Trent Bellinger

Stats 10 Lab 2

Exercise 1

Part a

```
flint <- read.csv('flint.csv')
```

Part b

```
dangerous_locs <- flint$Pb >= 15
mean(dangerous_locs)
```

```
[1] 0.04436229
```

Part c

```
north_reg <- flint$Region == 'North'
mean(flint$Cu[north_reg])
```

```
[1] 44.6424
```

Part d

```
mean(flint$Cu[dangerous_locs])
```

```
[1] 305.8333
```

Part e

```
mean(flint$Pb)
```

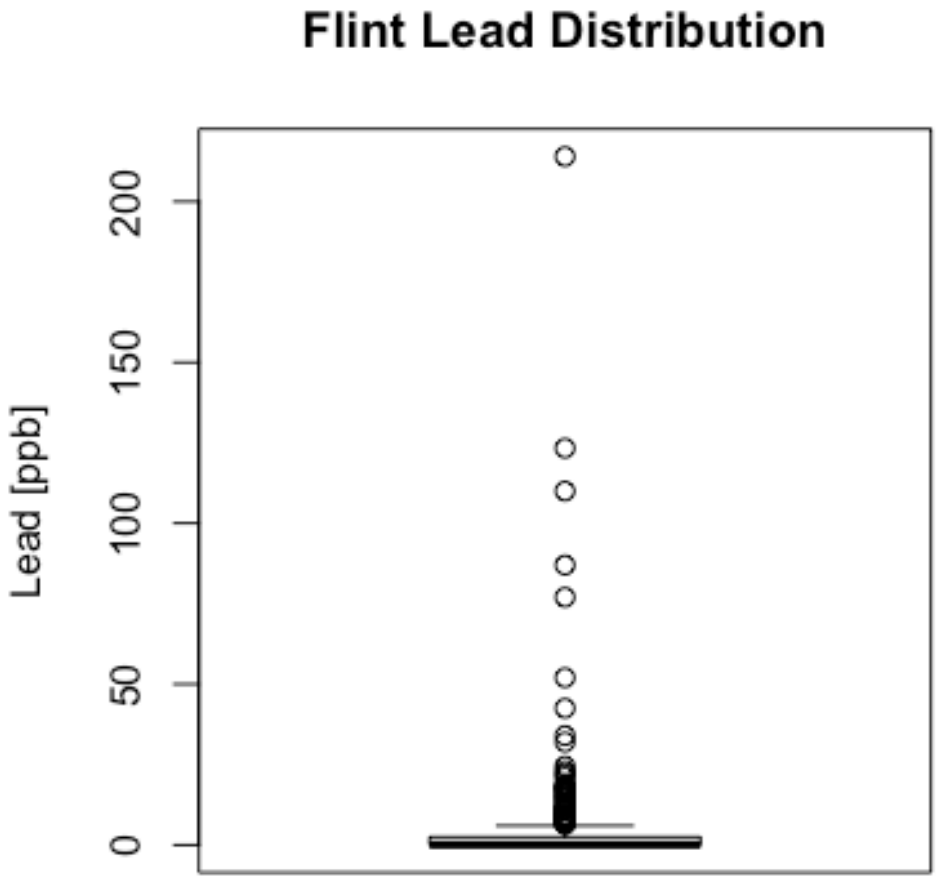
```
[1] 3.383272
```

```
mean(flint$Cu)
```

```
[1] 54.58102
```

Part f

boxplot(flint$Pb, main = 'Flint Lead Distribution', ylab = 'Lead [ppb]')



Part g

No, the mean does not seem to be a good measure of central tendency. There are outliers in the data and the data has a significant right skew. A better measure of central tendency would be the median.

```
median(flint$Pb)
```

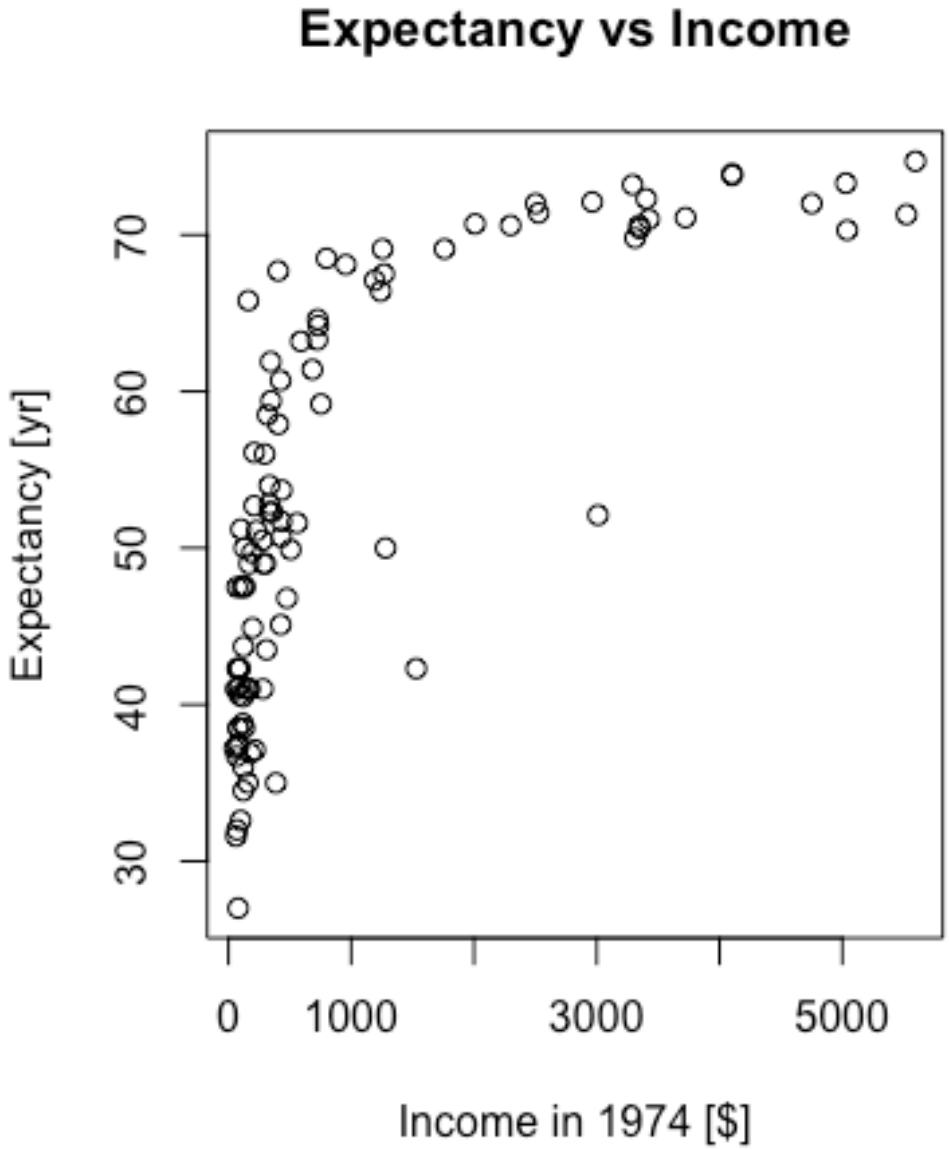
```
[1] 0
```

Exercise 2

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

Part a

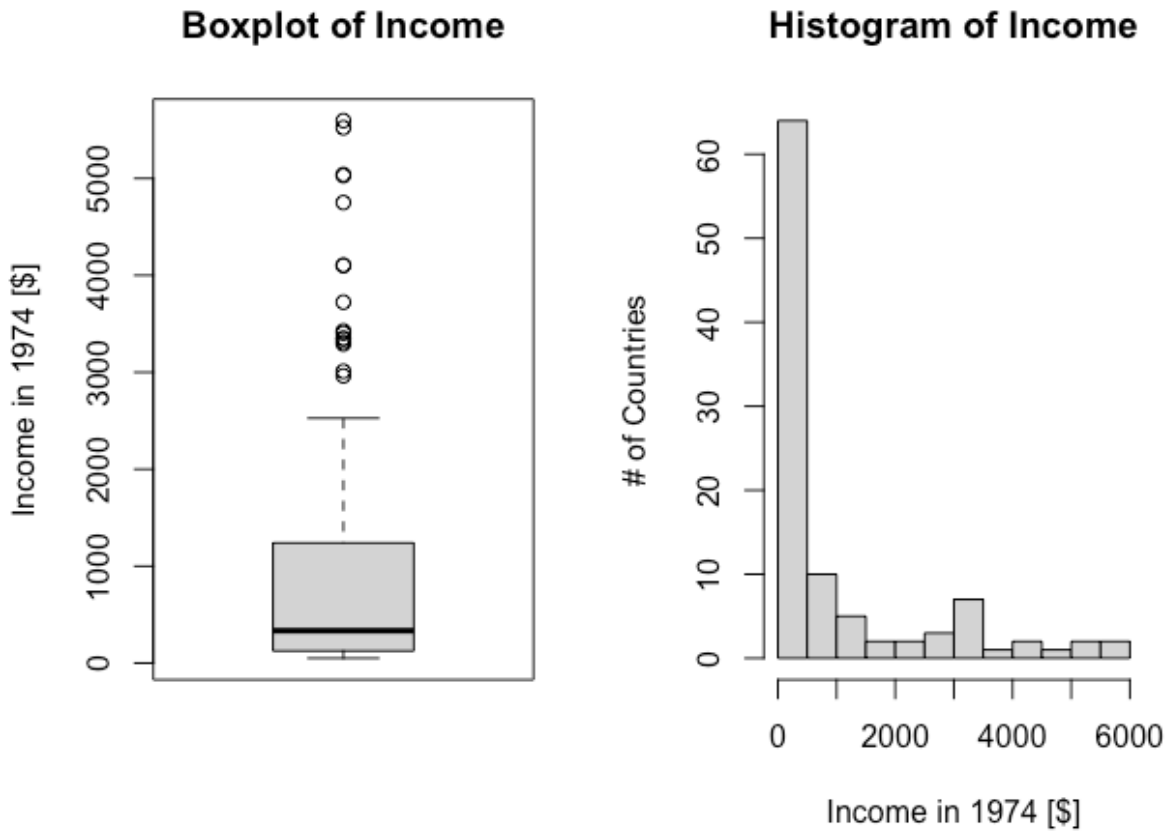
```
plot(Life~Income, data = life, xlab = 'Income in 1974 [$]', ylab = 'Expectancy [yr]',
    main = 'Expectancy vs Income')
```



So there is an increase in life expectancy as income increases. However, this increase tapers off as income further increases.

Part b

```
par(mfrow=c(1,2))
boxplot(life$Income, ylab = 'Income in 1974 [$]', main = 'Boxplot of Income')
hist(life$Income, xlab = 'Income in 1974 [$]', ylab = '# of Countries',
    main = 'Histogram of Income')
```



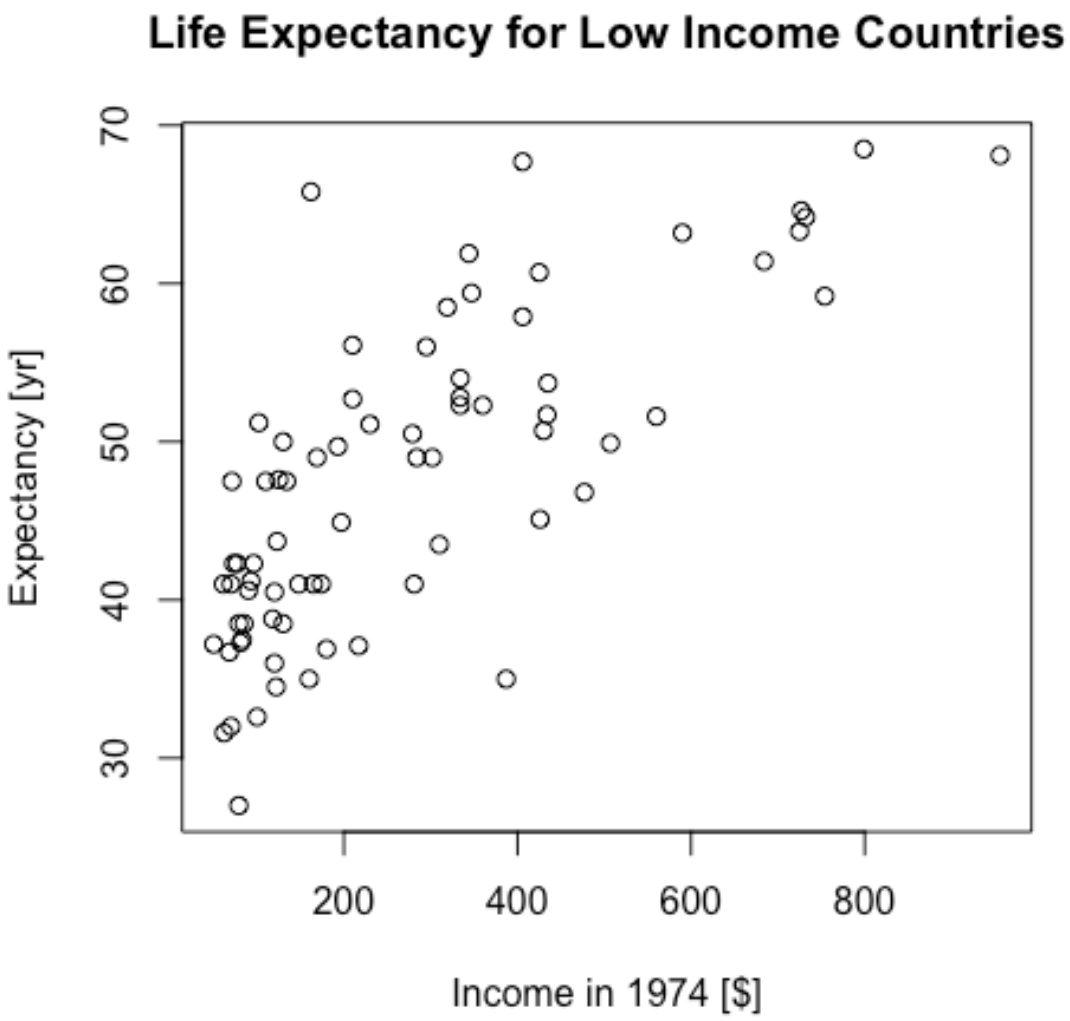
Yes there are many outliers in the dataset as shown by the boxplot. Furthermore our data is right skewed, which means the histogram tail goes off to the right.

Part c

```
life_gte1000 = life[life$Income >= 1000,]
life_lt1000 = life[life$Income < 1000,]
```

Part d

```
plot(life_lt1000$Income, life_lt1000$Life, xlab = 'Income in 1974 [$]', ylab = 'Expectancy [yr]',
     main = 'Life Expectancy for Low Income Countries')
```



cor(life_lt1000$Income, life_lt1000$Life)

[1] 0.752886

The correlation between income and life for countries with less than $1000 per capita income is about 0.752.

Exercise 3

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

Part a

summary(maas$lead)

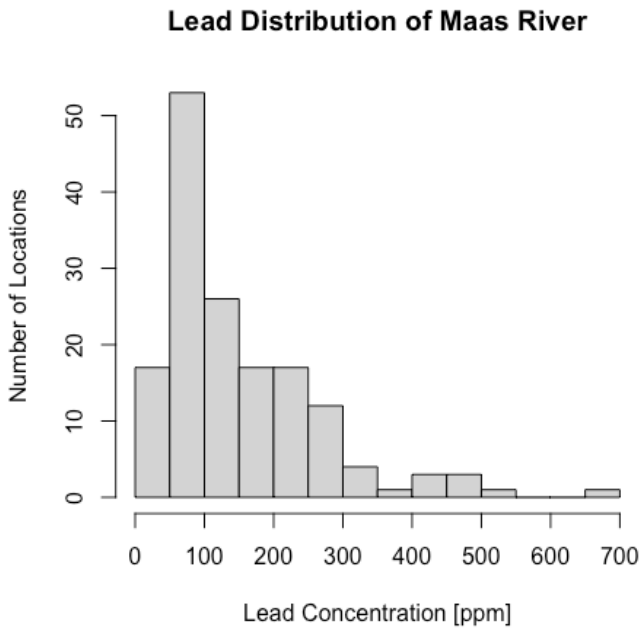
```
   Min.   1st Qu.  Median   Mean   3rd Qu.   Max.
   37.0    72.5     123.0     153.4    207.0     654.0
```

summary(maas$zinc)

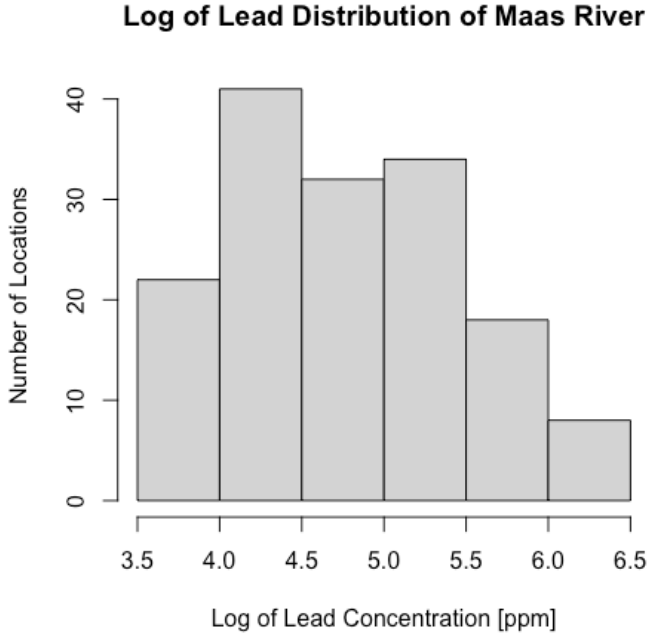
```
   Min.   1st Qu.  Median  Mean   3rd Qu.   Max.
  113.0   198.0    326.0    469.7    674.5     1839.0
```

Part b

```
hist(maas$lead, xlab = 'Lead Concentration [ppm]', ylab = 'Number of Locations',
    main = 'Lead Distribution of Maas River')
```

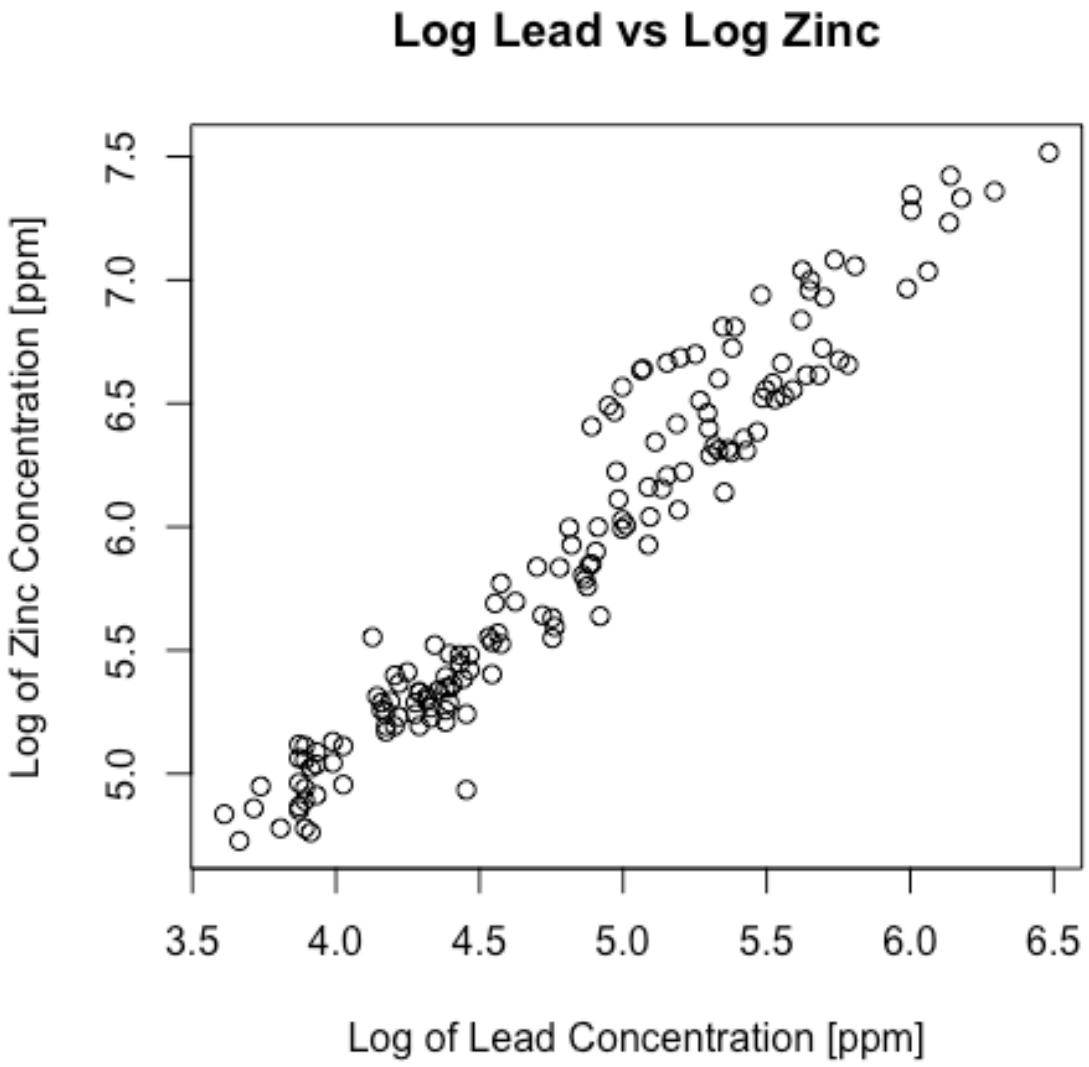



```
hist(log(maas$lead), xlab = 'Log of Lead Concentration [ppm]', ylab = 'Number of Locations',
    main = 'Log of Lead Distribution of Maas River')
```



Part c

```
plot(log(maas$lead), log(maas$zinc), xlab = 'Log of Lead Concentration [ppm]',
    ylab = 'Log of Zinc Concentration [ppm]', main = 'Log Lead vs Log Zinc')
```

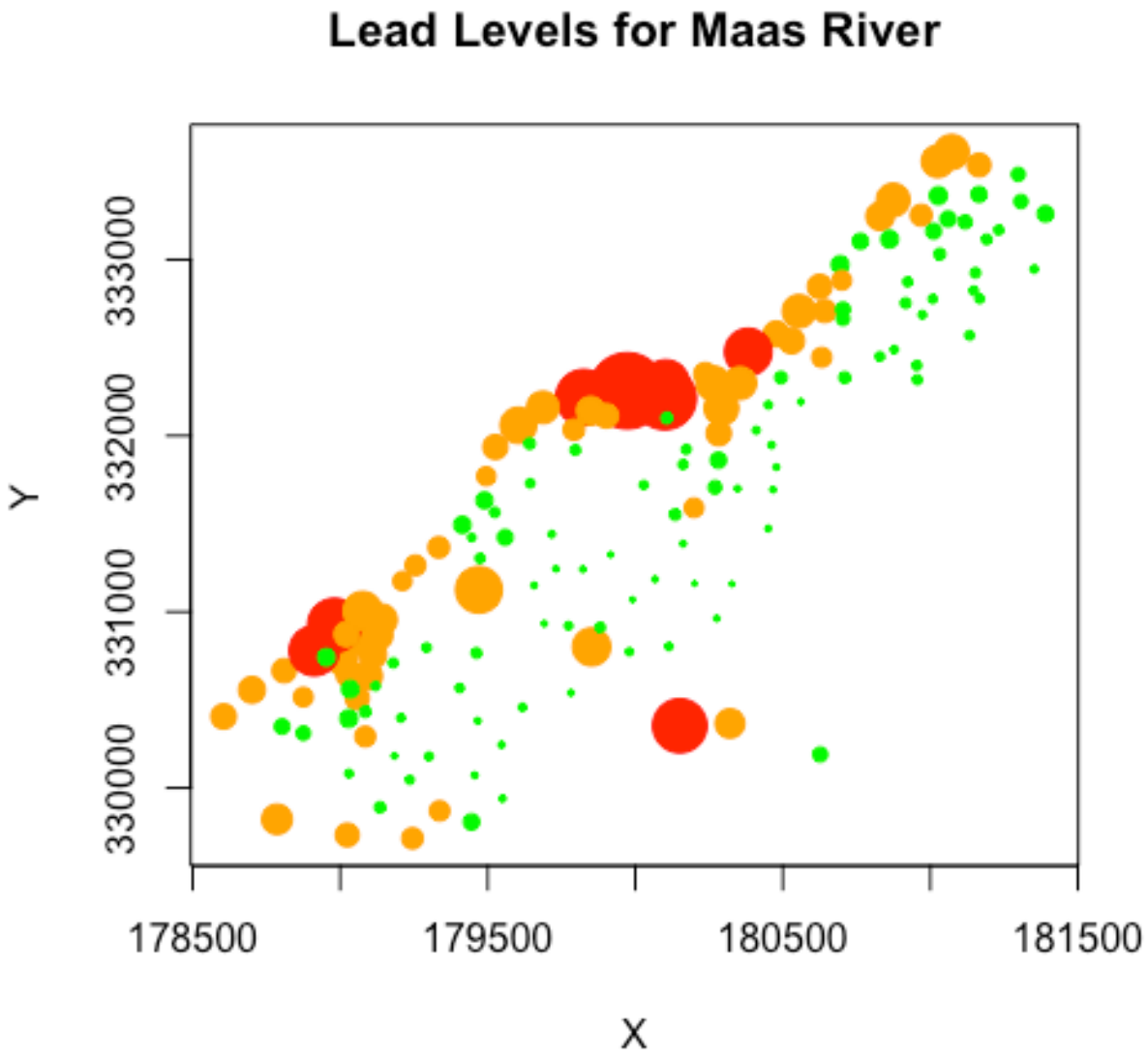


We see a strong positive relationship between log lead and log zinc. This means there is not much variation across the line of best fit for Log Lead vs Log Zinc.

Part d

```
lead_colors <- c('green', 'orange', 'red')
lead_levels <- cut( maas$lead, c(0, 150, 400, max(maas$lead) + 1) )
```

```
plot(maas$x, maas$y, cex = maas$lead / mean(maas$lead),
     col =lead_colors[as.numeric(lead_levels)],
     main = 'Lead Levels for Maas River', pch = 19, xlab = 'X', ylab = 'Y')
```

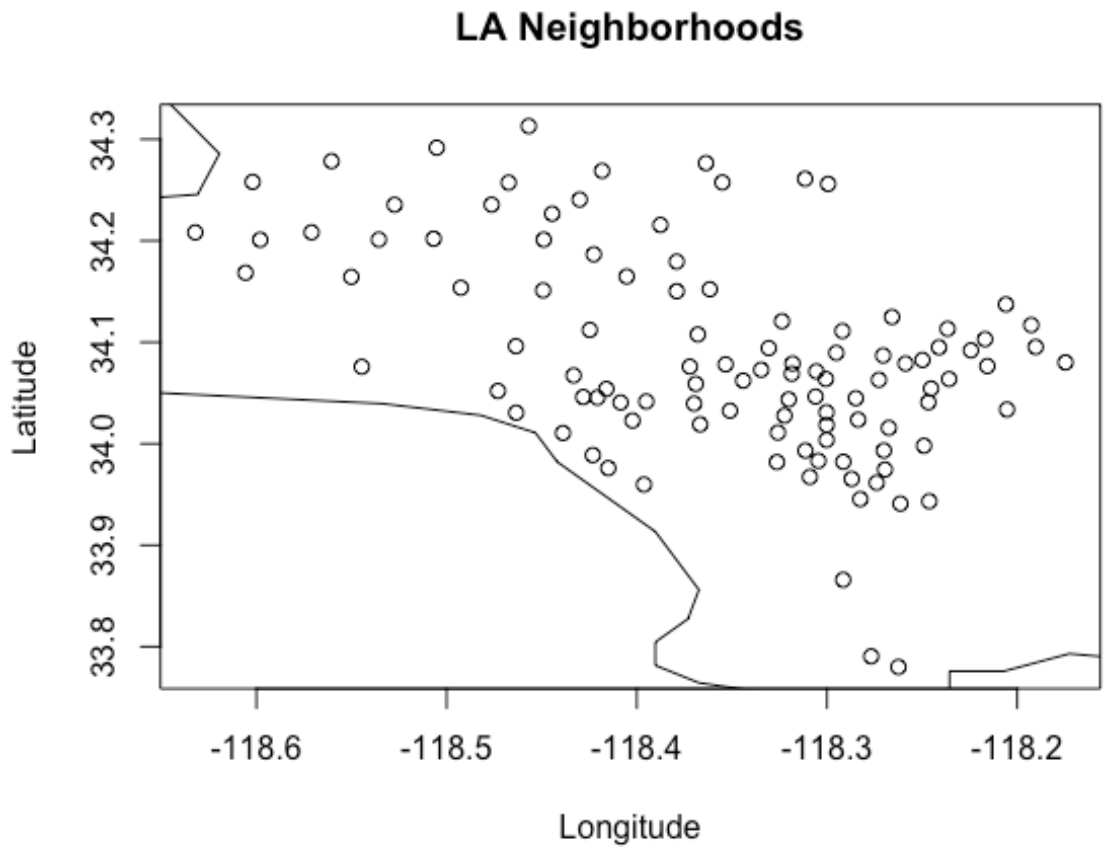


Exercise 4

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

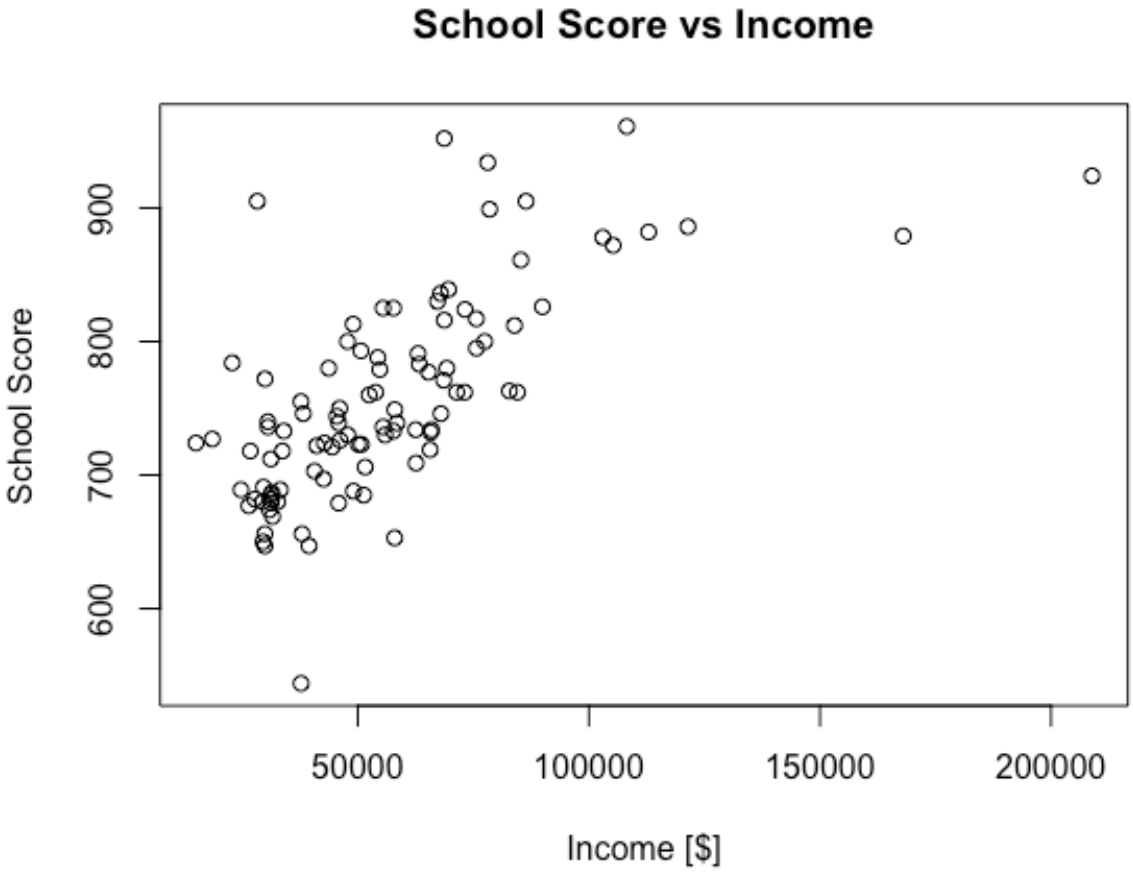
Part a

```
plot(LA$Longitude, LA$Latitude, xlab = 'Longitude', ylab = 'Latitude',
    main = 'LA Neighborhoods')
map("county", "california", add = TRUE)
```



Part b

```
LA_Schools <- LA[LA$Schools != 0,]
plot(LA_Schools$Income, LA_Schools$Schools, ylab = 'School Score', xlab = 'Income [$]',
    main = 'School Score vs Income')
```



We see a positive relationship between resident income and school performance in LA neighborhoods. In particular, there is a stronger linear relationship between these two variables for incomes ranging between 50,000 and 100,000 dollars.