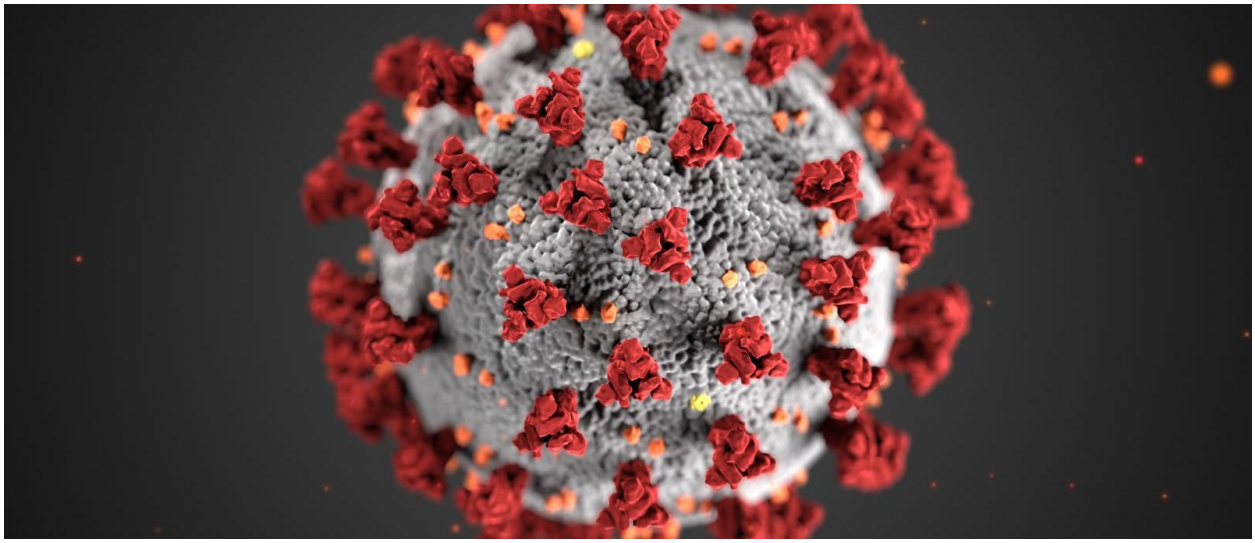
BST 263: Statistical Learning
Gabrielle LaRosa, Genevieve Lyons, Franklin Yang, Rebecca Youngerman

Socio-Economic Status and COVID-19: A New York City Case Study

A Machine Learning Approach to Identify Communities At Risk



Abstract

Background. It is well documented that those of a lower socio-economic status are at higher risk for infectious and chronic diseases. Many studies have shown that disease burden is disproportionately high for lower income communities, particularly those with minority populations.[1] These inequalities are not circumstantial but structural, reflecting a lack of resources and protections. During a public health crisis such as the one posed by COVID-19, these inequalities are amplified. Many of the jobs that cannot be done from home are held by those with less access to medical resources and lower income.

Methods. We used data related to socio-economic status (SES) and the prevalence of COVID-19 in New York City to examine the relationship between SES and disease burden. We complete a clustering analysis to identify at-risk communities based on income and other demographic factors, a LASSO regression to triangulate the primary SES factors that are correlated with disease burden, and a random forest classification to predict the disease burden after accounting for these SES.

Results. We analyzed data for 117 NYC zip codes. Our clustering analysis showed a distinct bimodal distribution and a highest silhouette score for two distinct SES clusters. The LASSO regression selected a subset of SES-defining coefficients, many of which were also seen as high indicators in the random forest. The LASSO regression had a misclassification rate under 0.20 and the random forest had an Out of Bag score of 0.831.

Conclusion. There is strong evidence that socio-economic status can largely classify a New York City community's disease burden for COVID-19. This analysis would be useful in informing resource allocation to heavily impacted communities during this unprecedented pandemic.

Data and Definitions

Zip Code Granularity. We were motivated to use zip code data because it is a finer level of granularity than county, city, and state-reported data. It allows us to link COVID prevalence to SES, which varies widely within a given county, and guards against many confounders that could affect other less granular analysis. However, this decision did come with its limitations; namely, the only available data were cumulative confirmed positive cases for residents of a given zip code for specific snapshots in time. We scraped the GitHub history to obtain daily incidence rates. While certain points in time are used throughout this report, we found that some dates were clearly inaccurately reported (e.g., spikes that are not mirrored in the aggregate data), some dates were missing, and some dates were reported on a lag (e.g., data for April 6, 2020 appears to have been reported on April 9, 2020).

Disease Burden. We have assessed disease burden as the severity of the pandemic in a given community (i.e., NYC zip code) when resources were at their lowest. April 6, 2020 was the peak of the pandemic in New York City with respect to confirmed cases, hospitalizations, and deaths.[2] The peak came 17 days after Governor Cuomo issued his

[1] https://www.ncbi.nlm.nih.gov/books/NBK425844/
[2] https://www1.nyc.gov/site/doh/covid/covid-19-data.page

stay-at-home order[3] and 5 days after Mount Sinai Health System set up a field hospital in Central Park to handle the overflow of patients.[4] At this point in time, New York City's health system was experiencing its heaviest load, and access to care was at its lowest. However, it appears that the data was reported on a lag, so we define **disease burden** as the cumulative number of confirmed positive cases per 10,000 residents that have been reported for residents of a given zip code in New York City as of April 9, 2020 (see Figure 3). Because the intention of this project is to identify communities at risk, we dichotomize this value into communities that are "High" vs "Low" burdened.[5] This data was sourced from COVID-19 data published by NYCHealth for 117 NYC zip codes.[6]

It's important to acknowledge that all data related to COVID-19 confirmed cases and deaths is currently being under-reported. Estimates of true confirmed cases and deaths suggest that the true figures may be 9 to 27 times greater than the counts that researchers are currently relying on for analysis.[7] However, confirmed positive cases is the only data point currently available at the zip code level. This analysis could be re-executed when the pandemic has run its course and we could ascertain more specific disease burden metrics; however, it is unlikely that the overall insights from this analysis will change.

Socio-Economic Status. Socio-economic status (SES) is defined statically as a combination of demographic, income, and health accessibility data. This data was sourced from the US Census data[8] and Health Cost Report Information System (HCRIS) data[9]. Examples of SES factors include: income, percentage of income derived from Social Security, percentage of population who had bachelor degrees, and many others. HCRIS data were used to control for the availability of medical resources in each zip code. In total, 78 covariates were included in our models to predict COVID positive rates per capita in each zip code.

Exploratory Trends and Observations

COVID Incidence

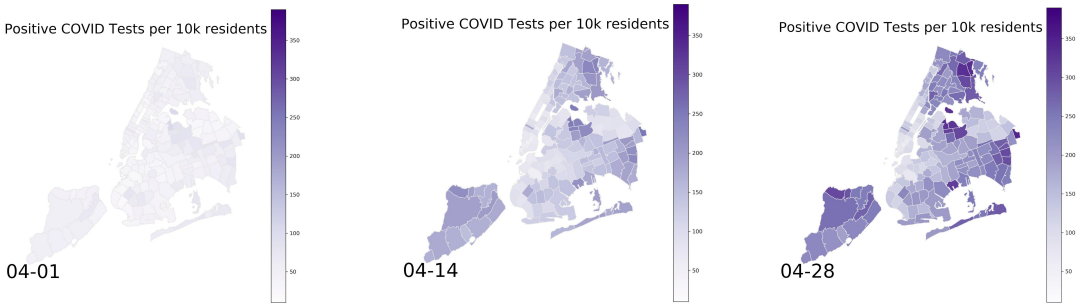
We began by looking at the geographical trend of positive COVID-19 cases over time. To do this, we generated a gif that shows the population-adjusted cumulative positive COVID-19 cases for each date that we had available data (4/1/2020 through 4/30/2020). The full gif, showing every day sequentially, is available here.

A small subset of the images that make up the gif is shown in Figure 1. A qualitative trend is immediately evident: in Manhattan and northern Brooklyn, more affluent areas of the city, incidence of COVID-19 is almost negligible at 4/28 compared to other areas. The Bronx, in northern NYC above Manhattan, and Queens, towards the east, seem to be hit particularly

[3] https://www.cnn.com/2020/03/20/politics/new-york-workforce-stay-home/index.html
[4] https://www.nytimes.com/2020/04/15/nyregion/coronavirus-central-park-hospital-tent.html
[5] The threshold for "High" vs "Low" disease burden is 85 cumulative cases in a given zip code as of April 9th, 2020. This threshold was drawn by inspecting the data and identifying the natural split. (Figure 3)
[6] https://github.com/nychealth/coronavirus-data
[7] https://www.medrxiv.org/content/10.1101/2020.04.18.20070821v1
[8] https://data.census.gov/cedsci/
[9] https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/Cost-Reports

hard. Both of these areas are home to large immigrant and working class populations; Queens is referred to as the “melting pot” of the five boroughs.

Figure 1: COVID-19 Cases in NYC Over Time



Socio-Economic Status Correlations

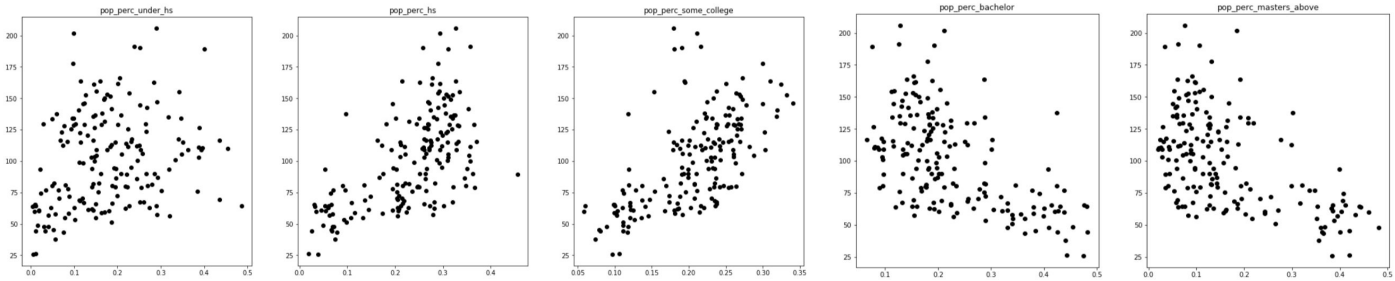
We continued our exploratory data analysis by examining simple scatterplots of the relationship between positive COVID tests per 10,000 residents and socio-economic factors from the census data. Here we note some of our most interesting findings.

Figure 2: Scatterplots: Disease Burden vs Education

From left to right y-axes: Percent of residents with less than high school education; percentages with only a high school education, percentage with some college education; percentage with a Bachelor’s; percentage with a Master’s or above

y-axis: Positive COVID-19 tests per 10,000 residents

Each observation is one zip code.



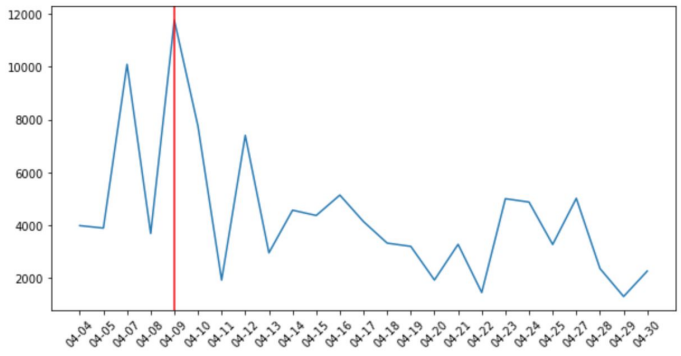
In Figure 2, we see that there is a strong positive correlation with disease burden when considering the percent of residents who have completed less than a high school degree, only a high school degree, or some college; and a negative correlation when looking at the percent of residents who have completed a Bachelor’s degree and who have completed a Master’s degree. There is also a distinct negative correlation between income and disease burden (see Appendix A). This and many other results from this scatterplot analysis (which

can all be found in "ScatterAndLine.ipynb" file) point to a positive correlation between disease burden and lower socio-economic status. The results seen for education are just one example of the well known-phenomenon of the social determinants of health that we can easily see with this data.[10]

Time Trends

We also explored time trends to inform our analyses. We generated these time-specific data by subtracting each previous date's cumulative count for discrete incidence, using the scraped GitHub history. In Figure 3, we can see the peak on April 9 for positive COVID tests, affirming our disease burden definition as defined in the introduction.

Figure 3: Time Trend of Positive COVID-19 Tests per 10,000 Residents - NYC [11]



Identifying At-Risk Communities via Clustering Analysis

We first used K-Means Clustering to identify SES clusters in NYC. The purpose of this analysis is to determine if we can discern which communities will bear the brunt of the disease burden using SES alone. The latent variable is the SES cluster, which we hypothesize will drive heightened disease burden (due to lack of resources, systemic racism/classism, health disparities, etc.).

We started by normalizing and standardizing all of our predictors. We then performed principal components analysis (PCA) to account for 90% of the variability in the data. Finally, when fitting the clusterer, we fine-tuned the number of clusters.

We decided to perform dimension reduction via PCA prior to clustering to account for 90% of the variability in our data, thereby excluding outliers that may skew our clustering results. PCA is particularly effective because many of our covariates are highly collinear (e.g., income and education). We observe in Figure 4 that the first 31 principal components cover 90% of the variance in our dataset (compared to 78 total covariates).

[10] https://www.rwjf.org/en/our-focus-areas/topics/social-determinants-of-health.html

[11] These figures do not include April 26, 2020 as this date was clearly an unreasonable outlier based on our analysis. The cumulative count of total and positive tests spiked on this date and then plummeted the next day. With the date removed, the cumulative trend is sensible. Additionally, some dates (4/06 and 4/02) were unavailable.

We compared the clustering results with and without PCA. Clustering with PCA outperformed clustering without PCA with respect to the silhouette score (see Appendix B). Because of the increased silhouette score, the benefits of a more parsimonious and lower dimensional model, and the additional robustness to outliers, we decided to use PCA prior to clustering.

We fine-tuned the appropriate number of clusters using the silhouette score. Figure 5 shows the highest silhouette score is associated with n = 2 clusters. This supports our belief that there are high and low risk communities based on SES, as we discussed previously.

Figure 4: Proportion of Variance Explained by PCA

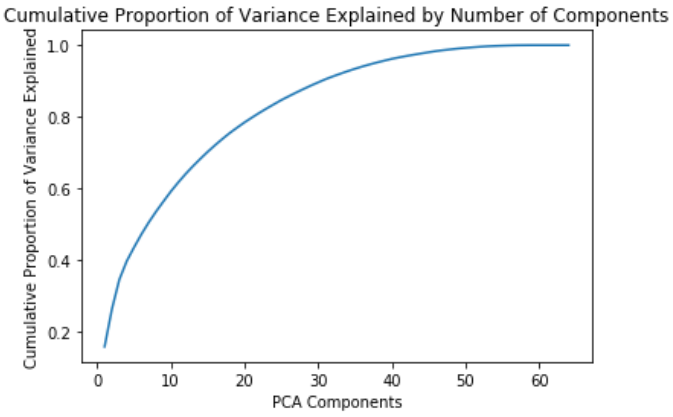
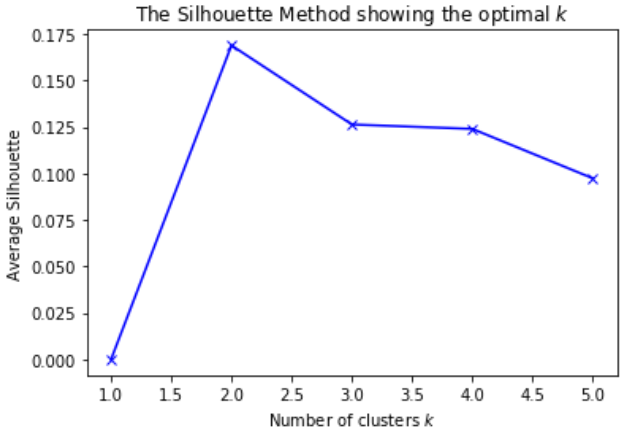
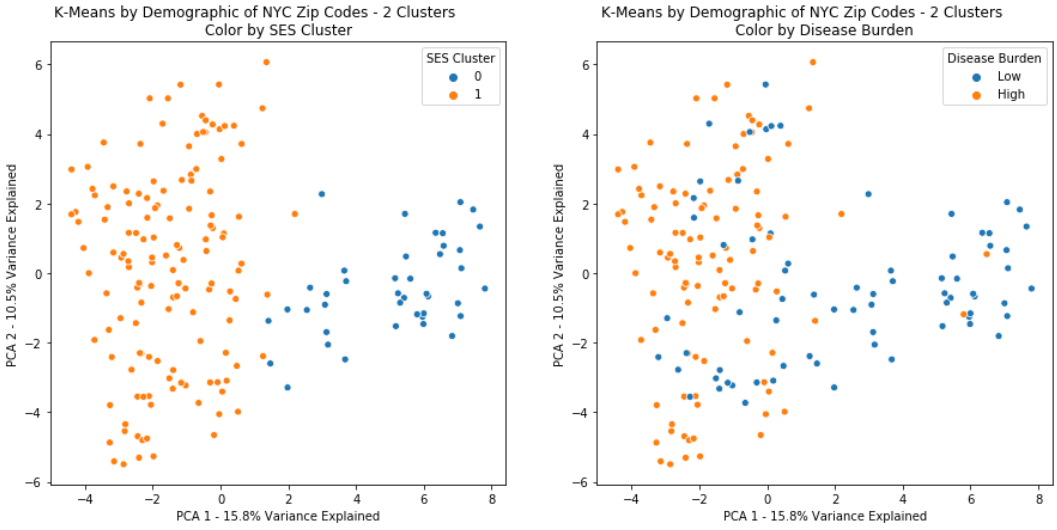


Figure 5: Silhouette Method to Select Number of Clusters



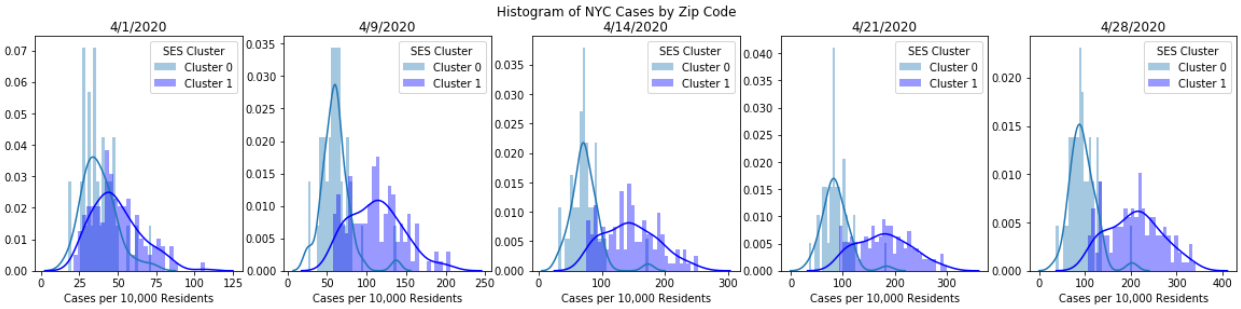
The results of our clustering analysis strongly support the idea that NYC's neighborhoods are organized into different SES clusters that correlate with disease burden. In Figure 6, we visually inspect our clusters, and compare them to the disease burden categories. It is immediately obvious how similar the SES clusters align to disease burden.

Figure 6: Results of K-Means Clustering Compared to Disease Burden



Furthermore, it's interesting to see the way that the disease burden shifts for SES clusters over time. In Figure 7, we see the histogram of disease burden over time becoming increasingly bimodal over time, and the individual distributions are clearly being drawn from the respective SES cluster. On 4/1, for example, there is not as prominent a difference in disease burden for the SES clusters, but by 4/28, we see a distinct bimodal distribution with the low-risk SES cluster centered at less than 100 cases per 10,000 residents, and the high-risk SES cluster centered at over 200 cases per 10,000 residents. These results reinforced our decision to dichotomize disease burden for the LASSO and Random Forest analyses.

Figure 7: Disease Burden by SES Cluster over Time



A few summary characteristics of each SES cluster are found in Table 1. We can clearly see the characteristics of the clusters -- for example, household income for the high disease burden cluster is double that of the low disease burden cluster, and percent with a master's degree is over three times higher.

Table 1: Summary Characteristics by SES Cluster

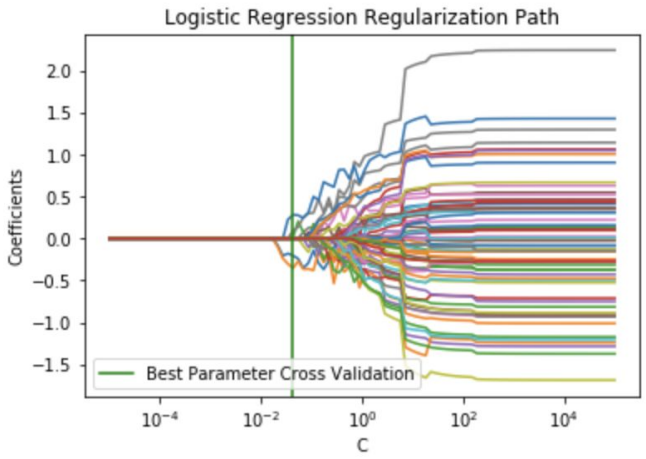
	High SES / Low Disease Burden	Low SES / High Disease Burden
Median Household Income	$ 109,247	$ 52,985
Population	34,730	51,178
Percent White	71%	40%
Percent Black	6%	26%
Percent with Less Than High School Education	6%	21%
Percent with Bachelor's Degree	39%	18%
Percent with Master's Degree or Higher	35%	11%
Percent US Citizens	71%	59%
Percent Transporting to Work in Car	7%	36%
Percent on Supplemental Security Income or Food Stamps	13%	35%
Hospital Beds Available	175	77

Identifying Important SES via Lasso Regression

Defining our SES clusters incorporated 78 total covariates. For the purposes of interpretability, we wanted to understand which of these covariates are primarily driving the relationship. To that end, we constructed a LASSO logistic regression model to discern the features that are most important when understanding the relationship between SES and COVID-19 disease burden.

We started by normalizing and standardizing all of our predictors. Next, we investigated the LASSO regularization path (Figure 8). Visually, it looked like most of our predictors were driven to zero between a log(α) of -2 and 0. We conducted a GridSearch for the optimal value of 0.043 and ran LASSO against our training data (see Appendix C for cross-validation results). Our final model had a misclassification rate of 19.8% in our test data and resulted in 4 variables: (1) percent of population transporting to work on a bike, (2) percent of population with a bachelor's degree, (3) percent of population with some college education, and (4) percent of population with high school education. When we relaxed our hyper-parameter slightly, we found that the percentage of the population that works at home and the percentage of Asians in the community were also significant predictors. See Appendix C for the logistic regression coefficients.

Figure 8: LASSO Regularization Path

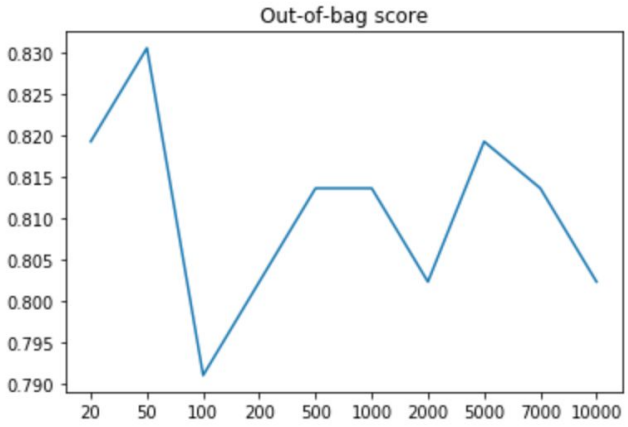


Predicting Burden via Random Forest Regression

We constructed a Random Forest Classifier using the standardized 78 SES predictors and the dichotomized outcome of low and high disease burden communities defined in our earlier analyses. The goal of the random forest classifier is to predict whether a given zip code will fall into high or low risk of disease burden and identify the SES predictors that provide the most insight into the prediction. Given that we are working with 78 features on 177 zip codes, we chose to use a Random Forest model as this method tends to perform well with small datasets.

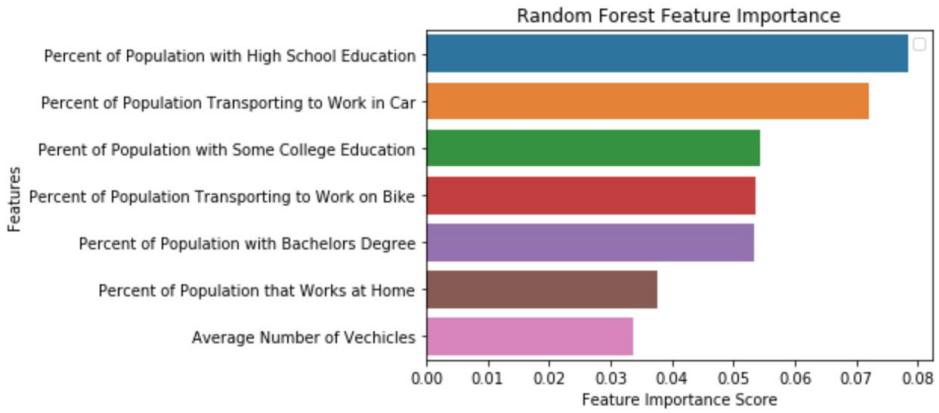
After standardizing the features in our dataset, we use the out-of-bag (OOB) score as the metric to tune the hyperparameter for the number of trees used in the random forest. Using a grid search approach, we found that using 50 trees to grow the random forest maximizes the OOB score (Figure 9). Our final random forest classifier using 50 trees and default values for all other parameters achieves an OOB score of 0.831.

Fig. 9: Hyperparameter Tuning Plot for 'B' Parameter



The most important predictors in the random forest classifier are the percent of population with a high school degree, percent of population that uses a car to transport to work, and percent of population with some college education (Figure 10). Overall, an OOB score of 0.831 is a relatively accurate rate of predicting disease burden in a given zip code considering that our model is only utilizing socio-economic factors. We see this reflected in the most important features, as subpopulations with lower education status tend to be disproportionately burdened by the effects of COVID-19.

Figure 10: Feature Importance in Random Forest



Conclusions and Recommendations

The goal of our analysis was to examine the relationship between socio-economic status and disease burden of COVID-19 by zip code in New York City. We accomplish this task

through three independent approaches: K-Means clustering, LASSO logistic regression, and Random Forest Classification. All three approaches agree with the overall narrative that areas with lower SES are associated with higher disease burden.

The LASSO and Random Forest analyses both reveal important insights about the most significant SES when predicting disease burden: education status and means of commuting to work. Education status is widely held as one of the most important social determinants of health. Both the LASSO Regression and Random Forest obtained accuracy rates around 80%-85%. This is relatively high accuracy considering that these models were built solely on socio-economic factors, as opposed to environmental factors, social network information, or clinical variables.

A very important note about the features identified as highly predictive is that many of the covariates included in our analyses are highly collinear. This is expected in variables all centering around the socio-economic status of a community, and is directly accounted for in the clustering analysis with PCA. However, due to this collinearity, the results of LASSO and Random Forest feature importance may be unstable. For example, we found that the means of commuting to work is highly correlated to income (see "ScatterAndLine.ipynb"). We could correct for this with PCA as we did in the clustering analysis, but we then lose interpretability.

We also must be careful not to slip into the ecological fallacy by generalizing these conclusions to the patient level, as our results can only be generalized to the zip code level. These results would be much more powerful if we could conduct this analysis at the patient level.

A major limitation of this study was small sample size. Small sample size was a trade-off in order to obtain the detail of zip-code level analyses, as we believe an SES analysis at a broader level would misrepresent the heterogeneity of each community. Once more data is collected on COVID-19, we would be interested in performing these analyses on communities across the United States to increase the scope of the study while retaining the granular details at the zip code level.

Our analysis sheds light on the socio-economic factors that largely define New York City's disease burden for COVID-19. Moving forward, we suggest that the government allocates more disease protective resources to the communities in the high risk cluster. Ideally, these resources would be allocated prior to a major disease outbreak as preventative measures, but further allocation of resources after the initial outbreak may still help flatten the curve in the communities that are disproportionately affected by COVID-19. Overall, this framework may be useful in protecting high risk communities in this unprecedented pandemic and inform policy for future outbreaks of disease.

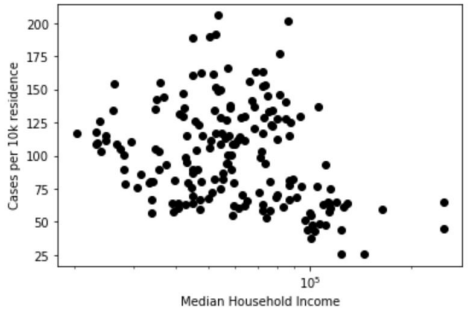
Appendix

A. Exploratory Trends and Observations

y-axes: Log-scaled median household income

x-axis: Positive COVID tests per 10,000 residents

Each observation is one zip code



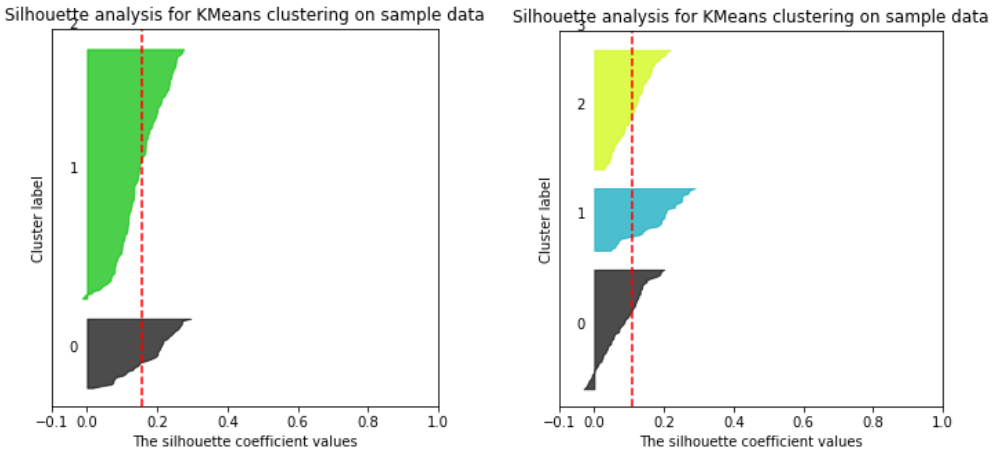
See “ScatterAndLine.ipynb” for a complete collection of our exploratory data analysis.

B. Identifying At-Risk Communities via Clustering Analysis

Silhouette Scores without PCA:

For n_clusters = 2, the average silhouette_score is 0.15698431663200013.

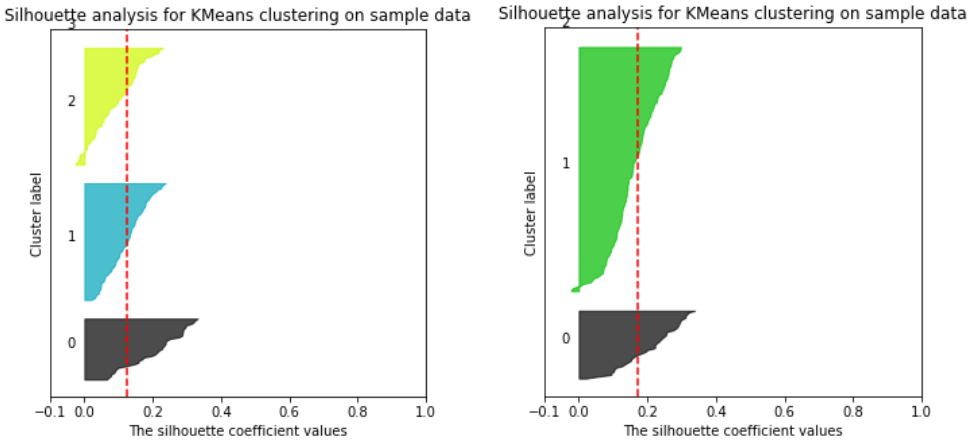
For n_clusters = 3, the average silhouette_score is 0.11024537091017984.



Silhouette Scores with PCA:

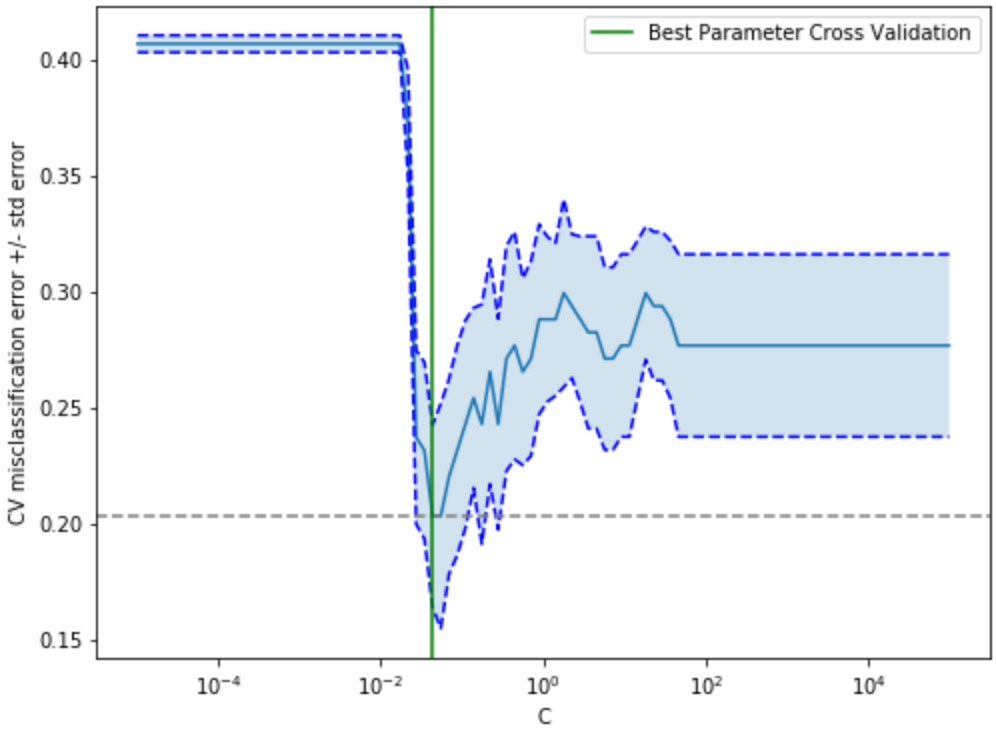
For n_clusters = 2, the average silhouette_score is 0.1724920427649748.

For n_clusters = 3, the average silhouette_score is 0.1262821588175884.



C. Identifying Important SES via Lasso Regression

Cross-validation chart to identify optimal C:



Final variable coefficients for non-zero variables

```
Percent of Population Transporting to Work on Bike    0.058518
Percent of Population with Bachelors Degree           0.033279
Perent of Population with Some College Education     -0.199920
Percent of Population with High School Education     -0.348719
dtype: float64
```