1. A counselor at ABC University wanted to find out the amount of outstanding tuition (in CAD$) ABC University international students carry in the Spring semester. A random sample of 50 ABC University international students was drawn in March to investigate this.

a) Identify the researcher. [1 mark]

The researcher = Counselor at ABC University.

b) Provide a description of the objective. [1 mark]

What=to find out the amount of outstanding tuition of international students at ABC University in the spring semester (currency $CAD)

c) Identify the subjects of interest. [2 marks] *Note: Make sure you also include the when and where, if available.*

Who (Subjects) = International students

When= Spring Semester

Where= ABC University

d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. [1+1 marks]

Variable = the amount of outstanding tuition

Units of measurement = CAD$

e) Identify the type of the variable and the corresponding scale of measurement. [1+1 marks]

Note: Marks will be deducted with missing words (like "variable" or "scale").

Type = numerical variable

Scale =ratio scale

f) Provide a description of the population of interest. [1 mark]

Population=All the international students at ABC University in the spring semester

g) Identify the most appropriate sampling method. [1 mark]

Simple Random Sampling Method

h) Provide a description of the sample. [1 mark]

A random sample of 50 ABC University international students in March

i) Is there an issue with the selection bias? Briefly justify your answer by comparing the (target) population and the sampling frame. [0+2 marks]

Sampling frame = All the international students at ABC University in the spring semester.

The sampling frame matches the target population. So, there is no issue with the selection bias.

2. Marisol lives in the City of Vancouver. She was planning to buy an electric vehicle (EV) this summer and wondering what percentage of EV owners have installed a Level 3 Supercharger at home. A sample of 20 EV owner living in her neighbourhood was drawn in June to investigate this.

a) Identify the researcher. [1 mark]

The researcher = Marisol at the City of Vancouver

b) Provide a description of the objective. [1 mark]

What=To find out what percentage of EV owners have installed a Level 3 Supercharger at home this summer in the City of Vancouver.

c) Identify the subjects of interest. [2 marks]

Note: Make sure you also include the when and where, if available.

Who = EV owners

When= this summer

Where= City of Vancouver

d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. [1+1 marks]

Variable = Binary indicator of the presence or absence of a Level 3 Supercharger installed at home by an EV owner.

Classes = Yes or Not

e) Identify the type of the variable and the corresponding scale of measurement. [1+1 marks]

Note: Marks will be deducted with missing words (like "variable" or "scale").

Type = Categorical variable

Scale =Nominal scale

f) Provide a description of the population of interest. [1 mark]

All the EV owners in the City of Vancouver this summer.

g) Identify the most appropriate sampling method. [1 mark]

Convenience Sampling Method, because she chose her neighbourhood

h) Provide a description of the sample. [1 mark]

A sample of 20 EV Owner living in her neighborhood in this summer

i) Is there an issue with the selection bias? Briefly justify your answer by comparing the (target) population and the sampling frame. [0+2 marks]

There is an issue with the selection bias because the sample was taken in her neighborhood (sampling frame) and is not the most accurate representation of the City of Vancouver (target population).

3. The director overseeing all senior homes in the Fraser Health Authority (FHA) wanted to know how many falls seniors have in 2023 that leads to major hip or lower body surgery. Ten seniors were randomly selected from each senior home in the FHA region to form the sample.

Note: This is a tricky question. Make sure you spend some time thinking about how to differentiate between a categorical variable from a numerical variable.

a) Identify the researcher. [1 mark]

The director overseeing all senior homes in the FHA.

b) Provide a description of the objective. [1 mark]

To find out how many falls seniors led to major hip or lower body surgery in the FHA region in 2003.

c) Identify the subjects of interest. [2 marks]

Note: Make sure you also include the when and where, if available.

Who (Subjects) = elderly people

Where = Senior homes in the FHA region

When= 2023

d) Provide a description of the variable of interest. Please also provide the units of measurement if it is a numerical variable or list out all possible classes if it is a categorical variable. [1+1 marks]

Variable = Falls that lead to major hip or lower body surgery

Units of measurement= number of falls

e) Identify the type of the variable and the corresponding scale of measurement. [1+1 marks]

Note: Marks will be deducted with missing words (like "variable" or "scale"). Note: I suggest you provide some reasoning to justify your choice of the type of variable.

Type= Numerical variable

Scale= Ratio scale

f) Provide a description of the population of interest. [1 mark]

All the elderly people who lived in senior homes in the FHA region in 2003

g) Identify the most appropriate sampling method. [1 mark]

Stratified Random Sampling Method

4. Students are generally confused with the stratified random sampling method and the cluster sampling method. I hope students have a better understanding of them after completing this question. There are two situations in the questions: one method is used in each of the two situations. In each of the following two situations, (1) identify the sampling method used, (2) describe how you would define the non-overlapping groups, and (3) outline the three main steps.

Note: In practice, geographic factor should not be used to do the stratification. It is used here only for the sake of illustration.

a) To get an idea of how much detached houses in the city of Vancouver cost these days, Joe went to a real estate web site to collect some information. He divided the city of Vancouver into 20 communities; randomly selected 5 detached houses from each community; and the 100 (20 communities times 5 houses per community) detached houses formed the sample. [1+1+3 marks]

1) Stratified Random Sampling Method

2)The non-overlapping groups are strata, in this case, the strata are the 20 communities, and then 5 houses from each community.

3) First divide the population into strata. Second, select some random subjects from each stratum. Finally, form the sample with all the chosen subjects (homogenous group)

b) To get an idea of how much detached houses in the city of Vancouver cost these days, Joe went to a real estate web site to collect some information. He divided the city of Vancouver into 20 communities; randomly selected 5 of the communities; and all detached houses (that are listed on the web site) within those 5 communities were chosen in the sample. [1+1+3 marks]

1) Cluster Sampling Method.

2) The non-overlapping groups are clustered, in this case, the 20 communities, and then 5 communities were chosen.

3) First divide the population into clusters. Second, select some clusters. Finally, form the sample with all the subjects from the chosen cluster (heterogeneous group).

5. Consider a movie theatre with 30 rows of 20 seats in each row. There are some prize giveaways before the movie starts. Identify the most appropriate sampling method used below. *Note: In practice, geographic factor should not be used to do the stratification. It is used here only for the sake of illustration.*

a) Two random rows will be drawn (from among all 30 rows). Everybody in the two selected rows (40 of them) will get a prize. [1 mark]

Cluster Sampling Method

b) Each person entering the theatre will have to write down their names on a piece of paper and then put the paper in a big bag. Twenty (20) names were randomly drawn for the prize. [1 mark]

Simple Random Sampling Method

c) Two random winners will be drawn in each row, for all 30 rows. [1 mark]

Stratified Random Sampling Method

d) The first 30 movie goers who enter the theatre get a prize. [1 mark]

Convenience Sampling Method

e) A random person is chosen among the first five who enter the theatre and prizes are given to every 5th person thereafter. [1 mark]

Systematic Sampling Method

6. General public underestimate the power of statistics. In particular, misuse of statistics could have a devastating impact on individuals or organizations. Your task is to do a Google search about the 1936 US Presidential Election between the incumbent Democratic candidate Franklin D. Roosevelt and Republican candidate Alf Landon. You want to pay special attention to the pre-election prediction between the two organizations – Literary Digest and

Gallup Poll. Answer the following questions.

Note: You might get slightly different values from different websites. So, there are no standard

Answers here.

a) What was the success rate (predicting the correct election outcome) of Literary Digest prior to 1936? [1 mark]

The **Literary Digest** had an impressive **success rate of 100%** in predicting the outcome of **five consecutive U.S. presidential elections** from **1916 to 1932** before the 1936 election.

However, this record was broken in **1936**, when the magazine incorrectly predicted that **Alf Landon** would defeat **Franklin D. Roosevelt** in a landslide. In reality, Roosevelt won by a massive margin.

This failure is a famous example of **sample bias**: although the Digest surveyed over **2 million people**, it relied on **lists of automobile owners, telephone directories, and its subscribers**—a population skewed toward wealthier Americans, who were more likely to vote Republican during the Great Depression. George Gallup, using a **much smaller but more scientifically selected sample**, correctly predicted Roosevelt's victory.

b) What were the 1936 pre-election prediction by both Literary Digest and Gallup Poll? Please focus on your answer as

who would win the 1936 Presidential election and by what percentage of popularity vote. [2 marks]

- Literary Digest: Predicted Alf Landon would win with 57.08% of the popular vote.
- Gallup Poll: Predicted Franklin D. Roosevelt would win with 55.7% of the popular vote.

c) what was the official result of the 1936 US Presidential Election? [1 mark]
Franklin D. Roosevelt won with 60.8% of the popular vote

Now let us focus on the Literary Digest only from this point on.

d) In relation to selection bias, what did the Literary Digest do (or not do) to wrongly predict the selection results? Please provide as much details as possible. [2 marks]

Selection bias occurs when the sample is not representative of the population, mostly was answered by republicans.

What the Literary Digest did:

- "Opted for quantity, paying little attention to the method of selection." (1)
- It used **telephone directories and automobile registration lists** to collect addresses and send out **10 million questionnaires**.
- During the Great Depression (1930s), only **wealthier Americans** had **telephones and cars**.
- This led to an **overrepresentation of affluent, Republican-leaning voters** in the sample and **underrepresentation of working-class and poorer Democratic voters**.

Their sample was **biased toward Republican voters**, which **skewed the prediction** in favor of Alf Landon.

e) In relation to non-response bias, what did the Literary Digest do (or not do) to wrongly predict the election results? Please provide as much details as possible. [2 marks]

What the Literary Digest did:

- The magazine **mailed over 10 million ballots**, but only about **2.4 million responded** (~24% response rate).
- Those who responded were **wealthier**, **more politically engaged**, and more likely to be **Republican voters**.

- Those who didn’t respond (over 7 million people) were **largely working-class and poor Americans**, who were suffering the most during the Great Depression and were more likely to support **Franklin D. Roosevelt** (Democrat).
- Because the magazine **only counted those who replied**, it **ignored the political voice of the silent majority**, which leaned Democratic.

f) In relation to response bias, what did the Literary Digest do (or not do) to wrongly predict the election results? Note that The Great Depression started in 1929 and last till around 1939. So, general public did not know when the Great Depression would end. So, imagine to whom you would point your fingers when time is tough. [2 marks]

What the Literary Digest did:

- The survey was conducted during the **Great Depression** (1929–1939), when many people were **financially devastated** and blamed the **Republican Party** for the economic collapse.
- However, **the Digest’s respondents were mostly wealthy or middle-class individuals**, who owned **cars or phones** (because that’s how the Digest gathered its mailing list).
- These people may have **expressed support for the Republican candidate, Alf Landon**, either to preserve their interests or out of social identification, even if others around them were suffering.
- Additionally, those **truly suffering (unemployed, poor)** were **not heard** in the survey, so the **responses didn’t reflect the anger or political shift** among that demographic.

Conclusion: The Digest’s sample was not only limited, but also captured responses from a group that was less likely to be objective or representative, especially given the intense public blame directed at the Republicans during the Depression.

g) Now focus on modern days. What are your thoughts about 2016 US Presidential Election between Democratic candidate Hilary Clinton and Republican candidate Donald Trump, in relation to the three biases that we have learned? [2 marks]

Selection Bias:
Many pollsters failed to include enough non-college-educated white voters, a group that strongly supported Trump.

Samples often overrepresented urban, educated populations that leaned toward Clinton.

Non-Response Bias:
Some Trump supporters distrusted mainstream media and pollsters, and therefore did not participate in polls.

This undercounted support for Trump, especially in rural areas and swing states.

Response Bias:
Due to social pressure, some Trump voters might have hidden their support, fearing backlash. This is often called the "shy Trump voter" effect.

As a result, they either said they were undecided or claimed support for Clinton when asked in public surveys.

While Clinton won the popular vote, Trump won key Electoral College states due to these polling blind spots. The 2016 election showed that polling is still vulnerable to the same biases that plagued the Literary Digest in 1936—just in new forms.

From this point on, you are expected to use R (no python, no Excel and no other tools).

7. A police officer from Vancouver Police Department wanted to find out the percentage of drivers who were distracted (defined as using their phone while driving or waiting at the traffic lights) during the day. The officer took a random sample of 80 drivers **to** investigate this. The results can be found in “DANA4800_HW1_Q07_Data.xlsx” on BrightSpace.

A) Create a Frequency Table of the variable “Distracted” using the table() function. [1 mark]

Note: When copy-and-pasting text output from R to Word document, for example, make sure

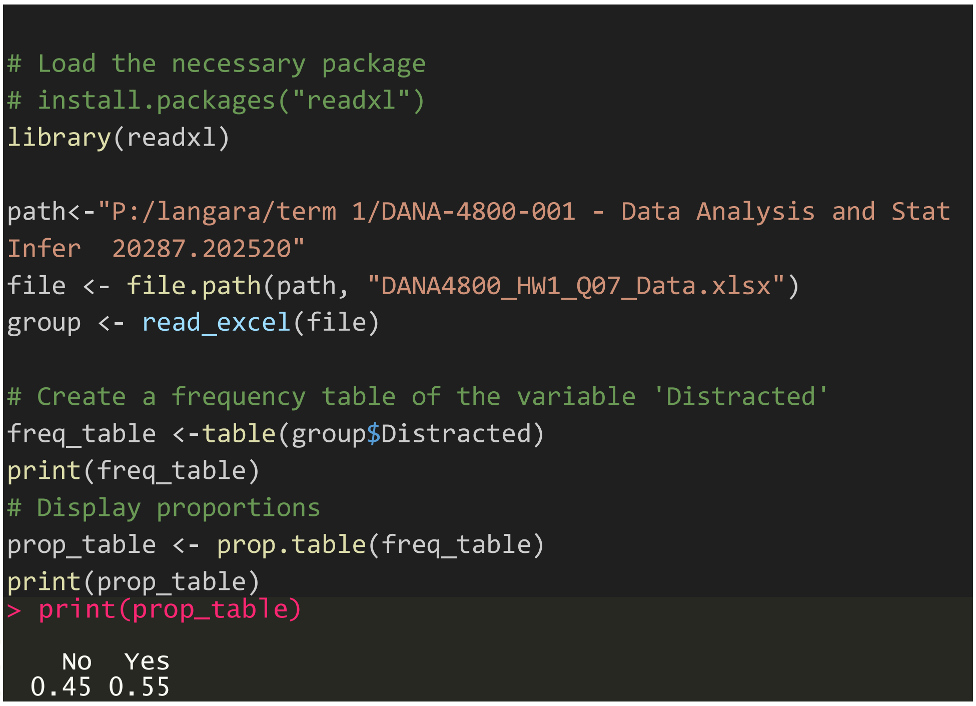
you use “fixed-width fonts”, like Courier New. Otherwise, the output does not look right or

aligned properly.

```
# 18-08-2025
#
# DANA4800_lancheros_leonardo_HW1.docx
# r studio

# Load the necessary package
# install.packages("readxl")
library(readxl)

path<-"P:/langara/term 1/DANA-4800-001 - Data Analysis and Stat
Infer  20287.202520"
file <- file.path(path, "DANA4800_HW1_Q07_Data.xlsx")
group <- read_excel(file)

# Create a frequency table of the variable 'Distracted'
freq_table <-table(group$Distracted)
print(freq_table)

No Yes
 36  44
```

B) Create a Probability Table of the variable “Distracted” using the proportions() function. [1 mark]

```
# 18-08-2025
#
# DANA4800_lancheros_leonardo_HW1.docx
# r studio
```



```
# Load the necessary package
# install.packages("readxl")
library(readxl)

path<-"P:/langara/term 1/DANA-4800-001 - Data Analysis and Stat
Infer  20287.202520"
file <- file.path(path, "DANA4800_HW1_Q07_Data.xlsx")
group <- read_excel(file)

# Create a frequency table of the variable 'Distracted'
freq_table <-table(group$Distracted)
print(freq_table)
# Display proportions
prop_table <- prop.table(freq_table)
print(prop_table)
> print(prop_table)

   No  Yes
0.45 0.55
```

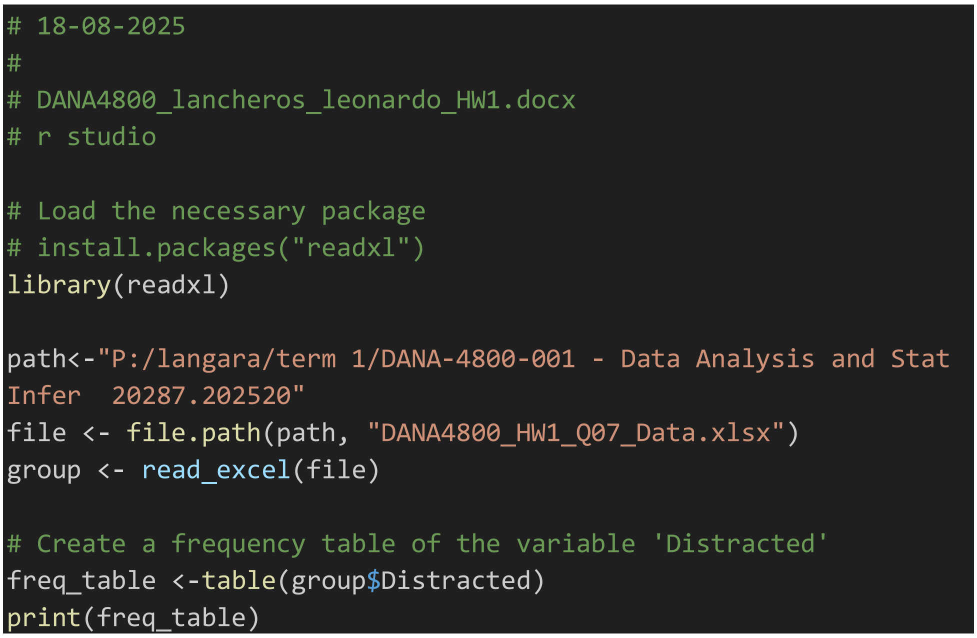
C) Provide a description of the parameter of interest. [2 marks]

The parameter of interest is the proportion of all drivers **who are distracted** during the day.

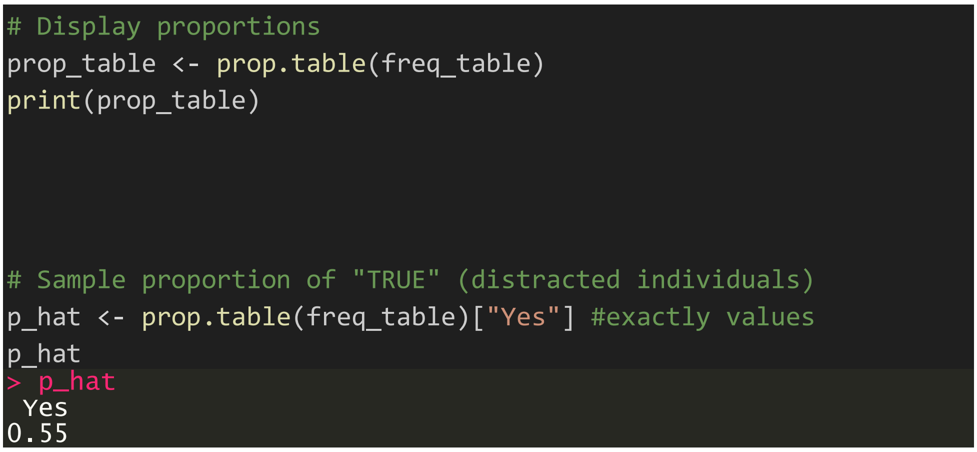
D) Provide a description of the corresponding statistic. [2 marks]

The statistic is the proportion of 80 randomly selected drivers **who are distracted** during the day.

E) Calculate the value of the most appropriate statistic. [1 mark]



```
# 18-08-2025
#
# DANA4800_lancheros_leonardo_HW1.docx
# r studio

# Load the necessary package
# install.packages("readxl")
library(readxl)

path<-"P:/langara/term 1/DANA-4800-001 - Data Analysis and Stat
Infer  20287.202520"
file <- file.path(path, "DANA4800_HW1_Q07_Data.xlsx")
group <- read_excel(file)

# Create a frequency table of the variable 'Distracted'
freq_table <-table(group$Distracted)
print(freq_table)
```



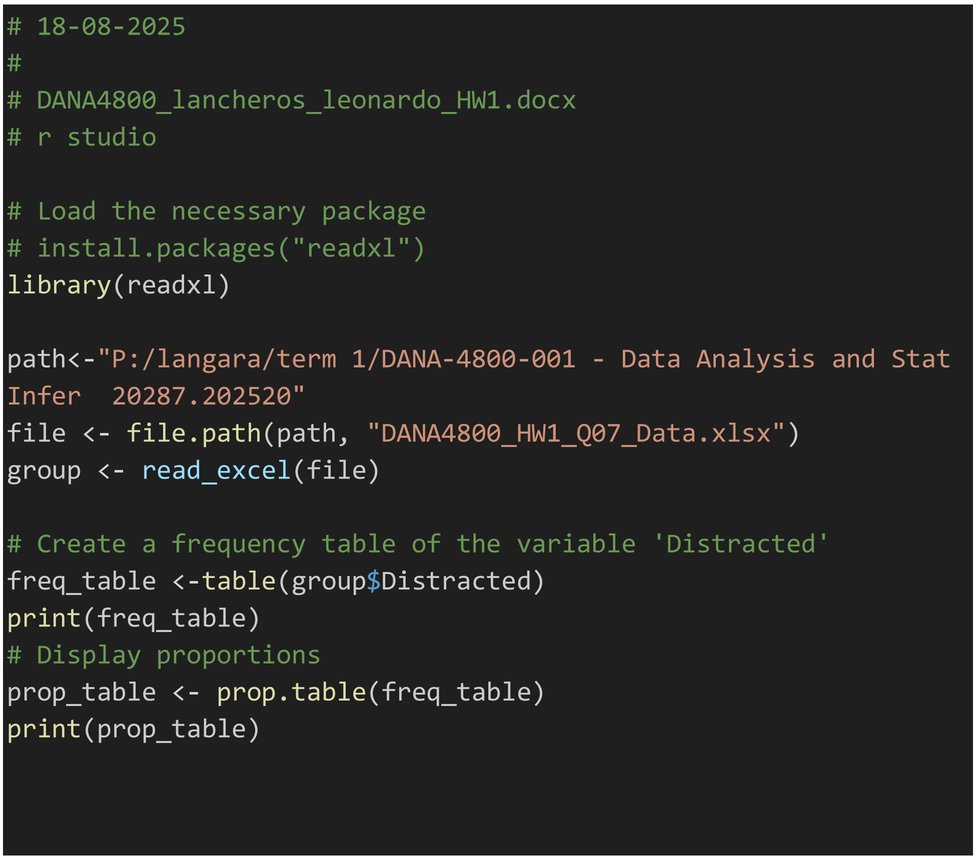
```
# Display proportions
prop_table <- prop.table(freq_table)
print(prop_table)




# Sample proportion of "TRUE" (distracted individuals)
p_hat <- prop.table(freq_table)["Yes"] #exactly values
p_hat
> p_hat
 Yes
0.55
```

F) Produce a Pie Chart using the pie() function, with the “clockwise” arguments set as TRUE, and “Yes” goes before “No”. Please also submit the code to produce such graph using fixed-width fonts. [2+1 marks] Page 4 of 7

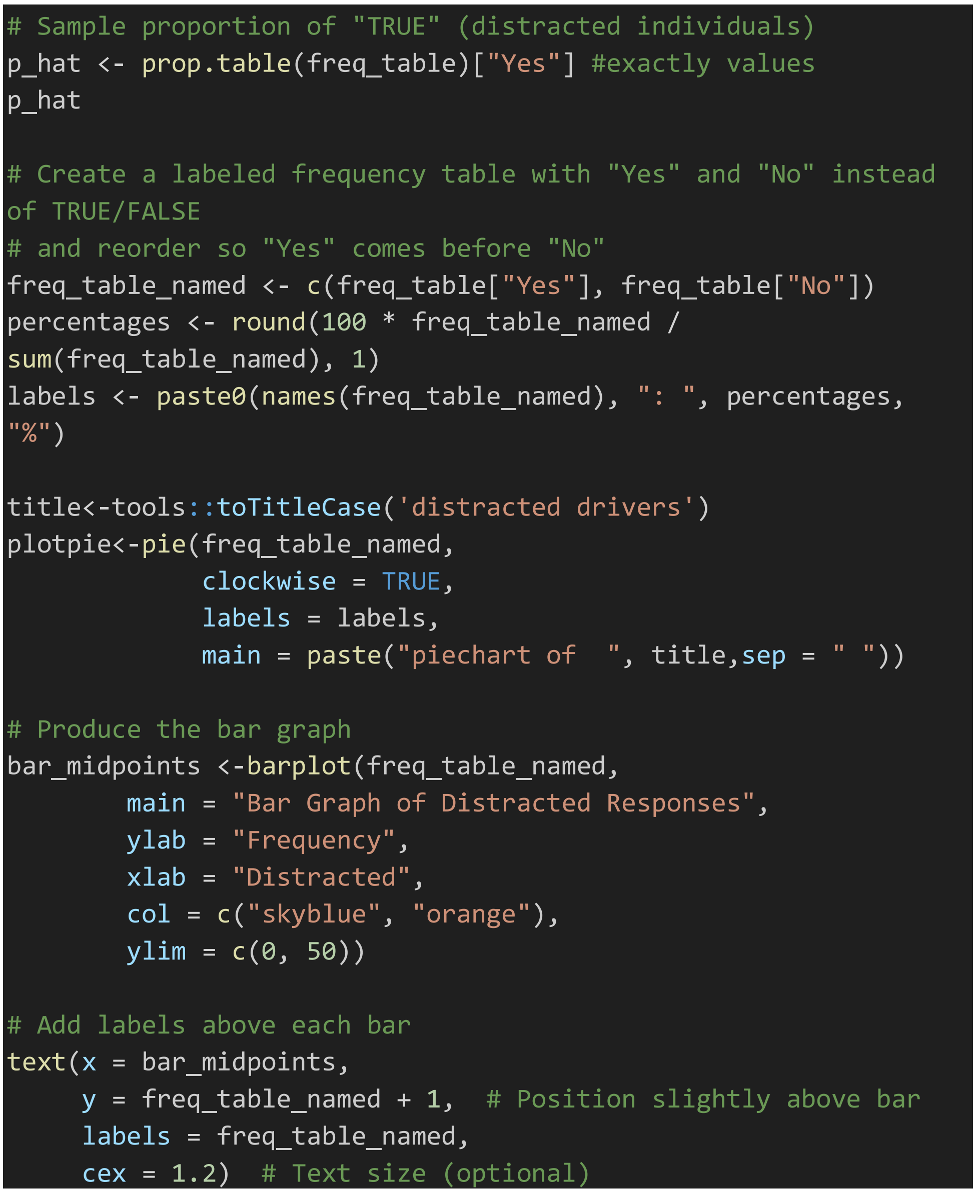
DANA 4800 Homework #1

Note: Please make sure you personalize the pie chart by including the Main Title, and add labels

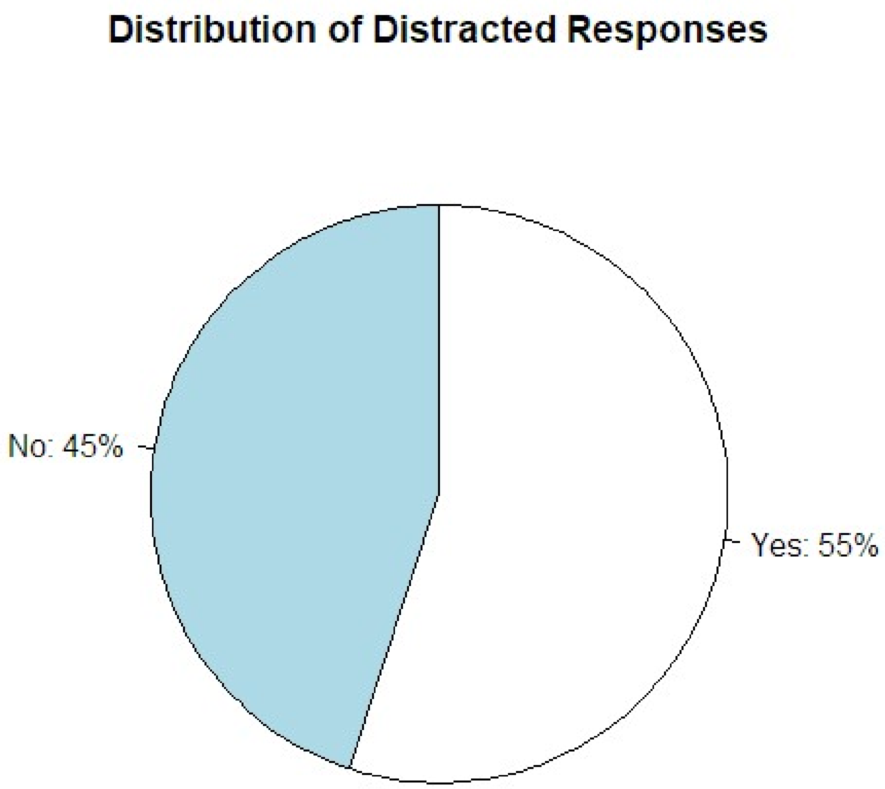
to axes (if applicable) etc.



```
# 18-08-2025
#
# DANA4800_lancheros_leonardo_HW1.docx
# r studio

# Load the necessary package
# install.packages("readxl")
library(readxl)

path<-"P:/langara/term 1/DANA-4800-001 - Data Analysis and Stat
Infer  20287.202520"
file <- file.path(path, "DANA4800_HW1_Q07_Data.xlsx")
group <- read_excel(file)

# Create a frequency table of the variable 'Distracted'
freq_table <-table(group$Distracted)
print(freq_table)
# Display proportions
prop_table <- prop.table(freq_table)
print(prop_table)
```



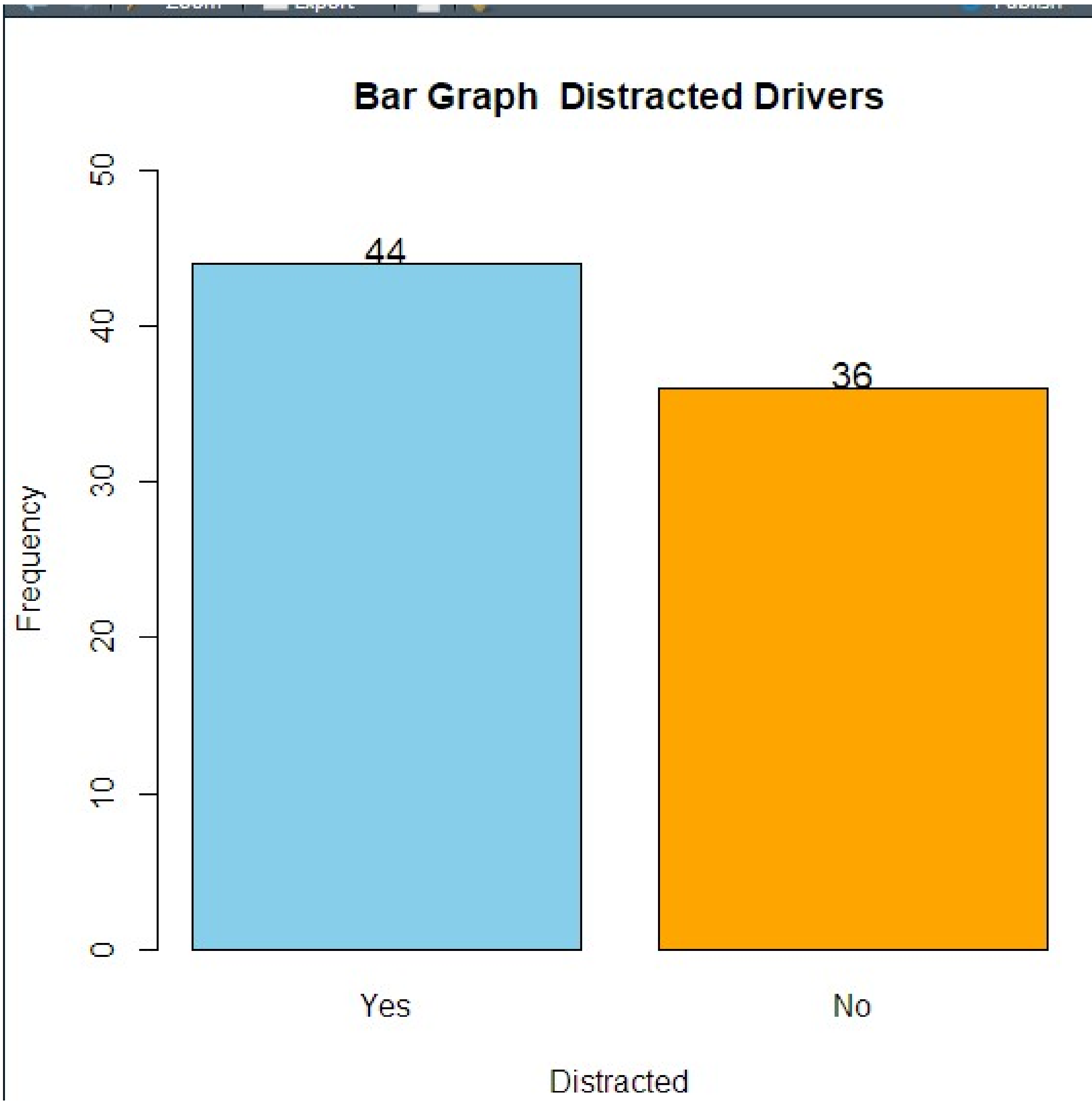
```r
# Sample proportion of "TRUE" (distracted individuals)
p_hat <- prop.table(freq_table)["Yes"] #exactly values
p_hat

# Create a labeled frequency table with "Yes" and "No" instead
of TRUE/FALSE
# and reorder so "Yes" comes before "No"
freq_table_named <- c(freq_table["Yes"], freq_table["No"])
percentages <- round(100 * freq_table_named /
sum(freq_table_named), 1)
labels <- paste0(names(freq_table_named), ": ", percentages,
"%")

title<-tools::toTitleCase('distracted drivers')
plotpie<-pie(freq_table_named,
             clockwise = TRUE,
             labels = labels,
             main = paste("piechart of  ", title,sep = " "))

# Produce the bar graph
bar_midpoints <-barplot(freq_table_named,
        main = "Bar Graph of Distracted Responses",
        ylab = "Frequency",
        xlab = "Distracted",
        col = c("skyblue", "orange"),
        ylim = c(0, 50))

# Add labels above each bar
text(x = bar_midpoints,
     y = freq_table_named + 1,  # Position slightly above bar
     labels = freq_table_named,
     cex = 1.2)  # Text size (optional)
```



G) Provide a description of the above pie chart. [1 mark]

The above pie chart shows the proportion of individuals who reported being distracted with “Yes” and no distracted with “No”. The chart is drawn clockwise, with yes appearing before No.
The 55% of the sample is with value of “Yes”, this denotes drivers **are distracted** during the day.

H) Produce a Bar Graph using the barplot() function, “Yes” goes to the left of “No”. Please also submit the code to produce such graph. Please use fixed-width fonts. [2+1 marks]
Note: A graph directly copied from Excel without any annotation will get a zero.



I) Which graph is better to use here? Briefly justify your answer using statistical reasoning. [0+1 mark]

Any of these charts/graphs are suitable for this data.

8. Whenever there is a major concert in a city, the hotel rate during that time normally go up. A random sample of 20 hotels in downtown Vancouver was drawn during the time of major concert and the rate of a hotel room per night (based on two double-bed rooms) was recorded.

and the rate of a hotel room per night (based on two double-bed rooms) was recorded.

286	378	245	292	244	314	298	282	281	317
319	237	289	275	285	227	270	322	274	293

a) Identify the subjects of interest. [1 mark]

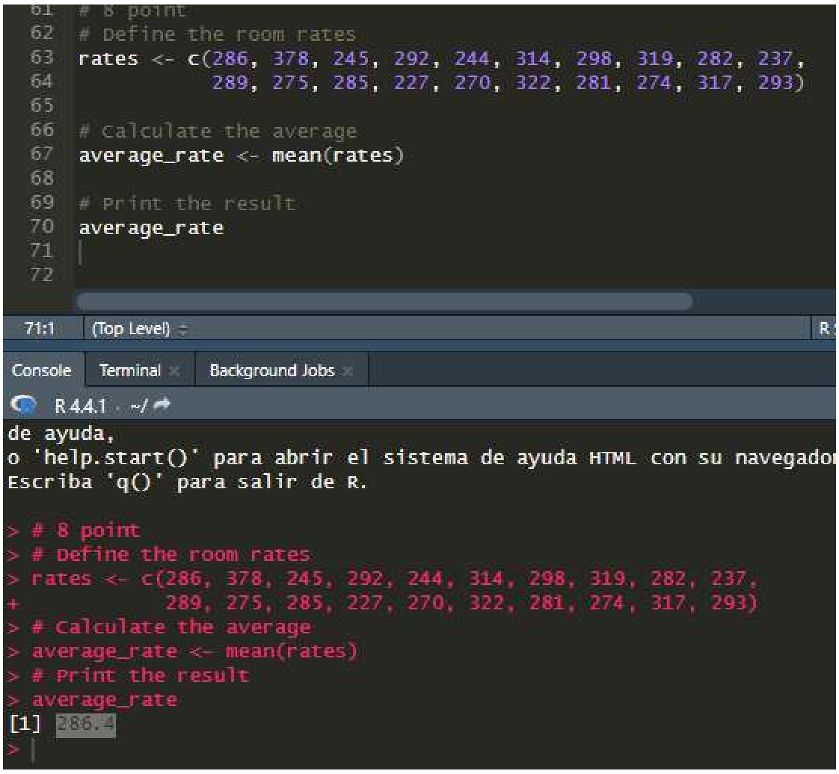
Who (Subjects) = hotels in downtown Vancouver

When= during the time of major concert

Where= vancouver

b) Calculate the average hotel room rate manually. Please show all work. [1 mark]

```
# 8 point
# Define the room rates
rates <- c(286, 378, 245, 292, 244, 314, 298, 319, 282, 237,
           289, 275, 285, 227, 270, 322, 281, 274, 317, 293) #
Room rates data

# Calculate the average
average_rate <- mean(rates) # Mean of room rates
```



```
# 8 point
# Define the room rates
rates <- c(286, 378, 245, 292, 244, 314, 298, 319, 282, 237,
           289, 275, 285, 227, 270, 322, 281, 274, 317, 293)

# Calculate the average
average_rate <- mean(rates)

# Print the result
average_rate
```

```
de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador
Escriba 'q()' para salir de R.

> # 8 point
> # Define the room rates
> rates <- c(286, 378, 245, 292, 244, 314, 298, 319, 282, 237,
+            289, 275, 285, 227, 270, 322, 281, 274, 317, 293)
> # Calculate the average
> average_rate <- mean(rates)
> # Print the result
> average_rate
[1] 286.4
> 
```

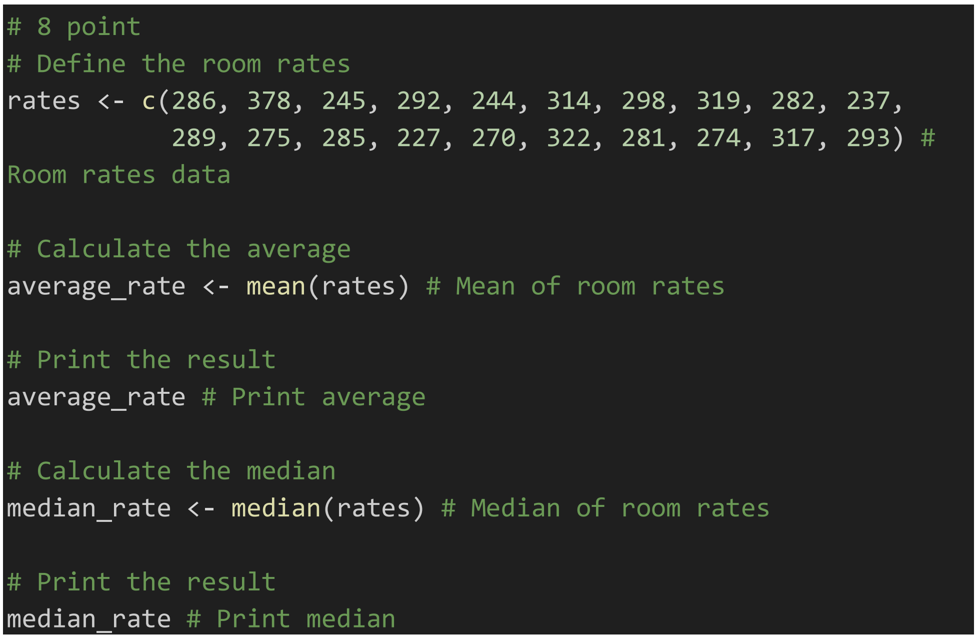
c) Provide a description of the statistic in part (b). [2 marks]

Statistic=the average of price of a hotel room per night (based on two double-bed rooms)

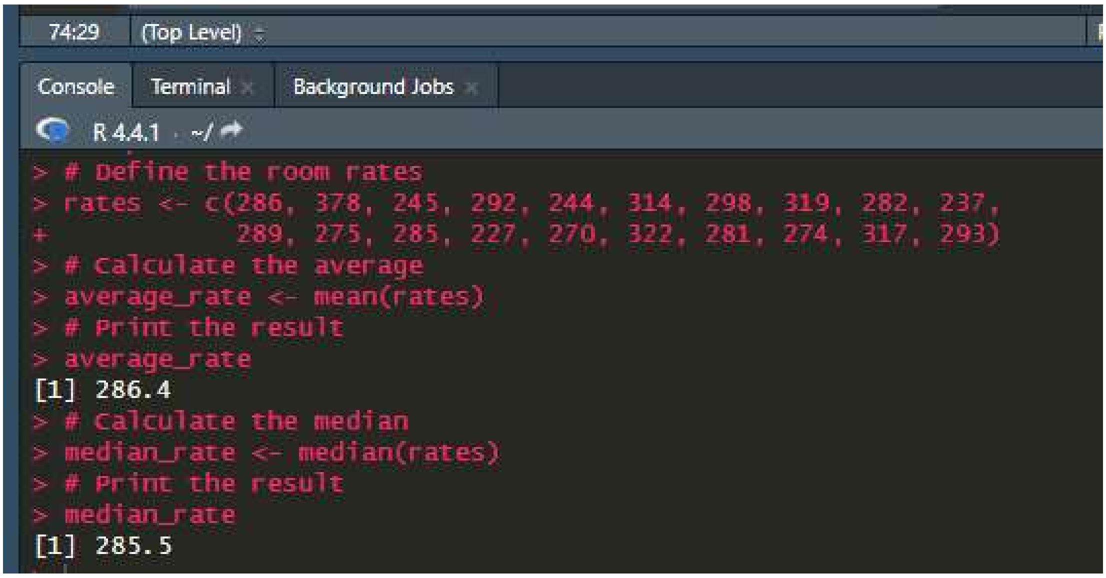
during the time of major concert

d) Find the median hotel room rate manually. Please show all work. [2 marks]

R job

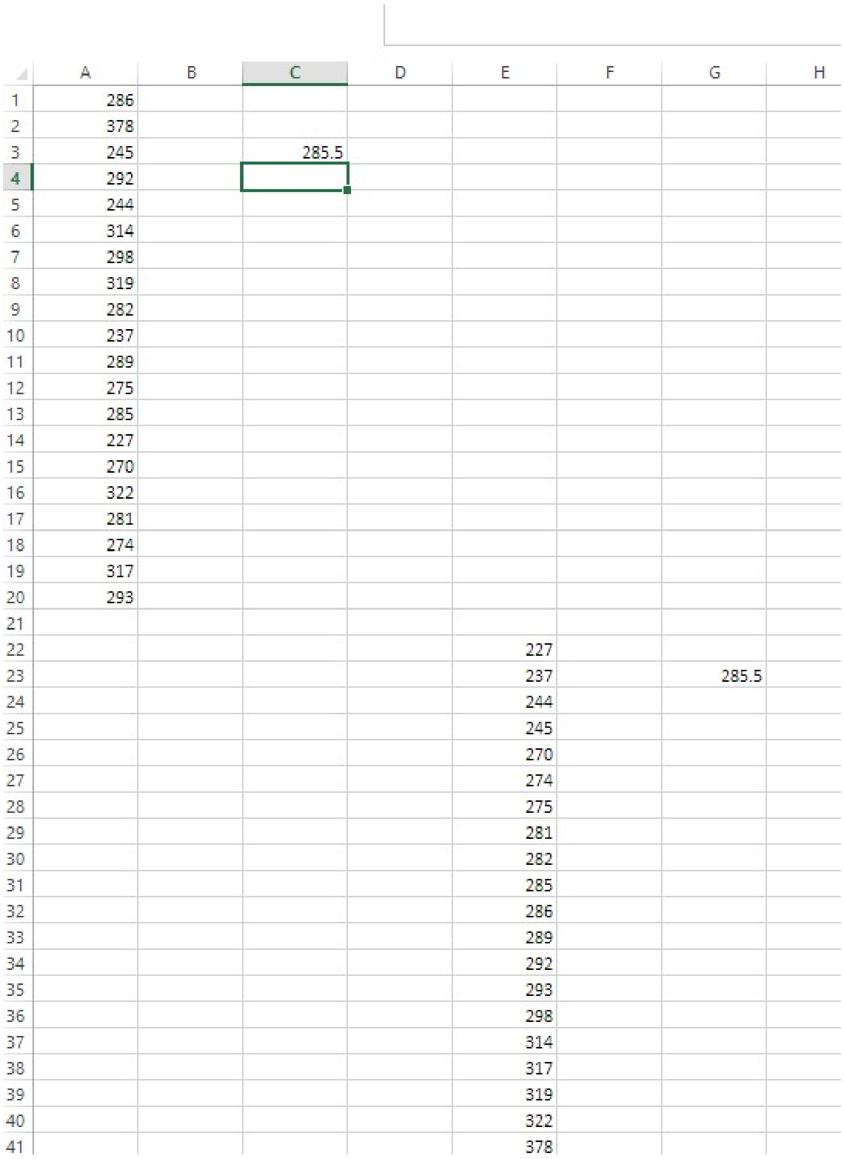


```
# 8 point
# Define the room rates
rates <- c(286, 378, 245, 292, 244, 314, 298, 319, 282, 237,
           289, 275, 285, 227, 270, 322, 281, 274, 317, 293) #
Room rates data

# Calculate the average
average_rate <- mean(rates) # Mean of room rates

# Print the result
average_rate # Print average

# Calculate the median
median_rate <- median(rates) # Median of room rates

# Print the result
median_rate # Print median
```



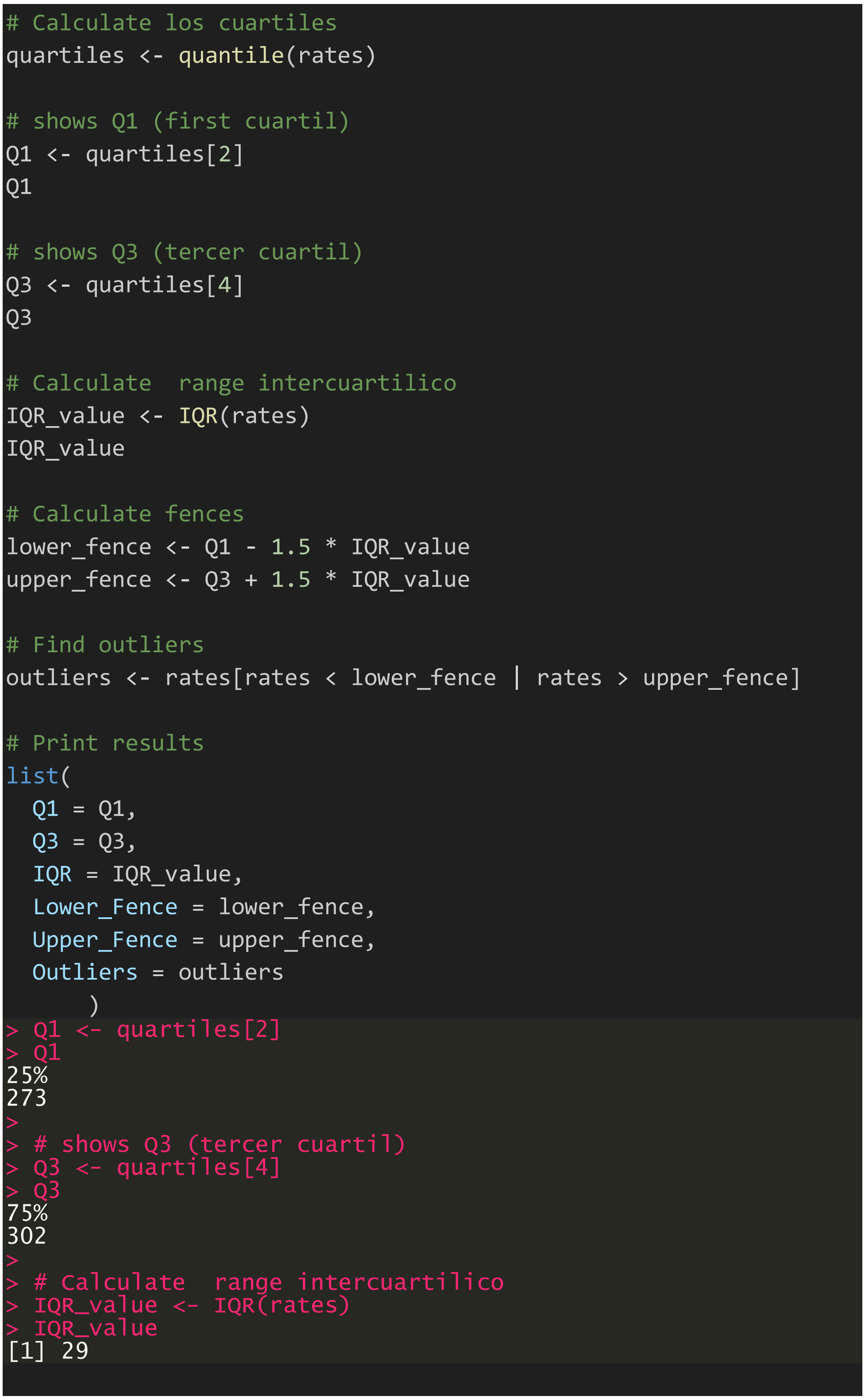
```
> # Define the room rates
> rates <- c(286, 378, 245, 292, 244, 314, 298, 319, 282, 237,
+            289, 275, 285, 227, 270, 322, 281, 274, 317, 293)
> # Calculate the average
> average_rate <- mean(rates)
> # Print the result
> average_rate
[1] 286.4
> # Calculate the median
> median_rate <- median(rates)
> # Print the result
> median_rate
[1] 285.5
```

Excel job



	A	B	C	D	E	F	G	H
1	286							
2	378							
3	245		285.5					
4	292							
5	244							
6	314							
7	298							
8	319							
9	282							
10	237							
11	289							
12	275							
13	285							
14	227							
15	270							
16	322							
17	281							
18	274							
19	317							
20	293							
21								
22					227			
23					237		285.5	
24					244			
25					245			
26					270			
27					274			
28					275			
29					281			
30					282			
31					285			
32					286			
33					289			
34					292			
35					293			
36					298			
37					314			
38					317			
39					319			
40					322			
41					378			

e) Use the method in our notes to find the first quartile and the third quartile. Then find the interquartile range. [1+1+1 marks]

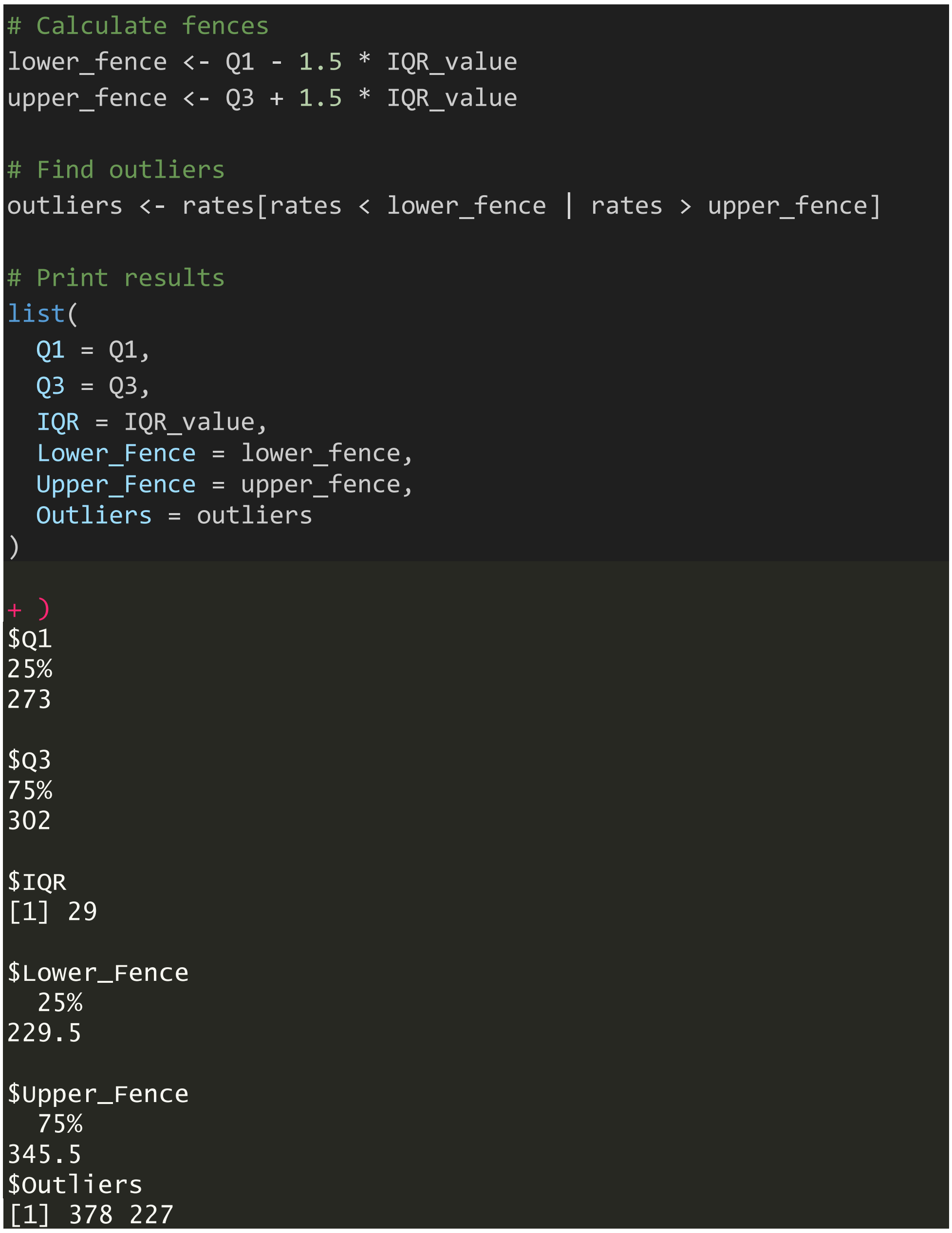


```r
# Calculate los cuartiles
quartiles <- quantile(rates)

# shows Q1 (first cuartil)
Q1 <- quartiles[2]
Q1

# shows Q3 (tercer cuartil)
Q3 <- quartiles[4]
Q3

# Calculate  range intercuartilico
IQR_value <- IQR(rates)
IQR_value

# Calculate fences
lower_fence <- Q1 - 1.5 * IQR_value
upper_fence <- Q3 + 1.5 * IQR_value

# Find outliers
outliers <- rates[rates < lower_fence | rates > upper_fence]

# Print results
list(
  Q1 = Q1,
  Q3 = Q3,
  IQR = IQR_value,
  Lower_Fence = lower_fence,
  Upper_Fence = upper_fence,
  Outliers = outliers
     )
```

```
> Q1 <- quartiles[2]
> Q1
25% 
273 
> 
> # shows Q3 (tercer cuartil)
> Q3 <- quartiles[4]
> Q3
75% 
302 
> 
> # Calculate  range intercuartilico
> IQR_value <- IQR(rates)
> IQR_value
[1] 29
```

```
[1] 29
```

f) Determine if there is/are any outlier(s), using IQR. [3 marks] The following two parts require the use of R. 293 g) Use the summary() function to find the five-number summary of the hotel rates. [1 mark]



```
# Calculate fences
lower_fence <- Q1 - 1.5 * IQR_value
upper_fence <- Q3 + 1.5 * IQR_value

# Find outliers
outliers <- rates[rates < lower_fence | rates > upper_fence]

# Print results
list(
  Q1 = Q1,
  Q3 = Q3,
  IQR = IQR_value,
  Lower_Fence = lower_fence,
  Upper_Fence = upper_fence,
  Outliers = outliers
)

+ )
$Q1
25%
273

$Q3
75%
302

$IQR
[1] 29

$Lower_Fence
  25%
229.5

$Upper_Fence
  75%
345.5
$Outliers
[1] 378 227
```

9. A first-year Langara student wanted to know the weekly expenses (in CAD$) of typical Langara students. To investigate this, she got a random sample of 100 students this term. The data set is in “Expense” worksheet of the file “DANA4800_HW1_Q09_Data.xlsx” on BrightSpace.

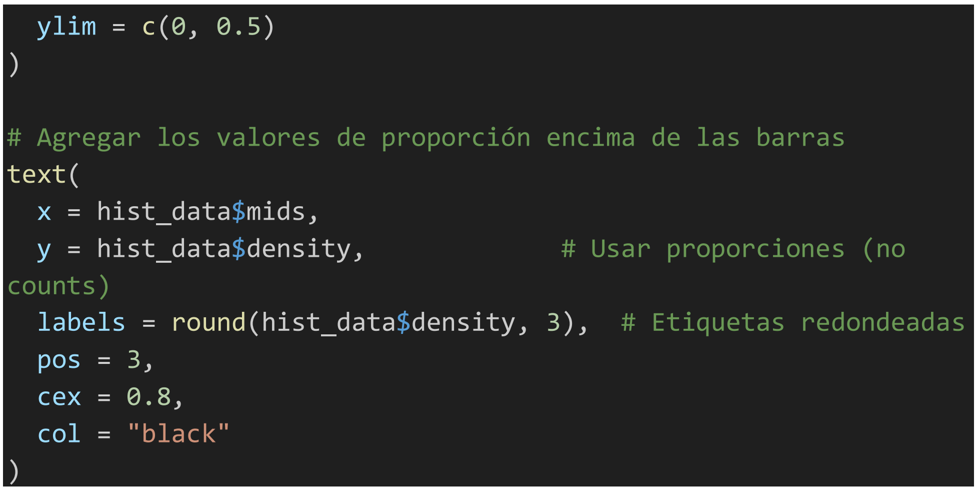
A) Use the summary() function to find the Five-Number Summary. [1 mark]

```
# 9. A first-year Langara student wanted to know the weekly
# expenses (in CAD$) of typical Langara students.
# To investigate this, she got a random sample of 100 students
# this term. The data set is in “Expense” worksheet of the file
# “DANA4800_HW1_Q09_Data.xlsx” on BrightSpace.

file_HW1_Q09 <- file.path(path, "DANA4800_HW1_Q09_Data.xlsx")
group2 <- read_excel(file)
summary(group2)
```

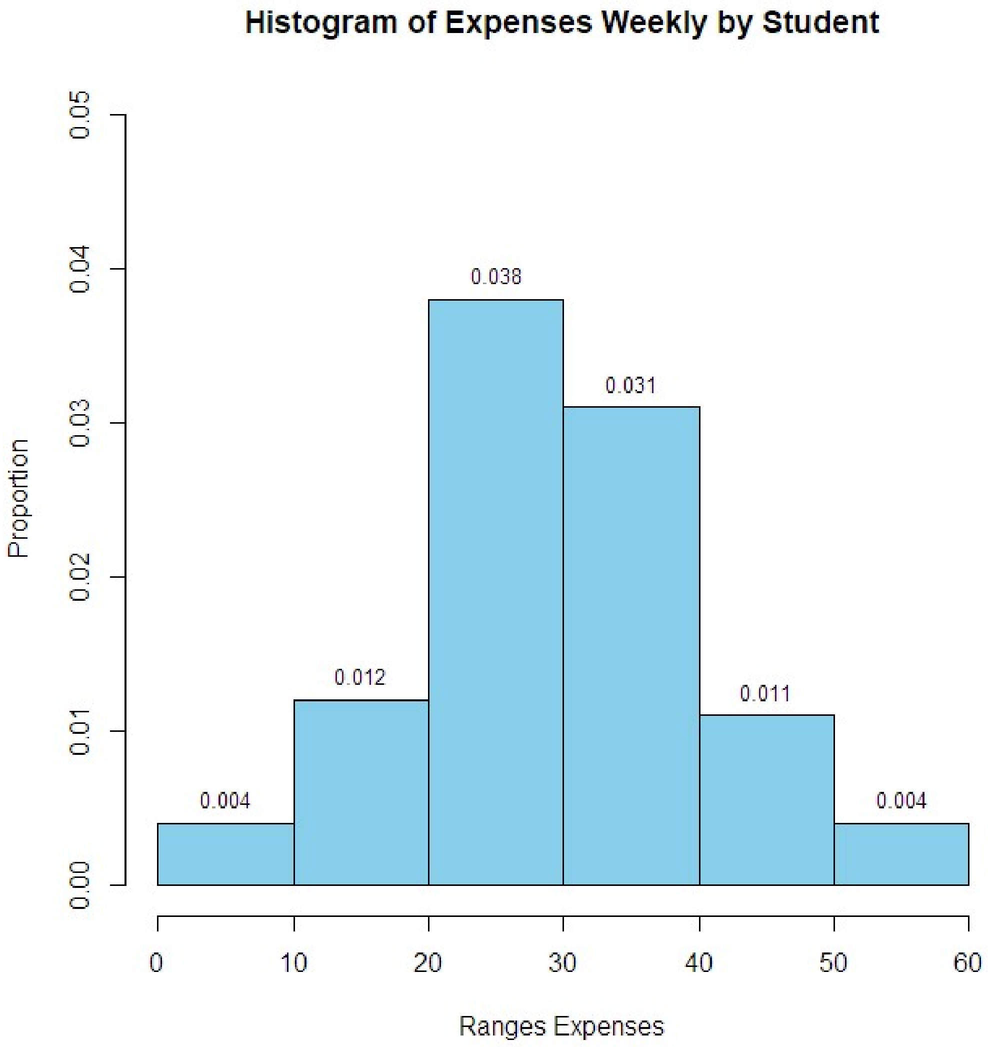
```
    Expense     
 Min.   : 3.00  
 1st Qu.:23.75  
 Median :29.50  
 Mean   :29.90  
 3rd Qu.:36.25  
```

b) Use the hist() functions to produce a Histogram with proportion on the y-axis with 6 bars only. Specifically, please make sure there are 6 bars (1-10, 11-20, …, 51-60), chart title included and axes labelled properly. [4 marks]
Note: Make sure you label your graph and axes appropriately.

```
# Assuming the data is in the first column, extract it:
data <- group2[[1]]  # Adjust if your column name is known,
e.g., group2$Score
data
# Create histogram with 6 specific breaks and proportion on y-
axis
titleHistogramStudent <- tools::toTitleCase("Histogram of
Expenses Weekly by Student")

# Crear histograma y guardar el objeto (proporciones)
hist_data <- hist(
  data,
  breaks = seq(0, 60, by = 10),
  freq = FALSE,                # Proporciones
  main = titleHistogramStudent,
  xlab = "Ranges Expenses",
  ylab = "Proportion",
  col = "skyblue",
  border = "black",
```



```
  ylim = c(0, 0.5)
)

# Agregar los valores de proporción encima de las barras
text(
  x = hist_data$mids,
  y = hist_data$density,              # Usar proporciones (no
counts)
  labels = round(hist_data$density, 3),  # Etiquetas redondeadas
  pos = 3,
  cex = 0.8,
  col = "black"
)
```



c) Provide a description of the above graph. [2 marks] d) Use the following segment of code to produce a Density Curve. Provide a description of the shape (only). [2+1 marks] e) Use the boxplot() function to make a horizontal Boxplot. [2 marks] Note: Make sure you label your graph and axes appropriately.

The histogram titled "Histogram of Expenses Weekly by Student" displays the distribution of weekly student expenses using 6 intervals (bins) from approximately 0 to 60 units on the x-axis, labeled as "Ranges Expenses". The y-axis represents the proportion of students within each expense range.

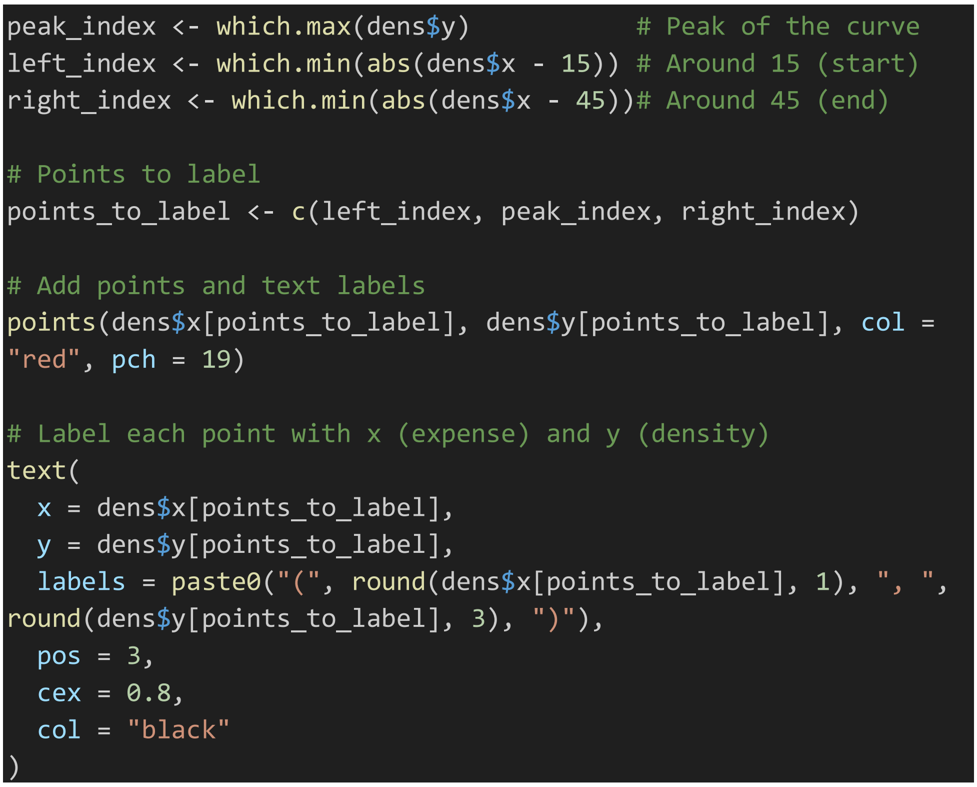
The most common expense range is between 20 and 30, with approximately 3.8% of the students falling into that bin.

The distribution is fairly centered between 20 and 40 units, with smaller proportions in the outer bins (e.g., 0–10 and 50–60).

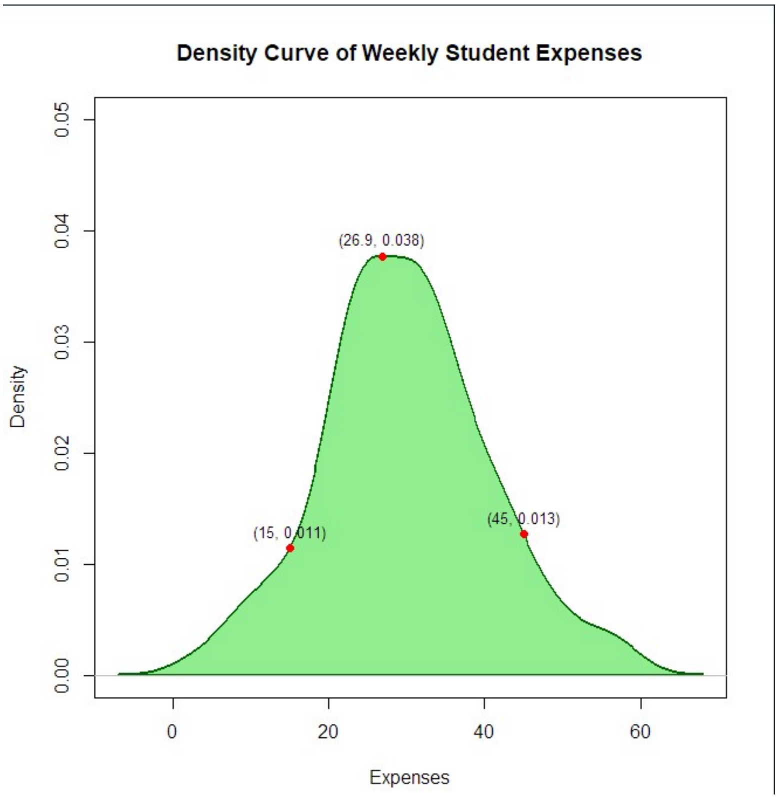
Overall, it appears to have a slightly bell-shaped but low-frequency distribution, suggesting that most students spend within a moderate range weekly.

d) Use the following segment of code to produce a Density Curve. Provide a description of the shape (only). [2+1 marks] e) Use the boxplot() function to make a horizontal Boxplot. [2 marks] Note: Make sure you label your graph and axes appropriately.

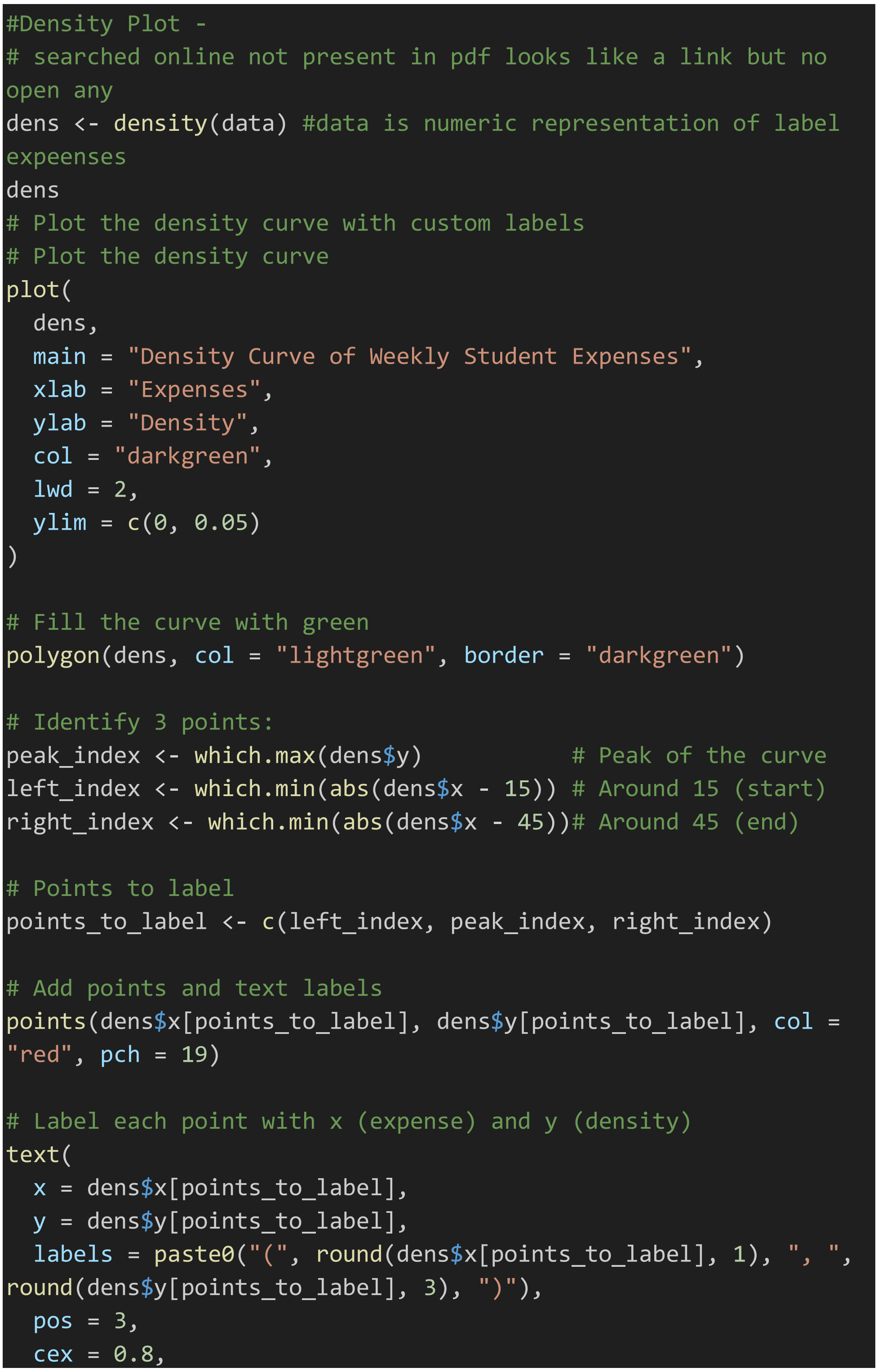
```
#Density Plot -
# searched online not present in pdf looks like a link but no
open any
dens <- density(data) #data is numeric representation of label
expeenses
dens
# Plot the density curve with custom labels
# Plot the density curve
plot(
  dens,
  main = "Density Curve of Weekly Student Expenses",
  xlab = "Expenses",
  ylab = "Density",
  col = "darkgreen",
  lwd = 2,
  ylim = c(0, 0.05)
)

# Fill the curve with green
polygon(dens, col = "lightgreen", border = "darkgreen")

# Identify 3 points:
```



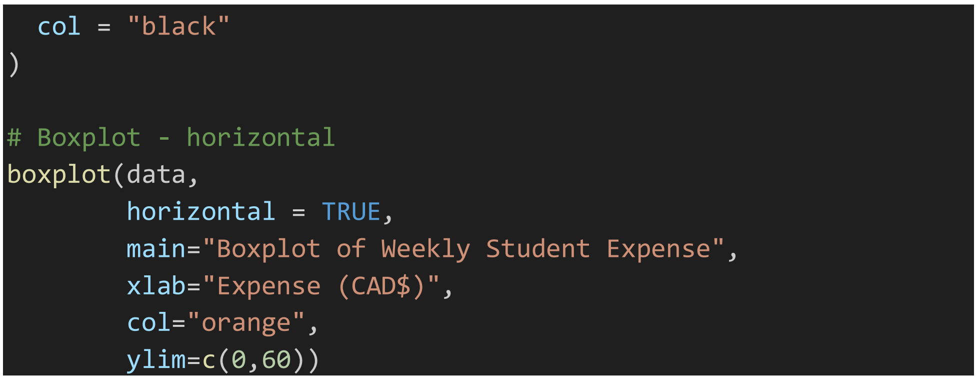
```
peak_index <- which.max(dens$y)           # Peak of the curve
left_index <- which.min(abs(dens$x - 15)) # Around 15 (start)
right_index <- which.min(abs(dens$x - 45))# Around 45 (end)

# Points to label
points_to_label <- c(left_index, peak_index, right_index)

# Add points and text labels
points(dens$x[points_to_label], dens$y[points_to_label], col =
"red", pch = 19)

# Label each point with x (expense) and y (density)
text(
  x = dens$x[points_to_label],
  y = dens$y[points_to_label],
  labels = paste0("(", round(dens$x[points_to_label], 1), ", ",
round(dens$y[points_to_label], 3), ")"),
  pos = 3,
  cex = 0.8,
  col = "black"
)
```



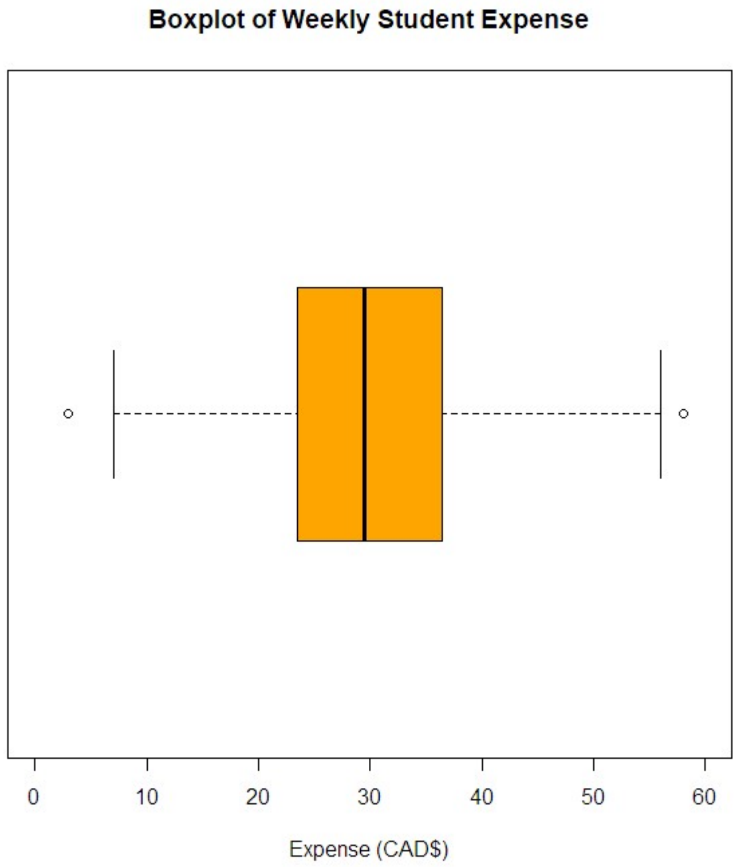
e) Use the boxplot() function to make a horizontal Boxplot. [2 marks] Note: Make sure you label your graph and axes appropriately.



```
#Density Plot -
# searched online not present in pdf looks like a link but no
open any
dens <- density(data) #data is numeric representation of label
expeenses
dens
# Plot the density curve with custom labels
# Plot the density curve
plot(
  dens,
  main = "Density Curve of Weekly Student Expenses",
  xlab = "Expenses",
  ylab = "Density",
  col = "darkgreen",
  lwd = 2,
  ylim = c(0, 0.05)
)

# Fill the curve with green
polygon(dens, col = "lightgreen", border = "darkgreen")

# Identify 3 points:
peak_index <- which.max(dens$y)           # Peak of the curve
left_index <- which.min(abs(dens$x - 15)) # Around 15 (start)
right_index <- which.min(abs(dens$x - 45))# Around 45 (end)

# Points to label
points_to_label <- c(left_index, peak_index, right_index)

# Add points and text labels
points(dens$x[points_to_label], dens$y[points_to_label], col =
"red", pch = 19)

# Label each point with x (expense) and y (density)
text(
  x = dens$x[points_to_label],
  y = dens$y[points_to_label],
  labels = paste0("(", round(dens$x[points_to_label], 1), ", ",
round(dens$y[points_to_label], 3), ")"),
  pos = 3,
  cex = 0.8,
```



```
  col = "black"
)

# Boxplot - horizontal
boxplot(data,
        horizontal = TRUE,
        main="Boxplot of Weekly Student Expense",
        xlab="Expense (CAD$)",
        col="orange",
        ylim=c(0,60))
```



References

1. Bruce P, Bruce A, Gedeck P. *Practical Statistics for Data Scientists*. 2nd ed. Sebastopol (CA): O’Reilly Media; 2020.