Project 1 Group 2 Write up

Michael Sarkis,Xige Huang,Yanqin Shen,Lingyu Zhou

8/31/2021

Background, motivation, objective

At the beginning of our project we wanted to see if socio-economic factors had any influence on the impact Covid. Some original ideas for socio-economic data to use included looking at unemployment, education scores, public assistance percentages and median income. We also looked how to quantify the effect of covid, which could be measured through hospitalizations, deaths or cases. We ultimately decided to use median-income alongside case per capita for our socio-economic analysis due to the availability of quality data for all US counties. We also decided to include political voting information as it was decided that a trend between covid cases and political voting was likely.

The overall objective of our project was to use visualizations to look at Covid-19 case data on a county level and see if there was any connection between median income or democrat voting percentage and the Covid cases per 1000 people.

Data acquiring and cleaning

We obtained our data from many different sources. The county-level Covid-19 cases data came from the New York Times. The median income data was obtained from the U.S. bureau of labor statistics. The urbanization index was gathered from the National Center for Health Statistics and the political voting data was access from Tony McGovern's public github. We chose to use the percentage of people who voted for Democrats in each county to make the data consistent. For further analysis, we also decided to include the region each county belongs based upon their Census region designation.

In order to merge datasets together we used county FIPS code, a 5 digit code used by the federal government to numerically identify counties. Using FIPS code is safer than county and state names because some states have same county names. Unfortunately, county FIPS codes have different formats in different datasets, as there have been multiple editions of the FIPS code system. After standardizing the FIPS code a merge became much easier. However, In the Covid case data, some observations had missing FIPS Codes. Most of these observations have unknown county names and were ultimately ignored. There were also two counties with missing FIPS code: Joplin and Kansas City in Missouri. Since the information of these two counties are also missing in other datasets, we decided to exclude them.

The other issue in the covid case data was that New York City was considered a single county despite consisting of five counties. To fix this decided to combine the information from counties in New York City together in the other datasets. For household median income we just calculated the average.

Finally, we took a 7 day average of covid cases at each of the days we chose for our peaks(July 20th 2020, January 8th 2021 and August 27th 2021) and merged all the datasets into the final dataset.

Visualizations

Scatterplot

The "County View" tab is designed to give customized scatterplots comparing Covid cases to political voting or median income on county level. The visualization allows users to select from 3 time periods, which corresponds to the 3 main peaks in Covid cases in the US that we found through background research. The region and urban index filters also allow the user to customize the visualization to a smaller subset of the data. Furthermore, the user can click and drag on one or multiple points that they are interested in to see more detailed information including county name, state, region, case per 1000 people and the factor they have previously chosen.

Map

The "Map View" tab is designed to give customized map plots on Covid cases, unemployment and median income for each Covid case peak. The median-income and unemployment data are static and do not change between time periods. However, Covid cases do change between time periods and selecting a specific peak displays a map of the 7-day average of Covid cases from the specified date for every county. Blank counties and states in the map are caused by a lack of county or state level data of the variable.

Visualizations Insights

Overall, our visualizations demonstrate a clear negative relationship between cases per 1000 people and median income as well as between cases per 1000 people and percentage of population voted for the Democrats in the 2020 election in all of the three peaks. This means that counties with higher income level or proportion of people supporting the Democratic party tend to have less cases during the pandemic. The relationships are more flat in peak 1 and more negative in peak 3.

Among all the regions, we observe some interesting trends for the South region and the Northeast region regarding the trend between cases per 1000 people and the political factor – percentage of people voted for the Democratic party and that between cases and the socioeconomic factor – median income.

For the South region, some of the counties from the lower- to mid- income group had significantly more cases than counties in other regions in peak 3. This was especially unique because for most counties there were fewer cases in peak 3 than there were in peak 2. Meanwhile higher income counties in the South region did have cases in peak 3. This also corresponds to what we can observe from the map view where cases from peak 3 are concentrated in the southwest part of Florida. Consequently, the slope between cases and median income for the South is much steeper in peak 3 than that in peak 2.

As for the Northeast region, we surprisingly observe a positive slope between cases and median income in peak 2, meaning that counties with higher median income were having larger number of average cases; this was not true for all other regions as they were either essentially flat or negatively sloping. Moreover, the slope between cases and percentage voted for the Democrats is also almost completely flat for the Northeast Region in peak 2.

Conclusion

This visualization can show a wide range of information about county level covid cases. The insights in the previous section are only a few of the many interesting correlations and trends that can be seen from this app.