Friedrich Alexander University, Erlangen Nuremberg
MADE - Data Report - SS24

Nipun Arora – 23084364

1. **Main Question** - Does temperature changes with the increare/decrease in pollutants in air?

2. **Data Sources:**

- Data has been taken from Kaggle and a custom pipeline is created to fetch the data from the source.
- Two datasets have been used that hides the answer to our question
- You can import the dataset by running 'pipeline.sh' in the project folder.
- Links to the dataset are given below:-
  - Dataset 1: https://www.kaggle.com/datasets/vanvalkenberg/historicalweatherdataforindiancities
  - Dataset 2: https://www.kaggle.com/datasets/abhisheksjha/time-series-air-quality-data-of-india-2010-2023
- The data is in format of csv's that are collected from multiple cities in India.

3. **Data Licences**

- The dataset 1 holds a CC0: Public Domain licence that has no copyrights it is available in public domain.
- The dataset 2 hold the CC BY-NC-SA 4.0 licence that allows us to freely copy and share it.

4. **Data Description**

- The dataset 1 has about 9 separate csv's that describe the temperature and precipitation data of Indian cities.
- The dataset 2 has about 454 separate csv's that describes the AQI.
- Considering both the csv's, 6 Indian cities have been selected who's data we will study in our analysis.
- The cities selected are : Delhi, Mumbai, Chennai, Bangalore, Lucknow, and Jodhpur.
- The purpose of selecting these cities was purely based on availability of the data.

5. **Data Processing**

- All the preprocessing steps can be found in the 'preprocessing.ipynb' file that is present in the repository.
- The dataset 1 has columns: `['time', 'tavg', 'tmin', 'tmax', 'prcp']` that were later converted into: `['date', 'tavg', 'tmin', 'tmax', 'prcp']`.

- The dataset 2 has columns: `['From Date', 'To Date', 'PM2.5', 'PM10 (ug/m3)', 'NO (ug/m3)', 'NO2 (ug/m3)', 'NOx (ppb)', 'NH3 (ug/m3)', 'SO2 (ug/m3)', 'CO (mg/m3)', 'Ozone (ug/m3)', 'Benzene (ug/m3)', 'Toluene (ug/m3)','Eth-Benzene (ug/m3)', 'MP-Xylene (ug/m3)', 'O Xylene (ug/m3)', 'Temp (degree C)', 'RH (%)', 'WS (m/s)', 'WD (deg)', 'SRs(W/mt2)', 'BP (mmHg)', 'VWS (m/s)']` in which the columns `'From Date'` and `'To Date'` was replaced with 'date'.
- The dataset 2 had data based on hourly intervals which were converted into daily intervals by averaging the data based on their date.
- 6 files from both the datasets were taken into account and will be used for our analysis.
- The files with same cities were later merged to form a new datasets (csv's) that will contain the data of both pollutants and temperature and can be found in the 'data' folder.
- Additional data cleaning hasn't been performed to preserve the data and this can later be tackled in our further analysis.