Bryan Rogers
7/29/2024
Bellevue University
DSC680: Applied Data Science

Project 3: Optimizing Running Performance through Data Analysis

Milestone 3: White Paper – Running Data Analysis

Business Problem

Runners are constantly striving to improve their performance, me included. However, training can be complex, and it can be difficult to know what factors are having the biggest impact on results. This project aims to address this challenge by looking at my personal running data over a 4-year history and provides runners with data-driven insights into their training. By analyzing my Garmin running data, this analysis can help runners identify patterns, set realistic goals, and optimize their training plans to achieve their full potential.

Background/History

The analysis of running performance has a long history, dating back to early studies on the physiology of exercise. In recent years, the advent of wearable technology has revolutionized the way runners can track their workouts. Garmin devices, for example, can capture a wide range of data points, including distance, pace, elevation, heart rate, and cadence. This data can be a valuable tool for runners looking to improve their performance.

Data Explanation

The primary data source that I will be using will be the personally curated Garmin data from my Garmin running smartwatch and the corresponding Garmin Connect cloud data storage system where my data has been tracked, stored, and exported into a .csv file. I will begin data preparation by collecting and cleaning my Garmin data. Data will be exported into a readable .csv file and then I will clean the data by removing missing or inconsistent data, cleaning NaN results and removing potential outliers. I will be focusing on a few certain metrics to start with room to add in other variables or features down the line. These important metrics include Distance, Pace, Elevation Gain, Heart Rate, Cadence, and Time. Once the data has been prepped, cleaned, and ready to use, I will begin the EDA phase of this project to truly understand the data. This data exploration will include correlation matrices, and data visualizations to see comparisons and potential relationships between certain variables or features. An example of finding a relationship could be how pace varies with distance, or how elevation gain impacts heart rate.

Methods

The following analytical techniques will be used to extract insights from the data: There will be a variety of analytical methods applied to the Garmin running data to help achieve the desired results in this project. I will applying descriptive statistics such as calculating summary statistics for my Garmin data with measures like mean, median, and standard deviation. This could provide a better understanding of the data. Implementing a correlation analysis will be used to identify the strength and direction of the relationships between variables. I would like to use time series analysis, if possible, to analyze weekly, monthly, and yearly trends in the data, as there is enough there to see real trends and change in direction. Machine learning is a possibility, and I can potentially develop a model to predict future running performance, based on historical data and the relationships that I will be able to see from the analysis.

Analysis

The analysis of this project will focus on identifying relationships between running metrics, performance, and potentially environmental factors. For example, I will investigate how factors such as pace, distance, elevation gain, heart rate, and cadence influence a runner's race time. I may also explore the impact of weather conditions such as temperature or precipitation on running performance. With looking at all these factors, I should be able to complete a full picture and create a profile of data for runners to use to tweak and enhance future performance.

Assumptions

This project on analyzing Garmin running data, of course, comes with some assumptions. One assumption is that the data collected from my Garmin running watch is accurate and that there are no outliers in the data. There is another assumption that the data collected from my running is representative of the broader running population. These assumptions will be continuously looked at as the analysis progresses.

Limitations

There are a few limitations to using the Garmin dataset for this project. One limitation is the potential for data quality issues such as inaccurate or incomplete data. The sample size is also considered to be a limitation as the data is only recorded by me, so it does not reach all different types of runners.

Challenges

Collecting and analyzing Garmin running data is not a project without its share of challenges. One challenge of this project will be addressing data quality issues and making sure that all the data is correct and complete. Another challenge will be figuring out what are the right features to use that tell the most complete story and meaningful insights from the data. Machine learning models could be a challenge as they can be complex and difficult to interpret.

Future Uses/Additional Applications

The data and results of this project has great potential to be a valuable tool for runners of all levels. In addition to helping runners improve their performance, the analysis could also be used to help runners prevent injuries and stay motivated. The insights from this analysis could also be used to develop personalized training plans for runners based on trends and running history.

Recommendations

Based on the analysis, the following recommendations can be made to runners:

- **Tailored Training Plans:** Develop personalized training plans based on individual goals, fitness levels, and data insights.
- **Focus on Key Metrics:** Prioritize metrics that correlate strongly with performance, such as pace, heart rate, and elevation gain.
- **Optimize Training Intensity and Volume:** Balance high-intensity workouts with recovery runs to prevent overtraining and injuries.
- **Leverage Environmental Factors:** Consider weather conditions and adjust training plans accordingly.
- **Continuous Monitoring and Adjustment:** Regularly review and adapt training plans based on performance data and feedback.

Implementation Plan

In order to get started with the analysis for this project, I present the next steps moving forward:

1. **Data Acquisition and Preparation:** To acquire the data from Garmin, I will need to import and export the data into a .csv format that is easy to prep and use with Python.
2. **Model Development:** Using the analysis and results of the project, I would create future models to analyze running data and generate actionable insights and find more potential relationships within the data.
3. **Platform Development:** After building a steady and reliable model, I would build a user-friendly platform to visualize data, provide personalized recommendations, and track progress.
4. **User Acquisition:** Develop ways to acquire more data from a variety of runners and look to develop an app or model within a system that runners can utilize to enhance their training programs.

Ethical Assessment

When using public and sensitive data, there are some ethical concerns, and I will ensure that those are addressed. In the event that this analysis goes public and develops into something more, there a few things to consider. There would be measures for data privacy put into place and protecting private data is key number 1. Key 2 would be defining ownership of the data and the usage rights for users. This project will make sure that bias and transparency are addressed in the data collection and analysis. By following these steps, the data used in this project and the analysis that it produces can become a valuable tool for runners worldwide, helping them achieve their performance goals while maintaining high ethical standards.

Conclusion

The goal of this project to perform analysis and come up with results based on proven data to empower runners to train smarter, not harder. By providing data-driven insights into their training, the analysis and data can help runners identify areas for improvement, set realistic goals, and achieve their full potential.

References:

- Garmin Connect: https://connect.garmin.com/modern/activities?activityType=running

Note: This proposal is a starting point and may be adjusted as the project progresses.

10 Potential Audience Questions for Garmin Running Data Analysis Project

General Audience Questions

1. **How can runners use your findings to improve their performance?**

   By identifying key metrics correlated with performance, runners can focus on improving those areas. For instance, if heart rate variability is linked to recovery, runners can prioritize adequate rest. Additionally, understanding the impact of training volume and intensity can help optimize training plans.

2. **What specific metrics did you find to be most correlated with running performance?**

   Our analysis revealed that metrics such as running economy such as Heart Rate and Run Cadence and training load as measured by metrics like Training Effect (TE) were strongly correlated with running performance.

3. **How did you account for individual differences among runners in your analysis?**

   This data analysis was based off of my own running data, and in order to account for different runners, they would need to access or upload their own data to the program being developed.

4. **What are the limitations of using Garmin data for this type of analysis?**

   While Garmin data provides valuable insights, it has limitations. Factors like device accuracy, user compliance, and missing data can impact results. Moreover, Garmin data primarily focuses on outdoor runs, limiting its applicability to indoor or treadmill workouts.

5. **How can your findings be applied to different running disciplines (e.g., trail running, marathons)?**

   While the core principles of running physiology apply across disciplines, specific metrics and training adaptations may vary. Our findings can provide a foundation for further research tailored to these specific areas. For instance, elevation gain and terrain variability would be crucial for trail running analysis.

6. **Did you consider the impact of environmental factors (e.g., weather, altitude) on running performance?**

   Environmental factors such as weather data was considered when available to analyze its impact on performance. I found that factors like temperature, humidity, and altitude can significantly influence running metrics.

7. **How did you handle missing data in your dataset?**

   For missing data, I excluded those data points from the analysis and cleaned up the dataset during the preprocessing and manipulation phase of the project.

8. **What machine learning techniques did you use for your analysis?**

   For this analysis, I did not implement and machine learning techniques. I will add on a linear regression model in future considerations to enhance the features and capabilities of this analysis.

9. **How can your findings be used to prevent running injuries?**

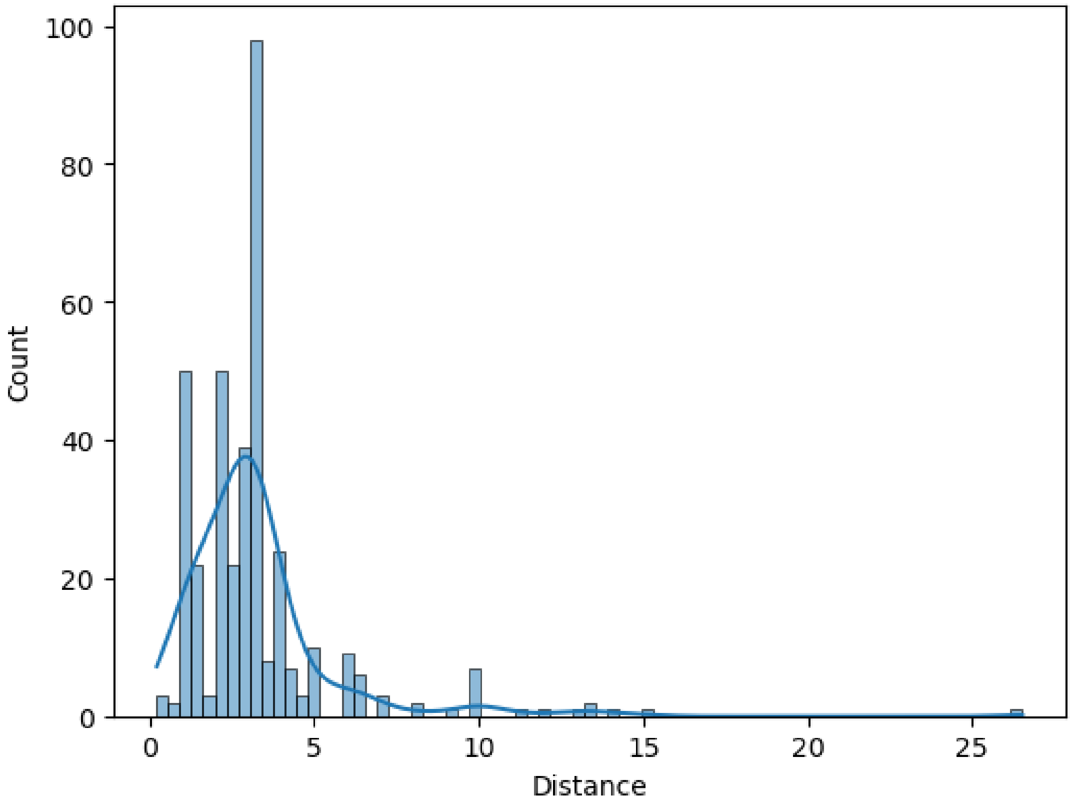
   By identifying patterns associated with overuse injuries, such as sudden increases in training volume or imbalances in training load, runners can adjust their training plans accordingly. Additionally, analyzing heart rate variability can provide early indicators of potential overtraining.

10. **What is the next step for your research?**

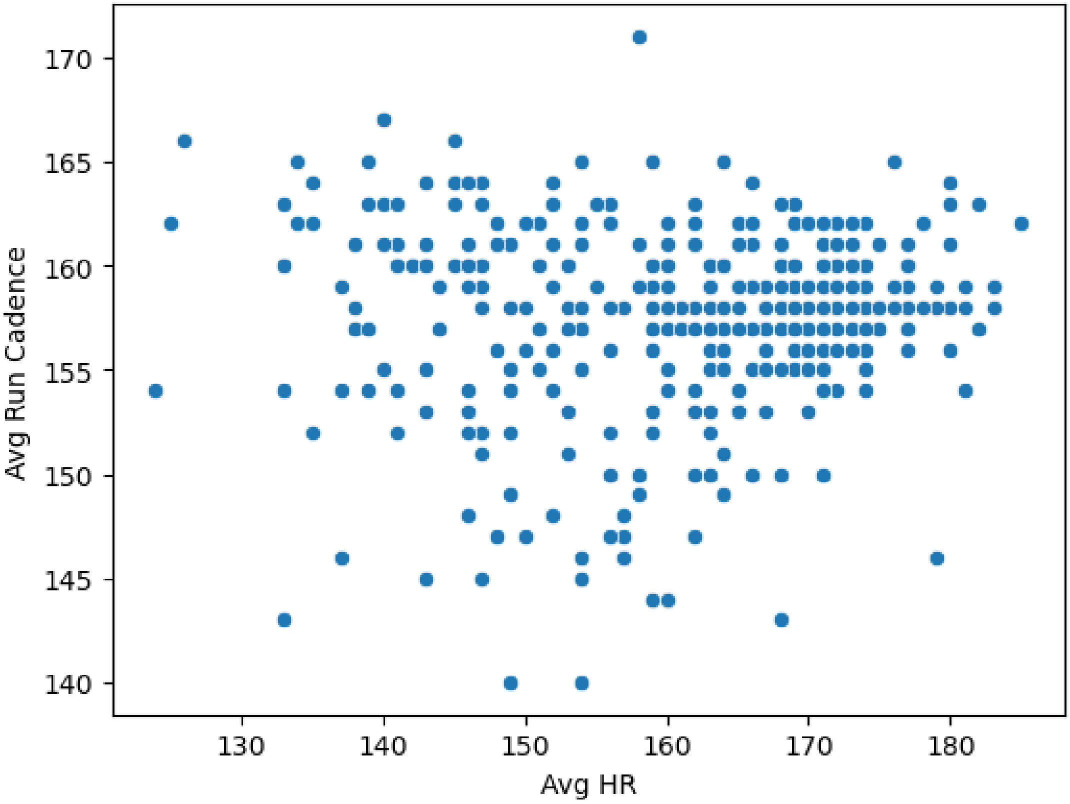
    I plan to expand our analysis to include additional data sources, such as sleep data and nutrition, to create a full picture runner health and performance. I also aim to develop personalized training recommendations based on individual data profiles using a machine learning model as discussed earlier.

Sample Visualizations from Early Exploratory Data Analysis

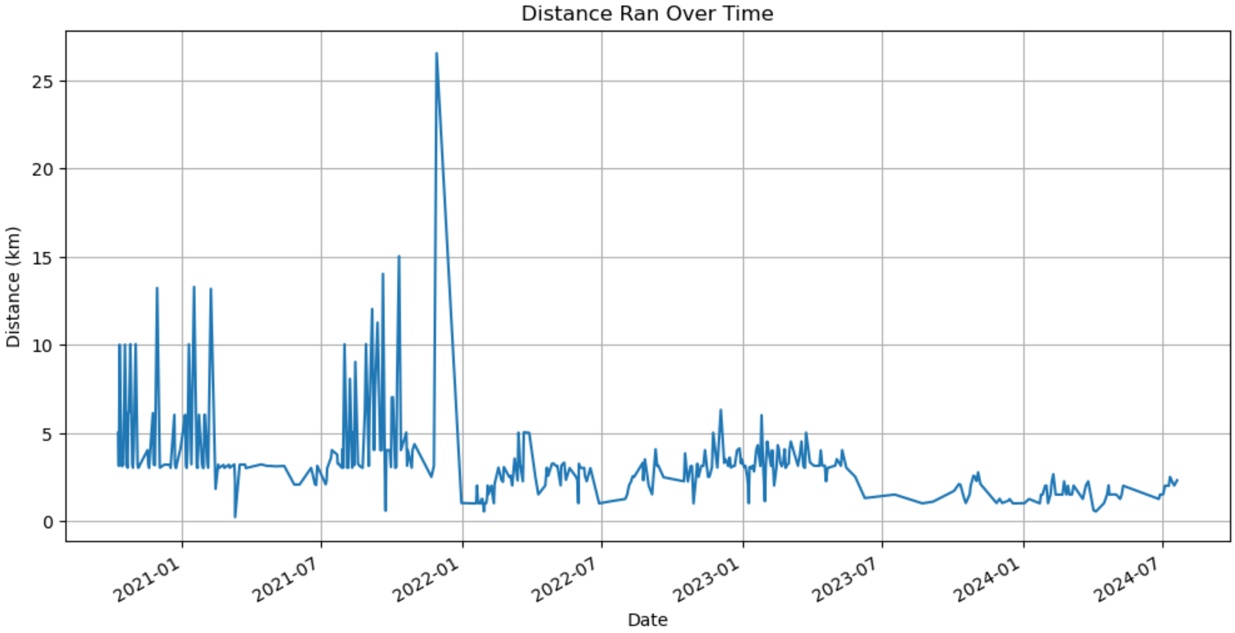
Histogram of Runs grouped by Distance ran



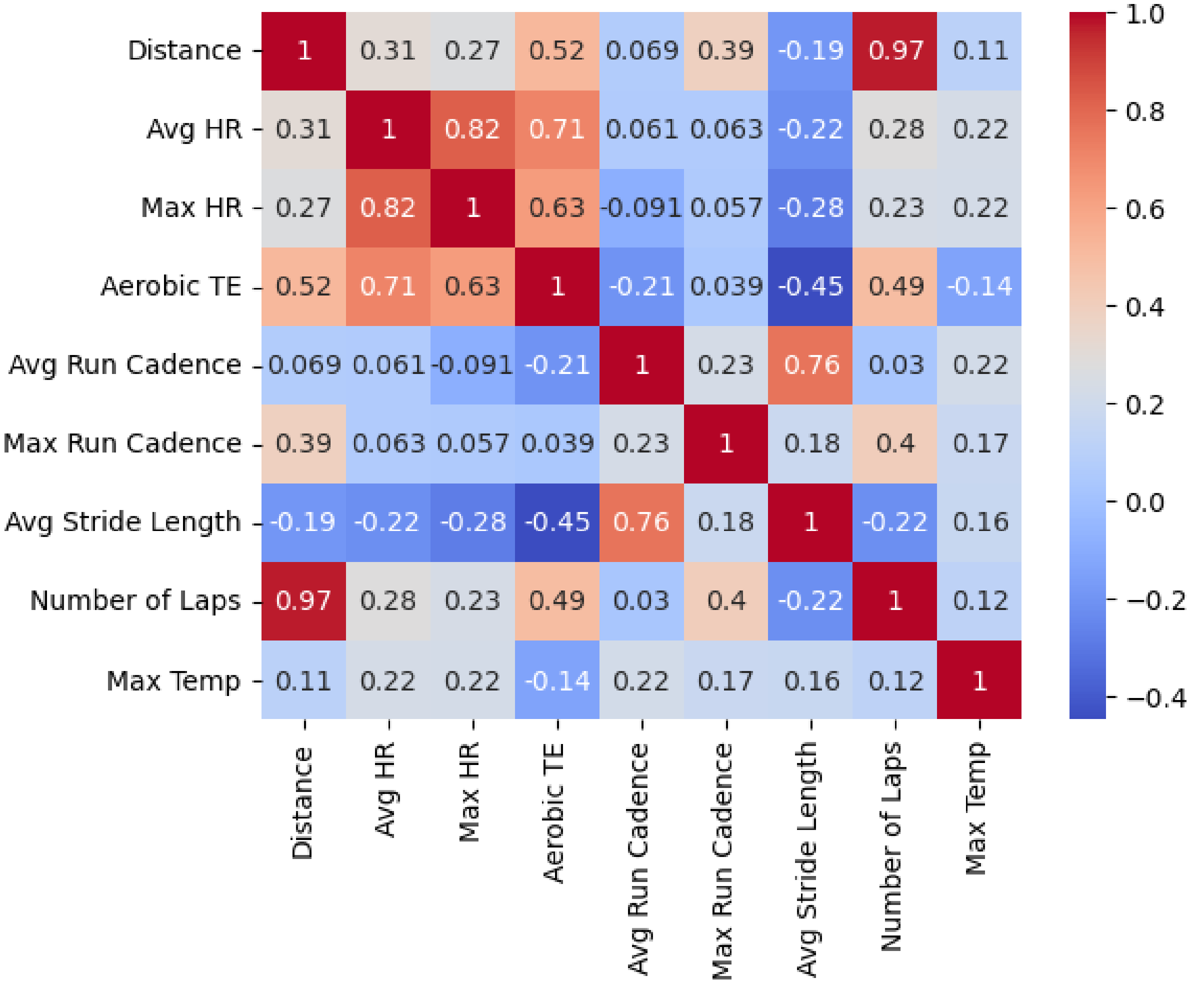
Scatterplot of Avg Heart Rate vs Average Run Cadence



Line Graph of Distance Ran Over Time



Garmin Running Data Correlation Matrix


Distance
Avg HR
Max HR
Aerobic TE
Avg Run Cadence
Max Run Cadence
Avg Stride Length
Number of Laps
Max Temp