**Hemanth Matta**

**Webster University**

**CSDA 6010 DATA ANALYTICS PRACTICUM**

**SPRING 2025**

**FINAL PROJECT (Case-2)**

**JP WANG**

**Executive Summary**

The Fitness Center Analytics project aims to leverage data-driven insights to optimize fitness programs by analyzing workout habits, health metrics, and user behaviors. The goal is to enhance member engagement, improve retention, and develop personalized workout plans tailored to individual fitness needs, ultimately maximizing fitness outcomes.

The project commenced with a comprehensive data preprocessing phase, during which key attributes were defined, missing values were addressed, and the dataset was thoroughly explored for inconsistencies. Exploratory Data Analysis (EDA) was then conducted to uncover trends in BMI distributions, age groups, heart rate variations, calorie expenditure, and hydration needs. This phase involved visualizations and statistical summaries that provided critical insights into the fitness patterns and behaviors of members.

A detailed correlation analysis was performed to assess the relevance of various predictors, ensuring that only variables with meaningful relationships to calorie expenditure were considered. To preserve the integrity of the dataset, all attributes were retained, allowing for a deeper and more comprehensive understanding of member behavior. Dimension reduction techniques were also evaluated, but no attributes were excluded to ensure that the analysis was as holistic as possible. For clustering, the data partitioning approach utilized the entire dataset, as there was no predefined outcome variable

To segment members based on their fitness behavior, demographics, and health metrics, K-Means clustering and hierarchical clustering were employed. These methods successfully identified distinct fitness groups, enabling the development of tailored fitness programs. The silhouette score for hierarchical clustering confirmed the strong group separation, indicating that the clustering approach was effective and meaningful.

For predictive modeling, both multiple linear regression and random forest regression were applied to quantify the impact of various factors on calorie expenditure. While the random forest model captured non-linear relationships and complex interactions effectively, multiple linear regression provided a clear and interpretable understanding of the key factors influencing calorie burn. Ultimately, the multiple linear regression model was chosen for evaluation due to its strong predictive capability and interpretability.

This project provides FitLife Wellness with actionable insights that can be used to create personalized workout plans, optimize calorie burn strategies, and enhance overall member experience. By leveraging these findings, FitLife Wellness can increase member engagement, improve retention, and deliver more targeted fitness programs that lead to better health outcomes

**TABLE OF CONTENTS**

1. INTRODUCTION ……………………………………………………………………….…6

2. BUSINESS AND ANALYTICAL……………………………………….............................6

2.1. BUSINESS PROBLEM……………………………………………………………6

2.2. BUSINESS GOAL…………………………………………………………………6

2.3. ANALYTICAL GOAL…………………………………………………………….6

2.4. ANALYTICAL APPROACH…………………………………................................6

3. DATA EXPLORATION AND PREPROCESSING………………………………………….7

3.1. ATTRIBUTE DEFINITION……………………………………………………….7

3.2. TYPES OF DATA………………………………………………………………….7

3.3. DETAILED ANALYSIS ON EACH ATTRIBUTE…………….………………….9

4. DATA ENGINEERING AND TRANSFORMATION ………………………………………13

5. PREDICTOR ANALYSIS AND RELEVANCY…………………………………………….19

5.1. PREDICTOR RELAVANCY FOR REGRESSION……………………………….19

5.2. PREDICTOR RELANCY FOR CLUSTERING …………………………………….26

6. DIMENSION REDUCTION………………………………………………………………...26

6.1. DIMENSION REDUCTION FOR REGRESSION………………………………….26

6.2. DIMENSION REDUCTION FOR CLUSTERING………………………………….28

7. DATA PARTITIONING……………………………………………………………………...30

8. MODEL SELECTION………………………………………………………………………31

8.1. MODEL SELECTION FOR CLUSTERING…………………………………………31

8.2. MODEL SELECGTION FOR REGRESSION ……………………………………….41

9. MODEL IMPROVEMENT AND EVALUATION…………………………………………...42

10. OBSERVATION……………………………………………………………………………..47

11. CONCLUSION………………………………………………………………………………53

**LIST OF FIGURES**

1. Types of data……………………………………………………………………………....9

2. Gender distribution………………………………………………………………………10

3. Workoutype distribution ………………………………………………………………...10

4. Experience level………………………………………………………………………….11

5. distribution of all continuous attributes………………………………………………….12

6. BMI distribution…………………………………………………………………………16

7. Age group………………………………………………………………………………...17

8. distribution of new attributes…………………………………………………………….17

9. missing values………………………………………………………………..............….19

10. age, gender, weight, workout type, session duration vs calories burned………………...19

11. Height, max bpm, avg bpm, resting bpm and fat percentage vs calories burned……….20

12. Water intake, workout frequency, experience level, BMI, workout intensity…………...22

13. Weight-to-Height Ratio, Heart Rate Range, Heart Rate Reserve, Heart Intensity Ratio, Calories per kg vs Calories Burned……………………………………………………………...23

14. Hyderation need ratio, age group, BMI category vs Calories Burned………………...…25

15. correlation matrix……………………………………………………………………...…27

16. data partition………………………………………………………………………….….30

17. z-score normalization……………………………………………………………………32

18. Elbow method plot………………………………………………………………………33

19. silhouette plot for K-means clustering………………………………….……………….34

20. Hierarchical clustering dendrogram…………………………………….………............36

21. silhouette plots for Hierarchical clustering……………………………………...............36

22. number of clusters……………………………………………………………………….36

23. Multi linear regression…………………………………………………………………...38

24. Evaluation metrics for Regression………………………………………………………39

25. random forest…………………………………………………………………………....39

26. Evaluation of metrics for random forest…………………………………………...……40

27. Forward selection……………………………………………………………….............43

28. silhouette score vs number of clusters…………………………………………..............44

29. Linkage methods………………………………………………………………………...47

30. clustering centroids interpretation…………………………………………………….…49

**1. Introduction:**

FitLife Wellness is a leading fitness center dedicated to promoting better health and wellness for its members. With the increasing focus on personalized fitness experiences, the company is eager to harness the power of data to enhance the effectiveness of its programs. By collecting anonymized data from members’ fitness trackers, gym attendance, and health assessments, FitLife Wellness aims to gain deeper insights into its members' workout habits and overall health.

This data-driven approach will help the company understand how different workout types impact fitness progress and member engagement, enabling it to create more tailored fitness plans. The goal is to use these insights to improve the fitness journey for each member, ensuring they achieve optimal health outcomes.

**2. Business and Analytical:**

**2.1. Business Problems:**

* **Lack of Member Insights & Engagement:** FitLife Wellness struggles to understand member preferences and workout behaviors, making it difficult to create an engaging and motivating fitness experience.
* FitLife Wellness lacks a clear understanding of the key factors that influence calorie expenditure during workouts. Without this insight, it is difficult to design personalized and effective workout programs that maximize fitness outcomes and help members achieve their health goals.
  1. **Business Goal:**
* **Lack of Member Insights & Engagement:** Enhance member engagement and satisfaction by gaining a deeper understanding of workout preferences, behaviors, and health trends. Utilize data-driven insights to recommend suitable workout plans for desired fitness outcomes and actively engage customers to improve retention.
* Gain insights into the factors that impact calorie expenditure by analyzing workout, health, and demographic data. Use these insights to develop personalized, data-driven workout plans that optimize calorie burn and overall fitness progress for members.

**2.3 Analytical Goals:**

* **Member Segmentation:** Use unsupervised learning models to group members based on workout behavior, health metrics, and demographics. Segmentation will help design tailored fitness programs for different fitness levels.
* **Calorie Prediction Model:** Develop a supervised regression model to predict Calories Burned based on given predictors

**2.4 Analytical Approaches:**

**For Addressing Varying Fitness Needs:**

Approach:

Data Preparation: Clean and preprocess data by handling missing values, scaling numerical variables, and encoding categorical variables.

Segmentation: Apply clustering algorithms to group members based on key features like workout frequency, session duration, age, BMI, and workout preferences.

Interpretation: Analyze the segments to identify patterns and characteristics that will guide the creation of personalized workout plans for each group.

**For Gaining Insights on Workout Impact:**

Approach:

Data Preprocessing: Clean and preprocess the data, ensuring that missing values are handled, categorical variables (workout type) are encoded, and numerical variables are scaled.

Regression Modeling: Build regression models to predict calorie expenditure based on variables such as workout type, session duration, and heart rate metrics.

Evaluation: Evaluate model performance using metrics like R-squared, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) to assess the accuracy of predictions.

Insights: Identify the most significant factors influencing calorie expenditure and use this information to optimize workout programs for better results.

**3. Data exploration and preprocessing:**

This step involves examining the dataset to understand its structure, identifying missing values, and performing necessary transformations

**3.1 Attribute definition:**

1) Age: The age of the member in years.

2) Gender: The gender of the member, either Male or Female.

3) Weight (kg): The weight of the member in kilograms.

4) Height (m): The height of the member in meters.

5) Max\_BPM: The maximum heart rate recorded during a workout session, measured in beats per minute.

6) Avg\_BPM: The average heart rate during the workout session, measured in beats per minute.

7) Resting\_BPM: The resting heart rate of the member measured before the workout session, in beats per minute.

8) Session\_Duration (hours): The total duration of the workout session, measured in hours.

9) Calories\_Burned: The total number of calories burned during the workout session.

10) Workout\_Type: The type of workout performed during the session, such as Yoga, Cardio, HIIT, or Strength training.

11) Fat\_Percentage: The percentage of body fat in the member’s body.

12) Water\_Intake (liters): The total amount of water consumed by the member during the session, measured in liters.

13) Workout\_Frequency (days/week): The average number of days per week the member works out.

14) Experience\_Level: The experience level of the member, categorized as 1 = Beginner, 2 = Intermediate, and 3 = Advanced.

BMI (Body Mass Index): The body mass index of the member, calculated using the formula: Weight / (Height²).

**3.2. Types of data:**

There are 3 types of data in the file.

**1. Integer:**

Age, Weight (kg), Max\_BPM, Avg\_BPM, Resting\_BPM, Workout\_Frequency, Experience\_Level

**2. Character:**

Gender, Workout\_Type

**3. Number**

Height, Session\_Duration, Calories\_Burned, Fat\_Percentage, Water\_Intake (liters), BMI

**Categorical Attributes:**

Ordinal:

* Experience\_Level

Nominal:

* Gender
* Workout\_Type (Yoga, Cardio, HIIT, Strength)

Binary:

* None in this dataset

**Continuous Attributes:**

* Age, Weight, Height, Max\_BPM, Avg\_BPM, Resting\_BPM, Session\_Duration, Calories\_Burned, Fat\_Percentage, Water\_Intake, Workout\_Frequency, BMI

A screenshot of a computer

AI-generated content may be incorrect.

Fig1: Types of data

**3.3. Detailed analysis on each attribute:**

**GENDER:**

* The fitness center has 511 male members and 462 female members, indicating a nearly balanced distribution, with males making up a slightly higher proportion.
* The near-equal representation suggests that the fitness center attracts both genders evenly, which is a positive sign of inclusivity.

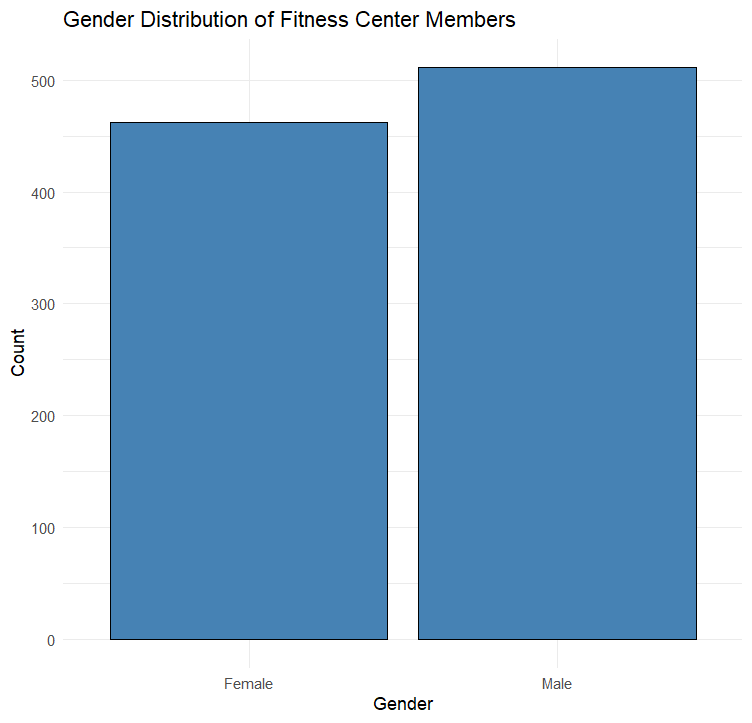


Fig2: gender distribution

**Workout type:**

* The dataset contains four primary workout types: Cardio (255), HIIT (221), Strength (258), and Yoga (239).
* Strength training is the most common workout type, while HIIT has the least number of sessions.
* The distribution is balanced, indicating a diverse range of workout preferences among users.

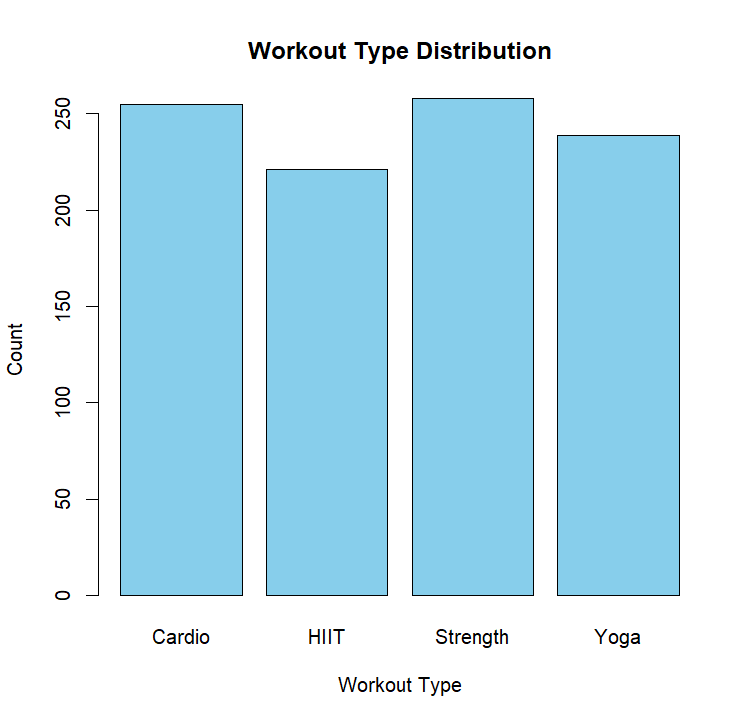


Fig3: workout type distribution

**Experience level:**

* Beginner (Level 1): 376 members fall into this category, indicating that a significant portion of the fitness center’s members are new to working out.
* Intermediate (Level 2): 406 members make up the largest group, showing that most members have some experience and are progressing in their fitness journey.
* Advanced (Level 3): 191 members are at an advanced level, meaning they have extensive training experience.
* Since over 78% (Beginner + Intermediate) of members are still progressing, personalized training plans and motivation strategies could enhance their experience.
* Advanced-level members are relatively fewer, suggesting the need for specialized programs to retain and attract highly experienced fitness enthusiasts.

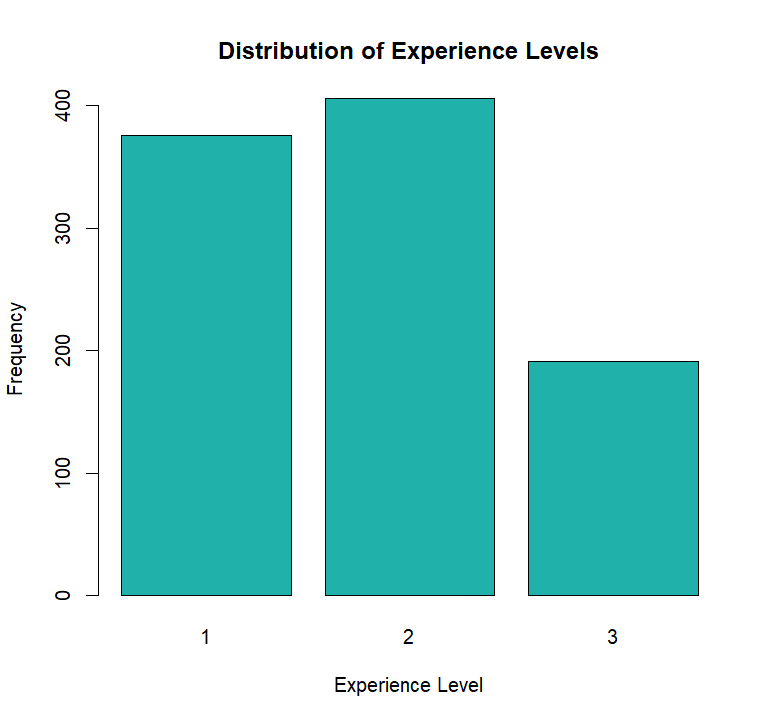


Fig4: experience level

**Continuous attributes insights:**

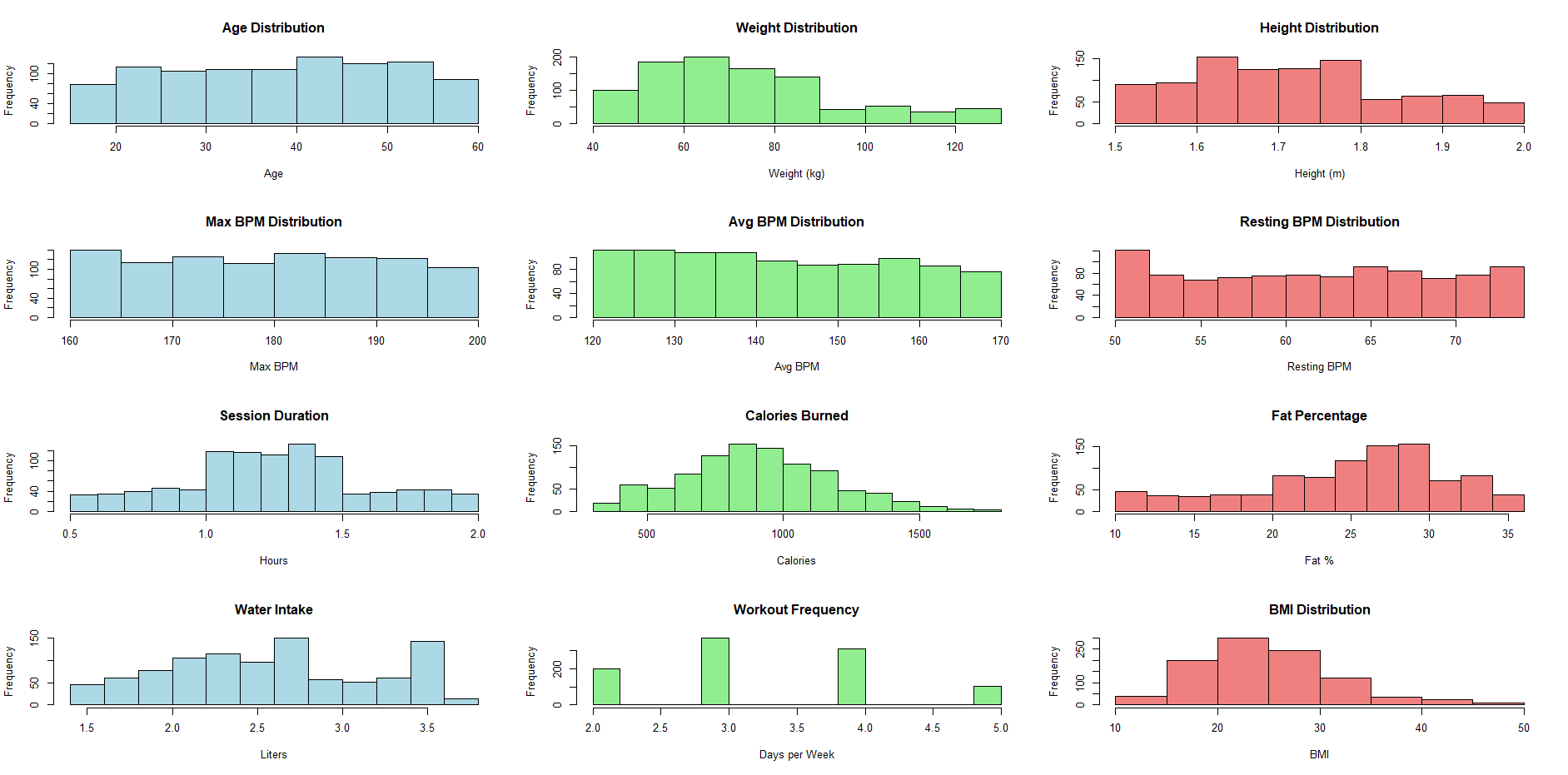


Fig5: distribution of all continuous attributes

1. Demographics:

Age Distribution: The dataset includes individuals aged approximately 18 to 60 years, with the most common age range being 30-45 years.

Height Distribution: Most individuals fall within 1.6m to 1.8m, with fewer participants below 1.5m or above 1.9m.

Weight Distribution: The majority of individuals weigh between 50 kg and 90 kg, with a peak around 70-80 kg. Very few individuals weigh above 120 kg.

2. Heart Rate Metrics:

Max BPM (Beats Per Minute): Typically ranges from 160 BPM to 200 BPM, with a balanced distribution across bins.

Average BPM: The majority of values fall between 120 BPM and 170 BPM, suggesting variations in exercise intensity.

Resting BPM: Most individuals have a resting heart rate in the range of 50 to 70 BPM, which is within the normal healthy range.

3. Workout & Fitness Habits:

Session Duration: Most workout sessions last between 1.0 and 1.5 hours, with very few extending beyond 2 hours.

Workout Frequency: The majority of individuals work out 3 to 4 days per week, with fewer people exercising only 2 days or up to 5 days.

Calories Burned: Most individuals burn 500-1000 calories per session, with some burning over 1500 calories, indicating intense workouts.

4. Health Metrics:

Fat Percentage: The majority of individuals have a fat percentage in the range of 20-30%, with fewer people below 15% or above 35%.

BMI Distribution: Most participants have a BMI between 20 and 30, with a peak around 25-27, indicating that many individuals are in the "Overweight" category (BMI > 25).

5. Hydration & Water Intake:

Water Intake: The most common daily water intake is 2.0 to 3.5 liters, with a smaller group consuming below 2 liters or above 3.5 liters.

**4. Data Transformation:**

**To enhance the analysis, new attributes were created in the dataset.**

**1. BMI\_CAT (Body Mass Index Category):** This categorizes individuals based on their BMI (Body Mass Index). The categories typically reflect whether a person is underweight, normal weight, overweight, or obese.

* This ratio is an indicator of body composition (fat vs. muscle). A higher ratio may suggest that someone has more body mass in relation to height, which could correlate with overweight or obesity. This could be used in designing personalized fitness or nutritional strategies.

**2. Weight to Height Ratio:** This ratio measures how much a person weighs relative to their height.

* Fitness Level: A larger heart rate range may indicate better cardiovascular fitness, as the heart can handle both higher and lower rates effectively.
* Formula: Weight to Height Ratio = Weight\_kg / Height\_m

**3. Heart Rate Range (Max BPM - Resting BPM):** The difference between a person's maximum heart rate (when exerting maximum effort) and their resting heart rate (when at rest).

* This measure can indicate how hard a person is working during their workouts. If the difference is small, the person might not be pushing themselves hard enough during exercise.
* Formula: Heart Rate Range = Max BPM – Resting BPM

**4.** **Heart Rate Reserve (Max BPM - Avg BPM):** The difference between a person’s maximum heart rate and their average heart rate during exercise.

* This measure can indicate how hard a person is working during their workouts. If the difference is small, the person might not be pushing themselves hard enough during exercise.
* Formula: Heart Rate Reserve = Max BPM – Avg BPM

**5. Heart Intensity Ratio (Avg BPM / Max BPM):** This is the ratio of the average heart rate during exercise to the maximum heart rate capacity.

* A higher ratio indicates that the person is exercising at a high intensity relative to their maximum capacity. This can be used to evaluate whether the person is pushing themselves enough during workouts.
* Heart Intensity Ratio = Avg BPM / Max BPM

**6. Calories per kg:** This measure indicates how many calories are burned per kilogram of body weight.

* This data can help tailor nutritional and exercise plans to optimize energy expenditure and weight management.
* Calories per kg = Calories Burned / Weight kg

**7. Workout Intensity Score:** A composite score that reflects the intensity of a person’s workouts, combining average BPM, session duration, and workout frequency.

* This can provide an overview of how intense a person’s workout routine is. It allows trainers or fitness apps to adjust training loads based on individual fitness levels.
* Formula: Workout Intensity Score = (Avg BPM / Resting BPM) \* Session Duration \* Workout Frequency

**8. Age Group:** Categorizes people into different age groups (18-25, 26-35, 36-45, 45+).

* It can help identify patterns in fitness behavior or outcomes across different age groups. For instance, younger individuals may burn more calories or have a higher heart rate intensity compared to older individuals.

**9. Hydration Need Ratio**: This ratio compares water intake to session duration, indicating how well a person is hydrating relative to the time spent exercising.

* Helps identify if an individual is hydrating adequately during their workouts. A low ratio might indicate dehydration, which could affect performance and recovery.
* Formula: Hydration Need Ratio = Water Intake / Session Duration

**BMI CAT:**

Distribution of BMI Categories

* The majority of individuals fall into the Normal category (366 members).
* 242 members are categorized as Overweight, indicating a significant portion may need weight management strategies.
* 197 members fall into the Obese category, suggesting a potential focus area for weight loss programs.
* 168 members are classified as Underweight, which may require targeted nutrition and muscle-gain interventions.

Implications for Workout Programs:

* The gym may need to offer diverse workout programs catering to different BMI categories.
* Obese and Overweight members may benefit from weight-loss-focused training (e.g., HIIT, strength training).
* Underweight individuals might need programs emphasizing muscle gain and balanced nutrition.

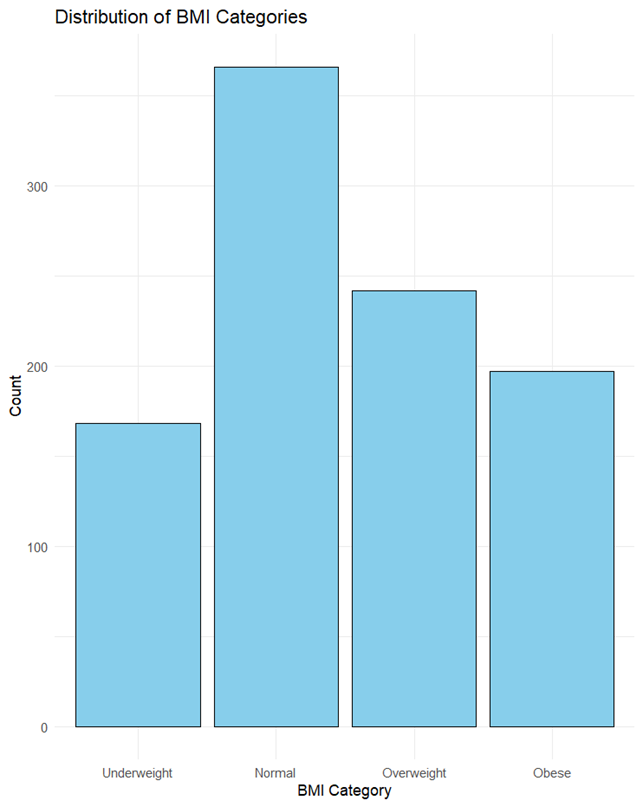


Fig6: BMI distribution

**Age group:**

Majority of Gym Members Are Older:

The largest age group is indicating that a significant portion of gym members are older adults.

This suggests a need for low-impact workouts, flexibility training (e.g., Yoga), and strength training tailored for injury prevention and longevity.

Balanced Distribution Among Other Age Groups:

36-45 age group is the second largest, showing a strong middle-aged demographic.

26-35 age group represents young professionals, likely focused on fitness for weight management and performance.

18-25 age group is the smallest, possibly due to younger individuals having different fitness habits or preferring alternative workout methods (e.g., sports, home workouts).

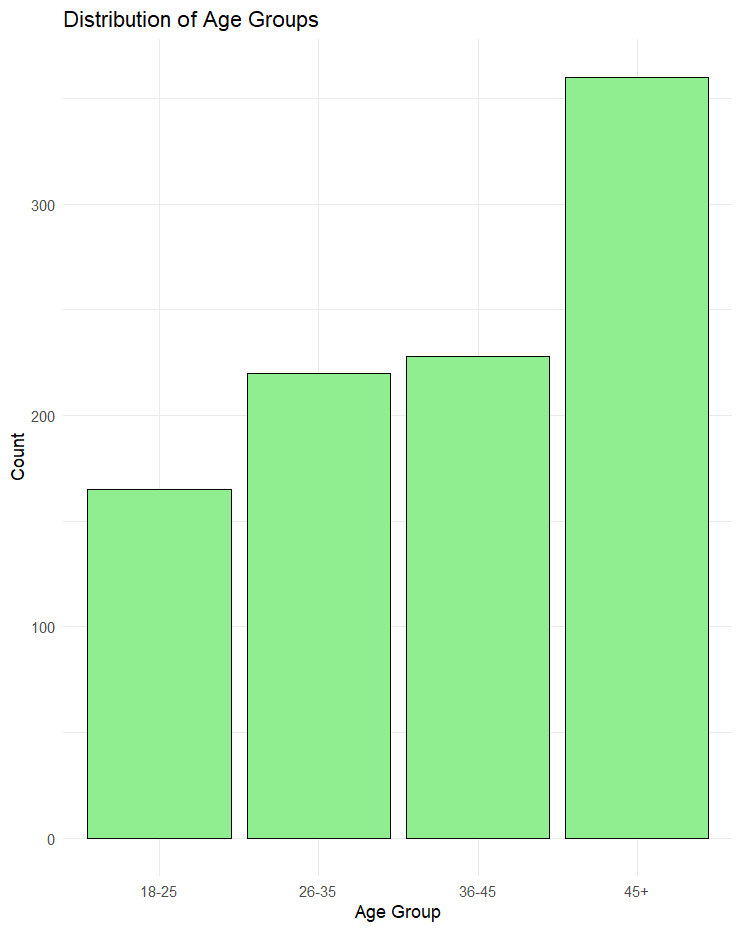


Fig7: age group

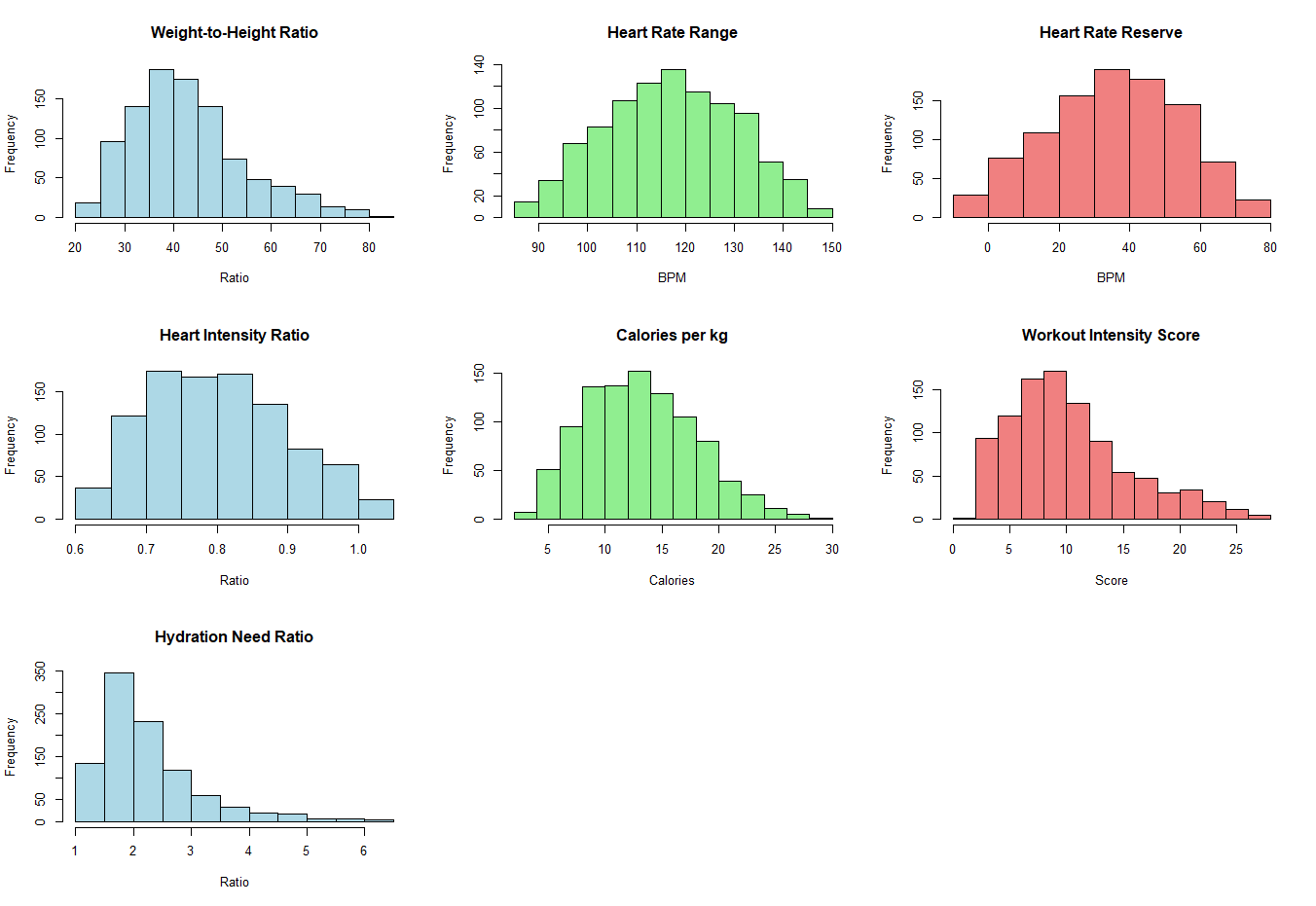


Fig8: distribution of new attributes

Weight-to-Height Ratio:

* The distribution is slightly right skewed, with most values between 30 and 60.
* This suggests that the majority of individuals have a weight-to-height ratio within a common range, while a few outliers have higher ratios, possibly indicating higher BMI values.

Heart Rate Range (BPM):

* The data is normally distributed, centered around 120–130 BPM, which aligns with a moderate-intensity workout zone.
* Most individuals maintain a heart rate within this range, showing a balanced approach to cardiovascular exercise.

Heart Rate Reserve (BPM):

* The distribution is right-skewed, with most values between 20 and 60 BPM.
* Higher heart rate reserve values suggest a higher fitness level, as a greater difference between resting and maximum heart rate indicates better cardiovascular efficiency.

Heart Intensity Ratio:

* The distribution is fairly uniform, ranging between 0.7 and 1.0, meaning that workout intensities vary widely.
* A large proportion of individuals maintain a heart intensity ratio above 0.8, suggesting high-intensity training is common.

Calories Burned per kg

* The data is normally distributed, peaking around 10–15 calories per kg.
* This suggests that most individuals burn calories within this range, with fewer participants achieving extreme calorie expenditure.

Workout Intensity Score:

* The histogram shows a slight right skew, with most values between 5 and 15.
* A few individuals engage in very high-intensity workouts, but the majority remain in moderate intensity, indicating a balanced workout regimen.

Hydration Need Ratio

* This distribution is highly right-skewed, meaning most individuals have hydration needs around 1–2, but some require significantly more.
* Those with higher hydration needs may engage in more intense workouts or have higher body mass.

**3.4. Checking missing values:**

There are no NAs in the dataset

A screenshot of a computer

AI-generated content may be incorrect.

Fig18: missing values

**5. Predictor analysis and relevancy:**

Predictor analysis helps identify which variables (predictors) are most relevant for predicting a target variable. By evaluating the relationship between predictors and the target, it can select the most significant predictors for building efficient models.

**5.1 Predictor analysis for regression model:**

**age, gender, weight, workout type, session duration vs calories burned:**

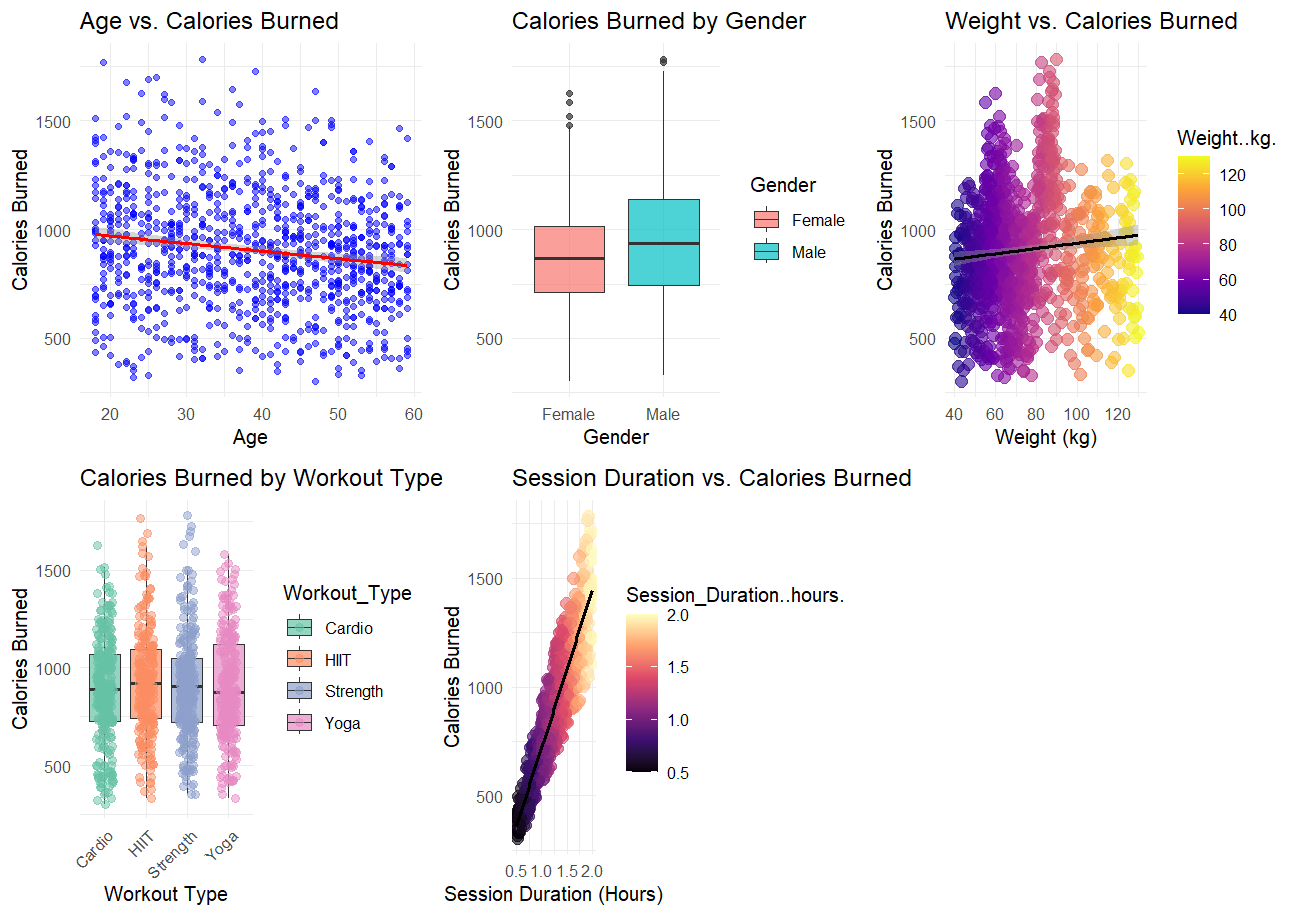


Fig10: age, gender, weight, workout type, session duration vs calories burned:

Age vs. Calories Burned

The first scatter plot shows a slight negative correlation between age and calories burned. As age increases, calorie expenditure tends to decrease, but the effect is not very strong. This could be due to variations in fitness levels, workout intensity, and BMI across different age groups.

Calories Burned by Gender

The box plot compares calories burned between males and females. While both genders have a similar range, the median calories burned appears slightly higher for males. This may be attributed to differences in body composition and workout intensity. However, the overlap suggests that gender alone is not a strong predictor of calories burned.

Weight vs. Calories Burned

The third scatter plot shows a weak positive correlation between weight and calories burned. Heavier individuals tend to burn slightly more calories, which is expected since carrying more body mass generally requires more energy expenditure during exercise. However, weight alone is not a strong determinant of calorie burn, as workout type and duration play a more significant role.

Calories Burned by Workout Type

The fourth box plot compares calories burned across different workout types (Cardio, HIIT, Strength, Yoga).

HIIT and Cardio have a higher median calorie burn, indicating that they are more intense workout styles.

Strength training shows a moderate calorie burn, likely due to a combination of resistance and endurance factors.

Yoga has the lowest calorie burn, as it typically involves lower-intensity movements compared to other workout types.

This confirms that workout type plays a key role in determining calorie expenditure.

Session Duration vs. Calories Burned

The last scatter plot shows a strong positive correlation between session duration and calories burned. As workout duration increases, the number of calories burned also increases in a nearly linear fashion. This suggests that session duration is one of the most influential factors in determining calorie expenditure.

**Height, max bpm, avg bpm, resting bpm and fat percentage vs calories burned:**

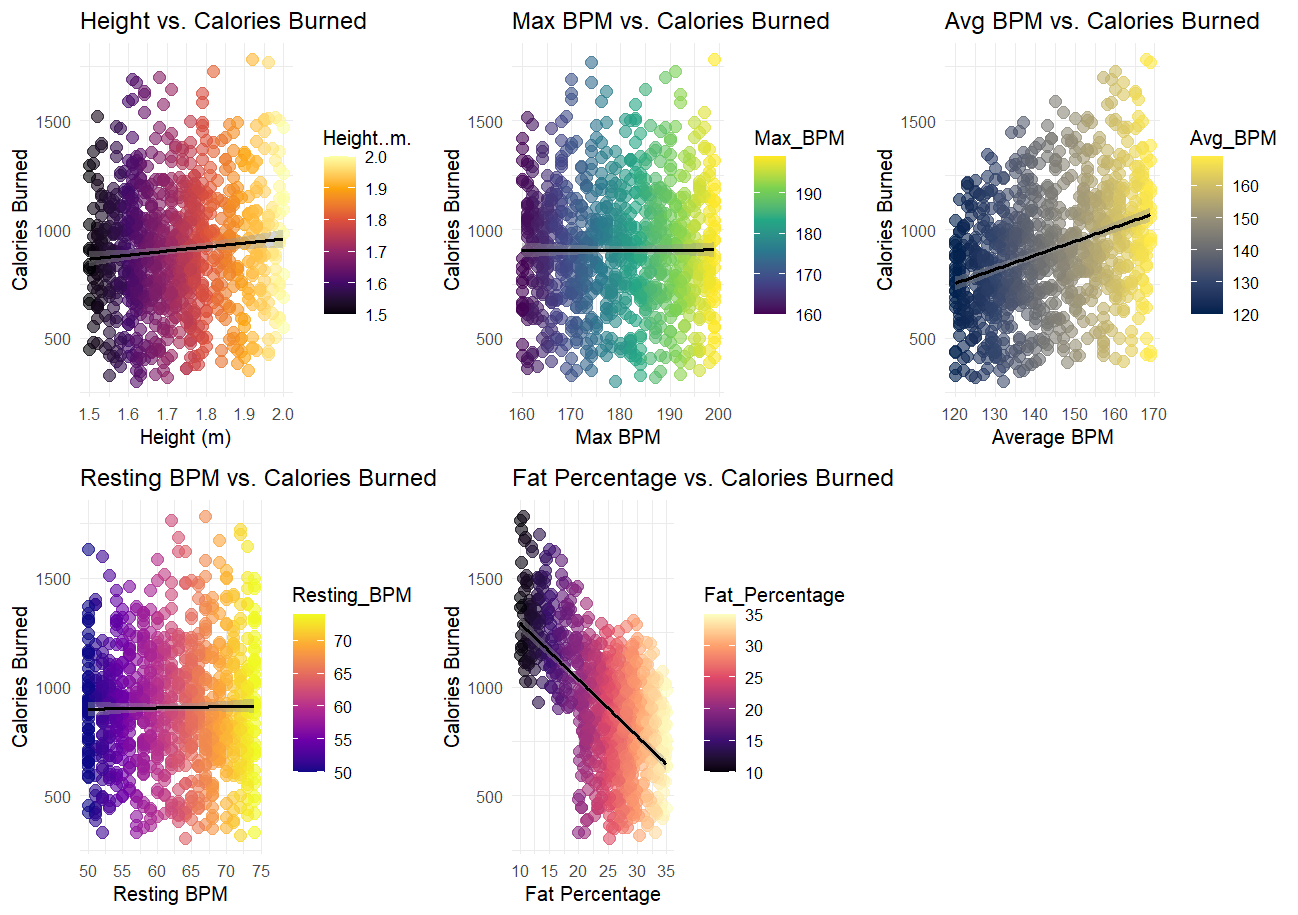


Fig11: Height, max bpm, avg bpm, resting bpm and fat percentage vs calories burned

Height vs. Calories Burned

The first plot examines the relationship between height and calories burned. The trend line appears almost flat, indicating that height has a weak or negligible correlation with calories burned. The spread of data points suggests that calorie expenditure is more influenced by other factors, such as workout intensity and duration, e.t.c.,

Max BPM vs. Calories Burned

The second plot explores the association between maximum heart rate (Max BPM) and calories burned. There is no clear upward or downward trend, implying that Max BPM does not strongly impact calorie expenditure. This could be because Max BPM is more of a physiological limit rather than an indicator of workout intensity.

Avg BPM vs. Calories Burned

The third plot reveals a positive correlation between average BPM and calories burned. As the average BPM increases, the number of calories burned also tends to rise. This suggests that individuals who maintain a higher heart rate during workouts tend to burn more calories, likely due to engaging in more intense or sustained physical activity.

4. Resting BPM vs. Calories Burned

The fourth plot shows the relationship between resting BPM and calories burned, where the trend line is nearly flat. This indicates that resting heart rate does not have a significant direct impact on calories burned. Since resting BPM primarily reflects baseline cardiovascular fitness rather than workout intensity, it is not a strong predictor of calorie expenditure.

5. Fat Percentage vs. Calories Burned

The final plot displays a strong negative correlation between fat percentage and calories burned. As fat percentage increases, the number of calories burned tends to decrease. This suggests that individuals with higher fat percentages may have lower metabolic efficiency or engage in lower-intensity workouts, leading to reduced calorie expenditure.

**Water intake, workout frequency, experience level, BMI, workout intensity:**

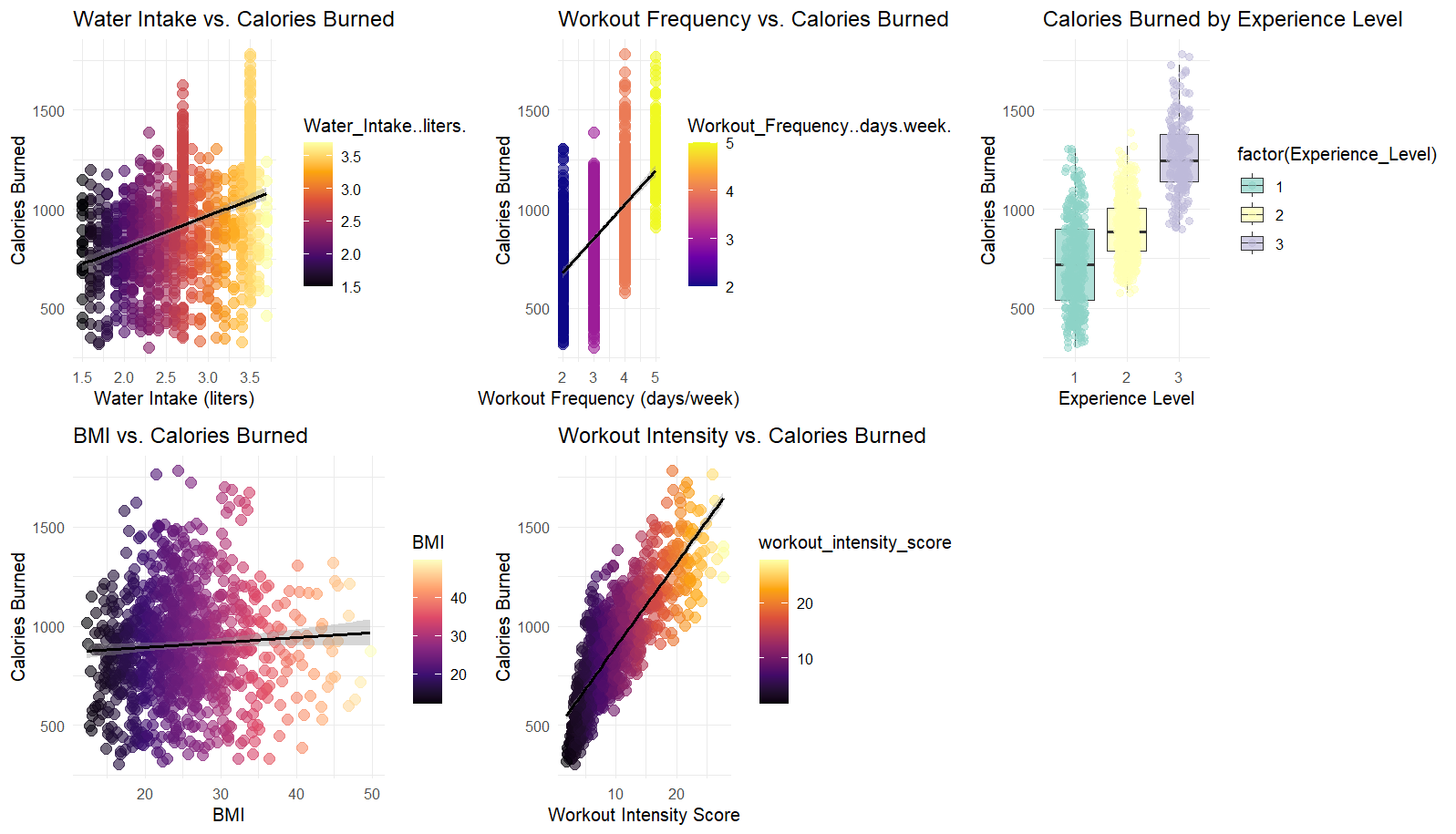


Fig12: Water intake, workout frequency, experience level, BMI, workout intensity

Water Intake vs. Calories Burned

The first scatter plot shows a positive correlation between water intake and calories burned. Individuals who consume more water tend to burn more calories, which may indicate that higher water intake is associated with longer or more intense workout sessions. However, the relationship is not very strong, suggesting other factors also influence calorie burn.

Workout Frequency vs. Calories Burned

The second scatter plot demonstrates a positive correlation between workout frequency (days per week) and calories burned. As individuals work out more frequently, they tend to burn more calories. This is expected, as consistent physical activity contributes to higher energy expenditure over time. The trend suggests that workout frequency is a relevant predictor of calorie burn.

Calories Burned by Experience Level

The third box plot compares calories burned across different experience levels (1 = Beginner, 2 = Intermediate, 3 = Advanced). Advanced members (Experience Level 3) tend to burn more calories than beginners and intermediates. This may be due to their ability to sustain higher workout intensities or engage in longer sessions. The difference in calorie expenditure across experience levels highlights its potential as an important predictor.

BMI vs. Calories Burned

The fourth scatter plot shows a weak positive correlation between BMI and calories burned. While individuals with higher BMI tend to burn slightly more calories, the effect is minimal. This suggests that BMI alone is not a strong predictor of calorie expenditure, and workout-related factors play a more significant role.

Workout Intensity vs. Calories Burned

The final scatter plot displays a strong positive correlation between workout intensity score and calories burned. As workout intensity increases, calories burned also rise significantly. This confirms that workout intensity is one of the most influential factors in determining calorie expenditure.

**Weight-to-Height Ratio, Heart Rate Range, Heart Rate Reserve, Heart Intensity Ratio, Calories per kg vs Calories Burned**

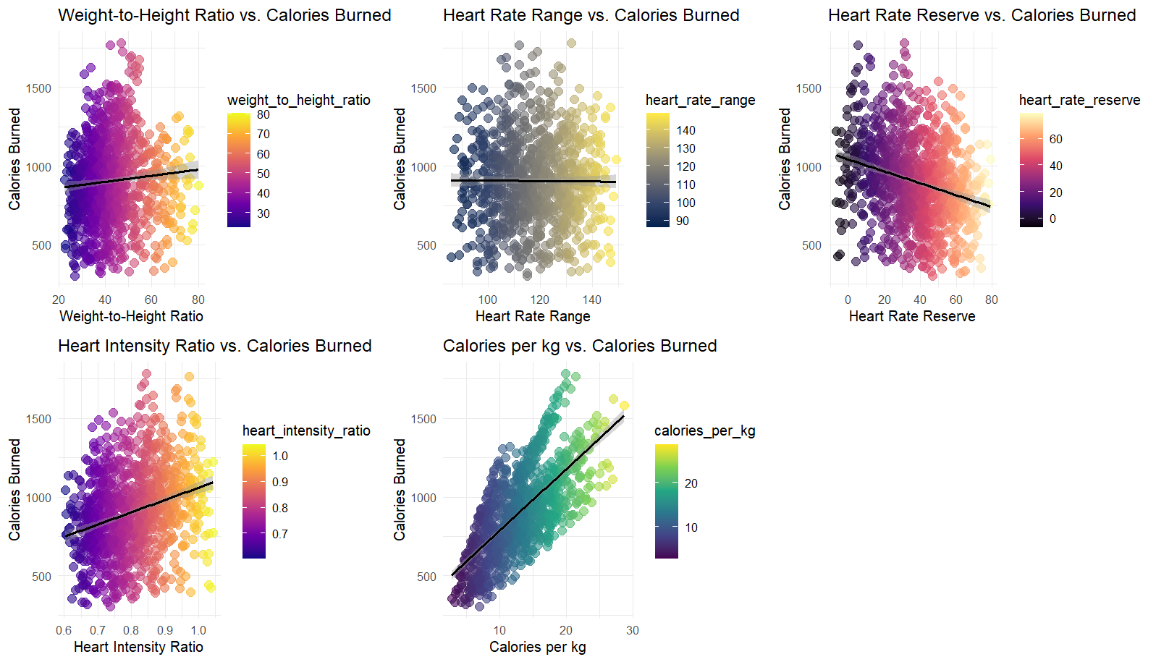


Fig13: Weight-to-Height Ratio, Heart Rate Range, Heart Rate Reserve, Heart Intensity Ratio, Calories per kg vs Calories Burned

Weight-to-Height Ratio vs. Calories Burned

The first scatter plot shows a weak positive correlation between weight-to-height ratio and calories burned. While individuals with a higher weight-to-height ratio tend to burn slightly more calories, the effect is not strong. This suggests that other factors, such as workout intensity and duration, play a larger role in determining calorie expenditure.

Heart Rate Range vs. Calories Burned

The second plot examines the relationship between heart rate range (Max BPM - Resting BPM) and calories burned. The nearly flat trend line indicates no strong correlation between these variables. This suggests that heart rate fluctuations alone may not be a reliable predictor of calorie burn, as exercise intensity and duration may have a greater impact.

Heart Rate Reserve vs. Calories Burned

The third plot shows a negative correlation between heart rate reserve (Max BPM - Avg BPM) and calories burned. As the heart rate reserve increases, calories burned tend to decrease. This could indicate that individuals with a smaller difference between their max and average BPM (i.e., maintaining a higher sustained heart rate) tend to burn more calories, likely due to consistent workout intensity.

Heart Intensity Ratio vs. Calories Burned

The fourth plot displays a positive correlation between heart intensity ratio (Avg BPM / Max BPM) and calories burned. This suggests that individuals who sustain a higher average heart rate relative to their maximum during workouts tend to burn more calories, reinforcing the idea that maintaining workout intensity is a key factor in calorie expenditure.

Calories per kg vs. Calories Burned

The final plot shows a strong positive correlation between calories per kg (Calories Burned / Weight) and total calories burned. This indicates that individuals who burn more calories per kilogram of body weight also tend to burn more calories overall, making this one of the most relevant predictors.

**Hyderation need ratio, age group, BMI category vs Calories Burned**

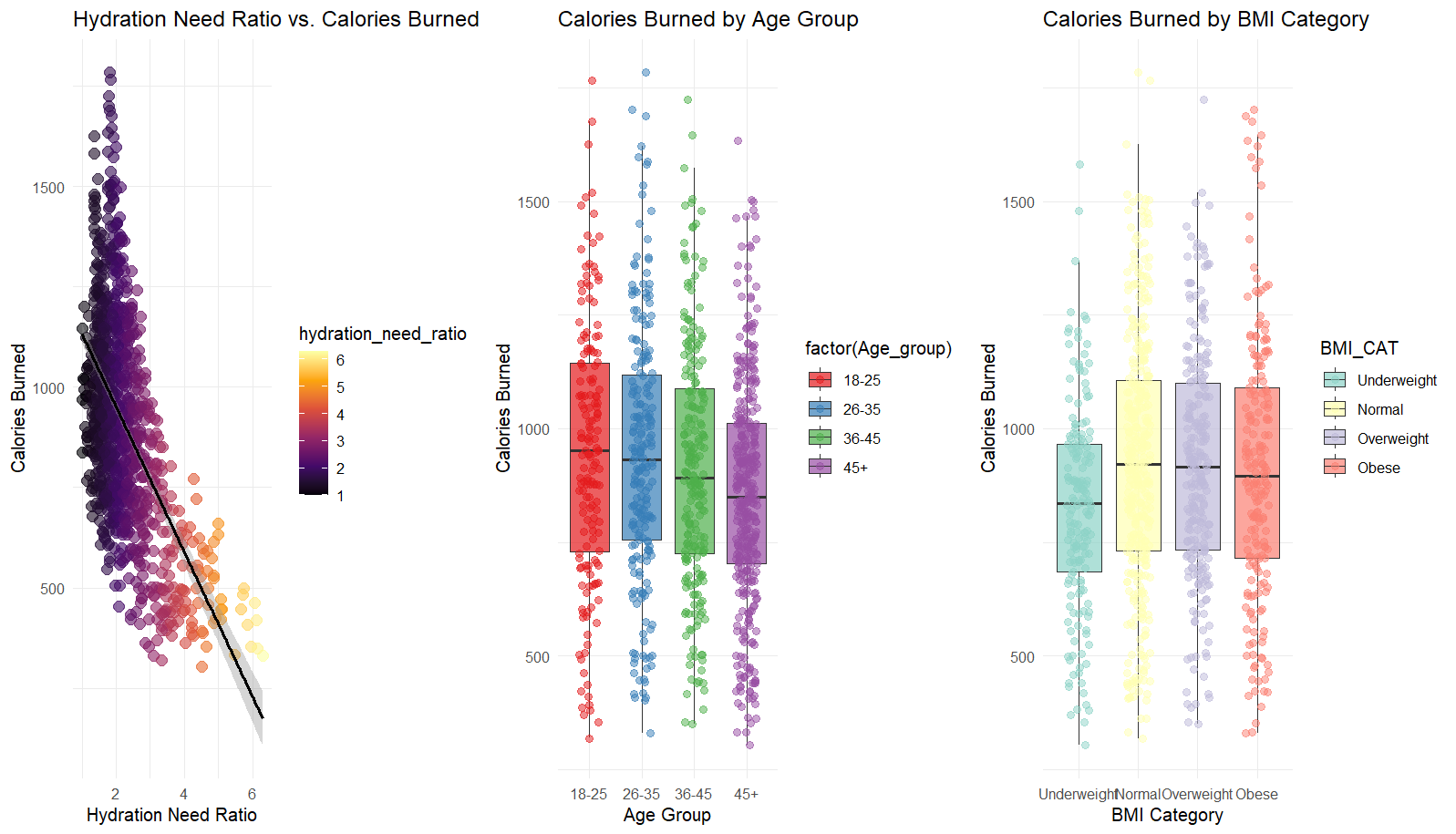


Fig14: Hyderation need ratio, age group, BMI category vs Calories Burned

Hydration Need Ratio vs. Calories Burned

The first scatter plot shows a strong negative correlation between hydration need ratio and calories burned. As the hydration need ratio increases, the number of calories burned decreases. This suggests that individuals with higher water intake per session duration tend to burn fewer calories, possibly indicating lower workout intensity or shorter exercise sessions.

Calories Burned by Age Group

The second box plot compares calories burned across different age groups (18-25, 26-35, 36-45, 45+). The distribution appears similar across all age groups, with slight variations in the median values. While younger individuals (18-25) show a slightly higher spread of calories burned, age does not seem to be a strong predictor, as there is considerable overlap between the groups.

Calories Burned by BMI Category

The third box plot illustrates the distribution of calories burned across BMI categories (Underweight, Normal, Overweight, Obese). The median calorie burn appears similar across all groups, with slightly higher variability in the overweight and obese categories. This suggests that BMI alone does not strongly influence calorie expenditure, and other factors like workout intensity and duration play a more significant role.

**5.2. Predictor Relevance in Clustering:**

Since clustering is an unsupervised learning technique, there is no predefined outcome variable to measure the direct impact of individual predictors. Instead, all attributes contribute to defining patterns and identifying meaningful segments. Each variable captures unique aspects of a member’s fitness journey. Retaining all available predictors ensures that valuable insights are not overlooked, maximizing the ability to uncover hidden patterns in the data.

**6. Dimensional reduction:**

Dimension reduction is a crucial step in data preprocessing, especially when dealing with high-dimensional datasets. It helps in removing redundant and less relevant variables, improving model efficiency, and reducing computational complexity. The key objective is to retain as much information as possible while minimizing the number of features.

**6.1. Dimension reduction for regression:**

Correlation Analysis:

A correlation analysis was conducted to examine the relationship between all attributes and Calories Burned. This analysis helped identify which variables have a strong or weak association with the target variable.

A threshold of 0.2 was set, meaning that only variables with an absolute correlation value greater than 0.2 (correlation > 0.2) were considered. This approach ensured that the focus remained on attributes with a meaningful impact on calorie expenditure while removing less relevant factors.

A screenshot of a computer

AI-generated content may be incorrect.

Fig15: correlation matrix

**Selected Variables:**

The following variables met the threshold criteria and were retained for further analysis based on their significant correlation with Calories Burned:

Session Duration (0.908): The strongest predictor, indicating that longer workouts lead to higher calorie burn.

Workout Intensity Score (0.823): Higher workout intensity correlates with increased calorie expenditure.

Experience Level (0.694): More experienced individuals tend to burn more calories.

Calories per kg (0.692): A strong indicator of calorie burn efficiency based on body weight.

Workout Frequency (0.576): More frequent workouts contribute to higher calorie burn.

Water Intake (0.357): Hydration plays a role in metabolic processes affecting calorie burn.

Average BPM (0.340): Higher average heart rate is associated with increased energy expenditure.

Heart Intensity Ratio (0.279): A measure of workout effort, positively impacting calorie burn.

Heart Rate Reserve (-0.259): A negative correlation suggests that individuals with lower reserves burn more calories.

Hydration Need Ratio (-0.571): Indicates that lower hydration needs correlate with higher calorie expenditure.

Fat Percentage (-0.598): A negative correlation, implying that individuals with a higher fat percentage burn fewer calories.

**Removed Variables (correlation ≤ 0.2):**

The following variables had weak correlations and were excluded from the analysis:

Age (-0.154)

Weight (kg) (0.095)

Height (m) (0.086)

Resting BPM (0.016)

Max BPM (0.002)

BMI (0.059)

Weight-to-Height Ratio (0.080)

Heart Rate Range (-0.007)

Gender (Male: 0.150, Female: -0.150)

Workout Type (Cardio: -0.045, HIIT: 0.040, Strength: 0.011, Yoga: -0.004)

BMI Categories (Normal: 0.060, Overweight: 0.032, Obese: 0.014)

Age Groups (18-25: 0.072, 26-35: 0.074, 36-45: 0.020, 45+: -0.139)

By removing variables with weak correlations, the model focuses on attributes that have a substantial impact on calorie burn. Variables such as Session Duration, Workout Intensity, Experience Level, and Heart Rate Metrics provide meaningful insights into workout effectiveness. Conversely, factors like Age, Height, Weight, and BMI Categories showed minimal correlation, indicating they are less predictive of calorie expenditure. This refined selection improves model efficiency and interpretability.

**6.2. Dimension Reduction for Clustering:**

For clustering, selecting the right variables is crucial to ensure meaningful segmentations. From a business perspective, the goal is to group members based on fitness behavior, demographics, and health metrics to personalize workout plans and optimize fitness progress. Here’s how can approach variable selection:

**Key Selection Criteria:**

Demographics: Age and gender help differentiate members based on physiological differences in fitness needs.

* Age
* Gender

Body Composition & Fitness Levels: Metrics like BMI, fat percentage, and weight-to-height ratio are important for grouping members based on body type.

* BMI (Body Mass Index)
* Fat\_Percentage
* Weight\_to\_Height\_Ratio

Workout Patterns & Intensity: Variables related to workout duration, frequency, and intensity help in segmenting members based on engagement levels.

* Workout\_Frequency (days/week)
* Session\_Duration (hours)
* Workout\_Type
* Workout\_Intensity\_Score

Cardiovascular Performance: Heart rate metrics (Max BPM, Avg BPM, Resting BPM) are useful in clustering based on endurance and fitness levels.

* Max\_BPM
* Avg\_BPM
* Heart\_Rate\_Range (Max BPM - Resting BPM)
* Heart\_Intensity\_Ratio (Avg BPM / Max BPM)

Calorie Expenditure: Calories burned per session and per kg help in identifying efficiency and effectiveness of workouts.

* Calories\_Burned
* Calories\_per\_kg

Hydration & Recovery: Water intake and hydration need ratio are useful for identifying members’ recovery habits.

* Hydration\_Need\_Ratio

Removed Variables:

Resting BPM:

Reason for Exclusion: Resting heart rate is inherently captured by the relationship between Max\_BPM (Maximum Heart Rate) and Avg\_BPM (Average Heart Rate). Specifically, Heart\_Rate\_Range (Max BPM - Resting BPM) and Heart\_Intensity\_Ratio (Avg BPM / Max BPM) already provide valuable insights into cardiovascular fitness and workout intensity. Including Resting\_BPM separately introduces redundancy without adding significant new information to the clustering process.

Heart Rate Reserve (Max BPM - Avg BPM):

Reason for Exclusion: Heart\_Rate\_Reserve essentially measures the difference between the maximum heart rate and the average heart rate during exercise. Since Heart\_Rate\_Range already captures this difference in a more intuitive manner (Max BPM - Resting BPM), there is no added value in retaining both. The Heart\_Rate\_Range metric provides a clearer, more concise representation of cardiovascular effort during workouts, making Heart\_Rate\_Reserve unnecessary for clustering.

BMI CAT (Body Mass Index Category):

Reason for Exclusion: The BMI\_CAT categorizes individuals based on their BMI into predefined categories (e.g., underweight, normal weight, overweight, obese). However, BMI itself, which is a continuous variable, provides more granular and precise information on body composition. Since BMI\_CAT is essentially a transformation of BMI, it is redundant and does not add additional insight for clustering, where continuous variables are preferred for capturing finer distinctions.

Age Group (Age Category):

Reason for Exclusion: Age\_Group is a categorical variable derived from Age, segmenting the population into predefined age brackets (e.g., 18-25, 26-35). However, Age itself, being a continuous variable, already provides a more detailed and flexible segmentation. Categorical variables like Age\_Group can obscure the fine distinctions present in the data, especially when using clustering techniques that benefit from continuous measures.

**7. DATA PARTITION:**

To build a robust predictive model, to split the dataset into training and testing subsets. This allows us to train the model on a portion of the data and evaluate its performance on unseen data.

For regression:

will divide the dataset into three parts:

Training Set (70%) – Used to train the machine learning model and help it learn patterns.

Validation Set (15%) – Used to fine-tune the model and adjust parameters to avoid overfitting.

Testing Set (15%) – Used to evaluate how well the model performs on completely new data.

A 70-15-15 split balances learning, tuning, and evaluation.

The validation set helps improve model performance before testing.

The test set remains untouched until the final evaluation, ensuring an unbiased performance check.

A computer code with blue text

AI-generated content may be incorrect.

Fig16: data partition

for Clustering:

Unlike supervised learning, where data is split into training and testing sets, clustering involves unsupervised learning, meaning there is no outcome variable to predict. Therefore, traditional data partitioning is not required. Instead, the entire dataset will be used to identify natural groupings within the data.

**8. Model selection:**

**8.1. Model selection for clustering:**

K-Means Clustering:

K-Means clustering is selected for segmenting members based on fitness behavior, demographics, and health metrics. This algorithm efficiently groups individuals with similar characteristics, allowing the identification of distinct fitness segments. K-Means is scalable and performs well with large datasets, making it suitable for analyzing member data to develop personalized workout plans. The model helps uncover underlying patterns that contribute to optimizing fitness programs and tailoring recommendations for different member groups.

**Necessity of Normalization:**

In this project, data normalization is crucial because the dataset contains variables with different scales and measurement units. For example:

Age ranges from around 20 to 80, while Weight (kg) is between 40 and 100.

Calories Burned can be in the hundreds or thousands, while Water Intake (liters) is in a small range (e.g., 1-4).

Binary categorical variables (e.g., Gender, Workout Type) are represented as 0 or 1.

Since K-Means clustering uses Euclidean distance, features with larger numerical ranges (e.g., Calories Burned) will dominate distance calculations, making other important features (e.g., Water Intake, Experience Level) less impactful. Z-score normalization (standardization) ensures all features contribute equally by transforming them to have:

Mean = 0

Standard deviation = 1

This prevents bias in cluster assignments and improves the interpretability of segmentation results.

For this analysis, Z-score Standardization is chosen to ensure that all variables have equal importance in distance calculations. The transformation is performed as follows:

* X is the original value
* μ is the mean of the feature
* σ is the standard deviation

A close up of a text

AI-generated content may be incorrect.

Fig17: z-score normalization

**Determining the Optimal Number of Clusters using the Elbow Method:**

To ensure that the K-Means clustering model accurately segments fitness center members, the Elbow Method was used to determine the optimal number of clusters. The Elbow Method helps identify the ideal number of clusters by analyzing the Within-Cluster Sum of Squares (WSS), which measures the variance within each cluster.

A plot of WSS vs. the number of clusters (K) was generated, where the point at which the WSS curve starts to level off (forming an "elbow") indicates the best value of K. This is the point where adding more clusters no longer significantly reduces intra-cluster variance, ensuring a balance between model complexity and interpretability.

Based on the Elbow Method plot, the optimal number of clusters was selected. This ensures that the segmentation effectively groups members with similar fitness behaviors, demographics, and health metrics, allowing for personalized workout plan recommendations.

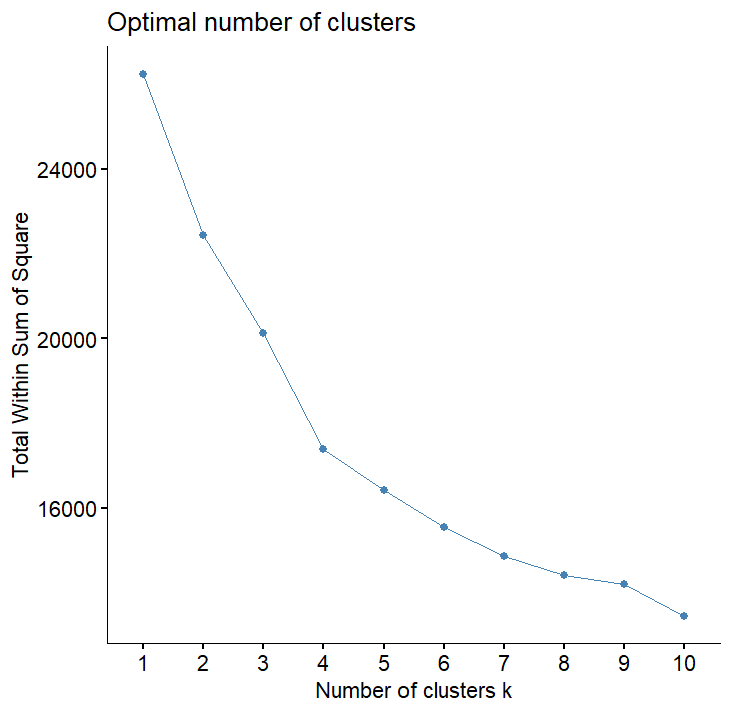


Fig18: Elbow method plot

The Elbow Method plot shows a sharp decline in Total Within-Cluster Sum of Squares (WSS) from K = 1 to K = 3, indicating that adding these clusters significantly reduces intra-cluster variance and improves segmentation. However, after K = 3, the rate of decline slows down, meaning that adding more clusters results in only marginal improvements in WSS while increasing model complexity.

By selecting K = 3, achieve a balance between compactness and interpretability ensuring that the clusters are distinct enough to capture meaningful differences in fitness behavior while avoiding over-segmentation that could lead to unnecessary complexity. This optimal choice allows for effective member segmentation, enabling tailored workout recommendations based on fitness levels, demographics, and health metrics.

After determining that 3 clusters is the optimal choice, the K-Means clustering algorithm was applied to segment the fitness center members.

Based on your clustering results from the fitness\_clustering$Cluster table, where the cluster sizes are as follows:

* Cluster 1: 399 members
* Cluster 2: 383 members
* Cluster 3: 191 members

A screen shot of a computer

AI-generated content may be incorrect.

Fig: Number of clusters

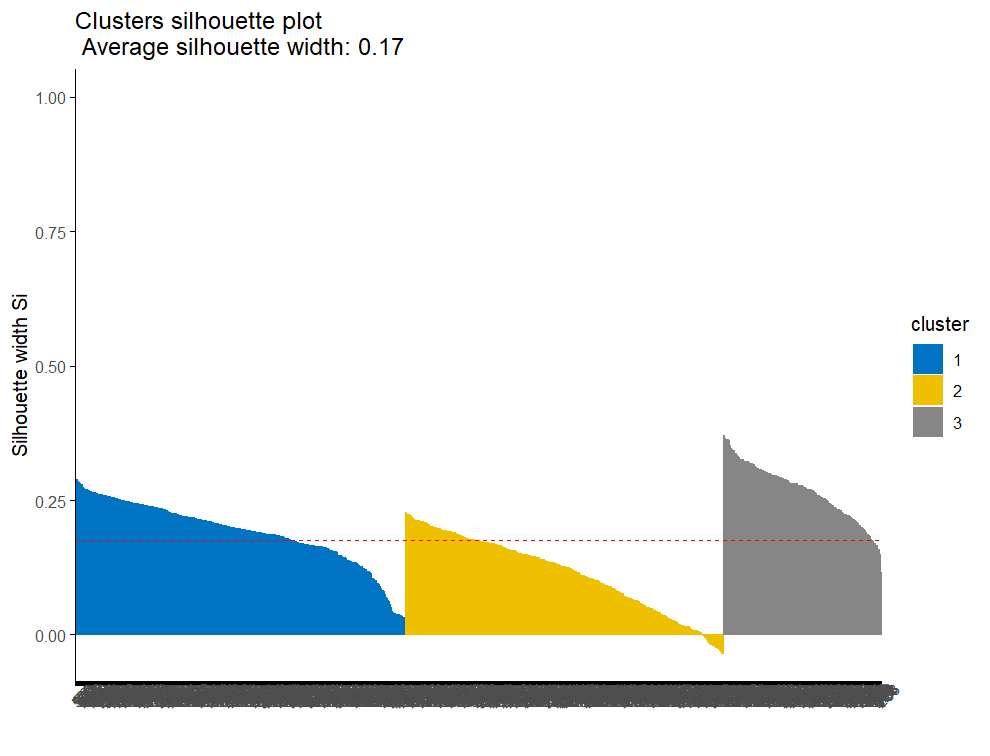


Fig19: silhouette plot for K-means clustering

Interpretation of K-Means Clustering Results

The silhouette analysis evaluates how well data points fit within their assigned clusters.

The dataset has 973 data points grouped into 3 clusters.

The average silhouette width for each cluster is:

* Cluster 1 (399 points): 0.19
* Cluster 2 (383 points): 0.11
* Cluster 3 (191 points): 0.27

Key Insights:

Silhouette width ranges from -1 to 1, where values closer to 1 indicate well-defined clusters, and values close to 0 or negative suggest overlap or poor clustering.

Cluster 3 (0.27) has the best separation, meaning its points are more distinct.

Cluster 2 (0.11) has the lowest silhouette width, suggesting weaker cluster definition and possible overlap with other clusters.

The overall mean silhouette score is 0.17, indicating a moderate clustering structure but with room for improvement.

Conclusion:

The clustering structure is somewhat meaningful but could be improved. Adjusting the number of clusters may enhance separation and improve results.

Hierarchical Clustering:

Hierarchical clustering is included to provide a structured view of member segmentation. Unlike K-Means, it creates a hierarchy of clusters, allowing for better visualization through dendrograms. This helps in understanding relationships between different fitness segments, identifying subgroups within broader clusters, and refining personalized workout recommendations. Hierarchical clustering is particularly useful when determining the optimal number of clusters and analyzing how members with similar workout patterns are related.

The hierarchical clustering process is implemented using the hclust() function, which applies the Ward.D2 method to minimize variance within clusters. The dist() function calculates Euclidean distances between data points to determine their similarity.

To visualize the clustering structure, a dendrogram is created using the as.dendrogram() and color\_branches() functions, which highlight the three clusters with distinct colors.

Clustering quality is evaluated using the silhouette() function, which computes silhouette values for each point. These values indicate how well data points fit within their assigned clusters. A summary() function is used to display the average silhouette width for each cluster, and the fviz\_silhouette() function is applied to generate a silhouette plot. This plot helps assess the cohesion and separation of clusters, guiding decisions on the optimal number of clusters.

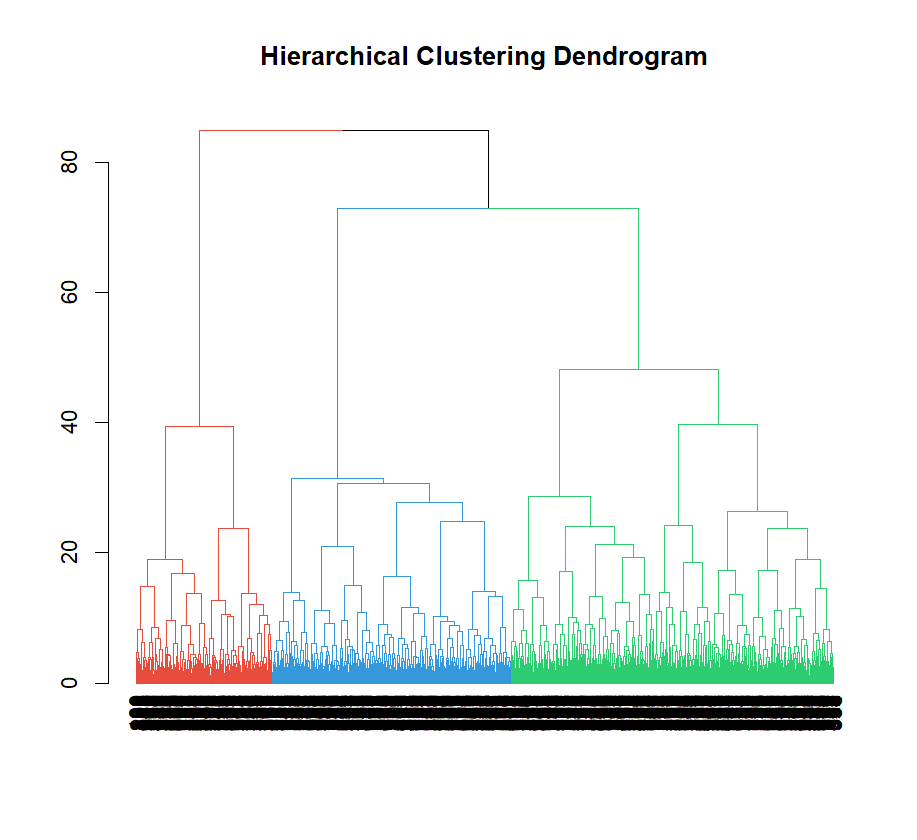


Fig20: Hierarchical clustering dendrogram

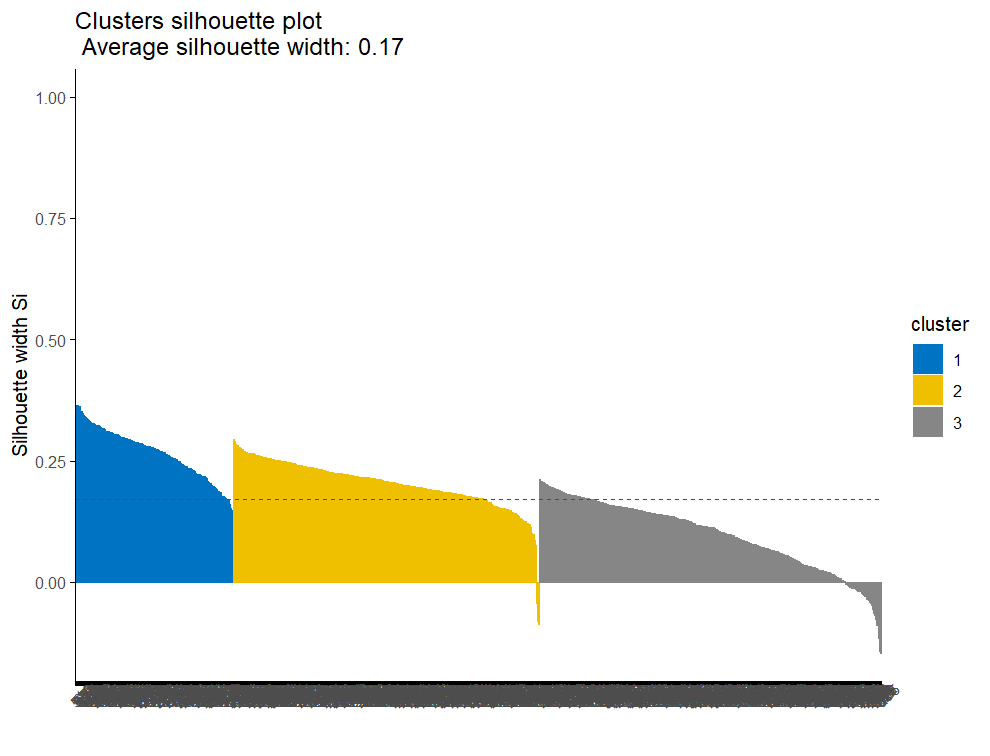


Fig21: silhouette plots for Hierarchical clustering

A black text on a white background

AI-generated content may be incorrect.

Fig22: number of clusters

The average silhouette width for each cluster is:

Cluster 1 (191 points): 0.27

Cluster 2 (369 points): 0.20

Cluster 3 (413 points): 0.10

The silhouette analysis for hierarchical clustering indicates moderate clustering quality with:

* The average silhouette score is 0.17, which suggests that the overall clustering is not very strong.
* Cluster 1 (0.27) has the best-defined separation, meaning its members are well-grouped.
* Cluster 2 (0.20) has a moderate silhouette score, indicating some overlap with other clusters.
* Cluster 3 (0.10) has the weakest separation, meaning many points may be misclassified or belong to multiple clusters.
* Some points have negative silhouette values, indicating that they may be incorrectly assigned to a cluster.

**8.2. Model selection for Regression.**

Multiple linear Regression:

Multiple linear regression is selected to predict calorie expenditure based on workout type, session duration, and health metrics. This model quantifies the relationship between multiple independent variables and calorie burn, helping determine which factors significantly influence energy expenditure. The interpretability of multiple linear regression enables fitness experts to identify key workout attributes that maximize calorie burn, aiding in the optimization of training programs.

For the linear regression model, a training dataset (train\_Fitness) is used to fit the model with the lm() function, where Calories\_Burned is the dependent variable, and all other variables in the dataset are included as predictors. The model's performance is evaluated by examining the summary of the model, which provides insights into the coefficients, R-squared value, p-values, and residuals. These metrics help assess the significance of each predictor in explaining the variation in Calories\_Burned and the overall fit of the model. The summary allows for interpretation of the strength and direction of the relationships between the predictors and the target variable.

A screenshot of a calories burned

AI-generated content may be incorrect.

Fig23: Multi linear regression

The evaluation metrics for the Multiple Linear Regression model are as follows:

Mean Absolute Error (MAE): 51.3

This value represents the average magnitude of errors in the predictions. On average, the model's predictions are off by approximately 51.32 units. A lower MAE indicates better accuracy, but it’s important to compare it with the scale of the target variable.

Mean Squared Error (MSE): 4092.6

This metric penalizes larger errors more heavily, as it squares the differences between predicted and actual values. The lower the MSE, the better the model's predictions. In this case, the MSE suggests some variability in the model’s performance.

Root Mean Squared Error (RMSE): 63.97

RMSE is the square root of MSE and provides the error in the same units as the target variable. It suggests that the model's predictions are, on average, about 63.97 units away from the actual values. Again, lower values are better.

R-squared: 0.94

This indicates that 94% of the variance in the target variable is explained by the model. An R-squared value closer to 1.0 suggests that the model is a good fit for the data.

Conclusion:

The model has strong performance, with an R-squared value of 0.94, indicating that it explains most of the variability in the target. The MAE, MSE, and RMSE values show a moderate level of prediction error, but they are still relatively low, suggesting that the model is making accurate predictions overall.

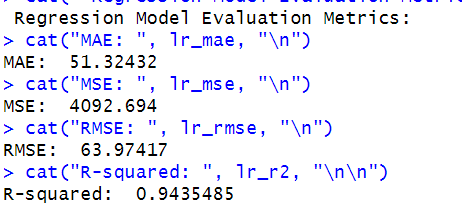


Fig24: Evaluation metrics for Regression

**Random forest:**

Random Forest regression is selected to predict calorie expenditure based on workout type, session duration, and health metrics. This ensemble learning method builds multiple decision trees and averages their predictions, improving accuracy and reducing overfitting. Unlike linear regression, Random Forest can capture complex, non-linear relationships between independent variables and calorie burn, making it a powerful tool for fitness analytics.

For the Random Forest model, a training dataset (train\_Fitness) is used to fit the model using the randomForest() function, where Calories\_Burned is the dependent variable, and all other available features serve as predictors. The model's performance is assessed using key evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²). These metrics help determine the model's predictive accuracy and its ability to generalize to unseen data. Additionally, feature importance analysis identifies the most influential factors affecting calorie burn, providing valuable insights for optimizing workout programs based on key fitness attributes.

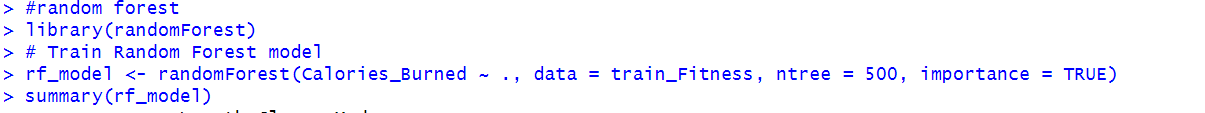


Fig25: random forest

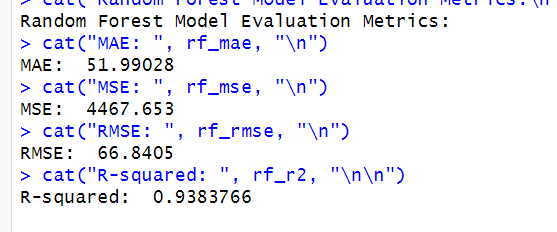


Fig26: Evaluation of metrics for random forest

Evaluation Metrics for the Random Forest Model

Mean Absolute Error (MAE): 51.99

This value represents the average magnitude of errors in the predictions. On average, the model's predictions deviate by approximately 52.99 units from the actual values. A lower MAE suggests better accuracy, though it should be evaluated relative to the scale of the target variable.

Mean Squared Error (MSE): 4467.653

MSE measures the average squared differences between predicted and actual values, penalizing larger errors more heavily. The MSE of 4467.653 indicates some variability in predictions but remains relatively low, suggesting reasonable performance.

Root Mean Squared Error (RMSE): 66.84

RMSE, the square root of MSE, provides an error measurement in the same units as the target variable. This means the model's predictions are, on average, about 66.95 units away from the actual values. A lower RMSE indicates better predictive accuracy.

R-squared (R²): 0.9383

The R² value of 0.93 indicates that 93% of the variance in calorie expenditure is explained by the model’s predictors. This suggests a strong fit, meaning the model effectively captures the relationship between workout attributes and calorie burn.

Conclusion:

The Random Forest model demonstrates strong predictive performance, with an R² of 0.93, showing that it explains most of the variance in the target variable. While the MAE, MSE, and RMSE values indicate some degree of error, they are relatively low and comparable to those of the Multiple Linear Regression model. Random Forest’s ability to capture complex, non-linear relationships makes it a valuable tool for predicting calorie expenditure, particularly when interactions between variables are significant.

**Report on model performance:**

**Regression:**

**Model selection:**

The goal is to identify the most effective workout types and intensities that maximize calorie expenditure and overall fitness improvements. To achieve this, two models Multiple Linear Regression (MLR) and Random Forest were evaluated based on their predictive performance.

|  |  |  |
| --- | --- | --- |
| Metric | Multiple Linear Regression | Random forest |
| MAE (Mean Absolute Error) | 51.32 | 51.99 |
| MSE (Mean Squared Error) | 4092.69 | 4467.653 |
| RMSE (Root Mean Squared Error) | 63.97 | 66.84 |
| R-squared | 0.94 | 0.93 |

Interpretation:

Analysis of Model Performance

Predictive Accuracy:

* Both models achieve a high R² value (0.94), meaning they explain 94% of the variance in calorie expenditure.
* MLR has slightly lower error values (MAE, MSE, RMSE) compared to RF, indicating marginally better predictive accuracy.

Interpretability vs Complexity:

* MLR provides clear relationships between input variables (e.g., session duration, workout intensity) and calorie burn, making it easier to interpret and apply insights.
* RF is more complex and less interpretable, as it captures non-linear relationships but functions as a black-box model.

Generalization & Practical Use:

* MLR is simpler and more interpretable, making it easier to derive actionable fitness recommendations.
* RF may capture complex interactions but doesn’t offer significant improvement in predictive performance.

Choosing the Best Model:

Since the primary objective is to understand the key workout factors influencing calorie burn and provide practical fitness recommendations, the Multiple Linear Regression (MLR) model is the preferred choice. It offers strong performance, high interpretability, and actionable insights, making it the best model for this project.

**Clustering:**

**Model Selection**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Clustering Method | Overall Average Silhouette Score | Cluster 1 | Cluster 2 | Cluster 3 |
| K-Means Clustering | 0.17 | 0.19 | 0.11 | 0.27 |
| Hierarchical Clustering | 0.17 | 0.27 | 0.20 | 0.10 |

Both clustering techniques resulted in an overall average silhouette width of 0.17, indicating similar clustering quality. However, Hierarchical Clustering produced higher silhouette scores in two out of three clusters (0.27 and 0.20) compared to K-Means (0.27 and 0.19). This suggests that Hierarchical Clustering maintains better-separated and well-defined clusters.

Additionally, Hierarchical Clustering does not require specifying the number of clusters in advance, making it more suitable for exploratory analysis. However, it can be computationally intensive for larger datasets. In contrast, K-Means is more scalable and efficient but requires predefined cluster numbers, which may affect the results if not chosen correctly.

Based on the silhouette analysis, Hierarchical Clustering is the preferred method for this dataset, as it provides better-defined clusters while offering flexibility in determining the optimal cluster structure.

**9. Model improvement and Evaluation**

**Regression:**

Model Improvement using Forward Selection:

Forward Selection was used to identify the most relevant predictors for the Calories\_Burned variable in the multiple linear regression model.

Null Model: The process begins by defining a null model that contains no predictors, only the intercept (mean of the dependent variable, Calories\_Burned). This was created using the function lm(Calories\_Burned ~ 1, data = train\_Fitness).

Full Model: A full model was constructed using the function lm(Calories\_Burned ~ ., data = train\_Fitness) to include all available predictors (features) in the dataset.

Forward Selection Process: The stepAIC function from the MASS package was applied for forward selection. This function evaluates the Akaike Information Criterion (AIC) to determine the best-fitting model by adding one predictor at a time. The function stepAIC() considers all variables and adds the one that reduces the AIC the most. The process continues iteratively until no further improvements are possible.

Final Model: The output of stepAIC() provides the final model with only those predictors that significantly contribute to predicting Calories\_Burned. The selected variables are identified based on their AIC values and statistical significance.

A screenshot of a computer

AI-generated content may be incorrect.

Fig27: Forward selection

Model Comparison: The R-squared and Adjusted R-squared values are almost identical, indicating that the set of predictors after forward selection explains almost the same amount of variance as the full model.

Residuals: The residual standard errors are also very close, meaning that the predictive performance is similar for both models.

Significance: The forward selection method removed non-significant predictors (e.g., Experience\_Level, Heart\_Rate\_Reserve, etc.), resulting in a more parsimonious model, which is likely to generalize better.

Conclusion:

Forward selection has streamlined the model, removing predictors that were not adding significant value to the prediction of Calories Burned. Although the model's explanatory power and residuals haven't changed much, the reduced number of predictors should lead to better generalization and simpler interpretation.

**Clustering:**

Model Improvement: Analyzing the Silhouette Scores of Hierarchical Clustering

To evaluate the quality of our hierarchical clustering model,utilized the silhouette scores, which measure how similar each point is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a higher score indicates that the data points are well-matched within their own clusters, and poorly matched to neighboring clusters.

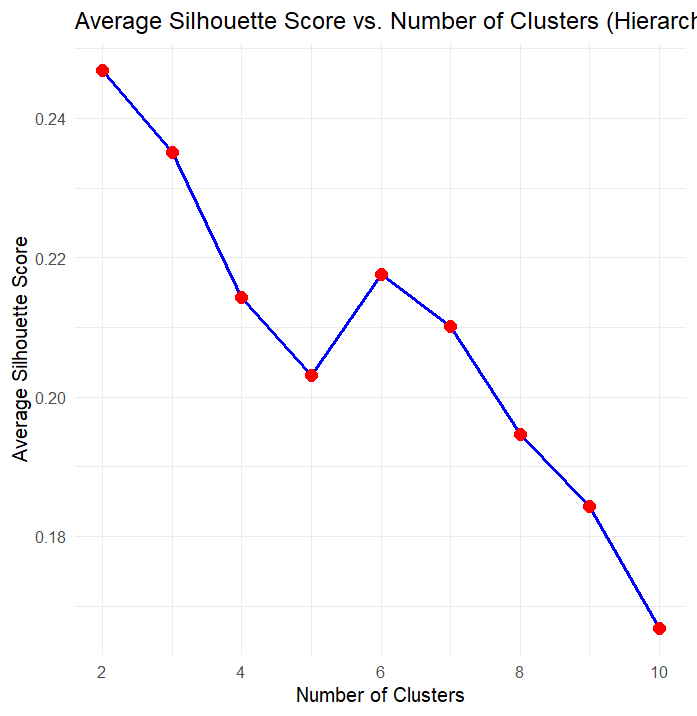


Fig28: silhouette score vs number of clusters

**Results:** After performing hierarchical clustering for different numbers of clusters (k = 2 to 10), the following average silhouette scores were observed:

* For 2 clusters, the silhouette score was 0.246.
* For 3 clusters, it decreased slightly to 0.235.
* As the number of clusters increased, the silhouette scores continued to decline, reaching as low as 0.166 for 10 clusters.

This pattern suggests that while 2 clusters provide some separation, increasing the number of clusters beyond 2 does not significantly improve the clustering quality. The silhouette scores decrease progressively as more clusters are added, indicating that additional clusters do not enhance the model's ability to distinguish data points effectively.

**Interpretation:**

* **Clusters = 2:** The silhouette score of 0.246 suggests a basic level of separation, but the clustering might be too simplistic and not capture the full complexity of the data.
* **Clusters = 3:** With a slight decrease to 0.235, the 3-cluster solution offers a better balance between complexity and meaningful separation compared to 2 clusters, making it a more effective choice.
* **Clusters = 4 to 6:** The silhouette scores continue to drop, with values ranging from 0.214 to 0.217. This indicates that while more clusters provide more segmentation, they do not meaningfully improve data separation and may introduce unnecessary complexity.
* **Clusters = 7 to 10:** As the number of clusters increases, the silhouette scores decrease further, with values approaching 0.166, suggesting overfitting and diminishing returns from adding more clusters.

Based on the silhouette analysis, the model performs optimally with 3 clusters. While 2 clusters offer a simple approach, they may not fully capture the complexity of the data. The 3-cluster solution strikes a better balance by providing a reasonable level of separation between groups, with a silhouette score that is higher than those for 4 or more clusters.

Thus, 3 clusters are preferred as the optimal number for this dataset, as it delivers meaningful segmentation without the risk of overfitting or unnecessary complexity.

**Using Manhattan distance:**

Using Manhattan distance in your hierarchical clustering model offers several advantages, particularly when dealing with high-dimensional or sparse data. Manhattan distance, also known as city block distance, computes the sum of the absolute differences between corresponding features of two data points. This approach is less sensitive to outliers compared to Euclidean distance, making it well-suited for datasets with varying scales or where features are not necessarily linearly related. Additionally, Manhattan distance is computationally less intensive when handling datasets with a large number of dimensions, as it avoids the complexity of squaring the differences as seen in Euclidean distance. It also performs better in scenarios where the data points are expected to align along axis-aligned grids, such as in clustering based on categorical or discrete features. Given these characteristics, Manhattan distance is a good choice for your clustering task, offering robustness and efficiency, particularly when the data has complex or varied structures.

**Model improvement by different linkage methods:**

After performing hierarchical clustering for different linkage methods (single, complete, average, and ward.D2) and cutting the tree into 3 clusters, the following silhouette scores were observed:

Single linkage: 0.11

Complete linkage: 0.21

Average linkage: 0.17

Ward.D2 linkage: 0.24

Interpretation:

* Single linkage (0.11): The silhouette score is quite low, indicating poor cluster separation and potential chaining effects, where clusters are linked through single points.
* Complete linkage (0.21): The silhouette score is moderate, suggesting better separation between clusters compared to single linkage, but it is not the highest among the methods tested.
* Average linkage (0.17): This method also provides a reasonable separation, but its silhouette score is not significantly better than complete linkage.
* Ward.D2 linkage (0.24): This method provides the highest silhouette score, indicating the best clustering separation and cohesion. The ward.D2 method tends to minimize the variance within clusters, leading to more compact and well-separated clusters.

Given the results from both the linkage method testing and the initial silhouette analysis with 2-10 clusters, it is evident that the 3-cluster solution with Ward.D2 linkage is the most effective. The silhouette score of 0.24 indicates that the 3-cluster configuration offers the best separation of data points while maintaining compact clusters.

Conclusion for improvement model:

Based on the silhouette analysis, the optimal number of clusters for this dataset is 3, and the best linkage method is Ward.D2, which produces the highest silhouette score (0.24). This configuration provides the most meaningful and well-separated clusters, balancing cluster cohesion and simplicity. While other linkage methods (such as complete and average linkage) also provide reasonable separation, Ward.D2 linkage stands out as the best option for improving the clustering model.

Therefore, using the 3-cluster solution with Ward.D2 linkage as the final model for this dataset. Further refinements could involve experimenting with other clustering algorithms or tuning the feature space, but this configuration offers the best balance between clustering quality and interpretability.

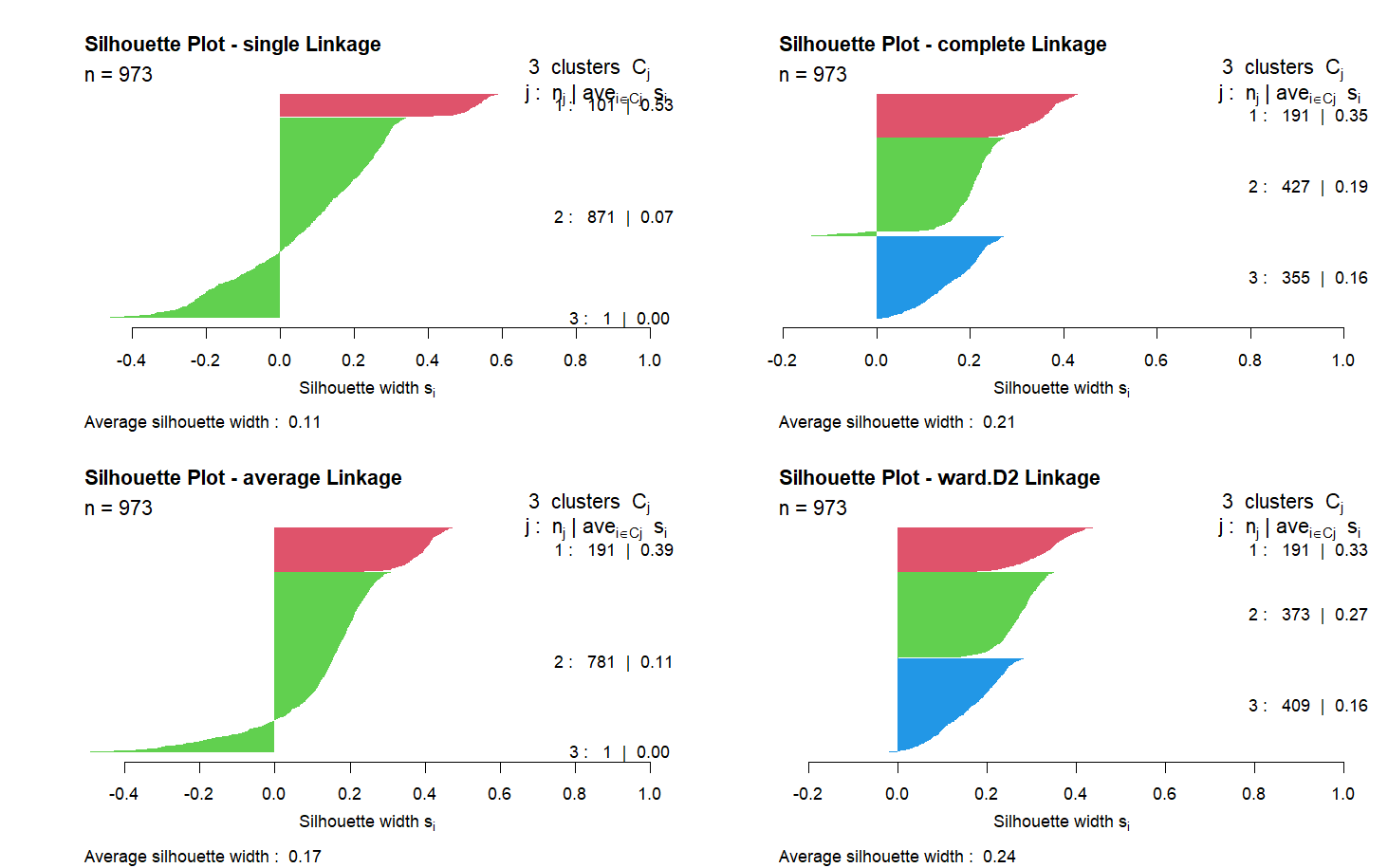


Fig29: Linkage methods

**10. Observations:**

**Regression:**

**Coefficients Interpretation:**

The regression analysis offers valuable insights into how various factors influence calorie expenditure, directly addressing the business problem of gaining insights on workout impact. These findings support FitLife Wellness in achieving its goal of identifying the most effective workout types and intensities to maximize calorie burn and overall fitness improvements.

**1. Session Duration:**

Coefficient (596.2552): Session duration has a significant positive impact on calorie expenditure. For every additional hour of workout, members burn approximately 596 more calories, highlighting the importance of encouraging longer, well-structured sessions.

FitLife Wellness can recommend extended workout sessions for members targeting higher calorie burn, particularly those on weight loss or endurance programs. Personalized session duration suggestions can enhance member satisfaction and outcomes.

**2. Workout Intensity (Average BPM):**

Coefficient (6.1229): A 1-unit increase in Average BPM results in burning approximately 6 additional calories. This confirms that higher workout intensity is closely linked to increased calorie expenditure.

FitLife Wellness should promote heart rate monitoring for members to maintain an optimal intensity zone. Personalized training programs can be designed for individuals aiming to maximize calorie burn through cardio-intensive workouts.

**3. Hydration:**

Water Intake Coefficient (74.0843): Every additional liter of water consumed is associated with burning around 74 extra calories.

Hydration Need Ratio Coefficient (-39.1568): A higher hydration need ratio negatively impacts calorie burn, reducing calories by approximately 39 calories per unit increase.

Proper hydration is essential for maximizing workout effectiveness. FitLife Wellness can introduce hydration tracking as part of the fitness app, offering reminders and personalized hydration goals for members to optimize performance and calorie expenditure.

**4. Fat Percentage:**

Coefficient (-3.5417): Higher body fat percentage reduces calorie burn, with an estimated 3.54 fewer calories burned for every 1% increase in fat percentage. This indicates that leaner individuals tend to burn calories more efficiently during exercise.

FitLife Wellness can design specialized fat-reduction programs combining cardio and strength training for members with higher body fat percentages. Promoting body composition tracking and progress visualization can further engage members.

**5. Experience Level:**

Coefficient (-13.6988): More experienced members burn fewer calories compared to beginners, with approximately 13.7 fewer calories burned per unit increase in experience level. This is likely due to increased workout efficiency among experienced members.

FitLife Wellness can offer progression-based training programs for experienced members focusing on muscle-building or endurance, ensuring they remain engaged. For beginners, calorie-focused programs can be emphasized to maintain motivation.

**Clustering:**

**A screenshot of a graph

AI-generated content may be incorrect.**

Fig30: clustering centroids interpretation

The clustering results provide valuable insights into different member profiles based on their workout behaviors, health metrics, and demographics. These clusters can help FitLife Wellness design more personalized and engaging fitness experiences for its members. The three distinct clusters identified represent unique member segments, each with specific preferences and behaviors.

Cluster 1: Active & Hydrated Members

Age, Weight, Height, and BMI:

Members in this cluster have average values for age, weight, and BMI (e.g., BMI ≈ 25). They likely represent individuals with balanced physiques and established fitness routines.

Max BPM and Avg BPM:

They exhibit moderately high heart rates during workouts, reflecting consistent cardiovascular engagement (e.g., avg BPM ≈ 130).

Session Duration:

Their workouts are moderately long, averaging 1.47 hours per session, suggesting they maintain a regular fitness habit.

Calories Burned & Hydration:

This group burns relatively high calories per session, and hydration levels are well-managed (hydration ratio ≈ 0.95), showing good awareness of hydration needs.

Workout Frequency & Type:

They work out around 1.33 times per week on average, with a preference for cardio-based activities.

Fat Percentage & Hydration Needs:

Their fat percentage is moderate to low, and hydration needs appear well-balanced.

Interpretation:

This segment reflects dedicated fitness enthusiasts who may benefit from more personalized, progressively challenging workout programs focusing on cardiovascular health and calorie optimization. FitLife could also offer nutritional guidance to enhance their fitness outcomes.

Cluster 2: Beginner Members

Age, Weight, Height, and BMI:

Members in this cluster tend to have slightly higher BMI values (e.g., avg BMI ≈ 29.5) compared to other groups, suggesting they may be newer to fitness or working towards weight management.

Max BPM and Avg BPM:

They exhibit lower average and max heart rates during workouts (e.g., avg BPM ≈ 115), indicating lower workout intensity.

Session Duration:

Their average session duration is about 1.12 hours, slightly shorter than Cluster 1.

Calories Burned & Hydration:

They burn fewer calories per session, and hydration intake appears lower than recommended (hydration ratio ≈ 0.85), signaling an opportunity for hydration education.

Workout Frequency & Type:

Members in this group exercise less frequently, with preferences leaning towards Strength or Yoga-based workouts, which are generally lower intensity compared to cardio or HIIT.

Fat Percentage & Hydration Needs:

This group shows a higher fat percentage and slightly elevated hydration needs, consistent with their higher BMI and lower activity level.

Interpretation:

This segment represents individuals at the early stages of their fitness journey. FitLife can support them with beginner-friendly programs focused on building strength, flexibility, and gradually improving cardiovascular fitness. Providing guidance on hydration and weight management will further support their progress.

Cluster 3: Experienced & Fitness-Oriented Members

Age, Weight, Height, and BMI:

Members in this cluster have a lower BMI than other groups (e.g., avg BMI ≈ 23.5), indicating a leaner, more athletic body composition.

Max BPM and Avg BPM:

They display higher average and max heart rates (e.g., avg BPM ≈ 140), reflecting that they engage in more intense workouts.

Session Duration:

Their workouts average 1.12 hours per session, similar to Cluster 2, but their intensity level is higher.

Calories Burned & Hydration:

They burn moderate calories per session (relative to their body size and workout duration), with hydration needs well-balanced (hydration ratio ≈ 0.94), showing good hydration management.

Workout Frequency & Type:

Members in this cluster work out regularly and at higher intensity, preferring Strength and HIIT workouts.

Fat Percentage & Hydration Needs:

Their fat percentage is lower and hydration needs are balanced, supporting a well-rounded fitness routine.

Interpretation:

This group consists of experienced, fitness-oriented members who would benefit from advanced programs focusing on strength training, endurance building, and high-intensity interval training (HIIT). FitLife could provide specialized challenges, performance tracking tools, and tailored nutritional advice to help them push their fitness goals further.

**11. Conclusion:**

The regression and clustering analyses conducted provide valuable insights into how FitLife Wellness can optimize its workout programs to achieve better fitness outcomes for its members.

The regression analysis provided valuable insights into the factors influencing calorie expenditure, supporting FitLife Wellness in its goal of optimizing workout programs for better fitness outcomes. The model demonstrated a strong predictive capability, explaining approximately 94.9% of the variance in calories burned.

Key findings indicate that session duration and workout intensity (measured by average BPM) are the most significant contributors to calorie burn. Additionally, factors like hydration levels, fat percentage, and experience level play a notable role. Proper hydration enhances calorie expenditure, while higher body fat percentage and greater experience levels are associated with reduced calorie burn, reflecting the impact of body composition and workout efficiency.

To address the business problem of gaining insights on workout impact, FitLife Wellness can implement data-driven strategies to:

* Recommend personalized workout plans based on members' fitness goals.
* Provide heart rate-based training programs to optimize calorie burn.
* Encourage proper hydration management to enhance performance.
* Develop targeted fat reduction programs for improved calorie expenditure.
* Offer progression-based training for experienced members to ensure continuous improvement.

The clustering analysis further identified three distinct segments within the member base: Active & Hydrated Members, Beginner Members, and Experienced & Fitness-Oriented Members. Understanding these segments allows FitLife Wellness to tailor fitness programs to meet the unique needs of each group, enhancing member satisfaction and engagement.

Tailored Fitness Programs:

For Active & Hydrated Members: Design advanced, high-intensity programs that challenge cardiovascular and muscular endurance while incorporating personalized nutrition and recovery strategies.

For Beginner Members: Provide introductory programs focusing on building foundational strength, flexibility, and cardiovascular fitness. Educate these members on hydration, recovery, and calorie management to ensure balanced progress.

For Experienced & Fitness-Oriented Members: Develop specialized regimens that include strength training and HIIT to maintain high fitness levels. Offer advanced tracking tools to monitor progress and prevent plateaus.

By leveraging these insights from both regression and clustering analyses, FitLife Wellness can foster a more engaging and motivating fitness experience. This will not only help optimize fitness programs but also contribute to long-term member retention and improved health outcomes.