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', 'NO', 'NO2', 'NOx', 'NH3', 'SO2', 'CO', 'Ozone', 'Benzene', 'Toluene','Eth-Benzene', 'MP-Xylene', 'O Xylene', 'Temp', 'RH', 'WS (m/s)', 'WD', 'SRs', 'BP', 'VWS']` 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.

6. Data Pipeline

- To pull the data from it source (kaggle) a custom data pipeline was made that can be used using 'pipeline.sh' in the project directory.
- It runs an underlying python script that pulls the data into the data folder.
- Rest of the tasks were performed in the 'preprocessing.ipynb' file. Rest of the steps are already addressed in [5].

7. Errors and Difficulties

- Pulling the dataset from kaggle was rather a tideous task but managed to do so by creating a throwaway account.
- The current dataset(s) has some **nan** values that need to be address during the final analysis. Although they have not been removed as we might loose a lot of important information that might be useful for our analysis.
- During the merge, since it's an 'inner' join it automatically handles missing data files.

8. Results and Limitations

- Since I didn't do much analysis about the data, the main answer to our question is still unknown.
- However, we now have a good pipeline and a process that can help us lead to our destination.
- We now have 6 data files in form of csv's that contain a time series data of most important cities in India.
- They depict the amount of pollutants and the temperature data of the cities.
- The datatype of our data was chosen to be csv's because they are widely used and can be easily interpreted with/without coding knowledge.
- The things that are still in process are a work for the future.

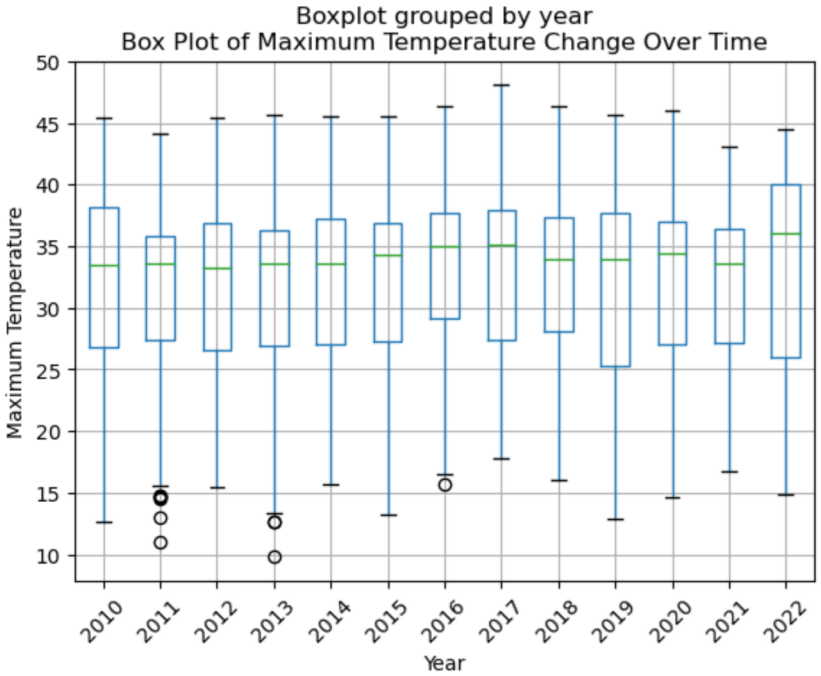
9. Future Work

- Making a robus pipeline that does all the tasks.
- Dealing with ‘nan’ values.
- Doing an in-depth Exploratory Data Analysis with newly created dataset.

10. Descriptive Statistics

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4589 entries, 0 to 4588
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   date             4589 non-null   object
 1   tavg             4587 non-null   float64
 2   tmin             4348 non-null   float64
 3   tmax             4540 non-null   float64
 4   prcp             1250 non-null   float64
 5   PM2.5 (ug/m3)    3208 non-null   float64
 6   PM10 (ug/m3)     1755 non-null   float64
 7   NO (ug/m3)       3549 non-null   float64
 8   NO2 (ug/m3)      3547 non-null   float64
 9   NOx (ppb)        3581 non-null   float64
 10  NH3 (ug/m3)      2519 non-null   float64
 11  SO2 (ug/m3)      3038 non-null   float64
 12  CO (mg/m3)       2921 non-null   float64
 13  Ozone (ug/m3)    3096 non-null   float64
 14  Benzene (ug/m3)  1379 non-null   float64
dtypes: float64(14), object(1)
memory usage: 537.9+ KB
```

It depics the data samples in Delhi Dataset



A box plot showing slight upward trend of the maximum temperature over the years.