Final Report of Traineeship Program 2024

On

"Running Performance Analysis Using Fitness Tracker Data"

By

Viraj Balasaheb Patil

MEDTOUREASY

28th June 2024

ACKNOWLDEGMENTS

The traineeship opportunity at MedTourEasy provided an excellent platform for me to learn and understand the complexities of Data Visualizations in Data Analytics. This experience significantly contributed to both my personal and professional growth. I am deeply grateful for the chance to interact with numerous professionals who guided me throughout the project, making it a valuable learning experience.

Firstly, I would like to extend my heartfelt gratitude to the Training & Development Team at MedTourEasy for offering me the opportunity to undertake my traineeship at their esteemed organization. I am especially thankful for their efforts in helping me comprehend the intricacies of the Data Analytics profile and providing training that enabled me to execute the project effectively and ensure client satisfaction. I appreciate the time they devoted to me despite their busy schedules.

Additionally, I want to thank the entire MedTourEasy team and my colleagues for creating a productive and supportive working environment.

Table of Contents

Section	Subsection	Description
10. References		
11. Fun Facts		

Abstract

This project focuses on analyzing fitness tracker data to enhance running performance. With the surge in popularity of fitness trackers, runners worldwide are collecting vast amounts of data through devices such as smartphones and watches. This data, exported from Runkeeper in CSV format, provides insights into various aspects of running activities, including speed, distance, intensity, and heart rate.

The primary objectives of this project include importing, cleaning, and preprocessing the fitness tracker data, followed by a thorough analysis to answer key questions related to running performance. The analysis aims to evaluate whether training goals were met, assess progress over time, identify best achievements, and compare performance with peers.

Data visualization plays a crucial role in this project, helping to identify trends and patterns through various graphical representations. Visual tools such as line charts, bar graphs, and histograms are used to depict trends in distance and heart rate, compare annual distance totals with set goals, and illustrate the distribution of different metrics.

The findings from the analysis are used to assess progress towards training goals and provide recommendations for future training. The project concludes with a detailed summary of the results, highlighting key insights and offering strategies for improving running performance based on data-driven evidence.

This project not only demonstrates the practical application of data analytics and visualization techniques but also underscores their importance in achieving fitness goals and enhancing athletic performance.

1. Introduction

About Company

MedTourEasy, a global healthcare company, provides you the informational resources needed to evaluate your global options. MedTourEasy provides analytical solutions to our partner healthcare providers globally.

Overview of the Project

With the surge in the popularity of fitness trackers, runners around the world are leveraging gadgets like smartphones and smartwatches to monitor their activities. This project aims to analyze data exported from Runkeeper to answer key questions about running performance and progress.

Importance of Analyzing Fitness Tracker Data

Fitness tracker data provides valuable insights into various aspects of running, such as speed, distance, intensity, and overall progress. By analyzing this data, runners can set realistic goals, track their achievements, and make informed decisions to improve their training routines.

Objectives of the Report

Main Objectives of the Report

1. Import, Clean, and Preprocess the Fitness Tracker Data
- Importing Data: Load the fitness tracker data from Runkeeper CSV files into a structured format using a DataFrame, ensuring it's ready for analysis.
- Cleaning Data: Remove unnecessary columns, handle missing values, and correct any inconsistencies to prepare the data for accurate analysis.
- Preprocessing Data: Transform the data into a usable format, including parsing dates, setting appropriate indices, and converting data types as needed.

2. Analyze the Data to Answer Questions Related to Running Performance
- Performance Metrics: Calculate and analyze key metrics such as speed, distance, duration, and heart rate to evaluate running performance.
- Achievement Analysis: Determine whether training goals were met by comparing actual performance metrics to set targets.
- Progress Evaluation: Assess improvements over time by examining trends and patterns in the data.

3. Visualize the Data to Identify Trends and Patterns
- Trend Analysis: Use line plots, bar charts, and other visual tools to identify trends over time, such as changes in running distance, speed, and heart rate.
- Distribution Analysis: Create histograms and density plots to understand the distribution of various metrics and identify any outliers or anomalies.
- Goal Comparison: Visualize the comparison between actual performance and set goals to clearly see how well goals were achieved.

4. Assess Progress Towards Training Goals
- Goal Achievement: Evaluate the extent to which annual and weekly training goals were met by analyzing performance data.
- Long-term Progress: Analyze performance over different time periods to understand long-term trends and progress, identifying areas of improvement.

5. Provide Recommendations Based on the Analysis
- Training Adjustments: Offer suggestions for modifying training routines based on the analysis to help align with fitness goals.
- Goal Setting: Provide recommendations for setting realistic and achievable goals based on past performance and identified trends.
- Performance Improvement: Give tips and strategies for improving specific aspects of running, such as increasing distance or speed, based on data-driven insights.

2. Background

History of Fitness Trackers

Fitness trackers have evolved significantly over the past decade. Initially, they were simple pedometers, but today, they are sophisticated devices capable of tracking a wide range of activities and metrics, including heart rate, GPS location, and even sleep patterns.

Importance of Data Collection in Fitness and Running

Collecting data on fitness activities allows individuals to quantify their performance and progress. It helps in setting measurable goals, staying motivated, and making adjustments to training plans based on objective feedback.

3. Data Collection

Source of the Data (Runkeeper CSV File)

The dataset used in this project was exported from Runkeeper, a popular fitness tracking app. The data is stored in a CSV file format, where each row represents a single training activity.

Description of the Dataset

The dataset includes various columns capturing details of each training activity, such as date, type of activity, distance, average speed, and heart rate.

Variables Included in the Dataset

- Date
- Type
- Distance (km)
- Duration
- Average Speed (km/h)
- Average Heart Rate (bpm)
- Climb (m)
- Calories Burned (excluded in analysis)
- Notes (excluded in analysis)

4. Methodology

Data Preprocessing

Data Import

The data was imported into a Pandas DataFrame using the read_csv function with parse_dates and index_col parameters to ensure the 'Date' column was correctly parsed as dates and set as the index.

```
import pandas as pd

runkeeper_file = '/content/cardioActivities.csv'
df_activities = pd.read_csv(runkeeper_file, parse_dates=['Date'], index_col='Date')
```

Cleaning Procedures

Unnecessary columns were identified and dropped to focus the analysis on relevant metrics.

```
cols_to_drop = ['Friend\'s Tagged', 'Route Name', 'GPX File', 'Activity Id', 'Calories Burned', 'Notes']
df_activities.drop(columns=cols_to_drop, inplace=True)
```

Handling Missing Values

Missing values in the 'Average Heart Rate (bpm)' column were handled by filling them with the mean heart rate for each type of activity.

```
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()

df_walk['Average Heart Rate (bpm)'].fillna(110, inplace=True)
df_run['Average Heart Rate (bpm)'].fillna(int(avg_hr_run), inplace=True)
df_cycle['Average Heart Rate (bpm)'].fillna(int(avg_hr_cycle), inplace=True)
```

Tools and Software Used

The analysis was conducted using Python, primarily with the Pandas library for data manipulation and Matplotlib for visualization. Other tools included Statsmodels for time series analysis and the warnings module to handle plotting warnings. Google Colab was used as the development environment for its ease of use and collaboration features.

5. Data Analysis

Analysis Approach Overview

The analysis involved summarizing the running data, creating visualizations to identify trends, and comparing the results against personal training goals.

Summary Statistics of Running Data

The dataset was split into specific DataFrames for different activities (Running, Walking, Cycling) and summary statistics were calculated for each.

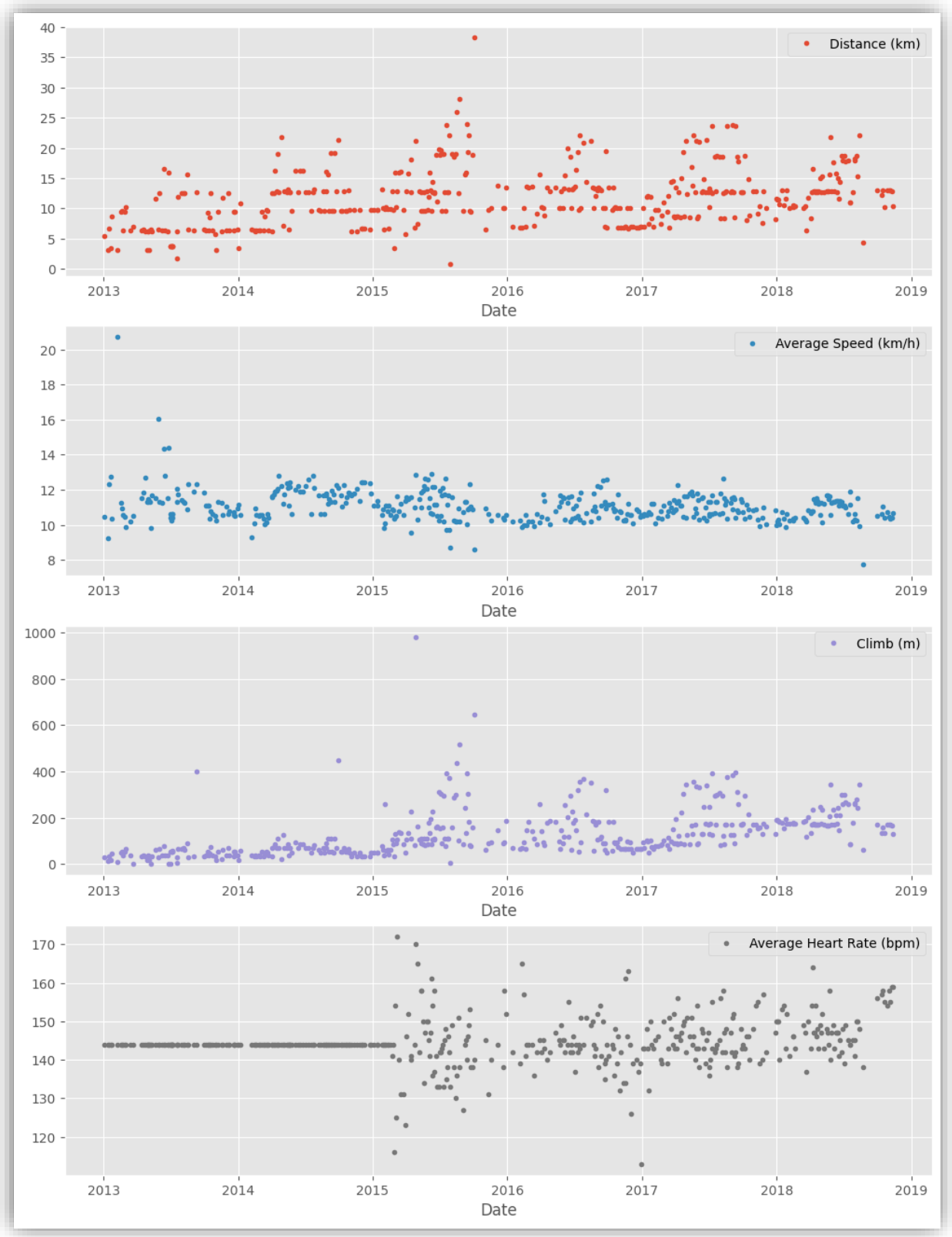
```
df_run = df_activities[df_activities['Type'] == 'Running'].copy()
df_walk = df_activities[df_activities['Type'] == 'Walking'].copy()
df_cycle = df_activities[df_activities['Type'] == 'Cycling'].copy()

df_run.describe()
```

Visualizations of Running Activities

Various visualizations were created to understand the data better. For instance, subplots of running metrics over time, annual and weekly statistics, and histograms of heart rate distribution.

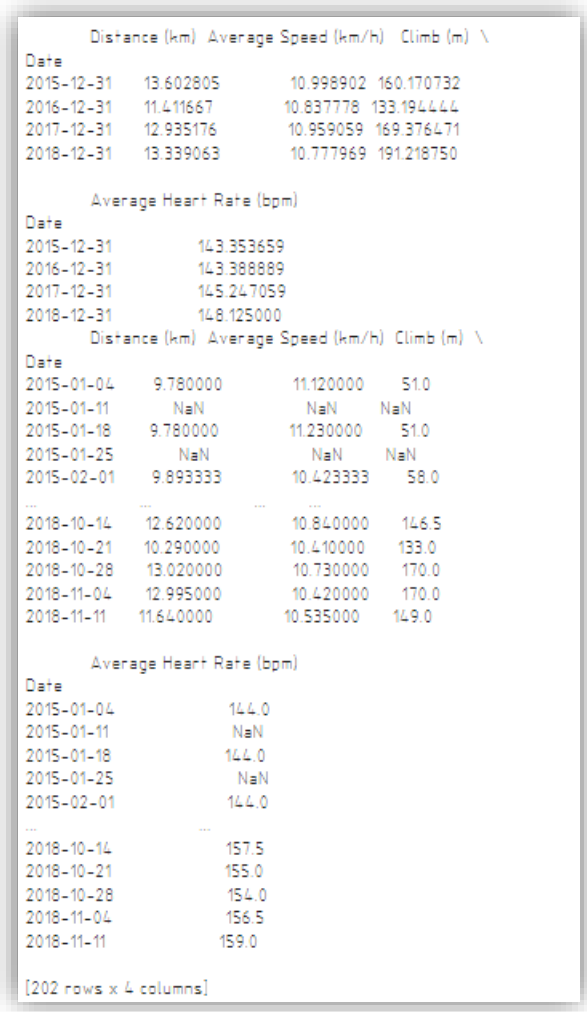
Subplots of Running Metrics Over Time

runs_subset_2013_2018.plot(subplots=True, sharex=False, figsize=(12,16), linestyle='none', marker='o', markersize=3)
plt.show()



Annual and Weekly Statistics

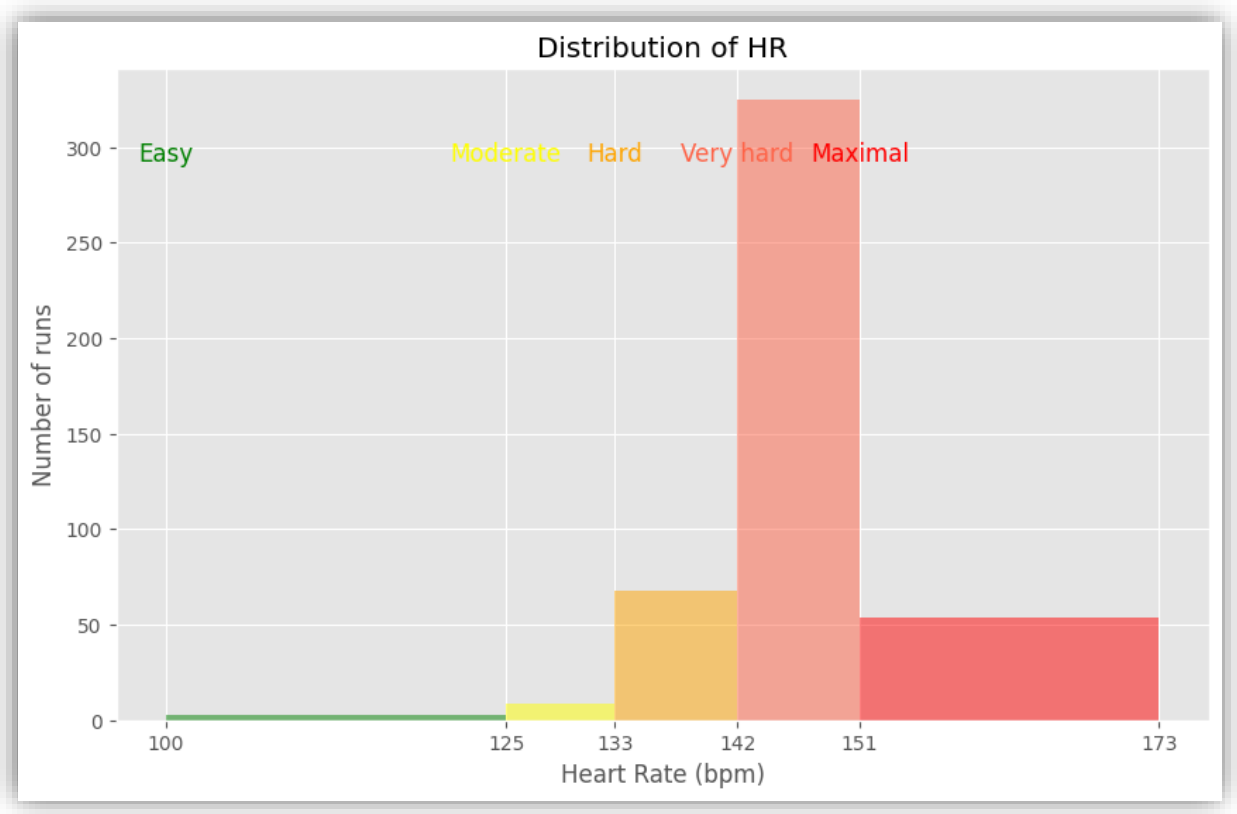
annual_stats = runs_subset_2015_2018[numeric_cols].resample('A').mean()
weekly_stats = runs_subset_2015_2018[numeric_cols].resample('W').mean()



```
            Distance (km)  Average Speed (km/h)   Climb (m)  \
Date
2015-12-31      13.602805            10.998902  160.170732
2016-12-31      11.411667            10.837778  133.194444
2017-12-31      12.935176            10.959059  169.376471
2018-12-31      13.339063            10.777969  191.218750

            Average Heart Rate (bpm)
Date
2015-12-31                143.353659
2016-12-31                143.388889
2017-12-31                145.247059
2018-12-31                148.125000
            Distance (km)  Average Speed (km/h)  Climb (m)  \
Date
2015-01-04       9.780000            11.120000       51.0
2015-01-11            NaN                  NaN        NaN
2015-01-18       9.780000            11.230000       51.0
2015-01-25            NaN                  NaN        NaN
2015-02-01       9.893333            10.423333       58.0
...                   ...                  ...        ...
2018-10-14      12.620000            10.840000      146.5
2018-10-21      10.290000            10.410000      133.0
2018-10-28      13.020000            10.730000      170.0
2018-11-04      12.995000            10.420000      170.0
2018-11-11      11.640000            10.535000      149.0

            Average Heart Rate (bpm)
Date
2015-01-04                144.0
2015-01-11                  NaN
2015-01-18                144.0
2015-01-25                  NaN
2015-02-01                144.0
...                         ...
2018-10-14                157.5
2018-10-21                155.0
2018-10-28                154.0
2018-11-04                156.5
2018-11-11                159.0

[202 rows x 4 columns]
```

Heart Rate Distribution

```
fig, ax = plt.subplots(figsize=(10, 6))
n, bins, patches = ax.hist(df_run_hr_all, bins=hr_zones, alpha=0.5)
for i in range(0, len(patches)):
    patches[i].set_facecolor(zone_colors[i])
```



Comparison with Personal Training Goals

Annual totals for distance were compared against predefined goals to evaluate progress.

```
df_run_dist_annual = df_run['2013':'2018'].resample('A')['Distance (km)'].sum()
```

Progress Analysis Over Time

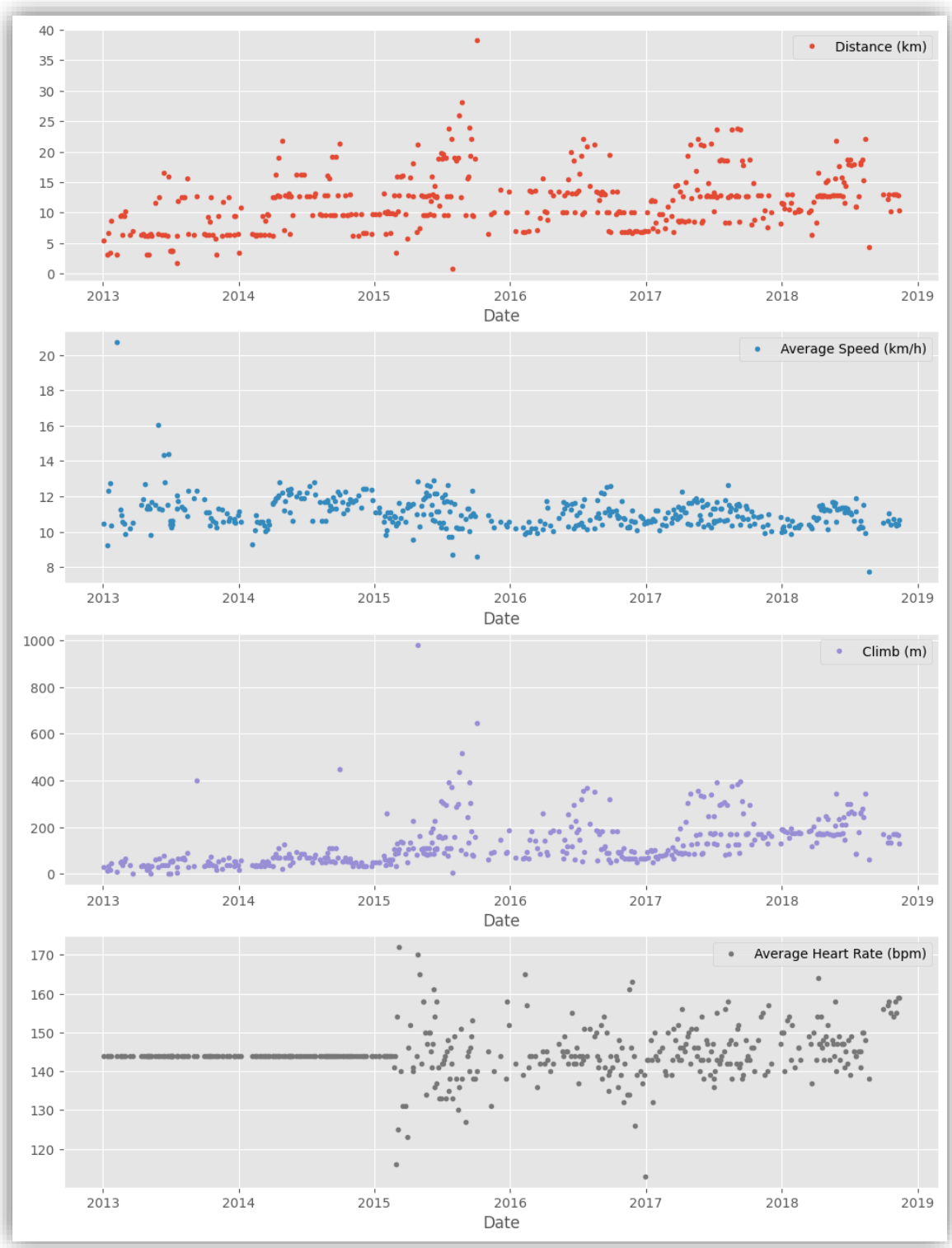
Seasonal decomposition was applied to weekly running distance data to identify trends and patterns over time.

```
decomposed = sm.tsa.seasonal_decompose(df_run_dist_wkly, extrapolate_trend=1, period=52)
```

6. Results and Discussion

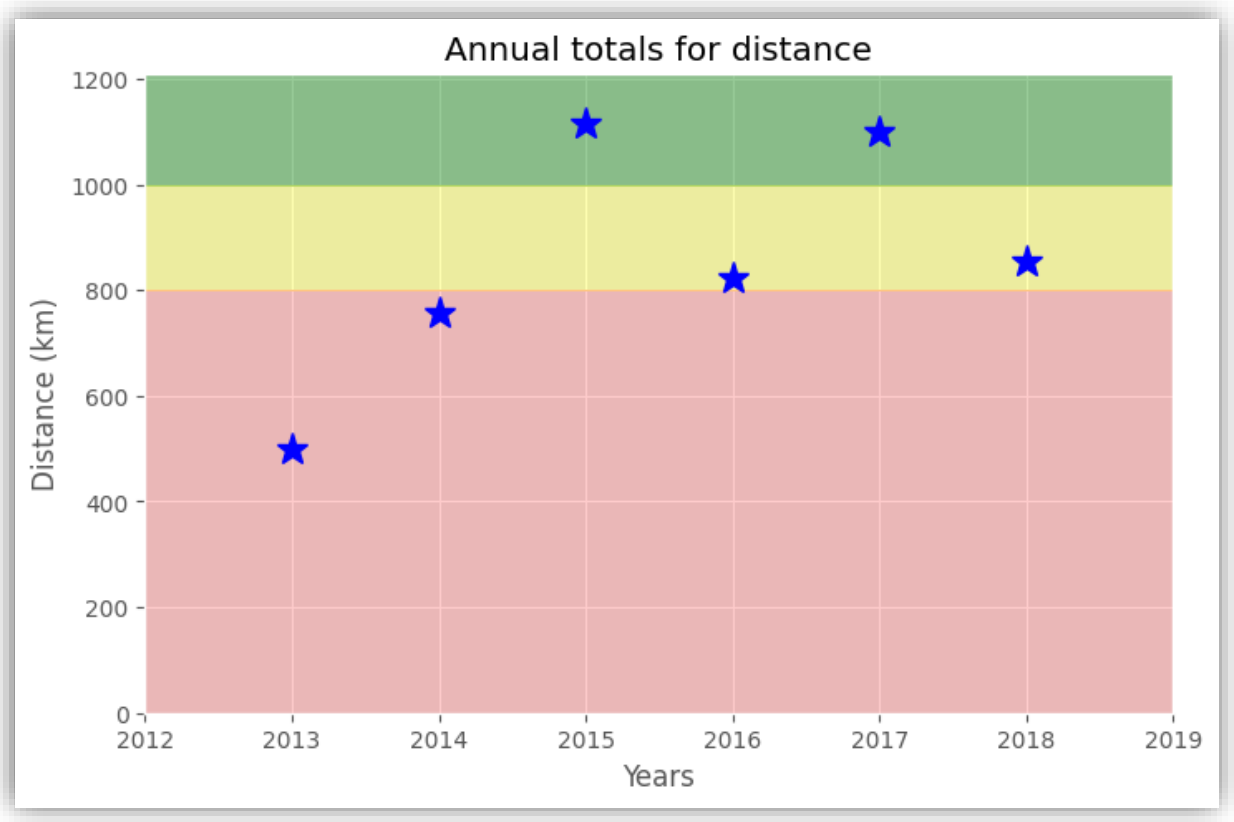
Detailed Analysis of Running Statistics

The analysis revealed key insights about running performance, such as average distance, speed, and heart rate over different periods.



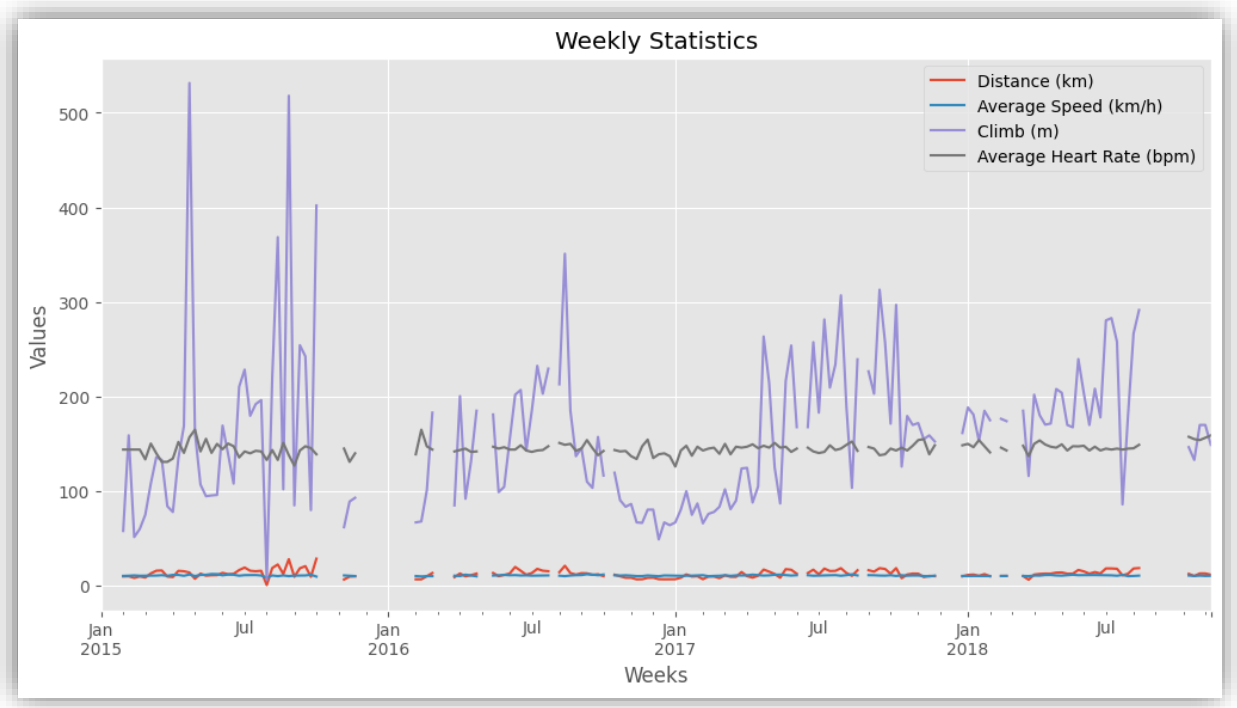
Comparison with Set Goals

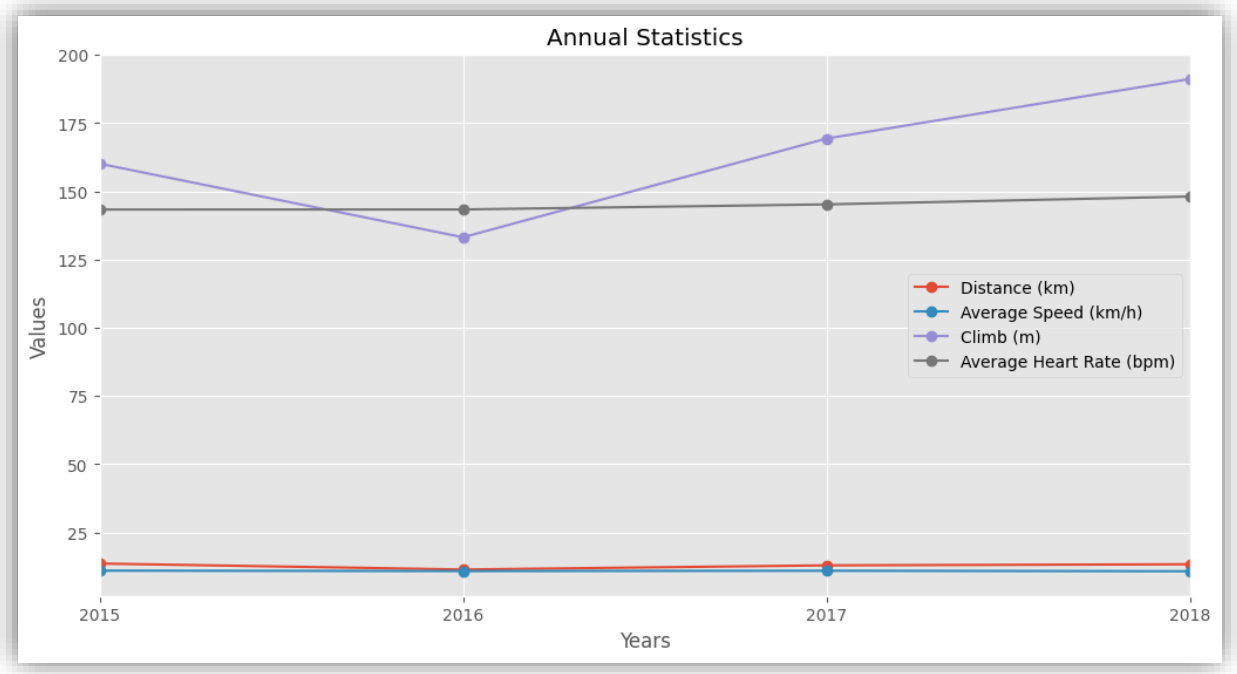
The annual totals for distance showed how the training aligned with the set goals, indicating areas of improvement or consistency.



Interpretation of Visualizations

The visualizations highlighted trends and patterns in the running data, providing a clear picture of performance over time.





Insights Gained from Data Analysis

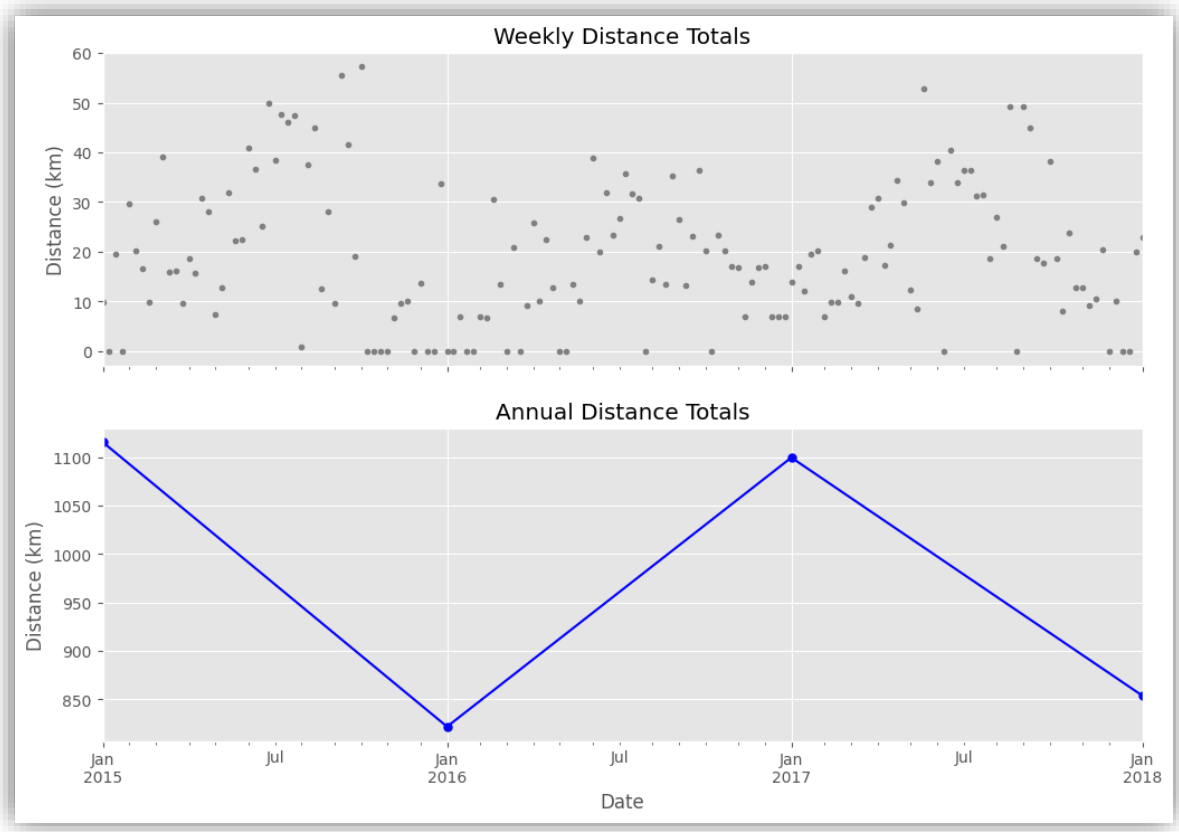
The analysis provided actionable insights, such as the effectiveness of training routines, areas needing improvement, and overall progress towards fitness goals.

7. Conclusion

Summary of Findings

The project successfully demonstrated how fitness tracker data can be analyzed to gain valuable insights into running performance and progress.

Weekly and Annual Distance Totals:




```
Weekly Distance Totals:
count    202.000000
mean      19.258713
std       14.280139
min        0.000000
25%        9.527500
50%       17.865000
75%       28.817500
max       57.260000
Name: Distance (km), dtype: float64
Annual Distance Totals:
count       4.000000
mean      972.565000
std       156.447657
min       821.640000
25%       845.685000
50%       976.595000
75%      1103.475000
max      1115.430000
Name: Distance (km), dtype: float64
```

Implications for Personal Training and Goal Setting

The findings can help runners set more realistic and achievable goals, tailor their training plans, and stay motivated.

Limitations of the Study

The analysis was limited to data from Runkeeper and may not generalize to other fitness tracking platforms. Additionally, some metrics had missing values that could affect the analysis.

Future Directions for Analysis or Improvements

Future analyses could incorporate data from multiple sources, explore more advanced machine learning techniques, and include more detailed goal-setting frameworks.

8. Recommendations

Recommendations Based on Findings

- Regularly review and update training goals based on data insights.
- Focus on areas needing improvement, such as increasing distance or speed gradually.
- Use visualizations to stay motivated and track progress over time.

Suggestions for Improving Fitness Tracking and Goal Achievement

- Ensure consistent data collection by regularly syncing fitness trackers.
- Set specific, measurable, and realistic goals.
- Use data-driven insights to adjust training plans and avoid overtraining.

9. Appendices

Raw Data Sample

A sample of the raw data is included to provide an overview of the dataset structure.

Additional Charts and Graphs

Additional visualizations and charts are provided to supplement the analysis.

Code Snippets

Key code snippets used in the analysis are included for reference.

10. References

- Pandas and Matplotlib official documentation
- Statsmodels official documentation

11. Fun Facts

- Forrest Gump would need approximately 61 pairs of running shoes to cover his total run distance of 19,024 miles.
- The average lifetime distance for a pair of running shoes is around 500 kilometers.