AIR POLLUTION, CLIMATE CHANGE, AND OUR HEALTH

by

KATHIA VARGAS

A master’s capstone submitted to the Graduate Faculty in Liberal Studies in partial fulfillment of the requirements for the degree of Master of Science, The City University of New York

2022

Air Pollution, Climate Change, and Our Health

by

Kathia Vargas

This manuscript has been read and accepted by the Graduate Faculty in Liberal Studies in satisfaction of the capstone requirement for the degree of Master of Science.

____________________ Date	____ Aucher Serr Thesis Advisor
____________________ Date	____ Matthew Gold Executive Officer

THE CITY UNIVERSITY OF NEW YORK

Abstract

Air Pollution, Climate Change, and Our Health

by

Kathia Vargas

Advisor: Aucher Serr

Climate change is a subject that is creating a lot of controversies nowadays. From newspapers to researchers, there are big efforts going on trying to bring awareness about the effects of air pollution and climate change over time. It is recommended that governments all over the world, and people from all communities act by taking care of the environment because the situation might turn out to be irremediable. There is a quote by Leonardo Dicaprio stating, *"Climate change is real. It is happening right now; it is the most urgent threat facing our entire species and we need to work collectively together and stop procrastinating.* (DiCaprio)*"* The words in this quote express the reason why I decided to research this subject for my capstone.

My goal with this research is to explore air pollution, climate change, and other sources of data to try to understand more about the effects of climate change over time. Also, I aim to explain the basic concept of the Air Quality Index and its relationship with climate change to educate an audience that might be interested in learning more about this subject.

Table of contents

List of figures

Digital manifest

I- Dissertation Whitepaper (PDF)

II- WARC Files

a) Project Website:

Archived version of https://kathiavf16.github.io/ms_final/.

III- Code and other deliverables:

a) Zip file containing the contents of the GitHub repository at the time of deposit (https://github.com/kathiavf16/ms_final)

A Note on Digital Specifications

This website can run on any server since it was built using HTML, CSS, and JavaScript. The visualizations on the website were built using d3.js and Datawrapper. All the visualizations, except for the interactive heatmap, were where created using Datawrapper. The heatmap was created using d3.js and the version 6 is recommended. The website files can be found in this GitHub repository: https://github.com/kathiavf16/ms_final

CHAPTER 1: INTRODUCTION

Before starting with this research project, my concept of climate change was somewhat limited to the thought that climate change was mostly about global warming due to the negligent attitude of governments and individuals by taking for granted the importance of caring for the environment. After putting this capstone project together, I have learned there are many factors that need to be understood to get the full picture of the impact of this phenomenon. First, I would like to start with a proper definition of climate change. According to the National Geographic Society, "Climate change is the long-term alteration of temperature and typical weather patterns in a place. Climate change could refer to a particular location or the planet as a whole. Climate change may cause weather patterns to be less predictable" (Thiessen). "Climate change has also been connected with other damaging weather events such as more frequent and more intense hurricanes, floods, downpours, and winter storms" (Thiessen). By reading this definition, it is clear that the climate change subject can be complex to digest. There is the aspect of understanding concepts such as temperature, weather patterns, and other natural behaviors, but what causes climate change? The National Geographic Society states that "the cause of current climate change is largely due to human activity, like burning fossil fuels, like natural gas, oil, and coal" (Thiessen).

If there is plenty of information indicating climate change is largely due to human activity, why haven't the powerful nations taken drastic decisions to remediate this situation? One of the main points I would like to address in this research is the negative impact of climate change on the health of the current and future generations. The World

Health Organization (WHO) states, “climate change is impacting human lives and health in a variety of ways. It threatens the essential ingredients of good health - clean air, safe drinking water, nutritious food supply, and safe shelter - and it has the potential to undermine decades of progress in global health” (“Environment and health”). This statement alarms me because according to this definition, climate change is the main threat that might end the possibility of a habitable environment for humans on earth. The fact that climate change is impacting the quality of clean air is one of the most concerning from my perspective. Humans cannot live for long without breathing air. The thought of inhaling polluted air throughout a lifetime is scary.

To have a better understanding of the negative impact of air pollution first let’s define what it is. The National Geographic Society states that “air pollution consists of chemicals or particles in the air that can harm the health of humans, animals, and plants. Pollutants in the air take many forms. They can be gases, solid particles, or liquid droplets” (Muegel). “Pollution enters the Earth's atmosphere in many different ways. Most air pollution is created by people, taking the form of emissions from factories, cars, planes, or aerosol cans. Second-hand cigarette smoke is also considered air pollution” (Muegel). As described with climate change, air pollution is also mostly generated by people. It’s worth mentioning The WHO claims “air pollution poses a major threat to health across the globe. Almost all of the global population (99%) are exposed to air pollution levels that put them at increased risk for diseases including heart disease, stroke, chronic obstructive pulmonary disease, cancer, and pneumonia” (“Air pollution data portal”).

Undoubtedly, 99% of the global population is a large number, according to this data, mostly everyone is exposed to polluted air. The purpose of putting together these definitions was to explain the reason why the title of this capstone project is “Air Pollution, Climate Change, and Our Health.” I believe it is quite important for everyone to have a clear understanding of the meaning and relationships between climate change, air pollution, and our health.

CHARTER 2: EXPLORATORY DATA ANALYSIS

Before using any publicly available data, it is a good practice to perform an Exploratory Data Analysis (EDA). According to IBM, “The main purpose of the EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, and find interesting relations among the variables” (IBM).

Software engineer Kaushik Katari explains, “EDA uses data visualization to draw meaningful patterns and insights. It also involves the preparation of data sets for analysis by removing irregularities in the data” (Katari). With the purpose of the explanation above, I have used the EDA approach to source, clean, analyze, and identify the critical patterns of multiple datasets compiling a meaningful dataset to support this capstone project.

- **DATA SOURCING AND CLEANING**

To perform the data cleaning and processing, I picked an array of tools and used them according to my needs. The main tool I used was Python within the Anaconda Navigator and Jupyter Notebooks. I chose Python because it is one of the more powerful tools available, as of now, to perform robust EDA. Some of the libraries I used are Pandas for data exploration and analysis, Numpy which supports Pandas by adding more features to deal with complex multidimensional arrays, and Matplotlib and Seaborn for data visualizations.

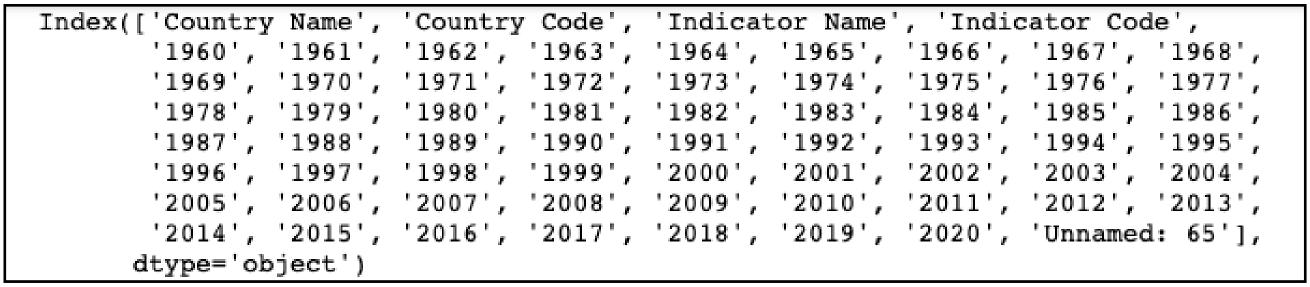
Using a reliable datasource for this project was very important to me because of the importance of the climate change subject. Reading the dataset overview before using it is always a good practice because it helps you to clear any concerns you might have regarding the data. The first data source I worked with was the World Development Indicators dataset from The World Bank Data Catalog. After researching, I concluded The World Bank Data Catalog is one of the most reliable data sources to consult when looking to get accurate information about climate environmental emissions. The World Bank Group is a very well-known organization whose mission is committed to reducing poverty and rising to the challenges of climate change. The World Bank states "the primary World Bank collection of development indicators was compiled from officially-recognized international sources. Also, they indicate this dataset presents the most current and accurate global development data available and includes national, regional, and global estimates" (The World Bank Data Catalog).

When downloading the data, The World Bank Data Catalog offers several format options to choose from. For this project, I selected the .csv format because it is the most commonly used when working with the Pandas library. The download .zip folder contained six .csv files. From these six files, I only worked with 3 which are the

WDIData, the WDISeries, and the WDICountries.

The first step in the data cleaning and processing was to look at the shape of the data. The first dataset I loaded was the WDIData assuming it contained the most important information about the emissions. Using the function pandas.info(), which returns the basic info about the dataset, resulted in 383,838 rows and 66 columns making me realize the WDIData was a very large dataset. Then, to have a better overview of the features of the data, I called pandas.columns() returning the following columns:

graph 1: columns in the WDIData dataset

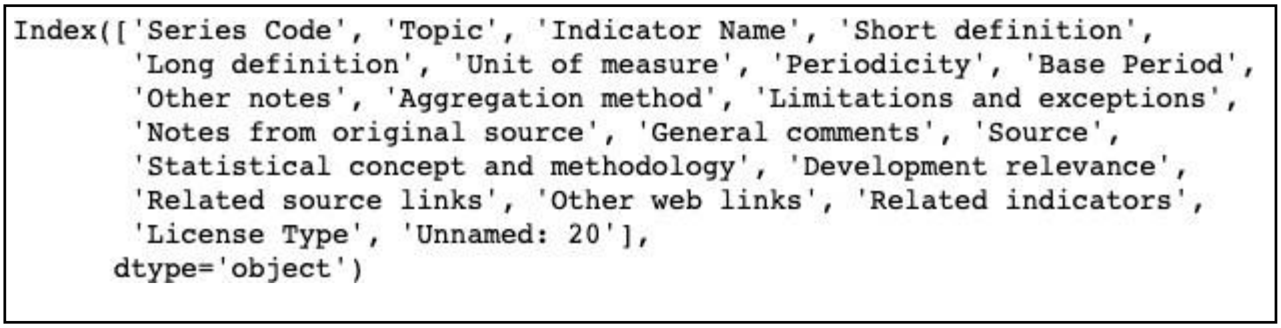


```
Index(['Country Name', 'Country Code', 'Indicator Name', 'Indicator Code',
       '1960', '1961', '1962', '1963', '1964', '1965', '1966', '1967', '1968',
       '1969', '1970', '1971', '1972', '1973', '1974', '1975', '1976', '1977',
       '1978', '1979', '1980', '1981', '1982', '1983', '1984', '1985', '1986',
       '1987', '1988', '1989', '1990', '1991', '1992', '1993', '1994', '1995',
       '1996', '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004',
       '2005', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013',
       '2014', '2015', '2016', '2017', '2018', '2019', '2020', 'Unnamed: 65'],
      dtype='object')
```

By looking at the columns, it was clear the WDIData dataset was composed of data from the years 1960 to 2020. Also, it is important to note the columns “Country Name”, “Indicator Name”, and “Indicator Code”. This result was positive, but it was clear more information was missing, and I needed to use another dataset to make sense of this data puzzle.

The second dataset I used was the WDISeries. When calling pandas.info() the result was a total of 1,429 rows and 21 columns. Using pandas.columns() this was result:

graph 2: columns in the WDISeries



```
Index(['Series Code', 'Topic', 'Indicator Name', 'Short definition',
       'Long definition', 'Unit of measure', 'Periodicity', 'Base Period',
       'Other notes', 'Aggregation method', 'Limitations and exceptions',
       'Notes from original source', 'General comments', 'Source',
       'Statistical concept and methodology', 'Development relevance',
       'Related source links', 'Other web links', 'Related indicators',
       'License Type', 'Unnamed: 20'],
      dtype='object')
```

Finding the columns “Series Code”, “Topics”, and “Short Definition” was exactly what I was looking for because they provided the details I needed to complement the previous dataset.

For the next step, since I was interested in working with climate-related data, I decided to create a filtering data frame containing only the indicators with the keyword “Environment” in the Topic column. Filtering the data by the “Environment” topic was positive because it reduced the number of rows from 1,429 to 138, which made a significant difference. Then, the next step was filtering for topics containing the word “Emissions", and by doing this, the dataset was reduced to 42 rows from 138, making it a small and manageable filtering dataset.

Now that I had the filtering dataset ready, I was able to merge the WDIdata with the WDISeries dataset. For this final dataframe, I only kept the columns “Country Name”, “Country Codes”, “Indicator Name”, “Indicator Code”, “Year”, and “Indicator Value”. The shape of the dataset after the merging was 676,368 rows and 6 columns. Then, after proceeding with dropping the NaN values, the dataset was reduced more than half to only 325,858 rows.

This large dataset contained information not only about the countries but also the world regions, which led me to proceed to separate the data in two dataframes "Countries" and "Regions". This decision was beneficial in aspects I would discuss further in the paper.

Taking into account the focus of this research is mostly about the harmful effects of air pollution, it was my goal to only analyze air pollution-related indicators. For this purpose, I used the five major indicators from the Air Quality Index (AQI). The AQI is the index for reporting air quality, and it is composed mostly of the following five pollutants: ground-level ozone (O3), particulate matter (PM2.5 and PM10), carbon monoxide (CO2), sulfur dioxide (SO2, and nitrogen dioxide (N2O).

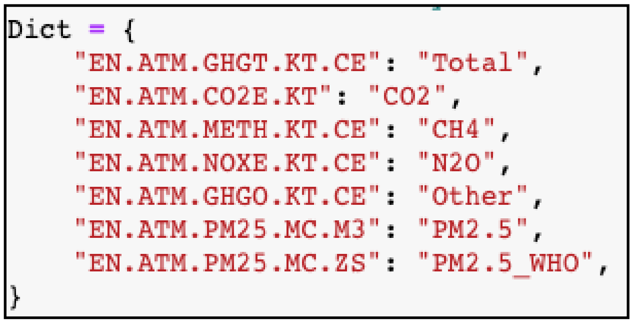
I will briefly define the five pollutants mentioned above. **Ozone**: According to the US Environmental Protection Agency (EPA), "Ozone can be good or bad, depending on where it is found. Ozone at ground level (O3) is a harmful air pollutant, because of its effects on people and the environment, and it is the main ingredient in "smog" ("Ground-level Ozone Basics | US EPA"). Unhealthy levels of ozone happen most likely on hot sunny days. People with asthma are at the greatest risk to be affected by breathing air containing high levels of ozone. **PM:** The EPA defines "*the PM term as a mixture of solid particles and liquid droplets found in the air. Some particles, such as dust, dirt, soot, or smoke, are large or dark enough to be seen with the naked eye. Others are so small they can only be detected using an electron microscope*" ("Particulate Matter (PM) Basics | US EPA"). "*The size of particles is directly linked to their potential for causing health problems. Small particles less than 10 micrometers in diameter pose the greatest problems because they can get deep into your lungs, and some may even get into your* *bloodstream. People with heart or lung diseases, children, and older adults are the most likely to be affected by particle pollution exposure"*(“Health and Environmental Effects of Particulate Matter (PM) | US EPA”).

The other three pollutants in the AQI fall into the group of greenhouse gas emissions which are the gasses that trap heat in the atmosphere. Here is the EPA definition: *“**Carbon dioxide ($CO2$)** is the primary greenhouse gas emitted through human activities. The main human activity that emits CO2 is the combustion of fossil fuels (coal, natural gas, and oil) for energy and transportation, although certain industrial processes and land-use changes also emit CO2. **Methane (CH4)** is emitted during the production and transport of coal, natural gas, and oil. Methane emissions also result from livestock and other agricultural practices, land use, and the decay of organic waste in municipal solid waste landfills. **Nitrous oxide (N2O)** is emitted during agricultural, land use, industrial activities, combustion of fossil fuels and solid waste, as well as during treatment of wastewater*” (“Overview of Greenhouse Gases | US EPA”).

Now having a basic concept of the AQI components, I will return to the data cleaning. My next step was to retrieve the necessary elements to create the emissions datasets that I would use for the data analysis. After looking deeply into the fields, I found supporting data for all the pollutants except for ground-level ozone (O3). I believe the reason why there is no data for O3 “is because O3 is not emitted directly into the air, but is created by chemical reactions between oxides of nitrogen (NOx) and volatile organic compounds (VOC)” (“Ground-level Ozone Basics | US EPA”). It would have been ideal to find information for the five components, but I considered the data available was sufficient to support the purpose of the analysis.

To filter the data, I created a dictionary containing the emission indicators below.

graph 3: emission's dictionary



```
Dict = {
    "EN.ATM.GHGT.KT.CE": "Total",
    "EN.ATM.CO2E.KT": "CO2",
    "EN.ATM.METH.KT.CE": "CH4",
    "EN.ATM.NOXE.KT.CE": "N2O",
    "EN.ATM.GHGO.KT.CE": "Other",
    "EN.ATM.PM25.MC.M3": "PM2.5",
    "EN.ATM.PM25.MC.ZS": "PM2.5_WHO",
}
```

After some more data cleaning and data exploration, I was able to create two datasets, one with the greenhouse gas pollutants (N2O, CO2, CH4), containing data from 2018, the latest year provided in the dataset, and the other with the particulate matter indicators, specifically, the PM2.5 mean annual exposure, and PM2.5 indicating the percentage of population exposed to levels exceeding WHO guidelines, with data from 2017 because there was not data available for 2018.

The fields included in the greenhouse gas dataset are "Country Name", "Income Group", "Region", "Sub-region", "Indicator Value", "Population", and "No. of Tonnes per Capita". To get some of these fields, I needed to get data from other data sources, and perform extra calculations. For instance, the "Income Group" data comes from the WDICountries dataset. The "Region" and "Subregion" data are from the ISO-3166-Countries-with-Regional-Codes dataset in this Github Repository, and the Population data is from the United Nation World Population Prospects online database.

I decided to include each country's population because I was interested in calculating the number of pollution tonnes generated per citizen. This idea was inspired by a visualization I found on the Visual Capitalist Website. This visualization served me as an inspiration and guide in the initial approach when thinking about how I wanted to visualize the data on the interactive site. Another important decision I made was to limit the number of countries in the analysis. I only kept 49 out of 192 countries showing the higher pollution indicator values. The other countries contained less than one percent of the values, making them irrelevant for the purpose of the analysis.

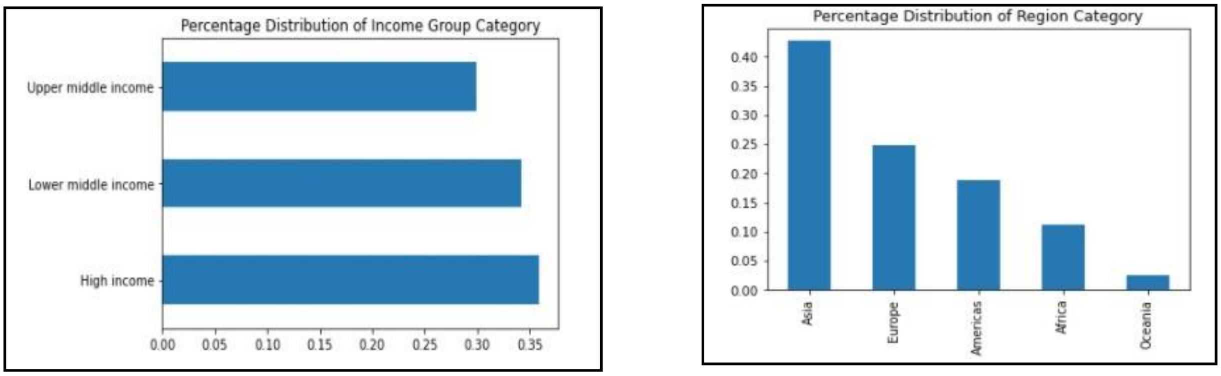
The particulate matter (PM) dataset is separated from the greenhouse gas pollutants because it contains different data types and it is from 2017, as opposed to the former that is from 2018. Also, this dataset has fewer columns containing only the "Country Name", "Indicator Name", "Indicator Value", "Region", and "Subregion" fields. In the next section, I would be discussing working with these cleaned datasets, and the exploring and finding of important patterns that would support the Air Pollution, Climate Change and Our Health interactive site.

- **STATISTICAL EXPLORATORY DATA ANALYSIS**

After putting the two datasets together, I proceeded to work on a new Jupiter workbook exclusively to focus on the statistical part of the exploratory data analysis (EDA). From now on, I will be explaining some of the techniques I used to analyze the data. This methodology was referenced from the article "Exploratory Data Analysis in Python" from the Towards Data Science website.

According to the author, there are different types of analysis. I have approached the three mentioned in the article, which are the univariate, bivariate, and multivariate. "The univariate analysis is when we analyze data over a single variable or column from a dataset. Referring to our data, the greenhouse gas and PM datasets only contain unordered variables. An unordered variable is a categorical variable that has no defined order" (Katari). The "Income Group" and "Region" are unordered fields. They are unordered because they are divided into subcategories, in the case of Region, we have Europe, North America, Africa, etc. There is no measure given to any value in this variable. When analyzing "Region" and "Income Group" of the greenhouse gas dataset columns this was the result:

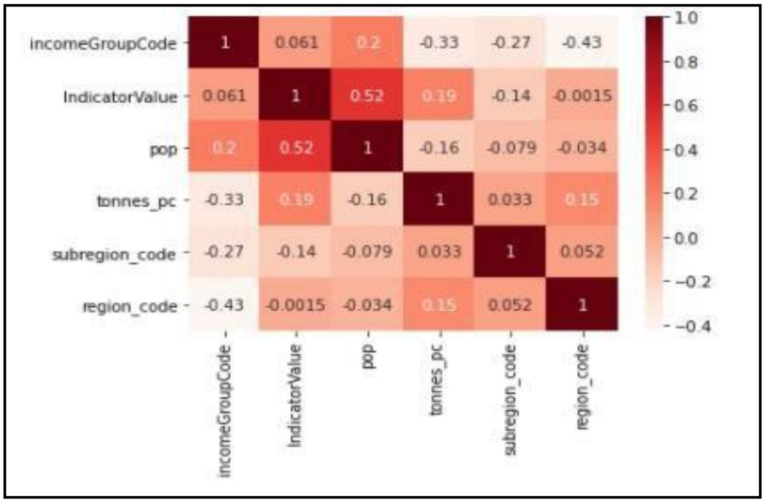
graph 4: Distribution of Income Group and Region



By the bar plots above, we can conclude that 35% of the 49 countries in the dataset fall into the high income category, and around 42% of them are located in Asia. I ran the same analysis for the PM dataset and it was slightly different, 30% of the 49 countries fall into the high income category, and most of them are located in the regions of Europe and Central Africa with 25%, and Sub-Saharan Africa with 24%.

As Kaushik Katari explains, "when we analyze data by taking two columns or variables into consideration it is known as Bivariate Analysis. **Numeric-numeric analysis**: Analyzing two numeric variables from a dataset is known as numeric-numeric analysis. We can analyze it in three different ways, using Scatter plots, Pair Plots and

Correlation Matrix " (Katari). My preferred way is using a correlation matrix. According to Tim Bock, “A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables" (Bock and Kelly). When running a correlation matrix for the greenhouse gas dataset this was the result:

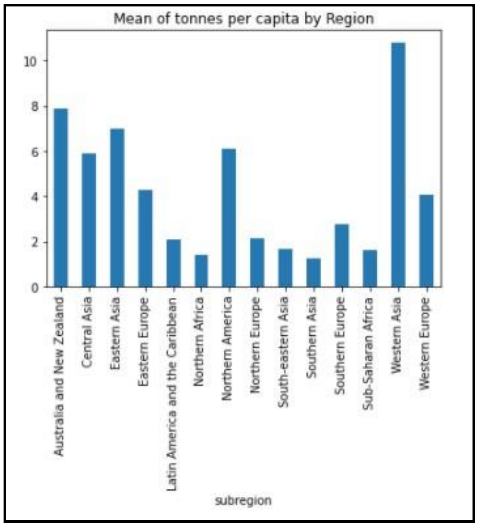


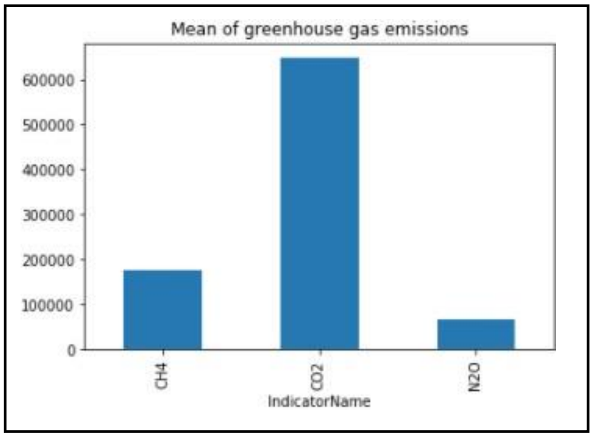
graph 5: correlation matrix for the greenhouse gas dataset

By scanning quickly through the matrix, it is clear there is a somewhat positive correlation between population and indicator value, which represents the total amount of emissions per country, with a coefficient of 0.52. The closer the value is to one, the stronger the positive correlation. For this result, I can interpret there is a somewhat strong positive correlation between the population and the total amount of emissions, but more research is needed before making any serious assumptions. **Numeric-categorical analysis:** “Analyzing one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyze them mainly using mean, median, and box plots” (Katari). To run the numeric-categorial analysis, I used the “Indicator

Name,” “Indicator Value,” and “Tonnes per Capita” fields to look for the mean values and this was the result:

graph 6: numeric-numeric analysis approach

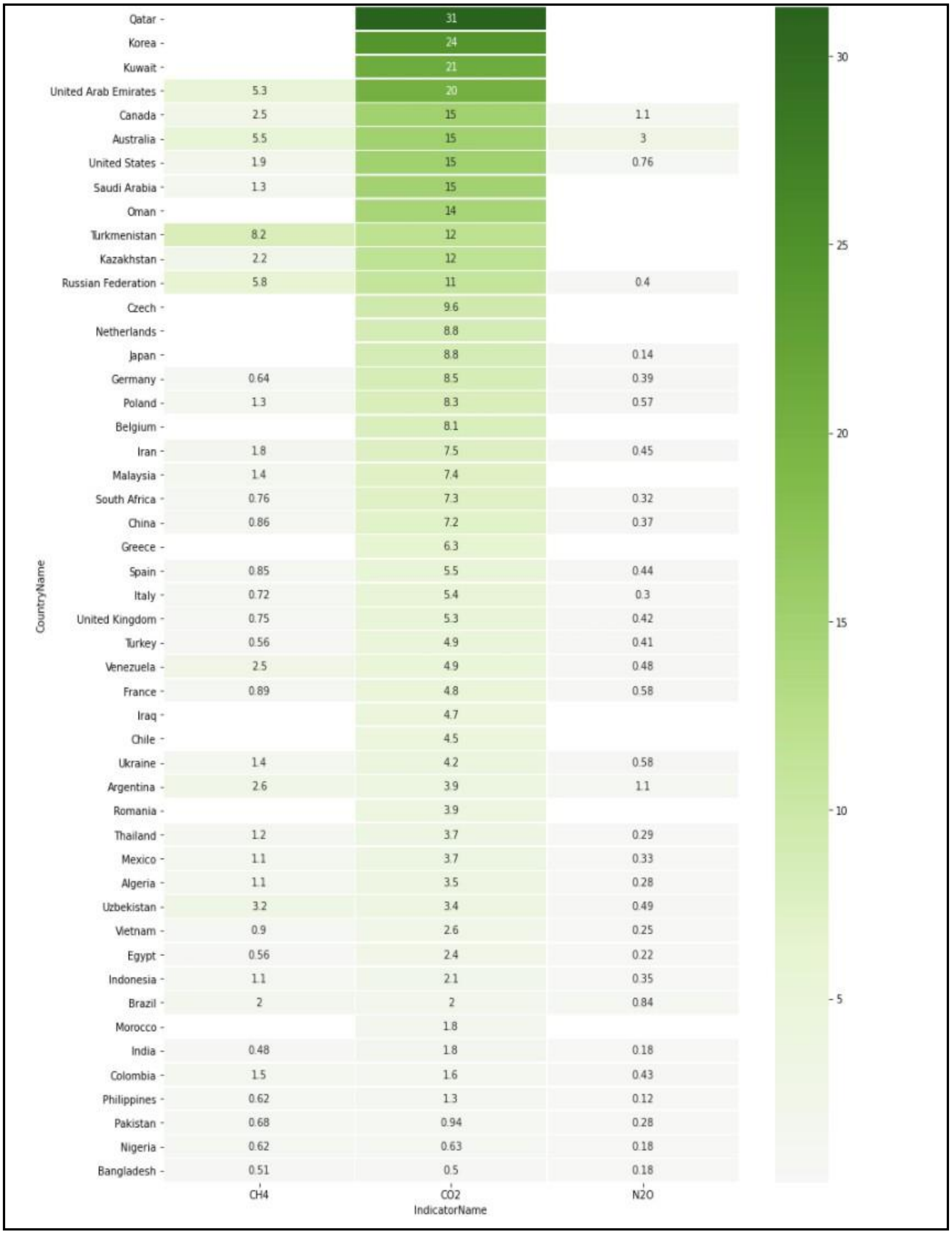




The chart from the left, plotted using “Tonnes per Capita " and “Subregion,” shows that Western Asia is the subregion with the highest mean of tonnes per capita. On the other hand, in the graph from the right, plotted using the “Indicator Name” and “Indicator Values,” shows CO2 is the most abundant pollutant of the three.

Lastly, to wrap up with the statistical data exploration, I will explain the multivariate analysis. “If we analyze data by taking more than two variables or columns into consideration, it is known as **Multivariate analysis**” (Katari). Let’s see how “Country Name”, “Indicator Value ", and “Tonnes per capita” vary with each other.

graph 7: multivariate analysis



By first creating a pivot table with these three columns, and creating a heatmap, the result was a very powerful visualization. By looking at this viz, I can see the tonnes per capita for the 49 countries broken down by the three components in the greenhouse gas dataset. I can interpret CO2 as having stronger values compared to the other pollutants. Also, the top four countries producing more CO2 tonnes per capita are Qatar, Korea, Kuwait, United Arab Emirates. The values of CH4 are stronger than the N2O values. Similarly, I can use the multivariable analysis to plot the PM dataset, and most likely I would get similar results. The detailed exploratory data analysis is available for both datasets in the Jupyter Notebook hosted on the capstone GitHub repository.

To conclude this chapter, this is how conducting an exploratory data analysis can help us to look at our data from different aspects and perspectives. The more we explore the data, the more insight we get from it. In the next chapter, I would be explaining the design and development of the Climate Change, Air Pollution, and Our Health interactive site.

CHARTER 3: DESIGN AND DEVELOPMENT OF THE WEB-BASED INTERACTIVE NARRATIVE

After taking the World Development Indicators dataset and doing the cleaning and exploratory data analysis, the next step was to share my findings with the world. To get to the final outcome of the website, it was not a straightforward process. I went through many design iterations, and even I changed my mind on how I wanted the website to flow after it was nearly implemented. In this section, I would be explaining in detail everything I went through during the development process.

First, I would like to start by explaining the technology used. My website was built based on HTML, CSS, and JavaScript. For the interactive visualizations I used d3.js, which is "a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS" (Bostock). I chose d3.js because it is one of the most flexible and powerful libraries to create interactive visualizations. d3.js is known for having a steep learning curve, but it is well documented and there are plenty of resources available to get you started. Another visualization tool I used is Datawrapper. Datawrapper is an online data visualization service that allows you to create beautiful charts with only a few clicks. They offer this service for free, and they claim their mission is to facilitate data visualization for everyone. I personally trust their site because well-known organizations such as Moody's Financial Services and *The New York Times* use their services. Here is the link for you to learn more about Datawraaper.

At first, my initial idea was a website having scrolling effects. After intensively looking at different scrolling libraries, I chose to try the "Page Scroll Effects" from the CodyHouse". This library was very user friendly and offered different scrolling styles by just adjusting simple CSS code. I was able to come up with a scrolling effect that would create an effect similar to an open book, transitioning in a way that when you scroll it, it would move to a different page by each section representing a new page. This effect was visually appealing at first, but when I started loading the visualizations and the text, the sections were too small compared to the size of the charts. Adjusting the layout was a little challenging because the scrolling effect was not flowing as smooth as before. After giving it some thought, I decided that implementing a new layout style was a better solution to this problem, and I moved on to searching for a static layout instead. I kept on searching, and I was not satisfied with the result, and I moved on to create my own layout from scratch. This was a good decision because when creating a layout from scratch you have more control on how the elements flow, and there is no need to learn or modify code as it is the case when working with a customizable template.

The layout design was simple and clean. It is composed of three main sections. The top navigation bar showing the project title, the left side navigation bar containing the important links pointing to sources such as the white paper, the Jupiter notebooks with the data cleaning and statistical EDA, the link to the Github repository, and some information about me including my LinkedIn profile. The third section is the body of the site containing the project information.

The design of the body is similar to reading an academic article. It starts with an introduction. Here I give a brief description of the goals and the audience of the project. Then, I move on to explain important information about the data sources, its associated links, and the definition of the AQI and the main five air pollution indicators.

To explain my findings, I start by giving a general overview about how the air pollution emissions have been increasing over time. To explain the growth of the emissions, I used an interactive line chart showing the growth from 1960 to 2018. Here we can clearly see that the CO2 emissions are the most abundant among this group, and have been increasing significantly fast over the last twenty years. The other pollutants also have been increasing over time, but not as fast as CO2.

To have a better understanding of the distribution, I decided to create a heatmap illustrating the total amount of emissions per country using the three greenhouse gas pollutants. I chose the heatmap because it offers a good visual perspective of all the data highlighting the important information. By using a heatmap is it nearly impossible not to get the right message, which in this case is identifying the countries with the highest emission values. On the left Y axis are the countries, and on the X axes are the pollutants (CO2, N2O, and, CH4), sorted in descending order starting with the highest CO2 emission, which in this case is China.

The heat map shows the total amount of emissions per country, but I think this is not a fair way to get to the conclusion of who is responsible for generating the most emissions. I believe calculating emissions per capita would give us a clearer picture because air pollution emissions are mostly generated by human action, and it is worth mentioning the result of the correlation matrix indicating that there is a positive correlation between the population and the total amount of emissions. When comparing the results calculating the tonnes per capita vs total of emissions, I get a different picture. As the heatmap shows, the three top countries with the highest amount of emissions are China, the United States, and India. However, the three top countries in terms of tonnes per capita are Qatar, North Korea, and Kuwait.

This result makes sense because the total amount of emissions is relative to the population. The larger the population, the larger the total amount of emissions. For instance, Qatar, North Korea, and Kuwait are very small countries compared to China, the United States, and India. Having such large numbers of tonnes per citizen is concerning taking into account the harmful effects of CO2. To visualize this result, I have plotted two dot charts, one showing the top ten countries by the total amount of CO2 emissions, and the top 10 countries by tonnes per capita.

Another aspect of the data I wanted to look at was the variation of the amount of emission generated per year. Are these countries making any progress in terms of reducing emissions? To get this result I looked at the year change for 2017 vs 2018. I only focused on the top ten and the bottom ten countries with the highest and lowest emission values referencing the 49 countries in the greenhouse gas dataset. The result was, from the top ten countries, Omar had the largest increase going from 0.52% in 2017 to 4.96% in 2018. On the other hand, Saudi Arabia had the largest reduction going from -2.88% in 2017 to -4.83% in 2018. When looking at the bottom ten countries, Nigeria had the largest increase from 4.15% in 2017 to 15.72% in 2018, and Brazil had the highest decrease going from 1.79% in 2017 to -4.92% in 2018.

Moving on to analyzing the PM dataset, I took a different approach by plotting a choropleth map. Choropleth maps "provide a way to visualize values over a geographical area, which can show variation or patterns across the displayed location" (The Data Visualization Catalogue). The shades of the map represent the percentage of the population exposed to levels of PM2.5 exceeding the WHO guidelines. I think this is one of the most impactful visualizations in the project because almost the entire map is red with most of the countries exceeding the WHO guidelines by 100%. This is very concerning considering how harmful PM2.5 emissions are for our health. Also, I included a table showing the percentage breakdown by region. According to this data, North America is the less polluted region in terms of PM2.5 with only 34% of the population being exposed to levels exceeding the WHO guidelines. It is worth mentioning that

Canada is the winning country with PM2.5 0% exposure. This concludes the data visualization part of the project.

In the next chapter, I am going to reflect about how this capstone project relates to the previous coursework I learned in the M.S in Data Analysis and Visualization program. Also, I will be evaluating how I feel about the result of this project, and the possible ideas I might add to the site in the future.

CHARTER 4: CAPSTONE PROJECT REFLEXION AND EVALUATION

Now that I have put together this exploratory data analysis and interactive site, I would like to reflect on how the knowledge acquired in the M.S in Data Analysis and Visualization program helped me to make this possible. Before pursuing this masters program, I already had previous knowledge of web design and data warehousing because I had a Bachelors in Computer Information Systems. However, I have to acknowledge pursuing this program was a good decision that has enhanced my career in many aspects. Before pursuing this program, I was feeling lost and insecure regarding my level of skill and knowledge in the information technology field compared to the competition, but now thanks to the knowledge acquired, that is no longer the case. The way the Data Analysis and Visualization program is structured is one of the positive aspects about it. The program requires 30 credits, of which 15 are elective, giving you enough room to choose other classes that you might be interested in. In my case, most of the classes I picked were data analysis and visualization oriented. This was a good decision because now I have expertise using the most relevant visualization tools in the industry. Before, I had no knowledge about d3.js or web interactivity using JavaScript. Thanks to the four data visualization courses I took which are Data Analysis Methods, Visualization and Design, Interactive Visualization, Advanced Data Visualization Studio I feel I am an expert in the subject.

Other courses in the program that made a big impact were Working With Data: Fundamentals, Data Analysis Method, and Advanced Data Analysis. I would say these three classes complement each other. In Working with Data, I learned the fundamentals of how to work data with Python using Anaconda and the Jupyter Notebooks. The Data Analysis Method class was very important because I learned about statistical analysis and hypothesis testing, which is a subject I enjoy working on as a result of taking the class. Lastly, in the Advanced Data Analysis class, the knowledge learned in the two previously mentioned classes was put in practice because it was about the basics of Machine Learning, which requires Python, Jupyter Notebooks, and statistical analysis.

As you can notice, the skills I have mentioned above were necessary to be able to put this capstone project together. Honestly, I feel very satisfied with everything I learned by pursuing this program. Another thing I would like to mention is that, before this master's degree, it was very difficult for me to write detailed long essays due to English not being my first language, but now even that aspect of my skill set has improved thanks to taking the Data, Culture, and Society class, which was intensive writing oriented.

Moving on to evaluating the final outcome of the capstone project, I have to say I feel very satisfied with what I have put together. While doing my research, I found very interesting information, and powerful visualizations regarding climate change, but nothing similar to what I have put together. I feel that this interactive website delivers the basic concepts on how climate change and air pollution are related to each other in a fun and interesting way. Taking the approach of using the AQI pollutants, it was a good idea because it gave me the guidance to only focus on one subject without going out of the scope, which I think can be very easy since the climate change subject is broad.

Using interactive visualization with d3.js and Datawrapper, helped in making the delivery of this subject more digestible in terms of improving the user experience. Also, there is the aspect of the datasets I have put together. I feel they might be useful to an audience looking for a simplified version of the air pollution emissions specifically. It was a lot of work going from the World Development Indicators dataset, to have the final greenhouse gas and the particulate matter datasets. By having the Jupyter notebook with the detailed exploratory data analysis available to the public, I think it might add value as well to an audience looking for guidance on how to conduct EDA using different approaches.

If I had more time to work on improving the current version of this project, I believe it might be useful adding a section addressing the impact climate change and air pollution is having on our health using interactive visualizations. For instance, extend the research to find data related to mortality rates or any other related harms air pollutants are causing. Doing this research with the purpose of making a stronger point, not only showing how emissions are increasing overtime, but also to show the actual facts of the negative impact of air pollution and climate change.

Works Cited

"Air pollution data portal." *WHO | World Health Organization*, https://www.who.int/data/gho/data/themes/air-pollution. Accessed 13 December 2021.

Bock, Tim, and Andrew Kelly. "What is a Correlation Matrix?" *Displayr*, https://www.displayr.com/what-is-a-correlation-matrix/. Accessed 21 December 2021.

Bostock, Mike. "d3.js." *D3.js - Data-Driven Documents*, https://d3js.org/. Accessed 22 December 2021.

The Data Visualization Catalogue. "Choropleth Map - Learn about this chart and tools to create it." *The Data Visualization Catalogue*, https://datavizcatalogue.com/methods/choropleth.html. Accessed 28 December 2021.

Data Viz Project. "Sunburst Diagram." *Data Viz Project*, https://datavizproject.com/data-type/sunburst-diagram/. Accessed 24 December 2021.

DiCaprio, Leonardo. "Leonardo DiCaprio Quote: Climate change is real." *Quotefancy*, https://quotefancy.com/quote/946624/Leonardo-DiCaprio-Climate-change-is-real-It-is-happening-right-now-It-is-the-most-urgent. Accessed 20 January 2022.

"Environment and health." *WHO | World Health Organization*, https://www.who.int/data/gho/data/themes/public-health-and-environment. Accessed 13 December 2021.

"Ground-level Ozone Basics | US EPA." *US Environmental Protection Agency*, 5 May 2021, https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics. Accessed 16 December 2021.

"Health and Environmental Effects of Particulate Matter (PM) | US EPA." *US Environmental Protection Agency*, 26 May 2021, https://www.epa.gov/pm-pollution/health-and-environmental-effects-particulate-matter-pm. Accessed 16 December 2021.

IBM. "What is Exploratory Data Analysis?" *IBM*, https://www.ibm.com/topics/exploratory-data-analysis. Accessed 19 December 2021.

Katari, Kaushik. "Exploratory Data Analysis(EDA): Python." *Towards Data Science*, 21 August 2020, https://towardsdatascience.com/exploratory-data-analysis-eda-python-87178e35b14. Accessed 19 December 2021.

Muegel, Trudy. "air pollution." *National Geographic Society*, 4 April 2011, https://www.nationalgeographic.org/encyclopedia/air-pollution/. Accessed 13 December 2021.

"Overview of Greenhouse Gases | US EPA." *US Environmental Protection Agency*, https://www.epa.gov/ghgemissions/overview-greenhouse-gases. Accessed 16 December 2021.

"Particulate Matter (PM) Basics | US EPA." *US Environmental Protection Agency*, 26 May 2021, https://www.epa.gov/pm-pollution/particulate-matter-pm-basics. Accessed 16 December 2021.

Thiessen, Mark. "Climate Change." *National Geographic Society*, 28 March

2019, https://www.nationalgeographic.org/encyclopedia/climate-change/. Accessed 13 December 2021.

The World Bank. *World Bank Group - International Development, Poverty, & Sustainability*, https://www.worldbank.org/en/home. Accessed 15 December 2021.

The World Bank Data Catalog. *World Development Indicators*. 2021. *World Development Indicators*, https://datacatalog.worldbank.org/search/dataset/0037712.