**Battle of the Neighborhoods**

**An Exploratory Study on Air Pollution in Urban Areas**

**By**

**Md. Saiful Bari Siddiqui**

6thJanuary, 2021

1. **Introduction**

**1.1 Background**

Air Pollution is a burning problem nowadays, especially in urban areas. So It is necessary to examine what is responsible for poor air quality. A good way to do that is to determine what differentiates the cities with Least polluted air from those with Most polluted air. In this project we've used population data and the internal surroundings of 16 such cities to determine what the difference is between the two groups.

**1.2 Target Audience**

This study pretty much concerns every one, Because of the fact that everyone consumes air and is affected by it. **But the primary audience of this study is the City authorities and Urban Planners who are responsible for designing and Planning the structure of a city**, so that they have an insight on what could possibly lead to a City which is Pollution free.

1. **Problem Statement**

The core question that we want to answer is **What is responsible for poor air quality in many Cities in the World?** But in this study, we will attempt to address 2 more specific Problems. These are:

1. **Do the internal structure and surroundings of these cities have a significant impact on Air Pollution? If it does, what kind of structures in a city correlates to the city being more polluted?**
2. **What role does population have in case of Pollution? Does Higher Population Density mean More Pollution?**
3. **Data**

To understand and explore the following Open Data were used:

1. List of most-polluted cities by particulate matter concentration: <https://en.wikipedia.org/wiki/List_of_most-polluted_cities_by_particulate_matter_concentration>
2. A Comparative list of the Most and Least Polluted Cities in the World: <https://www.rd.com/list/most-and-the-least-polluted-cities/>
3. Foursquare Developers Access to venue data: <https://foursquare.com/>
4. Population Data from Wikipedia: <https://en.wikipedia.org/wiki/List_of_cities_proper_by_population_density>
5. Latitude & Longitude of Cities using geopy library: <https://github.com/geopy/geopy>

Using this data will allow exploration and examination to answer the questions. The venue data will be used to properly determine the common internal surroundings of each city and determine if there is some kind of correlation to pollution. The Population data will be used to compare the population density to examine if cities with higher pollution are also more densely populated. The lists of most polluted and least polluted cities will be used as reference. From these, we will use the 8 most polluted and 8 least polluted cities to compare the common venues within the cities retrieved from Foursquare location data and the population densities of the cities got from Wikipedia and other sources. The geopy library was used to extract the latitude & longitude values of the cities.

## Methodology

### **4.1 Analytic Approach**

There are 2 different stages in our Analysis. In the 1st stage we compare the common venues within the most polluted and least polluted cities and determine if these surroundings have any effect on pollution. In the 2nd stage we compare the population densities of those cities to determine if a higher population density results in more pollution.

#### **First Stage**:

1. Gathering all resources & importing necessary libraries.
2. Creating a list each for most & least polluted cities in the world using relevant data sources.
3. Getting the Latitude & Longitude values of the cities using Python geocoder library an putting all the data in a dataframe.
4. Finding out all the nearby venues within a certain radius in each city, categorizing them & putting all the results in a dataframe.
5. Finding out the most common venues from each city using Python pandas toolkit.
6. Comparing the common venues to find out if there is any pattern of differences between the most & least polluted cities.

#### **Second Stage**:

1. Gathering the population densities of all the listed cities and creating a dataframe using this data.
2. Creating a visualization of the population data using the dataframe, separating most polluted & least polluted cities.
3. Using the visualization, finding out if higher population density in cities corresponds to more pollution.

### **4.2 Exploratory Data Analysis**

## Importing Necessary Libraries[¶](http://localhost:8888/notebooks/Battle_of_the_Neighbourhoods.ipynb#Importing-Necessary-Libraries)

## 

## Creating Dataframe

## 

## 

## 

## Explore all Cities

## 

## 

## 

## 

## Analyzing Each City

## 

## 

## Grouping rows by City and by taking the mean of the frequency of occurrence of each category

## 

## 

## Printing each City along with the top 10 most common venues

## 

## 

## Checking Population Density of each City from [www.wikipedia.com](http://www.wikipedia.com) & creating a DataFrame

## 

## Results

## *Each City along with the top 10 most common venues*

## *Most Polluted Least Polluted*

## 

## 

## 

## 

## 

## 

## 

## 

### ***As we can see, despite being far apart in terms of Pollution, all 16 cities have pretty much same types of venues consisting of Restaurants, Hotels, Coffee Shops etc. So, it can be concluded that Pollution must correlated to something other than the internal surroundings of a City***.

## *Plotting the Population Density of each City*

## 

### ***Here, Red bars represent Most Polluted Cities & Green ones represent Least Polluted Cities. As we can see, Cities with higher Population Density tend to be more polluted. This is one of the reasons of pollution and it correlates much more than the internal surroundings of a City to Pollution.***

1. **Discussions**

In this study, the following issues were worth discussing:

1. One important drawback of analyzing with common location data is that **it doesn't account for relative weights of the venues**. For example it gives equal weight to a tannery and a coffee shop. But clearly a tannery is a way more significant venue than a coffee shop,specially in our case. **So it is recommended to address this issue in future studies and if possible come up with a way to single out important venues like tanneries or factories.**
2. The radius limitation gives rise to a problem too. In many cases the entities responsible for air pollution are situated at the edge of the cities. So our analysis can't properly include them. **It is recommended to address this issue in future studies too.**
3. In case of population density, even though in most cases a higher population density means more pollution, there are some pretty significant outliers too. For example Honolulu has a way higher population density than Hotan despite being much less polluted. So clearly population density alone can't explain pollution. Other factors, like Waste Management systems should be considered as pointed out in our data sources. **It is recommended to study what other factors in a city leads to more pollution in future works.**
4. **Conclusion**

In this study, a comprehensive approach was taken to study what factors are behind air pollution in many cities. Analyzing location,common venues and population data of **8 Most Polluted** & **8 Least Polluted** cities, we can share at least 2 concluding remarks:

1. **The most common structures and venues within a city are pretty much the same in most of them and they have little effect in making a city a polluted one. Some significant venues might seem to be exceptions but the regular surroundings have little to no impact in this case, hence uncorrelated to pollution.**
2. **Even though there can be many other factors, Densely populated cities tend to be more polluted than less densely populated cities. So, there is a positive correlation between population density & pollution.**

City authorities and Urban planners should take note of these remarks,plan accordingly and there should be further study in order to solve the problem of air pollution in big cities.

1. **Appendix**

For further information on this study, check the *Github* repository : <https://github.com/Saiful185/Data_Science_Capstone>