**Visualization Description and Reflection**

**Part1 Common Analysis**

**Neel Shah**

**Visualizations:**

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

**A graph of a graph

Description automatically generated**

This visualization is depicting the number of fires occurring every 50 mile distance from Del Rio, Texas up to the max specified distance (1250 Miles). Each bin in this histogram is 50 miles. The X-axis for this graph is the Distance from Del Rio (in Miles) and the Y-axis is the number of fires corresponding to that distance. The reader can read this graph as analyzing the number of fires close and far away from Del Rio. The left part of the graph shows a smaller number of fires which makes sense as Del Rio is a remote location on the border of Mexico. Thus, most of the US files will be far away from the city and which is evident by looking at the right part of the graph where the bars almost cross 10,000 fires.

All the counts above 5000 fires correspond a distance of more than 900-1000 miles from Del Rio. This graph was processed from the wildfire data extracted using GeoJSON from the USGS. This data was extracted for each fire and then filtered for fires within 1250 miles of distance from Del Rio. The dataset had information like fire year, fire type, area burnt, and we calculated the distance of the fire from Del Rio.

Overall, the graph gives agreeable insight about fires corresponding to Del Rio.

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

**A graph with blue lines and numbers

Description automatically generated**

This visualization is a time series graph of total acres burned per year for the fires occurring in the specified distance from Del Rio, Texas. The X-axis in the graph represent the year of fire and the Y-axis is the area burned in acres due to fire. This graph is created from the wildfire data extracted for USGS using geojson. The data contains information for each fire and we extracted all those fires close to Del Rio. The data (Acres burned) is then aggregated per year to show this line chart.

We can see that there is a clear pattern with how the area of ground burned per year is increasing as the time passes and we can see a clear exponential increase. The line, although with some ups and downs, is comparatively increasing upwards towards the recent years. We can also find a sharp spike in the total area burned in 2011 and 2017.

Fires from 2006 to 2020 are of very high intensity and also more in number, thus resulting in very drastically large land burned in comparison to fires from 1963 to 1990.

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

**A graph with blue and red lines

Description automatically generated**

This visualization is a time series graph containing the fire smoke estimate created for Del Rio and the AQI estimate for Del Rio. The graph is basically a comparison between the annual smoke estimate derived from the wildfire data and the actual annual AQI values from the monitoring sites near Del Rio. The X-axis of the graph represents the year of the fire and the Y-axis Is basically the value of Smoke Estimate/AQI.

This graph is created using 2 data frames. The fire data is extracted from USGS and is basically a combined data for all fires in the US. We extract all those fires which are within 1250 miles from Del Rio and started after 1963. The second data frame is the AQI data which is extracted using US EPA AQI for the AQI values from the monitoring stations around Del Rio. The Smoke Estimate is a derived variable created using Acres or area burned and the distance of fire from Del Rio. This is derived from the 1st data frame and then aggregated annually by taking the average of the smoke estimate. The AQI values extracted from the 2nd data frame is monitored daily which then is aggregated to an annual AQI value by taking average of the top 30 daily AQI values. Finally, this line chart is created by showcasing the annual smoke estimate and annual AQI value for each year from 1963 to 2020.

We can see that there is clear correlation between these two values. Even though due to lack of monitoring stations in Del Rio and AQI values starting from 1988, the lines are following a clear pattern and also seasonality. The Blue line is the Smoke Estimate and the Red line is the AQI value. The reader can also clearly identify the close relation and see matching of spikes, for example, in the year of 2011. We can also see that the values are in the 0-60 which is the indicator for Good Air Quality. This also matches to our initial finding that the fires are very far away from Del Rio thus confirming that there is very little or negligible impact of the fires on Del Rio.

**Reflection:**

This assignment had 3 steps on a high level which were:

1. Wildfire Data extraction for a specific city
2. AQI Data Extraction for a specific city
3. Data Analysis and Predictive Model Deployment

All these 3 steps were time-consuming and needed extreme assumptions. The data extraction for wildfire data involved downloading 2.8 GB of data which contained information for all ~135k fires/features. The data had to be filtered for Del Rio, Texas which was my assigned city. The filter included fires within 1250 miles of distance from Del Rio which started after 1963. So I used the sample code provided by the Prof to extract specific fires which took around 2 hours to run and I got 70861 fires corresponding to Del Rio. I also generated a Smoke Estimate based on the Area of land burned in Acres and Distance from Del Rio. The final annual smoke estimate was generated by averaging all the daily smoke estimates for a year.

The 2nd part included data extraction of AQI data which was an API call to US EPA to get AQI values for Del Rio. I couldn’t get any monitoring stations around Del Rio so I had to use the bounding box method. I found 1 station within 200 miles radius but when I reran the code to get the summary of data, I could only get data after 2005 which was not that useful. Thus I increased the radius to 250 miles and got 5 stations. I finalized the monitoring station at Brewster, Texas, and extracted AQI data from 1988 to 2020 (All I could get). The final annual AQI was then calculated based on averaging the top 30 daily AQIs for that specific year.

The last part was analysis and prediction. The Analysis was done on the smoke estimates and the AQI data per year. For Prediction, I decided to go with ARIMA model to get a time series forecast for the smoke estimates. Even though the Auto Arima provided me some values of p,q and d, I tweaked them to get a good prediction with values (4,2,3).

I was able to see clear correlation between smoke estimates and AQI for Del Rio. I observed most of the fires for Del Rio where far away and thus affected very less on the Air quality of Del Rio. Also Del Rio is on border on US and Mexico, thus I feel there can be further investigation if we get Mexico fires data. I also learnt that based on predictions, even by 2049, the air in Del Rio is going to be safer and the AQI will be between 0-60 which is very good according to the internet.

Overall the whole assignment was very assumption based, for example, the smoke estimate actually is based on so many factors like air, humidity, intensity of fire, human factors, etc., but we didnt have all these factors to get the final metric. I was able to discuss the approach with some of my collogues which helped me realize some different ways to approach the problem and devise the solution. Especially in the part where we needed to create the smoke estimates, there were numerous approaches one could have done that but due to collaboration, I was able to gain help for others to finalize my formula as well as implement my own method with additional tweaks. Also the data extraction was time consuming and involved multiple aspects. Collaborating with others helped me with checking my approach and gave me a reference point. I reused some of the multi-threading code which I discussed with some of my friends which helped me speenden up the process.

One more example was use of Google collab for AQI extraction. My local machine showed 2 hours 25 mins to run the whole extraction for AQI values from 1963 to 2020. But after using Google Collab, I was able to get the results in just 45 minutes.