Data Analytics Capstone

Wearable Technology - Data Analysis

Ezra Gershman

Introduction

Bellabeat is a **high-tech manufacturer of health-focused products for women**. They are a successful small company, but they have the **potential to grow** even further and become a larger player in the **wearable smart device industry**. The co-founder and Chief Creative Officer of Bellabeat, ***Urška Sršen***, believes that **analyzing smart device data** could help unlock new growth opportunities for the company.

The **Bellabeat IVY+ Health Tracker** is the only tracker designed and engineered for women, which correlates **menstrual cycle data, lifestyle habits, and biometric readings**. It reveals a comprehensive and accurate state of your body and mind.[1] The IVY+ sells on the Bellabeat website for €249.00 which converts roughly to **$275.00**. The IVY+ Health Tracker has the following features:

- **Long Battery Life** (up to 8 days)
- **Pedometer**
- **Automatic Distance and Activity Tracker**
- **Calorie Tracker**
- **Automatic Sleep Tracker**
- **24/7 Heart Rate Monitor**
- **Breathing Rate Monitor**
- **Temperature Monitor**
- **Stress Level Monitor**
- **Hypoallergenic**
- **Waterproof** (up to 5 ATM)
- **Advanced Cycle Tracker**
    a. Menstrual predictor
    b. Ovulation tracking
    c. Symptom input
- **Compatible with iOS 14.0+ and Android 8.0+**

[1]As described on the Bellabeat website.

Ask

- **We want to understand how users engage with health tracking features in smart devices, particularly for the Bellabeat IVY+.**
    - **Understanding Engagement**: Our goal is to determine how effectively users interact with the IVY+ and its features. This includes identifying which features—such as menstrual tracking and sleep monitoring—users find valuable and which ones they tend to overlook or underutilize.
    - **Identifying Gaps:** By analyzing data from other smartwatches and fitness trackers, we can spot areas where users may struggle or where their needs aren't fully met. This understanding can reveal why some users might stop using their devices altogether, helping us pinpoint specific aspects of the IVY+ that could be improved to enhance overall user satisfaction and retention.
- **In essence**, the goal is to **leverage data insights** to create a **better product experience** that keeps users engaged and satisfied with the IVY+.

Prepare

This analysis utilizes multiple datasets stored on my local computer, with regular backups maintained in Google Drive to ensure data security and accessibility. This dual storage method minimizes the risk of data loss while adhering to best practices in data management. Each dataset provides valuable insights into user engagement with health-tracking features and consumer behavior, which are crucial for understanding the impact of wearable technology on health. Personal identifiers will be managed to protect user privacy, and data integrity will be verified through consistency checks for missing or erroneous values.

- **Fitabase FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius):**
  This dataset is stored on Kaggle and comprises multiple CSV files organized into two primary export periods: 3.12.16-4.11.16 and 4.12.16-5.12.16. It captures daily metrics, including activity, heart rate, calories, steps, sleep, and weight logs. The data is structured in long format, with each row representing a unique observation per day. Potential biases may arise due to the specific demographic of Fitbit users, which could affect the generalizability of findings. This dataset is essential for analyzing user engagement with the Bellabeat IVY+.

- **One year of Fitbit ChargeHR data - Alket Cecaj:**
  Authored by Alket Cecaj and stored on Kaggle, this dataset includes a single CSV file capturing daily activity metrics for one individual over a year. It details various daily metrics, such as calories burned, steps taken, and minutes of different activity levels. Last updated in 2018, the dataset may not fully represent current user behaviors. This dataset helps answer research questions by providing comprehensive daily metrics crucial for analyzing user engagement with health tracking features.
- **Fitness tracker data (2016 - present) [2450+days] (CC0: Public Domain) - Damir Gadylyaev:**
  Stored on Kaggle, this dataset consists of two primary CSV files: 01_Steps.csv and 02_Sleep.csv. Each file contains multiple entries for daily metrics on steps and sleep patterns, organized in long format. The single-user focus may introduce data bias, limiting generalizability. This dataset provides comprehensive metrics essential for analyzing user engagement over time.
- **LifeSnaps Dataset (Attribution 4.0 International):**
  The LifeSnaps Dataset includes a primary CSV file named daily_fitbit_sema_df_unprocessed.csv, featuring anthropological data from 71 participants over four months. The long format data captures daily metrics from Fitbit devices, including physical activity and sleep patterns. This multi-modal dataset provides insights into user behavior and well-being, although participant demographics may affect generalizability. This dataset is vital for analyzing user engagement with health tracking features in diverse contexts.
  doi:10.1038/s41597-022-01764-x
  Accessed data via Kaggle.
- **Dataset from *JAMIA* Research Study:**
  Hosted on the Dryad digital repository, this dataset accompanies the research paper "Assessing the Impact of Wearable Health Devices on User Engagement" published in JMIR (doi:10.1093/jamiaopen/ooab054). It provides detailed daily metrics on two users' step and sleep behaviors: a father and daughter. While the dataset's representativeness may be influenced by specific demographics and collection periods, it is essential for analyzing engagement patterns and usage trends.
  doi:10.7272/Q6NG4NWT
  Accessed data via Dryad.
- **Amazon Product Scraping Dataset:**
  This dataset includes details such as product names, descriptions, prices, ratings, and customer reviews, offering insights into consumer behavior and preferences. It was scraped using a Python program I developed with Selenium, enabling a deeper understanding of market trends and product characteristics.

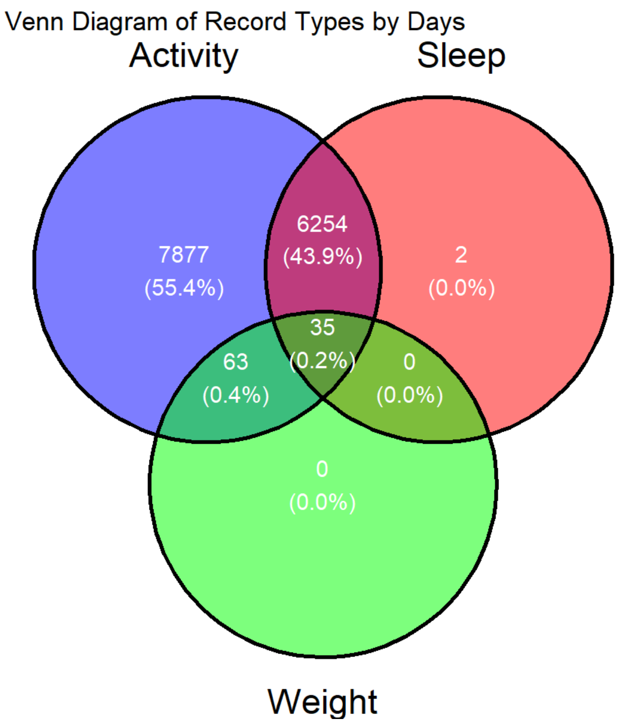
Processing

In order to prepare the various datasets for analysis, I used R and RStudio for their robust data manipulation capabilities, which were very helpful when handling large datasets. Initially, the first step involved merging the daily sleep, activity, and weight log datasets for all users into a single comprehensive user daily activity log. This process required ensuring consistent date formats and correctly aligning records so that, for each day and user, there was a unique record. To maintain data integrity, I ensured that all entries were accurate, consistent, and complete. Additionally, I standardized formats across datasets and addressed any missing values, thereby enhancing the reliability of the data for subsequent analysis.

Following this, I standardized the datasets to resolve naming inconsistencies and streamline the data for analysis. This included renaming columns for clarity and merging records while addressing any duplicates. To ensure that the data is clean, I conducted checks for missing values and duplicate entries before finalizing the dataset. Furthermore, I documented the entire cleaning process, outlining the steps taken, which will facilitate review and sharing of results in future analyses. Overall, this structured approach ensures that the dataset is not only clean but also reliable for meaningful insights.

Analysis

Analysis of User Logging Patterns



The Venn diagram illustrates the overlap between days when users logged activity, sleep, and weight records. This provides a comprehensive view of how users engaged with the different health-tracking metrics across unique days.

Activity Logging: The set of days with activity logs is notably large, comprising 99.9% of the total days logged (55.4% for activity-only days, 0.4% overlapping with weight-only, 43.9% with sleep-only, and 0.2% overlapping with both sleep and weight). This indicates that activity tracking is a central feature of user engagement with the wearable device, which is to be expected. Only less than 0.1% of the recorded days lack any fitness tracker information, suggesting that nearly all logged days include some level of physical activity monitoring.

Weight Logging: Interestingly, there are no days in the set of days with weight logs that are not accompanied by other types of logs. Specifically, every day that includes weight data also intersects with either activity logs (0.4%) or logs for all three metrics—activity, sleep, and weight (0.2%). This suggests that users who track their weight are likely to be more engaged with a broader range of health metrics, using their wearable devices to monitor multiple aspects of their health simultaneously, rather than focusing solely on weight.

Key Statistics of User Logging Behavior

Engagement with Data Logging:

- **Mean Unique Days:** Users logged data for an average of **129 days**, reflecting a moderate level of engagement with the health-tracking platform.
- **Median Unique Days:** The median is **69 days**, indicating that half of the users logged data for 69 days or fewer. The difference between the mean and median suggests that a small number of users are highly engaged, which skews the average upward.
- **Minimum Unique Days:** The user with the fewest logs recorded data for only **8 days**, indicating low engagement or potentially limited use of the device.
- **Maximum Unique Days:** One user logged data for **2417 days**, representing a high level of commitment and long-term tracking behavior. This user is likely a dedicated, daily user of the platform.

Variability in Logging Patterns:

- **Standard Deviation of Unique Days:** A high standard deviation of **315.08** suggests considerable variation in the number of days users log data. This supports the observation that while many users log data sporadically, a few users maintain consistent tracking over extended periods.

Activity Tracking:

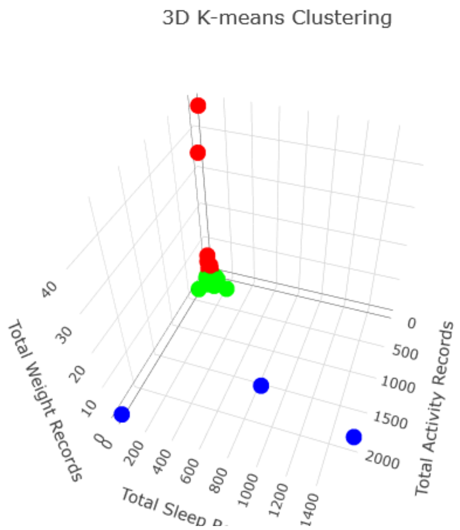
- **Mean Activity Records**: On average, users logged **130 activity records**, highlighting consistent engagement with tracking physical activities.
- **Median Activity Records**: The median value is **69 records**, suggesting that a few highly engaged users are contributing a large proportion of activity logs, while others are less consistent.

Sleep Tracking:

- **Mean Sleep Records**: Users logged an average of **57 sleep records**, showing that sleep tracking is less frequent compared to activity tracking.
- **Median Sleep Records**: A median of **31 days** indicates that many users recorded sleep data for **31 days or fewer**, suggesting inconsistency in sleep logging among a significant portion of the user base.

Weight Tracking:

- **Mean Weight Records**: The low average of **0.9454545** suggests that weight logging is rare among users, with many users opting not to record this metric.
- **Median Weight Records**: A median of **0** indicates that half of the users did not log any weight records, pointing to a notably low level of engagement with weight tracking features.



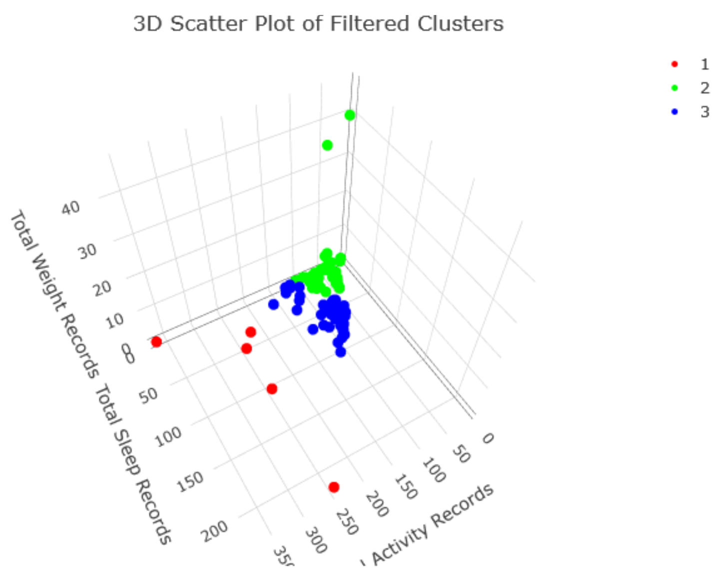
Key Statistics of User Logging Behavior

Clustering is a data analysis technique that groups a set of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. The goal of clustering in this analysis is to identify distinct user segments based on their engagement with health-tracking features, allowing for tailored insights and strategies that address the diverse needs and behaviors of different user groups.

- Red (Cluster 1): Characterized by the existence of any weight records whatsoever. The average number of sleep and activity logs is fairly normal.
- Green (Cluster 2): Close to the center of the plot, indicating mid-range amount of usage for activity and sleep.
- Blue (Cluster 3): Spread towards vastly higher values of total sleep and activity records indicating high usage.

Cluster	Total Users	Activity Records	Sleep Records	Weight Records
1	53	51.9	14.8	1.96
2	54	105	59.1	0
3	3	1960	790	0

Upon reviewing the clusters, it became evident that individuals in Cluster 3, characterized by having more vastly more records, significantly skewed the results. Consequently, we needed to filter out the Cluster 3 users from our clustering to ensure a more accurate representation of the remaining data. These new clusters can hopefully give us some insights into the general behavior patterns of different kinds of users.



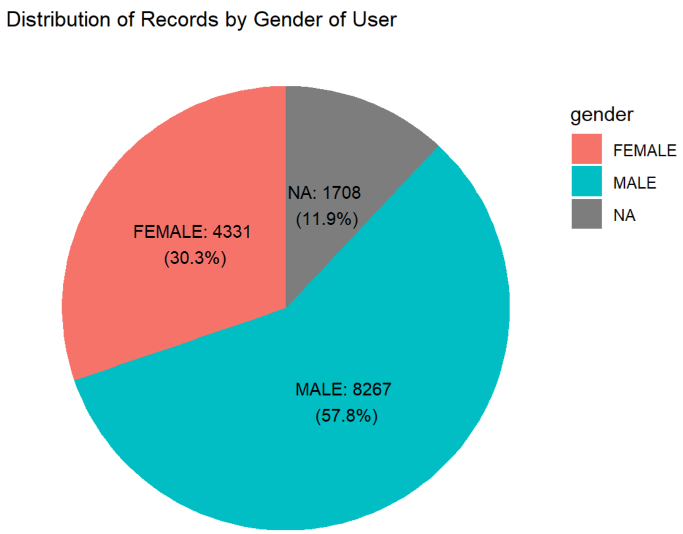
- Red (Cluster 1): Concentrated mainly at higher amounts of activity records and sleep records.
- Green (Cluster 2): Close to the center of the plot, indicating low-range amounts of records for those users for activity and sleep. This is also the cluster where users with any amount of weight records can be found.
- Blue (Cluster 3): Generally mid-range amount of records for both activity and sleep.

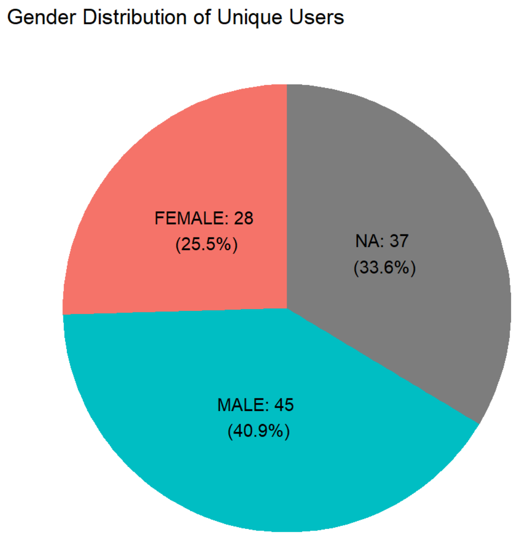
New Cluster	Total Users	Activity Records	Sleep Records	Weight Records
1	5	264	83.4	0
2	48	50.5	11.5	2.17
3	54	86.6	55.7	0

Data Regarding Female Users

In our analysis, isolating data from female users is essential given that Bellabeat specializes in health-focused products specifically designed for women. Understanding how female users engage with their activity trackers. By focusing on female-specific data, we can identify which features resonate most with this demographic and where potential gaps in user experience exist. I have made two pie charts to break down the demographics of our data by gender.

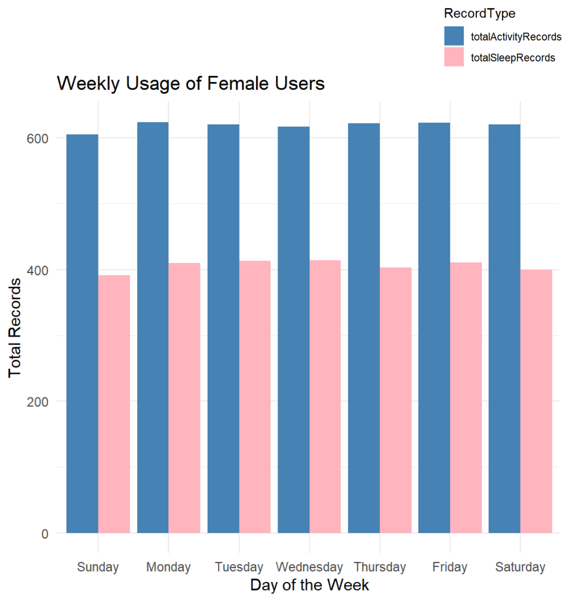
The first pie chart depicts the gender distribution of unique users, highlighting the representation of female users within the dataset. The second pie chart shows the total records logged by gender, indicating differences in user engagement levels.

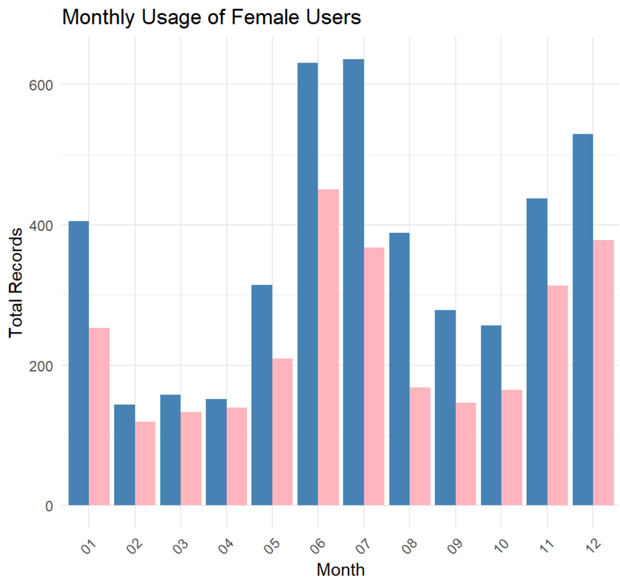




The metrics suggest that while there is a healthy representation of female users, their engagement levels are lacking compared to male users. This gap highlights the potential for targeted marketing campaigns or product enhancements that specifically address women's health concerns and promote features like menstrual tracking, sleep monitoring, and stress management.

To do further analysis on the data from the female users we will filter the data to include only records where the gender is identified as “FEMALE” or “UNKNOWN.” This approach allows us to focus on these two groups for further analysis; for example, we can look at the days of the week or the months that we have data from female users, and can use that to inform our approach.



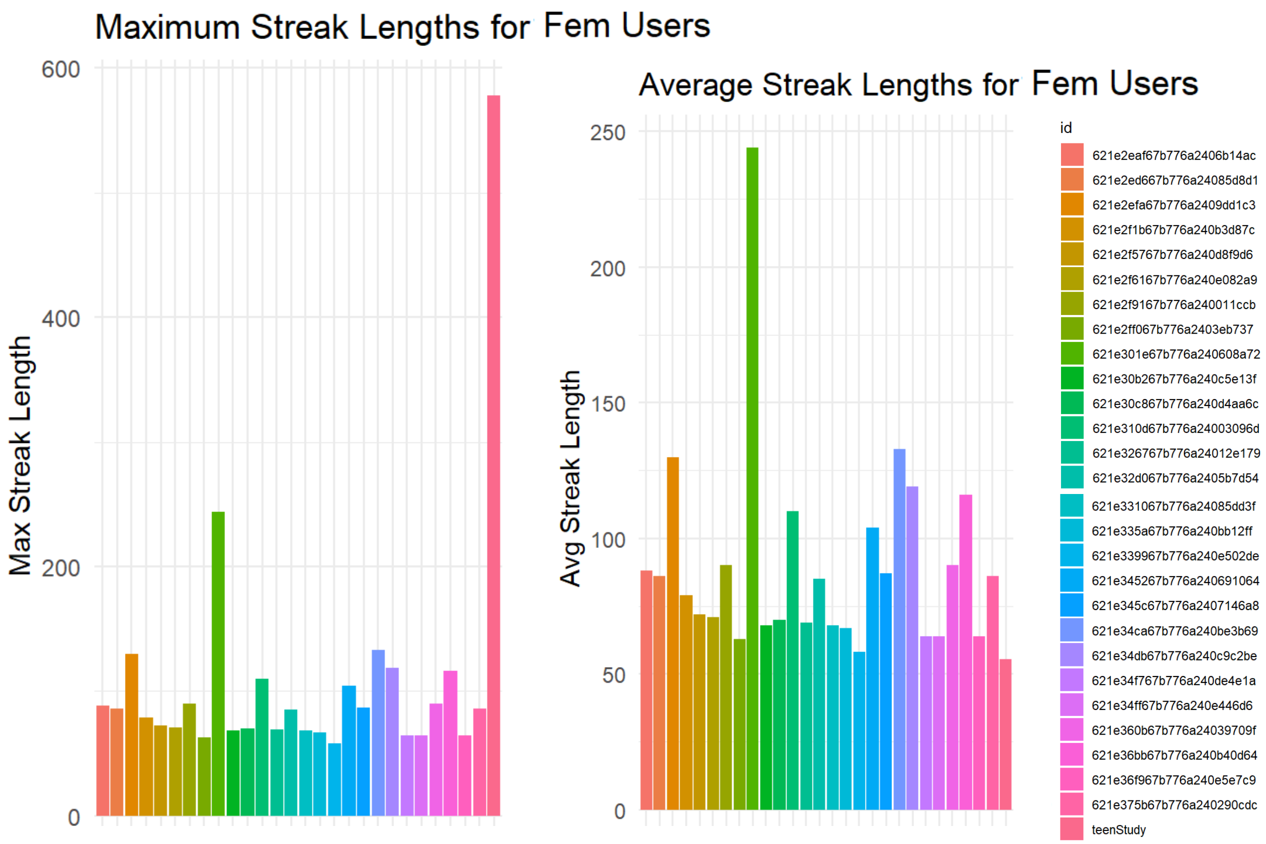


Upon analyzing this data, three observations are clear:

- **Lack of Weight Records:** There are no users with recorded weight data in this subset, which aligns with expectations as weight records were initially sparse across the datasets.
- **Even Distribution of Usage Across Days of the Week:** The analysis of records across different days of the week indicates that user activity and sleep tracking are relatively consistent. There is no significant variation in usage patterns from one day to another, suggesting that day of the week does not strongly influence users’ tracking behavior.
- **Monthly Usage Trends:** The monthly analysis reveals two notable peaks in usage: May-July and November-January. The high usage during summer months may be attributed to higher physical activity levels during the summer, as users are more active and likely to spend time outdoors, which encourages them to track their activities and sleep more regularly. Similarly, the high usage in the winter months could be due to users receiving new fitness trackers as gifts, or a renewed

focus on health and fitness as part of New Year's resolutions, leading to a temporary boost in tracking activity and sleep data.

While our analysis highlights key trends in female user engagement, it's also essential to examine the consistency of usage over time, particularly through the lens of tracking streaks, which can indicate consecutive days of logged data. In the following section, we will delve into the streak analysis for female users, exploring both the duration of streaks and the implications for user engagement.

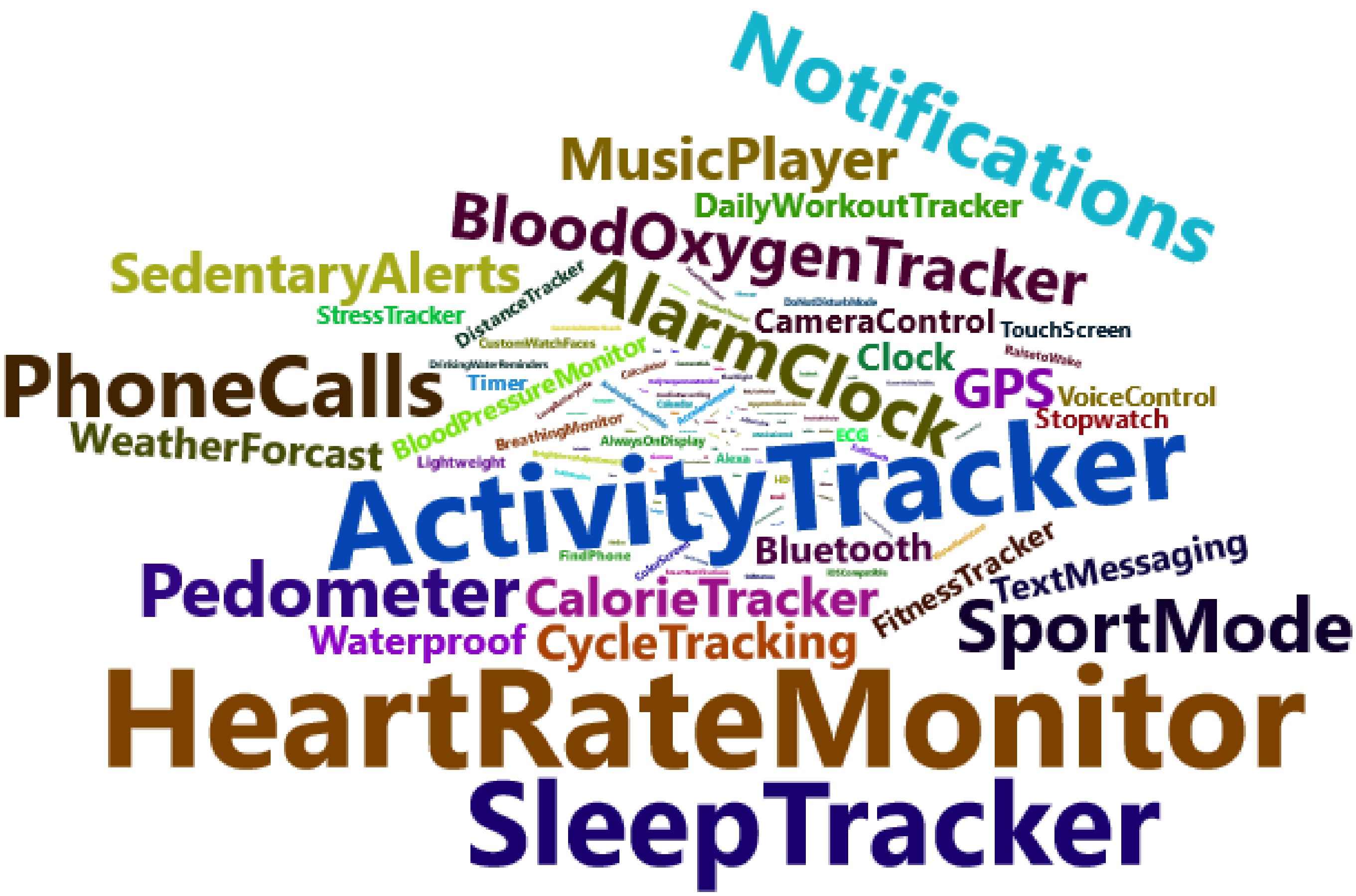


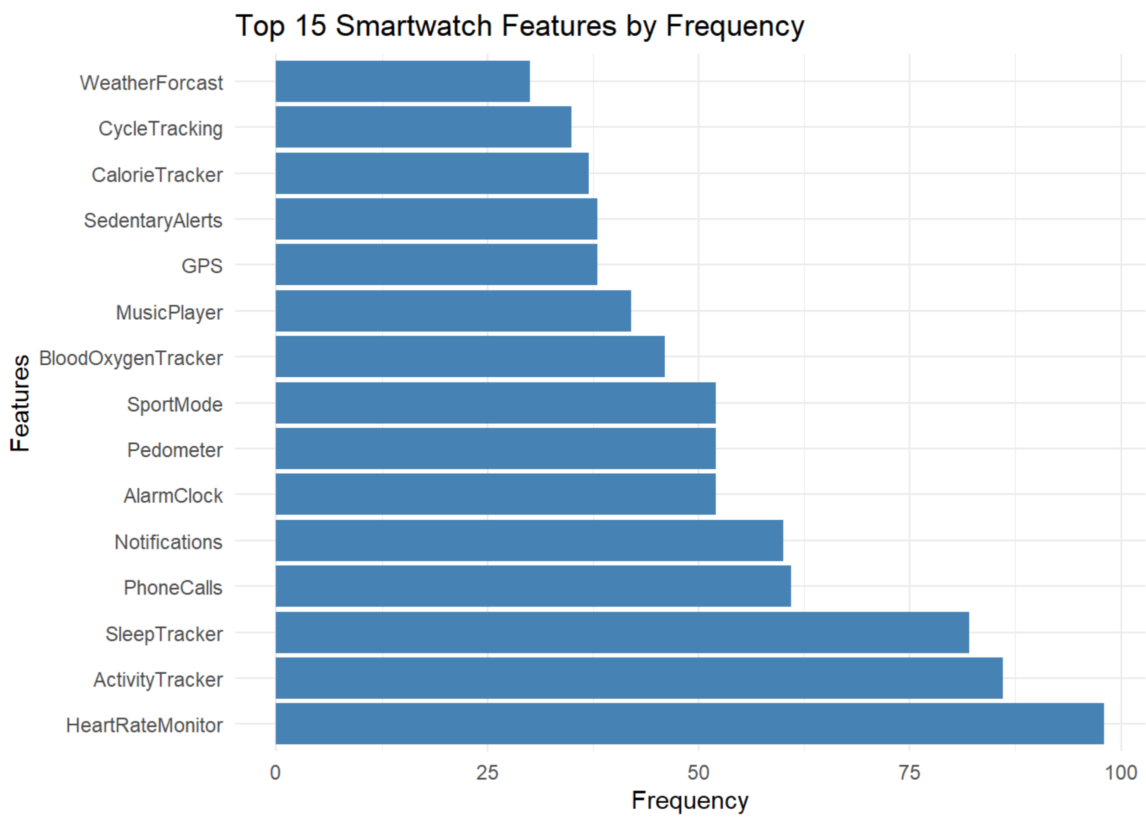
Most users have average streak lengths ranging from 50 to 150 days. A few users deviate significantly, with one user (green bar) having a notably higher average streak length of around 250 days. This indicates a consistent usage pattern or engagement over time compared to others. While many users show maximum streak lengths around 100 to 200 days, one user (pink bar) stands out with a maximum streak length reaching around 600 days. This suggests a period of sustained engagement for that user, even though their average streak length may be lower.

The users with higher average streak lengths tend to have relatively high maximum streak lengths, but not necessarily the highest. For example, the user with the highest average streak length in the second graph (around 250 days) doesn't have the highest maximum streak length in the first graph. The user with the highest maximum streak (around 600 days) has a substantial deviation between their average and maximum streak lengths, suggesting a single or few long streaks rather than consistent long-term behavior.

Smartwatch Data from Amazon

I wrote a Python Selenium script in order to scrape the top 100 products from specific categories on Amazon, gathering detailed information such as product names, prices, ratings, and other relevant data. The 'special features' of the top 100 smartwatches and fitness trackers were separated, tokenized, and standardized to provide insights into their usage and popularity among consumers.The frequency of each feature from the top 100 smartwatches and fitness trackers was counted and summarized into a data frame, which was then used to generate a visually engaging word cloud to illustrate the most common features. Additionally, the top ten features were compiled in a bar graph for easy viewing and comparing. For an interactable display of the word cloud, see the attached RMarkdown notebook.





The word cloud highlights the most frequently mentioned features that are relevant to smartwatches in general, which can be compared against the specific features of the Bellabeat IVY+.

Terms like **HeartRateMonitor**, **ActivityTracker**, **SleepTracker**, and **BloodOxygenTracker** reflect the high interest in health and wellness functionalities that users expect from wearable devices. Additionally, the word cloud's emphasis on **Cycle Tracking** highlights the importance of menstrual health management, which is a critical feature of the IVY+. The IVY+ includes advanced functionalities like **Menstrual Predictor**, **Ovulation Tracking**, and **Symptom Input**, tailored specifically for women. This alignment emphasizes Bellabeat's commitment to addressing the unique health needs of its female user base.

Takeaways

Final Conclusion Based on the Analysis

The analysis reveals that while Bellabeat's IVY+ Health Tracker aligns well with user interests and habits in health tracking features, there is significant room for growth in user engagement, particularly among female users.

The data suggests that female users, despite being a key target demographic, are not fully utilizing the comprehensive tracking capabilities of their fitness trackers, such as sleep monitoring and activity tracking. Furthermore, the absence of weight records across the female user dataset points to potential gaps in feature adoption. Although fitness tracking is a standard feature in most wearable tech, Bellabeat has a unique opportunity to differentiate itself from its competitors by focusing on its advanced cycle tracking, menstrual predictor, and stress monitoring features.

Engagement is consistent throughout the week but shows peaks during summer and the beginning of the year, suggesting seasonal influences like increased outdoor activities and New Year’s resolutions. The streak analysis also highlights that some users maintain regular engagement over extended periods, which could be leveraged to encourage consistent usage among a broader group.

How Can the Team and Business Apply These Insights?

- **Targeted Feature Education**: Bellabeat can focus on educating female users about the full range of features available on the IVY+, emphasizing the benefits of consistent tracking. For example, highlighting how cycle tracking can integrate with sleep and stress monitoring may encourage users to take advantage of these interconnected functionalities.
- **Marketing Campaigns**: The identified peaks in usage during summer and winter can guide the timing of marketing campaigns, such as fitness challenges in summer and wellness-oriented promotions around the New Year. Emphasizing features like activity tracking during and stress or sleep monitoring could resonate with users' seasonal interests.

- **Enhanced User Onboarding**: Implementing an improved onboarding experience that introduces new users to key features, like weight tracking or stress level monitoring, can help drive initial engagement and familiarity with the device's capabilities.

What Are the Recommended Next Steps for Stakeholders?

- **Feature Development**: Consider adding reminders or alerts for features that are underutilized, such as weight tracking, to encourage more consistent use. Additionally, introducing gamification elements (e.g., achievement badges for consistent tracking streaks) could increase long-term engagement. Although the IVY+ does not have a screen, this could be implemented through the companion app.
- **User Feedback Surveys**: Conduct surveys or interviews with female users to understand why certain features like weight tracking are not widely used and gather suggestions for feature improvements. This qualitative data can complement the existing quantitative analysis and provide deeper insights into user preferences.
- **Focus on Retention**: Use insights from the streak analysis to develop retention strategies. For instance, reward programs for maintaining consistent tracking streaks could help motivate users to engage more regularly over time.

What Additional Data Could Expand the Analysis?

- **Detailed Demographic Data**: Additional demographic information such as age, location, or fitness goals could provide a more nuanced understanding of how different groups of women use the IVY+. This would allow for further segmentation and more targeted feature development.
- **User Feedback Data**: Gathering user feedback directly through surveys or app reviews could help identify reasons behind the underutilization of certain features, like weight tracking, and guide improvements.