Homework 1

Erik Lee

Mon Jul 2 13:24:31 2018

###Homework policy: This homework is due by 5pm (EST) on the due date. Homework is to be handed in via the course website in pdf format. Although we prefer you use Rmarkdown or Word, you do not need to type the homework; there are many ways (scanner in the library or phone apps) to convert written homework into a pdf file. Ask the teaching staff if you need assistance.

###Late homework will not be accepted. You are encouraged to discuss homework problems with other students (and with the instructor and TFs, of course), but you must write your final answer in your own words. Solutions prepared "in committee" or by copying someone else's paper are not acceptable.

###Please keep your computer output to a minimum and focus on the required answer. The easiest way to put your computer output into your homework is to cut and paste it into a Word file and use the font "courier new".

Problem 1

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

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.

The issues with the 60 selected customers are the time dependency of Saturday morning and biased order of the selected set of 60. The time dependency limits the type of customer being observed. This does not consider customers who shop during the afternoon and evening, which can be argued as busier hours, or those who shop on weekdays, which may exhibit different habits. To correct this, the manager should select customers throughout a given day, across the week, or in a given interval (few days, weeks, or months) to get a better sense of the habits of a standard customer. Selecting the first 60 also creates a bias as customers who come in sequence may be part of groups, which purchase items together. It would be better to select individuals randomly to provide a better set of data to study spread and commonalities. Also it would be helpful specify the type of customer studied; children may still be considered a customer, but since they are unlikely to make their own purchase, this will result in null data points.

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 household income.

Selecting 10 households in southwest Oak Lawn creates a poor sample for representing the incomes across Oak Lawn. Because the households come from a specific region, the prices of the homes and the incomes from those families are likely to be similar. This creates a homogenous group limiting the range and variation of incomes and biasing the results from the rest of the population of Oak Lawn. This can cause the observed incomes to be lower or higher than what is true about the rest of Oak Lawn's incomes. To fix the limited subset, the survey should select houses, evenly spread across town, in a random fashion to get a better sense of the income spread of Oak Lawn. This will take into account wealthier and poorer regions/neighborhoods and give a better sense whether the incomes of Oak Lawn tend toward higher levels or lower ranges of income, based on the clustering of the data.

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 and older. The interviewer asks the respondents, "Do you favor harsher penalties for individuals who sell guns illegally?"

The framing of the question asked by the interview can cause bias. Choosing "harsher" to describe severity of the penalty cause a person to think emotionally rather than objectively. As such, the person being asked may immediately jump to a conclusion without carefully considering the question, either by immediately agreeing for harsher penalties, or sympathizing with a criminal by disagreeing. The sentence is also structured so that the penalties are mentioned before the "individuals selling guns illegally", giving emphasis to the penalties rather than the criminals. The question can be reworded as "Should individuals, who sell guns illegally, receive stricter/stronger penalties?" This gives guilty persons fair consideration for or against, and using a neutral word like stricter or stronger does not sway opinion to on side or the other.

Problem 2

Suppose you are back in high school and are 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.

This procedure of sampling should not be used. Selecting her friends from math class provides a biased group of participants, which does not represent her class of 314 students. This group will likely vote for her either unanimously or close to full, bar some friends who do not like her. This result provides a false assumption that she would win in a landslide, when in reality her percentage of votes would be lower.

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 a good procedure for polling the class. It employs equally random selection among the entire class to create a realistic sample of the students. Randomly sampling removes any bias and provides a result that would look similar to the actual election. There is an equal chance a given student will vote for her or another candidate, and the percentages will reflect the candidates approval.

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

This is a decent way for polling the students in class. By asking every student during lunch time, this would allow the entire class to be polled and give an accurate prediction. Since students enter lunch in a random fashion, the percentage polled will reflect the percentage in the actual election. There are a few instances that could create bias such as if students enter lunch in groups, or a student is able to hear the response given from the student before him or her. However, these can be controlled by separating a section of the line so that students are asked individually and are not aware of other students' responses. The main issue is that asking each student takes a long time, but can be expedited if multiple people, from the candidate's campaign, are surveying at the same time.

Problem 3

In R, read in the results of a small survey done by visitors to a regional mall. This can be done as follows. We also show you below how to obtain some information about the data set.

```r
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/smallsurvey.csv")
# number of rows
nrow(mydata)
```

```
## [1] 30
```

```r
# number of columns
ncol(mydata)
```

```
## [1] 10
```

```r
# names of the variables
names(mydata)
```

```
##  [1] "id"              "gender"          "residence"       "politicalparty"
##  [5] "numbchildren"    "age"             "income"          "jobhappy"
##  [9] "tvhours"         "radiohours"
```

```r
# for example, mean of the income variable
mean(mydata$income)
```

```
## [1] 45.4
```

a) How many rows of data are in this data set?

```r
nrow(mydata)
```

```
## [1] 30
```

There are 30 rows in the dataset.

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

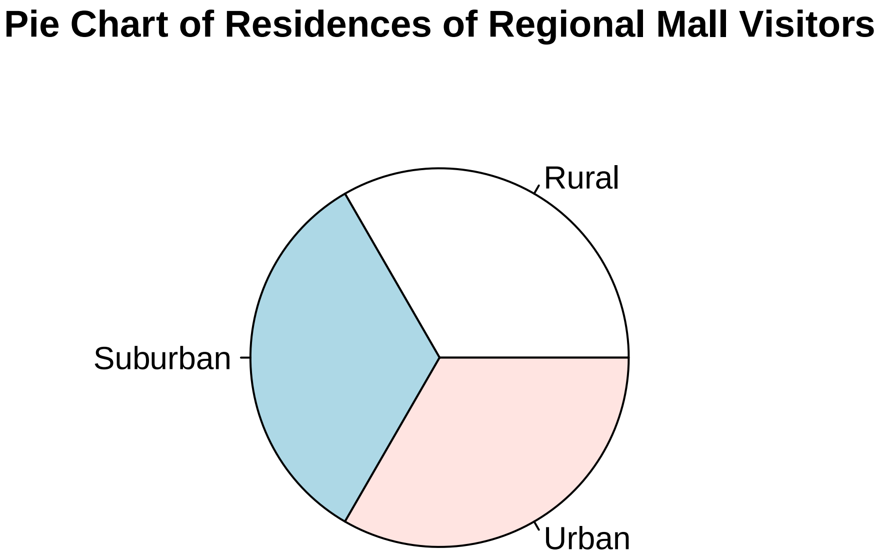
```r
ncol(mydata)
```

```
## [1] 10
```

There are 10 variables/columns in the data set.

c) 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 `pie` command. Comment on the graph.

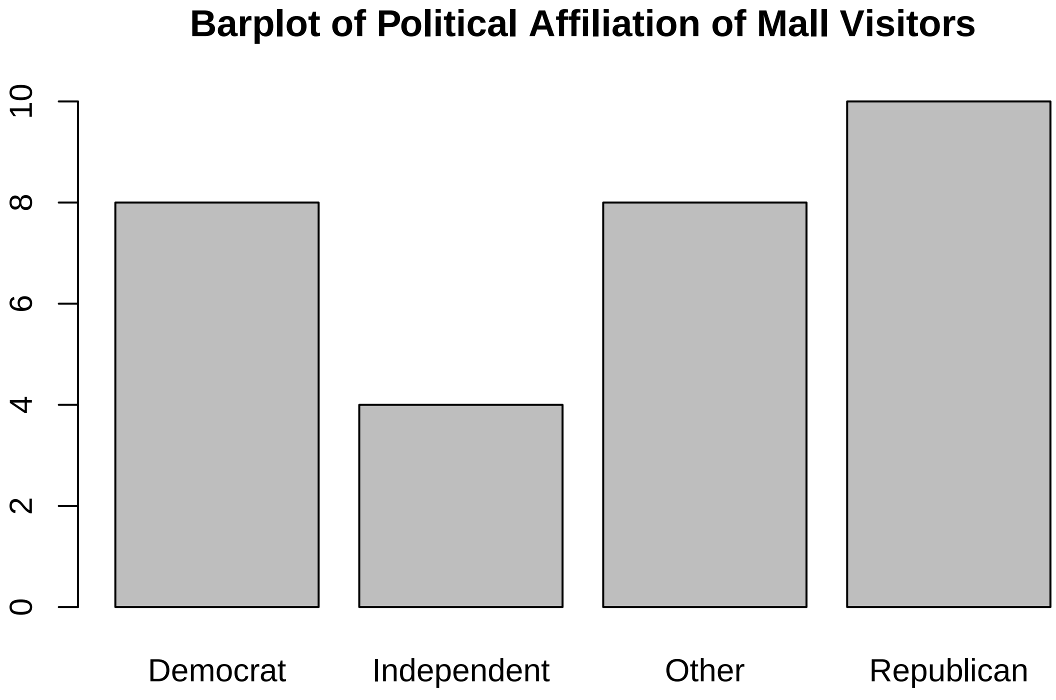
```r
# Creates a summary of the number of different residences Rural, Suburb, Urban
sum = summary(mydata$residence)
# Create labels of each type of Residence for the pie chart
lbls = levels(mydata$residence)
# Construct pie chart with summary, label, and main title
pie(x=sum, labels = lbls, main="Pie Chart of Residences of Regional Mall Visitors")
```



The Pie Chart of Residences has an equal proportion of each type of Residence. From the summary, there are 10 of each type of residence (Rural, Suburb, Urban), labeled respectively. This results in an equally divided pie chart with a third of the pie for each Residence type

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

```
# Get summary of each type of political party from my data
polsum = summary(mydata$politicalparty)
# Create bar plot using the summary of political party, add in title to describe Political Affiliation
barplot(polsum, main="Barplot of Political Affiliation of Mall Visitors")
```



Histograms only accept numberical values to x. For political affiliation, the categories are represented by strings for each party (Democrat, Independent, Other, Republican).

e) Find the average of the income variable.

```r
mean(mydata$income)
```

```
## [1] 45.4
```

Average income is $45.4 (assumed to be thousands)

f) We can subset data in different ways (see handout on class site for how to do this). Compare the average income and standard deviation of income for men and women.

```r
males=subset(mydata,mydata$gender=="M")
females=subset(mydata,mydata$gender=="F")
# Average income $ (mean) for males and females
mean(males$income) # male avg. income
```

```
## [1] 53.4
```

```r
mean(females$income) # female avg. income
```

```
## [1] 37.4
```

```r
# Standard of Deviation for incomes of males and females
sd(males$income) # male std. dev. income
```

```
## [1] 15.55
```

```r
sd(females$income) # female std. dev. income
```

```
## [1] 12.02
```

Based on the data, men have a higher average income at $53.4 to the women's average of $37.4. Men also have a higher standard of deviation of avg. income of 15.55 to the income standard of deviation of women of 12.02.

g) 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?

```r
# Creat subsets of each jobhappy group
highjobhappy=subset(mydata,mydata$jobhappy>=8) # 8 or higher jobhappy
lowjobhappy=subset(mydata,mydata$jobhappy<=3) # 3 or lower jobhappy
# Find the average income (mean) for each subset
mean(highjobhappy$income) # 8 or higher
```

```
## [1] 37.25
```

```r
mean(lowjobhappy$income) # 3 or lower
```

```
## [1] 51.42
```

Individuals of the 8 or higher jobhappy subset have a lower average income of $37.25 as compared to the 3 or lower jobhappy subset with an average income of $51.42.

Problem 4

This question uses an old data set on cars from Consumer Reports. To load the data into R enter the following command

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/cars10.csv")
#Always good to know the variable names
names(mydata)
```

```
##  [1] "make"         "price"        "mpg"          "headroom"
##  [5] "trunk"        "weight"       "length"       "turn"
##  [9] "displacement" "gear_ratio"   "foreign"
```

```
#Calculate some means and medians
mean(mydata$price)
```

```
## [1] 6165
```

```
median(mydata$price)
```

```
## [1] 5006
```

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

```
mean(mydata$price)
```

```
## [1] 6165
```

Mean price is $6165

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

```
median(mydata$price)
```

```
## [1] 5006
```

Median price is 5006.

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

Since the mean is greater than the median, the data is scewed to the right and the distribution is asymmetrical. This indicates that data points that are greater than the median (middle point) have greater values/spread, with the possibility of having outliers, compared to the datapoints lower than the median.

d) Calculate the mean price of automobiles separately for the domestic and foreign cars and compare the results. Note that `foreign` is coded "Foreign" for foreign cars and "Domestic" for domestic cars.

```
foreigncars=subset(mydata,mydata$foreign=="Foreign")
domesticcars=subset(mydata,mydata$foreign=="Domestic")
# Means of subsets
mean(foreigncars$price)
```

```
## [1] 6385
```

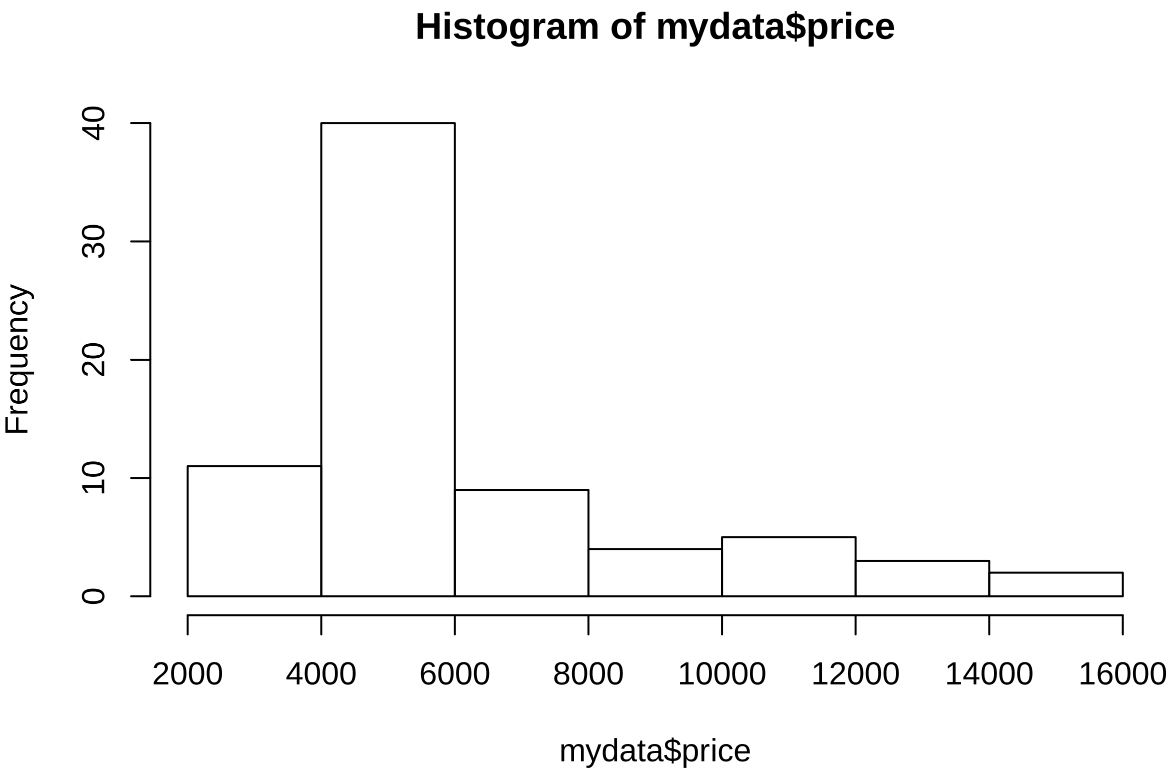
```
mean(domesticcars$price)
```

```
## [1] 6072
```

Foreign cars have a higher mean price of $6385 compared to domestic mean price of $6072.

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

```
hist(mydata$price)
```



The histogram of prices of cars is skewed to the right. The highest frequency for price is at $4000-6000. There are several data points that have low frequency but high price when the price is above $8000.

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

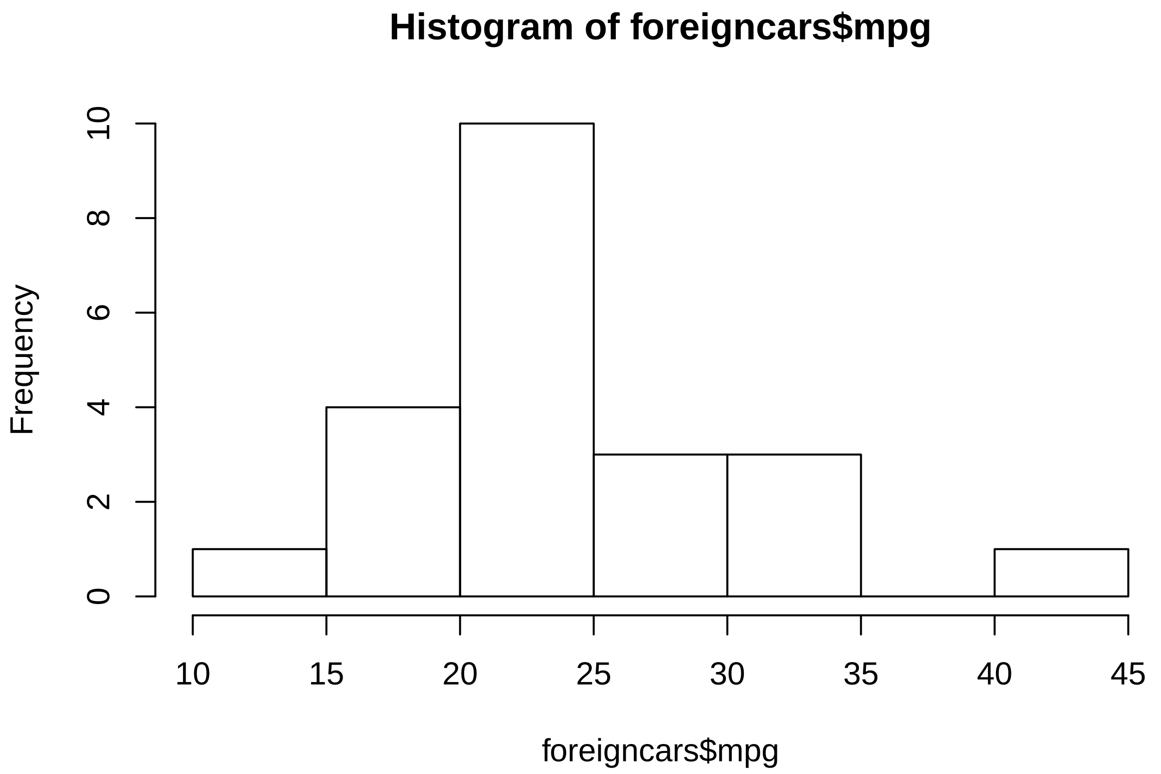
```
# Summaries for foreign and domestic cars mpg for mean and medians
summary(foreigncars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    14.0    21.0    24.5    24.8    27.5    41.0
```

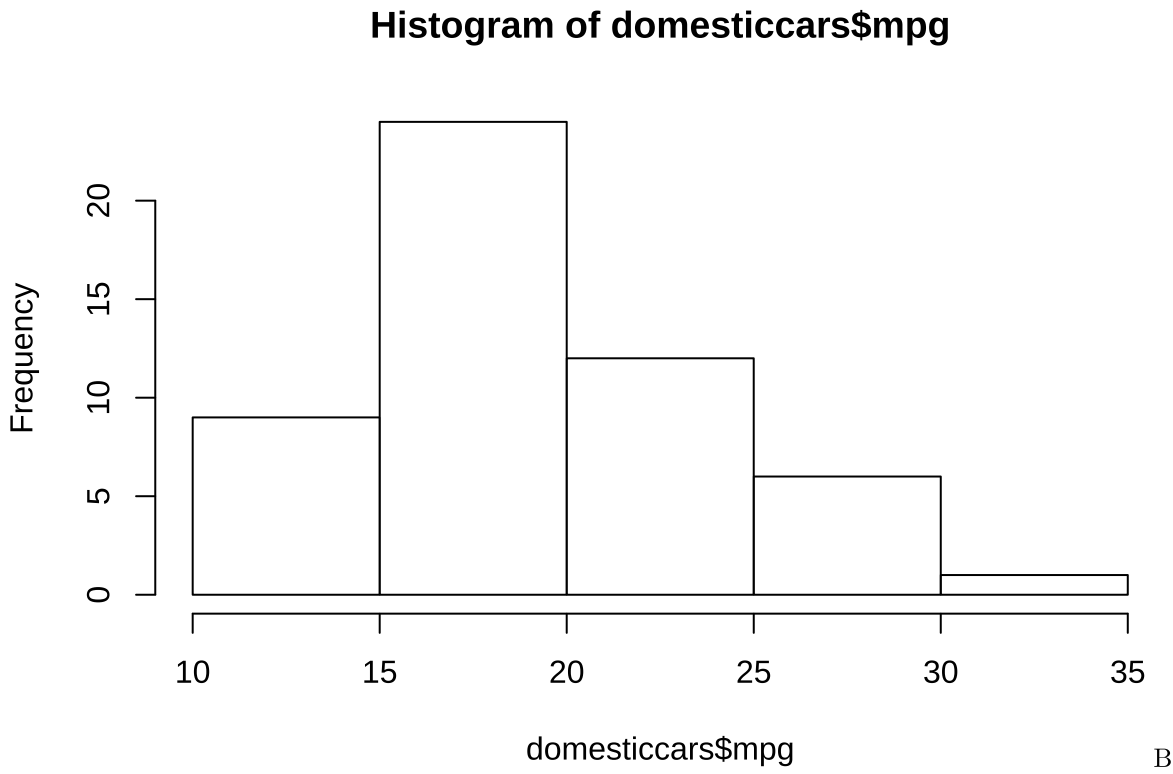
```
summary(domesticcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    12.0    16.8    19.0    19.8    22.0    34.0
```

```
# Histograms for each respective mpgs
hist(foreigncars$mpg)
```



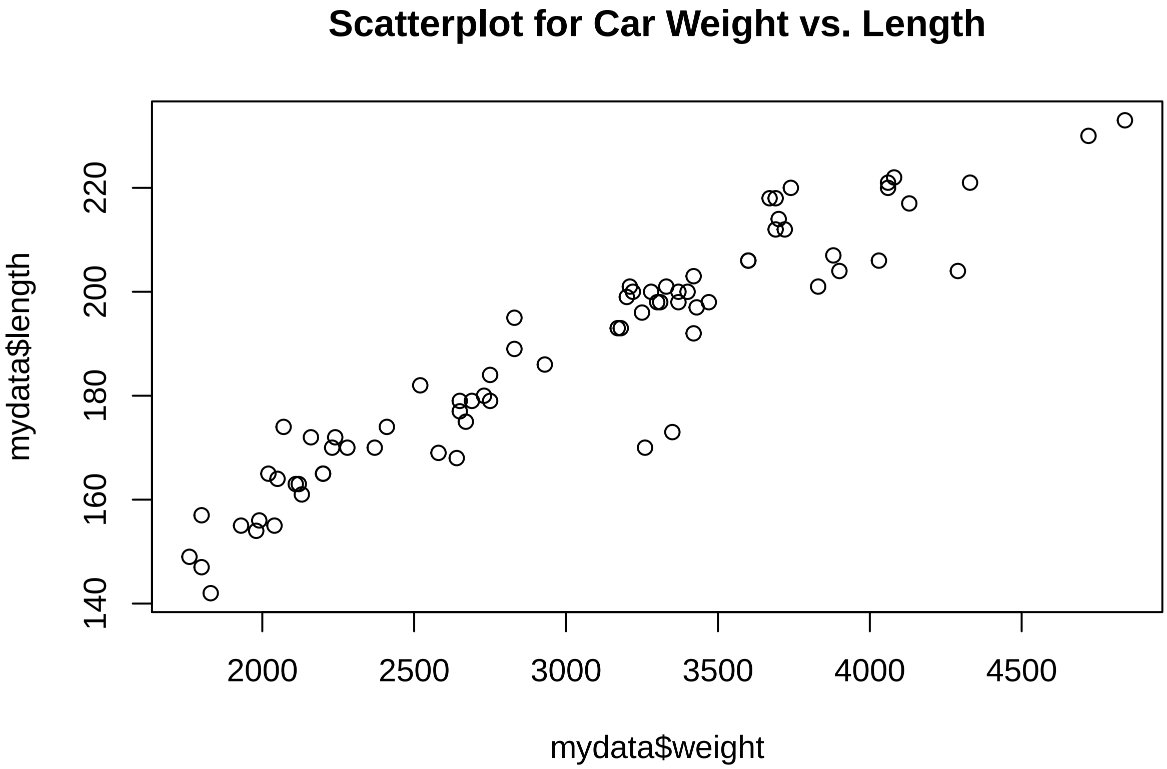
```
hist(domesticcars$mpg)
```



Both foreign cars and domestic cars have distributions skewed to the right. This can be seen as both subsets as the means are larger than the medians and have histrograms with higher value datapoints on the right. Foreign cars have a higher median of 24.5 mpg and mean of 24.8 mpg, while domestic cars have a median of 19.0 mpg and mean of 19.8 mpg. As such, the graph of foreign cars have data points shifted to right with higher average values. A possible outlier also exists for foregin cars with the 41.0 mpg max.

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

```
# Scatter plot of car weight vs. length for my data
plot(mydata$weight, mydata$length, main = "Scatterplot for Car Weight vs. Length")
```



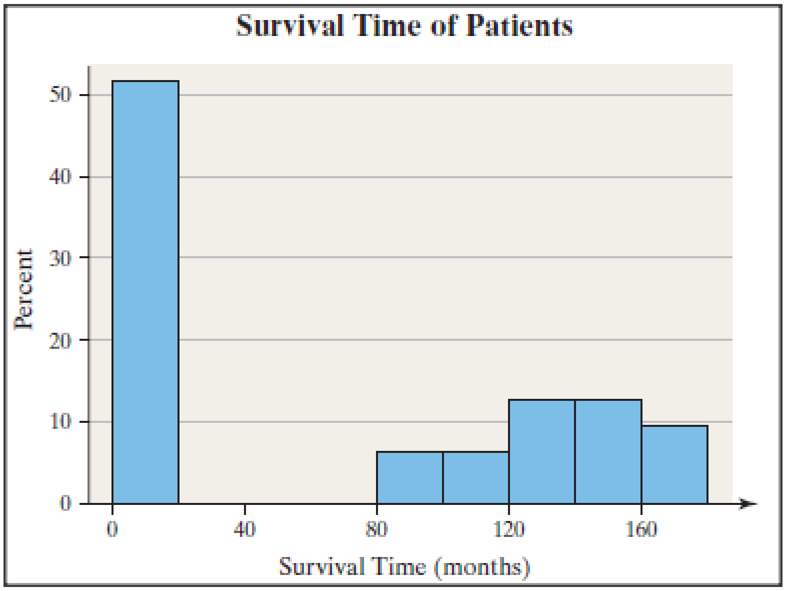
Based on the scatter plot, there seems to be a positive association between car weight and car length. According to this association, cars with lower weights are shorter in length and cars with higher weights are longer in length in the plot.

Problem 5

(thought exercise) 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.]

I would tell my friend that although the median survival time is 11 months, he/she should consider the mean survival time of 69 more. The average indicates that half the people who are diagnosed do live 69 months or more. And he/she has an equally good chance to live more than 69 months based on this data. The median is not as helpful in this case because it is much lower than the mean, indicating the data is skewed to the right. And while the middle value (median) is 11 months, the mean of 69 months tells a different story in that the latter half of the data points have much longer survival durations (>69 months) to bring the mean up to 69 months.

As seen in the Survival Times of Patients, the graph shows a skew to the right with half the data points are are between 0-20 months and the rest have a cluster of data points above 80 months. The cluster of data points greater than 80 months are spread up to 180 months as a max. It is equally likely that my friend’s diagnosis will fall in the latter range of data points and his/her predicted survival length could be between 80 and 180 months based on the data.



Problem 6

(thought exericse) 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.

This is possible because Seth has an IQ higher than the average IQ at Yale, but lower than the average IQ at Harvard. Seth’s IQ is lower than the Harvard average (mean) and by removing his IQ data point, the mean increases. At Yale, the average IQ (mean) is increased since Seth’s IQ is higher than Yale’s average, and factors into the mean by increasing the average IQ.

Problem 7

(thought exercise) 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.

Assuming that the average diameter is correct for the new sample of tires, the manager should be happy with the production line. Having a 0 standard of deviation for diameter, means that all the tires are the exact same diameter. A standard deviation of 0 means that each data point is the same as the mean, and so each tire has the same diameter as the average tire. This ensures that when the cars are built, they will have a consistent alignment/height because the diameters of the tires are all the same. And if a tire is a replacement for the ohter tires, the diameter match the one being replaced and keep the height of the car.

Problem 8

Use this data set for the following question {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 mean.

```
# Standard deviation for x
sd(x)
```

```
## [1] 15.81
```

```
# Mean for x
mean(x)
```

```
## [1] 30
```

Standard deviation is 15.81 and mean is 30.

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

```
y=c(15,25,35,45,55)
sd(y)
```

```
## [1] 15.81
```

```
mean(y)
```

```
## [1] 35
```

Std. dev. is 15.81 and mean is 35.

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

```
z=c(5,15,25,35,45)
sd(z)
```

```
## [1] 15.81
```

```
mean(z)
```

```
## [1] 25
```

Std. dev is 15.81 and mean is 25.

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

```
a=c(50,100,150,200,250)
sd(a)
```

```
## [1] 79.06
```

```
mean(a)
```

```
## [1] 150
```

Std. dev. is 79.06 and mean is 150.

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

```
b=c(2,4,6,8,10)
sd(b)
```

```
## [1] 3.162
```

```
mean(b)
```

```
## [1] 6
```

Std. dev. is 3.162 and mean is 6.

f) Generalize the results of parts b through e. When adding and subtracting the standard of deviations stayed the same as the original set of x's standrard of deviation. The mean when adding 5 (part b) to all values increased by 5 compared to set x. And the mean when subtracting 5 (part c) to all values decreased by 5 compared to set x.

When multiplying by a factor of 5, the standard of deviation and mean increased compared to the original set of x. When dividing by a factor of 5, the standard of deviation and mean decreased as compared to set x.

Problem 9

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. Decide 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. You do not have to give an explanation.

a) The average salary is below $60,000.
Cannot tell

b) The median salary is below $60,000. Cannot tell

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

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

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

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

Problem 10

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?

```
12.3-(15.7*2) # lower value; mean - (s.d.*2)
```

```
## [1] -19.1
```

```
12.3+(15.7*2) # upper value; mean + (s.d.*2)
```

```
## [1] 43.7
```

Between -19.1% and 43.7%

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

```
12.3-(15.7*3) # lower value; mean - (s.d.*3)
```

```
## [1] -34.8
```

```
12.3+(15.7*3) # upper value; mean + (s.d.*3)
```

```
## [1] 59.4
```

Between -34.8% and 59.4%

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

No, an investor should not be surprised with a negative return. Based on Chebychev's Inequality and the standard of deviation of 15.7%, a significant portion of the potential return values are in a negative range. The values from less than 0 all the way out to the minimum negative value of Chebychev's inequality (-19.1 in part a and -34.8 in part b) have a likely chance to be a return value. And her return has an chance of being a negative value within this range.

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 the data is mound shaped data, when plotted as a historgram.

Problem 11

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

```r
x_1=2
x_2=-1
x_3=0
2 + sum((5*x_1), (5*x_2), (5*x_3))
```

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

Answer is 7.

####Problem 12 We have data on the amount of the dinner bill and the resulting tip from a local restaurant. Read the data in as follows:

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

```
## [1] "Bill"   "Tip"    "Credit" "Guests" "Day"    "Server" "PctTip"
```

Let's first build a variable that has the tip percentage:

```
tiper=100*mydata$Tip/mydata$Bill
summary(tiper)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.67   14.29   16.20   16.62   18.19   42.19
```

The median tip is a bit above 16% and more than 25% of the people tip more than 14%. Someone tipped an amazing 42%. [my bad-I just realized there is a variable in this data set called PctTip which is the same thing I just defined. Oh well.]

How many people tipped above 40%? Looks like two people:

```
sum(tiper>40)
```

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

Suppose we want to remove these big tippers from our data set. One way to do so is to make a new data set with these 2 points removed:

```
dim(mydata)
```

```
## [1] 157   7
```

```
newdata=subset(mydata,tiper<40)
dim(newdata)
```

```
## [1] 155   7
```

```
newtiper=100*newdata$Tip/newdata$Bill
summary(newtiper)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.67   14.28   16.07   16.30   18.03   31.79
```

The code above shows we started with 157 rows of data, and when we delete the two largest tippers our new data set has 155 rows of data. Note that we have to create a new variable for tip percentage for the new data set, and this new variable has a max less than 40.

a) Using the box plot rule, how many Tip values are considered outliers (use the original data set).

```
length(boxplot.stats(mydata$Tip)$out)
```

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

There are 9 outliers for Tip values in the original data.

b) Using the box plot rule, how many tiper (tip percentage) values are considered outliers (use the original data set).

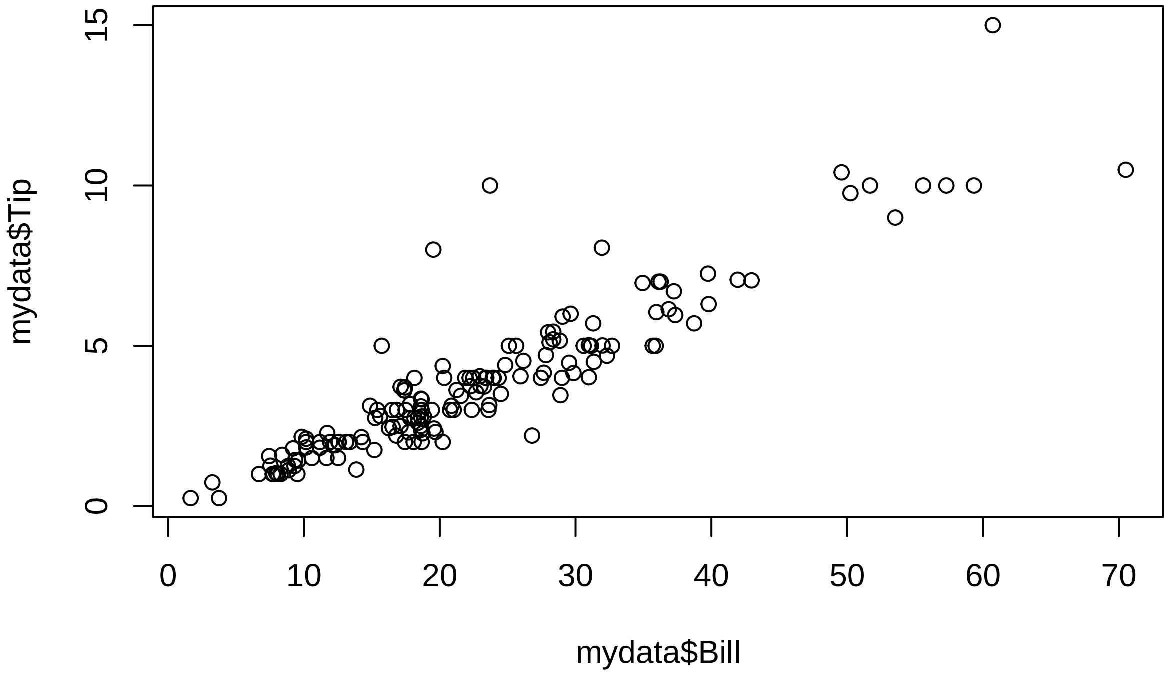
```
length(boxplot.stats(tiper)$out)
```

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

There are 8 outliers in the Tiper, accroding to the box plot rule.

c) Using the original data set, what is the correlation between dinner bill and tip?

```
# Graph of bill vs. tip
plot(mydata$Bill, mydata$Tip)
```



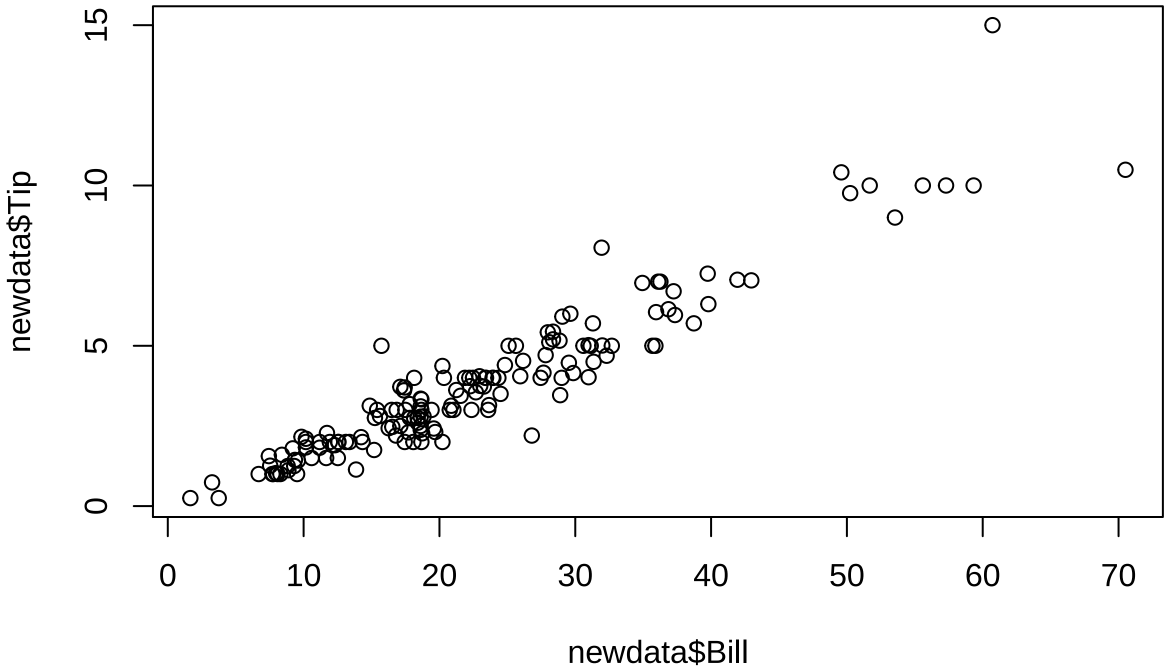
```
# Correlation of bill and tip
cor(mydata$Bill, mydata$Tip)
```

```
## [1] 0.9151
```

The graph shows a positive relationship between bill and tip. The correlation function reveals the correlation coefficient to be 0.9151. This indicates that bill and tip have a strong positive correlation. Higher bills with higher tips, and lower bills with lower tips.

d) Using the data set with the two largest tip percentages removed, what is the correlation between dinner bill and tip? Is this number the same as from part (c)? Explain.

```
# Graph of bill vs. tip
plot(newdata$Bill, newdata$Tip)
```



```
# Correlation of bill and tip
cor(newdata$Bill, newdata$Tip)
```

```
## [1] 0.9462
```

The correlation coefficient for the new data, without the two outliers, is 0.9462. This value is higher than the original data, indicating a stronger positive relationship between bill and tip. Outliers, like the two points, can affect the correlation coefficient. These values will have have an effect on the standard deviation of x (bill) and standard deviation of y (tips) and this has the affect of decreasing the correlation coefficient. Without the outliers, the coefficient increases.

Problem 13

This question moves us in the direction of understanding that just because two variables are uncorrelated does not mean they are independent.

a) Explain in words what a correlation of 0 implies. Correlation of 0 implies no relationship between the two variables.

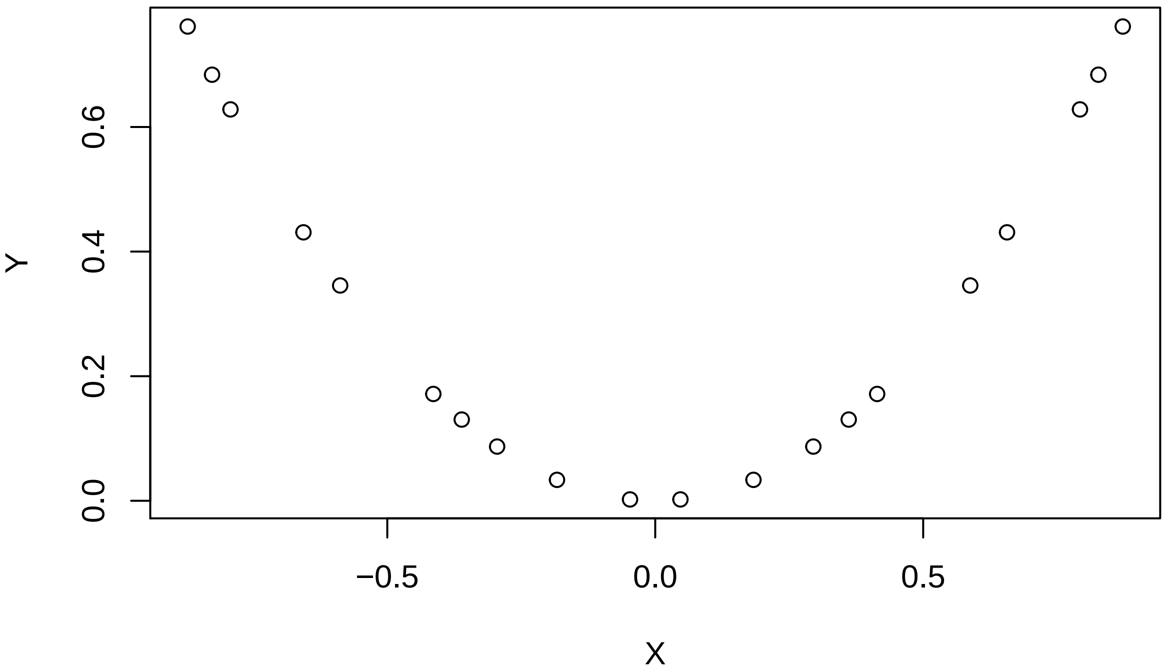
A correlation of 0 implies that there is no relationship between variables.

b) Load the blas data set into R and find the correlation of X and Y

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

c) Plot the data-does it agree with your definition?

```
#
plot(mydata)
```



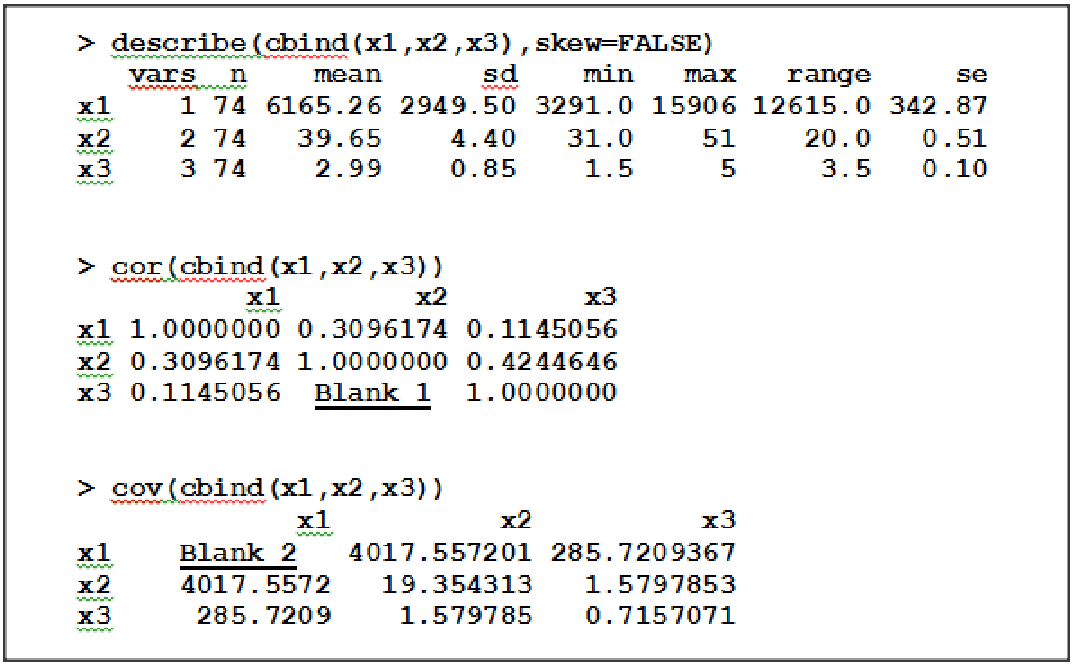
```
# Correlation coefficient of 0
cor(mydata$X, mydata$Y)
```

```
## [1] 1.041e-20
```

This does not exactly follow the definition from part A about a 0 coefficient. That definition assumes that we are looking for a "linear" relationship. However, when we plot this data, and while we do not see a linear relationship, X and Y seem to have a "quadratic" relationship, with the curve resembling a parabola. If we check the correlation coefficient the value is 1.041e-20, which is a small decimal close to 0. This just says that X and Y do not have a linear relationship, but nothing about another type of relationship.

Problem 14

Fill in the blanks using the definition of covariance and correlation



```
> describe(cbind(x1,x2,x3),skew=FALSE)
   vars  n    mean      sd    min   max   range     se
x1    1 74 6165.26 2949.50 3291.0 15906 12615.0 342.87
x2    2 74   39.65    4.40   31.0    51    20.0   0.51
x3    3 74    2.99    0.85    1.5     5     3.5   0.10


> cor(cbind(x1,x2,x3))
          x1        x2        x3
x1 1.0000000 0.3096174 0.1145056
x2 0.3096174 1.0000000 0.4244646
x3 0.1145056  Blank 1  1.0000000


> cov(cbind(x1,x2,x3))
             x1           x2          x3
x1    Blank 2   4017.557201 285.7209367
x2    4017.5572   19.354313   1.5797853
x3     285.7209    1.579785   0.7157071
```

a) Blank1 equals 0.4244646

b) Blank2 equals 8699550.25 (s.d.ˆ2 = 2949.50ˆ2) cov(x_1) = var(x_1,x_1)

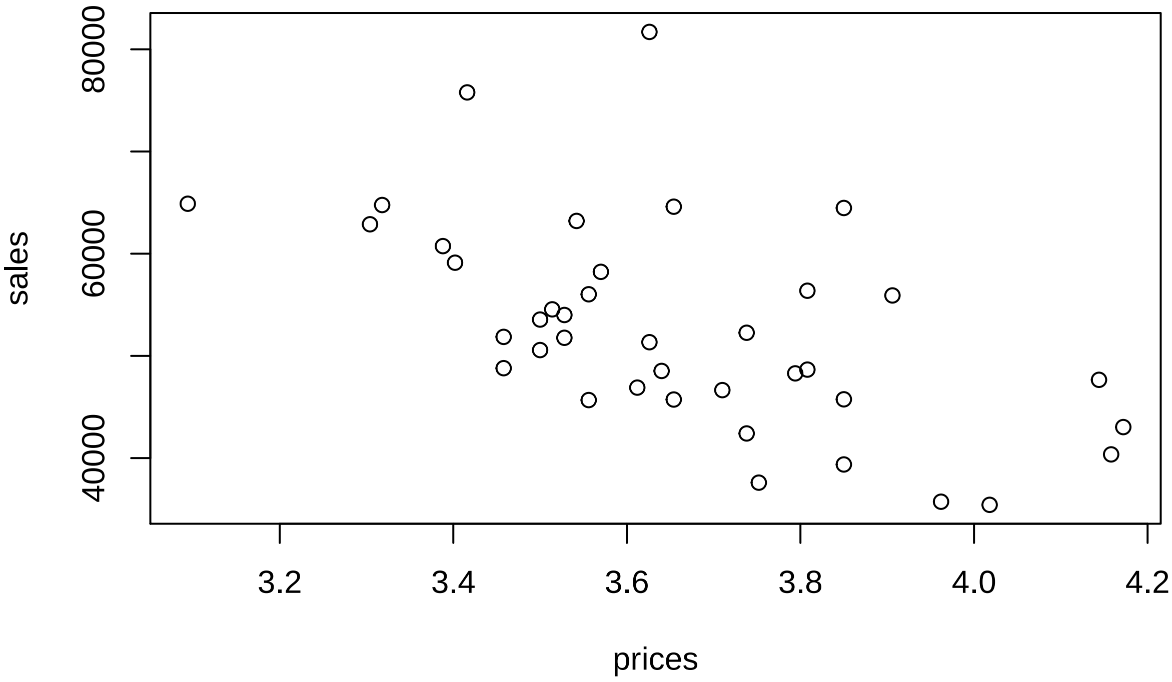
Problem 15

We have data on frozen pizza sales (in pounds) and average price ($/unit) from Dallas Texas for 39 recent weeks. Load the class survey data into R using the command

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/pizzasales1.csv")
```

a) Using price as the explanatory variable and sales as the response variable, run a regression and write down the linear equation relating sales to price from the output.

```
prices=mydata$price #X
sales=mydata$sales #Y
# Scatterplot of Prices vs. Sales
plot(prices,sales)
```



```
# Least squares method for finding regression coefficients b_0 and b_1
lm(sales~prices)
```

```
##
## Call:
## lm(formula = sales ~ prices)
##
## Coefficients:
## (Intercept)       prices
##      141866       -24369
```

The linear regression equation for the sales is Y = 141866 - 24369*X, where b_0 the y_intercept is 141866 and b_1 the slope is -24369.

b) What does the slope mean in this context? The slope shows a negative linear relationship between frozen pizza price and sales. Each time the price goes increases by 1(/unit) the sales decrease by 24369 pounds of pizzas. This makes sense as more expensive pizzas sell less and cheaper pizzas are a better deal. The slope shows that lower priced pizzas have more sales and higher price pizzas have lower sales, based on the data.

c) What does the y-intercept mean in this context? Is it meaningful? The Y intercept in this context indicates that The Y-intercept indicates the Y (sales) value when when X (price, $/unit) is 0 in the graph.

The Y-intercept is 141866 pounds of pizza. This means that when the price is $0/unit already 141866 pounds have been sold. This is not very helpful because the Y-intercept in this case is only adjusting for the linear regression and does not necessarily mean that 141866 pounds are sold when the pizza is free. It is probably unlikely the so much frozen pizza is sold for free.

d) What do you predict the sales to be if the average price charged was $3.50 for a pizza?

```
# Calculating sales in pounds using the linear regression equation above
141866 - (24369*3.50)
```

```
## [1] 56574
```

According to the linear regression equation, 56574 pounds of frozen pizza are predicted to be sold at $3.50.

e) If the sales for a price of $3.50 turned out to be 60,000 pounds, what would the residual be?

```
# Residual = observed - predicted; e = y - y_hat
60000-56574
```

```
## [1] 3426
```

The residual (e) is 3426 pounds of frozen pizza. This is the difference between the observed sales (60,000 pounds) and the predicted sales (56,574 pounds).

Problem 16

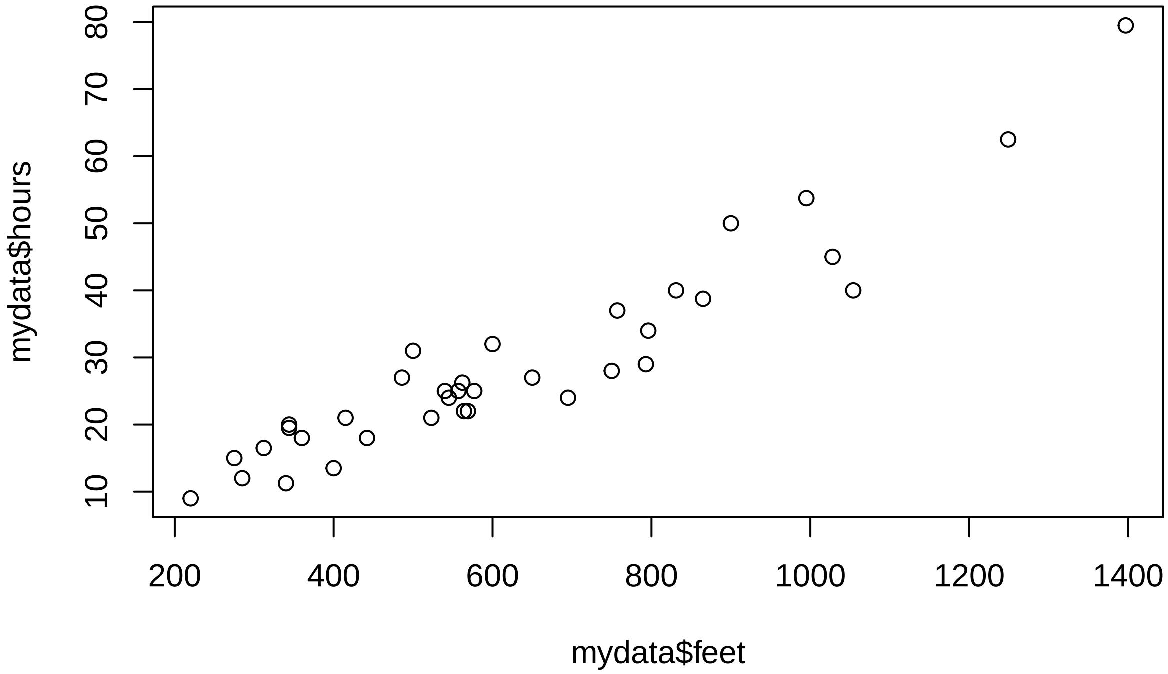
The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours (Y). In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved as the independent variable (X) and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and in which the travel time was an insignificant portion of the hours worked. The data may be loaded into R as follows

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

Use R to answer the questions below.

a) Create a scatter diagram of the data.

```
plot(mydata$feet, mydata$hours)
```



b) a least squares regression line to this data and interpret the slope.

```
lm(mydata$hours~mydata$feet)
```

```
##
## Call:
## lm(formula = mydata$hours ~ mydata$feet)
##
## Coefficients:
## (Intercept)  mydata$feet
##     -2.3697       0.0501
```

The slope is 0.0501 hours/cubic feet. According to the regression line, as the distance moved increases by 1 cubic feet, the number of labor hours increases by 0.0501.

c) Predict the labor hours for a 500 cubic feet move using the estimated regression equation developed in part (b).

```
# From the regression equation of part b
-2.3697+(0.0501*(500))
```

```
## [1] 22.68
```

For 500 cubic feet moved, the predicted labor hours is 22.68 hours.