Stats 10 Lab 2
Name: Brandon Truong

Section 1

a)

```
> flint <- read.csv('~/UCLA Coursework/STATS 10/flint.csv')
> head(flint) #testing to see if read properly
  ï..Latitude Longitude Pb  Cu Region
1    43.09414 -83.60974  0   0  North
2    43.09054 -83.70344  0 130  North
3    43.08601 -83.71996  4 170  North
4    43.08100 -83.75415  0   0  North
5    43.07435 -83.70043  0   0  North
6    43.07399 -83.71788  0   0  North
> class(flint)
[1] "data.frame"
```

b)

The proportion of locations with dangerous lead levels is 0.04436229, or 4.4%.

```
> library(mosaic)
> dangerousPb_indicator = (flint$Pb >= 15)
> tally(~dangerousPb_indicator, format="proportion")
dangerousPb_indicator
      TRUE      FALSE
0.04436229 0.95563771
> mean(flint$Pb>=15)
[1] 0.04436229
```

c)

The mean copper levels is 44.6424 ppb in the Northern region.

```
> north_flint <- flint[flint$Region=="North",]
> mean(north_flint$Cu)
[1] 44.6424
```

d)

The mean copper levels in locations with dangerous Pb levels is 305.8333 ppb.

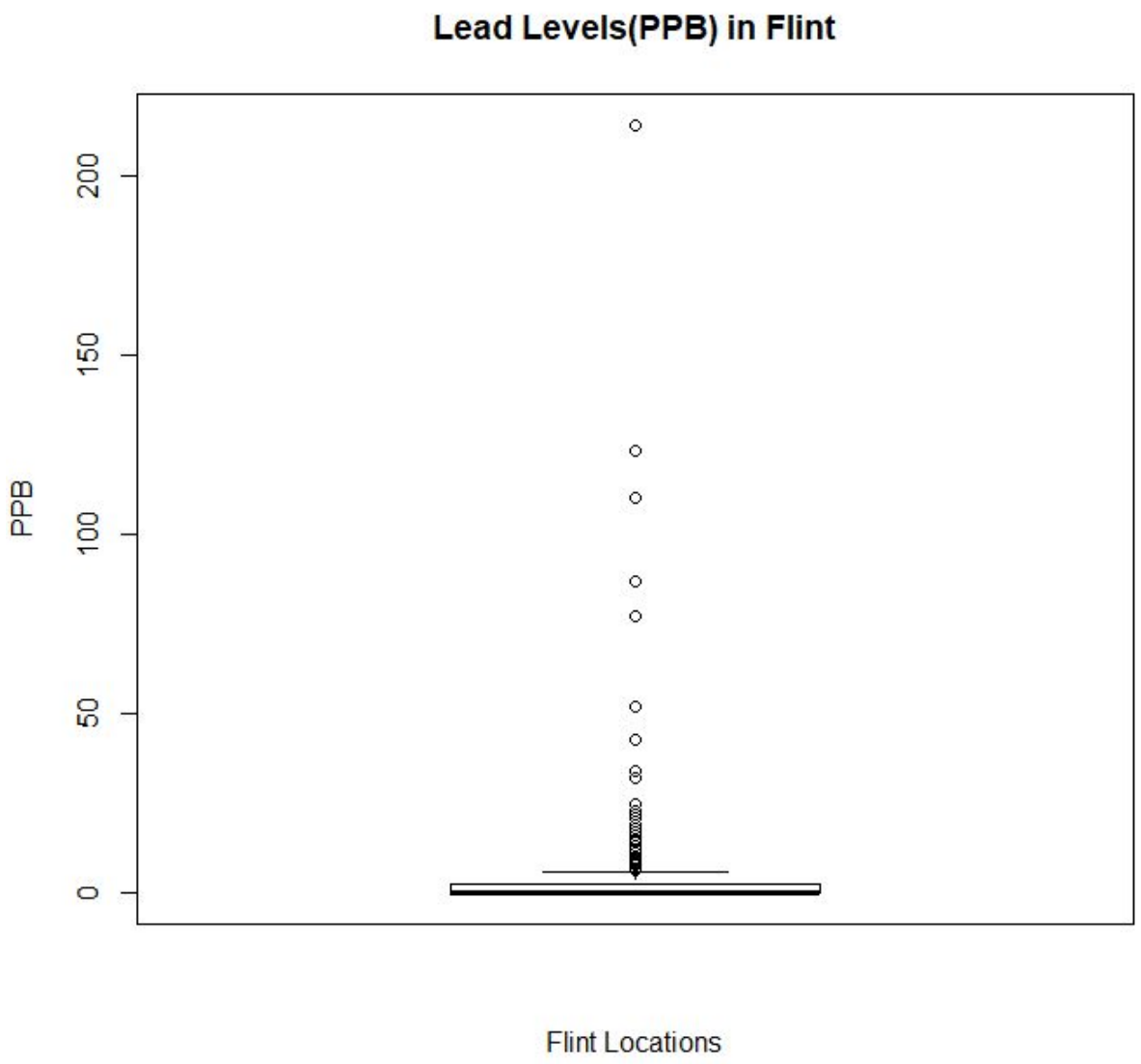
```
> dangerousPb_flint <- flint[flint$Pb>=15,]
> mean(dangerousPb_flint$Cu)
[1] 305.8333
```

e)

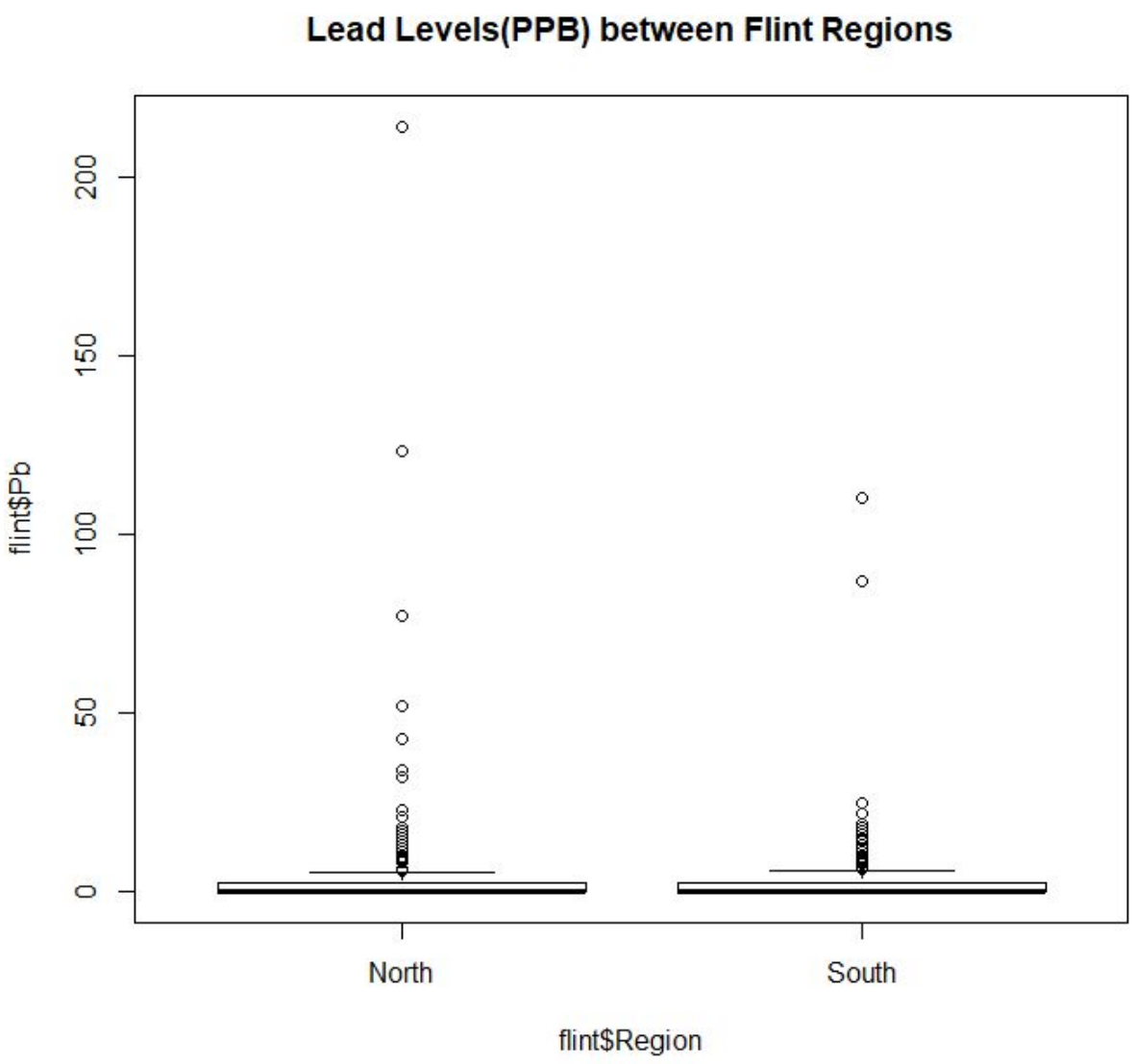
The mean lead level is 3.383272 ppb and copper level 54.58102 ppb.

```
> mean(flint$Pb)
[1] 3.383272
> mean(flint$Cu)
[1] 54.58102
```

f)




```
> boxplot(x = flint$Pb, main="Lead Levels(PPB) in Flint", xlab="Flint Locations",ylab="PPB")
```



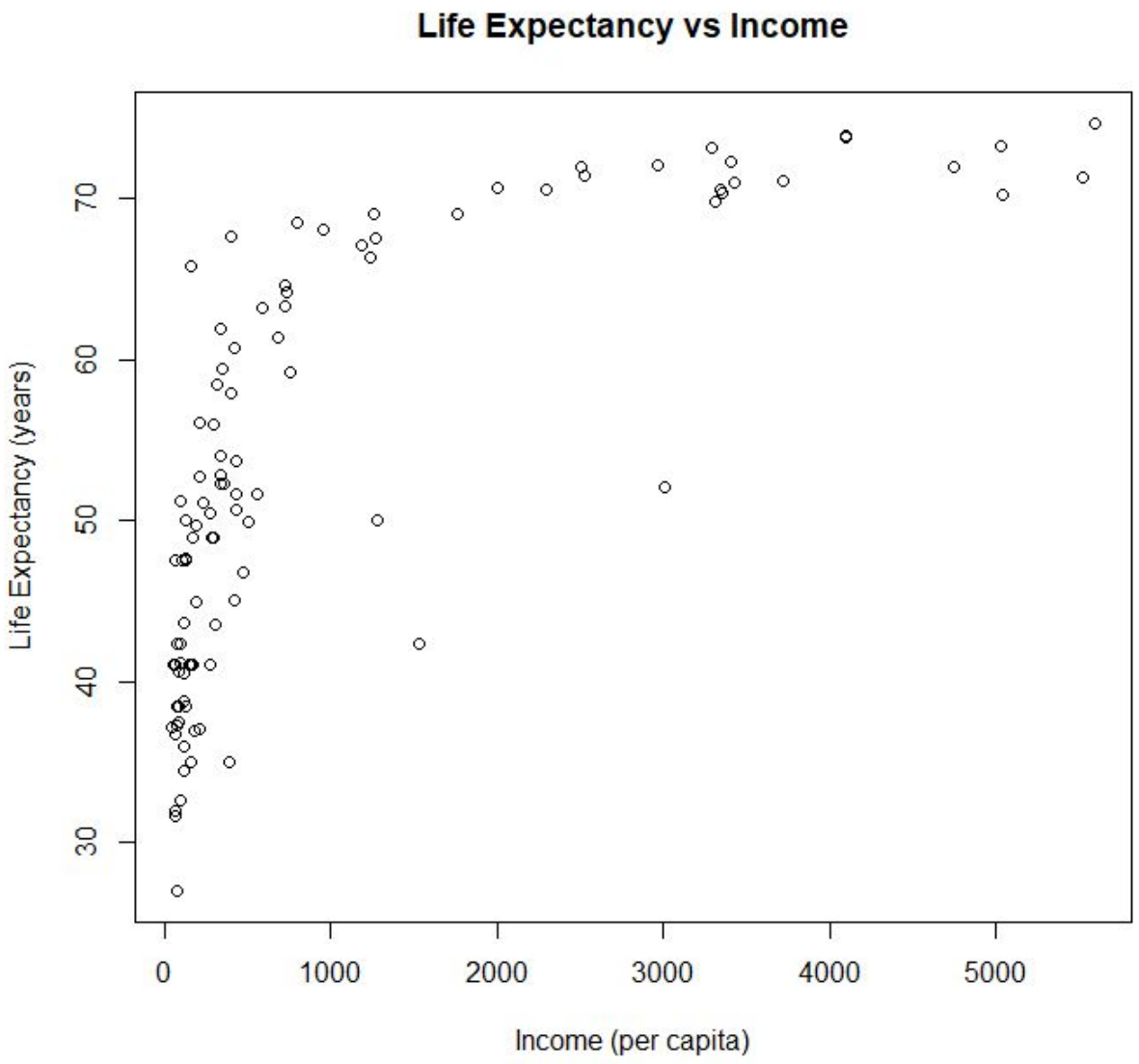

```
> boxplot(flint$Pb~flint$Region, main="Lead Levels(PPB) between Flint Regions")
```

g)
Based on the boxplot of part f, we can see that there are many outliers in the data. This means that the mean would not be a good measure of center of data, since both mean and standard deviation are not resistant to outliers, thus making the data skewed towards higher values. Median and IQR are better measures since they are resistant to outliers.

```
> mean(flint$Pb)
[1] 3.383272
> median(flint$Pb)
[1] 0
```

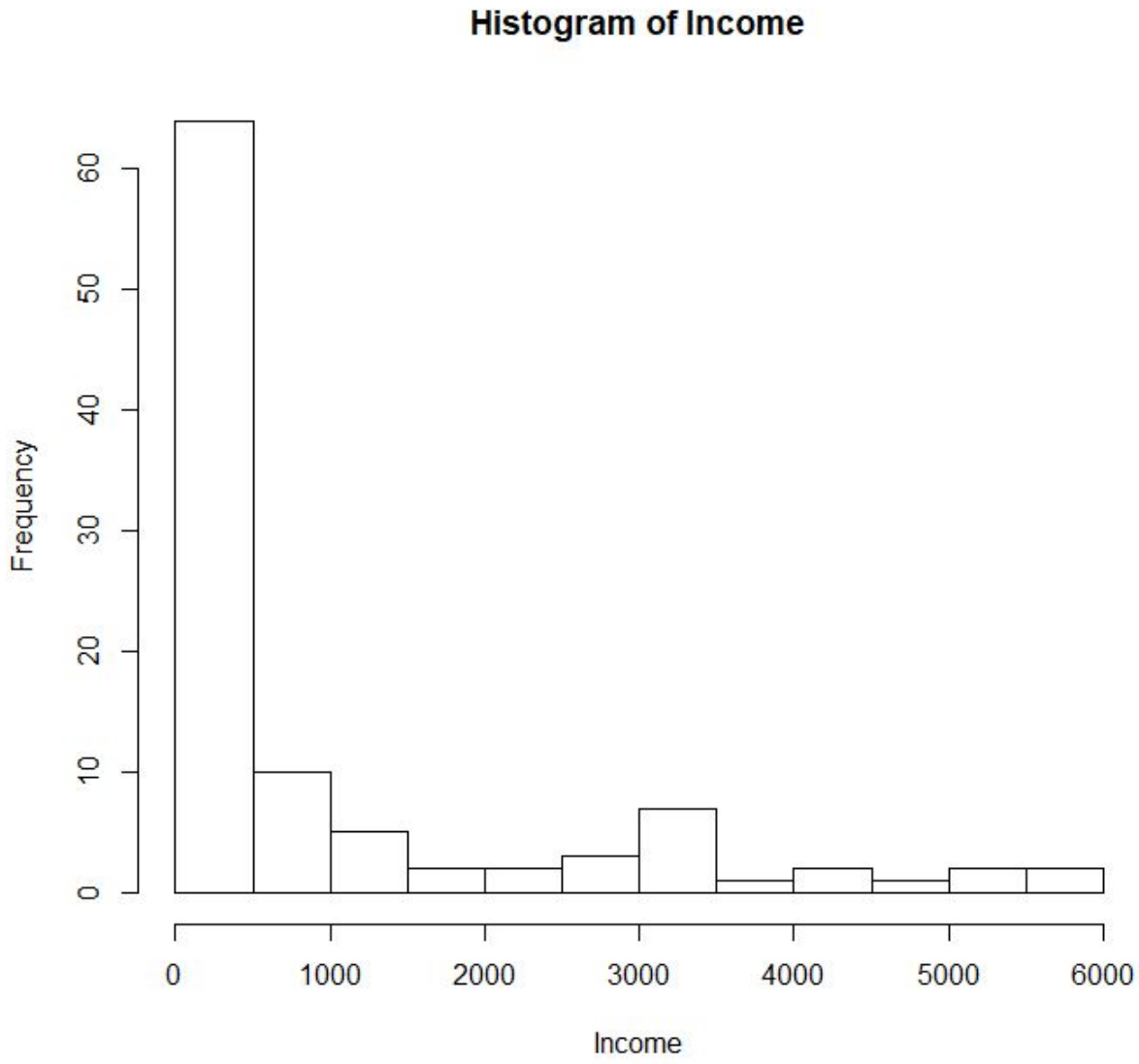
Section 2

a)
Income appears to be positively associated with Life Expectancy up to a point, creating an almost logarithmic graph. After hitting 1000 income, it appears to level off, increasing only slightly from a life expectancy of around 70 years. Since this is only data without a control or treatment, this implies an observational study and that we can only conclude correlation, not causation.

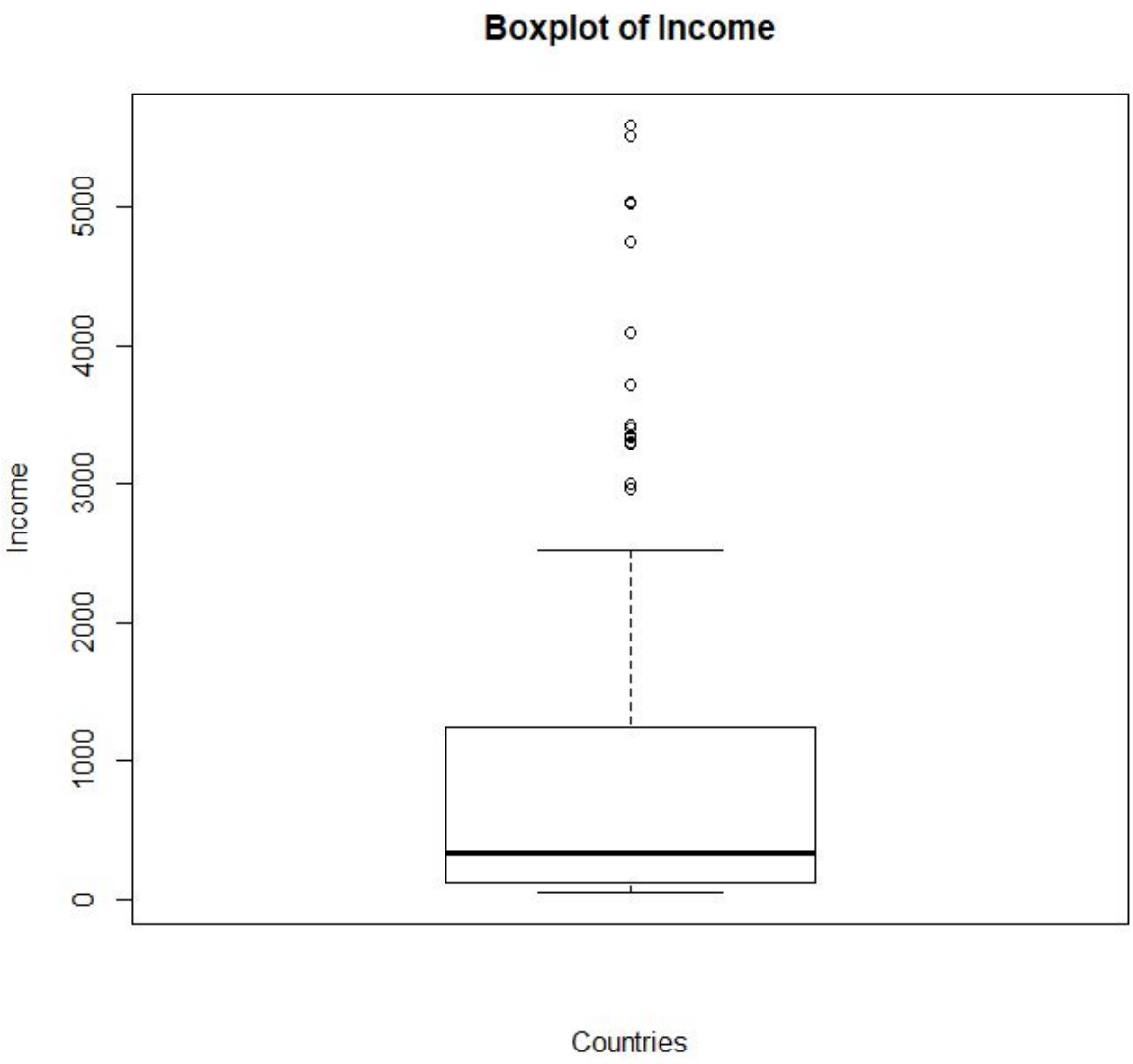



```
> plot(x=life$Income, y=life$Life,xlab="Income (per capita)",ylab="Life Expectancy (years)",main="Life Expectancy vs Income")
```

b)
We can use the histogram to tell that the graph is right-skewed, with outliers in the right. Boxplot allows us to recognize the outliers (indicated by a dot), which are located to the right. By using Q3+1.5IQR, we can find that outliers begin after 2912.5 Income. This means that there are 16 potential outliers.




```
> hist(life$Income,xlab="Income",ylab="Frequency",main="Histogram of Income")
```




```
> boxplot(life$Income,xlab="Countries",ylab="Income",main="Boxplot of Income")
```

c)

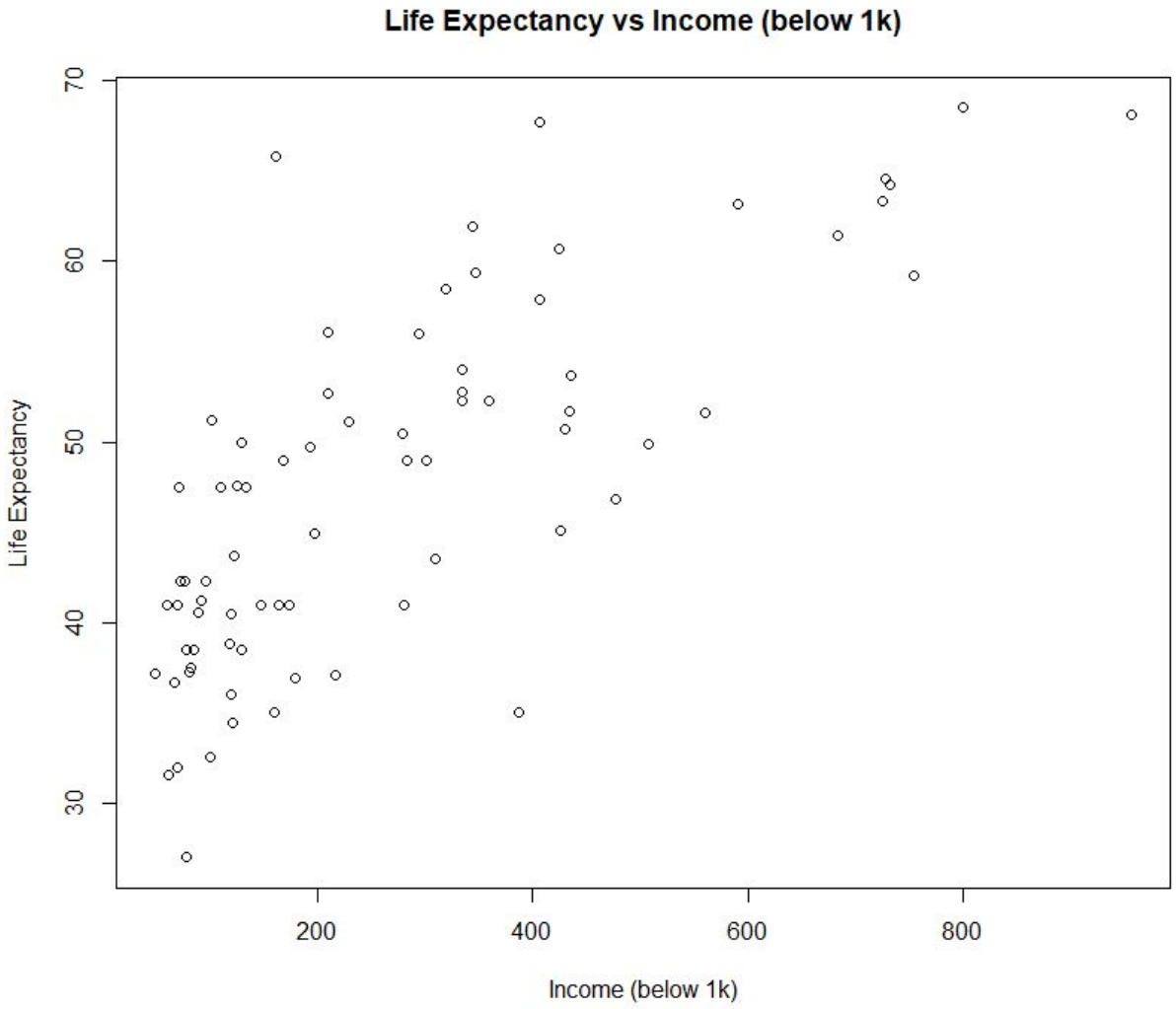
```
> below1kInc <- life[life$Income<1000,]
> above1kInc  <- life[life$Income>=1000,]
```

d)

The correlation (r) is .752886

```
> cor(below1kInc$Income, below1kInc$Life)
[1] 0.752886
```

```
> plot(x=below1kInc$Income, y=below1kInc$Life,xlab="Income (below
1k)",ylab="Life Expectancy",main="Life Expectancy vs Income (below 1k)")
```



Section 3

```
> maas<-read.table("http://www.stat.ucla.edu/~nchristo/statistics12/soil.txt",
header=TRUE)
```

a)

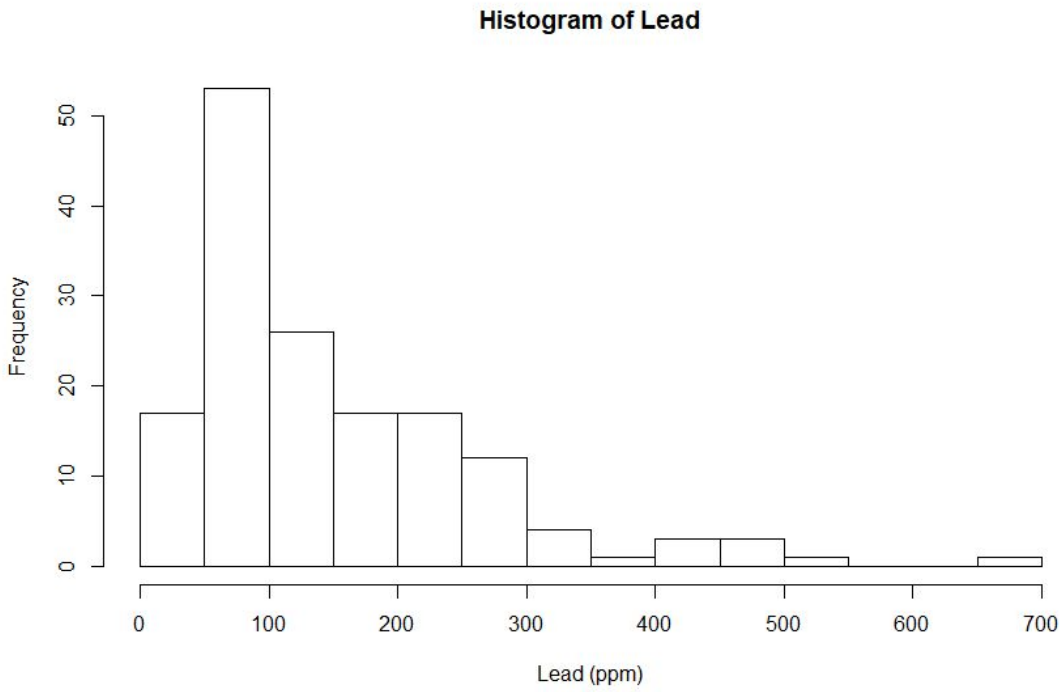
Lead

```
> summary(maas$lead)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   37.0    72.5   123.0   153.4   207.0   654.0
```

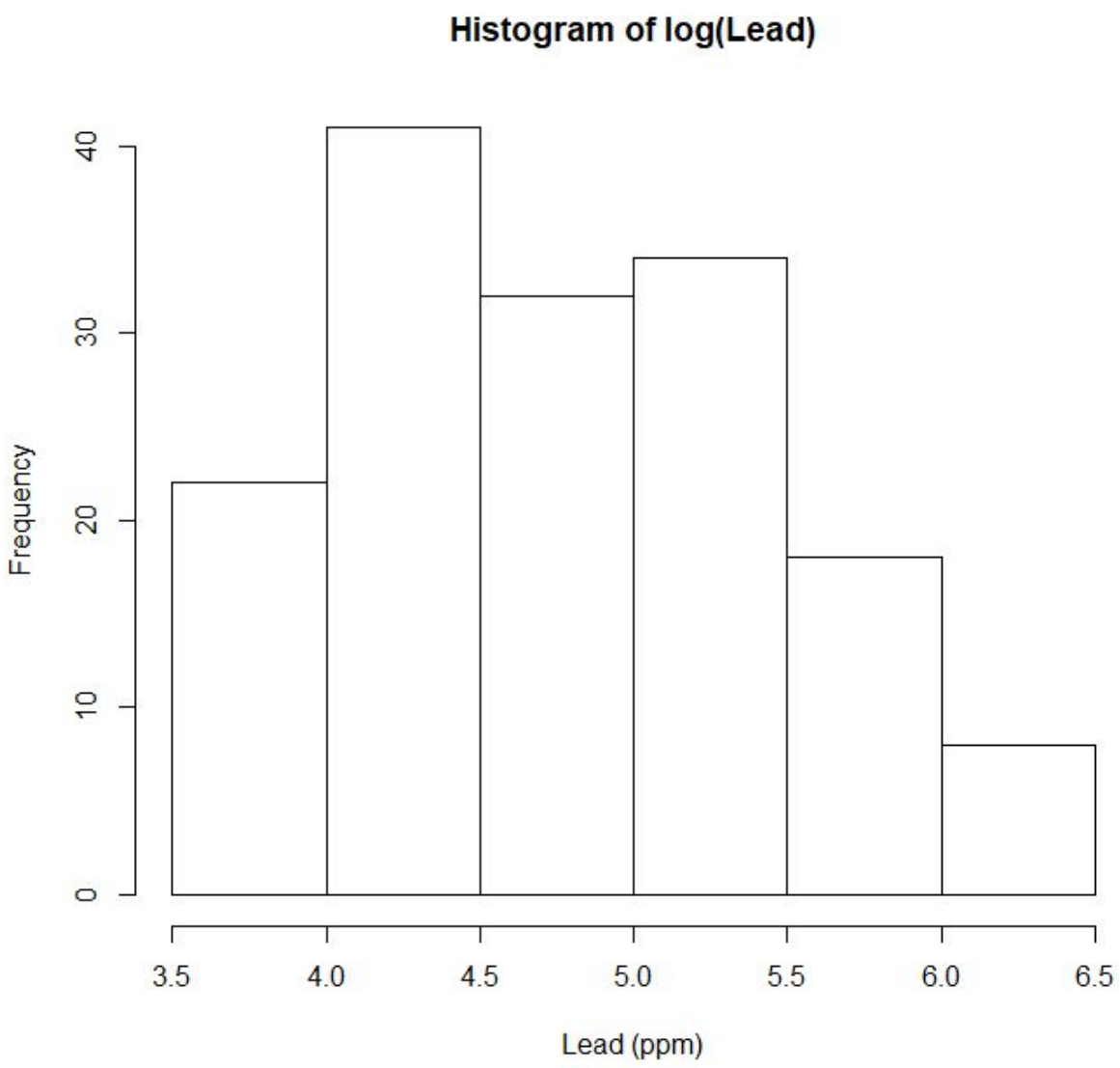
Zinc

```
> summary(maas$zinc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  113.0   198.0   326.0   469.7   674.5  1839.0
```

b)

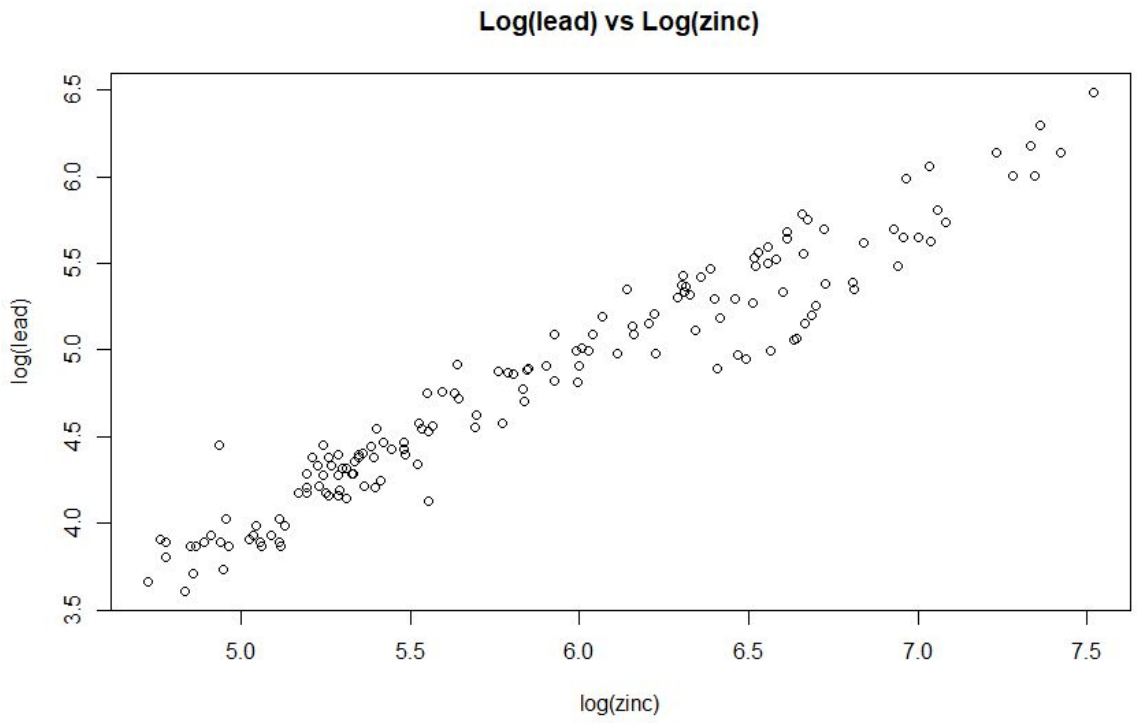



```
> hist(maas$lead, xlab="Lead (ppm)",main="Histogram of Lead")
```



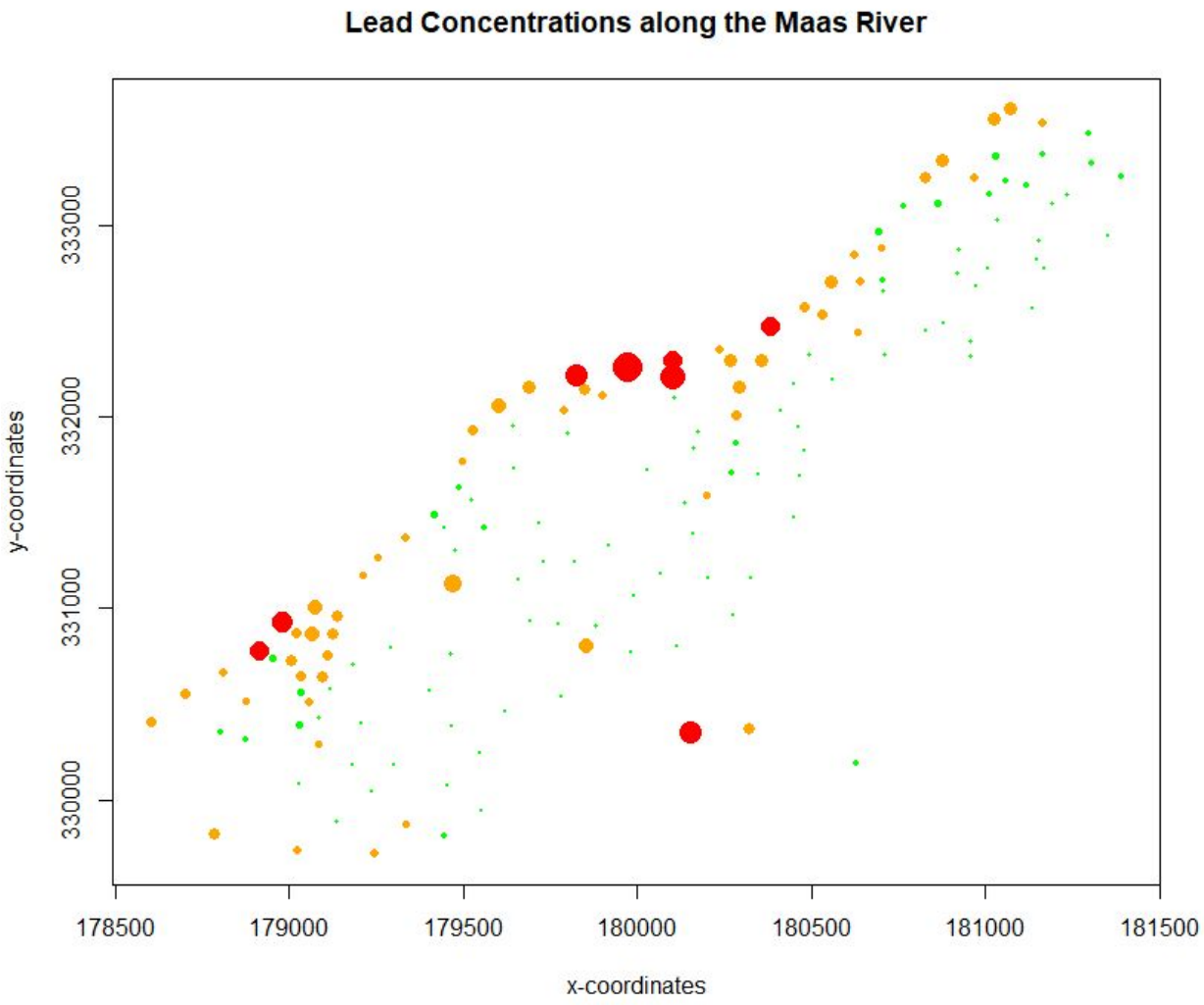

```
> hist(log(maas$lead),xlab="Lead (ppm)",main="Histogram of log(Lead)")
```

c)
We notice when we plot log(lead) against log(zinc) that their relationship is linear, creating a positively associated graph. This means that the two data sets have near equal variance due to the linearity of the graph. This also means that their relationship is symmetric, meaning we can flip them and still have the same correlation coefficient.




```
> plot(x=log(maas$zinc),y=log(maas$lead),ylab="log(lead)", xlab="log(zinc)",
main="Log(lead) vs Log(zinc)")
```

d)

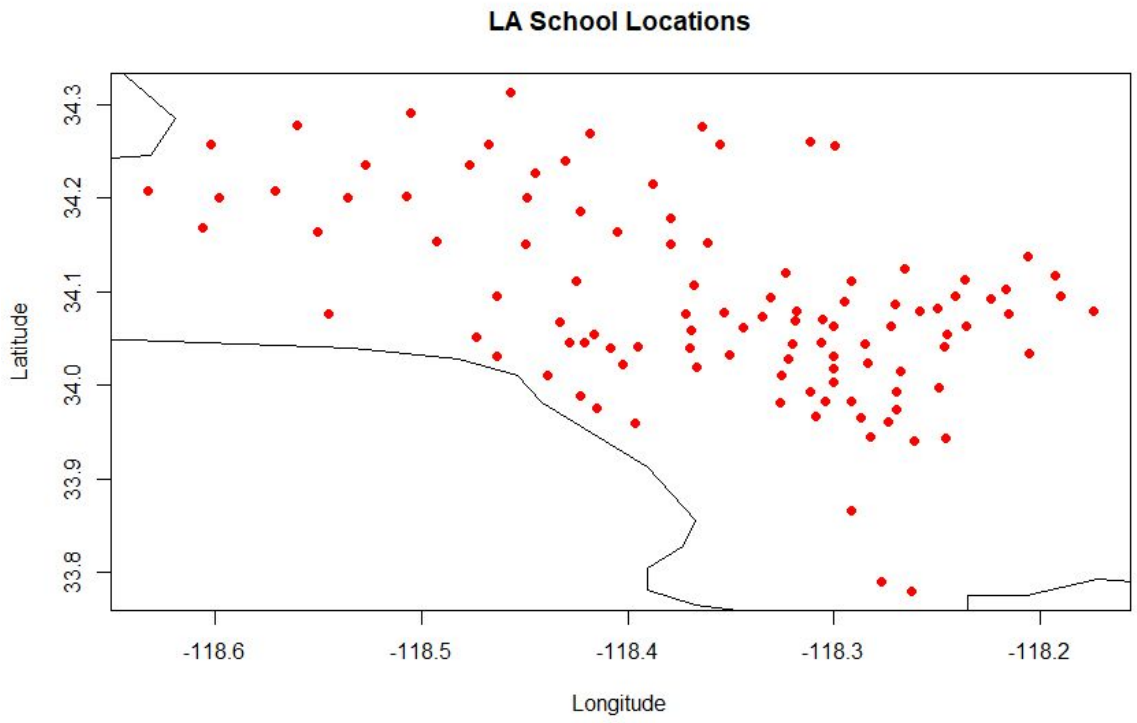



```
> lead_colors <- c("green","orange","red")
> lead_levels <- cut(maas$lead, c(0,150,400,1000))
> plot(maas$x,maas$y,xlab="x-coordinates",ylab="y-coordinates",main="Lead
Concentrations along the Maas River","n")
>points(maas$x, maas$y, cex=maas$lead/mean(maas$lead)/1.5,
col=lead_colors[as.numeric(lead_levels)], pch=19)
```

Section 4

```
LA<-read.table("http://www.stat.ucla.edu/~nchristo/statistics12/la_data.txt",
header=TRUE)
> library(maps)
```

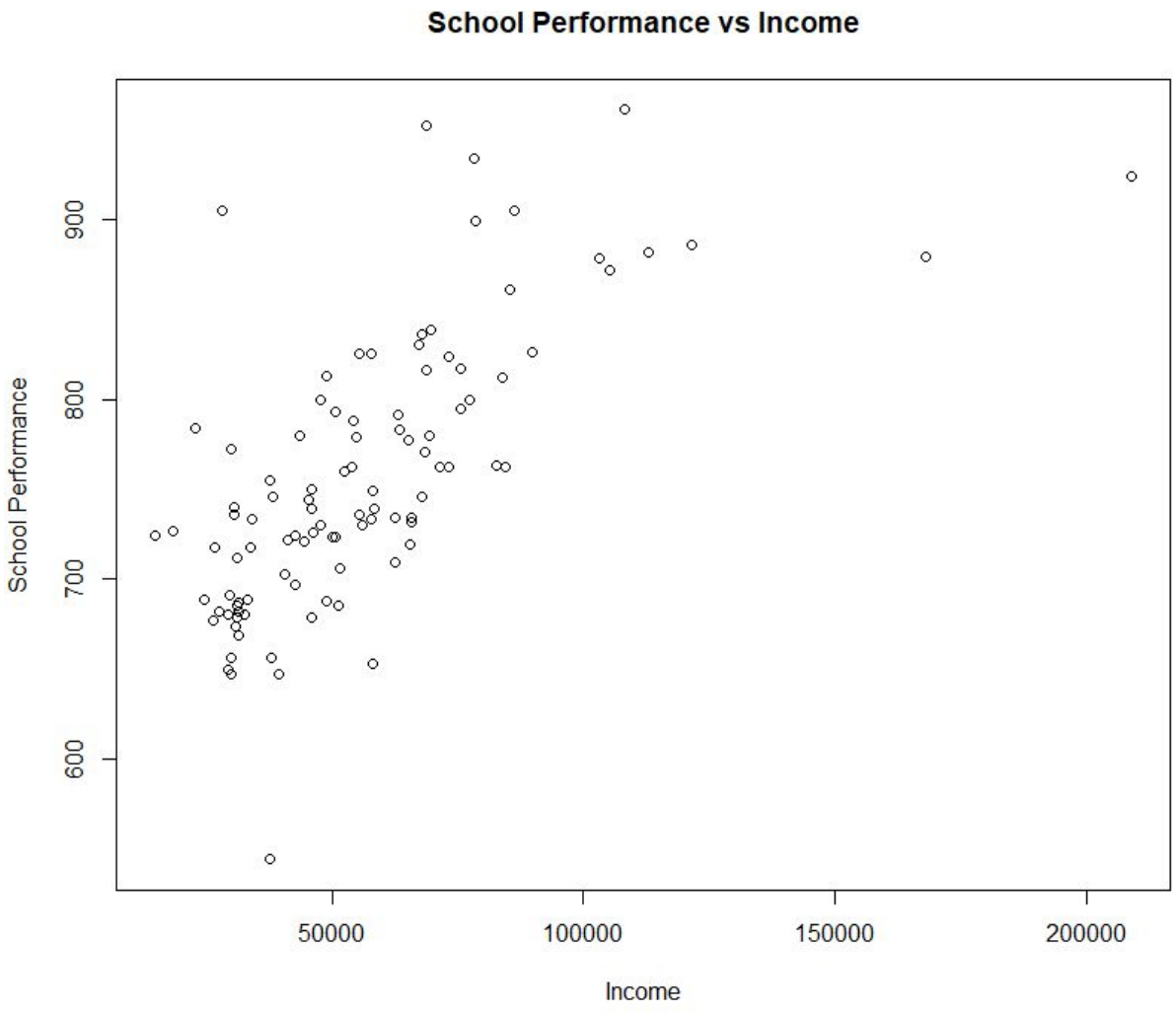
a)




```
> plot(x=LA$Longitude, y=LA$Latitude, ylab="Latitude", xlab="Longitude",
main="LA School Locations", pch=19, col="red")
> map("county", "california", add = TRUE)
```

b)

We can see through the scatterplot that the relationship between school performance is moderately linear, with a positive association between the two variables. This means that as one increases, the other generally increases as well and vice versa. There is a bit of variance within the data, which we can find with > cor(LA.subset$Schools,LA.subset$Income) which gives us [1] 0.6869965=r. THis means that r^2 is around .47, meaning 47% of the variance in performance is explained by income. However, since this is an observational study, we can only conclude correlation, with no implication of causation.




```
> LA.subset <- LA[LA$Schools>0,]
> plot(x=LA.subset$Income, y=LA.subset$Schools, xlab="Income",ylab="School
Performance",main="School Performance vs Income",pch=21,col="black")
```