Eng. Andrea Cendón Durán

Certificate of Google Analysis Capstone: Complete a Case Study

25, October 2025

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

Part I. Business Context and Problem Definition

I.I Business context

Bellabeat is a company that manufactures high-tech smart products focused on health and women's empowerment by collecting data on their activity, stress, reproductive health, and sleep. The company was founded in 2013 by Urška Sršen and Sando Mur, and it has grown rapidly, to the point that by 2016 it had opened stores all around the world. They launched five main products: the Bellabeat app, Bellabeat membership, Leaf, Time, and Spring. The company currently uses different information networks to promote themselves. At this point, they want to grow more, so they consider that the analysis of the users' data will give them new opportunities to influence the marketing strategies and achieve their objectives.

I.II Business Task

Analyze the data from the company to identify user behavior patterns and factors influencing their engagement. The goal is to present the analysis results and the recommendations to the stakeholders to strengthen the marketing strategies.

I.III Objectives of the analysis

- Identify trends among users.
- Determine features or sections that promote participation.
- Provide data-driven recommendations for marketing strategies.

I.IV Stakeholders

Table 1.- List of stakeholders and their function for this project in the company Bellabeat.

Stakeholders	**Function**
Urška Sršen	Cofounder and Chief Creative Officer
Sando Mur	Mathematician and Bellabeat's cofounder

Part II. Data sources used

The data for this project is located on Kaggle, a platform that contains thousands of public datasets and cloud notebooks that enable collaboration, learning, and model building for data science problem solving.

The project data is called Fitbit Fitness Tracker Data and was published by Arashnic Möbius. It was collected from December 3 to 5, 2016, through a survey sent via Amazon Mechanical Turk, and in total, they obtained data from 30 users. The credibility of this

data is based on the fact that it comes from a known source (Fitbit) and is publicly available, so it can be considered reliable. However, it is essential to consider that these are voluntary users and not a scientific sample, which lowers its credibility. The data is considered objective, qualitative, and consistent; however, it has not been updated since 2016. Another point to consider is that the sample size (30 users) is not very large, which indicates a bias in the research. In addition, the lack of information on the representation of all countries and ages that use Fitbit or only a portion of them may generate a significant bias in the results. On the other hand, we do not have any more details about the dataset, so we cannot update or expand the data. This data was used for the project's purpose and because the course requires it. The data can be found in the following link:

https://www.kaggle.com/datasets/arashnic/fitbit?resource=download-directory

From Kaggle, a compressed folder (.zip) with all the data was imported, and the eleven CSV files were extracted to a folder called Proyecto_Final_Google (Final_Google_Project) located on my desktop. For each step of the project, folders and files were created with appropriate names so that anyone could easily find the information and understand the process. RStudio was used to analyze the structure of the extracted datasets, clean the information, and filter it.

The dataset used in this project is named *Daily_activity_merged.csv*, which contains users' complete and representative physical information. This file includes the following variables: ID, total steps, total distance (including both tracker-registered and app-logged data), calories burned, and minutes spent at different activity levels (very active, moderately active, light, and sedentary). Daily activity data allows for the analysis of general trends and activity habits to understand users' behavior and propose marketing strategies based on real data. It is worth noting that the information is in long format, as there are several entries per user, and to protect their privacy, the dataset only includes identification IDs.

To verify and ensure that each user's ID was real and that their data was complete, 10 complementary data sets were analyzed. Most of the complementary files (datasets of hours and minutes) were not considered because they have shorter time intervals (minutes and hours), while the project focuses on analyzing daily activity patterns. These datasets do not give additional information to the selected file; they only increase data redundancy. The complementary datasets and the reason why they are not included in the main analysis are described below:

- *Heartrate_seconds_merged.csv*

The dataset contains heart rate recordings per second on different days from 14 users. Since the users represent less than 50% of the project's user sample, it limits the information and reduces the validity of the study, making it unrepresentative of the project.

- Hourly datasets (*hourlyCalories_merged.csv, hourlyIntensities_merged.csv, Hourly_steps_merged.csv*)

These files contain records of calories burned, activity intensity, and the number of steps for different hours and days for 34 users, which is already captured by day in the Daily Activity dataset.

- Minute datasets (*minuteCaloriesNarrow_merged.csv, minuteIntesitiesNarow_merged.csv, minuteMETsNarrow_merged.csv, minuteSleep_merged.csv, minuteStepsNarrow_merged.csv*)

These datasets contain the calories burned, intensity of activity, energy expended relative to rest or METs (Metabolic Equivalent of Task), sleep records, and steps per minute for 34 users. The files were discarded because the time intervals are small compared to the daily study level of the main analysis, and they do not add new information. In the case of *minuteSleep_merged.csv*, it was not included because the project is based on physical activity behavior rather than sleep duration. Also, it has only 23 users as a sample, which limits the information for the results of the project.

- weightLogInfo_merged.csv

This dataset shows the weight records in kilograms and pounds, the percentage of fat, and the BMI of eleven users on different days. The file was discarded due to the small percentage of participants (11 users) and because there is not the same number of records per user or per day.

Part III. Process

The process of cleaning and filtering the information in the dataset was performed in RStudio; the procedure can be found in the RMarkdown file called Case_Study_Bellabeat_Prepare_Process_ACD.Rmd, and the code file is named Case_Study_Prepare_Process_code_ACD.R.

After analyzing the structure of the dailyActivity_merged.csv file, it was observed that it contained 457 records from 35 users; the format of the date variable was fixed (changing it from character to date); and that there were no duplicate information or missing values. However, an inconsistency was identified: the original documentation mentioned that the sample had 30 women, but the file had records for 35 IDs.

To analyze this problem, the 10 files available in the repository were imported, and their IDs were compared. As a result, it was identified that ID 4388161847 only appeared in the main file, and its consistency was not verified in the other files. While observing its eight records, no activity was seen, suggesting a recording error or that the device was not used. These records do not contribute significant information to the project and could produce biases in subsequent results. Therefore, these observations were removed, leaving 34 IDs and 449 records.

Validation consistency between values

i. Tracker Distance

Tracker Distance represents the total distance recorded by the device at different activity levels (very active, moderate, light, and sedentary). To evaluate its consistency, it was compared with the sum of activities per level. A new variable was created to represent

the differences between the two metrics, and it has a maximum value of 7.41 and a minimum value of -6.73, indicating considerable dispersion and the presence of outliers. Seventy records with outliers were obtained using Tukey's rule. This rule is a statistical method for analyzing outliers in the data set and evaluating whether or not they are significant in the project (Tukey, 1977). In addition, a comparison of means was made between all records and only those with outliers; it was concluded that observations with outliers should be treated as they have a high mean value compared to the mean. Through manual review of these records, two clear cases were identified. The first consisted of records where Tracker Distance was equal to 0, but the activity levels per minute had values greater than zero, which is due to synchronization or data export errors. The second case showed that Tracker Distance had a value lower than the sum of the activity level in minutes, which is attributed to synchronization failures. Tukey's rule was used again, and the remaining observations were removed; the data set was left with 379 records and 34 users.

ii. Total Distance

Subsequently, the Total Distance variable was evaluated, which is the sum of Tracker Distance and Logged Distance (distance entered by the user). A variable was created with the difference between the two, and a minimum value of -4.828 was obtained, indicating duplication of information or synchronization errors when recording data manually. The Tukey rule was applied to obtain the eight records with outliers. One of the two users who had these types of values has seven records, indicating that the discrepancy was due to user error. These observations were removed to avoid bias in future analyses, and after filtering, we were left with 371 records and 34 IDs.

iii. Remove registers that have zero calories burned by day

Five observations from five different IDs were identified with a value of zero in the variable for calories burned per day. The case is impossible, because the device also accounts for calories burned when the user is not performing functions (basal metabolism) (FitBit Help Center, 2023). These errors can be attributed to partial use of the device or data synchronization issues. These records were deleted, leaving 366 observations from 34 users.

iv. Evaluation of the unrecorded minutes

The variable *Minutos_no_registrados* was generated, which is the difference between the total minutes in a day (1440) and the sum of the activity levels per minute (very active, fairly active, lightly active, and sedentary active). These values allowed us to identify periods when the user does not use the device and enable us to estimate adherence. These records were saved in a separate CSV file for later analysis focused on the device's adherence. Tukey's rule was applied to evaluate inadequate data recording that generates outliers, and those records were deleted. After the process, we retained 358 records and 34 users.

v. Validation of the minimum calories burned in a day

It was established that the minimum limit of calories burned per day to validate the registers is 1000 kcal. An adult woman normally expends between 1200 and 1400 kcal

on average per day, considering basal metabolism and the daily activities (Mifflin & St Jeor, 1990; Frankenfield et al., 2005). However, the value of 1000 kcal was chosen because women have different characteristics that can vary the values of energy burned without activity, avoiding the elimination of potentially valid cases. There were fourteen observations obtained from different users that had values below the limit. These atypical values are attributed to data recording failures or errors in the use of the device. Finally, the data set remains with 344 registers from 33 users.

The reduction of IDs means that one of them has only atypical values, so they were removed. Despite three more IDs than expected, they were kept because their data was valid. All the errors removed from the data represent 24.727% of the observations in the original data set.

At the end, we preserved:

- The filtered dataset is for behavioral and activity-based analyses with 344 records and 33 users.
- The unfiltered dataset was retained for the adherence study, to include the intermittent use, with 358 records and 34 users.

Part IV. Analysis

1. Number of records per ID

To assess the consistency of device use, we obtained and evaluated the daily records for each ID. The data show that the minimum number of entries is 5, while the maximum is 28, with a median of 11 records. The use is relatively consistent over the period analyzed, as indicated by half of the users having records for between 8 and 11 days. Only five of the users have less than 8 records, suggesting low usage or possible abandonment of the device. The findings show a positive commitment to the device, but it is necessary to find opportunities to encourage consistent use and increase retention. Appendix 1 presents the graph corresponding to this analysis.

2. Records per day of week and type of day (Weekday or Weekend).

The number of registrations was evaluated according to the day of the week and the type of day (weekday or weekend). The results show that there are more registrations during the week (223) than during the weekend (121). However, the fact that weekdays have more records than weekend days does not necessarily mean that they have more activity.

To determine which days, have the most activity, the average number of records per type of day was calculated (per weekday and weekend day). The results show that the average activity on weekends is 60.5 records, while the average on weekdays is 44.6. The results indicate greater activity during weekends.

At the individual level, Sunday is the day with the highest number of records (61), followed by Saturday with 60 records. In conclusion, weekends have the highest levels of use, suggesting that users may be more interested in monitoring their activity during their free time. The graph in Appendix 2 shows the number of records obtained per day of the week.

3. Adherence analysis based on minutes of use

To analyze the daily adherence of users with the device, the type of use was classified according to the minutes recorded in four categories: very high, high, moderate use, and low use. In this analysis, the unfiltered file was used to avoid excluding partial use records. Table 1 shows the classification of use categories.

Table 1. Classification of usage criteria by minutes, based on scientific criteria for choosing the limits of each category (Troiano et al., 2007; Migueles et al., 2017; Evenson et al., 2015; Lyden et al., 2021).

Category	Range of minutes per day	Range of hours per day	Interpretation
Low use	< 300 min	< 5h	Occasional use, low adhesion.
Moderate use	300-720 min	5 – 12 h	Regular, but no continuous, use.
High	720 -1080 min	12 – 18 h	Consistent use, adequate adherence.
Very High	> 1080 min	> 18 h	Almost continuous use, high adhesion.

The adherence results show that 56% of the days had very high usage, 35% had high usage, 4.6% had moderate usage, and 2.7% had low usage. In general, 92.6% of the records show high or very high usage, which means that adherence among the users is high, exceeding 12 hours of device usage. On the other hand, 7.3% represents moderate and low use, which means that there are few incomplete records, since more than 8 to 10 hours of daily use are considered reliable records for adherence analysis with wearables (Troiano et al., 2007; Migueles et al., 2017). In conclusion, the results show that users employed the device consistently during the period studied. Figure 1 shows the percentages by category of use.

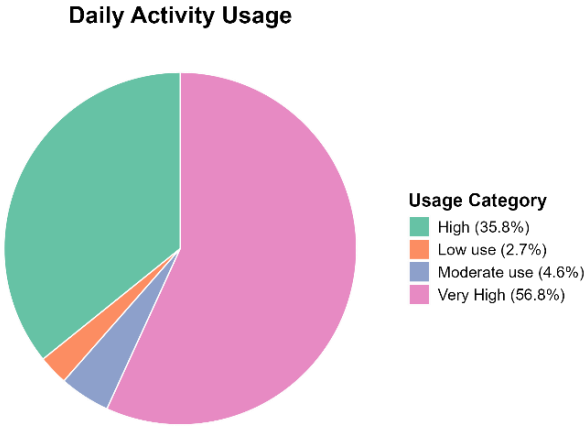


Figure 1.- The pie chart shows the results in percentage of the daily usage of the device, classified into four categories: very high, high, moderate, and low.

4. Compare Adherence vs. Number of Records

This analysis aims to determine whether a relationship exists between ID records and device usage time. This comparison helps us confirm adherence by examining if users who use the device for longer periods of time tend to record data more often.

We combined the Filter file and the Unfiltered file. Since the Unfiltered file did not meet the minimum data quality standards, we were left with 33 users. The choice ensures that the data can be reliable. Figure 2 provides a scatter plot that shows the relationship between the number of records per ID and the percentage of device use.

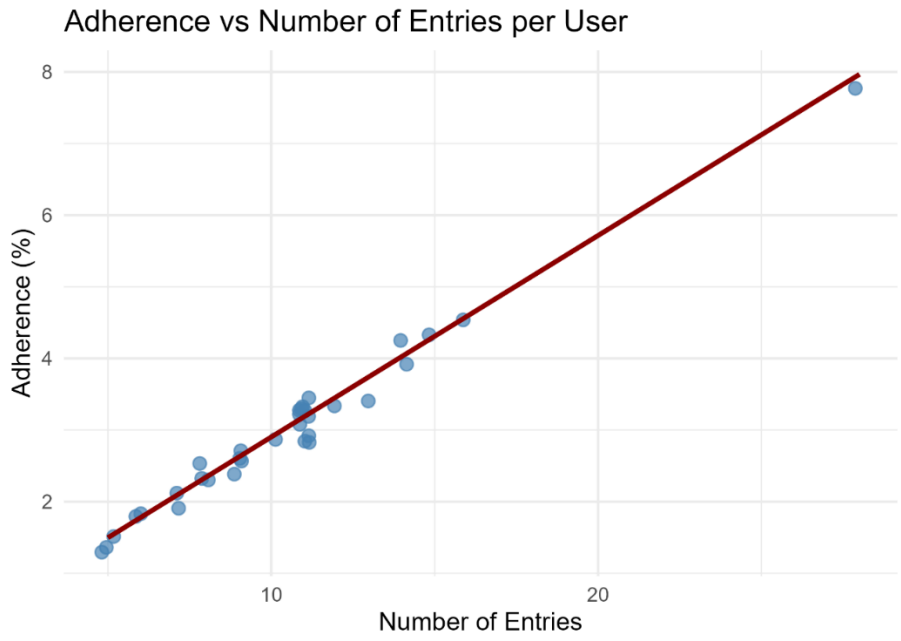


Figure 2.- Scatter plot showing the relationship between the percentage of time each ID uses the device and the number of records per user. Different IDs are represented by the blue dots, while the trend line is shown in red. The darker blue dots represent an overlap of IDs that have the same values on both axes.

According to the graph, the majority of users have records that vary between 5 to 16 and an adherence level between 0% and 5%. One of the users stands out with 28 records and an adherence level of nearly 8%. A clear correlation between the variables is shown by the points' distribution close to the increasing trend line: the higher the number of records, the higher the adherence percentage.

To confirm the relation between the variables, we performed the Pearson's correlation coefficient, which indicates that the variables have a nearly perfect relationship, with a value of 0.994.

5. Mobility Profile of the Users

In this analysis, users were classified according to their level of physical activity based on distance. To avoid bias, they were classified on a daily basis (for each record), and then the most predominant profile for each user was observed. Subsequently, the number of IDs representing each profile and their percentages were obtained; these are shown in

Figure 3. The mobility profiles are as follows: sedentary (less than 0.9 miles), light (0.9 to less than 3 miles), moderate (3 to less than 4.7 miles), and high (more than 4.7 miles). The distance ranges for each category were obtained from the literature (Tudor-Locke et al., 2011; Oja et al., 2018) and converted from kilometers to miles, because our data is in miles.

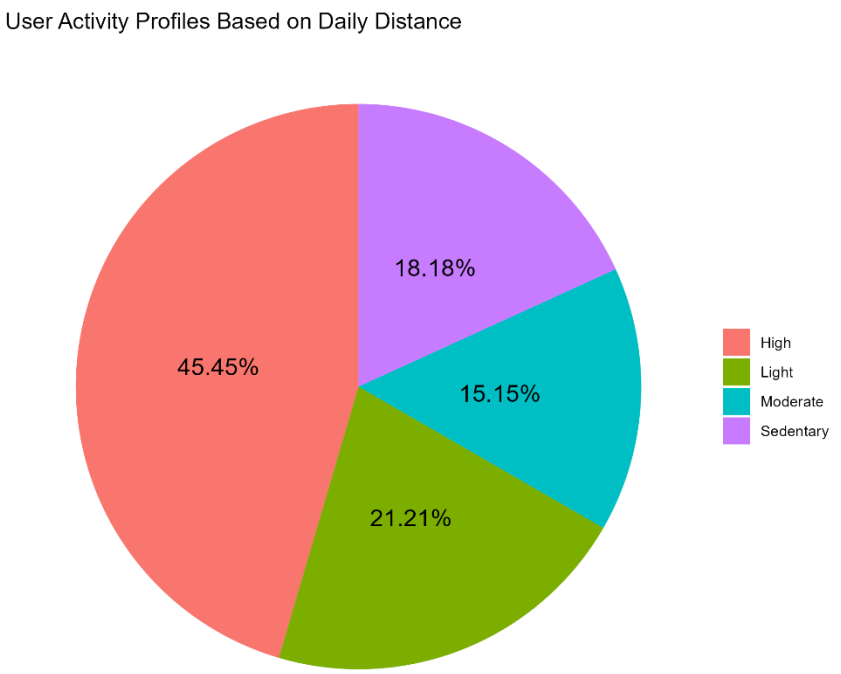


Figure 3.- Shows the percentages of users for each type of mobility profile.

The results show that 15 IDs have a high activity distance profile, and they represent 45.4% of the users. Seven IDs were classified as having a light activity distance profile (21.2%), five IDs have a moderate profile (15.2%), and six IDs have a sedentary profile (18.2%). The results indicate that almost half of the users have high activity. However, it is essential to note that, as the sample of users is small (33 IDs), there may be a bias in the results. Additionally, the profile that represents the smallest number of users is moderate activity.

6. Relationship between Total Steps and Calories Burned Daily

This analysis confirms that users who have more steps recorded by the device burn more calories. Figure 4 shows that most records are between 0 and 17000 steps, with a range of calories burned between 1000 kcal and 4000 kcal per day. It should be noted that records with around zero steps per day but burned more than 1200 kcal per day may be due to exercises that do not require walking, such as Pilates, Yoga, or weightlifting. The linear model increases to the right of the graph, indicating a positive relationship between the variables, which means that the higher the number of steps, the greater the number of calories burned.

Between 0 and 10000 steps, the confidence interval is tighter, which means less dispersion in the data. However, above 10000 steps, the variability increases; this can be caused by the different metabolisms that each user has or the intensity of the activity.

To analyze this relationship, Pearson's correlation was performed. The value obtained was 0.54, indicating that the relationship between the variables is positive but moderate with the linear trend. In conclusion, the results suggest that the greater the number of steps, the greater the daily caloric expenditure, but other factors also play a role (such as the physiological characteristics of the users or the duration of the activity).

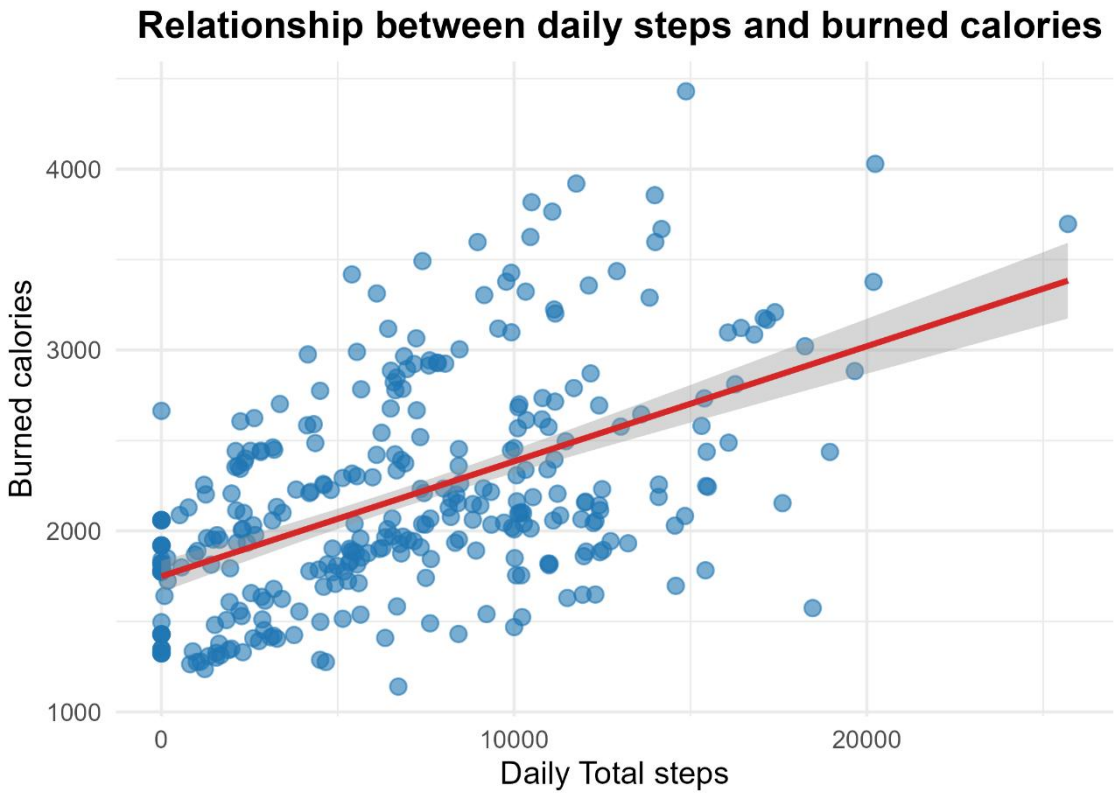


Figure 4. -The scatter plot shows the relationship between the total daily steps and the burned calories. The blue dots represent the registers for each user, the red line represents the linear model, and the grey shadow represents the confidence interval of the linear model.

7. Comparison between average steps on weekends vs weekdays.

The objective of the analysis is to determine whether the average number of steps taken by users is higher on weekdays or weekends. To avoid bias due to differences in the number of records between IDs, the analysis was performed at the individual level before obtaining the overall average, ensuring a fair representation of all users and analyzing the actual behavior of the group.

The results showed that users have a slight tendency to walk more on weekdays (6,910 steps ± 756 standard error) than on weekends (6,536 steps ± 824 standard error); the results can be seen in Figure 5. The difference between the two averages is not very large (374 steps), and there is an overlap in the error bars; it can be said that the difference is not statistically significant. The difference between the averages is attributed to individual differences in movement patterns.

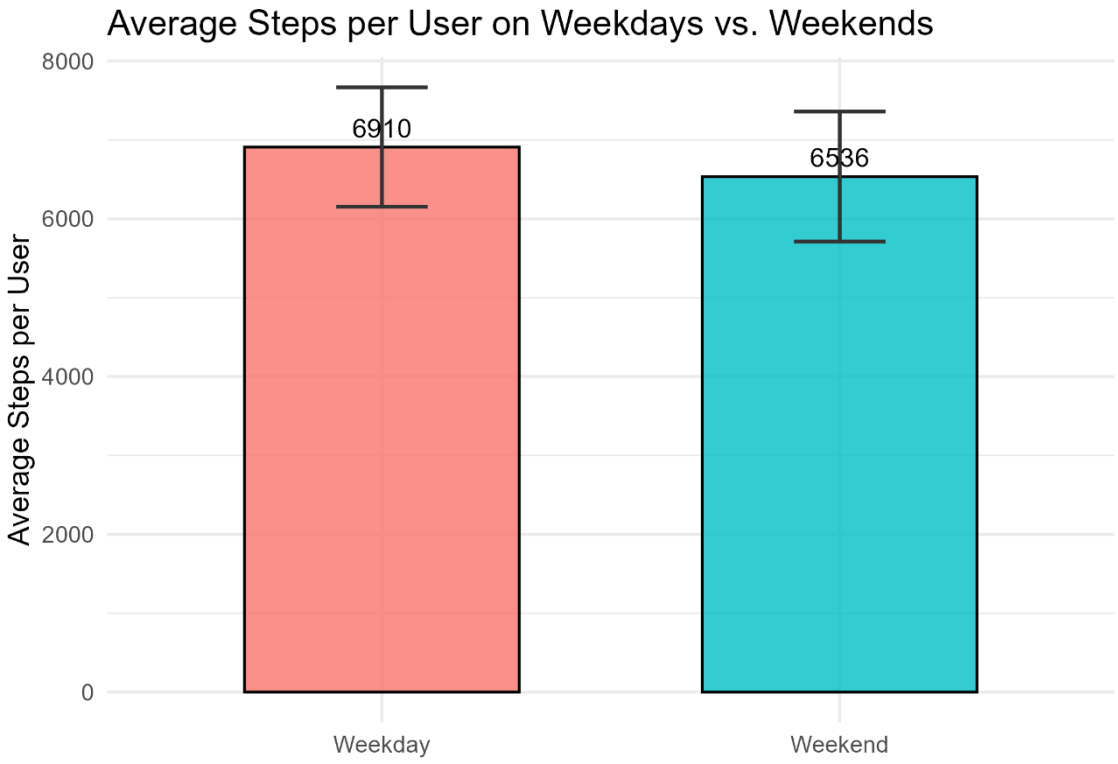


Figure 5.- The graph shows the total average of steps per user during weekdays and weekends. The bars represent the daily average of days per type of day, and the black lines represent the standard error.

8. Classification by type of day and activity profile

The analysis of the records was categorized by type of day (weekday or weekend) and by the level of activity in distance (high, moderate, light, and sedentary). The results show that during weekdays, the predominant mobility profiles are high active distance (48.4%) and sedentary active distance (20.2%). The least frequent profile is moderate active distance (14.3%), followed by light active distance (17%). On the other hand, during weekends, the predominant activity profile is high active distance (40.5%) and light active distance (24%). The least frequent profiles during the weekend are moderate active distance (18.2%) and sedentary active distance (17.4%). Figure 6 shows the results.

The type of day is compared, and it can be observed that the highest percentages in both cases are for high activity distance; however, on workdays, the records in this category increase by 7.9%. The light active profile increases by 7% from workdays to weekends. In the case of the moderately active profile, the percentage during weekends increases by

3.9% compared to workdays. The percentage of the sedentary active profile is 2.8% higher on workdays.

These results can be attributed to the fact that on workdays, users have a structured routine that increases their sedentary lifestyle and have set schedules for physical activity or displacement. On the other hand, on weekends, the type of profiles indicates greater variability in movement patterns.

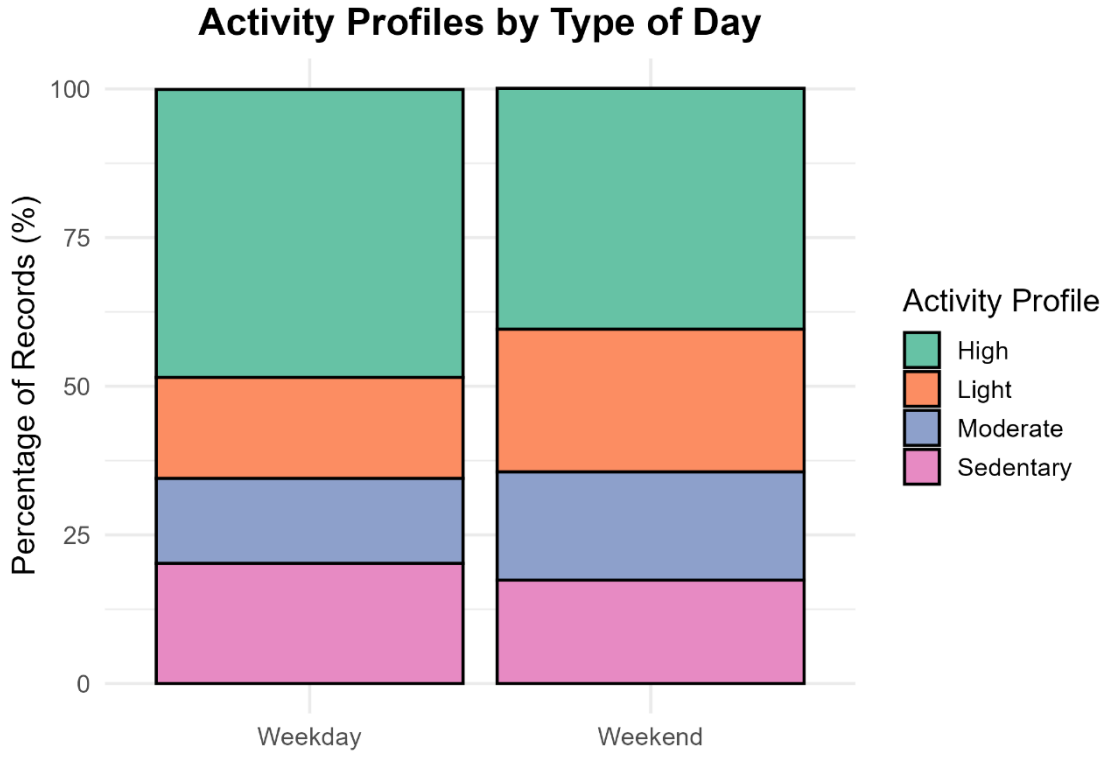


Figure 6.- The percentage distribution of the mobility categories in distance based on the type of day (weekday and weekend).

Part IV. Bibliography

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). **Systematic review of the validity and reliability of consumer-wearable activity trackers**. *International Journal of Behavioral Nutrition and Physical Activity, 12*(1), 159. https://doi.org/10.1186/s12966-015-0314-1

Fitbit Help Center. (2023). *How does my Fitbit device calculate my calorie burn?* Fitbit

Frankenfield, D., Roth-Yousey, L., & Compher, C. (2005). *Comparison of predictive equations for resting metabolic rate in healthy nonobese and obese adults: A systematic review*. *Journal of the American Dietetic Association, 105*(5), 775–789. https://doi.org/10.1016/j.jada.2005.02.075. jandonline.org

Mifflin, M. D., St Jeor, S. T., Hill, L. A., Scott, B. J., Daugherty, S. A., & Koh, Y. O. (1990). *A new predictive equation for resting energy expenditure in healthy individuals*. *The American Journal of Clinical Nutrition, 51*(2), 241–247. https://doi.org/10.1093/ajcn/51.2.241. PubMed

Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nyström, C., Mora-Gonzalez, J., Löf, M., ... & Ortega, F. B. (2017). **Accelerometer data collection and processing criteria**. *Sports Medicine, 47*(9), 1821-1845. https://doi.org/10.1007/s40279-017-0724-3

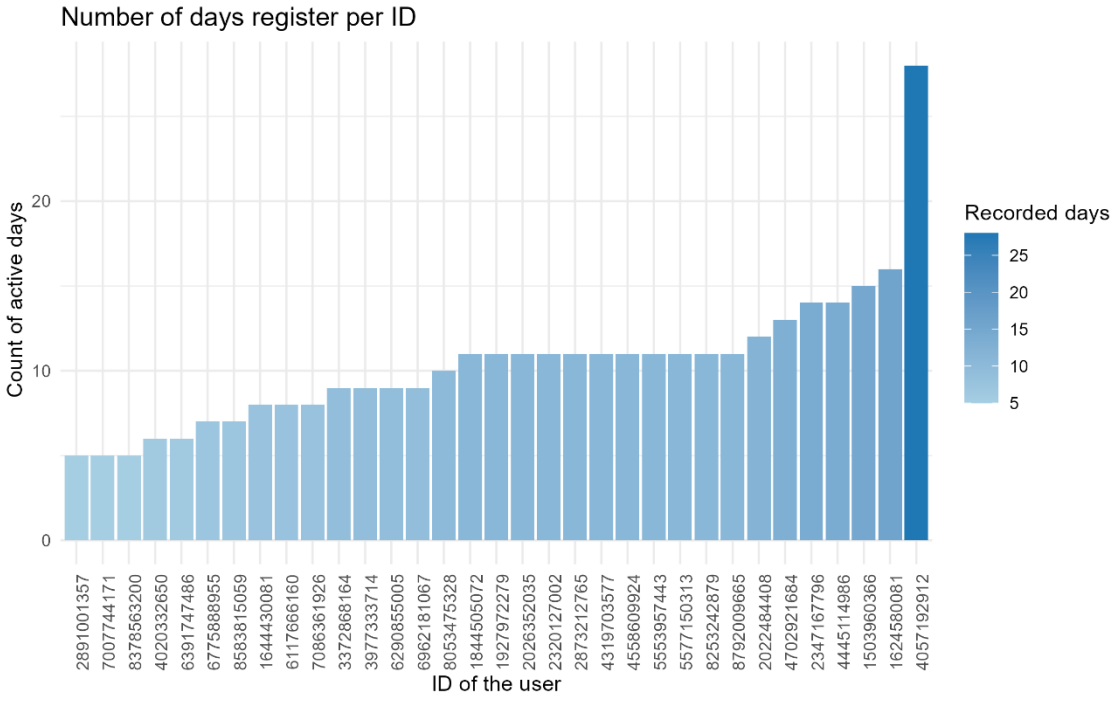
Lyden, K., Keadle, S. K., Strath, S. J., & Staudenmayer, J. (2021). **Estimation of physical activity using wrist-worn accelerometers**. *Medicine & Science in Sports & Exercise*, 53(1), 115-123. https://doi.org/10.1249/MSS.0000000000002431

Oja, P., Kelly, P., Pedisic, Z., et al. (2018). **Daily walking and cardiovascular health.** *British Journal of Sports Medicine.*

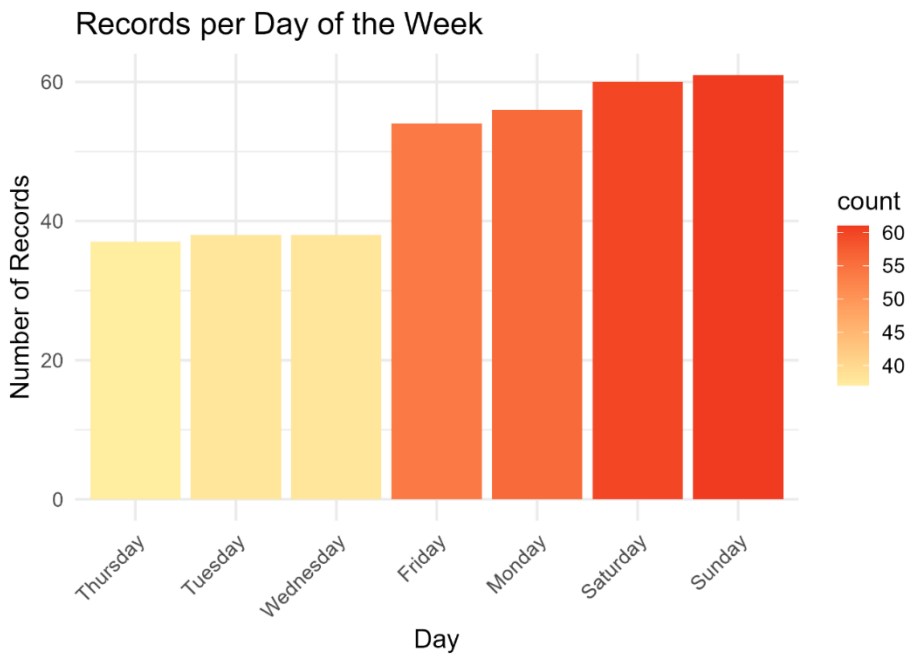
Troiano, R. P., Berrigan, D., Dodd, K. W., Masse, L. C., Tilert, T., & McDowell, M. (2007). **Physical activity in the United States measured by accelerometer**. *Medicine & Science in Sports & Exercise*, 40(1), 181-188. https://doi.org/10.1249/mss.0b013e31815a51b3

Tudor-Locke, C., Craig, C. L., Brown, W. J., et al. (2011). **How many steps/day are enough? for adults.** *International Journal of Behavioral Nutrition and Physical Activity.*

Part V. Appendix



Appendix 1.- The bar chart shows the number of days each ID has been registered. Each bar represents an ID; the intensity of the blue color increases with the number of records for that user.



Appendix 2.- The bar chart represents the number of records per day of the week. The yellow bars show low user activity, the orange bars show medium activity, and the red bars show high activity.