Mini Project 1

PSTAT100: Data Science Concepts and Analysis

Kai Barker

STUDENT NAMES

- Kai Barker (kaibarker)
- Luke Drushell (drushell)
- Rafael Romero (romeroflores)

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

```r
library(readxl)
library(ggplot2)
library(dplyr)
library(writexl)
library(tidyverse)
```

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:

From the collected data, there are many hypothesis that could be tested. Three that stood out were, is there a correlation between the hours of sleep that students get, and their level of stress? Is there a relationship between the number of hours a student spends studying and their GPA? Is there a relationship between the students stress level, and the amount of times they get exercise each week?

❗ 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:

```r
survey_data <- read_excel("Survey.xlsx")
survey_data
```

```
# A tibble: 84 x 16
   Gender Academic  Weight Heigh Sleep        Exercise Water Alcohol Meals Diet 
   <chr>  <chr>      <dbl> <dbl> <chr>        <chr>    <chr> <chr>   <chr> <chr>
 1 Male   Junior       165  58   6-8 hours    5-7 days More~ Yes     3 me~ Omni~
 2 Male   Junior       185 511   More than 8~ 3-4 days 4-6 ~ Yes     3 me~ Omni~
 3 Male   Junior       185  72   4-6 hours    3-4 days 2-4 ~ Yes     2 me~ Omni~
 4 Female Sophomore    115   5   6-8 hours    5-7 days 4-6 ~ Yes     2 me~ Omni~
 5 Female Senior       185  57   More than 8~ 1-2 days 4-6 ~ Yes     3 me~ Omni~
 6 Female Junior       106  62   More than 8~ 1-2 days 2-4 ~ No      2 me~ I ea~
 7 Male   Junior       145   5.7 More than 8~ 5-7 days More~ Yes     More~ Omni~
 8 Male   Senior       165   6   More than 8~ 1-2 days More~ Yes     3 me~ None 
 9 Female Junior       135  57   6-8 hours    0 days   Less~ Yes     2 me~ Omni~
10 Male   Senior       390   5   Less than 4~ 0 days   Less~ Yes     1 me~ Omni~
# i 74 more rows
# i 6 more variables: FastFood <chr>, Stress <dbl>, SocialMedia <chr>,
#   Studying <chr>, GPA <dbl>, Health <chr>
```

```r
#rename height since its spelled wrong
survey_data$Height <- survey_data$Heigh

head(survey_data)
```

```
# A tibble: 6 x 17
  Gender Academic Weight Heigh Sleep Exercise Water Alcohol Meals Diet  FastFood
  <chr>  <chr>     <dbl> <dbl> <chr> <chr>    <chr> <chr>   <chr> <chr> <chr>   
1 Male   Junior      165    58 6-8 ~ 5-7 days More~ Yes     3 me~ Omni~ 1-2 tim~
2 Male   Junior      185   511 More~ 3-4 days 4-6 ~ Yes     3 me~ Omni~ 1-2 tim~
3 Male   Junior      185    72 4-6 ~ 3-4 days 2-4 ~ Yes     2 me~ Omni~ 1-2 tim~
4 Female Sophomo~    115     5 6-8 ~ 5-7 days 4-6 ~ Yes     2 me~ Omni~ 0 times 
5 Female Senior      185    57 More~ 1-2 days 4-6 ~ Yes     3 me~ Omni~ 0 times 
6 Female Junior      106    62 More~ 1-2 days 2-4 ~ No      2 me~ I ea~ 0 times 
# i 6 more variables: Stress <dbl>, SocialMedia <chr>, Studying <chr>,
#   GPA <dbl>, Health <chr>, Height <dbl>
```

```r
summary(survey_data$Weight)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  106.0   135.0   160.0   159.0   173.5   390.0 
```

```r
summary(survey_data$Height)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  5.000   5.675  29.550  57.827  67.500 511.000 
```

```r
summary(survey_data$GPA)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  1.300   3.300   3.500   3.478   3.800   4.000       3 
```

We can see from the dataset that there is only three numerical variables. While many of the variables include numerical values, and value that can be categorized, such as the number of meals per day, does not count as a numerical variable. Further, students stress level was determined to not be a numerical variable as it was not continuous, and was limited to only ten options. Thus our numerical variables were Weight, Height, and GPA.

We can observe that the Weight variable had a minimum value of 106 and a maximum value of 309, with a mean value of 159. A challenge with this data is that not all responses may be in the same units, or honest. With that in mind, our mean could be more affected by the maximum and minimum values.

The height variable is a little bit more difficult to interpret, as the there were inconsistencies with the collection of the data. Some respondents input their height in centimeters, some in inches and others in a combination of feet' inches" or feet.inches. This lead to the data being difficult to interpret. We can see that the minimum height is 5, while the maximum height is 511, with a mean height of 57.827. Realistically, someone is not 5 inches tall, nor 511 inches tall. We can assume that the 511 is five foot eleven, while the minimum is five foot even.

From the summary of students GPA, we can observe that there was a minimum GPA of 1.3, and a maximum GPA of 4.0, with a mean of 3.478 and 3 missing values. The missing values are likely due to students who did not want to report their GPA. Further, some students may not have been honest in their responses, skewing the data. However, the mean GPA value seems plausible in a real world context.

❗ 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)

From the dataset, we can see that there are some missing values in the Alcohol, GPA, Health columns.

b) The missing data could be attributed to multiple factors. For the alcohol variable, some students may not have wanted to answer, as alcohol consumption can be a personal subject. They might also feel uncomfortable answering such a question on a school related survey. For the GPA variable, some students may not want to disclose their GPA because they think it is too low, and there is social pressure to be within a certain GPA range. Similarly, the health variable, related to how many days per week a student exercises may be too personal for someone to feel comfortable disclosing to their peers. Thus, while there many reasons for why these data are missing, it is most likely due to students digression. That would make these data MNAR, or missing not at random, as the missingness is related to the sensitive nature of the variable itself.

c)

To handle the missing values in the dataset, we can use drop_na to get rid of these rows that are missing data. While there are multiple methods to deal with the missing data, we have discussed in question one, our interest in the data correlating to some of these variables, thus we should drop rows that are missing these data.

```r
SurveyMissing <- survey_data %>%
  drop_na(Alcohol, GPA, Health)

head(SurveyMissing)
```

```
# A tibble: 6 x 17
  Gender Academic Weight Heigh Sleep Exercise Water Alcohol Meals Diet  FastFood
  <chr>  <chr>     <dbl> <dbl> <chr> <chr>    <chr> <chr>   <chr> <chr> <chr>
1 Male   Junior      165  58   6-8 ~ 5-7 days More~ Yes     3 me~ Omni~ 1-2 tim~
2 Male   Junior      185 511   More~ 3-4 days 4-6 ~ Yes     3 me~ Omni~ 1-2 tim~
3 Male   Junior      185  72   4-6 ~ 3-4 days 2-4 ~ Yes     2 me~ Omni~ 1-2 tim~
4 Female Senior      185  57   More~ 1-2 days 4-6 ~ Yes     3 me~ Omni~ 0 times
5 Female Junior      106  62   More~ 1-2 days 2-4 ~ No      2 me~ I ea~ 0 times
6 Male   Junior      145   5.7 More~ 5-7 days More~ Yes     More~ Omni~ 0 times
# i 6 more variables: Stress <dbl>, SocialMedia <chr>, Studying <chr>,
#   GPA <dbl>, Health <chr>, Height <dbl>
```

d) The bias for the missingness is likely due to the sensitive nature of the responses to those variables. See part **b)** of this question for further explanation.

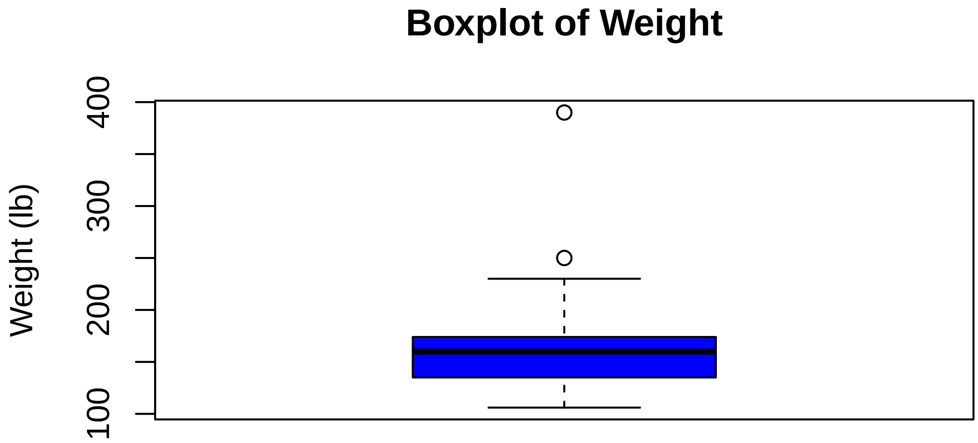
! Question 4

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:

Detect Outliers: With a boxplot we can locate outliers by analyzing the whiskers of the plot and noticing the points that lie outside of them.

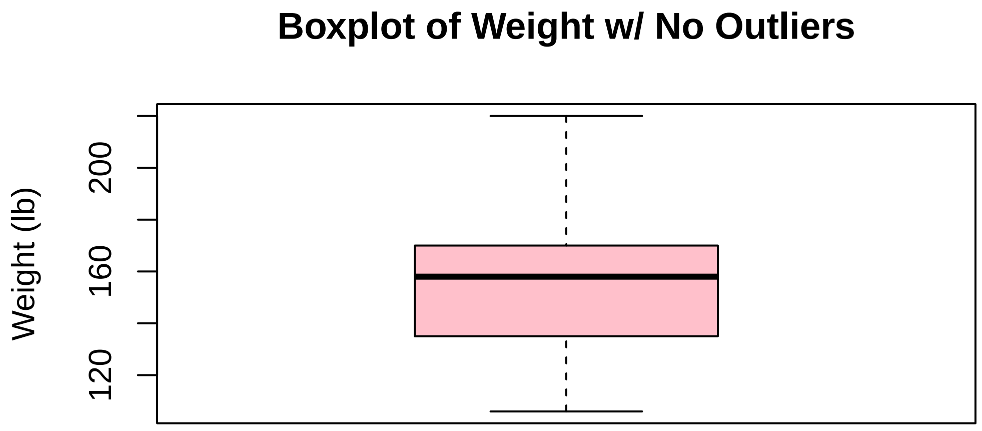
```r
boxplot(survey_data$Weight, main="Boxplot of Weight", col ="blue",ylab= "Weight (lb)")
```



We notice two outliers that fall outside the whiskers. Above the 230 mark with `visual inspection`.

Handling Strategy: We can simply remove the outliers over the weight of 230 since they are extreme and do not accurately represent the population.

```r
upper_limit<- 230
survey_data_no_outliers <- survey_data[survey_data$Weight < upper_limit, ]
boxplot(survey_data_no_outliers$Weight,main="Boxplot of Weight w/ No Outliers", col ="pink",ylab= "Weig
```



Consequences of Outliers: Outliers can have a great impact on the results of our analysis as they can `affect our mean, variance, and standard deviation` by providing an inaccurate measure due to an inflation or deflation of our statistics. Also, they can cause our results to be `biased` or lead to `inaccurate p-values` if they violate statistical test assumptions such as normality.

❗ 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:

To investigate the relationship between sleep duration and perceived stress levels, with must first calculate **the average stress for each sleep duration category**.

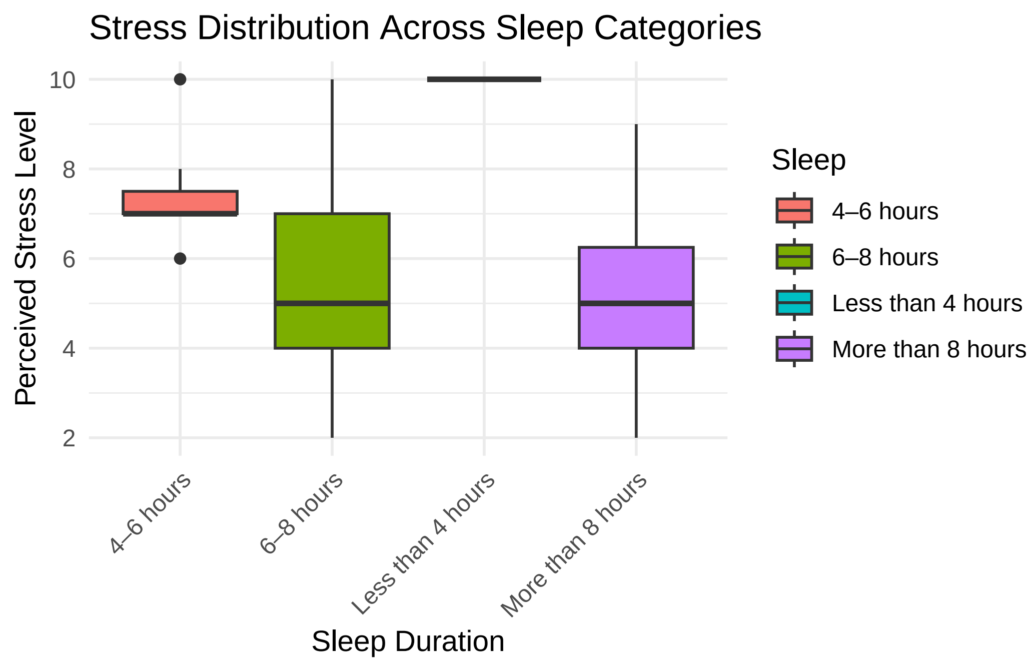
```r
survey_data %>% select(Sleep, Stress) %>%
  group_by(Sleep) %>%
  summarise("Average Stress Level by Group" = mean(Stress))
```

```
# A tibble: 4 x 2
  Sleep              `Average Stress Level by Group`
  <chr>                                        <dbl>
1 4-6 hours                                     7.43
2 6-8 hours                                     5.64
3 Less than 4 hours                            10
4 More than 8 hours                             5.05
```

Comments: We notice that on average, those who sleep less than 4 hours have the highest stress level of 10. We can also notice that those who sleep more than 8 hours have the lowest average stress level.

We can also look at a **boxplot** to visualize how the stress level changes based on sleep.

```r
survey_data %>% ggplot(aes(y=Stress,x=Sleep,fill=Sleep)) +
  geom_boxplot() +
    labs(title = "Stress Distribution Across Sleep Categories",
       x = "Sleep Duration",
       y = "Perceived Stress Level") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for readability
```

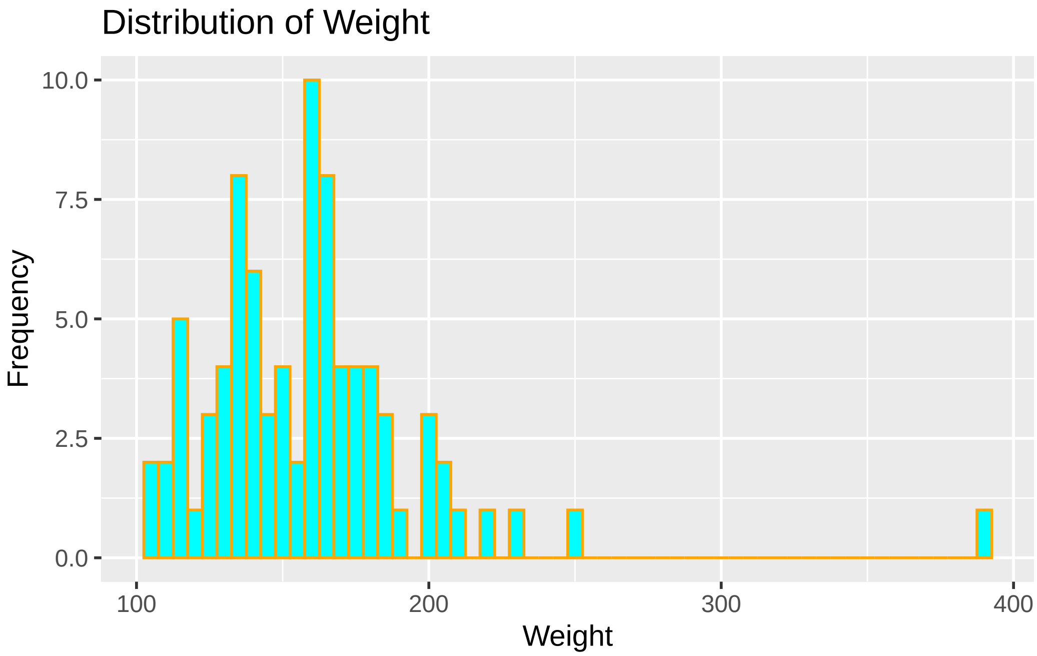


Conclusion: We can confidently state that there is a significant correlation between an increase in sleep duration and and decrease in perceived stress levels. 4-6 hours of sleep and less than 4 hours of sleep have the highest amount of perceived stress levels, which supports this claim. It is important to note, though, that 6-8 hours of sleep and more than 8 hours of sleep have notably long whiskers, which indicates that although they are getting good sleep, some may still be experiencing high stress levels. Regardless, their average stress level is still significantly lower than these stress levels, which again supports the correlation between high sleep duration and lower stress levels.

Question 6

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

```r
survey_data_revisions <- read_excel("Survey.xlsx")

ggplot(survey_data_revisions, aes(x = Weight)) +
  geom_histogram(binwidth = 5, fill = "cyan", color = "orange") +
  labs(title = "Distribution of Weight",
       x = "Weight",
       y = "Frequency")
```



While there is evidently one outlier, what is much more interesting is the dip found around 150, it seems a lot of people hang just below and just above that range, but very infrequently actually in that area.

! Question 7

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

- $\text{BMI} < 18.5$ → `"Underweight"`
- $18.5 \quad \text{BMI} < 25$ → `"Normal"`
- $25 \quad \text{BMI} < 30$ → `"Overweight"`
- $\text{BMI} \quad 30$ → `"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.

Starting by standardizing the height to inches:

```r
# Adjust the height column
survey_data_revisions$Heigh <- sapply(survey_data_revisions$Heigh, function(h) {
  if (h < 3) {
    # they typed in meters
    return(h * 39.3701)
  } else if (h < 30) {
    # ft.inches format (5.11 means 5 feet 11 inches)

    parts <- strsplit(as.character(h), "\\.")[[1]]
    feet <- as.numeric(parts[1])
    inches <- ifelse(length(parts) > 1, as.numeric(parts[2]), 0)

    # adjust if they had the audacity to put partial inches
    if (inches >= 12) { inches <- inches/10 }


    return(feet * 12 + inches)
  } else if (h >= 30 && h < 100) {
    # already in inches
    return(h)
  } else if (h >= 100 && h < 350) {
    # they typed in centimeters
    return(h * 0.393701)
  } else {
    # ftin format (511 means 5 feet 11 inches)

    h_str <- as.character(h)
    feet <- as.numeric(substr(h_str, 1, 1))
    inches <- as.numeric(substr(h_str, 2, nchar(h_str)))

    # adjust if they had the audacity to put partial inches
    if (inches >= 12) { inches <- inches/10 }

    return(feet * 12 + inches)
  }
})
```

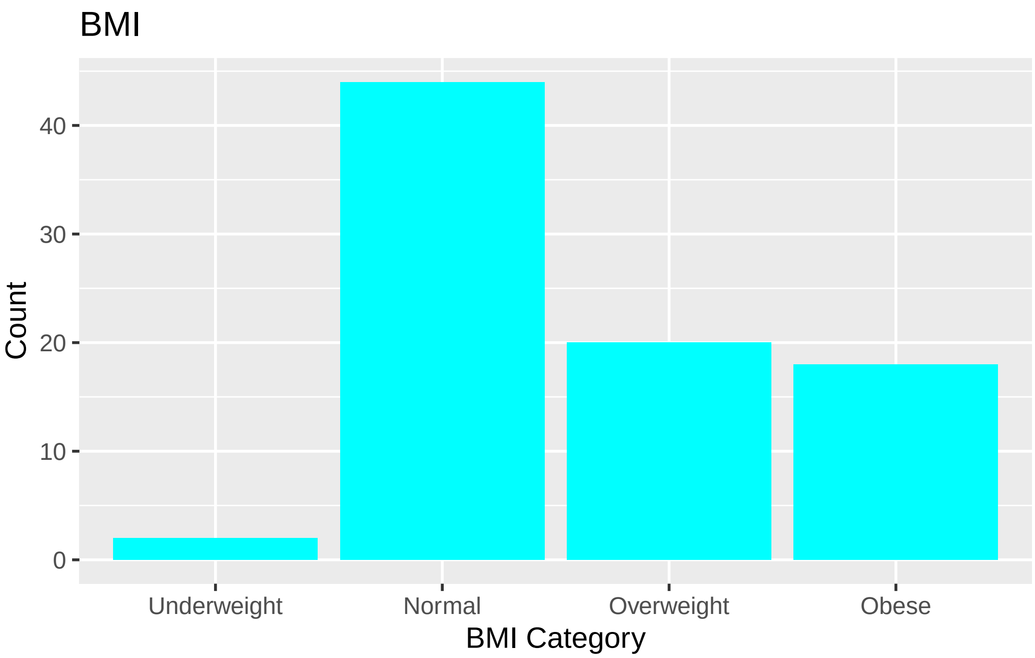
```

write_xlsx(survey_data_revisions, "Survey_Height_Revised.xlsx")
```

a)

```
BMI <- survey_data_revisions$Weight / survey_data_revisions$Heigh^2 * 704
```

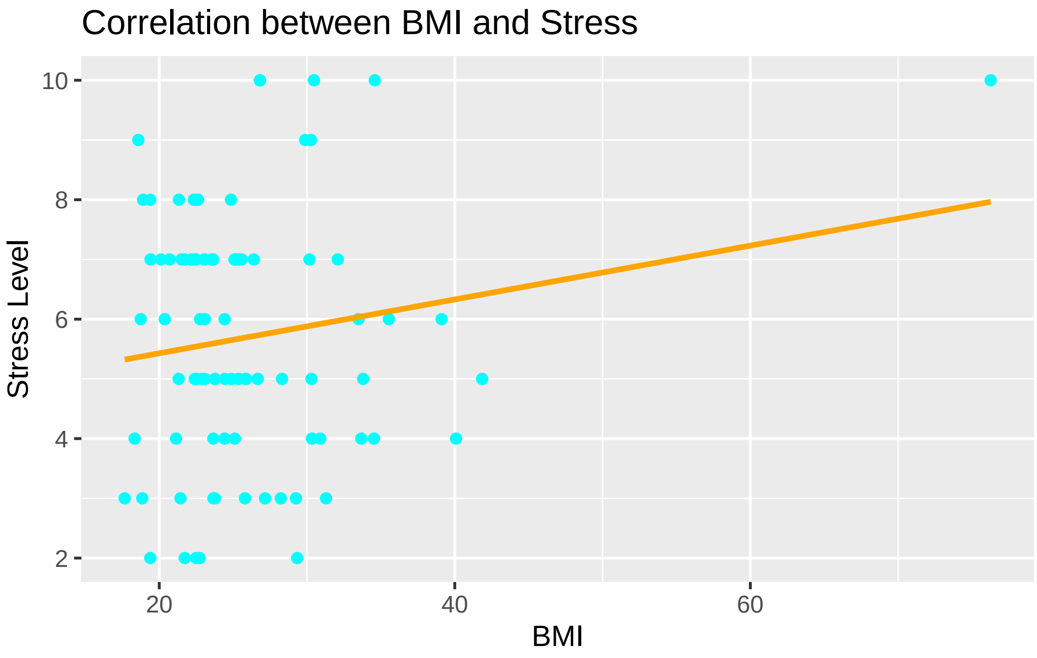
b)

```
bmi_category <- cut(
  BMI,
  breaks = c(-Inf, 18.5, 25, 30, Inf),
  labels = c("Underweight", "Normal", "Overweight", "Obese")
)

bmi_dataframe <- data.frame(Category = bmi_category)

ggplot(bmi_dataframe, aes(x = Category)) +
  geom_bar(fill = "cyan") +
  labs(title = "BMI", x = "BMI Category", y = "Count")
```

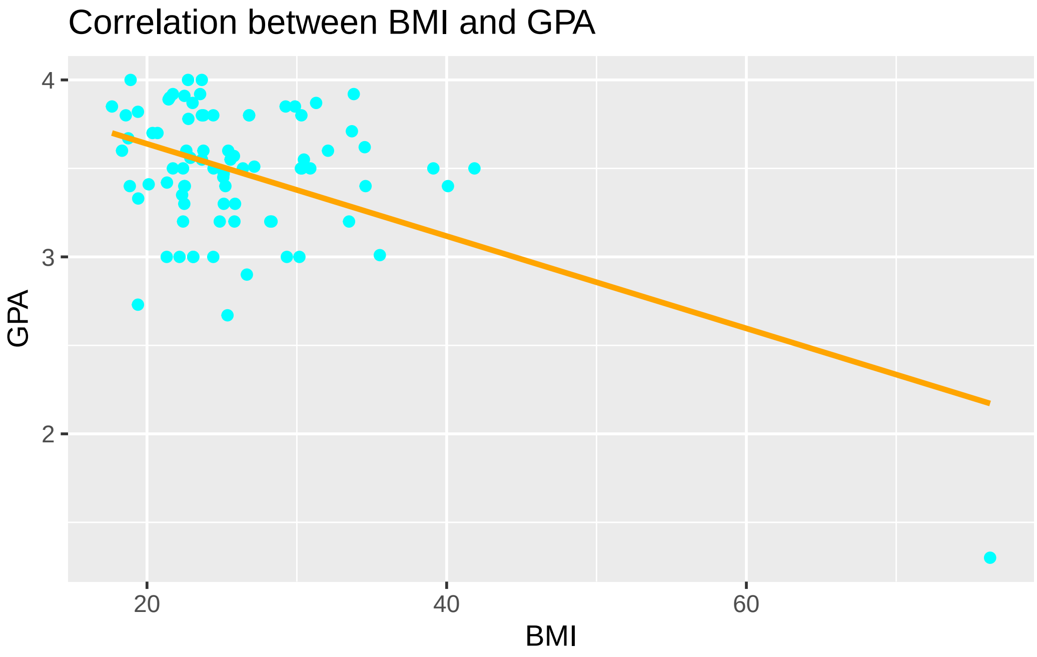


c) Explore the correlations between **BMI**, **GPA**, and **Stress**. Use visualizations and correlation statistics to summarize the relationships.

```
Stress <- survey_data_revisions$Stress
GPA <- survey_data_revisions$GPA

ggplot(survey_data_revisions, aes(x = BMI, y = Stress)) +
  geom_point(color = "cyan") +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  labs(title = "Correlation between BMI and Stress",
       x = "BMI",
       y = "Stress Level")
```



```r
ggplot(survey_data_revisions, aes(x = BMI, y = GPA)) +
  geom_point(color = "cyan") +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  labs(title = "Correlation between BMI and GPA",
       x = "BMI",
       y = "GPA")
```


Correlation between BMI and GPA
GPA
BMI