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

Assignment 3

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 93 pts = 21 + 48 + 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)

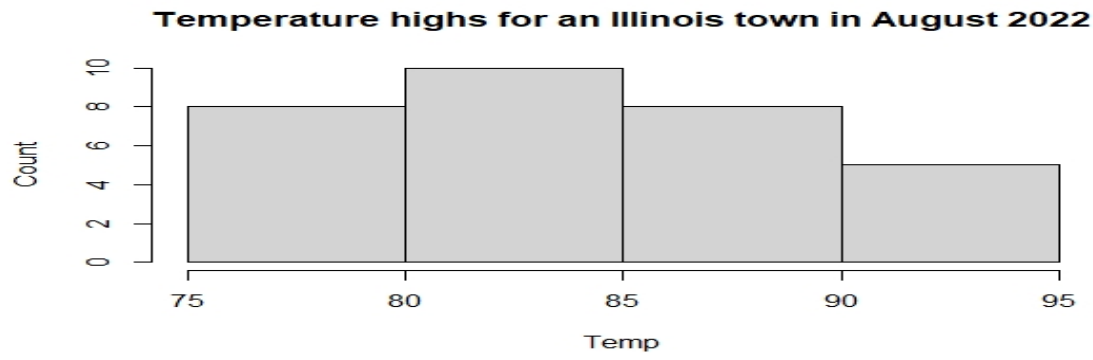
https://drive.google.com/drive/folders/1_8qcBjQVMfZggF42UYJuHQzMoBcyAy0Q?usp=drive_link

Part I: Quantitative Data [21 pts]

1) The following are the temperature highs for an Illinois town in August 2022.

78	75	83	85	87	91	78	91	85	84	80	79	83
85	86	80	88	94	89	86	88	84	81	85	83	80
79	87	88	91	95								

a) Create a histogram of this data (include a table of bins and counts).



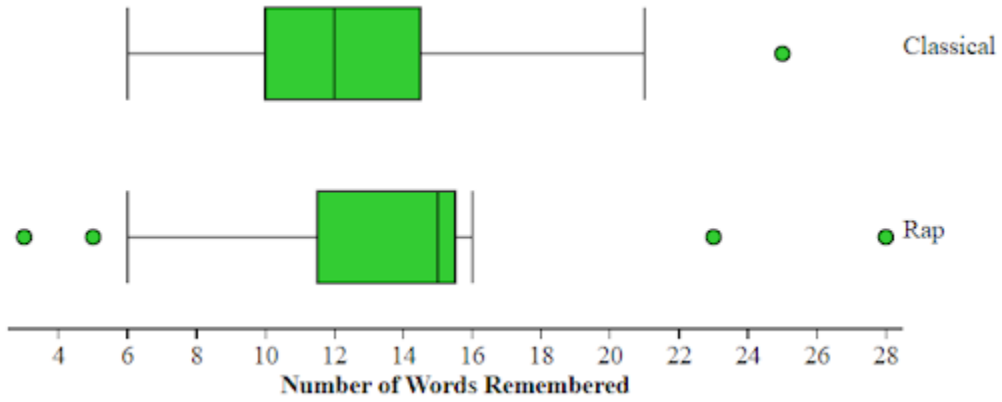
b) Create a stem and leaf plot for this data.

```
The decimal point is at the |

74 | 0
76 |
78 | 0000
80 | 0000
82 | 000
84 | 000000
86 | 0000
88 | 0000
90 | 000
92 |
94 | 00
```

c) Which graph do you think displays the data the best? Why?

The Histogram displays the data the best because it provides the change in temperature and frequency. The stem plot does not provide a great amount of change

2) Amy and Bob want to know if listening to different types of music while studying will help you remember the material better. They randomly assigned a group of students to two different groups: one group studied a list of words while listening to classical music while the second group studied the same list of words while listening to rap music. There was a total of 30 words on the list and each group studied the list while listening to the music for 5 minutes. They were then asked to write down as many words as they could remember. The data is displayed in the boxplots below.



a) Approximate the interquartile range for each set of data. Why is this the appropriate measure of spread to use for these two data sets? **(3 pts)**

Answer:

The interquartile range for classical is 22-6 which is 16 and for rap it is 16-6

b) Write **three** sentences comparing the number of words remembered between each of the two groups. **(9 pts)**

Answer:

The rap group has a higher median number of words remebered compared to classical. The rap group also has a tighter range of words remebered compared to classical. Lastly, rap has a large number of outliers compared

Part II. Summary Statistics and Boxplots [48 pts = 16 x 3 pts]

1) Below is a list of calories and cholesterol amounts in 4 randomly selected menu items from 4 different fast food companies. [The data is accessible via a .csv file saved under the course shared folder "Course Contents" as "calories_and_cholesterol_amounts.csv"]

Company	*Menu Item*	*Calories*	*Cholesterol (mg)*
McDonald's	Bacon, Egg, & Cheese Biscuit	460	215
McDonald's	Big Mac®	590	85
McDonald's	Filet-O-Fish®	390	30
McDonald's	Cheeseburger	300	40
Burger King	Whopper® Sandwich with Cheese	740	115
Burger King	Cheeseburger	280	45
Burger King	Crispy Chicken Sandwich	670	60
Burger King	Bacon, Egg & Cheese Biscuit	400	170
Wendy's	Baconator®	960	155
Wendy's	Bacon Double Stack®	440	65
Wendy's	Classic Chicken Sandwich	490	75
Wendy's	Sausage, Egg, and Cheese Biscuit	580	285
Chick-fil-A	Chicken Biscuit	460	45
Chick-fil-A	Bacon, Egg, and Cheese Biscuit	420	180
Chick-fil-A	Grilled Chicken Sandwich	390	75
Chick-fil-A	Chicken Nuggets (8 count)	250	80

a) Using R, find the following summary statistics for CALORIES for each company. **(6 pts = 2 x 3 pts)**

Answer:

	Mean	**Min**	**Q1**	**Med**	**Q3**	**Max**	**Std. Dev**
McDonald's	435	300	368	425	492	590	122
Burger King	522.	280	370	535	688.	740	218
Wendy's	618	440	478	535	675	960	236.
Chick-fil-A	380	250	355	405	430	460	91.3

Which company has the greatest calorie variability in the distribution?

Answer: The company with the greatest calorie Variability is Wendys

b) There are many methods for determining outliers. Two methods frequently used are:

Rule #1: An outlier is a value greater than 1.5 × IQR above the third quartile or more than 1.5 × IQR below the first quartile.

Rule #2: An outlier is a value located 2 or more standard deviations above, or below, the mean.

Using rule #1, are there any outliers in the Wendy's distribution? Show your work.

Answer:

```
# Extract Wendy's calorie data
wendys_calories &lt;- data %&gt;%
  filter(Company == "Wendy's") %&gt;% select(Calories) %&gt;% unlist()


# Show the extracted Wendy's calories
print(wendys_calories)


# Rule #1: Calculate Q1, Q3, and IQR
Q1 &lt;- quantile(wendys_calories, 0.25)  # First quartile
Q3 &lt;- quantile(wendys_calories, 0.75)  # Third quartile
IQR &lt;- Q3 - Q1                        # Interquartile range


# Calculate upper and lower bounds for outliers
upper_bound_rule1 &lt;- Q3 + (1.5 * IQR)
lower_bound_rule1 &lt;- Q1 - (1.5 * IQR)


# Identify outliers for Rule #1
outliers_rule1 &lt;- wendys_calories[wendys_calories &lt; lower_bound_rule1 | wendys_calories &gt; upper_bound_rule1]
print(outliers_rule1)
# Print results for Rule #1
cat("Rule #1: Using IQR Method\n")
cat("Q1:", Q1, "\n")
cat("Q3:", Q3, "\n")
cat("IQR:", IQR, "\n")
cat("Upper Bound for Outliers:", upper_bound_rule1, "\n")
cat("Lower Bound for Outliers:", lower_bound_rule1, "\n")
cat("Identified Outliers:", outliers_rule1, "\n\n")
```

According to rule 1 there are no outliers

Using rule #2, are there any outliers in the Wendy's distribution? Show your work.

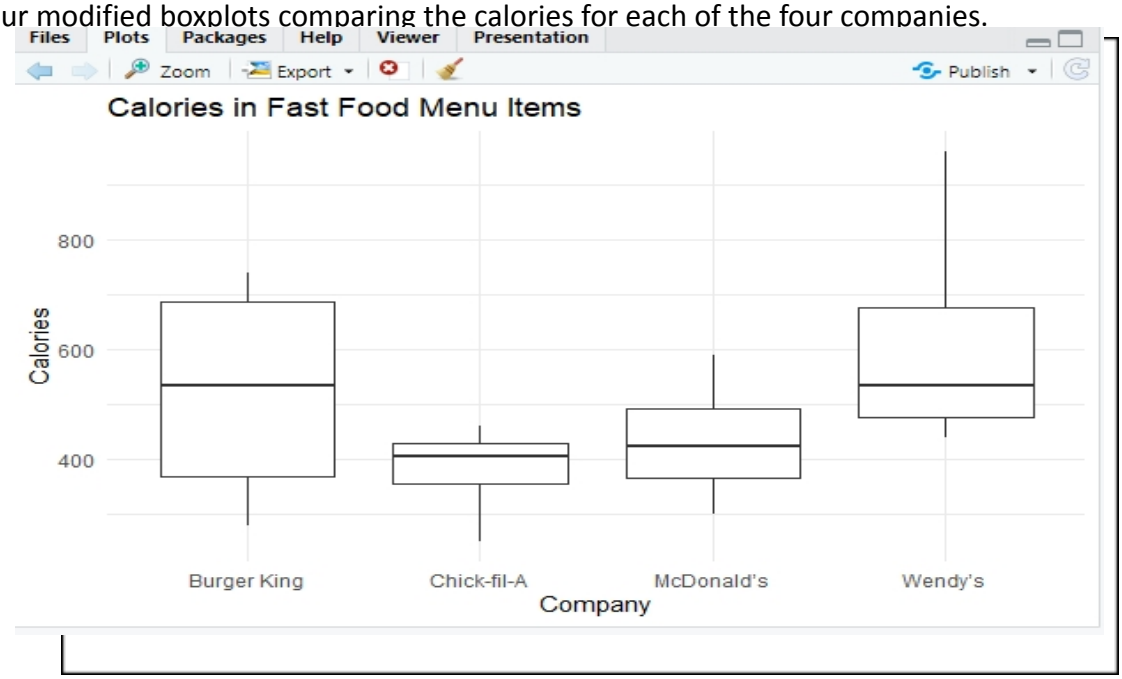
Answer:

```
# Rule #2: Calculate mean and standard deviation
mean_calories &lt;- mean(wendys_calories)            # Mean
std_dev_calories &lt;- sd(wendys_calories)           # Standard deviation


# Calculate upper and lower bounds for outliers
upper_bound_rule2 &lt;- mean_calories + 2 * std_dev_calories
lower_bound_rule2 &lt;- mean_calories - 2 * std_dev_calories


# Identify outliers for Rule #2
outliers_rule2 &lt;- wendys_calories[wendys_calories &lt; lower_bound_rule2 | wendys_calories &gt; upper_bound_rule2]


# Print results for Rule #2
cat("Rule #2: Using Standard Deviation Method\n")
cat("Mean:", mean_calories, "\n")
cat("Standard Deviation:", std_dev_calories, "\n")
cat("Upper Bound for Outliers:", upper_bound_rule2, "\n")
cat("Lower Bound for Outliers:", lower_bound_rule2, "\n")
cat("Identified Outliers:", outliers_rule2, "\n\n")
```

There are no outliers

c) Draw four modified boxplots comparing the calories for each of the four companies.

Answer:



d) In your town, McDonald's and Burger King are on the Northside and Wendy's and Chick-fil-A are on the Southside. Using R, find the following summary statistics for CHOLESTEROL for the Northside and Southside fast food restaurants. **(6 pts = 2 x 3 pts)**

	Mean	**Min**	**Q1**	**Med**	**Q3**	**Max**	**Std. Dev**
Northside	97.5	45	56.25	87.5	128.75	170	56.93564
Southside	95	45	67.5	77.5	105	180	58.7367

Which region has the greatest cholesterol variability in the distribution?

Answer:

South Side has greatest variability

e) Using rule #1, are there any outliers in the Southside's distribution? Show your work.

Answer:

There is one outlier

Using rule #2, are there any outliers in the Southside's distribution? Show your work.

Answer:

```
# Rule #2: Identify outliers based on standard deviation
mean_cholesterol &lt;- mean(southside_cholesterol, na.rm = TRUE)
sd_cholesterol &lt;- sd(southside_cholesterol, na.rm = TRUE)


# Determine the lower and upper bounds for outliers
lower_bound_rule2 &lt;- mean_cholesterol - 2 * sd_cholesterol
upper_bound_rule2 &lt;- mean_cholesterol + 2 * sd_cholesterol


# Identify outliers for Rule #2
outliers_rule2 &lt;- southside_cholesterol[
  southside_cholesterol &lt; lower_bound_rule2 |
    southside_cholesterol &gt; upper_bound_rule2
]


# Output Rule #2 Results
cat("Rule #2 Outliers (Standard Deviation Method):", outliers_rule2, "\n")
```

There are no outliers

f) Remove the value of 285mg from the Southside cholesterol's data set. Use R to find the following values again:

	Mean	Min	Q1	Med	Q3	Max	Std. Dev
Southside	95	45	67.5	77.5	105	180	58.7367

What values changed the most? What values changed the least?

Answer (Most):

The Max changed

Answer (Least):

Everything else

Part III Extra Credit Questions (21 pts)

Extra Credit Question (20pts)

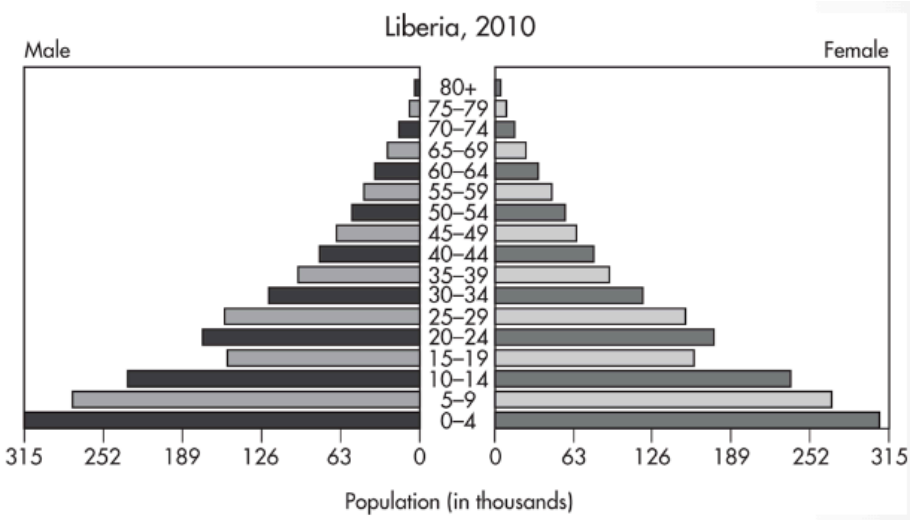
(1) (6 pts) To analyze the social media behavior differences between boys and girls, Mr. P's Statistics class was asked to count the number of text messages that they sent over a three-day weekend. The following table summarizes the data:

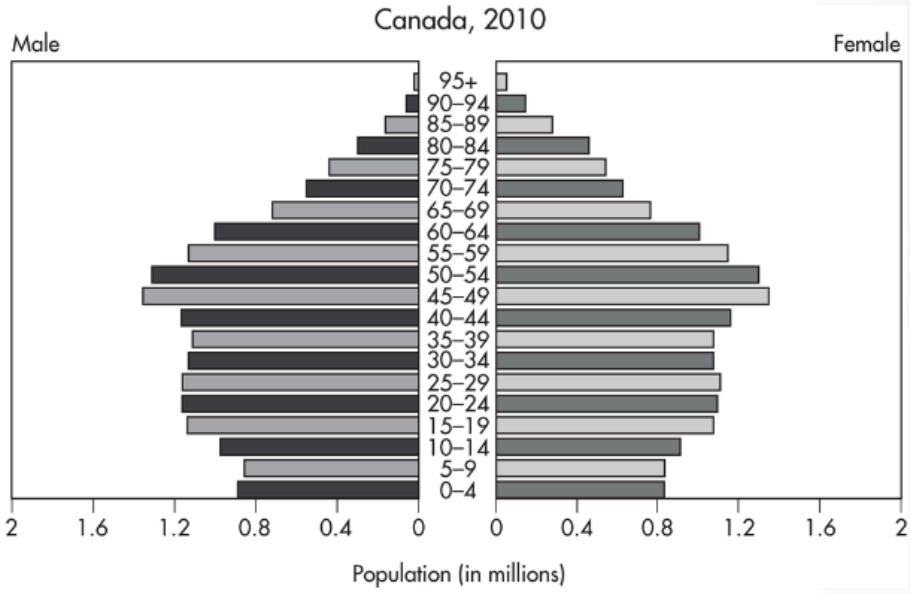
	Values under Q_1	Q_1	Median	Q_3	Values over Q_3
Females	15, 43, 100	130	175	358	450, 573, 1098
Males	3, 59	72	183	273	293, 337

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:

(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:

1. 20-24 because it is in the it is scewed towards the lower ages
2. It is impossible because population age changes everyday
3. Canada because it is measured in millions
4. They both live similair ages
5. There is a sharp decline on ages 15-19 most likely the age of child soldiers to be conscience to shoot and fight.

THE END