



MSc in Data & Artificial Intelligence

Advanced Statistical Programming - ACCA7022_BIRM

‘Component Two’

“Data Analysis”

Submitted To

Dr Aliyu Sambo

Mail: a.sambo@uwtsd.ac.uk

Submitted By

Tareque Hasnat Tuhin

Mail: 2513674@student.uwtsd.ac.uk

“Submission Date”

20 March, 2026

Introduction

When I first got this assignment, I felt nervous because the dataset has over 9000 rows. I decided to take it one task at a time. Before writing code, I read about the dataset online. I learned that CO comes from cars, NO₂ from traffic, and O₃ forms when sunlight hits other pollutants. This helped me later when explaining my results.

My Visualisations

For time trends, I used line plots because they show changes over time clearly.

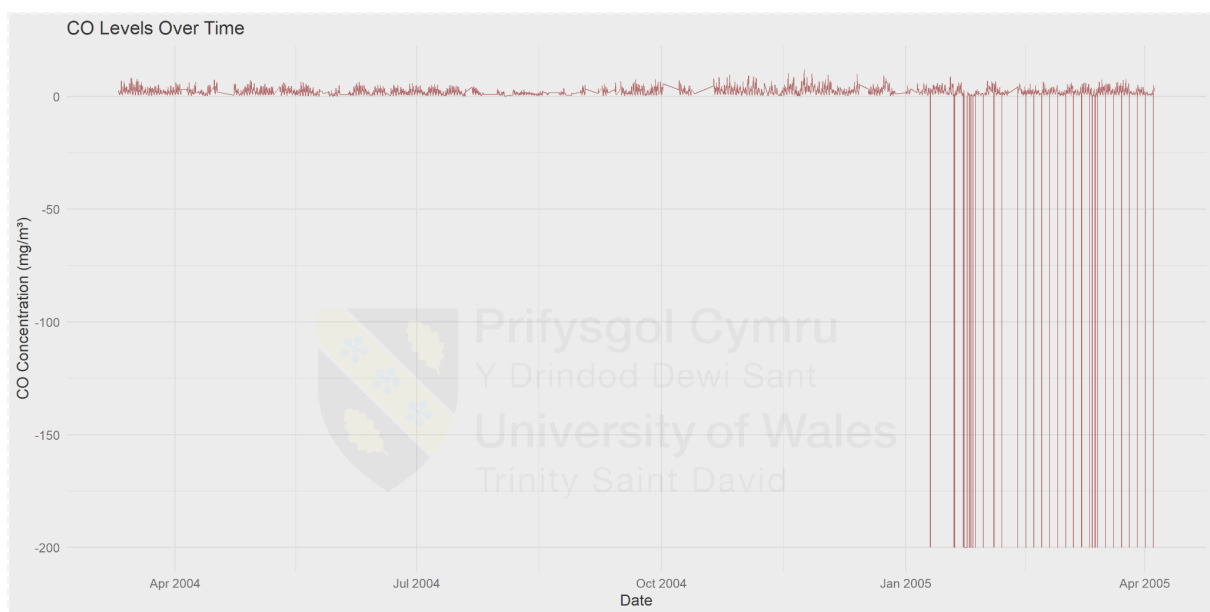


Figure 01: CO over time—I can see peaks that might be rush hour times when more cars are on the road.

Code I Used:

```
# --- Plot 1: CO Levels Over Time ---
p1 <- ggplot(plot_data, aes(x = DateTime, y = CO.GT.)) +
  geom_line(color = "darkred", linewidth = 0.3, alpha = 0.7) + # Line plot with transparency
  labs(title = "CO Levels Over Time",
        x = "Date",
        y = "CO Concentration (mg/m³)") +
  theme_minimal() # Minimal theme for clean look
ggsave("CO_timeseries.png", p1, width = 12, height = 6) # Save plot as PNG
print(p1) # Display plot
```

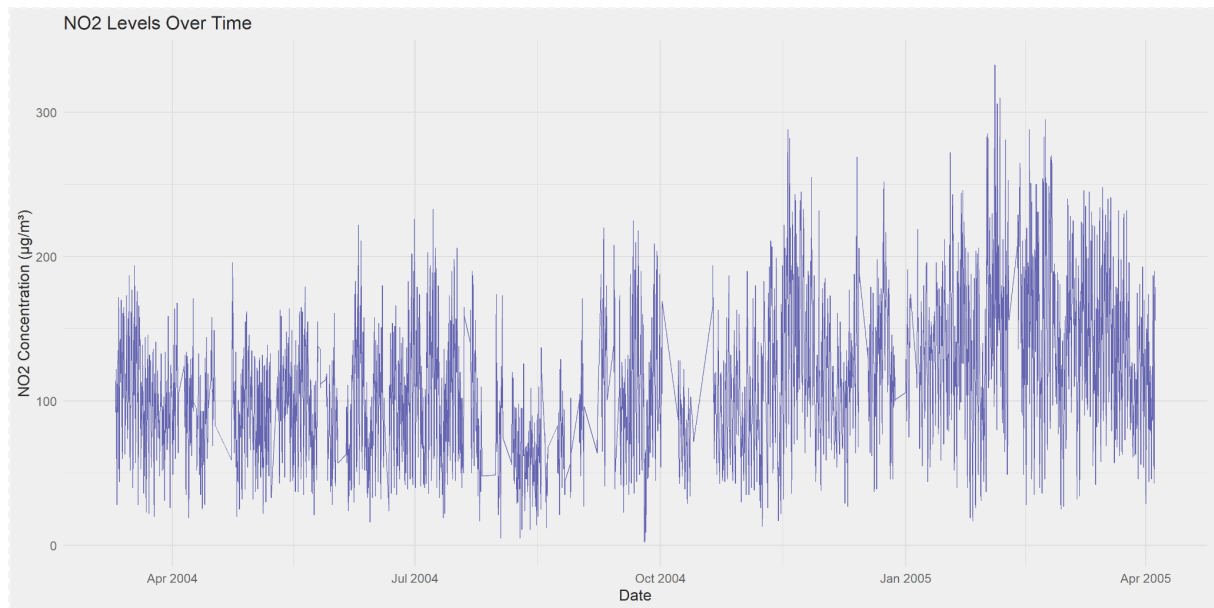


Figure 02: *NO₂ over time—similar pattern to CO, which makes sense because both come from traffic.*

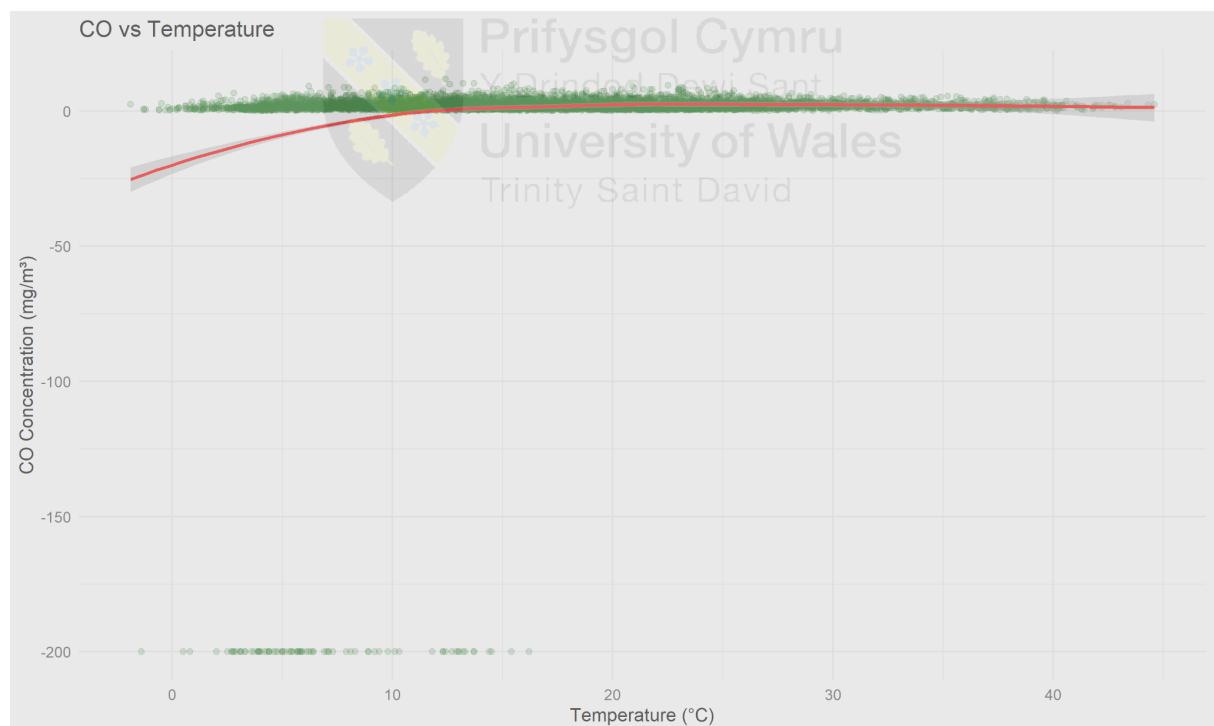


Figure 03: *CO vs. Temperature—more CO when it is cold, probably from heating and cold air trapping pollution.*

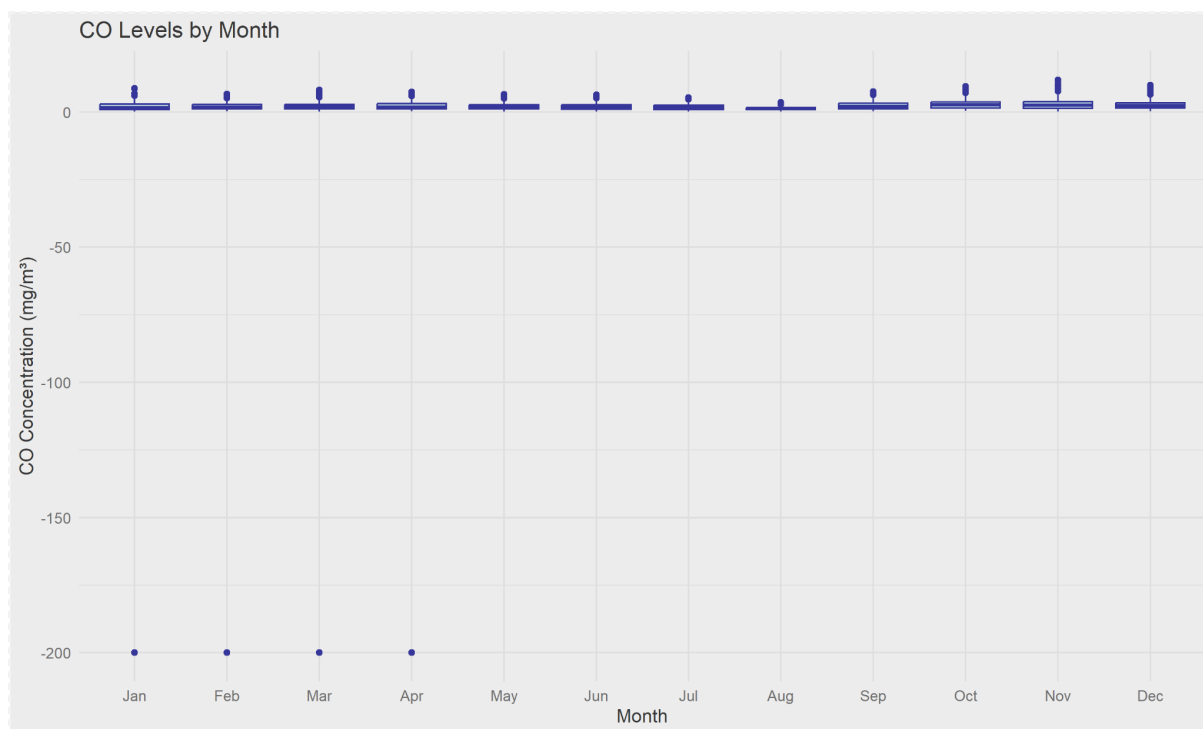


Figure 04: CO by month—winter months show higher CO levels.

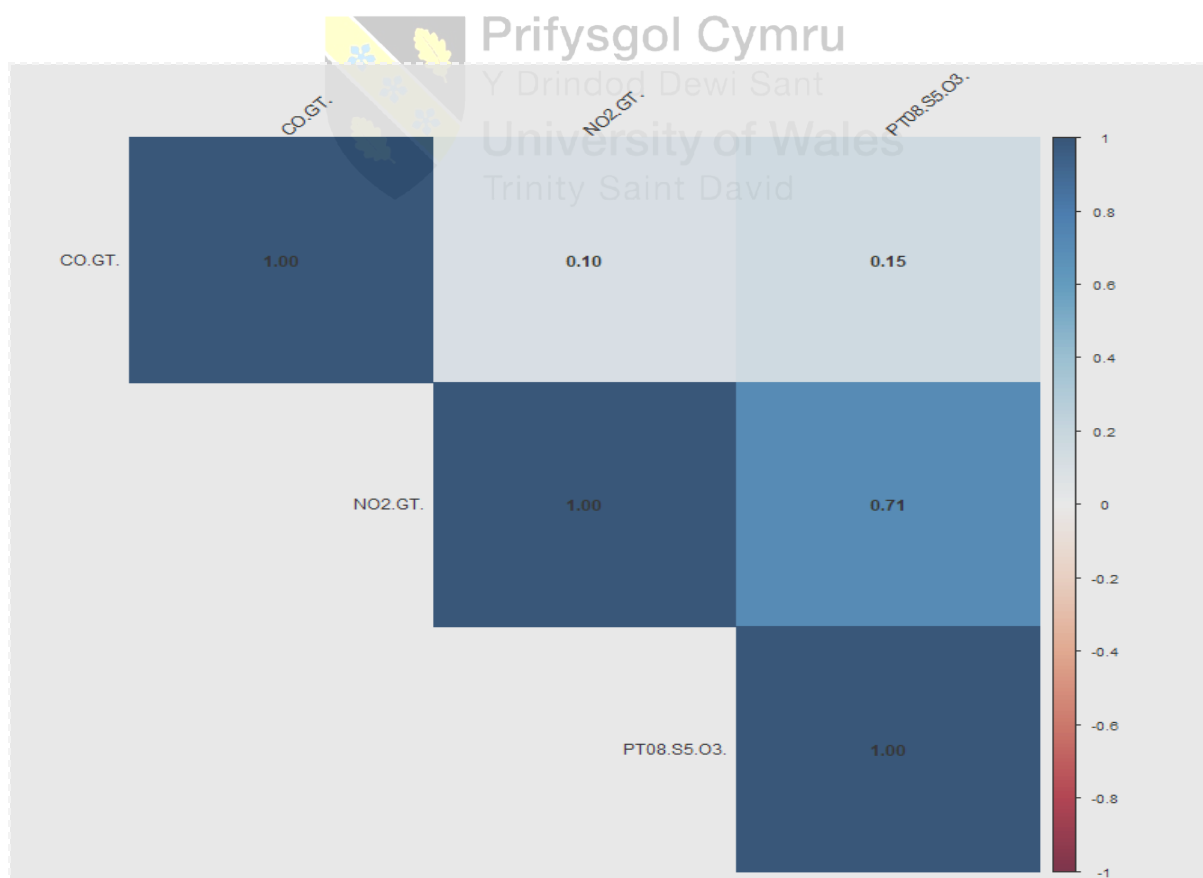


Figure 05: Correlation between pollutants—red means positive, blue means negative. CO and NO₂ are positively correlated.

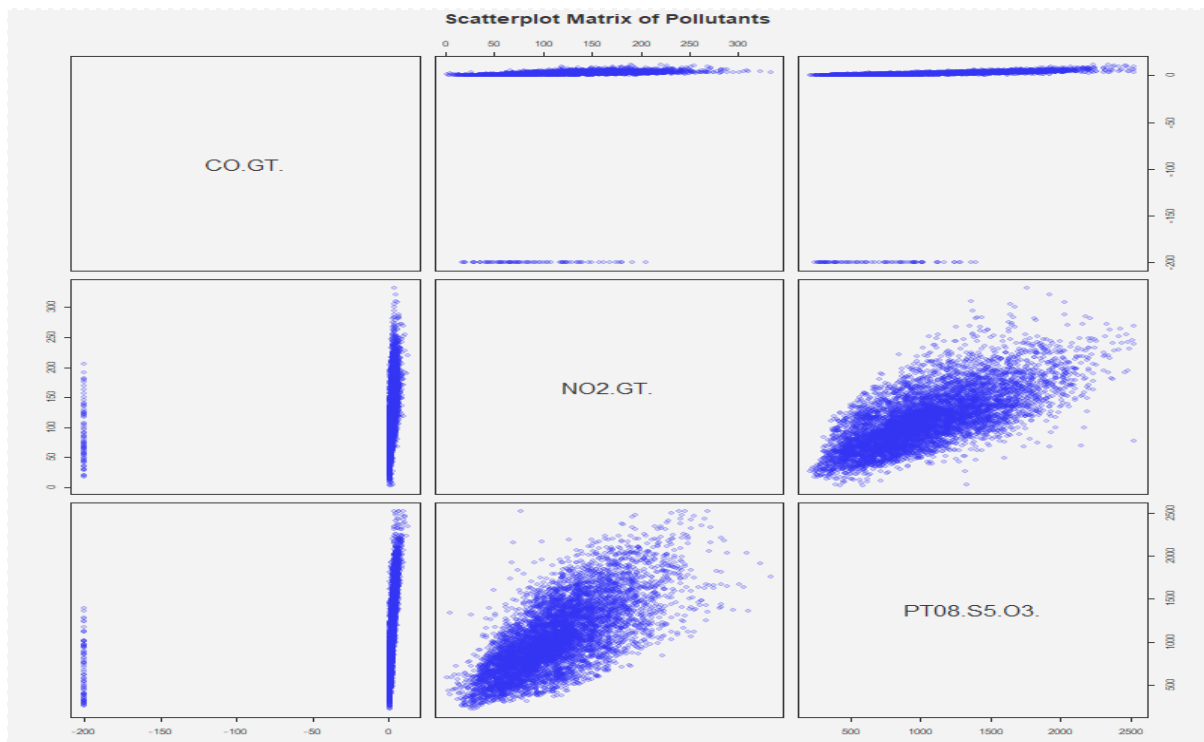


Figure 06: Scatterplot matrix—shows all relationships in one view.

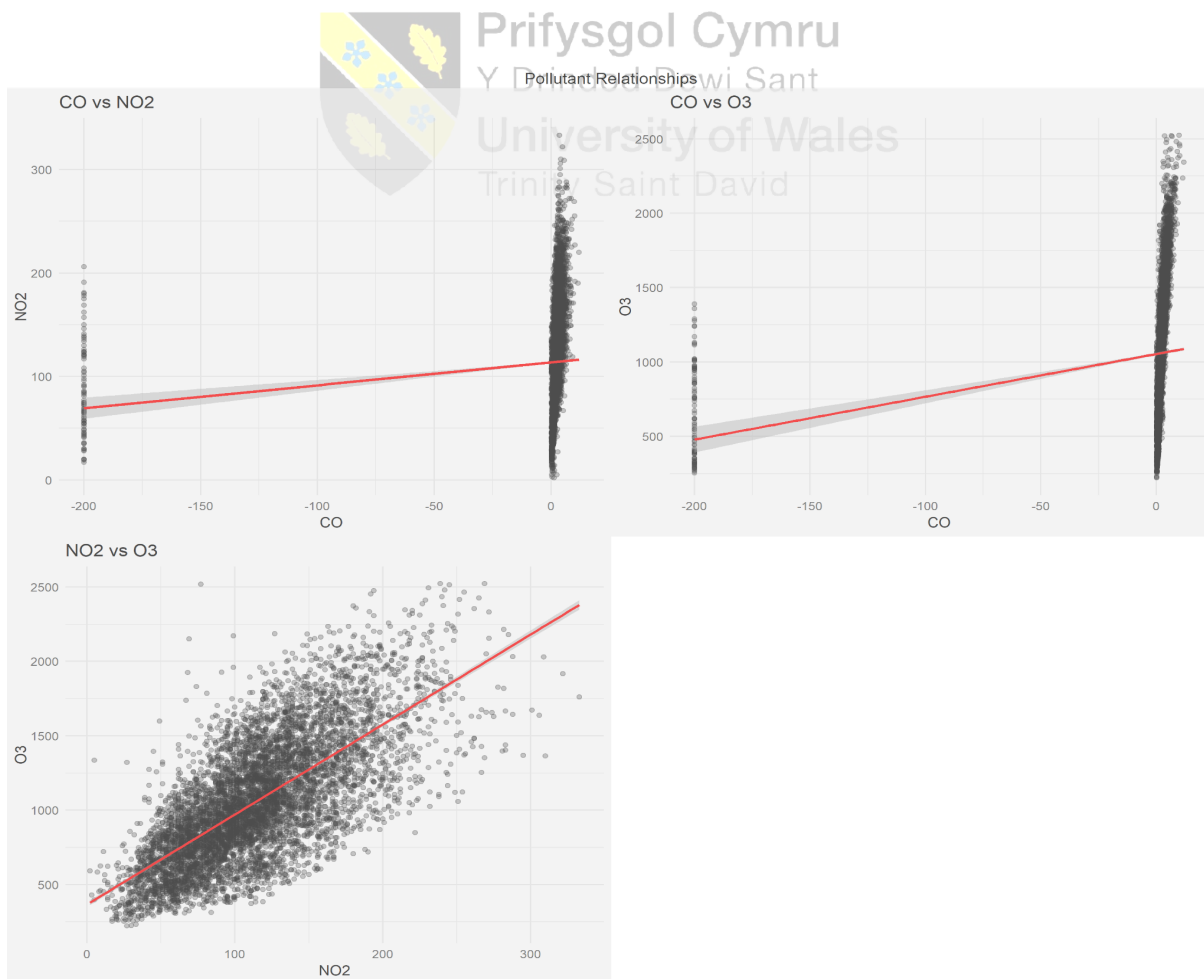


Figure 07: Individual scatterplots—each shows the relationship between two pollutants.

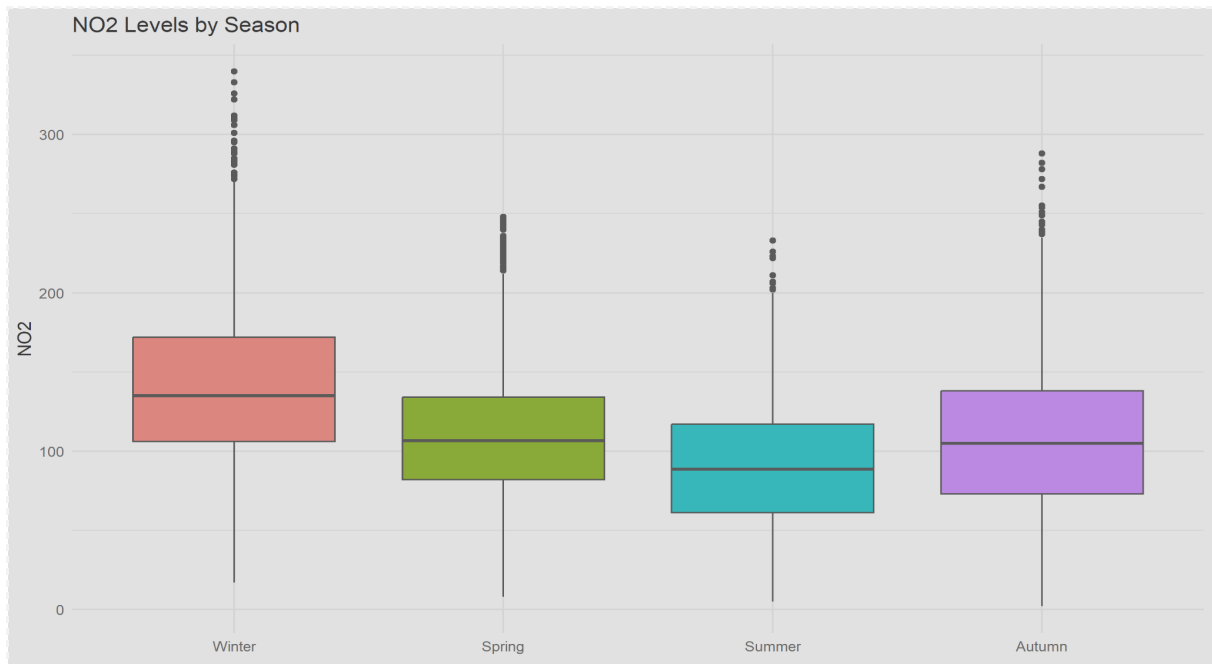


Figure 08: NO_2 by season—winter has the highest levels, and summer has the lowest.

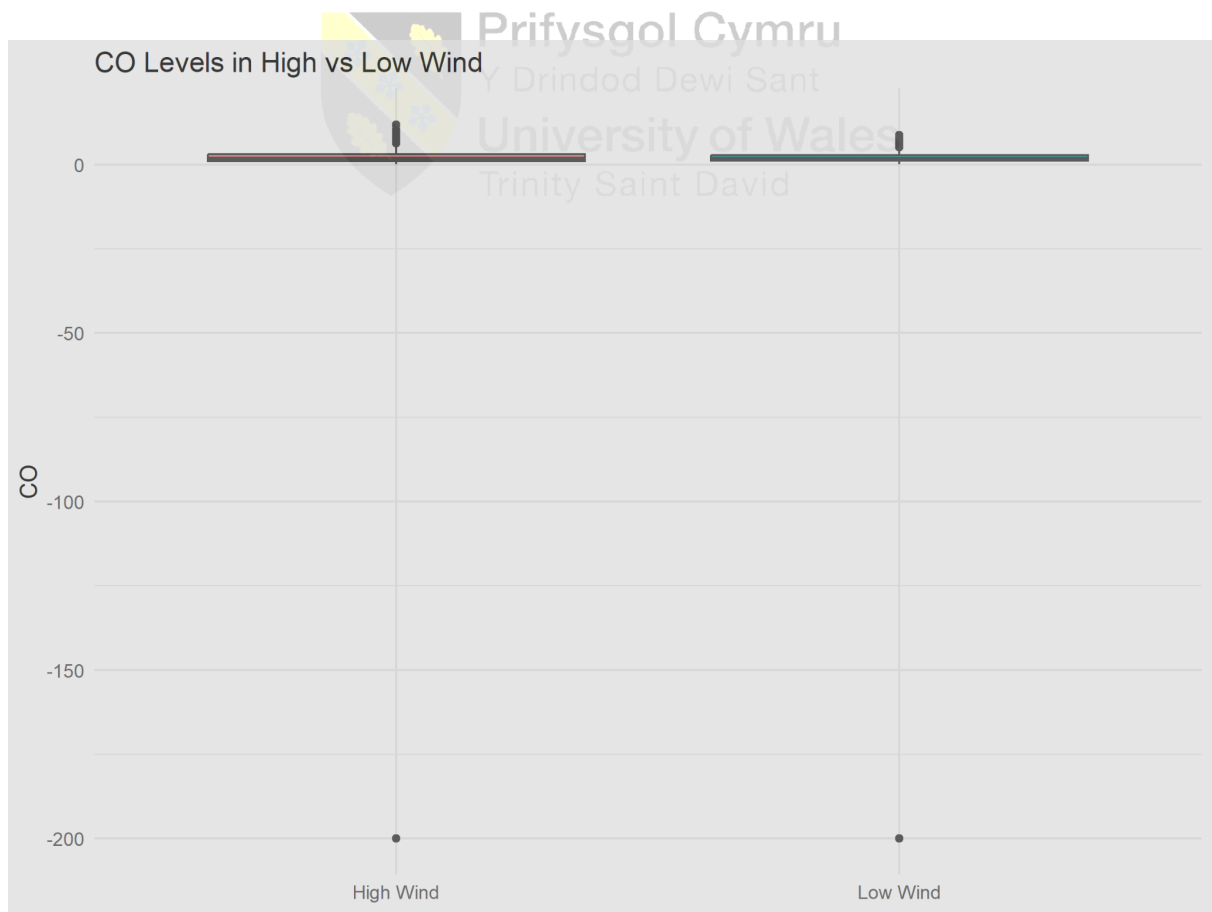


Figure 09: CO by humidity—lower CO when humidity is high, maybe from rain cleaning the air.

Problems I Faced

RStudio Kept Freezing: When plotting 9000 points, my laptop slowed down. I used all of this. This fixed the problem.

I Deleted My Code: I lost two hours of work. Now I save every 10 minutes and keep different versions.

Column Names Confused Me: I did not understand “PT08” columns. After reading, I learned “GT” means ground truth. I used **CO.GT** and **NO2.GT** for analysis.

Seasons in Wrong Order: Boxplots showed seasons alphabetically. I used `factor()` to set the correct order.

Could Not Find Saved Plots: I learned about working directly and now use `setwd()` at the start.

The correlation matrix had NA: I added `use = “complete.obs”` to fix this.

File Would Not Load: I used `file.chose()` to select my file manually. It worked.

Extra Empty Columns: I kept only the first 15 columns to remove warnings.

What I Learned

Real data is messy. I had to clean it, check it, and think about what I was doing. I learned to check assumptions before running tests. Just because a test is common does not mean it is right for my data.

Grapes are not just for looking pretty. They tell stories. The boxplots showed seasonal patterns before I ran statistical tests.

I also learned to be flexible. When wind speed was missing, I used humidity instead. This is what real data analysts do.

What I Would Do Better Next Time

I would check column names first. I wasted time guessing.

I would check missing data earlier.

I would write smaller code chunks and test each part.

I would back up to cloud storage.

I would explore data more before running tests.

Use of Generative AI

I asked QueenAI and ChatGPT to explain why my dates were wrong, and it told me about lubridate. I also asked about t-tests vs. Wilcoxon. I did not copy any code. I read explanations and wrote my own code. I used AI like a tutor, not to do the work for me.

Conclusion

This project was hard, but I learned a lot. I made mistakes, but I fixed them. I learned to check my data first before running tests. I feel more confident with R now. I am happy with what I did.

