Data Analysis Report

A. Project Overview

Project Title: Bellabeat user habit unveiled.

Objective: Provide marketing insights for strategic growth

Team Members: Daffa Arkananta

Date Started: November 26th, 2023.

Date Completed: November 28th, 2023.

B. Data Collection

a. Tools Used
   1. SQL
   2. Google BigQuery
   3. Excel
   4. Power Query
   5. Tableau

b. Data Sources
   Source 1**:** Fitbit Fitness Tracker Data
   Description: The 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. Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.
   URL or Location:
   https://www.kaggle.com/datasets/arashnic/fitbit/versions/1?resource=download
   Data Format: Multiple CSV files.

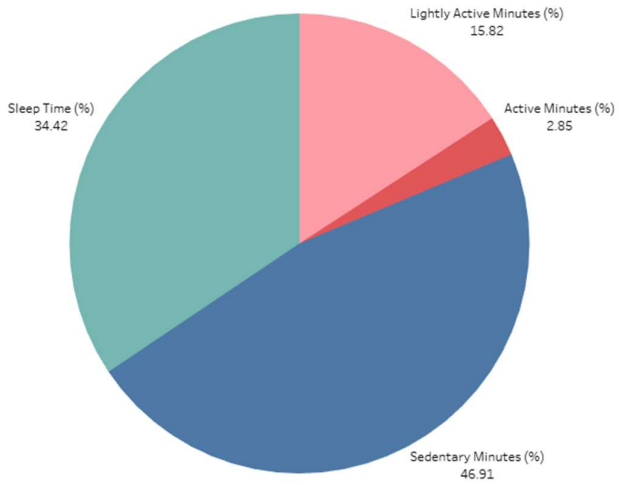
c. Data Preparation
   Missing Values: There are no missing values in the dataset.
   Duplicate Entries: There are no duplicate entries in the dataset.
   Data Transformation:
   1. Changing formats
      There was one big problem when dealing with the dataset. The date and time column were not compatible with SQL format. To make it compatible with SQL, Excel and Power Query was used to change the values from MM/DD/YYYY to DD/MM/YYYY.

   2. Merging datasets
      Multiple CSV files that have been cleaned were merged into 2 files, daily activities, and hourly activities. SQL was used in the merging process. There are 2 options for merging datasets. The first one is joining two or more datasets by using INNER JOIN or LEFT JOIN function. The second one is joining two or more datasets using LEFT JOIN function in a nested query. In this case, the

latter one was chosen because it has much less time needed to process the query, thus increasing the preparation phase efficiency. Code used in the pull process can be seen in the appendix.

C. Exploratory Data Analysis (EDA)

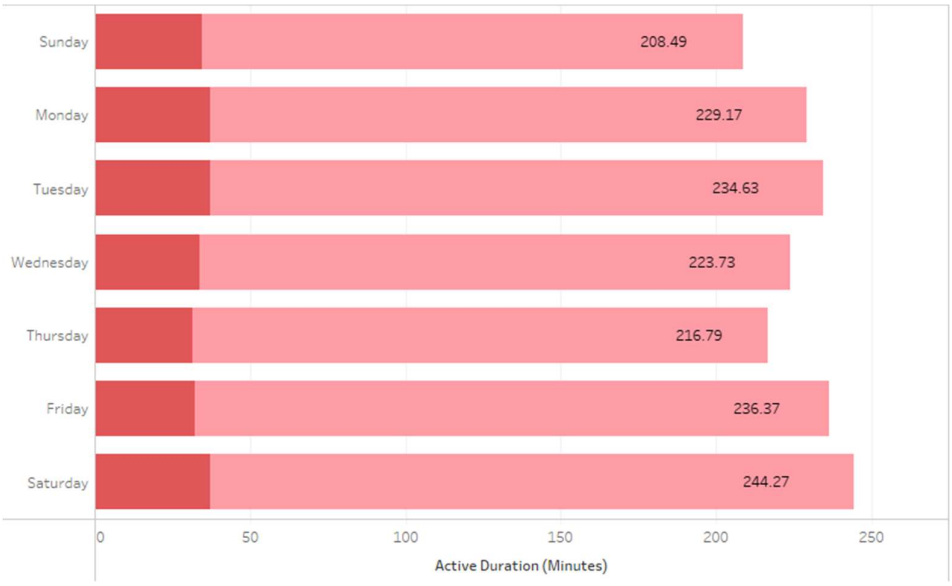
a. Data Visualization for Daily Activities



The pie chart visualization describes users daily active state by percentage. The parameter used in the visualization is the average Values for sedentary minutes, sleep time, lightly active minutes, and active minutes in one day. From the visualization, we can conclude that:

1. Users are using smart devices to track activities, such as sitting, walking, exercising, and sleeping.
2. On a daily average:
   - Users spend around 11 hours (46.91%) of their time sitting or being inactive.
   - Users spend around 4 and a half hours (18.67%) daily exercising.
   - The remaining time, 8 and a half hours (34.42%), is spent sleeping.

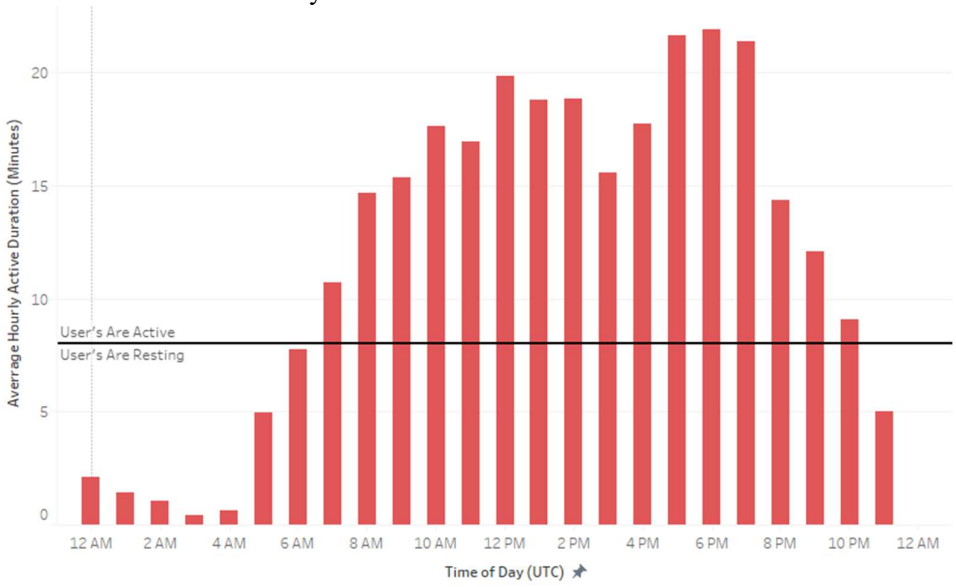
b. Data Visualization for Daily User Active Duration



The bar chart visualization points out the most active day in a week that is defined by user active duration in each day. The parameters used were days in a week and the average active and lightly active duration in each day of week. From the visualization, we can conclude that:

1. Saturday & Friday are the most active days of the week.
2. Sunday & Thursday are the least active days of the week.

c. Data Visualization for Hourly User Active Duration



The bar chart visualization gives us insight into the user's active hours in a day. The parameters used in the visual are the time of day in UTC and average hourly active and lightly active duration. From the visualization, we can conclude that:

1. 5 PM to 7 PM is when users are the most active.
2. Users are inactive or sleeping starting from 10 PM to 6 AM.

D. Conclusions

a. Key Findings

1. User's daily activities are dominated by being inactive, sitting, or staying still.
2. Saturday is the most active day while Sunday is the least active day.
3. Users tend to be more active from 5 to 7 PM and rest or sleep from 10 PM to 6 AM

b. Recommendations

1. Categorizing user's active rate from very active, active, moderately active, lightly active, and sedentary to make a more personalized app and ads from each user's device.
2. Displaying marketing ads on mobile phones at 3 PM where people are mostly on a break from work. Other options are leisure time at 9 PM.
3. Displaying marketing ads on busy streets at 6 to 8 AM and 5 to 7 PM when people are mostly moving to and/or from work with an ad's frequency increase on Friday to prepare for Saturday activities.

c. Steps to take

1. Collect more data about regions and locations to estimate the busiest location to maximize marketing ads efficiency.
2. Build an AI model based on personalized user habits to make a more personalized app and make customers feel more engaged.
3. Adding more features for the app to better understand user activities, such as running, cycling, walking, or even swimming. The input could be manual by user or automated by detecting heartrate, movements, phone's gyro sensor and GPS.

E. Appendix

a. SQL Query used in pulling the daily summary data from BigQuery

```
// SELECT
        Id,
        ActivityDate,
        TotalSteps,
        TotalDistance,
        VeryActiveMinutes + FairlyActiveMinutes +  LightlyActiveMinutes AS
      TotalActiveMinutes,
        Calories
FROM
        noaa-asfc-open-data.Capstone.daily_activity
```

b. SQL Query used in merging and pulling the hourly summary datasets from BigQuery

```
// SELECT
        Id,
        DateTime,
        Calories,
        (
          SELECT
```

```
    TotalIntensity
  FROM
    noaa-asfc-open-data.Capstone.hourly_intensities
  WHERE
    hourly_intensities.Id = hourly_calories.Id
    AND
    hourly_intensities.DateTime = hourly_calories.DateTime),
 (
  SELECT
    StepTotal
  FROM
    noaa-asfc-open-data.Capstone.hourly_steps
  WHERE
    Hourly_steps.Id = hourly_calories.Id
    AND
    hourly_steps.DateTime = hourly_calories.DateTime)
FROM
    noaa-asfc-open-data.Capstone.hourly_calories
ORDER BY
    Id      //
```