# Predictive analysis of health care devices based on physical activity

*Manusha Gadde, Anusha Perla, Sai Anmisha Doddamreddy, Sai Shivani Chirag*

## Introduction

The evolution of technology has made it possible for users to collect multifaceted information about their lifestyles and fitness routine. Wearable fitness and health monitoring gadgets have grown in popularity and have become one of the fastest-growing segments of the smart device market. Increasingly, firms are offering fitness tracker users integrated health and activity monitoring solutions. Recently, insurance companies have improved their consumers' access to health and condition monitoring. However, the massive amount of sensitive data acquired by monitoring products and its accessibility by third-party service providers places critical security and privacy constraints on the solutions deployed (Fereidooni et al.,2017).

Fitness applications for smartphones have grown highly popular among fitness fanatics and the public. The data collected by these apps have been shown to be beneficial in managing daily fitness and tracking progress toward health goals via workout summaries and insights. However, these insights are typically confined to weekly or monthly summaries of the user's activity (Bhargava & Nabi,2020). The study employs both graphical and statistical methods to analyze the fitness dataset to understand the routine and progress of the participants.

## Objectives of the study

To identify any trends in smart device usage and to understand how these trends apply to Bellabeat customers and influence Bellabeat marketing strategy.

## Methodology

These data sets were generated by 30 respondents' own fitness trackers between December 3 and December 5, 2016, as part of a distributed survey conducted through Amazon Mechanical Turk. Data about users' daily activity, steps is gathered, and this information can be used to explore and evaluate their actions. It collects data on users' daily activity, steps which can be utilized to investigate and analyze their behaviors. The data was obtained from Kaggle using the link; https://www.kaggle.com/datasets/stanley888cy/google-project-02

Various statistical methods are employed in R to investigate the relationship between the fitness data variables. The data sets were evaluated for null values, and then cleaned data by removing the null values. We have adjusted the date into a united date format. Dropped the "distance" variable as the "minutes" (time period) and "steps" are more representative of activities duration.

Adjusted the activity hour a united datetime format in Hour_Calorie and Hour_Step files. Later data frame was made by combining Hour_Calorie and Hour_Step with an inner-join with "Id" and "datetime" columns. Added in_date, in_weekday, in_hour identification to each data row to check the trends. The study focuses only on the variables that are relevant to our research. Visuals such as histograms, bar graphs, and potentially box plots for the data are generated in R software. To get a better view of patterns in health tracker device usage over time.

The study considers the number of exercises by year as one way of understanding the patterns. A linear regression model that best predicts outcomes using explanatory factors, an elimination strategy was utilized. The model fitting procedure was done recursively to attain the model with the best explanatory ability of the outcome variable. The linear regression model with the highest adjusted R^2 value was considered. This method is useful for determining whether the various variables have any significant correlations. Finally, we want to focus on the association between user activity type and calories consumed.

In conducting the study, the following assumptions were made about the variables;

1. "Distance" will be interpreted as the meter length detected by the device's GPS system, not as human input.

2. "Time" shall be taken to refer to the minute period during which the user is wearing the gadget. Each record day, it is believed that users will wear the gadget for 24 hours.

3. The term "steps" refers to the number of steps taken while walking. It refers to the degree of difficulty associated with activities or exercises, in the study's view. As a result, the exercise will be more active, resulting in a higher step count.

4. The term "calories" refers to the amount of energy consumed during a specified time period.

5. The terms "Very active," "Fairly active," "Lightly active," and "Sedentary" are used to categorize activity according to exercise intensity. The degree of activity increases in intensity from "Sedentary" to "Lightly active" to "Fairly active" to "Very active".

6. Each user is chosen at random from the population.

## Normality

The calories used under various levels of activities does not appear to follow the normal distribution due to the lack of symmetry. Most of the fitness devices users are only active for just less than 21.16 minutes per day. The average amount of time an individual is very active is 21.16 minutes. There are a few individuals who are very active for over 100 minutes a day and most of them for less than 5 minutes. The histogram of the distribution of sedentary minutes indicates that most individuals have set their active minutes to a bare minimum and are inactive for as many as over 991.21 minutes a day.

The normality assumptions for the calorie residuals were checked using the quantile plots; there is close conformity to the normal distribution of residuals as shown by the Q-Q norm plots since residuals are consistent with the predicted quantiles line.

## Results and discussions

The data analysis was concentrated on physical activity. Exercise activity investigates the association between type of exercise and calorie consumption. Additionally, this study makes use of the linear regression model and statistical model to analyze and predict the data quantitatively.

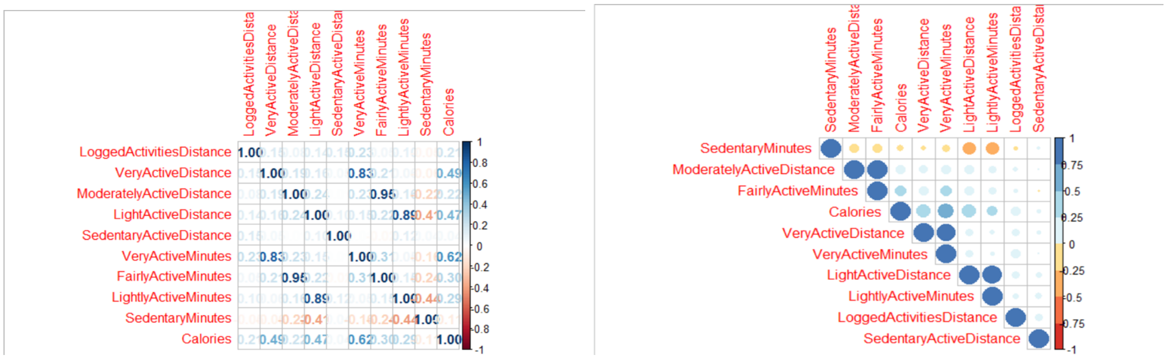
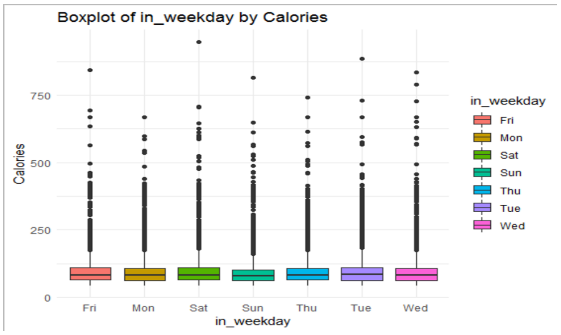
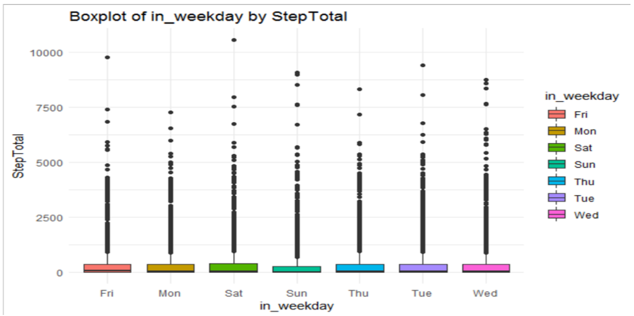


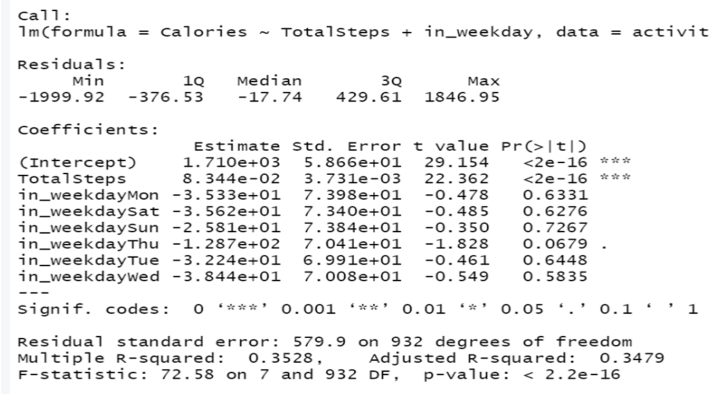
Figure 1: Pearson's correlation coefficients

There is a high positive correlation between calories used and very active minutes. An increase in sedentary minutes is associated with decreasing calorie usage.





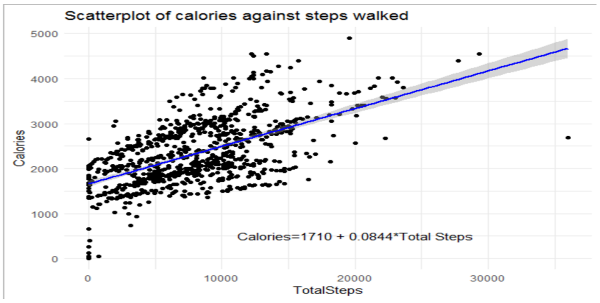
The distribution of calories against weekdays does not indicate significant variation in median hourly calories with respect to the day of the week. Therefore, calorie variation is independent of the day of the week.The median and the upper quartiles for the number of steps walked on Sunday appear to be fairly lower than that for the rest of the days. People are definitely waking less on Sundays than they do any other day of the week.

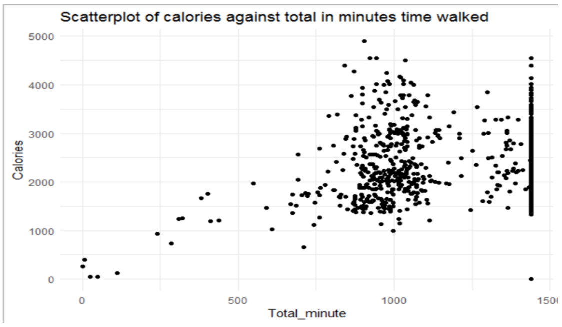


```
Call:
lm(formula = Calories ~ TotalSteps + in_weekday, data = activity_cleaned)

Residuals:
     Min       1Q   Median       3Q      Max 
-1999.92  -376.53   -17.74   429.61  1846.95 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.710e+03  5.866e+01  29.154   <2e-16 ***
TotalSteps      8.344e-02  3.731e-03  22.362   <2e-16 ***
in_weekdayMon  -3.533e+01  7.398e+01  -0.478   0.6331    
in_weekdaySat  -3.562e+01  7.340e+01  -0.485   0.6276    
in_weekdaySun  -2.581e+01  7.384e+01  -0.350   0.7267    
in_weekdayThu  -1.287e+02  7.041e+01  -1.828   0.0679 .  
in_weekdayTue  -3.224e+01  6.991e+01  -0.461   0.6448    
in_weekdayWed  -3.844e+01  7.008e+01  -0.549   0.5835    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 579.9 on 932 degrees of freedom
Multiple R-squared:  0.3528,	Adjusted R-squared:  0.3479 
F-statistic: 72.58 on 7 and 932 DF,  p-value: < 2.2e-16
```

The linear regression model was fitted to present the relationship between the calories used, day of the week and the total steps taken while walking. The day of the week is inconsequential in determining the calorific usage.  As the number of steps taken while walking increases by one, the calorie usage is predicted to increase by a factor of 0.0834. The parameters estimate of 0.0835, p=0.000<0.05 confirms the significance of the number of steps walked on the calory consumption. The p-value is quite small for the linear regression line. As a result, the total number of steps taken is significantly impactful to the number of calories ingested. A plot of this linear regression line on a graph, would be obtained using the formula;

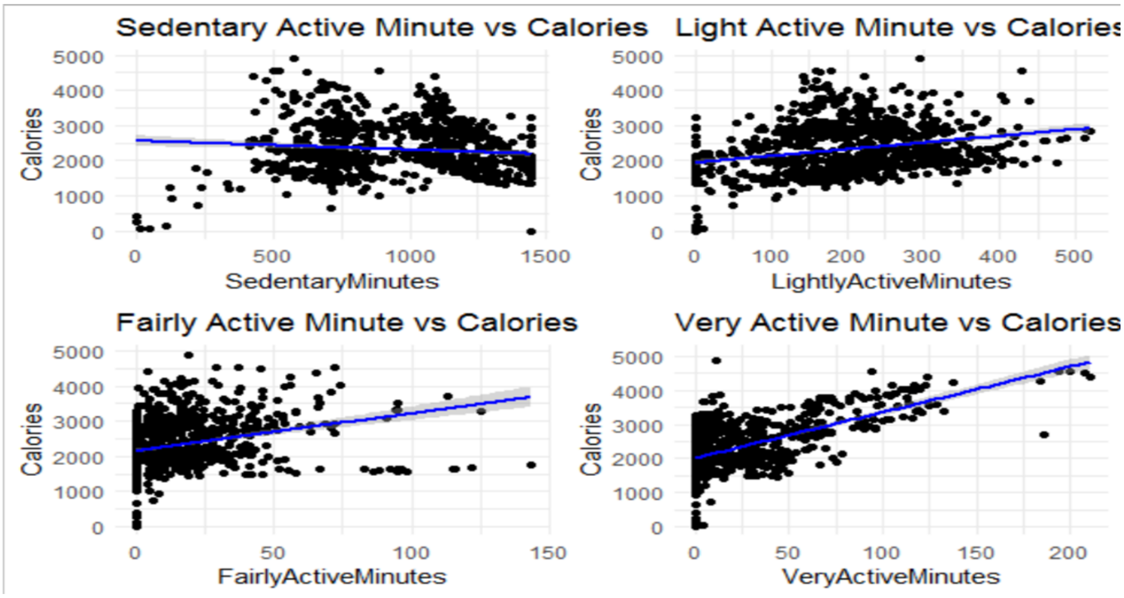
**Calories=1710+ 0.0834 * Total Steps.**





With a slope coefficient of 0.0844, each step can burn around 0.0844 calories.  The scatter plot further underscores the nature and the strength of the relationship between the two variables.

According to the graph, there is no discernible relationship between minutes and calories. This is due to the fact that there are numerous types of activity. As a result, the total minute variable cannot be utilized for correlation wholesomely. Thus, the data are categorized as "Very Active," "Fairly Active," "Lightly Active," and "Sedentary," and a plot is generated for each category.



Isolating total time in terms of activity intensity results in a graph with distinct slopes for regression lines. This is completely sensible, given different degrees of activity and burns varying amounts of calories. In theory, more active people should consume more energy than those who are more sedentary. As a result, a multivariable regression on the relationship between each type of activity and calories can be derived.

Sedentary minutes and total steps have a strong negative correlation, this further explains why an increase in sedentary minutes is associated with low calorie consumption.

**Analysis of variance for the calorie consumption against level of activity**

One-way anova, paired t test analysis were performed. All the different levels of activity have a significant influence on the calory consumption among the fitness devices users. As can be seen from the outcome of anova and t test , both the R-squared and factor p-values are quite small. This demonstrates that, despite the great degree of variability in the data, there is a tendency that all of those activities are associated with calories burned.

**Calories=1241 + 12.95 VeryActiveMinutes+ 3.676 FairlyActiveMinutes+ 2.01 LightlyActiveMinutes+ 0.3539 SedentaryMinutes**

Taking a look at the formula; each minute, the Very Active burn 12.95 calories. This is 3.5 times the amount of time spent Fairly Active, 6 times the amount spent Lightly Active, and 36.5 times the amount spent Sedentary.

**Data visualization**

The line charts showing average steps and activity minutes over a week appear to lack any discernible trend or pattern, with the exception of a few daily fluctuations. And, unexpectedly, people do not appear to be exercising more on Sunday. According to the data, Sunday has the fewest average steps and the second-fewest average activity minutes of the entire week.

The distribution of average calories per hour and the average steps per hour confirms that indeed the two variables are functions of time and that they assume a similar trend. The average calory per hour trend is in almost perfect synchrony with the average steps per hour over the entire time of observation. Their patterns are strikingly similar. Again, this demonstrates that the number of steps taken correlates with the number of calories burned. According to the graphical picture above, users steadily increase their activities from morning to evening. Additionally, the study indicated that people are most active between 12:00 and 20:00, when their calories burned, and the steps taken are much higher.

## Conclusion

The assessment of an individual's or a group's long-term health and wellness using longitudinal fitness data is still in its infancy. This type of study may be beneficial for providing tailored health recommendations for an individual and may contribute to the development of healthier communities (Guo et al.,2017). The study investigated the relationship between calory consumption, level of activity as well as intensity and number of steps walked every day. There was a positive association between calory consumption and the number of steps walked. The high-intensity activity was associated with the highest extent of calory consumption. The days of the week had a similar calory consumption and lower calorie consumption. The following specific deductions can be made from the fitness device study; The quantity of calories consumed is positively connected to the number of steps taken by users. Calorie intake is also related to the type of activity. The more vigorous the activity, the more calories burned. On Sundays, users often do not take more steps or engage in more active exercise. Rather than that, they'll take fewer steps and be less active on Sunday. The types of activity will not vary over the course of a week. There is no evidence that consumers' fitness habits will change. People are most active and burn the most calories throughout the afternoon and evening hours of the day.

## References

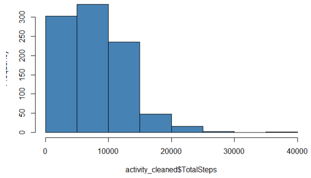
Bhargava, Y., & Nabi, J. (2020). The opportunities, challenges and obligations of Fitness Data Analytics. *Procedia Computer Science*, *167*, 1354-1362.

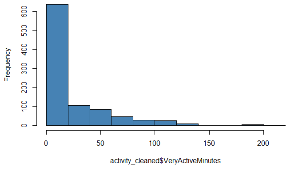
Fereidooni, H., Frassetto, T., Miettinen, M., Sadeghi, A. R., & Conti, M. (2017, July). Fitness trackers: fit for health but unfit for security and privacy. In *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)* (pp. 19-24). IEEE.

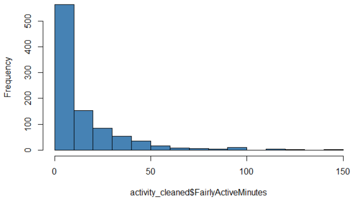
Guo, L., Sharma, R., Yin, L., Lu, R., & Rong, K. (2017). Automated competitor analysis using big data analytics: Evidence from the fitness mobile app business. *Business Process Management Journal*.

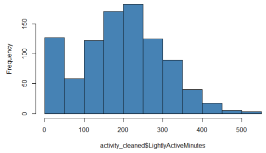
Data source: Link- https://www.kaggle.com/datasets/stanley888cy/google-project-02

# APPENDIX

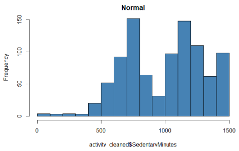




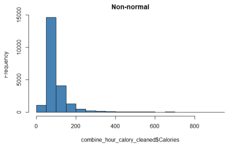




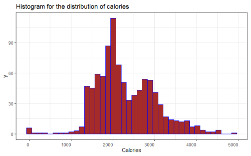
Normal



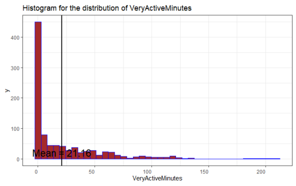
Non-normal



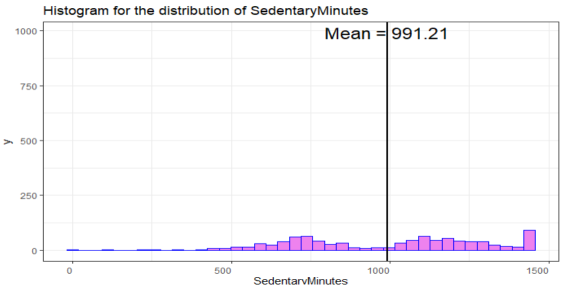
Histogram for the distribution of calories



Histogram for the distribution of VeryActiveMinutes

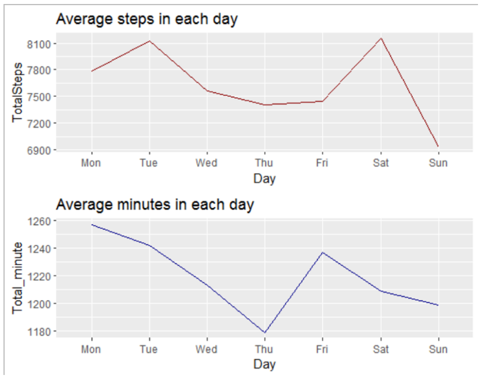


Histogram for the distribution of SedentaryMinutes



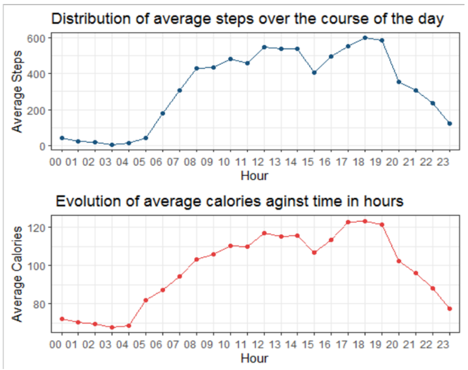
Average steps in each day

Average minutes in each day



Distribution of average steps over the course of the day

Evolution of average calories aginst time in hours


Average Steps
Hour
Average Calories
Hour