Assignment 2 Key

Makena Pollon

2023-11-06

General Note: For Part 1 problems, any form of subsetting to get the desired results is fine. Alternative methods like R functions from tidyverse/dyplr etc. are not allowed.

Exercise 1

Work with lead and copper data obtained from the residents of Flint, Michigan from January- February, 2017. Data are reported in PPB (parts per billion, or μg/L) from each residential testing kit. Remember that "Pb" denotes lead, and "Cu" denotes copper. You can learn more about the Flint water crisis at https://en.wikipedia.org/wiki/Flint_water_crisis.

a. Download the data from the course site and read it into R. Or use online data link. When you read in the data, name your object "flint".

Note: Either import method is fine.

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

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
```

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

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

```
## [1] 44.6424
```

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
```

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

e. Report the mean lead and copper levels.

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

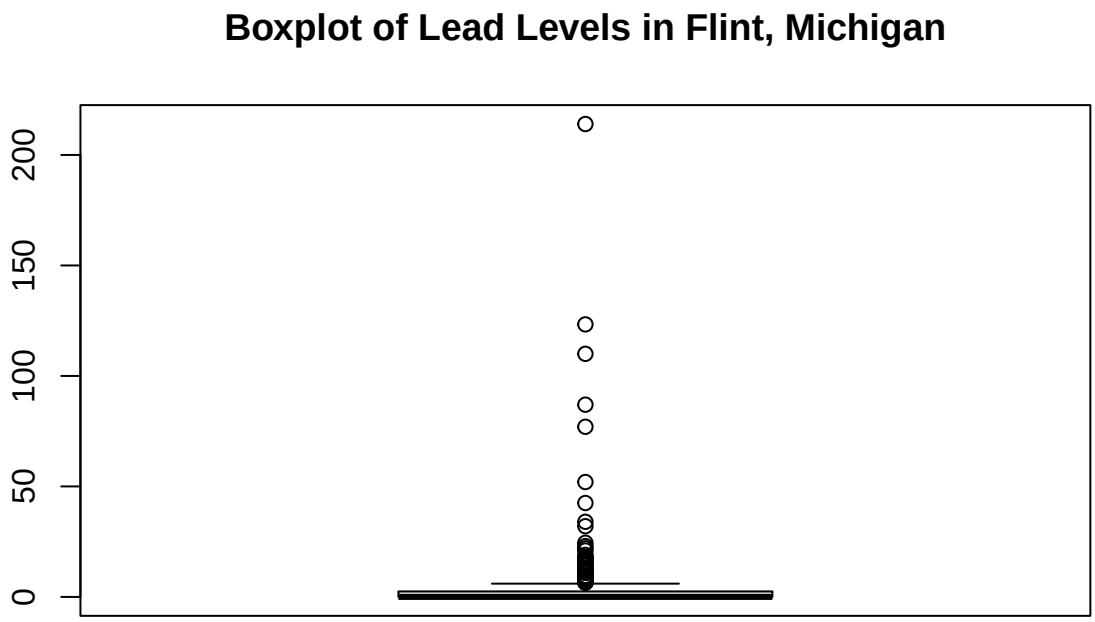
```
## [1] 54.58102
```

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

```
## [1] 3.383272
```

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

```
boxplot(flint$Pb, main = "Boxplot of Lead Levels in Flint, Michigan")
```



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 it does not, the boxplot gives good indication that the distribution is heavy on outliers, so the median would be better since it isn't as heavily influenced by the values of the outliers unlike the mean.

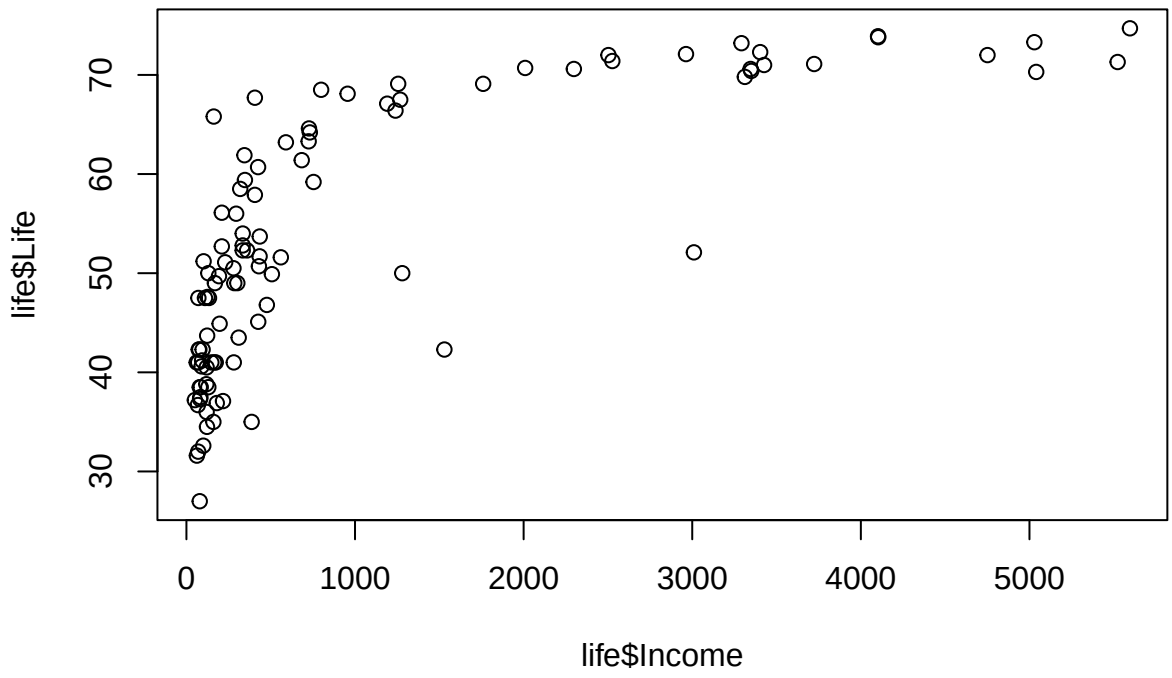
Exercise 2

The data here represent life expectancies (Life) and per capita income (Income) in 1974 dollars for 101 countries in the early 1970's. The source of these data is: Leinhardt and Wasserman (1979), New York Times (September, 28, 1975, p. E-3). They also appear on Regression Analysis by Ashish Sen and Muni Srivastava.

```
life <-read.table("https://ucla.box.com/shared/static/rqk4lc030pabv30wknx2ft9jy848ub9n.txt", header = TR
```

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)
```

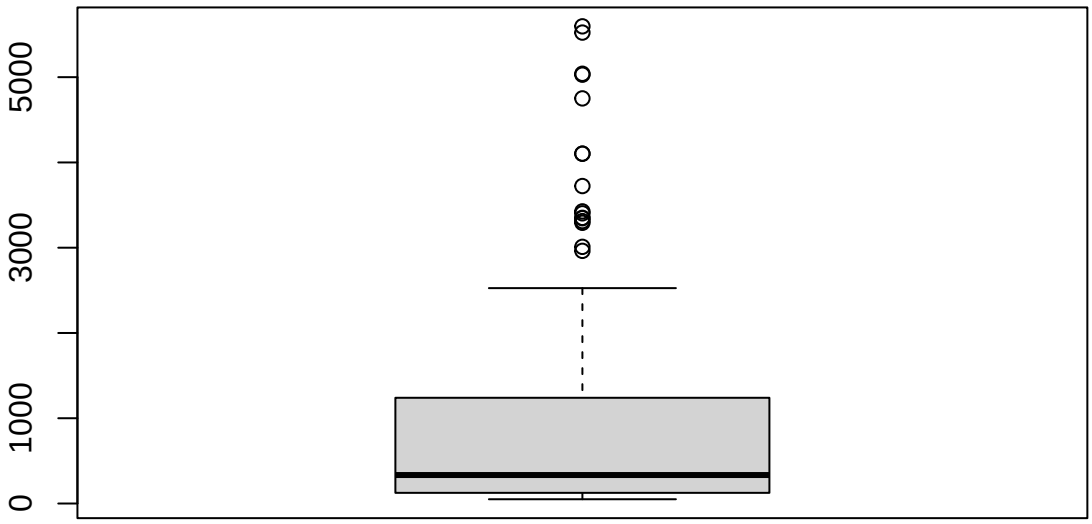


The relationship between life and income seems to be logarithmic in that an increase in income is positively correlated with an increase in life expectancy but there are diminishing returns as income increases.

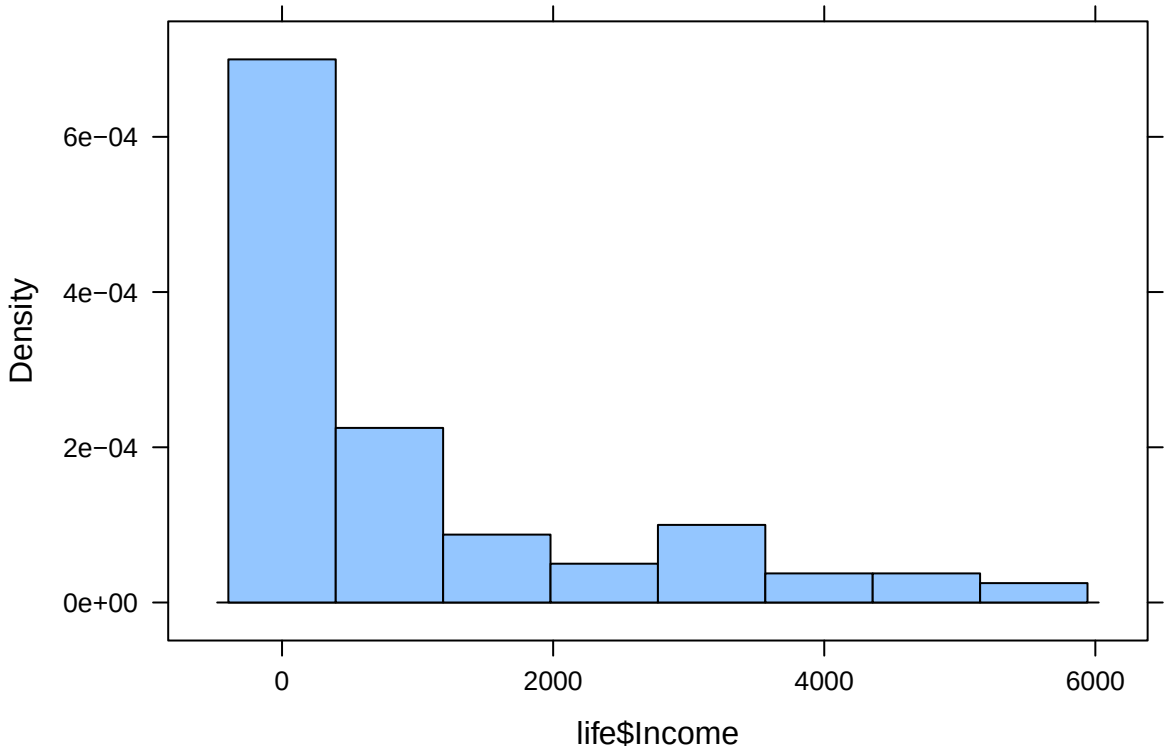
Note: As long as the answer does not say that the relationship is linear.

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

```
boxplot(life$Income)
```




```
histogram(life$Income)
```



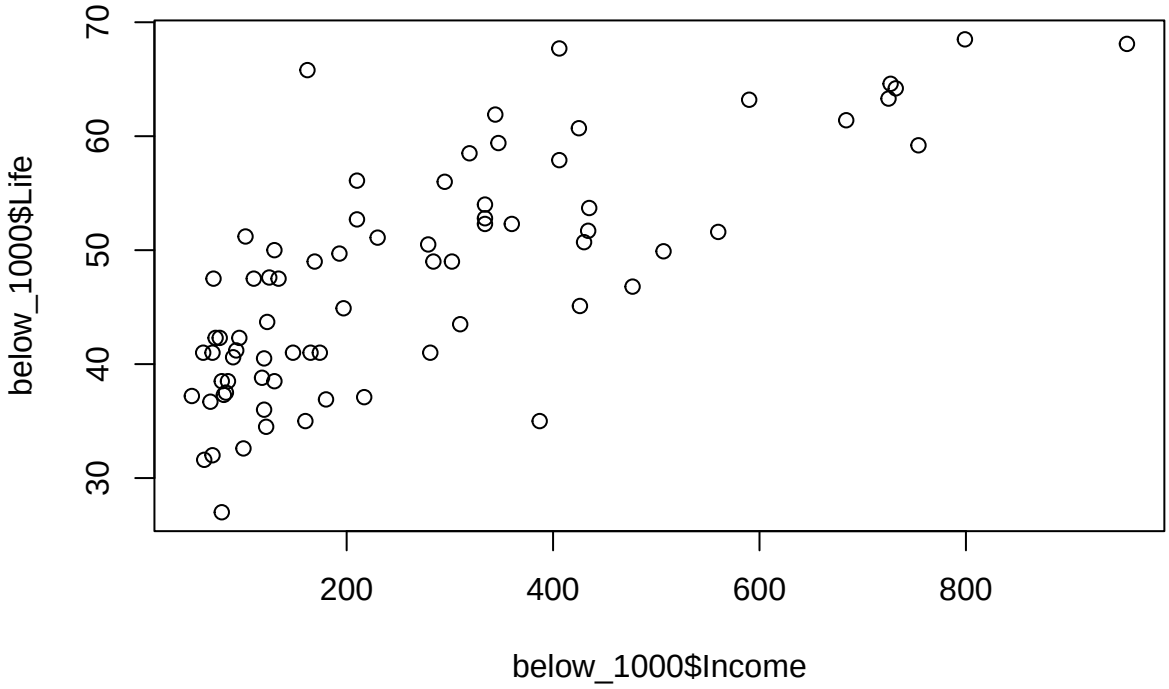
Yes, there appear to be outliers among those who had the higher income values.

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.

```
below_1000 <- life[life$Income < 1000,]
atleast_1000 <- life[life$Income >= 1000,]
```

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()

```
plot(below_1000$Life ~ below_1000$Income)
```




```
cor(below_1000$Life ~ below_1000$Income)
```

```
## [1] 0.752886
```

Exercise 3

The Maas river data contain the concentration of lead and zinc in ppm at 155 locations at the banks of the Maas river in the Netherlands.

```
maas <- read.table("https://ucla.box.com/shared/static/tv3cxooyp6y8fh6gb0qj2cxihj8klg1h.txt", header = '
```

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
```

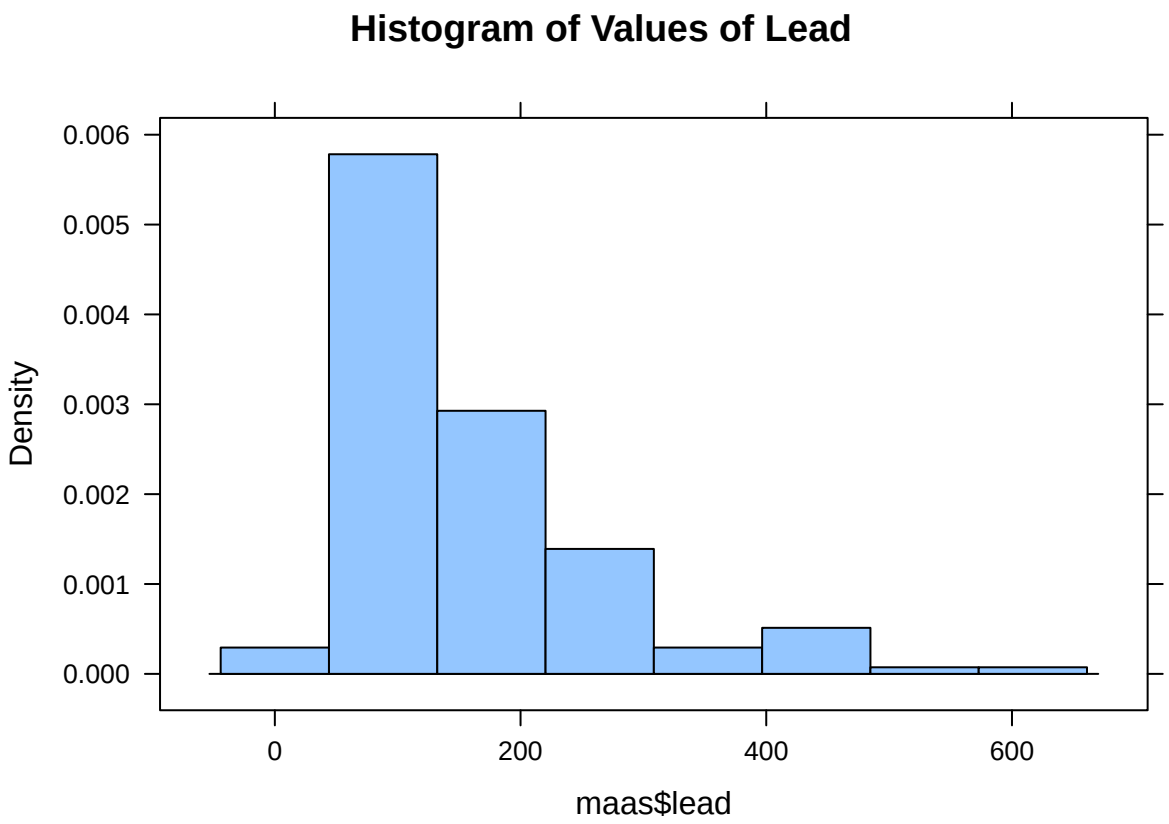
```
summary(maas$zinc)
```

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

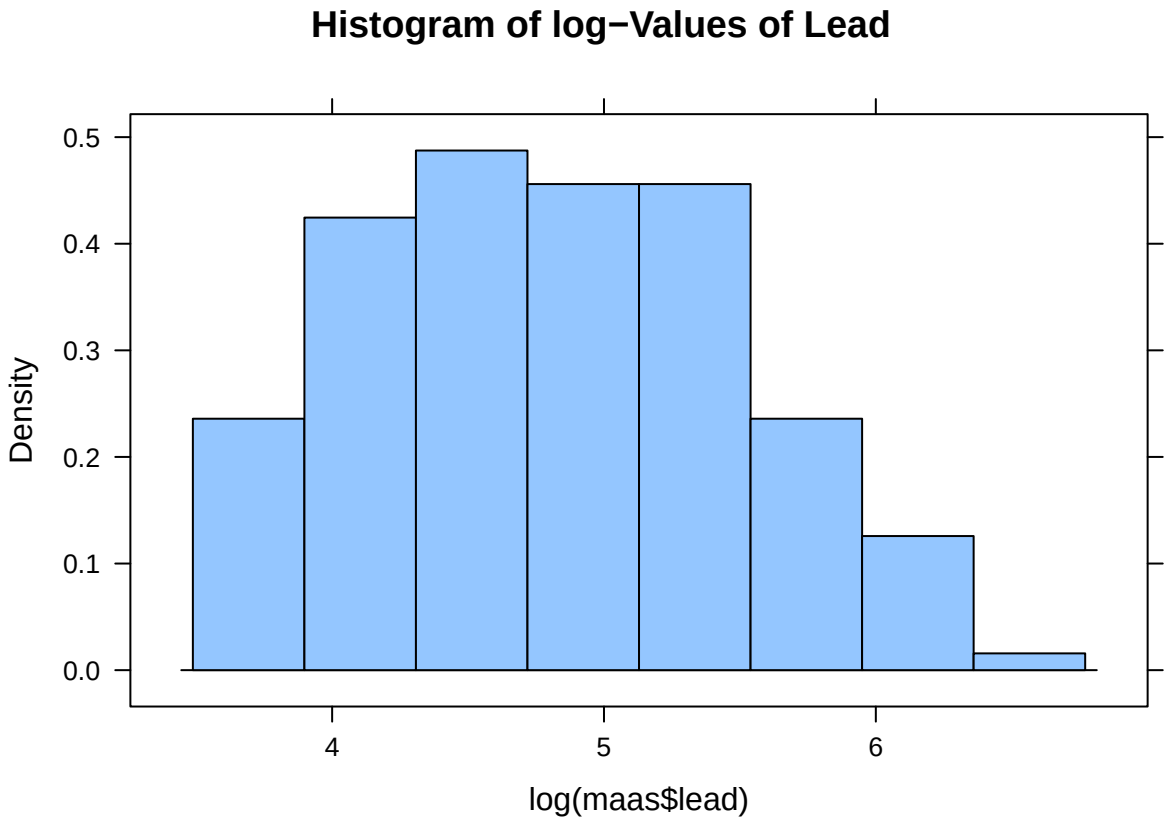
b. Plot two histograms: one of lead and one of log(lead).

Note: Titles not required.

```
histogram(maas$lead, main = "Histogram of Values of Lead")
```

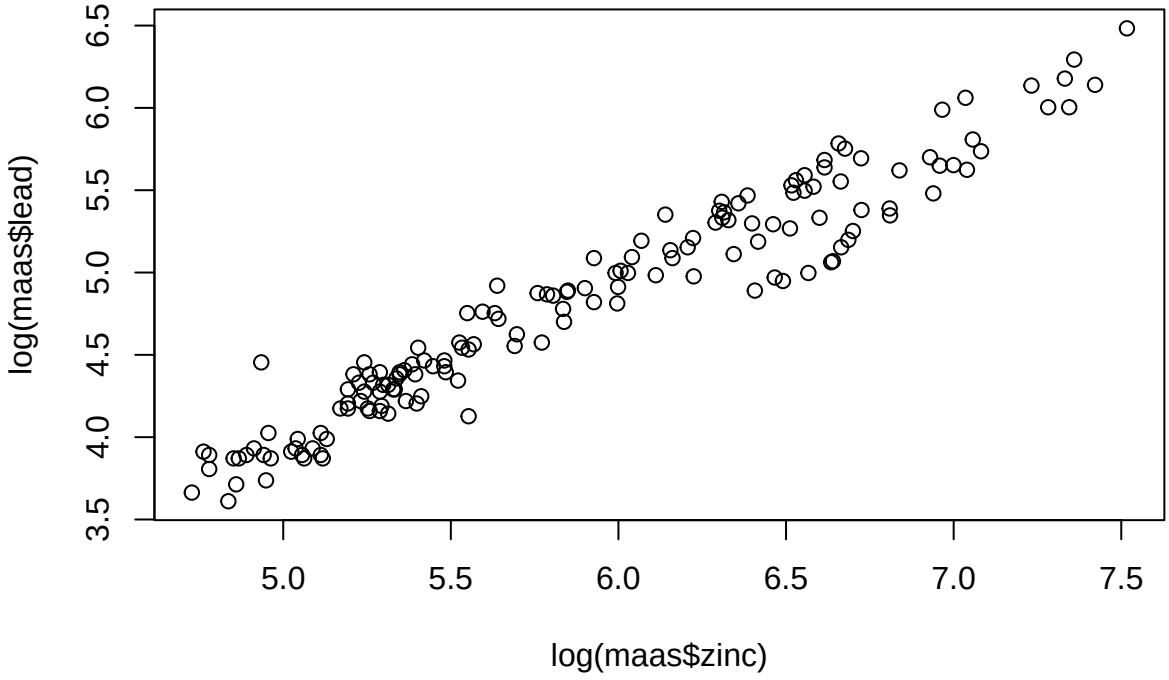



```
histogram(log(maas$lead), main = "Histogram of log-Values of Lead")
```



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

```
plot(log(maas$lead) ~ log(maas$zinc))
```




```
cor(log(maas$lead) ~ log(maas$zinc))
```

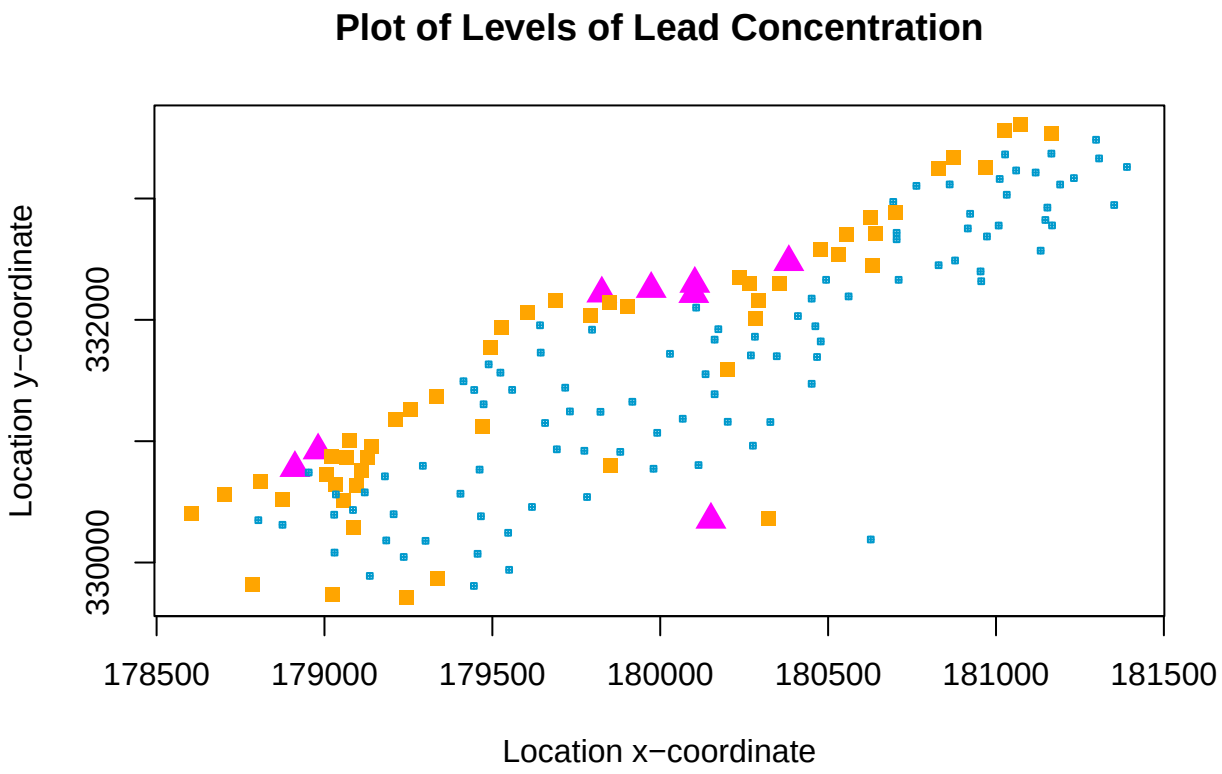
```
## [1] 0.9671621
```

- Direction trends positive/increasing.
- Linear form.
- Small amounts of vertical scatter so there appears to be a strong association.

d. The level of risk for surface soil based on lead concentration in ppm is given on the table in Assignment 2. Use similar techniques to the ones outlined in said assignment to give different colors and sizes to the lead concentration at these 155 locations.

Notes: Any color, point size, point style combination is fine. Point style itself isn't necessary to include but is here since it was in the original assignment file example code. The cut function in the "mylevels" object can take any upper bound value that encompasses the largest lead value i.e. 654. Plot title and labels are **not** required.

```
mycolors <- c("#0099CC", "orange", "#FF00FF") #can be changed to other colors
mylevels <- cut(maas$lead, c(0, 150, 400, max(maas$lead))) #the levels, can be changed to other values
mystyle <- c(12,15,17) #the point size, can be changed to other values
mysize <- c(.4, 1, 1.6)
plot(maas$x, maas$y, col=mycolors[as.numeric(mylevels)], pch= mystyle[as.numeric(mylevels)],
     cex = mysize[as.numeric((mylevels))], main = "Plot of Levels of Lead Concentration",
     xlab = "Location x-coordinate", ylab= "Location y-coordinate")
```



Exercise 4

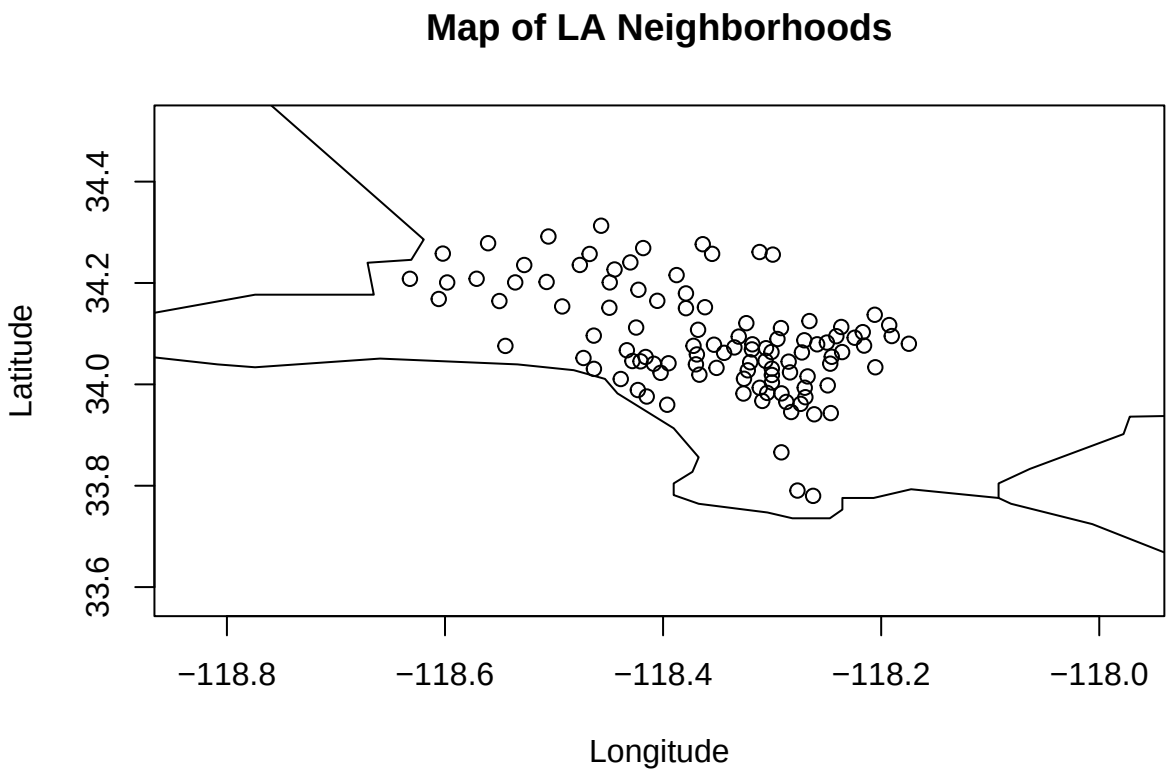
The data for this exercise represent approximately the centers (given by longitude and latitude) of each one of the City of Los Angeles neighborhoods. See also the Los Angeles Times project on the City of Los Angeles neighborhoods at: http://projects.latimes.com/mapping- la/neighborhoods/.

```
LA <- read.table("https://ucla.box.com/shared/static/d189x2gn5xfmcic0dmnhj2cw94jwvqpa.txt", header=TRUE
```

a. Plot the data point locations. Use good formatting for the axes and title. Then add the outline of LA County.

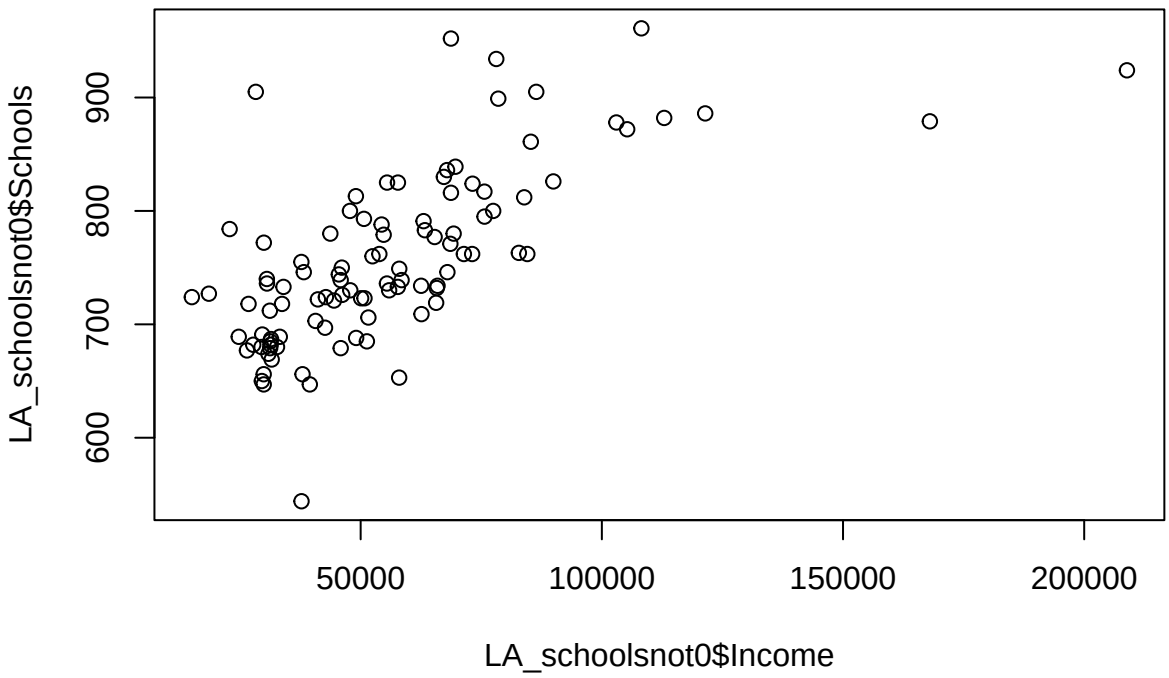
Note: Some form of axes scaling required because of "good formatting for the axes".

```
#Need maps package to run
plot(LA$Latitude ~ LA$Longitude, main = "Map of LA Neighborhoods", xlab = "Longitude", ylab = "Latitude
     xlim = c(min(LA$Longitude - .2), max(LA$Longitude + .2)),
     ylim = c(min(LA$Latitude - .2), max(LA$Latitude + .2)))
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_schoolsnot0 <- LA[LA$Schools != 0,]

plot(LA_schoolsnot0$Schools ~ LA_schoolsnot0$Income)
```



- Direction trends positive/increasing.
- Linear form, with two large outliers in the income bracket.

- Moderate amounts of vertical scatter so there appears to be a not-so-strong association.

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

```
customer_data <- read.csv("https://ucla.box.com/shared/static/y2y8rcie7mjw2h5t92x9dfcp133tc90h.csv")
```

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

Note: Loop not required; any code using is.na() and sum(is.na()) is accepted. Amount of missing values in each column not required, just the columns themselves.

```r
print(paste0("Missing values in data: ", sum(is.na(customer_data))))
```

```
## [1] "Missing values in data: 22"
```

```r
for (i in colnames(customer_data)){
  print(paste0("The variable, ", i, ", has ", sum(is.na(customer_data[[i]])), " missing values."))
}
```

```
## [1] "The variable, cust_id, has 0 missing values."
## [1] "The variable, age, has 10 missing values."
## [1] "The variable, gender, has 0 missing values."
## [1] "The variable, income, has 5 missing values."
## [1] "The variable, education, has 0 missing values."
## [1] "The variable, marital_status, has 0 missing values."
## [1] "The variable, purchase_amt, has 7 missing values."
```

22 total missing values. There are 10 missing values in the age column, 5 in the income column, and 7 in the purchase amount column.

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

Note: Credit for any code that shows effort in finding class of each variable.

```r
for (i in colnames(customer_data)){
  print(paste0("Variable: ", i, "; Class: ", class(customer_data[[i]])))
}
```

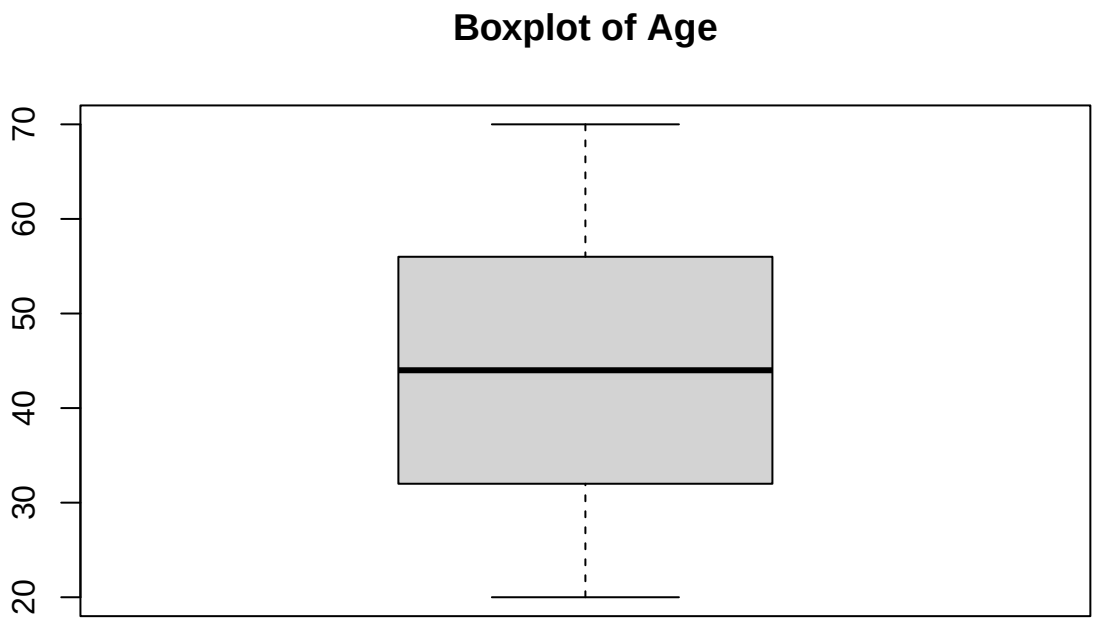
```
## [1] "Variable: cust_id; Class: character"
## [1] "Variable: age; Class: integer"
## [1] "Variable: gender; Class: character"
## [1] "Variable: income; Class: integer"
## [1] "Variable: education; Class: character"
## [1] "Variable: marital_status; Class: character"
## [1] "Variable: purchase_amt; Class: integer"
```

No classes should be changed. All quantifiable, numerical variables are already integer class and all identifying or categorical variables are already character class.

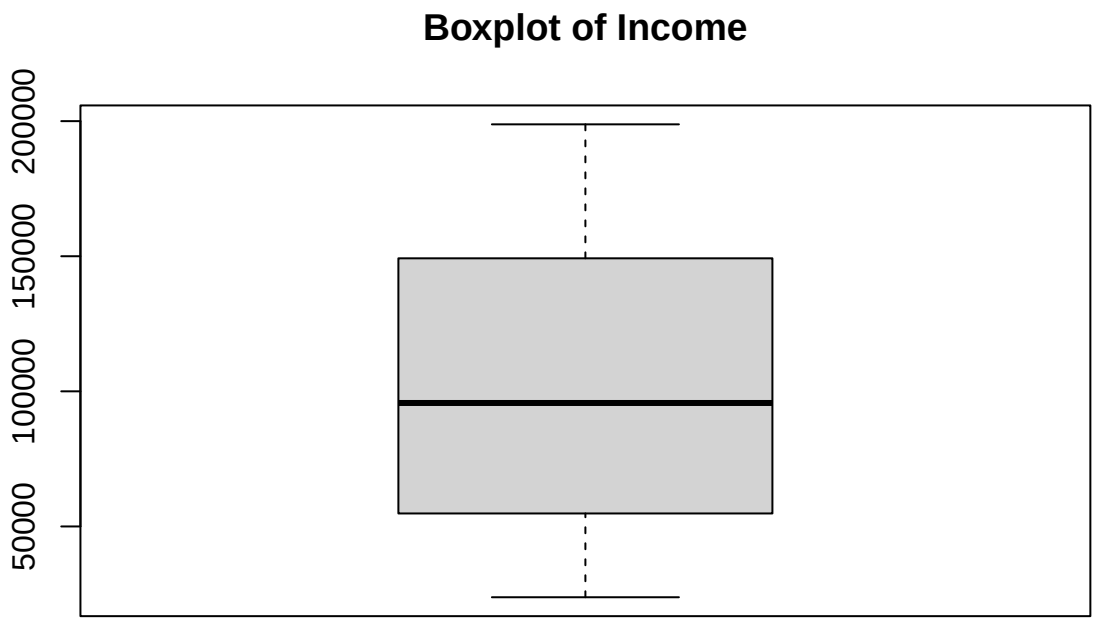
c. Do any numerical variables have outliers or extreme values? If so, how would you handle them? Provide your analysis in R for identifying outliers (e.g., visualization, numerical summary statistics). This is an open-ended question, so please feel free to use any appropriate methods to identify and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

Note: Creating data subset without NA values not required. For numerical methods, can also take Z-Score approach. At least one method and relevant conclusion needs to be mentioned and shown.

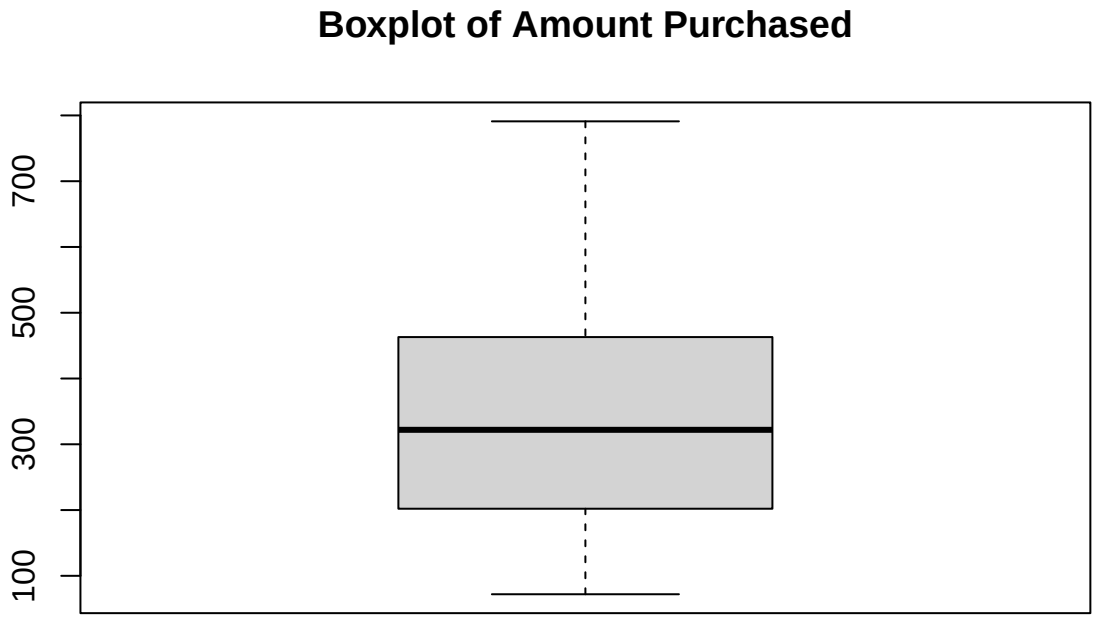
```r
# boxplot(customer_data$age)
# boxplot(customer_data$income)
# boxplot(customer_data$purchase_amt)

customer_data_noNA <- na.omit(customer_data)

boxplot(customer_data_noNA$age, main = "Boxplot of Age")
```



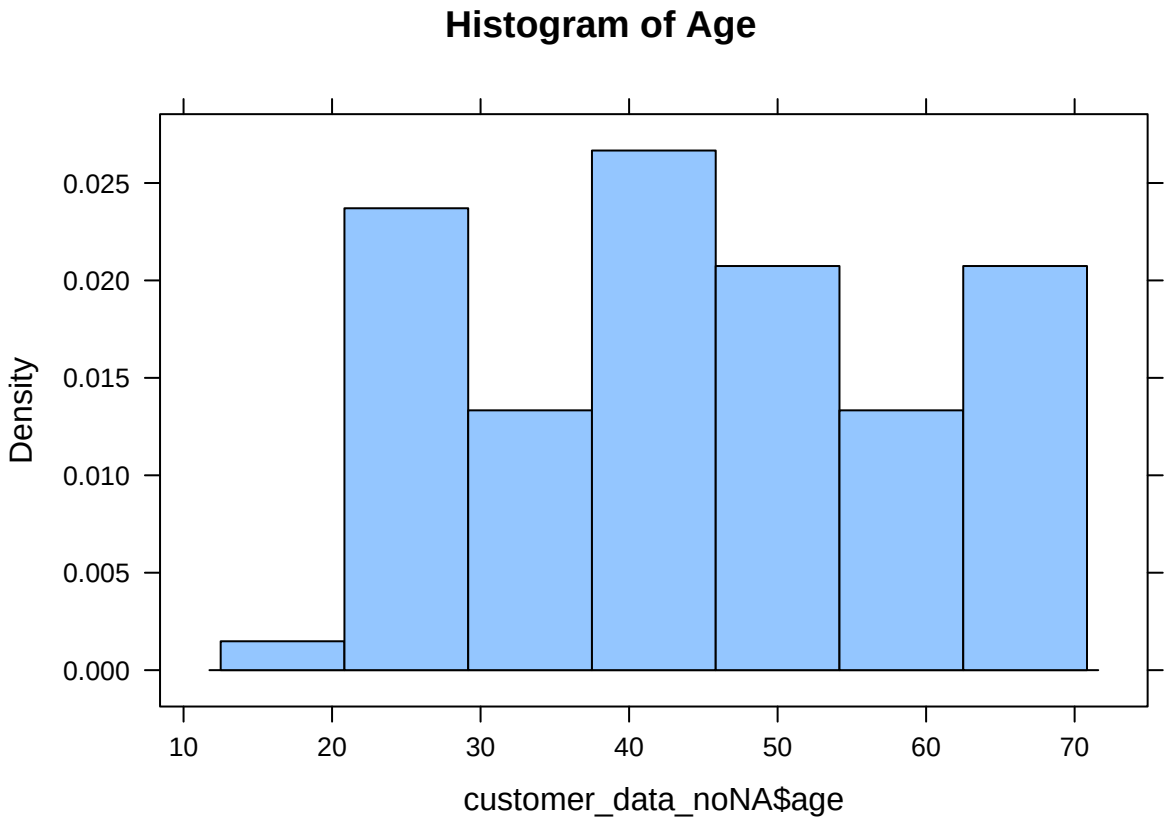

```r
boxplot(customer_data_noNA$income, main = "Boxplot of Income")
```



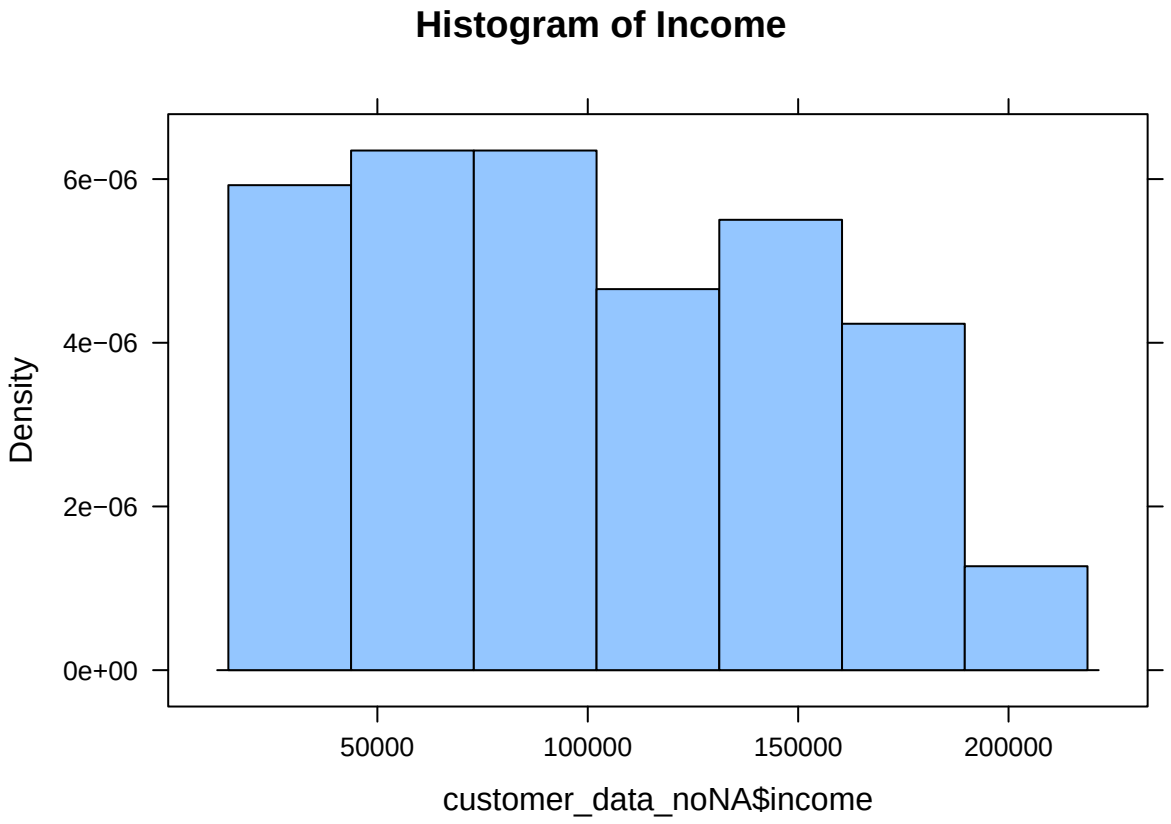

```
boxplot(customer_data_noNA$purchase_amt, main = "Boxplot of Amount Purchased")
```



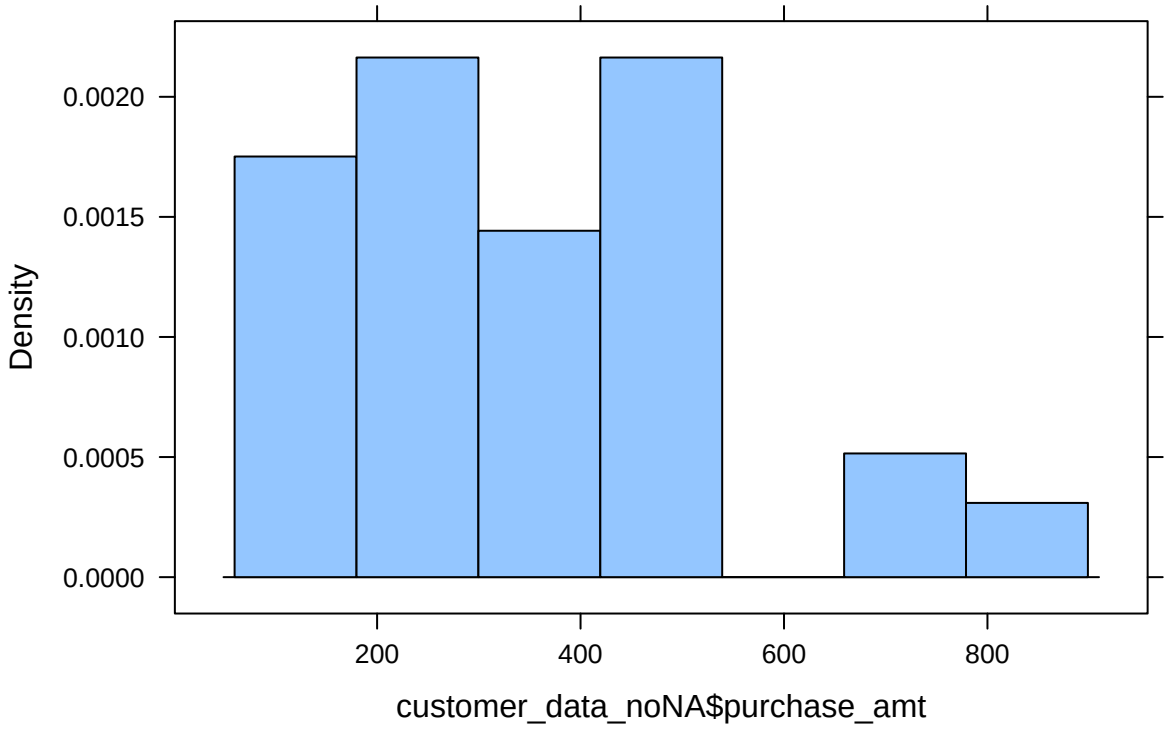

```r
histogram(customer_data_noNA$age, main = "Histogram of Age")
```




```
histogram(customer_data_noNA$income, main = "Histogram of Income")
```




```
histogram(customer_data_noNA$purchase_amt, main = "Histogram of Amount Purchased")
```




```
summary(customer_data_noNA$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.0    32.0    44.0    44.8    56.0    70.0
```

```
summary(customer_data_noNA$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23798   54804   95688  100683  149228  198808
```

```
summary(customer_data_noNA$purchase_amt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    72.0   202.0   322.0   345.7   463.0   791.0
```

Boxplots of our numerical variables show evidence that Age is unimodal, symmetric but both Income and purchase amount are slightly right skewed. Histograms of all 3 variables don't make a strong case for any of the 3 distributions being unimodal and symmetric. 5-number summaries show that there are no outliers. This conclusion can be shown by calculating the IQR for each variable and then checking $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$.

Part 2

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

Figure 1: Image1.

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

$$\frac{\text{Total responding "About Right"}}{\text{Total Participants}} = \frac{310 + 290}{900} = \frac{2}{3}$$

A majority of students feel positive about their body weight, so yes, in general the students are happy regarding this metric.

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)

A bar graph/ bar chart / grouped bar chart because we're examining the relationship between 2 categorical variables.

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

$$\frac{\text{Female students who feel "About Right"}}{\text{Total Female Students}} = \frac{310}{310 + 130 + 30} \approx 0.6596$$
$$\frac{\text{Male students who feel "About Right"}}{\text{Total Male Students}} = \frac{290}{290 + 68 + 72} \approx 0.6744$$

From the proportions, we see that the male students are more likely to feel "about right" than the female students.

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?)

$$\frac{\text{Female students who feel "Overweight"}}{\text{Total Female Students who don't feel "About Right"}} = \frac{130}{130+30} = 0.8125$$

$$\frac{\text{Male students who feel "Overweight"}}{\text{Total Male Students who don't feel "About Right"}} = \frac{68}{68+72} = 0.4857$$

From the proportions, we see that among the students who don't feel "about right" about their body image, an overwhelming majority of the female students see themselves as overweight where a little under half of the male students hold the same views of their body image. Conversely, a small proportion of the female students, i.e. $(1 - .8125)$, view themselves as underweight where a slight majority of male students , i.e. $(1 - .4857)$, view themselves in the same regard. **Note:** Only need to comment on one comparison, either overweight or 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.

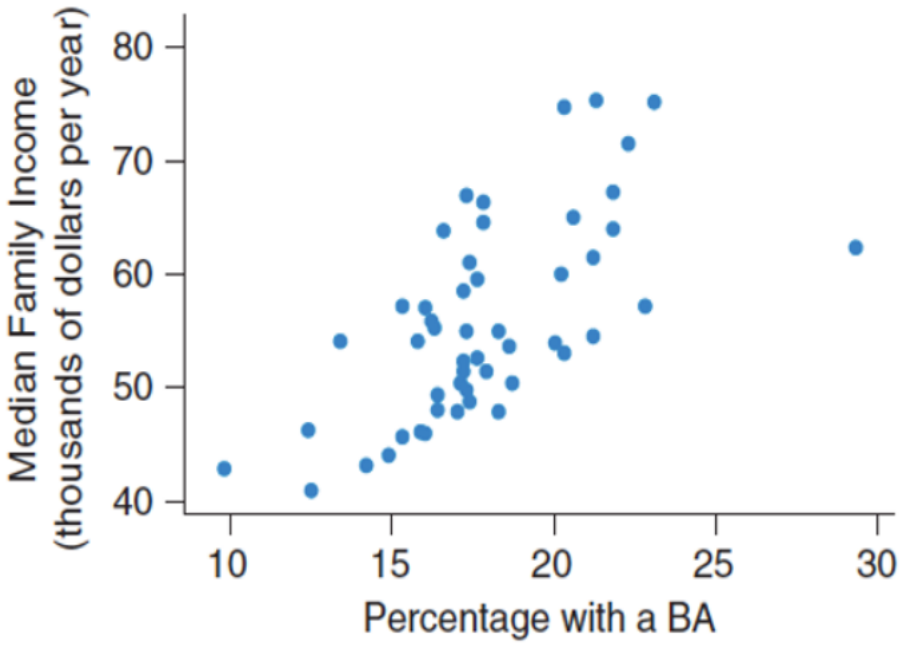


Figure 2: Image3.

- Positive/increasing direction.
- Linear form.
- Sizeable verticle scatter so weak strength.
- Possible outlier at the largest observed x-value.

Any explanation that surmises that median income is positively correlated with an increase in the amount of college educated persons in the 25 and older population. The outlier represents a break in this trend of

constant linear growth and there's perhaps a ceiling/upper limit that would be hit even as larger percentages of the population become college educated.

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).

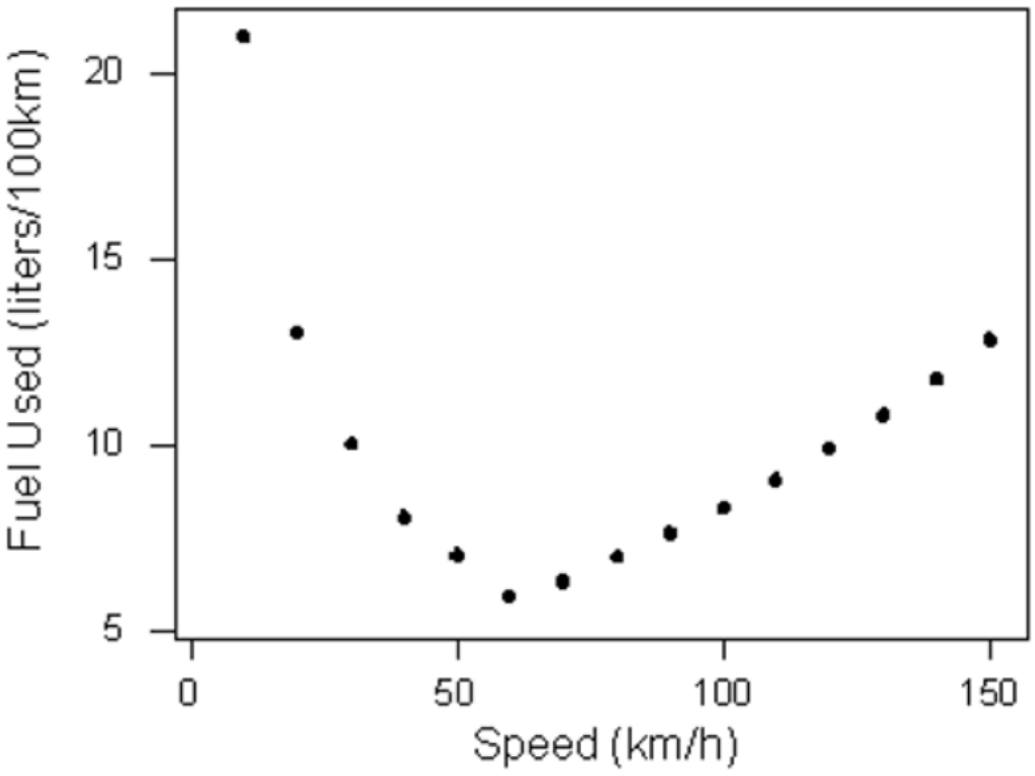


Figure 3: Image4.

- This is not a linear function.
- No comments necessary on direction because that depends on the derivative.
- Can mention that the shape is concave-up.
- No noticeable scatter, so there's is a strong non-linear relationship.
- There do not seem to be any outliers.

Any explanation that mentions how the relationship between fuel used and speed driven is **not** linear.

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?

Explanatory: Starting median salary Response: Mid-career median salary

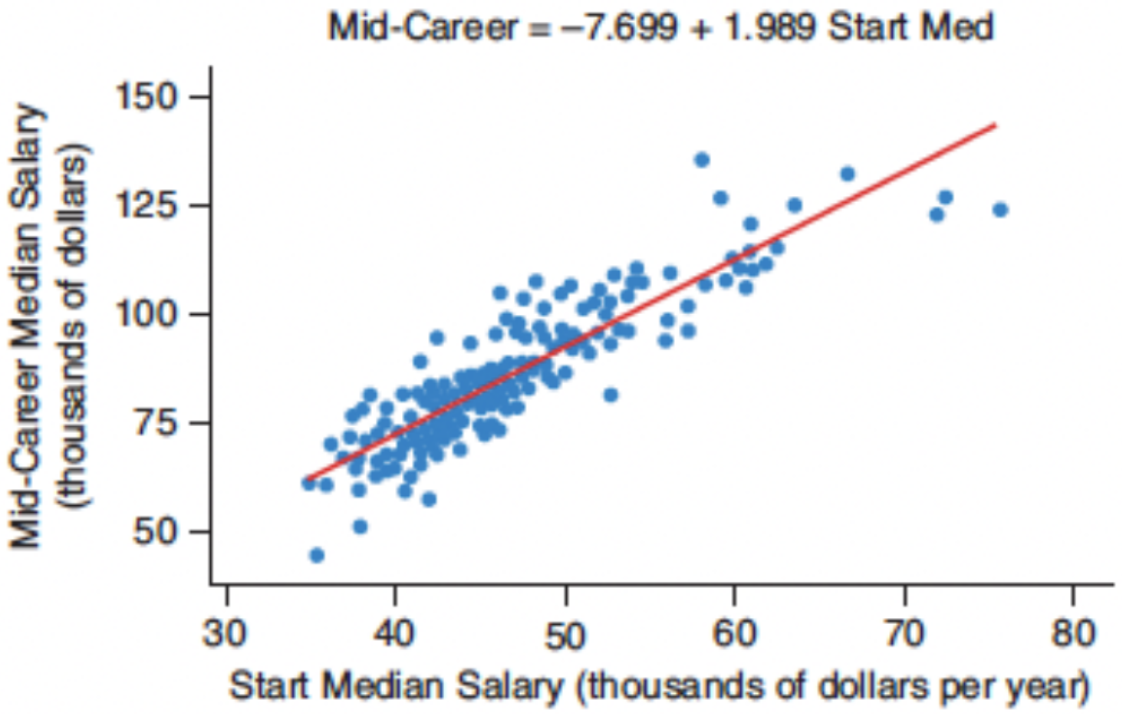


Figure 4: Image7.

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

There could be outliers with the starting salaries of employees on both the low-end and high-end. This could skew the mean value, so we use the median instead because it is not as sensitive to outliers.

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.

Yes, because a median starting salary of $$$60k is within the range of the data that is fit by the regression line. We can plug this value into the regression equation to get the expected mid-career salary e.g.

$$\begin{aligned} \text{Mid-career} &= -7.699 + 1.969 * \text{Start Med} \\ &= -7.699 + 1.969 * 60 \\ &\approx 110.44 \end{aligned}$$

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.

No, because a median starting salary of $$$100k is out of the range of the data within which our regression line was fit.

Problem 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.

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	

Figure 5: Image10.

a. Calculate slope and intercept of the regression line.

$$b = r\frac{s_y}{s_x} = .95 * \frac{46.34}{2.32} \approx 18.9754$$

$$a = \bar{y} - b\bar{x} = 141.67 - (18.9754 * 11.03) = -67.629$$

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

$$y = a + bx \rightarrow y = -67.629 + 18.9754x$$

Our regression line equation tells us that for every unit increase of percent alcohol content, an additional 18.9754 calories will be added to the wine serving.

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

Coefficient of determination: $r^2 = 0.95^2 = .9025$. Around 90% of the variation in the calories of the 5 ounce serving of wine is explained by the alcohol content of that wine.

\textbf{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)}

We essentially added a massive outlier to the data, so how will this effect the r^2 value and slope?

For r^2, the goodness of fit will be worse (so the r^2 value is lower) given that the line will no longer fit the data as well i.e. less of the variation of y is now explained by x.

For the slope, it depends where the value falls. An outlier below the data will cause the slope to shrink lower, whereas an outlier above the data will cause the slope to increase.

For 20% alc with 0 calories, we have a data point on the very far right which is also far below the data, which will pull the regression line down aka decrease slope.

Note: States that the slope and r^2 decreases and provides justification.

Problem 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.

Yes, this will cause undercoverage because the "talk therapy" group will no longer receive an important part of the sample.

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.

No, this is not acceptable because it will induce bias. This violates the principle of having a blinded study. A violation of this sort could cause the doctor to make decisions that bias the results of the experiment.

c. What improvements to the plan would you recommend?

Any study proposal that mentions meeting the key features of a controlled experiment 1) sample size, 2) random assignment, and 3) blinding. Placebo not necessary.