FitBut Fitness Tracker Report

Pranay Goel

October 2023

1 Summary

- In this project, I will use the data set, FitBit Fitness Tracker Data provided by Kaggle, to parse, transform, explore, and analyze this dataset to gain insight into the relationships between physical activity, heart rate and amount of sleep.
- I created diagrams to find correlations, distributions, and outliers between different variables in this data set.
- I use Python Pandas, NumPy, Plotnine, and Scipy libraries in this project.

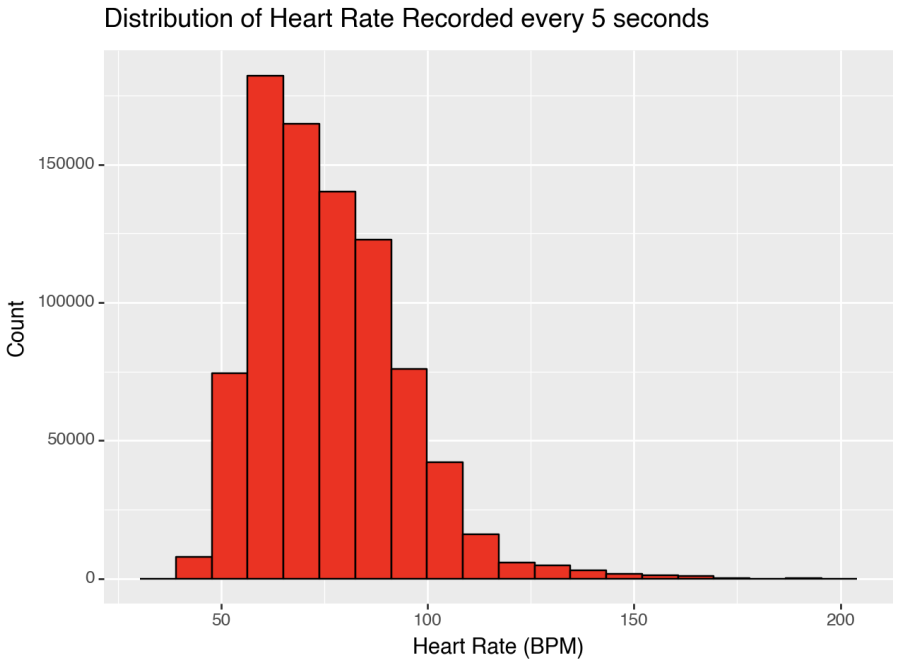
2 Introduction

My goal in this project is to analyze different variables in the different datasets provided by the Fitbit fitness tracker data to find correlations between independent values and formulate and test hypotheses about the data. I start with distributions and outliers and find the distributions and outliers of recorded heart rate, daily sleep duration, and daily steps. I created some scatter plots to find the correlation between different variables including Total Minutes Asleep to Total Time in Bed and Total Minutes Asleep to Sedentary Minutes..

3 Methodology

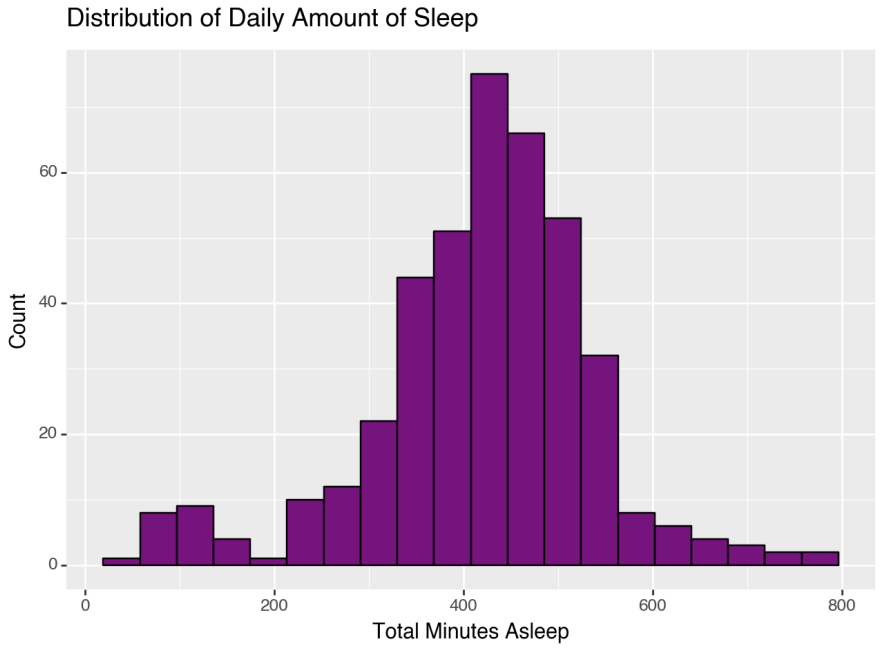
3.1 Distributions and Outliers

A. I first want to plot the distribution of different users' heart rates recorded every 5 seconds for one month. Since there are 2483658 rows in this dataset, I decided to measure only 5 different users' heart rates. This brings the dataset down to 844486 rows which is much more manageable. I plot a histogram with a consistent 20 bins for every variable.



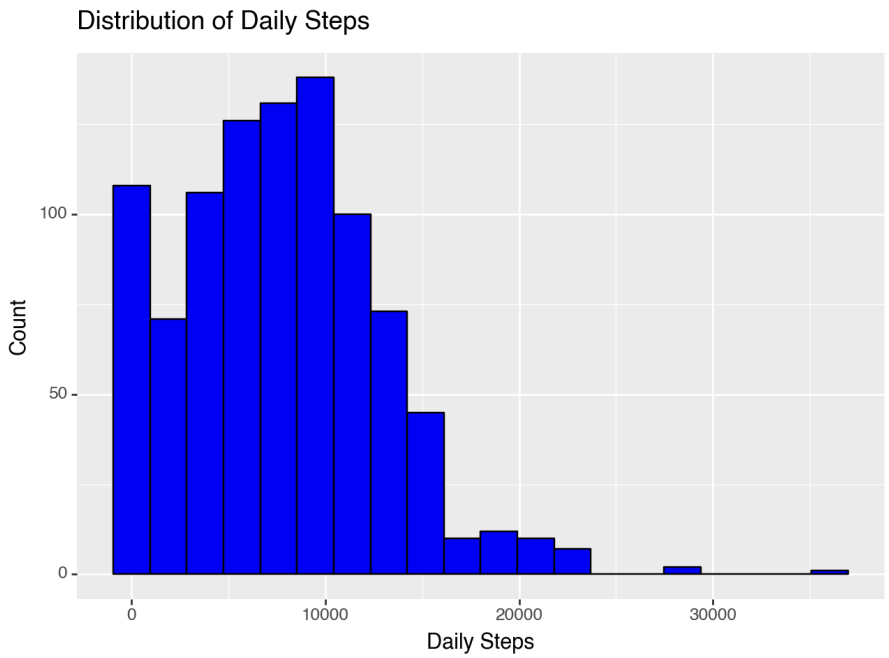
Looking at the plot, we can see a slightly skewed right distribution with the mean being around 76 bpm. This skewness is caused by several outliers in the higher bpm range. This makes sense in the data because people will have short spurts where their heart rates will increase rapidly due to many factors such as exercise or a sudden mood change.

B. Now I find the distribution for daily sleep duration. This data was in another CSV file which I read into a data frame. There are 24 users' total minutes asleep each night recorded in the data frame.



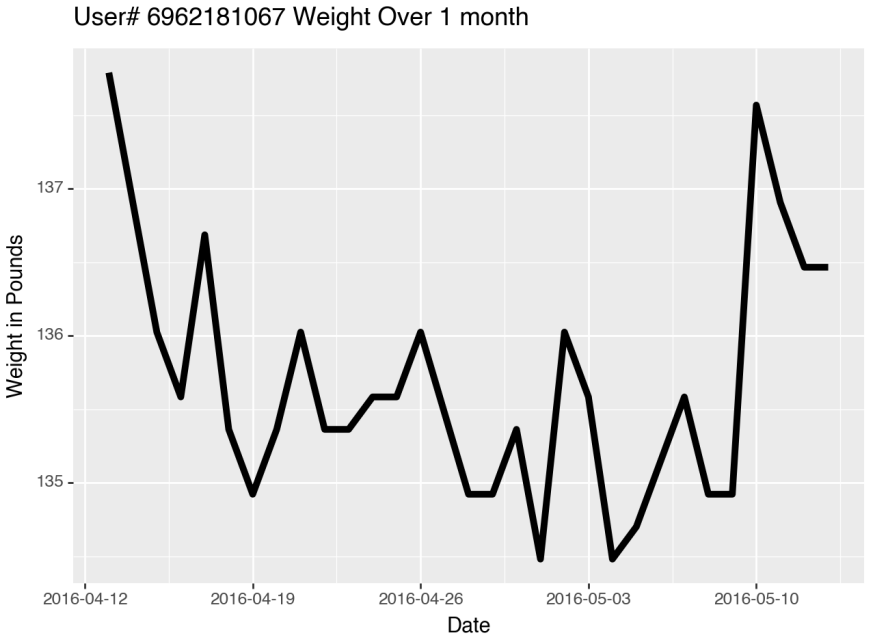
Analyzing this plot we can see that is essentially a normal distribution with a few outliers in both less and above the mean. We approximate the mean of total minutes asleep to be around 425 minutes of sleeping per night or just over 7 hours. There are many reasons there could be outliers in this data from sleeping longer on weekends to different peoples' personal preferences in the amount of hours of sleep they need so we can disregard the outliers.

C. I move onto daily steps distribution from 33 different users whose Fitbit recorded their number of steps per day.



This plot is significantly skewed to the right with outliers very far to the left. This type of distribution could be for many reasons such as people recorded in this data set have a more sedentary lifestyle and have more inactive days than active ones. It can also be caused by a few users being overly active and pulling the distribution to the right. The mean of this data set is bout 7600 steps.

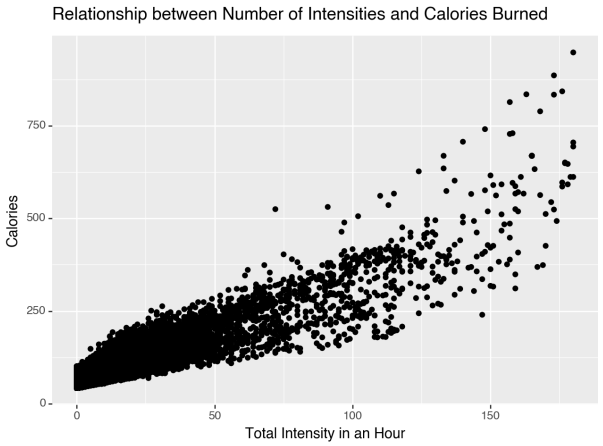
D. The last distribution I found was the weight change of the user with the most amount of weight entries. This data set only included 8 different users with many of them recording 1-2 days. There is only one user who recorded the entire month, so I plot their weight change throughout the month.



This is a fluctuating type of distribution with no clear trend. This makes sense because we only looked at a singular case and weight loss has many factors such as diet, exercise, hydration, and more making a trend impossible to find.

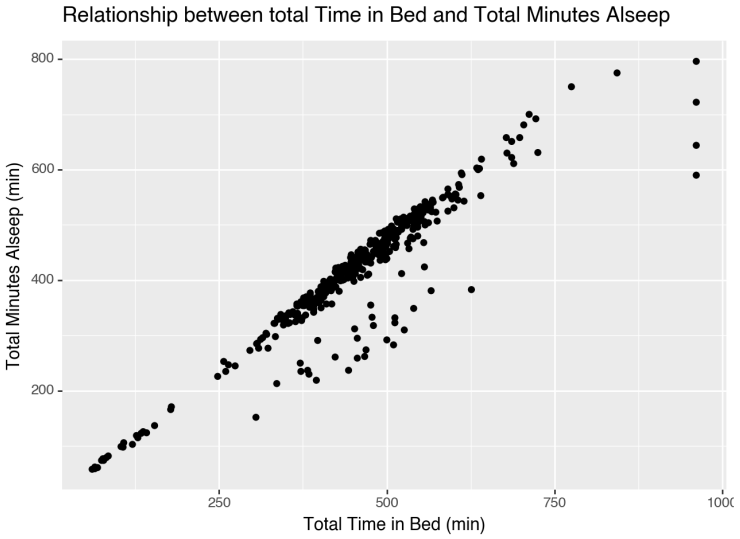
3.2 Correlation and Plots

1. I started looking at correlations between different variables. I start with two plots that show the correlation between steps and calories and intensity and calories. I plotted a scatter plot between these variables to see the difference in their correlation and find which one is more relevant and stronger.



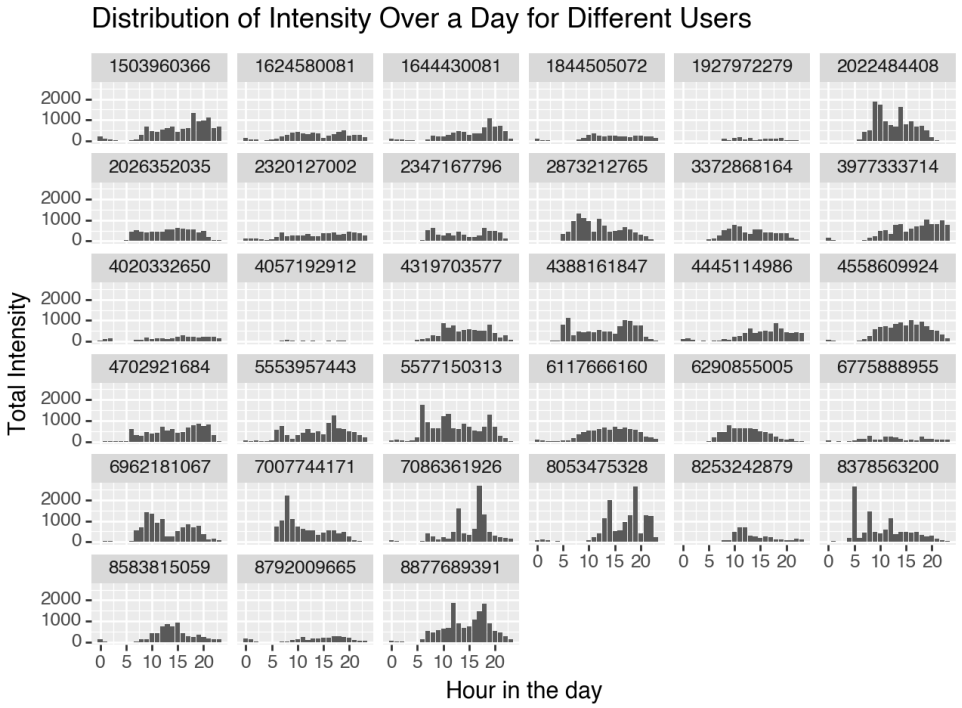
Analyzing these plots we can see that both graphs have a positive correlation which makes sense since the more steps one takes and the more intensities there are per hour will increase the amount of calories lost. Total intensities per hour seem to be more relevant than the total steps per hour because there is a stronger and tighter linear pattern with the data than the total steps and seems to have fewer outliers out of place.

2. I also plotted the scatter plot between Total Minutes Asleep and Total Minutes in Bed to find their correlation to see how strong these variables are



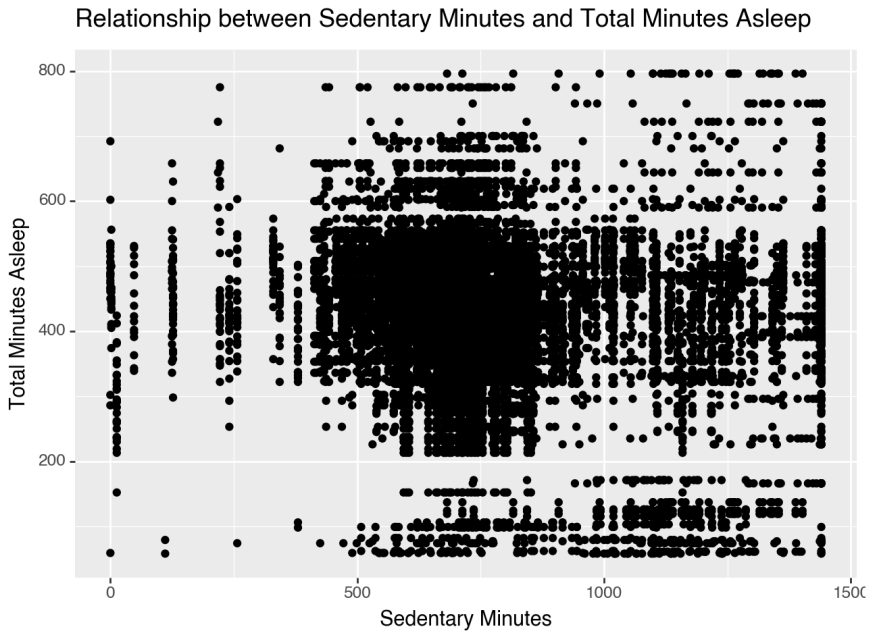
This plot indicates a very strong correlation since almost all of the points are exactly 1:1. This makes sense as most people who are in bed are asleep. There are a few outliers that can be explained by people who take longer to fall asleep than others.

3. I then plotted bar graphs to display the distribution of intensity throughout the day. I used facet wrap to show all of the different users to see if there were any differences between the users in what time intensities peak.



In this plot, we can see every single user's total intensity in each hour of the day. Every single user has a slightly different type they peak intensity but on average it seems to be a normal distribution where the peak intensity happens during the middle of the day.

4. I plotted Total Minutes Asleep and Sedentary Minutes to find their correlation.



This graph is a clear sign of no correlation since no pattern can be easily identified in this scatter plot. This is because sedentary minutes track the minutes someone is sitting or lying down while awake, not sleeping meaning that there will be no correlation with the total minutes asleep which results in a scatter plot like this where there is no clear pattern.

4 Conclusion

In this report, I took the Fitbit Fitness Tracker Data provided by Kaggle and parsed, transformed, explored, and analyzed this dataset. I accomplished this by interpreting certain distributions and outliers, merging multiple different CSV files into one data frame to use for future plots, and creating numerous scatter plots to determine the correlation between different variables and offer insight into the reason for the correlation. All of these different plots and tests have provided me with an accurate idea of the relationship between exercise, sleep, and energy expenditure.