DSA 210 TERM PROJECT
How Air Quality Affects Running Performance

1) INTRODUCTION

As someone who enjoys running, I often find myself confused by the fluctuations in my performance during outdoor running sessions. Some days, my performance feels effortless, while on others, I struggle to complete my planned route without any major changes in my routine. I always suspected this had something to do with an external factor, such as air quality. Living and running on the Sabancı University campus, where air quality can noticeably differ, led me to believe air pollution might be influencing my endurance, breathing, and overall efficiency in running.

This personal curiosity is a great foundation for my DSA210 term project. I wanted an answer so that I could understand the frequent changes in my performance. If it turned out that there was a correlation between my performance and air quality, I would no longer get discouraged when I had these “bad performance days”.

The aim of my project was to explore the impact of air quality on running performance by analyzing data from personal running sessions alongside publicly available pollution metrics, such as the Air Quality Index (AQI) and specific pollutant concentrations (PM2.5, NO_2). By combining personal performance data with publicly available environmental data, I wanted to investigate whether poor air quality correlates with measurable changes in pace, heart rate, and perceived effort.

The project applied data science techniques such as exploratory data analysis, hypothesis testing, and machine learning to uncover any hidden patterns in the data. The goal was not only to satisfy my curiosity but also to contribute data-driven insights to a broader conversation on how environmental conditions impact physical activity and endurance sports.

The following hypotheses have been tested throughout the project:

Null Hypothesis (H_0): Air quality has no significant effect on my running performance.
Alternative Hypothesis (H_1): Poor air quality negatively impacts my running performance.

2) DATA COLLECTION

To investigate the potential impact of air quality on running performance, I collected data from multiple outdoor running sessions under controlled conditions.

- All running performance data, including heart rate, pace, and distance, was collected using a Samsung Galaxy Watch.
- Each run was recorded on the same route within the Sabancı University campus to eliminate performance differences caused by terrain.
- I ensured consistency in running conditions by scheduling all sessions at approximately 7 AM, before breakfast, which aligns with my usual training routine. This timing helped minimize daily variations in temperature, humidity, and personal hydration or nutrition levels.
- Only outdoor runs were included in the dataset; treadmill or indoor workouts were excluded to maintain environmental consistency.

Air quality was investigated using AQI, PM2.5, NO_2, temperature, and humidity data as indicators. PM2.5 and NO_2 are two of the most critical pollutants affecting air quality and respiratory health, particularly during outdoor activities like running. PM2.5 refers to fine particulate matter that can penetrate deep into the lungs, while NO_2 is a harmful gas emitted from vehicle exhaust and industrial processes. Even though temperature and humidity do not seem to be directly related to air quality, they could have some potential effect on my performance, maybe combined with air quality data. Therefore, they were also recorded each day at around 7 AM.

Running performance was analyzed using my pace (in min/km), the distance I ran that session, and my heart rate (BPM). The aim was to see how these data related to the air quality metrics and to understand if there really was a correlation.

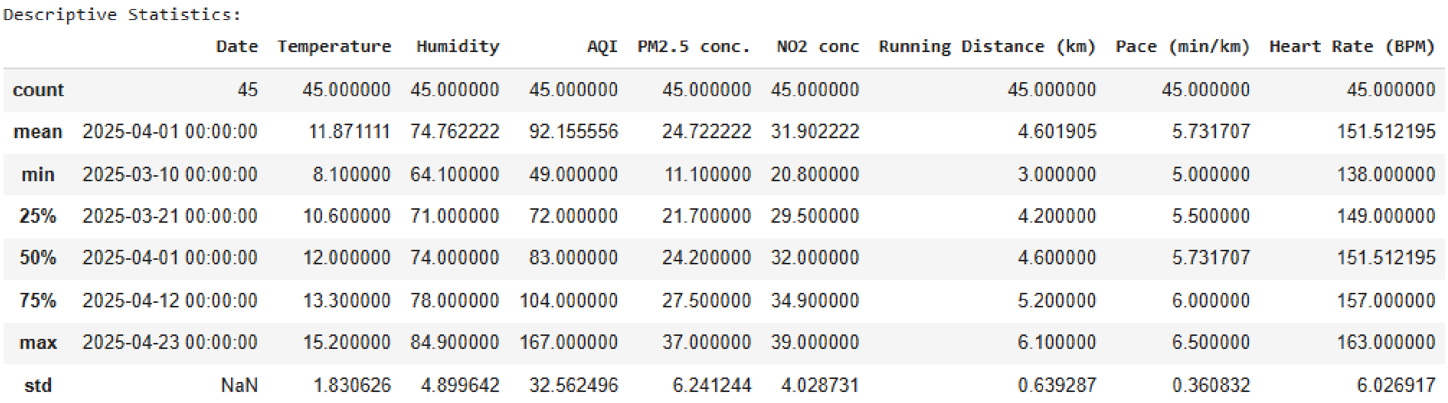
To complement the performance data, I gathered real-time air quality metrics from weather.com, which provides up-to-date values for AQI, PM2.5, NO_2, temperature, and humidity. I manually logged this information into an Excel spreadsheet within 30 minutes of each run to ensure that the environmental readings accurately reflected the conditions during the session.

The final dataset included the following variables:

Date: The specific day of each run.
Air Quality Index (AQI): A general measure of pollution levels.
PM2.5 and NO$_2$ Concentrations: Pollutant levels with known effects on respiratory function.
Temperature & Humidity: Environmental conditions recorded at the time of the run.
Running Distance (km): Kept consistent across sessions for comparative analysis.
Pace (min/km): Used to evaluate running efficiency.
Heart Rate (BPM): Used as an indicator of cardiovascular effort and strain.

I collected every entry in an Excel file daily. Even though I followed my usual training program, which consists of daily morning runs, there have been exceptional days when I could not run due to being sick or simply having a busy schedule. On those days, I kept recording air quality data, but kept the performance data empty to be later processed. Additionally, there has been one instance where there was a technical problem with my watch, so my pace and heart rate values were missing. All these missing values were cleaned by imputation, where they were filled with their corresponding mean values.

After I collected my data in around 40 training sessions, I analyzed the descriptive statistics of my set to get more information on it and obtained the following chart:

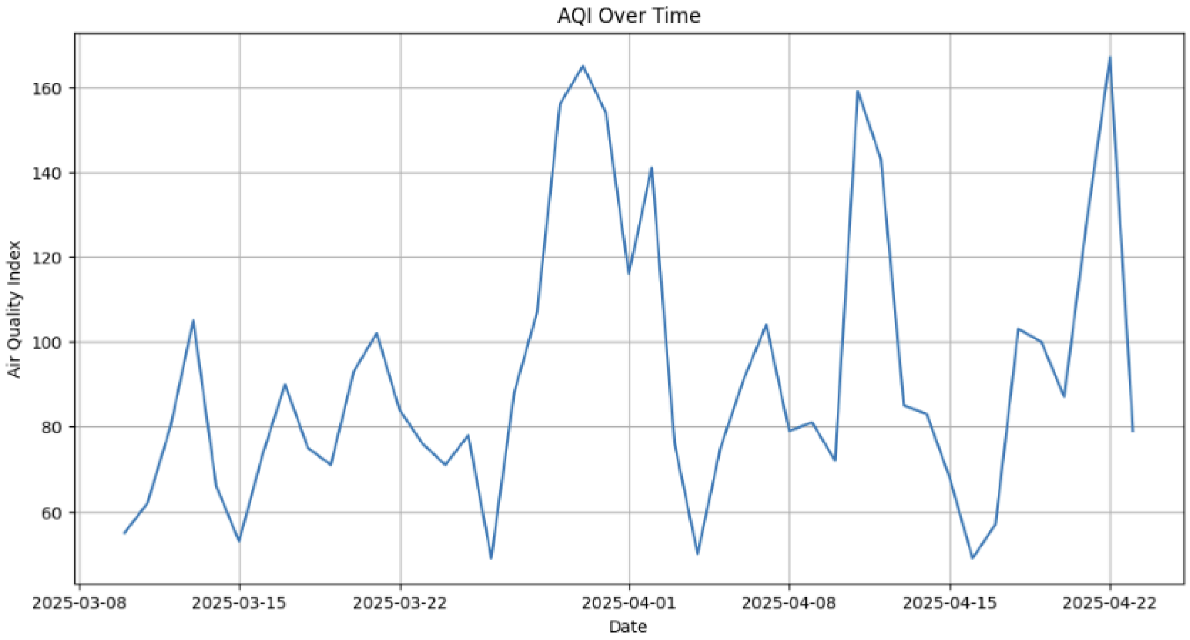


Descriptive Statistics:

	Date	Temperature	Humidity	AQI	PM2.5 conc.	NO2 conc	Running Distance (km)	Pace (min/km)	Heart Rate (BPM)
count	45	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000	45.000000
mean	2025-04-01 00:00:00	11.871111	74.762222	92.155556	24.722222	31.902222	4.601905	5.731707	151.512195
min	2025-03-10 00:00:00	8.100000	64.100000	49.000000	11.100000	20.800000	3.000000	5.000000	138.000000
25%	2025-03-21 00:00:00	10.600000	71.000000	72.000000	21.700000	29.500000	4.200000	5.500000	149.000000
50%	2025-04-01 00:00:00	12.000000	74.000000	83.000000	24.200000	32.000000	4.600000	5.731707	151.512195
75%	2025-04-12 00:00:00	13.300000	78.000000	104.000000	27.500000	34.900000	5.200000	6.000000	157.000000
max	2025-04-23 00:00:00	15.200000	84.900000	167.000000	37.000000	39.000000	6.100000	6.500000	163.000000
std	NaN	1.830626	4.899642	32.562496	6.241244	4.028731	0.639287	0.360832	6.026917

3) EXPLORATORY DATA ANALYSIS (EDA)

After collecting the data on daily logged AQI numbers, I initially obtained the following simple plot to see how AQI changed over time:

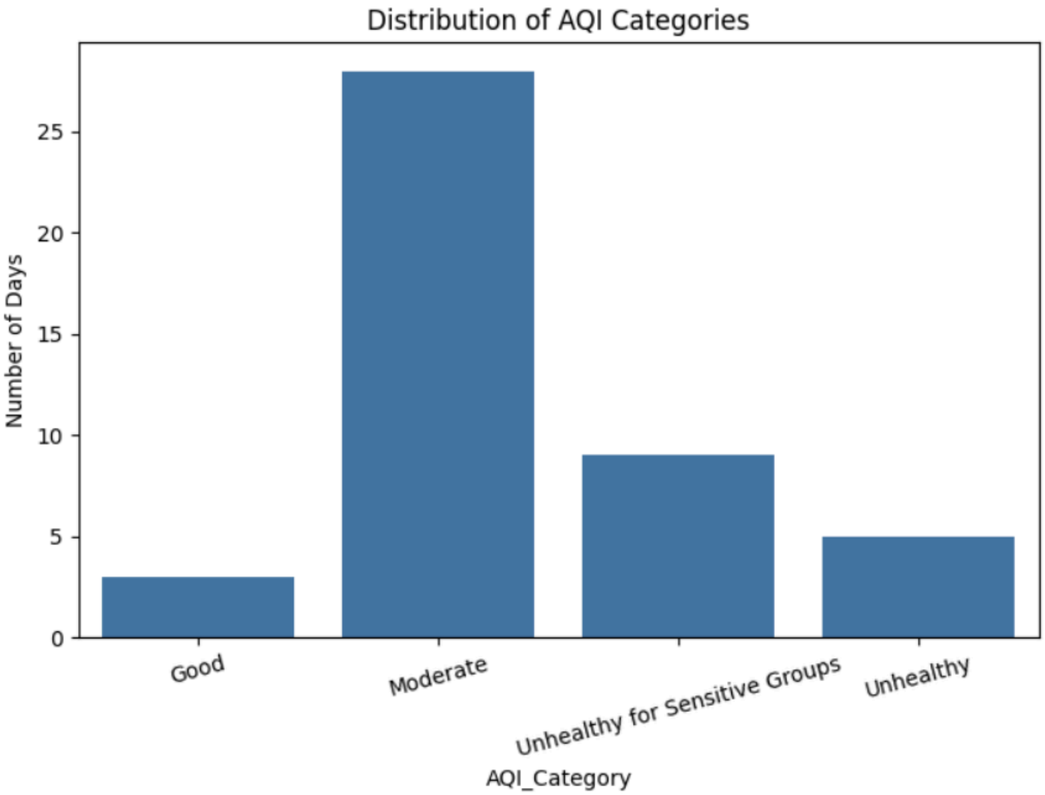


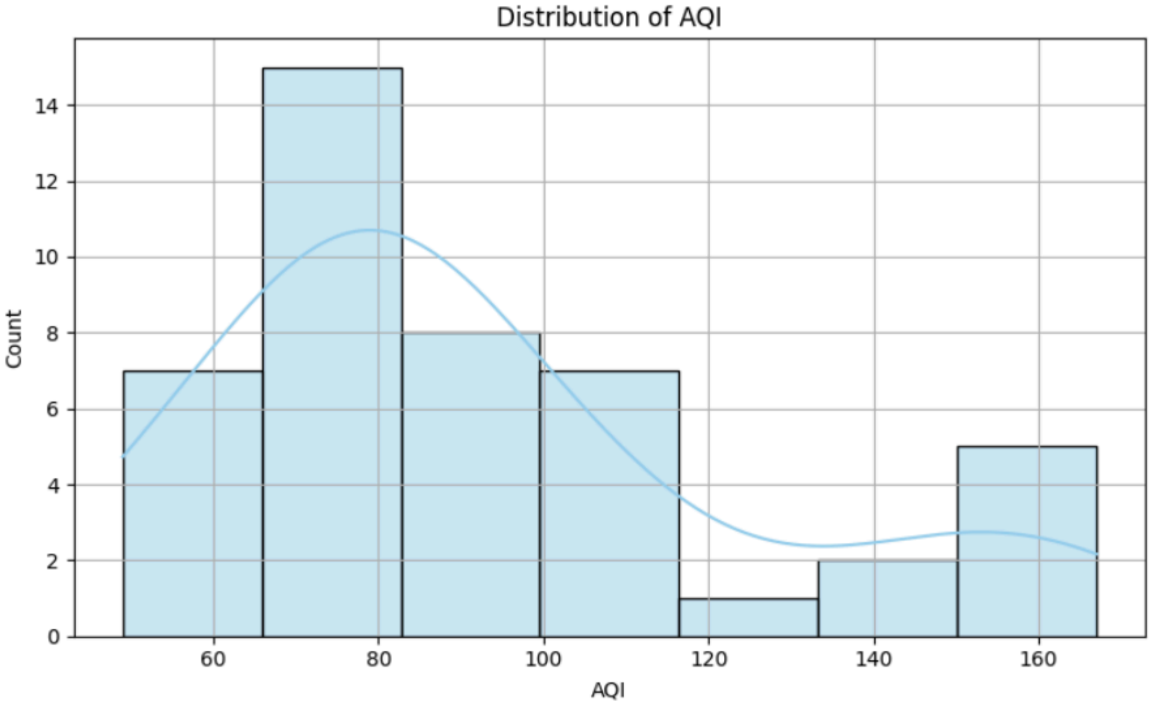
As expected, the AQI values in the Sabancı campus showed high variability every morning. This meant that this data could be further explored in order to see if AQI had relevance to my performance.

In order to categorize the different AQI values, I used the following definitions for different air quality levels:

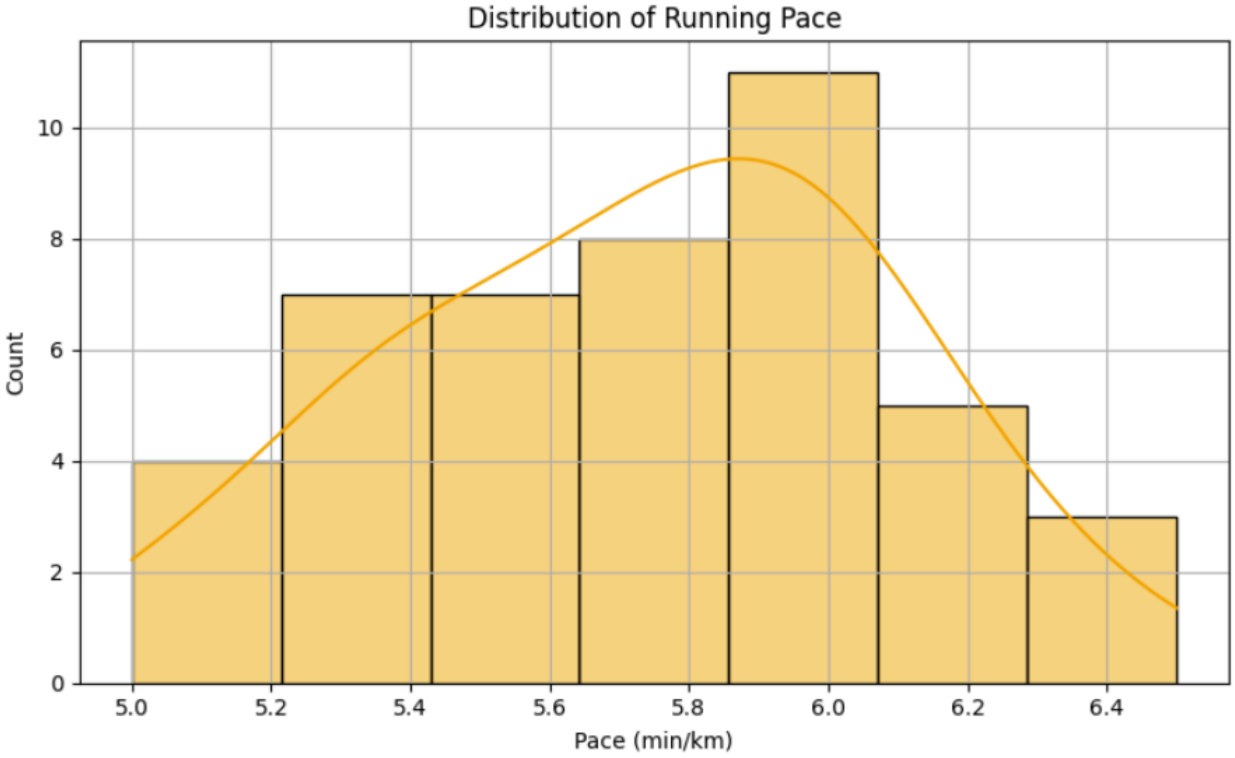
AQI < 50: **Good**
50 < AQI < 100: **Moderate**
100 < AQI < 150: **Unhealthy for Sensitive Groups**
AQI > 150: **Unhealthy**

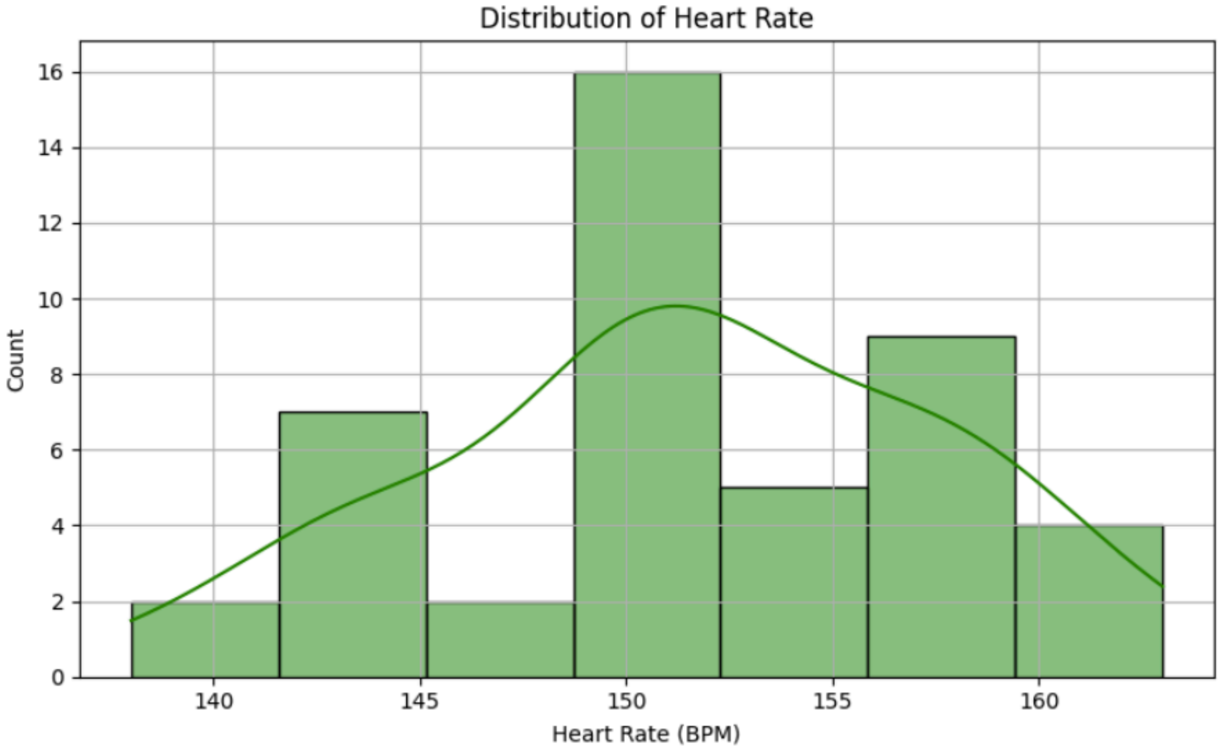
To see how many days in total were suited for which category, the histogram below was plotted. As can be seen from the histogram, the majority of the air quality values fell under the “Moderate” category, followed by the “Unhealthy for Sensitive Groups” category. The least occurring category was the “Good” air quality category, unfortunately.



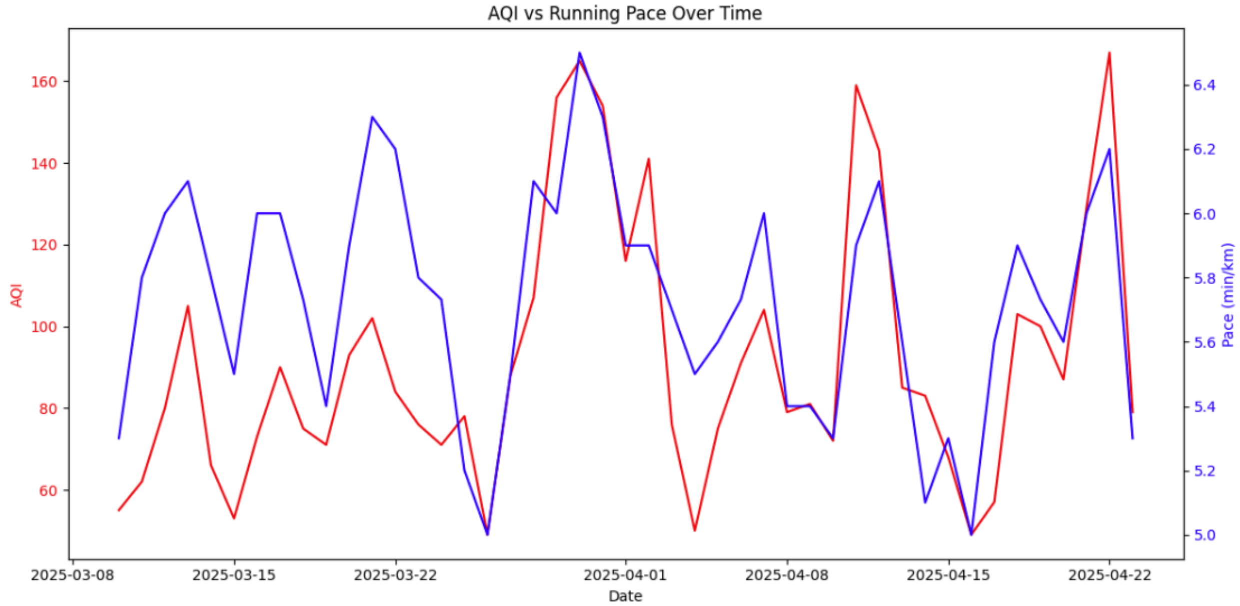


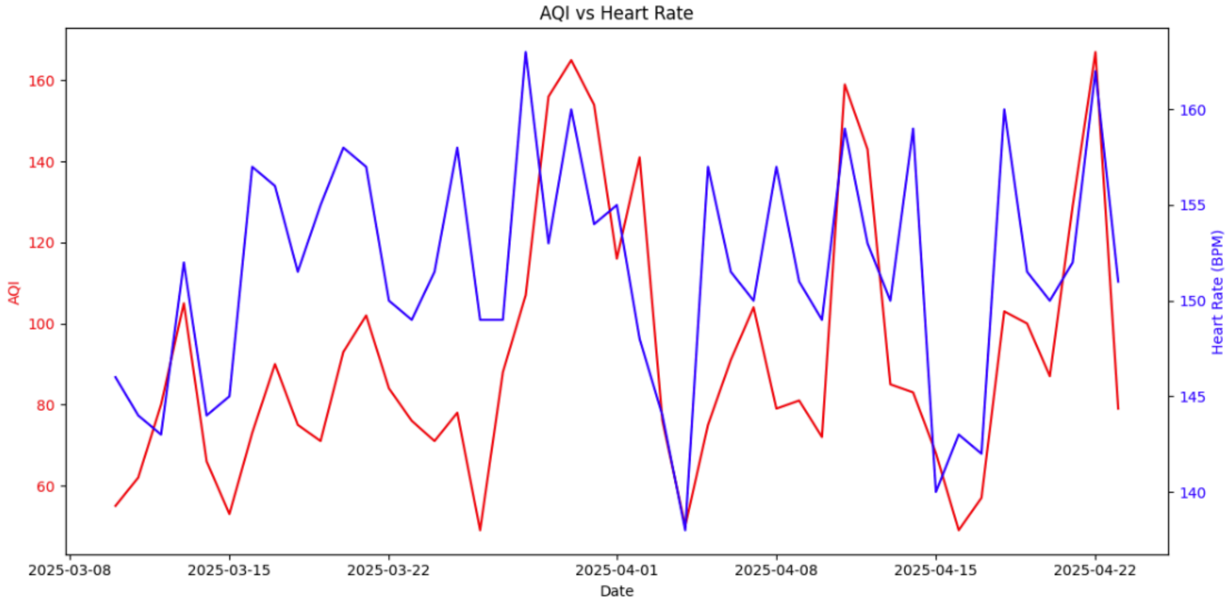
The histograms below display the distributions of running pace and heart rate across all sessions. Pace values mostly fall between 5.0 and 6.4 min/km, with a peak around 5.9, showing a slight skew toward slower runs, possibly linked to environmental factors. Heart rate readings are more tightly clustered around 150 BPM, indicating consistent levels. Both distributions appear roughly near-normal, supporting their use in further statistical analysis of how air quality may influence running performance.



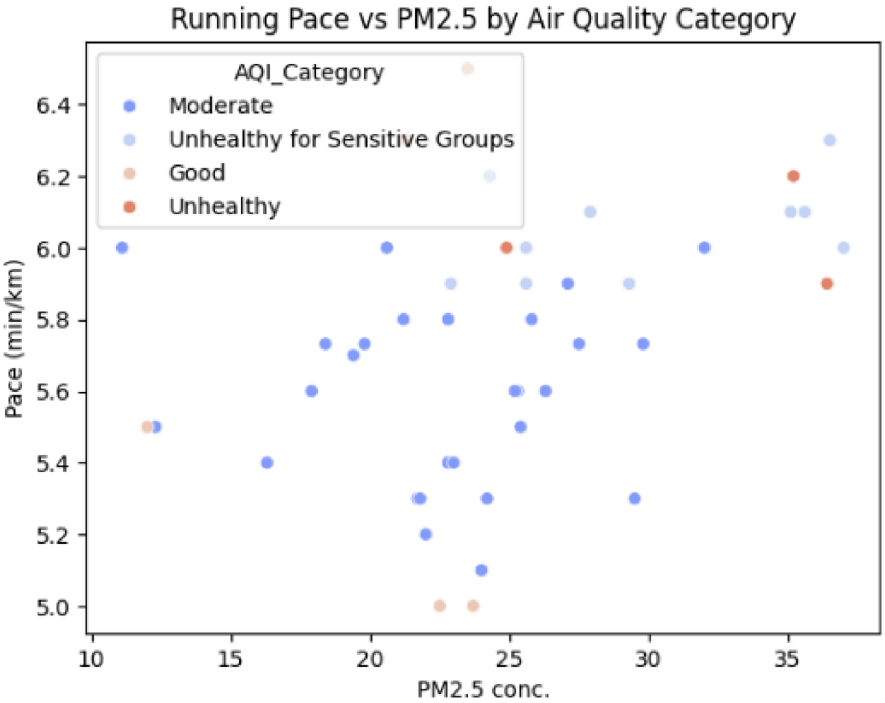


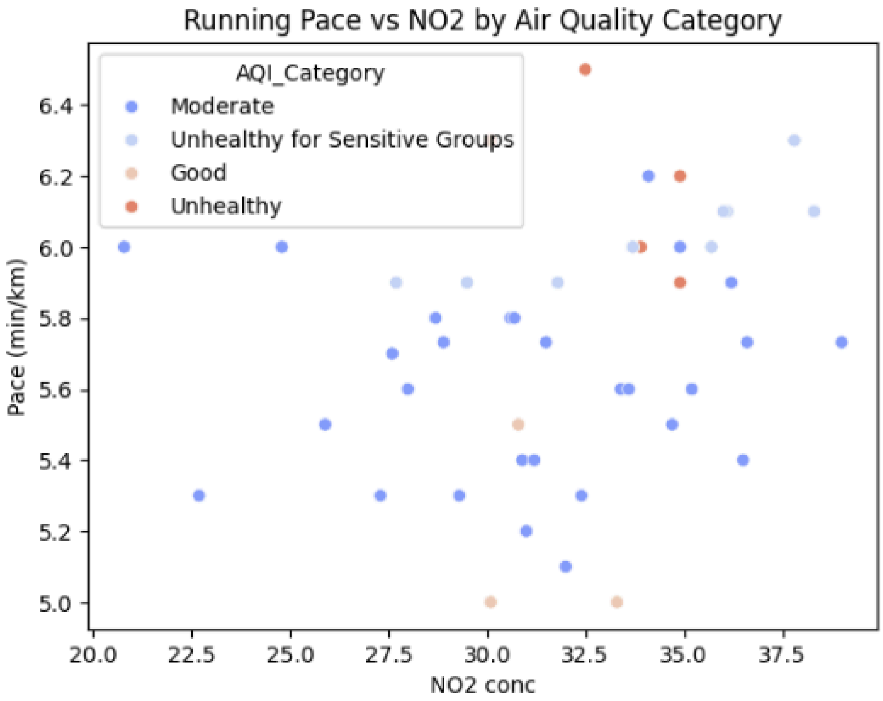
To see how AQI vs Pace and AQI vs Heart Rate data compare, the plots were combined in the following visualizations:



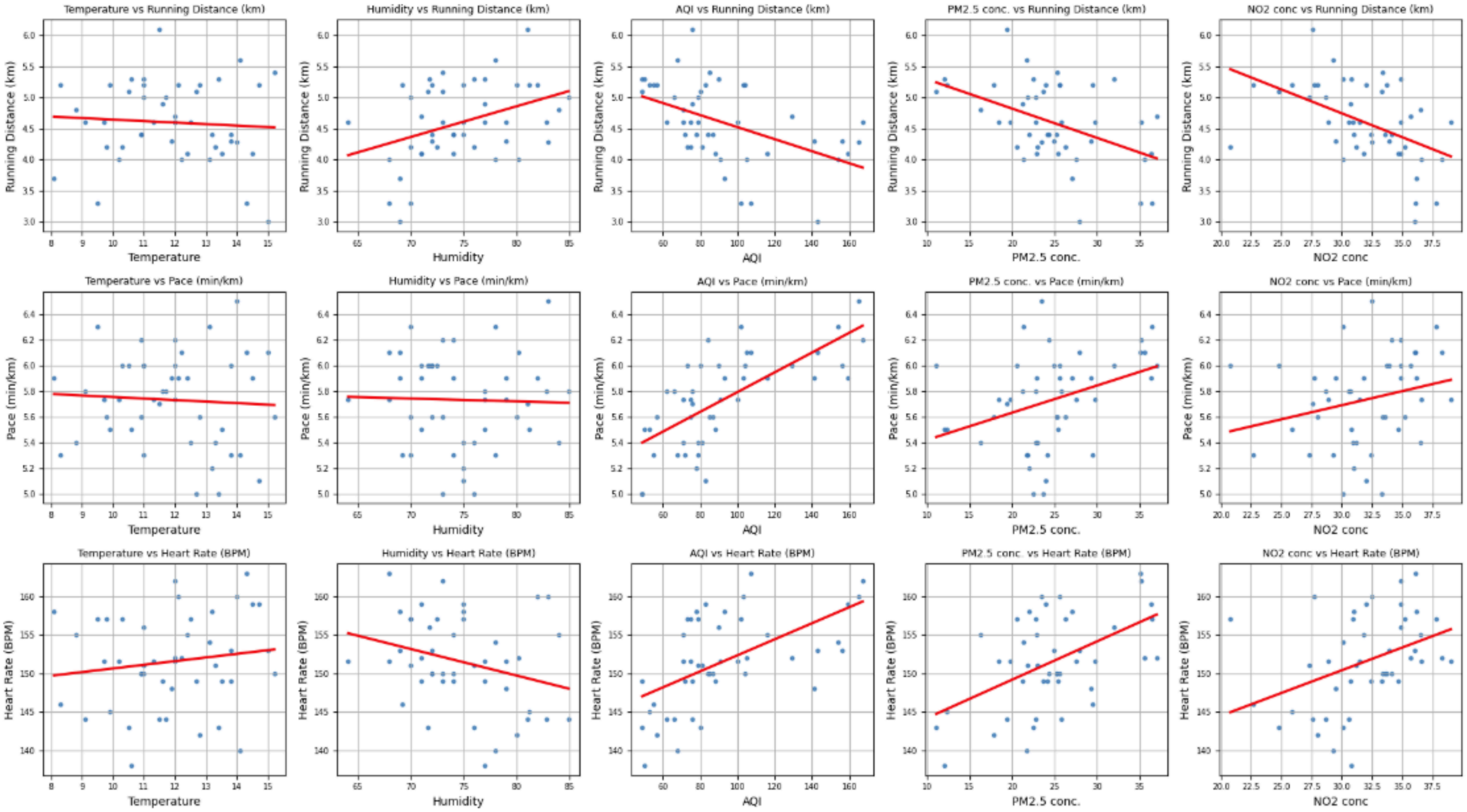


It was important to see how my running performance changed as the pollutant concentrations differed each day. Therefore, I decided to analyze a significant performance metric, like pace, together with AQI and pollutant concentrations. The scatter plots below illustrate that, with points color-coded by overall AQI category. In both the PM2.5 and NO_2 plots, there seems to be an upward trend, indicating slower running paces as pollutant levels increase, especially in “Unhealthy for Sensitive Groups” and “Unhealthy” categories. This suggests that high PM2.5 and NO_2 concentrations may alter performance, likely due to reduced oxygen consumption or high breathing discomfort. Even though these visualizations show the importance of pollutant-specific air quality measurements as physical performance predictors, there still needs to be further evaluation on how much performance is affected by them by conducting a correlation analysis later on.



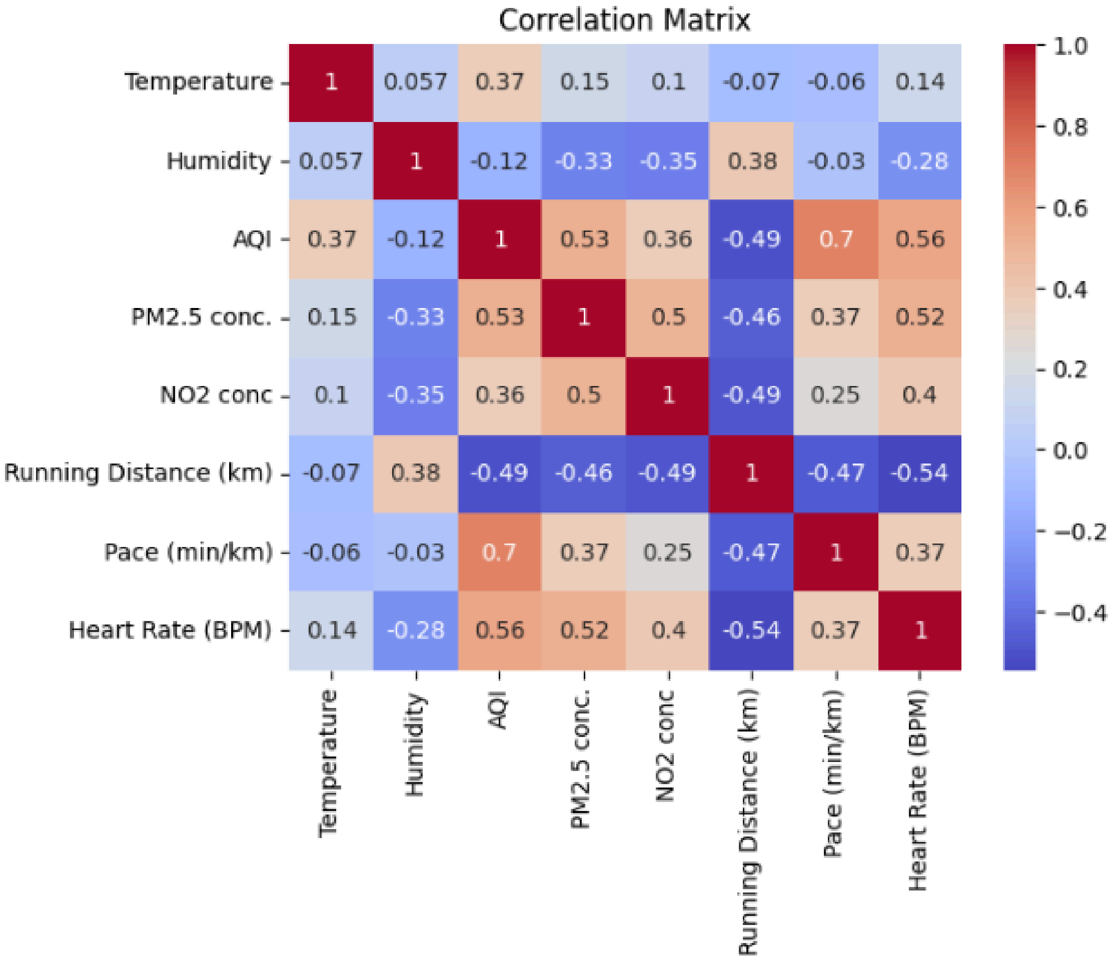


All combinations were plotted in order to see each relationship individually. There are many plots in the following visualization, and the correlation analysis in the following section will help understand which of these can meaningfully contribute to the project.



4) CORRELATION ANALYSIS

To observe the relationship between environmental factors and my running performance, I generated a correlation heatmap. Correlation plots were essential in identifying linear associations between numerical variables, which revealed the factors that might influence each other. This process would also guide the selection of my future predictors for ML modeling. In this analysis, I included both the air quality data and the performance data.

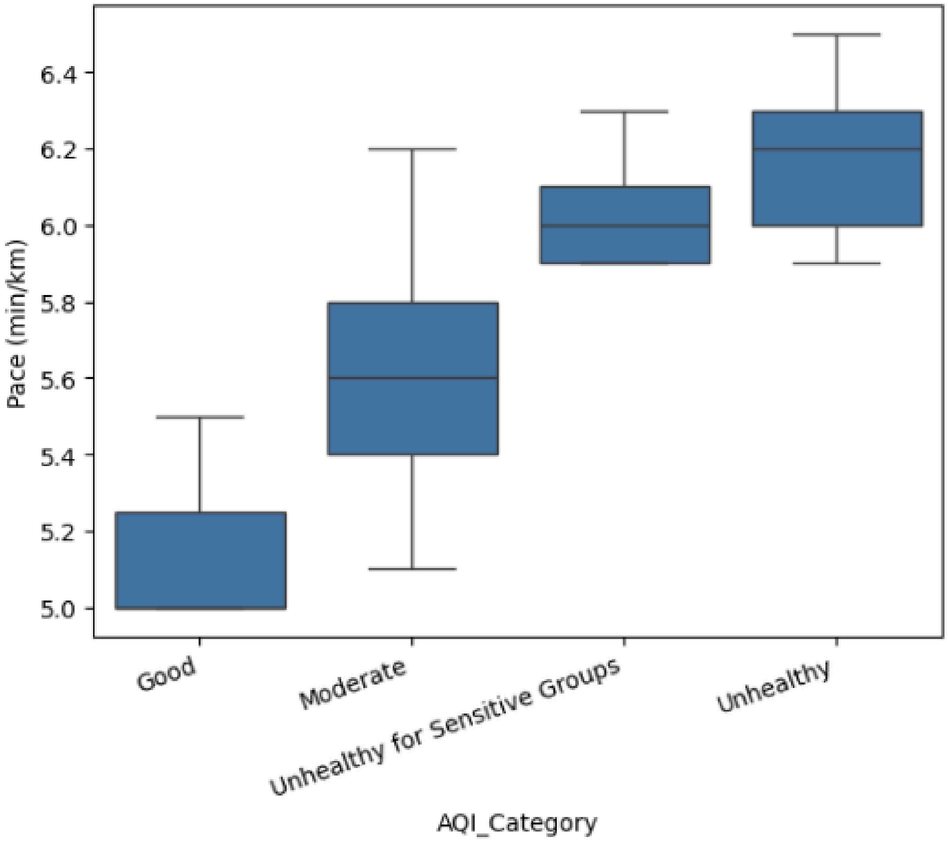


Correlation Matrix

	Temperature	Humidity	AQI	PM2.5 conc.	NO2 conc	Running Distance (km)	Pace (min/km)	Heart Rate (BPM)
Temperature	1	0.057	0.37	0.15	0.1	-0.07	-0.06	0.14
Humidity	0.057	1	-0.12	-0.33	-0.35	0.38	-0.03	-0.28
AQI	0.37	-0.12	1	0.53	0.36	-0.49	0.7	0.56
PM2.5 conc.	0.15	-0.33	0.53	1	0.5	-0.46	0.37	0.52
NO2 conc	0.1	-0.35	0.36	0.5	1	-0.49	0.25	0.4
Running Distance (km)	-0.07	0.38	-0.49	-0.46	-0.49	1	-0.47	-0.54
Pace (min/km)	-0.06	-0.03	0.7	0.37	0.25	-0.47	1	0.37
Heart Rate (BPM)	0.14	-0.28	0.56	0.52	0.4	-0.54	0.37	1

In order to explore how the variables relate to each other, I created a correlation matrix using Pearson coefficients and visualized it as a heatmap. In this plot, warmer colors show stronger positive relationships, while cooler tones represent negative ones. A few patterns stood out right away: for example, AQI had a clear positive correlation with both pace (0.70) and heart rate (0.56). This means that on days with worse air quality, I tended to run slower, and my heart rate was higher, possibly due to reduced breathing efficiency. I also noticed that PM2.5 and NO_2 levels followed a similar pattern, suggesting that these pollutants might be playing a direct role in how hard my body has to work during a run. On the other hand, running distance had negative correlations with all of the pollution indicators, which could imply that I might unintentionally shorten my sessions under worse air conditions. Overall, pace had the strongest correlation with AQI, and this led me to use pace as the most dominant performance indicator in future analysis.

I additionally created a box plot to further validate the relationship between air quality and pace, and compared running pace across the different AQI levels. The box plot made the relationship between air quality and running pace even clearer. As the air quality worsened, my running pace consistently increased, meaning I was running slower on more polluted days. The median pace was fastest under “Good” conditions and slowest under “Unhealthy”. The spread of the data varied in each category, having varied most under the “Moderate” category. These backed up what I saw in the correlation matrix; poor air quality seems to have a noticeable effect on how efficiently I can run.



The correlation analysis results explained offer strong support for the alternative hypothesis that pollution impacts performance, and lay the foundation for the more detailed hypothesis testing in the next section.

5) HYPOTHESIS TESTING

After exploring trends in my running performance and air quality, I wanted to statistically test whether the differences I observed were meaningful. To do this, I used two commonly applied hypothesis testing methods: a one-way t-test and a one-way ANOVA.

First, I focused on comparing the mean running pace between two extreme air quality categories: Good and Unhealthy. I used a one-sided independent t-test, which allowed me to check whether my pace was significantly slower on days with unhealthy air quality. The result was a very high t-statistic (around 5.41) and a p-value well below 0.05 (with p = 0.0008), which led me to reject the null hypothesis. This suggests that the difference in running pace between clean-air and polluted-air days is statistically significant, and not just due to chance.

Digging a little deeper, I also did a one-way ANOVA test to compare my pace across all four AQI categories: Good, Moderate, Unhealthy for Sensitive Groups, and Unhealthy . This was done to further verify the decision to reject the null hypothesis. ANOVA was great for this kind of analysis because it helped determine whether at least one group's mean is significantly different from the others. The test provided a very small p-value (around 2.82E-7), giving strong evidence to reject the null hypothesis and confirming that air quality category has a significant effect on running pace.

Overall, with the information I obtained from the hypothesis tests, my decision was to reject the null hypothesis. This confirmed that air quality indeed affects my running pace and performance, since pace was the highest correlated performance metric to air quality.

6) INSIGHTS ON EDA & HYPOTHESIS TESTING:

- Running pace tends to be slower on days with worse air quality. Especially on days when AQI was high, my average pace increased, which suggests I might be running more slowly in polluted conditions. Additionally, pace showed the highest correlation with AQI in the correlation heatmap. This supported my alternative hypothesis.
- Heart rate also seemed to go up slightly when AQI, PM2.5, or temperature increased. This could mean that my body was working harder when the air isn't as clean or when it was hotter outside.
- Running distance didn't change as dramatically, but on some high-AQI or very hot days, I ran shorter distances. It was proved in the correlation analysis that distance was negatively correlated with all the air quality metrics.

- Both the T-test and ANOVA test showed that there is a statistically significant difference in my pace between different AQI categories, which means the variation I saw in the plots is not just by chance. So, the null hypothesis was rejected.

With the hypothesis testing results confirming a statistically significant relationship between air quality and running performance, the next step was to explore whether this relationship could be modeled and predicted using machine learning techniques, which will be explained in the following section of the report.

7) MACHINE LEARNING

After confirming that air quality had a statistically significant impact on my running performance, I took the project one step further and explored if I could predict my pace, heart rate, or distance based on that day's air conditions, using machine learning. The goal was to be able to forecast my performance under different conditions, so that I could know of a "bad performance day" before it happened and not get demotivated.

During my analysis, the dataset was used to predict three performance metrics: Pace, heart rate, and running distance. I selected five commonly used ML methods for testing:

1. Linear Regression
2. Ridge
3. Lasso
4. Random Forest
5. Gradient Boosting

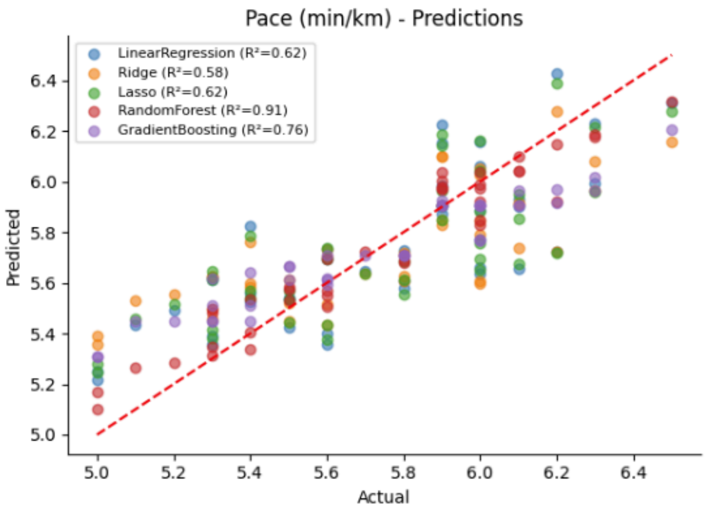
I included linear methods (like Ridge and Lasso) to establish baseline performance, while Random Forest and Gradient Boosting were chosen for their ability to model nonlinear relationships and handle interactions, which I suspected might exist between pollutants and performance responses.

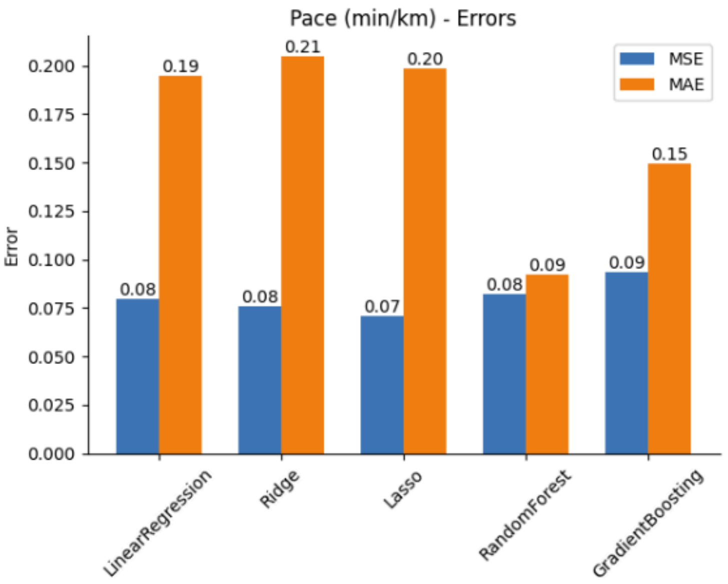
I built a model using scikit-learn, which included:

- SimpleImputer: For handling missing values using median imputation
- StandardScaler: To normalize input features
- GridSearchCV: To use a K-fold cross-validation method, where K=5.

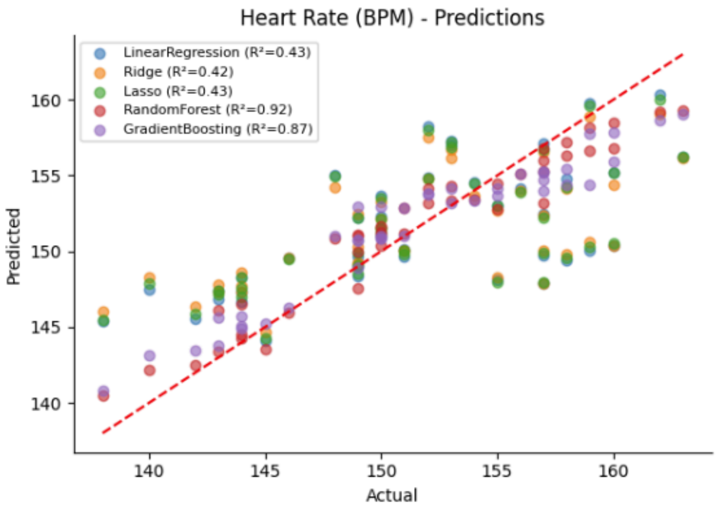
I used environmental factors like temperature, humidity, AQI, PM2.5, and NO_2 as input features. Each model was trained and evaluated using three metrics:

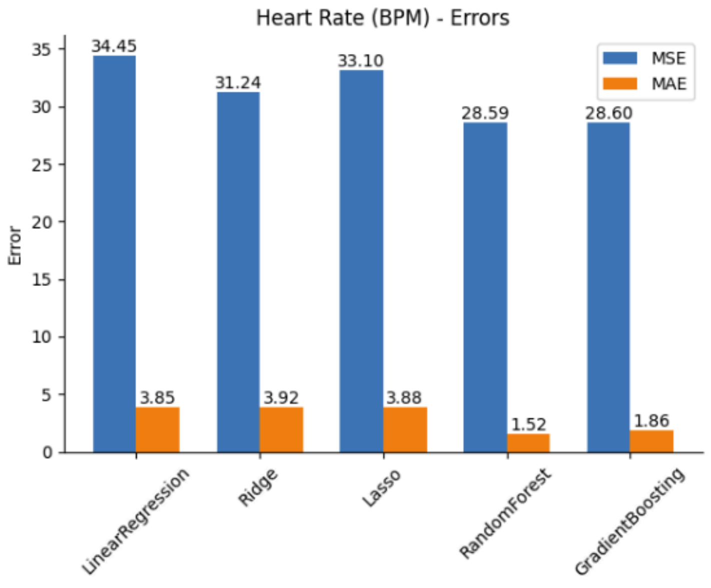
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- R^2 Score





The scatter plot above shows the predicted vs actual pace values for each model. The linear models performed moderately (R^2 ~0.62), while Random Forest ($R^2 = 0.91$) and Gradient Boosting ($R^2 = 0.76$) were significantly better. The bar chart next to it summarizes the errors, where ensemble models produced the lowest MSE and MAE, confirming their suitability for this task.



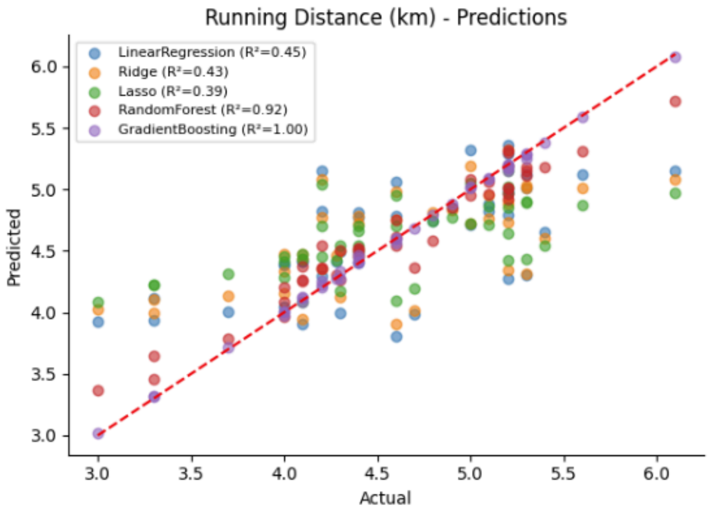


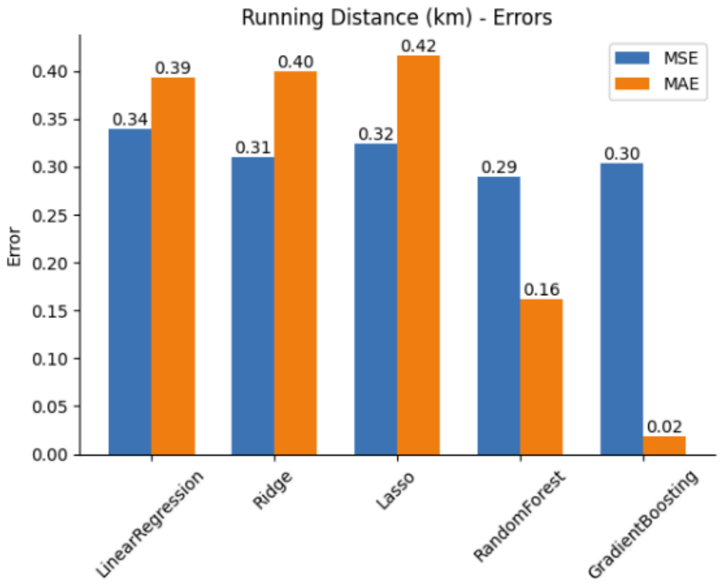
A similar pattern appeared in the heart rate predictions. The ensemble models again outperformed the linear ones, with Random Forest achieving an R² of 0.92 and Gradient Boosting close behind at 0.87. This suggests that heart rate, like pace, is quite responsive to air quality and environmental data when modeled using tree-based methods.

The same procedure on running distance produced a great result, with Gradient Boosting achieving great accuracy, and Random Forest following it closely. This could reflect the structured characteristic of my running sessions, where environmental stressors on bad days made me stop early, a behavior that the models captured very well.

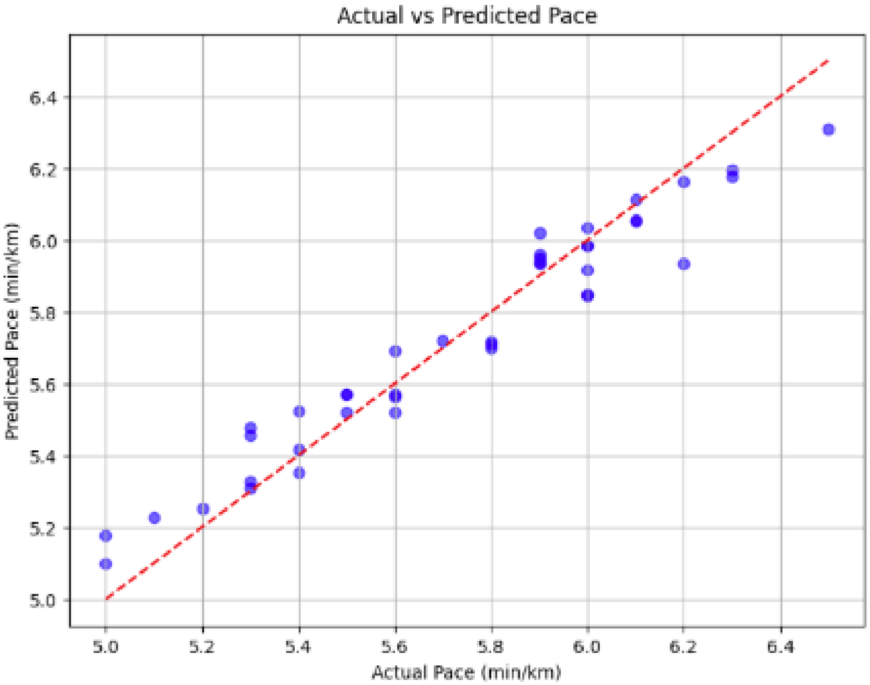
Based on the error metrics and R² scores, the most suitable models for my dataset were Random Forest and Gradient Boosting. To continue, I used one of these models, the Random Forest model, to see the difference between actual vs predicted performances.

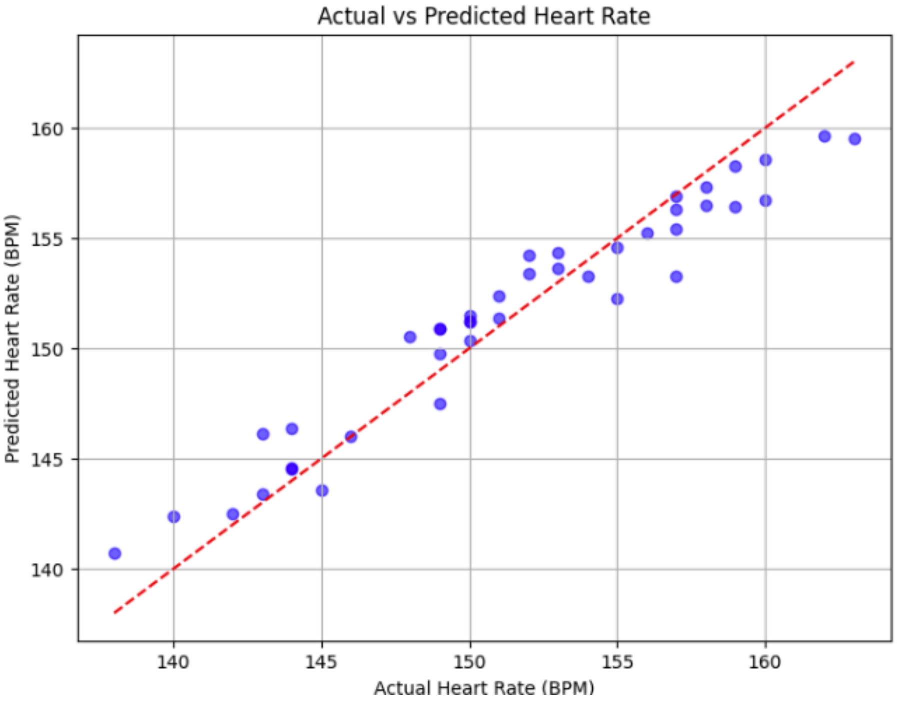
The following final plots helped me see how well the Random Forest model performed in predicting my running pace, heart rate, and distance. In each case, the points are mostly lined up along the red diagonal, which means the predicted values were very close to the actual ones. For pace and heart rate, the model captured the overall trend well, with only a few scattered points falling noticeably off the line. The running distance predictions were especially accurate, and most of them matched the actual distances I ran, particularly between 4 and 5.5 kilometers. Seeing these results visually confirmed what the earlier error metrics showed: using air quality

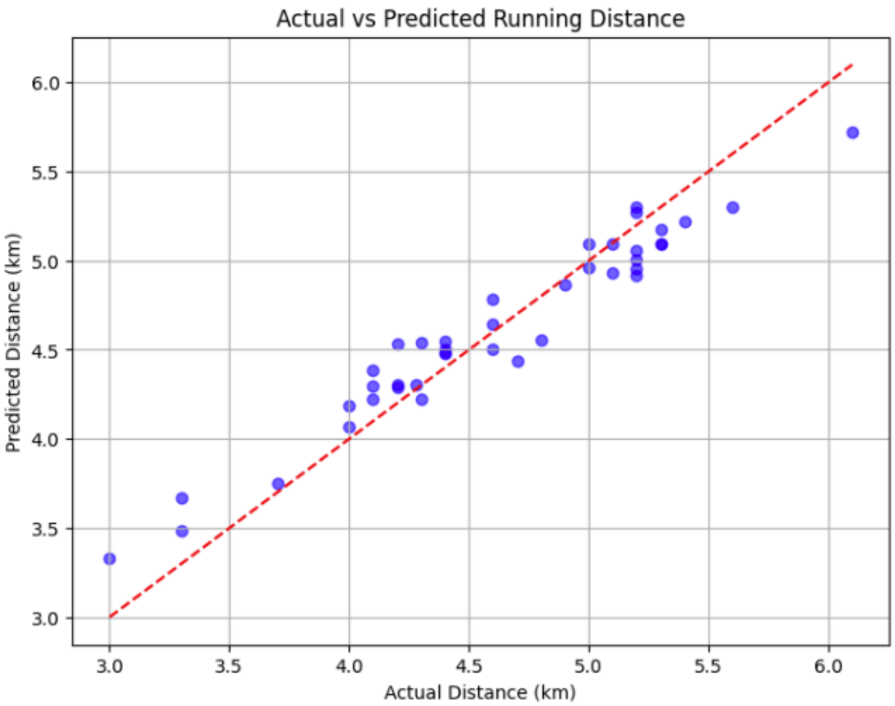




and other environmental features, it’s really possible to make meaningful predictions about how I’ll perform on a run. This whole process took the project from just exploring a pattern to actually building something useful and predictive.







8) CONCLUSION

This project started with a simple question that came from a personal experience: Why do I feel so different on some runs on the Sabancı campus, compared to others, even when nothing about my routine has changed?

Through this project, I was able to turn that curiosity into a data science project, where I collected my own performance data, daily environmental conditions, and applied statistical and machine learning methods to understand what was really happening.

The results were clear. There is a strong relationship between air quality and my running performance. Days with higher AQI, PM2.5, and NO_2 levels were consistently linked to a lower pace, higher heart rate, and sometimes even shorter running distances. These weren't just vague trends; they were statistically significant, and they later became predictable with models like Random Forest and Gradient Boosting. Being able to predict this impact was a meaningful outcome for me. It means that I now have a way to anticipate "bad running days" and not blame myself when my performance dips.