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.

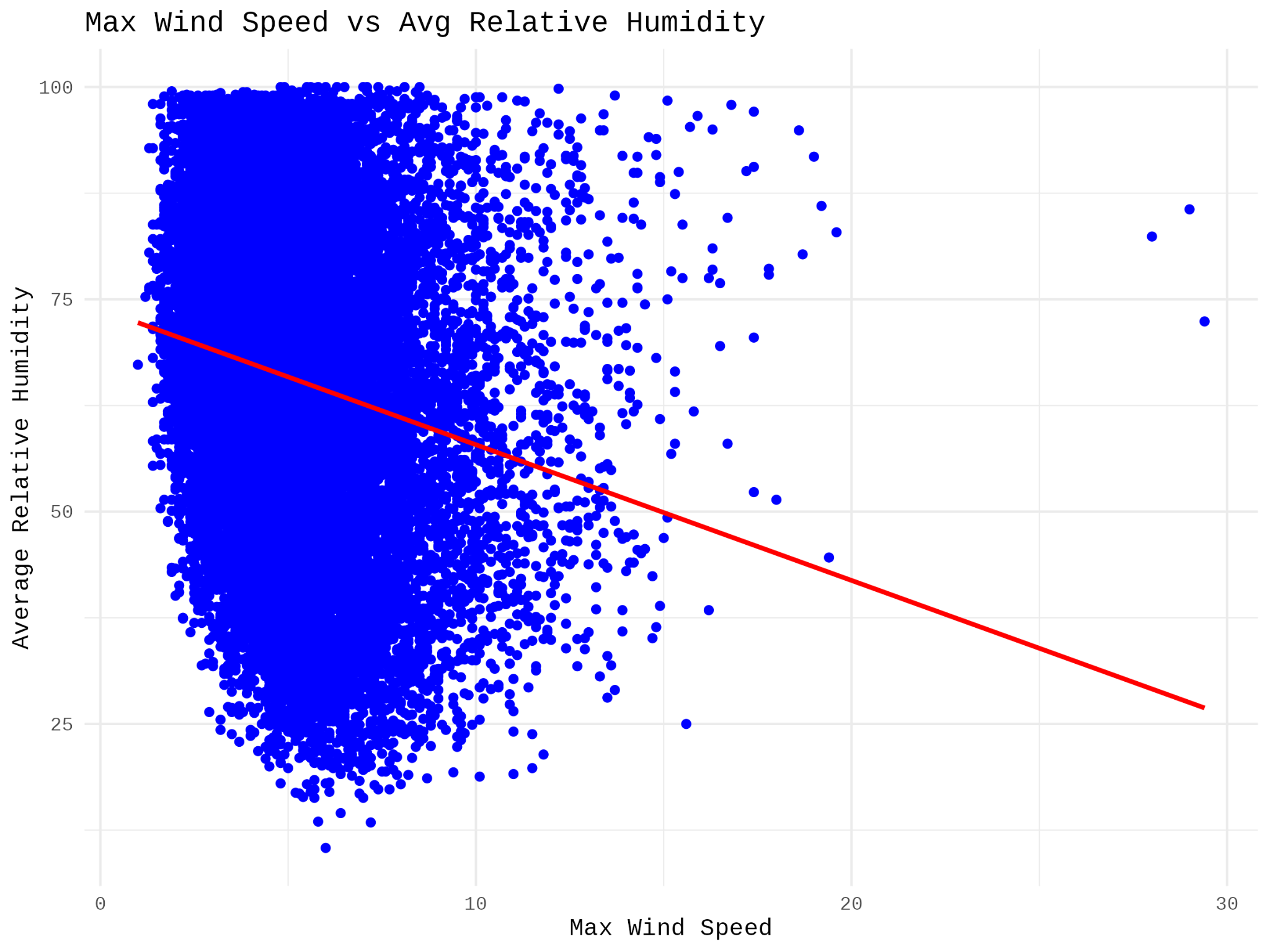
## 2.2 Why the Research Question is of Interest

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.