Outline: Comparative Analysis of Top Uses for Wearable Healthcare Related Trackers

Customer: Bellabeat

Industry: Google Data Analysis Career Certification Capstone

Links:

Tableau: https://public.tableau.com/app/profile/brianna.goulbounrne/viz/GoogleCareerCertificationCapstoneBellaBeat/Sheet4

Changelog:

Author: Brianna Goulbourne

Published Date: December 2023

1. ASK Phase

1.1 Client information:

The client, Bellabeat is a wellness corporation with a focus on providing specifically designed wearables and other health related products crafted for women.

1.2 Stakeholders:

Urška Sršen, cofounder, Chief Creative Officer and Sando Mur, Cofounder, would like a thorough analysis of some of the other top brands of healthcare's wearables and wellness related smart products, to learn about how consumers are using these smart devices.

The analyst conducted a comparative analysis of wearable usage of a similar product sold by another fitness brand

1.3 Business task:

Through data-driven analysis the analyst hopes to identify a set of recommendations to help guide a marketing structure for the company based on any observed trends in healthcare smart device usage.

2. Prepare Phase

2.1 Data Usage:

The data used in this analysis is the FitBit Fitness Tracker Data downloaded from Kaggle where it was uploaded by a user with the name Mobius. This dataset contains responses obtained from 30 Fitbit users and includes data about physical activity, heart, caloric intake, and sleep.

2.2 Data storage, licensing, privacy, security, and accessibility:

This data was obtained from Kaggle and made Public and available for use by Mobius under the CC0: Public Domain Creative Common License. During the uploading of this dataset, the owner has waived all rights under the copyright law. This leaves this dataset open to copy, modification, distribution, and commercial purposes without the need to seek for permissions. The author has cited the following reference in relation to the database; Furberg, Robert; Brinton, Julia; Keating, Michael ; Ortiz, Alexa [https://zenodo.org/record/53894#.YMoUpnVKiP9]

2.3 Data organization

This file consists of several excel sheets with large amounts of data that is organized in a both long and wide format. For the purposes of this analysis, I have chosen to operate only with long data and have formatted it to only include only the necessary information needed for my evaluation. The following are the spreadsheets that we will be working with prior to data cleaning:

- dailyActivity_merged
- weightLogInfo_merged
- sleepDay_merged
- heartRate_seconds_merged
- dailySteps_merged
- dailyCalories_merged

2.4 Data credibility and limitations

The dataset we are operating with is indeed coming from a reliable source. The dataset was recommended by the client and uploaded to a public platform with cited sources attached. A few limitations of the data are that it was obtained in 2016, making it outdated. The dataset also contains only 30 participants, while this is the smallest number that can represent a population, a larger sample size would have led to less bias.

3. Process Phase

3.1 Data cleaning:

The sample has been cleaned via the removal of any inconsistencies, missing, duplicated or inaccurate data, and the fixing of any typos or spelling errors. These modifications have been conducted using both SQL and Microsoft Excel. A few examples of processes conducted within this phase are documented below. An in-depth documentation of all data alterations can be found in the attached changelog.

- SQL:
  - Used ‘SELECT DISTINCT’ query to count how the number of unique ID’s in all datasets.

- This brought back 33 unique ID’s codes within the daily activity, daily calories, and daily_steps datasets. The daily_sleep data set contained 24 unique ID’s. The daily_heart rate contained 7 unique values, and the weight_log dataset contained 8 unique values.

3.2 Ensuring data integrity:

Consistency: Upon initial analysis of the data, there were a several inconsistency errors such as a lack of proper decimal placement across certain values. This was corrected during the cleaning phase and all values have been limited to 2 decimal places. There was also an inconsistency in the noted amount of people surveyed, and the analysis will take place with the three datasets that all contain 33 units of surveyed data.

Completeness: All relevant data has been included and cleaned. The following datasets have been omitted due to a lack of relevance or due to repetitive contents;

- dailyIntensities_merged.csv
- hourlyCalories_merged.csv
- hourlyIntensities_merged.csv
- hourlySteps_merged.csv
- minuteCaloriesNarrow_merged.csv
- minuteCaloriesWide_merged.csv
- minuteIntensitiesNarrow_merged.csv
- minuteIntensitiesWide_merged.csv
- minuteMETsNarrow_merged.csv
- minuteSleep_merged.csv
- minuteStepsNarrow_merged.csv
- minuteStepsWide_merged.csv
- weightLogInfo_merged.csv
- sleepDay_merged.csv
- heartRate_seconds_merged.csv

- The following data sets will be used for further analysis:
  - Daily_activity_merged.csv was renamed as daily_activity
  - Daily_steps_merged.csv was renamed as daily_steps
  - Daily_calories_merged.csv was renamed as daily_calories

Accuracy: Since the data was obtained via survey, there is no true measure of accuracy.

3.3 Chosen Tools: SQL, Tableau and Excel

The tools that I have chosen to use for this analysis are SQL, tableau, and a little bit of Microsoft excel for my initial analysis. SQL is great for managing large datasets and skimming through them quickly and efficiently. Tableau will provide a great the tools I need to create a detailed and comprehensive visualization to present to the client.

4. Analyze and Share Phase

4.1 Data preparation in SQL

We begin our analysis by creating a brand-new dataset folder within the one of the most widely used structured query language platforms BigQuery. The bellabeat_capstone_project folder dataset was established. Into this dataset, the following three tables were uploaded into the dataset; daily_activity, daily_calories, and daily_steps.

4.2 Establishment of primary uses of Fitbit device

The analysis began by establishing what the primary usage of this device was. This was calculated using the SELECT DISTINCT ID FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_calories ` query in SQL. The pie chart below (Figure 4.1) shows the distinct count of users for each category.

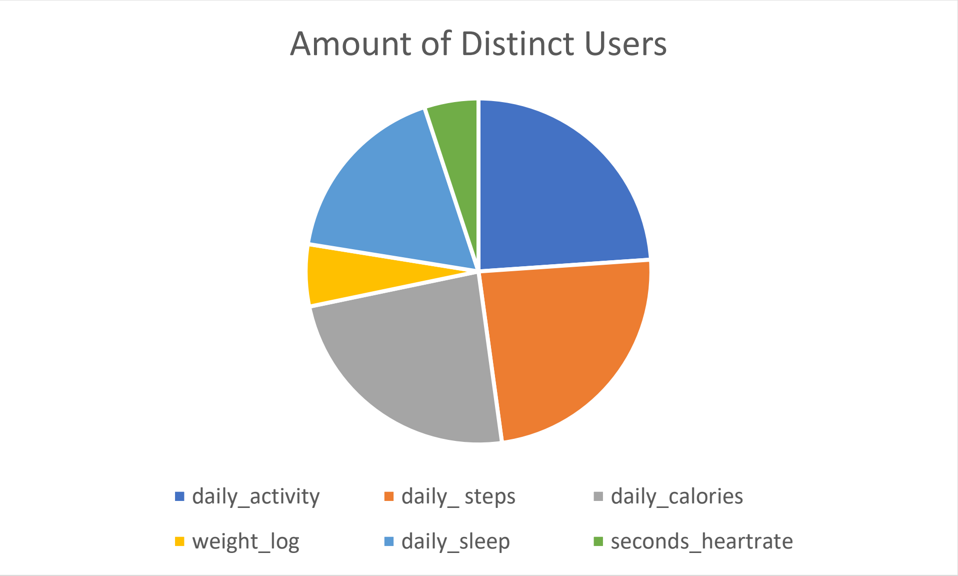


Figure 4.1

4.3 User Summary Statistics

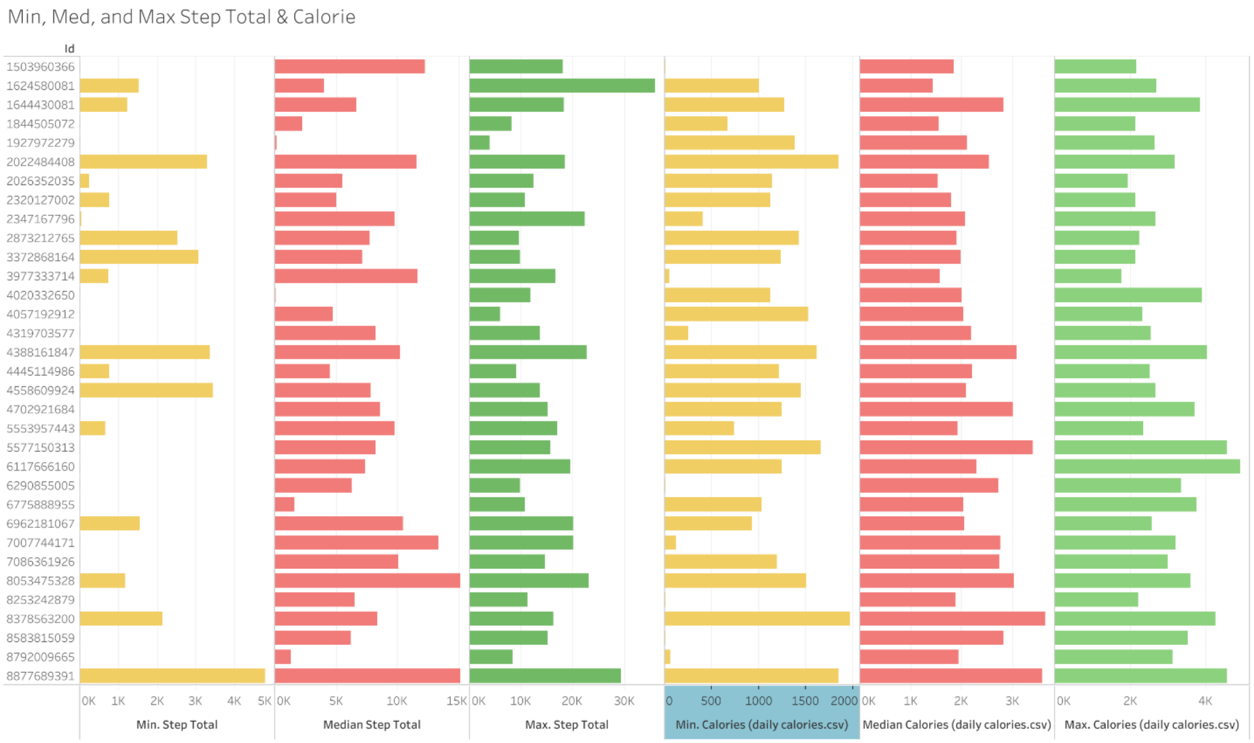
We begin our analysis with obtaining some summary statistics including total number of rows, distinct values, a maximum, minimum, and mean value for the tables. These values were obtained using the following SQL statements:

```
SELECT COUNT (*) as count_daily_activity
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity `
SELECT COUNT (*) as count_daily_calories
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_calories `
SELECT COUNT (*) as count_daily_steps
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_steps`

SELECT MAX(totaldistance)
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity `
SELECT MIN(totaldistance)
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity
```

```
SELECT AVG (totaldistance)
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity `

SELECT MIN(steptotal) FROM `bigquery-practice-
101099.bellabeat_capstone_project.daily_steps`
SELECT Max(steptotal) FROM `bigquery-practice-
101099.bellabeat_capstone_project.daily_steps`
SELECT AVG (steptotal)
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_steps`

SELECT MIN(calories) FROM `bigquery-practice-
101099.bellabeat_capstone_project.daily_calories `
SELECT MAX(calories) FROM `bigquery-practice-
101099.bellabeat_capstone_project.daily_calories `
SELECT AVG(calories) FROM `bigquery-practice-
101099.bellabeat_capstone_project.daily_calories `
```

The following table (Figure 4.2) contains the data acquired from the SQL statements.

Figure 4.2



4.4 User Frequency

I then wanted to evaluate how many times each user logged any data to get a sense of how often people were using their Fitbit. The following SQL statement was used to do so.

```
SELECT id, count(id) AS total_id_count
```

```
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity `
                                        group by id
```

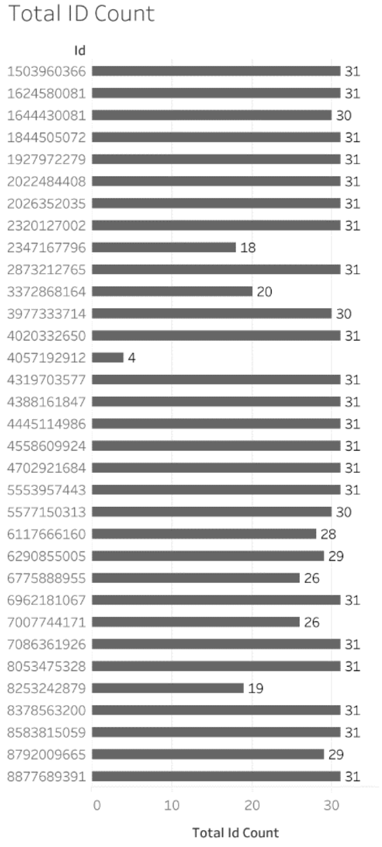
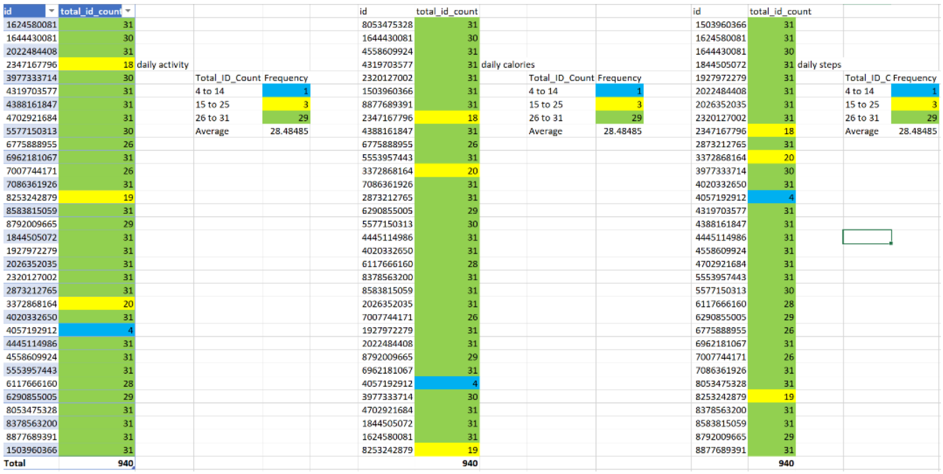


Figure 4.4



id	total_id_count				id	total_id_count				id	total_id_count			
1624580081	31				8053475328	31				1503960366	31			
1644430081	30				1644430081	30				1624580081	31			
2022484408	31				4558609924	31				1644430081	30			
2347167796	18	daily activity			4319703577	31	daily calories			1844505072	31	daily steps		
3977333714	30		Total_ID_Count	Frequency	2320127002	31		Total_ID_Count	Frequency	1927972279	31		Total_ID_C	Frequency
4319703577	31		4 to 14	1	1503960366	31		4 to 14	1	2022484408	31		4 to 14	1
4388161847	31		15 to 25	3	8877689391	31		15 to 25	3	2026352035	31		15 to 25	3
4702921684	31		26 to 31	29	2347167796	18		26 to 31	29	2320127002	31		26 to 31	29
5577150313	30		Average	28.48485	4388161847	31		Average	28.48485	2347167796	18		Average	28.48485
6775888955	26				6775888955	26				2873212765	31			
6962181067	31				5553957443	31				3372868164	20			
7007744171	26				3372868164	20				3977333714	30			
7086361926	31				7086361926	31				4020332650	31			
8253242879	19				2873212765	31				4057192912	4			
8583815059	31				6290855005	29				4319703577	31			
8792009665	29				5577150313	30				4388161847	31			
1844505072	31				4445114986	31				4445114986	31			
1927972279	31				4020332650	31				4558609924	31			
2026352035	31				6117666160	28				4702921684	31			
2320127002	31				8378563200	31				5553957443	31			
2873212765	31				8583815059	31				5577150313	30			
3372868164	20				2026352035	31				6117666160	28			
4020332650	31				7007744171	26				6290855005	29			
4057192912	4				1927972279	31				6775888955	26			
4445114986	31				2022484408	31				6962181067	31			
4558609924	31				8792009665	29				7007744171	26			
5553957443	31				6962181067	31				7086361926	31			
6117666160	28				4057192912	4				8053475328	31			
6290855005	29				3977333714	30				8253242879	19			
8053475328	31				4702921684	31				8378563200	31			
8378563200	31				1844505072	31				8583815059	31			
8877689391	31				1624580081	31				8792009665	29			
1503960366	31				8253242879	19				8877689391	31			
Total	940					940					940			

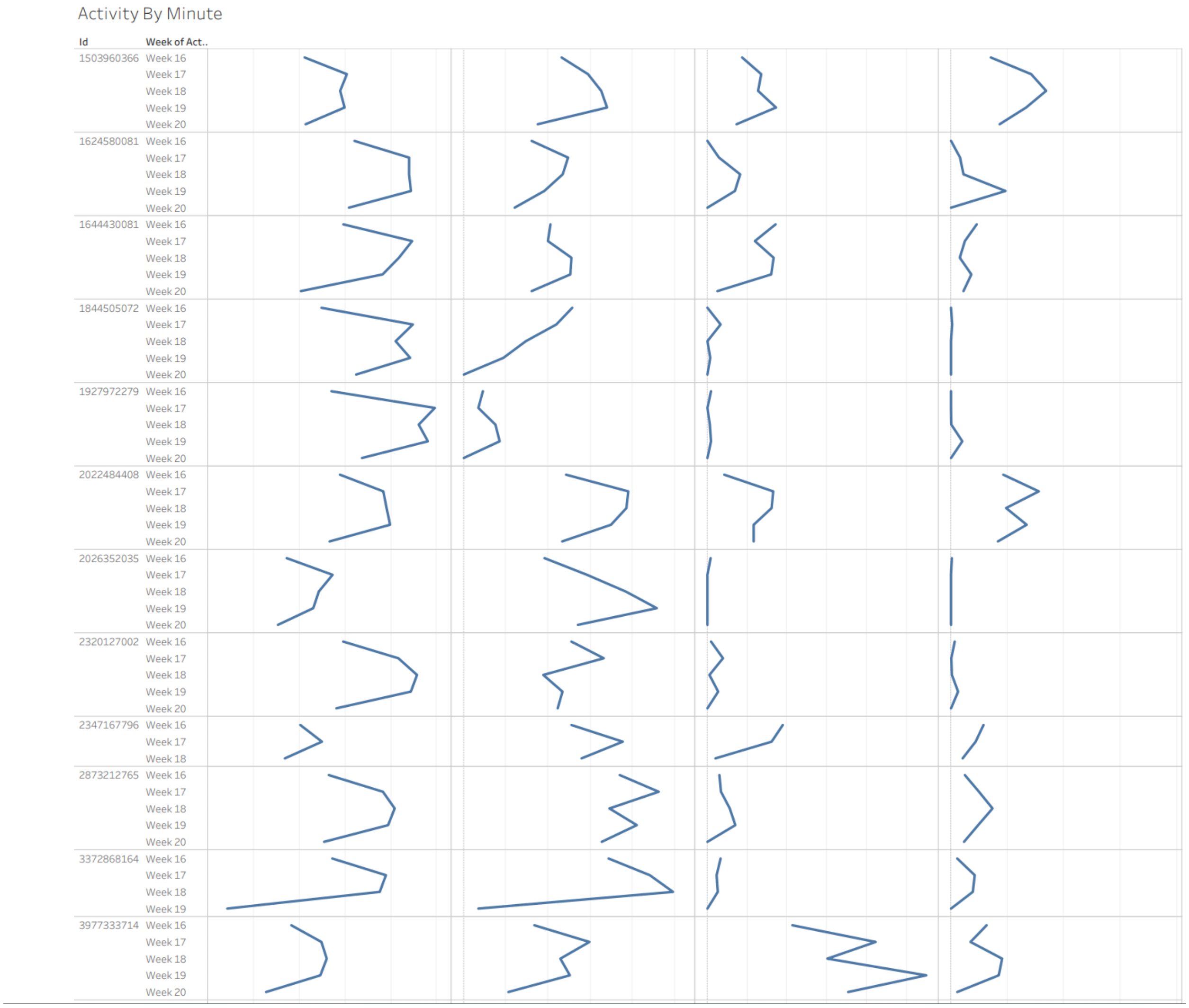
Figure 4.5

Graph (Figure 4.4) displays total ID count for total distance within the daily_activity table. Using Tableau Public, we were able to input the data obtained from SQL and generate a bar chart to analyze the frequency in which users were logging their data.

Conditional formatting was used (Figure 4.5) to analyze the data obtained from SQL in Excel where user frequency is denoted by color.

4.5 User Frequency per week

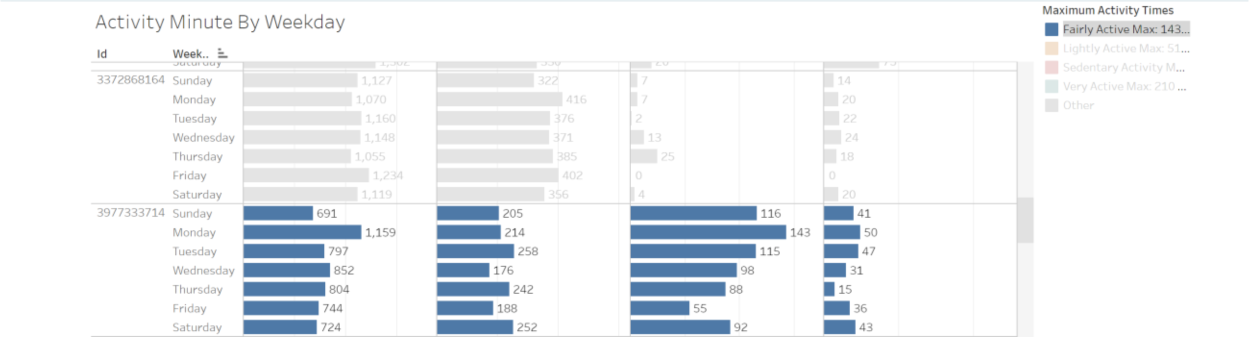
I then wanted to see if the frequency of use increased across the weeks of the trial, to see if using a wearable device encouraged users to remain active longer and for father distances.

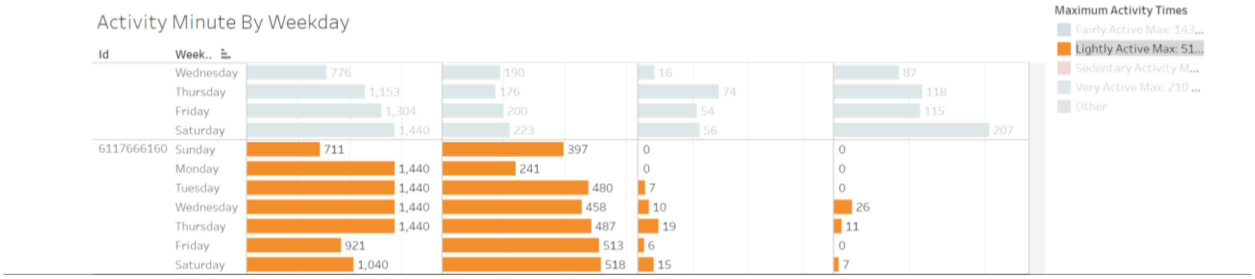


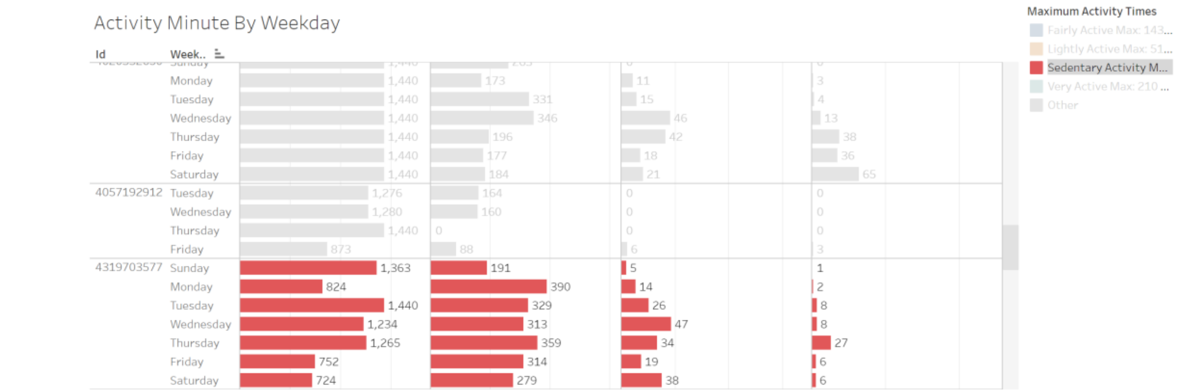
I also investigated whether there was an increase in usage on weekends opposed to weekends. The following SQL queries were conducted to find the maximum activity conducted per minute within each category of activity level.

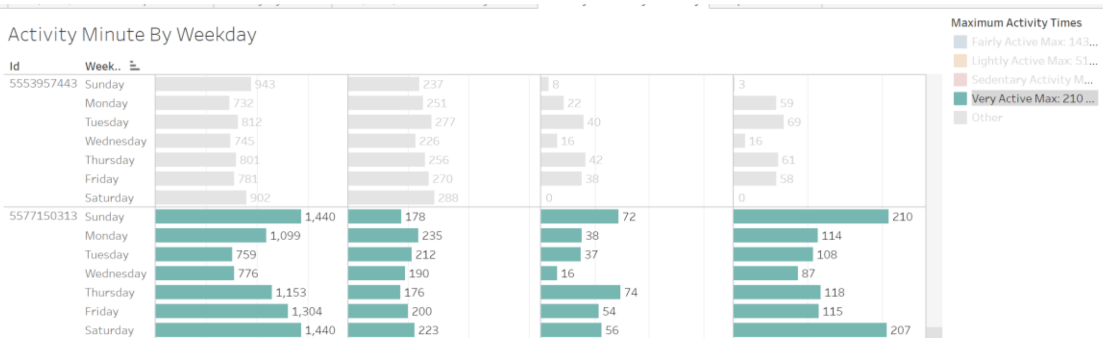
```
SELECT ID, activitydate, MAX(SedentaryMinutes) AS max_sedentary_min
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity `
group by id, activitydate
order by max_sedentary_min desc


SELECT ID, activitydate, MAX(fairlyactiveMinutes) AS max_fairlyactive_min
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity `
group by id, activitydate
order by max_fairlyactive_min desc


SELECT ID, activitydate, MAX(lightlyactiveMinutes) AS max_lightlyactive_min
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity `
```

```
group by id, activitydate
order by max_lightlyactive_min desc


SELECT ID, activitydate, MAX(veryactiveminutes) AS max_very_active_min
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity `
group by id, activitydate
order by max_very_active_min desc
```









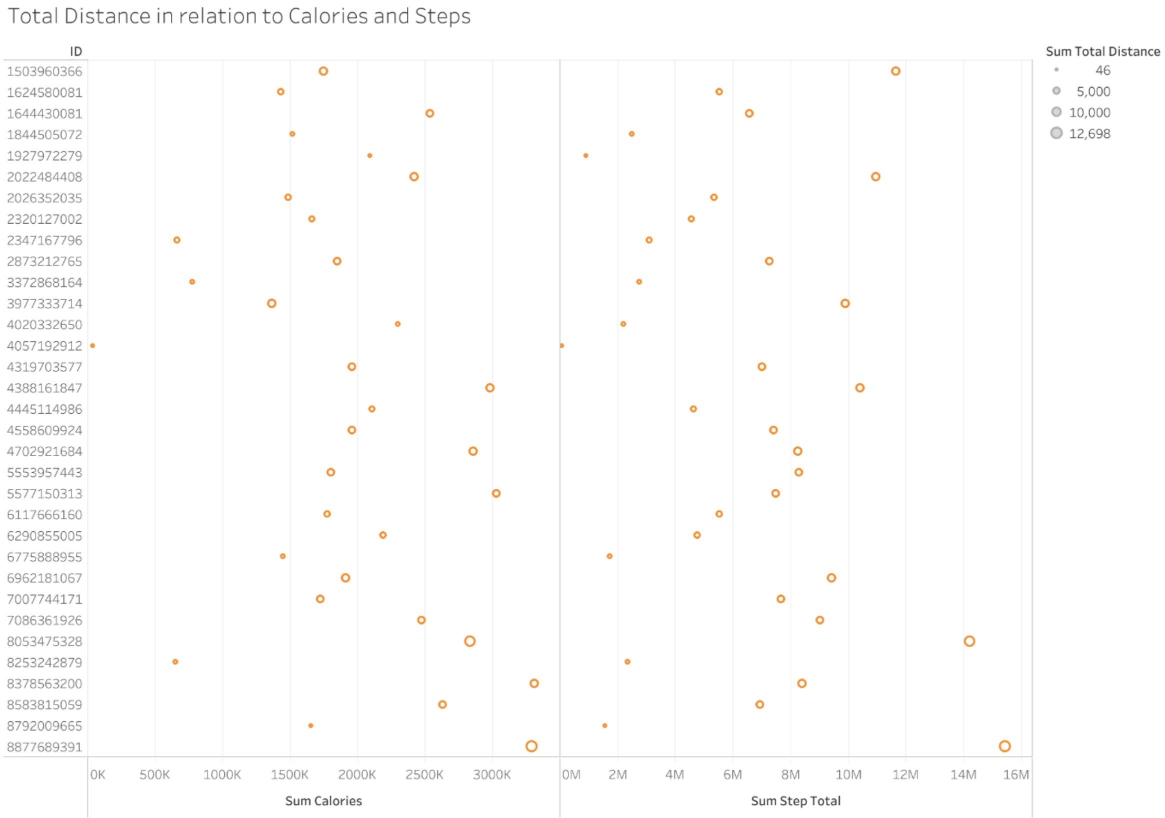
4.6 Dataset comparison’s

I then wanted to evaluate all three tables as a whole and identify any trends.

To do this, I created JOINs within the SQL platform to aggregate any similar data from the daily_activity and daily_steps tables. I used a left join to pull matching data from the two tables using ID numbers to link the two. I queried for the sum’s of both distance and step total to evaluate these data points for each distinct ID. The query is written below:

```
SELECT DISTINCT DA.ID, sum (DA.totaldistance) as sum_total_distance, sum(DS.steptotal)
as sum_step_total
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity ` as DA
left join bigquery-practice-101099.bellabeat_capstone_project.daily_steps as DS
ON DS.ID = DA.id
group by DA.ID
order by sum_total_distance desc
```

```
SELECT DISTINCT DA.ID, sum (DA.totaldistance) as sum_total_distance, sum(DC.calories)
as sum_calories
FROM `bigquery-practice-101099.bellabeat_capstone_project.daily_activity ` as DA
left join bigquery-practice-101099.bellabeat_capstone_project.daily_calories as DC
ON DA.ID = DC.id
group by DA.ID
order by sum_total_distance desc
```



5. ACT Phase

Based on the analysis conducted I have concluded the following:

Conclusion 5.1:

Due to low numbers of participation for certain datasets, we only proceeded forward in our analysis only with the daily_activity, daily calories, and daily_steps datasets as these seem to be what users are using their Fitbit for the most. The other datasets do not contain an adequate amount of data to move forward with analysis. I would recommend that Bellabeat conduct their own study with a larger sample size for further analysis. I would also recommend that the client focus mostly on developing software to monitor step, distance, and calories as this seems to be what users are most interested in logging.

	Daily_steps	Daily_activities (total distance)	Daily_calories
MAX	36019	28.03	4900
MIN	0	0	0
AVG	7637.910638	5.489702128	2303.609574
COUNT	33	33	33

Conclusion 5 .2:

It seems that a vast majority of users logged their data an average of 28 times over the 30-day period. This remained consistent across the additional tables as well. I suggest that the client tap into this market of current wearable users to maintain a high frequency of uses. I believe that pushing the narrative of women-centric wearable to women already interested in the use of wearable devices will be the best course of action.

Conclusion 5.3:

Sedentary Minutes and Fairly Active maximums were both during weekdays on Tuesday and Monday respectively. Very active and lightly active maximums were both fell on a weekend, Sunday, and Saturday respectively. Pushing notification's out to users to increase movement and usage during those days' weekdays may increase and promote usage on days with high levels of sedentary minutes.

This data also shows that most users hit a peak around week 19 after a couple weeks of using the product.

Conclusion 5.4:

There seems to be a direct correlation between total distance and number of calories burned and number of steps taken. Creating a software that show's the correlation between these factors may influence customers to use these products for a longer duration of time.