DEATHS CAUSED BY AIR POLLUTION

DATA201 Report
2023

Report By:

Sara Mann, Josh Lowe, Matthew Hock

02

PHOTO CREDIT:
MARKUS SPISKE

INTRODUCTION

Air pollution is a worldwide hazard posing risks to human health and the environment. The World Health Organization (WHO) reports that annually, nearly seven million deaths globally are attributed to air pollution. Everyday polluted air is being inhaled by 91% of the global population. We can make the choice to either reduce it or increase it.

QUOTE

Environmental pollution is an incurable disease. It can only be prevented.

- Barry Commoner

03

DATA SOURCES WE USED

The data collected was obtained from the World Health Organization (WHO) website, OpenAQ, The World Bank and Our World in Data. We gathered a few different data sets each analyzing the following:

2019 Ambient air pollution attributable death rate (per 100,000 population)– WHO

- *This data set looks at the death rates by multiple health conditions (diseases) caused by air pollution in different countries. Gender comes in as a specific factor here.*

2019 Ambient air pollution attributable death rate (per 100,000 population, age-standardized)– WHO

- *This data set looks at deaths that are from all causes attributed to ambient outdoor air pollution in different age groups.*

World Air Quality – OpenAQ

- *This data set showcases different air pollutants (CO, PM10, PM2.5, SO2, NO2, O3) present in the air. Our primary focus was PM2.5 as we discovered it to be one of the deadliest pollutants.*

PM2.5 Air Pollution (mean annual exposure) – The World Bank

- *We needed more data on only PM2.5 so we collected this data set which shows the mean annual exposure of PM2.5 air pollutant in different countries.*

Life Expectancy VS Healthcare Expenditure - Our World in Data

- *We wanted to see if healthcare expenditure had an effect on air pollution death rates so we collected this data source.*

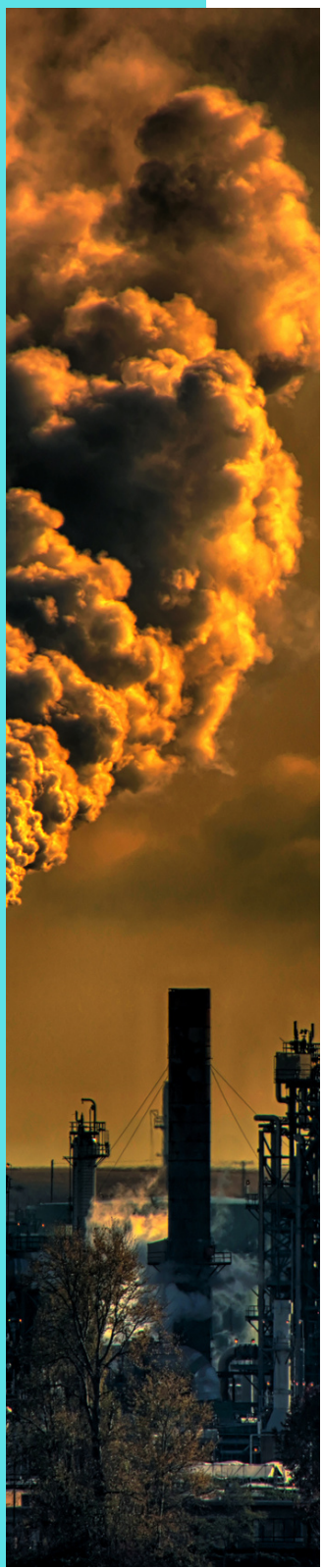
World Health Organization

openaq

Our World in Data

04

WHY WE CHOSE THOSE DATA SOURCES

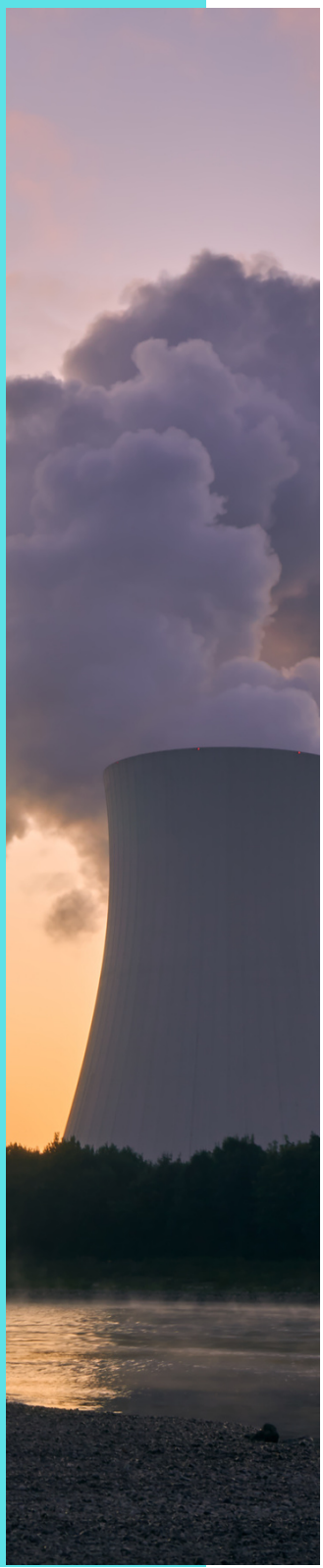


6.7 million total deaths linked to outdoor and indoor air pollution - State of Global Air Report, 2019

As the population continues to increase year by year, so does air pollution. The effect it has on the human body and environment is hazardous and WHO elaborates on the extent of that. According to WHO, at least 1 in 10 people die from air-pollution related diseases and poses a greater threat to life expectancy than smoking, HIV or War. The Ministry of Environment NZ states that “An average person inhales around 14,000 litres of air everyday,” hence if air quality is poor, it matters significantly. These are alarming facts that gained our attention in researching the topic more. Therefore, the data sources selected were based on these distressing facts.

05

THE TARGET WE CHOSE



The overall intent of our project was to analyse deaths caused by air pollution and what factors play a part in this. Since even this 'narrow' focus is so broad, we came up with specific questions that we wanted to answer with our wrangling.

The questions:

- Does gender affect deaths caused by pollution?
- What is the biggest air pollution-related disease?
- Does age affect deaths caused by pollution?
- Does higher concentrations of PM2.5 lead to a higher death count (rate) in the different countries/continents?
- Does health expenditure in different countries/continents affect death counts (rate) caused by pollution?

The datasets we sourced from WHO, OpenAq, The World Bank and Our World in Data each gave us the relevant and necessary information to answer the questions.

> *WHO reports:*
>
> - *1 in 10 people die from air pollution-related disease.*
> - *9 out of 10 people living in places with polluted air.*

06

THE OVERALL PROCESS

FAST PROTOTYPING

Finding suitable topic with data. Trello to keep track.

...Messy Data...

DATA WRANGLING

The different tools:
RMD, Github, Jupyter

RELATIONAL DATA MODELLING

Creating combined data frames.

...Tidy Data...

DATA VISUALISATION

Plots - bar charts, pie graphs, scatter diagrams
Maps - Static + Interactive

RESULTS/CONCLUSION

What did our results tell us? What did we find? Any patterns/trends?

07

TECHNIQUES WE USED

Throughout our wrangling process, we used tidyverse functions as they share a common design philosophy that results in reproducible code and output.
For the WHO data, we were able to smoothly load the csv file using the read_csv function which was a great start. During the data wrangling process the main problems that needed to be dealt with were:

1. The column names needed to be fixed so they represented what data they hold in their column.
2. Deleting and re-ordering columns (removing NAs)
3. Changing the column class types so they match what data they hold (e.g. numeric).
4. Transforming the data from a long format to a wide format.

To solve problem 1, we set all the column names = FALSE, since most of them needed to be changed. It was then a case of using the rename() function for each column.

To solve problem 2, we first had to identify all the 'unnecessary' columns that were irrelevant with our intent of the data. Lucky for us, the columns that were irrelevant had all the NA values. Then we used, select() function to do the job. Re-ordering was done using the relocate() function.

Following on from problem 1 and 2, came problem 3: the transformed columns were not reflecting the correct class type. Via the sapply() and as.numeric() functions, we were able to fix this with no major problems.

With each cell containing one value and each row containing one variable, we had a tidy data set. In terms of interpretability and readability, we wanted to transform the dataset into a wide format. Using the separate() and pivot_wide() functions we were able to achieve this.

08

TECHNIQUES WE USED

The PM2.5 data was opened in R using the read_excel function. This dataset had multiple pollutant types and multiple observations for different cities per country. So, the wrangling process involved tidying and condensing the data into two columns: country, and PM2.5 value.

Preparation: Firstly, the data column needed to be converted into proper date values, using mutate and the as.Date function. Negative pollution data values needed to be filtered out too, as an air pollution measurement cannot be negative. A new column called 'year' was created, which took the year value from the date column. After this, pollution values in ppm (parts per million units) were removed, by filtering to only keep PM2.5 in $\mu g/m^3$, as it is difficult and unreliable to convert $\mu g/m^3$ into ppm.

Condensing: Because there were multiple data observations per country, we grouped the data by the Country label, using the functions mean, pivot_wider, and group_by. This resulted in one individual row per country. The PM2.5 values were averaged, so the cell value of a country's PM2.5 value is the average of multiple observations. Next, PM2.5 values for the year 2019 were selected, and `Country Label` column was renamed to 'Country', for general tidiness.

The Problem: By the time this large dataset (40,000 + rows) had been wrangled to our specific requirements, the were far too many NA values for this data to be useful.

The solution: We found another dataset based on PM2.5 values for countries of the world, from the World Bank website. After, selecting complete rows, filtering to 2019 values, and renaming many country string names to match the other dataset, these two data sets were merged using a left_join() by the country names. This is can be considered a positive outcome, as there are now more PM2.5 pollution observations averaged into the country values of our data, potentially increasing accuracy, and enhancing our investigation.

09

RELATIONAL DATA MODEL

The relational data model is built based on the fact that all of our datasets contain country names. This can be used to relate them. A table was established to hold all the unique countries that are in the other datasets. This “Country Table” has a primary key called “CountryID” which can be used as a foreign key in other tables.

The other tables in the data model all have primary keys which are used as unique identifiers. They also all have the foreign key of “CountryID” which connects them all together. When they need to be used for graphing they are temporarily joined based on “CountryID”.
For instance:

```
temp_air_pollution <- left_join(country_table2, air_pollution_df, by = "CountryID") %>%
  select(-CountryID) %>% select(-MeasurementID)
```

Country	Continent	Year	Cause	Female	Both sexes	Male
Afghanistan	Eastern Mediterranean	2019	Trachea, bronchus, lung cancers	0.810	1.530	2.320
Afghanistan	Eastern Mediterranean	2019	Ischaemic heart disease	80.330	83.950	87.270
Afghanistan	Eastern Mediterranean	2019	Stroke	35.630	31.670	27.240
Afghanistan	Eastern Mediterranean	2019	Total	139.800	142.600	145.000
Afghanistan	Eastern Mediterranean	2019	Chronic obstructive pulmonary disease	8.830	9.470	10.200

OVERVIEW OF DATA MODEL (PRIMARY KEYS UNDERLINED)

PM2.5 Table: PM2.5 Values, CountryID, ValueID

Country Table: Country, CountryID

Life Health Table: Life Expectancy, Health Expenditure Per Capita, Population, CountryID, ValueID

Air Pollution Deaths by Age: Deaths from Pollution (Under 5 years), Deaths from Pollution (5-14 years), Deaths from Pollution (15-49 years), Deaths from Pollution (50-69 years), Deaths from Pollution (70+ years), Total_Deaths, CountryID, Value ID

Air Pollution by Disease: Continent, Year, Cause, Female, Both sexes, Male, CountryID, MeasurementID

10

RELATIONAL DATA MODEL TABLES

COUNTRY TABLE

CountryID	Country
1	Afghanistan
2	Africa Eastern and Southern
3	Africa Western and Central
4	Albania
5	Algeria

LIFE HEALTH TABLE

Live Expectancy	Health Expenditure Per Capita	Population	CountryID	ValueID
63.56500	285.5581	37769496	1	1
75.43900	1616.1779	2820604	12	2
73.10200	605.9345	10232761	15	3
80.01900	1880.5751	1494195	17	4
72.80600	123.2870	165516224	18	5

AIR POLLUTION DEATHS BY AGE

Deaths from Pollution (Under 5 years)	Deaths from Pollution (5-14 years)	Deaths from Pollution (15-49 years)	Deaths from Pollution (50-69 years)	Deaths from Pollution (70+ years)	Total_Deaths	CountryID	ValueID
36.7446820	0.858006500	8.883144	104.00544	424.9321	575.4234	1	1
15.9478650	0.362357700	9.737727	149.92294	823.9526	999.9235	12	2
42.9591640	0.984215100	10.660836	171.38441	1036.6786	1262.6672	15	3
5.7574315	0.174419750	9.417181	111.66634	1033.5146	1160.5300	17	4
38.8547820	1.057153500	7.319665	115.47842	524.2246	686.9346	18	5

PM2.5 TABLE

2019	CountryID	ValueID
26.709	1	1
30.879	2	2
63.357	3	3
18.639	4	4
16.917	5	5

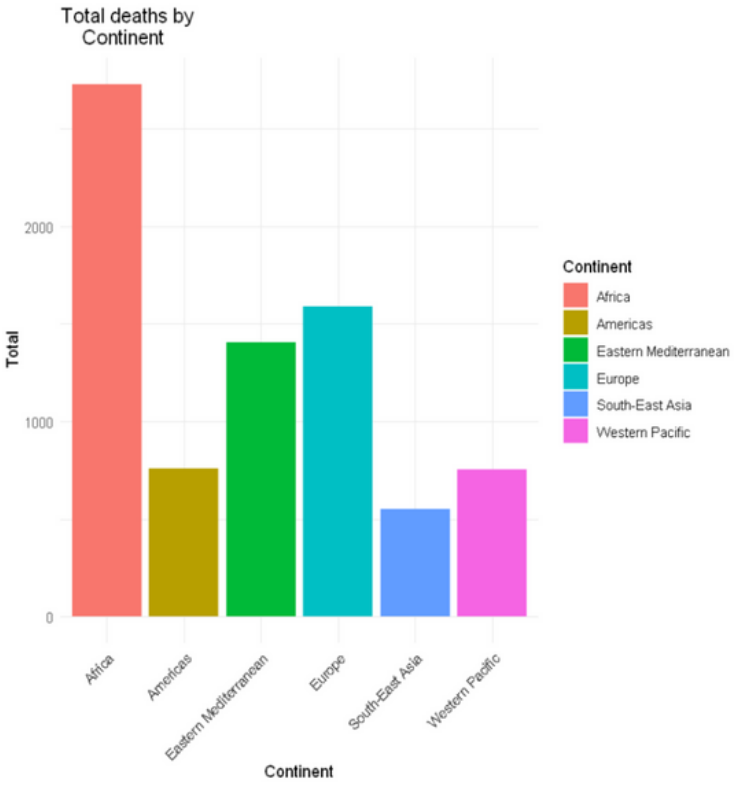
AIR POLLUTION DEATHS BY DISEASE

Continent	Year	Cause	Female	Both sexes	Male	CountryID	MeasurementID
Eastern Mediterranean	2019	Trachea, bronchus, lung cancers	0.810	1.530	2.320	1	1
Eastern Mediterranean	2019	Ischaemic heart disease	80.330	83.950	87.270	1	2
Eastern Mediterranean	2019	Stroke	35.630	31.670	27.240	1	3
Eastern Mediterranean	2019	Total	139.800	142.600	145.000	1	4
Eastern Mediterranean	2019	Chronic obstructive pulmonary disease	8.830	9.470	10.200	1	5

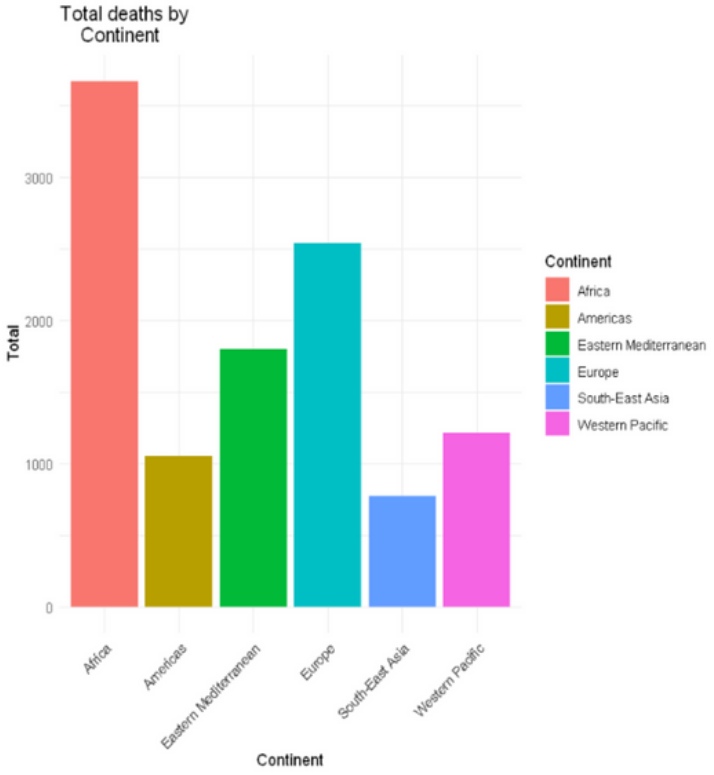
11

DATA VISUALISATION

VISUALISING MALE VS FEMALE DEATH COUNT DUE TO DIFFERENT AIR POLLUTION-RELATED DISEASES



Female Data - Death Count from each continent



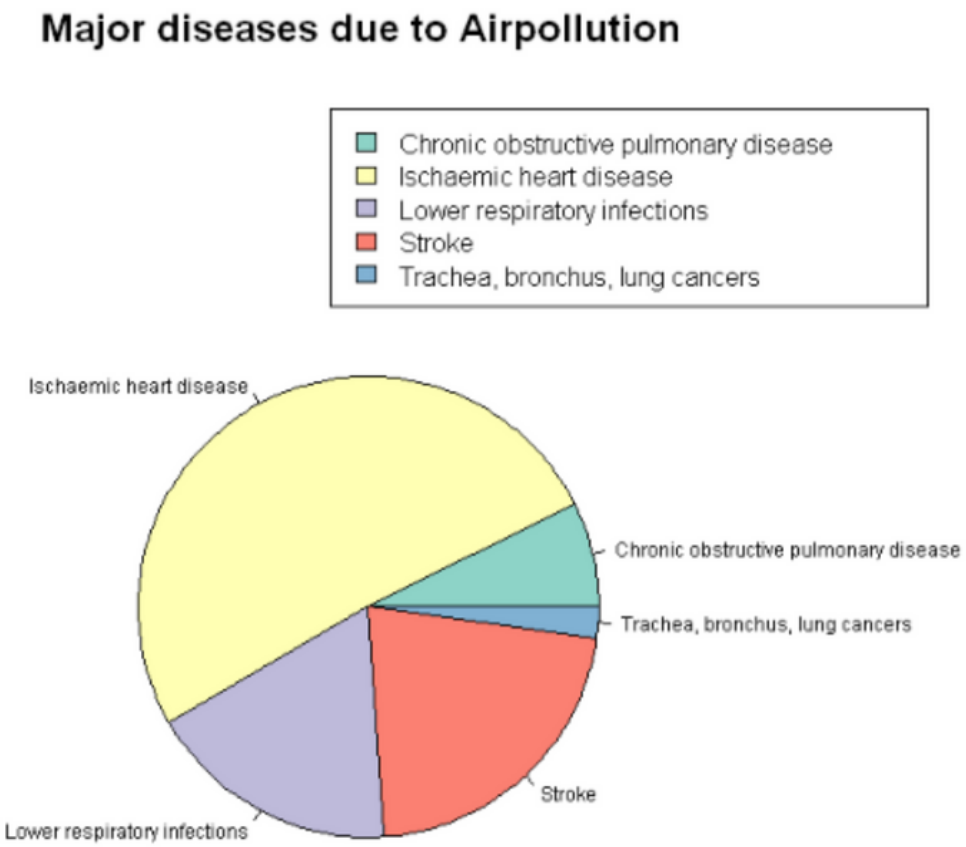
Male Data - Death Count from each continent

The bar graphs indicate that males tend to have higher deaths due to air pollution-related diseases (2019) looking at the y-axis. Africa has the highest death count followed by Europe and Easter Mediterranean. The question is can we justify this result? After further research, the conclusion was indeed valid. Males tend to have higher death counts than females due to more employment in hazardous occupations (construction - outdoor with bad air quality, machinery etc). Notice that the data is based off heart-related diseases. Males due to their risk-taking nature, have a higher chance of getting heart-related diseases.

Having Europe and Eastern Mediterranean with high death counts was surprising. But again, we're looking at heart-related diseases where drinking and smoking are the worst in these regions (statistically proven).

12

VISUALISING WHAT THE HIGHEST AIR POLLUTION-RELATED DISEASE IS

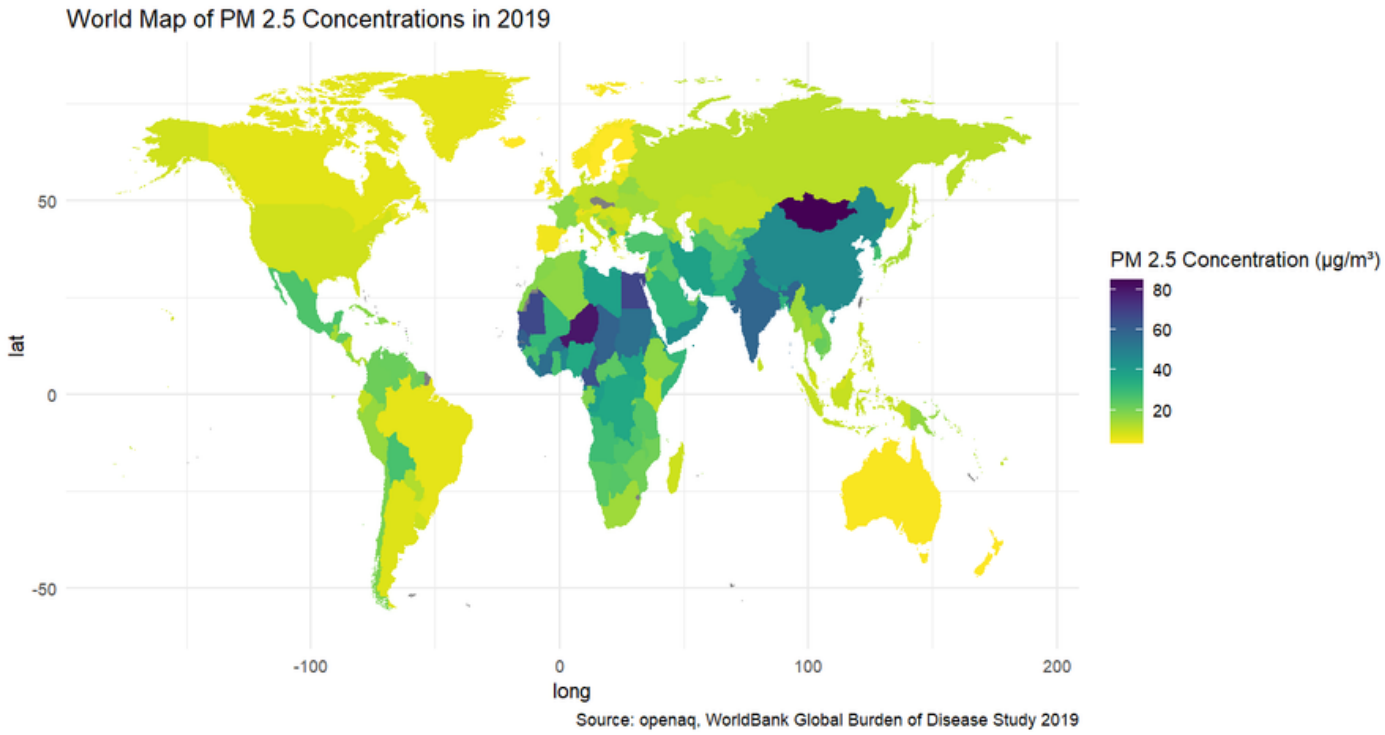


Female and Male data death count

Here we can see an overview of the diseases caused by air pollution in 2019 and the extent of them. We found that Ischaemic heart disease caused the most deaths followed by Stroke, Lower respiratory infections, Chronic obstructive pulmonary disease and finally Trachea, bronchus lung cancers.

13

VISUALISING GLOBAL AIR POLLUTION

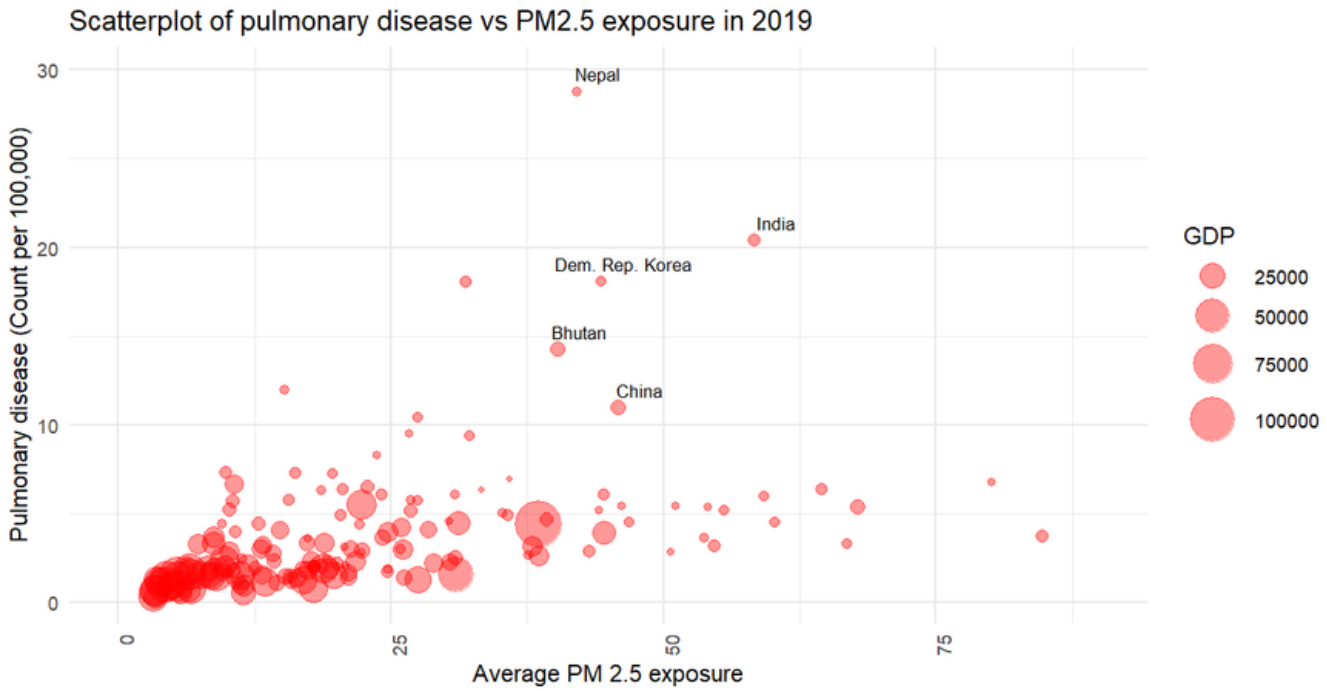


World map of PM2.5 values, made with the ggplot R package

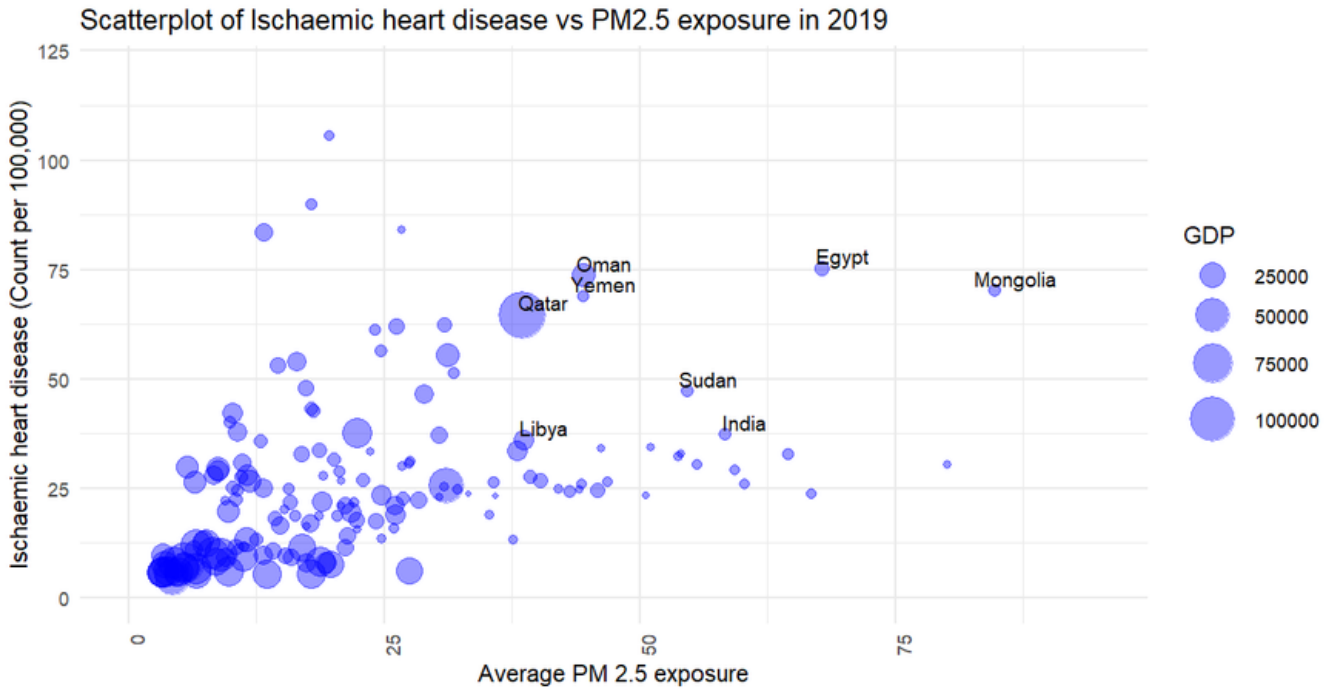
This world map of PM2.5 air pollution, generated using the ggplot package in RStudio, illustrates the disparities in air quality between countries. The countries are coloured based off a colour gradient ranging from yellow to dark blue/purple, representing the severity of PM2.5 concentration. Countries like Canada, Norway, Sweden, Brazil, Argentina, Australia, and New Zealand appear in shades of yellow, indicative of safe air quality. On the other hand, countries such as Niger, Egypt, India, China, and Mongolia are seen in dark blue and purple, signaling extremely high PM2.5 levels, and so hazardous air quality.

14

CORRELATION BETWEEN GLOBAL AIR POLLUTION AND HEALTH CONDITIONS



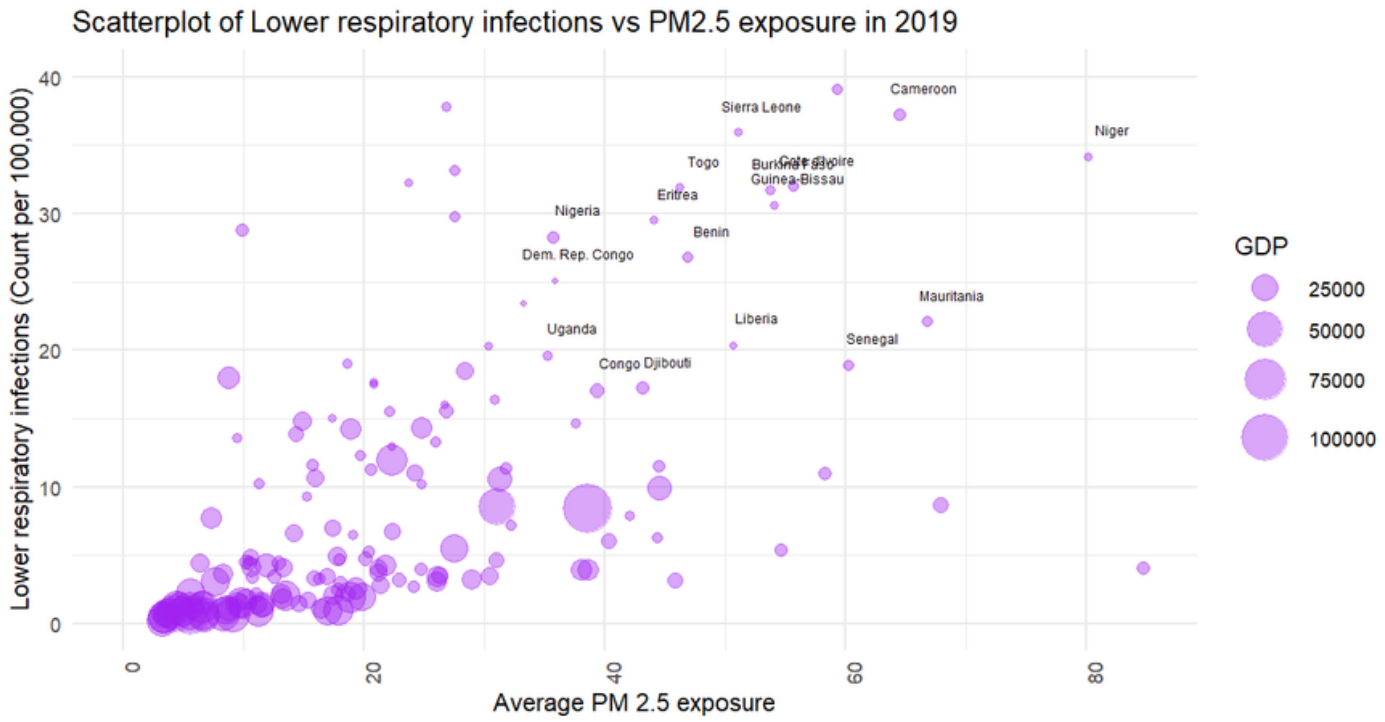
Scatterplot of chronic obstructive pulmonary disease against PM2.5 concentrations in 2019.



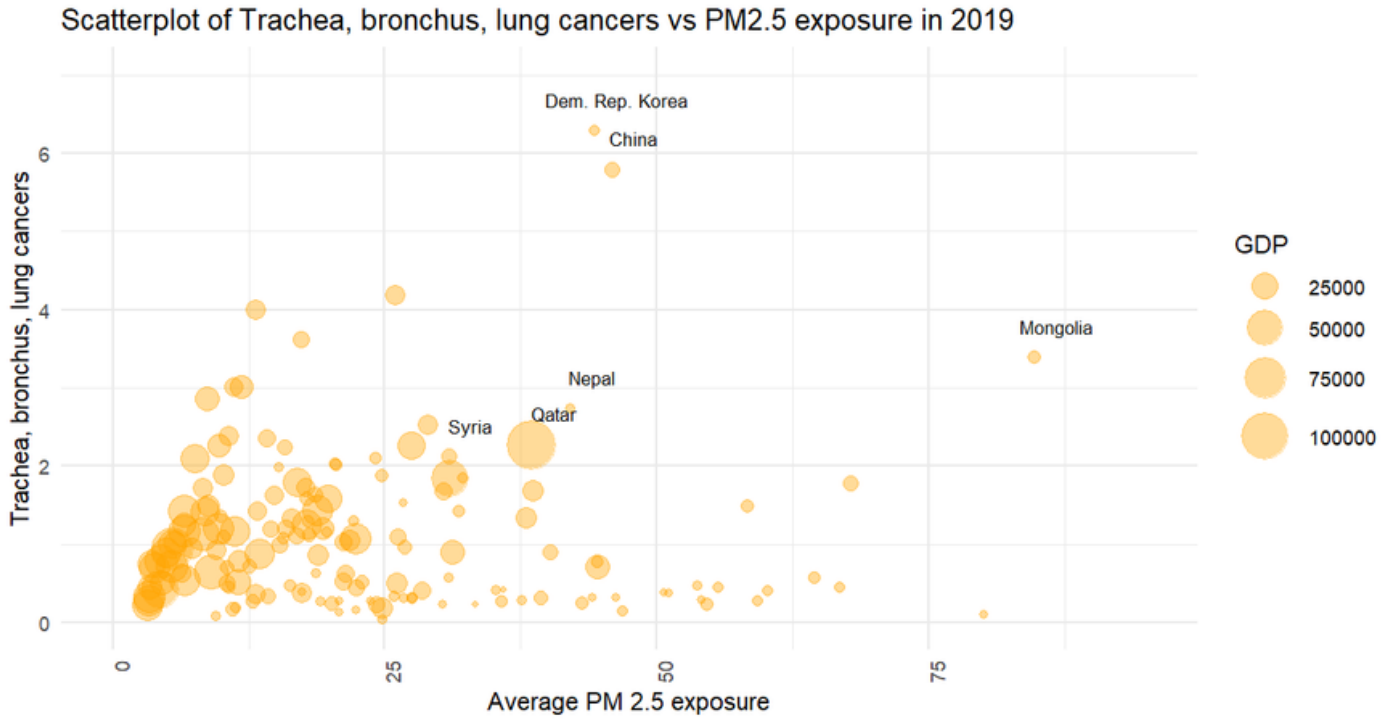
Scatterplot of ischaemic heart disease against PM2.5 concentrations in 2019.

15

CORRELATION BETWEEN GLOBAL AIR POLLUTION AND HEALTH CONDITIONS



Scatterplot of Lower Respiratory infections against PM2.5 concentrations in 2019.



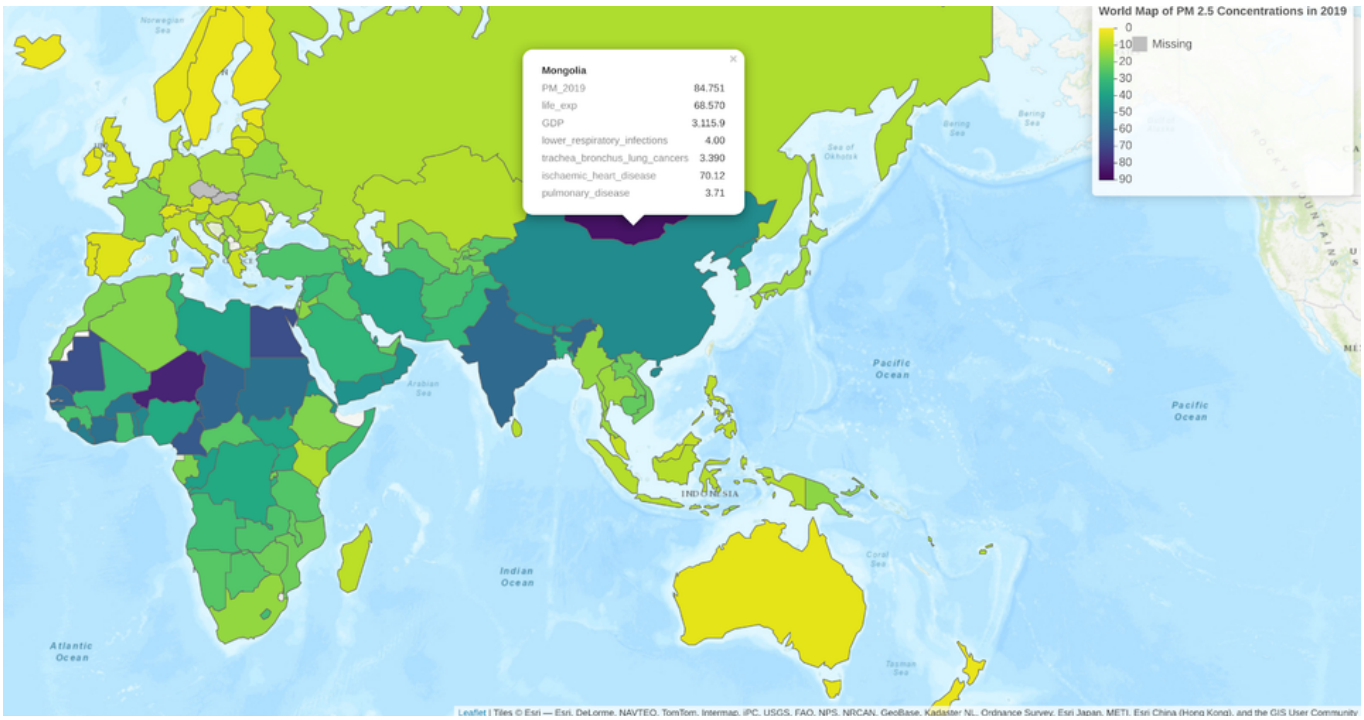
Scatterplot of trachea, bronchus, and lung cancers against PM2.5 concentrations in 2019.

16

CORRELATION BETWEEN GLOBAL AIR POLLUTION AND HEALTH CONDITIONS

These scatterplots demonstrate a correlation between PM2.5 concentrations, air quality-related illnesses, and GDP of the countries. The larger data points (indicating higher GDP) tend to cluster close to the origin, where there are lower PM2.5 concentrations, and fewer air quality-related illnesses.

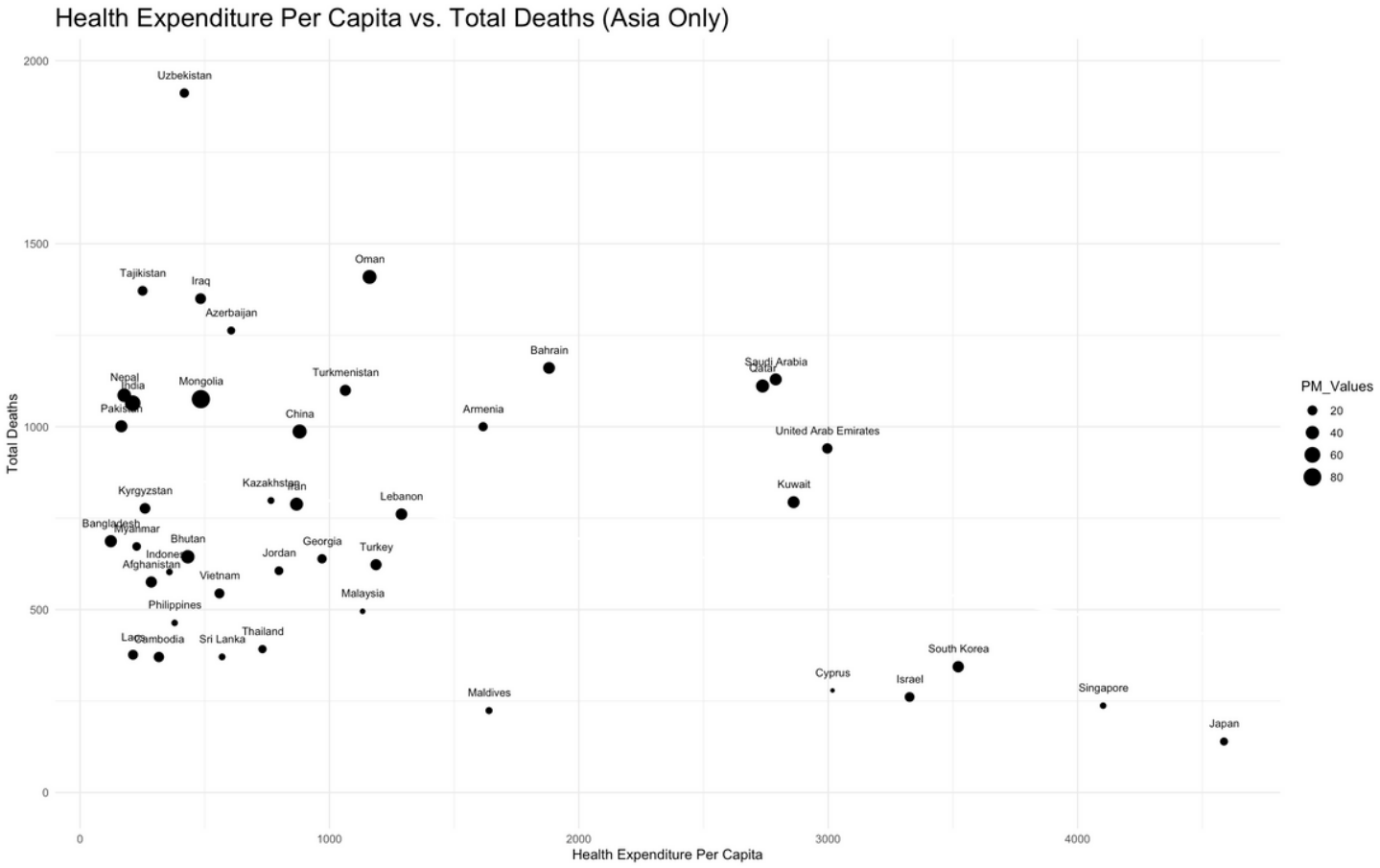
On the other hand, countries with smaller data points (indicating lower GDP) tend to be scattered outwards towards the top right corner, where PM2.5 concentrations are higher, and there are more counts of air quality-related illnesses. This reasonably consistent pattern between these plots indicate a correlation between GDP, PM2.5 concentrations, and air quality-related illnesses.



Screenshot of the interactive map, showing the pop-up information when a country is selected.

17

CORRELATION BETWEEN HEALTH EXPENDITURE AND DEATHS BY AIR POLLUTION



Scatterplot of total deaths by air pollution against health expenditure

This scatter plot demonstraights the relationship between health expenditure and deaths due to air pollution. It can be seen that as health expenditure increases, total deaths decreases. There are some countries that don't follow this trend, for instance Saudi Arabia and Qatar have large expenditures on health but still have a relatively high death rate, this is possibly because they have large amounts of PM2.5 (which can be seen by the size of the points).

18

DIFFICULTIES WE HAD TO OVERCOME TO ACHIEVE THE TARGET MODEL

- The first difficulty we faced was finding a suitable topic alongside some data that was interesting to us and easily explainable to other people. Exploring some sites, we found that the data was already clean or too complex. However, after some digging we managed to find four data sets that were appropriate and met the task at hand. We assigned each dataset to someone and any problems that cropped up, we would re-group and discuss for solutions.

- The next stage was to find a primary objective, i.e. what we want the data to show. Since the topic air pollution is so broad and heavy, this was a big challenge. We thought it would be a good idea to come up with some questions to answer. Since the first datasets we found were showing deaths caused by air pollution-related diseases we decided to delve deeper into this. We thought of the different factors (gender, age, air quality and health expenditure) and subsequently managed to come up with some solid objectives/questions. Beginning the wrangling process, we discovered some issues with the selected data sets which was our next challenge:

- Some (NOT all) of the data sets contained frequent NAs (missing data values) where simply omitting them would affect the data negatively. We re-evaluated which data sets we should continue to work with and how we can still implement a relational database. We decided to narrow the air pollutant factors to solely PM2.5 to make things simpler and completing our goals would be more straightforward. We gathered some new data (PM2.5 for multiple countries in different years and health expenditure vs life expectancy) and finally began the wrangling process.

19

DIFFICULTIES WE HAD TO OVERCOME TO ACHIEVE THE TARGET MODEL

- During the data manipulation, there were some problems with dplyr functions such as the select() function where it would be temperamental. It was a case of patience and finding alternatives to the functions.

- Despite finding appropriate data for the project, there were some inaccuracies across the data sets. In the data set displaying major air pollution-related diseases and deaths, we found that Europe and Eastern Mediterranean had a high death count. This was inconsistent with the PM2.5 and Health expenditure vs Life Expectancy data sets but the justification was that the diseases measured were heart-related diseases. Europe and Eastern Mediterranean have high drinking and smoking rates, therefore heart-related diseases/deaths are immense in number in these regions.

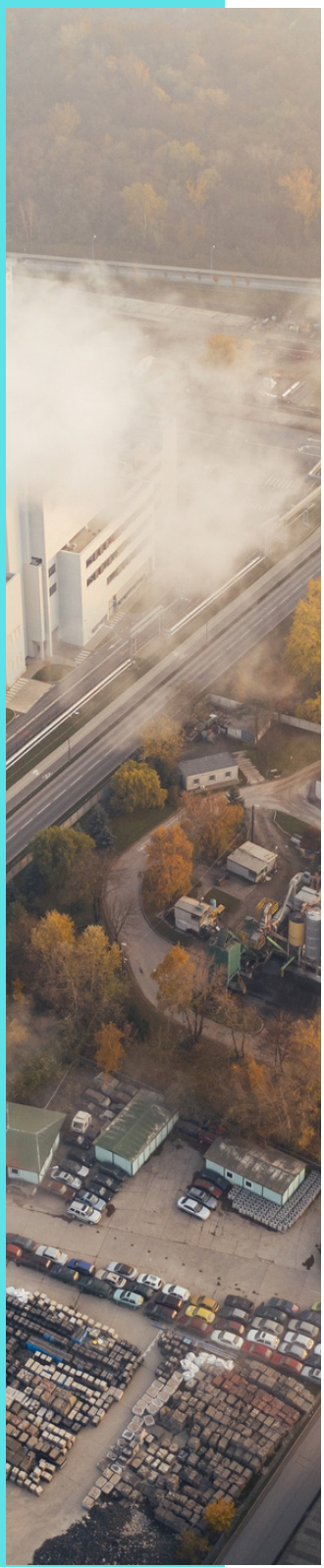
- Only one of us had used Github before, so it was a tough process to set up and understand for a couple of us. We'd often forget to use it as well, however we did partially use it by setting it up for the project.

- Since maps were made with both ggplot and tmap packages in R, these packages required different string names of countries, in order to be recognized and subsequently plotted. This involved manually adjusted the string names of countries and adjusting where needed. For instance, 'United States' needed to be renamed to 'USA'.

- Furthermore, since string names were the common element between out wrangled datasets, the country string names needed to be identical between datasets in order to construct the relational data model.

20

WHAT WE ACHIEVED AND FAILED TO DO



WHAT WE ACHIEVED

- Wrangle five messy data sets into tidy, readable and ‘easy to interpret’ data frames.
- Join the five data sets into a relational data base model using country as the common key.
- Created standard and interactive plots and maps to relay our idea effectively and answer our objective questions.
- Communicated well to have a smooth-running project.
- Learnt new concepts with Trello and GitHub (to some extent with GitHub) and used them efficiently to aid completing the project.

WHAT WE FAILED TO DO

- The GitHub process could’ve been a lot better and we failed to consistently update it with regards to code sharing and editing.
- We failed to make many variable names understandable to help make sense of our code and maintain tidiness.
- The bar and pie chart to answer if gender was a factor and what disease caused the most deaths could have had better labelling. We failed to make it obvious what the overall message was in the bar graphs specifically.

21

CONCLUSION + RESULTS

To summarise, we used a variety wrangling techniques from the tidyverse library both from the course and some prior knowledge to help answer our key objectives/questions.

Our analysis revealed that gender and age affects air pollution-related deaths. Males tend to have higher death count than females which was credible. After further research, we found that reasons could be males being employed in hazardous environments and their high risk-taking nature leads to heart-related conditions.
The most lethal disease causing the most air pollution-related deaths was found to be Ischaemic heart disease followed by Stroke and Lower respiratory infection.

Countries in southeast Asia and Central Africa tended to have significantly higher PM2.5 air pollution, and there is a correlation between this and GDP.

Generally countries that spent more on health per capita had lower deaths due to air pollution. There were some exceptions to this trend, however these countries also had high levels of PM2.5 air pollution meaning despite their high health expenditure, the air pollution levels likely led to an overwhelming amount of air pollution related deaths.

Although we faced an abundance of challenges including finding a suitable topic with sufficient data, coming up with key objectives and having issues with the wrangling process, we were able to wrangle four data sets to construct valid answers to our objectives/questions. We were also able to model and visualise the data insightfully with the purpose to answer our objectives/questions.

The entire group project assignment was extremely valuable and beneficial in reference to our learning and now we can take these skills and apply them in the future.