Google Data Analytics Capstone Project: How Can a Wellness Technology Company Play It Smart?

The following documentation follows the optional Capstone project provided by the Google Data Analytics Course. It follows through the eight stages of data analysis which are Ask, Prepare, Process, Analyze, Share and Act.

This Capstone Project was carried out with the help of R programming language, which is a data-centric, accessible language used to organize, modify, clean data frames and create insightful data visualizations. Let's get into it!

Phase I: Ask

This step involves defining the problem, asking effective questions, and confirming stakeholders' expectations and desired outcomes.

Bellabeat is a high-tech manufacturer of beautifully designed health-focused smart products for women since 2013. Inspiring and empowering women with knowledge about their health and habits, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for females.

The co-founder and Chief Creative Officer, Urška Sršen is confident that an analysis of non-Bellabeat consumer data (i.e., Fitbit fitness tracker usage data) would reveal more growth opportunities.

The business objectives are as follows:

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

The stakeholders are:

1. **Urška Sršen**: Bellabeat's co-founder and Chief Creative Officer
2. **Sando Mur**: Mathematician, Bellabeat's co-founder, and a key member of the Bellabeat executive team
3. **Bellabeat marketing analytics team**: A team of data analysts guiding Bellabeat's marketing strategy.

Phase II: Prepare

This stage involves documenting how and where the data is generated and collected, identifying the use of different types of formats, types, and structures, and making sure the data is unbiased and credible.

The data is publicly available on Kaggle: Fitbit Fitness Tracker Data and is stored in over 18 CSV files. The data consists of health data generated by 30 respondents related to their activity, sleep, stress, menstrual cycles, and mindfulness habits. The data is 3rd party as it was collected by Amazon Mechanical Turks and presented to Bellabeat.

The time frame of the survey was 12 March 2016 to 12 May 2016.

The data however face some key limitations:

1. Data is collected in 2016. Users' daily activity, fitness and sleeping habits, diet, and food consumption may have changed since then. Data may not be timely or relevant.
2. The sample size is low, and the findings may not be an accurate depiction of the entire population.
3. Since the data was collected through a survey, I am unable to ascertain its integrity/accuracy i.e. A user may have been embarrassed to release statistics that may suggest bad habits such as laziness.

The data can therefore be referred to as bad data as per the ROCCC metrics:

1. **Reliable (low) –** The sample size of 30 is not enough to support reliability since representation is not efficiently exhibited.
2. **Original (low) –** The data was collected by a third party hence authenticity is low.
3. **Comprehensive (medium) –** The data seems to match the metrics that Bella Beat's program collects from its consumer thus critical information to answer the problem are captured.
4. **Current (low) –** The data collected is outdated and thus may not be relevant.
5. **Cited (low) –** The data is collected by a third party hence the reliability of its citation might be low.

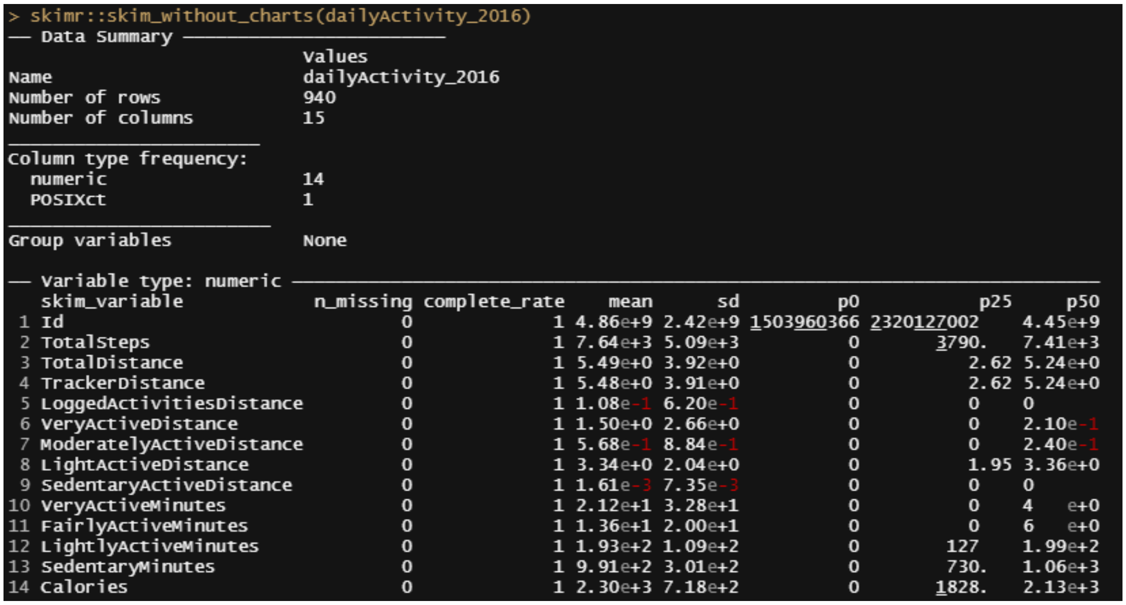
Even though the dataset is bad data and would not be recommended to make key business decisions, it will be used for this case study for learning purposes.

The following file that was selected for analysis was *DailyActivity_merged* – A summary of users' daily usage statistics i.e., Total Steps, Total distance covered, etc.

PHASE III: Process

This phase involves exploring the data, transforming the data into a usable format, cleaning the data, performing a preliminary statistical overview, and verifying and reporting on cleaning the data. The dataset's naming convention wasn't up to standards therefore the name was updated to *dailyActivity_2016.*

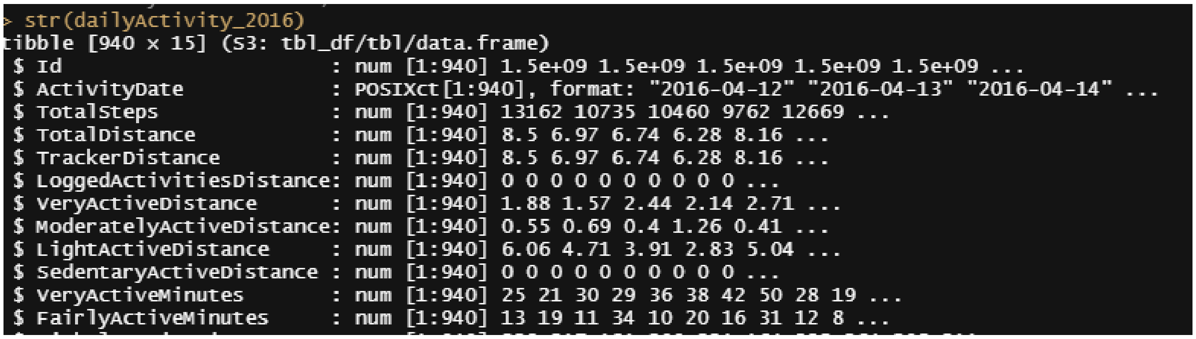
First, we check to see whether the data has missing values:



```
> skimr::skim_without_charts(dailyActivity_2016)
── Data Summary ────────────────────────
                           Values
Name                       dailyActivity_2016
Number of rows             940
Number of columns          15
_______________________
Column type frequency:
  numeric                  14
  POSIXct                  1
________________________
Group variables            None

── Variable type: numeric ──────────────────────────────────────────────────────────────────────
   skim_variable             n_missing complete_rate    mean      sd         p0        p25     p50
 1 Id                                0             1 4.86e+9 2.42e+9 1503960366 2320127002 4.45e+9
 2 TotalSteps                        0             1 7.64e+3 5.09e+3          0      3790. 7.41e+3
 3 TotalDistance                     0             1 5.49e+0 3.92e+0          0       2.62 5.24e+0
 4 TrackerDistance                   0             1 5.48e+0 3.91e+0          0       2.62 5.24e+0
 5 LoggedActivitiesDistance          0             1 1.08e-1 6.20e-1          0       0    0
 6 VeryActiveDistance                0             1 1.50e+0 2.66e+0          0       0    2.10e-1
 7 ModeratelyActiveDistance          0             1 5.68e-1 8.84e-1          0       0    2.40e-1
 8 LightActiveDistance               0             1 3.34e+0 2.04e+0          0       1.95 3.36e+0
 9 SedentaryActiveDistance           0             1 1.61e-3 7.35e-3          0       0    0
10 VeryActiveMinutes                 0             1 2.12e+1 3.28e+1          0       0    4   e+0
11 FairlyActiveMinutes               0             1 1.36e+1 2.00e+1          0       0    6   e+0
12 LightlyActiveMinutes              0             1 1.93e+2 1.09e+2          0     127    1.99e+2
13 SedentaryMinutes                  0             1 9.91e+2 3.01e+2          0     730.   1.06e+3
14 Calories                          0             1 2.30e+3 7.18e+2          0    1828.   2.13e+3
```

According to the summary done below, the *dailyActivity_2016* dataset contains 940 rows, and 15 columns contain no missing values. Let's check for the data types of the columns, shall we? There seems to be a problem.Columns*TotalDistance*,*TrackerDistance*,*(Very/Moderately/Light)ActiveDistance,* are supposed to be of type double rather than num.



```
> str(dailyActivity_2016)
tibble [940 x 15] (S3: tbl_df/tbl/data.frame)
 $ Id                      : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
 $ ActivityDate            : POSIXct[1:940], format: "2016-04-12" "2016-04-13" "2016-04-14" ...
 $ TotalSteps              : num [1:940] 13162 10735 10460 9762 12669 ...
 $ TotalDistance           : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
 $ TrackerDistance         : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
 $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveDistance      : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
 $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
 $ LightActiveDistance     : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
 $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
 $ VeryActiveMinutes       : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
 $ FairlyActiveMinutes     : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
```

However, in this case, this is not a problem in R because the data type ‘numeric’ represents both ‘double’ and/or ‘integer’. (double function - RDocumentation). This is a good observation to make as other programming languages might not share the same sentiments as R.

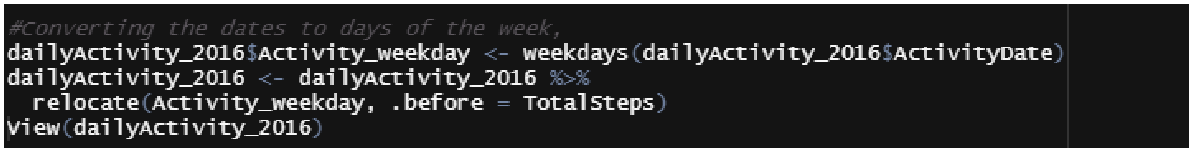
The data set is at this point considered to be clean. On to the next phase, analyze and share!

PHASE IV and PHASE V: Analyze and Share

These stages involve the transformation of data, identifying patterns, drawing inferences, and presenting visualizations.

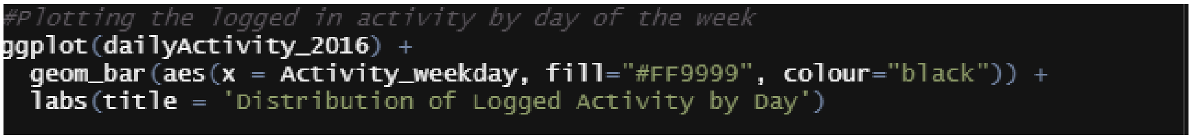
According to this data set, the data team viewed this problem from the perspective of the consumer asking questions such as: “What are users doing with this application and when are they most active on it”, “Are users very active on this application or use it in moderation”. Questions like these helped in trying to find the answers that the data team was looking for.

The column *ActivityDate* to reference days of the weeks so that the team would find out on which days is the application being used the most.



```
#Converting the dates to days of the week,
dailyActivity_2016$Activity_weekday <- weekdays(dailyActivity_2016$ActivityDate)
dailyActivity_2016 <- dailyActivity_2016 %>%
  relocate(Activity_weekday, .before = TotalSteps)
View(dailyActivity_2016)
```

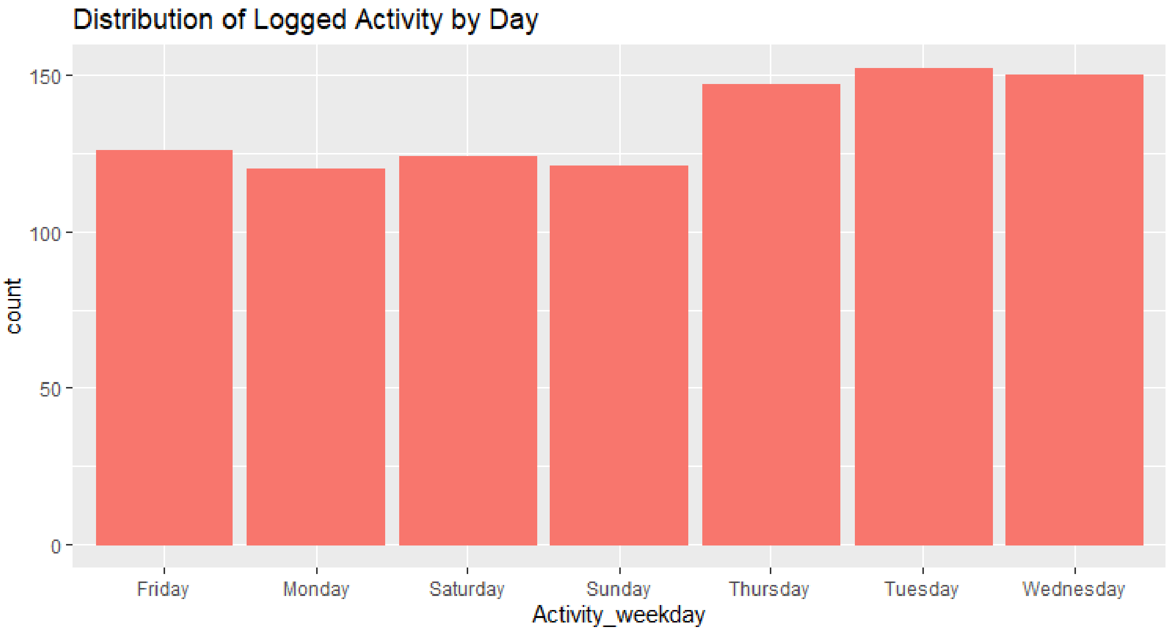
The new column *Activity_weekday* was plotted against the total number of logged-in ‘actions’ (any scenario where the user logged into the application even if it was to check the application’s settings).



```
#Plotting the logged in activity by day of the week
ggplot(dailyActivity_2016) +
  geom_bar(aes(x = Activity_weekday, fill="#FF9999", colour="black")) +
  labs(title = 'Distribution of Logged Activity by Day')
```

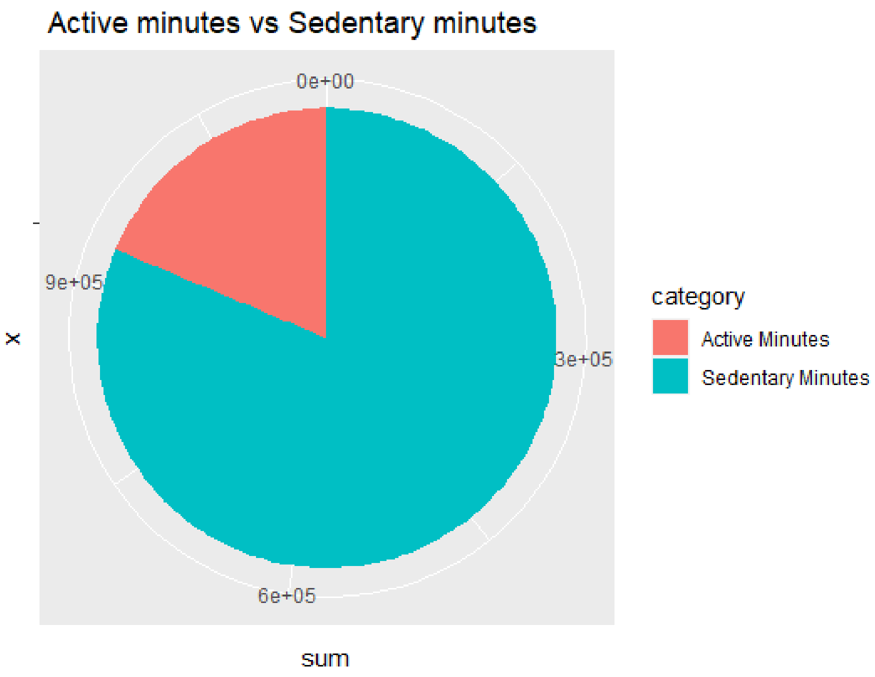
The conclusions made from the graph:

1. Users registered the highest logged-in usage statistics on Tuesday, Wednesday, and Thursday.
2. Users tend to use the application more on weekdays than on weekends.
3. Monday recorded the lowest number of log-ins throughout the week



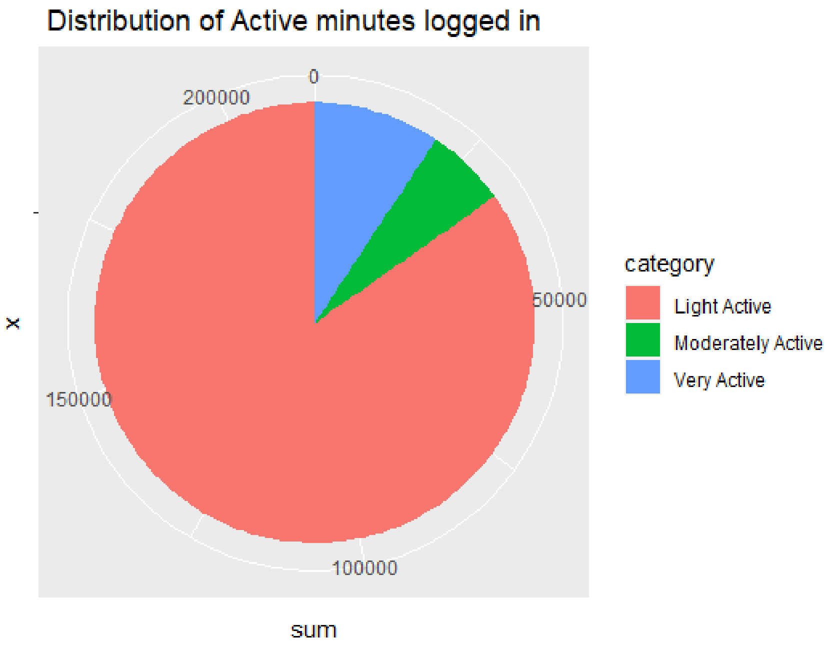
Before going further in, some assumptions about the usage could be made. According to (Lesser, 2021), research was carried out among individuals to prove whether the term ‘Monday Blues’ is a real thing. It was confirmed that individuals tend to exhibit low levels of motivation to do their normal routines on Monday. This could account for the low usage statistics on Monday.

This led to the question, are all of the minutes being tracked active or sedentary minutes? The following pie chart was made to answer this question.



The conclusion drawn from this pie chart was that Sedentary minutes took the majority of the pie share meaning that users were using the application to track stationary activity most of the time.

This was further supported by the pie chart below that segments the activity category based on how many minutes were logged into the application.



The pie chart confirms that the majority of the users logged in for light activity which could include Sedentary activity to some degree.

PHASE VI: ACT

The findings/trends identified were:

1. Weekdays (Tuesday through Thursday) registered the highest number of activity through.
2. Monday registered the lowest number of Logged in activity.
3. Most of the users logged into the app were sedentary users therefore the application was not used as for what it was intended.

The following were the recommendations made by the data analysis team.

1. According to (What are streaks on SnapChat?, n.d.), incorporating a ‘streak’ based model, where users can choose to share the activity with friends and family, may increase interest where users try their best to register positive activity statistics so that they can share progress.
2. On days where usage is low (Weekends), a notification motivation encouraging users to exercise is heavily advised.
3. The application could integrate an education module where consumers would be advised on the advantages of exercise and its health benefits in a bid to encourage fitness routines.

Thank You!