



ACCA7022

Advanced Statistical Programming

Data Analysis Assignment Documentation Report

Programmer: **Mohammad Lakhi**

Student ID: **2512430**

Lecturer: **Dr. Aliyu Sambo**

Submission Date: 20th March, 2026

1. Introduction

This summary describes the method of analysing the AirQualityUCI dataset with R. It describes the visualisation decisions taken during the analysis, the difficulties encountered and lessons learned, and potential improvements for future work.

2. Visualization

In this assignment, I have generated many types of plots to investigate the AirQualityUCI dataset. Each plot type was chosen depending on the nature of the data and the specific subject under analysis.

Time trends were shown by line plots with a smooth trend line. This was the best option because the dataset includes hourly readings recorded over a whole year. A line plot vividly illustrates how pollution levels grow and fall over time. The `geom_smooth()` algorithm introduced a loess trend line, which helped identify seasonal trends even when the raw hourly data was noisy. For example, the CO time series plot was created using:

```
ggplot(AirQuality, aes(x = DateTime, y = CO.GT.)) +  
  geom_line(color = "steelblue", alpha = 0.6, size = 0.4) +  
  geom_smooth(method = "loess", color = "red", se = FALSE)
```

I used scatterplots to demonstrate the correlations between pollution and weather variables. I choose them since both variables were continuous and I wanted to check if there was a consistent pattern between them. The scatterplot of CO and temperature revealed a negative link, which made sense to me because colder months result in more fuel consumption and traffic.

For group comparisons in Tasks 4 and 5, I used boxplots. I preferred boxplots for these jobs since they display the median, spread, and outliers all at once. When I compared NO2 levels between seasons, the boxplot immediately revealed that winter had significantly greater levels than summer, which was simple to explain and discuss.

For the correlation analysis in Task 3, I used a heat map. I selected this because I had five pollutants to compare, and a heatmap displays all of the correlations in a single colour-coded image. It was considerably simpler than creating ten different scatterplots.

3. Challenges and Things Learned

The most difficult challenge I faced at first was properly loading the dataset. The AirQualityUCI file uses semicolons as separators and commas as decimal points, which differs from the standard format I was familiar with. When I first used `read.csv()`, all of the numeric columns came in as text, and nothing worked. Following some investigation, I discovered the fix:

```
AirQuality <- read.csv(  
  "AirQualityUCI.csv",  
  sep = ";",  
  dec = ",",  
  header = TRUE,  
  stringsAsFactors = FALSE  
)
```

The second challenge was dealing with missing values. Instead of leaving missing sensor readings blank, the dataset assigns them the value -200. My summary statistics appeared completely incorrect until I realised this and replaced all -200 values with zeros. This was a good lesson to always read the dataset documentation before beginning any analysis.

```
AirQuality[AirQuality == -200] <- NA
```

The third thing I learned was about selecting the appropriate statistical test. I'd heard of the t-test and ANOVA before, but I didn't really understand when to use them. This assignment taught me that I needed to use the Shapiro Wilk test to first determine normality. Because my data was not normally distributed, I applied the Wilcoxon and Kruskal Wallis tests instead. This felt like a significant improvement in my understanding of statistics.

4. Improvements for Future Work

If I were to repeat this analysis, I would do three things differently:

- First, I would compute effect sizes using the hypothesis test results. Knowing whether a difference is statistically significant is useful, but knowing how large that difference is in practical terms is more important. To add this information, I would use measures such as eta-squared.
- Second, I would distinguish between weekday and weekend data in the hourly analysis. My boxplot in Task 2 showed clear rush hour CO levels, but I couldn't tell if the pattern changed on weekends. Separating the two groups

would provide much stronger evidence on whether traffic is the primary source of pollution.

- Third, I would create a single combined plot that displays all pollutants across seasons. This would make seasonal comparisons much easier to see, resulting in a cleaner and more professional finished product.

5. Conclusion

This project significantly improved my practical R skills. The most important lessons I learned were how to deal with real-world data issues such as incorrect formats and missing values, as well as how to select the appropriate statistical test based on evidence. After completing this assignment, I feel more confident working with large datasets.