US Air Pollution - Environmental Dataset

Mogollon Monster - Executive Summary

Team Members: Nibedita Bal, Zach Paquin, Upendra Rajendran, David Terracino

Introduction

In the fast growing world with technological advancements in IT services, Manufacturing, Supply Chain and so forth are being continuously integrated to provide betterment on "comfort living". A major challenge faced by government or local bodies is to control and regulate the air quality standards in the living zone. With the technological inventions in pollution monitoring, governments or local bodies are trying to find out an effective solution to regulate the Air Quality Index (AQI). The AQI is calculated for six common air pollutants: nitrogen dioxide (NO2), ozone (O3), sulfur dioxide (SO2), carbon monoxide (CO), and particulate matter of diameters less than 2.5µm and 10µm (PM2.5 and PM10). The dataset for this project contained AQI values for the first four pollutants listed (NO2, O3, SO2, CO). This project focuses on investigating the correlation pollutants and its respective AQI and building a prediction model based on the results of the regression analysis on EPA governed data.

The dataset was prepared by creating subsets of the data based on the region of the United States where the measurements were taken. The Season and time of day (TOD) fields were created based on the measurement date and measurement hour, respectively. Each team member was assigned one region to perform multiple regression to model the relationship between each pollutant's AQI and the variables within the dataset. The team members created the best model they could by transforming variables, adding interaction terms and/or adding second order terms. Since each team member had a different region of the United States, the models are expected to be slightly different. Each model was validated using 10-fold cross validation. This report details the analysis of four of the regions of the United States.

The results of our analysis show seasonal trends in the pollutants across all regions. For example, the Midwest region in the spring season has a higher NO2 AQI than in winter and summer, and the Pacific West region showed higher CO AQI values in the winter than any other season. The time which the measurements were recorded also showed trends across all regions analyzed. The SO2 AQI tended to be higher in the morning for the Pacific West region; the same holds true for NO2 in the Southwest region. These trends differ slightly between the regions, and which could be explained by the difference in climate for each of the regions. The Southwest United States is known for being warmer and more humid than the Midwest. The time of day trends make sense as well because people are commuting to work in the mornings and evenings. The following paragraphs will detail the major findings for each region analyzed.

Northeast

Nitrogen Dioxide (NO2) is known to follow seasonal patterns. The chemical gets more potent in colder temperature when compared to warmer seasons. Unlike Nitrogen Dioxide, Carbon Monoxide is a toxic gas that cannot be detected with naked eyes / sense. when produced, it typically causes headaches, vomiting, and dizziness. However, prolonged exposure will cause death (Centers for Disease Control). Following findings were exciting and worth doing a deep examination, It shows a correlation that the amount of Ozone (O3) and Sulfur Dioxide (SO2) produced are statically significant to each other. If one chemical is created, the odds of the other chemical being produced are very high. In addition, every unit of O3 produced, they also create -4.082e-04 units of Sulphur Dioxide. This relationship suggests that more Ozone is getting exposed than Sulphur Dioxide.

Southwest

The time of day and seasons for the Nitrogen Dioxide (NO2) model provided the most significant information. NO2 levels increase throughout the day and decrease during the evening and night time hours. This observation makes sense when we think of commuter traffic times and traffic patterns during the day. The major cities of Charlotte NC, and Winter Park Florida exhibited higher betas for NO2 which is also expected as major cities are population and traffic centers. Conversely those two cities were much lower in Ozone. Seasonally the data suggest that NO2 levels are higher in the spring and summer months and lowest during winter. More data with ambient temperatures would potentially be useful to uncover additional patterns with respect to NO2.

Seasonally the O3 betas were negative during the spring and summer, but positive in the winter time. Both pollutants are caused mainly by emissions from cars and trucks, and from the coal-burning power plants that supply a large amount of NC electricity. The Sulphur Dioxide (SO2) model with respect to seasonal results is more intuitive demonstrating a reduction in levels during the winter months and higher concentrations in spring and peaking in summer. The city feature revealed higher SO2 concentrations in most southern cities. This holds to the idea of the industrial complex being inside or close to the city limits and not in rural areas.

Carbon Monoxide (CO) levels seasonally demonstrate better values in the winter than in warmer months. Many of the cities as well have a low beta relative to other air pollutants. This may be due to increased mass transit or policies in effect are making an impact in bigger cities such as Raleigh and Charlotte but not Winston-Salem. Furthermore Raleigh is more of a tech industry, health conscience area and Charlotte more of a finance district than industrialized factories or plants. In closing, further data containing ambient temperatures throughout the day in a time series model may be particularly useful to increase analysis of this dataset. Additional team data will also be compounded for a collective analysis and potential new models will foster new insights.

Pacific West

Based on the R squared values for the models for the Pacific West Region, the CO model has the best prediction capabilities. Each model had a high R squared value, the lowest of which being the O3 model at 0.925. There are outliers in the dataset, which are identified in the residual plots. For the most part, the residuals are normal, homoscedastic, independent, and average to 0. A potential next step of this project would be to go back and perform different types of regression for each of the pollutants. Looking at the residuals vs. fitted plot for the NO_2 model, it might be prudent to attempt piecewise regression for the NO2.AQI. When the AQI gets around 100, there is an inflection point leading to a steep, downward drop. For the O_3 model, polynomial regression should be attempted. Piecewise regression should be attempted for SO_2 and CO as well.

All of the pollutants, except NO_2, had the Season and TOD variables in their model. According to the Utah Department of Environmental Quality, $PM_{2.5}$, or Particulate Matter with a diameter less than 2.5µm, is the main component of Utah's wintertime pollution. They also state that in the summer, nitrogen oxides, such as NO_2, close to hot roads forms O_3, and that nitrogen oxides are a direct component to $PM_{2.5}$ pollution. Also, since the main source of nitrogen oxide pollution is the burning of fossil fuels, that could explain why the TOD field was not included; the burning of fossil fuels happens nearly every hour of every day [1]. Although the source is specific to Utah, other states in the Pacific West region have similar climates, such as Montana and Nevada. This could be a factor as to why Season and TOD were not included in the NO_2 model for this region. $PM_{2.5}$ data for this region would be helpful. Defining the relationship between NO_2, O_3 and $PM_{2.5}$ might give more insight to better model NO_2 and O_3 for the Pacific West region.

Midwest

Among 4 pollutants, No2 has the best predictive capabilities in selecting the total AQI in the Midwest region. The best model chosen among the four response variables was the NO2 model. Even though NO2 and O3 have adjusted R2 nearly the same, O3 cant be taken as the best model because of the heteroscedasticity nature of O3 residuals. The sum of the residuals for NO2 is -6.56e-11which is almost zero, it means the model is a good model. The beta values for O3.1st.Max.Value.ppm is very high for NO3, which means this pollutant is pretty significant calculating NO2 AQI. The correlation between actual and prediction value is 0.99 and high correlated variables were removed to avoid multicollinearity. The adj R square of this model was 0.816. After adding second-order terms and interaction terms to 0.822 this means 82%of the variability in NO2.AQI is explained by the model. The P values look good for all variables. So we can reject the null hypothesis and accept the alternative hypothesis.

Some states are highly positive beta coefficients which indicate if the categorical variables are set to 1 then they directly impact the total AQI and these variables are trustable. For example, State North Dakota has more impact on NO2 AQI in comparison to other states, and the spring season has more NO2 than season winter and summer in the Midwest region

Conclusion

The results of this project show that different factors affect the AQI for each pollutant based on the region where the measurements were taken, which makes sense due to different state climates and environmental regulations. Additionally, seasonal and time trends were observed across all models and regions. Based on the results of this project, it is suggested to acquire additional data to better predict the AQI values for a pollutant. This additional data includes the ambient temperature and relative and absolute humidity for the date and time the measurement was recorded, the elevation level where the measurement was recorded, as well as PM2.5 and PM10 data. An analysis to determine the relationships between the air pollutants is critical in developing an accurate model. Another suggestion is to revisit the current models and perform advanced regression, such as piecewise or polynomial regression, to try and create a more accurate model. It is recommended to perform a time series analysis to try to identify if/when new regulations were passed for each region to try and explain some outliers found in the analysis.

References

[1] Utah Department of Environmental Quality, "Understanding Utah's Air Quality," 13 June 2019. [Online]. Available: https://deq.utah.gov/communication/news/featured/understanding-utahs-air-quality. [Accessed 30 May 2020].