

# Mini Project 1

PSTAT100: Data Science Concepts and Analysis

Instructor: Ali Abuzaid

## STUDENT NAME

- Xinyao Song (6194773)
- STUDENT 2 (NetID 2)
- STUDENT 3 (NetID 3)

## Instructions

- This mini project aims to familiarize you with real-life data sourced from various resources.
- The mini project includes narrative questions. While these questions are primarily based on lecture material and prerequisites, they may also require independent thinking and investigation.
- Collaborate in groups of **2-3** students from the **same section**; individual submissions **will not be accepted**.
- Please use the provided MP 1.qmd file to type your solutions and submit the completed assignment as a PDF file. You can utilize RStudio for this purpose. For guidance, refer to the [Tutorial: Hello, Quarto](#).
- Submit your answers via **Gradescope**.
- Ensure that all R code, mathematical formulas, and workings are presented clearly and appropriately.
- All figures should be numbered, and axes must be labeled.

## Due Date

**Due Date:** Sunday, May 4, 2025, 11:59 PM

💡 Survey (Use the attached 'Survey.xlsx' file)

The data was collected during the first lecture from 85 students in the PSTAT 100 class in Spring 2025. The survey has 6 main sections namely Personal Information, Physical Measurements, Health Habits, Diet and Nutrition, Mental Health and Stress, and Academic Life.

These sections are interconnected - for example, physical measurements may relate to health habits and diet, while academic life may impact mental health and sleep patterns. The goal is to explore the relevant variables that may affect students' performance.

! Question 1

Based on your understanding of the collected data, propose three questions or hypotheses that can be explored or tested using the data.

ANSWERS TO QUESTION 1:

**Hypothesis 1:** Whether the amount of sleep a student gets is associated with their GPA.

$H_0$  : The amount of sleep a student gets does not affect with their GPA ( $\mu_{<4 \text{ hours}} = \mu_{4-6 \text{ hours}} = \mu_{6-8 \text{ hours}} = \mu_{>8 \text{ hours}}$ ).

$H_A$  : The amount of sleep a student gets does not affect with their GPA (at least one sleep group has a different mean GPA).

**Hypothesis 2:** Whether the amount of physical exercise a student does affects their sleep duration.

$H_0$  : There is no association between the amount of exercise and the amount of sleep (Exercise  $\perp$  Sleep).

$H_A$  : There is an association between the amount of exercise and the amount of sleep (Exercise  $\not\perp$  Sleep).

**Hypothesis 3:** Whether a student's stress level influences their likelihood of consuming alcohol.

$H_0$  : There is no association between stress level and alcohol consumption (Stress  $\perp$  Alcohol).

$H_A$  : There is an association between stress level and alcohol consumption (Stress  $\not\perp$  Alcohol).

## ! Question 2

Load the data into R, obtain basic descriptive statistics for the numerical variables, and provide your interpretation of the results.

### ANSWERS TO QUESTION 2:

```
1 library(readxl)
2 library(dplyr)
3
4 survey <- read_excel("Survey.xlsx")
5 numeric_var <- survey %>% select(where(is.numeric))
6 summary(numeric_var)
```

Weight	Heigh	Stress	GPA
Min. :106.0	Min. : 5.000	Min. : 2.000	Min. :1.300
1st Qu.:135.0	1st Qu.: 5.675	1st Qu.: 4.000	1st Qu.:3.300
Median :160.0	Median : 29.550	Median : 5.500	Median :3.500
Mean :159.0	Mean : 57.827	Mean : 5.702	Mean :3.478
3rd Qu.:173.5	3rd Qu.: 67.500	3rd Qu.: 7.000	3rd Qu.:3.800
Max. :390.0	Max. :511.000	Max. :10.000	Max. :4.000
			NA's :3

Among all the variables in the Survey dataset, there are a total of four numerical variables, which include weight, height, stress, and GPA.

- From the summary statistics above, we can see that the weight is likely measured in pounds, with the minimum weight being 106 pounds and the maximum being 390. We can see that there is a significant difference between the third quartile and the maximum weight value, this may suggest that the data is right-skewed or that there are outliers on the higher end.
- On the other hand, the height variable seems a little inconsistent in formatting given the minimum value being 5 and the maximum value being 511. It's possible that most data were recorded in feet-inches, and that entries like 511 were meant to represent "5 feet 11 inches" but were entered without a decimal. There is also a huge spread between the third quartile and the maximum value, which may suggest that there are multiple entry errors that need to be fixed.
- While the stress levels are divided into 10 levels, we can see that they are rather even throughout each quartile. One thing to note is that while there is a maximum stress level of 10, the minimum stress inputted is 2, not 1.
- Finally, for GPA, we can see that nearly 75% of the data lies above a 3.3 GPA. There seems to be a huge gap between the minimum value and the 1st quartile. Since this is a survey, we might need to take a closer look into voluntary bias, as students with lower GPA may not be willing to share their data.

### ! Question 3

- a- Identify any missing values in the **Survey** dataset.
- b- Assess the missing data mechanism in the dataset.
- c- How would you handle the identified missing values? Save the treated dataset as **SurveyMissing**?
- d- Can you address any potential bias for this missing values?

### ANSWERS TO QUESTION 3:

a.

```
colSums(is.na(survey))
```

Gender	Academic	Weight	Heigh	Sleep	Exercise
0	0	0	0	0	0
Water	Alcohol	Meals	Diet	FastFood	Stress
0	2	0	0	0	0
SocialMedia	Studying	GPA	Health		
0	0	3	5		

From the summary above, we see that there are 2 missing values in Alcohol, 3 missing values in GPA, and 5 missing values in Health.

b.

- Alcohol:

#### **! Question 4**

- a- Detect and handle outliers in the **Weight** variable of the dataset.
- i- Justify your outlier detection method (e.g., IQR, Z-score, or visual inspection) and explain your chosen handling strategy (e.g., winsorization, capping, or removal).
- ii -Discuss the potential impact of these outliers on the analysis if left unaddressed.

#### **ANSWERS TO QUESTION 4:**

*Replace this line with your answers*

### **! Question 5**

Investigate the relationship between sleep duration (**Sleep**) and perceived stress levels (**Stress**). Address the following:

- a- Calculate the average stress level for each sleep duration category and comment.
- b- Do certain sleep ranges (e.g., “4–6 hours”) correlate with higher/lower stress?
- c- Use a boxplot to compare stress distributions across sleep categories and comment!

### **ANSWERS TO QUESTION 5:**

*Replace this line with your answers*

### **! Question 6**

Create a visualization to show the distribution of Weight, comment!

### **ANSWERS TO QUESTION 6:**

*Replace this line with your answers*

### ! Question 7

a. Define a new variable  $BMI = \left( \frac{\text{Weight (lb)}}{\text{Height (in)}^2} \right) \times 703$ , and classify students into BMI categories based on Centers for Disease Control and Prevention (CDC) guidelines:

- $BMI < 18.5 \rightarrow \text{"Underweight"}$
- $18.5 \leq BMI < 25 \rightarrow \text{"Normal"}$
- $25 \leq BMI < 30 \rightarrow \text{"Overweight"}$
- $BMI \geq 30 \rightarrow \text{"Obese"}$

b- Visualize and describe the distribution of students' BMI categories using a bar chart.

c- Explore the correlations between **BMI**, **GPA**, and **Stress**. Use visualizations and correlation statistics to summarize the relationships.

### ANSWERS TO QUESTION 7:

*Replace this line with your answers*