Gym-Members Dataset Analysis

Disclosure

*This dataset is **artificial**, meaning **some insights may not accurately reflect reality**. Before beginning the analysis, we **must identify the limitations** of what can be meaningfully inferred versus what is clearly a **byproduct of dataset fabrication**.*

*For example, when plotting **Weight vs. Height** for males and females, a clear pattern emerges:*

- ***Females** were assigned **heights between 1.5m and 1.8m** and **weights between 40 kg and 80 kg**.*
- *This artificial constraint means the graph **does not provide real-world insights**, as the data distribution is predefined.*

*Due to such constraints, **certain analyses will be excluded**, as they **do not contribute meaningful conclusions**.*

Summary:

Gym-Members per gender:

	Males	Females
Amount	511	462
Percentage	52.5%	47.5%

Table 1: Amount and percentage of male and female Gym-Members.

Gym-Members per WorkOut Type:

	Strength	Cardio	Yoga	HIIT
Amount	258	255	239	221
Percentage	26.5%	26.2%	24.6%	22.7%

Table 2: Amount and percentage of gym-members that participate in the different WorkOut Types available at the gym.

Attribues general characteristics:

Index	Data Type	#missing	Duplicate	#Unique	Min	Max	Avg	Std dev	Top Value	Freq
Age	int64	0	0	42	18	59	38.7	12.2	N/A	N/A
Gender	object	0	0	2	N/A	N/A	N/A	N/A	Male	511
Weight	float64	0	0	523	40	129.9	73.8	21.2	N/A	N/A
Height	float64	0	0	51	1.5	2	1.7	0.1	N/A	N/A
Max BPM	int64	0	0	40	160	199	179.9	11.5	N/A	N/A
Avg BPM	int64	0	0	50	120	169	143.8	14.3	N/A	N/A
Resting BPM	int64	0	0	25	50	74	62.2	7.3	N/A	N/A
Session Duration	float64	0	0	147	0.5	2	1.2	0.3	N/A	N/A
Calories Burned	float64	0	0	621	303	1783	905.4	272.6	N/A	N/A
WorkOut Type	object	0	0	4	N/A	N/A	N/A	N/A	Strength	258
Fat Percentage	float64	0	0	239	10	35	25	6.2	N/A	N/A
Water Intake	float64	0	0	23	1.5	3.7	2.6	0.6	N/A	N/A
WorkOut Frequency	int64	0	0	4	2	5	3.3	0.9	N/A	N/A
Experience Level	int64	0	0	3	1	3	1.8	0.7	N/A	N/A
BMI	float64	0	0	771	12.3	49.8	24.9	6.7	N/A	N/A

Table 3: General characteristics derived from an exploratory data analysis (EDA) of the database.

Correlation Matrix:

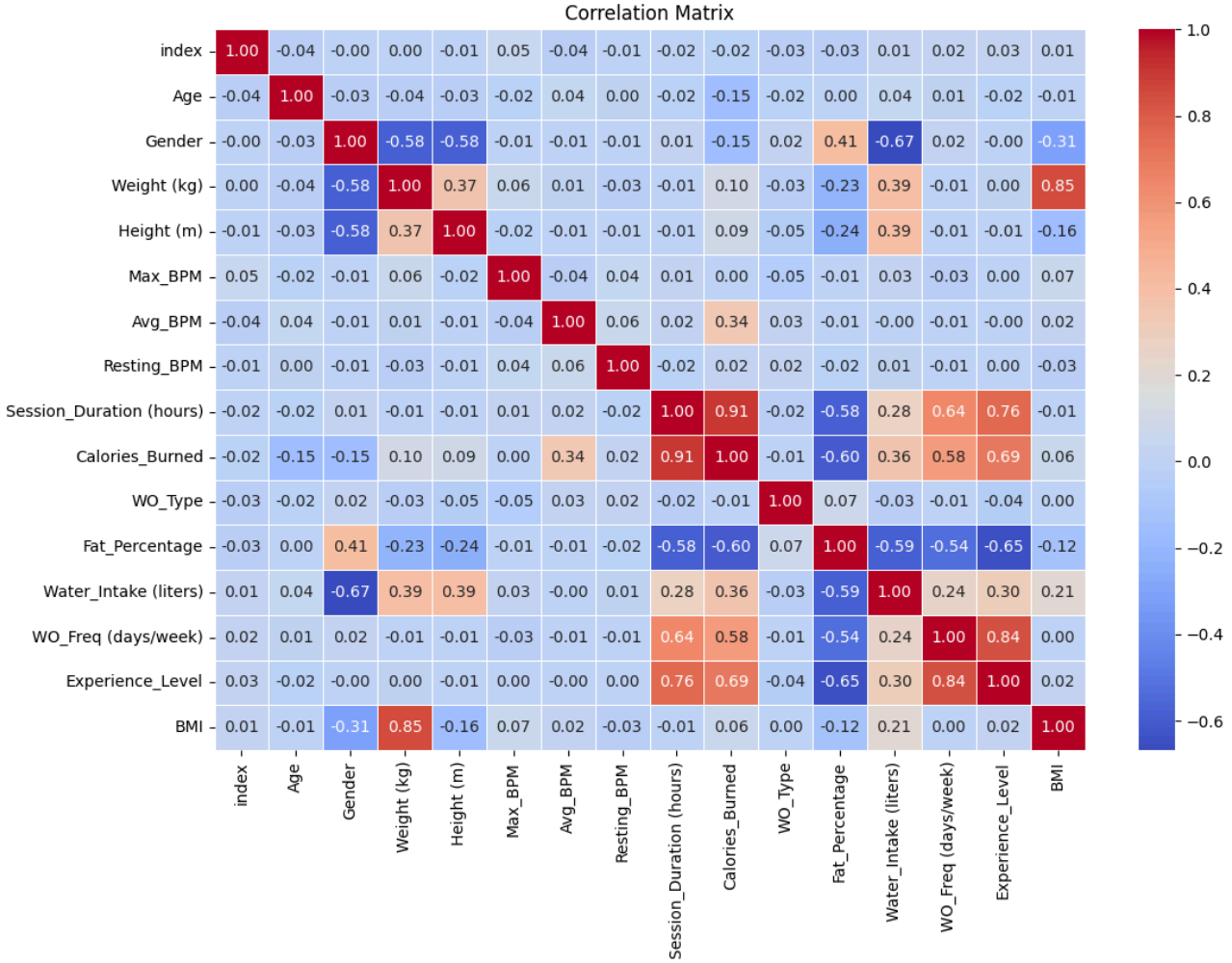

Figure 1: Correlation matrix of all the attributes of the database.

By analyzing **correlation_Matrix.png**, we can observe various relationships between attributes. A correlation above **|0.2|** is considered meaningful:

- **Positive correlation (≥ 0.2):** If one attribute increases, the other tends to increase.
- **Negative correlation (≤ -0.2):** If one attribute increases, the other tends to decrease.
- **Weak or no correlation (-0.2 to 0.2):** The attributes are largely independent.

General characteristics:

- **Age**, as expected, seems to have no distinguishable correlation with any other variable.
- For **gender** I established that Male is -1 and Female is 1 so it can be used in the correlation matrix. A high positive correlation with gender indicates a stronger

association with being Female, while a strong negative correlation suggests a stronger association with being Male.
    - **Female correlations**: Fat Percentage shows a positive correlation of 0.41, meaning higher fat percentage is associated with being Female.
    - **Male correlations**: Weight (-0.58), Height (-0.58), Water Intake (-0.67), and BMI (-0.31) show negative correlations, meaning higher values for these features are associated with being Male.

Body:

- **Weight** seems to be correlated to being Male (-0.58), positively correlated to Height (0.37), negatively correlated to Fat Percentage (-0.23), positively correlated to Water Intake (0.39), and highly correlated to BMI (0.85).
- **Height** seems to be correlated to being Male (-0.58), positively correlated to Weight (0.37), negatively correlated to Fat Percentage (-0.24), positively correlated to Water Intake (0.39), and highly correlated to BMI (0.85).
- **Fat Percentage** seems to be correlated to being Female (0.41), negatively correlated to Weight (-0.23), negatively correlated to Height (-0.24), negatively correlated to Session Duration (-0.58), negatively correlated to Calories Burned (-0.60), negatively correlated to Water Intake (-0.59), negatively correlated to WorkOut Frequency (-0.54), and negatively correlated to Experience Level (-0.65).
- **BMI** seems to be correlated to being Male (-0.31), positively correlated to Weight (0.85), and positively correlated to Water Intake (0.21).

Heart Rate:

- **Max BPM** seems to have no distinguishable correlation with any other variable.
- **Average BPM** seems to only be positively correlated to Calories Burned (0.34).
- **Resting BPM** seems to have no distinguishable correlation with any other variable.

Workout related attributes:

- **Session Duration** seems to be highly correlated to Calories Burned (0.91) (as expected), negatively correlated to Fat Percentage (-0.58), positively correlated to Water Intake (0.28), positively correlated to WorkOut Frequency (0.64), and highly correlated to Experience Level (0.76).
- **Calories Burned** seems to be positively correlated to Average BPM (0.34), highly correlated to Session Duration (0.91), negatively correlated to Fat Percentage (-0.60), positively correlated to Water Intake (0.36), positively correlated to WorkOut Frequency (0.58), and positively correlated to Experience level (0.69).
- **WorkOut Type** seems to have no distinguishable correlation with any other variable.
- **Water Intake** seems to be correlated to being Male (-0.67), positively correlated to Weight (0.39), positively correlated to Height (0.39), positively correlated to Session Duration (0.28), positively correlated to Calories Burned (0.36), negatively correlated to Fat Percentage (-0.59), positively correlated to WorkOut Frequency (0.24), positively correlated to Experience Level (0.30), and positively correlated to BMI (0.21).
- **WorkOut Frequency** seems to be positively correlated to Session Duration (0.64), positively correlated to Calories Burned (0.58), negatively correlated to Fat

Percentage (-0.54), positively correlated to Water Intake (0.24), and highly correlated to Experience Level (0.84).

- **Experience Level** seems to be positively correlated to Session Duration (0.76), positively correlated to Calories Burned (0.69), negatively correlated to Fat Percentage (-0.65), positively correlated to Water Intake (0.30), and highly correlated to WorkOut Frequency (0.84).

Bar Graphs:

Divided by gender:

Age Groups: The distribution of males and females across age groups appears similar.

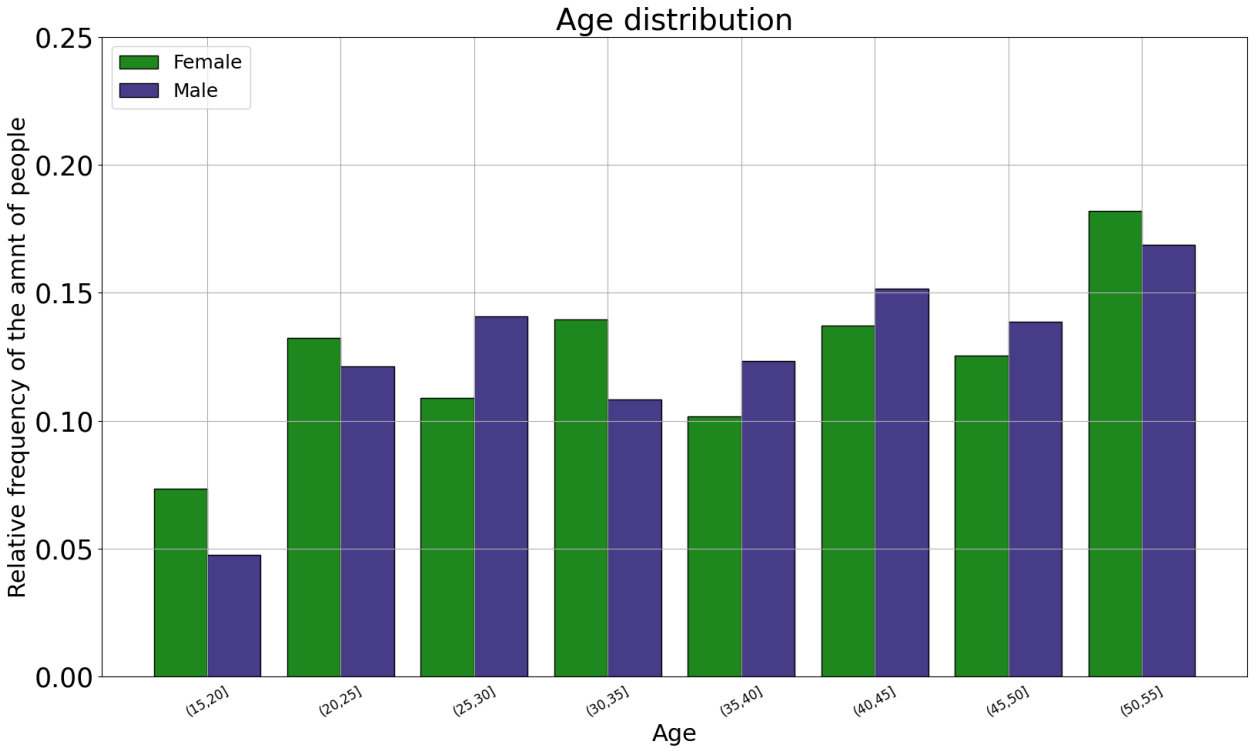

Figure 2: Age distribution of males and females.

Height & Weight:

- Females tend to be **shorter** than males. Females average around 1.62 m and males around 1.77 m.

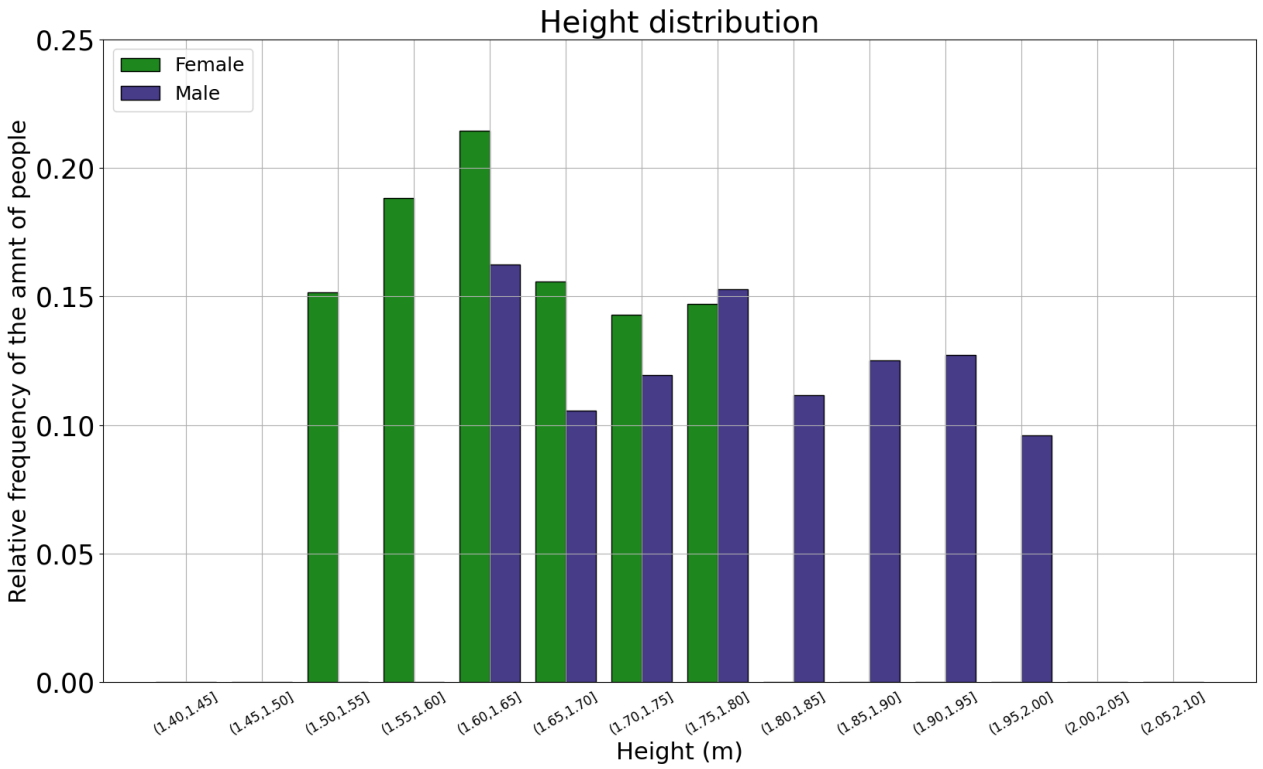

Figure 3: Height (m) distribution of males and females.

- Males generally weigh **more**, with a **high dispersion** and a **peak around 80-90 kg**. Females' **peak** is **around 60 kg**. The fact that females of higher weight don't go to the gym could be, for example, from the fear of being judged by others. However, this database can't answer this question.

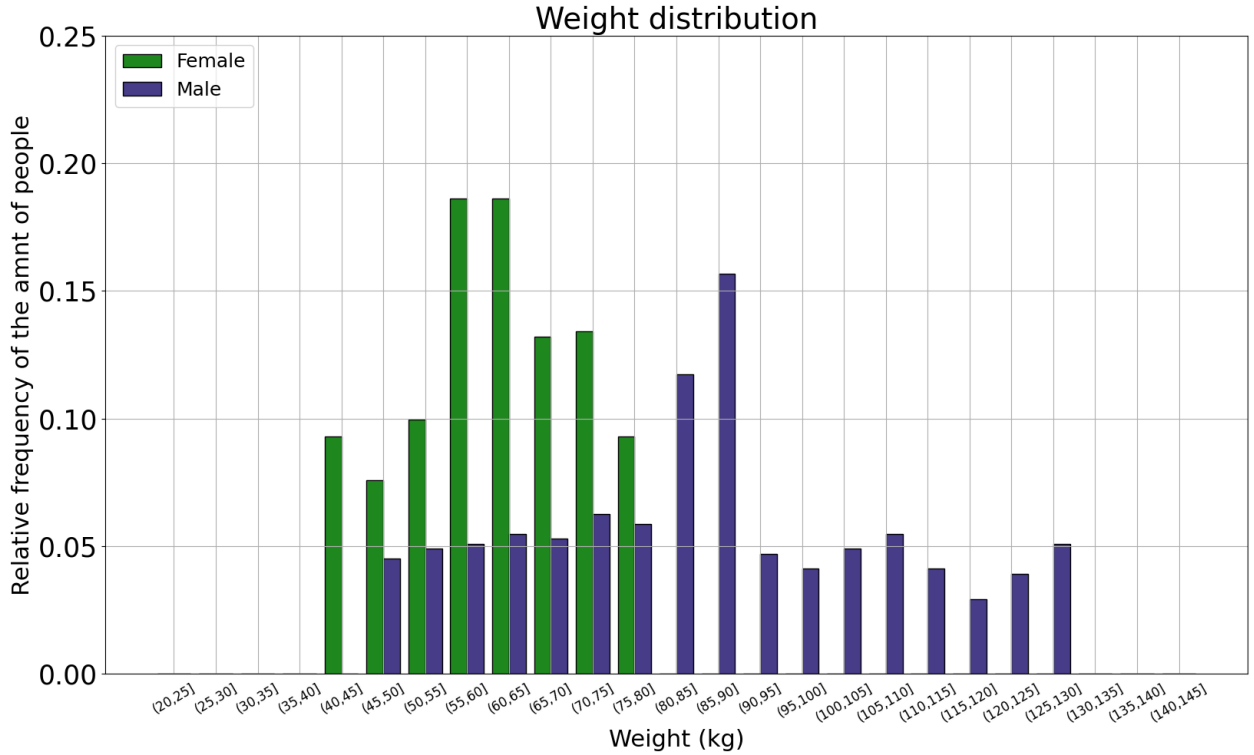

Figure 4: Weight (kg) distribution of males and females.

Heart Rate: Both **Max BPM** and **Resting BPM** are similar for males and females.

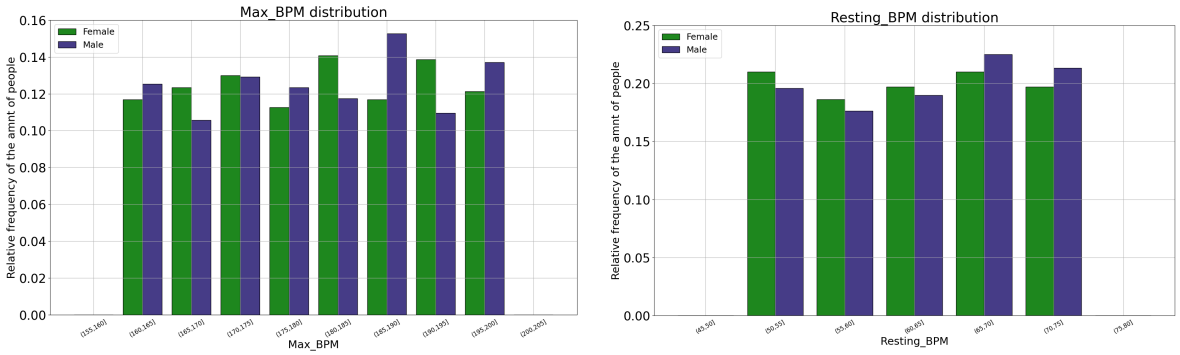

Figure 5: Max BPM and Resting BPM distribution of males and females.

Session Duration:

- Three distinct session durations are observed for both genders:
  - **30 min to 1 hour**
  - **1 hour to 1:30 hours** (most popular, with 61% of the gym members)
  - **1:30 hours to 2 hours**

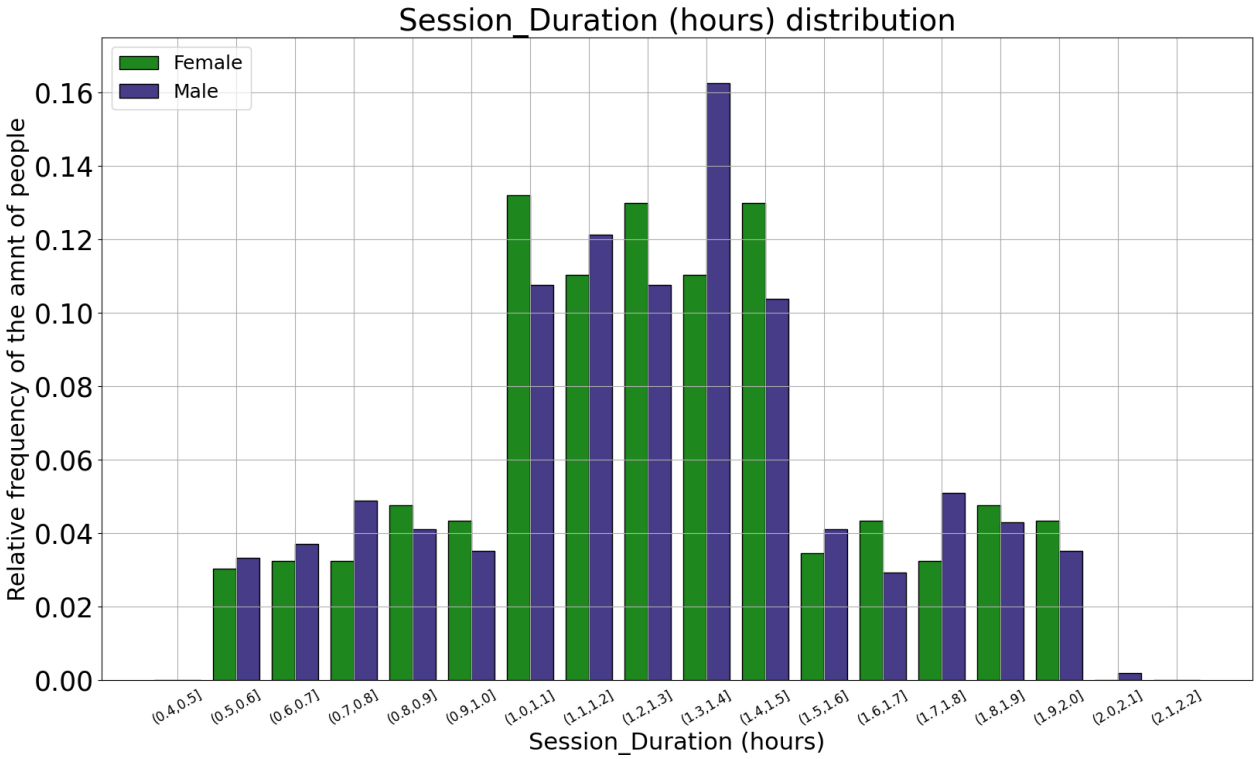

Figure 6: Session duration (hours) distribution of males and females.

Fat Percentage:

- Females tend to have a **higher fat percentage** than males.
- Both genders show **two distinct fat percentage groups**:
    - **Males:** 19.7% fall between **10-16%**, while the rest range from **20-32%**.
    - **Females:** 19.2% fall between **14-20%**, while the rest range from **24-36%**.

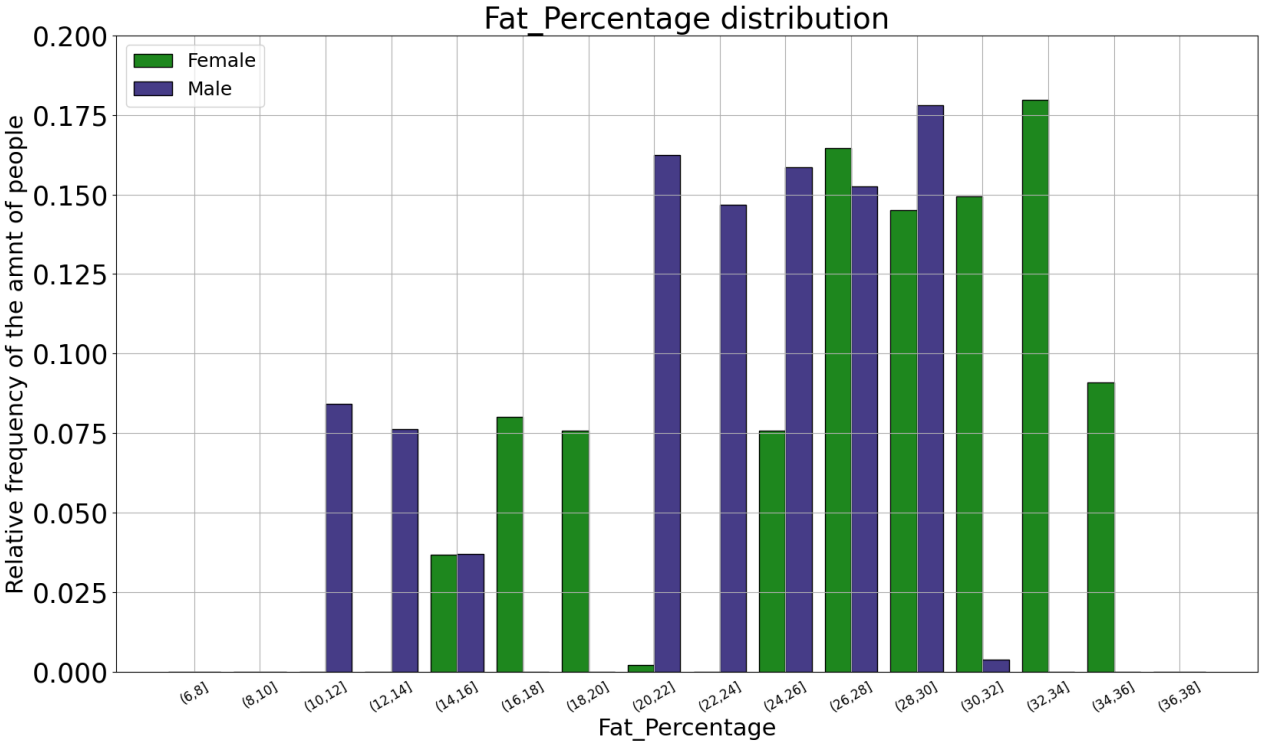

Figure 7: Fat percentage distribution of males and females.

Water Intake:

- Males tend to drink more water than females.
- High percentages are observed in specific ranges:
    - **27% of females** drink **2.6 to 2.8 L**.
    - **32% of males** drink **3.4 to 3.6 L**.

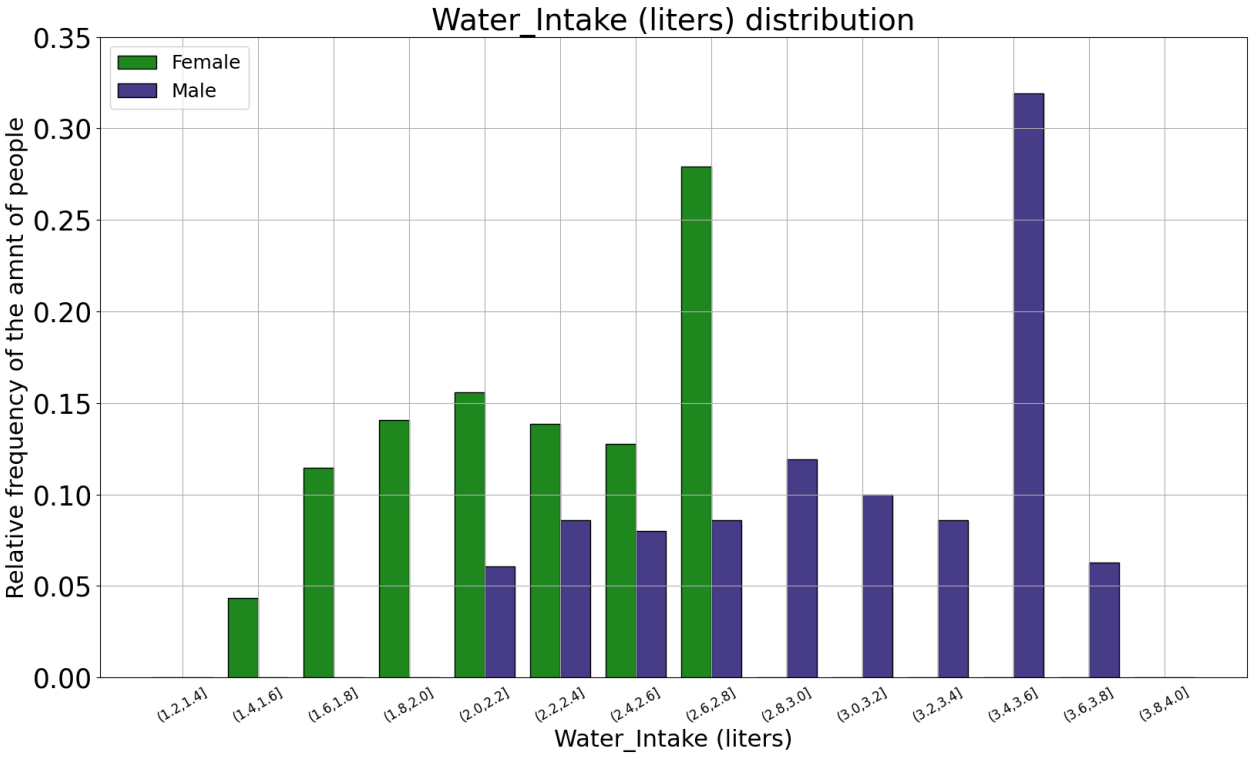

Figure 8: Water intake (L) distribution of males and females.

Workout Frequency & Experience Level: Similar distributions for both genders.

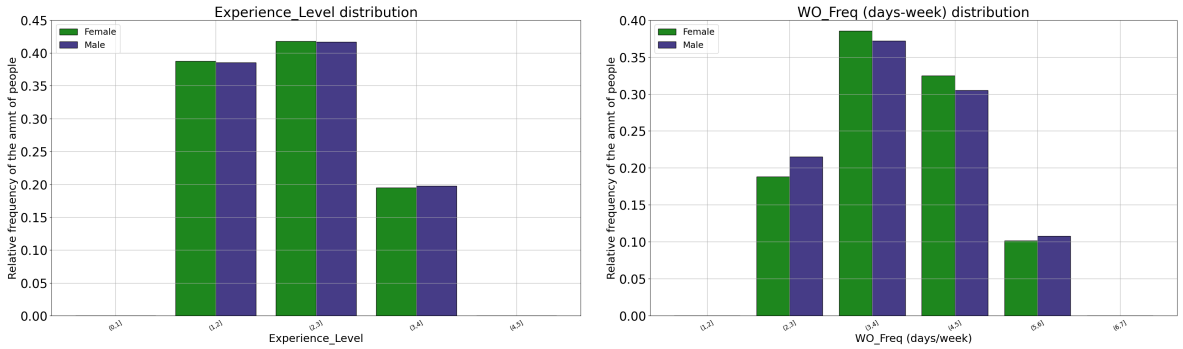

Figure 9: Workout frequency and experience level distribution of males and females.

BMI Distribution:

- Females: **~20 ± 10** (more centered distribution).
- Males: **~25 ± 20** (more dispersed).
- Both follow a **Gaussian distribution**.

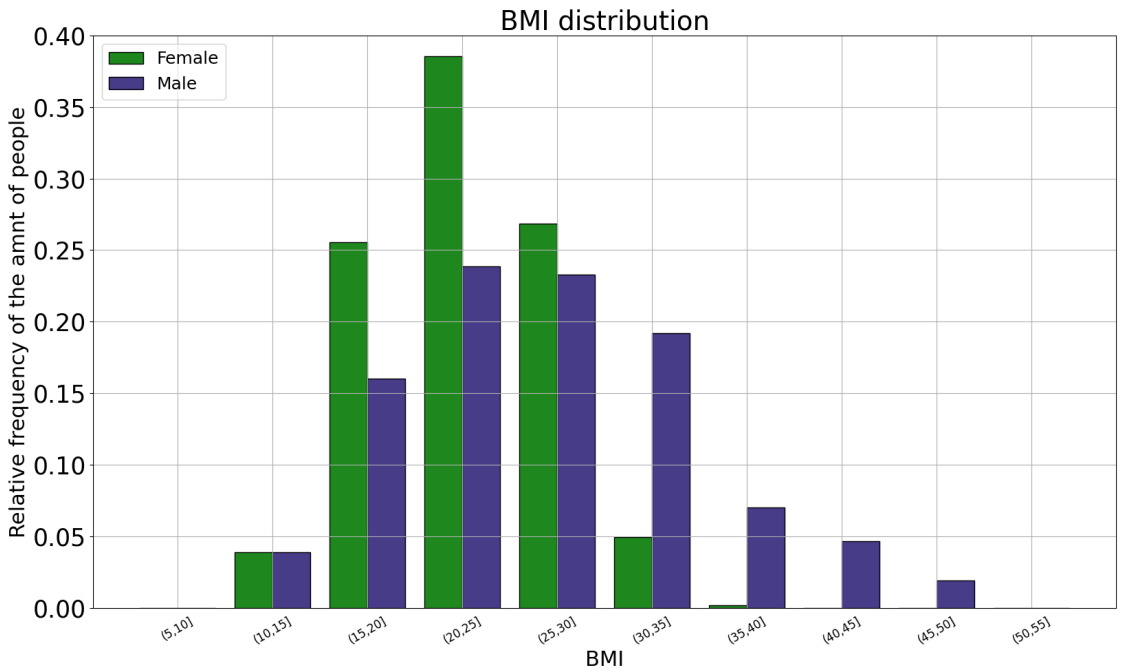

Figure 10: BMI distribution of males and females.

Divided by WorkOut Type:

Age Distribution:

- **Yoga:** 52% of attendees are **between 35-50**.
- **HIIT:** 36% are either **20-25 or 50-55**.
- **Cardio:** 7% are **15-20** years old, while the rest are **evenly distributed between 20-55**.
- **Strength Training:** Participation increases with **age**.

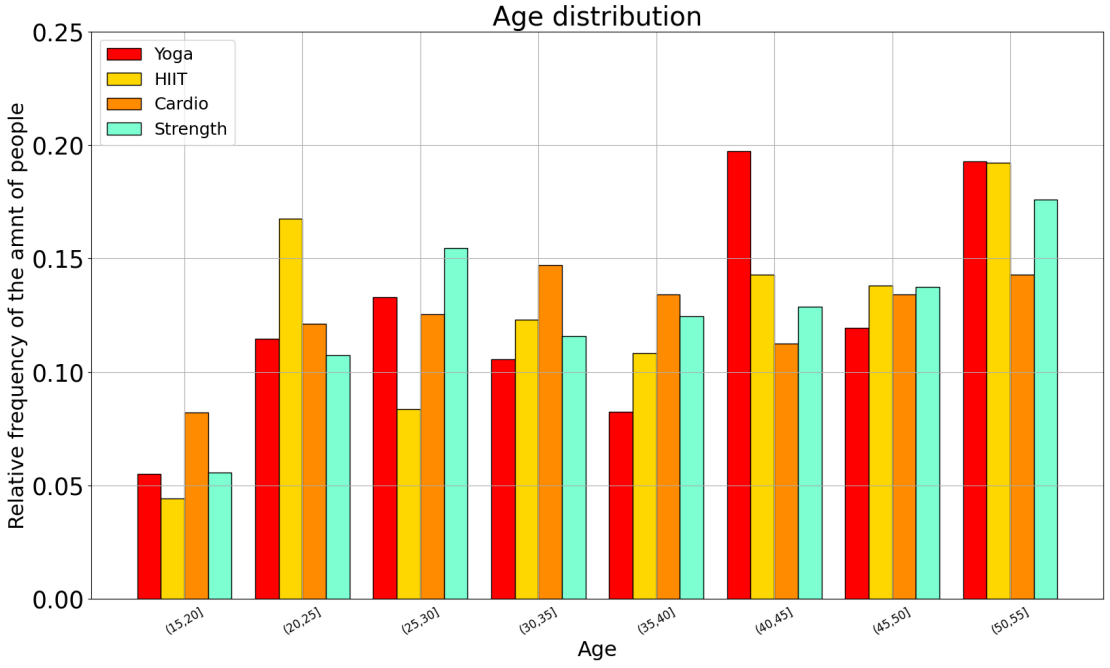

Figure 11: Age distribution of workout types.

Other Factors (Calories Burned, Resting BPM, Session Duration, Water Intake, and Max BPM): These metrics appear **independent of the workout type**.

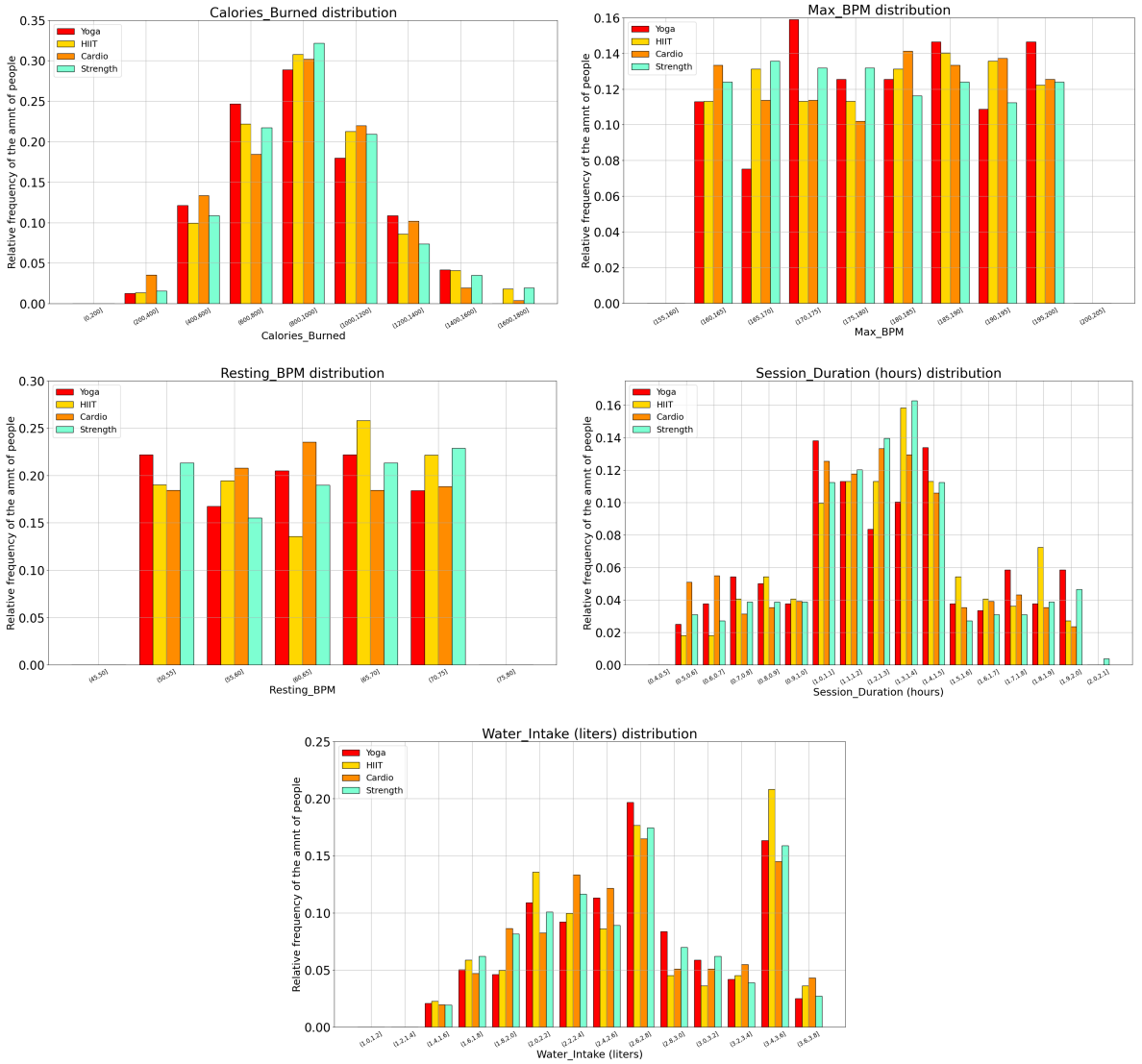

Figure 12: Calories burned, max BPM, resting BPM, session duration and water intake distribution of workout types.

Joint Plot Analysis

- **Session Duration & Calories Burned:**
  - A **strong positive correlation** is observed: **longer sessions lead to higher calorie burn**, as expected.
  - This trend is **consistent across both males and females**.

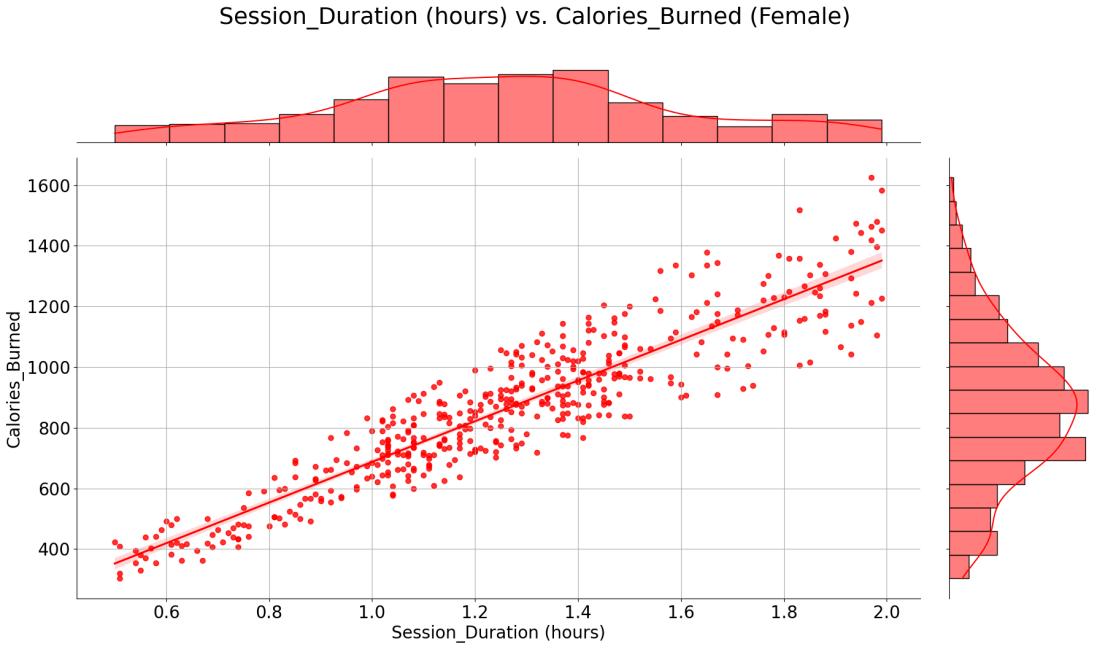

Figure 13: Joint plot comparing session duration and calories burned, for females.

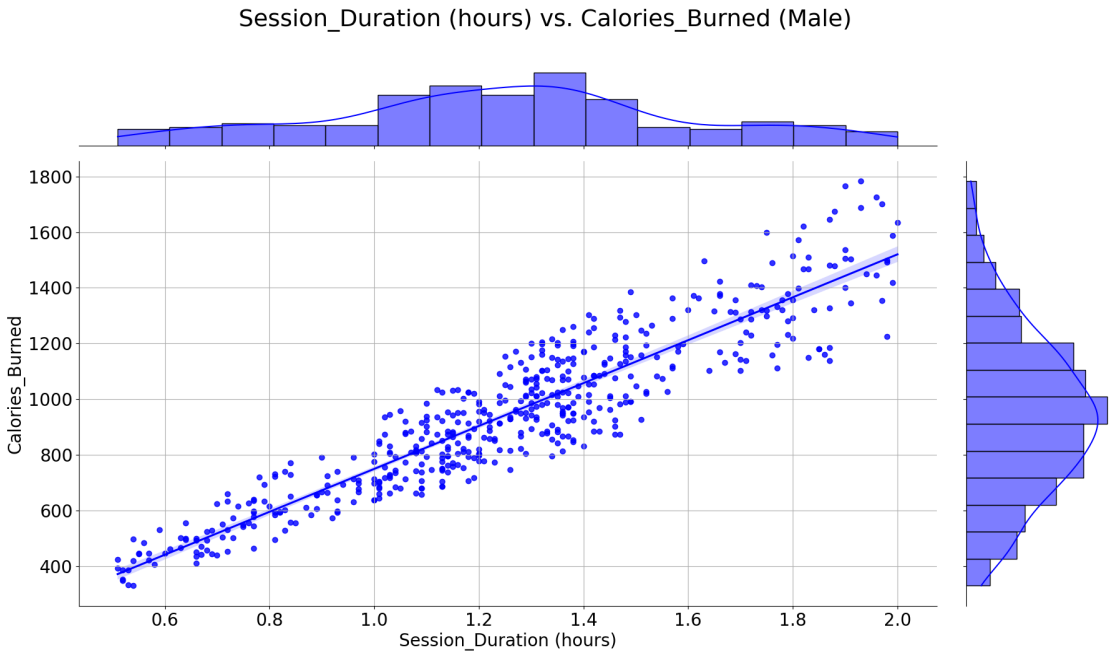

Figure 14: Joint plot comparing session duration and calories burned, for males.

- **Weight & BMI:**
  - As expected, **weight is positively correlated with BMI — heavier individuals** tend to have **higher BMI values**.
  - This holds true **regardless of gender**.

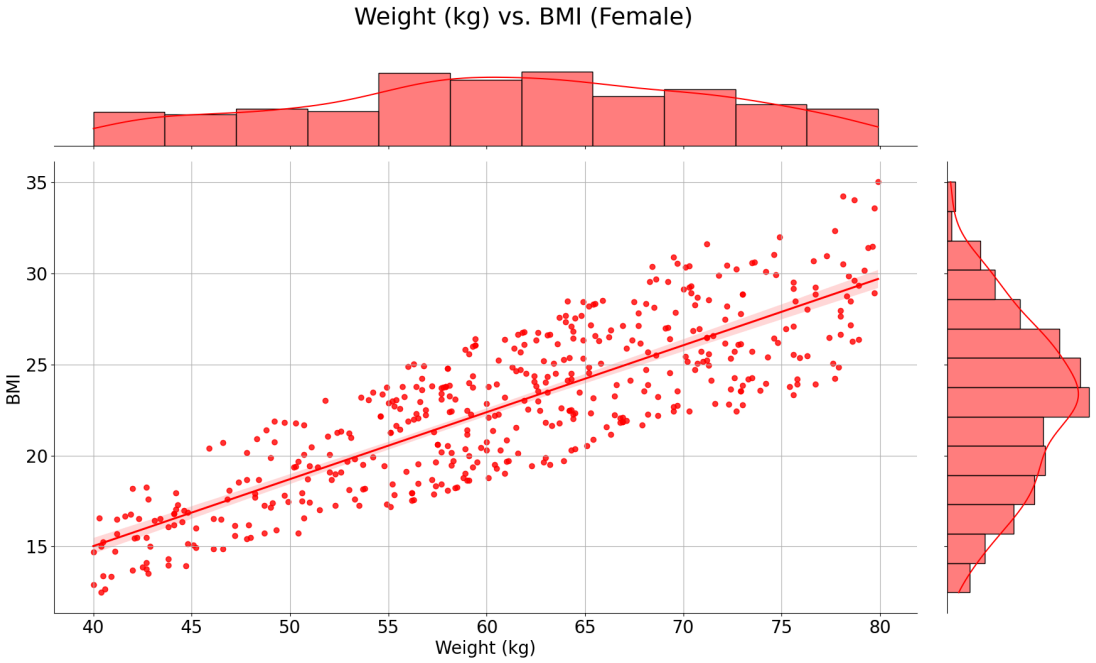

Figure 15: Joint plot comparing weight (kg) and BMI, for females.

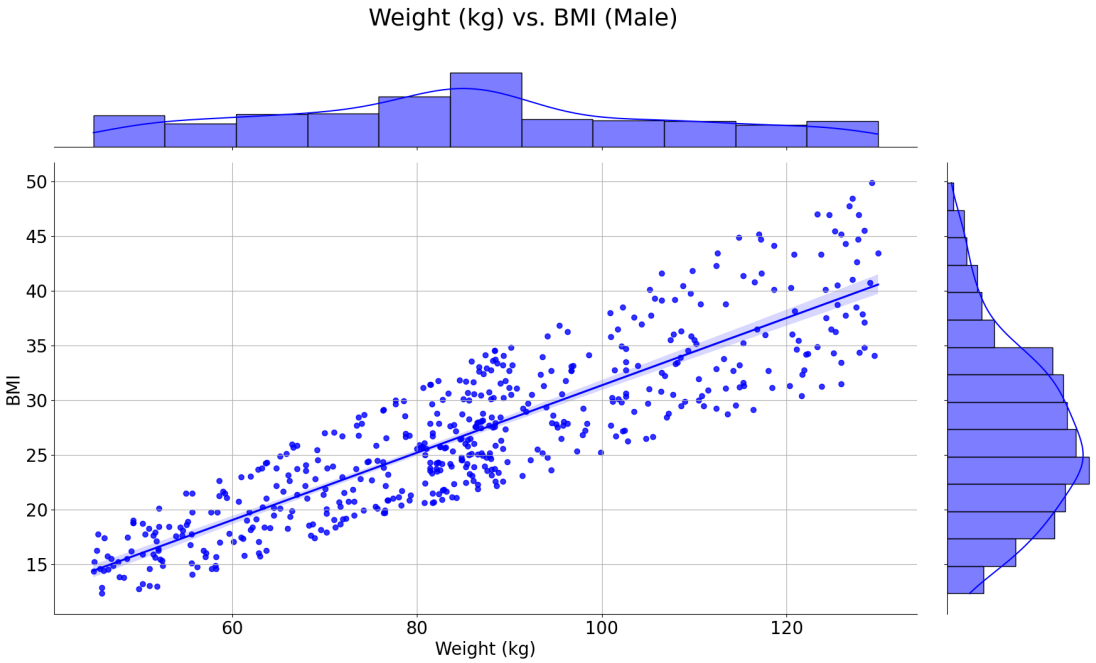

Figure 16: Joint plot comparing weight (kg) and BMI, for males.

- **Weight & Resting BPM:**
  - **No significant correlation** is found between weight and resting BPM.
  - This suggests that **individuals with different weights can have similar resting heart rates**.

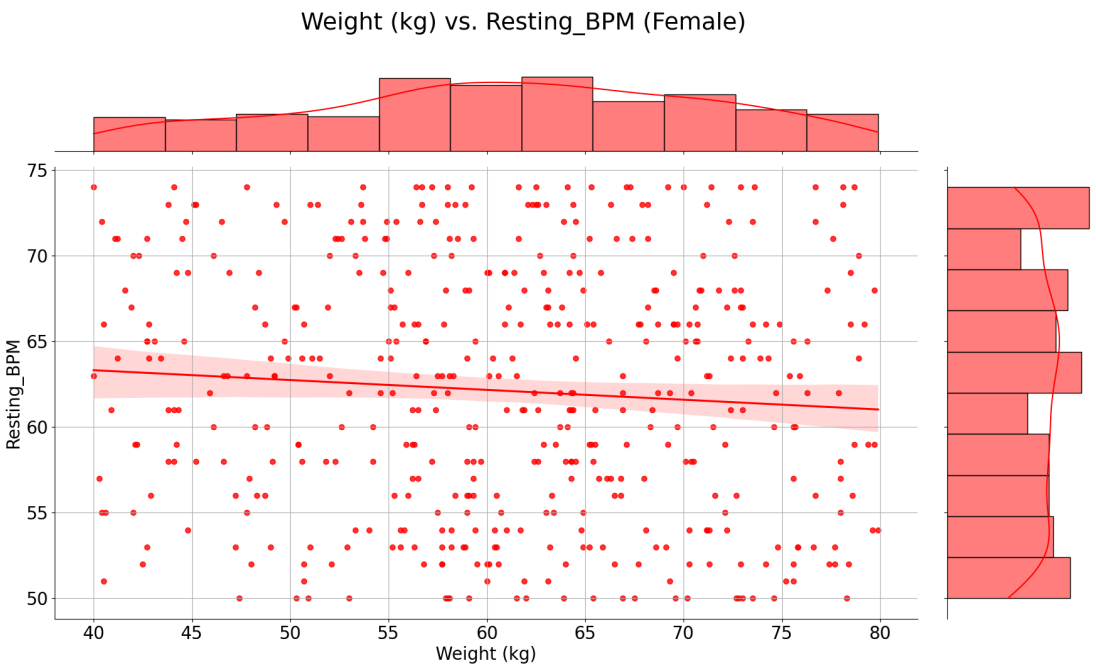

Figure 17: Joint plot comparing weight (kg) and Resting BPM, for females.

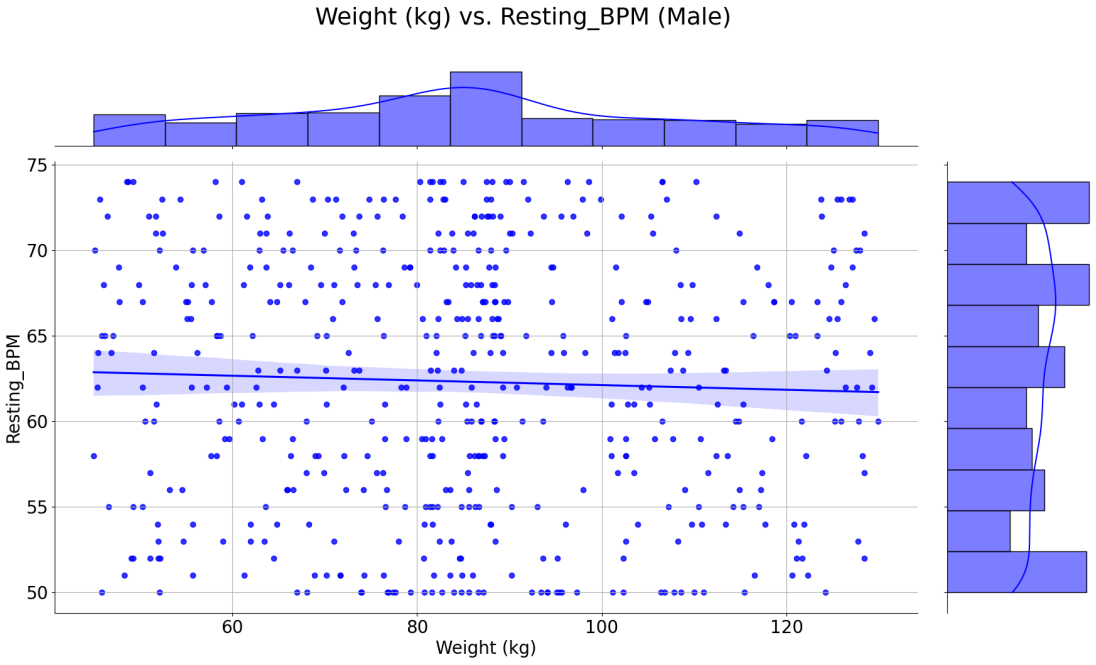

Figure 18: Joint plot comparing weight (kg) and Resting BPM, for males.

TensorFlow:

I wanted to make a model that could predict the amount of calories a new gym member would burn, depending on numerical features, such as: Age, Weight, Height, Session Duration, etc. and categorical features: Gender and WorkOut Type.

In order to train the model, I splitted the database:

- **80%** of it was used to **train** the **model**. **80%** of that was used to actually **train** the model and **20%** was used to **validate** it.
- **20%** was **held** till the end to evaluate the model using data not used for training.

I **scaled** all of the **numerical values** using the following formula:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

where μ and σ is the mean and the std of the attribute, respectively. I also **took out** a few **outliers** to **improve** the **fitting** of the model and **transformed** the **categorical columns** into numbers so they can also be **used** in the **training** of the model.

I built a **neural network** using **TensorFlow** with the following architecture:

- **Three hidden layers** with **ReLU activation**, each having 10, 30, and 64 neurons, respectively.
- A **Dropout layer (0.2)** to prevent overfitting.
- A **linear activation output layer** since this is a regression problem.
- The model was compiled with **Mean Squared Error (MSE)** as the loss function and **Adam optimizer** for efficient training.
- **Up to 200 epochs** to allow sufficient learning.
- **Batch size of 32**, meaning weights are updated after processing 32 samples at a time.
- An **early stopping mechanism** was applied to halt training when the validation loss stopped improving, preventing overfitting.

After training, I evaluated the model using **Relative Root Mean Squared Error (Rel RMSE), and R-squared (R^2).**

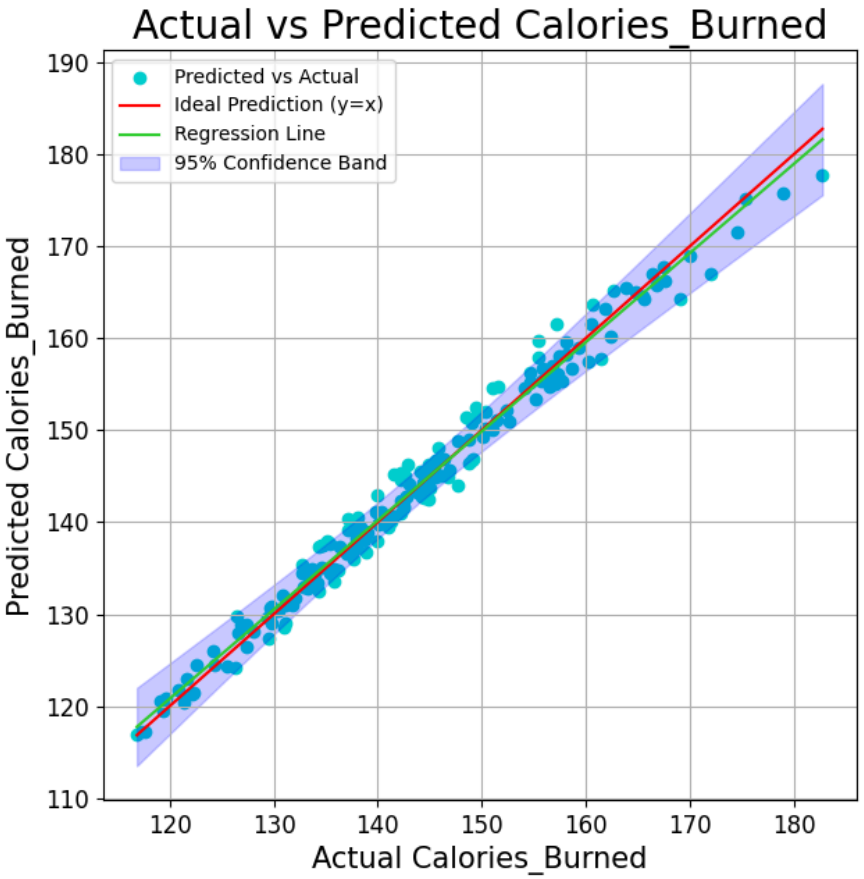

Figure 19: Graph comparing predicted calories burned vs. actual calories burned.

The model achieved an **R^2 of 0.982**, indicating that it captures a significant portion of the variance in calorie expenditure. Additionally, it has a **relative RMSE of 0.012**, meaning that the model's error is about 1.2% of the average calories burned, which suggests a reasonably accurate prediction performance. This makes it a useful tool for estimating caloric burn for new gym members based on their physical attributes and workout details.