PROJECT REPORT

ANISH DIXIT

INTRODUCTION

This course project involves analyzing wildfires in the US and trying to estimate the impact these wildfires and the smoke produced have on the environment, society and lives of residents around. More and more frequently, summers in the western US have been characterized by wildfires with smoke billowing across multiple western states. There are many proposed causes for this: climate change, US Forestry policy, growing awareness, just to name a few. Regardless of the cause, the impact of wildland fires is widespread. There is a growing body of work pointing to the negative impacts of smoke on health, tourism, property, and other aspects of society.

Investigating the spread and occurrence of wildfire smoke is essential for comprehending the dynamics of these natural disasters. Wildfires are often exacerbated by climate change, and their smoke contributes to air pollution. Analyzing the socio-economic impact provides insights into the broader implications of climate change on local environments and communities. Studying the spread of wildfire smoke allows for a better understanding of its health implications. This includes respiratory problems, cardiovascular issues, and the potential long-term health effects on the affected population. By identifying the socio-economic impact of wildfire smoke, the analysis can contribute to more effective risk mitigation strategies. This could involve improved emergency response planning, public health interventions, and land management practices to reduce the frequency and severity of wildfires. Understanding the economic consequences on nearby cities helps in allocating resources efficiently. This information is vital for local governments, emergency services, and healthcare providers to allocate funds and personnel appropriately.

The analysis can contribute to building resilient communities by providing data-driven insights into the socio-economic vulnerabilities that arise from exposure to wildfire smoke. It can serve as a tool for raising public awareness about the potential socio-economic impact of wildfires. This information is crucial for fostering a sense of responsibility among residents and encouraging proactive measures to reduce the risk and impact of wildfires. The findings can inform the development of policies and regulations aimed at mitigating the socio-economic impact of wildfire smoke. This includes zoning regulations, building codes, and land-use planning to minimize the vulnerability of communities to wildfires.

BACKGROUND/RELATED WORK

The US Environmental Protection Agency has a dedicated Air Quality System. AQS contains ambient air sample data collected by state, local, tribal, and federal air pollution control agencies from thousands of monitors around the nation. It also contains meteorological data, descriptive information about each monitoring station (including its geographic location and its operator), and information about the quality of the samples. Their API is the primary place to obtain row-level data from the EPA's Air Quality System (AQS) database.

RESEARCH QUESTION: How good is my smoke estimate with actual air quality index and is there a decent correlation between the two?

The Nebraska Department of Revenue provides population census data that I have used in my research. In addition to this, the main sources of obtaining health data were the Public Health Inventory Reports generated by the Nebraska Department of Health and Human Services which provide detailed insights, tables and graphs about yearly cancer cases in Nebraska. They also drill it down by gender, region, type of cancer and many other parameters which give users a great overview. This was also my primary dataset for cancer cases occurrences.

The impact of wildfires on the environment is a hotly debated topic because of the potential implications of this on the overall climate. How wildfire smoke affects the health of people around is one of the important factors and touched upon in quite a few research works. The US Environmental Protection Agency has authored an article titled 'Wildland Fire Research: Health Effects Research' which talks about the effects of smoke from wildfires, which can range from eye and respiratory tract irritation to more serious disorders, including reduced lung function, bronchitis, exacerbation of asthma and heart failure, and premature death. Children, pregnant women, and the elderly are especially vulnerable to smoke exposure. Emissions from wildfires are known to cause increased visits to hospitals and clinics by those exposed to smoke.

RESEARCH QUESTION: The above research clearly shows that wildfire smoke can have very detrimental health effects, and that is why I chose to model smoke estimates with lung cancer cases in my work and find correlations/causations between the same. I also try to use smoke estimates to predict future trends in lung cancer cases in the region.

METHODOLOGY

I followed a step by step approach across multiple parts in order to execute the ideas I had in mind for this project. This compartmentalization of tasks helped in creating a clear logical flow which from a human centered perspective, even a layman can understand without too much difficulty.

1. Wildfire Data Acquisition

- In the Python Notebook Wildfire_Data_Acquistion.ipynb, we first read in wildfire data from the JSON file of Combined wildfires, **USGS_Wildland_Fire_Combined_Dataset.json**, which is obtained from the US Geological Survey. This is a massive data file of 2.8GB.
- Professor McDonald's GeoJSON reader library and helpful code, which we are licensed to use, helped me undertake these processing steps.
- I then filter the data to only include those fires which are after 1963 and within 1250 miles of North Platte, Nebraska.
- I calculate a smoke estimate based on distance, area of fire and fire intensity as well.

$$SmokeEstimate = \frac{(AreaBurnt * FireIntensity)}{DistanceFromCity}$$

Smoke should be inversely proportional to distance (more the distance, lesser the smoke spread), directly to area burnt and also to fire intensity (wildfires should be weighted more than prescribed fires, I have chosen 2x). The intensity parameter is weighted as follows:

```
fire_intensity['Wildfire'] = 2
fire_intensity['Likely Wildfire'] = 1.75
fire_intensity['Unknown - Likely Wildfire'] = 1.5
fire_intensity['Prescribed Fire'] = 1.25
fire_intensity['Unknown - Likely Prescribed Fire'] = 1
```

- A massive intermediate JSON is generated named **smoke_estimate_NP.json**, which has wildfire features, distance from city and smoke estimate value.
- The JSON data has data from 1963-2020.

2. AQI Data Acquisition

- In the Ipython Notebook ,**AQI_Data_Acquisition.ipynb**, I get Air Quality index data from the US AQS (Air Quality System) API implemented by the EPA. Professor McDonald's helper notebook, licensed freely to use, helps in this process.
- I set up an account with EPA, created an API key and then repetitively called the API to get air quality indices around North Platte.
- Now, this city does not have any weather stations close by, so we create a bounding box for 50,100,150,200,250 miles around the city and get data from weather stations around there. Yearly API calls are made and AQI values are averaged out.
- Finally **aqi.csv** is generated with yearly data from (1985-2022) and AQI values for each year.

3. Wildfire Data Analysis

- In the Python Notebook **Wildlife Data Analysis.ipynb**, we read in **smoke_estimate_NP.json** and create visualizations, corresponding to different kinds of impact of the fires.
- The first graph is a histogram of the number of fires occurring every 50 miles from North Platte.
- The second graph charts out the area of land burnt by wildfires per year.
- The third chart tries to correlate air quality index with calculated smoke estimate and finds similarities in the two.
- We also create predictions for smoke estimates 25 years into the future, using the Prophet library by Facebook for Time Series Forecasting.
- The predicted smoke estimate values and corresponding years are saved in **future_predicted_smoke_estimate.csv**

4. Extension Plan

I planned to extend this work by trying to apply it to certain socio-economic factors, giving it a human centered angle, which can actually impact the community of North Platte as well. The focus area I am choosing to base my analysis on is ***Healthcare.*** I chose this domain because I believe the smoke estimate calculated in the common analysis would have the maximum impact and correlation with the health of the residents of North Platte. As mentioned in the motivation, standalone wildfires can affect the health and well being of the individuals residing in the city, and can have detrimental effects on the long term fitness as well. Of course, wildfires are not the only contributors to this wane of health, it is a much larger systemic problem, but it can still serve as one indicator. *I intend to study occurrences of bronchial cancer and tuberculosis in the North Platte area across years, try to map and correlate it to the yearly smoke estimate in part 1, and predict how feature frequency of cancer cases/mortality rates would look.*

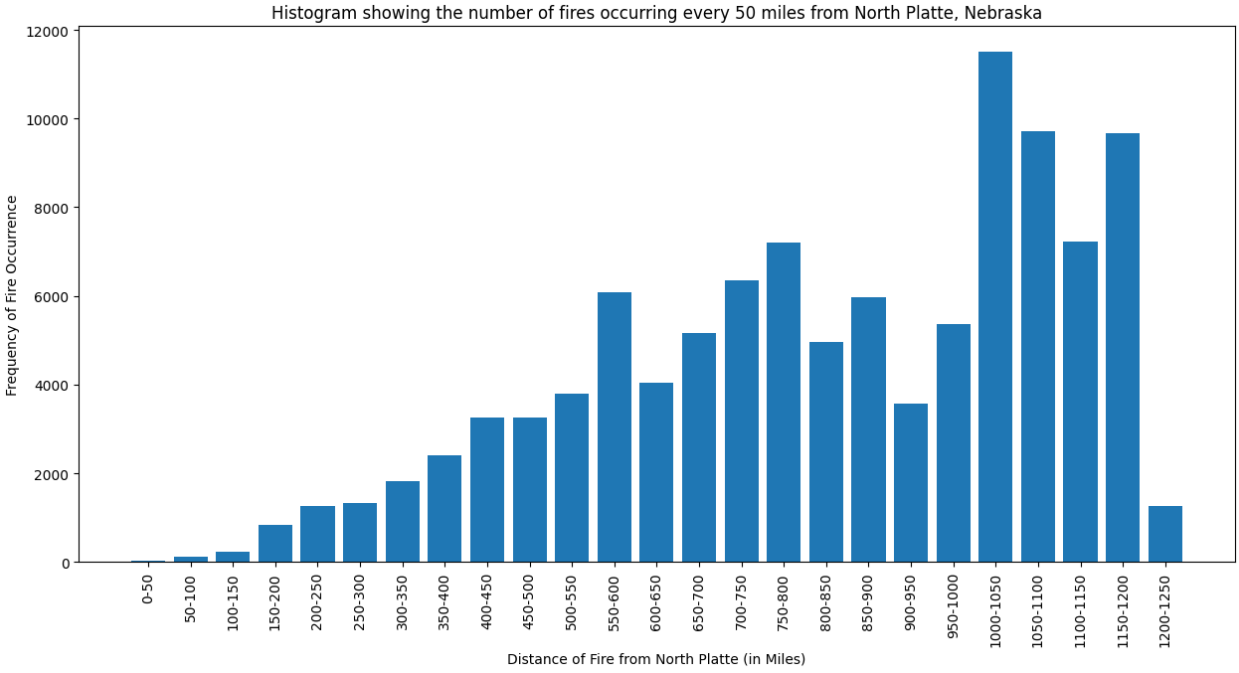
I decided to expand upon these diseases specifically as these are ***respiratory illnesses***. They can be directly brought about, cause or if inherently present, intensified by excessive smoke in the region, which my previous work can help with. A future prediction can help medical services be prepared for sudden spikes in cases in a cyclic manner if any seasonality exists, or if there is simply a generic upward trend in the number of cases found. This work could help the health services in the North Platte city and Lincoln county better understand how wildfire occurrences could be a potential harbinger of respiratory diseases and how the population is being affected by the same. Efforts can be made to better handle these humanitarian problems.

5. Modeling Smoke Impact on Respiratory Illnesses

- In the Python Notebook **Smoke Impact + Respiratory Illness Modeling.ipynb**, we read in population data from the MUN_1_Cert_Pop_Summary.xlsx and population_1990s.csv and merge it. We also read in cancer cases data from cancer_data.csv.
- I then have a look at a comparison of total detected lung cancer cases and fatalities by plotting a bar chart.
- I also read the smoke estimates calculated in **smoke_estimates_np.json**.
- I plot Correlations between cancer occurrences and smoke estimates and find good positive correlations.
- Finally, I try to predict future trends in lung cancer cases by using a Linear Regression model with smoke estimate as a predictor and obtain interesting trends.
- These values are stored into **lung_cancer_predictions.csv** with yearly values.

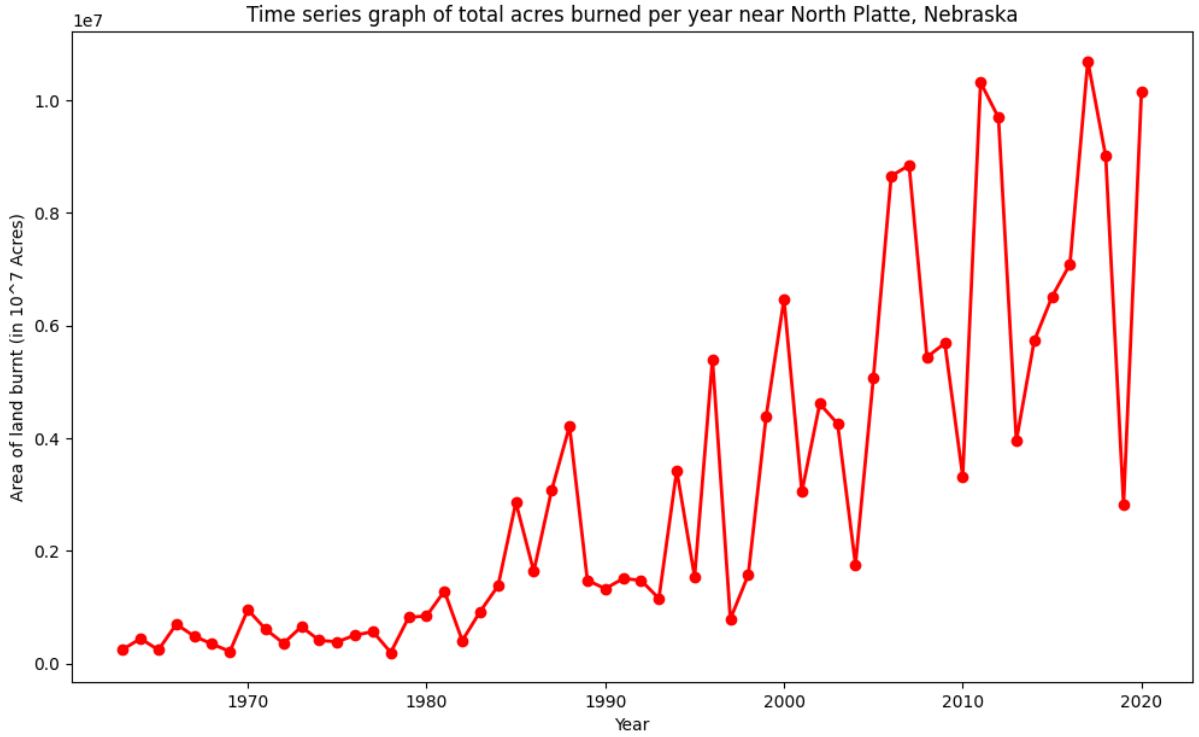
FINDINGS

In the first step of acquiring Wildfire data, I found around 100000 fire occurrences within 1250 miles of North Platte, across 50 years from 1963 to 2023. First off, I wanted to have a look at how the fires vary by distance from the city. This was important because if there are a lot of fires very close to the city, it is an alarming fact and the administration needs to account accordingly. The results can be seen in this figure:



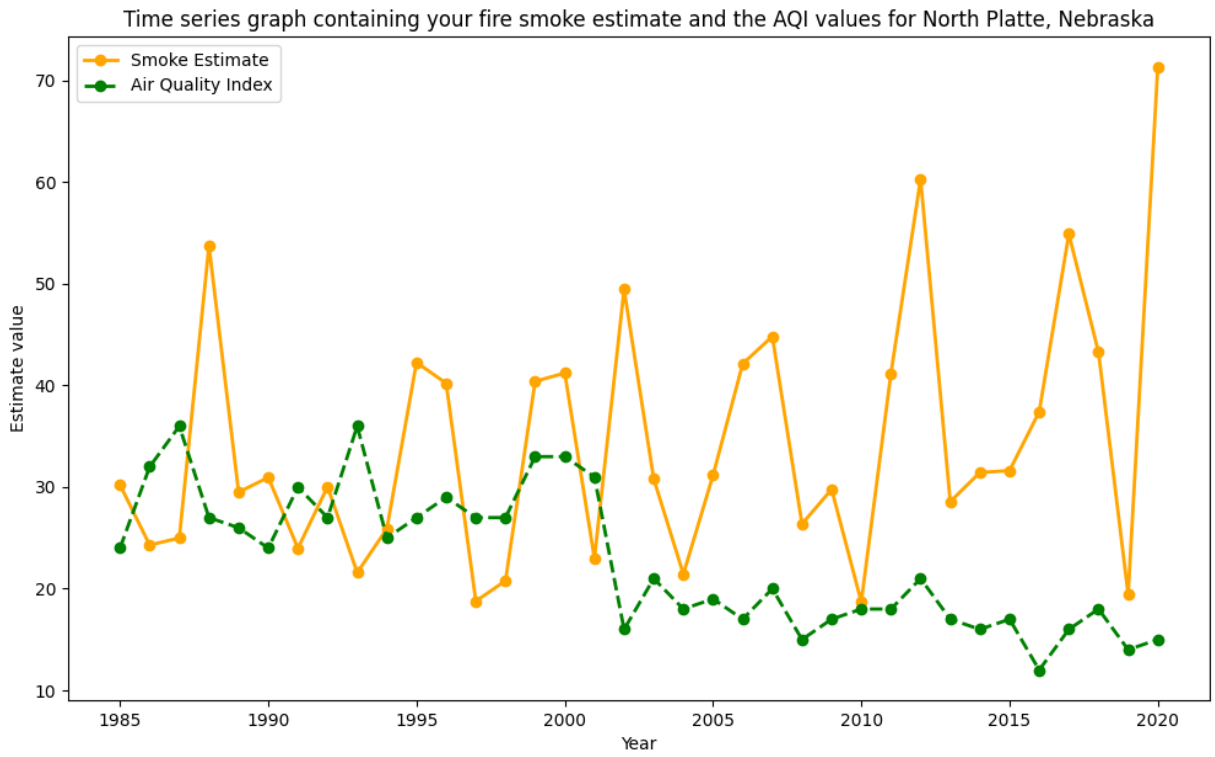
This histogram represents the frequency of occurrence of fires, arranged distance wise in bins of 50 miles, from North Platte, Nebraska. The data came from the filtered GeoJSON from USGS which includes the fire date, distance from city, smoke estimate, area burnt, fire type and other such parameters. The features used for this visualization is a count of the number of fire occurrences within each 50 mile distance bucket. The X axis represents the distance from North Platte in miles. The y axis represents the count of the number of wildfires. We can see that it is fairly well distributed, with more fires being present farther away from the city. The maximum number of fires are at a distance of 1000-1500 miles from the city. I can infer that North Platte, Nebraska is not a big wildfire center because there are not a lot of fires closer to the city.

The second analysis tries to model the geological impact of the wildfires. I look at the number of acres of land burnt historically. This would give me an idea of how recent trends in fire occurrences look like. If recent trends are pointing to more areas burnt by fires, it is problematic for agriculture, livestock rearing, human settlements etc.



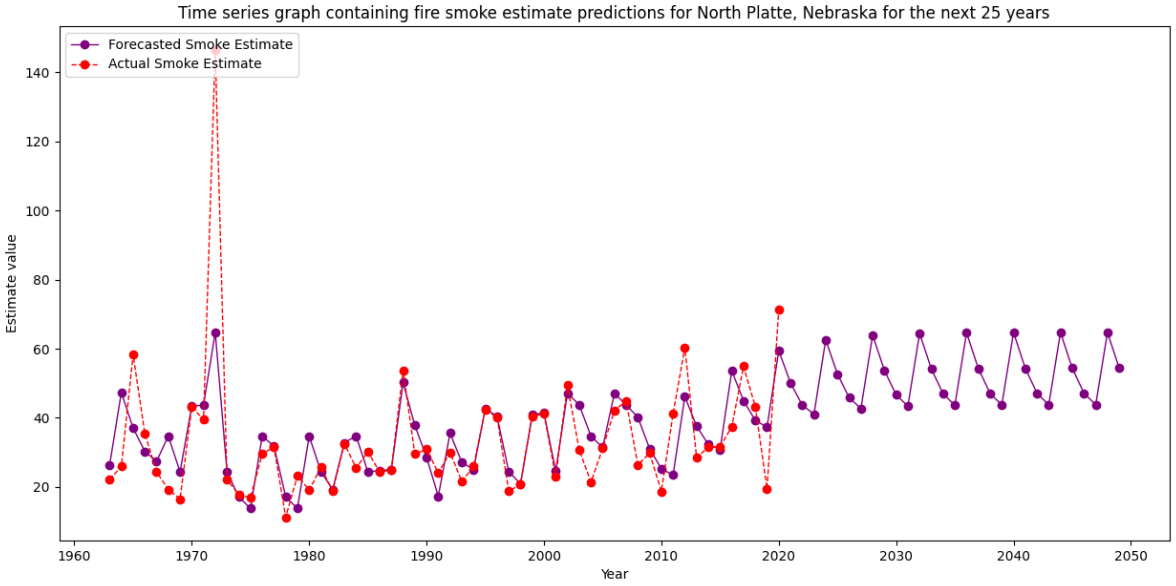
This time series graph represents the area burnt by fires close to North Platte. The data came from the filtered GeoJSON from USGS which includes the fire date, distance from city, smoke estimate, area burnt, fire type and other such parameters. The features used for this visualization is a summation of the total area burnt by the fire for each year. The X axis represents the years of the fires (1963-2020). The y axis represents the total area of land burnt in 10^7 acres for better visualization. We can clearly see that the amount of land area being burnt is increasing with time. The 2000s had a way higher burn impact than the 1900s. I can infer that the impact of fires around North Platte, Nebraska has grown significantly. This kind of correlates with the increased number of fires that we saw in the first graph, and probably fire intensity has also increased substantially, which does not bode well for North Platte city in general.

The final exploratory analysis tries to establish correlations between actual Air Quality Index data and my calculated smoke estimates. If there is a good similarity, it means that my smoke estimate is a good indicator of the real world impact wildfire smoke would have on the city of North Platte.



This time series graph represents the values of smoke estimates generated by me and the air quality indices, across years. The data came from the filtered GeoJSON from USGS which includes the fire date, distance from city, smoke estimate, area burnt, fire type and other such parameters. It also included the AQI data from the API, which has been aggregated by year to give one single value. The features used for this visualization are smoke estimate, aqi, fire year. The X axis represents the years of the fires (1985-2020). The y axis represents the estimated values. The orange solid line shows my smoke estimate and the green dotted line shows the actual AQI values. We can see some sort of correlation between the two lines charted. In the initial years, the smoke estimate and aqi overlap significantly. Later on the AQI drops a bit (which is great! While the smoke estimate actually goes up. The smoke estimate going up can be attributed to the higher number of fires and higher area burnt, as we say in the first two graphs. The discrepancy might arise from the fact that we are considering fires from up to 1250 miles away, but we are considering air quality only up to 200 miles away. Maybe the number of fires did increase, but the city still succeeded in maintaining a good air quality!

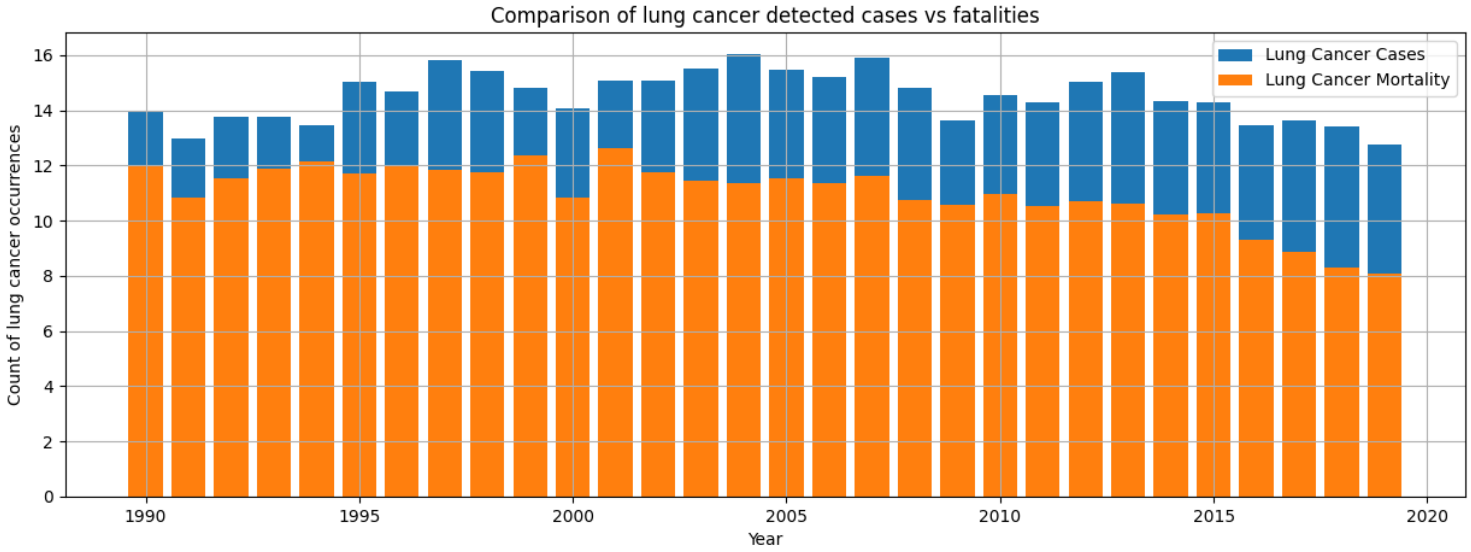
My next endeavor was to forecast future trends in smoke estimates. For this, I used Facebook's Prophet library for Time Series forecasting. I computed average yearly smoke estimate values and used the Prophet model to calculate estimates for the next 25 years, using Exponential Smoothing and ensuring Trend and Seasonality. This graph shows what it looks like going forward:



This time series graph represents the values of smoke estimates generated by me and the predicted values, across years. The data came from the filtered GeoJSON from USGS which includes the fire date, distance from city, smoke estimate, area burnt, fire type and other such parameters. The features used for this visualization are smoke estimate, fire year. The X axis represents the years of the fires (1963-2050). The y axis represents the estimated values of smoke impact. The pink dotted line shows my calculated smoke estimate and the purple solid line shows the predicted smoke estimate values. We can see some sort of correlation between the two lines charted. In particular, for the available data up till 2020, the purple line beautifully models the pink line, mirroring the trend almost perfectly. This is indicative of a very good model fit. In the future, the smoke is estimated to gradually increase, with seasonal cycles. I think the forecasted values are too regular and of course in the real world it cannot be as perfect as this and will have irregularities. However, it shows that smoke impact will worsen soon, and the administration of North Platte needs to take measures to ensure the people are not adversely affected.

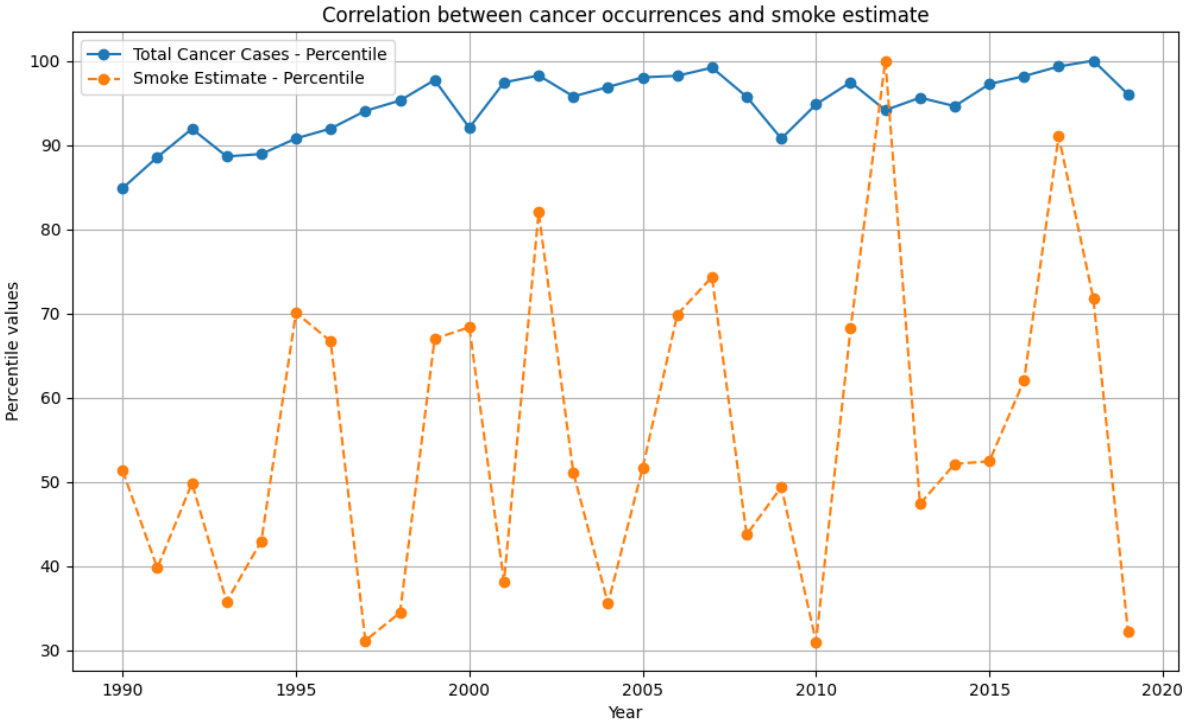
After planning out an extension, I set about collecting my dataset. My population data was in a format where I had an excel spreadsheet with different sheets corresponding to population data for different cities in Nebraska for different years. I had to pull together data across these sheets and filter to only keep North Platte census data for population counts. The cancer data I planned to use was also unstructured and had to be captured from tables within PDF reports. I manually created CSV files from these PDF reports with the values I needed.

Now, I first wanted to have a look at existing trends relating to lung cancer cases in Nebraska and North Platte. Taking the city level total lung cancer cases and fatalities, I plotted out a stacked bar chart showing total counts.



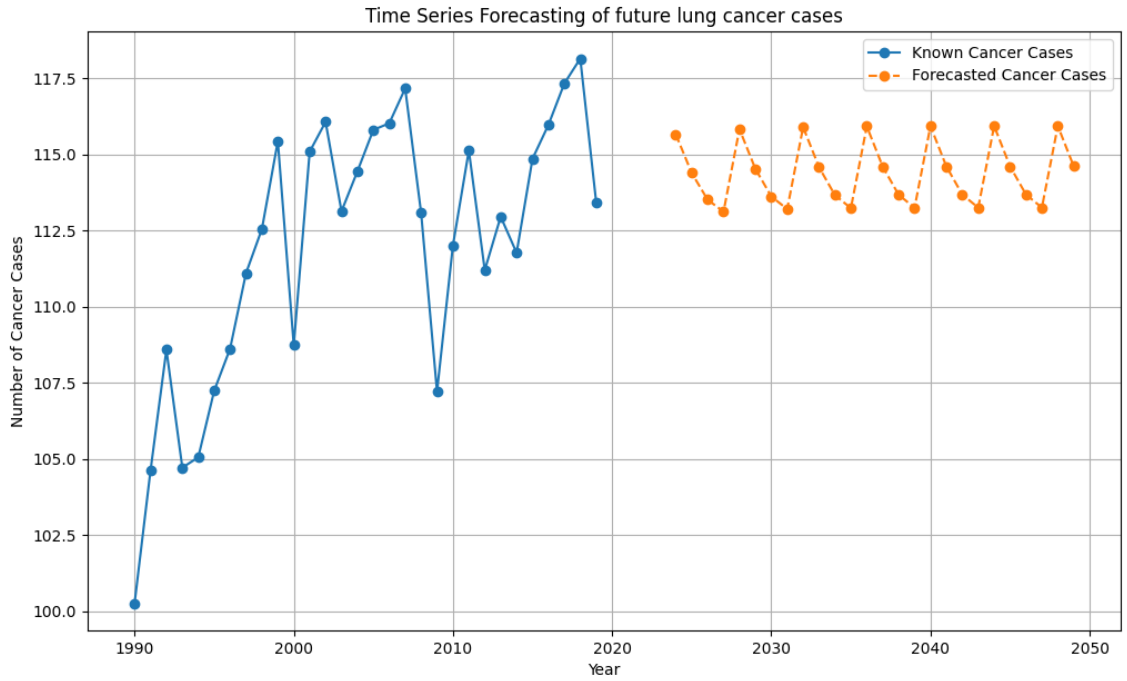
This time series graph represents the number of lung cancer detected cases and fatalities in North Platte. The data came from the cancer cases reports from the Nebraska Health Dept. The features used for this visualization are lung cancer cases, lung cancer mortalities and year. The X axis represents the years (1990-2020). The y axis represents the number of cases. The blue bars represent the total number of bronchial cancer cases and the orange charts show the number of fatalities. It is a good thing to visually observe that both these numbers are reducing in recent years, meaning that total cancer occurrences around the area are reducing. However, one alarming fact is that the mortality rate is really high. A large population of the people who get affected by lung cancer are dying, this needs to be fixed by the local healthcare administration using medical measures as are necessary.

The final exploratory analysis tries to establish correlations between actual lung cancer cases data and my calculated smoke estimates. If there is a good similarity, it means that my smoke estimate is a good cause of lung cancer happening to the residents of North Platte.



This time series graph represents the number of lung cancer cases and smoke estimates in North Platte. The data came from the cancer cases reports from the Nebraska Health Dept and previously calculated smoke estimates from USGS wildfires data. The features used for this visualization are lung cancer cases, smoke estimate and year. The X axis represents the years (1990-2020). The y axis represents the percentile values. The orange dotted line shows my smoke estimate and the blue solid line shows the actual AQI values. We can see some sort of correlation between the two lines charted. To quantitatively confirm this, I calculated the Pearson correlation coefficient between the 2 estimates, and I obtained a value of 0.32, which confirmed positive correlation. This was great because it meant wildfire smoke is in some way contributing to lung cancer in the area and the authorities can act accordingly.

My final endeavor was to forecast future trends in lung cancer occurrences. For this, I used a Linear Regression model with smoke estimate as a predictor. I used Scikit Learn Linear Regression models for this and fitted it to my data. It is a decent fit and this graph shows what the trends for the next few years look like:



This time series graph represents the count of lung cancer cases from the Nebraska Public Health Reports and the predicted future values, across years. The data came from the Nebraska Public Health Reports and smoke estimates from the USGS datasets. The X axis represents the years of the fires (1990-2050). The y axis represents the count of lung cancer cases. The blue solid line shows actual cancer cases estimate and the orange dotted line shows the predicted smoke estimate values. We can see some sort of correlation between the two lines charted. In the future, the lung cancer is estimated to gradually increase, with seasonal cycles. I think the forecasted values are too regular and of course in the real world it cannot be as perfect as this and will have irregularities. However, it shows that lung cancer will be more prevalent, and the administration of North Platte needs to take measures to ensure the people are not adversely affected.

DISCUSSION/IMPLICATIONS

Over the course of implementing the various parts of this project, I have uncovered some very interesting findings, trends and future predictions, relating to wildfires and the impact of their smoke, specifically around North Platte, Nebraska. Some of these clearly warrant steps from the local administration to curb the adverse effects in the future. Some results are also promising, probably stemming from previous actions undertaken by the authorities. One of these is the fact that we saw that there are lesser fires closer to the city of North Platte. We also have a disturbing trend of very high land area burnt by the wildfires in recent years. In order to reduce this, some steps include clearing vegetation, creating defensible zones, maintaining lifebreaks and safe landscaping. Other steps include using satellites, drones and other surveillance systems to monitor and detect fires in their early stages, encouraging communities to report signs of potential fires promptly, and mobilizing firefighting teams quickly to contain small fires before they escalate. Employing aircraft, such as water bombers or helicopters, to drop water or fire retardants on the fire to slow its progression and conducting controlled or prescribed burns in strategic areas during periods of low fire risk to reduce fuel loads and minimize the risk of larger, uncontrolled fires would also help. We also saw that the Air Quality index is reducing, and this is a very positive trend, which can be continued to follow air quality standards and regulations.

We also looked at trends in lung cancer occurrences. The positive fact is that the total number of lung cancer cases are reducing, and following air quality regulations, discouraging smoking or vaping can help in this. However, we also saw that the fatality rate is very high. This can be attenuated by Raising public awareness about the dangers of smoking, Promoting lung cancer screening programs and increasing awareness about the symptoms of lung cancer and encouraging individuals to seek medical attention if they experience persistent symptoms, such as coughing, chest pain, or shortness of breath. Improving access to healthcare services, especially for underserved populations and ensuring that individuals have access to affordable and timely medical care, including cancer screenings and treatment, would also help in this cause.

All of these problems have been analyzed from a human centered perspective. I took special care to ensure that going beyond the geological or environmental concerns of wildfire, I looked at how it would impact the residents of the area. This is precisely why I chose to look at respiratory illnesses. I have tried prioritizing the well-being and needs of the affected communities. This approach recognizes that the ultimate goal of the analysis is to enhance the quality of life for individuals and communities facing the challenges posed by wildfire smoke.

LIMITATIONS

Despite all the methods implemented in this study and my utmost efforts to perfect them, there are undoubtedly some limitations and shortcomings in this study that would hamper reproducibility and reduce the applicability of this work in the real world. I have tried to list out all of these limitations as honestly as I can.

1. While calculating the smoke impact around North Platte, we have considered wildfires from up to 1250 miles away. This is a HUGE radius and we also saw that most of our wildfires are far away from the city. Realistically, it is very unlikely that smoke from these wildfires would have any impact on the city in itself, hence our smoke estimate may not be very relevant.
2. The smoke estimate calculation formula is arbitrary and is based only on a few parameters: distance, area and fire intensity. In reality, these cannot explain the entire smoke effect. Many other parameters such as wind speed, wind direction, land topography etc. would also affect how smoke would travel.
3. The AQI data I got only had particulate pollutants and no gaseous data, hence it cannot accurately represent real values. Another factor is that I did not find any weather stations close to North Platte, hence I took stations from up to 200 miles away, which cannot really extrapolate well.
4. We also have not considered fire season in either our wildfire or AQI data. The USA fire season runs from May to October, but because we do not have good month level data, we do not consider this, which is an oversight.
5. The GeoJSON dataset is massive and the data acquisition process needs good error handling. Some fires do not have the 'ring' parameter and instead have a 'curved ring' which needs to be handled and dealt with separately.
6. The Time Series forecasting using Facebook's Prophet library introduces trend and seasonality into future predictions without being sure whether the historical data actually have any trend. Considering the fact that the smoke estimate was built off a formula I created, I do not think it would have well defined trends.
7. Lung cancer occurrences cannot be attributed only to smoke, hence using a regression model with only smoke as a predictor is a gross oversimplification. A plethora of factors such as the individual's immunity, smoking practices, etc., working conditions play a part in determining how they contract lung cancer.
8. Linear regression has normality and homoscedasticity assumptions, and we are not checking these before applying it to our data.

However, I believe that these limitations do not discredit the work as a whole. The statistical analysis conducted is sound and the human centered design of this analysis ensures that it does end up benefiting the people of the affected region ultimately.

CONCLUSION

To summarize the project, this undertaking involves analyzing wildfires in the US, specifically around North Platte, Nebraska, and trying to estimate the impact these wildfires and the smoke produced have on the environment, society and lives of residents around. I calculated a smoke impact estimate to quantify the wildfire smoke's effect on the city. One of my research questions involves trying to correlate my smoke estimates with the actual Air Quality Index in the region. I also found that there are lesser fires close to the city, but also that wildfires have burnt huge amounts of land around the city in recent years. I tried to forecast future trends in smoke and found that smoke in the city will increase in the coming years and the local authorities need to administer measures to ensure the residents are not affected.

I also tried to give a human centered angle to my analysis by trying to gauge if respiratory diseases such as lung cancer in the area were a by-product of the smoke in the air. I found that though the overall lung cancer cases in the area were going down of late, the fatality rates were very high and this needed fixing. Finally, I tried to forecast future lung cancer occurrences using smoke estimates as a predictor and concluded that cases could well increase in the future.

Throughout this project, I have tried to follow the best human centered data science practices. Human-centered data science involves integrating principles of human-centric design, ethics, and user engagement into the entire data science process. Some important steps I followed were prioritizing ethical considerations throughout the data science lifecycle, understanding the real-world problems and challenges faced by end-users, formulating data science problems based on the actual needs and experiences of the people who will be impacted by the solutions, regularly evaluate the impact of data science solutions on individuals and communities and assessing how well the solutions meet the intended goals and whether any unintended consequences have emerged.

REFERENCES

- Creative Commons License: https://creativecommons.org/licenses/by/4.0/
- Wildfire Polygons Metadata Explanation: https://www.sciencebase.gov/catalog/file/get/61aa537dd34eb622f699df81?f=__disk__d0%2F63%2F53%2Fd063532049be8e1bc83d1d3047b4df1a5cb56f15&transform=1&allowOpen=true
- Wildfire GeoJSON module (Prof. David McDonald) https://drive.google.com/file/d/1TwCkvdaw0MxJzW7NSDg6XxYQ0dvaS44I/view
- Ipython Notebook for geodetic distance computation: https://colab.research.google.com/drive/1Dx6dEU0vqOFRlH-whnxevtE8MrRjfjAH?usp=drive_link
- AQI Metadata Explanation: https://www.airnow.gov/sites/default/files/2020-05/aqi-technical-assistance-document-sept2018.pdf
- AQI FAQs: https://www.epa.gov/outdoor-air-quality-data/frequent-questions-about-airdata
- Ipython Notebook to call EPI Air Quality History Data: https://colab.research.google.com/drive/14ni6Z1YPPGgitIY2WiOZuV0Dnjl3014n?usp=drive_link
- Facebook Prophet Library documentation: https://facebook.github.io/prophet/docs/quick_start.html
- Nebraska Public Health Reports: https://dhhs.ne.gov/Pages/Public-Health-Reports.aspx
- Nebraska Revenue Department Statistics: https://revenue.nebraska.gov/research/statistics/local-government-data
- EPA Research Study on Wildfire Effects on Health: https://www.epa.gov/air-research/wildland-fire-research-health-effects-research

DATA SOURCES

- Historical Wildfire Occurrences Dataset: https://www.sciencebase.gov/catalog/item/61aa537dd34eb622f699df81
- US EPA API - Air Quality IndexDataset: https://aqs.epa.gov/aqsweb/documents/data_api.html
- City Allotment: https://docs.google.com/spreadsheets/d/1cmTW5fgU3KyH6JbrRao-qWjzu2GovKk_BkA7a-poGFw/edit#gid=1247370552
- Cancer Cases in Nebraska Report (2006): https://dhhs.ne.gov/Reports/Cancer%20Incidence%20and%20Mortality%20in%20Nebraska%20-%202006.pdf
- Cancer Cases in Nebraska Report (20013): https://dhhs.ne.gov/Reports/Cancer%20Incidence%20and%20Mortality%20in%20Nebraska%202013.pdf
- Cancer Cases in Nebraska Report (2019): https://dhhs.ne.gov/Reports/Cancer%20Incidence%20and%20Mortality%20in%20Nebraska%202019.pdf
- Nebraska Population Data: https://revenue.nebraska.gov/sites/revenue.nebraska.gov/files/doc/MUN_1_Cert_Pop_Summary.xlsx