# Air Quality in São Paulo

## Dataset:

The dataset contains hourly pollutant measurements recorded by CETESB at various air quality monitoring stations across São Paulo, Brazil, covering the period from August 5, 2013, to September 9, 2020.

Available in: https://www.kaggle.com/datasets/amandalk/sp-air-quality

## Objective:

The general objective of this project is to analyze the variation in nitrogen dioxide ($NO_2$) levels, measured in micrograms per cubic meter ($\mu g/m^3$), in São Paulo throughout the day. Specific goals include identifying the main causes of $NO_2$ concentration, analyzing its behavioral patterns, and studying regional differences across the city.

## Language and Libraries:

The project was developed using Python. Data visualization employed Matplotlib and Seaborn, known for their flexibility and ability to create informative and visually appealing charts. Pandas was utilized for handling tabular data due to its efficiency and versatility. Lastly, Datetime provided robust tools for manipulating and formatting temporal data.

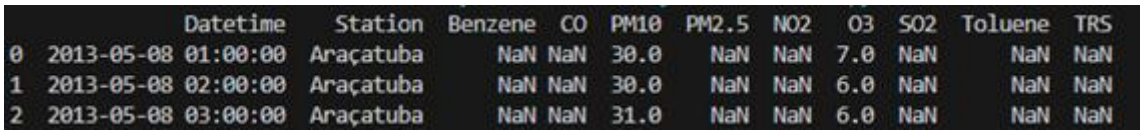
## Exploratory Analysis:

The dataset consists of 11 columns grouped into three main categories: collection time, city, and pollutants. The collection time is represented by the Datetime column in ISO 8601 format ('YYYY-MM-DD HH:MM:SS'), recording the date and time of each measurement. The Station column identifies 59 different monitoring stations. The remaining columns include pollutants: ‘Benzene’, ‘CO’, ‘PM10’, ‘PM2.5’, ‘NO2’, ‘O3’, ‘SO2’, ‘Toluene’, and ‘TRS’, measured in micrograms per cubic meter ($\mu g/m^3$).

Moreover, the presence of missing values (NaN) was identified in some pollutant columns, including the ‘NO2’ column. To minimize the impact of these

missing data on the analyses, the possibility of excluding incomplete records was evaluated. However, since the missing values represented more than 5% of the total records, it was decided to impute the missing data. The imputation was performed by replacing the missing ‘NO2’ values with the median of this variable for each monitoring station identified in the ‘Station’ column. The choice of the median as the imputation method for the missing values in the NO2 column was based on its robustness against extreme values (outliers). After this step, outliers were identified considering the subgroups in the ‘Station’ column, using the interquartile range (IQR) for each collection station. The lower limit (Q1 – 1.5 x IQR) and upper limit (Q3 + 1.5 x IQR) were calculated separately for each group, ensuring a more precise identification of extreme values, which were subsequently removed from the analysis.

**Figure - print(air_quality.head(3))**



| | Datetime | Station | Benzene | CO | PM10 | PM2.5 | NO2 | O3 | SO2 | Toluene | TRS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2013-05-08 01:00:00 | Araçatuba | NaN | NaN | 30.0 | NaN | NaN | 7.0 | NaN | NaN | NaN |
| 1 | 2013-05-08 02:00:00 | Araçatuba | NaN | NaN | 30.0 | NaN | NaN | 6.0 | NaN | NaN | NaN |
| 2 | 2013-05-08 03:00:00 | Araçatuba | NaN | NaN | 31.0 | NaN | NaN | 6.0 | NaN | NaN | NaN |

Fonte: Air Quality in São Paulo.py

## Data Manipulation:

After the exploratory data analysis, the manipulation stage was carried out using the Pandas library for handling tabular data, Datetime for working with temporal information, and Matplotlib/Seaborn for data visualization. First, a new column called "Hour" was created, from which the collection times recorded in the Datetime column were extracted. This transformation was performed using the functionalities of the Datetime module, allowing efficient conversion and access to the time elements.

Next, an additional table was created from two dictionaries. The first one was constructed from the unique values of the collection stations found in the "Station" column of the original DataFrame. The second dictionary categorized these stations based on their geographical locations, assigning them to five regions of the São Paulo state: Capital, Metropolitan Area, Interior, Vale do Paraíba, and Coast.

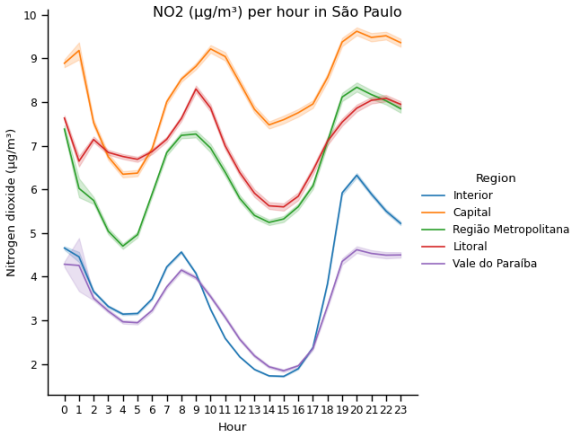
Subsequently, using the join functionalities of the Pandas library, a left "merge" operation was performed between the original DataFrame and the new table. This allowed the addition of the "Region" column, with the respective regional

classifications, to the main DataFrame. This procedure enriched the data by incorporating useful geographical information for later analyses.

Finally, the Matplotlib and Seaborn libraries were employed for visualizing the manipulated DataFrame. A line relational plot was created, where the "Hour" column was represented on the horizontal axis, and the values from the "NO2" column were plotted on the vertical axis. The plot was segmented by region, allowing the analysis of nitrogen dioxide concentration trends in different locations of São Paulo state over time.

## Data Analysis:

**Graphic - NO2 (µg/m³) per hour in São Paulo**




Fonte: Air Quality in São Paulo.py

The data analysis reveals a day-night pattern in the concentration of nitrogen dioxide ($NO_2$) in the atmosphere of São Paulo. During the morning period, a significant peak in concentration is observed between 6 AM and 10 AM, coinciding with the increased vehicle traffic typical of the morning rush hour. The second peak occurs between 5 PM and 8 PM, again associated with heavy traffic at the end of the workday. On the other hand, the periods of lower $NO_2$ concentration occur during the early morning, between 1 AM and 5 AM, and in the early afternoon, between 11 AM and 3 PM, reflecting reduced vehicle activity at these times and greater pollutant dispersion due to higher temperatures and solar radiation.

Regarding regional differences, the Metropolitan Region, Capital, and Coast areas exhibit the highest $NO_2$ concentrations throughout the day. This phenomenon can be attributed to higher population density, intense traffic, and industrial activities in these regions. In contrast, the Vale do Paraíba and Interior regions record the lowest $NO_2$ concentrations, as they have lower population densities and less intense industrial

activities. Therefore, the data supports the variations in the pollutant's behavior at different times of the day and in different regions, fulfilling the objectives of understanding the causes, patterns, and specificities of $NO_2$ concentrations.

## Conclusion:

The results indicate that vehicular traffic is the primary contributor to nitrogen dioxide emissions in São Paulo, directly influencing concentration peaks. Furthermore, regional differences highlight the need for local policies aimed at mitigating the pollutant's impacts, especially in the most affected areas such as the Metropolitan Region, Capital, and Coast. In this context, adopting less polluting fuels is essential for reducing emissions. Alternatives like electric vehicles, which do not emit pollutants directly, and promoting the use of Natural Gas, which emits less $NO_2$, can significantly contribute to improving air quality. Finally, encouraging sustainable public transportation, promoting biofuels, and implementing strict industrial emission controls are crucial for achieving substantial reductions in air pollution and improving public health.