STATS 10 ASSIGNMENT 2

PART I

EXERCISE 1:

a. Download the data from the course site and read it into R. Or use online data link: read.csv("https://ucla.box.com/shared/static/e9xuft4h3p8fdi4ydoj2hhujee0vmopb.csv") When you read in the data, name your object "flint"

```
> flint<- read.csv("flint.csv")
> View(flint)
> head(flint)
  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
```

b. The EPA states a water source is especially dangerous if the lead level is 15 PPB or greater. What proportion of the locations tested were found to have dangerous lead levels?

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

Therefore, 4.43% is the proportion of areas with dangerous lead levels.

c. Report the mean copper level for only test sites in the North region.

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

Therefore, the mean copper level for only test sites in the North region is 44.6426.

d. Report the mean copper level for only test sites with dangerous lead levels (at least 15 PPB).

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

Therefore, the mean copper level for only test sites with dangerous lead levels is 305.8333.

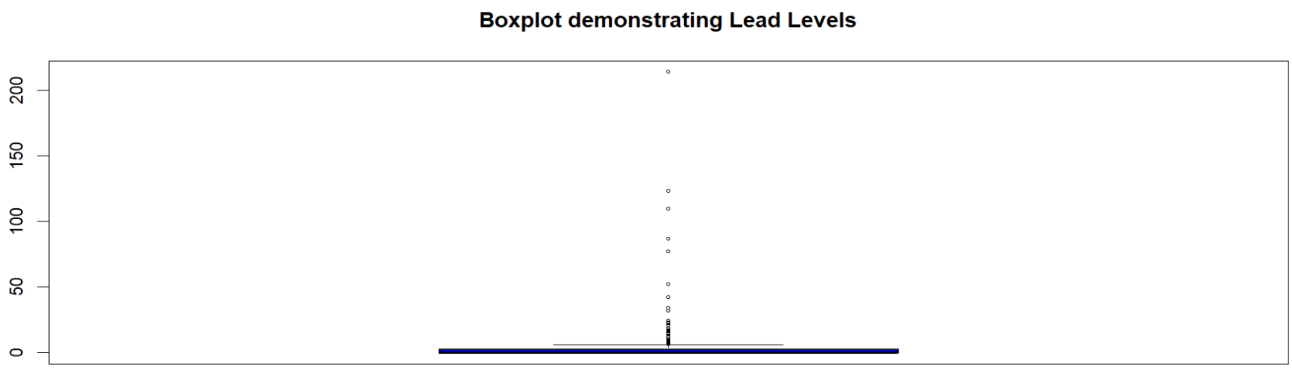
e. Report the mean lead and copper levels.

```
> mean_lead<- mean(flint$Pb)
> mean_lead
[1] 3.383272
> mean_copper<- mean(flint$Cu)
> mean_copper
[1] 54.58102
```

The mean lead and copper levels are 3.383272 and 54.58102 respectively.

f. Create a box plot with a good title for the lead levels.

```
> boxplot(flint$Pb, main="Boxplot demonstrating Lead Levels", col="b
lue", cex= 0.5)
```



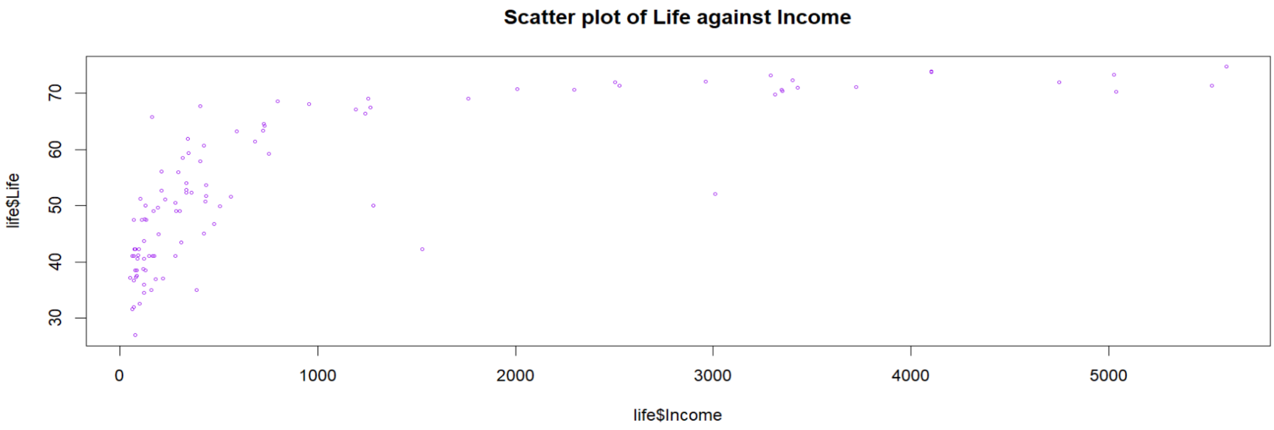
g. Based on what you see in part (f), does the mean seem to be a good measure of center for the data? Report a more useful statistic for this data.

No, mean is not a good measure of the central tendancy for the data because of the vast distribution of the values in the data set. As it can be observed that the data is skewed with many outliers, hence, the median is a better measure of the center.

EXERCISE 2

a. Construct a scatterplot of Life against Income. Note: Income should be on the horizontal axis. How does income appear to affect life expectancy?

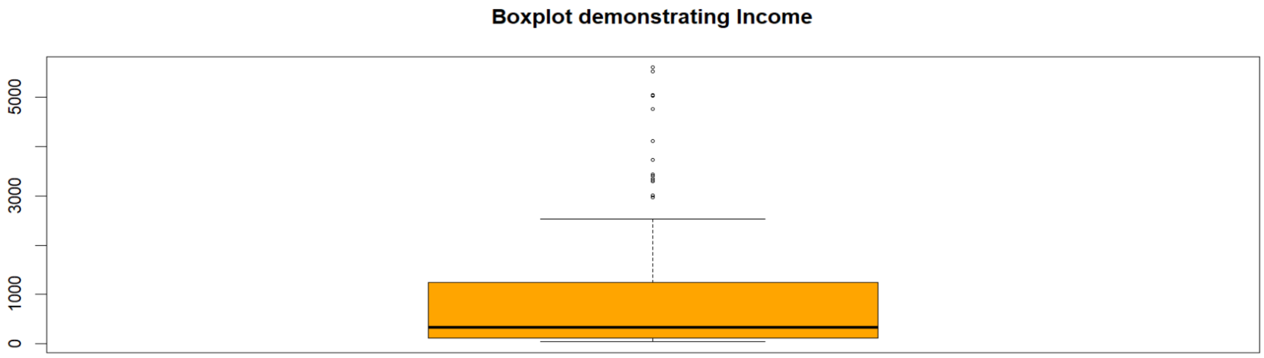
```
> plot(life$Life~ life$Income, cex=0.5, col= "purple", main="Scatter
plot of Life against Income")
```



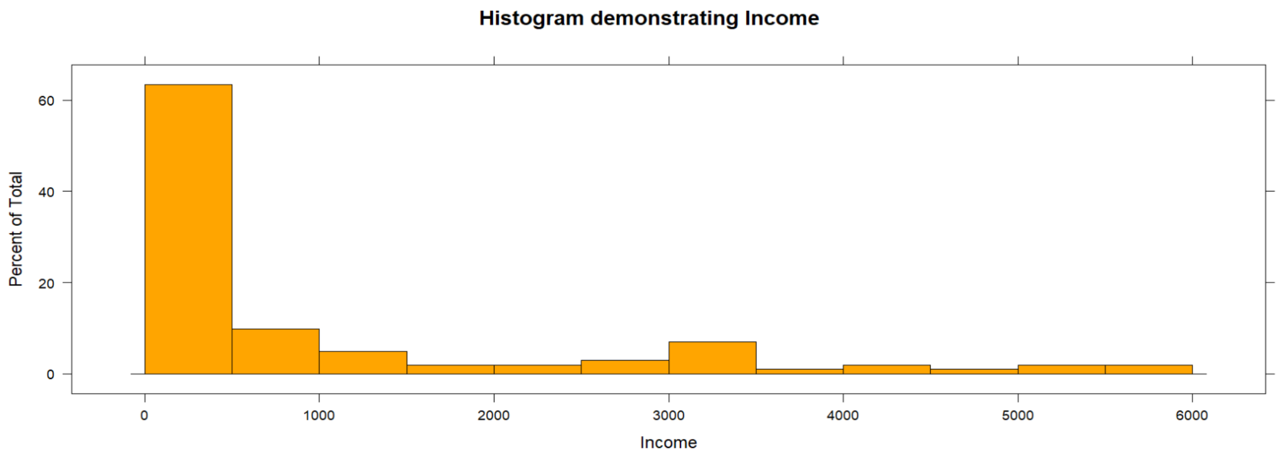
There is a positive non-linear relationship between the variables. It can be observed that people with higher per capita income tend to have greater life expectancies.

b. Construct the boxplot and histogram of Income. Are there any outliers?

```
> boxplot(life$Income, main="Boxplot demonstrating Income", col="ora
nge", cex= 0.5)
```




```
> histogram(life$Income, main="Histogram demonstrating Income", col="orang
e", breaks=10, xlab="Income")
```



Yes, as it can be clearly seen in both the Boxplot and the Histogram that there are outliers present in the dataset.

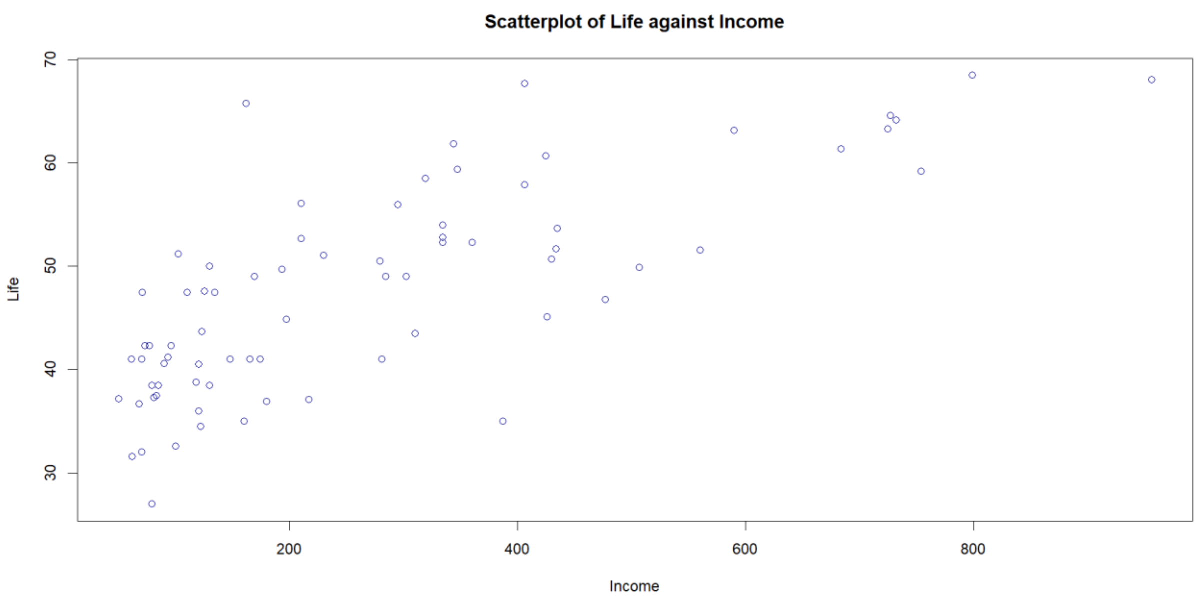
c. Split the data set into two parts: One for which the Income is strictly below $1000, and one for which the Income is at least $1000. Come up with your own names for these two objects.

```
> life_low <-life[life$Income < 1000, names(life)]
> life_high <- life[ life$Income >= 1000, names(life) ]
```

life	101 obs. of 3 variables
life_high	27 obs. of 3 variables
life_low	74 obs. of 3 variables

d. Use the data for which the Income is below $1000. Plot Life against Income and compute the correlation coefficient. Hint: use the function cor()

```
> low_income<- life[life$Income<1000,]
> high_income<-life[life$Income>=1000,]
> plot(Life~Income, data= low_income, col="dark blue", main="Scatterplot o
f Life against Income")
> cor(low_income$Life, low_income$Income)
[1] 0.752886
```



The correlation coefficient is 0.752886.

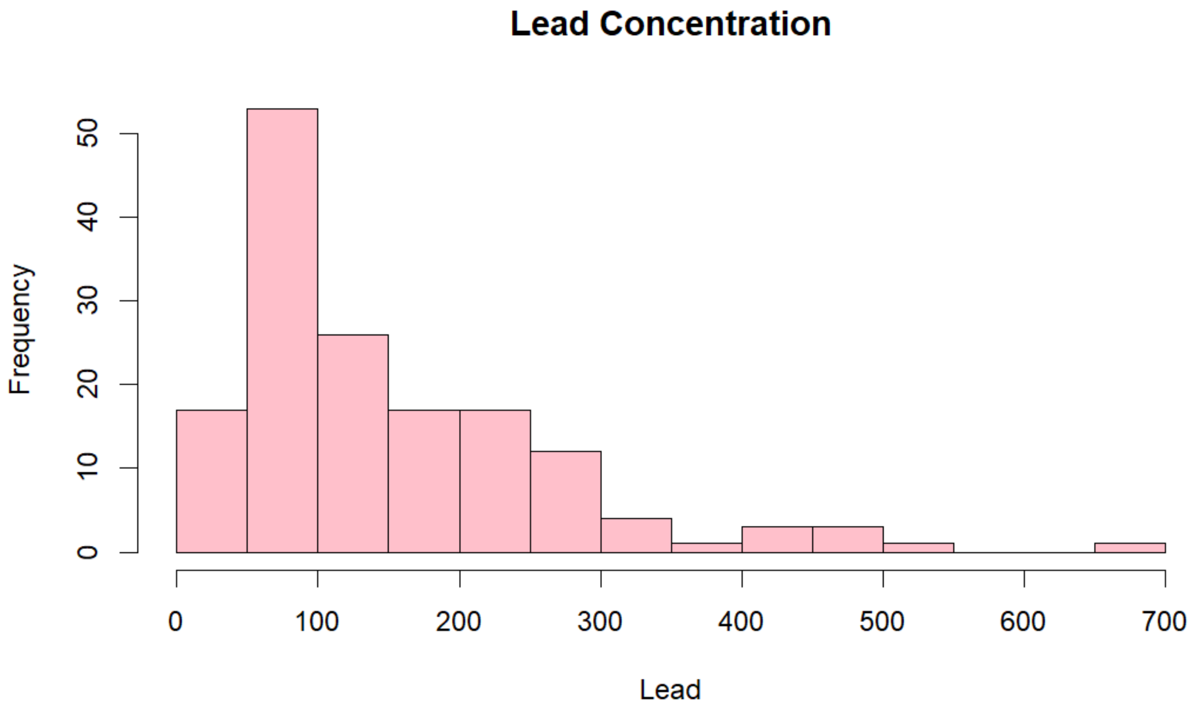
EXERCISE 3

a. Compute the summary statistics for lead and zinc using the summary() function.

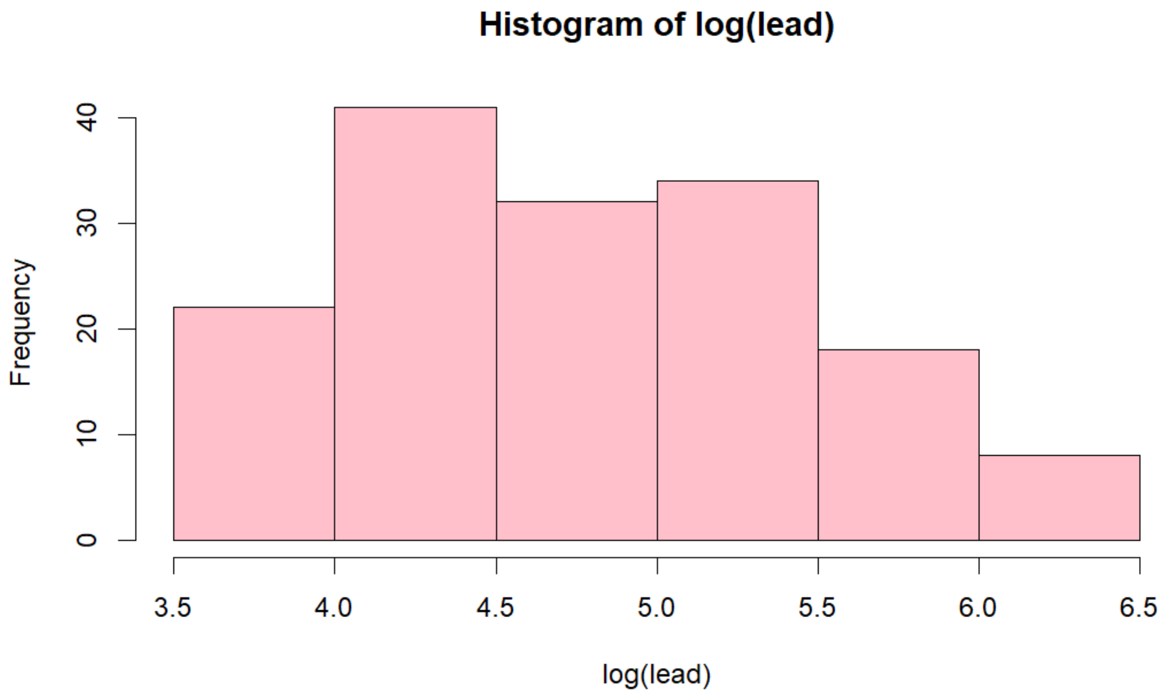
```
> summary(maas$lead)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   37.0    72.5   123.0   153.4   207.0   654.0 
> summary(maas$zinc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  113.0   198.0   326.0   469.7   674.5  1839.0 
```

b. Plot two histograms: one of lead and one of log(lead).

```
> hist(maas$lead, main = "Lead Concentration", xlab = "Lead", col="pink"
)
```

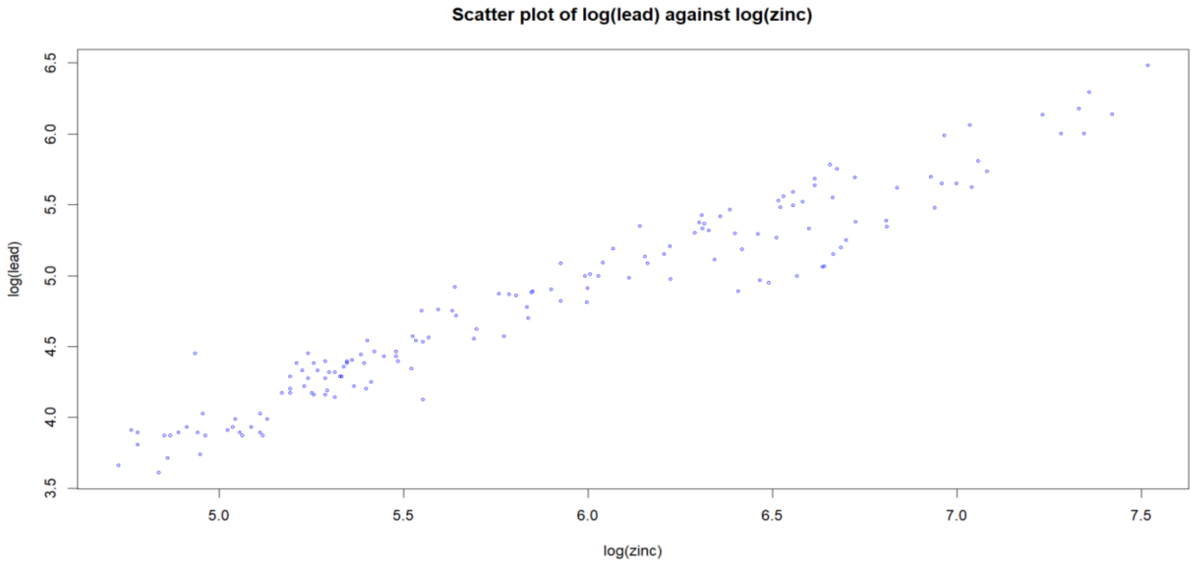



```
> lead <- maas$lead
> hist(log(lead), main = "Histogram of log(lead)", col="pink")
```



c. Plot log(lead) against log(zinc). What do you observe?

```
plot(log(maas$lead)~ log(maas$zinc), cex=0.5, col= "blue", main="Scatter p
lot of log(lead) against log(zinc)", ylab="log(lead)", xlab= "log(zinc)")
```



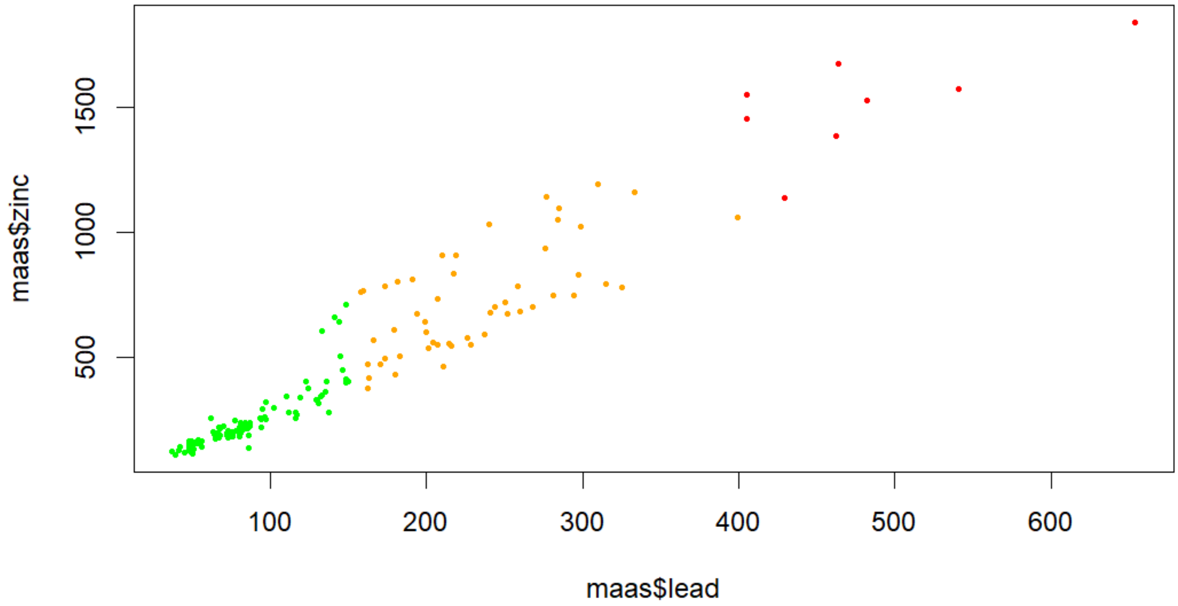
Observation: There is a strong positive linear trend between the two variables - the log of lead and the log of zinc.

d. The level of risk for surface soil based on lead concentration in ppm is given on the tab le below:

Mean concentration (ppm)	Level of risk
Below 150	Lead-free
Between 150-400	Lead-safe
Above 400	Significant environmental lead hazard

Use similar techniques to give different colors and sizes to the lead concentration at these 155 locations.

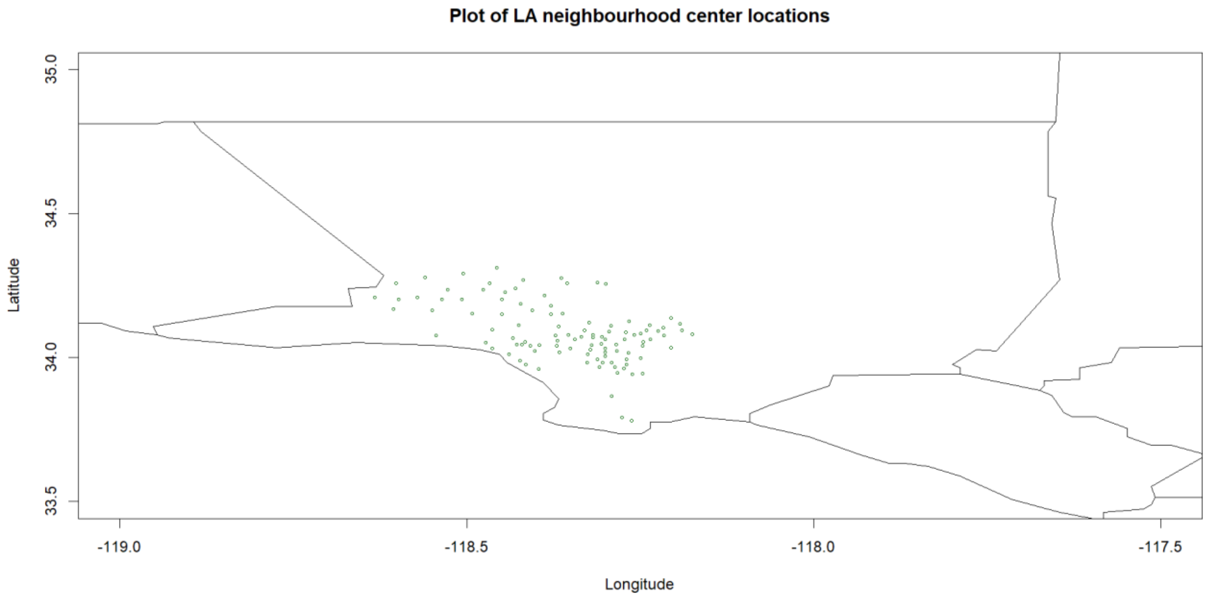
```
> mycolors <- c("green", "orange", "red")
> mylevels <- cut(lead, c(0, 150, 400, 1000))
> plot(maas$lead, maas$zinc, col=mycolors[as.numeric(mylevels)], cex=0.5, pch=19)
```



EXERCISE 4

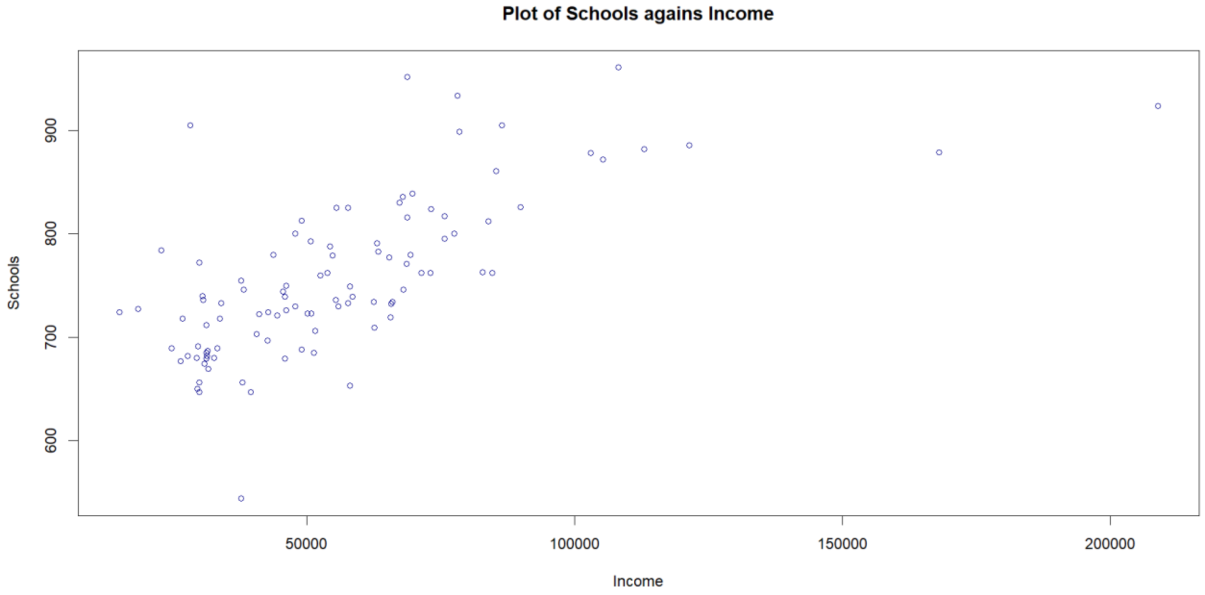
a. Plot the data point locations. Use good formatting for the axes and title. Then add the outline of LA County by typing: map("county", "california", add = TRUE)

```
> plot(Latitude ~ Longitude, data = LA, main = "Plot of LA neighbourhood center locations", xlim = c(-119, -117.5) , ylim = c(33.5, 35), cex=0.5, col="darkgreen")
> map("county", "california", add = TRUE)
```



b. Do you see any relationship between income and school performance? Hint: Plot the variable Schools against the variable Income and describe what you see. Ignore the data points on the plot for which Schools = 0. Use what you learned about subsetting with logical statements to first create the objects you need for the scatter plot. Then, create the scatter plot. Alternate methods may only receive half credit.

```
> LA_sch<- LA[LA$Schools!=0,]
> plot(Schools ~ Income, data= LA_sch, ylab="Schools", xlab="Income"
, main="Plot of Schools agains Income", cex= 0.8, col="darkblue")
```



There is a positive nonlinear relationship between the schools and income. Outliers can also be observed.

EXERCISE 5

In this exercise, you will work with a dataset containing information about customers of a retail store.
The dataset includes the following variables:

a. Customer ID: unique identifier for each customer
b. Age: age of the customer in years
c. Gender: gender of the customer (M for male, F for female)
d. Income: annual income of the customer in dollars
e. Education: education level of the customer (high school, some college, college degree, graduate degree)
f. Marital status: marital status of the customer (single, married, divorced, widowed)
g. Purchase amount: the total amount the customer spent at the store in the past year

a. Are there any missing values in the dataset? If so, how many are there and which variables have missing values?

```
> customer_data <- read.csv("https://ucla.box.com/shared/static/y2y8rcie7m
jw2h5t92x9dfcp133tc90h.csv")
> is.na(customer_data)
      cust_id   age gender income education marital_status
 [1,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [2,]   FALSE FALSE  FALSE   TRUE     FALSE          FALSE
 [3,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [4,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [5,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [6,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [7,]   FALSE FALSE  FALSE   TRUE     FALSE          FALSE
 [8,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
 [9,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[10,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[11,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[12,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
[13,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[14,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[15,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[16,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[17,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[18,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[19,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[20,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[21,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[22,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[23,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[24,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[25,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[26,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[27,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[28,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[29,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[30,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[31,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[32,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[33,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[34,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[35,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[36,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[37,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[38,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
```

```
 [39,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [40,]   FALSE FALSE  FALSE   TRUE     FALSE          FALSE
 [41,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
 [42,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [43,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [44,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [45,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [46,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [47,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [48,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [49,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [50,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [51,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [52,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
 [53,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [54,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [55,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [56,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [57,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
 [58,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [59,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [60,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [61,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [62,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [63,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [64,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [65,]   FALSE FALSE  FALSE   TRUE     FALSE          FALSE
 [66,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [67,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
 [68,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [69,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [70,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [71,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [72,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
 [73,]   FALSE  TRUE  FALSE   TRUE     FALSE          FALSE
 [74,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [75,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [76,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [77,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [78,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
 [79,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [80,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [81,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [82,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [83,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [84,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [85,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [86,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [87,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [88,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [89,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [90,]   FALSE  TRUE  FALSE  FALSE     FALSE          FALSE
 [91,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [92,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [93,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [94,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [95,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [96,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [97,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [98,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
 [99,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
[100,]   FALSE FALSE  FALSE  FALSE     FALSE          FALSE
       purchase_amt
  [1,]        FALSE
  [2,]        FALSE
  [3,]        FALSE
  [4,]        FALSE
  [5,]        FALSE
  [6,]        FALSE
```

```
 [7,]          TRUE
 [8,]         FALSE
 [9,]         FALSE
[10,]         FALSE
[11,]         FALSE
[12,]         FALSE
[13,]         FALSE
[14,]         FALSE
[15,]         FALSE
[16,]         FALSE
[17,]         FALSE
[18,]         FALSE
[19,]         FALSE
[20,]          TRUE
[21,]         FALSE
[22,]         FALSE
[23,]         FALSE
[24,]         FALSE
[25,]         FALSE
[26,]         FALSE
[27,]         FALSE
[28,]         FALSE
[29,]          TRUE
[30,]         FALSE
[31,]         FALSE
[32,]         FALSE
[33,]         FALSE
[34,]         FALSE
[35,]         FALSE
[36,]         FALSE
[37,]         FALSE
[38,]         FALSE
[39,]         FALSE
[40,]         FALSE
[41,]         FALSE
[42,]         FALSE
[43,]         FALSE
[44,]         FALSE
[45,]          TRUE
[46,]         FALSE
[47,]         FALSE
[48,]         FALSE
[49,]         FALSE
[50,]         FALSE
[51,]         FALSE
[52,]         FALSE
[53,]         FALSE
[54,]         FALSE
[55,]         FALSE
[56,]         FALSE
[57,]         FALSE
[58,]         FALSE
[59,]         FALSE
[60,]         FALSE
[61,]         FALSE
[62,]         FALSE
[63,]         FALSE
[64,]         FALSE
[65,]         FALSE
[66,]         FALSE
[67,]         FALSE
[68,]         FALSE
[69,]         FALSE
[70,]         FALSE
[71,]         FALSE
[72,]         FALSE
[73,]         FALSE
[74,]         FALSE
[75,]         FALSE
```

```
 [76,]         FALSE
 [77,]          TRUE
 [78,]          TRUE
 [79,]         FALSE
 [80,]          TRUE
 [81,]         FALSE
 [82,]         FALSE
 [83,]         FALSE
 [84,]         FALSE
 [85,]         FALSE
 [86,]         FALSE
 [87,]         FALSE
 [88,]         FALSE
 [89,]         FALSE
 [90,]         FALSE
 [91,]         FALSE
 [92,]         FALSE
 [93,]         FALSE
 [94,]         FALSE
 [95,]         FALSE
 [96,]         FALSE
 [97,]         FALSE
 [98,]         FALSE
 [99,]         FALSE
[100,]         FALSE
> sum(is.na(customer_data))
[1] 22
> sum(is.na(customer_data$cust_id))
[1] 0
> sum(is.na(customer_data$age))
[1] 10
> sum(is.na(customer_data$gender))
[1] 0
> sum(is.na(customer_data$income))
[1] 5
> sum(is.na(customer_data$education))
[1] 0
> sum(is.na(customer_data$marital_status))
[1] 0
> sum(is.na(customer_data$purchase_amt))
[1] 7
```

Yes, there are missing values in the dataset. There are 22 missing values in the dataset. The variables with missing values are age, income, and purchase_amt.

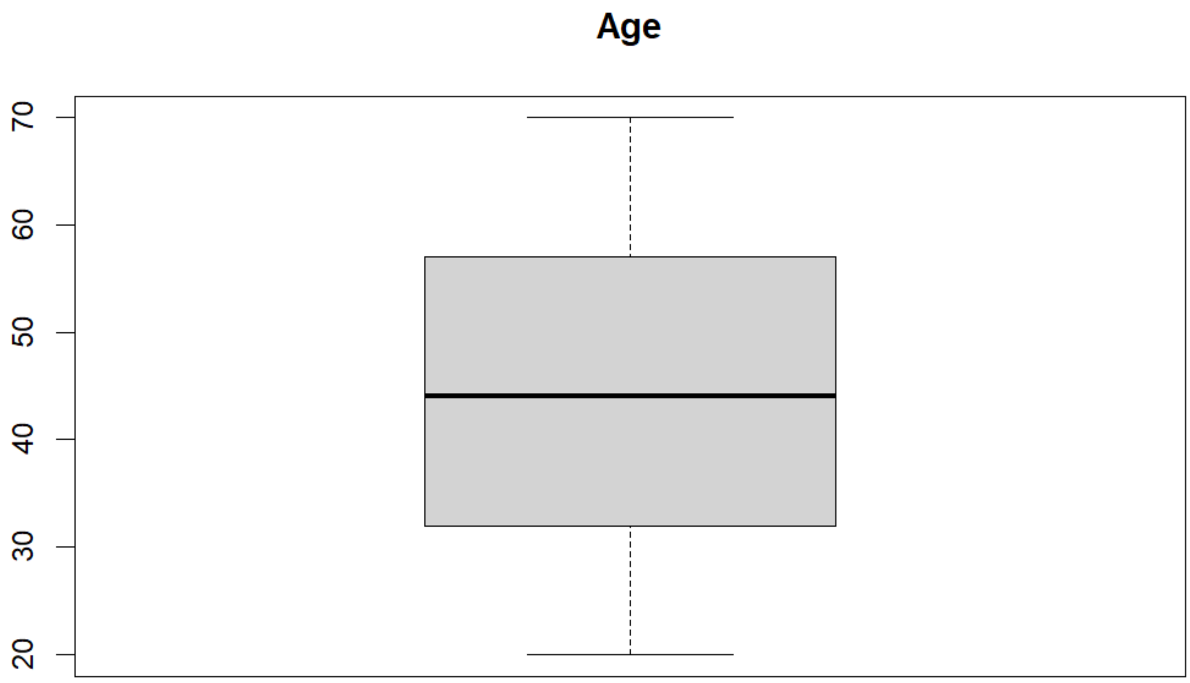
b. What is the data type of each variable? Are there any variables that should be converted to a different data type?

```
> class(customer_data$cust_id)
[1] "character"
> class(customer_data$age)
[1] "integer"
> class(customer_data$gender)
[1] "character"
> class(customer_data$income)
[1] "integer"
> class(customer_data$education)
[1] "character"
> class(customer_data$marital_status)
[1] "character"
> class(customer_data$purchase_amt)
[1] "integer"
```

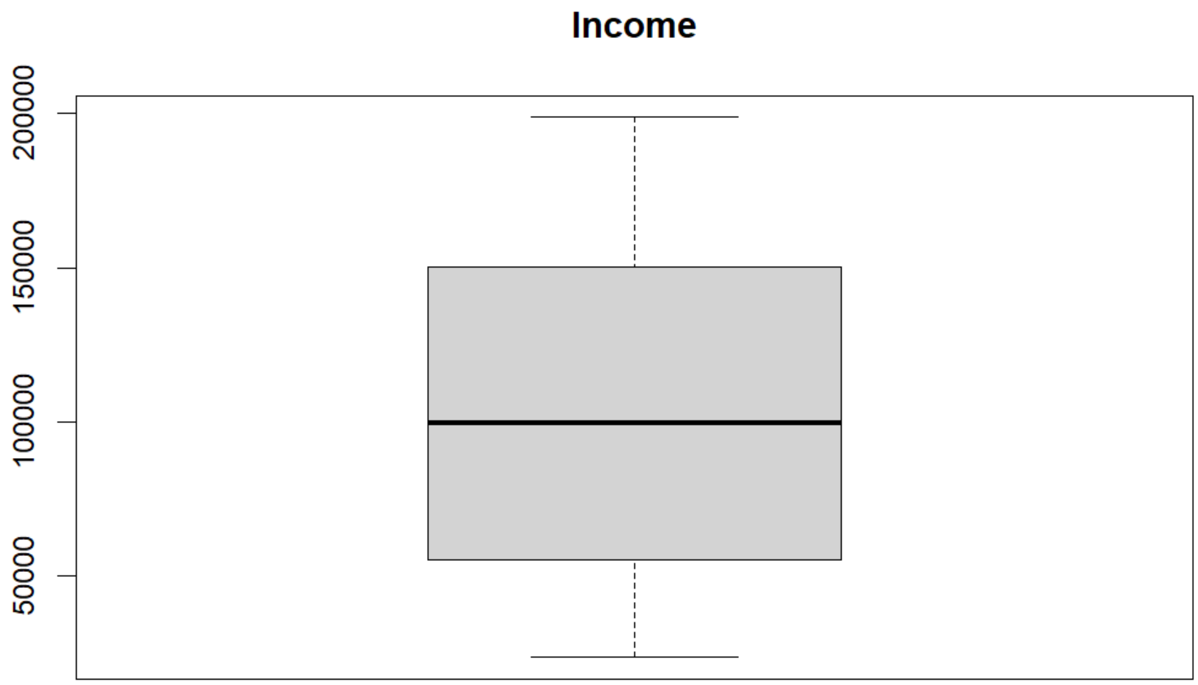
No, there are no variables for which their data types need to be changed.

c. Do any numerical variables have outliers or extreme values? If so, how would you handle them? Provide your analysis in R for identifying outliers (e.g., visualization, numerical summary statistics). This is an open-ended question, so please feel free to use any appropriate methods to identify and deal with any outliers or extreme values in the dataset.

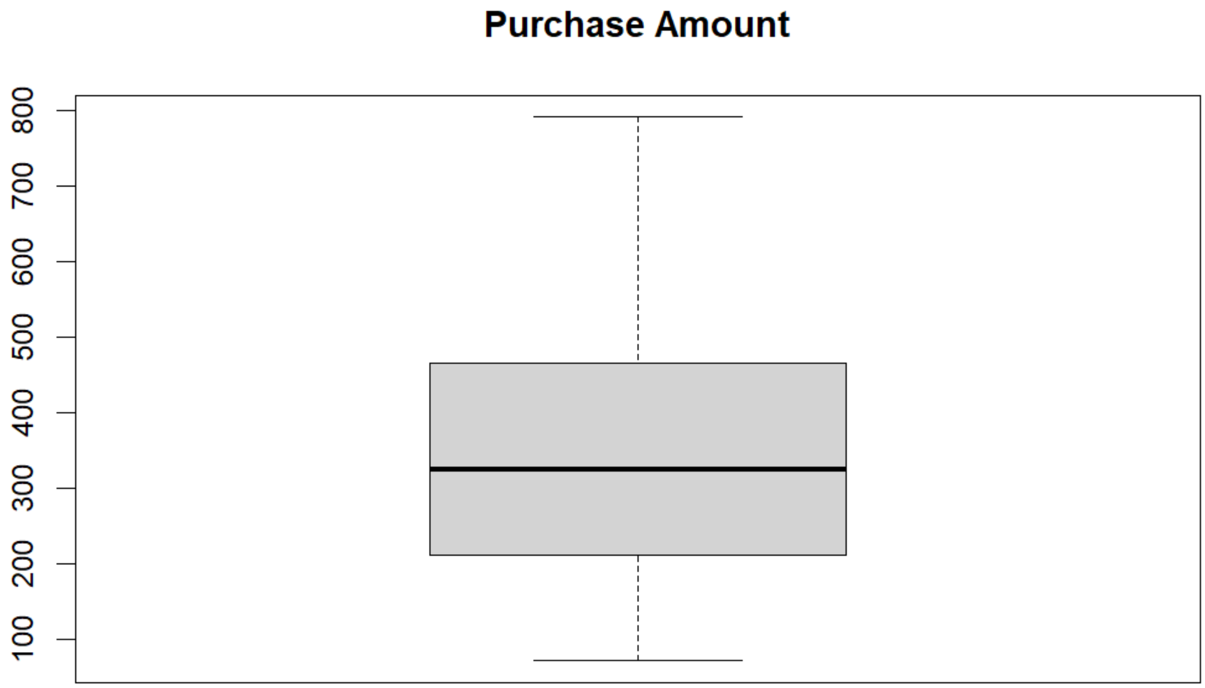
```
> boxplot(customer_data$age, main = "Age")
```




```
> boxplot(customer_data$income, main = "Income")
```




```
> boxplot(customer_data$purchase_amt, main = "Purchase Amount")
```



PART II

Exercise 1

A study was done random sample of 900 college students. The researcher wants to find out if gender would affect people's body image. The two-way table below summarizes the two variables.

Two-way table		Body Image			
		About right	Overweight	Underweight	Total
Gender	Female	310	130	30	
	Male	290	68	72	
	Total				900

a. In general, are students happy with their body weight? (Hint: Students that are happy with their body weight responded "about right.")

b. If the researcher wants to compare the differences in body image between females and males. What graph would best visualize the data for this purpose? Explain. (No need to draw the actually plot)

c. Are female students more likely to feel they are about right than male students? Explain with numerical evidence.

d. For students who do not feel 'about right' with their body image, are there any differences between the two gender groups? (Hint: are they more likely to feel there are overweight or underweight? Do female students and male students feel the same way?)

a) % of total students who votes 'about right': $\frac{\text{no. of students who voted about righ}}{\text{total no. of students}} \times 100$

$\frac{310+290}{900} \times 100 = 66.67\%$

66.67% of students responded 'about right'.

Hence, in general the students are happy with their body weight as only 33.33% are not happy.

b) If the researcher wants to compare the differences in body image between females and males, the best graph to visualize it is a bar chart.
It is the best way to visualize categorical variables.
It would allow us to compare percentages of females and males who responded different things.

c) % of females that voted about right = $\frac{\text{no. of female responded} \times 100}{\text{total no. of females}}$

$= \frac{310}{470} \times 100 = 65.96\%$.

% of males that voted about right = $\frac{\text{no. of males responded} \times 100}{\text{total no of males}}$

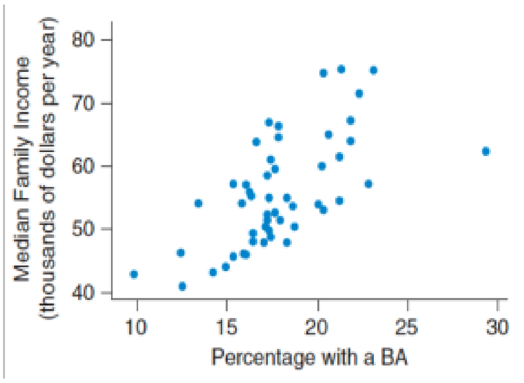
$= 67.44\%$.

Hence as % of males that responded about right is greater than % of females, male students are more likely to feel they are about right than females

d) % of females feeling overweight = $\frac{130}{470} \times 100 = 27.66\%$
% of females feeling underweight = $\frac{30}{470} \times 100 = 6.39\%$
% of males feeling overweight = $\frac{68}{430} \times 100 = 15.81\%$
% of males feeling underweight = $\frac{72}{430} \times 100 = 16.74\%$
∴ larger proportion of male student are more likely to feel underweight than overweight whereas large proportion of the female students are more likely to feel overweight than underweight

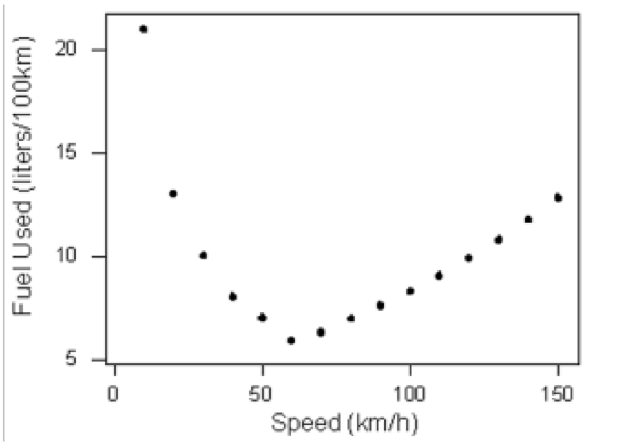
Exercise 2
For each of the scatterplots shown, provide a written description that includes the direction, form, and strength of the relationship, along with any outliers that do not fit the general trend. In addition, explain what these characteristics mean in the context of the data.

a. Data on 50 states taken from the U.S. Census shows how the median family income is related to the population (25 years or older) with a college degree or higher.



b. Consider the relationship between the average amount of fuel used (in liters) to drive a fixed distance in a car (100 km), and the speed at which the car is driven (in km per hour).

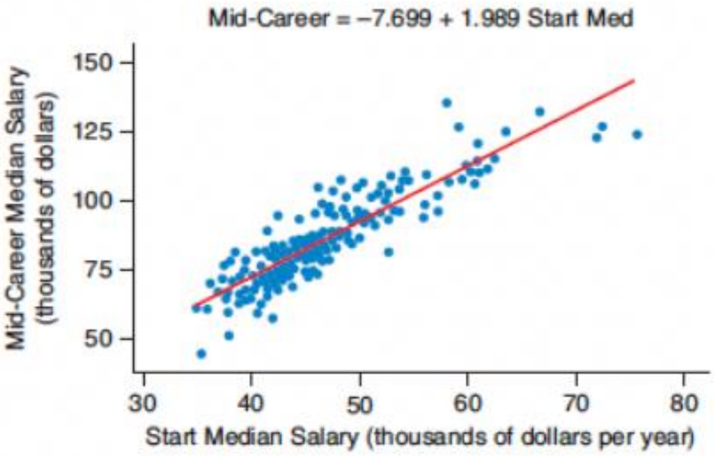
c.



a) This scatterplot depicts a linear but ~~weak~~ moderately strong and positive correlation
The outlier in this data is around 29% of the population
with a BA who make around \$62000 as their family inc per year
It shows that people (25 year or older) who are educated with a BA
~~(a)~~ tend to have a higher median family income per annum. As the points
are not close together it is moderately strong correlation. Outliers further
weaken the relationship and impact measures such as mean and standard deviation
b) It is not linear and neither positive nor negative. However,
it is a strong non linear plot. The only outlier is around 10km/h
with a high fuel usage at around 22 L/100km. The
~~Exercise 3~~ graph shows that the fuel consumption
initially dec when the speed of car is below 60km/h. However,
as the speed inc beyond 60km/hr, the fuel consumption
seems to inc consistently with a linear trend. The outlier
indicates that fuel used at 10km/h is unusually high.

Exercise 3

A researcher collected data on the median starting salaries and the median mid-career salaries for graduates at a selection of colleges. (Source: The Wall Street Journal, Salary increase by salary type, https://www.wsj.com/public/resources/documents/info-Salaries_for_Colleges_by_Type-sort.html). The data points and the fitted least squares regression line are displayed in the graph below.



a. What is the explanatory variable and response variable?

b. And why do you think the median salary is used instead of the mean?

c. Can the median mid-career salary be estimated given a median starting salary of 60 (in thousands of dollars)? Please explain why or why not, and show your calculation and explanation if possible.

d. Can the median mid-career salary be estimated given a median starting salary of 100 (in thousands of dollars)? Please explain why or why not, and show your calculation and explanation if possible.

Exercise 3

a) ~~explanatory~~ response variable: mid career median salary measured in thousands of dollars.
~~Response~~ explanatory variable: Start median salary measured in thousands of dollars per year.

b) median salary is used instead of the mean because salary distributions are usually skewed, making median a more meaningful measure of centre. median does not get affected by outliers as much as the mean does.

c) Yes we can estimate it using the fitted least square regression line.

mid career = $-7.699 + 1.8989$ start mid

mid career = $-7.699 + 1.986(960)$

$= 111.641$

$\therefore$, the estimated median mid-career salary is 111.641 (in thousands of dollars)

d) The fitted least square regression line only extends till 75 (in thousands of dollars). Therefore, we can't estimate the median mid career salary for a median starting salary as we cannot extrapolate. Making an assumption of the behavior of the plot might result in wrong results.

Exercise 4

Assume that the relationship between the calories in a five-ounce serving and the % alcohol content for a sample of wines is linear. Use the % alcohol as the explanatory variable, and fit a least squares regression line.

a. Calculate slope and intercept of the regression line.

b. Report the equation of the regression line and interpret it in the context of the problem.

c. Find and interpret the value of the coefficient of determination.

d. Suppose a new point was added to your data: a wine that is 20% alcohol that contains 0 calories. How will that affect the value of r and the slope of the regression line? (No calculation needed)

Data table (Source:healthalicious.com)

Calories	% alcohol
122	10.6
119	10.1
121	10.1
123	8.8
129	11.1
236	15.2

Table of summary statistics

	Calories	% alcohol
Mean	141.67	11.03
Std. Dev.	46.34	2.32
r	0.95	

Exercise 4)

a) slope of regression line: $b = r \frac{sy}{sx}$

$$= \frac{0.95 \times 46.34}{2.32} = 18.98\%$$

Intercept of regression line: $a = \bar{y} - b\bar{x}$

$$= 141.67 - 18.98 \times 11.03$$

$$= -67.\cancel{23}\,683$$

b) equation of the regression line: $y = a + bx$

calories = $-67.\cancel{2}\cancel{8}3 +$ 18.98 alcohol.

$y = -67.63 + 18.98x$

The equation can be interpreted as: for every one percent increase in alcohol content, the calories in beverage is expected to increase by 18.98%. The slope of regression line is positive.

c) The coefficient of determination (R^2) measures the proportion of the total variation in the dependent variable (calories) that is explained by the independent variable (% alcohol).

$R^2 = r^2 = 0.95^2 = 0.9025$

This means that 90.25% of the variation in calories can be explained by the variation in % alcohol.

d) As the new point is not consistent with the existing data, the value of r will change. Outlier weakens the relationship. The correlation coefficient and slope of regression line will ~~increase~~ decrease as the new point being an outlier will have a large influence and weaken the relationship. ~~r inc because it will have to pass through the new point at 20% with 0 calories.~~ b is affected by r. $b = r\frac{s_y}{s_x}$

b is directly proportional to r. a dec in r will lead ~~to~~ a decrease in b.

Exercise 5

A doctor who believes strongly that antidepressants work better than "talk therapy" tests depressed patients by treating half of them with antidepressants and the other half with talk therapy. The doctor recruited 100 patients for the study. After six months' treatment, the patients will be evaluated on a scale of 1 to 5, with 5 indicating the greatest improvement. The doctor is designing the study plan.

a. The doctor wants to put the most severe patients in the antidepressants group because he is concerned about those patients' conditions. Will this affect his ability to compare the effectiveness of the antidepressants and the "talk therapy"? Explain.

b. The doctor asks you whether it is acceptable for him to know which treatment each patient receives. Explain why this practice may affect his ability to compare the two groups.

c. What improvements to the plan would you recommend?

a) Yes it will affect his ability to compare the effectiveness of the antidepressants and the talk therapy as the group of severely depressed patients is not the representative of the overall sample, hence the study might result in a inaccurate results. The patients should be randomly assigned to the 2 treatments and not according to the severity of their conditions. Severity of the patient's conditions is a confounding variable that will impact the result of the study. Bias can occur from non-randomized assignment and influence results in a particular direction.

b) This practice may affect his ability to compare the two groups as he will be incline to interpret the observations in the favor of the results that he is expecting which is that antidepressants work better than talk therapy. This confirmation bias may result in the doctor concluding that antidepressants work better despite the actual result. Both researchers and participants should be unaware and follow a double blind format.

c) To prevent any kind of bias, the study should be made double blind, ~~the~~ neither the doctor nor the patient should know which treatment they are undergoing. The patients should not be aware of the original hypothesis made by the doctor that antidepressants work better than talk therapy. The doc should have a large sample size to ensure full range of variability in the subjects being studied. The comparison group or talk therapy group could be given harmless pills as placebos in order to control for possible differences between groups that may occur because some subjects are more likely than others to ~~expect~~ expect their treatments to be effective