This project analyzes physical activity data from smart device users to help Bellabeat optimize its marketing strategy. Bellabeat is a high-tech company that creates health-focused products for women. The main goal of this study is to analyze consumer behavior using open Fitbit data and identify growth opportunities for the product.

The dataset was provided by Mobius on Kaggle and is licensed under the CC0 public domain. This dataset contains personal fitness tracker data from thirty Fitbit users. Thirty eligible Fitbit users consented to share their personal tracker data, including minute-level information on physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate, which can be used to study user habits.

The description of this dataset says: *"This dataset was created by respondents during a distributed survey on Amazon Mechanical Turk between December 3, 2016, and December 5, 2016. Thirty eligible Fitbit users agreed to share their personal tracker data, including minute-level information on physical activity, heart rate, and sleep monitoring".*

However, in reality, the data contains information from December 4, 2016, to December 5, 2016, and 33 users participated in the survey instead of 30. Because of this information, I consider the received data to be unreliable.

I downloaded the data as a zip file and extracted 18 documents in .csv format. I used 10 files for this analysis.

Data Processing

To perform this work, I uploaded the following files:

1. dailyActivity_merged
2. dailyCalories_merged
3. dailyIntensities_merged
4. dailySteps_merged
5. hourlyCalories_merged
6. hourlyIntensities_merged
7. hourlySteps_merged
8. minuteSleep_merged
9. sleepDay_merged
10. weightLogInfo_merged

I cleaned the data using Google Sheets:

1. I checked the files for empty cells. In the “weightLogInfo_merged” file, I marked empty cells in the “Fat” column as "NA";
2. I removed duplicate rows. In the "minuteSleep_merged" file, I deleted 543 duplicates, leaving 187,978 unique rows. In the "sleepDay_merged" file, I removed 3 duplicates, leaving 410 unique rows;
3. I organized and sorted all the data;
4. I formatted the date columns as MM/DD/YYYY and the time columns as HH:MM:SS. I used the TRIM function to remove extra spaces at the beginning, end, and between words. I separated the "Date and Time" column into two individual columns using the "Split text to columns" tool in the Data menu.

After cleaning the data, I went back to each file and double-checked all the information.

Data Analysis

After cleaning in Google Sheets, I used BigQuery (SQL) to join the tables and verify that the data was consistent.

I created my own dataset and several tables to store the data.

For the analysis, I used four tables: `dailyActivity_merged`, `dailyCalories_merged`, `dailyIntensities_merged`, and `dailySteps_merged`. I chose them because they all have the same "Id" and "ActivityDate" columns. I used these columns to join the tables and make sure the data was the same.

```
SELECT activity.calories,
calories.calories
FROM `composite-rhino-472219-m9.BellaBeat.dailyActivity_merged` activity
INNER JOIN `composite-rhino-472219-m9.BellaBeat.dailyCalories_merged` calories
ON activity.id = calories.id
AND activity.ActivityDate = calories.ActivityDay

SELECT activity.SedentaryMinutes,
intensities.SedentaryMinutes
FROM `composite-rhino-472219-m9.BellaBeat.dailyActivity_merged` activity
INNER JOIN `composite-rhino-472219-m9.BellaBeat.dailyIntensities_merged` intensities
ON activity.id = intensities.id
AND activity.ActivityDate = intensities.ActivityDay

SELECT activity.TotalSteps,
steps.StepTotal
FROM `composite-rhino-472219-m9.BellaBeat.dailyActivity_merged` activity
INNER JOIN `composite-rhino-472219-m9.BellaBeat.dailySteps_merged` steps
ON activity.id = steps.id
AND activity.ActivityDate = steps.ActivityDay
```

The analysis showed that the data in the calories, intensities, and steps tables perfectly matches the `dailyActivity_merged` table.

Next, I checked the number of unique users using the `COUNT(DISTINCT id)` function.

```
SELECT COUNT(DISTINCT id) AS unique_id
FROM `composite-rhino-472219-m9.BellaBeat.dailyActivity_merged`
```

I repeated the query for all tables, changing only the table name. The number of unique users was less than 33 in the following tables:

1. minuteSleep_merged = 24
2. sleepDay_merged = 24
3. weightLogInfo_merged = 8

Because of these results, I decided to work with these 4 tables:

1. dailyActivity_merged
2. hourlyCalories_merged
3. hourlyIntensities_merged
4. hourlySteps_merged

I wanted to see how often each user logged into the system. I used a `COUNT(id)` query grouped by ID:

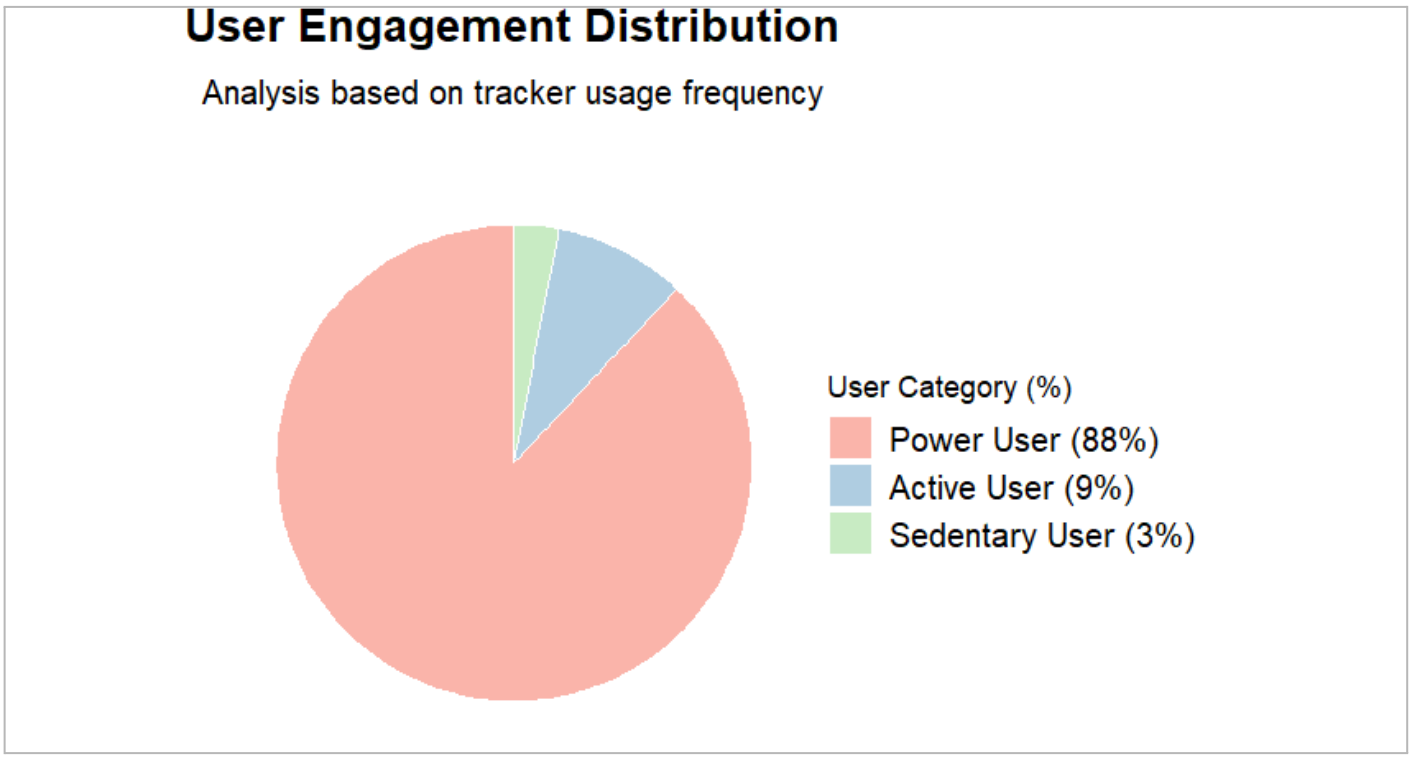
```
SELECT id,
COUNT(id) AS total_id
FROM composite-rhino-472219-m9.BellaBeat.dailyActivity_merged
GROUP BY id
```

The results showed that users logged in between 4 and 31 times. I divided this range into four equal groups to categorize the users:

1. **Sedentary Users**: 4 to 10;
2. **Casual Users**: 11 to 17;
3. **Active Users**: 18 to 24;

4. **Power Users**: 25 to 31.

```
SELECT id,
COUNT(id) AS total_uses,
CASE
WHEN COUNT(id) BETWEEN 4 AND 10 THEN "sedentary user"
WHEN COUNT(id) BETWEEN 11 AND 17 THEN "casual user"
WHEN COUNT(id) BETWEEN 18 AND 24 THEN "active user"
WHEN COUNT(id) BETWEEN 25 AND 31 THEN "power user"
END AS user_classification
FROM composite-rhino-472219-m9.BellaBeat.dailyActivity_merged
GROUP BY id
```

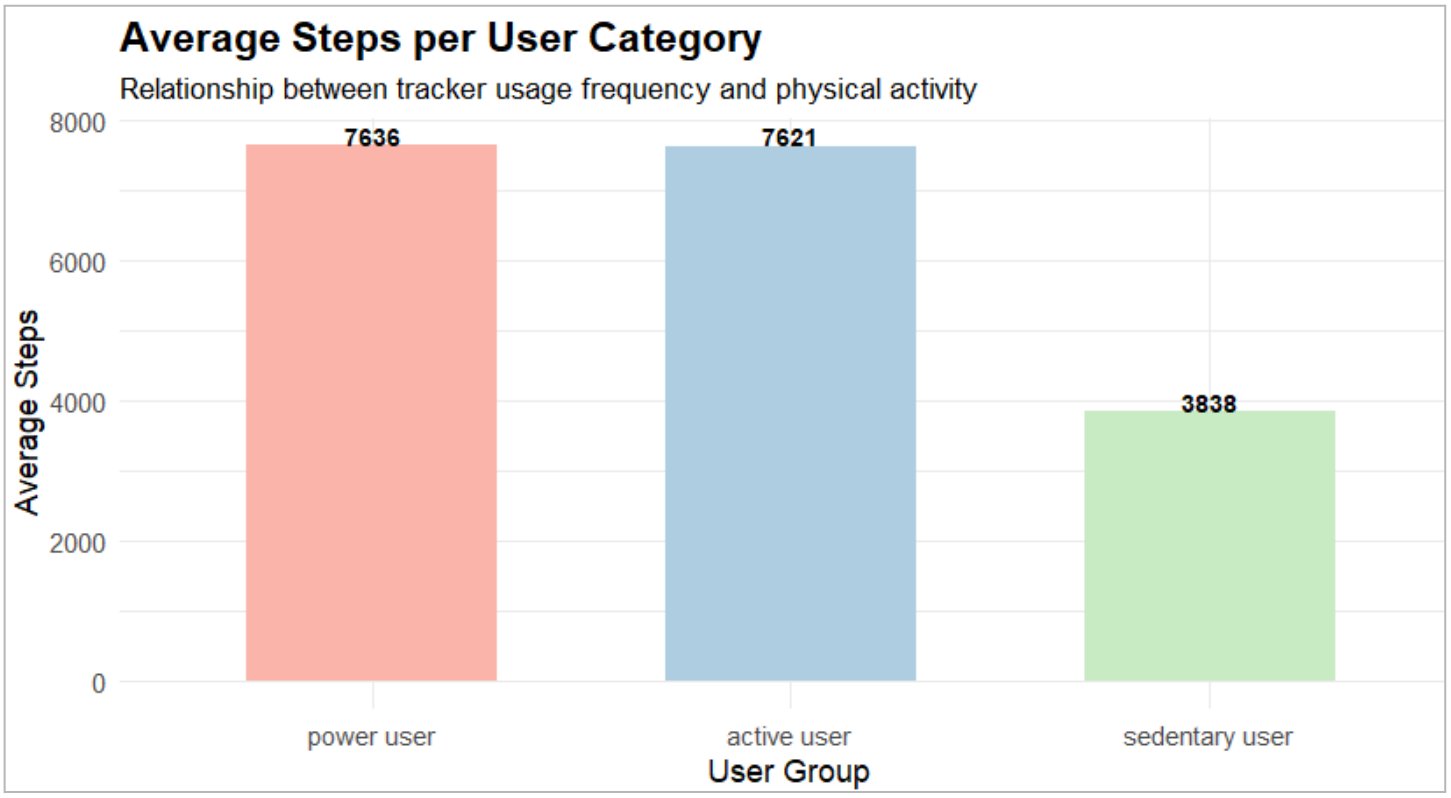


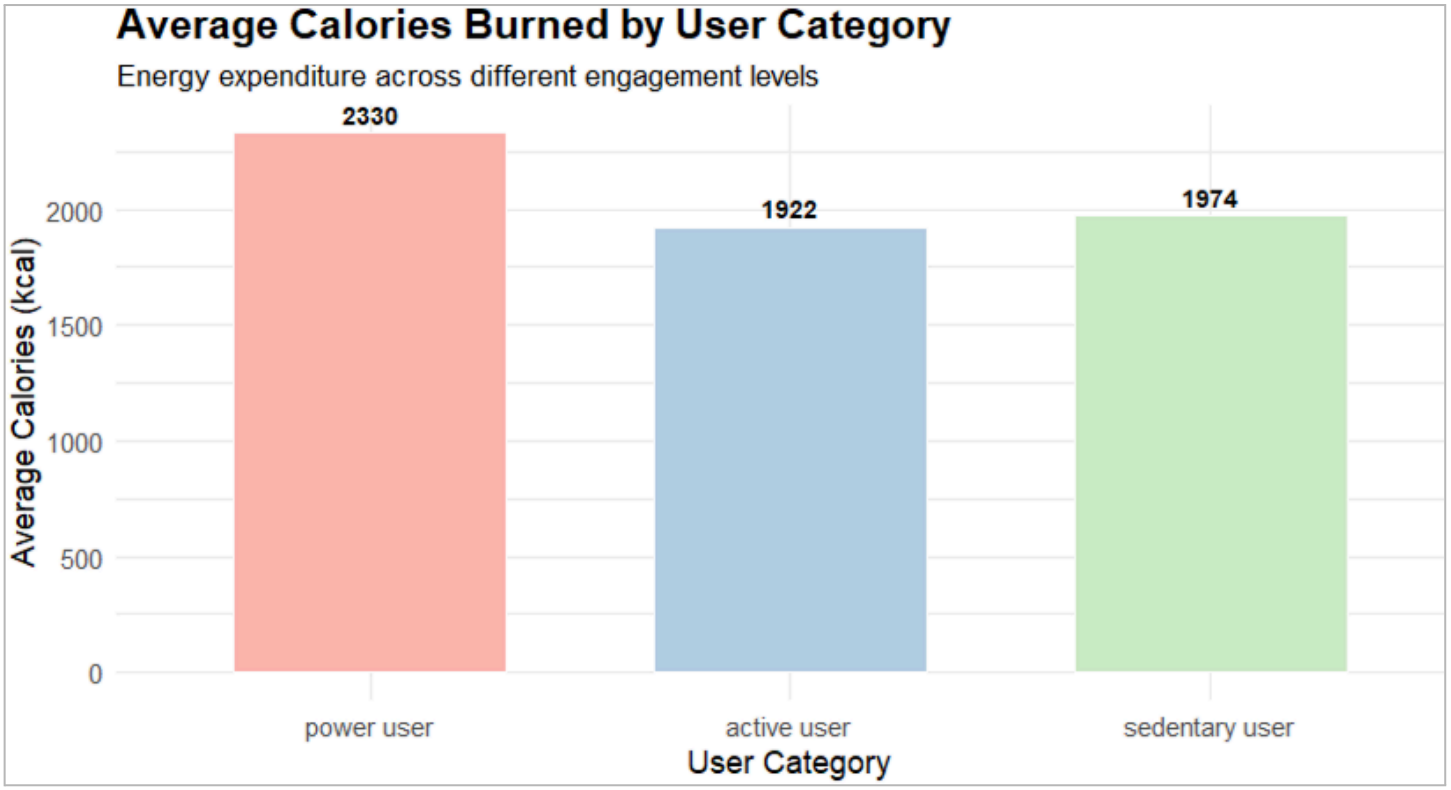
High Retention Rate: 88% of users in the sample are "Power Users," meaning they use their smart devices almost every day. Business Recommendation: Bellabeat should focus on advanced features and long-term health trends, as their core audience is highly engaged and consistent in tracking their data.

After that, I checked the minimum, maximum, and average values for steps, calories, total distance, and other metrics.

```
SELECT id,
ROUND(MIN(TotalSteps), 2) AS min_total_steps,
ROUND(MAX(TotalSteps), 2) AS max_total_steps,
ROUND(AVG(TotalSteps), 2) AS avg_total_steps,
ROUND(MIN(TotalDistance), 2) AS min_total_distance,
ROUND(MAX(TotalDistance), 2) AS max_total_distance,
ROUND(AVG(TotalDistance), 2) AS avg_total_distance,
ROUND(MIN(Calories), 2) AS min_calories,
ROUND(MAX(Calories), 2) AS max_calories,
ROUND(AVG(Calories), 2) AS avg_calories,
ROUND(MIN(VeryActiveMinutes), 2) AS min_very_active_minutes,
ROUND(MAX(VeryActiveMinutes), 2) AS max_very_active_minutes,
ROUND(AVG(VeryActiveMinutes), 2) AS avg_very_active_minutes,
ROUND(MIN(FairlyActiveMinutes), 2) AS min_fairly_active_minutes,
ROUND(MAX(FairlyActiveMinutes), 2) AS max_fairly_active_minutes,
ROUND(AVG(FairlyActiveMinutes), 2) AS avg_fairly_active_minutes,
```

```
ROUND(MIN(LightlyActiveMinutes), 2) AS min_lightly_active_minutes,
ROUND(MAX(LightlyActiveMinutes), 2) AS max_lightly_active_minutes,
ROUND(AVG(LightlyActiveMinutes), 2) AS avg_lightly_active_minutes,
ROUND(MIN(SedentaryMinutes), 2) AS min_sedentary_minutes,
ROUND(MAX(SedentaryMinutes), 2) AS max_sedentary_minutes,
ROUND(AVG(SedentaryMinutes), 2) AS avg_sedentary_minutes,
FROM composite-rhino-472219-m9.BellaBeat.dailyActivity_merged
GROUP BY id
ORDER BY id
```



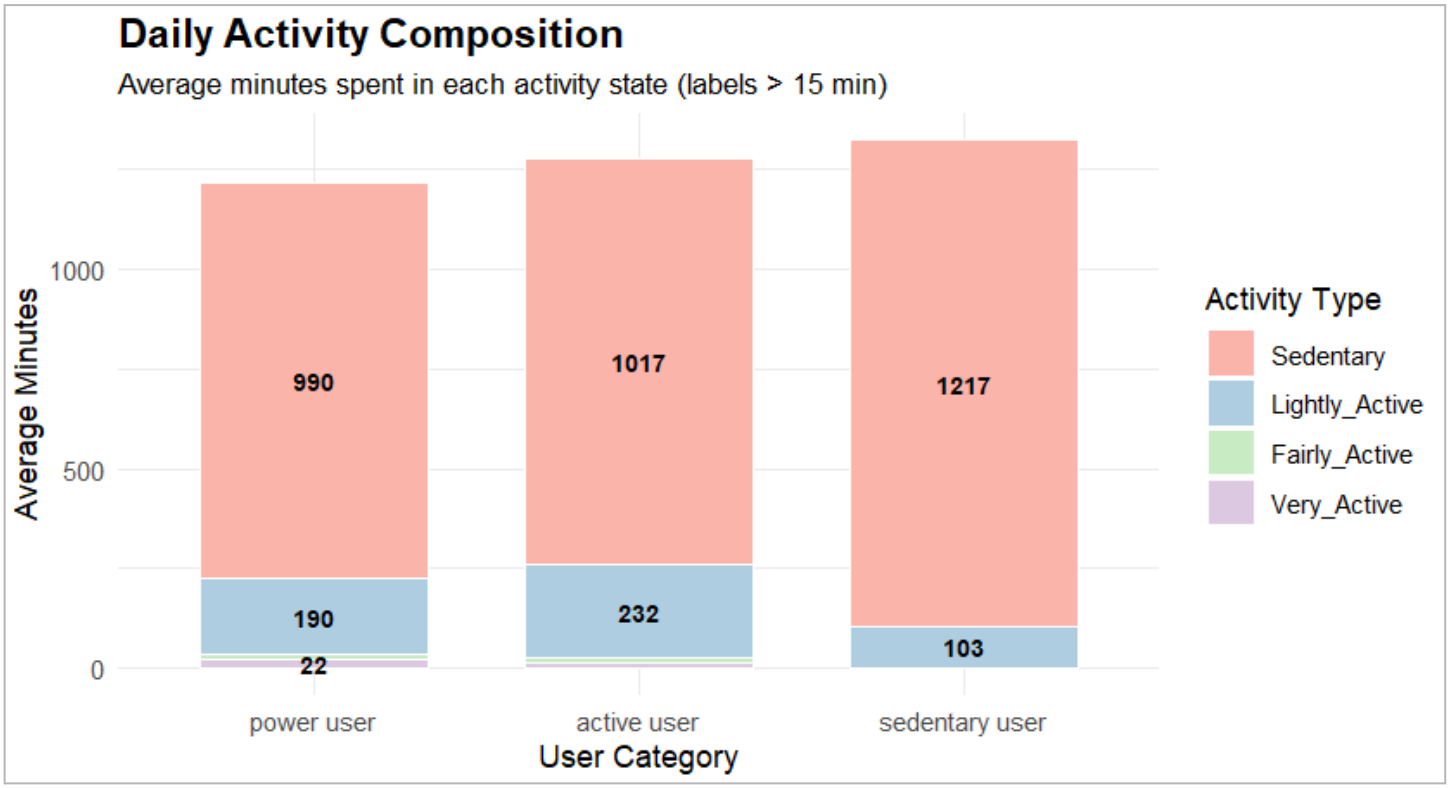


Direct Correlation: There is a clear positive correlation between device usage frequency and physical activity. Power Users (those tracking 25–31 days) consistently achieve the highest daily step counts and caloric burn.

Activity Thresholds: On average, Power Users exceed the recommended daily activity levels, while Sedentary Users fall significantly below, showing a direct link between "tracking habits" and "fitness outcomes."
Engagement Impact: The data suggests that high engagement with the Bellabeat app/tracker acts as a catalyst for a more active lifestyle, as users who wear the device daily are more likely to reach their health goals.

I also looked at the average number of minutes spent on active movements.

```
SELECT id,
ROUND(AVG(VeryActiveMinutes), 2) AS avg_very_active_minutes,
ROUND(AVG(FairlyActiveMinutes), 2) AS avg_fairly_active_minutes,
ROUND(AVG(LightlyActiveMinutes), 2) AS avg_lightly_active_minutes,
ROUND(AVG(SedentaryMinutes), 2) AS avg_sedentary_minutes,
FROM composite-rhino-472219-m9.BellaBeat.dailyActivity_merged
GROUP BY id
ORDER BY 2, 3, 4, 5 DESC
```

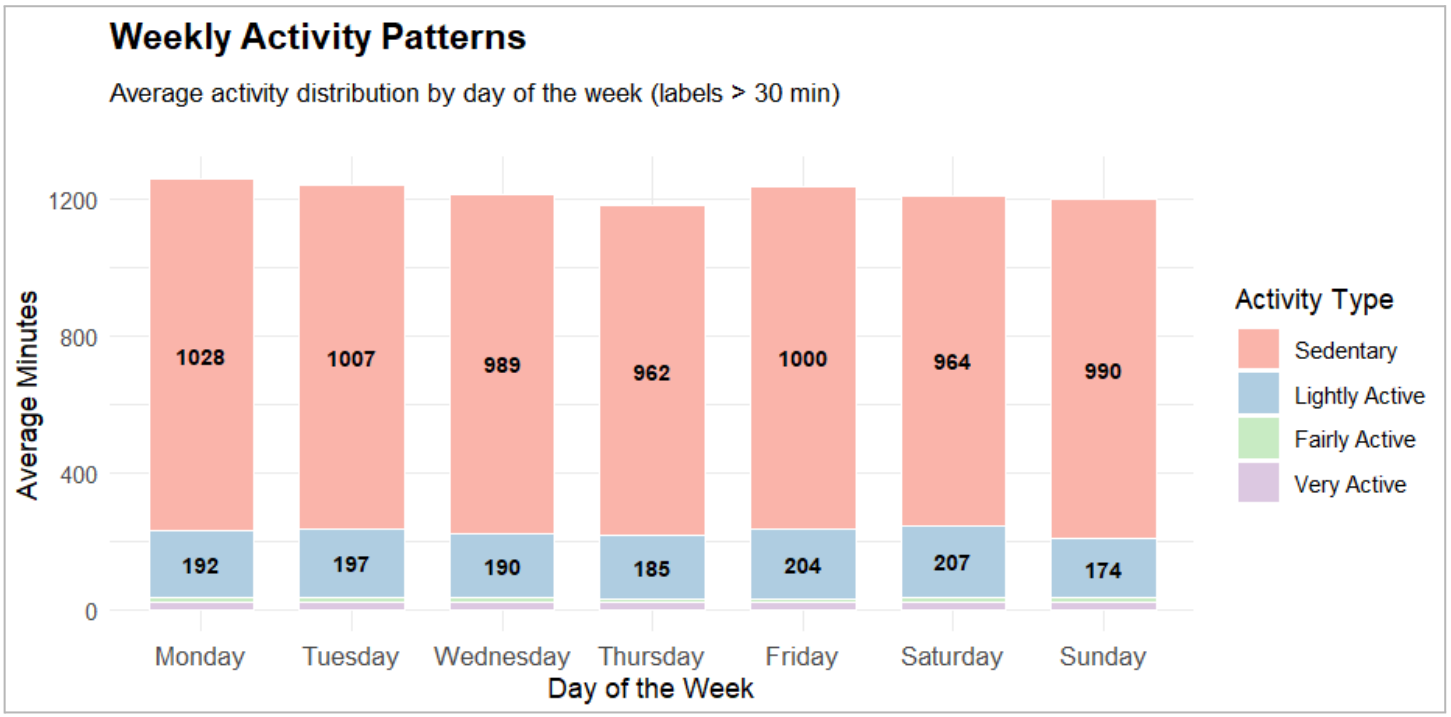


Even "Power Users" spend an average of 990 minutes per day being sedentary. This highlights that wearing a tracker doesn"t automatically eliminate sedentary behavior, but it does help users incorporate more "Very Active" sessions compared to other groups.

I decided to change my query slightly and check on which days of the week users are more active or inactive.

```
SELECT FORMAT_DATE("%A", ActivityDate) AS day_of_the_week,
ROUND(AVG(VeryActiveMinutes), 2) AS avg_very_active_minutes,
ROUND(AVG(FairlyActiveMinutes), 2) AS avg_fairly_active_minutes,
ROUND(AVG(LightlyActiveMinutes), 2) AS avg_lightly_active_minutes,
ROUND(AVG(SedentaryMinutes), 2) AS avg_sedentary_minutes
FROM `composite-rhino-472219-m9.BellaBeat.dailyActivity_merged`
GROUP BY day_of_the_week
```

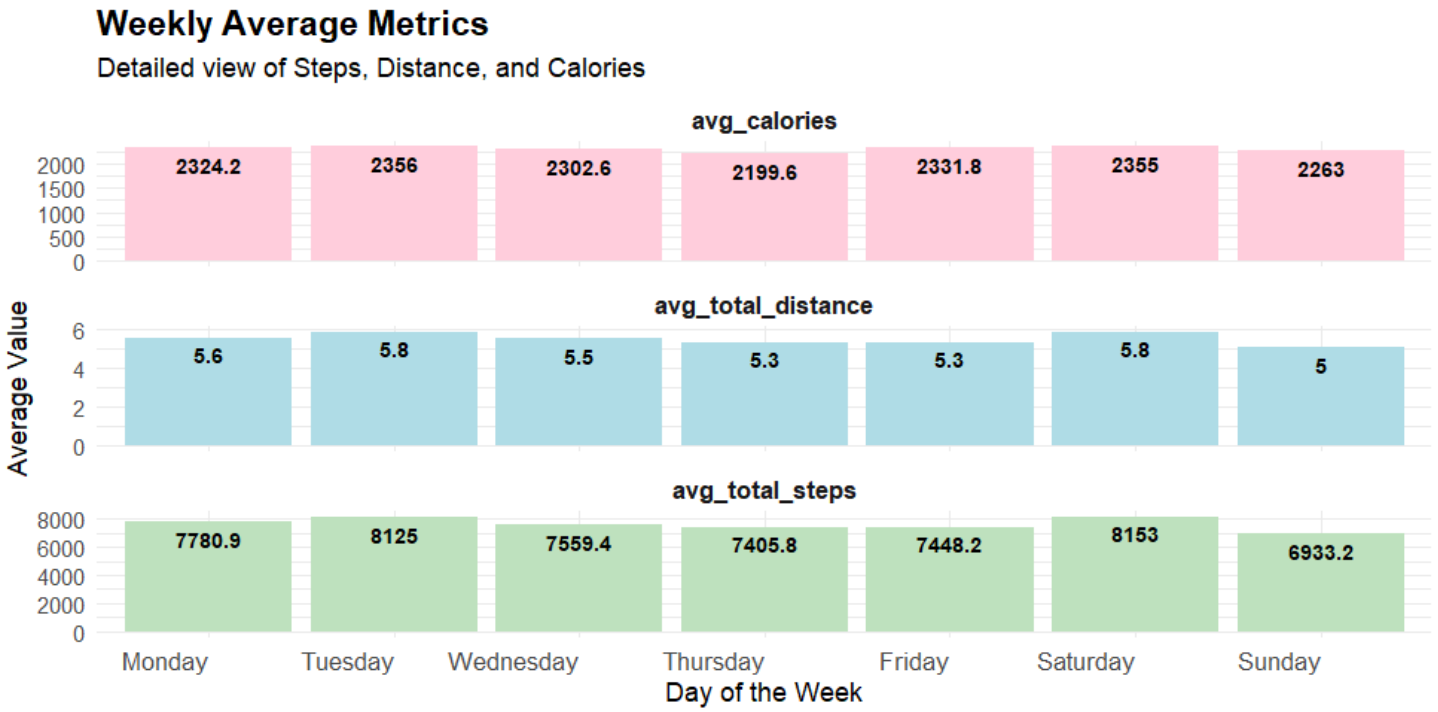
```
ORDER BY 2, 3, 4, 5 DESC
```



Daily Activity Intensity: Analysis shows that users spend the vast majority of their day in a Sedentary state. Time dedicated to "Very Active" and "Fairly Active" intensity often stays below the 30-minute daily goal. This 30-minute benchmark is the official recommendation from the World Health Organization (WHO) to maintain heart health and prevent disease.

I checked the average values for steps, distance, and calories for each day of the week.

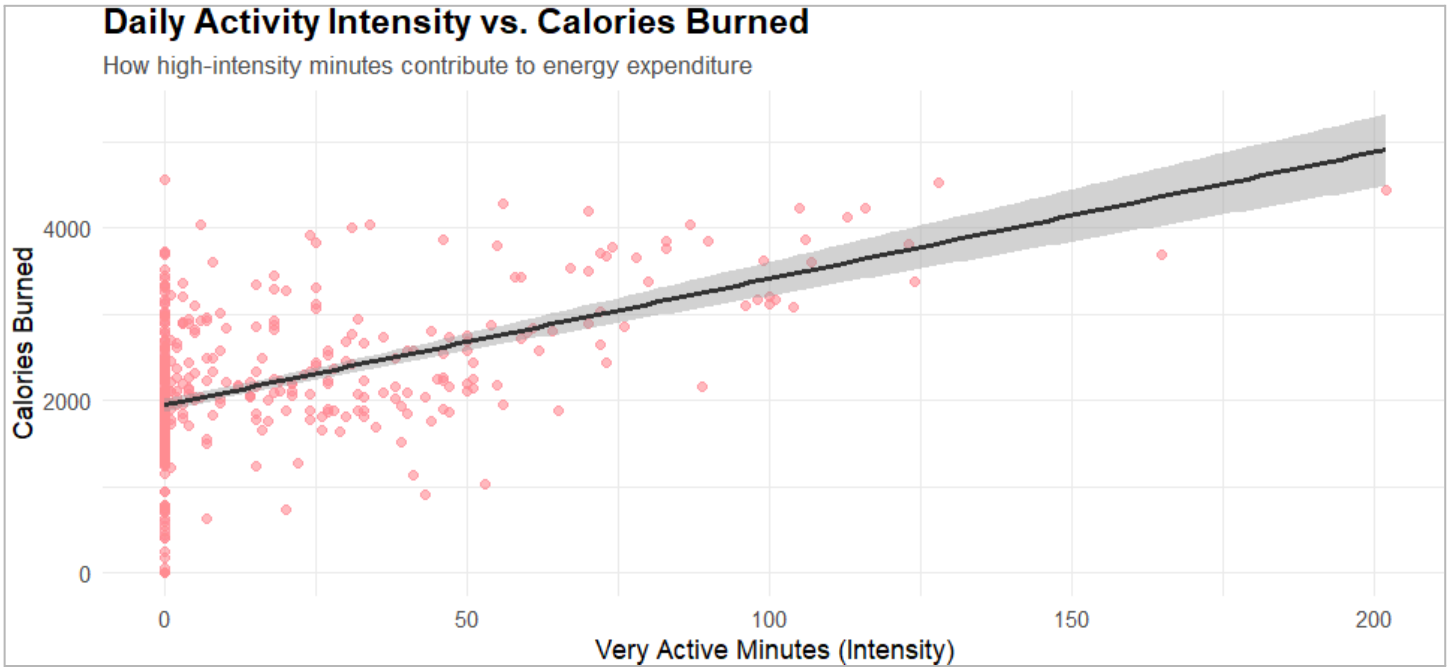
```
SELECT FORMAT_DATE("%A", ActivityDate) AS day_of_the_week,
ROUND(AVG(TotalSteps), 2) AS avg_total_steps,
ROUND(AVG(TotalDistance), 2) AS avg_total_distance,
ROUND(AVG(Calories), 2) AS avg_calories
FROM `composite-rhino-472219-m9.BellaBeat.dailyActivity_merged`
GROUP BY day_of_the_week
ORDER BY 2 DESC
```



Everything is Connected: The data shows that steps, distance, and calories move together. On Tuesdays and Saturdays, users have their best results. They walk more, cover more distance, and burn the most calories (around 2,300 kcal on average).
Step Goals: On most days, users do not reach the 7,500-step goal. Health experts from Harvard Medical School say that 7,500 steps is the minimum number needed to stay healthy and reduce medical risks.
Weekly Patterns: Users are less active in the middle of the week (Wednesday and Thursday) and on Sundays. This shows that their activity depends on their work schedule and the weather. When they don"t go for a walk, their calorie burn drops significantly.
Daily Energy: Even on slow days like Sunday, the tracker shows that users still burn calories. This is because the device counts both their basic body functions (metabolism) and small daily movements.

After reviewing the daily activity table, I decided to use it to compare the number of calories burned with the time spent on these activities.

```
SELECT dailyactivity.id AS dailyactivity_id,
calories.id AS calories_id,
COUNT(dailyactivity.calories) AS dailyactivity_calories,
COUNT(calories.calories) AS total_hourly_calories
FROM `composite-rhino-472219-m9.BellaBeat.dailyActivity_merged` dailyactivity
LEFT JOIN `composite-rhino-472219-m9.BellaBeat.hourlyCalories_merged` calories
ON dailyactivity.id = calories.id
GROUP BY dailyactivity.id, calories.id
```



We compared the time spent on high-intensity exercise with the number of calories burned. The data shows a clear result: the more "Very Active Minutes" a user has, the more calories they burn. While light activity is good, high-intensity workouts are much more effective for weight loss. This means Bellabeat can encourage users to add at least 10–15 minutes of intense exercise to their daily routine to see better results.

Conclusions and Recommendations:

1. Technical Audit and Data Limitations

Before looking at the business recommendations, it is important to note some critical issues in the dataset:

- The task stated that the data covers December 3 to December 5, 2016. In reality, the data only covers December 4 to December 5, 2016 (only 2 days).
- Instead of the 30 respondents mentioned, the data actually includes 33 users.
- **Reliability:** Because the observation period is extremely short (48 hours), this data is not representative of long-term habits. The results might be affected by specific days of the week or random activity bursts.
- The data is considered **insufficiently reliable** for large strategic decisions, but it is useful for identifying general user behavior trends.

2. Key Analytical Insights

Despite the limited data, the analysis identified the following patterns:

- **Intensity vs. Calories:** There is a direct link between high-activity minutes and calorie burn. Intensive exercise is much more effective for reaching fitness goals than just walking.
- **The "30-Minute" Problem:** Most users do not reach the 30-minute daily goal for active exercise recommended by the WHO. Users spend most of their time in a **Sedentary** state.
- **Step Threshold:** The average number of steps is around 7,500. According to research (e.g., Harvard Medical School), this is the minimum threshold for health benefits, but Bellabeat users often do not reach it.

3. Strategic Recommendations for Bellabeat

Based on the identified trends, I suggest the following steps for product development:

- The app should encourage users to replace 20 minutes of light walking with 10 minutes of intense training to get better calorie results.
- Implement notifications with links to WHO recommendations (30 minutes of activity) and scientific data about 7,500 steps. This will build user trust in the app as an expert assistant.
- It is necessary to increase the automatic data collection period to 30+ days to avoid the errors found in this analysis and provide users with personalized monthly reports.
- Add a "warm-up" reminder if the device records more than 2 hours of sedentary time to help users overcome sedentary habits.