Geography of the Pandemic: How COVID-19 Spreads in the US

INTRODUCTION

Since the early stages of the pandemic, COVID-19 has spread to all corners of the world and has affected the U.S. in an especially bad way. It is important to determine what factors are contributing to the spread of COVID-19. Are some geographical factors among them? This brings up to the following question I want to explore:

How does geography affect the spread of COVID-19 in the US?

DATA

First, I need to obtain data on COVID-19 daily cases in each U.S. state and choose geographical criteria that can be used to group states by their similarity according to the criteria. Then, I can analyze which geographical criterion is better correlated to daily infection rates numbers. The two criteria I analyze in this poster are *distance* and *urbanization*. Distance may be relevant because people travel more often between states that are closer together, thereby spreading the virus to neighboring states. On the other hand, high urbanization (high proportion of people living in cities) causes more close contacts between people, leading to higher spread rates.

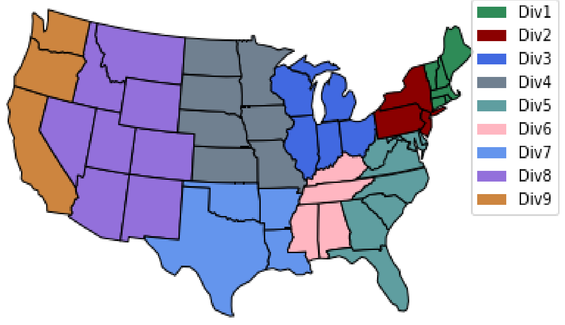
I use the following *data sources*:

For COVID-19 Cases -- data from the *COVID Tracking Project* site, which provides COVID-19 current and historical Data for the US.

I use the following *state grouping types*:

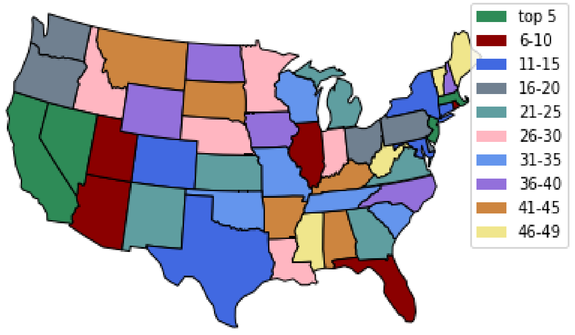
Grouping by Distance – data extracted from Wikipedia's *List of regions of the United States* under Census Bureau-designated regions and divisions. States in the same census division are generally closer compared to states from different ones.

Census Bureau Divisions of the U.S.

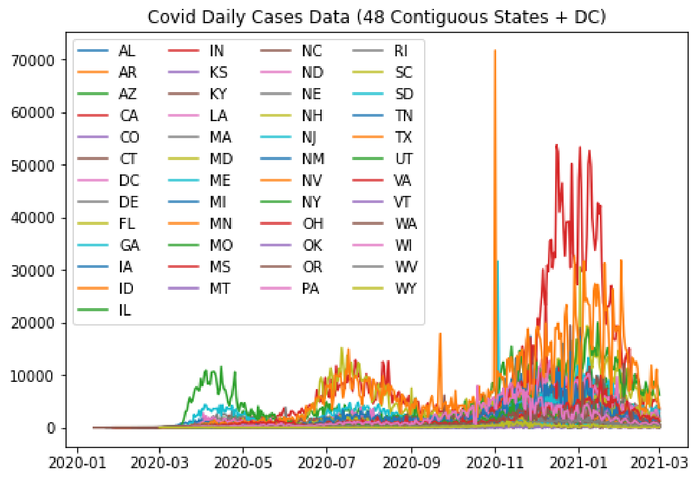


Grouping by Urbanization – data found on *ICIP*'s *Community Indicators Program* site. Shows what percent of a state's population lives in an urbanized area. The states in each urbanization percentile-based division have similar rates of urbanization.

Urbanization-Based Divisions of the U.S. States



ANALYZING DATA



The visualization of the daily COVID-19 data is shown above. Here, I am analyzing the 48 contiguous states as well as Washington D.C.'s data through the period 1/24/2020 – 3/1/2021. To process the data, I use the Python modules: *matplotlib, pandas, cartopy, and scipy.stats*. My analysis for the *two state grouping types* is made up of three steps.

Step 1: In each division, compute the case-number correlations between every two states in that division (top right figure).

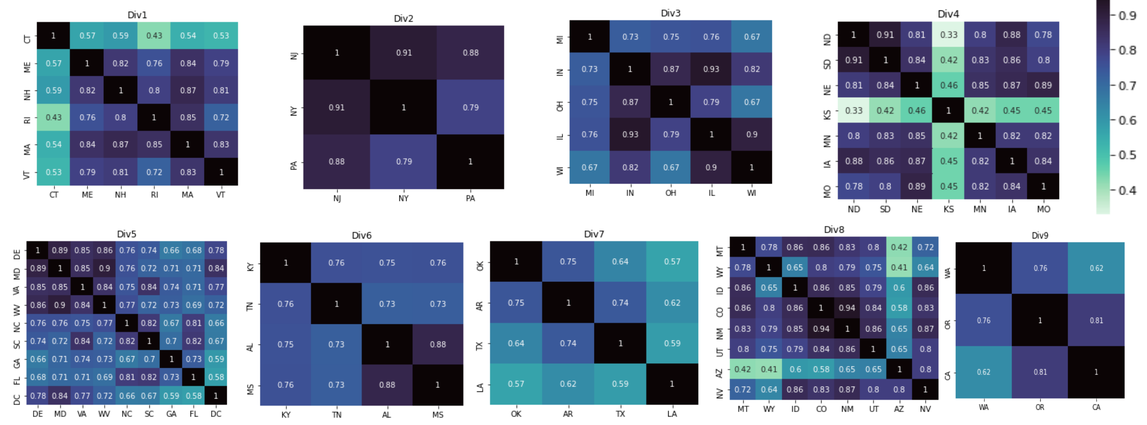
Step 2: For each state, find the mean of each correlation with the other states in its division (the figure to the right).

Step 3: I visualize the numbers found in step 2 into the U.S. map (shown on the figure below).

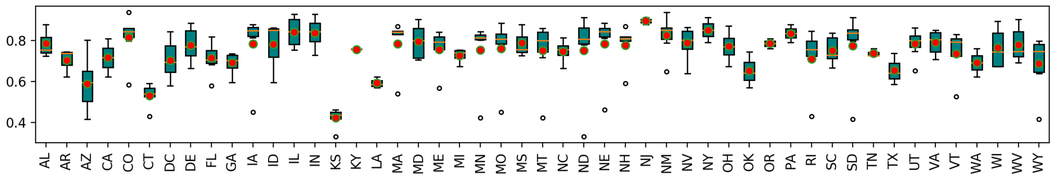
As a "base case" used here for comparison, I also use the same process of three steps without grouping, which means all states are in one group.

Step 4: For an accurate comparison, at the end I visualize the results from Step 2 in a side-by-side bar-plot at the bottom, which shows the average correlation for each grouping type.

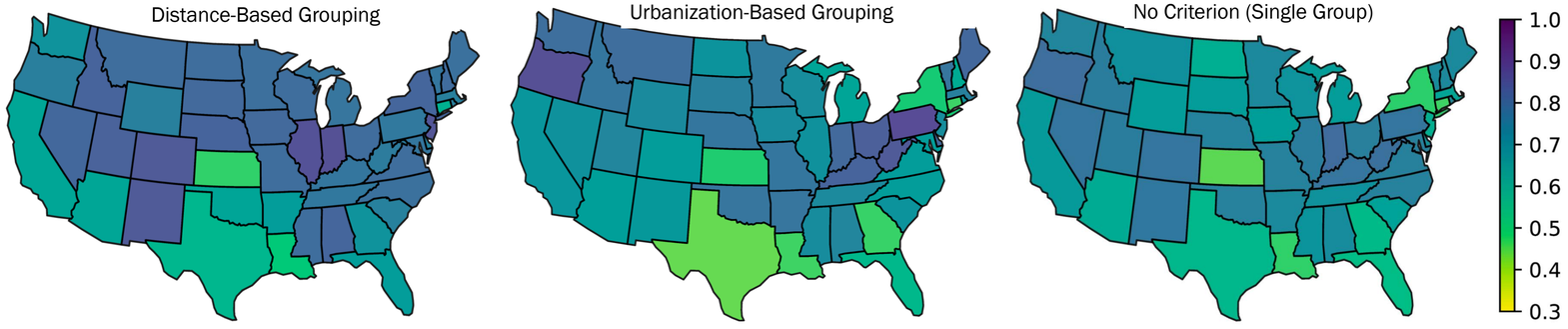
Step 1: Computing In-Division Correlation



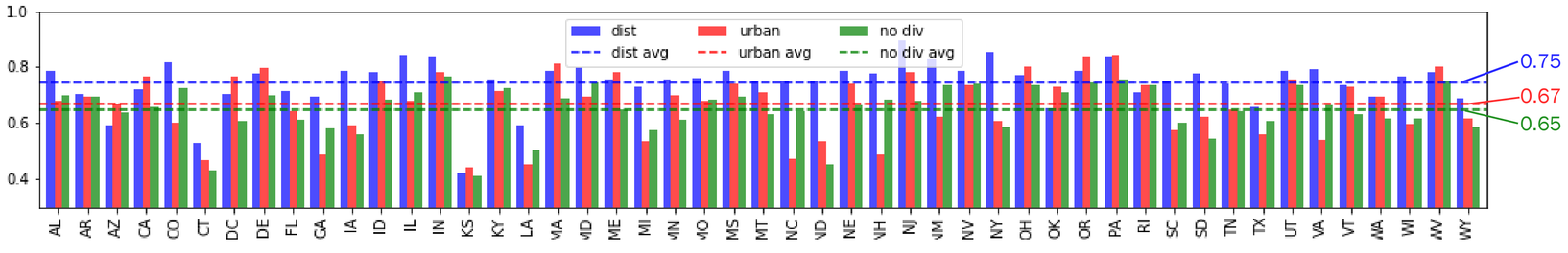
Step 2: Computing the Mean Correlation for Each State (orange dots show the means)



Step 3: Plotting the State Averages on the U.S. Map for the Grouping and the Base Case



Step 4: Side-By-Side Comparison of the State Means (bars) and the Grouping Means (dashed lines)



CONCLUSION

By visually inspecting the maps created in Step 3, it appears that the distance-based grouping has the highest correlation values, while for the other two there is no clear winner. In a different perspective, the bar plot from Step 4 allows me to look at the values side-by-side for each state. The blue bars are highest most of the time, which means that the distance-based correlations are highest, however, it is hard to determine which of the red (urban) and the green (no criterion) grouping is better. In order to make an accurate comparison, I look at the 3 averages. The distance-based one is the best, while the urban-based outperforms the no-criterion one just slightly. I conclude that some geographical features such as geographical distance are clearly correlated to the spread pattern of COVID-19, while for others such as urbanization there is very little gain in correlation compared to the base case. Another observation is that there is a significant correlation (0.65) even in the base case.