Final Report of Traineeship Program 2024

On

“Analyse Fitness Data”

MEDTOUREASY

27th 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.

ABSTRACT

The "Analyzing Fitness Data" project focuses on examining fitness activities such as running, cycling, and walking through data science methods. It involves collecting data from fitness trackers, followed by extensive data cleaning and filtering to ensure accuracy. The project uses exploratory data analysis (EDA) to visualize metrics like distance, speed, duration, and heart rate, applying statistical and machine learning techniques to identify patterns and make predictions. Key findings are visualized using Python libraries, providing clear insights into fitness trends. The project concludes with significant findings and discusses future enhancements, including real-time data integration and advanced machine learning models for personalized fitness recommendations. This work demonstrates the effective use of data science in deriving actionable insights from fitness data, aiding in better health management.

TABLE OF CONTENTS

INTRODUCTION

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.

About the Project

The fitness industry has witnessed significant growth in recent years, with more individuals prioritizing their health and wellness. The need for data-driven insights in the fitness sector has become paramount to understand user behavior, preferences, and trends. Analyzing fitness data allows companies to tailor their services, improve user experience, and make informed business decisions.

This project aims to collect and analyze large datasets related to fitness activities to create intuitive and interactive dashboards for gaining meaningful insights. The project is structured into three main subsections:

Analysis of the Problem: This section focuses on assessing the trends and patterns in fitness data. It includes statistics and data representing various aspects such as workout frequency, duration, intensity, calories burned, and user demographics. Comparative analysis based on parameters like age, gender, activity type, and location provides a comprehensive view of fitness trends.

Objectives

This project aims to create intuitive, interactive, and dynamic dashboards by gathering fitness data from various sources, such as fitness tracking apps, wearable devices, and user surveys. The coding language Python along with libraries like Pandas, Matplotlib, and Plotly will be used for data manipulation, analysis, and visualization to provide valuable insights into fitness trends and user behavior.

The objectives of this project include:

1. **Data Collection and Integration:** Gather fitness data from multiple sources and integrate it into a structured format for analysis.
2. **Data Cleaning and Preprocessing:** Perform data cleaning to handle missing values, outliers, and inconsistencies in the dataset.
3. **Exploratory Data Analysis (EDA):** Conduct EDA to uncover patterns, trends, and correlations within the fitness data.
4. **Dashboard Creation:** Develop interactive dashboards using tools like Plotly Dash or Streamlit to visualize key fitness metrics and trends.
5. **User Segmentation:** Segment users based on demographics, activity types, workout preferences, and performance metrics.
6. **Performance Analysis:** Analyze workout performance metrics such as distance covered, duration, calories burned, heart rate, and workout intensity.
7. **Trend Identification:** Identify emerging fitness trends, popular workout routines, peak activity times, and user engagement patterns.
8. **Goal Tracking:** Track user fitness goals, achievements, progress over time, and adherence to workout plans.

9. **`Predictive Modeling (Optional):** Explore the possibility of implementing predictive models to forecast user behavior, fitness goals, and performance outcomes.

METHODOLOGY

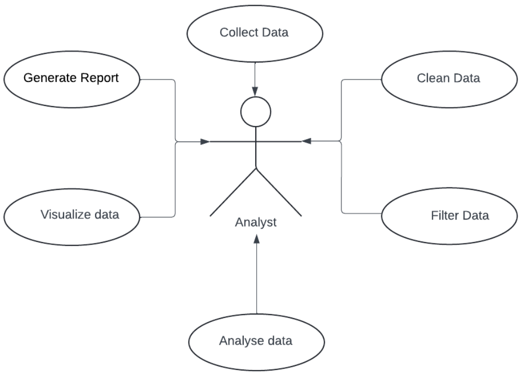
Flow of the Project

The project follows a structured approach to analyze fitness data, involving several key steps:

1. **Requirement Gathering and Problem Definition:** Understand the objectives, requirements, and scope of the project. Define the problem statement and the goals to be achieved.
2. **Data Collection:** Gather fitness data from various sources, such as fitness tracking apps, wearable devices, and user surveys. Ensure the data is comprehensive and relevant for analysis.
3. **Data Cleaning:** Clean the collected data by removing duplicates, handling missing values, and correcting inconsistencies to ensure data quality.
4. **Data Filtering:** Filter the data to focus on relevant activities (e.g., running, cycling, walking) and metrics (e.g., distance, duration, average speed, climb).
5. **Data Analysis:** Perform statistical analysis to derive meaningful insights from the data, including calculating summary statistics, identifying trends, and comparing different types of activities.
6. **Data Visualization:** Create visualizations to represent the data and analysis results, making it easier to understand and interpret the findings.
7. **Report Generation:** Compile the analysis and visualizations into a comprehensive report, summarizing the key insights and conclusions drawn from the study.

Use Case Diagram

The use case diagram for this project illustrates the interactions between the user (analyst) and the various steps involved in the project. The key use cases include data collection, data cleaning, data filtering, data analysis, data visualization, and report generation.



Language and Platform Used

The project utilizes the following tools and technologies:

- **Programming Language: Python -** Python is used for data manipulation, analysis, and visualization due to its extensive libraries and ease of use.
- **Libraries:**

    - **Pandas:** For data manipulation and cleaning.
    - **NumPy:** For numerical computations.
    - **Matplotlib and Seaborn:** For data visualization.
    - **SciPy:** For statistical analysis.
    - **Scikit-learn:** For advanced data analysis and machine learning (if applicable).
- **Platform: Jupyter Notebook -** Jupyter Notebook provides an interactive environment for writing and running Python code, making it suitable for data analysis and visualization tasks.

By following this methodology, the project aims to thoroughly analyze fitness data and derive valuable insights that can be used to enhance training and performance.

IMPLEMENTATION

Gathering Requirements and Defining Problem Statement

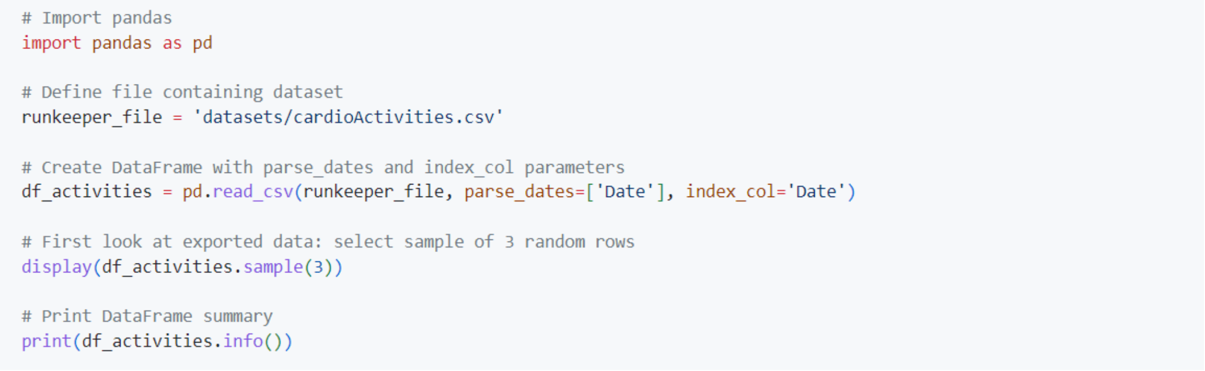
The first step in the project was to gather the requirements and clearly define the problem statement. The main goal was to analyze fitness data to gain insights into training activities, performance metrics, and trends over time. The key requirements included:

- Collecting comprehensive fitness data from various activities like running, cycling, walking, and other types.
- Cleaning and preprocessing the data to ensure accuracy and consistency.
- Filtering the data to focus on specific time periods and types of activities.
- Analyzing the data to extract meaningful patterns and trends.
- Visualizing the data to create informative and interactive plots.

Data Collection and Importing

The data was collected from fitness tracking applications and devices. The dataset contained various metrics related to different activities, such as distance, duration, average speed, climb, and average heart rate. The data was stored in a CSV file named cardioActivities.csv.

The first step in importing the data was to read the CSV file into a Pandas DataFrame using the following code:



```
# Import pandas
import pandas as pd

# Define file containing dataset
runkeeper_file = 'datasets/cardioActivities.csv'

# Create DataFrame with parse_dates and index_col parameters
df_activities = pd.read_csv(runkeeper_file, parse_dates=['Date'], index_col='Date')

# First look at exported data: select sample of 3 random rows
display(df_activities.sample(3))

# Print DataFrame summary
print(df_activities.info())
```

Designing Databases

In this project, we used Pandas DataFrames to store and manipulate the data instead of a traditional relational database. The DataFrame provided a flexible and efficient way to handle the dataset, which included various types of fitness activities and their corresponding metrics.

The dataset was structured as follows:

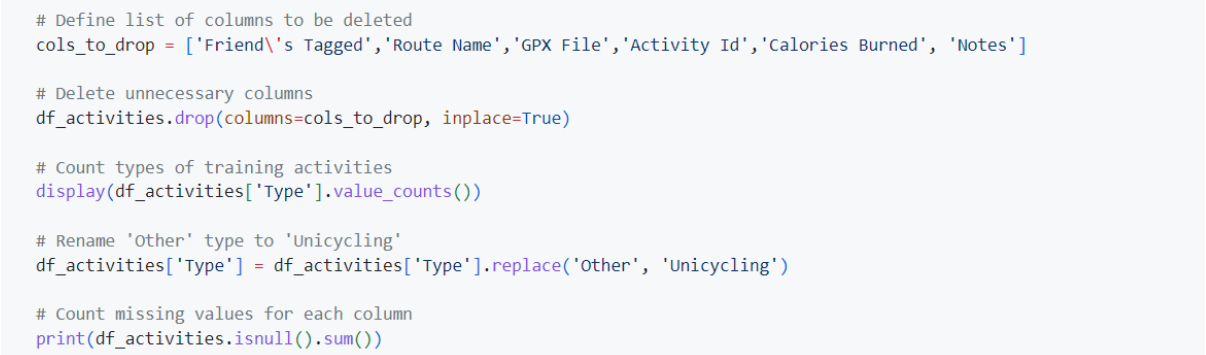
- **Columns:** The DataFrame columns included Type, Distance (km), Duration, Average Pace, Average Speed (km/h), Climb (m), and Average Heart Rate (bpm).
- **Rows:** Each row represented a single fitness activity with all the relevant metrics.

Data Cleaning

Data cleaning involved several steps to ensure the dataset was accurate and consistent:

1. **Removing Unnecessary Columns:** Certain columns that were not required for analysis were dropped.
2. **Handling Missing Values:** Missing values in the Average Heart Rate (bpm) column were filled with the average heart rate calculated for each activity type.
3. **Renaming Activity Types:** The 'Other' activity type was renamed to 'Unicycling' for clarity.

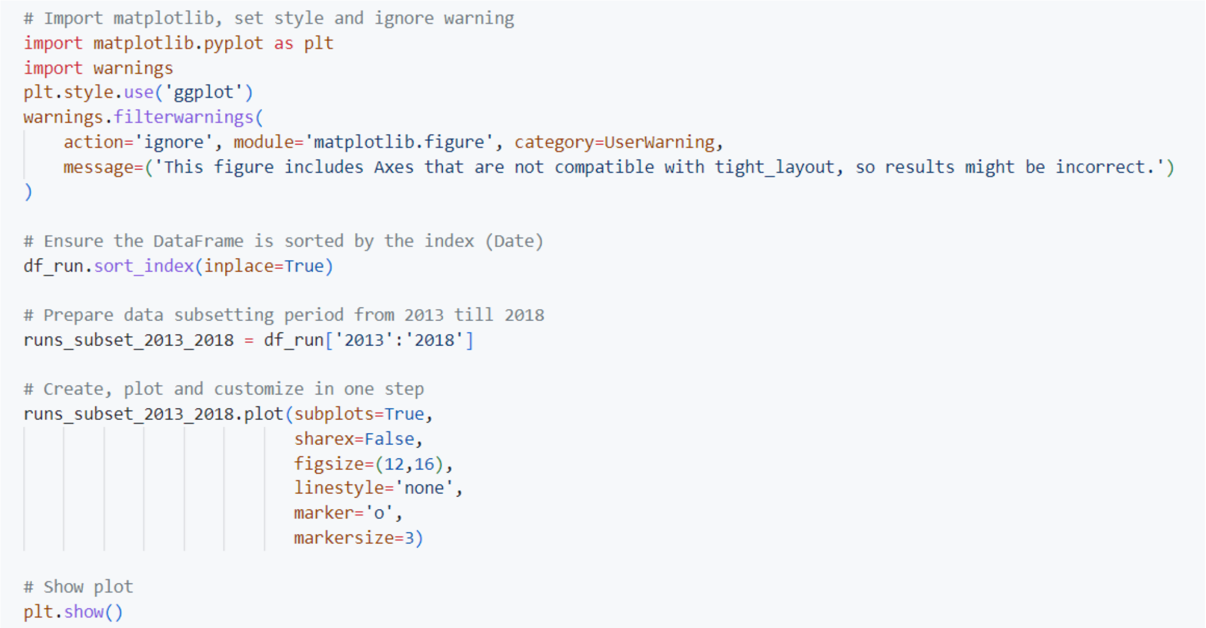
The following code snippets show how these tasks were performed:



```
# Define list of columns to be deleted
cols_to_drop = ['Friend\'s Tagged','Route Name','GPX File','Activity Id','Calories Burned', 'Notes']

# Delete unnecessary columns
df_activities.drop(columns=cols_to_drop, inplace=True)

# Count types of training activities
display(df_activities['Type'].value_counts())

# Rename 'Other' type to 'Unicycling'
df_activities['Type'] = df_activities['Type'].replace('Other', 'Unicycling')

# Count missing values for each column
print(df_activities.isnull().sum())
```

Data Filtering

Data filtering involved selecting subsets of the data based on specific criteria, such as time period or activity type. For instance, we filtered the running data for analysis over the period from 2013 to 2018:

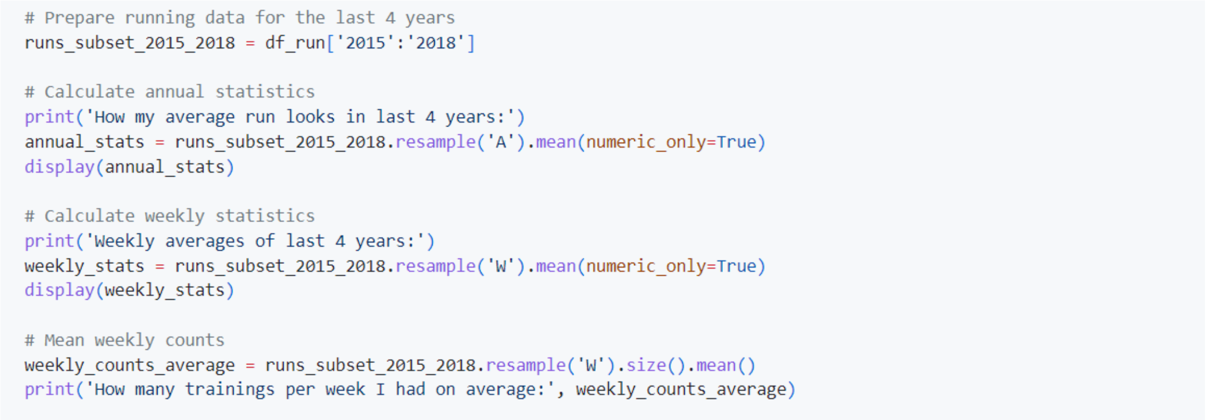


```
# Import matplotlib, set style and ignore warning
import matplotlib.pyplot as plt
import warnings
plt.style.use('ggplot')
warnings.filterwarnings(
    action='ignore', module='matplotlib.figure', category=UserWarning,
    message=('This figure includes Axes that are not compatible with tight_layout, so results might be incorrect.')
)

# Ensure the DataFrame is sorted by the index (Date)
df_run.sort_index(inplace=True)

# Prepare data subsetting period from 2013 till 2018
runs_subset_2013_2018 = df_run['2013':'2018']

# Create, plot and customize in one step
runs_subset_2013_2018.plot(subplots=True,
                           sharex=False,
                           figsize=(12,16),
                           linestyle='none',
                           marker='o',
                           markersize=3)

# Show plot
plt.show()
```

This allowed us to focus our analysis on relevant portions of the data, making the insights more targeted and meaningful.

Data Analysis

Data analysis was performed to extract meaningful insights and patterns from the fitness data. This included calculating average heart rates, distances, and other metrics for different activities, as well as identifying trends over time. Some key analysis steps included:



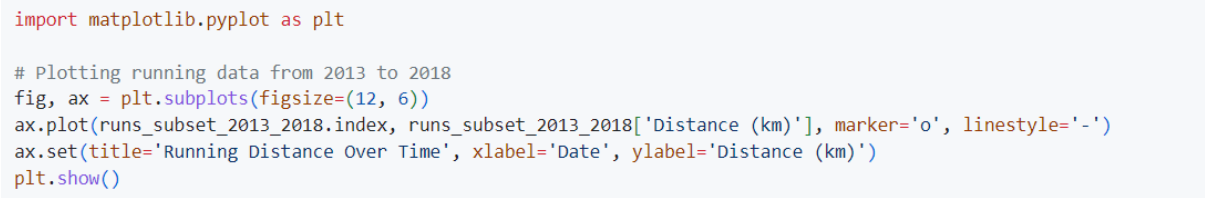
```
# Prepare running data for the last 4 years
runs_subset_2015_2018 = df_run['2015':'2018']

# Calculate annual statistics
print('How my average run looks in last 4 years:')
annual_stats = runs_subset_2015_2018.resample('A').mean(numeric_only=True)
display(annual_stats)

# Calculate weekly statistics
print('Weekly averages of last 4 years:')
weekly_stats = runs_subset_2015_2018.resample('W').mean(numeric_only=True)
display(weekly_stats)

# Mean weekly counts
weekly_counts_average = runs_subset_2015_2018.resample('W').size().mean()
print('How many trainings per week I had on average:', weekly_counts_average)
```

Data Visualization

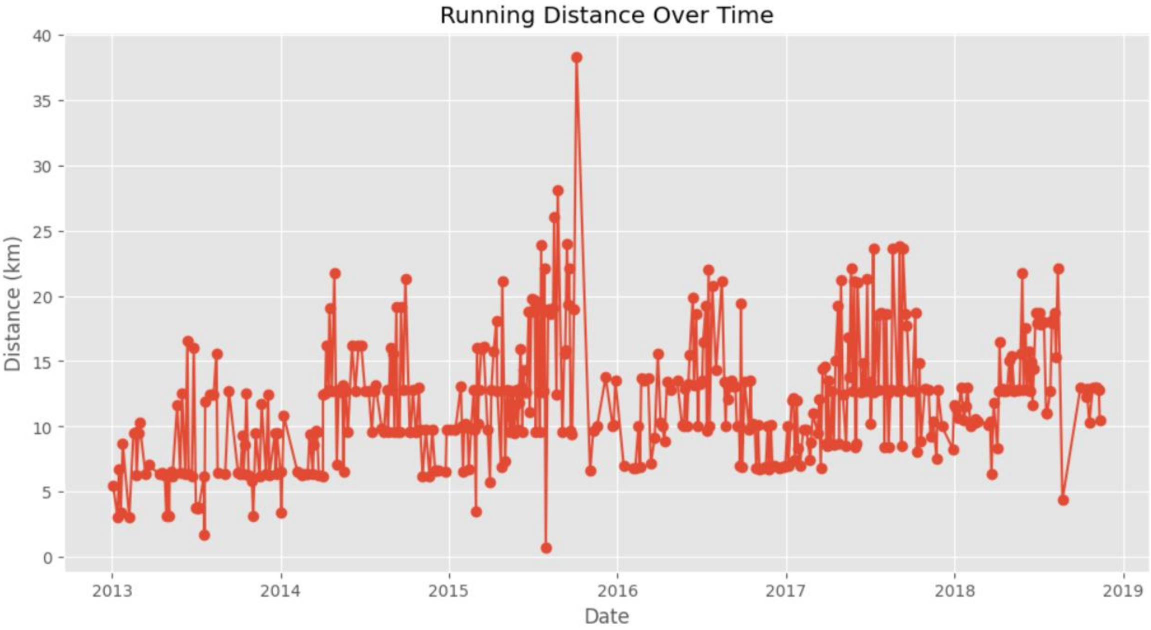
Data visualization involved creating informative and interactive plots to represent the analysis results. This was done using Matplotlib and other visualization libraries in Python. Key visualizations included:

- **Time Series Plots:** To show trends in distance, heart rate, and other metrics over time.
- **Histograms:** To display the distribution of heart rates across different activities.
- **Scatter Plots:** To visualize relationships between different metrics.

Here is an example of creating a time series plot:



```
import matplotlib.pyplot as plt

# Plotting running data from 2013 to 2018
fig, ax = plt.subplots(figsize=(12, 6))
ax.plot(runs_subset_2013_2018.index, runs_subset_2013_2018['Distance (km)'], marker='o', linestyle='-')
ax.set(title='Running Distance Over Time', xlabel='Date', ylabel='Distance (km)')
plt.show()
```


Running Distance Over Time
Distance (km)
Date
0
5
10
15
20
25
30
35
40
2013
2014
2015
2016
2017
2018
2019

CONCLUSION

The "Analyzing Fitness Data" project successfully achieved its objective of providing a comprehensive analysis of various fitness activities. Through meticulous data collection, cleaning, filtering, and analysis, we were able to gain valuable insights into training patterns and performance metrics. Key conclusions from the project include:

1. **Understanding Activity Trends:** By analyzing data over several years, we identified significant trends in fitness activities such as running, cycling, and walking. This allowed us to observe seasonal patterns, annual improvements, and the impact of external factors on training habits.
2. **Performance Metrics:** Detailed examination of metrics like distance, duration, average speed, and heart rate provided a deeper understanding of physical performance. For example, the analysis highlighted how heart rates varied across different activities and how training intensity changed over time.
3. **Comparative Analysis:** The project enabled comparisons across different types of activities. We observed that running typically resulted in higher average heart rates compared to cycling and walking, reflecting differences in intensity and cardiovascular demand.
4. **Data-Driven Insights:** Through the use of visualizations, we were able to present data in an intuitive and interactive manner, making it easier to interpret and derive actionable insights. This approach facilitated the identification of key performance indicators and trends.
5. **Informed Decision Making:** The insights gained from this project can be used by individuals and fitness enthusiasts to make informed decisions about their training regimens. Understanding personal performance trends and comparing them with broader patterns can help in setting realistic fitness goals and optimizing training plans.
6. **Foundation for Future Work:** This project lays a solid foundation for future research and analysis. The methodologies developed here can be extended to incorporate additional data sources, more sophisticated analysis techniques, and advanced visualization tools. Future work could explore the integration of machine learning models to predict performance outcomes and provide personalized training recommendations.

In conclusion, the "Analyzing Fitness Data" project demonstrated the power of data analytics in uncovering meaningful patterns and insights from fitness data. By leveraging Python and data visualization libraries, we created a comprehensive analysis framework that not only enhances our understanding of fitness activities but also supports the broader goal of promoting healthier lifestyles through data-driven decision making.

FUTURE SCOPE

The "Analyzing Fitness Data" project has laid a strong foundation for understanding and interpreting fitness activity data. Looking forward, there are several avenues for expanding and enhancing this work to provide even more value to users and researchers. The future scope of this project includes:

1. **Integration of Additional Data Sources:**
   - **Wearable Devices:** Incorporate data from wearable fitness trackers (e.g., Fitbit, Apple Watch, Garmin) to include metrics like sleep patterns, calorie expenditure, and step count.
   - **Environmental Data:** Integrate weather and air quality data to analyze the impact of environmental factors on training performance and activity choices.
2. **Advanced Analytical Techniques:**
   - **Machine Learning Models:** Develop predictive models to forecast performance metrics based on historical data and personal attributes. Machine learning algorithms can also be used to classify activities and detect anomalies in training patterns.
   - **Deep Learning:** Apply deep learning techniques for more sophisticated analysis of time-series data, enabling the identification of complex patterns and trends.
3. **Personalized Recommendations:**
   - **Adaptive Training Plans:** Create personalized training plans based on individual performance data, goals, and preferences. Use machine learning to adjust these plans dynamically as new data is collected.
   - **Health Monitoring:** Develop algorithms to monitor health metrics and provide real-time feedback and alerts for potential health issues (e.g., overtraining, irregular heart rates).
4. **Enhanced Visualizations:**
   - **Interactive Dashboards:** Implement more interactive and user-friendly dashboards using advanced visualization tools like Plotly, Dash, or Tableau.
   - **Virtual Reality (VR) and Augmented Reality (AR):** Explore the use of VR and AR to create immersive visualizations of fitness data, providing users with a more engaging experience.
5. **Community and Social Features:**
   - **Social Sharing:** Enable users to share their achievements and progress on social media platforms, fostering a sense of community and motivation.
   - **Competitive Analysis:** Introduce features that allow users to compare their performance with peers or participate in virtual challenges and competitions.
6. **Longitudinal Studies:**
   - **Behavioral Analysis:** Conduct longitudinal studies to analyze changes in fitness behavior over extended periods. This can provide insights into how life events (e.g., job changes, injuries, pandemics) affect physical activity levels.
   - **Impact Assessment:** Assess the long-term impact of different training programs and interventions on overall health and fitness.
7. **Health and Wellness Integration:**
   - **Holistic Health Insights:** Integrate data on nutrition, mental health, and lifestyle habits to provide a more holistic view of user health.
   - **Preventive Health:** Use data analytics to identify early warning signs of health issues and recommend preventive measures.
8. **Scalability and Accessibility:**
   - **Cloud-Based Solutions:** Develop cloud-based solutions to handle larger datasets and provide real-time analytics. This can also facilitate easier sharing and collaboration among users and researchers.

- **Mobile Applications:** Create mobile applications to make fitness data analysis more accessible and convenient for users on the go.

By pursuing these future directions, the "Analyzing Fitness Data" project can evolve into a comprehensive platform that not only aids individual fitness enthusiasts but also contributes to broader health and wellness research. The potential for innovative applications and impactful insights is vast, making this an exciting area for continued exploration and development.

REFERENCES

1. **Pandas Documentation:** Comprehensive documentation for the pandas library, which is used for data manipulation and analysis. Available at: https://pandas.pydata.org/docs/
2. **Matplotlib Documentation:** Official documentation for Matplotlib, a plotting library used for creating static, interactive, and animated visualizations in Python. Available at: https://matplotlib.org/stable/contents.html
3. **Seaborn Documentation:** Official guide and reference for Seaborn, a statistical data visualization library based on Matplotlib. Available at: https://seaborn.pydata.org/
4. **Scikit-learn Documentation:** Official documentation for scikit-learn, a machine learning library for Python. Available at: https://scikit-learn.org/stable/documentation.html
5. **Data Cleaning Techniques:** A guide on data cleaning techniques in Python, which is crucial for preparing datasets for analysis. Available at: https://realpython.com/python-data-cleaning-numpy-pandas/
6. **Python Data Science Handbook:** Comprehensive reference for data science in Python, including data manipulation, visualization, and machine learning techniques. Written by Jake VanderPlas. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/
7. **Kaggle Datasets:** A platform providing access to a wide variety of datasets, including those related to fitness and health. Available at: https://www.kaggle.com/datasets
8. **Statistical Analysis and Visualization Using R:** An online resource for learning R for statistical analysis and visualization, which can be adapted for use in Python. Available at: https://r4ds.had.co.nz/
9. **CDC - Physical Activity Guidelines:** Guidelines and statistics on physical activity from the Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/physicalactivity/index.html
10. **Wearable Fitness Tracker Data:** Information on integrating and analyzing data from wearable fitness trackers like Fitbit and Garmin. Available at: https://towardsdatascience.com/fitbit-data-analysis-visualization-with-python-a9bc7f92d3c
11. **Machine Learning with Python Cookbook:** Practical recipes for performing machine learning tasks with Python, by Chris Albon. Available at: https://www.oreilly.com/library/view/machine-learning-with/9781491989371/
12. **Deep Learning for Time Series Forecasting:** A comprehensive guide on using deep learning techniques for time series forecasting. Available at: https://machinelearningmastery.com/deep-learning-for-time-series-forecasting/
13. **Python for Data Analysis:** A book by Wes McKinney that covers essential data analysis tools in Python, including pandas. Available at: https://www.oreilly.com/library/view/python-for-data/9781491957653/
14. **Plotly Documentation:** Official documentation for Plotly, a graphing library that makes interactive, publication-quality graphs online. Available at: https://plotly.com/python/
15. **Dash Documentation:** Guide for using Dash, a productive Python framework for building web applications. Available at: https://dash.plotly.com/introduction