National College of Ireland

Bachelor of Science in Computing

Data Analytics

2020/2021

Sean Burke

X17132118

X17132118@student.ncirl.ie

Exploratory Analysis on Air Pollution

Technical Report

Contents

Executive Summary

1.0 Introduction

1.1. Background

Air pollution has become an increasing problem since the mid-19th century, the growing popularity of burning fossil fuels as a means of producing power has been growing ever since. The current popularity of alternative options such as using electricity to replace the combustion engine in electric cars has proven that it is no longer necessary to produce such large amounts of pollution via burning these fossil fuel.

It is still true that cars that run on either petrol or diesel less costly to buy when considering second hand options. It is also true that the charge life on an electric vehicle makes it less appealing, and even the charging stations are not available in all countries. The electric option would be considered only viable in developed countries around the world. Developing regions such as remote African countries where the introduction of electric vehicles is far from the highest priority, would certainly not be using an electric option.

The remote areas are not where the problem lies in this global crisis, it is the overpopulated cities where there are factories and motor vehicles running continuously. An example of this would be Hong Kong, China. This is a city that has high levels of pollution in the air, to the point where citizens that travel around the city are commonly seen wearing face masks to prevent the inhalation of toxic particles caused by pollution. China alone contributes to roughly one quarter of the global emissions.

The spread of pollution is evident as particle matter from the California fires had made its way across the United States of America to New York, as stated by (Milman, 2021). Wind patterns are speculated to a major contributor to the spread of particles in the air (including the hazardous particles emitted by combustion engines).

1.2. Aims

The aim of this exploratory analysis is to find out how much of a factor does wind have on the spread of air pollution. By following common wind patterns, the analysis should show that there is a correlation between areas along the same wind direction. A strong correlation along the same wind pattern would prove:

Hypothesis: Wind patterns have a significant impact on the pollution levels in areas that the wind travels through.

Considering that there are other factors that dictate the pollution levels in an area, such as:

- Population
- Factory's
- Building sizes
- Economy

I think it would be sufficient to have an alpha value of 0.1 when testing for significance.

1.3. Technology

The technology I will be using in this project is:

- Spyder for Python programming
- RStudio for R programming
- GitHub for pushing the code to an online source
- Libraries such as rvast for web scraping and stringr for selecting data from data frames containing strings.
- Microsoft Word of writing my report
- Microsoft Excel for manipulating and storing my data

1.4. Structure

This document contains many different sections that show how the analysis was performed and concluded in this report. The main sections are as follows:

- Data – This is an overview of the data that was used in this analysis, where it came from, how it was used/ manipulated, and all its different variables.
- Methodology – What methods were used when going from data to analysis.
- Analysis – What was used to help analyse the data, Graphs, models etc.
- Results – What were the results from the analysis and explain how they were obtained.
- Conclusion – What were the conclusions gathered from the results. Did the results prove/ disprove anything, what was learned from the analysis?
- Further research – If given more time and resources, what would be done with this project. Is there anything you would like the chance to do differently?
- References – list of references from the academic research that was done on this project.
- Appendices - The project proposal and all the journals that were

2.0 Data

The data that was used in this analysis is separated into two sections:

1. The data that was used at initial stages of testing, which includes many different types of pollution values. This was data that was later made redundant by data used in the second section. This data can still be made useful by showing that the data obtained by OpenAQ was accurate when compared to the yearly average that was obtained by IQAir. If the data didn't match what was expected, then one or both sources would be faulty data.
2. The second section contains the data that was used for the results and conclusions. This data consisted of just PM 10 values from three different locations within two countries. The reason for this was so that the locations could be picked based on the wind patterns of that area and compared against an area that the wind pattern does not go through.

There are two data sources for this project:

1. OpenAQ (OpenAQ, 2021)
2. IQAir (IQAir, 2020)

Section 1

This section contains the data that was used in the initial stages of the project, this data showed many different types of particle matter that was collected by the city of cork.

OpenAQ is a non-profit organization that collects data on air quality around the world in efforts to inform people around the world about the increasing rise in air pollution. OpenAQ provides an open access API for all the collected data. The data files can also be downloaded as .csv files under specific countries, cities, and area zones.

The data that was collected for this project was from Cork city between July 2019 and December 2021. This data was downloaded as a csv file for testing the data, the data contains a data frame of 11 columns and 58065 rows. The dataset contains a mixture of character, numeric, and integer values that make up the data frame, each column consisted of:

- Location ID
- Location
- City
- Country
- Universal Time Coordinated
- Local Time Coordinated
- Parameter that was measured
- Parameter value
- Unit of the value
- Latitude of location
- Longitude of location

Once that data was acquired, it was imported into R Studio to view. The data was then analysed with the summary method on the dataset to show if there was any outlying data as seen in figure 1 below

```
> cork_data <- read.csv("cork_data.csv")
> View(cork_data)
> summary(cork_data)
   locationId     location            city             country             utc                local              parameter        
 Min.   :7967   Length:58065       Length:58065       Length:58065       Length:58065       Length:58065       Length:58065      
 1st Qu.:7967   Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Median :7967   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :7967                                                                                                                    
 3rd Qu.:7967                                                                                                                    
 Max.   :7967                                                                                                                    
     value               unit              latitude       longitude      
 Min.   :-1731.7707   Length:58065       Min.   :51.91   Min.   :-8.475  
 1st Qu.:    0.3467   Class :character   1st Qu.:51.91   1st Qu.:-8.475  
 Median :    4.0000   Mode  :character   Median :51.91   Median :-8.475  
 Mean   :   12.0831                      Mean   :51.91   Mean   :-8.475  
 3rd Qu.:   17.0000                      3rd Qu.:51.91   3rd Qu.:-8.475  
 Max.   : 2089.1853                      Max.   :51.91   Max.   :-8.475
```

Figure 1

From this data it is clear to see that there are some outliers/faults in the data, considering the minimum value of the value column has a value of -1731.7707 and there should be no negative values for the results.

Considering the Value, Latitude, and longitude have decimal points it is clear to see that these classes are numeric and not integers because of the decimal point values so they will not have to be adjusted.

Upon inspecting the data frame values, the data was not the same parameter type, the parameters were:

- Pm10
- Co
- O3
- So2

The data needed for this project would have to be all the same type of parameter to get accurate results when comparing the data to the scraped data. Failing having the same parameter, a formula would have to be used to convert parameters.

The second dataset was a table scraped from the web. This data contained the worlds most polluted countries between the years 2018 and 2020, the data is based on the average pm2.5 value per country. The data contains a data frame of 6 columns and 106 rows, the dataset contains string values for each column that makes up the data frame, each column consists of:

- Rank of highest polluted country/ region
- Country/ region
- 2020 average
- 2019 average
- 2018 average
- Population of the country/ region

Web scraping is the extraction of data from websites, this is done with using a library for managing the scraped data, in this case using the “rvest” library for r. A link is stored as a variable and then the page variable is created using rvest’s read html function on the link. Then using the page variable, the desired html nodes are selected so that not all data from the page is imported.

```
11 link <- "https://www.iqair.com/world-most-polluted-countries"
12 page <- read_html(link)
13 scrape_data <- page %>% html_nodes(".mat-elevation-z8") %>% html_table() %>% .[[1]]
```

Figure 2

3.0 Methodology

Pre-processing is the process of:

- Cleaning data
- Data Integration
- Data transformation
- Data reduction or dimension reduction

Before pre-processing the data might not be suitable for the analysis you are trying to receive. The analysis may not require all columns provided, the values in some datasets could be missing or be not within the scope of what the analysis is trying to prove.

Pre-processing initial data

In this section not all columns were necessary. Instead of using all 11 columns, the data frame was narrowed down to 5 columns:

1. City name
2. Country
3. Local time coordinated
4. Value
5. Parameter

This was done by creating a vector and assigning the corresponding data frame vector to the newly created vector, resulting in five new vectors containing the relevant data from the dataset, as shown in the Figure below.

```
27 cork_city <- cork_pm2.5$city
28 cork_country <- cork_pm2.5$country
29 cork_local <- cork_pm2.5$local
30 cork_parameter <- cork_pm2.5$parameter
31 cork_value <- cork_pm2.5$value
```

Figure 3

From here we have the vectors that will be used in the new dataset for the analysis, but there is still the issue of negative values in the dataset. The decision was made to remove the negative values and replace those values with 0, this is because there were only 86 values out of the 13438 total pm10 values that were negative. This would not negatively skew the results from being accurate.

To replace the negative values with 0, a new vector was created that used the “which” method to get all the values that were greater than or equal to 0. The new vector contained 13352 values that had a value of 0 or greater.

As mentioned in the data section, the values were not all the same parameter type, so it was decided to use the pm10 values as there was no pm2.5 values provided. The parameter class has character values, so to obtain only the pm10 values the dataset had to be refined down by using the “stringr” library to detect the data that had a parameter of “pm10” by using the string detect method on the data frames parameter column.

The next step is to get the max length of all the vectors going into the new data frame. Considering the length of the four other vectors are the same length of 13438, which is greater than the length of the vector of values that are greater than 0, the max length variable was saved as 13438.

The max length variable was then used to add a “0” to every value position if there was no assigned value, by repeatedly adding “0” to every row that did not have a value. In this case it would add a “0” to the 86 rows that had negative values.

```
cork_parameter <- c(cork_parameter, rep(0, maxlength - length(cork_parameter)))
cork_refined <- c(cork_refined, rep(0, maxlength - length(cork_refined)))
```

Figure 4

The data is now refined to 5 columns of data that contain the city name, Country, time and date of the recorded data, value of pm10 equal or greater than 0, and a parameter type of pm10. From here the summary function is run on the refined dataset to see what the minimum value is set to 0 and the length of the columns is 13438

```
> summary(refined_data)
  cork_city         cork_country        cork_local         cork_refined    cork_parameter
 Length:13438       Length:13438       Length:13438       Min.   :  0.00   Length:13438
 Class :character   Class :character   Class :character   1st Qu.:  3.00   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median : 10.00   Mode  :character
                                                          Mean   : 12.88
                                                          3rd Qu.: 18.00
                                                          Max.   :622.00
```

Figure 5

As seen in figure 5 above, the minimum value is now at “0” and the length of the columns is 13438.

The scraped data was pre-processed to show just the Ireland data from the scraped dataset so that the two datasets could be compared against each other. This is the best method to see if the data is accurate or not.

The stringr library was used in the same process as the last dataset. The Ireland string was selected from the country/ region column and the corresponding data with it. The different Ireland columns were stored as separate variables to show the value as numeric type, then the average pm2.5 values from the year 2018 – 2020 were added to a vector. Another vector was made for the names of each corresponding value and then both vectors were added to a data frame.

Pre-processing final data

In this section the data was taken from three different areas in Dublin Ireland, and three different areas in southwestern India. For this data to be valid, the data must be recorded on the same day as the other areas within that region.

Excel

Upon the initial analysis of the data, the time was in the wrong format from what was needed from the dataset, the local variable was altered with a function to take just the first ten characters from the string. This allowed the date function to be used in R when the data was read in. This process was repeated for the five other csv files.

	locationId	location	city	country	utc	local	paramete	value	unit	latitude	longitude
2	12043	Haji Colon	Raichur	IN	2021-01-25T23:30:00+00:00	25/01/2021	pm10	54	Âµg/mÂ³	16.20814	77.34842
3	12043	Haji Colon	Raichur	IN	2021-01-25T13:30:00+00:00	25/01/2021	pm10	227	Âµg/mÂ³	16.20814	77.34842

Figure 6

R Studio

The csv files were then read into R studio, the data was viewed first to see if the days and months matched up (some of the data has the first entry for February, which is not what is wanted). If the data has the first entry from the next month, the first row is then removed from the dataset. From here the data is then further refined to the data needed:

- The date
- The value
- Latitude
- Longitude

After using some of the in built formatting functions in r the days were separated and a new data frame was made. The data frame contained all the individual days, the values, latitude, and longitude that was associated with the row from the dataset. The data was uneven in both the quantity of recorded data per day and days per month (some datasets would have four entries for the 4th day of January and 10 entries for the 20th of January, some dates were not recorded at all). An example of this would be the south Dublin entries having a total of 120 entries, but only covering the last 6 days of January. To solve this issue and not skew the data I got the mean value for each day recorded in January from all the different datasets. Once I had all the mean values, I was able to pick the dates that all three areas had in common, an example of this is Dublin having all three areas using the data from the last 6 days of January.

	comparison_fingal	comparison_sd	comparison_dc
1	17.278275	7.707184	9.636286
2	15.096776	12.161773	11.046250
3	11.256462	10.776913	11.370979
4	10.893854	9.057198	9.408208
5	11.332939	12.881154	15.331917
6	8.973872	8.782958	11.686239

Figure 7

The data was now ready for analysis.

4.0 Analysis

There were two methods I employed for the initial analysis:

1. Ggplot
2. Conversion algorithms

In my analysis I made use of the mean method to print out the mean of the value column. From this value I was able to find the pm2.5 value through a conversion algorithm.

$$pm2.5 = 0.75 \times pm10$$

This value was then saved as a variable. The value was then added to the data frame of Ireland average data so that the values could be plotted and compared with a Ggplot.

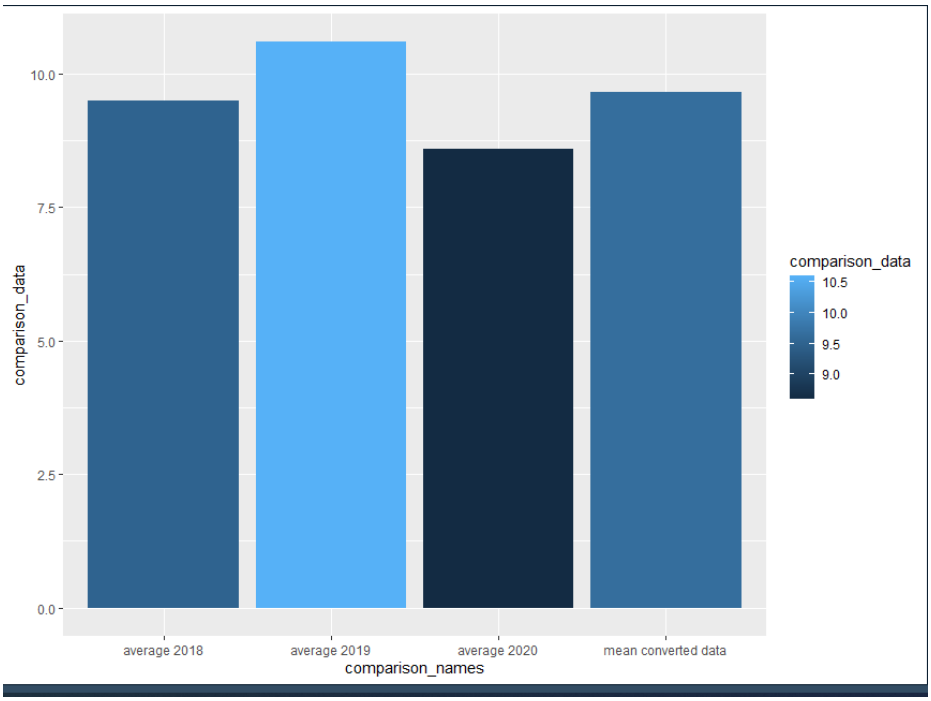


Figure 8

This column based Ggplot was used because the data frame consisted of two vectors, the name which was used on the x axis and the data values which was used on the y axis. This Ggplot can graphically represent how close the values are from each other based on the height of each column. I chose to add the colour fill on the data values to better illustrate the differences between the data.

Ggplot was also used on the values of the Cork city dataset to show the count of pm10 recorded between 0 and 100 µg/m³ with a bin range of 2. Considering the mean µg/m³ value is between 8 – 10 µg/m³ since 2018, the data is expected to populate mostly around that area on the plot.

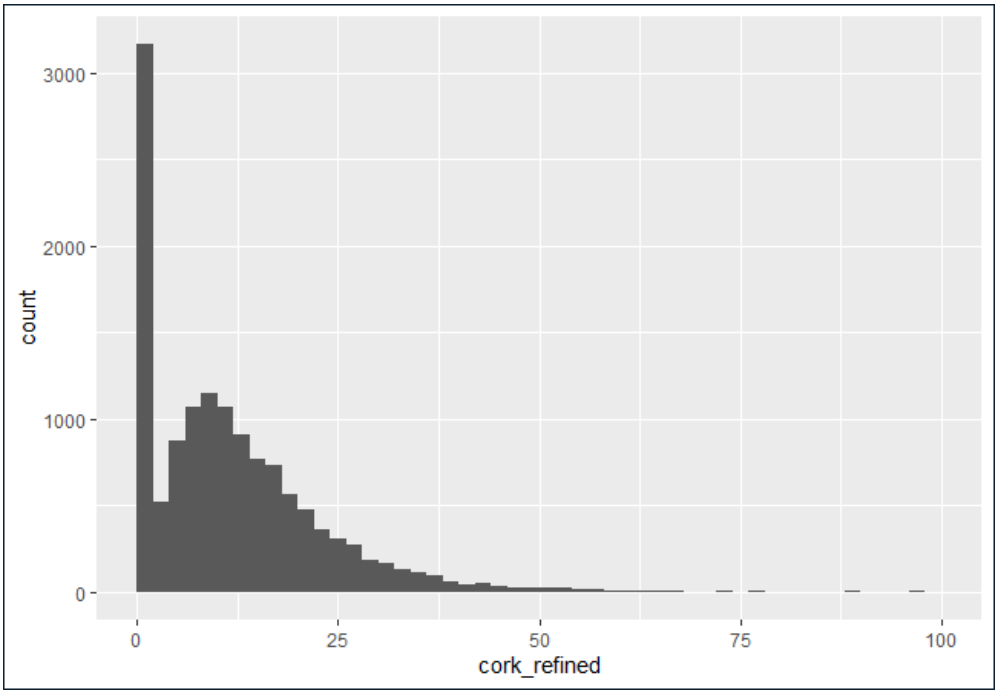


Figure 9

This plot was used to graphically represent the bin values that most of the data gathering at. There were other plots that could be used such as the area or density plots, but I felt that the histogram best showcased a visual aid to receive results from.

Data analysis from the second set of data

Before the analysis is made on these datasets, the wind direction must be obtained to know which dataset will be the dependant set. The global wind patterns dictate that in Ireland, specifically in Dublin, wind patterns travel from the south, and occasionally the southeast during the month of January. When analysing the Indian patterns, I opted to stay away from the northern area of India, as monsoons could lead to inaccurate results. The areas that were chosen were in the southwestern cities of India where the wind patterns come from the north-eastern direction.

There were three methods used to analyse the data from the pre-processed datasets:

1. Multi-scatterplot
2. Linear regression
3. Multiple regression

Multi-scatterplot

Once the data was pre-processed, the data could then be compared against each other. Using a scatterplot on all three areas covered in both Dublin and India, the data was able to show which areas showed visual evidence of correlation and which areas showed no correlation. Ideally the points would follow an imaginary line of correlation that travels somewhere between the bottom left and top right corner of the plot. If the plots are not traveling along this line, the plots would be uncorrelated. This method saves time on other

analysis options as it gives a good visual representation of if there is or is not correlation between the data sets.

Linear regression

The linear regression model is a means of comparing two sets of data against each other to determine a level of correlation. This is given through many different factors provided by the model. After the function is ran, the initial data received is the residuals, which determine how far on average the points are away from the line. Ideally the Min and Max values would be equidistant from zero in the best-case scenario. The standard error of the estimates and the t-value are both provided to show how the p values were calculated. Depending on the alpha value desired by the hypothesis, the p value will dictate if the alpha value is satisfied, in this case the desired level of confidence is 90%.

Multiple regression

Multiple regression is very similar to linear regression, the key difference is that instead of two data sets being compared, there are three or more. In multiple regression there is the dependant variable in which the model is proving correlation with two or more independent variables. In the case of this analysis the area in which the wind travels through last would be the dependant variable, to see if the other areas are significant pollution rates when combined. Multiple regression shows the p values for each independent variable against the dependent too, so It would be clear to see outliers in the data.

5.0 Results

From the analysis of data that was chosen the following results were obtained:

Initial results

Although these results were not used for the hypothesis or conclusions, these results were still important to the analysis. The initial results will aid in viewing to accuracy of the data provided by the data sources. As can be seen by the histograms the mean converted data from the cork dataset shows that the average mean particle matter is between 8 – 10 $\mu g/m^3$ which is exactly what was expected. When taking the size of the dataset into account, there is no reason to believe that the data is in any way false. The purpose of getting new data for this analysis was because, combining different particle matters was proving too inconsistent upon further evaluation of other datasets. The second method yielded better results and a more straightforward approach.

Final results

The second set of results are as follows:

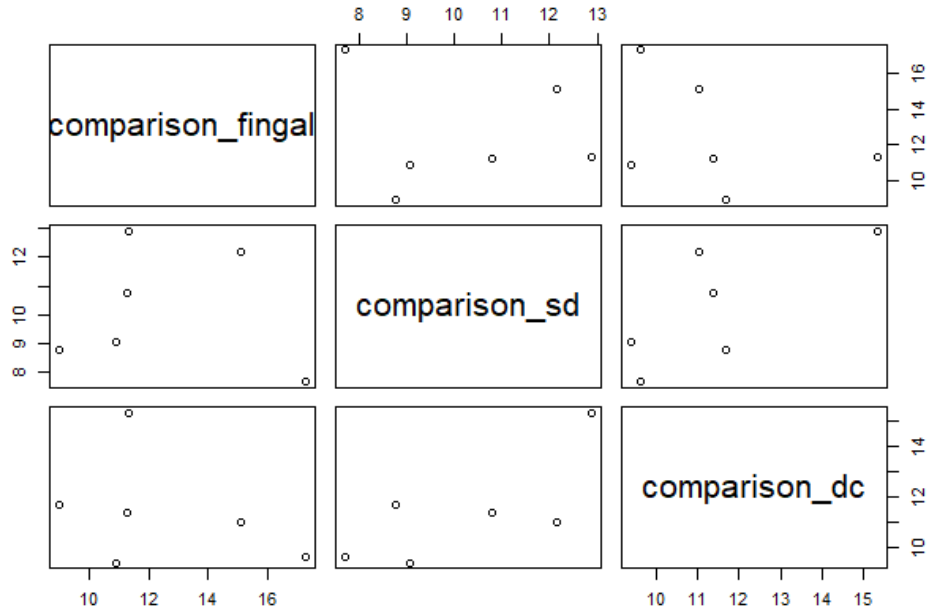


Figure 10

In the Figure above the last six days of the month of January were compared against eachother on a scatterplot in the hopes that there would be some correlation between at least two out of the three data sets. It can be seen when looking at the plots between Dublin city and south Dublin there are points going from the bottom left of the plot to the top right (except for one outlier) these two sets could be considered positively correlated. To prove this, the linear regression model was then used on the data sets. It is clear from the results of the linear regression model that what was visible in the plot was correct, there is a significant correlation between Dublin City and South Dublin.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.5337     3.6577   0.966   0.3887
comparison_sd  0.7704     0.3518   2.190   0.0937 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.608 on 4 degrees of freedom
Multiple R-squared:  0.5452,    Adjusted R-squared:  0.4315
F-statistic: 4.795 on 1 and 4 DF,  p-value: 0.09371
```

Figure 11

The figure below represents the corelation between Fingal and South Dublin, unlike the results from the figure above, there is very little correlation between the two data sets. The P- value is too high for there to be significant correlation.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.3453     7.7793   1.844    0.139
comparison_sd  -0.1832     0.7483  -0.245    0.819

Residual standard error: 3.42 on 4 degrees of freedom
Multiple R-squared:  0.01476,   Adjusted R-squared:  -0.2316
F-statistic: 0.05991 on 1 and 4 DF,  p-value: 0.8187
```

Figure 12

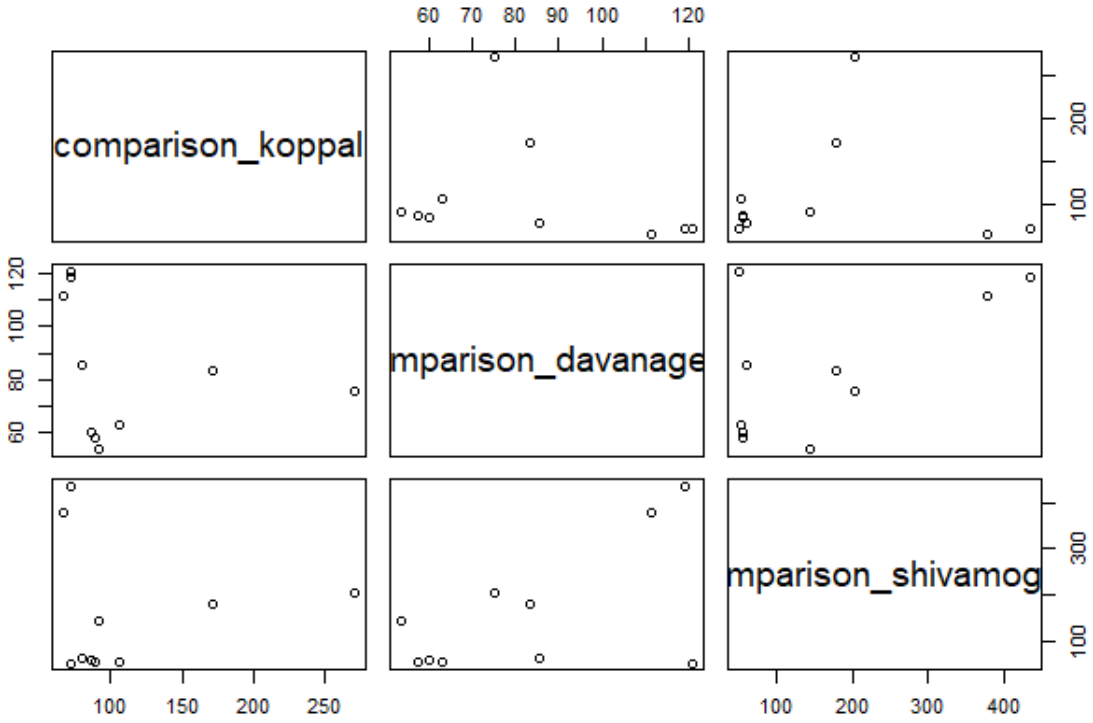


Figure 13

As can be seen in the figure above, the city of Davanagere seems to have a significant correlation with Shivamogga, whereas Koppal seems to not have much correlation with Shivamogga. The linear model is then used to put numeric values to the assumptions.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -100.888    136.618  -0.738   0.4813
comparison_davanagere    3.162      1.579   2.002   0.0803 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 122.6 on 8 degrees of freedom
Multiple R-squared:  0.3338,    Adjusted R-squared:  0.2505
F-statistic: 4.008 on 1 and 8 DF,  p-value: 0.08027
```

Figure 14

As can be seen in the above figure, there is a significant correlation between Shivamogga and Davanagere, even more so than the test on the Dublin data.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       93.99144   17.58627   5.345  0.00069 ***
comparison_koppal -0.09977    0.13944  -0.716  0.49464
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.6 on 8 degrees of freedom
Multiple R-squared:  0.06015,   Adjusted R-squared:  -0.05733
F-statistic: 0.512 on 1 and 8 DF,  p-value: 0.4946
```

Figure 15

As can be seen in the above figure, there is no significant difference between the two cities. The P – value is two high for there to be any meaningful correlation.

6.0 Conclusions

From the scatterplot and the linear regression results we can reject the null hypothesis in favour of the alternative that there is a significant correlation between Dublin city and south Dublin pollution rates. The second figure represents an area that is not within the wind pattern direction as can be seen by the latitude and longitude values. This further backs the hypothesis that wind has a significant impact on spreading air pollution.

To prove that the results were not by coincidence, the same test was done on three cities in India, and the results followed the same pattern. The wind pattens from the North East travels past Davanagere into Shivamogga and the data from the linear regression gives a confidence rate of 91% correlation. The second linear regression was done on Koppa and Davanagere. Koppal as can be seen by the coordinates is north east of Davanagere, but it is not east enough so that the wind patterns travel in line with Davanagere.

From the results shown I can say with a high degree of confidence that wind patters have a very high correlation with carrying air pollution.

7.0 Further Development or Research

With additional time and resources, I would like to make a database that contains the current pollution rate of the major cities of each country and where that pollution is traveling to in reference to the wind speed and wind patterns.

I would think that in the case of land locked countries, governments would like to know where and how pollution is entering their country. Especially countries that are already having a positive impact on a sustainable eco friendly future such as Sweden.

There is a lack of restrictions to how some countries are negatively effecting the climate, If this information is made more public, it might be the push some countries need to impact pollution rates further.

8.0 References

Jerrett, M., Burnett, R.T., Ma, R., Pope, C.A., Krewski, D., Newbold, K.B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E.E., Thun, M.J., 2005. Spatial Analysis of Air Pollution and Mortality in Los Angeles: Epidemiology 16, 727–736. https://doi.org/10.1097/01.ede.0000181630.15826.7d

Milman, O., 2021. New York air quality among worst in world as haze from western wildfires shrouds city. The Guardian.

OpenAQ [WWW Document], n.d. . OpenAQ. URL https://openaq.org/ (accessed 12.22.21).

Stern, A.C., 1977. Air Pollution: The Effects of Air Pollution. Elsevier.

Zhang, Z., Arshad, A., Zhang, C., Hussain, S., Li, W., 2020. Unprecedented Temporary Reduction in Global Air Pollution Associated with COVID-19 Forced Confinement: A Continental and City Scale Analysis. Remote Sens. 12, 2420. https://doi.org/10.3390/rs12152420

9.0 Appendices

9.1. Project Proposal

National College of Ireland

Project Proposal

An exploratory data analysis on air pollution

BSHCEDA4

Data Analytics

2021/2022

Sean Burke

X17132118

X17132118@student.ncirl.ie

Contents

Objectives

The objective of this project is to apply what I have learned in my data analysis modules to a topic that Interests me. I have chosen the topic of air pollution.

I plan to gather data from an API from https://docs.openaq.org , This API holds the world air quality data that also contains many datasets within countries so I will be able to provide a more refined overview within each country.

I'm also planning to source another dataset via web scraping. Currently the data set I'm looking to use is the most polluted countries 2020 https://www.iqair.com/world-most-polluted-countries .

My aim is to be able to provide values and graphs that report my findings comparing different countries and cities. Find out the "Why" on the high or low values of air pollution in different parts of the world.

Background

I chose to undertake this project because, air pollution has become an increasing problem over the years and has many negative impacts.

I plan to use programming languages like R and Python to take in the data sets that I have chosen (either via reading in from csv files, accessing an API, or Web scraping). From there I will clean the data to show that specific data that I plan to use (not every part of the dataset will be used). Once I have my cleaned the data, I can compare the data against other cleaned data with the aid of graphs and then give a detailed report on my findings

State of the Art

The first academic Article looked at was: "The Spatial analysis of air pollution and mortality in Los Angeles" (Jerrett et al., 2005)this study was done to determine how a high particle mass 2.5 could result in a higher risk of people in that area getting ischemic heart disease of lung cancer.

The second academic article I looked at was: "Air Pollution: The effects of air pollution" (Stern, 1977), this article shows the different negative effects that air pollution causes, e.g. vegetation, Indoor air quality, Ozone etc.

The third academic article I looked at was : "Unprecedented Temporary Reduction in Global Air Pollution Associated with COVID-19 forced confinement: A Continental and City Scale Analysis" (Zhang et al., 2020), This study shows how a global lockdown shown huge signs of improvement in air quality such as a reduction of 49% for NO2 emission in London. Showing how changes for a relatively small amount of time can make a huge difference.

The premise of my project was to get data from all over the world, not just the countries, but the cities in those countries. I want to find out the reasons why certain countries are so heavily polluted. Could there be small changes made to apply a big impact or are these small changes even feasible in these countries (for example poverty, culture etc.)

This differs from the articles I have provided because they are more based around the premise of the effects of air pollution and a report on how the COVID-19 pandemic has forced countries into better air quality.

Data

The data I will need for this project is the fine particle matter values (usually pm2.5, pm10) of each of the countries/ cities I want to gather data from. I'm going to gather this data by downloading the .csv files and using R to get comfortable with the data as I'm quite comfortable with using R for projects, once I know the data is usable in R, I would like to attempt using the api to gather the data. I would preferably use python for most of this project as I would like to use this project as a learning experience to learn a new language like python. This data will be used for the current year dataset.

The second dataset needed will come from the iqair website, specifically the table data that shows the most polluted countries from 2020, and this data will be obtained via web scraping. Web scraping is the process of extracting data from a website, the data is collected and exported into a chosen format, an example of this is table data.

Methodology & Analysis

I plan to us an Agile approach to this project. The agile methodology provides guidance on how to choose methods and procedures within the project.

I plan to use sprint cycles starting from December to make sure that the project stays on track, and If I'm falling behind or the project needs to be adjusted, it can be updated for the next sprint cycle.

At the start of each month, I will assess what stage I'm currently on in the sprint cycle and then adjust specific topics that I had aimed to be done at certain dates.

Technical Details

On the technical side of this project, the technology I will be using are:

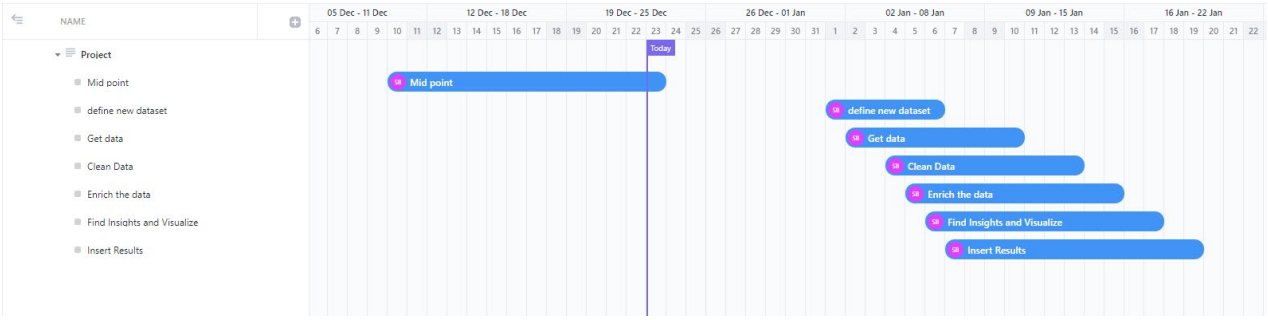
- Python3
- R
- RStudio
- Spyder
- Microsoft Excel
- Microsoft Word

Both R and Python are very useful when it comes to data analysis, they provide the user with the necessary tools packages to take in data in its raw form and manipulate it into a data frame and represent that data using graphical aided plots. Considering some .csv files could have hundreds of thousands of rows of data, an analyst needs to be able to show how that data can be represented

Formulas that I will use in this project are the pm10 to pm 2.5 conversion formula

$$pm2.5 = 0.75 \times pm10$$

Project Plan



The plan as it stands is to us a sprint cycle-based system to complete each section over the course of 2 weeks, every two weeks I will be reviewing the work I have completed and the results I have found. Upon reviewing those results, I will decide on what is to be done next cycle. The cycle will be updated in January when I next have a discussion with my project supervisor.

9.2. Reflective Journals

Month: October

What?

Reflect on what has happened in your project this month?

This month I had to think of an idea on what I wanted to do my project on. The Project idea I decided to go with was an exploratory data analysis on air pollution across different countries / cities across the world, from 2019 to present day data. I have three different data sets that I can use so far.

I recorded my project pitch video on my plans for the project. This video is then going to be viewed by a project supervisor that has been assigned to me. Currently I have not received any information regarding the status of my project pitch.

We talked about how to approach the project pitch in the Software Project class's and were also given access to past student's projects and seen how they approached them.

So What?

Consider what that meant for your project progress. What were your successes? What challenges remain?

Successes:

- Project pitch.
- Good understanding of how to write the project proposal.

Challenges remain:

- Receive feedback from my supervisor.
- Once I've received my feedback, I might be asked to change certain parts.
- I could be asked to do a completely different project if my project proposal doesn't meet the requirements for a valid project.
- Write the project proposal

Now What?

What can you do to address outstanding challenges?

- I have already sent my project supervisor an email to plan out a suitable date to talk about the project.
- I've written down some potential amendments that I could make to the project if It's not complex enough.
- If asked to do a different project, I am more than happy to do so, although I would like to do a project based on something I chose, I would prefer to do a project that matched the criteria needed for a successful project and a good grade.
- Using the template provided, and with some guidance from my project supervisor I'll be able to write a project proposal that accurately defines how I want to approach my project.

Month: November

What?

Reflect on what has happened in your project this month?

The Project has been pitched and accepted

This month I was in talks with my supervisor about what must be done for the midpoint presentation and where would be a good point to start.

CA1 was also given out by the other modules during this month

So What?

Consider what that meant for your project progress. What were your successes? What challenges remain?

Successes:

- Keep on top of work life and college assignments
- Maintaining a solid work schedule for the project so far

Challenges remain:

- Midpoint presentation

- Elaborate more on what I'm trying to achieve in my project and how I aim to achieve this.

Now What?

What can you do to address outstanding challenges?

- Follow the rubric and the template provided by the college
- I need to follow up on some ideas that I've been hoping to integrate into the project and see how feasible they are.

Month: December

What?

Reflect on what has happened in your project this month?

This month I submitted my midpoint presentation on the project so far. I covered:

- A completed project proposal.
- An in-depth overview of the data that I will be using.
- The Methodology behind using the data.
- The analysis that will be performed on the pre-processed data.
- The use of equations and visual aids to show the analysis.
- Making a slide show for my midpoint presentation.
- Presenting the midpoint presentation in 10 minutes.

So What?

Consider what that meant for your project progress. What were your successes? What challenges remain?

So far, I have completed the mid-point presentation and showcased what work I've done so far. The plots that I provided accurately show the expected results when compared against other summarised data from another dataset. Proving that both the data and the equation used are accurate enough to report my findings. I'm happy with the progress so far as there was a lot of time pressure between work and other modules.

The challenges I would like to tackle next are:

- Get a more refined Gannt chart on what the time schedule should look like between now and the project turn in point. Have goals set out for each month and fall backs in case those goals cannot be completed.
- Focus on countries that have poor air quality, specifically the countries that boarder each other so that I can perform a better analysis on them by researching the issues that may cause the air pollution.

  I feel like this is enough work to get me through to the end of January considering exams and assignments.

Now What?

What can you do to address outstanding challenges?

Currently, my plan is to complete the assignments and exams for my other modules, and once this is done, I'll contact my project supervisor about setting up a meeting to talk about my midpoint presentation.

Research landlocked countries that border other countries with similar high air pollution rates. Get the data from these countries and perform an analysis as before on this data. Find academic papers that explain the high pollution values.

Month: January

What?

In the past month I had TABA/ exams for three classes from the first semester. This was followed by a two week break between the first and second semester. In these two weeks I was working full time in my job and continued my learning on code academy courses to brush up on my R and Python knowledge. I received a message from my project supervisor regarding a 1 to 1 session and it was planned for the end of January when classes were back in session.

I received a deadline for project profiles and Bio's information which is due by February 14th.

So What?

I made good progress on my code academy courses and learned many techniques that I will use in my project. I also had a meeting with my boss to look for part time, so I could concentrate more on my studies. From this the start of February I will only be working 3 days a week so that I have time to concentrate on my project and be more available to meet with my project supervisor. I did not make much progress on the challenges I set out for myself since the last reflection report, which is the reason I am looking for more study time.

I will need between 1 and 3 non copyright images. I will also need a current portrait photo of myself for this bio. The information itself will not take too long to fill out, I know a photographer that would be able to take the pictures I'll need for the project, but I'll have to crop them to the appropriate size.

Now What?

I have set out time every Tuesday, Thursday, and Saturday in which I will be using just for my college work. I am going to go more in depth by making a Gantt chart, not only for my final year project, but for all my other modules too.

I will continue to work on my online courses, but I will be more focusing on getting the project profile and Bio for the next two weeks as It is very important to me that I showcase the project well.

I'll also get in contact with Keith (project Supervisor) about my new work schedule and let him know of the days that I'll be available for meetings, as I could not make the last one due to not being home in time from work.

Month: February

What?

In the past month I had two meetings with my project supervisor. The first meeting we discussed the grade that I received from the midpoint presentation. The topics we discussed were:

- Percentages from the different sections of the grading rubric of the midpoint presentation.
- Where I need to improve going forward.

- Comments on what is deemed important and left out of the data during pre-processing.

The second meeting we discussed that wind patterns should be included into the data, as the dispersion of air pollution can be dependent on wind patterns, especially in highly polluted areas. I researched academic papers, scientific articles and blogs that shown the effects that wind has over air particles such ass the pm10 pollution particles.

I have set aside time each week that I will be able to commit to college work.

So What?

- I need to incorporate the research that was made on wind pattern effects on air pollution to my analysis, one of my initial points was to see if neighbouring countries are affected by the pollution of other highly polluted countries. Using the latitude and longitude that was removed from the midpoint data along with the standard wind patterns I will attempt to prove that there is a noticeable difference in air pollution from the pollution expected to be in the city, rather than the current levels of pollution.
- I need to go into more detail on each of the sections of where I lost marks on the midpoint presentation.
- Provide weekly updates on my findings and what I'm doing with this project with the project supervisor.

Now What?

- I will research a highly polluted city, and using the coordinates from my dataset I can pin point a reference pollution level at a specific date. From here I will find out common weather patterns along that latitude and longitude where I can pick my new points of reference and see what the pollution levels are like on these points.
- A lot of the midpoint presentation can be improved upon. There are many sections that are too short/ not gone into enough detail in to receive the marks that I'm looking for.
- Keep in contact with my project supervisor about what I'm planning to do, gather his thoughts on the matter, this will be done either via Teams meetings or email.

Month: March

What?

In the past month I fell into many unfortunate circumstances. In the last week of February, I was informed that many of the staff in my workplace received covid and I was no longer able to take the time off work that I needed. I ended up having to work full time on site to make up for the people missing for the next two weeks. At the end of these two weeks, I started to feel unwell, two positive antigen tests and eventually a PCR had shown that I had covid. My symptoms were more on the severe side.

- Pains in my chest
- Difficulty breathing
- Very bad headaches

I still to this day have some of these symptoms. I was in contact with the college about my situation and was able to get an extension on my assignments.

So What?

- Over the next month I have a lot of catching up to do regarding the assignments and project work. I have two assignments until the 14th of April which are worth 50%
- Regarding the project work, I will need to contact my project supervisor to discuss a realistic approach to finishing the workload in time. We had fallen out of contact which was my fault due to the circumstances.

Now What?

- I will be focusing on my two assignments over the next two weeks.
- I will have to contact my project supervisor about where I am currently in the project and what needs to be done.
- I will need to reassess the Gannt chart to adjust the timelines.

Month: April

What?

In the past month I had submitted 4/4 of my semester assignments for the Advanced Business Data Analysis and Data and Web Mining. Due to the extensions, I received these due dates spanned between the 8th of April to the 20th. This resulted in downtime on doing project work that I had set out but was able to resume progress on the 21st.

I was able to:

- Use python to import the data through an API
- Pre-process the data in python
- I added the wind variable, as it was proven to be vital data when drawing a conclusion on how it will affect the spread of pollution.

So What?

- Adding Python to the list of languages, and utilizing the API made the project more complex, and shows the use of multiple languages.
- The wind variable was researched through numerous academic journals and how it effects small particle matters, this was vital in obtaining the knowledge for conclusions
- I gained success in complexity and valuable information derived from the academic journals.

Now What?

The due date is 15th of May, I must make sure that I have:

- Fulfilled all the requirements on the grading rubric.
- Have a detailed report of my findings and explain my methods and results clearly
- Add more complexity to the analysis, using referenced journals state how conclusions were made from the results