Bellabeat data analysis case study by Damien

Table of Contents

Introduction

This case study is part of the Google Data Analytics Professional Certificate course that I completed in 2024 that has the following fictive scenario:

I'm a junior data analyst working for the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat creates health-focused smart devices that collect data on physical activities, sleep, weight changes of women to help them gain knowledge about their own health and habits. I have been asked to focus on a specific Bellabeat product and analyse smart device data to gain insight into how consumers are using their smart devices. My discoveries will help guide a new marketing strategy for the company.

The stakeholders

The main stakeholders involed in this project are:

- The founders of the company Urška Sršе and Sando Mur
- The bellabeat marketing analytics team whose mission is to collect, analyse, report the data to helps guide the marketing strategy

Business tasks

This case study will focus on completing the following objectives:

- Analyse the data that was generated by Bellabeat smart devices in order to understand how their products are being used
- Identify trends using this data to help the marketing team make better marketing decisions to promote these products and find more opportunities for growth

Prepare the data

Where is the data from?

The data used in this analysis comes from the following public dataset on kaggle:

https://www.kaggle.com/datasets/arashnic/fitbit

How is the data organised?

The dataset is organised into two folders:

- mturkfitbit_export_3.12.16-4.11.16
- mturkfitbit_export_4.12.16-5.12.16

Both folders combined have 2 months worth of data stored in multiple tables from March to May 2016. These tables contain information about various users such as their number of steps, their daily activities, calories burnt, numbers of hours slept, weight information. The data is stored in long format with each user ID having multiple rows

How was the data generated and is it reliable?

The dataset was generated by respondents to a distributed survey via “Amazon Mechanical Turk” between March and April 2016. 30 users agreed to submit their personal tracker data. However, there is not enough explanation about how the survey was carried. There's an inconsistency between the number of users that allegedly submitted their data and the number of IDs found within the dataset. While 30 unique users ID were expected, 33 were found within the dataset. For these reasons, the data doesn't appear to be very reliable.

Is the data recent?

As mentioned in the previous section, the data is from 2016 which makes it outdated

Is the data comprehensive?

The dataset doesn't have any metadata providing more context about the table columns. However, a data dictionary was seperatly made giving very detailed information about each table and column.

Licensing concerns

The dataset uses CC0 license which means it can be used without any particular restriction.

Process the data

This section will explain various techniques to clean the data and make it usable for analysis using Excel. As mentioned previously, the dataset is split into two folders including Fitabase Data 4.12.16-5.12.16 which has the following excel documents:

- dailyActivity_merged.csv
- dailyCalories_merged.csv
- dailyIntensities_merged.csv
- dailySteps_merged.csv
- heartrate_seconds_merged.csv
- hourlyCalories_merged.csv
- hourlyIntensities_merged.csv
- hourlySteps_merged.csv
- minuteCaloriesNarrow_merged.csv
- minuteCaloriesWide_merged.csv
- minuteIntensitiesNarrow_merged.csv
- minuteIntensitiesWide_merged.csv
- minuteMETsNarrow_merged.csv
- minuteSleep_merged.csv
- minuteStepsNarrow_merged.csv
- minuteStepsWide_merged.csv
- sleepDay_merged.csv
- weightLogInfo_merged.csv

Upon further inspection, some the tables are redundant. For instance, dailyActivity_merged.csv is a combination of dailyCalories_merged.csv, dailyIntensities_merged.csv, dailySteps_merged.csv and dailyActivity_merged.csv. For the analysis, the focus will be on the Fitabase Data 4.12.16-5.12.16 folder that has the following tables:

- dailyActivity_merged.csv
- sleepDay_merged.csv

dailyActivity_merged.csv table

The dailyActivity_merged_to_clean table includes information about daily number of steps, exercice intensity and distance, calories burned from different users.

The table contains the following columns:

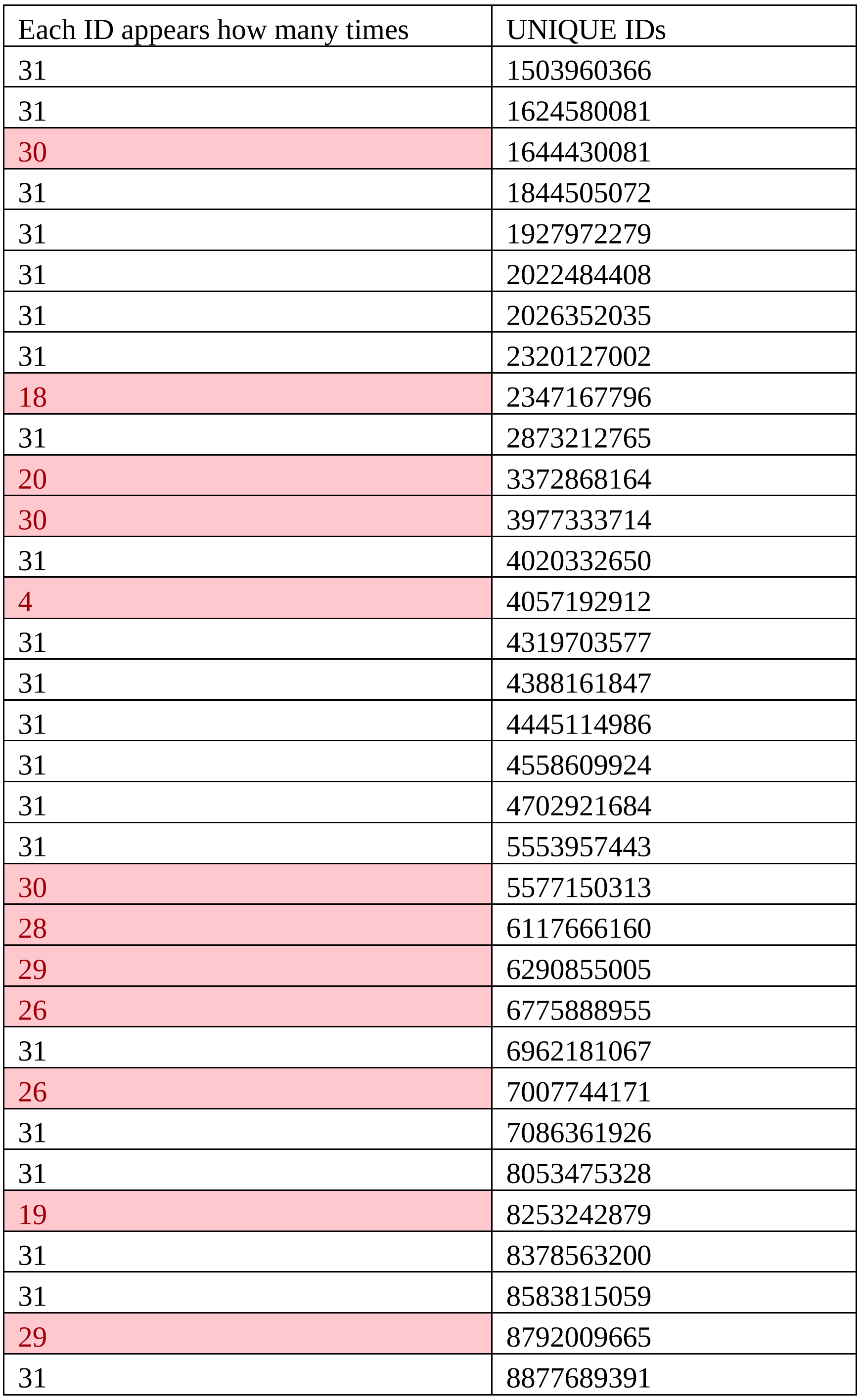
- Id
- ActivityDate
- daily_steps
- total_distance
- TrackerDistance
- LoggedActivitiesDistance
- VeryActiveDistance
- ModeratelyActiveDistance
- LightActiveDistance
- SedentaryActiveDistance
- VeryActiveMinutes
- FairlyActiveMinutes
- LightlyActiveMinutes
- SedentaryMinutes
- calories_burned

Using the UNIQUE function on Excel combined with COUNTIF on the ID column reveals that there are 33 unique IDs

Some of the table names were changed for better clarity

The table should contain data from the 12/04/2016 – 12/05/2016 or a month worth for each unique ID. It means there should be for each user ID, 31 rows, each row corresponding to a day. Using COUNTIF can help return the number of times each ID appears on the table as such:

=COUNTIF("IDs Column range", "ID value")

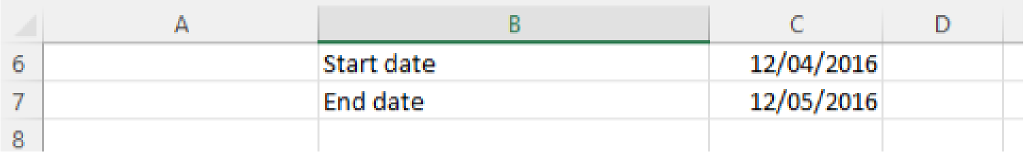


Each ID appears how many times	UNIQUE IDs
31	1503960366
31	1624580081
30	1644430081
31	1844505072
31	1927972279
31	2022484408
31	2026352035
31	2320127002
18	2347167796
31	2873212765
20	3372868164
30	3977333714
31	4020332650
4	4057192912
31	4319703577
31	4388161847
31	4445114986
31	4558609924
31	4702921684
31	5553957443
30	5577150313
28	6117666160
29	6290855005
26	6775888955
31	6962181067
26	7007744171
31	7086361926
31	8053475328
19	8253242879
31	8378563200
31	8583815059
29	8792009665
31	8877689391

As seen on the above image, some IDs appear less than 31 times, meaning for these IDs there's missing data. For example, 4057192912 only appears 4 times in the table. Due to the lack of data for this specific user, I deleted the 4 rows associated with it

The ActivityDate columns should only include dates between the 12/04/2016-12/05/2016. To verify if there's any date outside of this range, the following can be done:

1. Create cells that contain a start and end date



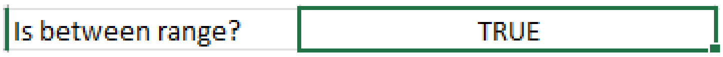
	A	B	C	D
6		Start date	12/04/2016	
7		End date	12/05/2016	
8				

2. Create a formula that returns true if all the dates fall within the start and end date as such:

=AND(F2:F941>=$C$6,F2:F941<=$C$7)

`$C$6``` contains the start date while ```$C$7``` the end date
Range F2:F941 corresponds to the activity date.

The formula results in the following:



Is between range?	TRUE

Added an extra column day_of_the_week. =TEXT(B:B,"dddd") allows to translate each date into a day of the week. B:B refers the ActivityDate column

Checked for duplicates and found none

SleepDay_Merged

Unique Ids
1503960366
1644430081
1844505072
1927972279
2026352035
2320127002
2347167796
3977333714
4020332650
4319703577
4388161847
4445114986
4558609924
4702921684
5553957443
5577150313
6117666160
6775888955
6962181067
7007744171
7086361926
8053475328
8378563200
8792009665

The SleepDay column was changed to MM/DD/YYYY format

Checked for duplicates and found none

Analyse the data and share the results

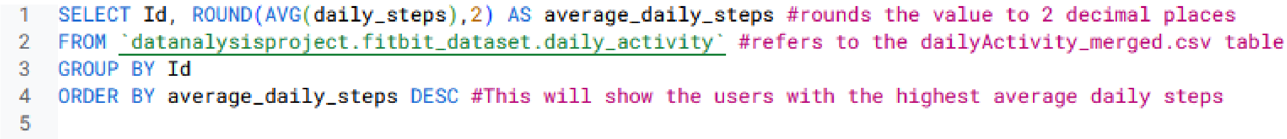
This section will focus on using SQL to run queries and Tableau to create visuals based on the findings

How many daily steps on average did the users make throughout the month?

According to this document [1], lifestyles can be categorised into the following categories based on the number of daily steps:

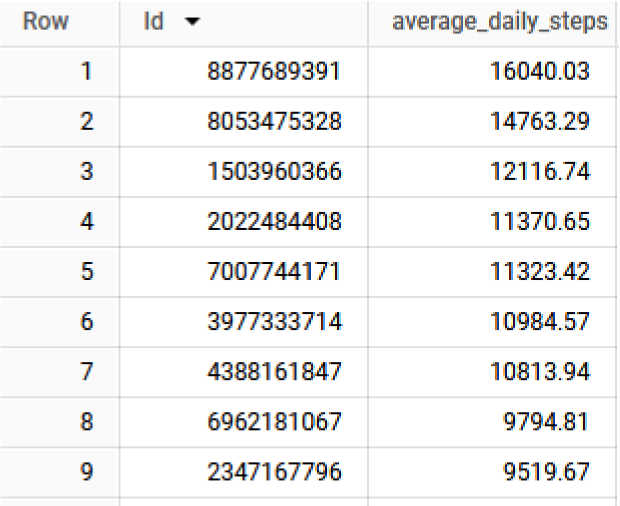
- **Under 5000 steps/day**: Sedentary lifestyle
- **5000-7499/day**: Low active lifestyle
- **7500-9999**: somewhat active lifestyle
- **10,000-12,499**: active lifestyle

Let's check the average number of steps done by users via the following query:



```
SELECT Id, ROUND(AVG(daily_steps),2) AS average_daily_steps #rounds the value to 2 decimal places
FROM `datanalysisproject.fitbit_dataset.daily_activity` #refers to the dailyActivity_merged.csv table
GROUP BY Id
ORDER BY average_daily_steps DESC #This will show the users with the highest average daily steps

```

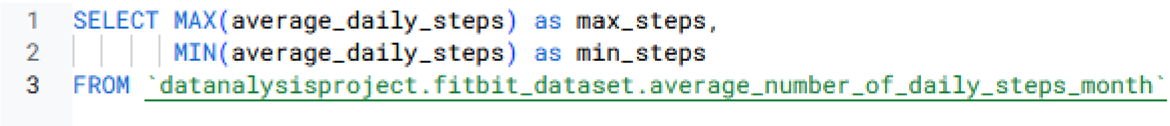
The query results are then stored in a separate table:



Row	Id	average_daily_steps
1	8877689391	16040.03
2	8053475328	14763.29
3	1503960366	12116.74
4	2022484408	11370.65
5	7007744171	11323.42
6	3977333714	10984.57
7	4388161847	10813.94
8	6962181067	9794.81
9	2347167796	9519.67

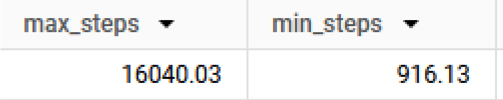
Find the maximum and minimum values

Now, it is possible to find the minimum and maximum amount of steps using this query:



```
SELECT MAX(average_daily_steps) as max_steps,
       MIN(average_daily_steps) as min_steps
FROM `datanalysisproject.fitbit_dataset.average_number_of_daily_steps_month`
```

This results in the following:

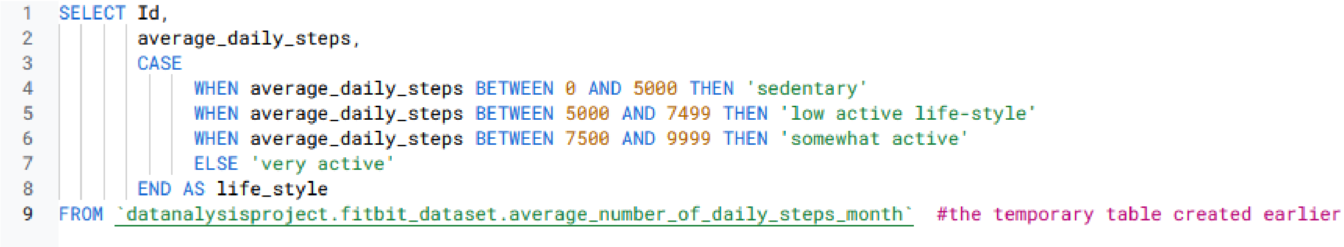


max_steps	min_steps
16040.03	916.13

The maximum value is 16040.03 steps and minimum 916.13 steps

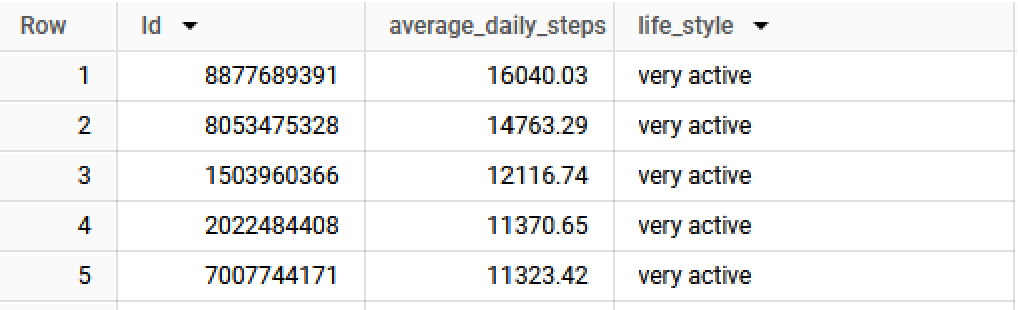
Associate the number of steps with a specific lifestyle

Now let's associate for each user, a category based on their average daily steps:



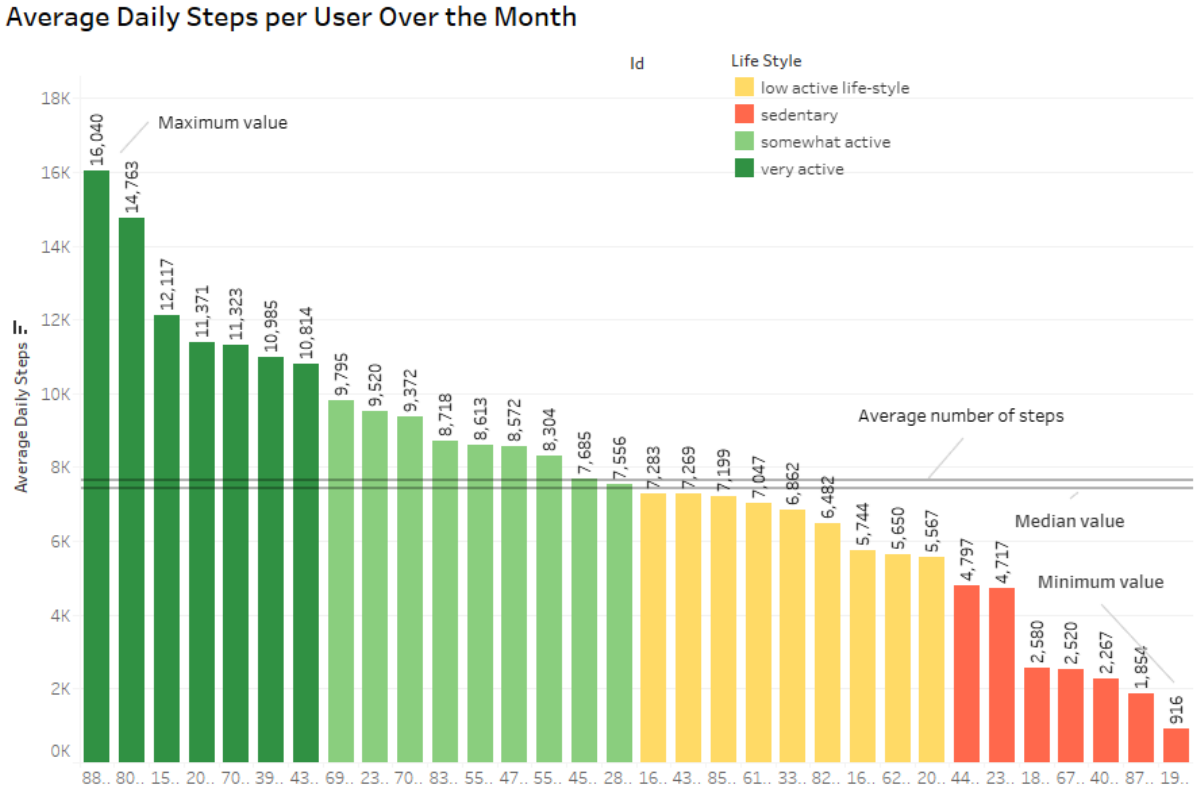
```
SELECT Id,
       average_daily_steps,
       CASE
           WHEN average_daily_steps BETWEEN 0 AND 5000 THEN 'sedentary'
           WHEN average_daily_steps BETWEEN 5000 AND 7499 THEN 'low active life-style'
           WHEN average_daily_steps BETWEEN 7500 AND 9999 THEN 'somewhat active'
           ELSE 'very active'
       END AS life_style
FROM `datanalysisproject.fitbit_dataset.average_number_of_daily_steps_month`  #the temporary table created earlier
```

There's now a life style category associated with every user:



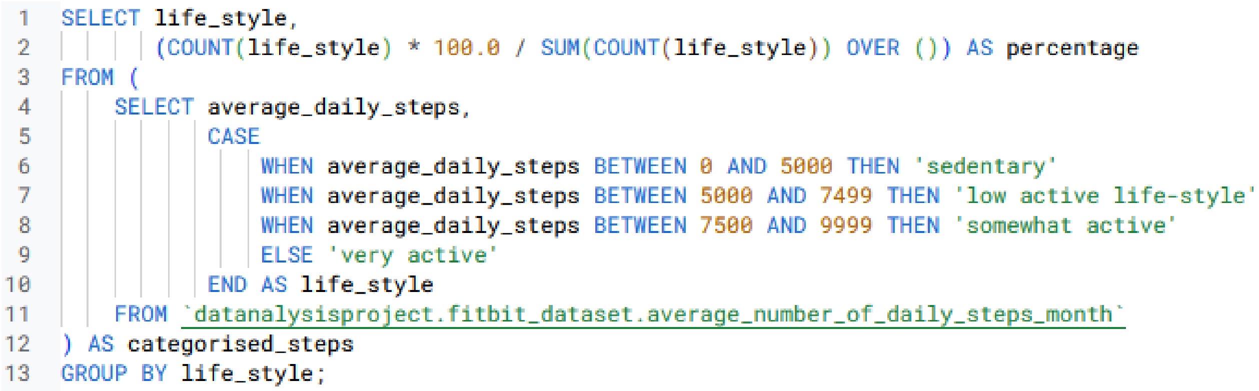
Row	Id	average_daily_steps	life_style
1	8877689391	16040.03	very active
2	8053475328	14763.29	very active
3	1503960366	12116.74	very active
4	2022484408	11370.65	very active
5	7007744171	11323.42	very active

Based on the above table, let's create a visualisation showing the average daily step per ID:



It appears there's very little disparity between categories with some sort of uniform distribution, each of them having a similar proportion of users.

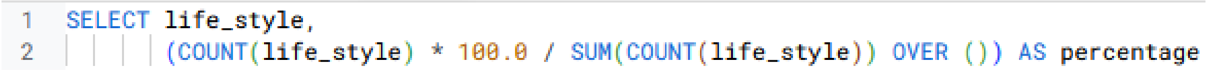
Show the proportion of each life style category via a pie chart



```sql
SELECT life_style,
       (COUNT(life_style) * 100.0 / SUM(COUNT(life_style)) OVER ()) AS percentage
FROM (
    SELECT average_daily_steps,
           CASE
               WHEN average_daily_steps BETWEEN 0 AND 5000 THEN 'sedentary'
               WHEN average_daily_steps BETWEEN 5000 AND 7499 THEN 'low active life-style'
               WHEN average_daily_steps BETWEEN 7500 AND 9999 THEN 'somewhat active'
               ELSE 'very active'
           END AS life_style
    FROM `datanalysisproject.fitbit_dataset.average_number_of_daily_steps_month`
) AS categorised_steps
GROUP BY life_style;
```

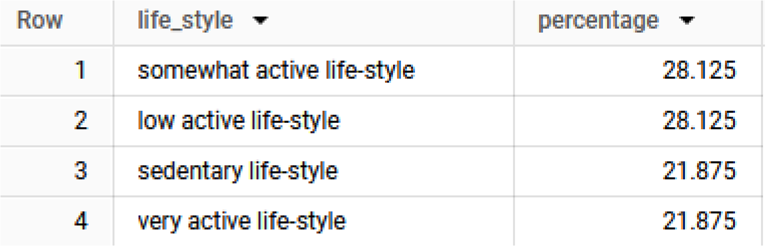
Similarly to the previous SQL query, the **inner query** will associate for each user, a category based on their average daily steps

Let's break the **outer query**:



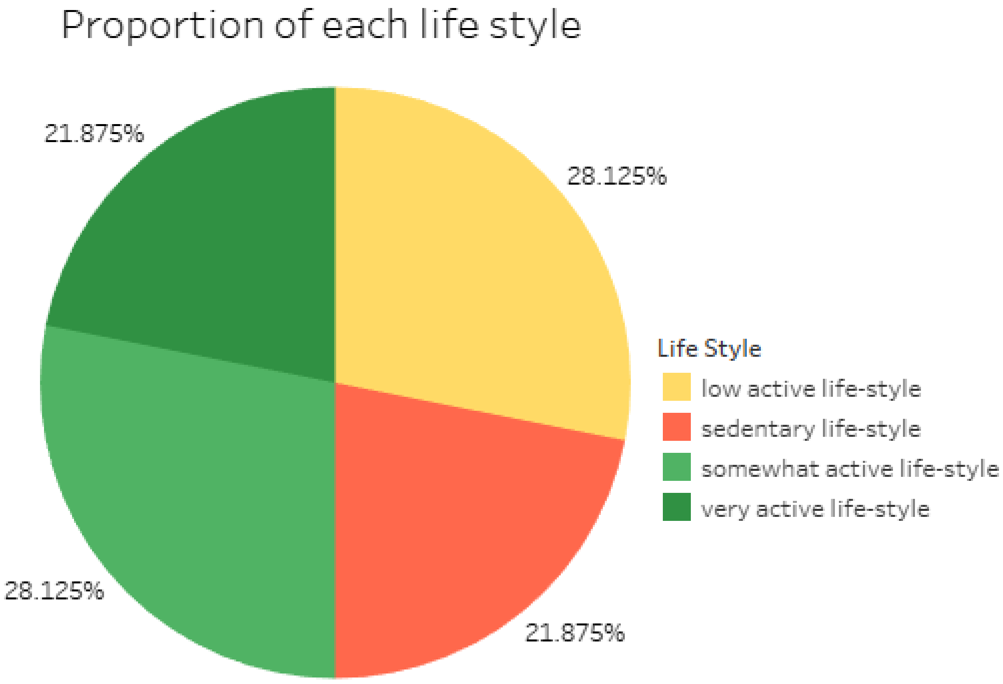
```sql
SELECT life_style,
       (COUNT(life_style) * 100.0 / SUM(COUNT(life_style)) OVER ()) AS percentage
```

First, the number of times each lifestyle occurs is counted and multiplied by 100. The OVER() clause "defines a window or user-specified set of rows within a query result set" according to this document [2]. It will perform calculations across these specified set of rows that are related to the current row. Since OVER () has no argument, it means that this window will be applied through the entire result set. It will compute the SUM of COUNTS across the entire result set which results in the following table:



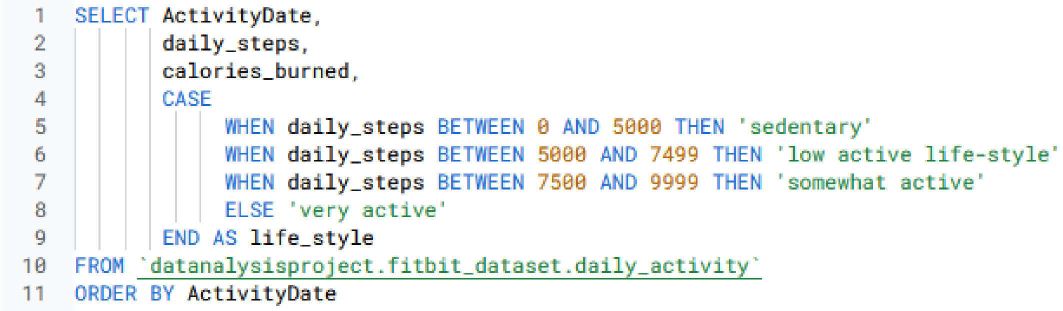
Row	life_style ▾	percentage ▾
1	somewhat active life-style	28.125
2	low active life-style	28.125
3	sedentary life-style	21.875
4	very active life-style	21.875

From that table, the following Pie chart is generated:



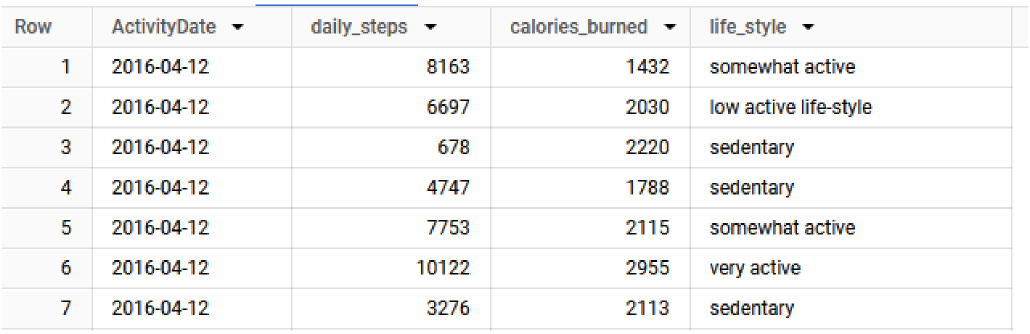
The analysis reveals that 21.875% of users were very active, while 28.125% were somewhat active, resulting in a combined total of 50% of users engaging in some level of activity. Additionally, 28.125% of users led a low-active lifestyle, and only 21.875% were considered sedentary. This indicates that nearly 80% of users demonstrated some level of activity based only on their step count.

Is there a correlation between calories burned and daily steps?



```
SELECT ActivityDate,
       daily_steps,
       calories_burned,
       CASE
           WHEN daily_steps BETWEEN 0 AND 5000 THEN 'sedentary'
           WHEN daily_steps BETWEEN 5000 AND 7499 THEN 'low active life-style'
           WHEN daily_steps BETWEEN 7500 AND 9999 THEN 'somewhat active'
           ELSE 'very active'
       END AS life_style
FROM `datanalysisproject.fitbit_dataset.daily_activity`
ORDER BY ActivityDate
```

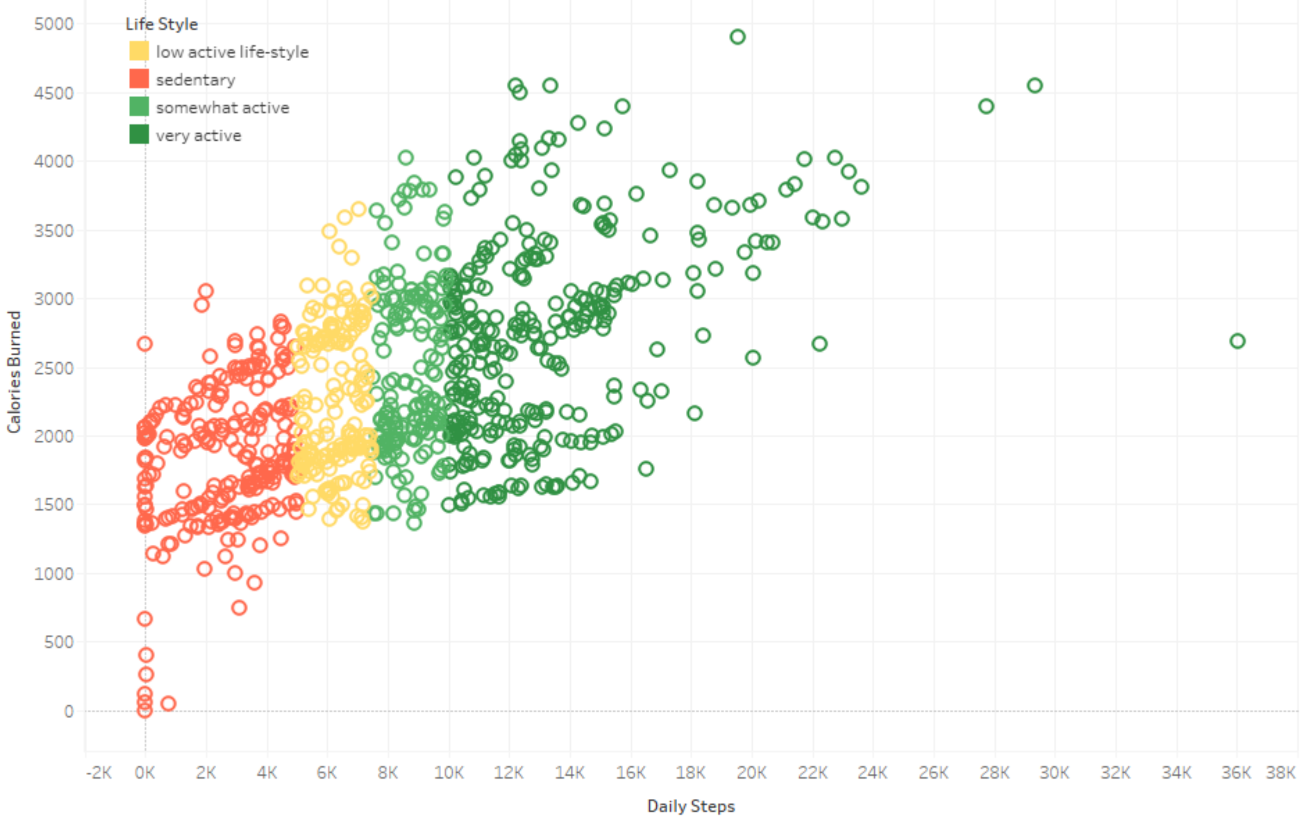
The above query generates a table that includes the activity date, the number of calories burned, and the daily step count along with its corresponding category:



Row	ActivityDate	daily_steps	calories_burned	life_style
1	2016-04-12	8163	1432	somewhat active
2	2016-04-12	6697	2030	low active life-style
3	2016-04-12	678	2220	sedentary
4	2016-04-12	4747	1788	sedentary
5	2016-04-12	7753	2115	somewhat active
6	2016-04-12	10122	2955	very active
7	2016-04-12	3276	2113	sedentary

Based on that table, the following graph was created:

Correlation between daily steps and calories burned for each user

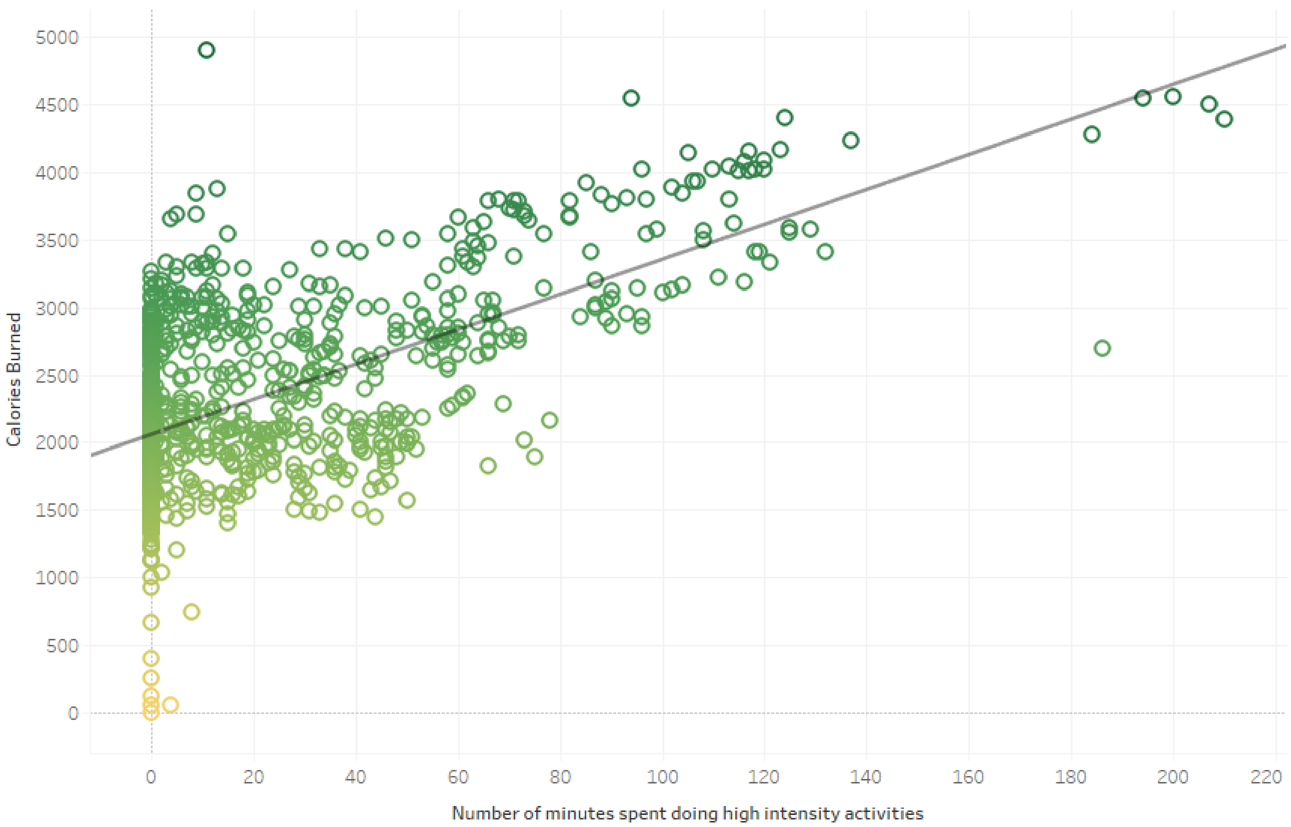


There is a positive correlation between the number of steps and calories burned, as the graph indicates that walking more steps daily generally lead to more calories burned. However, a closer look reveals that one user who walked around 22,000 steps burned approximately 4,000 calories, while another user burned the same number of calories with just 8,000 steps. This suggests that simply walking more steps does not always translate into significantly higher calorie burn. Additionally, the number of steps alone doesn't provide enough context about how they were achieved. For example, reaching the same step count through high-intensity activities would lead to burning more calories in a shorter period compared to low-intensity walking

Is there a correlation between high intensity exercices and calories burned?

Let's generate a similar graph but this time, the focus will be on the number of minutes spent on high intense activities and calories burned:

Correlation between minutes spent doing high intensity activities and calories burned

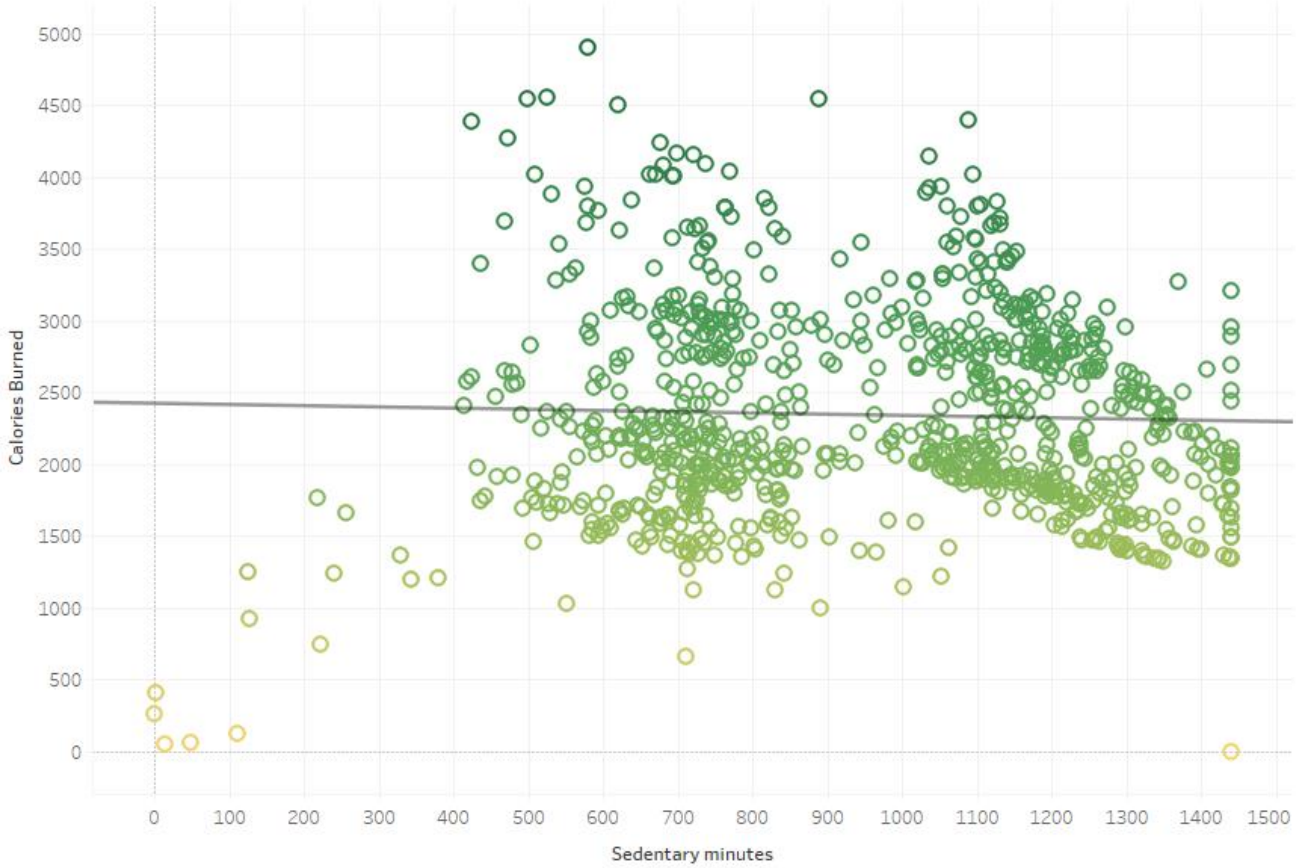


There's also a positive correlation between the number of minutes spent on high intense activities and calories burned. The graph seems to confirm the idea that spending more time on high intense activities lead to burning more calories.

Is there a correlation between sedentary time in minutes and calories burned?

Let's generate a graph which will show the correlation between sedentary time in minutes and calories burned:

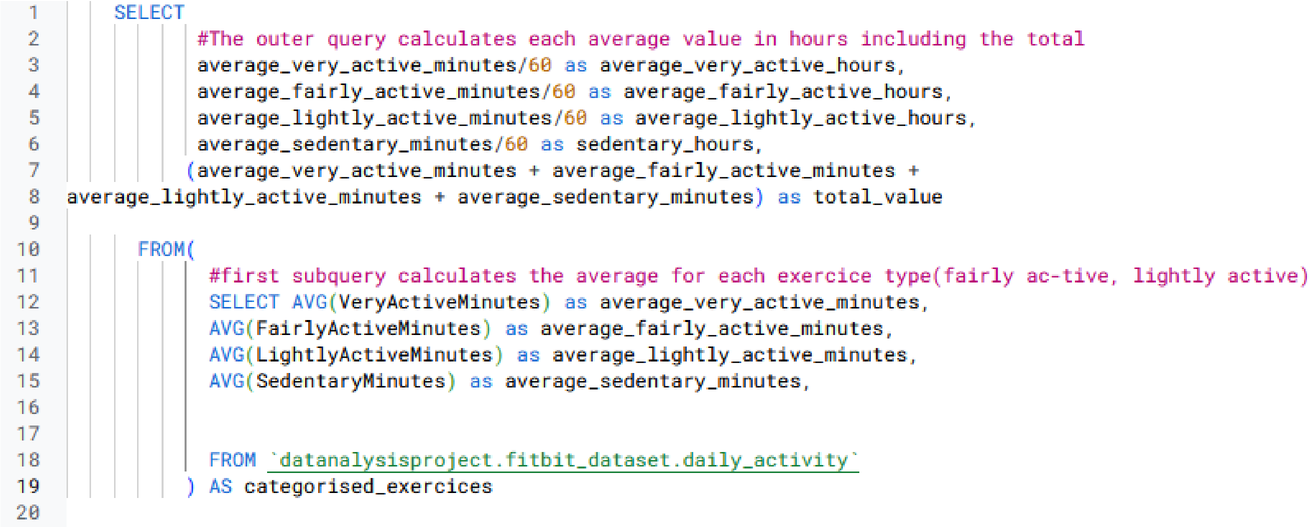
Correlation between sedentary time in minutes and calories burned



Similarly to the previous graph, there's also a positive correlation between sedentary time and calories burned. It appears that users who had the most sedentary lifestyle burned the least number of calories.

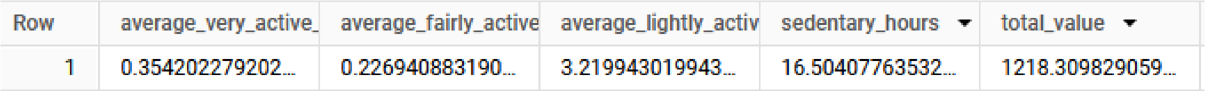
What's the average amount of time in hours spent on each exercice type?

To figure out the amount of time spent on each activity type (sedentary, light-active, fairly active, very active) the following query was used:



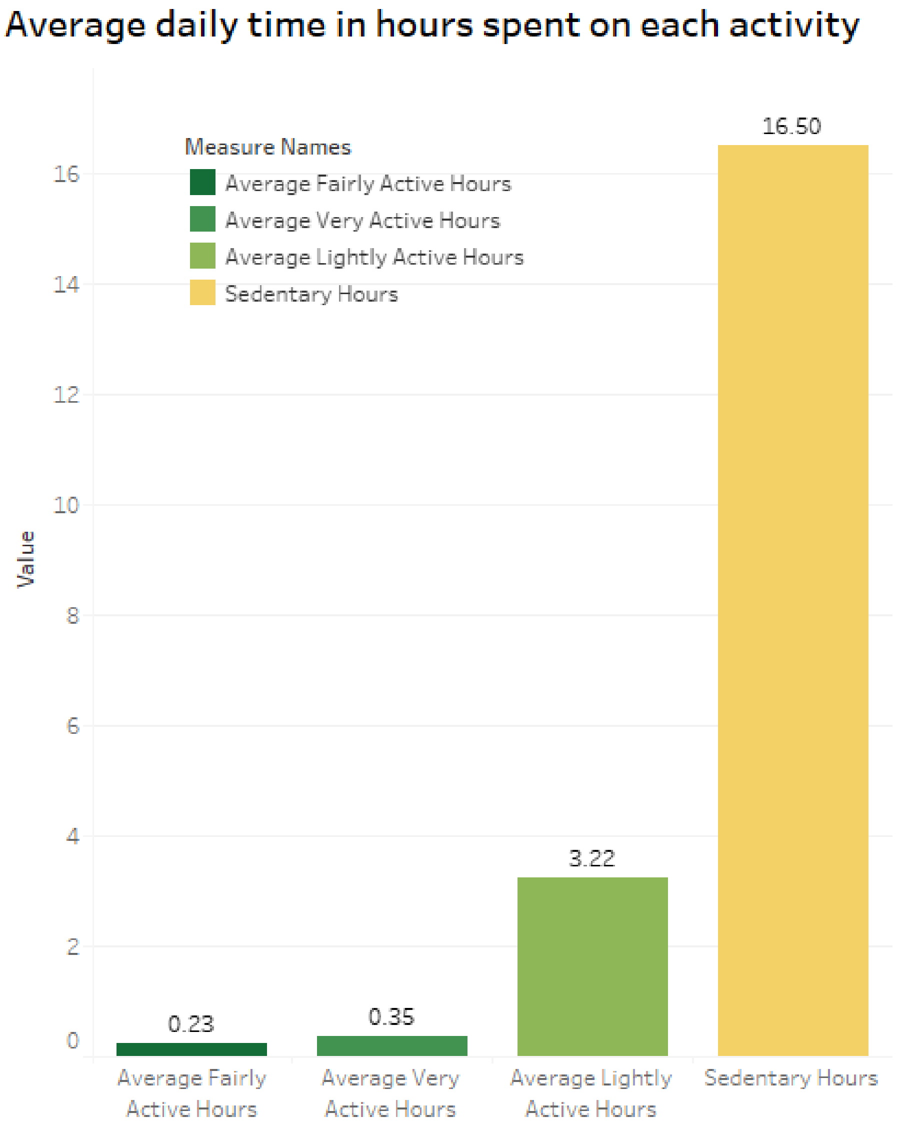
```
    SELECT
          #The outer query calculates each average value in hours including the total
          average_very_active_minutes/60 as average_very_active_hours,
          average_fairly_active_minutes/60 as average_fairly_active_hours,
          average_lightly_active_minutes/60 as average_lightly_active_hours,
          average_sedentary_minutes/60 as sedentary_hours,
         (average_very_active_minutes + average_fairly_active_minutes +
average_lightly_active_minutes + average_sedentary_minutes) as total_value

      FROM(
           #first subquery calculates the average for each exercice type(fairly ac-tive, lightly active)
           SELECT AVG(VeryActiveMinutes) as average_very_active_minutes,
           AVG(FairlyActiveMinutes) as average_fairly_active_minutes,
           AVG(LightlyActiveMinutes) as average_lightly_active_minutes,
           AVG(SedentaryMinutes) as average_sedentary_minutes,


           FROM `datanalysisproject.fitbit_dataset.daily_activity`
         ) AS categorised_exercices
```

This results in the following table with the average time spent on each activity type for all users:



Row	average_very_active_	average_fairly_active	average_lightly_activ	sedentary_hours ▼	total_value ▼
1	0.354202279202...	0.226940883190...	3.219943019943...	16.50407763532...	1218.309829059...

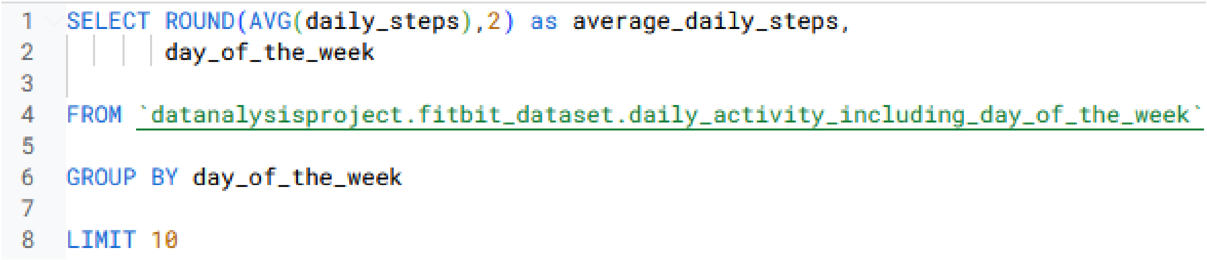
Then, we can generate the following graph based on the above data:



On average, users spent 3.22 hours engaged in light exercises, while the majority of their time, approximately 16.50 hours, was spent in sedentary activities

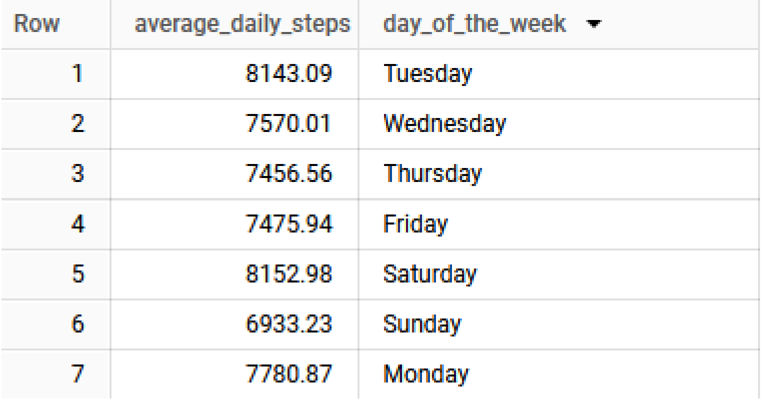
Is there a correlation between daily steps and days of the week?

Do users do more steps during certain days of the week? For example, are they most active on the weekend or during the week?

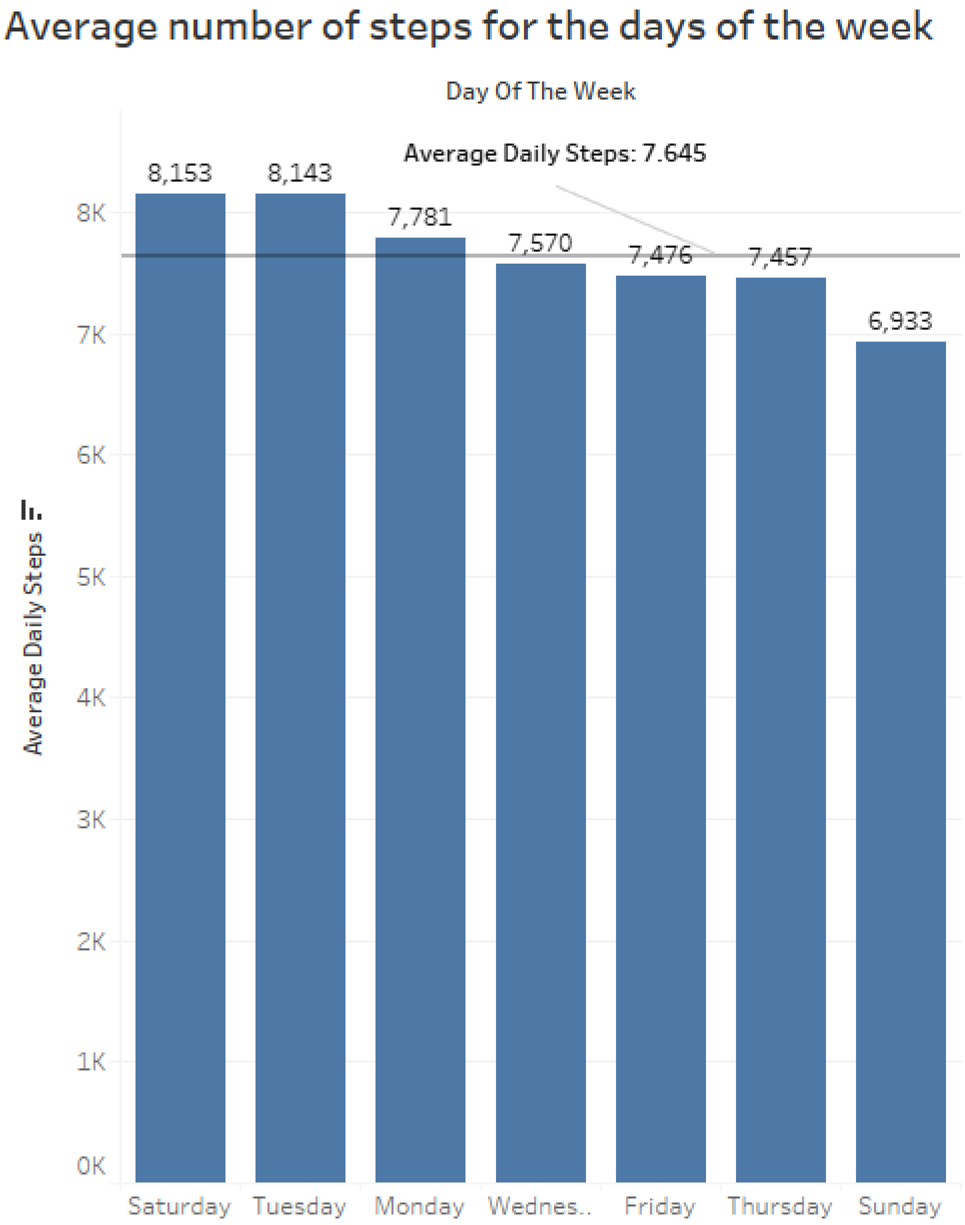


```
SELECT ROUND(AVG(daily_steps),2) as average_daily_steps,
       day_of_the_week

FROM `datanalysisproject.fitbit_dataset.daily_activity_including_day_of_the_week`

GROUP BY day_of_the_week

LIMIT 10
```

A table is generated which groups the average daily steps by day of the week:



Row	average_daily_steps	day_of_the_week ▾
1	8143.09	Tuesday
2	7570.01	Wednesday
3	7456.56	Thursday
4	7475.94	Friday
5	8152.98	Saturday
6	6933.23	Sunday
7	7780.87	Monday



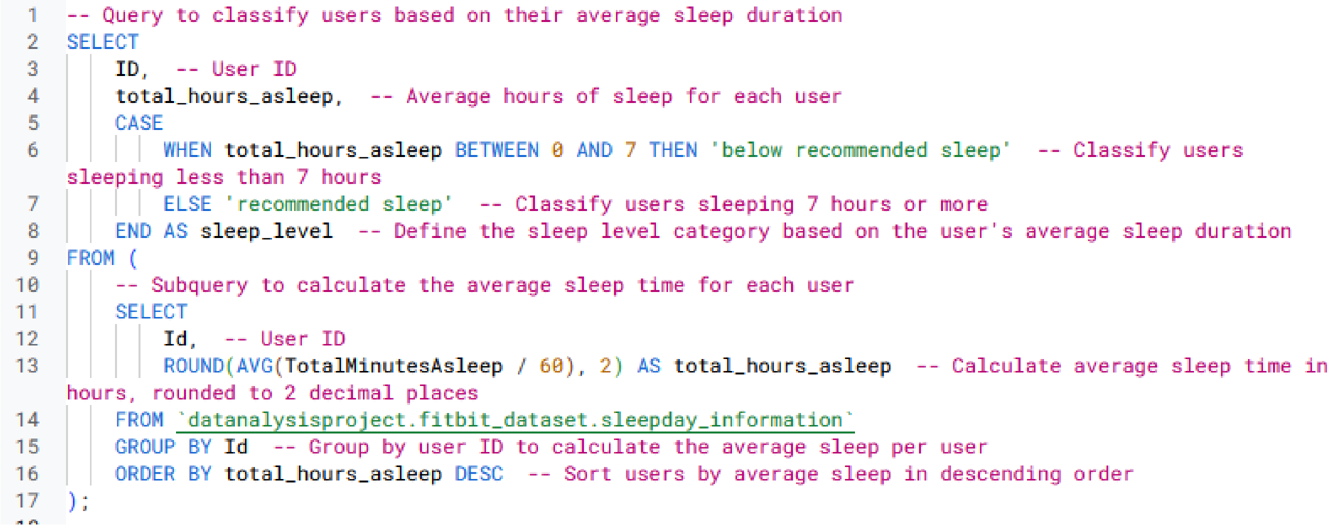
Average number of steps for the days of the week

Users walked the most on Saturdays, averaging 8,153 steps, while on Sundays the average dropped to 6,933 steps. This could be due to people typically taking Sundays to relax.

How many hours a night do people sleep on average?

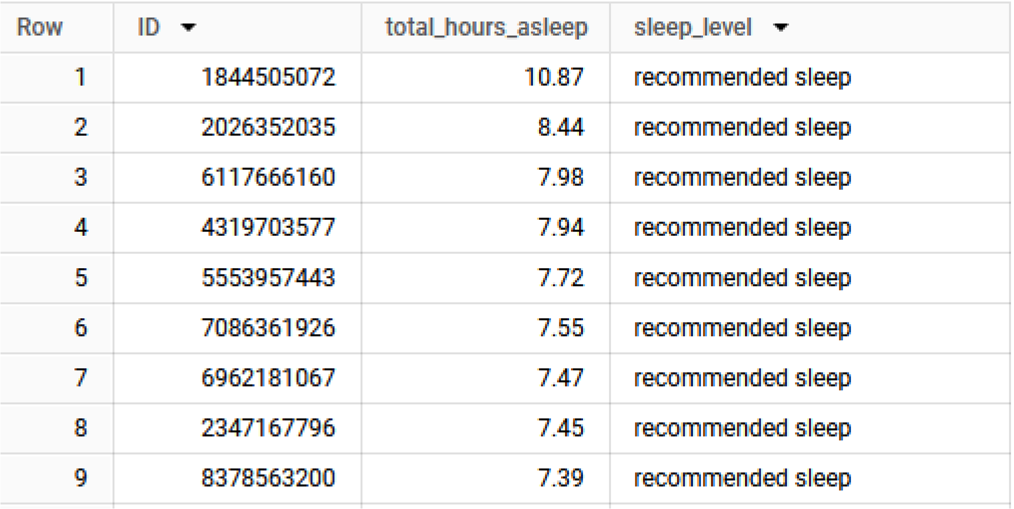
According to this document [3], “Adults should sleep 7 or more hours per night on a regular basis to promote optimal health. Sleeping less than 7 hours per night on a regular basis is associated with adverse health outcomes, including weight gain and obesity, diabetes, hypertension, heart disease and stroke, depression, and increased risk of death.”

We can get the sleep pattern of users via the following query:

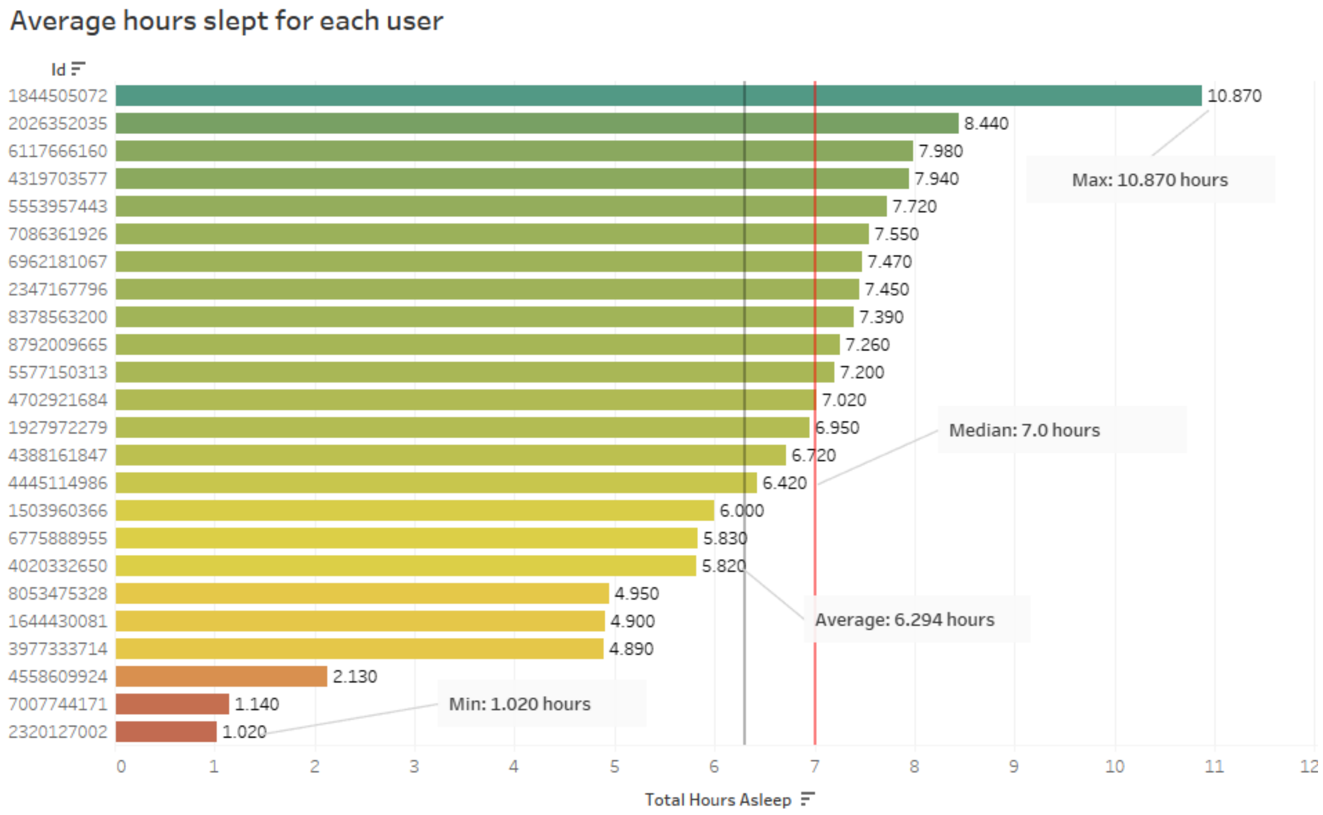


```sql
-- Query to classify users based on their average sleep duration
SELECT
    ID,  -- User ID
    total_hours_asleep,  -- Average hours of sleep for each user
    CASE
        WHEN total_hours_asleep BETWEEN 0 AND 7 THEN 'below recommended sleep'  -- Classify users sleeping less than 7 hours
        ELSE 'recommended sleep'  -- Classify users sleeping 7 hours or more
    END AS sleep_level  -- Define the sleep level category based on the user's average sleep duration
FROM (
    -- Subquery to calculate the average sleep time for each user
    SELECT
        Id,  -- User ID
        ROUND(AVG(TotalMinutesAsleep / 60), 2) AS total_hours_asleep  -- Calculate average sleep time in hours, rounded to 2 decimal places
    FROM `datanalysisproject.fitbit_dataset.sleepday_information`
    GROUP BY Id  -- Group by user ID to calculate the average sleep per user
    ORDER BY total_hours_asleep DESC  -- Sort users by average sleep in descending order
);
```

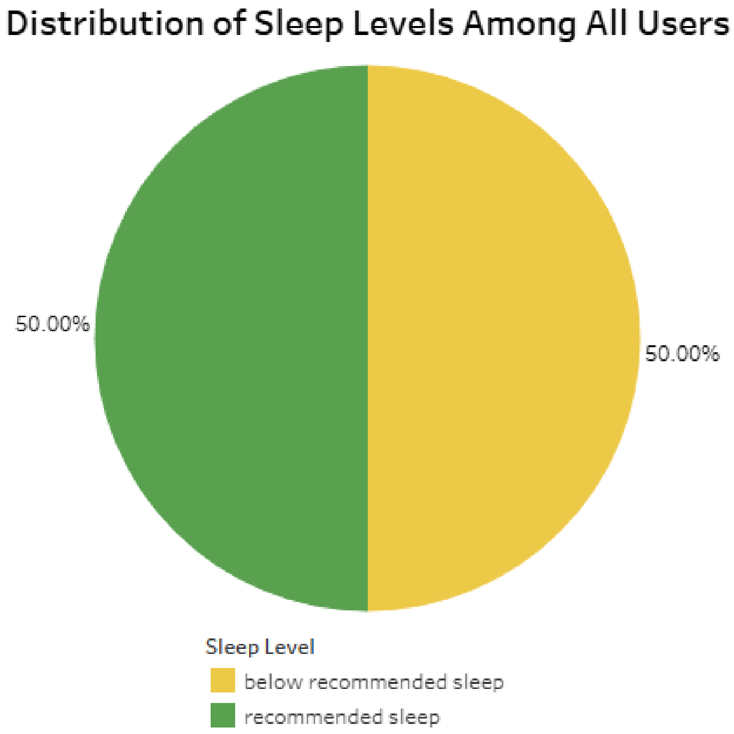
This results in the following table:



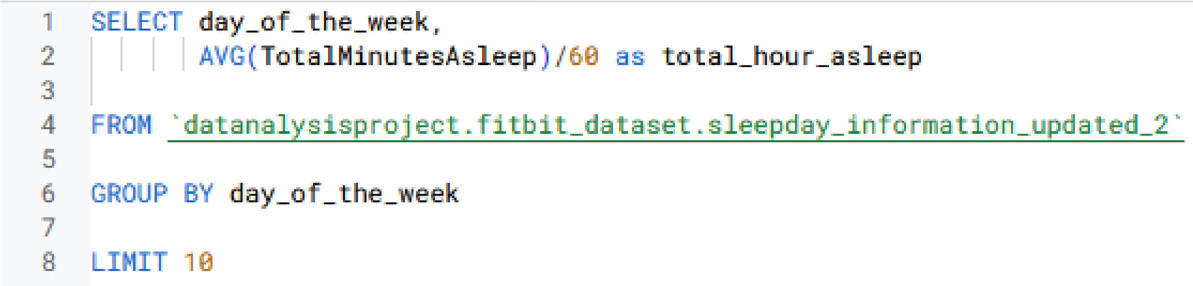
Row	ID ▾	total_hours_asleep	sleep_level ▾
1	1844505072	10.87	recommended sleep
2	2026352035	8.44	recommended sleep
3	6117666160	7.98	recommended sleep
4	4319703577	7.94	recommended sleep
5	5553957443	7.72	recommended sleep
6	7086361926	7.55	recommended sleep
7	6962181067	7.47	recommended sleep
8	2347167796	7.45	recommended sleep
9	8378563200	7.39	recommended sleep



The maximum sleep duration was approximately 10.87 hours, while the minimum was around 1 hour. On average, people slept about 6.3 hours. There is a roughly equal number of people getting sufficient sleep and those who are not:

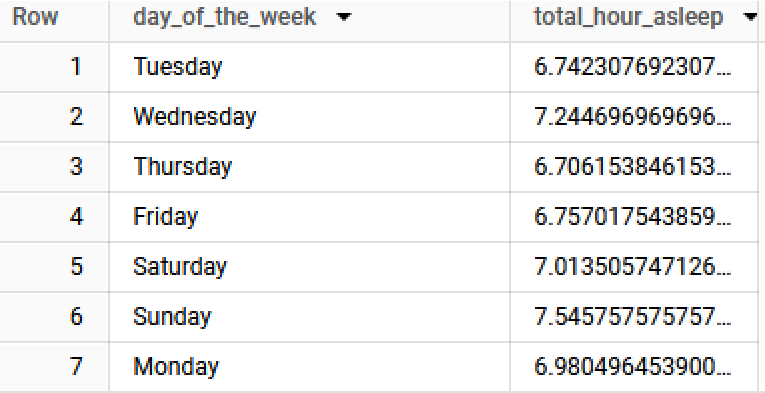


Is there a correlation between daily steps and days of the week?

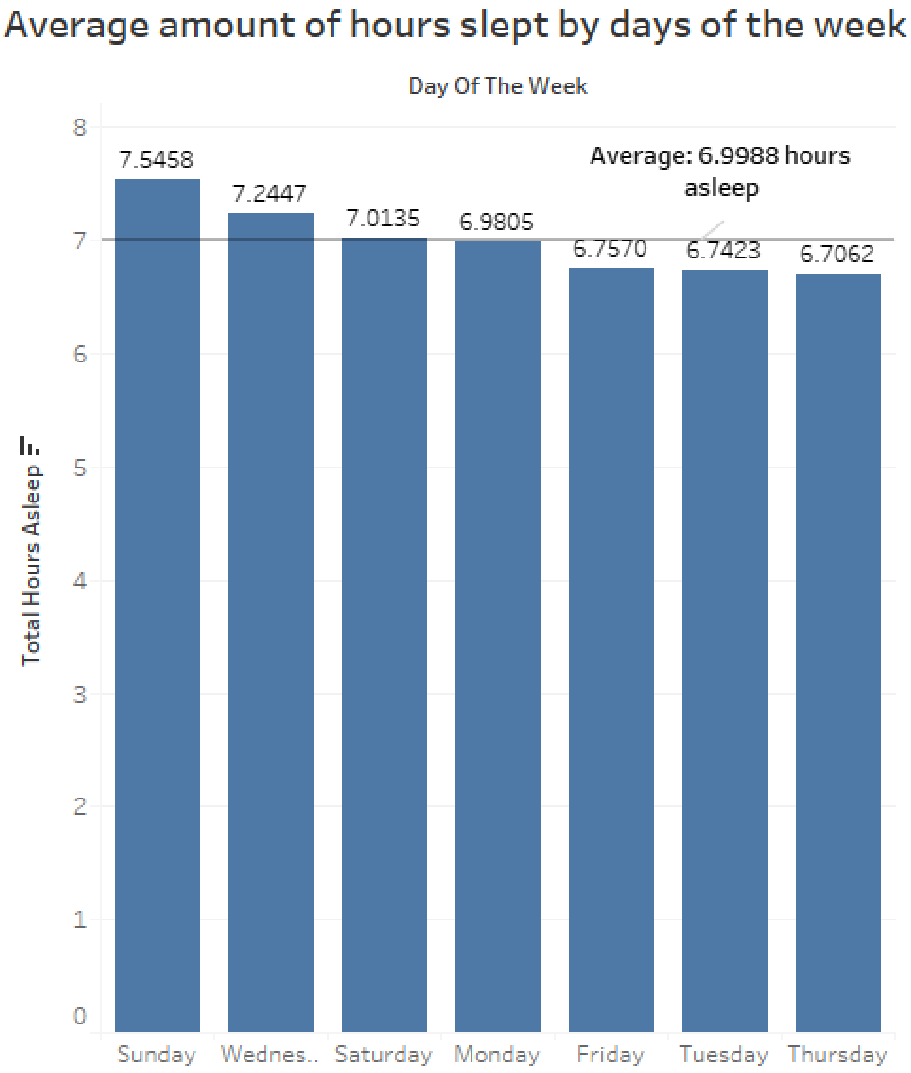


```sql
SELECT day_of_the_week,
       AVG(TotalMinutesAsleep)/60 as total_hour_asleep

FROM `datanalysisproject.fitbit_dataset.sleepday_information_updated_2`

GROUP BY day_of_the_week

LIMIT 10
```

The above query results in the following table:



Row	day_of_the_week	total_hour_asleep
1	Tuesday	6.742307692307...
2	Wednesday	7.244696969696...
3	Thursday	6.706153846153...
4	Friday	6.757017543859...
5	Saturday	7.013505747126...
6	Sunday	7.545757575757...
7	Monday	6.980496453900...



Users slept the most on Sundays, averaging 7.5458 hours, while on Thursday the average dropped to 6,7062 hours.

Conclusions and recommendations

- There is a balanced distribution of activity levels among users, with some being very active while others lead a more sedentary lifestyle, based on daily step counts
- The analysis reveals a positive correlation between daily steps and calories burned. While increasing step count generally results in higher calorie burn, high-intensity exercises lead to even greater energy expenditure. Users with sedentary habits burn fewer calories overall.
- Globally, users spend 80% of their time in sedentary activities, with only 16% of their time dedicated to light exercises. This highlights a significant portion of inactivity.
- Most active days are Saturdays with the least active days being on Sundays as highlighted by the average daily steps taken by users
- There is a roughly an equal number of people getting sufficient sleep and those who are not. Only 50% of people slept the minimum recommended amount of time of 7 hours. Furthermore, users sleep the most on Sundays and the least on Thursday.

References

- My main recommendation is for Bellabeat to continue promoting their products as valuable tools for tracking essential metrics which can help individuals maintain a healthy lifestyle. The results of this study can be shared to the public to show a general lack of physical activity. It’s important to highlight the negative effects and potential risks caused by sedentary lifestyles, while emphasising that these products can help develop healthier habits
- Furthermore, it would be interesting to communicate that, people do not sleep enough which can contribute to health issues. Bellabeat could introduce features in their devices that can help improve and monitor sleep quality such as sleep tracking reminder and sleep tips.
- Have a rewards system that would give some kind of prize for people who increase their activity level by regularly walking more steps or doing more High intense exercises.