CS 216 Project Prototype

Group Members: Nhu Do (ntd6), Raksha Doddabele (rsd35), Pierce Forte (phf7), Joe Cusano (jgc28), Caleb Sanford (cis12)

Introduction and Research Questions

In our project, we plan to examine the correlation between economic health and drug abuse within North Carolina. Since last time, we have decided to focus on counties in North Carolina. North Carolina has been cited as a state with "one of the biggest drug problems", having the 10th highest number of deaths due to drug use in 2019, according to The News & Observer (Specht, 2019). Additionally, it is a state which we all call home for a large portion of the year, which makes it more interesting for use to analyze its data. According to the World Health Organization, low socioeconomic status is a significant risk factor for opioid overdose (2018). We will quantify economic health using data about GDP per capita, unemployment, income, and type of area (metropolitan vs. non- metropolitan) . Additionally, we will quantify drug abuse using CDC drug poisoning data. With these different sources of information, our goal is to identify areas that are most affected by drug use during different economic conditions. We hope to use this information to make predictions for this year (2020) regarding drug abuse and overdoses in regions based on economic data.

While our project proposal offered two primary research questions, upon the suggestion of our mentor, we have chosen one question to focus on for the purposes of this prototype.

Research question

1. Can we predict drug use based on region and economic data?
    a. One of our primary goals in this project is to use the information we collect to make predictions about other regions in the future. We will make predictions for the various counties in North Carolina. The methodology we develop and use will hopefully be able to applied to other states.
    b. We feel that these findings would be valuable to a broader community because they could allow regions to work toward preventing increases in drug abuse before they occur. For example, if it has been shown that in a given region, residents tend to drastically increase their drug use during a recession, and a recession seems to be approaching, the community could make an effort to target the drug abuse before it worsens.

We will pursue our other research question (How do the people in different types of regions – for example, rural and urban – change their use of drugs as economic conditions change?) as time allows after thoroughly exploring our first question.

Preliminary Results and Methods

Thus far in our project, we have successfully found and linked three data sets containing drug poisoning, median income, unemployment, and GDP data organized by county from 2007-2017. We linked these datasets by merging rows based on shared county codes (FIPS) and years. During this step, data wrangling was the most difficult part, as much of the data was previously organized in a manner that did not allow simple merging based on a column as we have done in class before. We had to convert the county unemployment data set from long to wide format to merge it with the other data sets. We also

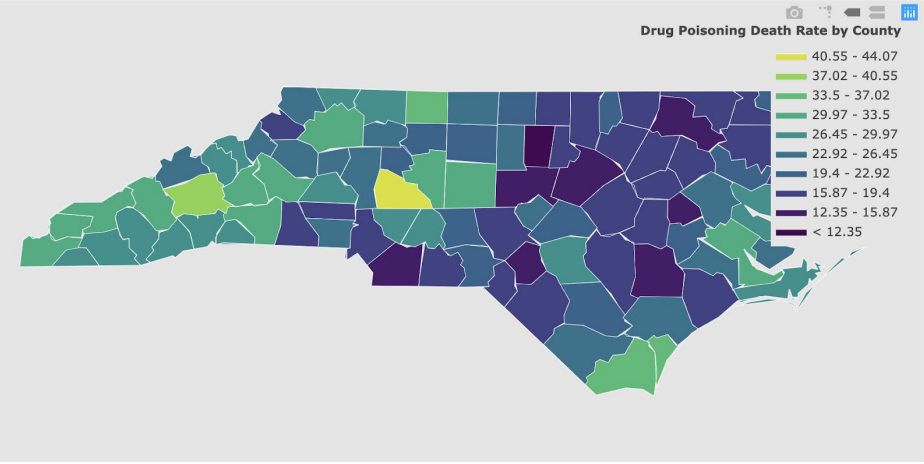
created a column to designate whether a county was considered metropolitan or not using one hot encoding to facilitate easier prediction later on. Part of our resulting data frame is shown below and our code can be downloaded here: https://github.com/calebsanfo/CS216-Final.

	FIPS	Year	GDP	State	FIPS State	County	Population	Model-based Death Rate	Standard Deviation	Lower Confidence Limit	Upper Confidence Limit	Metro Code	Unemployment Rate	Median Income
0	37001	2007	5306622	North Carolina	37	Alamance County, NC	144,712	10.705740	1.778055	7.693931	14.650021	1	5.1	$50,480.00
1	37003	2007	795721	North Carolina	37	Alexander County, NC	36,609	18.836393	3.640613	12.904131	27.130432	1	5.2	$49,138.00
2	37005	2007	319131	North Carolina	37	Alleghany County, NC	11,061	17.426167	3.830240	11.439603	26.384358	0	5.2	$39,735.00
3	37007	2007	762762	North Carolina	37	Anson County, NC	26,664	11.725714	2.486966	7.789710	17.498755	0	7.3	$38,023.00
4	37009	2007	709296	North Carolina	37	Ashe County, NC	26,506	16.769407	3.419842	11.287649	24.643601	0	5.0	$41,864.00

Another difficulty in creating this data frame is that much of the data we have been able to find online cover different years. As such, it was difficult to find a period of years in which several data sets overlapped so we could combine them. Moreover, it was challenging to find economic or drug data which is organized by county, as most data is only organized by state or region type (such as urban or suburban) due to confidentiality concerns.

After creating our dataframe, we worked on creating initial exploratory visualizations to see if they showed any strong correlations between different variables in the data frame. First, we created several maps showing population, GDP, unemployment rate, and Model-based Death Rate by county for 2017. We created these maps because we thought they might help us visualize general correlations between drug poisoning death rates and economic factors by county. A map showing model-based drug poisoning death rate by county is shown below.

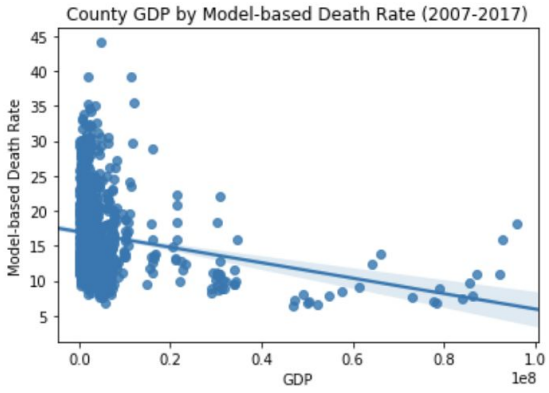
After creating the maps we created several scatterplots with linear regression model fits to try and



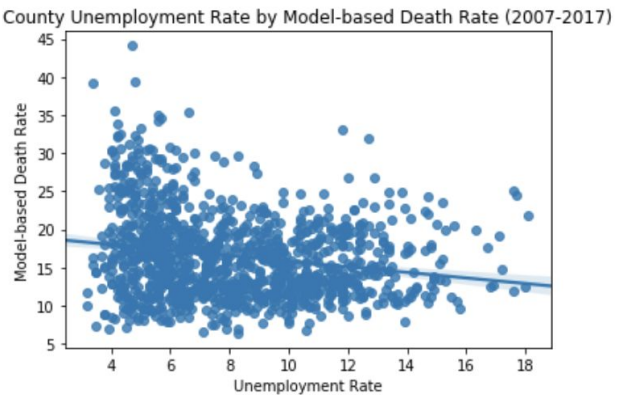
visualize correlations between economic factors and drug poisoning death rates. We also calculated the r-squared values between unemployment and death rates and between GDP and death rates. The scatterplots and r-squared values all pointed to a fairly weak correlation between economic factors and drug poisoning death rates by county in North Carolina. The r-squared values were less than 0.05. A plot

of drug poisoning death rates compared to GDP and a plot of drug poisoning death rates compared to unemployment rate are shown below along with the r-squared value.

```
r_squared: 0.046090872491197366
```




```
r_squared: 0.03913324160285298
```



We tried performing some log transformations on the data to see if that revealed a greater correlation, but doing so only increased the r squared values slightly. The weak correlations suggest that it may be very difficult to create a linear predictive model which uses economic factors to predict drug poisoning deaths.

Reflection and Next Steps

Thus far, it has been particularly challenging to find relevant data with overlapping years in regard to our topic. As such, we have had fewer factors to compare in conducting our analysis, which has made it difficult to identify strong correlations among the data, particularly between economic conditions and drug use.

Despite these issues, we have successfully created numerous figures and maps to portray our findings. However, these findings suggest a weak correlation between economic factors and drug poisoning deaths in North Carolina counties. We will search for more data which may show a stronger correlation and try to transform our current data in ways that may reveal a stronger correlation.

Moving forward, we would like to find additional data regarding drug use in North Carolina's counties. Due to the difficulties with confidentiality mentioned above, we were only able to find one dataset containing drug-related deaths; however, if we are able to find more data – particularly data that includes not only drug-related deaths but also drug use – perhaps we will be able to identify stronger correlations among the data.

Additionally, given that our second research question focuses primarily on region type (like urban or suburban) and more data was available with this information, we would like to take advantage of these additional resources and explore them further, hopefully allowing us to identify new correlations.

Moving forward, we want to see if we can transform our current data or find new data which will reveal greater correlations between drug use and economic factors. If we can find a greater linear correlation, we will create a linear predictive model which will use economic factors to predict drug use. If we can not find a linear correlation, we will see if using a non-linear model, such as a neural network,

will be helpful. If that does not work, we will see if we can make a model which uses other information, such as education or crime, to predict drug use.

Datasets Used

https://fred.stlouisfed.org/release/tables?rid=397&eid=1079693
https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/
https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/

Git Repo

https://github.com/calebsanfo/CS216-Final

Citations

1. Specht, Paul A. "Fact Check: Is North Carolina No. 2 in the Nation in Drug Overdose Deaths?" *Newsobserver*, Raleigh News & Observer, 11 Apr. 2019, www.newsobserver.com/news/politics-government/article229118439.html.
2. World Health Organization. "Information Sheet on Opioid Overdose" *Management of Substance Abuse*, World Health Organization, Aug. 2018, https://www.who.int/substance_abuse/information-sheet/en/