Name (Last, First):	Student ID:
Lopez Sepulveda,	U59702827

Assignment 5

METCS544A3A4_F2024

Instructions:

1. **This assignment has no specific R programming questions, but you're encouraged to use R to plot and graph the data and calculate relevant statistic summaries.**
2. For answering programming questions, please use Adobe Acrobat to edit the pdf file in two steps **[See Appendix: Example Question and Answer]**:
    a. Copy and paste your R code as text in the box provided (so that your teaching team can run your code);
    b. Screenshot your R console outputs, save them as a .PNG image file, and paste/insert them in the box provided.
    c. Show all work - credit will not be given for code without showing the code in action by including the screenshot of R console outputs.
3. To answer non-programming questions, please type or handwrite your final answers clearly in the boxes. Show all work - credit will not be given for numerical solutions that appear without explanation in the space above the boxes.
4. **[Total 60 pts = 57 + 3 Extra Credit pts]**

Grading Rubric

Each question is worth 3 points and will be graded as follows:

3 points: Correct answer with work shown

2 points: Incorrect answer but attempt shows some understanding (work shown)

1 point: Incorrect answer but an attempt was made (work shown), or **correct answer without explanation (work not shown)**

0 points: Left blank or made little to no effort/work not shown

Reflective Journal [3 pts]

(Copy and paste the link to your live Google doc in the box below)

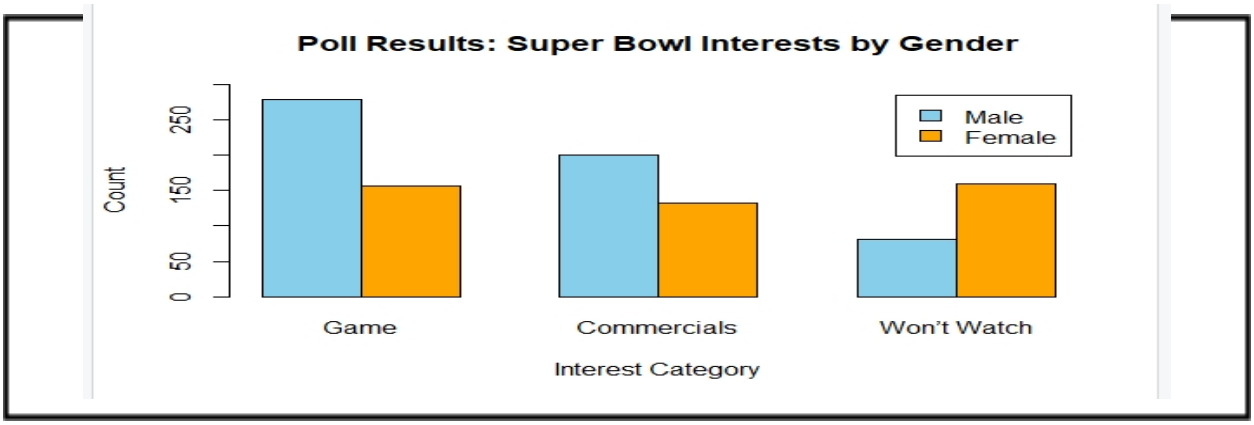
https://drive.google.com/drive/folders/1_8qcBjQVMfZggF42UYJuHQzMoBcyAy0Q?usp=drive_link

Part I. Exploring Two Categorical Variables (33 pts)

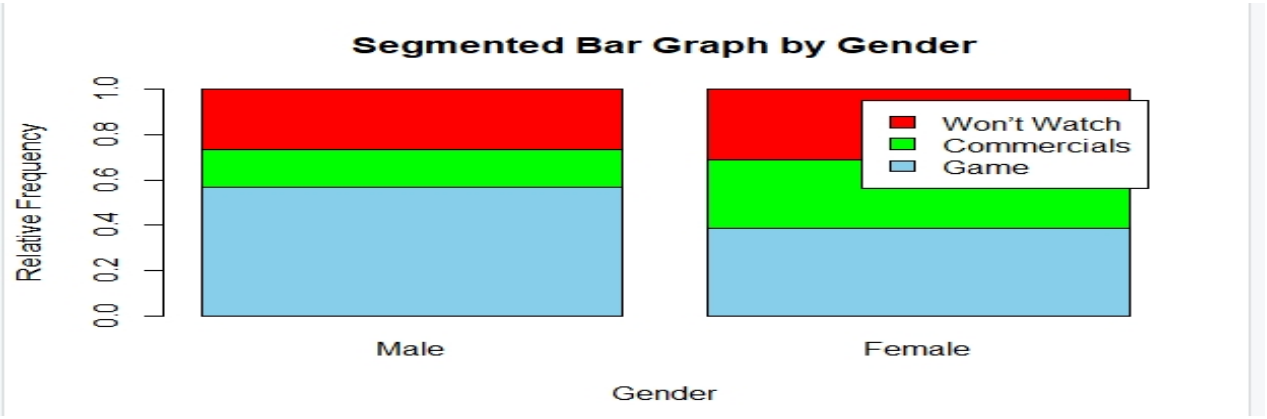
1) **(12 pts)** The table shows the results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

	Male	Female	Total
Game	279	200	479
Commercials	81	156	237
Won't Watch	132	160	292
Total	492	516	1008

a) Display the data with a side-by-side bar chart.



b) Display the data with a segmented bar graph (by making a relative frequency table first).



c) Based on your investigation, would you agree or disagree that there is an association between a person's sex and their interest in the Super Bowl?

Based on the graphs, It is clear a lot of women watch for the commercials and not the game itself.

2) **(9 pts)** A public opinion survey explored the relationship between age and support for increasing the minimum wage. The results are summarized in the two-way table below.

	Yes	*No*	*No Opinion*	*Total*
18 to 30	25	20	5	50
31 to 60	20	35	20	75
Over 60	55	15	5	75
Total	100	70	30	200

a) Give an example of a joint relative frequency and interpret it in context.

One example is in Yes from 18 to 30 is 25 which shows that 25 over 200 particpants are years 18 to

b) Give an example of a marginal relative frequency and interpret it in context.

The total number of people who voted yes over all age ranges from 18-65+ is 100. 100 over 200

c) Give an example of a conditional relative frequency and interpret it in context.

18-30 year olds that voted yes is only 25 out of 100 that voted yes among all age groups that voted

3) **(12 pts)** The table below shows the three most popular social media platforms and the distribution of the ages on those platforms. The data is based on MAU (monthly active users). A random sample of 100 active users from each platform was taken, and their age was recorded.

		Social Media Platform			
		Facebook	YouTube	WhatsApp	Total
Age Group	0 – 17	6	23	17	46
	18 – 26	28	40	18	86
	27 – 35	37	30	28	95
	36 – 49	23	5	17	45
	50 and older	6	2	20	28
		100	100	100	300

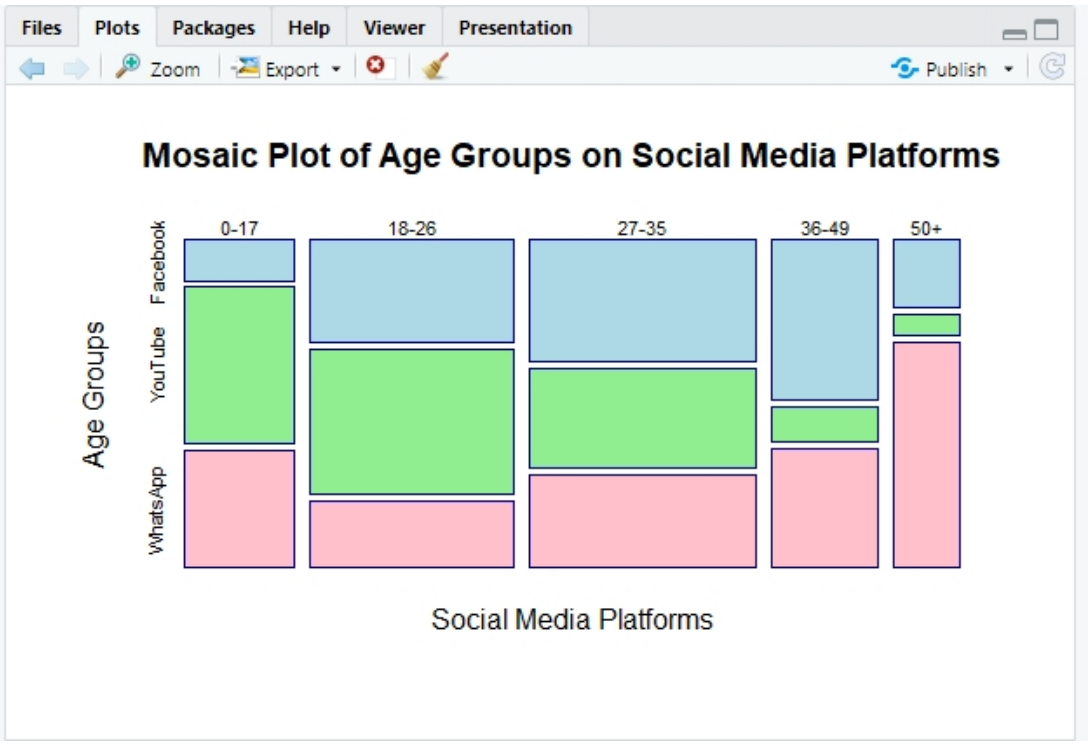
Create a mosaic plot of the data above. Show ALL work and calculations used to create the plot.

Answer:

```
# Input data as a matrix
age_data &lt;- matrix(c(6, 23, 17,  # 0-17 age group
                     28, 40, 18, # 18-26 age group
                     37, 30, 28, # 27-35 age group
                     23, 5, 17,  # 36-49 age group
                     6, 2, 20),  # 50 and older
                   nrow = 5, byrow = TRUE)


# Assign row and column names
rownames(age_data) &lt;- c("0-17", "18-26", "27-35", "36-49", "50+")
colnames(age_data) &lt;- c("Facebook", "YouTube", "WhatsApp")


# Create a mosaic plot
mosaicplot(age_data,
           main = "Mosaic Plot of Age Groups on Social Media Platforms",
           xlab = "Social Media Platforms",
           ylab = "Age Groups",
           col = c("lightblue", "lightgreen", "pink", "orange", "yellow"),
           border = "darkblue")
```



Part II. Exploring Two Variable Data: Scatterplots and Correlation (24 pts)

1) **(12 pts)** A student wonders if tall women tend to date taller men. She measures herself, her dormitory roommate, and the women in the adjoining rooms; then, she measures the next man each woman dates. Here are the data (heights in inches):

Women	66	64	66	65	70	65
Men	72	68	70	68	71	65

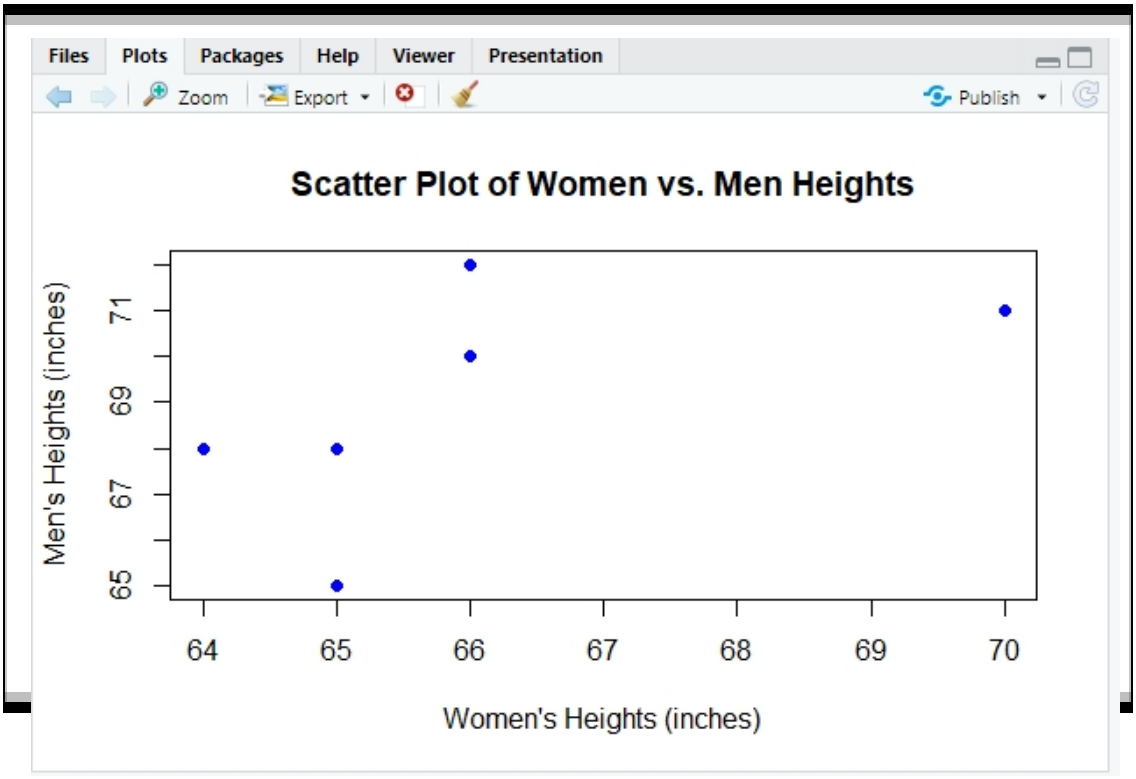
a) Is there a clear explanatory variable and response variable in this setting? If so, tell which is which. If not, explain why not.

Answer:

There is no clear explanatory variable as it is unclear if the women are pursuing men of a certain height based on their height alone or there are other variables or the men pursuing

b) Make a well-labeled scatterplot of these data.

Answer:



c) Based on the scatterplot, describe the pattern, if any, in the relationship between the heights of women and the heights of the men they date.

Answer: There seems to be little to no relationship between the two.

d) Suppose another 70-inch-tall female who dated a 73-in-tall male were added to the data set. How would this influence ***r***?

Answer: It would make r closer to 1 but not exactly reach 1.

D

2) From tax records, it is relatively easy to determine the amount of liquor consumed per capita and the number of cigarettes consumed per capita for each of the 10 provinces of Canada. These are plotted on a scatterplot and a high positive correlation is found. Which of the following is correct?

(A) This implies that heavy smoking causes people to drink more.
(B) This implies that heavy drinking causes people to smoke more.
(C) We cannot conclude cause and effect, but this also implies that there is a high positive correlation between cigarette smoking and alcohol consumption for individuals.
(D) This could be an example of a correlation caused by a common cause because both activities are highly correlated with average family income and average income varies widely among the provinces.
(E) We cannot conclude cause and effect, but this also implies that the same individuals both smoke and consume liquor.

C

3) Suppose a study finds that the correlation coefficient relating family income to SAT scores is r = +1. Which of the following are proper conclusions?

I. Poverty causes low SAT scores.
II. Wealth causes high SAT scores.
III. There is a very strong association between family income and SAT scores.

(A) I only
(B) II only
(C) III only
(D) I and II
(E) I, II and III

A

4) An agricultural economist says that the correlation between corn prices and soybean prices is r = 0.7. This means that

(A) when corn prices are above average, soybean prices also tend to be above average.
(B) there is almost no relation between corn prices and soybean prices.
(C) when corn prices are above average, soybean prices tend to be below average.
(D) when soybean prices go up by 1 dollar, corn prices go up by 70 cents.
(E) the economist is confused, because correlation makes no sense in this situation.

C

5) If data set A of (x, y) data has correlation r = 0.65, and a second data set B has correlation r = –0.65, then

(A) the points in A fall closer to a linear pattern than the points in B.
(B) the points in B fall closer to a linear pattern than the points in A.
(C) A and B are similar in the extent to which they display a linear pattern.
(D) you can't tell which data set displays a stronger linear pattern without seeing the scatterplots.
(E) a mistake has been made—r cannot be negative.

THE END