Hemalatha_Subbiah_Project_Step1

Hemalatha Subbiah

2023-02-12

R Markdown

Introduction

An activity tracker is a type of electronic device that helps monitor some type of human activity, such as walking or running, sleep quality or heart rate.Better's research shows that almost three-quarters of people who wear fitness trackers do so to monitor their progress, while 62% of wearers use them to increase their motivation to exercise. Another 46% want to understand their body better by tracking things like their heart rate, steps taken and calories burned.Activity trackers are devices that translate movement into different forms of data. Most trackers will provide estimates of steps, distance, and active minutes.

Problem Statement

This data science project aims to help data scientists develop an intelligent model for how can your own personal analysis data assist you in living a better life? To solve this project related to data science, the popular Kaggle dataset containing activity tracker transaction made in September 2016 by individuals.This Kaggle data set contains personal fitness tracker from thirty activity tracker users. The dataset contains 18 .csv files, which is used are about activity, calories, intensity, steps, and sleep time. Activity tracker collect continuous physiological measurement and are generating gigabytes of data every single minute. Fitbit reports that they have over 150 billion hours of heart rate recorded, and over 6 billion recorded nights of human sleep. While this data is extremely useful for gathering information at the population-level, how can your own personal analysis assist you in living a better life? Do activity trackers really help to better your health?All the information exists at your fingertips (or on your wrist), and we can make it actionable

Research questions

1. Do people have Awareness to relate the data to personal health ?
2. What are some trends in smart device usage?
3. How could these trends apply to activity trackers customers?
4. How could these trends help influence good health Business task?
5. Identify potential opportunities for growth and recommendations for the Bellabeat marketing strategy improvement based on trends in smart device usage.
6. Is it focused own women from all countries?
7. IS tracking physical activity, mental state, menstrual cycle helps them to improve health better?
8. Is it possible to collect a large amount of data about personal activity relatively inexpensively?
9. Do people have Awareness to relate the data to personal health ?

Approach

Business understanding Generate Your Hypotheses Study the data Clean the data Engineer the features Model Fitting Making Prediction

How your approach addresses (fully or partially) the problem

Business understanding : In business understand phase we basically Understands the business process,Define and Frame the business problem,define the business objective and agree on success criteria.In my project how the activity trackers really help to better your health and the business task is to identify themes/trends in how people currently use their smart devices and relate to their own health.

Data understanding : Understand data touch points in the context of business process and gather knowledge on where data originates from, how it gets processed, what decisions are being made, where it is getting stored and how it flows to downstream.Deep dive into business meaning of the data being leveraged as well as knowledge present in existing system in form of rules. For this project this dataset would have been more reliable, original, current, and cited; albeit data privacy would have to be carefully guarded.However, since this is a hypothetical scenario and this is the only dataset available, I'll make do.

Data preparation and Cleaning :Good data hygiene is so important for business. For starters, it's good practice to keep on top of your data, ensuring that it's accurate and up-to-date.As part of data cleaning I want to a) Get rid of unwanted observations b) Fix structural errors - Removed inconsistent capitalization, which often occur during manual data entry c) Remove unwanted outliers- Removed few outliners in the data d) Fix contradictory data errors e) Type conversion and syntax errors f) Validate your dataset for null values and condition them. All the above steps will be completed as part of data sets I picked for this project.

Modeling :Build predictive model variables and do feature engineering and fit an closest model to the problem solution.

Validation : Validating the model by training the model. Deployment : The concept of deployment in data science refers to the application of a model for prediction using a new data. Building a model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data science process.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

The dataset used for this analysis is FitBit Fitness Tracker Data hosted on Kaggle or Zenodo. It comprises .csv files of various fitness metrics measured from different users at different times, stored in a wide format. The fitness tracker data was provided by 30 respondents to a paid distributed survey on Amazon Mechanical Turk in 2016.

Limitations of DataSet: Data was collected in 2016, hence data may not be relevant to modern trends. Small sample size of only 30 participants. Data does not include demographics about the sample such as sex, age, or geographical location. This may not be a good representation of the population of women globally who would use a similar product.Survey style of data collection may be subject to response bias. Integrity and accuracy of data is not clear.

Inital observations of these CSVs within Mircosoft Excel shows that these files contain acitvites, calory records, physical acitivity records, step record, sleep monitoring, heart rate, weight and BMI calculations.

Using simple unique formula against unique ID of users bring out the fact that these files contain the above mentioned data for anywhere between 8 to 33 users. Another point to be noted here is the fact that some of these numbers are manual input of users, such as weight in the weightLogInfo_merged.csv file.

```
## Set the working directory to the root of your DSC 520 directory
## Load the `data/r4ds/week-6-housing.csv` to
setwd("C:/MastersCourse/RAssignemtents/data")

dailyActivity_data <- read.csv("dailyActivity_merged.csv", header = TRUE)
head(dailyActivity_data)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.44                     0.40
## 4                        0               2.14                     1.26
## 5                        0               2.71                     0.41
## 6                        0               3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

```
dailyCalories_data <- read.csv("dailyCalories_merged.csv", header = TRUE)
head(dailyActivity_data)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.44                     0.40
## 4                        0               2.14                     1.26
## 5                        0               2.71                     0.41
## 6                        0               3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

```
dailySteps_data <- read.csv("dailySteps_merged.csv", header = TRUE)
head(dailySteps_data)
```

```
##           Id ActivityDay StepTotal
## 1 1503960366   4/12/2016     13162
## 2 1503960366   4/13/2016     10735
## 3 1503960366   4/14/2016     10460
## 4 1503960366   4/15/2016      9762
## 5 1503960366   4/16/2016     12669
## 6 1503960366   4/17/2016      9705
```

```
sleepDay_data <- read.csv("sleepDay_merged.csv", header = TRUE)
head(sleepDay_data)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

```
weightLogInfo_data <- read.csv("weightLogInfo_merged.csv", header = TRUE)
head(weightLogInfo_data)
```

```
##           Id                  Date WeightKg WeightPounds Fat   BMI
## 1 1503960366  5/2/2016 11:59:59 PM     52.6     115.9631  22 22.65
## 2 1503960366  5/3/2016 11:59:59 PM     52.6     115.9631  NA 22.65
## 3 1927972279  4/13/2016 1:08:52 AM    133.5     294.3171  NA 47.54
## 4 2873212765 4/21/2016 11:59:59 PM     56.7     125.0021  NA 21.45
## 5 2873212765 5/12/2016 11:59:59 PM     57.3     126.3249  NA 21.69
## 6 4319703577 4/17/2016 11:59:59 PM     72.4     159.6147  25 27.45
##   IsManualReport        LogId
## 1           True 1.462234e+12
## 2           True 1.462320e+12
## 3          False 1.460510e+12
## 4           True 1.461283e+12
## 5           True 1.463098e+12
## 6           True 1.460938e+12
```

Required Packages

library("tidyverse") library("car") library("ggplot") library("dplyr") library("ggplot2") library("tidyr") library("dply")

Plots and Table Needs

ggplot : histogram : Density Curves : Box plot : Line Plot : Scatter Diagram :

Questions for future steps :

1.A time series analysis of your heart rate to forecast your future heart rate and comparing it with normal healthy heart rate may lead to think of healthy lie style.

2. What changes to dietary and life style choices after watching the data ?

3.Predicting cholestrol depending on factors like calorie in take, weight, no. of steps walked every day, distance covered, heart rate bpm, types of type of physical excercise,Although one of these activities you have to track outside of Activity Trackers. What all more factors as you see fit?

4.What are the effects of using these devices and correlating them to Relationship satisfaction or quality of life?

5.What type of decisions will our data science feature drive?

6.What metric will we use to call this project a success and how will we measure it?

7.what do they currently use and what is the baseline (current) value of that metric?

8.What the outcome of this project success?