Data Report

CHANGE IN AIR QUALITY ACROSS THE AMERICAS & THE PRIMARY CONTRIBUTORS

Muhammad Umar Naeem | Methods of Advanced Data Engineering | 11-11-2024

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 identifying 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.

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.

Open Data

All the datasets are taken from EDGAR - Emissions Database for Global Atmospheric Research. EDGAR is a multipurpose, independent, global database of anthropogenic emissions of greenhouse gases and air pollution on Earth.

EDGAR doesn't explicitly provide any open data license but the reason why its data sets are considered open and available for public usage is that it is published by the **European Commission**, specifically the **Joint Research Centre (JRC)**, which is part of the European Union's scientific and knowledge service. As a public entity, the European Commission has a general policy of making its research and data openly accessible to promote transparency and scientific research. The Commission further recommends all EU countries to put publicly funded research results in the public domain to strengthen science and the knowledge-based economy.

Source where openness of data from EU commission initiatives like EDGAR is discussed:
https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science/open-access_en

Fig: EDGAR – an EU Commission Research Initiative

Choice of Data

The datasets from **EDGAR** are already well maintained. They don't have missing values, and no additional pre-processing is required. Another option is the datasets from the **World Health Organization** (WHO). But I found that their datasets lack complete data, and they are also not that practical to work with. For example, in the context of determining the trends in the air quality recently say past 10 years, WHO will have separate versions of the same dataset, with each version containing data at a particular time whereas **EDGAR** has all the yearly data in a single excel workbook which would require very minimal effort for data analysis.

Fig: WHO Air Quality DB

C_group_IM24_sh	Country_code_A3	Name	Substance	Y_1970	Y_1971	Y_1972	Y_1973	Y_1974	Y_1975	Y_1976	Y_1977	Y_1978	Y_1979	Y_1980	Y_1981
Rest Central America	ABW	Aruba	PM2.5	0.0298345	0.0291749	0.026632	0.0290004	0.0277859	0.0281773	0.0260948	0.0286491	0.0284992	0.0292331	0.0304082	0.0297574
India +	AFG	Afghanistan	PM2.5	6.8653409	6.7532607	7.6714027	7.8283287	8.0274617	8.3420017	8.5038925	8.1616985	8.4103095	8.1892006	7.9628301	7.8389663
Southern_Africa	AGO	Angola	PM2.5	48.753503	48.567962	49.430089	49.501251	49.364778	49.757714	50.2313	49.99725	51.615445	51.775474	52.520121	52.822178
Rest Central America	AIA	Anguilla	PM2.5	0.000614	0.0006178	0.0006857	0.0006423	0.0007356	0.0009761	0.0008682	0.0009127	0.0010128	0.0012308	0.0013108	0.0022467
Int. Aviation	AIR	Int. Aviation	PM2.5	10.763955	10.763955	11.414023	11.903051	11.453669	11.076072	11.095375	12.09862	12.533686	12.964181	12.887864	12.848823
Central Europe	ALB	Albania	PM2.5	12.680422	12.67682	12.886512	13.135104	13.264342	13.380787	13.645093	13.872106	14.589345	15.582309	15.580462	15.214506
Rest Central America	ANT	Netherlands Antilles	PM2.5	2.9868713	2.9972378	2.9467014	2.9889617	2.6956376	2.0525917	2.3508619	2.3598961	1.8913021	2.0626867	1.7389975	1.6450979
Middle_East	ARE	United Arab Emirates	PM2.5	0.8481008	0.9226005	1.0925643	1.3845522	1.5542959	2.1871335	2.9202478	3.5285075	3.826772	4.684543	8.237113	9.8066433
Rest South America	ARG	Argentina	PM2.5	101.12599	104.36058	100.93061	106.0181	104.7003	104.78497	106.36387	103.79587	110.00918	113.17479	104.15289	113.23962

Fig: EDGAR PM 2.5 Dataset

Data Pipeline

Python is used for data pipeline. The requests library is used to end http requests to fetch data. All the download urls give a zip file with the dataset in the excel format. The built-in python library `**zipfile**` is used to unzip the response from the get requests & then the excel file that is the dataset is found & saved in the `data` dir. The **tenacity** library is used to build a retry mechanism in case there are issues in fetching data from the links & saving it locally.

Since there are 7 datasets to be used in the project, their corresponding download urls have been placed in a links folder inside the project dir based on their category of use.

The script `**data_extraction.py**` takes care of automating the data extraction process by automatically fetching the urls from the `links` folder & then saving them to the `data` dir.

The datasets are already cleaned & ready for use out of the box for the purpose of this project so no additional pre-processing is required. The datasets do contain additional data which can be removed in this data pipeline but the organization of data in the datasets allows for the additional data to be ignored instead of wasting time and effort in removing it. During the analysis, we can simply pick the data that we need from the datasets.

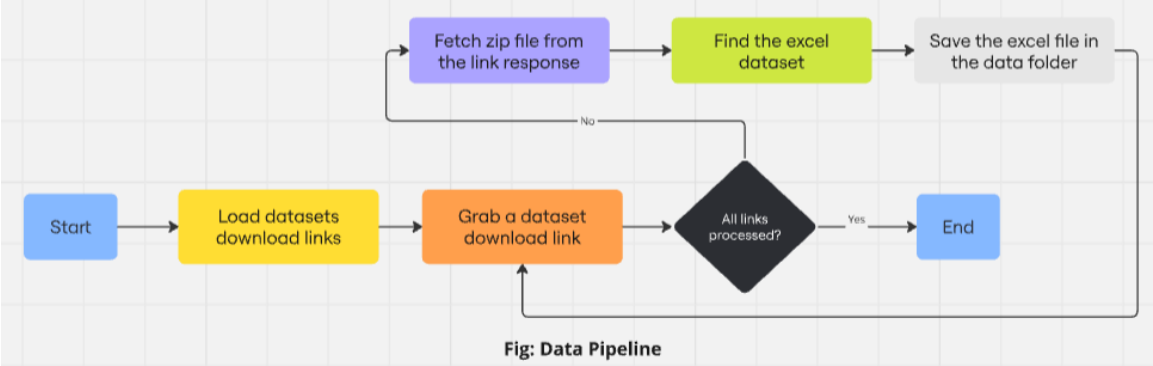

Fig: Data Pipeline

A retry decorator is added around the function that sends the get request with the download links, finds the zip file, unzips it & saves the excel format-based dataset in the data folder.

Result of the Pipeline

The result of a successful pipeline run is the local availability of all the datasets. All of them are excel based files found in the data folder. Each excel file contains several sheets, we will be using the final sheet which contains the totals for each country.

Excel files are quite easy to use with the **Pandas** library in **Python**. A sheet of choice can also be selected very easily.These sheets are fully labelled and ready for use straight away. The data is structured, accurate, complete, consistent, timely and relevant in the context of the project. Due to the data being conformant to these dimensions, there are no known limitations in the data.

The potential issues in the final report could be stumbling upon unusual trends in air quality that may not be easily explainable with pollutants' emissions. These unusual trends can be because of meteorological factors (temperature inversions, seasonal variations), wildfires, volcanic eruptions, urbanization, climate changes, etc. which are out of the scope of the project.