Stats 12 Lab 2 Submission

Name: Brian Tehrani

UID: 604715464

Section 1

a)

```
> #a read in flint water data from csv file
> flint <- read.csv(file = "flint.csv")
```

b)

```
> #b proportion of lead that is greater than 15 PPB
> mean(flint$Pb >= 15)
[1] 0.04436229
```

c)

```
> #c mean Cu lvls for only Northern region
> mean(flint$Cu[flint$Region == "North"])
[1] 44.6424
```

d)

```
> #d mean Cu lvl for only dangerous conc.
> mean(flint$Cu[flint$Danger_Cu == "Y"])
[1] 330.6597
```

e)

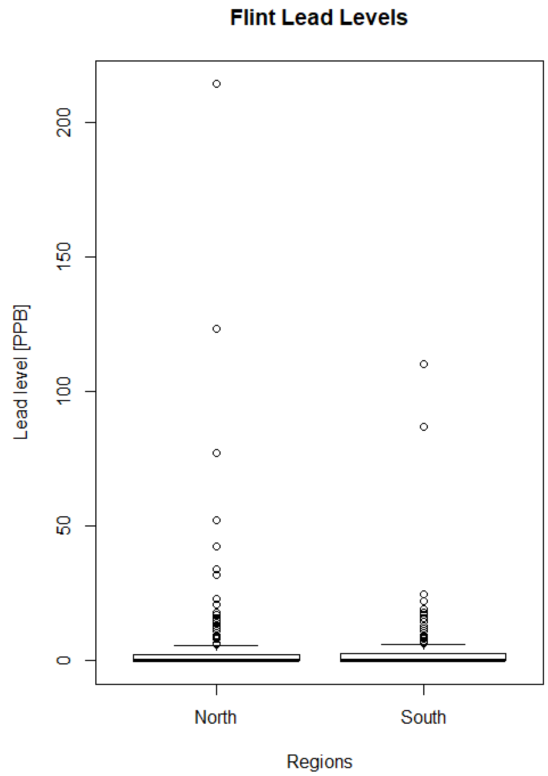
```
> #e Pb observations above 0 & Below  Danger lvl
> mean(flint$Pb < 15 & flint$Pb > 0)
[1] 0.4011091
```

f)

```
> #f Mean Pb and Cu lvls
> mean(flint$Pb)
[1] 3.383272
> mean(flint$Cu)
[1] 54.58102
```

g)

```
> #g Boxplot of Pb lvls
> flint_Pb_North <- flint$Pb[flint$Region == "North"]
> flint_Pb_South <- flint$Pb[flint$Region == "South"]
> boxplot(flint_Pb_North,flint_Pb_South, main = "Flint Lead Levels",
xlab = "Regions",
+         ylab = "Lead level [PPB]", names = c("North", "South"))
```

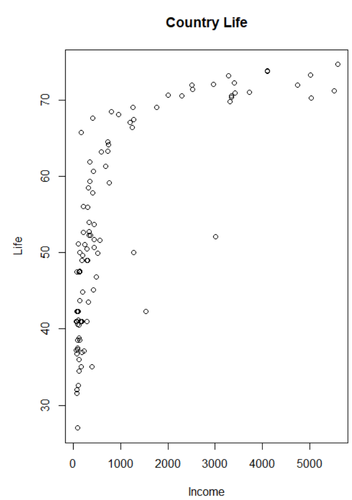


h) The above boxplot shows a skewed right distribution and thus the mean would not be a good measure of center for the data. A more useful statistic for the data would be the median.

Section 2

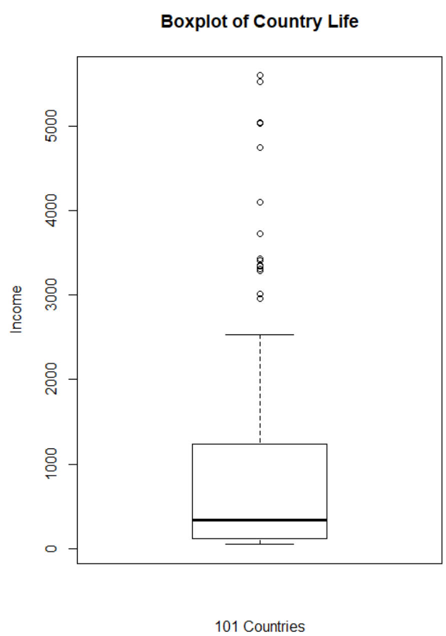
a)

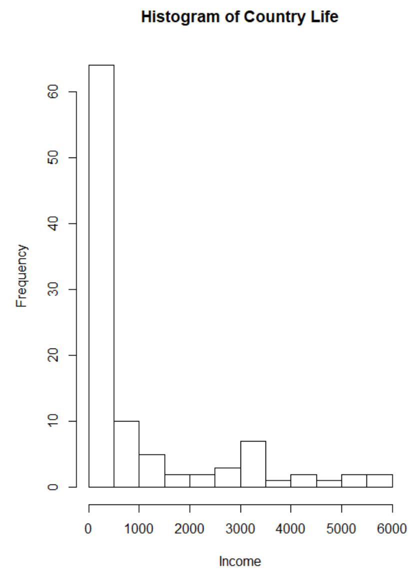
```
> plot(life$Income, life$Life, ylab = "Life", xlab = "Income", main = "Country Life")
> # When income increases, Life expectancy increases
```



There is a strong relationship that increasing income increases life Expectancy b)

```
> boxplot(life$Income, main = "Boxplot of Country Life", ylab = "Income",
+   xlab="101 Countries")
> hist(life$Income,main = "Histogram of Country Life", xlab = "Income")
> #Are there any outliers? => yes
```





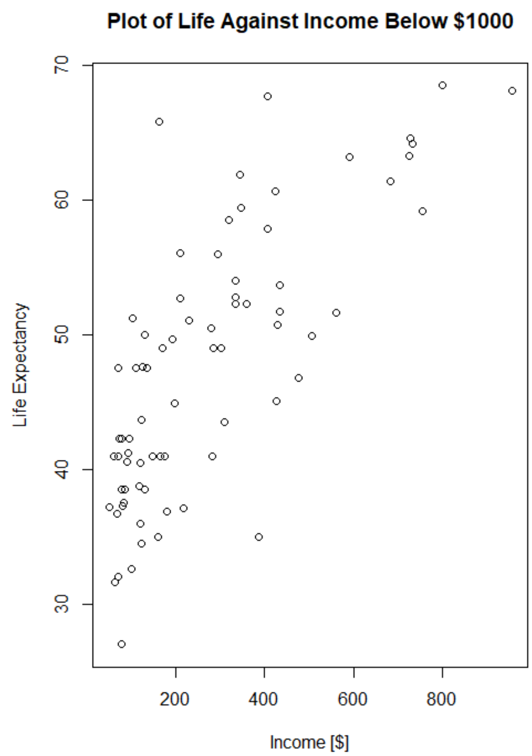
The boxplot of the incomes shows that there are outliers.

c)

```
> income_more_1000 <- life$Income[life$Income >= 1000]
> income_less_1000 <- life$Income[life$Income < 1000]
```

d)

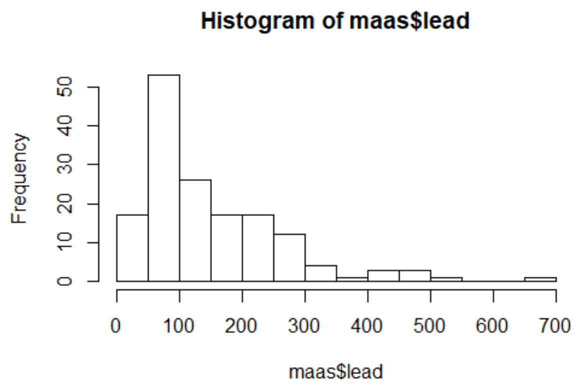
```
> life_less_1000 <- life$Life[life$Income < 1000]
> plot(income_less_1000,life_less_1000, main = "Plot of Life Against Income Below $1000
"
+      , ylab = "Life Expectancy", xlab = "Income [$]")
> cor(income_less_1000,life_less_1000)
[1] 0.752886
```

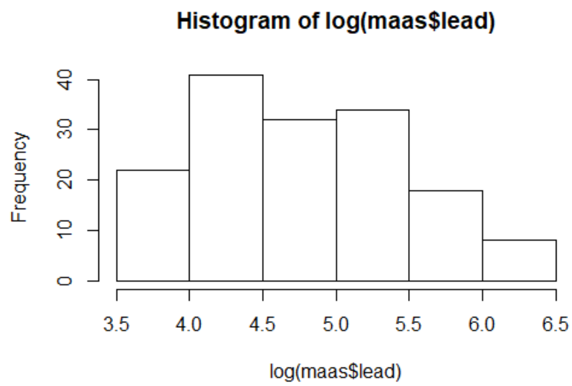


Section 3

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
> IQR(maas$zinc)
[1] 476.5
```

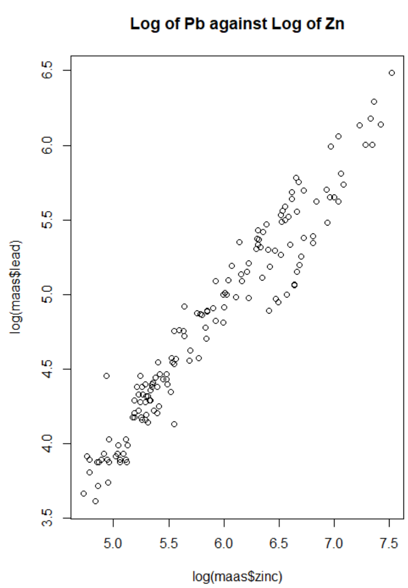
b)
```
> par(mfcol=c(2,1))
> hist(maas$lead)
> hist(log(maas$lead))
```





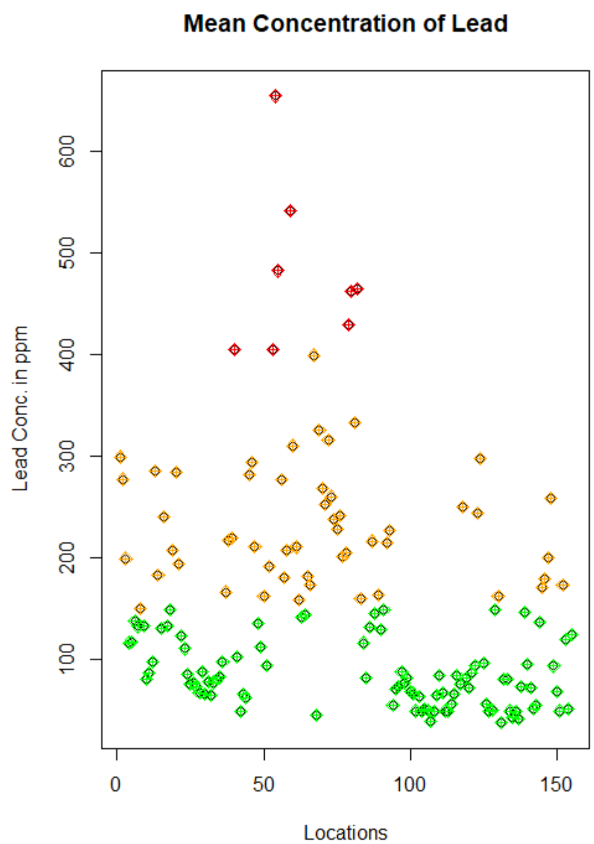
c)

```
> #plot of logs of lead against zinc
> plot(log(maas$zinc), log(maas$lead), main = "Log of Pb against Log of Zn")
> # there is a positive correlation between the two variables
```



The relationship between the two plots shows a near strong positive relationship between the two variables d)

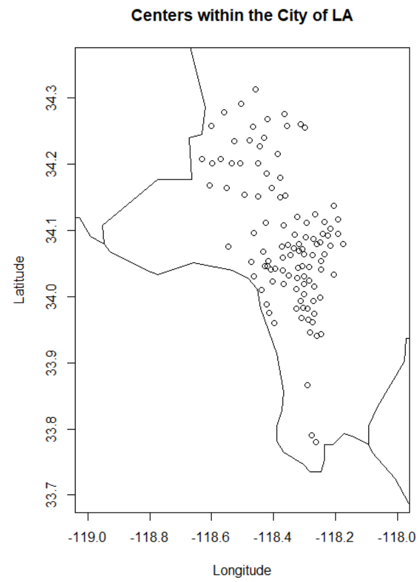
```
> risk_colors <- c("green", "orange", "red")
> risk_lvl <- cut(maas$lead,c( 0, 149, 399, 100000))
> as.numeric(risk_lvl)
  [1] 2 2 2 1 1 1 1 2 1 1 1 1 2 2 1 2 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 3
 [41] 1 1 1 1 2 2 2 1 1 2 1 2 3 3 3 2 2 2 3 2 2 2 1 1 2 2 2 1 2 2 2 2 2 2 2 2
2 2 3 3
 [81] 2 3 2 1 1 1 2 1 2 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 2 1 1
[121] 1 1 2 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 1 1 1 2 1 1 1
> plot(maas$lead, main = "Mean Concentration of Lead", ylab = "Lead Conc. in
ppm",
+       xlab = "Locations")
> points(maas$lead, col=risk_colors[as.numeric(risk_lvl)], pch=9)
```



Section 4

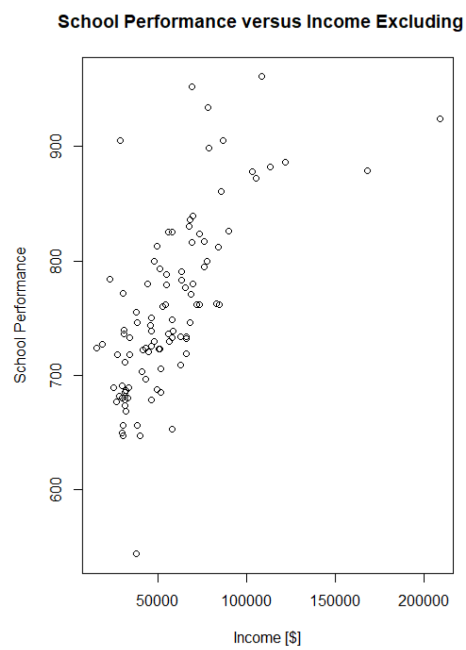
a)

```
> LA <- read.table("http://www.stat.ucla.edu/~nchristo/statistics12/la_data.txt", heade
r=TRUE)
> library(maps)
> plot(LA$Longitude, LA$Latitude, ylim = c(33.7, 34.35), xlim = c(-119, -118),
+ xlab = "Longitude", ylab = "Latitude", main = "Centers within the City of LA")
> map("county", "california", add = TRUE)
```



b)

```
> LA_Schools_not_Zero <- LA$Income[LA$Schools != 0]
> LA_Income_Schools_not_zero <- LA$Schools[LA$Schools !=0]
> plot(LA_Schools_not_Zero, LA_Income_Schools_not_zero, main = "School
Performance versus Income Excluding 0",
+       xlab = "Income [$]", ylab = "School Performance")
```



There is a positive relationship that as income increases, school performance increases.

c)

```
> mean_LA_Income <- mean(LA$Income)
> sd_LA_Income = sd(LA$Income)
> z <- (100000-mean_LA_Income)/ sd_LA_Income
> z
[1] 1.35168
```

d)
```
> total <- length(LA$Income)
>
> # Function ... nSD is the number of SD you are looking at
> pData <- function(nSD){
+   lo = mean_LA_Income - nSD/2*sd_LA_Income
+   hi = mean_LA_Income + nSD/2*sd_LA_Income
+   percent = sum(LA$Income>=lo & LA$Income<=hi)/total *100
+ }
> print(paste("Percent of data within 1 SD is ",pData(1),"%", sep=""))
[1] "Percent of data within 1 SD is 46.7289719626168%"
> print(paste("Percent of data within 2 SD is ",pData(2),"%", sep=""))
[1] "Percent of data within 2 SD is 84.1121495327103%"
> print(paste("Percent of data within 3 SD is ",pData(3),"%", sep=""))
[1] "Percent of data within 3 SD is 92.5233644859813%"
```

The empirical rule states that the distribution of data in a standard normal distribution fits the 68-95-99.7 rule where 68% of the data falls within 1 standard deviation from the mean, and 95% and 99.7% od the data falls within 2 and 3 standard deviations away from the mean respectively. This data does not follow the trend of the empirical rule as the data distributes about 46.7-84.1-92.5.