Analysis of air pollutant emission

1. Introduction: Research Problem

1.1 Background

With the “Fridays for future” strike going on, we can see an increasing attention on climate change among the younger generation. Extreme weather, heavy flood and drought arouse the public’s anxiety about the global warming. One of the major drive of climate change is air pollutant emission. Green House Gases (GHGs), carbon dioxide (CO_2), aerosol particles and so on are acting as heating intermediate in the atmosphere. Human activities, industry productions, road vehicles as well as biogenic and oceanic sources are emitting chemical species into the atmosphere. After certain import chemical and micro-physical processes , these chemicals exert different influence on the climate.

To slow the pace of climate change, different protocols, for instance the Montreal Protocol and the Paris agreement, have been discussed and applied to different countries, Many countries are trying to cut down their emissions and to reach the goal made in protocols.

1.2 Problem

The first question to cut down emission is to evaluate emissions from each country., including the amount, emission categories, key pollutants, the drive behind the trend of emission values. This is project will focus on to the above mentioned points to analyze air pollutants from some countries. Besides, caparison between/among countries is also done to find target countries, which emits relatively more air pollutants and its emission should be strictly monitored and controlled.

1.3 Interests

policy makers could be interested to see such a report about their own country’s emission values and trend. Thus they can contribute to making policies on industry emission standards, public transport and so on, to decrease air pollutant emission. This may also interest environmentalists and even ordinary people, by providing them hints to cut down air pollutant emissions from their own perspective.

2. Data

2.1 Data source

There many online database about air pollution, built by environmental protection bureau and public affair organizations. Since I am most interested in air pollutant emissions in European countries, I obtained the air emission data by accounts from 37 countries from here, including some countries outside Europa. This dataset includes emission data from 2011 to 2017. Emissions from different sources, such as industry and land use and so on, and the amount of each air pollutants are also contained in the dataset. The information of all listed counties, like country area and population, are produced with the package countryinfo.

2.2 Data cleaning

First, data are downloaded from the website and uploaded to the IBM cloud. While reading in data, the dataset is the dataset is simplified by only keeping columns 'country', 'pollutant', 'activity', 'year' and 'value'. The same air pollutant from different sourced are combined and separated with comma in the ‘activity’ column.

I acquire the total emission of all air pollutants for each year and for each country to discuss the emission trend in all countries and to categorize these countries. Not all countries documented their

emissions from 2011 to 2017. Missing emission values are then replaced by mean values from other years. Countries are clustered based on their average emission over years.

Second, country emission data and their geographical data are combined to visualize the clusters of countries on a map.

Third, emission data and countryinfo data are merged to get average emission per million people ("density_pop") and per square kilometer ("density_area"). By sort the data and apply linear regression, relations between emisison and land/population can be revealed.

2.3 Delivering messages

In the end, the Foursquare location data will be used to help me out if I would like to present my analysis in schools as an appeal to the young to make contribution to protect our environment and climate, for example in Cologne. It would be helpful as well for me to how to organize my trip if I would like to deliver these messages in different schools.

Schools 5 km meters away from the Cologne central station will be chosen as the target schools. They are divided into 6 groups according to their locations so that better efficiency can be achieved by crossing these schools.

3. Exploratory Data

There are two main dataframes of emission values: df_data and df_country. df_data contains the emission values of different air pollutants from different years and different countries, in spite of combined emission sources. Whereas df_country summarized all values of different air pollutants as the total emisison for each country in each year. df_data is more for analysis air pollutants in specific country and df_country fits better for comparisons among countries.

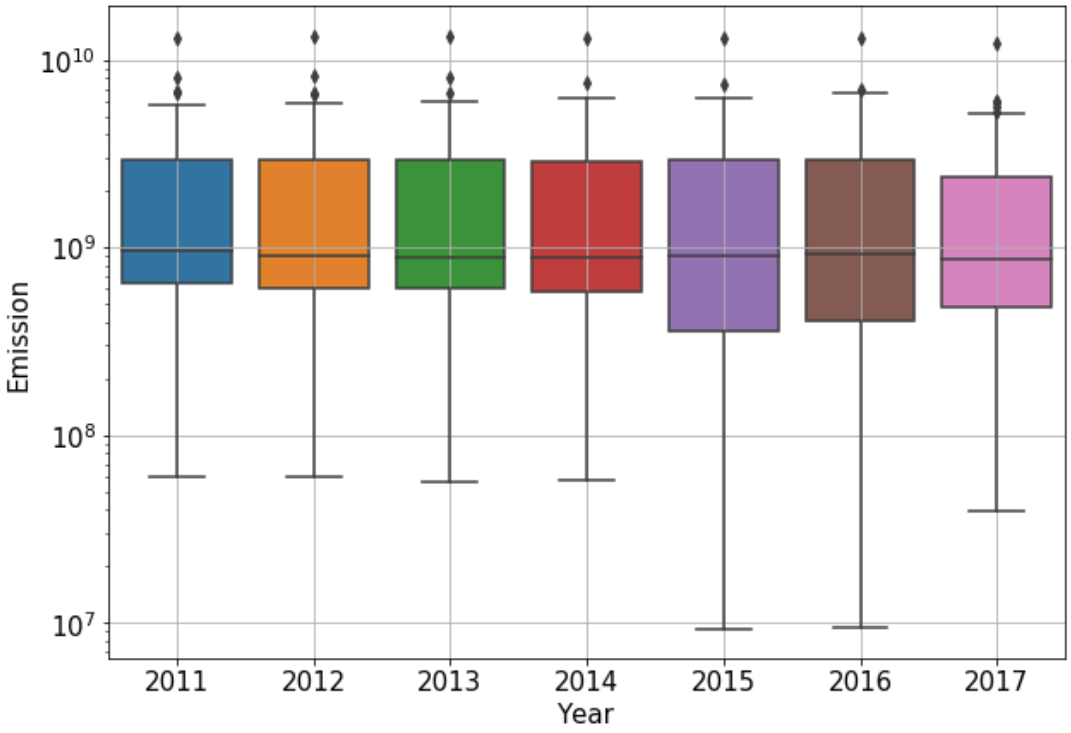
df_data:

	year	country	pollutant	value	activities
0	2011	Australia	GHG	1749560300	IND-TOTAL, A, B, C, C10-C12, C13-C15, C16-C18,...
1	2011	Austria	ACG	1243390	IND-TOTAL, A, B, C, C10-C12, C13-C15, C16-C18,...
2	2011	Austria	CH4	36131775	IND-TOTAL, A, B, C, C10-C12, C13-C15, C16-C18,...
3	2011	Austria	CO	2816107	IND-TOTAL, A, B, C, C10-C12, C13-C15, C16-C18,...
4	2011	Austria	CO2	650233746	IND-TOTAL, A, B, C, C10-C12, C13-C15, C16-C18,...
5	2011	Austria	GHG	376583182	H52, C22-C23, C22, A02, H53, D, M72, C33, HH04...
6	2011	Austria	HFC	6647100	IND-TOTAL, IND-TOTAL, A, A, A01, A01, A02, A02...
7	2011	Austria	N2O	16799090	IND-TOTAL, A, A01, A02, A03, B, C, C10-C12, C1...
8	2011	Austria	NH3	947678	IND-TOTAL, IND-TOTAL, A, A, A01, A01, A02, A02...
9	2011	Austria	NMVOC	636920	IND-TOTAL, A, A01, A02, A03, B, C, C10-C12, C1...

df_country:

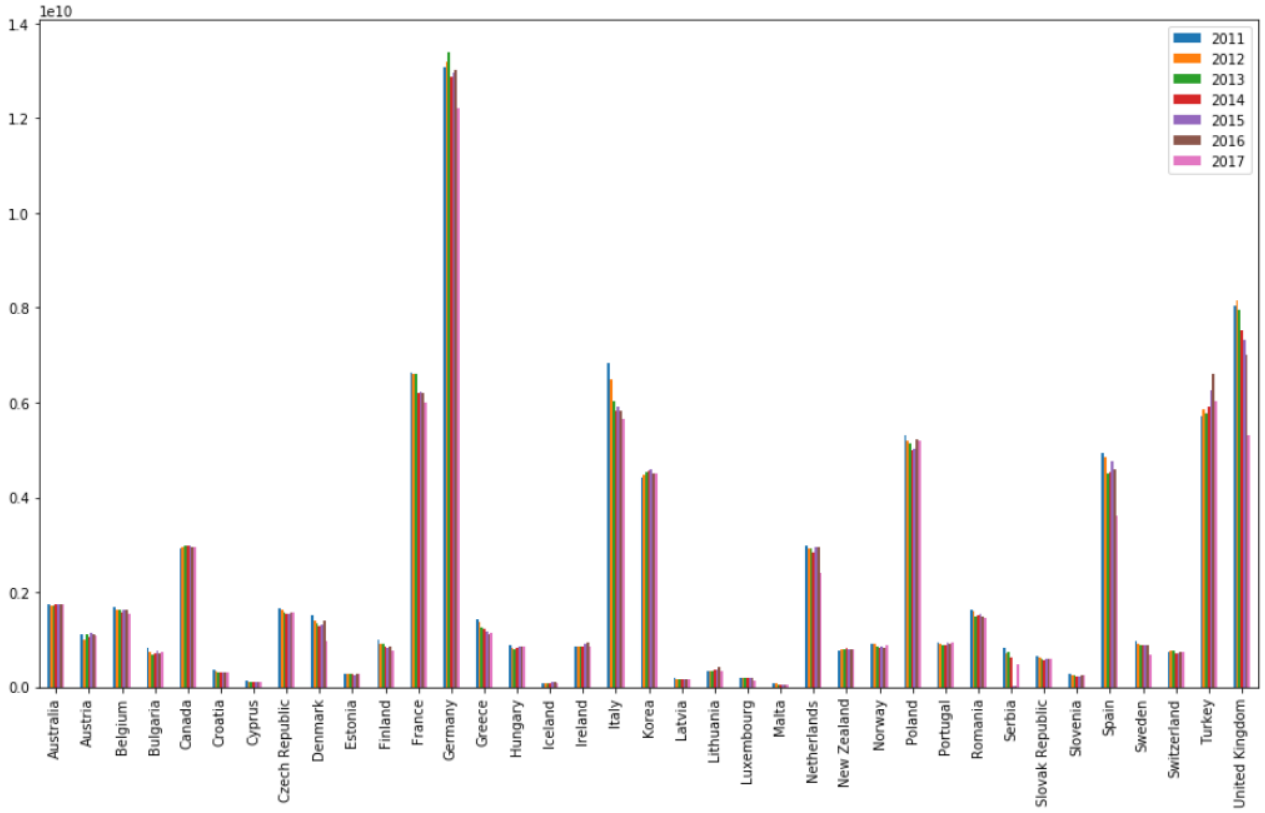
	2011	2012	2013	2014	2015	2016	2017
Australia	1.749560e+09	1.701095e+09	1.717634e+09	1.730529e+09	1.739198e+09	NaN	NaN
Austria	1.098418e+09	9.923948e+08	1.119370e+09	1.064454e+09	1.131147e+09	1.105221e+09	1.069875e+09
Belgium	1.677021e+09	1.629567e+09	1.640869e+09	1.570333e+09	1.620684e+09	1.621602e+09	1.548004e+09
Bulgaria	8.081259e+08	7.444715e+08	6.732812e+08	7.100709e+08	7.501578e+08	7.182643e+08	7.383424e+08
Canada	2.919352e+09	2.942326e+09	2.984098e+09	2.989698e+09	2.988086e+09	2.944435e+09	NaN

3.1 Descriptive statistics among countries

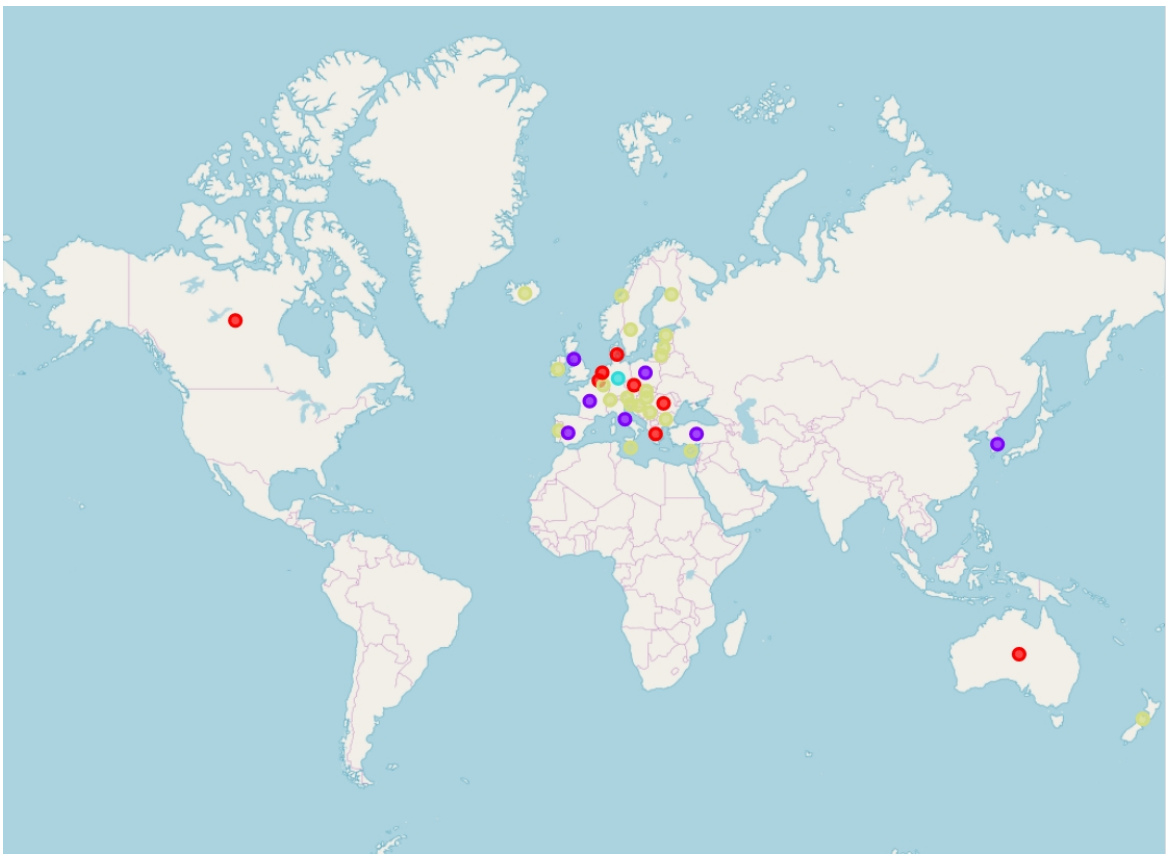


Above is a box plot from total emission values in df_country (NaN are replaced by mean values as indicated in section 2). The x-axis is the year and y-aixs shows the emission (Tonne). This boxplot shows the min. and max. values from each year. We can see that all countries' mean emission didn't vary much from 2011 to 2017. However, the max. value from each year showed a slightly decreasing trend. So did the min. emission, especially in 2015 and 2016, some country's sum emisison were largely cut down. Some countries meanwhile are regarded as outliers, since they were producing far more air pollutants than their counterparts. Now let's see who was emitting the most and the least air pollutants among these countries with a bar plot.

Bellow is the bar plot. In this plot, countries are plotted along x-axis and y-aixs again shows the emission (Tonne). Different colors indicate different years. It clearly says that Germany produced the most air pollutants in the list. The amount produced by United Kingdom, France, Italy and Turkey was almost half of that from Germany. Malta, Cyprus and Iceland emitted the least air pollutants in the past several years.

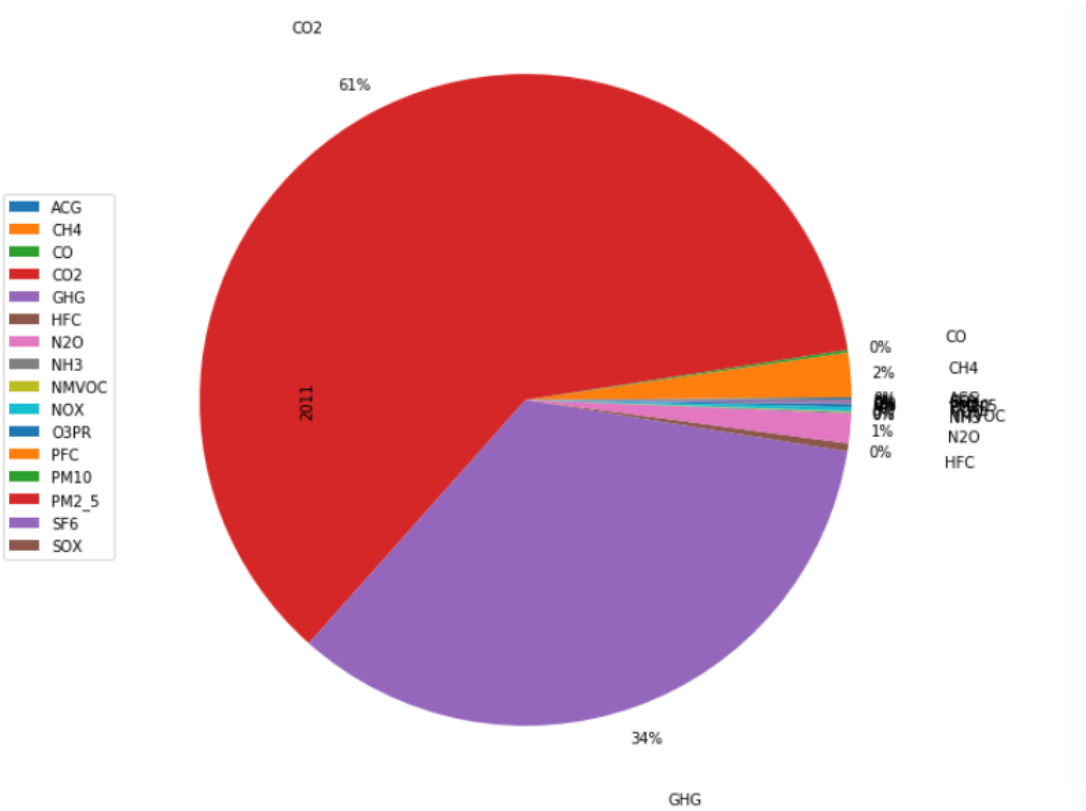


These countries can then be clustered into 4 groups according to their mean emission in the 7 years. They are classified as low (cluster label 3), medium (0), high (1) and extremely high (2) emission groups. After getting the coordinated of all countries, the geographical distribution of these 4 groups are then visualized on map as following:



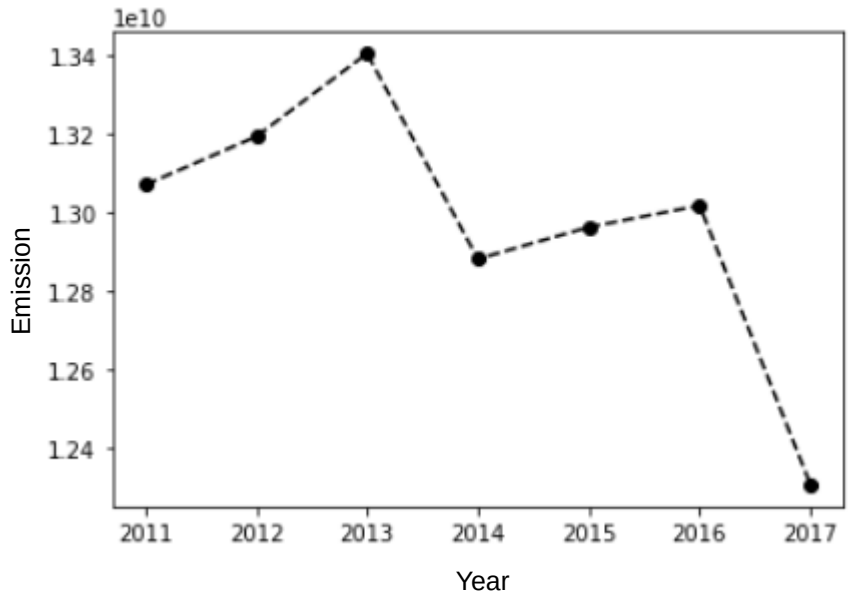
3.2 Exploring emissions from one country and compare with some other countries

Given that Germany emitted the largest of amount of air pollutants, it's curios to know which pollutant matters most in terms of its contribution to the total emission. Here an example is given from year 2011.

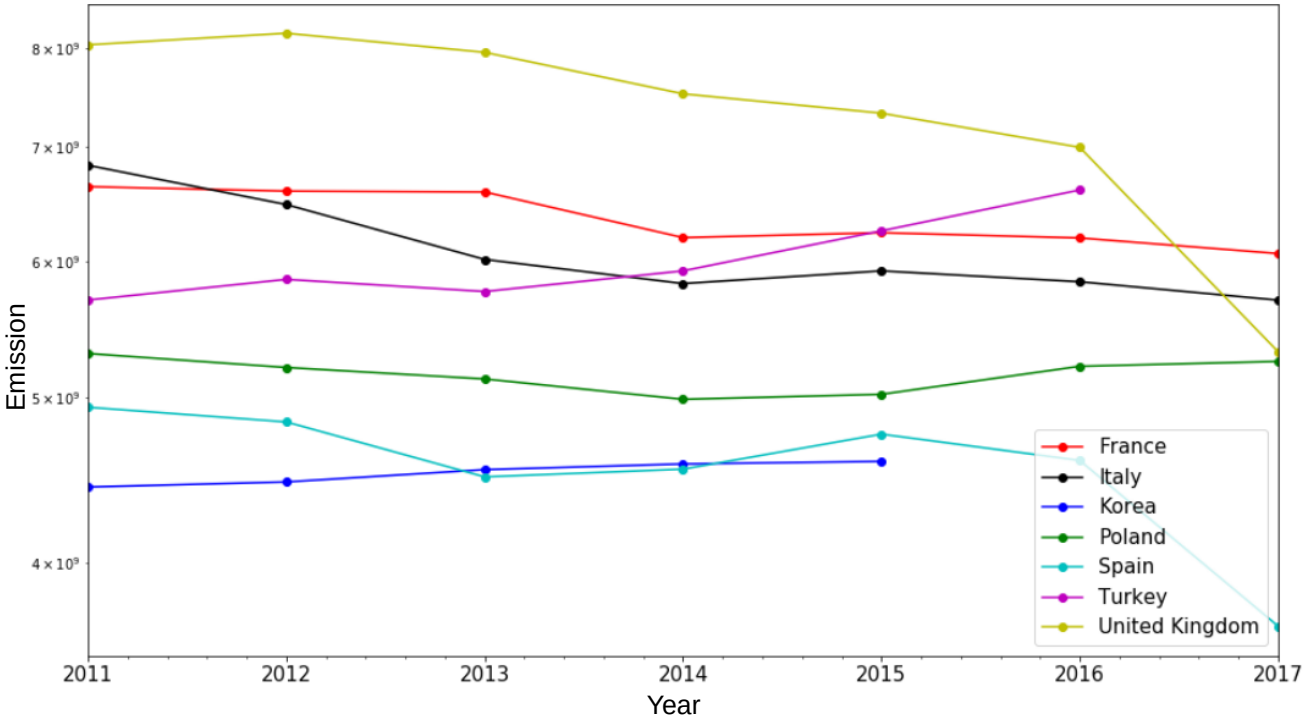


This pie plot tells that CO_2 (61 %) and Green House Gas (GHG, 34 %) were the main constitutes of the air pollutants in 21011. Only 5 % of the total emission was composed of other pollutants. The correlation coefficient of CO_2 and GHG is 0.999703 and 0.99995 respectively, which reveals that the total emission in Germany in 2011 was high positively correlated with CO_2 and GHG emissions. Similar correlation can be found from other years.

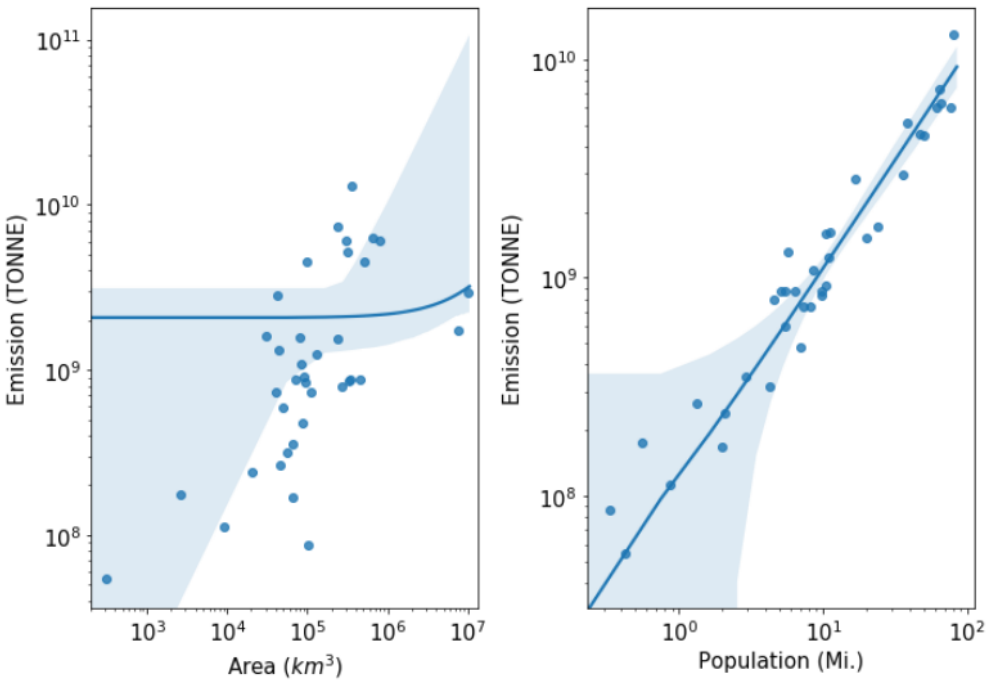
The total emission in 2018 and 2019 in Germany can be predicted from previous data collection. It seems that there is a strong decrease in the amount every three years. And the three years cycle saw a gradual increase. Thus, the emission in 2018 and 2019 could be slightly higher than that in 2017.



Comparison among countries in the same cluster is also feasible. For example, the following plot shows the emission from countries in cluster 1. Clear drops exist in UK and Spain from 2012 to 2017. Other countries show slowly decrease during this time period, except that the emission in Turkey was gradually increasing and quite stable in Korea.



3.3 Linear Regression model



The above plot displayed the fitting of the linear regression model. The R2-score of this model is -0.65 for Area and 0.99 for Population. It reveals that a strong linear regression was found between averaged total emission and population ,and emission values were not linearly related to a country's

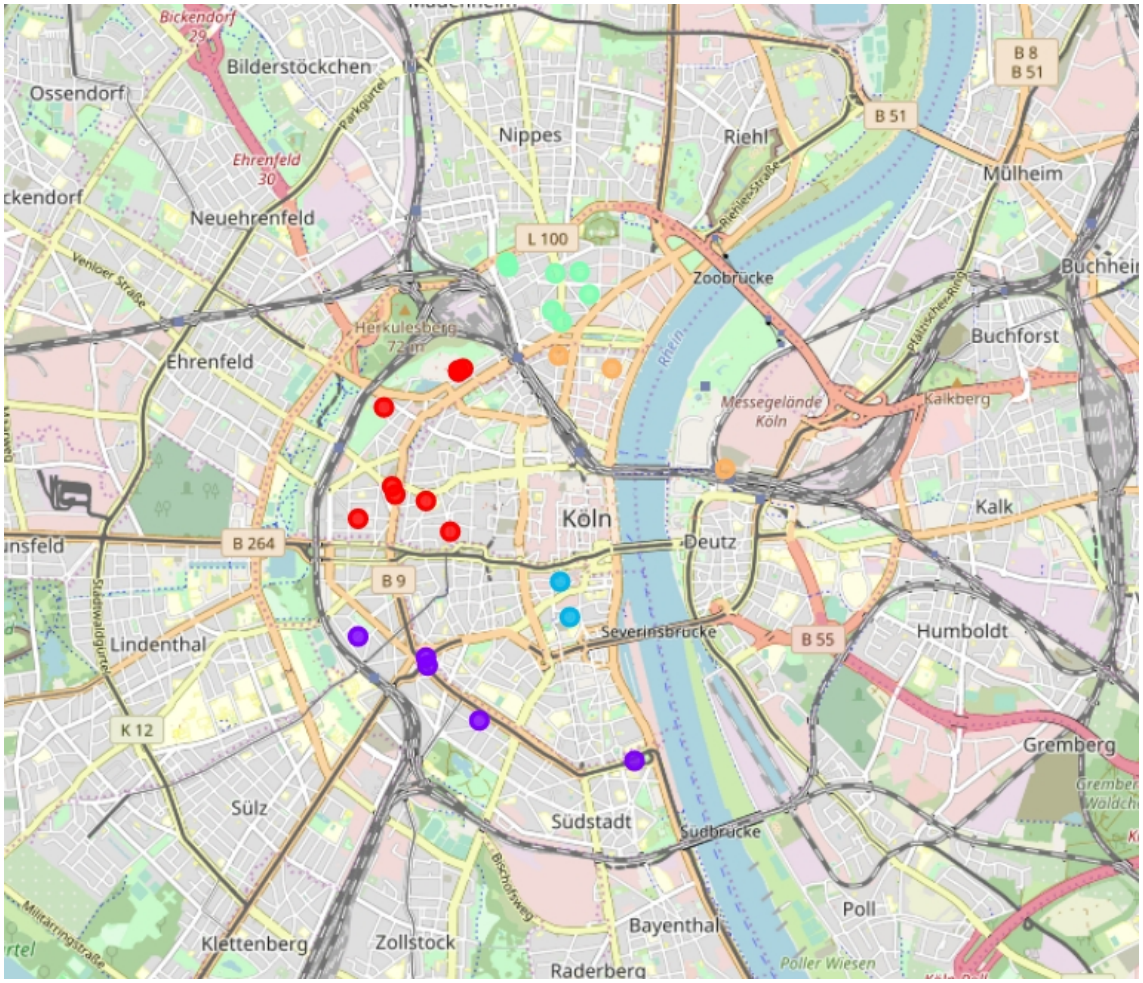
area. When a country has a larger population, it is more likely to produce more air pollutants. Thus it really calls for every single person's attention and contribution to reduce air pollutant emission.

4. The application of Foursquare API

Since air pollution and climate change is a highly focused issue of the society. I would like to draw more attention from the younger generation to care about our climate as well as policy makers to plan ahead in decreasing air pollutant emissions. Let's say that I plan to deliver my analysis to students in Cologne, considering that the emission in Germany is extraordinary more than other countries in the list. At first, I have to count on the Foursquare API to search schools within 2 kilometer of Cologne central station. Below is a part of a list of 260 schools within the searching area.

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng	postalCode	state	id
0	Célestin-Freinet-Schule	Elementary School	Dagobertstr. 79	DE	Köln	Deutschland	NaN	1126	[Dagobertstr. 79, 50668 Köln, Deutschland]	[{'label': 'display', 'lat': 50.94834705513646...	50.948347	6.962592	50668	Nordrhein-Westfalen	5167c10de4b0e8420d8653f8
1	Kaiserin Augusta Schule	High School	Georgsplatz 10	DE	Köln	Deutschland	NaN	756	[Georgsplatz 10, 50676 Köln, Deutschland]	[{'label': 'display', 'lat': 50.93166586986782...	50.931666	6.958150	50676	Nordrhein-Westfalen	4c0531b49a7920a18e42d279
2	Bali-Dojang Kampfkunst Schule	Martial Arts Dojo	Neusser Str. 81	DE	Köln	Deutschland	NaN	1833	[Neusser Str. 81, 50670 Köln, Deutschland]	[{'label': 'display', 'lat': 50.954697, 'lng':...	50.954697	6.956593	50670	Nordrhein-Westfalen	5ac5097ebd897e2f561af1d4
3	Paul Maar Schule	None	NaN	DE	Köln	Deutschland	NaN	521	[Köln, Deutschland]	[{'label': 'display', 'lat': 50.934038, 'lng':...	50.934038	6.957092	NaN	Nordrhein-Westfalen	4d5ba8f8590b224bae27986d
4	Königin-Luise-Schule	High School	Alte Wallgasse 10	DE	Köln	Deutschland	Palmstraße	1208	[Alte Wallgasse 10 (Palmstraße), 50672 Köln, D...	[{'label': 'display', 'lat': 50.93948902184421...	50.939489	6.942840	50672	Nordrhein-Westfalen	4ca0f20f8afca0931f2d1716

To facilitate my business trip, I would love to visit schools which are not far from each other within one or two days so that I won't waste much time on the way. Therefore I divided these schools into 5 groups so that each time I visit schools within one groups to save commuting time.



5. Conclusion

In this study, I analyzed the air pollutant emissions from 37 countries during 2011-2017, including most European countries, Australia, Canada, Korea and so on. I separated these countries into 4 emission categories: low, medium, high and extremely high emission. Analysis was more focus on Germany due to that it emitted extremely large amount of air pollutants during these years. To pose less threat to climate change, reduction of CO_2 and GHG emissions in Germany shall be seriously considered, in fact that the two species comprised 95 % of the total emission. Based on previous emission volumes, a prediction that Germany would emit slightly more air pollutants than 2017 was made. Among the high emission countries, emission reduction was observed in countries like France and Spain. But increasing emission was also shown in Turkey. A linear regression model was also applied to analyze the relationship between mean emission of each country and its area and population of the country. Larger as the the population is, a country tends to emit more air pollutants. Last but not least, I chose 50 schools within 5 km around Cologne central station to deliver my analysis with Foursquare API. I divided these 50 schools into 6 clusters to simplify my travel plan.

6. Prospects

Considering more countries for such an analysis can provide us more insights into how to reduce air pollutant emission, which calls for a large and complete dataset. The impact of population on air pollutant emission was briefly discussed. Other factors, such as economy, road traffic, agriculture, *etc.*, could be further explored apart from a country's area and population. More models may be applied to feature studies so that prediction based on model fitting is robust. And in this case, the model was not trained due to its limited data size. To better understand the influence of population on air pollutant emission, population data from these countries during the study period might be collected to test the linear regression model.