CSE 545: Big Data Analytics Project Report: SDG 13 - Climate Action

Deepika Gonela 115129713

Pandre Vamshi 115077003

Krishna Tej Alahari 115192436

Akshith Reddy Kota 114855927

1. Introduction

The Climate Action Sustainable Development Goal aims to address the global challenge of climate change by focusing on strengthening resilience to climate-related hazards, integrating climate change measures into national policies, improving education and awareness, mobilizing resources for developing countries, and promoting effective climate change planning and management in vulnerable communities. The goal recognizes the urgent need for action to mitigate and adapt to the impacts of climate change, and emphasizes the importance of collaboration and capacity building to achieve a sustainable future for all.

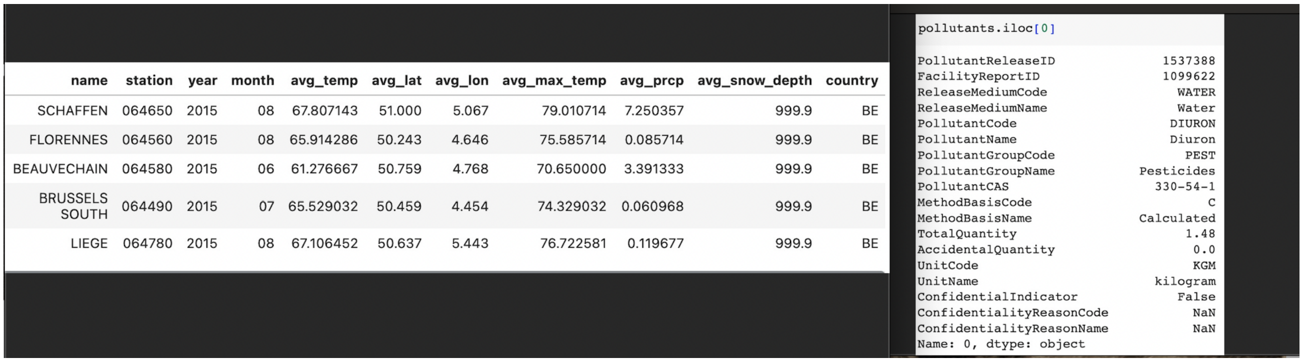
Our work includes finding similar facilities with similar emission profiles across the European Union so that similar policies can be made by the authorities to mitigate the environmental effects. We also find the correlations between many different pollutants and change in various climate variables. Finally, we also create a predictive model where we can predict the effect on a particular climate variable by a particular emission profile. In essence, we find facilities which pollute our environment in a similar way. Then we find what kind of pollutants released are actually causing a change in the environment and finally create a predictive model to predict the effect by a particular emission profile.

2. Sustainable Development Goal and Background

The sustainable development goal 13 focuses on taking urgent action to combat climate change and its impacts. The goal emphasizes the need for global cooperation to strengthen resilience and adaptive capacity to climate-related hazards, integrate climate change measures into national policies, and raise awareness and capacity on climate change mitigation and adaptation. The background for this goal is the pressing need to address the adverse impacts of climate change, such as rising sea levels, extreme weather events, and droughts, which disproportionately affect vulnerable communities, including persons with disabilities. The goal acknowledges that climate change is a cross-cutting issue that requires a holistic approach and collaboration across sectors and stakeholders.To achieve this, we are finding ways to aid the policymakers across many different sectors.

3. Data Description

We used the data from the European Pollutant and Transfer Register (E-PRTR). Furthermore, for getting the nearest weather station data for each of the facilities, we have used the data from the Daily Global Historical Climatology Network which is around 30GB labeled data and contains data from over around 100,000 weather stations across the world. In conclusion, emission profiles were created for each of the facilities in the E-PRTR dataset and nearest weather station data was merged with these emission profiles to do further computation. Also, the weather data around each facility was obtained from the nearest weather stations to the facility. The distance was calculated using Haversine Distance from the latitude and longitudes data of facilities and the weather stations.



name	station	year	month	avg_temp	avg_lat	avg_lon	avg_max_temp	avg_prcp	avg_snow_depth	country
SCHAFFEN	064650	2015	08	67.807143	51.000	5.067	79.010714	7.250357	999.9	BE
FLORENNES	064560	2015	08	65.914286	50.243	4.646	75.585714	0.085714	999.9	BE
BEAUVECHAIN	064580	2015	06	61.276667	50.759	4.768	70.650000	3.391333	999.9	BE
BRUSSELS SOUTH	064490	2015	07	65.529032	50.459	4.454	74.329032	0.060968	999.9	BE
LIEGE	064780	2015	08	67.106452	50.637	5.443	76.722581	0.119677	999.9	BE

```
pollutants.iloc[0]

PollutantReleaseID                   1537388
FacilityReportID                     1099622
ReleaseMediumCode                      WATER
ReleaseMediumName                      Water
PollutantCode                         DIURON
PollutantName                         Diuron
PollutantGroupCode                      PEST
PollutantGroupName                Pesticides
PollutantCAS                        330-54-1
MethodBasisCode                            C
MethodBasisName                   Calculated
TotalQuantity                           1.48
AccidentalQuantity                       0.0
UnitCode                                 KGM
UnitName                            kilogram
ConfidentialIndicator                  False
ConfidentialityReasonCode                NaN
ConfidentialityReasonName                NaN
Name: 0, dtype: object
```

Sample Weather stations data and facility data.

4. Methods

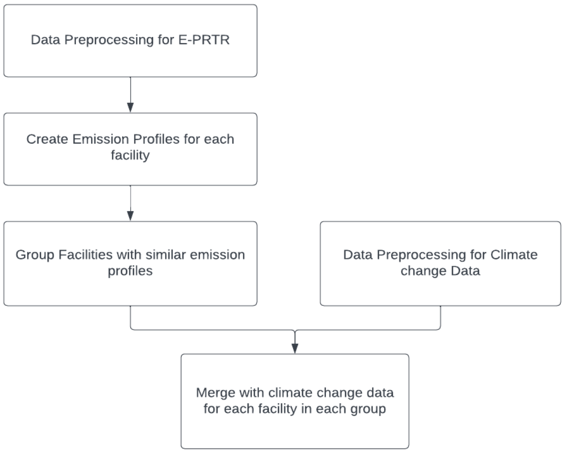
4.1. Emission Profile

Firstly, we merged the data from different datasets(dbo.PUBLISH_FACILITYREPORT.csv, dbo.PUBLISH_POLLUTANTRELEASE.csv, and dbo.PUBLISH_POLLUTANTRELEASEANDTRANSFERREPORT.csv). Groups finalData by 'FacilityID' and applies aggregation functions to various columns to obtain the sum of pollutant quantities. Extracts the pollutant columns from finalData and converts them to vectors to form emissionsProfiles.

4.2. Similarity Search

After the emission profiles are created, we have used Locality Sensitive Hashing to group the similar profiles. Firstly, we computed the signature matrix using Minhashing. We created MinHash objects for the signature matrix and built an LSH Forest using these objects. We retrieved similar groups by this created LSH Forest. Overall, we utilized MinHash and LSH techniques to efficiently find similar groups of items based on their MinHash signatures.

4.3. Data Preprocessing and Feature Engineering:


Data Preprocessing for E-PRTR
Create Emission Profiles for each facility
Group Facilities with similar emission profiles
Data Preprocessing for Climate change Data
Merge with climate change data for each facility in each group

In order to find the nearest station to a particular facility to see if there is any significant change in any of the major climate variables, we have processed the weather stations data from the Daily Global Historical Climatology Network. We have used the data from all the weather stations across the entire European Union. Essentially, we have collected the average temperature, average maximum temperature and average precipitation from all the weather stations across the entire European Union. This was done with the help of Big Query and further processing was performed using PySpark.

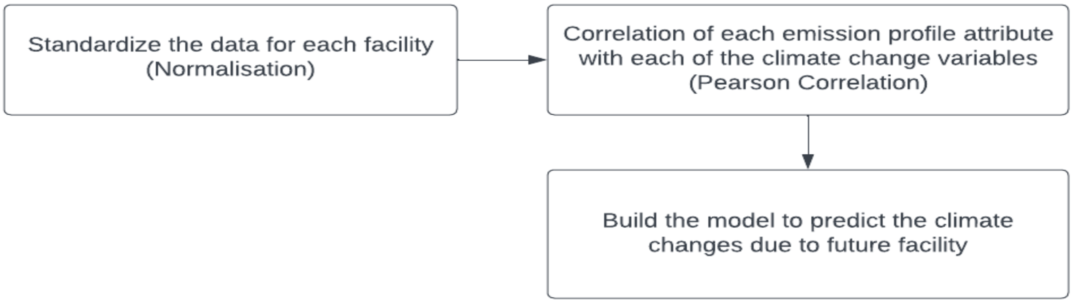
For data preprocessing, we have used Spark RDDs. Irrelevant data like all the other climate variables except from the ones mentioned above were removed. The dataset was filtered to only get the weather stations present in the European Union. Then, we have averaged the three variables (temperature, maximum temperature and precipitation) over the course of 4 years (2015 to 2018). Finally, we found the change in average temperature, change in average maximum temperature and change in average precipitation.

4.4. Merging the data:

In this step, we have merged the changes in climate variables with the emission profiles created in the initial step. We have used the latitudes and longitudes for both facilities and weather stations to get the nearest weather station for each of the facilities. This way, we can get the change in climate variables from the nearest weather station to a particular facility. We have used the Haversine distance formula to get the nearest weather station. The final output is in the following form.

FacilityID	Avg Temp change	Avg MaxTemp change	Avg Precipitation change
9281	-0.13213517665130325	-0.43583589349719176	0.0022748655913978205
25074	0.5886586700782317	0.6343004624479818	0.011937111291597767
182565	0.6064026363421533,	0.18599920733390718,	-7.4817809050716315

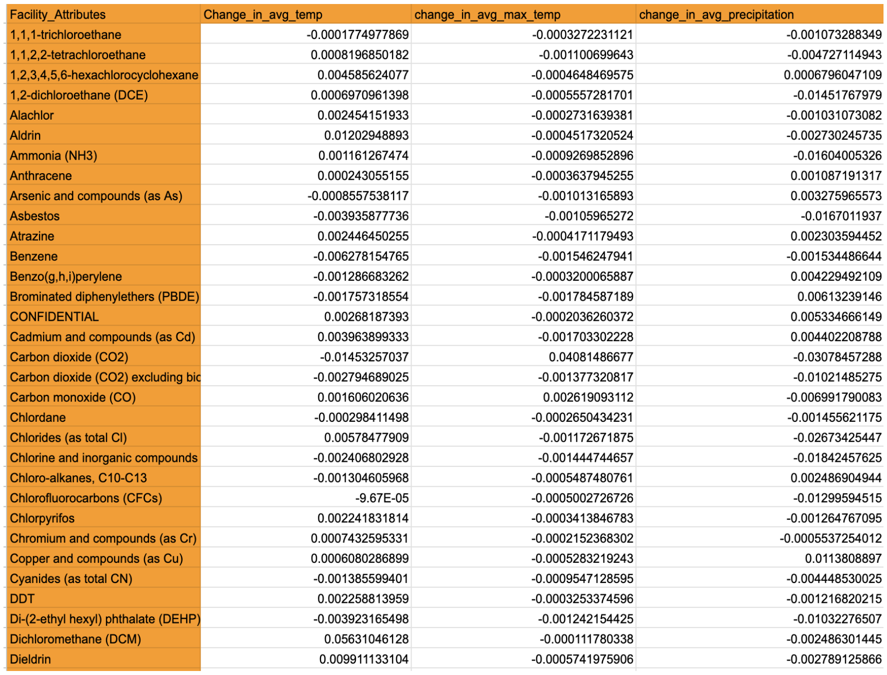
4.5. Creating a predictive model:


Standardize the data for each facility (Normalisation)
Correlation of each emission profile attribute with each of the climate change variables (Pearson Correlation)
Build the model to predict the climate changes due to future facility

To Perform this we have used LSTM (long short-term memory) Neural Network as it can process multiple sequences of time series data. The optimizer used is Adam (adaptive moment estimation) as it has faster computation time and less parameters for tuning. The loss function we used is MSE(mean squared loss). The data is split in 85:15, for training purposes. The training is done in the batches of 128, with 40 nodes and 15 epochs. It is a mapping of 89 input attributes to 3 output attributes. The model resulted in ~0.694 accuracy for the 15% test data set, with threshold for regression being 0.5. The built model is such that it could predict the climate variables like, change in average temperature, change in precipitation etc, given a new facility. This would help in determining the effect of the facility on the climate.

4.6. Creating a Correlation Matrix:

Correlation is calculated, after taking an average of all the involved emission profile attributes across all years for a particular facility, and this is done for all the facilities, eventually arriving at emission profile variables versus climate variables correlation matrix.



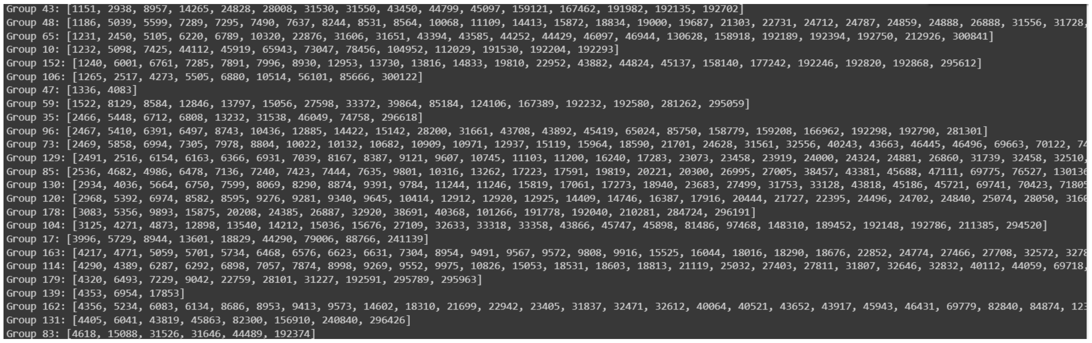
Facility_Attributes	Change_in_avg_temp	change_in_avg_max_temp	change_in_avg_precipitation
1,1,1-trichloroethane	-0.0001774977869	-0.0003272231121	-0.001073288349
1,1,2,2-tetrachloroethane	0.0008196850182	-0.001100699643	-0.004727114943
1,2,3,4,5,6-hexachlorocyclohexane	0.004585624077	-0.0004648469575	0.0006796047109
1,2-dichloroethane (DCE)	0.0006970961398	-0.0005557281701	-0.01451767979
Alachlor	0.002454151933	-0.0002731639381	-0.001031073082
Aldrin	0.01202948893	-0.0004517320524	-0.002730245735
Ammonia (NH3)	0.001161267474	-0.0009269852896	-0.01604005326
Anthracene	0.000243055155	-0.0003637945255	0.001087191317
Arsenic and compounds (as As)	-0.0008557538117	-0.001013165893	0.003275965573
Asbestos	-0.003935877736	-0.00105965272	-0.0167011937
Atrazine	0.002446450255	-0.0004171179493	0.002303594452
Benzene	-0.006278154765	-0.001546247941	-0.001534486644
Benzo(g,h,i)perylene	-0.001286683262	-0.0003200065887	0.004229492109
Brominated diphenylethers (PBDE)	-0.001757318554	-0.001784587189	0.00613239146
CONFIDENTIAL	0.00268187393	-0.0002036260372	0.005334666149
Cadmium and compounds (as Cd)	0.003963899333	-0.001703302228	0.004402208788
Carbon dioxide (CO2)	-0.01453257037	0.04081486677	-0.03078457288
Carbon dioxide (CO2) excluding bic	-0.002794689025	-0.001377320817	-0.01021485275
Carbon monoxide (CO)	0.001606020636	0.002619093112	-0.006991790083
Chlordane	-0.000298411498	-0.0002650434231	-0.001455621175
Chlorides (as total Cl)	0.00578477909	-0.001172671875	-0.02673425447
Chlorine and inorganic compounds	-0.002406802928	-0.001444744657	-0.01842457625
Chloro-alkanes, C10-C13	-0.001304605968	-0.0005487480761	0.002486904944
Chlorofluorocarbons (CFCs)	-9.67E-05	-0.0005002726726	-0.01299594515
Chlorpyrifos	0.002241831814	-0.0003413846783	-0.001264767095
Chromium and compounds (as Cr)	0.0007432595331	-0.0002152368302	-0.0005537254012
Copper and compounds (as Cu)	0.0006080286899	-0.0005283219243	0.0113808897
Cyanides (as total CN)	-0.001385599401	-0.0009547128595	-0.004448530025
DDT	0.002258813959	-0.0003253374596	-0.001216820215
Di-(2-ethyl hexyl) phthalate (DEHP)	-0.003923165498	-0.001242154425	-0.01032276507
Dichloromethane (DCM)	0.05631046128	-0.000111780338	-0.002486301445
Dieldrin	0.009911133104	-0.0005741975906	-0.002789125866

4.7. Hypothesis Testing:

The correlation matrices and p-value matrices were calculated for all the pollutant variables vs climate change variables. We then identified the top 3 correlation values in each correlation matrix. Next, we conducted hypothesis testing for each group to suggest policies. For example, In one group the release of mercury significantly increases the temperature. So, we can suggest ways to reduce the mercury emission to that group.

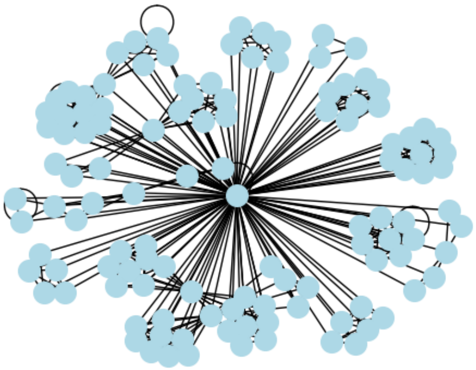
5. Results:

5.1. Similarity Search:



```
Group 43: [1151, 2938, 8957, 14265, 24828, 28008, 31530, 31550, 43450, 44799, 45097, 159121, 167462, 191982, 192135, 192702]
Group 48: [1186, 5039, 5599, 7289, 7295, 7490, 7637, 8244, 8531, 8564, 10068, 11109, 14413, 15872, 18834, 19000, 19687, 21303, 22731, 24712, 24787, 24859, 24888, 26888, 31556, 31728,
Group 65: [1231, 2450, 5105, 6220, 6789, 10320, 22876, 31606, 31651, 43394, 43585, 44252, 44429, 46097, 46944, 130628, 158918, 192189, 192394, 192750, 212926, 300841]
Group 10: [1232, 5098, 7425, 44112, 45919, 65943, 73047, 78456, 104952, 112029, 191530, 192204, 192293]
Group 152: [1240, 6001, 6761, 7285, 7891, 7996, 8930, 12953, 13730, 13816, 14833, 19810, 22952, 43882, 44824, 45137, 158140, 177242, 192246, 192820, 192868, 295612]
Group 106: [1265, 2517, 4273, 5505, 6880, 10514, 56101, 85666, 300122]
Group 47: [1336, 4083]
Group 59: [1522, 8129, 8584, 12846, 13797, 15056, 27598, 33372, 39864, 85184, 124106, 167389, 192232, 192580, 281262, 295059]
Group 35: [2466, 5448, 6712, 6808, 13232, 31538, 46049, 74758, 296618]
Group 96: [2467, 5410, 6391, 6497, 8743, 10436, 12885, 14422, 15142, 28200, 31661, 43708, 43892, 45419, 65024, 85750, 158779, 159208, 166962, 192298, 192790, 281301]
Group 73: [2469, 5858, 6994, 7305, 7978, 8804, 10022, 10132, 10682, 10909, 10971, 12937, 15119, 15964, 18590, 21701, 24628, 31561, 32556, 40243, 43663, 46445, 46496, 69663, 70122, 74
Group 129: [2491, 2516, 6154, 6163, 6366, 6931, 7039, 8167, 8387, 9121, 9607, 10745, 11103, 11200, 16240, 17283, 23073, 23458, 23919, 24000, 24324, 24881, 26860, 31739, 32458, 32510,
Group 85: [2536, 4682, 4986, 6478, 7136, 7240, 7423, 7444, 7635, 9801, 10316, 13262, 17223, 17591, 19819, 20221, 20300, 26995, 27005, 38457, 43381, 45688, 47111, 69775, 76527, 130136
Group 130: [2934, 4036, 5664, 6750, 7599, 8069, 8290, 8874, 9391, 9784, 11244, 11246, 15819, 17061, 17273, 18940, 23683, 27499, 31753, 33128, 43818, 45186, 45721, 69741, 70423, 71805
Group 120: [2968, 5392, 6974, 8582, 8595, 9276, 9281, 9340, 9645, 10414, 12912, 12920, 12925, 14409, 14746, 16387, 17916, 20444, 21727, 22395, 24496, 24702, 24840, 25074, 28050, 3160
Group 178: [3083, 5356, 9893, 15875, 20208, 24385, 26887, 32920, 38691, 40368, 101266, 191778, 192040, 210281, 284724, 296191]
Group 104: [3125, 4271, 4873, 12898, 13540, 14212, 15036, 15676, 27109, 32633, 33318, 33358, 43866, 45747, 45898, 81486, 97468, 148310, 189452, 192148, 192786, 211385, 294520]
Group 17: [3996, 5729, 8944, 13601, 18829, 44290, 79006, 88766, 241139]
Group 163: [4217, 4771, 5059, 5701, 5734, 6468, 6576, 6623, 6631, 7304, 8954, 9491, 9567, 9572, 9808, 9916, 15525, 16044, 18016, 18290, 18676, 22852, 24774, 27466, 27708, 32572, 3278
Group 114: [4290, 4389, 6287, 6292, 6898, 7057, 7874, 8998, 9269, 9552, 9975, 10826, 15053, 18531, 18603, 18813, 21119, 25032, 27403, 27811, 31807, 32646, 32832, 40112, 44059, 69718,
Group 179: [4320, 6493, 7229, 9042, 22759, 28101, 31227, 192591, 295789, 295963]
Group 139: [4353, 6954, 17853]
Group 162: [4356, 5234, 6083, 6134, 8686, 8953, 9413, 9573, 14602, 18310, 21699, 22942, 23405, 31837, 32471, 32612, 40064, 40521, 43652, 43917, 45943, 46431, 69779, 82840, 84874, 123
Group 131: [4405, 6041, 43819, 45863, 82300, 156910, 240840, 296426]
Group 83: [4618, 15088, 31526, 31646, 44489, 192374]
```

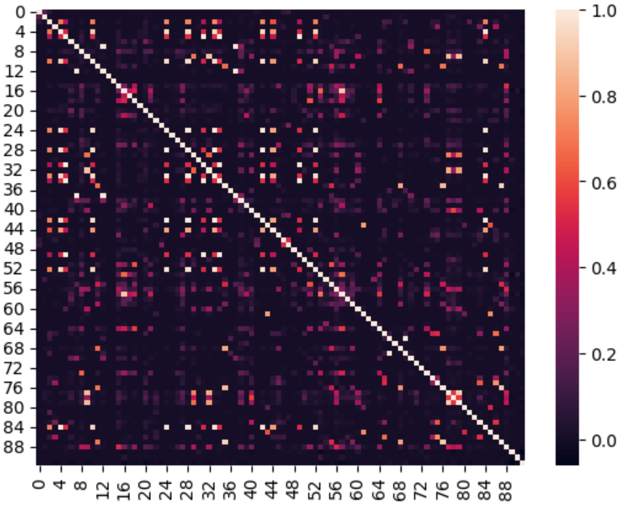
The above picture shows the groups of similar facilities. Similarity search was done on the emission profiles for each of the facilities.



Similar Groups

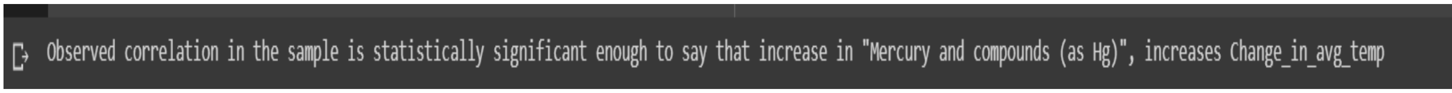
The above network signifies the relationships among the facilities across the entire European Union. The clusters that can be seen in the network depict the facilities that are similar in emitting the pollutants and thereby polluting the environment.

5.2. Correlation between Attributes: The below correlation heatmap signifies the correlation between the pollutants in the emission profiles along with climate change variables. Last 3 columns signifies how each emission element is correlated with climate change variables. Rest of the columns signify how each emission element is correlated with each emission element. This correlation also needs to be taken into consideration as it helps if the gasses are being released by the same material, thereby helping the government to have better policies on materials used, and not just the emissions.



Correlation HeatMap with Pollutants and Climate Change

5.3. Hypothesis Testing:



```
Observed correlation in the sample is statistically significant enough to say that increase in "Mercury and compounds (as Hg)", increases Change_in_avg_temp
```

6. Conclusion

It is crucial to analyze which emissions released by facilities have an impact on climate change. Below are the takeaways from our project:

- Identify the key emissions that have an influence on climate change variables and take measures to prevent them.
- When a new facility is constructed and we happen to know the emission profile of that facility, we can use the constructed predictive model to predict how this facility may affect the environment.
- We also found the facilities that are similar in the sense they release pollutants into the environment. This can aid policymakers in making similar policies for those similar facilities and mitigate the harmful effects on the environment.
- Also, hypothesis testing on each group will give a better idea on what policies need to be imposed on the new facility, as in maybe an existing policy that has been imposed on the rest of the facilities in the group.

7. References

[1] European Transfer Release and Transfer Register
[2] Gemstat dataset
[3] Research Paper - http://www.jatit.org/volumes/Vol100No24/27Vol100No24.pdf
[4] How to query the NOAA GHCN Daily Weather Data
[5] Locality Sensitive Hashing
[6] LSTM model in Pytorch