Covid-19: Health Disparities and Patient Symptoms

Izzie Lau, Kendall Kikkawa, Sina Ghandian, Trisha Sanghal

Abstract

Our primary objective this weekend was to identify and understand some of the factors that have contributed to disparities in Covid-19 outcomes between different segments of the American population, and to analyze the progression of these disparities over time (under Covid-19). Our analysis focused on the following key factors: primary language, race, education, economic status, and disability status.

In order to augment our findings, we attempted to build a predictive model using these key features that would provide us with Covid-19 projections by county. Our process and limitations are described in more detail in this paper.

We then looked into the symptoms data and tried to improve upon existing models to produce a more accurate Covid-19 predictor (this part of our project is unrelated to health disparities, and was undertaken just for fun!).

Key Trends

Background Information and Assumptions

Our first step towards our goal was to choose datasets with enough granularity to make inferences about segments of our population, but also with enough breadth to generalize to a population as large as the United States. We chose Dataset 3 from Vertical 2 (“ACS Subset County Level Economic Data for COVID-19 - Demographic and Economic Indicators”) and Dataset 4 from Vertical 2 (“Abridged County Level Data”) to suit the focus of our research.

Dataset 3 contains county-level data on general population distributions filtered by demographic and economic indicators such as race, age, education, and employment. It was collected as part of the American Community Survey (ACS), a nationwide survey that collects information about the United States’ social, economic, housing, and demographic characteristics.

Dataset 4 contains county-level data on Covid-19 infection, mortality, and case rates, as well as statistics about each county’s access to healthcare, Social Vulnerability Index (SVI), and intervention policies.[1]

We then merged these two datasets by county and state. It is important to note that this merged dataset included information on only 133 counties (and there are approximately 3000 counties in the United States). As such, correlations drawn here may or or may not be representative of U.S. counties in general. We must also note that we rounded our correlation-coefficient values in this paper to four decimals for the sake of readability, without loss of generality. We have also applied the log function to all of the y-axes in our visualization graphs for the sake of reducing the effect of outliers on our results (with the exception of our visualizations depicting changes in disparities over time).

[1] Altieri, N., Barter, R., Duncan, J., Dwivedi, R., Kumbier, K., Li, X., Netzorg, R., Park, B., Singh, C., Tan, Y., & others (2020). Curating a COVID-19 data repository and forecasting county-level death counts in the United States. arXiv preprint arXiv:2005.07882.

Findings

1. Primary Language

The primary language of each county of this dataset falls into one of three categories: "Language Other than English", "Limited English", and "Only English."

To understand the relationship between the dominant language of a county and that county's corresponding Covid-19 rates, we first plotted the proportions of people in each category against the log of the proportion of people who tested positive for Covid-19 in each county (totalled over the observed time frame of 1/22/2020 to 10/15/2020). We found a positive correlation between "Language Other than English" and proportion of people with Covid-19 (0.3541), a positive correlation between "Limited English" and proportion of people with Covid-19 (0.3973), and a negative correlation between "Only English" and proportion of people with Covid-19 (-0.3541). This implies a potential negative relationship between a dominant county-wide usage of English and Covid-19 cases.

We then looked into the progression of this relationship over time. To do this, we first grouped the counties into four quartiles based on one of the language categories (e.g. "Language Other than English"). Then, for each day that Covid-19 proportions were observed, we calculated the average proportion for all of the counties in each quartile. Next, we plotted the average proportions for each quantile over time. This process was repeated for the other two language categories as well. These were our findings:

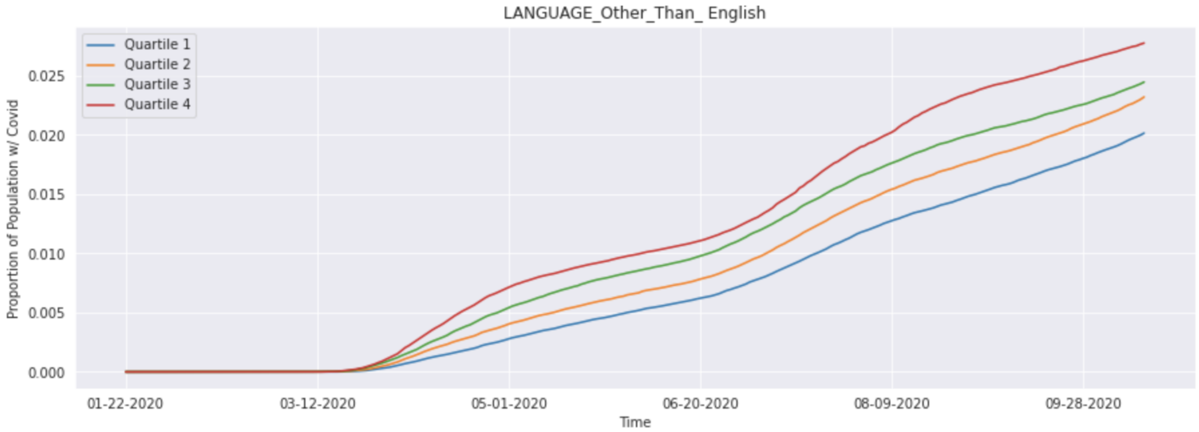

Figure 1.1
Proportion of Covid-19 per quantile of population whose primary language is a language other than English from January to October (2020).

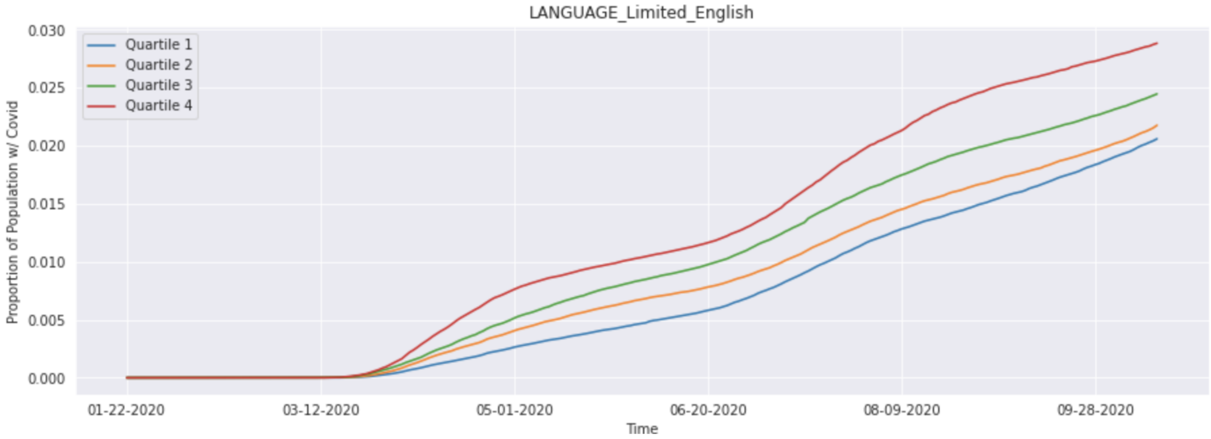

Figure 1.2
Proportion of Covid-19 per quantile of population whose primary language is limited English from January to October (2020).

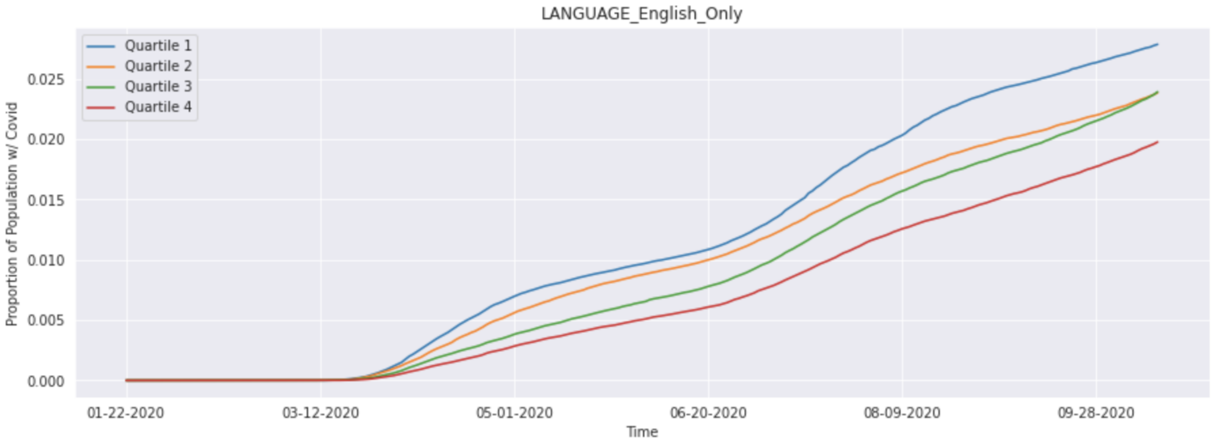

Figure 1.3

Proportion of Covid-19 per quantile of population whose primary language is limited English from January to October (2020).

These plots demonstrate that the proportion of people with Covid-19 grows faster on average in counties with a lower proportion of dominant English speakers. For example, in the last plot, counties in the fourth quartile of proportion of people in the "English Only" language category have the smallest relative growth in proportion of people with Covid-19.

We also explored the relationship between language barriers and access to healthcare, which we quantified by the ratios of number of people in a county to the number of doctors, number of hospitals, and number of ICU beds. Here, we found the following correlations:

	Ratio for County Population / # MDs	**Ratio for County Population / # Hospitals**	**Ratio for County Population / # ICU Beds**
Language other than English	0.1296	0.1889	0.2077
Limited english	0.0978	0.1862	0.1776
English only	-0.1296	-0.1889	-0.2077

These correlations are relatively weak, but they match a general trend that counties with higher proportions of English as their dominant language also tend to have more access to healthcare resources (there are fewer people per doctor, hospital, and ICU bed).

2. Race

We first attempted to get a basic understanding of the relationships between race and Covid-19 cases, and between race and access to healthcare. To do this, we calculated the correlation between the proportion of people in each racial category against the log of the proportion of people who tested positive for Covid-19 in each county (totalled over the observed time frame of 1/22/2020 to 10/15/2020). We also calculated the correlation between the proportion of people in each racial category against ratios indicative of access to healthcare (namely, the ratios of number of people in a county to the number of doctors, number of hospitals, and number of ICU beds). These were our findings:

Race/Ethnicity	Correlation for Proportion of People with Covid-19	Ratio for County Population / # MDs	Ratio for County Population / # Hospitals	Ratio for County Population / # ICU Beds
White	-0.2141	0.2793	-0.2413	0.0090
Hispanic/Latino	0.3970	0.3081	0.03598	0.01048
Black	0.3180	-0.2857	0.1522	-0.2273
Asian	-0.2601	-0.1857	0.1493	0.2664
Pacific Islander	-0.2373	0.1022	-0.0032	0.0370
American Indian	-0.0209	-0.01672	-0.1553	-0.1076
Other	0.3261	0.1385	0.1426	0.1313

Looking first at the proportion of people with Covid-19, we can see a relatively strong positive correlation with the proportion of the county that is "Hispanic/Latino" or "Black" and a relatively strong negative correlation with the proportion of the county that is "White". Based on these specific measures of healthcare access, there doesn't seem to be a consistent pattern in correlation between race and healthcare access.

We then looked into the progression of this relationship over time, specifically for the racial categories with the highest positive and negative correlation with Covid-19 cases (the progression visualizations for the remaining groups are also included in our code). To do this, we first grouped the counties into four quartiles based on the proportion of people in the selected racial categories. Then, for each day that Covid-19 proportions were observed, we calculated the average proportion for all of the counties in each quartile and plotted them over time. These were our findings:

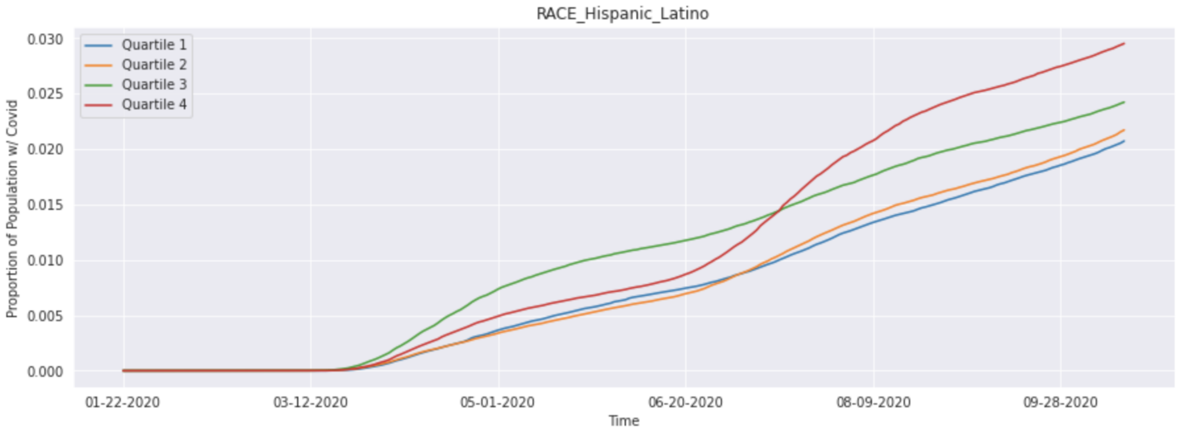

Figure 2.1
Proportion of Covid-19 per quantile of population who identifies as Hispanic/Latino from January to October (2020).

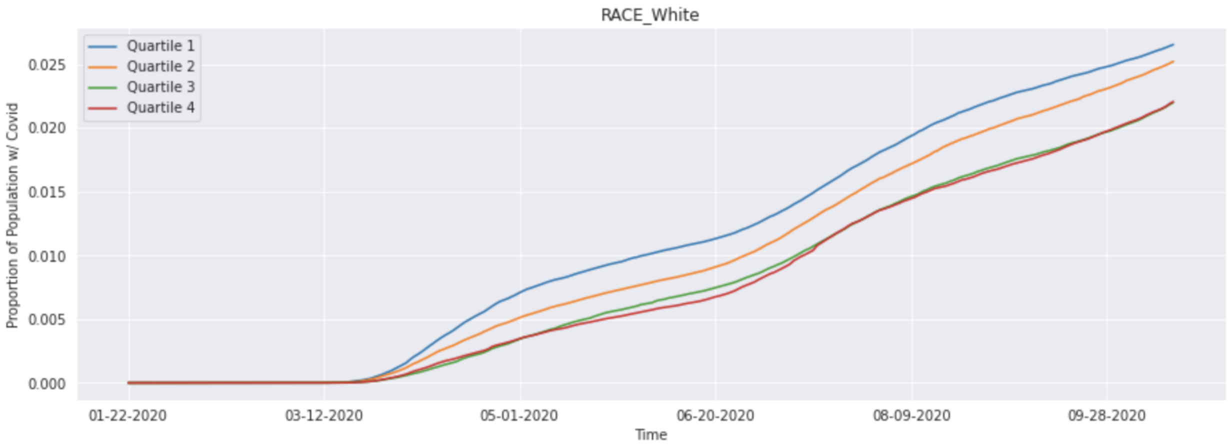

Figure 2.2
Proportion of Covid-19 per quantile of population who identifies as White from January to October (2020).

As we can see, for racial minorities groups like the “Hispanic/Latino” group, counties with a higher proportion of these minority groups had larger increases in Covid-19 cases.

3. *Education*

People over 25 years old in each county fall into one of four education categories: “Less than High School Graduate”, “High School Graduate”, “Some College”, and “Bachelor’s Degree or Higher”. To understand the relationship between education levels and Covid-19 rates, we first plotted the proportions of people in each category against the log of the proportion of people who tested positive for Covid-19 in each county (totalled over the observed time frame of 1/22/2020 to 10/15/2020).

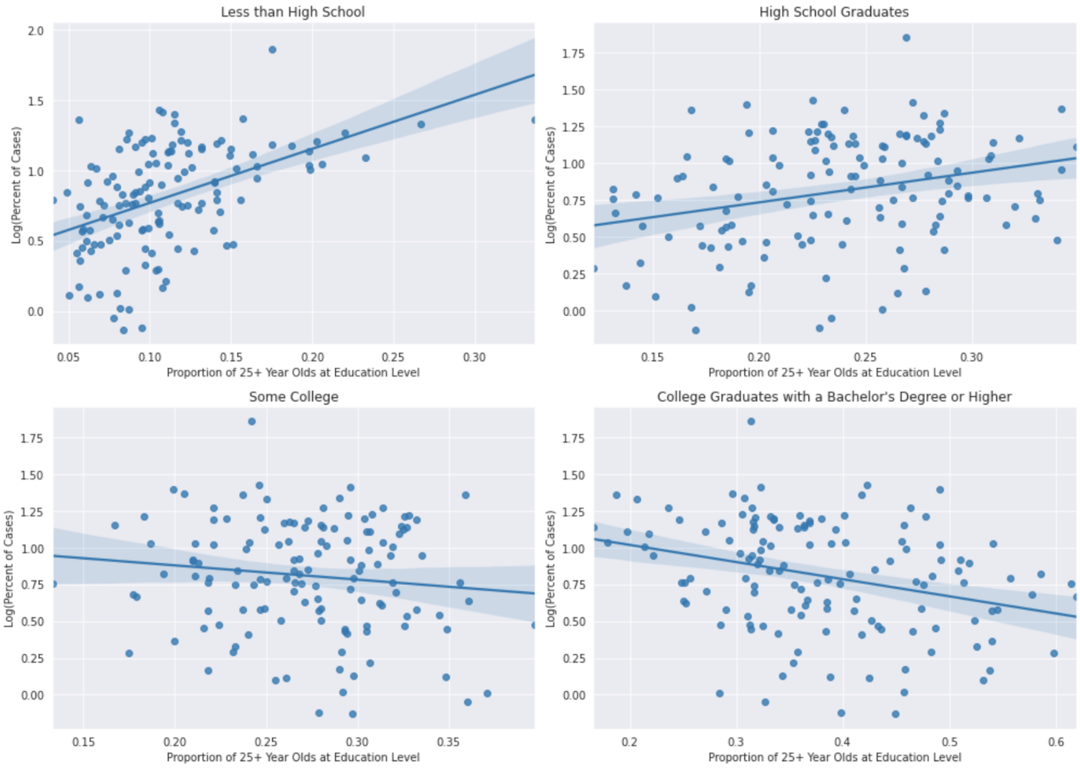

Figure 3
Proportion of 25-year-old and above adults diagnosed with Covid-19 based on highest education level.

We found a positive correlation between “Less than High School Graduate” and proportion of people with Covid-19 (0.4630), a positive correlation between “High School Graduate” and proportion of people with Covid-19 (0.2826), a negative correlation between “Some College” and proportion of people with Covid-19 (-0.1226), and a negative correlation between “Bachelor’s Degree or Higher” and proportion of people with Covid-19 (-0.3055). This implies a potential negative relationship between county education levels and Covid-19 cases.

We then looked into the progression of these relationships over time. Here are visualizations for the two categories that were most strongly correlated with cases of Covid-19:

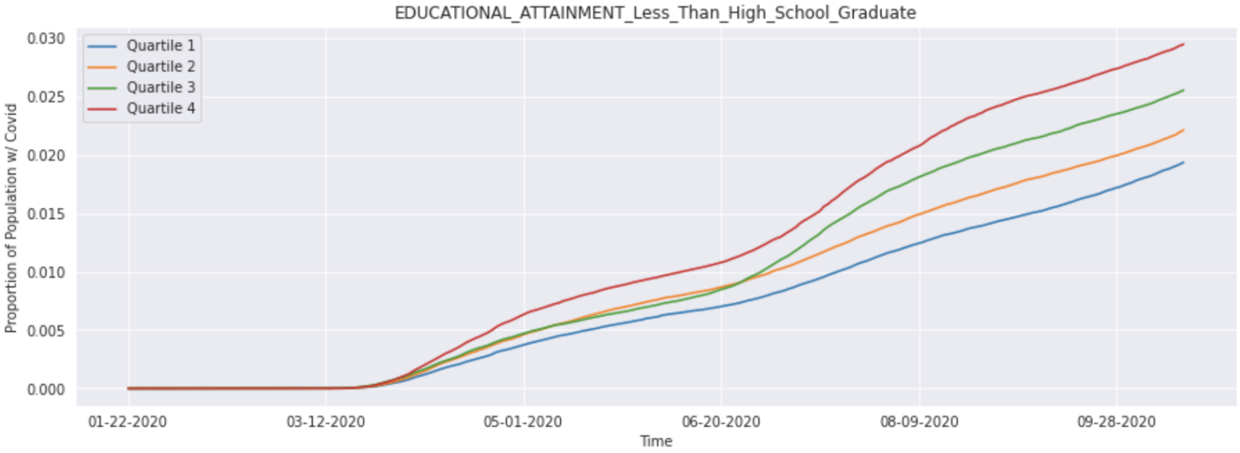

Figure 4.1
Proportion of population per quantile of population whose highest education level is less than high school graduate from January to October (2020).

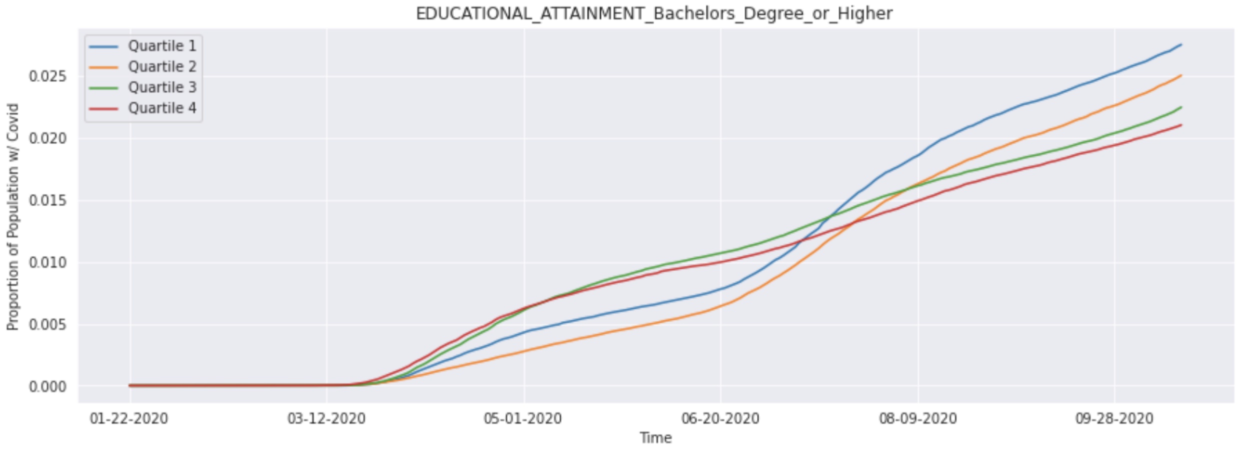

Figure 4.2
Proportion of population per quantile of population whose highest education level is a bachelor’s degree or higher from January to October (2020).

These plots demonstrate that the proportion of people with Covid-19 grows slower on average in counties with a higher average education level. For example, in the last plot, counties in the fourth quartile of proportion of people in the “Bachelor's Degree or Higher ” education category have the smallest relative growth in proportion of people with Covid-19.

There are multiple additional factors that could play a role in the existence of this trend. First, it could be that individuals with a higher education level are also more likely to have higher salaries, more access to healthcare, and less need to work in-person. Alternatively, it could be that individuals with a higher level of education are more likely to engage in safer social-distancing measures, which are hypothesized to reduce the likelihood of contracting Covid-19.

4. *Economic Status*

There were a couple factors we considered when looking into economic status: mean income (with various adjustments), poverty status, housing metrics (e.g. size of a house, number of people per room), and the proportion of people on Medicare out of those who are eligible for it. To gain a baseline understanding, we again calculated the correlation coefficients of these variables with the proportion of people who tested positive for Covid-19 in each county (totalled over the observed time frame of 1/22/2020 to 10/15/2020). These were our findings:

	Proportion of People with Covid-19
Income 12 Months (Mean Dollars Earned)	-0.2467
Income 12 Months (With Social Security)	-0.0153
Income 12 Months (Social Security Mean Dollars)	-0.3258
Income 12 Months (With Supplemental Security Income)	0.2030
Income 12 Months (Mean Supplemental Security Dollars)	-0.0931
Income 12 Months (With Cash Public Assistance)	-0.0187
Income 12 Months (With Retirement Income)	-0.3342
Income 12 Months (Mean Retirement Income Dollars)	-0.1718
Income 12 Months (With Food Stamps)	0.3411
Poverty Status (Below 100 Percent Level)	0.3796
Poverty Status (100 to 149 Percent Level)	0.4446
Poverty Status (Above 150 Percent Level)	-0.4166
Housing HH Size Owner	0.3158
Housing HH Size Renter	0.3158
Proportion Enrolled Medicare	-0.1597

There are a few interesting things to note here. Namely, there seems to be a relatively strong correlation between poverty status and the proportion of people with Covid-19, and between housing size and the proportion of people with Covid-19. We decided to leave our analysis of economic status at this since there has been much previous research about the relationship between Covid-19 disparities and economic status. However, this isn't to say there isn't anything new to learn!

5. Disability Status

Another group that could be disproportionately impacted by Covid-19 are those that are disabled. Our data included the proportion of a county which was disabled. We first explored whether there was a direct correlation between this proportion and the number of cases per capita within that county.

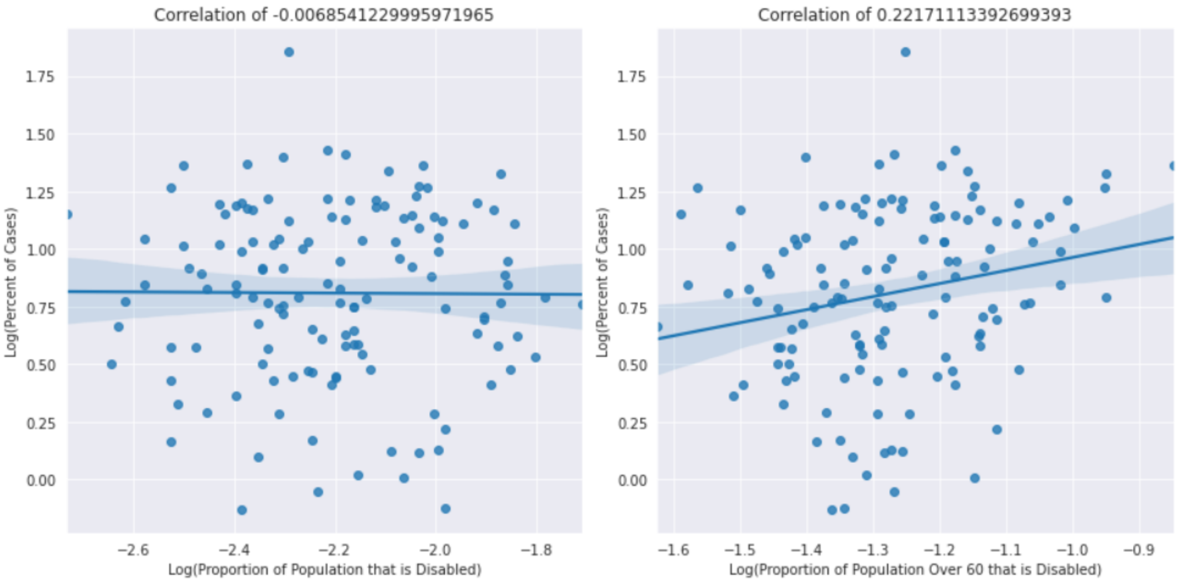

Figure 5

Correlation between the proportion of population that is disabled and the corresponding proportion of total Covid-19 cases, without and with a lower bound of age 60.

Since there was no correlation amongst the general population, we focused on the subpopulation that was over 60 to account for age as a confounding variable. Doing this, we found that a higher proportion of disabled people amongst the elderly population was mildly correlated with an increase in the number of positive cases. While it may seem that those who are disabled are not likely to be impacted in a disproportionate way, the conclusivity of our results was limited by the extent of the data which was available. For instance, one area of interest was the availability of special accommodations infrastructure and whether it might be correlated with the number of Covid-19 cases. However, there was not a reliable way to represent this infrastructure's presence utilizing solely our data. While the data possessed features that might be correlated with the presence of such infrastructure, such as the average income level of the county, these are indirect and therefore are not as valuable in proving a relationship.

Model

General Summary

We have two primary problems to model, as summarized below:

1. Classification of a county as high or low risk utilizing underlying makeup of population as features.
2. Regression of total positive percentage of cases on Oct.15 per county, using the same features.

In order to better understand the disproportionate effects of COVID-19 on different subpopulations, we attempted to construct a model that would have input features which summarized the counties in the same way done within our data. The model would then predict if some novel county would be at a high or low risk of developing a relatively large number of positive cases.

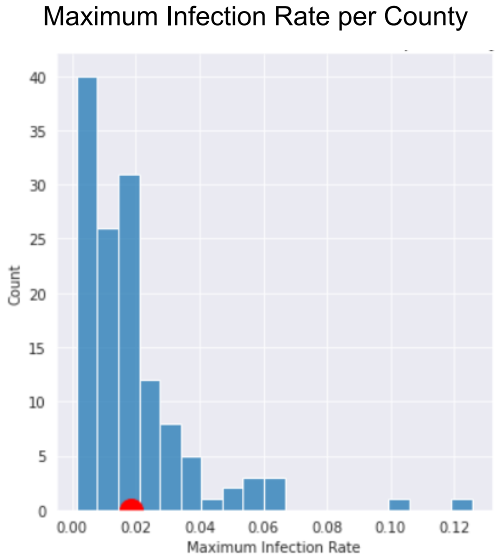

Figure 6
Distribution of maximum infection rate per county.

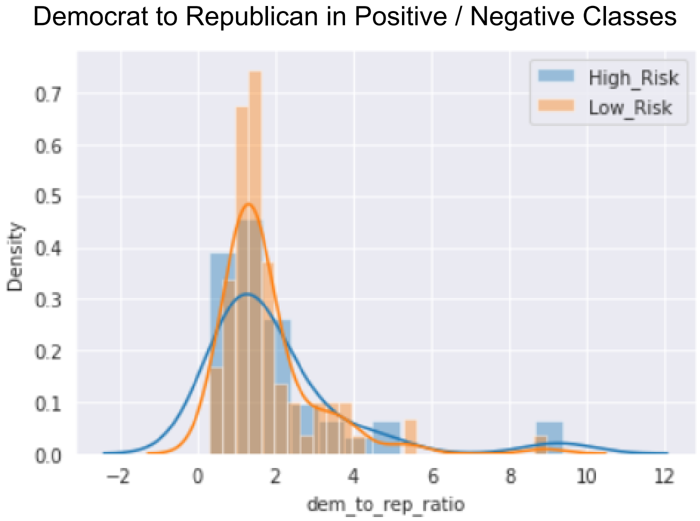

Figure 7
Distribution of Democrat to Republican in positive/negative classes.

The first concern was creating some way to classify the counties within our dataset as “high-risk” or “low-risk”. We hypothesized that a high-risk county would be one in which the maximum proportion of positive cases exceeded the median of the distribution of maxes across all counties.We then began to construct a feature matrix, utilizing all of the data that summarized a county’s constituency and was agnostic to case counts. Since we are intent on exploring the disparity in impact to different groups, we chose not to include any information such as the date at which the county had its first case, which would generally be more indicative of worse outcomes. Categorical data was one-hot encoded, the data was standardized, and split into training and testing sets at 80/20%.

A few models were trained in attempts to classify these groups. The first was a 5-fold cross-validated logistic regression, and the second was an xgboost gradient boosted tree classifier. Both models performed very poorly on our data, hardly achieving consistent AUCs of 0.5. We hypothesize a few reasons for this outcome. Firstly, the total number of sample points with respect to county is very small, (*n=133*), therefore our training set is too small to come to any generalizable conclusions, and our test set is similarly too small to test our hypothesis, creating large variance in the cross-validation scores (AUC standard deviation ~0.06). In addition, there’s a large amount of noise and collinearity between features, which ultimately does not make the data descriptive enough to correctly classify our points of interest. We validated this hypothesis by conducting PCA to find that our 161 input features could be combined to 23 principal components to explain 95% of the sample variance. This figure depicts one of the most significant features to distinguish classes, although the test and training sets are so small that this outcome was likely due to random chance. In addition, the assumptions we made to construct our binary threshold may have also been incorrect, as there are a number of confounding variables that make it difficult to discretize “higher risk”.

A second modeling approach we took was one in which we could use the same feature set to predict the number of total positive cases per capita (scaled to a percentage) on October 15th, the last date in our dataset. We used an xgboost regressor, and a random forest regressor to make these predictions, then evaluated the results using a standard root mean squared error. We found the XGB regressor to predict the outcome variable with an RMSE of 1.02%, which is nearly one standard deviation of the response variable’s distribution (mean of 2.37%, standard deviation of 0.896).

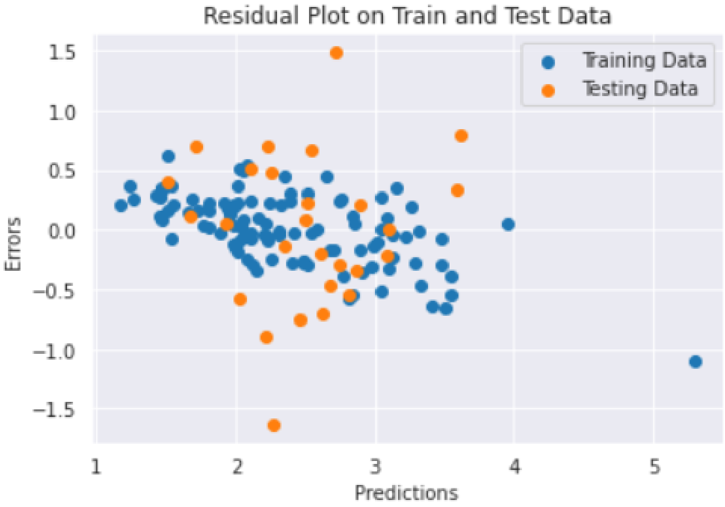

Figure 8
Residual plot for the Random Forest Regressor on training data vs. testing data.

We then fit a Random Forest Regressor and found we achieved an RMSE of 0.63%. This result is the most promising of the models that we explored in this vertical and has a number of potential implications. To validate this result further, we visually inspected the residual plot of both the training and testing data, and found relative similarity between the two, save for some outliers that likely decreased the model's performance. The high predictive accuracy of this model indicates that it is relatively possible with a description of the population distribution alone to predict how heavily impacted a county may become with COVID-19, relative to other counties. Below is the permutation feature importance plot generated by the random forest regression model trained on the data. With it, we can both validate some of our hypotheses, such as the positive correlation between low educational attainment and higher COVID-19 prevalence. But we can also understand other predictors that we did not catch at a higher level, such as the impact of diabetes on positive cases, a possible indication that those who possess diabetes and other preexisting conditions are more susceptible to this virus than others.

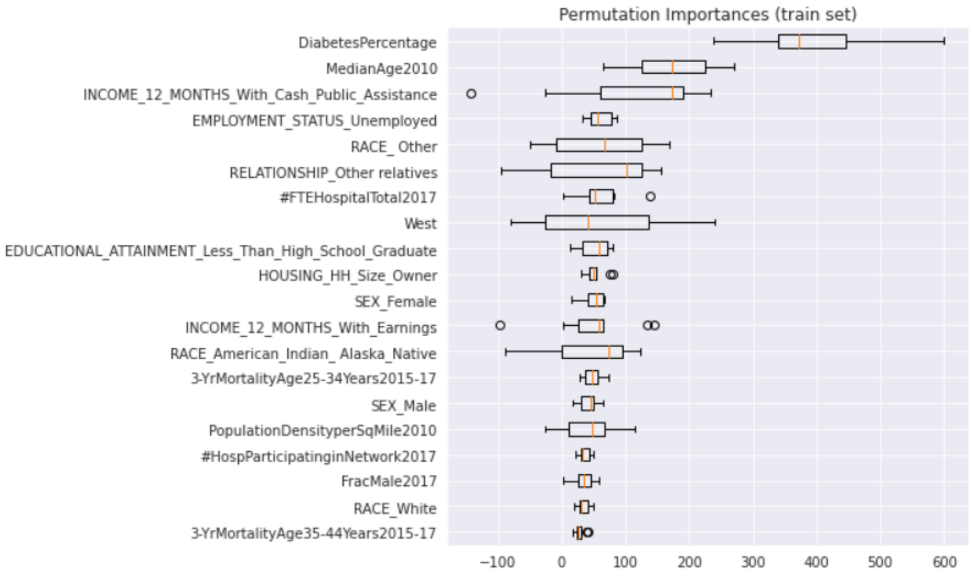

Figure 9
Top 20 Feature importance plot for the Random Forest Regressor.

Patient Symptoms

After analyzing county-level data, we analyzed the Enya.AI COVID-19 symptoms data. Through exploratory data analysis, we were able to identify several key observations that improved our understanding of the FeverIQ diagnostic tests, and the relationships between the variables in them. Through initial pairplots, we noticed that the 4 symptom vectors (cold, flu, the COVID symptom, and the COVID neurological symptom) were highly correlated with each other. Through time series analysis, we also inferred that there may be more than a simple correlation, as the four vectors seemed to behave very similarly through time, especially during the month of June.

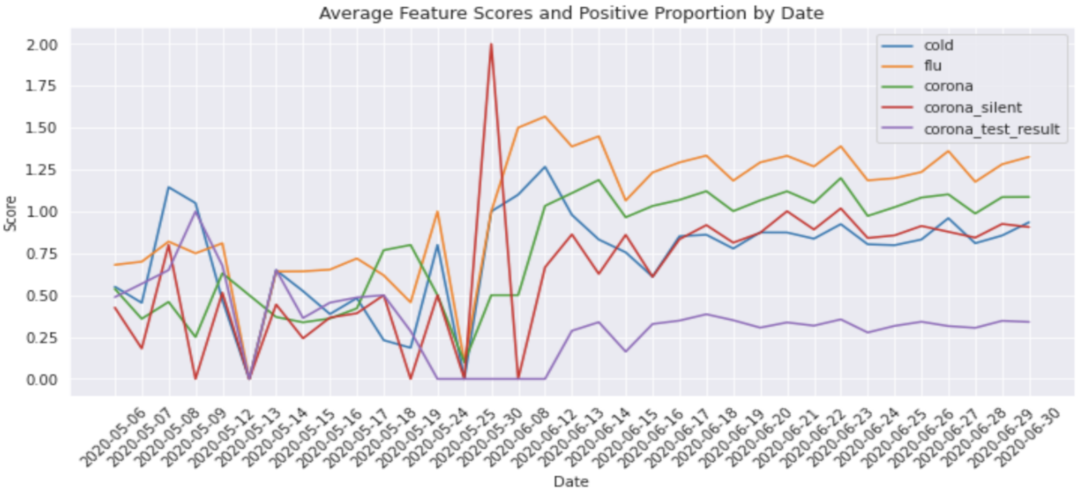

Figure 10
Average feature scores and the proportion of positive Covid-19 cases over the period of May 6, 2020 to June 30, 2020.

Based upon the observed correlations between each of the four symptom vectors and the test results, we started our analysis by constructing simple logistic regression models. We compared the ability of one symptom to serve as a predictor of a positive case. Below are the results by using each of the four vectors (corona_silent was treated as 6 binary columns).

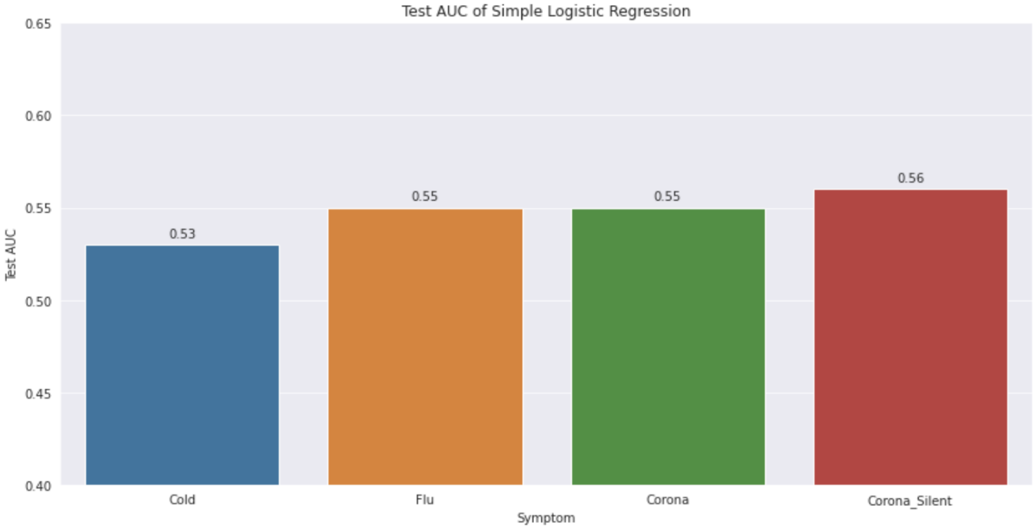

Figure 11
Summary of Simple Logistic Regression classification performance across different symptoms.

We observe that each of the symptoms do achieve better than 50%, indicating that there is some correlation between common symptoms and COVID-19. However, we sought to take our analysis deeper to understand how combinations of these features may yield more predictive power.

Through further EDA, we observed other attributes of the dataset that would help us create our more complex design matrices. First, we noticed that the corona_silent score (neurological symptom) was discrete on the range [0.6]. We also hypothesized that the country and date features could contain signals about the COVID trends in a particular country. We considered joining the FeverIQ dataset with other datasets that were given, but we didn't have the same granularity of information about all countries (only had other data on the U.S.), so we decided to engineer features from the six initial columns.

These observations inspired several feature engineering options, outlined below:

Description	**Number of Continuous Features**	**Number of Categorical (Binary) Features**	**Number of Interaction Features**	**Total Number of Features**
Maintain the 4 original symptom vectors. Only transformation is one-hot-encoding the corona neurological symptom.	3	6	0	9
In addition to the features in the previous iteration, this design matrix included an interaction term for each 2 column combination of the previous vectors, including continuous and categorical interactions (example: cold*flu, cold*corona_silent_1)	3	6	25	34
Instead of including interaction terms, this design matrix utilizes the country, month, and day of the week as categorical variables. All of these features are then one-hot-encoded	3	151	0	154

We compared the performance of these design matrices across five different models: K-Nearest-Neighbors (KNN), Logistic Regression, Support Vector Classifier (SVC), Gradient-Boosted Trees (XG-Boost), and Bagged Logistic Regression. We used a randomized train:test split of 80:20 across all models, we applied 10-Fold cross validation in each case to ensure model stability, and we used these cross-validated models to make predictions on the test set. The bar chart below outlines the performance results.

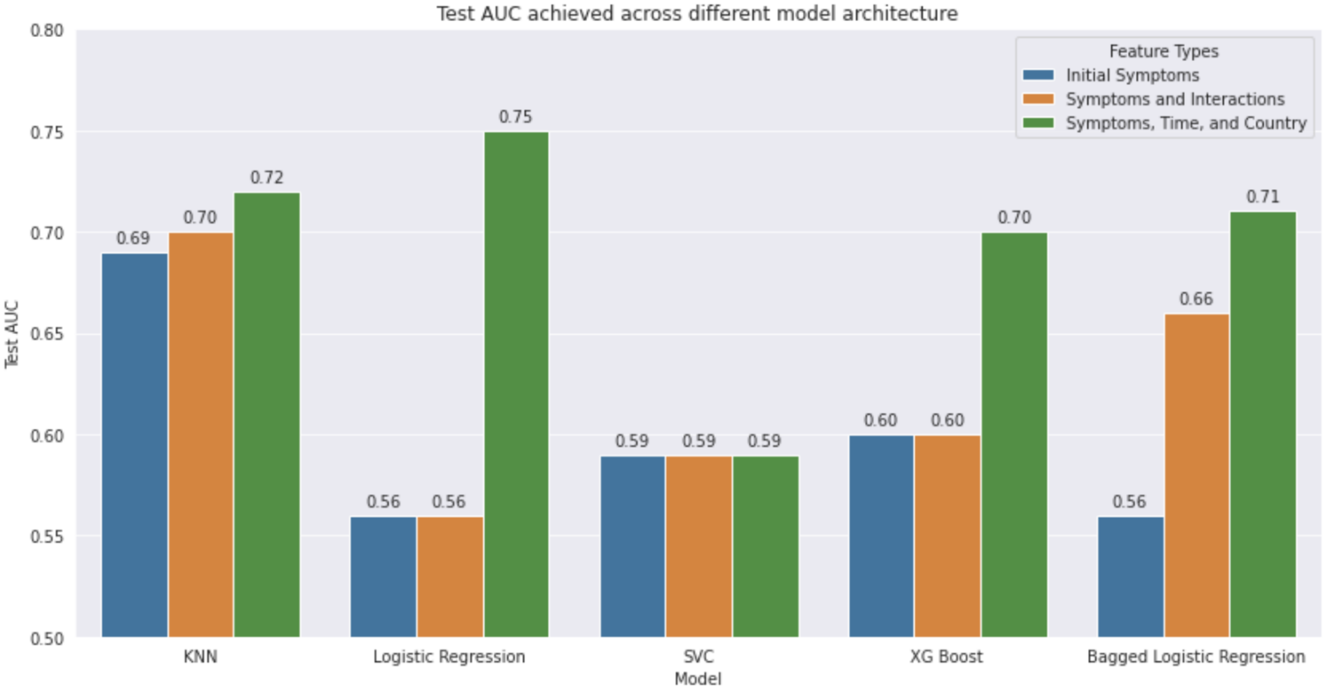

Figure 12

Summary of COVID-19 classification performance across different model types and design matrices.

While our highest performing model was a Logistic Regression model (Test AUC of 0.75) that contained features corresponding to country, month, and day of the week, we should acknowledge that this model could lead to overfitting. In particular, because there are only a select number of individuals from each country in the dataset, our model may have learned noise instead of true signal. For example, the model may have picked up on which countries had a higher proportion of COVID-19 cases in the dataset, rather than identifying the true trends that may be going on, or understanding the nuanced relationship between features.

Given the structure of the data, the *most reliable predictive model* would be the KNN model that achieved a 0.70 AUC on test data (the KNN architecture was also quite stable across different combinations of explanatory variables). This model is the most reliable test because its features are entirely composed of symptom features, but it has the highest predictive power due to the interaction terms.

We should note that an AUC of 0.70 is not near a suitable threshold for what a COVID test should be in practice. However, this simple test based on only four features displays the signal that common illnesses, like the cold and the flu, may contain. While the symptom data alone is not enough, it can be used as a screening mechanism to help flag people that are at risk of catching or spreading the virus. Hopefully, these symptoms can be measured much more rapidly and measurement mechanisms can become much more readily available.

In an ideal world, we would recommend that FeverIQ collect other data related to a patient's condition, or maybe data about the amount of time they have spent outside or around other people. All of this information could be leveraged to make more informed predictions. However, if FeverIQ were to collect more data along the same lines as this dataset, then we would recommend using a more complex model that may be able to infer more signal, such as a simple feedforward neural network. While we may sacrifice model interpretability, we believe that the post important task related to testing is to screen potentially positive and negative individuals, not to understand every single root cause. The world needs to slow the spread of COVID-19 as quickly as possible, and this tool can help accomplish that.

Conclusion

In this paper, we identified a subset of the factors that have caused disparities in Covid-19 outcomes between subgroups of the population of the United States. We conducted a thorough analysis of the dependency between the proportion of Covid-19 cases by county and five key features:

1. Primary language
2. Race
3. Education
4. Economic status
5. Disability status

We discovered that factors such as a higher proportion of non-English language as a primary language and a higher proportion of minority populations such as Black or Hispanic communities were strong indicators for that county's corresponding proportion of Covid-19 cases. In addition, we introduced time series data into our project.

Finally, we developed two models to supplement our findings: (1) a random forest regressor that predicts the "risk" of a county using a large feature set and (2) a logistic regression model that added a temporal feature to the FeverIQ data. For the latter, we ended up with an AUC-ROC score of 0.75, which was a significant improvement upon the model sited in the FeverIQ COVID-19 Symptom Tracking paper[2]. However, we acknowledge that this model may have fit to the noise of location and date rather than the symptoms themselves. Despite this, we were still able to achieve an AUC-ROC score of 0.70 with a symptom-interactive model. This classifier opens up data collection possibilities that incorporate other symptoms, patient information, or geographic COVID-19 trends, as well as new possibilities for simple and accessible coronavirus screening.

Overall, we faced a couple of challenges during our process. While designing the scope of our project, we ran into issues with managing and visualizing the large amounts of data that we were given. In addition, we had to find the balance between focusing too much on a handful of features (and not considering the wide scope of factors that contribute to the effect of Covid-19) and wanting to analyze every single feature that we were given. We tackled these issues by creating small exploratory data analysis (EDA) experiments in which we pattern-matched between the granularity of different datasets and made preliminary visualizations to see if there were any immediately noticeable correlations.

Furthermore, we ran into technical challenges as we conducted the correlation coefficient experiments between counties' features and their corresponding proportion of total Covid-19 cases. We at first had trouble drawing conclusions from our data because large outliers would skew the dataset. We accounted for this by taking the log function of the y-axis (proportion of total Covid-19 cases) to reduce the effect of outliers on the correlation coefficient.

Most importantly, we want to share what we've learned from working with this data and our proposals for steps moving forward. Through our analysis of the effect of factors such as race, economic status, and primary language, we were able to show that there is a clear statistical disparity between groups of certain demographics in regards to the proportion of Covid-19 cases. This was shown through statistically significant correlation coefficients and varying growth rates of Covid-19 between the quartiles of demographic groups.

[2] Ranjan et. al. "Fever IQ - A Privacy-Preserving COVID-19 Symptom Tracker with 3.6 Million Reports" *medRXIV*; doi: doi.org/10.1191/2020/09.23.20200006

We want to go one step further and suggest tangible solutions for ways in which we can alleviate these disparities.

1. **Counties with a larger proportion of a non-English speaking population are at higher risk to test positive for Covid-19 (r = 0.3541).**
   - Counties can take actionable measures to inform their non-English communities by translating announcements about quarantine regulations.
   - Counties can provide alternative forms of media to alert its members of safety measures (i.e. sending mail translated in the residence's primary language as denoted by the census)[3].

2. **Counties with a larger proportion of residents living in poverty status (below the 100% level) are more likely to have a higher proportion of Covid-19 cases (r = 0.3796) than that of those living significantly above the poverty status (r = -0.4166).**
   - Counties with a larger proportion of residents living below the poverty level can enact extra aid policies (i.e. free Covid-19 testing centers, emergency relief funds).

3. **Counties with a larger proportion of Black residents are more likely to have a higher proportion of Covid-19 cases (r = 0.3180).**
   - Counties can account for the racial disparity by investing into free Covid-19 testing centers and public health care in primarily-Black neighborhoods.

Overall, analyzing this dataset allowed us to find strong evidence that there is indeed a strong relationship between socioeconomic and demographic factors, and that these cause disparities in how different subgroups of America suffer from and recover from the Covid-19 pandemic. Moving forward, we urge that this data be incorporated into future policies to alleviate the strain that disadvantaged and minority groups face in the midst of the pandemic.

[3] US Census Bureau. "American Community Survey: Why We Ask About Language Spoken at Home." *US Census Bureau*, www.census.gov/acs/www/about/why-we-ask-each-question/language/.

References

1. Altieri, N., Barter, R., Duncan, J., Dwivedi, R., Kumbier, K., Li, X., Netzorg, R., Park, B., Singh, C., Tan, Y., & others (2020). Curating a COVID-19 data repository and forecasting county-level death counts in the United States. arXiv preprint arXiv:2005.07882.
2. Ranjan et. al. "Fever IQ - A Privacy-Preserving COVID-19 Symptom Tracker with 3.6 Million Reports" medRXIV; doi: doi.org/10.1191/2020/09.23.20200006
3. US Census Bureau. "American Community Survey: Why We Ask About Language Spoken at Home." US Census Bureau, www.census.gov/acs/www/about/why-we-ask-each-question/language/.