Urban Virus: An exploration of the effects of public transit density, race and age on COVID-19 transmission and severity

Zain Asghar, Soumya Karwa, Suyash Kothari and Udayveer Sodhi

Brown University Data Science
2023

Abstract

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[1], population density[2] and transit stops[3] to determine the impact of transit stop density, race and age on COVID-19 cases. 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. To determine the existence of a significant relationship between COVID-19 case density and public transit stop density, 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. Statistical and machine learning approaches alike showed no significant relationship between transit stop density and COVID-19. For the final hypothesis in question–that there is a significant relationship between race and COVID-19–both statistical and machine learning approaches confirmed initial beliefs. 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. 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.

---

[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.