Project 2: Exploratory Data Analysis

MATH346: Mathematical & Statistical Software

28th November 2025

Dataset:

Exercise (Gym Members Exercise Tracking)

I. Introduction

The dataset we have chosen for this project is the “Gym Members Exercising Tracking” dataset. It has 973 entries/observations across 15 features/variables, which are:

- Age
- Gender
- Weight (kg)
- Height (m)
- Maximum Beats Per Minute (BPM)
- Average BPM
- Resting BPM
- Session Duration (hours)
- Calories Burned
- Workout Type
- Fat Percentage (%)
- Water Intake (litres)
- Workout Frequency (times per week)
- Experience Level
- Body Mass Index (BMI)

All these variables are either related to physical health or exercise, and since those are two very closely related topics that influence one another in real life, we expect to find relationships between the features in our data. We will conduct exploratory data analysis in Rscript using the techniques covered in the MATH346 course, including but not limited to plotting various graphs, analyzing them, and computing correlation between variables. Our methodology and analysis are documented in detail throughout this report, as well as our key takeaways in the conclusion, and an appendix including our code.

II. Methodology

To clean our data, we decided to spot outliers, missing or unusual values to handle them accordingly. We will find such values using boxplots, histograms, and the built-in R missing value function (is.na). Each variable will be assessed separately but with reasonable judgement.

2.1 Continuous Variables

For continuous variables, like Age, Weight, Height, etc., we will plot both boxplots and histograms to illustrate the spread of our data using the functions from the ggplot2 library geom_boxplot() and geom_histogram(). Histograms will help us find skewness, clustering, and gaps in the data, whereas boxplots will help us find outliers using the IQR rule. The graphs will be zoomed in on the right and left ends to find outliers using the coord_cartesian() function from the ggplot2 library. Moreover, when unusual patterns are found, we will zoom in on those areas too. Once we spot outliers and unusual values, we will judge based on the context of the variable and whether they are impossible values, like negatives ones, or extreme but plausible values, and they will be removed from the dataset when necessary.

2.2 Categorical Variables

For Categorical variables, like Gender, Workout Type, etc., we will be plotting bar charts to visualize the frequency of categories in our variables using the ggplot2 library function geom_bar(). This visualization will show us categories that are rare or unexpected, categories that are a result of a typo, and categories that have a very low frequency, which could be due to values that don't fall into any valid category. When incorrect or unusual entries are found they are either overwritten or removed depending on the context of the variable and its frequency.

2.3 Missing Values

To find missing values, we will use the is.na(variable) and count() functions. After finding them we will evaluate whether the missingness of the values is random or systematic. Based on this we will handle the missing values by imputing median values of the variable if its systematic or remove the entire entry if it is random.

Overall, these methods for cleaning our data will help our exploratory data analysis be consistent for the tasks required in Parts A – F.

III. Analysis

3.A Variation within a single variable

The first continuous variable we chose was BMI, and we found that the BMI of the people in the dataset was mostly right skewed (figure 1), with the mode being around 24. There aren't many people with high BMI, but that is to be expected as such large BMI values (40 and up) are very abnormal, especially since data set covers people that work out at least once a week.

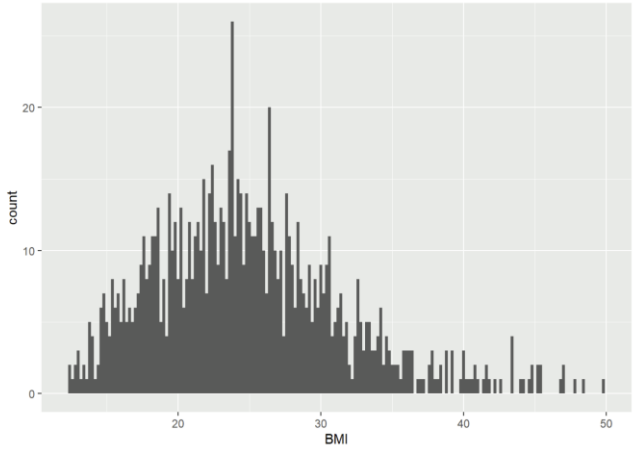


Figure 1

The second continuous variable we chose was session duration, where we found very clear clustering of the data (Figure 2), with the largest cluster being in the center of the plot, ranging from 1-1.5 hours, along with two smaller clusters from ~0.5-0.9 and ~1.5-2.0, respectively. This shows that a third of the people do intense sessions, possibly body builders, a third a moderate amount, and a third do light exercises.

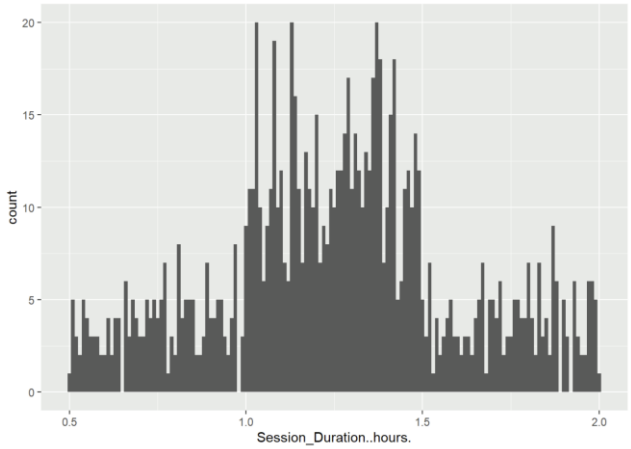


Figure 2

The categorical value we decided to analyze is the workout type, where the bar chart (Figure 3) shows the distribution of workout types across the data set. The counts are relatively close, ranging from ~230-260, with Strength (258) and Cardio (255) being the most common, and

HIIT being the least common. This makes sense as HIIT is a very intense workout and would be done less frequently and by less people, on the other hand, strength and cardio are more open to all people of all experiences and can be done at a higher frequency.

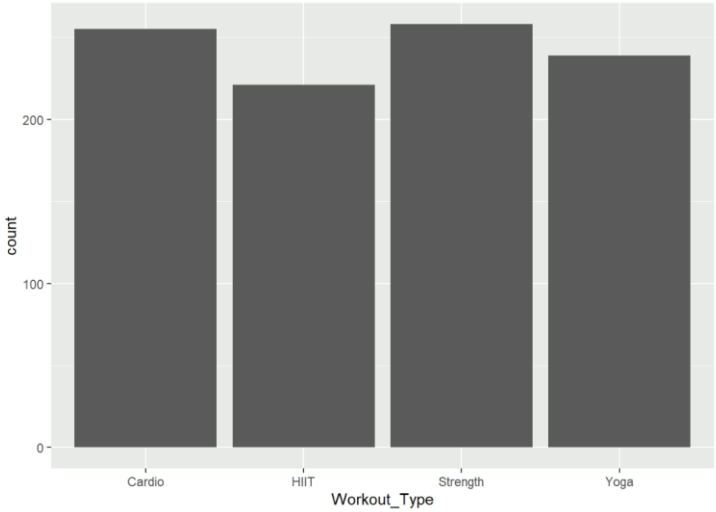


Figure 3

3.B, C Unusual Values/Outliers

While the BMI distribution is mostly dense within the normal, human range for BMI, there are some extreme values on either side that are biologically rare and unlikely (Figure 4). Based on medical research, BMI < 15 and BMI > 40 are considered extreme. The box plot, however, only flags the high extreme BMI values as outliers, since the low extreme BMI values aren't mathematically far off from the rest of the data and fit within 1.5 times the IQR below Q1. However, these low BMI values are medically absurd and so are considered outliers as well.

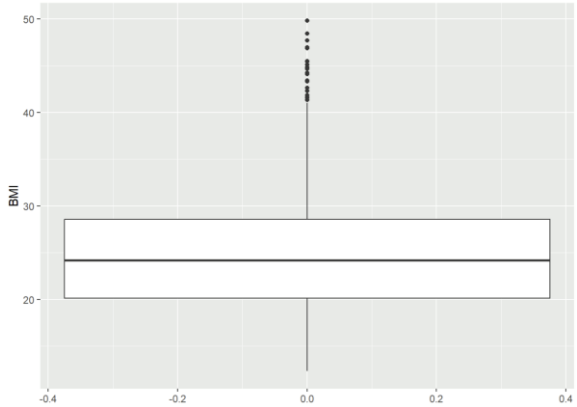
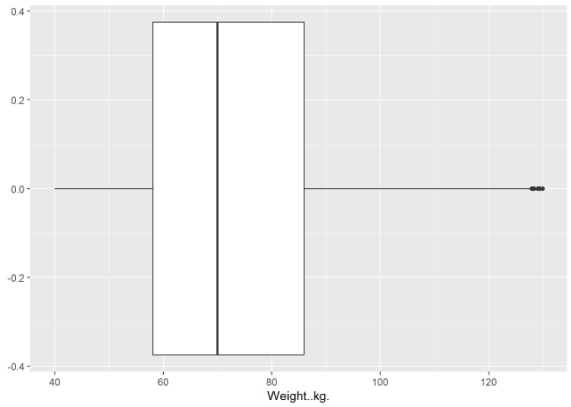


Figure 4

A few variables contain outliers that we decided to keep, as they are plausible and possible, and are most likely correct data.

First, the weight box plot (Figure 5) detected some outliers according to the IQR standard, however the differences between the outliers and non-outliers beside them are not drastically big, and they don't seem to be biologically unusual/impossible, so removing them would not benefit the data set much. There aren't impossible values such as negative weights either.



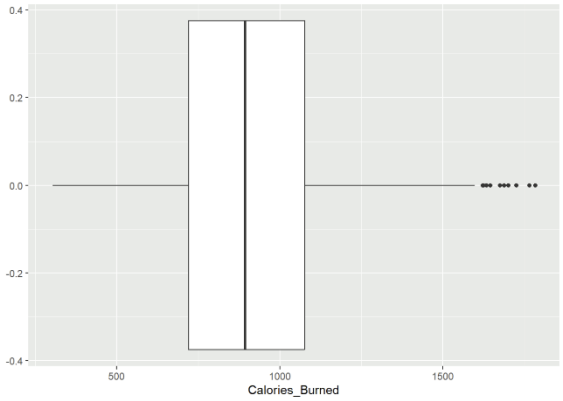


Figure 5

Figure 6

Second, the boxplot (Figure 6) and histogram for the Calories Burned variable shows that there are some outliers with values between 1600 and 1800, however, these values are medically plausible so removing them will only lead to less data and inaccurate population representation. Moreover, there are no negative and no unusual values.

The rest of the variables did not contain any outliers, and all data seemed plausible. This includes: Age (18-59), height (1.5-2 meters), average, max, and resting BPM, session duration (0.45-2.05 hours), Fat percentage (10-35%). Water intake also had no outliers and had a range of 1.5-3.75 Liters, which we found plausible due to the people exercising, some up to 2 hours. The categorical variables (experience level, Gender, workout type) were analyzed using bar charts, which also confirmed the lack of typos or improper categories.

There were also no missing values in any of the variables, and no typos in any of the categorical variables.

3.D Covariation Between Categorical and Continuous Variables

Figure 5 shows how more calories are burned (higher median) in higher intensity workouts like HIIT, Strength and Cardio compared to Yoga. The boxes are nearly symmetric, with Yoga having a slight right-skew. The range of Yoga is the largest and this is likely because there are many types of yoga, some of which are more fast-paced and cause higher calorie burn. The outliers above Cardio, Strength, and HIIT suggest that very extreme workouts can be done through these types, leading to even more calorie burn. Across all 4 categories, there's near-consistent variability, with HIIT having the largest and Strength having the lowest.

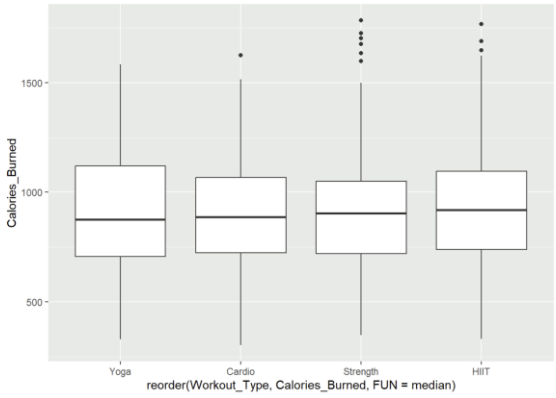
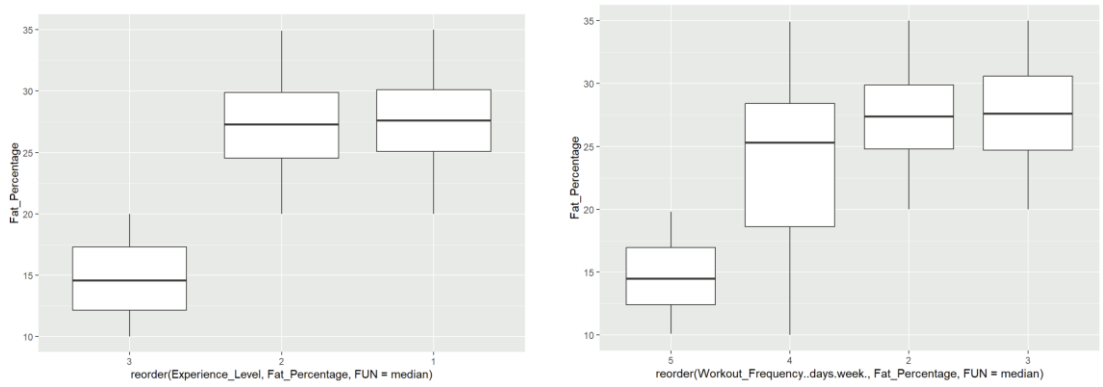


Figure 7

Both Figures 6 and 7 show a similar result. The median fat percentage decreases with higher experience level and weekly workout frequency. This is an expected result, as people who are more experienced in fitness and workout more frequently are likely more muscular and carry less fat than average. In figure 6, the experience levels are all symmetric with no indications of skewness. Levels 2 and 3 have a higher variability than level 1. In figure 7, frequencies 5,2 and 3 show similar variability and little to no skew, while frequency 4 has very large variability and a strong left skew. The variability at frequency 4 shows that both beginners and advanced people work out 4 times a week, as it includes people who have a high percentage of fat despite very frequent exercising.



Figures 8 & 9

Figure 8 shows that water intake generally increases with workout frequency. The variability is high in frequencies 2,3 and 4 with a slight right-skew across the 3 boxes. The box for frequency 5 shows very low variability, and an extreme left-skew. The median seems to overlap almost completely with Q3, indicating that the data of water intake above the median are all near-identical. The skew is logical as people who work out more frequently drink more water on average, although the large range hints to weakness in the covariation.

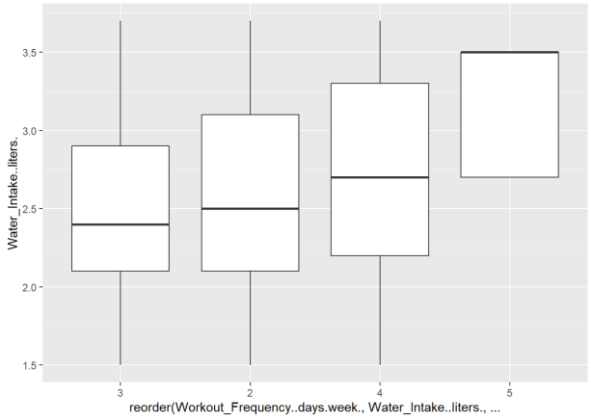


Figure 10

3.E Covariation Between Two Continuous Variables

In both Figures 9 and 10, there are many data points overlapping since in each, the two variables appear to be positively linearly correlated, leading this scatterplot to be overplotted. Below we have used alpha = 0.2 to fix this issue and increase visibility.

Figure 9 shows that Calories burned and session duration are clearly correlated in a strong, positive, linear relationship. This makes sense logically as the longer a workout is, the more the gym members' bodies will burn calories to produce energy. Our conclusion from the plot is supported by the numerical correlation being 0.908. This is an extremely strong positive correlation.

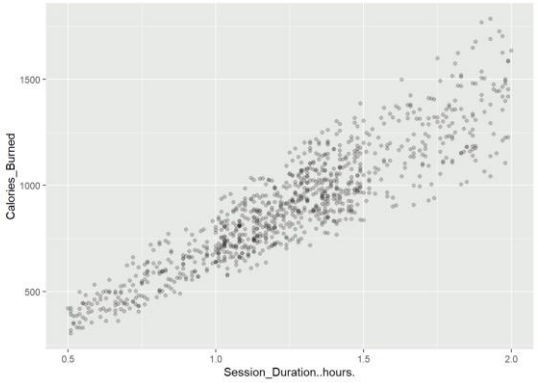


Figure 11

Figure 10 also shows that Weight and BMI also seem positively correlated, but less strongly than with calories burned and session duration, as this plot's relationship appears to have data points in a looser linear relationship. The positive covariance is expected, as weight is a big factor in determining BMI. But since height also has an effect on BMI, the relationship isn't very strong. Our conclusion from the plot is supported by the numerical correlation being 0.808. This is a strong positive correlation, but not as strong as the correlation between calories burned and session duration.

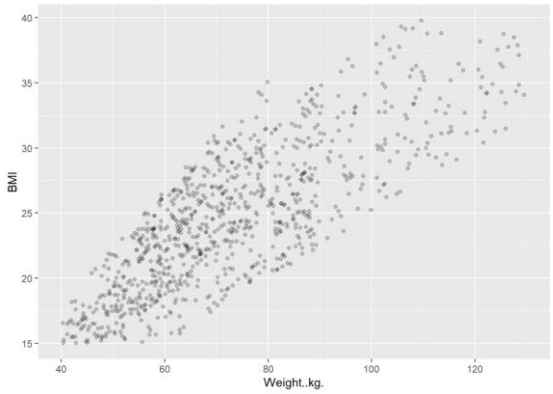


Figure 12

3.F Covariation Between Two Categorical Variables

Since the data set has 2 categorical variables that we have not yet compared, the heatmap (Figure 11) will represent the relationship between Gender and Workout Type. The darker shades represent low frequencies, and the lighter shades represent higher frequencies.

The heatmap doesn't show a strong association between Gender and Workout Type. Participation in the Workout Types are evenly and similarly distributed across the two genders with a slight difference, most probably due to the fact that the data set has more data about males than females. Although the shades for males are lighter, meaning a higher frequency of participation, the differences between choices comparing female and male are small (the range of counts are from 110 to 135), which means that no matter the gender, the preferences are the same. In conclusion, the Gender of the gym member does not influence their Workout Type preference.

Figure 13

IV. Conclusion

Through performing exploratory data analysis, we were able to find relationships and correlations between certain variables. For instance, we uncovered and demonstrated positive correlations between the members' calories burned and their session durations, as well as their weights and BMIs. We also found non-numerically quantifiable relationships, such as members' calories burned being higher in more intense workout types like HIIT.

One key takeaway was realizing that a point numerically being an outlier doesn't mean it's abnormal or unusual. For data with wide ranges like the calories burned, box plots could indicate the presence of outliers, despite the data points being realistic and biologically plausible. The opposite happens as well, with medically rare BMIs not being flagged as numerical outliers.

In conclusion, we were able to confirm the existence of correlations between certain features, understand when different plots and techniques are appropriate, and realize the importance of topic context when studying a dataset.

V. Appendix

Overview

Exploratory data analysis performed on an Exercise data set. Data consists of 973 observations with 15 features each.

```{r}
data = read.csv("C:/Program Files/RStudio/gym_members_exercise_tracking.csv")
head(data)
```

Part A – Variation Within Single Variables

Choose two continuous and one categorical variable. Create histograms, bar charts, and describe shapes, gaps, peaks, clusters.

```{r}
library(ggplot2)
library(dplyr)
ggplot(data) + geom_bar(mapping=aes(x=Workout_Type))
# [Categorical] This bar chart shows the distribution of workout types across the data set. The counts are relatively close, ranging from ~230-260, with Strength (258) and Cardio (255) being the most common.

data %>%
  count(Workout_Type)
ggplot(data) + geom_histogram(mapping=aes(x=BMI), binwidth = 0.2)
# [Continuous] This histogram shows that the distribution of BMI is right-skewed, with one large cluster that has a peak at ~24. There are a few gaps around the higher end of the BMI axis, but that is to be expected as such large BMI values are very abnormal– especially since this data set covers people that work out at least once a week.

ggplot(data) + geom_histogram(mapping=aes(x=Session_Duration..hours.),binwidth = 0.01)
# [Continuous] This histogram shows that the distribution of session duration in hours has very clear clustering. There is one large cluster with high peaks in the
```

center of the plot, ranging from 1-1.5 hours, along with two smaller clusters from ~0.5-0.9 and ~1.5-2.0, respectively.

```

## Part B – Detecting Unusual Values / Outliers

### *Plot with zoomed views. Identify extreme or impossible values. Clean using ifelse() and re-plot.*

```{r}

# BMI Variable - Outliers Detected

ggplot(data) + geom_histogram(mapping=aes(x=BMI), binwidth = 0.2)

# While the distribution is mostly dense within the normal, human range for BMI, there are some extreme values on either side that are biologically rare and unlikely. Based on medical research, BMI < 15 and BMI > 40 are considered extreme.

ggplot(data) + geom_histogram(mapping=aes(x=BMI), binwidth = 0.2) + coord_cartesian(xlim=c(11,16))

ggplot(data) + geom_histogram(mapping=aes(x=BMI), binwidth = 0.2) + coord_cartesian(xlim=c(40,50))

# Zoomed-in histogram plots of our extreme areas.

ggplot(data) + geom_boxplot(aes(y=BMI))

# The box plot only flags the high-extreme BMI values as outliers, since the low-extreme BMI values aren't mathematically far off from the rest of the data and fit within 1.5 times the IQR below Q1. However, these low BMI values are medically absurd, and so we will consider them outliers as well.

data_clean = mutate(data, BMI=ifelse((BMI<15|BMI>40),NA,BMI))

ggplot(data_clean) + geom_histogram(mapping=aes(x=BMI), binwidth = 0.2)

ggplot(data_clean) + geom_boxplot(aes(y=BMI))

# Data has been cleaned to replace abnormal BMI values by NA. Graphs were then re-plotted to show the effects of cleaning in reducing the skew slightly, improving the boxplot to recalculate IQR and have much less outliers, with the only new outlier (39.77) being more biologically plausible.

# Age variable

```
ggplot(data_clean) + geom_boxplot(aes(Age)) # boxplot for the Age variable

ggplot(data_clean) + geom_histogram(aes(Age), binwidth = 1)

ggplot(data_clean) + geom_histogram(aes(Age), binwidth = 1) +
coord_cartesian(c(50,70)) # Zoomed View of the 50 to 65 range of age

ggplot(data_clean) + geom_histogram(aes(Age), binwidth = 1) +
coord_cartesian(c(10, 30)) # Zoomed View of the 15 to 30 range of age

The boxplot (and histogram) for the Age variable show that there are no outliers
and that the range of Age is reasonably between 18 to 59 with no negative values.

Weight variable

ggplot(data_clean) + geom_boxplot(aes(Weight..kg.)) # boxplot for the Weight
variable

ggplot(data_clean) + geom_boxplot(aes(Weight..kg.)) + coord_cartesian(c(120, 140))
# Zoomed boxplot view of the 120 to 140 range of weight

ggplot(data_clean) + geom_histogram(aes(Weight..kg.), binwidth = 0.5)

ggplot(data_clean) + geom_histogram(aes(Weight..kg.), binwidth = 0.5) +
coord_cartesian(c(30,50)) # Zoomed view of the 30 to 50 range of weight

ggplot(data_clean) + geom_histogram(aes(Weight..kg.), binwidth = 0.5) +
coord_cartesian(c(120,140)) # Zoomed view of the 120 to 140 range of weight

The boxplot for the Weight variable shows that there are a handful of outliers,
on the RHS of the boxplot, according to the IQR standard, however the differences
between the outliers and non-outliers beside them are not drastically big, and
they don't seem to be biologically unusual/impossible, so removing them would not
benefit the data set much. There aren't any negative values either.

Height variable

ggplot(data_clean) + geom_boxplot(aes(Height..m.)) # boxplot for the Height
variable

ggplot(data_clean) + geom_histogram(aes(Height..m.), binwidth = 0.01) # histogram
for the Height variable

ggplot(data_clean) + geom_histogram(aes(Height..m.), binwidth = 0.01) +
coord_cartesian(c(1.45, 1.55)) # Zoomed view of the 1.45 to 1.55 range of height
```

```
ggplot(data_clean) + geom_histogram(aes(Height..m.), binwidth = 0.01) +
coord_cartesian(c(1.95, 2.05)) # Zoomed view of the 1.95 to 2.05 range of height


# The boxplot (and histogram) for the Height variable shows that there are no
outliers or negative values, moreover, the range of height is 1.5 to 2.0 meters
which is reasonable.


# BPM variables (Max, Average, & Resting)

ggplot(data_clean) + geom_boxplot(aes(y=Max_BPM)) + geom_boxplot(aes(y=Avg_BPM)) +
geom_boxplot(aes(y=Resting_BPM)) # boxplot for the BPM variables


# Max BPM

ggplot(data_clean) + geom_histogram(aes(Max_BPM), binwidth = 1) # histogram for
the Max BPM variable

ggplot(data_clean) + geom_histogram(aes(Max_BPM), binwidth = 1) +
coord_cartesian(c(155, 165)) # Zoomed view of the 160 to 165 range of Max BPM

ggplot(data_clean) + geom_histogram(aes(Max_BPM), binwidth = 1) +
coord_cartesian(c(195, 200)) # Zoomed view of the 195 to 200 range of Max BPM


# Average BPM

ggplot(data_clean) + geom_histogram(aes(Avg_BPM), binwidth = 1) # histogram for
the Average BPM variable

ggplot(data_clean) + geom_histogram(aes(Avg_BPM), binwidth = 1) +
coord_cartesian(c(115, 125)) # Zoomed view of the 120 to 125 range of Average BPM

ggplot(data_clean) + geom_histogram(aes(Avg_BPM), binwidth = 1) +
coord_cartesian(c(165, 170)) # Zoomed view of the 165 to 170 range of Average BPM


# Resting BPM

ggplot(data_clean) + geom_histogram(aes(Resting_BPM), binwidth = 1) # histogram
for the Resting BPM variable

ggplot(data_clean) + geom_histogram(aes(Resting_BPM), binwidth = 1) +
coord_cartesian(c(45, 55)) # Zoomed view of the 50 to 55 range of Resting BPM

ggplot(data_clean) + geom_histogram(aes(Resting_BPM), binwidth = 1) +
coord_cartesian(c(70, 75)) # Zoomed view of the 70 to 75 range of Resting BPM


# The boxplot (and histogram) for the Max, Average, and Resting BPM variables
shows that there are no outliers or negative values for all of them, moreover, the
```

range of the variables are a reasonable range for them, with 160 to 200 for Max BPM, 120 to 170 for Average BPM, and 50 to 70 for Resting BPM

Session Duration variable

ggplot(data_clean) + geom_boxplot(aes(Session_Duration..hours.)) # boxplot for the Session Duration variables

ggplot(data_clean) + geom_histogram(aes(Session_Duration..hours.), binwidth = 0.1) # histogram for the Session Duration variables

ggplot(data_clean) + geom_histogram(aes(Session_Duration..hours.), binwidth = 0.1) + coord_cartesian(c(0, 0.8)) # Zoomed view of the to range of Session Duration

ggplot(data_clean) + geom_histogram(aes(Session_Duration..hours.), binwidth = 0.1) + coord_cartesian(c(1.9, 2.3)) # Zoomed view of the to range of Session Duration

The boxplot for the Session Durations is almost symmetric with no outliers, no negative values, and no unusual durations. The histogram shows no outliers either, even when zoomed.

Calories Burned variables

ggplot(data_clean) + geom_boxplot(aes(Calories_Burned)) # boxplot for the Calories Burned variables

ggplot(data_clean) + geom_boxplot(aes(Calories_Burned)) + coord_cartesian(c(1600, 1800)) # Zoomed view of the 1600 to 1800 range of Resting BPM

ggplot(data_clean) + geom_histogram(aes(Calories_Burned), binwidth = 10) # histogram for the Calories Burned variables

ggplot(data_clean) + geom_histogram(aes(Calories_Burned), binwidth = 5) + coord_cartesian(c(0, 400)) # Zoomed view of the 0 to 400 range of Resting BPM

ggplot(data_clean) + geom_histogram(aes(Calories_Burned), binwidth = 5) + coord_cartesian(c(1600, 1800)) # Zoomed view of the 1600 to 1800 range of Resting BPM

The boxplot for the Calories Burned variable shows that there are some outliers with values between 1600 and 1800, however, these values are medically plausible so removing them will only lead to less data. Moreover, there are no negative and no unusual values. The histogram shows the same insight.

```
# Fat Percentage variable


ggplot(data_clean) + geom_boxplot(aes(Fat_Percentage)) # boxplot for the Fat Percentage Variable

ggplot(data_clean) + geom_histogram(aes(Fat_Percentage), binwidth = 0.1) # boxplot for the Fat Percentage Variable

ggplot(data_clean) + geom_histogram(aes(Fat_Percentage), binwidth = 0.1) + coord_cartesian(c(0, 12)) # Zoomed view of the 0 to 12 range of Fat Percentage

ggplot(data_clean) + geom_histogram(aes(Fat_Percentage), binwidth = 0.1) + coord_cartesian(c(32.5, 40)) # Zoomed view of the 32.5 to 40 range of Fat Percentage


# The boxplot for the Fat Percentage variable shows that there are no outliers, no negative values, and no unusual values in the data. The histogram shows the same insight.


# Water Intake variable


ggplot(data_clean) + geom_boxplot(aes(Water_Intake..liters.)) # boxplot for the Water Intake variable

ggplot(data_clean) + geom_histogram(aes(Water_Intake..liters.), binwidth = 0.1) # boxplot for the Water Intake variable

ggplot(data_clean) + geom_histogram(aes(Water_Intake..liters.), binwidth = 0.1) + coord_cartesian(c(0, 1.75)) # Zoomed view of the 1.5 to 1.8 range of Water Intake

ggplot(data_clean) + geom_histogram(aes(Water_Intake..liters.), binwidth = 0.1) + coord_cartesian(c(3.5,4.0)) # Zoomed view of the 3.7 to 4.0 range of Water Intake


# The boxplot for the Water Intake variable shows that there are no outliers, no negative values, and no unusual values in the data. It would make sense for the minimum Water Intake to be 1.5 liters since they're gym members, they would need a lot of water, so the range is understandable. The histogram shows the same insight.


# Workout Frequency variable


ggplot(data_clean) + geom_bar(aes(Workout_Frequency..days.week.)) # bar chart for the Workout Frequency variable
```

```
# Since the Workout Frequency is a categorical ordinal variable, a bar chart will
better represent the data. There aren't any unusual shapes and no outliers or
negative values show up. The range of Workout Frequency is from 2 to 5, this makes
sense because a gym member would likely workout more than once a week and will
also have at least 2 rest days in the week leaving 5 days for them to workout.


# Experience Level variable


ggplot(data_clean) + geom_bar(aes(Experience_Level)) # bar chart for the
Experience Level variable


# Since the Experience Level variable is a categorical ordinal variable, a bar
chart will represent the data better. There are only 3 levels of experience in the
data set. There are no outliers, negative, or unusual values for Experience
Levels.


# Gender Variable


data_clean %>%
  count(Gender)


# Gender      n
# <chr>     <int>
# Female        462
# Male     511


ggplot(data_clean) + geom_bar(aes(Gender)) # bar chart for the Gender variable


# Since it is categorical, a bar chart is better suited here. The graph shows that
there are only 2 genders in the data set, meaning there were no typos or unusual
entries for this variable. The count reconfirms this. Moreover, there are more
male gym members than female.


# Workout Type variable


data_clean %>%
```

```
  count(Workout_Type)


#Workout_Type      n
#   <chr>        <int>
#  Cardio            255
#   HIIT             221
# Strength         258
#   Yoga             239


ggplot(data_clean) + geom_bar(aes(Workout_Type)) # bar chart for the Workout Type Variable


# Since it is categorical, a bar chart is better suited here. The graph shows that there are 4 Workout Types in the data set, meaning there were no typos or unusual entries for this variable. The count reconfirms this. The spread of the types of Workouts the people do seem to balance, with each workout being done by almost quarter of different gym members. The Strength workout has the most count, followed by the Cardio, Yoga, then HIIT workouts. The balance between the workout types is good for fair comparison between other variables.
```


Part C — Missing Values Analysis

Identify missing values, visualize patterns, and explain whether missingness is systematic.


```{r}
sum(is.na(data))

# To check if the data had any missing values, we summed the instances of NA data (before we imputed extreme BMI). The result being 0 shows that there are no missing data in the original data set.
```


Part D — Covariation Between Categorical and Continuous Variables

Use boxplots, reorder categories, interpret medians, variability, skewness, and outliers.

```{r}

ggplot(data_clean) +
geom_boxplot(aes(x=reorder(Workout_Type,Calories_Burned,FUN=median),
y=Calories_Burned))

# This boxplot shows how more calories are burned (higher median) in higher intensity workouts like HIIT, Strength and Cardio compared to Yoga. The boxes are nearly symmetric, with Yoga having a slight right-skew. The range of Yoga is the largest and this is likely because there are many types of yoga, some of which are more fast-paced and causes higher calorie burn. The outliers above Cardio, Strength, and HIIT suggest that very extreme workouts can be done through these types, leading to even more calorie burn. Across all 4 categories, there's near-consistent variability, with HIIT having the largest and Strength having the lowest.


ggplot(data_clean) +
geom_boxplot(aes(x=reorder(Experience_Level,Fat_Percentage,FUN=median),
y=Fat_Percentage))

ggplot(data_clean) +
geom_boxplot(aes(x=reorder(Workout_Frequency..days.week.,Fat_Percentage,FUN=median
), y=Fat_Percentage))

# Both boxplots show a similar result. The median fat percentage decreases with higher experience level and weekly workout frequency. This is an expected result, as people who are more experienced in fitness and workout more frequently are likely more muscular and carry less fat than average. In the first plot, the experience levels are all symmetric with no indications of skewness. Levels 2 and 3 have a higher variability than level 1. In the second plot, frequencies 5,2 and 3 show similar variability and little to no skew, while frequency 4 has very large variability and a strong left skew.


ggplot(data_clean) +
geom_boxplot(aes(x=reorder(Workout_Frequency..days.week.,Water_Intake..liters.,FUN
=median), y=Water_Intake..liters.))

# This boxplot shows that water intake generally increases with workout frequency. The variability is high in frequencies 2,3 and 4 with a slight right-skew across the 3 boxes. The box for frequency 5 shows very low variability, and an extreme left-skew. The median seems to overlap almost completely with Q3, indicating that the data of water intake above the median are all near-identical.

```


Part E — Covariation Between Two Continuous Variables

Use scatterplots. Demonstrate overplotting and fix using alpha. Compute correlation and interpret.

```{r}

ggplot(data_clean) + geom_point(mapping = aes(x=Session_Duration..hours., y=Calories_Burned))

ggplot(data_clean) + geom_point(mapping = aes(x=Weight..kg., y=BMI))

# In both plots, there are many data points overlapping since, in each, the two variables appear to be positively linearly correlated, leading this scatterplot to be overplotted. Below we have used alpha = 0.2 to fix this issue, and increase visibility.


ggplot(data_clean) + geom_point(mapping = aes(x=Session_Duration..hours., y=Calories_Burned), alpha = 0.2)

# Calories burned and session duration appear to be clearly correlated in a strong, positive, linear relationship. This makes sense logically as the longer a workout is, the more the gym members' bodies will burn calories to produce energy.

cor(data_clean$Session_Duration..hours., data_clean$Calories_Burned)

# Our conclusion from the plot is supported by the numerical correlation being 0.908. This is an extremely strong positive correlation.


ggplot(data_clean) + geom_point(mapping = aes(x=Weight..kg., y=BMI), alpha = 0.2)

# Weight and BMI also seem positively correlated, but less strongly than with calories burned and session duration, as this plot's relationship appears to have data points in a looser linear relationship. The positive proportionality is expected, as weight is a big factor in determining BMI. But since height also has an effect on BMI, the relationship isnt so strong.

no_NA_data_clean = filter(data_clean, !is.na(BMI)) # since the NA values lead correlation to output NA

cor(no_NA_data_clean$Weight..kg., no_NA_data_clean$BMI)

# Our conclusion from the plot is supported by the numerical correlation being 0.808. This is a strong positive correlation, but less strong than calories burned and session duration.

```


Part F — Covariation Between Two Categorical Variables

Use geom_count or heatmaps to visualize category combinations and interpret relationships.


```{r}

library(ggplot2)
```

```
library(dplyr)

data_clean %>%

  count(Gender, Workout_Type) %>%

  ggplot(aes(Gender, Workout_Type, fill = n)) + geom_tile()

# Since the data set has 2 categorical variables that we have not yet compared, the heatmap will represent the relationship between Gender and Workout Type. The darker shades represent low frequencies and the lighter shades represent higher frequencies.

# The heatmap doesn't show a strong association between Gender and Workout Type. Participation in the Workout Types are evenly & similarly distributed across the two genders with a slight difference, most probably due to the fact that the data set has more data about males than females. Although the shades for males are lighter, meaning a higher frequency of participation, the differences are small (the range of counts are from 110 to 135). In conclusion, the Gender of the gym member does not influence their Workout Type preference.

```