Project Report

David Tran

University of Missouri - Kansas City

CS465R: Statistical Learning (Undergraduate)

Summary

Air quality is a concern that should be raised on a global scale and there should be more regulations to prevent constant air pollution. Pollutants can cause illnesses and blocks the atmosphere, which could cause changes in climate. One of those pollutants is particle matter that is under 2.5 micrometers (PM2.5). This project looks at cities within the United States and compares cities known for being clean with cities known for being highly polluted.

In order to get a comparable measurement, specific parameters were set. The five cities that will be compared are Kansas City, Bakersfield, Fairbanks, Honolulu, and Wilmington. With those cities, PM2.5 levels from the November 2019 will be used. The safety levels of those cities can be determined, and trends will appear. With those trends, predictions and correlations will be made.

There are various factors that could play in to air quality levels, such as population, population density, climate, and elevation. Higher population and population density signifies that more emissions will be created, such as from increase of cars. Climate has an effect because it can determine how air particles stay within the city's atmosphere. Elevation is also a factor because of PM2.5 size and the amount of atmosphere available above the city changes the density of the pollution.

Python is used to create plots for each city using Pandas and Matplotlib.pyplot libraries. The cities that are known to be clean had lower levels of PM2.5, while the cities that are known for being more polluted had higher levels of PM2.5. Based on all of the data collected, a correlation can only be directly made between the pollution level to the elevation of the city.

Table of Contents

Problem Statement / Proposal

Poor air quality due to pollution and other factors is an invisible danger that affects every human on Earth. There is a direct correlation between decreasing air quality levels and increase in pollution. There are multiple factors that can affect air quality levels, such as population, population density, elevation, climate, and industry. A specific danger is PM2.5, which is particle matter that is less than 2.5 micrometers in diameter. For perspective, human hair typically has a diameter of 50-70 micrometers.

PM2.5 occurs from numerous sources, such as power plants, motor vehicles, wood burning, forest fire, dust storms, etc. Because of their small size, PM2.5 can stay in the air for a longer period of time. When inhaled, these particles can bypass the nose and throat and end up directly deep in the lungs. Because of this, there are many health concerns that can be caused by PM2.5, which include heart and lung diseases, such as asthma, heart attack, and bronchitis. On a less severe scale, PM2.5 can cause things such as eye/nose/throat irritation, coughing, sneezing, runny nose, and shortness of breath.

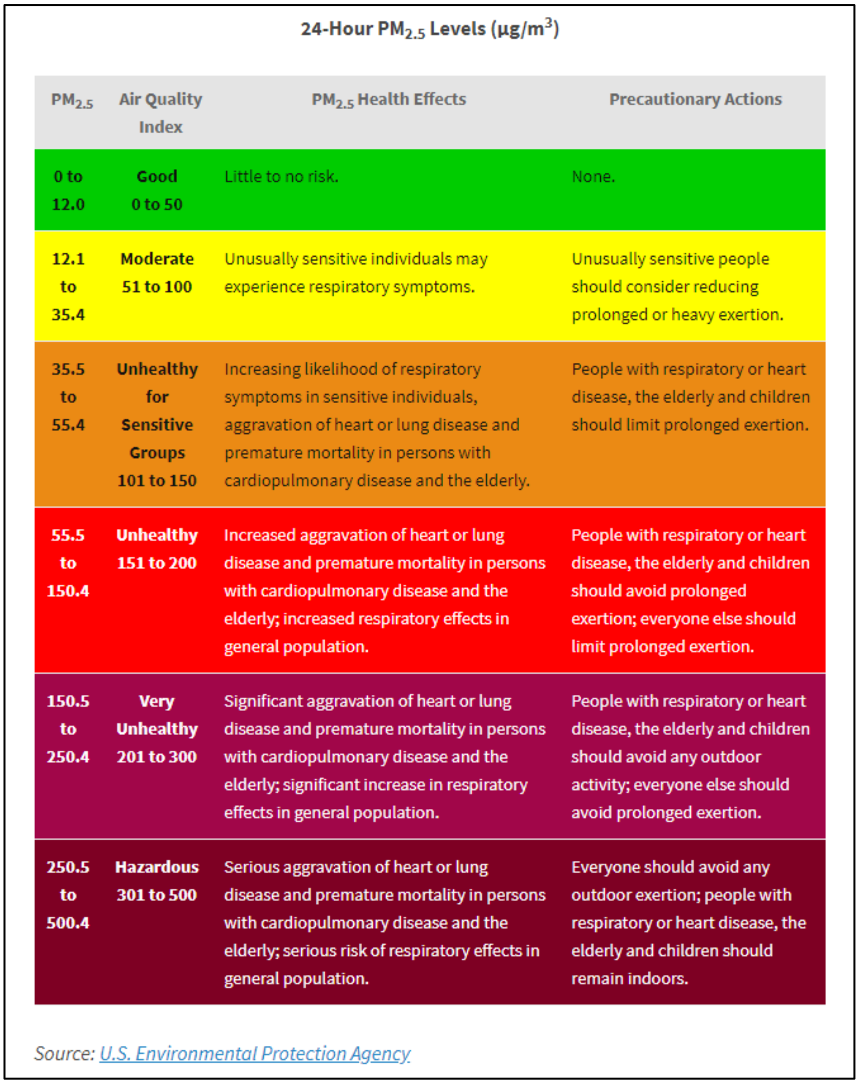
This project looks specifically at 5 cities within the United States, with 2 being known as very polluted and 2 known as being relatively clean. Those 5 cities are Kansas City, Bakersfield, Fairbanks, Honolulu, and Wilmington. In order to get comparable information, the dataset will be limited to all readings within a particular month, which is November 2019.

To present the dataset, the data points will be graphed using a scatter plot with the horizontal axis being time, and the vertical axis being the value of PM2.5. Using these points, a linear regression model was fitted to each city's graph. This line of best fit can be used to find a trend within the air quality levels of each city. Based on that trend, the current safeness and the future air quality of the city can be predicted.

With that information, air quality trends can be shown to be enforced by other factors. The city information can provide insight as to factors that could increase or decrease air quality. Those factors include population, population density, climate, altitude, and industry.

Literature Survey / Related Works

From the United States Environmental Protection Agency (EPA):



24-Hour $PM_{2.5}$ Levels (µg/m^3)

$PM_{2.5}$	Air Quality Index	$PM_{2.5}$ Health Effects	Precautionary Actions
0 to 12.0	Good 0 to 50	Little to no risk.	None.
12.1 to 35.4	Moderate 51 to 100	Unusually sensitive individuals may experience respiratory symptoms.	Unusually sensitive people should consider reducing prolonged or heavy exertion.
35.5 to 55.4	Unhealthy for Sensitive Groups 101 to 150	Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.	People with respiratory or heart disease, the elderly and children should limit prolonged exertion.
55.5 to 150.4	Unhealthy 151 to 200	Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid prolonged exertion; everyone else should limit prolonged exertion.
150.5 to 250.4	Very Unhealthy 201 to 300	Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid any outdoor activity; everyone else should avoid prolonged exertion.
250.5 to 500.4	Hazardous 301 to 500	Serious aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; serious risk of respiratory effects in general population.	Everyone should avoid any outdoor exertion; people with respiratory or heart disease, the elderly and children should remain indoors.

Source: U.S. Environmental Protection Agency

The U.S. Clean Air Act dictates that the government can regulate air emissions in order to control air pollution throughout the nation. Based on the EPA's findings for safety levels of PM2.5, concerns for the general public should be raised once daily PM2.5 levels reach at least 35.5. Pollution of all kinds, even PM2.5, can be created through basic living actions, such as driving or cooking.

Data Set / Data Collection

The data set is sourced from open air quality information provided by OpenAQ (https://openaq.org/). OpenAQ's API can be accessed through the link (https://api.openaq.org/v1/measurements), and when called on its own, the API will provide information about air quality for certain cities with measurements that were obtained at the time the API was called.

Parameters that can be used to acquire information from the API includes:

country	: limit results by certain country
city	: limit results by certain city
location	: limit results by a certain location
parameter	: limit parameters (available = pm25, pm10, so2, no2, o3, co) (Particle Matter, Sulfur Dioxide, Nitrogen Dioxide, Ozone, Carbon Monoxide)
has_geo	: filter items that do or do not have geographical information (true, false)
coordinates	: latitude and longitude to get measurements
radius	: radius around area in meters **(default = 2500)**
value_from	: value threshold for parameter
value_to	: value threshold for parameter
date_from	: results from certain date
date_to	: results to certain date
order_by	: sort by **(default = date)**
sort	: sort by (desc, asc) **(default = asc)**
include_fields	: output has extra fields (attribution, averagingPeriod, sourceName)
limit	: number of results returned (max is 10000) **(default is 100)**
page	: paginate through results (max is 100) **(default is 1)**
format	: date return format type (csv, json) **(default is json)**

Python Libraries:

Requests

Pandas

Matplotlib.pyplot

Prepossessing Steps

To generate readings for the graphs, the API had to be limited to:

5 cities (Kansas City, Bakersfield, Fairbanks, Honolulu, Wilmington)

Data points generated throughout the month of November 2019

Data points containing PM2.5

To specify the dataset for this project, the following parameters were used:

city	: limited to specific cities
parameter	: limited to pm25 (particle matter under 2.5 micrometers)
date_from	: limited to start of certain month
date_to	: limited to end of certain month
order_by	: sorted by date
sort	: sorted by ascending
limit	: up to 10,000 results
format	: formatted to json

The remaining parameters create too much noise for the project.

Build

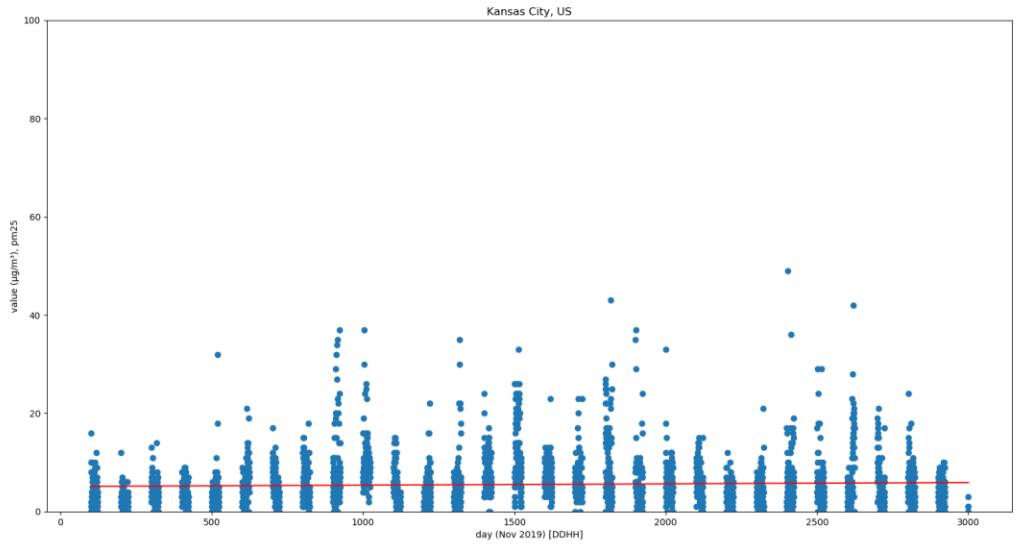
Using OpenAQ's API, the datasets obtained information for the 5 cities based on readings that occurred throughout the month of November 2019. With those datasets, graphs were created for each city in order to view their readings of PM2.5 over a course of a month.

To represent this data, scatterplots were created to visualize the information sets.

Graph Title	: City
Horizontal Axis	: day time of reading (DDHH)
Vertical Axis	: value of PM2.5 (units micrograms per meter cubed)

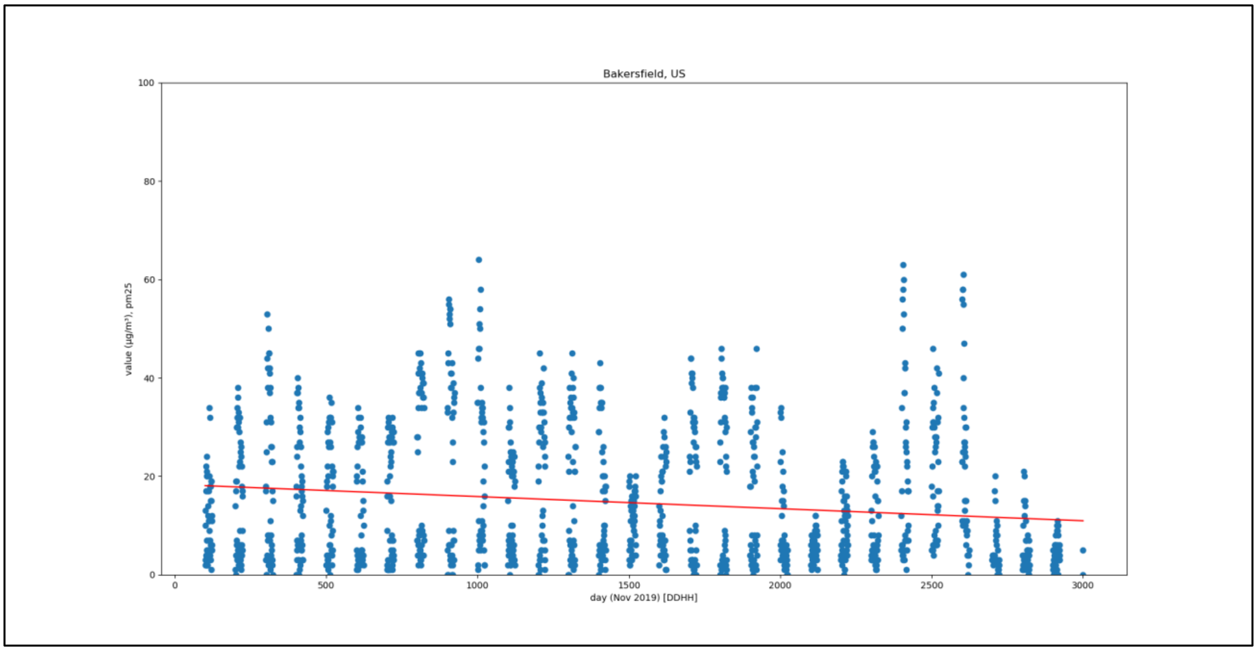
In order to occur a trend within the data sets, a line of best fit was developed.

Kansas City, Missouri Model: (3,676 data points) (Average = 5.53)

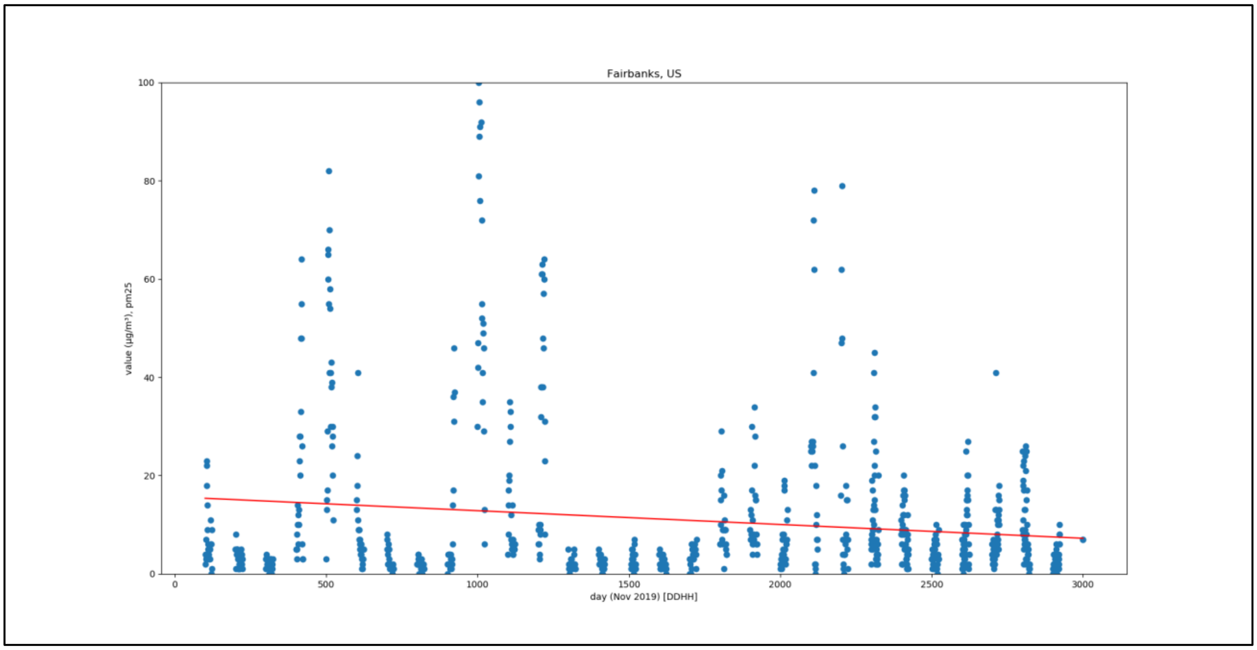


Polluted Cities:

Bakersfield, California Model: (1,356 data points) (Average = 14.66)

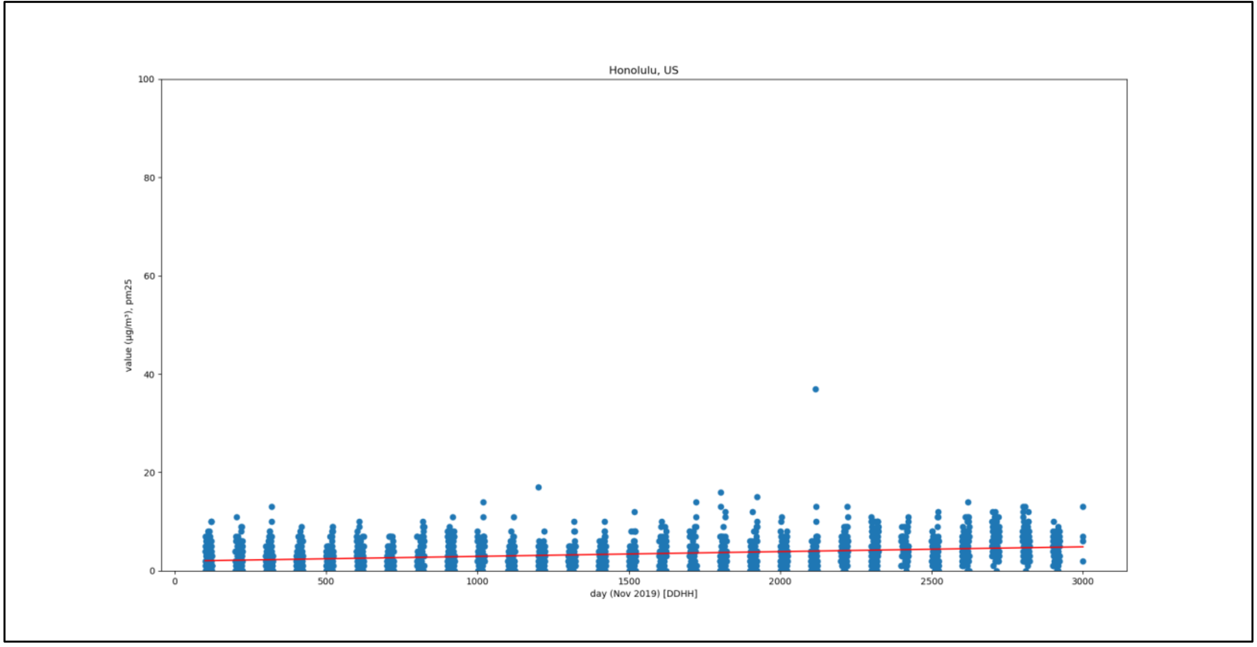


Fairbanks, Alaska Model: (835 data points) (Average = 10.82)

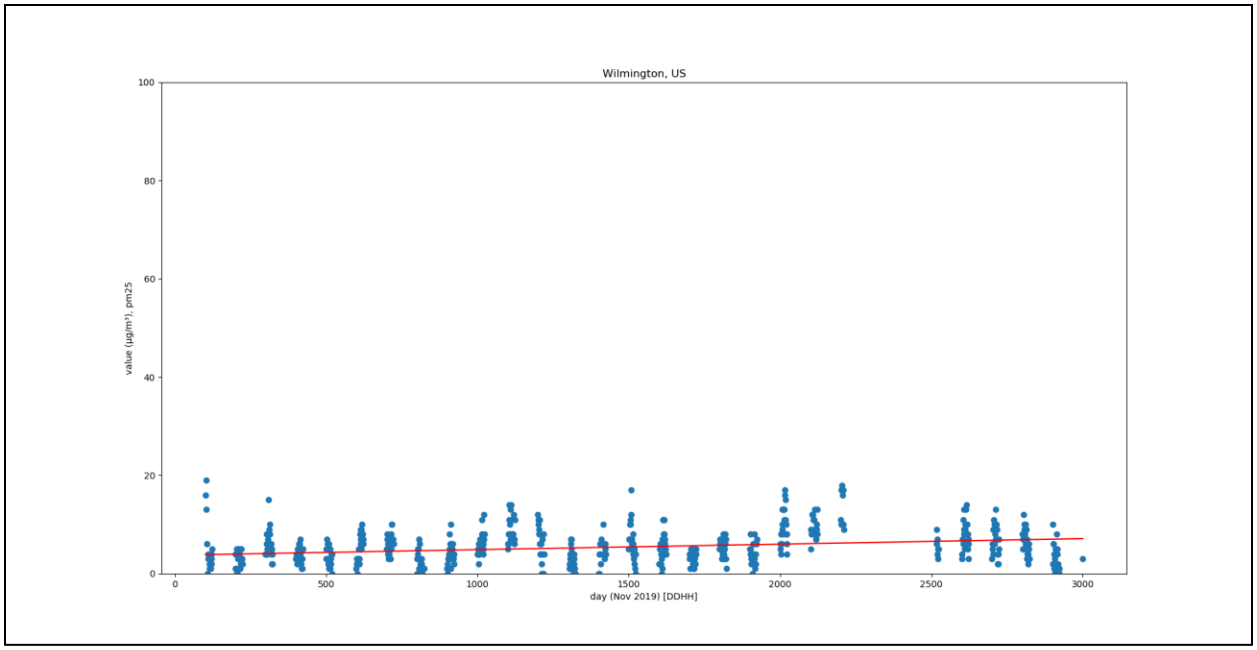


Cleaner Cities:

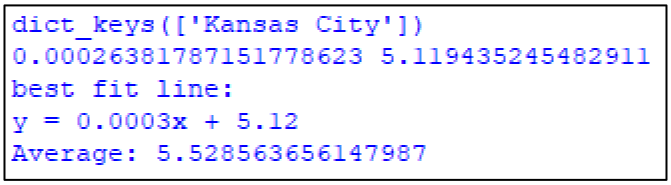
Honolulu, Hawaii Model: (2,705 data points) (Average = 3.44)



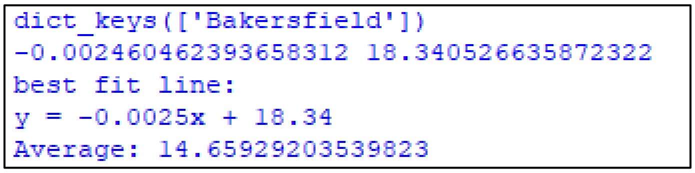
Wilmington, North Carolina Model: (604 data points) (Average = 5.34)



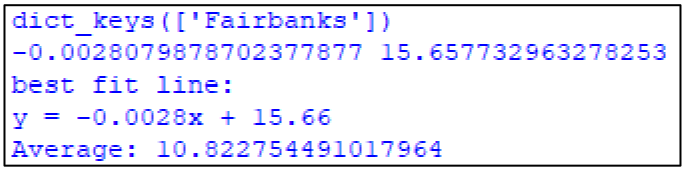
Results



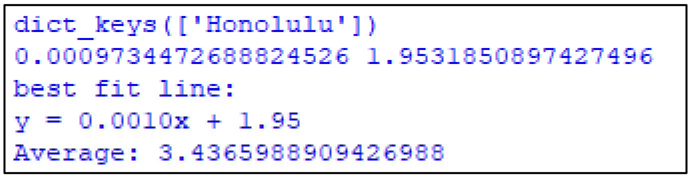
```
dict_keys(['Kansas City'])
0.0002638178715177862З 5.119435245482911
best fit line:
y = 0.0003x + 5.12
Average: 5.528563656147987
```



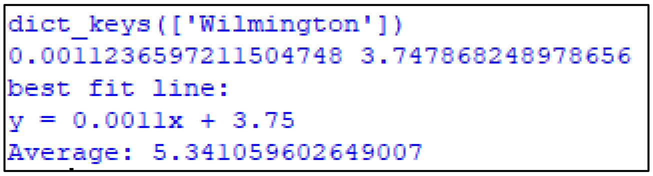
```
dict_keys(['Bakersfield'])
-0.002460462393658312 18.340526635872322
best fit line:
y = -0.0025x + 18.34
Average: 14.65929203539823
```



```
dict_keys(['Fairbanks'])
-0.0028079878702377877 15.657732963278253
best fit line:
y = -0.0028x + 15.66
Average: 10.822754491017964
```



```
dict_keys(['Honolulu'])
0.0009734472688824526 1.9531850897427496
best fit line:
y = 0.0010x + 1.95
Average: 3.4365988909426988
```



```
dict_keys(['Wilmington'])
0.0011236597211504748 3.747868248978656
best fit line:
y = 0.0011x + 3.75
Average: 5.341059602649007
```

Based on the graphs from the datasets, Kansas City, Honolulu, and Wilmington had a reading where PM2.5 level below the unhealthy limit of 35.5. While Bakersfield and Fairbanks had a reading where PM2.5 level above the unhealthy limit of 35.5. However, if we take the average of each city's value information, each city was below the unhealthy limit for the entire month.

Viewing the trends on each graph, the cities that are known to be "polluted" were decreasing in PM2.5 levels as the month went on, while the cities that are known to be "clean" were increasing in PM2.5 levels. This is probably due to the natural actions of air where the polluted particles tries to fit the entire space, which is the atmosphere.

Conclusion

Overall, Kansas City is relatively clean in terms of air pollution via PM2.5 levels. The data sets were for the month of November in 2019, so there might be other factors that could attribute to the PM2.5 levels.

An obvious factor could be population & population density (mi^2) specifically. Kansas City is a very populous city (491,918). Solely basing on population numbers, Kansas City should be compared to Bakersfield (383,579) and Honolulu (347,397). If based off density, then Kansas City (1,562.20) can be compared to Bakersfield (2,562.22) and Wilmington (2,372.97). Because these cities vary, air quality can only be said to be partially correlated by population and population density.

Another factor could be the climate, with Fairbanks being extremely cold (~10 °F) during November, and Honolulu being moderately hot (~80 °F) during November. To have a true comparison to Kansas City (~55 °F), it should be compared to relatively similar climates, such as Wilmington (~60 °F) and Bakersville (~65 °F). With Bakersville being a known polluted city and Wilmington being a known clean city, air quality levels can not be solely correlated to the city's climate.

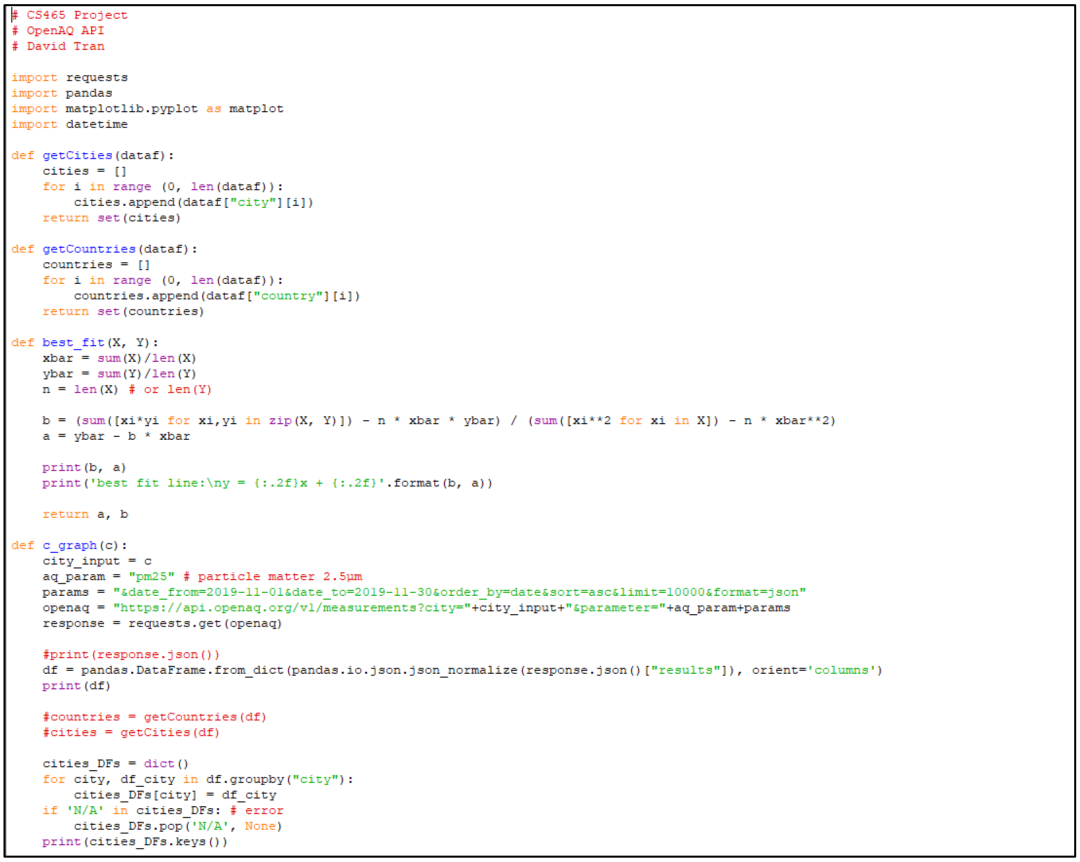
Elevation could be the most correlated point for air quality levels. The known clean cities have low elevation, Honolulu (19.69 ft) and Wilmington (30 ft).

For future comparisons, there are better ways to compare air quality levels. It is best to compare data sets for the entire year, to see if there is change based on month, which could relate to the climate. Data sets can be compared for only larger cities to see if there is a correlation to population and air quality levels. Based on apparent elevation correlation, data sets should be collected for cities with low vs high elevation.

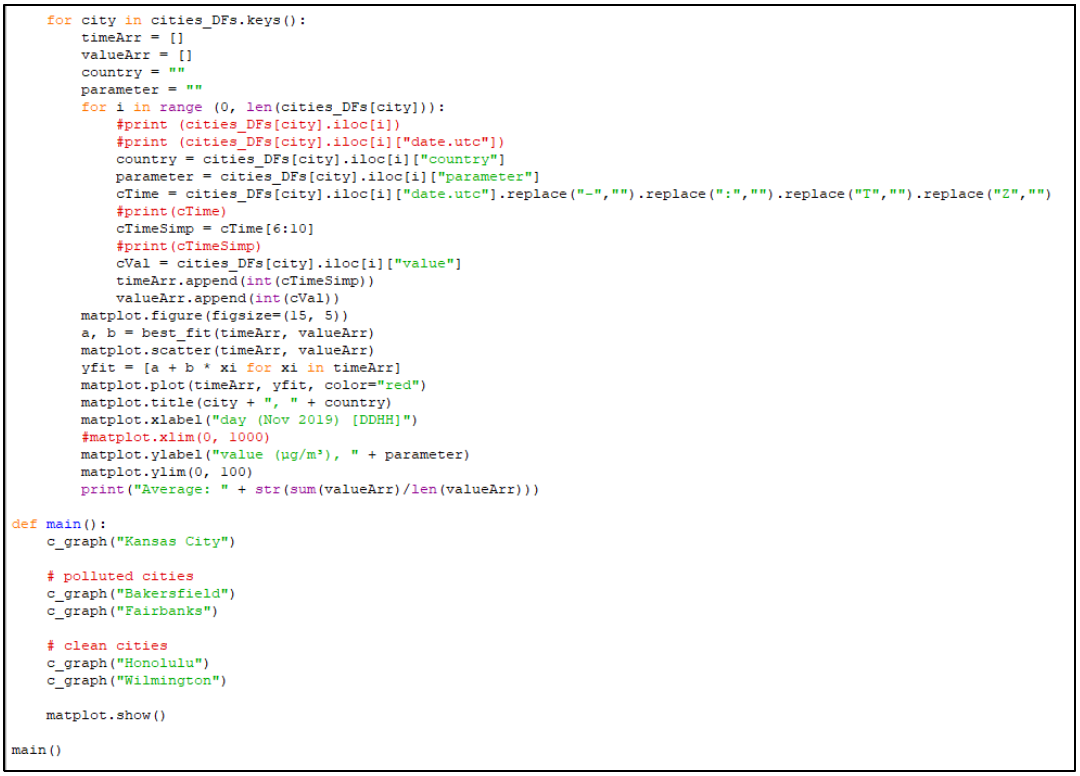
	Population	Pop/Density	Climate	Elevation	PM2.5
Kansas City	491,918	1,562.20 (mi^2)	54 / 36 (°F)	909′	5.53
Bakersfield	383,579	2,562.22 (mi^2)	68 / 44 (°F)	404′	14.66
Fairbanks	31,516	991.49 (mi^2)	10 / -4 (°F)	446′	10.82
Honolulu	347,397	5,737.79 (mi^2)	84 / 70 (°F)	19.69′	3.44
Wilmington	122,607	2,372.97 (mi^2)	64 / 51(°F)	30′	5.34

Climates are averages for November month.

Code



```python
# CS465 Project
# OpenAQ API
# David Tran

import requests
import pandas
import matplotlib.pyplot as matplot
import datetime

def getCities(dataf):
    cities = []
    for i in range (0, len(dataf)):
        cities.append(dataf["city"][i])
    return set(cities)

def getCountries(dataf):
    countries = []
    for i in range (0, len(dataf)):
        countries.append(dataf["country"][i])
    return set(countries)

def best_fit(X, Y):
    xbar = sum(X)/len(X)
    ybar = sum(Y)/len(Y)
    n = len(X) # or len(Y)

    b = (sum([xi*yi for xi,yi in zip(X, Y)]) - n * xbar * ybar) / (sum([xi**2 for xi in X]) - n * xbar**2)
    a = ybar - b * xbar

    print(b, a)
    print('best fit line:\ny = {:.2f}x + {:.2f}'.format(b, a))

    return a, b

def c_graph(c):
    city_input = c
    aq_param = "pm25" # particle matter 2.5µm
    params = "&date_from=2019-11-01&date_to=2019-11-30&order_by=date&sort=asc&limit=10000&format=json"
    openaq = "https://api.openaq.org/v1/measurements?city="+city_input+"&parameter="+aq_param+params
    response = requests.get(openaq)

    #print(response.json())
    df = pandas.DataFrame.from_dict(pandas.io.json.json_normalize(response.json()["results"]), orient='columns')
    print(df)

    #countries = getCountries(df)
    #cities = getCities(df)

    cities_DFs = dict()
    for city, df_city in df.groupby("city"):
        cities_DFs[city] = df_city
    if 'N/A' in cities_DFs: # error
        cities_DFs.pop('N/A', None)
    print(cities_DFs.keys())
```



```python
    for city in cities_DFs.keys():
        timeArr = []
        valueArr = []
        country = ""
        parameter = ""
        for i in range (0, len(cities_DFs[city])):
            #print (cities_DFs[city].iloc[i])
            #print (cities_DFs[city].iloc[i]["date.utc"])
            country = cities_DFs[city].iloc[i]["country"]
            parameter = cities_DFs[city].iloc[i]["parameter"]
            cTime = cities_DFs[city].iloc[i]["date.utc"].replace("-","").replace(":","").replace("T","").replace("Z","")
            #print(cTime)
            cTimeSimp = cTime[6:10]
            #print(cTimeSimp)
            cVal = cities_DFs[city].iloc[i]["value"]
            timeArr.append(int(cTimeSimp))
            valueArr.append(int(cVal))
        matplot.figure(figsize=(15, 5))
        a, b = best_fit(timeArr, valueArr)
        matplot.scatter(timeArr, valueArr)
        yfit = [a + b * xi for xi in timeArr]
        matplot.plot(timeArr, yfit, color="red")
        matplot.title(city + ", " + country)
        matplot.xlabel("day (Nov 2019) [DDHH]")
        #matplot.xlim(0, 1000)
        matplot.ylabel("value (µg/m³), " + parameter)
        matplot.ylim(0, 100)
        print("Average: " + str(sum(valueArr)/len(valueArr)))

def main():
    c_graph("Kansas City")

    # polluted cities
    c_graph("Bakersfield")
    c_graph("Fairbanks")

    # clean cities
    c_graph("Honolulu")
    c_graph("Wilmington")

    matplot.show()

main()
```

References

OpenAQ: https://openaq.org/

OpenAQ API: https://docs.openaq.org/#api-Measurements

Python Pandas Documentation: https://pandas.pydata.org/pandas-docs/stable/

Python Matplotlib Documentation: https://matplotlib.org/3.1.1/contents.html

PM2.5: https://blissair.com/what-is-pm-2-5.htm

EPA PM2.5 Trends: https://www.epa.gov/air-trends/particulate-matter-pm25-trends

EPA Particle Pollution: https://www.epa.gov/particle-pollution-designations

Top Polluted U.S. Cities: https://www.lung.org/our-initiatives/healthy-air/sota/city-rankings/most-polluted-cities.html

Top Populated U.S. Cities: https://www.nlc.org/the-30-most-populous-cities

Cities: https://www.marketwatch.com/story/four-in-10-americans-are-breathing-unsafe-air-and-these-8-cities-are-the-worst-2019-04-24

City Population Information: http://worldpopulationreview.com/us-cities/

Github Source Code: https://github.com/davidtstran/Statistical-Learning-OpenAQ