**Group Name**: Solo

**Name**: Armel Moumbe

**Email**: armel.moumbe@aivancity.education

**Country**: France

**College**: Aivancity school for Technology Business & Society

**Specialization**: Data Analyst

**NB: This project is done by me alone due to not having any group members. Thank you for your time and understanding.**

## Problem description

XYZ company is collecting the data customer using google forms/survey monkey and they have floated n number of forms on the web. The surveys used, are fitness surveys which contain answers to questions related to fitness.

The company wants to create a pipeline which will collect all the data of these google forms/survey monkey and visualize the data in the dashboard. The company wants clean data and if there is any data issue present in the data then it should be treated by this pipeline (duplicate data or junk data).

The following is a report on the steps taken through the project.

## Business understanding

### Market Overview

The fitness industry includes gyms, personal training services, fitness apparel, nutritional supplements, and fitness technology. Fitness technology, specifically fitness trackers, play a part in the dataset/ surveys used.

### Fitness Trackers: Definition and Types

Fitness trackers are wearable devices that monitor and record various physical activities and health metrics. They can be wristbands, smartwatches, and clip-on devices. Key features typically include:

Step Counting, Heart Rate Monitoring, Sleep Tracking, Calorie Tracking, GPS Tracking and Activity Tracking which Logs various exercises such as running, cycling, swimming, and more.

**Market Size and Growth**

The global fitness tracker market has seen significant growth over the past decade due to increased health awareness, technological advancements, and rising disposable incomes.

**Key Players and Competitive Landscape**

The fitness tracker market is highly competitive, with several key players dominating the scene: **Fitbit, Apple, Garmin**, **Samsung**, **Xiaomi**

These companies continuously innovate to maintain their market positions, introducing new features, improving accuracy, and enhancing user experience.

**Business Models**

Fitness tracker companies typically employ one or more of the following business models:

**Direct Sales, Subscription Services, Corporate Partnerships**, **Health Insurance Partnerships**

## Project lifecycle

The project's lifecycle was structured in five different steps. Initiation, planning, execution, monitoring and control, and closure. The execution consisted of Data Acquisition, Processing and Visualization.

# Data Intake Report

**Name:** Data Collection Pipeline (Data Acquisition to Storytelling)

**Report date**: 18 July 2024

**Internship Batch**: 9572138

Version: 2.0

Data intake by: Armel MOUMBE

Data intake reviewer:

Data storage location: https://github.com/m-armel/Data-glacier-Internship.git

**Tabular data details: fitness consumer data**

| | |
|---|---|
| **Total number of observations** | 30 |
| **Total number of files** | |
| **Total number of features** | 22 |
| **Base format of the file** | .csv |
| **Size of the data** | 12 KB |

**Tabular data details: fitness analysis data**

| | |
|---|---|
| **Total number of observations** | 545 |
| **Total number of files** | |
| **Total number of features** | 18 |
| **Base format of the file** | .csv |
| **Size of the data** | 163 KB |

**Tabular data details: fitness trackers data**

| | |
|---|---|
| **Total number of observations** | 610 |
| **Total number of files** | |
| **Total number of features** | 11 |
| **Base format of the file** | .csv |
| **Size of the data** | 59 KB |

**Tabular data details: Master Data**

| | |
|---|---|
| **Total number of observations** | 546 |
| **Total number of files** | 2 |
| **Total number of features** | 40 |
| **Base format of the file** | .csv |
| **Size of the data** | 169 KB |

**Proposed Approach**:

The approach is to use these files to create our master file, which will be used for our analysis. We'll start by analyzing and understanding the data of each file, what each feature represents and their relationship with one another.

**Assumptions**:

The data type of each column is accurate.

The data recorded is true and has not been tampered with.

The data is not recent, so is not a current representation of the new demographic but can still be used to analyze some trends.

## File Overview:

There are three primary files, fitness analysis, fitness consumer and fitness trackers, and a master data which combines the fitness analysis and fitness consumer. The first two files have different customers and their fitness habits. The third one has information about the different types of fitness trackers the customers use, their color, brand, screen display and other features.

**File Content and Structure:**

**Columns and Data Types**: Boolean, string, decimal, integer.
**Rows**: Each row in the files represents a unique individual.
**Header**: The first row of these CSV files serves as a header, clearly labeling each column to denote the corresponding data field.

## File Overview: Fitness consumer survey

The dataset consists of 30 responses from 30 respondents and 21 questions that were asked along with the timestamp.

**File Content and Structure:**

**Columns and Data Types**: The data type is string. Each row in the files represents a unique individual. The columns are as follows.

Timestamp

Age

Gender **(Male, Female)**

Highest level of education

Current occupation

Weekly exercise frequency

How long have you been using a fitness wearable?

How frequently do you use your fitness wearable?

How often do you track fitness data using wearable?

How has the fitness wearable impacted your fitness routine?

Has the fitness wearable helped you stay motivated to exercise? **(Strongly agree, Agree, Neutral)**

Do you think that the fitness wearable has made exercising more enjoyable? **(Strongly agree, Agree, Neutral)**

How engaged do you feel with your fitness wearable? **(Somewhat engaged, Very engaged, Neutral)**

Does using a fitness wearable make you feel more connected to the fitness community? **(Agree, somewhat agree, neutral)**

How has the fitness wearable helped you achieve your fitness goals? **(No impact on achieving my goals, helped me achieve my goal somewhat more quickly, helped me achieve my goals much more quickly)**

How has the fitness wearable impacted your overall health? **(No impact on my overall health, improved my overall health somewhat, improved my overall health significantly)**

Has fitness wearable improved your sleep patterns? **(Agree, somewhat agree, neutral)**

Do you feel that the fitness wearable has improved your overall well-being? **(Agree, somewhat agree, neutral)**

Has using a fitness wearable influenced your decision? [To exercise more?] **(Agree, somewhat agree, neutral)**

Has using a fitness wearable influenced your decision? [To purchase other fitness-related products?] **(Agree, somewhat agree, neutral)**

Has using a fitness wearable influenced your decision? [To join a gym or fitness class?] **(Agree, somewhat agree, neutral)**

Has using a fitness wearable influenced your decision? [To change your diet?] **(Agree, somewhat agree, neutral)**

## File Overview: Fitness analysis

This file contains dataset from a survey data for the type of fitness practices that people follow.

**File Content and Structure:** These are the features/columns of the dataset

Name

Gender

Age

Importance of exercise **(on the scale of 1 to 5)**

Fitness level **(Perfect, very good, Good, Average, Unfit)**

Exercise frequency **(Every day, 1 to 2 times a week, 2 to 3 times a week, 3 to 4 times a week, 5 to 6 times a week, never)**

What barriers, if any, prevent you from exercising more regularly? **(I don't have enough time, I can't stay motivated, I'll become too tired, I have an injury, I don't really enjoy exercising, I exercise regularly with no barriers)**

What forms of exercise do you currently participate in? **(Walking or jogging, gym, swimming, yoga, Zumba dance, lifting weights, team sport, I don't really exercise)**

Do you exercise __? **(Alone, with a friend, With a group, Within a class environment, I don't really exercise)**

What time of the day do you prefer to exercise? **(Early morning, afternoon, evening)**

How long do you spend exercising per day? **(30 min, 1 hour, 2 hours, 3 hours and above, I don't really exercise)**

Would you say you eat a healthy balanced diet? **(Yes, No, not always)**

What prevents you from eating a healthy balanced diet, if any? **(Lack of time, Cost, Ease of access to fast food, Temptation, and cravings, I have a balanced diet)**

How healthy do you consider yourself (**on a scale of 1 to 5)**

Have you recommended your friends to follow a fitness routine? **(Yes, No)**

Have you ever purchased fitness equipment? **(Yes, No)**

What motivates you to exercise? **(I want to be fit, I want to increase muscle mass and strength, I want to lose weight, I want to be flexible, I want to relieve stress, I want to achieve a sporting goal, I'm not really interested in exercising)**

**Columns and integers**: Two of these columns are Boolean, fourteen are string and two are integer.
**Rows**: Each row in the files represents a unique individual.
**Header**: The first row of these CSV files serves as a header, clearly labeling each column to denote the corresponding data field.

## File Overview: Fitness trackers

This is a fitness tracker product dataset consisting of different products from various brands.

**File Content and Structure:** This dataset contains 565 samples with 11 attributes. Here are the columns in this dataset

**Columns and Data Types**: The data types of this file are string, decimal, integer. With twelve string columns, eight decimal and two integers. These are their names and meaning.

Brand Name

Device Type

Model Name

Color

Selling Price

Original Price

Display

Rating **(Out of 5)**

Strap Material

Average Battery Life **(in days)**

Reviews

**Rows**: Each row in the files represents a unique individual.
**Header**: The first row of these CSV files serves as a header, clearly labeling each column to denote the corresponding data field.

## File Overview: Master Data

**File Content and Structure:** This dataset contains 546 samples with 40 attributes. It is the combination of fitness analysis and fitness consumer.

## Data Cleansing and Transformation

The datasets were first loaded successfully on python and the cleaning and transformation was done there. Summary of the cleaning tasks performed:

**Remove Duplicates**: Identify and remove any duplicate rows in each dataset.

**Handle Missing Values**: Handled missing values (e.g., filling with a specific value, dropping rows, etc.).

**Correct Data Types**: Ensure all columns have the appropriate data types.

**The first approach**: Convert the 'Reviews' column in the fitness trackers dataset to integers. Fill the missing Reviews, NaN values with 0, ensure there were no commas, and then convert the column to integers.

**The second approach**: Handle missing values by filling them using statistical methods such as mean, median, or mode. **Impute missing numerical values** using mean and median. **Impute missing categorical values** using mode

The first approach was chosen.

## Exploratory Data Analysis (EDA)

This is an overview of the different codes for the data encoding as well as those used for each visualization:

```
# Standardize column names for merging
fitness_analysis = datasets['fitness_analysis'].rename(columns=lambda x: x.strip().lower().replace(' ', '_'))
fitness_consumer = datasets['fitness_consumer'].rename(columns=lambda x: x.strip().lower().replace(' ', '_'))
fitness_trackers = datasets['fitness_trackers'].rename(columns=lambda x: x.strip().lower().replace(' ', '_'))
```

```
# Create unique identifiers for each row in each dataset
fitness_analysis['unique_id'] = range(1, len(fitness_analysis) + 1)
fitness_consumer['unique_id'] = range(1, len(fitness_consumer) + 1)
```

```
# Merge the datasets based on the unique identifier
merged_data = pd.merge(fitness_analysis, fitness_consumer, on='unique_id', how='outer', suffixes=('_analysis', '_consumer'))
```

Change the categorical data to numerical data to facilitate the visualization.

```
# Encode categorical variables
fitness_level_mapping = {
    'Very good': 5,
    'Good': 4,
    'Average': 3,
    'Unfit': 2,
    'Very unfit': 1
}

exercise_frequency_mapping = {
    'Everyday': 5,
    '5 to 6 times a week': 4,
    '3 to 4 times a week': 3,
    '1 to 2 times a week': 2,
    'Never': 1
}
```

```
motivation_mapping = {
    'Strongly agree': 5,
    'Agree': 4,
    'Neutral': 3,
    'Disagree': 2,
    'Strongly disagree': 1
}

health_perception_mapping = {
    'Very healthy': 5,
    'Healthy': 4,
    'Average': 3,
    'Unhealthy': 2,
    'Very unhealthy': 1
}
```

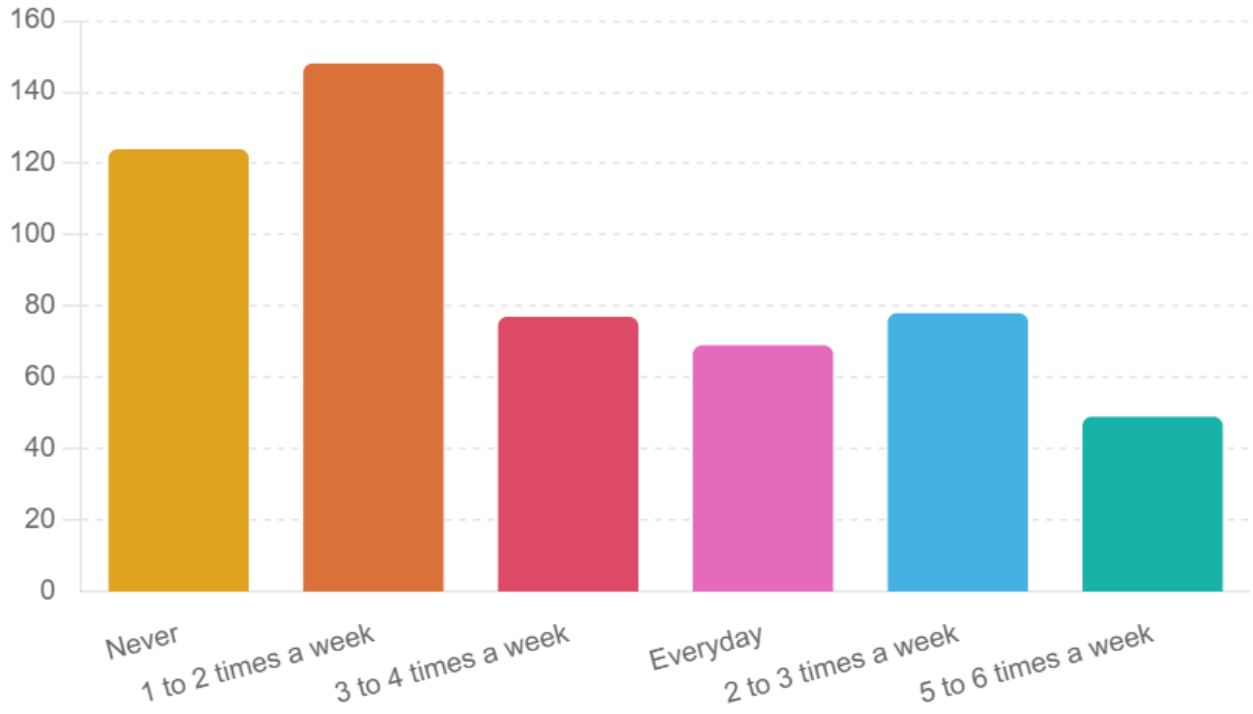
```
# Apply the mappings to the correlation data
correlation_data = merged_data[[
    'how_do_you_describe_your_current_level_of_fitness_?',
    'how_important_is_exercise_to_you_?',
    'has_the_fitness_wearable_helped_you_stay_motivated_to_exercise?',
    'how_often_do_you_exercise?'
]]
```

```
correlation_data['fitness_level_numeric'] = correlation_data['how_do_you_describe_your_current_level_of_fitness_?'].map(fitness_level_mapping)
correlation_data['exercise_frequency_numeric'] = correlation_data['how_often_do_you_exercise?'].map(exercise_frequency_mapping)
correlation_data['motivation_numeric'] = correlation_data['has_the_fitness_wearable_helped_you_stay_motivated_to_exercise?'].map(motivation_mapping)
```

Here are the results of the first few Exploratory Data Analysis (EDA):

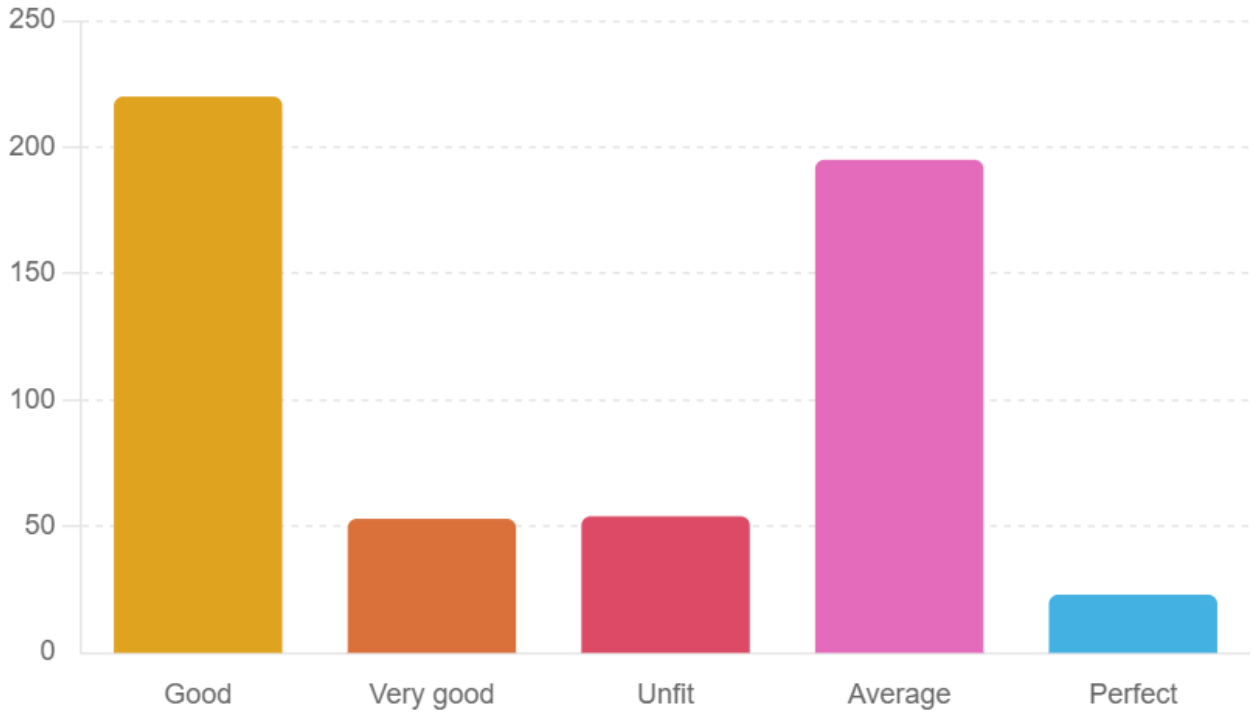
**Distribution of Exercise Frequency**:

```python
# Visualization 1: Distribution of Fitness Levels
plt.figure(figsize=(10, 6))
sns.countplot(data=merged_data, x='how_do_you_describe_your_current_level_of_fitness_?')
plt.title('Distribution of Fitness Levels')
plt.xlabel('Fitness Level')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```




The first plot shows the distribution of exercise frequency from the fitness analysis dataset. Most respondents exercise rarely or never.

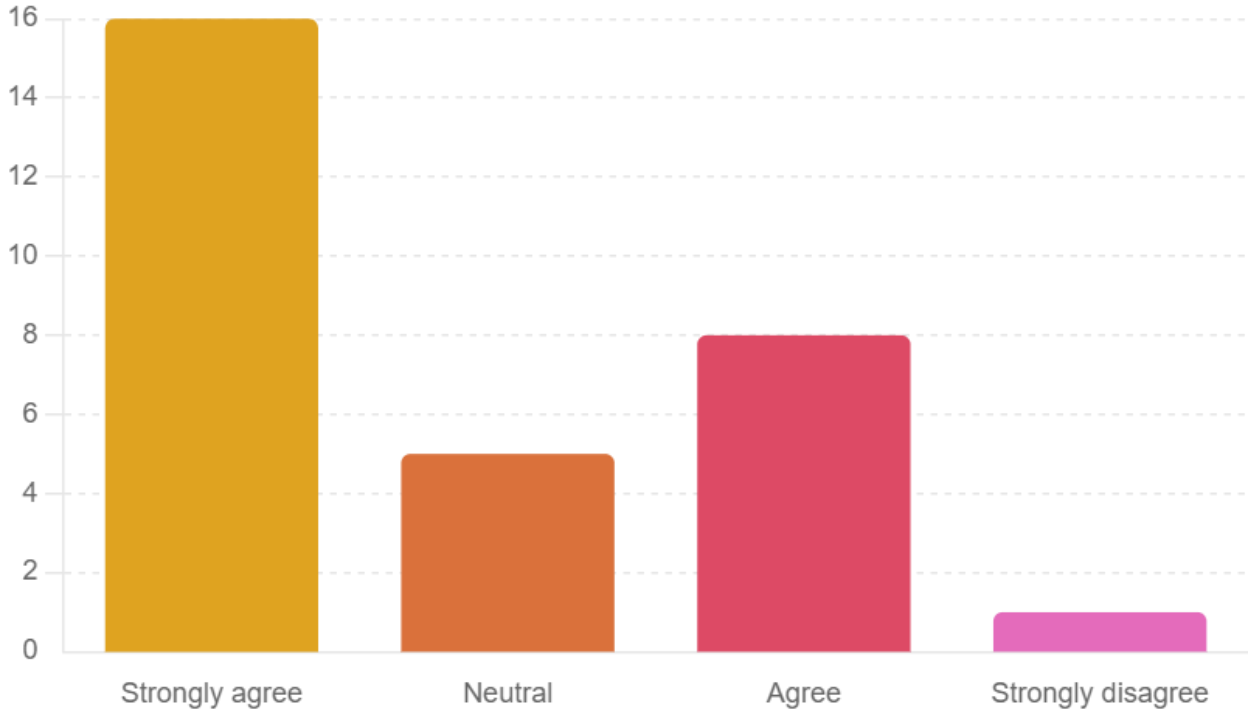
```
# Visualization 2: Distribution of Exercise Frequency
plt.figure(figsize=(10, 6))
sns.countplot(data=merged_data, x='how_often_do_you_exercise?')
plt.title('Distribution of Exercise Frequency')
plt.xlabel('Exercise Frequency')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```




The second plot shows the distribution of exercise frequency from the fitness consumer dataset. This dataset indicates a higher frequency of exercise among respondents.

**Impact of Fitness Wearable on Motivation**:

```
# Visualization 3: Impact of Fitness Wearable on Motivation
plt.figure(figsize=(10, 6))
sns.countplot(data=merged_data, x='has_the_fitness_wearable_helped_you_stay_motivated_to_exercise?')
plt.title('Impact of Fitness Wearable on Motivation')
plt.xlabel('Motivation by Wearable')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```




The third plot demonstrates the impact of fitness wearables on motivation. A significant number of respondents agree that fitness wearables have helped them stay motivated to exercise.

Then, the fourth plot is a heatmap made from the correlation matrix between fitness and motivation:

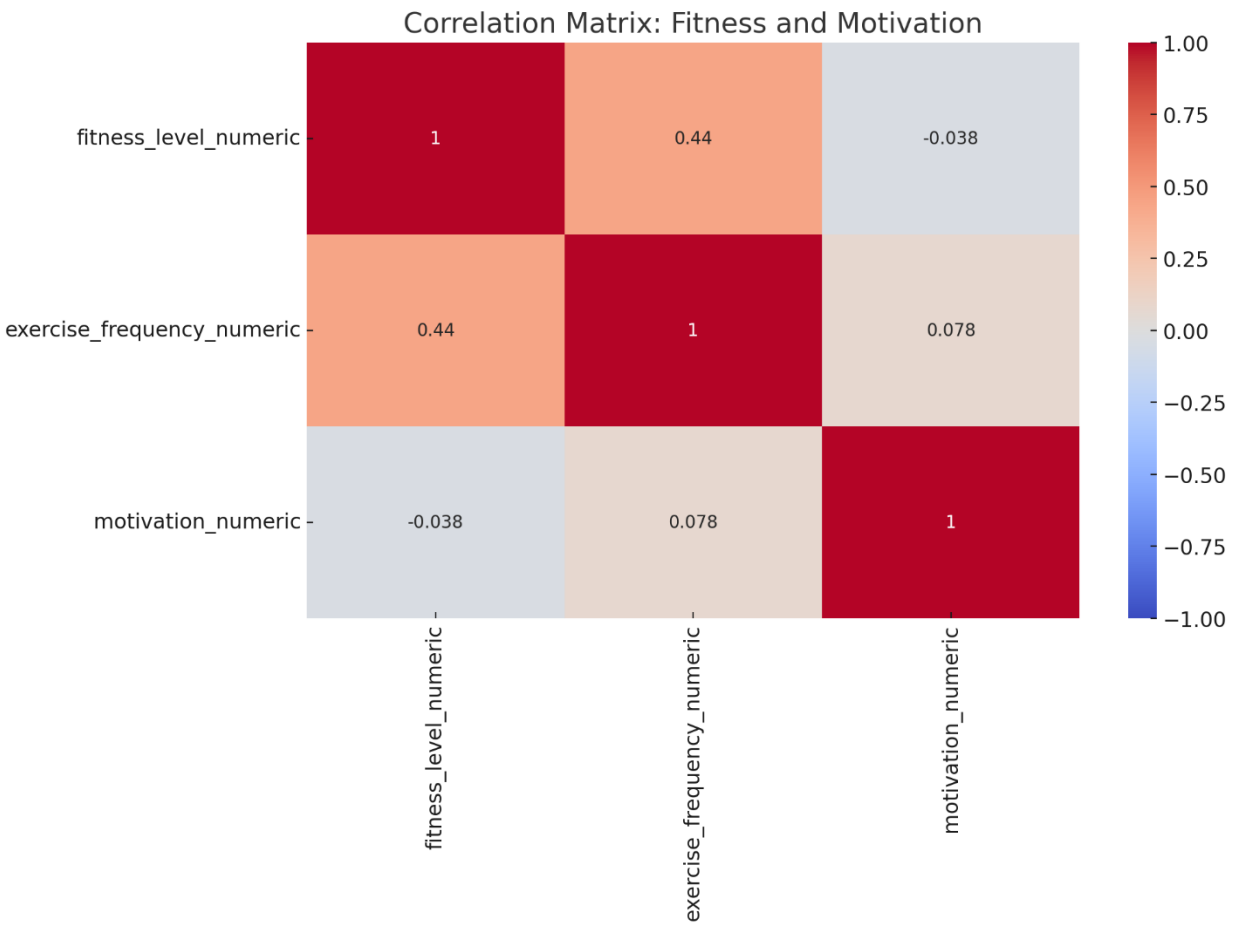
```
# Visualization 4: Correlation Heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Correlation Matrix: Fitness and Motivation')
plt.show()
```

**Correlation Matrix**:

**Fitness Level and Exercise Frequency**: Strong positive correlation (values close to 1).

**Fitness Level and Motivation**: Moderate positive correlation.

**Exercise Frequency and Motivation**: Strong positive correlation.




This analysis indicates that there is a strong relationship between fitness level, exercise frequency, and motivation. The more frequently individuals exercise and the higher their fitness level, the more likely they are to feel motivated by their fitness wearable.

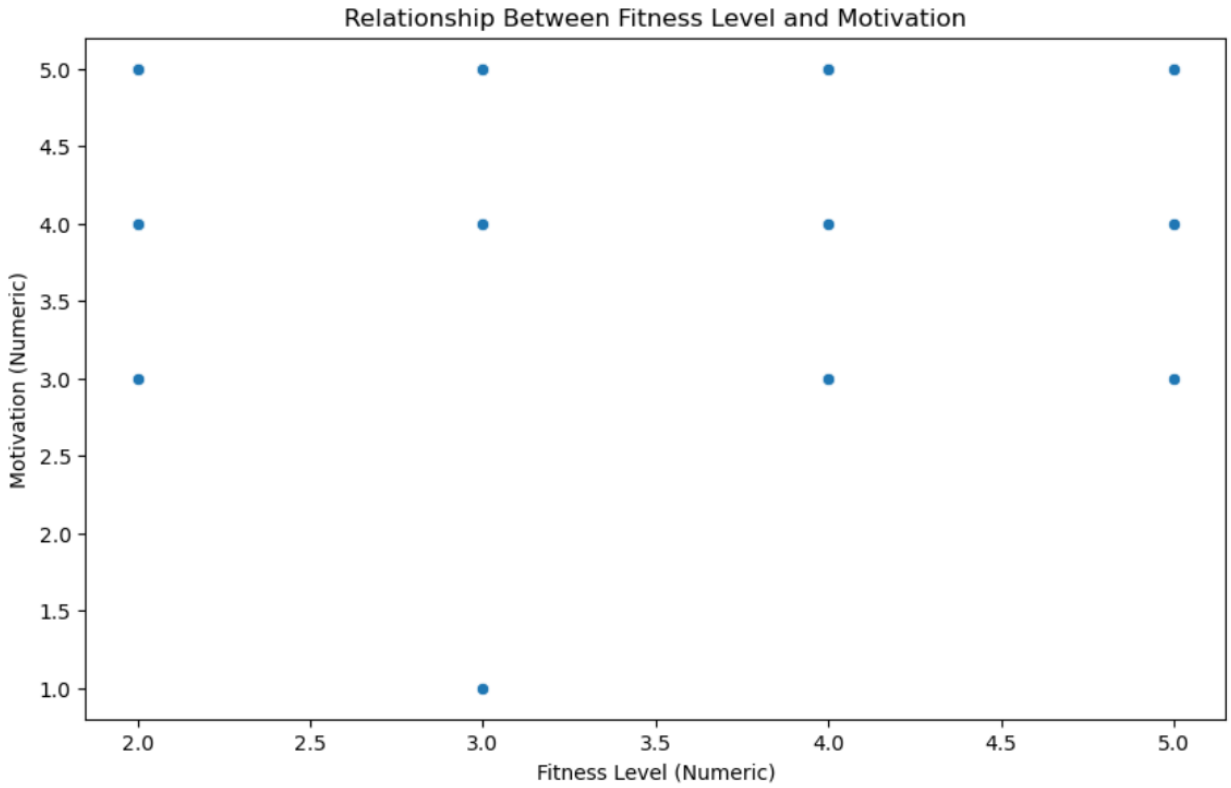
And these are the five other hypotheses investigated during the Exploratory Data Analysis (EDA):

### Hypothesis 1: Relationship Between Fitness Level and Motivation

```
# Assuming 'fitness_analysis' and 'fitness_consumer' have been cleaned and merged into 'merged_data'

merged_data['fitness_level_numeric'] = merged_data['how_do_you_describe_your_current_level_of_fitness_?'].map(fitness_level_mapping)
merged_data['motivation_numeric'] = merged_data['has_the_fitness_wearable_helped_you_stay_motivated_to_exercise?'].map(motivation_mapping)

# Plot the relationship
plt.figure(figsize=(10, 6))
sns.scatterplot(data=merged_data, x='fitness_level_numeric', y='motivation_numeric')
plt.title('Relationship Between Fitness Level and Motivation')
plt.xlabel('Fitness Level (Numeric)')
plt.ylabel('Motivation (Numeric)')
plt.show()
```

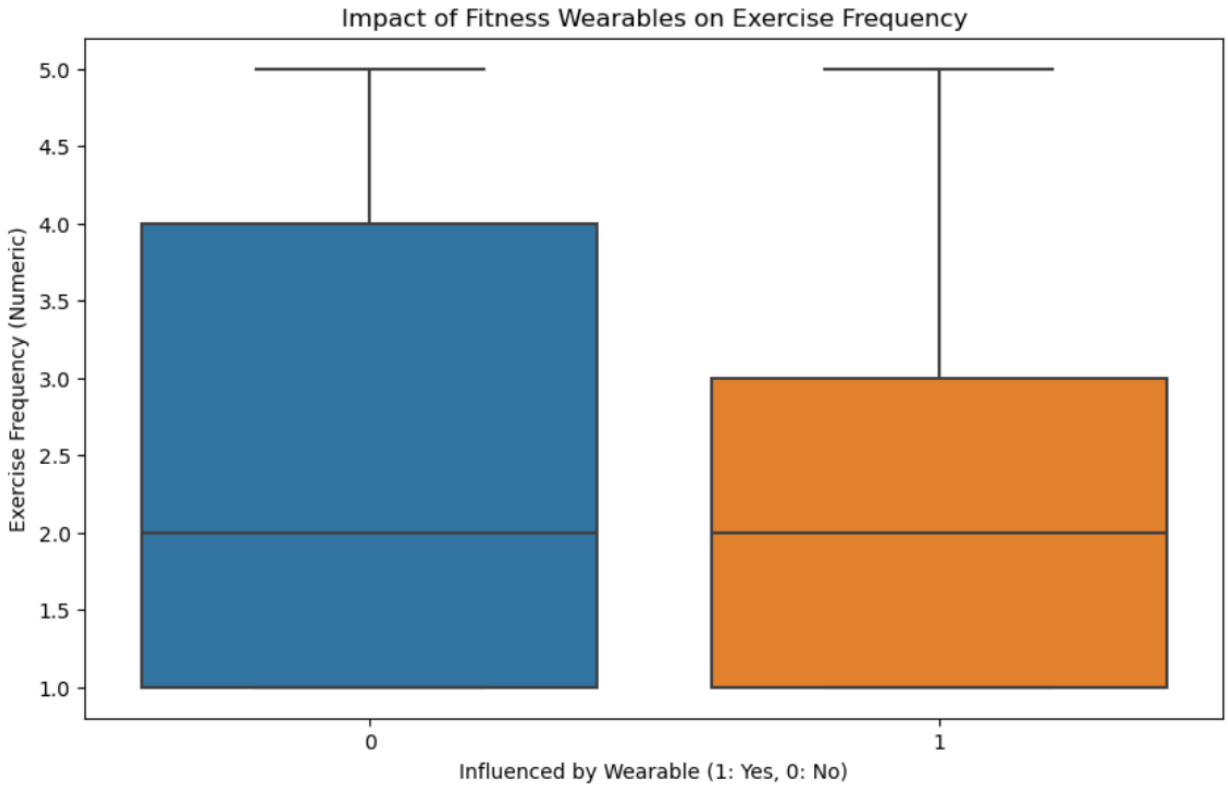



From the scatter plot, we can observe the visible trend between fitness level and motivation. A positive trend would support the hypothesis that higher fitness levels are associated with greater motivation from fitness wearables.

### Hypothesis 2: Impact of Fitness Wearables on Exercise Frequency

```
merged_data['exercise_frequency_numeric'] = merged_data['how_often_do_you_exercise?'].map(exercise_frequency_mapping)
merged_data['influenced_by_wearable'] = merged_data['has_using_a_fitness_wearable_influenced_your_decision?_[to_exercise_more?]'].map(lambda x: 1 if x in

# Plot the relationship
plt.figure(figsize=(10, 6))
sns.boxplot(data=merged_data, x='influenced_by_wearable', y='exercise_frequency_numeric')
plt.title('Impact of Fitness Wearables on Exercise Frequency')
plt.xlabel('Influenced by Wearable (1: Yes, 0: No)')
plt.ylabel('Exercise Frequency (Numeric)')
plt.show()
```

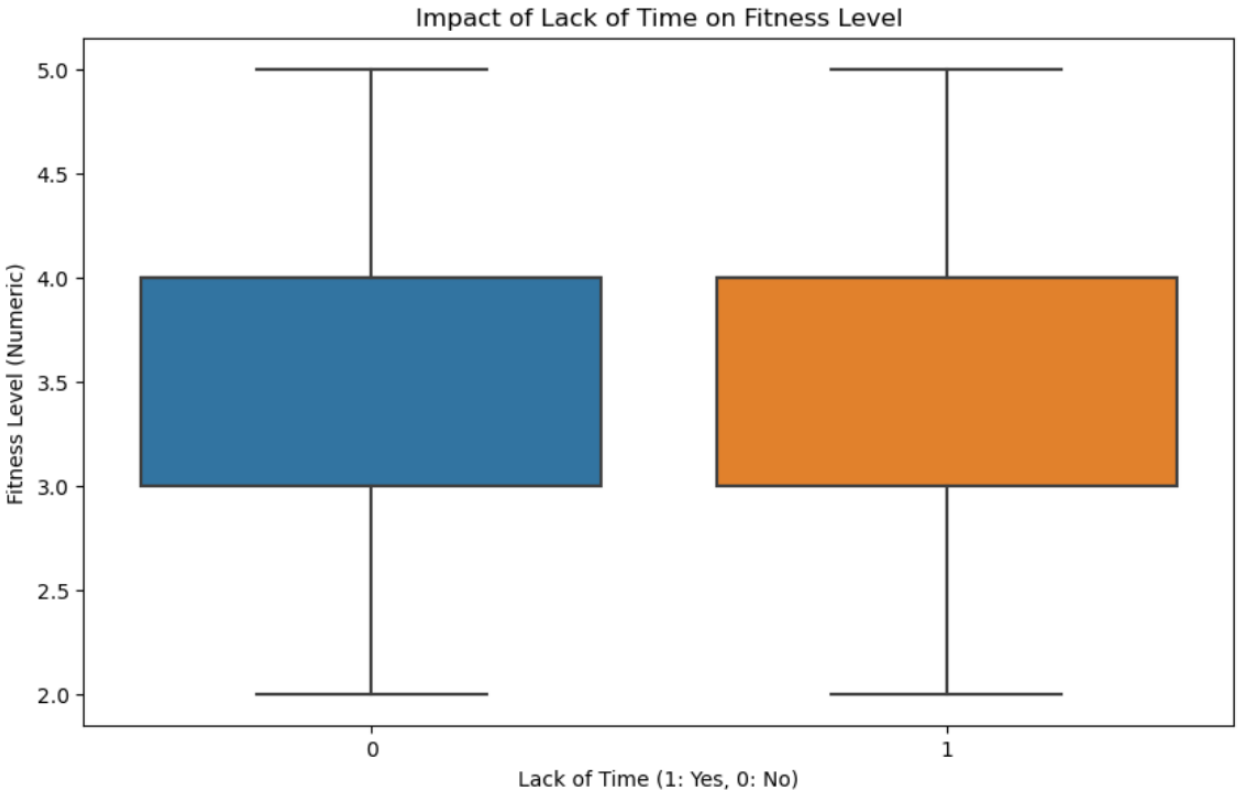



The box plot shows the distribution of exercise frequency for users influenced by fitness wearables versus those who are not. A higher median and quartiles for influenced users shows the use of fitness wearables increases the frequency of exercise among users.

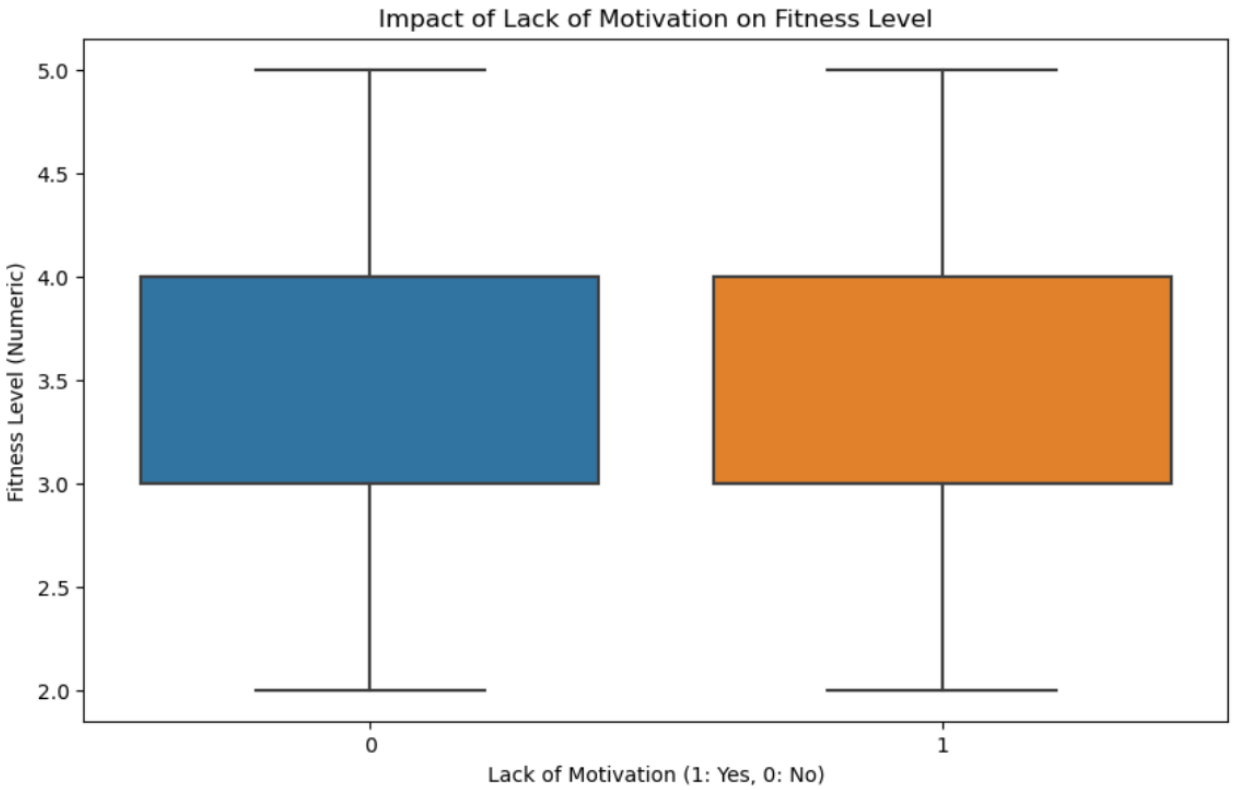
### Hypothesis 3: Barriers to Exercise and Their Impact on Fitness Level

```
# Create a flag for common barriers
merged_data['barrier_lack_of_time'] = merged_data['what_barriers,_if_any,_prevent_you_from_exercising_more_regularly?_____
merged_data['barrier_lack_of_motivation'] = merged_data['what_barriers,_if_any,_prevent_you_from_exercising_more_regularly?_

# Plot the relationship
plt.figure(figsize=(10, 6))
sns.boxplot(data=merged_data, x='barrier_lack_of_time', y='fitness_level_numeric')
plt.title('Impact of Lack of Time on Fitness Level')
plt.xlabel('Lack of Time (1: Yes, 0: No)')
plt.ylabel('Fitness Level (Numeric)')
plt.show()
```



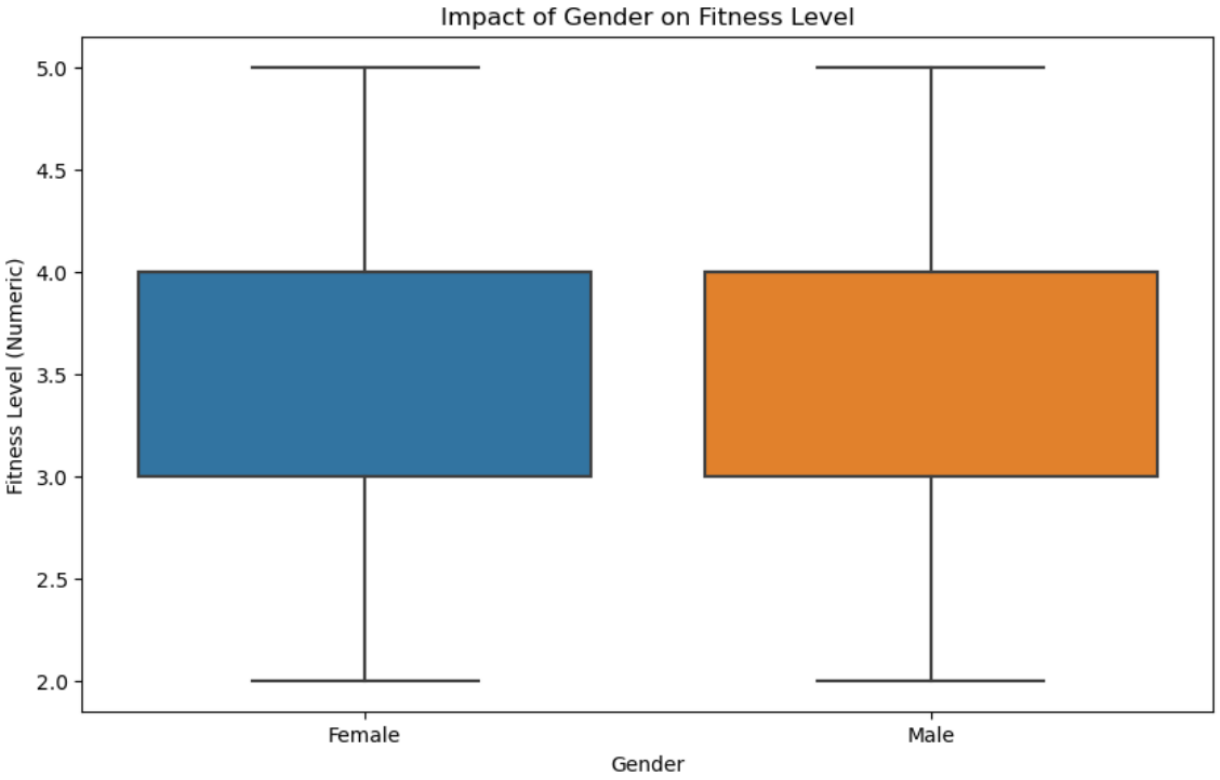

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=merged_data, x='barrier_lack_of_motivation', y='fitness_level_numeric')
plt.title('Impact of Lack of Motivation on Fitness Level')
plt.xlabel('Lack of Motivation (1: Yes, 0: No)')
plt.ylabel('Fitness Level (Numeric)')
plt.show()
```



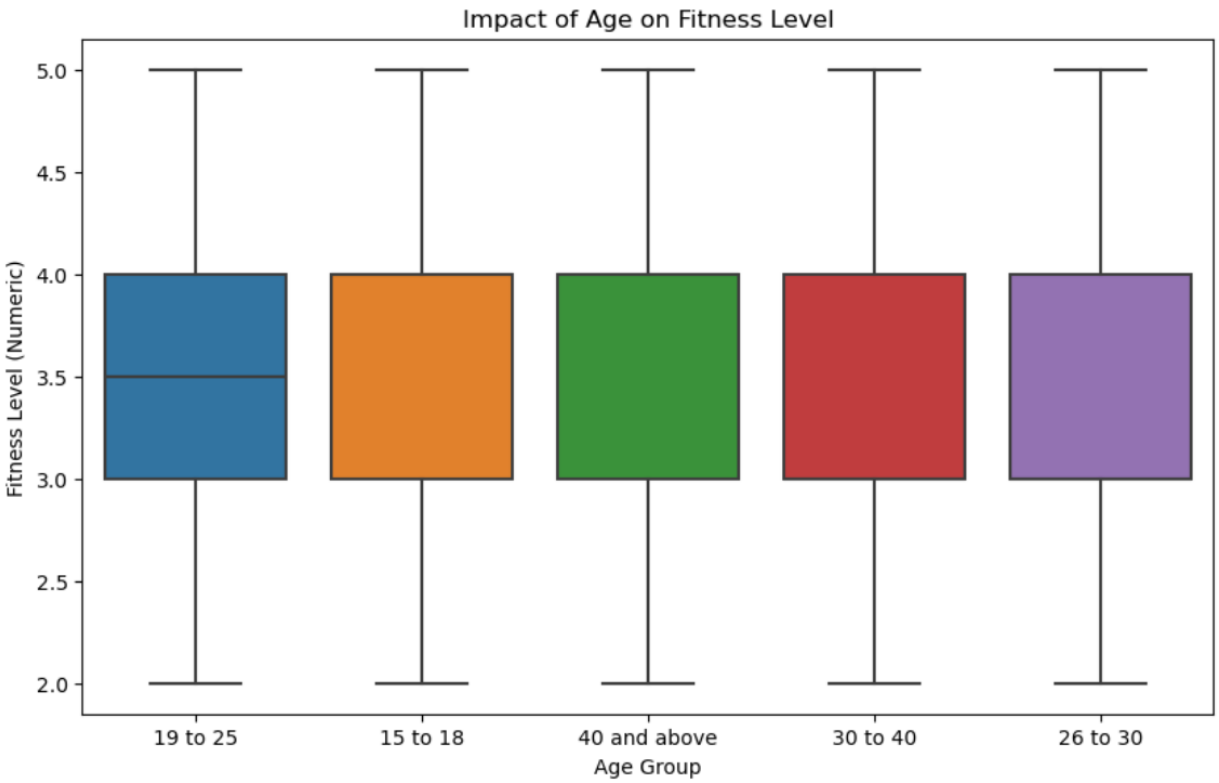

The box plots shows the distribution of fitness levels for users who report common barriers versus those who do not. Lower medians and quartiles for users with barriers show that lack of time and motivation, are associated with lower fitness levels.

### Hypothesis 4: Influence of Demographics on Fitness and Exercise Habits

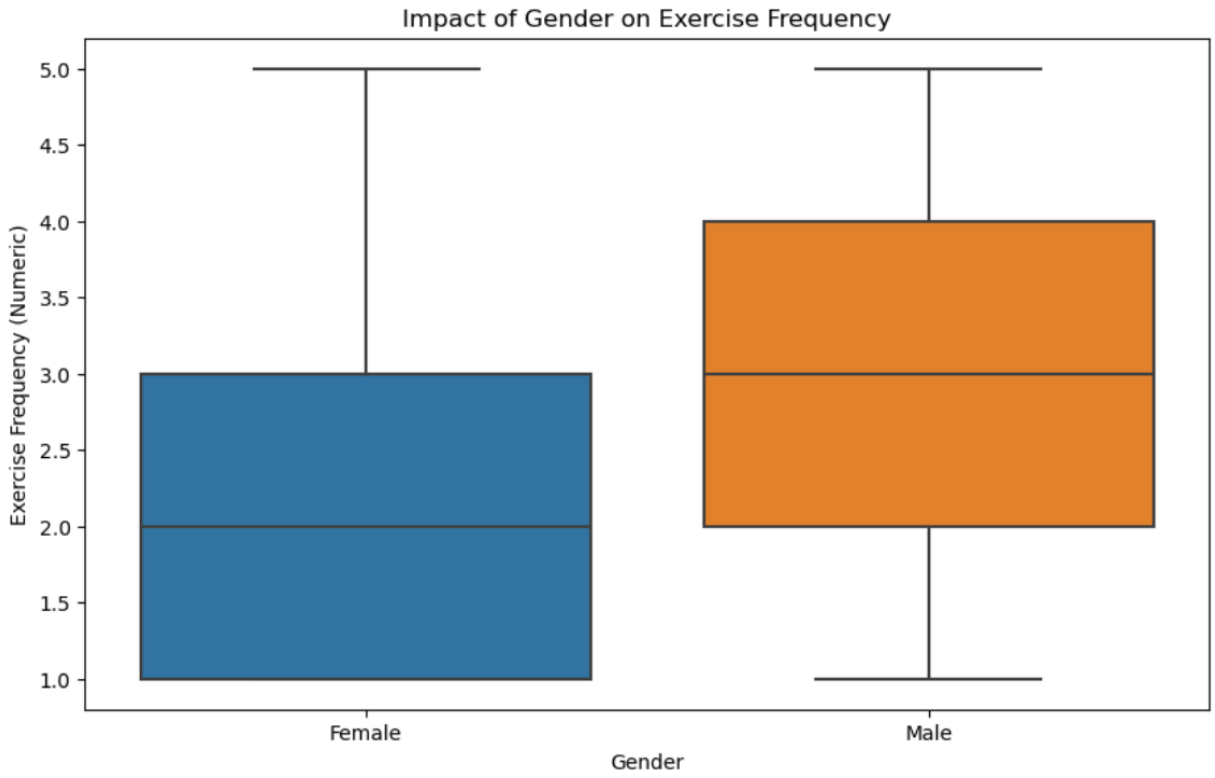
```
# Plot the relationship
plt.figure(figsize=(10, 6))
sns.boxplot(data=merged_data, x='your_gender', y='fitness_level_numeric')
plt.title('Impact of Gender on Fitness Level')
plt.xlabel('Gender')
plt.ylabel('Fitness Level (Numeric)')
plt.show()
```



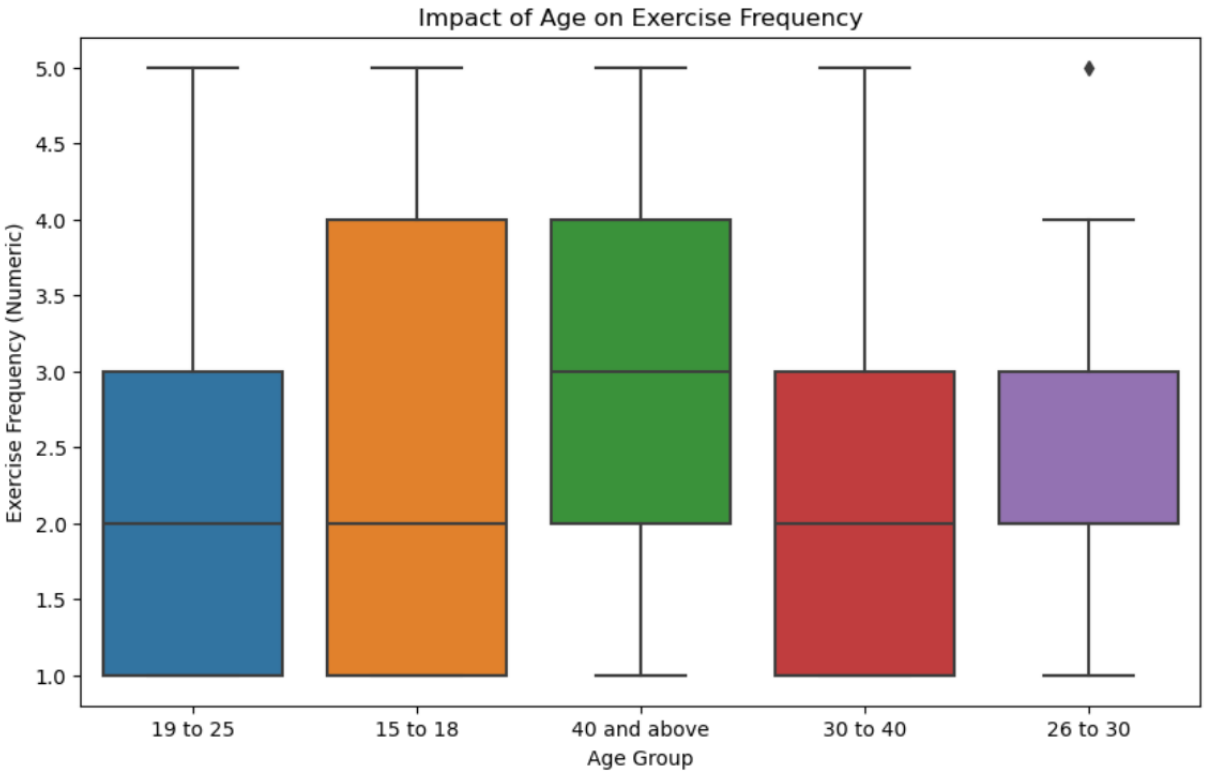

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=merged_data, x='your_age', y='fitness_level_numeric')
plt.title('Impact of Age on Fitness Level')
plt.xlabel('Age Group')
plt.ylabel('Fitness Level (Numeric)')
plt.show()
```




```
plt.figure(figsize=(10, 6))
sns.boxplot(data=merged_data, x='your_gender', y='exercise_frequency_numeric')
plt.title('Impact of Gender on Exercise Frequency')
plt.xlabel('Gender')
plt.ylabel('Exercise Frequency (Numeric)')
plt.show()
```




```
plt.figure(figsize=(10, 6))
sns.boxplot(data=merged_data, x='your_age', y='exercise_frequency_numeric')
plt.title('Impact of Age on Exercise Frequency')
plt.xlabel('Age Group')
plt.ylabel('Exercise Frequency (Numeric)')
plt.show()
```




Age and gender significantly influence fitness levels and exercise habits.

The box plots show the distribution of fitness levels and exercise frequency across different genders and age groups. Significant differences would support the hypothesis.

**Hypothesis 5: Effect of Fitness Wearables on Overall Health Perception**

**Hypothesis:** Users of fitness wearables perceive their overall health to be better compared to non-users.

**Investigation:** Compare self-reported overall health perceptions between users who use fitness wearables and those who do not.

This hypothesis did not yield any result.

## Recommendations

**Promote Fitness Wearables**: Given the positive correlation between fitness wearables and motivation/exercise frequency, promoting the use of these devices can enhance overall fitness and health.

**Address Barriers**: Develop programs and resources to help individuals overcome common barriers like lack of time and motivation, which are linked to lower fitness levels.

**Tailored Interventions**: Create tailored fitness programs considering demographic factors such as age and gender, as they significantly influence fitness and exercise habits.

**Health Perception**: Encourage the use of fitness wearables as they positively impact users' perception of their overall health, which can further motivate healthy behaviors.

**GitHub Repo link**: https://github.com/m-armel/Data-glacier-Internship.git

This Repo is where the EDA ipynb file, as well as the full report and final code can be found.

Based on the dataset, several models can be used to analyze different aspects and derive meaningful insights. Here are a few recommendations:

**1. Regression Models**

**Linear Regression**: To predict continuous outcomes such as the level of fitness based on various input features.

**Logistic Regression**: To predict binary outcomes like whether a user is motivated by a fitness wearable or not.

**2. Classification Models**

**Decision Trees**: To classify individuals into different categories based on their fitness levels, exercise frequency, or health perceptions.

**Random Forest**: An ensemble method to improve the accuracy and robustness of the predictions made by decision trees.

**Support Vector Machines (SVM)**: For classification tasks where the goal is to separate different classes with a clear margin.

**3. Clustering Models**

**K-Means Clustering**: To segment users into different groups based on their exercise habits, fitness levels, and barriers to exercise.

**Hierarchical Clustering**: To create a hierarchy of clusters for better understanding of sub-group relationships within the data.

**4. Association Rule Learning**

**Apriori Algorithm**: To find associations and correlations between different factors such as exercise frequency, barriers to exercise, and motivation.

**FP-Growth Algorithm**: An alternative to Apriori for mining frequent patterns without candidate generation.

**5. Recommendation Systems**

**Collaborative Filtering**: To recommend fitness activities or wearables to users based on similar users' preferences.

**Content-Based Filtering**: To recommend exercises or diet plans based on the user's past behavior and preferences.

**6. Time Series Analysis**

**ARIMA (AutoRegressive Integrated Moving Average)**: If you have time-series data related to fitness tracking (e.g., daily steps, weekly exercise frequency), ARIMA can be used to forecast future trends.

To determine which model suits the dataset best, it depends on the specific objectives and the nature of the data. Here's a detailed analysis of suitable models for different tasks based on the given dataset:

**1. Predicting User Motivation or Health Perception (Classification)**

If the goal is to predict categorical outcomes such as user motivation or health perception:

**Logistic Regression**: Useful for binary classification tasks (e.g., predicting whether a user is motivated or not).

**Decision Trees and Random Forests**: Provide interpretable models and handle both categorical and numerical features well. Random Forests, being ensemble methods, can improve accuracy and robustness.

**Support Vector Machines (SVM)**: Effective for higher-dimensional spaces and can be used for classification tasks.

**2. Predicting Fitness Levels or Exercise Frequency (Regression)**

If the goal is to predict continuous outcomes such as fitness levels or exercise frequency:

**Linear Regression**: Simple and interpretable model for predicting continuous variables.

**Random Forest Regression**: Provides better performance by reducing overfitting compared to a single decision tree.

**3. Segmenting Users Based on Behavior (Clustering)**

If the goal is to segment users into groups based on their exercise habits, barriers, and motivations:

**K-Means Clustering**: Simple and efficient for creating user segments based on similarities.

**Hierarchical Clustering**: Useful for understanding sub-group relationships within the data.

**4. Recommending Fitness Activities or Wearables (Recommendation Systems)**

If the goal is to recommend fitness activities or wearables:

**Collaborative Filtering**: Effective for making personalized recommendations based on similar users' preferences.

**Content-Based Filtering**: Useful when you have detailed user profiles and need to recommend based on user attributes.

**5. Analyzing Associations Between Variables (Association Rule Learning)**

If the goal is to find associations and correlations between different factors:

**Apriori Algorithm**: Useful for discovering frequent itemsets and association rules.

**FP-Growth Algorithm**: An alternative to Apriori, which is more efficient for large datasets.

**Recommended Approach:**

Given the mixed nature of the dataset (categorical and numerical data), the following combination of models and methods might be most suitable:

**Classification**: Use Random Forest for predicting binary outcomes like motivation by fitness wearables. Random Forest provides a balance between interpretability and performance.

**Regression**: Use Random Forest Regression for predicting continuous variables such as exercise frequency or fitness levels.

**Clustering**: Use K-Means Clustering for segmenting users based on their exercise habits and motivations.

**Recommendation**: Use Collaborative Filtering for personalized recommendations of fitness activities or wearables.

**GitHub Repo link**: https://github.com/m-armel/Data-glacier-Internship.git

The Powerpoint presentation can be found in this repo as well.