Bellabeat Case Study: Fitness or General Use

Joseph Robinson

2024-04-01

Executive Summary:

The purpose of this case study is to provide Bellabeat stakeholders insights into user preferences for industry-wide smart devices, to then use those insights to inform their marketing analytics team how best to proceed in the marketing of the Bellabeat Time product. The findings of this study are as follows:

- Users prefer the Apple Watch over the FitBit, based on their trend towards generalized use.
- Apple Watch uses accounted for 77% of activity in the data, and that activity was fairly equal between categories.
- FitBit users tended to use their device for fitness purposes, with the females aged 30 and under being a key demographic.
- Female users perform more steps by average than males users across both devices. Suggesting a tendency towards fitness.
- There is strong evidence to suggest marketing the Bellabeat Time as a general use smart device would cater to the widest demographic.
- However, if the fitness market is the goal, tailor your marketing strategy to female users 30 and under.

Ask

Guiding questions:

- What is the problem you are trying to solve?
- How can your insights drive business decisions?

Business task:

Urska Srsen tasked me to analyze smart device user data to gain insight into how consumers use non-Bellabeat smart devices. She, then, wants us to apply these finding to one Bellabeat product. The product I chose to showcase is Time, the Bellabeat market equivalent to the FitBit and Apple Watch.

Stakeholders:

- Urska Srsen: Bellabeat's co-founder and Chief Creative Officer

- Sando Mur: Mathematician and Bellabeat's co-founder; key member of the executive board
- Bellabeat marketing analytics team

Key questions:

- What are some trends in smart device usage?
- How could these trends apply to Bellabeat customers?
- How could these trends help influence the Bellabeat marketing strategy?

Assumptions:

- Apple Watch has brand recognition, which could skew results towards Apple products
- FitBit's market towards fitness
- Apple Watches are marketed more towards general use, and brand status

Prepare:

Guiding questions:

- Where is your data stored?
- How is the data organized? Is it long or wide format?
- Are there issues with bias or credibility in this data?

- How are you addressing licensing, privacy, security, and accessibility?
- How did you verify the data's integrity?
- How does it help you answer your question?
- Are there any problems with the data?

Data source:

FitBit Fitness Tracker Data: Pattern recognition with tracker data:: Improve Your Overall Health https://www.kaggle.com/datasets/arashnic/fitbit/data

At the outset of this assignment, the stakeholders requested that I first look at the FitBit Tracker Data data set, hosted on Kaggle, mentioned above. However, due to the following constraints, inconsistencies, and misalignments within that data set, after a discussion with the stakeholders, I decided to move away from this data set and secure a different data source all together.

Constraints, Inconsistencies, and Misalignments:

The FitBit Fitness Tracker Data data set hosted on Kaggle is a public domain second-party data source that was created by surveying 30 FitBit users, including minute-level output for physical activity, heart rate, and sleep monitoring. The data also breaks down user activity based on behavior and preferences. Things to note, data set inconsistencies were apparent from the onset, as the date range 03/12/2016-05/12/2016 posted on in the "Content" details on the Kaggle profile is misaligned to the actual start and end date of the data records, which are 04/12/2016 to 05/12/2016.

Additionally, on a more serious level, the "ID" column within data set parses the data to individuals "based on session", with no further explanation as to whether these are assigned statically or dynamically. Moreover, the "timestamp" column, as described in the "Content" description, but named "Date" in the data set is inconsistent between datasets, especially if you parse the date from time in separate columns to validate entries. One example of data misalignment is the "heartrate_seconds_merged" .csv, which has three columns, "Id", "Time", and "Value". If you parse the date and time in the "Time" variable you find all of the heart rate data records account for a 24-hour period between 04/12/2016 to 04/13/2016. Even more, if you inspect and filter the "Id" section you will find that there are only 8 unique values, out of 30 total users. To note, this does provide some insight into user preferences and user rate; however, the time-constraint on this data set limits the ability to aggregate this data to the other data sets calling into question the data's viability for the current task.

To test this, the data set was loaded with another into BigQuery to join them into one table in SQL:

SELECT

```
  heart_rate,
  total_steps,
  total_dist,
  very_active_mins,
  moderate_active_dist,
  light_active_dist_12,
  sedentary_mins,
  calories
```

FROM

```
  `jrob-capstone.bellabeat.heart_rate` AS heart
```

INNER JOIN `jrob-capstone.bellabeat.daily` AS daily

ON heart.id = daily.id

WHERE

```
  heart_rate != 0
```

Unfortunately, this merged data only served to further highlight how inconsistent and misaligned that data truly is. This was just one example of misaligned data sets within that source. Other notable constraints include, a lack of demographic details in all datasets, such as, age, and gender. Although, this analysis is industry-wide and tailored to non-Bellabeat smart devices, it would be best to include the demographics that Bellabeat markets to.

New data source:

Fuller, Daniel, 2020, “Replication Data for: Using machine learning methods to predict physical activity types with Apple Watch and Fitbit data using indirect calorimetry as the criterion.”, https://doi.org/10.7910/DVN/ZS2Z2J, Harvard Dataverse, V1

This data source, similar to the last one, surveyed 46 users (26 female), aged 21 to 44. The date source is broken into two .csv files, one for FitBit and the other for Apple Watch. The columns in both files were labeled “Heart”, “Calories”, “Steps”, “Distance”, “Age”, “Gender”, “Weight”, “Height”, and “Activity”. The “Activity” column is broken into the following ordinal categories, “0.Sleep”, “1.Sedentary”, “2.Light”, “3.Moderate”, and “4.Vigorous”. With a mix of quantitative and qualitative data types within data sets, this data source is more aligned to the business task, and even has a larger sample size.

Prepare

Guiding questions

- Where is your data stored?
- How is the data organized? Is it long or wide format?
- Are there issues with bias or credibility in this data?
- How are you addressing licensing, privacy, security, and accessibility?
- How did you verify the data’s integrity?
- How does it help you answer your question?
- Are there any problems with the data?

After I verified that is data source is viable for my needs, I decided to consolidate the two data sets into one .csv file in Excel. First, I opened both files, and changed the header row names to all lower case, so that it will be easier to ingest and run queries or write code in either BigQuery or R when it is clean. Additionally, I added in a “device” column, so that when I merge the two files the device, “FitBit” or “Apple Watch” will be attributed to the correct records within the data set. After that I copy and pasted the Apple Watch data set into the FitBit one and renamed the file “consolidated_fb_aw_data”. I did this so that I would not overwrite my raw data files, to ensure I have a back up in case something goes wrong down the line. Next, I imported the new .csv file into Google Sheets, because I am more familiar with their data cleaning tools, and I trimmed my data for any white space, removed duplicates and inspected my data for blanks and nulls. I had earlier established with my stakeholders that any null data can just be replaced with a “0” or “no response” in case any were revealed during the cleaning process. Once that was complete, I was ready to begin processing my data.

Processing:

Guiding questions:

- What tools are you choosing and why?
- Have you ensured your data’s integrity?
- What steps have you taken to ensure that your data is clean?

- How can you verify that your data is clean and ready to analyze?

- Have you documented your cleaning process so you can review and share those results?

Since I had already consolidated my data set into one file, my first thought was to ingest it into BigQuery to run queries on the data to draw out some insights. However, since my data is only 64,593 records and does not slow down Excel or Google Sheets, I opted to first look at my data on a pivot table in Google Sheets. I will then import my data into RStudio to further transform it, create data visualizations, and write my report using RMarkdown. I will, also, use Tableau to make a few visualizations to see which ones tell my data story the best. Lastly, I will make a PowerPoint presentation to present my findings.

Analyze:

Guiding questions:

- How should you organize your data to perform analysis on it?
- Has your data been properly formatted?
- What surprises did you discover in the data?
- What trends or relationships did you find in the data?
- How will these insights help answer your business question?

I imported my "consolidated_fb_aw_data" file into Google Sheets and made a pivot table to run some preliminary analysis on the data set. I first set up a table using the "device" and "calories" variables, placing "device" in row and "calories" as a value. I then looked at the SUM, AVERAGE, MIN, and MAX of the "calories" variable. I found that the "data"calories" variable favored FitBit over Apple Watch by a wide margin:

```
if(!require(tidyverse)) install.packages("tidyverse", repos =
"http://cran.us.r-project.org")

## Loading required package: tidyverse

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(tidyverse)

calorie_count <- read_csv("calorie_count.csv")
```

```
## Rows: 3 Columns: 5
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): device
## dbl (4): sum_calories, avg_calories, max_calories, min_calories
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
head(calorie_count)
```

```
## # A tibble: 3 × 5
##   device      sum_calories avg_calories max_calories min_calories
##   <chr>              <dbl>        <dbl>        <dbl>        <dbl>
## 1 Apple Watch       17935.        0.380         9.27   0
## 2 FitBit            28841.        2.39         12.6    0.00000944
## 3 Grand Total       46776.        0.790        12.6    0
```

Note, the “calories” variable is tied to each record and will vary upon the type of activity the user performed. With this in mind, I added in the “activity” variable into the column section of my pivot table:

```
activity_count <- read_csv("activity_count.csv")
```

```
## New names:
## Rows: 4 Columns: 7
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (7): avg_calories, activity, ...3, ...4, ...5, ...6, ...7
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `` -> `...3`
## • `` -> `...4`
## • `` -> `...5`
## • `` -> `...6`
## • `` -> `...7`
```

```
head(activity_count)
```

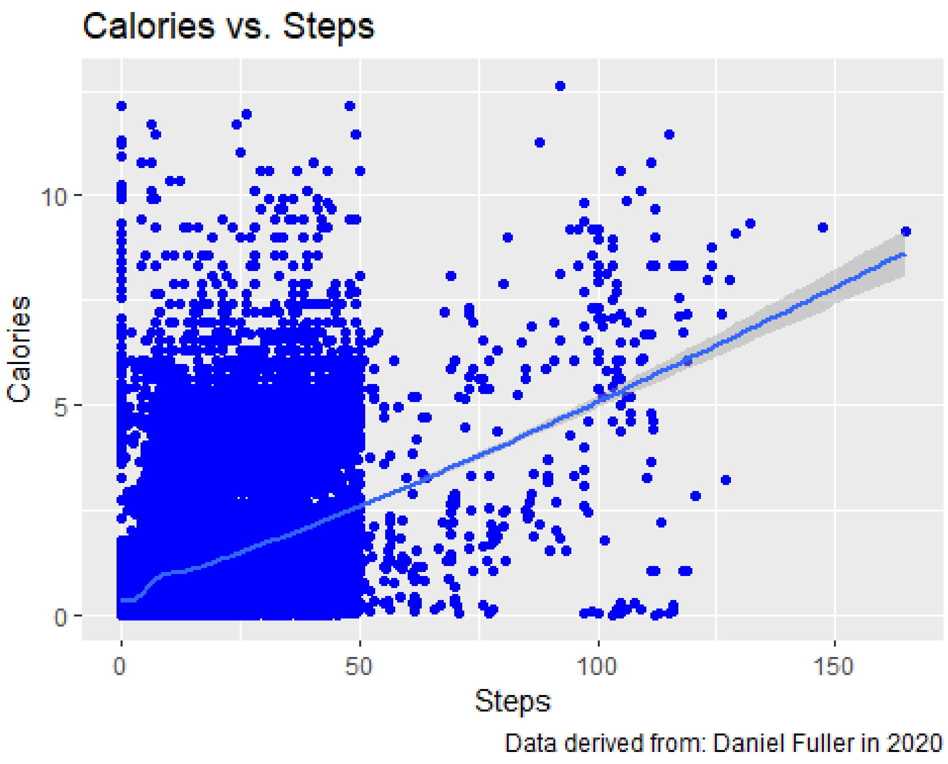
```
## # A tibble: 4 × 7
##   avg_calories activity    ...3        ...4        ...5        ...6
...7
##   <chr>        <chr>       <chr>       <chr>       <chr>       <chr>
```

```
<chr>
## 1 device       light       moderate    sedentary   sleep       vigorous
Gran…
## 2 Apple Watch  0.460574649 0.470258042 0.386217069 0.350629441 0.381184178
0.38…
## 3 FitBit       2.576773677 2.516304557 2.157042452 2.659511558 2.173511922
2.39…
## 4 Grand Total  0.907158617 0.886098691 0.697019685 0.935085079 0.845461847
0.78…
```

In this instance, the “calories” variable-split between devices stays constant, with FitBit having more calories than Apple Watch, and this information was further broken down categorically across each type of the activity variable, “sleep”, “sedentary”, “light”, “moderate”, and “vigorous”. (Note: I changed the variable names to lowercase and removed the number).

One insight to note is that users tend to burn more calories when they are either asleep or sedentary. This is important, because basing any analysis strictly on calorie burn could skew my results. So, to test this, I created another pivot table to test the relationship “calories” has with another variable, “steps”. The “steps” variable was placed as the row and “calories” as the value. I then created a scatter plot with “steps” in the x-axis and “calories” in the y-axis:

```
if(!require(ggplot2)) install.packages("ggplot2", repos = "http://cran.us.r-
project.org")
install.packages("ggplot2")

## Warning: package 'ggplot2' is in use and will not be installed

library(ggplot2)
fitbit_df<- read_csv("norm_fb_aw_data.csv")

## Rows: 59214 Columns: 10
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): gender, activity, device
## dbl (7): heart, calories, steps, distance, age, weight, height
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

fitbit_df %>%
  ggplot()+
  geom_point(mapping=aes(x=steps, y= calories), color="blue") +
  geom_smooth(mapping= aes(x=steps, y= calories))+
  labs(title= "Calories vs. Steps", caption= "Data derived from: Daniel
Fuller in 2020", x="Steps", y="Calories")

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

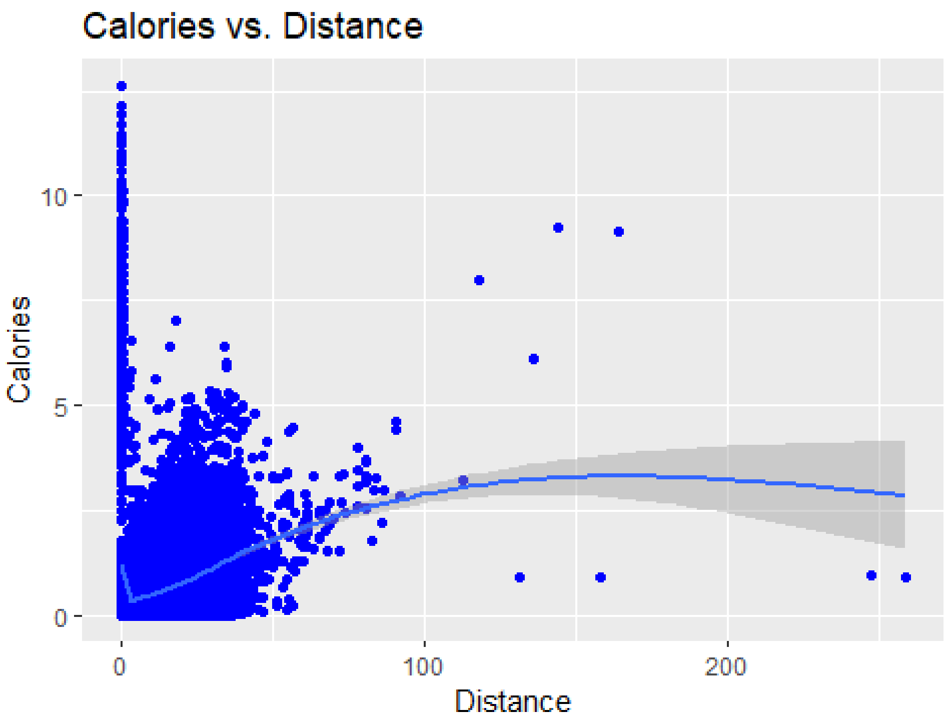

Calories vs. Steps

Data derived from: Daniel Fuller in 2020

As you can see from the trend line there is a positive correlation between the two variables.

I then traded out the “steps” variable for “distance” and made the same type of chart and had a similar trend:

```
fitbit_df %>%
  ggplot()+
  geom_point(mapping=aes(x=distance, y= calories), color="blue") +
  geom_smooth(mapping= aes(x=distance, y= calories))+
  labs(title= "Calories vs. Distance", caption= "Data derived from: Daniel
Fuller in 2020", x="Distance", y="Calories")

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Data derived from: Daniel Fuller in 2020

As you can see there is a lower positive correlation, however,this test showed that there is a relationship between calorie burn, steps, and distance that can now be explored using my other variables to find insights.

Now, I set back up my earlier pivot table, but this time subbed “calories” for its related “steps” variable”. At this point, I found there were inconsistencies in the data when it came to the “steps” count of the “sleep” and “sedentary” variables (Note: the data used for the above visualizations was already normalized based on the following findings). In some cases the “activity” is labeled as “sleep” but the steps indicated are >100. This could mean that the user mislabeled their activity, or that the activity accumulated both sleep and non-sleep activity into its calculation. I compensated for this error by filtering my data in the “sleep” and “sedentary” variables to normalize steps <= 50. I used BigQuery SQL, by running some WHERE clauses and combined my table with the filtered sleep and sedentary results.

Once the data set was normalized to account for the “sedentary” and “sleep” variables, I created a pivot chart with the “device” variable as the row, “activity” on column, and “steps” as the value. This column chart displayed the average steps between Apple Watch and Fitbit users by “activity”:

```
step_activity<- read_csv("step_activity.csv")

## New names:
## Rows: 4 Columns: 7
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
```

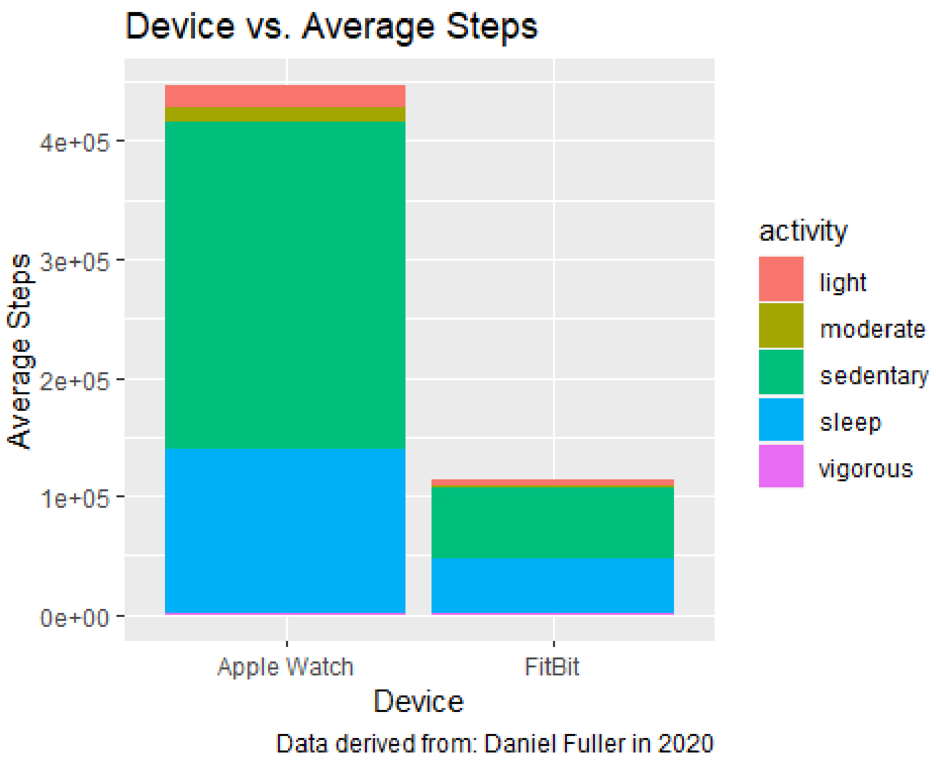
```
chr
## (7): avg_steps, activity, ...3, ...4, ...5, ...6, ...7
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `` -> `...3`
## • `` -> `...4`
## • `` -> `...5`
## • `` -> `...6`
## • `` -> `...7`
```

```
head(step_activity)
```

```
## # A tibble: 4 × 7
##   avg_steps   activity    ...3        ...4        ...5        ...6
...7
##   <chr>       <chr>       <chr>       <chr>       <chr>       <chr>
<chr>
## 1 device      light       moderate    sedentary   sleep       vigorous
Grand…
## 2 Apple Watch 11.10357472 10.94506912 8.519285277 7.775154751 10.45803802
8.461…
## 3 FitBit      29.91519598 23.17233006 11.99890521 12.91349566 6.617717136
13.41…
## 4 Grand Total 15.07341326 13.43015051 9.130003299 9.075841476 9.463256101
9.471…
```

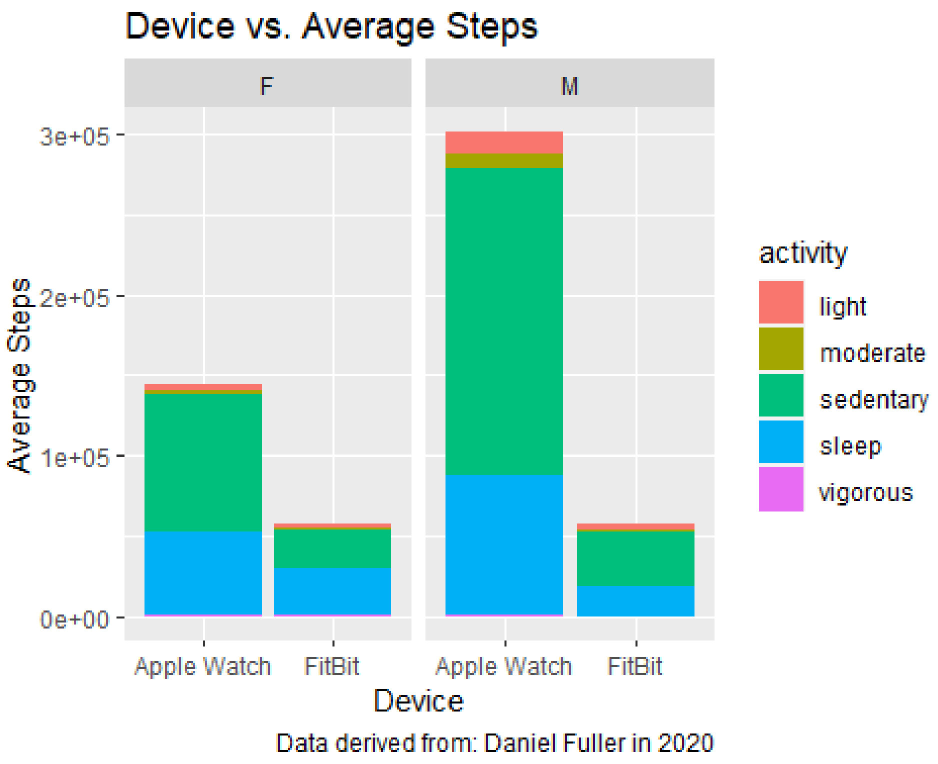
Note: I used the fitbit_df to create this plot, however, I used the mean of the “steps” variable for consistency.

```
fitbit_df %>%
  ggplot() +
  geom_col(mapping=aes(x=device, y=mean(steps), fill=activity))+
  labs(title= "Device vs. Average Steps", caption= "Data derived from: Daniel
Fuller in 2020", x="Device", y="Average Steps")
```



This plot provided stark contrast to how many users prefer the Apple Watch to the FitBit, whether that be due to brand recognition, or the user's preference for a general use smart device. Also, note the largest segments of "activity" are sedentary and sleep.

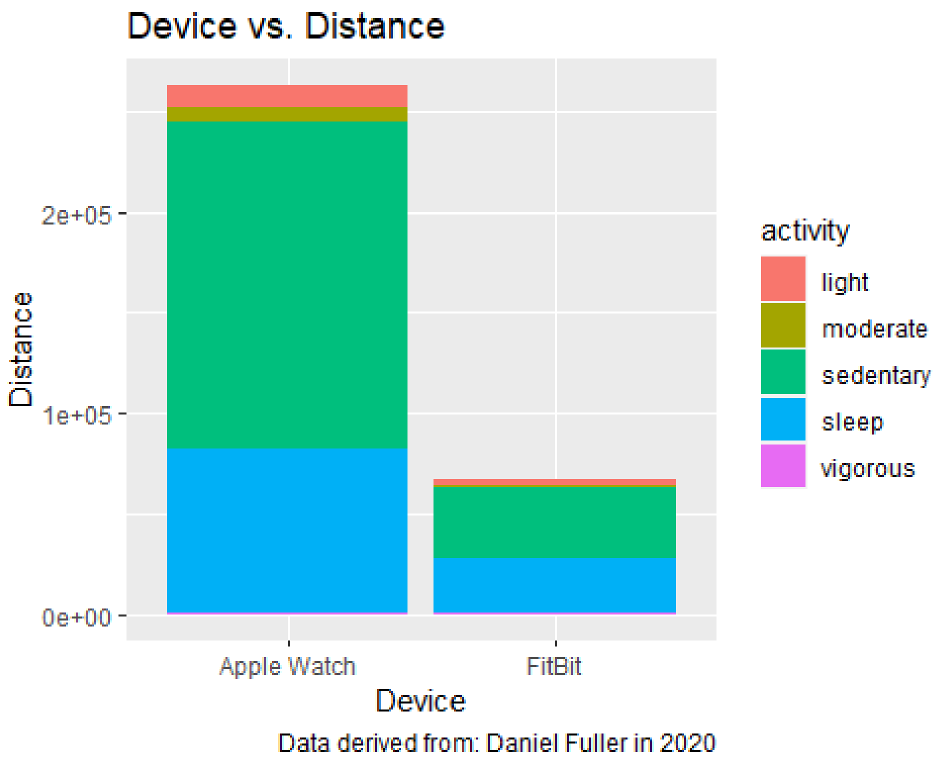
```r
fitbit_df %>%
  ggplot() +
  geom_col(mapping=aes(x=device, y=mean(steps), fill=activity))+
  facet_wrap(~gender)+
  labs(title= "Device vs. Average Steps", caption= "Data derived from: Daniel
Fuller in 2020", x="Device", y="Average Steps")
```



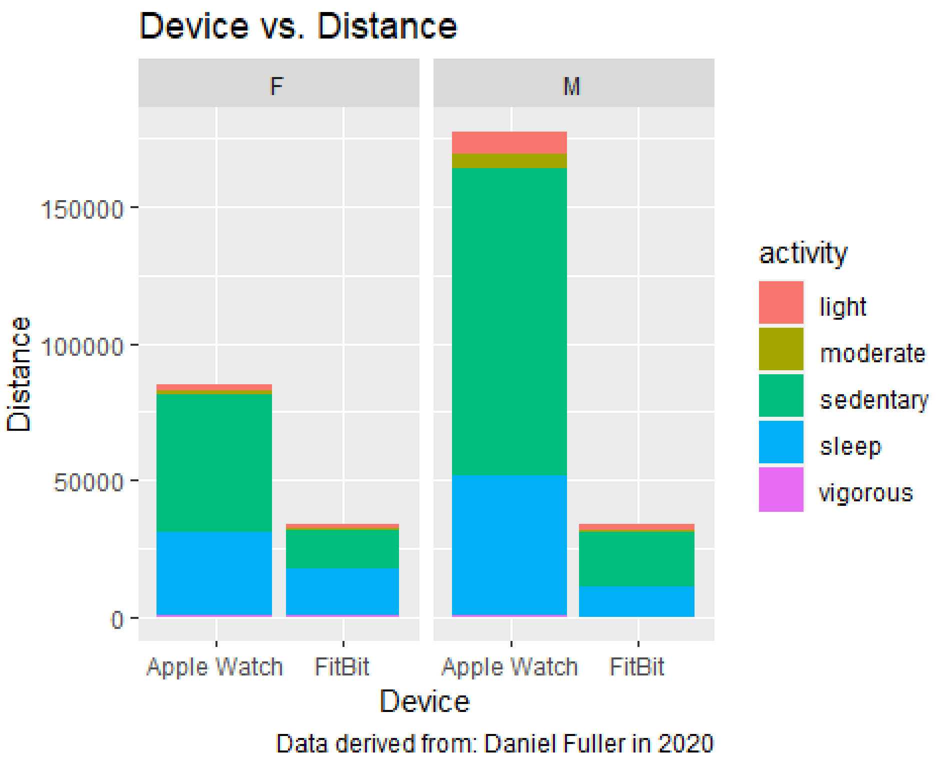
When split by gender, this plot suggests female users may not be as beholden to general use nor brand recognition. Although, Apple Watch is still the leading smart device, the degree by which female users choose it over the FitBit is less pronounced. Why this is significant is because there are more female users in the sample size 26 out of 46.

Next, I changed the steps value to distance, which surprisingly showed that Apple Watch users travel by average more distance with their devices by category. This is some more evidence behind the idea that general use smart device, without discounting brand recognition, is a key metric in the smart device industry.

```
fitbit_df %>%
  drop_na() %>%
  ggplot() +
  geom_col(mapping=aes(x=device, y=mean(distance), fill=activity))+
  labs(title= "Device vs. Distance", caption= "Data derived from: Daniel
Fuller in 2020", x="Device", y="Distance")
```



```
fitbit_df %>%
  drop_na() %>%
  ggplot() +
  geom_col(mapping=aes(x=device, y=mean(distance), fill=activity))+
  labs(title= "Device vs. Distance", caption= "Data derived from: Daniel
Fuller in 2020", x="Device", y="Distance")+
  facet_wrap(~gender)
```



Broken out by gender, there is a similar evidence to show the female demographic may not be as swayed by general use or brand recognition when it comes to smart device use. Lastly, it is important to note, again, that sedentary and sleep activity account for the stark majority across both devices and genders.

Share:

Guiding questions:

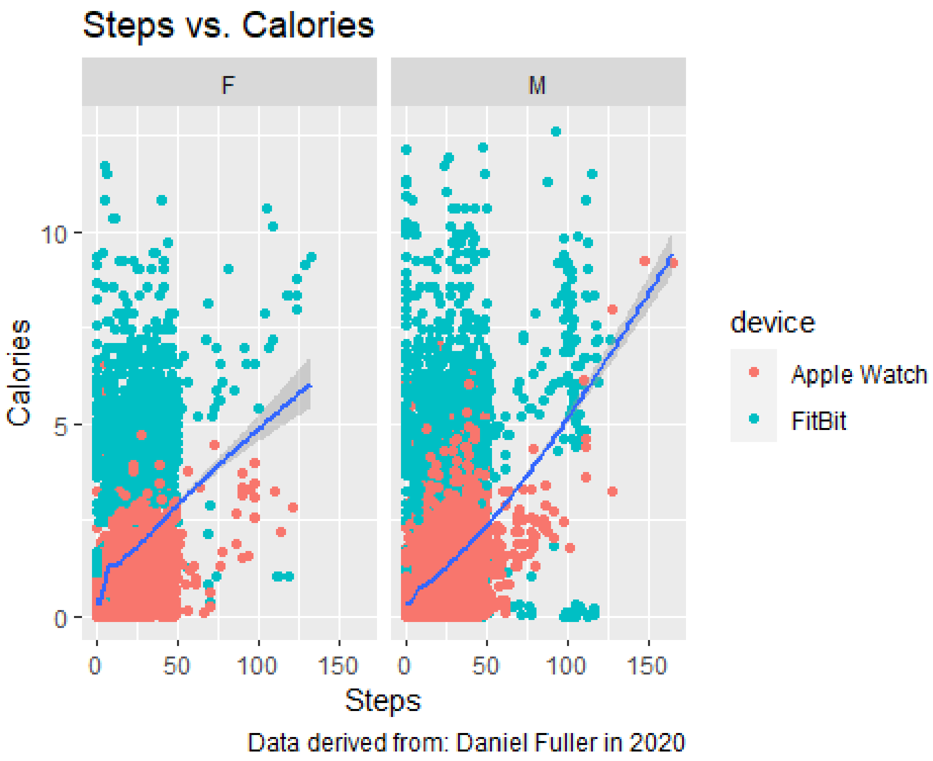
- Were you able to answer the business questions?
- What story does your data tell?
- How do your findings relate to your original question?
- Who is your audience? What is the best way to communicate with them?
- Can data visualization help you share your findings?
- Is your presentation accessible to your audience?

Plotting in R

(Note, although all of the plots above are written in R, a lot of the earlier analysis was done using the pivot tables and charts within spreadsheets. The following plots are the ones that lent further credit my earlier analysis.)

Once I loaded my normalized data in R (reducing the sedentary and sleep step count <=50), I renamed my data frame to fitbit_df before setting up the following scatter plot:

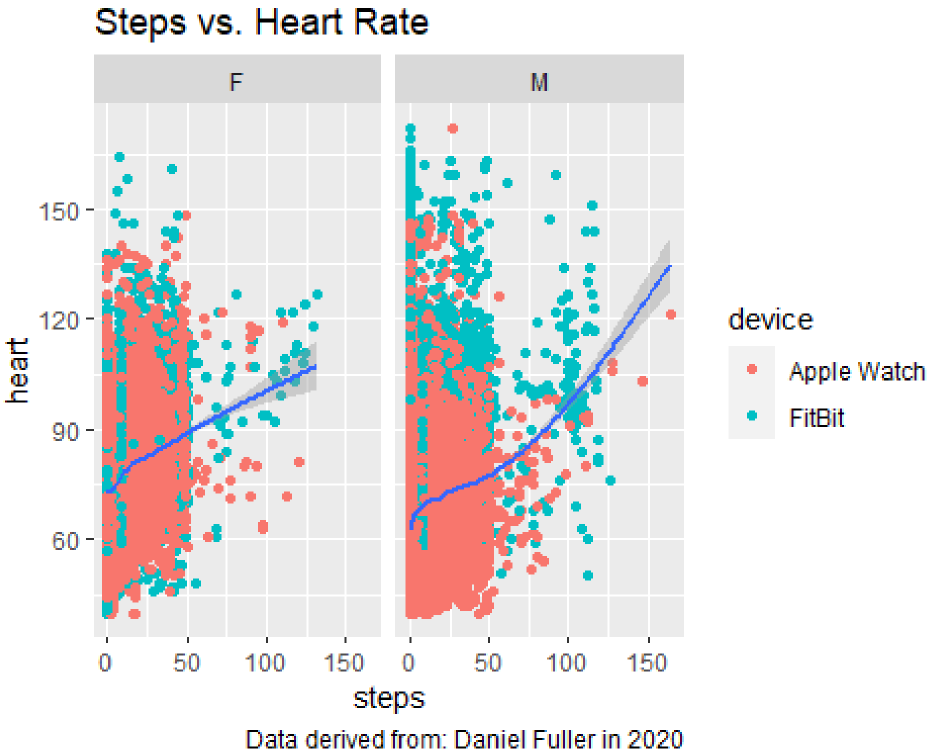
```
fitbit_df %>%
  ggplot() +
  geom_point(mapping=aes(x=steps, y= calories, color=device)) +
  facet_grid(~gender) +
  geom_smooth(mapping=aes(x=steps, y= calories))+
  labs(title= "Steps vs. Calories", caption= "Data derived from: Daniel
Fuller in 2020", x="Steps", y="Calories")

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



This plot has the “steps” variable on the x-axis and the “calories” variable on the y-axis. Moreover, it uses color to distinguish the “device” variable and splits the graph by “gender”. In this graph there is a strong correlation between the amount of steps taken and calorie burn. FitBit users tend to have a higher rate of calorie burn in relation to step count than Apple Watch users. Notably, female users seem to favor the FitBit for their fitness needs. This could have some relation to form factor, but most likely is due to FitBit marketing to the fitness industry.

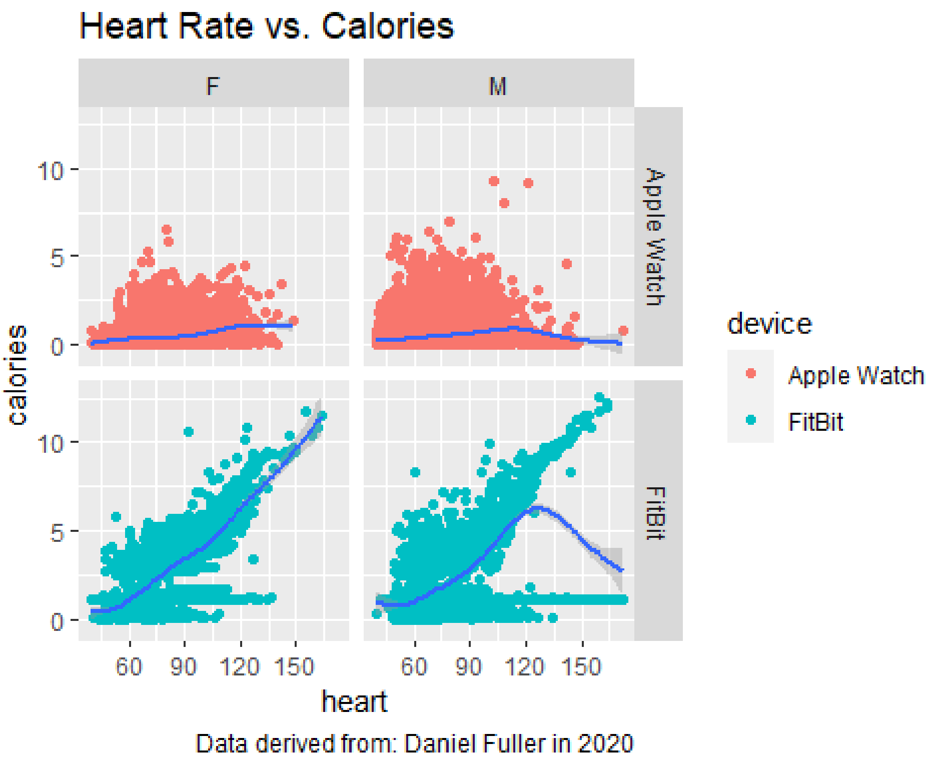
In this next plot I subbed the “steps” variable for “heart” or heart rate, which provides some evidence that FitBit users may just be more active than Apple Watch users:

```
fitbit_df %>%
  ggplot() +
  geom_point(mapping=aes(x=steps, y=heart, color=device)) +
  facet_grid(~gender)+
  geom_smooth(mapping=aes(x=steps, y= heart))+
  labs(title= "Steps vs. Heart Rate", caption= "Data derived from: Daniel
Fuller in 2020", x="steps", y="heart")

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



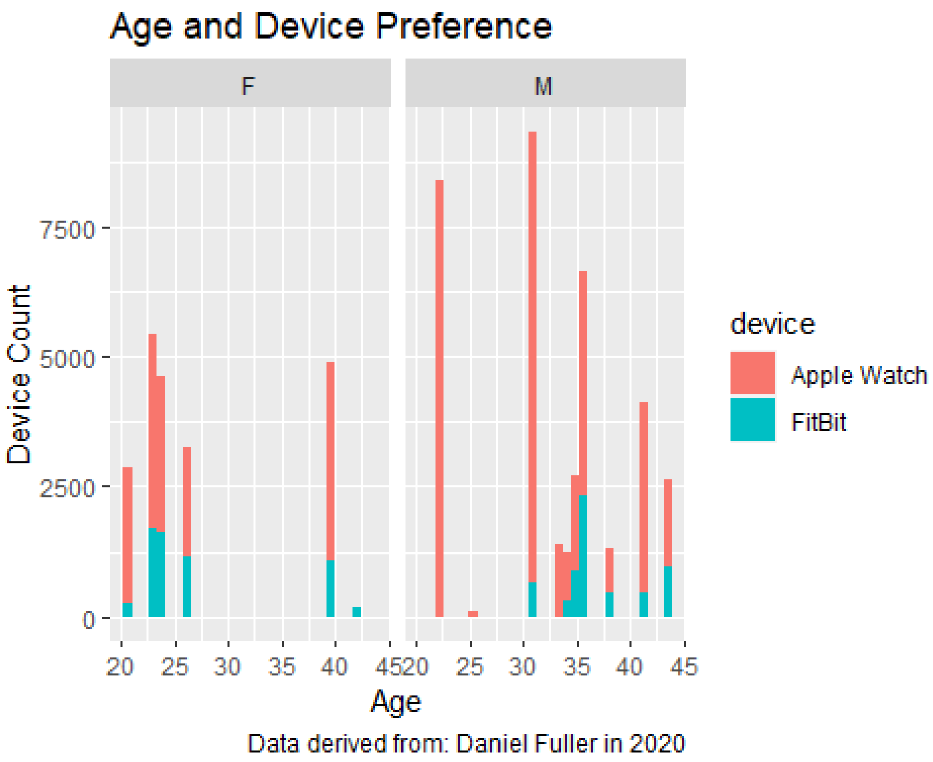
Female FitBit users tend to be more active, which could be a key indicator to market towards fitness.

```
fitbit_df %>%
  ggplot() +
  geom_point(mapping=aes(x=heart, y=calories, color=device)) +
    geom_smooth(mapping=aes(x=heart, y= calories))+
     facet_grid(device~gender)+
  labs(title= "Heart Rate vs. Calories", caption= "Data derived from: Daniel
Fuller in 2020", x="heart", y="calories")

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



This next graph places age on the x-axis, depicting user habits by both the age and gender demographics. Both genders seem to prefer the Apple Watch over the FitBit, which could infer that they prefer a more generalized device for personal use. In male users the preference for Apple watch is more profound in those < 35 and in female users the same holds for < 40. More data would need to be collected to make a true determination of this factor.

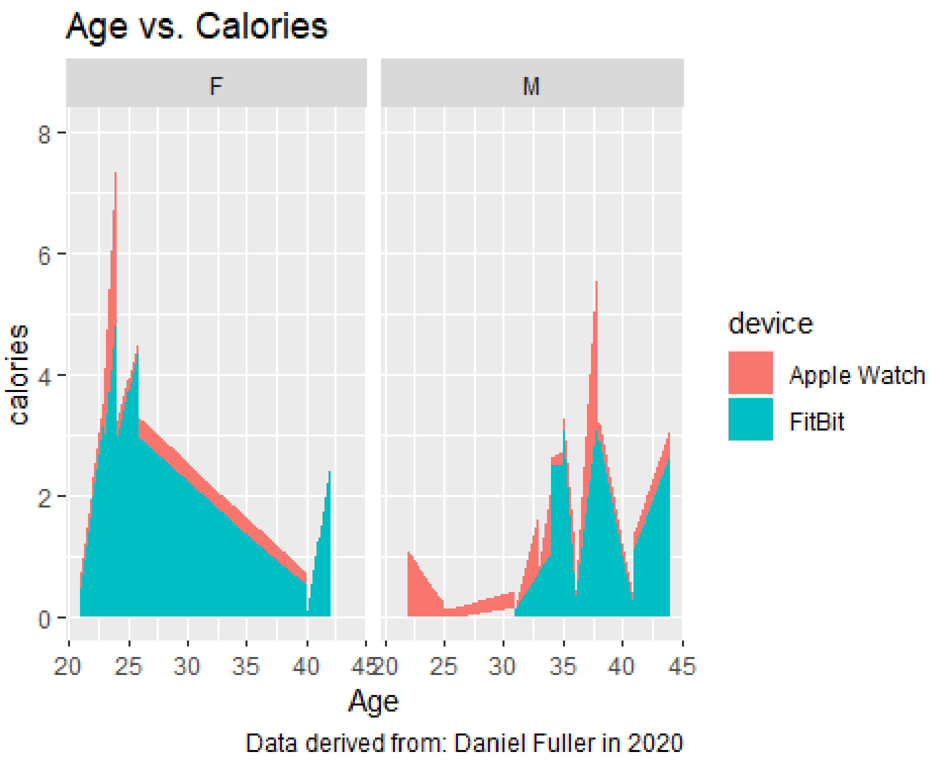
```
fitbit_df %>%
  ggplot() +
  geom_histogram(mapping=aes(x=age, fill= device))+
  facet_wrap(~gender)+
  labs(title= "Age and Device Preference", caption= "Data derived from:
Daniel Fuller in 2020", x="Age", y="Device Count")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Age and Device Preference

Data derived from: Daniel Fuller in 2020

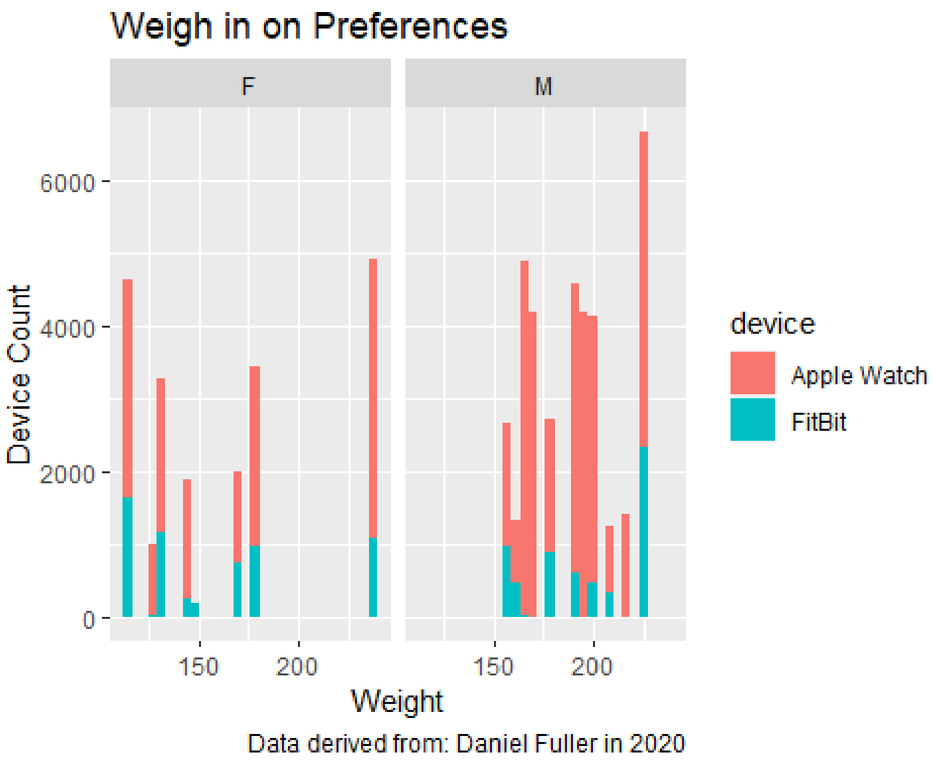
When you look at calorie burn based on age demographics there is a stark diversion in male and female user behavior. Males >= 35 tend to burn more calories, where women < 30 tend to burn more calories. If you plan to market to female users with a fitness focus, the <30 demographic is your market segment.

```
fitbit_df %>%
  ggplot() +
  geom_area(mapping=aes(x=age, y=calories, fill= device))+
  facet_wrap(~gender)+
  labs(title= "Age vs. Calories", caption= "Data derived from: Daniel Fuller
in 2020", x="Age", y="calories")
```



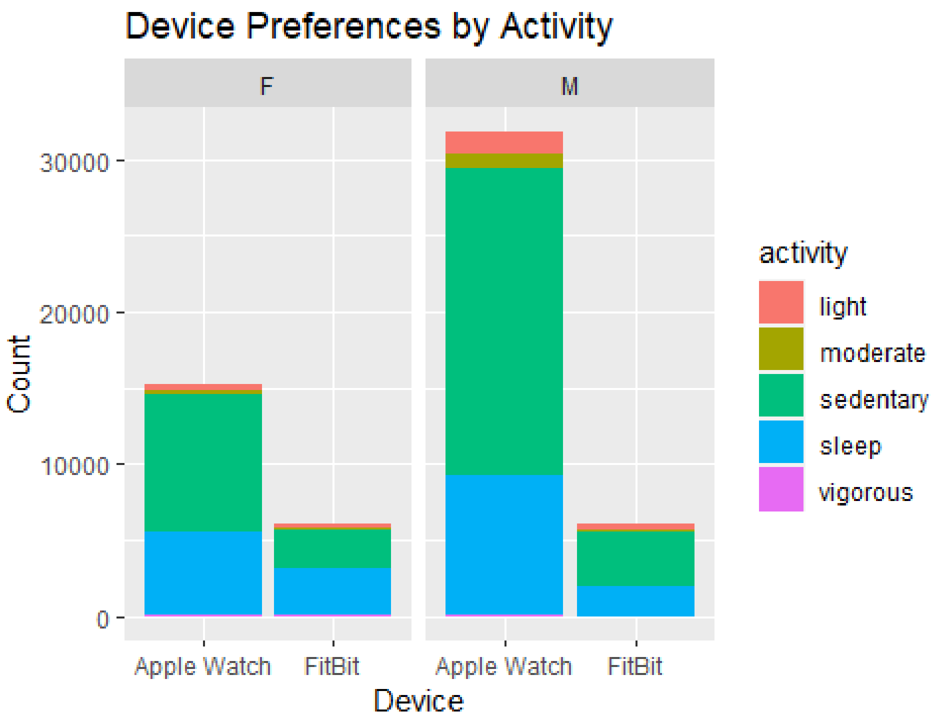
This next graph breaks user preferences by the “weight” variable, with FitBit users generally weighing less than Apple Watch users:

```
fitbit_df %>%
  ggplot() +
  geom_histogram(mapping=aes(x=weight, fill= device))+
  facet_wrap(~gender)+
  labs(title= "Weigh in on Preferences", caption= "Data derived from: Daniel
Fuller in 2020", x="Weight", y="Device Count")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This final graph places device on the x-axis and splits the graphs by gender and device. Additionally, it fills the bars based on activity. This graph shows that men are more likely to prefer a general-purpose device than women. Which is significant, since this sample size is skewed toward women with 26 out of 46 users are female.

```
fitbit_df %>%
  ggplot() +
  geom_bar(mapping=aes(x=device, fill= activity))+
  facet_wrap(~gender)+
  labs(title= "Device Preferences by Activity", caption= "Data derived from:
Daniel Fuller in 2020", x="Device", y="Count")
```



Data derived from: Daniel Fuller in 2020

Act:

Guiding questions:

- What is your final conclusion based on your analysis?
- How could your team and business apply your insights?
- What next steps would you or your stakeholders take based on your findings?
- Is there additional data you could use to expand on your findings?

Conclusion and Recommendations:

The purpose of this case study was to provide Bellabeat stakeholders insights into user preferences for industry-wide smart devices, to inform the decision-making of the marketing analytics team. The conclusion and recommendations of this study are as follows:

- Users prefer the Apple Watch over the FitBit, based on their trend towards generalized use.

- Apple Watch uses accounted for 77% of activity in the data, and that activity is fairly equal between categories.

- FitBit users tended to use their device for fitness purposes, with the females aged 30 and under being a key demographic.

- Female users perform more steps by average than males users across both devices. Suggesting a tendency towards fitness.

- There is strong evidence to suggest marketing the Bellabeat Time as a general use smart device would cater to the widest demographic.

- However, if the fitness market is the goal, tailor your marketing strategy to female users 30 and under.

Additional research

- Additional research should be conducted with a larger sample size to account for age exclusivity for female users.

- Other metrics and methods should be included in future study, like sales price, revenue, or a sentiment analysis

- Additional constraints include the currency of the data used. More recent data should be collected, and a longer period should be studied. One metric my data lacked was time and date entries. This type of data would be invaluable and establish more credibility.

- Lastly, future study should include more quantitative and statistical analysis, such as regression analysis.