Bellabeat Case Study

Regina

1/3/2022

Scenario

BellaBeat is 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, co-founder and Chief Creative Officer of BellaBeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company.

Step 1 - ASK - Identify the business task and key stake holders.

Business Task

The marketing analytics team has 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 discovered will then help guide marketing strategy for the company through high-level recommendations for BellaBeat's marketing strategy.

Stakeholders

- Urška Sršen: BellaBeat's cofounder and Chief Creative Officer
- Sando Mur: Mathematician and BellaBeat's cofounder; key member of the Bellabeat executive team
- BellaBeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

Step 2 - PREPARE - Fetch appropriate data

I will use a public data that explores smart device users' daily habits. (https://www.kaggle.com/arashnic/fit bit) on Kaggle. 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.

Step 3 - PROCESS - Choosing appropriate tools and data cleaning

As the dataset is large and has multiple variables, I decided to choose RStudio for analysis. Using R, I can process and analyse the data faster and visualize at the same time.

3.a)

Data Processing begins with installing required packages. Here I use tidyverse.

```
install.packages("tidyverse")
library(tidyverse)
```

3.b)

Next I view the Daily Activity data and its metadata.

```
daily_activity <- read_csv("dailyActivity_merged.csv")
colnames(daily_activity)
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"
```

```
tibble(daily_activity)
```

```
## # A tibble: 940 x 15
##        Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitie~
##     <dbl> <chr>             <dbl>         <dbl>           <dbl>             <dbl>
##  1 1.50e9 4/12/2016         13162          8.5             8.5                 0
##  2 1.50e9 4/13/2016         10735          6.97            6.97                0
##  3 1.50e9 4/14/2016         10460          6.74            6.74                0
##  4 1.50e9 4/15/2016          9762          6.28            6.28                0
##  5 1.50e9 4/16/2016         12669          8.16            8.16                0
##  6 1.50e9 4/17/2016          9705          6.48            6.48                0
##  7 1.50e9 4/18/2016         13019          8.59            8.59                0
##  8 1.50e9 4/19/2016         15506          9.88            9.88                0
##  9 1.50e9 4/20/2016         10544          6.68            6.68                0
## 10 1.50e9 4/21/2016          9819          6.34            6.34                0
## # ... with 930 more rows, and 9 more variables: VeryActiveDistance <dbl>,
## #   ModeratelyActiveDistance <dbl>, LightActiveDistance <dbl>,
## #   SedentaryActiveDistance <dbl>, VeryActiveMinutes <dbl>,
## #   FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,
## #   SedentaryMinutes <dbl>, Calories <dbl>
```

Step 4 - ANALYZE

I now begin the analysis.Following is my approach – * Group user data by user ID and analyse trends in average value of all statistics. * Group users in to various types ranging from 'very active' to 'sedentary' and analyze trends * Analyze trends in distance and calories to provide actionable insights

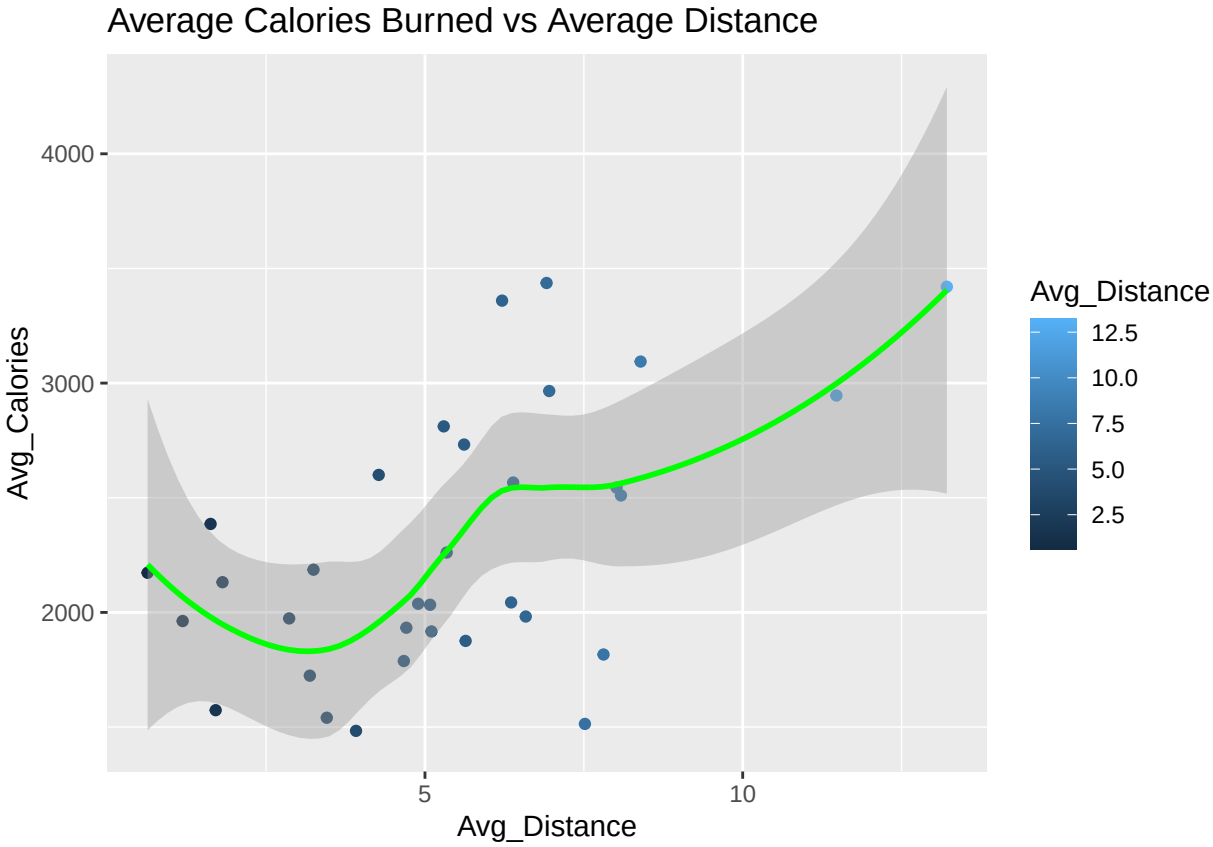
4.a)

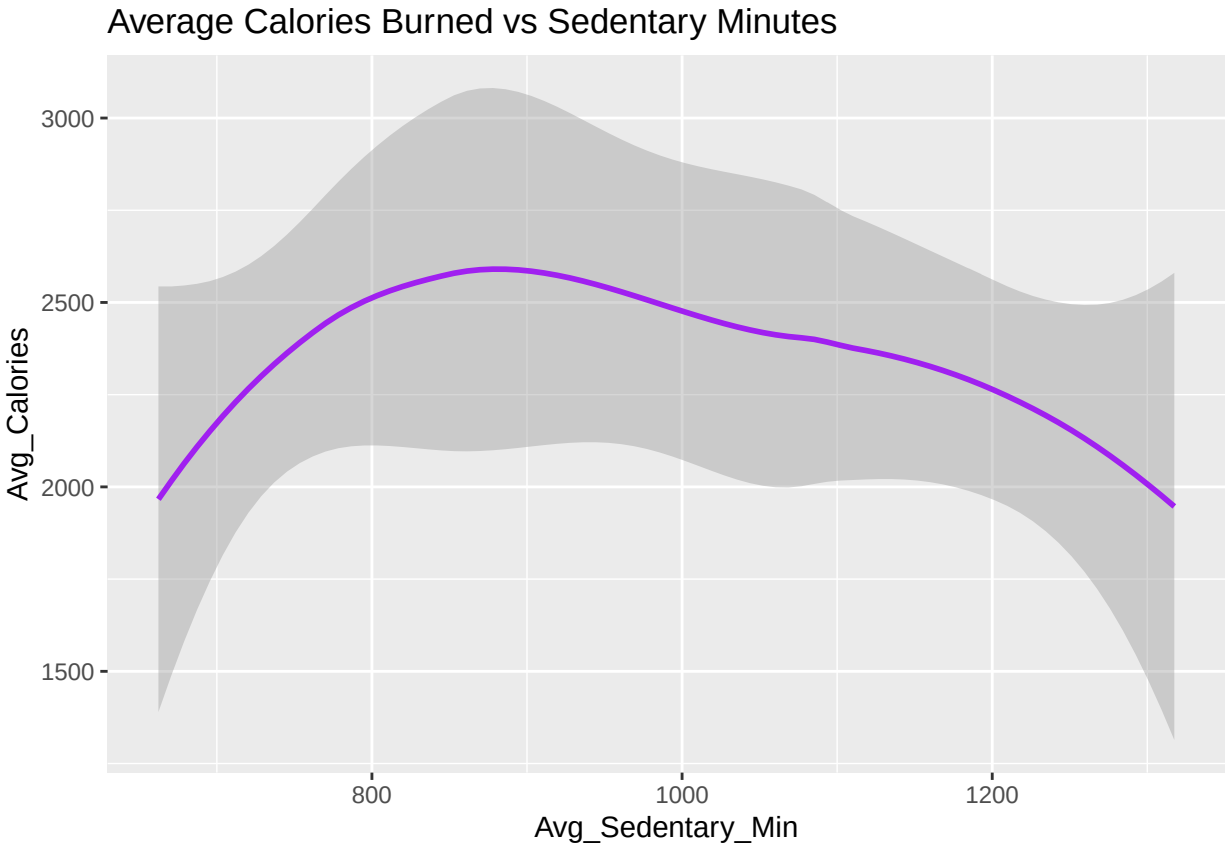
Create a new data frame grouping users by ID and analyzing general trends by visualizing.

```
## # A tibble: 6 x 11
##           Id Avg_Steps Avg_Distance Avg_VeryActive_Distance Avg_ModActive_Dista~
##        <dbl>     <dbl>        <dbl>                   <dbl>                <dbl>
## 1 1503960366    12117.        7.81                    2.86                0.794
## 2 1624580081     5744.        3.91                    0.939               0.361
## 3 1644430081     7283.        5.30                    0.730               0.951
## 4 1844505072     2580.        1.71                    0.00839             0.0490
## 5 1927972279      916.        0.635                   0.0958              0.0313
## 6 2022484408    11371.        8.08                    2.42                0.720
## # ... with 6 more variables: Avg_LightActive_Distance <dbl>,
```

```
## #   Avg_VeryActive_Min <dbl>, Avg_FairlyActive_Min <dbl>,
## #   Avg_LightActive_Min <dbl>, Avg_Sedentary_Min <dbl>, Avg_Calories <dbl>
```

From this data frame we now plot graphs showing relation between average distance and average calories burned and another plot showing relation between number of sedentary minutes and calories burned.



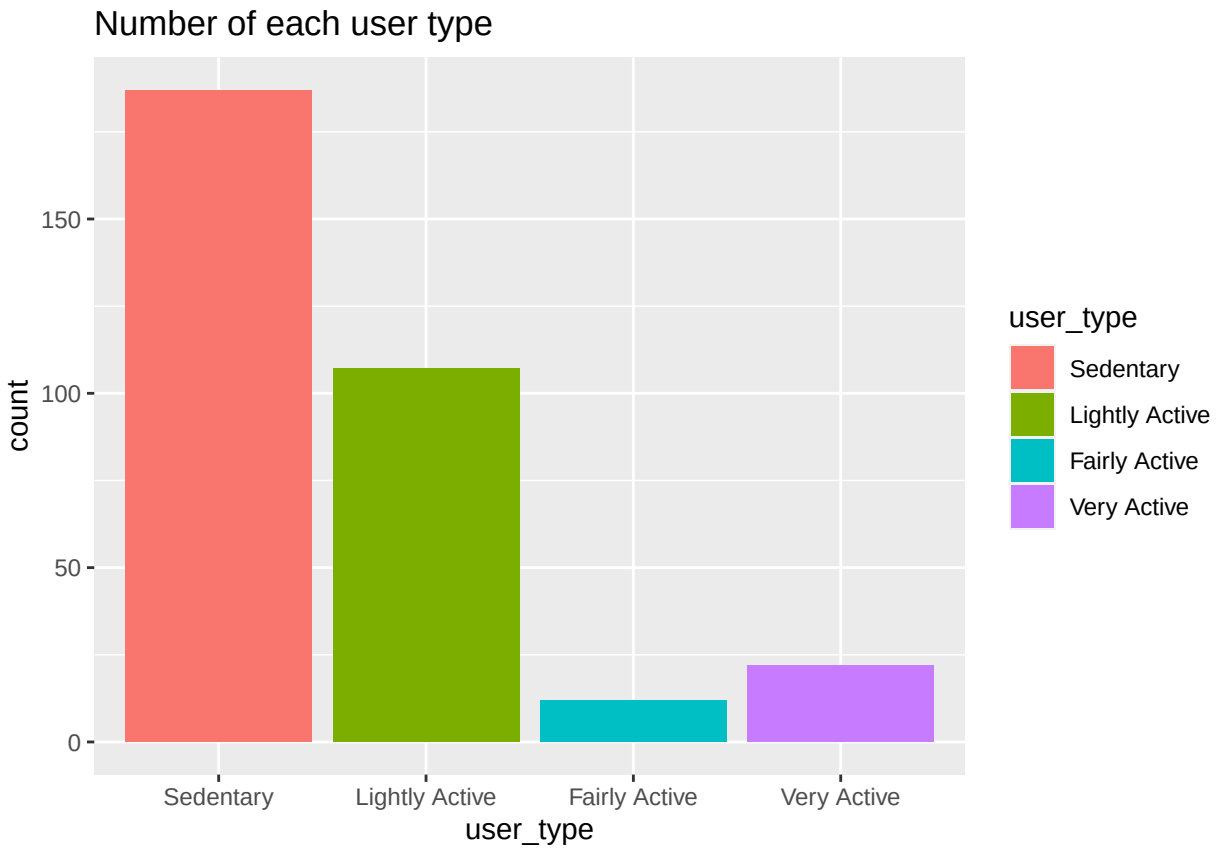


The plot confirms that as the distance covered increases more calories are burned and that longer sedentary periods lead to lesser calorie burn.

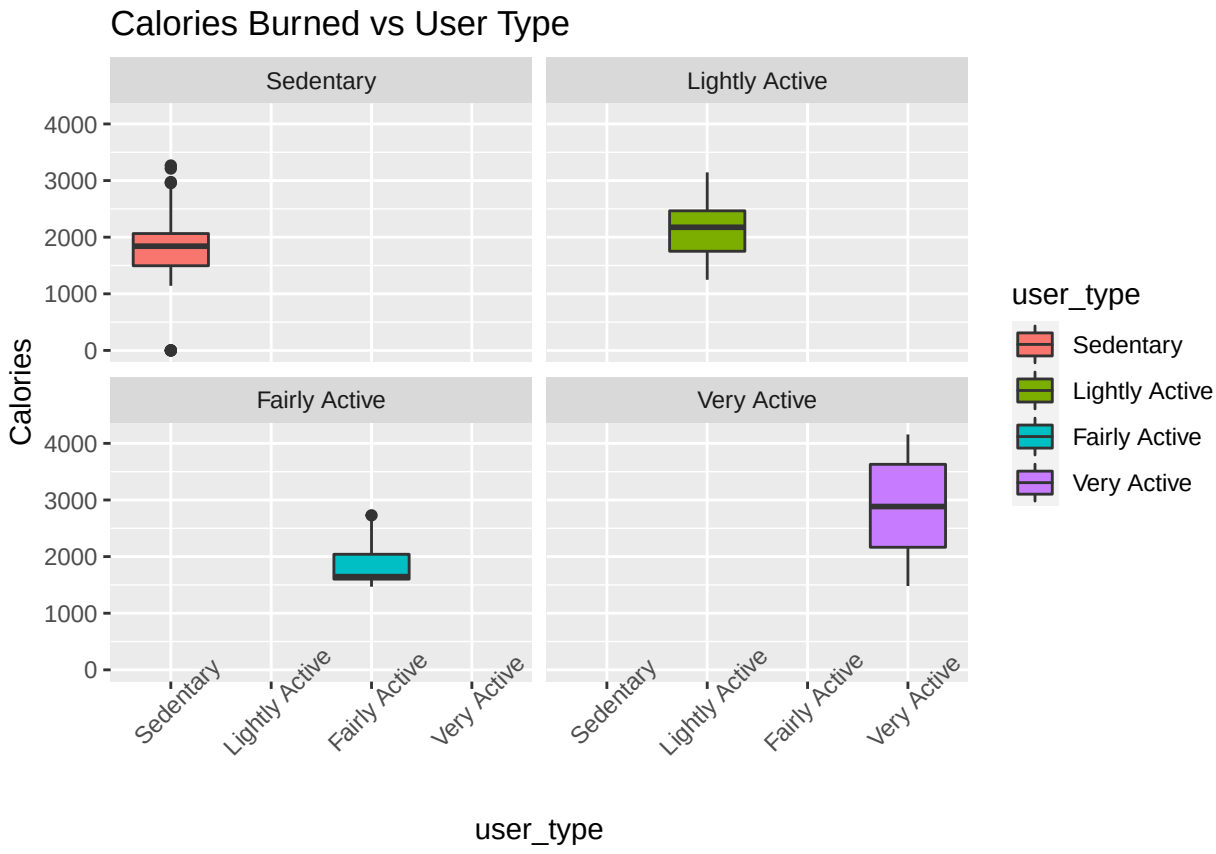
4.b) Next I categorize users into 4 types based on their daily activity intensities.

```
## # A tibble: 6 x 3
##   user_type      Calories         Id
##   <fct>             <dbl>      <dbl>
## 1 Lightly Active     1775 1503960366
## 2 Lightly Active     1837 1503960366
## 3 Sedentary             0 1503960366
## 4 Sedentary          1432 1624580081
## 5 Sedentary          1411 1624580081
## 6 Sedentary          1344 1624580081
```

Now I use bar graph to identify distribution of users.

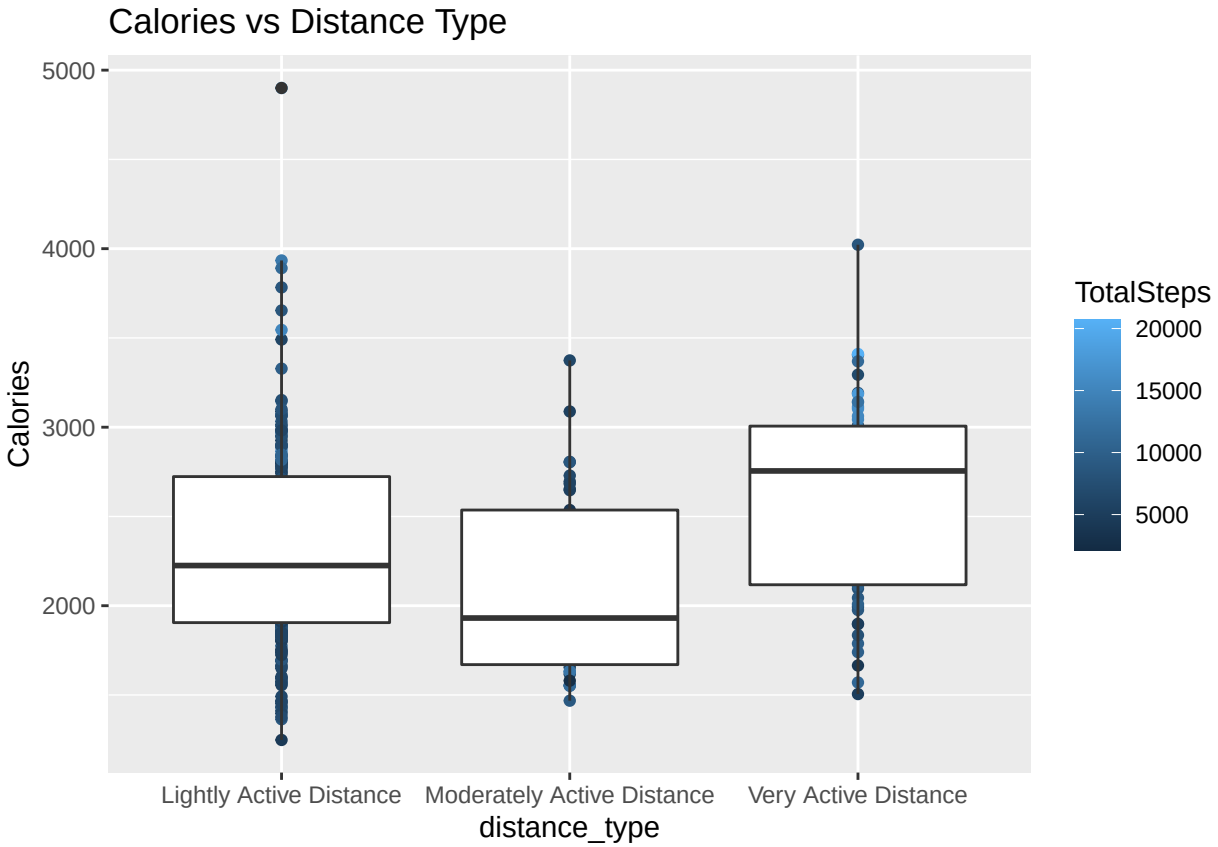


The graph shows that maximum number of users are of sedentary type, followed by lightly active, very active and fairly active. It would be interesting to find relationship between user type and calories burned.



As expected barring a few outliers the boxplot shows that maximum calories are burned by Very Active users and least by Sedentary users.

Now that I know that maximum users are sedentary type, it will be helpful if I can show some effective ways of burning calories. Thus, I create a new data frame based on distance type and show its relation to calories.



The box plot shows that maximum calories are burnt under very active distance which implies through activities like running, cycling etc. However, even the light activities like walking, jogging can burn equal amount of calories provided that the number of steps are more.

Step 5 - SHARE

In this step of the analysis the marketing analytics team shares the findings of the analysis and makes high level recommendations.

Findings

- There is a positive co-relation between physical activity and calories burned.
- Maximum type of users in the sample were of sedentary type.
- Most calories are burned by users that are very active,
- Fast paced activities burn most calories however slower activities can do the same if done for longer.

Recommendations

- Fitness data like number of steps, distance covered, calories burned must be collected and should give a daily status to users.
- Apart from descriptive analysis predictive analysis must also be done especially beacause majority users are sedentary and aiming for improvement - for example, daily update could read -“Congratulations! Your daily activity status for today is Fairly Active. Spend x more active minutes tomorrow to upgrade to Very Active status.You can do it!”
- Users should be able to set goals and based on goals and accordingly suggestions can be made for the type of activity they should perform.