MULTIVARIATE ANALYSIS PROJECT

GYM DATA

DR. MARWA SABRY

Nada Hamouda Mohamed Jamiel

**Introduction**

The gym dataset presents a comprehensive collection of variables related to gym members,

These factors can provide valuable insights into gym members' behaviors and health outcomes. The primary goal of this research is to explore the relationships between these variables and identify key components and patterns that can explain variability in gym members' characteristics. Specifically, we aim to answer three main research questions: (1) What are the principal components that explain the variability in gym members' characteristics, and how can we reduce the number of variables in the dataset while retaining most of the variance? (2) How effectively can Duration of each workout session and maximum heart rate discriminate between gym members who burn high versus low calories during workouts? (3) What are the multidimensional clusters of gym members combining physiological, behavioral, and demographic variables?

We will employ some Multivariate analysis techniques to uncover these insights.

**Data description :**

Age: Age of the gym member.

Gender: Gender of the gym member (Male or Female).

Weight (kg): Member’s weight in kilograms.

Height (m): Member’s height in meters.

Max\_BPM: Maximum heart rate (beats per minute) during workout sessions.

Avg\_BPM: Average heart rate during workout sessions.

Resting\_BPM: Heart rate at rest before workout.

Session\_Duration (hours): Duration of each workout session in hours.

Calories\_Burned: Total calories burned during each session.

Workout\_Type: Type of workout performed (e.g., Cardio, Strength, Yoga, HIIT).

Fat\_Percentage: Body fat percentage of the member.

Water\_Intake (liters): Daily water intake during workouts.

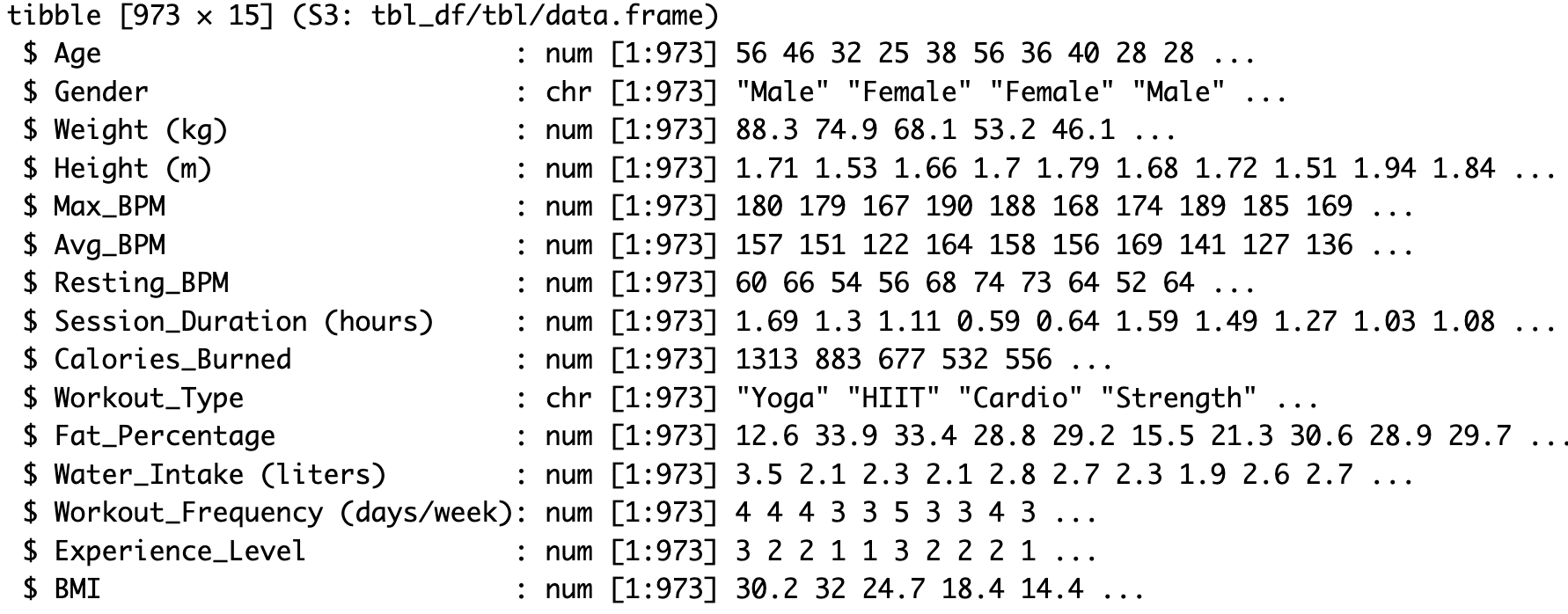
Workout\_Frequency (days/week): Number of workout sessions per week.

Experience\_Level: Level of experience, from beginner (1) to expert (3).

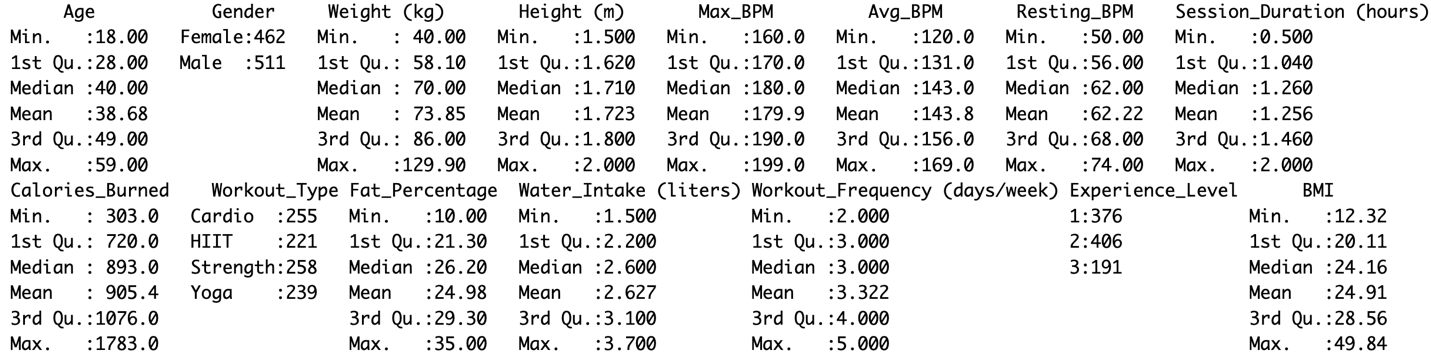
BMI: Body Mass Index, calculated from height and weight.

2) Data Descriptives:

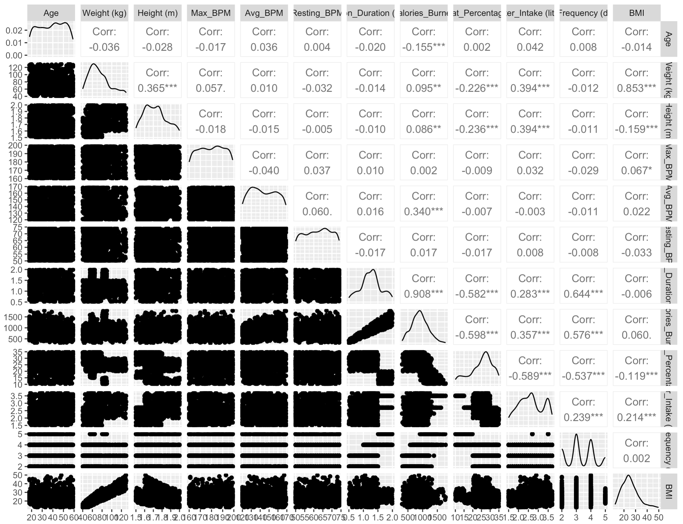
* The data consists of 973 observations,12 quantitative variables and 3 categorical variables.



* A summary statistic for the data contains the minimun,1st quantile,median,mean,3rd quantile and max for each of the quantitative variables , and the number of observations in each level of the factor variables.



* A screenshot of a graph

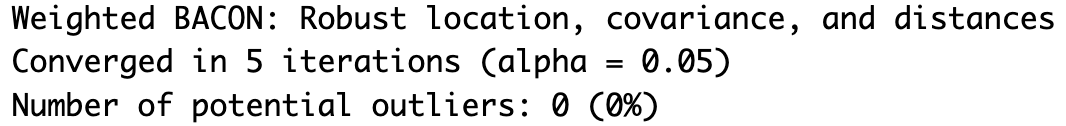
  Description automatically generatedCorrelation Analysis

The matrix shows that BMI and Weight are highly positively correlated suggesting that longer sessions burn more calories as well as the session of duration and the calories burned during the session. While fat percentage shows a moderate negative correlation with each of workout frequency , water intake and session duration.

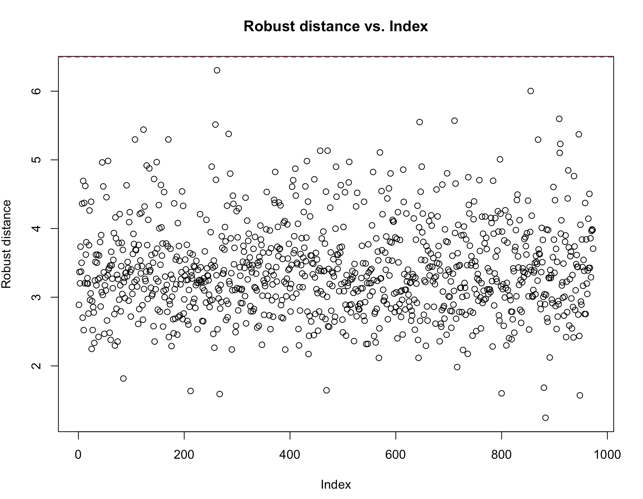
Also testing pearson correlation between work out type and experience level = 0.05135889 , which is also indicates a low relation between the two qualitative variables.

3)**Data Cleaning:**

* There are no missing values
* There are no duplicated observations
* The data consists of 12 quantitative variables and 3 categorical variables of type chr on R which we convert to fct.
* Using bacon method there is no potential outliers, however the boxplot suggests a few potential outliers.



A graph with a line and a box

Description automatically generated with medium confidence

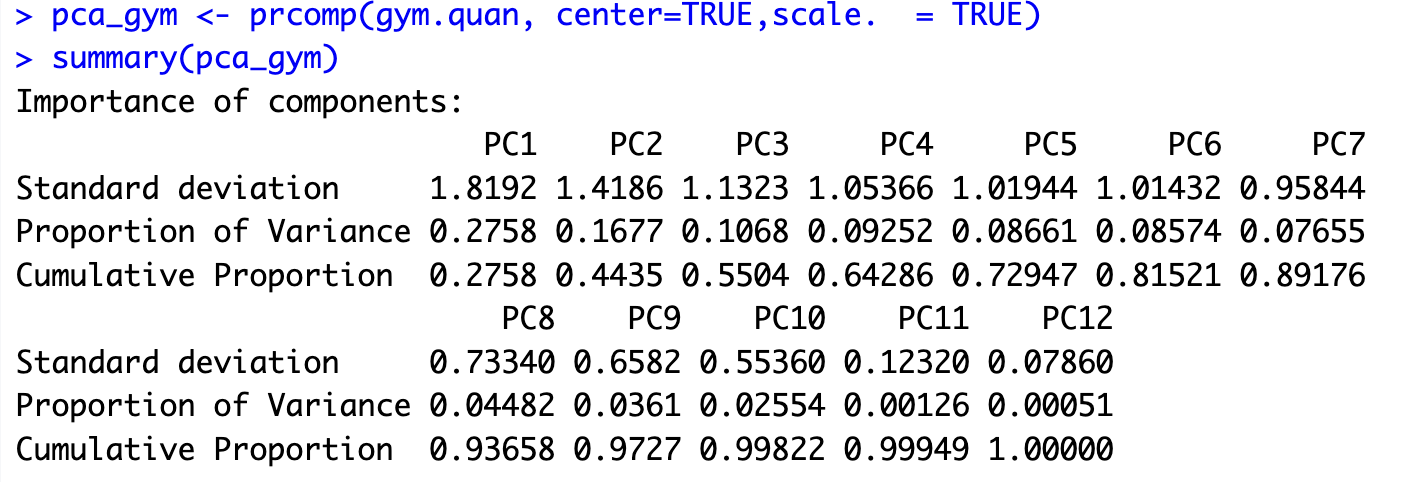
**4) Methodology and analysis**

**4.1) Principle Component Analysis**:

RQ: What are the principal components that explain the variability in gym members' characteristics? /How can we reduce the number of variables in the gym dataset while retaining most of the variance?

**Assumptions for PCA:**

1. **Linearity**: PCA assumes that the relationships between the variables are linear.
2. **Scale**: PCA is sensitive to the scale of the data. All variables should be on a similar scale, which is why standardization is often necessary, especially when the units of measurement are different like our case. (e.g., height in meters and weight in kilograms).
3. **Metric variables**: The variables should be continuous. since categorical variables exist, they should be excluded.So we have defined a variable gym.quan to include all quantitative variable.



Since it is a natural science applications, it is desired to explain at least 80%-90% of the total variance.

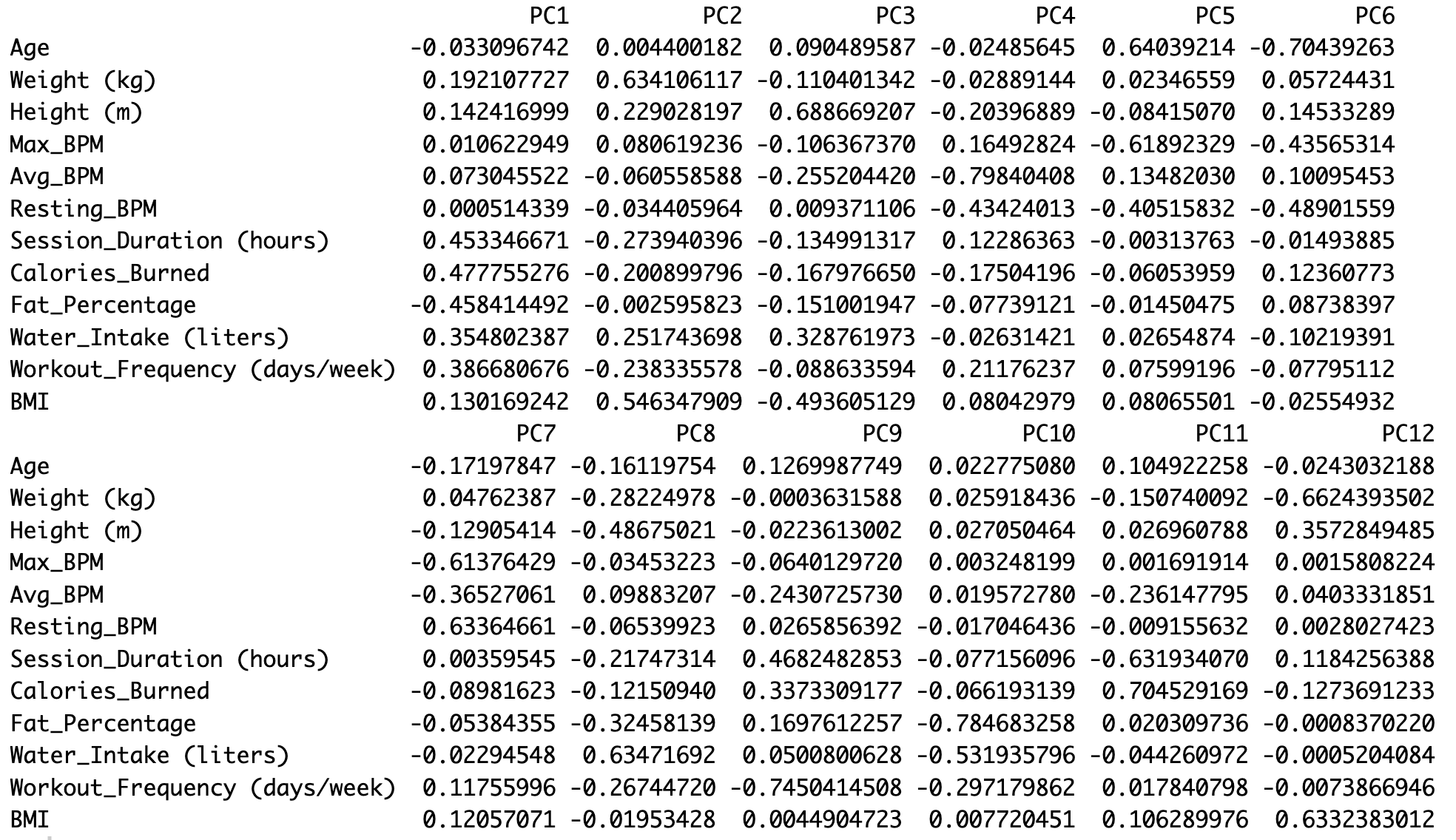
We decided on **Retaining 6 components** that explain the majority of the variation in the data, since the **Cumulative Variance Explained = 0.81521 (81.5%)**. This suggests that the PCA has successfully summarized the data into a smaller set of components without losing much of the variability in the data, we could then use these components for further analysis.

A graph of a graph with numbers and points

Description automatically generated with medium confidenceA line graph with numbers and points

Description automatically generated

Based on the scree plot we can retain 3 or 6 components.



Each principal component could me written as a linear combination of the original components:

PC1 = -0.0331 × Age + 0.1921 × Weight (kg) + 0.1424 × Height (m) + 0.0106 × Max\_BPM + 0.0730 × Avg\_BPM + 0.0005 × Resting\_BPM + 0.4533 × Session\_Duration (hours) + 0.4778 × Calories\_Burned - 0.4584 × Fat\_Percentage + 0.3548 × Water\_Intake (liters) + 0.3867 × Workout\_Frequency (days/week) + 0.1302 × BMI

The same logic applied to the 6 components…

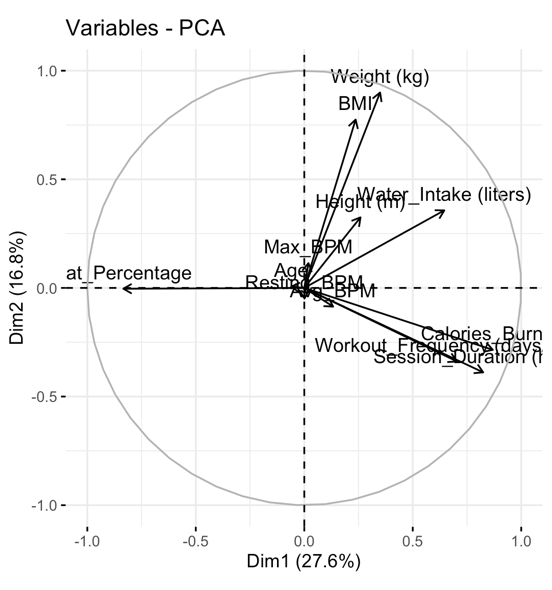
Pattern in the dataset revealed in the principal components :

On the 1st principal component: Session Duration, Calories Burned, Fat Percentage are associated with higher effects while resting heart beats is associated with the least effect.

On the 2nd principal component: Weight and BMI have the highest effect while fat pecentage has the lowest.

Height is mostly affect the 3rd principle while average BPM highly affect the 4th principle and age is affecting PC5 positively high while max BPM negatively affect it , lastly age also negatively affects PC6.

Weight and BMI are highly correlated to each other, as well as calories burned ,session duration and workout frequency are also highly correlated.

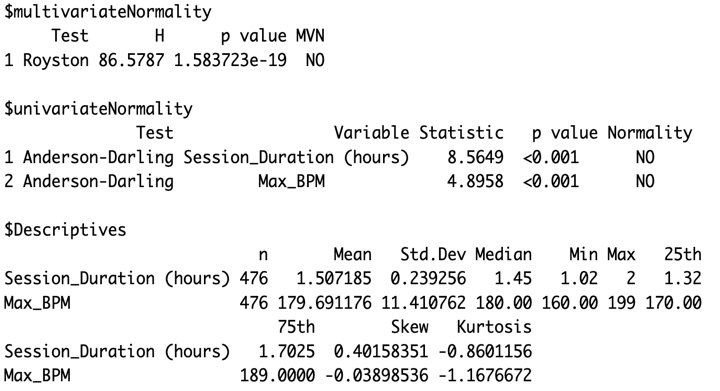
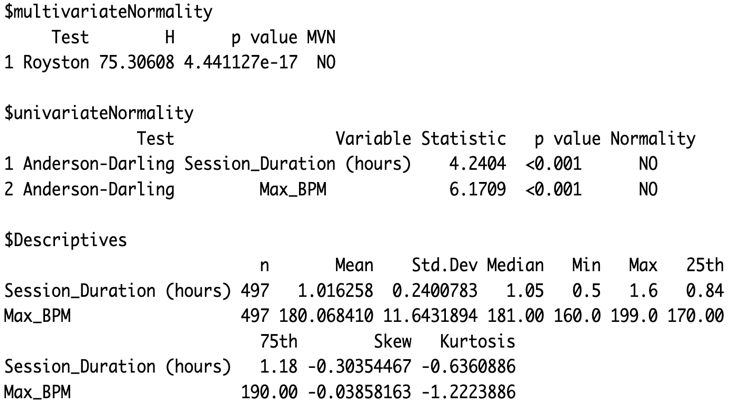
 A screen shot of a graph

Description automatically generated

**4.2) Discriminant Analysis:**

RQ: How effectively can Duration of each workout session and Maximum BPM discriminate between gym members who burn high versus low calories during workouts?

Check the normality assumption:



The **Royston test** for multivariate normality yielded a p-value< 0.05, indicating a significant departure from multivariate normality. This suggests that the assumption of multivariate normality is violated.

We have tried to remove the outliers and check if possible transformations can be applied using box-cox test but the result doesn’t differ.

The equality of variance covariance matrices:

H0: Σ1 = Σ2 H1: Σ1 ≠ Σ2

A black text on a white background

Description automatically generated

Using **Box’s M-tes**t for Homogeneity of covariance matrix, with p-value = 0.5012 > 0.05

We don’t reject H0 with significant 0.05, which means that the covariance matrix doesn’t differ between the 2 groups.

The equality of means between these groups:

H0: M1 = M2 H1: M1 ≠ M2

A number and numbers on a white background

Description automatically generated

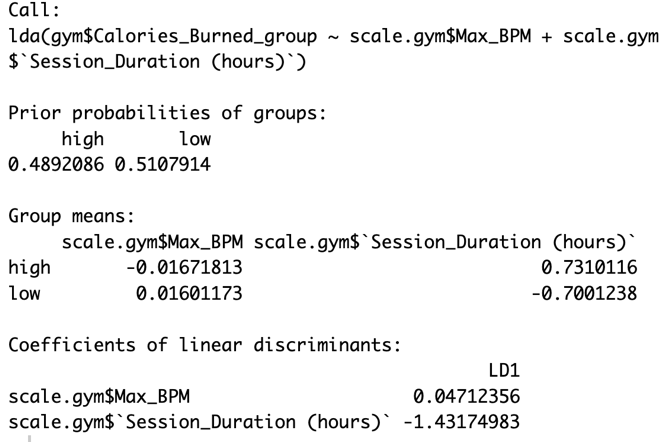
Using **Hoteling test** , with p-value = 0 < 0.05

We reject H0 with significant 0.05, which means that the means differ, and we could use discriminant analysis.

**Linear Discriminant Analysis:**

Since the normality assumption is not satisfied, and the variance covariance matrices are equal so we can apply **Fisher’s approach**:

A screenshot of a computer program

Description automatically generated 

We standardize the variables first to interpret the power of the variables in discriminating between the groups. However, using standardized variables, the most powerful variable in discriminating between members with high and low total calories burned during each session is the session duration(-1.4317) followed by the max heart rate (beats per minute) during workout sessions.

**For prediction :**

We use unstandardized coefficients as follows :

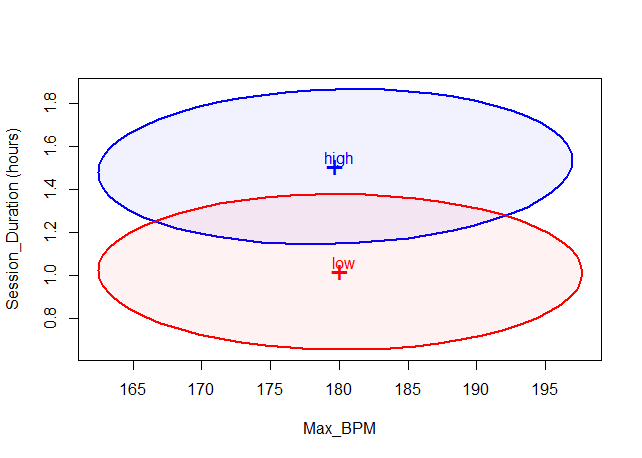
Y= 0.004088569 (max\_BPM) – 4.1738 (session duration)

A graph of blue and green bars

Description automatically generated with medium confidence

* the discriminant function for each group:

While the group means are distinct, there is some overlap between the two distributions. This indicates moderate separation, with potential misclassifications in the overlapping region. The discriminant function works reasonably well but could be improved for better group distinction.

* Probability of correct classification :

A close-up of a number

Description automatically generated

-For group 1 : 86%

-For group 2 : 87%

-Overall Classification Accuracy = 86.5%

We have tried different prior probabilities, and checked the classification table, but the number of misclassified observations increased.

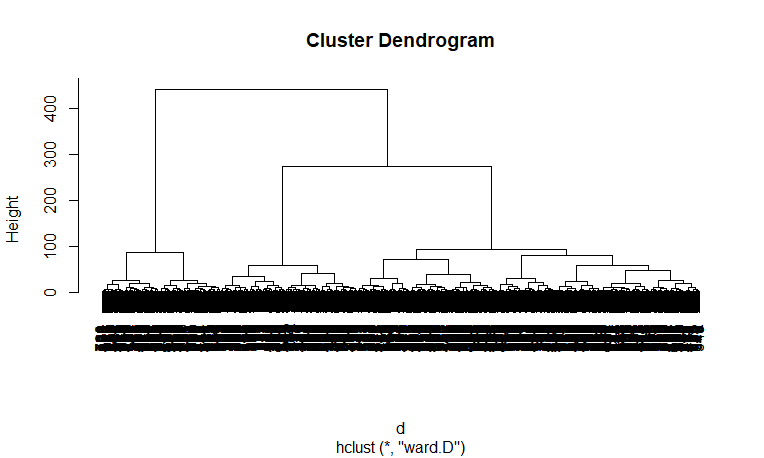
**4.3) Cluster Analysis**

RQ: What are the multidimensional clusters of gym members combining physiological, behavioral, and demographic variables?

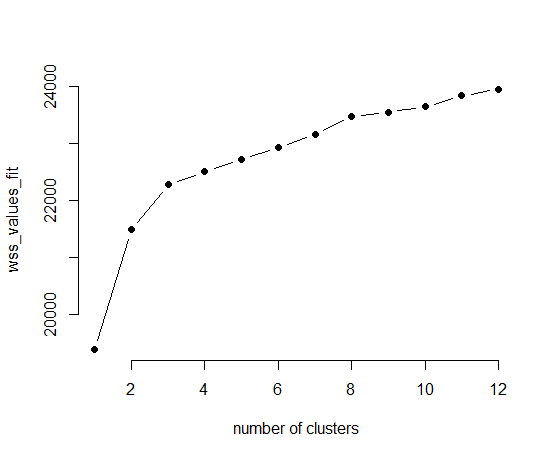
* We used both methods ; Agglomerative Clustering and DIANA , also tried different Linkage Methods like ( complete , ward’s , average ) ,
* although we reached best results using agglomerative cluster with Euclidean as a Similarity Measure and ward’s method as a linkage method .

at first we scaled our (quantitative) variables not to be affected by different measurement units in our data, then we checked the clustering tree to decide the number of clusters needed

A diagram of a cluster

Description automatically generated

-The tree was not clear as we have a large sample size.

- Using a dendrogram to analyze the hierarchical clustering results, we decided to retain **three clusters** as it shows a significant jump in the linkage distance when moving from three clusters to two and The variation between the clusters at the three-cluster level is substantial.

A number on a white background

Description automatically generated

-The number of observations in each cluster is acceptable.

- also elbow plot starts leveling off after third cluster

Results:

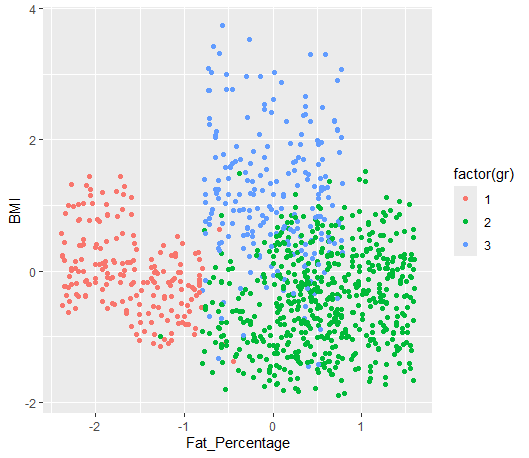
A white background with black and red text

Description automatically generated

From this results we can indicate that 3 clusters had divided our observations in a good way .

Cluster 1: includes members that have highest (resting\_BMP,session,calories burnt,water,workout) moderate (height,max\_BMP, BMI) lowest (age, weight,average\_BMP,fat)

Cluster2: includes members that have highest (average\_BMP,fat, BMI ) moderate (age,weight,session,workout) lowest (height, max\_BMP, resting\_BMP, calories burnt,water)

Cluster 3: includes members that have highest ( age, weight, height, max\_BMP) moderate (resting\_BMP, calories burnt, fat, water) lowest (average\_BMP, session,workout, BMI)

So this indicates :

- Cluster 1 could represent gym-goers focused on calorie burning and fat loss.

-Cluster 2 might represent casual exercisers or beginners with less focus on fitness intensity.

-Cluster 3 could include older or less active members with a focus on hydration and lower workout intensity.

-This graph indicates good separation between observations among the three clusters.

**5)conclusion**

This study analyzed a dataset of gym members to uncover patterns and relationships among various characteristics and behaviors. Using Principal Component Analysis (PCA), we reduced the complexity of the dataset, identifying six principal components that explain over 81% of the variability in the data. This allowed us to highlight key variables such as session duration, calories burned, and fat percentage as significant contributors. Additionally, we employed Linear Discriminant Analysis (LDA) to classify members into groups based on calories burned. The analysis revealed that session duration and maximum heart rate are the most powerful discriminators, achieving a classification accuracy of 86.5%. Despite some overlap between the groups, the model provides meaningful insights into gym members' performance and health metrics. And finally the cluster analysis results a good grouping between observations showing the three types of gym members.