# How Can Bellabeat, A Wellness Technology Company Play It Smart?

The analysis follows the 6 steps of Data Analysis taught in the Google course: Ask, Prepare, Process, Analyse, Share and Act.

**Step 1: Ask**

In this step, we define the problem and objectives of our case study and its desired outcome.

**1.0 Background**

Bellabeat is a high-tech manufacturer of beautifully-designed health-focused smart products for women since 2013. Inspiring and empowering women with knowledge about their health and habits, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for females.

The co-founder and Chief Creative Officer, Urška Sršen is confident that an analysis of non-Bellabeat consumer data (ie. FitBit fitness tracker usage data) would reveal more growth opportunities.

**1.2 Business Task:**

Analyze FitBit fitness tracker data to gain insights into how consumers are using the FitBit app and discover trends for Bellabeat marketing strategy.

**1.3 Business Objectives:**

* What are the trends identified?
* How could these trends apply to Bellabeat customers?
* How could these trends help influence Bellabeat marketing strategy?

**1.4 Deliverables:**

* A clear summary of the business task
* A description of all data sources used
* Documentation of any cleaning or manipulation of data
* A summary of the analysis
* Supporting visualizations and key findings
* High-level content recommendations based on the analysis

**1.5 Key Stakeholders:**

* **Urška Sršen:** Bellabeat’s co-founder and Chief Creative Officer.
* **Sando Mur:** Mathematician, Bellabeat’s co-founder, and a key member of the Bellabeat executive team.
* **Bellabeat marketing analytics team:** A team of data analysts guiding Bellabeat’s marketing strategy.

# STEP 2: PREPARE

In the Prepare phase, we identify the data being used and its limitations.

**2.1 Information on Data Source:**

* Data is publicly available on [Kaggle: FitBit Fitness Tracker Data](https://www.kaggle.com/arashnic/fitbit) and stored in 18 CSV files.
* Generated by respondents from a survey via Amazon Mechanical Turk between 12 March 2016 to 12 May 2016.
* 30 FitBit users consented to the submission of personal tracker data.
* Data collected includes physical activity recorded in minutes, heart rate, sleep monitoring, daily activity, and steps.

**2.2 Limitations of Data Set:**

* Data is collected 5 years ago in 2016. Users’ daily activity, fitness and sleeping habits, diet, and food consumption may have changed since then. Data may not be timely or relevant.
* Sample size of 30 FitBit users is not representative of the entire fitness population.
* As data is collected in a survey, we are unable to ascertain its integrity or accuracy.

**2.3 Is Data ROCCC?**

A good data source is ROCCC which stands for Reliable, Original, Comprehensive, Current, and Cited.

* Reliable — LOW — Not reliable as it only has 30 respondents
* Original — LOW — Third-party provider (Amazon Mechanical Turk)
* Comprehensive — MED — Parameters match most of the Bellabeat products’ parameters
* Current — LOW — Data is 5 years old and may not be relevant
* Cited — LOW — Data collected from the third party, hence unknown

Overall, the dataset is considered bad quality data and it is not recommended to produce business recommendations based on this data.

**2.4 Data Selection**

The following file is selected and copied for analysis.

dailyActivity\_merged.csv

* 1. **Tool**

We are using Python for data cleaning, transformation, and visualization.

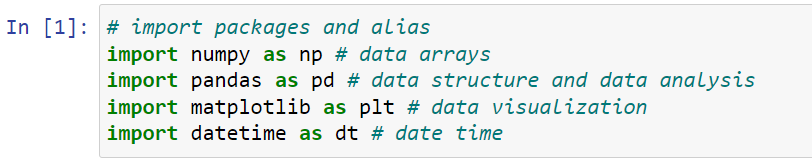
# STEP 3: PROCESS

Here, we will process the data by cleaning and ensuring that it is correct, relevant, complete, and free of error and outlier by performing:

* Explore and observe data
* Check for and treat missing or null values
* Transform data — format data type
* Perform preliminary statistical analysis

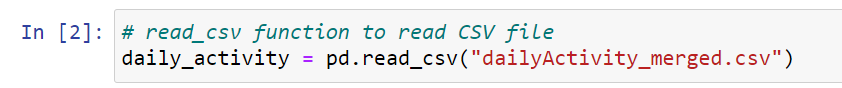
**3.1 Preparing the Environment**

The Python libraries are installed and aliased for easy reading.

****

**3.2 Importing data set**

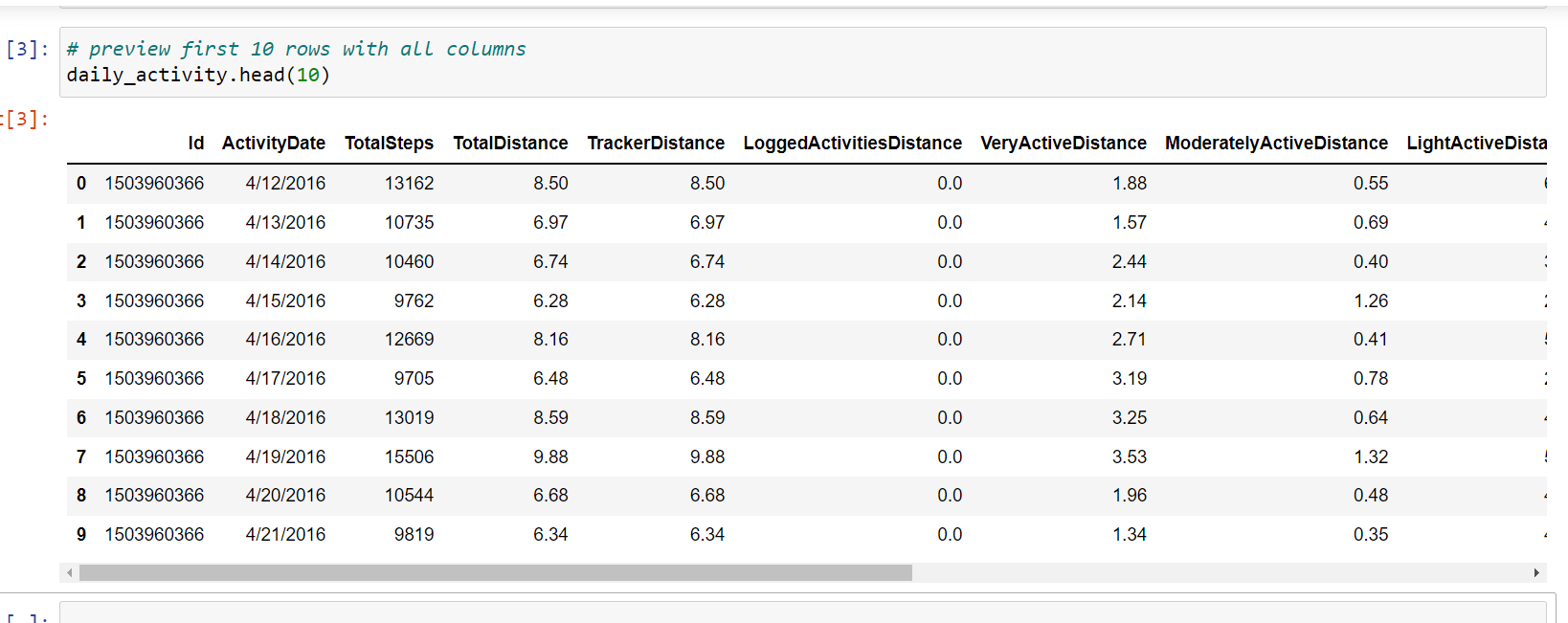
Reading in the selected file.

****

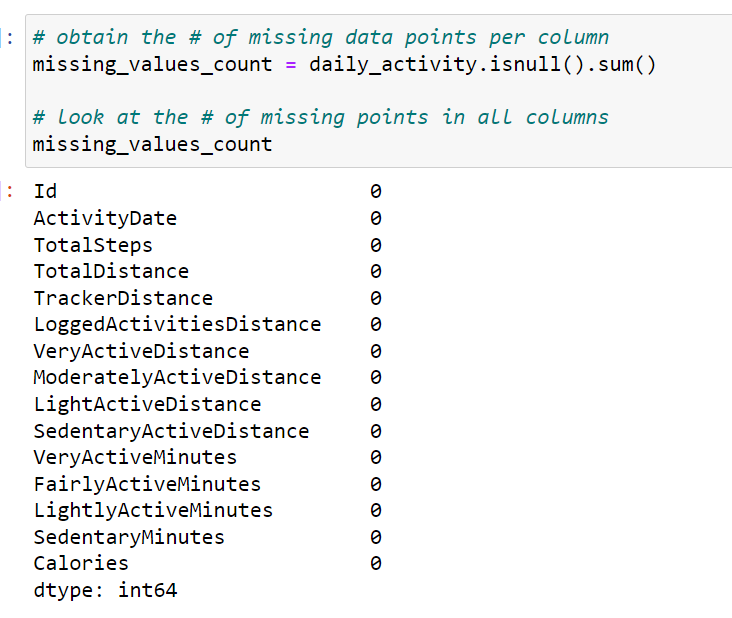
**3.3 Data cleaning and manipulation**

* Observe and familiarize with data
* Check for null or missing values
* Perform sanity check of data

Previewing the first 10 rows to familiarise yourself with the data.

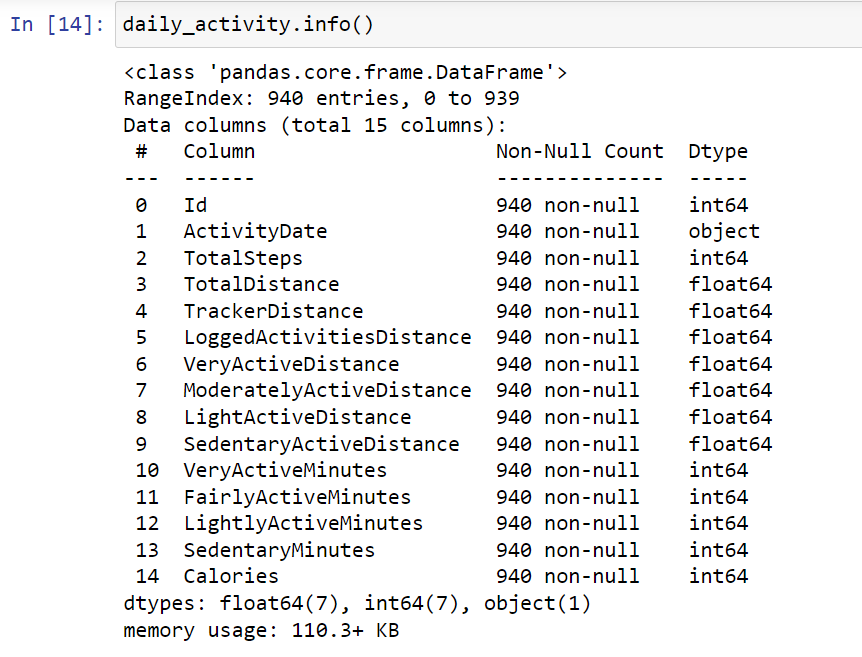
****

Then, I’m going to find out whether there are any null or missing values in the data.

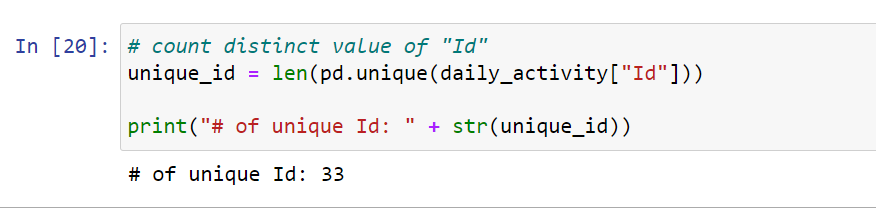


Finding out the basic information of the data:

* no. of rows and columns
* columns names
* non-null count
* data type



We are also going to count unique IDs to confirm whether data has 30 IDs as claimed by the survey.



From the above observations, we noted that

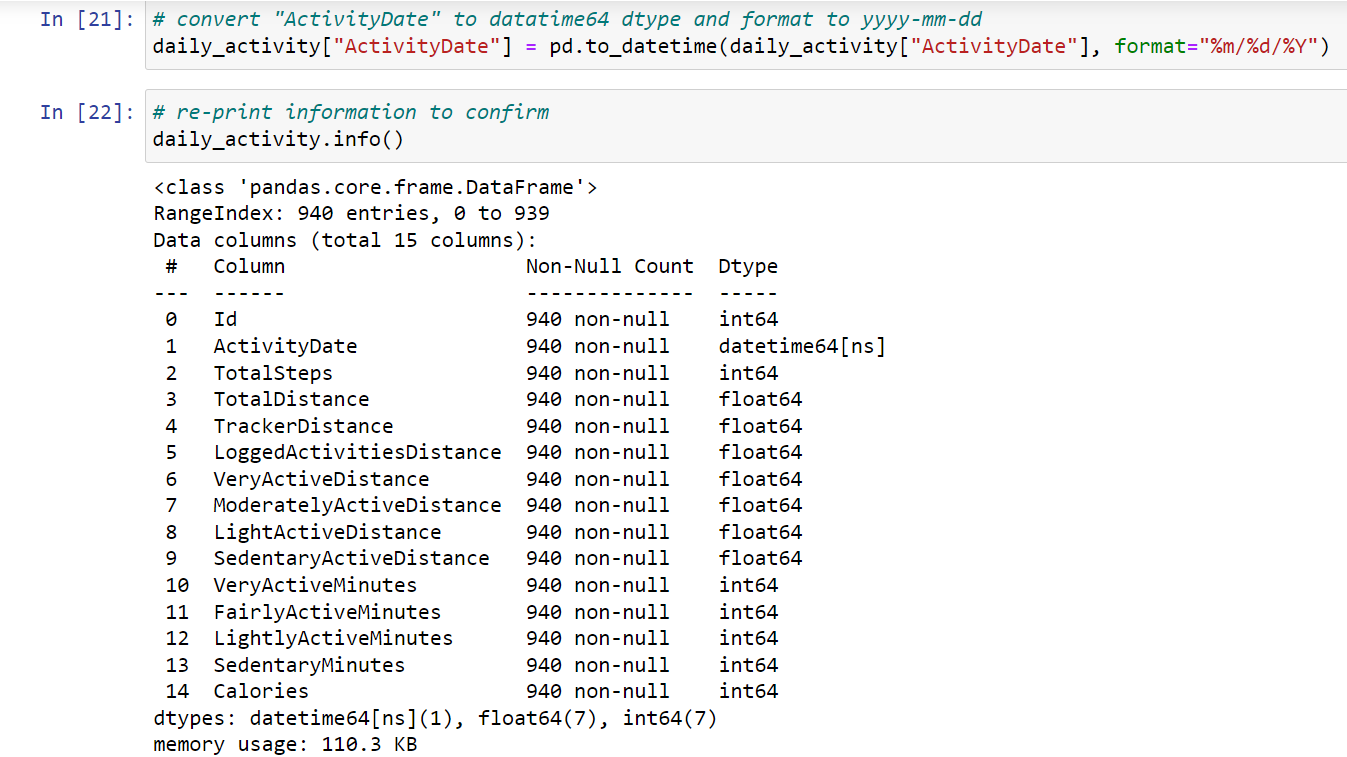
* There are no Null or missing values as stated under the ‘Non-Null Count’.
* Data has 15 columns and 940 rows.
* ActivityDate is wrongly classified as object dtype and has to be converted to datetime64 dtype.
* There are 33 unique IDs, instead of 30 unique IDs as expected. Some users may have created additional IDs during the survey period.

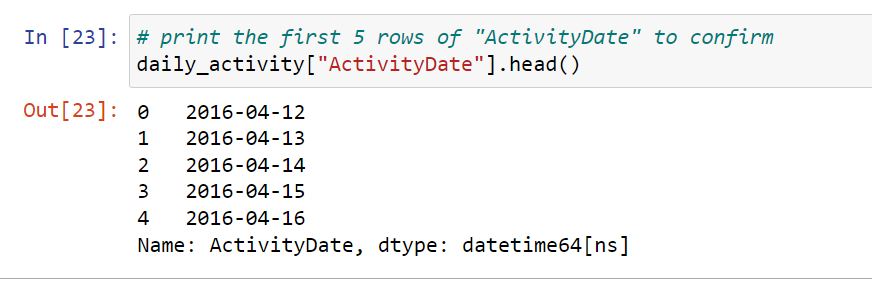
Now that we have identified the dirty data, we will then perform data manipulation/transformation.

* Convert ActivityDate to datatime64 dtype.
* Convert the format of ActivityDate to yyyy-mm-dd.
* Create a new column DayOfTheWeek by generating date in the form of the day of the week for further analysis.
* Create new column TotalMins being the sum of VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes and SedentaryMinutes.
* Create new column TotalHours by converting the new column TotalMins in no. 4 to the number of hours.
* Rearrange and rename columns.

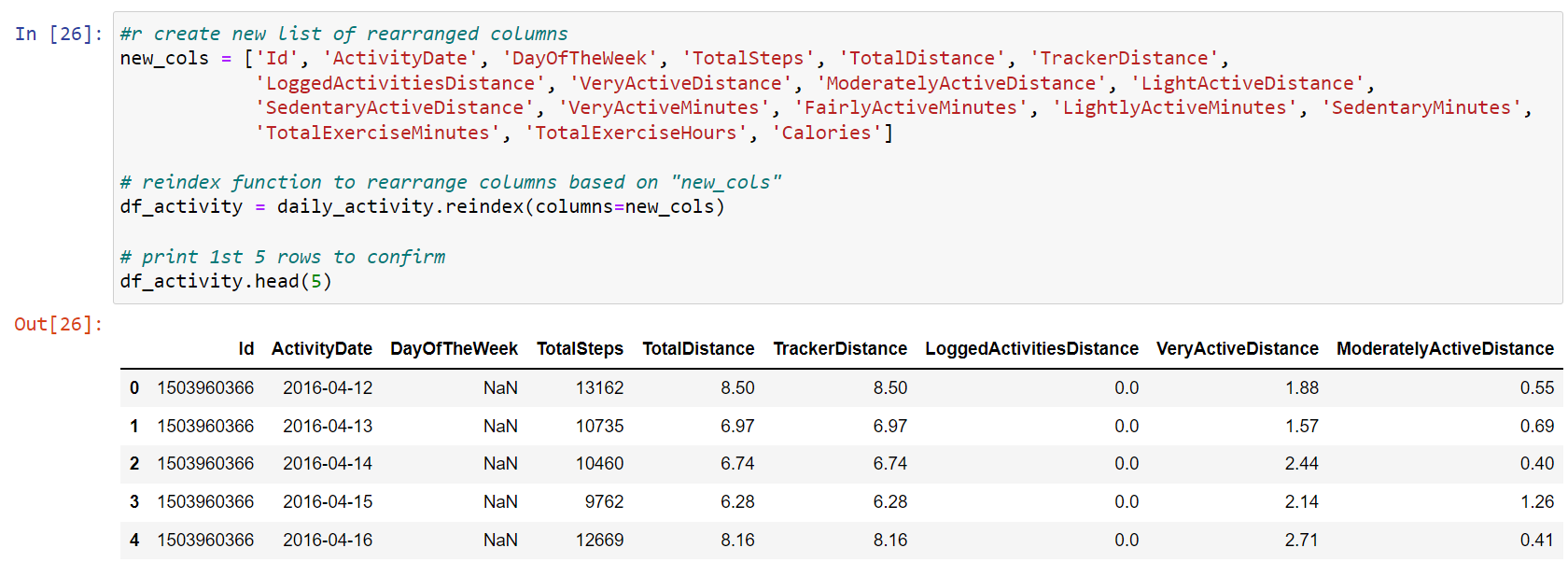
First off, we will convert ActivityDate from object to datatime64 dtype and then, set out to convert ActivityDate to yyyy-mm-dd.

Then, we confirm whether it has been updated to datatime64 dtype and ActivityDate to yyyy-mm-dd.

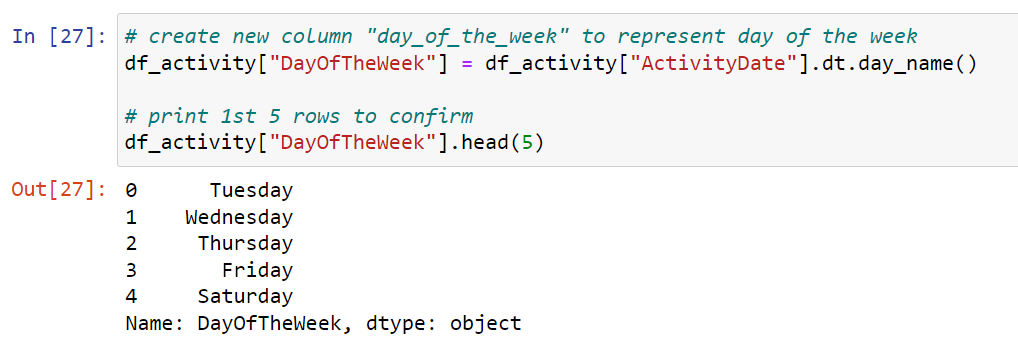




Creating a new list with rearranged column names and renaming daily\_activity to a shorter name df\_activity.



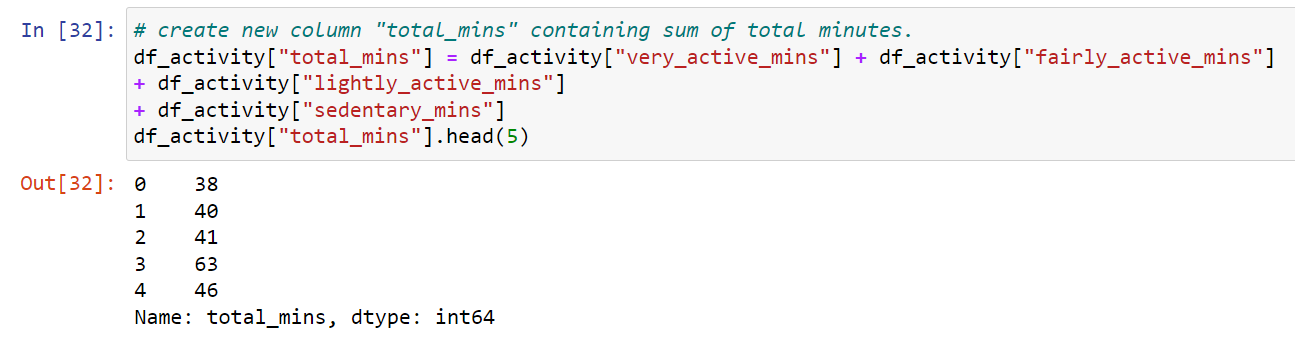
Create a new column by separating the date into a day of the week for further analysis.

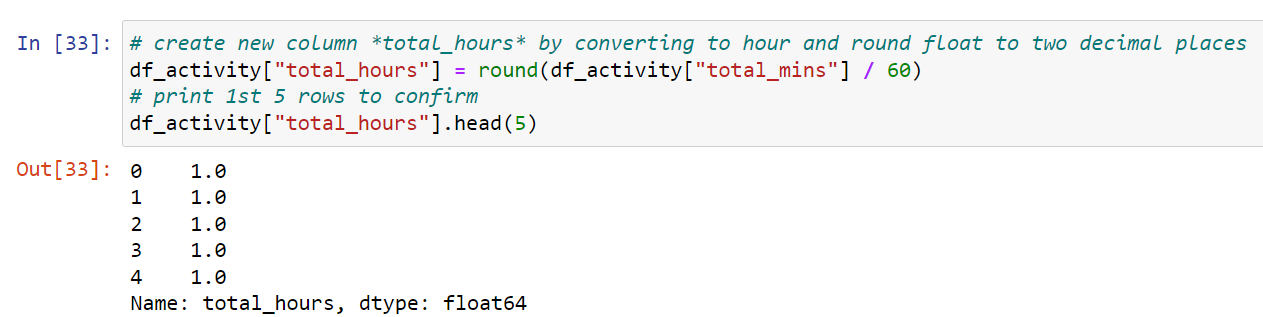


Rearranging and renaming columns from XxxYyy to xxx\_yyy.



Creating new column total\_mins being the sum of total time logged.

Creating a new column by converting total\_mins to several hours.



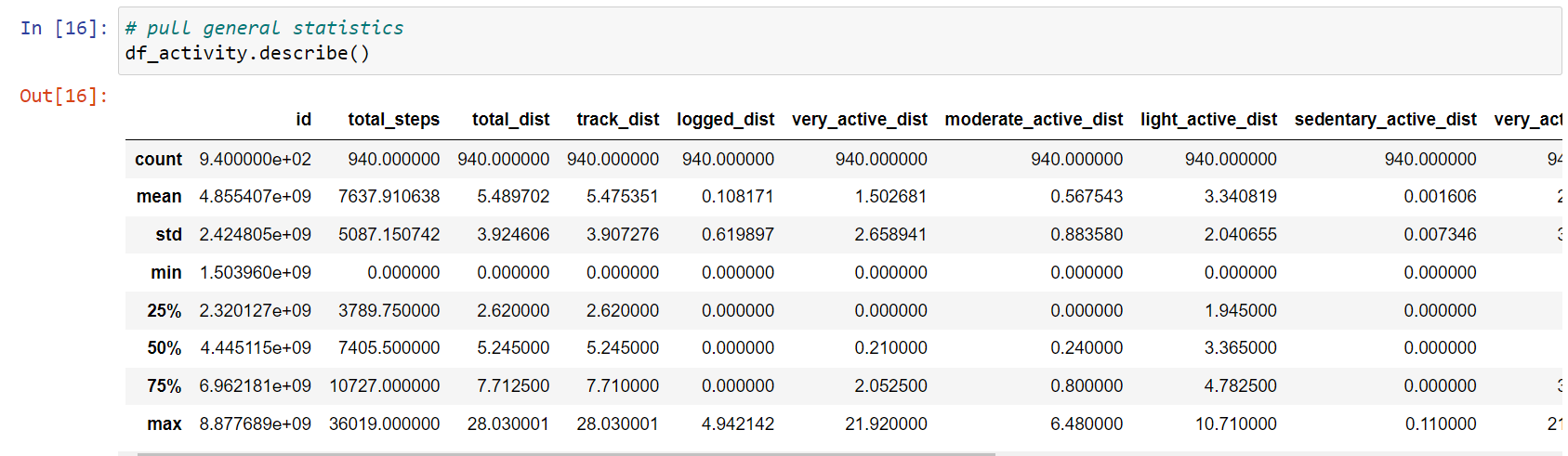
Data cleaning and manipulation are completed. Hence, data is now ready to be analyzed.

# STEP 4: ANALYSE

**4.1 Perform calculations**

Pulling statistics for analysis:

* count — no. of rows
* mean (average)
* std (standard deviation)
* min and max
* percentiles 25%, 50%, 75%



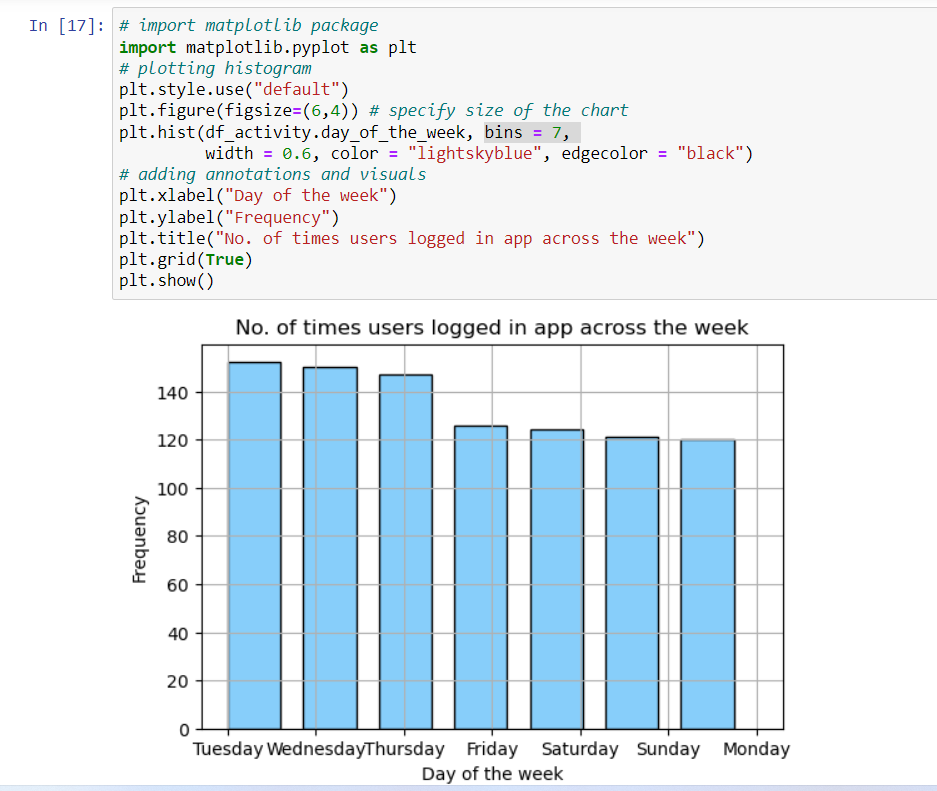
Interpreting statistical findings:

* On average, users logged 7,637 steps or 5.4km, which is inadequate. As recommended by CDC, an adult female has to aim for at least 10,000 steps or 8km per day to benefit from general health, weight loss, and fitness improvement. [Source: Medical News Today article](https://www.medicalnewstoday.com/articles/how-many-steps-should-you-take-a-day)
* Sedentary users are the majority logging on average 991 minutes or 20 hours making up 81% of total average minutes.
* Noting that the average calories burned are 2,303 calories equivalent to 0.6 pounds. Could not interpret in detail as calories burned to depend on several factors such as the age, weight, daily tasks, exercise, hormones and daily calorie intake. [Source: Health Line article](https://www.healthline.com/health/fitness-exercise/how-many-calories-do-i-burn-a-day#Burning-calories)

# STEP 5: SHARE

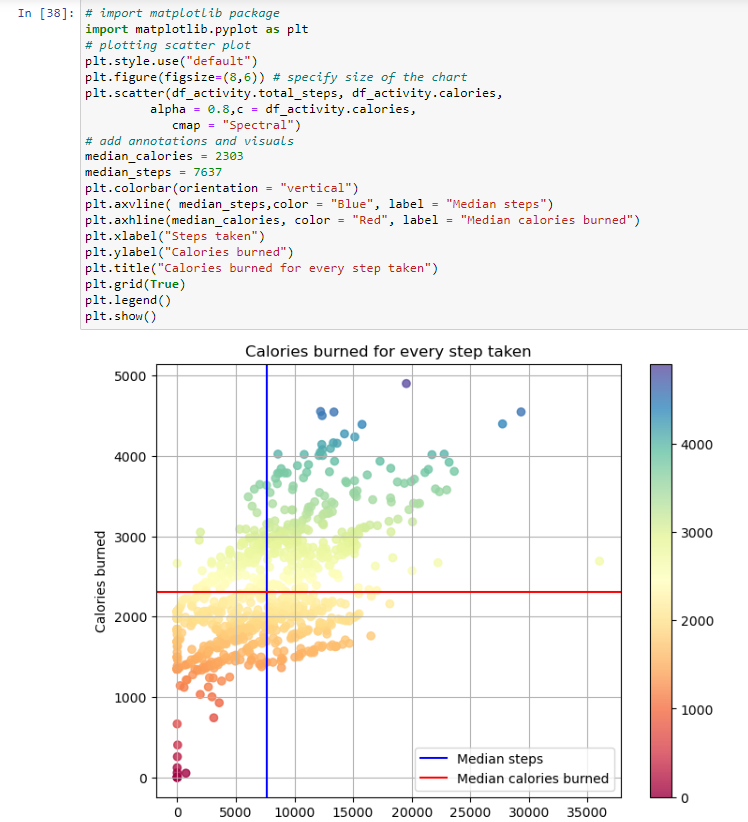
In this step, we are creating visualizations and communicating our findings based on our analysis.

**5.1 Data Visualization and Findings**

****

**Frequency of usage across the week**

* In this histogram, we are looking at the frequency of FitBit app usage in terms of days of the week.
* We discovered that users prefer or remember (giving them the doubt of benefit that they forgotten) to track their activity on the app during midweek from Tuesday to Friday.
* Noting that the frequency dropped on Friday and continue on weekends and Monday.

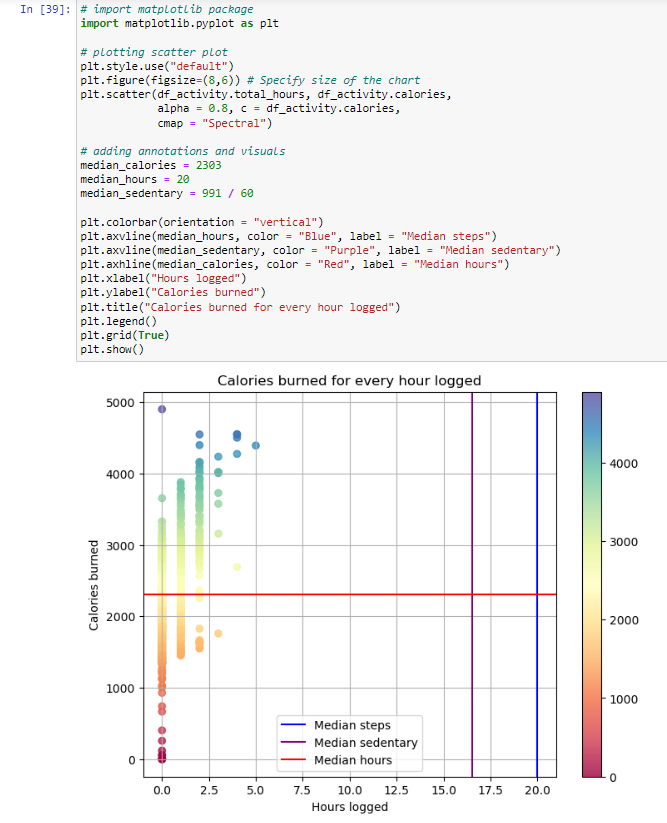
****

**Calories burned for every step taken**

From the scatter plot, we discovered that:

* It is a positive correlation.
* We observed that intensity of calories burned increase when users are at the range of > 0 to 15,000 steps with calories burn rate cooling down from 15,000 steps onwards.
* Noted a few outliers:

1. Zero steps with zero to minimal calories burned.
2. 1 observation of > 35,000 steps with < 3,000 calories burned.
3. Deduced that outliers could be due to natural variation of data, change in user’s usage or errors in data collection (ie. miscalculations, data contamination or human error).

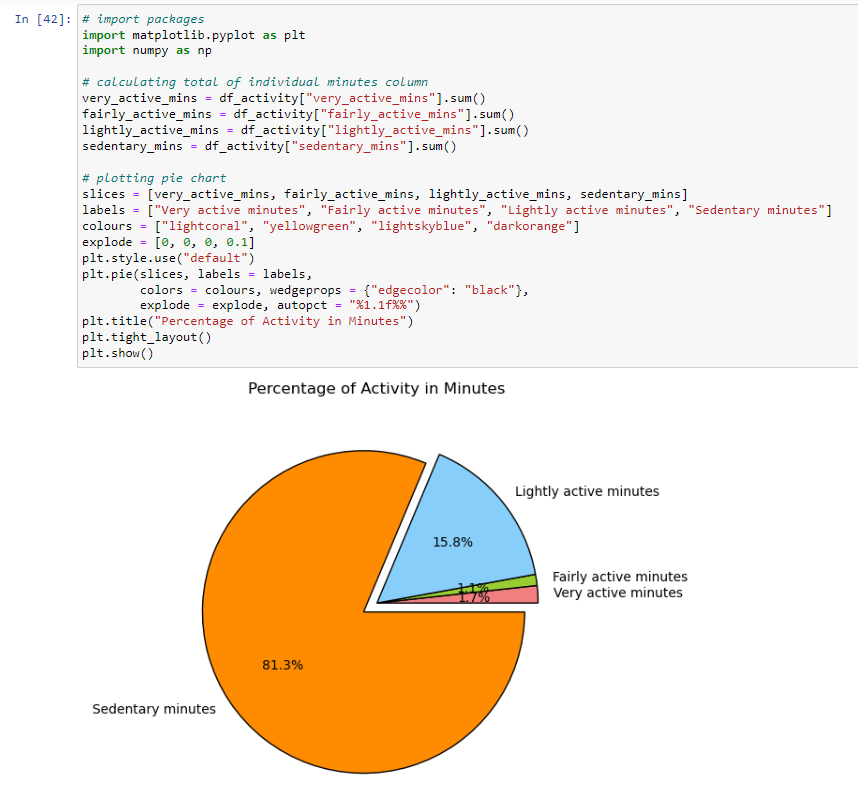


**Calories burned for every hour logged**

The scatter plot is showing:

* A weak positive correlation whereby the increase of hours logged does not translate to more calories being burned. That is largely due to the average sedentary hours (purple line) plotted at the 16 to 17 hours range.
* Again, we can see a few outliers:

1. The same zero value outliers
2. An unusual red dot at the 24 hours with zero calories burned which may be due to the same reasons as above.



**Percentage of Activity in Minutes**

As seen from the pie chart,

* Sedentary minutes takes the biggest slice at 81.3%.
* This indicates that users are using the FitBit app to log daily activities such as daily commute, inactive movements (moving from one spot to another) or running errands.
* App is rarely being used to track fitness (ie. running) as per the minor percentage of fairly active activity (1.1%) and very active activity (1.7%). This is highly discouraging as FitBit app was developed to encourage fitness.

# STEP 6: ACT

In the final step, we will be delivering our insights and providing recommendations based on our analysis.

Here, we revisit our business questions and share with you our high-level business recommendations.

**1. What are the trends identified?**

* Majority of users (81.3%) are using the FitBit app to track sedentary activities and not using it for tracking their health habits.
* Users prefer to track their activities during weekdays as compared to weekends - perhaps because they spend more time outside on weekdays and stay in on weekends.

**2. How could these trends apply to Bellabeat customers?**

Both companies develop products focused on providing women with their health, habit and fitness data and encouraging them to understand their current habits and make healthy decisions. These common trends surrounding health and fitness can very well be applied to Bellabeat customers.

**3. How could these trends help influence Bellabeat marketing strategy?**

* Bellabeat marketing team can encourage users by educating and equipping them with knowledge about fitness benefits, suggest different types of exercise (ie. simple 10 minutes exercise on weekdays and a more intense exercise on weekends), and calorie intake and burnt rate information on the Bellabeat app.
* On weekends, the Bellabeat app can also prompt notifications to encourage users to exercise.

The dataset and complete code can be found [here](https://www.kaggle.com/katiehuangx/case-study-bellabeat-using-python).