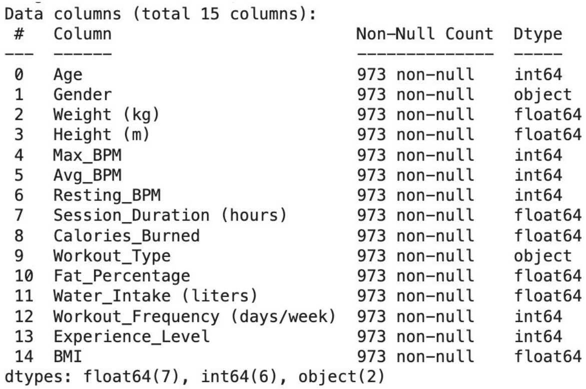
# Gym Members Exercise Analysis

## Applied Machine Learning - *Group 27*

Tattie Chitrakorn (tc3117), Eesun Moon (em3907), Akshara Pramod (ap4613), Tianjun Zhong ( tz2634), Jayanthi Yerchuru (jy3344)

# **Data Overview**: Gym Members Exercise Dataset [Link]

## ☑️ Feature Information



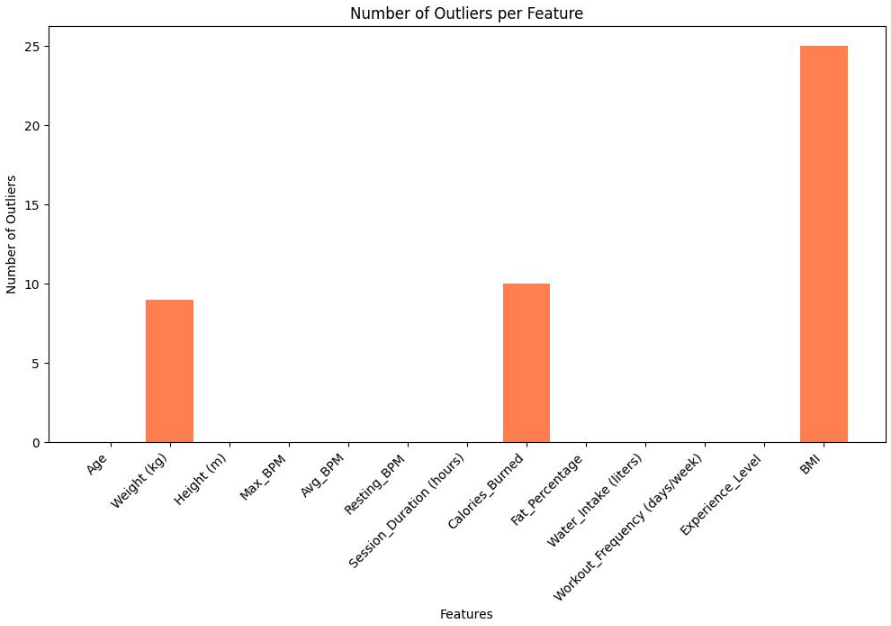
```
Data columns (total 15 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Age                            973 non-null    int64
 1   Gender                         973 non-null    object
 2   Weight (kg)                    973 non-null    float64
 3   Height (m)                     973 non-null    float64
 4   Max_BPM                        973 non-null    int64
 5   Avg_BPM                        973 non-null    int64
 6   Resting_BPM                    973 non-null    int64
 7   Session_Duration (hours)       973 non-null    float64
 8   Calories_Burned                973 non-null    float64
 9   Workout_Type                   973 non-null    object
 10  Fat_Percentage                 973 non-null    float64
 11  Water_Intake (liters)          973 non-null    float64
 12  Workout_Frequency (days/week)  973 non-null    int64
 13  Experience_Level               973 non-null    int64
 14  BMI                            973 non-null    float64
dtypes: float64(7), int64(6), object(2)
```

The dataset provides insights into gym members' physical characteristics and workout habits. This dataset contains **973 entries** with **no missing values** in any column, as indicated by the "Non-Null Count' for each feature.

[ Columns Overview ]

- Numerical Features:
  - Demographic: *Age*
  - Physiological: *Weight(kg)*, *Height(m)*, *BMI*, *BPM(Max, Average, Resting)*, *Fat Percentage*
  - Workout-Related: *Session Duration(hours)*, *Calories Burned*, *Water Intake(liters)*, *Workout Frequency(days/week)*, *Experience Level*
- Categorical Features: *Gender*, *Workout Type*

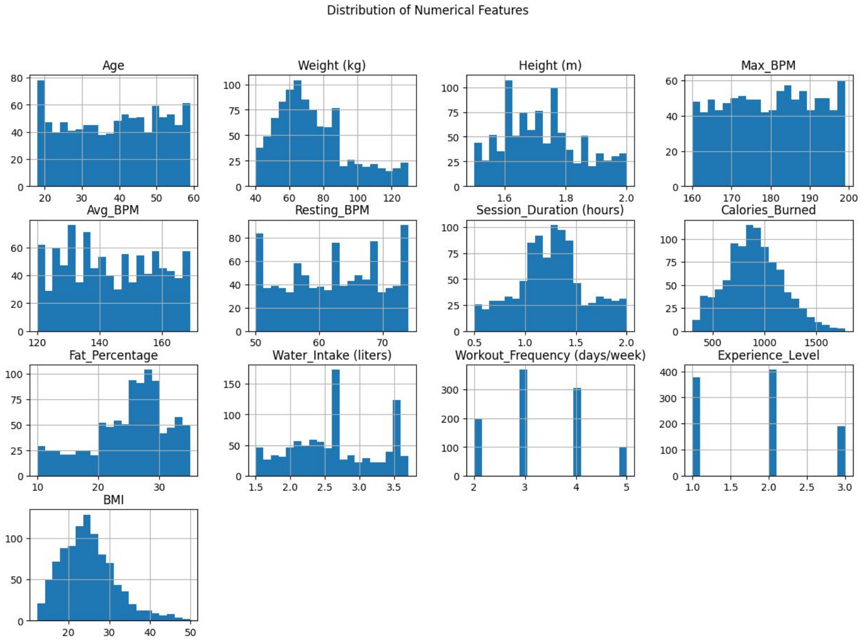
## ☑️ Outliers



- The bar chart illustrates the number of outliers identified for each feature using the **Interquartile Range (IQR)** method. Notably, *Weight(kg)*, *Calories_Burned*, and *BMI* have higher counts of outliers.
- To mitigate the impact of outliers in the dataset, **Robust Scaling** was applied. This method scales features based on the IQR, making it less sensitive to extreme values.

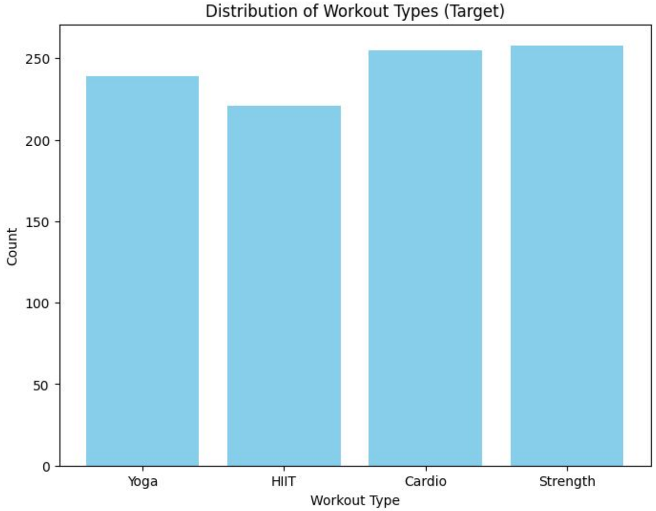
# Distribution Analysis

## ✅ Distribution of Numerical Features



- *Age*, *Weights(kg)*, *Calories_Burned*, and *BMI* have **varied scales**, which suggests that **standardization** could help align these features for modeling.
- *Calories_Burned* and *BMI* show **right-skewed distributions**, with outliers on the higher end. Applying **Robust Scaling** can help minimize the influence of these outliers by scaling based on the IQR.

## ✅ Target Variable Distribution: ***Workout Types***

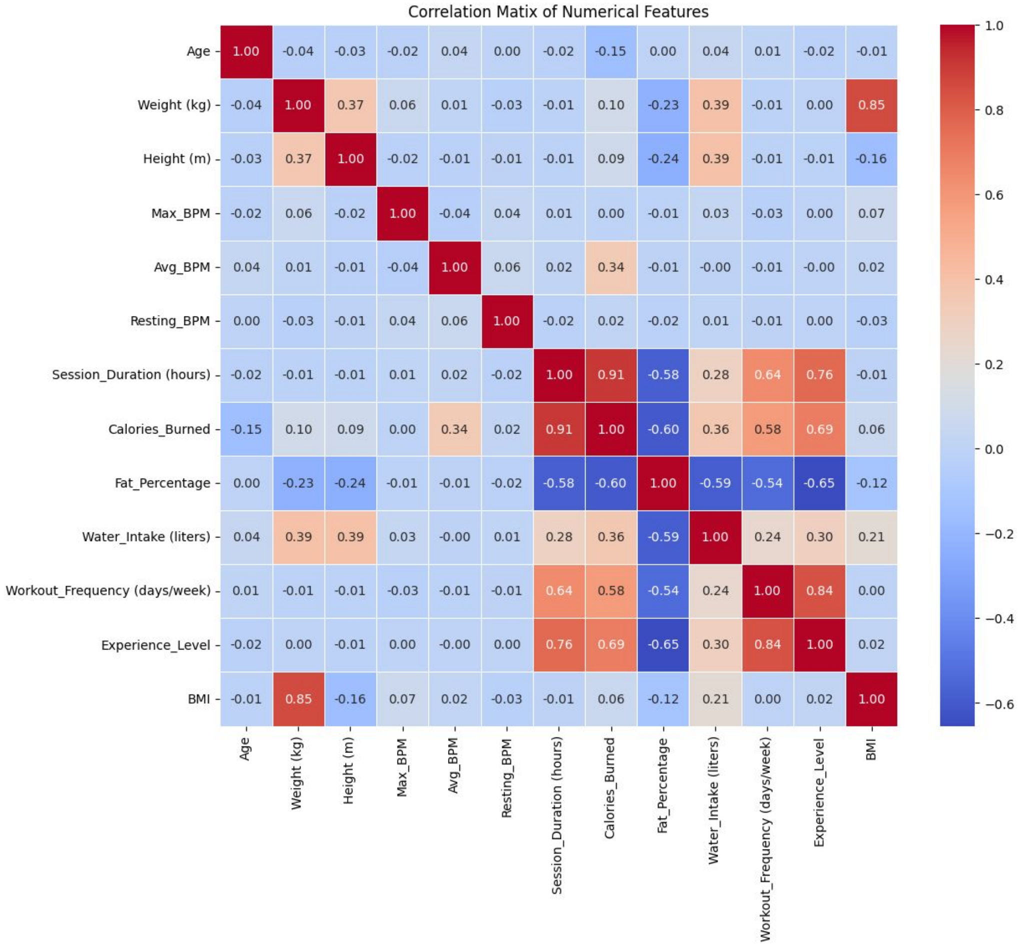


The bar chart illustrates the distribution of the target variable, *Workout_Type*, which is **evenly** distributed across four categories:

- Yoga: 239 members
- HIIT: 221 members
- Cardio: 255 members
- Strength: 258 members

The **near-equal distribution** across workout types indicates a balanced dataset, which reduces the risk of model bias toward a particular class.

# Correlation Analysis of Numerical Features



Correlation Matix of Numerical Features

| | Age | Weight (kg) | Height (m) | Max_BPM | Avg_BPM | Resting_BPM | Session_Duration (hours) | Calories_Burned | Fat_Percentage | Water_Intake (liters) | Workout_Frequency (days/week) | Experience_Level | BMI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | 1.00 | -0.04 | -0.03 | -0.02 | 0.04 | 0.00 | -0.02 | -0.15 | 0.00 | 0.04 | 0.01 | -0.02 | -0.01 |
| Weight (kg) | -0.04 | 1.00 | 0.37 | 0.06 | 0.01 | -0.03 | -0.01 | 0.10 | -0.23 | 0.39 | -0.01 | 0.00 | 0.85 |
| Height (m) | -0.03 | 0.37 | 1.00 | -0.02 | -0.01 | -0.01 | -0.01 | 0.09 | -0.24 | 0.39 | -0.01 | -0.01 | -0.16 |
| Max_BPM | -0.02 | 0.06 | -0.02 | 1.00 | -0.04 | 0.04 | 0.01 | 0.00 | -0.01 | 0.03 | -0.03 | 0.00 | 0.07 |
| Avg_BPM | 0.04 | 0.01 | -0.01 | -0.04 | 1.00 | 0.06 | 0.02 | 0.34 | -0.01 | -0.00 | -0.01 | -0.00 | 0.02 |
| Resting_BPM | 0.00 | -0.03 | -0.01 | 0.04 | 0.06 | 1.00 | -0.02 | 0.02 | -0.02 | 0.01 | -0.01 | 0.00 | -0.03 |
| Session_Duration (hours) | -0.02 | -0.01 | -0.01 | 0.01 | 0.02 | -0.02 | 1.00 | 0.91 | -0.58 | 0.28 | 0.64 | 0.76 | -0.01 |
| Calories_Burned | -0.15 | 0.10 | 0.09 | 0.00 | 0.34 | 0.02 | 0.91 | 1.00 | -0.60 | 0.36 | 0.58 | 0.69 | 0.06 |
| Fat_Percentage | 0.00 | -0.23 | -0.24 | -0.01 | -0.01 | -0.02 | -0.58 | -0.60 | 1.00 | -0.59 | -0.54 | -0.65 | -0.12 |
| Water_Intake (liters) | 0.04 | 0.39 | 0.39 | 0.03 | -0.00 | 0.01 | 0.28 | 0.36 | -0.59 | 1.00 | 0.24 | 0.30 | 0.21 |
| Workout_Frequency (days/week) | 0.01 | -0.01 | -0.01 | -0.03 | -0.01 | -0.01 | 0.64 | 0.58 | -0.54 | 0.24 | 1.00 | 0.84 | 0.00 |
| Experience_Level | -0.02 | 0.00 | -0.01 | 0.00 | -0.00 | 0.00 | 0.76 | 0.69 | -0.65 | 0.30 | 0.84 | 1.00 | 0.02 |
| BMI | -0.01 | 0.85 | -0.16 | 0.07 | 0.02 | -0.03 | -0.01 | 0.06 | -0.12 | 0.21 | 0.00 | 0.02 | 1.00 |

1. **High Positive Correlations**:
   - **[0.91]** *Session_Duration(hours)* **and** *Calories_Burned*: Indicates that longer workout sessions lead to more calories burned.
   - **[0.85]** *Weight(kg)* **and** *BMI*: A strong positive correlation, which is expected since BMI is partially derived from weight.
   - **[0.84]** *Workout_Frequency(days/week)* **and** *Experience_Level*: Shows that experienced individuals tend to work out more frequently.
2. **Moderate Positive Correlations**:
   - **[0.39]** *Water_Intake(liters)* **and** *Weight(kg)*: Heavier individuals may tend to drink more water, possibly due to greater hydration needs.
   - **[0.34]** *Calories_Burned* **and** *Avg_BPM*: Suggests higher average heart rates during workouts are associated with higher calorie burns, indicating more intense exercise.
3. **Negative Correlations**:
   - *Fat_Percentage* **with** *Experience_Level* **[-0.65]**, *Calories_burned* **[-0.60]**, and *Worout_Frequency(days/week)* **[-0.54]**: Implies that experienced individuals, those with higher calorie expenditure, or those who work out frequently tend to have lower fat percentages.
4. **Low or Near-Zero Correlations**:
   - *Age* **and** *Height(m)*: Shows minimal correlation with other features, suggesting they do not significantly influence other variables in this dataset.
   - *Max_BPM* and *Resting_BPM*: Shows low correlations with most features, suggesting these are more individual-specific metrics.

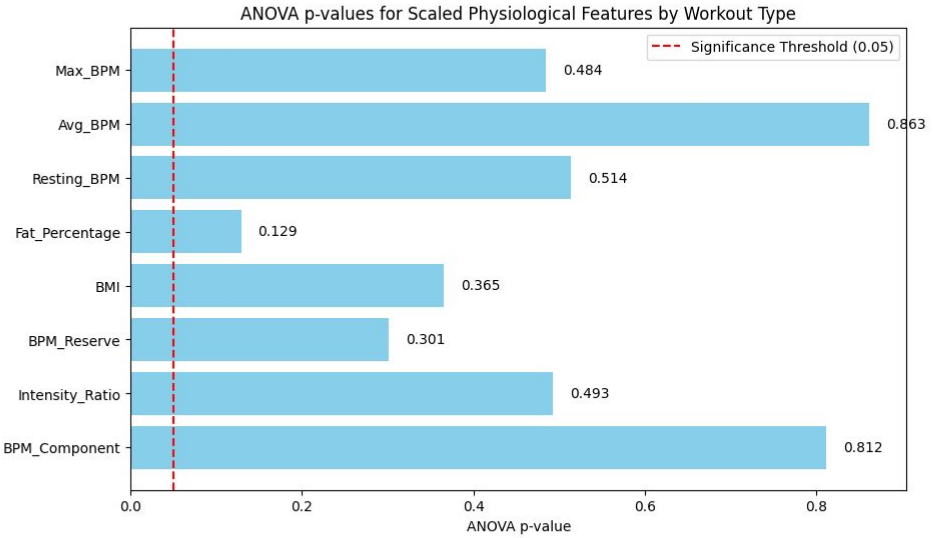
# Data Preprocessing & Insights - *Analysis of BPM, Fat Percentage, and BMI*

In this analysis, firstly, we focused on the three distinct BPM-related features: Max_BPM, Avg_BPM, and Resting_BPM. To gain a more comprehensive understanding of these heart rate metrics, we derived additional features that consolidate the highlight key aspects of these variables:

- ***BPM_Reserve****:* Calculated as ***Max_BPM - Resting_BPM***, this feature represents the difference between maximum heart rate during exercise and resting heart rate.
- ***Intensity_Ratio***: Defined as ***Avg_BPM / Max_BPM***, this feature indicates how close the average heart rate is to the maximum heart rate during exercise. A higher intensity Ratio suggests a more intense workout.
- **Principal Components Analysis (PCA)**: To retain the core information of all three BPM metrics while simplifying the analysis, we applied PCA, reducing these three variables into a single principal component.

After deriving these consolidated features, we applied **Standardization** to scale all BPM-related features, fat_percentages, and BMI to have mean 0 and a standard deviation of 1. This scaling step ensures that each feature contributes equally to the analysis without being influenced by differences in units or scales.

We conducted **an ANOVA test** on the scaled features to examine if there are statistically significant differences across workout types. The resulting ANOVA p-values provide insight into which features differ meaningfully between workout types, guiding us toward a deeper understanding of how physiological metrics related to exercise preferences.

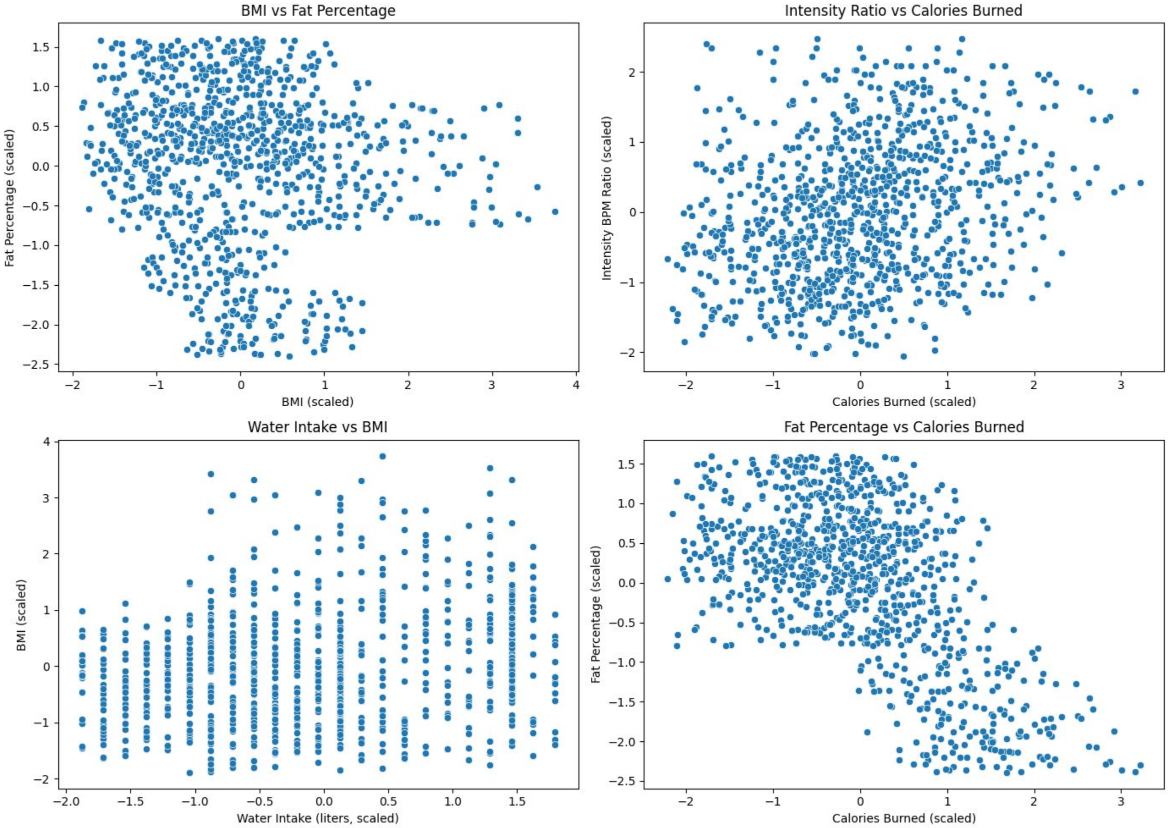


As shown in the **ANOVA p-values**, none of the original or derived physiological features show significant difference across workout types, as all p-values are above the 0.05 threshold. These results suggest that heart rate and body composition metrics, including BPM metrics, body fat percentage, and BMI, are **largely independent of workout type**.

Since none of these features show significant variation across workout types, they may have limited usefulness for predicting workout type preferences based on these physiological metrics alone. Therefore, **exploring other feature categories together** may yield more meaningful predictors.

# Exploring Additional Feature Relationships

## ☑ *BPM*(Max, Average, Resting), *Fat_Percentage*, and *BMI*


BMI vs Fat Percentage

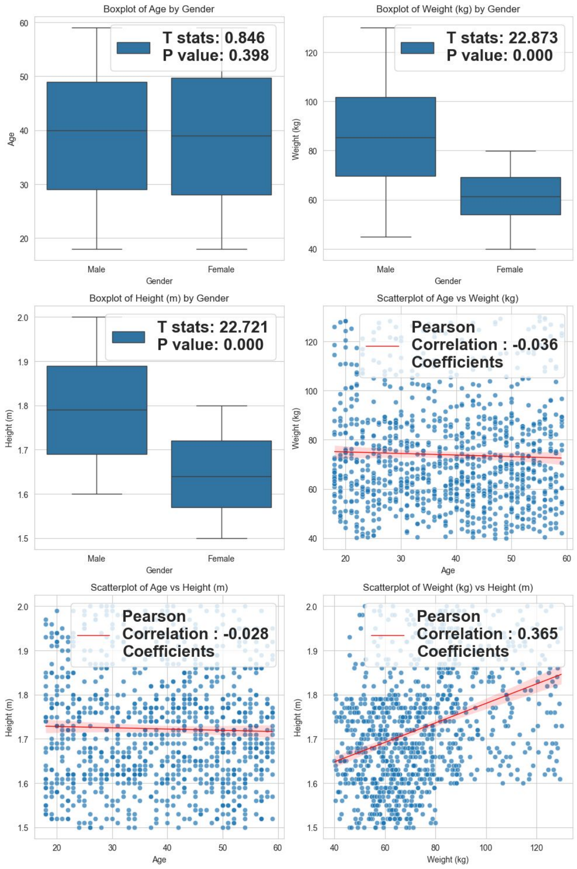
Intensity Ratio vs Calories Burned

Water Intake vs BMI

Fat Percentage vs Calories Burned

- ***BMI*** vs ***Fat_Percentage***: This plot reveals a positive correlation between *BMI* and *fat percentages*, as both metrics are related to body composition. Generally, higher *BMI* corresponds with a higher *fat percentage*. However, there are some variations, indicating that individuals with similar BMIs may have different fat composition, likely influenced by factors such as muscle mass.
- ***Intensity_Ratio*** vs ***Calories_Burned***: The *intensity ratio*, defined as *average BPM* relative to *maximum BPM*, shows no strong correlation with calories burned. This suggests that while *workout intensity* might impact *calorie burn* to some extent, it is not the sole determining factor. Other factors, such as *workout duration* and *type*, likely play significant roles in overall *calorie expenditure*.
- ***Water_Intake*** vs ***BMI***: There is no clear trend or strong correlation between *water intake* and *BMI*, suggesting that *BMI* does not directly relate to *water intake*. This also implies that personal hydration preferences that are not directly tied to *BMI*.
- ***Fat_Percentage*** vs ***Calories_Burned***: There is a slight inverse trend, where individuals with lower *fat percentages* appear to *burn more calories*. This may indicate that individuals with lower *body fat percentages* are more active or engage in more intense exercises, resulting in higher calorie burns.

# Bivariate Analysis Between *Demographic Features* and *Calories Burned*



**Independent t-test** between Gender and other 3 numeric demographic features:

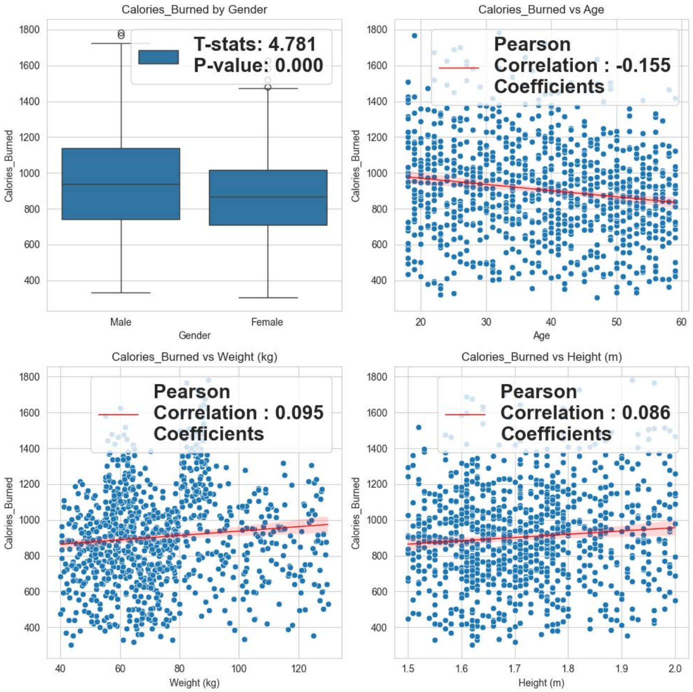
- Age: no significant difference in age between genders.
- Weight: significant difference in weight between genders, with males having higher weight distributions.
- Height: significant difference in height between genders, with males being generally taller.

**Pearson Correlation Coefficient** among the 3 numeric demographic features:

- Age vs Weight: Very weak negative correlation, no meaningful relationship.
- Age vs Height: Very weak negative correlation, no relationship with height.
- Weight vs Height: Moderate positive correlation, indicating taller individuals tend to weigh more.

**Pearson Correlation Coefficient** between Age, Height, Weight, and the target variable Calories_Burned

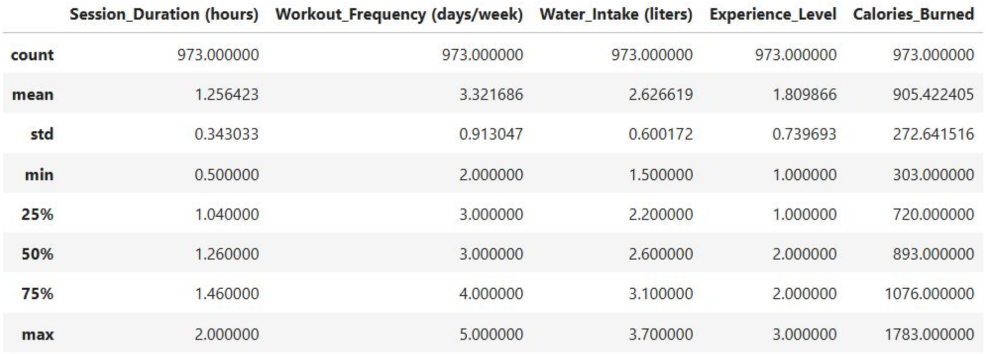
- Calories Burned vs Age: Weak negative correlation (-0.155), suggesting a slight decrease in calories burned with age
- Calories Burned vs Weight: Very weak positive correlation (0.095), indicating a minimal relationship.
- Calories Burned vs Height: Very weak positive correlation (0.086), indicating a minimal relationship.



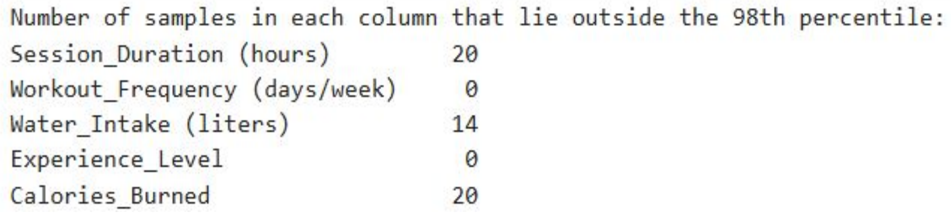
**Independent t-test** between Gender and target variable Calories Burned

- The t-test shows a significant difference in calories burned between genders, with males generally burning more calories than females.

# Descriptive Statistics and Outlier Analysis

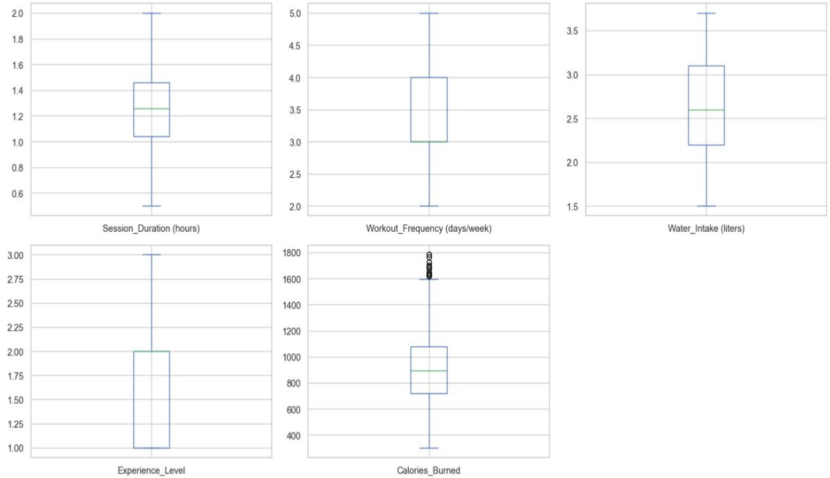


| | Session_Duration (hours) | Workout_Frequency (days/week) | Water_Intake (liters) | Experience_Level | Calories_Burned |
|---|---|---|---|---|---|
| **count** | 973.000000 | 973.000000 | 973.000000 | 973.000000 | 973.000000 |
| **mean** | 1.256423 | 3.321686 | 2.626619 | 1.809866 | 905.422405 |
| **std** | 0.343033 | 0.913047 | 0.600172 | 0.739693 | 272.641516 |
| **min** | 0.500000 | 2.000000 | 1.500000 | 1.000000 | 303.000000 |
| **25%** | 1.040000 | 3.000000 | 2.200000 | 1.000000 | 720.000000 |
| **50%** | 1.260000 | 3.000000 | 2.600000 | 2.000000 | 893.000000 |
| **75%** | 1.460000 | 4.000000 | 3.100000 | 2.000000 | 1076.000000 |
| **max** | 2.000000 | 5.000000 | 3.700000 | 3.000000 | 1783.000000 |



```
Number of samples in each column that lie outside the 98th percentile:
Session_Duration (hours)         20
Workout_Frequency (days/week)     0
Water_Intake (liters)            14
Experience_Level                  0
Calories_Burned                  20
```

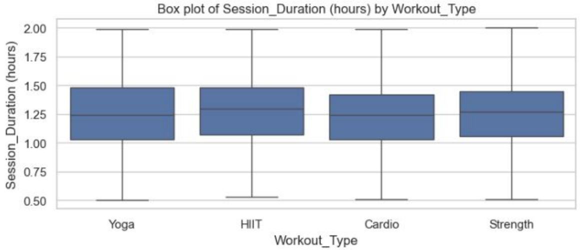
- The table summarizes the key statistics (mean, standard deviation, percentiles) for the columns, indicating *Session_Duration*, *Workout_Frequency*, *Water_Intake*, *Expereience_Level*, and *Calories_Burned*. These Statistics provide a high-level view of the data distribution within each feature, which may help identify potential patterns or abnormalities.

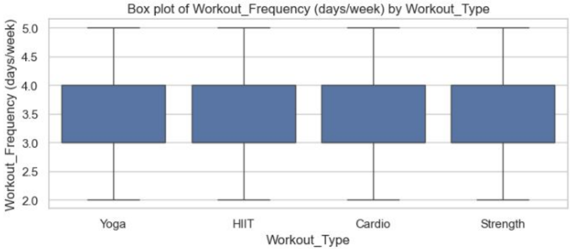


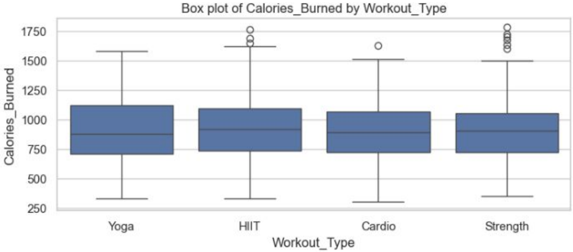
- The box plots display the distribution of each feature, highlighting possible outliers. Features such as Calores_Burned show a few extreme values.
- Approximately 100 samples lie outside the 98th percentile in one or more columns. Removing these would result in significant data loss, so future examination of these high values may reveal patterns or insights.

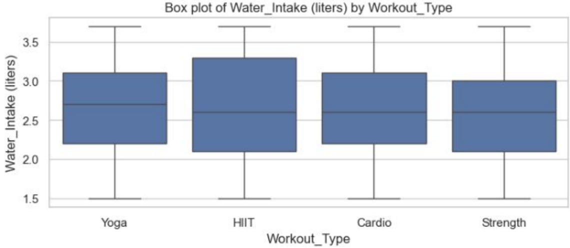
# Spearman Correlation and Distribution Insights

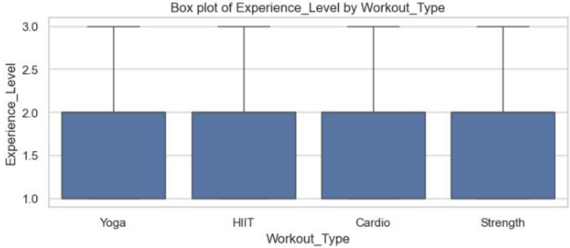
**Spearman's correlation** measures the **rank-order** association between two variables, where +1 indicates a perfect positive monotonic relationship, -1 indicates a perfect negative monotonic relationship, and 0 means little to no monotonic relationship.

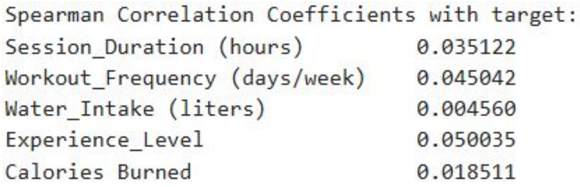

Box plot of Session_Duration (hours) by Workout_Type


Box plot of Workout_Frequency (days/week) by Workout_Type

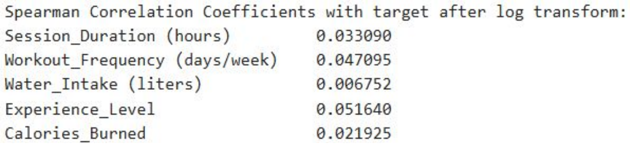

Box plot of Calories_Burned by Workout_Type


Box plot of Water_Intake (liters) by Workout_Type

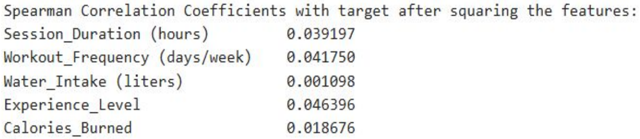

Box plot of Experience_Level by Workout_Type



```
Spearman Correlation Coefficients with target:
Session_Duration (hours)          0.035122
Workout_Frequency (days/week)     0.045042
Water_Intake (liters)             0.004560
Experience_Level                  0.050035
Calories_Burned                   0.018511
```



```
Spearman Correlation Coefficients with target after log transform:
Session_Duration (hours)         0.033090
Workout_Frequency (days/week)    0.047095
Water_Intake (liters)            0.006752
Experience_Level                 0.051640
Calories_Burned                  0.021925
```



```
Spearman Correlation Coefficients with target after squaring the features:
Session_Duration (hours)         0.039197
Workout_Frequency (days/week)    0.041750
Water_Intake (liters)            0.001098
Experience_Level                 0.046396
Calories_Burned                  0.018676
```

- **Box Plot Distributions**: The plots show similar distributions of features across different *workout types*, which visually supports the interpretation that these features do not vary significantly with *Workout_Type*.
- **Spearman Correlation Results**: The low Spearman coefficients (close to 0) for each feature with respect to *Workout_Type* indicate that there is little to no monotonic relationship between these features and the target variable. Even after applying transformations (logarithmic and squared), the correlations remain low, suggesting that these features alone might lack predictive power for differentiating workout types.
- Given the low correlation, **exploring non-linear relationships** or **examining combinations of features** may yield better predictive insights.

# Machine Learning Techniques

**Initial exploration** – Based on the preprocessing techniques suggested, we applied scaling using robust scaling and standard scaling, encoding using one-hot encoding for *gender*, and label encoding for *workout_type*. After preprocessing, we trained vanilla versions of the following models to get a rough understanding of their performance: **Logistic Regression, Decision Trees, Random Forest, XGBoost,Naive Bayes, KNN, and 5-layered Neural Nets**. We trained these baseline models setting the target variable as **Workout_Type**. However, our preliminary findings indicated minimal correlation between the features and the target, resulting in poor performance and generalization. This led us to reconsider our initial problem statement.

**Further analyses** – Our initial investigation led us to consider other target variables and problem statements. We explored three different target variables,which demonstrated some kind of relation with the other features. These target variables are as follows: **regression on Calories_Burned, multi-class classification on Expereince_Level and regression on Session Duration (hours)**. Following all the data-preprocessing steps and model trainings and evaluations, these three labels showed some strong correlation with rest of the features and provided better training and evaluation results.

**Next steps** – Moving forward, we plan to leverage the strengths of our initial findings, incorporate more sophisticated feature engineering, and try to build on a more nuanced problem statement by improving our dataset or setting a target variable that can show some strong relation with rest of the features. This optimization would help us to address more complex fitness-related questions and improve the model's generalizability.

**Performance Metrics of Different ML Models (Target: Workout_Type)**

| Model | Accuracy | Precision (Avg) | Recall(Avg) |
|---|---|---|---|
| Logistic Regression | 0.24 | 0.24 | 0.25 |
| Decision Tree | 0.33 | 0.32 | 0.32 |
| Random Forest | 0.26 | 0.26 | 0.27 |
| KNN | 0.27 | 0.27 | 0.28 |
| XGBoost | 0.24 | 0.24 | 0.24 |
| Naive Bayes | 0.22 | 0.22 | 0.23 |
| 5-Layered Neural Net | 0.33 | 0.29 | 0.31 |

**Performance with different Target Variables**

| Target | Performance (Random Forest Regressor) |
|---|---|
| Session Duration | R squared score: 0.9576 |
| Experience Level | Accuracy : 0.91 |
| Calories Burned | R squared score: 0.9635 |