

# Playing it Smart in the Wellness Industry with Data Science



Jesutimilehin Onayemi  
24th November, 2023

# Outline

- Summary
- Introduction
- Methodology
- Conclusion
- Recommendation
- Appendix

# Summary

- Exploratory Data Analysis with scatterplot.
- Exploratory Data Analysis with correlation.
- Model Development/Predictive Analysis

# Introduction

- Welcome to the Bellabeat case study! In this case study, I am taking on the role of a data scientist working for Bellabeat, a high-tech manufacturer of health-focused products for women.

# Introduction

- Scenario/ Problem Statement

I am working at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. I have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights I discover will then help guide marketing strategy for the company.

## Part 1

# Methodology

# Methodology

- Getting the data
- Data Wrangling
- Exploratory Data Analysis
- Model development and Predictive Analysis

# Getting the Data

FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius): This Kaggle data set contains personal fitness tracker from thirty fitbit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits.



# Getting the Data

- The data was downloaded from Kaggle:  
<https://www.kaggle.com/datasets/arashnic/fitbit>
- Some of the data are organized in wide format while others are in long format.
- The data source is: <https://zenodo.org/record/53894#.X9oeh3Uzaao>,
- License: CC0: Public Domain

# Data Wrangling

- Loaded all the eighteen csv files associated with this project into pandas dataframes in Python and viewed their contents using the head function. Also viewed some of the smaller sized csv files on Microsoft Excel.
- I discovered that the 'dailyActivity\_merged.csv' file contained merged data from other csv files in the project. In other words, it is a combination of some other data files associated with the project, so we could work with this 'dailyActivity\_merged' file, while not necessarily having to work with these other files. Other very distinct datasets, are 'weightlogInfo\_merged', and 'SleepDay\_merged'

# Data Wrangling

- After loading the dailyActivity\_merged file into a dataframe, I checked for the shape of the dataset.
- I then checked for null values in this dataset, there were none
- I checked the data types of the features in the dataset, and converted the 'Id' column to the correct data type
- I checked for the number of respondents in this data set, In this case, there were thirty-three respondents.
- One of the respondents only had four records and this was too inconsistent with the number of records of the other patients, so I deleted this respondent's records from the dataframe before analysis.

# Exploratory Data Analysis

The daily activity dataset consists of fifteen features. Most of these features are records of different activities performed by the respondent, how much of these activities they did or how long they did them. The last feature is 'Calories', which indicates the amount of calories the respondent burns per day. The aim of some people on some wellness apps and devices is to lose weight, keep in shape or stay fit, by burning calories the required amount of calories you can achieve these goals, So, I explored the relationship between certain features and the amount of calories burnt per day.

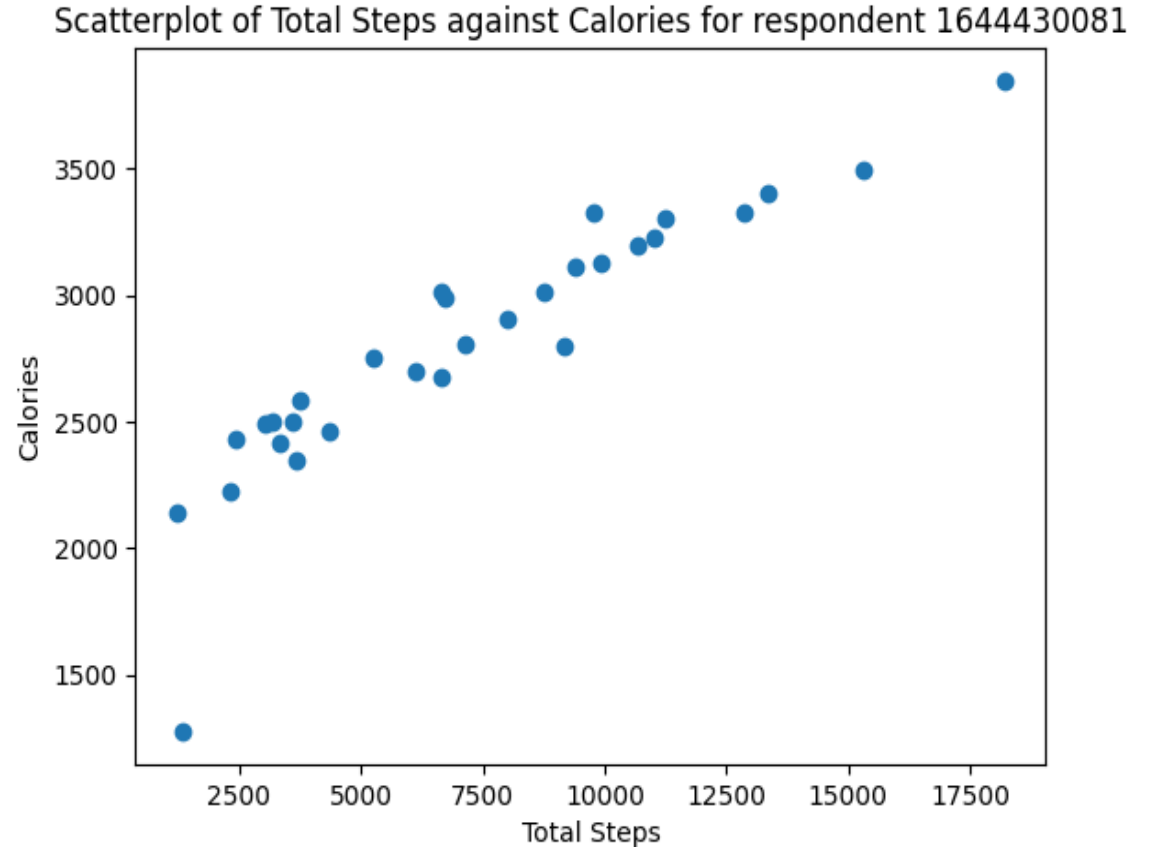
## Part 2

Insights drawn from the Exploratory Data Analysis

# Exploratory Data Analysis

- Total Steps vs Calories

The first one is Total Steps taken by the respondents per day. I used a scatter plot to visualize this feature with the corresponding amount of Calories burnt per day. The picture on this page shows one plot out of thirty-two for this category. This one is for the respondent with Id = 166443081.



# Exploratory Data Analysis

- Total Steps vs Calories

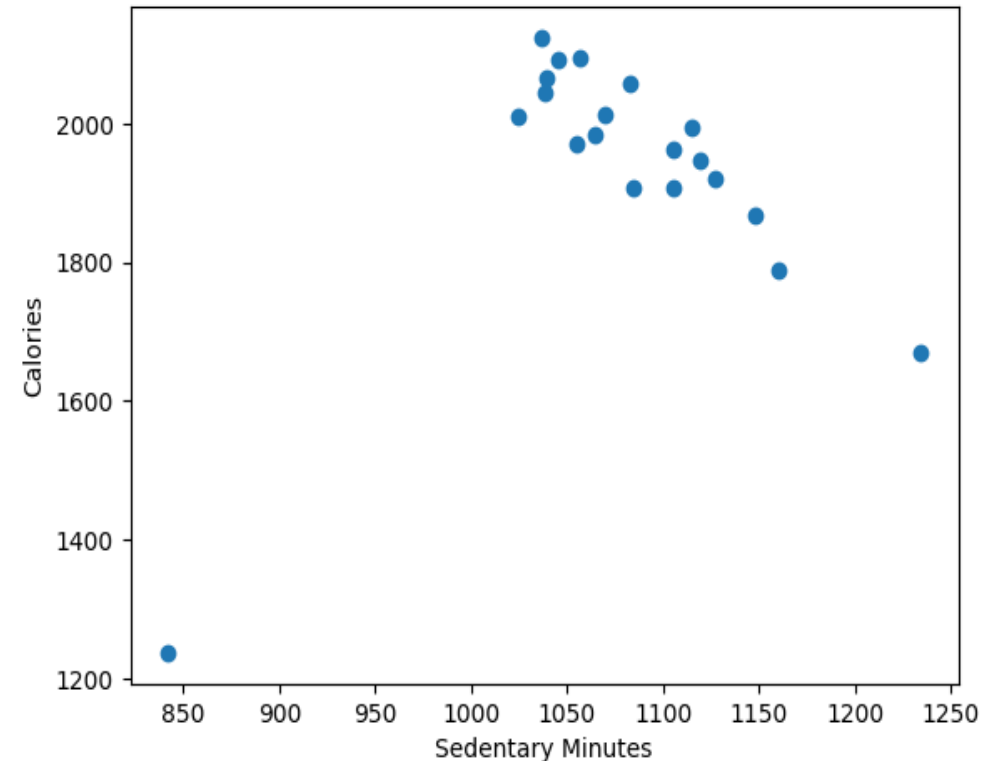
The rest of the plots in this section for the other respondents are more or less like this, indicating a positive linear relationship between the two entities measured.

# Exploratory Data Analysis

- Sedentary Minutes vs Calories

The second feature explored is, Total amount of time spent in Sedentary Activity per day. I used a scatter plot to visualize this feature with the corresponding amount of Calories burnt per day. The picture on this page shows one plot out of thirty-two for this category. This one is for the respondent with Id = 3372868164.

Scatterplot of Sedentary Minutes against Calories for respondent 3372868164





# Exploratory Data Analysis

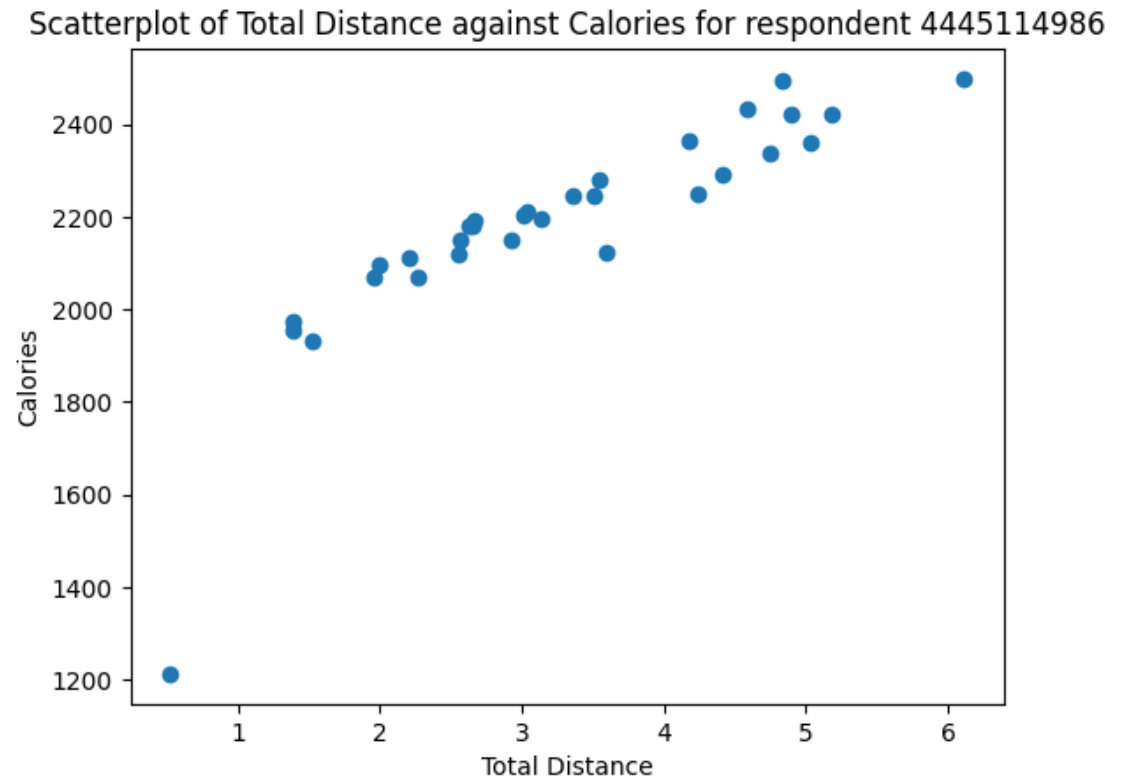
- Sedentary Minutes vs Calories

The relationship between Sedentary minutes and amount of calories burnt is not exactly clear from this set of visualizations, few are negatively related, while others are either in a circular or horizontal arrangement, which do not stipulate any relationship.

# Exploratory Data Analysis

- Total Distance vs Calories

The second feature explored is, Total Distance travelled in a day. I used a scatter plot to visualize this feature with the corresponding amount of Calories burnt per day. The picture on this page shows one plot out of thirty-two for this category. This one is for the respondent with Id = 444511496.



# Exploratory Data Analysis

- Total Distance vs Calories

To see the full python code and the results, follow the URL address:

[https://github.com/Jesutimilehin-Onayemi/Playing-it-Smart-in-the-Wellness-Industry-with-Data-Science/blob/main/dailyActivity\\_Analysis.ipynb](https://github.com/Jesutimilehin-Onayemi/Playing-it-Smart-in-the-Wellness-Industry-with-Data-Science/blob/main/dailyActivity_Analysis.ipynb)

# Exploratory Data Analysis

- Correlation

In order to view the linear relationship all other features have with the amount of calories burnt per day, I used the `corr()` function. The result is shown on the next page, we find out that, 'TotalDistance' 'TotalSteps', 'LightActiveDistance', 'VeryActiveDistance', 'VeryActiveMinutes' are the most correlated features to the amount of Calories burnt. Therefore I used these features in building the model

# Exploratory Data Analysis

The screenshot shows a Jupyter Notebook with the following components:

- File Explorer:** Contains files named `dailyActivity_Analysis.ipynb`, `dailyCalories_merged.ipynb`, and `heart_rate.ipynb`.
- Code Cell:** Contains a Python script that defines a list of variables and calculates their correlations with 'Calories'.
- Output:** A correlation matrix showing the relationship between various activity metrics and 'Calories'.
- Text:** A concluding statement about the variables most correlated with 'Calories'.

```
'ModeratelyActiveDistance', 'LightActiveDistance', 'SedentaryActiveDistance', 'VeryActiveMinutes', 'FairlyActiveMinutes', 'LightlyActiveMinutes', 'SedentaryMinutes', 'Calories'], dtype='object')

dailyact_df[['TotalSteps', 'TotalDistance', 'TrackerDistance', 'LoggedActivitiesDistance', 'VeryActiveDistance', 'ModeratelyActiveDistance', 'LightActiveDistance', 'SedentaryActiveDistance', 'VeryActiveMinutes', 'FairlyActiveMinutes', 'LightlyActiveMinutes', 'SedentaryMinutes', 'Calories']].corr()
```

ModeratelyActiveDistance	LightActiveDistance	SedentaryActiveDistance	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories
0.507105	0.692208	0.070505	0.667079	0.498693	0.569600	-0.327484	0.591568
0.470758	0.662002	0.082389	0.681297	0.462899	0.516300	-0.288094	0.644962
0.470277	0.661365	0.074591	0.680816	0.463154	0.514713	-0.289343	0.645313
0.076527	0.138302	0.154996	0.234443	0.053860	0.102135	-0.046999	0.207595
0.192986	0.157669	0.046117	0.826681	0.211730	0.059845	-0.061754	0.491959
1.000000	0.237847	0.005793	0.225464	0.946934	0.162092	-0.221436	0.216790
0.237847	1.000000	0.099503	0.154966	0.220129	0.885697	-0.413552	0.466917
0.005793	0.099503	1.000000	0.008258	-0.022361	0.124185	0.035475	0.043652
0.225464	0.154966	0.008258	1.000000	0.312420	0.051926	-0.164671	0.615838
0.946934	0.220129	-0.022361	0.312420	1.000000	0.148820	-0.237446	0.297623
0.162092	0.885697	0.124185	0.051926	0.148820	1.000000	-0.437104	0.286718
-0.221436	-0.413552	0.035475	-0.164671	-0.237446	-0.437104	1.000000	-0.106973
0.216790	0.466917	0.043652	0.615838	0.297623	0.286718	-0.106973	1.000000

from here, we see that TotalDistance, TrackerDistance, TotalSteps, VervActiveMinutes are the most correlated with Calories. These are the features that we will

## Part 3

# Model Development/ Predictive Analysis

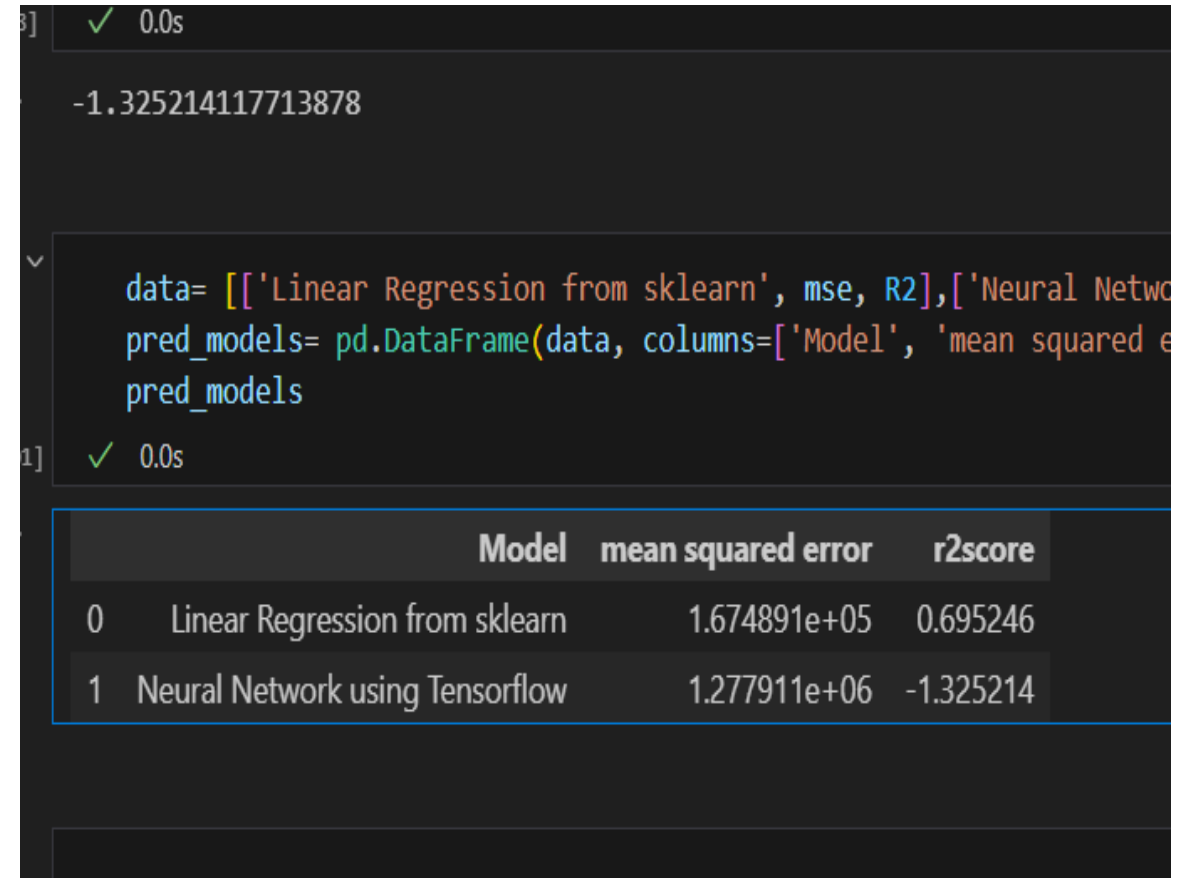
# Model Development

- I built two models to relate the input features to Calories. And in the end I chose the better one.

The first is a LinearRegression model using sklearn while the other is a Neural Network model using Tensorflow

# Model Development

- The screenshot to the right is a summary of the metrics of the model developed, as seen in the screenshot, Linear Regression using sklearn has a smaller mean squared error value than the Neural Network model. Also, its R2 score is better than the neural network model, this indicates that Linear Regression model is the best model



The screenshot shows a Jupyter Notebook interface. At the top, a cell with a green checkmark and '0.0s' execution time displays the value -1.325214117713878. Below this, a code cell contains the following Python code:

```
data= [['Linear Regression from sklearn', mse, R2],['Neural Network using Tensorflow', mse, R2]]
pred_models= pd.DataFrame(data, columns=['Model', 'mean squared error', 'r2score'])
pred_models
```

Below the code cell, another cell with a green checkmark and '0.0s' execution time displays a DataFrame table:

	Model	mean squared error	r2score
0	Linear Regression from sklearn	1.674891e+05	0.695246
1	Neural Network using Tensorflow	1.277911e+06	-1.325214



# Conclusion

1. 'TotalDistance' , 'TotalSteps', 'LightActiveDistance', 'VeryActiveDistance', 'VeryActiveMinutes' are the most correlated features to the amount of energy expended(Calories) per day.
2. They all have a positive correlation with the Calories column.
3. The total amount of minutes spent in Sedentary activity was the only negatively correlated feature with Calories, but the correlation coefficient very small.

# Recommendation

1. The essence of the model creation is to have a means by which the amount of Calories burnt can be computed from the input features. From this, Bellabeat can introduce a feature into the Bellabeat app, where the app suggests various combinations of all the input features; the total kilometres tracked, the number of steps taken, kilometres travelled during light activity, kilometres travelled during very active activity, minutes spent during very active activity, to give an approximate value of calories that will be burnt, and also an equivalent of weight that will be lost by doing these combination of activities. This feature of course will be personalized, and the example I just gave will be for a user that wants to lose weight.

# Recommendation

2. The company should develop a smart weighing balance that connects to the Bellabeat app, and with the combination of the height measurement of each user, the BMI of the user can be calculated, by doing this the user's BMI can be monitored over time, and a feature should also be included which tells them which range their BMI falls into whether Underweight, healthy weight, Overweight or obese. The app after making this information available should also be able to give brief medical advice depending on the BMI of the user, and also receive further instruction on how the user wants to proceed, whether it's to lose weight, stay fit or something else. On receiving the information on how the user wants to proceed, The app should be able to quickly generate a fitness program for the user, whether long-term or short-term.

# Recommendation

3. From the details about the case study, Bellabeat already monitors users' sleep with a device in conjunction with the app, and the dataset has already giving us an insight into how sleep is being measured. What Bellabeat should do differently, is make their app more engaging and personal. It can go ahead to notify a user during the time period when she has to go to bed, if she is to get the required hours of sleep before the next morning.

# Recommendation

4. A feature should be integrated into the bellabeat app, such that, in conjunction with a smart device, the heartrate of a patient can be measured. And the device should be able to notify the user when their heartrate is above the healthy level.

# Appendix

- To view all the code, jupyter notebooks, and associated project information used in this case study, go to the URL address below:

<https://github.com/Jesutimilehin-Onayemi/Playing-it-Smart-in-the-Wellness-Industry-with-Data-Science/tree/main>

Thank you!

