Analysis By: Jay Naresh Dhanwant

IIT(BHU) Varanasi

Jaynaresh.dhanwant@gmail.com

LinkedIn: Jay Dhanwant

Link to my website

Description of the Data: Pollution trend of 30 countries were provided from the year 1990 to 2017

Assumption: Assuming that the analysis is made with the perspective of an executive from some global pollution control organisation.

CONTENTS

Problem Statement 1:

To determine how the countries performed with reference to the increase in pollution (Top 5 and bottom 5 performers)

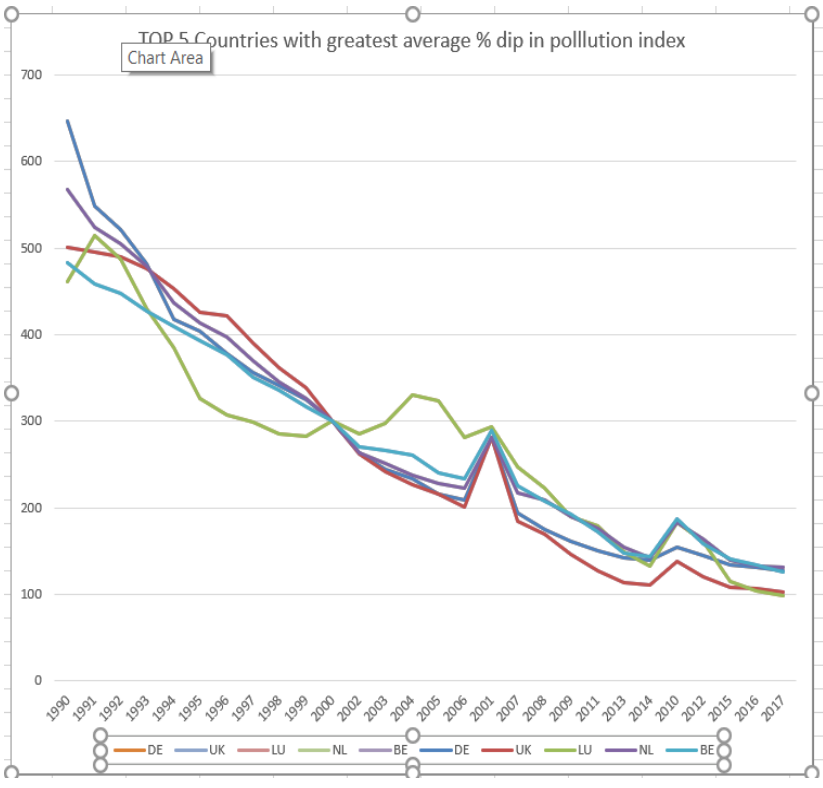
Motivation behind this problem statement: Being an executive of an pollution control community, I will be interested in knowing about the countries that performed best, with the help of statistical parameters, so that I can study the measures and policies adopted by these countries, and similarly identify bottom 5 countries to study the problem with their pollution control system.

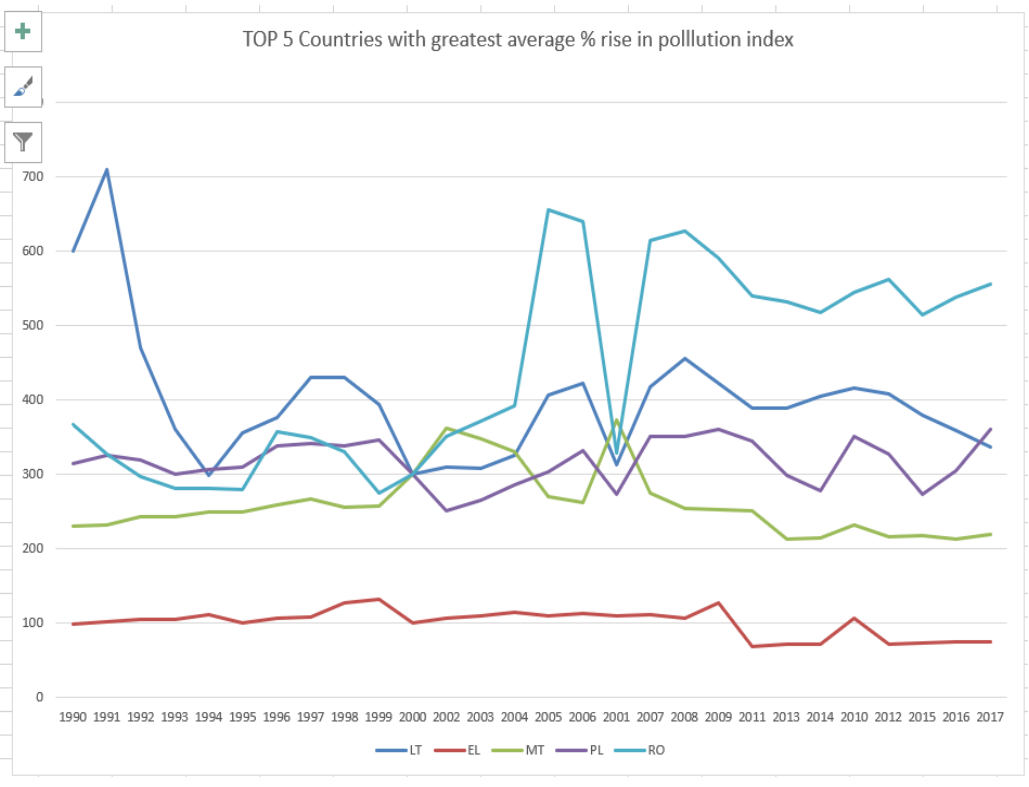
Metrics for evaluation and Data manipulation:

1. A metric – Year over Year was defined which gives the percent increase in pollution each year with respect to previous year, this gives us the idea of how the country performed for tackling the air pollution. {A high YoY% will mean an increment in pollution due to poor performance}
2. Another metric Average YoY% over a country was defined giving the average performance of the country in regards of the pollution. {A high average YoY% will mean poor performance of the country averaged over the years}

Analysis:

1. Net pollution index was calculated for all the cities
2. YoY% and average YoY was calculated
3. Ranking the country: Sorting based on average YoY% (Smallest to largest
   {Most negative growth would mean highest dip in average pollution over the years}





Insights:

1. Top 5 countries which the organization should approach and study are DE(Germany), UK(United Kingdom), LU(Luxembourg), NL(Netherlands), BE(Belgium)
2. Bottom 5 countries where the we can look for the flaws in the policies are RO(Romania) , PL(Poland) , MT(Malta) , EL(Greece) , LT(Lithuania)
3. **Lithuania** did a pretty good job in decreasing the population levels in the early 5 years but had a drastic surge in the pollution
4. Pollution of all the better performing nations spiked twice, once around the year 2001 and once around the year 2012

Problem Statement 2:

To find top 3 years when the pollution spiked and top 3 years when the pollution was controlled

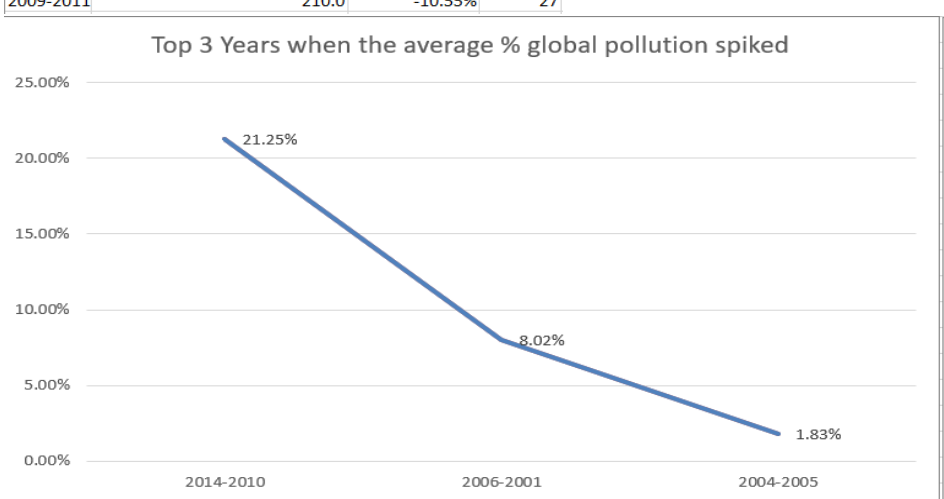
Motivation behind this problem statement:

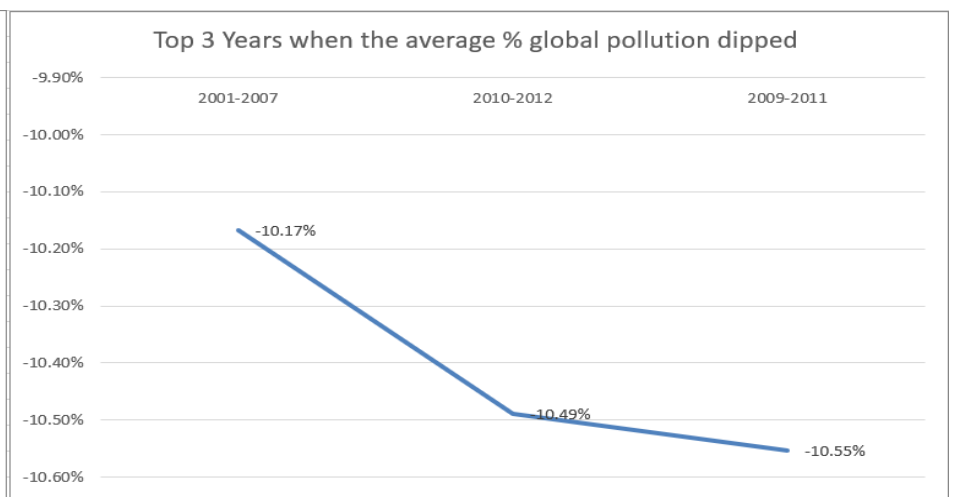
There can be specific years in which pollutions relatively rose across the world regardless of country. This can give us insights to what really happened to the world as a whole, eliminating the factor of policies adopted by specific countries.

Metrics for evaluation and Data Manipulation:

1. A transpose of the data was taken with appropriate manipulations.
2. Average pollution over the countries for all the years was calculated
3. Increase percent of the pollution was calculated similar to YoY percent
4. Data was ranked and visualizations were done

Years	Average Pollution index glo	Increase in p	Rank
2014-2010	223.7	21.25%	1
2006-2001	291.6	8.02%	2
2004-2005	280.2	1.83%	3
1995-1996	341.9	-0.21%	4
2016-2017	176.3	-0.91%	5
2002-2003	279.8	-0.96%	6
2015-2016	178.0	-1.51%	7
2003-2004	275.2	-1.64%	8
1997-1998	323.1	-1.68%	9
1994-1995	342.6	-2.39%	10
2013-2014	184.5	-2.66%	11
1993-1994	351.0	-2.84%	12
1990-1991	395.0	-3.40%	13
2005-2006	270.0	-3.65%	14
2000-2002	282.5	-3.70%	15
1996-1997	328.6	-3.88%	16
2007-2008	251.7	-3.94%	17
1992-1993	361.2	-4.19%	18
1998-1999	309.3	-4.29%	19
1991-1992	377.1	-4.54%	20
1999-2000	293.3	-5.15%	21
2008-2009	234.7	-6.72%	22
2011-2013	189.5	-9.73%	23
2012-2015	180.7	-9.77%	24
2001-2007	262.0	-10.17%	25
2010-2012	200.2	-10.49%	26
2009-2011	210.0	-10.55%	27





Insights:

1. In the period between 2010 and 2014, we see a highest surge in pollution growth
2. Highest dip in pollution was observed in the period between 2001 and 2007

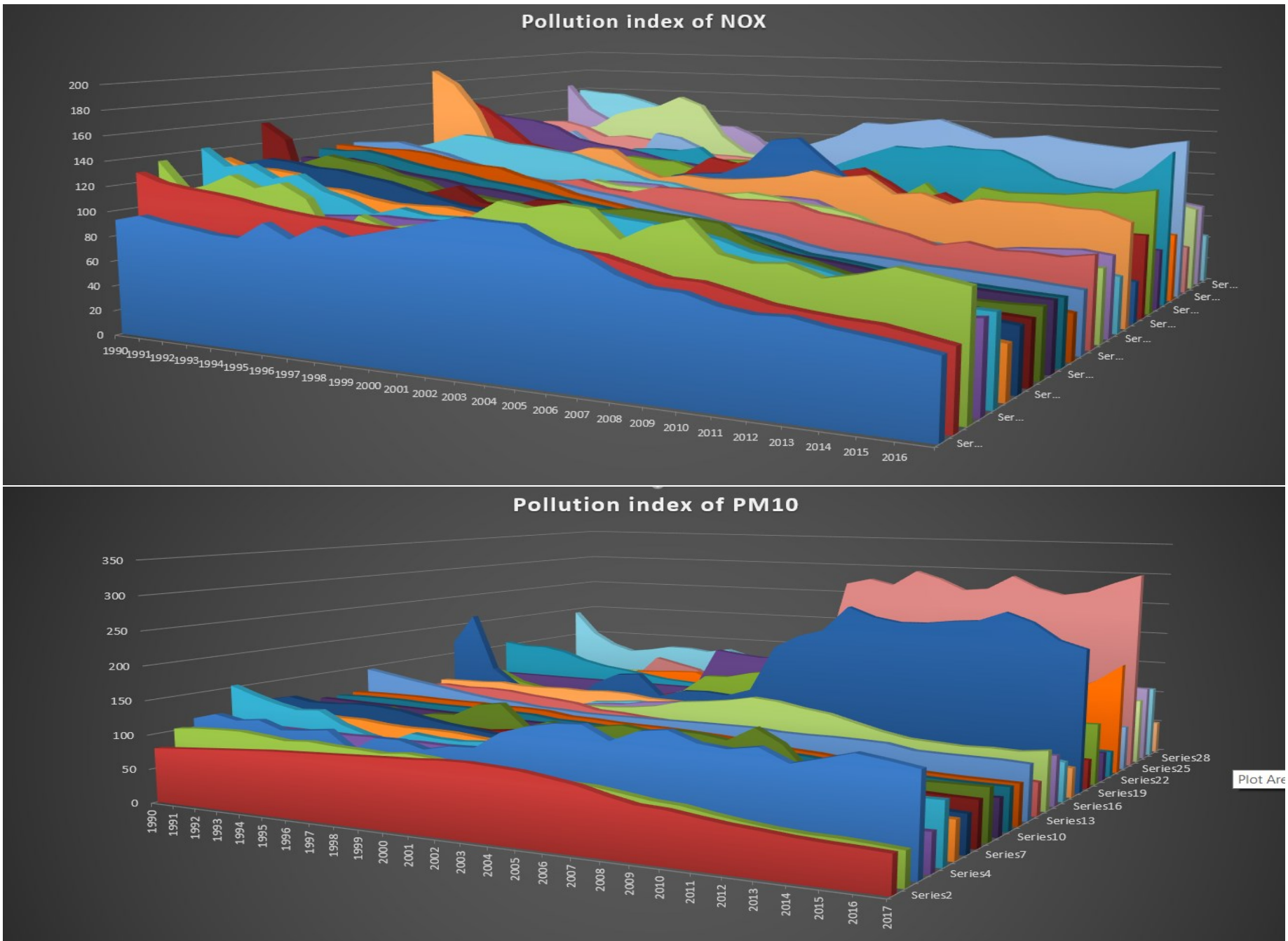
Problem Statement 3:

To observe the trend in the pollution growth over the years

Motivation behind this problem statement:

Visualization of the trends over the year can give us very valuable insights and is crucial for deciding the kind of machine learning model to use.

Visualization:



Insights:

1. An exponential decrease of the pollutant NMVOC was observed, with some outliers around 2009
2. There was a sudden and abrupt rise in the PM10 pollution after 2008 in some cities, but the pollution index for this pollutant was mostly stable.
3. There was a linear decrement in the pollution level of NOX, which some exceptions. These anomalies can be studied further to better analyse the population growth

Problem Statement 4:

Correlation of the different features of the data, to study the dependencies in pollution rate of various countries and the dependency of one pollutant with other two.

	DE	UK	LU	NL	BE	AT	FR	DK	EU28	FI	EE	EU27_2	IT	ES	SE	LV	CZ	HU	SK	IE	PT	CY	HR	SI	BG	LT	EL	MT	PL	RO
DE	1																													
UK	0.97	1																												
LU	0.918	0.911	1																											
NL	0.993	0.987	0.926	1																										
BE	0.979	0.993	0.941	0.992	1																									
AT	0.961	0.954	0.981	0.964	0.977	1																								
FR	0.965	0.998	0.919	0.982	0.993	0.958	1																							
DK	0.941	0.99	0.925	0.967	0.99	0.96	0.993	1																						
EU28	0.966	0.997	0.928	0.983	0.996	0.967	0.998	0.996	1																					
FI	0.96	0.992	0.93	0.976	0.994	0.969	0.996	0.996	0.999	1																				
EE	0.888	0.89	0.866	0.888	0.911	0.925	0.891	0.905	0.91	0.922	1																			
EU27_2	0.963	0.994	0.931	0.98	0.996	0.969	0.996	0.996	1	0.999	0.914	1																		
IT	0.902	0.973	0.886	0.932	0.965	0.926	0.98	0.989	0.982	0.985	0.888	0.983	1																	
ES	0.926	0.981	0.927	0.953	0.981	0.957	0.986	0.996	0.991	0.992	0.899	0.992	0.992	1																
SE	0.979	0.993	0.935	0.991	0.997	0.974	0.992	0.987	0.995	0.993	0.919	0.994	0.963	0.978	1															
LV	0.894	0.868	0.968	0.89	0.911	0.97	0.875	0.889	0.892	0.899	0.891	0.897	0.844	0.891	0.903	1														
CZ	0.9	0.916	0.954	0.925	0.949	0.957	0.921	0.944	0.938	0.939	0.877	0.942	0.911	0.947	0.939	0.935	1													
HU	0.8	0.834	0.882	0.811	0.866	0.906	0.849	0.883	0.874	0.889	0.911	0.884	0.889	0.903	0.86	0.91	0.908	1												
SK	0.835	0.848	0.863	0.855	0.883	0.898	0.844	0.88	0.874	0.877	0.897	0.88	0.853	0.881	0.873	0.887	0.939	0.907	1											
IE	0.804	0.887	0.856	0.833	0.891	0.885	0.899	0.93	0.914	0.924	0.886	0.921	0.954	0.95	0.89	0.84	0.888	0.951	0.858	1										
PT	0.845	0.926	0.873	0.876	0.925	0.904	0.941	0.96	0.947	0.955	0.88	0.952	0.981	0.974	0.923	0.844	0.899	0.93	0.836	0.98	1									
CY	0.817	0.89	0.88	0.851	0.906	0.904	0.904	0.94	0.921	0.931	0.893	0.928	0.951	0.955	0.9	0.872	0.929	0.962	0.901	0.971	0.974	1								
HR	0.469	0.529	0.584	0.484	0.567	0.617	0.547	0.612	0.59	0.615	0.739	0.606	0.652	0.648	0.564	0.652	0.667	0.875	0.755	0.811	0.735	0.815	1							
SI	0.774	0.875	0.715	0.832	0.856	0.765	0.866	0.894	0.871	0.866	0.752	0.869	0.896	0.881	0.848	0.668	0.824	0.715	0.8	0.817	0.842	0.844	0.531	1						
BG	0.774	0.724	0.77	0.767	0.77	0.795	0.719	0.727	0.738	0.73	0.674	0.74	0.673	0.721	0.744	0.798	0.834	0.726	0.794	0.634	0.642	0.684	0.476	0.594	1					
LT	0.466	0.329	0.414	0.435	0.366	0.415	0.3	0.281	0.321	0.309	0.423	0.317	0.187	0.247	0.383	0.401	0.352	0.203	0.433	0.128	0.102	0.149	0.034	0.155	0.294	1				
EL	0.381	0.478	0.523	0.426	0.506	0.526	0.497	0.56	0.53	0.542	0.567	0.544	0.597	0.606	0.494	0.535	0.635	0.753	0.706	0.736	0.681	0.739	0.819	0.499	0.435	0.034	1			
MT	0.036	0.113	0.247	0.047	0.147	0.237	0.158	0.211	0.173	0.211	0.293	0.189	0.273	0.254	0.137	0.355	0.231	0.524	0.226	0.443	0.414	0.455	0.653	0.096	0.086	-0.43	0.482	1		
PL	-0.04	-0.03	-0.11	-0	-0.03	-0.11	-0.06	-0.04	-0.04	-0.06	-0.04	-0.04	-0.07	-0.04	-0.02	-0.17	0.029	-0.13	0.111	-0.08	-0.1	-0.02	-0.01	0.155	-0.09	0.361	0.18	-0.47	1	
RO	-0.74	-0.81	-0.64	-0.75	-0.77	-0.71	-0.83	-0.79	-0.8	-0.81	-0.69	-0.8	-0.83	-0.79	-0.78	-0.6	-0.62	-0.66	-0.51	-0.75	-0.83	-0.72	-0.41	-0.66	-0.42	0.061	-0.34	-0.31	0.313	1

	NMVOC	*NOX*	*PM10*
NMVOC	1		
NOX	0.998527	1	
PM10	0.996388	0.995717	1

Insights:

1. Pollution of some countries are highly correlated
   Significance: A surge in certain pollutant in a certain country will mean a rise in the pollution of the highly correlated countries in near future.
2. Some of the highly correlated countries include Luxembourg, Germany, United Kingdom and Netherlands.
3. Though contamination of these pollutant varies In magnitude, they are highly correlated.
   Measurement of one pollutant can give a fair estimate of other two.

Problem Statement 5:

Predictions:

1. Predicting the pollution trend for the future
2. Detection of one pollutant can be tedious and costly than the others, building the model to predict to estimate the contamination level of one pollutant based on other two.

Metrics:

Average global pollution index is defined, this average is over the countries, and over the pollutants

Index = Σ Σ(Pollution Index i in country j)/ (number of countries*number of pollutants)

Analysis: (Fig 1: Prediction of Pollution over the years, Fig 2: Prediction of specific pollutant)

1. After doing parameter tuning, a regression model was applied for the future prediction of population
2. Taking advantage of the high correlation of pollutant, similar model was build for the prediction of the categories of the pollutants.