Note: The first 3 questions are from Q10-12 in HW1.

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_HW2_Q1_Data.xlsx” on BrightSpace.

   a) 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

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

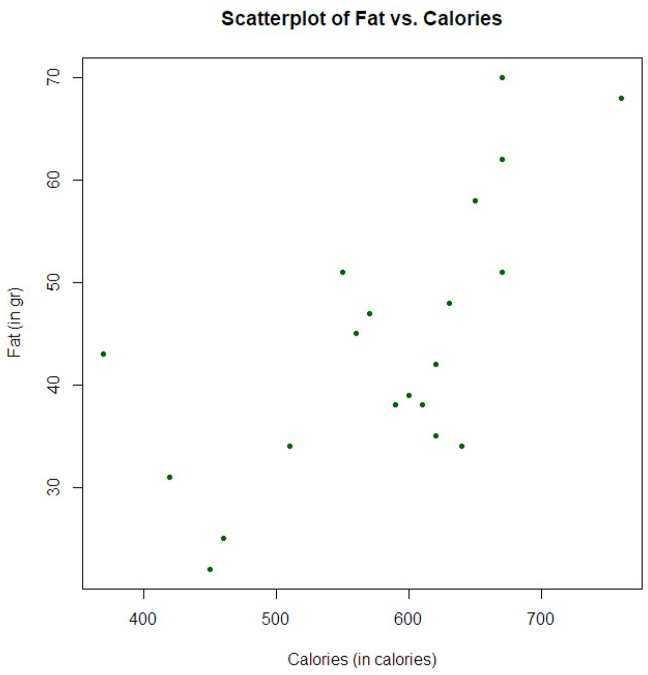
      Variable1 **The calories =independent (measured in calories)**

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

   c)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")
```

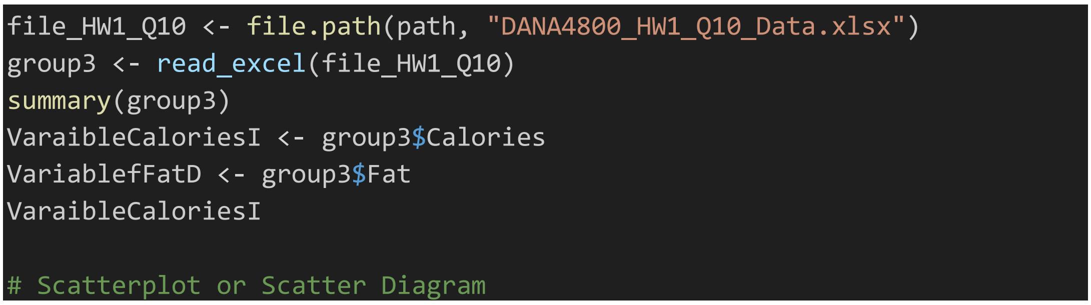


d) Provide a description of the above Scatterplot. **[2 marks]**

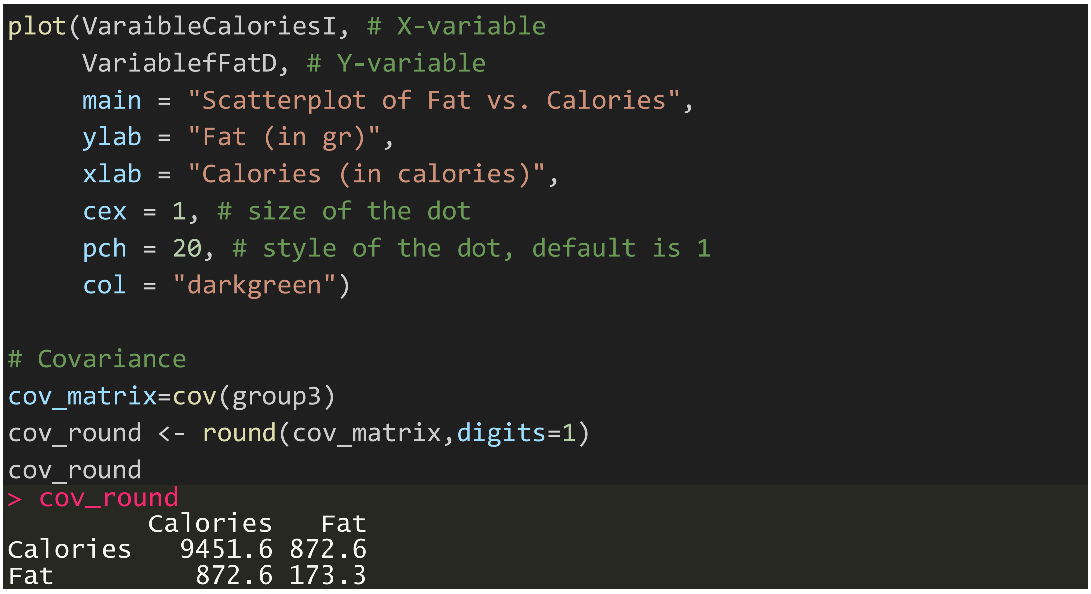
Note: Please make sure the title and axes are properly labeled.

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

e) 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]**



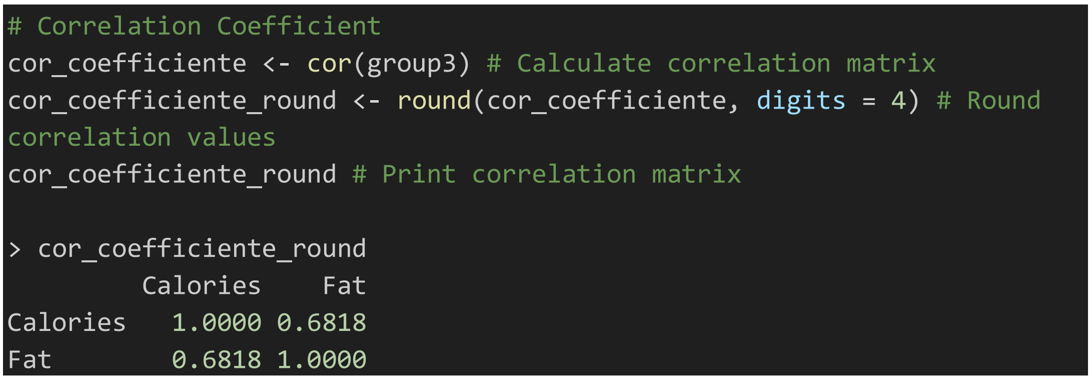
```
file_HW1_Q10 <- file.path(path, "DANA4800_HW1_Q10_Data.xlsx")
group3 <- read_excel(file_HW1_Q10)
summary(group3)
VaraibleCaloriesI <- group3$Calories
VariablefFatD <- group3$Fat
VaraibleCaloriesI

# 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
```

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

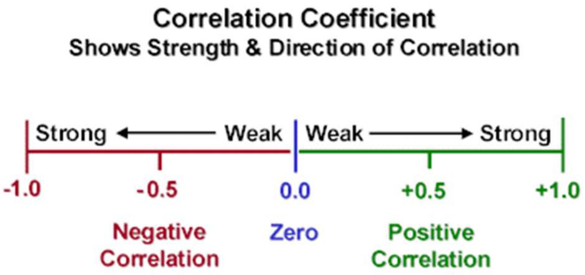
Correlation coefficient for far and calories is **R=0.6818**



```
# 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
```

g) 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.



(LO,2025)

2. 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_HW2_Q2_Data.xlsx" on BrightSpace.

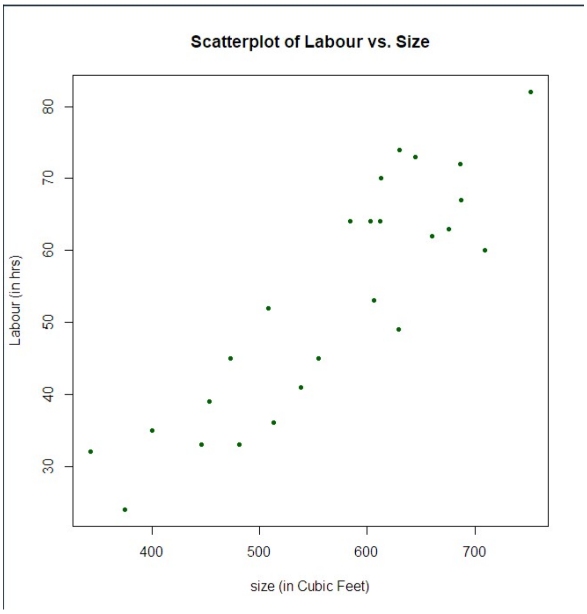
   a) Provide a description of the subjects of interest. **[1 mark]**

      Who (Subjects) = Apartment moves
      Where = Vancouver, downtown
      When = previous calendar year

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

      Variable1 the **Labour** = dependent (measured in Hours)
      Variable2 the **Size** = independent (measured in cubic feet)

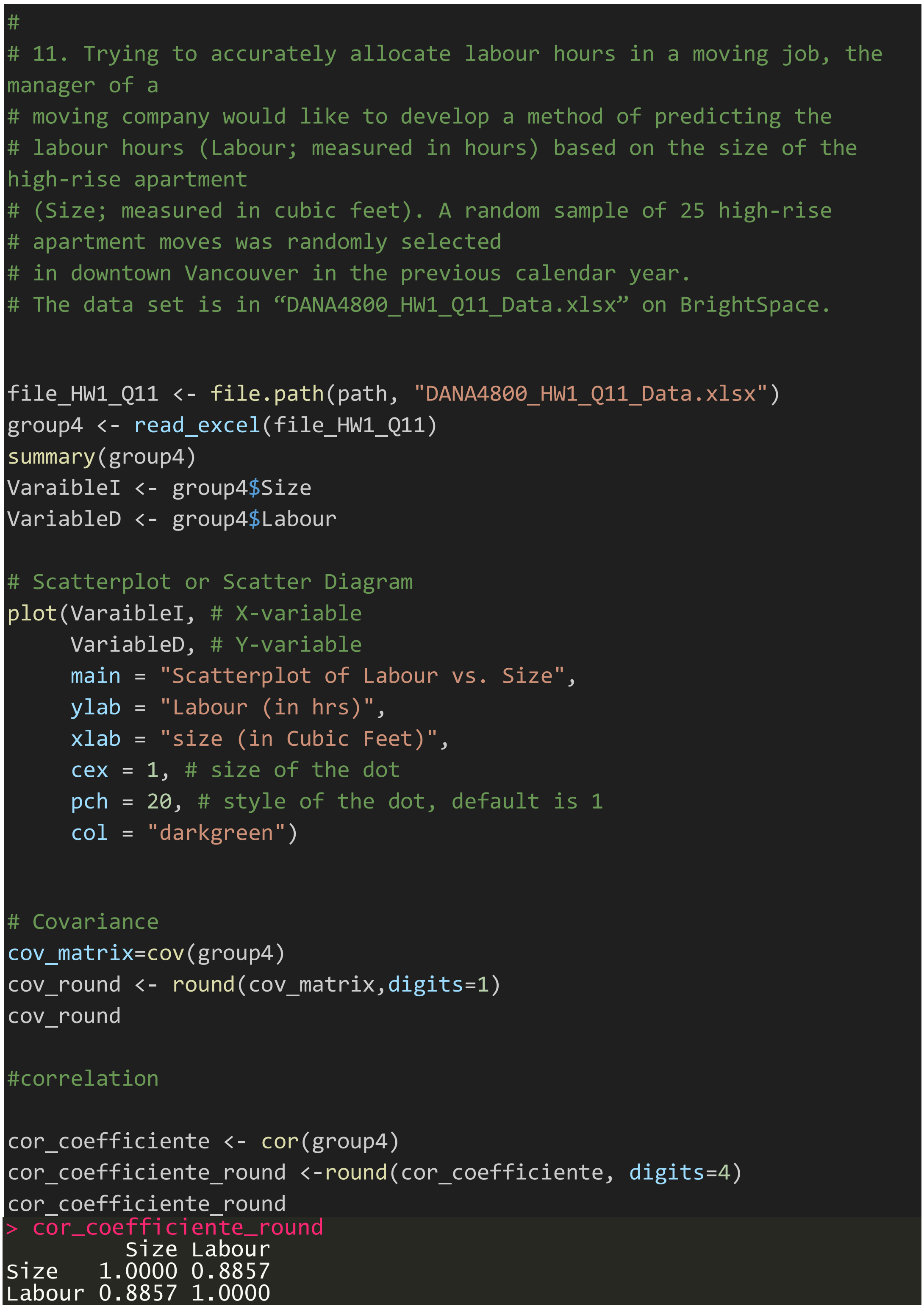
   c) Use the *plot()* function to produce a Scatterplot, based on the roles you defined in the above part. **[2 marks]**
   Note: Make sure you label your graph and axes appropriately.



   d) 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.

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



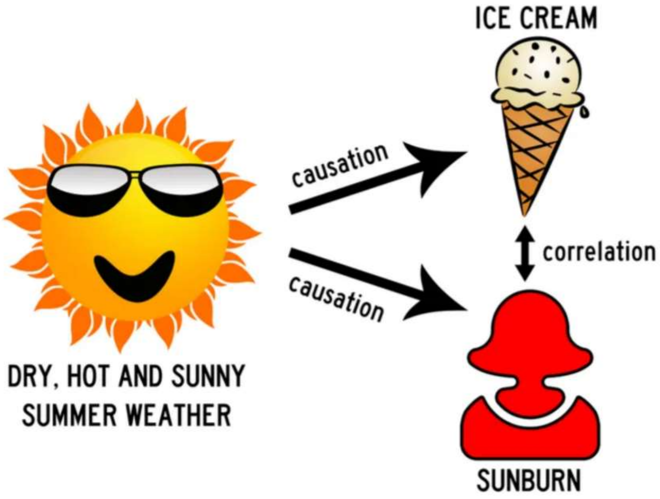
```
#
# 11. Trying to accurately allocate labour hours in a moving job, the
manager of a
# moving company would like to develop a method of predicting the
# labour hours (Labour; measured in hours) based on the size of the
high-rise apartment
# (Size; measured in cubic feet). A random sample of 25 high-rise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

f) 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.

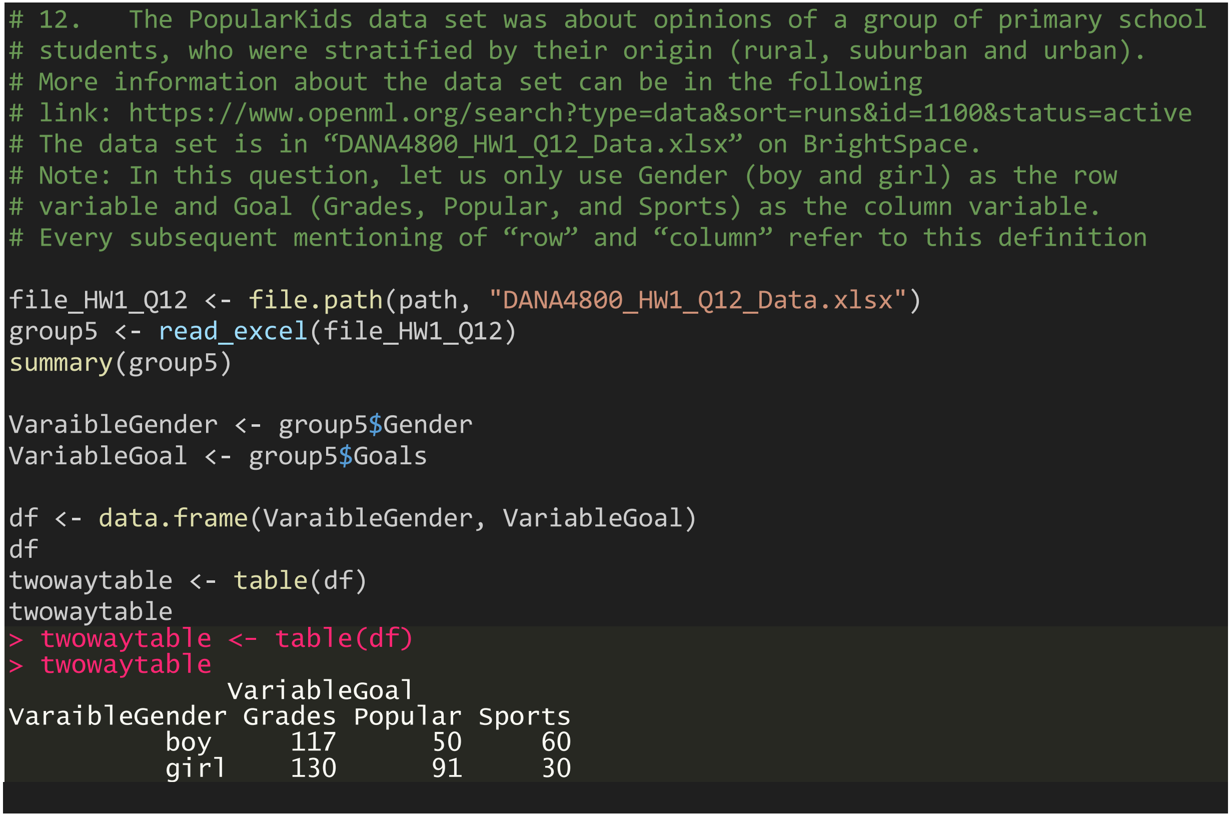


(LO,2025)

g) 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,

3. 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
*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.*
The data set is in "DANA4800_HW2_Q3_Data.xlsx" on BrightSpace.
   a) 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
```

   b) 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

		Goals			
		Grade	Popular	Sports	
Gender	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

c) Manually calculate the χ^2-statistic. **[3 marks]**
Please use the above two-way table with frequency to answer the following 3 questions.

$$\chi^2 = \sum \frac{(f-e)^2}{e}$$

$$X^2 - statistic = \frac{(117-117.3)^2}{117.3} + \frac{(50-67)^2}{67} + \frac{(60-42.7)^2}{42.7} + \frac{(130-129.7)^2}{129.7} + \frac{(91-74)^2}{74} + \frac{(30-47.3)^2}{47.3}$$

$$X^2 - statistic = 21.6$$

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.

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

$$popular\ \% = \frac{50}{478} = 10.5\%$$

e) Find the percentage of boys whose main goal is being popular. **[1 mark]**

$$\text{popular\%} = \frac{50}{227} = 22\%$$

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

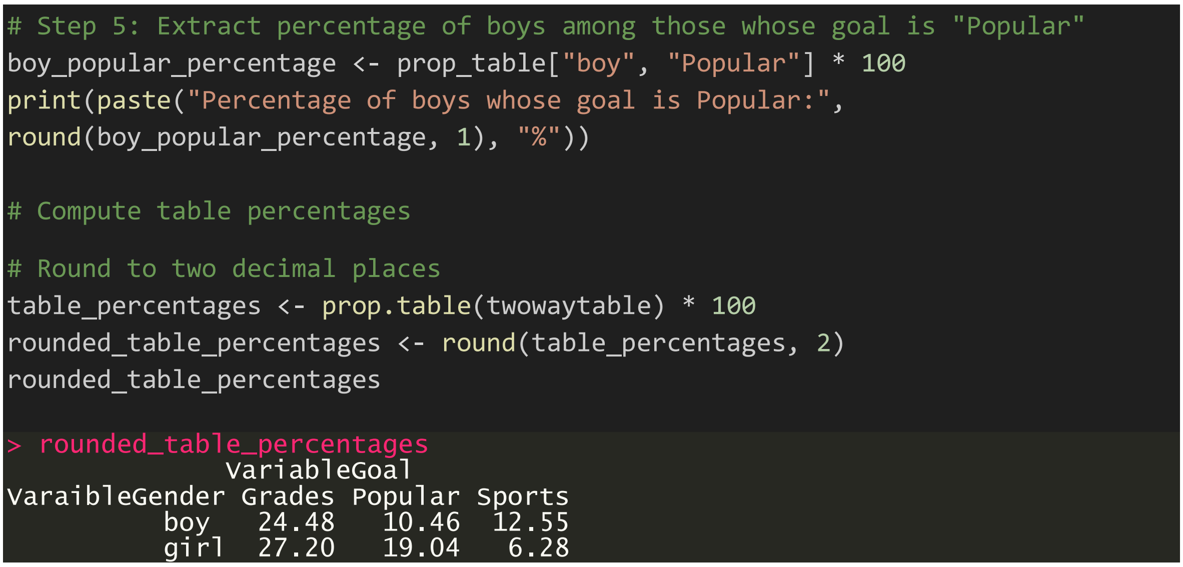
Note: There are the same 3 MARGIN options in the proportion() function. Please look up the R documentation for details.

$$\% = \frac{50}{141} = 35.5\%$$

```
# 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_po
pular_percentage, 1), "%"))
[1] "Percentage of boys whose goal is Popular: 35.5 %"
```

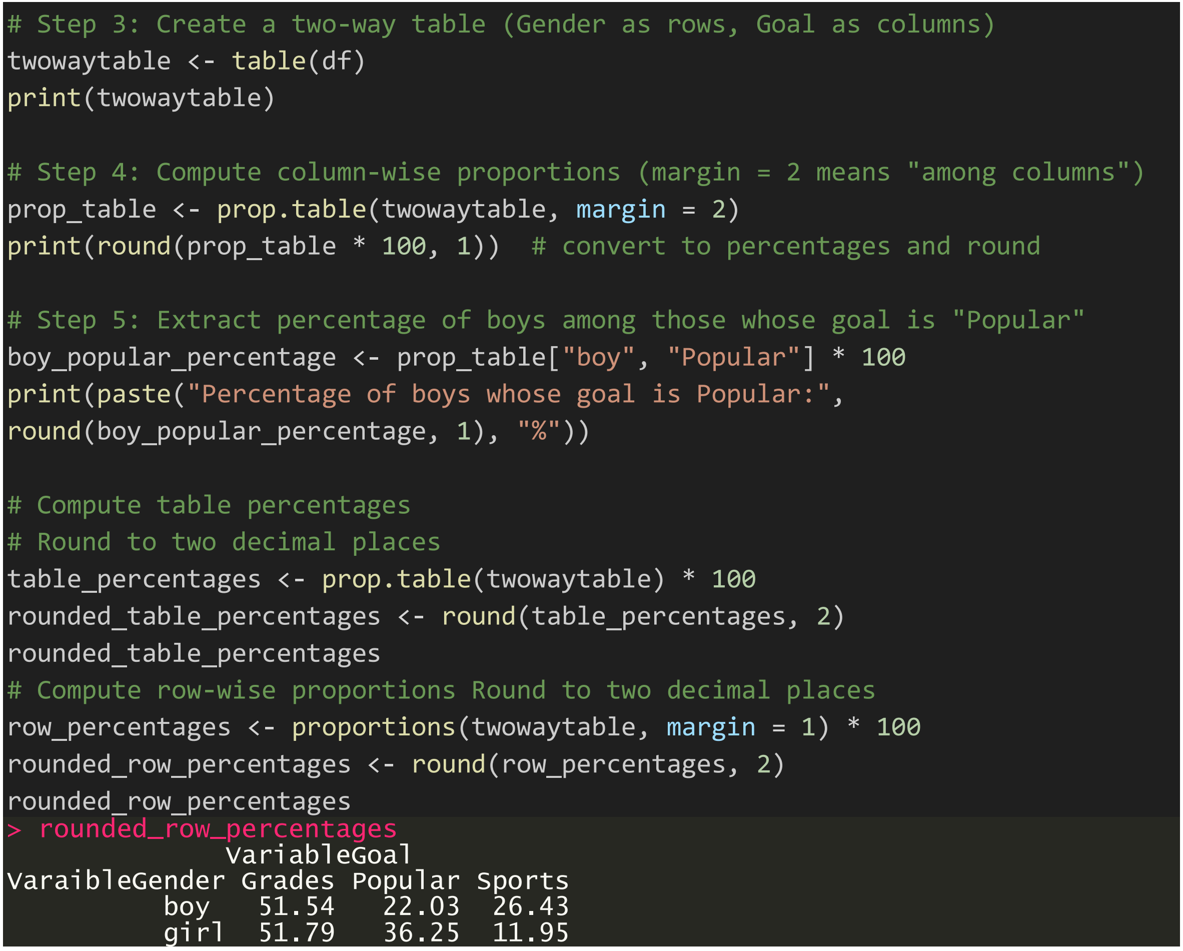
g) Use the *proportions()* function to produce a two-way table with Table Percentages. Please keep only two decimal places. **[1 mark]**

```
  # 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
```



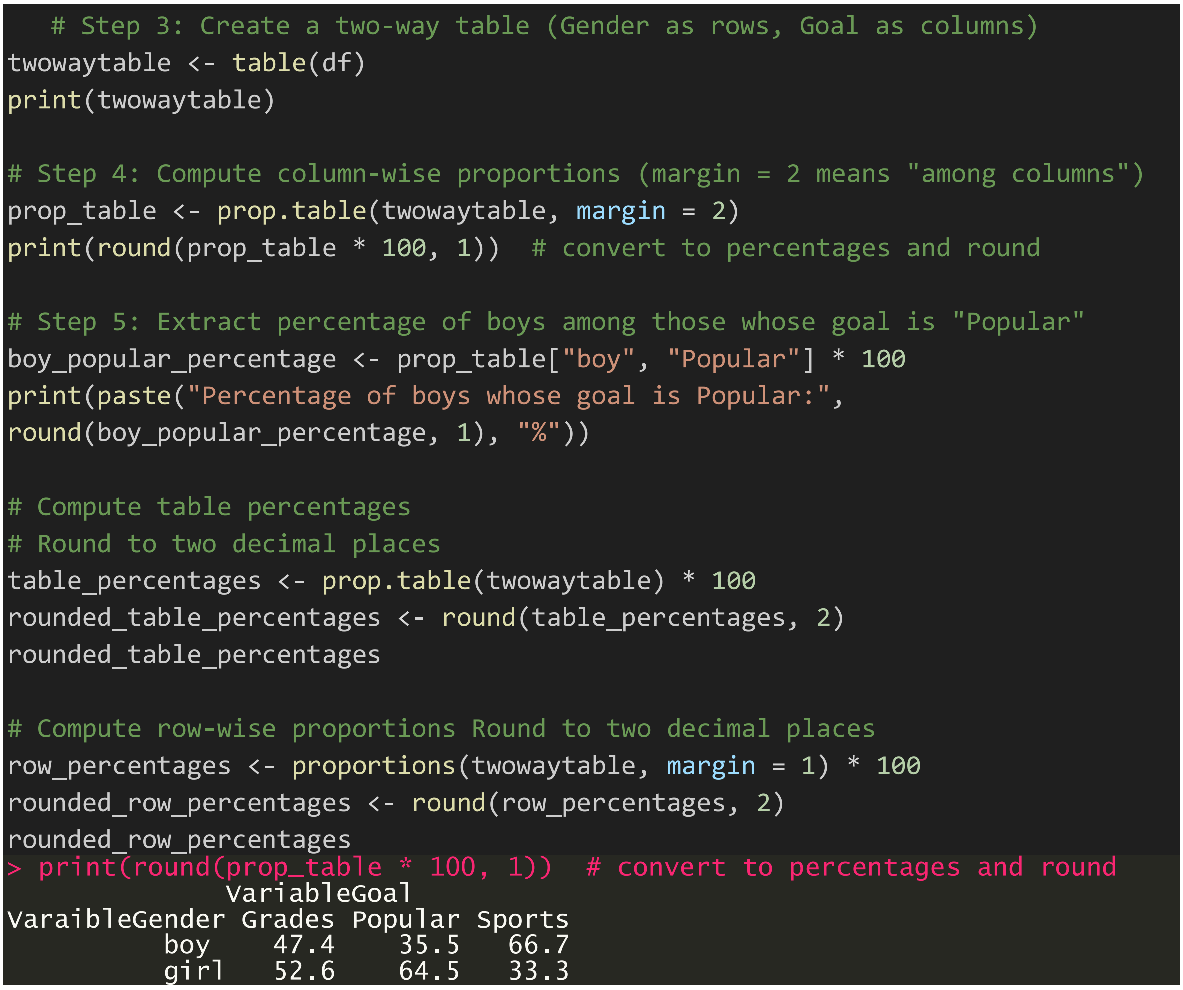
```
# Step 5: Extract percentage of boys among those whose goal is "Popular"
boy_popular_percentage <- prop_table["boy", "Popular"] * 100
print(paste("Percentage of boys whose goal is Popular:",
round(boy_popular_percentage, 1), "%"))

# 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
```

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



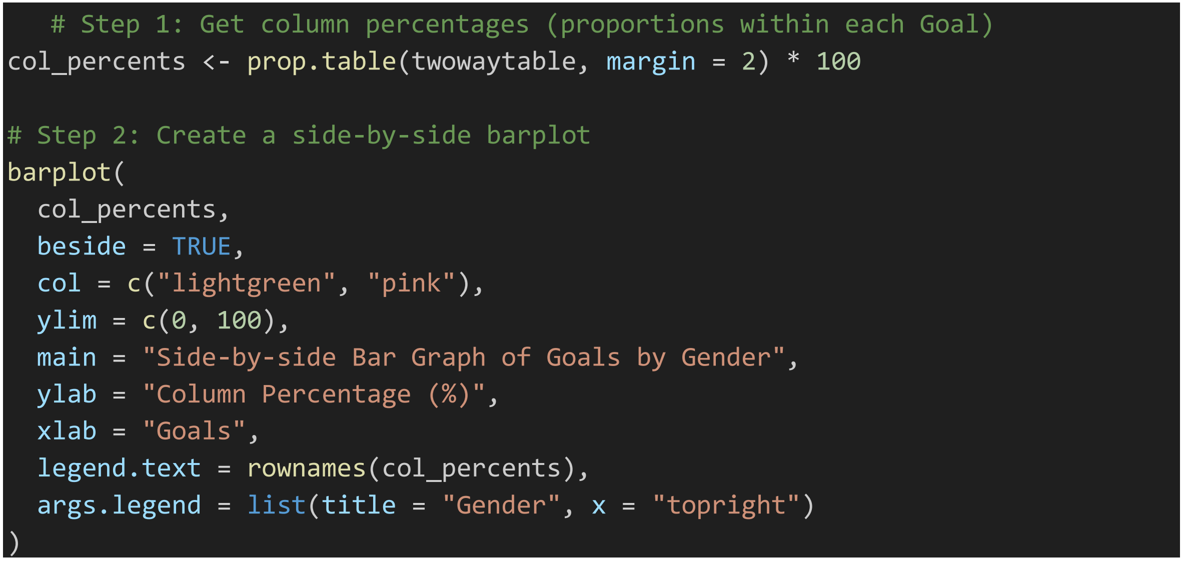
```
# Step 3: Create a two-way table (Gender as rows, Goal as columns)
twowaytable <- table(df)
print(twowaytable)

# 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
```

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

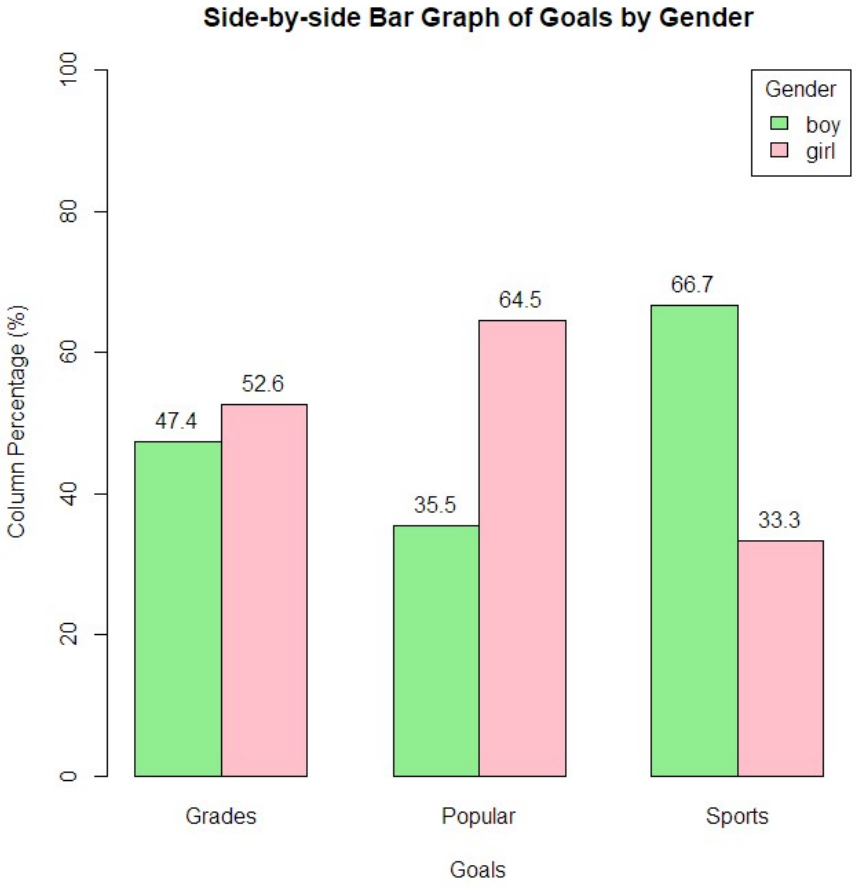


```
   # Step 3: Create a two-way table (Gender as rows, Goal as columns)
twowaytable <- table(df)
print(twowaytable)

# 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
```

j) 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")
)
```



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

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".

4. In a local high school, 25% of all Grade 11 students play basketball and 20% of all Grade 11 students play volleyball. It is also estimated that 5% of all Grade 11 students play both sports. Use a two-way table, or otherwise any other method, to find the following probabilities.

a) When a Grade 11 student is randomly selected, what is the probability that one is playing either basketball or volleyball? Provide an interpretation of the answer. **[2+2 marks]**

- P(B)=0.25 → probability the student plays basketball
- P(V)=0.20P(V) = 0.20P(V)=0.20 → probability the student plays volleyball
- P(B∩V)=0.05 → probability the student plays both

	Basketball	**No basketball**	
Volleyball	**5**	**15**	**20**
No volleyball	**20**	**60**	**80**
	25	**75**	**100**

$$P(B \cup V) = P(B) + P(V) - P(B \cap V) = 0.25 + 0.20 - 0.05 = 0.40$$
$$= 0.40$$

The probability a student play basketball or volleyball is 40%

b) When a Grade 11 student is randomly selected, what is the probability that one is playing neither basketball nor volleyball? **[1 mark]**

$$\mathbf{P(Neither) = 1 - P(Basketball\ or\ Volleyball) = 1 - 0.40 = 0.60}$$

The probability a student don't play basketball or volleyball is 60%

c) When a Grade 11 student who is playing volleyball is randomly selected, what is the probability that one is also playing basketball? **[2 marks]**

P(Basketball|Volleyball)= P(Basketball and Volleyball) / P(Volleyball)

$$P(B \mid V) = 0.05/0.20 = 0.25$$

d) When a Grade 11 student who is playing basketball is randomly selected, what is the probability that one is also playing volleyball? **[2 marks]**

The probability is 20%

$$\mathbf{P(V \mid B) = 0.5/0.25 = 0.20}$$

e) Define event A as Grade 11 students playing basketball and event B as Grade 11 students playing volleyball. Are the two events A and B mutually exclusive? Briefly justify your answer using some probability calculations. **[0+2 marks]**

A=Basketball

B=Volleyball the events are not mutually exclusive

$$P(A\ \cap B) = P(A) + P(B) - P(A\ \cup B)$$

$$P(A\ \cap B) = 0.25 + 0.20 - 0.40 = 0.05$$

$$P(A\ \cap B) \neq 0$$

f) Define event A as Grade 11 students playing basketball and event B as Grade 11 students playing volleyball. Are the two events A and B independent? Briefly justify your answer using some probability calculations. **[0+2 marks]**

Special case: When 2 events are independent to each other

$$\boldsymbol{P(A\ and\ B) = P(A) \times P(B)}$$

$$P(A\ \cap B) = 0.05$$

$$\boldsymbol{P(A) \times P(B) = 0.25 * 0.20 = 0.05}$$

Since $\boldsymbol{P(A\ \cap B) = P(A) \times P(B)}$, events A and B are **independent**.

5. Suppose that 85% of shoppers buy coffee at this local café, 65% of shoppers buy baked goods (like muffins and cookies etc.) at this local café, and 90% of shoppers buy either coffee or baked goods.

a) Construct a two-way table of the situation above. **[2 marks]**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.85 + 0.65 - 0.90 = 0.60$$

		Buy Coffee		
		Yes (A)	**No (A^c)**	
Buy Baked Goods	**Yes (B)**	60	5	65
	No (B^c)	25	10	35
		85	15	100

Please try to use the two-way table method to do the following 3 parts.

A = "shopper buys coffee"

B = "shopper buys baked goods"

b) When a shopper visiting this local café is randomly selected, what is the probability that one is buying coffee and baked goods? **[1 mark]**

$$\boldsymbol{P(A \cap B) = 60/100}$$

The probability is 60%.

c) Given a shopper visiting this local café is buying baked goods, what is the probability that one is also buying coffee? **[1 mark]**

$$\boldsymbol{P(A \mid B) = \frac{P(A \cap B)}{P(B)} = 60/65 \approx 0.923}$$

The probability is 92.3%.

d) Given a shopper visiting this local café is buying coffee, what is the probability that one is also buying baked goods? **[1 mark]**

$$\boldsymbol{P(B \mid A) = \frac{P(A \cap B)}{P(A)} = 60/85 \approx 0.706}$$

The probability is 70.6%.

6. Suppose that a city has three airports (Airport A, Airport B, and Airport C). Suppose Airport A handles 50% of the air airline traffic, Airport B and Airport C handle 30% and 20% of the air traffic respectively. Suppose the detection rates of weapon (W = weapon) at the three airports are 0.95,
0.55 and 0.40 respectively.

a) Define all necessary events, assign a symbol (or letter) to each one and identify all six probabilities. **[3 marks]**

P(A)=0.50
P(B)=0.30
P(C)=0.20

A=Passenger at Airport A
A^c=Passenger not at Airport A
B=Passenger at Airport B
B^c=Passenger not at Airport B
C=Passenger at Airport C
C^c=Passenger not at Airport C

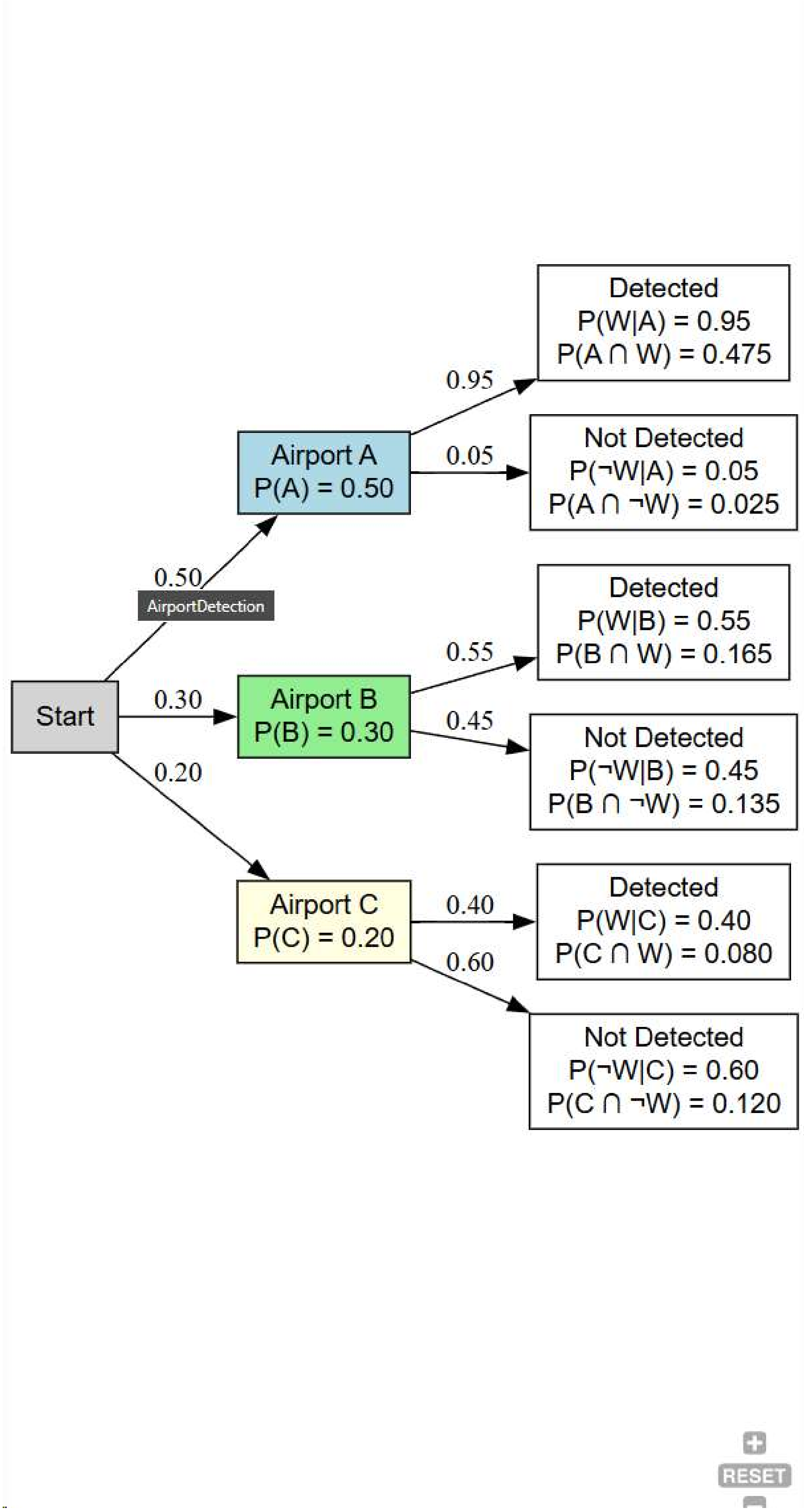
W= Passenger with a weapon
W^c: A weapon is **not** detected

P(W|A)= Probability of detecting a passenger with a weapon at Airport A
P(W^c|A)= Probability of not detecting a passenger with a weapon at Airport A
P(W|B)= Probability of detecting a passenger with a weapon at Airport B
P(W^c|B)= Probability of not detecting a passenger with a weapon at Airport B
P(W|C)= Probability of detecting a passenger with a weapon at Airport C

P(W^c|C)= Probability of not detecting a passenger with a weapon at Airport C

b) Draw a tree diagram showing the above situation. Please also compute all 6 joint probabilities (on the far right). **[2+2 marks]**
https://dreampuf.github.io/GraphvizOnline **code**

```
digraph AirportDetection {
    rankdir=LR;  // Left to right layout
    node [shape=rectangle, style=filled, fillcolor=lightgray, fontname="Arial"];
    Start [label="Start"];

    // Airport nodes
    A [label="Airport A\nP(A) = 0.50", fillcolor=lightblue];
    B [label="Airport B\nP(B) = 0.30", fillcolor=lightgreen];
    C [label="Airport C\nP(C) = 0.20", fillcolor=lightyellow];

    // Detection outcomes for A
    AW [label="Detected\nP(W|A) = 0.95\nP(A ∩ W) = 0.475", fillcolor=white];
    AnW [label="Not Detected\nP(¬W|A) = 0.05\nP(A ∩ ¬W) = 0.025",
fillcolor=white];

    // Detection outcomes for B
    BW [label="Detected\nP(W|B) = 0.55\nP(B ∩ W) = 0.165", fillcolor=white];
    BnW [label="Not Detected\nP(¬W|B) = 0.45\nP(B ∩ ¬W) = 0.135",
fillcolor=white];

    // Detection outcomes for C
    CW [label="Detected\nP(W|C) = 0.40\nP(C ∩ W) = 0.080", fillcolor=white];
    CnW [label="Not Detected\nP(¬W|C) = 0.60\nP(C ∩ ¬W) = 0.120",
fillcolor=white];

    // Connections
    Start -> A [label="0.50"];
    Start -> B [label="0.30"];
    Start -> C [label="0.20"];

    A -> AW [label="0.95"];
    A -> AnW [label="0.05"];

    B -> BW [label="0.55"];
    B -> BnW [label="0.45"];

    C -> CW [label="0.40"];
    C -> CnW [label="0.60"];
}
```


Start
0.50
AirportDetection
0.30
0.20
Airport A
P(A) = 0.50
Airport B
P(B) = 0.30
Airport C
P(C) = 0.20
0.95
0.05
0.55
0.45
0.40
0.60
Detected
P(W|A) = 0.95
P(A ∩ W) = 0.475
Not Detected
P(¬W|A) = 0.05
P(A ∩ ¬W) = 0.025
Detected
P(W|B) = 0.55
P(B ∩ W) = 0.165
Not Detected
P(¬W|B) = 0.45
P(B ∩ ¬W) = 0.135
Detected
P(W|C) = 0.40
P(C ∩ W) = 0.080
Not Detected
P(¬W|C) = 0.60
P(C ∩ ¬W) = 0.120
RESET

c) If a passenger is randomly selected at one of the three airports and is found to be carrying a weapon through the boarding gate, what is the probability that the passenger is using Airport A? **[2 marks]** Note: Please try to use the Bayes' Theorem to do this.

$$\boldsymbol{P(A \cap W) = P(A)\ x\ P(W|A) = 0.50 \times 0.95 = 0.475}$$
$$\boldsymbol{P(B \cap W) = P(B)\ x\ P(W|B) = 0.30 \times 0.55 = 0.165}$$
$$\boldsymbol{P(C \cap W) = P(C)\ x\ P(W|C) = 0.20 \times 0.40 = 0.080}$$

$$\boldsymbol{P(\,A \mid W\,)} = \frac{\boldsymbol{P(A \cap W)}}{\boldsymbol{P(A \cap W) + P(B \cap W) + P(C \cap W)}}$$

$$\boldsymbol{P(\,A \mid W\,)} = \frac{0.475}{0.475 + 0.165 + 0.08} = 0.660$$

The probability is 66%

d) Repeat part (c) with Airport C? **[2 marks]** Note: Please try to use two-way table to do this.

%	Weapon (W)	without Weapon (W^C)	Total
Airport A	47.5	2.5	50
Airport B	16.5	13.5	30
Airport C	8	12	20
Total	72	28	100

$$\boldsymbol{P(\,C \mid W\,)} = \frac{0.08}{0.475 + 0.165 + 0.08} = 0.11$$

The probability is 11.1%

e) Note that P(A) < P(A|W). (Note also that P(C) > P(C|W).) What implication does it have? Try to explain this to the Director of Aviation at Transport Canada, who has no clue what Bayes' Theorem is about. **[2+2 marks]**

When P(A)<P(A|W), it means that although 50% of passengers go through Airport A, a detected weapon makes it more likely the person came from A due to its high detection rate (95%).

When P(C)>P(C|W), even though 20% of traffic is from Airport C, a detected weapon makes it less likely the person came from C because of its low detection rate (40%).

7. Suppose that a certain disease is present in 20% of the population, and that there is a screening test designed to detect if this disease is present. If the screening test is applied to those who have this disease, 75% of the time it will give a positive result. Another way of saying it is that the "**true positive rate**" is 75%. Also, if the screening test is applied to those who do not have this disease, 60% of the time it will give a negative result. It is also called "**true negative rate**".

a) Define all necessary events, assign a symbol (or letter) to each one and identify all three probabilities. **[3 marks]**

A = person has the disease
A^c = person does NOT have the disease
B = person tests positive
B^c = person tests negative

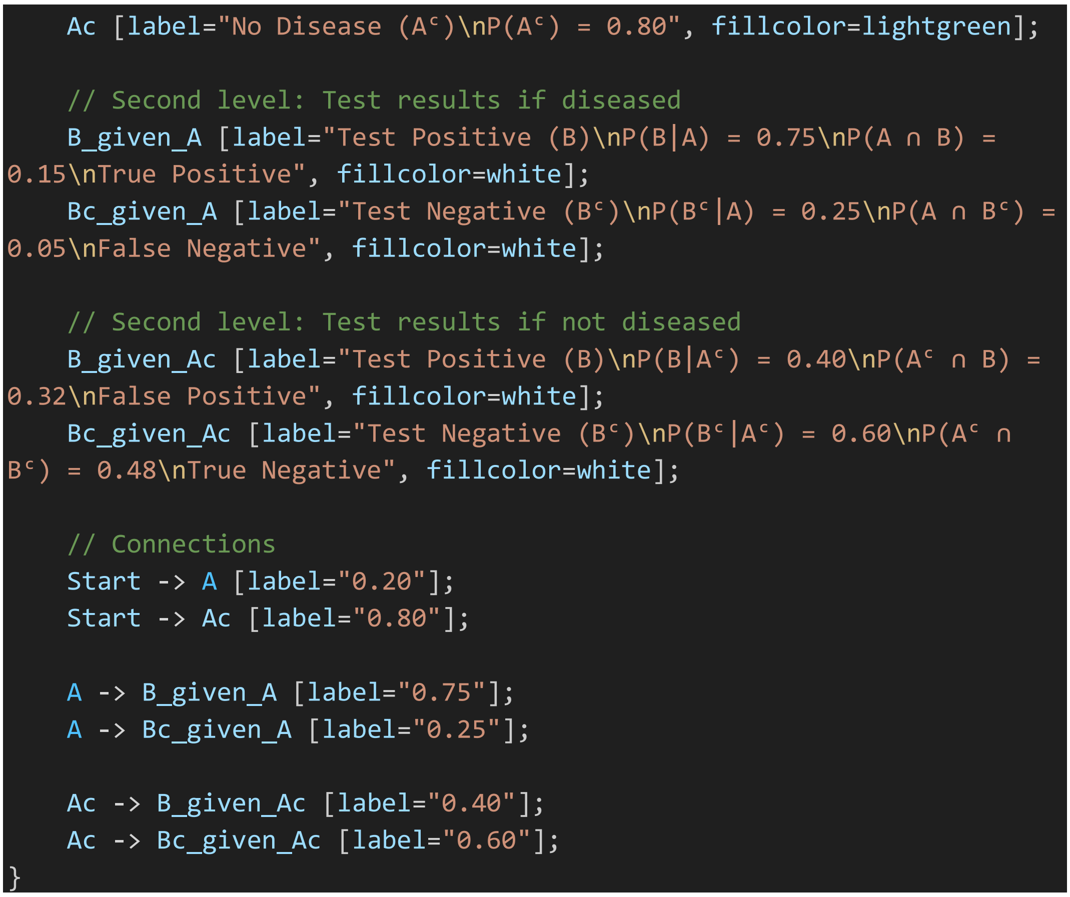
b) Google the terms "**false negative rate**" and "**false positive rate**". (1) Provide a definition using the context, (2) identify them using the event and symbol defined above, and (3) find their values in this question. **[2+2 marks]**
Note: There might be a few conflicting courses. So, please try not to rely on one webpage only.

"A **false negative error**, or false negative, is a test result which wrongly indicates that a condition does not hold." (Wikipedia,2025)

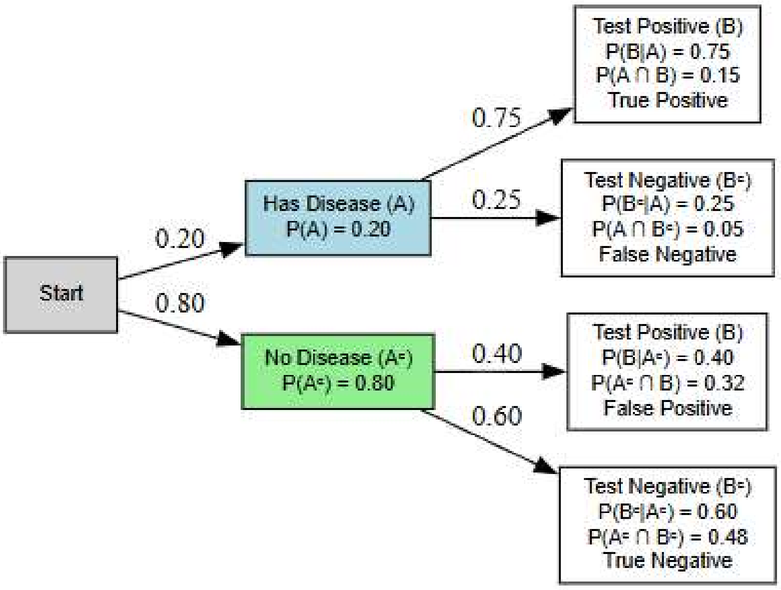
"A **false positive error**, or false positive, is a result that indicates a given condition exists when it objectively does not." (Wikipedia,2025)

c) Draw a tree diagram showing the above situation. Please also compute all 4 joint probabilities (on the far right) and identify clearly the four new terms (true positive rate, false negative rate, true negative rate, and false positive rate) on the tree diagram. **[2+2 marks]**

```
digraph DiseaseTestTree {
    rankdir=LR;
    node [shape=box, style=filled, fontname="Arial", fontsize=10];

    Start [label="Start", fillcolor=lightgray];

    // First level: Disease or No Disease
    A [label="Has Disease (A)\nP(A) = 0.20", fillcolor=lightblue];
```



```
    Ac [label="No Disease (Aᶜ)\nP(Aᶜ) = 0.80", fillcolor=lightgreen];

    // Second level: Test results if diseased
    B_given_A [label="Test Positive (B)\nP(B|A) = 0.75\nP(A ∩ B) =
0.15\nTrue Positive", fillcolor=white];
    Bc_given_A [label="Test Negative (Bᶜ)\nP(Bᶜ|A) = 0.25\nP(A ∩ Bᶜ) =
0.05\nFalse Negative", fillcolor=white];

    // Second level: Test results if not diseased
    B_given_Ac [label="Test Positive (B)\nP(B|Aᶜ) = 0.40\nP(Aᶜ ∩ B) =
0.32\nFalse Positive", fillcolor=white];
    Bc_given_Ac [label="Test Negative (Bᶜ)\nP(Bᶜ|Aᶜ) = 0.60\nP(Aᶜ ∩
Bᶜ) = 0.48\nTrue Negative", fillcolor=white];

    // Connections
    Start -> A [label="0.20"];
    Start -> Ac [label="0.80"];

    A -> B_given_A [label="0.75"];
    A -> Bc_given_A [label="0.25"];

    Ac -> B_given_Ac [label="0.40"];
    Ac -> Bc_given_Ac [label="0.60"];
}
```



d) When a person is tested positive, what is the probability that they have the disease? **[2 marks]**

$$\boldsymbol{P(B|A) = \frac{0.15}{0.32 + 0.15} = 0.319}$$

The probability is 32%

e) When a person is tested negative, what is the probability that they do not have the disease? **[2 marks]**

$$\boldsymbol{P(A^c|B^C) = \frac{P(A^C \cap B^c)}{P(B^c)} \frac{0.48}{0.48 + 0.05} = 0.9057}$$

The probability is 90.6%

f) In a 2x2 two-way table (a two-way table with two rows and two columns that contain frequencies or counts only), the "**odds ratio**" is defined as "the product of the number of true positives and the number of true negatives divided by the product of the number of false negatives and the number of false positives". Use this definition to find the odds ratio in this question. **[2 marks]**

Note: "True positive rate" is a probability, but "number of true positives" is a count, which you can find from the two-way table.

	Test Positive (B)	**Test Negative (B^c)**	**Total**
Has Disease (A)	0.15 (True Positive)	0.05 (False Negative)	0.20
No Disease (A^c)	0.32 (False Positive)	0.48 (True Negative)	0.80
Total	0.47	0.53	1.00

$$\boldsymbol{Odds\ Ratio} = (\boldsymbol{TP} \times \boldsymbol{TN}) / (\boldsymbol{FN} \times \boldsymbol{FP}) = \frac{0.15*0.48}{0.05*0.32} = \frac{0.072}{0.016} = 4.5$$

g) Note that the above odds ratio is greater than 1. Please look up the internet to find the meaning when the odds ratio bigger than 1 (and when the odds ratio is less than

1). What is the implication of "odds ratio greater than 1" here, in terms of the "power" or "ability" of the screening test to identify the disease. **[2 marks]**

"An odds ratio > 1 means the event is more likely in the exposed group; < 1 means it's less likely compared to the non-exposed group."

An odds ratio greater than 1 indicates that the test has a greater ability to correctly identify the disease.

h) What does it mean when the odds ratio is equal to one? Come up with a "hypothesis" or "postulation" about the relationship between "odds ratio being one" and "the independence relation between two categorical variables". Then, create a two-way table that has an odds ratio of 1, and shows that the two binary categorical variables are independent to each other. **[2+2 marks]**

Note: I am sure you can find thousands of websites about this, but I would challenge you to figure this part yourself without the help of internet. Believe in your own ability. :)

An odds ratio of 1 indicates that the odds of an event happening are the same across groups — in other words, there's no association between the two variables. This suggests the variables are independent of each other.
In this case, let's say we're analyzing whether Grade 11 students play **basketball (B)** or not (BC), and whether they play **volleyball (A)** or not (AC).
Based on the following 2x2 table:

	Play Volleyball (A)	Don't Play Volleyball (AC)	**Total**
Play Basketball (B)	40(a)	10	50
Don't Play Basketball (BC)	40	10(d)	50
Total	80	20	100

The **odds ratio** is: a·d/ b·c

$$OR = \frac{40 * 10}{40 * 10}$$

This confirms that playing basketball and playing volleyball are **independent events**. In your opinion, students choose to play these sports **completely independently**, and the data supports that—there is **no relationship** between the two activities.

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
3. Steve Brunton. (2024, December 2). Bayes' Theorem (with Example!) [Video]. YouTube. https://www.youtube.com/watch?v=akClB1J6b28
4. Wikipedia contributors. (2025, May 16). False positives and false negatives. Wikipedia. https://en.wikipedia.org/wiki/False_positives_and_false_negatives
5. Tenny, S., & Hoffman, M. R. (2024). Odds ratio. In *StatPearls*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK431098/