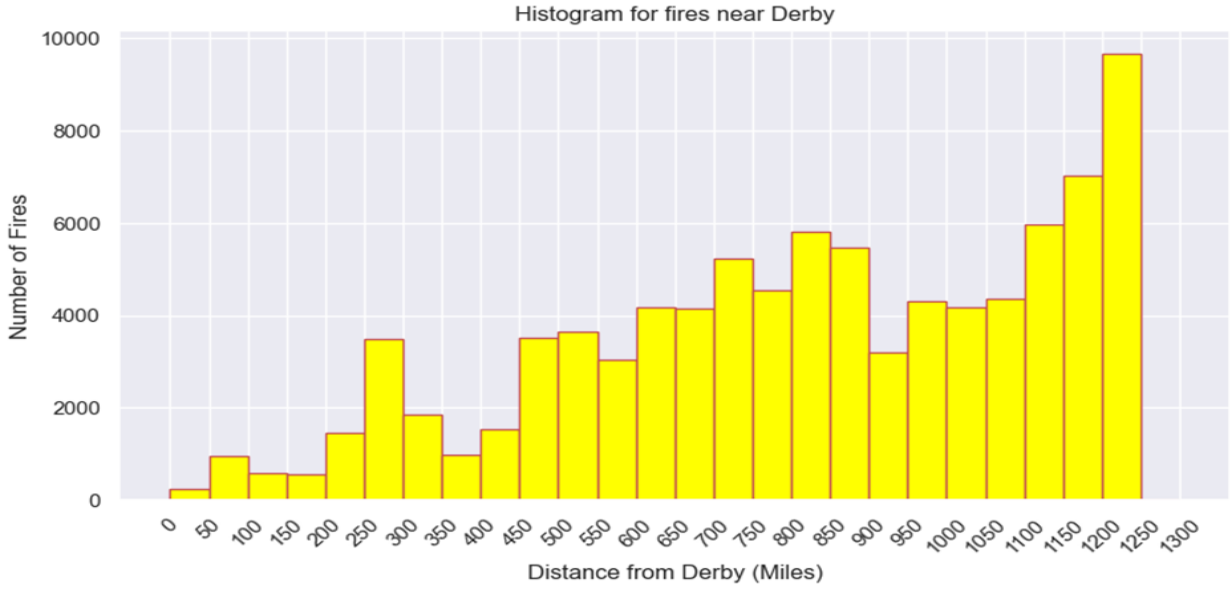
Common Analysis - DATA512

Prerit Chaudhary

Visualization 1

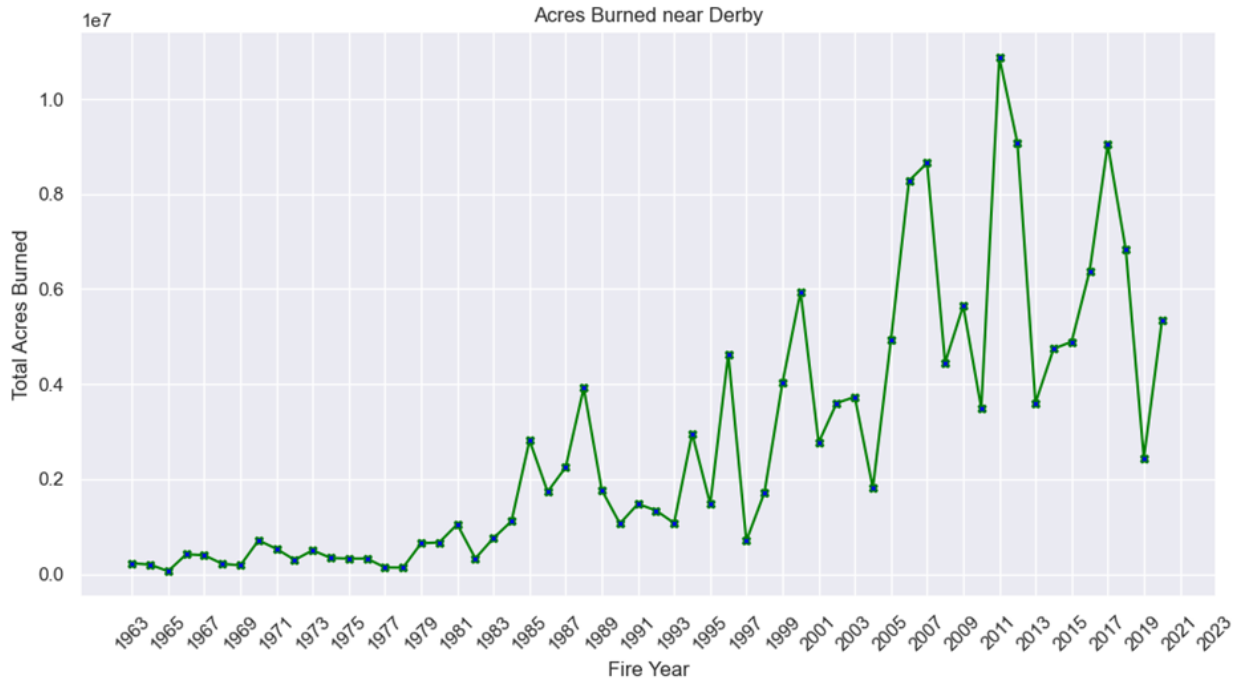
Produce a histogram showing the number of fires occurring every 50 mile distance from your assigned city up to the max specified distance



The above visualization is a plot which is representing the count of fires at an interval of 50 miles. The total range of distance considered is 1250 miles from Derby, Kansas and the data has been analyzed for the years between 1963 and 2023. The histogram has the count of fires on the y-axis and has distance from Derby in miles on the x-axis. The distance on the x-axis has been binned by 50 miles to improve the readability of the histogram and ensure reproducibility. The dataset for this analysis can be downloaded from the following link. The data was downloaded, read, cleaned and filtered to prepare the subset of data required for analysis. From the plot, it can be observed that the number of fires within a 200 mile distance from Derby are very less as compared to other distances. The count of fires up to 500 miles are also lesser in number and remain pretty much constant. The number of fire cases increases significantly as the distance from Derby increases above 500 miles and the cases seem to be increasing (not consistently) as the distance from Derby increases. The number of cases peak between 1200 – 1250 miles with approximately 10000 cases of fire. The lesser number of cases near Derby can be attributed to the awareness spread by the Fire Department of Wichita and the green spaces present in Derby which is the largest suburb in Wichita. The increase in the number of fire cases might be due to the terrain around Derby and the industrial areas outside the suburb which might be contributing to the increased number of cases.

Visualization 2

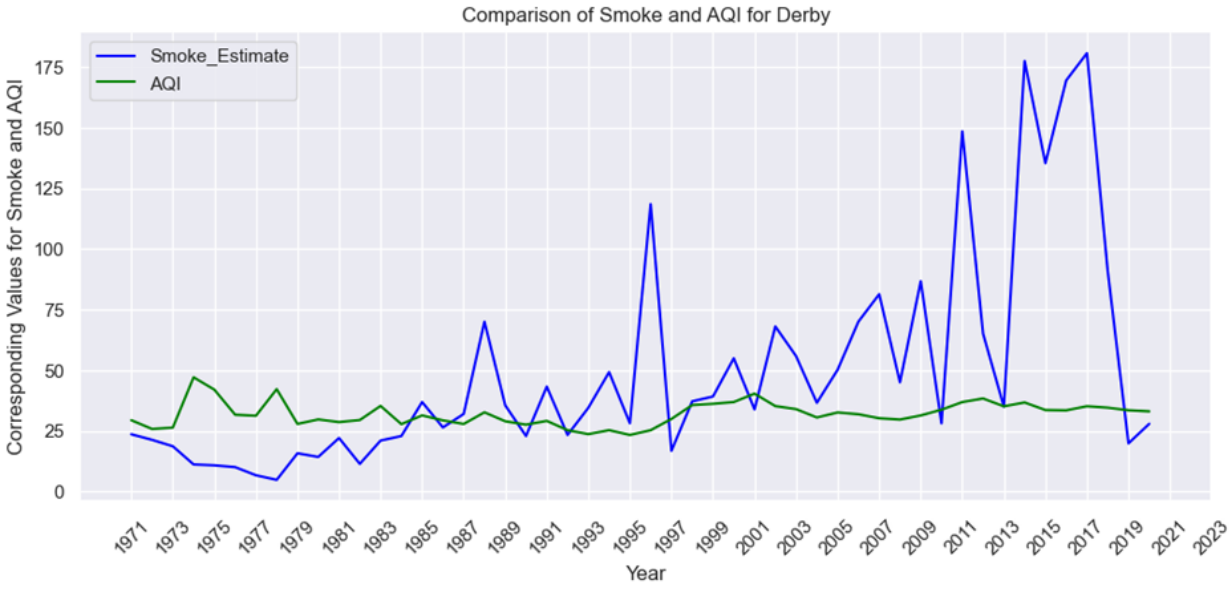
Produce a time series graph of total acres burned per year for the fires occurring in the specified distance from your city.



The above plot is a time series plot representing the total acres burned year over year by fire near Derby. The data for this analysis considered the areas within 1250 miles from Derby, Kansas and the data has been analyzed for the years between 1963 and 2023. The line plot has the area of acres burned on the y-axis between the period 1963 – 2023 which has been represented on the x-axis. The plot also includes markers to make it easier to gain insights about a specific year. A caveat to this analysis which can be noticed on the first glance of the plot is the absence of data beyond 2020. This means that there was no data available beyond 2020 for the total acres burned by fire which is within 1250 miles from Derby, Kansas. One possible hypothesis for this can be the commencement of the pandemic which had the world in a lockdown. The years on the x-axis have been represented at an interval of two years to ensure readability and reproducibility. The dataset for this analysis can be downloaded from the following link. The data was downloaded, read, cleaned and filtered to prepare the subset of data required for analysis. From the plot, it can be observed that the acre of land burned because of fire was significantly lesser till 1984 and almost remained constant. Beyond 1984, there has been a rise in the total acres burned because of fire which can be due to commercialization and increasing industrial activities. The acres of land burned due to fires can be seen to fluctuate after 1984 but overall it has been on the rise with only a few years seeing a dip. The total acres burned due to fire peak in 2011. Another key insight is that after 2011, the city has been taking measures to control the impact of fire as the acres of land seem to be decreasing at a slow pace ever since 2011.

Visualization 3

Produce a time series graph containing your fire smoke estimate for your city and the AQI estimate for your city.



The above plot is a time series plot representing the AQI and Smoke Estimate line plots over the years near Derby, Kansas. The data for this analysis considered the areas within 1250 miles from Derby, Kansas and the data has been analyzed for the years between 1963 and 2023. The two line plots have the values for AQI (represented by the green line) and the Smoke Estimate (represented by the blue line) on the y-axis between the period 1963 – 2023 which has been represented on the x-axis. A caveat to this analysis which can be noticed on the first glance of the plot is the absence of data beyond 2020. This means that there was no data available beyond 2020 for the total acres burned by fire which is within 1250 miles from Derby, Kansas. One possible hypothesis for this can be the commencement of the pandemic which had the world in a lockdown. The years on the x-axis have been represented at an interval of two years to ensure readability and reproducibility. The dataset for Smoke Estimate analysis can be downloaded from the following link. The data was downloaded, read, cleaned, and filtered to prepare the subset of data required for analysis. The formula for smoke estimate was based on the size and distance of the nearest point from Derby. The estimate was aggregated over a year to generate one final value for that year. The data for AQI was fetched through an API which had information regarding the particulate and gaseous pollutants. The two dataframes were then merged to do a comparative analysis year over year. From the plot, it can be observed that the AQI has remained almost constant over the years and as represented in the plot has been in the green zone which is between 0-50. The guidelines for AQI can be found here. The Smoke Estimate has increased ever since 1981 before which it was consistently low. One key trend is the decrease in Smoke Estimate beyond 2017 but it cannot be extrapolated with confidence as we only have a few data points beyond 2017.

Reflection

The aim of the project was to understand the impact of wildfires in different US Cities. I was assigned to the city of Derby, Kansas which is one of the largest suburbs in Wichita. The takeaways from this submission again highlighted the significance of a Human-Centered Approach in Data Science and Analysis. I collaborated with my peers like Ishank and Sagnik where we had multiple brainstorming sessions to tackle different challenges. This involved navigating through kernel failures, high response times in data retrieval, and dealing with messy datasets. Learning from students with a software engineering background was invaluable in structuring and optimizing the data pipeline effectively. I also understood the importance of a structured and modular approach of writing code and storing data in a structured format which can then be used for different downstream analysis. I went through multiple iterations of fetching data as I hadn't thought about the exact format and columns I needed in my final data

The second half of the assignment delved into interpreting trends and patterns revealed by the three visualizations we created using the filtered data. A key insight that was observed was the absence of data beyond 2020 for fires within 1250 miles of Derby. A possible hypothesis that I could think of was linked with the onset of the pandemic, imposing a global lockdown. Derby being one of the largest suburbs in Wichita is known for its green patches which is reflected by the less number of fire cases being reported in the proximity of Derby. However, the cases increase significantly as we move away which might be indicative of commercialization and industrialization.

There has been a significant rise in the acres of land being burned all over the world over the last few years, due to a number of factors such as climate change, industrialization, commercialization etc. The data which was filtered for Derby also reflected a surge in acres of land burned in recent years. The area of land being burned had been on a constant rise and it peaked in 2011 in Derby. However, there is a noticeable shift post-2011 where it can be seen that the acres of land being burned has started to decrease though not consistently. This would be indicative of the effective measures that would have been enforced to mitigate wildfires once this reached a breaking point. I aim to deep dive into this hypothesis of what measures were taken post 2011 in my extension plan and would try to find correlation with different industry-specific trends to validate my hypotheses and understand the factors contributing to this positive change.

Despite the challenges, the Air Quality Index (AQI) in Derby has consistently remained in the green zone, signaling the area's abundant tree coverage. This not only aligns with the community's commitment to environmental preservation but also emphasizes the potential correlation between green spaces and air quality.

For the forecasting model, I explored various techniques such as Regression, ARMA, ARIMA models but then ended up using SARIMA which is also able to capture seasonality. Though initially I haven't used months of the year to see seasonality but I believe this might be useful in my extension plan.