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.

1. 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).

1. 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

1. 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).

1. 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

1. 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.

1. 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.

1. 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.

1. 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.

1. 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.

1. 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.

1. 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.

1. 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).**

1. 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.

1. 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

1. 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)

* + 1. No Yes
    2. 0.45 0.55

1. 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.

1. 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.

1. 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

1. 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)

# 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))

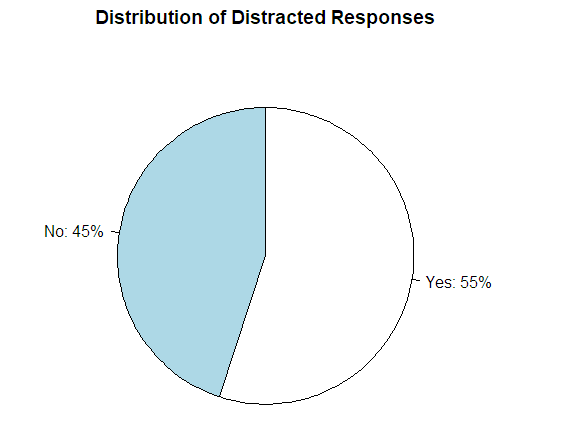
# Add labels above each bar

text(x = bar\_midpoints,

     y = freq\_table\_named + 1,  # Position slightly above bar

     labels = freq\_table\_named,

     cex = 1.2)  # Text size (optional)



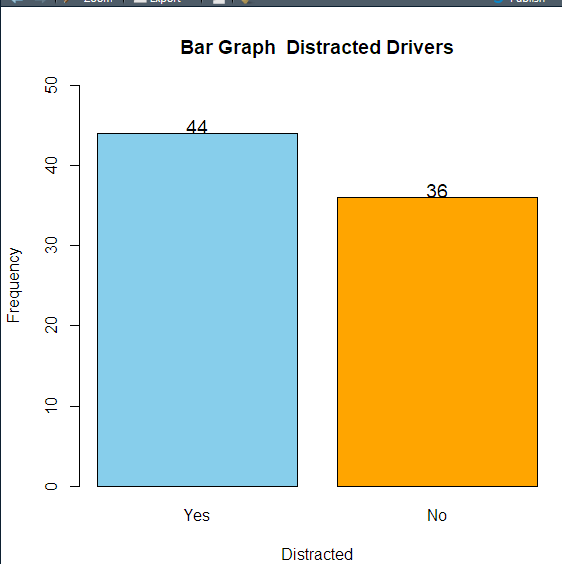
1. Provide a description of the above pie chart. [1 mark]

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.

1. 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.



1. 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.

1. 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.



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

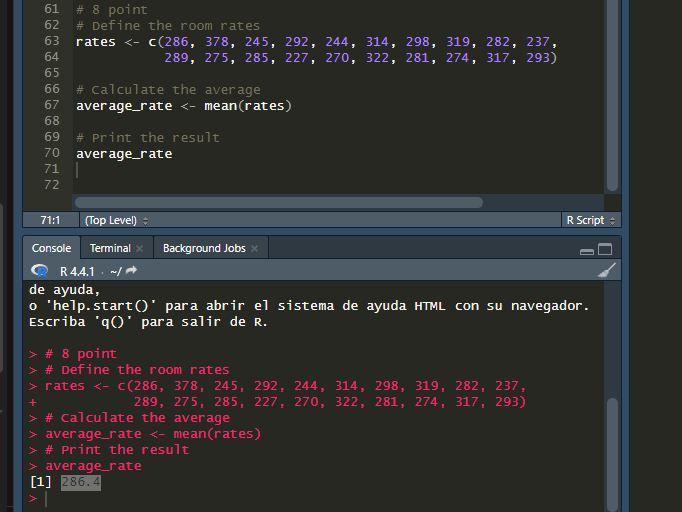
# Define the room rates

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



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

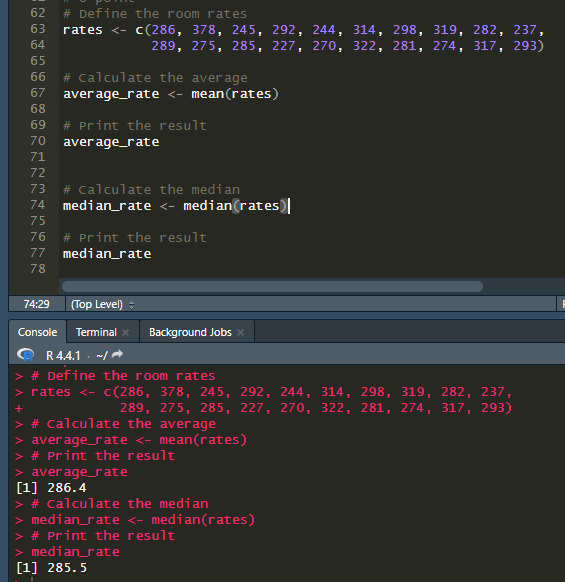
average\_rate # Print average

# Calculate the median

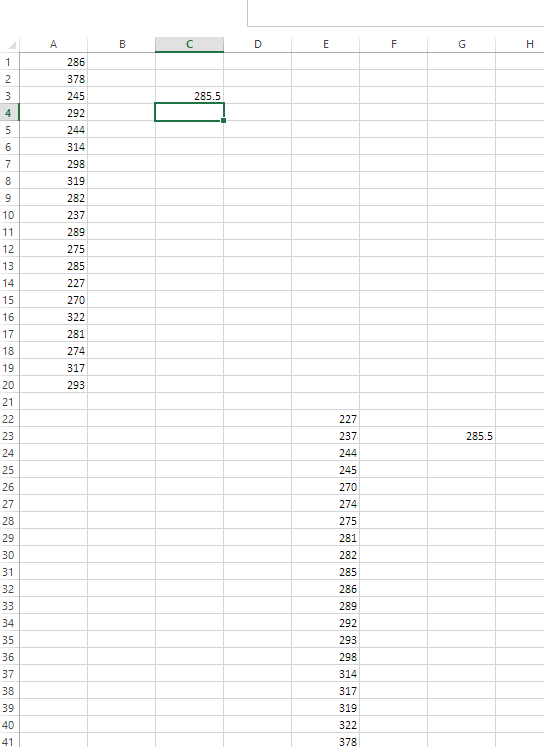
median\_rate <- median(rates) # Median of room rates

# Print the result

median\_rate # Print median



Excel job



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]

# 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

1. 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)

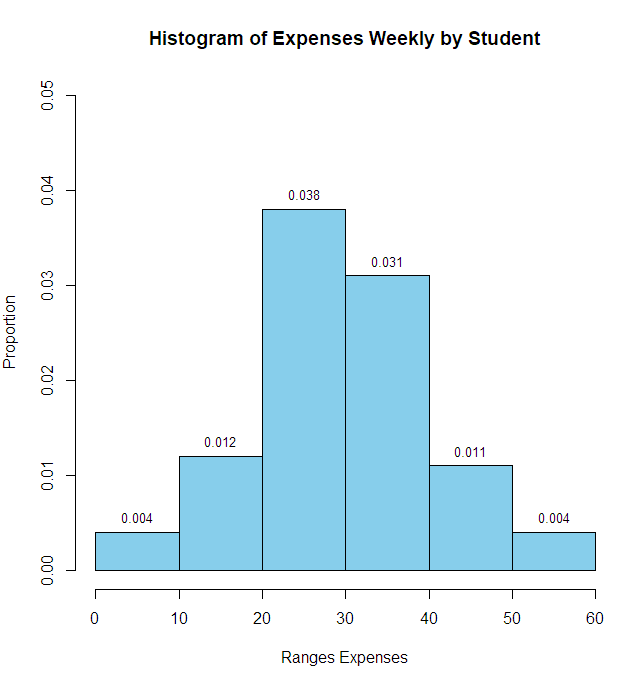
  labels = round(hist\_data$density, 3),  # Etiquetas redondeadas

  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],

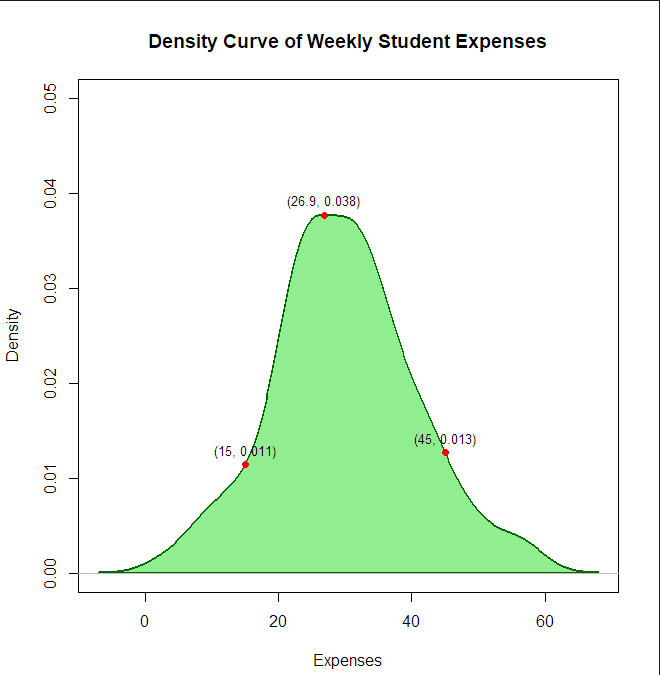
  labels = paste0("(", round(dens$x[points\_to\_label], 1), ", ", round(dens$y[points\_to\_label], 3), ")"),

  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,

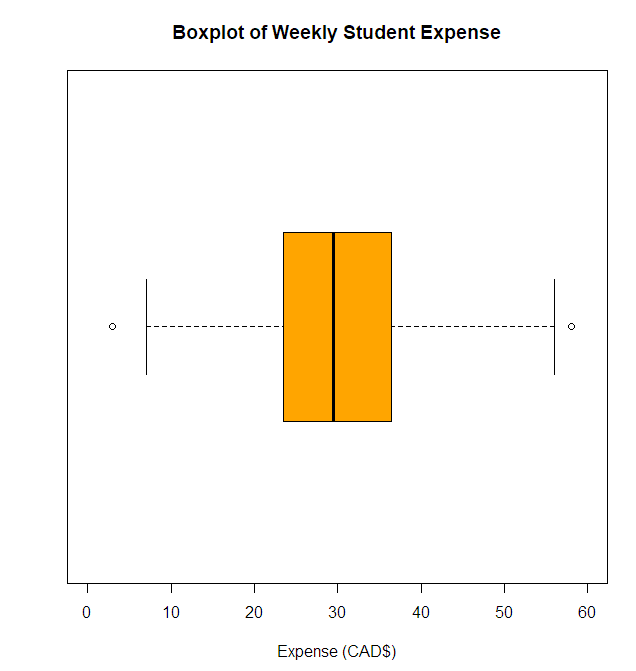
        horizontal = TRUE,

        main="Boxplot of Weekly Student Expense",

        xlab="Expense (CAD$)",

        col="orange",

        ylim=c(0,60))



1. A dietician wanted to find out how the total fat content (**Fat**; measured in grams per serving) is dependent on the amount of calories (**Calories**; measured in calories) among chicken burgers made from different fast food chains in Canada. A random sample of 20 chicken burgers was collected from different fast food chains (one burger per fast food chain) and the information was recorded. The data set is from “DANA4800\_HW1\_Q10\_Data.xlsx” on BrightSpace.
   1. Provide a description of the subjects of interest. **[1 mark]**

**Who (Subjects)** = Chicken burgers (one per fast food chain)  
**Where** = Fast food chains across Canada

* 1. Identify the role (or use) of the two variables. **[1 mark]**

**Variable1** The calories =independent (measured in calories)

**Variable2** the fat content = dependent (measured in grams per serving)

* 1. Use the *plot()* function to produce a scatterplot, based on the roles you defined in the above part. **[2 marks]**

*Note: Make sure you label your graph and axes appropriately.*

# 10.    A dietician wanted to find out how the total fat content (Fat; measured

# in grams per serving) is dependent on the amount of calories (Calories; measured

# in calories) among chicken burgers made from different fast food chains in Canada.

# A random sample of 20 chicken burgers was collected from different fast food chains

# (one burger per fast food chain) and the information was recorded.

# The data set is from “DANA4800\_HW1\_Q10\_Data.xlsx” on BrightSpace.

file\_HW1\_Q10 <- file.path(path, "DANA4800\_HW1\_Q10\_Data.xlsx")

group3 <- read\_excel(file\_HW1\_Q10)

summary(group3)

VaraibleCaloriesI <- group3$Calories

VariablefFatD <- group3$Fat

VaraibleCaloriesI

# Scatterplot or Scatter Diagram

plot(VaraibleCaloriesI, # X-variable

     VariablefFatD, # Y-variable

     main = "Scatterplot of Fat vs. Calories",

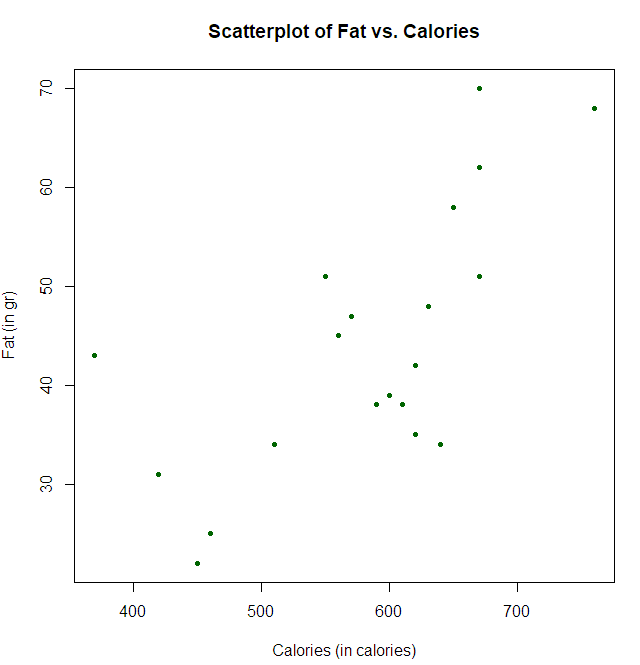
     ylab = "Fat (in gr)",

     xlab = "Calories (in calories)",

     cex = 1, # size of the dot

     pch = 20, # style of the dot, default is 1

     col = "darkgreen")



* 1. Provide a description of the above Scatterplot. **[2 marks]**

*Note: Please make sure the title and axes are properly labeled.*

|  |  |
| --- | --- |
| Direction | It shows a positive relationship because as the number of calories increases, the fat content also increases. |
| Strength: | it seems to have a strong relation, because they are close near to invisible line |
| Outliers: | Potentially one outlier, the first one on the left, over 40gr in fat less than 400 calories is the only one not share similar behavior than others |
| Form: | for the most part, it has a linear relation |

* 1. Use the *cov()* function to find the Variance-Covariance matrix. Please keep one decimal place only and identify which number is the covariance and which numbers are the variances of what. **[1+1 marks]**

file\_HW1\_Q10 <- file.path(path, "DANA4800\_HW1\_Q10\_Data.xlsx")

group3 <- read\_excel(file\_HW1\_Q10)

summary(group3)

VaraibleCaloriesI <- group3$Calories

VariablefFatD <- group3$Fat

VaraibleCaloriesI

# Scatterplot or Scatter Diagram

plot(VaraibleCaloriesI, # X-variable

     VariablefFatD, # Y-variable

     main = "Scatterplot of Fat vs. Calories",

     ylab = "Fat (in gr)",

     xlab = "Calories (in calories)",

     cex = 1, # size of the dot

     pch = 20, # style of the dot, default is 1

     col = "darkgreen")

# Covariance

cov\_matrix=cov(group3)

cov\_round <- round(cov\_matrix,digits=1)

cov\_round

> cov\_round

Calories Fat

Calories 9451.6 872.6

Fat 872.6 173.3

* 1. Use the cor() function to find the Correlation Coefficient. Please keep four decimal places and identify the value of the correlation coefficient. **[1 mark]**

Correlation coefficient for far and calories is **R=0.6818**

# Correlation Coefficient

cor\_coefficiente <- cor(group3) # Calculate correlation matrix

cor\_coefficiente\_round <- round(cor\_coefficiente, digits = 4) # Round correlation values

cor\_coefficiente\_round # Print correlation matrix

> cor\_coefficiente\_round

         Calories    Fat

Calories   1.0000 0.6818

Fat        0.6818 1.0000

* 1. Provide the description the above correlation coefficient (or most appropriate statistic in this situation). **[2 marks]**

The correlation coefficient **R=0.6818** indicates a **strong positive relationship** between fat and calories. This means that as the amount of fat increases, the number of calories tends to increase as well, which is consistent with the trend shown in the scatterplot.

1. Trying to accurately allocate labour hours in a moving job, the manager of a moving company would like to develop a method of predicting the labour hours (**Labour**; measured in hours) based on the size of the high-rise apartment (**Size**; measured in cubic feet). A random sample of 25 high-rise

apartment moves was randomly selected in downtown Vancouver in the previous calendar year. The data set is in “DANA4800\_HW1\_Q11\_Data.xlsx” on BrightSpace.

* 1. Provide a description of the subjects of interest. **[1 mark]**

**Who (Subjects)** = Apartment moves  
**Where** = Vancouver, downtown

**When = previous calendar year**

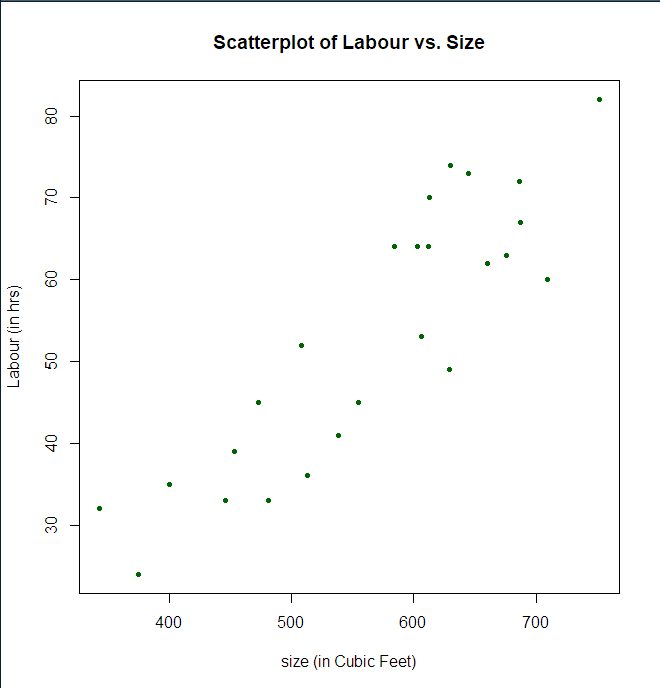
* 1. Identify the role (or use) of the two variables. **[1 mark]**

**Variable1** The **Labour** =independent (measured in Hours)

**Variable2** the **Size** = dependent (measured in cubic feet)

* 1. Use the *plot()* function to produce a Scatterplot, based on the roles you defined in the above part. **[2 marks]**

*Note: Make sure you label your graph and axes appropriately.*



* 1. Provide a description of the above scatter diagram. **[2 marks]**

|  |  |
| --- | --- |
| Direction | It shows a positive relationship because as the number of labour hours increases with the value of Apartment size. |
| Strength: | it seems to have a strong relation, because they are close near to invisible line |
| Outliers: | Seems there are two outliers, one at beginning near 60 hours of labour with approximately 200 ft3 and the another one are |
| Form: | The relationship appears to be linear, with most points following a straight-line pattern across the graph. |

* 1. Use the cor() function to find the correlation coefficient. Please keep four decimal places and identify the value of the correlation coefficient. **[1 mark]**

#

# 11. Trying to accurately allocate labour hours in a moving job, the manager of a

# moving company would like to develop a method of predicting the

# labour hours (Labour; measured in hours) based on the size of the high-rise apartment

# (Size; measured in cubic feet). A random sample of 25 high-rise

# apartment moves was randomly selected

# in downtown Vancouver in the previous calendar year.

# The data set is in “DANA4800\_HW1\_Q11\_Data.xlsx” on BrightSpace.

file\_HW1\_Q11 <- file.path(path, "DANA4800\_HW1\_Q11\_Data.xlsx")

group4 <- read\_excel(file\_HW1\_Q11)

summary(group4)

VaraibleI <- group4$Size

VariableD <- group4$Labour

# Scatterplot or Scatter Diagram

plot(VaraibleI, # X-variable

     VariableD, # Y-variable

     main = "Scatterplot of Labour vs. Size",

     ylab = "Labour (in hrs)",

     xlab = "size (in Cubic Feet)",

     cex = 1, # size of the dot

     pch = 20, # style of the dot, default is 1

     col = "darkgreen")

# Covariance

cov\_matrix=cov(group4)

cov\_round <- round(cov\_matrix,digits=1)

cov\_round

#correlation

cor\_coefficiente <- cor(group4)

cor\_coefficiente\_round <-round(cor\_coefficiente, digits=4)

cor\_coefficiente\_round

> cor\_coefficiente\_round

Size Labour

Size 1.0000 0.8857

Labour 0.8857 1.0000

* 1. Upon seeing the above correlation coefficient, an assistant reported it to the manager and said the following. Identify two major flaws of the statements. Briefly justify your answers. **[2+2 marks]**

*“Because the correlation coefficient 0.8857 cubic feet per hour is close to one, the reason of working long work hour is because of the high-rise apartment size only. ”*

**1st Flaw:**   
The correlation coefficient **does not have units** (e.g., "cubic feet per hour").  
**Justification:** Correlation measures the **strength and direction** of a linear relationship between two variables and is **unitless**. It only reflects how variables move together, not their actual measurements.

**2nd Flaw:** “**Correlation does not imply causation”.**  
**Justification:** A high correlation (like 0.8857) **does not prove that one variable causes the other**. There could be other factors influencing labour hours besides apartment size, such as layout complexity, materials used, or worker efficiency.

* 1. Provide the description the most appropriate statistic used in this situation (about apartment moving). **[2 marks]**

Statistic = the correlation coefficient between labor and size apartment of 25 high-rise apartment moves was randomly selected in downtown Vancouver in the previous calendar year,

1. The PopularKids data set was about opinions of a group of primary school students, who were stratified by their origin (rural, suburban and urban). More information about the data set can be in the following link: <https://www.openml.org/search?type=data&sort=runs&id=1100&status=active>

The data set is in “DANA4800\_HW1\_Q12\_Data.xlsx” on BrightSpace.

*Note: In this question, let us only use* ***Gender*** *(boy and girl) as the row variable and* ***Goal*** *(Grades, Popular, and Sports) as the column variable. Every subsequent mentioning of “row” and “column” refer to this definition*

1. Use the *table()* function to produce a Two-Way Table (or Contingency Table) with frequency. **[1 mark]**

# 12.   The PopularKids data set was about opinions of a group of primary school

# students, who were stratified by their origin (rural, suburban and urban).

# More information about the data set can be in the following

# link: https://www.openml.org/search?type=data&sort=runs&id=1100&status=active

# The data set is in “DANA4800\_HW1\_Q12\_Data.xlsx” on BrightSpace.

# Note: In this question, let us only use Gender (boy and girl) as the row

# variable and Goal (Grades, Popular, and Sports) as the column variable.

# Every subsequent mentioning of “row” and “column” refer to this definition

file\_HW1\_Q12 <- file.path(path, "DANA4800\_HW1\_Q12\_Data.xlsx")

group5 <- read\_excel(file\_HW1\_Q12)

summary(group5)

VaraibleGender <- group5$Gender

VariableGoal <- group5$Goals

df <- data.frame(VaraibleGender, VariableGoal)

df

twowaytable <- table(df)

twowaytable

> twowaytable <- table(df)

> twowaytable

VariableGoal

VaraibleGender Grades Popular Sports

boy 117 50 60

girl 130 91 30

1. Use the frequency table from part (a), calculate and enter expected frequencies in the following table. Please keep one decimal place in all entries. **[2 marks]**

### **📊 Calculation Table for Expected Frequencies**

| **Gender** | **Goal** | **Observed (O)** | **Expected (E)** | **Calculation (Row Total × Column Total) / Grand Total** |
| --- | --- | --- | --- | --- |
| **Boy** | Grades | 117 | 117.3 | (227 × 247) / 478 = 117.3 |
| **Boy** | Popular | 50 | 67.0 | (227 × 141) / 478 = 66.9 → 67.0 |
| **Boy** | Sports | 60 | 42.7 | (227 × 90) / 478 = 42.7 |
| **Girl** | Grades | 130 | 129.7 | (251 × 247) / 478 = 129.7 |
| **Girl** | Popular | 91 | 74.0 | (251 × 141) / 478 = 74.1 → 74.0 |
| **Girl** | Sports | 30 | 47.3 | (251 × 90) / 478 = 47.3 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Gender** | **Goals** | | | | |
|  | **Grade** | **Popular** | **Sports** |  |
| **Boy** | **117 (117.3 )** | **50 (67.0)** | **60 (42.7)** | **227** |
| **Girl** | **130 (129.7)** | **91 (74.0)** | **30 (47.3)** | **251** |
|  | **247** | **141** | **90** | **478** |

1. Manually calculate the 𝜒2-statistic. **[3 marks]**

*Please use the above two-way table with frequency to answer the following 3 questions.*

*Hint: You are expected to do this manually. But you could also use the margin.table() function to find the marginal totals first. There is an argument called MARGIN with three options. Please look up the R documentation for details.*

1. Find the percentage of students who are boys and their main goal is being popular. **[1 mark]**

1. Find the percentage of boys whose main goal is being popular. **[1 mark]**
2. Among the students whose main goal is being popular, find the percentage of them who are boys. **[1 mark]**

*Note: There are the same 3 MARGIN options in the proportion() function. Please look up the R documentation for details.*

# 12. The PopularKids data set was about opinions of a group of primary school

# students, who were stratified by their origin (rural, suburban and urban).

# More information about the data set can be in the following

# link: https://www.openml.org/search?type=data&sort=runs&id=1100&status=active

# The data set is in “DANA4800\_HW1\_Q12\_Data.xlsx” on BrightSpace.

# Note: In this question, let us only use Gender (boy and girl) as the row

# variable and Goal (Grades, Popular, and Sports) as the column variable.

# Every subsequent mentioning of “row” and “column” refer to this definition

file\_HW1\_Q12 <- file.path(path, "DANA4800\_HW1\_Q12\_Data.xlsx")

group5 <- read\_excel(file\_HW1\_Q12)

summary(group5)

VaraibleGender <- group5$Gender

VariableGoal <- group5$Goals

df <- data.frame(VaraibleGender, VariableGoal)

df

# Step 3: Create a two-way table (Gender as rows, Goal as columns)

twowaytable <- table(df)

print(twowaytable)

# Step 4: Compute column-wise proportions (margin = 2 means "among columns")

prop\_table <- prop.table(twowaytable, margin = 2)

print(round(prop\_table \* 100, 1))  # convert to percentages and round

> print(paste("Percentage of boys whose goal is Popular:", round(boy\_popular\_percentage, 1), "%"))

[1] "Percentage of boys whose goal is Popular: 35.5 %"

1. Use the *proportions()* function to produce a two-way table with Table Percentages. Please keep only two decimal places. **[1 mark]**

# Step 3: Create a two-way table (Gender as rows, Goal as columns)

twowaytable <- table(df)

print(twowaytable)

# Step 4: Compute column-wise proportions (margin = 2 means "among columns")

prop\_table <- prop.table(twowaytable, margin = 2)

print(round(prop\_table \* 100, 1))  # convert to percentages and round

# Step 5: Extract percentage of boys among those whose goal is "Popular"

boy\_popular\_percentage <- prop\_table["boy", "Popular"] \* 100

print(paste("Percentage of boys whose goal is Popular:", round(boy\_popular\_percentage, 1), "%"))

# Compute table percentages

# Round to two decimal places

table\_percentages <- prop.table(twowaytable) \* 100

rounded\_table\_percentages <- round(table\_percentages, 2)

rounded\_table\_percentages

> rounded\_table\_percentages

VariableGoal

VaraibleGender Grades Popular Sports

boy 24.48 10.46 12.55

girl 27.20 19.04 6.28

1. Use the *proportions()* function to produce a two-way table with Row Percentages. Please keep only two decimal places. **[1 mark]**

# Step 3: Create a two-way table (Gender as rows, Goal as columns)

twowaytable <- table(df)

print(twowaytable)

# Step 4: Compute column-wise proportions (margin = 2 means "among columns")

prop\_table <- prop.table(twowaytable, margin = 2)

print(round(prop\_table \* 100, 1))  # convert to percentages and round

# Step 5: Extract percentage of boys among those whose goal is "Popular"

boy\_popular\_percentage <- prop\_table["boy", "Popular"] \* 100

print(paste("Percentage of boys whose goal is Popular:", round(boy\_popular\_percentage, 1), "%"))

# Compute table percentages

# Round to two decimal places

table\_percentages <- prop.table(twowaytable) \* 100

rounded\_table\_percentages <- round(table\_percentages, 2)

rounded\_table\_percentages

# Compute row-wise proportions Round to two decimal places

row\_percentages <- proportions(twowaytable, margin = 1) \* 100

rounded\_row\_percentages <- round(row\_percentages, 2)

rounded\_row\_percentages

> rounded\_row\_percentages

VariableGoal

VaraibleGender Grades Popular Sports

boy 51.54 22.03 26.43

girl 51.79 36.25 11.95

1. Use the *proportions()* function to produce a two-way table with Column Percentages. Please keep only two decimal places. **[1 mark]**

# Step 3: Create a two-way table (Gender as rows, Goal as columns)

twowaytable <- table(df)

print(twowaytable)

# Step 4: Compute column-wise proportions (margin = 2 means "among columns")

prop\_table <- prop.table(twowaytable, margin = 2)

print(round(prop\_table \* 100, 1))  # convert to percentages and round

# Step 5: Extract percentage of boys among those whose goal is "Popular"

boy\_popular\_percentage <- prop\_table["boy", "Popular"] \* 100

print(paste("Percentage of boys whose goal is Popular:", round(boy\_popular\_percentage, 1), "%"))

# Compute table percentages

# Round to two decimal places

table\_percentages <- prop.table(twowaytable) \* 100

rounded\_table\_percentages <- round(table\_percentages, 2)

rounded\_table\_percentages

# Compute row-wise proportions Round to two decimal places

row\_percentages <- proportions(twowaytable, margin = 1) \* 100

rounded\_row\_percentages <- round(row\_percentages, 2)

rounded\_row\_percentages

> print(round(prop\_table \* 100, 1)) # convert to percentages and round

VariableGoal

VaraibleGender Grades Popular Sports

boy 47.4 35.5 66.7

girl 52.6 64.5 33.3

1. Use the *barplot()* function to produce a Side-by-Side Bar Graph, with the variable Goals on the x-axis, column percentages on the y-axis, and including a legend. The title of the graph should say “Side-by-side Bar Graph of Goals by Gender”. **[2 marks]**

# Step 1: Get column percentages (proportions within each Goal)

col\_percents <- prop.table(twowaytable, margin = 2) \* 100

# Step 2: Create a side-by-side barplot

barplot(

  col\_percents,

  beside = TRUE,

  col = c("lightgreen", "pink"),

  ylim = c(0, 100),

  main = "Side-by-side Bar Graph of Goals by Gender",

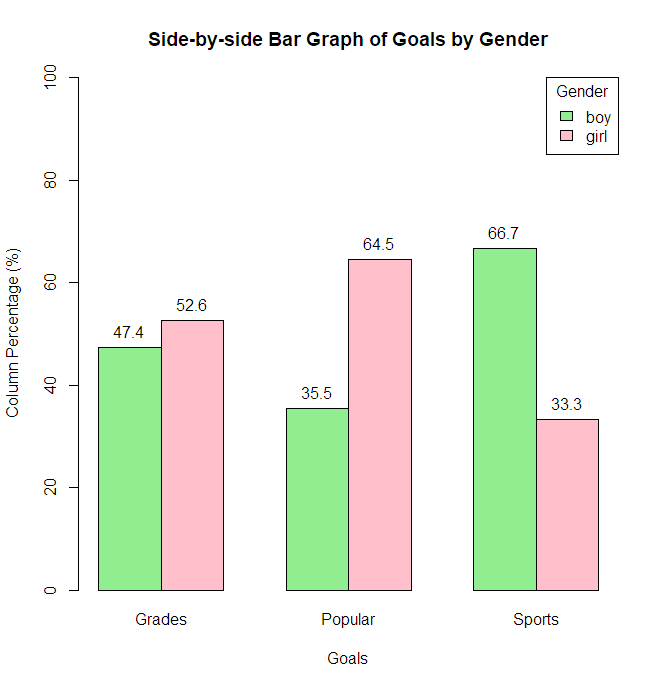
  ylab = "Column Percentage (%)",

  xlab = "Goals",

  legend.text = rownames(col\_percents),

  args.legend = list(title = "Gender", x = "topright")

)



1. Provide a description of the above graph. **[1 mark]**

The side-by-side bar graph shows how boys and girls are distributed within each goal category (Grades, Popular, Sports):

Grades: Slightly more girls than boys aim for good grades.

Popular: A noticeably higher percentage of girls (64.5%) than boys (35.5%) prioritize being popular.

Sports: A significantly higher proportion of boys (66.7%) than girls (33.3%) select sports as their main goal.

The graph reveals gender differences in goal preferences. Girls are more likely to choose "Grades" or "Popularity", while boys are more inclined toward "Sports".

# References

1. Bruce P, Bruce A, Gedeck P. *Practical Statistics for Data Scientists*. 2nd ed. Sebastopol (CA): O’Reilly Media; 2020.
2. Michael LO, *Module 05:EDA – Two Variables,*2025