Geospatial Analysis of U.S. Energy Infrastructure

Website

https://sites.google.com/sdsu.edu/usinfrastructureanalysis/home

Video Link Capstone Project Video.mp4

Github: https://github.com/Arbazmohammad/BDA-600-Capstone-Project

Group Members

Arbazuddin Mohammad: Data Preprocessing, EDA, Visualization, Modeling, Website, Report
Email: amohammad1049@sdsu.edu

Dev Shah: Data Preprocessing, Spatial Analysis, Website, PPT, Video, Report
Email: dshah7971@sdsu.edu

San Diego State University

Course

BDA 600 Big Data Analytics Capstone Seminar

Instructor

Dr. Gabriela Fernandez

Dr. Ming-Hsiang Tsou

May, 2025

Abstract

In this project, we looked at the condition of infrastructure across the United States and tried to find out which areas are more at risk when it comes to natural disasters like floods, hurricanes, or earthquakes. We used tools like ArcGIS and Tableau to map and analyze things like bridge conditions, population data, and hazard zones. Our main goal was to figure out where the weak points are and who might be most affected, especially people living in low-income or rural areas. The project also shows how things like income, location, and natural risk come together to make some communities more vulnerable than others. Our findings can hopefully help decision-makers understand where improvements are most needed so they can plan better and make things safer for everyone.

Keywords- Infrastructure, Risk, ArcGIS, Natural Disasters, Inequality, Mapping, Public, Tableau, Machine Learning, Safety

Problem Statement

Across the United States, there are a lot of places where the infrastructure is either getting old or not strong enough to deal with natural disasters like earthquakes, floods, or hurricanes. This becomes a big problem because when something like a major storm hits, bridges can collapse, roads can get washed out, and power can go down for days. These kinds of events don't affect everyone the same way, people living in poorer areas or far from big cities usually have a harder time recovering or even getting help.

Even though there are reports about these issues, we felt like there wasn't a clear, big-picture view that combines infrastructure conditions with natural hazard risks and social factors like income and population. So, in this project, we tried to bring all that data together. We wanted to see which parts of the country are more exposed to danger and don't have strong infrastructure to handle it. By doing this, we hope to help planners, city officials, or anyone interested see where help is needed most.

Research Goals

The main goal of our project was to find out which areas in the U.S. are most at risk when it comes to weak infrastructure and natural disasters. We wanted to see how things like income, population, and location all connect to the condition of infrastructure and the types of hazards each area might face.

To break that down a bit more, our goals were:

Map out risk areas – We used ArcGIS to map places where there's a higher chance of natural disasters and compared that with data on bridges, roads, and other infrastructure.

Look at who's affected most – We compared infrastructure risk with population and income data to find out if certain communities (like low-income or rural ones) are hit harder than others.

Find patterns – We looked for any clear connections between bad infrastructure, hazard zones, and socio-economic factors to understand what's going on.

Share what we found – By putting everything into charts, maps, and dashboards, we hope others can use our work to make better decisions about where to invest in repairs or improvements.

Literature Review

Before starting our project, we looked at a bunch of research papers and government websites to understand what people have already said about infrastructure and risk. A lot of studies talk about how bad infrastructure can cause big problems during natural disasters, especially in areas that don't have much money or resources to begin with.

Infrastructure Risk

Infrastructure risk usually means how likely it is that something like a road, bridge, or power line will fail, especially during a disaster like an earthquake or flood. The Federal Emergency Management Agency (FEMA) has a lot of data showing which areas are more likely to face hazards. Other papers also talk about how older systems are more likely to break down and cost more to fix.

Mapping and Spatial Tools

GIS (Geographic Information Systems) tools like ArcGIS are really helpful for projects like this. We saw in other research that mapping risks using layers, like population, income, and hazard zones, can show where the biggest problems are. This helped us decide how to approach our project and what tools to use.

Social and Economic Factors

Another big theme in the research is that infrastructure problems don't affect everyone equally. Poorer areas often have older bridges or fewer resources to fix things. Some studies showed that people in rural areas or communities of color are more likely to live in places with bad infrastructure and higher disaster risk. That's something we wanted to explore more in our own project.

Gaps in the Research

While there's a lot of good info out there, we noticed that not many studies bring everything together, like infrastructure data, hazard maps, and social data all in one place. That's what we tried to do with this project.

Data Processing

Data Collection

To start the project, we had to gather a bunch of different datasets and then clean and organize them so we could actually use them together. A lot of this took time because not all data is formatted the same way, and sometimes there were missing values or weird column names.

First, we collected data from places like FEMA (for hazard risks), the U.S. Census (for population and income info), and DOT (Department of Transportation) for infrastructure data like bridges and roads. We also used some shapefiles and map layers from open government sources for ArcGIS.

After we had all the data, we cleaned it up by doing things like:

1. Removing columns we didn't need
2. Fixing missing values (or just removing rows if too much was missing)
3. Making sure all the locations had the right coordinates so we could map them

We also created smaller versions of the datasets focused on the areas we were interested in the most, like California and a few other high-risk states. This made it easier to run analysis and create visuals without crashing the tools we were using.

Once everything was clean, we converted the files into formats that worked better with ArcGIS and Tableau (like CSV or GeoJSON), and then we were ready to start making maps and charts.

Dataset Highlights

1. **Indicators**: We used indicators like infrastructure age, bridge condition scores, population density, and hazard risk levels (earthquake, flood, etc.). These helped us figure out which areas are most vulnerable.
2. **Mapping and Analysis Tools**: The data was made ready for use in tools like ArcGIS and Tableau so we could create interactive maps and spot patterns.
3. **Research and Planning**: Our cleaned datasets helped us compare regions and understand where infrastructure needs the most attention. This also gave us something visual to present to others, which was a big help.

Data Preprocessing

Before doing any visualizations or analysis, we had to clean the data. Some of the steps we followed:

- **Irrelevant Data Removal**: We trimmed down the datasets to only include the states or counties we were focusing on. In some cases, we were only looking at California or specific disaster-prone areas.
- **Handling Missing and Invalid Data**: Some records had missing values (like missing bridge condition scores). We either filled in what we could or just removed the row if it wasn't usable.
- **Error Correction**: A few fields had mismatched data types or typos, so we fixed those to make sure everything would load properly into our tools.

Data Cleaning

Once we got all our data files, we realized they weren't exactly ready to use. Some files had extra columns we didn't need, and others were missing important values like bridge conditions or population numbers. So before doing any analysis, we had to clean things up.

Removing Irrelevant Data: We started by deleting rows that didn't relate to our focus areas. For example, if we were only looking at California and Texas, we removed other states to make the files smaller and easier to work with.

Fixing Missing or Null Values: A few columns had missing info, especially in older datasets. In some cases, we were able to fill in missing values by checking against other data (like county-level averages). But if too much data was missing, we just dropped those rows completely so they wouldn't mess up our charts or maps.

Standardizing Columns: Some datasets used different names or units for the same thing, like "State_Name" vs "state" or using percentages instead of counts. We fixed these so that when we combined files, everything lined up properly.

Fixing Location Data: A few files had errors in the latitude and longitude columns, or were missing coordinates altogether. We used ZIP code shapefiles and some geocoding tools to fix or fill in missing location points.

Final Check: Before using the data in ArcGIS or Tableau, we loaded it into Excel or Google Sheets to do a final scan, just to make sure columns were labeled right and there weren't any weird outliers or formatting issues.

Dataset Segmentation and Extraction

We created smaller versions of the data focused on key regions. For example:

- One file just for California
- Another focused on areas with high FEMA hazard scores
- A national summary for comparisons

Data Transformation for Visualization

Optimization for Visual Tools: To make the data work better with ArcGIS and Tableau, we saved it in formats like CSV or GeoJSON. We also made sure the coordinates were correct for mapping. Some datasets needed us to convert ZIP codes into shapefiles so they could be visualized on a map.

Repository and Accessibility

Open Source Sharing: Once everything was ready, we saved the cleaned datasets into our shared drive and Google site.

Data Visualization

Once our data was cleaned and ready, we started making maps and charts to actually see the patterns we were talking about. This part helped a lot because sometimes just looking at numbers doesn’t tell the full story, but a map or graph can make the problem super clear.

Mapping with ArcGIS

We mapped all the power plants and transmission lines and visualized it in different ways below to get some clear insights off of it.

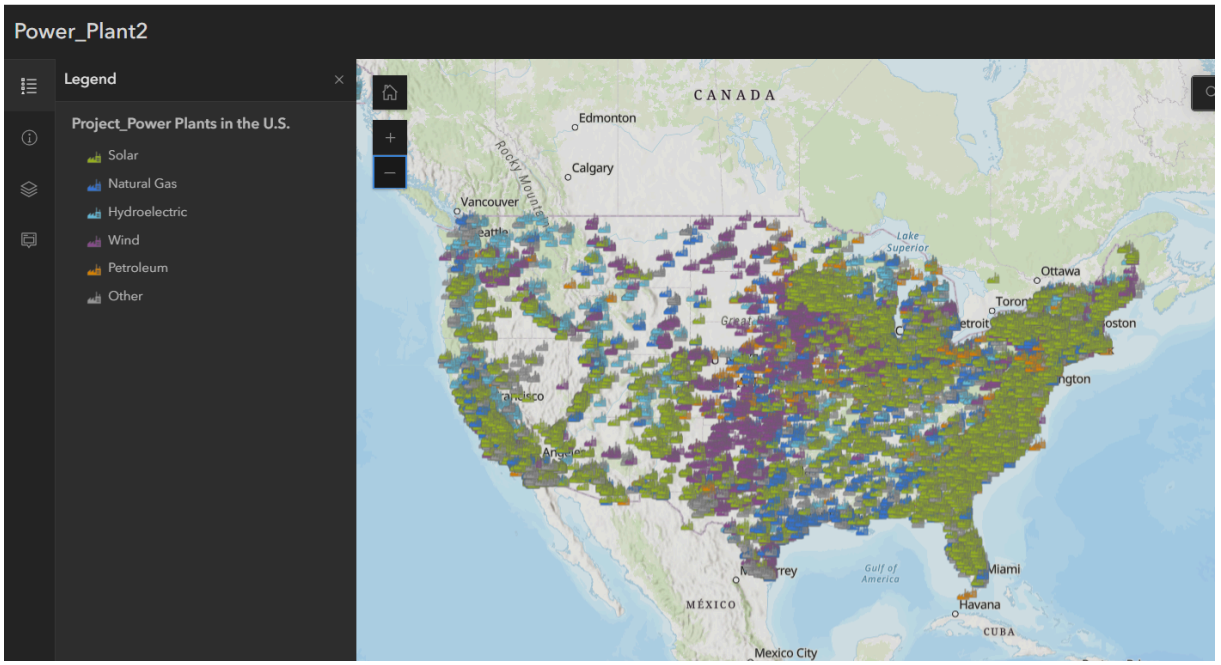


fig.1

Figure 1: Power plants in the US

The map provides a comprehensive visualization of power plants in US, here you can see that most of the power that has been generated across US uses the primary fuel as Solar(green coloured), then comes Natural Gas(Dark Blue), then it goes down the descending order of Hydroelectric, Wind, Petroleum and Other primary fuels used. We have in total of 20 primary fuel used to generate electricity in power plants.

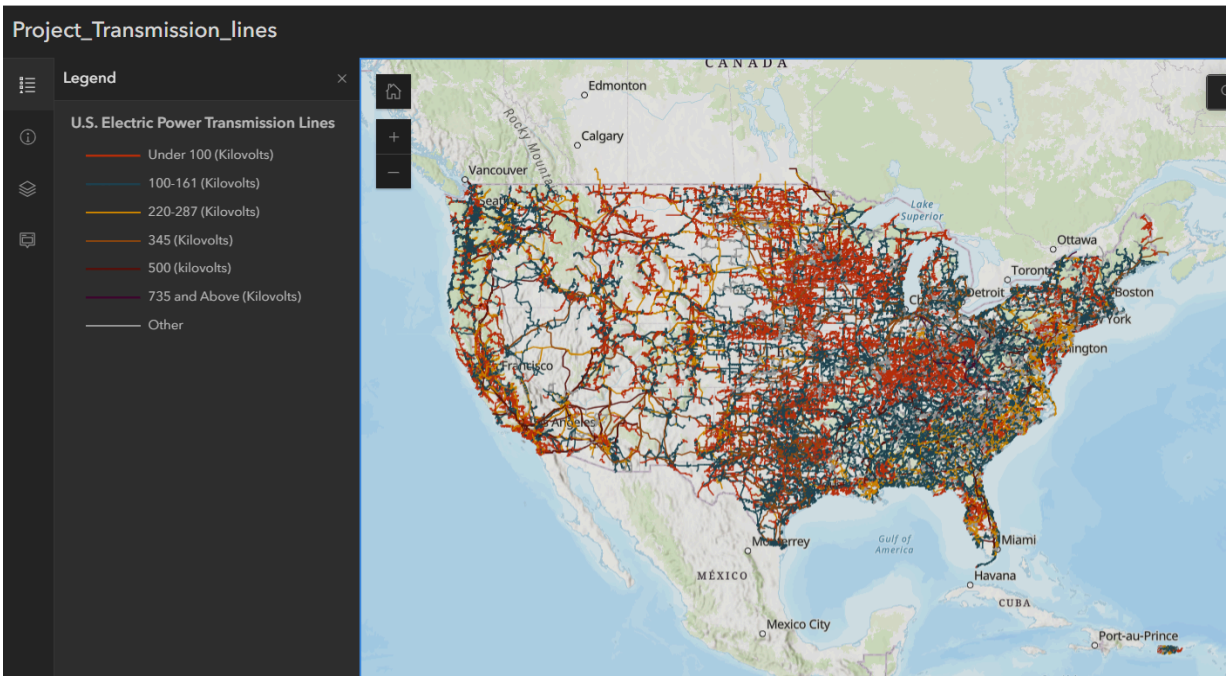


fig. 2

Figure 2: Transmission lines in the US

The map provides a comprehensive visualization of transmission lines in US, here you can see that most of the transmission lines that has been scattered across US carries Under 100KV which is in red colour, then comes 100-161KV in blue colour, then it goes down the descending order of 220-287KV, 345KV, 500KV, 735 and more KV.

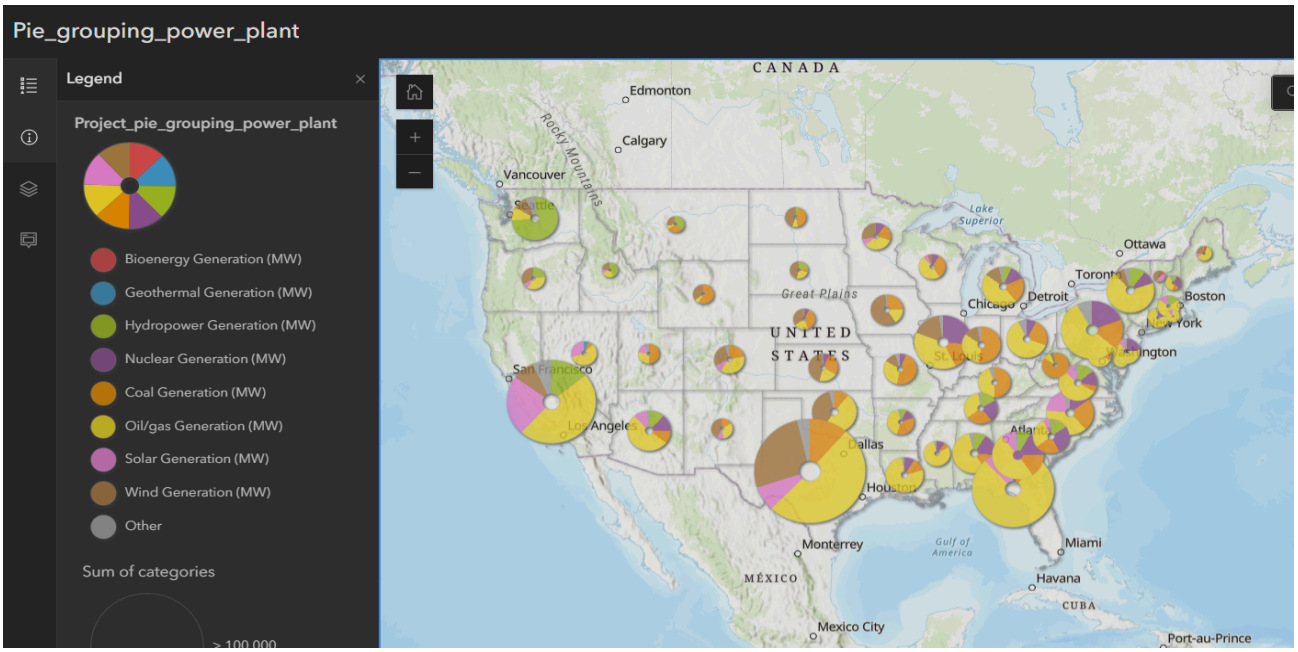


fig. 3

Figure 3: Pie chart distribution of Power plants in the US

This map visualizes the pie chart distribution of power plants in a way that each state showcases how the power distribution is divided amongst all the types of primary fuels that has been used to generate electricity here. As it is visible, Texas has the highest generation of energy which constitutes of more than 50% percent of Oil/Natural Gas primary fuel. In all the other states, it is visible that oil/Natural Gas constitutes of the maximum primary fuel that has been used in comparison of all the other primary fuels. So, Oil/Natural Gas is abundant everywhere.

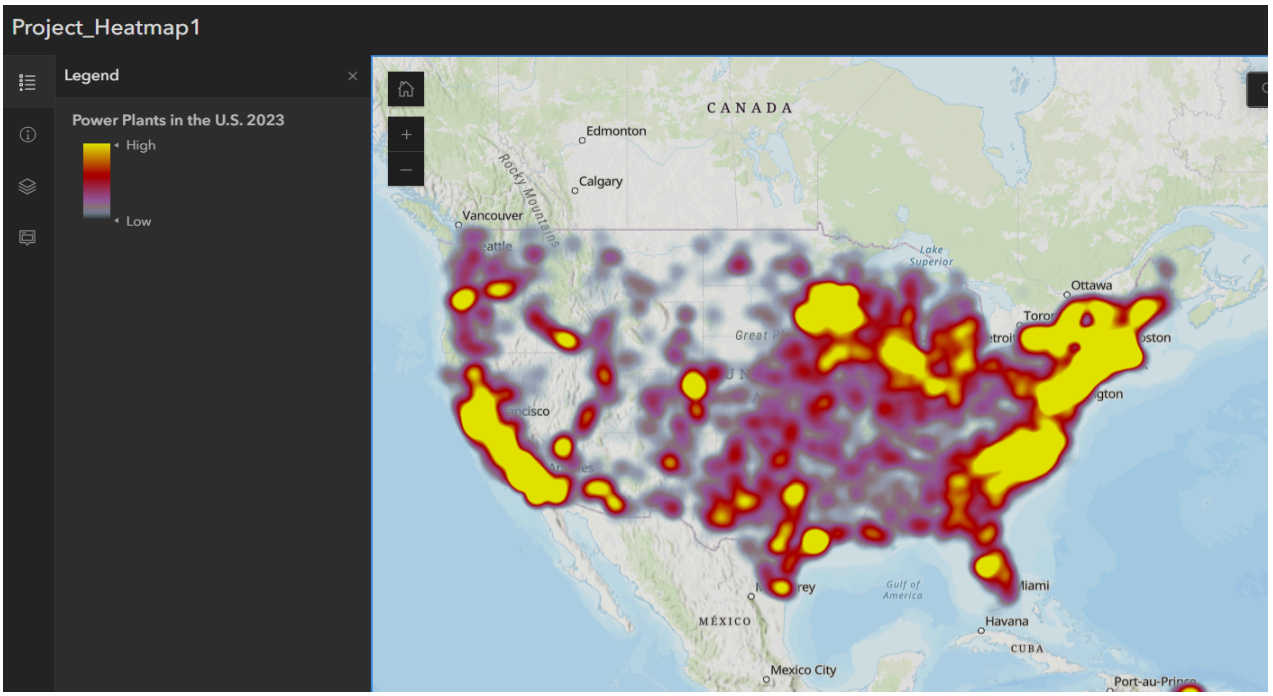


fig. 4

Figure 4: Power Plants Heatmap

This map visualizes the overall distribution of power plants in US. As we can see, the east side and west side of US has the highest generation of electricity across the US considering the maximum number of people live there. And, then it's scattered all the other areas across the US.

Visualization

To support our analysis, we used Tableau to create charts and dashboards that helped us explore relationships between infrastructure quality, natural hazards, and demographic factors. These visuals made it easier to compare states and see patterns that aren't always obvious in raw data.

fig. 5

Figure 5: Word Cloud Visualization for States

Utilizing word cloud visualization techniques, we created a graphical representation highlighting the states with the highest levels of power generation and gas emissions from power plants. The word cloud visually emphasizes states based on the magnitude of their energy output and infrastructure development.

The results revealed that Texas and Florida emerged as the leading contributors, showcasing significant power generation capacity alongside substantial energy infrastructure. These states not only dominate in total electricity production but also play a crucial role in the nation's overall energy landscape.

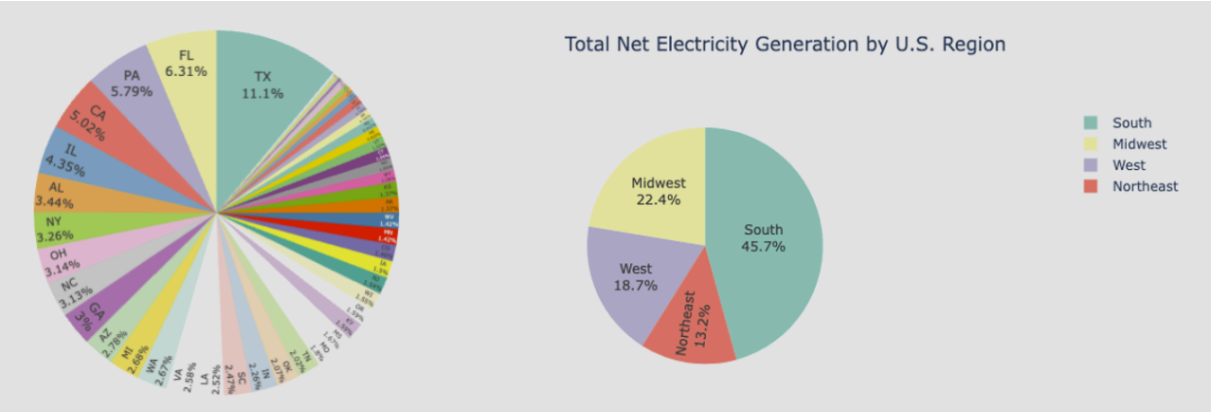


fig. 6

Figure 6: Pie chart visualization for states

The pie charts illustrates the distribution of total net electricity generation across U.S. States and regions. The South emerges as the dominant contributor, accounting for 45.7% of the nation's total power generation where Texas and Florida being major contributor, reflecting its expansive energy infrastructure and high demand. The Midwest follows with 22.4%, driven by a mix of traditional and renewable energy sources. The West contributes 18.7%, while the Northeast accounts for 13.2%, highlighting regional differences in generation capacity and energy consumption patterns. This regional breakdown provides crucial insights into where the nation's energy production is most heavily concentrated.

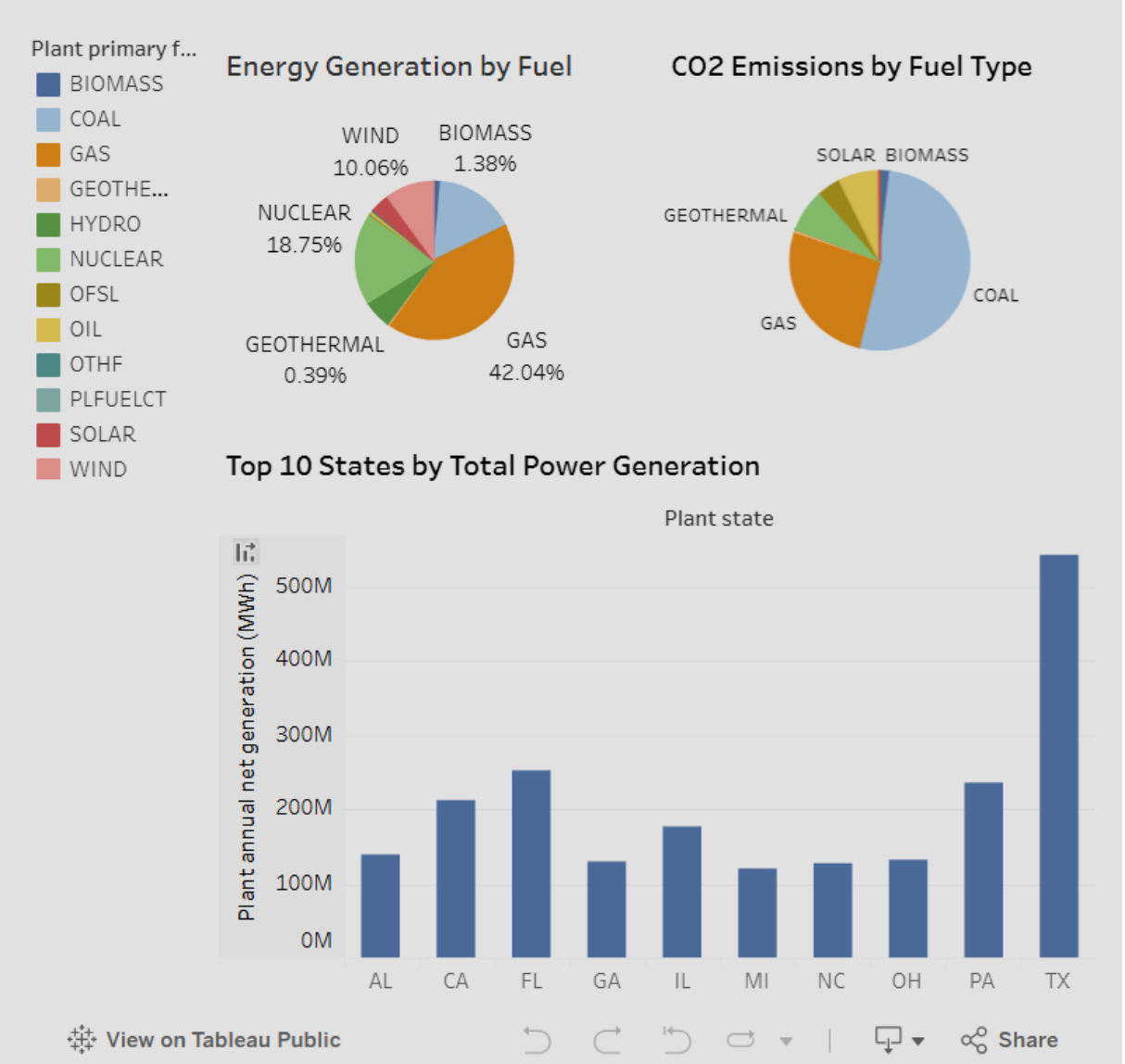


fig. 7

Figure 7: Tableau Dashboard

This dashboard offers a multi-dimensional view of the U.S. power generation landscape and related CO_2 emissions by fuel type.

Energy Generation by Fuel:

The first pie chart shows that gas is the dominant source of energy generation, contributing 42.04% of the total electricity, followed by nuclear (18.75%) and wind (10.06%). Renewable sources like solar and biomass make up smaller shares.

CO_2 Emissions by Fuel Type:

The second pie chart highlights that coal is the largest contributor to CO_2 emissions among all fuel types, despite not being the top source for power generation. This reflects the heavy environmental impact of coal usage compared to cleaner energy sources like gas and wind.

Top 10 States by Total Power Generation:

The bar chart identifies Texas (TX) as the clear leader in total power generation, significantly outpacing other states. States like Florida (FL), California (CA), and Pennsylvania (PA) also contribute heavily to the national energy grid.

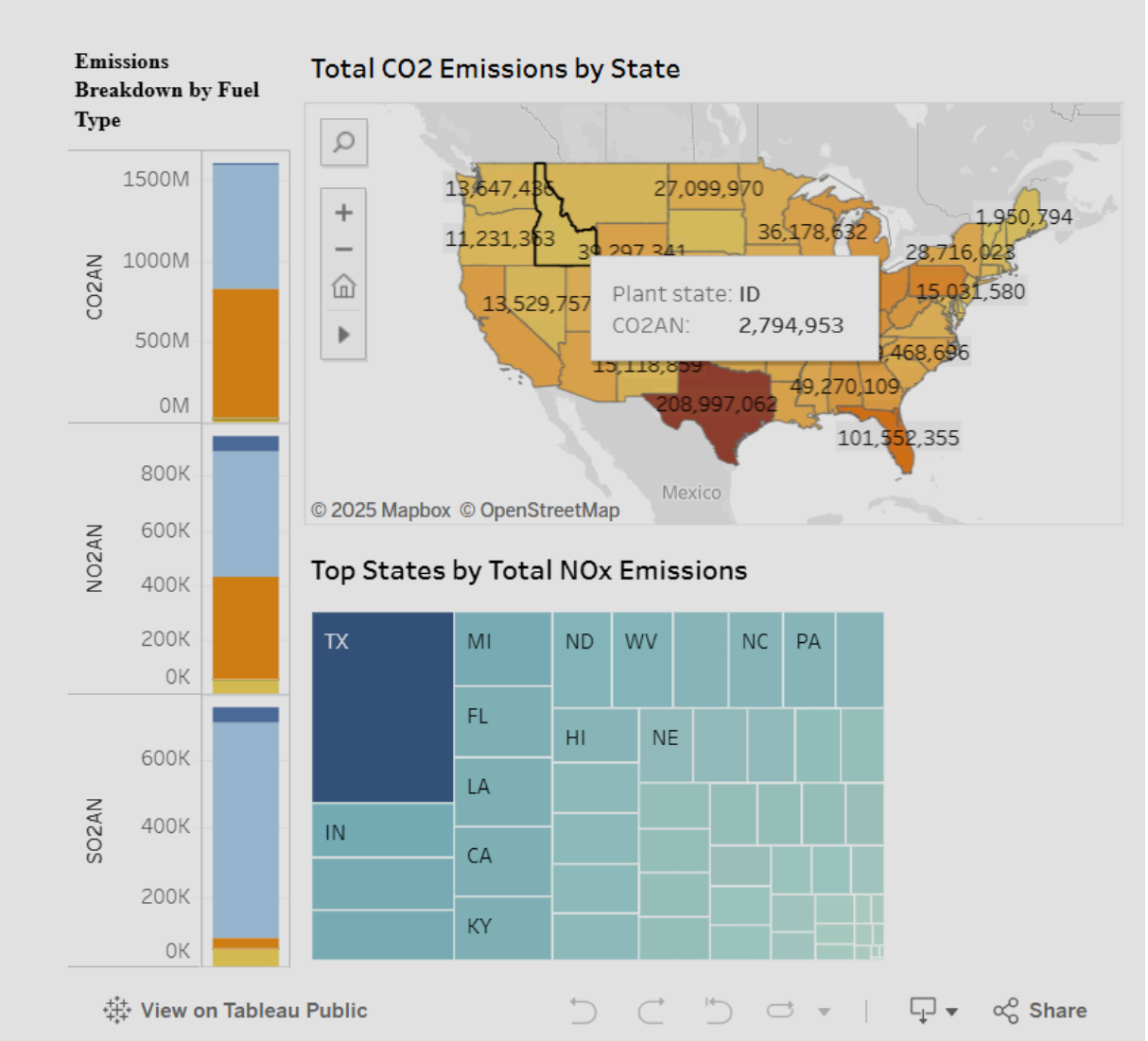


fig. 8

Figure 12: Tableau Dashboard Visualization

This dashboard presents a detailed view of CO_2, NO_2, and SO_2 emissions across different U.S. states, highlighting key contributors to national pollution levels.

Emissions Breakdown by Fuel Type (Left Bar Charts):

The side-by-side bar charts illustrate the total emissions for CO_2, NO_2, and SO_2, grouped by different fuel types. It shows that certain fuel sources contribute disproportionately to specific types of emissions, with CO_2 emissions being particularly high for fossil fuel-based generation.

Total CO_2 Emissions by State (Top Map):

The map visualization shows the distribution of CO_2 emissions across U.S. states. Texas clearly stands out as the highest emitter, contributing 208 million tons of CO_2, significantly more than any other state. Other notable high CO_2 emitting states include Pennsylvania, Florida, and Illinois.

Top States by Total NO_x Emissions (Bottom Tree Map):

The treemap shows the states ranked by their total NO_x emissions. Texas (TX) again leads by a large margin, followed by states like Indiana (IN), Michigan (MI), and Florida (FL).
The size of each block represents the relative contribution of each state to the overall NO_x emissions.

Results and Discussion

Data Modeling

Feature Selection: The top features exhibiting the strongest correlation with the emission levels are:

Key predictors identified based on correlation analysis include

- annual NOx emissions
- ozone season NOx emissions
- annual SO_2 emissions
- annual CO_2 emissions

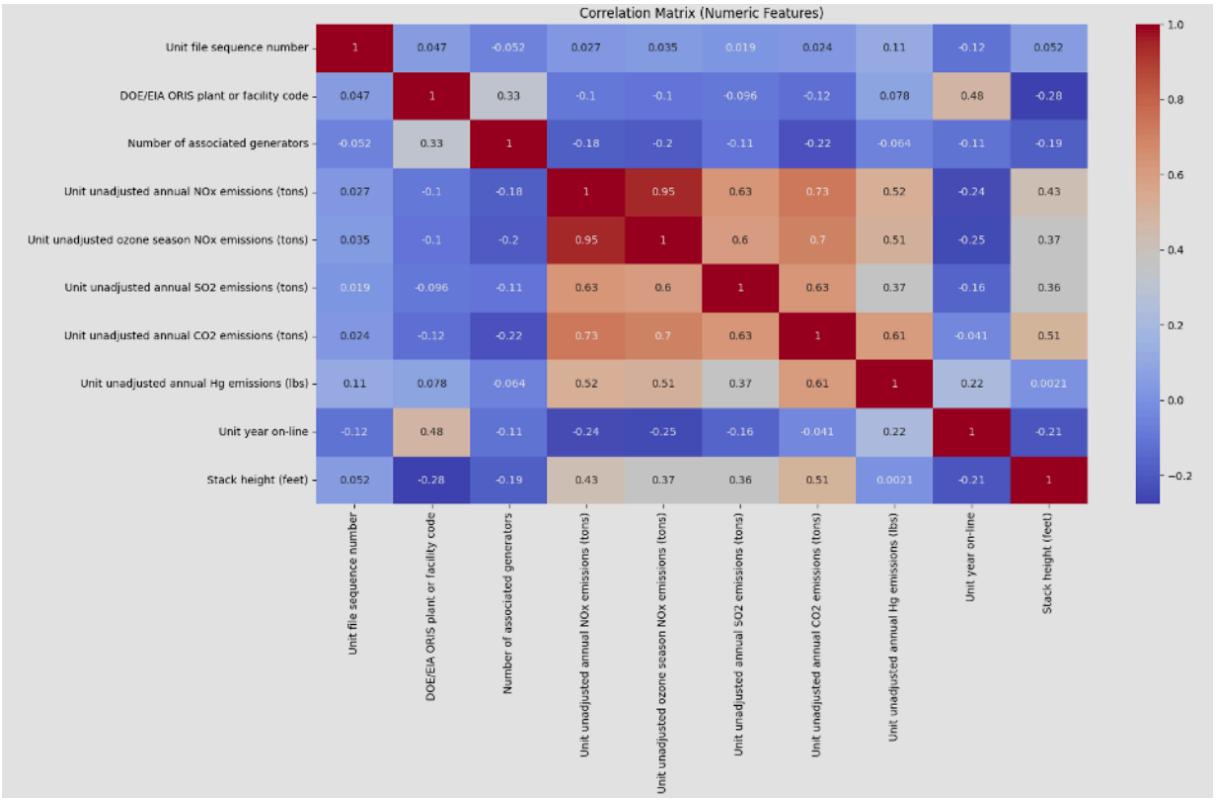


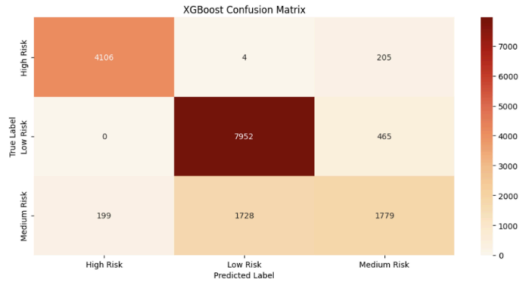
fig. 9

Figure 9: Correlation matrix for different models

Comparison of Different Models for Pollution Risk Classification:

The models we used during classification are:-

1. XGBoost



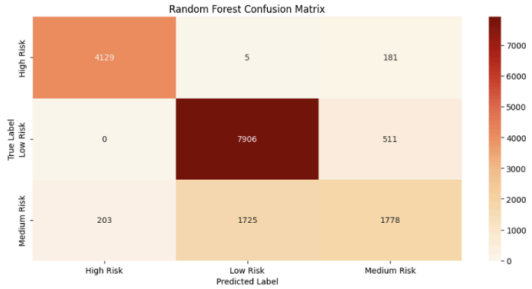

```
--- XGBoost ---
              precision    recall  f1-score   support

   High Risk       0.95      0.95      0.95      4315
    Low Risk       0.82      0.94      0.88      8417
Medium Risk       0.73      0.48      0.58      3706

    accuracy                           0.84     16438
   macro avg       0.83      0.79      0.80     16438
weighted avg       0.83      0.84      0.83     16438

Accuracy: 0.8417690716632193
```

2. Random Forest



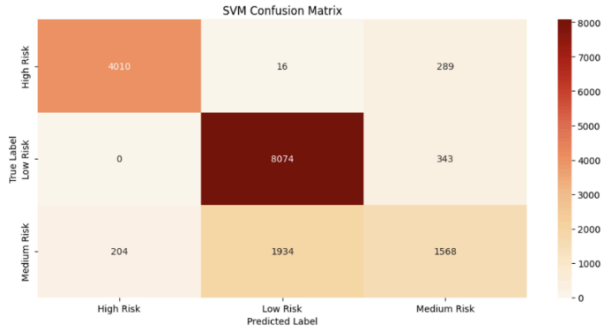

```
--- Random Forest ---
              precision    recall  f1-score   support

   High Risk       0.95      0.96      0.96      4315
    Low Risk       0.82      0.94      0.88      8417
Medium Risk       0.72      0.48      0.58      3706

    accuracy                           0.84     16438
   macro avg       0.83      0.79      0.80     16438
weighted avg       0.83      0.84      0.83     16438

Accuracy: 0.8403090400292006
```

3. SVM



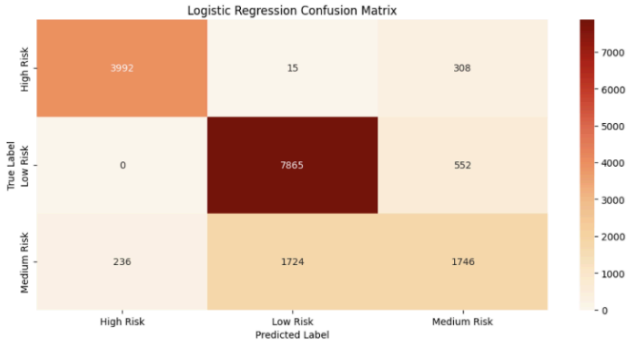

```
--- SVM ---
              precision    recall  f1-score   support

   High Risk       0.95      0.93      0.94      4315
    Low Risk       0.81      0.96      0.88      8417
 Medium Risk       0.71      0.42      0.53      3706

    accuracy                           0.83     16438
   macro avg       0.82      0.77      0.78     16438
weighted avg       0.82      0.83      0.81     16438

Accuracy: 0.8305146611509916
```

4. Logistic Regression



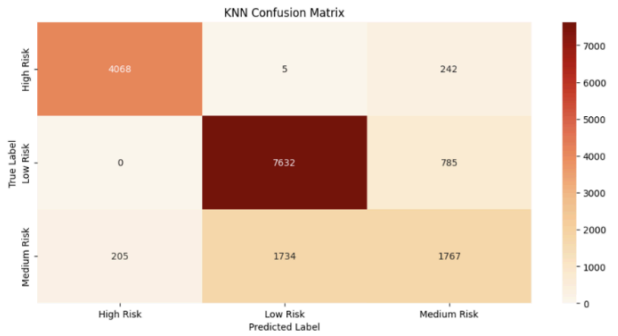

```
--- Logistic Regression ---
              precision    recall  f1-score   support

   High Risk       0.94      0.93      0.93      4315
    Low Risk       0.82      0.93      0.87      8417
 Medium Risk       0.67      0.47      0.55      3706

    accuracy                           0.83     16438
   macro avg       0.81      0.78      0.79     16438
weighted avg       0.82      0.83      0.82     16438

Accuracy: 0.8275337632315367
```

5. KNN



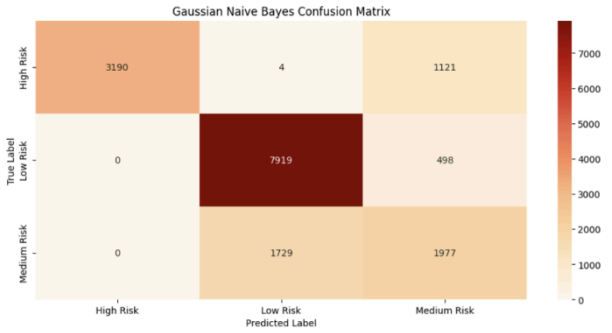

```
--- KNN ---
              precision    recall  f1-score   support

   High Risk       0.95      0.94      0.95      4315
    Low Risk       0.81      0.91      0.86      8417
 Medium Risk       0.63      0.48      0.54      3706

    accuracy                           0.82     16438
   macro avg       0.80      0.78      0.78     16438
weighted avg       0.81      0.82      0.81     16438

Accuracy: 0.8192602506387638
```

6. Gaussian Naive Bayes




```
--- Gaussian Naive Bayes ---
              precision    recall  f1-score   support

   High Risk       1.00      0.74      0.85      4315
    Low Risk       0.82      0.94      0.88      8417
 Medium Risk       0.55      0.53      0.54      3706

    accuracy                           0.80     16438
   macro avg       0.79      0.74      0.76     16438
weighted avg       0.81      0.80      0.79     16438

Accuracy: 0.7960822484487163
```

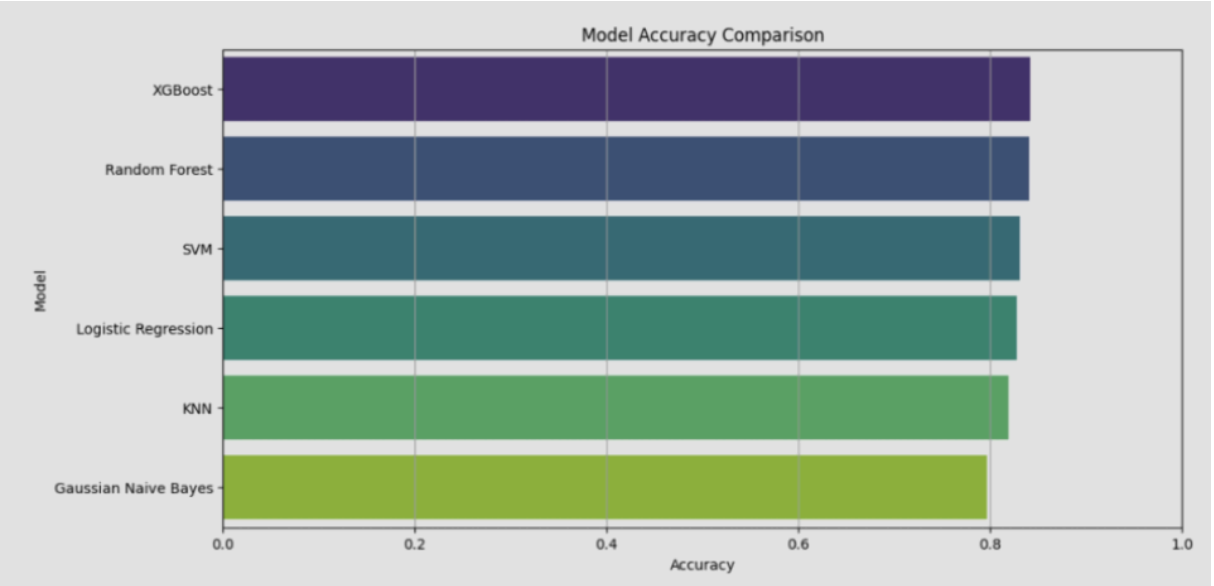


fig. 10

Figure 10: Horizontal Bar graph for different models and their accuracy

XGBoost achieved an accuracy of 84.17%, demonstrating strong capability in distinguishing between high and low pollution risks. With its powerful boosting technique and regularization, it maintained balanced performance across all classes, although moderate struggles were observed in medium risk classification.

Random Forest reached an overall accuracy of 84.03%, offering highly reliable predictions, especially for "High Risk" and "Low Risk" categories.Its ensemble of independent trees made it robust to overfitting and capable of capturing complex patterns in the emission data.

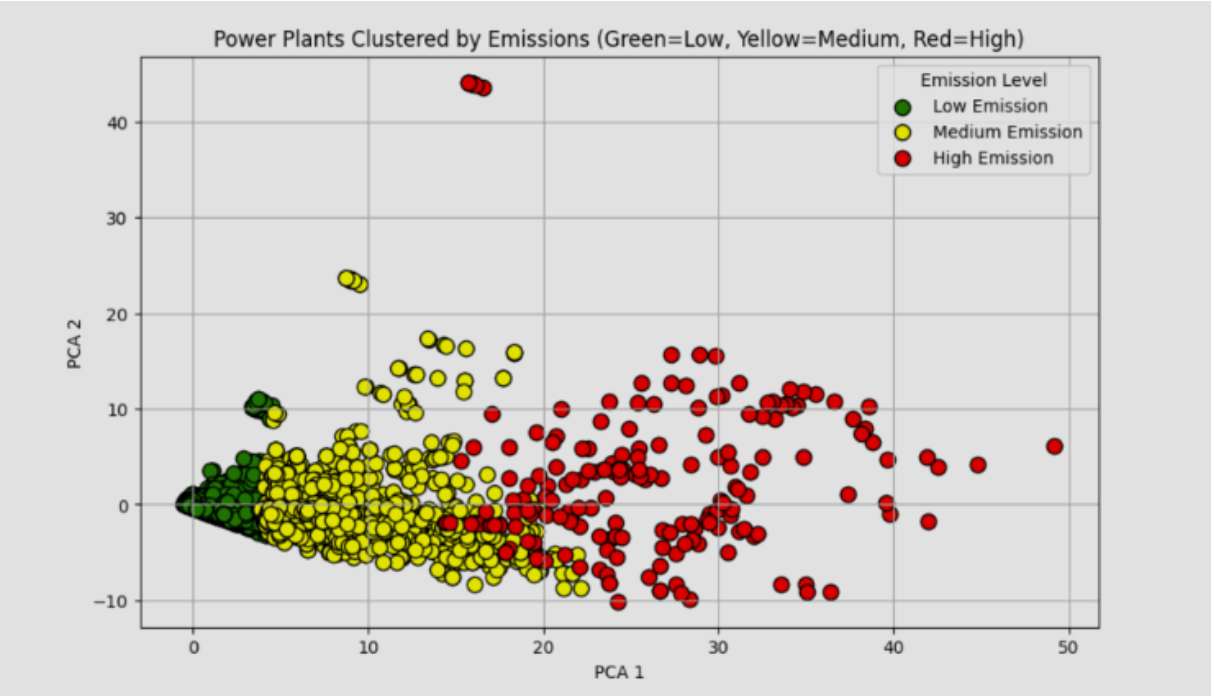


fig. 11

Fig. 11

Using techniques PCA for dimensionality reduction and KMeans clustering, power plants were grouped into three distinct emission levels: Low (Green), Medium (Yellow), and High (Red). The visualization shows a clear trend where plants with higher NOx, SO_2, and CO_2 emissions cluster further apart.

Team Members and Assigned Tasks

1. Dev Shah

- Focused on data preprocessing and spatial analysis using GIS tools
- Designed and built the project website
- Contributed to the PowerPoint presentation, final report, and project video

2. Arbazuddin Mohammad

- Worked on data preprocessing and exploratory data analysis (EDA)
- Handled data visualization and modeling tasks
- Contributed to the website design and report writing

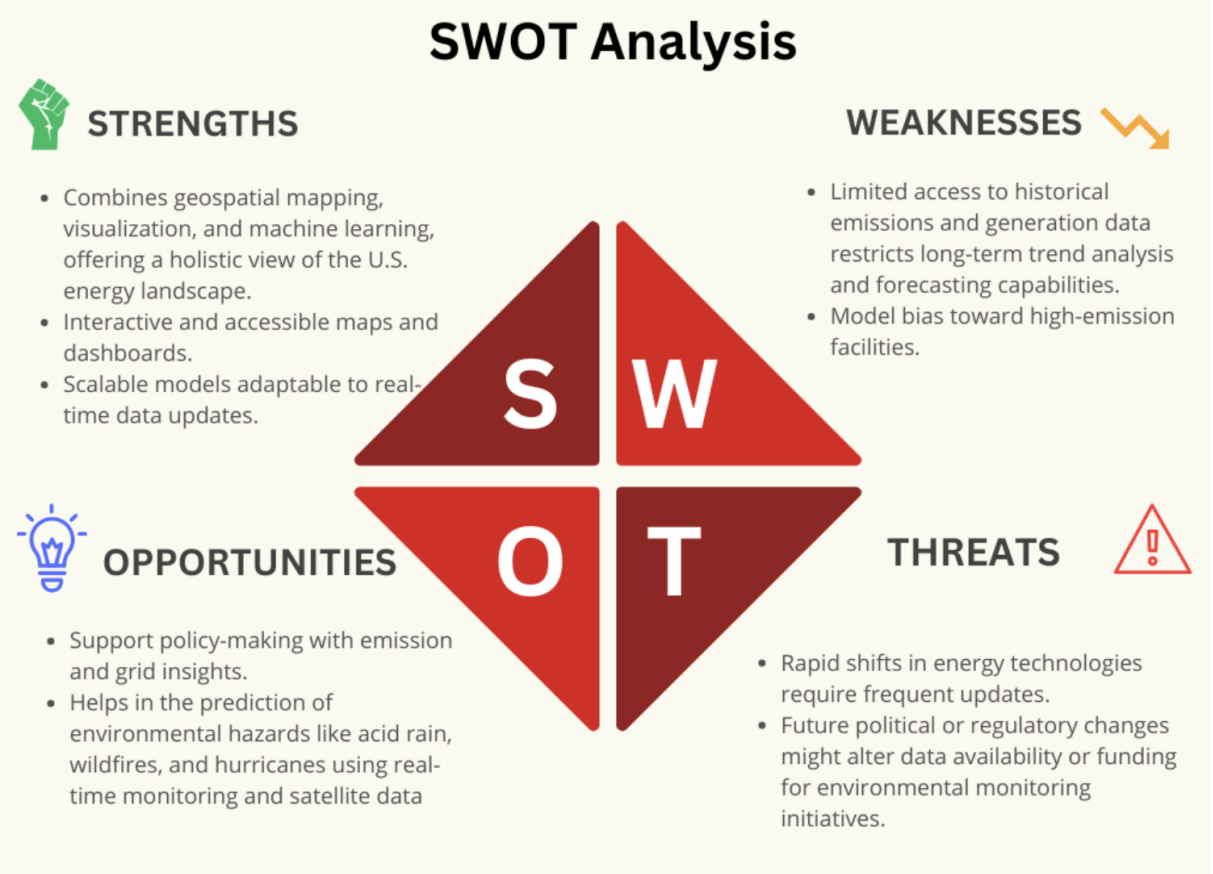
Conclusions

Overall, our project gave us a clearer picture of how infrastructure and disaster risks are connected across different parts of the U.S. By using mapping tools like ArcGIS and building visualizations in Tableau, we were able to spot patterns and highlight areas that are more vulnerable, especially places with aging infrastructure or higher exposure to natural hazards.

One of the key takeaways is that lower-income and rural communities often face more risk but have fewer resources to deal with it. This shows how important it is to use data not just for research, but to help support better planning and policies that make infrastructure safer and more fair for everyone.

There's still a lot of room to expand this work, like bringing in more real-time data, working with local agencies, or focusing on specific types of infrastructure in future projects. But even with what we've done, we hope our findings can make a small impact and help guide smarter, data-informed decisions moving forward.

SWOT Analysis



Strengths

One big strength of our project is that it combines geospatial tools like ArcGIS with interactive dashboards, which makes the data easy to explore and understand. The system is also pretty flexible, it can adapt to real-time data updates and scale up as needed. This gives a complete view of the infrastructure and hazard situation across the country.

Weaknesses

A main issue we ran into was that older historical data on emissions and energy generation isn't always available or easy to access. That makes it harder to do long-term trend analysis. Also, some models might have a slight bias, especially toward high-emission sites, depending on the input data.

Opportunities

There’s a lot of potential for this kind of work to support public policies, like helping with emission planning or improving the grid. The same tools can also help predict environmental problems like wildfires or hurricanes by using real-time data and satellite feeds.

Threats

Technology in energy and mapping is changing fast, so tools and models need regular updates. Plus, changes in political leadership or regulations could reduce funding or limit access to key data sources, which could slow down future research or development.

Citation/Sources

[1] Federal Emergency Management Agency. National Risk Index. FEMA, 2021. https://www.fema.gov/flood-maps/national-risk-index

[2] U.S. Department of Transportation. National Bridge Inventory Data. FHWA, 2022. https://www.fhwa.dot.gov/bridge/nbi/ascii.cfm

[3] U.S. Census Bureau. American Community Survey and Population Estimates. U.S. Department of Commerce, 2020. https://www.census.gov/data.html

[4] Environmental Protection Agency. State and Local Energy Emissions Inventory Tool. EPA, 2021. https://www.epa.gov/statelocalenergy

[5] Tableau Software. Tableau Public – Free Data Visualization Tool. 2024. https://public.tableau.com

[6] Esri. ArcGIS Online – Mapping and Spatial Analytics Platform. Esri, 2024. https://www.arcgis.com