Final Report of Traineeship Program 2024

On

“Analysis of Fitness Data”

MEDTOUREASY

24th June 2024

ACKNOWLDEGMENTS

The traineeship 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 traineeship 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 traineeship 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

Acknowledgments

Abstract

Sr. No.	Topic
1	Introduction
	1.1 About the Company
	1.2 About the Project
	1.3 Objectives and Deliverables
2	Methodology
	2.1 Flow of the Project
	2.2 Language and Platform Used
3	Implementation
	3.1 Obtain and review raw data
	3.2 Data Preprocessing
	3.3 Dealing with missing values
	3.4 Plot running data
	3.5 Running statistics
	3.6 Visualization with averages
	3.7 Did I reach my goals?
	3.8 Am I progressing?
	3.9 Training intensity
	3.10 Detailed Summary Report
	3.11 Fun Facts
4	Sample Screenshots and Observations
5	Conclusion
6	Future Scope
7	References

ABSTRACT

With the explosion in fitness tracker popularity, runners all of 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, I’ve create import, clean, and analyze my data to answer the above questions.

1. Introduction

1.1 About the 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.

1.2 About the Project

In the realm of health and fitness, data is king. The project in question revolves around a sophisticated fitness data tracker, designed to capture a comprehensive array of metrics that are pivotal to an individual's fitness journey. As an analyst, the role transcends mere data collection; it involves delving into the nuances of each recorded session, extracting meaningful insights that answer critical questions for the user.

The tracker meticulously logs the duration, pace, and intensity of the user's runs, painting a vivid picture of their daily exertions. This data is then analyzed to determine if the user's training goals are being met, whether they are incremental daily goals or ambitious long-term objectives. Progress is tracked over time, with sophisticated algorithms detecting trends and patterns in the user's performance, providing a clear trajectory of their fitness journey.

Achievements are not just numbers; they are milestones that mark the user's commitment and hard work. The tracker highlights personal bests, celebrating victories both big and small, from the fastest mile run to the longest continuous workout. These achievements serve as both a record and a motivator, pushing the user to set new goals and strive for greater heights.

Comparison with others adds a social dimension to the fitness regimen. The tracker offers anonymized data aggregation, allowing users to see where they stand in relation to the broader fitness community. This feature fosters a sense of camaraderie and healthy competition, encouraging users to not only pursue their personal best but also to see how they stack up against their peers.

In essence, this project is not just about tracking fitness; it's about understanding it. It's about providing actionable insights that empower users to make informed decisions about their health and fitness routines. It's a tool that not only records data but also interprets it, transforming numbers into narratives that guide users towards their fitness aspirations. The ultimate goal is to create a seamless interface between human endeavor and digital analysis, where every step taken is a step understood, and every milestone reached is a story told.

Embarking on a project is always an exciting journey, and I have a comprehensive roadmap laid out with 11 tasks. Starting with the crucial step of obtaining and reviewing raw data, I’m setting a strong foundation for accuracy and insight. Data preprocessing then ensures that the information is primed for analysis, leading to the meticulous handling of missing values, which is often a pivotal step in maintaining the integrity of your dataset.
Plotting running data will bring to life the trends and patterns that are hidden within the numbers, while running statistics will quantify the story your data is trying to tell. Visualization with averages will further enhance the narrative, providing a clear and digestible means of communicating complex information. The introspective tasks of asking 'Did I reach my goals?' and 'Am I progressing?' are essential for self-evaluation and steering the project towards its desired outcomes.

Training intensity will reflect the vigor and dedication poured into the project, and the detailed summary report will encapsulate all the findings and learnings in a comprehensive manner. Lastly, the inclusion of fun facts will not only add a layer of engagement but also serve as a reminder that data, at its core, can be fascinating and enjoyable to explore. This structured approach to your project ensures that every angle is considered, every number is scrutinized, and every conclusion is well-founded. It's a testament to the meticulous planning and effort that is the hallmark of a successful project.

1.3 Objectives and Deliverables

The primary objective of this project is to harness the power of data to extract meaningful insights and track performance against established goals. To achieve this, we will embark on a systematic journey through various stages of data analysis. Initially, we will obtain and meticulously review raw data to ensure its quality and relevance. Following this, we will engage in data preprocessing to refine the dataset, enhancing its suitability for analysis.

Addressing missing values is crucial; we will implement robust strategies to handle any gaps in the data, ensuring the integrity of our dataset. Plotting running data will allow us to observe trends and patterns over time, providing a visual narrative of the data's story. Running statistics will be computed to grasp the underlying distributions and variances within the data.

Visualization with averages will then be employed to distill complex data into understandable and actionable information. We will constantly evaluate our progress by asking critical questions such as "Did I reach my goals?" and "Am I progressing?" These reflections are essential for maintaining the project's direction and focus.

Training intensity will be analyzed to gauge the effort and resources invested, ensuring alignment with the project's objectives. A detailed summary report will encapsulate all findings, presenting them in a comprehensive and accessible manner. Lastly, we will sprinkle our report with fun facts to engage and enlighten stakeholders, adding an element of enjoyment to the data-driven journey.

Through these tasks, our objectives are clear: to transform raw data into a wellspring of insights, to track our trajectory towards our goals, and to foster an environment where data informs decisions and strategies. This structured approach ensures that every step taken is a stride towards data-driven excellence.

2. METHODOLOGY

2.1 Flow of the Project

The methodology and flow of a project are crucial for its success, providing a structured approach to achieving the project's objectives. Here's a comprehensive methodology for your project, incorporating the tasks you've outlined:

1. **Obtain and Review Raw Data**: Begin by gathering all necessary data, ensuring it's relevant and comprehensive. This stage involves meticulous scrutiny of the data sources for accuracy and reliability.

2. **Data Preprocessing**: Cleanse the data to improve its quality. This includes formatting, correcting errors, and standardizing to ensure consistency across the dataset.

3. **Dealing with Missing Values**: Identify any gaps in the data and address them appropriately, either by imputation or by analyzing the impact of their absence.

4. **Plot Running Data**: Visualize the data trends over time with running plots. This will help in understanding the dynamics and patterns within the data.

5. **Running Statistics**: Perform statistical analysis to summarize the data's main characteristics, often with the intent of making further analysis simpler or clearer.

6. **Visualization with Averages**: Create visual representations that incorporate average values to identify norms and deviations within the dataset.

7. **Evaluation of Goals**: Assess whether the initial project goals have been met. This involves a comparison of the project outcomes against the set objectives.

8. **Progress Assessment**: Regularly evaluate the project's progression to ensure it's on track. This includes reviewing milestones and deliverables against the project timeline.

9. **Training Intensity**: If the project involves a training component, analyze the intensity and effectiveness of the training modules.

10. **Detailed Summary Report**: Compile a comprehensive report detailing the findings, methodologies, and outcomes of the project.

11. **Fun Facts**: Incorporate interesting and engaging facts related to the project's subject matter to add an element of enjoyment and enhance the readability of the report.

This flow ensures a thorough and systematic approach to the project, allowing for clear tracking of progress and outcomes. It's designed to be adaptable, so you can tailor it to the specific needs and nuances of your project. Remember, the key to a successful project is not just in following the methodology, but also in being flexible and responsive to the data and findings as they evolve. Good luck with your project!

2.2 Language and Platform Used

2.2.1 Language: Python

Python is a powerhouse in the realm of data analysis, offering an intuitive syntax coupled with a rich ecosystem of libraries like pandas for data manipulation, NumPy for numerical computing, and Matplotlib for visualization. These tools streamline the data analysis workflow, from cleaning and transforming data to performing complex statistical analyses. Python's versatility and ease of use make it an ideal choice for data analysts who seek to uncover insights and drive decision-making through data-driven evidence.

2.2.2 Google Colab

Google Colab has revolutionized the way data analysis projects are approached, offering a cloud-based platform that is both powerful and accessible. With its seamless integration with Google Drive and other Google services, Colab provides an environment where data scientists can share, collaborate, and execute Python code in real-time. Its compatibility with popular libraries like Pandas, NumPy, and Matplotlib makes it an ideal choice for exploratory data analysis, allowing for quick insights into data sets and fostering a more efficient workflow. The ability to access high-performance computing resources such as GPUs and TPUs for complex computations is another standout feature, making it a go-to tool for data analysts worldwide. Whether you're a seasoned professional or just starting out, Google Colab's user-friendly interface and robust capabilities can significantly enhance the productivity and scope of your data analysis projects.

2.2.3 Package: Matplotlib

Matplotlib is a widely-used Python library for data visualization, which is particularly useful in data analysis projects. It provides an extensive range of plotting options, from histograms to scatter plots, and is designed to work well with NumPy, a library for numerical computations in Python. Matplotlib's versatility allows for the creation of complex plots with relatively simple code, making it an invaluable tool for exploratory data analysis and the communication of results through visualizations.

2.2.4 Python Library: Panda

The Python library Pandas is an essential tool for data analysis, providing robust data structures like Series and DataFrame. These structures are designed for efficient data manipulation and analysis, enabling tasks such as data cleaning, transformation, and aggregation. Pandas supports various data formats, including CSV and Excel, and integrates well with other libraries for scientific computing. Its functionality for handling missing data, merging datasets, and time series analysis makes it a go-to library for data scientists and analysts. For more detailed information, the official pandas documentation provides a comprehensive overview.

2.2.5 Statsmodels.api` Module

The `statsmodels.api` module in Python is a comprehensive library for estimating and interpreting various statistical models. It is particularly useful in the field of econometrics and data analysis, providing tools for linear and non-linear regression, time-series analysis, and hypothesis testing. The library supports specifying models using R-style formulas and integrates seamlessly with pandas DataFrames, making it a powerful tool for statistical computation within a Python environment. With an extensive list of result statistics available for each estimator, `statsmodels.api` facilitates detailed statistical data exploration and the validation of models against established statistical packages, ensuring accuracy and reliability in data analysis projects.

3. IMPLEMENTATION

The project's structure is meticulously organized into 11 distinct tasks, each with a clear explanation and execution plan. This systematic approach ensures that every phase of the project is addressed with attention to detail, facilitating a smooth workflow and clear communication among team members. By breaking down the project into manageable tasks, I had focused on delivering quality results in each segment, ultimately leading to the successful completion of the project as a whole.

3.1 Obtain and review raw data

To manage your training activities data with Python, you'll start by importing the pandas library as 'pd'. Then, you'll load your dataset from a CSV file into a DataFrame named 'df_activities', making sure to parse the dates and set the 'Date' column as the index. Once loaded, you can display three random entries from this DataFrame to get a glimpse of your data. Finally, you'll use the 'info()' method to print out a concise summary of your DataFrame, providing insights into the types and counts of data you're working with.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=75ikip6JPKd4&line=2&uniqifier=1

3.2 Data preprocessing

To process your data, you'll start by removing any columns that aren't needed from the dataframe named 'df_activities'. This is done by applying the 'drop()' function and specifying the 'cols_to_drop' list in the 'columns' parameter. Next, you'll determine how many instances of each activity type there are by using the 'value_counts()' method on the 'Type' column. Then, you'll update the dataframe by changing any 'Other' values in the 'Type' column to 'Unicycling' with the help of the 'str.replace()' method. Lastly, you'll identify any missing data by counting the number of null values in each column using 'isnull().sum()'. These steps will ensure your dataset is clean and ready for analysis.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=W-vfyUibPKd5&line=11&uniqifier=1

3.3 Dealing with missing values

To handle missing data in a dataset, one can use mean imputation, which involves replacing missing values with the mean value of the entire feature column. For example, to address missing 'Average Heart Rate (bpm)' values for cycling activities, one would first calculate the mean heart rate for all cycling activities and assign this value to a variable named 'avg_hr_cycle'. Next, one would filter the main dataframe to include only cycling activities, create a copy of this filtered data, and assign it to 'df_cycle'. Then, the missing heart rate values in 'df_cycle' can be filled with the integer value of 'avg_hr_cycle' using the 'fillna()' method. Finally, to understand the extent of missing data, one would count the missing values across all columns in the 'df_run' dataframe.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=-YBGDknLPKd5&line=7&uniqifier=1

3.4 Plot running data

To visualize running data spanning from 2013 to 2018, one would begin by segmenting the dataset 'df_run' to include only the entries within this timeframe. Given that the dataset is organized with the most recent entries at the top, the process would involve filtering the data accordingly. The filtered subset would then be stored in a new variable named 'runs_subset_2013_2018'. When it comes to plotting this data, it's essential to enable the subplot feature by setting the 'subplots' argument to 'True', adhering to the PEP 8 style guidelines which suggest avoiding spaces around the '=' in keyword arguments. Finally, to display the plot, one would use the command 'plt.show()' to render the visualization on the screen. This sequence of steps will result in a clear graphical representation of the running data over the specified years.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=eSzSUBzFPKd5&line=8&uniqifier=1

3.5 Running statistics

To understand my running performance, I'm looking to calculate the yearly and weekly averages for the distance I've run in kilometers, my average speed in kilometers per hour, the elevation I've climbed in meters, and my average heart rate in beats per minute. I'll focus on the data from 2015 to 2018, which I'll refer to as 'runs_subset_2015_2018'. For the annual averages, I'll use the 'resample' function with an 'A' alias, and apply the 'mean' method to this subset. Similarly, I'll calculate the average weekly statistics by resampling with a 'W' alias and using the 'mean' method twice. Lastly, I'll determine the average number of weekly trainings by filtering the 'Distance (km)' column and applying both 'count' and 'mean' methods, naming the outcome 'weekly_counts_average'.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=se8D1uvoPKd6&line=3&uniqifier=1

3.6 Visualization with averages

I need to organize some data and create a visual representation. First, I'll take the distance and heart rate data from the 'runs_subset_2015_2018' dataset and store them in 'runs_distance' and 'runs_hr' variables, respectively. Then, I'll generate two subplots that share the same x-axis. I'll do this by calling the 'plt.subplots()' function, specifying that I want 2 rows of plots, with a shared x-axis, and a figure size of 12 by 8 inches. The outputs will be stored in 'fig', 'ax1', and 'ax2'. Next, I'll plot the distance data on the first subplot, using 'ax1'. Finally, on the second subplot, 'ax2', I'll draw a horizontal line representing the average heart rate using the 'axhline()' function, setting the color to blue, the line width to 1, and the line style to a dash-dot pattern.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=SAQ4hLxvPKd6&line=8&uniqifier=1

3.7 Did I reach my goals?

First, I'll filter the 'df_run' dataset to include only the data from 2013 to 2018, focusing specifically on the 'Distance (km)' column. Next, I'll calculate the annual totals by resampling and summing the data, which I'll then assign to 'df_run_dist_annual'. After that, I'll create a visual representation by setting up a plot with dimensions of 8.0 by 5.0 inches. To highlight a specific range, I'll add a horizontal span from 0 to 800 km with a red color and a transparency set at 0.2. Finally, I'll display the plot to review the annual distance trends.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=1nj7fZ70PKd6&line=1&uniqifier=1

3.8 Am I progressing?

To analyze my running progress, I'll start by importing the 'statsmodels.api' as 'sm' to use its functionalities. Then, I'll focus on the data from 2013 to 2018, specifically looking at the 'Distance (km)' column. I'll organize this data on a weekly basis and use the backfill method to handle any missing values, ensuring a continuous dataset. With this prepared data, I'm going to create a visual representation—a plot that clearly shows the distance I've covered each week and the overall trend, setting the size of the plot to be 12 by 5 inches for a clear view.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=0YWa7NcqPKd6&line=3&uniqifier=1

3.9 Training intensity

I'm going to create a histogram that shows the distribution of heart rates. First, I'll take a subset of the 'df_run' data from March 2015 to 2018 and select the 'Average Heart Rate(bpm)' column, naming this new subset 'df_run_hr_all'. Then, I'll generate a plot using 'plt.subplots()', setting the size of the figure to 8 by 5 inches. I'll assign the output to 'fig' and 'ax'. Next, I'll customize the x-axis tick labels using 'ax.set_xticklabels()', with the labels set to 'zone_names', rotated at a -30-degree angle for better readability, and aligned to the left. Finally, I'll display the plot with 'plt.show()'.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=GP0Rdy9zPKd7&line=10&uniqifier=1

3.10 Detailed summary report

I'm creating a summary report by first concatenating the 'df_run' DataFrame with 'df_walk' and 'df_cycle' using the append() function. After that, I'll sort the combined DataFrame based on the index in descending order and assign it to 'df_run_walk_cycle'. Next, I'll group 'df_run_walk_cycle' by activity type and select the columns in 'dist_climb_cols', then sum the results and assign them to 'df_totals'. Finally, I'll use the stack() method on 'df_summary' to display a compact, reshaped form of the full summary report.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=rlGNYefHPKd7&line=2&uniqifier=1

3.11 Fun facts

I'm calculating the average number of shoes an instructor goes through in a lifetime by dividing the total kilometers they've run by the number of pairs of shoes they've worn out. Then, I'll estimate how many shoes Forrest

Gump might have used on his epic run by taking the total kilometers he ran and dividing it by the average number of kilometers I get from one pair of shoes, based on the instructor's experience. It's a fun way to understand the durability of running shoes and the incredible distances covered by runners.

To view the completed task as described above, please refer to the following cell in the Google Colab platform. This reference link will direct you to the specific section where the task has been executed, allowing for a comprehensive review of the methodologies and results obtained. Ensure that you have the necessary access permissions to view the content within the Google Colab environment.

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc#scrollTo=hL80d3ZWEE4k&line=9&uniqifier=1

4. SAMPLE SCREENSHOTS AND OBSERVATIONS

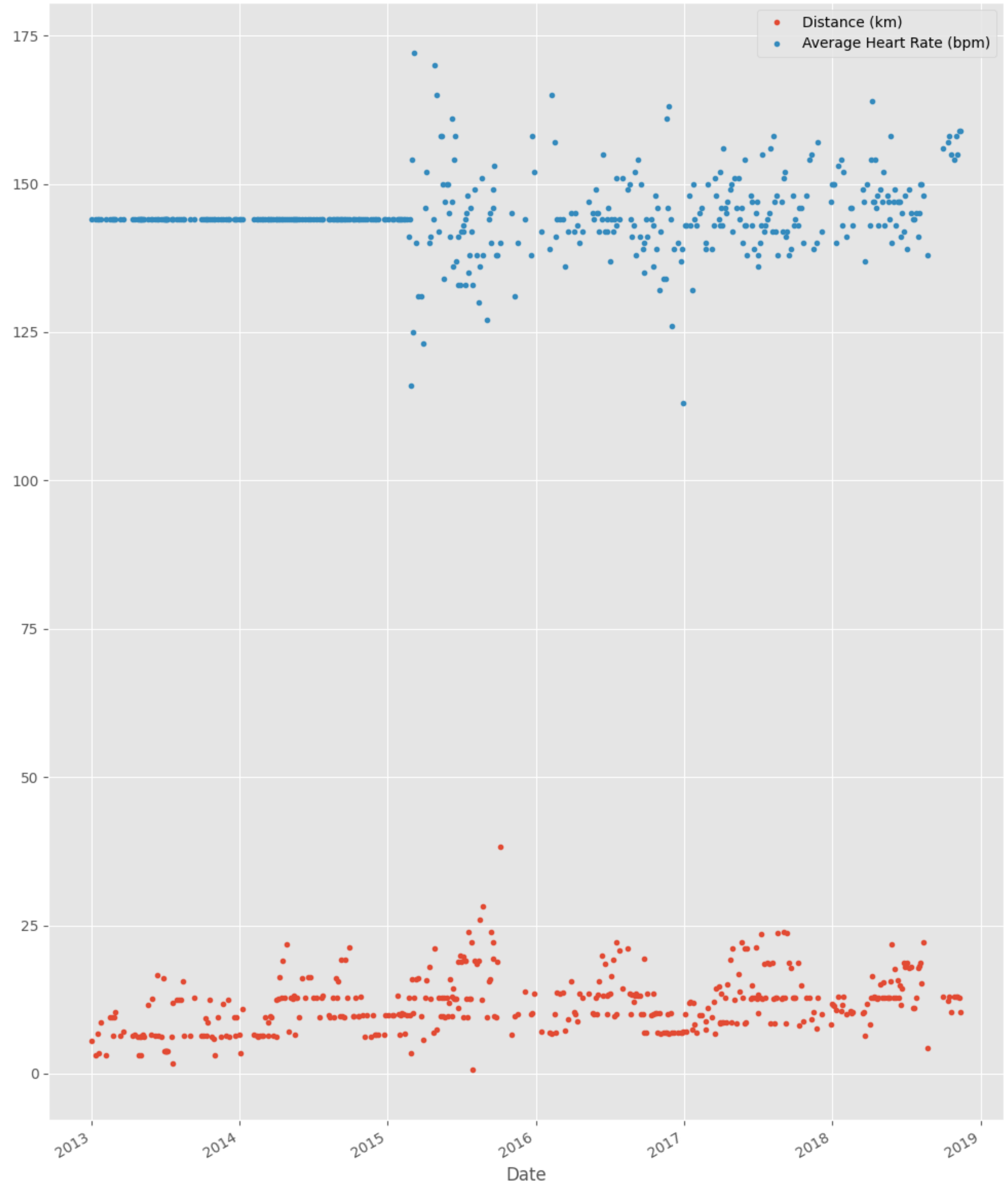

Figure 1“Analyzing the Relationship: Distance Run vs. Average Heart Rate (2013-2019)”

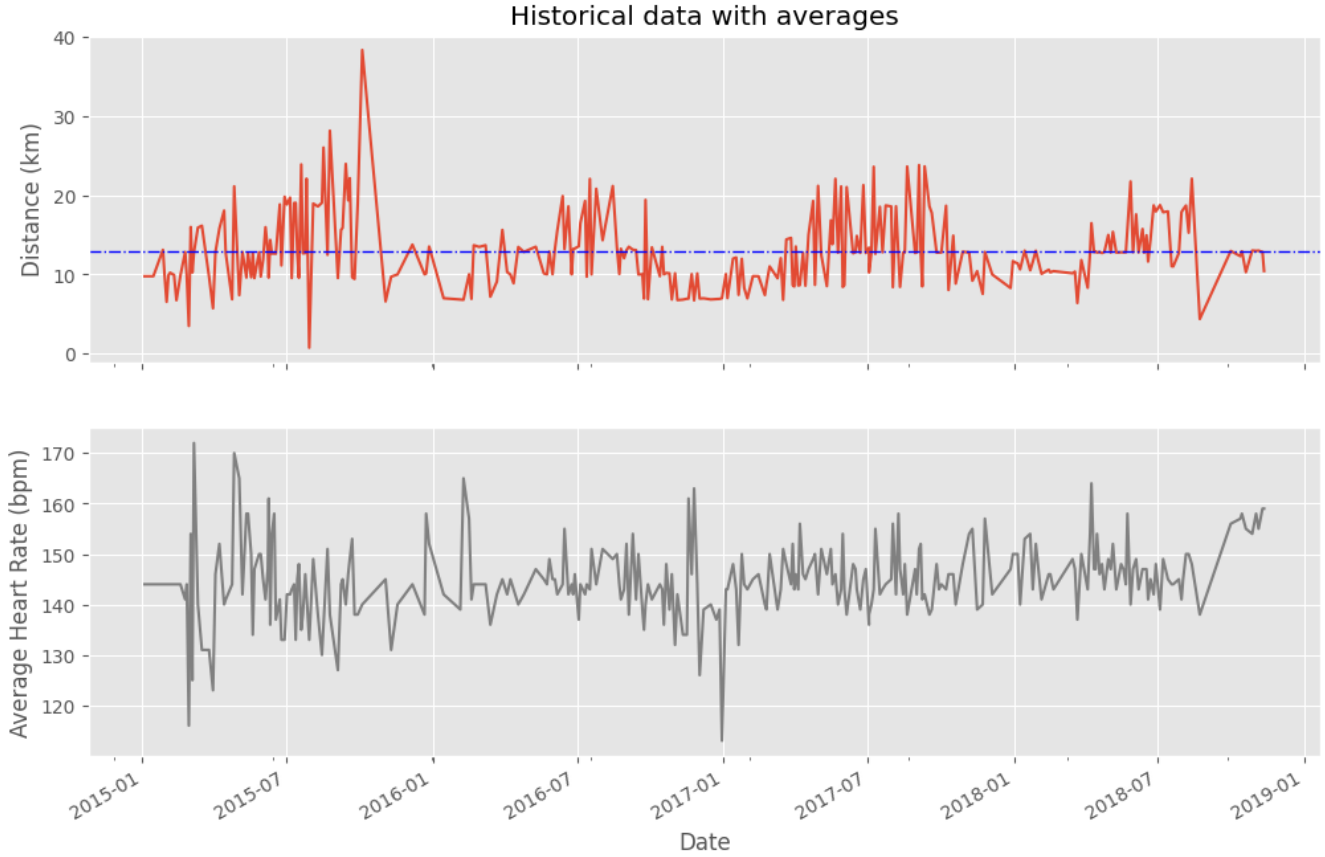

Figure 2"Analyzing the Relationship: Distance Run vs. Average Heart Rate (2013-2019)"

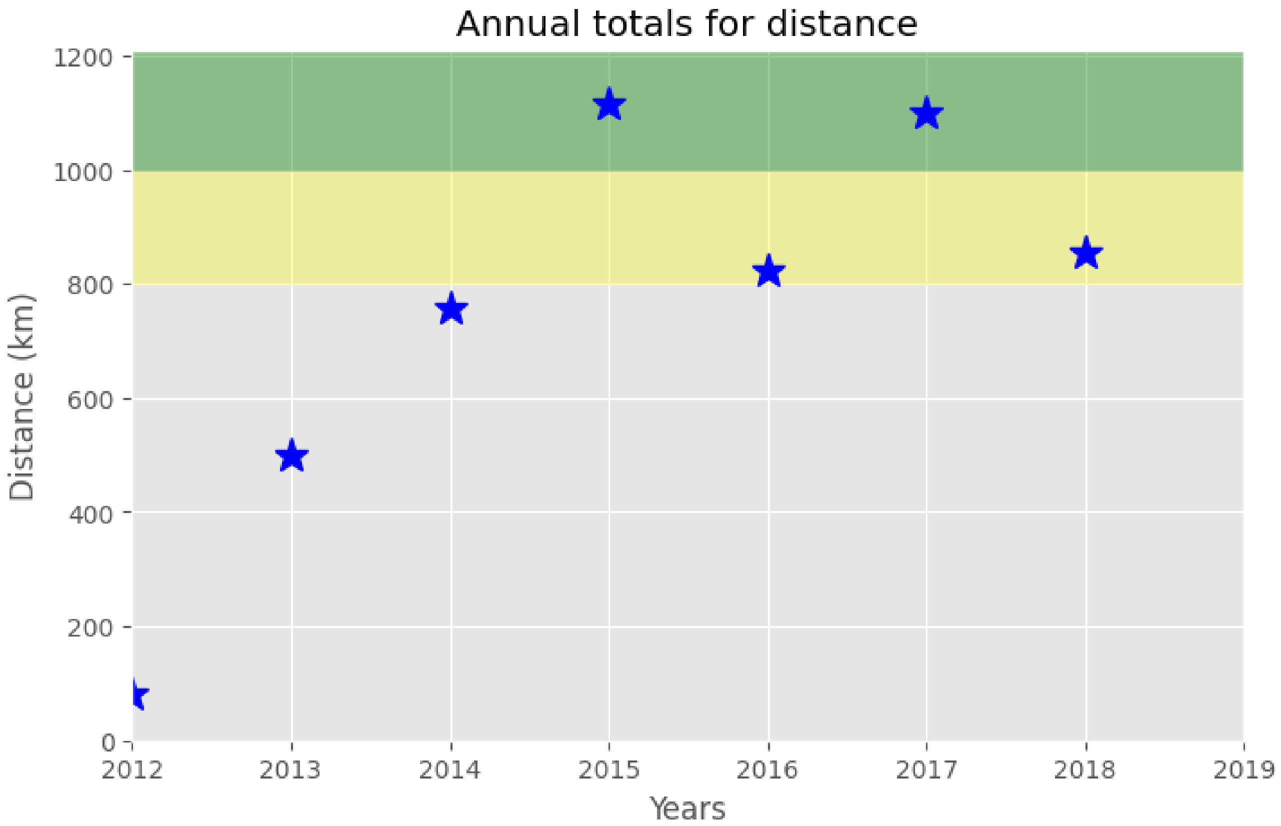

Figure 3“Annual Running Distance Totals: Tracking Progress from 2012 to 2018”

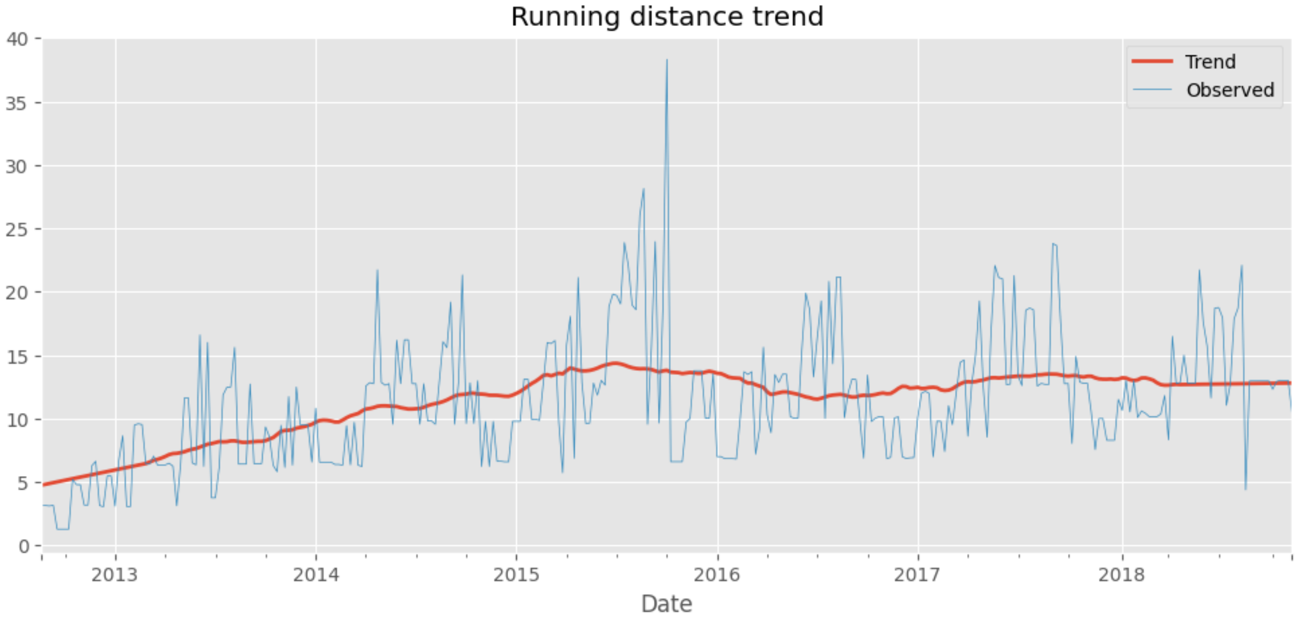

Figure 4“Trend Analysis of Running Distances and Heart Rates from 2013 to 2018”

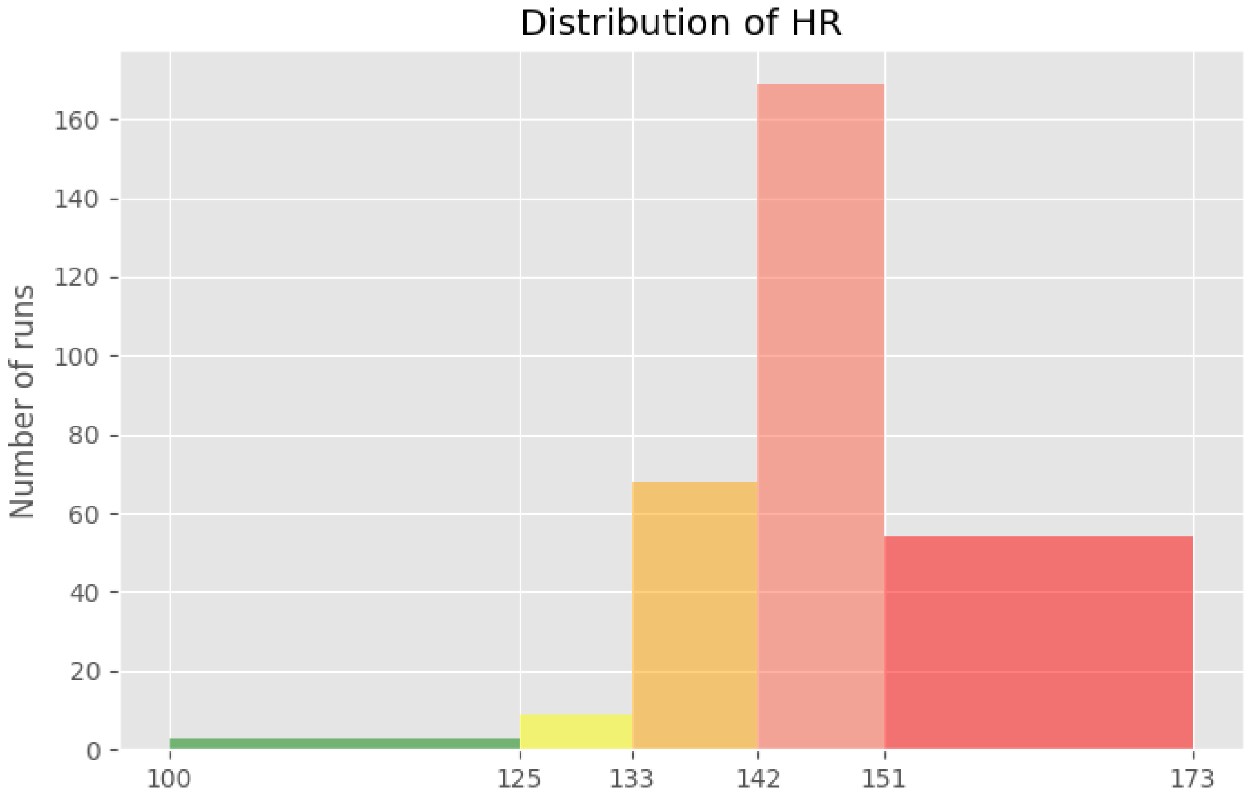

Figure 5“Training Intensity Zones: Heart Rate Distribution Analysis”

5. CONCLUSION

In conclusion, the data analysis project on the fitness tracking device has provided valuable insights into the patterns and trends of physical activity among users. The findings suggest that such devices can significantly contribute to the promotion of a healthier lifestyle by providing users with real-time feedback and personalized data. Looking ahead, the future scope of this project could include the integration of more advanced analytical tools to predict future trends in fitness behaviors. Additionally, exploring the correlation between fitness tracking data and various health outcomes could provide deeper understandings of the impact of physical activity on overall well-being. The potential for these devices to be used in preventive healthcare is immense, and further research could lead to more tailored and effective health recommendations for individuals. Ultimately, the continued evolution of wearable technology and data analytics will likely play a pivotal role in shaping the future of personal health and fitness.

6. FUTURE SCOPE

Looking ahead, the future scope of fitness trackers is poised for even greater advancements. With the integration of AI and machine learning, these devices will not only track but also predict health trends, offering personalized recommendations for health optimization. The incorporation of more advanced biometric sensors will allow for the monitoring of a wider range of physiological markers, such as blood glucose levels and oxygen saturation, making them indispensable tools for preventive healthcare. Furthermore, the expansion of their capabilities to include mental health tracking, through stress and mood indicators, will provide a more holistic view of an individual's well-being. As technology progresses, we can anticipate a new era of fitness trackers that are more intuitive, comprehensive, and seamlessly integrated into our daily lives, revolutionizing the way we approach health and fitness.

7. REFERENCES

Certainly! You can find the reference link to my project below:

https://colab.research.google.com/drive/1WHEs_hOWGv_8NSWWku6r-JbgjwM7fysc?usp=sharing