## Technical Review

We have recently celebrated 50th anniversary of Earth Day and the theme for Earth Day 2020 is climate action. Climate change and air pollution are two of the significant environmental problems we are facing today and are closely interlinked. Emissions of pollutants into the air contributes to the changes to the climate by affecting the amount of incoming sunlight that is reflected or absorbed by the atmosphere, with some pollutants warming and others cooling the Earth.

This inspired us to choose the theme for our team’s ETL project as **Climate Action – Reduce Air Pollution** and we wanted to further explore the various air pollutants and their distribution.

***Data sources:***

As a first step, to identify the data that may be relevant to the problem area, we had to search and identify usable data sources and we chose the public data sets listed below.

* Kaggle - US Pollution Data - <https://www.kaggle.com/sogun3/uspollution> - (2000-2016) CSV

The original data for this data set comes primarily from the well documented and high quality U.S. EPA AQS (Air Quality System) database: <https://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html>.

* OECD - Air Emissions by Source - <https://stats.oecd.org/index.aspx?lang=en> - (2000-2016) CSV

This dataset provides selected information on national emissions of traditional air pollutants. They concern man-made emissions of sulphur oxides (SOx), nitrogen oxides (NOx), particulate matter (PM), carbon monoxide (CO) and volatile organic compounds (VOC).

***Data Transformation – Cleanup & Analysis:***

The United States Environmental Protection Agency (EPA) defines air emissions which are or could be harmful to people as criteria pollutants. These six criteria pollutants are carbon monoxide (CO), lead (Pb), nitrogen dioxide (NO2), ozone (O3), particulate matter (PM), and sulfur dioxide (SO2).

(Source: <https://www.epa.gov/criteria-air-pollutants>)

There are many sources of emissions and can be grouped into four major categories: point, mobile, biogenic, and area.

* Point sources – Factories, electric power plants etc.
* Mobile sources - Cars, trucks lawn mowers, airplanes etc.

(Source: <https://www3.epa.gov/airquality/emissns.html>)

To transform the public data and use it in our study we performed the following:

* Used Pandas functions in Jupyter Notebook to load all source CSV files.
* Reviewed the files and transformed into data frames
* Removed information which was not relevant to the focus of this study.
* Generated new csv files (6 in total) representing the cleaned data

sets.

Even though air pollution is prevalent globally we have decided to limit our exploration to United states geographically and to the three major pollutants (Nitrogen Dioxide, Sulphur Dioxide, and Carbon Monoxide).

A more detailed description of data cleaning is described below per dataset

* ***Air Emissions by Source Dataset:***
  + Removed the columns for Country, COU, Pollutant, YEA, Unit Code, PowerCode Code, PowerCode, Reference Period Code, Reference Period, Flag Codes, and Flags.
  + Dropped rows that had data for other pollutants and extracted only the data for pollutants Nitrogen Dioxide, Sulphur Dioxide, and Carbon Monoxide.
  + Renamed the pollutants so that they matched our other dataset. We also changed the name for "VAR" and "Variable" columns to "pol\_var" and "pol\_variable".
  + Separated the data into three csv files, based on pollutant.
* ***US Pollution Dataset:***
  + Removed State Code, County Code, Site Num, Address, County, O3 Units, O3 Mean, O3 1st Max Value, O3 1st Max Hour, O3 AQI, NO2 1st Max Value, NO2 1st Max Hour, NO2 AQI,SO2 1st Max Value, SO2 1st Max Hour, SO2 AQI, CO 1st Max Value, CO 1st Max Hour, CO AQI
  + Aggregated the Mean columns for each pollutant and separate by year (columns) and pollutants (rows)
  + Converted the values column from Parts per Billion to Gram/Ton
  + Segregated the data into three csv files based on distribution of pollutants across the various states and cities and mean values of the pollutants.

We chose PostgreSQL to be the type of final production database to load the newly generated CSV files from the data frames (representing the cleaned data sets) and created a database named air\_pollution\_db.

Using the quick database website, we generated a database schema to create the tables needed to store the csv files. We then did an initial connection to the Postgres database using PG admin and generated a set of six tables in the production database as listed below:

* mod\_pollution\_year\_city
* mod\_pollution\_year\_state
* mod\_cleaned\_air\_emission\_by\_source\_co
* mod\_cleaned\_air\_emission\_by\_source\_no2
* mod\_cleaned\_air\_emission\_by\_source\_so2
* mod\_mean\_us\_pollution

We then populated the tables with the CSV files generated from the dataframes in jupyter notebook and designed queries, to determine the distribution of air pollutant levels geographically in united states over time.

***Query 1: Most Polluted Cities:***

-- ***Cities with the highest NO2 Pollution levels in the last 5 years:***

SELECT years, city, no2\_mean FROM mod\_pollution\_year\_city

WHERE years BETWEEN '2012' AND '2016'

ORDER BY no2\_mean DESC;

-- ***Cities with the highest CO Pollution levels in the last 5 years:***

SELECT years, city, co\_mean FROM mod\_pollution\_year\_city

WHERE years BETWEEN '2012' AND '2016'

ORDER BY co\_mean DESC;

***-- Cities with the highest SO2 Pollution levels in the last 5 years:***

SELECT years, city, so2\_mean FROM mod\_pollution\_year\_city

WHERE years BETWEEN '2012' AND '2016'

ORDER BY so2\_mean DESC;

***Query 2: Least Polluted Cities:***

***-- Cities with the least NO2 Pollution levels in the last 5 years:***

SELECT years, city, no2\_mean FROM mod\_pollution\_year\_city

WHERE years BETWEEN '2012' AND '2016'

ORDER BY no2\_mean ASC;

***-- Cities with the least CO Pollution levels in the last 5 years:***

SELECT years, city, co\_mean FROM mod\_pollution\_year\_city

WHERE years BETWEEN '2012' AND '2016'

ORDER BY co\_mean ASC;

***-- Cities with the least SO2 Pollution levels in the last 5 years:***

SELECT years, city, so2\_mean FROM mod\_pollution\_year\_city

WHERE years BETWEEN '2012' AND '2016'

ORDER BY so2\_mean ASC;

***Query 3: Most Polluted States:***

***-- States with the highest NO2 Pollution levels in the last 5 years:***

SELECT years, state, no2\_mean FROM mod\_pollution\_year\_state

WHERE years BETWEEN '2012' AND '2016'

ORDER BY no2\_mean DESC;

***-- States with the highest CO Pollution levels in the last 5 years:***

SELECT years, state, co\_mean FROM mod\_pollution\_year\_state

WHERE years BETWEEN '2012' AND '2016'

ORDER BY co\_mean DESC;

***-- States with the highest SO2 Pollution levels in the last 5 years:***

SELECT years, state, so2\_mean FROM mod\_pollution\_year\_state

WHERE years BETWEEN '2012' AND '2016'

ORDER BY so2\_mean DESC;

***Query 4: Least Polluted States:***

***-- States with the least NO2 Pollution levels in the last 5 years:***

SELECT years, state, no2\_mean FROM mod\_pollution\_year\_state

WHERE years BETWEEN '2012' AND '2016'

ORDER BY no2\_mean ASC;

***-- States with the least CO Pollution levels in the last 5 years:***

SELECT years, state, co\_mean FROM mod\_pollution\_year\_state

WHERE years BETWEEN '2012' AND '2016'

ORDER BY co\_mean ASC;

***-- States with the least SO2 Pollution levels in the last 5 years:***

SELECT years, state, so2\_mean FROM mod\_pollution\_year\_state

WHERE years BETWEEN '2012' AND '2016'

ORDER BY so2\_mean ASC;

***Query 5: Highest Pollution Level in all years:***

***-- Years of the highest level of NO2:***

SELECT years, no2\_mean FROM mod\_mean\_us\_pollution

ORDER BY no2\_mean DESC LIMIT 5;

***-- Years of the highest level of SO2:***

SELECT years, so2\_mean FROM mod\_mean\_us\_pollution

ORDER BY so2\_mean DESC LIMIT 5;

***-- Years of the highest level of CO:***

SELECT years, co\_mean FROM mod\_mean\_us\_pollution

ORDER BY co\_mean DESC LIMIT 5;

***Query 6: Lowest Pollution Level in all years:***

***-- Years of the lowest level of NO2:***

SELECT years, no2\_mean FROM mod\_mean\_us\_pollution

ORDER BY no2\_mean ASC LIMIT 5;

***-- Years of the lowest level of SO2:***

SELECT years, so2\_mean FROM mod\_mean\_us\_pollution

ORDER BY so2\_mean ASC LIMIT 5;

***-- Years of the lowest level of CO:***

SELECT years, co\_mean FROM mod\_mean\_us\_pollution

ORDER BY co\_mean ASC LIMIT 5;

**Query 7: Pollution level by pol\_variable for all Pollutants (CO, NO2,SO2):**

***-- Pollutant - NO2:***

SELECT pol, pol\_var, pol\_variable, sum(value)

FROM mod\_cleaned\_air\_emission\_by\_source\_no2

GROUP BY pol, pol\_variable, pol\_var

ORDER BY sum(value) DESC

***-- Pollutant - SO2:***

SELECT pol, pol\_var, pol\_variable, sum(value)

FROM mod\_cleaned\_air\_emission\_by\_source\_so2

GROUP BY pol, pol\_variable, pol\_var

ORDER BY sum(value) DESC

***-- Pollutant – CO:***

SELECT pol, pol\_var, pol\_variable, sum(value)

FROM mod\_cleaned\_air\_emission\_by\_source\_co

GROUP BY pol, pol\_variable, pol\_var

ORDER BY sum(value) DESC

**Summary:**

The datasets we used have data listed only from years 2000-2016. Some of the findings from our study are listed below:

* Most Polluted Cities were Salt lake City (NO2), Phoenix (CO), Arden-Arcade (SO2) for the year 2016.
* Least Polluted Cities were Capitan (NO2), York (CO), Houston (SO2) for the year 2016.
* Most Polluted States were Utah and Colorado(NO2), Florida and Utah (CO), Ohio (SO2) for the year 2016.
* Least Polluted States: Tennessee (NO2), Wyoming (CO), Wyoming (SO2)
* Highest Pollution Level in last all years: Year 2001 for NO2, Year 2000 for SO2, and Year 2000 for CO
* Lowest Pollution Level in last all years: Year 2015 for NO2, Year 2016 for SO2 and Year 2014 for CO
* Pollution level categorized by emission sources for all Pollutants (CO, NO2,SO2): Man-made Emissions was found to be the source with the highest pollutant level for all 3 pollutants followed by Mobile Sources and Road Transport.

Based on the above query results, it would seem that air pollution level seems to be on downward trend for the pollutants (CO, No2 and SO2) in the United States. However, the datasets we have examined did not have information on the other major air pollutants such as greenhouse gases, toxic air pollutants, Ozone and data from more recent years would also be needed in reaching such a conclusion. It would also help to have some more data sets describing the distribution of major indoor and outdoor air pollutants.

Most Polluted