# Stat 104: Quantitative Methods
# Homework 1 SOLUTIONS

1) For the following surveys, discuss any problems you think exist and suggest how to fix the issues.

   **Answers may vary slightly.**

   a. A retail store manager wants to conduct a study regarding the shopping habits of his customers. He selects the first 60 customers who enter his store on a Saturday morning.

      **Answer: This is a particularly bad sample because it narrowly selects customers that come in early on a Saturday morning, whose shopping habits may be very different from his average customer. This represents a selection bias stemming from the use of a convenience sample. I would recommend that he surveys as many customers as possible, at all hours of operations, for an entire month.**

   b. The village of Oak Lawn wishes to conduct a study regarding the income level of households within the village. The village manager selects 10 homes in the southwest corner of the village and sends an interviewer to the homes to determine the household income.

      **Answer: This is another example of selection bias. The survey participants are isolated geographically and will likely not represent a true cross section of the village. I would make the recommendation to randomize selection of the households and increase the sample size.**

   c. An antigun advocate wants to estimate the percentage of people who favor stricter gun laws. He conducts a nationwide survey of 1,203 randomly selected adults 18 years old or older. The interviewer asks the respondents, "Do you favor harsher penalties for individuals who sell guns illegally?"

      **Answer: This question is clearly biased and leading. It's actually somewhat difficult for a pro-gun individual to answer this question accurately because of the way it is framed. It does not provide an actual indication of persons that favor stricter gun law. Instead, I would simply ask "Do you think our current laws restricting the use of guns should be made more strict, less strict, or stay the same?"**

2) A bank with branches in a large metropolitan area is considering opening its offices on Saturday, but it is uncertain whether customers will prefer (1) having walk-in hours on Saturday or (2) having extended branch hours during the week. Listed below are some of the ideas for gathering data. For each, indicate what (if any) biases (problems) might result.

   **Answers may vary slightly.**

   a. Put a big ad in the newspaper asking people to log their opinions on the bank's Web site.
   **Answer: This is an example of voluntary response bias since individuals are choosing whether or not be in the sample. It is likely that those who choose to respond to the newspaper ad are different than the individuals who do not respond.**

   b. Randomly select one of the branches and contact every customer at that bank by phone.
   **Answer: This selection process will result in a convenience bias since only one branch is being selected. It is likely that branches differ and this one specific branch will not be representative of all the branches of this bank.**

   c. Send a survey to every customer's home, and ask the customers to fill it out and return it.
   **Answer: This will result in a voluntary response bias since most of the bank customers will not respond to the survey and those who do respond are likely different than those who do not respond.**

   d. Randomly select 20 customers from each branch. Send each a survey, and follow up with a phone call if he or she does not return the survey within a week.
   **Answer: This study design is the best of the four and will result in minimal bias compared to the other designs. If some of the branches have a significantly greater number of customers, we might want to select more customers from those branches.**

3) Suppose you are back in high school and the campaign manager for your friend who is running for senior class president. You would like to know what proportion of students would vote for her if the election was held today. The class is too big to ask everyone (314 students). Comment on whether or not each of the following sampling procedures should be used. Explain why or why not.

   a. Poll everyone in your friend's math class.
   **There could be selection bias because your friend's math class may not be representative of the entire class. Better to randomly select members, because friend's math class could be an advanced class.**

b. Assign every student in the senior class a number from 1 to 314. Then, use a random number generator to select 30 students to poll.

**This is the best design, because it is a random experiment that controls for all the other biases that we have seen**

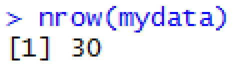
c. Ask every student who is going through the lunch line in the cafeteria who they will vote for.

**This is in part convenience bias, but also exhibits selection bias, because people in the lunch line will be of a certain socioeconomic standing (in order to afford lunch), and perhaps will represent similar groups of people (friends often clump together in the lunch line).**

4) R Practice, Part 1. In R, read in the results of a small survey done by visitors to a regional mall. This is done with the following command in the R command window

mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/smallsurvey.csv")
You can see the data with the command `View(mydata)`

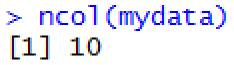
a. How many rows of data are in this data set? (the nrow(mydata) command could be useful here but remember the first row has the variables names).



```
> nrow(mydata)
[1] 30
```

**Answer: 30**

b. How variables are in this data set? (the ncol(mydata)command could be useful here).



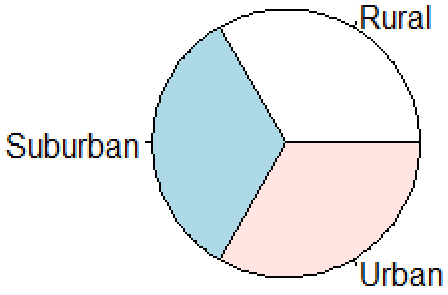
```
> ncol(mydata)
[1] 10
```

**Answer: 10**

c. How many categorical variables are in this data set?

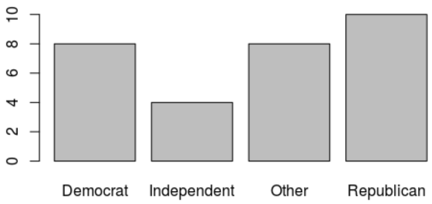
**There are 3 categorical variables in this data set: gender, residence, political party.**

d. One way to examine categorical variables is with a pie chart. Produce a pie chart of where people live (the residence variable) by using the following command. Comment on the graph: pie(table(mydata$residence))



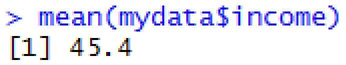
**The residence variable seems to be evenly split between suburban, rural, and urban categories.**

e. Another way to examine categorical variables is with a bar chart. Produce a bar chart of political affiliation (the politicalparty variable) by using the following command. Comment on the graph-why can't we use a histogram for this variable? barplot(table(mydata$politicalparty))



**There are 10 Republicans, 8 Democrats and Other, and 4 Independents. We cannot use a histogram for this variable because histograms are characterized by columns that represent a group defined by a quantitative variable (like a range or a single value). With categorical variables, there are no quantity to plot, so we use a bar chart.**

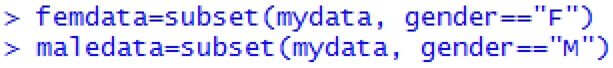
f. Find the average of the income variable.



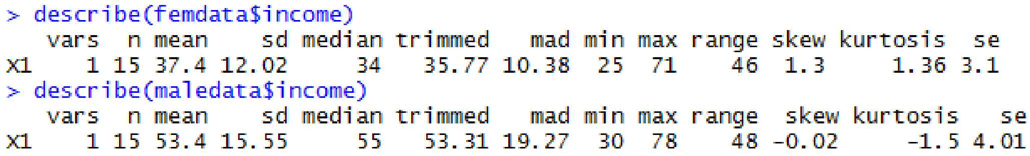
```
> mean(mydata$income)
[1] 45.4
```

**Answer: 45.4**

g. We can subset data in different ways. We could create a new data set just for all the females respondents by creating femdata=subset(mydata,gender=="F"). As another example, one could create a new data set for those people that have income over 50 with the command newdata=subset(mydata,income>50).
Compare the average income and standard deviation of income for men and women.



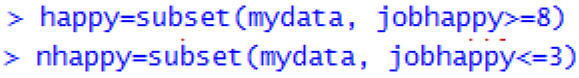
```
> femdata=subset(mydata, gender=="F")
> maledata=subset(mydata, gender=="M")
```



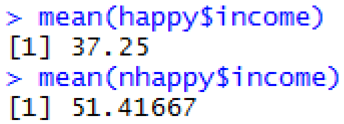
```
> describe(femdata$income)
   vars  n mean    sd median trimmed   mad min max range skew kurtosis  se
X1    1 15 37.4 12.02     34   35.77 10.38  25  71    46  1.3     1.36 3.1
> describe(maledata$income)
   vars  n mean    sd median trimmed   mad min max range  skew kurtosis   se
X1    1 15 53.4 15.55     55   53.31 19.27  30  78    48 -0.02     -1.5 4.01
```

**Answer:**
**Female average income: 37.4**
**Male average income: 53.4**
**Female std. dev:12.0226**
**Male std. dev: 15.5462**

h. The variable jobhappy measures on a 1-10 scale how happy someone is with their job. Compare the average income for someone with a jobhappy rating of 8 or more versus the average income of someone with a jobhappy rating of 3 or less. What do you find?



```
> happy=subset(mydata, jobhappy>=8)
> nhappy=subset(mydata, jobhappy<=3)
```



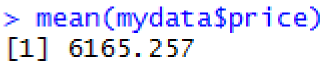
```
> mean(happy$income)
[1] 37.25
> mean(nhappy$income)
[1] 51.41667
```

**Answer: Average income for jobhappy 8 or more is 37.25. Average income for jobhappy 3 or less is 51.42. Those who have lower jobhappy score have higher incomes, it seems.**

5) R Practice, part 2. This question uses an old data set on cars from Consumer Reports. To load the data into R, enter the following command in R's command line:

`mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/cars10.csv")`
To see what is in this data set, you can enter the R command `View(mydata)`

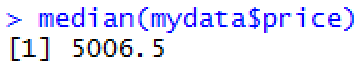
a. Calculate the mean price of automobiles in the data set.



```
> mean(mydata$price)
[1] 6165.257
```

**The mean price of automobiles in this data set is $6165.26.**

b. Calculate the median price of automobiles in the data set.



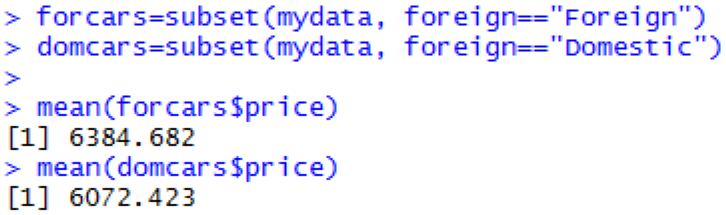
```
> median(mydata$price)
[1] 5006.5
```

**The median price of automobiles in this data set is $5006.50.**

c. What does the difference between the mean and median price indicate about the shape of the distribution for the price?

**Since the mean> median, the distribution of price is right-skewed.**

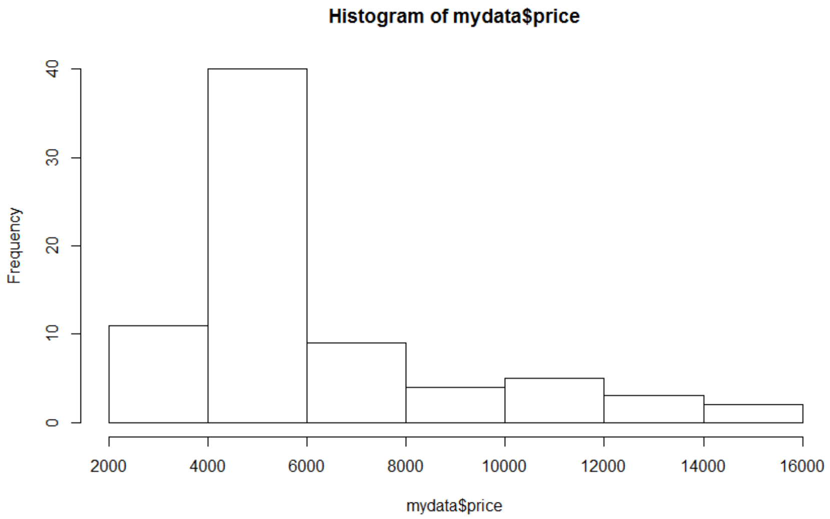
d. Calculate the mean price of automobiles separately for domestic and foreign cars and compared the results.



```
> forcars=subset(mydata, foreign=="Foreign")
> domcars=subset(mydata, foreign=="Domestic")
>
> mean(forcars$price)
[1] 6384.682
> mean(domcars$price)
[1] 6072.423
```

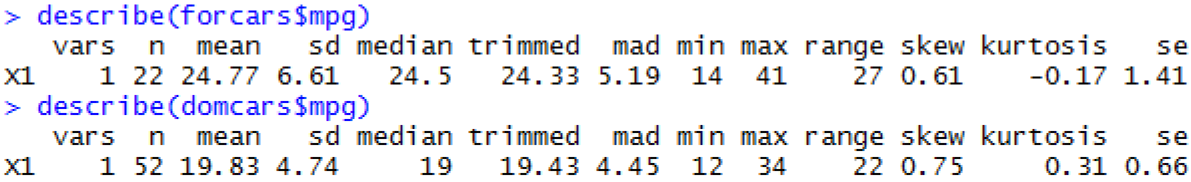
**The mean price of foreign cars ($6384.68) is greater than the mean price of domestic cars ($6072.42).**

e. Make a histogram of the price of cars. What shape does the histogram take? (Is it symmetric? Skewed?)

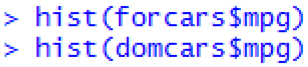


**The histogram is right-skewed.**

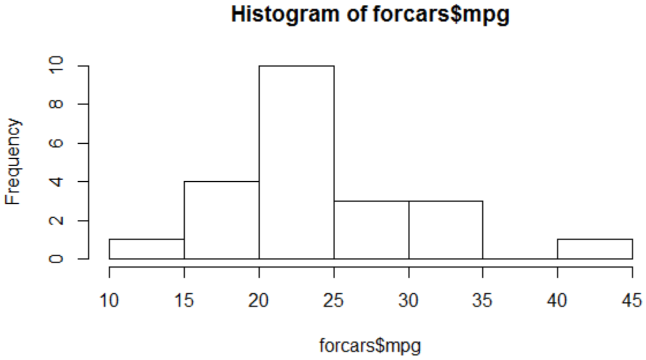
f. Discuss the difference in distributions for mpg for foreign and domestic cars. [do this by comparing means, medians, and histograms.]

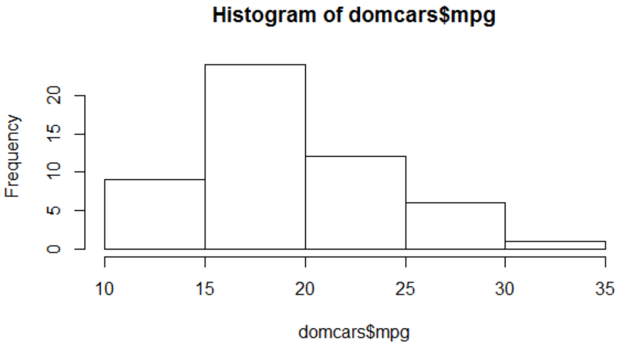


```
> describe(forcars$mpg)
   vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 22 24.77 6.61   24.5   24.33 5.19  14  41    27 0.61    -0.17 1.41
> describe(domcars$mpg)
   vars  n  mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 52 19.83 4.74     19   19.43 4.45  12  34    22 0.75     0.31 0.66
```



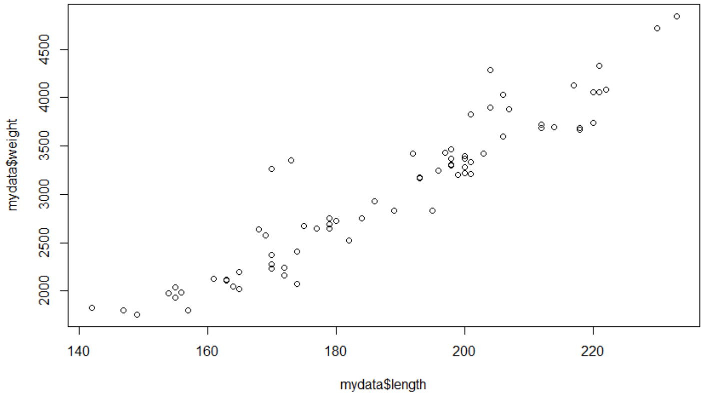
```
> hist(forcars$mpg)
> hist(domcars$mpg)
```





**The mean mpg for foreign cars is 24.77 and the median is 24.5. Domestic cars have both a lower mean (19.83) and median (19) mpg. Comparing the histogram, we observe that foreign cars have greater mpg and an overall wider range compared to domestic cars.**

g. Make a scatterplot of the variables weight and length. Does there appear to be any association between the variables?



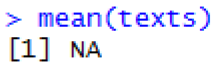
**Yes, there does appear to be an association. As car length increases, the weight of the car also increases.**

6) R practice, part 3. For this question we will use the following data set.

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/stat111survey.csv")
```

```
Create the following variable (which is number of texts students send per day)
texts = mydata$texts
```

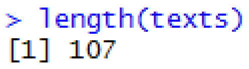
a. Using the `mean` command, find the mean number of texts. Uh oh, you should get a weird response- what is it?



```
> mean(texts)
[1] NA
```

**Answer: returning NA error (missing data)**

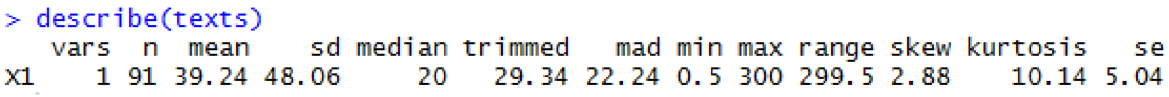
b. Use the command `length(texts)` to find how many data points are in the variable texts.



```
> length(texts)
[1] 107
```

**Answer: 107**

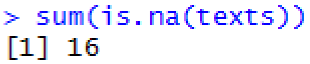
c. Use the command `describe(texts)` to get the summary statistics. How does the n from this output compare to what you found in (b).



```
> describe(texts)
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 91 39.24 48.06     20   29.34 22.24 0.5 300 299.5 2.88    10.14 5.04
```

**Answer: From the describe command, we see that n=91. This is less than the number reported in b (n=107).**

d. Do the command `sum(is.na(texts))` which counts the number of values that are missing. How many values are missing? Does this agree with (b) and (c)?



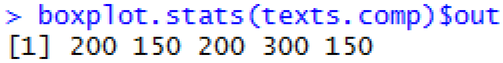
```
> sum(is.na(texts))
[1] 16
```

**Answer: There are 16 missing values. This agrees with b and c, since 16+91=107.**

e. Create a new variable text.comp = texts[complete.cases(texts)]. This removes all the missing data.

**No answer here- just run the command in R.**

f. Using the boxplot outlier rule, how many outliers does the data texts.comp have?



```
> boxplot.stats(texts.comp)$out
[1] 200 150 200 300 150
```

**Answer: 5 outliers**

7) Unfortunately, a friend of yours has been diagnosed with cancer. You obtain a histogram of the survival time (in months) of patients diagnosed with this form of cancer as shown in the figure below. The median survival time for individuals with this form of cancer is 11 months, while the mean survival time is 69 months. What words of encouragement should you share with your friend from a statistical point of view? [It is also recommended you read the essay "the median isn't the message" found on the course web site.]

**Answer: To encourage them, I would say that about half of people diagnosed with this type of cancer live for many years after their diagnosis, and that although the average is around 5 years, that many live much longer than that. The mean is heavily skewed and must not be relied upon in this case for any indication of what might happen. If my friend can get through the next few years, then their prospects looks pretty good for many years after.**

8) When my friend Seth transferred from Harvard to Yale, many of his friends remarked that the average student IQ increased at both places. Is this possible and if so, how? Briefly explain.

**Answer: It is certainly possible. If Seth's IQ is lower than the mean at Harvard, but higher than the mean at Yale, then the mean IQ at both schools would increase.**

**For example, assume that Seth had an IQ of 100, and the following is true:**

|  | Mean IQ | *N* |
|---|---|---|
| **Harvard** | **130** | **100** |
| **Yale** | **70** | **100** |

**If Seth transfer's his below-Harvard-average IQ to Yale, then the following will happen:**

**Harvard's cum. IQ scores - Seth's IQ = 13000 - 100 = 12900**

**Harvard's new cum. IQ scores / Harvard's new population = 12900/99 = 130.30**

**Yale's cum. IQ score + Seth's IQ = 7000+100 = 7100**

**Yale's new cum. IQ scores / Yale's new population = 7100/101 = 70.2970**

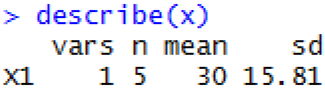
**Both school's average IQ increased by about 0.3 points when Seth transferred.**

9) Suppose the diameters of a sample of new tires coming off one production line turned out to have a standard deviation of 0. Would the manufacturer be happy or unhappy, assuming the average diameter was correct? Explain.

**Answer: The manufacturer would be very happy. The only way that the standard deviation could be 0 would be if all the tires had exactly the same diameter, which is the consistency of product the manufacturers would want.**

10) Use this data set for the following questions {10,20,30,40,50}. Feel free to use R for this problem. You can define this data set in R with the command `x=c(10,20,30,40,50).`

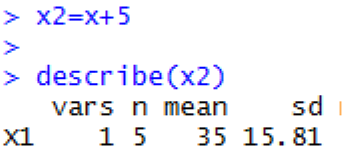
a. Find the standard deviation and the mean.



```
> describe(x)
   vars n mean    sd
X1    1 5   30 15.81
```

**Answer: The mean is 30 and the standard deviation is 15.81.**

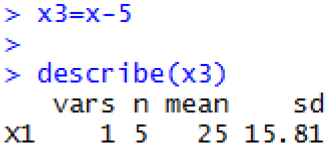
b. Add 5 to each value, and then find the standard deviation and mean.



```
> x2=x+5
>
> describe(x2)
   vars n mean    sd
X1    1 5   35 15.81
```

**Answer: The mean is 35 and the standard deviation is 15.81.**

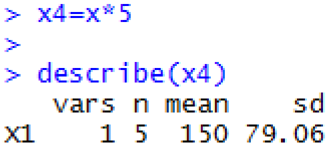
c. Subtract 5 from each value and find the standard deviation and mean.



```
> x3=x-5
>
> describe(x3)
   vars n mean    sd
X1    1 5   25 15.81
```

**Answer: The mean is 25 and the standard deviation is 15.81.**

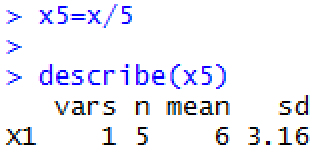
d. Multiple each value by 5 and find the standard deviation and mean.



```
> x4=x*5
>
> describe(x4)
   vars n mean    sd
X1    1 5  150 79.06
```

**Answer: The mean is 150 and the standard deviation is 79.06.**

e. Divide each value by 5 and find the standard deviation and mean.



```
> x5=x/5
>
> describe(x5)
   vars n mean   sd
X1    1 5    6 3.16
```

**Answer: The mean is 6 and the standard deviation is 3.16.**

f. Generalize the results of parts b through e.

**Answer: The general rules are: Var(a+bX)=$b^2$Var(X) SD(a+bX)=b*SD(X) Mean(a+bX)=a+b*[Mean(X)]**

11) A company has 30 employees, including a director. The lowest salary among the 30 employees is $22,000. The director's salary is $180,000, which is more than twice as much as anyone else's salary. Decided for each of the following statements about the 30 salaries, whether it is true, false, or you cannot tell *on the basis of the information at hand*.

a. The average salary is below $60,000.

**Answer: Can't tell. We do not know how the salaries are spread out.**

b. The median salary is below $60,000.

**Answer: Can't tell. We do not know how the salaries are spread out.**

c. If all salaries are increased by $1,000, that adds $1,000 to the average.

**Answer: True. Finding the average is adding everything up and then dividing by the # of observations. By adding 1,000 to each observation, the mean will increase by 1,000.**

d. If the director's salary is doubled, and all other salaries remain the same, that increases the average salary.

**Answer: True. By increasing the value of one of the observations, the average will also increase.**

e. If the director’s salary is doubled, and all other salaries remain the same, that increases the median salary.

**Answer: False. By doubling the director’s salary, that does not increase the middle value.**

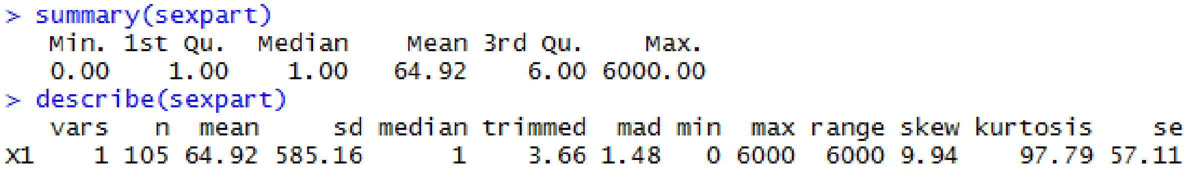
f. The standard deviation of the salaries is larger than $180,000.

**Answer: False. The standard deviation cannot be larger than 180,000.**

12) In this problem we will look at the sexual partner data set mentioned in class. Load it into R using the command

```
mydata=read.csv(“http://people.fas.harvard.edu/~mparzen/stat100/sexpart.csv”)
sexpart=mydata$x
```

a. Compare the standard deviation and IQR as measures of spread on the full dataset. Which measure do you think is more appropriate to describe the spread in the data set?



```
> summary(sexpart)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    1.00    1.00   64.92    6.00 6000.00 
> describe(sexpart)
   vars   n  mean     sd median trimmed  mad min  max range skew kurtosis    se
X1    1 105 64.92 585.16      1    3.66 1.48   0 6000  6000 9.94    97.79 57.11
```

**IQR= 6-1 = 5**

**Std dev=585.16**

**The better measure here is IQR, because SD is heavily affected by outliers.**

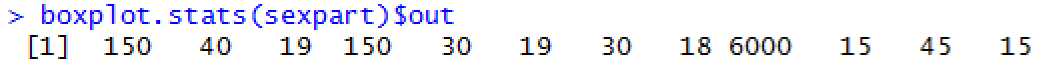
b. Compare which points are flagged as outliers using the two methods discussed in class (Z score and boxplot method).

**Z-score**: Any data point more than 2 standard deviations away from the mean is considered an outlier: 64.92 +/-(2*585.16)= any data point below-1105 or above 1235. According to this method, **the 6,000 point is an outlier.**

**Boxplot method:** Any point < Q1-1.5(Q3-Q1) or any point > Q3+1.5(Q3-Q1) would be considered an outlier. **12 outliers identified** (see below)

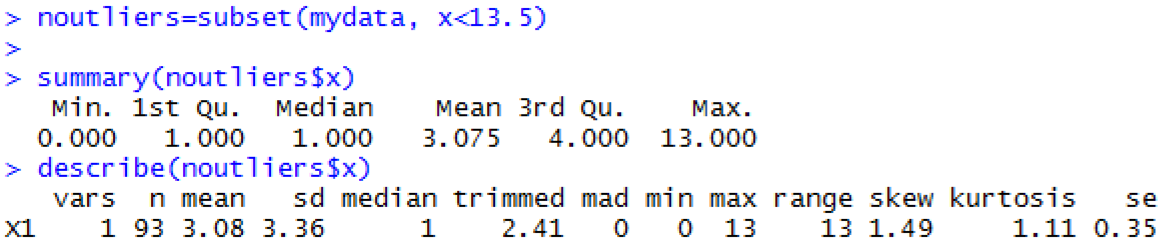
Q1-1.5(Q3-Q1) = 1-1.5(5)= -6.5

Q3+1.5(Q3-Q1)= 6-1.5(5)=13.5



```
> boxplot.stats(sexpart)$out
 [1]  150   40   19  150   30   19   30   18 6000   15   45   15
```

c. Remove the outliers flagged using the boxplot methods. Recalculate the IQR and standard deviation of this smaller dataset. Are the values closer to each other now?



```
> noutliers=subset(mydata, x<13.5)
>
> summary(noutliers$x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.000   3.075   4.000  13.000
> describe(noutliers$x)
   vars  n mean   sd median trimmed mad min max range skew kurtosis   se
X1    1 93 3.08 3.36      1    2.41   0   0  13    13 1.49     1.11 0.35
```

**New SD is 3.36 and new IQR is 3. These values are now much closer to each other.**

13) A mutual fund has a mean rate of return of about 12.3%, with a standard deviation of 15.7%.

a. According to Chebyshev's Inequality, at least 75% of returns will be between what values?

$$1-\frac{1}{k^2}=0.75, k=2$$

$$12.3\ \pm 2(15.7)=(\mathbf{-19.1}\%, \mathbf{43.7}\%)$$

b. According to Chebyshev's Inequality, as least 88.9% of returns will be between what two values?

$$1-\frac{1}{k^2}=0.889, k=3$$

$$12.3\ \pm 3(15.7)=(\mathbf{-34.8}\%, \mathbf{59.4}\%)$$

c. Should an investor be surprised if she has a negative rate of return? Why?

**The investor should not be surprised with a negative return, as it is possible to obtain negative returns. One SD away from the mean would land her in negative values.**

d. If we were going to use the Empirical Rule, what would we need to assume about the returns?

**We would need to assume that returns are symmetric.**

14) Suppose $x_1$=2, $x_2$=-1, $x_3$=0. Find $2+\sum_{i=1}^{3} 5x_i$ and $1/\sum_{i=1}^{3} x_i^2$.

$$2+\sum_{i=1}^{3} 5x_i = 2+[(5*2)+(5*-1)+(5*0)] = \mathbf{7}$$

$$\frac{1}{\sum_{i=1}^{3} x_i^2} = \frac{1}{[2^2+-1^2+0^2]} = \frac{1}{[4+1+0]} = \frac{\mathbf{1}}{\mathbf{5}}$$

15) Suppose $\bar{x} = 11$ and define $y_i = 2x_i - 5$. Find the numerical value of $\bar{y}$.

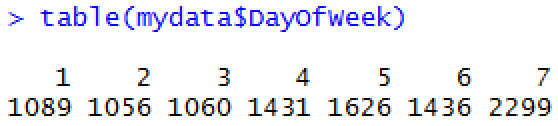
**To find the mean of Y, we are looking at transformations of the mean of x. Because the original mean is 11, the new mean will be 11*2-5=17.**

16) We have a data set that explores airline on time performance of domestic flights operated by large air carriers. The information was compiled from the Bureau of Transportation Statistics. We will only by analyzing data from randomly selected flights from November 2008 which is in the data set airlin2008NovS.csv. The variable names and definitions are listed in another file on the course web site.

You can read the data set into R as follows

```
mydata=read.csv(http://people.fas.harvard.edu/~mparzen/stat104/Airline2008NovS.csv)
```

a. Which day of the week has the most flights? Use the following R command to help answer the questions: `table(mydata$DayOfWeek)`

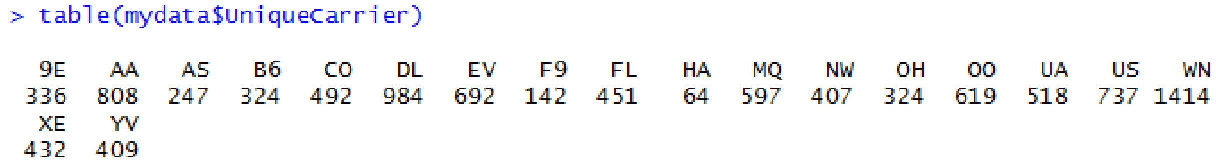


```
> table(mydata$DayOfWeek)

   1    2    3    4    5    6    7 
1089 1056 1060 1431 1626 1436 2299 
```

**Answer: The 7th day, which is Sunday, has the most flights.**

b. How many unique carriers are in this data set?

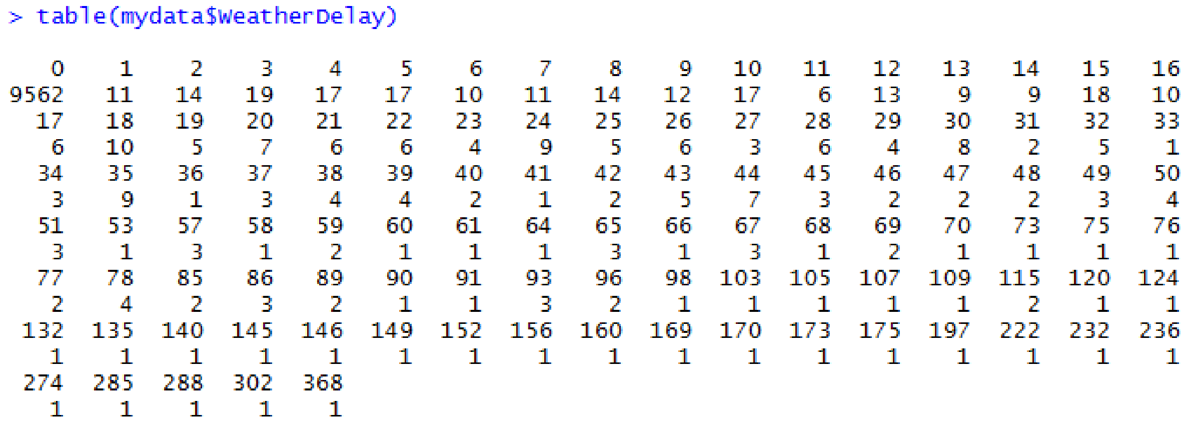


```
> table(mydata$UniqueCarrier)

  9E   AA   AS   B6   CO   DL   EV   F9   FL   HA   MQ   NW   OH   OO   UA   US   WN 
 336  808  247  324  492  984  692  142  451   64  597  407  324  619  518  737 1414 
  XE   YV 
 432  409 
```

**Answer: There are 19 unique carriers in this data set.**

c. How many flights in this data set had a zero minute weather delay?

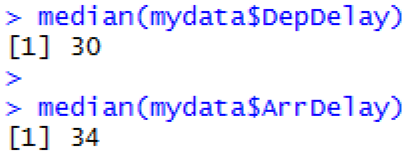


```
> table(mydata$WeatherDelay)

   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16 
9562   11   14   19   17   17   10   11   14   12   17    6   13    9    9   18   10 
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33 
   6   10    5    7    6    6    4    9    5    6    3    6    4    8    2    5    1 
  34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50 
   3    9    1    3    4    4    2    1    2    5    7    3    2    2    2    3    4 
  51   53   57   58   59   60   61   64   65   66   67   68   69   70   73   75   76 
   3    1    3    1    2    1    1    1    3    1    3    1    2    1    1    1    1 
  77   78   85   86   89   90   91   93   96   98  103  105  107  109  115  120  124 
   2    4    2    3    2    1    1    3    2    1    1    1    1    1    2    1    1 
 132  135  140  145  146  149  152  156  160  169  170  173  175  197  222  232  236 
   1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1 
 274  285  288  302  368 
   1    1    1    1    1 
```

**Answer: 9562 flights in this data set had a zero minute weather delay.**

d. Which is larger, the median departure delay or the median arrival delay?



```
> median(mydata$DepDelay)
[1] 30
>
> median(mydata$ArrDelay)
[1] 34
```

**Answer: The median arrival delay is larger.**