Data Glacier

Your Deep Learning Partner

Data Analysis: Data collection pipeline

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

Data cleaning

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

Data cleaning

This is how the data was cleaned.

```
import pandas as pd

# Load the datasets
file_paths = {
    'fitness_analysis': 'C:/Users/Armel/OneDrive/Documents/Glacier work/Group project/fitness_analysis.csv',
    'fitness_consumer': 'C:/Users/Armel/OneDrive/Documents/Glacier work/Group project/fitness_consumer.csv',
    'fitness_trackers': 'C:/Users/Armel/OneDrive/Documents/Glacier work/Group project/fitness_trackers.csv'
}
```

```
datasets = {name: pd.read_csv(path) for name, path in file_paths.items()}
```

```
# Remove duplicates from each dataset
for name, data in datasets.items():
    datasets[name] = data.drop_duplicates()
```

```
# Fitness Trackers Dataset: Handle missing values and correct data types
fitness_trackers = datasets['fitness_trackers']

# Remove commas from 'Reviews' column and convert to integers
fitness_trackers['Reviews'] = fitness_trackers['Reviews'].fillna('0').str.replace(',', '').astype(int)
```

```
# Ensure there are no more issues with the data
fitness_trackers.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 606 entries, 0 to 609
Data columns (total 11 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Brand Name                    606 non-null    object
 1   Device Type                   606 non-null    object
 2   Model Name                    606 non-null    object
 3   Color                         606 non-null    object
 4   Selling Price                 606 non-null    object
 5   Original Price                606 non-null    object
 6   Display                       606 non-null    object
 7   Rating (Out of 5)             551 non-null    float64
 8   Strap Material                606 non-null    object
 9   Average Battery Life (in days) 606 non-null   int64
 10  Reviews                       606 non-null    int32
dtypes: float64(1), int32(1), int64(1), object(8)
memory usage: 54.4+ KB
```

Data cleaning

This is how the data was cleaned.

```python
# Update datasets dictionary with the cleaned fitness_trackers data
datasets['fitness_trackers'] = fitness_trackers
```

```python
# Handle missing values and correct data types for fitness_analysis dataset
fitness_analysis = datasets['fitness_analysis']

# Checking for missing values
missing_values_analysis = fitness_analysis.isnull().sum()
```

```python
# Handle missing values for fitness_consumer dataset
fitness_consumer = datasets['fitness_consumer']

# Checking for missing values
missing_values_consumer = fitness_consumer.isnull().sum()
```

```python
# Fill missing values if appropriate (assuming dropping rows with missing values for simplicity)
fitness_analysis = fitness_analysis.dropna()
fitness_consumer = fitness_consumer.dropna()
```

```python
# Update datasets dictionary with the cleaned data
datasets['fitness_analysis'] = fitness_analysis
datasets['fitness_consumer'] = fitness_consumer
```

```python
# Displaying first few rows of each cleaned dataset
cleaned_previews = {name: data.head() for name, data in datasets.items()}
cleaned_previews
```

```
{'fitness_analysis':                    Timestamp Your name  Your gender  Your age  \
0  2019/07/03 11:48:07 PM GMT+5:30   Parkavi       Female  19 to 25
1  2019/07/03 11:51:22 PM GMT+5:30  Nithilaa       Female  19 to 25
2  2019/07/03 11:56:28 PM GMT+5:30  Karunya v      Female  15 to 18
3   2019/07/04 5:43:35 AM GMT+5:30    Anusha       Female  15 to 18
4   2019/07/04 5:44:29 AM GMT+5:30  Nikkitha       Female  19 to 25

   How important is exercise to you ?  \
0                                   2
1                                   4
2                                   3
3                                   4
4                                   3

  How do you describe your current level of fitness ?  \
0                                               Good
1                                          Very good
2                                               Good
```

Data cleaning

This is the data after being cleaned. **Fitness analysis**

	Timestamp	Your name	Your gender	Your age	ortant is exercise to you ?	ır current level of fitness ?	ɔw often do you exercise?	ease select all that a
1	3 11:48:07 PM GMT+5:30	Parkavi	Female	19 to 25	2	Good	Never	:ime;I can't stay motiv
2	3 11:51:22 PM GMT+5:30	Nithilaa	Female	19 to 25	4	Very good	Never	ı time;I'll become too
3	3 11:56:28 PM GMT+5:30	Karunya v	Female	15 to 18	3	Good	1 to 2 times a week	I can't stay motiv
4	)4 5:43:35 AM GMT+5:30	Anusha	Female	15 to 18	4	Good	3 to 4 times a week	I don't have enough
5	)4 5:44:29 AM GMT+5:30	Nikkitha	Female	19 to 25	3	Unfit	Never	I can't stay motiv
6	)4 6:23:37 AM GMT+5:30	Girija	Female	40 and above	5	Average	3 to 4 times a week	; regularly with no bai
7	)4 6:33:21 AM GMT+5:30	Srinivasan	Male	40 and above	3	Good	1 to 2 times a week	ɔn't really enjoy exerc
8	)4 7:40:51 AM GMT+5:30	Ranjani	Female	15 to 18	3	Unfit	Never	ɔn't really enjoy exerc
9	)4 8:06:17 AM GMT+5:30	Bupesh R	Male	19 to 25	5	Unfit	3 to 4 times a week	ɔn't really enjoy exerc
10	)4 8:09:02 AM GMT+5:30	Sudhan	Male	15 to 18	5	Very good	Everyday	; regularly with no bai
11	)4 8:10:44 AM GMT+5:30	Revanth	Male	15 to 18	4	Very good	3 to 4 times a week	I don't have enough
12	)4 8:11:42 AM GMT+5:30	Ashwin	Male	19 to 25	5	Unfit	3 to 4 times a week	I don't have enough
13	)4 8:12:40 AM GMT+5:30	Gurjyot Singh	Male	15 to 18	4	Unfit	Never	I can't stay motiv
14	)4 8:13:38 AM GMT+5:30	Harshita Jain	Female	15 to 18	3	Average	1 to 2 times a week	me too tired;Less sta
15	)4 8:19:03 AM GMT+5:30	Hari Vishwa	Male	19 to 25	5	Average	1 to 2 times a week	e too tired;I have an i
16	)4 8:19:11 AM GMT+5:30	Harini sri	Female	15 to 18	3	Average	Everyday	I can't stay motiv
17	)4 8:24:57 AM GMT+5:30	Raghul Prashath.K.A	Male	15 to 18	3	Unfit	Never	ɔn't really enjoy exerc
18	)4 8:27:53 AM GMT+5:30	RJ	Male	15 to 18	5	Good	1 to 2 times a week	I'll become too
19	)4 8:28:19 AM GMT+5:30	Pranesh s	Male	19 to 25	3	Unfit	Never	e too tired;I have an i
20	)4 8:29:31 AM GMT+5:30	Prasath M	Male	15 to 18	3	Good	Everyday	ɔn't really enjoy exerc

Data cleaning

This is the data after being cleaned. **Fitness consumer**

Timestamp	What is your age?	What is your gender?	ighest level of education?	your current occupation?	o you exercise in a week?	using a fitness wearable?	se your fitness wearable?
:19 PM GMT+5:30	18-24	Male	ollege or associate degree	Student	5 or more times a week	Less than 6 months	Daily
:46 PM GMT+5:30	Under 18	Male	Bachelor's degree	Student	5 or more times a week	Less than 6 months	3-4 times a week
:46 PM GMT+5:30	18-24	Female	Bachelor's degree	Student	Less than once a week	Less than 6 months	Rarely
:07 PM GMT+5:30	25-34	Female	ollege or associate degree	Employed part-time	3-4 times a week	6-12 months	3-4 times a week
:32 PM GMT+5:30	18-24	Male	Bachelor's degree	Student	1-2 times a week	Less than 6 months	Daily
:56 PM GMT+5:30	18-24	Female	Master's degree	Employed full-time	5 or more times a week	1-2 years	Daily
:50 PM GMT+5:30	18-24	Male	Bachelor's degree	Student	Less than once a week	1-2 years	1-2 times a week
:08 AM GMT+5:30	18-24	Female	Bachelor's degree	Student	Less than once a week	Less than 6 months	Daily
:14 AM GMT+5:30	18-24	Male	High school diploma	Employed part-time	1-2 times a week	Less than 6 months	1-2 times a week
:50 AM GMT+5:30	35-44	Male	High school diploma	Employed full-time	Less than once a week	6-12 months	Daily
:26 AM GMT+5:30	35-44	Female	ate or professional degree	Self-employed	5 or more times a week	More than 2 years	Daily
:01 PM GMT+5:30	18-24	Female	Bachelor's degree	Self-employed	1-2 times a week	Less than 6 months	3-4 times a week
:50 PM GMT+5:30	25-34	Female	High school diploma	Employed part-time	Less than once a week	6-12 months	3-4 times a week
:33 PM GMT+5:30	45-54	Prefer not to say	Master's degree	Unemployed	3-4 times a week	6-12 months	1-2 times a week
:23 PM GMT+5:30	55-64	Prefer not to say	ate or professional degree	Retired	5 or more times a week	6-12 months	3-4 times a week
:08 PM GMT+5:30	45-54	Female	Bachelor's degree	Self-employed	Less than once a week	Less than 6 months	3-4 times a week
:21 PM GMT+5:30	25-34	Female	ollege or associate degree	Unemployed	3-4 times a week	Less than 6 months	3-4 times a week
:35 PM GMT+5:30	25-34	Male	ollege or associate degree	Self-employed	1-2 times a week	6-12 months	1-2 times a week
:32 AM GMT+5:30	25-34	Female	ate or professional degree	Self-employed	3-4 times a week	1-2 years	3-4 times a week
:39 AM GMT+5:30	55-64	Female	ate or professional degree	Retired	1-2 times a week	Less than 6 months	1-2 times a week

Data cleaning

This is the data after being cleaned. **Fitness trackers**

Brand Name	Device Type	Model Name	Color	Selling Price	Original Price	Display	Rating (Out of 5)
Xiaomi	FitnessBand	Smart Band 5	Black	2,499	2,999	AMOLED Display	4.1
Xiaomi	FitnessBand	Smart Band 4	Black	2,099	2,499	AMOLED Display	4.2
Xiaomi	FitnessBand	HMSH01GE	Black	1,722	2,099	LCD Display	3.5
Xiaomi	FitnessBand	Smart Band 5	Black	2,469	2,999	AMOLED Display	4.1
Xiaomi	FitnessBand	Band 3	Black	1,799	2,199	OLED Display	4.3
Xiaomi	FitnessBand	Band - HRX Edition	Black	1,299	1,799	OLED Display	4.2
Xiaomi	FitnessBand	Band 2	Black	2,499	2,499	OLED Display	4.3
Xiaomi	Smartwatch	Revolve	Black	12,349	15,999	AMOLED Display	4.4
Xiaomi	Smartwatch	RevolveActive	Black	12,999	15,999	AMOLED Display	4.4
Xiaomi	FitnessBand	Smart Band 3i	Black	1,270	1,599	OLED Display	4.2
OnePlus	FitnessBand	n Harrington Edition Band	Blue	3,299	3,999	AMOLED Display	4.3
OnePlus	FitnessBand	Band	Dual Color	2,499	2,799	AMOLED Display	4.2
FitBit	Smartwatch	Versa 2	Grey, Pink, Black	11,999	14,999	AMOLED Display	4.3
FitBit	Smartwatch	Sense	Black, Pink, Beige	21,499	22,999	AMOLED Display	4.2
FitBit	Smartwatch	Versa 3	Black, Blue, Pink	17,999	18,999	AMOLED Display	4.3
FitBit	FitnessBand	Charge 4	m Blue, Black, Rosewood	9,999	9,999	PMOLED Display	4.2
FitBit	FitnessBand	Inspire	Maroon	7,990	7,999	LED Display	4.2
FitBit	FitnessBand	Inspire 2	Desert Rose, Lunar White	6,999	7,999	PMOLED Display	4.4
FitBit	FitnessBand	Lunar	White	10,899	10,999	AMOLED Display	4.7
FitBit	FitnessBand	Charge 4	Granite Reflective	10,999	11,999	PMOLED Display	4.2

EDA Summary

```python
fitness_level_mapping = {
    'Very good': 5,
    'Good': 4,
    'Average': 3,
    'Unfit': 2,
    'Very unfit': 1
}
```

```python
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
}
```

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

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

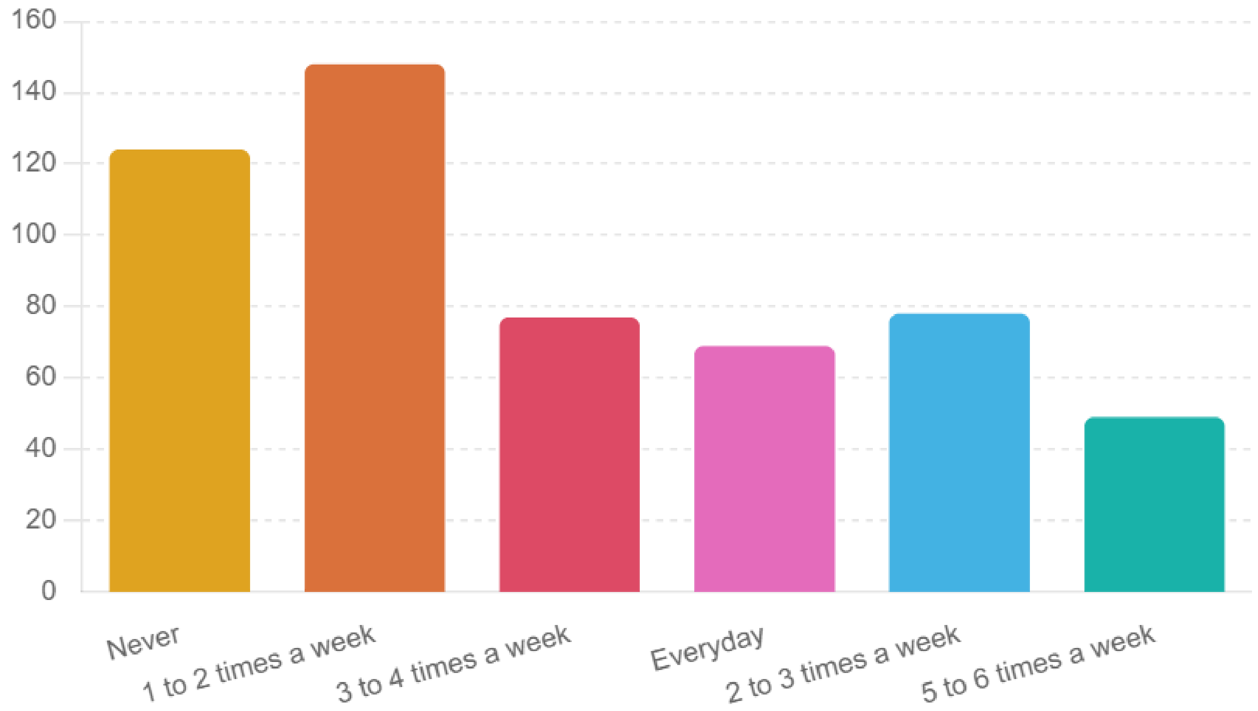
Encoding categorical data

To produce better visualizations, some of the categorical data needed was changed to numerical data.

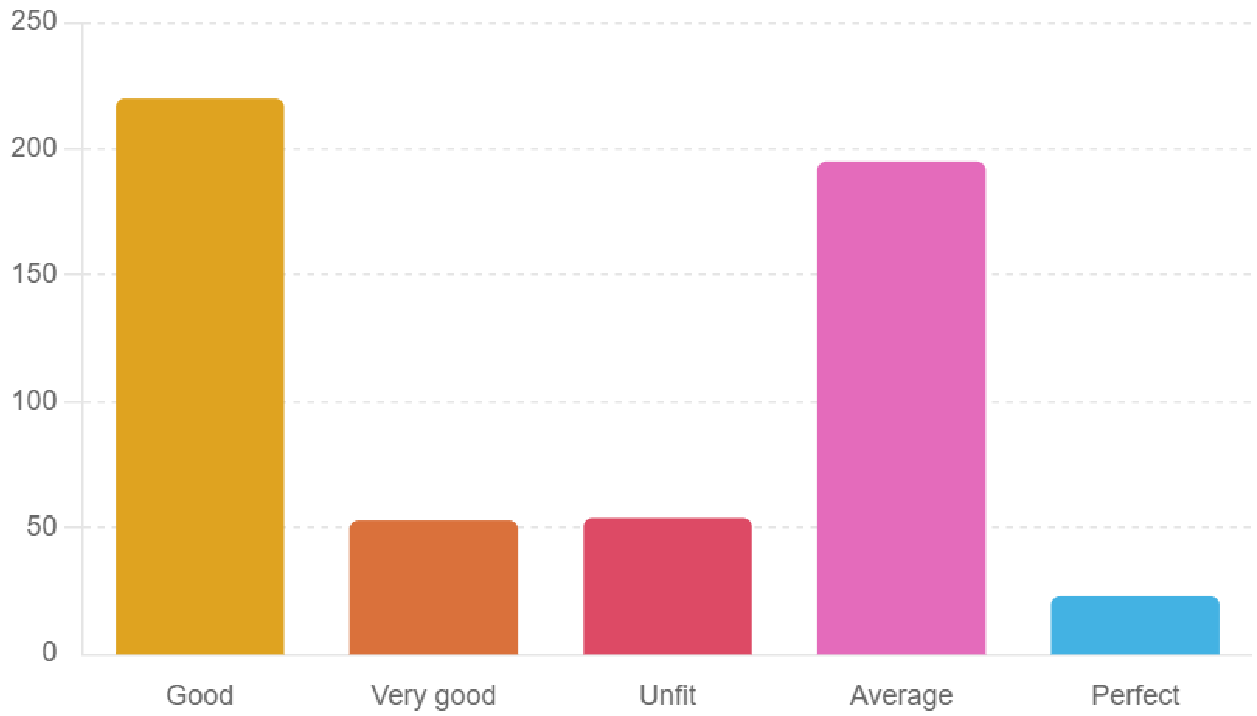
EDA Summary

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

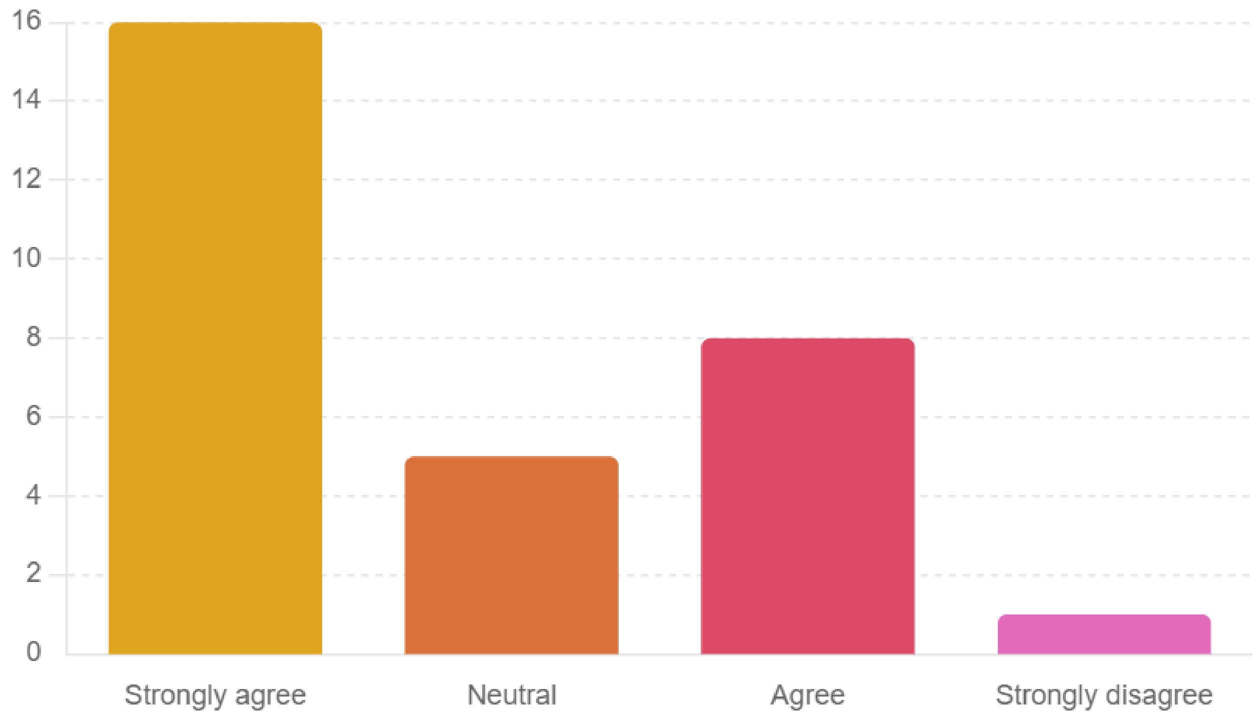


This second plot shows the distribution of exercise frequency from the fitness consumer dataset. This dataset indicates a higher frequency of exercise among respondents.

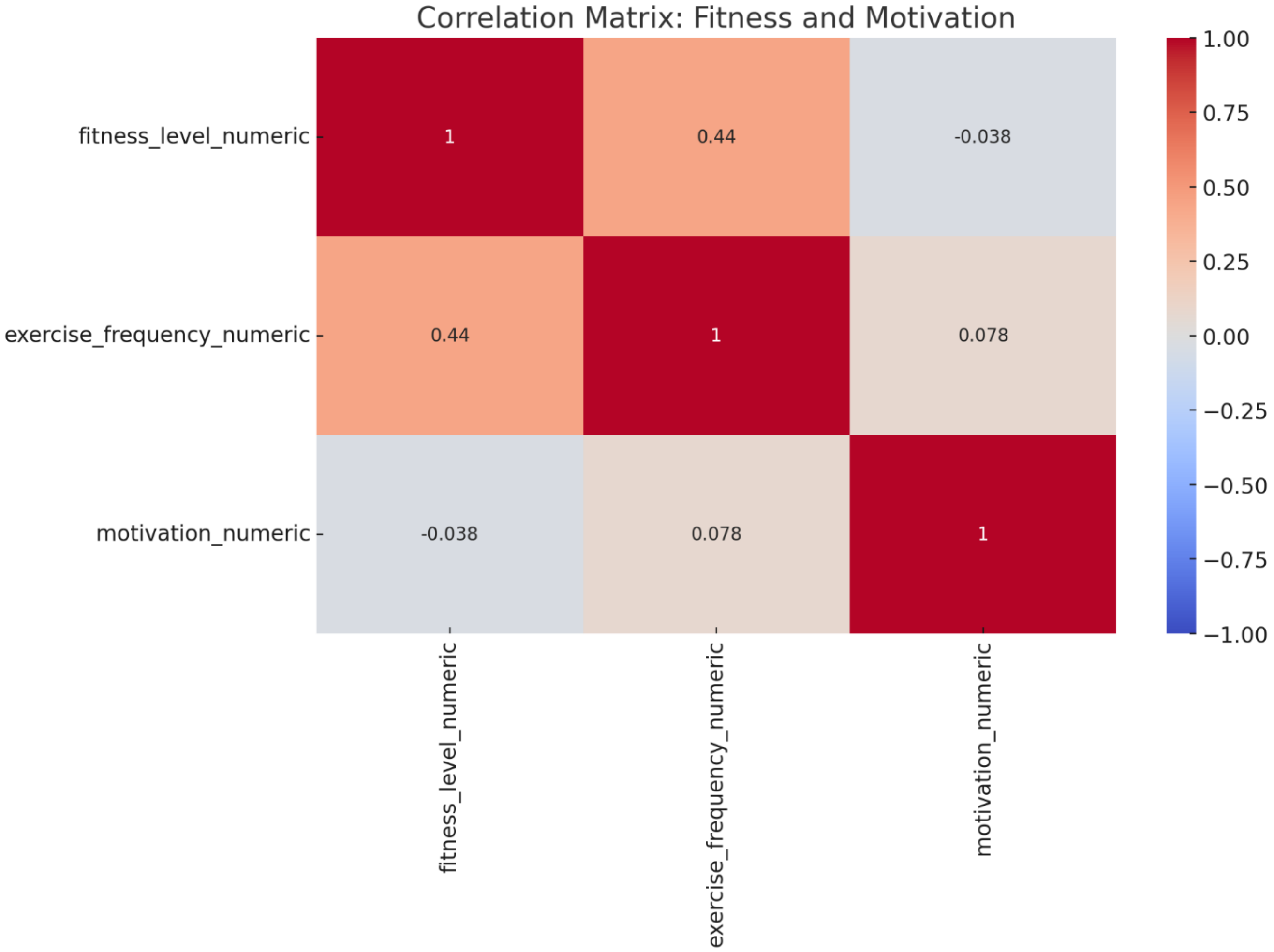
EDA Summary

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

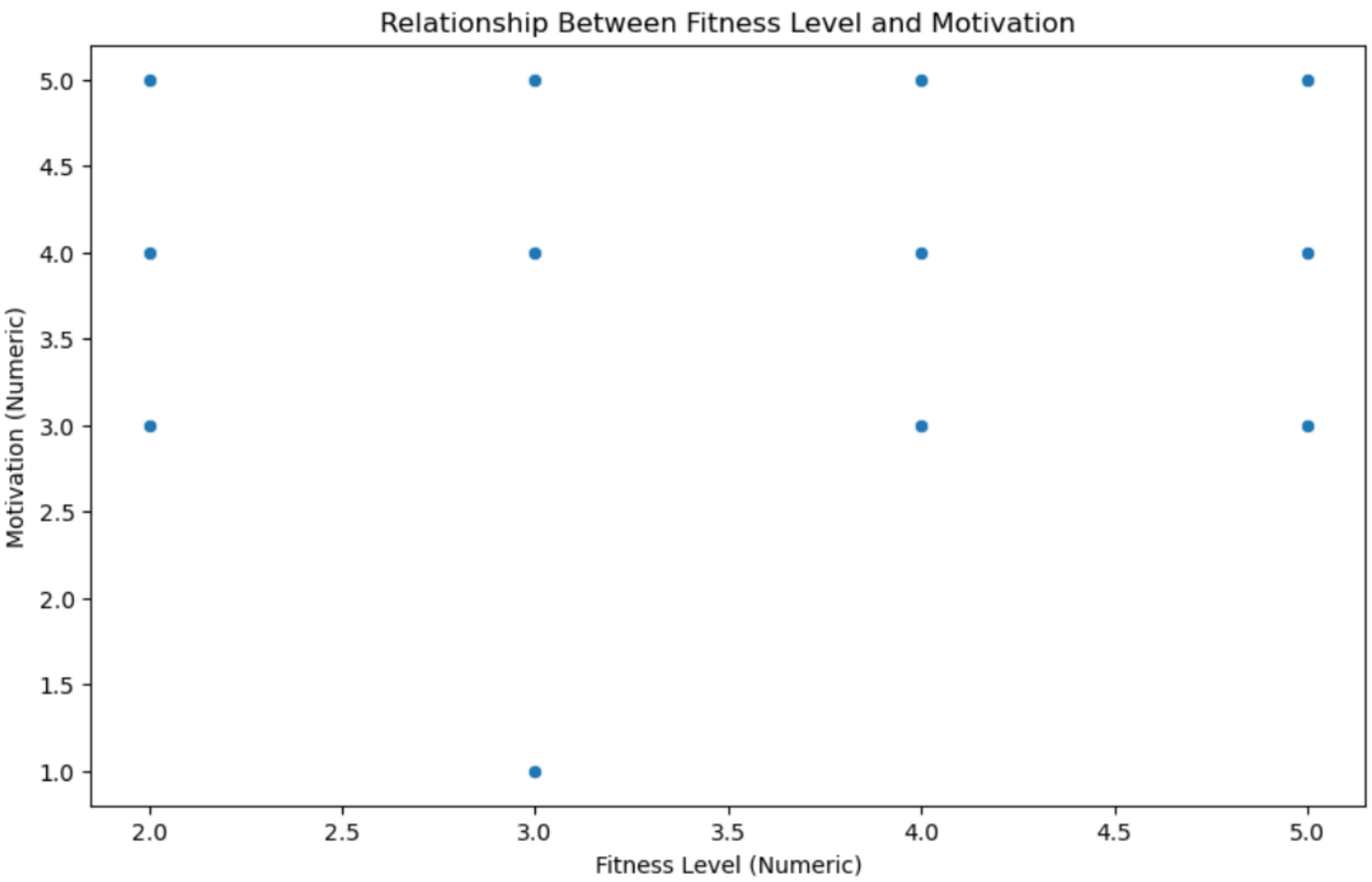


This fourth plot shows the correlation matrix and heatmap between fitness and motivation.

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



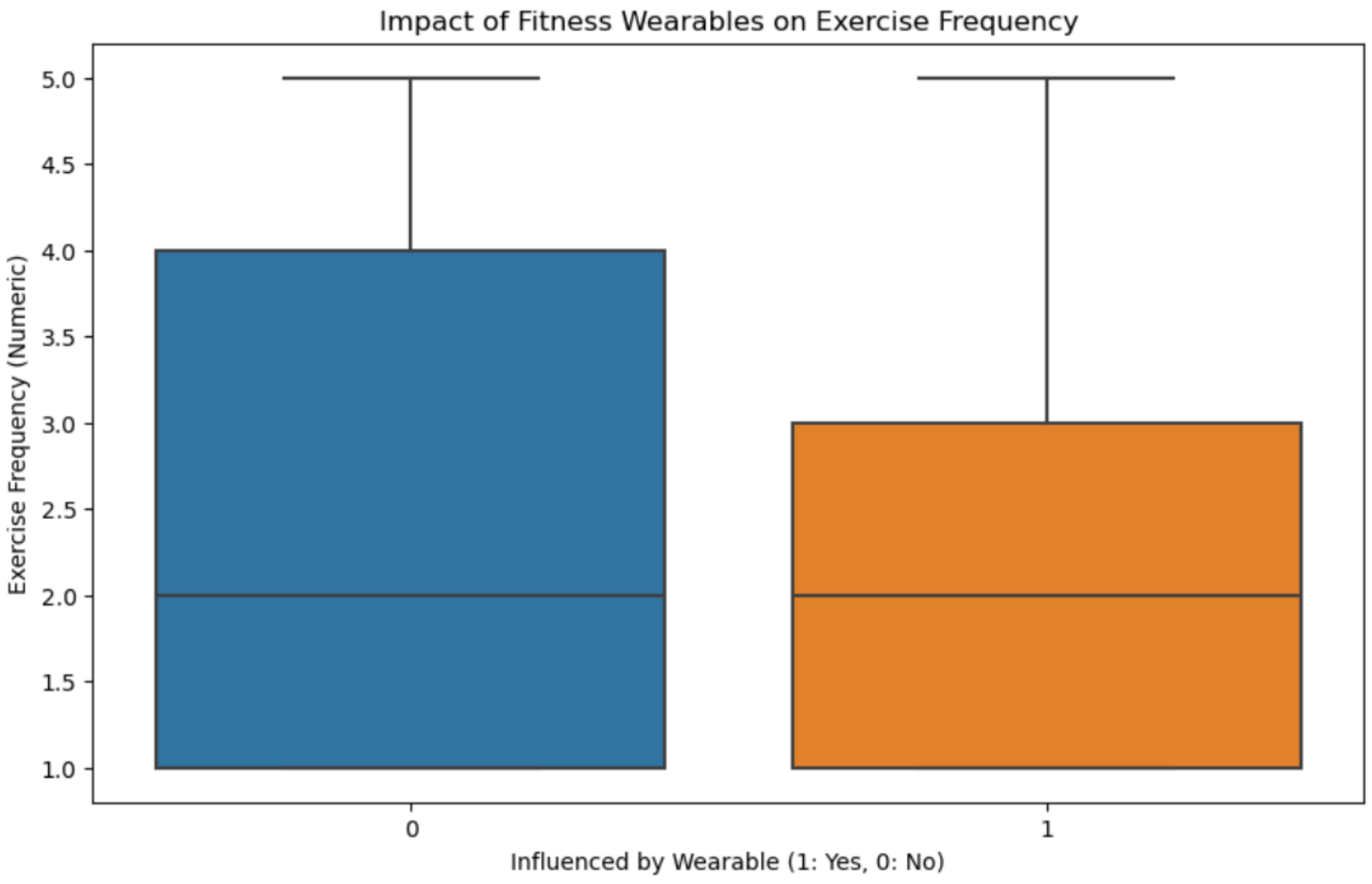
Relationship Between Fitness Level and Motivation

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.

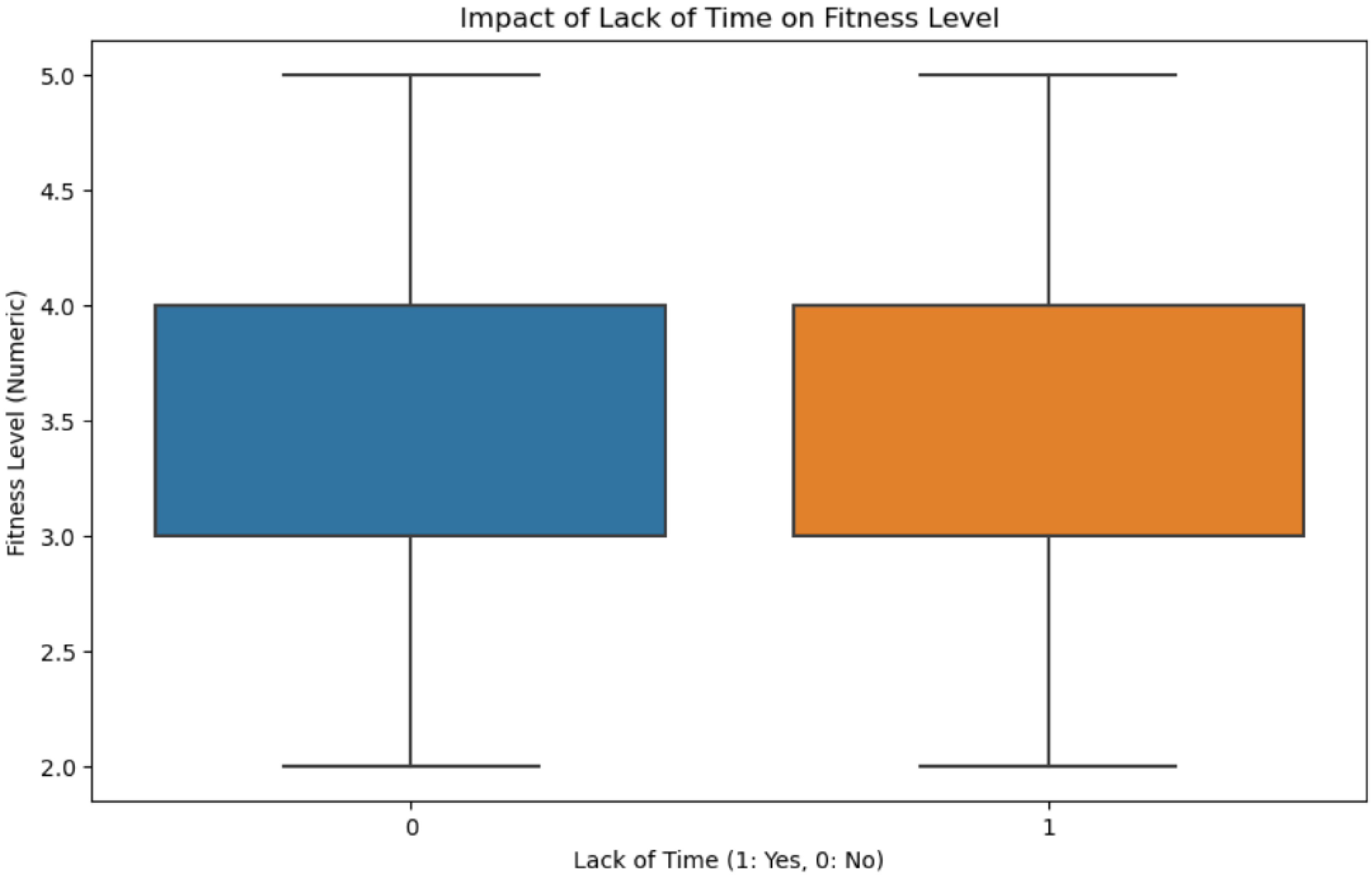
EDA Summary

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.



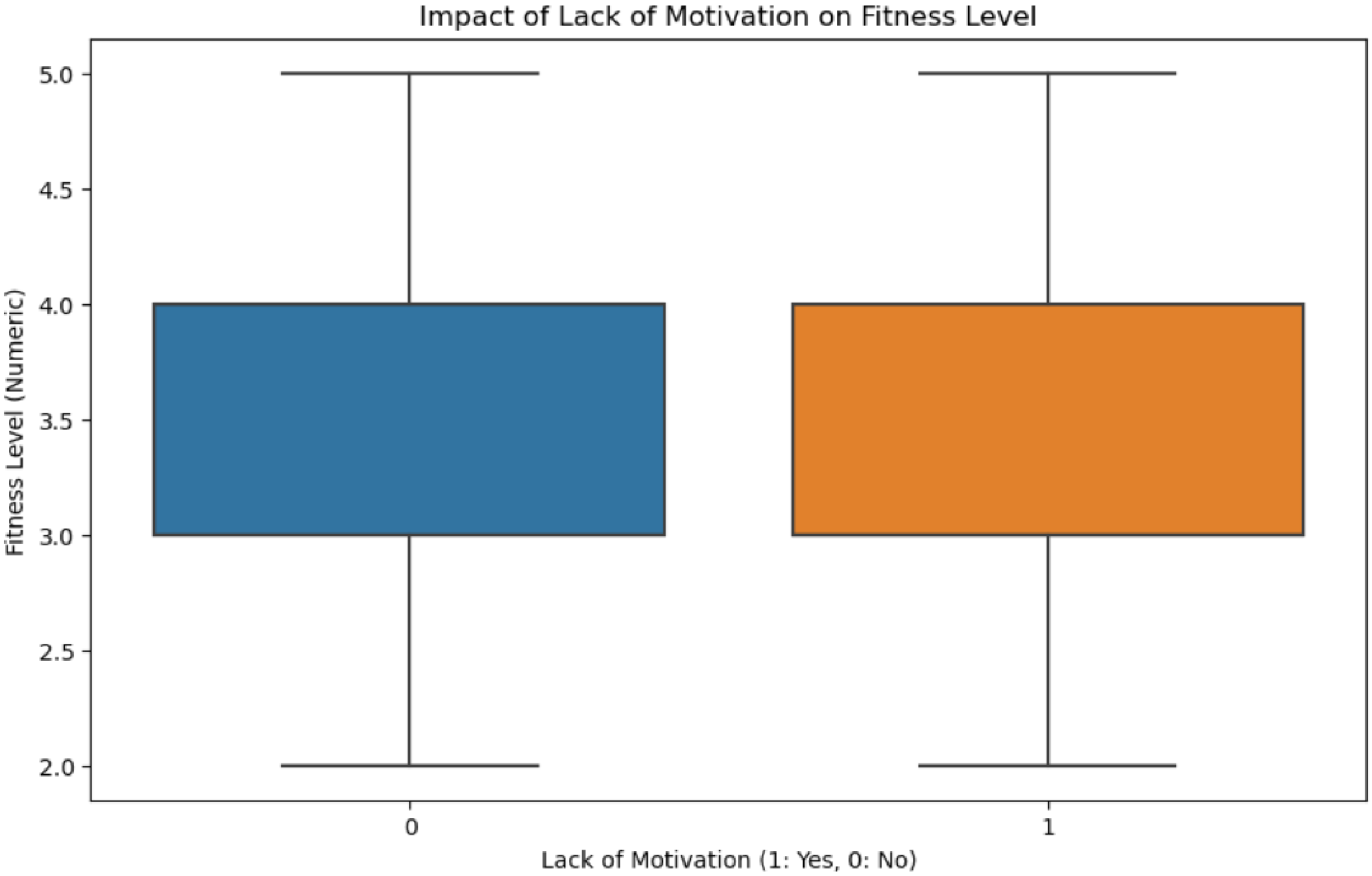
EDA Summary



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

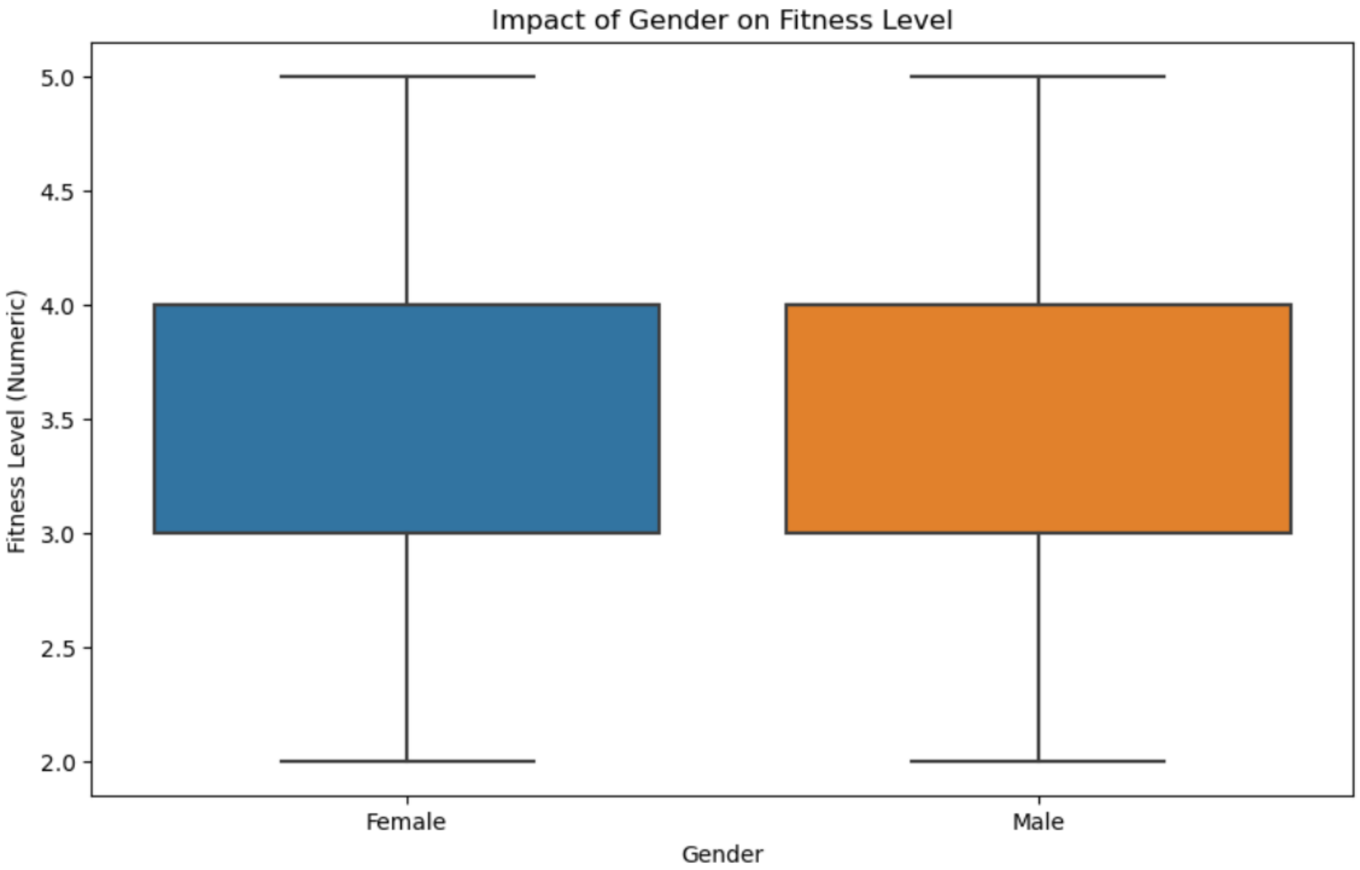
The box plots shows the distribution of fitness levels for users who report common barriers versus those who do not. Lower medians and quartiles for users with barriers show that lack of time and motivation, are associated with lower fitness levels.

EDA Summary

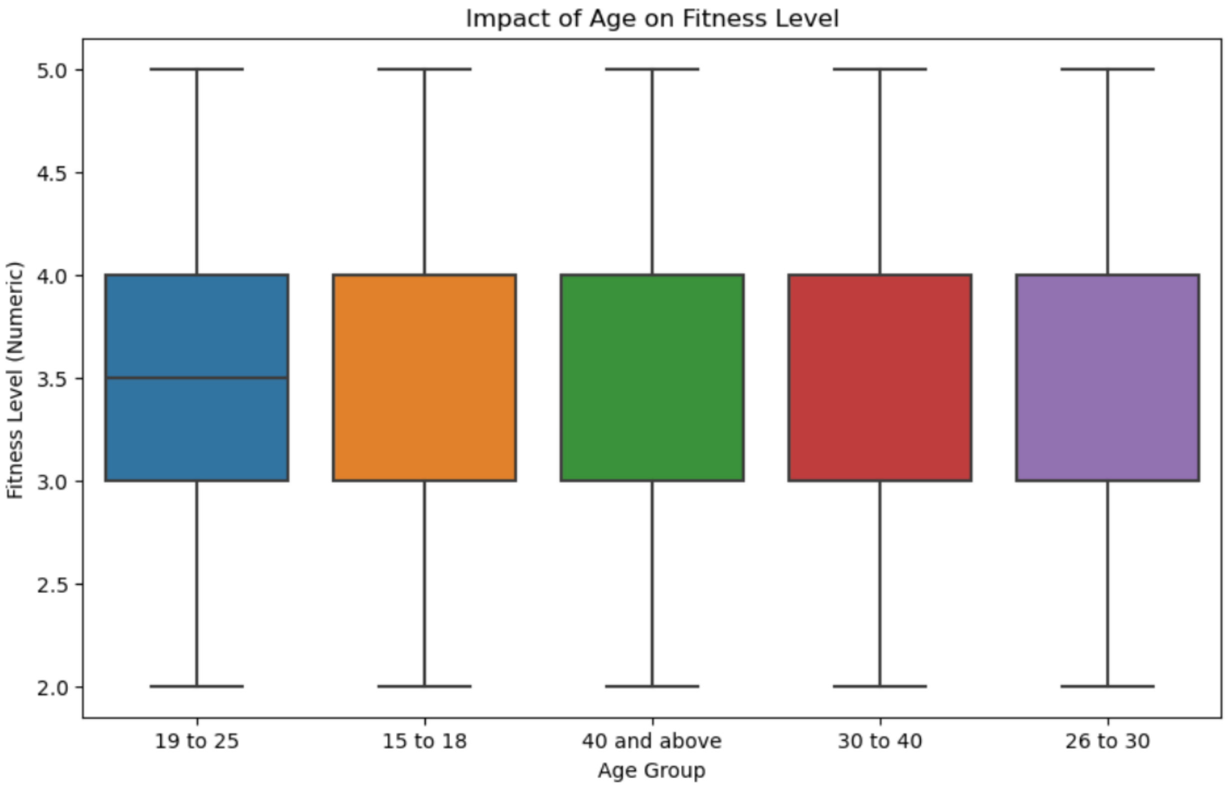
Influence of Demographics on Fitness and Exercise Habits

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.

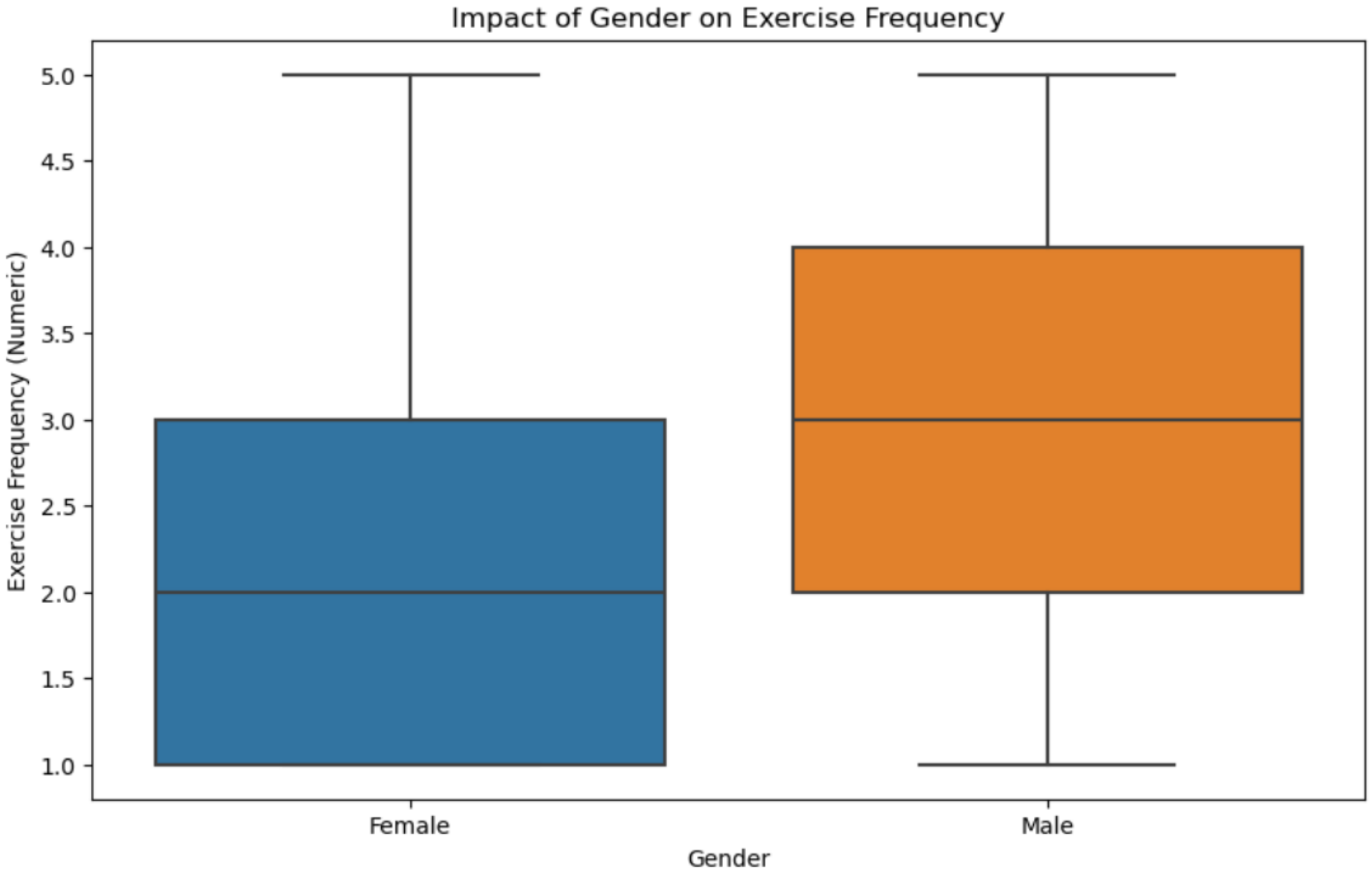
The box plots shows the distribution of fitness levels and exercise frequency across different genders and age groups. Significant differences would support the hypothesis.

EDA Summary

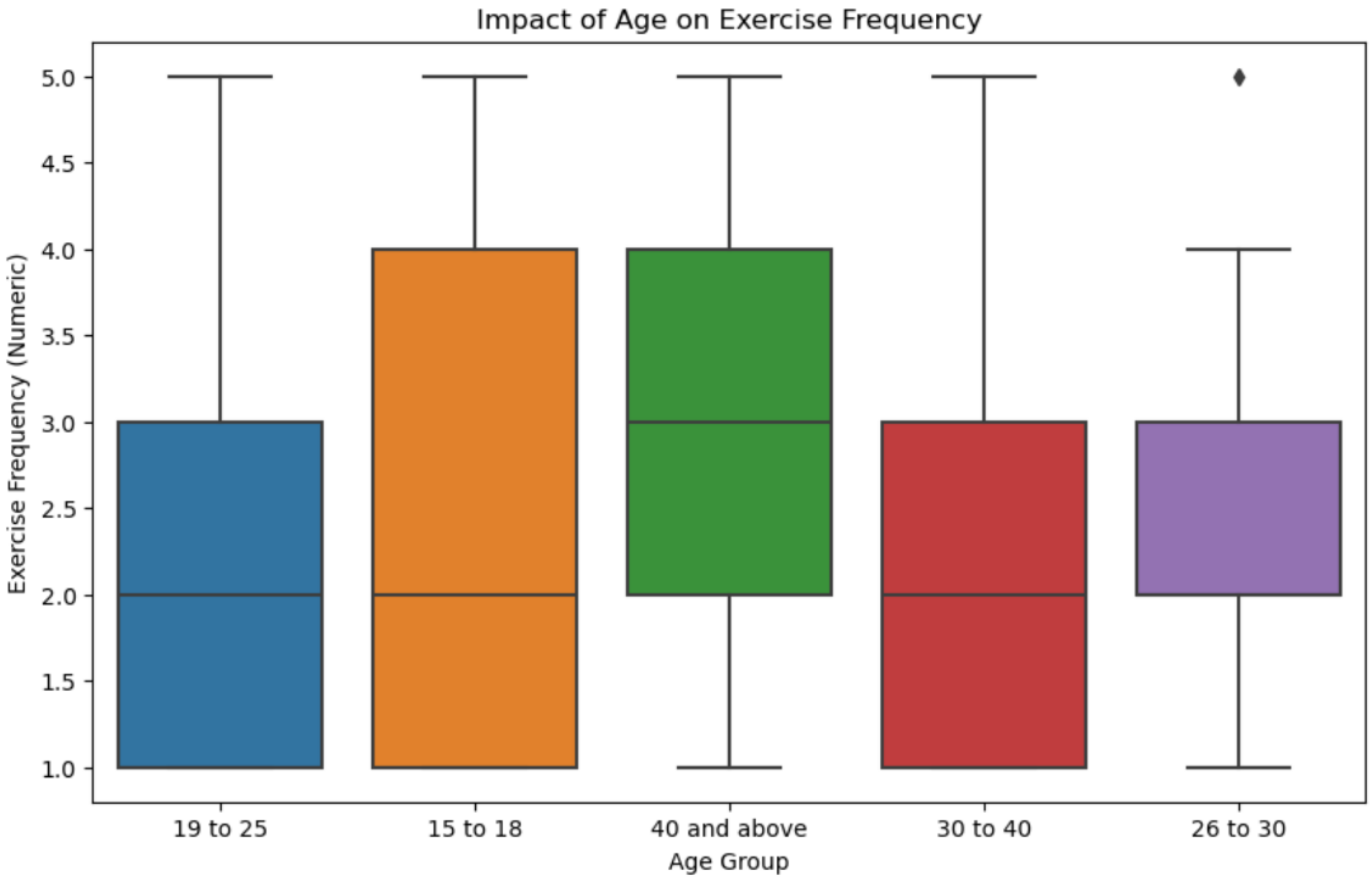
Influence of Demographics on Fitness and Exercise Habits

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.

Data Glacier
Your Deep Learning Partner

Thank You