Final Report of Internship Program 2021

On

"Analyse Fitness Data"

MEDTOUREASY

28th March 2021

ACKNOWLDEGMENTS

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 Analytics; 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 & Developement Team of MedTourEasy who gave me an opportunity to carry out my internship at their esteemed organization. Also, I express my thanks to the team for making me understand the details of the Data Analytics profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and also for spearing 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

Sr. No.	Topic
1	Introduction
	1.1 About the Company
	1.2 About the Project
2	Methodology
	2.1 Flow of the Project
	2.2 Use Case Diagram
	2.3 Language and Platform Used
3	Implementation
	3.1 Gathering Requirements and Defining Problem Statement
	3.2 Data Collection and Importing
	3.3 Designing Databases
	3.4 Data Cleaning
	3.5 Data Filtering 3.6 Defining Visuals
6	Conclusion
7	Future Scope

ABSTRACT

To put it simply, physical activity and exercise is important for everyone. Children, adolescents, and adults of all ages need regular physical activity. Physical activity promotes good health, and you should stay active throughout all stages of your life regardless of your body type or BMI. Understanding the benefits of physical fitness and knowing how active you should be can help you maintain good health and improve your overall quality of life. Here are a few benefits of regular physical activity that demonstrate the importance of physical fitness. With the explosion in fitness tracker popularity, runners all of the world are collecting data withgadgets (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?

We need to encourage your family to be more active, and challenge yourself to meet daily or weekly physical activity goals. Play outdoor sports with the whole family, schedule time each day to go to the gym, or pick up healthy, active hobbies like hiking or cycling

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

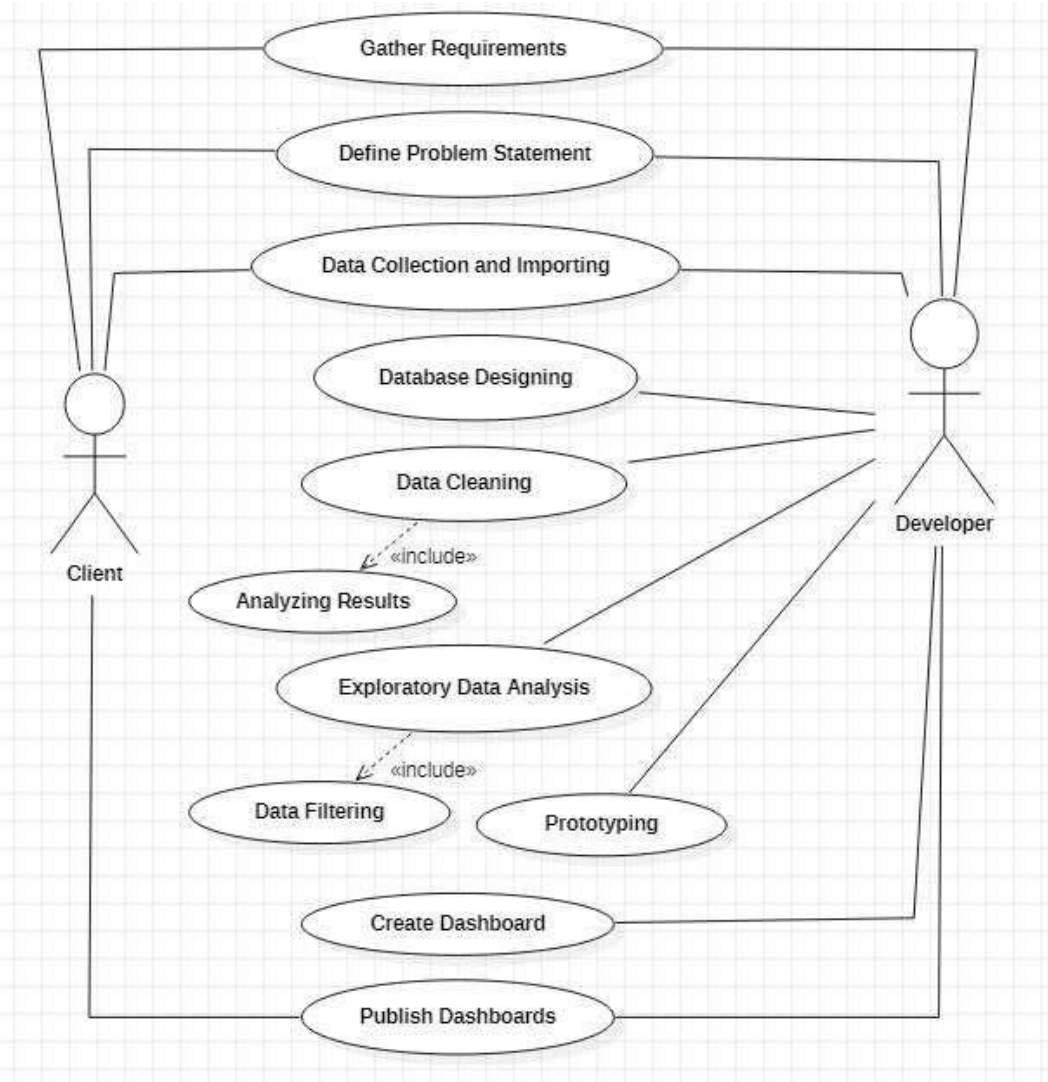
To put it simply, physical activity and exercise is important for everyone. Children, adolescents, and adults of all ages need regular physical activity. Physical activity promotes good health, and you should stay active throughout all stages of your life regardless of your body type or BMI. Understanding the benefits of physical fitness and knowing how active you should be can help you maintain good health and improve your overall quality of life. Here are a few benefits of regular physical activity that demonstrate the importance of physical fitness. With the explosion in fitness tracker popularity, runners all of the world are collecting data withgadgets (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?

Methodology

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

2.1 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..

2.2 Language and Platform Used

2.2.1 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 indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.[

Python is dynamically-typed and garbage-collected It supports multiple programming paradigms, including structured(particularly, procedural), object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library.

Guido van Rossum began working on Python in the late 1980's, as a successor to the ABC programming language, and first released it in 1991 as Python 0.9.0.Python 2.0 was released in 2000 and introduced new features, such as list comprehensions and a garbage collection system using reference counting and was discontinued with version 2.7.18 in 2020. Python 3.0 was released in 2008 and was a major revision of the language that is not completely backward-compatible and much Python 2 code does not run unmodified on Python 3.

2.2.2 IDE: Jupyter Notebook

Project Jupyter is a project and community whose goal is to "develop open-source software, open-standards, and services for interactive computing across dozens of programming languages". It was spun off from IPython in 2014 by Fernando Pérez. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are **Ju**lia, **Pyt**hon and **R**, and also a homage to Galileo's notebooks recording the discovery of the moons of Jupiter. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab.

2.2.3 Package: Pandas

In computer programming, **pandas** is a software library written for the Pytho programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.[Its name is a play on the phrase "Python data analysis" itself.Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010.

2.2.4 Package: Numpy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

IMPLEMENTATION

3.1 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.

3.2 Data Collection and Importing

Data collection is a systematic approach for gathering and measuring information from a variety of sources in order to obtain a complete and accurate picture of an interest area. It helps an individual or organization to address specific questions, determine outcomes and forecast future probabilities and patterns.

Data importing is referred to as uploading the required data into the coding environment from internal sources (computer) or external sources (online websites and data repositories). This data can then be manipulated, aggregated, filtered according to the requirements and needs of the project.

Packages Used:

read.csv (): It is a wrapper function for read.table() that mandates a comma as separator and uses the input file's first line as header that specifies the table's column names. Thus, it is an ideal candidate to read CSV files. It has an additional parameter of url() which is used to pull live data directly from GitHub repository.

read_excel (): It calls excel_format() to determine if path is xls or xlsx, based on the file extension and the file itself, in that order.

Sample Code:

```
import pandas as pd

# ... YOUR CODE FOR TASK 1 ...


# Define file containing dataset

df_activities =pd.read_csv("D:/Data Science/Project/Medtour
easy/cardioActivities.csv",parse_dates=["Date"])
```

3.4 Data Cleaning

"Quality data beats fancy algorithms"

Data is the most imperative aspect of Analytics and Machine Learning. Everywhere in computing or business, data is required. But many a times, the data may be incomplete, inconsistent or may contain missing values when it comes to the real world. If the data is corrupted then the process may be impeded or inaccurate results may be provided. Hence, Data cleaning is considered a foundational element of the basic data science.

Data Cleaning means the process by which the incorrect, incomplete, inaccurate, irrelevant or missing part of the data is identified and then modified, replaced or deleted as needed.

With reference to the Aanalysis of fitness dataset, it may contain many null values or incorrect value simply because of inconsistency in reporting cases and testing statistics by countries and states. Hence various functions are used to clean this data.

Packages Used:
Functions Used:

Is.null(): In Python, missing values are represented by the symbol **NA** (not available). Impossible values (e.g., dividing by zero) are represented by the symbol **NaN** (not a number). This function is used to check if a dataset contains NA values or not.

info(): The info() function is used to print a concise summary of a DataFrame. This method prints information about a DataFrame including the index dtype and column

Sample Code:

```
# First look at exported data: select sample of 3 random rows

df_activities.sample(3)

#display(...)
```

3.5 Data Filtering

Data filtering is the method of choosing a smaller portion of the data set and using that subset to view, analyze and evaluate data. Generally, filtering is temporary –the entire data set is retained, but only part of it is used for calculation. It is also called subsetting or drill down data wherein data is extracted with respect to certain defined logical conditions. Filtering is used for the following tasks:

- Analyzing results for a particular period of time.
- Calculating results for particular groups of interest.
- Exclude erroneous or "bad" observations from an analysis.
- Train and validate statistical models.

With respect to Fitness dataset, the data needs to be filtered according to certain conditions like years to analyse the progress of our activities,

Mean(): function can be used to calculate mean/average of a given list of numbers. It returns mean of the data set passed as parameters.
Arithmetic mean is the sum of data divided by the number of data-points. It is a measure of the central location of data in a set of values which vary in range. In Python, we usually do this by dividing the sum of given numbers with the count of number present.

Copy(): Sometimes, there is a need to reuse any object, hence copy methods are always of a great utility. Python in its language offers a number of ways to achieve this. This particular article aims at demonstrating the copy method present in the list. Since the library.

Sample Code:

```
# Calculate sample means for heart rate for each training activity type

avg_hr_run = df_activities[df_activities['Type'] == 'Running']['Average Heart
Rate (bpm)'].mean()

avg_hr_cycle = df_activities[df_activities['Type']=="Cycling"]['Average Heart
Rate (bpm)'].mean()
```

3.6 Defining Visuals

Data visualization is presenting data in a graphical or pictorial format. It allows decision-makers to see visually presented analytics, so that they can grasp difficult concepts or identify new patterns. In interactive visualizations, technology can be used to dig in charts and graphs for more detail, interactively modifying what data one can see and how it works.

Because of the way in which the human brain processes information, it is easier to visualize large amounts of complex data using charts or graphs than to poring over spreadsheets or reports. Data visualization is a quick, easy and universal way of conveying concepts. Data visualization can also:

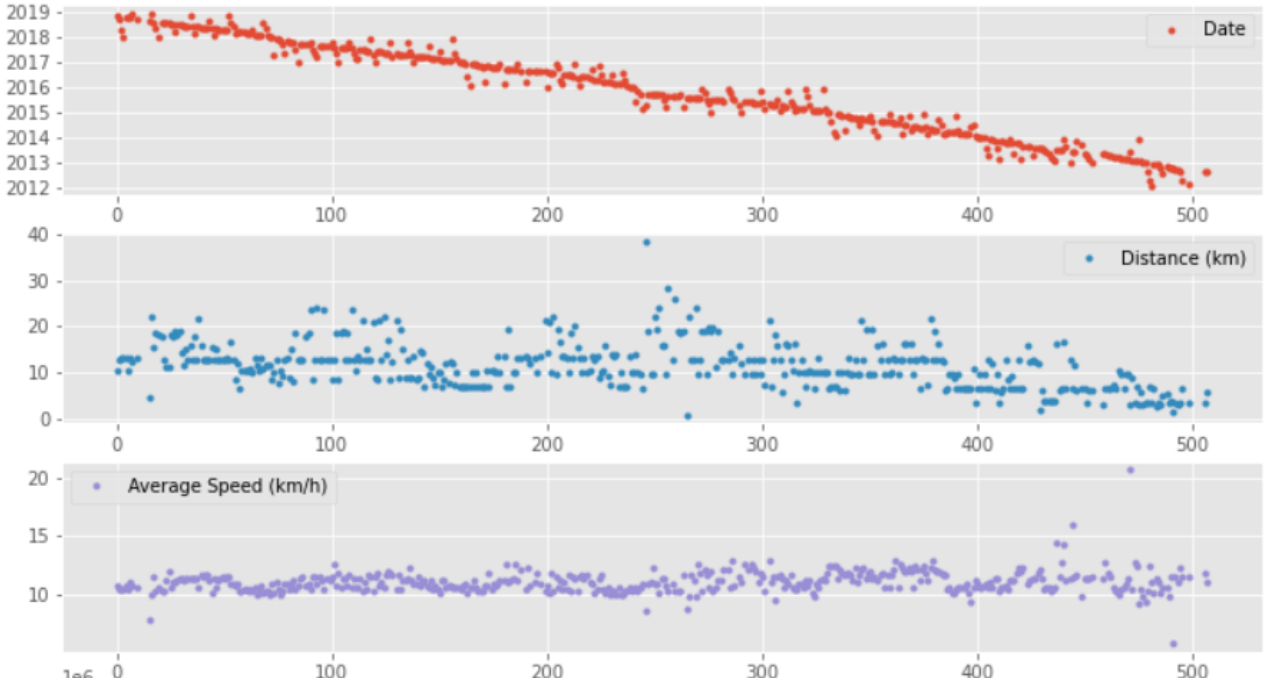
- Identify areas that need attention or improvement.
- Clarify which factors influence customer behaviour.
- Help you understand which products to place where.
- Predict sales volumes.

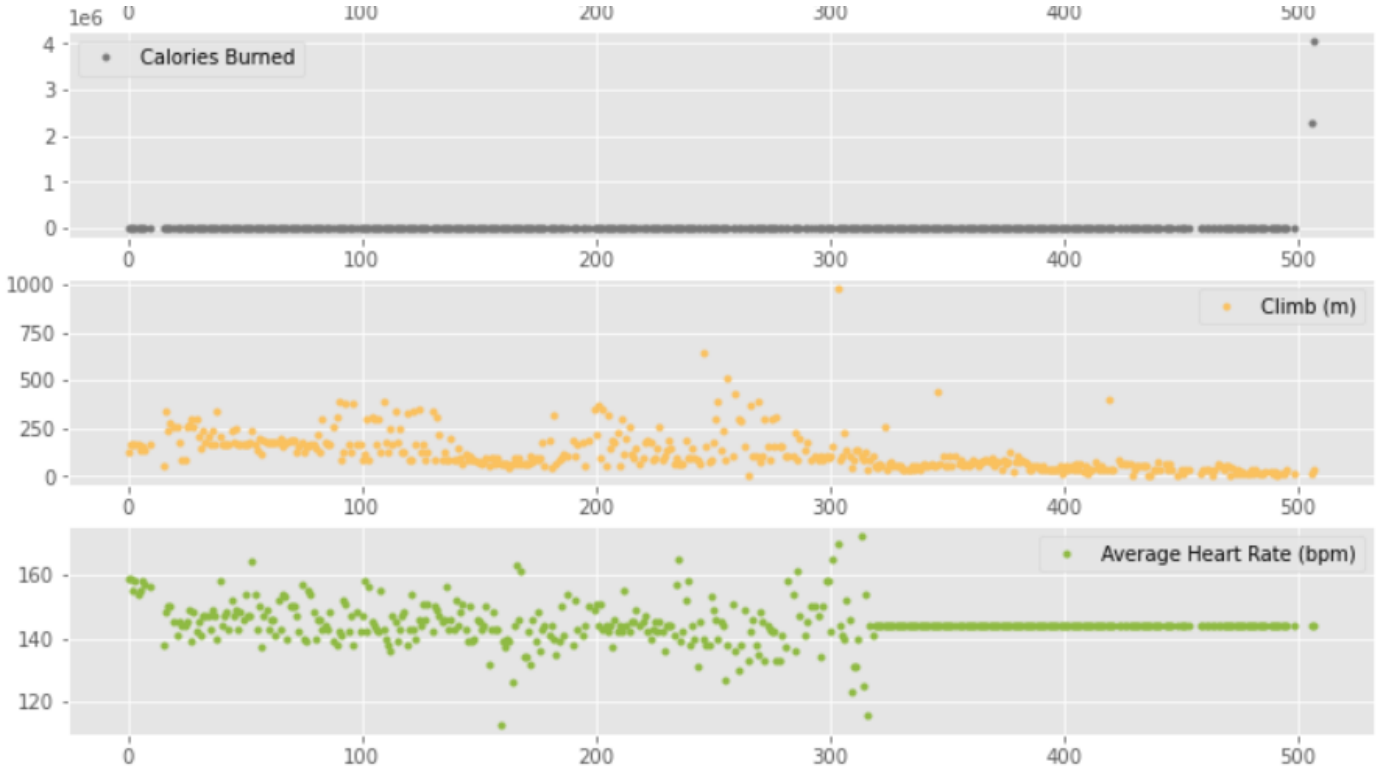
Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits.

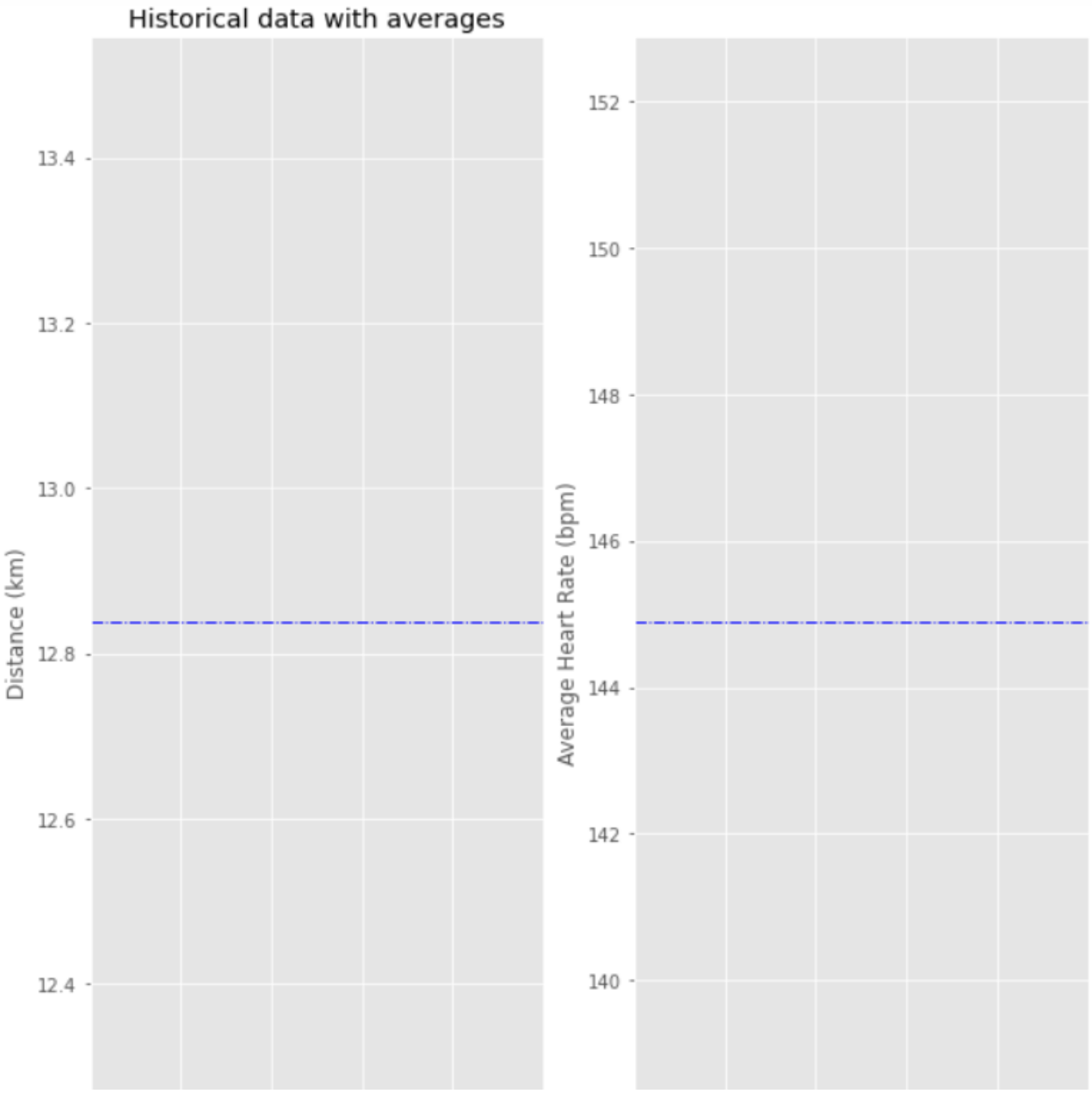
We can create our first plot! As we found earlier, most of the activities in my data were running (459 of them to be exact). There are only 29, 18, and two instances for cycling, walking, and unicycling, respectively. So for now, let's focus on plotting the different running metrics.

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.

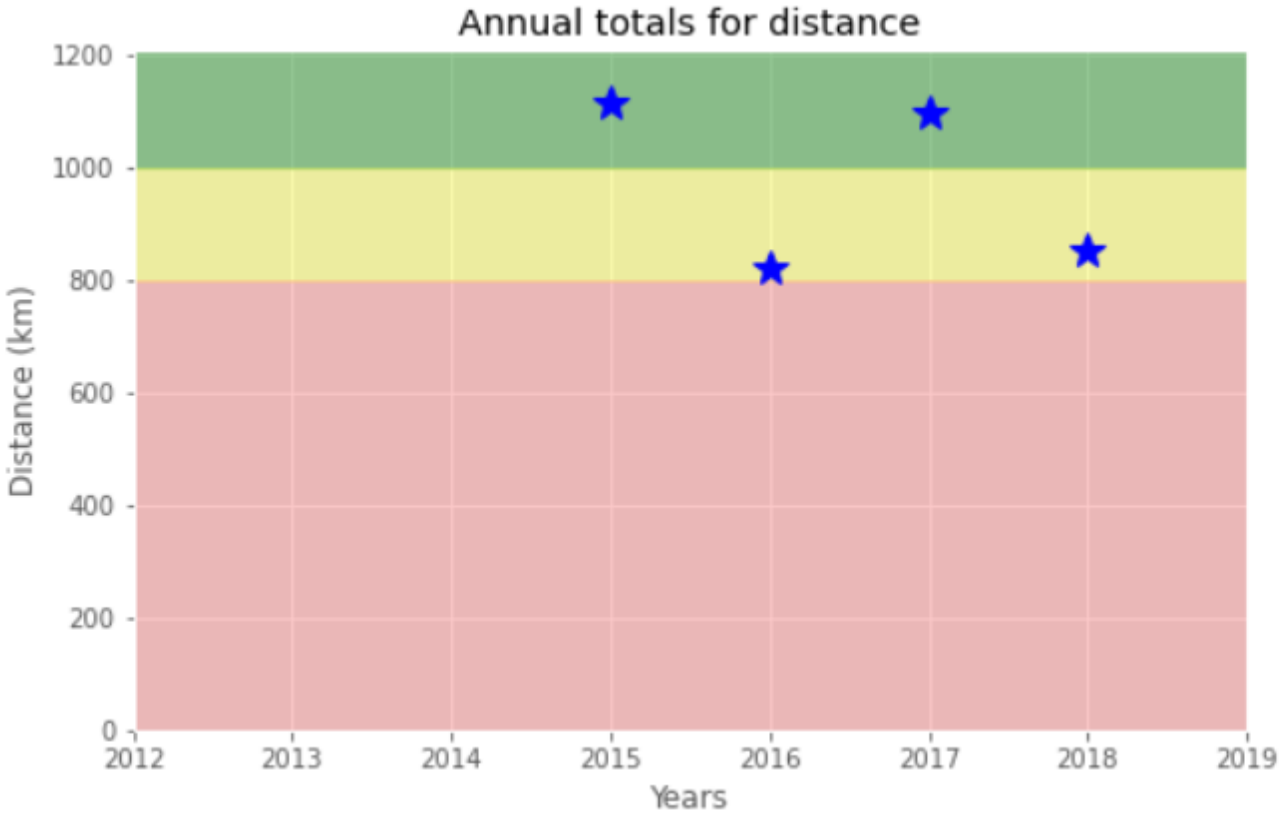

Date
Distance (km)
Average Speed (km/h)
2019
2018
2017
2016
2015
2014
2013
2012
40
30
20
10
0
15
0
100
200
300
400
500
1e6

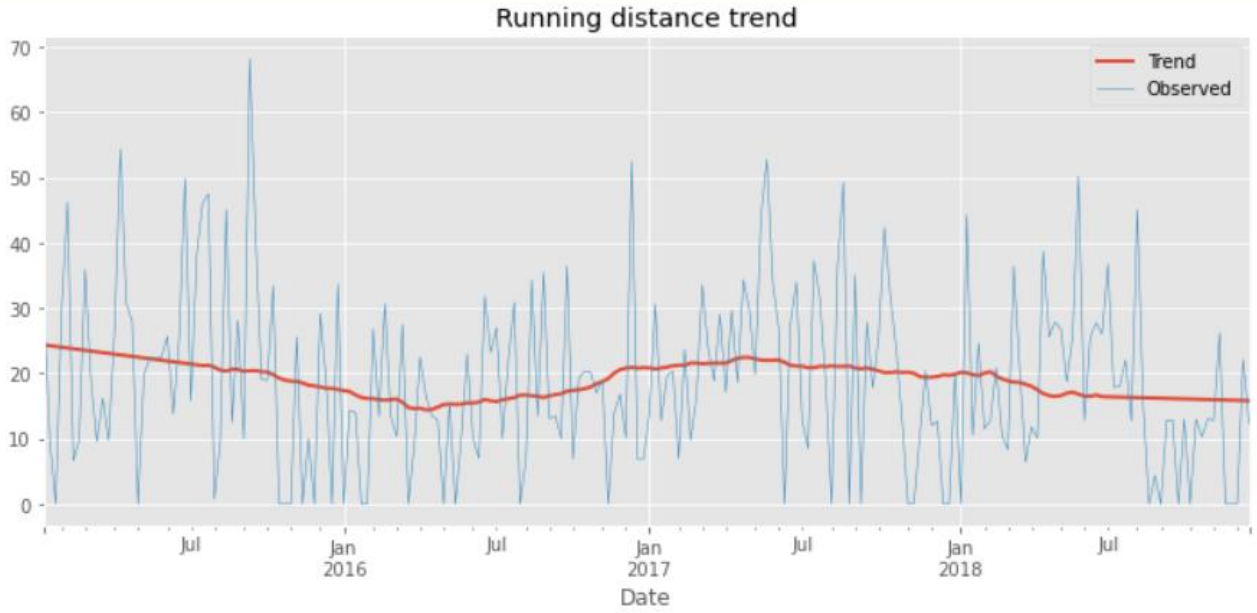


long term averages of my distance run and my heart rate with their raw data to visually compare the averages to each training session. Again, we'll use the data from 2015 through 2018.

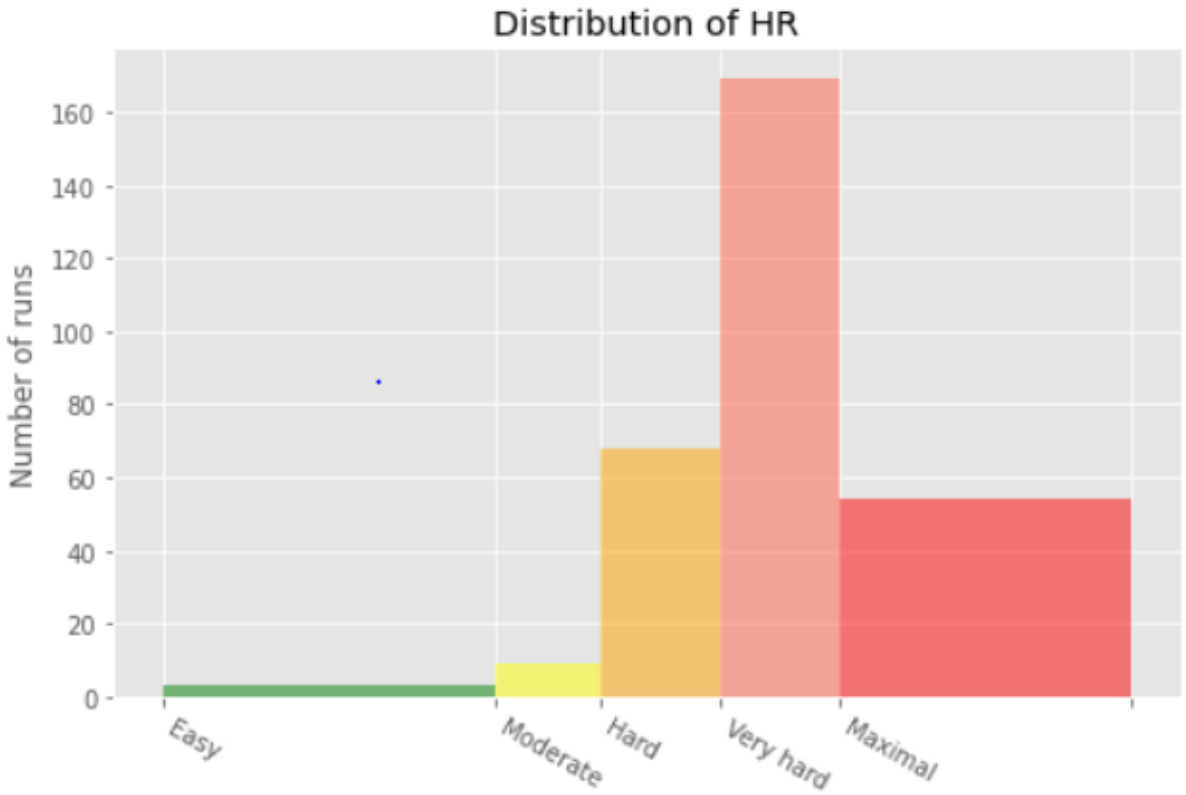

Historical data with averages
Distance (km)
13.4
13.2
13.0
12.8
12.6
12.4
Average Heart Rate (bpm)
152
150
148
146
144
142
140

set a target goal of running 1000 km per year to visualize the annual running distance (km) from 2013 through 2018 to see if I reached my goal each year. Only stars in the green region indicate success.





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.



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.

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%	10.420000	89.000000	9.785000
	50%	10.840000	145.000000	12.720000
	75%	11.335000	187.500000	14.745000
	count	303.000000	303.000000	303.000000
	max	12.880000	982.000000	38.320000
	mean	10.902772	162.900990	12.839142
	min	7.750000	6.000000	0.760000
	std	0.680800	103.356725	4.620825
	total	NaN	49359.000000	3890.260000

In [56]:

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: 167 pairs

.

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

CONCLUSION AND FUTURE SCOPE

The health and fitness industry has become extremely diverse in the range of services and facilities it offers; varying from large scale leisure centres and gymnasiums, to individual personal trainers who travel from one client to the next in their cars., Some services specialize in offering structured classes, others are informal; some cater for a particular demographic, and others for anyone.This group of people can make up a great proportion of your clientèle. They generally do not like loud music, so keep it soft or at a medium level. Most also prefer low-key surroundings so avoid bright lights, mirrors, or posters of super-fit young people for example. Older participants are often more concerned about joint movement and flexibility than weight loss and can relate better to an older instructor who shows respect and individual attention to their requirements. It is very common that this type of population attends for social reasons - to meet others and relax.This can be a hard group to cater for because they can vary considerably in the type of sports they are involved in, and elite athletes may require very specialised fitness components. The triathlete, swimmer, runner or aerobics-orientated person will want a hard, advanced class to challenge their fitness.

The type of sport and the fitness components to be improved will indicate the type of training employed. You may find elite athletes require personal training, one-on-one, to best achieve their goals.