# Final Report of Internship Program 2021

*On*

# *"ANALYSE FITNESS DATA"*

## MEDTOUREASY, NEW DELHI

30th January 2021

# ACKNOWLEDGMENTS

The internship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies of the subject of Data Visualizations in Data science; and also, for personal as well as professional development. I am very obliged for having a chance to interact with so many professionals who guided me throughout the internship project and made it a great learning curve for me.

Firstly, I express my deepest gratitude and special thanks to the Training Head of MedTourEasy, Mr. Ankit Hasija who gave me an opportunity to carry out my internship at their esteemed organization. Also, I express my thanks to him for making me understand the details of the Data Scientist profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and also for sparing his valuable time in spite of his busy schedule.

I would also like to thank the team of MedTourEasy and my colleagues who made the working environment productive and very conducive.

# TABLE OF CONTENTS

# ABSTRACT

With the explosion in fitness tracker popularity, runners all over the world are collecting data with gadgets (smartphones, watches, etc.) to keep themselves motivated. They look for answers to
questions like:
•How fast, long, and intense was my run today?
•Have I succeeded with my training goals?
•Am I progressing?
•What were my best achievements?
•How do I perform compared to others?
This data was exported from Runkeeper.The data is a CSV file where each row is a single training activity. In this project, you'll create import, clean, and analyze my data to answer the above questions.

## 1.1 About the Company

MedTourEasy, a global healthcare company, provides you the informational resources needed to evaluate your global options. It helps you find the right healthcare solution based on specific health needs, affordable care while meeting the quality standards that you expect to have in healthcare. MedTourEasy improves access to healthcare for people everywhere. It is an easy to use platform and service that helps patients to get medical second opinions and to schedule affordable, high-quality medical treatment abroad.

## 1.2 About the Project

With the explosion in fitness tracker popularity, runners all over the world are collecting data with gadgets (smartphones, watches, etc.) to keep themselves motivated. They look for answers to questions like:
•How fast, long, and intense was my run today?
•Have I succeeded with my training goals?
•Am I progressing?
•What were my best achievements?
•How do I perform compared to others?
This data was exported from Runkeeper. The data is a CSV file where each row is a single training activity. In this project, you'll create, import, clean, and analyze my data to answer the above questions.Hence, this project aims at collecting and analyzing large data sets to create intuitive and interactive dashboards for representing fitness in order to gain meaningful insights.

- *Analysis of the problem*: This is done to analyze the fitness statistics. It contains statistics and data representing the problem – avg pace, avg heart rate, distance. Also, it contains many comparative statistics with respect to parameters like type, duration, etc

The above section has been represented in the ipython notebook.

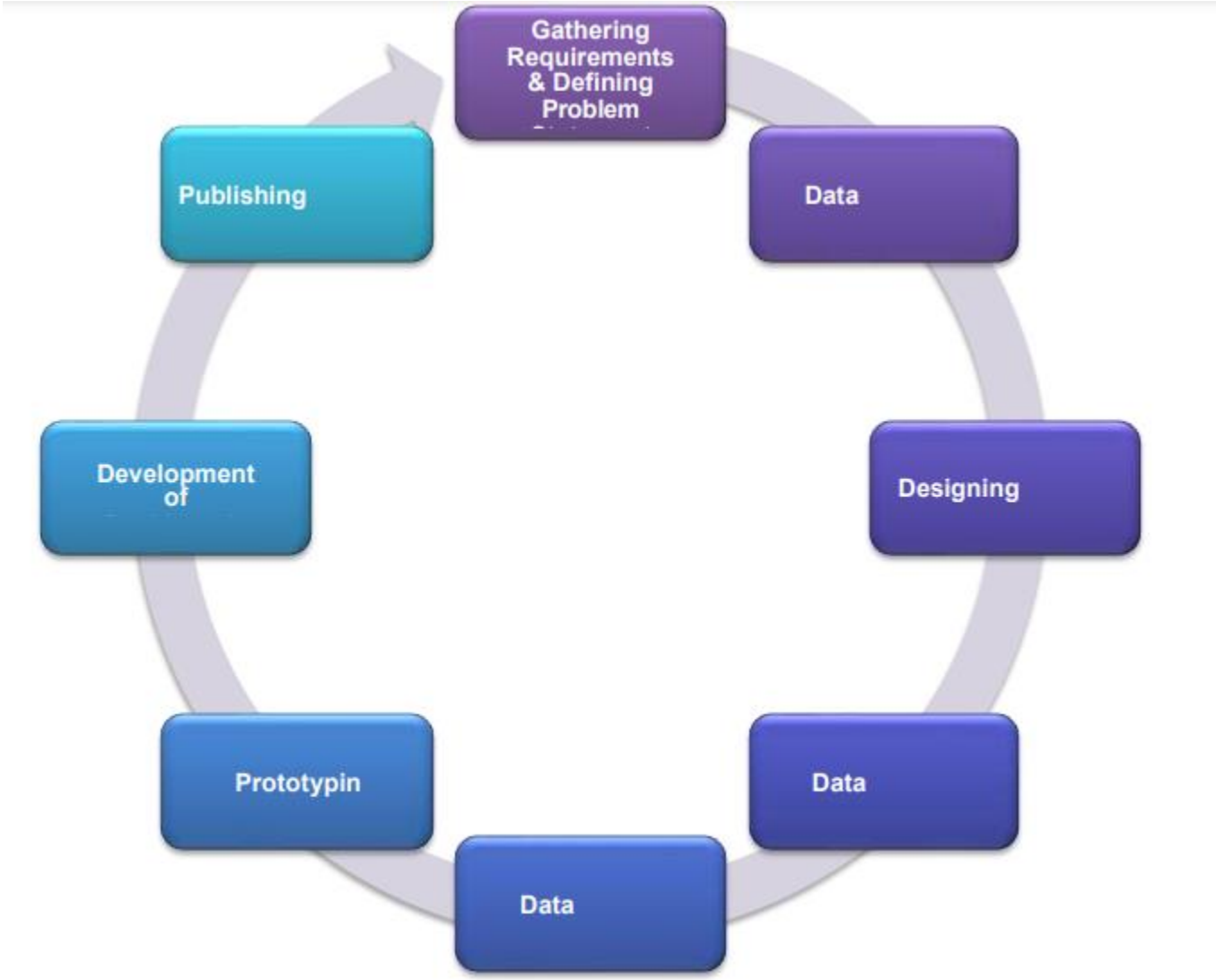
## 1.3 Objectives

This project focuses on creating easily understandable, interactive and dynamic dashboards by gathering data of running from runkeeper app. and using the coding language python.

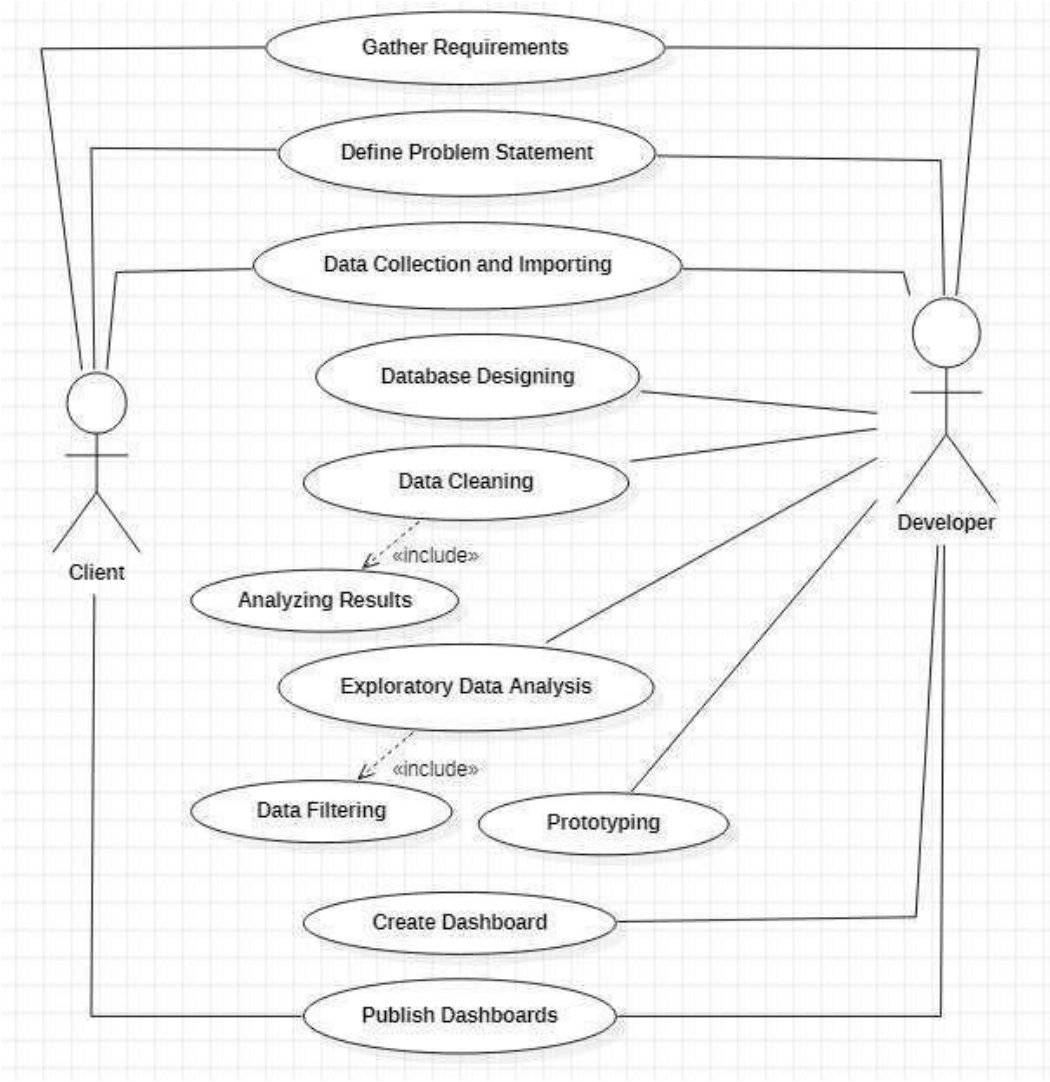
# I. METHODOLOGY

## 2.1 Flow of the Project

The project followed the following steps to accomplish the desired objectives and deliverables. Each step has been explained in detail in the following section.



## 2.2 Use Case Diagram




Above figure shows the use case of the project. There are two main actors in the same: The Client and Developer. The developer will first gather requirements and define the problem statement then collecting the required data and importing it. Then the developer will design databases so as to identify various constraints and relations in the data. Next step is to clean the data to remove irregular values, blank values etc. Next, exploratory data analysis is conducted to filter the data according to the requirements of the project. Then a prototype of the dashboards is created using PowerBI to get a clear view of the visualizations to be developed. Finally, a dashboard is developed and analyzed to publish the results to the client.

## 2.3 Language and Platform Used

### 2.3.1 Language: Python

Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.. The important features of Python are:

- Free and Open Source
- Object-Oriented Language
- GUI Programming Support
- High-Level Language
- Extensible feature
- Python is Portable language
- Python is Integrated language

### 2.3.2 IDE: Jupyter Notebook

The IPython Notebook is now known as the Jupyter Notebook. It is an interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media. Major features are:

- Interactive shells (terminal and Qt-based).
- A browser-based notebook interface with support for code, text, mathematical expressions, inline plots and other media.
- Support for interactive data visualization and use of GUI toolkits.
- Flexible, embeddable interpreters to load into one's own projects.
- Tools for parallel computing.

# II. IMPLEMENTATION

## ● Gathering Requirements and Defining Problem Statement

This is the first step wherein the requirements are collected from the clients to understand the deliverables and goals to be achieved after which a problem statement is defined which has to be adhered to while development of the project.

## ● Obtain and review raw data

A popular GPS fitness tracker called Runkeeper analyses running data and it exports data.One key feature of runkeeper app is: its excellent data export. Anyone who has a smartphone can download the app and analyze their data like we will in this notebook. After logging your run, the first step is to export the data from Runkeeper. Then import the data and start exploring to find potential problems. After that, create data cleaning strategies to fix the issues. Finally, analyze and visualize the clean time-series data. We have used seven years worth of training data, from 2012 through 2018. The data is a CSV file where each row is a single training activity. Let's load and inspect it.

## ● Data preprocessing

The column names Runkeeper provides are informative, and we don't need to rename any columns.But, we do notice missing values using the isna() method. What are the reasons for these missing values? It depends. Some heart rate information is missing because cardio sensors won’t be used always. In the case of the Notes column, it is an optional field and it is sometimes left blank. Also, Route Name column is only used once, and Freind’s tagged column is never used.We'll fill in missing values in the heart rate column to avoid misleading results later, but right now, our first data preprocessing steps will be to:

1. Remove columns not useful for our analysis.
2. Replace the "Other" activity type to "Unicycling" because that was always the "Other" activity.
3. Count missing values.

## 4. Dealing with missing values

There are 214 missing entries for my average heart rate. We can't go back in time to get those data, but we can fill in the missing values with an average value. This process is called *mean imputation*. When imputing the mean to fill in missing data, we need to consider that the average heart rate varies for different activities (e.g., walking vs. running). We'll filter the DataFrames by activity type (Type) and calculate each activity's mean heart rate, then fill in the missing values with those means.

## 5. Plot running data

We now create the first plot! As we found earlier, most of the activities in data were running (459 of them to be exact). There are only 29, 18, and two instances for cycling, walking, and unicycling, respectively. So for now, let's focus on plotting the different running metrics. An excellent first visualization is a figure with four subplots, one for each running metric (each numerical column). Each subplot will have a different y-axis, which is explained in each legend. The x-axis, Date, is shared among all subplots.

## 6. Running statistics

No doubt, running helps people stay mentally and physically healthy and productive at any age. And it is great fun! When runners talk to each other about their hobby, they not only discuss results, but also discuss different training strategies.You'll know you're with a group of runners if you commonly hear questions like:

1. What is your average distance?
2. How fast do you run?
3. Do you measure your heart rate?
4. How often do you train?

Let's find the answers to these questions in my data. If you look back at plots in previous task, you can see the answer to, *Do you measure your heart rate?* Before 2015: no. To look at the averages, let's only use the data from 2015 through 2018.
In pandas, the resample() method is similar to the groupby() method - with resample() you group by a specific time span. We'll use resample() to group the time series data by a sampling period and apply several methods to each sampling period. In our case, we'll resample annually and weekly.

## 7. Visualization with averages

Let's plot the long term averages of distance run and heart rate with their raw data to visually compare the averages to each training session. Again, we'll use the data from 2015 through 2018.In this task, we will use matplotlib functionality for plot creation and customization.

## 8. Was the goal reached?

To motivate a person to run regularly, They set a target goal of running 1000 km per year. Let's visualize annual running distance (km) from 2013 through 2018 to see if the goal was reached each year. Only stars in the green region indicate success.

## 9. Is the person progressing?

Let's dive a little deeper into the data to answer a tricky question: is the progress in

terms of my running skills?

To answer this question, we'll decompose weekly distance run and visually compare it to the raw data. A red trend line will represent the weekly distance run. We are going to use statsmodels library to decompose the weekly trend.

## 10. Training intensity

Heart rate is a popular metric used to measure training intensity. Depending on age and fitness level, heart rates are grouped into different zones that people can target depending on training goals. A target heart rate during moderate-intensity activities is about 50-70% of maximum heart rate, while during vigorous physical activity it's about 70-85% of maximum.We'll create a distribution plot of my heart rate data by training intensity. It will be a visual presentation for the number of activities from predefined training zones.

## 11. Detailed summary report

With all this data cleaning, analysis, and visualization, let's create detailed summary tables of training. To do this, we'll create two tables. The first table will be a summary of the distance (km) and climb (m) variables for each training activity. The second table will list the summary statistics for the average speed (km/hr), climb (m), and distance (km) variables for each training activity.

## 12. Fun facts

To wrap up, let's pick some fun facts out of the summary tables and solve the last exercise. These data (my running history) represent 6 years, 2 months and 21 days. And I remember how many running shoes I went through–7.
FUN FACTS
- Average distance: 11.38 km
- Longest distance: 38.32 km
- Highest climb: 982 m
- Total climb: 57,278 m
- Total number of km run: 5,224 km
- Total runs: 459
- Number of running shoes gone through: 7 pairs

The story of Forrest Gump is well known–the man, who for no particular reason decided to go for a "little run." His epic run duration was 3 years, 2 months and 14 days (1169 days). In the picture you can see Forrest's route of 24,700 km.

FORREST RUN FACTS
- Average distance: 21.13 km
- Total number of km run: 24,700 km
- Total runs: 1169
- Number of running shoes gone through: ...

Assuming Forest and I go through running shoes at the same rate, figure out how many pairs of shoes Forrest needed for his run.

# III. SAMPLE SCREENSHOTS

## Dataset:

| Date | Activity Id | Type | Route Name | Distance (km) | Duration | Average Pace | Average Speed (km/h) | Calories Burned | Climb (m) | Average Heart Rate (bpm) | Friend's Tagged | Notes | GPX File |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013-01-19 09:58:50 | 98321fac-a333-47d7-b568-1c609096a08f | Running | NaN | 3.39 | 15:56 | 4:42 | 12.75 | 190.0 | 15 | NaN | NaN | NaN | 2013-01-19-095850.gpx |
| 2015-06-22 18:19:53 | 2b3bb63b-e5b7-419e-9b53-744ce195fa83 | Running | NaN | 18.84 | 1:41:17 | 5:22 | 11.16 | 1294.0 | 156 | 141.0 | NaN | NaN | 2015-06-22-181953.gpx |
| 2017-05-10 18:40:04 | 0edf90b1-2417-413a-96b6-368d01df8677 | Running | NaN | 8.53 | 43:24 | 5:05 | 11.79 | 595.0 | 87 | 146.0 | NaN | TomTom MySports Watch | 2017-05-10-184004.gpx |

## Dataset Info:

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Activity Id              508 non-null    object
 1   Type                     508 non-null    object
 2   Route Name               1 non-null      object
 3   Distance (km)            508 non-null    float64
 4   Duration                 508 non-null    object
 5   Average Pace             508 non-null    object
 6   Average Speed (km/h)     508 non-null    float64
 7   Calories Burned          508 non-null    float64
 8   Climb (m)                508 non-null    int64
 9   Average Heart Rate (bpm) 294 non-null    float64
 10  Friend's Tagged          0 non-null      float64
 11  Notes                    231 non-null    object
 12  GPX File                 504 non-null    object
dtypes: float64(5), int64(1), object(7)
```

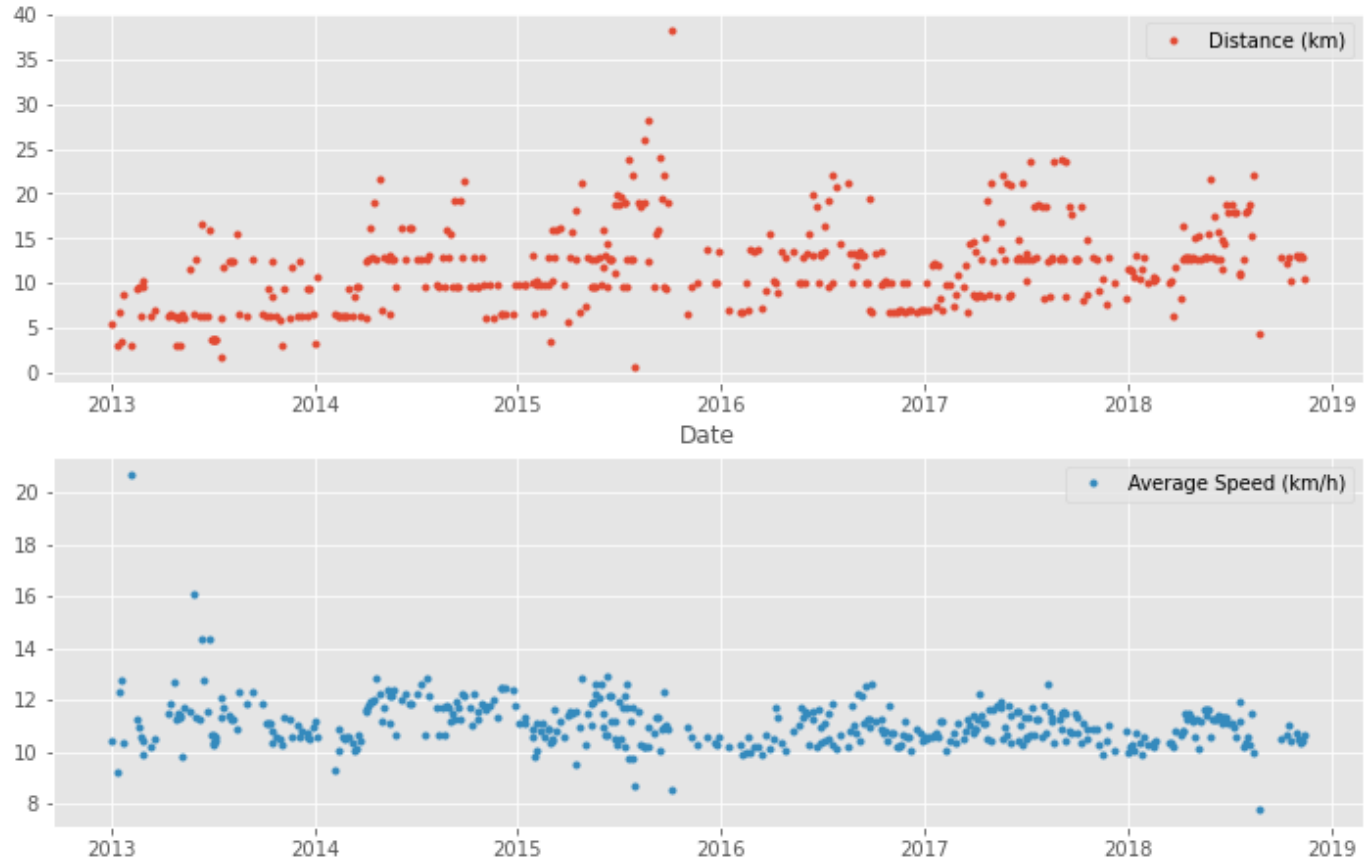
## Total values and Missing Values in each column:

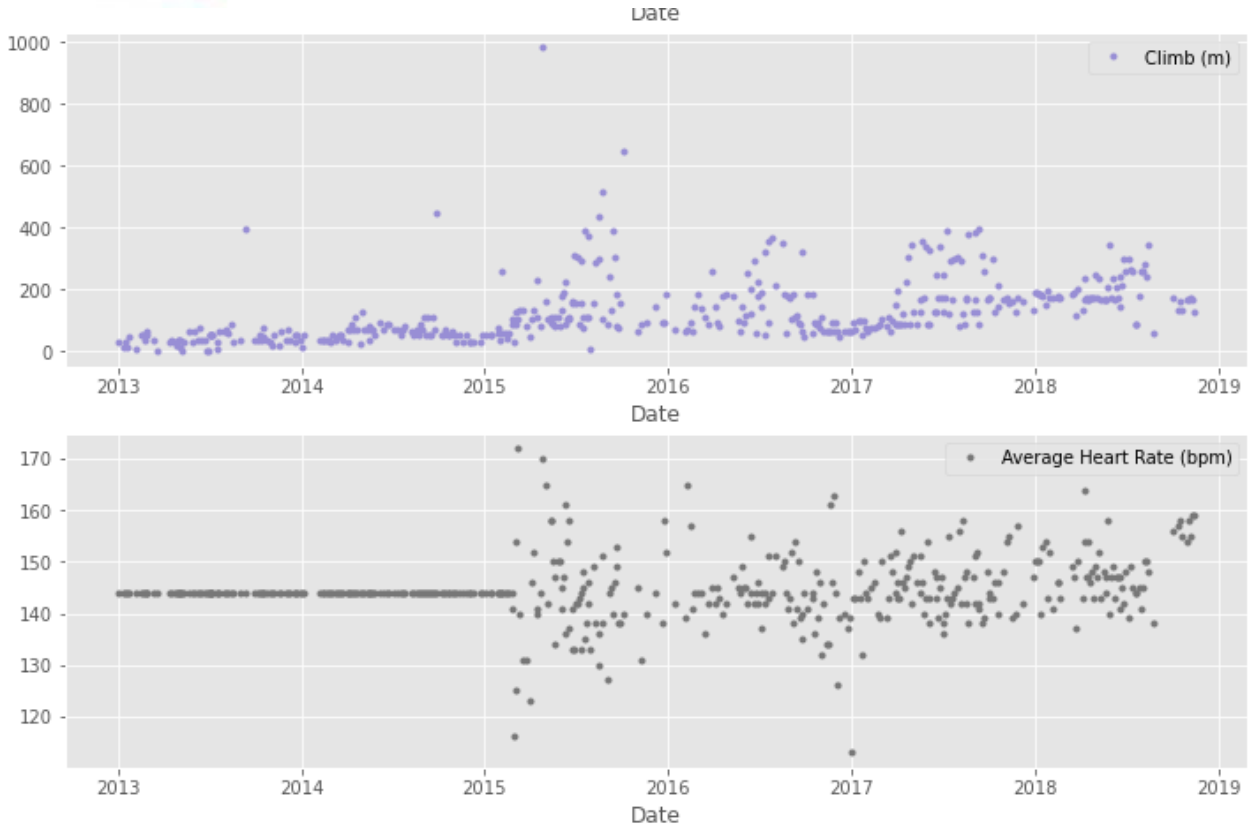
```
Running    459
Cycling     29
Walking     18
Other        2
Name: Type, dtype: int64

Type                          0
Distance (km)                 0
Duration                      0
Average Pace                  0
Average Speed (km/h)          0
Climb (m)                     0
Average Heart Rate (bpm)    214
dtype: int64
```

## Missing values after filling columns with missing values

```
Type                        0
Distance (km)               0
Duration                    0
Average Pace                0
Average Speed (km/h)        0
Climb (m)                   0
Average Heart Rate (bpm)    0
dtype: int64
```

## Different parameters between 2013-2018



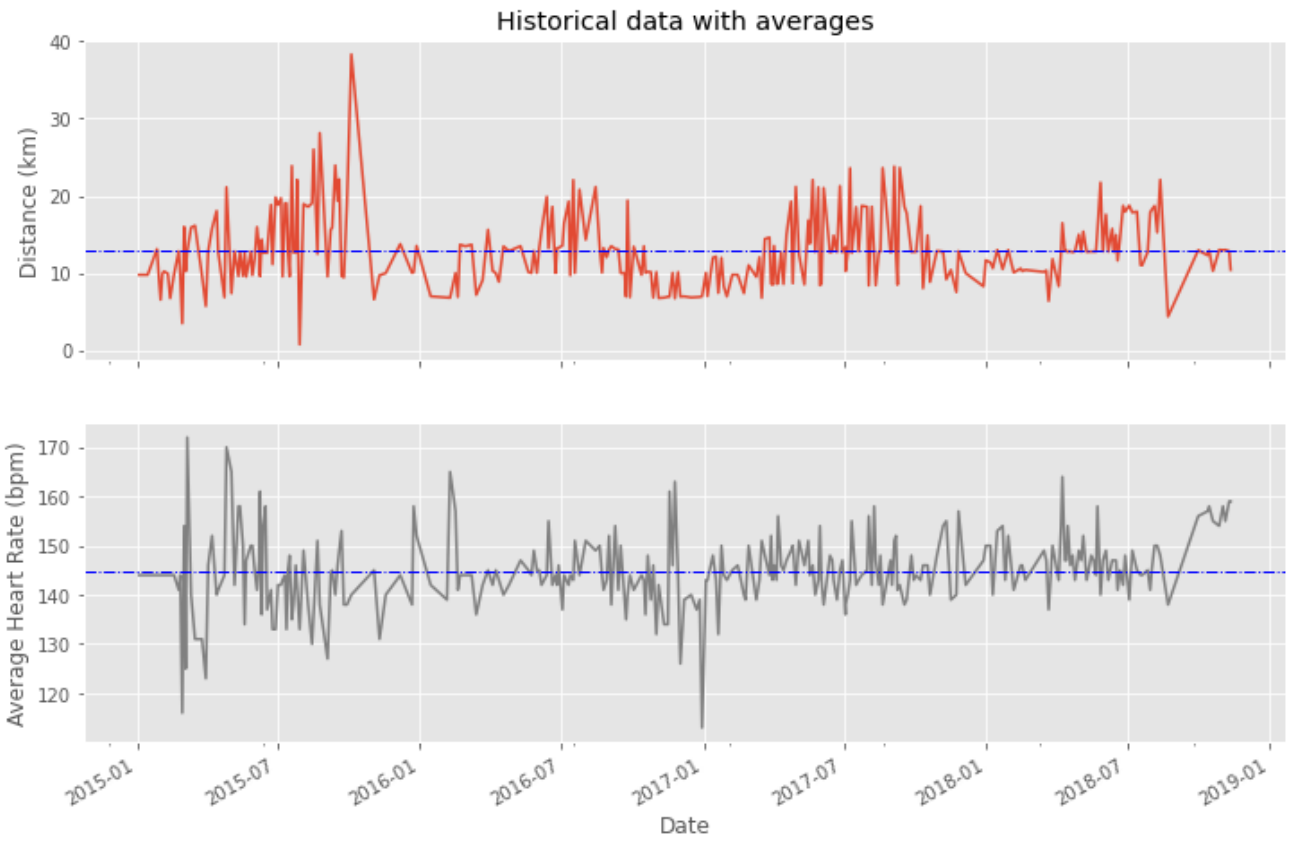



## Average run weekly and in last 4 years

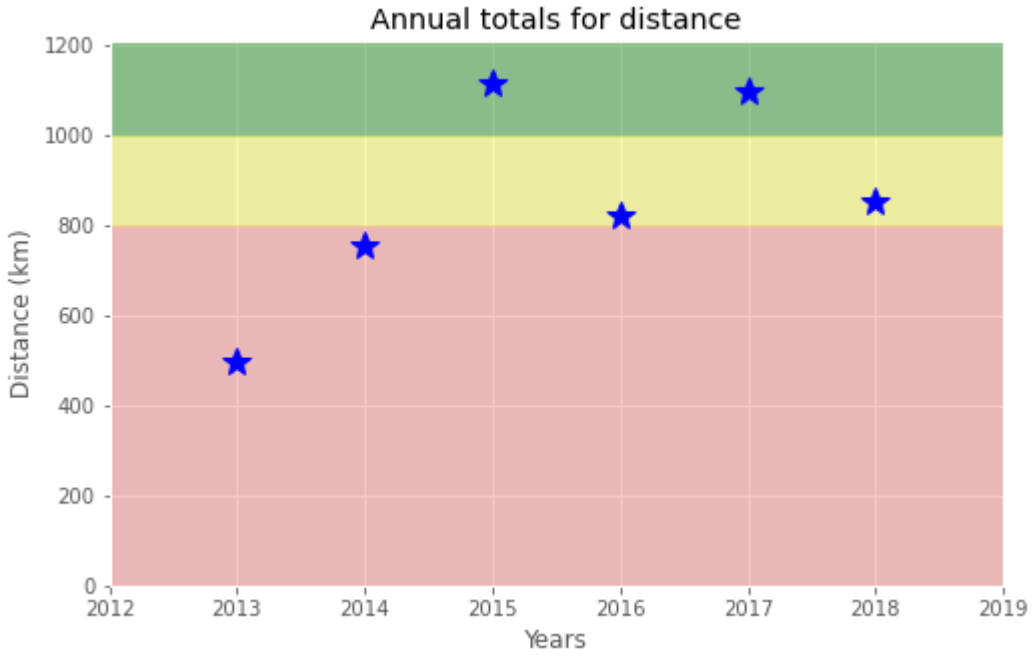
| Date | Distance (km) | Average Speed (km/h) | Climb (m) | Average Heart Rate (bpm) |
|---|---|---|---|---|
| **2015-12-31** | 13.602805 | 10.998902 | 160.170732 | 143.353659 |
| **2016-12-31** | 11.411667 | 10.837778 | 133.194444 | 143.388889 |
| **2017-12-31** | 12.935176 | 10.959059 | 169.376471 | 145.247059 |
| **2018-12-31** | 13.339063 | 10.777969 | 191.218750 | 148.125000 |

```
Weekly averages of last 4 years:

Distance (km)               12.518176
Average Speed (km/h)        10.835473
Climb (m)                  158.325444
Average Heart Rate (bpm)   144.801775
dtype: float64

How many trainings per week I had on average: 1.5
```

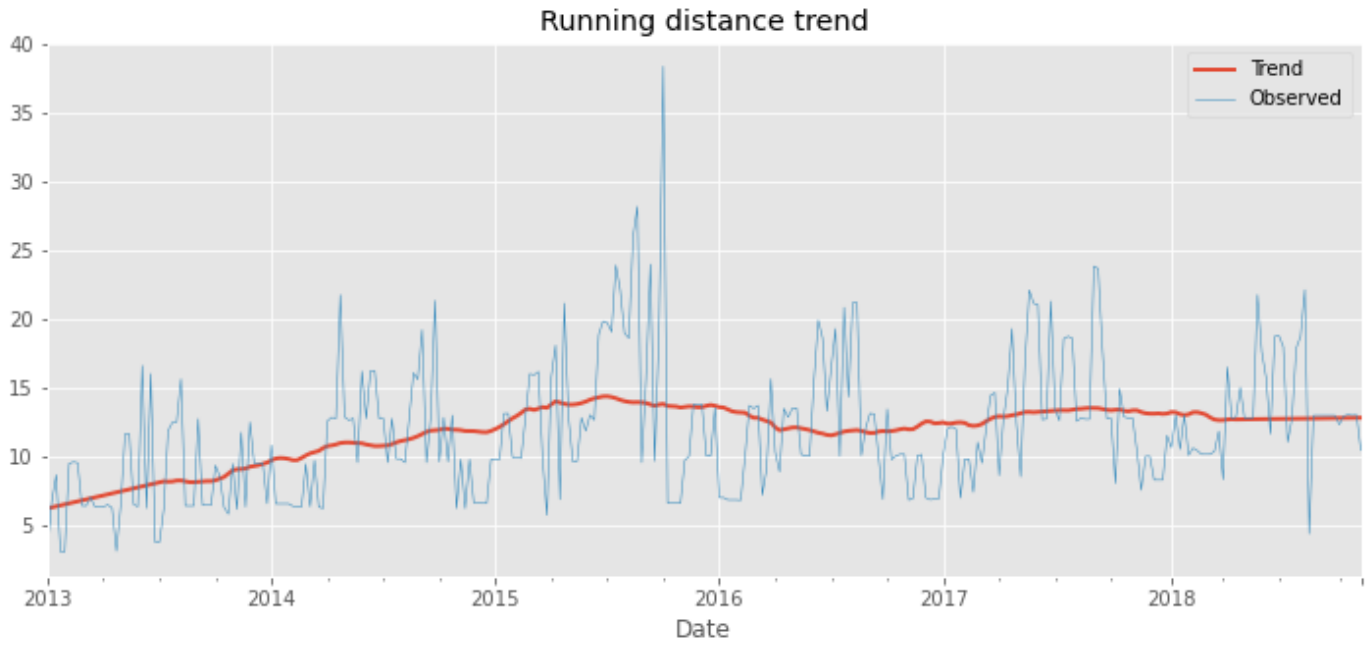
## Distance and Heart rate avg plot



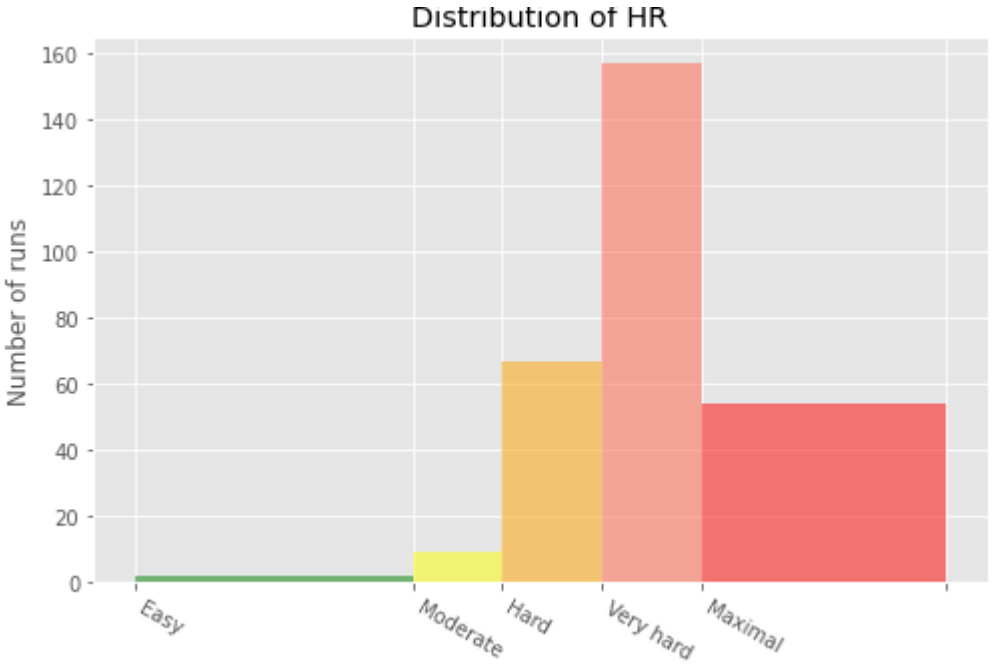

## Annual total distance to check if goal was reached



## Running distance trend




## Distribution of heart rates



# Detailed summary report

Totals for different training types:

| Type | Distance (km) | Climb (m) |
|---|---|---|
| Cycling | 680.58 | 6976 |
| Running | 5224.50 | 57278 |
| Walking | 33.45 | 349 |

Summary statistics for different training types:

| Type | | Average Speed (km/h) | Climb (m) | Distance (km) |
|---|---|---|---|---|
| Cycling | 25% | 16.980000 | 139.000000 | 15.530000 |
| | 50% | 19.500000 | 199.000000 | 20.300000 |
| | 75% | 21.490000 | 318.000000 | 29.400000 |
| | count | 29.000000 | 29.000000 | 29.000000 |
| | max | 24.330000 | 553.000000 | 49.180000 |
| | mean | 19.125172 | 240.551724 | 23.468276 |
| | min | 11.380000 | 58.000000 | 11.410000 |
| | std | 3.257100 | 128.960289 | 9.451040 |
| | total | NaN | 6976.000000 | 680.580000 |
| Running | 25% | 10.495000 | 54.000000 | 7.415000 |
| | 50% | 10.980000 | 91.000000 | 10.810000 |
| | 75% | 11.520000 | 171.000000 | 13.190000 |
| | count | 459.000000 | 459.000000 | 459.000000 |
| | max | 20.720000 | 982.000000 | 38.320000 |
| | mean | 11.056296 | 124.788671 | 11.382353 |
| | min | 5.770000 | 0.000000 | 0.760000 |
| | std | 0.953273 | 103.382177 | 4.937853 |
| | total | NaN | 57278.000000 | 5224.500000 |
| Walking | 25% | 5.555000 | 7.000000 | 1.385000 |
| | 50% | 5.970000 | 10.000000 | 1.485000 |
| | 75% | 6.512500 | 15.500000 | 1.787500 |
| | count | 18.000000 | 18.000000 | 18.000000 |
| | max | 6.910000 | 112.000000 | 4.290000 |
| | mean | 5.549444 | 19.388889 | 1.858333 |
| | min | 1.040000 | 5.000000 | 1.220000 |
| | std | 1.459309 | 27.110100 | 0.880055 |
| | total | NaN | 349.000000 | 33.450000 |

# IV. CONCLUSION AND FUTURE SCOPE

In today's world fitness is of utmost importance, and you need constant physical activities to keep yourself fit, Runkeeper apps keeps track of your running, heart rate, distance, notes and atc.

This project aimed at analyzing the current situation of the pandemic by creating intuitive and user interactive dashboards and drawing conclusions on the impact it will have on fitness. Currently, the project is in its last stage of development with the dashboards being developed and submitted for review and feedback.

With regards to the future work, it aims at regularly updating the dashboards with time and integrating it with their systems so as to continually draw conclusions and analyze the fitness statistics. This will enable them to keep track of runnings and provide different plot to help people increase fitness.

# V. REFERENCES

**Data Collection:**

**https://drive.google.com/file/d/1cnJjROpg2m4TYPK9NHEhlh3FsYaYkN3a/view**

**Code Link:**

**https://github.com/harshsomaiya18/MedTourEasy/blob/main/Data%20Science%20MTE%20Project.ipynb**