Exploring the impact of air quality and health factors on COVID-19 outcomes in England

Emer Buggy, Maariya Rachid Daud, Emma Horton, Punam Rattu, Rebecca Warren, Sabina Wellington

Introduction

Aims and objectives

This project aims to investigate the potential relationship between COVID-19 outcomes, air pollution, and health factors. The key questions that the project will address are:

- Do regional air quality measures play a role in COVID-19 outcomes e.g. severity or death?
- Are any health factors associated with an increase in COVID-19 cases or COVID-19 severity?
- If yes, can any of these health factors be attributed to air quality?

The impact of the COVID-19 pandemic has been significant and far-reaching and remains prevalent to this day. Therefore, a better understanding of factors that contribute to disease severity, such as air quality and health factors, can inform interventions aiming to improve COVID-19 outcomes.

Report roadmap

<table>
<tr><td rowspan="2">Select appropriate data sources</td><td rowspan="2">First data cleaning</td><td rowspan="2">Create SQL DB</td><td>Second MCS data cleaning</td><td rowspan="2">Add data to DB (MCS, API, COVID)</td><td rowspan="2">Analyse data from DB in notebooks</td><td rowspan="2">Analyse and draw conclusions</td></tr>
<tr><td>Pull API data</td></tr>
<tr><td colspan="4">Share files through GitHub</td><td colspan="3">Share files through Google Colab</td></tr>
</table>

Background

The COVID-19 pandemic has had a profound impact worldwide, causing widespread illness, death, and significant disruptions in various sectors. Gaining a comprehensive understanding of factors contributing to disease severity is essential for effective management and intervention strategies. Two critical areas of investigation in this project are air quality and health factors. Air quality is an indicator of air pollution, which has been linked to various respiratory and cardiovascular health issues. As COVID-19 is primarily a respiratory disease, exploring its potential association with air quality can provide insights into the interplay between environmental factors and disease severity. By analysing air quality data, we can assess whether it is predictive of COVID-19 severity or related deaths. Additionally, analysis of health factors, such as smoking habits, alcohol consumption, and pre-existing health conditions can help identify specific conditions or behaviours that may increase the risk of COVID-19 or impact its severity. Understanding these associations can inform targeted interventions in healthcare and public policies.

The target audience for our final report are public health policymakers and healthcare professionals. These individuals could include policymakers at the local, state, or national levels, as well as healthcare professionals such as doctors, nurses, and public health officials.

Steps/specifications

Data sources, gathering, and preprocessing

The data used in this project are summarised in Table 1.

Table 1: Summary of project data.

Data description	**Source**
COVID-19 hospitalisation (ICU/HDU admissions) and deaths by region	Office for National Statistics [1]
Pre-existing health conditions, and health factors (smoking habits, alcohol consumption), by region	Millenium Cohort Study [2]
General health ratings by region	Office for National Statistics [3]
API for air quality data	Air Quality by WeatherBit [4]

The Office for National Statistics (ONS) COVID-19 (referred to as 'covid') dataset contains infection data for a six-week period (Feb-Mar 2023), including overall hospital admission rates, intensive care unit (ICU) and high dependency unit (HDU) admission rates, and death rates, for nine defined regions of England (North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East, and South West) [1]. The reliability of covid data reporting and the accuracy of case numbers is important to the validity of this project's analysis. Although the ONS' reporting as a reliable statistics institution brings confidence in the accuracy of data, a potential limitation is the under-reporting of covid cases because of a lack of access to testing or positive tests not being reported by members of the public [2].

The data for health factors were obtained from two sources; the Millenium Cohort Study (MCS) [3] and ONS [4]. The MCS is a longitudinal study conducted in the UK to examine the development and well-being of children. The study began in 2000 and follows a nationally representative sample of around 19,000 children and their families. It has data on families' health, education, family background, social environment, and more. The MCS findings have already been used to analyse the impact of various variables on long-term outcomes [5], proving its reliability for investigating the potential connections between specific health factors and the number of covid cases or disease severity. For our project, we used the latest dataset from 2018. Non-relevant variables and data collected from non-English regions were excluded from our study.

The covid and health data were initially processed in excel and SQL. Pivot tables (excel) were used to convert the responses to binary data and deal with duplicate data entries from respondents in the same family. This was necessary to minimise skew in the analysis as some questions used responses from family units whereas some from each person within the said unit. Due to the large dataset and multiple SQL tables, the most efficient way to analyse the data was to create a new table (all_Covid_data) of average values and group the health conditions, covid deaths and ICU admissions region. Data from SQL tables were converted to CSV files, which could be easily imported into GitHub and Jupyter Notebook files. One of the key challenges during data visualisation was sharing the notebook within our team that can be edited by all team members. To combat this, we decided to use Google Colab, and imported our CSV files from GitHub. During the data visualisation process, Pandas and NumPy were used to manipulate the data frames, and Matplotlib and Seaborn were used to create graphs.

API

The Air Quality Index (AQI) data for each location was obtained via the RapidAPI host server [6]. AQI is a metric that classifies air quality on a numeric scale as follows: 0-50 ('Good'), 51-100 ('Moderate'), 101-150 ('Unhealthy for Sensitive Groups'), 151-200 ('Unhealthy') and 201-300+ ('Very Unhealthy) [7]. Following the generation of unique user keys, latitudinal and longitudinal coordinates were used to specify relevant regions. AQI values were obtained for two cities per region over a period of five days. Following this, the mean AQI value for each region was calculated.

Implementation and execution

Team roles

Each team member contributed to a different part of the project as shown in Table 2 below. More information can be found in the project activity log.

Table 2: Team member responsibilities.

Team member	**Main responsibilities**
Emer Buggy	Raw data collection and preprocessing, API
Emma Horton	API, data analysis & visualisation (health factors and air quality)
Punam Rattu	Raw data collection and preprocessing, Jira workspace admin
Rebecca Warren	Data analysis & visualisation (air quality and icu/deaths) , machine learning feasibility analysis, report editing
Sabina Wellington	API, data analysis & visualisation (air quality)
Maariya Daud	Data analysis & visualisation (health factors, covid deaths, hospital admissions)

Tools and libraries

Table 3 displays the tools and libraries used for the project.

Table 3: Tools and libraries used

Tool or library	**Use**
Pandas	Manipulation of dataframes
Numpy	Manipulation of dataframes
Matplotlib	Plotting and visualisations
Seaborn	Plotting and visualisations
Excel	Pivot Tables

Implementation process

All tables were created using SQL and exported as CSV files for data analysis. Initially, there was a consensus about using Jupyter Notebook to create the graphical visualisations; however, we noticed that it would be challenging to share and manipulate code between group members. To address this, we decided to use Google Colab, which proved a better option as the notebooks could be shared more easily amongst team members.

Agile development

To become ‘agile’, we used methods such as an iterative approach and code reviews. Our iterative strategy involved weekly video meetings and using a Jira board to assign tasks and track their progress. This ensured that all team members were aware of what others were doing and which small changes were being made. Moreover, the weekly meetings allowed us to discuss challenges and plan new approaches to problems. Code reviews involved using Github and Google Colab to share code and the teams’ Slack channel to provide feedback. Github and Google Colab allowed us to read and run one another’s code and view when changes had been made. The Slack channel allowed us to openly share feedback and suggestions for how the code could be improved. This method encouraged open collaboration and continuous improvements to each task.

Implementation challenges

The challenges that we faced included

- An API daily limit (air quality data) of 25 requests per day
- How to best share notebooks/data collaboratively so that all could use/edit
- How to structure our data tables to make analysis and visualisation easier

Results

Our analysis focused on nine geographical regions in England: (1) North East, (2) North West, (3) Yorkshire & The Humber, (4) East Midlands, (5) West Midlands, (6) East of England, (7) London, (8) South East and (9) South West. Using these regions, we studied the relationships between covid cases, covid-related deaths, ICU admissions, pre-existing health conditions and air quality. The covid data was taken from the ONS study over a 6-week period in Feb-Mar 2023 and the health data from the MCS 2018 dataset. To obtain air quality data, we chose two large cities in each region and used the API to collect real-time air quality index (AQI) values.

Covid cases and deaths

The data suggests that most regions had a similar number of covid-related deaths. Fig. 1 presents the percentage of national covid-related deaths per region (mean = 11.1%). However, the South East (15.6%) and North West (15.3%) were much higher than the North East (5.5%) and South West (7.3%). This trend may be accounted for by regional population or demographic variations. The same trend was not observed with covid-positive tests, as the North East and North West account for 12.6% and 12.9% of the national total, respectively. The correlation coefficient between positive tests and deaths was found to be 0.05, showing no correlation.

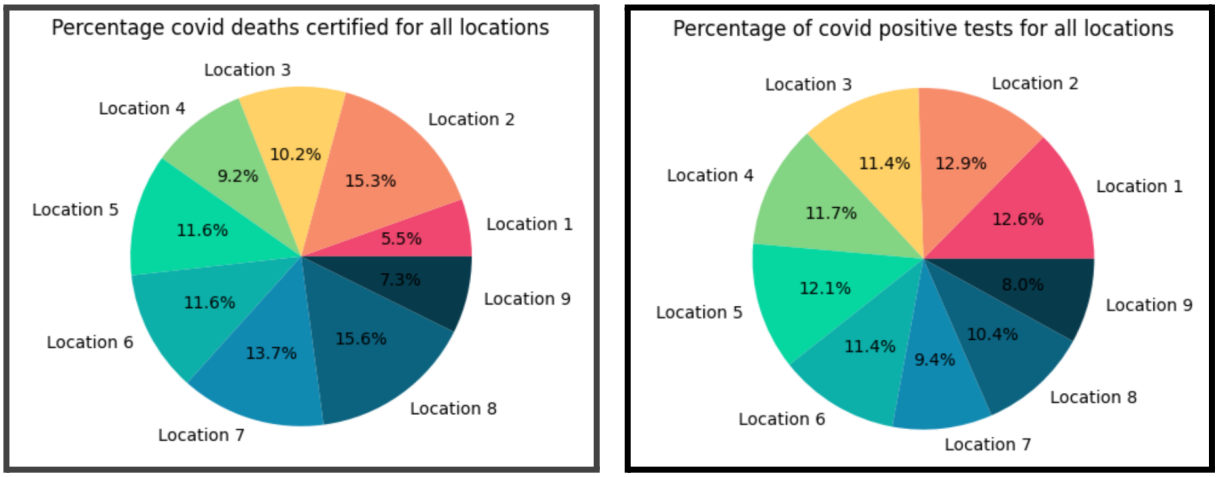


Fig. 1: Percentage of covid-related deaths (left) and covid-positive tests (right) for all locations ((1) North East, (2) North West, (3) Yorkshire & The Humber, (4) East Midlands, (5) West Midlands, (6) East of England, (7) London, (8) South East and (9) South West).

Air quality

Our analysis of the AQI data of 18 locations (2 per region) revealed notable variations in air quality between the cities (Fig. 2). Manchester and York had the lowest average AQI values, indicating better air quality, with values of 27 and 30, respectively. Cities North of London were classified as good with AQI ranges <50. In contrast, Hackney, Brighton, and Ealing had the highest AQI values of 79, 73, and 69, respectively, suggesting poorer air quality, despite falling within the moderate range. The high values in London (Hackney and Ealing), also positively correlated with increased ICU/HDU admissions and average deaths, compared to other regions.

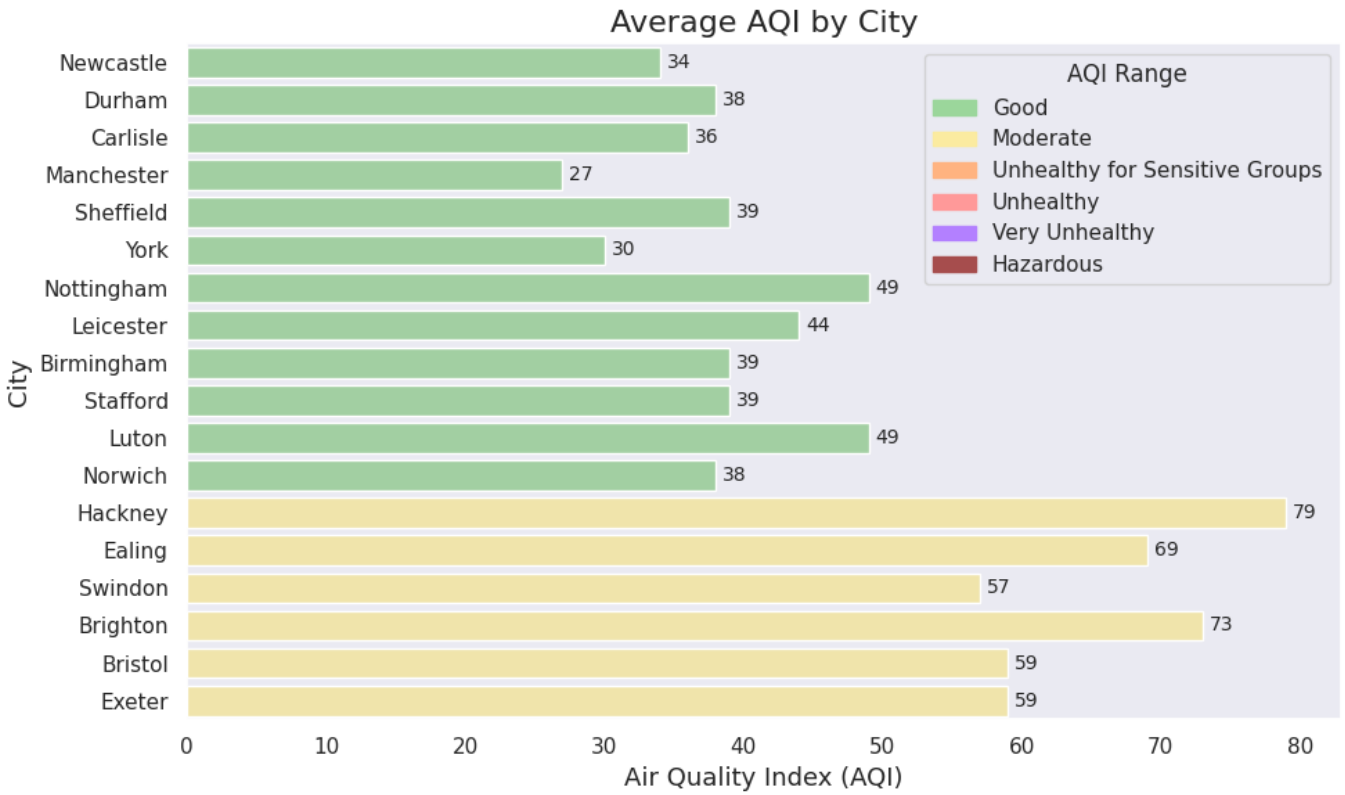


Fig. 2: Mean air quality index (AQI) across selected regions within England (n=18).

Air quality, ICU admissions and deaths

We investigated the link between air quality, covid deaths, and covid hospitalisations. For this analysis, covid-related deaths and ICU/HDU admission data were averaged over the six-week period. Average air quality for each region was calculated from the average values (over a five-day period) of the two cities in each area. Fig.3 illustrates the average ICU/HDU and deaths per region, with the shading representing the average air quality of that location.

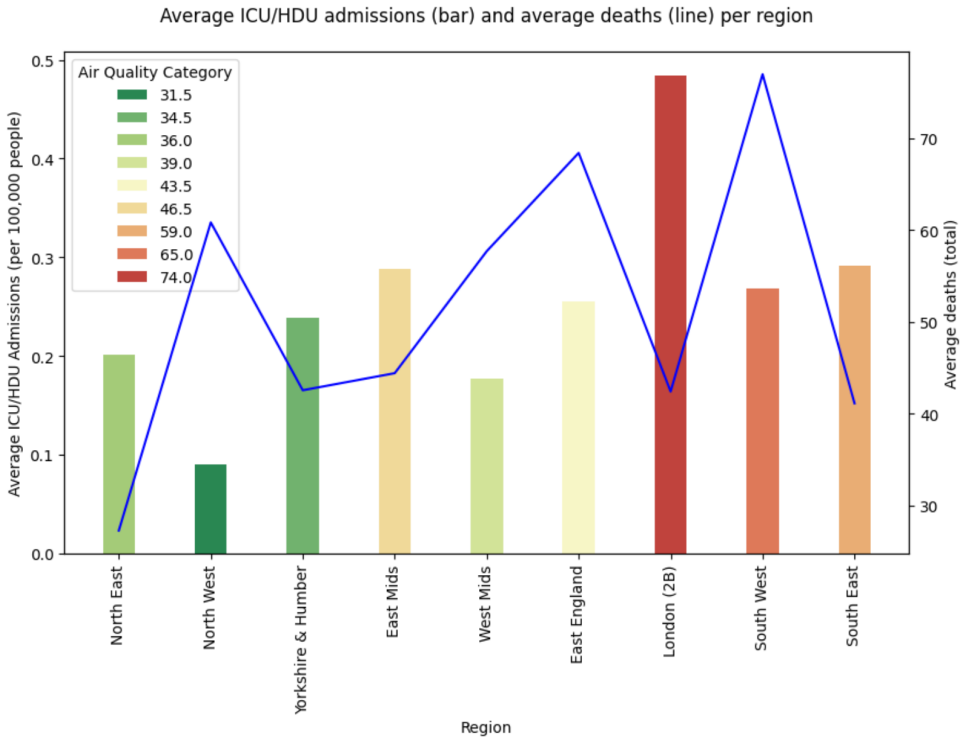


Fig. 3: Average ICU/HDU admissions and average deaths per region, coloured by air quality.

A positive correlation of 0.85 was found between air quality and ICU/HDU admissions, suggesting a strong link, and a much weaker correlation of 0.12 between air quality and deaths. These findings imply that health conditions exacerbated by poor air quality may play a role in the severity of covid infection.

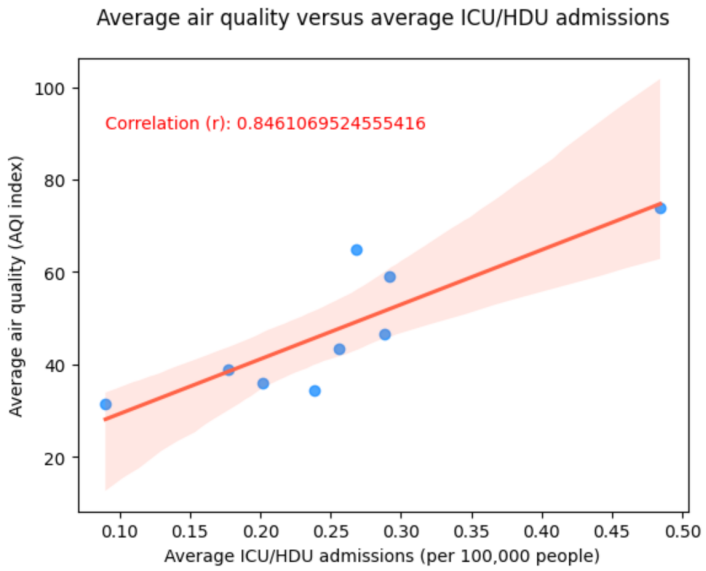
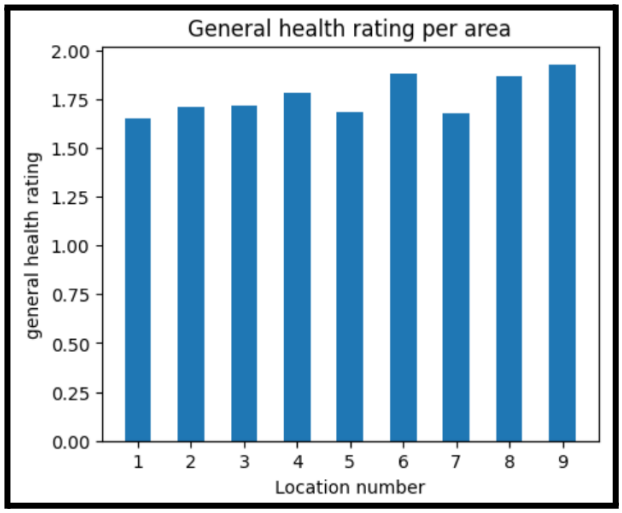


Fig. 4: Average air quality versus average ICU/HDU admissions

Health factors

The 'general health' rating (a numeric metric representing a person's overall health) of the population in all locations was similar (Fig. 5a). Just under half (47%) of the population had 'very good' health and only 5.5% of the population had 'bad' or 'very bad' health (Fig. 5b). Interestingly, the South East (8) had a better health rate compared to the North East (1), which may correspond to the higher number of covid-related deaths from that region, as discussed above.



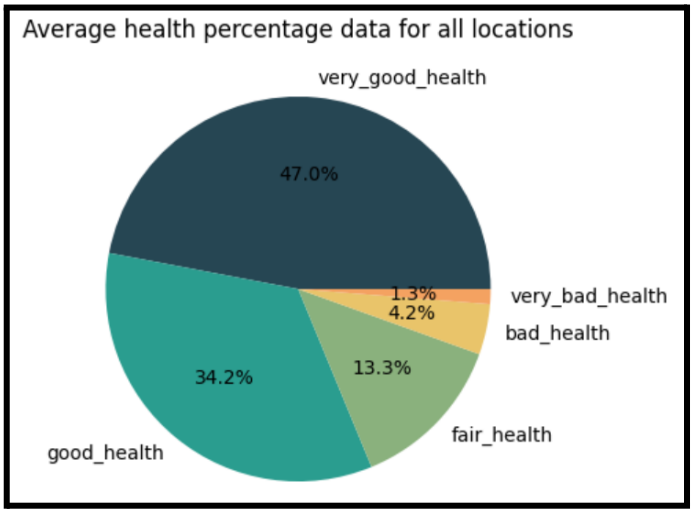


Fig. 5: (a) (left) 'General health' rating per region, (b) (right) Percentage of general health classifications (very bad, bad, fair, good, very good) for all regions.

Positive correlations were found between the majority of health factors and covid-related deaths (Fig. 6). In contrast, there was either a negative correlation or no correlation between health factors and covid-positive tests.

In terms of external factors, e-cigarette vape usage is high in all regions and its correlation coefficient with covid-deaths is 0.9, showing a very strong correlation. Hayfever also had a high correlation (0.8) whereas cancer had the lowest (0.3). This could be due to hayfever being a very common condition and cancer being relatively rarer. The South East (location 8) had the highest number of hayfever cases and also the highest number of covid-related deaths. Previous research has suggested that whilst hayfever does not directly have an influence on covid prevalence, it can play a role in asthma, an

illness that has been linked to more severe covid outcomes [8]. The correlation coefficient between asthma and covid-related deaths was 0.6.

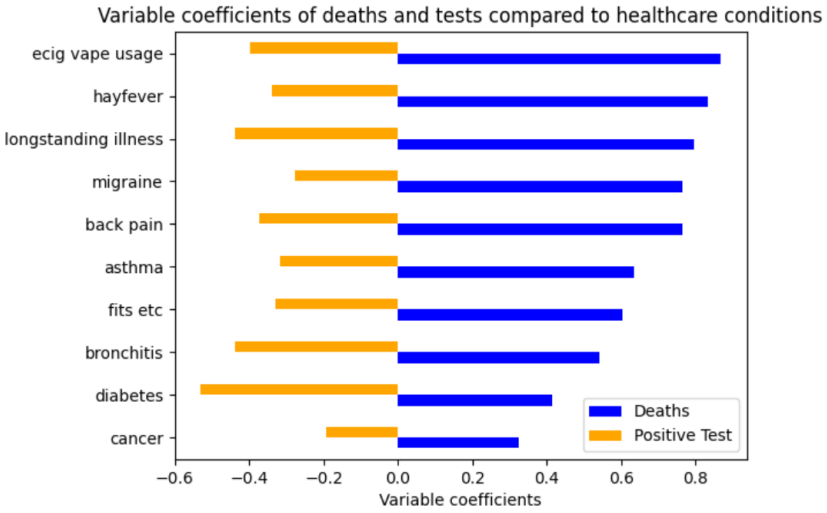


Fig. 6: Correlations between health conditions and covid-related deaths (blue) and covid-positive tests (orange).

Health factors and air quality

The data was examined to see if correlations existed between the health factors associated with covid-related deaths and air quality. Whilst no correlation existed for most health factors, a strong positive correlation was found (0.72) between the prevalence of fits etc. (where 'fits etc.' is an MCS classification) and the air quality of a region (Fig. 7).

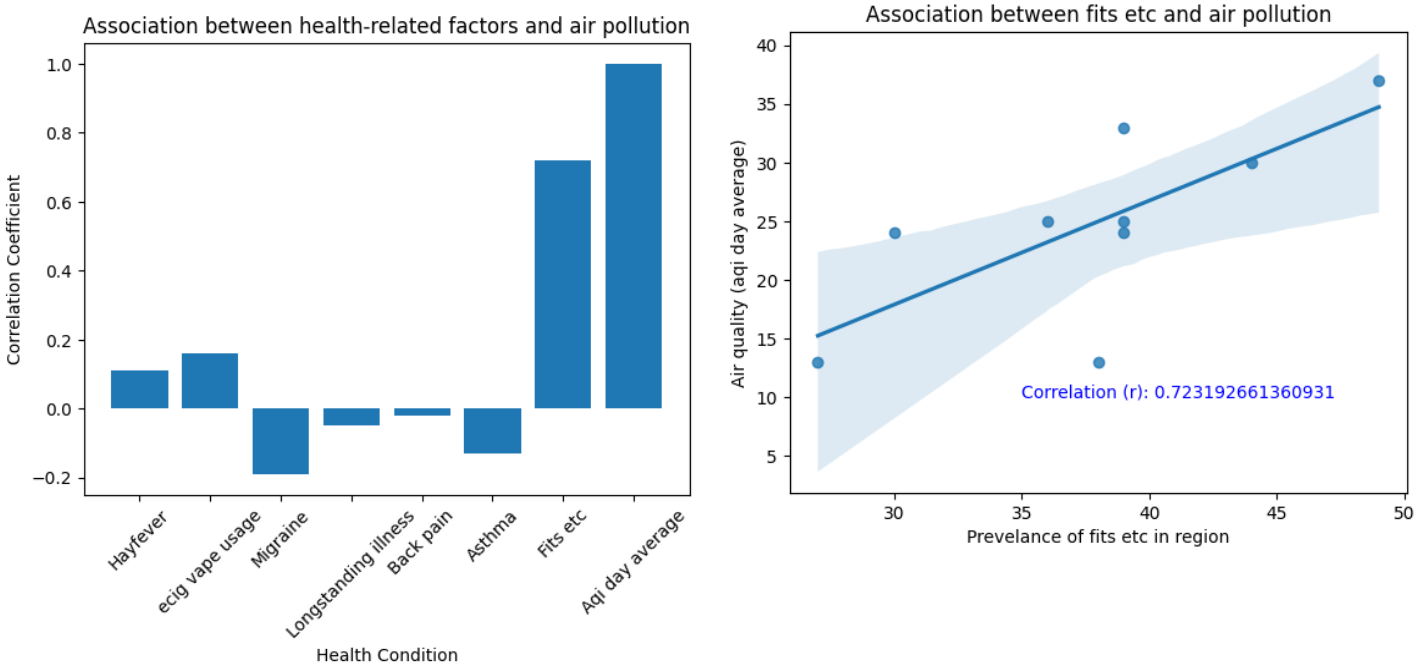


Fig 7: (left) Association between health related factors and correlation coefficient, (right) Average air quality versus prevalence of fits etc.

Machine learning feasibility

Applying machine learning (ML) models to averaged data with only nine rows poses significant challenges and limitations. ML models require enough data to effectively learn and generalise from the patterns present in the dataset. If the 'training' dataset is small, the risk of overfitting increases significantly as the model may struggle to discern genuine

patterns from noise, leading to poor generalisation of 'test' (unseen) data. Given more time, using the raw, unaveraged data would have been our preferred approach for applying ML algorithms. However, transforming the raw data into a suitable format required substantial preprocessing. Due to time constraints, these preprocessing steps could not be adequately performed, making utilising the raw data for ML analysis impractical.

Limitations of our data

We note the limitations of our data as follows:

- The covid ICU and deaths data only covered a 6-week period throughout the pandemic
- The air-quality values were averaged from only two locations per region and were collected in real-time as historical data were not available
- The small number of regional city data may not be be representative of regions as a whole

Conclusions

Our project explored the potential relationships between covid outcomes, air pollution, and health factors by conducting extensive analyses of selected datasets using python libraries, SQL, and other computational tools. We tackled the objectives successfully by setting and meeting pre-defined project milestones, working collaboratively, and dividing tasks and responsibilities. Additionally, we effectively used scientific tools and libraries (summarised in Table 3) to perform robust data manipulation, analysis, and visualisation, which enabled us to gain meaningful insights into relationships between covid, health factors, and air pollution. Incorporating an API to fetch data also enriched our analysis by broadening the scope of available information. Furthermore, we used a SQL database, which provided a structured framework to store and manage the diverse datasets used in this project.

In terms of scope for improvement, a longer data collection period could provide more comprehensive insights into the trends and patterns over time. Moreover, it should be possible to collect and incorporate historical air quality data into our analysis to assess long-term exposure to air pollution and its potential impact on covid and health factors. Lastly, the regions in England could be divided further to uncover the influence of health factors and air pollution on covid at a more localised level.

References

1. Office for National Statistics (ONS) 2023. *Coronavirus (COVID-19) latest insights.* Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19/latestinsights
2. Jones I. 2022. *Tens of thousands of UK Covid-19 cases missed from daily figures.* Available from: https://www.independent.co.uk/news/uk/office-for-national-statistics-open-university-omicron-b1993374.html
3. University College London, UCL Institute of Education, Centre for Longitudinal Studies. (2023). *Millennium Cohort Study. [data series].* 14th Release. UK Data Service. SN: 2000031, DOI: http://doi.org/10.5255/UKDA-Series-2000031
4. Office for National Statistics (ONS) 2023. *General health, England and Wales: Census 2021.* Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandwellbeing/bulletins/generalhealthenglandandwales/census2021
5. Millenium Cohort Study 2023. *Recent scientific publications.* Available from: https://cls.ucl.ac.uk/cls-studies/millennium-cohort-study/#publications
6. RapidAPI 2023. *Air Quality API (weatherbit).* Available from: https://rapidapi.com/weatherbit/api/air-quality
7. AirNow (2021). *AQI Basic | AirNow.gov*. [online] www.airnow.gov. Available at: Air Quality Index (AQI) Basics.
8. Institute for Quality and Efficiency in Health Care (IQWiG); 2020. *Hay fever or COVID-19: How do the symptoms differ?* Available from: https://www.ncbi.nlm.nih.gov/books/NBK556944/