Project Report For CS661: BIG DATA VISUAL ANALYTICS
2022-2023 Semester II

Seeing Beyond the Concrete: A Visual Journey into Rainfall and Air Pollution

Team members: Abhinav Anand, Akansh Agrawal, Anubhav Kalyani & Samanvay Lakhotia
Member emails: abhinav22@iitk.ac.in, akansh@iitk.ac.in, anukal@iitk.ac.in & samanvay@iitk.ac.in

IIT Kanpur

1. Introduction:

Visualizing air pollution data helps in the exploration of trends in air pollution levels across various regions and time periods. It will help in identifying the cause of high air pollution in certain highly polluted areas. Visualization of rainfall patterns can be used to understand the distribution of rainfall over geography, and over time, and hence identify impact of rainfall on agriculture and other sectors which are highly dependent on rainfall. These two factors help in developing the strategies for fighting against drought, excess rainfall, and climate changes by suitable prediction and statistical analysis of data. In this project, we have developed an interactive visualisation interface to analyse and forecast the rainfall and air quality distribution in India along with the study of identifying the relationship among the various parameters of air quality and the rainfall.

2. Tasks:

1. We began by applying various preprocessing techniques to improve the dataset quality. This involved addressing issues such as missing features using averaging techniques and other data cleaning methods. We also aggregated various complex datasets of GeoJSON files of India and its states, rainfall distribution patterns over a century and the Air Quality datasets for the further tasks to be performed.

2. Once the datasets have been preprocessed, we displayed the geographical distribution of the rainfall and air pollution data on the Indian map for the various regions. This further involved the application of focus and context feature to get information for particular regions/states.

3. An interactive time series plot for the rainfall distribution over a century for the different locations of India is generated to understand the trend and the underlying patterns. Further, various statistical modeling techniques have been used to forecast the rainfall patterns for each city.

4. The comparison of various parameters of Air Quality such as $PM10$, $PM2.5$, AQI, NO_2, SO_2 etc. is done using an interactive bi-variate plot against time for various locations of India. Moreover, various statistical modeling

techniques have been used to forecast the various air pollution parameters for various cities.

5. An interactive correlation matrix have been developed for the various parameters of air pollution and rainfall to draw the inference and analyse the trend among them. Moreover, this was a complex task since the rainfall data we have was from period 1901-2002 whereas the air pollution dataset was from 2015-2020. So, we used the statistical modeling technique developed in above task to generate the rainfall distribution data from 2002-2022 and use the overlapping interval and cities of the air quality dataset and the predicted/forecasted rainfall dataset to get the correlation plot.

6. An interactive Visualisation interface has been developed enabling various user-friendly features to visualise the entire tasks at various granularity levels. The system generated is easily deployable, and can be locally hosted as well. This interface can be used to study and predict the rainfall and air quality patterns, and identify the plausible reasons between the relationship of their various parameters, and is helpful in research and can also be utilised by the Indian Meteorological Department (IMD) to supervise, and advice the sectors of agriculture, weather, etc.

   The solution for each of the task performed is detailed in the coming section.

3. Proposed Solution:

We used Plotly and Dash [1] to build our web-app. Plotly has been used to create graphs, while Dash has been used to integrate the plots in an interactive web-app interface.

3.1. Dataset and its pre-processing:

We used datasets available from Kaggle ([2], [3], [4], [5], [6], [7], [8]) for both rainfall and air-quality. The air quality data ranged from 2015-2020, and contained data on various Air quality parameters (like $PM10$, $PM2.5$, AQI, etc.) of daily nature, for 26 Indian cities, while the rainfall data was monthly in nature, and ranged from 1901-2002, for 263 Indian cities.

We exerted considerable effort on the data-cleaning tasks. We handled missing data using the method of moving averages. We created new features in our datasets as well, where required. For eg- mapping each city to its corresponding state, using external data sources. To get the rainfall and air pollution parameters of a city or state mapped to the Indian map, we also used the Indian GeoJSON dataset along with the separate GeoJSON dataset for each state. Furthermore, it was necessary for us to carry out aggregation at various levels repeatedly. For eg- converting daily air quality data to monthly format, for comparison with the rainfall data, which was monthly in nature.

We also created a dataset by the intersection of Rainfall and Air Quality datasets, based on common cities.

3.2. Our Visualisations:

1. The first plot in the visualisation interface, is that of a map of India. This shows the entire rainfall and Air Quality data geographically at one go, so that the viewer can directly make complete sense of the **geographical distribution of the Rainfall and AQI data for India**. The cities for which the AQI data was available are marked using their latitude and longitude as red circular dots (Note that: AQI measuring equipment is expensive, hence very few cities can afford it, which is why very less number of cities have an AQI measuring equipment, and the number of cities in the AQI datasets are less in number). Additionally, the states are **shaded according to the amount of rainfall** they receive - the darker the colour is, the more rainfall they receive. For instance, we can immediately infer that the north-eastern states receive lots of rainfall, while Rajasthan recieves very scarce rainfall. For exact numerical data, we've implemented systems in place for both rainfall and AQI. Hovering over a state reveals the average rainfall in the state calculated using aggregation over all the cities in the state over a century of monthly rainfall data. Similarly, hovering over the red-marked cities reveals the average AQI value in the city. The plot is interactive, and allows for interactive exploration using **zooming**, **panning**, **azimuthal**, and **elevation** tilting features.

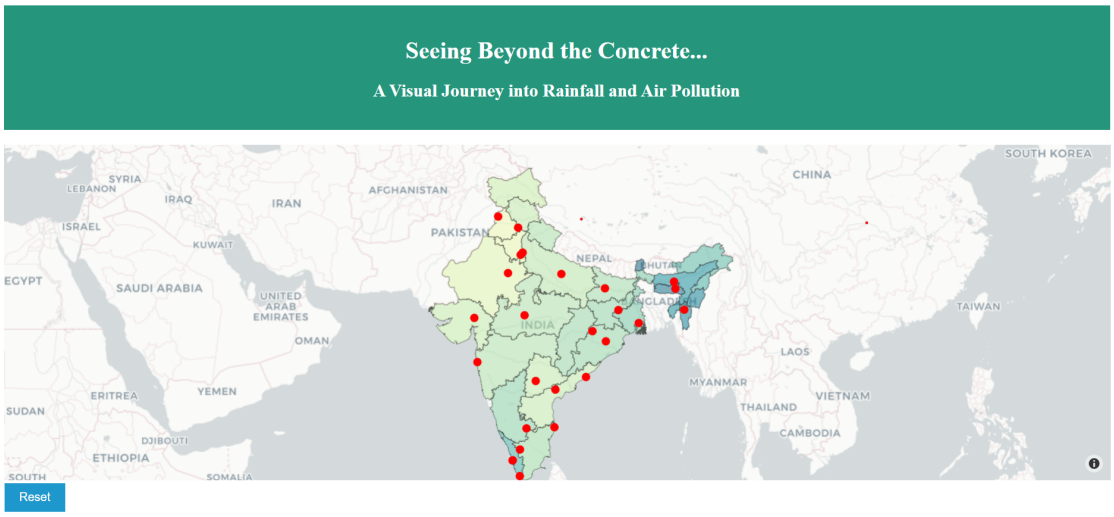

Figure 1: Map of India with reset button

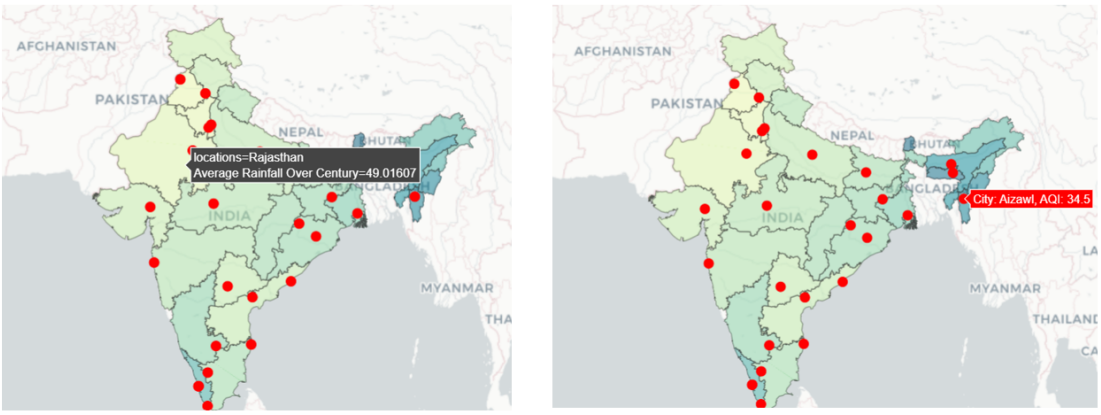

Figure 2: Left: Century-wise average rainfall information of a state when mouse is hovered over it, Right: Average AQI data of a city when a mouse is hovered over it

We include the concept of focus and context in our visualisation as well. When a state is clicked in the map, the entire map is zoomed to fit the state in the map. Here's an example for Rajasthan:

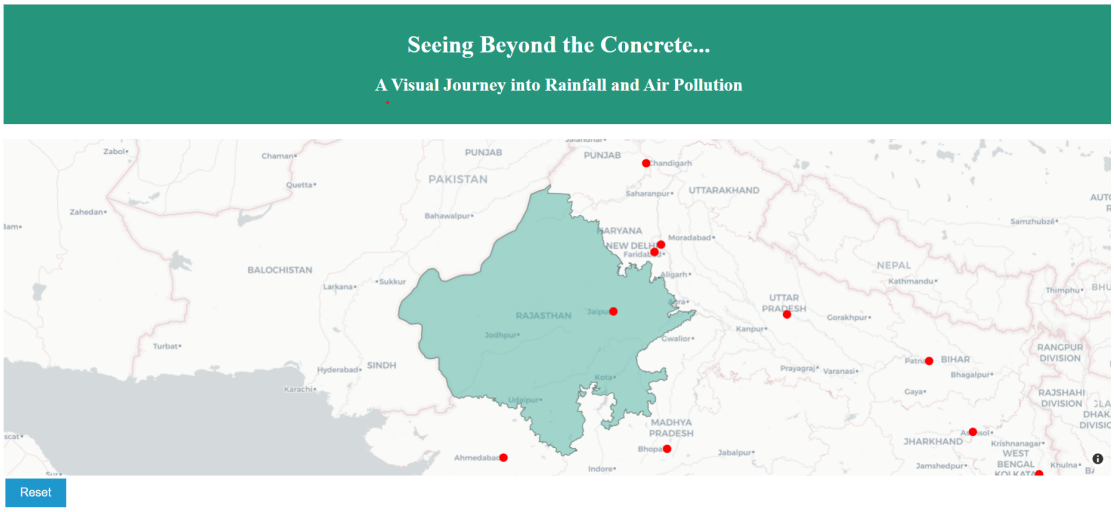

Figure 3: State of Rajasthan in focus when clicked

Now, since we are focusing on a particular state, all of the analysis below in the visualisation interface changes to that particular state only! All the interactive drop-downs in the below plots (that earlier included all Indian cities) now only include the cities within the selected state, since we're now only focusing on the particular state.

2. In the second plot, we have visualized the rainfall in a city over the past 100 years for a given month.

   This plot enables us to visualise the patterns and the cyclicity (not seasonality, since the period of fluctuations is more than 1 year) in the rainfall of India over a century.

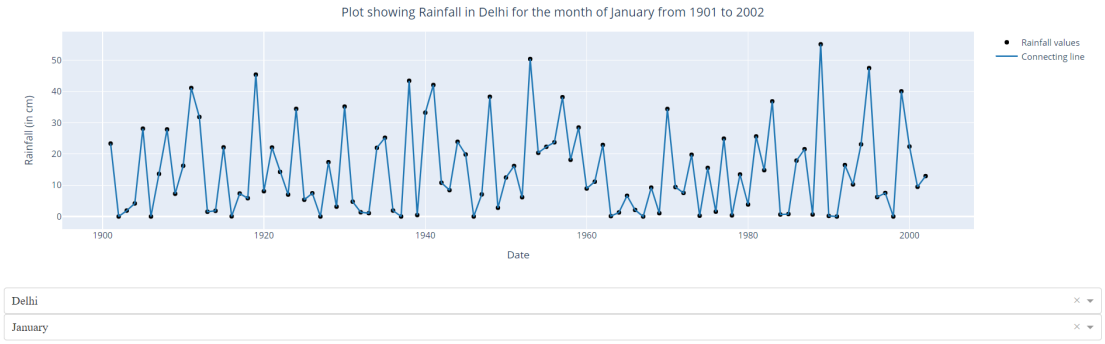

Figure 4: Rainfall in Delhi during January for the past 100 years

We have implemented two drop-downs in the plot: one for the city, and the other for the particular month that we want to look into.

3. In the third plot, we have forecasted the rainfall in a particular city. As described in the sub-section **3.3** below, we used the **Prophet** model for forecasting purposes, due to it's better predictions.

   We have first split the century-long month-wise data into 80% train and 20% test data. We train the **Prophet** model using the train data, and show it's predictions on both the train, as well as the test data in our plot. Alongside, we also show the confidence bounds of the predictions that the model makes. Again, the plot supports the zooming, and panning features, as present in the previous plots.

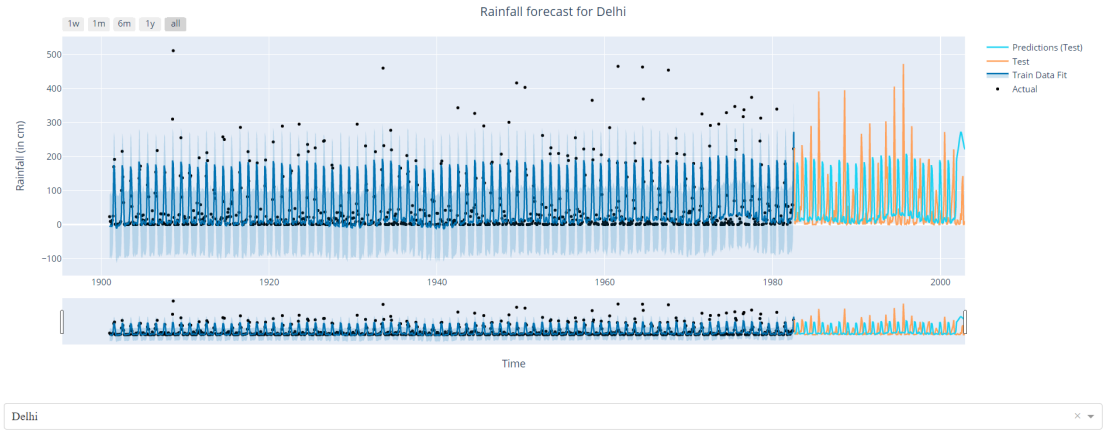

Figure 5: Rainfall forecast in Delhi for the next 20 years

Note that, we can **select the city** for which we want the predictions to be made, **using the dropdown box** at the bottom of the plot.

We have even further interesting interactive features in our plots! If we notice, we have a small subplot below. This enables us to implement the **brush-and-zoom** feature! We can simply drag the ends of the brush, and that zooms into the corresponding time-range on the x-axis.

Another interesting interactive feature are the buttons of '1w', '1m', '6m', '1y', and 'all'. Clicking on the buttons will respectively show only 1 week, 1 month, 6 months, 1 year, or all of the test data respectively.

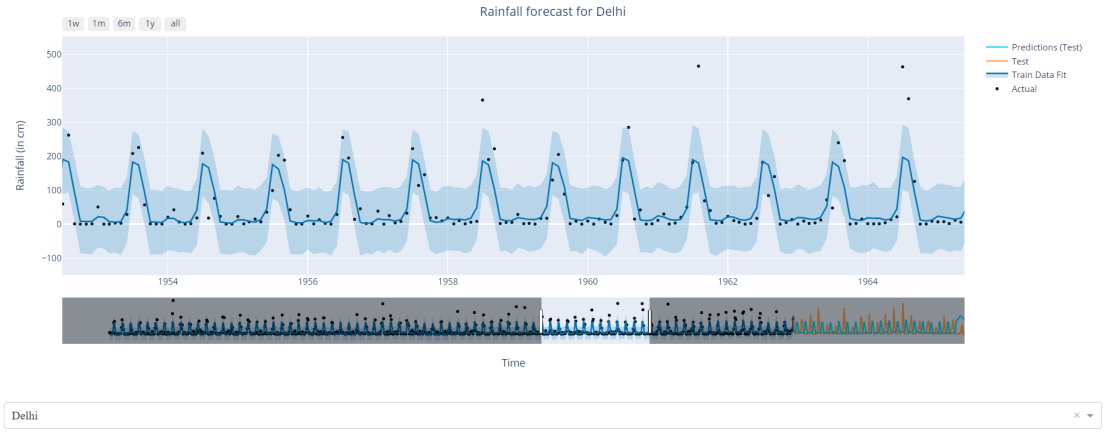

Figure 6: Showcasing the brush and zoom feature of the plot

4. In the fourth plot, we have visualized how two **air quality parameters** (AQI, $PM10$, $PM2.5$, NO_2, SO_2, CO, etc.) vary with respect to time using a **bi-variate plot**. The reason we have plotted two variables at once, is to be see the pattern and the correlation between the two air quality variables as well, while we can individually focus on the pattern, seasonality and variation of each of the air quality parameter with time by looking at the individual time-series plot.

   For example: In the figure below, we see the excellent correlation between AQI and $PM10$ in Delhi, where both fluctuate together. Similar plots are observed for multiple other bivariate air quality parameter combinations.

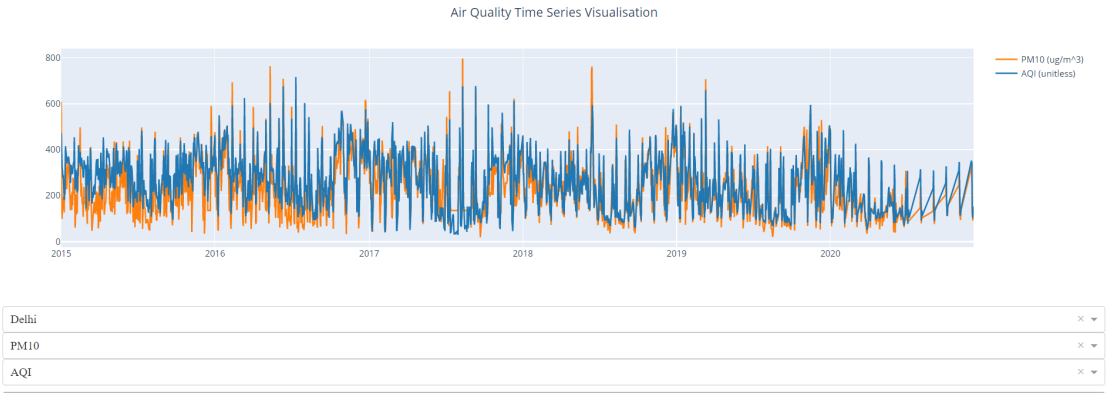

Figure 7: Variation of AQI levels and $PM10$ levels in Delhi

We offer 3 dropdown menus: One to select the city for analysis, and the other two, to select the air quality parameters for the bi-variate plot. Again,

the plot offers interactive features such as zooming and panning, for effective granular visualisation of the plotted data.

5. In the fifth plot, we create a visualisation of **AQI parameters** (AQI, $PM10$, $PM2.5$, NO_2, SO_2, CO, etc.) **forecasting** in a specific city. To carry out the predictions, we employed the **Prophet** model, which, as we describe in the subsection **3.3** below, provided superior forecasting results as compared to other methods.

   Initially, we divided the monthly data spanning 5 years from 2015-2020 into two portions - 80% for training purposes and 20% for testing. We then used the training data to train the **Prophet** model and plotted its predictions for both the training and test data. Moreover, we included the confidence intervals of the forecasts made by the model in the plot. Similar to the previous plots, this plot also supports zooming and panning functionalities. We have implemented the **brush-and-zoom**, as well as the '1w', '1m', '6m', '1y', and 'all' buttons as well, for effective visualisation of the forecasting plot.

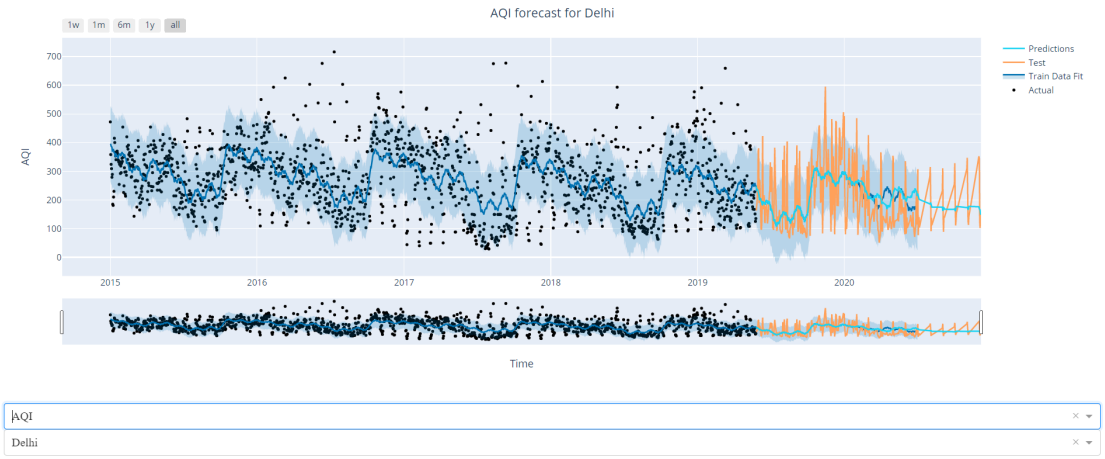

Figure 8: AQI forecast for Delhi

   Here, we have 2 dropdown menus: one for selecting the Air Quality parameter for which prediction is to be done, and the other for selecting the particular city for which we want the analysis.

6. In the sixth plot in our interactive visualisation interface, we have created a correlation matrix to visualize the correlation between the different parameters of air quality and rainfall for a city to see how parameters are correlated with each other. This especially enables us to visualise the relationship between Rainfall and Air Quality (the results of the analysis between the two is presented in the Results section below).

   The creation of the correlation matrix depended on the forecast created previously as well. The reason is: the rainfall data was monthly in nature, for a century from 1901-2002, while the air quality data was daily in nature, for

5 years only, from 2015-2020. Hence, in order to create a comparison between rainfall and air-quality, we used our forecasting model using **Prophet** to forecast the rainfall from 2002-2022. We also aggregated the air quality to a monthly level, in line with the rainfall data for comparison.

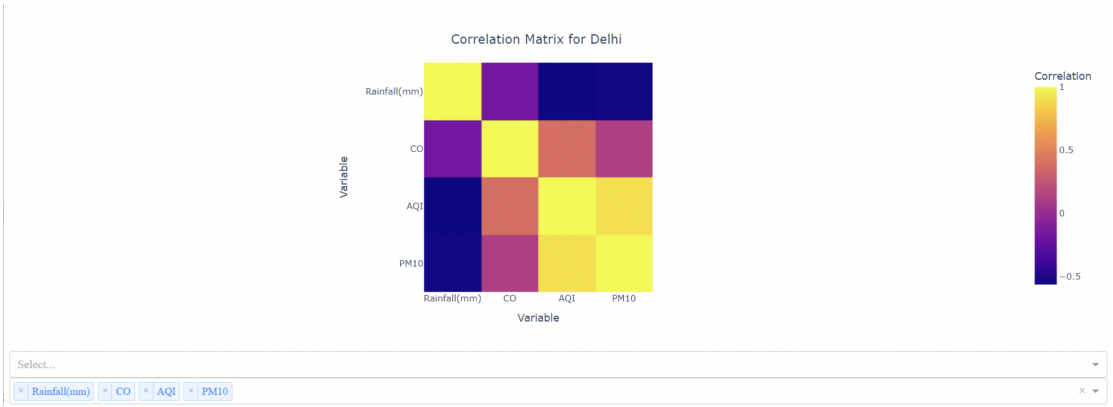

Figure 9: Correlation plot in Delhi for Rainfall, CO, AQI and PM10

We have included yet another useful new interactive feature along with the drop-down selectivity here. We have made a dynamic checkbox-like system, where you can select all the variables that we'd like to see the correlation amongst, and the correlation plot updates dynamically based on the selected variables!

3.3 Selection of Forecasting Model:

The preprocessed dataset for the rainfall (from 1901-2002) and air quality dataset (2015-2020) are used to get the forecasted interactive plots as discussed above. To select a model for our forecasting model, we implemented various statistical models such as **Holt-Linear** [9], **Holt-Winters** [10], Autoregressive Integrated Moving Average (**ARIMA**), Seasonal Autoregressive Integrated Moving Average (**SARIMA**) [11] and the **Prophet** ([12, 13]) model. We calculated the RMSE values using 20% test dataset of the original preprocessed data. The model with least RMSE was chosen. The RMSE for the 'Prophet' comes out to be the least in almost all the cities for both the rainfall and air quality datasets. The below table describes the RMSE values for AQI prediction of Delhi district using all these models.

Model	RMSE
ARIMA	114.018
SARIMA	114.097
Holt-Linear	596.57
Holt-Winters	127.22
Prophet	80.119

Figure 10: RMSE for the forecasting models

4. Results:

The visualisation system developed gives the insights of the trends of the rainfall and air pollution parameters. The parameters depicting air pollution such as AQI, $PM10$, $PM2.5$, SO_2, NO_2, etc. depicts the moderate to strong correlations in almost all the cities of India, confirming how air quality depletes (that is increase in AQI) over the years with the increase in emission of the above mentioned pollutants. Moreover, we are able to predict that how the AQI varies in future for various cities. Also, the rainfall time series forecast helps to identify the regions of heavy rainfall to the regions going to suffer draught. Additionally, we are successful to identify a trend between air pollution parameters and the rainfall.

City	Rainfall (cm)	AQI	Rainfall-AQI Correlation	Analysis
Guwahati	241.06	138.73	-0.57	Heavy Rainfall, Moderate AQI
Delhi	52.79	258.94	-0.56	Low Rainfall, AQI Poor
Patna	63.04	209.05	-0.54	Low Rainfall, AQI Poor
Hyderabad	76.53	105.3	-0.37	Moderate Rainfall, Moderate AQI
Jorapokhar	100.14	142.26	-0.35	Moderate Rainfall, Moderate AQI
Gurugram	51.78	214.02	-0.31	Low Rainfall, AQI Poor
Chandigarh	50.79	96.12	-0.25	Low Rainfall, Moderate AQI
Bengaluru	73.92	92.48	-0.23	Moderate Rainfall, Moderate AQI
Mumbai	115.65	73.61	-0.2	Moderate Rainfall, Low AQI
Ahmedabad	60.06	166.23	-0.16	Low Rainfall, AQI Poor
Bhopal	72.744	130.11	-0.09	Moderate Rainfall, Moderate AQI
Amravati	70.78	90.87	-0.05	Moderate Rainfall, Moderate AQI

Figure 11: Table showing rainfall is negatively correlated with AQI

We found out that rainfall is negatively correlated with AQI. This is shown in the above table for the major cities in India. The plausible reason behind this is that rainfall settles the air pollutants and helps in detoxifying the atmosphere. Coming to the cities such as Delhi, we observed that the low monthly rainfall over the years 2015-20, justifies the poor air quality in that region. Additionally, out study behind the industrial growth and urbanization ensures that cities not following the proper trend of negative correlation between AQI and rainfall confirms some heavy industrial or urbanization activities going on, which is a matter of concern and appropraite actions need to be taken immediately.

5. Conclusion:

In this project, we have developed an interactive visualisation interface which indeed helped to better analyse and get insights from the rainfall and air pollution datasets, thereby, looking beyond the concrete. The rainfall and air pollution parameters as seen have a hidden treasure of insights and their visualization, forecasting and correlation are the key gems to identify the impact on the various sectors. The study of these techniques help us to better develop the strategies for taking necessary actions against heavy rains, drought, air quality deterioration, and other climatic changes.

6. Link to source code:

The Github repository link :
https://github.com/Akansh-Agrawal/CS661A-Project.git

References

[1] https://dash.plotly.com/.

[2] https://www.kaggle.com/datasets/shrutibhargava94/india-air-quality-data.

[3] https://datasource.kapsarc.org/explore/dataset/district-wise-rainfall-data-for-india-2014/information/.

[4] https://www.kaggle.com/datasets/ravisane1/monthly-rainfall-data-india-of-a-century.

[5] https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india?resource=downloadselect=stations.csv.

[6] https://github.com/geohacker/india/blob/master/state/india_state.geojson.

[7] https://carto.com/.

[8] https://www.openstreetmap.org/copyright.

[9] https://towardsdatascience.com/forecasting-with-holts-linear-trend-exponential-smoothing-af2aa4590c18.

[10] https://www.analyticsvidhya.com/blog/2021/08/holt-winters-method-for-time-series-analysis/.

[11] https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/.

[12] https://medium.com/analytics-vidhya/time-series-forecasting-arima-vs-prophet-5015928e402a.

[13] https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/.