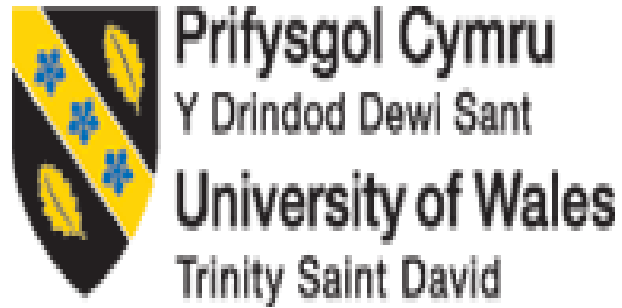




Program: Msc Data and Artificial Intelligence

Student ID: 2418784

Student Name: SADANAND TANNERU

Module Code: ACCA7022

Module: Advanced Statical Programming

Lecturer: Dr.Aliyu Sambo

1. Reflection

The coursework helped me in improving my technical proficiency in R as well as my analytical skills when dealing with actual environmental data. The first strategy I had was to divide the seven tasks that I had to do into small milestones and accomplish them in a sequence. I began by examining the data and getting familiar with its format as I realised that no analytical or visualisation would be accurate without appropriate data cleaning. This impacted my writing of the code because I concentrated on how to convert the raw data into a format accessible to R, before being allowed to manipulate the data by visualisations and statistical tests.

There were multiple choices that had to be made during the data import and cleaning phase. The raw data was of mixed type, had missed headers and measurements of pollutants in the form of character rather than numbers as there was a case of No data in one or more cells. I decided to apply `na_if()` with `mutate(across(...))` since I needed to transform all pollutant fields into numbers in one pipeline. This method was less clumsy and simpler to repeat than filling in values row-by-row. I also added a `DateTime` column by adding the separate `Date` and `Time` columns together since future tasks would need hour, month, and year filtering and it was much easier to work with an actual timestamp.

In creating visualisations, I was always reasoning to select the most understandable data trend representation. To analyze missing data (Task 2) I initially calculated missing percentages in a table, but I quickly realised that percentages alone were not impactful. Instead, a bar plot gave a visual impression that PM10 and PM25 had the highest number of missing data, which was easier to interpret and explain. In Task 3, when I was creating a function to display the levels of pollutants as time progressed, I used a line graph since it is the most intuitive type of representation of change over time. The annotation of highest and lowest observed numbers was not random, because I wanted the viewer to see the trend without having to check each point of the data separately.

In Task 4, I put several pollutants on the same plot to be able to compare them. I considered the idea of plotting each separately, but chose not to plot it since the idea was to demonstrate overlapping patterns and spikes. The use of the scatterplot with a regression line in Task 5 was dictated by the fact that correlation cannot be presented meaningfully in a bar chart or a line chart. Placing the correlation coefficient on the graph served to directly translate a statistical outcome to a visual one.

Tasks 6 and 7 (statistical testing) made me enhance my own interpretation skills. In the t-test, I compared the levels of ozone at low and high wind speeds. The interpretation of the output made me realize how assumptions and p-values are used to draw a conclusion on whether to reject or accept the

null hypothesis. ANOVA test was initially confusing, however, after I ran the Tukey post-hoc test, I was able to understand not only that a difference existed, but also where that difference was.

I encountered a number of problems in the course of action. A few of the graphs were initially plotted as blank due to NA values present in some fields of pollutants at some point in time. I solved this with drop na and filtering the data further. The data also included blank column names and text values like no data, which lead to the failure of the mutation step with the message Can not transform a data frame with NA or blank name. I managed this by providing temporary names and converting the non-numeric data to NA. I was given a warning later on coercion and a mistake in the t-test because of the absence of numeric values in filtered groups. Discarding rows where there were missing values and making sure that the type was correctly converted resolved the problems and made further analysis possible.

In an extension of this project, I would consider forecasting models like ARIMA or use machine learning to predict pollutants. I would also create an interactive dashboard in R Shiny that would allow the user to dynamically choose the pollutants and time ranges. All in all, this coursework has made me very confident in my ability to manage large datasets, write functions that are reusable in R, and interpret statistical output in a meaningful manner.