An exploration of the effects of public transit density, race and age on COVID-19 transmission and severity in New England

Urban Virus: zasghar1, skarwa, skothar7, usodhi

Goal and Hypotheses

In contrast to existing literature that explores the effects of vaccination history, underlying conditions on COVID-19 severity and transmission, this project explores COVID-19 in an urban context across certain quintessentially urban demographic and transit variables. Primarily, this investigation uses data on COVID-19 cases, population density and transit stops to determine the impact of transit stop density, race and age on COVID-19 cases. Specifically, the hypotheses explored are:

(1) That age group and COVID-19 contraction likelihood are correlated;
(2) That there is a positive relation between transit stop density and COVID-19 cases;
(3) That there is a significant relationship between race and COVID-19 cases.

Data

The exploration draws from data from three sources:

(1) COVID-19 case data for New England states from the Center for Disease Control (CDC).[1] Fields used were age, race, ethnicity, death (y/n). The data was spotty at random before cleaning, and death (y/n) was heavily skewed towards no deaths.
(2) Population density by FIPS Code from the Census Bureau.[2]
(3) Transit stops in New England counties from community-driven transit database Transitland.[3] These were in longitude and latitude before being converted to FIPS codes.

The transit stop and COVID-19 case data was standardized with population density and joined on county FIPS code with transit stop data for comparison and analysis.

Findings

Claim #1: Age and COVID-19 contraction likelihood are related.
Support for Claim #1: A statistical Chi-Squared independence test confirmed the hypothesis that age and COVID-19 contraction likelihood are related, corroborating existing literature with a critical value and test statistic of 5.99 and 28112 for Providence county.

Claim #2: There is no significant relationship between transit stop density and COVID-19 cases
Support for Claim #2: Linear regression, polyfit models and the Pearson correlation test (using a 0.05 significance level) were used, suggesting no significant correlation. Further, Kmeans clustering of New England counties using standardized attributes of transit stop density and case density at an optimal k-value of 3 yielded no clear plane dividing high transit density and high case density with low transit density and low case density. Finally, an SVM classifier to determine the predictability of a case death outcome (death or no death) with only transit stop data as inputs yielded a 51.8% average accuracy on 5-fold cross validation–a coin flip's accuracy for predicting outcomes. Given that most cases resulted in no death, inputs for SVM models were re-balanced using imblearn to prevent outcomes being biased towards over-represented labels.

---

[1] "COVID-19 Case Surveillance Public Use Data with Geography | Centers for Disease Control and Prevention." *CDC Data Sets*, 11 April 2023, https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4. Accessed 15 April 2023.
[2] "2020 Population Estimates FIPS Codes." *Census Bureau*, 25 October 2021, https://www.census.gov/geographies/reference-files/2020/demo/popest/2020-fips.html. Accessed 11 May 2023.
[3] "Transitland." *Interline Technologies*, https://www.transit.land/. Accessed 03 April 2023.

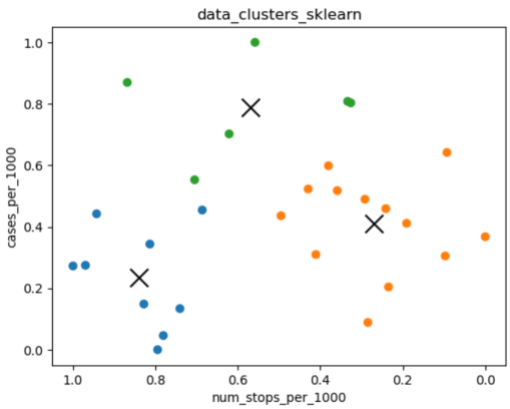

Figure 1: Kmeans clusters do not confirm hypothesis

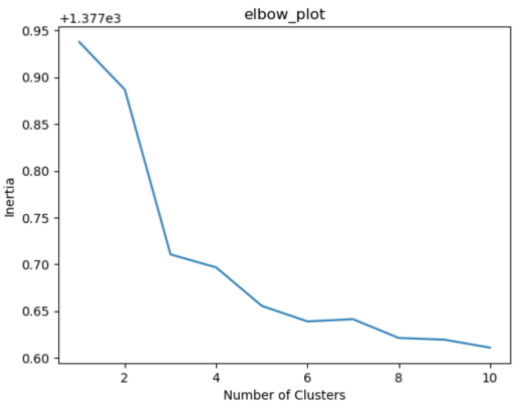

Figure 2: Elbow Plot for determining optimal k

Claim #3: There is a significant relationship between race and COVID-19
Support for Claim #3: A Chi-Squared independence test yielded a critical value of 11.07 and a test statistic of 1994.03 for the variables of race and COVID-19 cases in Providence county suggesting that there is a correlation between race and COVID-19 cases. Furthermore, an SVM classifier to determine the predictability of a case fatality (death or no death) on the basis of race alone yielded a better-than-chance accuracy across 5 folds of 65.6%, weakly corroborating the statistical result.

Claim #4: Multiple urban variables in composite are better predictors for COVID-19 fatality than any one of these variables.
Support for Claim #4: A final SVM was used to predict case death outcomes on the basis of race, sex, ethnicity and transit stop density, achieving an average accuracy of 71.6%, suggesting that multiple urban variables in composite are better predictors for COVID-19 fatality than any one variable.

This result was obtained after correcting imbalances

Socio-historical Context and Impact Report

Socio-historical context

Research the socio-historical context of your project to identify a few societal factors that could affect your data, prediction goal, and/or hypothesis. These factors might include current or historical policies, events, social conditions, larger societal systems, and more. Describe a few of the broader societal issues and their relationship to your data, prediction goal, and/or hypothesis.

Because our investigation pertains to finding the effects of variables on COVID-19 outcomes using large datasets, it is necessary to recognize that every record in the dataset is or was a real human being. Their circumstances and contexts were reduced into a finite set of columns that this investigation draws from, and so inferences drawn are necessarily narrow. Furthermore, even though the majority of alarming consequences of the pandemic have passed (at the time of investigation), the pandemic still affects real people, either directly or indirectly.

Summarize the most relevant technical or non-technical research that has already been conducted about your project topic. If relevant, what was the societal impact of existing research?

The COVID-19 pandemic was unexpected and uncertain, and thus attracted significant attention from academics and medical researchers alike. Due to the novel nature of the pandemic, most research was predicated on vast amounts of data collection, much of which included personal and private data. Indeed, data aggregation often used sources such as COVID-19 contact tracing apps, many of which "pose serious human rights risks".[4] Even beyond contact tracing apps, there are ethical concerns about many studies that conducted data collection while the pandemic was ongoing: an Oxford report poses the question, "Is it ethically appropriate to conduct my research study during the COVID-19 pandemic?".[5] This was a question of great importance to many researchers who conducted research that required effort and time on the part of COVID-19 patients. In addition to data collection risks, COVID-19 case outcomes themselves are partially a function of how many resources governments directed towards a specific county, potentially creating unfair outcomes and skewing results.[6]

Discuss the impact of your socio-historical research findings on your project.

While our data collection and analysis was conducted through secondary sources, the original data was inevitably conducted while the pandemic was still at its peak, potentially requiring patients to give up their

---

[4] Brown, D., & Toh, A. (2020, May 13). *Covid-19 apps pose serious human rights risks*. Human Rights Watch. https://www.hrw.org/news/2020/05/13/covid-19-apps-pose-serious-human-rights-risks?gclid=CjwKCAjwx_eiBhBGEiwA15gLN77kdw0qSuk75S602d2ZGwHds0-TEyP5WRXwlQfDy1r2c3Ku6yB4FxoC_-wQAvD_BwE

[5] Hensen, B., Mackworth-Young, C. R. S., Simwinga, M., Abdelmagid, N., Banda, J., Mavodza, C., Doyle, A. M., Bonell, C., & Weiss, H. A. (2021). Remote data collection for public health research in a COVID-19 era: ethical implications, challenges and opportunities. *Health Policy and Planning*, *36*(3), 360–368. https://doi.org/10.1093/heapol/czaa158

[6] *Ethics and COVID-19*. (n.d.). Who.int. Retrieved May 12, 2023, from https://www.who.int/teams/health-ethics-governance/diseases/covid-19

time and energy. Further, it is important to note that different demographics of people received varying levels of support from the state, affecting the severity of their experience with the pandemic. Given that part of this project investigates death outcomes and the impact of race on this, it is important not to interpret race as having inherently caused a different COVID-19 experience but also how race biased the state and other actors in their support for them. Results should be interpreted potentially as a lack of resources and a history of biased medical support rather than as something that was purely in the hands of the patient themselves.

Ethical considerations

What kind of underlying historical or societal biases might your data contain? How can this bias be mitigated?

Data used is inherently biased because transit stops and population density, for instance, are not factors that are neutral, yet our investigation treats these as neutral variables. A historic lack of funding and resources towards counties, racism or other factors certainly determined many of the variables we use at face-value. One way to tackle this in the future is to further investigate the causes of disparities between races and the causes for certain counties having higher transit stop densities than others. Only then can more complete conclusions be drawn about the factors that affect COVID-19 contraction and severity.

How could an individual or particular community's privacy be affected by the aggregation or analysis of your data?

COVID-19 patients certainly did not provide explicit permission to be analyzed as part of this project; the assumption, perhaps erroneously, is that the government's open-source data is free for all to use. Even though identifying information like name and contact are not fields in the databases employed, the data in composite does reveal access and demographic differences in counties, causing privacy concerns.

Is data being used in a manner agreed to by the individuals who provided the data?

Patients experiencing intense effects from COVID-19 are unlikely to have had the time to agree to have their data be used for research. Many permissions were overlooked during the pandemic in the name of public health—extenuating circumstances call for extenuating measures—yet researchers like ourselves must be mindful that there are real individuals behind the data.

What are possible misinterpretations or misuses of your project results and what can be done to prevent them?

One possible misinterpretation is that what is true in New England is universal; this is not the case. While race and ethnicity are certainly variables that affect health outcomes in New England, perhaps caste is a more appropriate variable to investigate in a country like India where caste has historically been the ascriptive identifier that has been used as a way to discriminate.

Works Cited

Ethics and COVID-19. (n.d.). Who.int. Retrieved May 12, 2023, from https://www.who.int/teams/health-ethics-governance/diseases/covid-19

Hensen, B., Mackworth-Young, C. R. S., Simwinga, M., Abdelmagid, N., Banda, J., Mavodza, C., Doyle, A. M., Bonell, C., & Weiss, H. A. (2021). Remote data collection for public health research in a COVID-19 era: ethical implications, challenges and opportunities. *Health Policy and Planning*, *36*(3), 360–368. https://doi.org/10.1093/heapol/czaa158

Brown, D., & Toh, A. (2020, May 13). *Covid-19 apps pose serious human rights risks*. Human Rights Watch. https://www.hrw.org/news/2020/05/13/covid-19-apps-pose-serious-human-rights-risks?gclid=CjwKCAjwx_eiBhBGEiwA15gLN77kdw0qSuk75S602d2ZGwHds0-TEyP5WRXwlQfDy1r2c3Ku6yB4FxoC_-wQAvD_BwE