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 5

Results and Discussion **5**

References **8**

Annex **8**
Exploratory univarate data analysis 8
Exploring correlations 8
Principal components analysis 8

Introduction

Real world problems are usually complex and dependent on a variety of factors, in order to be able to solve them we must rely on statistical learning methods. Statistical learning methods can be divided between supervised learning methods and unsupervised learning methods. Supervised learning methods are useful for classification and prediction, where the goal is to predict the value of a target variable using other known input variables. Unsupervised Learning methods are useful for clustering, association rule mining and dimensionality reduction, where the goal is to uncover patterns or structures from data without specified target outputs (1). Both of these methods rely on identifying structure in data but differ in whether a response variable guides that process.

In this project our goal is to be able to identify fitness patterns and performance across a diver gym experience levels using the *Gym Members Exercise Dataset* (2). This dataset has 15 variables with 973 observations each. Table 1 shows each variable and it's subsequent meaning. The purpose of the present project is to assess the differences between gym members considering their experience level, and how that impacts their performance.

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 start by taking a look at the way each of the fitness metrics and demographic variables relate to the members experience level to establish a baseline understanding. Then we perform a Principal Components Analysis and finally Clustering in order to define the underlying patterns in the data.

Exploratory Data Analysis

In order to prevent unwanted results when performing PCA we checked our data for missing values. As is possible to see from figure 5 this is not the case. Table 2 provides a quick assessment of descriptive statistics of the numeric variables in the Gym members dataset. The average gym member is a 39 yo male, 1.72 m tall, weighting around 74 kg, with 25% body fat and a body mass index of 25. Members average heart rate when working out is 144 beats per minute and when resting is 62 bpm and the maximum heart rate when working out on average is 180 beats per minute. Workout sessions tend to last 1 hour and 16 minutes spread between 3 workouts a week. The average member drinks 2.6 liters of water per workout. Also there aren't many differences in age between male and female members across the three levels of experience, expert male members tend to be younger than their female counterparts and the same is true for female intermediate members. Members who workout more days of the week usually tend to do longer sessions, see 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

Considering the probabilistic distribution of different variables from figures 8, 11, 12 and 13 it is possible to assess that the age and different heart rate measures of the gym members seem to follow an uniform distribution. While the remaining variables being reasonable symetric. Weight, Height and body mass index are slightly right skewed while fat percentage is slightly left skewed. The distribution of data can be assessed from the box-plot and histogram combination on figure 8-19 on the annex section of this report.

From figure 19 is possible to see that the strongest positive correlation between variables happens between the Calories Burned per sessions and the Session Duration, as expected longer sessions increase energy expenditure. Also workout frequency as is possible to assess from figure 7 also shows a positive strong correlation with the session duration. Fat percentage is strongly negatively correlated with the amount of calories burned per session, the amount of water intake per session, duration of session and even frequency of workout sessions. Indicating that higher levels of body fat are correlated with less time spent at the gym.

Methods

Having establish that there are no missing values on our dataset and some a priori relationships between variables it is possible to proceed to conduct a Principal Components Analysis followed by Clustering analysis using k-means method. Due to the different scales in which data is available we start by normalizing it.

Principal Components Analysis

PCA is used to reduce data dimensionality while preserving variance. It is a Matrix factorization technique aimed at data simplification. PCA seeks to find a sequence of best linear approximations to a multivariate dataset by projecting the data onto lower-dimensional subspaces (3). It provides insight into structure, variance and dimension reduction in unsupervised learning. It identifies a sequence of affine hyperplanes (subspaces) that approximate the data in a least-squares sense.

To solve PCA, we center the data matrix $\mathbf{X} = UDV^T$: U - left singular vectors (observations in PC space); D - Diagonal matrix of singular values; V - Right singular vectors (principal component directions). The principal components are the rows of **UD**, the principal axes are the columns of **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 shows the summary of the PCA, where you can see the variance explained by each principal component. The first principal component explains 28% , while the second principal component explains 17% and the third principal component explains 11% of the total variance. Together, the first three principal components explain 56% of the total variance.

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

From figure 1, it is possible to notice that starting from the third principal component, the variance explained by each component decreases considerably. Therefore, we can conclude that the first three principal components are the most important for explaining the variance in the data.

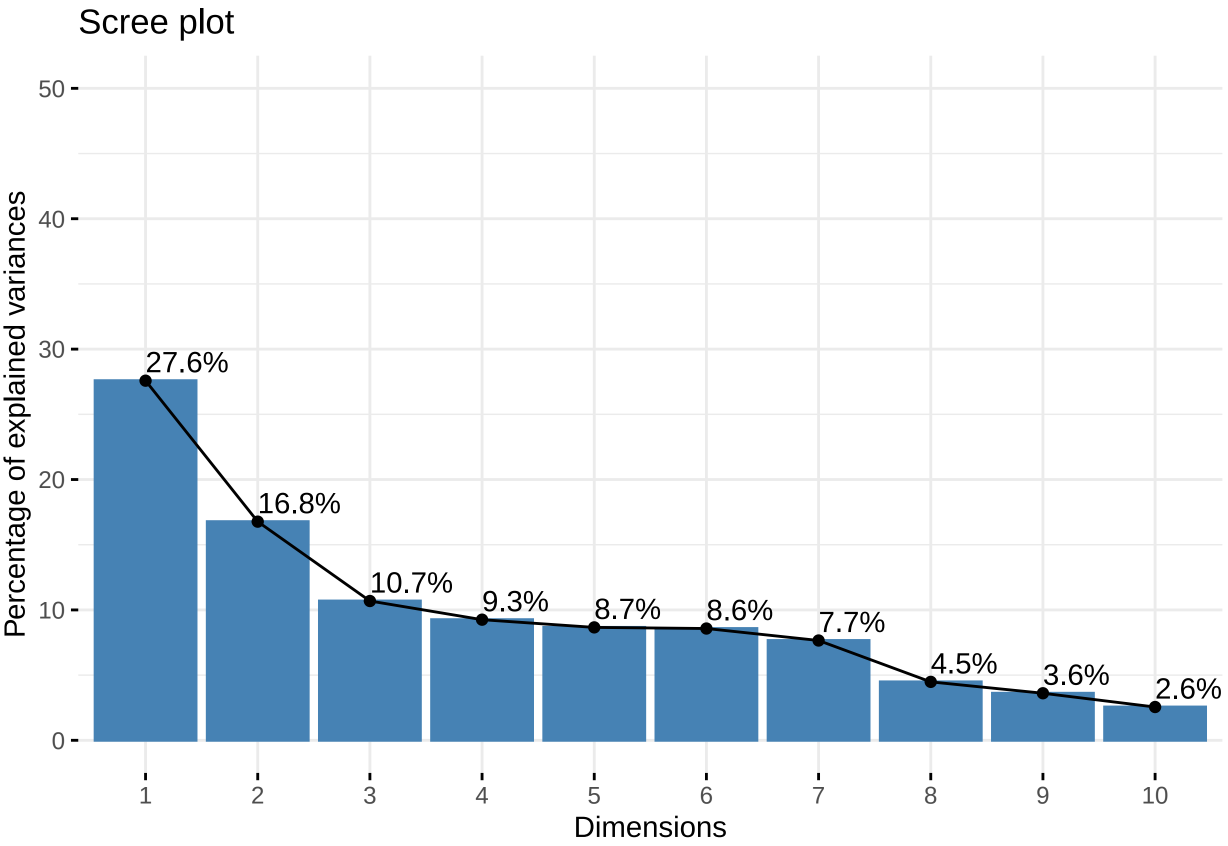


Figure 1: Scree plot - Principal components aggregated explained variance

Table 3 shows the three most influential variables in each principal component. The first principal component is mainly influenced by calories burned, percentage of body fat, and workout duration. The second principal component is mainly influenced by weight, body mass index, and workout duration. The third principal component is mainly influenced by height, body mass index, and daily water intake. To check the effect of each variable on the principal components please refer to table 4 on the annex section of this report and for a visual representation of the contribution of each variable see 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 observation into groups (clusters) such that intra-cluster similarity is high and inter-cluster similarity is low - without using outcome labels. Similarity and dissimilarity is related to how we compare observations and are the foundation of clustering methods. Most clustering methods rely on a dissimilarity matrix between each pair of observation. The dissimilarity is assessed by computing metrics such as euclidean distance

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

The choice of metrics affects cluster shape and euclidean distance typically favors spherical clusters like in K-means. Similarity and dissimilarity indices influence both clustering structure and result interpretation.

K-means is a popular and simple partitioning algorithm that minimizes the within-cluster sum of squares iteratively. It is performed in the next sequence of steps: 1. Choose the K number of clusters; 2. Initialize K centroids; 3. Assign each observation to the closest centroid; 4. Recompute centroids as the mean of all assigned points; 5. Repeat steps 3-4 until convergence. The output of which can be evaluated using internal validation indices. Using the elbow method we've established that 3 clusters would be the optimal number of clusters, for the purpose of this report, see figure 2. Finally we performed an ANOVA and subsequent Tukey test to assess significant differences between variables within clusters. To see the Tukey test results for each significant variable refer to the annex section of this report.

Mixture models generalize clustering by assuming that data is generated from a probabilistic mixture of distribution each representing a cluster. Given the overall distribution of our data being reasonably symmetric we have also used a gaussian mixture model. 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 method is a limiting case of GMMs when all $\sigma_k = \sigma^2 I$ and hard assignments are used. We've also used gaussian mixture models clustering in order to compare the results with the k-means, to see the summary of the model please refer to the annex section.

Results and Discussion

By performing the ANOVA followed by the Tukey tests, we were able to see that the variables significant for clustering are: Weight which is significantly different among the 3 clusters; Height which is significantly different among the 3 clusters; Session Duration which is significantly different between clusters 1-2 and between clusters 1-3, with no statistical evidence for a difference between clusters 3-2; Calories Burned which is significantly different among the 3 clusters; Fat Percentage which is significantly different among the 3 clusters; Water Intake which is significantly different among the 3 clusters; Workout Frequency which is significantly different between clusters 1-1 and between clusters 1-3, with no statistical evidence for a difference between clusters 3-2; BMI which is significantly different among the 3 clusters. Cluster 1 is characterized by heavier, taller individuals that subsequently have higher BMI and consume more water per workout. While

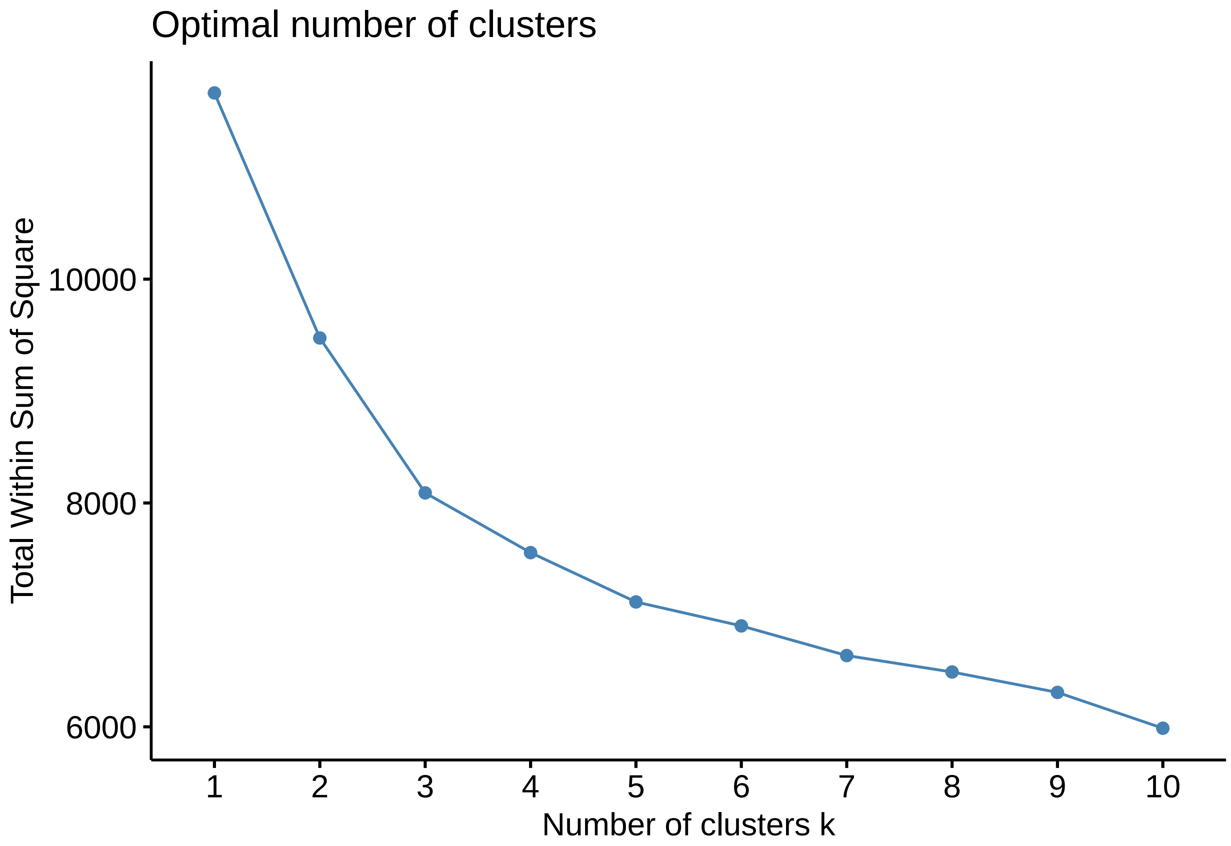
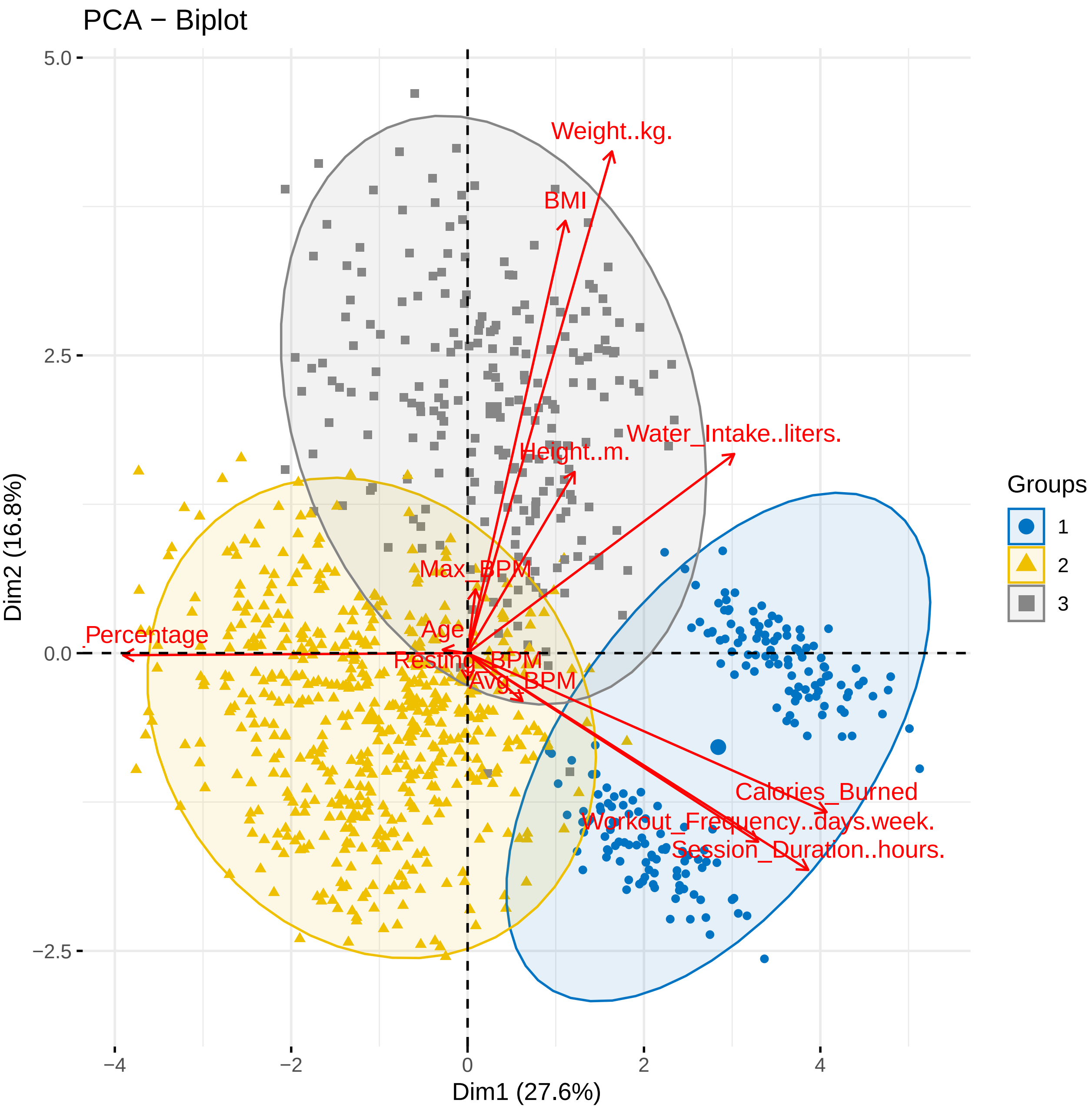


Figure 2: Assessing the optimal number of clusters from the total within sum of squares

Cluster 2 encompasses the majority of gym members and is mostly associated with older individuals with higher body fat percentage. Finally Cluster 3 encompass the high performance group of individuals who workout more frequently and during longer periods and subsequently burn more calories. To see a biplot showcasing these relations refer to figure 3.



To evaluate both clustering solutions we used silhouette indexes the plots of which can be seen in figure 4. For a more in-depth evaluation of the differences between clusters in both methods for each variable see figures 21-28 in the annex section. In the k-means clusters the most cohesive structure was found on cluster 3 while clusters 1 and 2 have flatter or declining silhouettes, which translates to the boundary between clusters being more ambiguous. In GMM clusters the most cohesive structure was found on cluster 1 in both methods cluster 2 performed the weakest. The average silhouette widths are slightly lower in the k-means cluster than in GMM's best cluster, but the cluster sizes are more balanced.

```
##   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  191          0.29
## 2       2  559          0.18
## 3       3  223          0.15
```

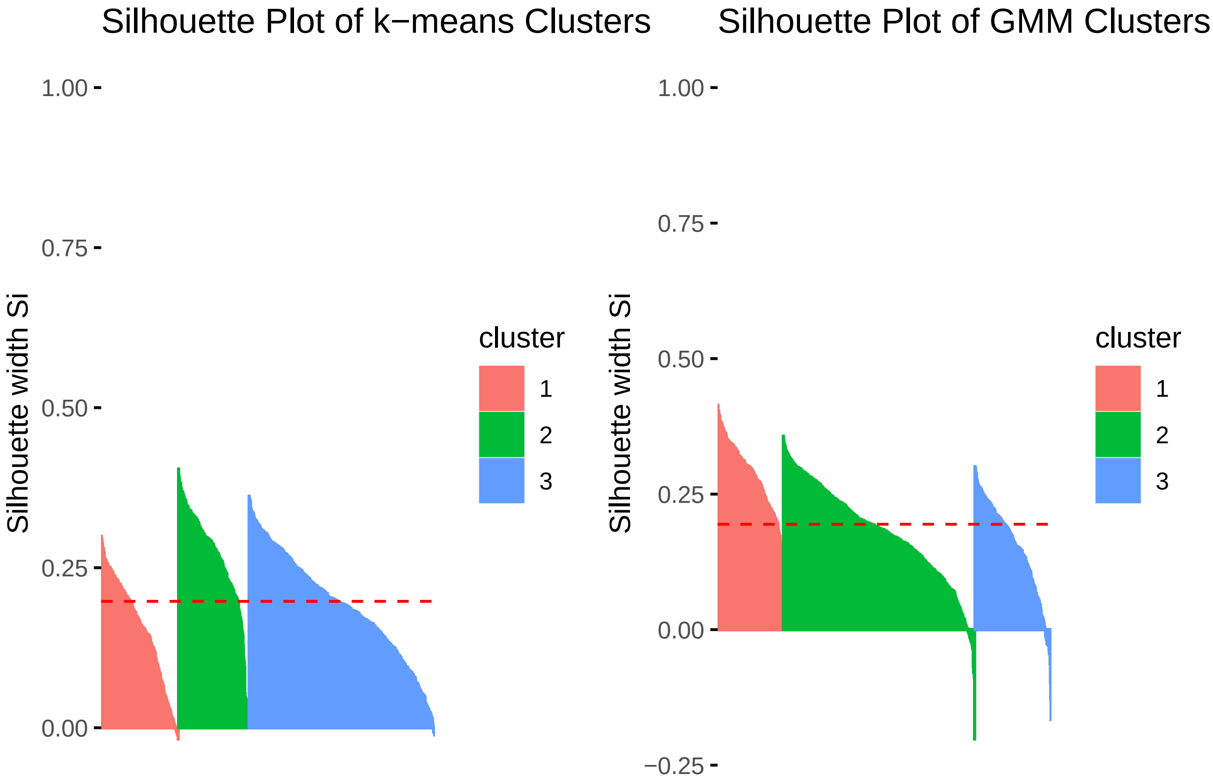


Figure 3: Comparison between silhouette plots of K-means Cluster (left) and GMM Clusters (right)

References

Annex

Exploratory univarate data analysis

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
Max_BPM	0.011	0.081	-0.106
Avg_BPM	0.073	-0.061	-0.255

	PC1	PC2	PC3
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       |
## |3-2 |3-2        |     -457.81     | [-496.88, -418.74]  |      0.000       |
##
```

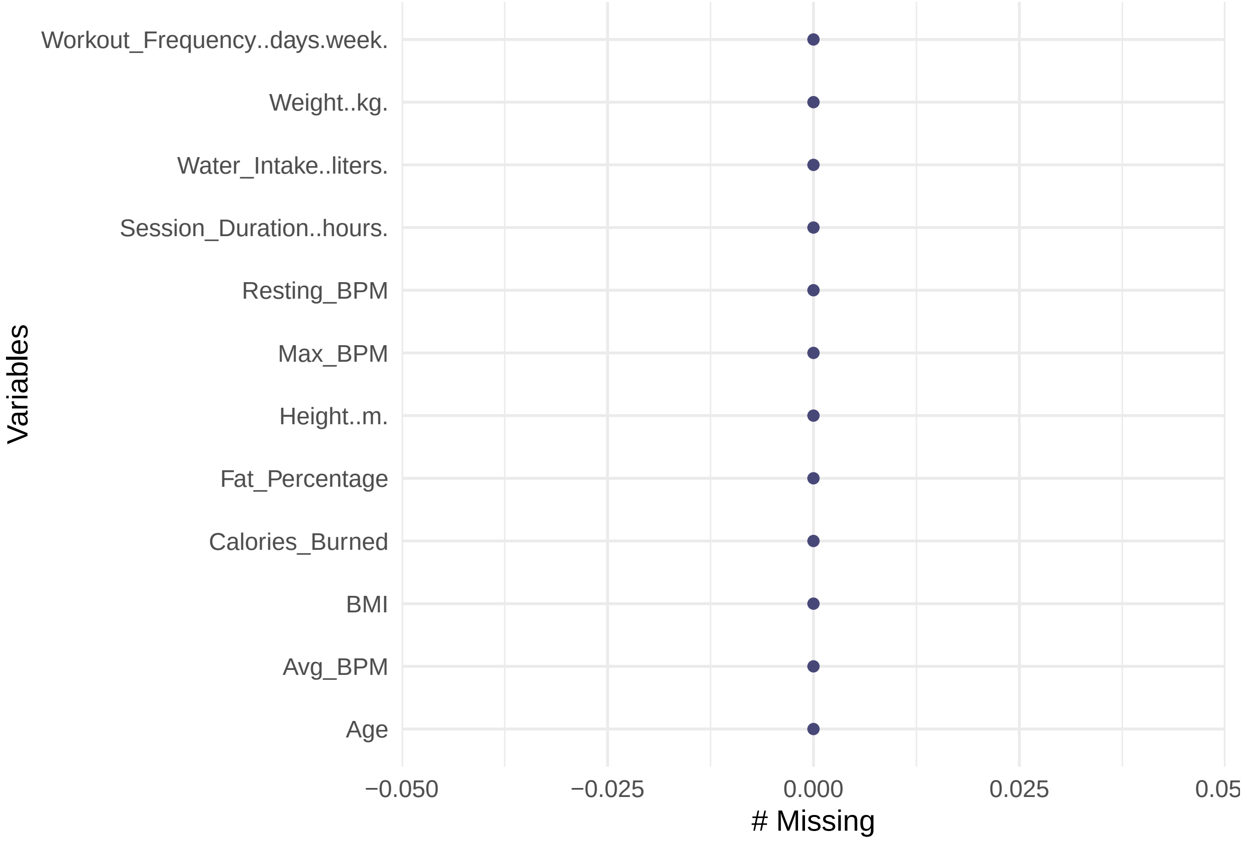


Figure 4: There are no missing values on this dataset.

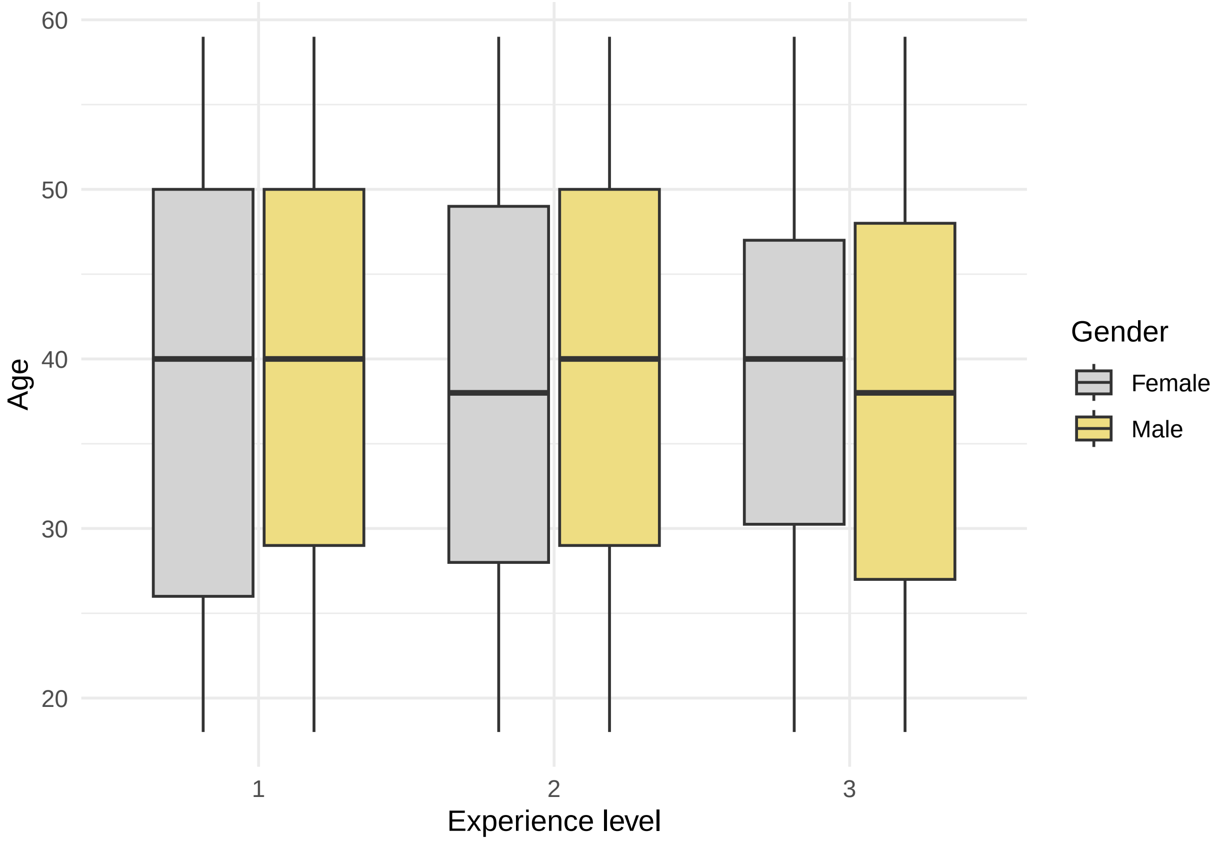


Figure 5: On average gym members are around 40yo. Expert level male member are on average younger than their female counterparts while the opposite is true for intermediate level members. Note that experience levels are organized from 1 - Beginner to 3 - Expert.

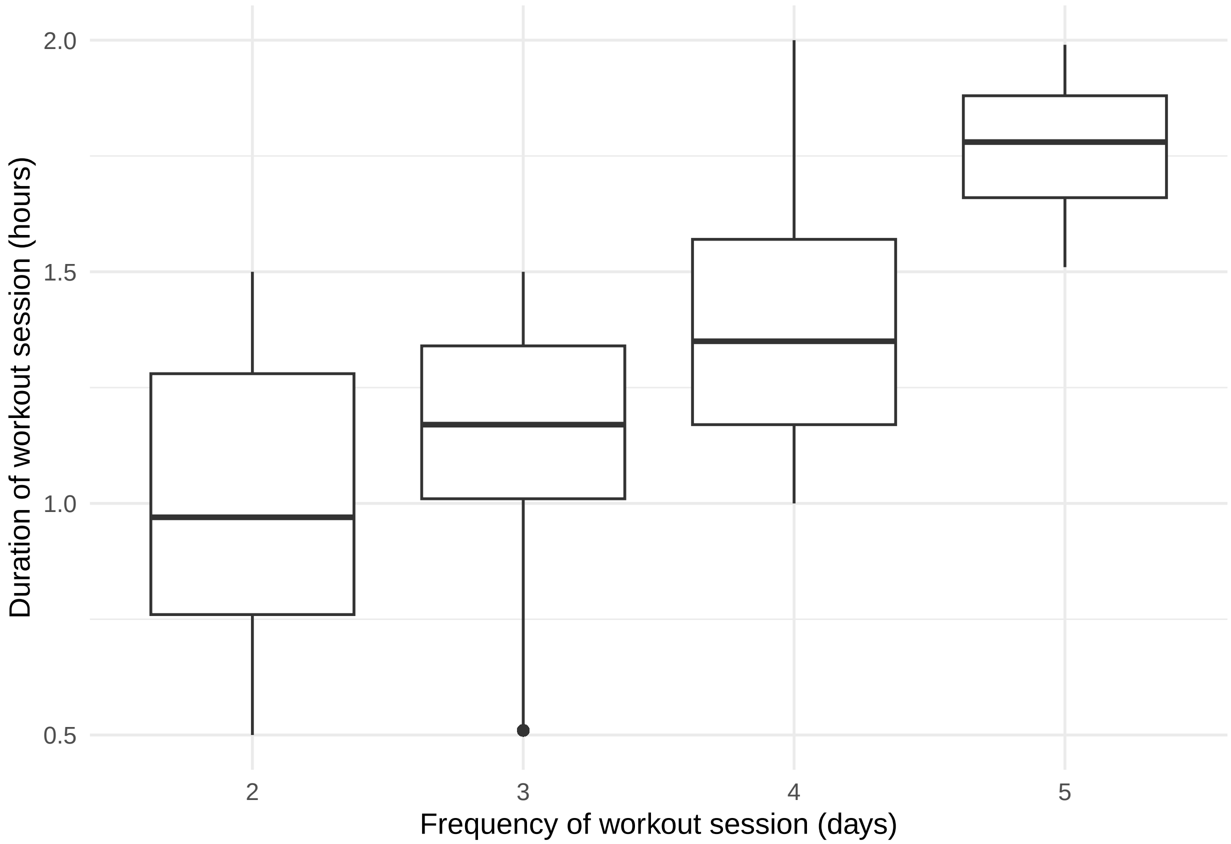


Figure 6: Members that workout with more frequency 4 to 5 days per week usually have longer workout sessions.

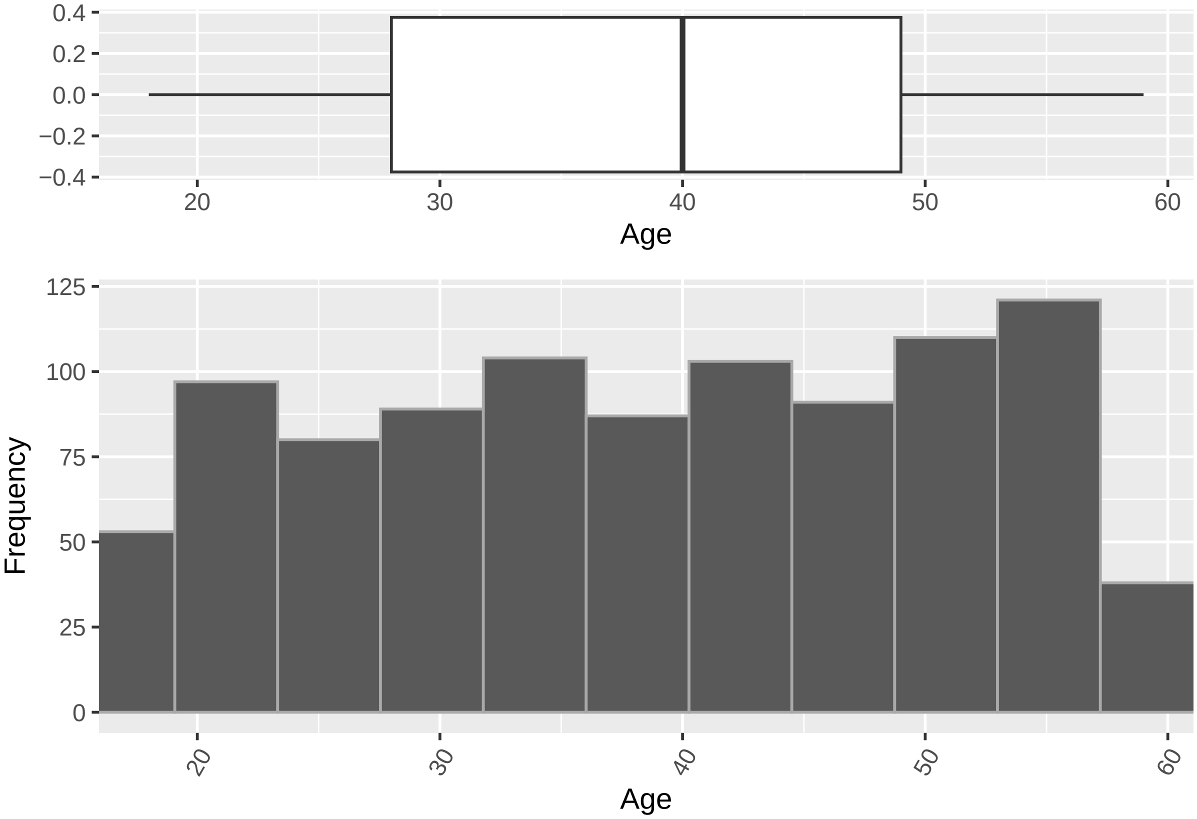


Figure 7: Histogram and Boxplot of Gym member's Age

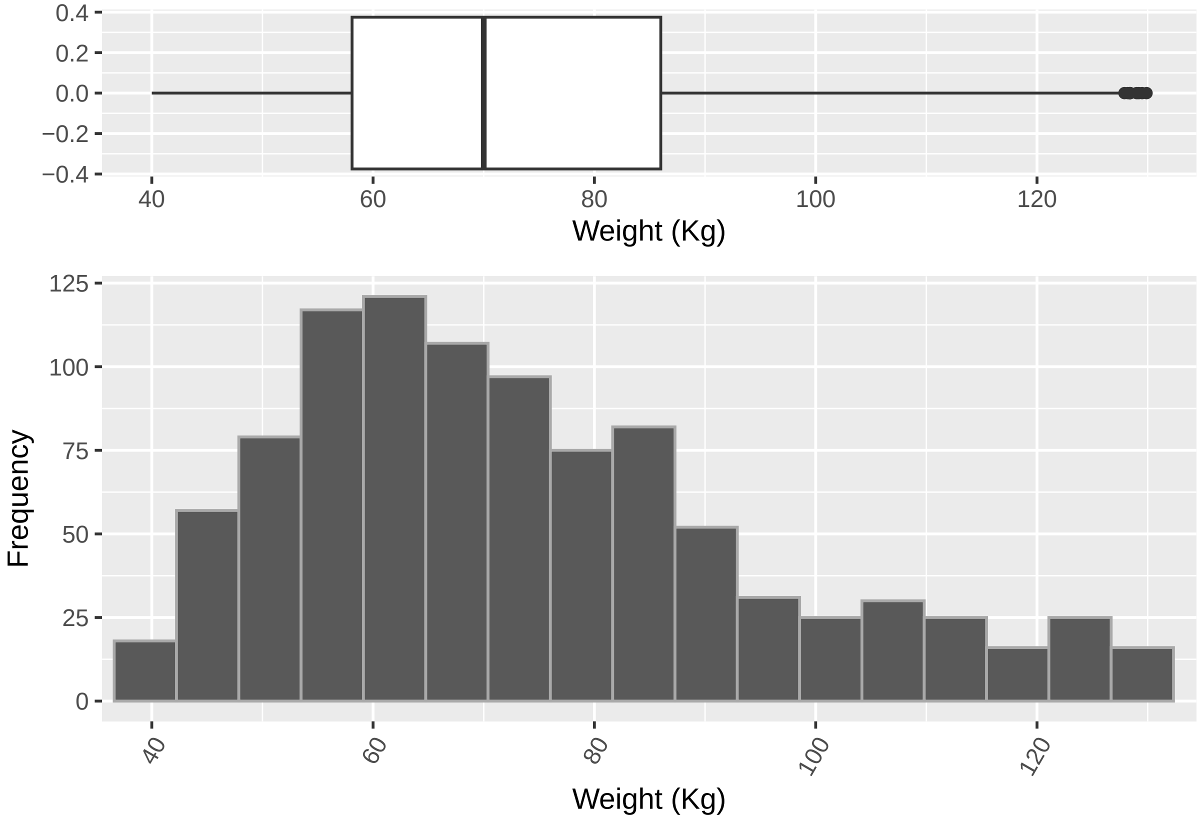


Figure 8: Histogram and Boxplot of Gym member's Weight

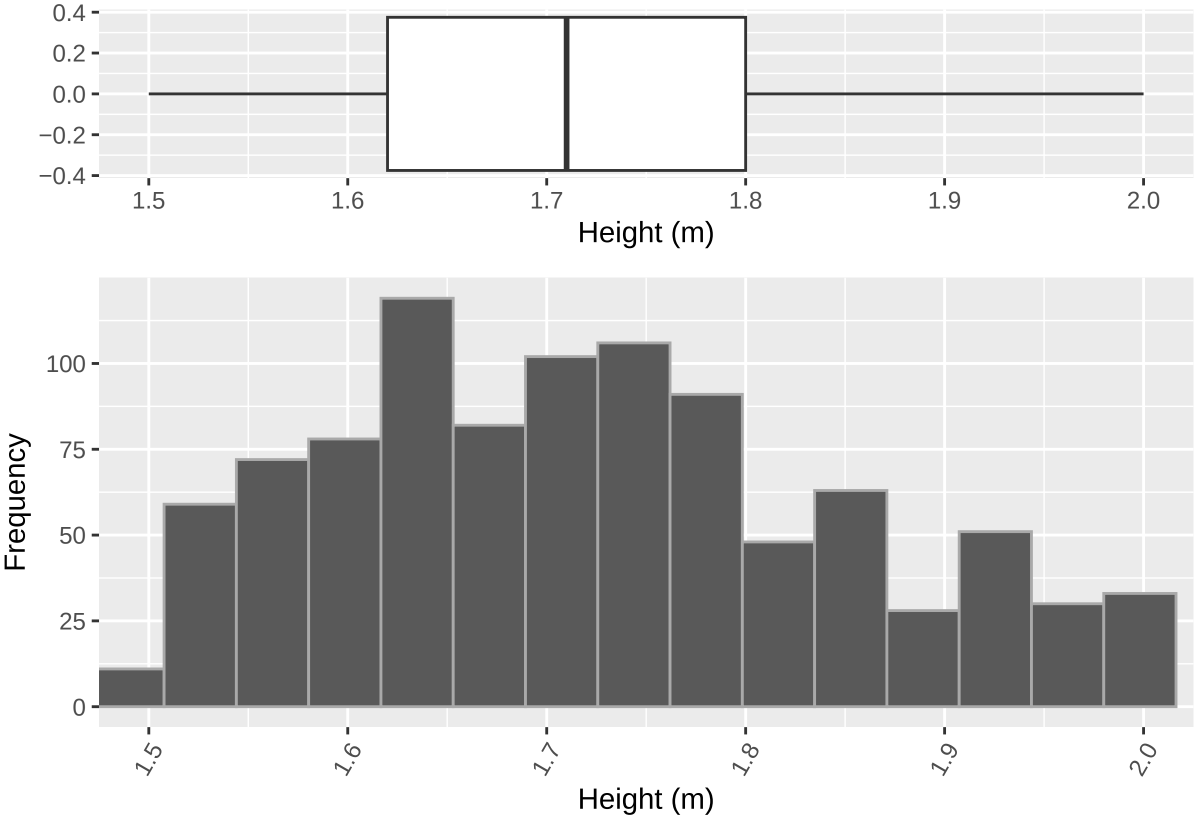


Figure 9: Histogram and Boxplot of Gym member's Height

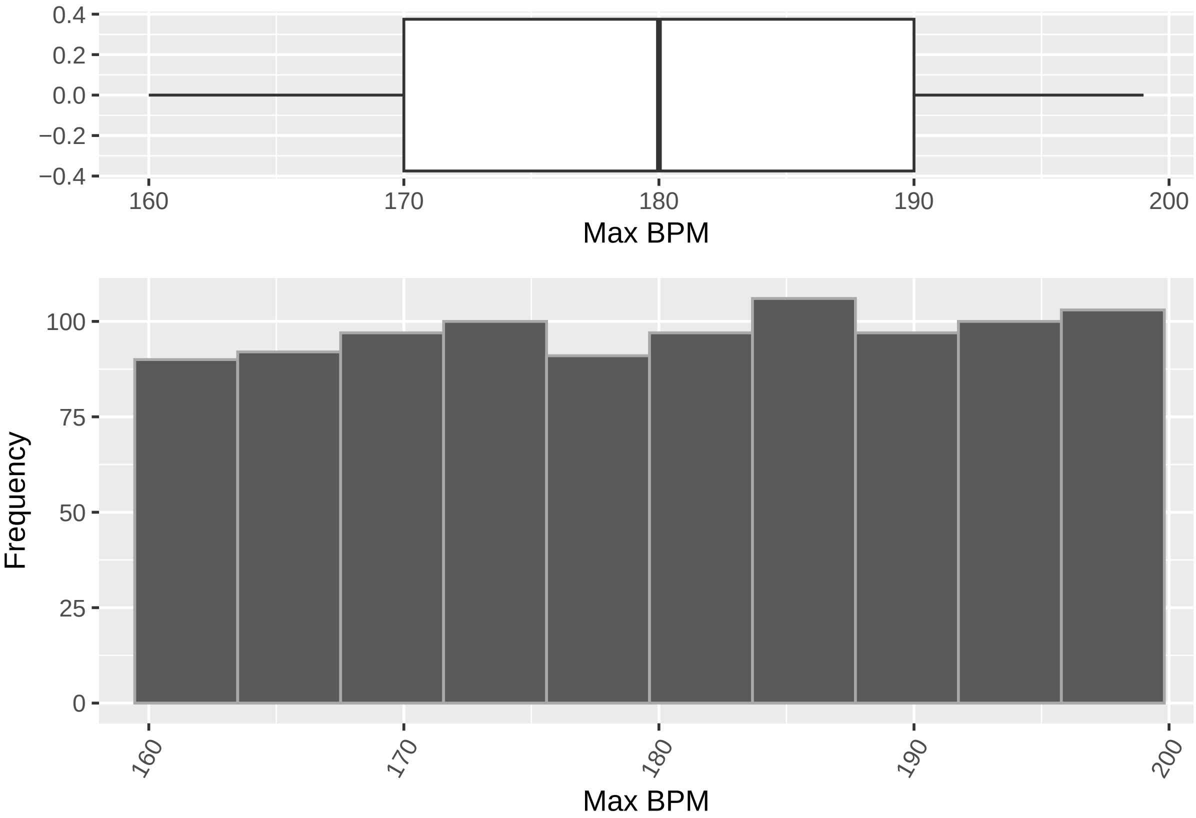


Figure 10: Histogram and Boxplot of Gym member's Maximum heart rate

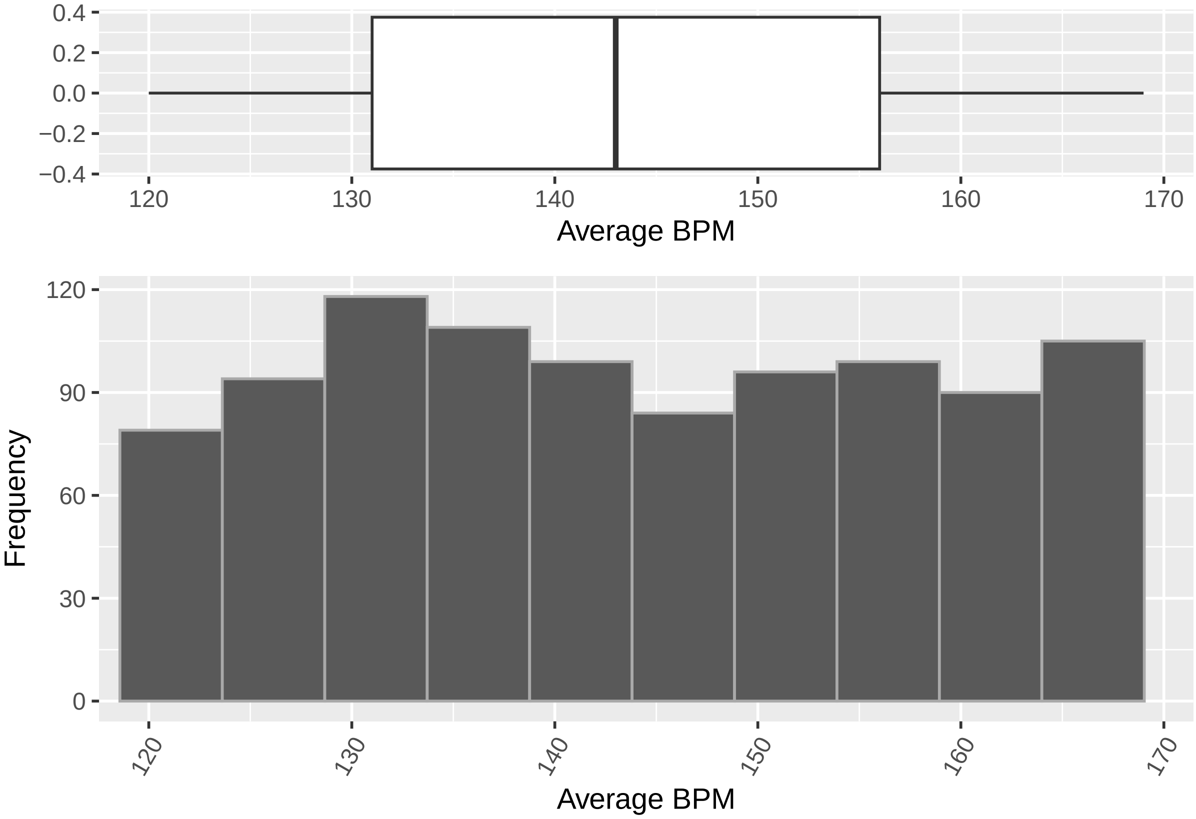


Figure 11: Histogram and Boxplot of Gym member's Average heart rate

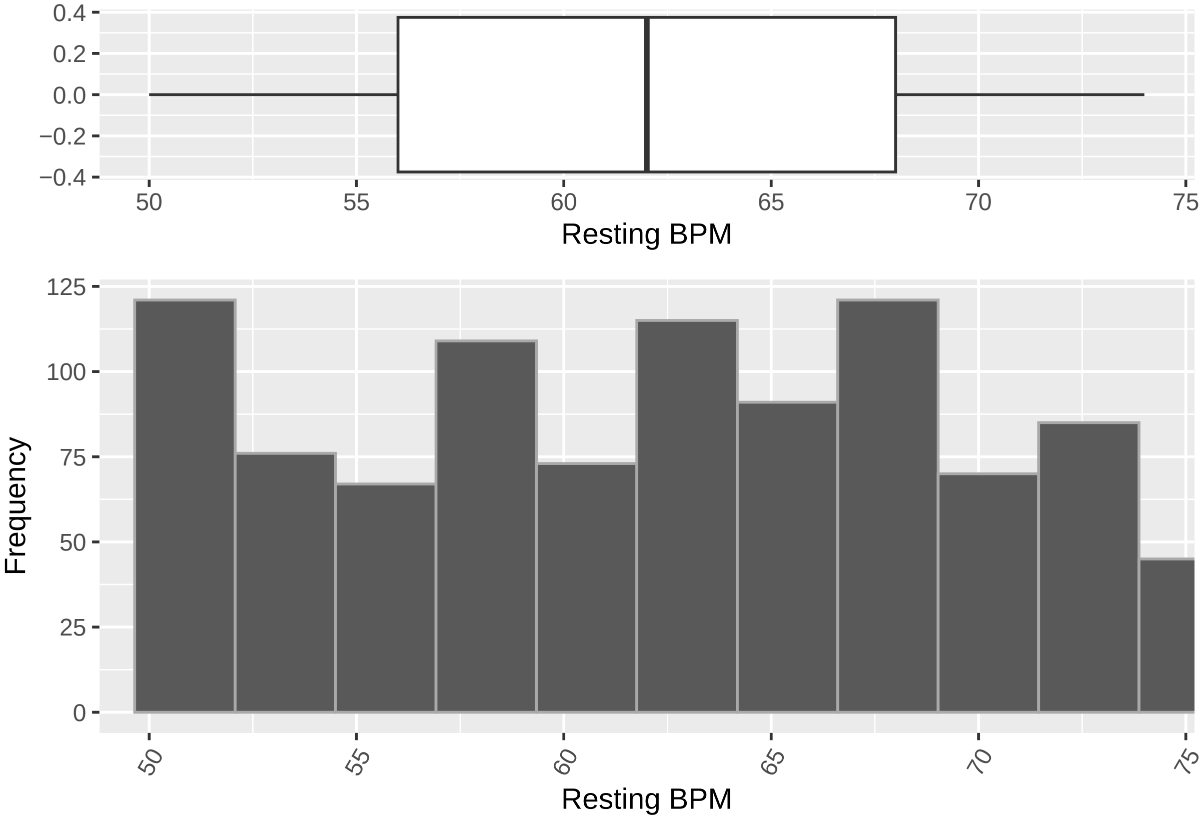


Figure 12: Histogram and Boxplot of Gym member's Resting heart rate

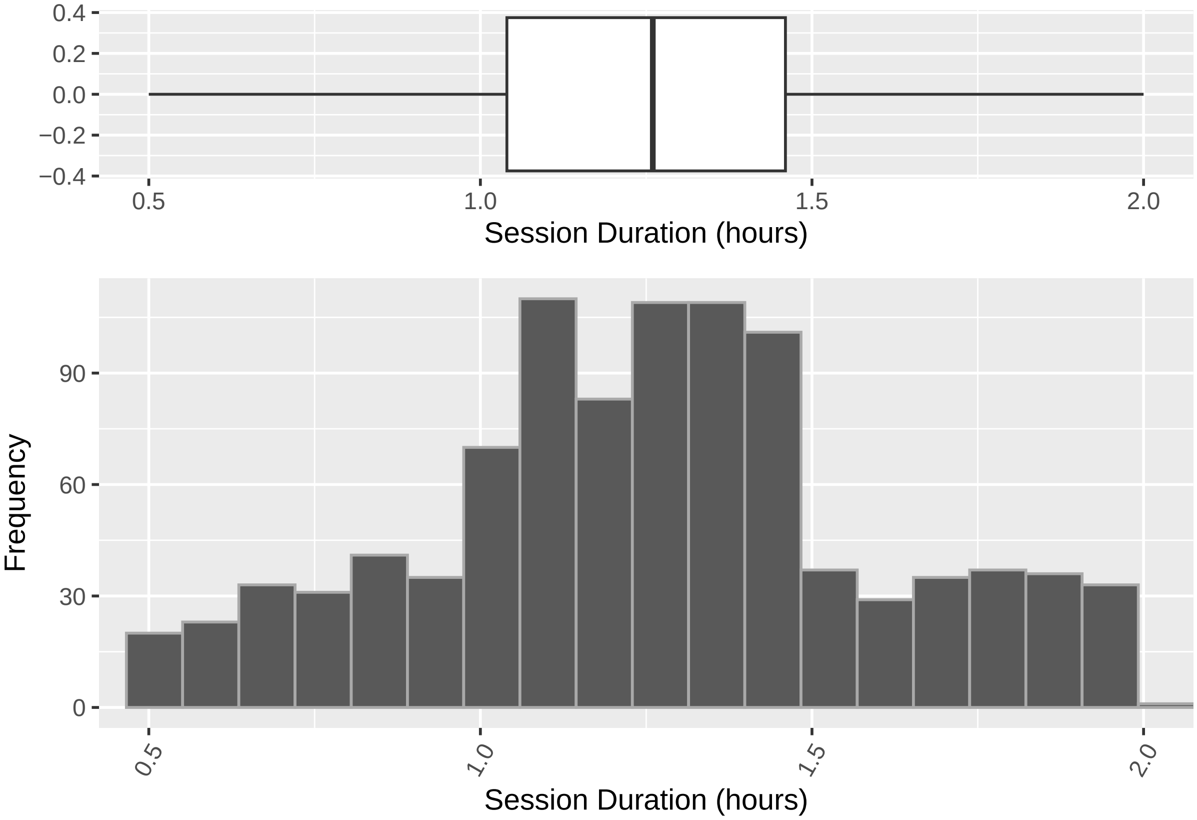


Figure 13: Histogram and Boxplot of Gym member's Training Sessions duration

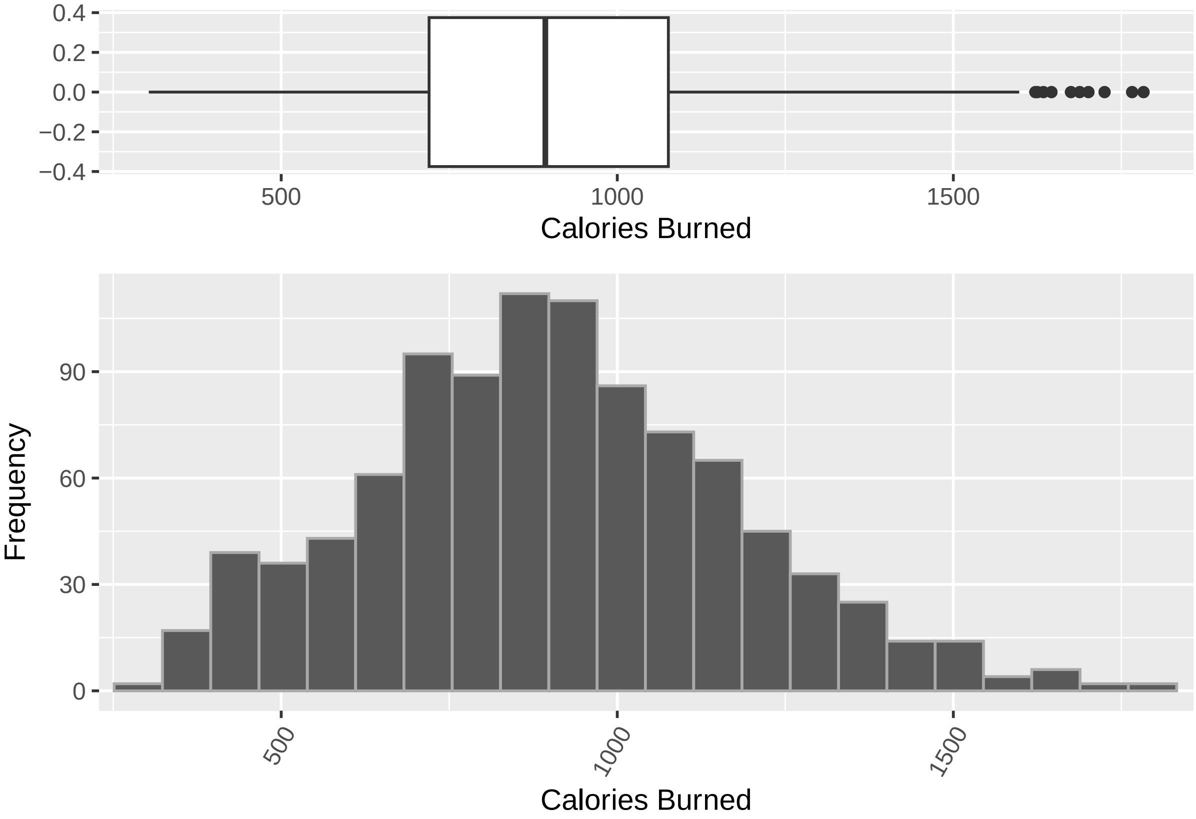


Figure 14: Histogram and Boxplot of Gym member's Calories burned during each session

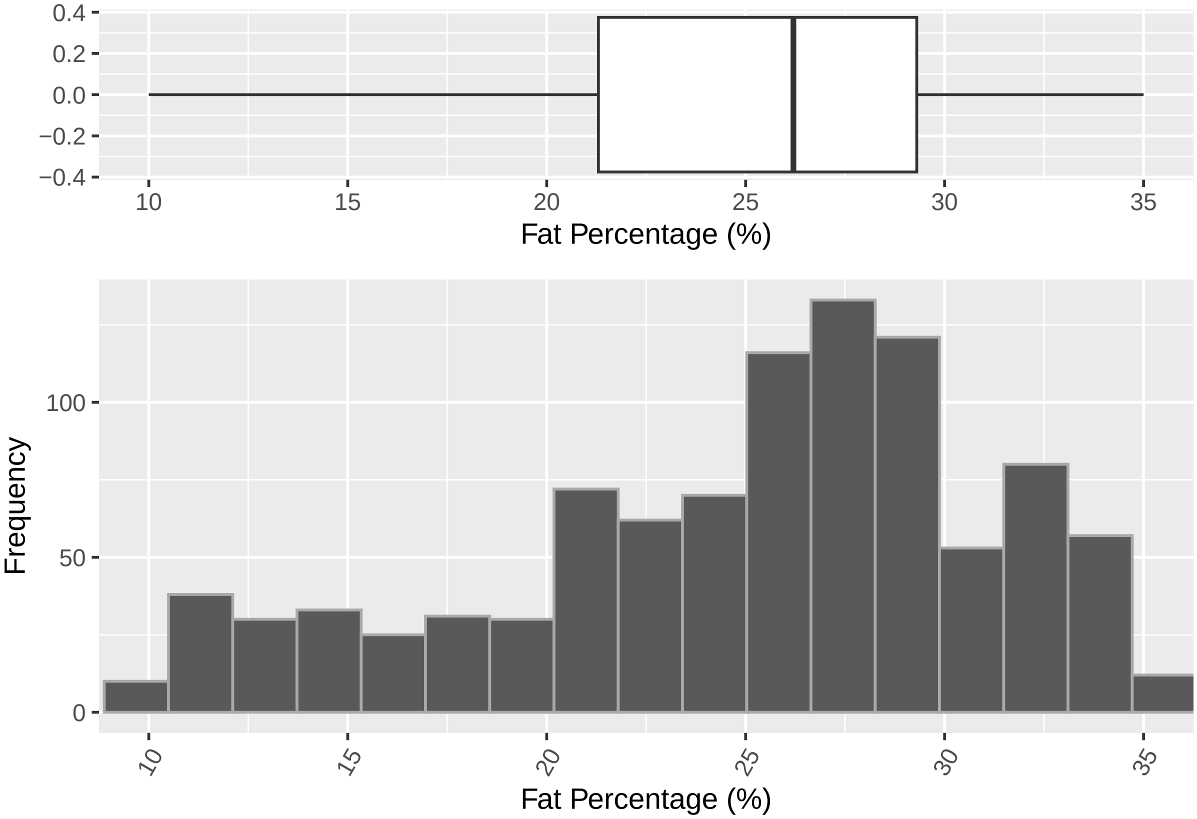


Figure 15: Histogram and Boxplot of Gym member's body fat percentage

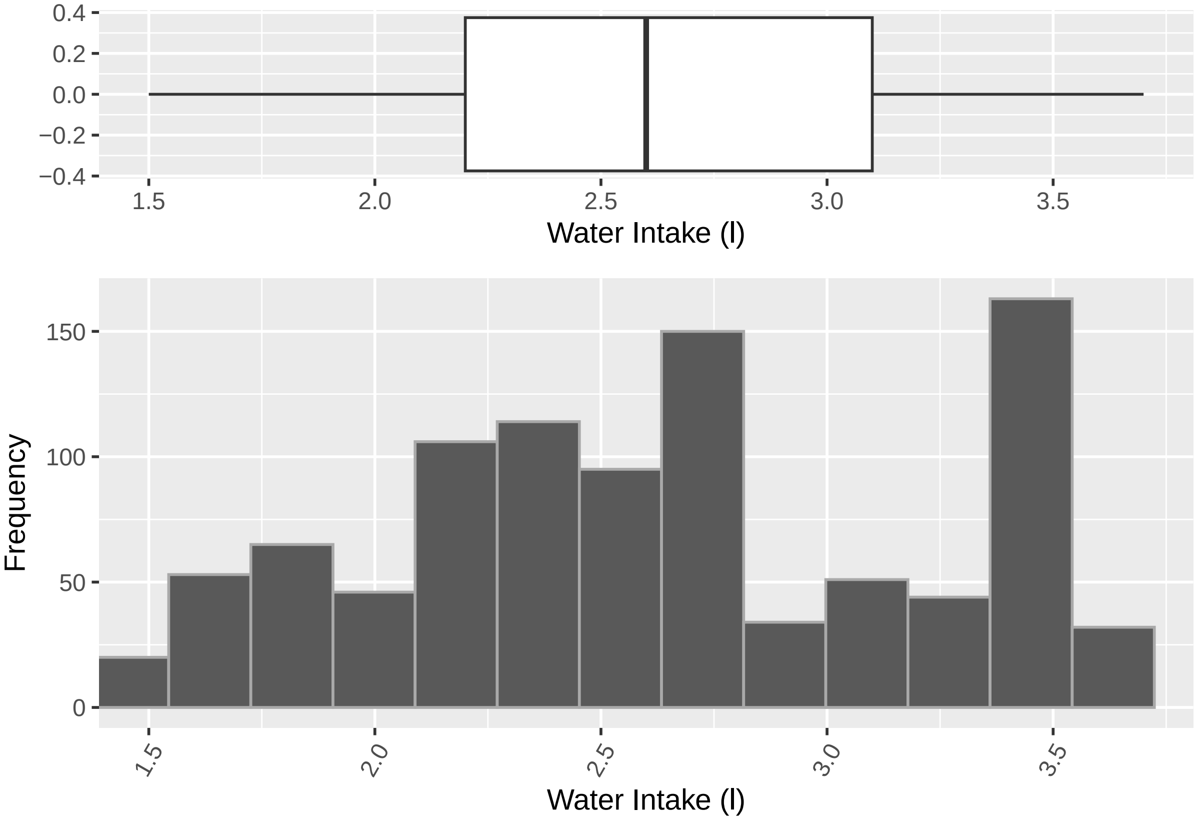


Figure 16: Histogram and Boxplot of Gym member's daily water intake during workouts

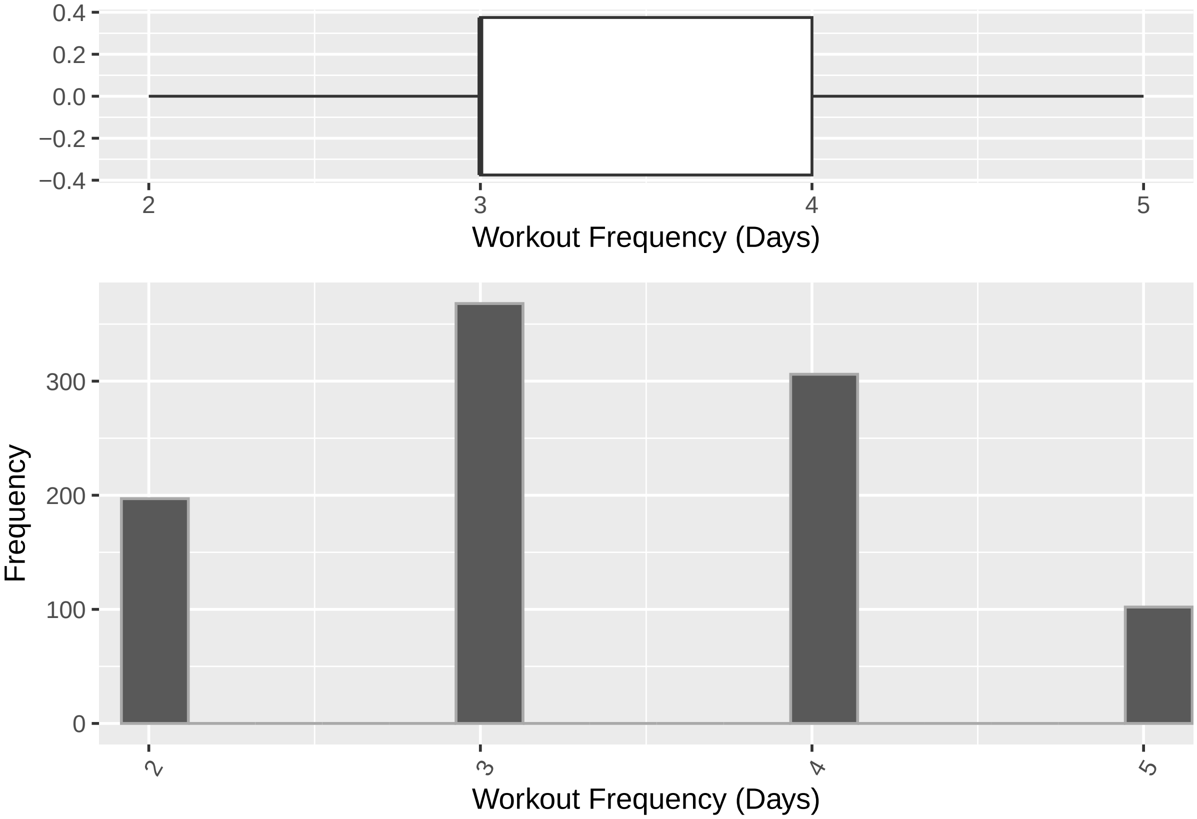


Figure 17: Histogram and Boxplot of Gym member's number of workout sessions per week

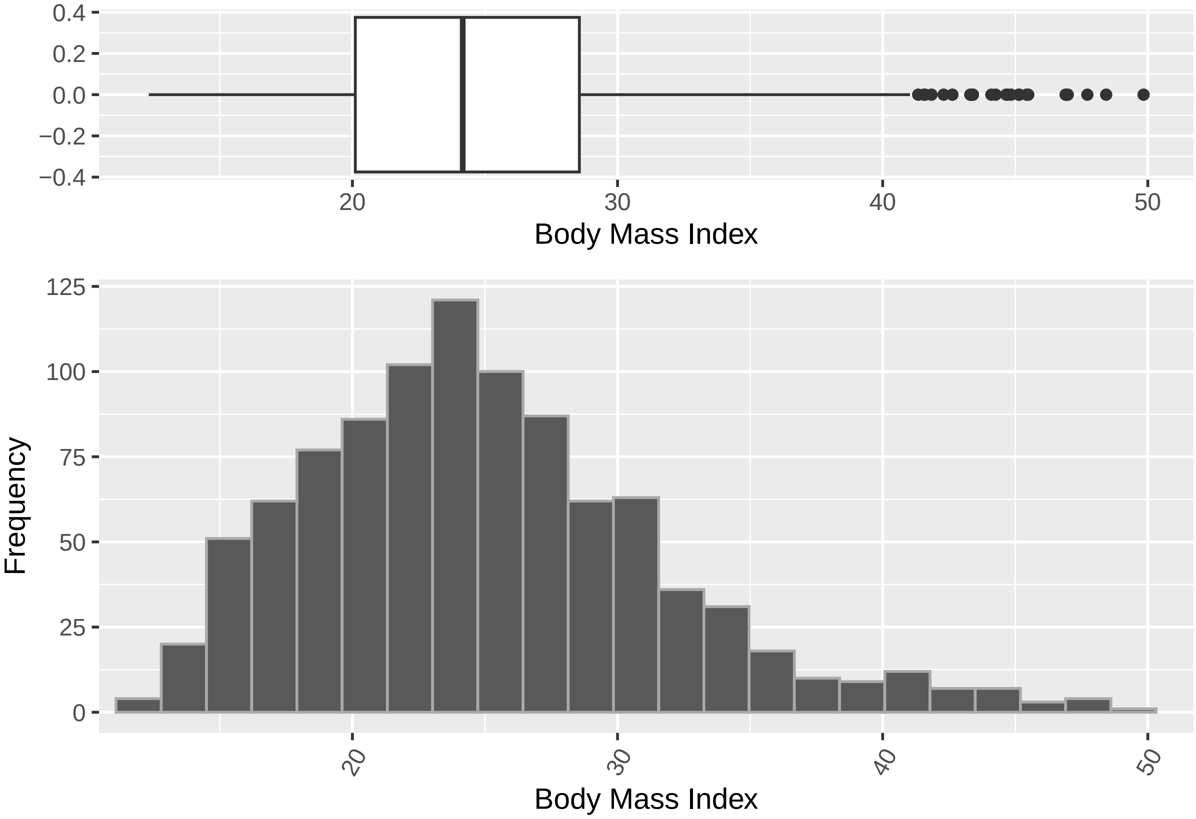


Figure 18: Histogram and Boxplot of Gym member's Body Mass Index

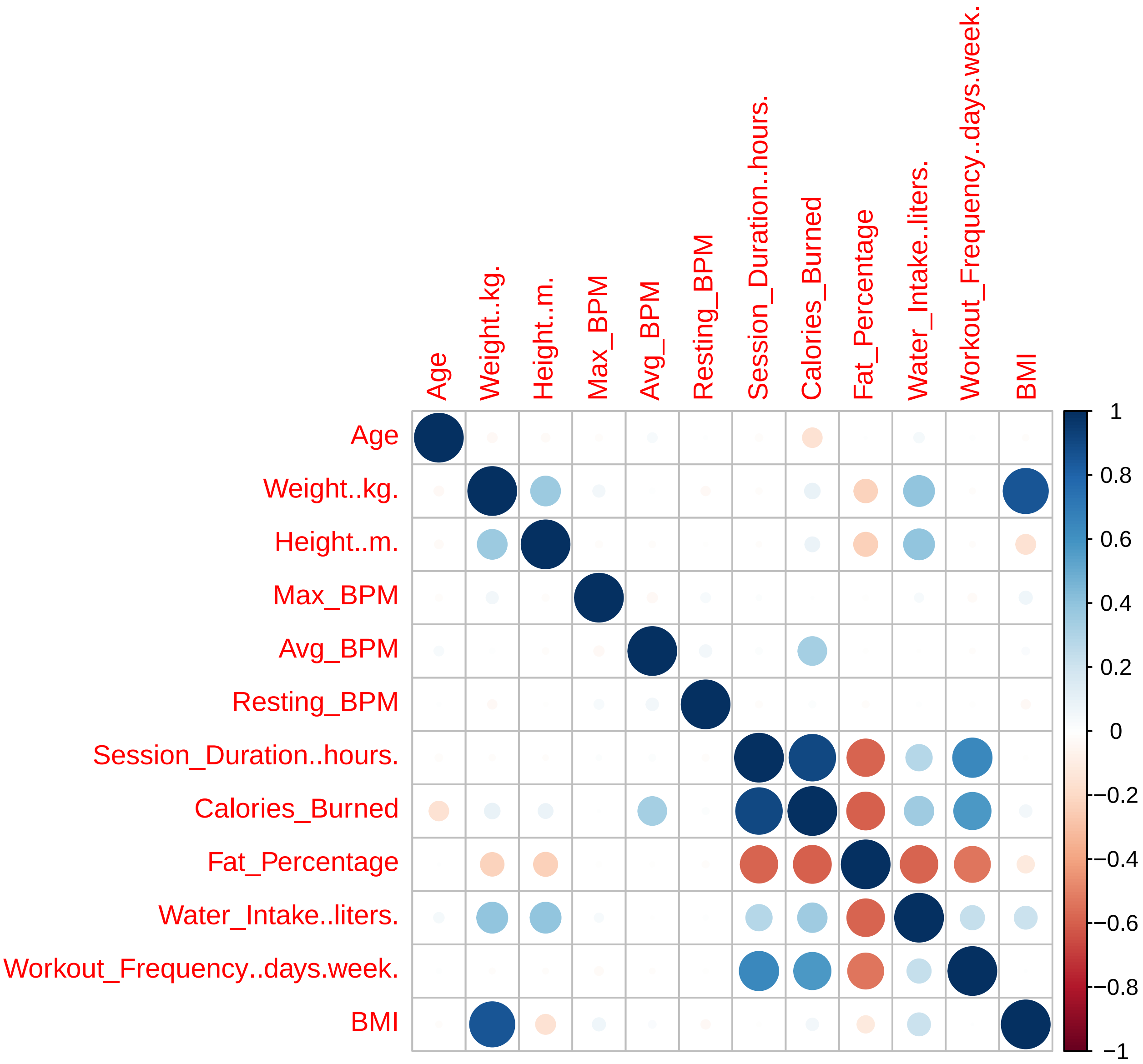


Figure 19: Correlation matrix

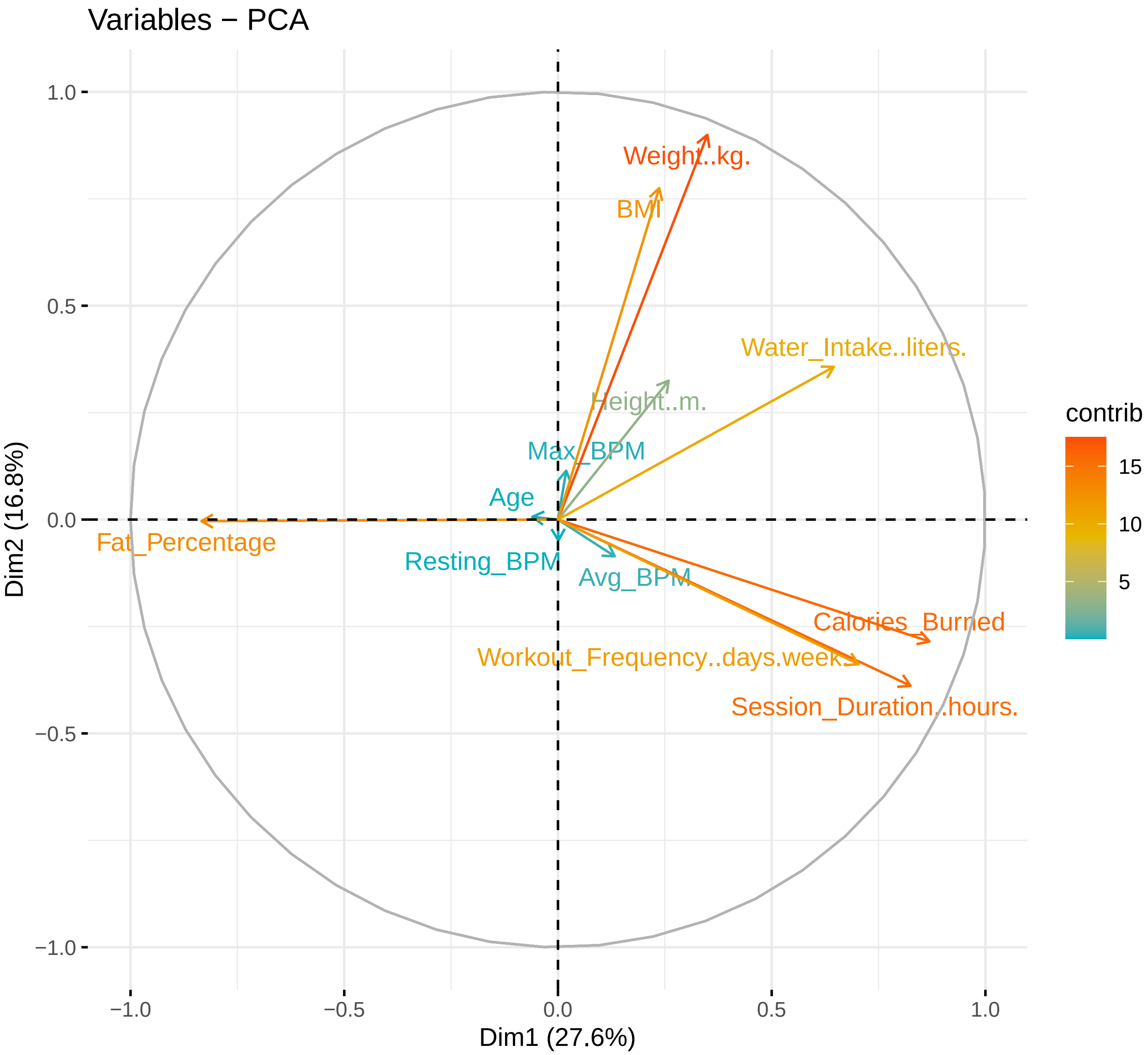
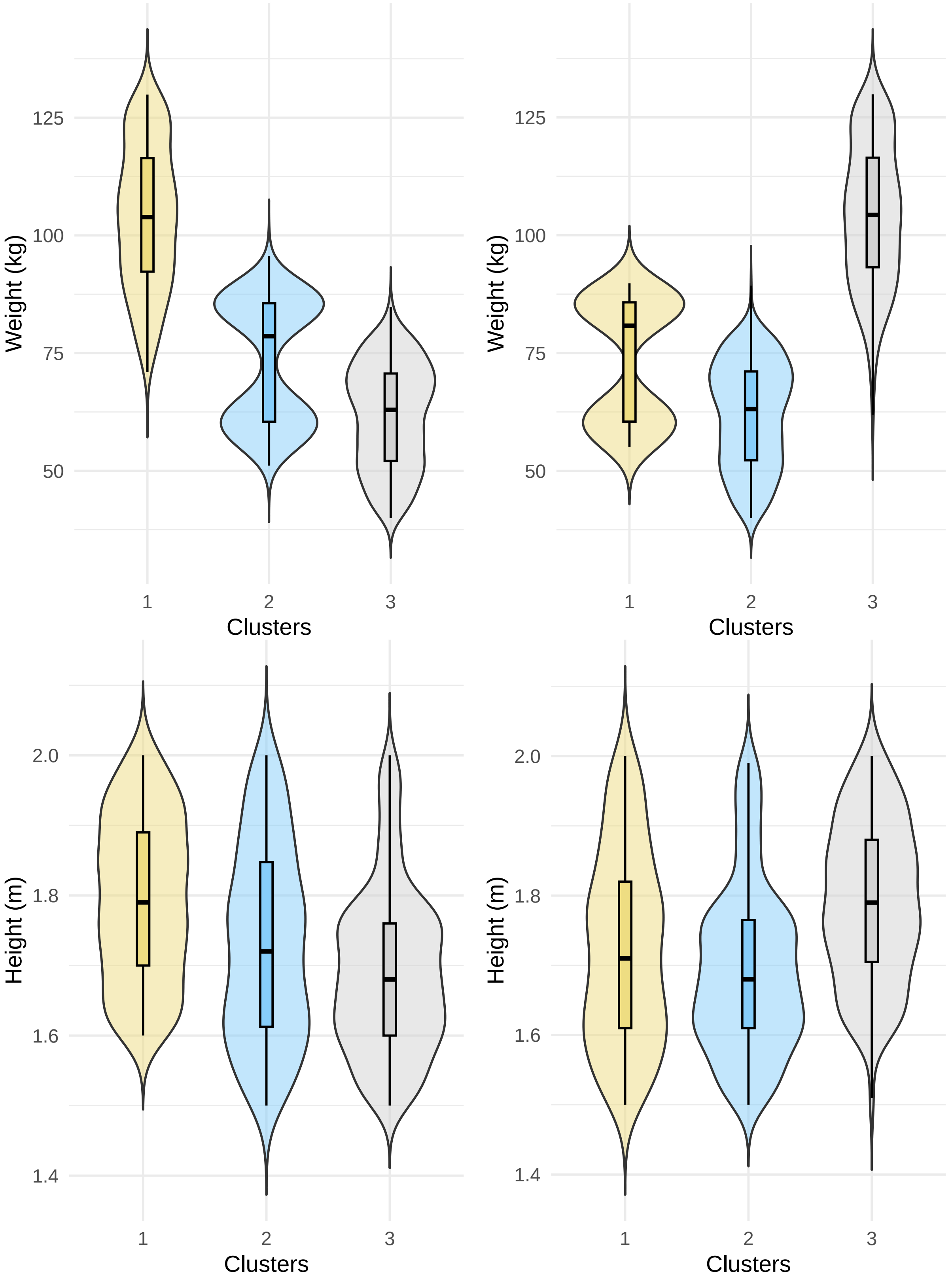


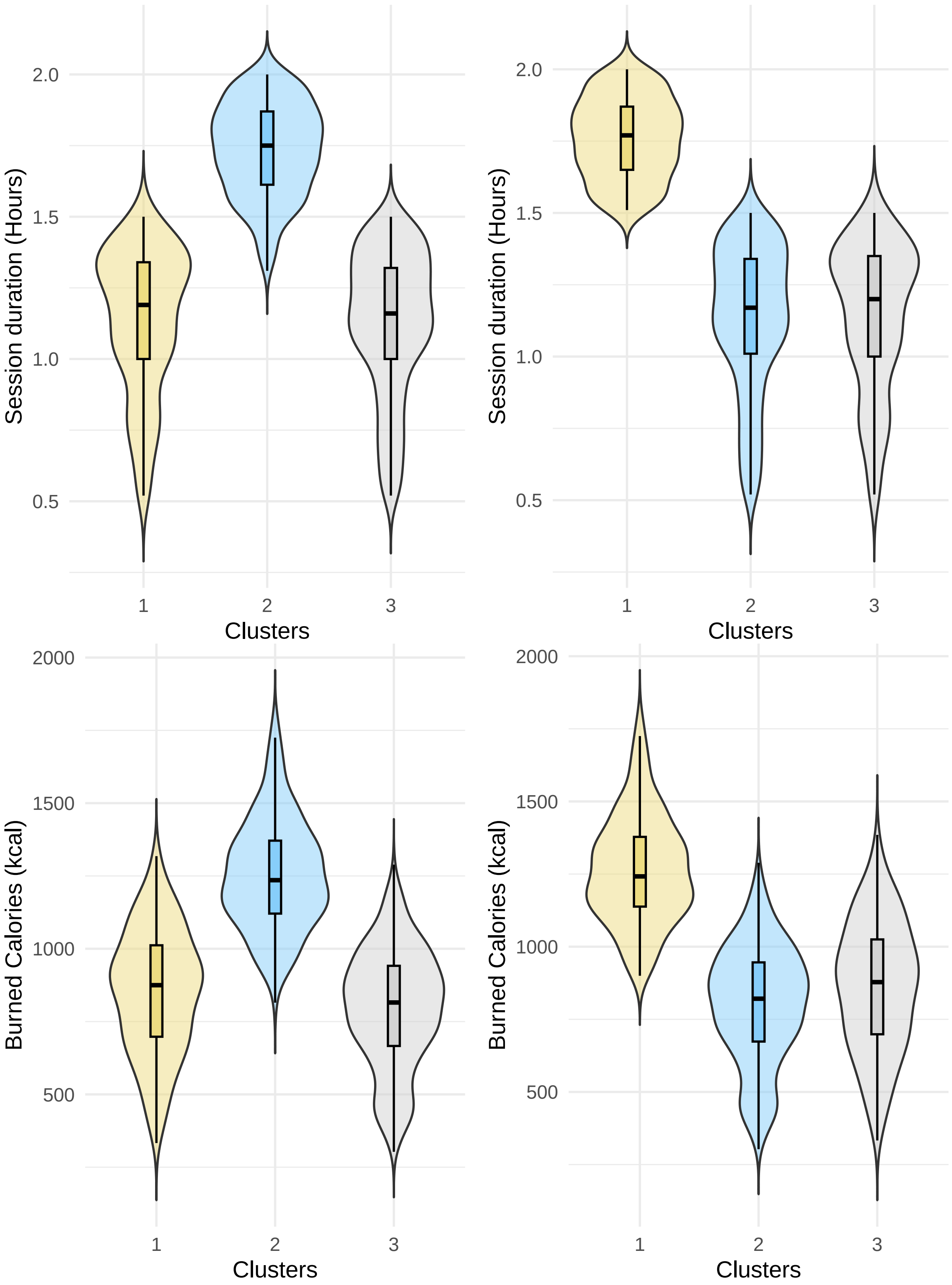
Figure 20: Variables effect on principle components

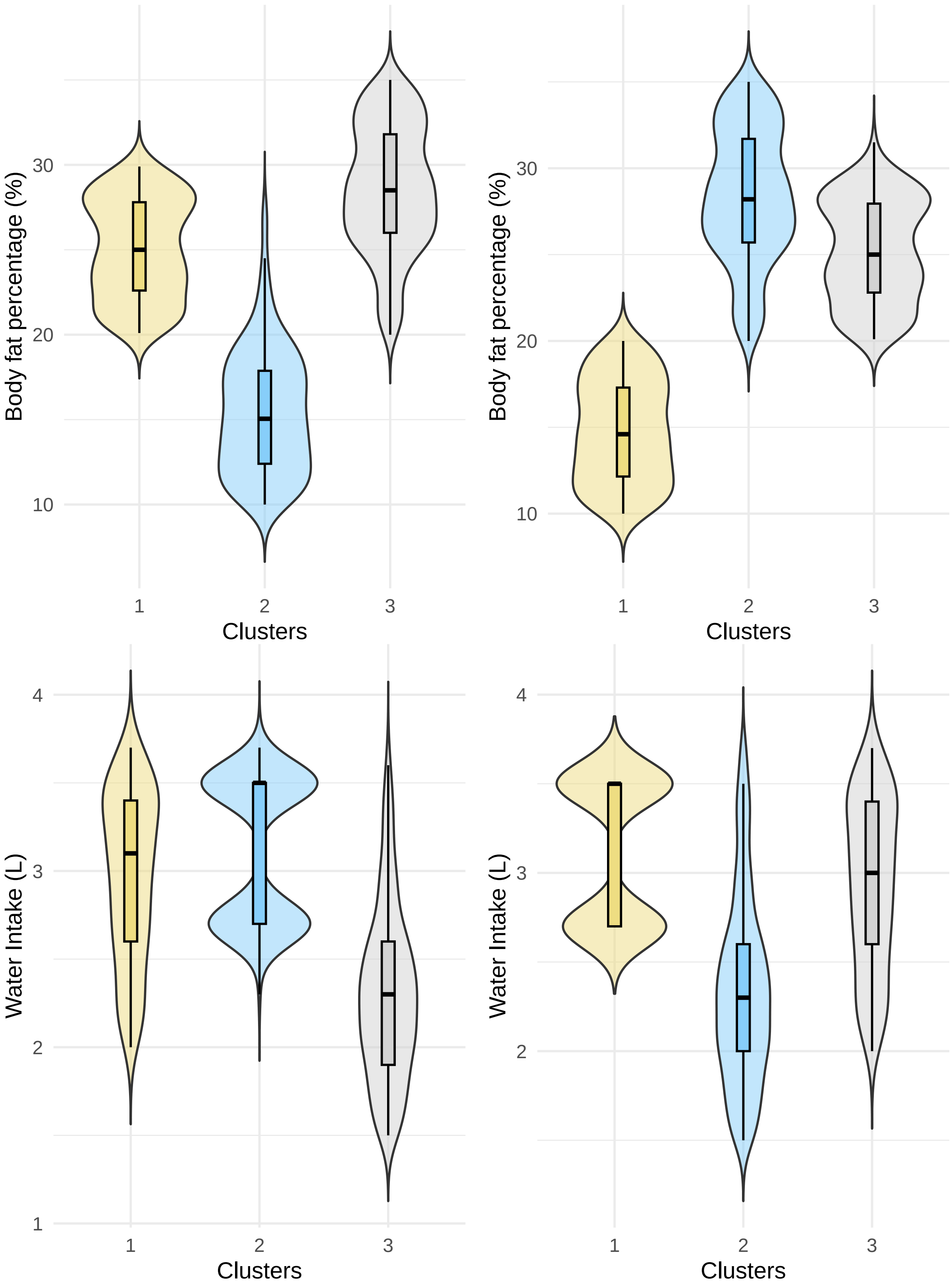
```
## [[5]]
##
##
## Table: Tukey test results for the variable: Mean_Difference (Fat_Percentage)
##
## |    |Comparison | Mean Difference | Confidence Interval | Adjusted p-value |
## |:---|:----------|:---------------:|:-------------------:|:----------------:|
## |2-1 |2-1        |      -9.67      |   [-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         |
```

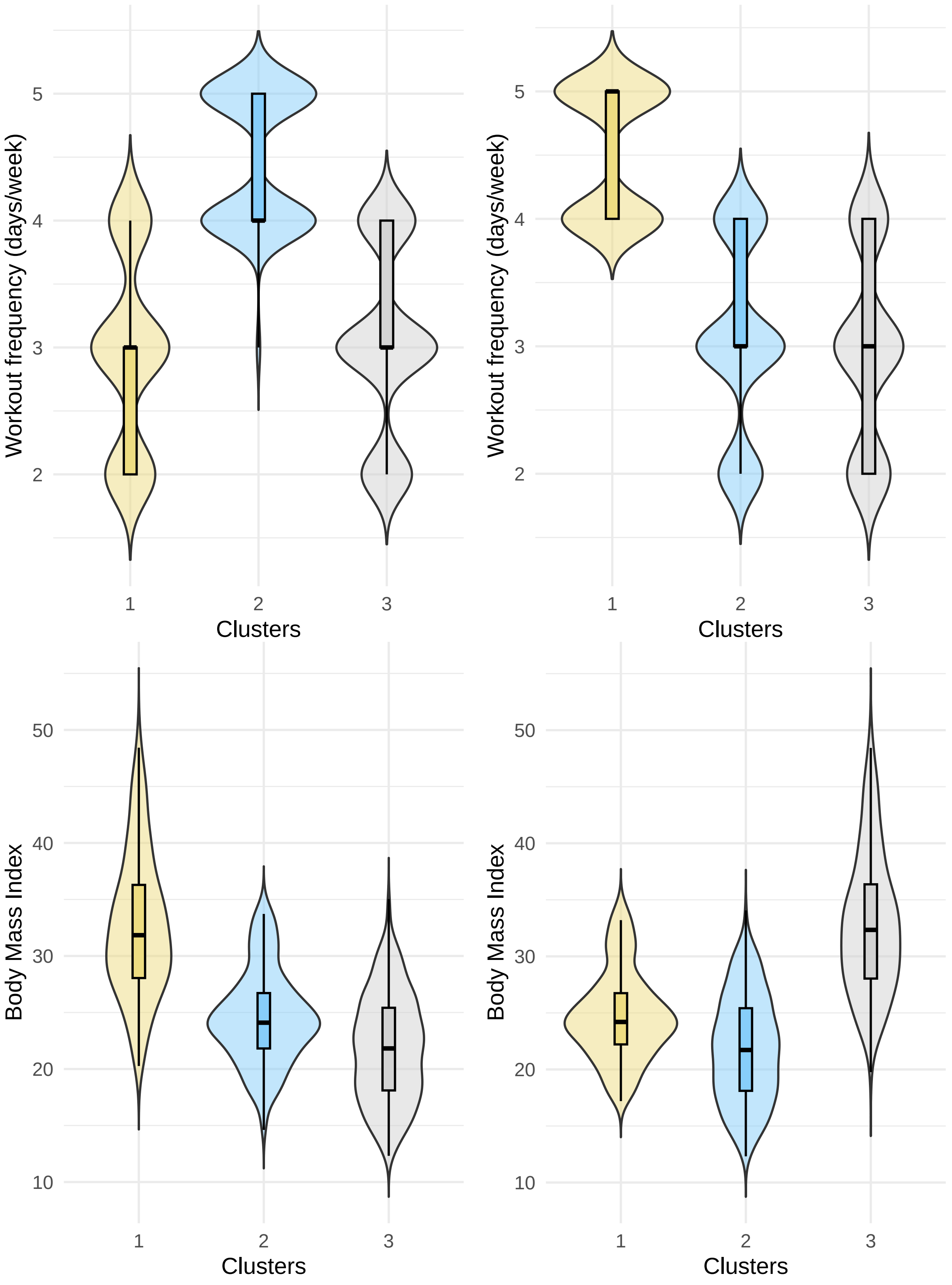
```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 3
## components:
##
##  log-likelihood   n  df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

```
## Clustering table:
##   1   2   3
## 191 559 223
```




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


Body fat percentage (%)
10
20
30
1
2
3
Clusters
Body fat percentage (%)
10
20
30
1
2
3
Clusters
Water Intake (L)
1
2
3
4
1
2
3
Clusters
Water Intake (L)
2
3
4
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