Bellabeat Case Study

By: Shubham Aswal

Introduction

Bellabeat is a high-tech manufacturer of health-focused products for women. The sole purpose of this case study is to determine the target audience for marketing an existing Bellabeat product, using an existing dataset of 33 Fitbit users.

Since a Fitbit has the closest resemblance to the Bellabeat Time, we will be using the dataset to determine the usage patterns and target audience for the latter.

The report is divided into the following sections:

1. Data Processing
    a. Detailed description of the processing done on the data to make sure it is clean and consistent before analysis.
2. Statistical Analysis & Summary
    a. Important statistical values from the datasets that have been chosen for analysis.
3. Visualizations & Analysis
    a. Finding trends and driving insights from the data and meaningful visualizations to better showcase the analysis.
4. Final Suggestions
    a. Recommending the suitable target audience for the Bellabeat Time.
5. Limitations
    a. The limitations of the datasets due to lack of more information.

Data Processing

Exactly 18 CSV files were provided for the purpose of the case study. They are all publicly available on Kaggle and hence all the data used was Third Party Data, hence, rigorous data processing had to be done in order to ensure that the final datasets used for analysis were clean and consistent.

Metadata

Most of the CSV files provided could be divided into 3 categories:

1. Daily measured data
2. Hourly measured data
3. Minutely measured data

Elimination of Unusable Data

It was clearly specified that the data provided would be for 30 users, after checking all the datasets it was found that some datasets didn't have the required number of users for consistent analysis, hence it was concluded that heartrate, weight log files and sleep data could not be used for analysis since they didn't have enough users (<25).

All data was divided into 3 folders; 'Daily_Data','Hourly_Data','Minute_Data', containing 4,5 and 8 CSV files respectively.

Daily Data Processing (Excel/Google Sheets)

Similar data cleaning techniques were used for cleaning all 5 files, namely:

1. Clear formatting
2. Remove duplicates
3. Trim Whitespace
4. Changing ActivityDay column from string to Date type
5. Rounding off all numeric values to 2 decimal places

In order to confirm that all 5 files had consistent (ID,ActivityDay) column pair entries, an inner join was performed on all the 5 files using excel and it was verified that the resultant merged file also contained the same number of entries, hence all daily data was now clean and consistent.

The merged CSV file was used for final analysis of daily data. Similar treatment was given to hourly and minutely gathered data, however, minute data was too large to be processed in excel and hence R was used for cleaning and processing of minute data, An R notebook file has been attached for explanation of the data cleaning process in R.

NOTE: *Even though minute data contained exactly 33 users and was consistent and clean, it was only cleaned in case it is required in the future, since minute level data doesn't provide any high-level insights into user patterns.*

Statistical Analysis & Summary

An overview of the 2 datasets that will be used for analysis:

Daily Data

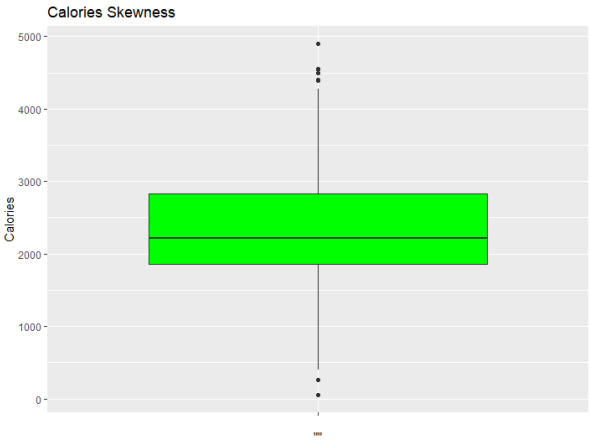
Summary of Daily Data

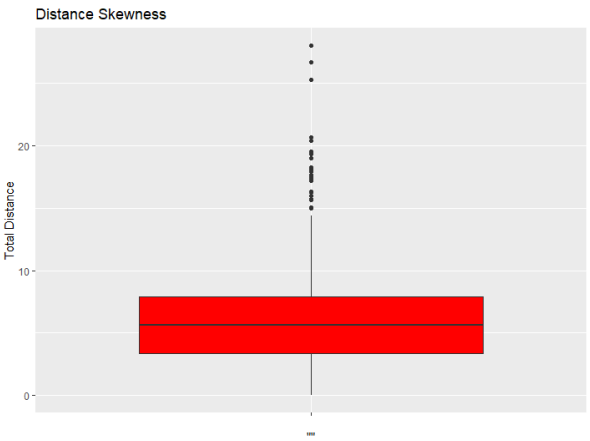
Hide

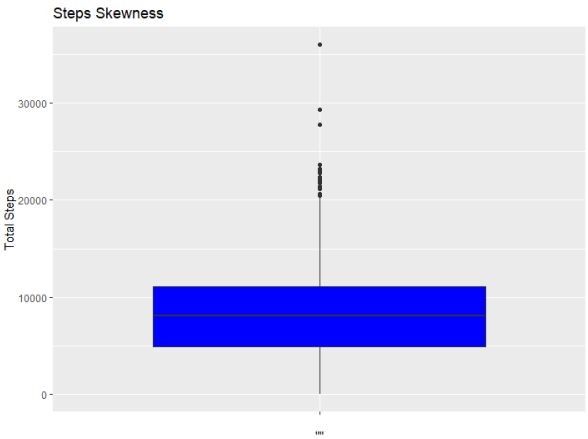
```
df=read_csv("C:\\Users\\sasuk\\OneDrive\\Desktop\\DataAnalyticsCourse\\Project\\Fitbase_Data_Modified\\Daily_data_Merged\\daily_data_merged.csv")
temp=df %>% select(TotalSteps,TotalDistance,Calories)
summary(temp)
```

```
   TotalSteps    TotalDistance       Calories
 Min.   :    0   Min.   : 0.000   Min.   :   0
 1st Qu.: 3790   1st Qu.: 2.620   1st Qu.:1828
 Median : 7406   Median : 5.245   Median :2134
 Mean   : 7638   Mean   : 5.490   Mean   :2304
 3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:2793
 Max.   :36019   Max.   :28.030   Max.   :4900
```

Skewness of the Daily Data:







The above plots visualize the skewness of a few important metrics present in daily data.

Hourly Data

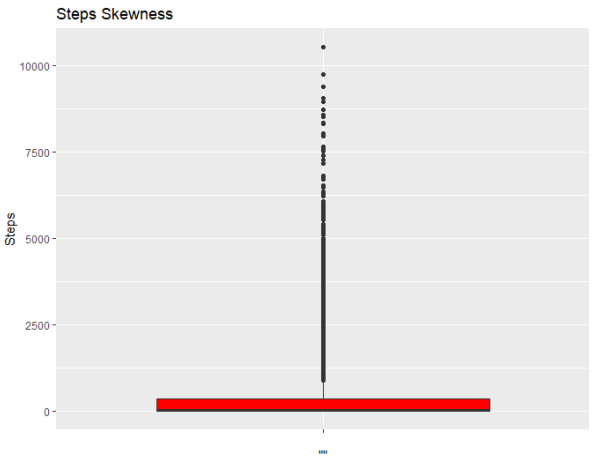
Summary of Minute Data

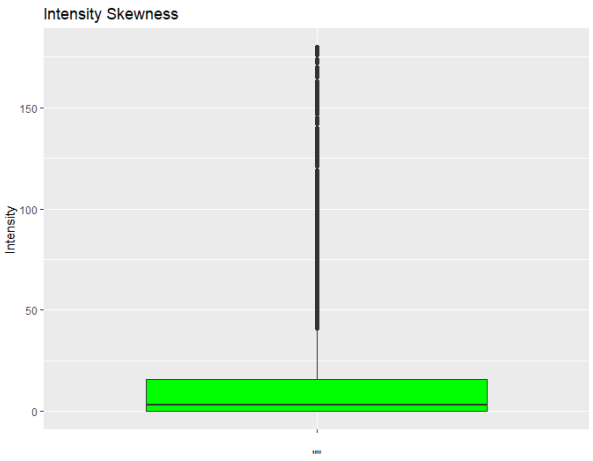
Hide

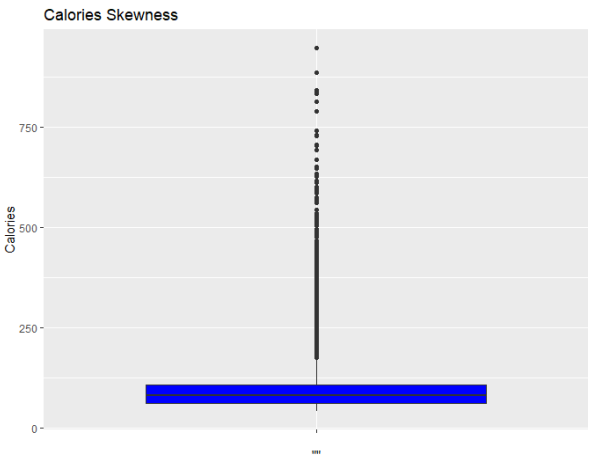
```
df=read_csv("C:\\Users\\sasuk\\OneDrive\\Desktop\\DataAnalyticsCourse\\Project\\Bellabeat_Case_Study\\Transformed_Data\\Hourly_Data_Merged\\hourly_merged.csv")
temp=df %>% select(Calories,`Total Intensity`,AverageIntensity,StepTotal)
summary(temp)
```

```
    Calories      Total Intensity  AverageIntensity   StepTotal
 Min.   : 42.00   Min.   :  0.00   Min.   :0.0000   Min.   :    0.0
 1st Qu.: 63.00   1st Qu.:  0.00   1st Qu.:0.0000   1st Qu.:    0.0
 Median : 83.00   Median :  3.00   Median :0.0500   Median :   40.0
 Mean   : 97.39   Mean   : 12.04   Mean   :0.2007   Mean   :  320.2
 3rd Qu.:108.00   3rd Qu.: 16.00   3rd Qu.:0.2700   3rd Qu.:  357.0
 Max.   :948.00   Max.   :180.00   Max.   :3.0000   Max.   :10554.0
```

Skewness of Hourly Data:







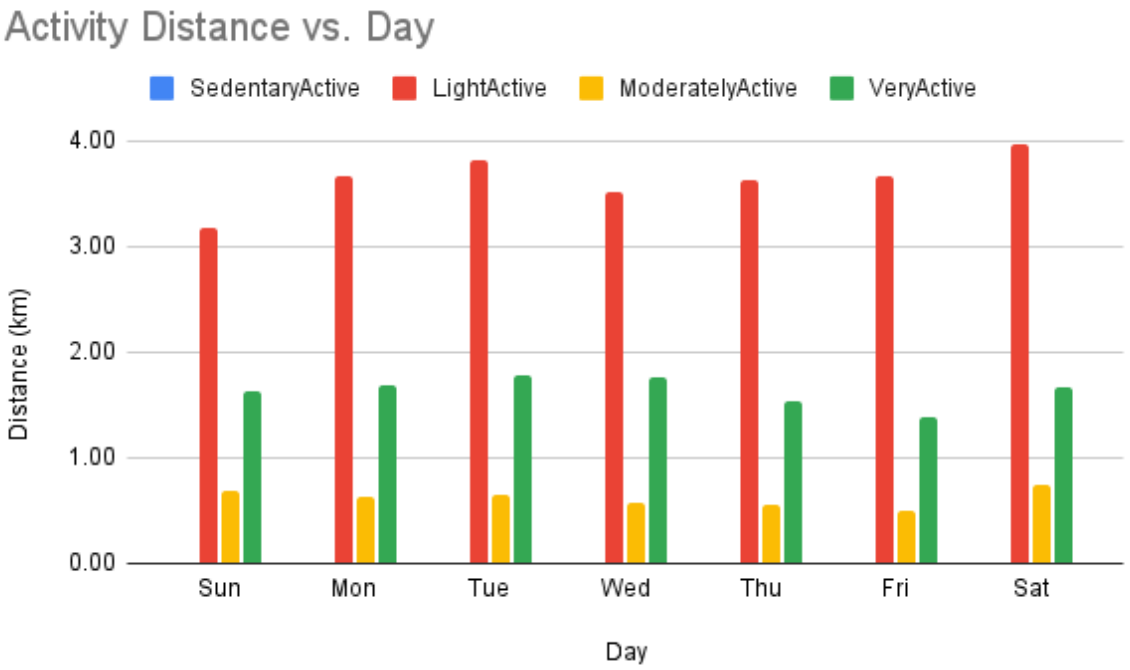
The above plots show the skewness of minute data.

Overall, except calories burnt, all data is heavily right skewed.

***Note**: R code files for all the boxplots can be found in their respective folders.*

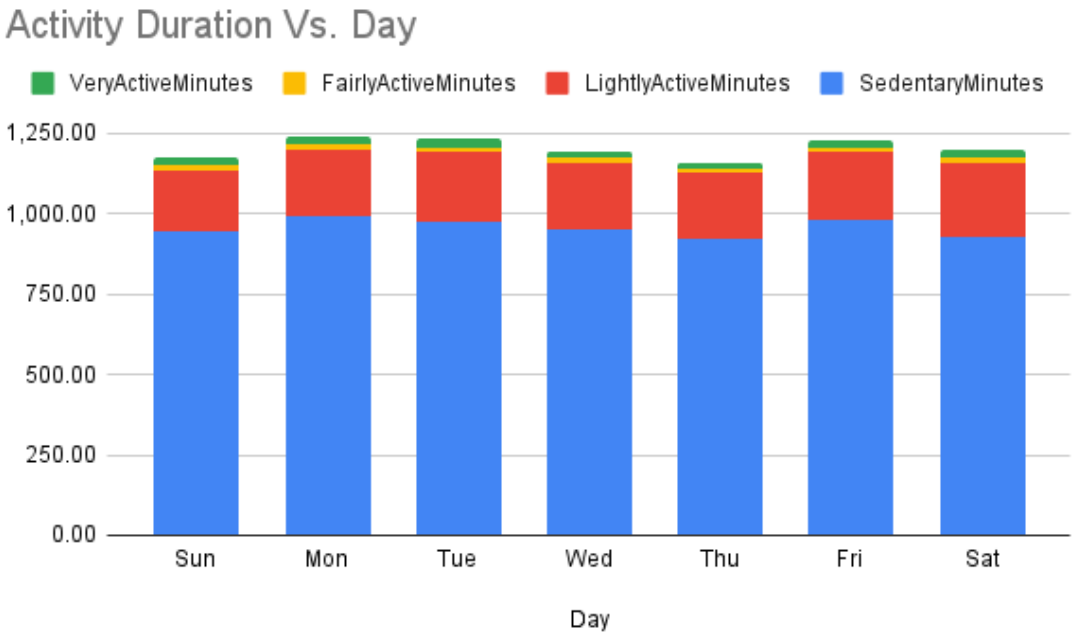
Visualizations and Analysis

Trends in Daily Data



Day	SedentaryActive	LightActive	ModeratelyActive	VeryActive
Sun	0.00	3.18	0.68	1.64
Mon	0.00	3.67	0.64	1.68
Tue	0.00	3.82	0.65	1.78
Wed	0.00	3.51	0.57	1.76
Thu	0.00	3.63	0.56	1.54
Fri	0.00	3.66	0.51	1.38
Sat	0.00	3.97	0.74	1.66

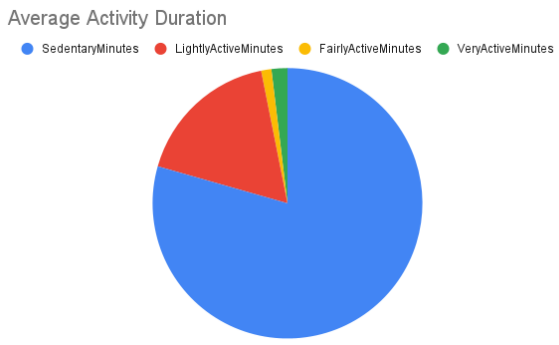
Average activity type by day, numerically and visually (Distance).
We can infer that Light Activity (ex. walking) is by far the most performed activity requiring movement throughout the day.
We are also going to notice a trend of low physical activity on Sundays in all the visualizations.

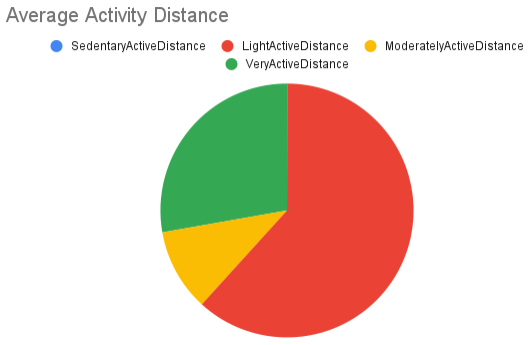


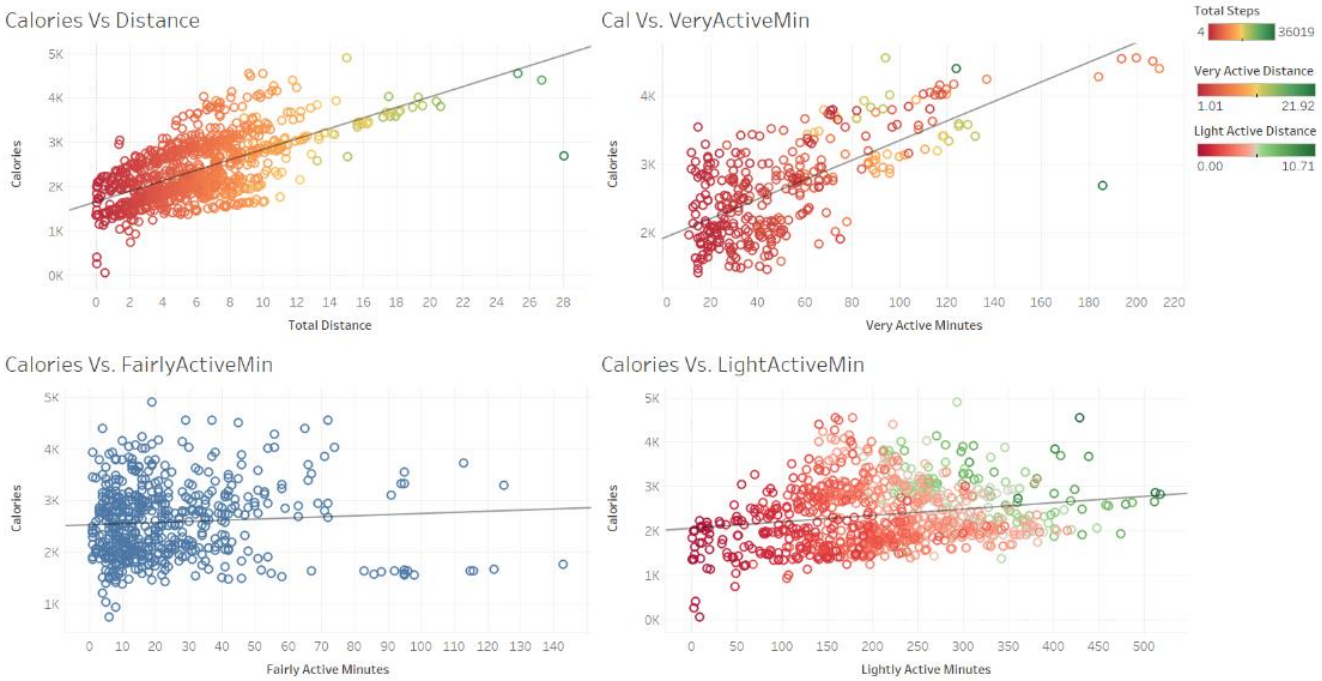
Day	SedentaryMinutes	LightlyActiveMinutes	FairlyActiveMinutes	VeryActiveMinutes
Sun	945.28	191.37	15.98	21.98
Mon	990.48	209.52	15.27	25.21
Tue	973.56	217.36	15.79	25.28
Wed	953.83	204.88	14.14	22.42
Thu	920.72	204.94	13.22	21.45
Fri	978.33	214.41	12.72	21.06
Sat	930.03	227.31	16.68	23.76

We can conclude by saying that Fitbit users spend most of their time sitting and the most frequent activity is walking.

The pie charts below also confirm these insights.







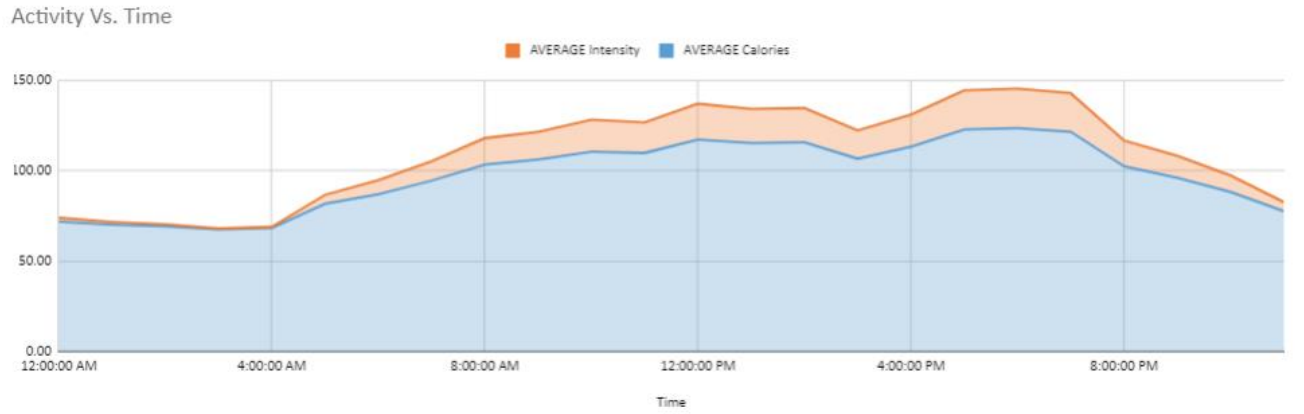
The above plots show the very obvious trend between calories burned and distance travelled, from the trend lines we can also roughly estimate how much activity is required to burn a certain amount of calories.

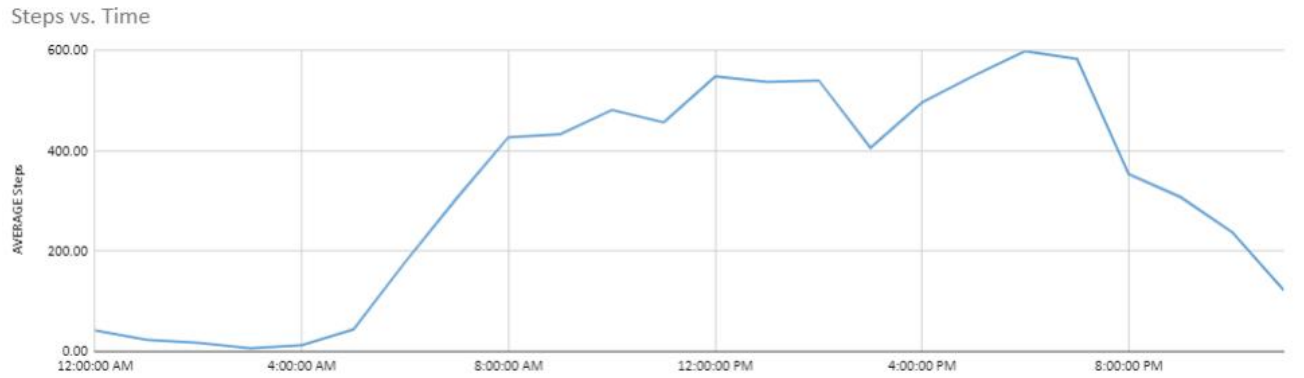
Trends in Hourly Data

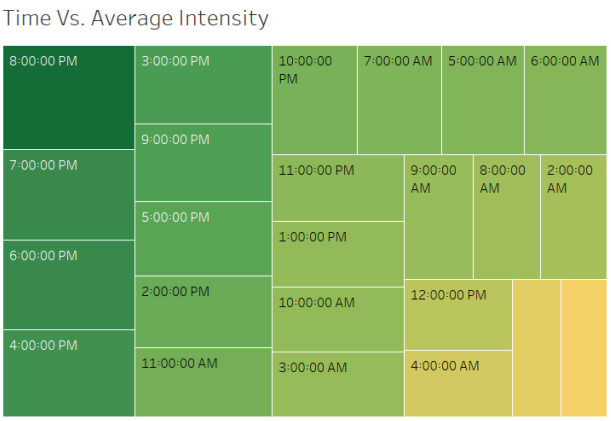
The visualizations shown below clearly indicate a trend in activity on the basis of time of day, indicating that 4 pm to 8 pm is when majority of the users spend walking or doing some sort of intensive work.

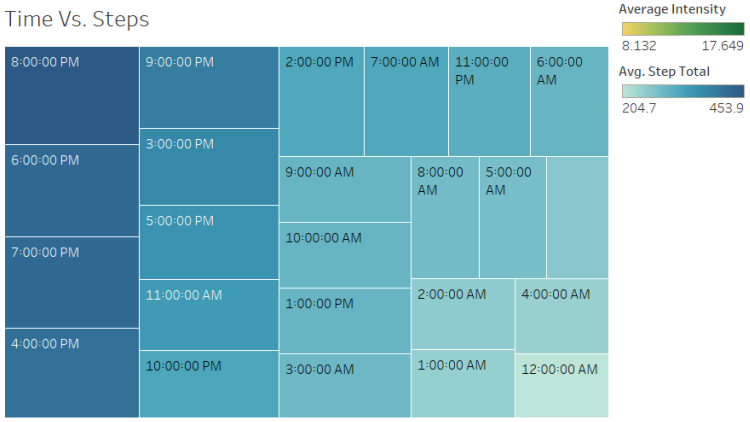
These trends make sense since 4 pm to 8 pm is usually the time when office workers commute back from work and go to the gym.

There is also the fact that the spike in evening is higher than the spike in the morning, which could mean that users like to do more physical activity in the evening.









Final Suggestions

1. Targeting women that have a “9 to 5” desk job seems to be the best marketing strategy.
2. The smart wearable can also be targeted towards women that like to go for evening walks/jogs.
3. As an added bonus, the trend line plotted between distance and calories burned can be improved further and be added as a feature to the Bellabeat app, pushing users to do X amount of specific activity to burn Y amount of calories, a regression model can be used here, alternatively the users can also set daily calory burn goals and the app can send push notifications on phone to encourage users to hit their targets.

Limitations

1. Lack of adequate sleep and weight data limited the depth of insights that could be generated.
2. In order to develop an accurate model for calory burn prediction, additional metrics such as age, weight and height also need to be obtained from the user.