CS 24200 Project 2 Report

Matthew Sindac

October 30, 2023

1 Summary

This technical report delves into a comprehensive data analysis of the Fitbit Fitness Tracker Data obtained from Kaggle. The dataset encompasses minute-level records of physical activity, heart rate, and sleep monitoring, offering a source of insights into human health and behavior. The objective of this project is to discover the relationships between exercise, sleep patterns, and energy expenditure. Additionally, we aim to explore and identify correlations between various variables within the dataset, showcasing their influence on personal well-being.

The analyses have given various amounts of conclusions that show the key aspects of the data. In this project, the data revealed significant relationships between exercise intensity and sleep quality, providing insights into how physical activity impacts restorative rest. Moreover, the findings reveal patterns between the number of steps taken and the duration of sleep, possibly indicating potential connections between daily activity and nighttime recovery.

Overall, this report serves as an exploration of the Fitbit Fitness Tracker Data, offering valuable insights into the relationships between exercise, sleep, and energy expenditure. These findings have the potential to inform health and wellness strategies, providing individuals with data-driven guidance for improving their overall well-being.

2 Introduction

In our exploration of the Fitbit Fitness Tracker Data, our objective is to discover valuable insights by jumping into the datasets using many different data processing tools, visualization techniques, and comparative analysis. This approach will enable us to unlock the full potential of this data as a whole.

This analysis is dedicated to identifying various relationships within the data, with the ultimate goal of offering guidance for personal fitness and overall well-being. By looking at various data points and employing analytical methods, we aim to reveal correlations and create hypotheses to better understand this dataset. Through this exploration, we will use data to create innovative interpretations that can contribute to the broader field of health and fitness.

3 Methodology

3.1 Distributions and Outliers

For the following processes, I imported relative datasets that were needed for the variables expected and created data visualizations in order to showcase the data in various different forms.

3.1.1 Heart Rate for Different Users

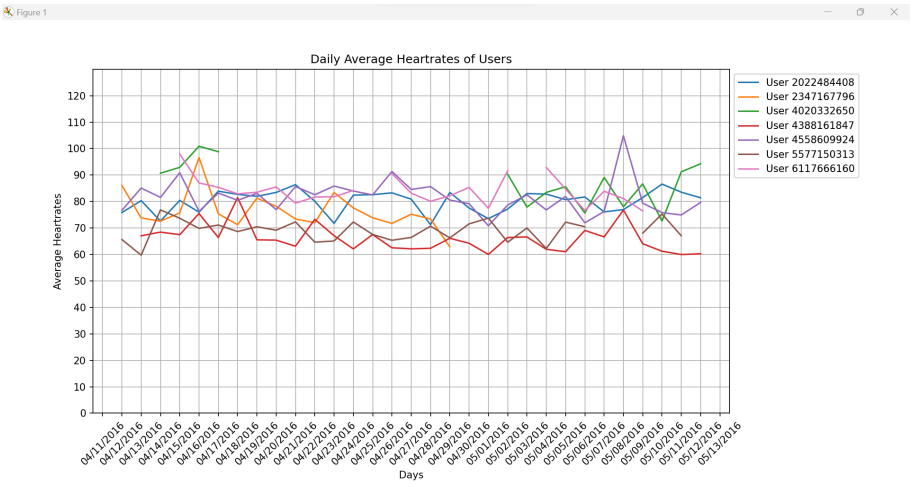
To gain insights into heart rate patterns for different users, I utilized a dataset named "heartrate_seconds_merged.csv" which contained heart rate measurements taken at 5 second intervals for each user. This dataset consisted of key attributes which was:

- Time: The timestamp of the heart rate measurement.
- Id: The unique identifier for each user.
- Value: The heart rate value recorded at a given time.

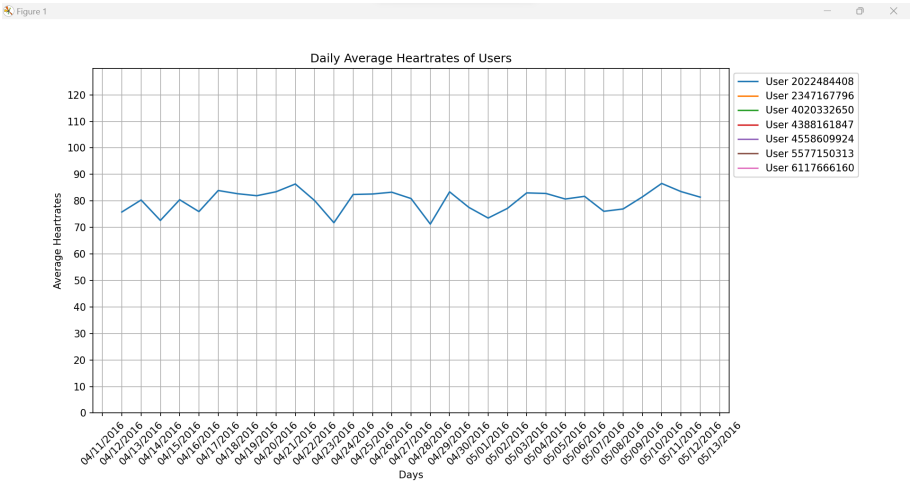
Starting off this process, I took 7 different users from this dataset using the data in the 'Id' column and gathered their heart rate data for the time period between April 12, 2016 and May 12, 2016. I resampled each of their data to calculate the daily average heart rates for each user, which was done by grouping the data by the user's ID and then gathering their average daily heart rate for that day.

Afterwards, I created a line graph in order to create a data visualization of the daily average heart rates of the selected users over time, where each user is represented by a distinct line on the graph. In this visualization, I set up the x-axis where it showcases the day by day period from April 12, 2016 through May 12, 2016 and the y-axis which represents the average heart rate of each user. I implemented keyboard shortcuts where users can press keys 1-7 to toggle the visibility of individual user's data lines, being able to showcase them one by one. Pressing '0' resets the graph to display all lines and users. I added a legend to depict each user's line and made sure to make the plot clear and informative.

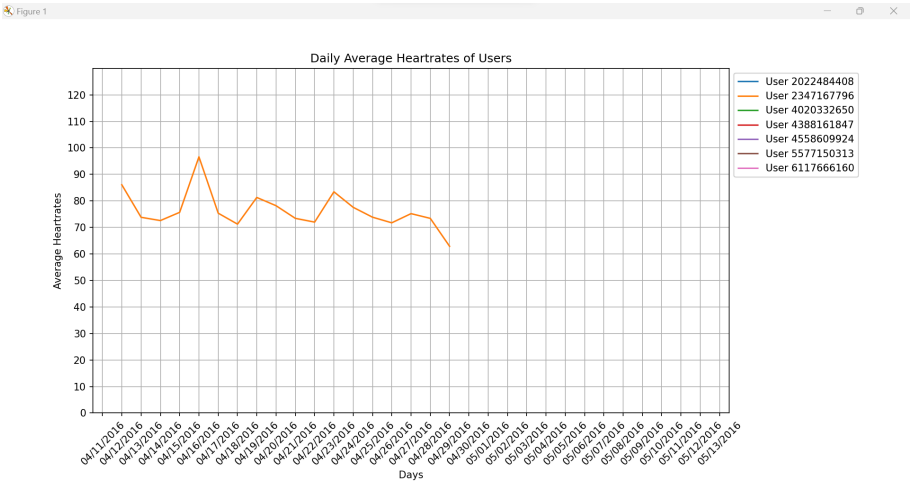
3.1.2 Visualizations for Heart Rate's of Different Users



(Heart rate graph of all users)



(Heart rate graph of user 2022484408)



(Heart rate graph of user 4558609924)

3.1.3 Daily Sleep Duration Analysis

To gain insights into the daily sleep duration of each user, we imported the sleep data from the "sleepDay_merged.csv" CSV file. This dataset contains the following attributes:

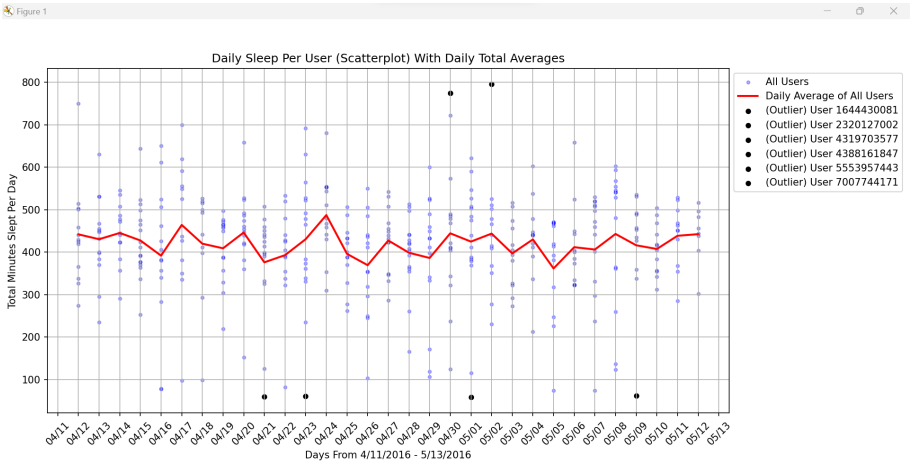
- SleepDay: A timestamp indicating the date and time of sleep data collection.
- Id: A unique identifier for each user.
- TotalMinutesAsleep: The total number of minutes each user slept on a specific day.

Since the "TotalMinutesAsleep" was already available for each day of each user, we proceeded to create a scatter plot based on these totals for each user for each day. In this visualization, the x-axis represents the date range, spanning day by day from April 12, 2016, to May 12, 2016, and the y-axis shows the total minutes slept per day. This scatter plot was complemented with a red line plot, showcasing the daily average of total minutes asleep for all users over the same time frame.

The scatter plot includes several significant elements:

- Blue Dots: These dots represent individual data points for all users, allowing us to see their sleep duration variations.
- Outliers: Outliers, or data points significantly different from the majority, are marked with black dots. The users with outliers are identified in the legend to the right of the plot. We determined these outliers by calculating z-scores based on "TotalMinutesAsleep" within the dataset and using a threshold of 3.

The outliers in this data were pretty surprising since it indicated that some users were either getting about an hour of sleep per day or up to thirteen hours of sleep per day which both deviate greatly from the mean amount of sleep per day for all users.



(Daily sleep graph of all users)

3.1.4 Analyzing Daily Steps

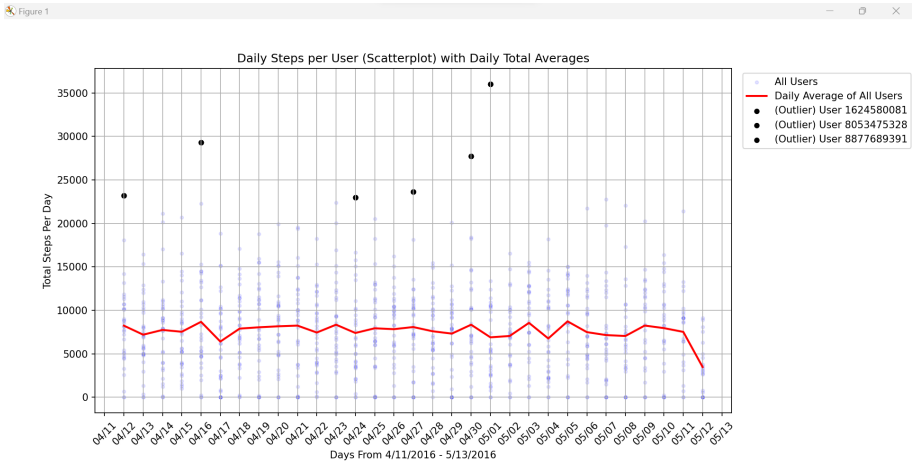
This section focuses on examining the distribution of daily steps recorded by users. We begin by importing data from the "dailySteps_merged.csv" file. The dataset contains essential attributes:

- ActivityDay: A timestamp denoting the date of recorded activity.
- Id: A unique identifier assigned to each user.
- StepTotal: The total number of steps taken by each user on a given day.

I computed the daily average of total steps for all users. Since the "StepTotal" column already contained total steps for each day, I just inputted it into the scatterplot. We visualized the daily steps data using a scatter plot. The x-axis represents the date, while the y-axis shows the total steps per day. A red line, indicative of the daily average of total steps for all users, was overlaid on the scatter plot for the period spanning April 12, 2016, through May 12, 2016. The scatter plot includes several significant elements:

- Blue Dots: These dots represent individual data points for all users, enabling us to observe variations in daily step counts.
- Outliers: Outliers, identified by black dots, signify data points significantly deviating from the majority. Users with outliers are listed in the legend on the right. We detected outliers by calculating z-scores based on the "StepTotal" column, employing a threshold of 3.

The outliers in this graph were pretty interesting since some of them were in between 22500 steps and 30000 steps with one extreme outlier being above 35000 steps. This means that some people gathered an extreme amount of steps in comparison to the average. One possible explanation for these observations is that some users may be marathon runners or engage in activities that involve covering greater distances than the typical person.



(Daily steps graph of all users)

3.1.5 Analyzing Weight Change for the User with the Most Records

In this section, I used the weight data recorded in the "weightLogInfo_merged.csv" dataset. I identified the user with the highest number of weight records in the dataset and created a visualization surrounding their weight change over a specific time range, from April 12, 2016, to May 12, 2016.

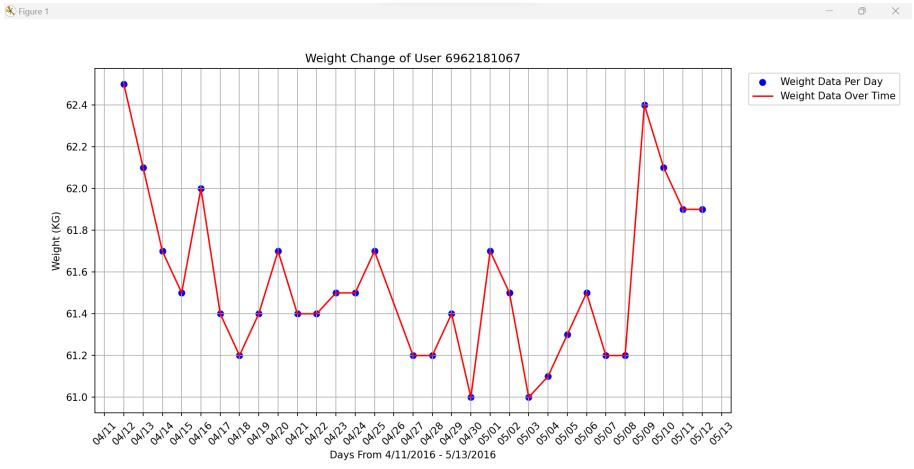
I began by reading in the dataset and determining the user with the highest number of weight records. This user's 'Id' was selected for further analysis.

The user's weight data was processed and formatted to ensure accurate representation. The 'Date' column was transformed into a datetime format, and an adjustment of one day was made to align the data correctly.

A scatterplot was utilized to visually present the user's weight measurements recorded on different days, denoted in blue. To provide insight into the trend of weight change over time, a red line plot was overlaid on the scatterplot.

Components of the Visualization:

- Blue Dots: These dots represent the individual weight measurements recorded on different days for the selected user. They allow us to visualize daily variations in weight.
- Red Line: The red line plot displays the trend in weight change over the specified time range, offering a clear visualization of the user's weight journey.



(Weight change graph of a singular user)

3.2 Data Processing

3.2.1 Merging Hourly and Minutely Dataframes

	Id	ActivityHour	StepTotal	TotalIntensity	\
0	1503960366	4/12/2016 12:00:00 AM	373	20	
1	1503960366	4/12/2016 1:00:00 AM	160	8	
2	1503960366	4/12/2016 2:00:00 AM	151	7	
3	1503960366	4/12/2016 3:00:00 AM	0	0	
4	1503960366	4/12/2016 4:00:00 AM	0	0	

	AverageIntensity	Calories	Date
0	0.333333	81	2016-04-12 00:00:00
1	0.133333	61	2016-04-12 01:00:00
2	0.116667	59	2016-04-12 02:00:00
3	0.000000	47	2016-04-12 03:00:00
4	0.000000	48	2016-04-12 04:00:00

(Hourly Dataframe)

First, I imported the necessary csv files that were to be merged which included 'hourlySteps_merged.csv', 'hourlyIntensities_merged.csv' and 'hourlyCalories_merged.csv'.

I merged each dataframe on the 'Id' column and the 'ActivityHour' column with an inner merge in order to filter out unmatched data within the dataset.

	Id	ActivityMinute	METs	Intensity	Calories	\
0	1503960366	4/12/2016 12:00:00 AM	10	0	0.7865	
1	1503960366	4/12/2016 12:01:00 AM	10	0	0.7865	
2	1503960366	4/12/2016 12:02:00 AM	10	0	0.7865	
3	1503960366	4/12/2016 12:03:00 AM	10	0	0.7865	
4	1503960366	4/12/2016 12:04:00 AM	10	0	0.7865	

	Date
0	2016-04-12 00:00:00
1	2016-04-12 00:01:00
2	2016-04-12 00:02:00
3	2016-04-12 00:03:00
4	2016-04-12 00:04:00

(Minutely Dataframe)

I did the same for the minutely dataframe where I imported the csv files

('minuteCaloriesNarrow_merged.csv', 'minuteIntensitiesNarrow_merged.csv' and 'minuteMETsNarrow_merged.csv').

I merged each dataframe on the 'Id' column and the 'ActivityMinute' column with an inner merge in order to filter out unmatched data within the dataset.

3.2.2 Activity, Sleep, Hourly and Minutely Dateframe Conversion

For the dataframe containing 'dailyactivity_merged.csv', 'sleepDay_merged.csv', along with the hourly and minutely dataframe, I created a new column named 'Date' and utilized the pandas command (.to_datetime) with specialized formatting for each specific dataframe to convert the time string to a datetime type and then saved each one to its respective new 'Date' column..

	Id	ActivityHour	StepTotal	TotalIntensity	\
0	1503960366	4/12/2016 12:00:00 AM	373	20	
1	1503960366	4/12/2016 1:00:00 AM	160	8	
2	1503960366	4/12/2016 2:00:00 AM	151	7	
3	1503960366	4/12/2016 3:00:00 AM	0	0	
4	1503960366	4/12/2016 4:00:00 AM	0	0	

	AverageIntensity	Calories	Date
0	0.333333	81	2016-04-12 00:00:00
1	0.133333	61	2016-04-12 01:00:00
2	0.116667	59	2016-04-12 02:00:00
3	0.000000	47	2016-04-12 03:00:00
4	0.000000	48	2016-04-12 04:00:00

(Hourly dataframe)

	Id	ActivityMinute	METs	Intensity	Calories	\
0	1503960366	4/12/2016 12:00:00 AM	10	0	0.7865	
1	1503960366	4/12/2016 12:01:00 AM	10	0	0.7865	
2	1503960366	4/12/2016 12:02:00 AM	10	0	0.7865	
3	1503960366	4/12/2016 12:03:00 AM	10	0	0.7865	
4	1503960366	4/12/2016 12:04:00 AM	10	0	0.7865	

	Date
0	2016-04-12 00:00:00
1	2016-04-12 00:01:00
2	2016-04-12 00:02:00
3	2016-04-12 00:03:00
4	2016-04-12 00:04:00

(Minutely dataframe)

	Id	SleepDay	TotalSleepRecords	TotalMinutesAsleep	\
0	1503960366	4/12/2016 12:00:00 AM	1	327	
1	1503960366	4/13/2016 12:00:00 AM	2	384	
2	1503960366	4/15/2016 12:00:00 AM	1	412	
3	1503960366	4/16/2016 12:00:00 AM	2	340	
4	1503960366	4/17/2016 12:00:00 AM	1	700	

	TotalTimeInBed	Date
0	346	2016-04-12
1	407	2016-04-13
2	442	2016-04-15
3	367	2016-04-16
4	712	2016-04-17

(Sleep dataframe)

	Id	ActivityDate	TotalSteps	TotalDistance	TrackerDistance	\
0	1503960366	4/12/2016	13162	8.50	8.50	
1	1503960366	4/13/2016	10735	6.97	6.97	
2	1503960366	4/14/2016	10460	6.74	6.74	
3	1503960366	4/15/2016	9762	6.28	6.28	
4	1503960366	4/16/2016	12669	8.16	8.16	

	LoggedActivitiesDistance	VeryActiveDistance	ModeratelyActiveDistance	\
0	0.0	1.88	0.55	
1	0.0	1.57	0.69	
2	0.0	2.44	0.40	
3	0.0	2.14	1.26	
4	0.0	2.71	0.41	

	LightActiveDistance	SedentaryActiveDistance	VeryActiveMinutes	\
0	6.06	0.0	25	
1	4.71	0.0	21	
2	3.91	0.0	30	
3	2.83	0.0	29	
4	5.04	0.0	36	

	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories	\
0	13	328	728	1985	
1	19	217	776	1797	
2	11	181	1218	1776	
3	34	209	726	1745	
4	10	221	773	1863	

(Daily activity dataframe)

	Date
0	2016-04-12
1	2016-04-13
2	2016-04-14
3	2016-04-15
4	2016-04-16

(Daily activity dataframe)

3.2.3 Average Heart Rate per Minute Dataframe

To obtain the average heart rate per minute, I first imported the csv file heartrate_seconds_merged.csv to gather the data necessary for this. Since the heart rate data is counted every five seconds and needs to be converted to minutes, I grouped the data in the heartratedata DataFrame by the 'Id' column

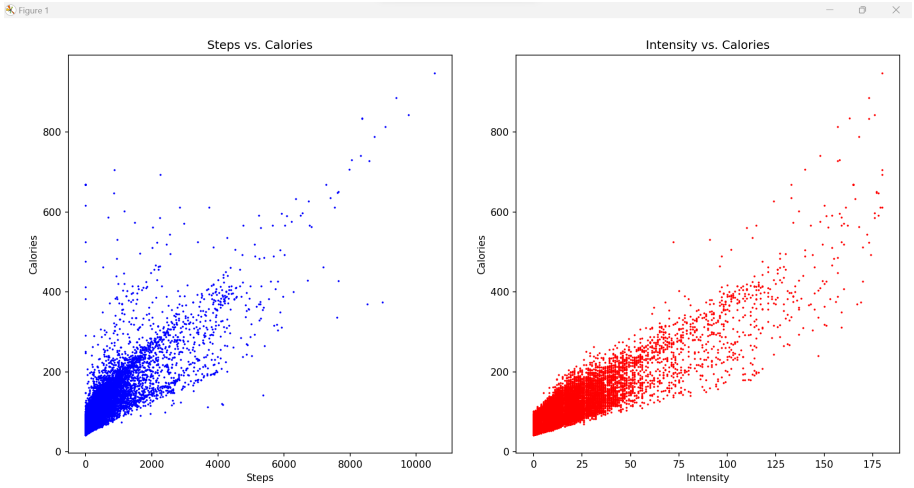
and then resampled the 'Value' column at one-minute (1 minute) intervals ('T'). This means it will calculate the mean heart rate value for each minute for each unique 'Id'. I reset the index in order to have it at the proper index at the end.

	Id	Time	Value
0	2022484408	2016-04-12 07:21:00	101.600000
1	2022484408	2016-04-12 07:22:00	87.888889
2	2022484408	2016-04-12 07:23:00	58.000000
3	2022484408	2016-04-12 07:24:00	58.000000
4	2022484408	2016-04-12 07:25:00	56.777778

(Average heart rate dataframe)

3.3 Correlation and Plots

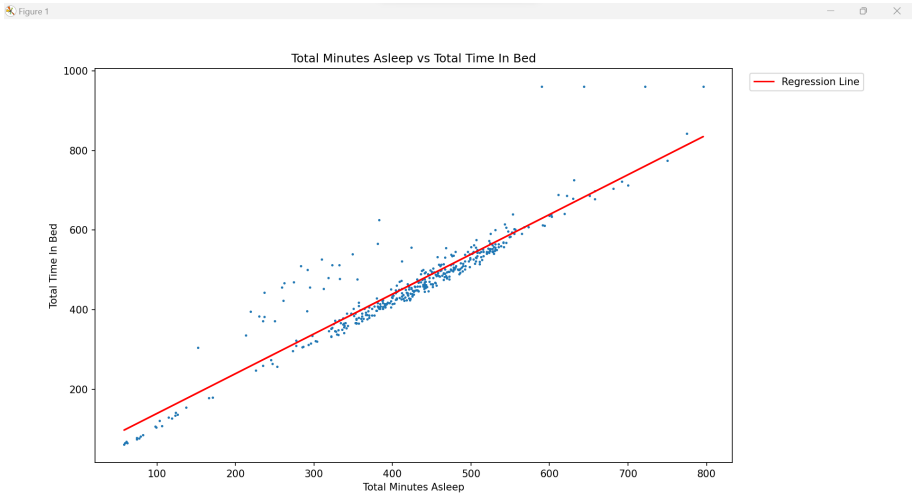
3.3.1 Relationships of Steps vs. Calories and Intensity vs. Calories



(Scatterplot comparison of Steps vs. Calories and Intensity vs. Calories)

The more relevant plot is the Intensity vs. Calories one. Since the datapoints are more condensed and less spread out as opposed to the scatterplot of Steps vs. Calories, the intensity plot is more relevant.

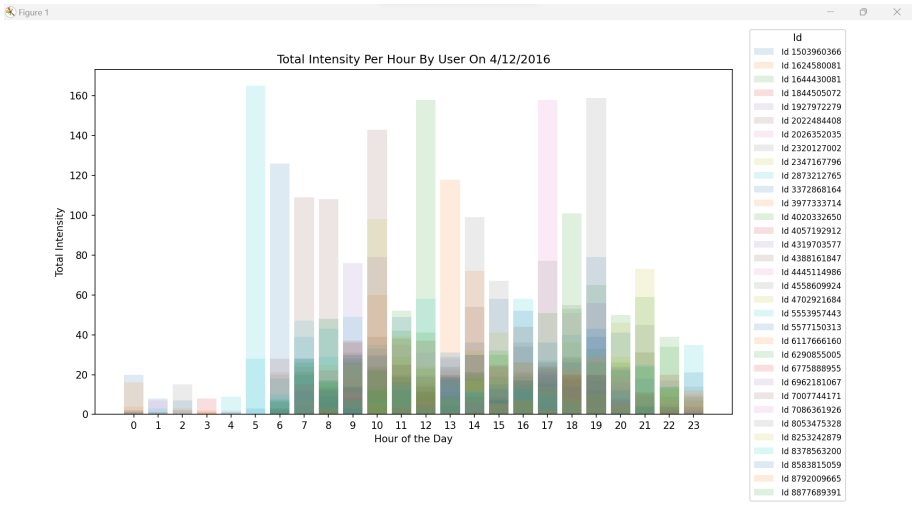
3.3.2 Scatterplot between TotalMinutesAsleep and TotalTimeInBed



(Scatterplot comparison of TotalMinutesAsleep vs. TotalTimeInBed)

I created a regression line within the scatterplot and it seems to show a correlation between the 'TotalMinutesAsleep' and 'TotalTimeInBed'. The more time you spend in bed, the more you sleep and vice versa. The points seems to be centered around the middle in between 400 minutes and 500 minutes which is around 7-8 hours. The few points that are away from the line of regression could have the following explanations. For example, if a user was sick, they would stay in bed to rest and not sleep. Another possibility is that a user was scrolling on their phone in bed for a substantial amount of time.

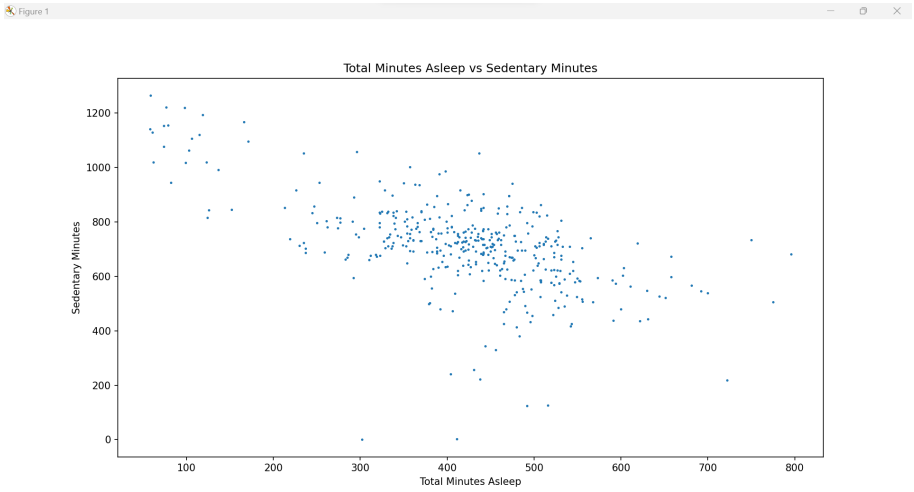
3.3.3 Distribution of Intensity On 4/12/2016 by Users



(Bar graph of the Distribution of Intensity On 4/12/2016 by Users)

Users showcase distinct variations in their intensity levels throughout the day. These variations are evident in the bar graphs, where each hour is represented by different shades of colors to reflect the varying total intensity. Notably, while users' intensity levels differ across hours, there is a common trend observed: intensity generally starts to rise from the sixth hour of the day, peaks at the eighteenth hour, and subsequently decreases from there.

3.3.4 Scatterplot between TotalMinutesAsleep and SedentaryMinutes



(Scatterplot of TotalMinutesAsleep vs. SedentaryMinutes)

There seems to be a negative correlation between Sedentary Minutes and Total Minutes Asleep. As Sedentary Minutes decrease, the Total Minutes Asleep increases. I think this was caused by the amount of activity level. The less time you spend staying still (inactive), the more you sleep. Since you are more active, you get more tired, which means you need more sleep.

3.4 T-Tests

For this T-Test hypothesis, I gathered two variables from the dataframe 'dailyActivity_merged.csv' which were the columns 'LightlyActiveMinutes' and 'SedentaryMinutes', where I believe that 'LightlyActiveMinutes' has the most significant difference for 'Calories' and 'SedentaryMinutes' has the least significant difference for 'Calories'. I will be testing them on an alpha value of 0.05.

H0 (for 'LightlyActiveMinutes') = There is no significant difference between the means of 'Calories' for the two groups.

H1 (for 'LightlyActiveMinutes') = There is a significant difference between the means of 'Calories' for the two groups.

I split the data of 'LightlyActiveMinutes' into two separate groups, one where 'LightlyActiveMinutes' is greater than the mean of 'LightlyActiveMinutes' and one where 'LightlyActiveMinutes' is less than or equal to the mean of 'LightlyActiveMinutes'. I used the scipy method (stats.ttest_ind) to compare the 'Calories' column of these two groups since they are independent of each other in which I got a tstat value of 5.7406402066972575 and a p value of 1.2730479146564296e-08. Since the p value is less than the alpha value of 0.05, we reject the null hypothesis and conclude that there is a significant difference between the means of the two 'LightlyActiveMinutes' groups. That means, in this context, the amount of calories burned is vastly different between these two groups.

H0 (for 'SedentaryMinutes') = There is no significant difference between the means of 'Calories' for the two groups.

H1 (for 'SedentaryMinutes') = There is a significant difference between the means of 'Calories' for the two groups.

I split the data of 'SedentaryMinutes' into two separate groups, one where 'SedentaryMinutes' is greater than the mean of 'SedentaryMinutes' and one where 'SedentaryMinutes' is less than or equal to the mean of 'SedentaryMinutes'. I used the scipy method (stats.ttest_ind) to compare the 'Calories' column of these two groups since they are independent of each other in which I got a tstat value of -1.8802978707120994 and a p value of 0.06037716807778881. Since the p value is greater than the alpha value of 0.05, we cannot reject the null hypothesis and we don't have enough evidence to conclude that there is a significant difference between the means of the two 'SedentaryMinutes' groups. That means that the calories burned in these two groups is not much different below the mean and above the mean.

4 Results

In this data analysis of the Fitbit Fitness Tracker Data, I explored aspects of the dataset to gain insights into the relationships between exercise, sleep, and energy expenditure. Here are some of the key findings:

4.1 Heart Rate Analysis for Different Users

I analyzed the daily average heart rates of seven selected users over a one-month period. This analysis revealed significant variations in heart rate patterns among different users. The visualization allowed us to track daily changes in heart rate, which could indicate things relating to physical activity, stress levels, or overall health.

4.2 Daily Sleep Duration Analysis

I examined the daily sleep duration of users over the same one-month period. The scatter plot displayed individual users' daily sleep patterns, with outliers identified as users deviating significantly from the average. This analysis suggested that the majority of users tended to have consistent sleep durations, but some experienced more variable sleep patterns.

4.3 Analyzing Daily Steps

I investigated the distribution of daily steps recorded by users. The scatter plot of daily step counts over the same one-month period displayed findings in physical activity. Outliers in this context represented users with exceptionally high or low daily step counts compared to the majority. This analysis provided insights into users' activity levels and highlighted potential areas for improvement in personal fitness routines.

4.4 Analyzing Weight Change for the User with the Most Records

I focused on the weight change of a singular user with the highest number of weight records in the dataset. By visualizing their weight measurements over a month, one could observe their weight fluctuation, which may be associated with dietary and lifestyle choices. This analysis emphasized the importance of tracking weight for personal health management.

4.5 Relationships of Steps vs. Calories and Intensity vs. Calories

I explored the relationships between daily steps, intensity, and calorie expenditure. The scatterplots displayed the connections between these variables, suggesting that intensity might have a more significant influence on calorie expenditure than daily step count. This finding could be valuable for individuals aiming to optimize their exercise routines for weight management and overall fitness.

4.6 Scatterplot between TotalMinutesAsleep and TotalTimeInBed

I visualized the relationship between the total minutes asleep and the total time in bed for users. The scatterplot, accompanied by a regression line, indicated a positive correlation, suggesting that more time spent in bed is associated with longer sleep duration. This finding may offer insights into sleep in general and the importance of creating a good sleep environment.

4.7 Distribution of Intensity On 4/12/2016 by Users

I examined the distribution of intensity levels throughout the day for various users. The bar graphs showed distinct variations in intensity levels, with a common trend of increasing intensity during the day, peaking around the eighteenth hour. This analysis can help users identify their most active periods and plan their workouts accordingly.

4.8 Scatterplot between TotalMinutesAsleep and SedentaryMinutes

I explored the relationship between total minutes asleep and sedentary minutes. The scatterplot revealed a negative correlation, suggesting that as sedentary minutes decrease, total minutes asleep tend to increase. This finding highlights the potential impact of physical activity and reduced sedentary behavior on sleep duration.

4.9 T-Tests

I performed t-tests to assess the significance of differences in calorie expenditure between two groups: one with higher 'LightlyActiveMinutes' and another with

lower 'LightlyActiveMinutes'. The results indicated a significant difference in calorie expenditure, emphasizing the role of light physical activity in burning calories. However, the t-test for 'SedentaryMinutes' showed no significant difference in calorie expenditure, suggesting that calories burned are consistent regardless of sedentary behavior.

4.10

In summary, this data analysis provides insights into the relationships between exercise, sleep, and energy expenditure. These findings can inform individuals about optimizing their health and wellness strategies, such as tailoring their physical activity levels and sleep patterns to achieve their fitness and well-being goals.

5 Conclusion:

In this data analysis of the Fitbit Fitness Tracker Data, there were many insights that were revealed regarding exercise, sleep patterns, and energy expenditure. These findings showcased impact of physical activity and sleep on personal well-being. The significance of these results lies in their potential to inform and guide individuals in their health and fitness goals. People are encouraged to find physical activity in order to experience positive long term effects on self.

6 References

Möbius. November 2022. *FitBit Fitness Tracker Data*.
https://www.kaggle.com/datasets/arashnic/fitbit/data