How Does Air Quality Affect COVID-19 Hospitalizations?

Introduction:

The introduction of COVID-19 as a common and widespread virus has prompted a variety of studies, many of which have given us key insight into its nature. Since COVID-19 is mostly known as a respiratory illness, I wanted to specifically investigate the effects of air quality on recent COVID hospital admissions. I hoped to find a way to predict new county COVID-based hospital admissions using county air quality and population data.

Background Research:

I found many studies that had been conducted on this topic. A widespread study by the National Library of Medicine found there to be significant correlations between air pollution and COVID-19 cases, and that air pollution may be associated with worse COVID-19 outcomes. It is detailed in the article *The impact of air pollution on COVID-19 incidence, severity, and mortality: A systematic review of studies in Europe and North America.* Another study from Harvard also showed an association between long term exposure to air pollution and higher COVID-19 mortality rates. In similar studies it has also been found that NO_2 concentration is positively associated with the transmission ability of COVID-19, that places with higher levels of nitrogen dioxide pollution in the five years before the pandemic had 22% more Covid-19 cases, while higher levels of small particle pollution saw a 15% rise in infection rates, and that particulate matter pollution is positively associated with increased cases of COVID-19. This study is detailed in the article *Coronavirus and Air Pollution.*

Data Exploration:

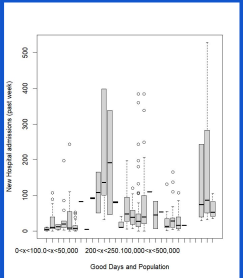
I combined air quality data from the EPA (Environmental Protection Agency), recent COVID hospitalization data from the CDC (Centers for Disease Control and Prevention), and county population data from the USDA ERS (U.S. Department of Agriculture, Economic Research Service). All data was selected for the year 2022. The air quality data contained 1001 data points over a wide spread of variables, including air quality maximums, good days, unhealthy days, and days with certain pollutants or particulate matter present in the air. The COVID hospitalization data provided 3227 data points on county by county covid hospital admissions from the past week. Its main two variables of interest were “New hospital admissions of confirmed COVID-19 in the past week,” and “New COVID-19 hospital admissions per 100,000 population in the past week.” Lastly, the county population data contained 3288 data points from 2020 to 2022, though I only used data from 2022. By numbering the data and using a random number generator, I selected 402 corresponding data points from all three datasets and compiled them into one dataset. Later, I omitted two data points from the sample; one lacked necessary information, and the other was an extreme outlier.

Hypothesis:

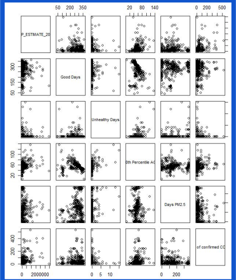
Going into this project with a wealth of studies at my back, I believed I would find a strong correlation between air quality and New Covid Hospital Admissions, though I was unsure what factors would affect the relationship most.

Inference Test:

At first, I wanted to run a two-way ANOVA, testing Good Days and Population (2022) against New Covid Hospital Admissions, but I soon realized that one of the most important checks, equal variance, was not met in the slightest by my dataset. Therefore, I switched my single two-way test into two one-way Welch’s ANOVAs, which meant the equal variance issue was nullified. This, however, also rid me of my interaction term between Good Days and Population, but I concluded that an accurate test with less information was better than an inaccurate test with more information.



Box plot (left) reveals unequal variance, so one of the major checks for two-way ANOVA is not met.
Bottom row of the scatter plot matrix (right) compares Population, Good Days, Unhealthy Days, 90th Percentile AQI, and Days PM2.5 to New hospital admissions of confirmed COVID-19 past week (total).



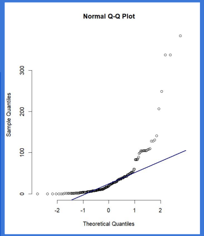
Two One-way Welch’s ANOVAs:

μ_1 = new county covid hospital admissions
μ_2 = good days
μ_3 = county population (2022)

Test 1: good days vs. new county covid hospital admissions
H_0: $\mu_1 = \mu_2$
There is no effect of county good days on new county covid hospital admissions.
H_a: $\mu_1 \neq \mu_2$
There is an effect of county good days on new county covid hospital admissions.
α = 0.05

Test 2: county population vs. new county covid hospital admissions
H_0: $\mu_1 = \mu_3$
There is no effect of county population on new county covid hospital admissions.
H_a: $\mu_1 \neq \mu_3$
There is an effect of county population on new county covid hospital admissions.
α = 0.05

Normal Probability Plot for two Welch’s ANOVAs.



Checks:

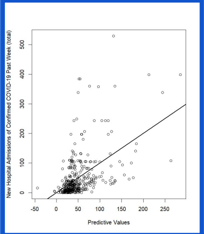
Normality: Normal probability plot does not appear normal; it has a major curve. We have a large enough sample size (n = 400 and with CLT, n>30). Proceed with caution.
Equal Variance: Variance on boxplots does not appear equal, but we’re running two Welch’s ANOVAs. Proceed.
Independent: Random Sample. Assume independence.

Source of Variation:	df:	F-statistic:	p-value:
Good Days:	df1 = 5 df2 = 51.08	2.91	0.0220
Population:	df1 = 4 df2 = 64.86	14.02	0.0000

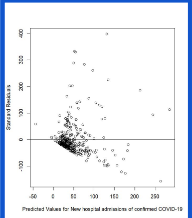
Conclusions:

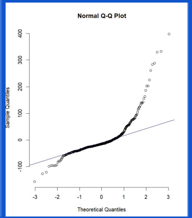
Since the p-value for Good Days (0.0220) is less than **α** (0.05), Good Days appear to have a statistically significant effect New Covid Hospital Admissions in the past week. Reject null hypothesis that there is no effect of county good days on county covid hospitalizations.

Since the p-value for Population (0) is less than **α** (0.05), Population appears to have a statistically significant effect on New Covid Hospital Admissions in the past week. Reject null hypothesis that there is no effect of county population on county covid hospitalizations.



Predictive model with regression line.





Residual plot (left) and normal probability plot (right) for Model Utility F-test.

Predictive Model:

After finding through my Welch's ANOVAs that Good Days and Population (2022) have a statistically significant influence on New Covid Hospital Admissions, I decided to create a predictive model that would hopefully help predict New Covid Hospital Admissions with air quality factors and population.

Model Utility:

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = 0$
H_a: At least one slope ≠ 0
α = 0.05

x_1 = Good Days
x_2 = Unhealthy Days
x_3 = County Population (2022)
x_4 = 90th percentile AQI
$x_5 = x_1 * x_3$ (interaction between Good Days and Population)
$x_6 = x_2 * x_3$ (interaction between Unhealthy Days and Population)
$x_7 = x_4 * x_3$ (interaction between 90th percentile AQI and Population)
$\hat{y}$ = predicted new covid hospital admissions (per county, past week)

β_1 = true slope regressing Good Days to predict new covid hospital admissions (past week).
β_2 = true slope regressing Unhealthy Days to predict new covid hospital admissions (past week).
β_3 = true slope regressing County Population for 2022 to predict new covid hospital admissions (past week).
β_4 = true slope regressing 90th percentile AQI to predict new covid hospital admissions (past week).
β_5 = true slope regressing the interaction between Good Days and County Population for 2022 to predict new covid hospital admissions (past week).
β_6 = true slope regressing the interaction between Unhealthy Days Days and County Population for 2022 to predict new covid hospital admissions (past week).
β_7 = true slope regressing the interaction between 90th percentile AQI and County Population for 2022 to predict new covid hospital admissions (past week).

b_1 = predicted slope regressing Good Days to predict new covid hospital admissions (past week).
b_2 = predicted slope regressing Unhealthy Days to predict new covid hospital admissions (past week).
b_3 = predicted slope regressing County Population to predict new covid hospital admissions (past week).
b_4 = predicted slope regressing 90th percentile AQI to predict new covid hospital admissions (past week).
b_5 = predicted slope regressing the interaction between Good Days and County Population for 2022 to predict new covid hospital admissions (past week).
b_6 = predicted slope regressing the interaction between Unhealthy Days Days and County Population for 2022 to predict new covid hospital admissions (past week).
b_7 = predicted slope regressing the interaction between 90th percentile AQI and County Population for 2022 to predict new covid hospital admissions (past week).

Chosen model:

$\hat{y} = 6.86 - 0.915x_1 - 3.105x_2 + 2.709x_3 - 0.0624x_4 + 00055x_5 - 0.008x_6 + 0.00515x_7$
$R^2 = 0.2688$
$R^2(\text{Adjusted}) = 0.2558$
$S_e = 62.43$
F-Statistic = 20.49

F-test For Model Utility:

Checks:

Normality: Curved normal probability plot. Sample size = 400 (CLT, n>30). Proceed with caution.
Equal Variance: Bad looking residual plot. Box plots show lack of equal variance. Proceed with caution.

Calculations:

k = number of predictors
n = number of data points
$R^2 = 0.2688$

$$F = \frac{R^2/k}{(1-R^2)/(n-(k+1))} \quad w/df = \frac{k}{n-(k+1)}$$

$$F = \frac{0.2688/7}{(1-0.2688)/(400-(7+1))} \quad w/df = \frac{7}{400-(7+1)}$$

$$F = 20.5864 \quad w/df = 0.017857$$

$$\text{p-value} = p(F > 20.5864)x2 = < 2.2x10^{-16} \approx 0$$

Conclusion:

Since the p-value (0) < **α** (0.05), we reject the null hypothesis that all slopes equal zero. Therefore, there is convincing evidence to suggest a useful linear relationship that uses Good Days, Unhealthy Days, 90th percentile AQI, County Population (2022), and several interactive terms to predict new covid hospital admissions (per county) from the past week.

Conclusion and Reflection:

In conclusion, it appears possible for us to create a statistically significant model that uses air quality and population to predict new COVID hospital admissions, though its accuracy may be limited and the conditions for its creation were sometimes violated. Therefore, we can infer that there are other significant factors besides air quality and population that affect new covid hospital admissions. My process on this project could have been improved in several ways. For one, I could have spread my data over a number of years instead of just 2022, because COVID also may have affected air quality in certain counties. For two, I could have gone about random selection in a more efficient way from the start instead of reverting to coding from manual selection about halfway through.