Merit America Google Data Analytics Certificate Capstone Project

Case Study: Bellabeat Analysis

Author: David Bastawros

Date: 02/26/2024

Email: dbastawros03@gmail.com

GitHub

LinkedIn

Introduction

Welcome to my Capstone project completed as part of the Google Data Analytics Certificate program in collaboration with Merit America. In this project, I assumed the role of a marketing analyst tasked with conducting a case study. The focus of my study was on Bellabeat, a distinguished high-tech manufacturer based in San Franscisco. Bellabeat has gathered data on activity, sleep, stress, and health with its devices. Bellabeat intends to tailor its products and services to meet the unique needs of its users using this data. The data was collected in 2016.

Important Details About Bellabeat

Key Stakeholders

- Sando Mur: Cofounder and Mathematician
- Urška Sršen: Cofounder and Chief Creative Officer

1) Ask - Key Questions

- What are some trends in consumers' use of the smart device?
- How can the insights from these trends be utilized to adapt Bellabeat's marketing approach?
- How can the the insights from these trends be utilized to adapt Bellabeat's products and services?

2) Prepare Data

In the end, I elected to focus on three main data sets: Daily Activity, Weight Log Information, and Sleep Day. It was evident that the data set had some limitations. There was a lack of information on each individual, such as gender or age, which is very important when looking at health statistics. ID was often labelled as an integer. There also appeared to be some empty data, such as logged_activities_distance and sedentary_activities_distance. I also had a concern about duplicate data, with total_distance and tracker_distance seeming identical.

3) Clean Data

I started with fixing Id's labelling. For empty data, I would've liked to contact the team who gathered the data to ask about it. Was this intentional, or was there a mistake? After that, I opted to verify if my suspicions regarding the similarity of the two columns (total_distance and tracker_distance) were accurate. I set up an additional column, distance_diff to measure the difference between the columns. At a glance they still seemed identical, but there was 940

different rows, and I decided to write a code to tell me whether all columns were the same.

In [9]:
```
da['distance_diff'] = da['TotalDistance'] - da['TrackerDistance']
```

In [10]:
```
da['distance_diff']
```

Out[10]:
```
0      0.0
1      0.0
2      0.0
3      0.0
4      0.0
      ...
935    0.0
936    0.0
937    0.0
938    0.0
939    0.0
Name: distance_diff, Length: 940, dtype: float64
```

In [11]:
```
def check_same_column_values(da, distance_diff):
    # Check if all values in the specified column are the same
    return da[distance_diff].nunique() == 1
```

In [12]:
```
result = check_same_column_values(da, 'distance_diff')
print("Are all values in the column the same?", result)
```

```
Are all values in the column the same? False
```

In the end, I found there to be a handful of rows that did have differing results, indicating that while the results of the columns often coincided, they were not identical. I then renamed the

In [21]:
```
da.isna()
any_null_values = da.isna().any().any()

if any_null_values:
    print("There are null values in the DataFrame.")
else:
    print("There are no null values in the DataFrame.")
```

```
There are no null values in the DataFrame.
```

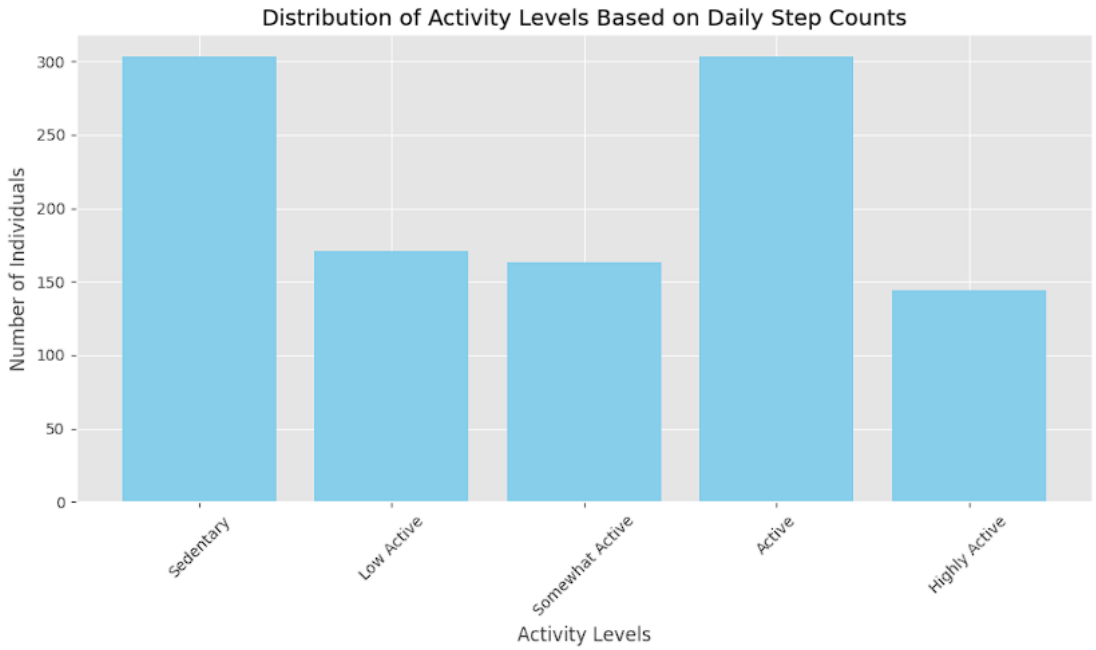
In [22]:
```
#Review for Duplicate Values
da.duplicated().sum()
```

Out[22]:
```
0
```

columns to all be in lowercase with underscores separating the words for readability. Finally, I examined for null and duplicate values.

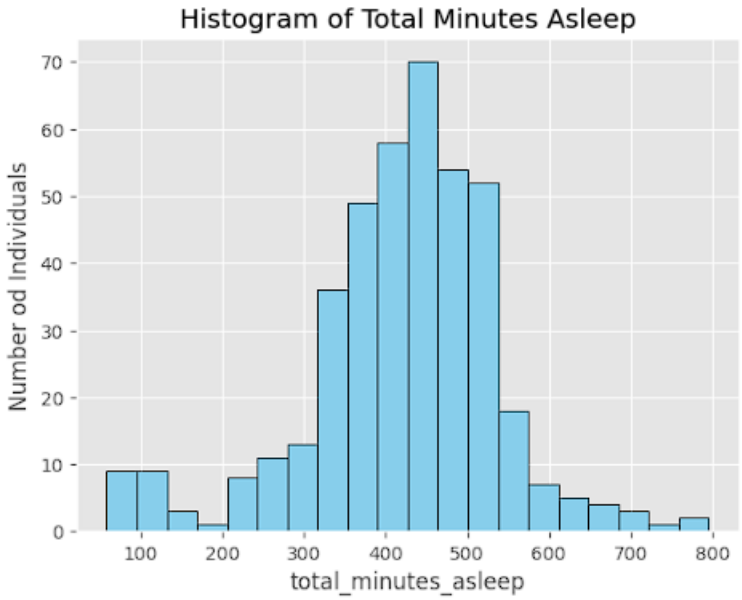
4) Analyze

Initially, I conducted an analysis of the activity levels chart, with a specific focus on the total step count. I proceeded to organize these step counts according to different activity levels, utilizing definitions outlined in a publication from the National Library of Medicine titled "How many steps/day are enough? Preliminary pedometer indices for public health" by Tudor-Locke C and Bassett DR Jr. (Sports Med. 2004;34(1):1-8. doi: 10.2165/00007256-200434010-00001. PMID: 14715035).

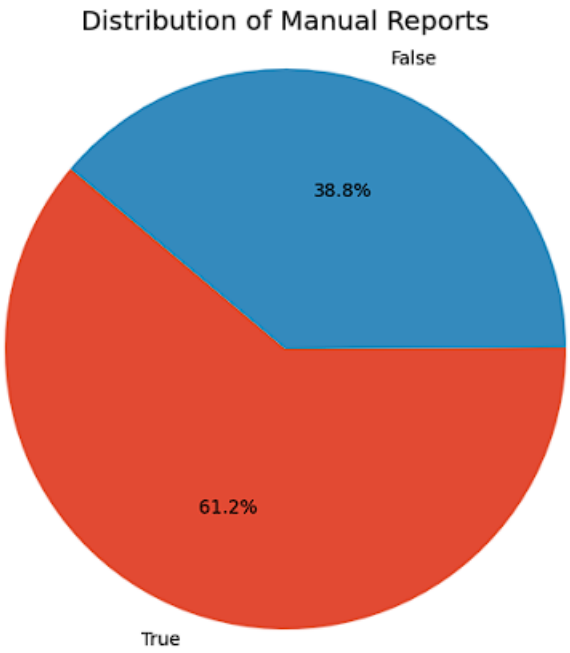


The bar chart illustrates the distribution of activity levels among 33 consumers, each of whom submitted multiple records reflecting various levels of activity. The majority of records indicate sedentary behavior or active. This diverse distribution underscores the variability in activity levels within the consumer group and provides insight into their overall physical activity patterns. Additional information regarding these individuals should be obtained if possible. It

would allow stronger conlusions to be drawn. I also established a histogram denoting sleep

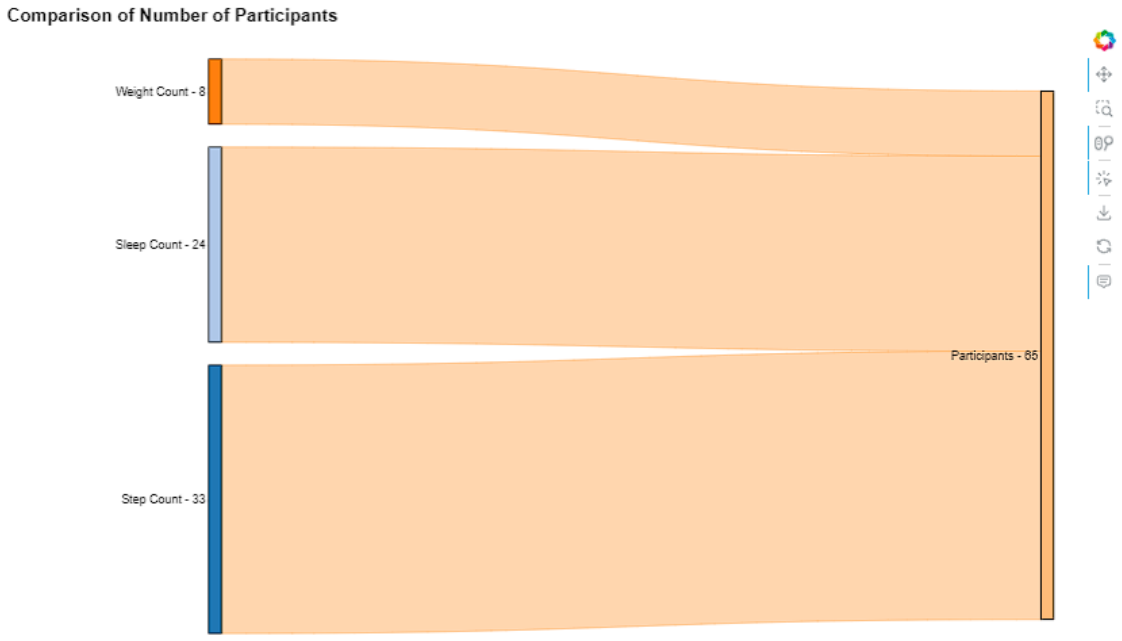


hours. This measures 24 individuals, and majority of nights consumers tend to sleep between 6 and 8 hours. This was followed by a pie chart on weight, but this time I decided to focus on



consumer participation. I found that majority of participants did not manually report their weight. Combine this with my final chart, a sankey diagram, and we see that this study only had 8

participants. This minute population, combined with the lack of data tailored to each indivdual



leaves us unable to draw many, if any, strong conclusions about Bellabeat's vast consumer base.

5) Act

It is imperative for Bellabeat to gather more comprehensive consumer data, which can be acquired through various means. Since the release of their app in 2014, updating the profile system to incorporate user input like height and weight would not only facilitate more accurate health outcomes for consumers but also furnish Bellabeat with enhanced information. Furthermore, incentivizing users to consistently track data, such as weight and sleep, could yield mutual benefits. Implementing a rewards program or enhancing notifications may serve as effective strategies in this regard. Additionally, tailoring Bellabeat products to individual user needs and comfort levels, coupled with incorporating features like sleep reminders, can further encourage user engagement and overall well-being.