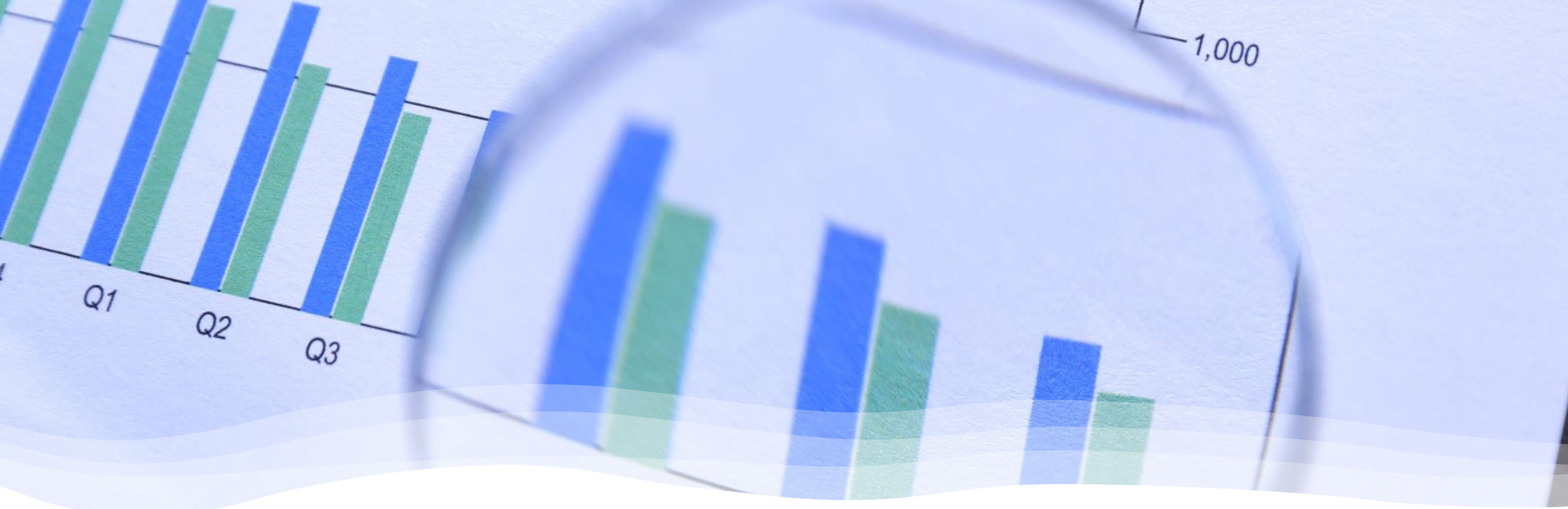


Data Processing & Data Analysis Part I for Case Study “How Can a Wellness Technology Company Play It Smart?”

By Kristin Lu
February 1, 2024

Data Preparation and Data Exploration

- The data used for this case study is the "**FitBit Fitness Tracker Data**" which was downloaded from 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**. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits.
- There are 18 files in the dataset. Not all the files were used for the analysis. The following is a description of the file used in part I of this analysis:
  - dailyActivity_merged.csv - this file contains the following columns: Id, ActivityDate, (in Short Date format "m/d/yyyy"), TotalSteps, TotalDistance, VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, SedentaryActiveDistance, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, Calories, etc.

Data Processing – Daily Activity

- Open dailyActivity_merged.csv and save it as an Excel Workbook.
- Built a Pivot Table to count number of rows associated with each consumers. Removed all rows associated with consumers with less than 15 days of data (15 days is 50% of 31 days, that is, the number of days from 4/12/2016 to 5/12/2016). Otherwise, the data might be biased. One consumer (4057192912) were impacted.
- Built a pivot table to sum up the total calories consumed by each consumer. Do the same to sum up the total calories consumed by each consumer in hourlyCalories_merged file. Compared these two sets of values. Delete rows associated with consumers whose total calorie values calculated here are significantly different from that values calculated in another file (difference greater than 1000 calories). Four consumers (8583815059, 6117666160, 4388161847, and 4319703577) were impacted.
- It's not clear that if sleeping time is considered SedentaryMinutes. Let's added a new column TotalMinutes to sum up VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes and SedentaryMinutes. There are 400+ rows have TotalMinutes field equal to 1440. Looks like sleeping time is considered SedentaryMinutes for some consumers, but not all the consumers. Anyway, remove rows with TotalMinutes less than 720 (half day).

Data Processing – Daily Activity

- Set a filter such that only rows with 0 TotalSteps value will be displayed. Noticed that there were quite a few rows with 0 value in the all the following fields: TotalSteps, TotalDistance, VeryActiveDistance, ModeratelyActiveDistance, LightlyActiveDistance, VeryActiveMinutes, FairlyActiveMinutes, and LightlyActiveMinutes – most of these rows have 1000+ in their Calories field, but few of these rows (3) just have 0 in their Calories filed. Looks like there is inconsistent way to calculate calorie consumed by each consumer. According to an article, people are still burn calories while they are in sedentary state or even sleep to maintain basic bodily functions. Remove all rows with 0 Calories field.
- Note: hence "calories burned" is not a reliable way to evaluate the effectiveness of one's walking/exercise. We will fill filter out the consumers who have quite a few rows with 0 TotalSteps while having 1000+ in Calories fields in some analysis.

Here is the article cited above: https://www.health.harvard.edu/staying-healthy/burning-calories-without-exercise?fbclid=IwAR1wfcE9be0GsymB5JHQDPv0HBoYn_LaRT3CBNmHUOpj13hwKBktGBdu6nw#:~:text=It's%20true%3A%20just%20sitting%20on,up%20watching%20TV%20or%20reading

Data Analysis: Daily Activity Tracking

Made some plots to see how TotalSteps and Total Distance data are distributed.

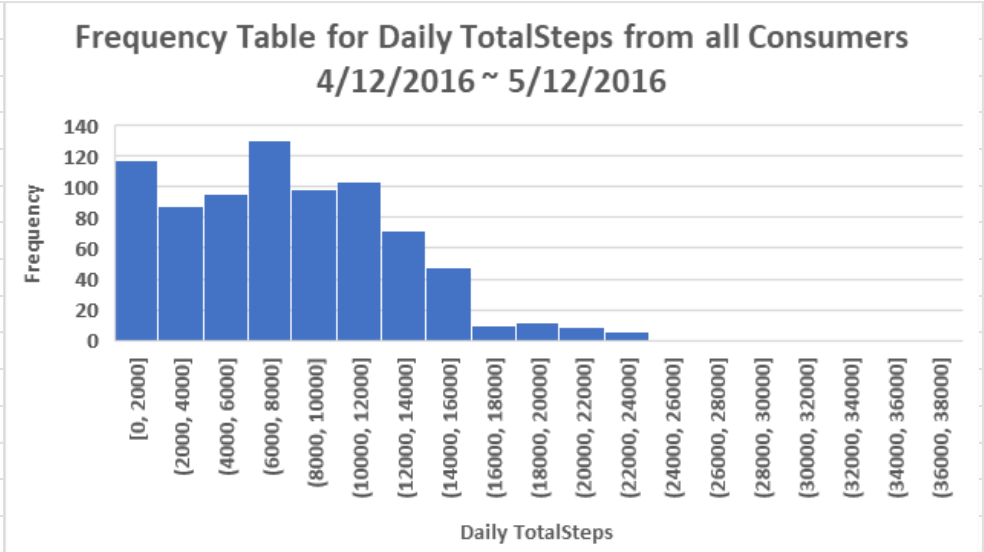
- The Daily TotalSteps histogram shows that the most common set of values for daily total steps are between **6,000 ~ 8,000** for consumers in this dataset. There is a long tail in the right side of the plot.
- The Daily TotalSteps boxplot shows that the **median value of daily total steps** is **7,396**. There are **some outliers above upper whisker** means there were some consumers who took much more steps than other consumers.
- The Daily TotalDistance histogram shows that the most common values for daily total distance are between **0 ~ 1.5 miles** in this dataset. The next common values are between **4.5** and **6 miles** for consumers in this dataset. There is a long tail in the right side of the plot.
- The Daily TotalDistance boxplot shows that the **median daily total distance** value is **5.19 miles**. There are **quite a few outliers above the upper whisker** means there were quite a few consumers who walked much longer in distance than other consumers.

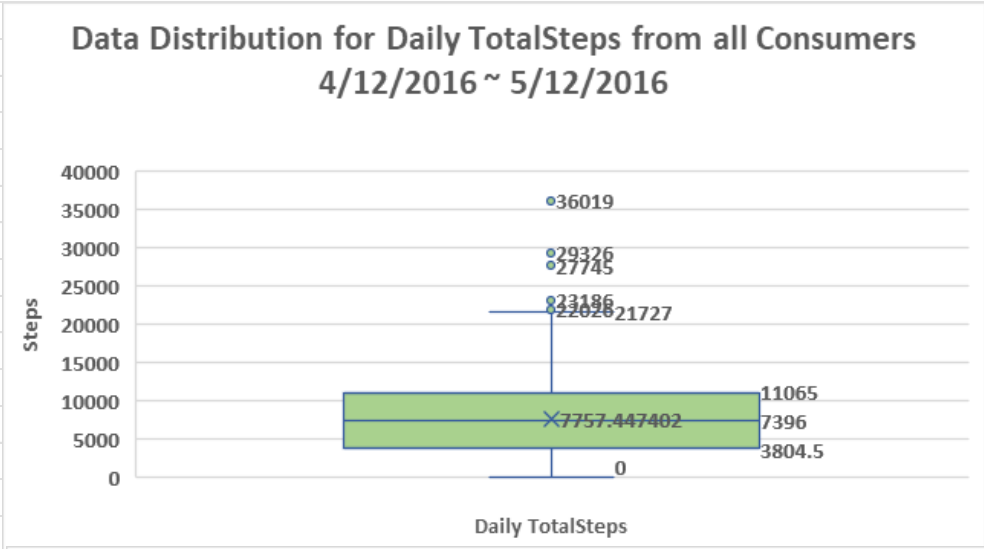
Data Analysis: Daily Activity Tracking

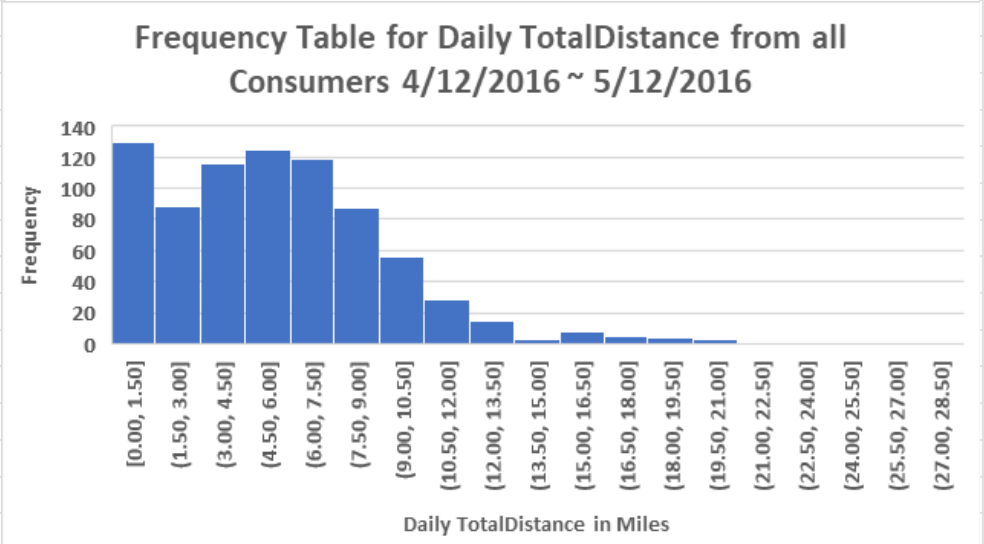
Made some plots to see how TotalSteps and Total Distance data are distributed.

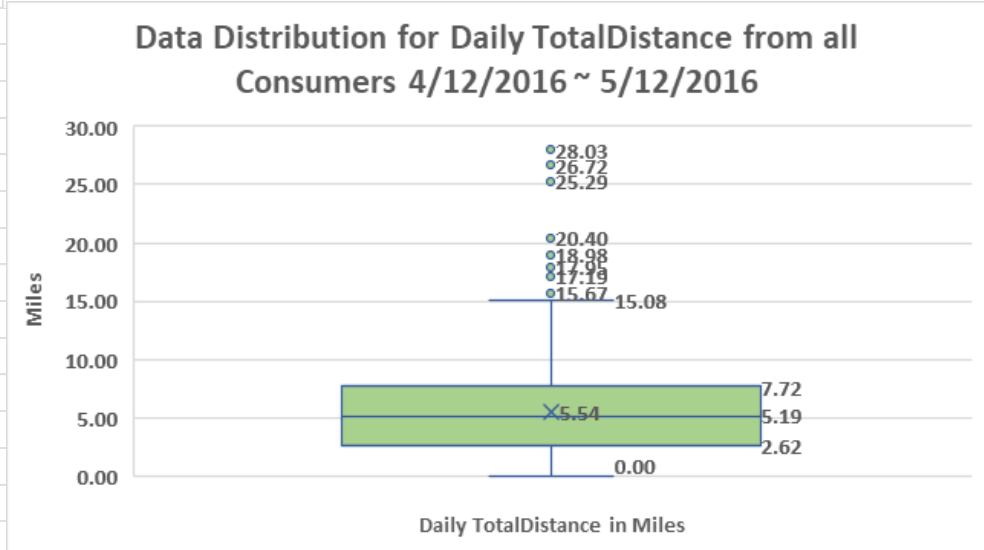
- The Daily TotalSteps histogram shows that the most common set of values for daily total steps are between **6,000 ~ 8,000** for consumers in this dataset. There is a long tail in the right side of the plot.
- The Daily TotalSteps boxplot shows that the **median value of daily total steps** is **7,396**. There are **some outliers above upper whisker** means there were some consumers who took much more steps than other consumers.
- The Daily TotalDistance histogram shows that the most common values for daily total distance are between **0 ~ 1.5 miles** in this dataset. The next common values are between **4.5** and **6 miles** for consumers in this dataset. There is a long tail in the right side of the plot.
- The Daily TotalDistance boxplot shows that the **median daily total distance** value is **5.19 miles**. There are **quite a few outliers above the upper whisker** means there were quite a few consumers who walked much longer in distance than other consumers.

Data Analysis: Daily Activity Tracking









Data Analysis: Daily Activity Tracking

Made some plots to see how VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes and SedentaryMinutes data are distributed.

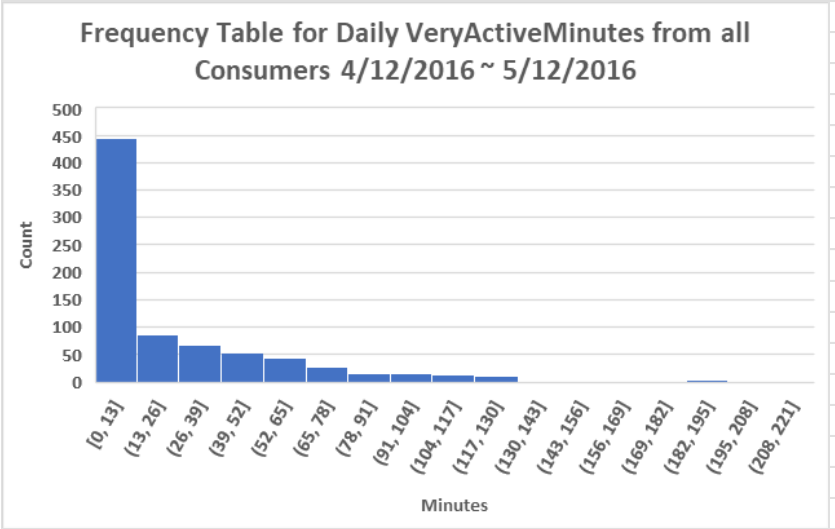
- The Daily VeryActiveMinutes histogram shows that the most common daily VeryActiveMinutes values are between **0 ~ 13** minutes for consumers in this dataset.

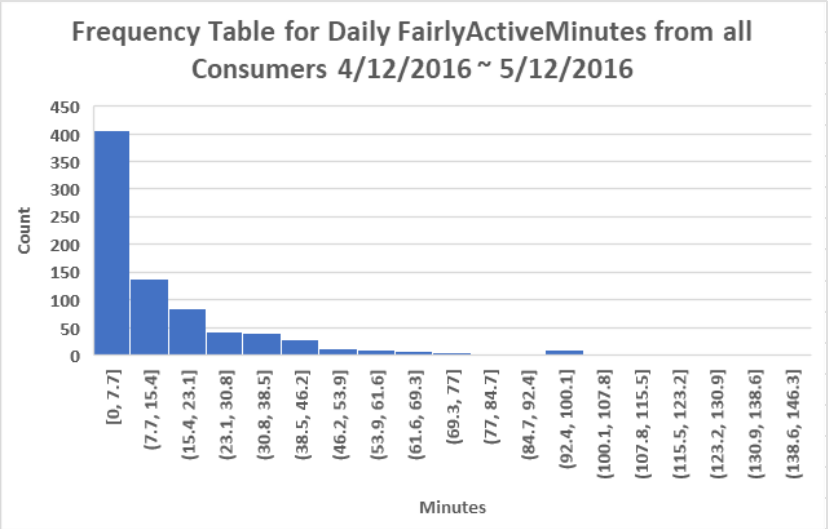
- The Daily VeryActiveMinutes boxplot shows that the **median daily VeryActiveMinutes** value is **6**. There are **some outliers above the upper whisker** means there were some consumers having much longer VeryActiveMinutes than other consumers.

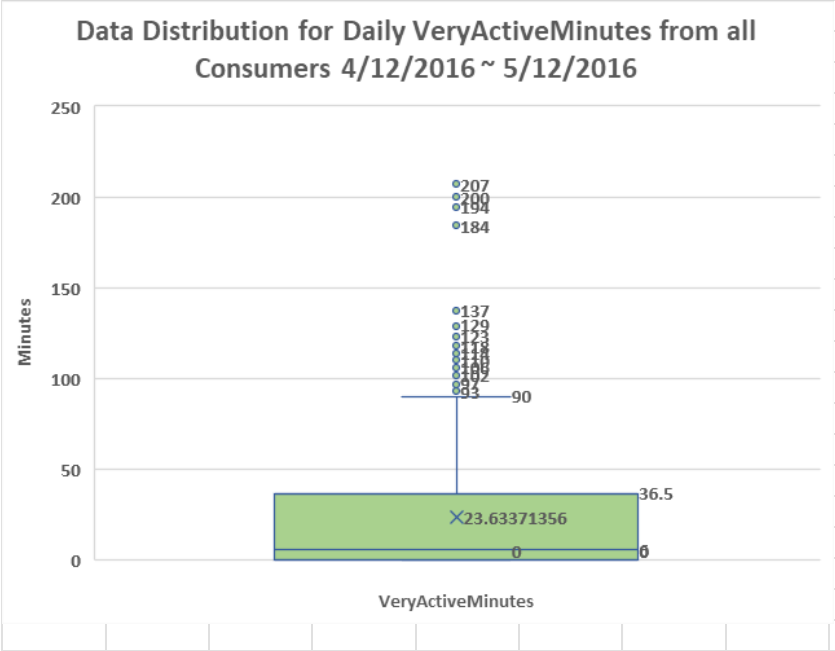
- The Daily FairlyActiveMinutes histogram shows that the most common values for daily FairlyActiveMinutes are between **0 ~ 7.7 minutes.**

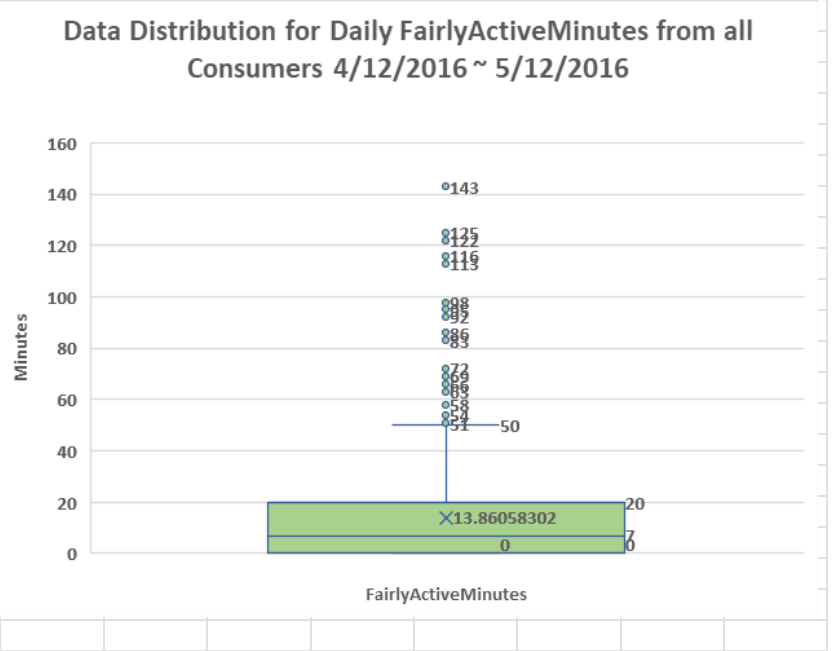
- The Daily FairlyActiveMinutes boxplot shows that the **median daily FairlyActiveMinutes** value is **7** minutes. There are **quite a few outliers above the upper whisker** means there were quite a few consumers having much longer FairlyActiveMinutes than other consumers.

Data Analysis: Daily Activity Tracking









Data Analysis: Daily Activity Tracking

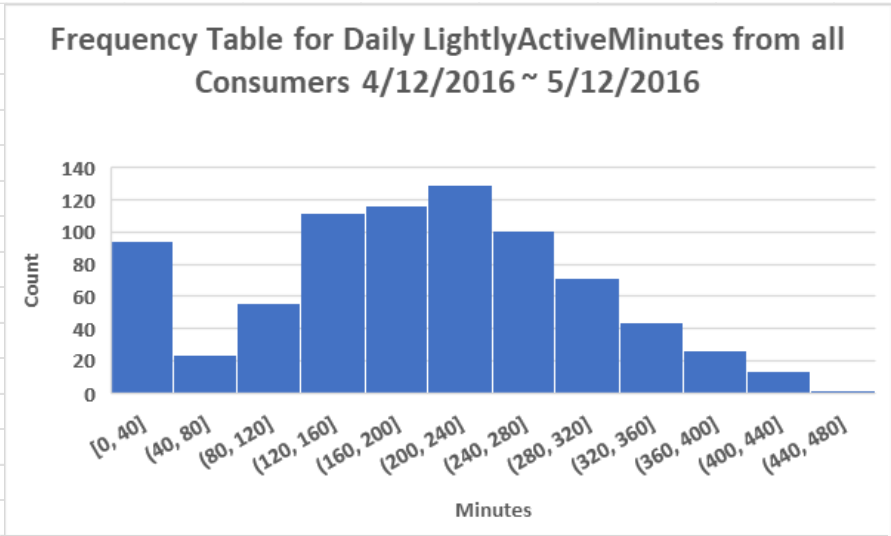
- The Daily LightlyActiveMinutes histogram shows that the most common daily LightlyActiveMinutes values are between **200 ~ 240** minutes for consumers in this dataset.

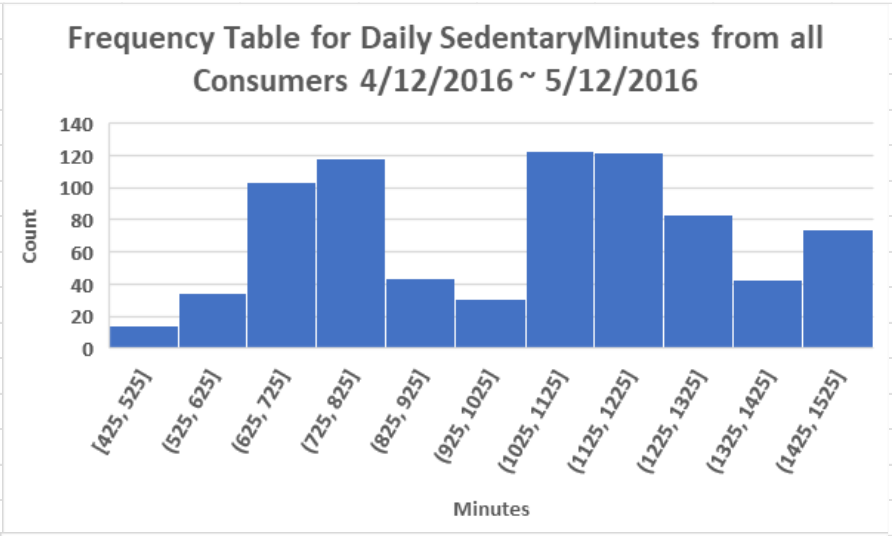
- The Daily LightlyActiveMinutes boxplot shows that the **median daily LightlyActiveMinutes** value is **199**.

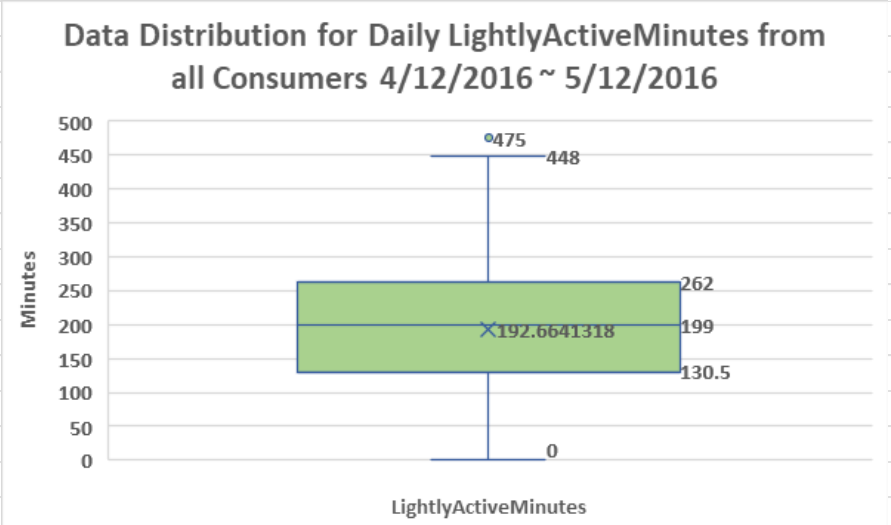
- The Daily SedentaryMinutes histogram shows that the most common values for daily SedentaryMinutes are between **1,025 ~ 1,125 minutes.**

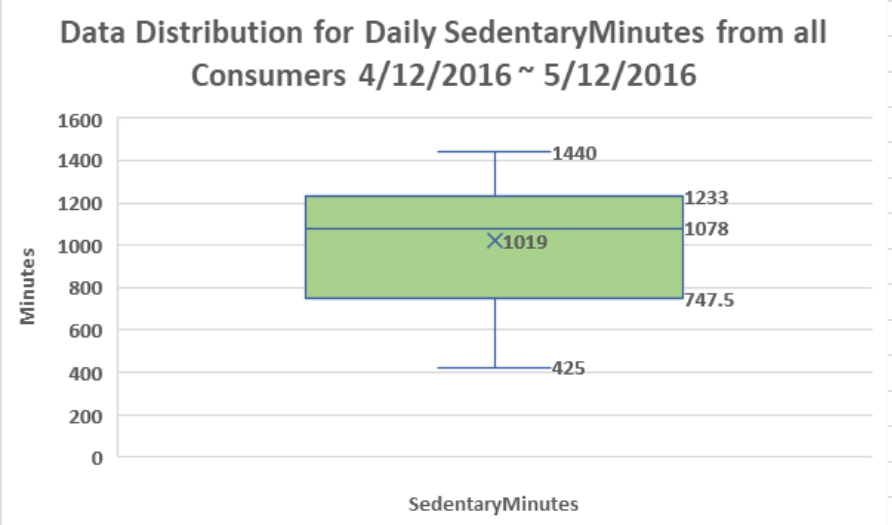
- The Daily SedentaryMinutes boxplot shows that the **median daily SedentaryMinutes** value is 1,078 minutes.

Data Analysis: Daily Activity Tracking





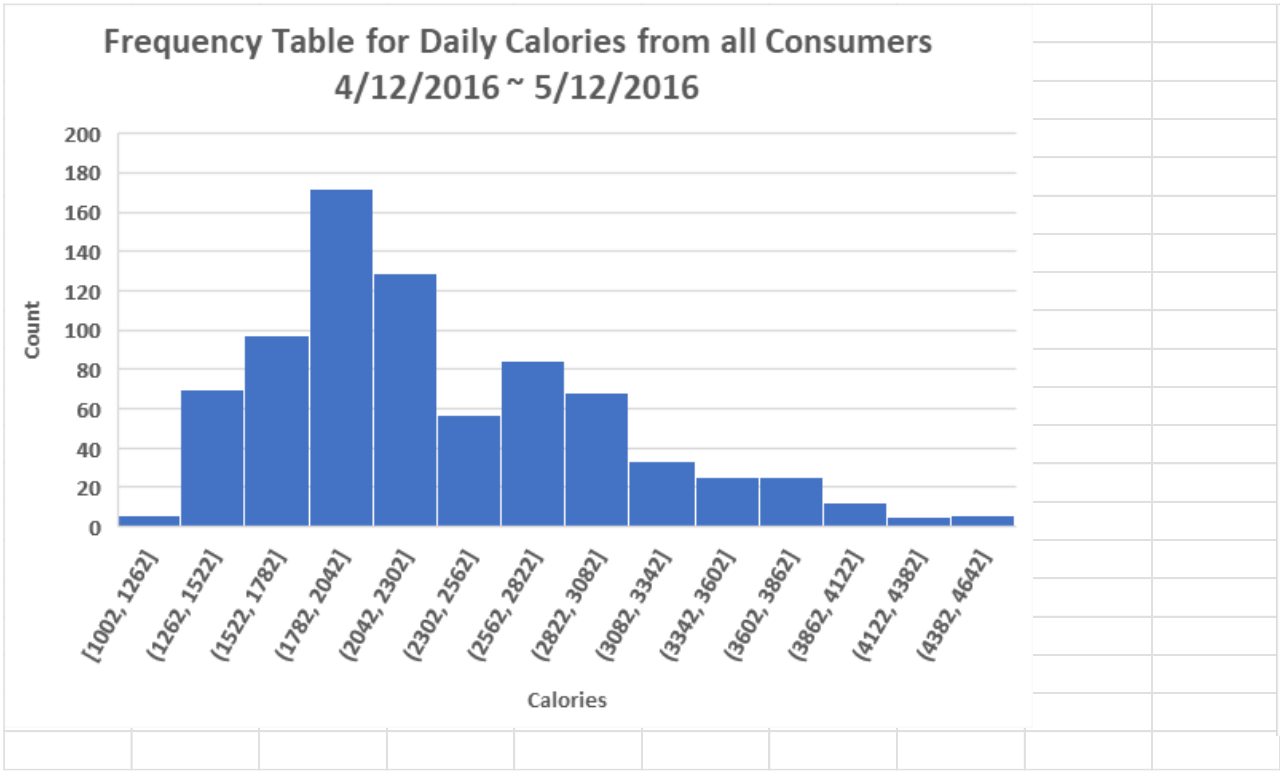


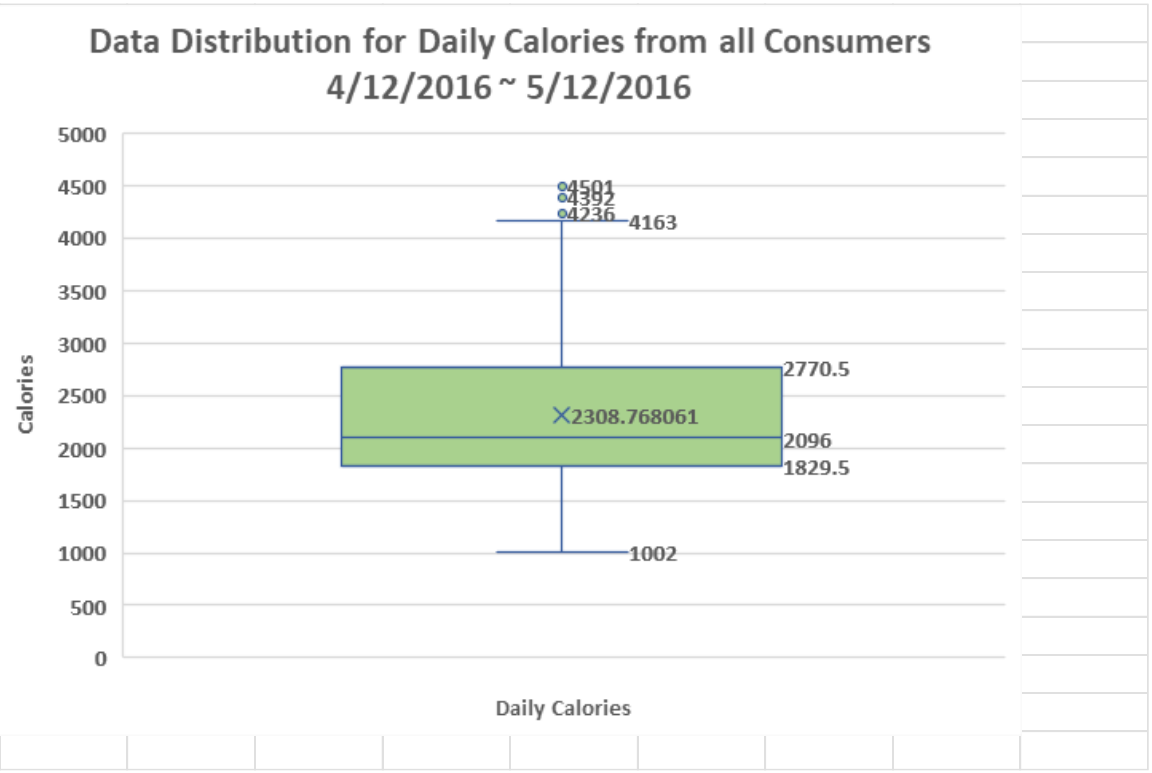


Data Analysis: Daily Activity Tracking

- Made some plots to see how Calories data are distributed.
  - The Daily Calories histogram shows that the most common set of daily Calories values are between **1,782 ~ 2,042 calories** for consumers in this dataset.
  - The Daily Calories boxplot shows that the **median daily calories** value is **2,096**. There are **some outliers above the upper whisker**, which means there were some consumers who burned much more calories than others.

Data Analysis: Daily Activity Tracking





Data Analysis: Daily Activity Tracking

- **Sorted** the worksheet by the **Calories** column and found the largest number in the column was **4,552**. This data related to consumer 5577150313.
- Input other numbers tracked for the day such as **TotalSteps**, **TotalDistance,** total walking time per activity level, and **weight** information (from another file weightLogInfo_merge.csv) into two **steps-to-calories converters** Walking Calorie Calculator & How to Count and Track Calories Burned Walking to get estimated values which were between **1,800 to 2,300.**
- Added **1,819** to those numbers. The calorie values got (**3,619 or 4,319**) are **still less than 4,552**. 1,819 is the calorie value found in the data tracking the activity of consumer 5577150313 on 5/7/2016 - the consumer did not walk that day (total steps and total distance were both 0). 1,819 could be the basal metabolic rate (or the energy/calories) for consumer 5577150313 (means the consumer needs that number of calories to perform basic body functions).
- Another consumer's daily activity data was input into the same calorie calculators mentioned above and the calculator's estimated calories were found to be lower than the calorie values found in this dataset.
- According to this article 6 Factors That Can Affect How Many Calories You Burn and this article What Affects How Many Calories You Burn? 6 Factors to Consider, factors like body weight, **muscle mass**, **age**, etc. can affect how much calories you burn. **No information on muscle mass or age was found in this dataset, so we were limited in performing the following tasks:**
  - **determining the accuracy of the calorie values in this dataset**
  - **discovering the relationship between calorie burned and other attributes tracked here** (like TotalSteps, TotalDistance, walking distance per activity level, time spent per activity level, etc.).

  Therefore, **we didn't do much analysis of calories in this report**.

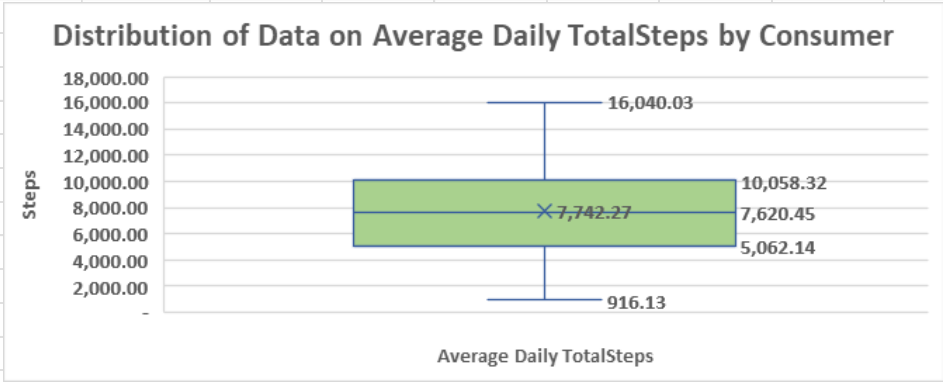
Data Analysis: Daily Activity Tracking

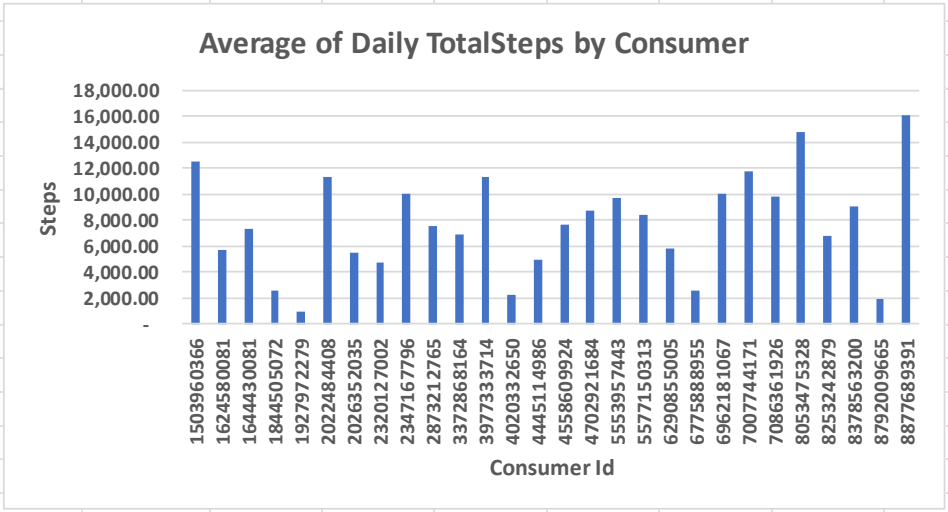
- Created a PivotTable. Aggregated data by consumer and calculated the average of daily TotalSteps for each consumer.
- Made a boxplot to check the distribution of data on the average of daily TotalSteps taken by each consumer and found that the **3rd quartile is around 10,058** which means nearly **75% of the consumers walked fewer than 10,000 steps a day**.
- Conditional formatting the Average of TotalSteps column to display the top 20% and bottom 20% values in different colors.

Data Analysis: Daily Activity Tracking

Row Labels	Average of TotalSteps	Average of TotalDistance
1503960366	12,520.63	8.07
1624580081	5,743.90	3.91
1644430081	7,282.97	5.30
1844505072	2,575.96	1.70
1927972279	916.13	0.63
2022484408	11,370.65	8.08
2026352035	5,456.07	3.39
2320127002	4,716.87	3.19
2347167796	10,077.18	6.73
2873212765	7,555.77	5.10
3372868164	6,861.65	4.71
3977333714	11,337.62	7.76
4020332650	2,267.23	1.63
4445114986	4,930.83	3.34
4558609924	7,685.13	5.08
4702921684	8,766.07	7.11
5553957443	9,676.31	6.34
5577150313	8,451.55	6.32
6290855005	5,851.32	4.43
6775888955	2,541.80	1.83
6962181067	10,001.73	6.73
7007744171	11,776.36	8.34
7086361926	9,766.07	6.67
8053475328	14,763.29	11.48
8253242879	6,842.28	4.93
8378563200	9,088.14	7.21
8792009665	1,919.93	1.23
8877689391	16,040.03	13.21
Grand Total	**7,757.45**	**5.54**

12,520.63
5,743.90
7,282.97
2,575.96
916.13
11,370.65
5,456.07
4,716.87
10,077.18
7,555.77
6,861.65
11,337.62
7,685.13
4,930.83
9,676.31
8,766.07
8,451.55
5,851.32
2,541.80
10,001.73
11,776.36
9,766.07
14,763.29
6,842.28
9,088.14
1,919.93
16,040.03





Data Analysis: Daily Activity Tracking

- Used a nested IF function to classify the “Activity Level” for each consumer based on the following guideline on steps and activity levels described in this article How Many Steps a Day Is Considered Active?
  - **Sedentary:** Less than 5,000 steps daily
  - **Low active:** About 5,000 to 7,499 steps daily
  - **Somewhat active:** About 7,500 to 9,999 steps daily
  - **Active:** More than 10,000 steps daily
  - **Highly active:** More than 12,500 steps daily
- Counted the number of consumers at each activity level and created a pie chart. **25%** of the consumers were at “**Somewhat Active**” level and **25%** of the consumers were at “**Sedentary**” level.
- According to this article How many steps should people take per day?, CDC recommends that most adults aim for **10,000** steps per day for health benefits. For most people, this is the equivalent of about 8 kilometers, or **5** miles.

Data Analysis: Daily Activity Tracking

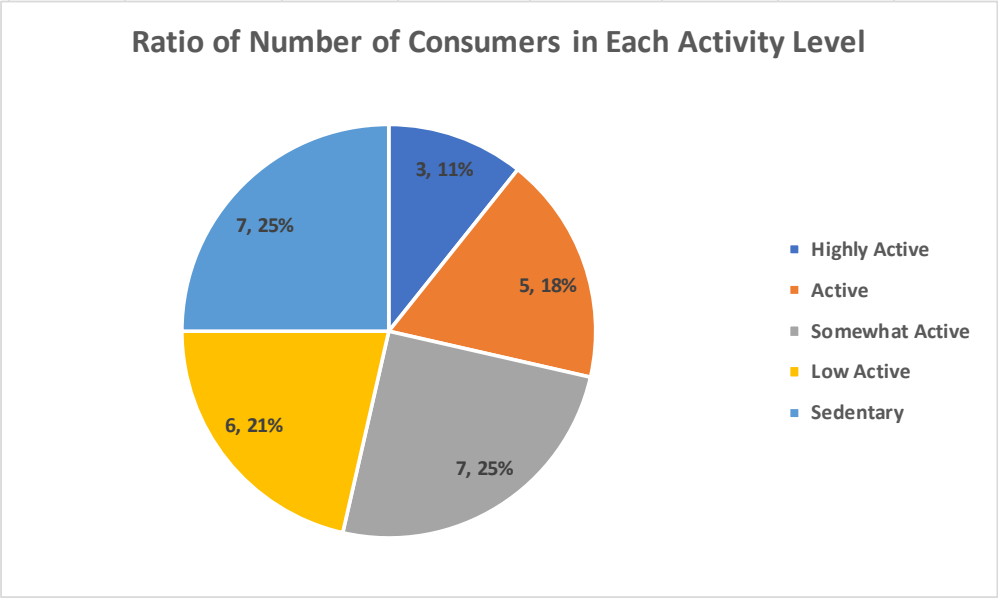
- Did some calculations and found that approximately **71%** of consumers in the dataset took **less than 10,000 steps** per day. According to this articles How many steps should people take per day?, most people in the United States only take 3,000–4,000 steps per day. Therefore, Bellabeat's marketing strategy should include **encouraging people to purchase and wear Bellabeat's smart devices to track/count the number of steps they take every day to ensure they achieve their goal of walking at least 10,000 steps a day.**

Data Analysis: Daily Activity Tracking

Row Labels	Average of TotalSteps	Activity Level
1503960366	12,520.63	Highly Active
1624580081	5,743.90	Low Active
1644430081	7,282.97	Low Active
1844505072	2,575.96	Sedentary
1927972279	916.13	Sedentary
2022484408	11,370.65	Active
2026352035	5,456.07	Low Active
2320127002	4,716.87	Sedentary
2347167796	10,077.18	Active
2873212765	7,555.77	Somewhat Active
3372868164	6,861.65	Low Active
3977333714	11,337.62	Active
4020332650	2,267.23	Sedentary
4445114986	4,930.83	Sedentary
4558609924	7,685.13	Somewhat Active
4702921684	8,766.07	Somewhat Active
5553957443	9,676.31	Somewhat Active
5577150313	8,451.55	Somewhat Active
6290855005	5,851.32	Low Active
6775888955	2,541.80	Sedentary
6962181067	10,001.73	Active
7007744171	11,776.36	Active
7086361926	9,766.07	Somewhat Active
8053475328	14,763.29	Highly Active
8253242879	6,842.28	Low Active
8378563200	9,088.14	Somewhat Active
8792009665	1,919.93	Sedentary
8877689391	16,040.03	Highly Active
Grand Total	**7,757.45**	

Activity Level	Count
Highly Active	3
Active	5
Somewhat Active	7
Low Active	6
Sedentary	7

Ratio of Consumer with Average TotalSteps < 10,000 per day **0.714286**



Data Analysis: Daily Activity Tracking

- Selected one consumer for each activity level from those whose activity was tracked daily from 4/12/2016 to 5/12/2016.
- Created a line chart of the total daily steps taken by these consumers.
- Created a line chart of calories consumed by these consumers per day.
- From these line charts, we can see that there is a **positive correlation** between the number of steps taken and the calories burned.
- However, according to some articles like 6 Factors That Can Affect How Many Calories You Burn and What Affects How Many Calories You Burn? 6 Factors to Consider, factors such as **age**, **body mass**, **weight** etc. can affect the number of calories you burn. **A younger/heavier person may burn more calories than an older/lighter person while performing the same exercise**.
- What's interesting in the line chart is that a consumer with a "sedentary" activity level (ID 4020332650) consumed more calories than consumers with a "low active" or "somewhat active" level. Consumer 4020332650 might be younger or heavier. Therefore, Bellabeat's marketing strategy should include **encouraging people who need to burn more calories (such as those who are heavier or older) to purchase and wear Bellabeat's smart devices to track calorie burned for health or other benefits** (e.g., looking younger, slimmer or more energetic).

Data Analysis: Daily Activity Tracking