

A dark blue vertical bar is on the left side of the page. A blue arrow points to the right from the bar, containing the date 5/1/2020.

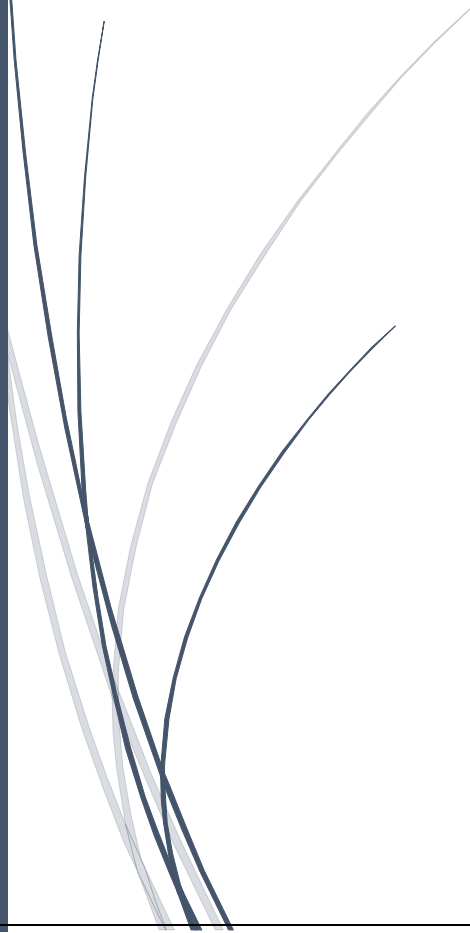
5/1/2020

FIT5147 – Data Exploration and Visualization

Semester 1, 2020

Data Exploration Project

Air Quality in India and its Causes - Report

Several thin, curved, light blue lines originate from the bottom left and sweep upwards and to the right, creating a decorative, organic shape.

Submitted by: Abhilash Kale [30254140]
Activity Number: 04-P1
Tutor Name: Farah Tasnuba Kabir

Table of Contents

1.	Introduction.....	1
2.	Data Wrangling.....	1
2.1	Ambient Air Quality for India Dataset.....	1
2.1.1	Missing Value Treatment.....	2
2.1.2	Air Quality Index	3
2.2	Dates for Diwali Dataset	3
2.3	Registered Vehicles in India Dataset.....	3
2.4	State-wise Registered Vehicles in India Dataset	4
2.5	Shapefiles – India.....	4
3.	Data Checking.....	4
3.1	Categorical Data.....	4
3.2	Numerical Data	5
3.2.1	Correlation	5
3.2.2	Outlier Detection.....	5
4.	Data Exploration	6
4.1	Distribution of different air pollutants in India through the time period	6
4.2	The role of location for the Air Quality in India	7
4.3	Different factors affecting the Air Quality in India	8
4.3.1	Vehicles in India	8
4.3.2	Type of area – Industrial, Residential and Sensitive	9
4.3.3	Diwali Festival Fireworks	9
5.	Conclusion	10
6.	Reflection.....	10
7.	Bibliography	11
	References.....	11
	Appendix	12

Table of Figures

Figure 1. Percentage of null values in the columns	2
Figure 2. KDE plots for comparison of before and after data imputation	2
Figure 3. Distribution of SPM throughout the years.....	3
Figure 4. Duplicate values in the type column	4
Figure 5. SPLOM for the 4 pollutants	5
Figure 6. Boxplots for SO ₂ , NO ₂ , RSPM and SPM.....	5
Figure 7. Descriptive statistics of the air pollutants.....	6
Figure 8. Distribution of pollutants in different states of India through the years	6
Figure 9. Air quality at various locations in India	7
Figure 10. Distribution of AQI for all states in India	7
Figure 11. Distribution of vehicles and the AQI through the years	8
Figure 12. Distribution of vehicles and AQI in India	8
Figure 13. Effect of different types of areas on the AQI	9
Figure 14. Pollution in India during Diwali and during the rest of the year	9
Figure 15. Level of AQI with respect to Diwali	10

1. Introduction

Pollution, climate change, global warming and environmental crisis are the growing fears for everyone in today's world. Globally, air pollution has been one of the most dangerous in terms of health risks. The growth of industries and the rise in the numbers of vehicles on road have major contributions towards it. According to the World Health Organization, 9 out of 10 people breathe the air which comprises high amounts of pollutants, and unfortunately, about 7 million people die each year due to the exposure to polluted air (Osseiran & Lindmeier, 2018). India, the world's second most populous country, is highly endangered because of its rising air pollution. Moreover, a news article also suggests that India's air quality is even more lethal than the leaders in global population, China (Anand, 2017).

As the time has passed, the growth in industrialisation and modernisation has been exponential in India, resulting in a massive increase in the air pollution. An analysis of these pollutants, resulting into a single valued Air Quality Index (AQI), and their causes through the different states in the country, will help us understand its rising levels and the effects it might create.

This report is an exploratory data analysis on India's ambient air quality data. It focuses on the major air pollutants and their levels which participate in polluting the air in India. The data is also explored by considering the causes of air pollution like industrial pollution, vehicular pollution and Diwali festival fireworks, spread throughout the nation. The analysis intends to find anomalies through distributions within the data and provide the appropriate treatments. The report further aims to answer the proposed research questions.

The data exploration is supported with the help of suitable visualizations and statistical explanations through R and Tableau Desktop.

2. Data Wrangling

2.1 Ambient Air Quality for India Dataset

The analysis is primarily based on the Historical Daily Ambient Air Quality Data for India during the years 1987 to 2015 ("Historical Daily Ambient Air Quality Data," 2017). This large data with 435,742 rows and 13 columns, is contributed by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).

The dataset is an aggregated version of the datasets available at:

<https://data.gov.in/catalog/historical-daily-ambient-air-quality-data>

The dataset is imported into RStudio and looking at the columns and the patterns its values exhibit, the following steps were taken as the pre-processing of the categorical data using R libraries:

- The values in 'sampling_date' column, can be considered as extra information as it is a near duplicate of the 'date' column, and hence, the column is dropped from the dataset. Similarly, considering the 'stn_code' column as extra information, it is also dropped.
- To shorten its name, the 'location_monitoring_station' column is renamed to 'station'.
- Date column is a string data type and needs to be converted into a date type column. For ease of merging of the dependent datasets, it is further separated into 3 different columns: day, month and year.
- Uttarakhand and Uttaranchal are two different names for the same state in India. The dataset consists differentiated information for this state under both these names. Hence, both the values are renamed as 'Uttarakhand' in the dataset to merge the separated information.

The data contains a lot of null values in various columns. We can check that using a column graph.

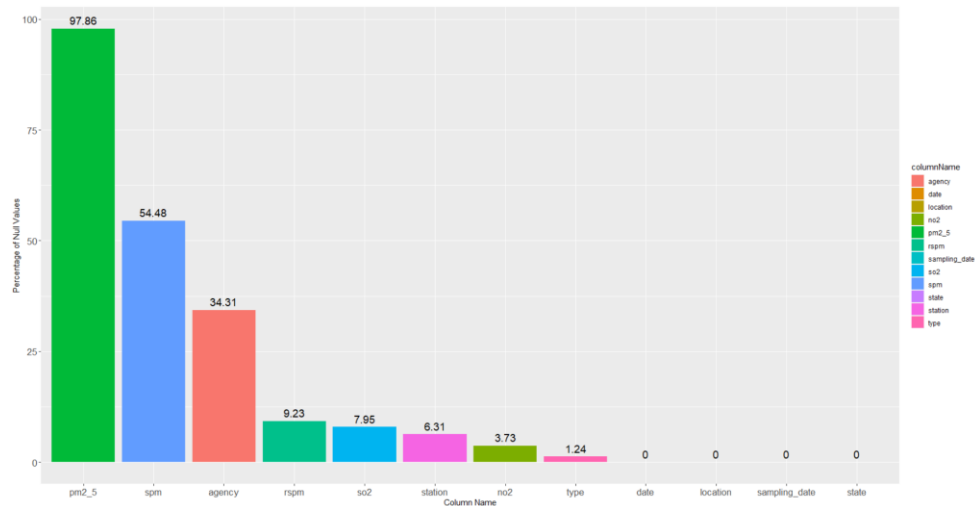


Figure 1. Percentage of null values in the columns

2.1.1 Missing Value Treatment

Considering the categorical data first, the null values in the ‘type’, ‘agency’ and ‘station’ columns were replaced with an ‘Unavailable’ value in the dataset in R, as the data is unavailable and also for easy categorization.

Now, for numerical data, as shown in figure 1, the pollutant data contains a significant amount of null values. The ‘pm2_5’ pollutant column consists of 97.86% of null values. Hence, instead of imputing values into these many rows, the dimension is reduced by dropping the column from the dataset to protect the data’s integrity and the exploration’s accuracy.

The missing values for the other 4 pollutants can be imputed using the forward fill or mean imputation techniques. Let’s explore the pollutants by comparing the actual data with possible imputed data using the kernel density estimation (KDE) plots to check their skewness along with the mean and median positions.

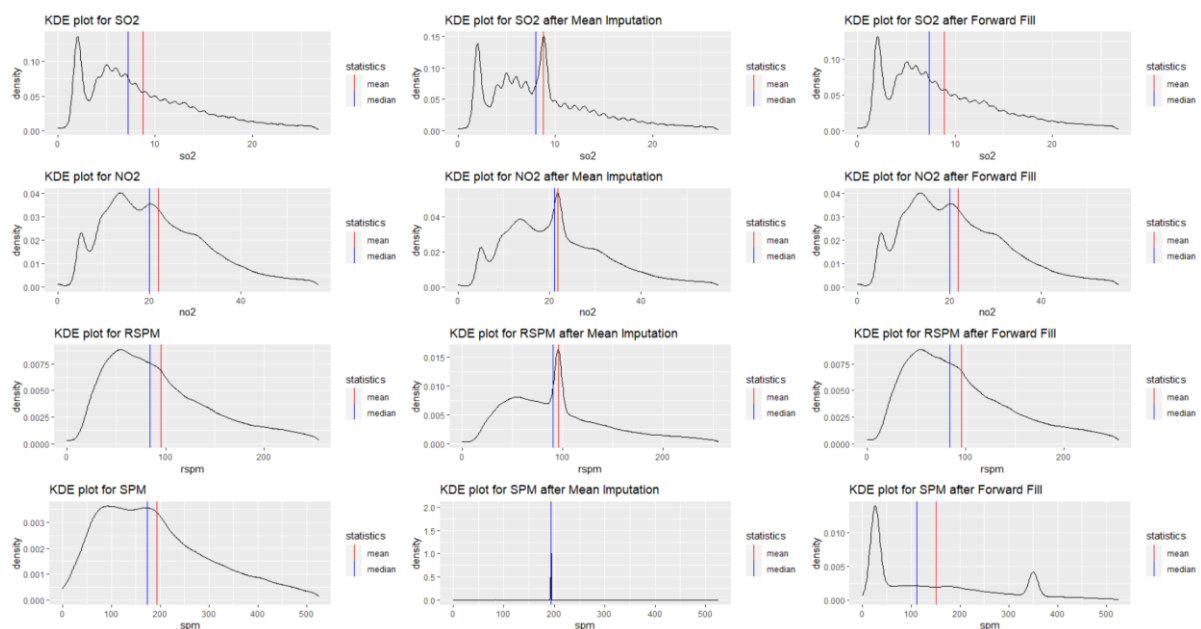


Figure 2. KDE plots for comparison of before and after data imputation

The data for all pollutants is positively skewed. We can clearly see that the distribution of the data is hardly affected with the forward fill imputation technique in comparison with the mean imputation technique. SPM is an exception, as it had 54.48% of null values and that creates plenty of assumed values in both the imputation methods. Hence, the null values for all the pollutants were imputed using the forward fill method in R.

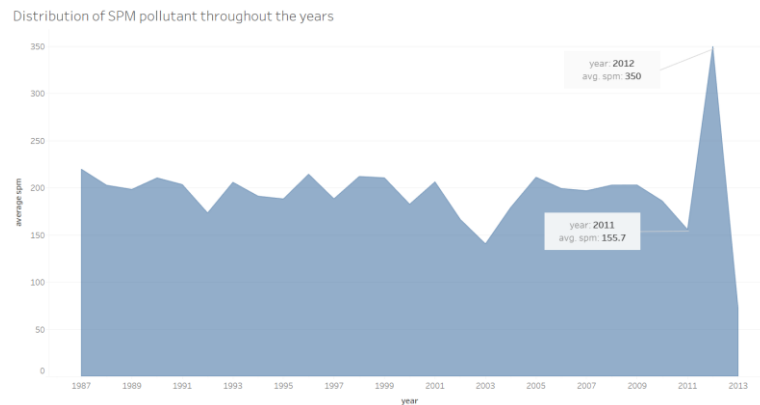


Figure 3. Distribution of SPM throughout the years

Although, figure 3 shows us that there has been an abnormal jump in the SPM readings for the year 2012. This is due to the imputation of values which were absent for this whole period. Hence, to avoid this, a mean imputation has been done only for the 2012 data for SPM pollutant based on the previous year's data.

2.1.2 Air Quality Index

It is tough to understand and visualize about the air quality using 4 different components. Air pollution is measured by the AQI which is derived using the values of all the individual pollutants. The AQI value for each record is calculated as per the Indian Government Standards. First, a sub-index for each of the pollutants is calculated based on the multiple conditions of ambient pollutant values, conforming standards and possible health influence. And, the worst sub-index resembles the overall AQI value ("National Air Quality Index," 2020). This computed column was derived and appended into the dataset in Python. The dataset can now be explored using the AQI column which showcases the levels of air quality, i.e. higher AQI value represents worse air quality.

2.2 Dates for Diwali Dataset

Next, a date dataset from the web, for India's festival of lights – 'Diwali', has been used to check its influence towards the air quality ("Diwali Date List," 2020).

The data is available at:

http://www.world-timedate.com/holidays/kali_puja_deepavali_date_list.html

- As the data is not in a format which can be used directly, it needs to be shaped based on the requirement of its merging with the primary dataset.
- The data needs to be structured into 3 columns – Year, Month and a Diwali festival check column for that particular year and month matching the actual date of the festival, with a yes or no value respectively.
- Hence, only the required data was scraped accordingly from the web using the Python's BeautifulSoup library. The data was then exported into a csv file, and then merged with the Ambient Air Quality dataset by the 'year' and 'month' columns in R.
- Post merging of these 2 datasets, the 'date' column consisted of 9 null values. As the date plays a crucial role for merging all the columns and also for answering the research questions, these 9 rows are dropped from the dataset to avoid any wrong interpretations.

2.3 Registered Vehicles in India Dataset

The third dataset, Registered Motor Vehicles in India during the years 1951 to 2013, has been used as a contributing cause towards India's increasing air pollution ("Total Number of Registered Motor Vehicles in India during 1951-2013," 2016). This data is contributed by the Ministry of Road Transport and Highways under NDSAP.

The dataset is available at:

<https://data.gov.in/resources/total-number-registered-motor-vehicles-india-during-1951-2013>

The data was wrangled after importing it in R by the following measures:

- Rename the columns appropriately and then pivot the separate columns of each of the type of vehicle into a single column, 'vehicle type' and the 'vehicle count' column is adjusted according to it.
- The dataset is filtered to be available from the year 1987. The dataset is then merged with the previously merged dataset, on the 'year' column.

2.4 State-wise Registered Vehicles in India Dataset

Another vehicle dataset, State-wise Total Registered Motor Vehicles in India from 2001 to 2011, has been used to know the counts of vehicles in the different state in India distinctly ("State-wise Total Registered Motor Vehicles In India," 2015). This separate year columns are pivoted into a single year column and the data is cleaned based on the requirements to remove the excessive rows which show total counts of the vehicles and keep only the relevant data for efficient data merging.

The dataset is available at:

<https://data.gov.in/catalog/state-wise-total-registered-motor-vehicles-india>

2.5 Shapefiles – India

Shapefiles for India have also been used to visualize the nation's air pollution story on a map ("Download data by country," 2020). The 'maptools' library in R has been used to load this data, and it is then fortified into a data frame. An additional 'id' column has been created in the above merged dataset for identifying the states and merging it with the shapefiles' data on the same column.

The shapefiles are available at:

<http://www.diva-gis.org/datadown>

3. Data Checking

3.1 Categorical Data

Let's take a closer look at the 'type' column.

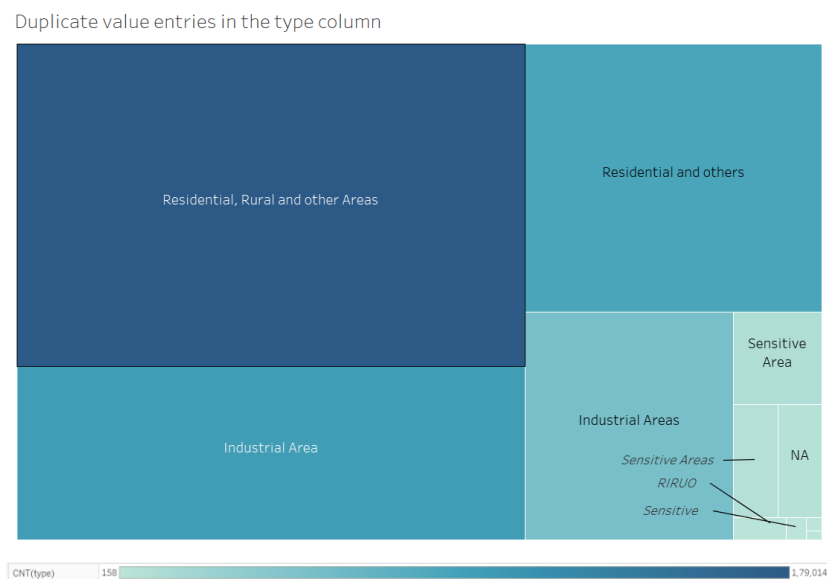


Figure 4. Duplicate values in the type column

This column consists of multiple duplicated values which actually should belong to a single category. In all, 3 such unique categories, 'Industrial Area', 'Sensitive Area' and 'Residential and others', are identified from the data. To find and assign these values to their respective categories, use of regular expression (REGEX) was done in R. Familiar patterns for a particular category were detected within the data and those values were assigned with the respective single category.

3.2 Numerical Data

3.2.1 Correlation

Let's start by plotting a Scatter Plot Matrix (SPLOM) for all the pollutants to check for any correlations and interdependencies within them.

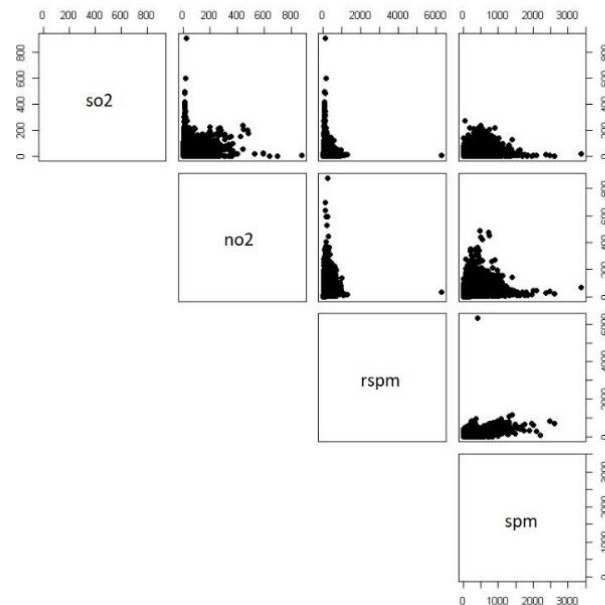


Figure 5. SPLOM for the 4 pollutants

Looking at the plot, we can clearly state that there is some correlation between the SO₂ and NO₂ pollutants. A similar trend can also be seen among NO₂ and SPM. A slight increase in the values of a pollutant can be observed when the other pollutant also shows an increase. But as the data might contain plenty of null values and outliers, we cannot get carried away with our speculations.

3.2.2 Outlier Detection

Now, let's have a look at each of the pollutants individually by revealing the outliers in the data using boxplots by excluding the null values for now.

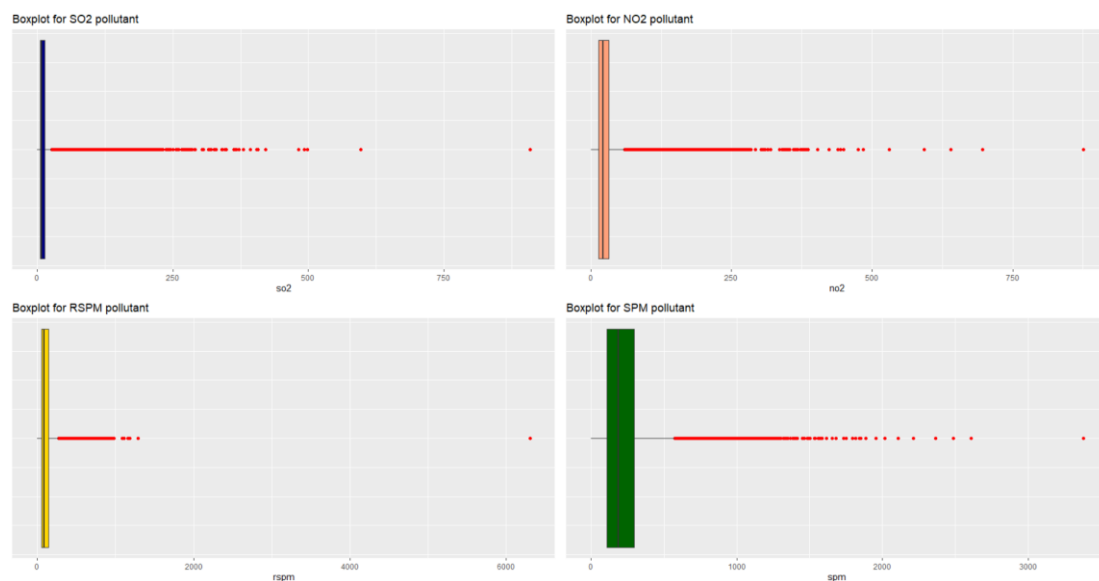


Figure 6. Boxplots for SO₂, NO₂, RSPM and SPM

And, following are the descriptive statistics for the 4 air pollutants including a count of null values.

	so2	no2	rspm	spm
Min.	: 0.00	: 0.00	: 0.0	: 0.0
1st Qu.	: 5.00	: 14.00	: 56.0	: 111.0
Median	: 8.00	: 22.00	: 90.0	: 187.0
Mean	: 10.83	: 25.81	: 108.8	: 220.8
3rd Qu.	: 13.70	: 32.20	: 142.0	: 296.0
Max.	: 909.00	: 876.00	: 6307.0	: 3380.0
NA's	: 34643	: 16230	: 40219	: 237380

Figure 7. Descriptive statistics of the air pollutants

A significant number of outliers can be seen for each of the 4 pollutants' data in the boxplots. The outliers are detected and removed using the Interquartile Rule for these variables in Python. The AQI column is also reassessed to ensure that all the data records hold an AQI value.

4. Data Exploration

4.1 Distribution of different air pollutants in India through the time period

Let's have a look at the different air pollutants and their spread in the country from 1987 through 2015.

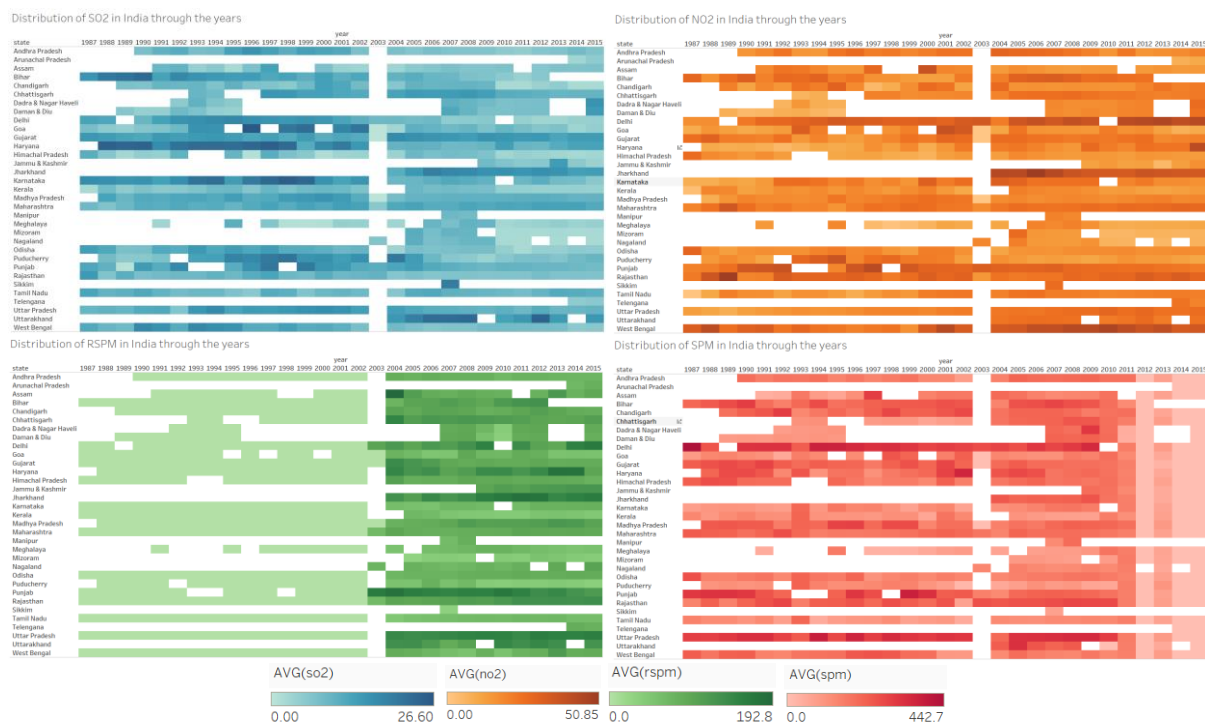


Figure 8. Distribution of pollutants in different states of India through the years

The facet grid of heatmaps shows us the overall spread of all the 4 pollutants in all the states of India. There has been no gradual increase or decrease in their levels and have been varying thoroughly through the entire period of time. Looking at the heatmaps, their spread seem more towards the Northern states. RSPM had no values recorded until the year 2002, and due to the forward fill imputation method used for the pollutants, its heatmap shows us missing readings accordingly. To preserve the skewness of the RSPM data, its values for early years were not imputed, and the information available from 2003 was held into account. Similarly, the SPM data was missing from 2012 onwards. Hence due to same imputation method used, a constant value for those years can be observed in the heatmap. The blank spaces in the heatmaps represent missing data for corresponding states or years. The scale legends for these plots also show us the range of values these pollutants lie in. These values and their conventions are completely different from each other, and hence, it's tough for us to determine the actual levels of air pollution that India is facing. Thus, the Air Quality Index (AQI) is determined and used as a common dimension for results.

4.2 The role of location for the Air Quality in India

Location plays a vital role in the distribution of the quality of air. Let's visualize this through the map of India.

Air Quality at various locations in India

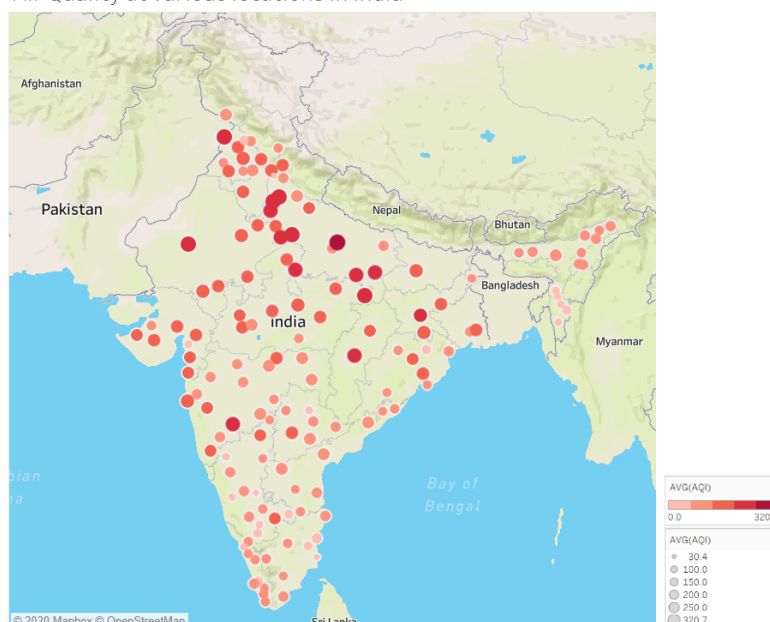


Figure 9. Air quality at various locations in India

The proportional symbol map from Tableau in figure 9 shows us the various locations throughout India, and how they are affected by air pollution considering their AQI. Looking at the map, we understand that the Northern part of India is highly affected and have the worst air quality. The Western and Eastern regions seem to be moderately affected, whereas, the North-Eastern and Southern parts of the nation are in a safe situation and have good air to breathe.

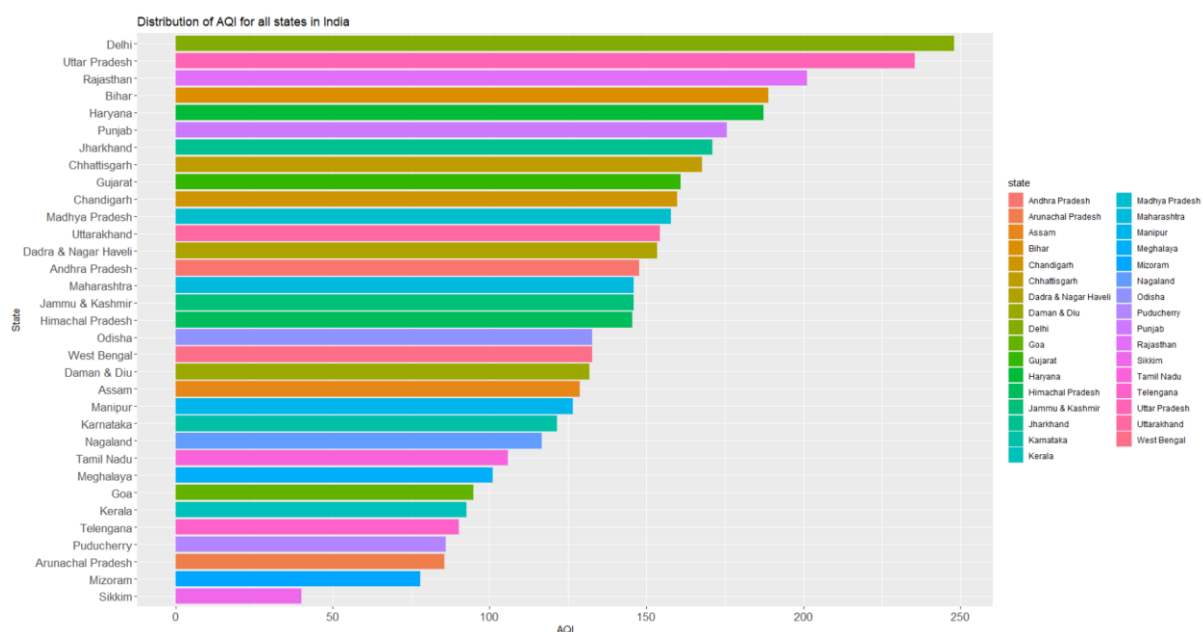


Figure 10. Distribution of AQI for all states in India

The bar graph from R for AQI in different states of India in figure 10, shows us that the Northern states, Delhi, Uttar Pradesh, Rajasthan, Bihar and Haryana are the respective top 5 leaders in terms of the worst air quality in India. Whereas, North-Eastern states, Sikkim, Mizoram and Arunachal Pradesh have the best air quality in comparison with all the states.

4.3 Different factors affecting the Air Quality in India

4.3.1 Vehicles in India

Studies suggest that 27% of the total air pollution in India is caused due to vehicles ("Want govt to build 1,600 km green wall along Aravalli, says activist," 2019). The count of vehicles registered in India has grown exponentially through the years in India.

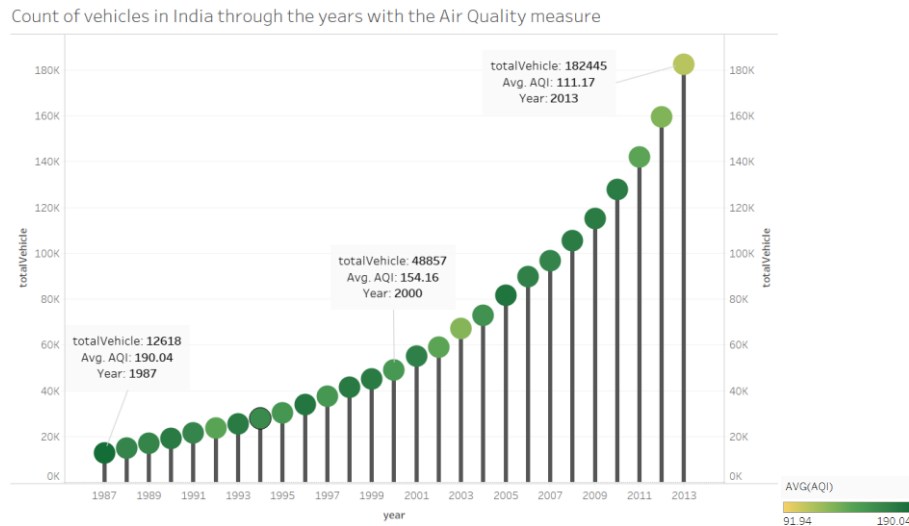


Figure 11. Distribution of vehicles and the AQI through the years

According to the lollipop chart (a version of column chart) from Tableau above, even though the vehicle count has grown at such high pace, the air pollution level has not been affected by this and has varied through all those years. Moreover, it seems like India has shown intent and have worked towards improving the air quality as the AQI levels have tapered down significantly to improve since 2011.

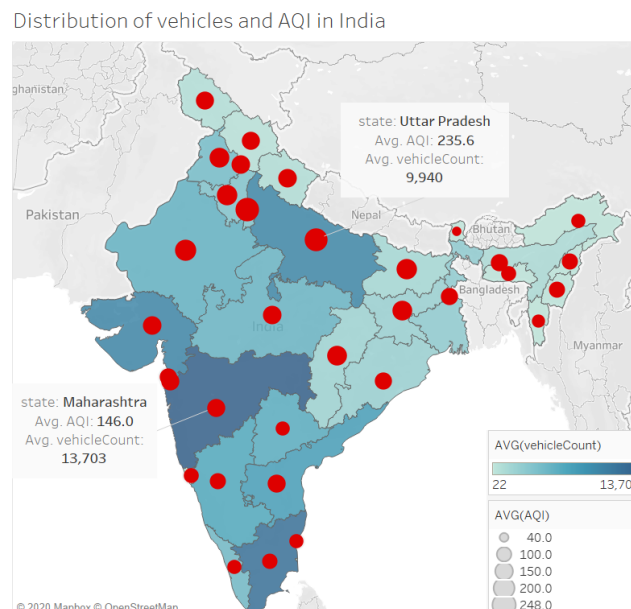


Figure 12. Distribution of vehicles and AQI in India

The choropleth and proportional symbol map from Tableau above shows us the air quality in different states in relation with the count of vehicles per state. Maharashtra, a state in the Western region, has the maximum count of registered vehicles in India. Although, its air quality still seems better than the states in North India. Comparatively, the count of vehicles is less in the Northern parts of India, except Uttar Pradesh. Uttar Pradesh shows a high vehicle count, as well as a high AQI level.

4.3.2 Type of area – Industrial, Residential and Sensitive

The type of area for which the pollution reading is recorded also plays a crucial role in determining the level of air quality. Let's have a look at it using a compound bar chart in Tableau.

Distribution of pollutants based on the different types of area

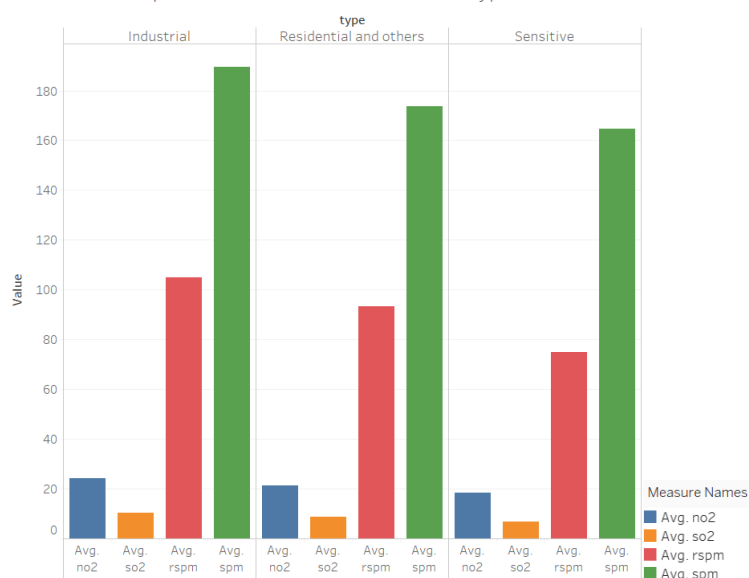


Figure 13. Effect of different types of areas on the AQI

Among the 3 types, Industrial areas are the ones with maximum levels of all the 4 pollutants, followed by Residential and others at the second spot, and Sensitive areas at the last with the least numbers. Studies suggest that 51% of the total air pollution in India is Industrial pollution ("Want govt to build 1,600 km green wall along Aravalli, says activist," 2019).

4.3.3 Diwali Festival Fireworks

In India, Diwali – festival of lights, is among the major festivals and celebrations for all people. People burst many fireworks and firecrackers all over the country during this 5-day long festival. Diwali is also considered as a contributor towards the air pollution in India, but with only 5% ("Want govt to build 1,600 km green wall along Aravalli, says activist," 2019). The comparative choropleth map in R and the bar chart in Tableau below show us that the AQI level in India is increased during Diwali, but by a very slim margin.

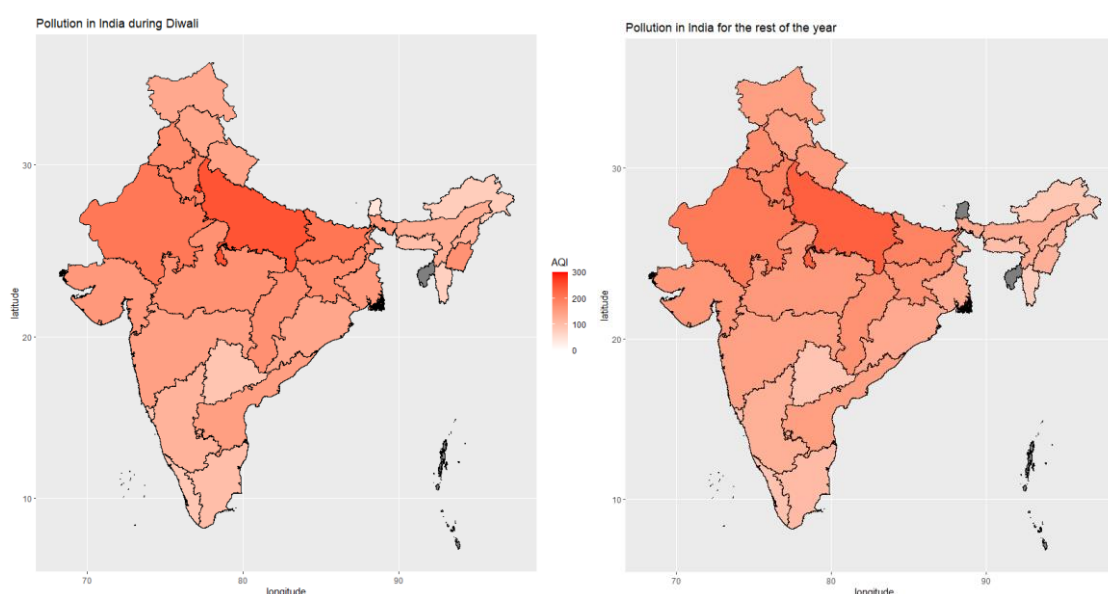


Figure 14. Pollution in India during Diwali and during the rest of the year

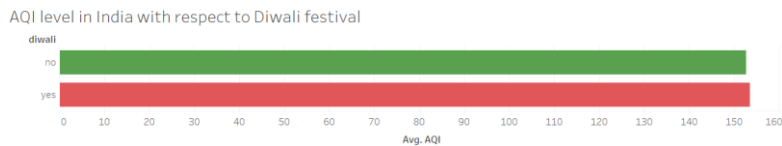


Figure 15. Level of AQI with respect to Diwali

5. Conclusion

The Historical Daily Ambient Air Quality data for India tells us a variety of stories after merging it with the dependent datasets. The 4 major pollutant readings tell us a composition summary of the air for a specific region and location. Although, as the 4 different variables were not leading to combined conclusive results, an Air Quality Index (AQI) dimension was computed and used as a common ground for the visualisations. Based on the outcomes, the Northern states of India have been the most polluted for all these years, whereas, the North-Eastern and Southern states have taken good care to keep the pollution levels down. High amounts of crop burning due to late withdrawal of monsoon every year in the Northern parts of India, can be a possible reason for its high pollution. On the contrary, the less air pollution in the North-Eastern and Southern regions can be a result of their high-altitude locations. Other factors like vehicles, industries and the Diwali festival also play important roles towards the increasing air pollution in India. Increasing number of vehicles has been a concern for India due to the various resulting problems that have grown, with air pollution being one of them. Industries produce a lot of smoke and harmful gas emissions. These gases are then accounted as air pollutants which cause the increase in the levels of dangerous air. Similarly, firecrackers burst during Diwali is also a known cause of pollution in the country. But, looking at the results, Diwali festival has not created a significant difference in the stages of air quality in India. Year data has not been consistent for the AQI, although, the pollution level has seemed to go down since 2011. The lowered pollution values for the recent years can also be a probable result of the new initiatives and awareness campaigns among the citizens of India from the Government.

6. Reflection

The data exploration project for India's Ambient Air Quality data is an attempt to understand the behaviour of the different air pollutants found in the atmosphere of India, with the various factors affecting it. All the datasets are first cleaned using various approaches and techniques, to check all columns for visible problems and to reduce the unnecessary dimensions, to merge them seamlessly, and to finally create a dataset ready for exploration. While dealing with numerous datasets, I learnt that cleaning the data into appropriate forms and formats is very important as the data needs to be free of jargon for accurate results after merging processes. After cleaning the data, it is visualized and checked for anomalies for both, categorical and numerical kinds. While plotting the data, I learnt that checking the data for errors post cleaning it, is crucial as we understand the data even more deeply. Use of statistical approaches is the key here for treating the errors logically. Data checking can also be suggested as a prior version of data exploration, where we aim to discover data issues which cannot be found by just looking at the data. The air quality results are dependent on 4 major pollutant dimensions. Hence, an AQI value is derived using a logic as per the Indian Government Standards, for deducing uniform results. This is one of the most important steps as the existence of different pollutants in the air varies depending on the region, type of area, and many other factors in any country.

Once the final version of the dataset is ready, it is explored using multiple statistical tests and visualizations, aiming to search for answers for the research questions, along with any new findings on the way. The visualizations intend to display and depict the optimum information with precise explanations. Learning these plots and tests and their correct use for the required outcomes from the data, was a significant take away for me. These are important to derive truthful conclusions from the data. Natural factors like location, altitude and weather affect a region's air pollution. Time is also a crucial factor which affects it. Although, the data does not show a consistent behaviour for years, but it does show for the different seasons or months. A seasonal factor can be considered to understand the air quality behaviour, as June to September is a rainy season in India, which might lower the AQI levels for that time period. Another known cause of air pollution in India, is crop burning. I would have loved to include another dataset comprising the crop burning statistics and research the findings. But due to insufficient open data available for this segment, it was not possible to relate the air quality with it.

7. Bibliography

References

- Anand, G. (2017). India's Air Pollution Rivals China's as World's Deadliest. *The New York Times*. Retrieved from https://www.nytimes.com/2017/02/14/world/asia/indias-air-pollution-rivals-china-as-worlds-deadliest.html?_r=0
- Diwali Date List. (2020). Retrieved from http://www.world-timedate.com/holidays/kali_puja_deepavali_date_list.php
- Download data by country. (2020). Retrieved from <http://www.diva-gis.org/gdata>
- Historical Daily Ambient Air Quality Data. (2017). Retrieved from <https://data.gov.in/catalog/historical-daily-ambient-air-quality-data>
- National Air Quality Index. (2020). Retrieved from <https://cpcb.nic.in/National-Air-Quality-Index/>; [https://cpcb.nic.in/displaypdf.php?id=bmF0aW9uYWwtYWlyLXF1YWxpdHktaW5kZXgvSG93X0FRSV9DYWxjdWxhdGVkLnBkZg](https://cpcb.nic.in/displaypdf.php?id=bmF0aW9uYWwtYWlyLXF1YWxpdHktaW5kZXgvSG93X0FRSV9DYWxjdWxhdGVkLnBkZg;); https://app.cpcbccr.com/AQI_India/
- Osseiran, N., & Lindmeier, C. (2018). 9 out of 10 people worldwide breathe polluted air, but more countries are taking action [Press release]. Retrieved from <https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action>
- State-wise Total Registered Motor Vehicles In India. (2015). Retrieved from <https://data.gov.in/catalog/state-wise-total-registered-motor-vehicles-india>
- Total Number of Registered Motor Vehicles in India during 1951-2013. (2016). Retrieved from <https://data.gov.in/resources/total-number-registered-motor-vehicles-india-during-1951-2013>
- Want govt to build 1,600 km green wall along Aravalli, says activist. (2019). *The Indian Express*. Retrieved from <https://indianexpress.com/article/cities/ahmedabad/want-govt-to-build-1600-km-green-wall-along-aravalli-says-activist-vijaypal-baghel-6182069/>

Appendix

1. Ambient Air Quality for India Dataset:

stn_code	sampling_date	state	location	agency	type	so2
1	150 February - M021990	Andhra Pradesh	Hyderabad	<NA>	Residential, Rural and other Areas	4.8
2	151 February - M021990	Andhra Pradesh	Hyderabad	<NA>	Industrial Area	3.1
3	152 February - M021990	Andhra Pradesh	Hyderabad	<NA>	Residential, Rural and other Areas	6.2
4	150 March - M031990	Andhra Pradesh	Hyderabad	<NA>	Residential, Rural and other Areas	6.3
5	151 March - M031990	Andhra Pradesh	Hyderabad	<NA>	Industrial Area	4.7
6	152 March - M031990	Andhra Pradesh	Hyderabad	<NA>	Residential, Rural and other Areas	6.4

no2	rspm	spm	location_monitoring_station	pm2_5	date
1	17.4	NA	NA	<NA>	NA 1990-02-01
2	7.0	NA	NA	<NA>	NA 1990-02-01
3	28.5	NA	NA	<NA>	NA 1990-02-01
4	14.7	NA	NA	<NA>	NA 1990-03-01
5	7.5	NA	NA	<NA>	NA 1990-03-01
6	25.7	NA	NA	<NA>	NA 1990-03-01

2. Dates for Diwali Dataset:

KALI PUJA / DEEPAVALI / DIWALI DATE LIST

from

1980 to 1999 :

Year :	Date :	Weekday :	Tithi :
1980	November 07, 1980	Friday	(Amavasya - Krishna Paksha)
1981	October 27, 1981	Tuesday	(Amavasya - Krishna Paksha)
1982	November 15, 1982	Monday	(Amavasya - Krishna Paksha)
1983	November 04, 1983	Friday	(Amavasya - Krishna Paksha)
1984	October 24, 1984	Wednesday	(Amavasya - Krishna Paksha)
1985	November 12, 1985	Tuesday	(Amavasya - Krishna Paksha)
1986	November 02, 1986	Sunday	(Amavasya - Krishna Paksha)
1987	October 22, 1987	Thursday	(Amavasya - Krishna Paksha)
1988	November 09, 1988	Wednesday	(Amavasya - Krishna Paksha)
1989	October 29, 1989	Sunday	(Amavasya - Krishna Paksha)

3. Web-scraped Dates for Diwali Dataset:

	year	month	diwali
1	1987	1	no
2	1987	2	no
3	1987	3	no
4	1987	4	no
5	1987	5	no
6	1987	6	no

4. Registered Vehicles in India Dataset:

	year	All.Vehicles	Two.Wheelers	Cars..Jeeps.and.Taxis	Buses	Goods.Vehicles	Others
1	1951	306	27	159	34	82	4
2	1956	426	41	203	47	119	16
3	1959	562	67	267	48	148	32
4	1960	605	76	282	54	157	36
5	1961	665	88	310	57	168	42
6	1962	749	116	340	60	189	44

5. State-wise Registered Vehicles in India Dataset:

	State.Union.Territory	X2001	X2002	X2003	X2004	X2005	X2006	X2007	X2008	X2009	X2010	X2011
1	Andhra Pradesh	3966	4389	5002	5720	6458	7218	6367	7208	8059	8923	10189
2	Arunachal Pradesh	21	21	21	21	22	22	22	22	22	22	145
3	Assam	542	596	657	727	815	914	1021	1116	1235	1384	1582
4	Bihar	949	1024	1121	751	1352	1432	1577	1739	1960	2357	2673
5	Chhatisgarh	857	948	1076	1216	1375	1541	1734	1935	2115	2436	2766
6	Goa	341	366	397	436	482	529	579	624	674	727	790

6. Final Merged Dataset:

year	state	month	day	date	location	agency	type	station	so2	si	no2	ni	rspm	rpi	
1	1987	Goa	1	8	08-01-1987	Vasco	Goa Pollution Control Board	Industrial	Unavailable	5.2	6.500	11.1	13.875	NA	0
2	1987	Goa	1	8	08-01-1987	Vasco	Goa Pollution Control Board	Industrial	Unavailable	5.2	6.500	11.1	13.875	NA	0
3	1987	Goa	1	8	08-01-1987	Vasco	Goa Pollution Control Board	Industrial	Unavailable	5.2	6.500	11.1	13.875	NA	0
4	1987	Goa	1	8	08-01-1987	Vasco	Goa Pollution Control Board	Industrial	Unavailable	5.2	6.500	11.1	13.875	NA	0
5	1987	Goa	1	8	08-01-1987	Vasco	Goa Pollution Control Board	Industrial	Unavailable	5.2	6.500	11.1	13.875	NA	0
6	1987	Bihar	4	12	12-04-1987	Sindri	Bihar Pollution Control Board	Industrial	Unavailable	16.3	20.375	16.5	20.625	NA	0
spm	spi	AQI	diwali	id	vehicleCountPerState	totalVehicle	vehicleType	vehicleCountPerType							
1	84	-32.5	13.875	no	11	NA	12618	Buses	245						
2	84	-32.5	13.875	no	11	NA	12618	Cars, Jeeps and Taxis	2007						
3	84	-32.5	13.875	no	11	NA	12618	Goods Vehicles	984						
4	84	-32.5	13.875	no	11	NA	12618	Others	1643						
5	84	-32.5	13.875	no	11	NA	12618	Two wheelers	7739						
6	341	291.0	291.000	no	5	NA	12618	Buses	245						

7. Code for Outlier Detection and Removal in Python:

```
Q1 = df['so2'].quantile(0.25)
Q3 = df['so2'].quantile(0.75)
IQR = Q3 - Q1
df = df[~((df['so2'] < (Q1 - 1.5 * IQR)) | (df['so2'] > (Q3 + 1.5 * IQR)))]

Q1 = df['no2'].quantile(0.25)
Q3 = df['no2'].quantile(0.75)
IQR = Q3 - Q1
df = df[~((df['no2'] < (Q1 - 1.5 * IQR)) | (df['no2'] > (Q3 + 1.5 * IQR)))]

Q1 = df['rspm'].quantile(0.25)
Q3 = df['rspm'].quantile(0.75)
IQR = Q3 - Q1
df = df[~((df['rspm'] < (Q1 - 1.5 * IQR)) | (df['rspm'] > (Q3 + 1.5 * IQR)))]

Q1=df['spm'].quantile(0.25)
Q3=df['spm'].quantile(0.75)
IQR = Q3 - Q1
df = df[~((df['spm'] < (Q1 - 1.5 * IQR)) | (df['spm'] > (Q3 + 1.5 * IQR)))]
```


8. Logic and code for calculating the Air Quality Index (AQI) in Python:

```
def calculate_si(so2):
    si = 0
    if (so2 <= 40):
        si = so2 * (50 / 40)
    elif (so2 > 40 and so2 <= 80):
        si = 50 + (so2 - 40) * (50 / 40)
    elif (so2 > 80 and so2 <= 380):
        si = 100 + (so2 - 80) * (100 / 300)
    elif (so2 > 380 and so2 <= 800):
        si = 200 + (so2 - 380) * (100 / 420)
    elif (so2 > 800 and so2 <= 1600):
        si = 300 + (so2 - 800) * (100 / 800)
    else (so2 > 1600):
        si = 400 + (so2 - 1600) * (100 / 800)
    return si
data['si'] = data['so2'].apply(calculate_si)
```

```
def calculate_ni(no2):
    ni = 0
    if(no2 <= 40):
        ni = no2 * 50 / 40
    elif(no2 > 40 and no2 <= 80):
        ni = 50 + (no2 - 14) * (50 / 40)
    elif(no2 > 80 and no2 <= 180):
        ni = 100 + (no2 - 80) * (100 / 100)
    elif(no2 > 180 and no2 <= 280):
        ni = 200 + (no2 - 180) * (100 / 100)
    elif(no2 > 280 and no2 <= 400):
        ni = 300 + (no2 - 280) * (100 / 120)
    else:
        ni = 400 + (no2 - 400) * (100 / 120)
    return ni
data['ni'] = data['no2'].apply(calculate_ni)
```

```
def calculate_rpi(rspm):
    rpi = 0
    if(rspm <= 30):
        rpi = rspm * 50/30
    elif(rspm > 30 and rspm <= 60):
        rpi = 50 + (rspm - 30) * 50 / 30
    elif(rspm > 60 and rspm <= 90):
        rpi = 100 + (rspm - 60) * 100 / 30
    elif(rspm > 90 and rspm <= 120):
        rpi = 200 + (rspm - 90) * 100/30
    elif(rspm > 120 and rspm <= 250):
        rpi = 300 + (rspm - 120) * (100 / 130)
    else:
        rpi = 400 + (rspm - 250) * (100 / 130)
    return rpi
data['rpi'] = data['rspm'].apply(calculate_rpi)
```

```
def calculate_spi(spm):
    spi = 0
    if(spm <= 50):
        spi = spm
    if(spm < 50 and spm <= 100):
        spi = spm
    elif(spm > 100 and spm <= 250):
        spi = 100 + (spm - 100) * (100 / 150)
    elif(spm > 250 and spm <= 350):
        spi = 200 + (spm - 250)
    elif(spm > 350 and spm <= 450):
        spi = 300 + (spm - 350) * (100 / 80)
    else:
        spi = 400 + (spm - 430) * (100 / 80)
    return spi
data['spi'] = data['spm'].apply(calculate_spi)
```

```
def calculate_aqi(si, ni, spi, rpi):
    aqi = 0
    if(si > ni and si > spi and si > rpi):
        aqi = si
    if(spi > si and spi > ni and spi > rpi):
        aqi = spi
    if(ni > si and ni > spi and ni > rpi):
        aqi = ni
    if(rpi > si and rpi > ni and rpi > spi):
        aqi = rpi
    return aqi
data['AQI'] = data.apply(lambda x: calculate_aqi(x['si'], x['ni'], x['spi'], x['rpi']), axis=1)
```