STATS 10 Assignment 2

Please submit both parts of the assignment in one single PDF file. You can use any PDF editor software to merge the two parts into one file. Please make sure that the questions are in the correct order and clearly labeled, and that the answers are legible and easy to read.
To submit your assignment, upload the PDF file under the designated assignment page on the course website before the deadline specified. Email or hard copy submissions are not accepted.

Part I

Include both the R commands and their corresponding outputs, results, or answers for all exercise questions in Part I.

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:

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

When you read in the data, name your object "flint".

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?

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

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.

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

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.

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. You can access these data in R using:

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

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

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

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.

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()*

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. You can read the data in R as follows:

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

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

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

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

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

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

The following commands give different colors and sizes on a scatterplot
For two variables: x, y

```
mycolors <- c("green", "orange", "red") #can be changed to other colors
mylevels <- cut(y, c(0, 100, 1000, 10000)) #the levels, can be changed to other values
mysize <- 19 #the point size, can be changed to other values
plot(x, y, col=colors[as.numeric(mylevels)], pch= mysize)
```

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

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/. You can access these data at:

```
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 by typing:

```
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.**

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

Load the data into R:

```
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?

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

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.

Part II

You may choose to type or write your answers electronically or scan your handwritten solutions. Please ensure that you show all steps and explanations to receive full credit, unless otherwise instructed.

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)

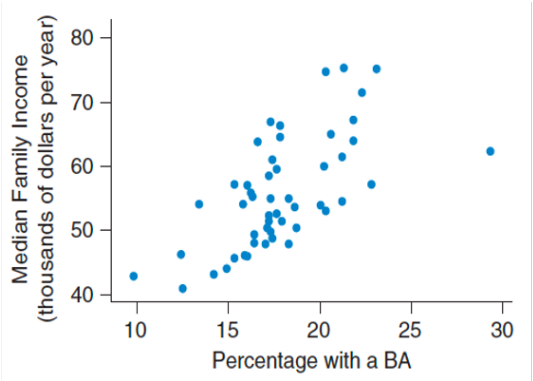
c. Are female students more likely to feel they are about right than male students? Explain with numerical evidence.

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?)

Exercise 2

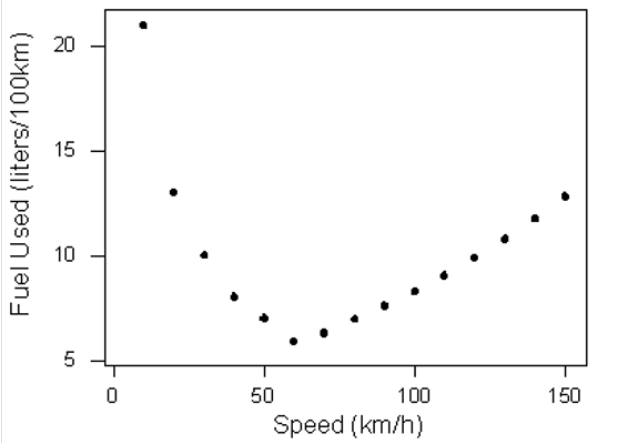
For each of the scatterplots shown, provide a written description that includes the direction, form, and strength of the relationship, along with any outliers that do not fit the general trend. In addition, explain what these characteristics mean in the context of the data.

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.



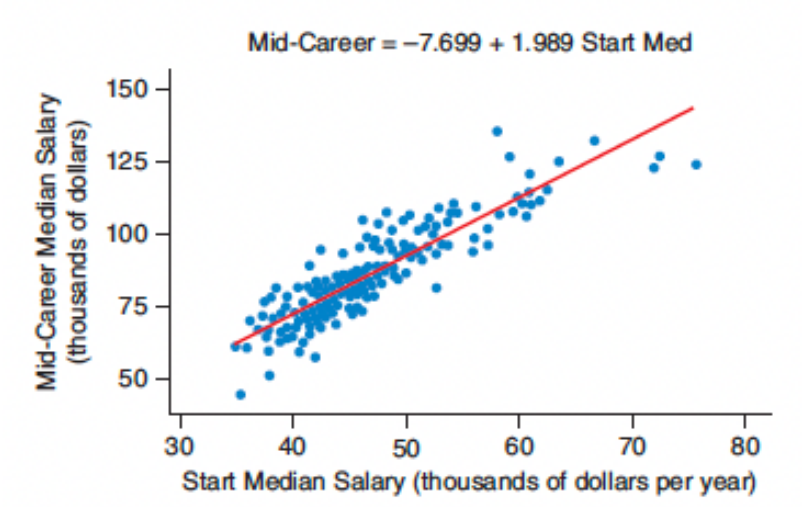
b. Consider the relationship between the average amount of fuel used (in liters) to drive a fixed distance in a car (100 km), and the speed at which the car is driven (in km per hour).

c.



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 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 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?