Part I

1a)

```
> flint <- read.csv("https://ucla.box.com/shared/static/e9xuft4h3p8fdi4ydoj2hhujee0vmopb.csv")
>
```

1b)

```
> mean(flint$Pb >= 15)
[1] 0.04436229
```

1c)

```
> copper_north = flint$Cu[flint$Region == "North"]
> mean(copper_north)
[1] 44.6424
```

1d)

```
> dangerous_copper = flint$Cu[flint$Pb >= 15]
> mean(dangerous_copper)
[1] 305.8333
```

1e)

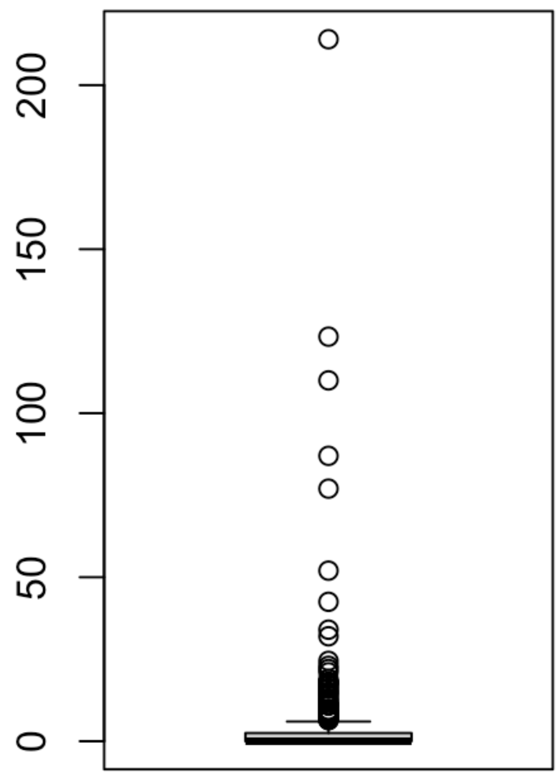
```
> mean(flint$Pb)
[1] 3.383272
> mean(flint$Cu)
[1] 54.58102
```

1f)

Command: boxplot(flint$Pb, main = "Flint Lead Levels (in PPB)")

Output:

Flint Lead Levels (in PPB)

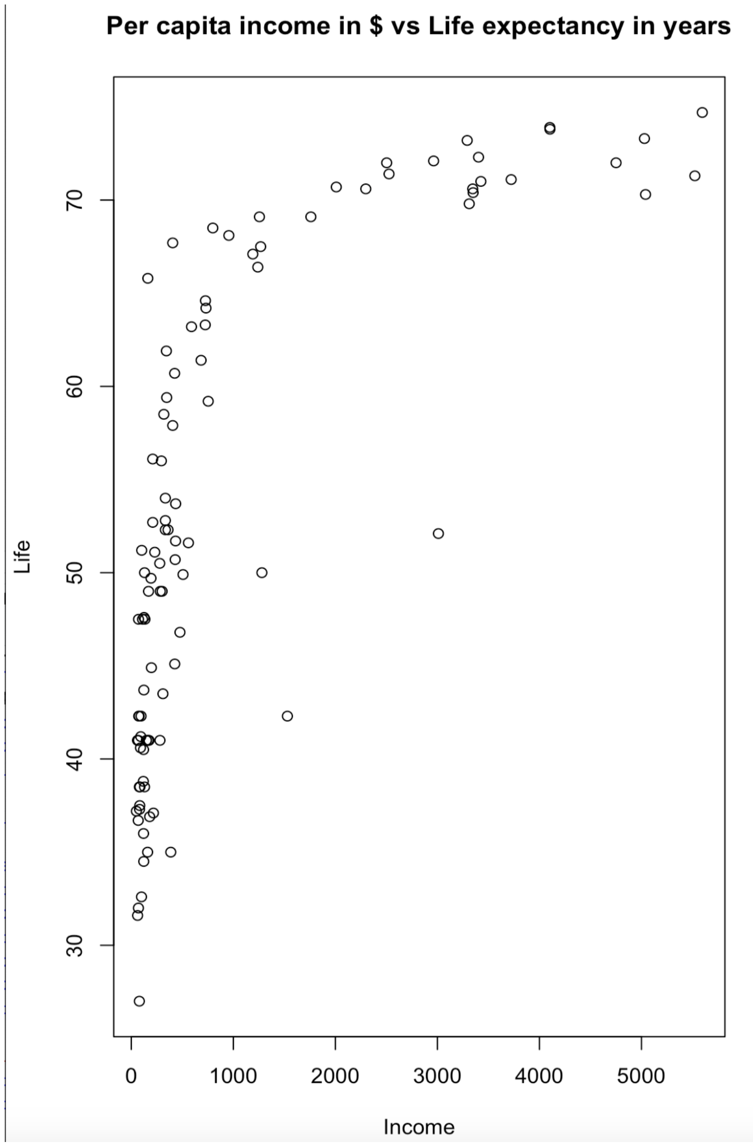


1g) The mean is not a good measure of center because due to the number of high outliers this data is skewed to the right. Therefore, we should use median as a measure of center instead.

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

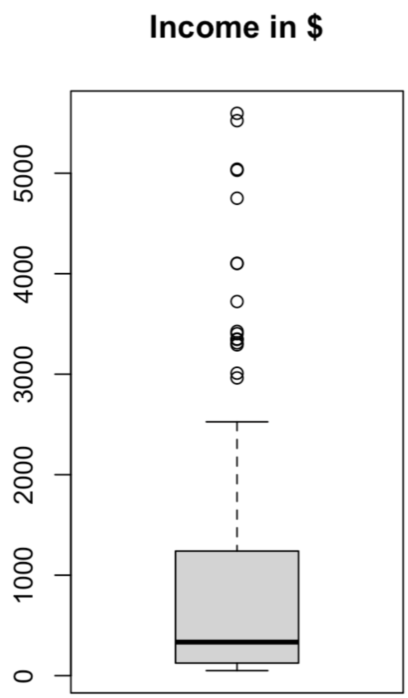
2a)

Command: plot(Life ~ Income, data = life, main = "Per capita income in $ vs Life expectancy in years")



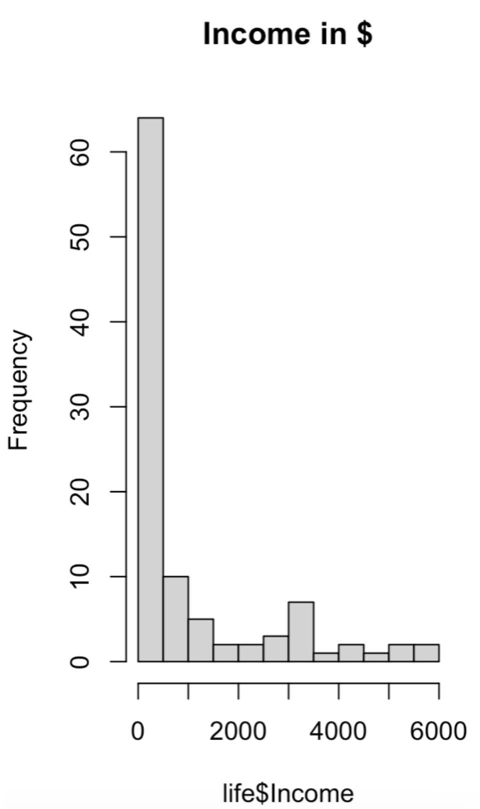
There seems to be a positive correlation between life and income - an increase in income generally corresponds to an increase in life expectancy.

2b)
Command: boxplot(life$Income, main = "Income in $")



There are multiple outliers on the high side.

Command: hist(life$Income, main = "Income in $")

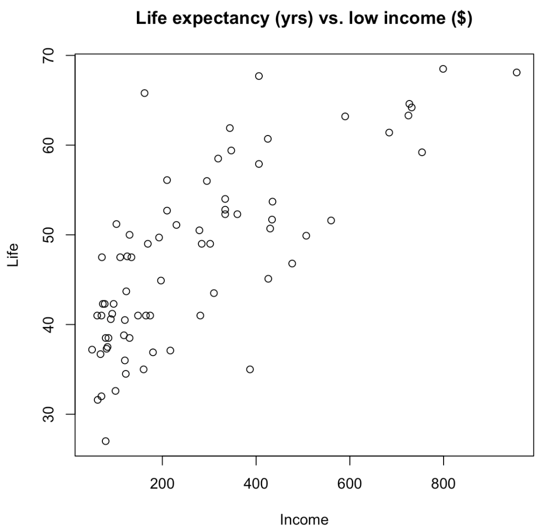


2c)

```
> below1000 <- life[life$Income < 1000,]
> atLeast1000 <- life[life$Income >= 1000,]
```

2d)

Command: plot(Life ~ Income, data = below1000, main = "Life expectancy (yrs) vs. low income ($)")



2e)

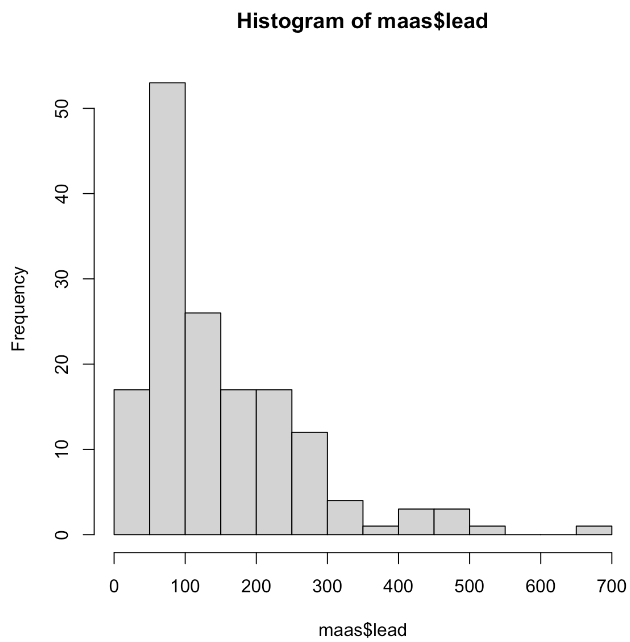
```
> cor(below1000$Life, below1000$Income)
[1] 0.752886
```

3a)

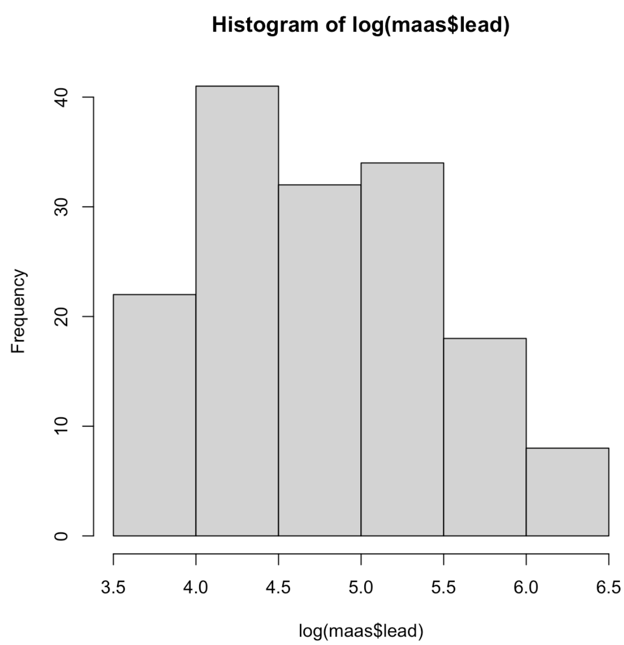
```
> summary(maas$lead)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   37.0    72.5   123.0   153.4   207.0   654.0 
> summary(maas$zinc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  113.0   198.0   326.0   469.7   674.5  1839.0 
```

3b)

Command: hist(maas$lead)

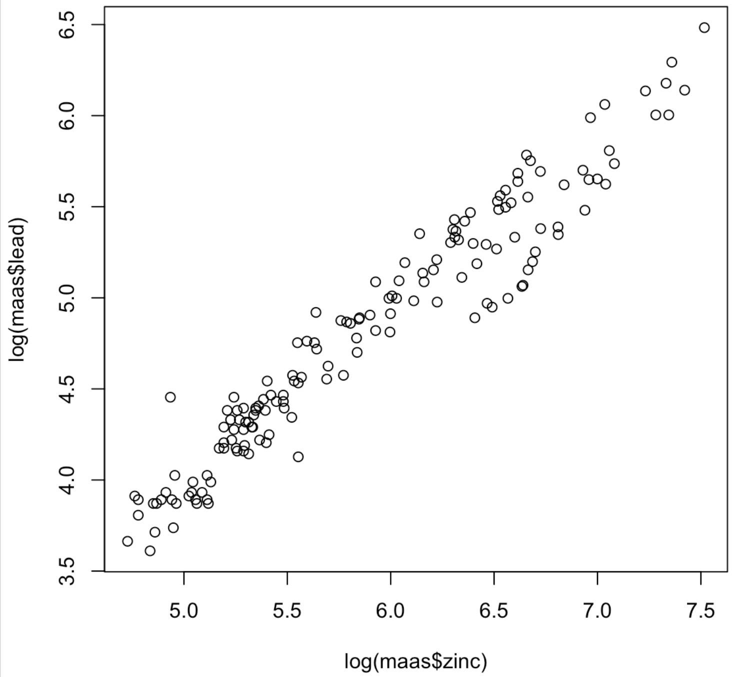


Command: hist(log(maas$lead))



3c)

Command: plot(log(maas$zinc), log(maas$lead))



There is a strong positive linear relationship between the two variables.

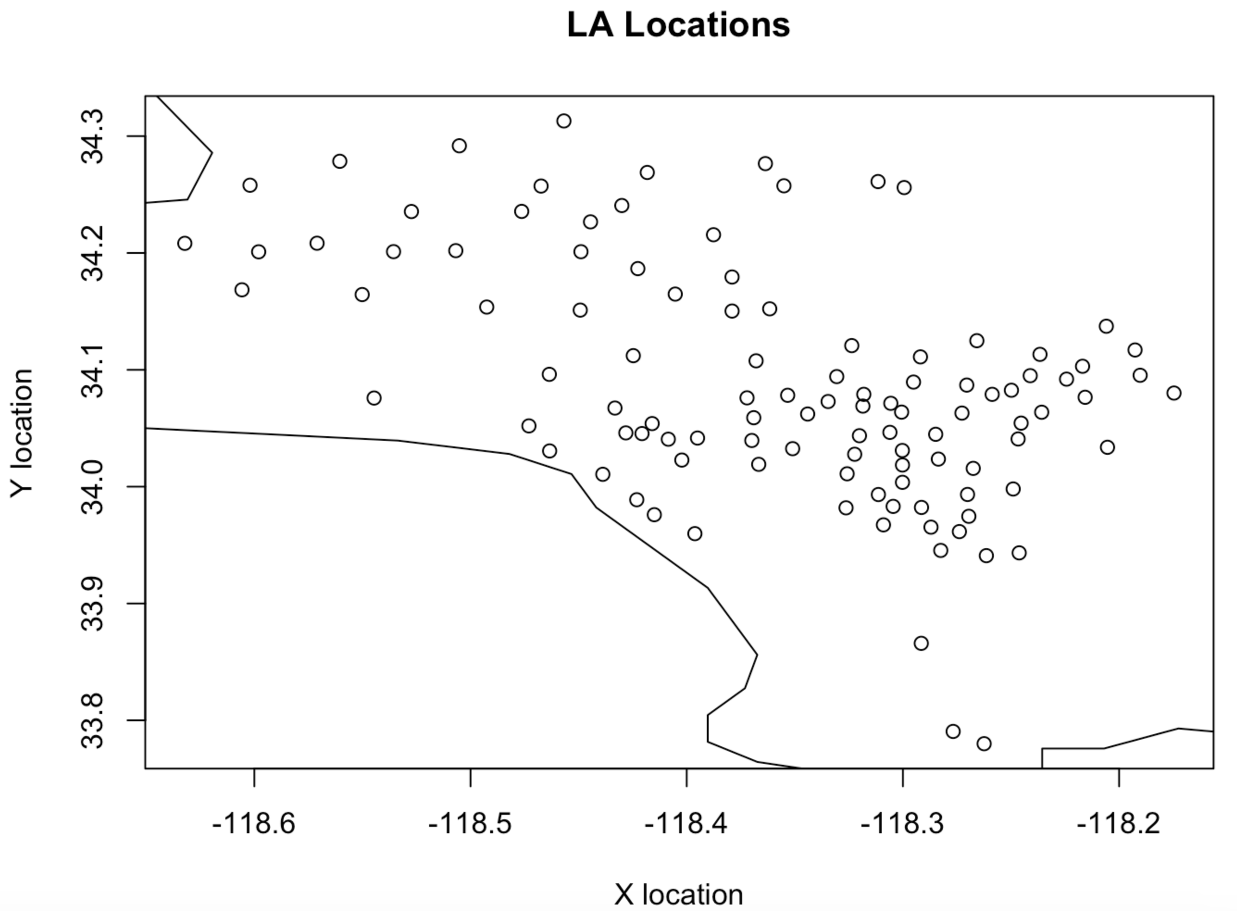
3d)
Commands:

```
leadMaas = maas$lead
safetyLevels = cut(leadMaas, c(0,150,400,700), labels = c("Lead-Free", "Lead-Safe", "Hazard"))
maas$safety = safetyLevels

plot(maas$y ~ maas$x, col = maas$safety, cex = 1, pch = 1, xlab = "X location", ylab = "Y
location", main = "Locations of lead safety")
```

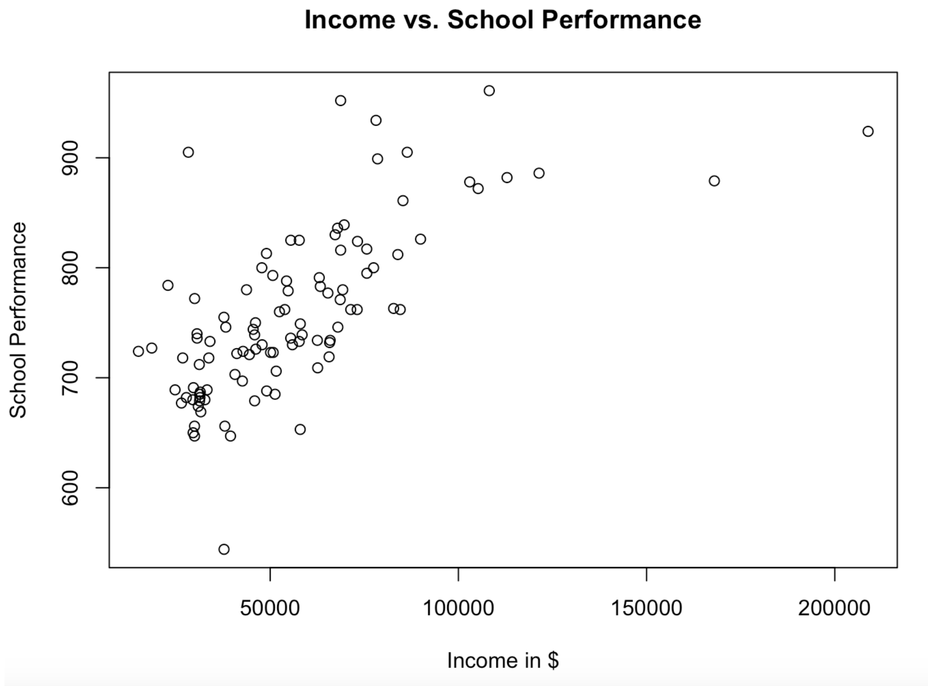
4a)
Commands:

library(maps)
LA <- read.table("https://ucla.box.com/shared/static/d189x2gn5xfmcic0dmnhj2cw94jwvqpa.txt", header=TRUE)
plot(LA$Longitude, LA$Latitude, pch = 1, xlab = "X location", ylab = "Y location", main = "LA Locations")
map("county", "california", add = TRUE)



4b)
Commands:

LAwithSchools = LA[LA$Schools != 0,]
plot(LAwithSchools$Income, LAwithSchools$Schools, xlab = "Income in $", ylab = "School Performance", main = "Income vs. School Performance")



> cor(LAwithSchools$Schools, LAwithSchools$Income)
[1] 0.6869965

Since the correlation coefficient between income and school performance is 0.69, there is a relatively strong positive linear (by glancing at the data) correlation between income and school performance. That means that an increase in income typically leads to higher school performance.

5a)

```
> summary(customer_data)
   cust_id               age           gender               income         education         marital_status
 Length:100         Min.   :20.00   Length:100         Min.   : 23798   Length:100         Length:100
 Class :character   1st Qu.:32.00   Class :character   1st Qu.: 55320   Class :character   Class :character
 Mode  :character   Median :44.00   Mode  :character   Median : 99637   Mode  :character   Mode  :character
                    Mean   :44.99                      Mean   :103425
                    3rd Qu.:56.75                      3rd Qu.:150030
                    Max.   :70.00                      Max.   :198808
                    NA's   :10                         NA's   :5
  purchase_amt
 Min.   : 72.0
 1st Qu.:211.0
 Median :325.0
 Mean   :356.2
 3rd Qu.:466.0
 Max.   :791.0
 NA's   :7
```

There are 10 missing values for age, 5 missing values for income, and 7 missing values for purchase amount.

5b)
Customer ID - categorical, Age - numerical, Gender - categorical, Income - numerical, Education - categorical, Marital status - categorical, Purchase amount - numerical

We can transform gender, education, and marital status into numerical data types in order to get a more accurate summary of the data values.

After running
customer_data$gender = as.factor(customer_data$gender)
customer_data$education = as.factor(customer_data$education)
customer_data$marital_status = as.factor(customer_data$marital_status)

summary(customer_data)

We get a much more accurate summary.

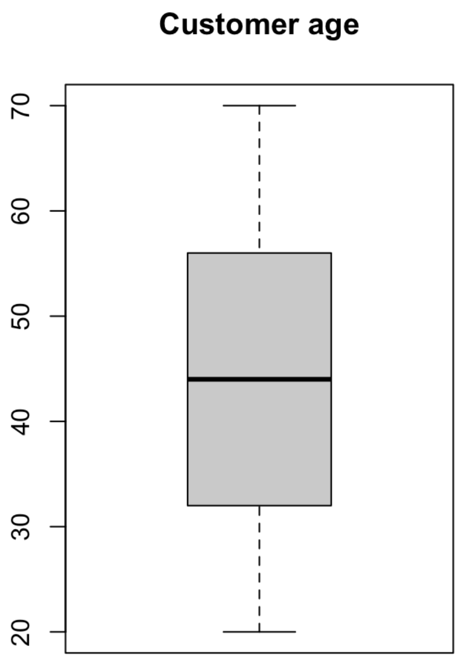
```
   cust_id               age          gender      income                education    marital_status  purchase_amt
 Length:100         Min.   :20.00   F:55   Min.   : 23798   college degree :25   divorced:30     Min.   : 72.0
 Class :character   1st Qu.:32.00   M:45   1st Qu.: 55320   graduate degree:31   married :25     1st Qu.:211.0
 Mode  :character   Median :44.00          Median : 99637   high school    :18   single  :18     Median :325.0
                    Mean   :44.99          Mean   :103425   some college   :26   widowed :27     Mean   :356.2
                    3rd Qu.:56.75          3rd Qu.:150030                                        3rd Qu.:466.0
                    Max.   :70.00          Max.   :198808                                        Max.   :791.0
                    NA's   :10             NA's   :5                                             NA's   :7
```

5c)
Let’s clean up our data and remove NA values first with

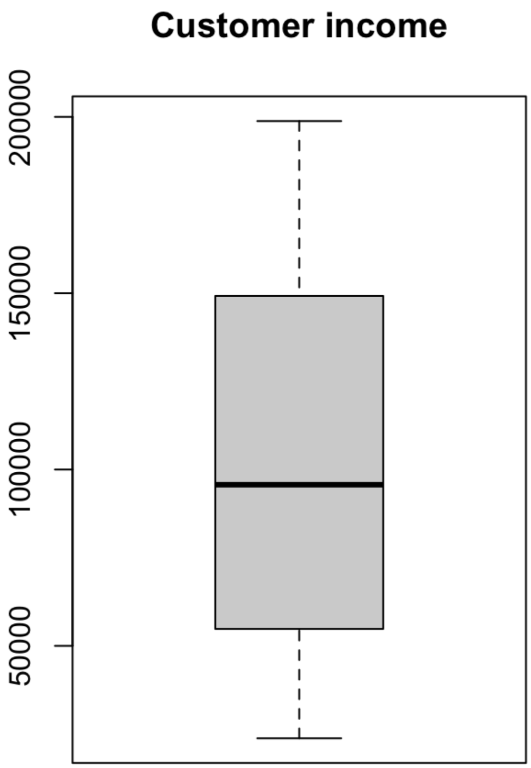
customer_data = na.omit(customer_data)

then run

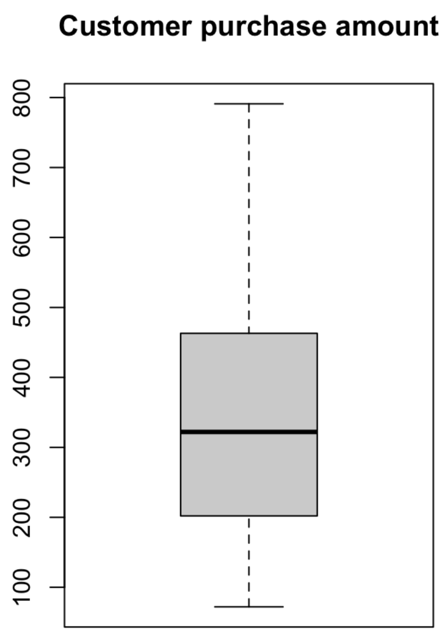
boxplot(customer_data$age, main = "Customer age")
boxplot(customer_data$income, main = "Customer income")
boxplot(customer_data$purchase_amt, main = "Customer purchase amount")



There are no outliers in age.



There are no outliers in income.



There are no outliers in purchase amount.

Part 2

1a) In general, students are happy with their body weight. Of the 900 sampled, 310+290 = 600 said they felt about right. 600/900 is ⅔, indicating a supermajority of the students are happy with their body weight.

1b) The best graph would be a grouped bar chart. This is because both gender and body image are categorical variables, and bar charts are appropriate for categorical variables.

1c) There are 310+130+30 or 470 women in the sample. 310/470, or 66% of women feel about right. There are 290+68+72 = 430 men in the sample. 290/430, or 67% of men feel about right. Therefore, female students are very slightly less likely to feel right than male students.

1d) 130/470 or 28% of women feel they are overweight, compared to 68/430 or 16% of men. 30/470 or 6% of women feel they are underweight, compared to 72/430 or 17% of men. For women who don't feel right, they are much more likely to feel overweight compared to underweight. On the other hand, for men who don't feel right, they are much more likely to feel underweight compared to overweight.

2a) There exists a positive, linear, and moderately strong relationship between family income and the percentage of the population with a college degree. However, there seems to be an outlier at (30,60). In general, we can expect that the higher the percentage of the population with a college degree is, the higher the median family income will be.

2b) There exists a somewhat positive, nonlinear, and weak relationship between average amount of fuel used and the speed at which the car is driven. There also seems to be an outlier at (10,22). Although there seems to be somewhat of a positive relationship, this is not meaningful given that the relationship is nonlinear. Because the relationship is nonlinear, we can’t really predict what an increase in speed would mean for fuel consumption.

3a) The explanatory variable is start median salary and the response variable is mid-career median salary.

3b) The median salary is probably used instead of the mean due to there being a significant skew in the data and potential outliers.

3c) We can estimate mid-career salary given a starting salary of 60,000 because that starting salary is within our data range. We can just find what y-value an x-value of 60,000 corresponds to, which is around 110,000. A median starting salary of $60,000 should yield a median starting salary of around $110,000.

3d) We cannot estimate mid-career salary given a starting salary of 100,000 because that starting salary is not within our data range (highest value is 80,000). If we tried to predict median mid-career salary values outside our data range, we would be extrapolating which could lead to very inaccurate results.

4a) Slope = r(Sy/Sx) = 0.95(46.34/2.23) = 19.74
Intercept = Ymean - Slope*Xmean = 141.67 - 0.95*(46.34/2.23)*11.03 = -76.07

4b) y = 19.74x - 76.07. For every increase of 1 percent in alcohol, the number of calories in a five-ounce serving of alcohol will increase by 19.74. It wouldn’t really make sense to interpret the intercept because having nonalcoholic wine with negative calories doesn’t make sense.

4c) The coefficient of determination = r^2 = 0.95^2 = 0.9025. 90.25% of the variation in calories can be explained by the percentage of alcohol in the 5-oz serving of wine.

4d) Both r and the slope of the regression line will decrease because outliers inherently decrease r and since it is an outlier on the low side the slope will decrease as well.

5a) This will negatively affect his ability to compare the effectiveness of the antidepressants. Having no random assignment exacerbates the effects of confounding variables (especially severity of symptoms in this case) and it could also lead to biases swaying the results.

5b) Without double-blinding, the doctor could subconsciously treat or evaluate the groups differently to influence the results towards what he wishes to see (confirmation bias).

5e) I would recommend randomly assigning the patients to the talk therapy group or to the antidepressant group. This makes the groups comparable and reduces the effects of confounding variables. I would also introduce double blindness. This reduces the possibility of patients acting differently because they know they are treated differently and also reduces the possibility of the doctor treating the patients differently in order to get his expected results.