Quantifying the Impact of Environmental Factors on Chronic Disease by Analysing the Vehicle Registration and Air Quality Indexs'

Vaibhav Sonia
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x22136860@student.nci.rl.ie

Pratik Shetty
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x21227578@student.ncirl.ie

Avis Massey
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x21199752@student.ncirl.ie

***Abstract*—This paper highlights the critical issue of hazardous air quality and its detrimental effects on human health. While unsafe levels of particulate matter are typically linked with newly industrialized and developing countries, this is a misconception, especially in densely populated urban areas. Despite vehicular exhaust being a major cause of air pollution, our understanding of its effects on public health is limited due to factors such as socioeconomic status, errors in measurement, and the tendency to avoid exposure to polluted areas. In order to tackle this problem, we investigate how emissions-cheating diesel vehicles, which emitted pollutants that were up to 150 times greater than those released by gasoline cars, were distributed throughout the United States from 2008 to 2015. By analyzing the corpus of vehicle registration data, we demonstrate that vehicular emissions contribute significantly to PM2.5 levels, posing a severe risk to human health and aggravating the risk of debilitating and fatal diseases. We argue that preventive investment and action are necessary to mitigate the harmful effects of vehicular emissions on the environment and public health.**

***Index Terms*—AQI, Vehicle Emission, Python, MongoDB, Postgr**

I. Introduction

The Air Quality Index of the United States is managed by the U.S. Embassy. The U.S. State Department, in collaboration with the U.S. Environmental Protection Agency (EPA), have installed air quality devices on some of its facilities to provide required information that may help protect the health of American personnel. The U.S Embassy is responsible for the operation of these devices which measures the particulate pollution with a diameter of 2.5 microns or less, also known as PM2.5, and convert hourly concentrations into the EPA's Air Quality Index (AQI), which is accessible to the public on the airnow.gov website [1]. According to a 2008 World Health Organization (WHO) report, ambient air pollution is responsible for 1.3 million premature deaths worldwide and in 2012 this figure increased to 3.7 million whereas in 2008 household air pollution was responsible for two million deaths, and by 2012 this number was almost doubled to 4.3 million. The combined effects of household and ambient air pollution in the year 2012 led to seven million premature deaths globally [2]. The Global Burden of Disease study conducted in 2013 stated air pollution led to various debilitating and deadly diseases, including lung cancer, heart disease, stroke, and Chronic obstructive pulmonary disease (COPD) contributing to excess mortality rates in the U.S. Air pollution is now the world's fourth-leading fatal health risk, causing one in ten deaths in 2013 [3].

In the United States, the issue of car exhaust as a significant contributor to air pollution has sparked a contentious academic and policy debate. Researchers from the International Council on Clean Transportation, George Washington University Milken Institute School of Public Health, and the University of Colorado Boulder conducted a recent study in 2010 and 2015 to assess the link between vehicle emissions and air pollution, as well as the resulting health impacts on a global, regional, national, and local level. Using data from vehicle registrations, the study estimated that ambient PM2.5 and ozone pollution caused approximately 361,000 premature deaths worldwide in 2010, with this figure rising to approximately 385,000 in 2015. Notably, the four largest vehicle markets (China, India, the European Union, and the United States) were found to be responsible for roughly 70% of these impacts in 2015 [4].

The research predominantly attempts to investigate whether vehicular emissions have contributed to air pollution exacerbating chronic diseases. For this we identified logically linked datasets and intend to merge them, the raw and unorganized data is retrieved via API. The data is in CSV (Comma-Separated Values) and JSON (JavaScript Object Notation) format and stored in an open-source document-based NoSQL database system MongoDB.

II. Literature Review

A literature review will be outlined below to give an overview of generalised research on the selected topic, including recent trends and developments.

The International Council on Clean Transportation (ICCT) has conducted a study on the global health impacts of vehicle exhaust. The study reports that vehicle exhaust emissions cause a significant number of premature deaths worldwide, with estimates ranging from 185,000 to 600,000 deaths per year. The report highlights, developing countries account for roughly 70% of global health. Additionally, the study provides an overview of the health impacts of vehicle exhaust emissions and it also analyses the economic costs associated with these health impacts. The authors estimate that the economic costs of vehicle emissions range from $1.4 trillion to $4.0 trillion per year. They emphasise the significance of shifting to cleaner transportation technologies such as electric and low-emission vehicles. The authors also emphasize the role of policy interventions, such as regulations and incentives, in encouraging the adoption of cleaner transportation technologies. The ICCT concluded that reducing vehicle exhaust emissions is an important step towards improving global public health and emphasises the need for global efforts to reduce transportation-related air pollution, such as policies that promote cleaner transportation technologies and infrastructure. [4]

Diane Alexander and Hannes Schwand in their research paper investigate the health effects of vehicle emissions on infants and children, focusing on the Volkswagen (VW) diesel exhaust cheating scandal. In the year 2008, a new generation diesel engine was introduced that was marketed to environmentally conscious consumers. By the year 2015, over 600,000 cars with clean diesel technology were sold in the United States. In the fall of 2015, however, it was discovered that a single "clean diesel" car could pollute as much nitrogen oxide (NOX; a precursor to fine particulate matter and ground-level ozone) as 150 equivalent gasoline vehicles after which we refer to vehicles with "clean diesel" technology as cheating diesel cars. The authors discovered that exposure to emissions from Volkswagen diesel vehicles had a significant impact on the respiratory health of children. They found that children living in areas with higher concentrations of emissions from VW diesel vehicles were more likely to be diagnosed with asthma or to have a nebulizer. Exposure to VW emissions is thought to have contributed to an additional 2,200 cases of childhood asthma in the United States. The authors concluded that the health costs associated with vehicle emissions cheating are substantial and have implications for public policy and highlight the need for more stringent regulations to reduce traffic-related air pollution and protect infant health. The study's authors concluded that reducing vehicle exhaust emissions is an important step towards improving global public health and emphasises the need for global efforts to reduce transportation-related air pollution, such as policies that promote cleaner transportation technologies and infrastructure. [5]

Overall, the Alexander and Schwand studies, as well as the ICCT, provide important insights into the significant impact of vehicle exhaust on public health. These findings support previous research by emphasising the importance of policies and regulations to reduce transportation-related air pollution and protect public health. More research, however, is required to fully comprehend the complex relationship between vehicle exhaust, air pollution, and chronic disease.

III. Methodology

A. Description of Dataset

The information on different datasets which were employed by us to perform analysis is enumerated below:

1) Dataset 1 : **Air Quality Measures in USA for the year 1999 - 2013**
The dataset provides information on air quality measures in different counties of the United States from 1999 to 2013. The Environmental Protection Agency (EPA) maintains a database called the Air Quality System (AQS) which contains data from approximately 4,000 monitoring stations around the country, mainly in urban areas. The data is obtained in CSV format from the open data portal of the Centers for Disease Control and Prevention (CDC) and contains 14 columns to name a few - *State name, County name, Report Year, Value, Stratification Level.* The reason for selecting this dataset is due to its provision of accurate information on the Air Quality Index (AQI) for each state. This information is imperative for our research because it allows us to interlink other datasets. The dataset's precise state-by-state AQI information provides us with a great deal of flexibility in our analysis, which is essential to quantify the impact of environmental factors on chronic disease.

2) Dataset 2: **Specific Chronic Condition in the USA for the Year 2013**
The Specific Chronic Conditions dataset includes data about 21 particular chronic health conditions experienced by people enrolled in Original Medicare. The data is retrieved in a JSON format and the information provided in the dataset includes the number of people with each condition and how often they use healthcare services. The data is sorted by location and individual chronic condition and has about 12 columns - *Bene_Geo_Lvl, Bene_Geo_Desc, Bene_Geo_Cd, Bene_Age_Lvl, Bene_Demo_Lvl, Bene_Demo_Desc, Bene_Cond, Prvlnc, Tot_Mdcr_Stdzd_Pymt_PC, Tot_Mdcr_Pymt_PC, Hosp_Readmsn_Rate, ER_Visits_Per_1000_Benes.* We opted to choose this data set because of its reliable information on the health conditions of the USA citizens. The precise data by state and individual chronic disease allow us to establish correlation with other data sets and give us the liberty to perform essential analysis to measure the impact of environmental factors on chronic disease.

3) Dataset 3: **Motor Vehicle Registrations in the USA for the year 1900 to 2020**
The Motor Vehicle Registrations dataset is a source of information on the trends of new motor vehicle registrations in the United States over time, providing insights into how the market has evolved in different states. By examining the number of new automobile, truck, motorcycle, and bus registrations across different years and states, it is possible to draw comparisons and analyze the relative popularity of these vehicle types. This data has the potential to be a valuable resource to gain a deeper understanding of the motor vehicle market in the United States and to develop informed strategies based on these insights. The data contains about 6 columns which are enlisted further: *Year, State, Auto, Bus, Truck, and Motorcycle.* We selected this data for its dependable information on vehicle registration through the year 1900 to 2020 which entails to be vital information for our analysis. As it contains the data in state-wise format it gives us the flexibility to interlink it with other two datasets which enables us to measure the impact of environmental factors on chronic disease.

B. Data Pre-Processing Algorithms

The flowchart illustrated in "Fig. 1" demonstrates the Extract Transform and Load or the (ETL) process and explains what are the steps involved in it. In the Extract stage of ETL, various data sources as enumerated in the aforementioned section are accessed using application programming interfaces (APIs) and libraries in Python. The data obtained is then stored in a MongoDB database (MongoDBCompass) for further transformation. In the Transform stage, the data from the staging database is transformed to meet specific requirements, which involves eliminating redundant and irrelevant information and then it is merged with other sources.

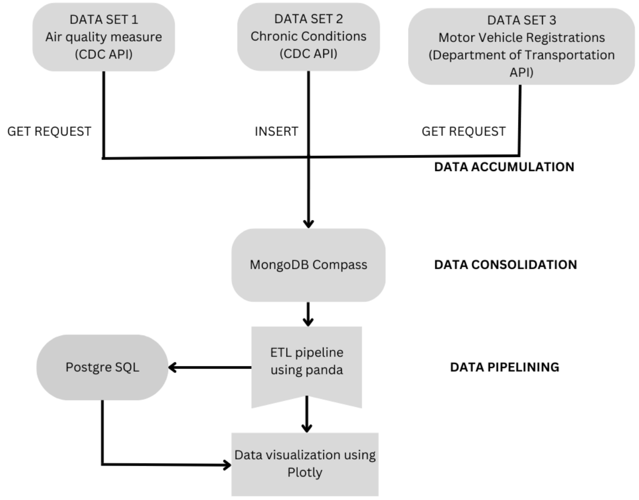

Fig. 1. Illustration of Data Gathering Process.

The data is then normalized and structured to facilitate analysis further. Finally, the transformed data is loaded into the final database called PostgreSQL (PGAdmin4) in batches. This automated process is used to investigate the relationship between vehicle registration, air quality and chronic disease in the USA. The data is analyzed to generate insights and reports on how vehicle emissions contribute to air pollution which poses a severe risk to human health.

1) **Data Gathering.**
1) Air Quality Measures in the USA for the year 1999 - 2013 and Specific Chronic Condition in the USA for the Year 2013:
The Environmental Protection Agency (EPA) maintains a database called the Air Quality System (AQS) which provides access to the CDC's Air Quality Measures in the USA for the years 1999 – 2013. The data is retrieved from the CDC's API, for which the code is written in Python and is read from their URL using the Pandas library and stored into a DataFrame named *"df1"*. This dataset contains information related to Medicare beneficiaries, including their demographic information, Medicare coverage, and chronic condition status. Similarly, the Centers for Medicare & Medicaid Services (CMS) a federal government website managed by the U.S. maintains a database on Specific Chronic Conditions in the USA. The data is read through the website's API and stored in a Pandas DataFrame named *"df2"*. This dataset contains information about the percentage of adults in the US with specific chronic conditions, based on data from the Behavioral Risk Factor Surveillance System (BRFSS). Both datasets are stored as DataFrames, which are two-dimensional labelled data structures with columns of potentially different types, similar to the spreadsheet.

2) Motor Vehicle Registrations in the USA for the year 1900 to 2020, :
United States Government provides computational access to USA's vehicle registration distribution in the United States. The data is recruited through API in JSON i.e., Java Script Object Notation format. The data confers the period in years, States, Automobiles, Bus, Trucks and Motorcycles which register as vehicles encompassing 6 columns. A Pandas DataFrame called *"df_VH"* is used to hold the data once it is read through the website's API. The dataset is kept in a two-dimensional labelled data structure with optional columns of various kinds. The response is further processed using *'JSON'* method. Using the python *'request.get(url)'* library, the data was requested in Python which provides the response data on the U.S Department of Transportation's on-time performance dataset. The code sourced or extracted the metadata from the dictionary using the *'Meta'* key. Necessary steps like retrieving the column names and listing the 'name' list containing all the column names

and further printing its length and the list itself.

2) **Data Consolidation**

The next step is to transform this unstructured data into a structured format by employing a database. Following a thorough analysis of various database programs, we have decided to utilize MongoDB and, specifically, MongoDB Compass - a graphical user interface (GUI) tool that provides a visual interface for MongoDB - to meet our specific needs. It allows users to interact with MongoDB databases and collections and provides features such as query building, schema analysis, data visualization, and data exploration. Since, it is a type of database that does not require a structured query, its schemaless nature makes it a convenient and efficient method of storing data without the need to define a schema beforehand, which can save time. To initialize a connection to the MongoDB database we have used a Python library called *Pymongo*. After establishing a successful connection, the cluster's databases and collections can be handled, modified, and queried using Pymongo commands. A new database named "projectDAP" is created in MongoDB using the MongoClient instance named *"client"*. Once the client has been created, using the *insert_many()* method the data is inserted into two different collections 'ChronicDisease' and 'AirQuality'. For the third dataset which is retrieved from JSON, we have stored the data in dictionary *(dict1)* and by using *'insert_one'* (italics) method the dictionary is inserted as a single document into the 'CarReg' collection.

3) **ETL Pipelining:**

Once the data has been retrieved, the datasets undergo a processing stage where data cleaning is performed. The details of this stage are described in the following sections.

1) **Air Quality Measures in USA for the year 1999 - 2013 :** After successfully inserting the data into MongoDB. The database is fetched from the '*AirQuality'* collection and using Pandas a dataframe is created and is assigned to the variable *'df_AQ'*. Since we are only interested in the data which has PM2.5 in microorganism, we filter the data with of the column name 'MeasureName', post-filtering the dataframe will only contain attributes that are associated with 'Annual average ambient concentrations of PM2.5 in micrograms per cubic meter (based on seasonal averages and daily measurement'. After successfully filtering the data we clean it by eliminating multiple columns namely '*id', 'MeasureId', 'MeasureType', 'StratificationLevel', 'StateFips', 'CountyFips', 'DataOrigin', 'MonitorOnly'. The 'ReportYear'* column is filtered to be greater than or equal to 2000 and less than or equal to 2012 since our study consists of the time period from 2000 to 2012. Hence clean data is ready to be visualized and processed for other computations.

2) **Specific Chronic Condition in the USA for the Year 2013** : Once the data has been successfully stored in the collection, The find() (italics) method is used to retrieve all the details from the collection - *"ChronicDisease"*. The list of dictionaries in the variable *'cursor1'* is then converted into a Pandas DataFrame using *'pd.DataFrame'* command and stored in variable *df_CD*. As our project aims to target chronic diseases which are caused due to air pollution, we eliminate all the unnecessary columns and modify the data which only contains diseases such as *'Asthma—COPD—Ischemic Heart Disease—Stroke'*. To achieve data in a more suitable format The 'astype()' method is utilized to convert the data type of multiple columns in the DataFrame from its default type to string. This is done to ensure that the data is suitable for visualization and can be used for further computations. Since the dataset comprised of State and County values together, split() function was used to retrieve only the state data. To make the data more readable we replaced the default names with more comprehensive names. *Eg: 'Bene_Age_Lvl': 'Age Group', 'Bene_Demo_Lvl': 'Patient Category'*. Finally using the *drop.na()* and *duplicate()* functions all Nan and duplicate values were eliminated with a resultant dataframe shape of (1940, 7). As a result, we now have a clean dataset that is suitable for visualization and further analysis.

3) **Motor Vehicle Registrations in the USA for the year 1900 to 2020 :** After successful data retrieval, the dataset is then cleaned and processed further. The Python library 'panda'(italic) is used to process the cleaning. Before successive cleaning and processing, the data is extracted from MongoDB database in the collection *'CarReg'* using the *'find()'* method of 'db'(italic) instance. A dataframe called *'df_VH'* is created using the list of datasets from MongoDB corpus or collection to *'pd.DataFrame()'*. It was done to ensure that the list of documents are converted into a data structure in the form of DataFrame. A quick review of the first few rows in the data is performed using the *'head()'* method. Using the *'drop()'* method, the less necessary columns have been discarded in order to acknowledge the appropriate structure of the data. Column names *'_id', 'sid', 'id', 'position', 'created_at', 'created_meta', 'updated_at', 'updated_meta', 'meta', 'motorcycles'* were eliminated. Moving ahead, using *'astype()'* method, the *'year'* column of the DataFrame is converted from default data type to integer to ensure the data is ready to be visualized and processed for other computations. The *'year'* column is filtered to be greater than or equal to 2000 and less than or equal to 2012 since our study consists of the time period from 2000 to 2012. Using *'dropna()'*, the

missing values and duplicate entries in the DataFrame are dealt with. The resultant DataFrame *'pp_df_VH'* includes all non-missing and non-duplicate entries from the original DataFrame.

4) **Data Storage**

The pre-processed data is then stored in a Structured Query Language - PostgreSQL database using for easy retrieval and management. The data is integrated with PostgreSQL databases using the *"psycopg2"* library. The reason for using PostgreSQL for this project is because it is an open-source, object-relational database system that is ACID (Atomicity, Consistency, Isolation and Durability) compliant. Its powerful query capabilities and ability to define custom data types make it a good choice for combining with MongoDB. Additionally, it offers features such as relational integrity, transactions, and stored procedures, which can be used to extend and enhance the functionality of MongoDB. Also, it stores data in tables with predefined schema, and its data model is designed to ensure data consistency and integrity. After successfully installing the necessary Python packages to connect to PostgreSQL, a connection is created to the PostgreSQL database. The connection for the PostgreSQL instance is then configured and is available to read or write data to the PostgreSQL database. Using the *psycopg2* library, a new database is created called "EVPOPULATION". The code which we have provided also checks if the database already exists, it will print a message indicating this, and if it doesn't exist, it will create the database. The set_isolation_level() method sets the isolation level to AUTOCOMMIT, which means that any SQL statements executed will be committed immediately rather than waiting for a transaction to be completed. The code also includes error handling to catch any exceptions that may occur during the connection process. Next, we insert the data from a pandas DataFrame into a PostgreSQL database table. First, the DataFrame is converted to a list of tuples, where each tuple represents a row of data. Then, an *INSERT* query template is created with placeholders for the table name, column names, and values. Next, a connection is established with the PostgreSQL database and a cursor is created to execute SQL commands. The code also checks whether the table already exists in the database and creates it if it doesn't. Then, the data is inserted into the table using the INSERT query template and the list of tuples. Now the inserted data is fetched from the table named 'merged_table' and loads it into a pandas dataframe. As a cursor was created before for SQL commands, we execute the following command *"SELECT query SELECT * FROM table_name"* to fetch all columns and rows from the table. After loading the fetched data from various sources, after merging or staging, into an operational database.

IV. Evaluation and Results

The final process of this project is that the data is retrieved and stored as a data frame in Python, and visualizations are created using colour palettes. The interactive data visualizations created using Plotly Dash, an open-source Python library, and Plotly.js, a widely-used JavaScript library, are displayed for analysis and evaluation of the results.

A. Visualizing the Air Quality Indexes

The evaluation of the air quality index is studied with the parameter of annual average ambient concentrations of PM2.5 in micrograms per cubic meter across all states in the USA between the years 1999 and 2013 as shown in "Fig. 2"

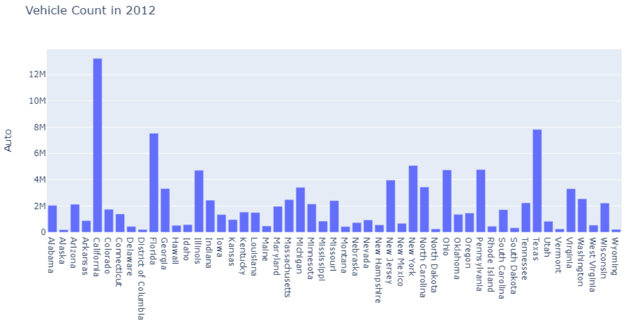

Fig. 2. Air Quality Index for the Year 1999 to 2013

It was observed that the state with the highest average concentrations of PM 2.5 in micrograms per cubic meter is Pennsylvania followed by the state of West Virginia, Georgia and the District of Columbia.

B. Visualizing Chronic Conditions

The analysis of the dataset on specific chronic conditions reveals that Asthma affects different age groups and genders differently. Young women are more prone to Asthma than older women as shown in'Fig. 3", and young men are more likely to have Asthma than older men and women. Furthermore, young women are more prone to Asthma than all other genders.

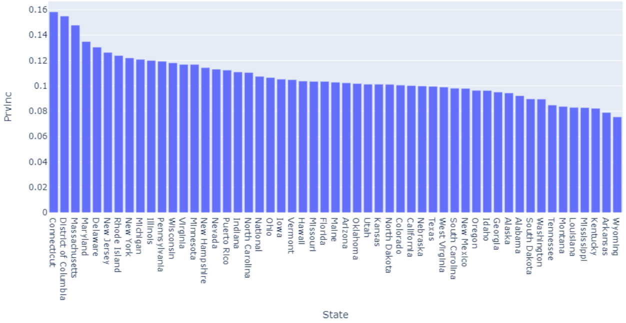

Fig. 3. Asthama Index for Females less than 65 years

The study also shows that the prevalence of Asthma varies across different states. In Puerto Rico and Rhode Island, older women above 65 years are more prone to Asthma as shown in‘Fig. 4”. In Connecticut, DC, and Massachusetts, young men below 65 years are more likely to have Asthma, whereas, in Puerto Rico and Rhode Island, older men above 65 years are more prone.

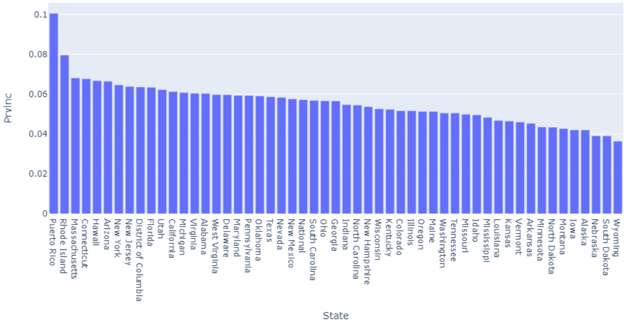

Fig. 4. Asthama Index for Females older than 65 years

These findings provide crucial insights into the demographics of Asthma and its prevalence across different age groups and genders in different states. The information can be used to create targeted Asthma prevention and management programs, especially in states where the prevalence of Asthma is higher in specific age groups and genders.

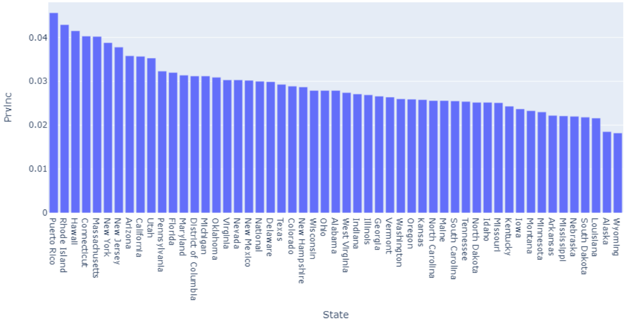

Fig. 5. Asthama Index for Males older than 65 years

Overall, this analysis can help healthcare providers and policymakers develop effective strategies for Asthma prevention, diagnosis, and management, ultimately improving the health outcomes of the affected population.

The analysis of the Chronic Obstructive Pulmonary Disease (COPD) dataset in the United States for the year 2013 revealed some interesting findings. COPD is a chronic inflammatory lung disease that causes obstructed airflow from the lungs and makes breathing difficult. The study discovered that women are more susceptible to COPD than men, which is a surprising finding given that COPD is generally associated with smoking, which is more common in men.

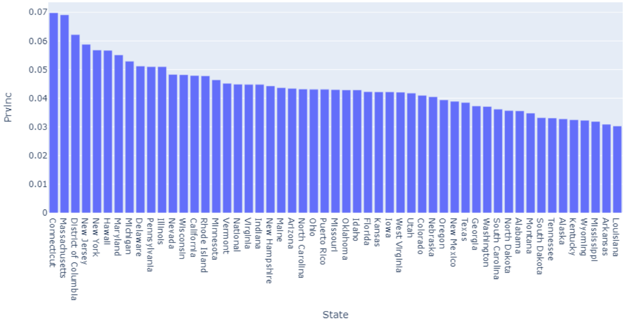

Fig. 6. Asthama Index for Males less than 65 years

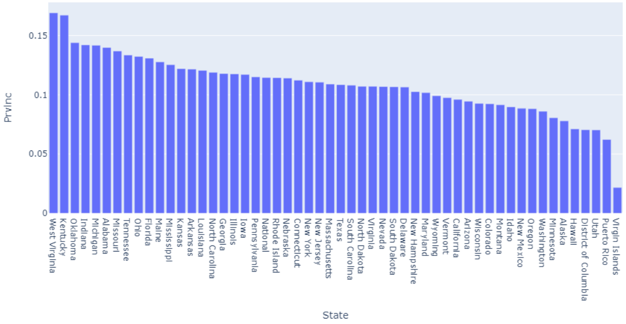

Fig. 7. COPD Index for Males

The study also discovered that the state of Puerto Rico has the highest prevalence of COPD among women, with a prevalence of 0.7, indicating that the disease is a major health concern for women in Puerto Rico as shown in‘Fig. 8” On the other hand, the state of Massachusetts has the highest prevalence of COPD among men, with a value of 0.04 prevalence as shown in‘Fig. 7”. This indicates that the disease is a significant health issue for men in Massachusetts.

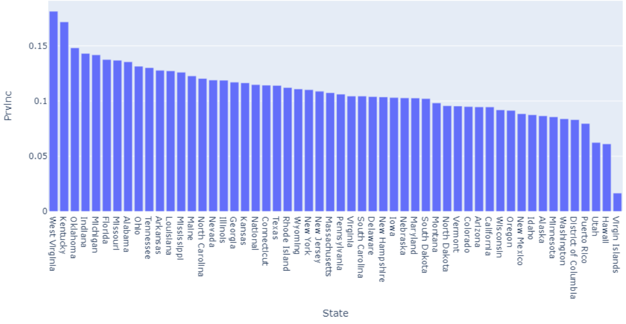

Fig. 8. COPD Index for Females

The results of this study can be compared to other chronic respiratory diseases like asthma. Unlike asthma, which affects more young people than the elderly, COPD is more prevalent among the elderly. The study found that women are more likely than men to develop COPD, whereas men are more likely than women to develop asthma. COPD is also more prevalent in some states than others, whereas asthma is more evenly distributed across the country. These comparisons emphasise the significance of understanding the demographics and geographic distribution of various chronic respiratory diseases in order to effectively target prevention and treatment efforts.

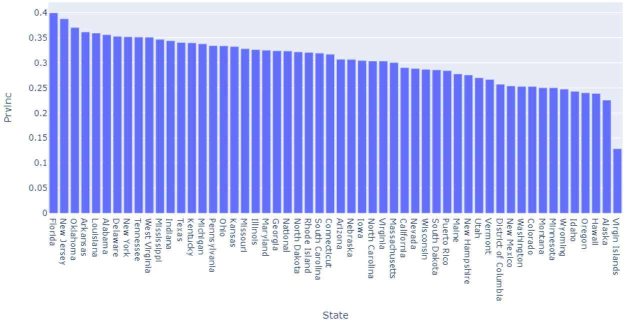

Fig. 9. Ischemic Heart Disease Index for Males

The dataset analysis revealed some interesting findings about the prevalence of Ischemic Heart Disease in the United States. It was discovered that men are more susceptible to Ischemic Heart Disease than women as shown in'Fig. 9" This finding contradicts previous findings for other chronic conditions, such as asthma and COPD, in which women were found to be more susceptible.

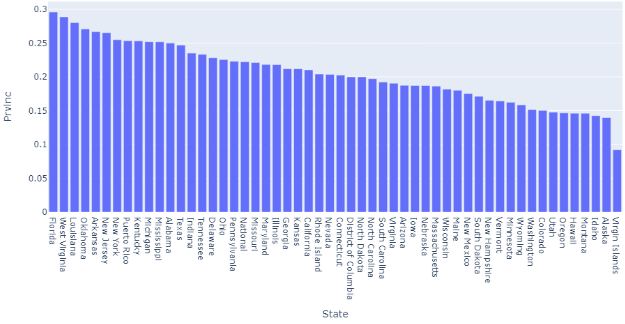

Fig. 10. Ischemic Heart Disease Index for Females

Florida was discovered to have a high prevalence of Ischemic Heart Disease in both men and women. However, women were found to have a higher prevalence with a value of 0.3993 than men with a value of 0.2958 as shown in'Fig. 10" This difference in prevalence rates could be attributed to various factors such as lifestyle choices, environmental factors, and genetic predisposition. Further analysis would be required to identify the exact reasons for this trend.

Surprisingly, the state of Virgin Islands had the lowest prevalence of Ischemic Heart Disease. This could be due to a number of factors, including a smaller population size, improved healthcare facilities, and lifestyle choices.

Overall, the findings indicate that Ischemic Heart Disease is a chronic condition affecting men more than women in the United States. Florida has a higher prevalence of Ischemic Heart Disease, with women being affected more than men. The information provided by the data will help policymakers and healthcare professionals design and implement targeted interventions to prevent and manage Ischemic Heart Disease.

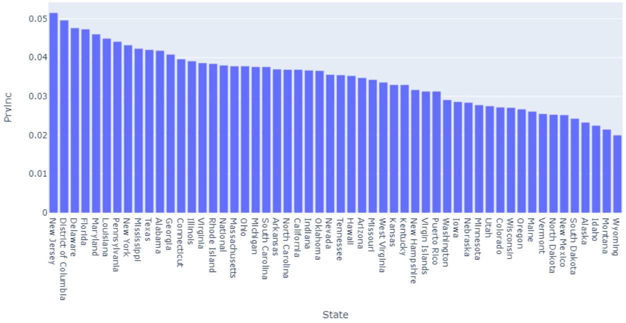

Fig. 11. Stroke Index for Males

According to an analysis of the dataset for heart disease and stroke in the United States, men are more prone to heart disease than women as shown in'Fig. 11". This is a significant discovery because it can assist healthcare professionals in developing targeted prevention and treatment strategies for men. Furthermore, it was discovered that heart disease and stroke are more prevalent in certain states, such as Louisiana and New Jersey, where women and men are more prone to these chronic conditions, respectively as shown in'Fig. 12"

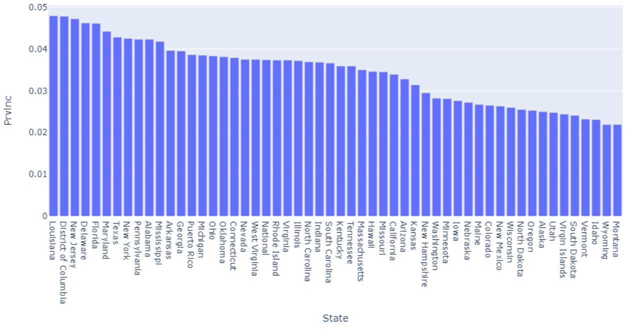

Fig. 12. Stroke Index for Females

The data also revealed that the state of Virgin Islands had the lowest prevalence of heart disease and stroke. This could imply that the healthcare system in Virgin Islands is more effective than other states at preventing and managing heart disease and stroke.

This analysis' findings are critical for healthcare providers, policymakers, and researchers to understand the distribution of chronic conditions in the United States. Targeted interventions to improve prevention, diagnosis, and treatment of heart disease and stroke can be implemented by identifying which populations and regions are most affected by these chronic conditions.

C. Visualizing Motor Vehicle Registration

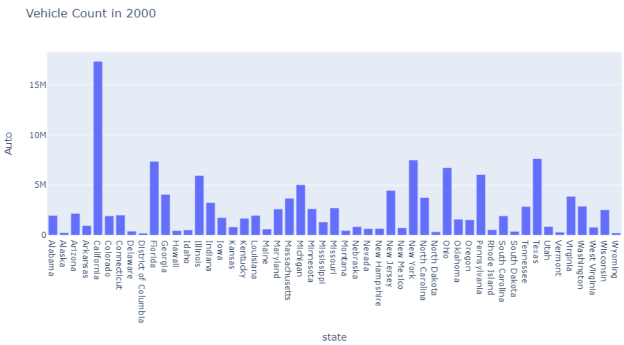

Fig. 13. Automobile Vehicle Count in the Year 2000

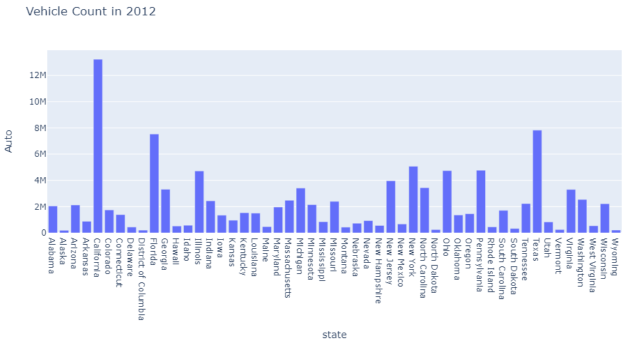

Fig. 14. Automobile Vehicle Count in the Year 2012

In order to compare the vehicle count, a quick visual plot is tooled after converting the default DataFrame to an integer using *‘as.type(int)*. The representation of vehicle entries in the year 2012 across all states in USA can is seen in the representation through plot in “Fig. 14”

In order to visualize the distribution of the vehicle types and their entries, the count of bus vehicles in the year 2012 across all states in the USA is represented in “Fig. 15”.

To understand the distribution of truck vehicles in the year 2012 across all states in the USA, a graphical representation is exhibited in “Fig. 16”.

To examine the data pattern and visualization between the two intervals of years 2000 and 2012, a graphical representation is exhibited which shows the count of vehicles in the year 2006 across all states in the USA. The DataFrame is converted to integer from default in order to ensure proper deployment of data and its representation in “Fig. 17”.

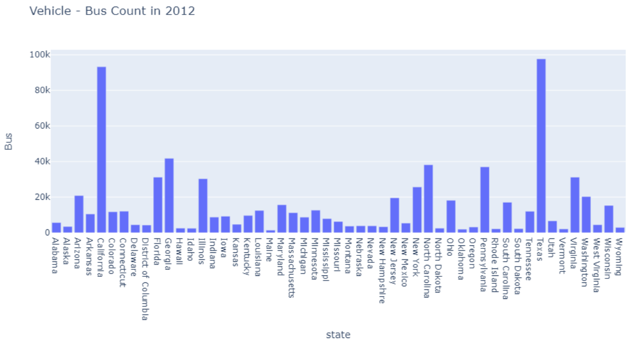

Fig. 15. Bus count in the Year 2012

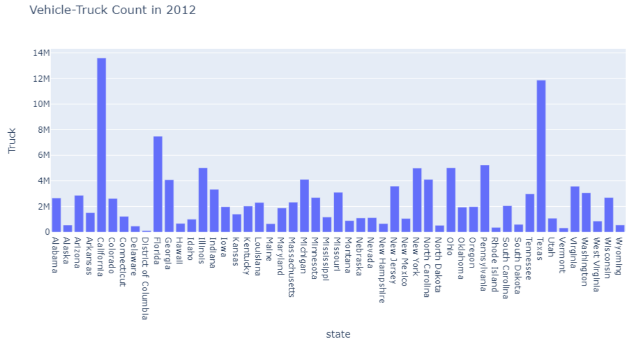

Fig. 16. Truck count in the Year 2012

The *‘reset_index’* reset the index of DataFrame ‘*averages’*. The step was taken to accommodate proper retrieval and optimization of the data. Studying the huge dataset, indexing is performed to ensure the efficient functioning of the query or any other task. Using *‘reset_index’*. A new Data Frame was created consisting of a default index and the *‘state’* column was employed as a regular column.

The three constituents of *‘auto’ ‘bus’* and *‘truck’* hold entries which were huge in number and hence accessing it as an integer will be a load. To counter this circumstance, using *as.type()* the data was grouped by state and the entries were converted from *int* to *float*. This ensured easy representation of DataFrame for further computation and visualization. In order

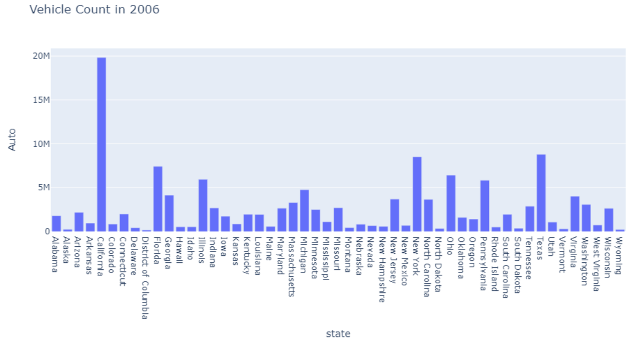

Fig. 17. Automobile count in the Year 2006

for easy display of the data, the entries in the columns were divided by 1000000 and the data was represented as a count of Automobiles, buses and trucks in a million across the state in the USA of all years 2000 to 2012.

A quick representation of the count of vehicles throughout the year 2000 to 2012. The entries are displayed in millions in "Fig. 18".

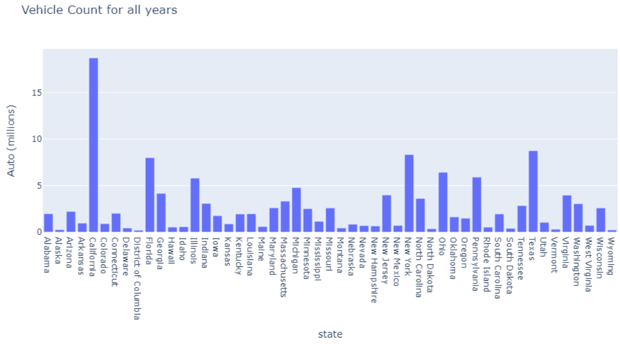

Fig. 18. Automobile count for all Years (in millions)

The entries being massive in quantity are represented by the median in order to be exhibited for visualization to identify trends and summary. Generally, averages are used as data to perform visualization and summary however while dealing with massive datasets it is vital to choose median over average since median entries are less affected by outliers which often skew the mean making it less accurate to represent. "Fig. 18" represents the median count of vehicles across all states in the USA.

D. Post Merge Analysis

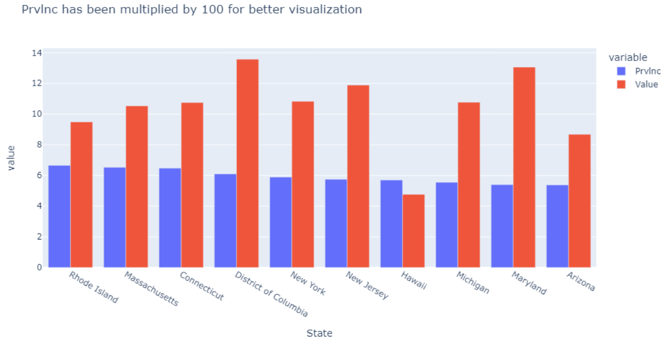

Fig. 19. Top 10 states with Asthama

The analysis of the three datasets provides valuable information about the relationship between air quality measures, specific chronic conditions, and motor vehicle registrations in the United States. The findings indicate a link between high AQI and chronic diseases such as asthma, COPD, ischemic heart disease, and stroke in several states. Furthermore, the number of registered vehicles in a state significantly contributes to high AQI and related diseases.

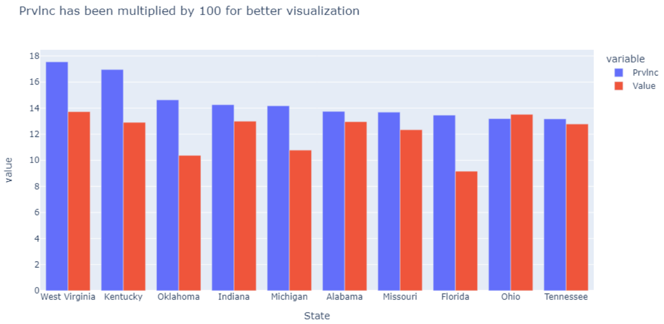

Fig. 20. Top 10 states with COPD

The findings show that the District of Columbia and Maryland have a significant link between high AQI and asthma, whereas West Virginia, Ohio, Indiana, Alabama, Kentucky, and Tennessee have a link between high AQI and COPD disease. West Virginia, Alabama, and Kentucky, on the other hand, share high AQI and ischemic heart disease, while Pennsylvania, the District of Columbia, Maryland, and Alabama share high AQI and the number of strokes.

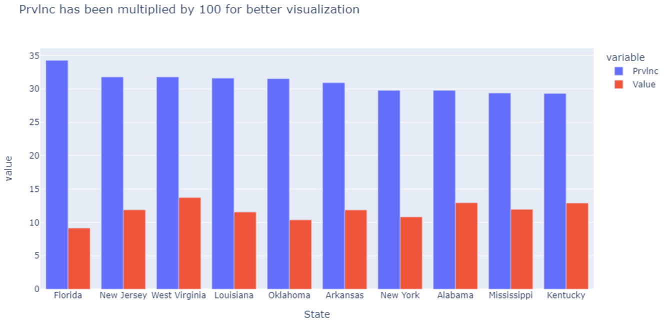

Fig. 21. Top 10 states with IHD

The data also revealed that the states with the highest AQI have registered more than 20 million vehicles, which may contribute to AQI and the diseases it causes. In contrast, states with the lowest AQI have registered approximately 8.9 million vehicles, which helps to reduce AQI and the diseases caused by it. According to the findings, the number of vehicles registered in a state plays a significant role in contributing to AQI and related diseases.

The analysis also revealed that Florida has a high number of AQI-related diseases but a lower number of vehicles, implying a very small significant relationship between vehicle count and AQI-related diseases. In contrast, while California has a high vehicle population, the number of AQI-related deaths has decreased due to the implementation of natural gas fuels and public transportation.

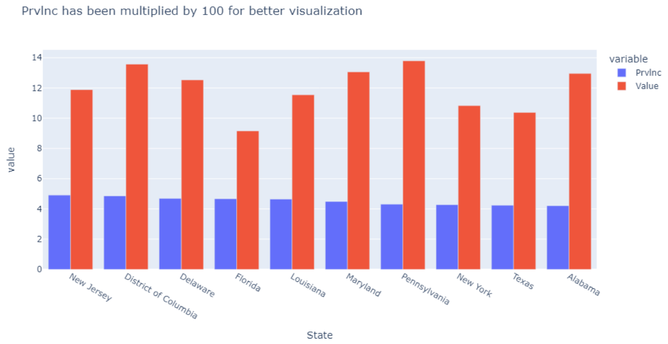

Fig. 22. Top 10 states with Stroke

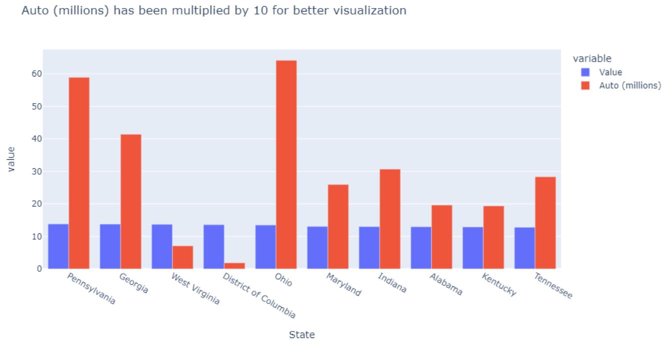

Fig. 23. Top 10 states with high AQI

Conclusion

Data from Centers for Disease Control (CDC) and Centers for Medicare & Medicaid Services (CMS) and Vehicle registration from the Vehicle Registration Distribution was collected to analyse how vehicle emission has been contributing to air pollution which eventually results in severe chronic diseases such as asthma, COPD, ischemic heart disease, and stroke. The findings suggest that there is a significant association between high AQI and chronic diseases. Moreover, the number of registered vehicles in a state plays a crucial role in contributing to the high AQI and related diseases. After careful discussion and examination of the three datasets, insightful information was obtained from our analysis.

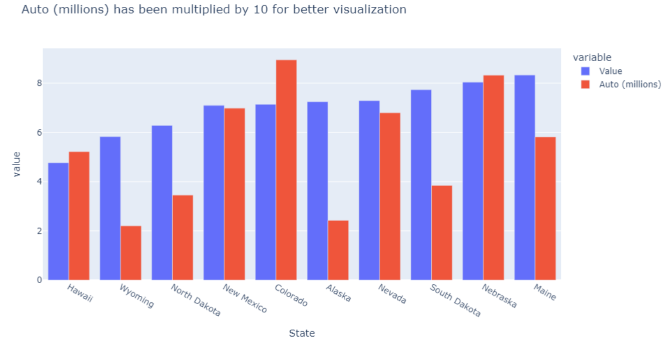

Fig. 24. Bottom 10 states with low AQI

1) **How does AQI contributes to exacerbate chronic diseases?**
The findings indicate a strong link between high AQI and asthma in the District of Columbia and Maryland. Similarly High AQI levels have also been linked to COPD disease in states such as West Virginia, Ohio, Indiana, Alabama, Kentucky, and Tennessee. The analysis also showed that ischemic heart disease is common in states with high AQI, such as West Virginia, Alabama, and Kentucky and in contrast to this it was observed that high AQI is associated with a higher incidence of stroke in Pennsylvania, the District of Columbia, Maryland, and Alabama.

2) **How does vehicle emission has contributed to AQI?**
The findings indicate a strong link between high AQI and asthma in the District of Columbia and Maryland. Similarly High AQI levels have also been linked to COPD disease in states such as West Virginia, Ohio, Indiana, Alabama, Kentucky, and Tennessee. The analysis also showed that ischemic heart disease is common in states with high AQI, such as West Virginia, Alabama, and Kentucky and in contrast to this it was observed that high AQI is associated with a higher incidence of stroke in Pennsylvania, the District of Columbia, Maryland, and Alabama. The analysis revealed that Florida has a high incidence of AQI-related diseases, but it has fewer vehicles compared to other states. This suggests that there is a slight significant correlation between the number of vehicles and AQI-related diseases. On the other hand, California has a high number of vehicles, but the number of deaths related to AQI is lower.

The overall analysis suggests that the analysis of air quality measures, chronic disease prevalence, and motor vehicle registrations in the United States has provided valuable insights into the relationship between high AQI levels and chronic diseases such as asthma, COPD, ischemic heart disease, and stroke. The findings imply that vehicle emissions contribute to high AQI levels and related diseases in some states. However, there is a fractional link between the number of vehicles and AQI-related diseases. The findings of this study can be used to guide public health policies and interventions aimed at reducing the negative effects of air pollution on public health.

Acknowledgment

We would like to express our gratitude to Professor Noel Cosgrave for clearing up all of our doubts throughout our coursework and guiding us at every step.

References

[1] ”Air Quality Monitor” The U.S Embassy and Consulate in Vietnam. https://vn.usembassy.gov/embassy-consulate/embassy/air-quality-monitor/ (accessed Apr. 20, 2023)

[2] ”Global Health Observatory data Repository” World Health Organization. https://apps.who.int/gho/data/node.main.152?lang=en (accessed Apr. 20, 2023)

[3] ”Global Burden of Air Pollution” Institute for Health Metrics and Evaluation. https://www.healthdata.org/infographic/global-burden-air-pollution (accessed Apr. 20, 2023)

[4] ICCT, “New Study Quantifies the Global Health Impacts of Vehicle Exhaust,” The International Council on Clean Transportation, 2018. [Online]. Available: https://theicct.org/publications/new-study-quantifies-global-health-impacts-vehicle-exhaust. (accessed Apr. 20, 2023).

[5] D. Alexander and H. Schwand, “The Impact of Car Pollution on Infant and Child Health: Evidence from Emissions Cheating,” NBER Working Paper Series, no. 25489, 2019, doi: 10.3386/w25489.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Trans. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.