MGT 6203 Group Project Final Report

BACKGROUND

Project Title: A Breath of Fresh Air: Investigating Air Quality and Its Impacts on Human Health

Team Members: Sarah Bartley, Zaid Hasan, Jayaprakash Jayakumar, Sarah Snader, Eli Wade

Background Information

Air quality has been studied for several decades, and a standard measurement system has been developed to outline air quality risk levels. This Air Quality Index (AQI) measures levels of ground level ozone, particulate matter, carbon monoxide, sulfur dioxide, and nitrogen dioxide. All of these compounds have been correlated with negative impacts on human health - specifically of the lungs, and related to increased susceptibility to infectious diseases.

Problem Statement

The purpose of this investigation is to analyze air quality data with hospitalization records in an area to determine if there is a correlation between poor air quality and an increased likelihood of requiring medical assistance. By understanding the relationship between air quality and respiratory health, we can provide valuable insights for policymakers and healthcare organizations to take action and improve public health.

Primary Research Question

How does the level of air pollution in different US counties correlate to the existence of respiratory diseases and mortality rates, and what are the key factors that contribute to this relationship?

Supporting Research Question

To what extent were areas with poor air quality more susceptible to COVID-19?

Initial Hypothesis

The initial hypothesis is that areas with poor air quality experience higher rates of hospitalizations due to respiratory diseases and also were more susceptible to COVID-19.

Approach

The datasets were cleaned and joined. Four datasets were used in this analysis: AQI dataset, hospitalization dataset, US population dataset, and COVID-19 dataset. Two separate joined datasets were created. The air quality dataset and COVID-19 dataset were aggregated into yearly values and joined on State, County, and year columns, and the hospitalization and US population datasets were joined to the hospitalization dataset by State, County, and year. Inner joins were used to eliminate NULL values.

For exploratory data analysis, a correlation matrix was created to assist in understanding correlations between variables, and heat maps were created for each independent and dependent variable to determine if commonality between maps existed.

Multivariate regression analysis was applied in R to find the degree of correlation between independent and dependent variables. Both linear regression models and log-linear regression models were tested, with either hospitalization or COVID case data used as independent variables. R squared was used to confirm model accuracy. Outliers were removed as necessary. Collinearity was analyzed with VIF analysis, and models were optimized as necessary.

OVERVIEW OF DATA

Data Cleaning

The AQI dataset was obtained from the EPA. This data was a series of daily datasets for each pollution molecule and the resulting AQI by county, so each of the individual molecule files were joined together using county and state. Rows that contained NULL values or duplicate dates in any of the air quality molecule columns were removed.

The hospitalization dataset was obtained from catalog.data.gov and contained US hospital admissions, DRG codes, and average charges per DRG per hospital by county. This dataset contained aggregated values for the entire year, rather than a daily granularity. This dataset includes data for all hospitalization records, and so it was filtered using respiratory diagnosis codes to remove any data unrelated to lung disease. The key for this table was hospital provider, rather than county. This was fixed by joining the hospitalization dataset with a dataset that includes county and zip codes so that the hospitalization dataset could be joined with the air quality dataset on county and state. This dataset was then inner joined with the air quality dataset on county and state.

The population dataset was obtained from the Census website, and has annual population data by state and county. This dataset was joined with the Hospitalization and AQI dataset on state and county.

Key variables in the AQI/Hospitalizations/Population joined dataset are: year; state; county; AQI (measured from CO, NO_2, ozone, particulate matter); number of days with good, moderate, and unhealthy AQI; population count; DRG (diagnosis) code; and number of hospital discharges.

A second joined dataset was made between the AQI dataset and the COVID-19 dataset. The COVID-19 dataset was obtained from github, and provides the daily number of cases and deaths per day by county. The number of cases and deaths was aggregated on a yearly basis. Mean, medium, min, and max number of cases and deaths per day in a year were created. The COVID-19 dataset used state abbreviations, while the AQI dataset has state names fully spelled out, so the COVID-19 dataset state column was converted to the full state name. The COVID-19 dataset was then joined with the AQI dataset on county, state and year. All data files were imported into Postgres database and SQL was used to clean, aggregate, and join them together.

Key variables in the AQI/COVID-19 joined dataset are: year; state; county; AQI; number of days with good, moderate, and unhealthy AQI; total COVID-19 cases; average cases per day; maximum cases in a day; average deaths per day; total deaths per year; and maximum deaths in a day.

Exploratory Data Analysis

To create the heat map, the hospitalizations dataset was used, and AQI was chosen as the independent variable. A dataset that correlates county names to numerical codes, or FIPS, was joined to the AQI/hospitalizations/population dataset. Note that due to the FIPS join, 2% of data was lost, and the county number for the heatmap dataset was 596. The US has a total of 3007 counties, so many counties are not depicted on the heat map range. As shown on the maps, the counties with data are represented by a color on the scale, while the counties without data are shaded in gray. The AQI heat map can be found in Figure 1.

The FIPS county codes were used as the geographical basis for the heat map. Two packages were uploaded in R (ggplot2, and usmap) to create the heat maps. A heat map was created for AQI and for hospital discharges. For the AQI heat map, median AQI was chosen as the value to be mapped to the FIPS location. For the hospital discharges heat map, to ensure that population bias was removed and to create a more accurate representation of hospitalizations and its correlation with AQI, we chose to take the negative log of the ratio of respiratory hospitalization discharges to a county's population and

map it against the FIPS codes. That being said, a correlation can be qualitatively identified between AQI and respiratory hospitalization discharges, especially in the western region of the US. The hospitalization heat map can be found in Figure 2.

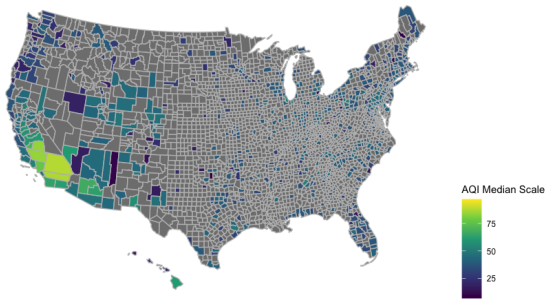

Figure 1: AQI Heatmap

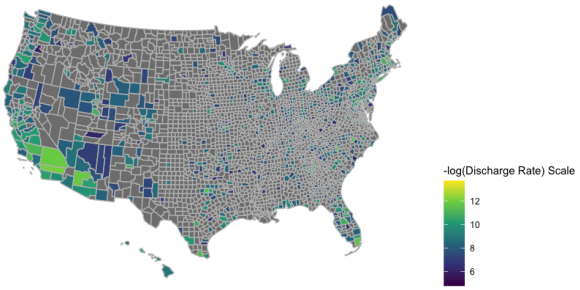

Figure 2: Hospitalization Heatmap

To gain a better understanding of the relationship between air quality index (AQI) and hospitalization rates, we created a scatter plot of the two variables. While the majority of data points appeared to be clustered in the middle range of the plot, a clear upward trend was evident, suggesting a positive correlation between median AQI and hospitalization rate. This depiction can be found in Figure 3.

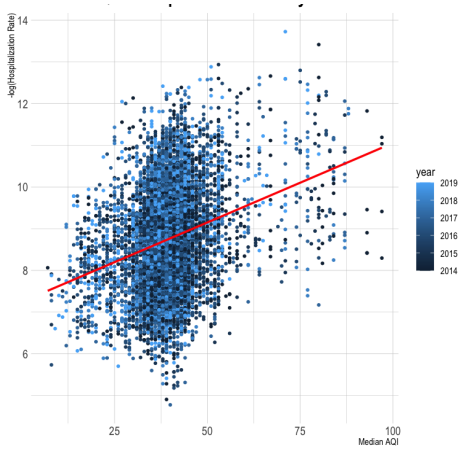

Figure 3: Median AQI vs Hospitalization Rate by Year

Two correlation matrices were created to view the coefficients for different variables. The correlation matrices can be found in Figures 4 and 5. The first was created based on the hospitalization dataset, where respiratory cases were summed by county, and adjusted per 100,000. There were not large correlations between any of the variables and respiratory-related hospitalizations. However, there were some small correlations with the number of unhealthy days of AQI, the median AQI, number of days with NO_2, and number of days with ozone.

The second correlation matrix that was created was to see the relationship between COVID-19 deaths and cases and AQI metrics. Median AQI, number of days with an AQI that is unhealthy for sensitive individuals, and number of moderate AQI days had positive correlations with the number of daily deaths and cases.

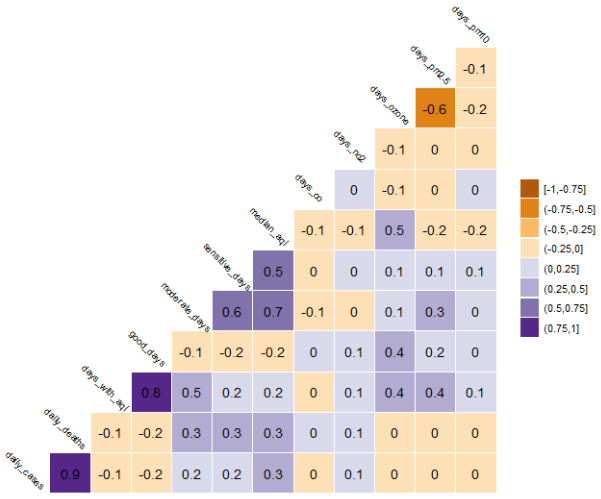

Figure 4: Hospitalization data correlation matrix

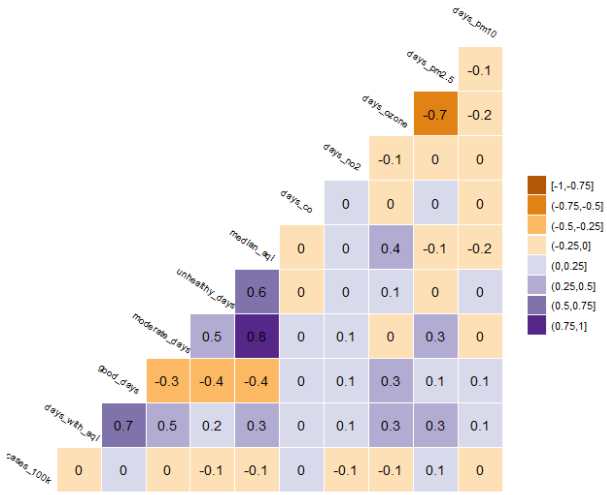

Figure 5: COVID-19 data correlation matrix

OVERVIEW OF MODELING

Initial Model Creation

For the first of two passes in model creation, several regression models were tested using the joined table “annual_aqi_hospital_population_joined” (which includes AQI and hospitalization records by county). Linear-linear, log-linear, linear-log, and log-log regression models were all tested. Several groups of predictor variables were tested using total_discharges, discharges_per_capita, and the log of these as response variables.

The first group of predictor variables were the following or the log of the following: good_days, moderate_days, unhealthy_sensitive_groups_days, unhealthy_days, very_unhealthy_days, hazardous_days, max_aqi, X90th_percentile_aqi, median_aqi. Four models were created regressing on total_discharges or log(total_discharges). Four models were created regressing on discharges_per_capita or log(discharges_per_capita). The log-log model regressing on log(discharges_per_capita) had the highest R-squared value of the eight models using these predictor variables, with an R-squared of 0.11836.

The second group of predictor variables were the following or the log of the following: days_co, days_no2, days_ozone, days_pm2.5, days_pm10. Four models regressed on total_discharges or log(total_discharges). Four models regressed on discharges_per_capita or log(discharges_per_capita). The log-log model regressing on log(discharges_per_capita) had the highest R-squared value of the eight models using these predictor variables, with an R-squared of 0.12922.

The third group of predictor variables included geographic, diagnostic, and AQI variables. A variable was created to combine State and County, and this new variable was then used to create indicator variables. Due to the size of the data and resultant runtime, only linear and logistic regression were tested. Two models regressed on total_discharges and log(total_discharges). Two models regressed on discharges_per_capita and log(discharges_per_capita). The logistic model regressing on log(discharges_per_capita) had the highest R-squared value of the four models using these predictor variables, with an R-squared of 0.83982. Not surprisingly, these models performed much better than the models of the previous two groups given the presence of diagnostic information in the predictor variables. This was addressed in the next set of models.

The fourth group of predictor variables were geographic and AQI variables, omitting diagnostic from the regression. Two models regressed on total_discharges and log(total_discharges). Two models

regressed on discharges_per_capita and log(discharges_per_capita). The logistic model regressing on log(discharges_per_capita) had the highest R-squared value of the four models using these predictor variables, with an R-squared of 0.41460. A second model was tested using the significant variables of this logistic model, but the only variables with a p-value < 0.05 were those that indicate the County/State, and the R-squared for the new model was lower than the first, with an R-squared value of 0.40814.

The third group of predictor variables had the best performance of the initial model creation. This model is useful in identifying variation in total discharges per capita in relation to the local AQI measurement. Further refinement of the hospitalization data models was required and is discussed in the next section of this report.

For the COVID-19 data, 6 models were created. Number of Good AQI days, number of moderate AQI days, number of days unhealthy for sensitive groups, number of unhealthy days, number of very unhealthy days, number of hazardous AQI days, maximum AQI, 90th percentile AQI, and median AQI were used as the independent variables. A linear regression was created for each of the following variables as independent variables: total cases per year, avg cases per day, max cases per day, avg deaths per day, total deaths per year, and max deaths per day. The best performing linear model was using total deaths per year as the independent variable, with an R squared value of 0.2739.

The linear models were all then converted into log-linear models. While the models for case counts improved, the models for death count generally got worse. The best performing model following the log transformation was average deaths per day, with an R squared of 0.2696, lowered from the linear model R squared of 0.2719. It is theorized that the death count models got worse and the case count models got better following a log transformation because case count grew at an exponential rate during the pandemic, while the death count did not. Generally speaking, the COVID-19 models did not perform very well, as they had low R-squared values.

Refining the Models

A second pass at model creation and refining for the hospitalization data was needed. The previous models did not include the state or county variables in the model, and the R-squared was low, with negative coefficient for the AQI variable. To account for other hidden factors influencing respiratory health, we decided to add the state and county variables to the model.

The annual air quality summary data consists of variables such as number of days with good, moderate, unhealthy, very unhealthy, hazardous AQI in a year, and then the median AQI for the year. For each state and county the number of days with available data varies, and to nullify this issue, a variable was calculated to determine the proportion of days in a year for each category. [eg: good_days_prop = (number of good days / number of days with AQI available)].

The hospitalization dataset has the annual summary for the number of admissions (~discharges) for different respiratory diseases, for each state and county. But the population for each county varies drastically, hence using the number of admissions directly might not be appropriate. So a transformed variable, total_discharges_per_1000s_pop = (total_discharges / population_in_1000s), was created to eliminate population variability for different counties.

The number of admissions for different diagnostic codes in respiratory diseases category to get total admissions for different respiratory diseases for a state, county in a year. State and county were concatenated to form a variable called state_county to use it as a categorical variable in the model to compare counties.

Initially, a linear model [model_1] was created with total_discharges_per_1000s_pop as the dependent variable, and good_days_prop, moderate_days_prop, unhealthy_sensitive_groups_days_prop,

unhealthy_days_prop, unhealthy_days_prop, very_unhealthy_days_prop, hazardous_days_prop, median_aqi and state_county (factor) as independent variables. Since the AQI data is at the annual level, to capture the seasonal variations, we thought it would be better to use the additional variables mentioned above along with the median AQI value.

Once all of these changes were implemented, the resulting model had an Adj R-squared of 0.9488. State_County had 752 levels and overall the model had 759 coefficients including the intercept, out of which 530 were significant at 95% confidence level. The VIF analysis showed some of the predictor variables were highly (directly) correlated, and the correlation values are shown in Table 1.

Variable	Good days	Moderate days	Unhealthy sensitive groups days	Median AQI
Good days	1	-0.976	-0.658	-0.831
Moderate days	-0.976	1	0.483	0.793
Unhealthy sensitive groups days	-0.658	0.483	1	0.610
Median aqi	-0.831	0.793	0.610	1

Table 1: Correlation matrix for the updated hospitalization model.

The median_aqi variable was highly correlated with the variables number of good days, moderate days, etc. This makes sense as the quality indicators such as good/moderate/etc were derived from the daily AQI values. Due to this, we decided to use only the median AQI value in the model for further iterations. To account for other factors we are adding county name to the linear model.

As a next step, we created a linear model [model_2] with total_discharges_per_1000s_pop as the response variable, and median_aqi & state_count as predictor variables. The adjusted R-squared for this model was 0.9478 and the F-statistic for regression was significant as well, similar to the previous model. After adding the state_county variable, the R-squared suggests the model explains the variability in the dataset well, but it could be inflated because of too many levels in the categorical variable. This model had 753 coefficients and 526 were significant at 95% confidence level. Median AQI was one of the significant variables with a positive coefficient of 0.02423, which can be interpreted as if the median AQI increased by one point, with state_county being constant, the number of hospital admissions would increase 24.22 people. The VIF analysis showed there is no presence of multicollinearity in this new model. For the model assumptions check, there was a deviation for the constant variance assumption, and normality assumption is not met in the model. The Cooks' distance plot also showed presence of a few influential and high leverage points present in the dataset. This data is shown in Figure 6. The presence of these influential points/outliers could be causing the violation in the model assumptions.

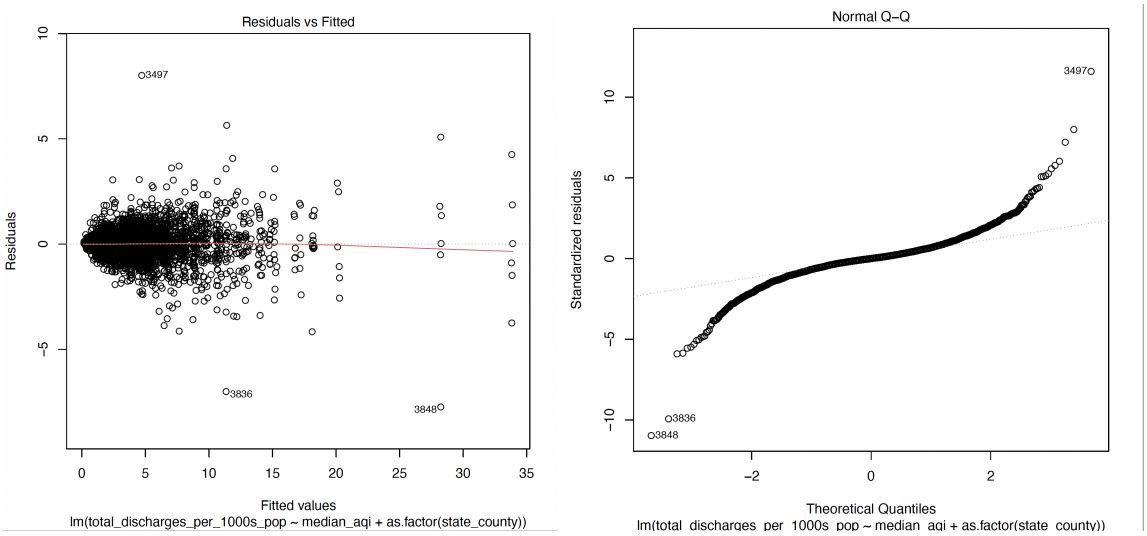

Figure 6: These plots show a violation of heteroskedasticity and normality.

To address the model assumption violations, we evaluated the log transformation on the response variable - total_discharges_per_1000s_pop (model_3). There was still presence of heteroscedasticity in the model, and the log transformation did not address it. This suggested that it could be because of the presence of influential points. As each county has a different population and different number of hospital admissions, intuitively some of them could be influencing the model.

To overcome this problem, we tried a couple of approaches instead of simply removing these influential points from the dataset. We created 4 subsets of datasets by splitting the data by total_discharges_per_1000s_pop, and median_aqi into 4 quartiles respectively. The box plot on these subset of data still showed presence of some outliers, and was not helpful to address the model assumption violation. As a next step, we removed the influential points using the cooks distance (>4), which removed 277 rows and 10 counties (originally 753 counties) from the dataset, which still provided sufficient data for the analysis.

As a next step, we created the linear model (model_4) on this dataset without influential points. This model had an adjusted R-squared of 0.9717, and had 591 coefficients significant at 95% confidence level (out of 744). The interpretation for median AQI was, if median AQI increases by one point, with state_county being constant, number of hospital admissions increases by 26.325 people. This was similar to the linear model (model_2). The model assumptions plot showed that there is no presence of heteroscedasticity in the dataset, and the normality assumption was also satisfied. Hence this linear model was a good fit for the dataset, and also explains the variability in the dataset well. This model (model_4) is chosen as our final model, and the boxplot and Q-Q plot can be found in Figure 7.

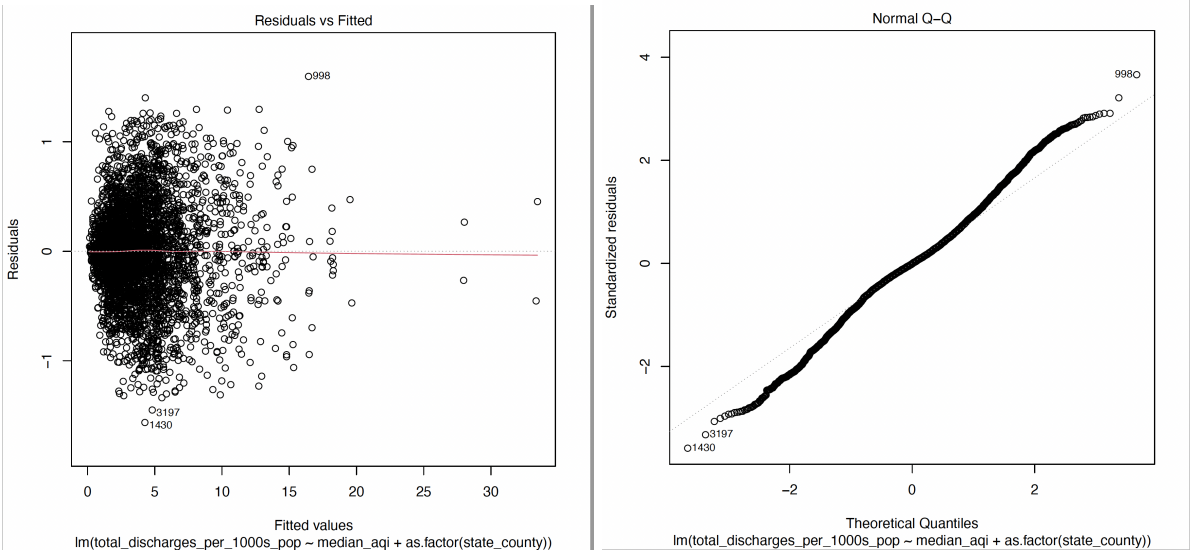

Figure 7: These plots show that the normality and heteroskedasticity assumptions are met in model_4.

To further improve this model, we added years as an additional factor variable to check the effect over the years. Interestingly, when we added the years as a predictor variable, the coefficient for median_aqi was not significant anymore. This suggests that it is sufficient to use median_aqi in the model. The general pattern of this data agrees with our initial hypothesis that AQI and hospitalization discharges are positively correlated. To validate this we plotted the median_aqi over the years for all counties and we could confirm this for most of the counties. This plot can be found in Figure 8.

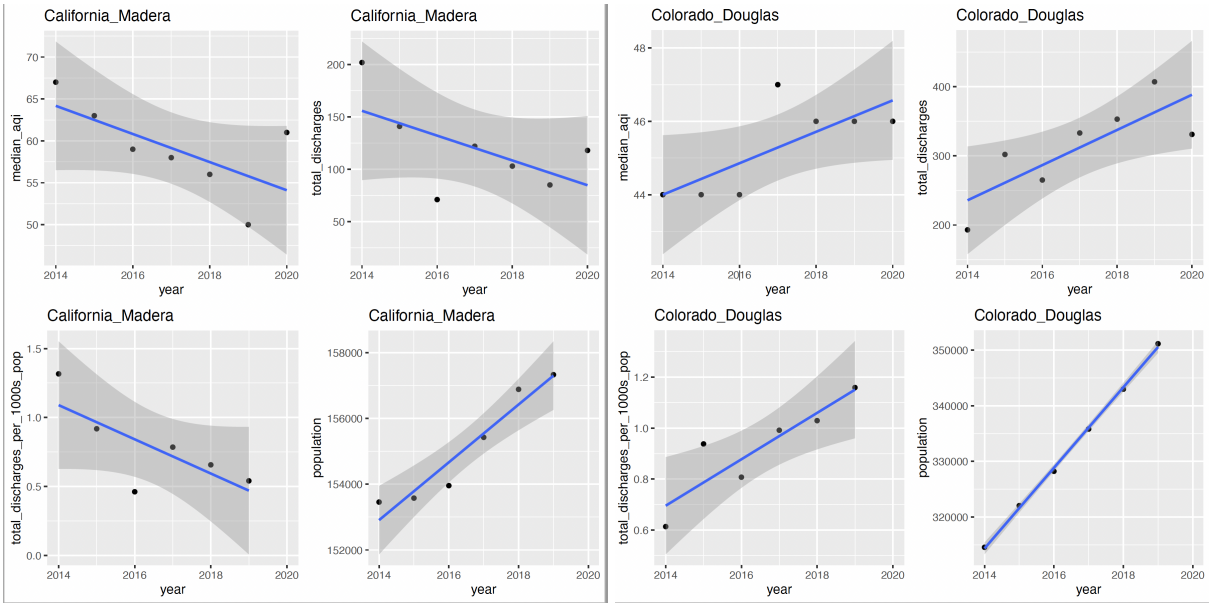

Figure 8: Median AQI and total discharges were plotted from 2014-2020 for each county. The data indicates a positive correlation between AQI and hospitalizations.

To refine the model for COVID-19 cases, a variance inflation factor (VIF) analysis was conducted on all COVID-19 models. VIF identified that only the variable X90th_percentile_aqi exhibited significant multicollinearity with the other predictor variables. Our chosen cut-off value for VIF was 5, which is commonly used as a threshold for detecting multicollinearity. The variable X90th_percentile_aqi had a VIF of 8.5, indicating that it was highly correlated with other predictors in the models.

To address this issue, we removed the X90th_percentile_aqi variable from the models and reran the analyses. Interestingly, we observed little to no change in the R-squared values, indicating that this variable was not contributing significantly to the predictive power of the models. This finding suggests that other variables included in the models may be better predictors of COVID-19 outcomes, and that the removal of X90th_percentile_aqi did not compromise the validity of the models.

Overall, our VIF analysis helped us to identify and address multicollinearity in our models, which is essential for ensuring the reliability of statistical analyses and the accuracy of our findings. Given that the COVID-19 model still had such a low R squared value following model refinement, we believe that there is not enough evidence to prove that poor air quality can be correlated with or cause increased levels of COVID-19 cases.

CONCLUSION

Based on our findings, we were able to determine that there is a positive correlation between AQI and hospitalization rate. The final model (model_4) correlates that for every 1 unit increase in AQI, there is a resulting increase of 26 additional people hospitalized for respiratory diseases, keeping the state and county constant. This model performed well in explaining the correlation between respiratory health and AQI. The R squared value was 0.9717, which indicates that the model can explain the variability in the dataset. The model satisfied all regression assumptions. However, the model's R-squared could be inflated because of the categorical variable (state_county), and since the analysis is at yearly level, there were not many data points per county. Perhaps, performing this analysis on a daily or weekly level would give more data points per county and could provide a better model with more accurate R-squared. Grouping the counties into rural and urban areas would be helpful to better measure the effect of AQI in health. These models were not sufficient to make any causal inferences, as other variables need to be included in the model.

The respiratory health is not exclusively dependent on AQI - there are other factors that impact respiratory health as well (such as age, BMI, smoking status, family history of respiratory illness, etc.) and would likely be highly correlated with respiratory illness. We anticipate that adding these variables would improve the model accuracy and provide more valuable insights on how specifically different factors impact respiratory health.

WORKS CITED

1) Liu, H., Hu, T., & Wang, M. (2021). Impact of Air Pollution on Residents' Medical Expenses: A Study Based on the Survey Data of 122 Cities in China. *Frontiers in Public Health*, *9*. https://doi.org/10.3389/fpubh.2021.743087
2) *The Impact of Air Quality on Hospital Spending*. (n.d.). RAND. https://www.rand.org/pubs/periodicals/health-quarterly/issues/v2/n3/06.html
3) Robinson, D. L. (2005). Air pollution in Australia: review of costs, sources and potential solutions. *Health Promotion Journal of Australia*, *16*(3), 213–220. https://doi.org/10.1071/he05213
4) https://aqs.epa.gov/aqsweb/airdata/download_files.html#AQI
5) https://github.com/nytimes/covid-19-data
6) https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html
7) https://catalog.data.gov/dataset/medicare-inpatient-hospitals-by-provider-and-service-9af02
    a) https://www.unitedstateszipcodes.org/zip-code-database/
    b) https://www.cms.gov/icd10m/version37-fullcode-cms/fullcode_cms/P0007.html
8) https://statisticsbyjim.com/regression/r-squared-too-high/
9) Dealing with categorical features of high cardinality - Medium article
10) Effect of many levels in categorical variable in model inference - Stackoverflow post

11) Collapsing categorical variables with many levels - Stackoverflow post