**Product Deliverable 2 - Midterm Report: Fitbit-X**

Adele Bloch, Sarah Cook, Kelley Liu, Wilson Tam, Daniel Hwang

**The Problem**

Wearable device companies collect millions of healthcare data points on its users every single day. One company in particular, Fitbit, makes wearable devices that track and measure users activity, including number of steps walked, heart rate, quality of sleep, steps climbed, and more.

**Problem #1:** Fitbit allows users to look at their own personal data, but only offers a limited scope of insights. Currently, Fitbit does not present a way for users to receive statistical correlations or long-term trend insights about their own data.

**Problem #2:** Fitbit automatically sets a user’s daily step count goal to 10,000 steps per day. Although users can adjust this goal, it is not realistic to try and hit the same goal every day. Some users will take rest days from exercising, and some users will tend to walk less on Monday’s, for example, simply because they have a busy schedule. It is important to have a means to adjust daily step count recommendations.

**Our Solution & User Validation**

**Solution #1:** Our solution is to provide a platform where users can upload their own Fitbit data to receive an in-depth analysis of correlations among variables and long-term trends. Our website will provide more insight than what the current Fitbit application provides. Ultimately, our website will give users a better understanding of their data, allowing them to better achieve their fitness and health goals.

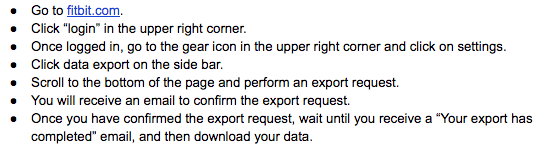
**Solution #2:** Our solution will also use machine learning to create a personalized daily step count recommendation based on user tendencies. The recommendation will take into account factors such as day of week, heart rate, amount of sleep, etc to produce an achievable step count recommendation each day.

To validate our idea, we created and sent a survey that received over 70 responses from our community. According to our research, of current Fitbit users, 60% do not feel as though they are receiving all of the information they would like from their Fitbit application. However, 86% would be interested in using a website where they can upload their Fitbit data to receive a more in-depth analysis of their data, including personalized insights and visual trends, to help them meet their fitness and health goals. We also found that on [Fitbit’s community forum](https://community.fitbit.com/t5/Feature-Suggestions/Export-RAW-data/idi-p/2703796), many users have suggested a feature that would let allow them to download their own raw data in order to find correlations on heart rate, sleep cycle, activity levels, etc. This is exactly the solution we plan to create.

**Datasets**

Our exploratory data analysis and machine learning models are built using data from specific Fitbit users, each of whom have thousands of personal data points. We obtained data from three users to begin our analysis, with the plan of extrapolating the code from our original users to any user who uploads their data to our website.

Users can download their data as follows:



**Data Analysis & EDA Results**

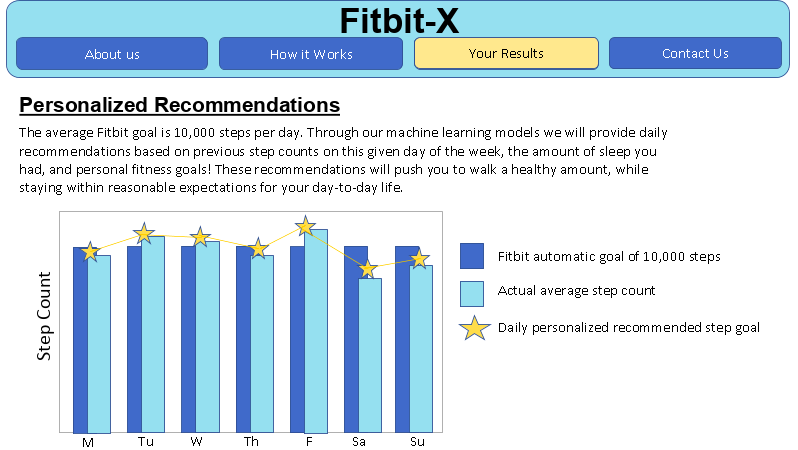
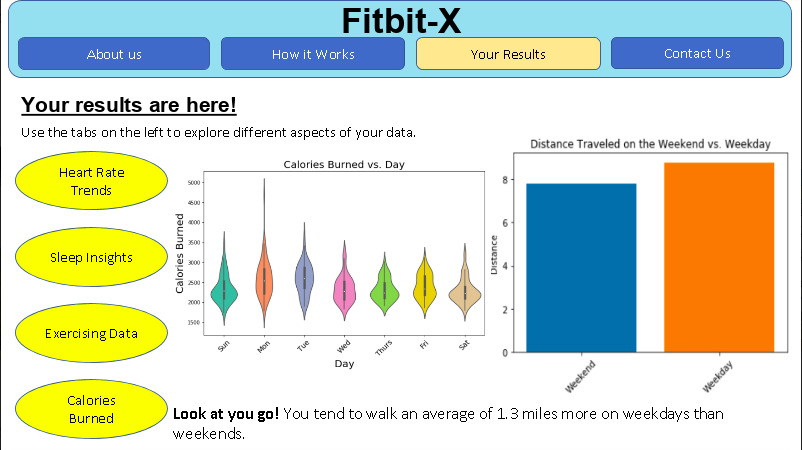
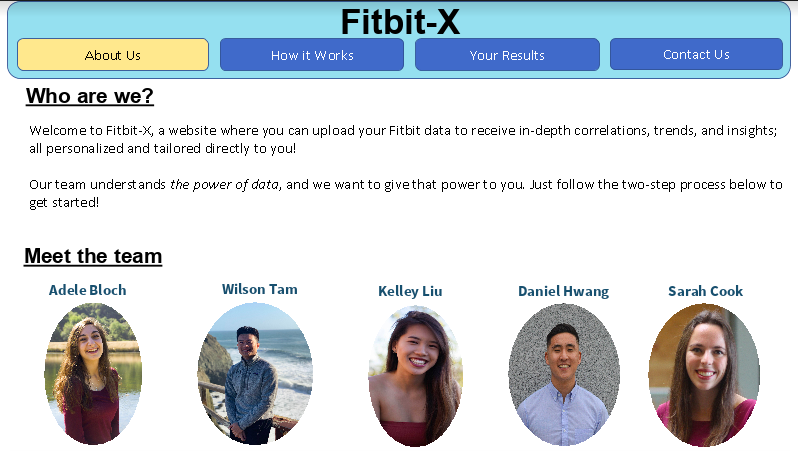
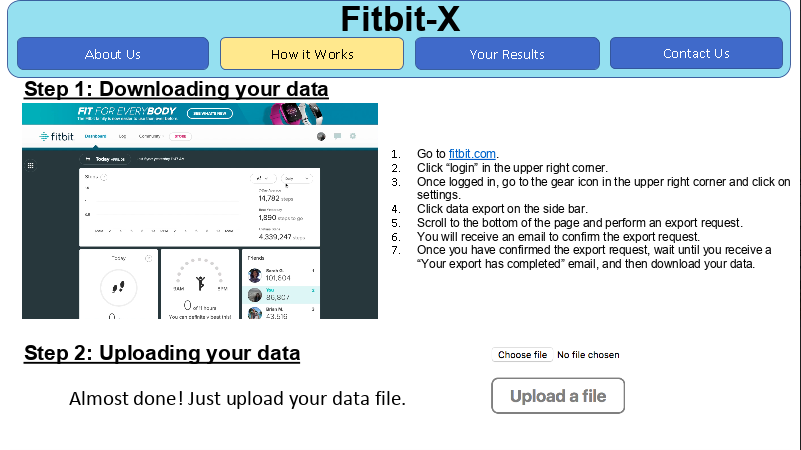
Fitbit users are able to download their data in two different ways, either as json or csv files. However, each of these downloads contain differing user data and formatting. The json files contain more granular data at an hour-by-hour level, while the csv files are less granular and contain day-by-day data. We wanted to explore and conduct exploratory data analysis on both types of files to see what insights we could extract.

The csv files were downloaded in a format that was easy to use. Because of this, we got started on EDA right away. For the csv file, we explored various functions that would allow us to see trends and correlations among different features. We grouped the data both by month and day of the week to understand how much a user traveled, burned calories, climbed floors, or was active (lightly, fairly, very) and sedentary over different periods of time.

However, the csv file did not have data on sleep or heart rate, which were two other features we wanted to explore. These features were included in the json files on a very granular level, however, the json features were formatted such that each day and measurement existed in separate files. In order to properly process the features the csv was missing, we had to figure out how to merge over 2,000 files together to extract separate dataframes for each of the features. This took a significant amount of time and code to be able to merge all of the files.

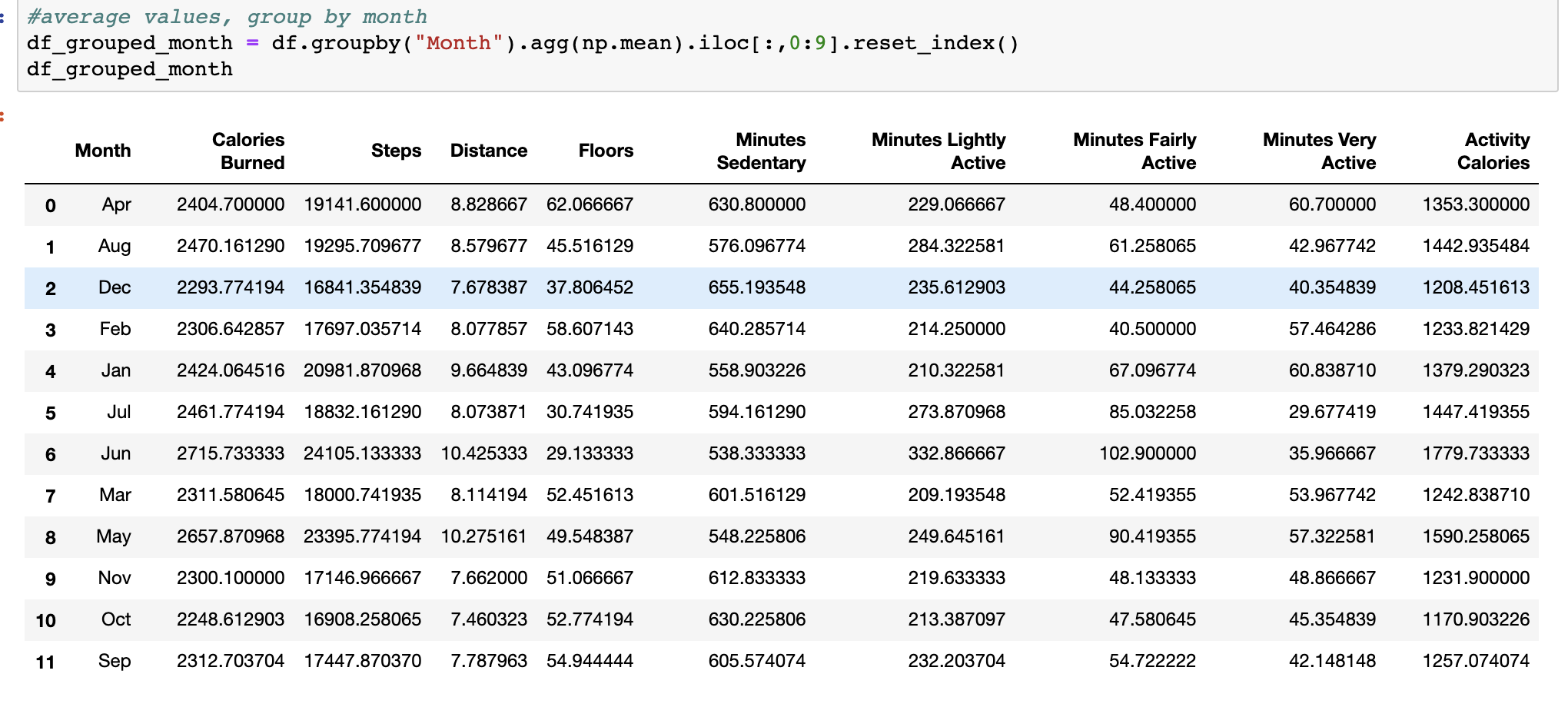
After successfully completing the json file merge, we will extract insights from the json files and combine them with the insights from the csv files. Ultimately, the json files have more granular data and additional features, so it will be valuable to be able to merge all of the insights we find.

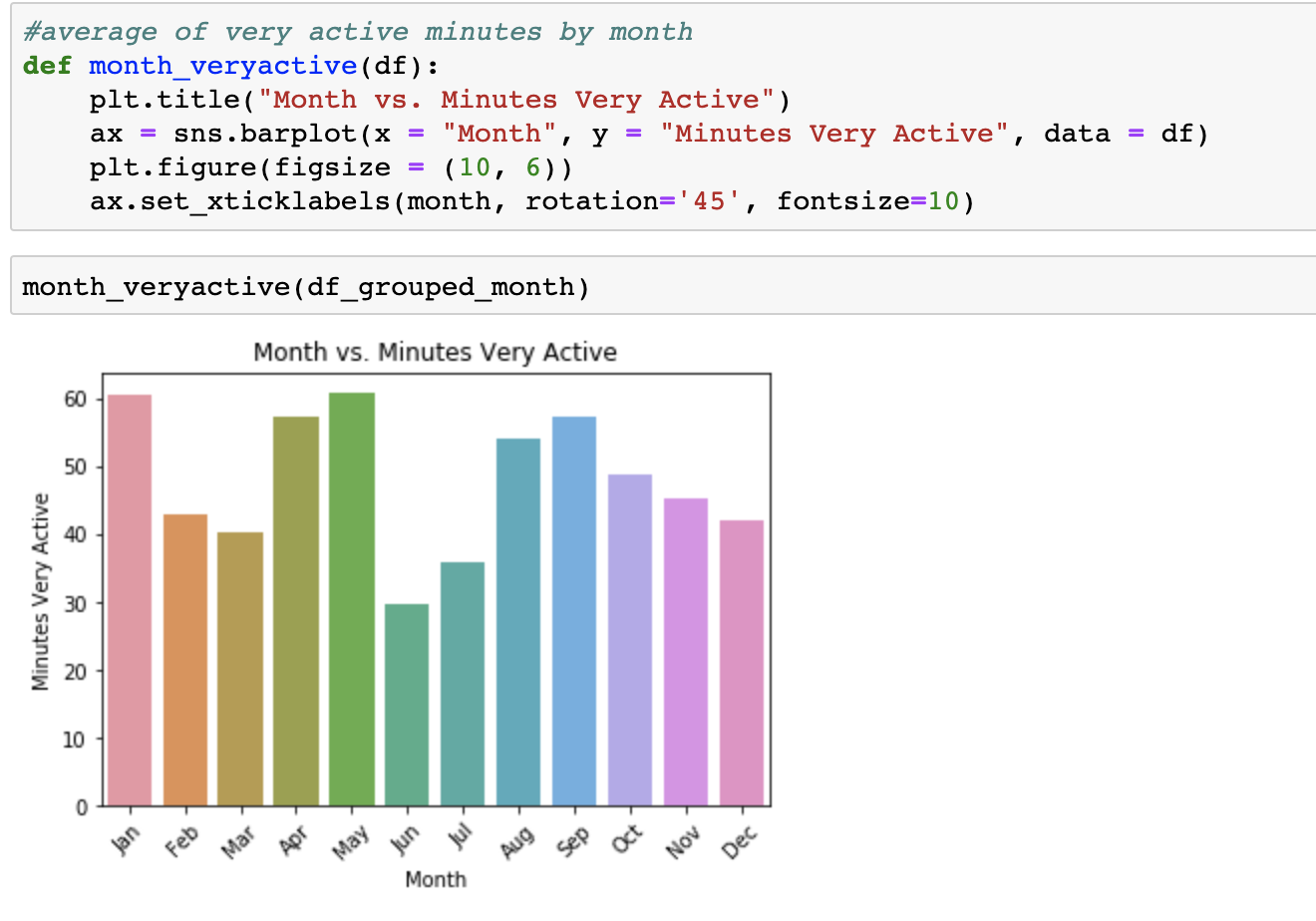
**System Architecture Overview**

Below is a model of the website we are trying to build. We needed to find a platform that would allow us to work with backend Python code. Similarly, no one on our team has experience with HTML or Javascript, so we also needed a platform that was easy to learn. We are currently considering working in either Django or Flask.****

**Code Snippets**

Below you can find some of our exploratory data analysis. The code below groups the cleaned dataset by month and averages the values for each of the features. This is then used to visualize certain features that will reveal specific trends and correlations.

This function below is used to plot the average minutes active by month for this specific user. The resulting plot reveals some interesting insights: in the summer months (June, July), the user has the least number of minutes per day spent very active, whereas January, April, May seem have a much higher average number of very active minutes each day. When these insights are shared with the user, we anticipate the user will have a better understanding of their data and daily lives.



You can find a full version of our code in the final section of this document in the link to our Github repository.

**Key Findings**

****

After cleaning the data, we plotted the following pairplot (as seen above), and included all features except month, numerical value for the day of the week, and the weekend/weekday encoding. With this, we were able to see the scatter plots of every feature plotted against one another to see which features had high or low correlation, based on the day of the week. For example, we can see that plotting the distance traveled vs. steps, calories burned vs. activity calories both have a linear positive correlation. On the other hand, minutes sedentary vs. minutes lightly active or distance vs. minutes lightly active do not seem to have much correlation.

Another thing we noticed in our data when comparing averaged values for weekdays and weekends was that weekdays tended to have higher values for users (calories burned, distance traveled, minutes very active). Ultimately, we plan to generate many of these exploratory data analysis results and insights for our users to better understand their own data.

**Bottlenecks & Challenges**

One challenge we are facing is the limitation of available Fitbit data. The only data we have access to is data that our friends are willing to give to us, so that may make it difficult for us to build comprehensive models that can account for every edge case. Additionally, different tiers of Fitbit products provide different data points. For example, one type of Fitbit does not provide any information about flights of stairs climbed, and some can’t track sleep as well as others. This may make it harder for us to create a product that can work for any Fitbit user.

Another challenge that our team has faced and discussed throughout this semester is the idea of building a website from scratch with backend Python code. None of us have web development experience, so this has always been a daunting task. However, we are going to start exploring Flask platforms in order to achieve a functional prototype where a user may upload their data and the EDA and recommendations will be displayed.

Finally, a bottleneck we have faced in our work so far is our ability to analyze the json files, which hold the most granular and comprehensive data. The json files are separated by day and data method (ie: heart rate, sleep, etc.), and to be able to work with them, we had to merge together over 2,000 data files per user. This requires a sufficient amount of time and space. However, we were able to figure this bottleneck out, it just took more time than we originally anticipated to be able to analyze all of the files.

Ultimately, we will let these challenges fuel us into our next steps!

**Performance Evaluation**

We are highly satisfied with the current EDA we have conducted. We were able validate our idea, and start to visualize and determine all the insights we will be able to provide our users. From looking at correlations between different features, discussing ideas from visualized distributions, and analyzing trends, we found the plethora of potential insights both exciting and fascinating. Our whole group truly believes that if we could provide this data to our users, users would have a better understanding of their own lives and health goals.

We were also proud of recovering from our first bottleneck, which was the format with which the data was downloadable. Because we are working with another company’s data, we were restricted by the data that is available to us, as well as the format that the data came in. However, we were successfully able to merge together over 2,000 json data files to provide a more comprehensive and in-depth analysis for our users, and successfully able to perform EDA on the csv files available to us too.

But with all these potential ideas, we realized that there is much work to be done to create a system that operates to a satisfactory level and encompasses what we ultimately envision providing to user. We have outlined 4 tangible next steps in the following section in order to create such a system.

**Next Steps**

1. Website Creation

Our team is excited to tackle is to create a basic user interface and working website, so the user may directly upload their Fitbit data and have personalized insights and trends on their screen within seconds (Reference section 5). Our team members are exploring Flask based websites in order to create a platform to upload, process, and display data. Currently, we have extremely basic outlines of a website through the Flask interface, but will be further refined as we near the final deadline.

2. Json and Csv Processor

We want our website to be able to process two different types of input Fitbit data. The Fitbit website allows a user to either download a csv file that has one month’s worth of data or a user may download a zipped file of json documents that have data down to the second for every day that the user has worn his or her Fitbit. This vast difference in input data type will be challenging to work with effectively. We are planning on having similar, but slightly modified, functions in order to make sure both cases are accounted for throughout the EDA.

3. ML Models

Our group has decided that we should prioritize providing non-ML related insights to our users, because we determined from our EDA that there is a lot of valuable information can be determined from our analysis (Reference sections 8 & 9). However, we still intend to incorporate machine learning in our project.

Currently, Fitbit automatically sets all users to a daily goal step count of 10,000 steps. However, [many Fitbit users](https://community.fitbit.com/t5/Feature-Suggestions/Ability-to-change-goals-per-day-without-changing-all-past-days/idi-p/72817?collapse_discussion=true&filter=location&location=idea-board:features&q=personalized%20step%20count%20goal&search_type=thread) wish they could vary their personal goal recommendations, since for example, there are some days that people are more likely to take a rest day or go on a hike. Our website will use a machine learning model to create personalized daily step count recommendations on a daily, weekly, and monthly basis for users, most likely by using linear regression. We plan to use this to inform users on how where they can increase their steps in order to reach their inputted fitness goals.

Another ML aspect we considered is to include a fitness level classifier. We will use CART to classify each user into one of X fitness levels. From there, we will provide personalized insights respective to each fitness level.

4. Additional Useful Features (i.e. Health reminder alerts)

If we have time, our group has also determined that it would be useful to have coded-in alerts for our users. One example would be if a user’s resting heart rate is comparatively high on certain days, our system would send a message with recommendations on how to lower resting heart rates. We realized there could be a lot of such recommendations, and plan to determine the most useful ones and add them along with other potential useful features.

**Link to Github Repository**

<https://github.com/sarahcook97/Fitbit-X>