**Final Report of Traineeship Program 2023**

**On**

**“*Analysis of Fitness Data*”**

**MedTourEasy**

**By: -Roshil Verma**

** ACKONWLEDGEMENTS**

The traineeship opportunity I had with MedTourEasy was truly invaluable in terms of my learning and understanding of the intricate subject of Data Visualizations, Data Analytics in Data Science. It has been a transformative experience for both my personal and professional development. I am sincerely grateful for the chance to interact with and learn from the highly skilled professionals at MedTourEasy who guided me throughout my traineeship project, making it an enriching learning journey.

I am immensely grateful for the mentorship and guidance provided by my project supervisor, who has been an invaluable source of knowledge and expertise. Their patient guidance and insights have been invaluable in helping me navigate through the challenges of the project and hone my skills in Data Science.

In conclusion, I am deeply grateful for the enriching traineeship opportunity with MedTourEasy, which has been instrumental in my learning and growth in the field of Data Science. I am thankful to the Training & Development Team, my project supervisor, and the entire team at MedTourEasy for their support, guidance, and mentorship. It has been an invaluable experience that will undoubtedly shape my future professional endeavours positively.

**TABLE OF CONTENTS**

Acknowledgments .................................... (i)

Abstract .....................................................(iii)

|  |  |  |
| --- | --- | --- |
| **S.no** | **Topic** | **Page No.** |
| **1** | Introduction |  |
|  | 1.1 About MedTourEasy | 5 |
|  | 1.2 About the Project | 5 |
|  | 1.3 Objectives | 6 |
| **2** | Methodology |  |
|  | 2.1 Flow of the Project | 7 |
|  | 2.2 Language and Platform Used | 7 |
| **3** | Implementation |  |
|  | 3.1 Gathering Requirements | 10 |
|  | 3.2 Data Collecting and Importing | 10 |
|  | 3.3 Data Preprocessing | 11 |
|  | 3.4 Data Cleaning | 11 |
| **4** | Plotting the Running Data | 13 |
| **5** | Running Statistics | 14 |
| **6** | Visualizing 2015 to 2018 data of distance and average heart rate with averages | 15 |
| **7** | Did I Reach My Goals? | 16 |
| **8** | Am I Progressing? | 17 |
| **9** | Training Intensity | 18 |
| **10** | Summary and End Notes | 19 |
| **11** | Fun Facts | 20 |

**ABSTRACT**

This abstract provides an overview of the concept of fitness, focusing on the importance of physical activity and exercise for promoting health and well-being. It begins by defining fitness as the ability to perform physical tasks efficiently and highlights its relevance in modern society where sedentary lifestyle and health issues are prevalent.

Fitness is a dynamic and individualized concept, with different individuals having varying fitness levels, goals, and health conditions. It emphasizes the importance of consulting with healthcare professionals or qualified fitness experts to design a safe and effective fitness program tailored to individual needs and abilities.

This project aims to collecting the data from the source provided and analyse the data to gain insights into the relationship between physical and overall health.

We know that the popularity of the fitness tracker. Runners around the world are collecting the data in different gadgets.

For this purpose, we take a data for the year 2012 to 2019 data. We could analyse and visualize the data and answering the questions which is asked by the people.

**1. INTRODUCTION**

**1.1 About MedTourEasy**

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. MedTourEasy provides analytical solutions to our partner healthcare providers globally.

**1.2 About the Project**

For the fitness purpose there are many gadgets which are popular for collecting the data. The runners are collecting their data (smartphones and watches, etc.) to keep themselves motivated. They look for the answer the question 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. In this project we create import, clean, and analyze the data and answering the given questions.

**A black and red logo

Description automatically generated with medium confidence1.3 Objectives**

This project focuses on the analysing the stored the fitness data of a persons in the many different gadgets such as smartphones, watches, and wrist bands. We have the data in the form of CSV with which importing the data and analysing and visualizing the data and find the answer how the smartphones and watches are analyse the data. We use Python programming language for analysing the data.

We are looking for the question which is mentioned in the “About the Project” section.

Finding the answer of this common questions.

For that purpose, we use the jupyter notebook and python packages like pandas, matplotlib, warning and statsmodel.

We use pandas to read the dataset. Pandas have the capability to read different types of files such as CSV, excel, json, etc. We are visualizing the data using the most popular library of python that is matplotlib. Python also have the more libraries for plotting the graphs and charts such as seaborn and plotly, etc.

We finds the training data given in this fitness data set has to be analysed and whatever shortcomings are there in it, such as missing values in the data and showing what is found in the data through plots and charts.

**A black and red logo

Description automatically generated with medium confidence2. METHODOLOGY**

**2.1 Flow of the Project**

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

`A diagram of data preparation

Description automatically generated with medium confidence

*CREDIT: - XENOSTACK*

**2.2 Language and Platform used**

**Python programming language**:

Python is a popular, high-level, interpreted, and dynamically typed programming language known for its simplicity, readability, and versatility. It was created by Guido van Rossum and first released in 1991. Python supports multiple programming paradigms, including object-oriented, procedural, and functional programming. It has a large standard library and a thriving ecosystem of third-party libraries and frameworks, which make it highly adaptable for various domains, such as web development, data analysis, scientific computing, machine learning, and more. Python's popularity is attributed to its ease of use, extensive community support, and widespread adoption in diverse industries.

1. **A black and red logo

   Description automatically generated with medium confidence**Python is a versatile and powerful programming language known for its simplicity and readability, making it a popular choice for beginners and experienced developers alike.
2. Python has a large ecosystem of libraries and frameworks, including NumPy for scientific computing, Django for web development, and TensorFlow for machine learning, making it suitable for a wide range of applications.
3. Python supports multiple programming paradigms, including object-oriented, procedural, and functional programming, giving developers flexibility in writing code in different styles.
4. Python has a large and active community of developers who contribute to its development and create a wealth of resources, such as documentation, tutorials, and forums, making it easy to find support and learn from others.
5. Python has widespread adoption in various industries, including data science, artificial intelligence, web development, scientific research, and finance, making it a highly sought-after skill in the job market.

**Jupyter Notebook:**

Jupyter Notebook is an open-source, web-based application that allows users to create, share, and run live code, equations, visualizations, and narratives in a single interactive document. It supports over 100 programming languages, including Python, R, Julia, and more. Here are five advantages of using Jupyter Notebook:

1. Interactive Computing: Jupyter Notebook provides an interactive computing environment where users can write and execute code in individual cells, making it ideal for exploratory data analysis, prototyping, and iterative development. Users can also visualize data and immediately see the results in the same document.
2. Reproducible Research: Jupyter Notebook promotes reproducible research by allowing users to share live code, data, and visualizations in a single document. This enables others to reproduce the results, verify the findings, and build upon the work, fostering collaboration and transparency in scientific and data-driven research.
3. Rich Media Integration: Jupyter Notebook supports rich media integration, allowing users to include images, videos, LaTeX equations, and interactive visualizations alongside code and narratives. This makes it a powerful tool for creating interactive, data-driven narratives, reports, and presentations.
4. Flexibility and Extensibility: Jupyter Notebook is highly flexible and extensible, with a large ecosystem of extensions and plugins available. Users can customize the environment to suit their needs, integrate with other tools and libraries, and extend its capabilities with additional functionality.
5. **A black and red logo

   Description automatically generated with medium confidence**Wide Adoption and Community Support: Jupyter Notebook has gained widespread adoption across various domains, including data science, machine learning, scientific research, and education. It has a large and active community of users and developers who contribute to its development, provide support through forums and mailing lists, and create a wealth of resources, tutorials, and examples, making it a robust and well-supported tool for data analysis, research, and education.

**Packages in Python**

1. **Pandas**: It is very popular library for exploratory data analysis. Pandas is a popular open-source data manipulation and analysis library for Python. It provides data structures such as Series (1-dimensional labelled arrays) and DataFrame (2-dimensional labelled arrays) that are designed for efficient data handling, cleaning, and analysis. Pandas is widely used in data science, machine learning, and data analysis workflows, and it offers numerous functionalities for data manipulation, exploration, and transformation.
2. **Matplotlib**: Matplotlib is a popular open-source data visualization library for Python. It provides a comprehensive collection of functions for creating static, animated, and interactive visualizations of data in various formats, including line plots, scatter plots, bar plots, histograms, heatmaps, 3D plots, and more. Matplotlib allows users to create publication-quality plots with extensive customization options for colors, fonts, annotations, and other visual elements.
3. **Statsmodel**: statsmodels is a Python library that provides a wide range of statistical models and statistical analysis tools for data analysis, econometrics, and machine learning. It is built on top of NumPy, SciPy, and pandas, and offers a comprehensive set of statistical methods for regression analysis, time series analysis, hypothesis testing, and more.

**Installation**

pip install pandas

pip install matplotlib

pip install statsmodel

**Import Packages**

import pandas as pd

import matplotlib.pyplot as plt

import statsmodel.api as sm

A black and red logo

Description automatically generated with medium confidence**3. IMPLEMENTATION**

**3.1 Gathering Requirements**

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 the process of gathering information or data from various sources, such as surveys, observations, or online platforms. It involves systematically collecting, organizing, and storing data for analysis and decision-making purposes. Data importing refers to the process of bringing data from external sources into a designated system or software for further use or analysis.



In the above picture we import the pandas library after that we use the pandas to read the file using ***pd.read\_csv****.* The file is in the form of CSV, If the file is in the form of other formate then pandas also support reading the different file formate such as excel, json, etc. For this we use ***read\_exc****el*  to read the excel file and ***read\_json*** to read the json files.

After that using the dataframe.info() to gain the information about the dataset such as Columns, Not-Null count and Dtype(data type). Checking the data type is important, in this dataset the data type is already in the fixed. Lets see how many different types of the data type. List the following datatypes,

1. **Integer (int)**: Represents whole numbers without any decimal point, e.g., 42, -10, 0.
2. **Floating-point (float)**: Represents numbers with a decimal point, e.g., 3.14, -0.5, 2.0.
3. A black and red logo

   Description automatically generated with medium confidence**String (str)**: Represents sequences of characters, enclosed in single (' ') or double (" ") quotes, e.g., 'hello', "world", "42".
4. **Boolean (bool)**: Represents either True or False, used for logical operations, e.g., True, False.
5. **List**: Represents an ordered collection of items enclosed in square brackets ([]), where items can be of different datatypes, e.g., [1, 2, 3], ['apple', 'banana', 'cherry'].
6. **Tuple**: Similar to lists, but enclosed in parentheses (()), and immutable, meaning their values cannot be changed after creation, e.g., (1, 2, 3), ('a', 'b', 'c').
7. **Dictionary**: Represents a collection of key-value pairs enclosed in curly braces ({}) where values can be accessed using their corresponding keys, e.g., {'name': 'Alice', 'age': 30}.
8. **Set**: Represents an unordered collection of unique items enclosed in curly braces ({}) without any duplicate values, e.g., {1, 2, 3}, {'apple', 'banana', 'cherry'}.

These are some of the commonly used datatypes in Python, and there are other specialized datatypes available in libraries and modules for specific use cases.

**3.3 Data Preprocessing**

Data preprocessing is the initial step in data analysis and involves cleaning, transforming, and organizing raw data to prepare it for further analysis. It typically includes tasks such as handling missing values, removing duplicates, standardizing data formats, and normalizing or scaling data to improve accuracy and reliability of downstream analyses. Data preprocessing is essential to ensure data quality and to extract meaningful insights from the data.

**3.4 Data Cleaning**

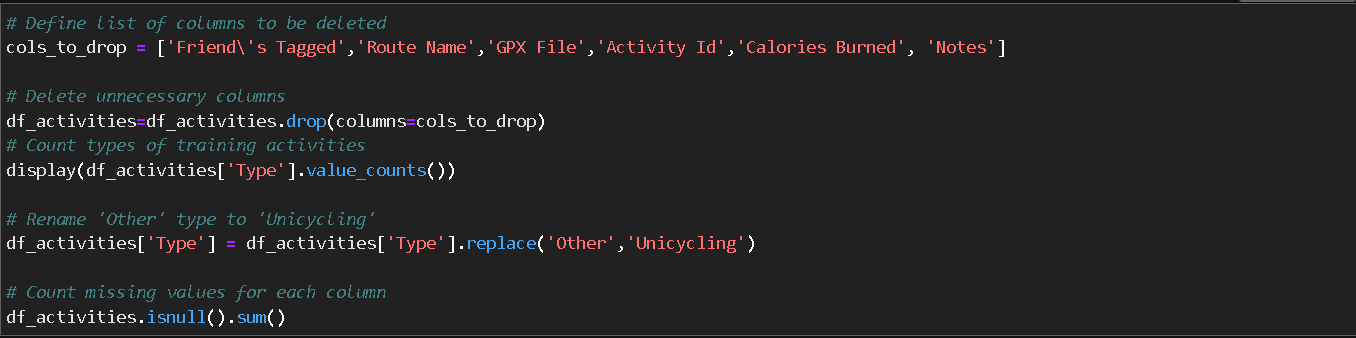
Data cleaning involves identifying and rectifying or removing errors, inconsistencies, and discrepancies in the data to ensure that it is accurate, reliable, and suitable for analysis or model training. This process may include tasks such as handling missing values, correcting data entry errors, standardizing formats, resolving inconsistencies, removing duplicate records, and validating data against predefined rules or business logic.

Function used:

**DataFrame.drop()**: To drop the column which is not used in analysis. For dropping the column using the parameter columns. In data frame dropping the column('Friend\'s Tagged','Route Name','GPX File','Activity Id','Calories Burned', 'Notes').

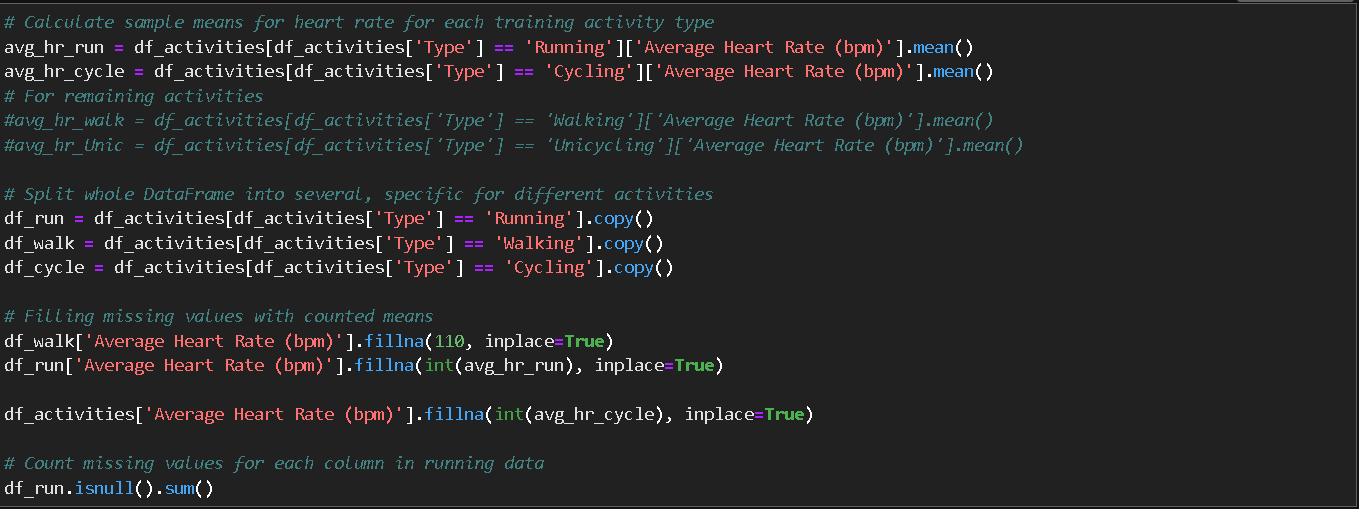
A black and red logo

Description automatically generated with medium confidence**DataFrame.replace():** To replace the value in the column. In the data frame replacing the value in the column Type from “Other” to “Unicycling”.

**DataFrame.isnull().sum()**: 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. In our data set, column Average heart Rate (bpm) contains 214 null value. 

**Dealing with missing values**: Dealing with missing values is a crucial step in data cleaning where incomplete or null values in data are addressed. It involves identifying and handling missing values through techniques such as imputation or deletion to ensure data accuracy. Appropriate handling of missing values is essential to avoid biased or incomplete analysis results.

In our dataset Average heart Rate (bpm) has null values. For this calculating the mean value of the column using **mean()** function. Now after calculating the mean value using the function **DataFrame[‘Average Heart Rate (bpm)’].fillna()** to fill the missing values. Then rechecking the missing value.



1. A black and red logo

   Description automatically generated with medium confidence**Plotting the Running Data**

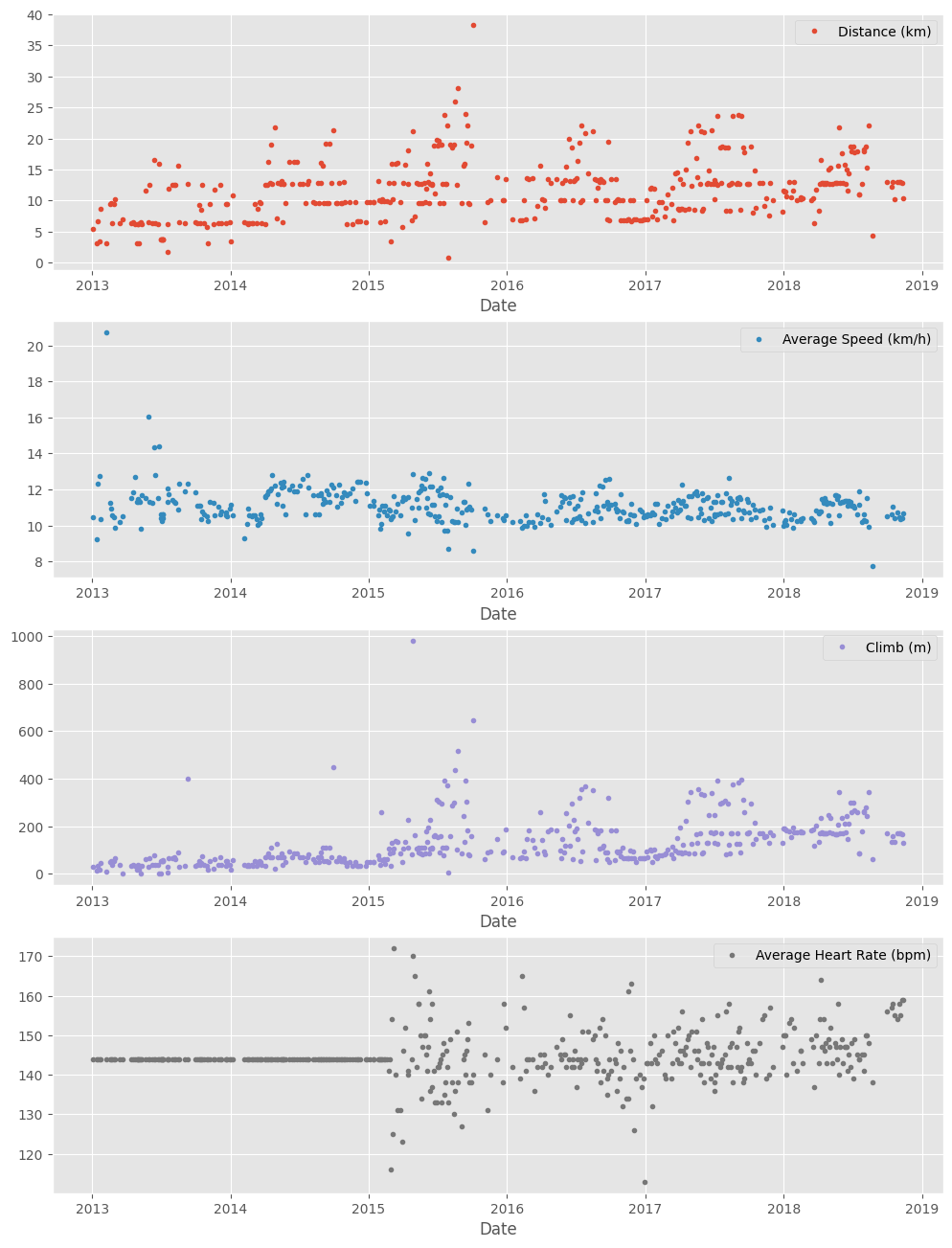
In the column Type most of the data were running(459). But before ploting the data first we import the required package and also filtering the date from 2013 to 2018.

Import packages:

**Matplotlib.pyplot:** This is the most popular library for ploting different type of visuals. We use the ‘ggplot’ style. And using the kind=’line’.

**%matplotlib inline** is a magic command in Python that is used in Jupyter notebooks to display matplotlib plots directly in the notebook output cell, allowing for inline visualization without the need to save plots to separate files. It enables the integration of matplotlib with Jupyter notebooks, making it convenient for exploratory data analysis and creating interactive data visualizations within the notebook environment.

**Output:**



1. **A black and red logo

   Description automatically generated with medium confidence**The highest distance covered in 2015 which is 38km.
2. The lowest distance is also covered in 2015 which is 0.7km.
3. The highest average speed in a year is covered in 2013.
4. The lowest average speed in a year is covered in 2018.
5. The highest climbing is covered in 2015 which is 982m.
6. The average heart rate is measured in 2015 which is 172 bpm.
7. **Running Statistics**:

Running is a popular form of exercise that offers numerous benefits for physical and menta well-being. When runners talk to each other about their hobbies, discuss our results and also discuss training strategies.

Runners common questions like:

* What is your avg. distance?
* How fast do you run?
* Do you measure your heart rate?
* How often do you train?

**resample()**:`resample()` is a function in Python used for time series data analysis that allows for changing the frequency or resampling of time series data. It is typically used to aggregate data at a lower or higher frequency, such as converting daily data to monthly data or vice versa. The function provides options for various resampling methods, such as mean, sum, or interpolation, to process the data during the resampling process.

We use resample() with alias ‘A’. which show annual. And finding the mean.

Code: **runs\_subset\_2015\_2018.resample(‘A’).mean()**

****

1. A black and red logo

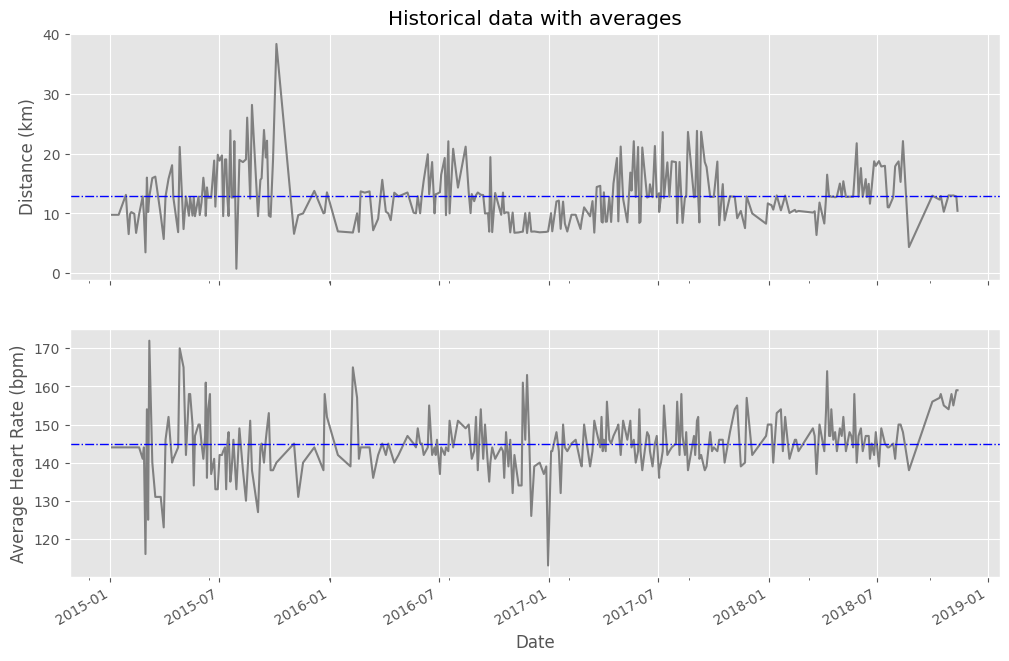
   Description automatically generated with medium confidence **Visualizing 2015 to 2018 data of distance and average heart rate with averages:**

Using the inline code of matplotlib

**Input:**

****

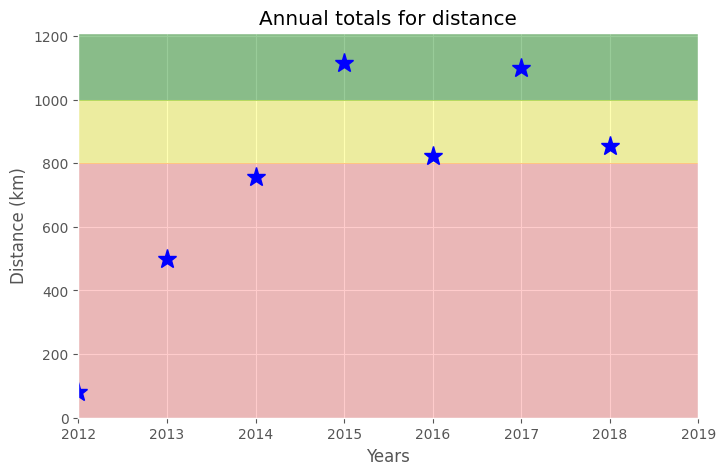
**Output:**

****

1. A black and red logo

   Description automatically generated with medium confidence **Did I Reach My Goals?**

In order to stay motivated and achieve my fitness goals, I decided to set a target of running 1000 km per year. To track my progress, I plotted a graph that shows my annual running distance (km) from 2013 to 2018, with the green region representing my target range. Only the stars falling within this green region indicate that I successfully achieved my goal for that year. This visualization provides a clear and visual way for me to assess my progress and keep myself accountable in pursuing my running target.



1. The target of running 1000 km per year was successfully achieved in the years 2015 and 2017.

1. A black and red logo

   Description automatically generated with medium confidence **Am I Progressing?**

To answer the questions, we’ll decompose the weekly distance run and visually compare it to the raw data.

A red trend line will represent the weekly distance run.

Package:

**Statsmodel** is a popular Python library for statistical modeling and econometrics. It provides a wide range of statistical models, including linear regression, time series analysis, and more. `statsmodels` is widely used for data analysis, modeling, and statistical inference in various fields such as finance, economics, social sciences, and more. It offers an extensive collection of tools and functions for statistical modeling, estimation, testing, and visualization, making it a powerful tool for data scientists and statisticians.

Using the statsmodel to decompose the weekly trend.

A picture containing text, screenshot, plot, diagram

Description automatically generatedWe see that the trend line increase from 2013 to 2015 and after that it moves constantly to 2018. So, we say that the running will higher than 10km per week from 2015 to 2018.

We also see that highest distance is covered in the year 2015.

After 2015 the progress will be constant.

1. A black and red logo

   Description automatically generated with medium confidence **Training Intensity**

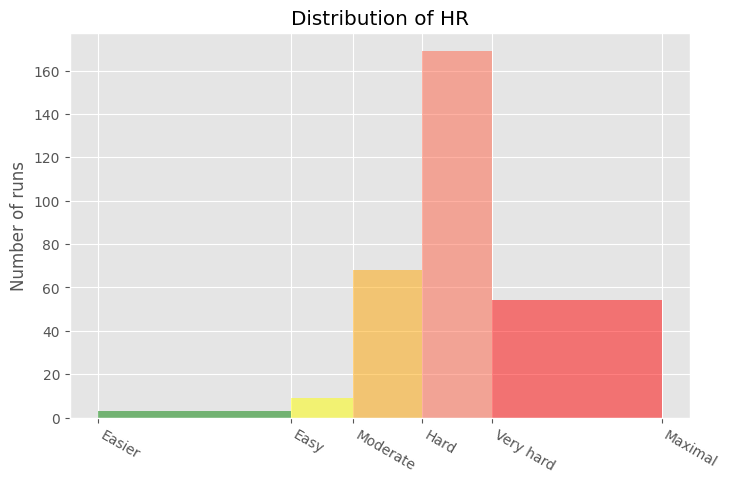
Heart rate is commonly used as a measure of training intensity, and different heart rate zones can be targeted based on age and fitness level to achieve specific training goals. The recommended target heart rate for moderate-intensity activities is generally around 50-70% of maximum heart rate, while for vigorous physical activity, it's typically around 70-85% of maximum heart rate. These heart rate zones are designed to help individuals optimize their training and improve overall fitness.

We will create a distribution plot of the heart rate data by training intensity.

**Input**:



**Output:**

****

1. A black and red logo

   Description automatically generated with medium confidenceLooking at the distribution plot of Heart Rate have high number of runs in the hard zone.
2. Easier zone have lowest heart rate.
3. **Summary**

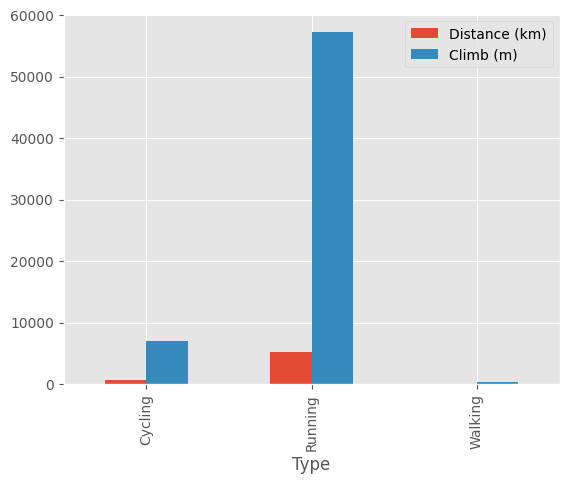
For the detailed summary we will create two tables. First table will be a summary of the distance (km) and climb (m) variables for each training activity. The second table will be the list of summary statistics for the average speed (km/h), distance(km) and climb(m) variables for each training activity.

**First table: Second table:**

**A screenshot of a computer

Description automatically generated with medium confidence**

**First table diagram:**

****

**A black and red logo

Description automatically generated with medium confidenceEnd Notes**

1. Running has the highest distance covered (5224.50 km), followed by cycling (680.58 km) and walking (33.45 km).
2. However, cycling has the highest total climb (6976 m), followed by running (57278 m) and walking (349 m).
3. Running covers the highest distance with the least amount of climb, indicating it may be the most efficient mode of exercise in terms of distance covered per climb.
4. Walking covers the least distance and has the least amount of climb, indicating it may be the least intense mode of exercise compared to cycling and running.
5. Cycling covers a moderate distance but has a relatively high amount of climb, indicating it may be a good exercise for building strength and endurance.
6. **Fun Facts**

Let’s pick some fun facts from the summary.

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 fact**

-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

**Forest Run Story**

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

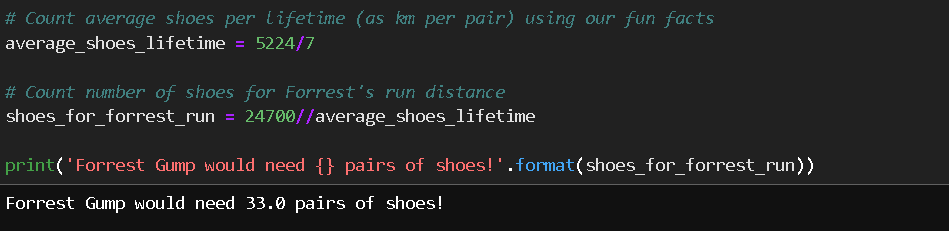
A black and red logo

Description automatically generated with medium confidence**Forest Map**



**Question**

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



First we count average shoes per lifetime using fun facts.

After that we count the shoes for the forests run distance using the forest run distance.

**Answer**

We need 33 pairs of shoes forest run.

A black and red logo

Description automatically generated with medium confidence

Thank You!