Name (Last, First):
Kunjanda
Sushma Arun

Student ID:
002473132

Assignment 4

NEU_COE_INFO6105_Fall2024

Instructions:

1. 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 or python code as text in the box provided (so that your teaching team can run your code);
   b. Screenshot your R or python 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 or python console outputs.
2. 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. **You're encouraged to use R or python to graph/plot the data and produce numerical summaries; please** append **your code and screenshot of the outputs at the end of your pdf submission.**
3. **[Total 111 pts = 21 + 12 + 54 pts + 24 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://docs.google.com/document/d/1ptEhnYHniNtT1yxDPcvXK7LpJaPGzi80BCSGZZhom7Y/edit?usp=sharing

Part I. Exploring One Variable Data: The Empirical Rule and Z-Scores (21 pts)

1. **(3 pts)** with z-scores above 2.5 on an IQ test are sometimes classified as geniuses. If IQ scores have a mean of 100 and a standard deviation of 15 points, what IQ score do you need to be considered a genius?

Answer:

$X = ?$, $z = 2.5$, $\mu = 100$ $\sigma = 15$
(Z-score above 2.5 on IQ test = genius)

$X = \mu + z\sigma$

$X = 100 + 2.5 * 15$

$= 100 + 37.5$

$= 137.5$

An IQ score higher than 137.5 would be considered a genius. ie., (IQ score >137.5)

2. **(6 pts)** An incoming freshman took her college's placement exams in French and mathematics. In French, she scored 82 and in math 86. The overall results on the French exam had a mean of 72 and a standard deviation of 8, while the mean math score was 68, with a standard deviation of 12. On which exam did she do better compared with the other freshmen?

Answer: Z score formula: $z = \frac{X - \mu}{\sigma}$

For French:

$X = 82$, $\mu = 72$, $\sigma = 8$

$z = \frac{82-72}{8} = \frac{10}{8} = 1.25$

For Math:

$X = 86$, $\mu = 68$, $\sigma = 12$

$z = \frac{86-68}{12} = \frac{18}{12} = 1.5$

Z score for French = 1.25
z score for Math = 1.5
Since Z score for Math is higher, she did better in Math compared to other freshmen.

3. **(12 pts)** A company's customer service hotline handles many calls relating to orders, refunds, and other issues. The company's records indicate that the median length of calls to the hotline is 4.4 minutes with an IQR of 2.3 minutes.
Hint: These questions are asking you what happens when shifting and scaling the data.

a) If the company were to describe the duration of these calls in seconds instead of minutes, what would the median and IQR be?

Answer: Median length of calls = 4.4 minutes
IQR = 2.3 minutes
→Convert data from minutes to seconds.

Median = $4.4\,\text{min} \times 60 \frac{\text{sec}}{\text{min}}$
= 264 seconds.

IQR = $2.3\,\text{min} \times 60 \frac{\text{Sec}}{\text{min}}$
= 138 seconds

Median = 264 seconds.
IQR = 138 seconds

b) In an effort to speed up the customer service process, the company decided to streamline the series of push button menus customers must navigate, cutting the time by 24 seconds. What will the median and IQR of the length of hotline calls become?

Answer: • When we subtract a constant value from all data points, the median will decrease by that amount.
New Median = 264 − 24 = 240 Seconds.

• IQR will remain same, because IQR measures spread which is not affected by shifting all values by a constant.
IQR = 138 seconds.

After reducing time by 24 seconds }
Median = 240 seconds.
IQR = 138 seconds (unchanged).

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

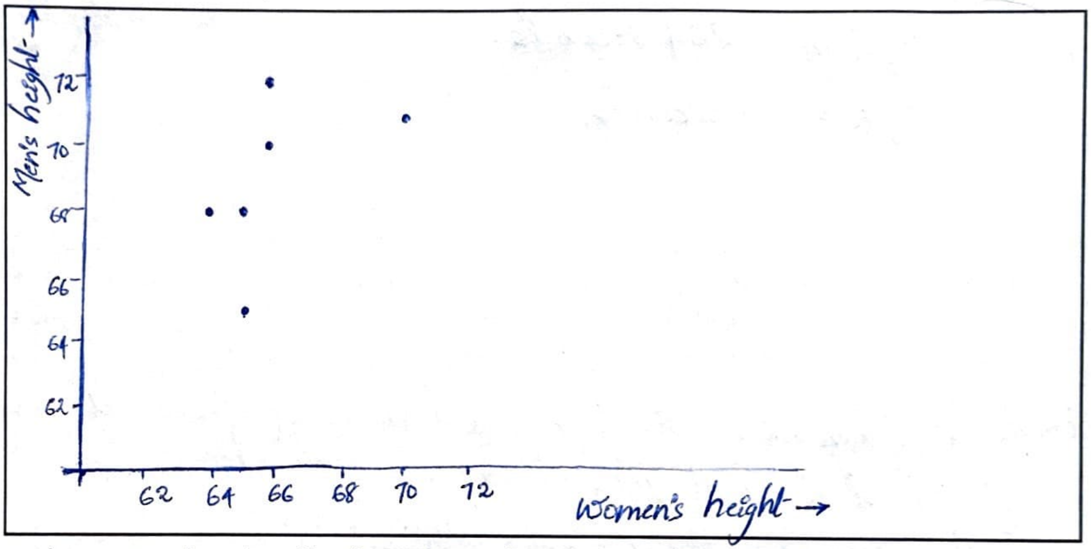
Women	66	64	66	65	70	65
Men	72	68	70	68	71	65

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: - Explanatory variable is the height of the women (since student is curious about how the height of the women affects their partner's height)
- Response variable is the height of the men (since we are observing how men's heights vary based on the women's height).

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.

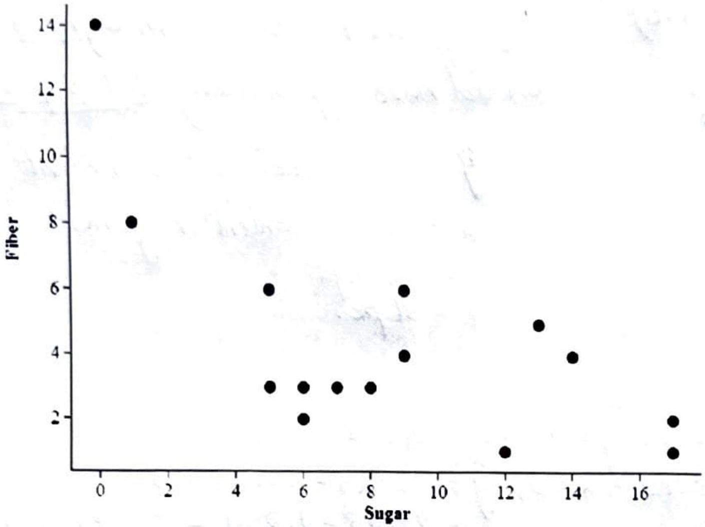
Answer: Based on data shown in scatterplot there is a weak positive association (i.e, weak positive linear correlation) between the heights of women & men they date. The taller women tend to date taller men, but relationship is not very strong or not perfectly linear as there are some variations.

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: Adding the new data point (70, 73) would likely increase/strengthen correlation coefficient r, as it fits the general trend of taller women dating taller men, strengthening the linear relationship (as the point is above the current line trend).

Part III. Exploring Two Variable Data: Linear Regression (54 pts)

1. Fiber helps regulate the body's use of sugars, helping to keep hunger and blood sugar in check. In children's cereal, which is usually loaded with large amounts of sugar, scientists wanted to investigate if a larger sugar content led to less nutritional value overall, such as lower fiber. A scatter plot of 15 randomly selected children's cereals, along with some selected summary statistics, are given below. Both sugar and fiber are measured in grams per serving.



Sugar	Fiber
$\bar{x} = 8.6$	$\bar{y} = 4.33$
$S_x = 5.18$	$S_y = 3.31$
$r = -0.654$	

a) Does the scatterplot indicate that it is okay to create a linear model? Explain.

Answer:
- Yes, it appears appropriate to create a linear model. The scatterplot shows a generally linear trend with points roughly following a downward sloping pattern.
- There is a clear negative correlation between sugar & fiber, as sugar content - increases the fiber content tends to decrease.

b) Find the slope of the least-squares regression line. Interpret this value in context.

Answer:

$$Slope = \frac{r.\,S_y}{S_x} \qquad r = -0.654 \quad S_y = 3.31 \quad S_x = 5.18$$

Slope of the least squares regression line. ie, $Slope = \frac{-0.654 \times 3.31}{5.81}$

$$= \frac{-2.16474}{5.81}$$

$$= -0.4179034749$$

$$\simeq -0.418 \text{ grams.}$$

Interpretation: For every additional gram of sugar, the fiber content decreases by approximately = 0.418 grams

c) What point must be on the least-squares regression line?

Answer: The point that must be on the least square regression line is the mean point $(\bar{x}, \bar{y})$ which is: $(\bar{x}, \bar{y}) = (8.6, 4.33)$

This is the "center" of the data and is always on the line of best fit.

d) Find the intercept of the least-squares regression line. Interpret this value in context.

Answer: $a = \bar{y} - b * \bar{x}$ b (Slope calculated earlier) $= -0.418$, $\bar{y} = 4.33$ $\bar{x} = 8.6$

$a = 4.33 - (-0.418 \times 8.6)$

$= 4.33 + 3.5948$

$= 7.9248 \simeq 7.925$

Interception: A cereal with 0grams suagar would have approximately 7.925 grams of fiber.

e) Write the equation of the linear model.

Answer: $\hat{y} = a + bx$

$\hat{y} = 7.925 - 0.418x$

$\hat{y}$ = predicted fiber content

x = sugar content in grams.

a = intercept

b = slope.

f) If you pick up a cereal and see that it has 3 grams of sugar, what is the predicted fiber content?

Answer: Equation of linear model: $\hat{y} = 7.925 - 0.418x$

Substitute $x = 3$, $\hat{y} = 7.925 - 0.418 \times 3$

$= 7.925 - 1.254 = 6.671$ grams

For 3 grams of sugar predicted fiber content approximately $= 6.67$ grams

g) What is the residual for the cereal with 9 grams of sugar and 6 grams of fiber?

Answer: To Calculate residual, we need predicted fiber gram for 9 grams of Sugar.

$\hat{y} = 7.925 - 0.418 \times 9 = 7.925 - 3.762 = 4.163$ gm

Residual = Observed fiber − predicted fiber $= 6 - 4.163 = 1.837$ gm

The residual is 1.837 grams of fiber.

h) Calculate the value of R^2 and interpret this value in context.

Answer: R^2 is the square of the correlation coefficient (r).

R^2 represents the proportion of the variance in the dependent variable (fiber) That is predicted from the independent variable (sugar).

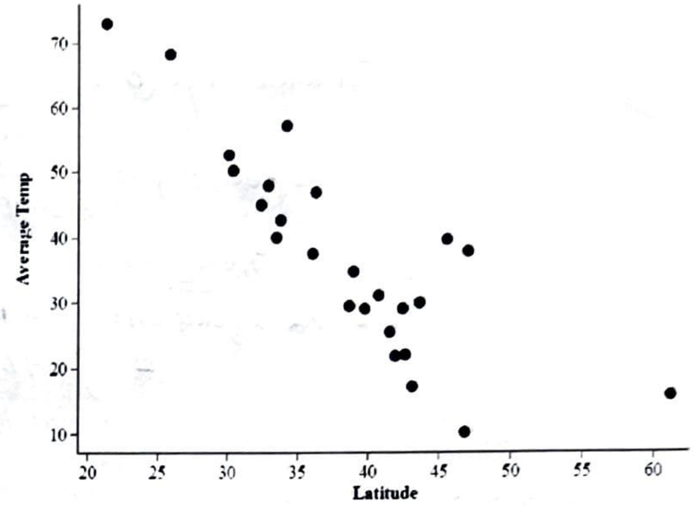
$$R^2 = (-0.654)^2 = 0.428$$

Interpretation: Approximately 42.8% of the variation in fiber content can be explained by the linear relationship with sugar content in these childrens cereals.

This indicates a moderate strength of linear relationship between sugar and fiber content.

2. We know that the further you get from the equator, the colder the climate becomes. For the following cities, we have assembled their latitude (angular distance from the equator, measured in degrees) and their average temperature in December 2021. The data and the scatterplot are given below.

City	Latitude (°N)	Dec Average Temp (°F)
Albany, NY	42.6	22.2
Anchorage, AK	61.2	15.8
Atlanta, GA	33.7	42.7
Austin, TX	30.3	50.2
Bismarck, ND	46.8	10.2
Boise, ID	43.6	30.2
Boston, MA	42.4	29.3
Charleston, SC	32.8	47.9
Chicago, IL	41.9	22
Cleveland, OH	41.5	25.7
Denver, CO	39.7	29.2
Honolulu, HI	21.3	73
Jackson, MS	32.3	45
Knoxville, TN	36	37.6
Las Vegas, NV	36.2	47
Los Angeles, CA	34.1	57.1
Madison, WI	43.1	17.3
Miami, FL	25.8	68.1
Newark, NJ	40.7	31.3
New Orleans, LA	30	52.6
Olympia, WA	47	38.1
Portland, OR	45.5	39.9
Roswell, NM	33.4	40
St. Louis, MO	38.6	29.6
Washington, DC	38.9	34.9



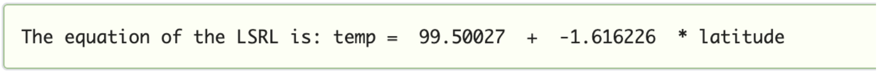
a) Describe the relationship between latitude and average December temperature in the US.

Answer:

There is a negative correlation between latitude and average December temperature. As latitude increases (further from the equator), the temperature decreases.

b) Find the LSRL using R.

Answer:



```
The equation of the LSRL is: temp =  99.50027  +  -1.616226  * latitude
```

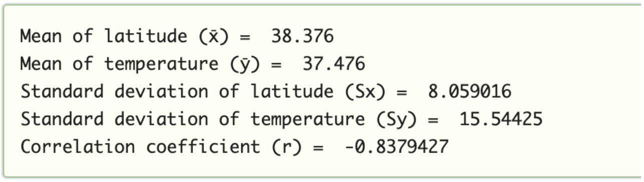
Answer:

b.

```
# Data
latitude <- c(42.6, 61.2, 33.7, 30.3, 46.8, 43.6, 42.4, 32.8, 41.9, 41.5,
              39.7, 21.3, 32.3, 36, 36.2, 34.1, 43.1, 25.8, 40.7, 30, 47,
              45.5, 33.4, 38.6, 38.9)
temp <- c(22.2, 15.8, 42.7, 50.2, 10.2, 30.2, 29.3, 47.9, 22, 25.7, 29.2,
          73, 45, 37.6, 47, 57.1, 17.3, 68.1, 31.3, 52.6, 38.1, 39.9, 40,
          29.6, 34.9)
# Linear model
model <- lm(temp ~ latitude)
# Extract the coefficients (Intercept and Slope)
intercept <- coef(model)[1]
slope <- coef(model)[2]
# Print the LSRL equation
cat("The equation of the LSRL is: temp = ", intercept, " + ", slope, " * latitude\n")
```

c.

```
# Data
latitude <- c(42.6, 61.2, 33.7, 30.3, 46.8, 43.6, 42.4, 32.8, 41.9, 41.5,
              39.7, 21.3, 32.3, 36, 36.2, 34.1, 43.1, 25.8, 40.7, 30, 47,
              45.5, 33.4, 38.6, 38.9)
temp <- c(22.2, 15.8, 42.7, 50.2, 10.2, 30.2, 29.3, 47.9, 22, 25.7, 29.2,
          73, 45, 37.6, 47, 57.1, 17.3, 68.1, 31.3, 52.6, 38.1, 39.9, 40,
          29.6, 34.9)
# Mean of latitude and temp
mean_latitude <- mean(latitude)
mean_temp <- mean(temp)
# Standard deviation of latitude and temp
sd_latitude <- sd(latitude)
sd_temp <- sd(temp)
# Correlation coefficient between latitude and temp
r <- cor(latitude, temp)
# Print summary statistics
cat("Mean of latitude (x̄ ) = ", mean_latitude, "\n")
cat("Mean of temperature (ȳ) = ", mean_temp, "\n")
cat("Standard deviation of latitude (Sx) = ", sd_latitude, "\n")
cat("Standard deviation of temperature (Sy) = ", sd_temp, "\n")
cat("Correlation coefficient (r) = ", r, "\n")
```



```
Mean of latitude (x̄) =  38.376
Mean of temperature (ȳ) =  37.476
Standard deviation of latitude (Sx) =  8.059016
Standard deviation of temperature (Sy) =  15.54425
Correlation coefficient (r) =  -0.8379427
```

c) Find the following summary statistics using R:

$\bar{x}$ = 38.376 °N $\bar{y}$ = 37.476 °F s_x = 8.059016 °N s_y = 15.54425 °F r = -0.8379427

d) Using the summary statistics from (c), show that the LSRL you find matches what you found in (b).

Answer: Calculate Slope (b) = $r \cdot \frac{S_y}{S_x}$

$= (-0.8379427) \cdot \frac{15.54425}{8.059016}$

$= -1.61622595$

Calculate Intercept (a) = $\bar{y} - b \cdot \bar{x}$

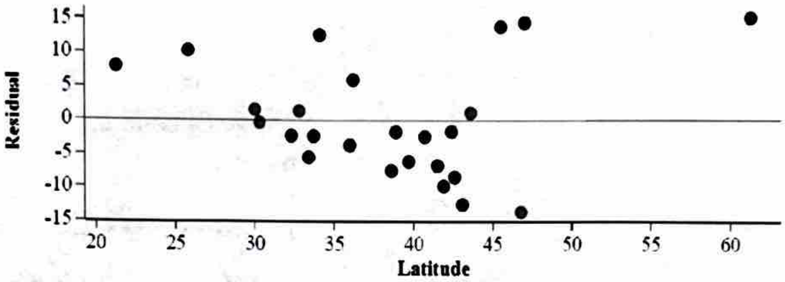
$= 37.476 - (-1.61622595)(38.376)$

$= 99.500287$

LSRL Equation: $\hat{y} = a + b \cdot x$

$\simeq \hat{y} = 99.5002 - 1.6162 * x$

e) Below is the residual plot for latitude and average December temperature.



Given the residual plot above, is a linear regression model appropriate for this data set? Explain.

Answer: The residual plot appears to show no clear pattern, with points scattered around the zero line. This suggets that a linear model is appropriate for this dataset. If there is a pattern in the residuals, it indicates that a non-linear model might be appropriate.

f) Kansas City, MO is located at 39.1°N and had an average December 2021 temperature of 42°F. What is the residual for this point?

Answer: LSRL equation: $\hat{y} = 99.5002 - 1.6162 * x$

Actual temp = 42°

where $x = 39.1$ (latitude of Kansas City)

$\hat{y} = 99.5002 - 1.6162 * 39.1$

$= 99.5002 - 63.19342$

$= 36.30678$

$\simeq 36.30°F$

Residual = Actual Temperature - Predicted Temperature.

$= 42 - 36.30$

$\simeq 5.7°F$

Part IV. Extra Credit Questions (21 pts)

(1) **(6 pts)** To analyze the social media behavior differences between boys and girls, Mr. P's Statistics class was asked to count the number of text messages that they sent over a three-day weekend. The following table summarizes the data:

	Values under Q_1	Q_1	Median	Q_3	Values over Q_3
Females	15, 43, 100	130	175	358	450, 573, 1098
Males	3, 59	72	183	273	293, 337

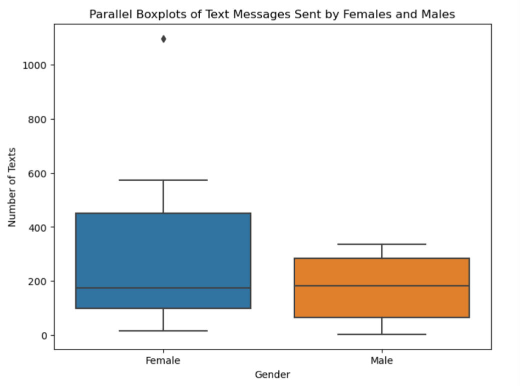
a. Construct parallel boxplots of this set of data.

b. Do the data indicate that females or males had the greater mean number of texts? Explain in detail (Shape, Outliers, Center, Spread; Conclusion).

Answer:

a.

```
import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

 # Data for females and males

females = [15, 43, 100, 130, 175, 358, 450, 573, 1098]

males = [3, 59, 72, 183, 273, 293, 337]

 # Creating a DataFrame for the data

data = pd.DataFrame({

    'Texts': females + males,

    'Gender': ['Female'] * len(females) + ['Male'] * len(males)
 })

 # Plotting the boxplot

plt.figure(figsize=(8, 6))

sns.boxplot(x='Gender', y='Texts', data=data)

plt.title('Parallel Boxplots of Text Messages Sent by Females and Males')

plt.xlabel('Gender')

plt.ylabel('Number of Texts')

plt.show()
```



b.

* Shape:Females: The data for females is skewed to the right, meaning most females sent fewer texts, but a few sent a lot more. There are some very large values (like 450, 573, and 1098 texts) that are much higher than most of the data, making the overall spread bigger.Males: The data for males is more evenly spread, but still has a few higher values (like 293 and 337 texts) that pull the data slightly to the right.However, the spread is not as extreme as for the females, and there are fewer outliers. The male data is more tightly grouped.

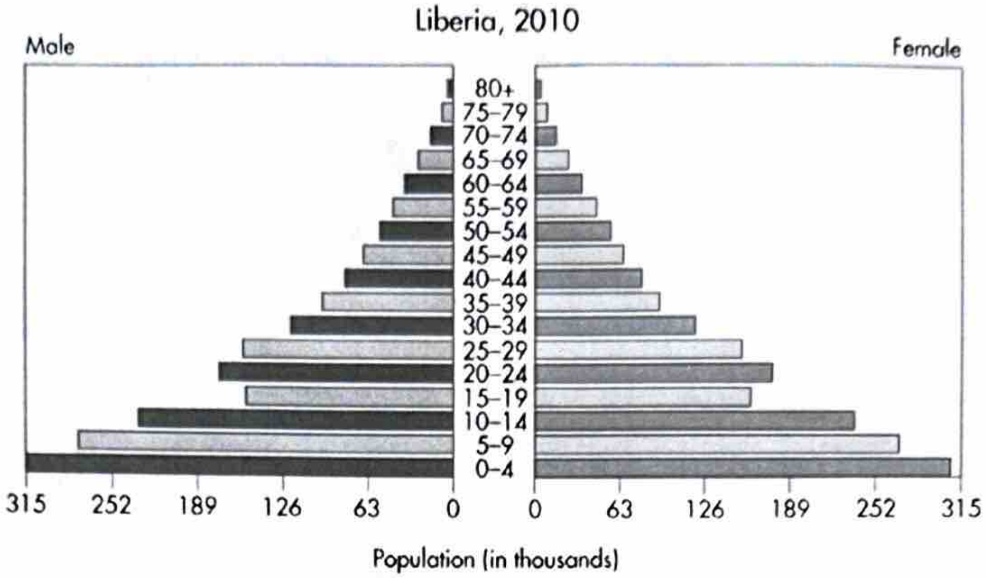
* Outliers: Females have higher outliers (450, 573, 1098), while males have fewer extreme values.

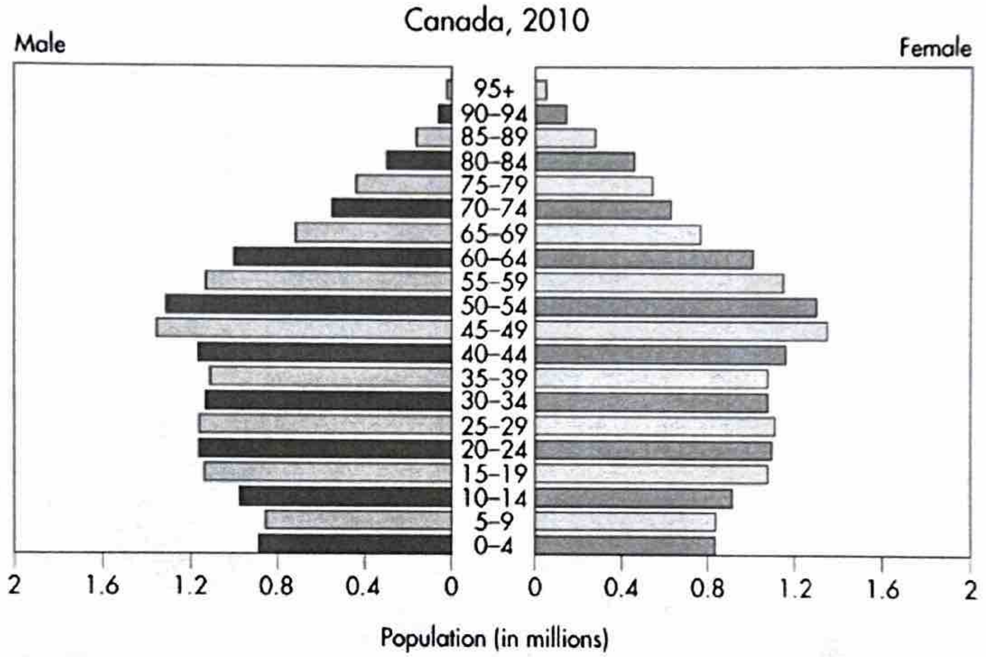
* Center: The median number of texts is slightly higher for males (183) compared to females (175).

* Spread: Females have a larger range (1083) and IQR (228), indicating more variability in the number of texts sent.
Males have a smaller range (334) and IQR (201), indicating more consistency in their texting behavior.
The standard deviation for females is expected to be high due to the large spread of values, while males have a lower expected standard deviation, reflecting their more uniform texting patterns.

Conclusion: While males have a higher median, the presence of extreme outliers suggests that some females sent significantly more messages.Therefore, females likely have a greater mean number of texts due to the influence of the higher outliers.

(2) **(15 pts)** Below are two population pyramids from the U.S. Census Bureau.





a. The approximate median age of the Liberian population falls in which of these intervals: 0–4, 15–19, 30–34, 40–44? Explain.

b. Explain why it is impossible to calculate the mean age of either population.

c. Which country has more children younger than 10 years of age? Explain.

d. Does the population pyramid indicate that Canadian men or Canadian women live longer? Explain.

e. In 2010, Liberia had recently come out of a civil war with the extensive use of child soldiers. How is this visible in the population pyramid?

Answer:

a.The median age of the Liberian population falls between 15–19. This can be inferred from the shape of the pyramid, where the population is largest in this age group, and the pyramid tapers down as you move toward older age groups.

b. It is impossible to calculate the mean age because the population pyramids only provide intervals of ages, not exact values for individuals. Without knowing the exact ages or how individuals are distributed within each interval, a precise mean cannot be determined.

c. Liberia has a larger proportion of children under 10 years of age, as seen by the wider base of the pyramid for ages 0-9 compared to Canada's pyramid, which has a much narrower base.

d. Canadian women live longer, as indicated by the fact that the topmost bars of the pyramid (85+ age groups) are wider for females than for males.This shows that there are more women than men in the older age groups.

e.The use of child soldiers is visible in the population pyramid as a distortion in the younger age groups (15–19) and possibly 10–14, where the population seems disproportionately affected (either in size or the distribution of males and females).

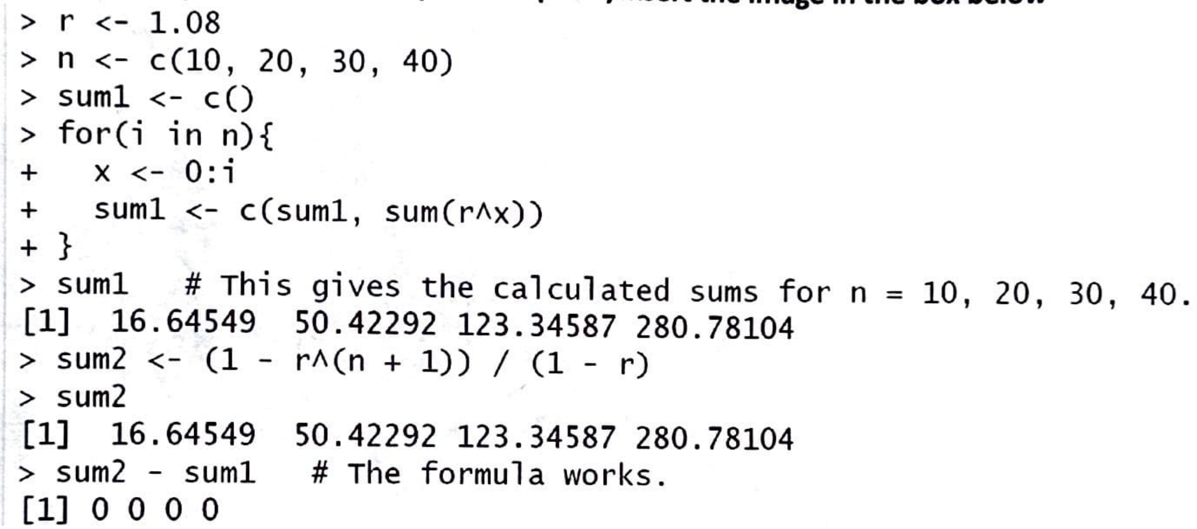
Appendix: Example Question and Answer for R programming questions:

Calculate the sum $\sum_{j=0}^{n} r^j$, where r has been assigned the value 1.08, and compare with $(1 - r^{n+1})/(1 - r)$, for $n = 10, 20, 30, 40$.

Answer: Copy and paste your R code in the box below (not an image but the text).

```
r <- 1.08
n <- c(10, 20, 30, 40)
sum1 <- c()
for(i in n){
  x <- 0:i
  sum1 <- c(sum1, sum(r^x))
}
sum1   # This gives the calculated sums for n = 10, 20, 30, 40.

sum2 <- (1 - r^(n + 1)) / (1 - r)
sum2

sum2 - sum1   # The formula works.
```

Screenshot of your R console outputs and paste/insert the image in the box below



```
> r <- 1.08
> n <- c(10, 20, 30, 40)
> sum1 <- c()
> for(i in n){
+   x <- 0:i
+   sum1 <- c(sum1, sum(r^x))
+ }
> sum1   # This gives the calculated sums for n = 10, 20, 30, 40.
[1]  16.64549  50.42292 123.34587 280.78104
> sum2 <- (1 - r^(n + 1)) / (1 - r)
> sum2
[1]  16.64549  50.42292 123.34587 280.78104
> sum2 - sum1   # The formula works.
[1] 0 0 0 0
```

THE END