Assessing gym members fitness level based gym experience

André Eiras, Diogo Cruz

2025-03-23

Contents

Introduction **2**
Exploratory Data Analysis 2

Methods **3**
Principal Components Analysis 3
Clustering 4

Results and Discussion **7**

References **12**

Annex **13**
Exploratory univarate data analysis 13
Exploring correlations 13
Principal components analysis 13

Introduction

Real-world problems are often complex and influenced by a wide range of factors. To effectively address such challenges, we must rely on statistical learning methods. These methods are typically categorized into supervised and unsupervised learning. Supervised learning methods are applied in tasks involving classification and prediction, where the objective is to estimate the value of a target variable based on known input variables. In contrast, unsupervised learning methods are used for clustering, association rule mining, and dimensionality reduction, aiming to discover hidden patterns or structures in data without a predefined target variable. While both approaches seek to identify structure within data, they differ in whether or not a response variable guides the analysis.

In this project, our objective is to identify fitness patterns and performance trends among gym members with varying levels of experience, using the Gym Members Exercise Dataset. The dataset comprises 15 variables and 973 observations. Table 1 provides a detailed description of each variable and its corresponding meaning. The overall aim of this analysis is to examine how experience level affects exercise performance, and to highlight key differences between gym members based on their level of training or familiarity with gym routines.

Table 1: Gym members dataset variable meaning.

Variables	Description
Age	Age of the gym member
Gender	Gender of gym member (binary)
Weight..kg.	Member's weight (kg)
Height..m.	Member's height (m)
Max_BPM	Maximum heart rate during workout sessions (bpm)
Avg_BPM	Average heart rate during workout sessions (bpm)
Resting_BPM	Heart rate at rest before workout (bpm)
Session_Duration..hours.	Duration of each workout sessions (hours)
Calories_Burned	Total calories burned during each session
Workout_Type	Type of workout performed (factors)
Fat_Percentage	Body fat percentage of the member
Water_Intake..liters.	Daily water intake during workouts
Workout_Frequency..days.week.	Number of workout session per week
Experience_Level	Experience level: 1-Begginer, 2-Intermediate, 3-Expert (factors)
BMI	Body Mass Index, calculated from height and weight

First, we begin by examining how each of the fitness metrics and demographic variables relates to the members' experience level, in order to establish a baseline understanding of the dataset. Following this initial exploration, we apply Principal Component Analysis (PCA) to reduce dimensionality and identify the most relevant features. Finally, we perform clustering analysis to uncover underlying patterns and groupings within the data based on shared characteristics.

Exploratory Data Analysis

To ensure reliable results when performing Principal Component Analysis (PCA), we first verified that the dataset contained no missing values. As shown in Figure 5, this condition is met. Table 2 presents descriptive statistics for the numeric variables in the Gym Members dataset. On average, a gym member is a 39-year-old male, standing 1.72 meters tall, weighing approximately 74 kg, with 25% body fat and a Body Mass Index (BMI) of 25. During workouts, the average heart rate is 144 beats per minute (bpm), resting heart rate is 62 bpm, and the maximum heart rate reaches an average of 180 bpm. Workout sessions typically last around 1 hour and 16 minutes, distributed over three sessions per week. On average, members consume 2.6 liters of water per workout session.

Additionally, there are no significant age differences between male and female members across the three experience levels. However, expert male members tend to be younger than their female counterparts, and the same pattern is observed among female intermediate members. Finally, members who train more frequently per week also tend to have longer workout sessions, as illustrated in Figure 7.

Table 2: Descriptive statistics for the numeric variables of the dataset.

Variable	Min	Mean	Median	SD	IQR	Max
Age	18.00	38.683453	40.00	12.1809279	21.00	59.00
Weight..kg.	40.00	73.854676	70.00	21.2075005	27.90	129.90
Height..m.	1.50	1.722580	1.71	0.1277199	0.18	2.00
Max_BPM	160.00	179.883864	180.00	11.5256860	20.00	199.00
Avg_BPM	120.00	143.766701	143.00	14.3451014	25.00	169.00
Resting_BPM	50.00	62.223022	62.00	7.3270599	12.00	74.00
Session_Duration..hours.	0.50	1.256423	1.26	0.3430335	0.42	2.00
Calories_Burned	303.00	905.422405	893.00	272.6415165	356.00	1783.00
Fat_Percentage	10.00	24.976773	26.20	6.2594188	8.00	35.00
Water_Intake..liters.	1.50	2.626619	2.60	0.6001719	0.90	3.70
Workout_Frequency..days.week.	2.00	3.321686	3.00	0.9130470	1.00	5.00
BMI	12.32	24.912127	24.16	6.6608794	8.45	49.84

Based on the probabilistic distribution of various variables shown in Figures 8, 11, 12, and 13, it appears that age and the different heart rate measures among gym members follow an approximately uniform distribution. In contrast, the remaining variables display reasonably symmetric distributions. Specifically, weight, height, and body mass index (BMI) are slightly right-skewed, while body fat percentage is slightly left-skewed. These distributional characteristics can be further assessed through the box plot and histogram combinations presented in Figures 8–19 in the Annex section of this report.

From Figure 19, we observe that the strongest positive correlation occurs between calories burned per session and session duration—a logical relationship, as longer sessions tend to result in higher energy expenditure. Additionally, workout frequency, as previously illustrated in Figure 7, also shows a strong positive correlation with session duration. Conversely, body fat percentage is strongly negatively correlated with several key variables: calories burned per session, water intake, session duration, and workout frequency. These correlations suggest that higher levels of body fat are associated with less frequent and shorter gym sessions.

Methods

Having established that there are no missing values in the dataset and having identified some preliminary relationships between variables, we proceed with a Principal Component Analysis (PCA) followed by clustering analysis using the k-means algorithm. Given that the variables are measured on different scales, we begin the analysis by normalizing the data to ensure that all features contribute equally to the results.

Principal Components Analysis

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of a dataset while preserving as much of its variance as possible. It is a form of matrix factorization aimed at simplifying complex data. PCA identifies a sequence of optimal linear projections that approximate a multivariate dataset by mapping it onto lower-dimensional subspaces. This method provides valuable insights into the structure, variance, and intrinsic dimensionality of the data, making it particularly useful in unsupervised learning. The goal is to identify a sequence of affine hyperplanes (subspaces) that best approximate the data in a least-squares sense.

To perform PCA, we center the data and decompose the data matrix using singular value decomposition (SVD): $\mathbf{X} = UDV^T$, where : U - contains the left singular vectors representing the observations in principal component (PC) space; D - Diagonal matrix of singular values; V - Right singular vectors (principal component directions). The principal components are the rows of $\mathbf{UD}$, the principal axes are the columns of $\mathbf{V}$, the eigenvalues of the covariance matrix $\mathbf{X^T X}$ are $\lambda_i = d_i^2$,and represent the variance explained by each component.The first PC direction v_1 maximizes variance:

$$v_1 = \arg \max_{v:||v||=1} Var(\mathbf{X}v)$$

Subsequent PCs are orthogonal to previous ones and maximize the remaining variance.

The following table summarizes the results of the PCA. The first principal component explains 28% of the total variance, the second explains 17%, and the third explains 11%. Combined, the first three principal components account for 56% of the total variance in the dataset.

Table 3: PCA synthesis

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	1.82	1.42	1.13	1.05	1.02	1.01	0.96	0.73	0.66	0.55	0.12	0.08
Proportion of Variance	0.28	0.17	0.11	0.09	0.09	0.09	0.08	0.04	0.04	0.03	0.00	0.00
Cumulative Proportion	0.28	0.44	0.55	0.64	0.73	0.82	0.89	0.94	0.97	1.00	1.00	1.00

As shown in Figure 1, the explained variance drops significantly after the third principal component. This indicates that the first three components capture the most meaningful variation in the dataset. Therefore, we can conclude that the first three principal components are the most relevant for summarizing the structure of the data and are sufficient for further analysis and dimensionality reduction.

Table 3 presents the three most influential variables for each of the first three principal components. The first principal component is primarily driven by calories burned, body fat percentage, and workout duration. The second component is most influenced by weight, body mass index (BMI), and again workout duration, indicating shared relevance across components. The third component is shaped mainly by height, BMI, and daily water intake. For a detailed breakdown of each variable's loading on the principal components, please refer to Table 5 in the Annex section. A visual representation of variable contributions is also available in Figure 20.

Table 4: The three most influential variables of each Principal Components

PC1	PC2	PC3
Calories_Burned	Weight..kg.	Height..m.
Fat_Percentage	BMI	BMI
Session_Duration..hours.	Session_Duration..hours.	Water_Intake..liters.

Clustering

The goal of clustering analysis is to partition a set of observations into groups (clusters) such that intra-cluster similarity is high and inter-cluster similarity is low—without relying on outcome labels. The concepts of similarity and dissimilarity are central to clustering, as they determine how observations are compared. Most clustering algorithms are based on a dissimilarity matrix, which quantifies the pairwise differences between observations. A common approach to assess dissimilarity is through the computation of metrics such as Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_k (x_{ik}, x_{jk})^2}$$

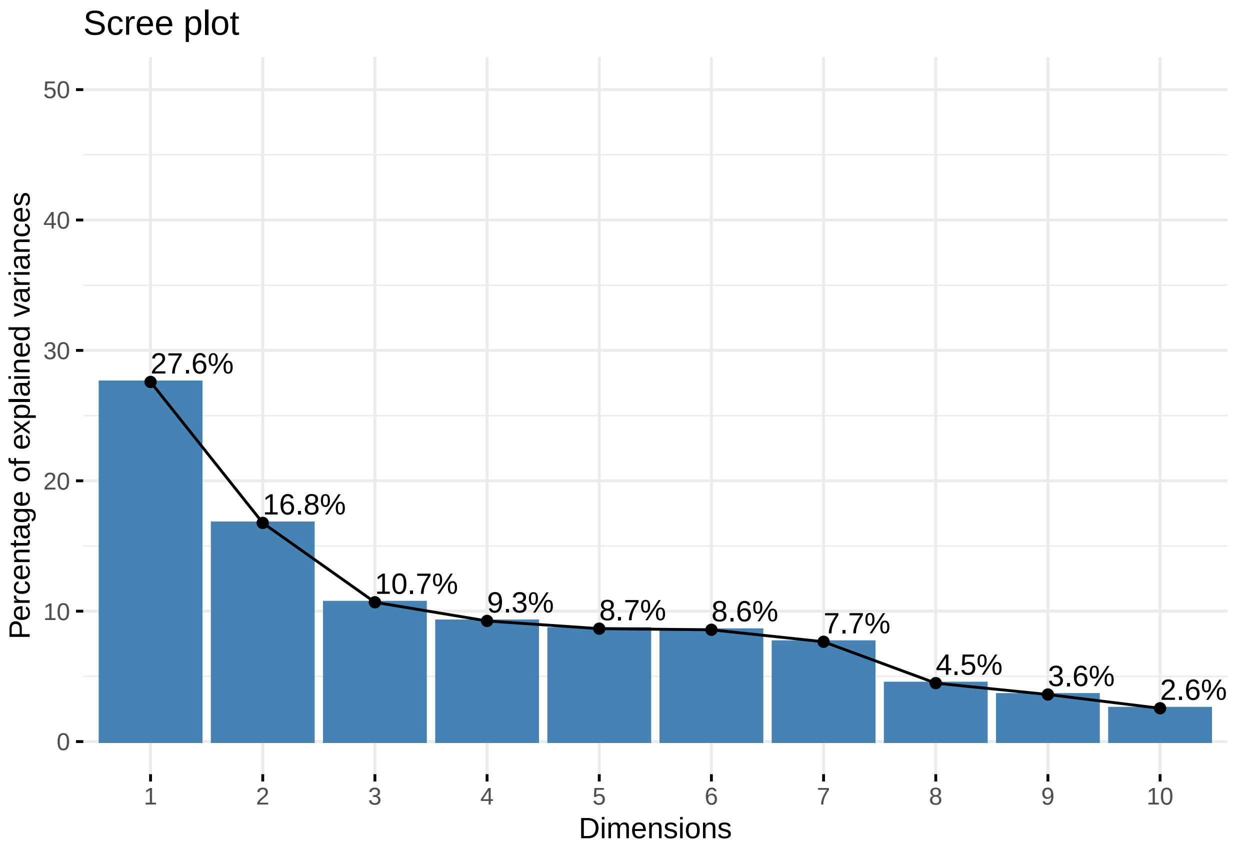


Figure 1: Scree plot - Principal components aggregated explained variance

The choice of distance metric significantly affects the shape of the resulting clusters. For instance, Euclidean distance—commonly used in clustering—tends to favor spherical clusters, as in the k-means algorithm. Both similarity and dissimilarity indices influence the clustering structure and the interpretation of results.

K-means is a widely used and conceptually straightforward partitioning algorithm that aims to minimize the within-cluster sum of squares through an iterative process. The algorithm follows these steps:

1. Select the number of clusters K;
2. Initialize K centroids;
3. Assign each observation to the nearest centroid;
4. Recompute centroids as the mean of assigned points;
5. Repeat steps 3–4 until convergence.

The quality of the clustering can be evaluated using internal validation indices. To determine the optimal number of clusters, we applied the elbow method, which indicated that three clusters would be appropriate for this analysis (see Figure 2). To further assess differences between clusters, we conducted an ANOVA, followed by a Tukey post-hoc test to identify statistically significant differences across variables. Results of the Tukey tests for each significant variable are available in the Annex section of this report.

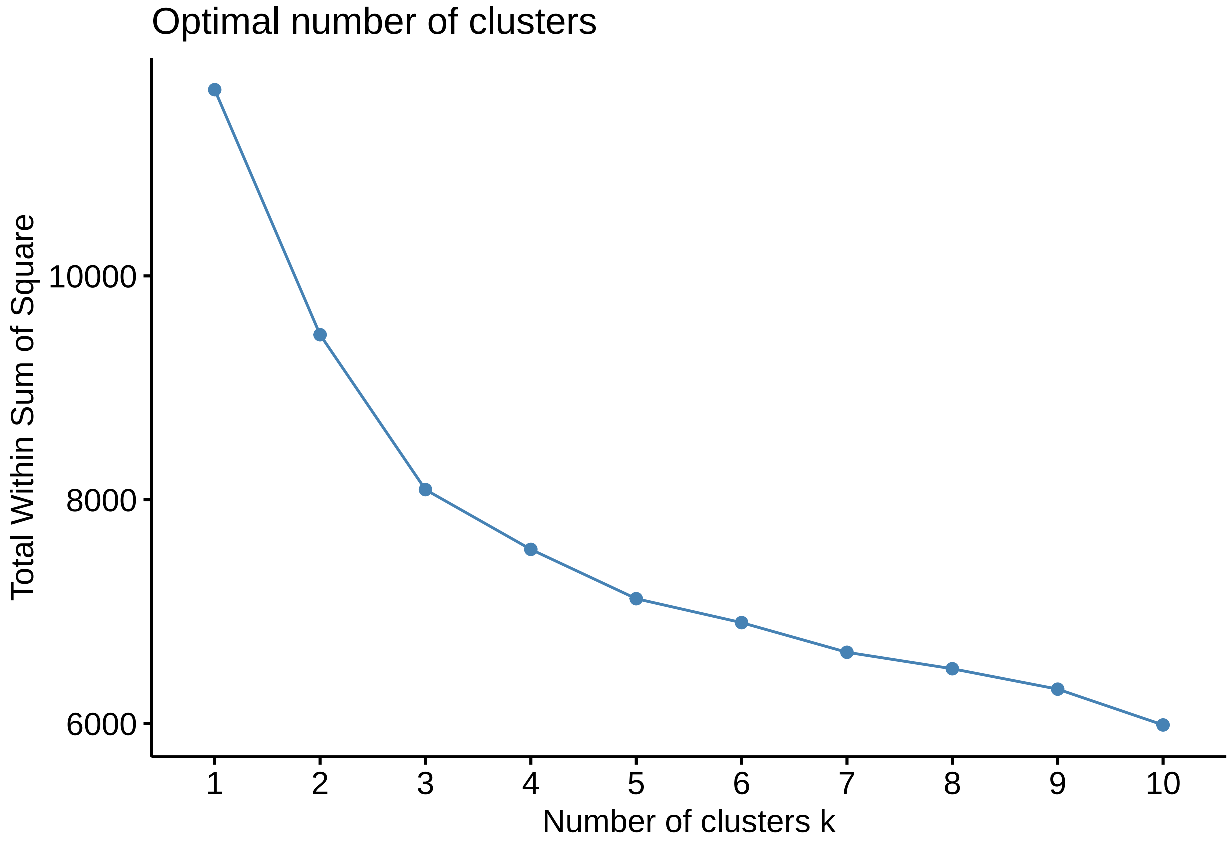


Figure 2: Assessing the optimal number of clusters from the total within sum of squares

Mixture models extend traditional clustering approaches by assuming that the data is generated from a probabilistic mixture of distributions, where each distribution corresponds to a distinct cluster. Given that the overall distribution of our data is reasonably symmetric, we also applied a Gaussian Mixture Model (GMM) as a probabilistic clustering method in our analysis. Gaussian mixture models assumes each data point x_i arises from:

$$p(x_i) = \sum_{k=1}^{K} \pi_k.N(x_i|\mu_k, \sigma_k)$$

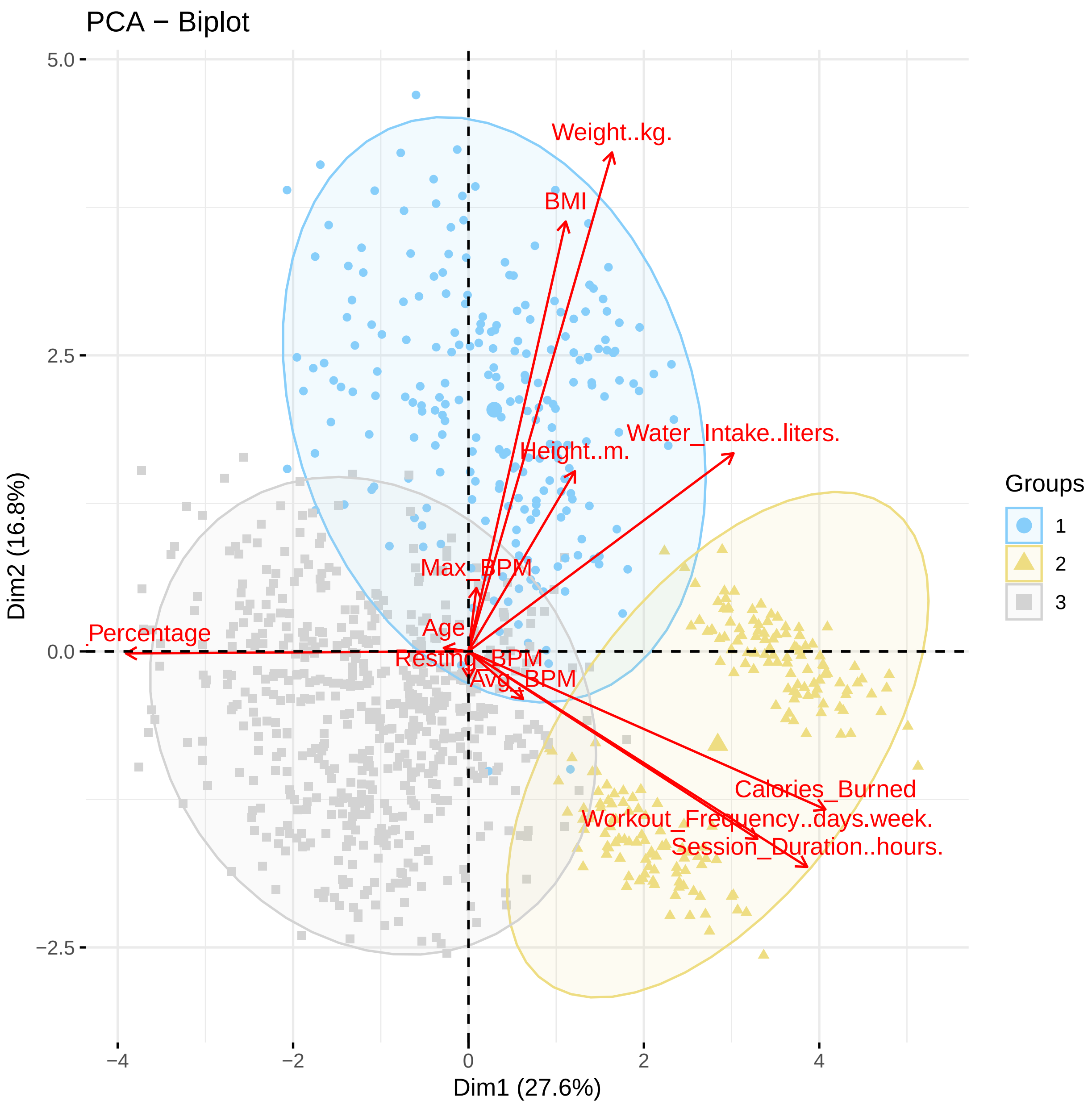
where π_k are mixing proportions, and N is a multivariate Gaussian. The k-means algorithm can be viewed as a limiting case of Gaussian Mixture Models (GMMs), specifically when all covariance matrices are equal and spherical, i.e., $\sigma_k = \sigma^2 I$ and hard cluster assignments are applied. In this project, we also applied GMM-based clustering to compare its results with those obtained from k-means. For a detailed summary of the GMM model and its output, please refer to the Annex section of this report.

Results and Discussion

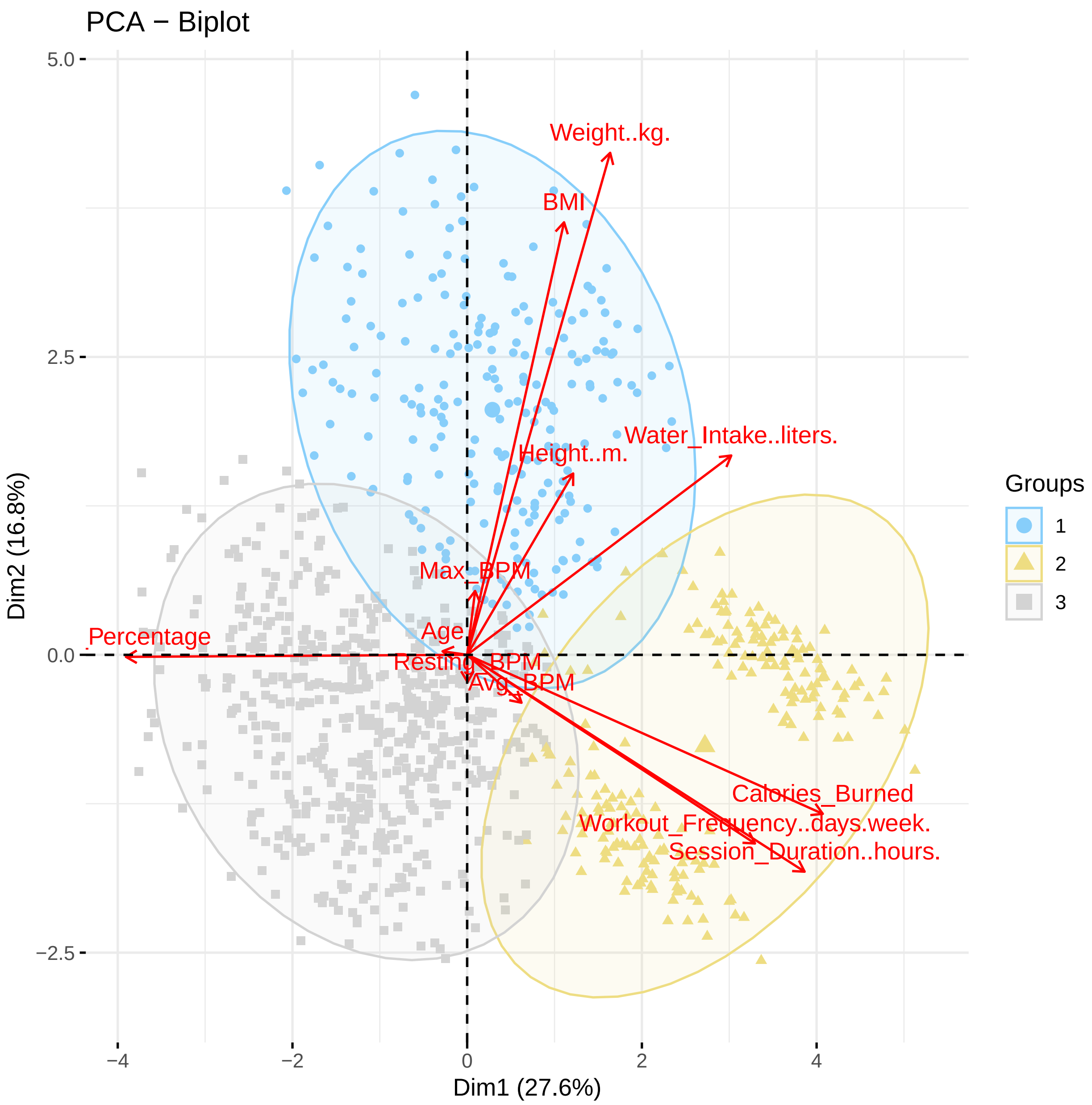
By conducting an ANOVA followed by Tukey post-hoc tests, we identified the variables that are significantly associated with clustering. The following variables showed statistically significant differences:

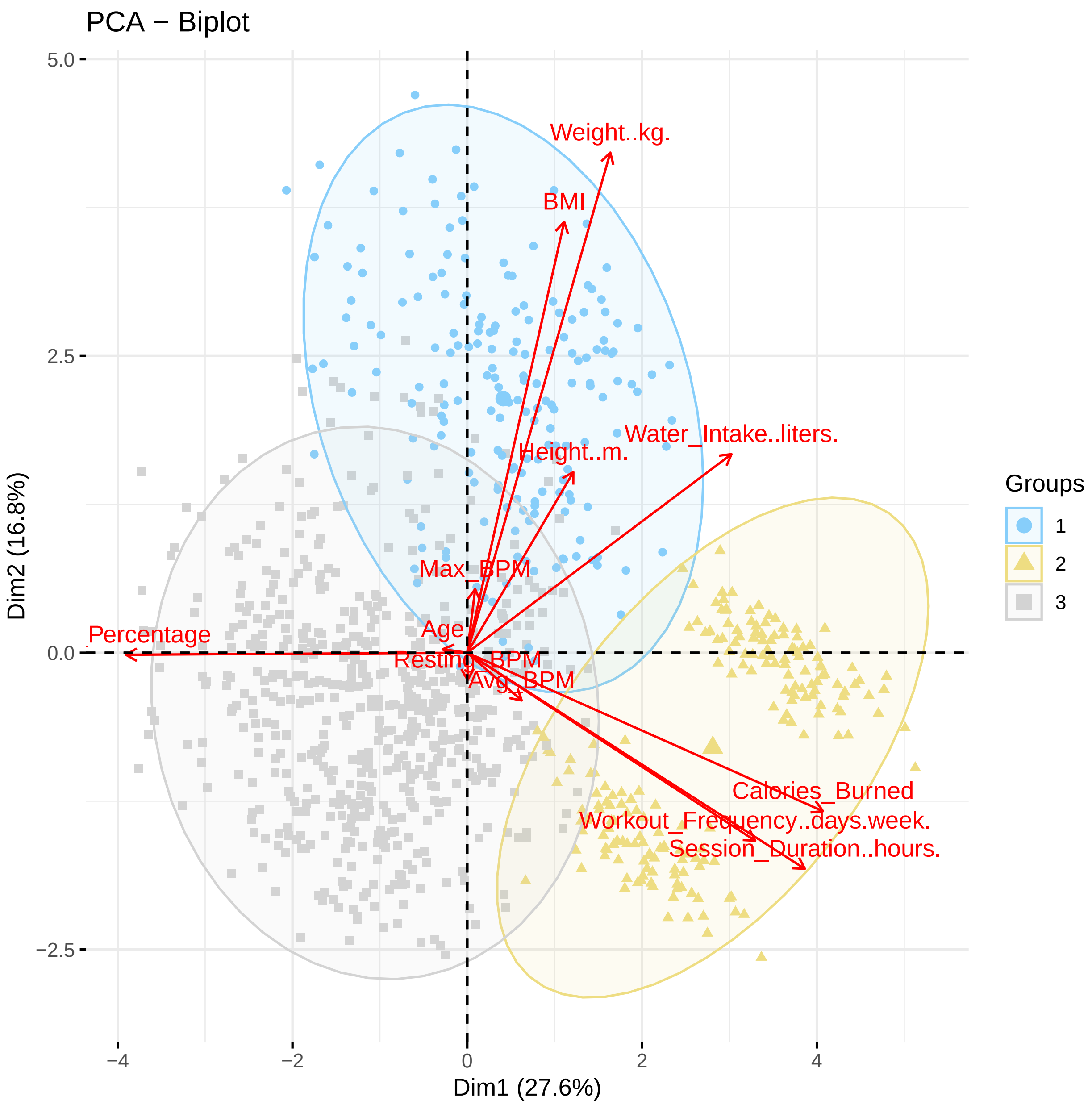
- Weight: significantly different across all three clusters.
- Height: significantly different across all three clusters.
- Session Duration: significantly different between clusters 1 and 2, and clusters 1 and 3, with no statistical evidence of a difference between clusters 2 and 3.
- Calories Burned: significantly different across all three clusters.
- Fat Percentage: significantly different across all three clusters.
- Water Intake: significantly different across all three clusters.
- Workout Frequency: significantly different between clusters 1 and 2, and clusters 1 and 3, with no significant difference between clusters 2 and 3.
- BMI: significantly different across all three clusters.

In terms of interpretation, Cluster 1 is composed of heavier, taller individuals who tend to have a higher BMI and consume more water per workout. Cluster 2 includes the majority of gym members, and is mainly associated with older individuals with a higher body fat percentage. Cluster 3 represents a high-performance group who work out more frequently and for longer duration, leading to greater calorie expenditure. For a visual representation of these relationships, refer to the biplot in Figure 3, Figure 4 and Figure 5. Figure 6 displays a dendrogram from the hierarchical clustering.



PCA − Biplot
Dim2 (16.8%)
Dim1 (27.6%)
5.0
2.5
0.0
−2.5
−4
−2
0
2
4
Weight..kg.
BMI
Water_Intake..liters.
Height..m.
Max_BPM
Age
Percentage
Resting_BPM
Avg_BPM
Calories_Burned
Workout_Frequency..days.week.
Session_Duration..hours.
Groups
1
2
3


PCA − Biplot
5.0
2.5
0.0
−2.5
−4
−2
0
2
4
Dim1 (27.6%)
Dim2 (16.8%)
Weight..kg.
BMI
Height..m.
Water_Intake..liters.
Max_BPM
Age
Percentage
Resting_BPM
Avg_BPM
Calories_Burned
Workout_Frequency..days.week.
Session_Duration..hours.
Groups
1
2
3



```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

To evaluate the quality of both clustering solutions, we used silhouette indices, with the corresponding plots shown in Figure 3. For a more detailed comparison of cluster differences across variables in both methods, refer to Figures 21–28 in the Annex section of this report. In the k-means clustering, the most cohesive structure was observed in Cluster 2, whereas Clusters 1 and 3 exhibited flatter or declining silhouette values, indicating less clearly defined boundaries between clusters. Cluster 2 showed the weakest performance in terms of cohesion.

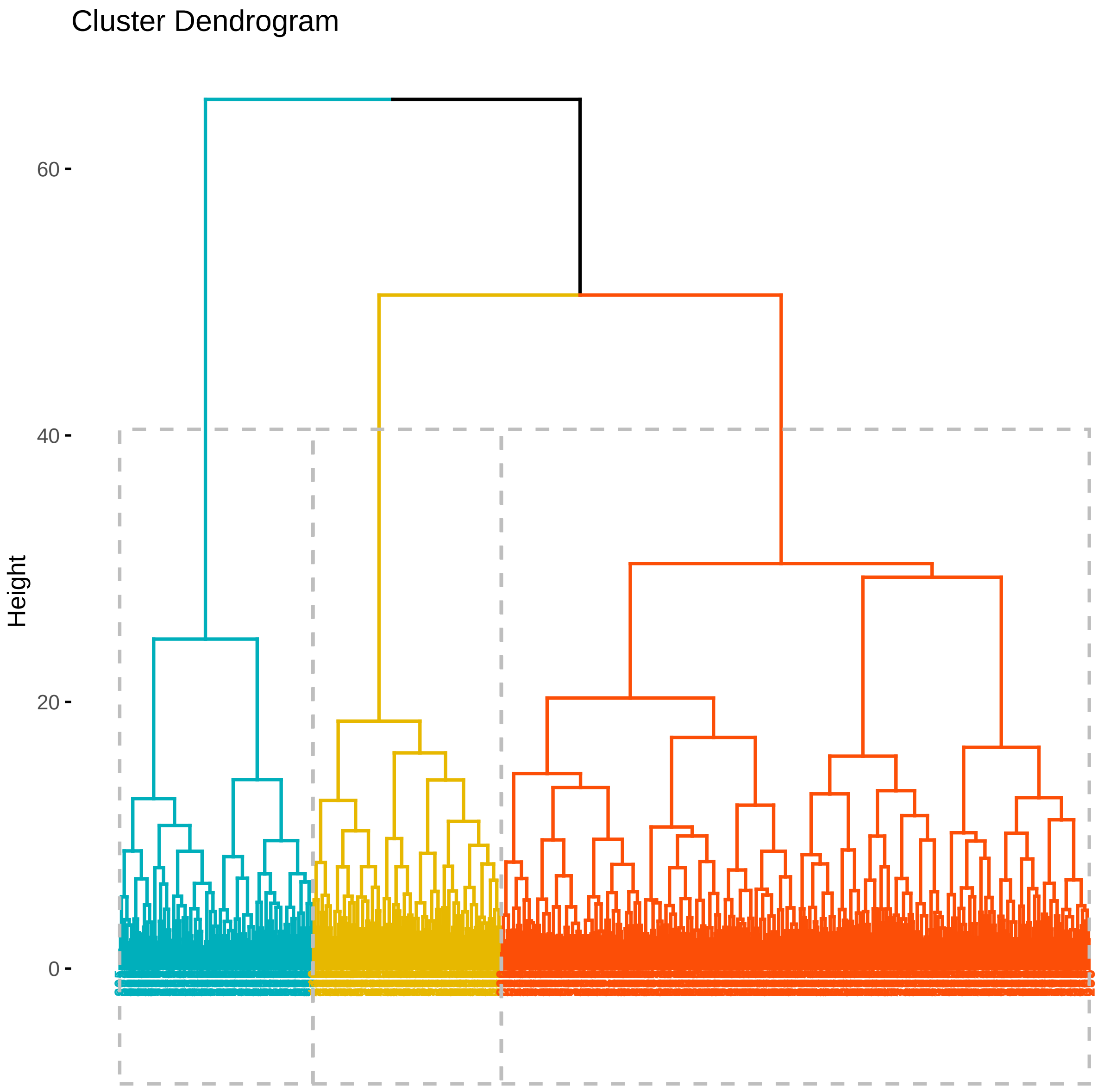


Figure 3: Dendogram of the hierarchical clustering with 3 clusters.

While the average silhouette widths were slightly lower for k-means compared to GMM's best-performing cluster, the k-means clusters exhibited more balanced sizes, which can be advantageous for interpretability and practical application.

```
##   cluster size ave.sil.width
## 1       1  225          0.15
## 2       2  206          0.26
## 3       3  542          0.19
```

```
##   cluster size ave.sil.width
## 1       1  223          0.15
## 2       2  191          0.29
## 3       3  559          0.18
```

```
##   cluster size ave.sil.width
## 1       1  189          0.16
## 2       2  194          0.28
## 3       3  590          0.17
```

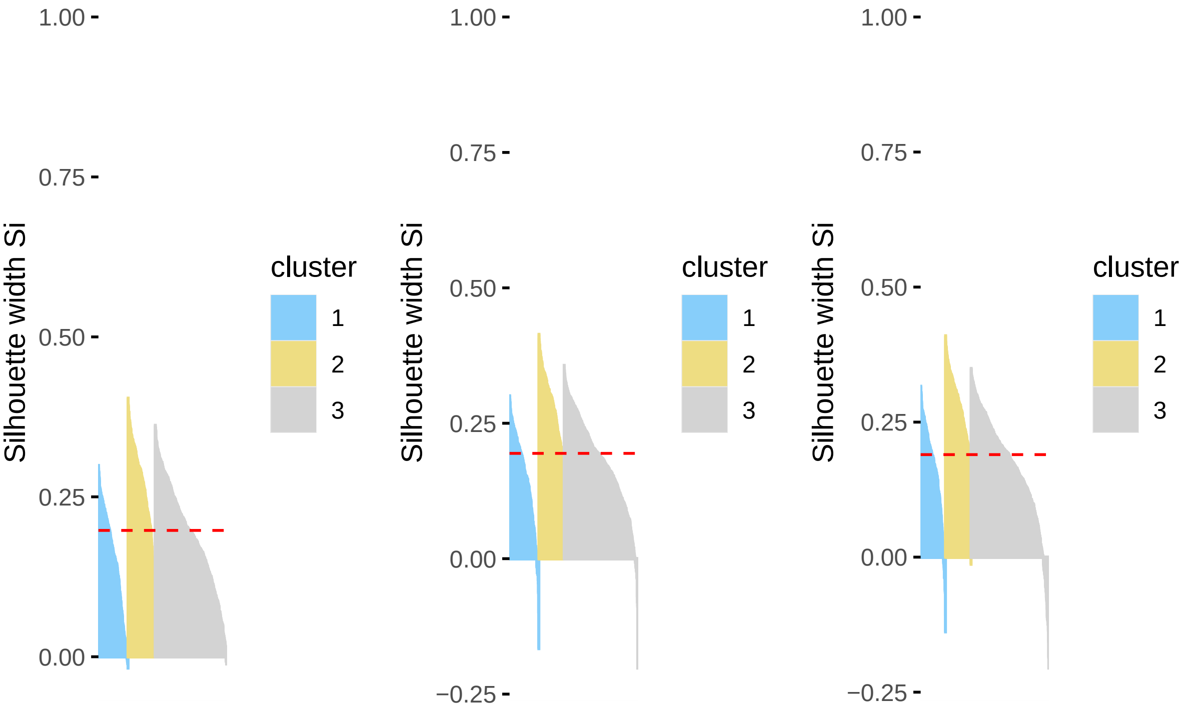


Figure 4: Comparison between silhouette plots of K-means Cluster (left, Average width 0.197), GMM Clusters (center, Average width 0.194) and Hierarchical Cluster (right, Average width 0.188)

References

Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining. MIT Press.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification (2nd ed.). Wiley-Interscience.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed., Chap. 14). Springer. https://doi.org/10.1007/978-0-387-84858-7

Liu, B. (2011). Web data mining: Exploring hyperlinks, contents, and usage data (2nd ed.). Springer. https://doi.org/10.1007/978-3-642-19460-3

Valakhorasani. (2022). Gym members exercise dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset

Annex

Exploratory univarate data analysis

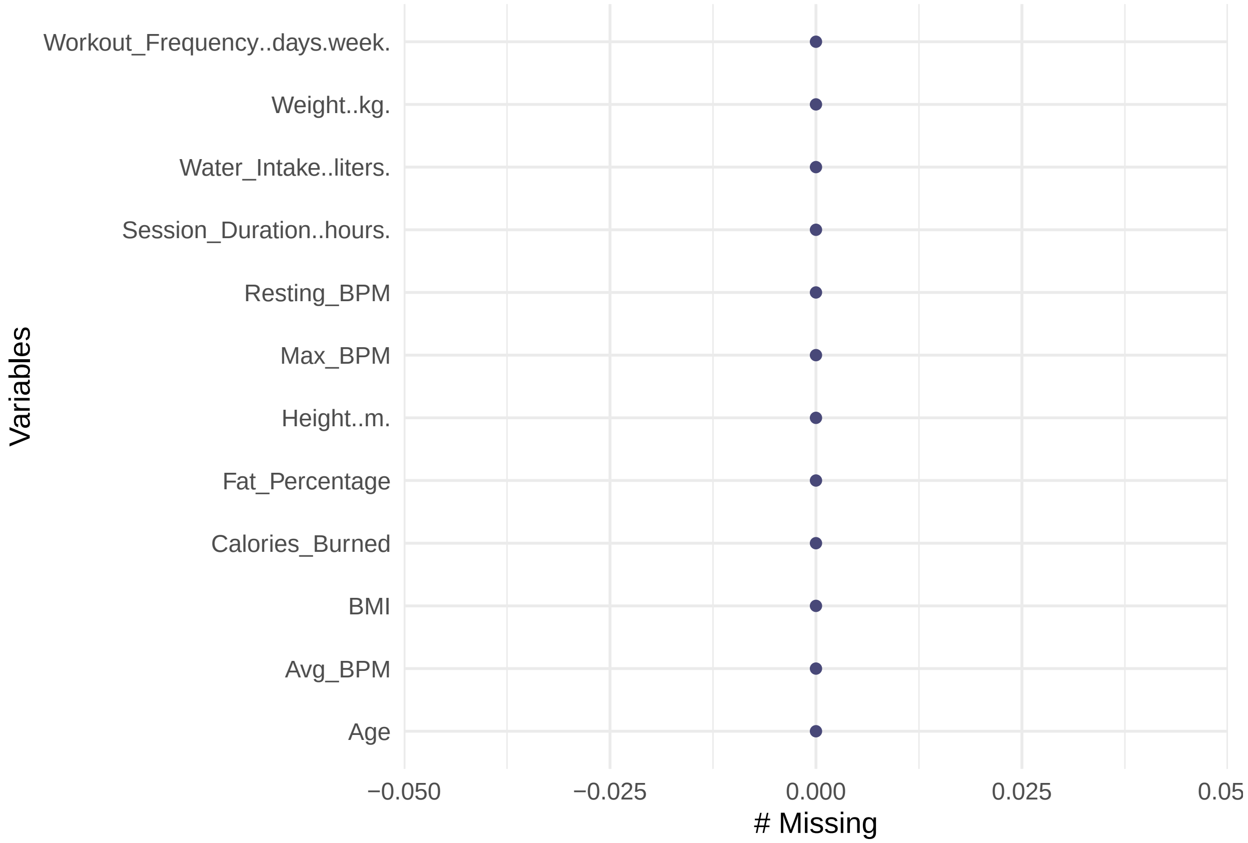


Figure 5: There are no missing values on this dataset.

Exploring correlations

Principal components analysis

Table 5: Loadings of the three first principal components

	PC1	PC2	PC3
Age	-0.033	0.004	0.090
Weight..kg.	0.192	0.634	-0.110
Height..m.	0.142	0.229	0.689

	PC1	PC2	PC3
Max_BPM	0.011	0.081	-0.106
Avg_BPM	0.073	-0.061	-0.255
Resting_BPM	0.001	-0.034	0.009
Session_Duration..hours.	0.453	-0.274	-0.135
Calories_Burned	0.478	-0.201	-0.168
Fat_Percentage	-0.458	-0.003	-0.151
Water_Intake..liters.	0.355	0.252	0.329
Workout_Frequency..days.week.	0.387	-0.238	-0.089
BMI	0.130	0.546	-0.494

```
## [[1]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (Weight..kg.)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |     -30.85      |  [-33.68, -28.02]   |        0         |
## |3-1 |3-1        |     -42.37      |   [-44.7, -40.04]   |        0         |
## |3-2 |3-2        |     -11.52      |   [-13.92, -9.12]   |        0         |
##
## [[2]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (Height..m.)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |      -0.06      |   [-0.09, -0.04]    |        0         |
## |3-1 |3-1        |      -0.10      |   [-0.13, -0.08]    |        0         |
## |3-2 |3-2        |      -0.04      |   [-0.06, -0.02]    |        0         |
##
## [[3]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (Session_Duration..hours.)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |      0.60       |    [0.55, 0.65]     |      0.000       |
## |3-1 |3-1        |      -0.01      |    [-0.05, 0.04]    |      0.888       |
## |3-2 |3-2        |      -0.61      |   [-0.65, -0.56]    |      0.000       |
##
## [[4]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (Calories_Burned)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |     401.88      |  [355.84, 447.91]   |      0.000       |
## |3-1 |3-1        |     -55.94      |  [-93.79, -18.08]   |      0.002       |
```

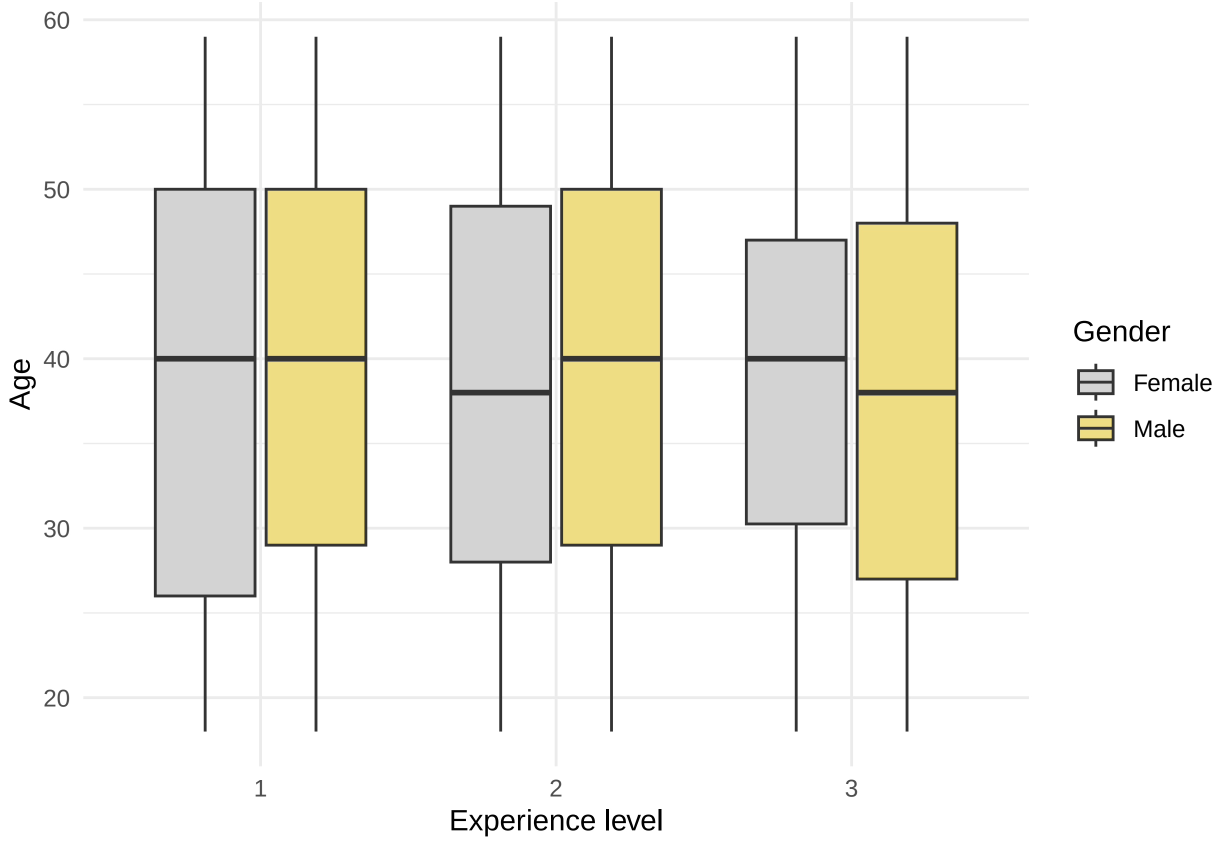


Figure 6: On average gym members are around 40yo. Expert level male member are on average younger than their female counterparts while the opposite is true for intermediate level members. Note that experience levels are organized from 1 - Beginner to 3 - Expert.

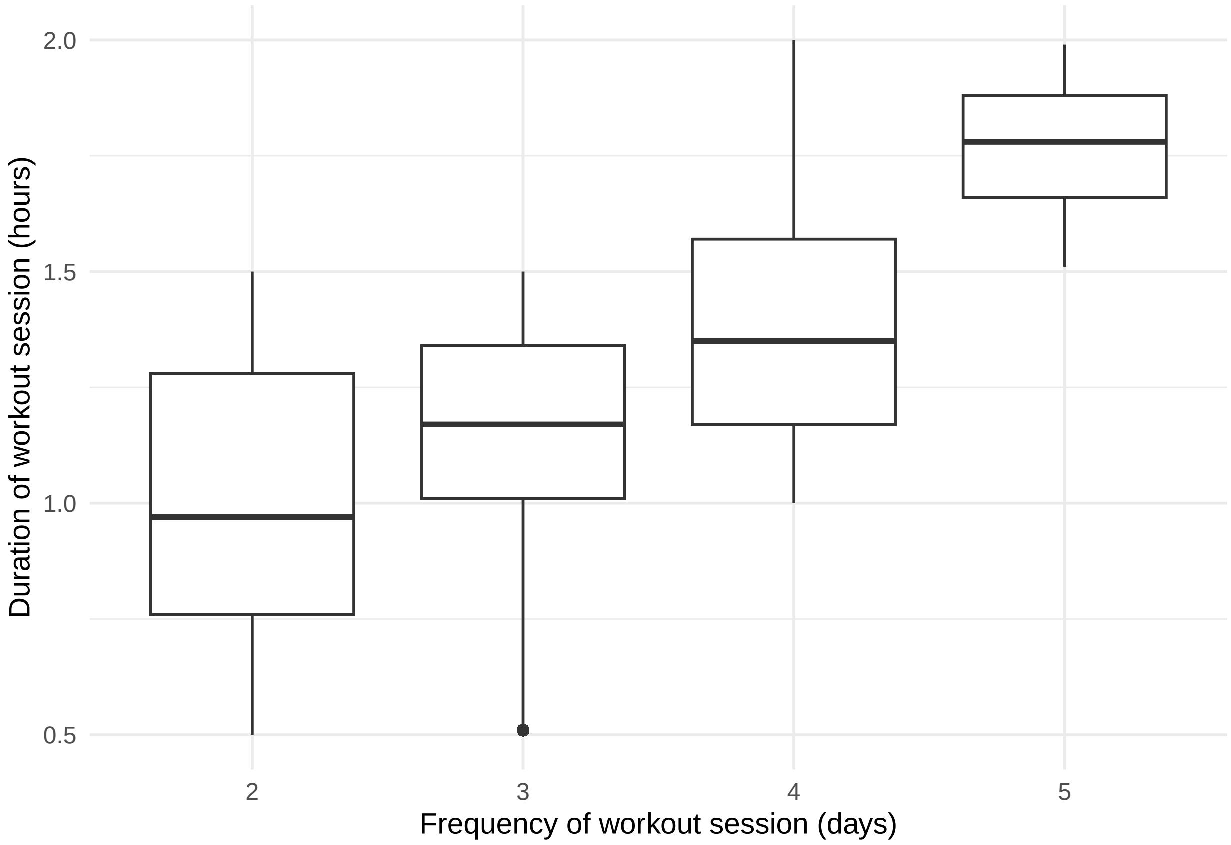


Figure 7: Members that workout with more frequency 4 to 5 days per week usually have longer workout sessions.

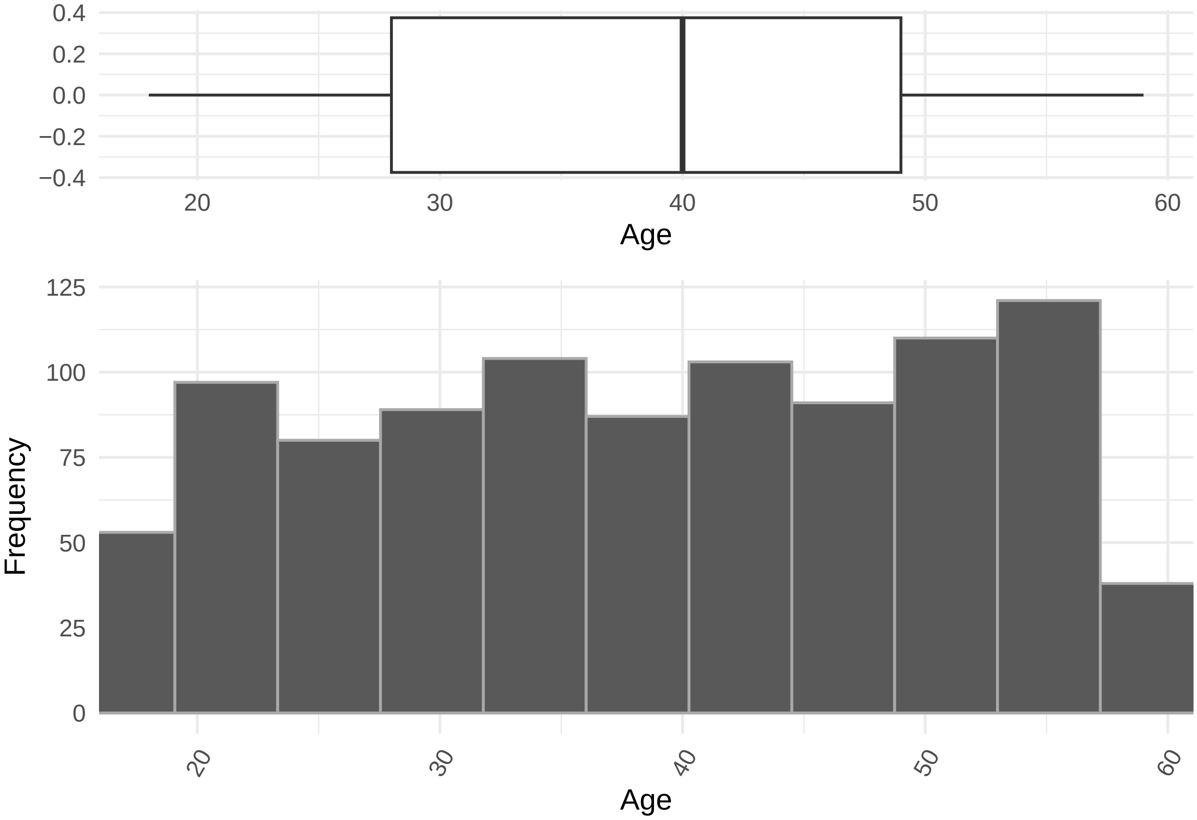


Figure 8: Histogram and Boxplot of Gym member's Age

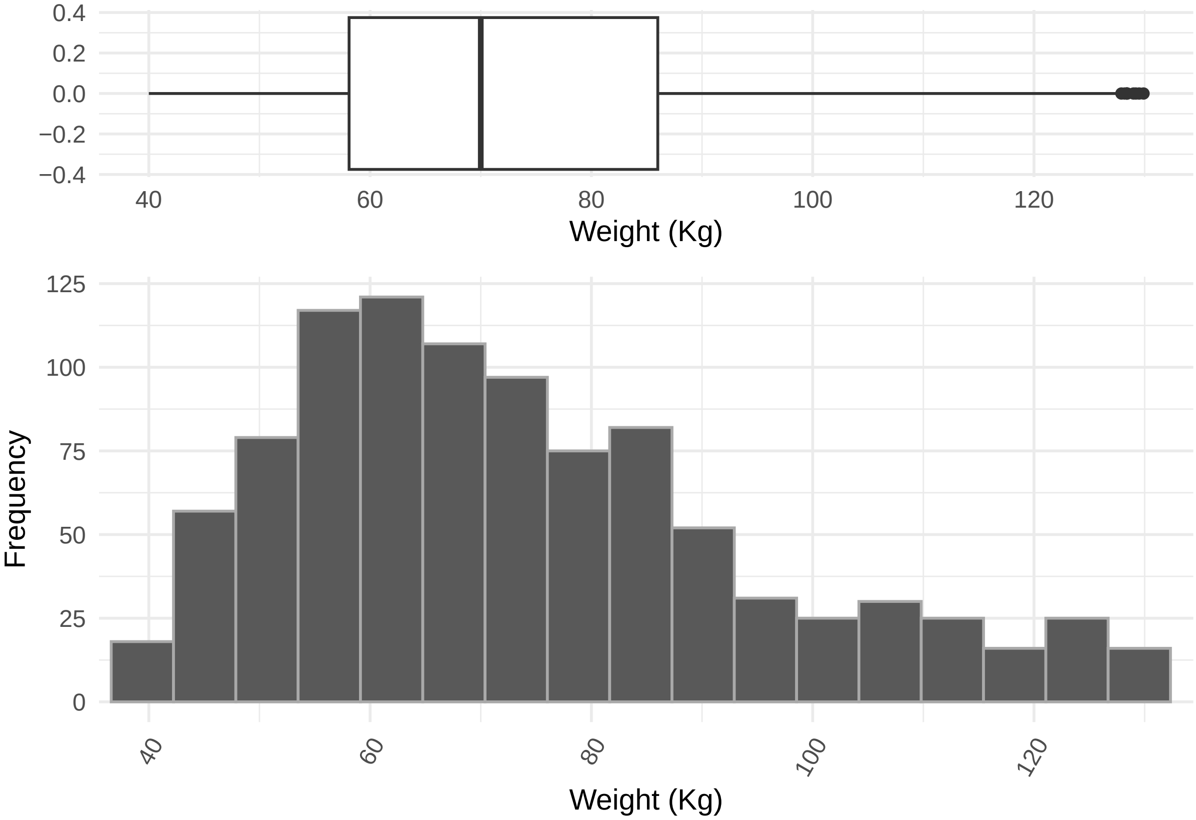


Figure 9: Histogram and Boxplot of Gym member's Weight

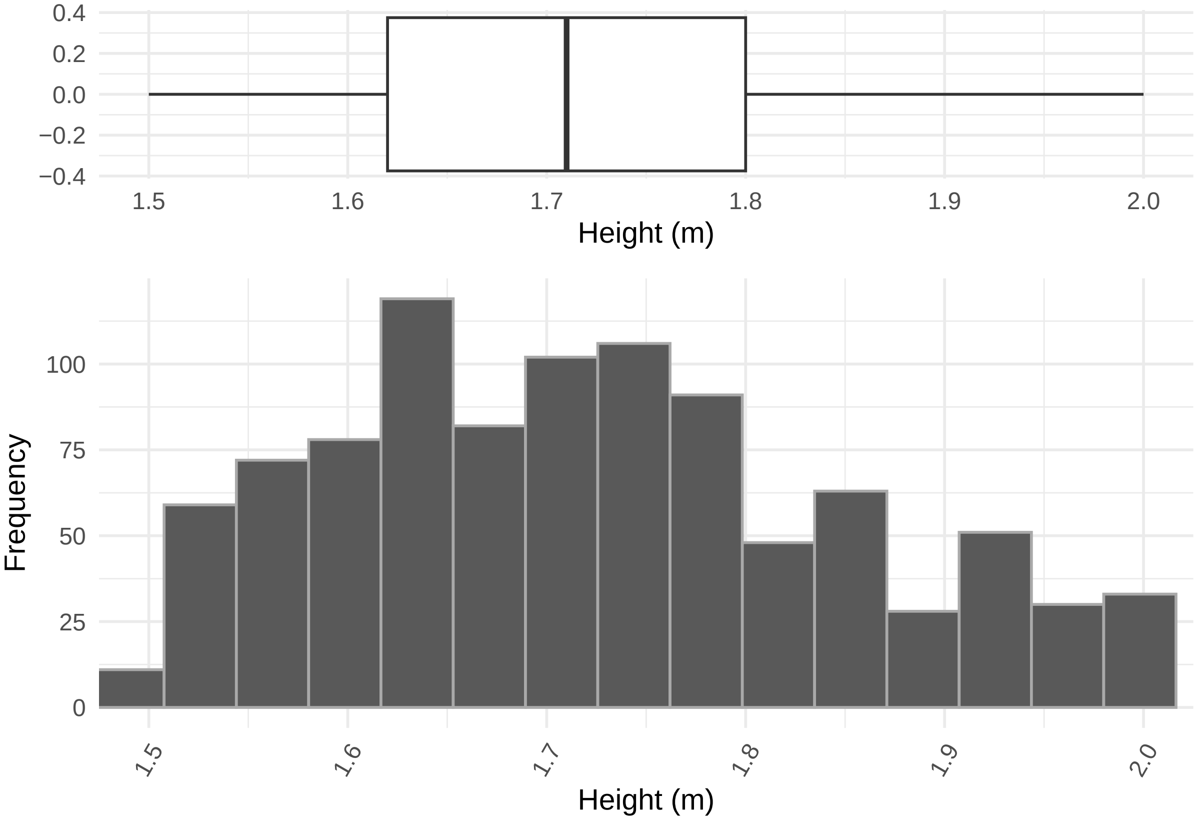


Figure 10: Histogram and Boxplot of Gym member's Height

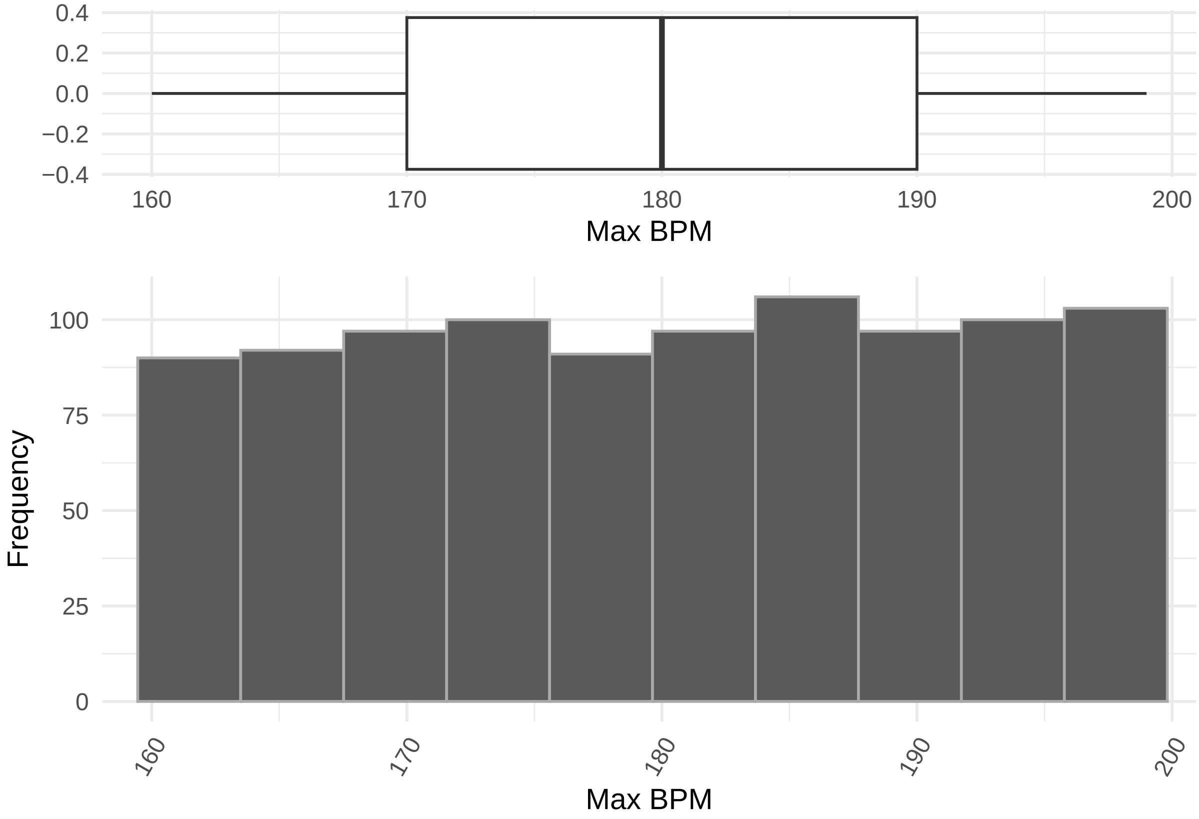


Figure 11: Histogram and Boxplot of Gym member's Maximum heart rate

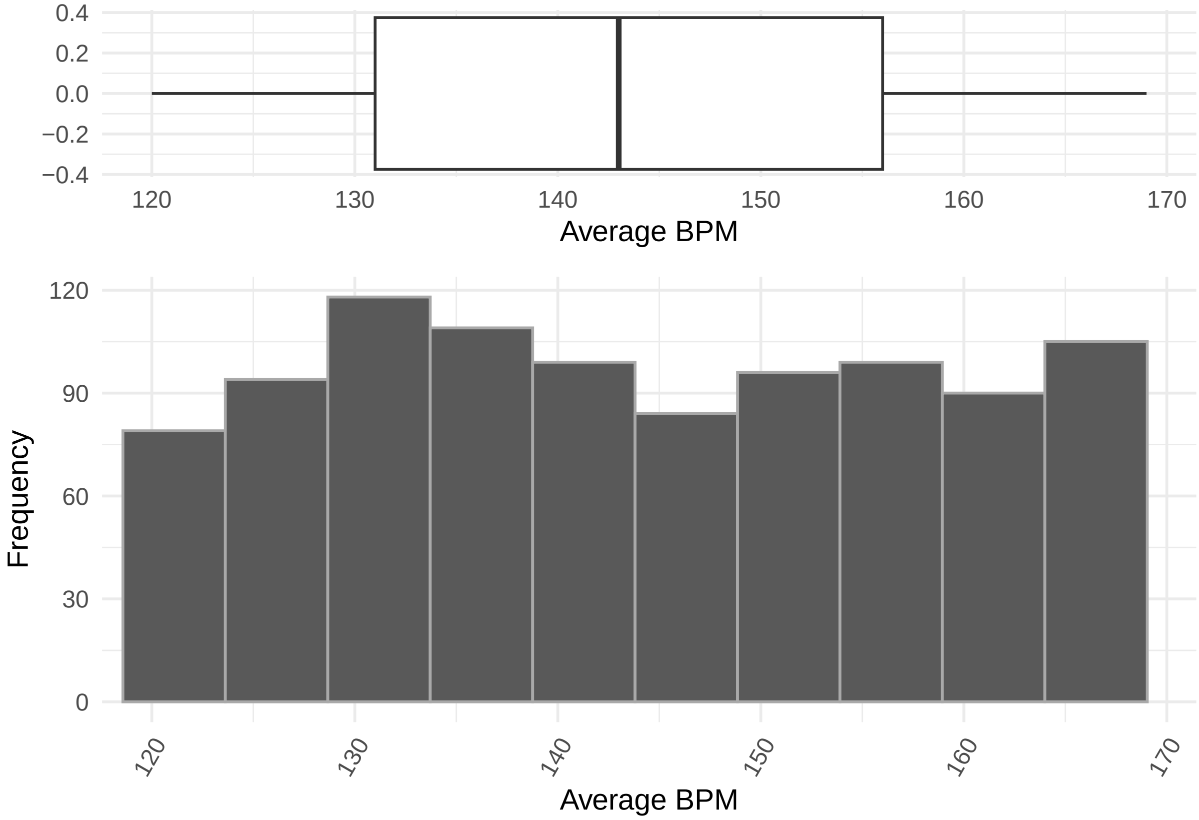


Figure 12: Histogram and Boxplot of Gym member's Average heart rate

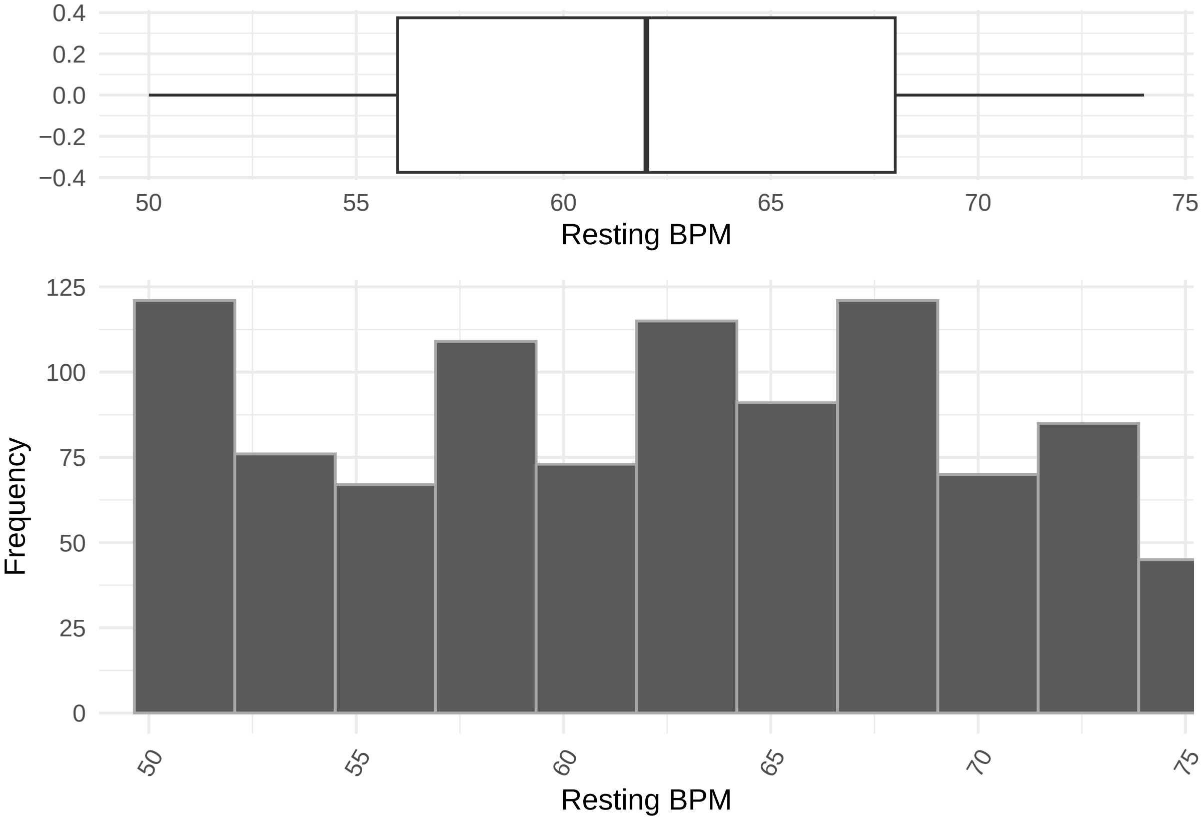


Figure 13: Histogram and Boxplot of Gym member's Resting heart rate

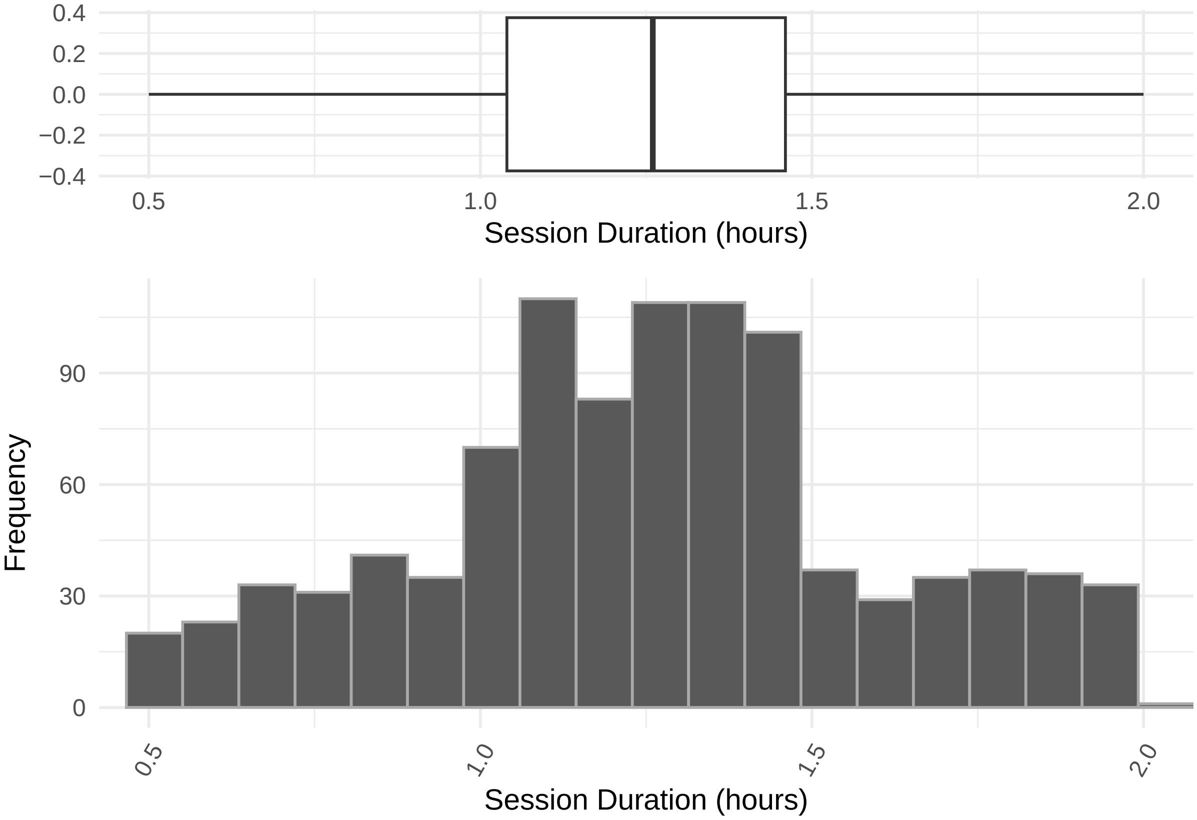


Figure 14: Histogram and Boxplot of Gym member's Training Sessions duration

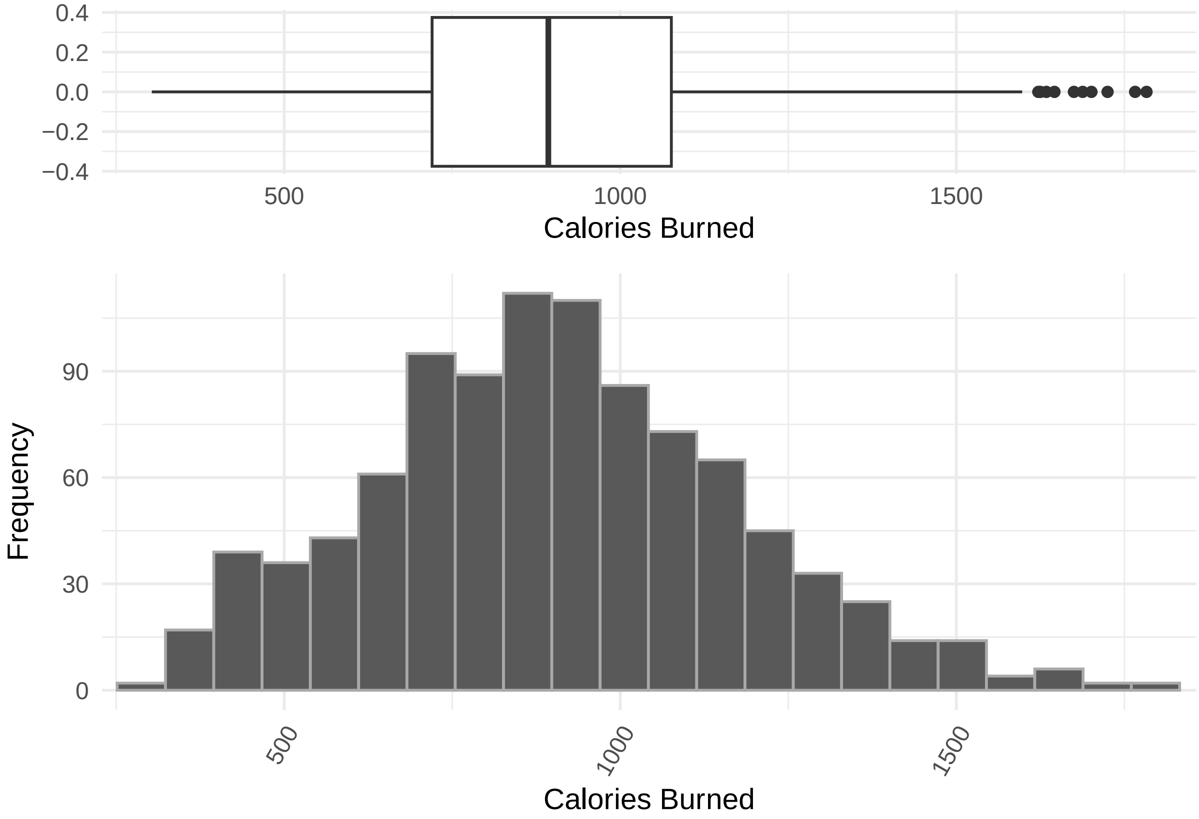


Figure 15: Histogram and Boxplot of Gym member's Calories burned during each session

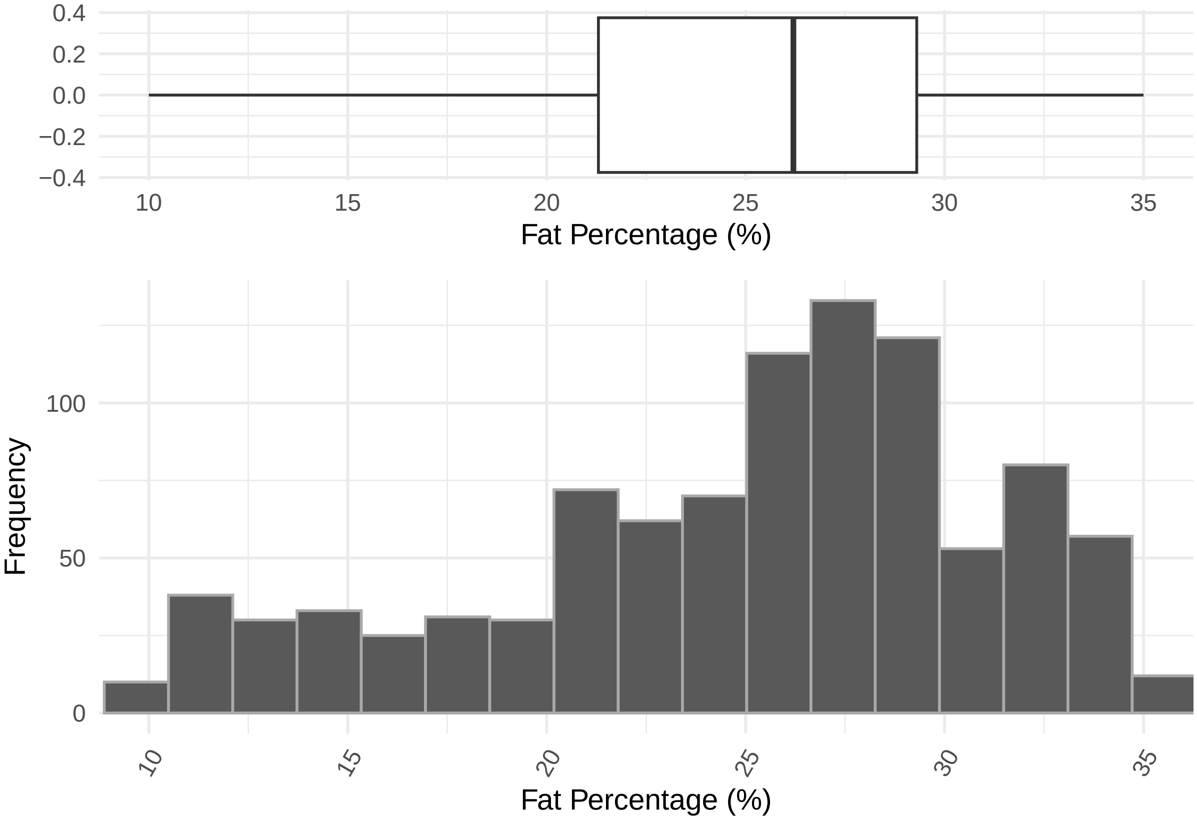


Figure 16: Histogram and Boxplot of Gym member's body fat percentage

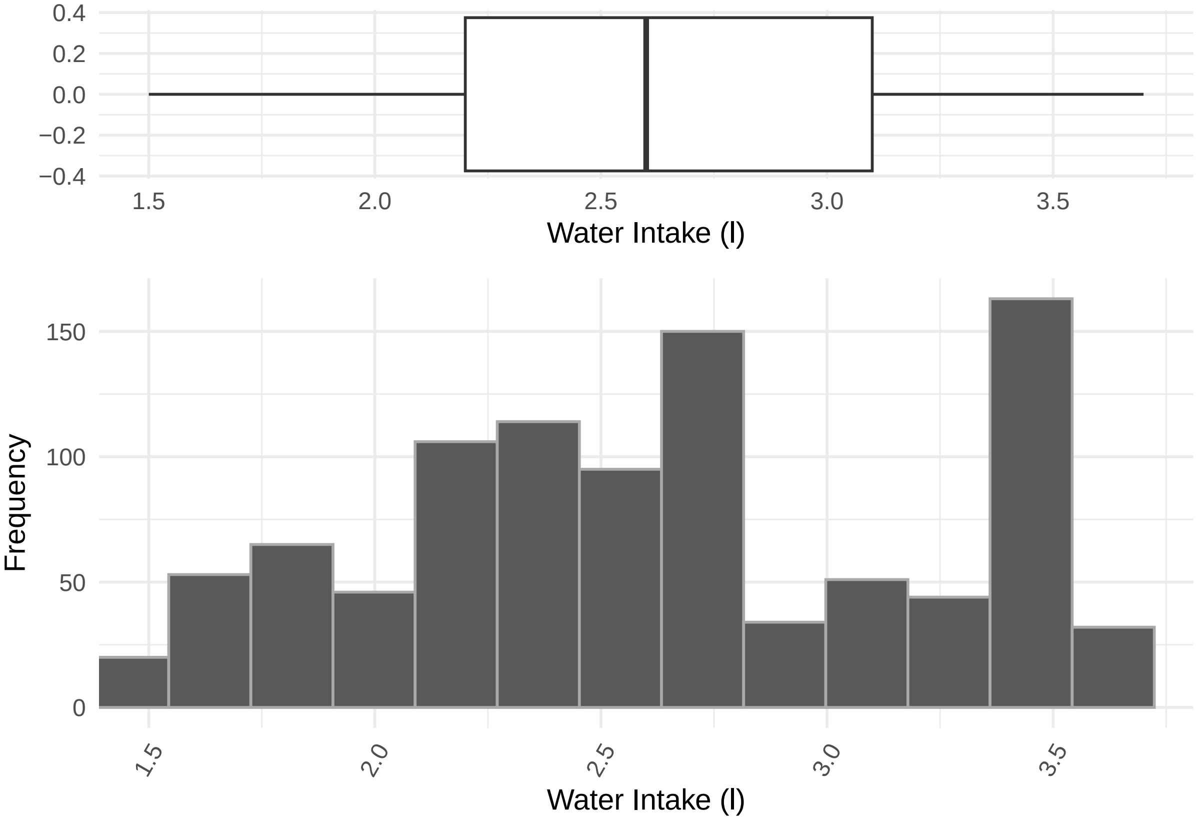


Figure 17: Histogram and Boxplot of Gym member's daily water intake during workouts

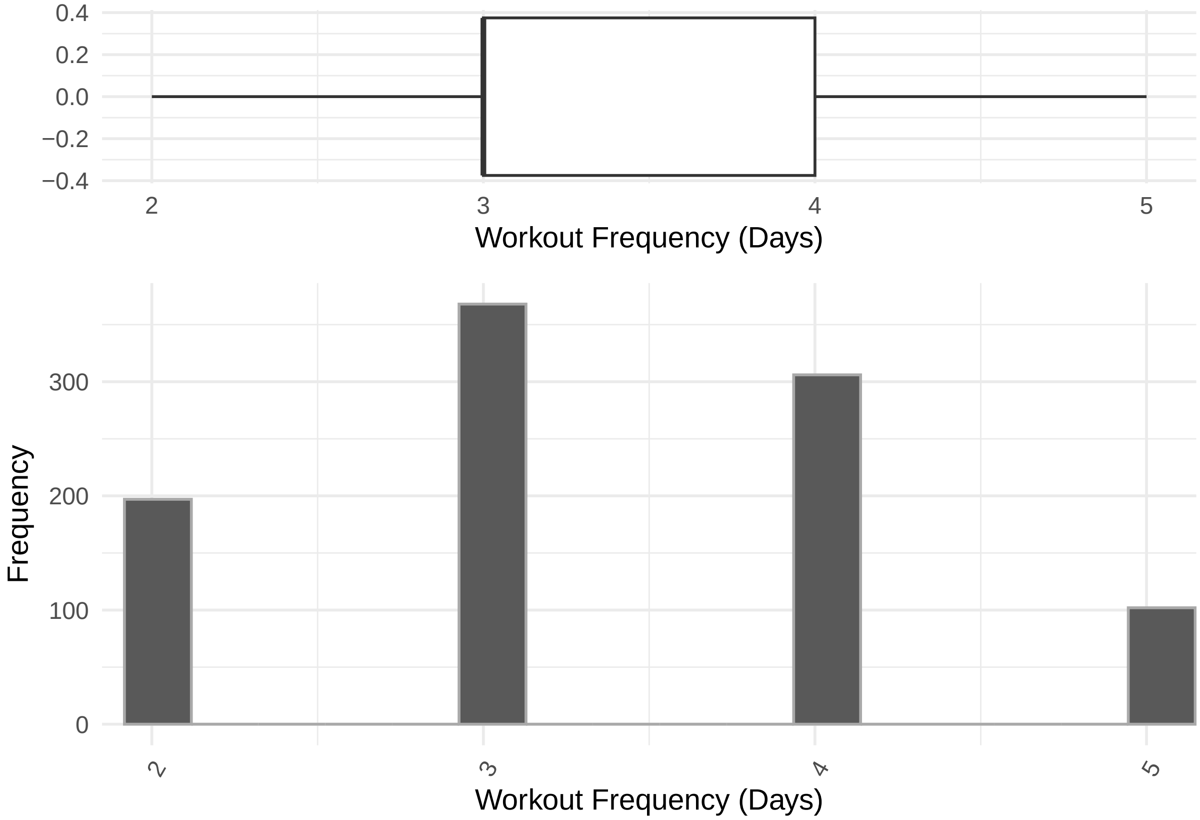


Figure 18: Histogram and Boxplot of Gym member's number of workout sessions per week

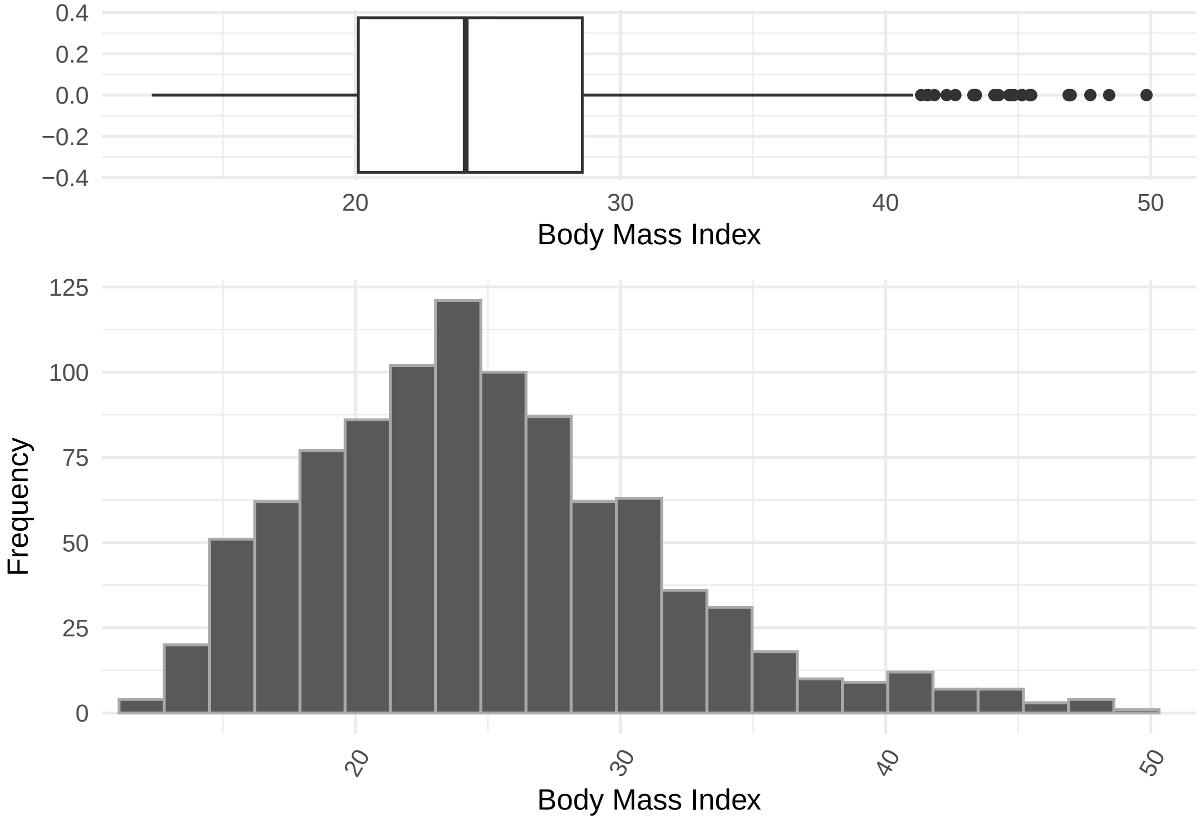


Figure 19: Histogram and Boxplot of Gym member's Body Mass Index

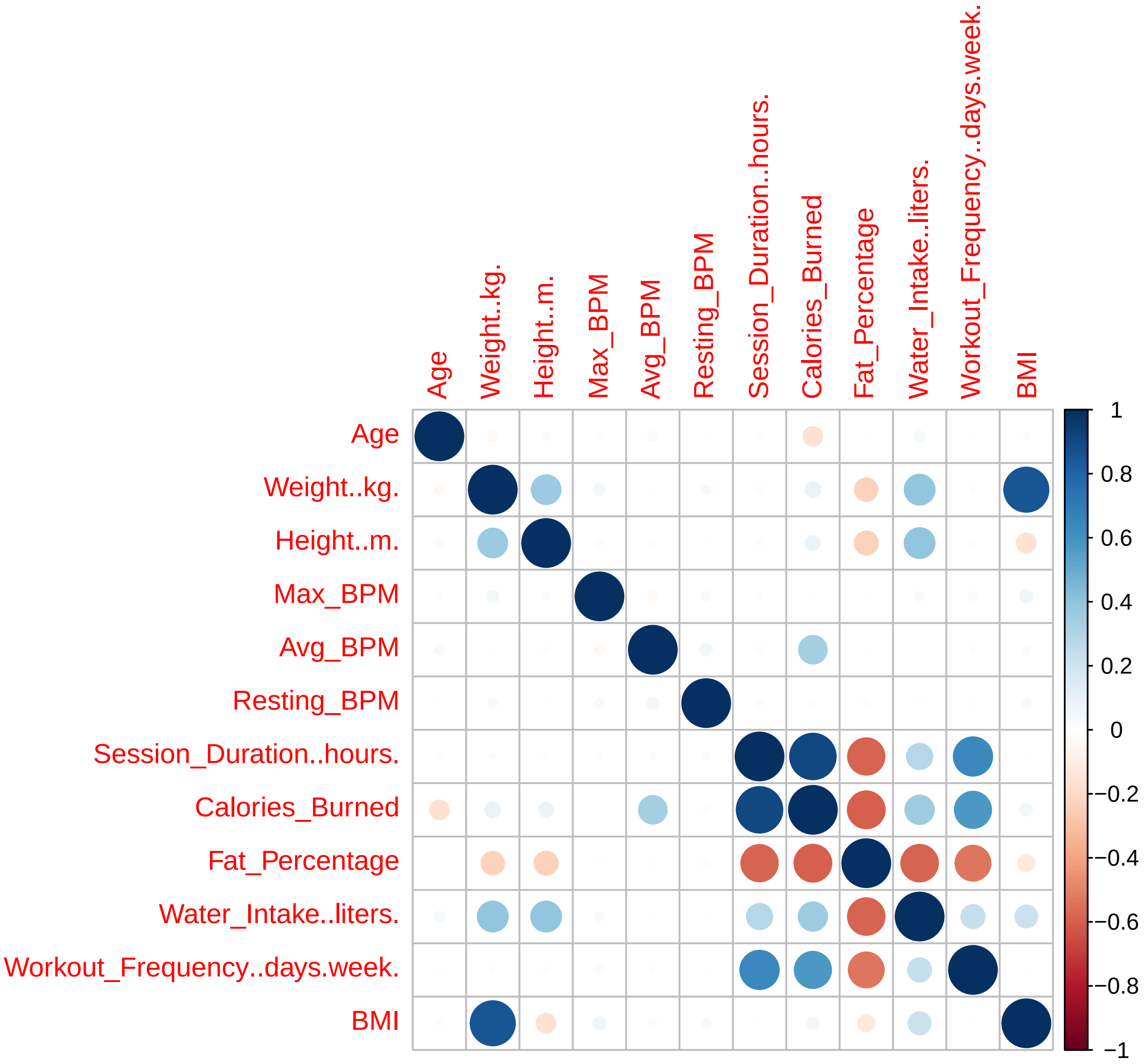


Figure 20: Correlation matrix

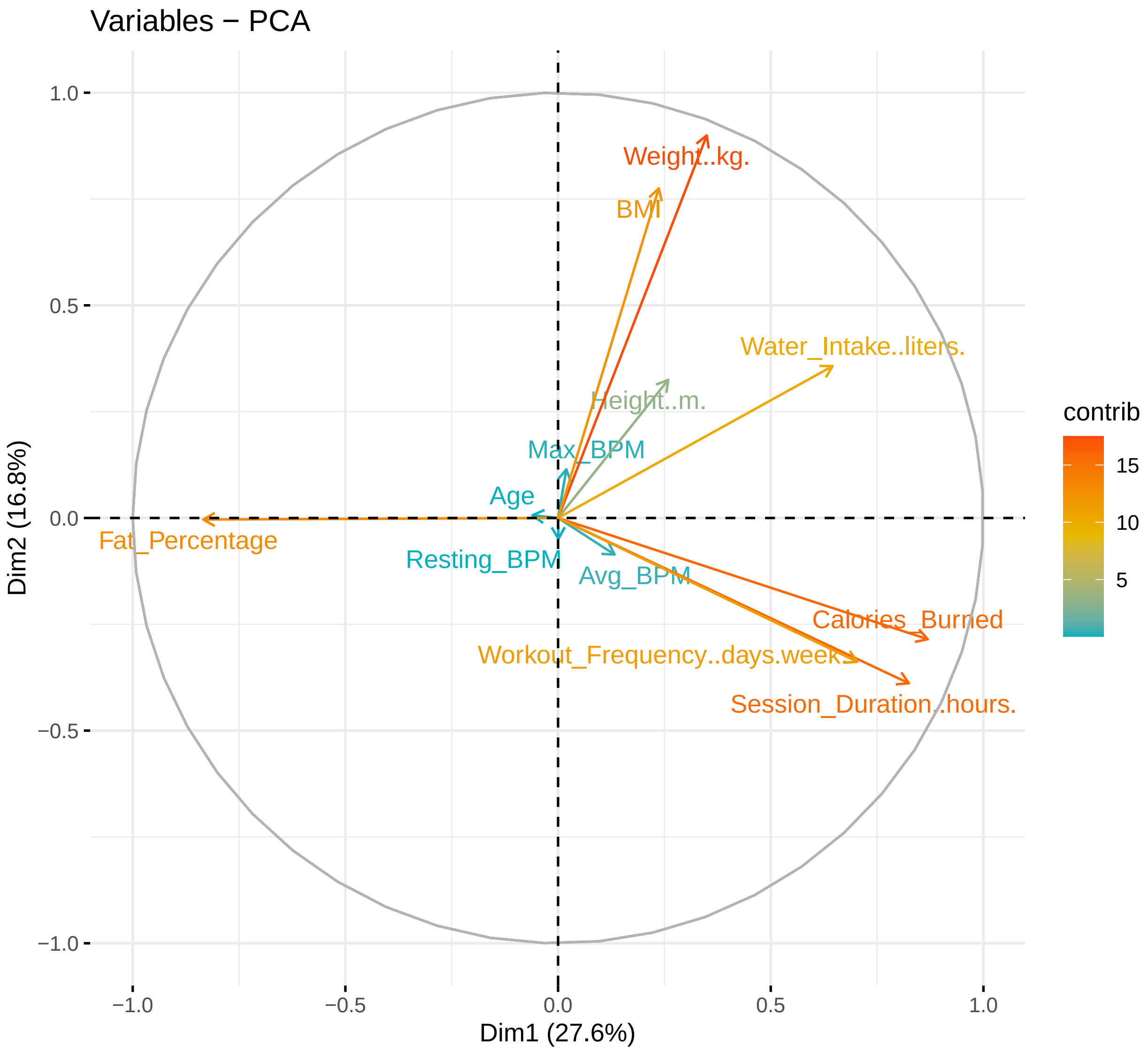
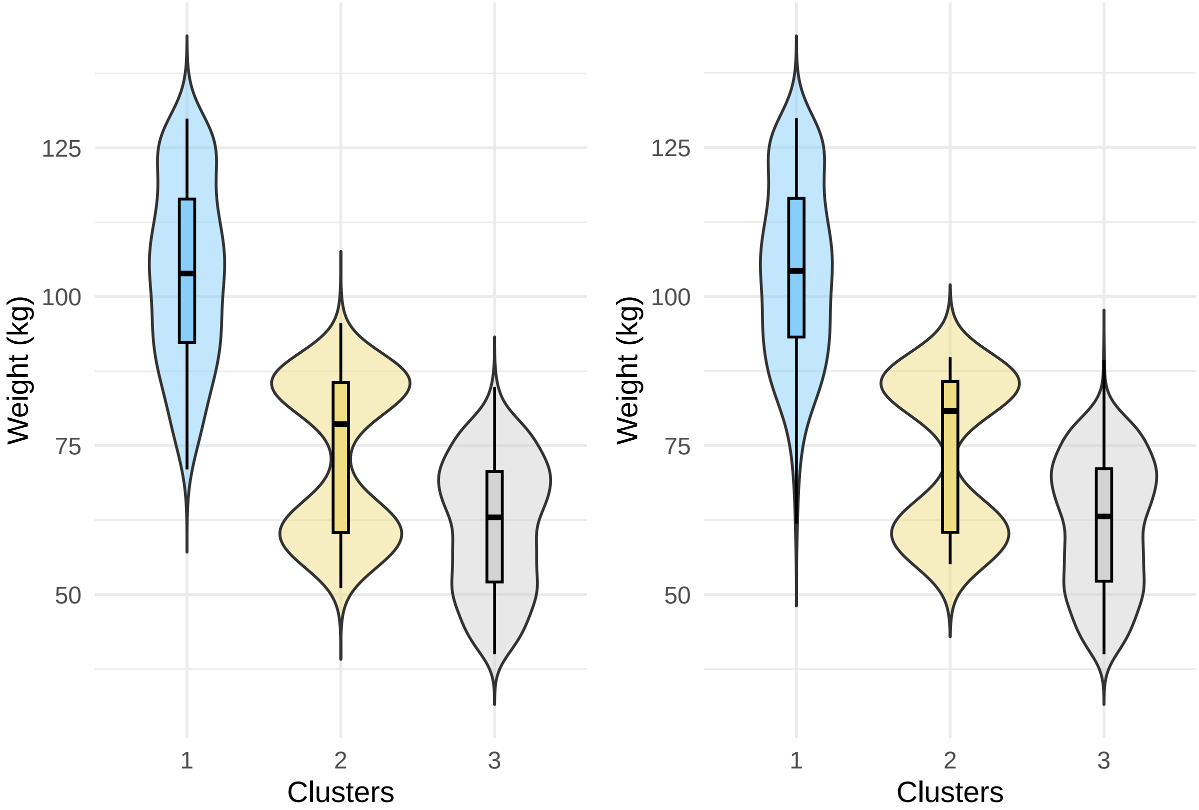
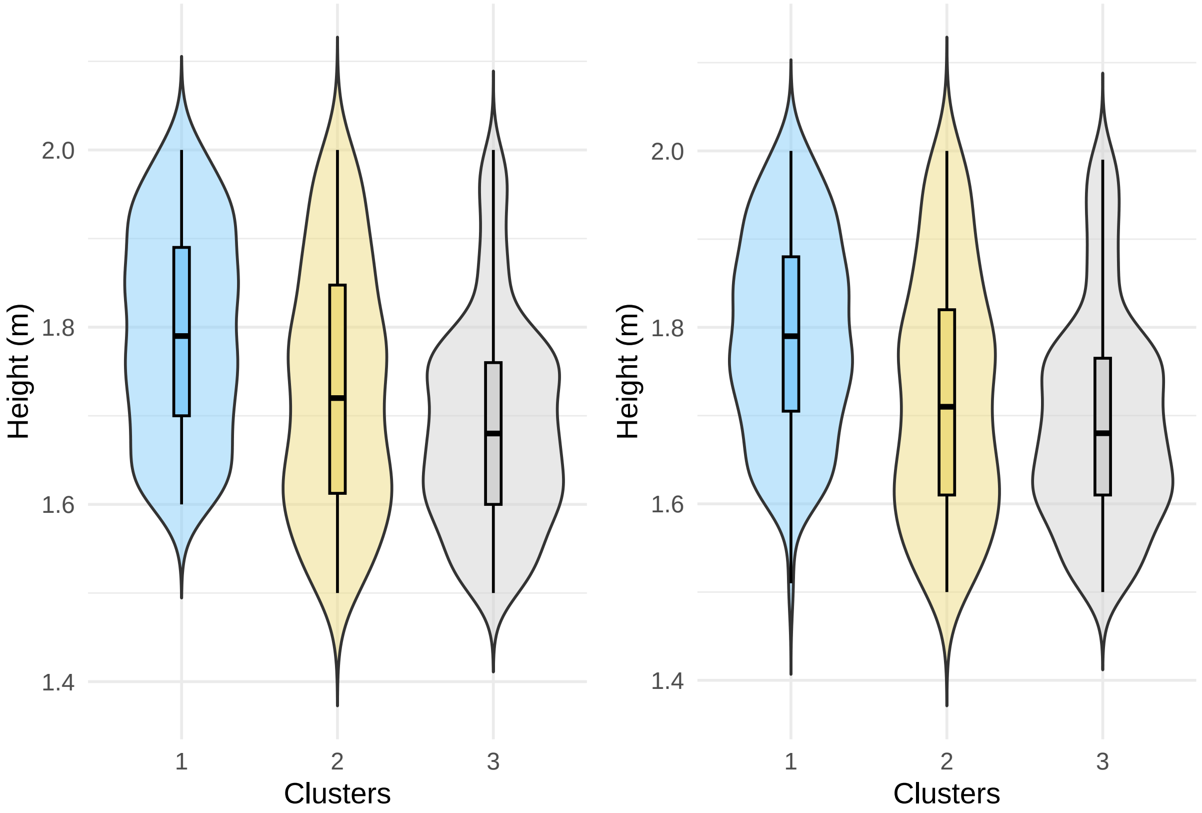
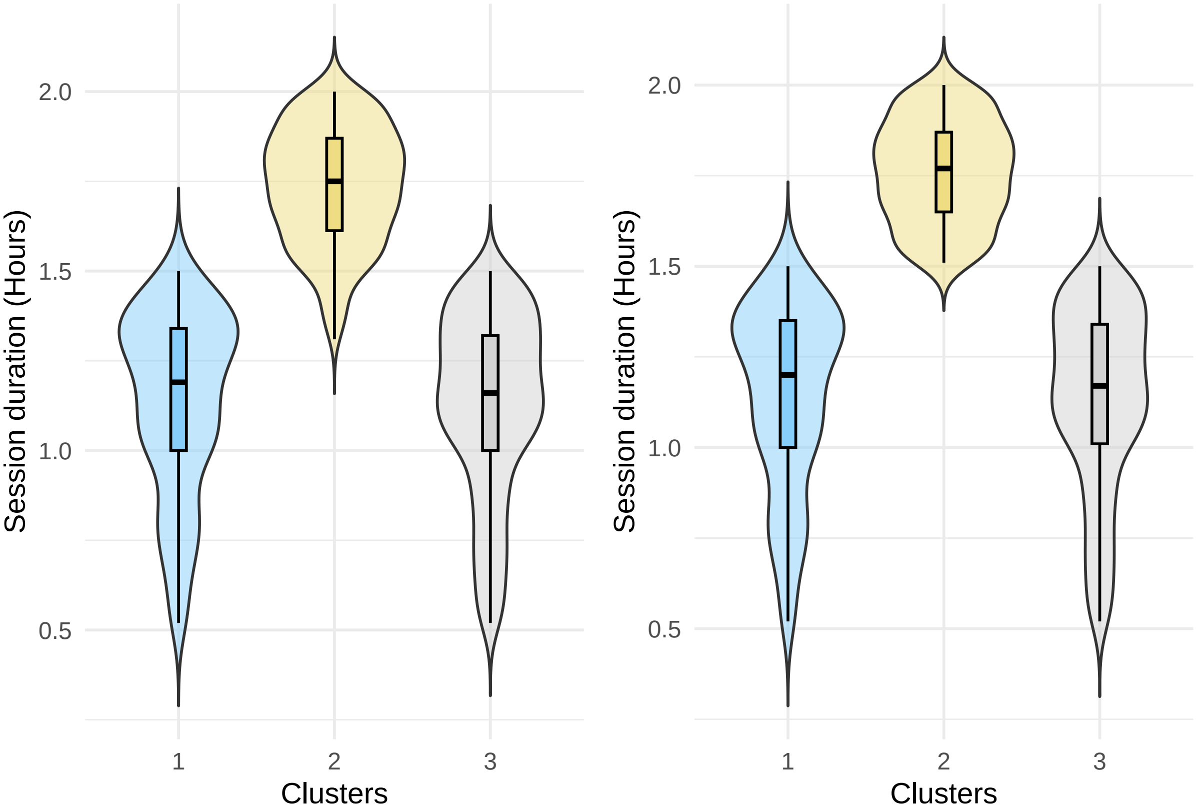


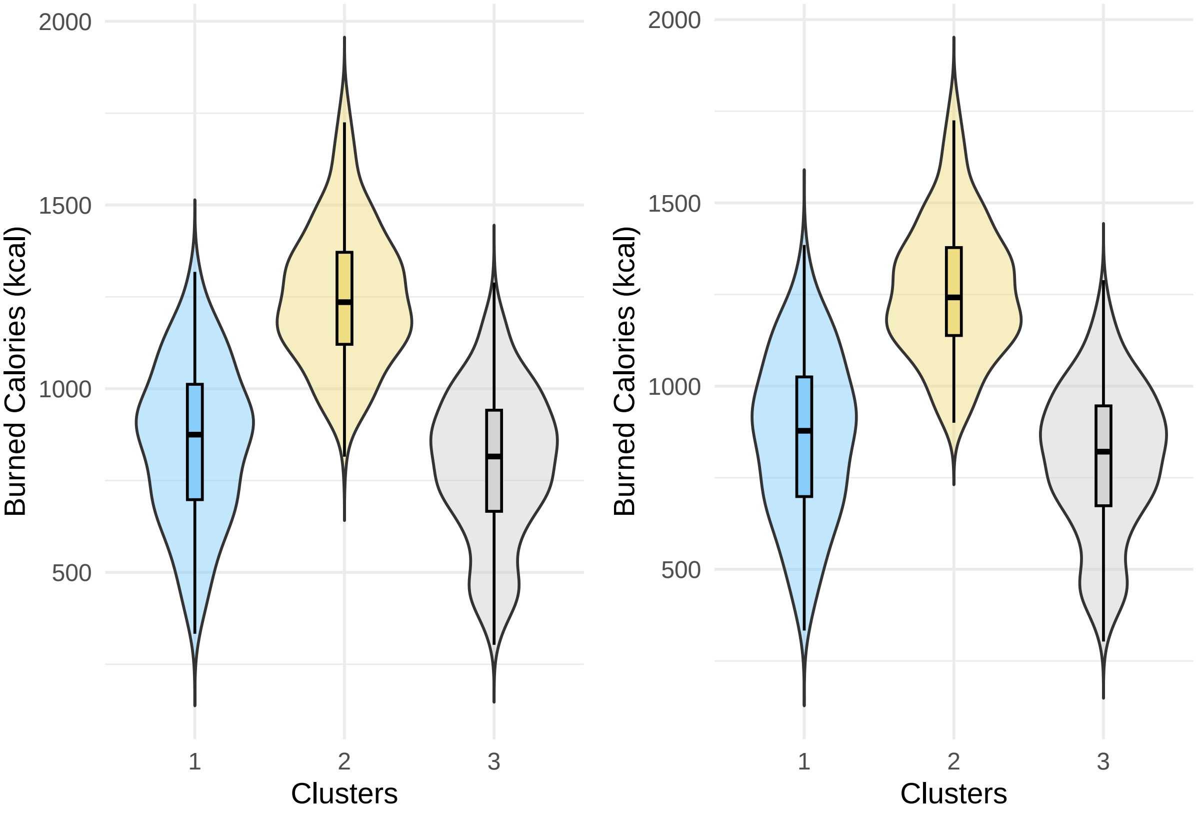
Figure 21: Variables effect on principle components

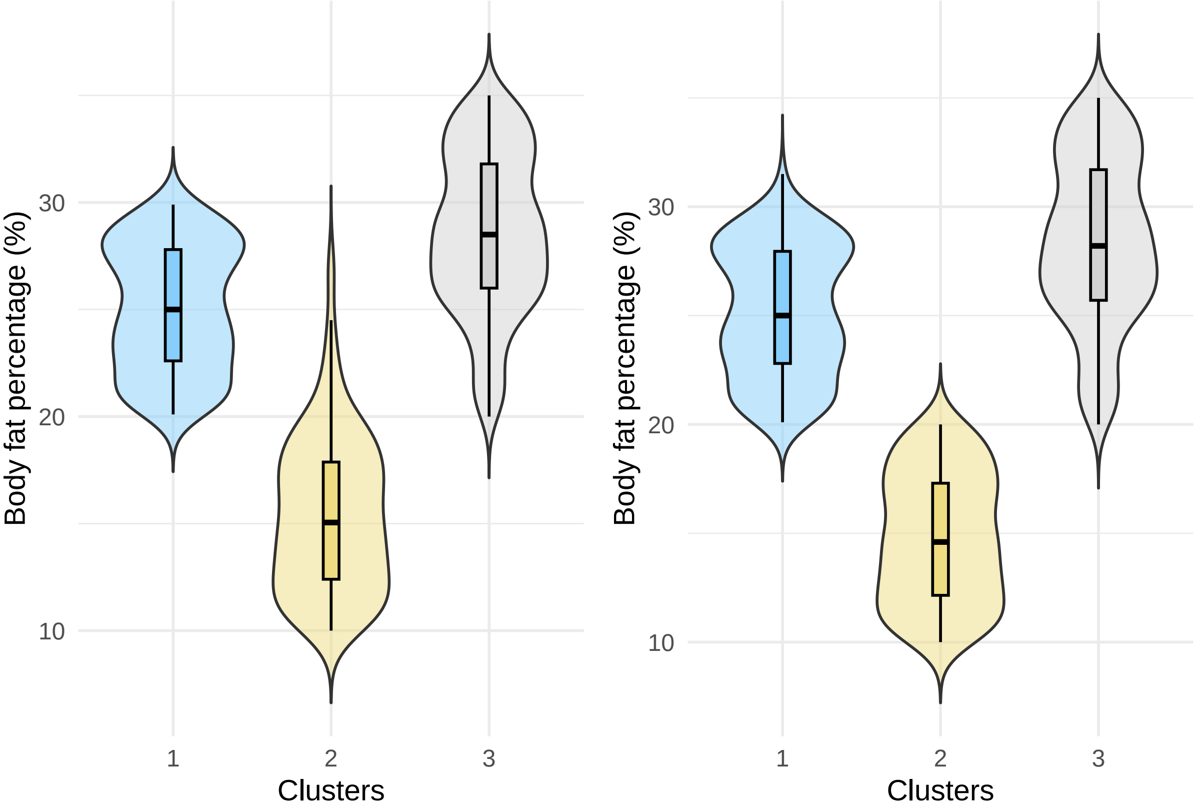
```
## |3-2 |3-2        |     -457.81     | [-496.88, -418.74]  |      0.000       |
##
## [[5]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (Fat_Percentage)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |      -9.67      |   [-10.47, -8.87]   |        0         |
## |3-1 |3-1        |      3.51       |    [2.85, 4.17]     |        0         |
## |3-2 |3-2        |      13.18      |    [12.5, 13.86]    |        0         |
##
## [[6]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (Water_Intake..liters.)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |      0.15       |    [0.05, 0.26]     |      0.002       |
## |3-1 |3-1        |      -0.68      |   [-0.76, -0.59]    |      0.000       |
## |3-2 |3-2        |      -0.83      |   [-0.92, -0.74]    |      0.000       |
##
## [[7]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (Workout_Frequency..days.week.)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |      1.53       |    [1.37, 1.68]     |      0.000       |
## |3-1 |3-1        |      0.08       |    [-0.05, 0.21]    |      0.329       |
## |3-2 |3-2        |      -1.45      |   [-1.58, -1.32]    |      0.000       |
##
## [[8]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (BMI)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |      -8.12      |   [-9.26, -6.99]    |        0         |
## |3-1 |3-1        |     -10.84      |   [-11.77, -9.9]    |        0         |
## |3-2 |3-2        |      -2.72      |   [-3.68, -1.75]    |        0         |

## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 3
## components:
##
##  log-likelihood   n  df       BIC       ICL
##       -10198.98 973 272 -22269.43 -22301.62
##
## Clustering table:
##   1   2   3
## 191 559 223
```

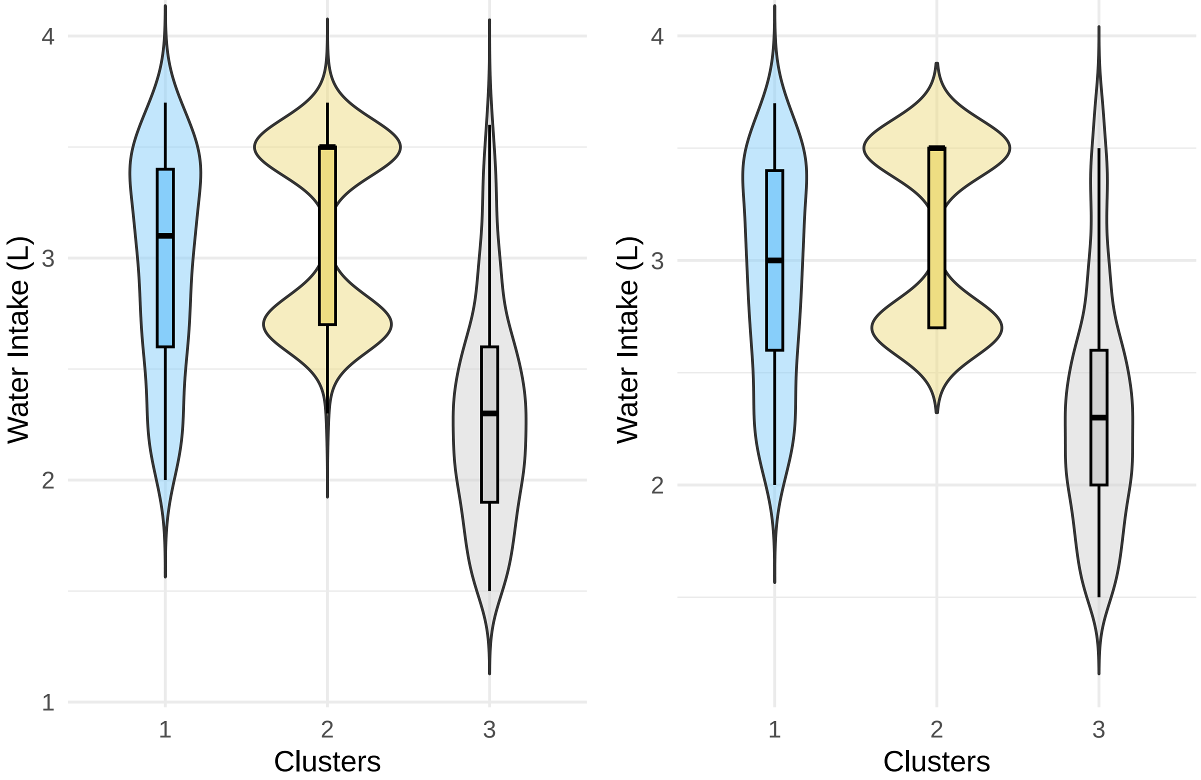


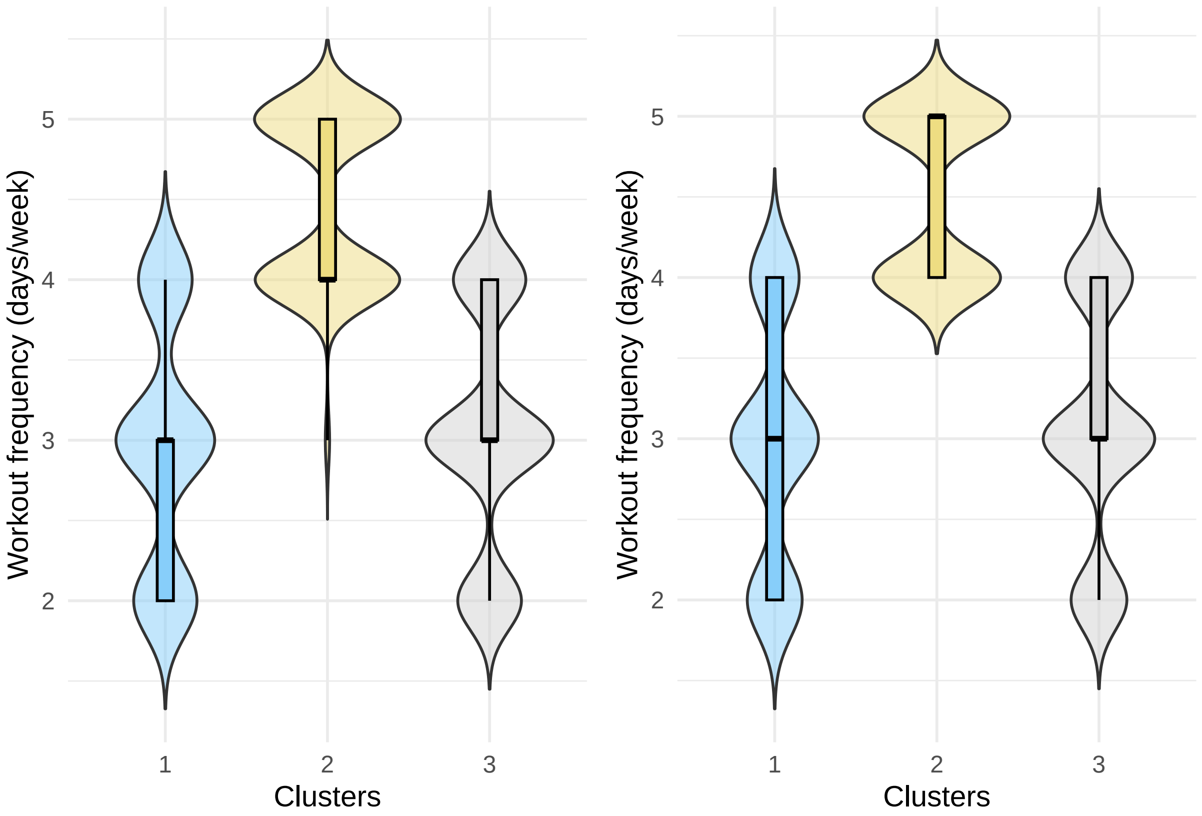

Height (m)
2.0
1.8
1.6
1.4
1
2
3
Clusters
Height (m)
2.0
1.8
1.6
1.4
1
2
3
Clusters

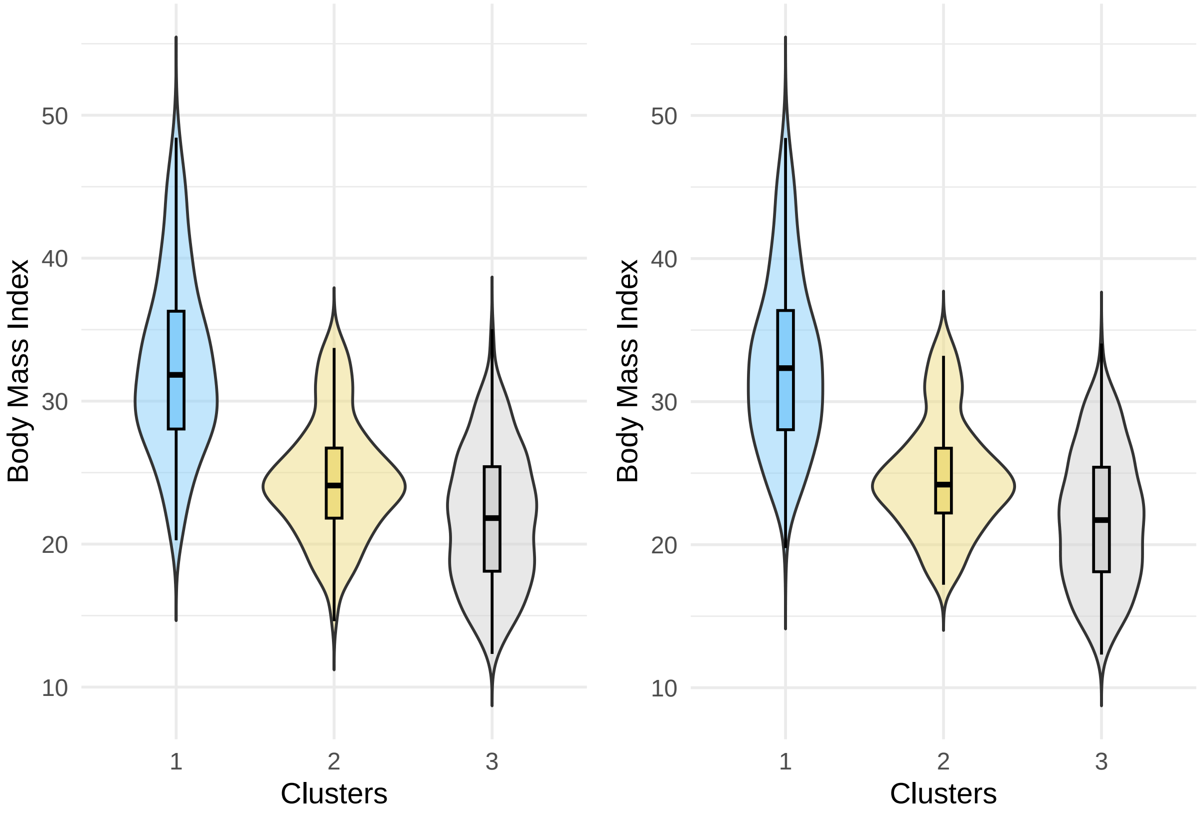

Session duration (Hours)
2.0
1.5
1.0
0.5
1
2
3
Clusters
Session duration (Hours)
2.0
1.5
1.0
0.5
1
2
3
Clusters


2000
1500
1000
500
Burned Calories (kcal)
1
2
3
Clusters
2000
1500
1000
500
Burned Calories (kcal)
1
2
3
Clusters


Body fat percentage (%)
30
20
10
1
2
3
Clusters
Body fat percentage (%)
30
20
10
1
2
3
Clusters


Water Intake (L)
4
3
2
1
1
2
3
Clusters
Water Intake (L)
4
3
2
1
2
3
Clusters


Workout frequency (days/week)
5
4
3
2
1
2
3
Clusters
Workout frequency (days/week)
5
4
3
2
1
2
3
Clusters


Body Mass Index
50
40
30
20
10
1
2
3
Clusters
Body Mass Index
50
40
30
20
10
1
2
3
Clusters