DATA ANALYSIS REPORT

Methods of Advanced Data Engineering

Muhammad Umar Naeem

23170038

Introduction

The project strives to answer the question: How has air quality in the major countries in North & South Americas changed over the past & what are the primary causes for such changes?

Air quality degradation is an important problem because it directly affects all living things including humans, animals & plants. It can lead to health risks, negative environmental impact, economic costs, poor quality of life, etc.

This project analyses the problem by looking at the air quality datasets & emissions datasets & then drawing useful insights & conclusions. The aim is to find strong correlations between air quality degradations & harmful emissions to determine which emissions have the most negative impact on air quality.

The results can help identify which countries in the Americas need to pay serious attention to their air quality improvement and as a next step which emissions (causes) they need to lower down.

Data Sources

The project requires the use of 7 datasets. They have been divided into 2 categories based on the purpose they will serve in this project:

1. Air quality metrics
2. Pollutants

The 3 datasets that provide information about the amounts of **PM 2.5, PM 10** and **NO_x** in the air will serve the purpose of air quality metrics. PM stands for particulate matter. 2.5 and 10 refer to the diameter of the particulate matter. Some particles with diameter less than 10 are inhalable by the lungs and pose serious health risks. PM 2.5 is described by the WHO as the most harmful air pollutant followed by PM 10. NO_x refers to Nitric Oxide (NO) and Nitrogen Dioxide (NO_2) which are most relevant for air pollution. These gases result in the formation of smog and acid rain, as well as affecting tropospheric ozone. The amounts of PM 2.5, PM 10 and NO_x in the air are the best indicators for air quality. The well known Air Quality Index (AQI) also uses the concentration of PM 2.5, PM 10 and NO_x in the formula for its calculation.

The remaining 4 datasets that provide information about the emissions of **CH_4 , NH_3 , Hg** and food serve the purpose of data sources about most prevalent air pollutants. Emissions of **CH_4 (methane)**, **NH_3 (ammonia)**, **Hg (mercury)**, and food-related emissions can contribute to the formation of **PM2.5**, **PM10**, and **NOx**, either directly or indirectly, through various processes. These pollutants will be explained in more detail in the analysis part.

Data Pipeline Result

The data pipeline results in the availability of the datasets in the form of excel files required for the project locally. Each excel file has several worksheets but we are interested in the **Totals By Country** sheet. It contains yearly total values for the concentrations of a substance by country which is perfect for our use case. All datasets contain yearly

values starting from 1970 and mostly ending in 2022. There are 2 exceptions in which ending years are 2023 and 2018. The exceptions don't impact the analysis in any way.

Data Analysis

The main goal in each of these scripts is to discover the trends of the specific substance across major American major American countries which helps us answer 1st part of the project question which is how has air quality changed across the Americas over the past 50 years.

To achieve this, 1st of all the saved datasets from the data pipeline are opened as Pandas data frames. Then a filter is applied to only keep the countries part of the American continents. Afterwards, some statistical techniques are applied to uncover the trends.

1. **Percent Change**: Percent change in amount of a substance from 1970 to end year (mostly 2022). It can be negative and positive. Negative percentage means the country has seen some improved air quality because all of the substances are harmful and vice versa.
2. **Absolute Change:** Given by the difference of amount of substance in the end and start year.
3. **Annual Change**: Average annual change over the years.
4. **Volatility:** Standard deviation of amount of substance over the years.
5. **Linear trend slope:** A linear regression line is fitted to find the linear slope. Countries with the steepest positive slopes have had the worst change in air quality and vice versa.
6. **Max and min amount years:** Records the years with the minimum and maximum amount of substance.
7. **Contribution 2022:** Contribution of the country's 2022 substance amounts to the total amount for all American countries.

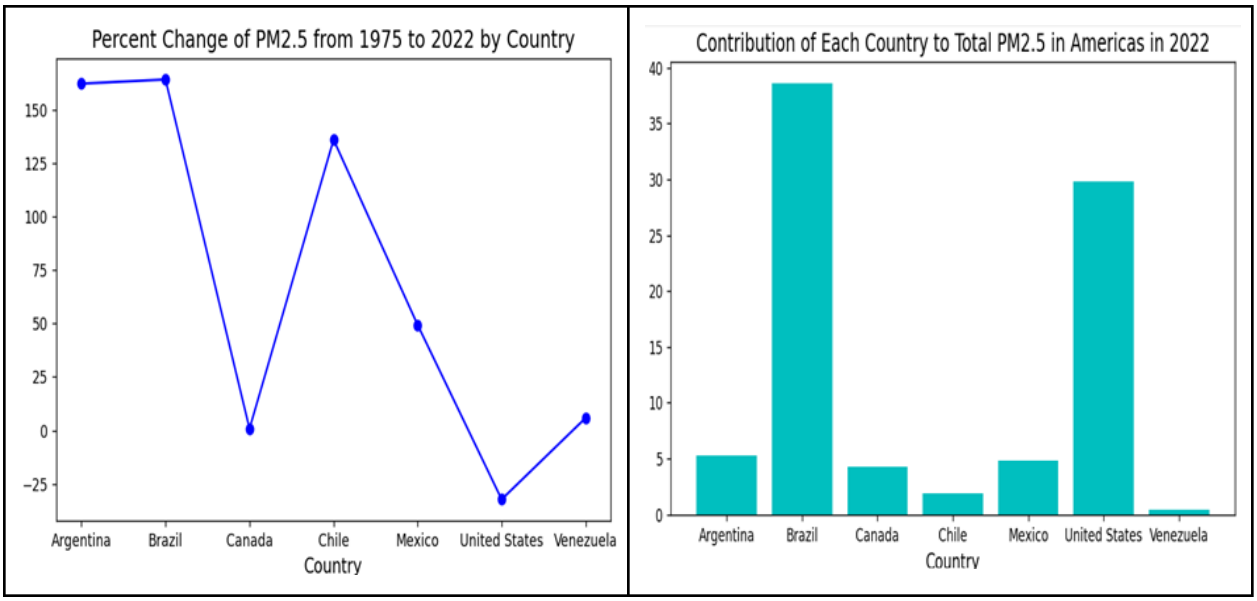
Ultimately, another filter on countries is applied on the data frame that contains the above mentioned statistics and trends to only include the most prominent countries in the Americas i.e.

- **North America**: United States, Canada, Mexico
- **South America**: Brazil, Argentina, Chile

The final data frames for all the substances that get used in the analysis to answer the 2nd part of the project question looks like as shown in the figure below.

Name	Y_2022	Percent_Change	Absolute_Change	Annual_Change	Volatility	Linear_Trend_Slope	Max_PM10_Year	Min_PM10_Year	Contribution_2022
Argentina	327.935870	112.219161	173.408886	3.334786	76.811761	4.672049	Y_2019	Y_1989	4.241849
Brazil	2901.982445	116.020989	1558.602590	29.973127	479.907311	28.659627	Y_2022	Y_1970	37.537134
Canada	306.233943	3.574397	10.568266	0.203236	34.491886	0.756142	Y_1999	Y_1975	3.961135
Chile	189.884474	177.929162	121.563297	2.337756	57.953990	3.356810	Y_2013	Y_1976	2.456155
Mexico	389.305710	49.923955	129.636927	2.493018	48.901647	2.755443	Y_2008	Y_1971	5.035668
United States	2245.460340	-27.575000	-854.933643	-16.441032	344.512910	-18.246654	Y_1988	Y_2020	29.045022
Venezuela	35.395344	17.301995	5.220799	0.100400	13.636656	0.480257	Y_2012	Y_1971	0.457839

Apart from this, a line graph for percent change and a bar graph for country's contribution to total substance amounts in the end year are plotted for better visualization of these trends.



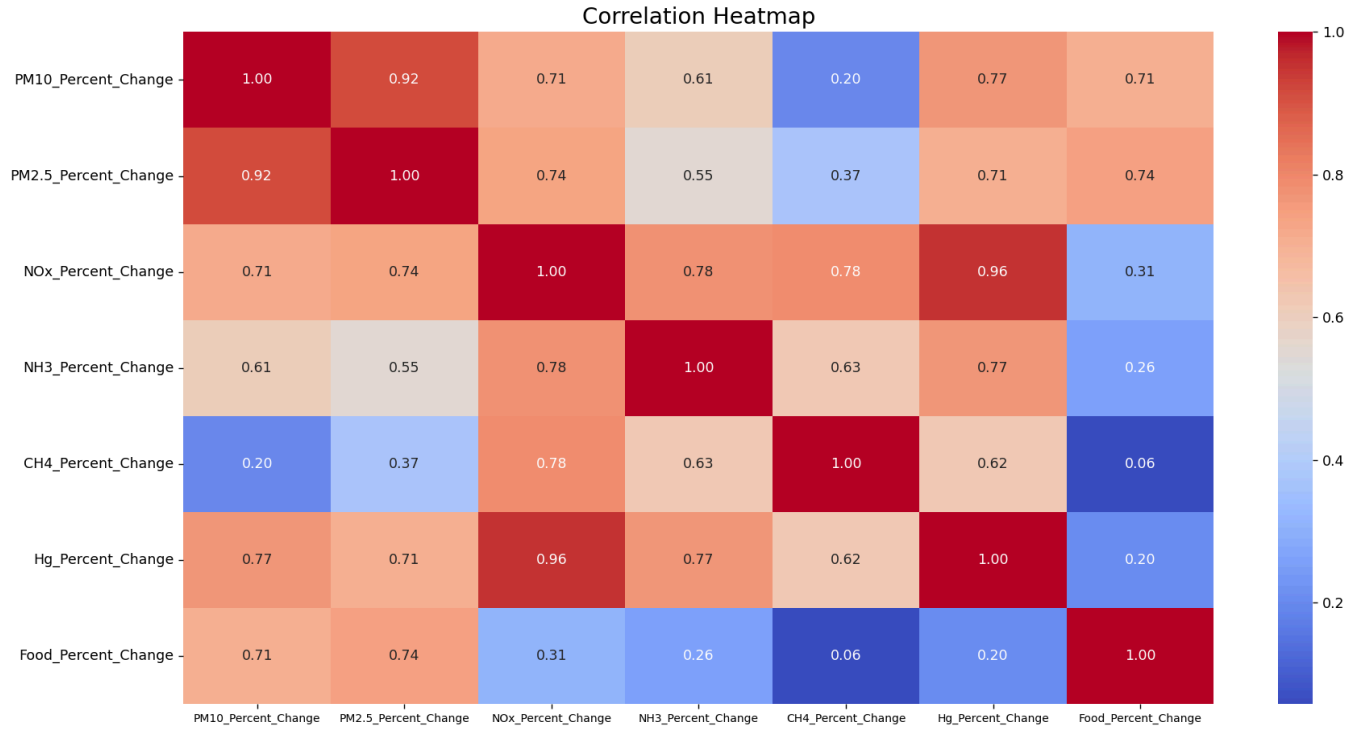
Some key observations from these trends are that the North American countries have seen some improvements in their air quality with Mexico being the exception. The USA had a steady reduction in the amount of substances whereas Canada saw almost constant amounts. In contrast, South American countries have mostly seen significant degradation of air quality. It is worth noting that despite the decline, the total amounts of substances in the USA are still very high and the USA appears as one of the largest contributors to the total pollutant amounts for the whole of the Americas along with Brazil partly due to their massive land areas. But on the other hand, Canada despite its huge land area has significantly smaller total pollutant amounts.

To answer the 2nd part of the project question which is to identify the causes and reasons for such trends, the final data frames for each data set are saved as a .csv file and then imported into a script called final_analysis.py. Here all the data frames are merged and a correlation matrix is computed to come to conclusions. For better understanding, a heat map is plotted for visualization. For each of the air quality metrics, a strong correlation can be observed with all but 1 pollutant. Thus, the trends in the air quality across Americas can indeed be associated with the pollutants that we picked in the research. But to better understand this, we need more information on what each pollutant is and how it is primarily produced in the first place to begin with.

1. **Methane (CH_4):** It contributes to the formation of secondary particulate matter through atmospheric reactions with energy extraction, coal mining, waste decomposition, etc as the main sources.
2. **Ammonia (NH_3):** It reacts with acidic gases (e.g., SO_2, NOx) to form secondary PM 2.5 and PM 10. Its main sources are decomposition of synthetic fertilizers, organic matter and industry releases.
3. **Mercury (Hg):** Mercury adsorbs onto particulate matter and its primary sources are coal combustion, gold mining, industrial processes, waste incineration and volcanic activities.
4. **Food emissions:** Cooking releases primary PM, while food waste decomposition contributes to secondary pollutants through gaseous emissions.

Some key observations from the correlation matrix are that for particulate matter PM 2.5 and PM 10, mercury and food emissions are the largest contributors. For NO_x emissions, mercury, methane and ammonia releases are the most prominent contributors. Moreover, the 3 air quality metrics that we chose for this research are also strongly correlated so that increase/decrease in air quality is mostly equally associated with them. Lastly, our selected pollutants that affect the air quality don't show strong correlation among each other which is a good sign because

that means that they are independent. Our statistical analyses and modeling efforts are hence more reliable, as they reduce the risk of multicollinearity because strongly correlated variables can distort results.



Conclusion

This research has given us a platform to explore the trends in the air quality across the major countries in the Americas. We were able to see its shifts over the past 50 years by looking directly at the concentration amounts of PM 2.5, PM 10 and NO_x in the air. Generally, North American countries have either seen small declines or some improvements in air quality over the observed time period whereas South American countries have seriously struggled and their air quality has significantly worsened. USA from the north and Brazil from the south, have emerged as the largest contributors to the total release of all the substances in the Americas so they need to pay serious attention to this alarming situation. Appropriate steps need to be taken to cut down the emissions of pollutants like Mercury, Ammonia, Methane and food as they have shown strong correlations with the air quality metrics. Independence among the pollutants suggests a diversity of emission sources, providing multiple avenues for intervention. Policymakers can implement parallel strategies across sectors (e.g., agriculture, transportation, industry) without concern for unintended interactions. It presents broader mitigation opportunities.

However, there have been some inevitable limitations and potential gaps in the research. Some other important pollutants like sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), etc. had to be left out because of the limited scope of the project. The detailed analysis could only be highlighted on a select few countries, potentially missing out on some interesting insights like countries with highest percent increase, lowest average, etc. The analysis could also not account for economic growth, industrialization, urbanization, or political factors influencing emission trends.