## Data Sources

1. FitBit Fitness Tracker Data <https://www.kaggle.com/datasets/arashnic/fitbit>

A dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID (column A) or timestamp (column B). Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.

## Critical information about data sources

### Where are the data stored ?

For now(24Jun2023), the data is stored in a subfolder of the project folder as both these 2 datasets do not take up a lot of space in the sense of modern computers/

### How is the data organized?

*Fibit dataset*

* The fitbit dataset is organized into multiple CSV files. Each representing a portion of the data(also a view of the data in specific time frames). Most of the data in the dataset is presented as long data. Only data within minute time frame is organized as both long and wide data alternatives.
* The fitbit dataset records 4 broad categories of data : activity, heart rate, sleep and weight data.
* There are 5 activity data : steps, distance travelled, time of movement, calories burned, and intensity of movement(METs). All activity data has atleast 3 views, i.e. daily, hourly, minutely, each in a different CSV file.
* Heart rates are recorded as per minute time frame.
* Weight and Sleep data are recorded as per day time frame.

**Dataset credibility assessment (ROCCC)**

|  |  |  |
| --- | --- | --- |
|  | Assessment | Score |
| Reliable | This dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. The datasource Amazon seem to be rather reputable and reliable. | 4.5 |
| Original | This is a secondary data | 4 |
| Comprehensive | This dataset contains data about steps,distance,time, calories, heart rate, sleep, weight, with multiple views. Which is pretty comprehensive to know the usage condition of sampled consumers. | 4 |
| Current | The time frame for this data is 03.12.2016-05.12.2016. Which is 7 years ago, so the effectiveness of this dataset is not that good compared to the other. | 3.5 |
| Cited | Yes | 4 |

## About licensing, privacy, security, and accessibility

* The license for both of these datasets are CC0 1.0, which means anyone can have the right to copy, modify, distribute and perform work on the data, even for commercial use. So there will present no issues regarding licensing using these datasets.
* As for the privacy, both these datasets do not contain any personal information that would lead to the identification of sample individuals. However, a full attention about data privacy is to be paid if there will be other dataset(s) included into the project.
* As for the security and accessibility, with regard to data ethics, also to show case my work in my portfolio, I will make all related materials public, including documentations, code, cleaned data, data visualizations, report. So, the main concern regarding data security for this project is to prevent any unauthorized access to the intermediate materials, which could potentially harm my analysis process. I will be using encryption features of document editing softwares to ensure data security during analysis process.

## Data integrity verification

*Methodology*

* accuracy
  + count outliers, null values, NA values.
  + look at the decimal precision
* completeness
  + brainstorm and list out all possible data columns that we might need in the project and see how much do the datasets' content overlap with the list.
* consistency
  + check for consistency between columns that should have arithmetic relation. (for instance, steps in an hour should be the sum of all minutely steps within that hour)
  + check for data format consistency
* trustworthiness
  + look at the statistical values of data and see if there present any phenomena that totally does not make sense.

**Existing problems & limitations with the data**

1. Sample size : sample size of 30 individuals, minimally sufficient but not significant in any meaningful way.
2. Time period : The data is about the time period between 03.12.2016-05.12.2016, users’ behavior might have changed.
3. Third party data : This is a public data on Kaggle, generated by respondents to a distributed survey via Amazon Mechanical Turk.Lack of demographic information
4. Lack of demographic information : The data does not contain any direct information about the user profile.