**Global Air Pollution 2010-2017**

**Introduction –**

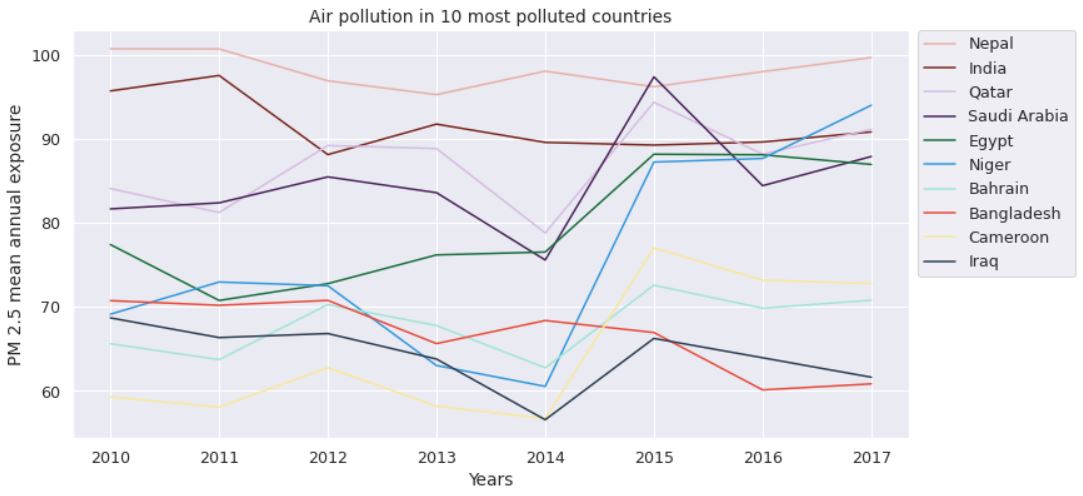
Air pollution is the need of the hour. According to WHO, air pollution kills about 13 people every minute in the world, more than 3 times the combined amount of Malaria, Tuberculosis and AIDS. It has also caused a tremendous increase in the global warming levels across the world, but the worst is yet to come. According to scientists, the global health is said to go down by 3% by 2025. Thus, it is a very important time to analyse and stop the unnecessary increment of air pollution. This project performs an analysis on the global air pollution levels from 2010-2017.

**Data sources and description** –

Credits to the dataset goes to - "Karl Weinmeister". Please find the dataset attached at - <https://www.kaggle.com/kweinmeister/pm25-global-air-pollution-20102017>. The data has been edited and modified to suit the purpose of the project. This data is a measure of mean annual exposure (micrograms per cubic meter) of PM2.5 across several countries from the year 2010 to 2017. PM2.5 also called as Particulate Matter 2.5, is a mixture of solid, liquid and aerosol particles that are suspended in the air. The 2.5 refers to their diameter of 2.5 micrometer. PM 2.5 is probably the most important factor in air pollution detection. For more information regarding PM 2.5 please visit - <https://www.airveda.com/blog/what-is-pm2-5-and-why-is-it-important>.

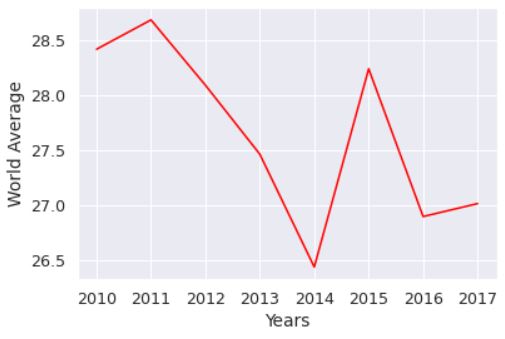
**Methodology –**

The data contains a total of 197 countries with their annual mean PM 2.5 exposure values from 2010 – 2017. The first step is the plot the data across the timespan. However, plotting 197 data lines in a single plot won’t make any sense. Hence, the data has been sorted according to the mean annual exposure of PM 2.5 and the top 10 most polluted countries have been plotted across the timespan of 2010 – 2017. According to WHO, the safe limit for mean annual PM 2.5 exposure is 35 micrograms per cubic meter. Following is the plot of the 10 most polluted countries in the timespan of 2010-2017 –

By observing the line plot, following points can be deduced -

* **Nepal** is the most polluted country across the timespan of 2010-2017.
* In 2015, **Saudi** **Arabia** caused more pollution than Nepal. However, the pollution level of Saudi Arabia decreased after 2015.
* **Niger** and **Cameroon** faced the highest rise in pollution level from 2010 to 2017. The pollution of **Niger** rapidly increased to become the 2nd most polluted country in 2017.
* **Bangladesh** and **Iraq** were able to reduce their annual pollution level across the timespan of 2010 to 2017.
* By observing the 10 countries, it can be seen that the pollution level varies from **60** **micrograms** to **100** **micrograms** per cubic meter. This hints at the fact that an overwhelming number of countries are doing good to moderately well in countering air pollution.

Since, mean annual exposure doesn’t give an idea of how good or bad a country performed in the timespan; the data was sorted according to the change in the pollution level from 2010 to 2017. The data was sorted and adjusted, where high positive value means better performance in countering air pollution while low negative values hints towards failure in combating air pollution. Following is the sorted list of top 5 and bottom 5 countries and their performance in combating air pollution.

As it can be seen from the lists, **Sri Lanka** and **China** did extremely well in combating air pollution across the years, while countries like **Cameroon**, **Niger** and **Nigeria** have failed extremely in abating air pollution.

There have been mixed performances from the countries. However from the following plot, we can safely infer that the world average of air quality has increased by a small 0.4%, after suffering a blow in 2015.

Further, unsupervised learning has been used to cluster the countries according to set pattern. The clustering has been done using K Means and it can be used to group the data into 3 clusters, namely –

* Lowly Polluted Countries (119).
* Highly Polluted Countries (16).
* Averagely Polluted Countries (62).

The biggest cluster is of 119 countries which are the ones with mean annual exposure of PM 2.5, well below the safety level of **35 µg/m3** set by WHO. The second biggest cluster is of averagely polluted countries whose mean exposure varies from **27 – 53 µg/m3**. The third cluster, which is the smallest cluster and comprises of only 16 countries are the actual pollution hotspots of the world and contribute to 37.31% of the world mean air pollution.

After the clustering has been performed, a classification model is trained on the data. The classification model used is Logistic Regression, with an average accuracy of 81%. For any new data of a country not present in the dataset, the model can predict the cluster label for the new data. This can help the country in taking necessary steps in abating further air pollution.