Combating America's Other Infodemic

Recommendations for a Targeted Approach to Address COVID Misconceptions, Physical Mobility and Disease Outcomes

West Coast Regional Datathon Spring 2021

Team 2

Ben Huckell, Kangyu Wang, Zhipeng Ye, Sheng Kun Zhou

February 21, 2021

1 Topic Question

Beginning in 2020, the world experienced in COVID-19 one of the most serious influenza pandemics since the Spanish flu. In the first global pandemic in this digital social media era, an unprecedented amount of digital information, regardless of validity, are widely available and perpetuated through multitudes of far-reaching platforms. As COVID-19 is also a novel situation to most of the public on an individual scale, misconceptions regarding the nature of the virus and methods of prevention can be understandably common. This phenomenon is notably prevalent in the United States, driven by diverse and intricate social, economic, political and technological contexts.

The concept of the need to combat COVID misinformation and educate against misconceptions is widely accepted, and yet efforts have largely fallen short. One notable reason could be that these efforts are general, unguided, and mostly qualitatively based. This is the topic we aim to investigate and quantitatively analyze. In this report, we leverage diverse data regarding COVID-19, public conception, misinformation, mobility and socioeconomic status to address the following questions:

1. Can we identify demographic groups based on socioeconomic status and/or geographical regions who are vulnerable to the effects of COVID-19 due to susceptibility to, or acceptance of, misinformation in addition to prevalence of misconceptions?

2. Across varying social, economical and geographical demographics, can we identify and quantify the vulnerabilities that each demographic has towards COVID-19, related misinformation or misconceptions?

Answering these questions could help provide valuable insight on how to combat and control the spread and effects of COVID-19 in the USA. Particularly, the answers to these questions will allow for targeted and localized education and information initiatives crafted to effectively reach and address individual communities and demographics suffering from COVID-19. Educational and informational resources can then be utilized more efficiently, and create more positive and impactful changes in vulnerable communities.

2 Executive Summary

To control the spread of COVID-19, in addition to traditional public health measures to restrict people's physical mobility, governments and big technology companies have also paid attention to misinformation, with the belief that false information about the disease makes people less likely to comply with government containment measures and thus increase the risk for COVID spread. We identified strong statistical relationships between people's increased physical mobility and worse outcomes of COVID, as measured by infection, hospitalization and death at the state level. Moreover, we discovered that the incorrect information problem to COVID containment turns our to be broader than misinformation from social or traditional media.

A lack of basic knowledge about the cause, spread and prevention of COVID-19 among people who are likely old, rural, without a high school diploma and less exposed to social media, turns out to be strongly associated with less reduction in non-essential mobility and worse disease outcomes. By contrast, despite being exposed to more misinformation from their frequent use of social media to obtain COVID-19 related information, people who are younger, more urban, and with higher education levels do not show a tendency to have high levels of physical movement during the pandemic, or suffer from more worse disease outcomes, even after controlling for the strictness of containment policies.

Although combating misinformation, especially on social media platform, has been the focus of governments and media and technology companies in their efforts to control the pandemic, our project should inform policy makers that the dissemination of correct information to people who are not necessarily exposed to much misinformation from the media, but nonetheless have more misconceptions about the basics of COVID-19, should also be a priority. We have identified key geographies and demographic groups for policy makers' consideration.

3 Technical Exposition

3.1 Data Pre-processing

We obtained our data from several different sources, and produced cleaned datasets to perform cross-sectional and panel regressions on. We define the time scope of this project to be from March 2020 to January 2021, and the geographical scope as all 50 states, plus the District of Columbia of the US. The process of cleaning and transforming data is described below.

3.1.1 Mobility

Google has made public datasets on movement of people available since February 15, 2020 to help combat COVID-19. For the US, both city and county level data are available. However, since other important datasets are only up to the state level, we opted to use state level data. We kept the daily structure of mobility data.

As part of our EDA, we looked at mobility for non-residential purposes (i.e. not staying at home/isolating). The subcategories under this are:

- Retail and Recreation
- Grocery and Pharmacy

- Parks
- Public Transit
- Workplaces

We then define our non-residential mobility change as the average of all these factors, recognizing that although some may be necessities (such as groceries, pharmacy, or workplaces for essential workers), these are still mobility factors that best reflect potential lack of compliance to public health measures and can contribute to COVID spread.

3.1.2 Misconception and Exposure to Misinformation

Using survey data from CovidStates, we constructed indicators for people's conceptions, as well as exposure to misinformation, about COVID-19 for each state in each round of survey.

- Misconceptions about COVID-19. It is measured as the average proportion of false statements about the prevention and treatment of COVID-19 identified as correct by respondents.
- Exposure to misinformation from social and traditional media. We took the average prevalence of false information by media source, weighted by the amount of usage of each media source for COVID-19 related information.

The dataset only provides data on misconceptions about COVID-19 from the ninth round of survey, which was conducted in August 2020. Data about exposure to misinformation is reported for every round of the survey. However, due to time constraint, we decided to aggregate all data by state, disregarding the time factor. Therefore, data about misconception and exposure to misinformation are only used in a cross-sectional framework. The dataset is complete, and from its standard representative sampling methodology, we concluded that the data is credible.

3.1.3 COVID Epidemiological Data

We obtained epidemiological data of COVID-19 from the the Covid Tracking Project. Using its API, we obtained daily data for each state of the US. We obtained the following variables from the dataset:

- New Deaths for each day ($\boldsymbol{DeathIncrease}$)
- New Hospitalizations for each day ($\boldsymbol{HopspitalizationIncrease}$)
- Current number of patients in ICU ($\boldsymbol{InICU}$)
- Current number of patients on ventilator ($\boldsymbol{OnVentilator}$)
- Test positivity rate for each day ($\boldsymbol{Pos.TestRate}$)

We then calculated deaths per thousand and hospitalization numbers per thousand. For cross-sectional analyses, we summed $\boldsymbol{DeathIncrease}$ and $\boldsymbol{HopspitalizationIncrease}$ for each to get total number of deaths and hospitalizations, and averaged $\boldsymbol{Pos.TestRate}$ through days for each state to get the mean test positivity rate for each state.

Choosing the right parameter to depict the severity of COVID-19 outbreak in each state is a tricky task. We did not use the absolute number of positive tests because people who are infected with COVID-19 but do not have symptoms may choose to not get tested. Also, in the early stages of the pandemic, getting the test was difficult and even symptomatic patients did not have access to the test.

We believe that number of deaths and hospitalizations are better, albeit still imperfect, measurements of severity of the outbreak. In the early stage of the outbreak when people with no or wild symptoms could not obtain COVID tests, hospitals prioritized severe patients for testing and treatment and thus their numbers were better reflected in the statistics. Also, deaths and hospitalizations are the actual toll that COVID-19 has on the society as a whole, and should be what policy makers aim to minimize. We also used test positivity rate as an indicator. We acknowledge that it not only depends on the number of people infected, but also the volume of testing available or required. It should be treated as a robust check against analyses on deaths and hospitalizations.

3.1.4 Demographics

Data on demographics was compiled from a variety of sources. These sources are described in the Appendix section, and include the Education Dataset, Economic Dataset, Temperature Dataset, Age Breakdown Dataset, Race Breakdown Dataset, and Urbanization Levels Dataset. By leveraging a large amount of the data available to us, we were able to classify demographics as accurately as possible. Since not all of the data was available as time series, we made the decision to complete our demographic analysis cross-sectionally.

In the process of cleaning the data, some datasets were quite clean and useable "as is", including the Temperature Dataset, and Urbanization Level Dataset.

The Education and Economic Datasets gave information at county level, so we simply extracted the information on the most recent year in the study, and aggregated or averaged (depending on the value of interest) by state. Education levels were given in the form of percentages of populations with differing levels of formal education. Since it was not clear to us which levels would be the most useful, we selected all four that were provided.

- Percentage of population with **less** than a high school diploma
- Percentage of population with **only** a high school diploma
- Percentage of population with some college or associates degree
- Percentage of population with **at least** a bachelors degree

We believe that including all of these metrics will give an accurate picture of the mean education levels within a state, since it is able to break down those in higher higher education brackets, lower education brackets, and varying levels in between these two extremes. It will allow us to classify proportions of the population that are not educated, somewhat educated, relatively educated, and very educated.

Finally, the Age Breakdown and Race Breakdown datasets give estimates of the number of people in a state, broken down by gender and race for integer values of age (up to 85, where 85+ is provided in lieu of continuing integer values). Since the ages provided were capped at 85, it was impossible to generate an accurate mean value, so we opted to calculate the median age value. For our race data, we chose to clean and process these values as percentages of the population

fitting into each category. In this sense, we are adding five new columns to our master demographic dataset, one for each race. The races included in the data are shown in the Appendix section.

One important caveat to note about the race data, is the fact that combinations of races are allowed. As such, the proportions included in our analysis may add up to greater than 100%. This should not cause any issues moving forward, but it is important to note, giving the rising amount of individuals who classify as more than one race in the United States.

3.2 Exploratory Data Analysis

3.2.1 Mobility

To begin with, we start to explore the google mobility data that we have processed. To visualize, we can take the year average of non-residential mobility change across the states. This is shown in the figure below.

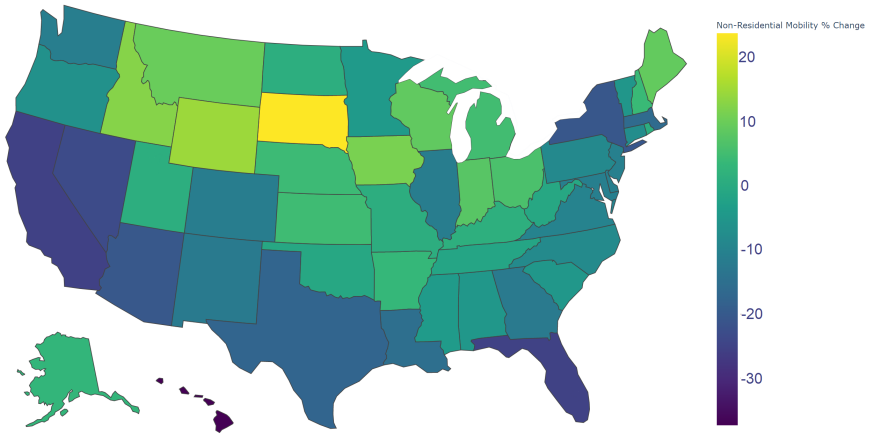


Figure 1: Year Average non-residential mobility % change

More interesting is to look at the evolution of change throughout the year, which we can visualize using the same data. We can also visualize the daily state wide data as a viewable time series. Please give the plot about 30 seconds to load, and any adjustments on the scroll bar will fix the auto-scaling of the colorbar.

While performing visual EDA on this time series, some drastic fluctuations were observed in some states, while others had relatively low change. To better explore and represent this, we plot the non-residential mobility rates of all states using a 15-day rolling average for smoothing:

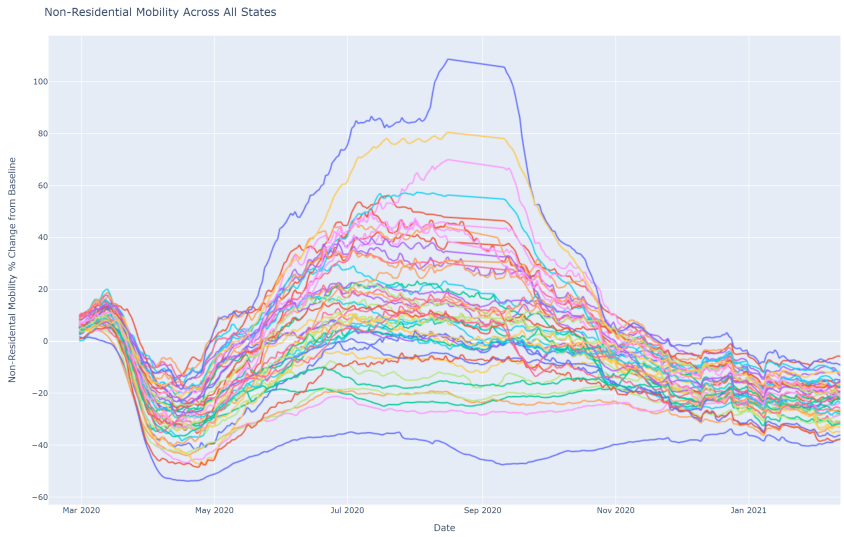


Figure 2: 15-day Rolling Average of Non-Residential Mobility % Change Across All States

As there are 50 states, the plot is quite crammed, although we can still notice that after the initial massive decrease in mobility (at the onset of the pandemic), some states saw relatively low changes in mobility, while others saw extreme fluctuations. The high resolution version of this plot, with legend, can be found at this link.

To better look at some prevailing and representative trends, we plot certain states of interest at both ends of the spectrum:

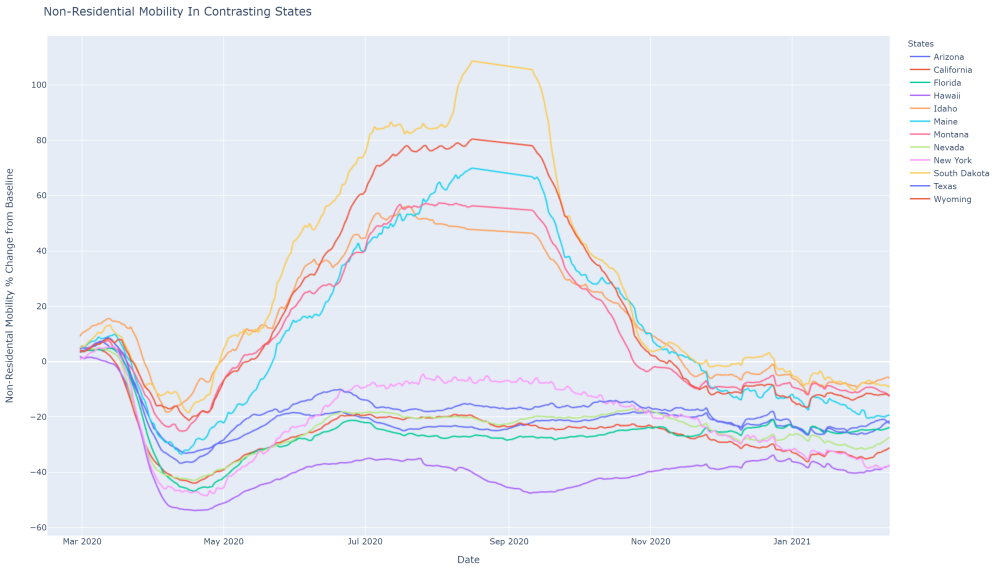


Figure 3: 15-day Rolling Average of Non-Residential Mobility % Change in States of Interest

The full definition hover-able version of this plot can be found at this link.

Looking at the states at the two ends of the spectrum, we can make some inferences. Most of the states experiencing large mobility fluctuations in the summer months are mostly northern

states that see larger seasonal fluctuations in temperature, with lower winter temperatures. This could contribute to the population of those states to naturally be more stagnant during the winter months, and more likely to alter their behaviour to be more mobile in the summer period. Also, as the baseline data was taken in January/February, this inference could be quite well defended.

On the other end of the spectrum, we see states that are known to be warmer and more consistently temperate saw mobility levels hold steadily below pre-pandemic numbers, and experienced very mild fluctuations in the summer. This trend seems to extend beyond the sociopolitical spectrum and level of lockdown measures, as states such as California/NY as well as Texas/Florida are abiding by this trend.

This EDA concludes in that beyond controlling for demographics, misinformation, misconception and other factors, state temperature/climate is a notable and causal factor for population mobility, which in turn is an empirical driver of COVID spread and severity. From this EDA, we then know to be mindful of a diverse range of factors to control for when analyzing the mobility of the states, with these factors being both social and natural.

3.2.2 Information

3.2.2.1 Misconceptions From the Misconceptions and Exposure to information dataset, we have data which gives an idea of the levels of misconceptions by states. The following chart is taken from the report which was published with this data.

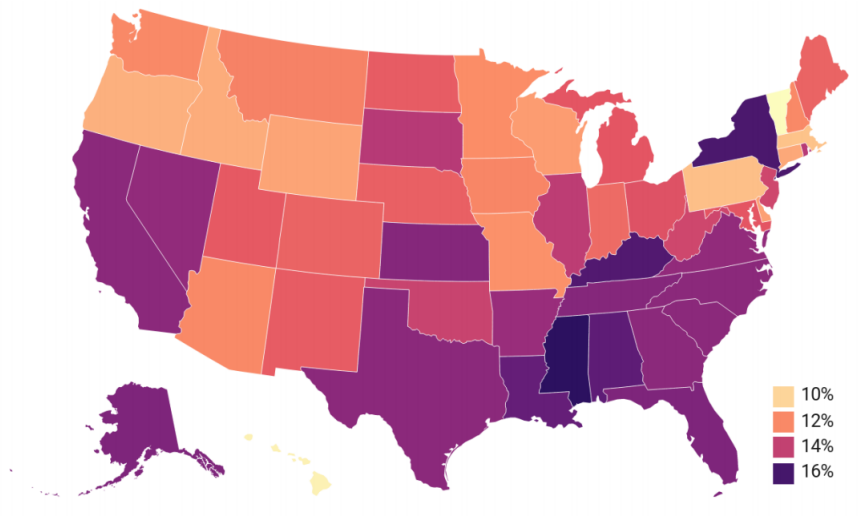


Figure 4: Misconception % by State

From a preliminary analysis, it appears that the southern states show much higher rates of misconceptions than the rest of the country. This is quite interesting since many of these states (Louisiana, Mississippi, Alabama, etc.) are at the same time poorer, less educated, more rural, and republican-leaning. In addition, we see that states with large urban centers such as California and New York show high rates of misconceptions about COVID, and that states in close geographic proximity show similar patterns.

3.2.2.2 Misinformation Similar to misconception, we can also plot our misinformation index overlayed on a map of the United States to highlight some similarities and difference. At first glance it seems as if southern states are still more prone than the rest of the country, but in this case by a much smaller margin. In addition, states such as California and New York (our urban centers), show the highest levels of misinformation.

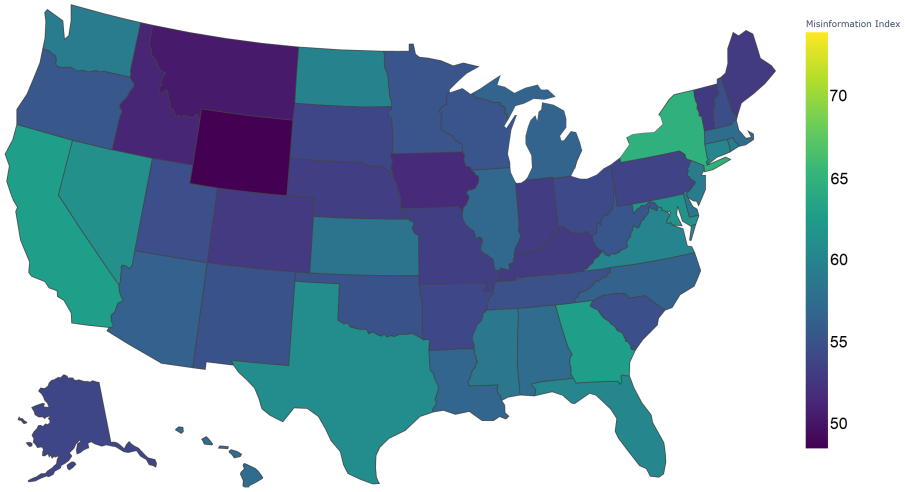


Figure 5: Misinformation Index by State

Overall, we can see quite clearly that there are general trends between misinformation and misconceptions, as one would intuitively expect. However, in this case the differences are much more interesting. The similarities between the states in the south is apparent, as well as the similarities between California and New York, but it is much less obvious how these two groups are related.

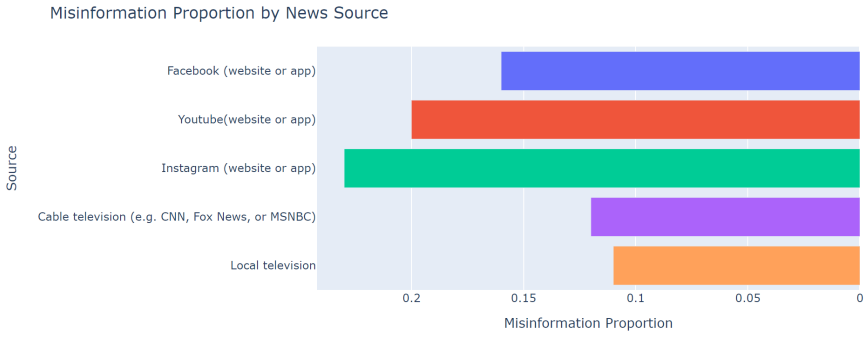
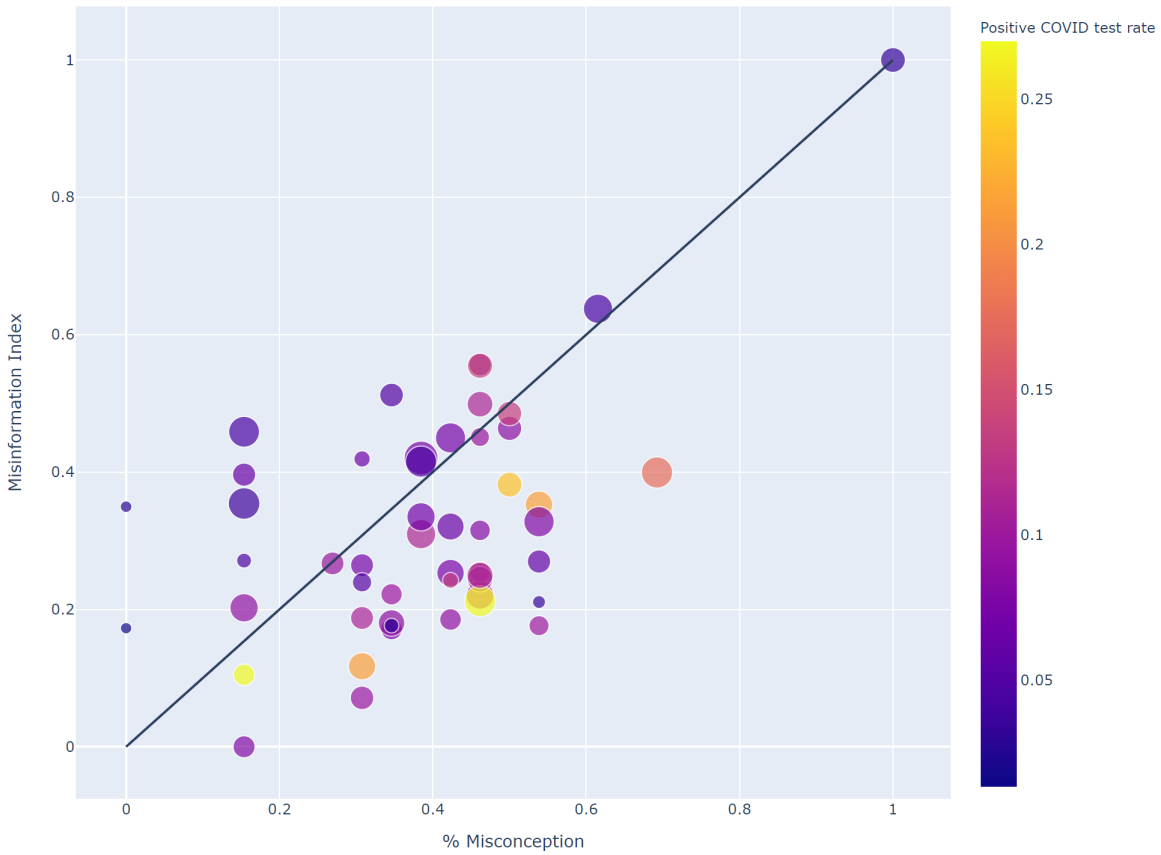


Figure 6: Proportion of false COVID-related information by News Source

Figure 7: COVID News Source by State

The two charts above give a good partial explanation. As shown in the bar-chart, social media platforms, especially Instagram and YouTube, have higher a high proportion of COVID-related information that is false. More urbanized areas such as Washington DC, California and New York have significantly higher proportion of residents who get information about COVID-19 from such sources, and are thus exposed to misinformation. By contrast, more rural states, like Wyoming and Vermont, are less exposed to Instagram and YouTube (levels of Facebook and traditional news usage are similar), and are thus less exposed to misinformation from the media.

To directly compare these two indices we first apply a normalization operation to map each index between 0 and 1. Then we plot each state on the following diagram, with the size of the point corresponding to the number of deaths per unit of population, and the color corresponding the positivity rate in that state. By then superimposing a diagonal line, we do see a noticeable difference between those points above and below the diagonal line. Those below the diagonal appear to be greater in number, with generally higher positivity but lower death rates, and are more affected by misconceptions as opposed to misinformation. Those above the diagonal appear to be smaller in number, with smaller positivity, higher death rates, and correspond to those states where misinformation is a larger driving factor. Overall, we limit ourselves from making any definitive conclusions at this point, but it is quite obvious that there are differences between misinformation and misconceptions in the United States.



3.2.3 Demographics

3.2.3.1 Factors Explored To explore the different socioeconomic, racial, and geographic factors affecting ones susceptibility to misinformation, misconceptions, and compliance of government movement restrictions, we gathered data from a number of sources. The contents of these datasets, how they were gathered, and who they were published by is all included in the Appendix sections. From our cleaning, wrangling, and feature engineering process described previously, we are able to leverage these datasets to provide state-level cross-sectional data on the following factors.

- Unemployment Rate
- Median Household Income
- Education Rates
- Urbanization Percentages
- Population
- Age
- Race

Before investigation, we can generate a correlation matrix for every factor in our analysis:

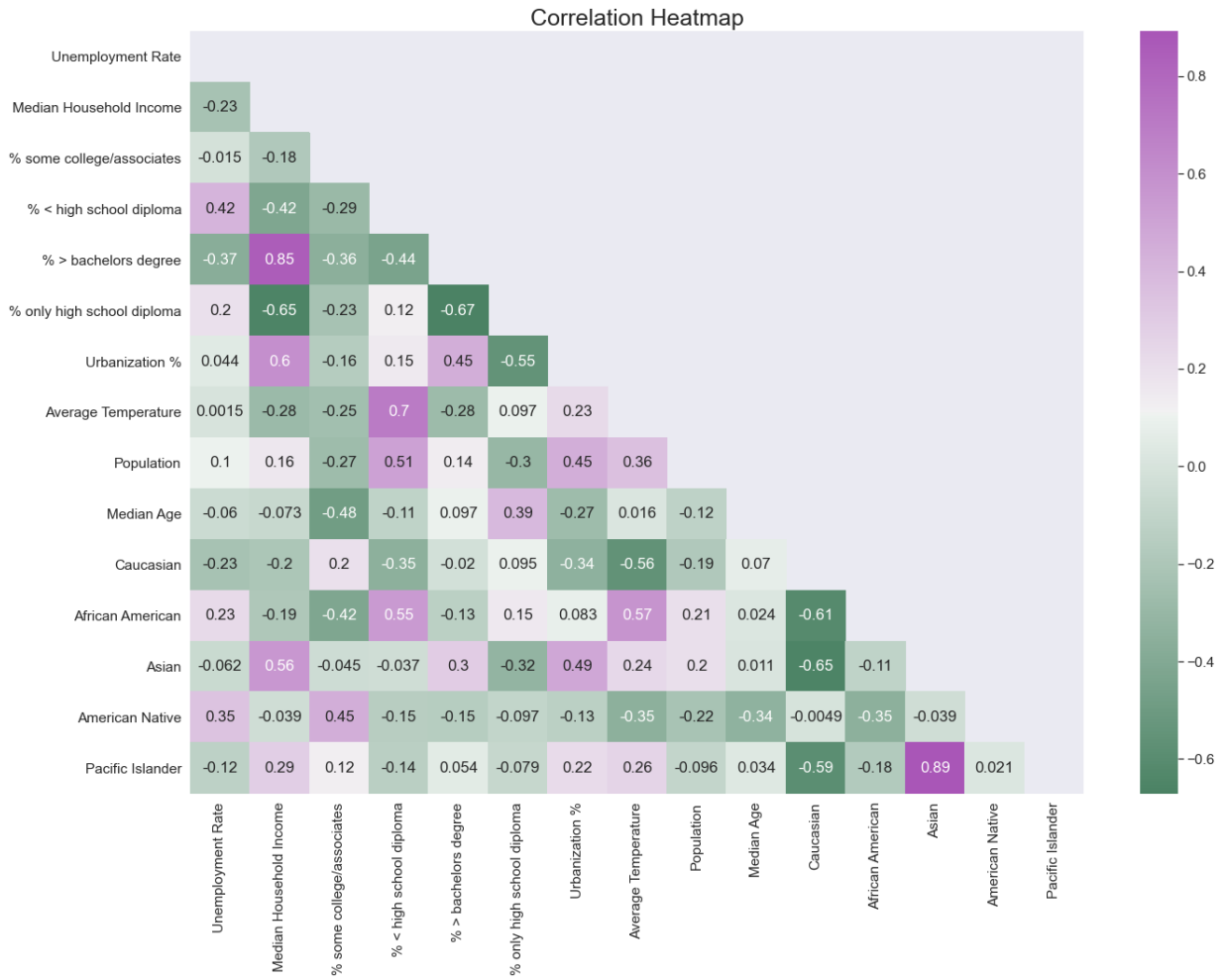


Figure 8: Correlation Heatmap of Socioeconomic Variables

For most factors, we see small but non-insignificant correlations. As a result, when running our regression analysis on these variables we will keep in mind that multicollinearity may be present in our findings. For highly correlated variables, such as various education levels with Median Household Income, we will try to avoid using both factors in the same model.

3.2.3.2 Exploration of Trends To explore if any of these factors appear to be correlated with misinformation, misconceptions, mobility, or COVID-19 severity, we overlay state plots onto a US map, and look for any noticeable trends or anomalies. Just from a preexisting intuitive understanding, we would assume that certain demographics would have varying levels of susceptibility to many of the variables we plan to examine.

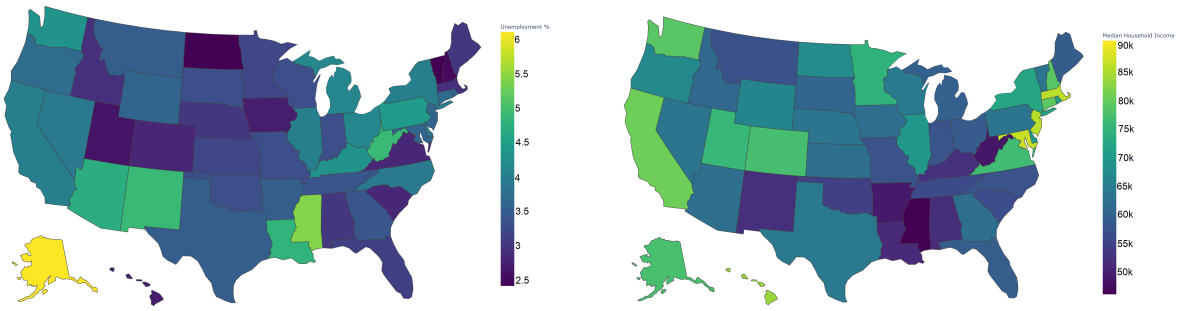


Figure 10: Unemployment (left), Median Household Income (right)

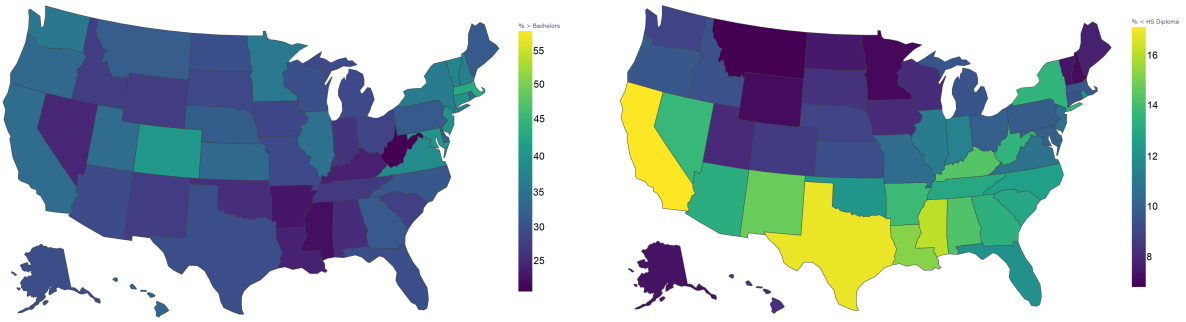


Figure 12: % with Bachelors (left), % with < High school diploma (right)

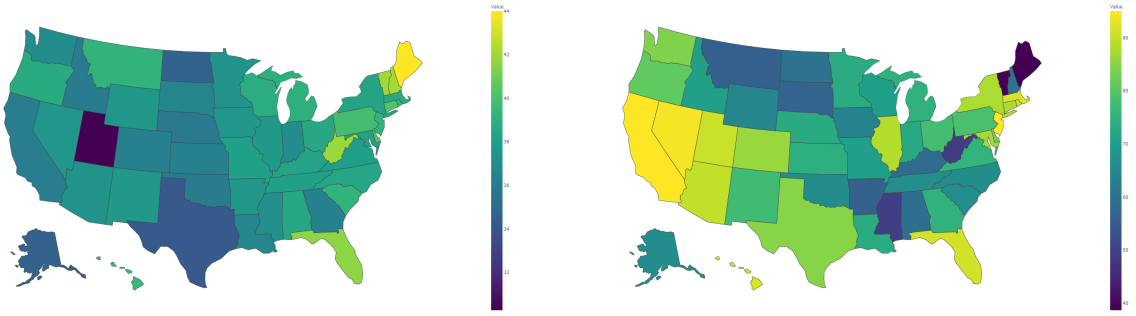


Figure 14: Median Age (left), Urbanization Level (right)

For this portion of the Exploratory Data Analysis, we leveraged a qualitative analysis of a number of trend maps that we were able to generate for many of the factors that we gathered. Our hypothesis was restricted to an intuitive understanding that some or all of these factors would have a noticeable impact. When compared to the mobility, misinformation, and misconception maps shown in previous sections, we do see that there are some noticeable trends. For example, misconceptions are high in regions of large high school dropout rates. We see that misinformation is high in regions of high unemployment and urbanization. Furthermore, mobility is much higher in colder, less urban, and lower median age states.

All of these insights are illuminating, but not rigorous. At this point, we have proven our hypothesis that some of these factors *are* related, but we have yet to prove exactly *how* they are. The proceeding rigorous analysis will be explained in the analytical modelling section.

3.3 Analytical Modelling Summary

Our exploratory data analysis revealed evidence of possible associations between demographics, misinformation, misconceptions, mobility, and ultimately COVID-19 outcomes. More importantly, preliminary findings as well as intuitive understanding gave us reason to believe there may be a causal effect chain between these associations.

We are aware that there are many factors contributing to the COVID outcomes and it is nearly impossible to control all of them. Simply linking the misinformation to the COVID through regression at high level may involve too much Omitted Variable Bias (OVB) in this case and undermine the causal relationship we want to investigate. Therefore, we would like to stay focused on proving one logical chain that we believe to be both important and interesting with statistical measures. In order to do so, we separated the whole process into steps: first we analyze COVID-Mobility relationship, and then we investigate Mobility-Misinformation relationship. To be rigorous, we also investigated COVID-Misinformation relationship and included in the Mobility-Misinformation section.

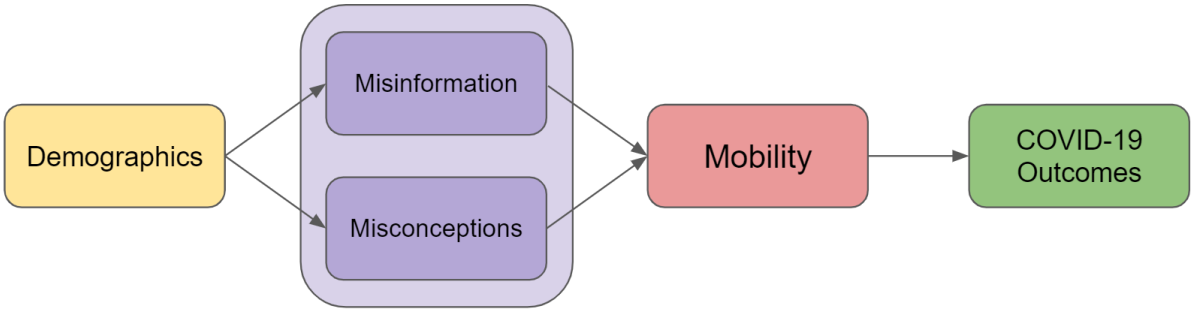


Figure 15: Causal Link Flow Chart

3.4 Mobility

3.4.1 Impact of mobility on COVID spread

To link the mobility data to COVID-19 data, we conducted regression analysis using the Google Mobility Data that we acquired from online sources ourselves and the US state-level COVID-19 data provided by the Covid Tracker Project. The Google Mobility Index Dataset is comprised of six dimension: Retail, Grocery, Parks, Transit, Workplace, Residential. The residential index describes how much people stay at home and a higher index means people comply with the lockdown policy better. For COVID data, we are aware that positive case number is related to tests conducted, and may be inaccurate or incomplete, so we decided to use metrics that could avoid this issue. We selected four metrics for COVID-19 data severity that could minimize the measurement error:

- Death Increase
- Hospitalized Increase
- Currently in ICU
- Currently on Ventilator

	Death Increase	Hospitalization Increase	In ICU	On Ventilator
R^2	0.085	0.031	0.068	0.032
Intercept p-Value	0	0	0	-0.002
Mobility (PCA) p-Value	0.015	0	0.382	0
AvgTemperature p-Value	0	0	0	0.171
Urbanization % p-Value	0	0	0	0
day since record % p-Value	0.01	0	0.726	0

Table 1: OLS with PCA result

We took two approaches to analyze the effect: one is the multi-variable ordinary-least-square (OLS) and the other one is the Panel Regression. We regressed each of the select COVID metrics against all five dimensions of mobility data plus ***day_since_record***, ***AverageTemperature***, and ***Urbanization%*** feature. ***day_since_record*** is calculated by the number of days since the creation of the first record for each state to isolate the time series effect, which is proven to be prevalent. Since the metrics show strong seasonality, we used the last 7 day sum (y-label) and last 7 day mean (x-metrics) to smooth the time-series data. As COVID usually incubates for up to 2 weeks, we also lagged our data by 14 days to compensate for this. We first run a single regression model on all state-level data against all 4 COVID severity metrics from above, and detected statistical significance on all independent variables, meaning all these mobility metrics as well as the time feature are impactful on COVID metrics. We then split up the dataset by state and used individual models to fit them. While not every state showed statistical significance, on average about 80 percent of the states showed significance for each of the mobility dimension. These results are available in the appendix.

We are fully aware of the inverse causal relationship that the COVID situation made people reduce their mobility, and used the 90 days leading techniques on dependent variables(COVID metrics). In other words, the dependent variable of each day corresponds to the COVID 90 days later, and give it enough time for the mobility to take effect.

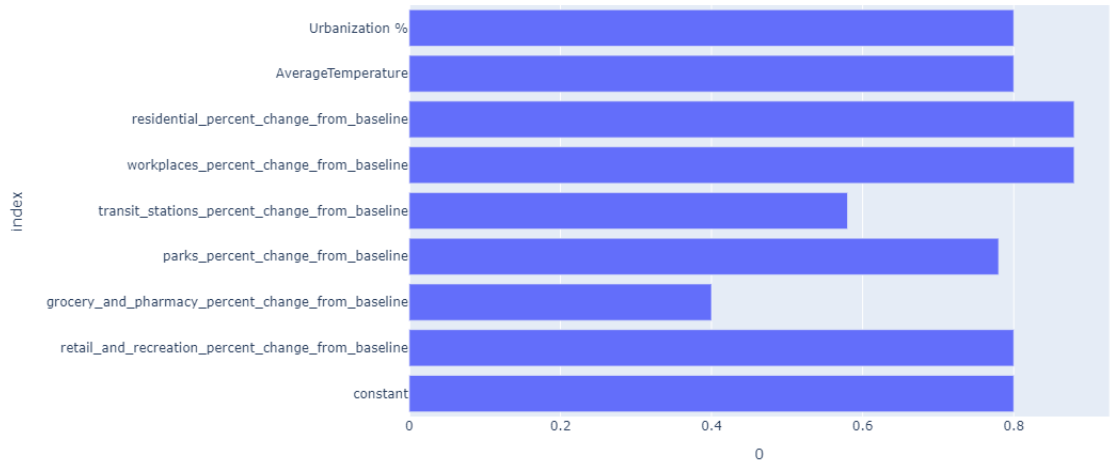


Figure 16: % of states showing statistical significance

We are also aware of the multi-collinearity problem within our features, and we therefore applied Principle Component Analysis (PCA) to summarize all the mobility features (capturing 87% variance), and run the regression once again. The result were even more clear with 90 percent of the states showing statistical significance. We also conducted Panel Regression and ran a Fixed Effect model using only comprised index and using PCA on all 4 COVID metrics. The model again shows statistical significance in the coefficient and proved the relationship between Mobility Index and the COVID stats.

Along with statistical measure, this would also makes sense in terms of how COVID spreads. The more outdoor activity leads to more chances of in-person interactions and therefore more infection opportunities. In addition, 90% of the states showed statistical significance including some of the less urbanized states, implicating the magnitude of COVID infectiousness and the necessity of lockdown policies.

3.5 Misconceptions, misinformation and disease outcomes

3.5.1 Methods

Current government measures to control the spread of COVID-19 mostly focus on people's physical movement. Government and big media companies have also paid attention to misinformation problem, which is particularly prevalent on social media. We believe that misconceptions and exposure to misinformation could have impact on both people's physical movements, as well as outcomes of the disease as measured by infection, hospitalization and deaths. We mainly used linear regression models to identify the relationship between variables of interest and establish the link between misconceptions, exposure to misinformation, people's physical movement and disease outcome.

3.5.2 Impact of misinformation on disease outcomes

It is intuitive to think that states where people have more misconceptions about COVID-19, or are more exposed to misinformation about COVID-19 from the media, should witness worse disease outcomes as a result. The following simple linear regressions indicate that this is probably true, but we need more control variables to reduce noise.

	Pos. Test Rate I	Deaths/Thousand I	Pos. Test Rate II	Deaths/Thousand II
Intercept	0.0341	0.0013	0.2633**	-0.0013
	(0.0523)	(0.0014)	(0.1081)	(0.0028)
R^2	0.0287	0.0655	0.0467	0.0620
R^2 Adj.	0.0089	0.0465	0.0272	0.0428
Misinformation Index			-0.0030	0.0001*
			(0.0019)	(0.0000)
% Misconception	0.4433	0.0177*		
	(0.3685)	(0.0095)		

Table 2: Misconception and Misinformation regression against COVID statistics

Other variables, including racial composition and urbanization would both complicate the re-

lationship between misinformation and COVID-19 outcomes. However, given that we only have 50 observations in the cross-sectional dataset, we took a conservative approach in adding new control variables. As shown in the following table, the positive relationship between COVID-19 misconceptions and severity of disease outcomes remain. For example, given that the standard deviation of pc_misconception across states is about 0.02, we can conclude that a standard deviation increase in pc_misconception is associated with 2.5 percentage points' increase in test passivity rate, and 0.0004 deaths increase per thousand.

Also interesting is the fact that although misconceptions are related to worse disease outcomes, exposure to misinformation from media sources does not have a significant impact on disease outcomes.

	Pos. Test Rate	Deaths/Thousand
Intercept	0.1768	-0.0006
	(0.1965)	(0.0052)
% Misconception	1.1730**	0.0216*
	(0.4678)	0.0125
Misinformation Index	-0.0046	0.0001
	(0.0033)	(0.0001)
Caucasian Race	0.0165	0.0017
	(0.0822)	(0.0022)
Urbanization Rate	0.0001	0.0000
	(0.0007)	(0.0000)
R^2	0.1584	0.1597
R^2 Adj.	0.0836	0.0850

Table 3: Regression Table for above factors of interest and COVID statistics

3.5.3 Impact of misinformation on mobility

As shown above, high physical movement is an extremely strong predictor of worse disease outcomes at the state level. Knowing that misconceptions about COVID-19 are also associated with worse disease outcomes, it is worth exploring whether people who have more misconceptions about COVID-19 are less mindful about reducing physical movement. The following table shows result of regressions of misconception and exposure to misinformation on physical movement in non-residential areas. Without controlling for level of strictness at each state, misconceptions are associated with more movement in non-residential areas. Specifically, the first model shows that a standard deviation increase in pc_misconception is associated with 3 percentage points' increase in non-residential mobility.

After controlling for strictness of government measures, the positive relationship between COVID-19 misconceptions and non-residential mobility remains, but the magnificence of impact decreases. However, we believe as state governments adjust its policies based on the perceived importance of reducing physical contact among residents, strictness of government measures is partially a variable internal to the level of COVID-19 misconception in the state.

Also worth noting is the fact that higher exposure to media misinformation is shown to be associated with **less** mobility in non-residential areas. A possible explanation is that although certain media sources, especially social media platforms do contain large amount of misinformation,

such information does not necessarily make people less likely to physically distance, or be careful in protecting themselves from COVID-19.

	Non-res % Change from baseline I	Non-res % Change from baseline II
AverageTemperature	-0.960***	-0.997***
	(0.240)	(0.216)
Intercept	93.952***	98.081***
	(18.023)	(16.213)
R^2	0.676	0.745
R^2 Adj.	0.648	0.716
Stringency Index		-0.358***
		(0.104)
Urbanization Rate	-0.234**	-0.226***
	(0.092)	(0.083)
Misinformation Index	-1.590***	-1.172***
	(0.410)	(0.387)
% Misconception	134.903**	64.356
	(62.803)	(59.939)

Table 4: Regression table for non-residential change in mobility

3.6 Identifying Vulnerable Demographics

Up until this point, we have been able to identify two main causal drivers of COVID-19 spread, hospitalization, and death - namely misconceptions and mobility. Rates of these metrics are able to give great insights into *what* the problems are, but they fail to identify *where* the problem is. Using compiled data from many of the external datasets we brought in, we are able to construct a likely "demographic makeup" of those that would be most susceptible to misconceptions, and more likely to disregard lockdown measures, leading to higher mobility rates. By doing so, we are not only able to pinpoint *where* vulnerable groups of individuals are, but we are also able to identify *which* driving factor they are most vulnerable to.

3.6.1 Misinformation/Misconceptions

For the first link in our proposed causal chain, we looked at whether an individual's demographic makeup makes them any more or less prone to misconceptions, or exposure to misinformation. We first ran an ordinary least squares regression model to determine if there were any causal effects between urbanization level, age, or education and misinformation and misconceptions. Due to some moderate levels of collinearity between these variables, we saw high VIF numbers. In addition, our model showed a somewhat large condition number.

To reduce some of the multicollinearity that may have been present, we first tried to apply an elastic net model to place a penalty term on high coefficients. Unfortunately the R^2 of the output model was reduced quite drastically, and many of the factors which we wanted to observe relationships of got unnecessarily eliminated. As a result, instead, we tried to standardize our predictor factors by applying a mapping between 0 and 1. This served a few purposes - first, as shown in the following table our VIF values dropped quite drastically, as well as our condition

number. Secondly, we are now able to look at the relative weightings of each component coefficient in our analysis, to determine the most impactful parameters.

	Before Normalization	After Nomalization
Median Age	421.728513	21.803313
Urbanization %	44.003164	10.745958
% > Bachelor's Degree	119.949468	13.603153
% < High School Diploma	23.349039	4.253047
% only High School Diploma	142.616335	9.542684

Table 5: VIF values, before and after normalization

We ran this regression model on both our **misinformation** index, as well as our **misconception** index. Surprisingly, the statistical significance of our predictor variables, as well as even the coefficients at times changed.

For **misconceptions** we found that both median age and urbanization percentage both showed statistically significant (below p=0.05) negative coefficients, while percentage of population with less than a high school diploma showed a statistically significant positive coefficient. This means that old, rural regions, with significant high school dropout rates are significantly more at risk for **misconceptions** than the rest of the country.

Conversely, for **misinformation**, we found a rather interesting different result. We found that percentage of population with less than a high school diploma, as well as percentage of population with greater than a bachelors degree, showed statistically significant positive coefficients, whereas those with only a high school diploma showed no strong directional coefficient. This means that misinformation is disproportionately affecting the lowest education levels, as well as the highest, but not those in the middle. We found this quite surprising.

In addition, we saw no significance for median age, but did see a weak positive urbanization percentage coefficient (below p=0.1). This was also very surprising to us, because the urbanization coefficient changed signs between misinformation,and misconceptions. Overall, this proves our original hypothesis - that misinformation and misconceptions are two different issues plaguing the United States. The visualization below shows the change in coefficients from each regression, and the results of both of these models are provided in the appendix.

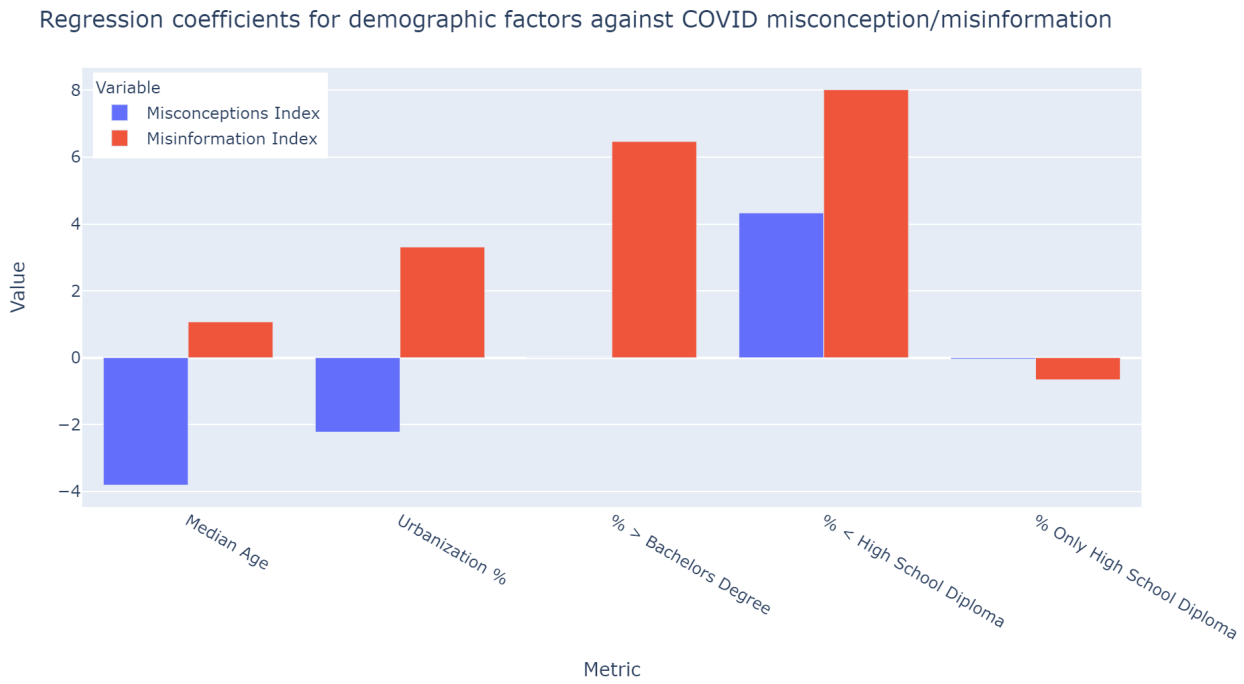


Figure 17: Regression coeffs. for demographic factors against misconception/misinformation

3.6.2 Mobility

The next step in our causal chain is mobility. We already proved that more misconceptions about COVID-19 are linked to increased rates of mobility, so we would expect to see a similar result for the demographics we have already identified.

To highlight any causal effects between our factors and the google mobility data that we gathered, we implemented an OLS using age, education, race, urbanization, and temperature. Our model produced very good results, achieving an $R^2 = 0.798$. We found that there was no significance to race (provided that we controlled for median household income), however there was a statistically significant negative correlation with factors median age and urbanization level. In addition, we included average temperature in the regression to help control for the temperature bias of the mobility data explained earlier, and we do see a negative coefficient, albeit with a p-value of 0.263 - so insignificant. The sign on this coefficient makes sense in our model.

Additionally, and perhaps most interestingly, we saw another strange relationship regarding education levels. We saw that percentage of population with less than a high school diploma as well as percentage of population with greater than a bachelors degree, showed very significant negative coefficients (p-values of 0.000, 0.003), while those with only some form of college or associates degree showed an insignificant, slighty negative coefficient. Once again, low and high education levels acted similarly. Here, those with high and low levels of education were more likely to stay home and listen to lockdown orders, whereas those with some form of education were less likely to.

In this instance, we see that the most at risk group when it comes to mobility are those that are young, living in non-urban areas, with moderate education levels. Once again, the results of this model, as well as coefficients and p-values are included in the appendix section.

3.6.3 COVID-19 Severity

Finally, we examine if there are any causal effects between demographic and actual COVID-19 outcome. Although we have already proven that the causal chain described at the beginning of this section does exist, we also tried to see if there were correlated effects while skipping certain layers. The COVID-19 metrics we used for this analysis was normalized Deaths per population, as well as overall positivity rate.

For this model, we were unsure which effects would possibly have the most effect on COVID-19 outcome, so we used all of the variables we had. Understanding that this may cause a considerable amount of multi-collinearity in out model, we applied a LASSO regression model to choose which of the factors were the most meaningful. From here, as was the case for misconceptions, we saw high VIF values and a high condition number. We applied the same normalization mapping technique to reduce this.

For both models, we found that many of the relationships that were previously linked to misinformation, misconceptions, and mobility are much less significant in this case. For positivity, we saw that the only statistically significant factors were average temperature (positive coefficient) and median age (negative coefficient). As for education and urbanization levels, there was no significant trend in either direction.

In the case of the deaths model, we saw that urbanization was the main significant factor ($p=0.012$), with higher levels of education showing lower death rates. In this case however, not all metrics of higher education were significant, but lower rates of education showed much less negative coefficients, indicating the relationship may not be as strong. All coefficients and p-values are again included in the appendix.

4 Conclusion and Discussion

4.1 Quantitative Conclusions

After identifying the links between lack of basic knowledge about COVID and more mobility in non-essential areas, as well as severity of disease outcomes as measured by infection, hospitalization and death, we have identified social-economic, demographic and geographic factors and groups associated with more misconceptions about COVID-19. Using cross-sectional and panel regressions, we modelled the links between increased physical mobility in public areas that worsen the effects of the pandemic. Using survey data, we then identified the relationship between prevalence of COVID-19 related misconceptions, as well as exposure to misinformation about COVID-19, and mobility and disease outcomes.

This analysis highlights that between exposure to misinformation, and possession of misconceptions about COVID-19, the latter is observably more strongly linked to potential lack of compliance to public health measures and therefore more severe pandemic effects. We have also identified demographics that are older, more rural, less educated (no high school diploma) and more economically disadvantaged are more likely to possess misconceptions about COVID-19, and in turn demonstrate more mobility in non-essential areas, less compliance to public health measures, and suffer more severe effects from the pandemic.

Below we include various plots that show our predictions as to which regions across the United States are most at risk of deaths from COVID-19. Each plot leverages one of the different models we

have created to predict which areas require the most direct response. As was mentioned throughout the report, even though these plots are just showing the geographic locations, each region is unique. We sought to prove more than just where we need to focus our efforts, but also to prove *how* we could help each region most effectively. One conclusions that we could draw, since these plots are rather dissimilar is that different demographics are at risk from different underlying issues. This is not a one-size-fits-all approach. Rather, we have shown that by looking into the susceptibilities of each demographic to different issues, we can hopefully craft a response that is best for everyone in reducing the severity of the COVID-19 pandemic. The combined visual map, with demographic, misconception and mobility data all taken together, and representing our final, full prediction, can be found at this link.

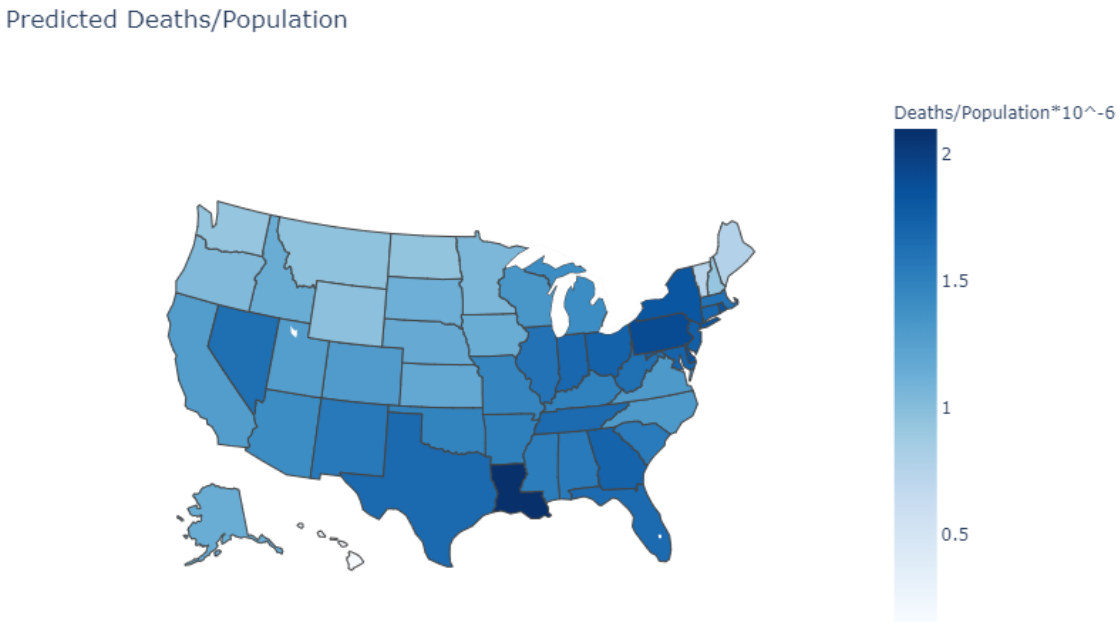


Figure 18: Predicted Deaths from Demographics

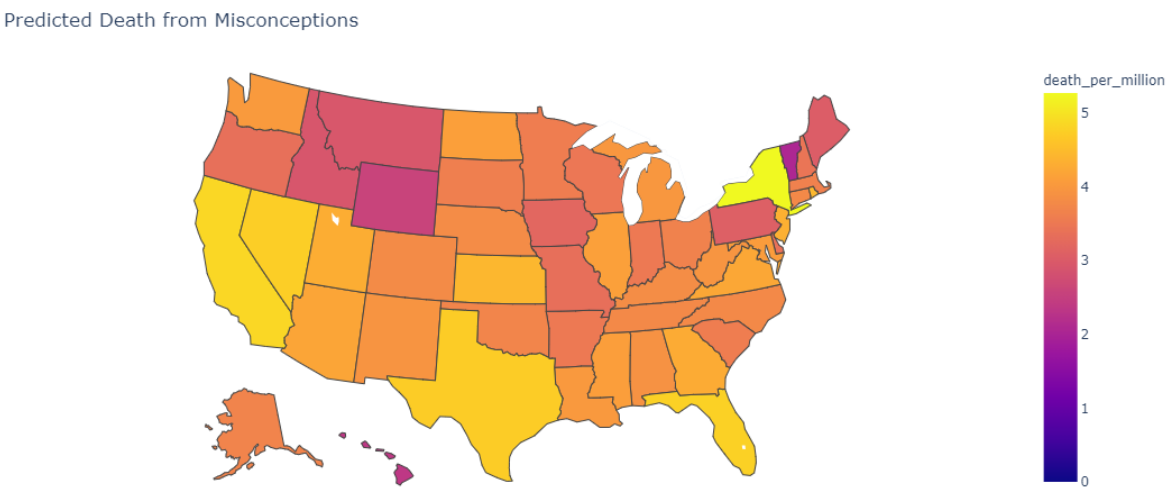


Figure 19: Predicted Deaths from Misconceptions

4.2 Recommendation

In this report, we were able to identify specific at risk demographics in the United States, and more specifically, what exactly they are at risk from that could lead to the worsening of the pandemic. We found that certain factors such as age, urbanization, and education often played a very significant role in predicting how susceptible an individual is to misconceptions, how willing they are to comply with mobility restrictions, and how likely they are to face negative effects from the COVID-19 pandemic.

Although statistically we have drawn the conclusion that misconceptions are a more direct driver of worsening COVID-19 effects, we do not downplay the contextual effects of misinformation. False information from the media and public sources, especially social media, is prevalent, has devastating potential, and in turn may result in new misconceptions in the future.

From the analysis was have completed, we are now in a position to provide relevant authorities, organizations and regulatory agencies guided and specific recommendations and strategies. Federal and state governments and other agencies could effectively use their resources to craft more impactful measures and regulations, such as focusing on education for demographics vulnerable to misconceptions, while simultaneously combating digital misinformation in areas that see more prevalence and spread. This targeted, two-pronged approach could address those most vulnerable to the pandemic, as well as prevent the formation of more misconception-vulnerable groups, ultimately alleviating the effects of not just COVID-19, but also possibly that of future pandemics.

5 References

- The Covid States Project by *Consortium for Understanding the Public's Policy Preferences Across States*. Available: https://covidstates.org/.
- The COVID Tracking Project at *The Atlantic*. Available: https://covidtracking.com/.
- Google LLC "Google COVID-19 Community Mobility Reports". Available: https://www.google.com/covid19/mobility/data_documentation.html?hl=en.
- Hale, Thomas, Noam Angrist, Emily Cameron-Blake, Laura Hallas, Beatriz Kira, Saptarshi Majumdar, Anna Petherick, Toby Phillips, Helen Tatlow, Samuel Webster (2020). Oxford COVID-19 Government Response Tracker, Blavatnik School of Government. Available: www.bsg.ox.ac.uk/covidtracker.

6 Appendix

6.1 Datasets

6.1.1 Mobility Dataset

As empirically viewed, mobility of population, particularly in shared areas, is a visible driving factor in the spread of COVID-19. Hence, many authorities have issued physical distancing and quarantining measures in an effort to reduce movement. These measures have seen varying degrees of success around the US, mainly due to the compliance of those under these measures.

To better understand mobility under pandemic health measures, starting on February 15, 2020, Google has been tracking mobility across all US states as well as countries worldwide to illustrate the change in human movement patterns during the pandemic. The change in mobility is compared to baseline day values from a 5-week period that is considered to be pre-COVID (Jan. 3 to Feb. 6), given as median value for each particular day of the week to account for changes in weekday/weekend routines. Although this is not a comprehensive representation of population movement, its use is nonetheless justified given the high prevalence of cellphone usage and breadth of Google's data collection methods in the US. Information in this dataset is relatively comprehensive and complete.

As part of our EDA, we looked at mobility for non-residential purposes (i.e. not staying at home/isolating). The subcategories under this are:

6.1.2 Education Dataset

In order to identify any relevant correlations between education levels and the provided COVID data, we explored an additional dataset, published by the US department of agriculture. This dataset provides metrics which describe various education metrics on a per county level. Data was gathered as a 5 year average of the Census Bureau's American Community Survey, which was administered every year from 2014 to 2018 (inclusive). Some of the relevant data that we have access to here is described in the following list:

- Percent of adults with less than a high school diploma

- Percent of adults with a bachelors degree or higher
- Percent of adults completing some college or associates degree
- Percent of adults with a high school diploma only

6.1.3 Economic Dataset

To isolate the effects that education have on misinformation and by extension covid response/results, we also bring in an economic dataset to help us control for correlated factors such as median household income and unemployment rates by state. This dataset contains primarily historical unemployment data by county, but it also includes 2019 metrics for household income and unemployment that will aide our analysis.

6.1.4 Misconceptions and Exposure to Information

This project aims to inform policy makers best approaches to disseminating accurate information about the prevention and treatment of COVID-19 to the public. The project comprises of a series of surveys on people's news consumption, disease prevention behavior, institutional trust, and beliefs when it comes to COVID-19. The survey uses representative samples from all 50 states in the US, and has been conducted for 13 rounds from April to November 2020. We obtained two sets of data from the project, namely people's conceptions about the disease (whether they are correct or incorrect), and people's news sources when it comes to COVID-19 information. The survey also included information on the prevalence of false information, as measured by the proportion of false statements that respondents identify as true, by news source.

6.1.5 Age Breakdown Dataset

In order to accurately classify the impact of age on many of our metrics of interest, we gathered age data from the United States Census Bureau from 2010-2019. This dataset included a comprehensive breakdown of estimated age, by integer age value, by state.

6.1.6 Race Breakdown Dataset

In addition to classifying age demographics, we also wished to examine the effects of race on many of our variables of interest. We gathered this data from the same source as the age data - from the United States Census Bureau. Race is broken up into five categories.

- White alone or in combination
- Black alone or in combination
- American Indian and Alaska Native alone or in combination
- Asian alone or in combination
- Native Hawaiian and other Pacific Islander alone or in combination

6.1.7 Temperature Dataset

Since there is a clear correlation between temperature and COVID-19 statistics, as well as a clear temperature bias from our mobility data, we found a cross-sectional temperature dataset which included average temperatures across geographic regions in the United States. In our case, we divided this information by state. This information comes from the National Centers for Environmental Information - Climate at a Glance. It uses weather stations across the country in conjunction to give accurate metrics on average temperature, so we believe it to be a very credible source.

6.1.8 Urbanization Levels

To accurately control for, and examine the links of varying urbanization levels across the country, we gathered a dataset from the United States Census Bureau, from their Decennial Census in 2010. This dataset gives estimates regarding the percentage of a states population living in urban areas. We do acknowledge that this dataset is now 10 years old, and may be slightly out of date, but we believe that it is still able to give us an accurate representation of the comparisons between states.

6.1.9 The COVID Tracking Project

To obtain a comprehensive and accurate picture of the prevelance of COVID-19 across US states over time, we collected data from the COVID Tracking Project, a source recommended by Citadel. We utilized its API here to download historical values for all states.

The COVID Tracking Project data contains a wide range of variables key to the project. Specifically, we have obtained information about:

- Number of new cases (confirmed and probable combined). We have also calculated daily increase in negative and positive test results.
- Number of new hospitalized patients
- Number of new deaths

We believe that the source is largely accurate and unbiased. However, it is partially generated by the left-leaning magazine *The Atlantic*. We will be mindful the party affiliation of the federal and state leadership when analyzing our data. Information in this dataset is relatively complete.

6.1.10 Covid-19 Policy Responses

To get an understanding of policy responses to contain the spread of COVID-19, we have also collected data from Covid-19 Policy Responses project the USA state level run by Blavatnik School of Government, Oxford University. The dataset contains topics including school, workplace, and public transport closing, cancellation of events, restrictions on gathering, stay-at-home orders, and restrictions on domestic and international movement.

In addition to these specific policy observations, we also used two composed indices, namely containment and health index and stringency index. Information in this dataset is comprehensive and complete.

6.2 Regression Results

6.2.1 Misinformation/Misconceptions

	Misconceptions Index	Misinformation Index
Const	15.6989***	47.7990***
	(1.4077)	(2.4298)
Median Age	-3.8078**	1.0737
	(1.6446)	(2.8387)
Urbanization %	-2.2237**	3.3105*
	(1.0702)	(1.8473)
% > bachelors degree	0.0027	6.4662**
	(1.7613)	(3.0401)
% < high school diploma	4.3312***	8.0138***
	(0.9596)	(1.6563)
% only high school diploma	-0.0397	-0.6586
	(1.8915)	(3.2648)
R^2	0.5023	0.5473
R^2 Adj.	0.4457	0.4959

Table 6: Regression table for demographic factors against COVID misconception

6.2.2 Mobility

	Non-Residential Mobility % Change from Baseline
const	110.7092***
	(40.5987)
Median Age	-1.3482**
	(0.5149)
% some college/associates	-0.5729
	(0.4044)
% > bachelors degree	-0.9268***
	(0.2931)
Caucasian	31.3454
	(26.9121)
African American	21.3547
	(27.5056)
Asian	-26.9274
	(43.8577)
Urbanization %	-21.0406**
	(8.5973)
% < high school diploma	-2.3268***
	(0.5595)
Average Temperature	-0.3503
	(0.3088)
R-squared	0.7982
R-squared Adj.	0.7528

Table 7: Regression table for demographic factors against non-residental mobility

6.2.3 COVID-19

	Positive COVID Test Rate	Death Rate
const	243.6985*	243.6985*
	(139.1915)	(139.1915)
Caucasian	-52.5136	-52.5136
	(134.5544)	(134.5544)
African American	9.1434	9.1434
	(82.6134)	(82.6134)
Asian	-179.1463	-179.1463
	(148.2705)	(148.2705)
Median Age	-20.9137	-20.9137
	(56.7967)	(56.7967)
% some college/associates	-96.6546**	-96.6546**
	(45.8150)	(45.8150)
Urbanization %	102.0892**	102.0892**
	(38.8392)	(38.8392)
% < high school diploma	-1.3855	-1.3855
	(46.0438)	(46.0438)
% > bachelors degree	-86.5299	-86.5299
	(53.1112)	(53.1112)
AverageTemperature	-24.4906	-24.4906
	(56.9773)	(56.9773)
R-squared	0.4295	0.4295
R-squared Adj.	0.3011	0.3011

Table 8: Regression table for demographic factors against COVID statistics