Data Analytics and Data Visualization Project Report

Study of the Air Quality of 1st and 3rd World Countries

Nathamayil Natesh, Rohaan Rangesh, Rohit Ramesh and Shravan Ganapathiraman

Scope, VIT Chennai

E-mail: nathamayil@gmail.com (Nathamayil Natesh), rohaan.rangesh@gmail.com (Rohaan Rangesh), rrohitramesh710@gmail.com (Rohit Ramesh), srivatsgana@gmail.com (Shravan Ganapathiraman)

Emissions of harmful gasses are produced in tons every single year and this has an adverse effect on human health and the surrounding environment. With respect to humans, it can cause respiratory tract problems and for the surrounding environment, is also impacted through outcomes like acid rain (due to sulfur dioxide emissions), polluted lakes and rivers and the depletion of the ozone layer. Some areas are worse than others and some areas which have negligible effects. This project aims to do a detailed study on the Air Quality Index of 1st World Countries (developed) vs Air Quality Index of 3rd World Countries (developing).

INTRODUCTION

We aim to analyze and display the researched dataset with respect to air quality throughout the years through suitable visualization techniques for a clearer understanding. To incorporate novelty features, we are planning to take the available datasets and use appropriate Machine Learning algorithms to predict how the Air Quality would look like over the period of next few years and also predict the death rates for the same. We will also draw correlations from different environmental, socioeconomic and economic factors to explain how the result can be supported with facts so that we have more credibility for our project. The main objective of this project is to extract the datasets ranging from many countries spanning 25+ years and also cities in India to visually display their air quality indexes, sum of indoor air pollution, sum of outdoor ozone pollution, and total air pollution with respect to human casualties. Other datasets include the emissions of various harmful gasses such as Nitrogen dioxide, Carbon Monoxide, Sulfur Dioxide, Ozone, Xylene and particulate matter 2.5 in various countries with respect to time. We then use Machine Learning models to predict the levels of Air Pollutants like Nitrogen Dioxide, Ozone and Sulfur Dioxide in the air. We also predict how the deaths due to air pollution would look like in the future.

MATERIALS AND METHODS

Machine Learning Models used

Linear Regression Model

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range rather than trying to classify them into categories. For our project, we have used Linear Regression to visualize how the pollutants look like over time, this helps us understand how the values are scattered across the plot.

Time-Series Analysis Model

One may utilize time series analysis to figure out what's causing trends or systemic patterns across time. It can reveal likely data changes, such as seasonality or cyclic behavior, allowing for a better understanding of data factors and improved forecasting. For our project, we used it to predict the levels of Air Pollutants like Nitrogen Dioxide, Ozone and Sulfur Dioxide in the air. We have used ARIMA forecasting algorithms to help us achieve this. An AutoRegressive Integrated Moving Average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

A statistical model is auto regressive if it predicts future values based on past values.

Gradient Boosting Algorithm

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model. Gradient boosting algorithms can be used for predicting not only continuous target variables (as a Regressor) but also categorical target variables (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.

LSTM Recurrent Neural Network

Time series prediction problems are a difficult type of predictive modeling problem. Unlike regression predictive modeling, time series also adds the complexity of a sequence dependence among the input variables. A powerful type of neural network designed to handle sequence dependence is called recurrent neural networks. The Long Short-Term Memory network or LSTM network is a type of recurrent neural network used in deep learning because very large architectures can be successfully trained. An LSTM model needs sufficient context to learn a mapping from an input sequence to an output value. LSTMs can support parallel input time series as separate variables or features. Therefore, we need to split the data into samples maintaining the order of observations across the two input sequences. After a prediction is made, it is fed back into the model to predict the next value in the sequence. With each prediction, some error is introduced into the model. To avoid exploding gradients, values are 'squashed' via (typically) sigmoid & tanh activation functions prior to gate entrance & output.

SOFTWARE AND LANGUAGES USED

- **Python:**
  Python is one of the most popular, interpreted high level language used heavily in machine learning and data science.
- **Tableau:**
  Tableau is a powerful visualization software that helps anyone see and understand their data. Connect to almost any database, drag and drop to create visualizations, and share with a click.

RESULTS AND DISCUSSION

Initial Visualization using Tableau

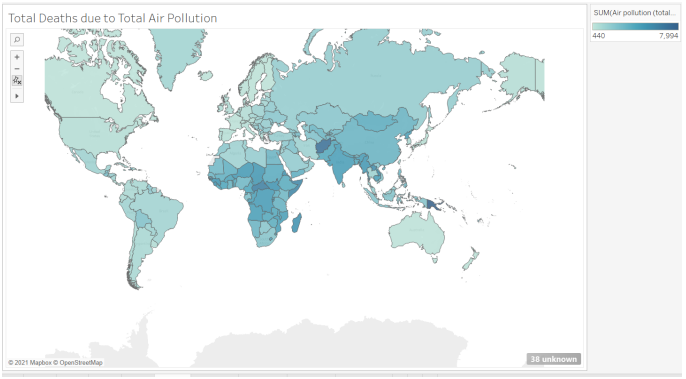

Figure 1. Map based on Longitude (generated) and Latitude (generated). Color shows sum of
Air pollution (total) (deaths per 100,000). Details are shown for the Country

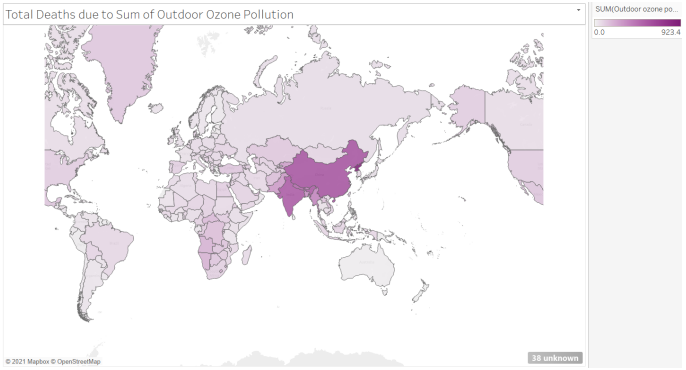

Figure 2. Map based on Longitude (generated) and Latitude (generated). Color shows the
sum of Outdoor ozone pollution (deaths per 100,000). Details are shown for the
Country.

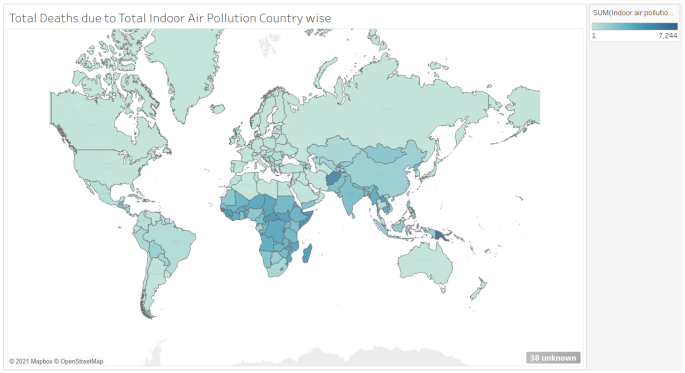

Figure 3. Map based on Longitude (generated) and Latitude (generated). Color shows the sum of Indoor air pollution (deaths per 100,000). Details are shown for the Country.

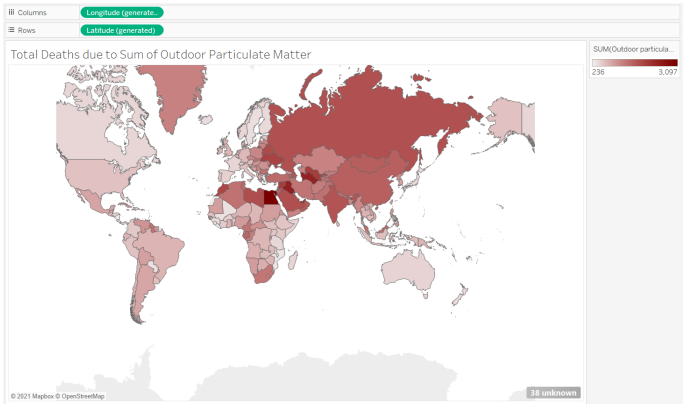

Figure 4. Map based on Longitude (generated) and Latitude (generated). Color shows sum of Outdoor particulate matter (deaths per 100,000). Details are shown for Country.

Prediction using ML Models

Linear Regression

It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables they are considering and the number of independent variables being used. In our project, you can see that we have used it to predict the Total deaths due to Air Pollution, the Deaths due to Indoor Air Pollution, Deaths due to Outdoor Particulate Matter and the Deaths due to Outdoor Ozone Pollution.

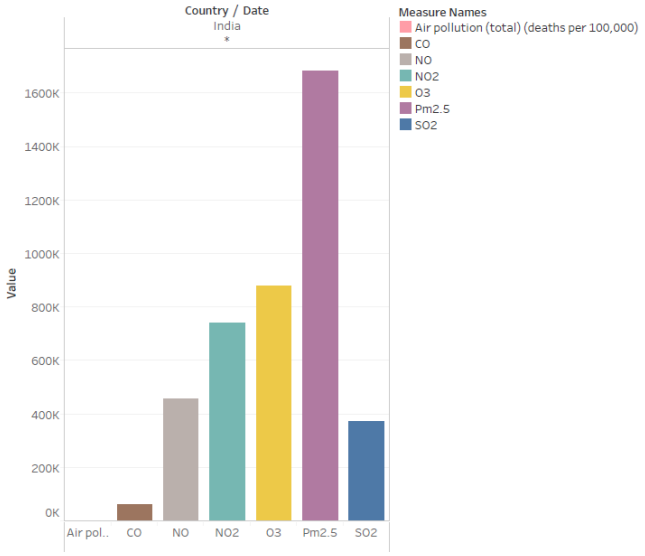

Figure 5. CO, NO2, NO, O3, Pm2.5, SO2 and Air pollution (total) (deaths per 100,000) for each Date (city_day) Year broken down by Country. Color shows details about CO, NO2, NO, O3, Pm2.5, SO2 and Air pollution(total) (deaths per 100,000). The data is filtered on Year and Date (city_day) Year. The Year filter ranges from 2015 to 2017. The Date (city_day) Year filter ranges from 2015 to 2020. The view is filtered on Country, which keeps India.

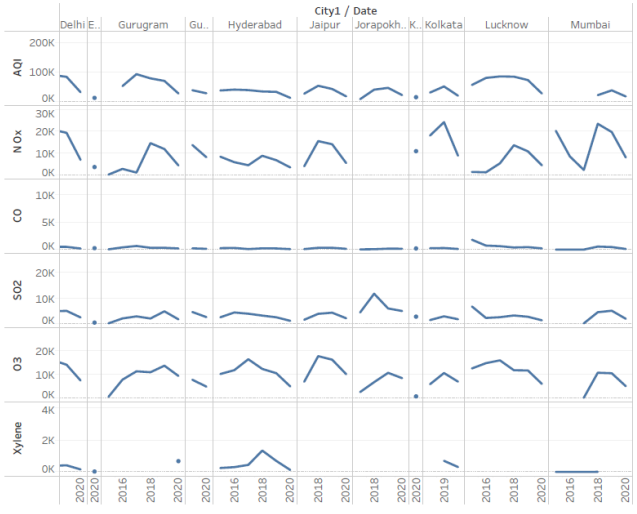

Figure 6. The trends of sum of AQI, sum of N Ox, sum of CO, sum of SO2, sum of O3 and sum of Xylene for Date Year broken down by City

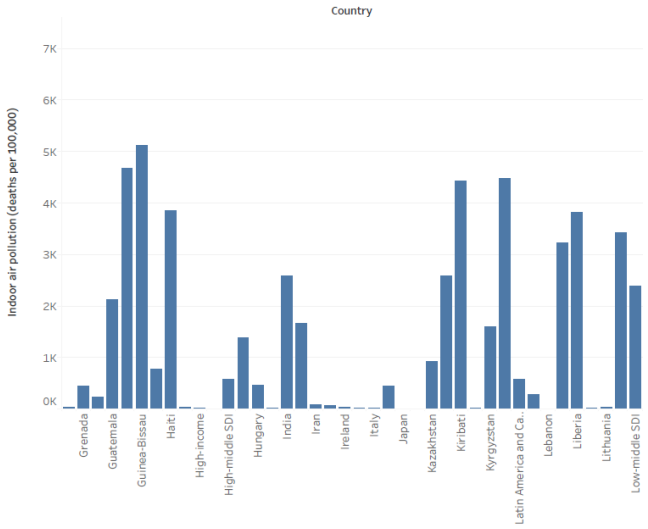

Figure 7. Sum of Indoor air pollution (deaths per 100,000) for each Country.

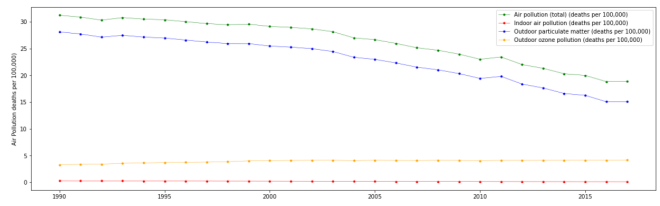

Figure 8. Plotting the Predicted Values

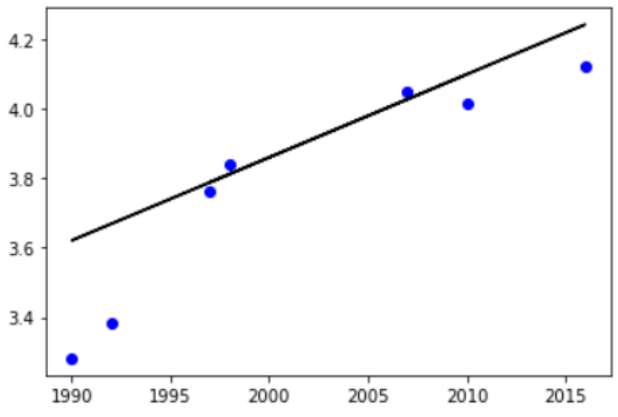

Figure 9. Model Fit

Gradient Boosting Algorithm

For our project we used it as a Regressor and computed the Mean Square Error with respect to the Air Pollutant factors from the dataset, they are namely Ozone and Sulfur Dioxide.

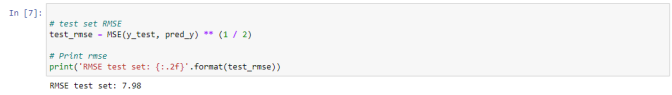

Figure 10. RMSE Test Set Value

Long Short-Term Memory:

The dataset was split into a training and testing model where the training dataset was trained using feature sets and then fitting it into the model under Keras in Python, and then testing it using the testing data and predicting the levels of O3 and SO2 for the next 60 days. We use this model to accurately predict and forecast the pollution levels of the future. This also gives us a high prediction accuracy when compared to the other models.

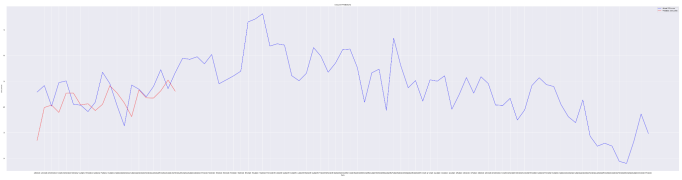

Figure 11. O3 Level Prediction in Delhi for the 60 days from the last date in the dataset

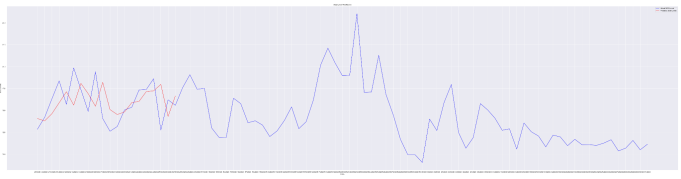

Figure 12. SO2 Level Prediction in Delhi for the 60 days from the last date in the dataset

Time Series

For Ozone Levels For this, we used out Indian Air Pollution Dataset and filtered the dataset to visualize India's most polluted (in terms of Air Quality) city, Delhi.

Now we try to implement our ARIMA Time series model and in the pre training of the model, we try to visualize the Original Series, 1st Differencing and 2nd Differencing.

In [69]:
```
df.head()
```

Out[69]:

	Unnamed: 0	City	Date	NO	NO2	NOx	CO	SO2	O3	Benzene	Toluene	Xylene
0	10229	Delhi	2015-01-01	69.16	36.39	110.59	15.20	9.25	41.68	14.36	24.86	9.84
1	10230	Delhi	2015-01-02	62.09	32.87	88.14	9.54	6.65	29.97	10.55	20.09	4.29
2	10231	Delhi	2015-01-03	25.73	30.31	47.95	10.61	2.65	19.71	3.91	10.23	1.99
3	10232	Delhi	2015-01-04	25.01	36.91	48.62	11.54	4.63	25.36	4.26	9.71	3.34
4	10233	Delhi	2015-01-05	14.01	34.92	38.25	9.20	3.33	23.20	2.80	6.21	2.96

Figure 13. Overview of the Dataset after filtering

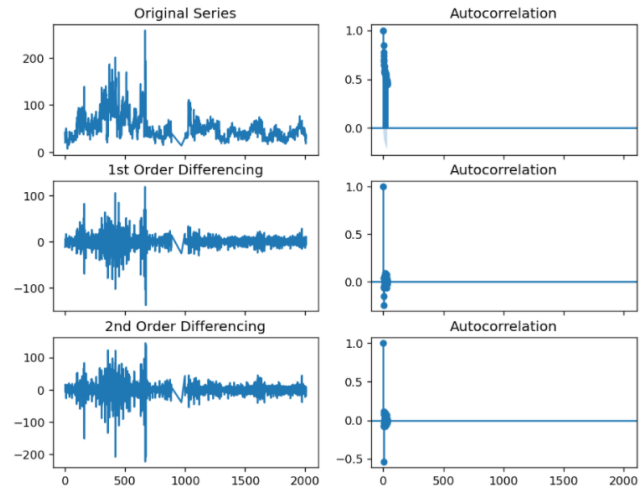

Figure 14. Plot for Original Series, 1st Differencing and 2nd Differencing

Now, after fitting our ARIMA model, we view the Residuals and Density.

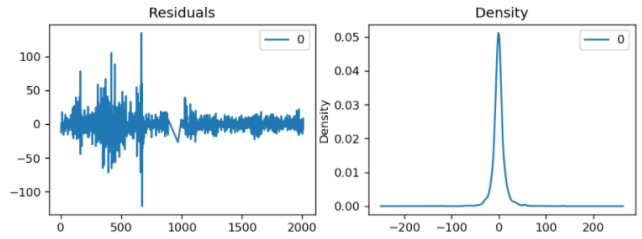

Figure 15. Plot for Residuals and Density

We have successfully predicted the Ozone levels in Delhi using our ARIMA Time series model. Now, similarly we predict the NO values and SO2 values and then plot it on the graph.

CONCLUSION

We have successfully predicted the Total deaths due to Air Pollution, the Deaths due to Indoor Air Pollution, Deaths due to Outdoor Particulate Matter and the Deaths due to Outdoor Ozone

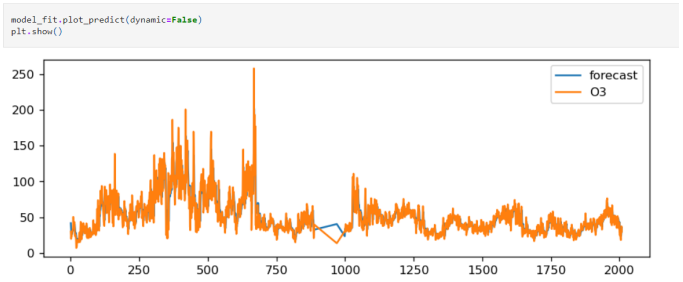

Figure 16. Plot for Ozone Pollutant where the blue line represents the Forecasted values

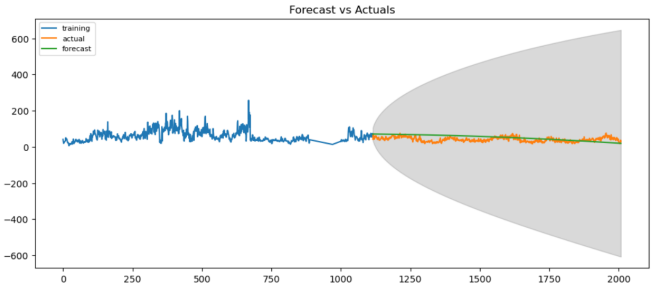

Figure 17. Forecast vs Actuals plot to see how the forecasting took place and how the training curve looks like vs the actual curve.

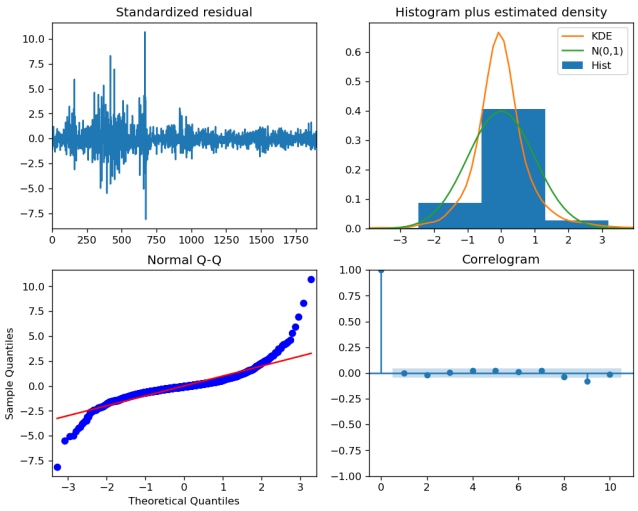

Figure 18. Plot to Visualize the Standardized residual, Histogram plus estimated density, Normal Q-Q and Correlogram

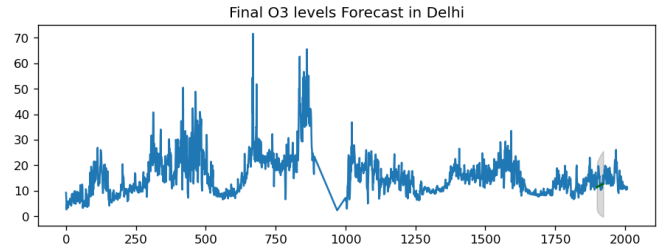

Figure 19. The final Ozone levels Forcasted for Delhi

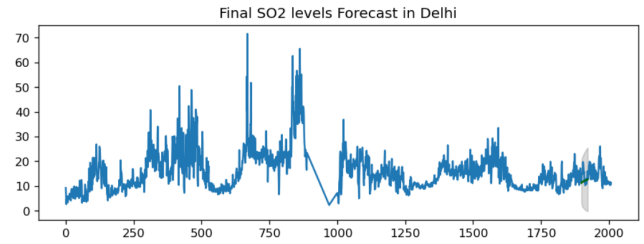

Figure 20. The final SO2 levels Forcasted for Delhi

Pollution. We have also predicted how the pollution levels of most commonly found pollutants like NO2 and O3 will look like using the models we mentioned above. And after predicting and successfully plotting our ML models, we got the accuracy values of the models to be what is shown in the table below.

Table 1. Summary of ML Models

Model	Accuracy
Linear Regression	97.6%
RMSE Test Set (Value)	17.89
Time Series (ARIMA)	92.82%
LSTM	98.23%

DATASETS USED:

1. https://www.kaggle.com/rohanrao/air-quality-data-in-india select=station_hour.csv
2. https://www.kaggle.com/akshat0giri/death-due-to-air-pollution-19902017
3. https://www.kaggle.com/open-aq/openaq