## 

## 

## 

## 

## 

Multivariate linear regression analysis of air quality among CASTNET sites

EPH 505: Biostatistics I

Yale School of Public Health

Rei Marshall

**Introduction**

The Clean Air Status and Trends Network (CASTNET) is an atmospheric monitoring program established by the 1990 Clean Air Act amendments to assess the effectiveness of emission reduction programs. The network tracks pollutant concentrations and other environmental deposits at 99 sites across the United States and Canada to gauge the current atmospheric health1. Pollutants such as carbon dioxide, nitrous oxide, and methane contribute to global warming through the “greenhouse effect,” when radiation from the sun is trapped and absorbed into the atmosphere rather than reflecting into outer space, raising the temperature of the earth’s atmosphere2. Research and data analysis on meteorological factors that exacerbate (or alleviate) poor air quality across geographical zones is necessary as climate change is proving fatal for millions across the globe3. Poor air quality is linked to respiratory illnesses, cancer, and other chronic diseases, and accurate data on the most effective emission reduction programs is important for progress toward reducing climate change rates and global disease burden. In this code build, we leverage CASTNET datasets, which support the evaluation of primary and secondary National Ambient Air Quality Standards, to identify meteorological risk factors for air quality across CASTNET sites and compare mean ozone and key trace gas levels among different study locations.

## **Methods**

Negative values in the air quality data, often resulting from sensor noise, calibration issues, or measurements near the detection limit, were carefully handled to maintain data integrity. We replaced negative concentrations with zero, a common approach in air quality analysis, assuming the pollutant was not present in those instances.4 Additionally, we managed outliers in trace gas levels and temperature data to reduce distortion in our analysis.

We analyzed correlations between ozone and each meteorological factor, extending this analysis to trace gases with sufficient data—carbon monoxide (CO), nitrogen oxides (NO), sulfur dioxide (SO₂), and reactive nitrogen oxides (NOy). These gases were chosen based on data availability; others lacked sufficient data for linear regression analysis. Following, we conducted a multivariate linear regression analysis using selected meteorological factors as independent variables to evaluate their combined and individual effects on each gas level. A multicollinearity assessment was done in order to select the best meteorological predictors on each model.

In our analysis of mean gas levels by zone (SITE\_ID), we examined the distribution of each gas to ensure robust comparisons. We tested for normality in gas level distributions, a key assumption for accurately interpreting mean differences across zones, and assessed homoscedasticity. Based on these results, we selected either parametric tests for normally distributed data or nonparametric tests when normality was not met.

**Results**

1. **Meteorological factors and air quality**

**Table 1** presents the adjusted correlation coefficients (aCoeff) between levels of ozone and trace gases (CO, NO, NOY, SO₂) and various meteorological factors, indicating statistically significant relationships (p<0.005) in several cases:

**Ozone**: Ozone levels are positively correlated with temperature (0.525) and solar radiation (0.0079), while showing a significant negative correlation with relative humidity (-0.357) and wetness (-0.3148).

**Carbon monoxide (CO)**: CO levels exhibit significant positive correlations with temperature (1.2190) and precipitation (1.3826), and a strong negative correlation with wind speed (-1.8835) and wind direction (-0.0356).

**Nitrogen Oxide (NO)**: NO levels are significantly affected by temperature (-9.06e-3), solar radiation (5.10e-4), precipitation (-1.53e-2), windspeed (-2.53e-2), and wind direction (-2.95e-4), indicating that these meteorological factors play a substantial role in its distribution.

**Reactive oxides of Nitrogen (NOy)**: NOY levels have a weak negative relationship with temperature (-0.0198), solar radiation (-0.0019), precipitation (-0.2215), windspeed (-0.3956), wind direction (-0.0089), and wetness (-0.5581), suggesting that these factors generally reduce NOY concentrations.

**Sulfur dioxide (SO₂)**: SO₂ levels show a slight positive correlation with temperature (0.0152) and wind speed (0.0736), but a significant negative correlation with relative humidity (-0.0087) and wetness (0.3321), implying that humid conditions and high wetness levels decrease SO₂ concentration.

These adjusted correlations highlight the influence of temperature, humidity, windspeed, and other meteorological factors on the levels of ozone and trace gases, with both positive and negative associations depending on the gas and factor involved.

1. **Geographical zones and air quality**

**Table 2** shows the average levels of ozone and trace gases (CO, NO, NOY, SO₂) in parts per billion (ppb) across various study sites (SITE\_ID). Ozone levels vary significantly by location, with Pinedale, WY (46.25 ppb) and Rocky Mountain National Park, CO (46.16 ppb) showing the highest averages, while Bondville, IL (32.03 ppb) has the lowest. CO levels, where available, also differ markedly, with Great Smoky Mountains, TN (165.36 ppb) showing the highest concentration. NO levels are generally low across sites, but Canyonlands National Park, UT (2.01 ppb) and Great Smoky Mountains, TN (2.16 ppb) report higher values compared to others. Significant variations are observed for NOY and SO₂ levels, with notable concentrations in Great Smoky Mountains, TN and Mackville, KY. All p-values are extremely low (<2e-16), indicating statistically significant differences in mean gas levels across sites. The F-statistics further confirm these differences across study sites for each gas.

**Discussion**

Ozone emissions have disastrous effects on human health and livelihood, as indicated by the results of the CASTNET monitoring. For example, a decrease in precipitation, associated with increased ozone in the atmosphere, will harm crop production. Aside from the obvious famine and job loss associated with a decreased amount of viable food crops cultivated by humans, low precipitation impacts plants that serve as food sources for many insects and other small animals, therefore impacting the entire food chain. Mirzabaev et al. described the converging impact of climate change related to drought and precipitation on low crop quantity, and how these factors converge to cause problems with hunger and adequate nutrition 5. (**Figure 1**)

Humidity is also related to human adaptability to climate change, as humidity in the atmosphere increases the severity of water-related natural disasters such as hurricanes and decreases human heat tolerance. Though one would expect that the environmental changes would lead to an increased amount of humidity, the data suggests that precipitation actually has a statistically significant negative correlation with increased ozone in the atmosphere. The data is in agreement with the National Center for Atmospheric Research’s recent findings related to moisture in dry regions of the United States; though atmospheric moisture was expected to increase based on the thermodynamic climate change models, the United States is experiencing a decrease in atmospheric moisture. The dryness may accelerate drought and encourage wildfire, and future research about the relationship between atmospheric pollutants and humidity is needed to understand this occurrence6.

## **Conclusion**

Current literature and analysis highlight the critical impact of air quality on both public health and climate change, emphasizing the harmful effects of air pollutants. In this study, we present statistical correlations between meteorological factors, ozone concentrations, and trace gas levels, along with average pollutant levels across seven study sites. Our findings reveal a strong relationship between ozone concentrations and temperature, suggesting an underlying mechanism contributing to climate change. Understanding this relationship is essential for developing effective solutions and serves as a predictive tool to estimate the future impacts of climate change if mitigation efforts are delayed or neglected.

## **References**

1. *Clean Air Status and Trends Network (CASTNET)*, www3.epa.gov/castnet/docs/CASTNET-Factsheet-2021.pdf. Accessed 2 Nov. 2024.
2. “Overview of Greenhouse Gases | US EPA.” *Environmental Protection Agency (EPA)*, 11 April 2024, https://www.epa.gov/ghgemissions/overview-greenhouse-gases. Accessed 3 November 2024.
3. “Air Pollution and Your Health.” *National Institute of Environmental Health Sciences*, U.S. Department of Health and Human Services, www.niehs.nih.gov/health/topics/agents/air-pollution. Accessed 2 Nov. 2024.
4. Jiang N, Akter R, Ross G, White S, Kirkwood J, Gunashanhar G, Thompson S, Riley M, Azzi M. On thresholds for controlling negative particle (PM2.5) readings in air quality reporting. Environ Monit Assess. 2023 Sep 12;195(10):1187. doi: 10.1007/s10661-023-11750-4. PMID: 37698727; PMCID: PMC10497433.
5. Alisher Mirzabaev, Rachel Bezner Kerr, Toshihiro Hasegawa, Prajal Pradhan, Anita Wreford, Maria Cristina Tirado von der Pahlen, Helen Gurney-Smith, Severe climate change risks to food security and nutrition, Climate Risk Management, Volume 39, 2023, 100473, ISSN 2212-0963, <https://doi.org/10.1016/j.crm.2022.100473>.
6. Hosansky, David. “Climate change isn't producing expected increase in atmospheric moisture over dry regions | NCAR & UCAR News.” *News*, 17 January 2024, https://news.ucar.edu/132936/climate-change-isnt-producing-expected-increase-atmospheric-moisture-over-dry-regions. Accessed 3 November 2024.

**Appendix**

**Table 1. Correlations between meteorological factors ad ozone/trace level gases**

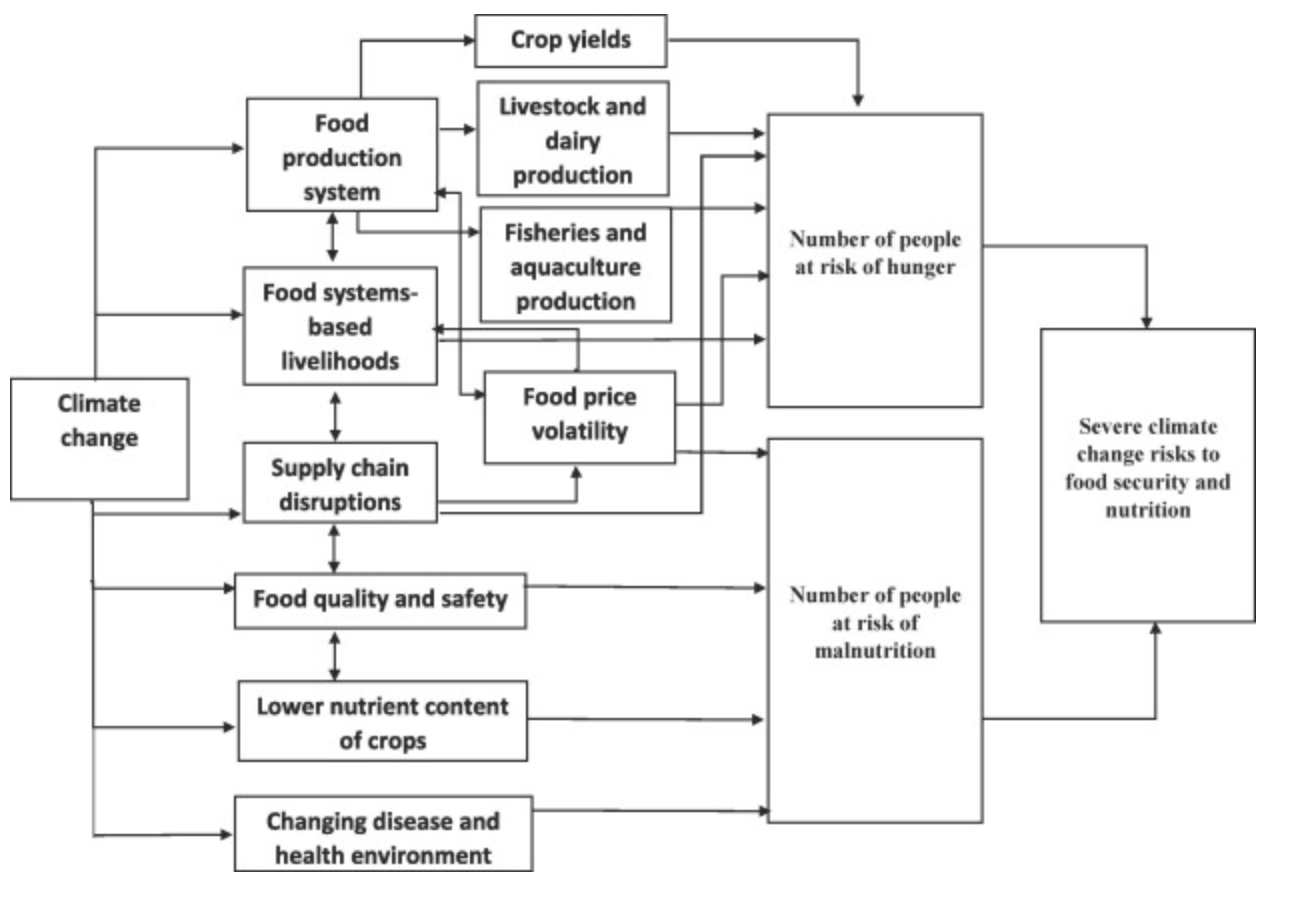
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Ozone** | | **CO** | | **NO** | |
|  | **uCoeff** | **aCoeff** | **uCoeff** | **aCoeff** | **uCoeff** | **aCoeff** |
| **Temperature** | 0.5978 | **0.525\*** | 0.3000 | **1.2190\*** | -0.0180 | **-9.06e-3 \*** |
| **Relative Humidity** | -0.7279 | **-0.357\*** | -0.0197 | **0.2001\*** | -0.0216 | **1.01e-4** |
| **Solar Radiation** | 0.5747 | **0.0079\*** | 0.0611 | -0.0014 | 0.0852 | **5.10e-4 \*** |
| **Precipitation** | 0.0158 | **1.3826\*** | -0.0390 | **-3.2600\*** | -0.0180 | -1.53e-2 |
| **Windspeed** | 0.0921 | **0.8144\*** | -0.2087 | **-1.8835\*** | -0.0508 | **-2.53e-2 \*** |
| **Wind Direction** | -0.0819 | **-0.0015\*** | -0.1578 | **-0.0356\*** | -0.0233 | **-2.95e-4 \*** |
| **Wetness** | -0.2726 | **-0.3148\*** | 0.0324 | 0.1838 | 0.0069 | **8.78e-2 \*** |
| uCoeff= Unadjusted correlation coefficients; aCoeff = Adjusted correlation coefficients in a regression model. \*Statistically significant (p<0.005) | | | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NOy** | | **SO2** | |
|  | **uCoeff** | **aCoeff** | **uCoeff** | **aCoeff** |
| **Temperature** | -0.0020 | **-0.0198\*** | 0.0457 | **0.0152\*** |
| **Relative Humidity** | 0.0220 | -0.0025 | -0.0480 | **-0.0087\*** |
| **Solar Radiation** | -0.0783 | **-0.0019\*** | 0.0268 | -0.0001 |
| **Precipitation** | -0.0267 | **-0.2215\*** | -0.00006 | -0.0292 |
| **Windspeed** | -0.1595 | **-0.3956\*** | 0.0403 | **0.0736\*** |
| **Wind Direction** | -0.1450 | **-0.0089\*** | -0.0156 | -0.0005 |
| **Wetness** | -0.0132 | **-0.5581\*** | 0.0079 | **0.3321\*** |
| uCoeff= Unadjusted correlation coefficients; aCoeff = Adjusted correlation coefficients in a regression model. \*Statistically significant (p<0.005) | | | | |

**Table2.** Average levels of Ozone and trace gases in **ppb** by study sites

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SITE\_ID** | **Ozone** | **CO** | **NO** | **NOY** | **SO2** |
| **Bondville, IL (BVL 130)** | 32.03 | 100.86 | 0.28 | 4.35 | 0.68 |
| **Canyonlands National Park, UT (CHC432)** | 41.32 | - | 2.01 | - | - |
| **Duke Forest, NC (DUK008)** | 36.01 | - | 0.28 | 2.47 | - |
| **Great Smoky Mountains, TN (GRS420)** | 41.20 | 165.36 | 2.16 | 3.76 | 2.86 |
| **Mackville, KY (MAC426)** | 34.07 | 157.64 | 1.65 | 4.27 | 2.26 |
| **Pinedale, WY (PND165)** | 46.25 | - | 0.24 | 0.83 | - |
| **Rocky Mountain National Park, CO (ROM206)** | 46.16 | - | 0.09 | 1.08 | - |
| **P-value** | <2e-16 | <2e-16 | <2e-16 | <2e-16 | <2e-16 |
| **A (F-statistic)** | 10123 | 5799 | 114.17 | 732.74 | 83.974 |
| **num df** | 6 | 2 | 6 | 5 | 2 |

**Figure 1: Impact pathways from** [**climate change**](https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/climate-change) **to hunger and malnutrition. (Alisher Mirzabaev et all. 2023.)**



# CODE

# close all, clear all

rm(list=ls())

graphics.off()

# set working directory

getwd()

setwd("/Users/reisplace/Desktop/YaleBiostats")

# Install required packages

#install.packages("readxl") # for reading Excel files

#install.packages("dplyr") # for data manipulation

#install.packages("doBy")

#install.packages("corrplot")

#install.packages("ggcorrplot")

#install.packages("car")

#install.packages("psych")

# Load the necessary libraries

library(readr)

library(readxl)

library(dplyr)

library(ggplot2)

library(tidyr)

library(corrplot)

library(doBy)

library(ggcorrplot)

library(reshape2)

library(car)

library(psych)

# 0. reading datasets from file

ozone\_data = read.csv("ozone\_2023.csv")

met\_data = read.csv("metdata\_2023.csv")

gas\_data = read.csv("hourly\_gas\_2023.csv")

sites\_data = read\_xlsx("activesuspendedcastnetsites.xlsx")

# Merging datasets by common keys

merged\_data = met\_data %>%

inner\_join(ozone\_data, by = c("SITE\_ID", "DATE\_TIME"))

merged\_data = merged\_data %>%

inner\_join(gas\_data, by = c("SITE\_ID", "DATE\_TIME"))

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#1. Are there meteorological factors that exacerbate (or ameliorate) air quality?

##1.OZONE LEVELS VS METEOROLOGICAL DATA

###Negative values in air quality data are often artifacts from sensor noise, calibration issues, or low-concentration measurements near the detection limit. Handling these values requires careful consideration to maintain data integrity.

###By excluding negative gas values, we ensure that our analysis focuses only on physically meaningful measurements, leading to more accurate insights into the relationship between meteorological factors and air quality.

####Set to Zero: Since negative concentrations are not physically possible, you can replace negative values with zero, which assumes that the pollutant is not present in those cases. This method is simple and commonly used in air quality analysis.

merged\_data = merged\_data %>%

mutate(VALUE = ifelse(VALUE < 0, 0, VALUE))

####removing outliers:

# Calculate Q1, Q3, and IQR for the TEMPERATURE column

Q1\_temp <- quantile(merged\_data$TEMPERATURE, 0.25, na.rm = TRUE)

Q3\_temp <- quantile(merged\_data$TEMPERATURE, 0.75, na.rm = TRUE)

IQR\_temp <- Q3\_temp - Q1\_temp

# Define the lower and upper bounds based on IQR

lower\_bound\_temp <- Q1\_temp - 1.5 \* IQR\_temp

upper\_bound\_temp <- Q3\_temp + 1.5 \* IQR\_temp

# Filter the data based on these bounds for TEMPERATURE

merged\_data <- merged\_data %>%

filter(TEMPERATURE >= lower\_bound\_temp & TEMPERATURE <= upper\_bound\_temp)

###Selecting relevant columns: ozone and meteorological factors

air\_quality\_vars = "OZONE.y"

met\_vars = c("TEMPERATURE", "RELATIVE\_HUMIDITY", "SOLAR\_RADIATION", "PRECIPITATION", "WINDSPEED", "WIND\_DIRECTION", "WETNESS")

##excluded sigma-theta because it also includes information on wind direction --> source: https://www3.epa.gov/castnet/docs/CASTNET\_Factsheet\_2019.pdf

### Ensuring that all columns are numeric

merged\_data = merged\_data %>%

mutate(across(all\_of(c(air\_quality\_vars, met\_vars)), as.numeric))

# Reshaping the data to long format for faceting

plot\_data = merged\_data %>%

select(OZONE.y, all\_of(met\_vars)) %>%

pivot\_longer(cols = all\_of(met\_vars), names\_to = "met\_var", values\_to = "Value")

# Plotting with facets

ggplot(plot\_data, aes(x = Value, y = OZONE.y)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE, color = "blue") +

labs(title = "Ozone Levels vs Meteorological Variables",

x = "Meteorological Variable Value", y = "Ozone Level(ppb)") +

facet\_wrap(~ met\_var, scales = "free\_x") +

theme\_minimal()

# Calculating correlation matrix

correlations = merged\_data %>%

select(all\_of(c(air\_quality\_vars, met\_vars))) %>%

cor(use = "complete.obs")

print(correlations)

# Visualizing correlation matrix

ggcorrplot(correlations, lab = TRUE)

# Linear regression model for ozone with meteorological factors

ozone\_model = lm(OZONE.y ~ TEMPERATURE + RELATIVE\_HUMIDITY + SOLAR\_RADIATION + PRECIPITATION + WINDSPEED + WIND\_DIRECTION + WETNESS, data = merged\_data)

summary(ozone\_model)

##assessing multicolinearity

vif\_values <- vif(ozone\_model)

print(vif\_values)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#Assessing correlations between meteorological parameters and types of trace-level gases

# Defining the meteorological variables

met\_vars = c("TEMPERATURE", "RELATIVE\_HUMIDITY", "SOLAR\_RADIATION", "PRECIPITATION", "WINDSPEED", "WIND\_DIRECTION", "WETNESS")

##Plotting the correlations between meteorological variables and trace-level gases

### Filter data for CO

CO\_data = subset(merged\_data, PARAMETER == "CO")

NO\_data = subset(merged\_data, PARAMETER == "NO")

NOY\_data = subset(merged\_data, PARAMETER == "NOY")

SO2\_GA\_data = subset(merged\_data, PARAMETER == "SO2\_GA")

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

####removing outliers:

# Filter outliers for CO data

CO\_data <- subset(merged\_data, PARAMETER == "CO")

describe(CO\_data$VALUE)

summary(CO\_data$VALUE)

IQR\_CO = IQR(CO\_data$VALUE, na.rm = TRUE)

upper\_bound\_CO <- 177 + 1.5 \* IQR\_CO

# Filter the data based on these bounds for TEMPERATURE

CO\_data <- CO\_data %>%

filter(VALUE <= upper\_bound\_CO)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Filter outliers for NO data

NO\_data <- subset(merged\_data, PARAMETER == "NO")

NO\_data <- NO\_data %>%

filter(VALUE <= 900)

describe(NO\_data$VALUE)

summary(NO\_data$VALUE)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Filter outliers for NOY data

NOY\_data <- subset(merged\_data, PARAMETER == "NOY")

describe(NOY\_data$VALUE)

summary(NOY\_data$VALUE)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Filter outliers for SO2\_GA data

SO2\_GA\_data <- subset(merged\_data, PARAMETER == "SO2\_GA")

describe(SO2\_GA\_data$VALUE)

summary(SO2\_GA\_data$VALUE)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### Reshaping the data to long format for easier plotting

CO\_long = melt(CO\_data, id.vars = "VALUE", measure.vars = met\_vars,

variable.name = "Meteorological\_Variable", value.name = "Value")

NO\_long <- melt(NO\_data, id.vars = "VALUE", measure.vars = met\_vars,

variable.name = "Meteorological\_Variable", value.name = "Value")

NOY\_long <- melt(NOY\_data, id.vars = "VALUE", measure.vars = met\_vars,

variable.name = "Meteorological\_Variable", value.name = "Value")

SO2\_GA\_long <- melt(SO2\_GA\_data, id.vars = "VALUE", measure.vars = met\_vars,

variable.name = "Meteorological\_Variable", value.name = "Value")

### Generate the combined plot with facets

plot\_pollutant <- function(data\_long, title) {

ggplot(data\_long, aes(x = Value, y = VALUE)) +

geom\_point(alpha = 0.6) +

geom\_smooth(method = "lm", se = FALSE, color = "blue") +

facet\_wrap(~ Meteorological\_Variable, scales = "free\_x") +

labs(title = title,

x = "Meteorological Variable Value", y = "Concentration") +

theme\_minimal()

}

### Generating plots for each pollutant

plot\_pollutant(CO\_long, "CO Concentration (ppb) vs Meteorological Variables")

plot\_pollutant(NO\_long, "NO Concentration (ppb) vs Meteorological Variables")

plot\_pollutant(NOY\_long, "NOy Concentration (ppb) vs Meteorological Variables")

plot\_pollutant(SO2\_GA\_long, "SO2 Concentration (ppb) vs Meteorological Variables")

##Calculating the correlation values for each gas pollutant

### List of gases to analyze

CO\_correlation\_data = CO\_data[, c("VALUE", met\_vars)]

# Calculate correlation matrix

correlation\_matrix <- cor(CO\_correlation\_data, use = "complete.obs")

# Display the correlation of VALUE with each meteorological variable

correlation\_matrix["VALUE", ]

NO\_correlation\_data = NO\_data[, c("VALUE", met\_vars)]

# Calculate correlation matrix

correlation\_matrix2 <- cor(NO\_correlation\_data, use = "complete.obs")

# Display the correlation of VALUE with each meteorological variable

correlation\_matrix2["VALUE", ]

NOY\_correlation\_data = NOY\_data[, c("VALUE", met\_vars)]

# Calculate correlation matrix

correlation\_matrix3 <- cor(NOY\_correlation\_data, use = "complete.obs")

# Display the correlation of VALUE with each meteorological variable

correlation\_matrix3["VALUE", ]

SO2\_correlation\_data = SO2\_GA\_data[, c("VALUE", met\_vars)]

# Calculate correlation matrix

correlation\_matrix4 <- cor(SO2\_correlation\_data, use = "complete.obs")

# Display the correlation of VALUE with each meteorological variable

correlation\_matrix4["VALUE", ]

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Defining the formula for the regression model

formula = VALUE ~ TEMPERATURE + RELATIVE\_HUMIDITY + SOLAR\_RADIATION + PRECIPITATION + WINDSPEED + WIND\_DIRECTION + WETNESS

# Fit models for each gas type by filtering the data for each parameter

CO\_model = lm(formula, data = CO\_data)

##assessing multicolinearity

vif\_values2 <- vif(CO\_model)

print(vif\_values2)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

NO\_model = lm(formula, data = NO\_data)

##assessing multicolinearity

vif\_values3 <- vif(NO\_model)

print(vif\_values3)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

NOY\_model = lm(formula, data = NOY\_data)

##assessing multicolinearity

vif\_values4 <- vif(NOY\_model)

print(vif\_values4)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

SO2\_GA\_model = lm(formula, data = SO2\_GA\_data)

##assessing multicolinearity

vif\_values5 <- vif(SO2\_GA\_model)

print(vif\_values5)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Display summaries of each model

summary(CO\_model)

summary(NO\_model)

summary(NOY\_model)

summary(SO2\_GA\_model)

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# close all, clear all

rm(list=ls())

graphics.off()

# 0. reading datasets from file

ozone\_data = read.csv("ozone\_2023.csv")

met\_data = read.csv("metdata\_2023.csv")

gas\_data = read.csv("hourly\_gas\_2023.csv")

sites\_data = read\_xlsx("activesuspendedcastnetsites.xlsx")

# Merging datasets by common keys

merged\_data = met\_data %>%

inner\_join(gas\_data, by = c("SITE\_ID", "DATE\_TIME"))

#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

#2. Are there geographical zones that tend to have better or worse air quality?

\_\_\_\_

#2.A

# Calculating average of ozone levels for each site

ozone = merged\_data %>%

group\_by(SITE\_ID) %>%

summarize(avg\_ozone = mean(OZONE.y, na.rm = TRUE))

# Histogram of OZONE values

ggplot(merged\_data, aes(x = OZONE.y)) +

geom\_histogram(bins = 50, color = "black", fill = "blue") +

labs(title = "Histogram of Ozone Values", x = "Ozone Value", y = "Frequency")

##data is normally distributed by visual methods

##Assessing normality for ozone levels by SITE\_ID

ggplot(merged\_data, aes(sample = OZONE.y)) +

facet\_wrap(~SITE\_ID) +

stat\_qq() +

stat\_qq\_line(color = "blue") +

labs(title = "Q-Q Plots of OZONE.y by SITE\_ID")

# Histogram for each SITE\_ID

ggplot(merged\_data, aes(x = OZONE.y)) +

facet\_wrap(~SITE\_ID) +

geom\_histogram(bins = 30, color = "black", fill = "gray") +

labs(title = "Histograms of OZONE.y by SITE\_ID", x = "OZONE.y", y = "Frequency")

# Perform Bartlett test for homoscedasticity

bartlett\_test\_result1 <- bartlett.test(OZONE.y ~ SITE\_ID, data = merged\_data)

# Display the result

print(bartlett\_test\_result1)

levene\_test\_result <- leveneTest(OZONE.y ~ SITE\_ID, data = merged\_data)

print(levene\_test\_result)

##output: p<0.05 --> heterocedasticity

# Perform Welch's ANOVA

welch\_anova\_result <- oneway.test(OZONE.y ~ SITE\_ID, data = merged\_data, var.equal = FALSE)

print(welch\_anova\_result)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Calculating daily average of CO levels for each site

CO = CO\_data %>%

group\_by(SITE\_ID) %>%

summarize(average\_CO = mean(VALUE, na.rm = TRUE))

# Histogram of CO values

ggplot(CO\_data, aes(x = VALUE)) +

geom\_histogram(bins = 50, color = "black", fill = "blue") +

labs(title = "Histogram of CO Values", x = "CO Value", y = "Frequency")

###data seems to be normally distributed by visual methods

##Assessing normality for CO levels by SITE\_ID

ggplot(CO\_data, aes(sample = VALUE)) +

facet\_wrap(~SITE\_ID) +

stat\_qq() +

stat\_qq\_line(color = "blue") +

labs(title = "Q-Q Plots of CO by SITE\_ID")

# Histogram for each SITE\_ID

ggplot(CO\_data, aes(x = VALUE)) +

facet\_wrap(~SITE\_ID) +

geom\_histogram(bins = 30, color = "black", fill = "gray") +

labs(title = "Histograms of CO by SITE\_ID", x = "CO", y = "Frequency")

# Perform Bartlett test for homoscedasticity

bartlett\_test\_result2 <- bartlett.test(VALUE ~ SITE\_ID, data = CO\_data)

# Display the result

print(bartlett\_test\_result2)

levene\_test\_result2 <- leveneTest(VALUE ~ SITE\_ID, data = CO\_data)

print(levene\_test\_result2)

##output: p<0.05 --> heterocedasticity

# Perform Welch's ANOVA

welch\_anova\_result2 <- oneway.test(VALUE ~ SITE\_ID, data = CO\_data, var.equal = FALSE)

print(welch\_anova\_result2)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Calculating daily average of NO levels for each site

NO = NO\_data %>%

group\_by(SITE\_ID) %>%

summarize(average\_NO = mean(VALUE, na.rm = TRUE))

# Histogram of NO values

ggplot(NO\_data, aes(x = VALUE)) +

geom\_histogram(bins = 50, color = "black", fill = "blue") +

labs(title = "Histogram of NO Values", x = "NO Value", y = "Frequency") +

xlim (0,1)

###data does not seem to be normally distributed by visual methods

# Histogram of NO values

ggplot(NO\_data, aes(x = VALUE)) +

geom\_histogram(bins = 50, color = "black", fill = "blue") +

labs(title = "Histogram of NO Values", x = "NO Value", y = "Frequency") +

xlim (0,1) +

ylim(0, 5500)

###data does not seem to be normally distributed by visual methods

##Assessing normality for NO levels by SITE\_ID

ggplot(NO\_data, aes(sample = VALUE)) +

facet\_wrap(~SITE\_ID) +

stat\_qq() +

stat\_qq\_line(color = "blue") +

labs(title = "Q-Q Plots of NO by SITE\_ID") +

xlim (0,1) +

ylim(0, 2)

# Histogram for each SITE\_ID

ggplot(NO\_data, aes(x = VALUE)) +

facet\_wrap(~SITE\_ID) +

geom\_histogram(bins = 30, color = "black", fill = "gray") +

labs(title = "Histograms of NO by SITE\_ID", x = "NO", y = "Frequency") +

xlim (0,1) +

ylim (0,10)

# Perform Bartlett test for homoscedasticity

bartlett\_test\_result3 <- bartlett.test(VALUE ~ SITE\_ID, data = NO\_data)

# Display the result

print(bartlett\_test\_result3)

levene\_test\_result3 <- leveneTest(VALUE ~ SITE\_ID, data = NO\_data)

print(levene\_test\_result3)

##output: p<0.05 --> heterocedasticity

# Perform Welch's ANOVA

welch\_anova\_result3 <- oneway.test(VALUE ~ SITE\_ID, data = NO\_data, var.equal = FALSE)

print(welch\_anova\_result3)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Calculating daily average of NOY levels for each site

NOY = NOY\_data %>%

group\_by(SITE\_ID) %>%

summarize(average\_NOY = mean(VALUE, na.rm = TRUE))

# Histogram of NOY values

ggplot(NOY\_data, aes(x = VALUE)) +

geom\_histogram(bins = 50, color = "black", fill = "blue") +

labs(title = "Histogram of NOY Values", x = "NOY Value", y = "Frequency") +

xlim (0,10)

###data does not seem to be normally distributed by visual methods

##Assessing normality for NOY levels by SITE\_ID

ggplot(NOY\_data, aes(sample = VALUE)) +

facet\_wrap(~SITE\_ID) +

stat\_qq() +

stat\_qq\_line(color = "blue") +

labs(title = "Q-Q Plots of NOY by SITE\_ID") +

xlim (0,10)

# Histogram for each SITE\_ID

ggplot(NOY\_data, aes(x = VALUE)) +

facet\_wrap(~SITE\_ID) +

geom\_histogram(bins = 30, color = "black", fill = "gray") +

labs(title = "Histograms of NOY by SITE\_ID", x = "NOY", y = "Frequency") +

xlim (0,10)

# Perform Bartlett test for homoscedasticity

bartlett\_test\_result4 <- bartlett.test(VALUE ~ SITE\_ID, data = NOY\_data)

# Display the result

print(bartlett\_test\_result4)

levene\_test\_result4 <- leveneTest(VALUE ~ SITE\_ID, data = NOY\_data)

print(levene\_test\_result4)

##output: p<0.05 --> heterocedasticity

# Perform Welch's ANOVA

welch\_anova\_result4 <- oneway.test(VALUE ~ SITE\_ID, data = NOY\_data, var.equal = FALSE)

print(welch\_anova\_result4)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Calculating daily average of SO2\_GA levels for each site

SO2 = SO2\_GA\_data %>%

group\_by(SITE\_ID) %>%

summarize(average\_SO2 = mean(VALUE, na.rm = TRUE))

# Histogram of SO2 values

ggplot(SO2\_GA\_data, aes(x = VALUE)) +

geom\_histogram(bins = 50, color = "black", fill = "blue") +

labs(title = "Histogram of SO2 Values", x = "SO2 Value", y = "Frequency") +

xlim (0,3)

###data does not seem to be normally distributed by visual methods

##Assessing normality for SO2 levels by SITE\_ID

ggplot(SO2\_GA\_data, aes(sample = VALUE)) +

facet\_wrap(~SITE\_ID) +

stat\_qq() +

stat\_qq\_line(color = "blue") +

labs(title = "Q-Q Plots of NOY by SITE\_ID")

# Histogram for each SITE\_ID

ggplot(SO2\_GA\_data, aes(x = VALUE)) +

facet\_wrap(~SITE\_ID) +

geom\_histogram(bins = 30, color = "black", fill = "gray") +

labs(title = "Histograms of SO2 by SITE\_ID", x = "SO2", y = "Frequency") +

xlim (0,10) +

ylim (0,1500)

# Perform Bartlett test for homoscedasticity

bartlett\_test\_result5 <- bartlett.test(VALUE ~ SITE\_ID, data = SO2\_GA\_data)

# Display the result

print(bartlett\_test\_result5)

levene\_test\_result5 <- leveneTest(VALUE ~ SITE\_ID, data = SO2\_GA\_data)

print(levene\_test\_result5)

##output: p<0.05 --> heterocedasticity

# Perform Welch's ANOVA

welch\_anova\_result5 <- oneway.test(VALUE ~ SITE\_ID, data = SO2\_GA\_data, var.equal = FALSE)

print(welch\_anova\_result5)