Data Analysis: Data collection pipeline

EDA Presentation and proposed modeling technique

Group Name: Solo
Name: Armel Moumbe
Email: armel.moumbe@aivancity.education
Country: France
College: Aivancity school for Technology Business & Society
Specialization: Data Analyst
GitHub Repo link: https://github.com/m-armel/Data-glacier-Internship.git

14-08-2024

Agenda

Executive Summary

Problem Statement & Approach

EDA Summary

Modeling Techniques

Recommendations

Executive Summary

XYZ company is collecting the data of customer using google forms/survey monkey and they have floated n number of forms on the web. These forms used are fitness forms. These forms contain consumer information, fitness wearables as well as various information on the frequency and intensity of the fitness activities.

These forms will be processed and used for the company's needs, but before that our objectives are as follows

Objective :

- Create a pipeline for the data collection.
- Make sure the data is usable i.e., Undergo data validations to make sure the data is correct
- Use EDA(Exploratory Data Analysis) to provide insights on how the data can be analyzed and the solutions we can derive from them.

Problem Statement & Approach

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).

Using the fitness datasets collected, various steps were used to accomplish the given goal

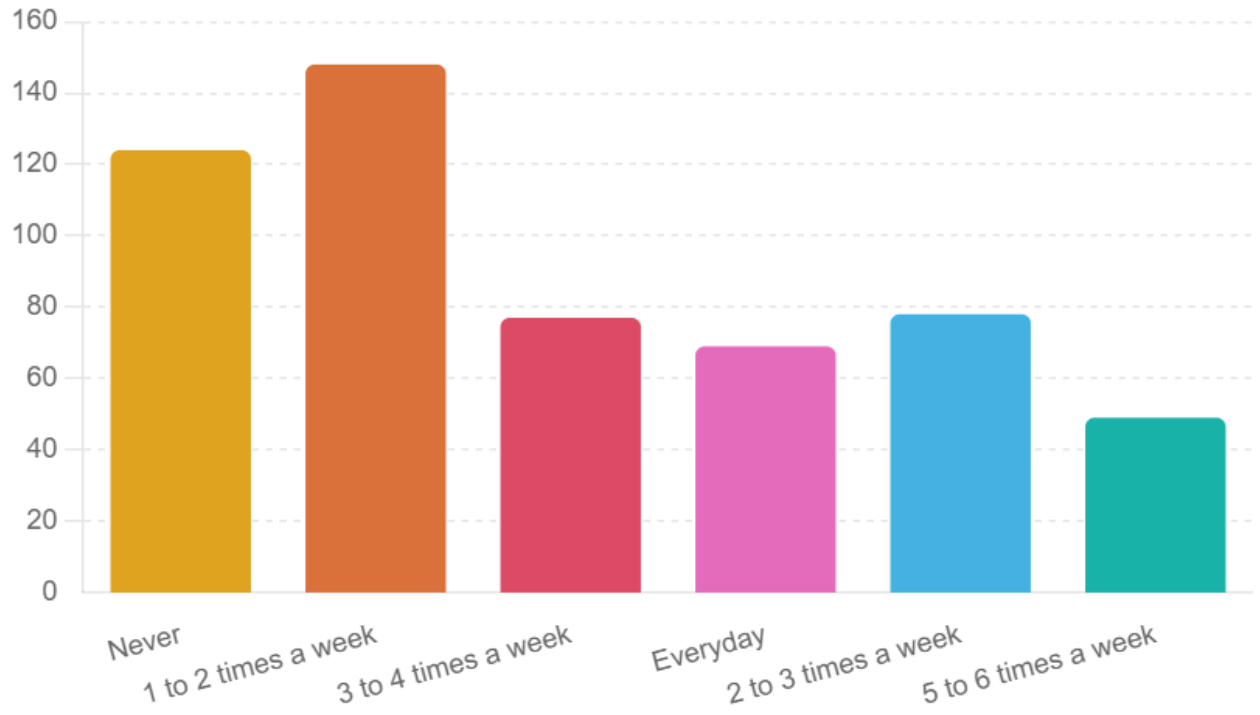
The approach has been divided into four parts:

- Data Understanding, Cleaning, Exploration and Integration
- Correlation and visualization of the datasets
- Finding modeling techniques and creating a dashboard
- Recommendations

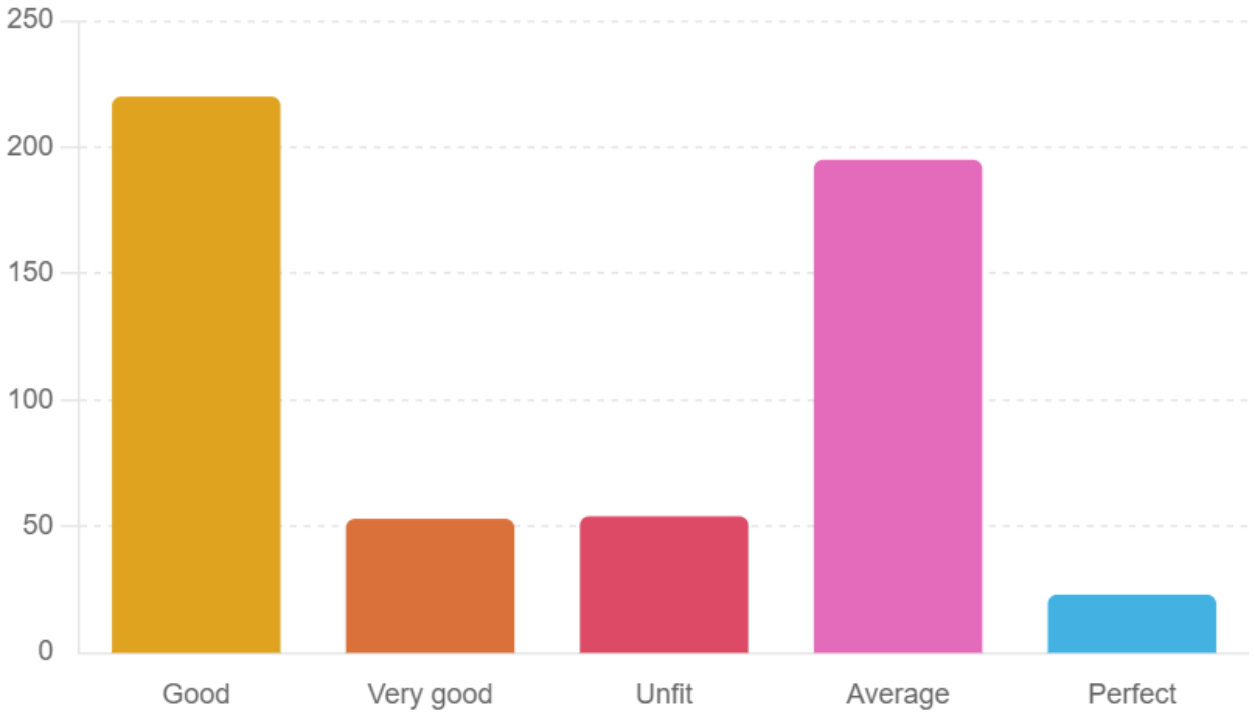
EDA Summary

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

Distribution of Exercise Frequency:
This first plot shows the distribution of exercise frequency from the fitness analysis dataset. Most respondents exercise rarely or never.



EDA Summary

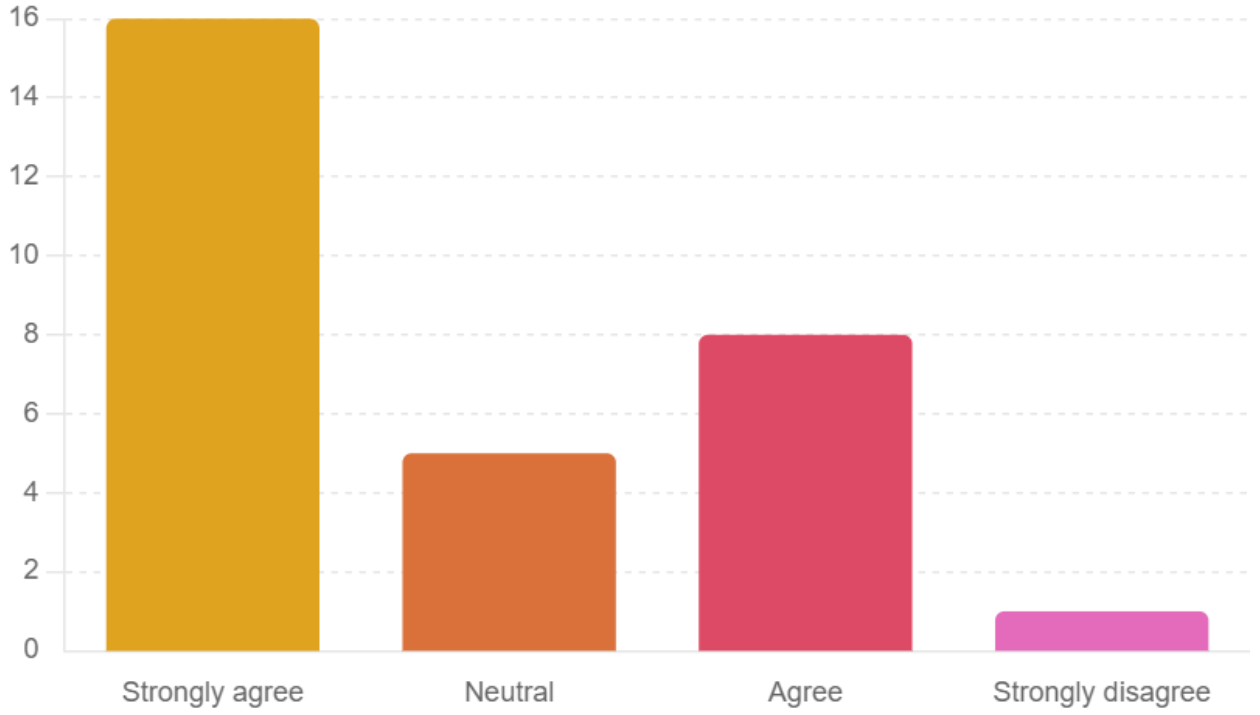


This second plot shows the distribution of exercise frequency from the fitness consumer dataset. This dataset indicates a higher frequency of exercise among respondents.

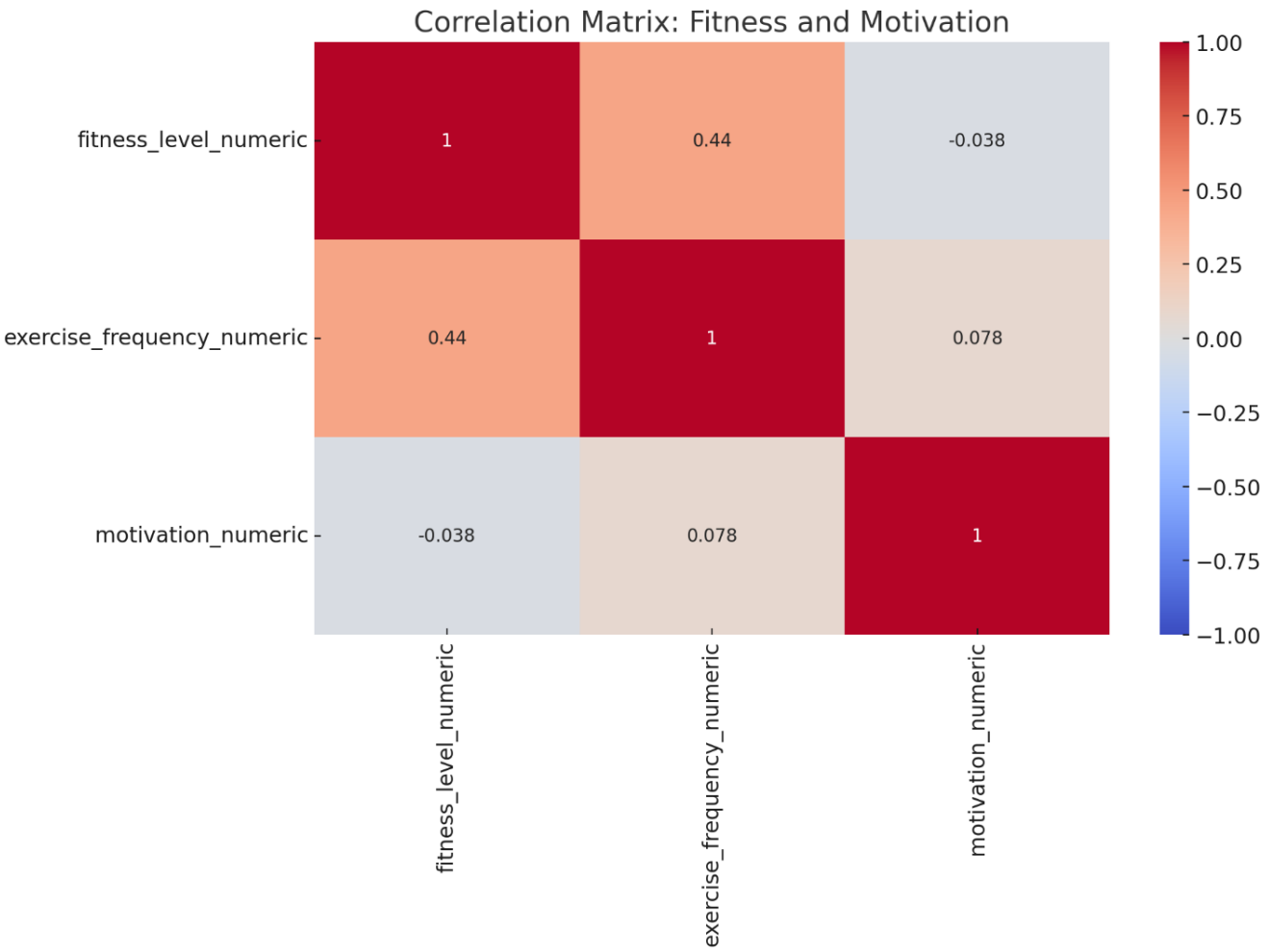
EDA Summary

Impact of Fitness Wearable on Motivation:

This 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.



EDA Summary

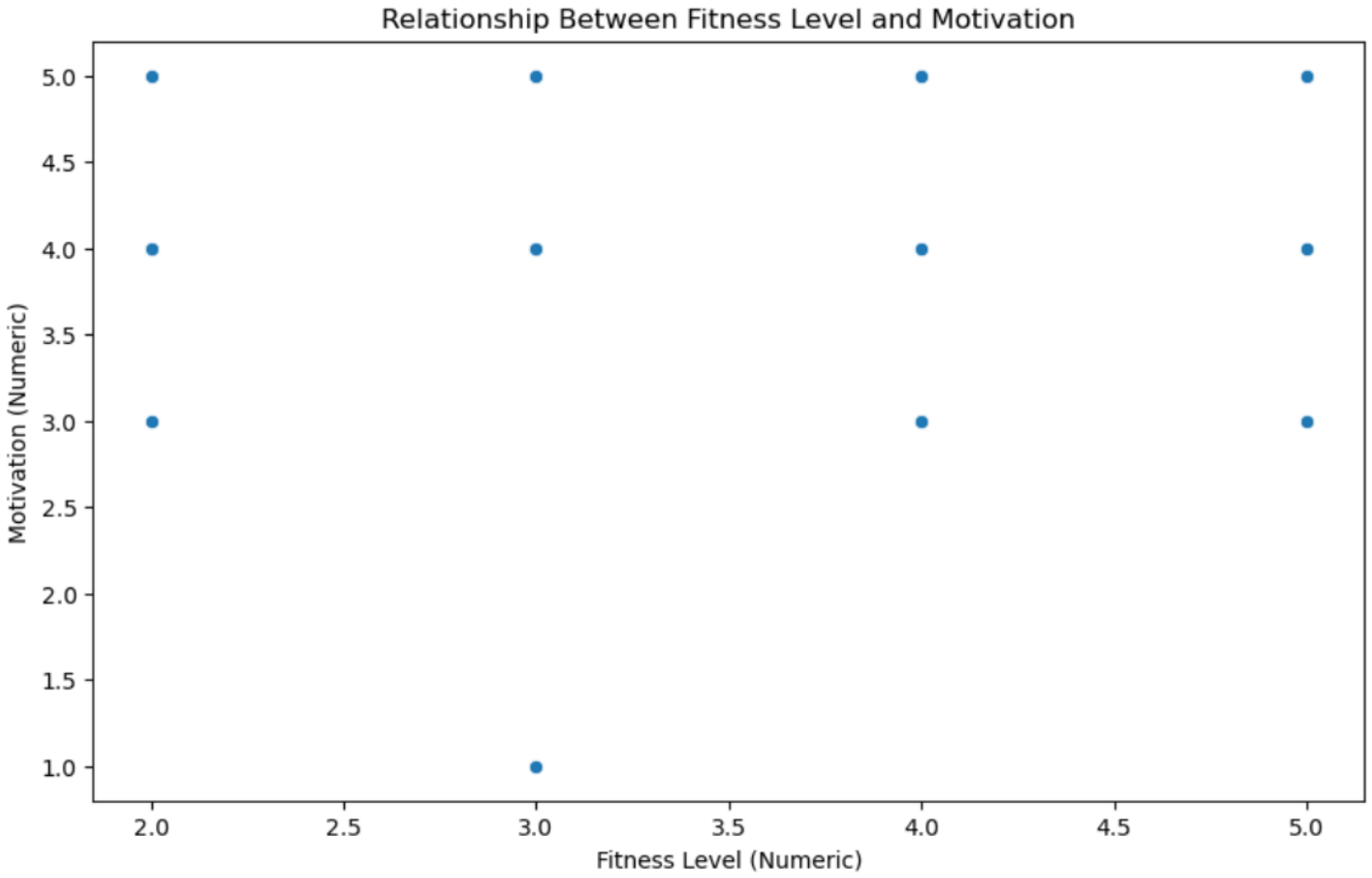


This fourth plot shows the correlation matrix and heatmap between fitness and motivation.

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.

EDA Summary



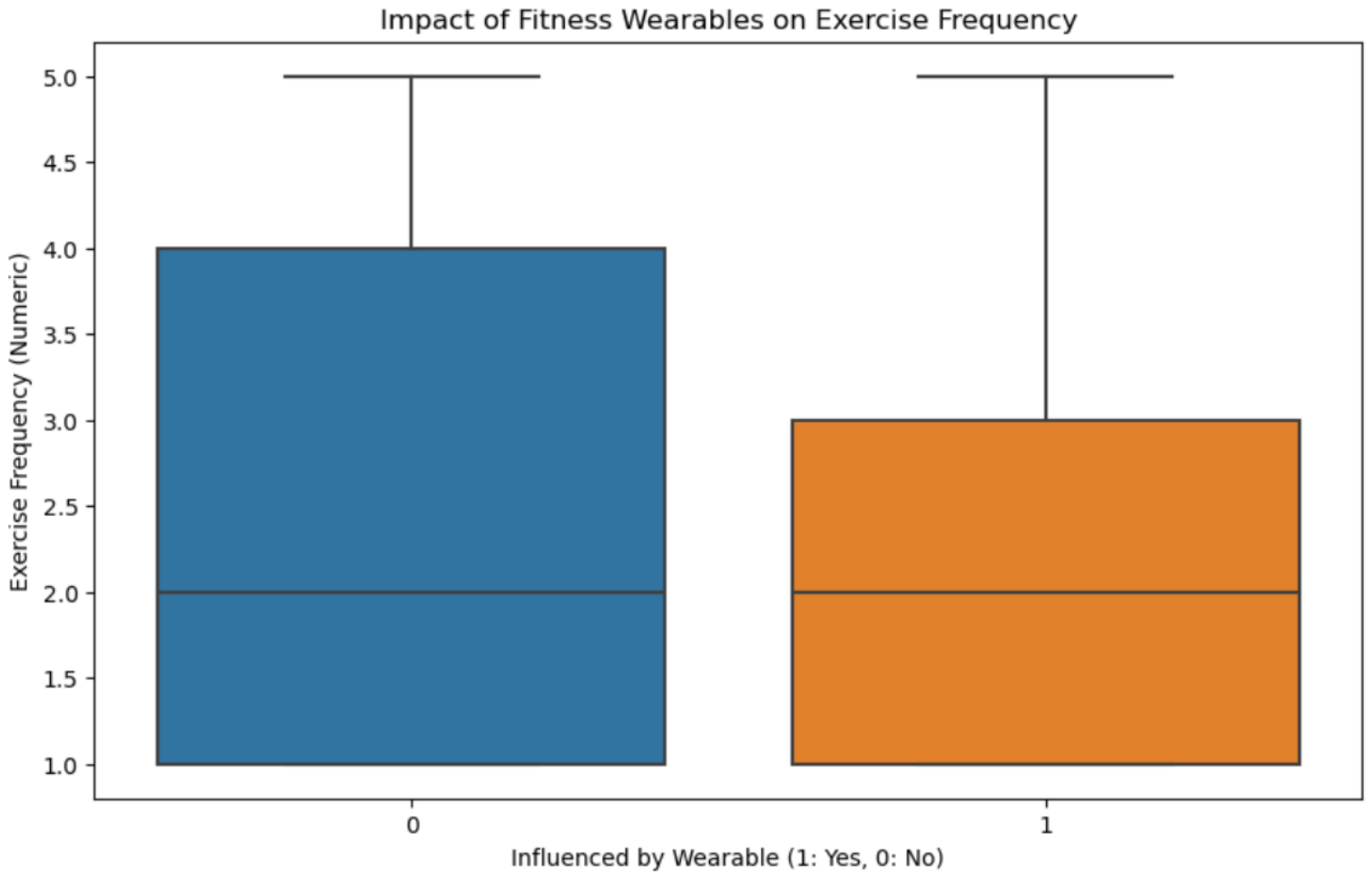
Relationship Between Fitness Level and Motivation

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.

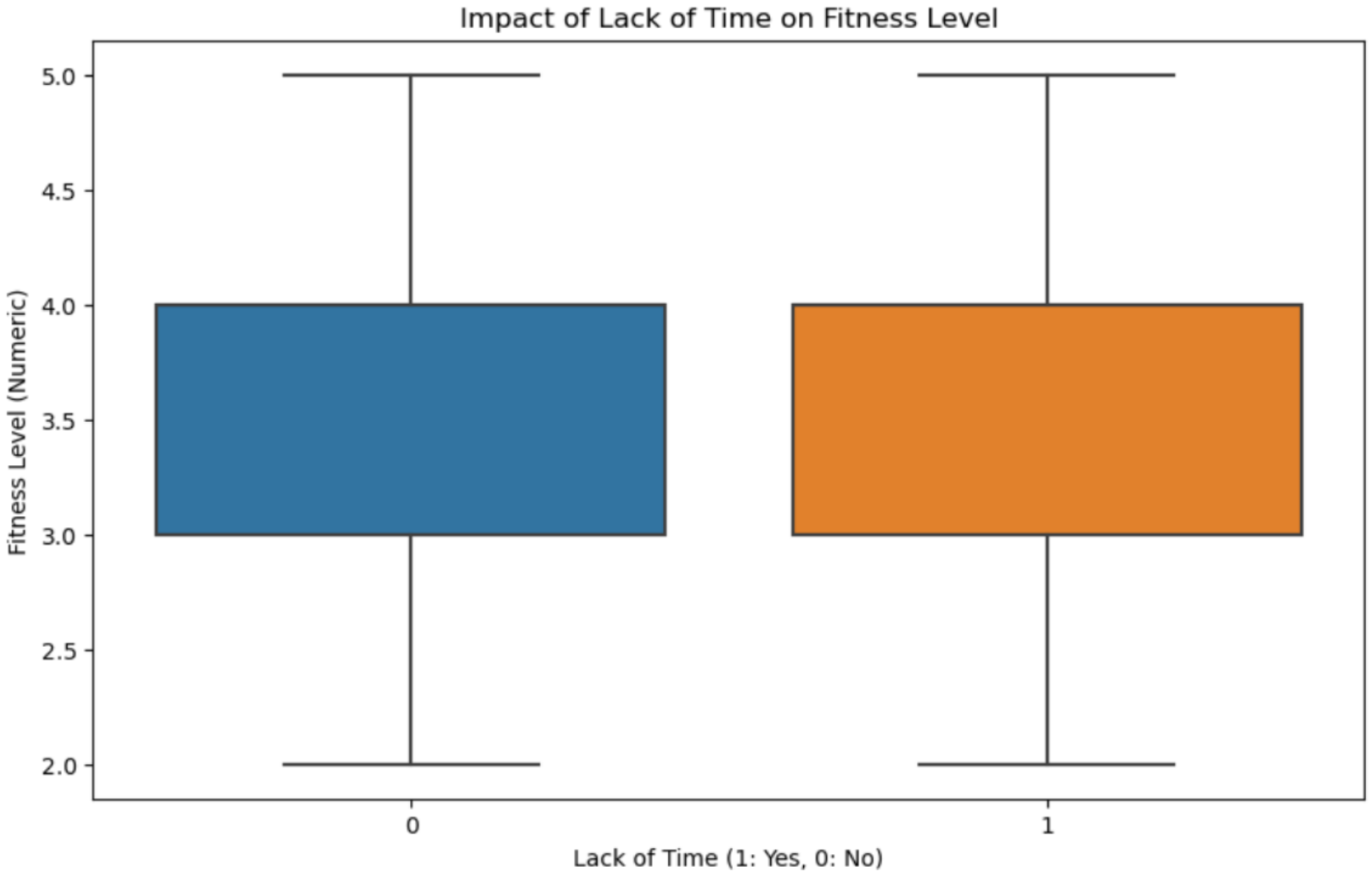
EDA Summary

Impact of Fitness Wearables on Exercise Frequency

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.



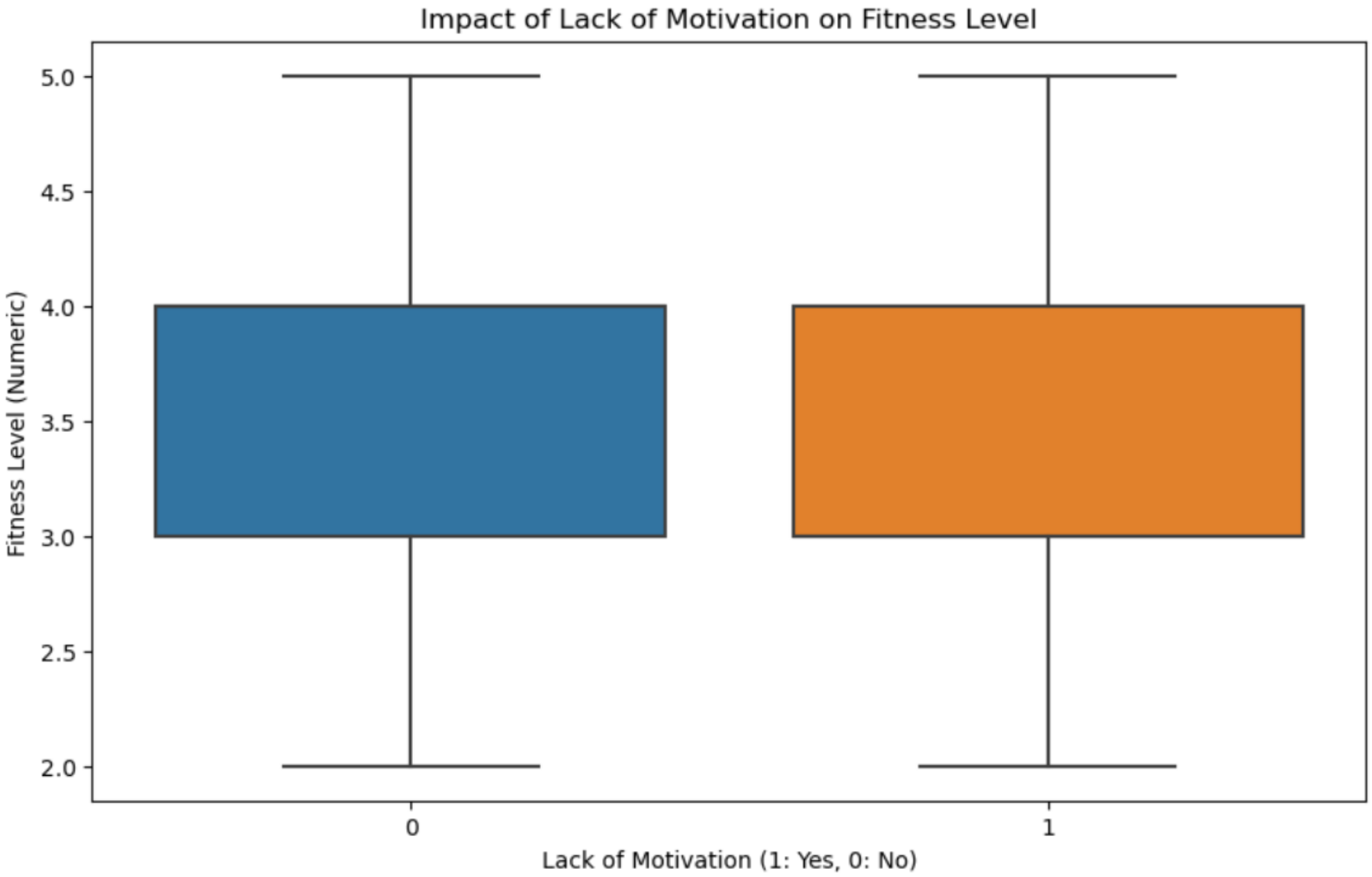
EDA Summary



Barriers to Exercise and Their Impact on Fitness Level

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.

EDA Summary



Barriers to Exercise and Their Impact on Fitness Level

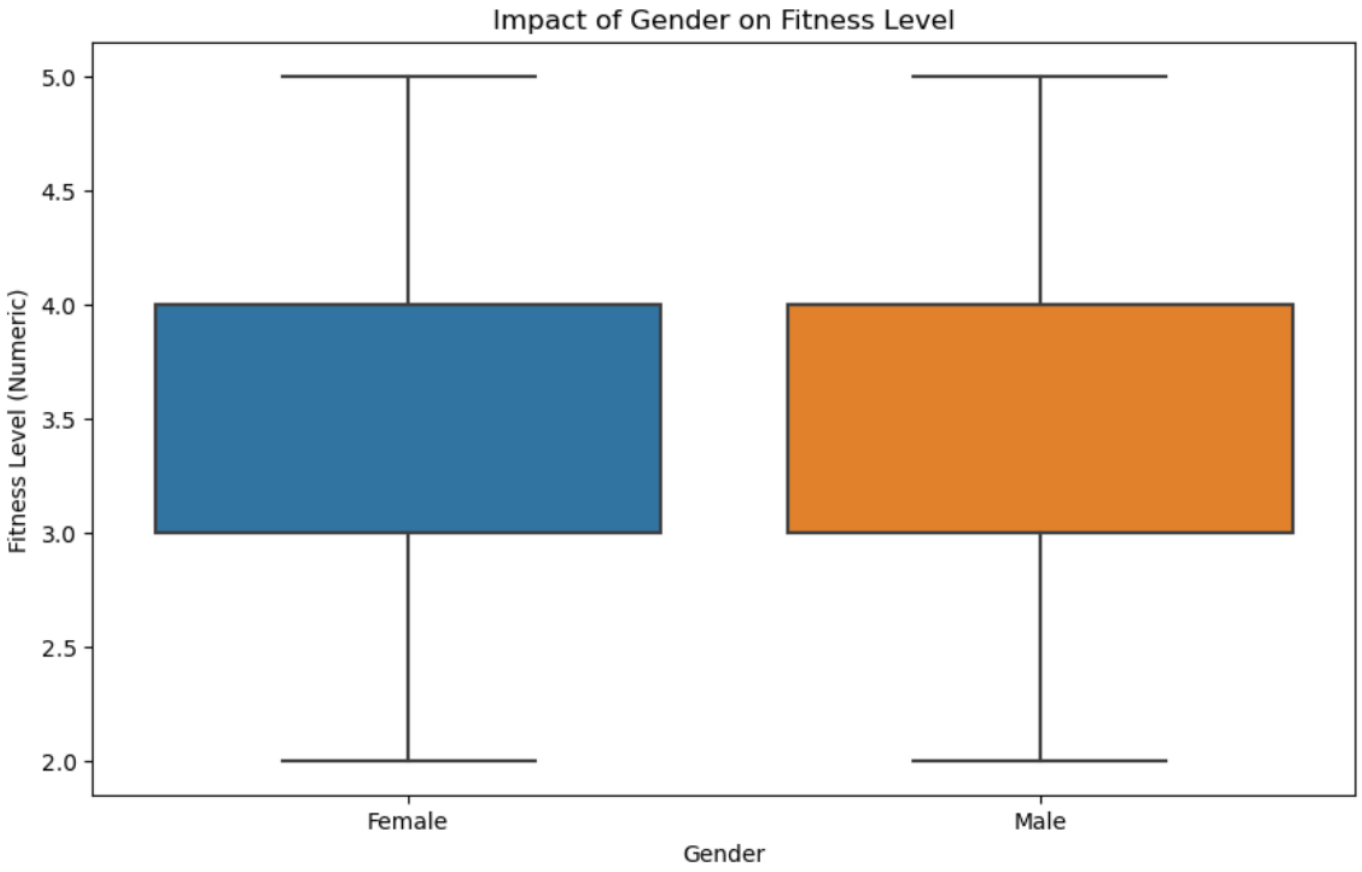
The box plots shows the distribution of fitness levels for users who report common barriers versus those who do not. Lower medians and quartiles for users with barriers show that lack of time and motivation, are associated with lower fitness levels.

EDA Summary

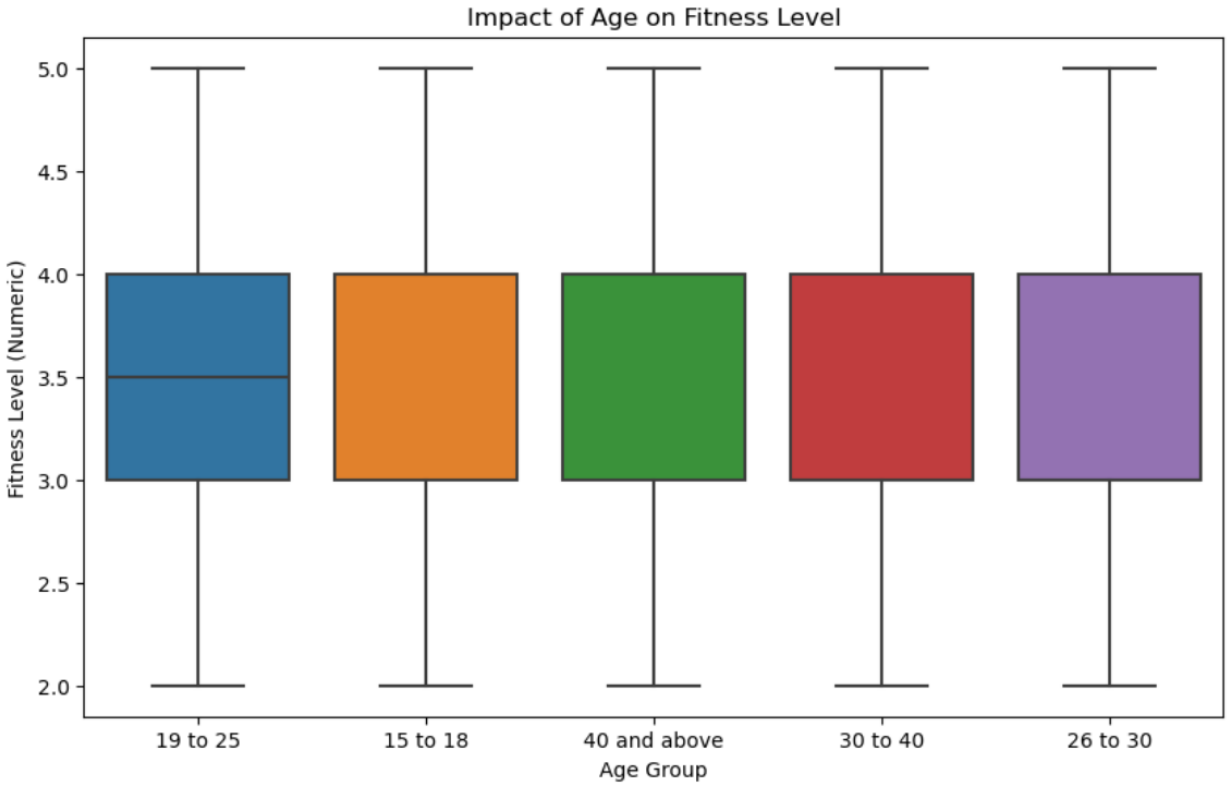
Influence of Demographics on Fitness and Exercise Habits

Age and gender significantly influence fitness levels and exercise habits.

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



EDA Summary



Influence of Demographics on Fitness and Exercise Habits

Age and gender significantly influence fitness levels and exercise habits.

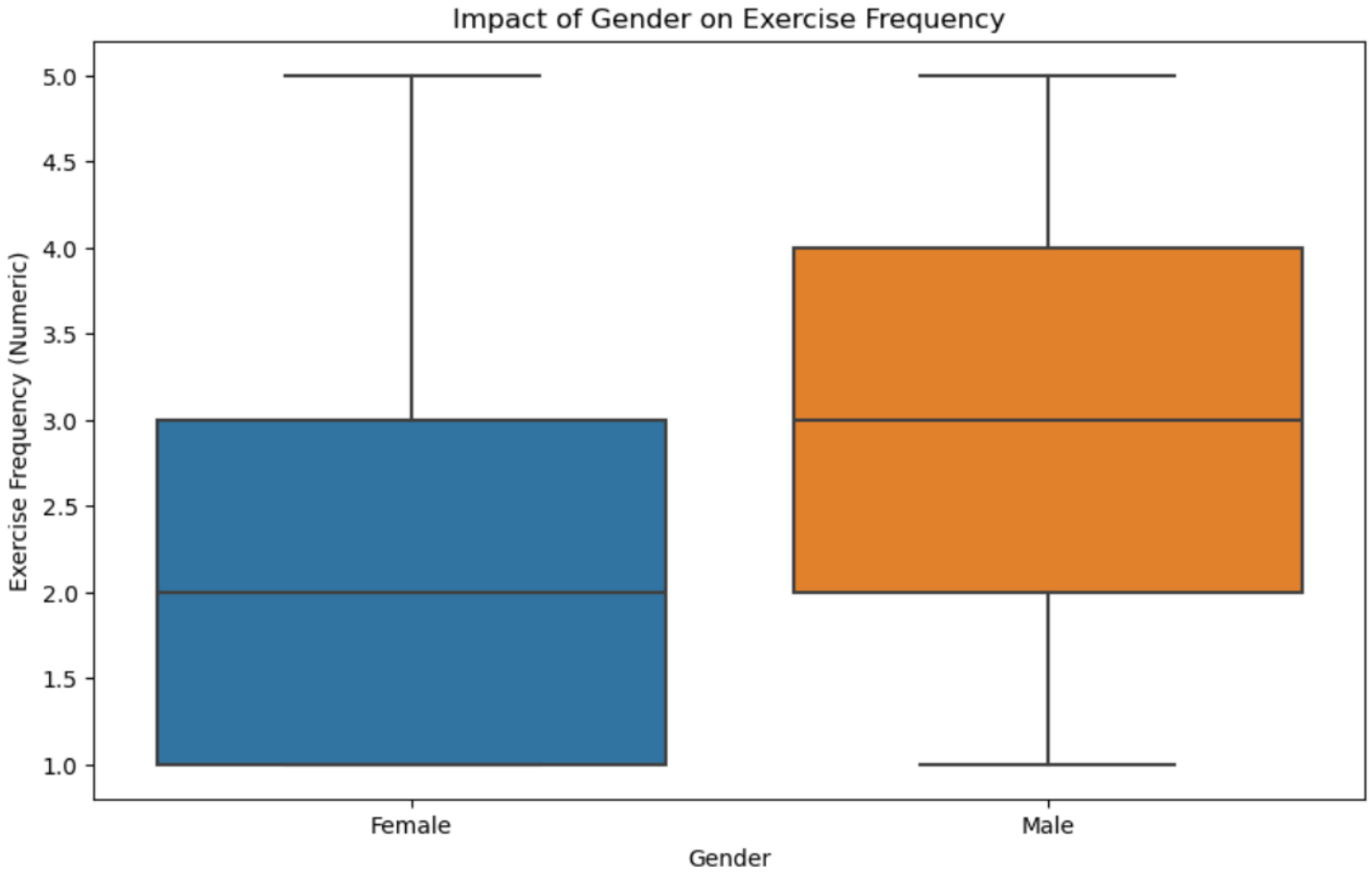
The box plots shows the distribution of fitness levels and exercise frequency across different genders and age groups. Significant differences would support the hypothesis.

EDA Summary

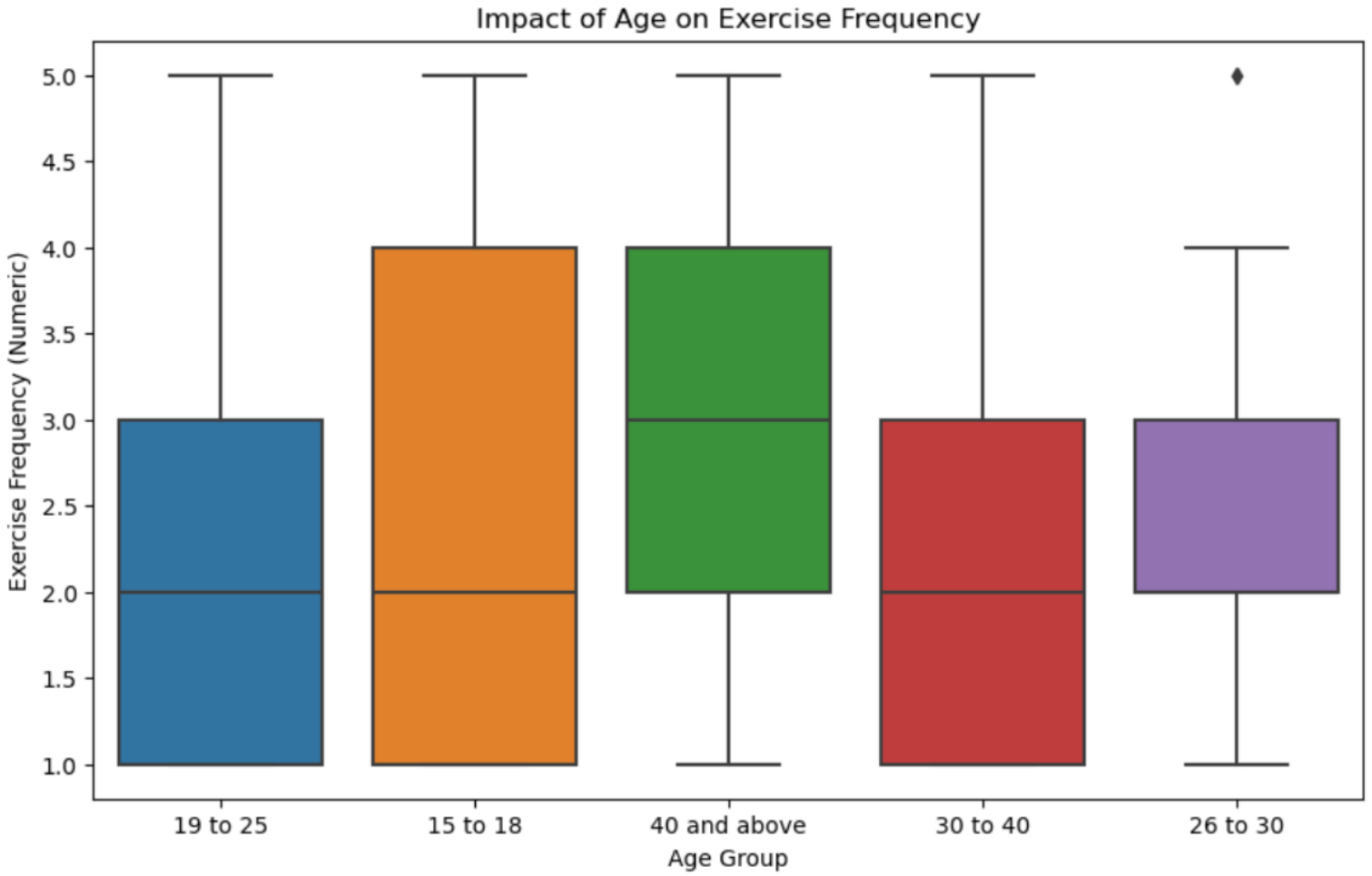
Influence of Demographics on Fitness and Exercise Habits

Age and gender significantly influence fitness levels and exercise habits.

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



EDA Summary



Influence of Demographics on Fitness and Exercise Habits

Age and gender significantly influence fitness levels and exercise habits.

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

Recommendations

These are some recommendations based on the analysis:

- **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.

Modeling Techniques

Based on the dataset, several models can be used to analyze different aspects and derive meaningful insights. Here are some of them:

- **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.

- **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.

- **Clustering Models**

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

Modeling Techniques

But to determine which model suits it best, it depends on specific objectives and the nature of the data. Here are some suitable models for different tasks based on the 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.

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.

Modeling Techniques

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.

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

1. **Classification**: Use Random Forest for predicting binary outcomes like motivation by fitness wearables. Random Forest provides a balance between interpretability and performance.
2. **Regression**: Use Random Forest Regression for predicting continuous variables such as exercise frequency or fitness levels.
3. **Clustering**: Use K-Means Clustering for segmenting users based on their exercise habits and motivations.

Thank You