Google Data Analytics Professional Certificate

Course 8: Google Data Analytics Capstone: Complete a Case Study

Case Study 2: How Can a Wellness Technology Company Play It Smart?

Name: Carlos Pereira

Ask

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

You will produce a report with the following deliverables:

1. A clear summary of the business task
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of your analysis
5. Supporting visualizations and key findings
6. Your top high-level content recommendations based on your analysis

Report

1. A clear summary of the business task

The stakeholder for this project is Urška Sršen, Bellabeat's cofounder and Chief Creative Officer. She knows that the available data and its analysis could reveal opportunities to grow. The business task is to identify possible trends between smart device usage and Bellabeat customers to focus the next Bellabeat marketing strategy into these trends. The insights identified in this study can lead to high-level conclusions for the stakeholders regarding audience targeting, the days of the week when users are most active, users' use of smart devices, and other metrics that can influence their next marketing strategy.

2. A description of all data sources used

The dataset used in this project is from Kaggle's "FitBit Fitness Tracker Data". It contains minute-level data and daily summaries from thirty FitBit users, as indicated in the Prepare Phase, covering distance, heart rate, sleep monitoring, calories, intensity, steps, among

others. The dataset is in the public domain provided by Mobius, and is organized into two time frames (12.03.2016 to 11.04.2016 and 12.04.2016 to 12.05.2016). The data is stored in CSV files and in narrow format, except for some CSV files that are in wide format.

This dataset fits the criteria of ROCCC (Reliable, Original, Comprehensive, Current and Cited), because according to Kaggle it is complete, has credibility, contains comprehensive information about the FitBit fitness tracker, receives annual updates, and is widely cited, evidenced by its 172K downloads.

Data integrity was ensured by reviewing the data type of each column before uploading it to SQL. String, date, integer, and float types were assigned depending on the case. Also, the minimum and maximum values for each numeric column in the narrow data were reviewed by sorting and checking if the values made sense in magnitude (for example, total steps in a day were approximately 19,000). Additionally, the same process was performed with the string columns to ensure there were no null values. [1]

This data and its validation allow for addressing the business task with reliable data. The only problem identified in the data is related to the daily activity data. There are rows with no calories registered, which may indicate a data anomaly. These rows could have been a misrecording by the smart device [2]. Also, the method used to choose the thirty users is not clear, so it is not possible to identify whether this could lead to a bias in the data.

3. Documentation of any cleaning or manipulation of data

As the dataset contains CSV files with over 1 million rows, making it a large dataset, BigQuery SQL was used for data cleaning and manipulation. To ensure the data is clean and ready to analyze, initial checks were performed in the BigQuery console:

```
SELECT *
FROM `case-study-name.name_database.csvfile_name`
```

After this command, each numeric and string column in each data table was sorted using the console interface to identify possible data errors, as mentioned in [1]. Second, the date data was provided in format Month/Day/Year Hour:Minute:Second “A/P M”. Since the “A/P M” is not a standard format accepted by SQL, this data was stored as a string.

Third, as identified in [2], there were rows with 0 calories. To clean the data, the next SQL query was performed, considering only the rows with Calories > 0. A total of 7 CSV files were reviewed; 5 were identified as affected and were subsequently cleaned.

```
SELECT *
FROM `case-study-coursera-458016.Fitabase_Data_1.minuteCaloriesNarrow_2`
WHERE Calories > 0
```

4. A summary of your analysis

The “Id” column is present in all the data tables and will serve as the reference column. For the daily activity data, it is first necessary to identify how many unique values the Id column contains to understand its magnitude, using the following command:

```
SELECT DISTINCT Id
FROM `case-study-coursera-458016.Fitabase_Data_1.Daily_Activity_1_cleaned`

SELECT DISTINCT Id
FROM `case-study-coursera-458016.Fitabase_Data_1.dailyActivity_2_cleaned`
```

This results in 35 unique Id values for the first data frame and 33 for the second.

For data analysis, combining both data frames is necessary as this allows for better visualization and understanding of the data. The analysis will focus on identifying trends and important insights regarding daily activity, heart rate, weight log information, sleep data, and hourly data such as calories and steps. The following commands were used to combine both data frames. After executing these commands, the resulting data was saved as a table in BigQuery and exported as a CSV file.

Daily_activity_Total

```
SELECT  *
FROM `case-study-coursera-458016.Fitabase_Data_1.Daily_Activity_1_cleaned` AS Table_A
UNION ALL
SELECT *
FROM `case-study-coursera-458016.Fitabase_Data_1.dailyActivity_2_cleaned` AS Table_B
```

HeartRate_Seconds_Total

```
SELECT  *
FROM `case-study-coursera-458016.Fitabase_Data_1.heartrate_seconds_1` AS Table_A
UNION ALL
SELECT *
FROM `case-study-coursera-458016.Fitabase_Data_1.heartrate_seconds_2` AS Table_B
```

WeightLogInfo_Total

```
SELECT  *
FROM `case-study-coursera-458016.Fitabase_Data_1.weightLogInfo_1`  AS Table_A
UNION ALL
SELECT *
FROM `case-study-coursera-458016.Fitabase_Data_1.weightLogInfo_2`  AS Table_B
```

Hourly-Calories and Steps_Total

```
SELECT  *
FROM `case-study-coursera-458016.Fitabase_Data_1.hourlyCalories_1`  AS Table_A
UNION ALL
SELECT *
FROM `case-study-coursera-458016.Fitabase_Data_1.hourlyCalories_2`  AS Table_B

SELECT  *
FROM `case-study-coursera-458016.Fitabase_Data_1.hourlySteps_1`  AS Table_A
UNION ALL
```

```
SELECT *
FROM `case-study-coursera-458016.Fitabase_Data_1.hourlySteps_2`    AS Table_B
```

First, the focus is in daily data: In the daily activity table the following calculations were performed to identify trends and relationships:

```
SELECT Id,
AVG(TotalSteps) AS avg_steps,
AVG(TotalDistance) AS avg_distance,
AVG(VeryActiveDistance) AS avg_veryactive_distance,
AVG(ModeratelyActiveDistance) AS avg_moderatelyactive_distance,
AVG(LightActiveDistance) AS avg_lightactive_distance,
AVG(SedentaryActiveDistance) AS avg_sedentaryactive_distance,
AVG(VeryActiveMinutes) AS avg_veryactive_minutes,
AVG(FairlyActiveMinutes) AS avg_fairlyactive_minutes,
AVG(LightlyActiveMinutes) AS avg_lightlyactive_minutes,
AVG(SedentaryMinutes) AS avg_sedentary_minutes,
AVG(Calories) AS avg_calories,

MAX(TotalSteps) AS max_steps,
MIN(TotalSteps) AS min_steps,
MAX(TotalDistance) AS max_distance,
MIN(TotalDistance) AS min_distance,
MAX(Calories) AS max_calories,
MIN(Calories) AS min_calories,

FROM `case-study-coursera-458016.Fitabase_Data_1.dailyActivity_Total`
GROUP BY Id
```

By Activity Date

```
SELECT ActivityDate,
AVG(TotalSteps) AS avg_steps,
AVG(TotalDistance) AS avg_distance,
AVG(VeryActiveDistance) AS avg_veryactive_distance,
AVG(ModeratelyActiveDistance) AS avg_moderatelyactive_distance,
AVG(LightActiveDistance) AS avg_lightactive_distance,
AVG(SedentaryActiveDistance) AS avg_sedentaryactive_distance,
AVG(VeryActiveMinutes) AS avg_veryactive_minutes,
AVG(FairlyActiveMinutes) AS avg_fairlyactive_minutes,
AVG(LightlyActiveMinutes) AS avg_lightlyactive_minutes,
AVG(SedentaryMinutes) AS avg_sedentary_minutes,
AVG(Calories) AS avg_calories,

MAX(TotalSteps) AS max_steps,
MIN(TotalSteps) AS min_steps,
MAX(TotalDistance) AS max_distance,
MIN(TotalDistance) AS min_distance,
MAX(Calories) AS max_calories,
MIN(Calories) AS min_calories,

FROM `case-study-coursera-458016.Fitabase_Data_1.dailyActivity_Total`
GROUP BY ActivityDate
```

Second, in the weight log info table the following calculations were performed:

```
SELECT Id,

AVG(WeightKg) AS avg_weightkg,
AVG(BMI) AS avg_bmi

FROM `case-study-coursera-458016.Fitabase_Data_1.weightLogInfo_Total`
GROUP BY Id
```

5. Supporting visualizations and key findings

Now, **Tableau** will be used to gain insights from the data, focusing on the summary obtained from **SQL**. This summary includes daily activity by date, Id and Total information (daily_activity_total), as well as heart rate data. Two data sources were created, as demonstrated next:

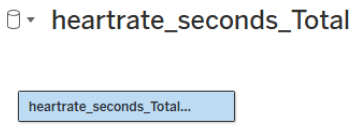

Figure 1. Data source for heart rate

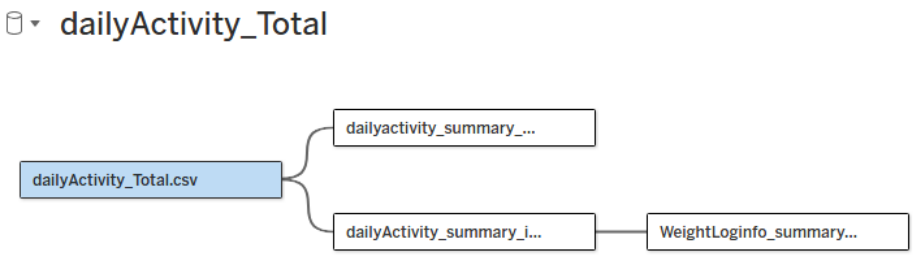

Figure 2. Data source for daily activity

Days with highest average distance and calories

According to Figure 3, the data indicates that the days where people walk or run the most on average are Wednesdays, Mondays and Tuesdays. However, the days with the highest average calorie burn are Saturdays, Fridays and Wednesdays. This supports the trend that smart device users tend to have exercise routines during the week, particularly on Wednesdays and Mondays, and also on weekends, such as Saturdays, as these days show the highest average distance and calorie burn.

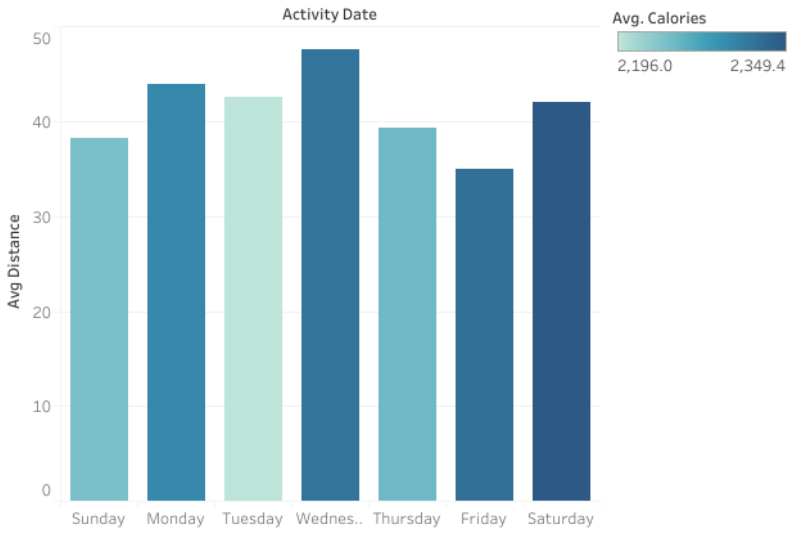

Figure 3. Average distance per day of the week

User calorie range and relationship with steps

Figure 4 presents a histogram displaying the distribution of daily calories observations. Each observation corresponds to the calories recorded by a user on a single day, so users who used the device across multiple days contribute multiple observations. The majority of the smart device users fall within the 1800-2000 calorie range. It is clear that as the number of steps increases, calorie burn also increases, so users can track their results and progress through activity. Furthermore, as the calorie range increases, the number of users decreases, which indicates that the data includes people who used the smart device for daily tracking and others for exercise.

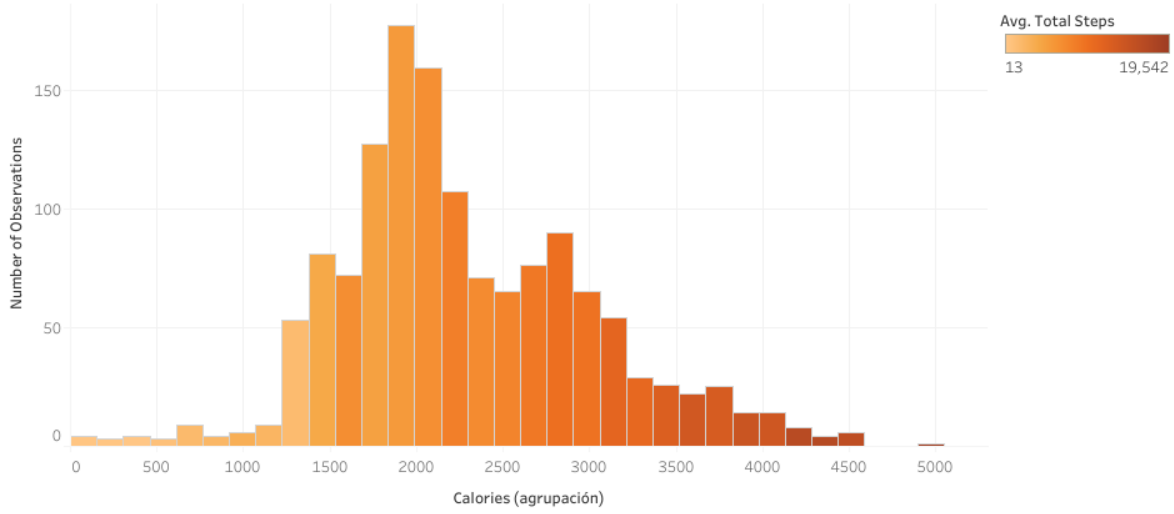

Figure 4. Histogram by calories (with focus on Steps)

Consistency of smart device use

As shown in Figure 5, the results in the first month varied significantly, but from April to May, the trend was more consistent. The reason for this behavior is shown in Figure 6, which presents the daily calorie registrations for each user. The Figure reveals that there were only 2 observations in the first month, but the number of users increased starting in late March. This resulted in higher variability for March compared to April and May.

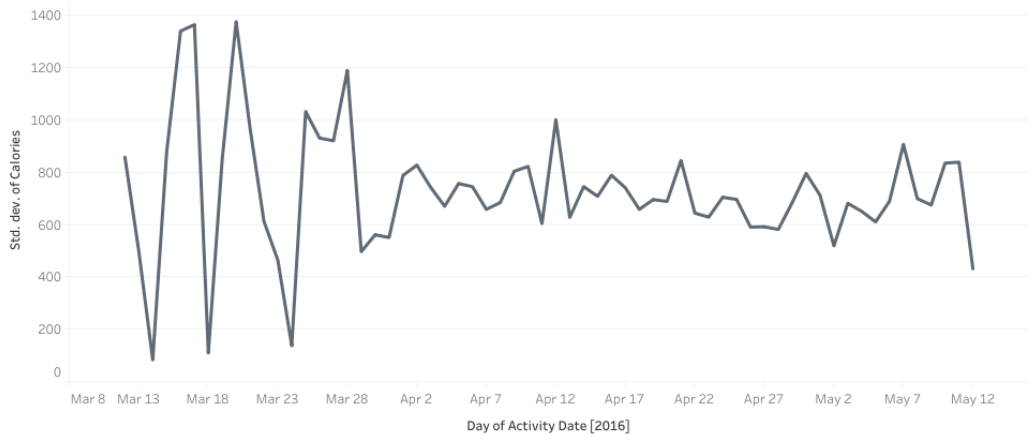

Figure 5. Standard deviation vs Activity Date

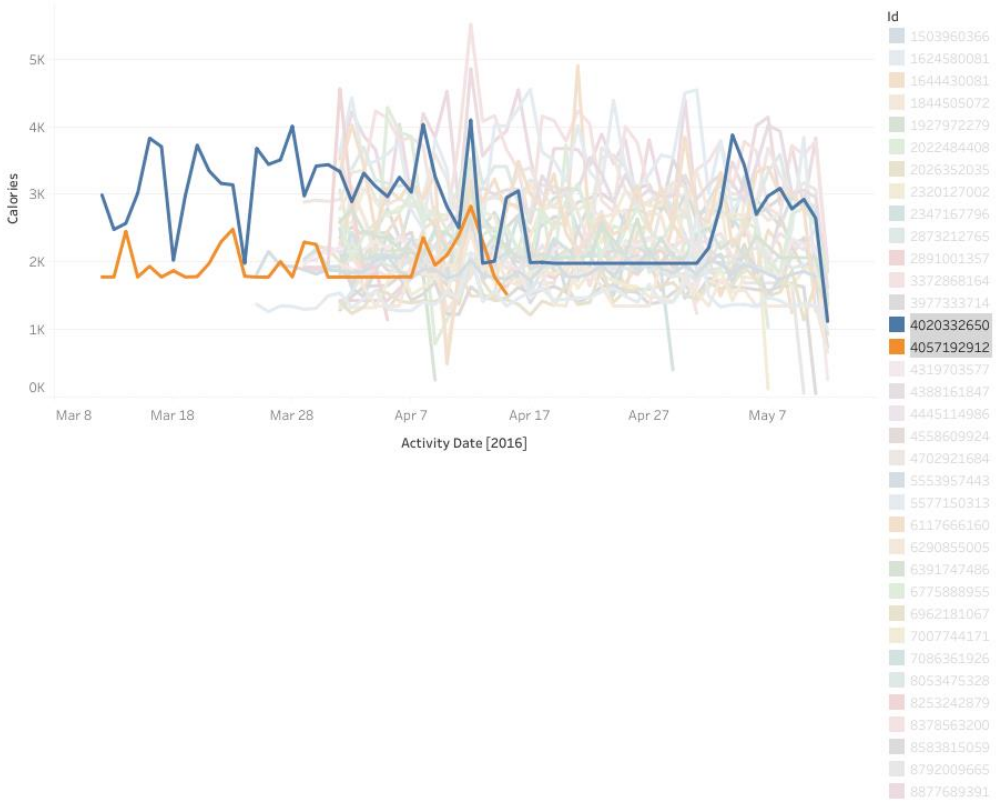

Figure 6. Calories vs Activity Date

Differences among smart devices users

Figure 7 shows average and maximum distance per user, highlighting the variability in smart device usage. Users who cover a higher distance on their active days naturally contribute to a higher average distance. It is observed that smart device use varies significantly among users: some exhibit low average distance, while others show high average distance. For this reason, the smart device should target a diverse user base and not be limited to a specific group (e.g., athletes).

To evaluate the consistency of use, three users with varying average distances from Figure 7 were selected: namely, the users with the highest and lowest average distance, and one with an intermediate average distance. Figure 8 shows the detailed activity patterns, specifically their total distance per activity date. It is clear that consistency of use depends on the individual user and is not uniform. These examples illustrate three different patterns of smart device use across their active days.

Additionally, Figure 9 provides further insight into user activity patterns by focusing on heart rate data. While one user exhibits heart rate spikes on weekends and Wednesday, another user shows a consistent heart rate throughout the week. This indicates that the smart device can be used for tracking exercise on weekends or for daily activity tracking.

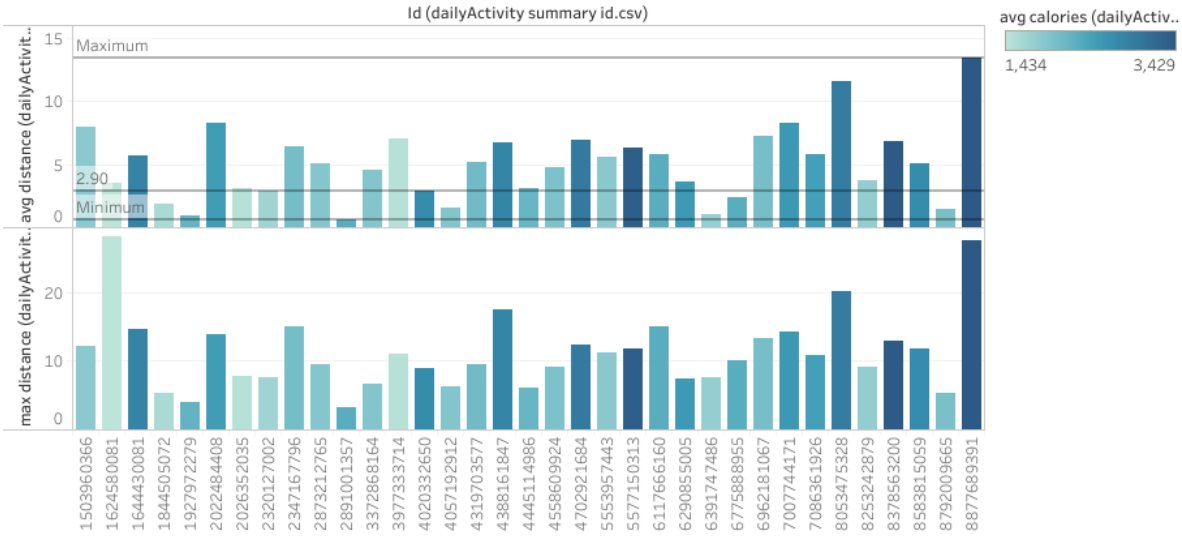

Figure 7. Average and Maximum distance by Id

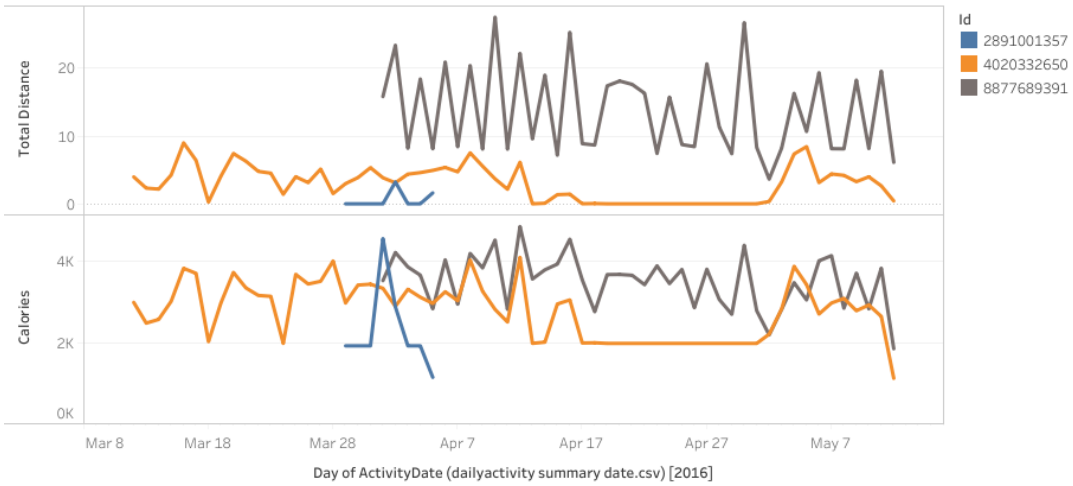

Figure 8. Total Distance and Calories vs Activity Date

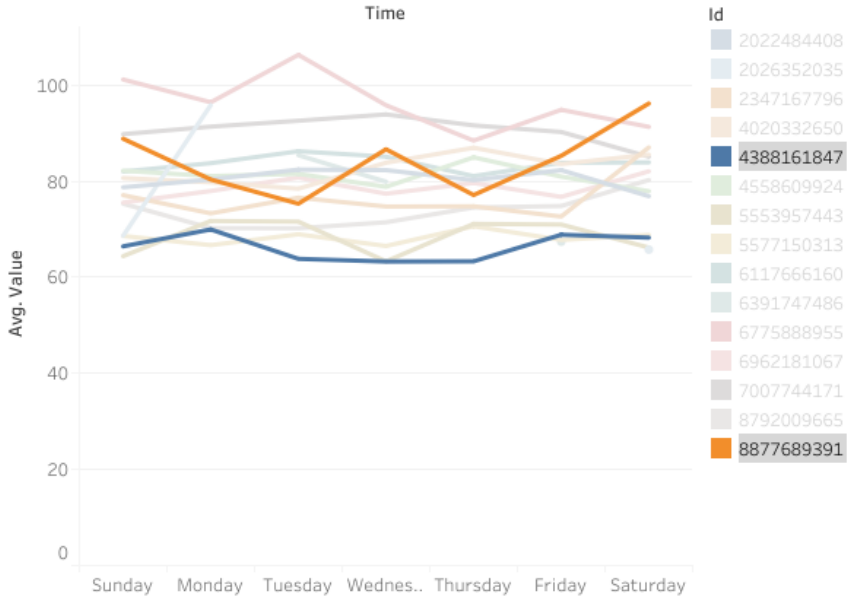

Figure 9. Average heart rate per day of the week

6. Your top high-level content recommendations based on your analysis

Based on this analysis, the following key conclusions are drawn:

1. Users walk or run the most on Wednesdays, Mondays (weekdays), and on Saturdays (weekends), and the major calorie burn also occurs on Saturdays.
2. Users can track their results and progress through activity with steps and calorie tracking.
3. Data variability is higher in March compared to April and May, likely due to the increase in the number of users starting in late March.

4. Smart device use varies significantly among users: some exhibit low average distance, while others show high average distance. For this reason, the smart device should target a diverse user base and not be limited to a specific group (e.g., athletes).
5. Consistency of use depends on the individual user and is not uniform across all users.
6. Smart devices are used by people for exercise and daily activity tracking.

Based on these conclusions, the following recommendations can be considered for the next Bellabeat marketing strategy:

1. Bellabeat can establish a physical presence in key locations, such as parks or running areas, on Mondays, Wednesdays, or Saturdays, as these are the most active days for users to exercise.
2. It is possible to implement a congratulatory alert when a user reaches a certain level of steps or calories burned. Since users pay attention to these metrics, this could be an extra motivation for them.
3. The general strategy should target a diverse user base and not be limited to a specific group (e.g., athletes).
4. Not only fitness enthusiasts use smart devices, but also people who want to track their daily activity. Therefore, highlighting the benefits of Bellabeat products in these areas could be attractive to this target audience. Additionally, the Spring Bellabeat product could be an interesting option for them, as hydration levels are important on a daily basis, highlighting the benefits of this product in online advertisements is recommended.
5. Users with higher average distance or calories should be a target audience for Bellabeat membership, as they may be interested in knowing more about their training, activity, and how to improve it. This could be implemented as a message in the Bellabeat app if they reached a certain level of distance or calories.
6. As the Bellabeat Time product is more like a classic watch and the Leaf offers more versatility in how it can be worn, the Time product should be targeted via the Bellabeat app or online channels for daily tracking, highlighting its benefits. The Leaf product, on the other hand, should be targeted via the Bellabeat app, online, and at fitness events, highlighting its benefits for accurate and more specific fitness metrics.

The trends identified previously are for smart device users and user behavior can vary significantly. Consequently, the trends observed in this data may differ from those of actual Bellabeat customers and their specific behavior when using a smart device.

This project is a practical case study for completing the Google Data Analytics Professional Certificate Course, just for learning purposes.