# Analysis of Pollution by Increase in Urban Population

DNSC 6211: Programming for Analytics

Suffyan Asad

**Abstract**

This study determines the relation between increase in Urban Population and increase in pollution. Two pollution indicators: PM 2.5 and $CO_2$ have been studied for this purpose in relation to % increase in Urban Population. This study tries investigates whether there is any direct relationship between the two by investigating the direct correlation between average of pollution indicators vs. increase in Urban Population, and then, the next step utilized in this study is two types of clustering: K-Means Clustering and Agglomerative Hierarchical Clustering to study whether there are any groups of countries based on how they are handling pollution.

# Contents

# 1 Introduction

The objective of this project is to investigate the relation of increasing pollution with the increase in population. The project investigates the relation of pollution indicators: $CO_2$ Emissions and PM 2.5 with increase in Urban Population. The objective is to determine whether the increase in Urban Population, that is prevalent in many developing countries of th world, is related to increase in pollution or not. In this study, two primary indicators of pollution: $CO_2$ Emissions and PM 2.5 are related with the increase in Urban Population over the last 2 decades. The project tries to determine the relation of this increase in pollution in relation with the growth in Urban Population.

# 2 Background

The study begins by downloading the $CO_2$ Emissions, Percentage of population living in areas with PM 2.5 above the UN guidelines, and Annual %age Increase in Urban Population. The used in the analysis is obtained from World Bank. The objective of this project is to determine whether the increase in population is related to increase in pollution.

The rise of many developing countries in Asia and Africa has been happening since the 1990s, and much of this development has been taking place in urban areas of most of these developing counties instead of all over the country, attracting many people to migrate to the urban areas from rural areas all over the country in search of better education, healthcare facilities, and better job and work opportunities.

Another phenomenon prevalent since the 1990s that is being discussed at many levels including the media and is the cause of debate is the increasing pollution and its effects, including global warming, specially in developing countries.

Therefore, this study using the World Bank Data on pollution and urban population, tries to determine if this is really the case and that pollution is increasing with the increase in Urban Population or not. And if there is any way to group countries according to their performance in handling pollution with increasing population.

# 3 Method

The overall question answered in the study, as mentioned before, is whether Increase in Pollution is related to Increase in Urban Population.
This study, investigates the following questions to determine this relationship:

- Finding a relation between pollution and urban population
- Finding a connection between increase in urban population and pollution
- Identify different types of countries based whether they are managing pollution well with increase in urban population or not.

Additionally, this analysis takes a look at only two pollution indicators i.e. $CO_2$ and %age of population living in areas with PM 2.5 above UN Guidelines. The other pollution indicators are not investigated.
Secondly, the data investigated is only of years 2013, 2011, 2010, 2005, 2000, 1995, 1990, and

the decades of 1980s and 1970s are not investigated because of lack of data in one or more of the indicators.

## 3.1 Workflow

The workflow of the analysis can be explained in the diagram as follows.

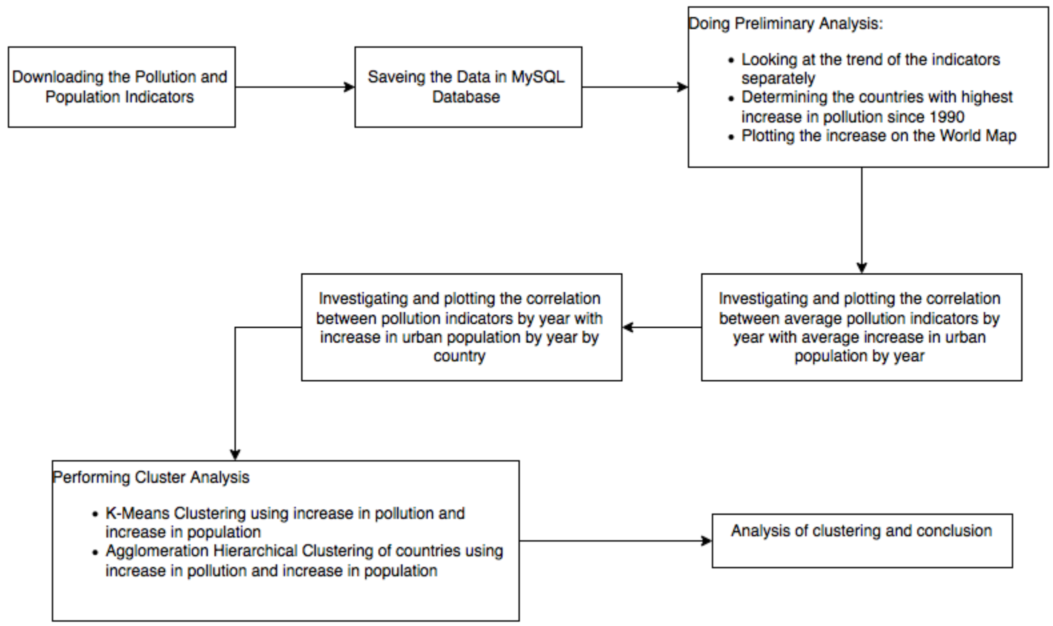


Figure 1: The project workflow

The study has been accomplished in following steps:

1. The first step is to download the Datasets from World Bank data source. Pollution and population indicators are downloaded.
2. Next step is to save the downloaded data in a MySQL database called ProjectDb so that it can be used in analysis.
3. The third step is performing preliminary analysis on average values of pollution indicators with the change in population to investigate if there is a general trend between in the change in urban population and in average pollution levels in the world.

   Additionally, change in each of the three indicators is determined from 1990 to 2013 (2011 in case of $CO_2$ as 2013 data is not available) to see which countries have the highest increase. Finally, this change is plotted on the world map to have an overall picture of the change.
4. Fourth step is to determine the correlation between the average pollution indicators by year with average increase in Urban Population to determine if there any positive relation between them.

5. Next, correlation between change in pollution and change in Urban Population is determined for each country, and see highest and lowest correlations, and plot it on the map.
6. Next, clustering is performed to determine clusters of countries by Change in Population and Change in Pollution to determine which groups are doing well and which groups of counties are not handling pollution properly. Two types of clustering is performed:
   - K-Means Clustering (performed using Python)
   - Agglomerative Hierarchical Clustering (performed using R)
7. Final step of the Analysis is to interpret the results of the correlation, and to plot it on the world map to visualize the groups.

All of this analysis is performed in a Python Notebook called "Suffyan_Asad_IndividualProjectNotebook" that states and executes all the steps.

Additionally, following Python and R script files and other files are used in analysis:

1. Correlation.py
2. Clustering.py
3. PlotDiffOnMap.py
4. PlotClustersOnMap.py
5. SaveSetInDb.py
6. HierarchicalClustering.R
7. WorldLow.svg (SVG file containing the world map used by Python scripts to plot on)

These files are required to be in the same folder as the notebook for the notebook to successfully execute.

## 3.2 Project structure

Following pollution indicators are used in the analysis:

1. **PM 2.5 Air Pollution Indicator Code:** EN.ATM.PM25.MC.M3
   **Years:** 2013, 2011, 2010, 2005, 2000, 1995
   Source of Data

   **Description:** The chosen PM 2.5 indicator indicates the percentage of population that lives in areas with PM 2.5 amounts beyond the guidelines mentioned by WHO. The source describes the measure as:
   "Percent of population exposed to ambient concentrations of PM2.5 that exceed the WHO guideline value is defined as the portion of a countrys population living in places where mean annual concentrations of PM2.5 are greater than 10 micrograms per cubic meter, the guideline value recommended by the World Health Organization as the lower end of the range of concentrations over which adverse health effects due to PM2.5 exposure have been observed." Source: PM2.5 air pollution, population exposed to levels exceeding WHO guideline value (% of total)

   I have chosen this measure since it directly indicates the people affected by pollution, since this measure is measuring the proportion of population exposed to high PM 2.5 pollution, and therefore it will give an indication of the proportion of population living with pollution.

2. $CO_2$ **Emissions (Metric Tons per capita) Indicator Code:** EN.ATM.CO2E.PC
   **Years:** 2011, 2010, 2005, 2000, 1995
   Source of Data

   **Description:** The source describes $CO_2$ Emissions indicator as:
   "Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring." Source: $CO_2$ Emissions Metric Tons Per Capital

   $CO_2$ emissions is one if the primary indicators of pollution, and it directly indicates the level of emissions from traffic and other sources present in an area (country in our case).

Next, to measure growth of urban population, following indicator has been chosen:

**Urban Population % of total**
**Indicator Code:** SP.URB.TOTL.IN.ZS
**Years:** 2013, 2011, 2010, 2005, 2000, 1995
Source of Data

**Description:** This measure gives the proportion of a country's population living in urban areas, as urban areas are generally associated with pollution, it will allow to see if population is linked with growth in urban population. The source describes it as:

"Urban population refers to people living in urban areas as defined by national statistical offices. It is calculated using World Bank population estimates and urban ratios from the United Nations World Urbanization Prospects."Source: Urban population (% of total)

## 3.3 Figures and Tables

Following is the basic description of tables and figures included in the report.

### 3.3.1 Description of Figures

| | Figure | Explanation |
|---|---|---|
| 1 | Figure 1 | Project workflow and steps of conducting the study |
| 2 | Figure 2 | PM 2.5 Trend Graph |
| 3 | Figure 3 | PM 2.5 change by country |
| 4 | Figure 4 | $CO_2$ Trend Graph |
| 5 | Figure 5 | $CO_2$ change by country |
| 6 | Figure 6 | Urban Population Trend Graph |
| 7 | Figure 7 | Urban Population Increase by Country |
| 8 | Figure 8 | Correlation - Average PM 2.5 and Average of Percent Urban Population |
| 9 | Figure 9 | Correlation - Average $CO_2$ and Average of Percent Urban Population |
| 10 | Figure 10 | Country-wise correlation between PM 2.5 and percent change in Urban Population |
| 11 | Figure 11 | Country-wise correlation between $CO_2$ and percent change in Urban Population |
| 12 | Figure 12 | Result of K-Means - Statistics of groups obtained by clustering |
| 13 | Figure 13 | K-Means clusters plotted on world map |
| 14 | Figure 14 | Agglomerative Hierarchical Clustering plotted on map |

Table A: Description of Figures used in report

### 3.3.2 Description of Tables

| | Table | Explanation |
|---|---|---|
| 1 | Table A | Description of Figures used in report |
| 2 | Table B | Description of Tables used in report |
| 3 | Table 1 | Average PM 2.5 Values by Year |
| 4 | Table 2 | $CO_2$ trend over the years |
| 5 | Table 3 | Urban Population trend over the years |
| 6 | Table 4 | Cluster of Countries handling pollution well with summary of indicators |
| 7 | Table 5 | Cluster of Countries not handling pollution well with summary of indicators |
| 8 | Table 6 | Clusters summary obtained by Agglomerative Hierarchical Clustering |
| 9 | Table 7 | Cluster of Countries not handling pollution well |
| 10 | Table 8 | Cluster Countries handling pollution well |

Table B: Description of Tables used in report

# 4 Discussion

## 4.1 Overall Trends

Initially, the overall trends are determined to have the sense of data.

### 4.1.1 Trend of PM 2.5 (1990 - 2013)

Average values of PM 2.5 is following the following trend:

| Year | PM 2.5 |
|---|---|
| 1990 | 52.1002924528 |
| 1995 | 53.4971273585 |
| 2000 | 54.7943113208 |
| 2005 | 56.2235566038 |
| 2010 | 57.6314150943 |
| 2011 | 57.9140801887 |
| 2013 | 58.4807877358 |

Table 1: Average PM 2.5 Values by Year

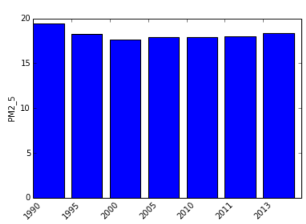


Figure 2: PM 2.5 Trend Graph

It can be observed that PM 2.5 values are higher in 1990, and then they drop and then again increase in recent years, but not as high as it was in 1990.

It can be seen in the following map that PM 2.5 has increased the most in the developing countries that have seen massive industrial development in 1990s and 2000s.

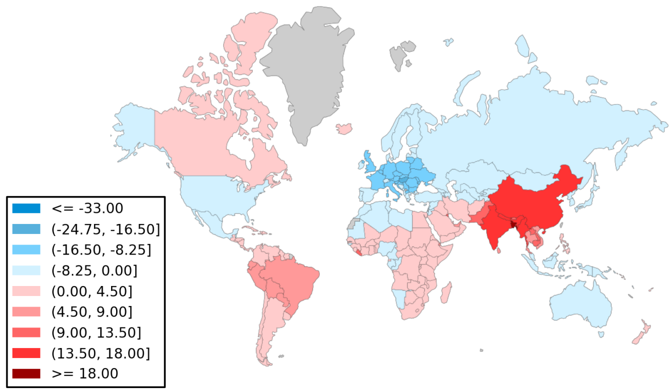


Figure 3: PM 2.5 change by country

#### 4.1.2 Trend of $CO_2$ (1990 - 2011)

Average values of $CO_2$ is following the following trend:

| Year | $CO_2$ |
|---|---|
| 1990 | 4.36052141128 |
| 1995 | 4.59247870215 |
| 2000 | 4.69128418694 |
| 2005 | 4.85424021814 |
| 2010 | 4.86414743968 |
| 2011 | 4.80523752319 |

Table 2: $CO_2$ trend over the years

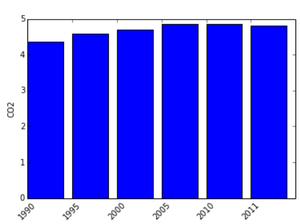

Figure 4: $CO_2$ Trend Graph

It can be observed that $CO_2$ values are gradually increasing from 1990 to 2005, and then they level off and drop slightly in 2011. The data is not available for 2013.

It can be seen in the following map that $CO_2$ has increased the most in the developing countries that have seen massive industrial development in 1990s and 2000s.

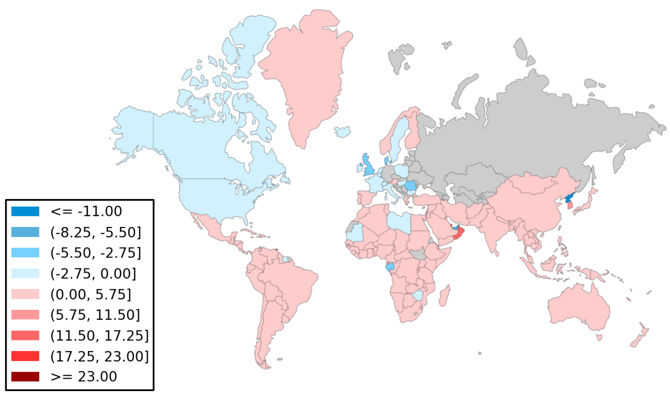


Figure 5: $CO_2$ change by country

#### 4.1.3 Trend of Urban Population (1990 - 2013)

Next is to observe the trend of %age increase in Urban Population:

| Year | $Urbanpopulation\%increase$ |
|---|---|
| 1990 | 52.1002924528 |
| 1995 | 53.4971273585 |
| 2000 | 54.7943113208 |
| 2005 | 56.2235566038 |
| 2010 | 57.6314150943 |
| 2011 | 57.9140801887 |
| 2013 | 58.4807877358 |

Table 3: Urban Population trend over the years

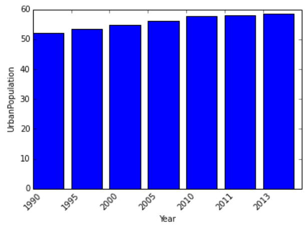


Figure 6: Urban Population Trend Graph

Average of Percent change in Urban Population has remained positive and increasing over the years from 1990 to 2013. It is clear from the growth in urban population is on an increasing trend.

It can be seen in the following map that Urban Population has incraesed all over the world, specially in the developing countries of Asia and Africa.

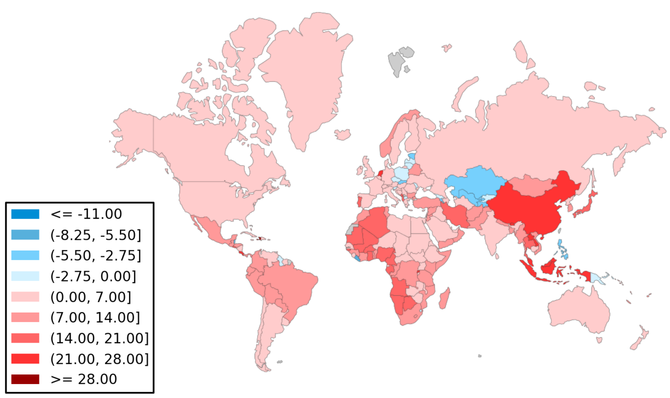


Figure 7: Urban Population Increase by Country

## 4.2 Analysis: Correlation

Now, after looking at the general trends, next step is to determine the correlations between the indicators to see if increase in Population and increase in Pollution go together or not. Initially, the correlation is determined between the average values of pollution indicators and urban population.

### 4.2.1 Correlation of Average Urban Population Increase with Average Pollution indicators

First, following is the plot of average PM 2.5 pollution levels and average of percent change of urban population:

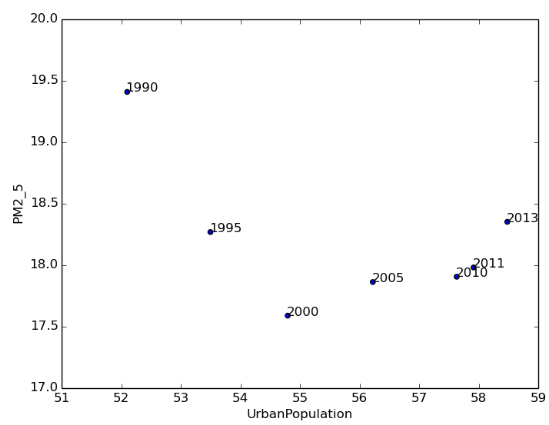


Figure 8: Correlation - Average PM 2.5 and Average of Percent Urban Population

The correlation between Average PM 2.5 and Average of Percent Urban Population is -0.551630, which is quite strong negative relation.

Next step is to determine the correlation between $CO_2$ value and average of percent change in Urban Population. Following is the graph between the average values of these two indicators:

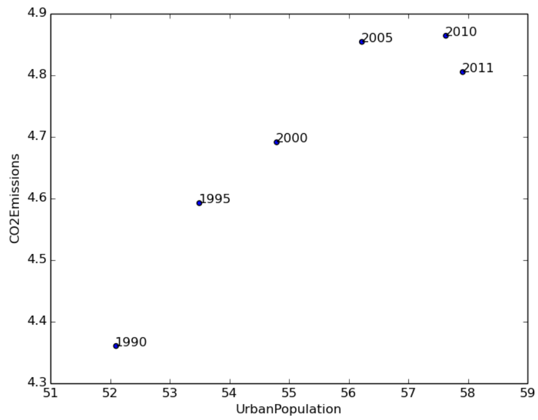


Figure 9: Correlation - Average $CO_2$ and Average of Percent Urban Population

The correlation is 0.93 which is a strong positive correlation indicating that $CO_2$ is related with Urban Population.
Due to the mixed results that have been obtained in the previous two correlation values, separate country wise-analysis is required, and therefore, the next step is to calculate the correlation by country. The results are mixed, meaning there are some very high positive correlations and some negative correlations in case of both computations, and it can be best observed when plotted on the world map:

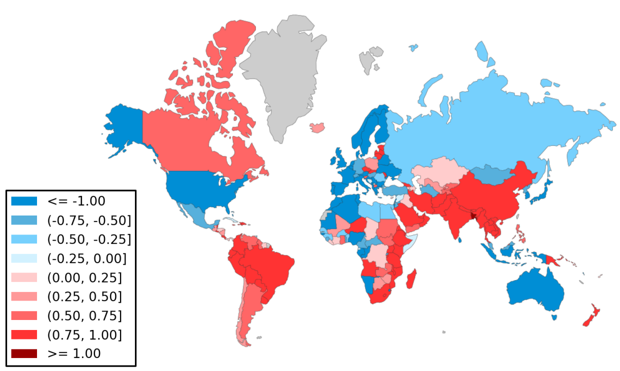


Figure 10: Country-wise correlation between PM 2.5 and percent change in Urban Population

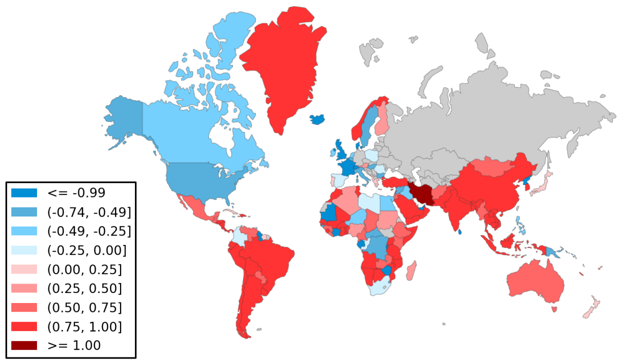


Figure 11: Country-wise correlation between $CO_2$ and percent change in Urban Population

From the figures it can be observed that the correlations are all over the range, both positive and negative extreme values are present, and this indicates that there is no general trend in increase in pollution and increase in population, rather, this result indicates that there are groups of countries, some are doing well, and some are not doing well. Therefore, I have decided to do cluster analysis to determine these groups and to understand the characteristics of each group separately.

## 4.3 Cluster Analysis

As stated before, the correlations obtained, specially country wise correlation between change in Urban Population and pollution indicators indicate that there is no global trend, rather there are countries with positive correlations and countries with negative correlations, meaning some countries are not handling pollution and some counties are handling the pollution well despite increase in urban population. The objective of cluster analysis is to identify these groups by grouping countries into groups based on their handling of pollution with increase in urban population. The urban population has been increasing in almost all the countries according to the datasets obtained from World Bank.

Following types of cluster analysis will be performed:

1. K-Means Clustering (Using Python)
2. Agglomerative Hierarchical Clustering (Using R)

Clustering is performed between increase in each indicator from 1990 to 2011 (2013 data is not present for $CO_2$ data).

### 4.3.1 K-Means Clustering

First step is to create a table called IndicatorChanges using the following SQL Query:

```
%%sql

Create Table IndicatorChanges
select
        P.country,
        P.`2011` - P.`1990` as PM2_5,
        C.CO2Emissions,
        U.UrbanPopulation
from PM2_5 as P
join
(
    select country, `2011` -`1990` as CO2Emissions
    from CO2Emissions
) as C on C.country = P.country
join
(
    select country, `2011` -`1990` as UrbanPopulation
    from UrbanPopulation
) as U on U.country = P.country;
```

Next, a column is added to save the result of clustering:

```
%%sql

alter table IndicatorChanges add Cluster INT;
```

Next, the Python script Clustering.py is invoked to perform the clustering, with k = 5 to form 5 clusters. Following clusters are obtained:

Cluster 0

| | PM2_5 | CO2Emissions | UrbanPopulation |
|---|---|---|---|
| count | 29.000000 | 29.000000 | 29.000000 |
| mean | -0.679001 | 0.695334 | 17.650172 |
| std | 5.372552 | 1.655707 | 3.131679 |
| min | -10.989098 | -3.668745 | 13.252000 |
| 25% | -2.948012 | 0.033112 | 14.981000 |
| 50% | -1.208376 | 0.433858 | 16.904000 |
| 75% | 1.432348 | 0.898067 | 19.197000 |
| max | 14.773588 | 4.791223 | 24.946000 |

Cluster 1

| | PM2_5 | CO2Emissions | UrbanPopulation |
|---|---|---|---|
| count | 10.000000 | 10.000000 | 10.000000 |
| mean | 7.906669 | 7.578184 | 5.945000 |
| std | 6.128856 | 8.510965 | 3.578763 |
| min | -0.586493 | 0.098500 | 0.305000 |
| 25% | 3.293005 | 0.386503 | 4.872750 |
| 50% | 7.913160 | 4.694099 | 5.877500 |
| 75% | 12.900047 | 12.587213 | 8.075250 |
| max | 16.576868 | 23.260179 | 11.414000 |

Cluster 2

| | PM2_5 | CO2Emissions | UrbanPopulation |
|---|---|---|---|
| count | 16.000000 | 16.000000 | 16.000000 |
| mean | -10.530099 | -3.386976 | 2.786813 |
| std | 8.517578 | 3.288317 | 2.542496 |
| min | -36.055413 | -11.085076 | -0.490000 |
| 25% | -12.401407 | -3.302253 | 0.686500 |
| 50% | -10.258090 | -2.139497 | 2.023000 |
| 75% | -7.381247 | -1.622569 | 4.727000 |
| max | 5.228208 | 0.251961 | 7.959000 |

Cluster 3

| | PM2_5 | CO2Emissions | UrbanPopulation |
|---|---|---|---|
| count | 48.000000 | 48.000000 | 48.000000 |
| mean | 0.619113 | 0.609425 | 8.670833 |
| std | 2.739489 | 1.020108 | 2.234301 |
| min | -6.408519 | -0.270755 | 4.421000 |
| 25% | -0.916124 | 0.060387 | 6.991000 |
| 50% | 0.208936 | 0.243316 | 8.508000 |
| 75% | 1.423074 | 0.796374 | 10.627500 |
| max | 7.902476 | 6.080383 | 12.690000 |

Cluster 4

| | PM2_5 | CO2Emissions | UrbanPopulation |
|---|---|---|---|
| count | 51.000000 | 51.000000 | 51.000000 |
| mean | -0.083613 | 0.070535 | 1.231176 |
| std | 2.590850 | 1.569762 | 3.596638 |
| min | -6.584674 | -7.591477 | -10.899000 |
| 25% | -1.126181 | -0.083966 | -0.367500 |
| 50% | 0.063758 | 0.064691 | 2.110000 |
| 75% | 1.107686 | 0.719827 | 4.065000 |
| max | 6.842911 | 5.211749 | 5.305000 |

Figure 12: K-Means clustering result

Note: K-means clustering is non-deterministic, and initial means are random, therefore running the notebook again can give different results, but the overall properties of the clusters should be similar.

### 4.3.2 Results of K-Means Clustering

The countries are divided into 5 clusters (1 - 4). The previous image shows that there are two extremes: Countries that are handling pollution well (low increase or decrease in pollution), and countries that are not handling it well. Following is the cluster of the countries that are handling pollutoin well:

| Country | Cluster | PM 2.5 | $CO_2$ Emissions | Urban Pollution |
|---|---|---|---|---|
| Austria | 2 | -10.7605556406 | 0.251961190007 | 0.093 |
| Belgium | 2 | -9.7556246461 | -1.79027337047 | 1.31 |
| Bulgaria | 2 | -11.4826265523 | -1.97584179905 | 6.261 |
| Denmark | 2 | -7.63659420446 | -2.5223699818 | 2.114 |
| France | 2 | -8.08638942137 | -1.23462476222 | 4.528 |
| Hungary | 2 | -16.9885233627 | -1.84747936928 | 3.51 |
| Italy | 2 | -11.1634756196 | -0.659146968248 | 1.718 |
| Korea, Dem. Rep. | 2 | -2.66917277018 | -9.13673078956 | 1.932 |
| Luxembourg | 2 | -12.1236444186 | -5.29984706896 | 7.959 |
| Poland | 2 | -13.2346957156 | -1.33071189841 | -0.49 |
| Romania | 2 | -13.5478375395 | -2.63638857609 | 0.743 |
| Singapore | 2 | -36.0554128132 | -11.0850755009 | 0.0 |
| Switzerland | 2 | -6.6152035967 | -1.71985524368 | 0.517 |
| United Arab Emirates | 2 | 5.22820754467 | -8.27732267232 | 5.324 |
| United Kingdom | 2 | -8.57985600699 | -2.62476526141 | 3.43 |
| United States | 2 | -5.0101810955 | -2.30315182879 | 5.64 |

Table 4: Countries handling pollution well

It can be observed that the countries handling pollution well are mostly rich and developed countries, and the increase in Urban Population is also not very high.

Next, lets look at the countries that are not managing pollution well:

| Country | Cluster | PM 2.5 | $CO_2$ Emissions | Urban Pollution |
|---|---|---|---|---|
| Bangladesh | 1 | 16.5768682069 | 0.225452225545 | 11.414 |
| Equatorial Guinea | 1 | -0.586492734815 | 8.58656621843 | 4.592 |
| India | 1 | 14.1910210276 | 0.869655741947 | 5.729 |
| Kuwait | 1 | 10.8813935281 | 4.63657758497 | 0.305 |
| Myanmar | 1 | 11.8877738341 | 0.0985000204245 | 7.365 |
| Nepal | 1 | 13.2374717692 | 0.12562526247 | 8.312 |
| Oman | 1 | 2.98118960103 | 13.9207620731 | 9.589 |
| Qatar | 1 | 4.94492673762 | 19.3068997887 | 6.026 |
| Saudi Arabia | 1 | 4.22845145155 | 4.7516200435 | 5.715 |
| Trinidad and Tobago | 1 | 0.724089594246 | 23.2601790623 | 0.403 |

Table 5: Countries not handling pollution well

These countries not handling pollution well are the developing countries of Asia and Middle East with huge increase in Urban Population, and the countries have gone through a lot of development recently, and clearly do not have budget and infrastructure to control pollution that comes with it.

Following is the plot of clusters on the world map.

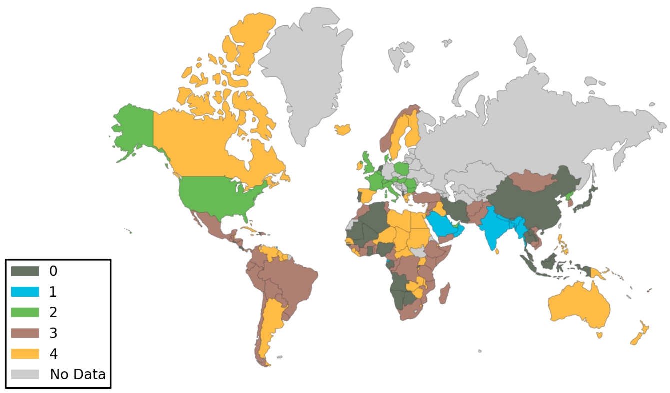


Figure 13: K-Means clusters plotted on map

#### 4.3.3 Agglomerative Hierarchical Clustering (Using R)

Next step is to perform using Agglomerative Hierarchical Clustering using R. R is used at this step since it has many clustering algorithms built in and many more analytics methods can be added using easily available libraries and packages.

The R Script HierarchicalClustering.R is used to perform the clustering and to store the results in the database. It requires following R packages:

1. RWorldMap (install.packages("rworldmap")
2. RMySQL

#### 4.3.4 Results of Agglomerative Hierarchical Clustering

Following is the summary of the clusters made by Agglomerative Hierarchical Clustering:

| Cluster | Avg Pm 2.5 | Avg $CO_2$ Emissions | Avg Urban Population |
|---|---|---|---|
| 1 | 0.700690629003 | 0.371388330969 | 8.67672727273 |
| 2 | -1.75862335336 | 0.637997568066 | 16.1741212121 |
| 3 | 0.789083314833 | 0.263026682989 | 1.80551666667 |
| 4 | -10.6870373354 | -2.08314492942 | 3.083 |
| 5 | 9.62806571961 | 6.3009869451 | 11.0981 |

Table 6: Clusters summary obtained by Agglomerative Hierarchical Clustering

It can be observed that cluster 4 is the countries that are handling pollution well, and cluster 5 is the countries that are not handling pollution well.

Agglomerative Hierarchical clustering identifies the following countries that are facing increased pollution:

| Country |
|---|
| Bangladesh |
| Bhutan |
| China |
| India |
| Lao PDR |
| Myanmar |
| Nepal |
| Oman |
| Qatar |
| Trinidad and Tobago |

Table 7: Countries not handling pollution well

Following are the countries that are handling pollution well, and again, they are the developing countries that have seen a low of growth and industrial development in 1990s and 2000s.

| Country |
|---|
| Austria |
| Belgium |
| Bulgaria |
| Denmark |
| France |
| Greece |
| Hungary |
| Ireland |
| Italy |
| Luxembourg |
| Malta |
| Poland |
| Romania |
| Singapore |
| Spain |
| Switzerland |
| United Kingdom |
| United States |

Table 8: Countries handling pollution well

And finally, the clusters plotted on the world map, using the RWorldMap package:

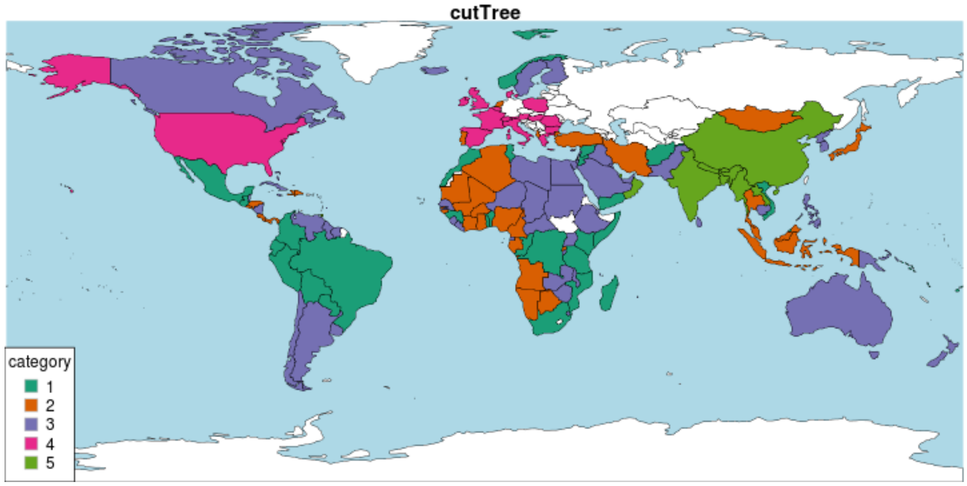


Figure 14: Agglomerative Hierarchical Clustering plotted on map

## 4.4 Learnings

This project was very valuable because it allowed me to explore many things in R and specially Python. The technological learning was immense, and beyond the books and tutorials, it allowed me to understand many subtleties of Python and R, and made me familiar with many libraries, specially mapping and charting.

### 4.4.1 Technical: Python and R

Following are the new items I have been introduced to:

1. Connecting Python with MySQL Database to write and retrieve data.
2. Using Numpy DataFrame to accomplish different tasks, manipulating them, loading DataFrame from MySQL Database, writing DataFrame to MySQL tables etc.
3. Editing SVG files of maps using Beautiful Soup to plot colors on a world map.
4. Creating Bar, Scatter and other charts in Python Matplotlib, changing size of the plots, adding and modifying legends on plots.
5. Connecting R with MySQL Database to read and write data.
6. Creating project report using LaTeX .

### 4.4.2 Analytics and Other items

Following are the items related to Analytics, writing and other skills I came across and practiced while working on the project:

1. Writing Project Reports, organizing reports, managing charts and tables, combining them to explain the results of the project.

2. Types of data available from World Bank data and got the glimpse of how this data can be used to make interesting studies on different items in many fields.
3. K-Means Clustering and Agglomerative Hierarchical Clustering.
4. Correlation

### 4.5 Challenges

Following were the technical challenges that I had to solve while working on this project, and the things that I wanted to do but could not do so.

1. Working with Python DataFrames, which was new to me, specially altering columns, groupings, saving/loading to MySQL Database.
2. Working with R to create world maps was new and difficult initially.
3. I wanted to include more pollution indicators but for many indicators the data was incomplete, or less data was available, otherwise I wanted to include them as well to get a better picture.
4. Old data was not available, otherwise a better long term picture of the way countries (or groups of countries) are handling pollution. I was interested to include the data from 1980s and 1970s but it is not available.
5. Other techniques of identifying groups in addition to two types of clustering used could be included.

## 5 Conclusion

The study of pollution vs. Urban Population increase indicates the fact the increase of pollution is not linear, PM 2.5 was higher in 1990s than 2013, whereas $CO_2$ levels have been increasing but this increase has been getting slower recently. Next, there is no general trend in increase in Urban Population and increase in pollution, rather, there are groups of countries that have various levels of success in handling pollution with the increase in urban population, which has been increasing in most countries since 1990s. Mostly the first-world countries, countries with large budgets, developed infrastructure and moderate increase in Urban Population have been handling pollution well, with low or negative pollution increase. On the other hand, the developing countries of Asia and Africa that have seen rapid development and industrialization, and rapid urbanization in the past 2-3 decades have not been able to handle pollution properly and the pollution levels are increasing in these countries with increase in population. K-Means and Agglomerative Hierarchical Clustering have both pointed at similar results.

## 6 References

Following tutorials, guides and examples on the internet were helpful in implementing this project:

1. Svg map of world
2. Creating and editing maps with python
3. Saving SVG as PNG in Python
4. Matplotlib Documentation
5. Matplotlib Bar charts

6. Matplotlib Scatter plots
7. Matplotlib Legends Example
8. RWorldMap
9. RWorldMap FAQ
10. R Colors Cheatsheet (Color Brewer paletts)