EXPLORE TRENDS IN EMISSIONS, POLLUTION, AND HEALTH METRICS ACROSS U.S. STATES WITH DATA ANALYTICS & VISUALIZATION.

Pratik Sunar
x23292512
MSCDAD_A
National College of Ireland

Shivansh Bhatnagar
x23237252
MSCDAD_A
National College of Ireland

Malav Naik
x23271779
MSCDAD_A
National College of Ireland

***Abstract*— The intersection of various environmental pollution and public health is a deep area of research, especially the mutual relationship between air pollution, emissions, and cardiovascular health. This comprehensive report analyzes the trends in air pollution from the year 2000 onwards showcases the correlation with emissions from various sectors in different states, and examines the patterns in heart health to show insights into the relationship between pollution levels and cardiovascular diseases. By integrating data analytics and visualization techniques, this study aims to broadcast insights into how Industries and air Pollution target public health, framing strategies and policy decisions to mitigate these risks.**

Keywords—Data, Analytics, Pollution, Heart, Health, Insights, Dashboard, Dagster, Pipeline, SQL, Python, Visualization.

1) INTRODUCTION

This report is based on three separate datasets—Pollution, Heart Care and Hospitals, and Emissions—integrating their interrelationships and connections with health, environmental quality, and socioeconomic variables. We Aimed to Analyze and Visualize a Dashboard to show the Impact of the Emission of Different types of fuel, demonstrate which sectors are contributing to these emissions then analyze the yearly/seasonal trend of air pollution in the same US states to show what the factors that are contributing in Air Quality Index (Just Like O3 Mean, CO3 Level) then We have performed comprehensive analytics of Healthcare dataset and got insights of Cardiovascular Health. And presented how Environmental pollution and emissions contribute to the Cardiovascular health of people in the USA.

The novel Questions we are solving are stated below:-

The research seeks to answer the following questions:

- What are the trends in air pollution from the 2000s, and how do they correlate with emissions data?
- Can patterns in heart health data reveal associations between pollution levels and cardiovascular diseases? Ease of Use.

2) RELATED WORK

Many studies have investigated the health effects of air pollution, especially cardiovascular diseases. Pope et al. (2002) showed that long-term exposure to PM2.5 was positively related to mortality from cardiovascular diseases [11]. In a similar study, Miller et al. (2007) found that exposure to higher levels of air pollution [2] was associated with an increased risk of heart disease in women. These studies have really outlined the health implications of air pollution. However, most of the previous studies have been limited either by geographical scope or breadth of health variables considered. Our analysis extends this by incorporating a broader range of health metrics and investigating state-wise variations within the United States. Further, Smith et al. (2015) showed that most states contribute a great percentage to the country's emissions and thus show the need for focused policies upon those states. Lee et al. (2018) investigated trends in emissions over time and consequent impacts on air quality

3) DATA PROCESSING METHODOLOGY:

a) Pollution

1. Description of the Dataset

The pollution dataset used for analyzing pollution trends contains yearly air quality data from various states in the United States spanning from 2000. It contains information on key air pollutants such as ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). The dataset has 170,018 rows and 20 columns.

The dataset was selected due to its comprehensive content of pollutants and its relevance to analyzing long-term trends in air quality in the real world. Which allowed to explore state-specific pollution levels and their evolution over time.

Dataset Reference:
https://www.kaggle.com/datasets/guslovesmath/us-pollution-data-200-to-2022

2. Data Processing Activities

2.1 Dataset Storage and Loading

- The raw dataset was provided in CSV format. It was first imported into PostgreSQL to work on a

structured database for efficient querying. PostgreSQL was chosen for its robustness in handling large datasets and its support for advanced SQL queries.

After storing in SQL, the database was extracted into a python environment using the "psycopg2" library to proceed with further analysis.

2.2 *Data Cleaning and Preprocessing*

- The data was explored and inspected for missing, null values and duplicates, which were handled using appropriate imputation techniques or dropped if insignificant.
- The dataset contained some redundant columns (e.g., Year, Month, which could be extracted from the Date field). While retaining these for further temporal analysis, proper care was taken to ensure consistency across derived and raw fields.
- All pollutant measurements were standardized to normalize form to uniform units where applicable to enable comparison of pollutant levels.

3. Data Processing Algorithms

The following algorithms and techniques were implemented:

- **Temporal Analysis**: Python's pandas and numpy libraries were utilized to compute rolling averages, annual trends, and seasonal variations in pollutant values.
- **Correlation Analysis**: Relationships between pollutants and AQI levels were explored using correlation matrices and scatter plots, implemented via the seaborn and matplotlib libraries

b) Health care and Hospitals

Description of the Dataset

The second dataset used in the project focuses on heart care and hospital data. It contains 10,000 records detailing patient demographics, health metrics, treatments received, and hospital information. The dataset includes attributes such as blood pressure, cholesterol levels, BMI, glucose levels, gender, and whether the patient has been diagnosed with heart disease.

This dataset was chosen to investigate potential associations between air pollution levels and cardiovascular health outcomes, a key objective of the study. Its structured format and inclusion of critical health parameters make it a valuable resource for exploring correlations between environmental and health variables

Dataset reference:
https://www.kaggle.com/datasets/samwash94/heart-care-and-hospital-data

2. Data Processing Activities

2.1 Dataset Storage and Loading

- Original Data in JSON Format: The dataset was originally provided in **JSON format**. It was stored in a **MongoDB** database using "pymongo", which was selected for its ability to efficiently handle JSON-like hierarchical data and its flexibility for dynamic querying.

2.2 Data Cleaning and Preprocessing

- **Year and Month Extraction**: First, the treatment date column, which was in date time format, was processed such that it yielded two new columns representing the year and month, respectively. This was an effort toward enabling temporal trends in treatments and heart disease prevalence, matching the time-based processing of the pollution dataset.
- **Correlation Analysis**: A correlation matrix was generated for all numerical variables (e.g., blood pressure, cholesterol, BMI, and glucose levels) to identify significant relationships and trends within the data. This analysis complements the correlation exploration conducted in the pollution dataset.
- **Null Value Checks**: The dataset was inspected for missing values, and it was confirmed that there were no null entries, ensuring data completeness.
- **Duplicate Removal**: Duplicate records were identified and removed to ensure that the analysis was not skewed by redundant data.

3. Data Processing Algorithms

- **Correlation Matrix**: Computed correlations among health-related numerical variables to identify significant patterns or risk factors related to heart disease. This parallels the pollutant correlations explored in the pollution dataset.
- **Categorical Grouping**: The creation of age_group enabled grouping and comparative analysis across different age brackets, providing insights into demographic trends.

c) Emission

1. Description of the Dataset

It includes carbon dioxide emissions classified by state, sector, and type of fuel in the 2000s. It permits detailed insight into the analysis of which sectors are really contributing toward greenhouse gas emissions under various fuels-petroleum, natural gas, and coal.

Dataset reference:
https://www.kaggle.com/datasets/samwash94/u-s-carbon-dioxide-emissions

2. Data Processing Activities

2.1 Dataset Storage and Loading

- This dataset was provided in JSON format and stored in a MongoDB database. MongoDB was chosen for its efficient handling of semi-structured data and dynamic querying capabilities.

3. Data Cleaning and Preprocessing

- **Null Value Checks**: The dataset was checked for missing values. No null values were found, ensuring the dataset was complete for analysis.
- **Duplicate Removal**: Duplicate entries were identified and removed to maintain the integrity of the analysis.
- **Top Emitting States Identification**: The dataset was filtered by grouping emissions values (value) by state-name and summing the total emissions for each state. The ten states with the highest total emissions were identified using the **largest** method.

4. Alignment with Pollution Dataset

- The temporal component (year) was retained to align emissions data with the pollution dataset, allowing for time-based comparative analyses.
- Sectoral and fuel-type breakdowns facilitated in-depth analysis of emissions sources, complementing pollutant-specific insights from the pollution dataset.

5. Data Processing Algorithms

- **Aggregation of Emissions by State**: The dataset was grouped by state-name and value was aggregated to identify states with the highest emissions. This parallels the state-level analyses conducted in the pollution dataset.
- **Sectoral and Fuel-Based Insights**: Emissions were analyzed across sector-name and fuel-name to uncover trends in carbon dioxide emissions by economic activity and energy source.

4) CHOICE OF TECHNOLOGIES

MongoDB / PostgreSQL

MongoDB was selected for storing the JSON dataset due to its compatibility with the hierarchical structure of the data and its ability to handle unstructured or semi-structured data formats efficiently. The document-oriented nature of MongoDB facilitated flexible data exploration and extraction and To Store Structured Data(CSV) we have used PostgreSQL database.

ETL (Extract Transform Load)

We have used ETL process to commute the data in between different sources. For Extracting semi-structured file(Json) File We have used MongoDB then fetch file to python and store data in PostgreSQL. And for Extracting CSV data we have used PostgreSQL and Load data to SQL.

- Extract:- Collected data from MongoDB / SQL to Python.
- Transform:- Performed Data Cleaning and Transformed Data.
- Load:- Loaded transformed Data, performed further analytics and later stored it to PostgreSQL.

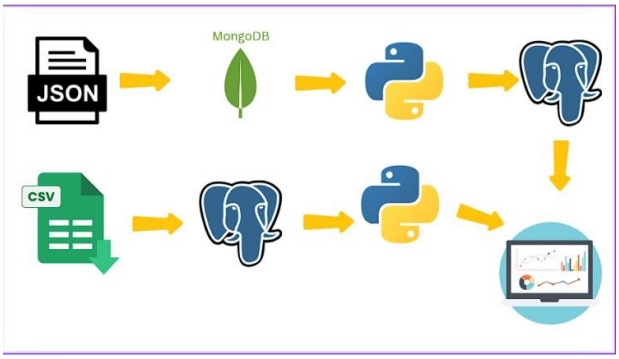


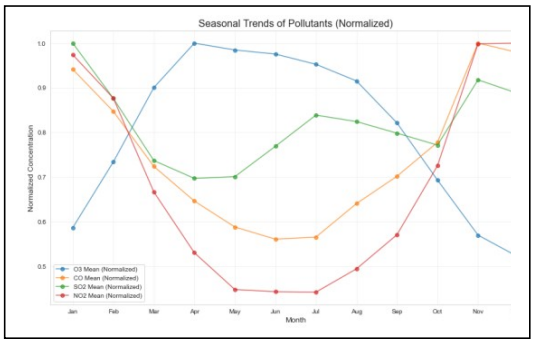
Fig: ETL(Extract Transform Load)

5) DATA VISUALIZATION METHODOLOGY

A. Pollution Dataset

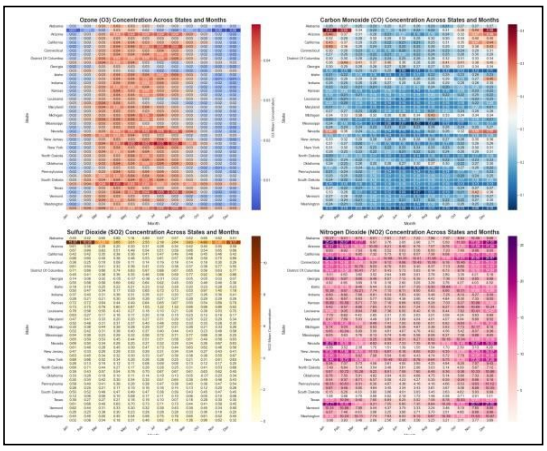
- Seasonal Trend of Pollutants

The line graph, "Seasonal Trends of Pollutants (Normalized)" shows the normalized concentration of O3, CO, SO2, and NO2 during the year. O3 starts high in January, decreases to a minimum in October, and increases again slightly. CO declines from January to April, remains steady during summer, and increases toward December. SO2 is lowest in May and June but increases by December. NO2 concentration is highest during winter, particularly during January and December of the year, whereas the lowest recorded months are from May to September. This is one such visualization of the seasonal fluctuation of these pollutants, important for environmental and public health strategies.[3]



- **Heatmap of Different pollutants**

Ozone (O_3)



1. **Peak Months**:

Concentration values are higher mostly between **May and August,** this reflects the probable trend connected with high temperature and sunshine in promoting ozone formation.

States like California, and Nevada show consistently high levels of O_3 during these months[8].

---

a) Carbon Monoxide (CO)

1. **General Trends**:

Most states show a higher concentration of CO during winter months, from November to February. This may be attributed to increased fuel combustion for heating and increased vehicular emissions during colder weather.

2. **High Concentration States**:

California, Arizona, Alabama, and Nevada all maintain high CO concentrations in general throughout the year, higher than 0.3 levels.

*b) Sulfur **Dioxide (SO_2)***

1. **Regional Variability**:

States like Alabama, Indiana, and Ohio have notably high SO_2 concentrations year-round, with monthly averages exceeding 0.6–1.0. These states likely have significant industrial activity.

2. **Peak Months**:

SO_2 levels seem to increase in the winter months, possibly due to increased coal burning for heating in those industrial areas.

---

c) Nitrogen Dioxide (NO_2)

1. **Consistently High Levels**:

The highest concentrations of NO_2 are typically seen in highly urbanized states with considerable vehicular emissions, often above 15–20 µg/m³ in states like California, Texas, and New York.

2. **Seasonal Peaks**:

NO_2 concentrations seem to be higher in winter, probably because of thermal inversions that trap the pollutants and increase energy consumption.

B. Heartcare and Hospital Data Dataset

- Most active Treatment Months

The visualization shows monthly treatment counts from 2000's, with March, **A**ugust, and October consistently being the most active months. Peaks are notable in March (2019, 2020, 2023) and August/October (2021, 2022), suggesting some seasonal or cyclical trends. In contrast, February and June often show lower fluctuation across the year[7].

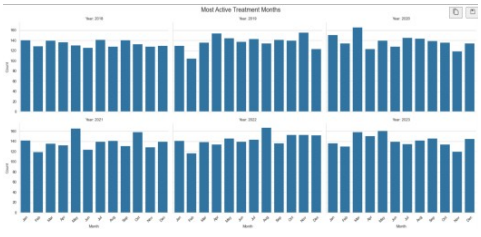


Fig : Yearly and Monthly active case

- State-Wise Heart Disease Cases

The bar chart demonstrates the counts of heart disease cases across different states. **Oregon and Michigan** show the highest number of heart disease cases, with values exceeding 120. On the other end, states like **Missouri, and New Jersey** have relatively lower-case counts, ranging below 90. The chart uses multiple colors to visually differentiate the states, but the general trend shows that most states have between 90 to 110 cases, with only a few outliers above or below this range.[5]

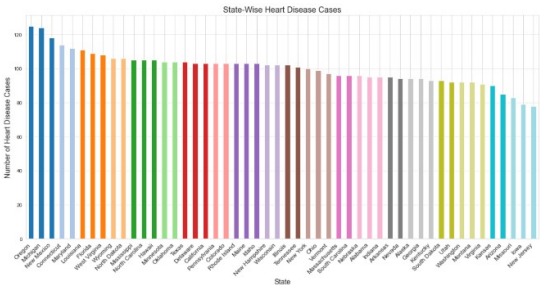


Fig : State-wise Heart Disease Cases

- State wise Average of Health Variables for Heart Disease Cases

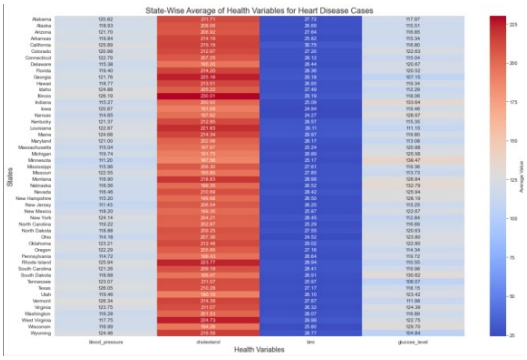


Fig: State-Wise Average of Health Variables for Heart Disease Cases

This Chart represents the Blood Pressure, Cholesterol, BMI and glucose levels of heart patients. Dark red colour represents higher levels of BP levels among states, similarly, dark colour in other columns represents a higher level of Cholesterol, BMI, and Glucose level.[6]

- Correlation Matrix of Heart disease and Health Variables

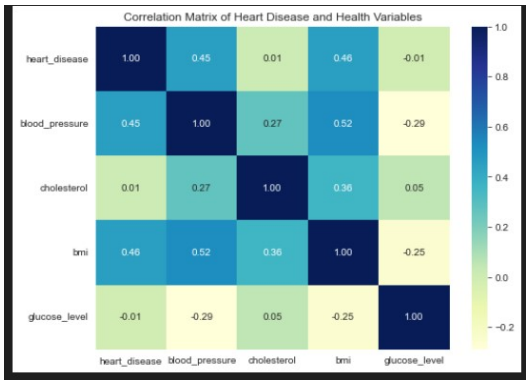


Fig: Correlation Matrix of Heart Disease and Health Variables

This Correlation Matrix Clearly shows a good and positive correlation of Heart Disease with Blood Pressure(0.45) and BMI(0.46). Also, BMI and Blood pressure also have strong correlation (0.52) while Cholesterol has very minor correlation, which states. Heart Disease has Major impact of Blood Pressure problems and bad BMI while glucose levels is not impact it[10].

Average Health Score by Top 20 States and Age Category

The below bar chart visualizes the average health scores across different age groups for the top 20 states. These graphs clearly plot the story of the health scores of different age groups in different states. As the Analyzed age group of 0-18 have the best health score in Illinois while West Virginia has the lowest for this particular age group. This visualization highlights the need for age-specific health policies to address the varying health care outcomes across different states of the USA.

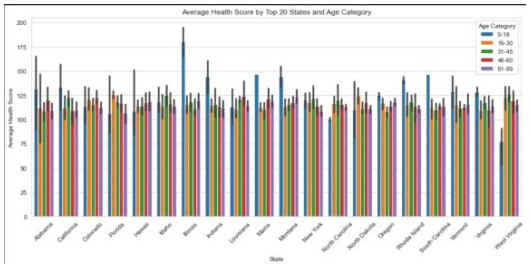


Fig: Average Health Score of States by Age-Group

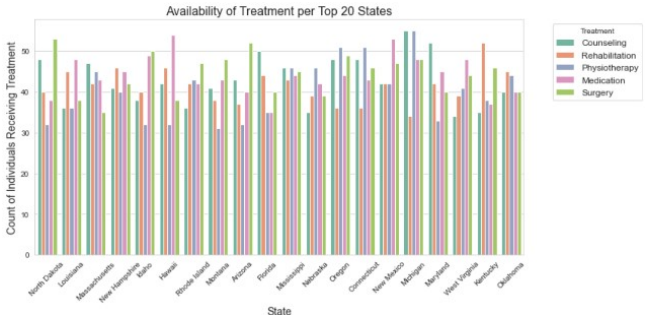


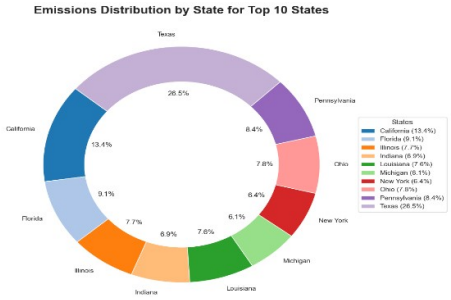
Fig: Availability of Treatments in States

C. Emission

- Total Emissions by Top 10 States

The bar chart shows the values of emissions for the ten states in the United States with the highest level of emissions. Texas has the highest value of emissions, far greater than the rest of the states, followed by California. The rest of the states-Florida, Michigan, New York, Ohio, and Pennsylvania-show a somewhat lower and closer value of emission. This chart highlights that efforts for the reduction of emissions should be made for these top-contributing states to get substantial improvements in the environment.[1]

- Emissions Distribution by States.



-

Fig: Emissions Distribution by States

This Chart depicts the distribution of the percentage of the United States' total emissions contributed by the top ten states. The leading state is Texas, contributing about 26.5%, while California stands second with 13.4%. While some other states, including Florida, Illinois, Indiana, Louisiana, Michigan, New York, Ohio, and Pennsylvania each contribute within the range from 6% to 9%. Texas and California dominate the overall levels from where efforts may most efficacious for the reduction of the trend[4].

- Fuel Type Distribution by States

Fig: Fuel Type Distribution in Top 10 States

The bar chart plots out the values of each fuel type emission - All Fuels, Coal, Natural Gas, and Petroleum. This plot shows Texas has ranked highest for top emissions followed by California. Interestingly, Natural Gas and Petroleum are the major pollutants in most of the states, with Texas having extremely high Natural Gas. This is a striking representation of the dominant drivers of emissions, especially from such fuel types, and underlines the need for targeted environmental and energy policies in the big states.

6) COMBINING VISUALIZATIONS INTO A DASHBOARD

- Environmental and Health Insights Dashboard

Generate a combined Dashboard using the “dash” and “plotly” libraries in Jupyter Notebook. The Environmental and Health Insights Dashboard provides a comprehensive analysis of emissions, pollution, fuel distribution, and health metrics across U.S. states. Designed to explore trends of Pollution Emissions impacting on Cardiovascular Health, it enables policy and decision-makers to make data-driven decisions for addressing environmental and public health challenges effectively[9]. Below is an explanation of the Dashboard and combined insights of all three datasets:-

1. Fuel type distribution

This Chart describes the proportion of fuel consumed including all types of fuel types in our study like, Petroleum, Natural Gas, Coal. This is the distribution by state.

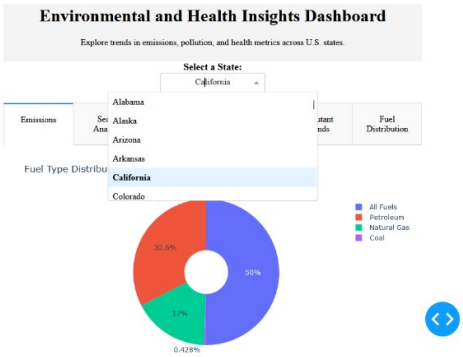


Fig: Fuel type Distribution

2. Sector Wise Emissions

This Bar graph is showing State-wise consumption of different type of fuels in different Sectors over time. In the next graphs, we will compare what is the impact of this state-wise fuel consumption on Air Pollution and Cardiovascular Health.

Fig: Sector Wise Emissions

3. Fuel Type Distribution of Year

We are Now showing Fuel type distribution by year. On x axis there is counts of Emission while on y axis we have plotted Year. Different color coding defines type of fuel on basis of state.

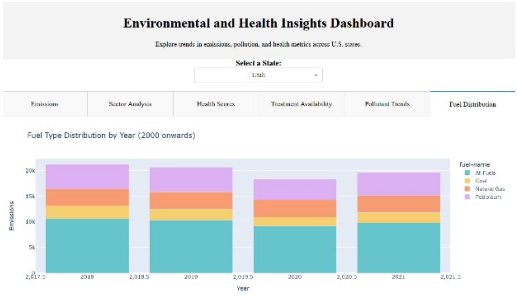


Fig: Fuel Type Distribution of Year

4. Yearly Trends of Pollutants

This Graph is plotting state-wise yearly trends of pollutants on basis of O3 Mean, CO Mean, SO2 Mean, NO2 Mean. These are the factor to determine Air Quality Index(AQI).

Fig: Yearly Trends of Pollutants

5. Health Score of States

After gathering above pollution and emission insights we have plotted below graph to calculate health score **(Blood_pressure + BMI + Cholestrol + Glucose_Level / 4)** by different states, to show impact of emissions and pollution on health of people.

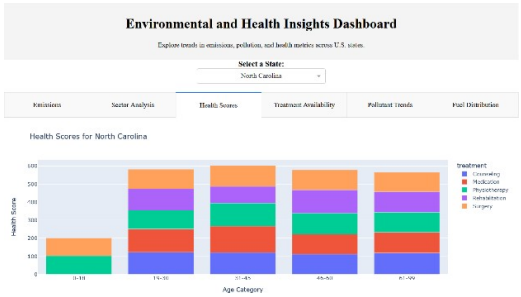


Fig: Health Scores of States

6. Treatment Availability in States

In Below Bar Graph We are showing the availability of treatment per states which comprises of Physiotherapy, Surgery, Rehabilitation, Counselling, Medication.

Fig: Availability of Treatment in States

7. Dashboard Conclusion

This dashboard integrates pollution data, fuel trends, and healthcare metrics to provide a practical view of environmental and public health performance. Where we have worked on 3 different datasets of Pollution, Emissions, and Cardiovascular Health of USA. First We have analysed which fuels combustion are most frequent in USA and what are the industries which are consuming those fuels and emitting gases yearly from Year 2000. Then We have analysed Yearly pollution trends in different states on which basis we have concluded the state with pollution and average people with Health Score(Less than 100 are Unfit / Between 100-160 are with Moderate health / More than 160 are Fit). Then we have analysed the availability of treatment state-wise.

7) RESULTS AND EVALUATION

Our Research concludes that 2022 is the year in which pollution level are at the peak on an average by calculating measure of Air Quality Index, O3 Mean in which November is the most polluted month and Colorado, and Alaska are the most polluted states. Texas, California and Illinois are the states that consume most All fuels(Coal, Petroleum, Natural Gas) and produce emissions. Oregon, Michigan, and Illinois are the states in which there are the most number of Heart Disease cases which is analysed by health score and Health Metrics. Illinois has the fittest youth(0-18), while Texas and Florida exhibit higher averages in Health Metrics which means the average population are not that fit. And Michigan is winning in giving treatment to the public in all sectors. Overall Texas is the polluted state has the Worst Health Metrics and doesn't stand in the top 20 statements with treatment facilities.

These findings demonstrated the clear picture that environmental factors, such as emissions and pollution, are directly correlated with cardiovascular health of people. The interactive dashboard allows for further exploration and validation of these results, offering a dynamic platform for user to engage with the data.

8) CONCLUSIONS AND FUTURE WORK

This project underscores the critical importance of examining the relationship between environmental factors and public health. The findings suggest that states with higher emissions and pollution levels also exhibit higher instances of heart disease, highlighting the need for targeted public health interventions and policies to mitigate these risks. The insights gained from this study have significant decision making capability for policymakers and public health officials, providing a robust evidence base to inform strategies aimed at reducing environmental health risks.

Future Works: To Extend the analysis that includes data from other countries to validate the findings on a global scale, that would provide broader insights into the relationship between environmental factors and health.

9) REFERENCES

[1] Kanchan, Amit Kumar Gorai, and Pramila Goyal, A Review on Air Quality Indexing System, Asian Journal of Atmospheric EnvironmentVol. 9-2, pp. 101-113, 2015

[2] Miller KA, Siscovick DS, Sheppard L, Shepherd K, Sullivan JH, Anderson GL, Kaufman JD. Long-term

exposure to air pollution and incidence of cardiovascular events in women. N Engl J Med. 2007 Feb 1;356(5):447-58. doi: 10.1056/NEJMoa054409. PMID: 17267905.

[3] Air Quality Index [Online], Available: en.wikipedia.org/

[4] AQI Basics[Online],Available: www.airnow.gov/aqi

[5] Newman JD, Bhatt DL, Rajagopalan S, Balmes JR, Brauer M, Breysse PN, Brown AGM, Carnethon MR, Cascio WE, Collman GW, Fine LJ, Hansel NN, Hernandez A, Hochman JS, Jerrett M, Joubert BR, Kaufman JD, Malik AO, Mensah GA, Newby DE, Peel JL, Siegel J, Siscovick D, Thompson BL, Zhang J, Brook RD. Cardiopulmonary Impact of Particulate Air Pollution in High-Risk Populations: JACC State-of-the-Art Review. J Am Coll Cardiol. 2020 Dec 15;76(24):2878-2894. doi: 10.1016/j.jacc.2020.10.020. PMID: 33303078; PMCID: PMC8040922.

[6] Brauer M, Casadei B, Harrington RA, Kovacs R, Sliwa K; WHF Air Pollution Expert Group. Taking a Stand Against Air Pollution - The Impact on Cardiovascular Disease: A Joint Opinion from the World Heart Federation, American College of Cardiology, American Heart Association, and the European Society of Cardiology. Glob Heart. 2021 Jan 28;16(1):8. doi: 10.5334/gh.948. PMID: 33598388; PMCID: PMC7845468.

[7] Abed Al Ahad M, Sullivan F, Demšar U, Melhem M, Kulu H. The effect of air-pollution and weather exposure on mortality and hospital admission and implications for further research: A systematic scoping review. PLoS One. 2020 Oct 29;15(10):e0241415. doi: 10.1371/journal.pone.0241415. PMID: 33119678; PMCID: PMC7595412.

[8] Q. Zhang, GN. Geng, SW. Wang, R. Andreas, and HE. KeBin, “Satellite remote sensing of changes in NOx emissions over China during 1996– 2010,” Chinese Science Bulletin, vol. 277, pp. 2857-2864, August 2012.

[9] Elias Dabbas, Interactive Dashboards and Data Apps with Plotly and Dash: Harness the power of a fully fledged frontend web framework in Python – no JavaScript required , Packt Publishing, 2021.

[10] B. George, "Extended Abstract: Articulating Environmental & Human Health in the Rust Belt," 2023 IEEE International Professional Communication Conference (ProComm), Ithaca, NY, USA, 2023, pp. 121-122, doi: 10.1109/ProComm57838.2023.00030. keywords: {Climate change;Environmental management;Government policies;Environmental monitoring;Health and safety;Human factors;Manufacturing industries;Public healthcare;Discourse analysis;environmental communication;environmental justice;environmental policy},

[11] Pope III CA, Burnett RT, Thun MJ, et al. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *JAMA.* 2002;287(9):1132–1141. doi:10.1001/jama.287.9.1132