Discovering Relationships between COVID-19 Spread and Green Spaces in New York City

Cynthia Chen, Lauren Chen, Iris Lang, Michelle Zhu

February 4, 2023

1 Problem Statement and Research Question

In 2020, as the COVID-19 pandemic spread rapidly worldwide, New York City (NYC) emerged as a major epicenter of the disease outbreak in the United States. According to Thompson et al. [2020], from March 2020 to May 2020, approximately 203,000 laboratory-confirmed COVID-19 cases were reported to the NYC Department of Health and Mental Hygiene. To combat such disease spread, we turn towards green spaces, which provide not only a way for people to physically distance but have also been shown to have immense mental health benefits on nearby residents as discussed in Wood et al. [2017]. In this project, we seek to discover underlying relationships between green space attributes in certain neighborhoods and COVID-19 rates in New York City. Identifying such relationships will not only lead to a better understanding of how green spaces impact the spread of disease in metropolitan environments, but also help generate recommendations for preparing cities for future pandemics.

Our **research question** is as follows: how did green spaces affect COVID-19 spread in NYC? What neighborhood-specific relationships can we discover that can help drive better city planning and green space related decisions to prepare cities for pandemics in the future? In considering these questions, we made sure to account for confounding variables including income and population density by visualizing their relationships with tree coverage.

We chose New York City as our location of analysis for several reasons. First, NYC is divided into several geographically and socioeconomically distinct neighborhoods, which provides grounds

for analysis of relationships between income, population density, and the existence of green spaces. Furthermore, NYC was hit especially hard by the COVID-19 pandemic in the spring of 2020, leading us to want to investigate how we could develop a model to generate recommendations on how to prevent such ravaging outbreak in the future in densely-populated areas.

In this report, we propose a regression model and statistical methods for investigating the relationship between green space features (tree cover, park access) and COVID-19 spread (hospitalization/death rates, percent positivity) in New York City. We first outline our data cleaning and feature extraction methods that we used to extract features involving green spaces and COVID-19 metrics by zip code in NYC. Next, we describe our modeling approach in detail and present the regression results from our modeling. Finally, we cover the limitations and future avenues of exploration of our work.

2 Methods

2.1 Data and Code

We investigated the following datasets in this project:

- **us_cities_zips.csv**: A dataset that contains data on US cities including population density and number of residents.
- **percent_cover_zipcode_tabulated_areas.txt**: A dataset mapping zip code to the percent of land that is public accessible parkland.
- **5_million_trees_us_cities.csv**: A dataset of 5 million city trees from cities across the US. The relevant features from this dataset include trunk diameter, tree condition, whether the tree is native, and the number of trees by zip code.
- An **external dataset on COVID-19 metrics in NYC** provided by the NYC Department of Health and Mental Hygiene (can be found at this link) that provides data on the COVID-19 disease starting from February 2020 to the present day in monthly increments. The relevant metrics given by this dataset include weekly case, hospitalization, and death rates.
- **urban_tree_canopy.csv**: A dataset built to help understand tree canopy and its assocation with income inequality across over 5000 US cities.

The code written for this project, including data preprocessing steps, feature extraction, and our statistical models, can be found at the following link: github.com/cynthia9chen/green-spaces.

2.2 Data Preprocessing and Feature Engineering

After data collection, the first step was to clean the data, extract the relevant features, and aggregate.

From the `us_cities_zips.csv` dataset, we extracted population and population density (people/km^2) values by zip code.

From `percent_cover_zipcode_tabulated_areas.txt`, we extracted `pc_cover`, the percent of land that is public accessible parkland.

In the `5_million_trees_us_cities.csv` dataset, each row represents a different tree. We first filtered for trees in New York City zip codes, namely those between 10001 and 11697, and then extracted the relevant features, listed below.

- `diameter_breast_height_CM`: trunk diameter in cm at breast height
- `condition`: tree condition as coded by the city-specific protocol converted to standardized conditions
- `native`: whether the tree is native to the state, not native, or of unknown status
- `tree_cover`: the number of trees by zip code

To encode the categorical features, such as `condition` and `native`, we used the following mapping functions:

`condition`	Encoded Value
poor	0
fair	1
good	2

`native`	Encoded Value
introduced	0
naturally_occurring	1

The final dataset had 682,853 rows.

To aggregate information on COVID-19 metrics in NYC, from our external NYC COVID-19 dataset, we imported three separate files (`deathrate-by-modzcta.csv`, `hosprate-by-modzcta.csv`, `percentpositive-by-modzcta.csv`) as pandas Dataframes. These Dataframes included information on the hospitatalization rate, death rate, and percent positivity rate per 100,000 residents, each by zip code, over the course of the pandemic (February 2020 to current). There are 177 NYC zip

codes represented in these datasets, and for each zip code, we calculated the average hospitalization rate (`hosp_avg`), average death rate (`death_avg`), average hospitalization rate (`percentpos_avg`). The assumption that average rates are a strong indicator of overall COVID-19 spread is discussed in Section 4.1.

Finally, we generated our overall features dataset by merging all Dataframe discussed above by zipcode into one large Dataframe. Since we investigated a variety of datasets where some zip codes were not represented in a certain dataset, we only kept information on zip codes that were represented in all the datasets we examined (dropping entries for zip codes that had missing or NaN values). Our overall feature set was used for our later modeling and regression analyses.

2.3 Multiple Linear Regression Model

Using R, we developed multiple linear regression models one for each COVID-19 metric. To start off, we performed exploratory data analysis: univariate and multivariate.

For univariate exploratory data analysis we plotted histograms for each individual response and predictor variable as shown in Figure 1, we see that the distributions for PC Cover and Tree Cover are right-skewed and the distribution for Condition Encoded is left-skewed. Therefore we transform the data such that it is normally distributed by taking the log.

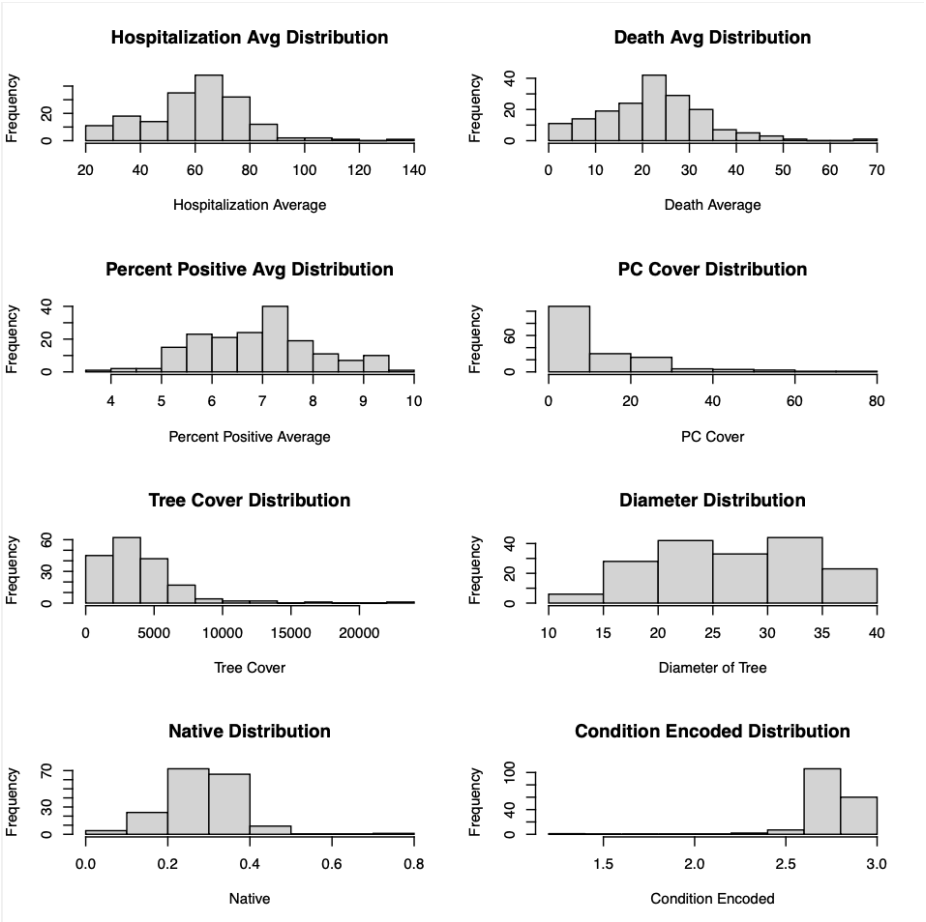

Figure 1: Exploratory data analysis conducted on single variables that we considered in our multi linear regression analysis.

Next, we performed multivariate exploratory data analysis by plotting scatter plots of all the predictor variables against each of the response variables. As shown in Figures 2, 3, and 4, we see that the relationship between each predictor variable and each response variable is weak to moderately positive. There does not seem to be any curves or any indication of using a nonlinear model.

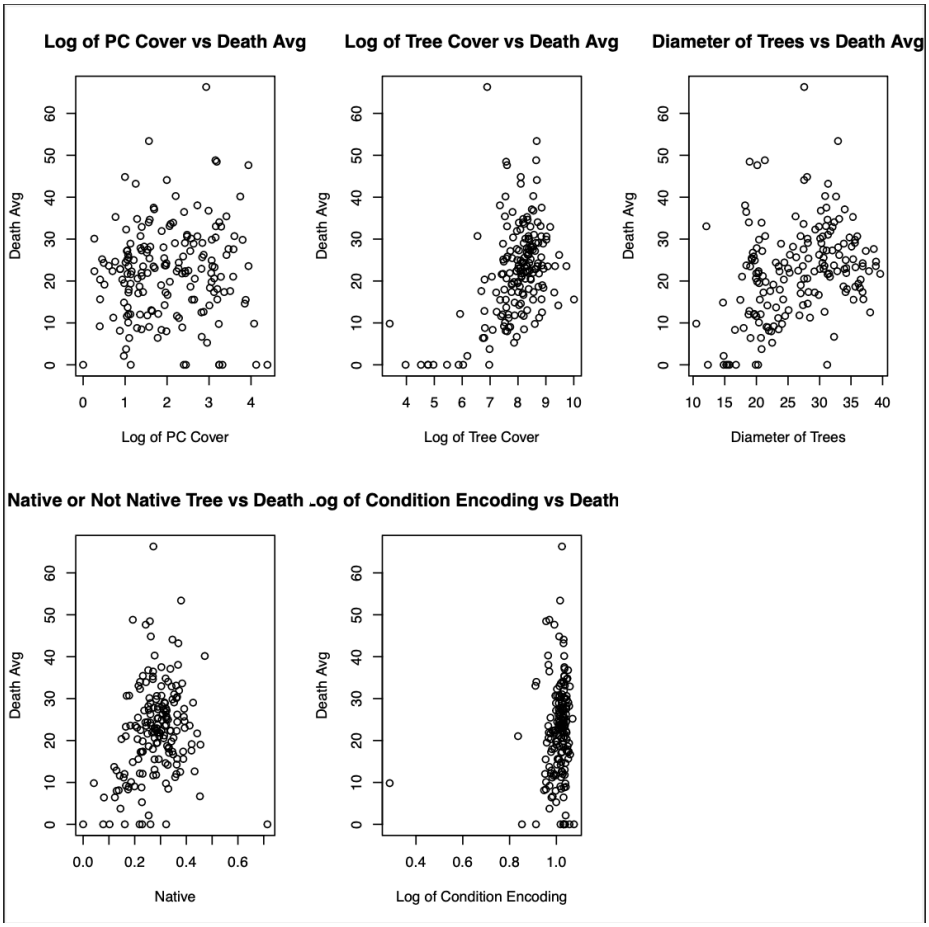

Figure 2: Multivariate EDA for Death Model

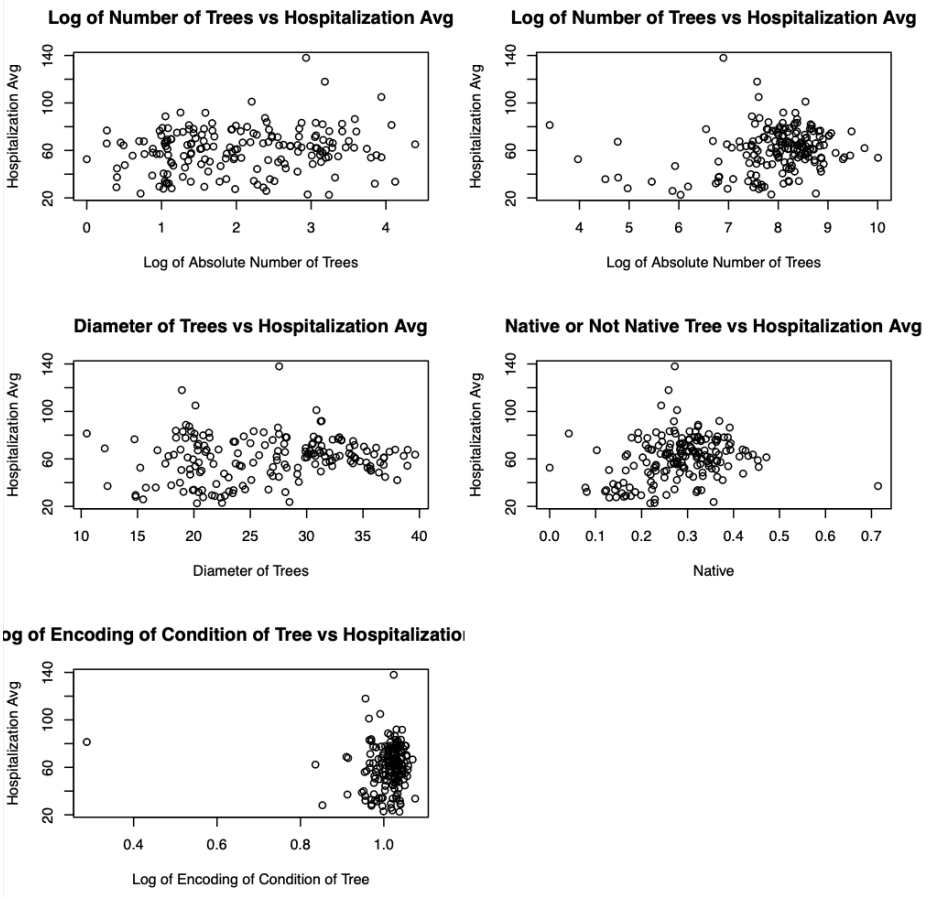

Figure 3: Multivariate EDA for Hospitalization Model

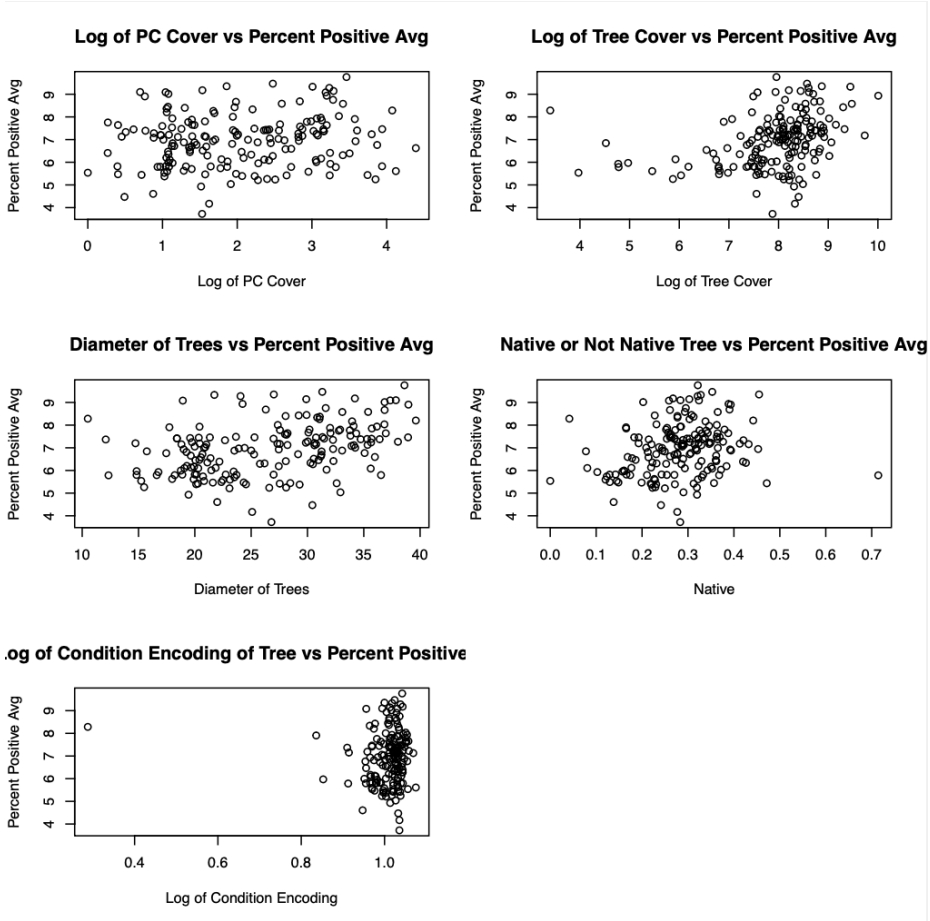

Figure 4: Multivariate EDA for Percent Positive Model

After exploring the data, we fitted each model. We first fitted each model with all of the predictor variables. Then, we used leave-one-out cross validation (LOOCV) to determine which predictor variables to keep in each model to give us the model with the lowest LOOCV.

2.4 Minimizing Effects of Confounding Variables

Given that our goal is to determine the relationship between green spaces and the COVID-19 pandemic, we want to minimize confounding factors. In particular, we analyze income levels and population density levels within the state of New York, and their relationships with tree coverage.

In the `urban_tree_canopy.csv` dataset, we obtain data for all 155107 census blocks in the state of New York. Each block contains a population density level (on a scale of 1-4, 1 being the least populus level), an income level (on a scale of 1-4, 1 being the highest income), and a mean tree cover

percentage of the said block. Section 3.4 details the results of our analysis and its impacts on our modeling.

3 Results

3.1 Regression Results: Model Coefficients and Model Evaluation

3.1.1 Hospitalization Rates

The best model for predicting COVID-19 hospitalization rates only used the variable `native`, and the p-value for that coefficient was 0.00135, which is extremely significant. The value of the coefficient suggests that having only naturally occurring trees in a zip code correspond to a 49.46% increase in COVID-19 hospitalization rates as compared to having only introduced trees. However, the R^2 value is 0.05748, which is low. This indicates that while the nativeness of a tree is significantly related to hospitalization rates, very little of the variance in hospitalization rates is captured by the nativeness of trees.

```
Call:
lm(formula = hosp_avg ~ native, data = combined, x = T)

Residuals:
    Min      1Q  Median      3Q     Max
-44.728 -10.886   0.101  11.641  78.034

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   46.505      4.494  10.348  < 2e-16 ***
native        49.458     15.182   3.258  0.00135 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.86 on 174 degrees of freedom
Multiple R-squared:  0.05748,   Adjusted R-squared:  0.05207
F-statistic: 10.61 on 1 and 174 DF,  p-value: 0.001351
```

Figure 5: Model output for predicting COVID-19 hospitalization rates.

3.1.2 Death Rates

The best model for predicting COVID-19 death rates used the number of trees, tree trunk diameters, nativeness of the trees, and tree condition. The p-value for the number of trees was 2.5×10^{-6}, which is extremely significant. The p-values for the other coefficients are relatively low as well with non exceeding 0.20. These results also make intuitive sense, as we see that trees in better condition are associated with lower death rates. The R^2 for this model is 0.2406. This is better than the previous model for hospitalization rates, suggesting a stronger relationship between green space features and COVID-19 death rates.

```
Call:
lm(formula = death_avg ~ log(tree_cover) + diameter_breast_height_CM +
    native + log(condition_enc), data = combined, x = T)

Residuals:
    Min      1Q  Median      3Q     Max
-19.262  -6.638  -0.728   3.959  49.413

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.9363    12.0572  -0.161   0.8726
log(tree_cover)              4.7974     0.9846   4.873  2.5e-06 ***
diameter_breast_height_CM    0.1869     0.1401   1.334   0.1840
native                      12.2197     9.1389   1.337   0.1830
log(condition_enc)         -22.2099    13.1335  -1.691   0.0926 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.844 on 171 degrees of freedom
Multiple R-squared:  0.2406,    Adjusted R-squared:  0.2229
F-statistic: 13.55 on 4 and 171 DF,  p-value: 1.298e-09
```

Figure 6: Model output for predicting COVID-19 death rates.

3.1.3 Infection / Percent Positive Rates

The best model for predicting COVID-19 percent positive rates used the percent of greenspaces accessible, number of trees, tree trunk diameters, and tree condition. We see that all the variables have coefficients that are significantly non-zero which means they have an effect in predicting percent positive COVID-19 rates. We see that the R^2 value is 0.2529 which is the highest out of the three models. This suggests that there is a stronger relationship between green space and tree features

and COVID-19 percent positive rates.

```
Call:
lm(formula = percentpos_avg ~ log(pc_cover + 1) + log(tree_cover) +
    diameter_breast_height_CM + log(condition_enc), data = combined,
    x = T)

Residuals:
     Min       1Q   Median       3Q      Max
-3.00329 -0.60144  0.04212  0.63377  2.31955

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.25795    1.27393   5.697 5.22e-08 ***
log(pc_cover + 1)           0.15961    0.07863   2.030 0.043911 *
log(tree_cover)             0.28428    0.09960   2.854 0.004848 **
diameter_breast_height_CM   0.06422    0.01432   4.486 1.33e-05 ***
log(condition_enc)         -4.57456    1.35322  -3.380 0.000897 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.013 on 171 degrees of freedom
Multiple R-squared:  0.2529,    Adjusted R-squared:  0.2354
F-statistic: 14.47 on 4 and 171 DF,  p-value: 3.369e-10
```

Figure 7: Model output for predicting COVID-19 percent positive rates.

3.2 Checking OLS Model Assumptions

In this subsection, we check that the assumptions for Ordinary Least Squares (OLS) regression hold.

Homoskedasticity and Normality: For each of the three models, we have the residual plots and normal Q-Q plots below.

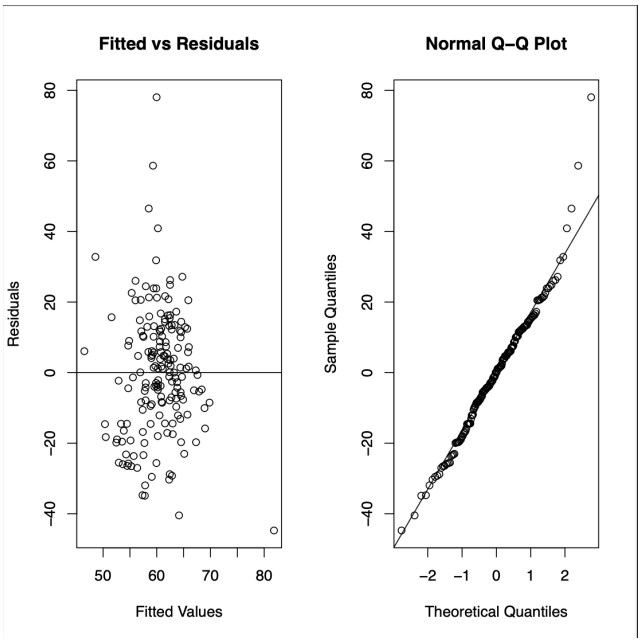

Figure 8: Residuals and Q-Q plot for predicting COVID-19 hospitalization rates.

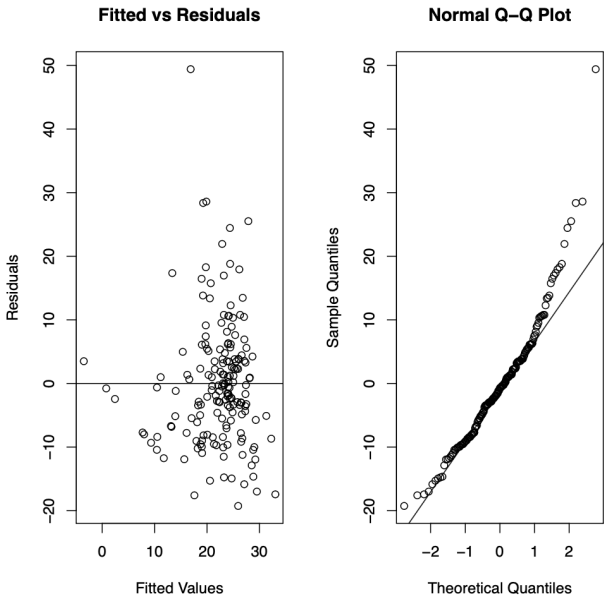

Figure 9: Residuals and Q-Q plot for predicting COVID-19 death rates.

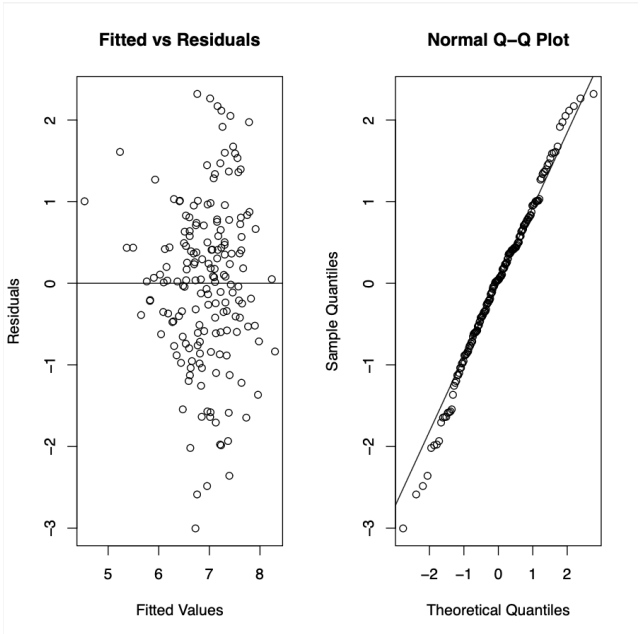

Figure 10: Residuals and Q-Q plot for predicting COVID-19 percent positive rates.

We can see in Figures 8, 9, and 10 that the residuals are randomly scattered above and below the horizontal 0 line. There does not seem to be any curved or fanning out pattern, and the variance seems to be constant throughout. Moreover, the linearity of the points in the Q-Q plots shows that the data is roughly normally distributed. This suggests that our model satisfies the homoskedasticity and normality assumptions.

No multicollinearity: To check for no multicollinearity, we calculated the correlation between all of the covariates used in the models, shown in Table 1 below. As we can see in the table, the magnitudes of all pairwise correlations are less than 0.5, so no two variables are highly correlated. Thus, our model satisfies the no multicollinearity assumption.

	pc_cover	diameter_breast_height_CM	condition_enc	native_enc	tree_count
pc_cover	1.000000	-0.051849	-0.180711	0.083486	0.027083
diameter_breast_height_CM	-0.051849	1.000000	0.442791	0.323314	0.403612
condition_enc	-0.180711	0.442791	1.000000	0.278807	0.225250
native_enc	0.083486	0.323314	0.278807	1.000000	0.273388
tree_count	0.027083	0.403612	0.225250	0.273388	1.000000

Table 1: The correlation matrix between all of the covariates used in the models.

3.3 Confounding Variables: Income and Population Analysis

In the state of New York, looking at all census blocks, each categorized into one of four different levels of income and one of four different levels of population density, in relation to tree percentage coverage, we see that there is a clear pattern of wealthier regions having higher percentages of tree coverage, and more densely populated regions having lower percentages of tree coverage.

This suggests that population density is a potential confounding factor, as it is positively correlated with tree coverage. Thus, when implementing our model to analyze the relationship between green spaces and characteristics of trees with the impacts of COVID in specific regions of New York City, we opted to exclude features such as population or population density in each zip code as predictors for our model.

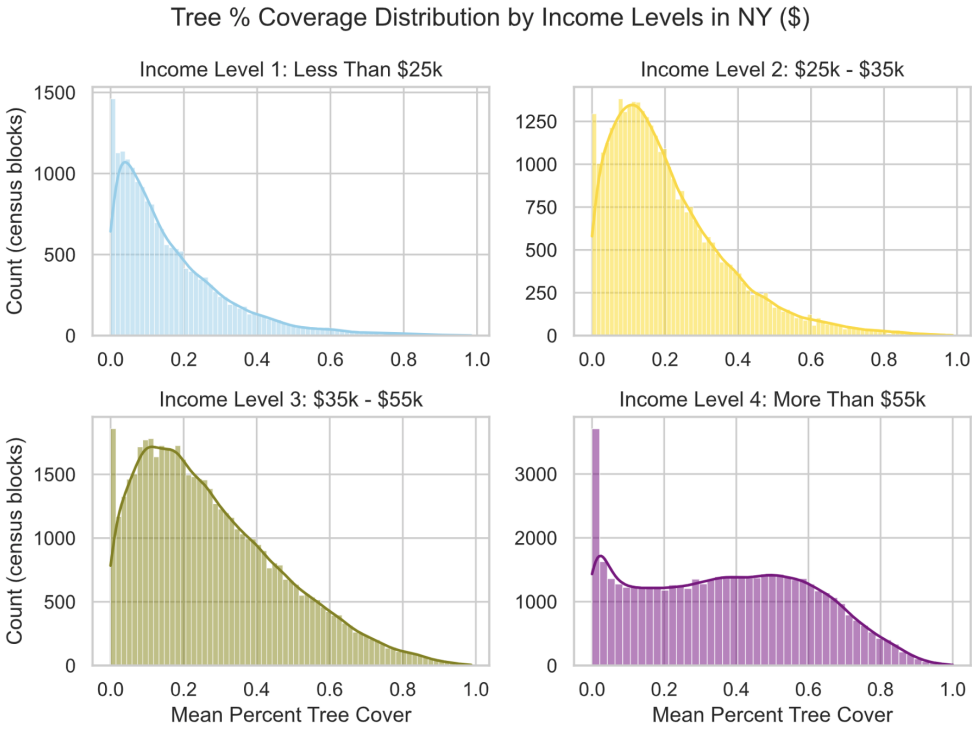

Figure 11: The distribution of tree percentage coverage by income level in NYC. We notice that mean tree cover percentages grow higher (with more values above 0.5) as we move from income level 1 to income level 4.

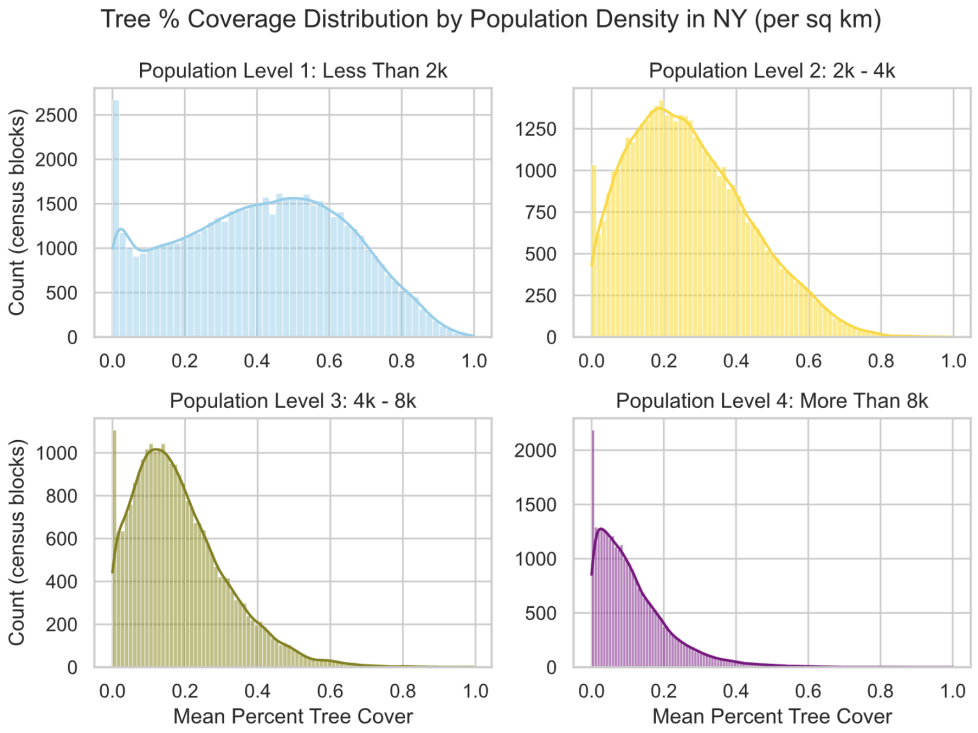

Figure 12: The distribution of tree percentage coverage by population density in NYC. We notice that as the population density increases, the mean tree cover decreases drastically as we might expect (more urban + densely packed areas might have less space for nature).

4 Discussion

4.1 Assumptions and Model Limitations

In this subsection, we outline the assumptions of our methods and why they may be reasonable for the purpose of our analysis.

We assumed that the average of hospitalization, death, and percent positivity rates over the course of the pandemic are a good indicator of COVID-19 spread in the region. Some further metrics beyond the average that could be explored are the max rate over the course of the pandemic, as well as averages across certain time periods (eg: by year) in order to account for waves of infection due to disease variants.

We also assumed that tree features do not change drastically between the most recent observation of the trees in the `5_million_trees_us_cities.csv` dataset, which were in 2015-2016, and the COVID-19 pandemic in 2020-2022.

Some limitations of our current model is that it does not account for inequality gaps and also does not account for differences across age groups. Both factors could play a big role in determining COVID-19 outcomes, and it would be important to investigate the effects of these variables to our results.

All of our analyses are also limited to the New York City area, which was a deliberate decision but also means that our results may differ for cities or areas that have different compositions from New York City. However, due to the applicable nature of our methods, with more time, we could extend our analyses to include other large cities and epicenters of COVID outbreak.

4.2 Extensions and Further Avenues of Exploration

The results from this project lend themselves to several avenues of future exploration. First, we would like to extend our analyses to other epicenters of the COVID-19 outbreak around the world including Shanghai, Los Angeles, and more. This could help reveal more insights into relationships between variables like tree cover, park cover, and tree density per block with epidemic metrics.

Furthermore, we could examine different subgroups of individuals within the city (age ranges, ethnicity, socioeconomic status) to create a more rigorous model that minimizes potential confounding variables. If given more time, we would have liked to incorporate our state-level analysis of New York and the effects of income inequality and population density on tree coverage into the model.

Lastly, a potential avenue for exploration includes a temporal analysis within a particular pandemic or outbreak. In this project, our metrics for hospitalizations, death rate, and infection rates are single numbers. This hinders our ability to analyze the pandemic from a dynamic lens. It can be reasonably assumed that tree composition in a geographic region will not vary drastically within a year, or even within a few years. However, in the span of a year, a pandemic or disease outbreak can rapidly spread, and its dynamics may change. Therefore, a temporal analysis would be effective in better understanding the effects of green spaces in relation to future disease outbreaks.

4.3 Conclusion

In our work, we propose a model for better understanding the relationship between green space features and COVID-19 spread in New York City. The COVID-19 pandemic has exposed flaws in the current green space and park system, including exacerbating existing inequalities in income. The benefits of our work arise not only from the insight into how tree attributes and COVID-19 rates affect each other, but also from the applicability of our methods. Though we only considered New York City in this report, we can easily generalize our methods to examine other areas and generate similar insights about the relationship between green spaces and disease spread. The results generated from this project can serve as a preliminary understanding of how green spaces affected the spread of COVID-19 cases in New York City. Ultimately, this can serve as a ground for recommending better governmental and infrastructural green space policies to mitigate the effects of disease spread in cities.

References

Corinne N Thompson, Jennifer Baumgartner, Carolina Pichardo, Brian Toro, Lan Li, Robert Arciuolo, Pui Ying Chan, Judy Chen, Gretchen Culp, Alexander Davidson, et al. Covid-19 outbreak—new york city, february 29–june 1, 2020. *Morbidity and Mortality Weekly Report*, 69(46): 1725, 2020.

Lisa Wood, Paula Hooper, Sarah Foster, and Fiona Bull. Public green spaces and positive mental health–investigating the relationship between access, quantity and types of parks and mental wellbeing. *Health & place*, 48:63–71, 2017.