TU Dublin Tallaght

Analysing Fitness Patterns and Predicting Calorie Expenditure for Gym Members

A Data Analysis and Predictive Modelling Report

By Alan Walker X00204054 & Roko Skugor X00203402

Computing with IT Management (TA\_KACTM\_B)

Abdelsalam Busalim

08.12.2024

Contents

[Executive Summary 3](#_Toc184592887)

[Methodology 3](#_Toc184592888)

[Data Cleaning 3](#_Toc184592889)

[Initial Data Exploration 3](#_Toc184592890)

[Data Analysis and Visualization 3](#_Toc184592891)

[Predictive Modelling 3](#_Toc184592892)

[Conclusion 4](#_Toc184592893)

[Introduction 4](#_Toc184592894)

[Methodology 5](#_Toc184592895)

[Data Description 5](#_Toc184592896)

[Tools Used 5](#_Toc184592897)

[Data Cleaning 5](#_Toc184592898)

[Initial Data Exploration 6](#_Toc184592899)

[Data Analysis and Visualization 7](#_Toc184592900)

[Demographic insights 7](#_Toc184592901)

[Predictive Modelling 9](#_Toc184592902)

[Objective 9](#_Toc184592903)

[Modelling 9](#_Toc184592904)

[Conclusion 10](#_Toc184592905)

[References 10](#_Toc184592906)

# Executive Summary

## Methodology

The report leveraged a structured approach to analyse a dataset of 973 records sourced from Kaggle. This dataset, formatted as a CSV, includes 18 variables encompassing demographics, physiological metrics, and workout details. A Jupyter Notebook was utilized for the project, enabling step-by-step documentation and the use of key Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn for data manipulation, visualization, and machine learning.

## Data Cleaning

The dataset underwent cleaning to ensure accuracy and consistency. Missing values were not an issue in this dataset; however, outliers were addressed using the Interquartile Range (IQR) method. Categorical data were standardized and encoded for model readiness, and numerical inconsistencies like negative values were corrected. Additional contextual columns, such as a categorized BMI indicator and readable experience levels, were created to enhance analysis. These efforts ensured a reliable and actionable dataset.

## Initial Data Exploration

Preliminary exploration revealed key insights:

* The average BMI of the dataset was 24.91, indicating that this weight distribution might be higher than the average.
* A strong correlation was observed between workout session duration and calories burned.
* Gender-based trends were identified, showing males tended to burn more calories than females for similar workout durations.
* A correlation matrix highlighted relationships such as higher heart rates correlating with increased calorie burn and shorter workout durations being associated with higher body fat percentages.

## Data Analysis and Visualization

Detailed visualizations, including scatter plots, histograms, and box plots, revealed:

* Males made up a slightly larger proportion (52.5%) of gym-goers compared to females.
* Experience levels significantly impacted calorie burn, with experts consistently outperforming beginners and intermediates.
* Gender differences were noted in workout preferences and calorie burn distribution, with males favouring yoga and females preferring HIIT workouts.
* The calorie burn distribution was slightly negatively skewed, with a few notable outliers likely tied to specific workout types or durations.

## Predictive Modelling

A Linear Regression model was developed to predict calories burned during workouts. The dataset was split into training (80%) and testing (20%) sets. The model achieved a Root Mean Square Error (RMSE) of 70 calories, indicating a reasonable accuracy given the dataset’s range (300–1600 calories). Predictions were most accurate within the 600–1000 calorie range but less reliable for higher values, likely due to limited data at those extremes.

## Conclusion

The analysis confirms that calorie burn is influenced by multiple factors, including workout duration, intensity, and experience level. Key findings include:

* Longer workouts, especially by experienced gym-goers, are more efficient in calorie burn.
* Gender differences in calorie burn are minor, emphasizing that consistency and experience are more critical than demographics.
* Workout type has a lesser impact on calorie burn than the act of exercising itself.
* These insights reinforce the value of consistent and informed exercise routines over specific workout methodologies for optimizing calorie burn and achieving fitness goals.

# Introduction

Data analysis is an essential tool when it comes to learning about reasons and coming to conclusions. In the context of gym-going and optimizing workouts, data analysis is essential to use as it can provide unique and valuable insights as to what variables should be prioritised and which ones cannot be affected much. The aim of this report is to find out what factors are the ones that influence the burning of calories the most, which demographics are affected most by the changing of variables, and to give a recommendation of what should be done based on an individual’s aspects. Besides this, it is important to check if the data being analysed follows common trends and meets expectations, or if what is common knowledge is not best practice.

# Methodology

## Data Description

The data that has been used for this report was a dataset found on Kaggle, the link of which can be found in the references. This CSV dataset that has been used contains 973 records, with each having 18 variables. These variables cover key areas that are important to analyse such as demographics (age, gender), physiological metrics (BMI, average heart rate) and workout details.

## Tools Used

We used a Jupyter notebook to record all our data input, cleaning, processing, output and analysing. This allowed us to annotate our code step by step as it was produced. The main libraries used for this project were:

* Pandas: This python library provides a useful way to interpret and clean data. Data is separated into rows and columns in a Dataframe, which can be easily manipulated and cleaned.
* NumPy: Useful for working with math, provides methods such as a square root method which streamline working with numbers.
* Matplotlib: Essential library used for data visualization.
* Seaborn: Like matplotlib, used for data visualization but has some differences & is a more intuitive tool.
* Scikit-learn: Very useful machine learning package which allows for the training of models such as linear regression and provides the tools for prediction.

## Data Cleaning

Once the data had been imported and passed into a Dataframe, the next step was to clean it. The process of data cleaning encompasses every step done to ensure quality, veracity and validity of data. This includes removing duplicate values, handling missing values, determining and dealing with outliers and formatting categorical values. The first step was to clean out missing values. Luckily, the dataset being used was whole and had no empty fields, so after confirming these facts, there was nothing else to be done. The next step was to identify outliers and handle them, which was done by removing the outliers using the IQR (Interquartile Range) Method.

After this, standardization of the categorical (Non-numerical) values was needed. For example, for the different workout types, the .strip() and .capitalize() methods would be used, which would remove extra whitespace & Capitalize the first letter respectively. To ensure consistency, all time units would be standardized to minutes and the column renamed.

Some important contextual data was also added. A categorical column called “BMI” was created based on the numerical BMI values provided by the dataset, to streamline the data processing and reduce the amount of calculations that needed to be done. The “Experience\_Level” column held numbers, which indicated it was Numerical. To make it easier to read this data, categorical values were mapped onto these numerical values. To ensure that the data represented within the Dataframe would be usable for when models would be trained on it, the workout type column was encoded using the “One Hot” encoding method.

The final few steps when cleaning the data were some minor touch-ups, such as checking and removing duplicates and verifying the data types of each column. If duplicates were not removed these could skew the data and make the data inaccurate, and if the data types were not appropriate (for example, weight being a String) these would mean the data would be incredibly harder to format. The final step was to ensure that the data was consistent in the sense of not having negative values where they shouldn’t be (like a negative amount of calories burned) and once these were all completed the cleaned Dataframe was saved and prepared to be used.

# Initial Data Exploration

A graph of a bar

Description automatically generated with medium confidence Once the data had been cleaned and prepared, some initial logic was applied to the dataset to summarize and describe it, giving an idea of where to start to find trends and dependent factors. Basic data manipulation techniques such as finding the mode, quartiles, mean and standard deviation provides an overview of what to expect from the data.

From the descriptive statistics, some important details can be inferred: from our data set, the average BMI is 24.91 which is bordering overweight. This data is above the average BMI of the general population. Another fact that can be determined from the data is that the amount of calories burned varies considerably from one entry to another. There seems to be a large amount of optimisation that can be done when it comes to burning calories in workouts.

Figure 1: Calories burned per experience level

A correlation matrix mapping correlation between variables.
 A chart that can be used to determine if the data makes sense is comparing the Experience Level to Calories burned. This is visualised in Figure 1, where it can be seen that the expected results match the data: The higher the experience level of an individual, the more calories that individual burns. This is likely because they can follow through and finish workouts more often, and these individuals know how to optimise their exercises to maximize burning calories.

A correlation matrix is an indispensable tool that can be used to visualize correlation between different variables. In this case, the following correlation matrix was produced. From Figure 2, the strongest correlations that can be identified are that the amount of calories burned has a strong positive correlation with a workout session, meaning that longer sessions will strongly increase the amount of calories burned. There are strong negative correlations between fat percentage and both calories burned and session duration, A histogram showing the distribution of calories burned.
but considering the previous fact that session duration and calories burned have such a strong correlation, the meaningful interpretation of these would be that an individual with a higher body fat percentage would have a tendency to work out for a short period of time. Finally, the slightly positive correlation is between calories burned and heart rate. A higher heart rate is indicative of a more intense workout, so it makes sense that this would correlate with more calories burned.

Figure 2: Correlation Matrix comparing variables

The next tool used to visualize the data, shown in Figure 3, was a histogram collecting the distribution of calories burned. This data would give an idea on any trends that could be inferred from the data such as a negative/positive skew, and any outliers or abnormalities. The first and most noticeable fact about this is that the distribution is slightly negatively skewed. This means that although is represented by a normal distribution, there is a slight tendency for individuals to burn less calories than what is expected. Another thing to note is there are some slight outliers, such as there being a spike in frequency when 1600 calories are burned, which may be an indicator of a certain type of workout being a longer duration, or experienced gym-goers knowing how to optimise burning calories.

Figure 3: Histogram showing frequency of calories burned

# Data Analysis and Visualization

## Demographic insights

A blue and red pie chart

Description automatically generated Once the data had been explored and the structure understood, it was time to go more in-depth as to what the data represented. One of the more influential data points that could be relevant was identified to be demographics, and as such the data was visualized by gender. In Figure 4, the first thing to note was that the data indicated that males had a higher propensity to go to the gym. Males make up 52.5% of the gender distribution of gym goers, which is not only above the average (50%), but also as there is a slightly larger number of females in the world than males, it means that the difference between number of male and female gym-goers is larger than this pie chart might indicate.

Figure 4: Gender Distribution in gyms

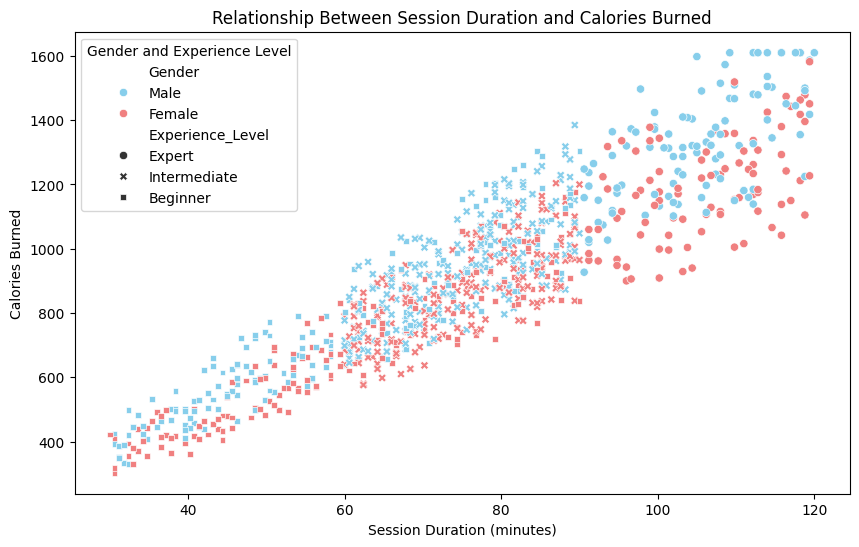
The next step taken was to find the relationship between session duration and calories burned, distinguishing by gender and experience level. Figure 5 presents this data on a scatter plot. The first most noticeable trend is that females have a tendency to burn less calories for workouts of the same duration as men, increasingly noticeable the longer a session lasts. The clear differences between experience levels is also shown here, as there is more or less a clear cut divide between beginner, intermediate and expert gym goers: beginner gym goers almost completely occupy the lower third of the graph, intermediate the middle third, and the experienced gym-goers take up the upper third. This is in line with the data and what is to be expected.

Figure 5: Scatter plot showing calories burned by session duration

A graph of different colored bars

Description automatically generatedA diagram of different colored squares

Description automatically generated This point that females burn less calories on average than men is further exacerbated by the box plot shown in Figure 6, but another interesting phenomena can be observed: not only do females on average have a lower amount of calories burned, but also there is a tighter spread between the calories burned, most visible in the “strength” workout type. Here, the interquartile range is very small, almost as small as the entirety of Q2 for their male counterparts. One thing to note is the fact that outliers can be observed in this box plot, so there might be some data inaccuracies.

Figure 6: Workout type distribution by gender

Figure 7: Box plot for calories burned per workout

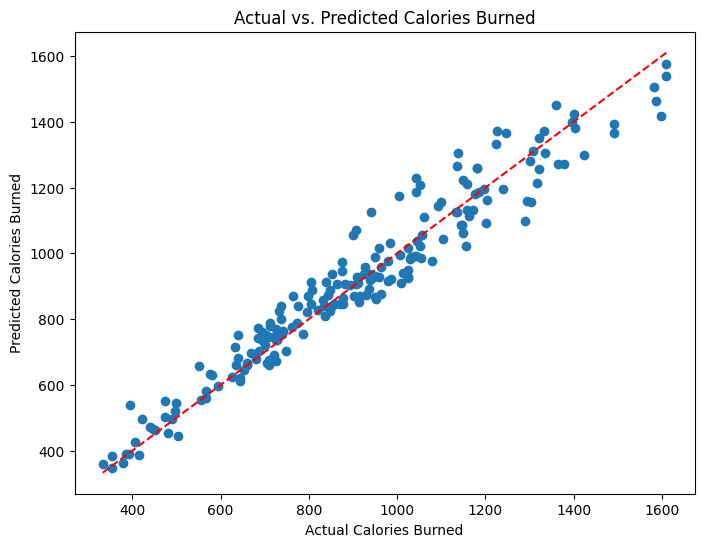
The final data visualisation technique we used for demographics was quite simple, just comparing the distribution of workout types by gender. There isn’t a lot of information that can be taken from this graph, as the difference between male and female distribution is about in line with the demographics shown in Figure 4, except for one small detail: males tend to prefer Yoga over females, females tend to prefer Hiit over males, and the least popular workout type overall is Hiit.

# Predictive Modelling

## Objective

After the data had been collected, cleaned and analysed, an attempt at prediction was made. The goal with this prediction was to be able to determine how many calories would be burned during a workout. The model chosen to carry out these predictions was a Linear Regression model. The most intuitive tool to execute the prediction is known as SciKit Learn, or in short sklearn, which provides many different types of models and verification methods. In our context, the tools we will be using are the Linear Regression model, the train/test split, the prediction method and the scoring methods (mean absolute error, mean squared error and RMSE).

## Modelling

 To start the process of training the model, the first step was to split the data into training and testing sets. These were split A computer screen with text on it

Description automatically generatedas training 80% - testing 20%. The Linear regression model was then initialized and trained using the allocated training and testing sets. Once the model had been fitted and tested, it was time to attempt predictions. This was done using the .predict() method.

Figure 9: Scatter Graph showing expected vs actual calories burned

Figure 8: The code to train the Linear Regression model

The final step was to evaluate the model, which was, as mentioned before, done using Mean Absolute Error, the Mean Square Error, and finally the most useful and accurate one, the Root Mean Square Error. The RMSE result gave us a value of 70, meaning that for each entry predicted, the prediction was 70 calories off of the actual value. This may sound like a lot but considering that the dataset ranged from 300 calories burned to 1600 calories burned, this is a relatively small margin of error. It can be deduced that the model is accurate at making predictions, and this is further exemplified by Figure 9. Some conclusions to be taken from this prediction graph is that the model excels between 600 and 100 calories burned, where the dots are closest bunched up with the regression line. Within this range, it can be expected that the model will make a very accurate prediction. Past 1000 calories the model starts to become less accurate with predictions, as the dots become quite sparse and spread apart. This could be an indicator that there is a lack of training data at these ranges, but more importantly, the higher the amount of calories burned (which correlates with exercise length), the harder it is to predict how many calories have been burned.

# Conclusion

From all of this data manipulation, prediction and visualization, it becomes apparent that the amount of calories burned and quality of exercise depends on a large number of factors, some of which are easily controllable, but others that cannot be controlled so easily. The main points to take away from this investigation are some that would be expected. The data seems to reinforce widely-accepted knowledge, with some particular surprises. Mainly, when it comes to burning calories, as is obviously known, the longer a duration of a workout, the more calories that will be burned. But the surprising fact is that with longer workout durations, there seems to be a larger amount of calories burned per minute. This is interpreted to be due to these individuals working out for longer durations having more experience (as there was a determination that workout length and experience level were correlated), and thus knowing how to burn that fat more efficiently. In terms of demographics, there was not a particular difference between both males and females, where both groups benefitted from the same factors, only with males able to burn slightly more fat doing the same workout for the same time, but it was only a slight difference. As for the type of workout used, the data shows that there isn’t a significant difference between workout types; instead, the important thing is to actually work out. The way to optimise calorie burning isn’t through some specific regime or method, rather it is by experience and consistency. Those individuals who have a lot of experience know how to burn their calories more efficiently with their time, and its because they have been consistent with going to the gym and doing the routines.

# References

Dataset – Kaggle (<https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset>)

Tools:

* NumPy (https://numpy.org/)
* SKLearn (https://scikit-learn.org/stable/)
* Pandas (https://pandas.pydata.org/)
* Seaborn (https://seaborn.pydata.org/)
* MatPlotLib (https://matplotlib.org/)