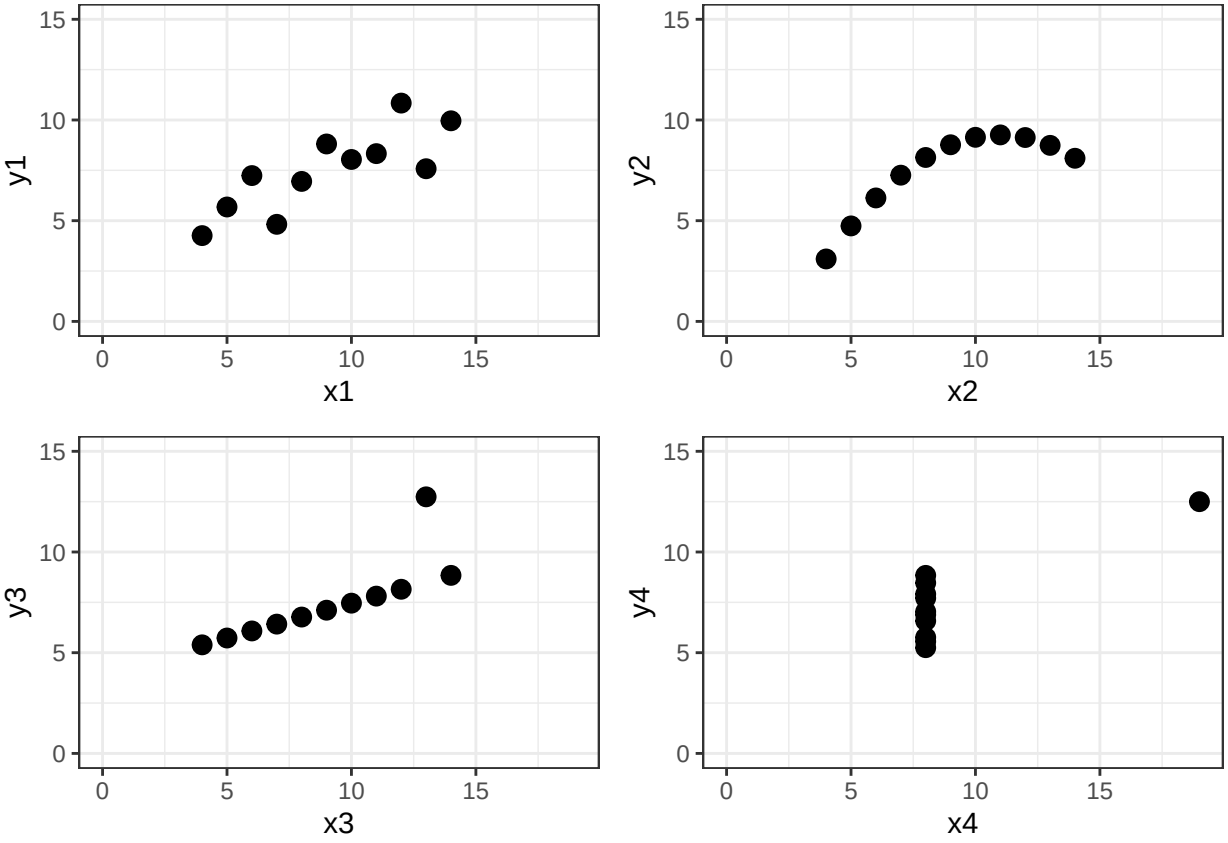
Part B

Based on the summary statistics the four data sets appear to be quite similar. The X's are all identical in mean, median, and standard deviation except for a small difference in the median for X4. The Y's are all quite similar too, with identical means and standard deviations, with small differences in their medians. It appears as though the X/Y data sets could be nearly identical to each other.

Part C

```
p1 <- ggplot(data = quartet, aes(x = x1, y = y1)) + geom_point(size = 3) + xlim(0, 19) + ylim(0, 15)
p2 <- ggplot(data = quartet, aes(x = x2, y = y2)) + geom_point(size = 3) + xlim(0, 19) + ylim(0, 15)
p3 <- ggplot(data = quartet, aes(x = x3, y = y3)) + geom_point(size = 3) + xlim(0, 19) + ylim(0, 15)
p4 <- ggplot(data = quartet, aes(x = x4, y = y4)) + geom_point(size = 3) + xlim(0, 19) + ylim(0, 15)

grid.arrange(p1, p2, p3, p4)
```



Part D

Based on the four plots generated, all 4 X/Y pairings are quite different. Each of the four plots indicates a different relationship, from a smooth curve to a set of similar points with a single outlier.

Part E

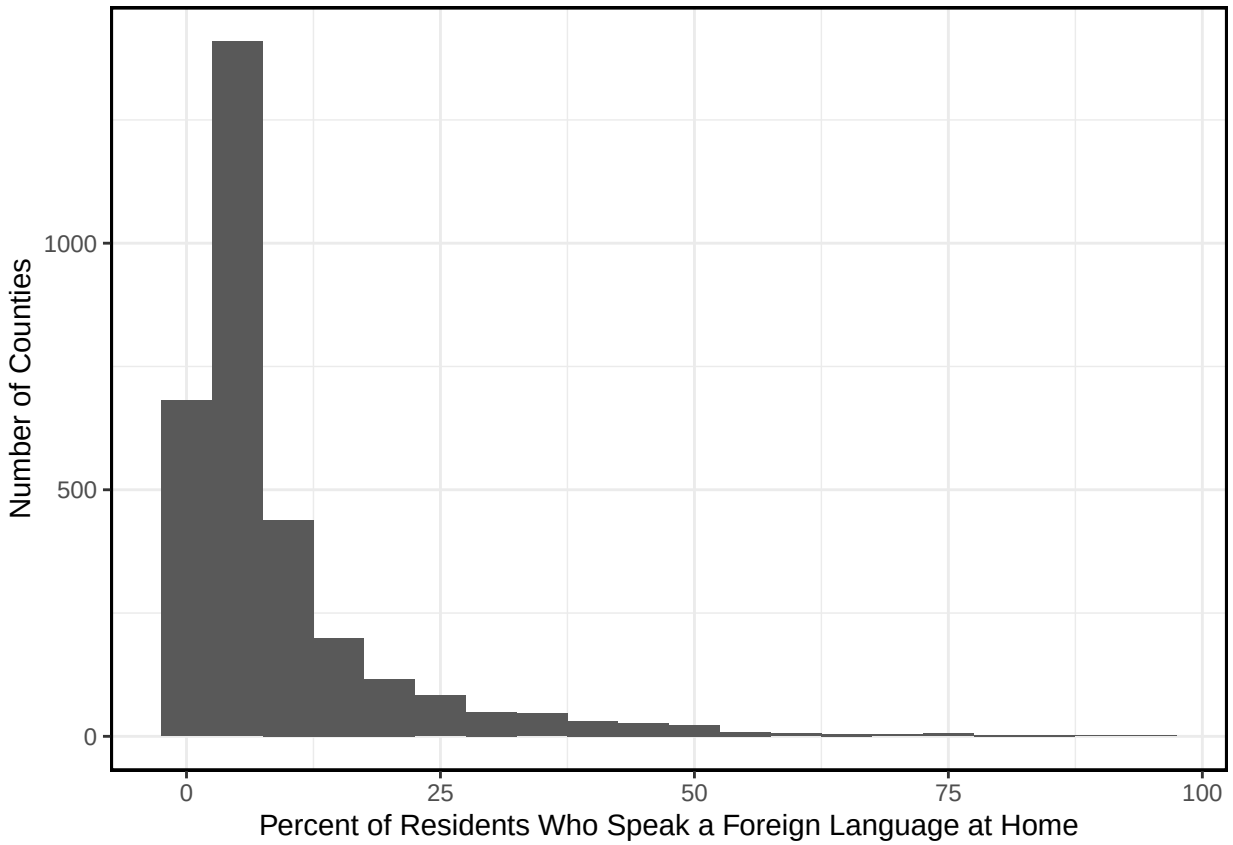
The stark difference between the nearly identical summary statistics and the very different plots demonstrates that trusting a set of summary statistics for the complete analysis of a data set could result in missing key information. More generally though, it is likely best to use a combination of analysis methods combining both summary statistics and plotting to get a well rounded picture of the data.

Question 2

```
countyData <- read.csv("Data/CountyData.csv")
countyData <- as_tibble(countyData)
```

Part A

```
ggplot(data = countyData) + geom_histogram(mapping = aes(foreign_spoken_at_home), binwidth = 5) + labs(
```



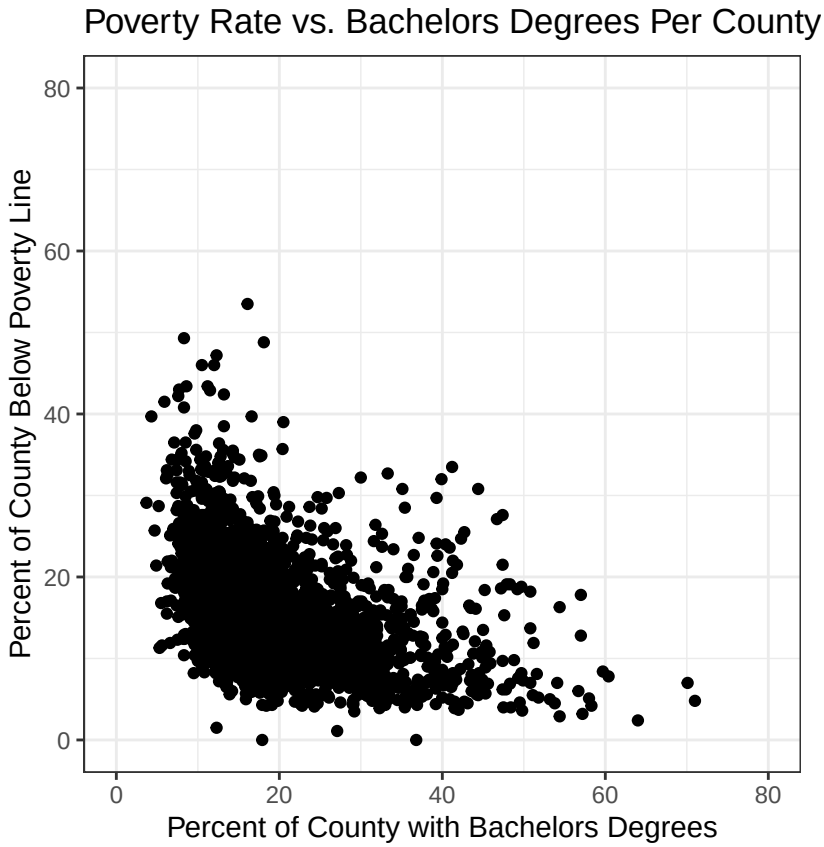
Part B

```
fedSpendingMedian <- median(countyData$fed_spending, na.rm = TRUE)
```

The median annual federal spending per county is $214,994.

Part C

```
ggplot(data = countyData) + geom_point(mapping = aes(x = bachelors, y = poverty)) + xlim(0, 80) + ylim(
```



The above plot shows that in general, counties where a larger portion of residents have bachelor degrees have lower poverty rates and counties with high poverty rates have a smaller portion of residents with bachelor degrees. Though it shows that poverty rates and bachelor degree rates are correlated, it isn't possible to tell whether they have any direct effect on each other (though it might be suspected).

Part D

```
over30 <- countyData %>% filter(age_under_18 > 30)
over30 <- nrow(over30)
percOver30 <- 100  * (over30 / nrow(countyData))
```

3.62% of counties in the US have a population with over 30% of people under the age of 18.

Question 3

Part A

New Mean House Price $(\bar{x}) = \frac{(24 \cdot 500000) + 700000}{25} = \$508,000$

Part B

New Standard Deviation (σ)

$(n-1)\sigma^2 = \sum_{i=1}^{n-1} (x_i - \bar{x})^2$

$$(n-1)\sigma_n^2 - (n-2)\sigma_{n-1}^2 = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 - \sum_{i=1}^{n-1}(x_i - \bar{x}_{n-1)})^2$$

$$= (x_n - \bar{x}_n)^2 + \sum_{i=1}^{n-1}[(x_i - \bar{x}_n)^2 - (x_i - \bar{x}_{n-1)})^2]$$

$$= (x_n - \bar{x}_n)^2 + \sum_{i=1}^{n-1}[x_i^2 + 2(x_i)(\bar{x}_n) + \bar{x}_n^2 + x_i^2 + 2(x_i)(\bar{x}_{n-1}) + \bar{x}_{n-1}^2]$$

$$= (x_n - \bar{x}_n)^2 + \sum_{i=1}^{n-1}[x_i^2 + 2(x_i)(\bar{x}_n) + \bar{x}_n^2 + x_i^2 + 2(x_i)(\bar{x}_{n-1}) + \bar{x}_{n-1}^2]$$

$$s_{n-1}^2 = 30000^2 = 900000000$$

$$s_{n-1}^2 = \frac{1}{n-2}\sum_{i=1}^{n-1}$$

Part C

The median will not decrease, but it is not possible to tell whether it will remain the same or increase with the information we have. When the new house is built, its price is above the new median so it is impossible for the median to shift down. If the values in the middle of the ordered set are the same, adding a single value to the upper half would have no effect. Since the data set initially had an even number of values, the median was originally an average of the two middle values. If there is a difference between the two, adding a value above the median would shift the median to the upper value.

Question 4

Part A

Condition: $y_i = (x_i + c) \cdot k$

Calculating the new mean:

$$\bar{y} = \frac{1}{n}\Big[((x_i + c) \cdot k) + ...((x_n + c) \cdot k)\Big]$$

$$\bar{y} = c + \frac{1}{n}\Big[(x_i \cdot k) + ...(x_n \cdot k)\Big]$$

$$\bar{y} = k\Big[c + \frac{1}{n}\Big[(x_i \cdot k) + ...(x_n \cdot k)\Big]\Big]$$

$$\bar{y} = k\Big[c + \bar{x}\Big]$$

Calculating the new variance:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(((x_i + c) \cdot k) - k(c + \bar{x}))^2$$

Part B