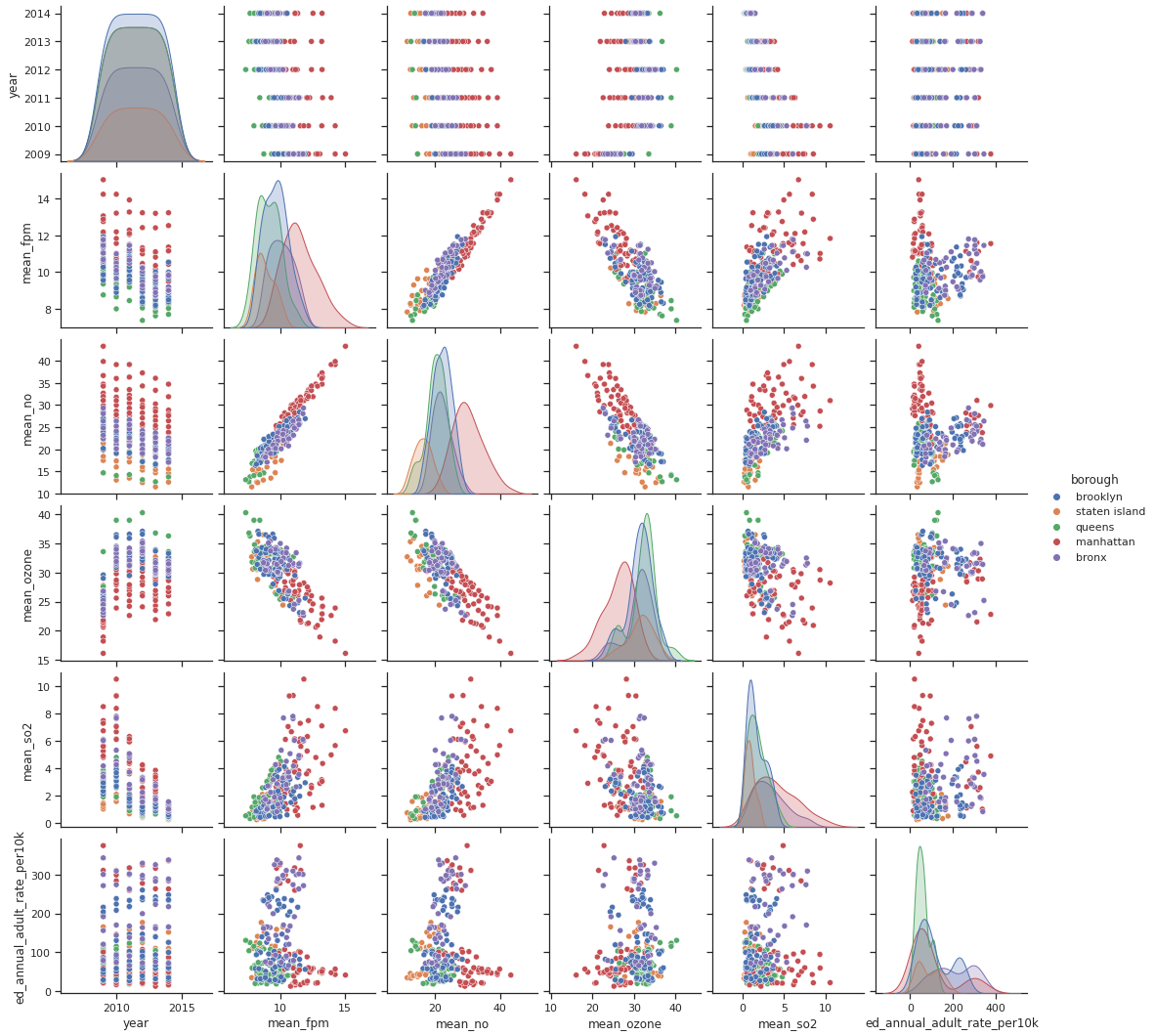
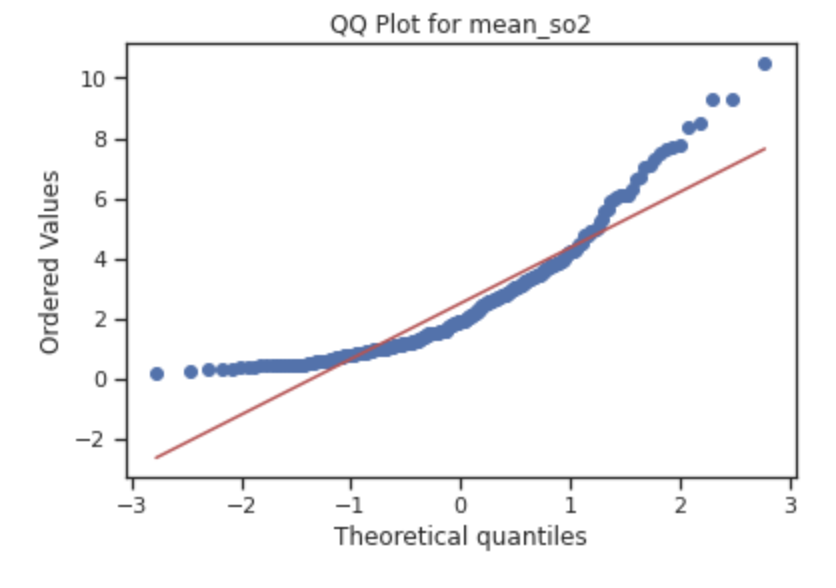
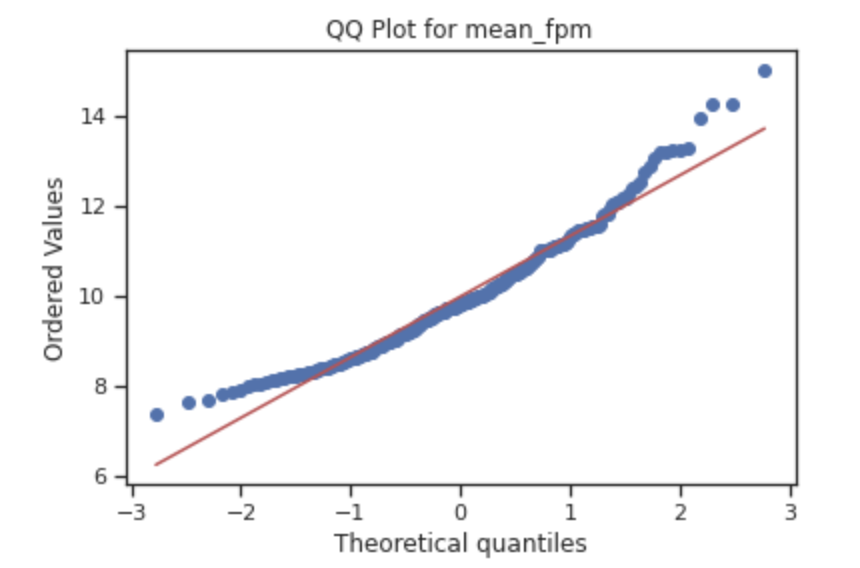
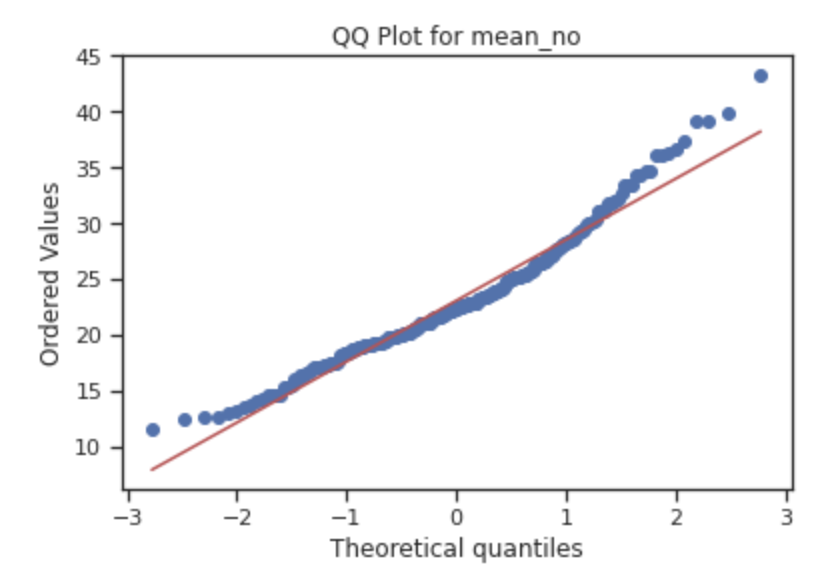
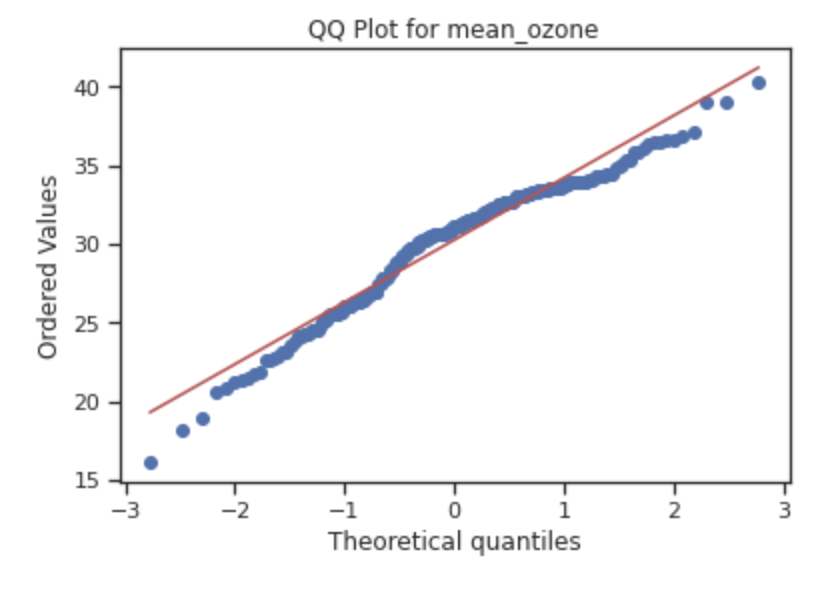
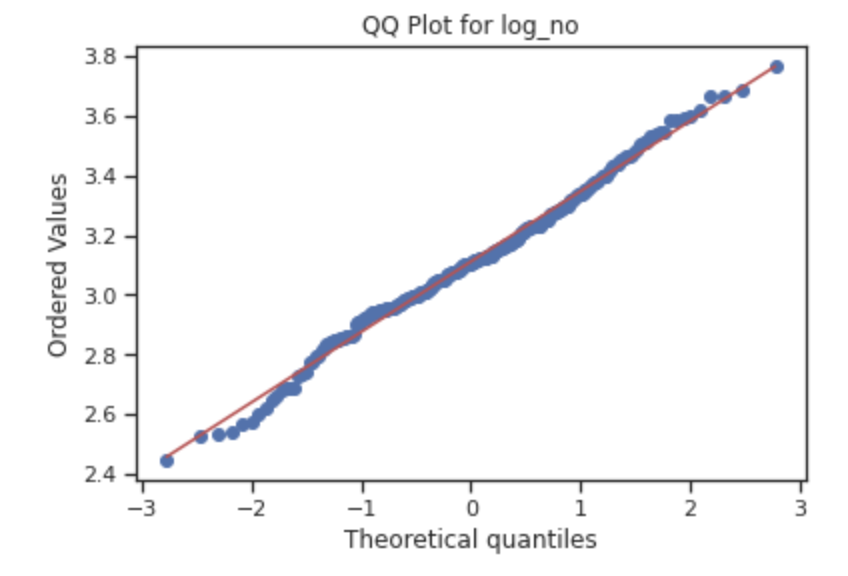
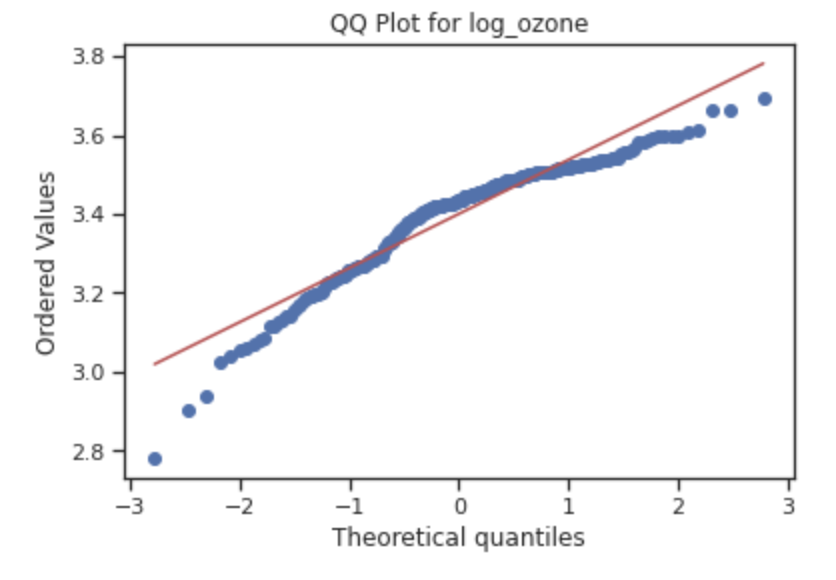
Testing Regression Assumptions:

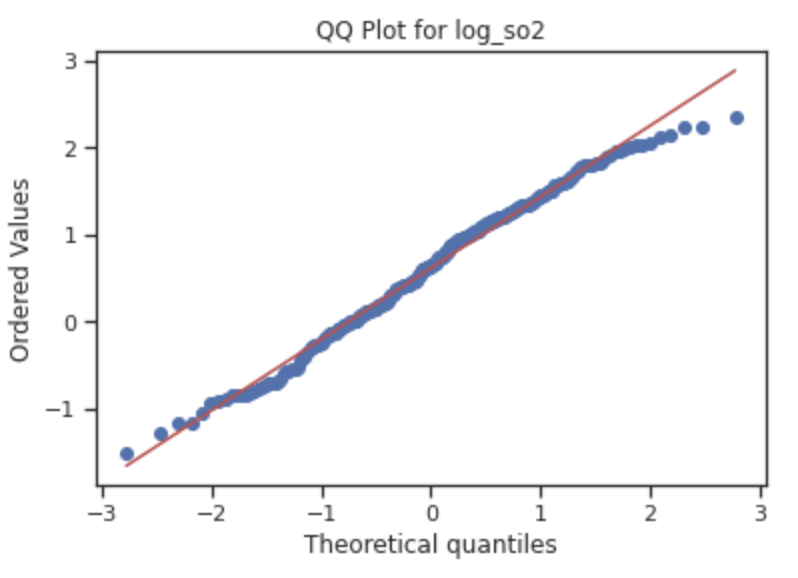
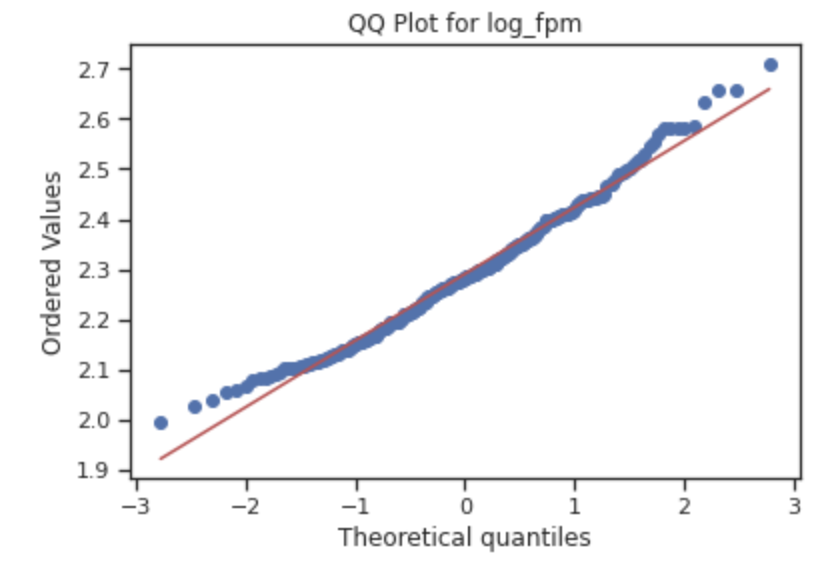


1. Linear relationship: Is there an apparent linear relationship between the predictors and the target variable? The scatter plots show that there is not a strong linear relationship between any of the predictor variables and the target variable of emergency department visits. The scatter plot may not tell the whole story therefore it is worth comparing models to determine if there are any predictors that demonstrate a linear relationship with the target variable.
2. Normal distribution of each variable: I used QQ plots to check for normal distribution of each variable.



The QQ plots show that the data for mean\_fpm, mean\_no, and mean\_so2 are right skewed and the data for mean\_ozone are left skewed. The variables were transformed using a log transformation to try to create distributions that more closely resemble a normal distribution. The following QQ plots are the represent the data after a log transformation. It is clear that the data are closer to a normal distribution after the transformation.





1. The third assumption of linear regression is that there is no autocorrelation. This means that when considering each predictor variable independently, one observation is not dependent on the previous observation. Each observation represents the yearly average of a certain air pollutant in a particular neighborhood.
2. No multicollinearity: This means that the predictor variables are not highly correlated with each other. We can see that mean\_fpm and mean\_no are highly correlated with each other, and therefore only one of them should be included in the model.
3. Homoscedasticity: this means that the variance of the observations is the same for all the data. We can see that is not the case with these data sets since the variability of the dependent variable tends to increase as the independent variable increases.

Regression Process:

After checking the assumptions of linear regression, the data was transformed using a log transformation and standardized. Several regression models were tested using ordinary least squares regression to determine the best model. The first few models tested related annual emergency department visits for asthma to each of the four air pollutants. These models showed r-squared values close to 0.0 and high AIC values over 700.

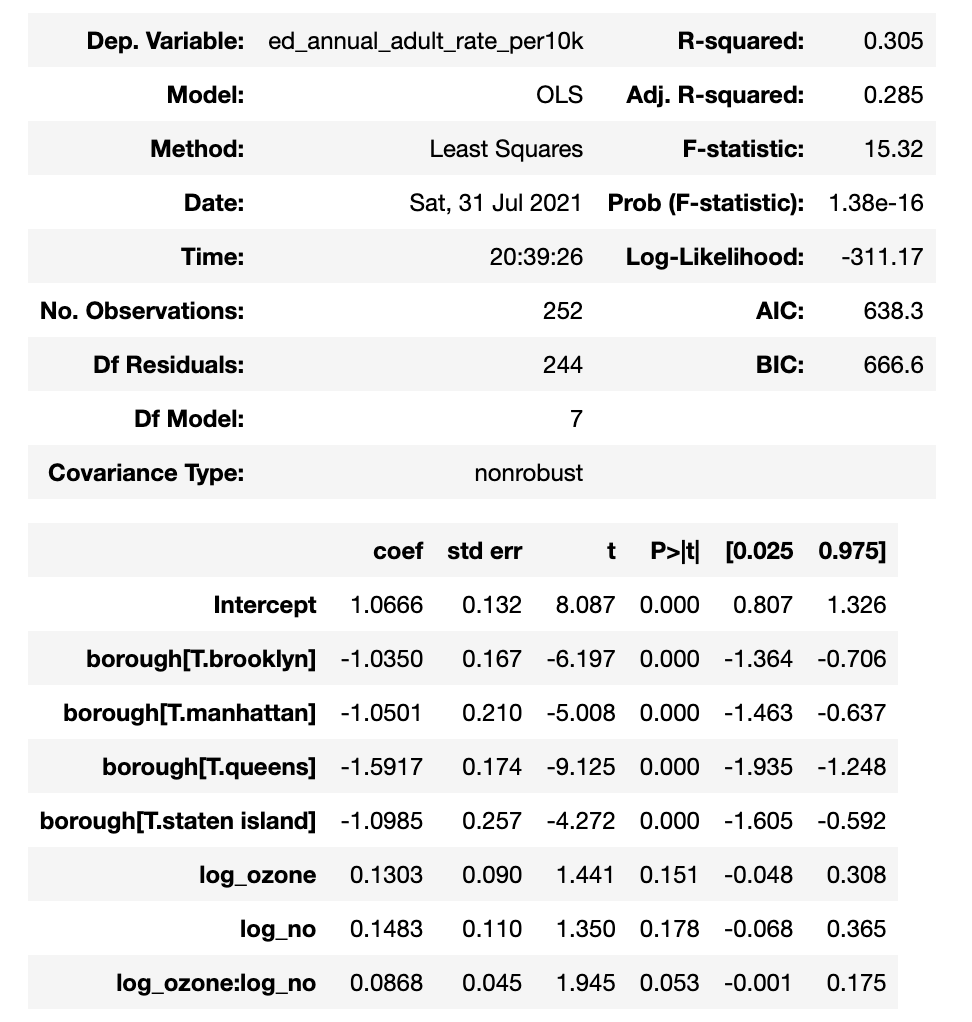
The next step was to include the categorical variable ‘borough’ into the model. Introducing this categorical variable raised the r-square value significantly and decreased the AIC value. Various models were tested using different combinations of variables. The table summarizes the results.

| Predictor Variables in Model | r-squared | AIC | Significant Variables |
| --- | --- | --- | --- |
| log\_no | 0.005 | 716.8 | none |
| log\_ozone | 0.003 | 717.3 | none |
| log\_so2 | 0.022 | 712.6 | log\_so2 |
| log\_fpm | 0.002 | 717.7 | none |
| borough | 0.278 | 642.2 | borough[T.brooklyn], borough[T.manhattan], borough[T.queens], borough[T.staten island] |
| borough, log\_ozone, log\_no, log\_ozone x log\_no | 0.305 | 638.3 | borough[T.brooklyn], borough[T.manhattan], borough[T.queens], borough[T.staten island] |

The model that had the highest r-squared and the lowest AIC is the model that includes the categorical variable ‘borough’, log-transformed mean\_ozone, log-transformed mean\_no, and the interaction of log\_ozone and log\_no.

This model indicates that asthma rates have some dependence on borough. There might be some underlying causes for that which could relate to social determinants of health such as income levels, number of people below the federal poverty line, and race.

The regression summary for the model is shown below:



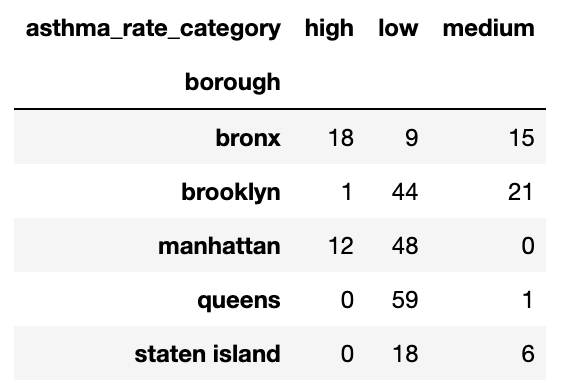
In the summary report, each borough is compared to The Bronx. The coefficient for each of the four boroughs are negative which means that Brooklyn, Manhattan, Queens and Staten Island have lower asthma rates than The Bronx.

The variables log\_ozone, log\_no, and the interaction of log\_ozone and log\_no have p-values greater than 0.05 which means that they are not significant.

The results of the regression analysis indicate that the air pollution factors of mean yearly fine particulate matter, nitrogen dioxide, sulfur and ozone are not significant predictors for asthma rates in New York City. Even though research shows that these air pollutants are closely linked to high asthma rates, our analysis did not reflect this conclusion. One of the possible reasons why our data does not show similar results to other studies of air quality and asthma rates is that the original datasets regarding air quality that are used in our study were aggregated to indicate the yearly mean levels of air pollutants. Our findings could be missing important information about daily air pollution rates that factor into high asthma rates.

Chi Square Analysis:

The results of the regression analysis indicate that borough is a significant predictor of asthma rates. For further investigation, a chi squared analysis was performed to test the null hypothesis that asthma rates in each borough are equal. In order to perform the chi squared test, asthma rates needed to be turned into a categorical variable. The asthma rates were transformed to a categorical variable indicating low (rate between 0 and 126), medium (rate from 126 to 252) and high (rate 252 or higher). Then a bivariate frequency table was created to show the number of low medium and high asthma rates in each borough.



From the table, it is clear that The Bronx has the most regions with high asthma rates, and Queens has the most regions with low asthma rates. The Chi Square analysis returns a p-value of . This value is significant, so we can reject the null hypothesis that asthma rates in each borough are the same. This result is consistent with the significance of the boroughs in the regression analysis.