Final Report

on

“Analyze your Runkeeper Fitness Data.”

MedTourEasy

From: Stefan Lacher, BA

Internship Program: Data Scientist Trainee

March 2021

Acknowledgments

At the beginning I would like to thank the MedTourEasy team for giving me the opportunity to do an internship with you. Also, thank you very much for the instructive time and opportunity to develop in the area of "Data Science".

During the internship at MedTourEasy, I was able to complete the “Data Scientist with Python” training and apply the skills I had learned with the “Analyze your Runkeeper Fitness Data” project. The internship helped me to expand my knowledge in the areas of data manipulation, data visualization, probability & statistics, importing & cleaning data and machine learning with the help of Python.

Thank you very much!

Abstract

The training analysis is particularly important for athletes. With the help of GPS fitness trackers, progress can be analyzed and viewed.

In this project, fitness data between 2012 and 2018 are recorded and analyzed with the help of Runkeeper. This is done with 11 tasks that indicate the way for the analysis. Sports such as running, cycling or walking are considered and questions such as “Did I reach my goals?” or “Am I progressing?” are answered.

Inhaltsverzeichnis

1 Introduction

As part of the “Data Scientist” internship, a project entitled “Analyze your Runkeeper Fitness Data” was carried out for “MedTourEasy”. "MedTourEasy" is a healthcare company that offers a platform for second opinions in the healthcare sector (see MedTourEasy.com).

The project is about fitness trackers. In recent years these devices have become more popular, for example runners all over the world are collecting data to make further progress. The questions to be answered in this project are:

How fast, long and intense was my run today?

Have I achieved my training goals?

Am i making progress?

What have been my best achievements?

How am I compared to others?

Within seven years, Runkeeper exported training data and saved it in a CSV file. These are to be analyzed in this project. (See Pavlenko 2021.)

In addition to answering the questions, the aim of this project is to create an exciting display of training data with Python, which is then interpreted.

2 Methodology

As mentioned earlier, this project uses Python. In order to be able to carry out the presentation and analysis of the data better, further packages are used. These are pandas, matplotlib and statsmodels.

Python is a universal, interpreted, high-level programming language that claims to be easy to read. In addition, the programming language is open source. (See python.org.) Another open source software is Jupyter Notebook. This software serves as a notebook for a wide variety of projects in the fields of data science and scientific computing. (See jupyter.org.) In this project Jupyter Notebook is used.

The packages used are discussed below. Pandas is a program library for Python that is used to manage and analyze data (see pandas.pydata.org). Matplotlib is also a program library that enables data to be represented (see matplotlib.org). Statsmodels is a Python package that can be used to estimate statistical models and perform statistical tests (statsmodels.org).

Now that the technical perspective has been considered, the procedure is explained below. This project is guided with the help of tasks. These are as follows:

1. Obtain and review raw data

2. Data preprocessing

3. Dealing with missing values

4. Plot running data

5. Running statistics

6. Visualization with averages

7. Did I reach my goals?

8. Am I progressing?

9. Training intensity

10. Detailed summary report

11. Fun facts

In the next chapter, the procedure and implementation of the tasks that have already been described are explained in detail.

3 Implementation

The implementation is the chapter where the procedure is explained in detail step by step.

3.1 Obtain and review raw data

It started with the import of pandas. The code “import pandas as pd” was used here. This package is required because it is used to import the data set, the CSV file. This is done with the line “pd.read_csv ()”. In the same step, the data record is saved as a data frame “df_activities”. To get a first insight into the data, the first three rows and all columns are shown. This can be seen in Table 1.

	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
Date													
2017-01-12 18:19:37	1a0c5ffe-b6ef-4c05-8126-acd92ab1b5a6	Running	NaN	12.00	1:08:08	5:41	10.57	860.0	100	148.0	NaN	TomTom MySports Watch	2017-01-12-181937.gpx
2017-09-05 18:34:21	5fb52c7a-0331-4382-9636-b9fca42a797f	Running	NaN	12.77	1:06:15	5:11	11.56	896.0	127	152.0	NaN	TomTom MySports Watch	2017-09-05-183421.gpx
2015-06-21 17:53:46	ee566558-df2c-4874-8470-4c96bc16e11b	Cycling	NaN	34.49	1:35:39	2:46	21.63	751.0	318	NaN	NaN	NaN	2015-06-21-175346.gpx

table 1 Sample of the Dataset

Here are three rows at different times. The columns can also be viewed here. These are 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 and GPX file. The ".info ()" method was then used to obtain further information about the data. Here it can be said that there are 508 rows and 13 columns. In addition, three data types could be considered, float, integer and object.

3.2 Data preprocessing

First of all, the column names are considered, which do not have to be renamed as they are informative. When looking at the info () method again, missing values were found. Since not all columns are important for the analysis, these are removed. These include Friend's Tagged, Route Name, GPX File, Activity Id,

Calories Burned and Notes. To remove these,.drop ’is used. Consequently, the different training activities and the number of times they are performed are considered. These are running, cycling, walking and other. Since Other is not very meaningful, this is renamed to Unicycling. Missing values are checked again and in this case only the Average Heart Rate (bpm) is displayed.

3.3 Dealing with missing values

Since the missing values of Average Heart Rate (bpm) can no longer be obtained, they are filled with an average value. This process is called mean imputation.

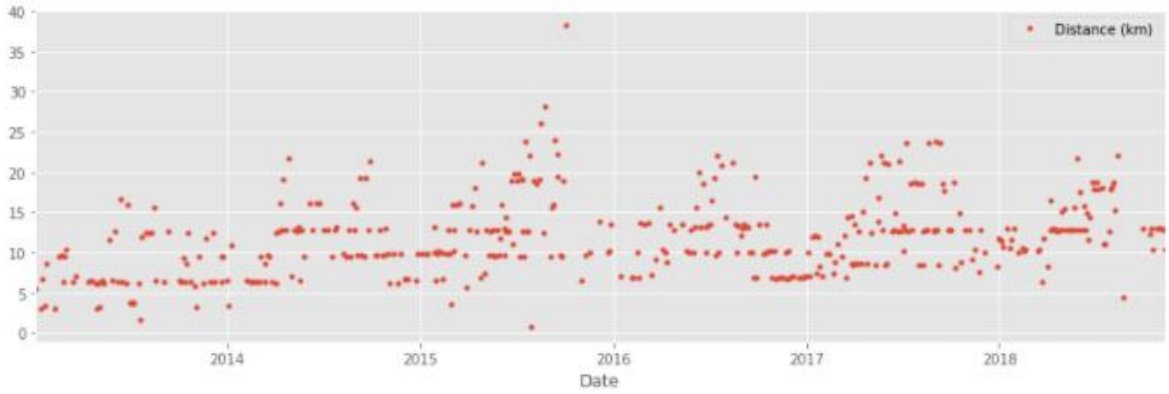
This is adapted to each type of training. It starts with the filtering of the activity and then the mean value of the heart rate is calculated. An example code:

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

The Dataframe is then divided into several according to the training activity and the missing values are filled with the calculated one. Finally, it is checked again for missing values and it is determined that there are no more.

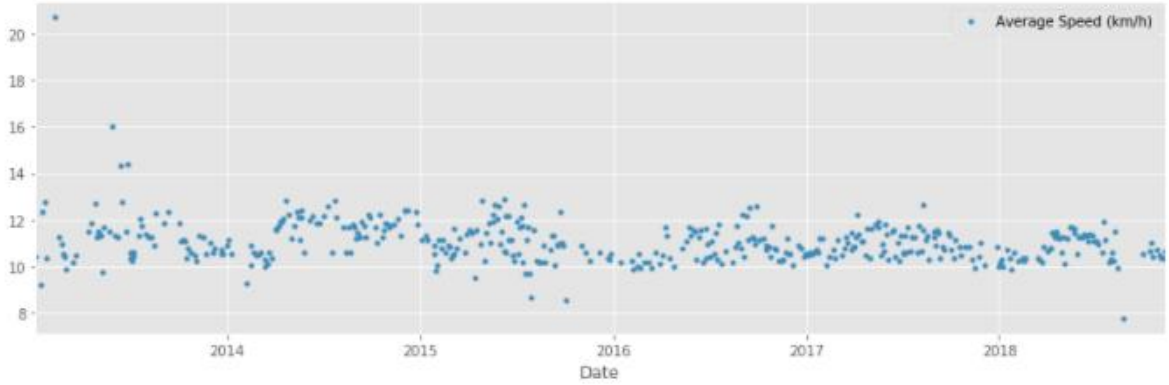
3.4 Plot running data

In the next step the running data are displayed. To do this, matplotlib.pyplot is imported as plt. The data is then narrowed down for the period between 2013 and 2018 and then presented as subplots. For all subplots, the X-axis is the date between 2013 and 2018, the Y-axis varies.



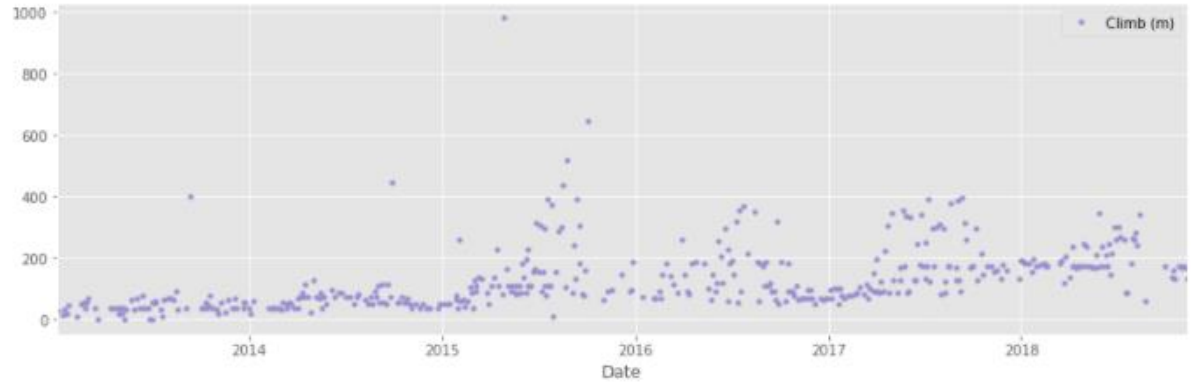
graph 1: Running Distance (km)

In this subplot you can see that the distance was smallest in the first year and has increased over the years. In 2015 there was an outlier where between 35 and 40 km was run once.



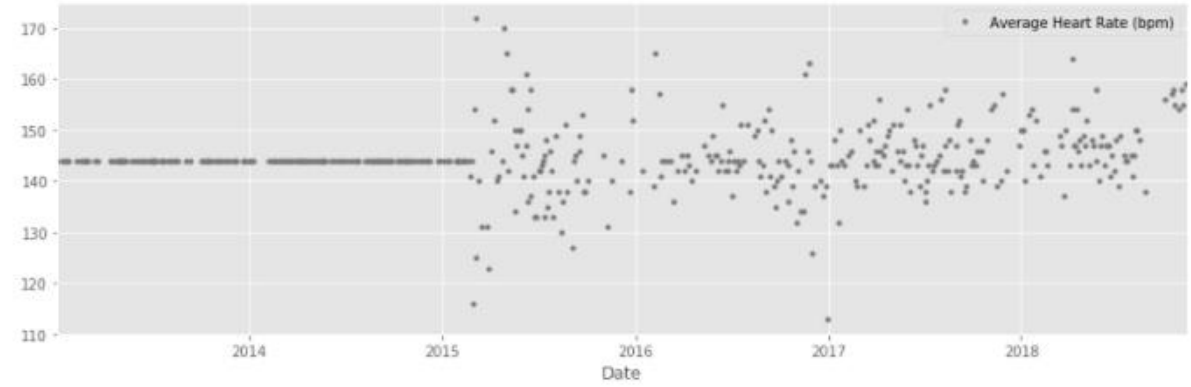
graph 2: Running Average Speed (km/h)

With regard to the average speed in km/h, it can be said that this has remained almost the same over the entire period. At the beginning an attempt was made to run faster, but the speed then remained on average between 10 and 12 km/h.



graph 3: Running Climb (m)

When climbing in m it can be seen that there has been an increase over the years. At the beginning it was between 0 and 200 meters, from 2015 it went higher more often.



graph 4: Running Average Heart Rate (bpm)

Interestingly, the Average Heart Rate in bpm between 2013 and 2014 is the average between 140 and 150 bpm. From 2015 there is a large spread, but the value remains on average between 140 and 150 bpm.

3.5 Running statistics

Frequently, runners are asked questions such as:

What's your average distance?

How fast are you running

Do you measure your heart rate?

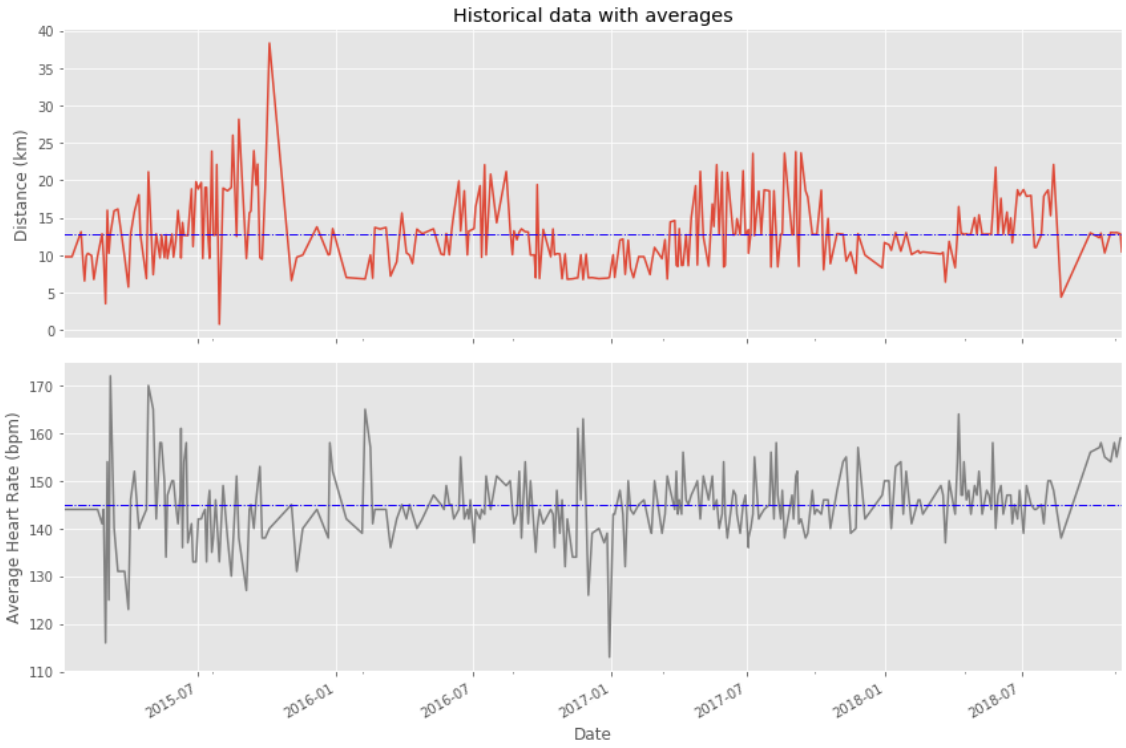
How often do you workout?

These questions are answered in this chapter.

We start with the preparation of the data by restricting them to the last 4 years 2015 to 2018. Then the Pandas method resample () is used, which carries out a grouping. This is broken down into years and weeks. The result shows that an average of 12.8 km per year is run with an incline of 163 m, a speed of 10.9 km/h and a heart rate of 145 bpm. When looking at the average data by weeks, the numbers change slightly. The distance is 12.5 km, an incline 158 m, speed 10.8 km/h and the heart rate 144.8 bpm. At the end it is calculated how many trainings were carried out per week, these are 1.5.

3.6 Visualization with averages

In the next step, the long-term distance and heart rate data are visually compared. Here, too, the data are prepared, starting with the time limitation between 2015 and 2018, followed by the creation of the distance and heart rate variables. Then 2 subplots are created, one looking at the distance and the other looking at the heart rate.

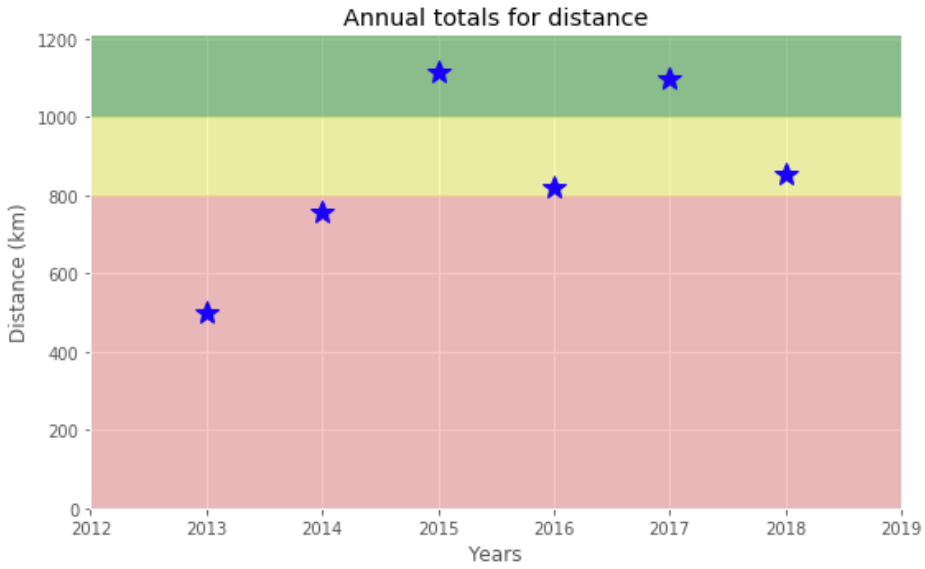


graph 5: Historical data with averages, Distance and Average Heart Rate

It can be seen here that the average heart rate over the period towards the end of the measurement had fewer outliers than at the beginning. Interestingly, it can be seen that fewer km were run in the end, but the heart rate was higher.

3.7 Did I reach my goals?

Since a comparison between the years is particularly interesting, this is visualized. The goal was to run 1000 km per year.

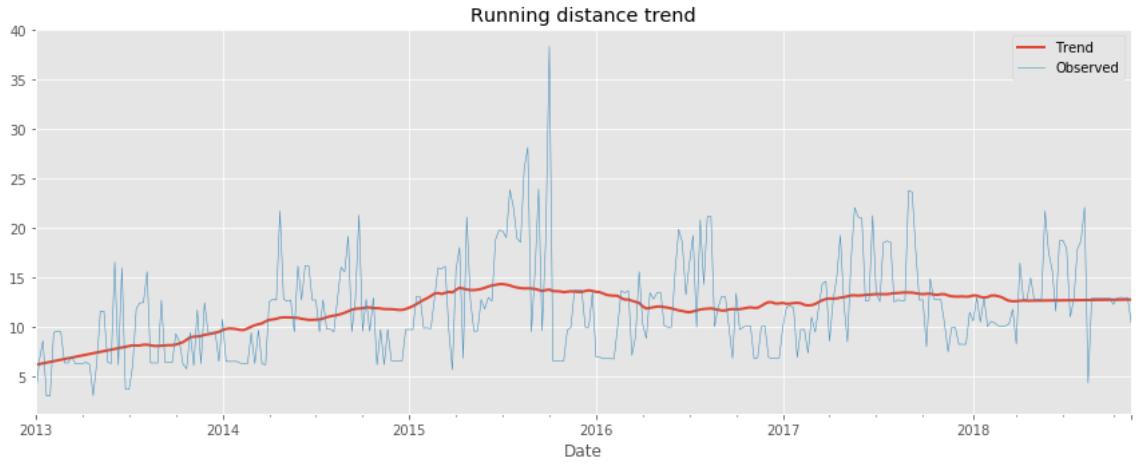


graph 6: Annual totals for distance

In this plot it can be seen that the red area is between 0 and 800 km, the yellow area between 800 and 1000 km and the green area between 1000 and 1200 km. The goal has only been achieved when the blue stars are in the green area. Thus, it can be said that in 2015 and 2017 the target was achieved, in 2016 and 2018 the target was almost achieved, and in 2013 and 2014 the target was not achieved.

3.8 Am I progressing?

So, the next question arises, namely whether one can see progress in terms of running skills. To answer this, the weekly run data is analyzed. These are compared on the one hand with a trend line in red and on the other hand with the observed data in blue. To make this possible, the statsmodels.api is used.



graph 7: Running distance trend

It can be seen here that there was an increase in km between 2013 and 2016 and that the km remained constant with slight fluctuations from 2016 to 2019. So it can be said that no progress is being made at the moment.

3.9 Training intensity

In the next step, the training intensity is considered. The heart rate is analyzed in this regard. To illustrate this graphically, a plot is created, whereby easy workouts are shown as green, moderate as yellow, hard, as orange, very hard pink and maximally as red.

graph 8: Distribution of heard rate

In this regard, it can be said that easy and moderate workouts were very seldom run. A very hard workout was performed most often with almost 160 runs. Then the training was hard and with a maximum of just over and under 60 runs.

3.10 Detailed summary report

In the last step, detailed overview tables are created. Two tables are created for this, the first containing a summary of distance in kilometers and incline in meters for cycling, running and walking. The second deals with summarizing statistics regarding average speed in km/h, ascent in m and distance in km.

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

table 2: Totals for different training types

It can be seen here that a total of 680 km and 6976 meters of altitude was covered for cycling, 5224 km and 57278 meters of altitude for running, and 33 km and 349 meters of altitude for walking. In the following, the table with the summarizing statistics will be looked at.

		Average Speed (km/h)	Climb (m)	Distance (km)
Type				
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

table 3: Summary statistics for different training types

First, the type of cycling is considered. It can be said that the maximum speed is 24 km/h and the minimum speed is 11 km/h. The average speed is 19 km/h. The maximum gradient was 553 m and the minimum gradient 58 m. Here, too, the average can be considered, which is 240 m. The maximum distance in km is 49 and the minimum 11 km. An average of 23 km was driven.

When running, the maximum speed is 20 km/h and the minimum speed is 5 km/h, with an average of 11 km/h. The maximum climb is 982 m, the minimum climb is 0 and the average is 124 m. The furthest distance is 38 km, the smallest is 0.76 km and the average is 11 km.

Looking at walking, it can be said that the maximum speed is 7 km/h, the minimum 1 km/h, but the average is 5 km/h. When walking, a maximum of 112 m, a minimum of 5 m and an average of 19 m gradient was covered. If the distance is considered, it can be seen that a maximum of 4 km, a minimum of 1.2 km and an average of 1.86 km have been completed.

3.11 Fun facts

Finally, some interesting facts from the data are considered. For example, 7 running shoes were used for the runs. Since some data is known from Forrest Gump, such as the total number of kilometers of 24,700 km, it would be interesting how many shoes he needed. Since the 7 shoes in the data were used for 5224 km, the lifespan of a shoe is calculated and then Forrest Gump outputs the number of shoes. These calculations result in 33 running shoes.

4 Conclusion

Many athletes in particular want to analyze their training in order to be able to make and observe progress. GPS fitness trackers are particularly helpful here.

This project analyzed the fitness data between 2012 and 2018 that was recorded using Runkeeper. Different areas were considered, such as running, cycling or walking. Different data such as the average heart rate, distances or speeds were illustrated using different display options. Questions like “Did I reach my goals?” Or “Am I progressing?” Were answered.

In the future, it would make sense to update the data over time or, for example, to consider other sports.

5 References

Jupyter.org 2021: About Us. Some information about the Jupyter Project and Community. https://jupyter.org/about (accessed on 19.03.2021).

Matplotlib.org: History. https://matplotlib.org/stable/users/history.html (accessed on 19.03.2021).

MedTourEasy.com 2021: MedTourEasy. Connecting Patients Worldwide. https://www.medtoureasy.com/ (accessed on 19.03.2021).

Pandas.pydata.org 2021: About pandas. https://pandas.pydata.org/about/ (accessed on 19.03.2021).

Pavlenko, Andrii 2021: Analyze your Runkeeper Fitness Data. https://learn.datacamp.com/projects/727 (accessed on 19.03.2021).

Python.org 2021: About Python. https://www.python.org/about/ (accessed on 19.03.2021).

Statsmodels.org 2021: About statsmodels. https://www.statsmodels.org/stable/about.html (accessed on 19.03.2021).

6 List of graphs

7 List of tables