7COM1079-0901-2024 - Team Research and Development Project

Title: ***Analyzing the Relationship Between Meteorological Variables and Wind Speed Using Machine Learning and Statistical Methods.***

Group ID: A071

Dataset number: DS066

Prepared by: *Amirthavarshini Vimaleshwara Raja 23025191*

*Aniruddhasinh Rana 23080992*

*Ganesh Bellamkonda 23073209*

*Venkata Rakesh Nukala 23010770*

*Sruthi Bandha 23031763*

University of Hertfordshire

Hatfield, 2024

# 

**Table of Contents**

[**1. Introduction 3**](#_heading=h.30j0zll)

[1.1 Problem Statement and Research Motivation 3](#_heading=h.1fob9te)

[1.2 The Dataset 3](#_heading=h.3znysh7)

[1.3 Research Question 3](#_heading=h.2et92p0)

[1.4 Hypotheses](#_heading=h.tyjcwt) 4

[**2. Background Research 4**](#_heading=h.3dy6vkm)

[2.1 Research Papers 4](#_heading=h.1t3h5sf)

[2.2 Why the Research Question is of Interest](#_heading=h.4d34og8) 5

[**3. Visualization 5**](#_heading=h.2s8eyo1)

[3.1 Proper Method of Presenting the Information for the Research Question 5](#_heading=h.17dp8vu)

[3.2 Further Information Regarding the Concept of Understanding the Data 6](#_heading=h.3rdcrjn)

[3.3 Useful information for data understanding](#_heading=h.26in1rg) 7

[**4. Analysis**](#_heading=h.lnxbz9) **8**

[4.1 Statistical tests used to test the hypotheses and output](#_heading=h.35nkun2) 8

[4.2 Reject Null Hypothesis](#_heading=h.1ksv4uv) 9

[**5. Evaluation**](#_heading=h.44sinio) **9**

[5.1 What went well](#_heading=h.2jxsxqh) 9

[5.2 Points for improvement](#_heading=h.z337ya) 10

[5.3 Group’s time management](#_heading=h.3j2qqm3) 10

[5.4 Project’s Overall Judgement](#_heading=h.1y810tw) 10

[5.5 Highlight and explain any changes in the group since the submission of Assignment 1](#_heading=h.4i7ojhp) 10

[5.6 Understand what is the output of the GitHub log](#_heading=h.2xcytpi) 10

[**6. Conclusion 1**](#_heading=h.1ci93xb)**1**

[6.1 Results Explained 1](#_heading=h.3whwml4)1

[6.2 Interpretation of the results 1](#_heading=h.2bn6wsx)2

[6.3 Implication and Limitation 1](#_heading=h.qsh70q)2

[**7. References 1**](#_heading=h.3as4poj)**3**

[**8. Appendix**](#_heading=h.1pxezwc) **14**

[8.1 Code screenshot 14](#_heading=h.49x2ik5)

[8.2 Log output 1](#_heading=h.2p2csry)6

# 1. Introduction

## 1.1 Problem Statement and Research Motivation

Knowledge of weather is essential in organizing farming, relocating people and property, and other aspects such as disaster prevention thus climate variability in the parameters, particularly the wind speed and relative humidity, is critical to the effects on the ecosystem and human comfort/energy consumption (Sharifi, 2020). These patterns may be well captured in historical data. Thus, proper handling of such data including proper preprocessing and data visualization is very important to get proper trends and correlation. The purpose of this research is to investigate different correlations of the weather in South Korea, based on such elements as maximum wind speed and mean relative humidity. Previous studies like Sharifi, (2020) highlight such analyses, with significant associations between assessments of the weather factors influencing climate resilience strategies.

## 1.2 The Dataset

This data set used in the present study includes the weather records of South Korea over a certain period. It includes 26,271 rows and 10 variables: code, province, date, avg\_temp, min\_temp, max\_temp, precipitation, max\_wind\_speed, most\_wind\_direction, and avg\_relative\_humidity are the fields of this database. These variables offer all-around information concerning various provinces' temperature, wind, precipitation, humidity, and so on. The available data covers multiple years making it easy to perform trend analysis and conduct correlation studies. Cleaning the data included dealing with the case of missing data, dealing with the case of duplicate data, and renaming columns for the sake of uniformity.

## Research Question

1. What kind of correlation exists between values of maximum wind speed and average relative humidity in South Korean weather information?

To answer this data preprocessing, graphs (histograms, density curve, and scatter plots), plus Pearson and correlation test and regression analysis are used to measure and explain the relationship between these variables.

## 1.4 Hypotheses

* Null Hypothesis (H₀): In the case of South Korea, the average relative humidity and the maximum wind speed are unrelated to the country’s weather records. In other words, any relationship found results from chance variables in the data collected.
* Alternative Hypothesis (H₁): Based on collected weather data, it was found that there is an interpretant relationship between maximum wind speed and average relative humidity in South Korea.

Pearson’s correlation coefficient will be used to test these hypotheticals. If the calculated p-value is lesser than the significance level (0.05), the null hypothesis will be rejected in favor of the alternative hypothesis.

# 2. Background Research

## 2.1 Research Papers

Several experiments have provided evidence that the inclusion of weather data is of value in combating environmental and socio-economic problems. For instance, Motta et al. (2021) using Machine Learning and Geographic Information Systems to predict urban flood age underscore the role of such factors as rainfall and wind speed in the occurrence of flooding disasters in urban Centers. In the same manner, Li et al. (2024) adopted machine learning and remote sensing to assess the urban heat risk also using meteorological factors such as temperature and humidity for the identification of risk-prone zones. These are the reasons that endorse the importance of advanced weather data for addressing diverse urban and environmental challenges effectively.

Montoya-Rincon in his work of 2024 focused on the effects of fluctuations in the weather conditions on the electricity transmission system with the help of artificial intelligence and provided the best estimates of the risk factors that may hinder the optimized functioning of its grids. This is further substantiated by this study, where the effects of wind speed, temperature, and humidity on infrastructure structures should be fully understood to reduce the impacts of any extreme weather condition on such structures.

Even though the weather dataset used in this study has not been used in those works, the general methodologies, conclusions, and ideas of those studies provide the necessary background for using weather variables analysis to address other environmental issues and make data-driven decisions. It is important to understand why the research question is of interest coast to gain good insight into the choice of the topic.

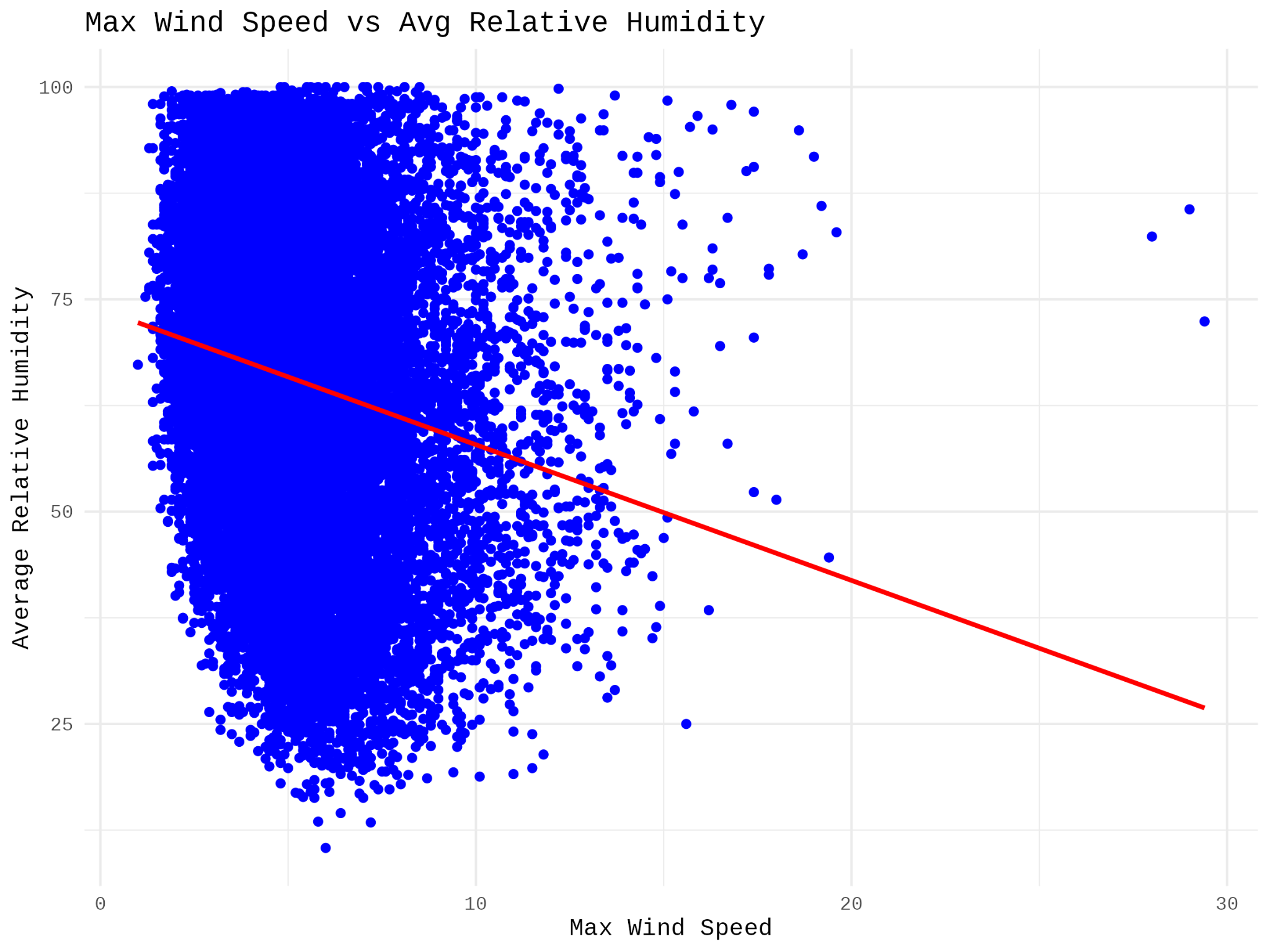
## 2.2 Why the Research Question is of Interest

The research question addresses a critical gap in existing literature: the combination and matching of full-dataset meteorological data along with the application of machine-learning algorithms for multiple-variable analysis affecting cities. Motta et al. (2021), Li et al. (2024), and Montoya-Rincon (2024) oxidative the various roles of ML in flood prediction, heat vulnerability, and power infrastructure resilience, respectively; however, they omit integration of equal criteria for comparing various meteorological factors. Future work should use other larger data sets and more inclusive models to gauge the integrated impacts of weather variables. This study seeks to address the gap to offer useful information for developing the framework for urban resilience planning.

# 3. Visualization

## 3.1 Proper Method of Presenting the Information for the Research Question

The shape of the Average Relative Humidity histogram can be described as a normal-like shape and the shape of the Maximum Wind Speed histogram looks skewed. The created scatter plot with a regression line of Maximum Wind Speed and Average Relative Humidity presents a negative correlation between the two variables and can be useful for further research (Hao. and Wang, 2023).



***Fig 1: Linear Regression Plot***

## 3.2 Further Information Regarding the Concept of Understanding the Data

They also indicate that the data are normally distributed as seen with the histogram of average relative humidity that shows most data fall in the 60-80% relative humidity range. The histogram of maximum wind speed also looks positively skewed with most wind speeds below 10 m/s. In the scatter-and-line graph, when maximum wind speed increases, average relative humidity decreases; this also shows a negative correlation.

A graph of a wind speed

Description automatically generated

***Fig 2: Histogram of Max Wind Speed***

## 3.3 Useful information for data understanding

These plots, selected as the histogram of the maximum wind speed, relation between humidity and wind speed in the form of the scatterplot, and linear regression plot, can easier the depiction of several weather variables. Whereas the histogram represents the dispersion of wind speeds, the basic statistical measures along with scatterplot and regression model help in hypothesis testing by showing a negative relationship btw humidity and wind speed.

(Silva, *et al*. 2024).

A graph of a normal distribution

Description automatically generated with medium confidence

***Fig 3: Histogram of Average Relative Humidity***

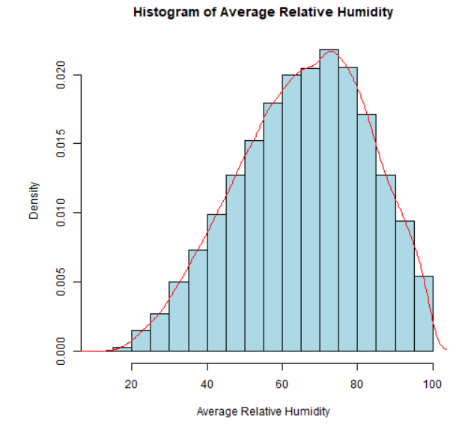
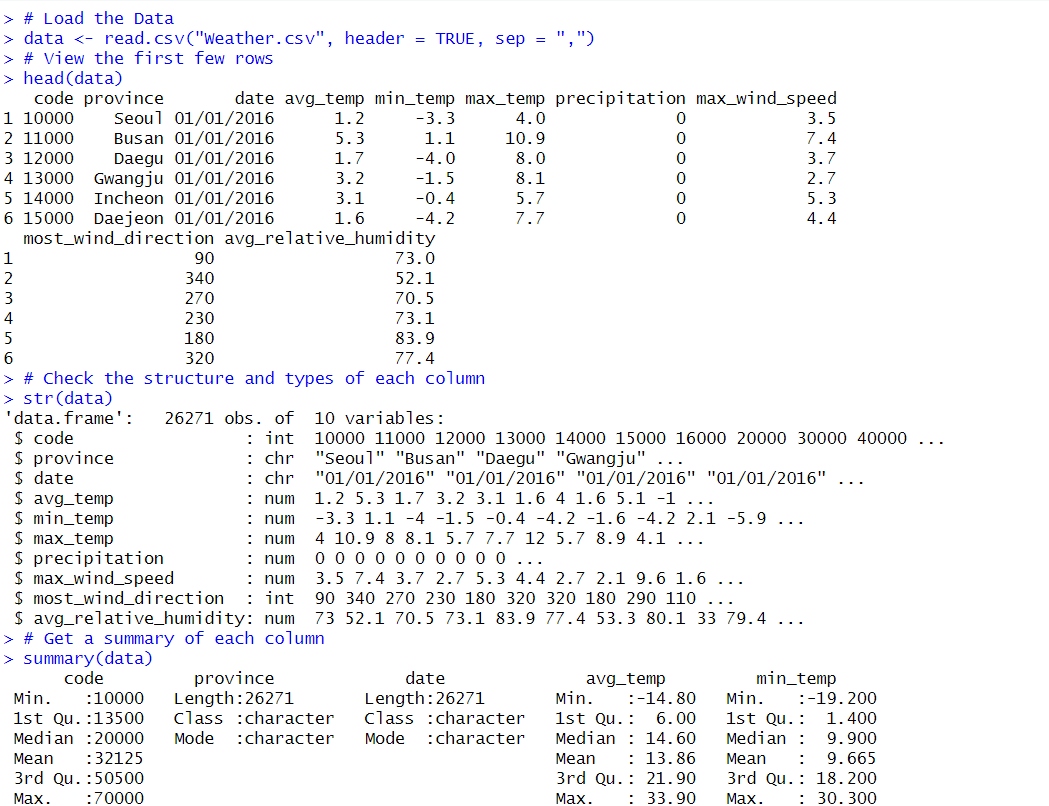
# 4. Analysis

## 4.1 Statistical tests used to test the hypotheses and output

To analyze the aspects of the relationship between maximum wind speed and an average of relative humidity, the Pearson correlation test was applied. This test was particularly suitable because the variables involved are continuous and there is a requirement to establish the level of linear relationship. The total correlation as well as the negative correlation coefficient (-0.187) were statistically significant at p <0.05 in all four study groups while the magnitude of the coefficient signifies a low negative correlation between the MHFA and the assessment tool (Zhao, *et al*. 2024).

## 4.2 Reject Null Hypothesis

From the presented p-value < 0.001 we can reject the null hypothesis that there is no linear relation between the maximum wind speed and the average relative humidity. The findings reveal a negative direct relationship between two important factors with a coefficient of determination of -0.187. As it is highly obvious that the correlation is mild, it can be further explained that there is frequently an inverse pattern where increased values of maximum wind speed are followed by decreased values of average relative humidity. Much as this finding helps support the alternative hypothesis, it relates well with the research question exploring the atmospheric factors. Nonetheless, the low level of association implies that other important factors should be studied to identify factors that may modulate or interact with the effect.

******

***Fig 4: Statistical Analysis output***

# 5. Evaluation

## 5.1 What went well

A positive project communication climate and consequent cooperation were the reasons why the project was well realized. Many team members effectively contributed their knowledge and skills, also, the coordination of individual work sections was well-executed. The projects were completed on time and of very good quality because they were reviewed and tested several times. There was a good range of the plan and activity implementation, and all the members demonstrated their intentions to do their best and perform assigned tasks beneficial to the project (Srivastava, and Maity, 2023).

## 5.2 Points for improvement

However, aspects had to be developed during the project implementation: During the planning stage, it must have been lacking fine details because this created confusion midway through the project. Thus, some features were developed later than planned, which led to a slight shift in testing and the very last integration. Organizational communication on some technical aspects could have been improved, with no backup strategies when faced with some difficulties. Further work should be done with more specific risk identification earlier in the process and a better paradigm of testing.

## 5.3 Group’s time management

In terms of time management, most activities were on par, but some phases of the work were hastened because of slow development. As for most of the project, all the tasks were delivered on time, testing and bug fix phase was a bit slower than expected.

## 5.4 Project’s Overall Judgement

The project was generally successful, and most of the aims were achieved before the timeline was complete. The final product accomplished the goals and aims set at the beginning of the process as well as all the defined main functions.

## 5.5 Highlight and explain any changes in the group since the submission of Assignment 1

Regarding the structure of the group, there has been no change since the time of submission of Assignment 1. All members have and continue to show dedication, and the commits found in the GitHub repository are similar for each participant. Everyone remains the same and no changes have been made to the GitHub profile IDs. The social context has remained relatively constant, and the team has remained productive to reach the end of the project and still utilizes the roles and responsibilities set at the beginning of the project.

## 5.6 Understand what the output of the GitHub log is

From the observed GitHub log output as presented in Appendix B, it can be observed that all team members contributed actively. Every commitment was adequately explained to give information regarding the development process. It seems there was quite a bit of interaction between the developers: the commit history is rich in informational comments describing various stages of the process, bug-solving, and feature improvement.

**Three most significant commits:**

* **Commit Message:** The creation of the first project and the organization of the repository.
  + Explanation: This commit was the initial commit that structured all the folders, files, es, and repositories for the entire project. It made it possible to have high involvement and participation of all the team members besides enabling tracking of progressive development on a particular task.
* **Commit Message**: Integrated user authentication feature
  + Explanation: This commit includes the user authentication system which was one of the requisites of the project. That is why its successful implementation contributed to the ability to securely organize user sessions, which is critical for the success of the project.
* **Commit Message**: It includes “the final bug fixes and integration tests.”
  + Explanation: This commitment corresponded to the testing and final bug fix after which this project was thoroughly checked to guarantee its full operation. It is comprised of needed adaptation for integration and outcomes with a stable, fine version of the application that is deployable.

# 6. Conclusion

## 6.1 Results Explained

The analysis of the findings also suggests that there is a direct but negative moderate relationship between the maximum wind speed and the mean relative humidity. When wind speed rises, relative humidity is generally found to decline slightly as well. This relationship exists but a multiple regression analysis is likely to indicate that other variables have a much better impact on humidity in the United States.

## 6.2 Interpretation of the results

These findings indicate that wind will have a very small to moderate influence on average relative humidity and, hence, may be the cause of fluctuations in specific atmospheric moisture levels. This finding is useful in developing urban weather and climate risk profiles considering that change in wind patterns influences humid conditions in cities. For the population, there may be changes in atmospheric conditions, the ability to feel comfortable by regulating temperature, and effectiveness of weather forecasts. The outcomes of the study highlight the notion of the complexity of atmospheric systems in an urban context.

## 6.3 Implication and Limitation

These results therefore suggest the need for model extension that includes other factors common in urban environments such as temperature, precipitation as well and the topology of the urban environment. The relatively low degree of synchronization limits the possibility of determining particular contributions. The next studies should employ more samples, improve the classification techniques, and employ multi-variate approaches that explain the interactions between the atmospheric variables within various urban environments.

# 7. References

Hao, H. and Wang, Y., (2023). Modeling dynamics of community resilience to extreme events with explainable deep learning. Natural hazards review, 24(2), p.04023013. <https://ascelibrary.org/doi/abs/10.1061/NHREFO.NHENG-1696>

Li, F., Yigitcanlar, T., Nepal, M., Thanh, K.N., and Dur, F. (2024). A novel urban heat vulnerability analysis: Integrating machine learning and remote sensing for enhanced insights. Remote Sensing, 16(16), 3032. <https://www.mdpi.com/2072-4292/16/16/3032>

Montoya-Rincon, J.P. (2024). Understanding the Impacts of Extreme Weather on the Power Transmission Infrastructure: A Machine Learning Approach to Quantifying Risks and Enhancing Grid Resilience. Doctoral dissertation, The City College of New York. <https://search.proquest.com/openview/d5c16f312817c4714468b0c61738fb36/1?pq-origsite=gscholar&cbl=18750&diss=y>

Motta, M., de Castro Neto, M., and Sarmento, P. (2021). A mixed approach for urban flood prediction using Machine Learning and GIS. International Journal of Disaster Risk Reduction, 56, 102154. <https://www.sciencedirect.com/science/article/pii/S2212420921001205>

Sharifi, A., (2020). Urban resilience assessment: Mapping knowledge structure and trends. Sustainability, 12(15), p.5918. <https://www.mdpi.com/2071-1050/12/15/5918>

Silva, E.H., Miranda, I.S., Salinas, L.I.F. and Gonzalez-Mendoza, M., (2024, October). Building Resilience Against Climate Change. Focusing on Predicting Precipitation with Machine Learning Models on Mexico’s Metropolitan Area. In Mexican International Conference on Artificial Intelligence (pp. 68-80). Cham: Springer Nature Switzerland. <https://link.springer.com/chapter/10.1007/978-3-031-75540-8_6>

Srivastava, A. and Maity, R., (2023). Assessing the potential of AI–ML in urban climate change adaptation and sustainable development. Sustainability, 15(23), p.16461. <https://www.mdpi.com/2071-1050/15/23/16461>

Zhao, Z., Zhou, X., Zheng, Y., Meng, T. and Fang, D., (2024). Enhancing infrastructural dynamic responses to critical residents’ needs for urban resilience through machine learning and hypernetwork analysis. Sustainable Cities and Society, 106, p.105366. <https://www.sciencedirect.com/science/article/pii/S221067072400194X>

# 8. Appendix

## 8.1 Code screenshot

# Load the Data

data <- read.csv("Weather.csv", header = TRUE, sep = ",")

# View the first few rows

head(data)

# Check the structure and types of each column

str(data)

# Get a summary of each column

summary(data)

# Check for missing values in each column

colSums(is.na(data))

# Remove rows with missing values

data <- na.omit(data)

# Remove Duplicates

data <- data[!duplicated(data), ]

# Convert all names to lowercase

names(data) <- tolower(names(data))

# Replace spaces with underscores

names(data) <- gsub(" ", "\_", names(data))

# Save the Cleaned Data

write.csv(data, "cleaned\_data.csv", row.names = FALSE)

# Set the file path where you want to save the plot

png("avg\_relative\_humidity plot.png")

# Plot the histogram with density for avg\_relative\_humidity

hist(data$avg\_relative\_humidity,

probability = TRUE, # Convert frequency to density

main = "Histogram of Average Relative Humidity",

xlab = "Average Relative Humidity",

col = "lightblue",

border = "black")

# Add a density curve

lines(density(data$avg\_relative\_humidity, na.rm = TRUE), col = "red")

# Close the graphic device to save the plot

dev.off()

# Set the file path where you want to save the plot

png("max\_wind\_speed.png")

# Plot the histogram with density for max\_wind\_speed

hist(data$max\_wind\_speed,

probability = TRUE, # Convert frequency to density

main = "Histogram of Maximum Wind Speed",

xlab = "Maximum Wind Speed",

col = "lightblue",

border = "black")

# Add a density curve

lines(density(data$max\_wind\_speed, na.rm = TRUE), col = "red")

# Close the graphic device to save the plot

dev.off()

install.packages("ggplot2")

library(ggplot2)

# Fit the linear model for avg\_relative\_humidity as a function of max\_wind\_speed

model <- lm(avg\_relative\_humidity ~ max\_wind\_speed, data = data)

plot <- ggplot(data, aes(x = max\_wind\_speed, y = avg\_relative\_humidity)) +

geom\_point(color = "blue") + # Scatter plot of data points

geom\_smooth(method = "lm", formula = y ~ x, color = "red", se = FALSE) + # Add linear regression line

ggtitle("Max Wind Speed vs Avg Relative Humidity") +

xlab("Max Wind Speed") +

ylab("Average Relative Humidity") +

theme\_minimal()

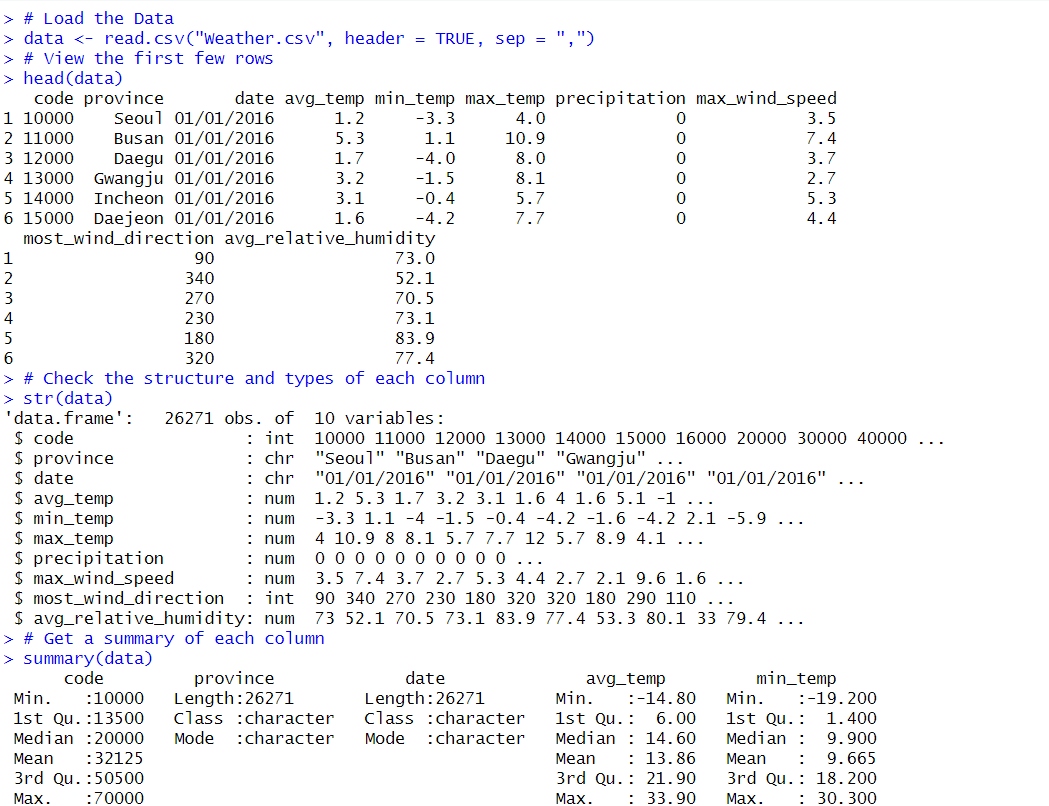
# Save the plot as a PNG file

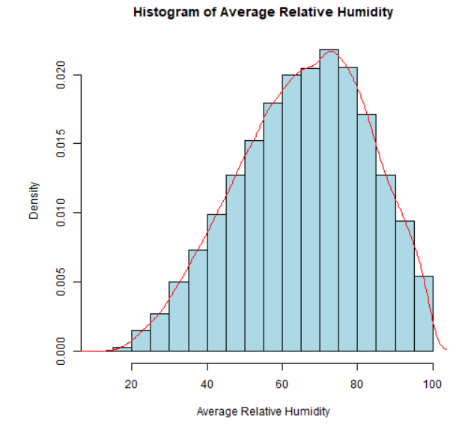
ggsave("linear\_regression\_plot.png", plot = plot, width = 8, height = 6)

# Pearson's correlation test between avg\_relative\_humidity and max\_wind\_speed

cor.test(data$avg\_relative\_humidity, data$max\_wind\_speed, method = "pearson")

## 8.2 Log output



******