# Google Data Analytics Course

## Product Analysis Case Study

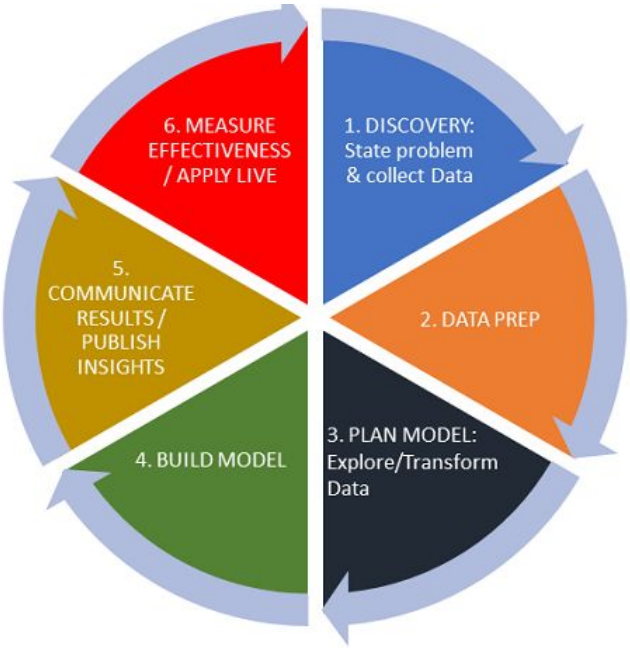
Cameron Kemske
12/9/2021

# Scenario

- Newly hired Business Analyst for *Bellabeat Inc.*
- Tasked with conducting a data analysis report for the CMO, Urska Srson and various members of the executive team.
- Problem
  - Urska is looking for new ways to improve the user interactions on Bellabeat smartphone application.
  - This application is similar to Apple Fitness and Fitbit products.
- Solution
  - Analyze data and create a report to showcase findings to improve marketing strategies.

# Deliverables

1. Clear Summary of Business Task **(Ask)**
2. Description of Data Source **(Prepare)**
3. Documentation of Cleaning **(Process)**
4. Summary of Analysis **(Analysis)**
5. Supporting Visualizations **(Share)**
6. High-Level Recommendations **(Act)**



# Step 1: “Ask” Phase - Key Questions

1. What are some trends in smart device usage?
    a. These device are used for health tracking, recording exercise, and constant connection to product’s parent company.
2. How could these trends apply to Bellabeat customers?
    a. Trends established from other smart devices can be used as proxies for gathering information for our own customers.
3. How could these trends help influence marketing strategies?
    a. Data collected comes directly from smart device users. This data can help with market research which leads to implementation of new strategies.

# Step 1 : “Ask” Phase – Business Task

- Analyze smart device usage from non-Bellabeat users to find insights:
  - Recording/logging trends
  - Activity trends
  - User participation trends

- **Based on the analysis, recommendations will be constructed to improve Bellabeat’s smart device application through new marketing strategies.**

# Step 2: “Prepare” Phase - Key Questions

- Where was the data located?
  - The Fitbit device datasets were located on Kaggle.com. Kaggle.com is an internet-based data community that offers free datasets for analysis.
    - daily_activities.csv
    - daily_intensities.csv
    - daily_calories.csv
    - daily_sleep.csv
    - weight_log_info.csv
- How do you determine credibility?
  - Kaggle.com is “highly trusted within the data science industry with Facebook, Walmart and Winton Capital using it for their own operations” *(Quora.com)*.
  - No reason to assume the data provided is unreliable, inaccurate, or compromised.

# Step 2: “Prepare” Phase - Key Questions cont.

- Do the datasets have integrity?
  - All datasets were uploaded in R Studio, using the view() function to check rows and columns for null values.
  - All data sets have a primary key that links them together **(“Id” variable).**
- Any problems with the data?
  - Zero values have been identified as “missed logs” from the user instead of the smart device error **(missing data prevents additional insights).**
  - Data has a time frame from mid-April to mid-May **(short time-frame to analyze).**
- How is the data organized?
  - The data is in wide format, and each variable has a value for that observation.

# Step 3 : “Process” Phase - Tools

- Tools
  - R Studio: cleaning, analyzing, and visualizing the datasets
  - SQL syntax package in R: assists in streamlining the cleaning process
  - Google Slides: constructing data report for stakeholders

- Data Cleaning Log
  - Please refer to *Appendix* for the R Markdown Cleaning Log.
    - This log reflects the changes made by the analyst in the cleaning process, from the SQL checks to the addition of a new variable in the dataset to be covered later.

# Step 3: “Process” Phase - SQL Example

Example Code Chunk #1

```
daily_activities2 <- daily_activities %>%
  select(Id, ActivityDate, Calories)
```

- The creation of the data frame with the specific variables to be cross referenced in the SQL check.

Example Code Chunk #2

```
sql_check1 <- sqldf('SELECT *
                     FROM daily_activities2
                     INTERSECT
                     SELECT *
                     FROM daily_calories')
count(sql_check1)
```

- The code is performing the cross reference to check similarities of observation between the data sets.

# Step 3: “Process” Phase – R Examples

- Additional functions used:
  - head() - to check the first six variables and to find out variable names
  - count() - to count the total number of observations in each data set
  - weekdays() - to convert the “date” data into the “day of the week” variable
  - mutate() - to add the additional “ day of the week” variable to the data set

head() function

| Id | ActivityDate |
|---|---|
| 1503960366 | 4/12/2016 |
| 1503960366 | 4/13/2016 |
| 1503960366 | 4/14/2016 |
| 1503960366 | 4/15/2016 |
| 1503960366 | 4/16/2016 |
| 1503960366 | 4/17/2016 |

count() function

| | n |
|---|---|
| 1 | 940 |

weekdays() function

| ActivityDate | dayofweek |
|---|---|
| 4/12/2016 | Tue |
| 4/13/2016 | Wed |
| 4/14/2016 | Thu |

mutate() function

| dayofweek | mean_by_day |
|---|---|
| Tue | 2356.013 |
| Wed | 2302.620 |
| Thu | 2199.571 |

# Step 4: “Analysis” Phase - Summary

- User Participation Analysis
  - Distinct users were decreasing in recording activity through the three remaining data sets.
  - Each different activity log in the Fitbit app is decreasing in usage by users.

| | Distinct Users | Decrease in Participation |
|---|---|---|
| Daily_activites | 33 | - |
| Daily_sleep | 24 | 27% |
| Daily_weight_log | 8 | 67% |

- Statistical Analysis
  - Used summary() to provide general statistics on the daily_activity data set.
  - Median and Mean values for “VeryActiveMinutes” vary by a substantial amount.
  - Potentially, several highly active users are raising the average.

| | TotalSteps | TotalDistance | SedentaryMinutes | VeryActiveMinutes |
|---|---|---|---|---|
| Minimum | 0 | 0 | 0 | 0 |
| 1st Quartile | 3790 | 2.26 | 729.8 | 0 |
| Median | 7406 | 5.245 | 1057.5 | 4 |
| Mean | 7638 | 5.49 | 991.2 | 32 |
| 3rd Quartile | 10727 | 7.713 | 1229.5 | 32 |
| Maximum | 36019 | 28.03 | 1440 | 210 |

# Step 4: “Analysis” Phase - Summary cont.

## Daily Activity Analysis #1

- Averages of the calories burned per day allows you to determine the amount of activity occurring on each day.
- Tuesday: highest calories burned @ 2356
- Thursday: lowest calories burned @ 2199

| Day | Mean Calories Burned |
|---|---|
| Mon | 2324 |
| Tues | 2356 |
| Wed | 2302 |
| Thu | 2199 |
| Fri | 2331 |
| Sat | 2354 |
| Sun | 2263 |

## Daily Activity Analysis #2

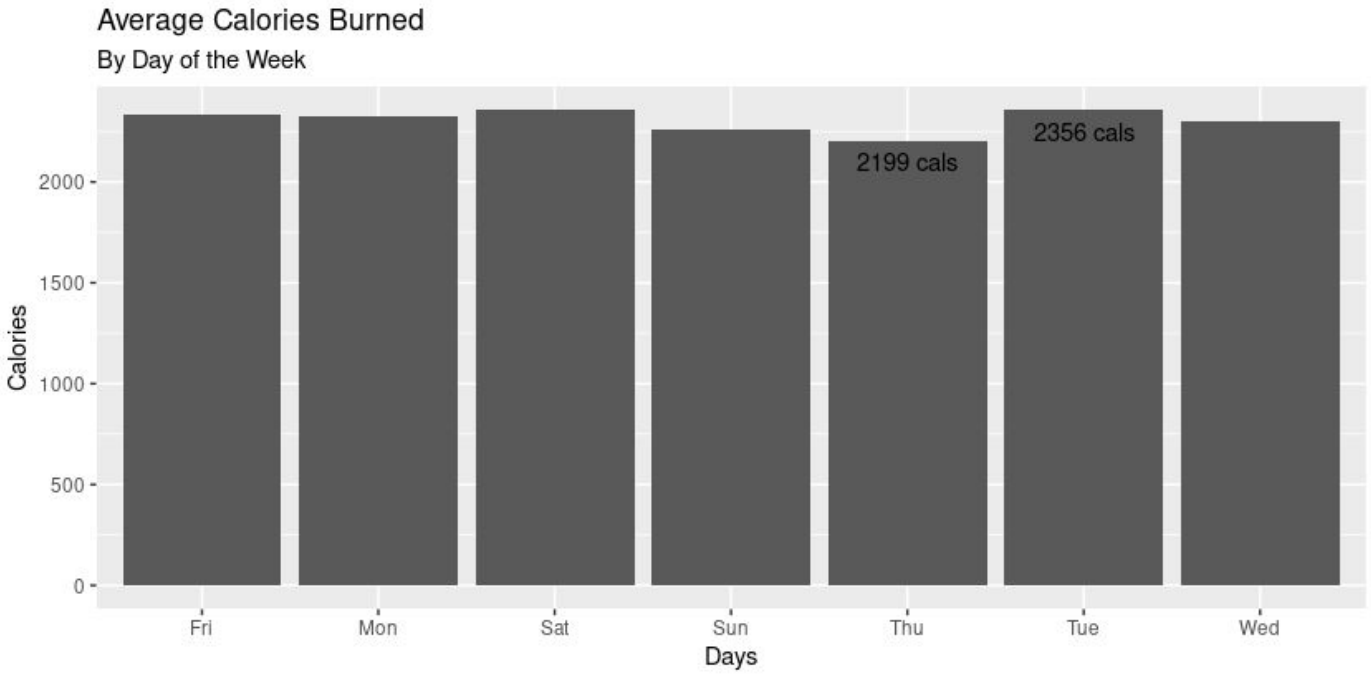
- The relationship between “Total Steps” and “Calories” has a somewhat strong positive relationship. The Correlation Coefficient is 0.59.
- However, the “goodness-of-fit” R value displays a .35 which means the linear trend line is fairly weak.

| SUMMARY OUTPUT | |
|---|---|
| ***Regression Statistics*** | |
| Multiple R | 0.59251635 |
| R Square | 0.35107562 |
| Adjusted R Square | 0.35038306 |
| Standard Error | 579.081699 |
| Observations | 940 |
| | |
| ***Coefficients*** | |
| Intercept | 1665.2170 |
| Calories | 0.0837 |

# Step 5: “Share” Phase - Summary

- Data Visualizations
  - Bar Chart - Mean Calories by Day of the Week
    - To showcase the difference and comparison between average calories burned for each day of the week.
  - Scatterplot - Total Activity by Day of the Week
    - To display the amount of unique user variation in calories burned for each week day.
  - Scatterplot - Total Steps & Calories: Positive Trendline
    - To show the relationship of calories burned to total steps logged.

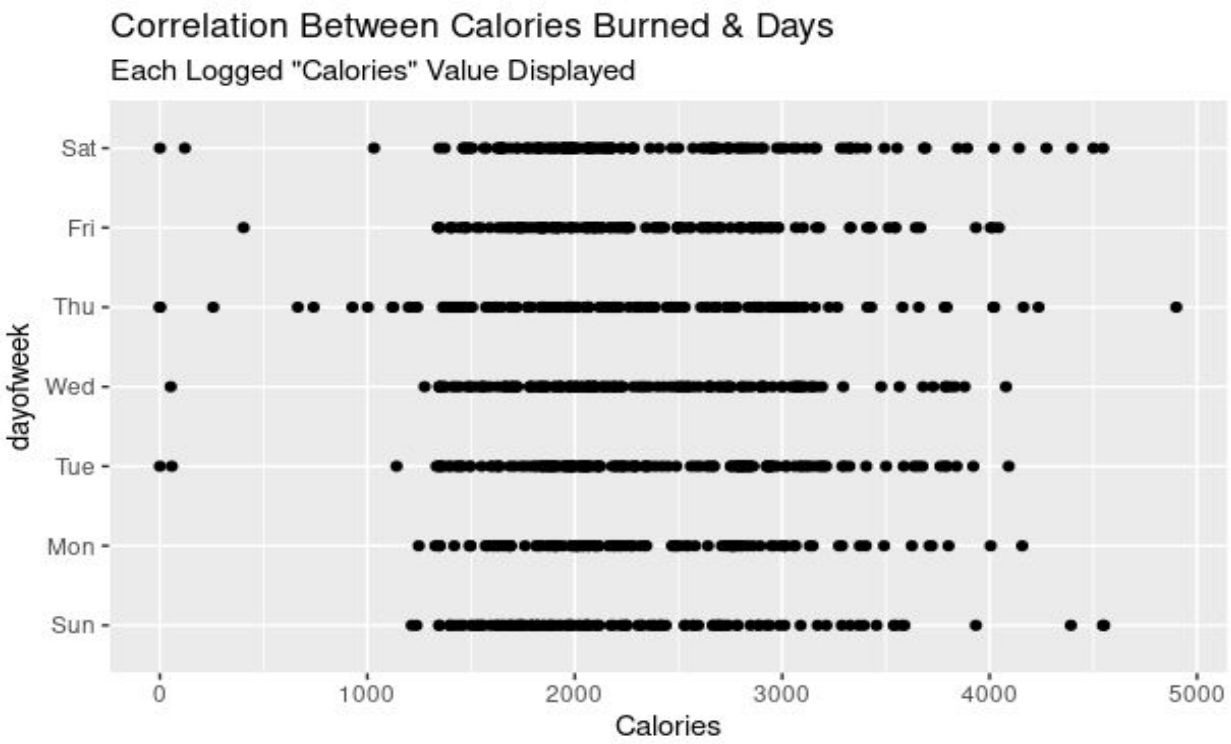
# Bar Chart - Mean Calories




- The difference between the mean calories for Tuesday and Thursday is 157.
- This is a substantial amount of calories that are not burned due to the user's activity.

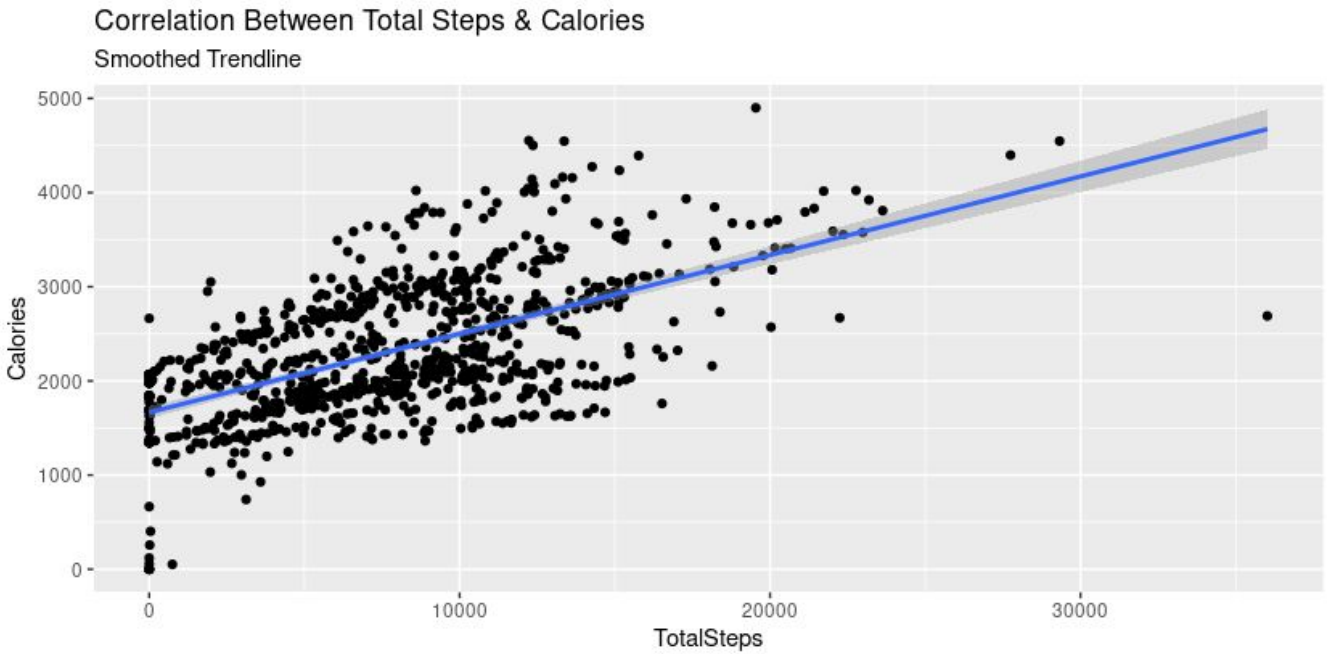
*Refer to Slide 12 for corresponding analysis*

# Scatterplot - Activity by Day




- Thursday shows the highest variation in calories burned with records trending towards the lower range.
- Saturday shows the second highest variation with records trending towards the upper range.

# Scatterplot - Trendline & Regression Analysis




- Linear Regression Equation
  - y = 0.0835x + 1665.7
- The relationship shows that the user can burned more calories by simply increasing steps taken for that day.

*Refer to Slide 12 for corresponding analysis*

# Step 6 : “Act” Phase – Main Recommendation

- Fitbit’s product analysis focused on:
  - User participation on recording their daily activity.
  - Relationship of calories burned compared to the days of the week.
  - Correlation of walking and calories burned.

- Based on the data collected:
  - Bellabeat should draft and create a notification system for the smartphone application that triggers when users are falling behind the average calories burned, did not record daily activity, and/or need extra encouragement due to lessening interaction with the application.
  - The alert system could reinforce the user’s perception that the application is attuned to their activities and has their best interest in mind.
    - Provides a truly “smart” application for the users.

# Challenges

1. Data from Fitbit did not disclose user gender for privacy reasons. This was limiting as Bellabeat's user base/target market is women.
    a. It would have been extremely helpful knowing which users were males and females as this would make for more precise analysis and actionable recommendations based on Bellabeat's target market.
2. Hard to derive actionable recommendations based on the data set.
    a. The data set was populated with records of different users' activities such as steps, active minutes, and calories burned. It was difficult to find a way to position the analysis to provide actionable insights besides a personalized alert system to encourage users to log their activities.
3. For simplicity, one merged data set was used, the other two were not utilized.
    a. Given more time, an identical analysis of "daily_sleep" and "daily_weight_log" should be performed and provide different recommendations based on the derived insights.

# Appendix

## R Markdown File

Please use the the R Markdown file titled *"GDA_CS_#2_RMD_V6"* located in the project's web page.

## Sources


"Refurbished Apple Watch Series 4 GPS + Cellular, 40mm Gold Stainless Steel Case with Stone Sport Band." *Apple*,

https://www.apple.com/shop/product/FTUR2LL/A/Refurbished-Apple-Watch-Series-4-GPS-Cellular-40mm-Gold-Stainless-Steel-Case-with-Stone-Sport-Band.

Posted by Stephanie Glen on May 6, 2019 at 10:00am View Blog. "The Lifecycle of Data." *Data Science Central*,

https://www.datasciencecentral.com/profiles/blogs/the-lifecycle-of-data.

"What Is Kaggle and How Exactly Should I Use It?" *Quora*,

https://www.quora.com/What-is-Kaggle-and-how-exactly-should-I-use-it.