An Exploratory Health and Fitness Data Analysis using Smart Watch

Akshya Ramesh, Rathnapriya Gopalakrishnan, Gokul Ragunandhan Narayanasamy

Abstract

Analog watches evolved into digital watches and digital watches led to smartwatches. Now smartwatches generate a plethora of information that can be leveraged to learn about the health and fitness of their users. Our aim is to explore and analyze this data to investigate the association among fitness metrics and how they are influenced by various physical activities.

Introduction

Fitness trackers like smartwatches give users important information that can help in the improvement of their health. Users can monitor their step count, heart rate, and calories burnt to determine which aspects of his/her health need to be improved. Advanced fitness trackers can even provide information about how your body reacts to stress, temperature, dehydration, blood glucose level, etc. A recent survey states that around half of the American population use smartwatches. 69% of them are willing to use smartwatches in order to get better health insurance. Around 92% of people with smartwatches use it to improve their health by tracking their daily activity. It has also been statistically proven that people using smart watches exercise on an average of 4.3 days/week when compared to non-users who only exercise for 3 days/week [10].

All these insights tell us that the amount of data collected by smartwatches is increasing day by day. Knowing its importance and its extensive usage, it is paramount that we analyze this data carefully to derive valuable insights. We also aim to test the precision of these smartwatch-collected data across various physical activities and determine whether it is consistent across various brands.

Related Work

We investigated the previous research on this dataset and looked into general visualizations of smartwatch data. We not only spotted a few drawbacks for some visualizations but also found a few visualizations very interesting and intuitive. We hope to improve upon these visualizations and keep in mind the lessons learnt to use our own visualizations for this dataset.

Here are a few visualizations that have be done on the dataset that we are planning to use and our critiques.

The below visualization fig 1 is used to represent how important the various features are. We can see that act_Running_7_METs have the highest importance and steps_times_distance has the lowest. This gives a comprehensive look into statistics of the different features and makes it easy to compare them. It would have been better to consider only the top 5 features for a single visualization. Since we have used so many categories, this visualization looks a little cluttered which can be avoided. So, in our visualization, we will try to limit the number of categories we plan to present in a single plot.

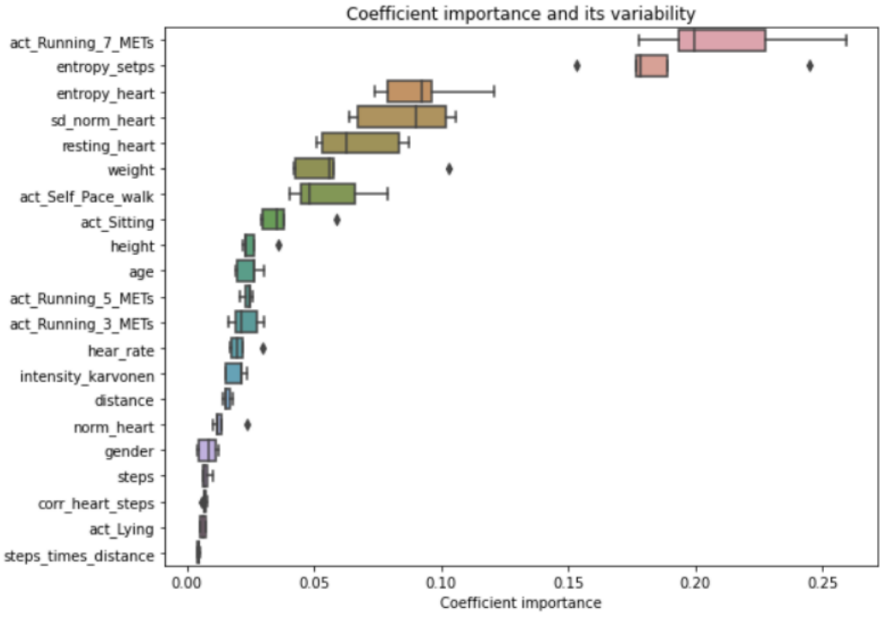


Fig 1: Box plots for features against importance [7]

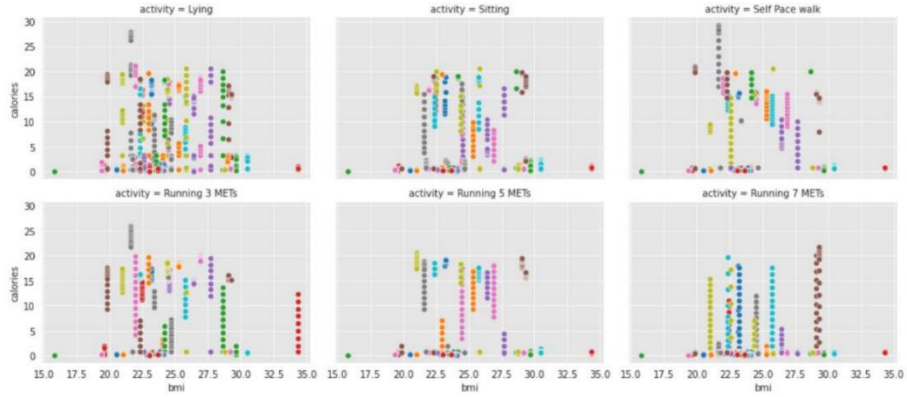


Fig 2: Scatter plot against calories, BMI and person [6]

Here, the plots have been created for BMI in the X axis and Calories in the Y axis, the color represents the persons performing the activity which is categorical data. This is not an advisable visualization as there are too many people visualized in one sub-plot and it is very hard to keep track of the progress made by each person. Either visualizing just a few people or highlighting a few people across the subplots to show consistency could have been more effective in delivering the message. The other issue we found is the lack of legends, we had to skim through the code to understand what was visualized here. Adding a legend could have improved the readability of the plot.

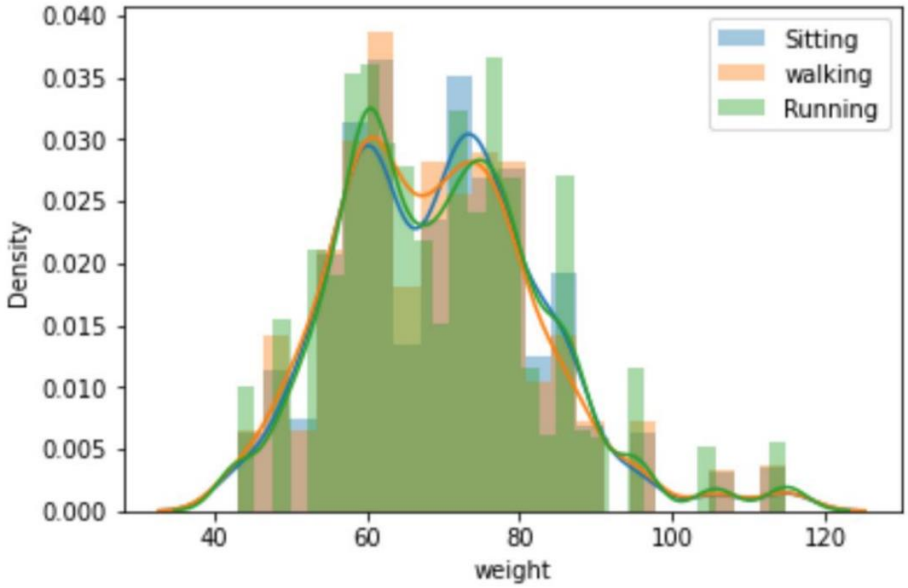


Fig 3: Histogram of weights [8]

This overlaid histogram visualization is used to analyze how the weight is distributed against sitting, walking, and running activities. The author has used different colors to represent various activities – Blue for sitting, peach for walking, and green for running. The main drawback of this visualization is not being able to see the distribution of data for the activities placed behind. Placing the histograms one behind the other is a good way to compare the data distribution but it also reduces the readability of the data. We strongly feel that the author could have increased the transparency of the colors or used distinct colors to represent each activity or even use 3x1 subplots where each row represents an activity.

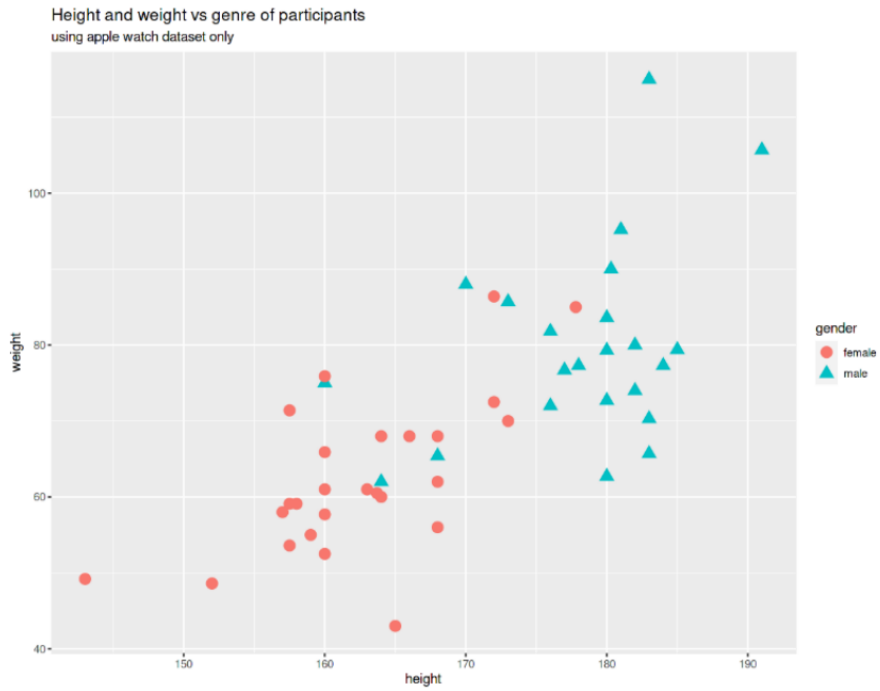


Fig 4: Scatter plot of height against weight for various genders [5]

The above visualization (fig 4) is used to see the relationship between height and weight. It is a scatter plot, and the color and shape attributes are used to differentiate between the different genders – male and female. We first thought that it is redundant to use shape for differentiation as it is already achieved though the usage of different colors. But then we realized that using different shapes helps when the visualization is transformed to black and white. The red and blue colors are too similar to each other, and the shape attribute proved to be essential in differentiating the two genders.

Fig 5: Dashboard for Apple Watch Data [3]

The above visualization is a dashboard of Apple Watch Data where different activities such as cycling running and swimming for different years 2018 to 2020 and various features such as minutes, distance, etc. are shown. It is very hard to interpret what these pie charts and grouped bar charts represent without proper titles and labels. The pie charts are also not labelled with their years and the middle pie chart sums up to 100.1%.

The below visualization (fig 6) is used to show the achieved goal in a week for features such as total sleep time, calories burnt, and total steps achieved daily. The author has also failed to add the legends. The first line of each color represents the minimum score, the middle line represents the target, and the third line represents the maximum score. There are many inconsistencies in the above figure such as, in the first circle the darker violet is used to represent the achieved score, but this is not consistent across different categories as lighter yellow is used to represent the achieved score for calories burnt and there are no variations in the blue sector.

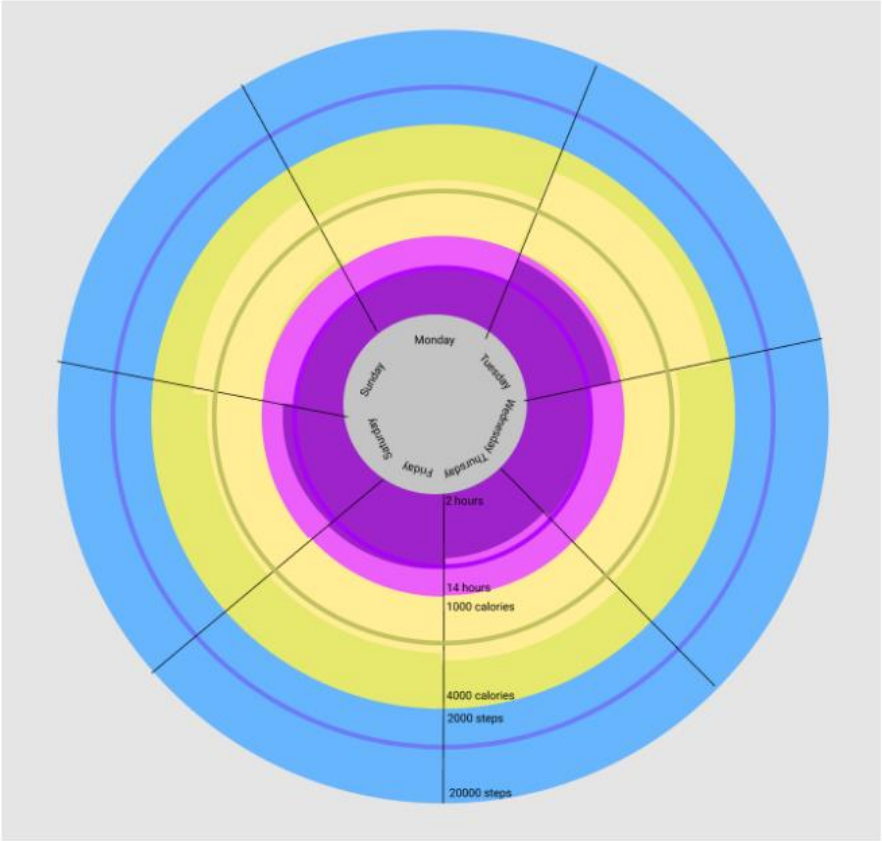


Fig 6: Visualization to represent weekly activity [9]

Dataset

Data source: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZS2Z2J

Data description

The data for this project is obtained from Harvard datasets where they conducted a study on participants wearing Apple watch Series 2 and Fitbit charge HR2. The participants were 23 men and 26 women, who performed a series of physical activities like lying, sitting, running 3, 5, and 7 meters. Metrics such as number of steps taken, distance covered, calories burned, are recorded and attributes such as age, gender, height and weight of the participants are noted down.

Data Explanation

Our data holds a mix of categorical and quantitative values, and a list of main attributes is provided below:

1. Age – This is a quantitative attribute that records the age of the participant and contains a range of values from 18 to 56. We combined them into age groups and treat it as categorical.
2. Gender – This is a categorical attribute that represents the gender of the participant (male/ female).
3. Height – Height of the participants in meters and is a quantitative attribute. This contains a range of values between 143 to 191 meters.
4. Weight – Weight of the participants is in kilograms and is a quantitative attribute. Contains range of values between 43 Kgs and 115 Kgs.

5. X1 – Is an identifier given for a person performing an activity and it is used to associate both apple and Fitbit data
6. Steps – This is a quantitative attribute and accounts for the steps taken while performing the activity.
7. Heart Rate – Records for the heart rate of the participant while performing the activity values ranges up to 194. This is a quantitative attribute.
8. Calories – This is a quantitative attribute and records the calories burnt while performing the activity. Its value ranges from 0.6 to 97.5.
9. Distance – This is a quantitative attribute and reports the distance travelled
10. Device – This is a categorical value that records if the activity was recorded using Fitbit or Apple watches.
11. Activity – This is a categorical value that records what kind of activity the participant is performing among the following and its durations
    - Lying down – 5 minutes
    - Sitting – 5 minutes
    - Self-paced walking/running on treadmill – 10 minutes
    - 3-meter running – 10 minutes
    - 5-meter running – 10 minutes
    - 7-meter running – 10 minutes

All these attributes will suffice to answer the following questions.

Objective

Our main objective is to use visualization to understand the relationship between various attributes and how different activities influence these features.

Visualizing the difference in the heart rate across six activities for Apple Watch

From Fig 7, we can see that the heart rate for lying activity ranges from 75 to 220 and the heart rate for sitting and self-paced walking follows the same distribution with values ranging from around 100 to 230. We can also see that as the level of running increases from 3m to 7m, the range of heart rate shifts higher. We believe apple does a fairly good job of capturing the different heart rates of each participant for running activities.

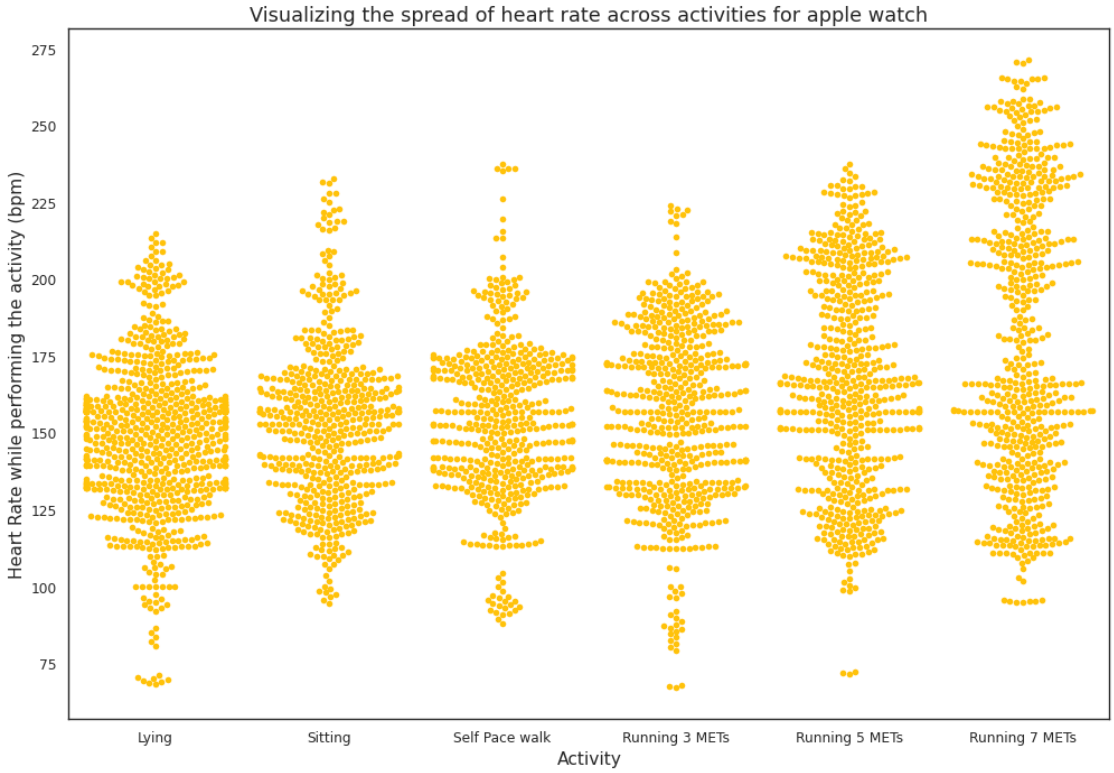


Fig 7: Visualizing heartrate across activity for Apple watch

Visualizing the difference in the heart rate across six activities for Fitbit watch

Fig 8: Visualizing heartrate across activity for Fitbit watch

The data collected from Fitbit for heart rate seems to follow the same distribution for most of the participants. We can see a lot of gaps in the plot and multiple points have the same value for each activity. All the activities have similar distributions including lying and 7-meter running which seems unlikely. To have a closer look at how these values differ for each device we have plotted a 1D scatter plot as seen in the below figure.

Visualizing the distribution of heart rate across different activities for Apple Watches and Fitbit

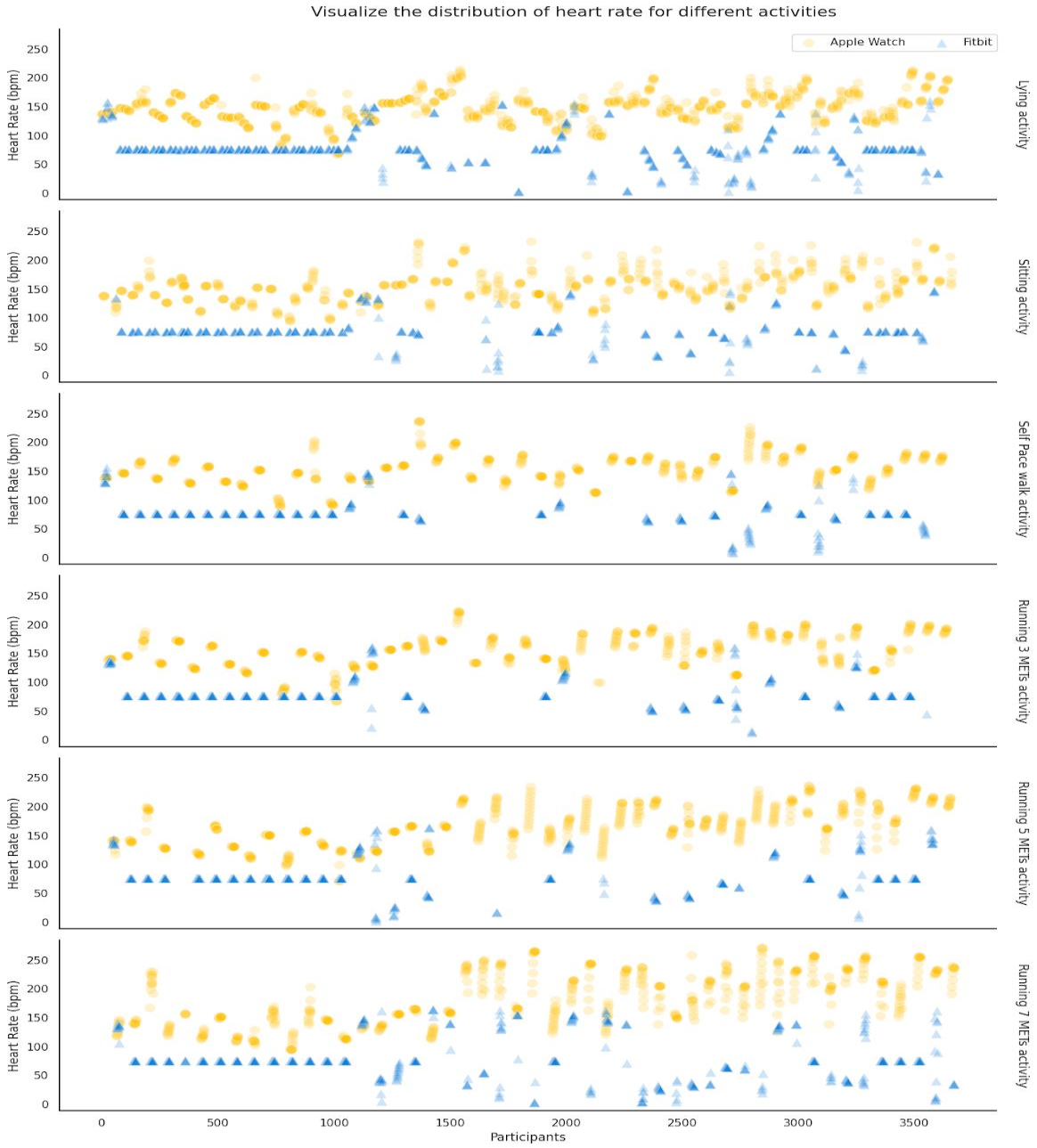


Fig 9: Visualizing the distribution of heart rate across different activities for Apple Watches and Fitbit

We used a 1D scatter plot to see how the data is distributed for Apple and Fitbit watches for each activity. We noticed that the heart rate recorded by the apple watches is always on the higher end, compared to the Fitbit watches. The Fitbit watches failed to capture the variation in individual participant's data as apple watches do. To understand this better we performed a violin plot comparing the two devices.

Violin plot to visualize the distribution of heart rate for apple and Fitbit watches across each activity

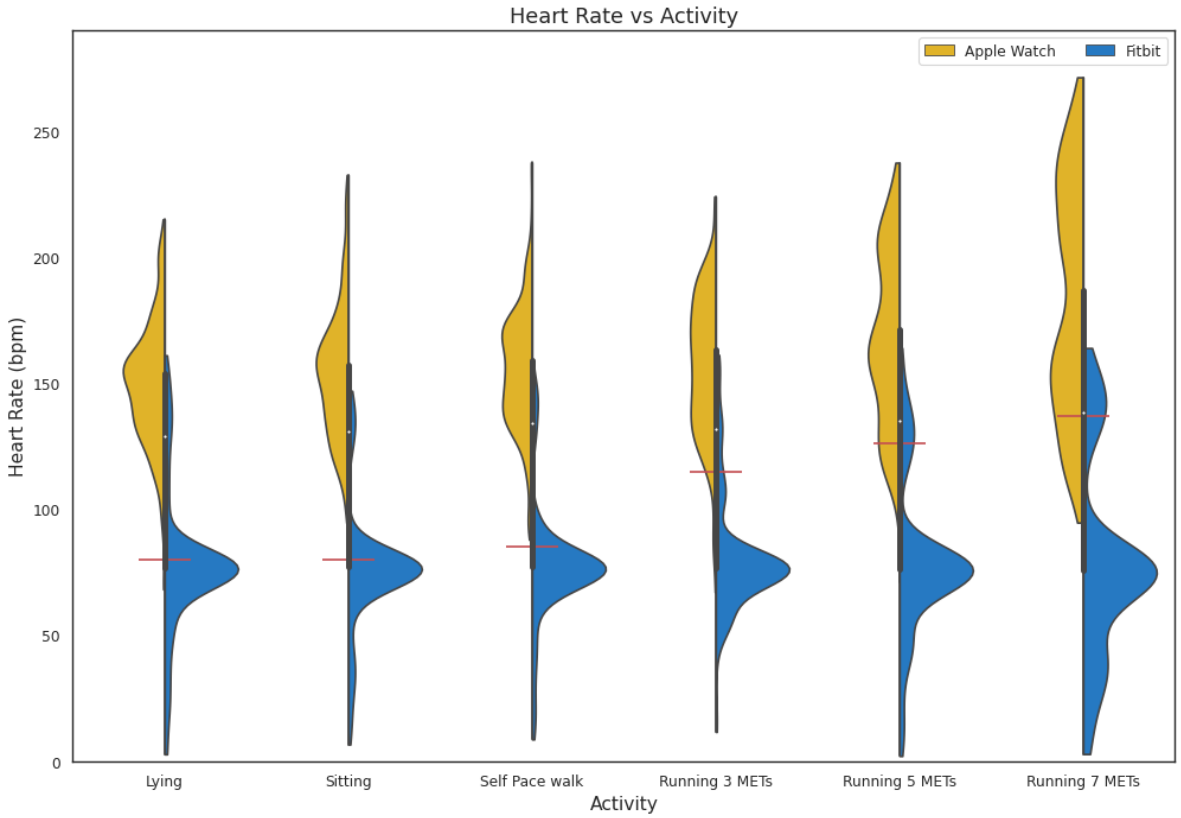


Fig 10: Violin plots for heart rate across each activity

This is a clutter-free visualization in comparison to the above 1D scatter and swarm plots. For apple watch data, we have the same distribution that we noticed in the swarm plot. We can see that lying, sitting, self-paced walking, and 3-meter running have almost the same distribution and a slight change can be observed for 5 and 7-meter running both occupying wider ranges from 0 to around 150. From this and the above visualizations, we can see that it records the heart rate to be around 0 for a few participants. Based on our research we found that the average heart rate for a normal human during Lying, and sitting is 80 bpm, and Self-pace walking is 85 bpm. We also found that for moderate activity the heat rate varies from 64% to 76% of the maximum heart rate, so we calculated the heart rate for 3, 5 and 7 meter running as 64%, 70%, and 76% of the maximum heart rate respectively. [13] These findings are marked as red lines in the above plot to represent an average value for each activity.

Violin plot to visualize the distribution of calories burnt for apple and Fitbit watches across each activity

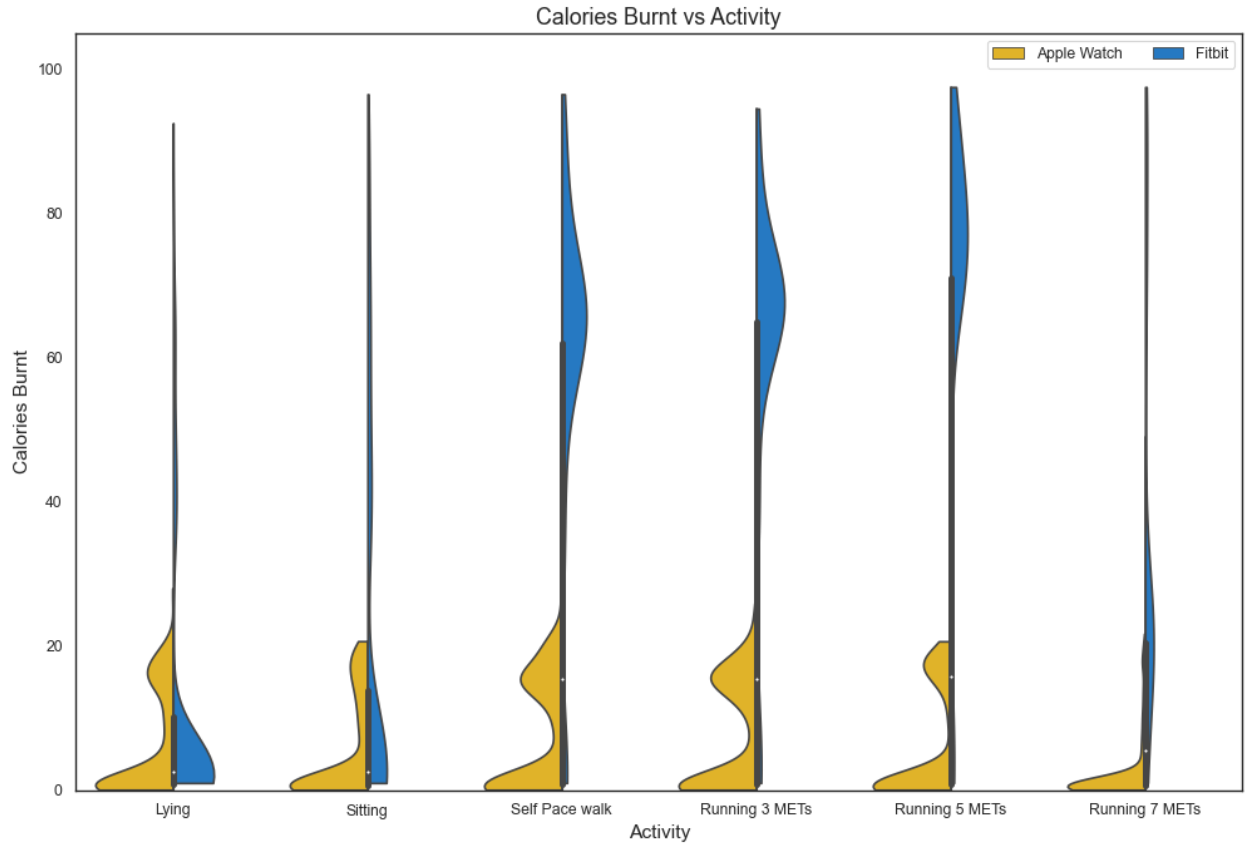


Fig 11: Violin plot for activity across calories burnt

For the calories burnt, data collected from Apple Watches falls below 25 cals for all the activities. It can be seen that lying and sitting burn more calories when compared to 7-meter running. A similar observation is noticed for the data collected from Fitbit where self-paced walking, 3 and 5-meter running burns more calories when compared to 7 meter running which has a similar distribution to lying and sitting.

Based on both observations we can conclude that either there was a miscalculation or error while recording the calories burnt or the experiment should have been conducted where people ran for more than 10 minutes, or people failed to participate in the 7-meter running as expected.

Is there a significant difference in the calories burnt when comparing 3-meter and 7-meter running?

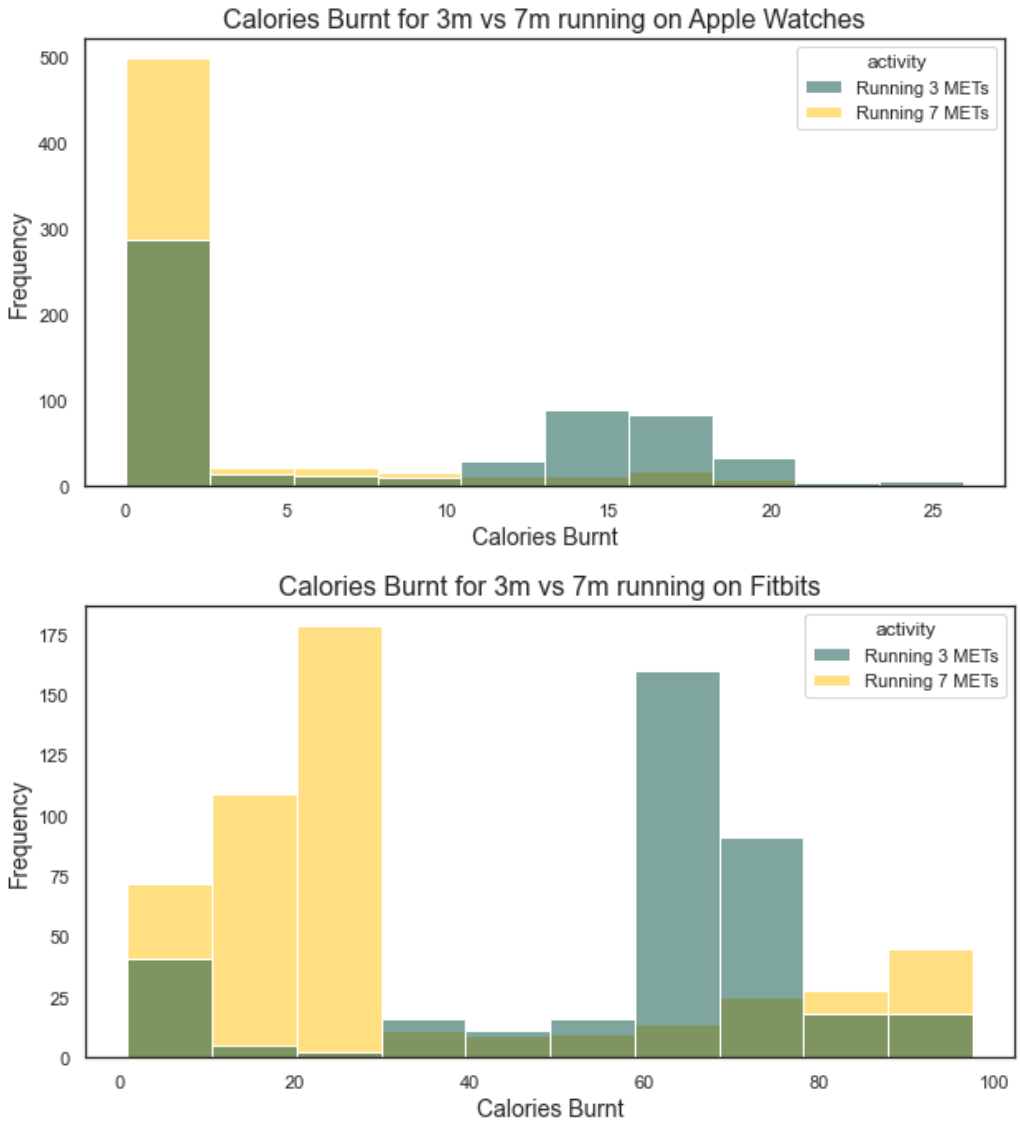


Fig 12: Histogram for calories burnt for 3- and 7-meter running

We expected the calorie burnt for the 7m running to be higher than that of 3m running. But from the above histogram, we can see that the calories burnt in the 3m running activity occupies a wider distribution than 7m running. We notice a lot of observations for 3m running range in the higher calorie burn zone like 15-20 Cals for Apple Watch and 40 - 70 Cals Fitbits.

The 7m running has a positive skew, having more observations in the 0 Cal zone. A lot of observations for the 7m running values occupy the low-calorie burn zone having values ranging between 0-3 Cal for Apple Watches and 0-30 Cal for Fitbits.

Visualizing the steps recorded for each activity

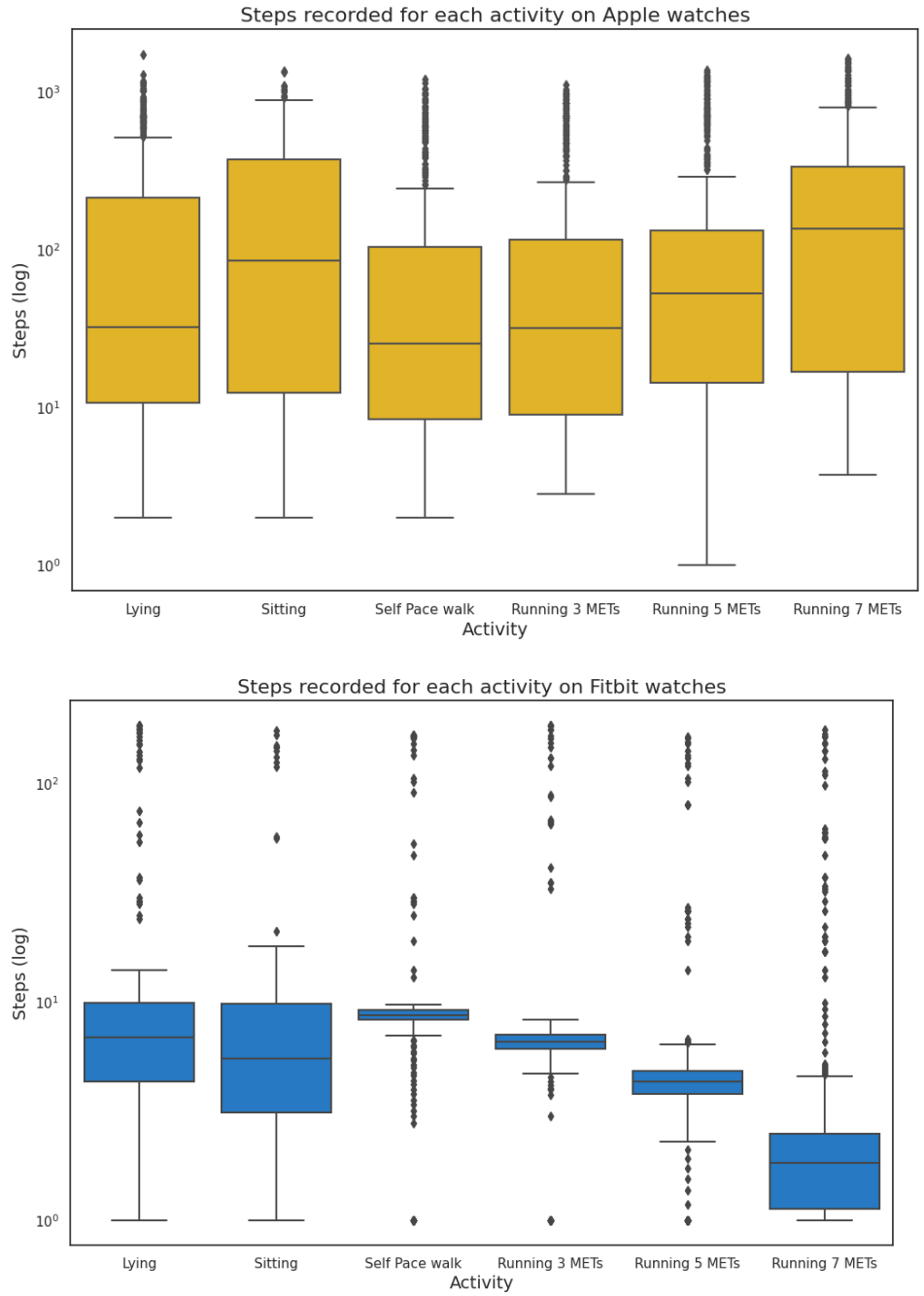


Fig 13: Box plots for steps taken for each activity

From the given box plots we can see that apple watches perform better in capturing the number of steps taken for 3m, 5m, and 7m running when compared to the Fitbit data, as we see a trend of the median and the range of steps gradually increase as the activity progresses from 3m to 7m running.

For the Fitbits, we see that as the activity becomes more extensive, that is from a 3m run to a 7m run, the number of steps taken decreases, which does not make sense. We expect the 7m run to have more steps taken than a 3m running. But the Fitbit data does not reflect this expectation.

For lying and sitting activities, both the watches perform poorly by recording that there are steps taken during these stationary activities, which is unlikely.

Visually explore the difference among various categories such as Male vs Female, different Age Groups and different BMI classes.

We now explore if attributes of the participants such as their gender (Male/ Female), their age group and their BMI category (Underweight/Normal/Overweight/Obese), has an influence on the Calorie Burn or Heart Rate across the 5 activities.

Age Grouping

We created a categorical variable, age group, from the continuous variable - age. We selected age range in intervals of 8 years, starting from 18 which is the youngest participant age that we have in our dataset. The chosen age groups are '18-26', '26-34', '34-42', '42-50', '50-56', The last bucket contains only 6 years as opposed to the 8 years in other buckets as the oldest participant in our data is 56 years old.

```
bins = [18, 26, 34, 42, 50, 56]
labels = ['18-26','26-34','34-42','42-50','50-56']
df['Age_Group'] = pd.cut(df['age'], bins=bins, labels=labels)
```

Fig 14: Python code for calculating age groups

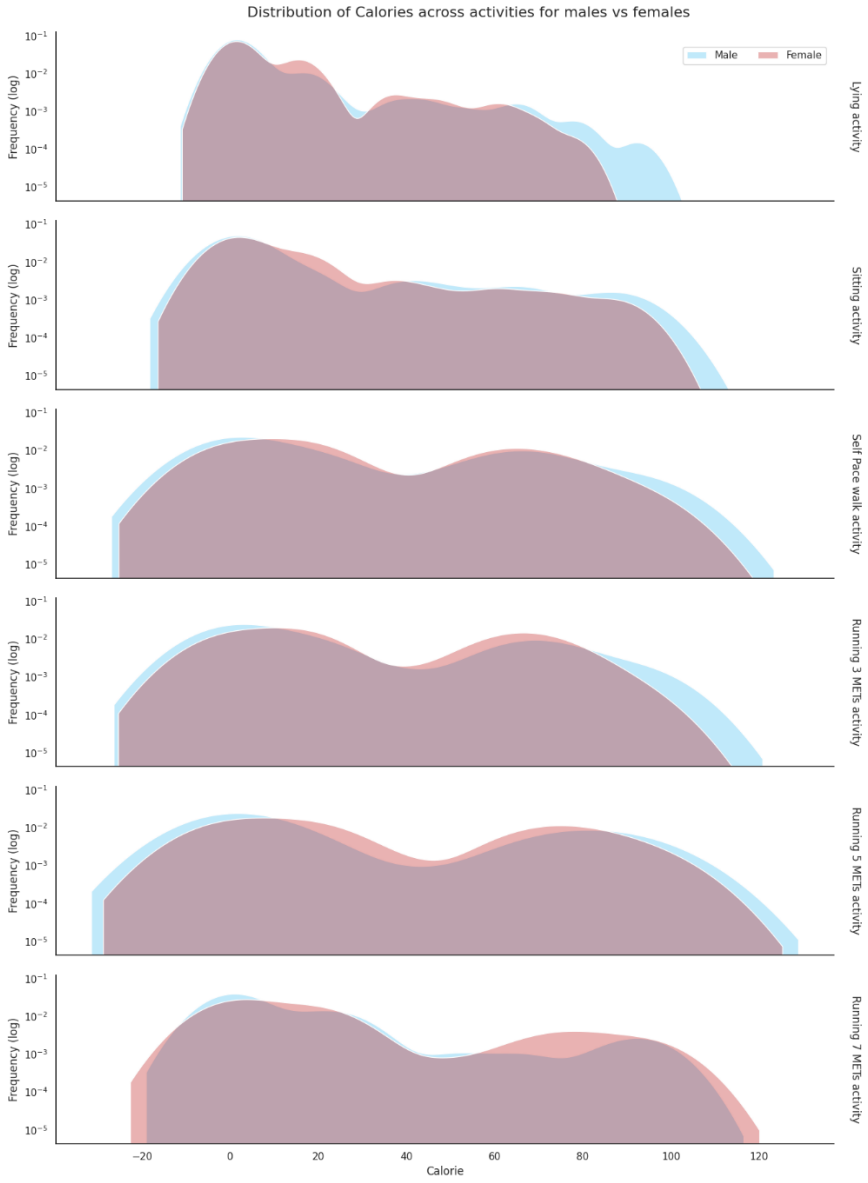
BMI Calculation

We calculated the BMI using the height and weight attribute for each participant using the formula below and based on our research we categorized them into underweight (BMI values fall below 18.5), normal (BMI ranges between 18.5 and 25), overweight (BMI ranges between 25 and 30), obese (BMI value above 30)[14].

```
weight = df['weight']
height = df['height']
df['bmi'] = df['weight']/(df['height']*0.01)**2
df['bmi_group'] = 'Obese'
df.loc[df['bmi'] < 18.5, 'bmi_group'] = 'Underweight'
df.loc[(df['bmi'] >=18.5) & (df['bmi']<25), 'bmi_group'] = 'Normal'
df.loc[(df['bmi'] >=25) & (df['bmi']<30), 'bmi_group'] = 'Overweight'
df['gender_class']=df['gender'].apply(lambda x: 'Female' if x==1 else 'Male')
df.head()
```

Fig 15: Python code for calculating the BMI

Visualize the similarity in heart rate and calories burnt for male and female across each activity



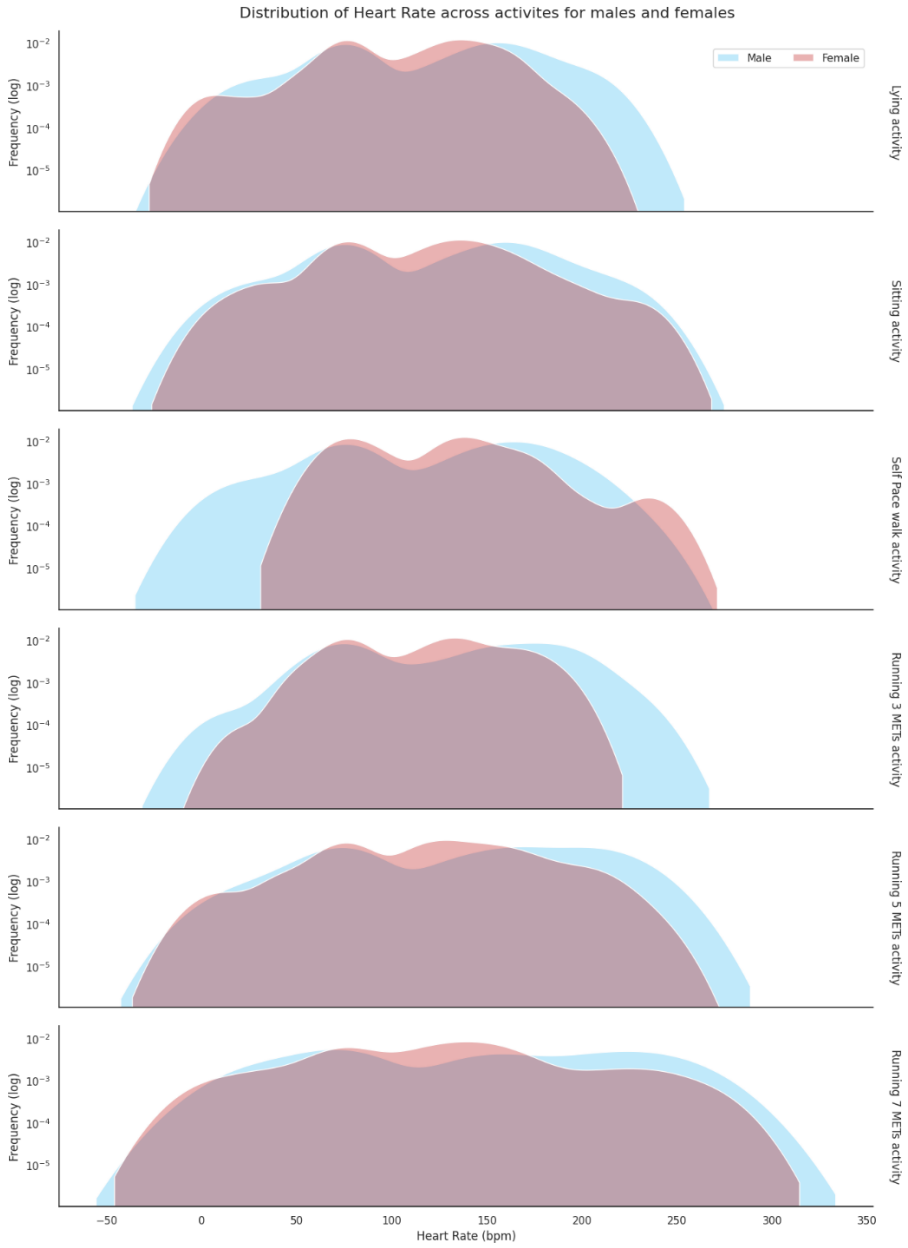


Fig 16: KDE plot for gender and calories and heart rate

We used a KDE plot to visualize the probability density of heart rate and calories burnt for male and female across each activity. We can see that both male and female have almost the same distribution of Calories burnt. For Heart Rate, in self-paced walk, the distribution for male, ranges lower than females and for 3m running, the distribution of male is wider than that of female. To better comprehend this difference, we plotted the below paired box plot.

Box plot to visualize the Heart Rate for each activity for both male and female

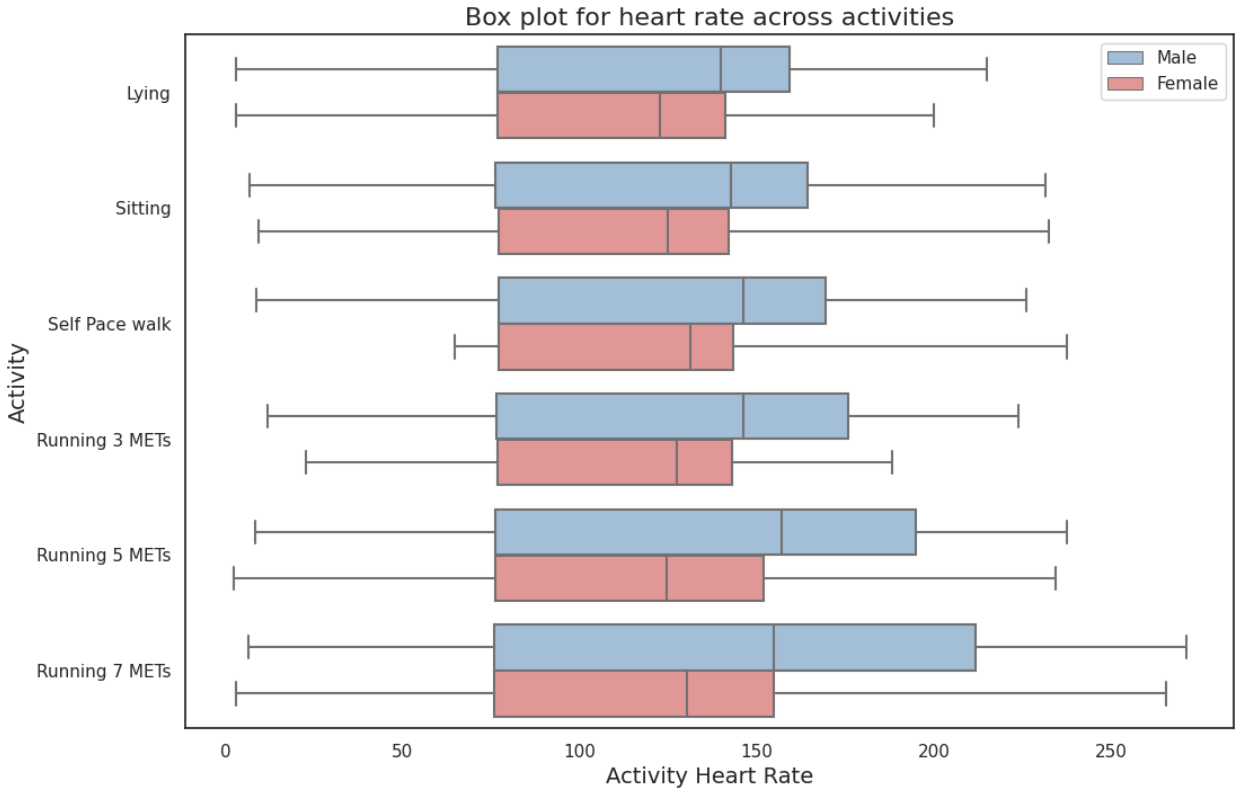


Fig 17: Paired box plot for gender, heartrate and activity

To get a better look at the how heart rate varies for male and female, we used a paired box plot. We can see that the median and the third quartile value of males always falls higher than females across all the activities. This proves the known fact that men have higher heartrate when compared to females. [11]

We tried overlaying the boxplot with a swam plot to visualize the individual data points for males and females to get a better look at how they are distributed, but it created a crowded and cluttered visualization, so we decided not to include it.

Analyze the proportion of male and female for each BMI category

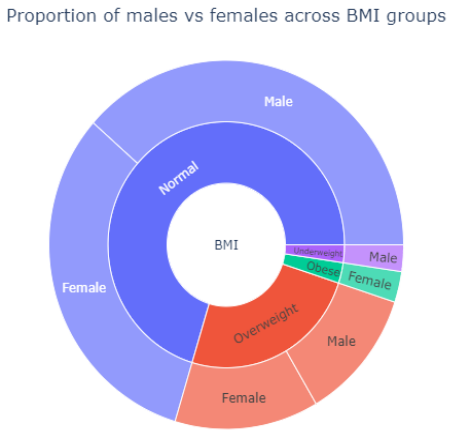


Fig 18: Sunburst plot for BMI and gender

We used a sunburst plot to visualize the proportion of male and female in each of the BMI category. We can see that we have almost equal observation for male and female under the normal and overweight category but has an unbalanced observation for the obese and underweight category. Since it is difficult to interpret the exact values through a donut chart, we corrected our mistake by plotting a stacked bar plot.

Stacked bar plot to analyze the proportion of male and female for each BMI category

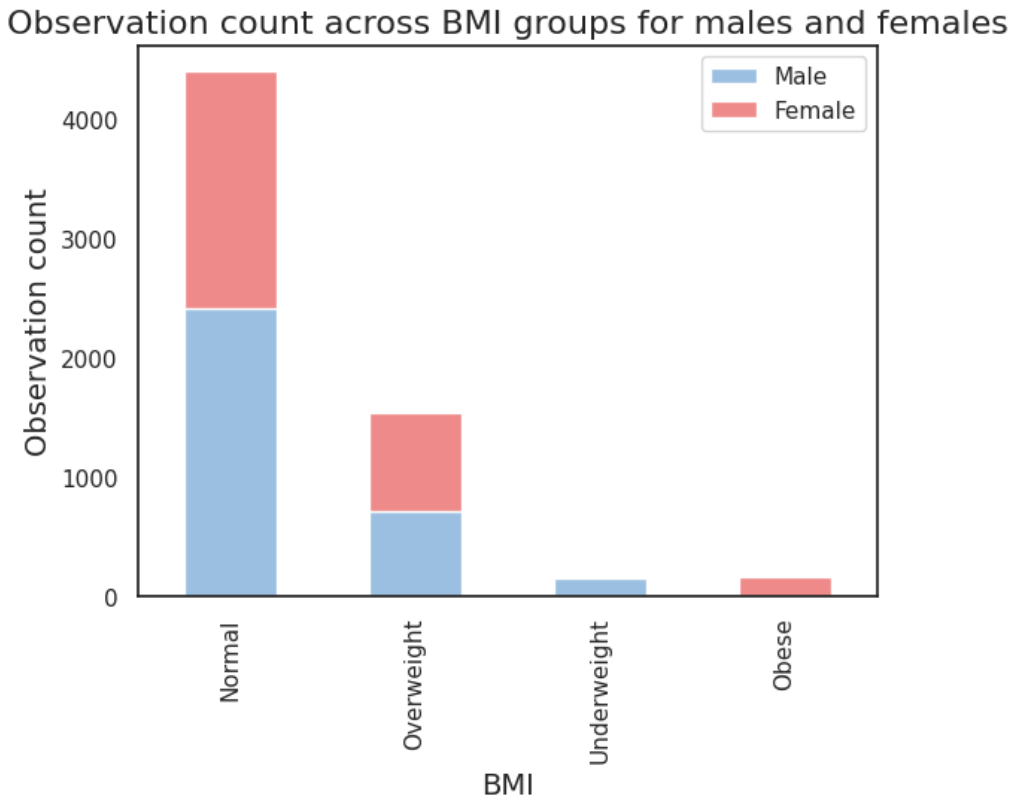


Fig 19: Stacked bar plot for number of observations, gender and BMI

It conveys the same information as the Sunburst chart, but we can better interpret this visualization. There are almost equal number of male and female participants under the normal and overweight BMI category. But it seems that this study did not include any underweight females and overweight males have very low observation under each of these categories.

Visualizing how Median BMI varies across age groups for male and female

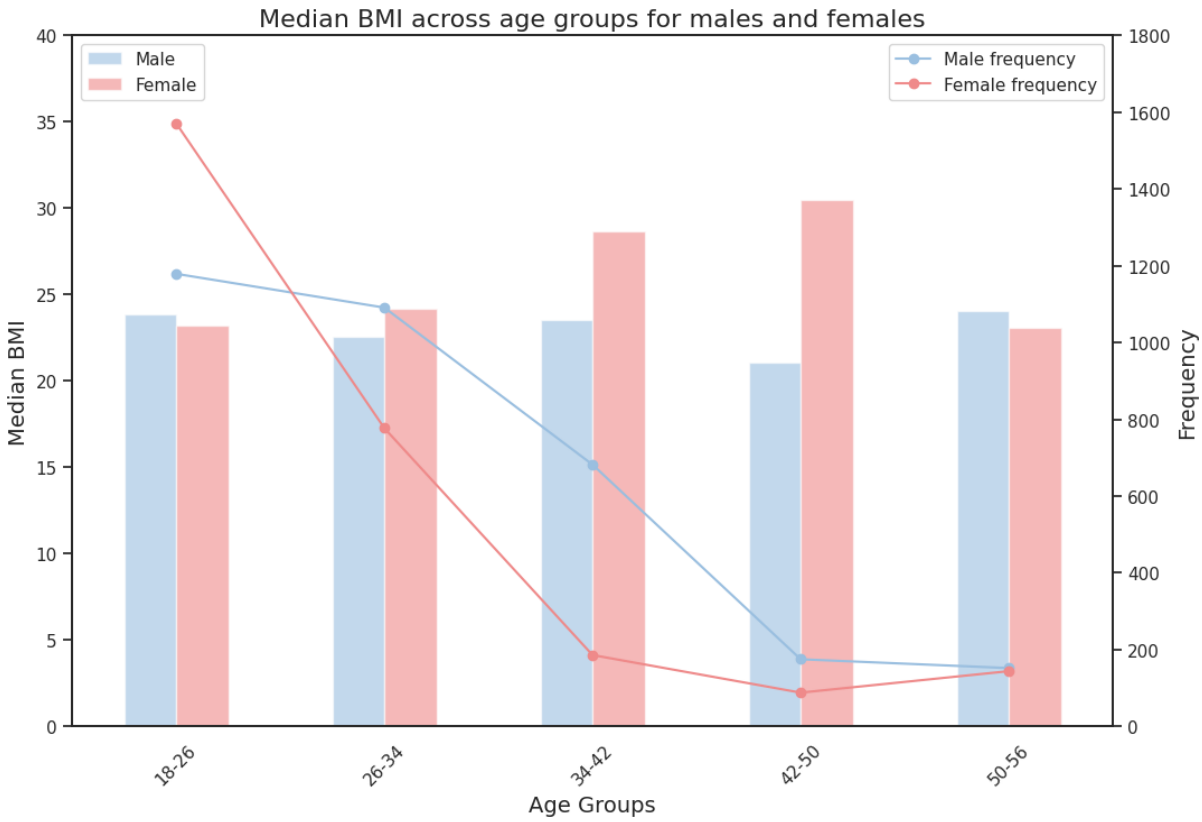


Fig 20: Combination of line and bar chart

The line plots show the number of observations in each age group for male and female. We notice that for higher age groups, we have fewer records. The number of female observations plummets from around 1200 for age group 26-34 to less than 100 for age group 42-50. A same trend follows for male observations.

The bar chart shows the median BMI for each age group for male and female. The female BMI gradually increases with age till 50 yrs. For the males, we have a alternating rise and drop in the median BMI.

We feel that having a balanced dataset with an equal number of observations for each age group could help in obtaining an unbiased and accurate analysis.

Lollipop visualization for median heart rate between male and female

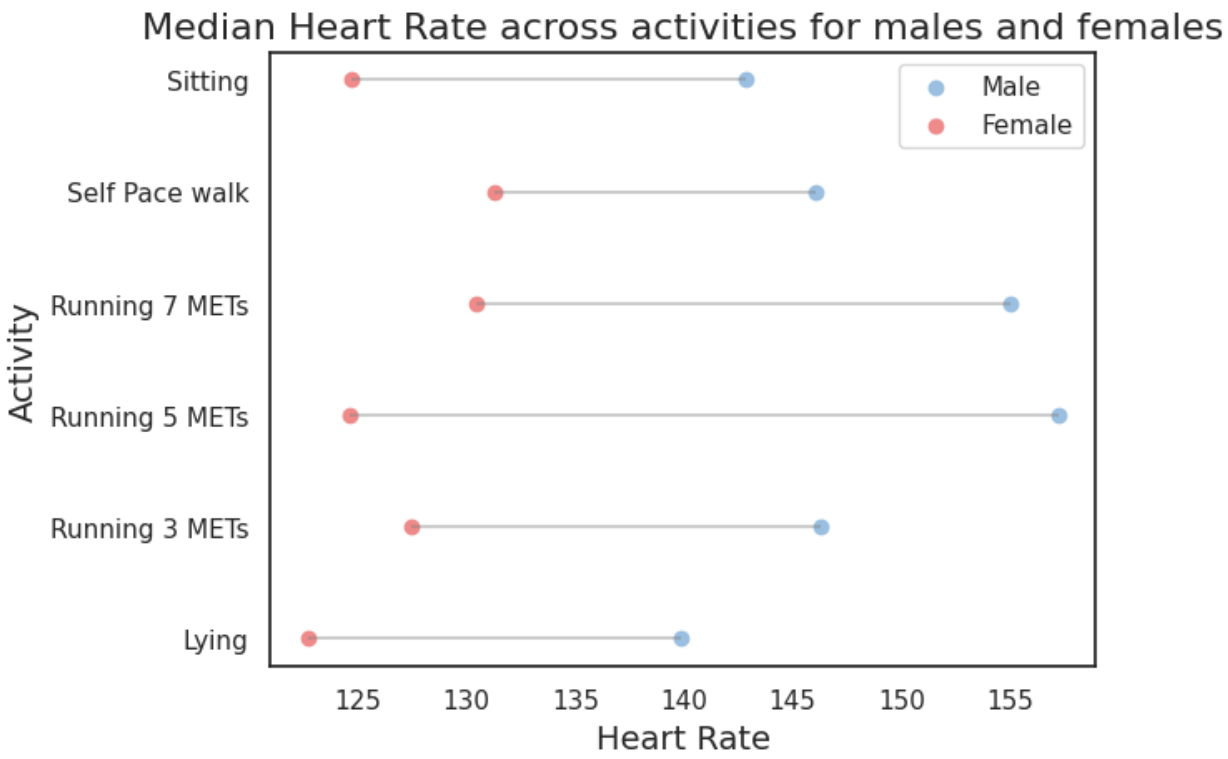


Fig 21: Lollipop chart for identifying the difference in heart rate for each gender

This visualization focuses on the difference in median heart rate between male and female across the activities. We see that the median heart rate of male is always higher than female which agrees which makes sense logically and agrees with the biological research on this subject. As we are focused on the difference in the heart rate values, our x-axis does not start from 0. We see that by far, 5m running shows a wider gap between male and female and self-paced walk shows the least difference.

Interactive Dashboard using Tableau

We created an interactive visualization on Tableau that can be used to view the median heart rate and calorie for different genders and age groups. You can interact with the visualization using the filters for devices and activities. Users can use these filters to drill down to a specific device and a specific activity and analyze the data for male and female or across different age groups. We used a color-blind friendly palette so that our dashboard is accessible to everyone.

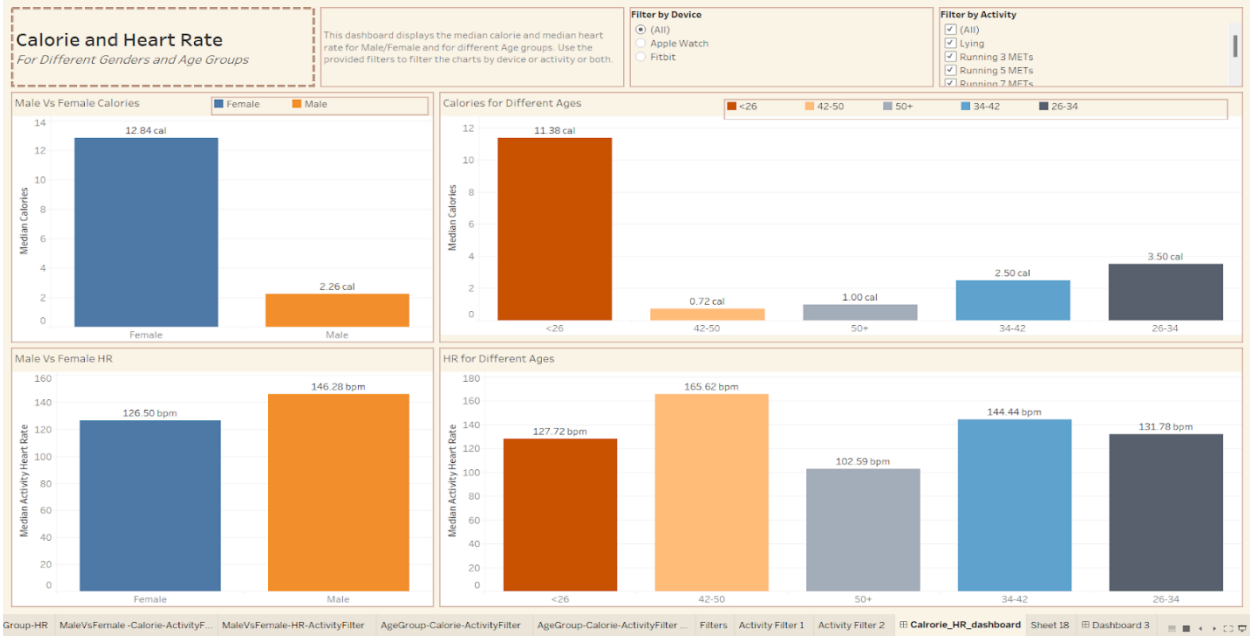


Fig 22: Snipped of the dashboard created in Tableau

Here is a link to the interactive dashboard created using Tableau: -

https://public.tableau.com/views/DV_Project_16695919079250/Calrorie_HR_dashboard?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Color Blind Friendly Palette

Throughout our project, we made sure to use color-blindness-friendly visualizations and tried maintaining a consistent color combination for attributes such as data of Apple with yellow and Fitbit with blue colors, representing female with pink and male with light blue colors and so on. We used the website **davidmathlogic** (https://davidmathlogic.com/colorblind), to check how our colors are being intercepted by people different types of color-blindness. Below we have provided the images of how our colors appear to different color-blind groups. We made sure that the color combinations we choose can be sufficiently differentiated across all groups of color-blindness.

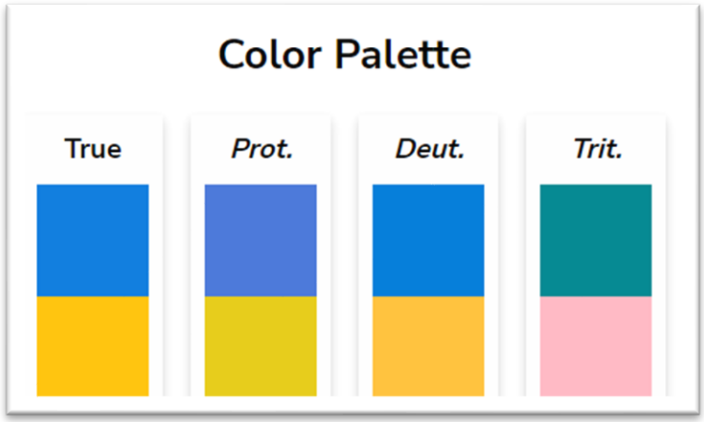


Fig 23: Color blind friendly palette used -1

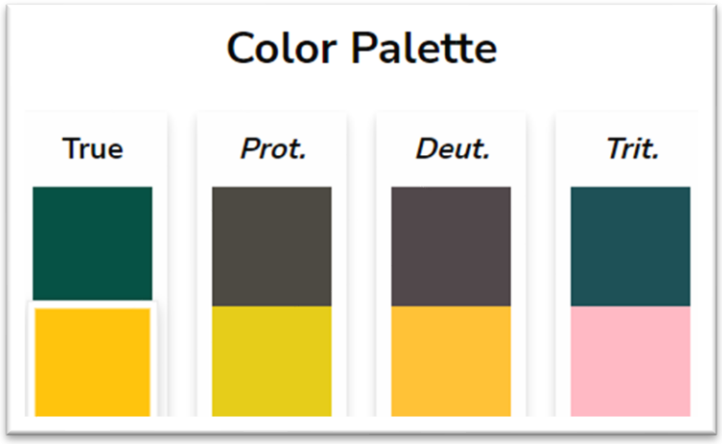


Fig 24: Color blind friendly palette used -2

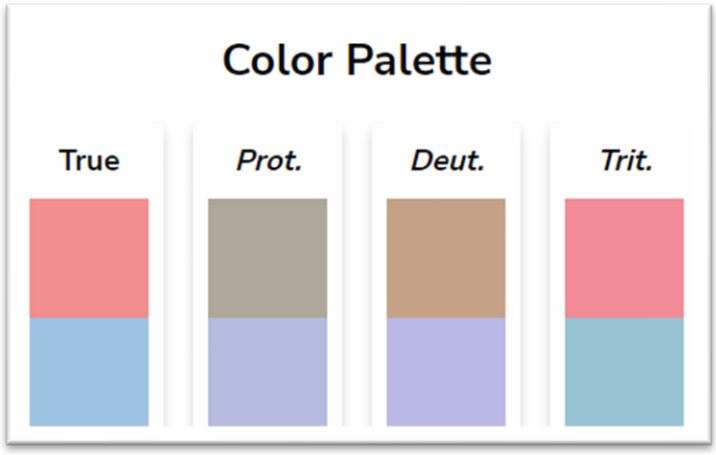


Fig 25: Color blind friendly palette used -3

Conclusion and Future Work

Below are a few interesting insights about the data that has been collected. We noticed that both Apple and Fitbit watches collect the data in different ways. For calories, Apple Watch records the active calories burnt during that activity without including the constant calorie burn into account whereas Fitbit records total calories burnt including the constant calories burnt [1]. But this still doesn't still give us a valid explanation of how the calories burnt follow the same distribution for few activities.

We can see that Apple watches are able to capture the heart rate better when compared to Fitbit. Because Apple Watches capture minute differences in the heart rate among participants as can be seen in the swarm plots, unlike Fitbit that has the same value for most of the participants.

We assumed that 7m running burns more calories when compared to 3m running, but the visualization tells a different story. This may be a result of either a wrong method of data collection or the participants did a poor job during the 7m running. We believe both Apple and Fitbit watches perform poorly in capturing the calories burnt as calories burnt for 7m shows lower than lying and sitting with is unlikely.

We could investigate why both Apple watch and Fitbit record that there are steps taken by a participant when they are either lying or sitting as a future work. Our visualization shows that the median heart rate for males is always higher than females irrespective of the activity which is proven through research.

[11] We also suggest using a balanced dataset with an equal number of observations for each age group in order to perform a more accurate analysis.

Reference

1. https://assets.researchsquare.com/files/rs-17022/v1/d5923374-d56c-4fe7-a036-949ecf41917e.pdf?c=1631831698
2. https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/ZS2Z2J/SEZCTK&version=1.0
3. https://towardsdatascience.com/data-analysis-of-your-applewatch-workouts-672fe0366e7c
4. https://www.kaggle.com/datasets/aleespinosa/apple-watch-and-fitbit-data/code
5. https://www.kaggle.com/code/aleespinosa/google-capstone-project-in-r-bellabeat
6. https://www.kaggle.com/code/eigenvalue42/fitbit-vs-apple-watch-classifying-activity
7. https://www.kaggle.com/code/aleespinosa/calories-regression-in-data-from-apple-watch-user
8. https://www.kaggle.com/code/kallelmedanis/how-to-predict-the-activity
9. https://medium.com/@evannwu_15820/fitbit-health-data-dashboard-3ee0da3c975c
10. https://www.valuepenguin.com/fitness-tracker-smartwatch-health-survey#:~:text=The%20vast%20majority%20(92%25),achievements%20cited%20by%20smartwatch%20users.
11. Saleem S, Hussain MM, Majeed SM, Khan MA. Gender differences of heart rate variability in healthy volunteers. J Pak Med Assoc. 2012 May;62(5):422-5. PMID: 22755301.
12. https://davidmathlogic.com/colorblind/#%23D81B60-%231E88E5-%23FFC107-%23004D40
13. https://www.webmd.com/fitness-exercise/what-you-need-to-know-about-running-heart-rate-zones
14. https://www.cdc.gov/obesity/basics/adult-defining.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obesity%20range.