Fitbit Activity Dashboard Development

Kyle Lam

1. Project Goal —

- The goal of this project is to analyze Fitbit users' health and activity metrics to create visualizations that explore relationships between physical activity, heart rate, weight, and energy expenditure across multiple time intervals.
- These visualizations aim to reveal actionable insights that support maintaining and improving human health.

2. Data Sources —

- This project uses data from two folders provided in the Fitbit Kaggle dataset:
    - Fitabase Data 3.12.16-4.11.16
    - Fitabase Data 4.12.16-5.12.16
- Each folder contains a collection of second-level, minute-level, hourly, and daily data for various health and activity metrics collected from consenting Fitbit users during each respective time period
    - Metrics included: activity, calories, intensities, steps, heart rate, METs (metabolic equivalent of task), sleep, sleep day, and weight logs.
- Source: Kaggle – Fitbit Fitness Tracker Data

3. Tools Used —

- Python: Used pandas for data cleaning, preprocessing, and initial exploration.
- Apple Numbers: Used for quick visual data inspection and note-taking in exploration.
- Tableau Used to create interactive dashboard visualizations using processed datasets.

4. Initial Observations, Planning, and Unknowns —

- I created master datasets corresponding to the fitbit datasets for 3/12/16 - 4/11/16 and 4/12/16 - 5/12/16 that populate them
- I found unique user id counts range from 15 to 35 across datasets despite Kaggle dataset content description stating "Thirty eligible Fitbit users consented to the submission of personal tracker data…"
    - <=30 makes sense for those who did not enable specific tracking features or did not want to submit, but 33 and 35 values are not consistent with the description
- Initial plan for data sources after extracting csv in Tableau is to join by id and time intervals: daily, hourly, minute, seconds
    - Then join all datasets by id
    - Only heartrate_seconds is tracked in seconds
        - Meaningful metric must be determined before joining and making a graph

                - Average such as bpm for each day/hour period could be used as a useful aggregate
- Fitbit dataset differences between the two monthly periods
    - Three “minute…Wide” datasets are only available in 4/12 - 5/12 datasets folder
        - Equivalent to corresponding narrow csv datasets in the same folder, but missing the 4/12 data
        - Can join with hourly csv datasets due to ActivityHour column, but with have missing data from 3/12 - 4/12 (first day missing)
    - dailyCalories, dailyIntensities, dailySteps, heartrate_seconds are only available in 4/12 - 5/12 datasets folder
- minuteSleep Date values do not start at the same time on the first day for each id
    - At least a few differ from consistent 12:00 am start value for ActivityMinute
    - Tracked in minute intervals, but some intervals of values start at “[hour]:30” rather than beginning of each minute “[hour]:00”
    - Minute intervals in Date are discontinuous, which means n/a values for several ActivityMinute rows if joined
        - Change between minute intervals starting at “[hour]:00” to “[hour]:30” or vice versa appear to indicate a new continuous interval discontinuous to the one it follows
        - Can omit seconds to join with csv datasets with ActivityMinute
            - Discontinuous intervals means no potential issue with duplicates (e.g. 12:00 and 12:30 if seconds are omitted)
    - Unknowns
        - Value column: 1 vs 2
        - Log id column changes after continuous interval of same value for same user id
- minuteMETs (metabolic equivalent of task)
    - Generated min and max MET value across users
    - Min and max of 0 and 189 collectively; mode of 10; thirteen 0 values
        - 1865956 MET values equal the mode of 10
        - Since baseline value of 10 is most frequent, the MET values give a possible and much more reasonable distribution of values if they are all divided by 10 since resting metabolism is 1
            - The upper range MET value becomes 18.9, which is unusually intense, but possible

5. Decisions Made —

- In progress

6. Challenges and Fixes —

- In progress

7. Limitations and Assumptions —

Assumptions:

- User IDs are unique and represent distinct individuals.
- Dates and timestamps are correctly standardized across datasets.
- Fitbit's estimates for METs, intensities, and calories burned are sufficiently accurate for trend visualizations and analysis.
- Incomplete data such as users not enabling certain tracking features is assumed to be random and not systematically biased.
- minuteSleep data lacks synchronization across IDs, with some logs starting mid-hour rather than at the beginning of the hour.
- The meaning of some variables (e.g., Value column in minuteSleep dataset) remains unclear without Fitbit's internal documentation.

Limitations:

- Strength-based activities may not be accurately represented since Fitbit relies on heart rate and motion detection, which is not necessarily conducive to estimating energy expenditure for such activities
- The number of unique user IDs exceeds the expected 30 in some datasets, which indicates data inconsistency or an error in the dataset collection description
- Some datasets are not available for the earlier time range (3/12–4/11)

8. **Next Steps (Possible Additional Features) —**

- In progress