Climate change and its impact on health

-Ankit Desai(atd363)
-Himanshu Vaghela(hkv214)
-Mukul Jangid(mdj327)
-Varun Dhuldhoya(vyd208)

Table of contents

Introduction

So what is climate?
It is the long term average of weather.

And it is this long-term average, the changes in it and the impact due to its change is what we have tried to study in the project.

The reason such a study is important is to understand how change in climate has effects on people that they would not be able to attribute to Climate should they not be able to see the data.

Thus this analysis provides a view of how a city or state or country can make use of the data it collects and the impacts on its citizen and is of prime importance to activists and Non-Profit Organizations in order to bring to light the effects of Climate Change.

Our project wanted to explore the results of climate data after processing, Collecting and storing of huge amounts of weather data. The data collection is done by the Meteorological department. The Meteorological departments use different types of sensors such as temperature, humidity etc. to get the data. The sensors volume and velocity of data in each of the sensor make the data processing time consuming and complex. We aim at analyzing the climate change and its impact on the environment as well as human health by using the available Big data technologies.

We have exploited the opportunities to mine large climate datasets, temperature datasets and a global health and death dataset with an emphasis on the visualization of the mined data. We applied Big Data technologies to analyze and draw correlations among the radical features causing climate change. Machine Learning (ML) is all about predicting future data based on patterns in existing

data. The last step of our project was to use the Machine Learning Algorithms to predict tomorrow's temperature, by feeding the test set to the model.

Datasets used

1.U.S. Pollution Data

The dataset provides us an insight into 17 years of pollution data from 2000-2016 which measures the Parts Per Billion (PPB) of 4 pollutants - NO2, SO2, CO, O3. We will clean and analyze this data to get an understanding of the pollution trends and the cities affected by it. The original dataset contains 28 fields:

1. State Code: The code allocated by US EPA to each state
2. County Code: The code of counties in a specific state allocated by US EPA
3. Site Num: The site number in a specific county allocated by US EPA
4. Address: Address of the monitoring site
5. State: State of monitoring site
6. County: County of monitoring site
7. City: City of the monitoring site
8. Date Local: Date of monitoring

The four pollutants (NO2, O3, SO2 and O3) each has 5 specific columns. For instance, for NO2:
• NO2 Units: The units measured for NO2
• NO2 Mean: The arithmetic means of concentration of NO2 within a given day
• NO2 AQI: The calculated air quality index of NO2 within a given day
• NO2 1st Max Value: The maximum value obtained for NO2 concentration in a given day
• NO2 1st Max Hour: The hour when the maximum NO2 concentration was recorded in a given day

Out of this data we take the mean and AQI for each pollutant along with state, city and Date Local

Source-US EPA(Environmental Protection Agency) Data

2.Climate Change: Earth Surface Temperature Data

This dataset goes back till 1750,of which we have extracted the relevant part of the data and use it to augment our pollutant data inorder to find a relation between the the level of pollutants and the temperatures and try to predict the temperature changes.
The dataset contains the following attributes:

- dt-date
- Average Temperature-Average temperature in C
- Average Temperature Uncertainty-Average uncertainty in measurement of temperature
- City-city names
- County-county names
- Latitude-latitude values
- Longitude-longitude values

Source-Berkeley Earth

3.NCHS Leading Causes of Death United States

This datset gives us a detailed view of the figures related to deaths in the US and we use these figures to understand the impact the change in pollutant AQI (Air Quality Index) levels and rising tempearture have had on the the US population.

The dataset has the following attributes:

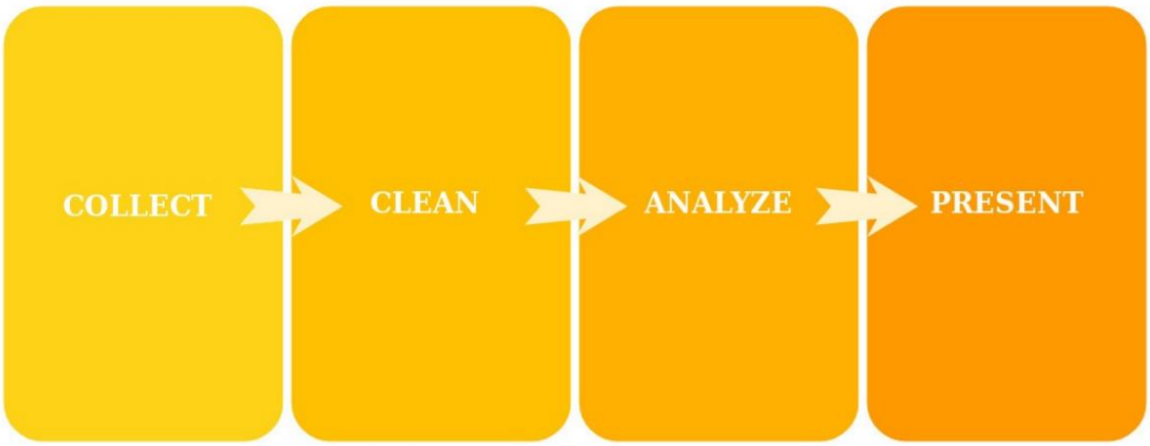
	Name	Description
1.	Year	1999-2015
2.	113 Cause Name	Disease Name
3.	Cause name	Disease Category
4.	State	State name
5.	Deaths	Total number of deaths
6.	Age-adjusted Death Rate	To make fairer comparisons between groups with different age distributions

Based on the different causes of death we decided to filter out deaths caused by certain factors like Accidents, Assaults, murder as we realized that they do not have any direct relationship with the pollution levels in the atmosphere

Source- US CDC(Centers for Disease Control and Prevention) Data

Steps Taken

Process
For analysis of each of the dataset, we had to go through the following steps.



a) For collection part we had data that was accumulated from the sources and available on Kaggle.

b) The cleaning part however took time. Most of the data had atleast one column value missing and we used technique ranging from removing the data points to replacing the value with the mean.
c) For analysis of the data we had to select and filter the columns and used pyspark sql to run queries on the data.
d) For presenting the data we learnt how to use matplotlib and seaborn library

Pollutant Level

We started by first taking the US pollution data set.
It has 22 columns ranging from state and county code to adresses of measuring station. We extracted only the required values of NO2, SO2, O3 and CO.
Then for the first initial analysis we just took each of the pollutants individually and plotted their means over the years. This helped us see that the pollution appeared to decrease. This made it a bit suspicious about the data but we have explained about this in the later slides.

Most Frequently Polluted Cities

Next to find the most frequently polluted cities, we first filtered out entries that had Max values of the pollutants less than 150. Below 150 the AQI(Air Quality Index) is considered safe and moderate.
Then we found the total pollutants in each of the cities and saw for each city how many times they appear in our data over the 17 years.
We ten sorted and plotted the data to find the top 10 most frequently appearing cities in this list.

City Wise Analysis

Then we tried another way of representation of data to extract the pollution over the years.
We took those 10 cities from the dataset and for each year we made a heat map of their mean pollution level.
This allows us to see time and pollution levels simultaneously for various cities.
This helps us analyse and compare the various if the pollution trend seen in some cities was a global trend or individual.
It also helps us identify which cities had similar pollution level through time.

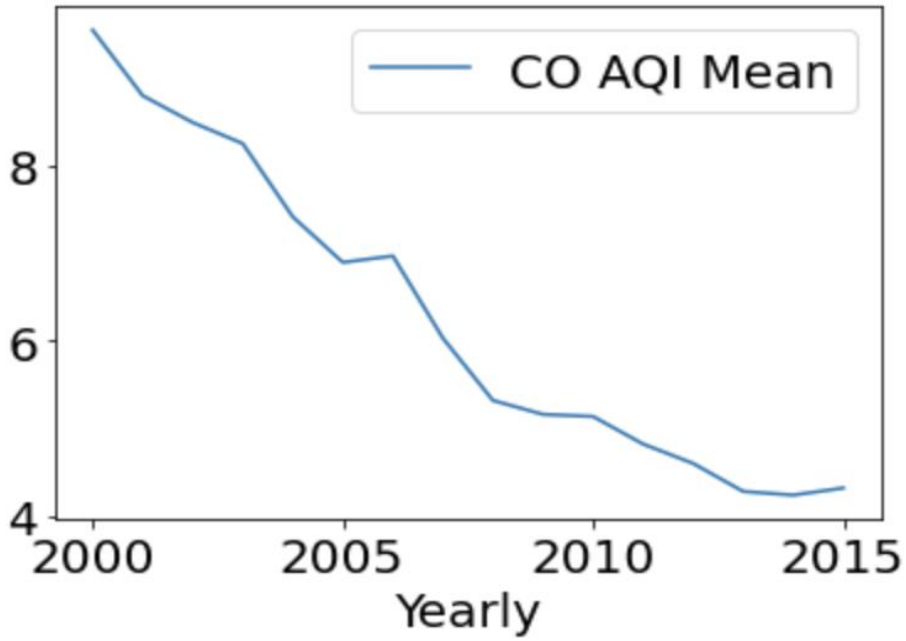
Pollution Constituents

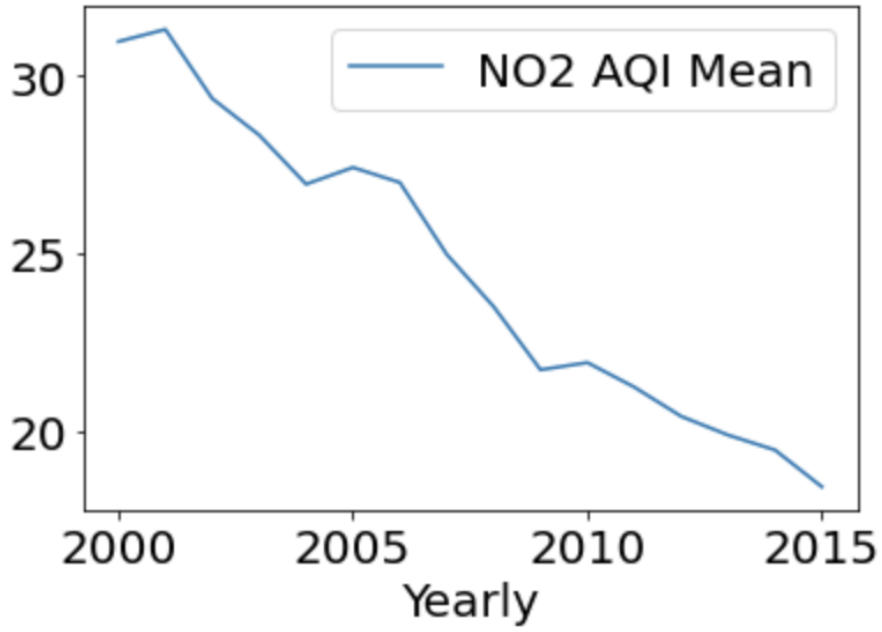
Next we tried to find the constituents of the total pollution in the year 2000.
We found the highest levels of pollutions recorded in the year 2000 and then for 10 cities which corresponded to it, made stacked bar chart.
This helps us understand SO2 and O3 were primary pollutants.
Such studies can be helpful in trying to understand the which pollutants require higher attention and how the environment affects the way the pollution of a place is.

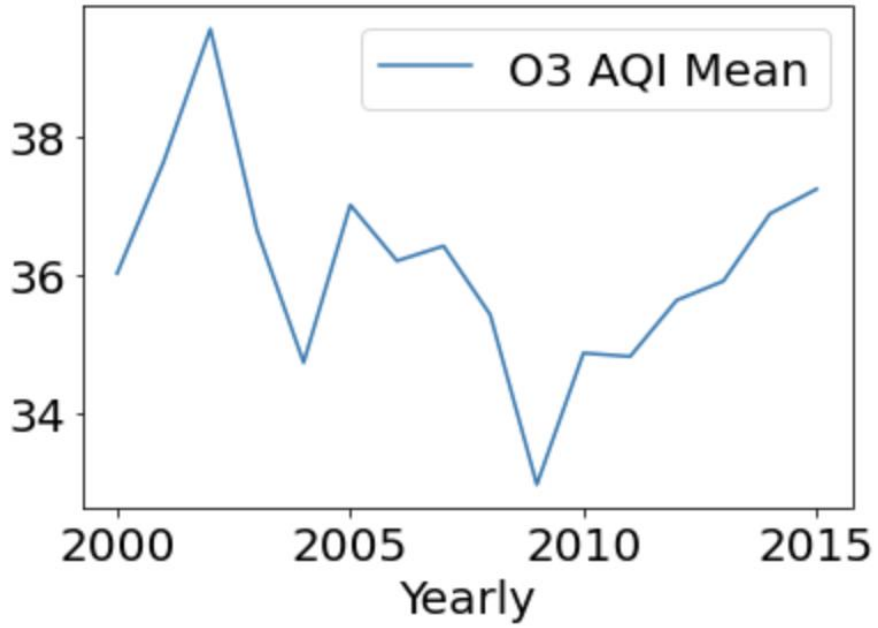
Pollution Data mapping 4 pollutants

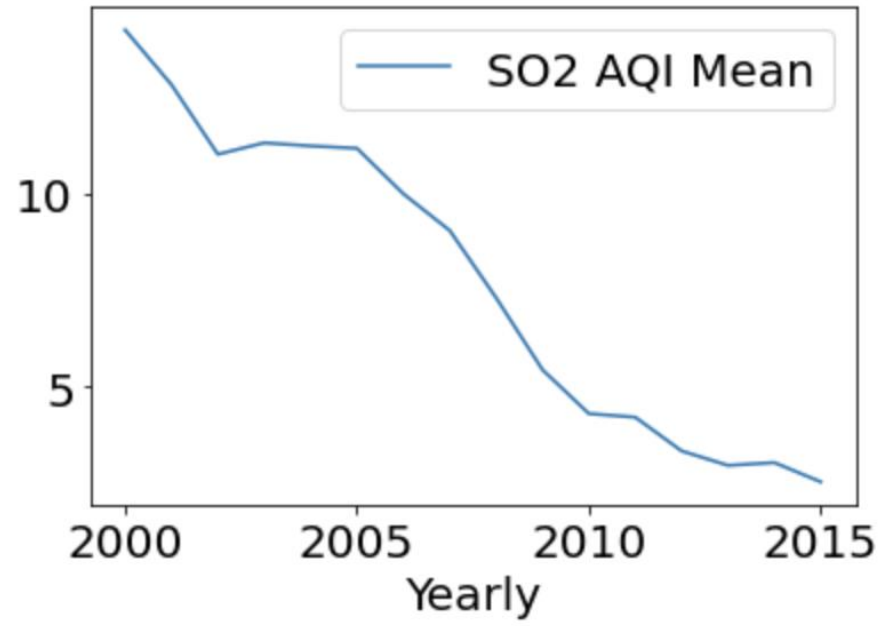
AQI- AIR QUALITY INDEX

The AQI is an index for reporting THE daily air quality. It tells us how clean or polluted the air is, and what associated health effects might be a concern for you. Thus, our main aim was to first calculate the average daily AQI for each city, then calculate the average monthly AQI levels for each city and filter out the cities with AQI levels lower than 80 because an AQI level above 150 is considered as dangerous by the Environment pollution agency. We used pyspark to analyze our dataset and the result we obtained was as follows.
Here we show the pollutant levels with respect to time by plotting the yearly AQI mean.


CO AQI Mean
8
6
4
2000
2005
2010
2015
Yearly

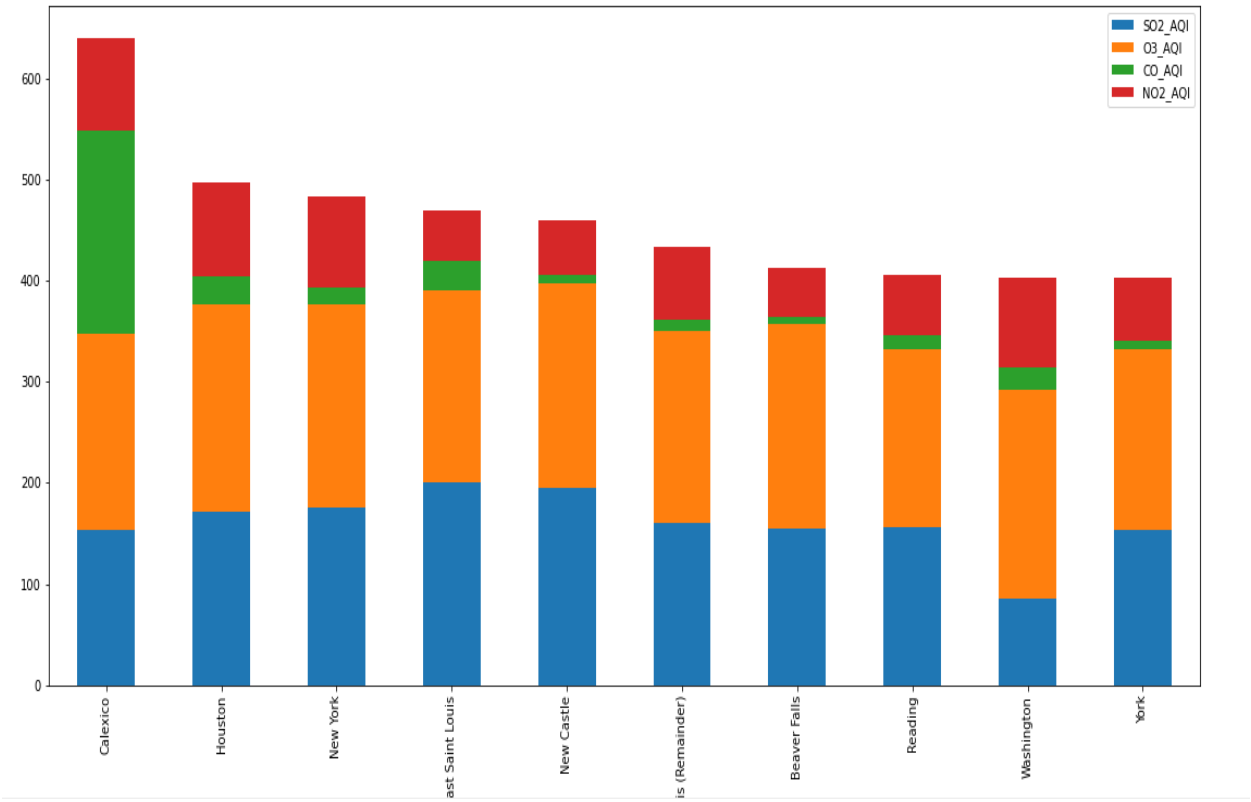

NO2 AQI Mean
30
25
20
2000
2005
2010
2015
Yearly


O3 AQI Mean
38
36
34
2000
2005
2010
2015
Yearly


SO2 AQI Mean
10
5
2000
2005
2010
2015
Yearly

Top Cities Max Polluted Bar Stacked

Here we can see the constituents of pollutants fot different cities stacked in a bar chart. This allows us to implement different pollution mitigation strategies for each city since we know which pollutant affects it the most

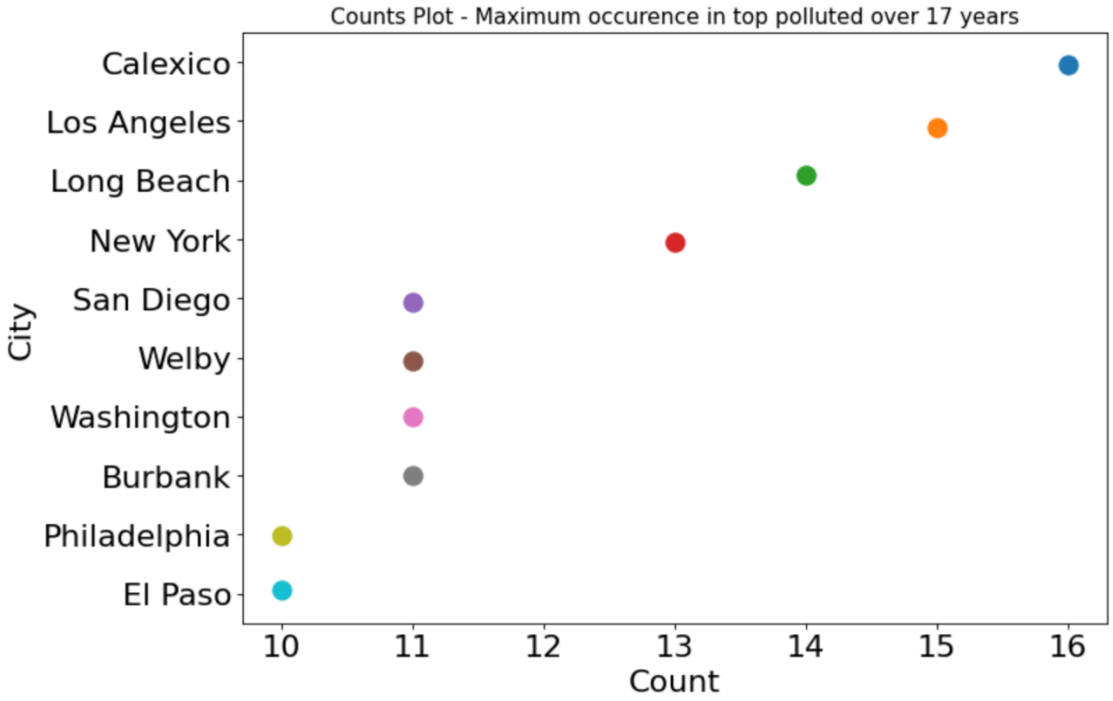


Rank of cities by appearance

Here we calculate the cities that have had the highest count of yearly AQI level for pollutants above 150 for the last 17 years

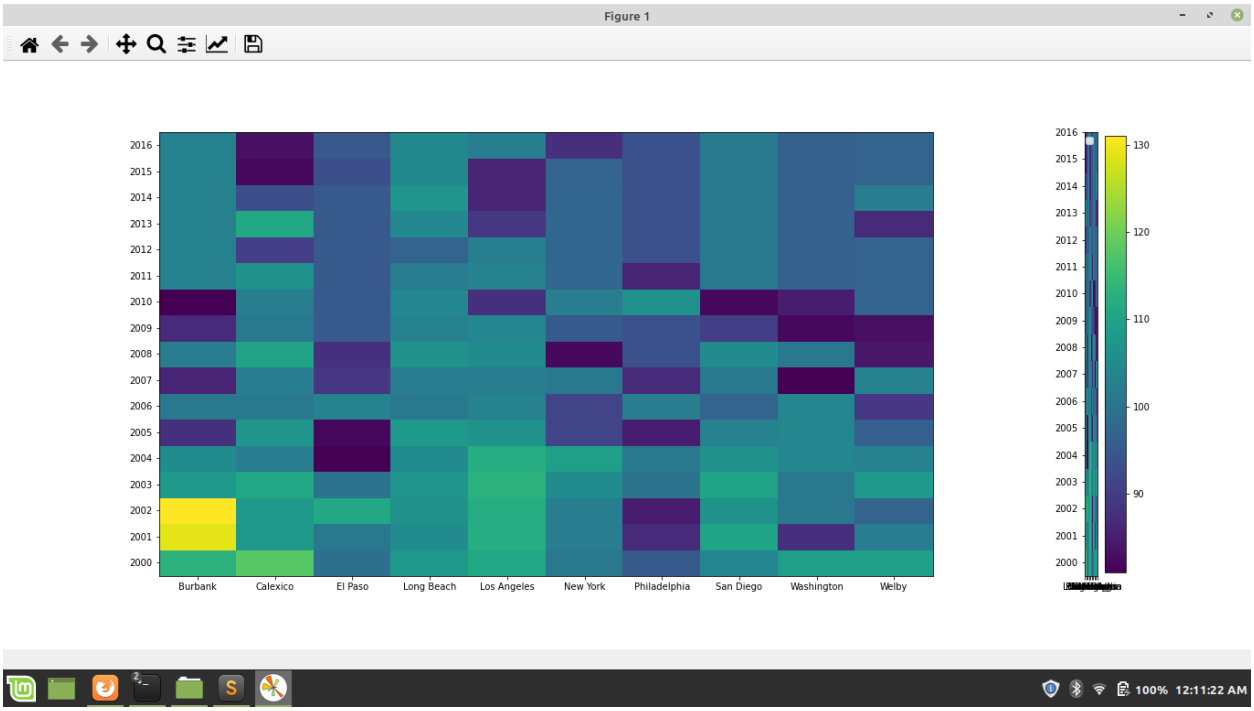
State	City	Count
California	Calexico	16
California	Los Angeles	15
California	Long Beach	14
New York	New York	13
California	San Diego	11
Colorado	Welby	11
District Of Columbia	Washington	11
California	Burbank	11
Texas	El Paso	10
Pennsylvania	Philadelphia	10

Then we visualize this


Counts Plot - Maximum occurence in top polluted over 17 years
Calexico
Los Angeles
Long Beach
New York
San Diego
Welby
Washington
Burbank
Philadelphia
El Paso
City
10
11
12
13
14
15
16
Count

Heat Map showing Pollution

A heat map is a good way to see the similarities between cities' pollutants trends.



Decision Tree with Actual vs. Predicted

A decision tree is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their costs, probabilities, and benefits. They can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically. Decision trees where the target variable can take continuous are called regression trees. We used the Decision tree for a regression analysis to predict the Average Temperature of different from based on their levels of pollution in air.
Our Target variable was: Temperature

This figure shows the decision tree modeling with only a mean squared error of 21.65%

```
Test Mean Squared Error = 21.657031520363805
Learned regression tree model:
DecisionTreeModel regressor of depth 5 with 55 nodes
  If (feature 1 <= 45.0)
   If (feature 1 <= 10.5)
    If (feature 3 <= 100.5)
     If (feature 1 <= 6.5)
      If (feature 3 <= 56.0)
       Predict: 9.752458333333333
      Else (feature 3 > 56.0)
       Predict: 11.381555555555558
     Else (feature 1 > 6.5)
      If (feature 0 <= 37.5)
       Predict: 9.854333333333335
      Else (feature 0 > 37.5)
       Predict: 14.522583333333332
    Else (feature 3 > 100.5)
     If (feature 3 <= 150.5)
      If (feature 0 <= 57.5)
       Predict: 17.455601851851853
      Else (feature 0 > 57.5)
       Predict: 16.02006388888888
     Else (feature 3 > 150.5)
      If (feature 2 <= 17.5)
       Predict: 18.09475
      Else (feature 2 > 17.5)
       Predict: 20.220930555555558
   Else (feature 1 > 10.5)
    If (feature 1 <= 32.0)
     If (feature 2 <= 12.0)
      If (feature 1 <= 18.0)
       Predict: 17.895759259259258
      Else (feature 1 > 18.0)
       Predict: 24.915215277777776
     Else (feature 2 > 12.0)
      If (feature 0 <= 46.5)
       Predict: 13.846708333333334
      Else (feature 0 > 46.5)
       Predict: 17.73110907605466
    Else (feature 1 > 32.0)
     If (feature 2 <= 14.5)
      If (feature 3 <= 85.5)
       Predict: 15.000916666666669
      Else (feature 3 > 85.5)
       Predict: 9.617475000000002
     Else (feature 2 > 14.5)
      If (feature 2 <= 15.5)
       Predict: 21.04595833333333
      Else (feature 2 > 15.5)
       Predict: 16.134452256944446
  Else (feature 1 > 45.0)
   If (feature 2 <= 34.5)
    If (feature 0 <= 80.5)
     If (feature 0 <= 66.5)
      If (feature 2 <= 9.5)
       Predict: 15.191780748663103
      Else (feature 2 > 9.5)
       Predict: 12.722854259259261
     Else (feature 0 > 66.5)
      If (feature 3 <= 95.0)
       Predict: 11.270083333333334
      Else (feature 3 > 95.0)
       Predict: 16.460012566137568
    Else (feature 0 > 80.5)
     If (feature 3 <= 175.5)
      If (feature 3 <= 142.5)
       Predict: 12.329833333333333
      Else (feature 3 > 142.5)
       Predict: 10.37367708333333
     Else (feature 3 > 175.5)
      Predict: 20.639999999999986
   Else (feature 2 > 34.5)
    If (feature 0 <= 41.5)
     If (feature 1 <= 59.5)
      Predict: 8.810125000000001
     Else (feature 1 > 59.5)
      Predict: 9.828000000000001
```

Conclusion

The objective of this project is to correlate pollution levels,rise in temperature and death counts for cities. This is important as WHO states that both the detection and measurement of health effects due to climate change are necessary as these act as evidence for creating national and international polices such as reduction of emission greenhouse gases. But while doing this project we encountered an unusual inference where over time due to awareness about climate change the pollutants in the air have reduced but despite this the number of deaths have been increasing. Hence we can see that there might be other factors affecting deaths.

Lessons we have learned:

- Data Cleaning, removing non-informative attributes, feature engineering using Spark
- Applying machine learning algorithms to the Big Data files to make future predictions using Spark MLlib.
- Visualization techniques such as stacked bar charts, seaborn strip plot, heatmaps.
- Teamwork