**Analysis of Calgary’s Historical Air Quality Data**

**Introduction**

The domain that we will be working with for this project is Environment and safety. Since the 1800s, anthropogenic factors have significantly contributed to the rise in pollution levels all over the globe. The increase in the production of heat trapping gasses (greenhouse gasses) over the last couple hundred years not only contributed to shifting temperature and weather patterns (climate change) but also to many airborne infections in humans and animals. Air quality data has been continually collected in Calgary from the 1980s by Alberta Environment and Parks to monitor the various parameters observed in the atmosphere and how they changed over the years. Air quality is continually changing in specific areas so this dynamic variability makes tracking air quality very important to take health precautions and eco-friendly actions to protect ourselves and the environment. In our project, we would like to analyze historical air quality data to see how the production of greenhouse gasses varied over the last 39 years in the city of Calgary and draw correlations between various factors within the data set to understand if there are any noticeable patterns in the data.

**Guiding Questions**

1. How did the environmental parameters such as greenhouse gasses (Carbon Dioxide, Methane, Nitrogen Dioxide and Nitric Oxide), air quality index, wind speed and temperature change over the course of 39 years?

We chose to answer this question to understand if there is any significant pattern observed in air quality between the 1980s and 2010s i.e before and after the industrial revolution. We will also test if there is any overlap between Calgary vs. global air quality patterns.

2. Is there any correlation between temperature and greenhouse gasses and between wind speed and greenhouse gasses over the 39 years?

This question was chosen to understand if greenhouse gasses affect temperature or wind speed as these gas have been known to have an impact on such environmental factors

3. What is the distribution of pollutants throughout Calgary? And was there a significant difference in the level/distribution of pollution in the year 2019 compared to the year 1980?

This question was chosen to understand if there is any difference in the amount of pollution across the City of Calgary and if there is any particular part of Calgary that experiences the most amount of pollution. This will also allow us to gain insight on the extent of residential/commercial pollution from various communities.

**Dataset**  
The dataset we have chosen for our project is “Historical Air Quality” which was collected by the Calgary Region Airshed Zone and submitted to Alberta Environment and Parks (AEP). This information is publicly available and can be used from the City of Calgary’s Open Data Portal.  
This dataset is in a structural tabular format and is organized into nine columns namely Station Name, Date, Method, Parameter, Average Daily Value, Units, Location and Count. The station name is divided into five stations positioned in Calgary Southeast, Central-Inglewood, Central, East and Northwest. The air quality data was collected from 1980 to 2019 using Instrumental, Sharp, calculated and Tape-Sampler methods. From the different stations various environmental factors such as Relative Humidity, Outdoor Air Temperature, Wind Speed as well as pollutant factors such Non-methane Hydrocarbons, Ozone, Hydrogen Sulphide, Methane and Carbon Dioxide to name some were measured. The stations also measured the Air Quality Index to test the overall quality of air in that particular area every day. All of these parameters were measured daily and their Average Daily Value was recorded for 39 years in their respective units. Finally, the location column recorded the exact location from which the parameters are measured.

For this project, we aim to use the columns Station name, Date, Parameters (Carbon Dioxide, Methane, Nitric oxide, Nitrogen Dioxide, Temperature and Wind speed), Average Daily Value and Units to answer our guiding questions. The overall raw data shows 422,928 rows of data, we plan to clean this data and only extract the data that we will need to answer our questions.

**Tasks**

For the data wrangling tasks of this project, we will focus on three major parts such as data cleaning, feature selection and manipulating data frames. In the data cleaning step, we will first clean spam and noise data and remove identifying data. Then fix any structural data, for example, remove space and non character text, fix number and number signs. Since date is an important factor to our dataset, we will correct any date and time format issues. When processing feature selection, we will remove columns that are not used or related to the guiding questions by merging or splitting columns. For the last step to K, we will sample data randomly if needed, this action depends on the size of filtered data from feature selection. Then we generate data frames based on usage of different visualizations.

For the visualization tasks, we plan to have four types of charts (scatter plot, line chart, multiple subplots and geographical plots) to best represent the answers to our guiding questions. Line charts will be used to display how the different parameters measured changed over time. To further analyze this, we plan to find appropriate global patterns data to compare Calgary air quality with the global air quality specifically PM2.5, Temperature and air quality index. We will have scatter plots and geographical plots to compare and identify correlations under certain conditions and a line chart to observe the overall changes in the amount of pollution over 39 years. To compare changes of the same pollutants in different parts of Calgary, we will display multiple subplots in one figure and to observe the extent of correlation between temperature/wind speed and greenhouse gasses, we will be using scatter plots. Finally, to statistically analyze distribution of pollutants across Calgary, boxplot/violin plot and geographical plots will be used. Furthermore, to execute these tasks, we plan to use libraries such as JupyterLab Environment, Panda, Numpy, Matplotlib libraries and other libraries for geophysical visualization.

**Tasks for each team member**

Jannatul: Write up an introduction of the project report ; data visualization graphs and write up for the guiding question 3.

Meghana: Report write up, data wrangling and data visualization for question 2 and Conclusion write up.

Serena: Data wrangling(data cleaning and feature selection), manipulating data frame, visualizing data and report write up for guiding question 1.

**References**

City of Calgary’s Open Data Portal *Historical Air Quality* [online]. Available at: <https://data.calgary.ca/Environment/Historical-Air-Quality/uqjm-jxgp/data>

Anon, Microsoft. *Microsoft Support*. Available at: https://support.microsoft.com/en-us/office/top-ten-ways-to-clean-your-data-2844b620-677c-47a7-ac3e-c2e157d1db19 [Accessed September 18, 2022].

Hillier, W. et al., 2021. What is data wrangling? A complete introductory guide. *CareerFoundry*. Available at: https://careerfoundry.com/en/blog/data-analytics/data-wrangling/ [Accessed September 18, 2022].

Anon, 2022. Data wrangling. *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Data\_wrangling [Accessed September 18, 2022].

Anon, Guide to data cleaning: Definition, benefits, components, and how to clean your data. *Tableau*. Available at: https://www.tableau.com/learn/articles/what-is-data-cleaning [Accessed September 18, 2022].

Anon, Stackplots and streamgraphs#. *Stackplots and streamgraphs - Matplotlib 3.6.0 documentation*. Available at: https://matplotlib.org/stable/gallery/lines\_bars\_and\_markers/stackplot\_demo.html#sphx-glr-gallery-lines-bars-and-markers-stackplot-demo-py [Accessed September 18, 2022].

Anon, Create and use a choropleth map. *Create and use a choropleth map-ArcGIS Insights | Documentation*. Available at: https://doc.arcgis.com/en/insights/latest/create/choropleth-maps.htm [Accessed September 18, 2022].

Anon, Tutorials: Documentation: Arcgis developers. *Documentation | ArcGIS Developers*. Available at: https://developers.arcgis.com/documentation/mapping-apis-and-services/tutorials [Accessed September 18, 2022].

Dhankhar, A., 2019. Mapping with matplotlib, Pandas, Geopandas and basemap in Python. *Medium*. Available at: https://towardsdatascience.com/mapping-with-matplotlib-pandas-geopandas-and-basemap-in-python-d11b57ab5dac [Accessed September 18, 2022].

Anon, Single-page. *Plotly*. Available at: https://plotly.com/python/reference/#choropleth [Accessed September 18, 2022].

Anon, Creating multiple subplots using plt.subplots#. *Creating multiple subplots using plt.subplots - Matplotlib 3.6.0 documentation*. Available at: https://matplotlib.org/stable/gallery/subplots\_axes\_and\_figures/subplots\_demo.html [Accessed September 18, 2022].

Denchak, Melissa. “Greenhouse Effect 101.” NRDC, 16 July 2019, www.nrdc.org/stories/greenhouse-effect-101#:~:text=Bednarski%20for%20NRDC-.

Ren, D., 1970. Effects of global warming on Wind Energy Availability. *AIP Publishing*. Available at: https://aip.scitation.org/doi/10.1063/1.3486072 [Accessed September 19, 2022].