Introduction:

My research involves the correlation analysis between pollutants in the air and weather conditions. Specifically, the pollutants that are being measured include CO(GT), NOx(GT), C6H6(GT), PT08.S1(CO), NMHC(GT), PT08.S2(NMHC), NO2(GT), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3). The target pollutants include carbon monoxide(CO), nitrogen oxides(NOx), and methane (C6H6). The weather conditions include temperature(T), relative humidity(RH), and absolute humidity(AH).

I hypothesize that there will be a positive correlation between higher temperatures and lower humidity with elevated concentrations of pollutants. This correlation indicates a link between weather conditions and the degradation of air quality.

The dataset I am using was acquired from Kaggle. The CSV file required cleaning to be ready for regression analysis usage. After cleaning the data, I completed multiple linear regression to analyze the correlation between the pollutants, independent variables, and the weather conditions, dependent variables. After I obtained all the values needed for analysis I made visualizations that represent the relationship between variables.

Background:

Air quality is a crucial indicator of how clean or polluted the air is. It also represents the overall health of the population that lives in the area. This is because people that live in the region are constantly breathing in whatever is in the air, and if there are many pollutants in the air then humans are forced to breathe in the toxic materials as well. The pollutants in the air also affect the environment by depositing toxins into plants and trees. Additionally, if there is rainfall in an area with a high concentration of pollutants, the rain will deposit acid and excess elements that damage habitats within the ecosystem.

This dataset contains 9357 instances of hourly averaged responses from five metal oxide chemical sensors that measured gas concentrations. This was done in a significantly populated city in Italy on a field at the same altitude as the roads. The data was taken from March 2004 and February 2005.

Data and Methods:

The methods that I have used will be put into a Python notebook and uploaded to GitHub. My workflow is also outlined within my code, making it easy to follow along with my processes to obtain results.

I had a few problems when completing all of my final work. First, the original dataset I wanted to use was unavailable after I had already chosen it. Additionally, after presenting my draft to the class, I decided to do multiple linear regression, rather than simple linear regression. I wanted to compare all the pollutants to three different variables to see the relationship between different environmental factors and all the pollutants. I also had lots of problems within my code in completing the multiple linear regressions and visualizations. For example, I wanted to implement a bubble chart but decided not to put it in my presentation due to the lack of information it conveyed. I felt like the bubble chart did not do a good job of visualizing the relationship between the variables.

The dataset used needed a lot of work to be usable for analysis. Cleaning the data was difficult as there were many columns with NaN values and due to the three independent variables I had to ensure the data was proper so every relationship would be on equal ground.

Results:

Correlation Analysis:

The interpretation of correlation coefficients helps understand the relationship between variables. Correlation Coefficients range from -1 to 1:Positive Correlation (0 to 1): As one variable increases, the other tends to increase as well. Negative Correlation (-1 to 0): As one variable increases, the other tends to decrease. The closer the coefficient is to -1 or 1 the stronger the correlation is strong. If the coefficient is close to 0, that means there is a weak correlation.

Here are the correlation results in the primary target pollutants:

CO(GT):

T: Correlation of -0.0689 (weak negative correlation, significant with p-value = 2.46e-11). As temperature increases, CO(GT) tends to decrease slightly, but the relationship is weak.

RH: Correlation of -0.0482 (weak negative correlation, significant with p-value = 3.05e-06). Relative humidity has a slight negative relationship with CO(GT).

AH: Correlation of -0.0459 (weak negative correlation, significant with p-value = 8.96e-06). Absolute humidity also has a weak negative relationship with CO(GT).

NOx(GT):

T: Correlation of -0.1385 (moderate negative correlation, significant with p-value = 2.85e-41). Temperature has a moderate negative effect on NOx(GT).

RH: Correlation of -0.0530 (weak negative correlation, significant with p-value = 2.89e-07). Relative humidity has a slight negative effect on NOx(GT).

AH: Correlation of -0.0958 (moderate negative correlation, significant with p-value = 1.52e-20). Absolute humidity has a moderate negative relationship with NOx(GT).

C6H6(GT):

T: Correlation of 0.9714 (very strong positive correlation, significant with p-value = 0.0). Temperature is highly positively correlated with C6H6(GT).

RH: Correlation of 0.9251 (strong positive correlation, significant with p-value = 0.0). Relative humidity is strongly positively correlated with C6H6(GT).

AH: Correlation of 0.9846 (very strong positive correlation, significant with p-value = 0.0). Absolute humidity has a very strong positive correlation with C6H6(GT).

Regression Analysis:

In my presentation, I chose to only use the intercept values. I chose the intercepts over the three individual coefficients relating to Temperature(T), Relative Humidity(RH), and Absolute Humidity(AH) because the intercept gives the base level of the pollutant when environmental factors don't have an influence. In real-world settings, this may not be practical, but it was easier to visualize within code and the presentation.

Another reason I chose not to use individual coefficients is because they represent a one-unit change in the respective independent variable. The coefficient values show the influence of each independent variable on pollutant values.

For example, for carbon monoxide(CO(GT)), the intercept was 31.0791, which means if T, RH, and AH are all zero then the predicted level of CO(GT) would be 31.0791. The coefficients from the regression model tell you how much CO(GT) would increase or decrease with a one-unit change in T, RH, or AH. For every one-unit increase in temperature (T), the value of CO(GT) is expected to decrease by approximately 2.00, making relative humidity and absolute humidity constant. This suggests that as the temperature increases, the CO(GT) concentration tends to decrease. For every one-unit increase in relative humidity (RH), the value of CO(GT) is expected to decrease by approximately 0.66, making temperature and absolute humidity constant. This indicates that higher humidity levels are associated with lower concentrations of CO(GT). For every one-unit increase in absolute humidity (AH), the value of CO(GT) is expected to increase by approximately 2.91, making temperature and relative humidity constant. This suggests that as the moisture in the air increases, the concentration of CO(GT) tends to increase as well.

Here are all the Regression Coefficients for Air Quality Pollutants

- CO(GT) Intercept: 31.0791, Temperature: -2.0048, Relative Humidity: -0.6599, Absolute Humidity: 2.9063
- PT08.S1(CO) Intercept: 890.7222, Temperature: 4.2836, Relative Humidity: 2.6813, Absolute Humidity: -1.5121
- NMHC(GT) Intercept: -96.5638, Temperature: -1.8978, Relative Humidity: -0.6362, Absolute Humidity: 2.8995
- C6H6(GT) Intercept: 5.4098, Temperature: 0.1615, Relative Humidity: 0.0171, Absolute Humidity: 0.8484
- PT08.S2(NMHC) Intercept: 724.3954, Temperature: 8.8043, Relative Humidity: 1.1990, Absolute Humidity: -5.3821
- NOx(GT) Intercept: 343.7566, Temperature: -8.0770, Relative Humidity: -0.8591, Absolute Humidity: 9.2116
- PT08.S3(NOx) Intercept: 1142.6391, Temperature: -8.5252, Relative Humidity: -3.4457, Absolute Humidity: 18.6825
- NO2(GT) Intercept: 221.9128, Temperature: -4.2961, Relative Humidity: -1.9150, Absolute Humidity: 6.8493
- PT08.S4(NO2) Intercept: 409.1585, Temperature: 33.9252, Relative Humidity: 9.4832, Absolute Humidity: -40.3601
- PT08.S5(O3) Intercept: 793.7094, Temperature: 2.8348, Relative Humidity: 3.6286, Absolute Humidity: -1.4950

- T(Temperature) Intercept: 5.3291e-15, Temperature: 1.0000, Relative Humidity: -1.7073e-16, Absolute Humidity: 2.4813e-16
- RH(Relative Humidity) Intercept: -4.9738e-14, Temperature: -3.6227e-16, Relative Humidity: 1.0000, Absolute Humidity: 3.3307e-16
- AH(Absolute Humidity) Intercept: -1.3323e-14, Temperature: 1.5881e-16, Relative Humidity: 2.2204e-16, Absolute Humidity: 1.0000

Some things to note about the models and results for the target pollutants include: that for CO the correlations with T, RH, and AH are weak. Even though they are statistically significant, these environmental factors do not properly explain the little variance. For NOx, the correlation with T and AH is moderate, while the correlation with RH is weak. For C6H6 the correlations with T, RH, and AH are very strong, particularly with AH (0.9846) and T (0.9714). This is consistent with the high R-squared (0.970). This suggests that the environmental factors in the model are very reliable predictors for C6H6.

The R-squared results are a good indication of whether the models were effective per pollutant.

- CO(GT): $R^2 = 0.031$, Adjusted $R^2 = 0.031$.
  - Interpretation: The model explains only about 3% of the variance in CO, which is very low.
- NOx(GT): $R^2 = 0.063$, Adjusted $R^2 = 0.063$.
  - Interpretation: The model only explains about 6% of the variance in NOx. Again, the relationship is weak, which may indicate that the environmental factors may not have the most effect.
- C6H6(GT): $R^2 = 0.970$, Adjusted $R^2 = 0.970$.

- Interpretation: With nearly 97% of the variance in C6H6 being explained by T, RH, and AH, this suggests a very strong relationship and that these predictors are highly relevant for predicting C6H6.

Some pollutants, like C6H6(GT) and PT08.S4(NO2), have high R^2 scores (close to 1.0), suggesting a better relationship with the environmental factors.

Pollutants like CO(GT) and NO2(GT) have lower R^2 scores, meaning that the environmental factors explain less of their variance. This could indicate the need for additional predictors or a more complex model for those pollutants. These models may include the inflow of traffic in an area, whether there are industrial factories around, and the population of residents that drive cars that emit high amounts of carbon.