The Effect of Air Quality on Human Health

Completed for the Certificate in Scientific Computation and Data Sciences
Summer 2023

Amulya Cherian
Bachelor of Science and Arts in Neuroscience
College of Natural Sciences
University of Texas at Austin

[Supervising Faculty's Signature]

_____

Dr. Layla Guyot
Assistant Professor of Instruction
Statistics and Data Sciences

Abstract

The quality of the air that humans breathe is undoubtedly important and impactful on human health. The landscape, industries, population densities, and urban densities vary greatly across the United States, which can differently shape the impact of air quality on human health in different parts of the U.S. This project attempts to understand the relationship between air quality and human health in the U.S. For air quality, this project examines the Environmental Protection Agency site monitoring data for various parameters such as ozone, black carbon, carbon monoxide, nitric oxide, and more. For health, two datasets are utilized: one dataset by the Centers for Disease Control and Prevention hosts information about the prevalence of chronic health conditions and the other dataset by the National Institute of Health has information about the prevalence of respiratory cancers across the U.S. All three datasets are based on the environmental and health data of the year 2019. This project visualizes the data and utilizes machine learning techniques to discern the relationship between air quality and the prevalence of chronic health conditions. The machine learning techniques utilized for this project include multiple linear regression modeling, kNN modeling, K-means clustering, and principal component analysis. These machine learning models automate the analysis of large datasets while developing predictive models.

Introduction

Air pollutants such as Particulate Matter (PM), nitrogen oxide, sulfur dioxide, Volatile Organic Compounds (VOCs), dioxins, and polycyclic aromatic hydrocarbons are all deleterious to both the environment and human health (Manisalidis et al., 2020). Short-term health effects include Chronic Obstructive Pulmonary Disease (COPD), asthma, and wheezing. Long-term health effects associated with air pollution include diabetes, lung cancer, and cardiovascular issues. The quality of the air varies around the world, and pollutants are prominent in areas with high urban density and/or high rates of road emissions (Manisalidis et al., 2020).

Three datasets were utilized to explore the relationship between the quality of air and human health in the United States during the year 2019. The Annual Concentration by Monitor 2019 dataset by the Environmental Protection Agency consisted of concentrations of various pollutants including criteria gases and particulates at monitoring sites across the nation (Environmental Protection Agency, 2020). This project focused on pollutants such as Black Carbon, carbon monoxide, nitric oxide, nitrogen dioxide, ozone, PM10 - LC, PM2.5 - LC, and sulfur dioxide. The U.S. Chronic Disease Indicators (CDI) dataset by the Centers for Disease Control and Prevention provided 124 health indicators and questions pertaining to chronic diseases (Centers for Disease Control and Prevention). Since air pollution has associations with chronic diseases, this project focused on asthma, diabetes, COPD, chronic kidney disease, and cardiovascular disease from this dataset. Lastly, the State Cancer Profiles dataset by the National Institute of Health contained the incidence rates for espiratory cancers for each state (NIH, 2020). The primary cancers examined with this dataset are lung and bronchus cancers.

The personal prediction prior to the start of this project was that there would be a strong, positive relationship between states with high levels of air pollutants and chronic diseases such as asthma, lung and bronchus cancer, and COPD.

Materials and Methods

Data visualization, cleaning, and analysis along with machine learning techniques were conducted with Python. The air quality dataset was filtered for certain air pollutants: ozone, PM2.5 - LC, PM10 - LC, black carbon PM2.5, nitric oxide, nitrogen dioxide, sulfur dioxide, carbon monoxide, and lead PM2.5 - LC. This dataset also had various sample durations, so it was further filtered for only one-hour sample durations. Examination of the units variable led to the realization that each parameter (air pollutant) had different units, so each parameter was converted into micrograms per cubic meter.

The CDI dataset was filtered for the year 2019 to match the other two datasets. Then, it was filtered for various health conditions: asthma, diabetes, cardiovascular disease, Chronic Obstructive Pulmonary Disease (COPD), and chronic kidney disease. To further focus on one specific group, the CDI dataset was filtered for specific questions that pertained to the age-adjusted and overall prevalence of the chronic disease for adults who are at least 18 years old.

The state cancer dataset included various respiratory cancers such as the nose, nasal cavity, and middle ear cancers, larynx cancer, lung and bronchus cancers, and pleura cancer. This dataset was filtered for lung and bronchus cancers. Then, all three datasets were merged together so that each row contained a state name, mean concentrations of various air quality parameters (micrograms per cubic meter), and age-adjusted prevalence rates of various chronic health conditions.

Various data visualizations were conducted to better understand the three datasets. A histogram was produced from the air quality dataset, which reflected the distribution of concentrations of various air pollutants. For the CDI dataset and state cancer dataset, it was helpful to view a geographic distribution of the prevalence of the various chronic health conditions on a map of the United States. Histograms of the numeric variables in the merged dataset alerted the need for normalization of the data. Moreover, a heatmap colored correlation matrix allowed an easy visualization of the varying strengths of correlations between the different numeric variables.

To conduct machine learning techniques on the data, the merged dataset was normalized so the variables had a slightly more Gaussian distribution than before. Many of the air quality parameters had an abundance of NaN values, so the parameters were filtered to include only carbon monoxide, nitric oxide, ozone, and sulfur dioxide. Then, multiple linear regression was applied to understand how these four air quality parameters influence the prevalence of asthma, cardiovascular disease, diabetes, chronic kidney disease, COPD, and lung/bronchus cancer. The predicted values and distribution of residuals were both visualized on a density plot to understand the multiple linear regression model. Additionally, the R-squared values and F-statistic values were calculated for each model.

The next machine learning model conducted is the K-means clustering, which requires the standardization of data. Once the data was standardized and each point was assigned a cluster, the silhouette score was calculated and the clusters were visualized with a cluster plot. Then, dimensionality reduction with principal component analysis was performed to calculate the variation percentages for each principal component (PC) and plot the variation attributed to PC 1 and 2.

Lastly, k-nearest neighbor (kNN) modeling was conducted on scaled data. Each model reflected the four air quality parameters and one health condition. Once the kNN model was trained to predict prevalence rate values for the health condition, the classification report, confusion matrix, and accuracy score were all produced.

Results

The boxplot displayed in Figure 1 depicts the distribution of the various air quality parameters. There is a narrow range of values for parameters such as ozone, carbon monoxide, and black carbon. Parameters such as nitrogen dioxide appear to have a much larger range of values. Sulfur dioxide has several outliers that were kept upon further inspection.

Next, there are the geographic heatmap distribution of the prevalence rates of various health conditions: asthma, cardiovascular disease, chronic kidney disease, chronic obstructive pulmonary disease, diabetes, and lung/bronchus cancer. These heatmaps enable a quick understanding of which states and regions of the U.S. are particularly prone to the corresponding chronic health condition.

Figure 1

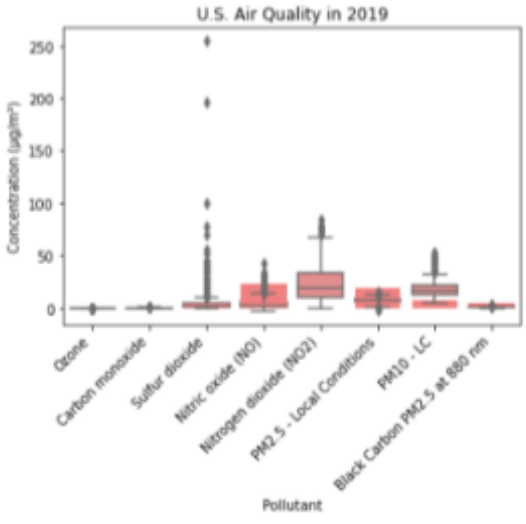


Figure 2

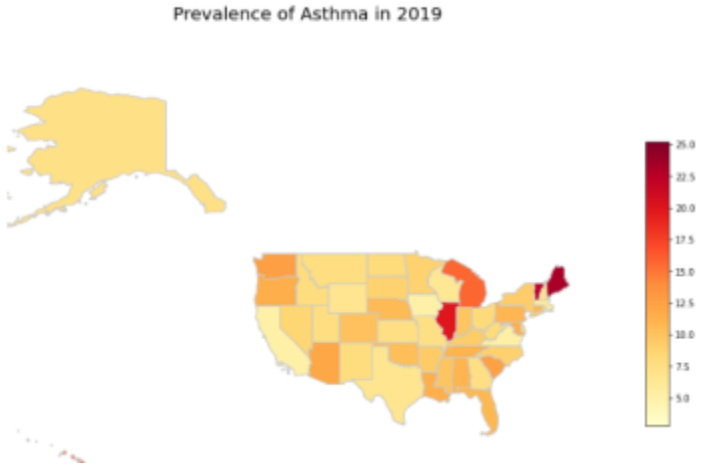


Figure 2 displays the distribution of prevalence rates for asthma among adults. The prevalence rates range from 3% to 25%. Visually, there isn't a strong pattern established among the states.

Figure 3

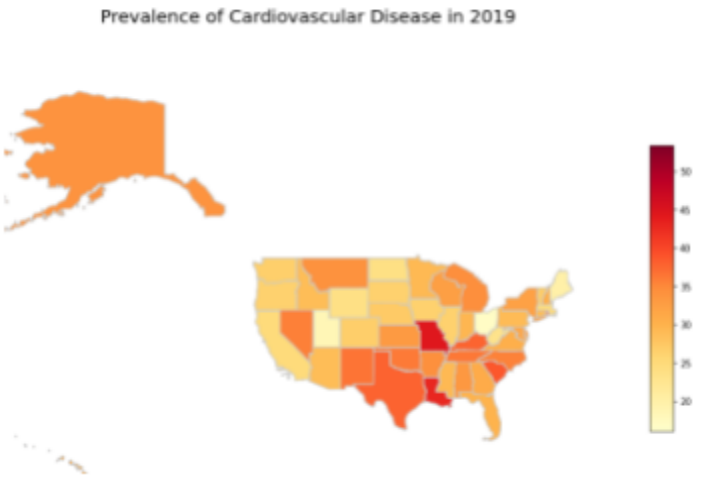


Figure 3 displays the distribution of prevalence rates for cardiovascular disease among adults. The prevalence rates range from 15% to 55%. It appears that the higher rates of cardiovascular disease are found in the southern region of the United States.

Figure 4

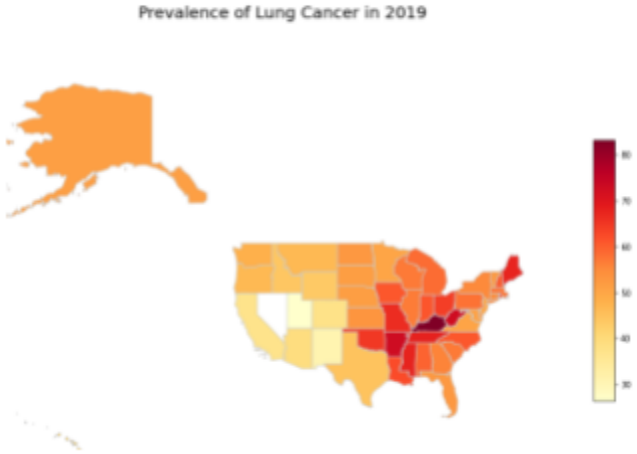


Figure 4 above displays the distribution of prevalence rates for lung cancer. The rate of lung cancer for Nevada is not included in the dataset, which is reflected in the heatmap above. States in the southern region appear to observe a high rate of lung cancer compared to the rest of the United States. The geographic heatmaps for chronic kidney disease, COPD, and diabetes did not possess a noticeable pattern of distribution. These heatmaps can be found in the attached document of computer code.

The correlation matrix below in Figure 5 exhibits the strength of the linear relationships between each variable. The lighter colors reflect a positive correlation, while the darker colors indicate a negative correlation between the two variables.

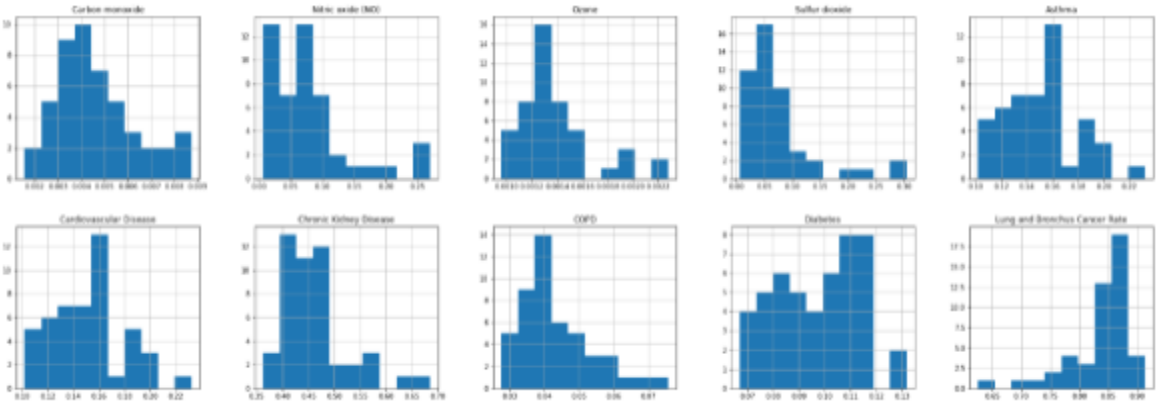
Figure 5

The strongest positive correlation between an air quality parameter and a chronic health condition is found between PM2.5 - LC and cardiovascular disease with a correlation value of

0.55. The weakest correlation value is -0.54, which is between cardiovascular disease and black carbon.

The air quality parameter variables were narrowed down to variables without NaN values, so that the machine learning techniques could be conducted with ease. These variables are carbon monoxide, nitric oxide, ozone, and sulfur dioxide.

Figure 6



The histograms in Figure 6 reflects the distribution of each variable after normalizing them. Despite the normalization of these variables, the distributions generally don't appear to be symmetrical. After normalizing the variables, multiple linear regressions were performed on each health condition variable with the air quality parameters as the independent variables. The output of these models such as the coefficients of each variable, y-intercept, and R-squared values are listed below in Table 1. The table also includes the F-statistic and its corresponding p-value for each MLR model.

Table 1

Output of the Multiple Linear Regression Models

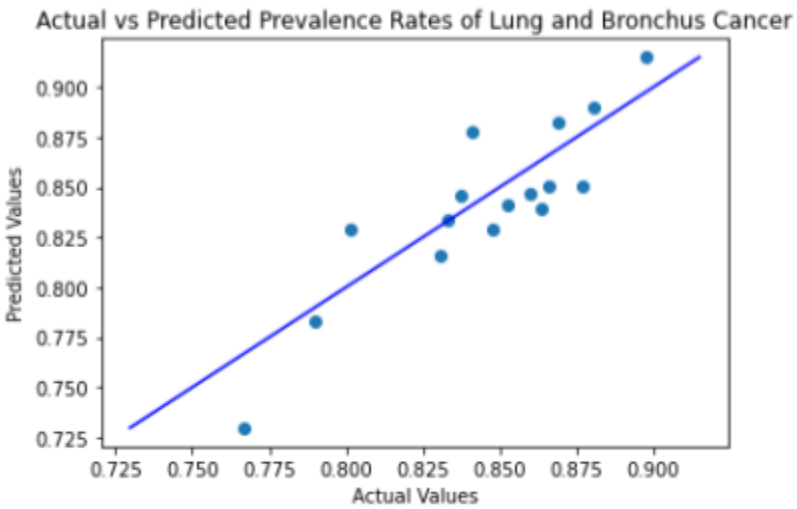
MLR Model #	Health Condition	Carbon Monoxide	Nitric Oxide	Ozone	Sulfur Dioxide	Intercept	F-statistic	R-squared
1	**Asthma**	0.258	0.037	57.05	-0.004	0.069	13.98	0.356
	p-values:	0.83	0.55	1.36e-13	0.53		1.94e-7	
2	**Cardiovascular Disease**	9.55	-0.11	169.61	0.045	0.189	41.67	0.759
	p-values:	0.11	0.66	1.46e-20	0.17		1.99e-14	
3	**Chronic Kidney Disease**	1.195	-0.018	17.04	0.006	0.015	1.56	0.329
	p-values:	0.29	0.60	1.04e-10	0.48		0.20	

4	**COPD**	-3.508	0.02	8.33	0.023	0.097	0.94	0.065
	p-values:	0.43	0.61	7.22e-10	0.16		0.45	
5	**Diabetes**	1.76	-0.04	48.17	-0.002	0.078	7.85	0.300
	p-values:	0.54	0.83	1.09e-12	0.47		7.28e-5	
6	**Lung/Bronchus Cancer**	-6.35	-0.08	-129.49	-0.188	1.067	47.29	0.777
	p-values	0.94	0.84	7.42e-9	0.31		2.22e-15	

The R-squared values for these six multiple linear regression models range from 0.065 to 0.777. The MLR model for lung/bronchus cancer has a R-squared value of 0.777, which indicates that this model is the strongest model fit among the six MLR models. The R-squared value for the MLR model for cardiovascular disease is 0.759. These two R-squared values indicate that the multiple linear regression models fit the data well. However, closer inspection of the independent variables indicate otherwise. For both cardiovascular disease and lung/bronchus cancer MLR models, the p-values for each variable are greater than 0.05 for carbon monoxide, nitric oxide, and sulfur dioxide. Ozone does exhibit a highly statistically significant p-value for both MLR models. Since most of these air quality variables are not statistically significant, the linear models appear to be a poor fit for the data.

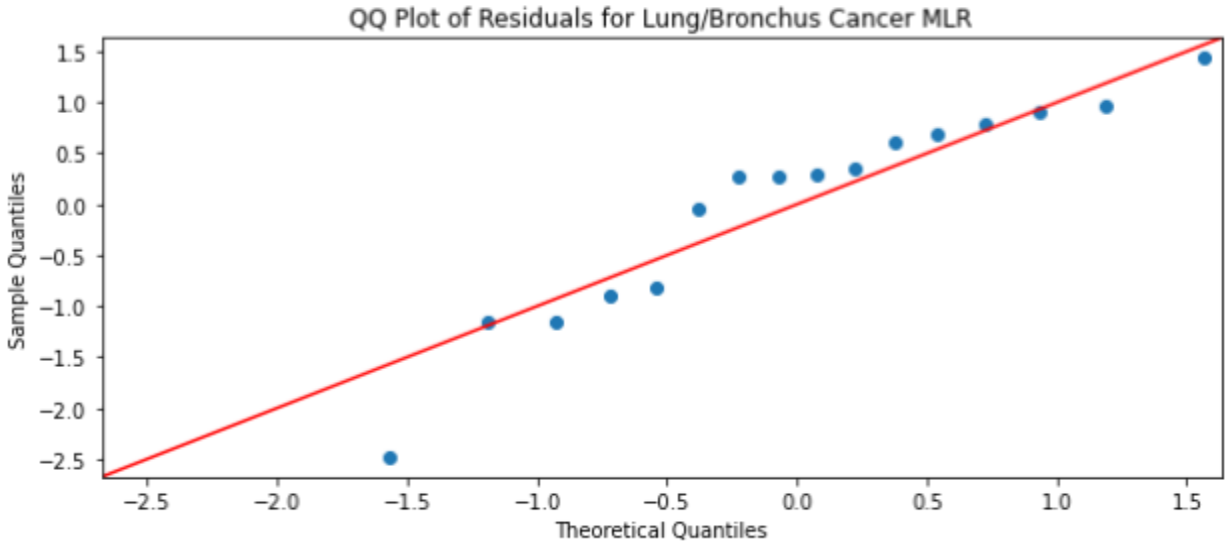
To understand if the MLR model fits the data better than a model with no independent variables, the F-statistic value and its corresponding p-value were also calculated for each model. The F-statistic values for the MLR models ranges from as low as 0.94 for the COPD MLR model to as high as 47.29 for the lung/bronchus cancer MLR model. For both COPD and chronic kidney disease MLR models, the p-values associated with the F-statistic values are greater than 0.05. This indicates that there is no statistical significance of these MLR models, so there is no general relationship between the air quality parameters and the chronic health condition. The MLR models for cardiovascular disease and lung/bronchus cancer both have high F-statistics with statistically significant p-values. The F-statistic for the cardiovascular disease MLR is 41.67, which indicates that there is a relationship between the air quality parameters and the prevalence of cardiovascular disease in the U.S. Similarly, the F-statistic for the lung/bronchus cancer MLR has a F-statistic of 47.29, which indicates that the prevalence of lung/bronchus cancer is influenced by the quality of the air. The scatterplot of the actual rates of lung/bronchus cancer versus the predicted values are depicted below, along with an identity line of y=x.

Figure 7



In Figure 7, the scatterplot shows how the actual and predicted values of lung/bronchus cancer prevalence. In the case of this MLR model, these values are somewhat close to the identity line. This indicates that the MLR model for lung/bronchus cancer can somewhat accurately predict the prevalence rate of cancer from the concentrations of air quality parameters.

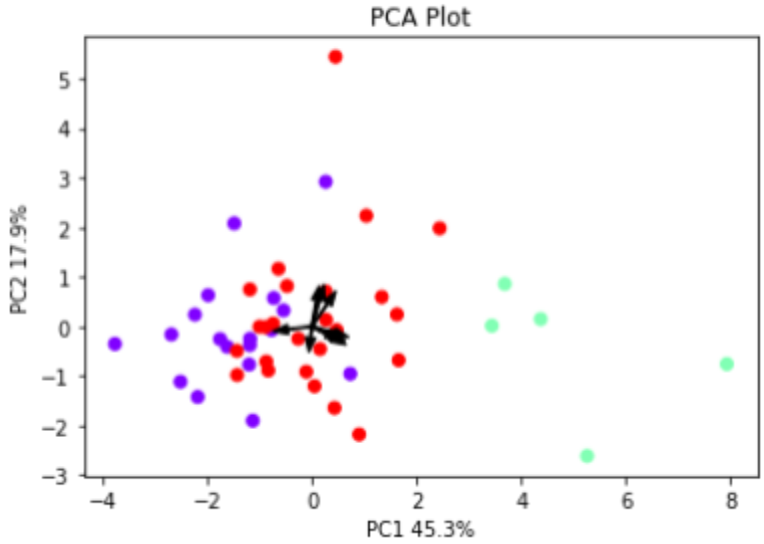
Figure 8



In Figure 8, the QQ plot of the residuals above for the lung/bronchus cancer MLR model show the residuals are somewhat aligned with the identity line, which indicates that the residuals are mostly symmetric in distribution. However, there are some outliers on the QQ plot.

After scaling the data, k-Nearest Neighbors (kNN) models were generated for each health condition to determine if this would serve as a better prediction model for the dataset. To understand the strength of each kNN model, the accuracy scores were calculated. The kNN model for COPD had the highest level of accuracy at 13.33%, however this is still quite low. Then, principal component analysis (PCA) was performed. Ten principal components were produced, with the first two components explaining 45.3% and 17.9% of the variance respectively. PC 1 and 2 are depicted in Figure 9.

Figure 9



The PCA plot above shows PC 1 on the x-axis and PC 2 on the y-axis, with PC 1 explaining 45.3% of the variance and PC 2 explaining 17.9% of the variance. There are two main clusters in addition to the smaller, more spread out cluster on the plot. The arrows on the biplot are quite short, which demonstrates that there is not much variance among the variables.

K-means clustering was performed to determine the cluster centroid values and the silhouette score. The silhouette score comes out to be 0.184. Since this score is positive and above zero, it indicates that the samples were assigned to the right clusters. However, since the score is closer to zero and not one, this model has overlapping clusters. Overall, the silhouette value indicates that the K-means clustering model is not strong.

Figure 10

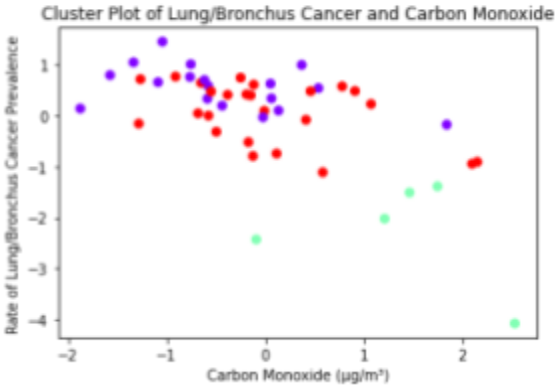
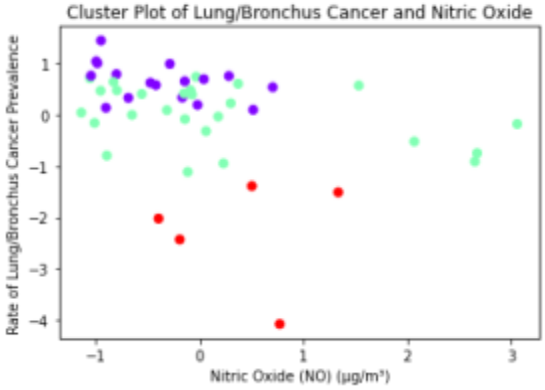


Figure 11



The cluster plot in Figure 10 reflects the K-means clustering model for the relationship between carbon monoxide and the rate of prevalence of lung and bronchus cancer. The cluster plot in Figure 11 belongs to the same model and instead reflects nitric oxide as the air quality parameter. These plots reflect the low silhouette score in that the clusters in both of these plots are overlapping each other and not densely situated.

Discussion

After merging the datasets together and conducting histograms of the variables, the data was visibly asymmetrical. After normalizing the data, the distributions of the variables shifted but still came across as fairly asymmetrical. Eventually, the data was later standardized for some machine learning techniques. The lack of symmetric distribution for each variable may have affected the visualizations and the results of the machine learning models.

The correlation matrix revealed both direct and inverse relationships between the air quality and health condition variables. These correlation coefficients hold some value in demonstrating the direction and strength of the linear relationships between variables. However, the correlation coefficients are easily influenced by the outliers. In the case of sulfur dioxide, for example, there were several outliers that needed to remain in the dataset.

According to the R-squared value, the multiple linear regression model was the strongest for lung and bronchus cancer. Carbon monoxide, nitric oxide, and sulfur dioxide all had high p-values, while sulfur dioxide had a statistically significant p-value. The overall model had a high F-statistic with a p-value that is less than 0.05, which indicates that there is a relationship

that exists between the air quality parameters and the prevalence of lung and bronchus cancer. Similarly, the MLR model for cardiovascular disease had a high F-statistic with a low p-value. The k-Nearest Neighbors models were another prediction model generated, but all the models had poor accuracy scores. The model for COPD had the highest predicting accuracy compared to the other models at 13.33%. The kNN model is generally better for predicting categories, so the low accuracy scores of the kNN models reflect its misuse in predicting numeric values. The MLR models were generally a much better fit for the evaluation and prediction of the data compared to the kNN models. However, the MLR models for COPD and chronic kidney disease were found to have no linear relationship between air quality parameters and the chronic health condition. Additionally, most of the MLR models had statistically insignificant correlation coefficients for nitric oxide, carbon monoxide, and sulfur dioxide parameters. This could also be attributed to the poor quality of data, especially given its mostly asymmetric distribution.

Principal Component Analysis (PCA) produced ten principal components. PC 1 explained 45.3% of the variance in the data, while PC 2 explained 17.9%. The vectors on the PCA plot indicate that there is very little variance among the variables. The clusters on the plot are not well-defined and mostly overlapping. It is possible that the lack of definition for the clusters may be picked up the other principal components. The overlap in clusters may also be occurring because the data points share a lot of similarities. The K-means clustering model had a similar output in that there were clusters with low density and some overlap with other clusters.

Since the ozone parameter had a statistically significant relationship with each chronic health condition, it may be worth isolating the ozone parameter from the other air quality parameters to further explore how it impacts the prevalence of these conditions. The machine learning models in this project were generally found to be low in predicting accuracy and fit for the data, which underscores the importance of having good, almost perfect data for ML to truly work. Statistical analysis may also help in understanding the data, especially when ML struggles to make sense of real data points.

Acknowledgements
This data science project would not have been possible without the help and support of Dr. Layla Guyot. Dr. Guyot's feedback and patience were incredibly valuable to my learning process.

References

Centers for Disease Control and Prevention. (n.d.). *Chronic Disease Indicators (CDI)*. Chronic Data. https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-/g4ie-h725

Environmental Protection Agency. (2020). *Annual Summary Data*. EPA. https://aqs.epa.gov/aqsweb/airdata/download_files.html

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, *8*(14), 1–13.

NIH. (2020). State Cancer Profiles. https://statecancerprofiles.cancer.gov/

VanderPlas, J. (2019). *Python Data Science Handbook | Python Data Science Handbook*. Github.io. https://jakevdp.github.io/PythonDataScienceHandbook/

Computer Code
The Python code for this project can be found in the PDF below:

SRP Code

Reflection

I started working on my certificate in Scientific Computation and Data Sciences in June 2022. Within the past year, I have taken several courses in data visualization and sciences. These courses have introduced me to a variety of languages, machine learning techniques, and statistical analyses. I was first exposed to Python for data sciences in NEU 365P: Programming and Data Analysis for Modern Neuroscience. I strengthened my data science skills with Python even more through the SDS 322E: Elements of Data Science course. With this project, I sought to strengthen my data science skills with Python, especially since the majority of my coursework has been with R programming.

The presence of air pollution is the unfortunate reality in nearly every country today. Air pollution is fueled by human activities such as utilization of fossil fuels and deforestation. Air pollution promotes climate change, which is often destroying habitats and homes. Knowing all of this, I was curious as to how the quality of air impacts our health.

The datasets I chose to work with were big, which made it difficult to wrangle and clean up for analysis. This was the first data science project that I have completed with Python, so it was challenging to figure out how to conduct certain techniques. With R, machine learning and data visualization requires very simple and little code. I found that these same processes are not as intuitively built with Python. It was challenging, but exciting to learn how to utilize Python for data science and visualizations for this project. Moreover, I learned a lot about the connection between air quality, climate change, and human health.