Yifan Lu

Dr. Dalzell

STA 381

December 8, 2022

Final Write up on Heat Maps

**Section 1: Goal**

The part of the project that I was in charge of is to create heat maps that looks at both Trial rate percent convictions and Total Trial percent dispositions from 2010-2021. There are two main purposes of composing these maps. First is by request of our client, to look at the geographical distributions of the variables of interests across all counties in North Carolina, and how does the distribution shifts across time. The second purpose is to determine whether there's a specific year where the distribution seemed as an anomaly that we should remove it in modeling phase in order to gain better modeling accuracy. To achieve these goals, I have created a heat map for each variable in each year, and comparing them horizontally can help us determine the general trend and identify the outlier, if any.

**Section 2: Data**

The data that I'm using come from holder and holder2, which are data of Trial rate percent convictions and Total Trial percent dispositions, respectively. In each data set, there are three columns, representing the year, county, and the percentage count of the variable of interest. Both holder and holder2 come from nostateNoNA file where the data consist of year, county, variables of interest, but also all the different charges. A few lines of codes were used so that for each year each county, we are looking at the total rate for all charges, rather than each individual charge. More will be explained the the next section.

**Section 3: Data Cleaning, Processing, and Creation**

The creation of holder and holder2 was described in the previous section. However, those data needs to be further processed before actually inputting into the map. For each holder, what I did was download them onto my local space, then use a "filter" in Excel. To do so, select the cells on the first row, which are the names of the columns, then hit from the tool bar there is an option called “Sort & Filter”, select “Filter” from the drop down menu. For each cell filtered, there will be a small triangle on the side which gives a pop up menu. Then for year, select one time interval each and save the other columns into another file. This will provide us with a data file that consists only two columns: county name and percentage count for variable of interest. Repeat such process for each year, and we will get a list of data sets for each variable per year. Detailed annotated steps are included in the appendix.

**Section 4: Heat Map**

Constructing the heat maps is actually the easier part. Once we figured out which programming package to use, all we need to do is to import the data that we have prepared for into the program and set a scale.

One final step before importing the data is to change the name of each columns so that we can combine the data we generated with the data from R packages. For example, we need to change column name from "County" to "subregion", and all county names to their lower cases. In this way we can perfectly merge our data with the ones provided in R package. Once we specify which state we want the map to be, and with the data we imported, we can already start to draw the graph. One thing to note is that, since our goal is to compare the graphs, when we produce each map, we have to make sure that the color scales are unified. By which I mean that for a certain percentage count in our variable of interest, the color for that value needs to be the same across all maps, otherwise we cannot see the change in color in a single county over time.

For each individual year's map, I set the color scale with increment of 0.03 and a maximum of 0.15. The disadvantage of choosing such a scale is that if the percentage goes over 0.15, the county's color will be grey. Fortunately, there are not that many occasions, and setting a lower scale will help us see a larger color difference between counties. The maps are shown below (maps’ implications will be discussed in the next section):

Maps for Trial Rate percent convictions:

| 1 | 2 |
| --- | --- |
| 3 | 4 |
| 5 | 6 |
| 7 | 8 |
| 9 | 10 |
| 11 |  |

Maps for trial percent dispositions:

| 1 | 2 |
| --- | --- |
| 3 | 4 |
| 5 | 6 |
| 7 | 8 |
| 9 | 10 |
| 11 |  |

Finally, after generating heat maps for each individual year, we are also generating a heat map that represents the sum of rate across all years. We could have also used a map showing weighted average for each county across those years, but that will require additional data collection and cleaning, while the sum of all years can elaborate just as much, but much simpler to do.

| Total | total |
| --- | --- |

**Section 5: Conclusions**

Recall that the primary goal for generating these heat maps is to look for outlier years that we might need to consider removing in the modeling phase. Looking at the 11 heat maps for each variable of interest, we can see a general trend, where the western and middle part of North Carolina tend to have higher trial rate percent convictions and trial percent dispositions. This trend is true across all times apart from year 2020-2021. Though COVID year tend to be the outlier, but if we look closer at the color differences between regions, we can see that even during the pandemic, the two rates still follow the same trend. There exists a overall decrease in numbers, which is why we may think the color for west and middle have changed, but there is still a slight difference that proves the trend to be real. Especially given the rebound in the following year, we concluded that such an issue should not be worried, and year 2020-2021 should still be included within the data frame.

**Section 6: Next Steps**

Since my main focus are heat maps, there are not a lot of room for this to be changed. If any, I suggest to look at two approaches. First is change the color scales of the maps so that each county can be assigned with a color, while the color contrast remains high enough for a trend to be spotted. There should be a critical value for this to happen, but it requires trial and error to find it. Second is to collect more data from other years in order to make sure my conclusion about trends across time is valid.

**Appendix - write up for heat maps**

Filtering through holder and holder2

|  | Select “Sort & Filter” |
| --- | --- |
|  | Select a single year and apply filter |
|  | Copy the last two columns into a new excel file to form the data we need for mapping |