CS 216 Final Project

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

Introduction and Research Questions

We examined the correlation between economic health and drug abuse within North Carolina and extrapolated those methods to the broader United States of America. In our prototype, we focused 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 quantified economic health using data about GDP, unemployment, income, and type of area (metropolitan vs. non- metropolitan) . Additionally, we quantified drug abuse using CDC drug poisoning data. With these different sources of information, our goal was to create a model that predicted drug poisoning deaths in counties based on economic data.

Based on our initial findings with data from the North Carolina dataset, we started to use similar methods to explore national data, under the direction of our mentor. Reasons for this pivot will be discussed below. Exploring national data allowed us to see if our previous methods are scalable and accurate. Additionally, it allowed for deeper data analysis and more accurate predictive modelling due to a greater amount of useful, available data.

Research Questions

1. Can we predict drug overdose rates in counties in North Carolina based on economic data?
    a. We feel that these findings would be valuable to a broader community because they would allow officials to use economic data to identify and aid counties which are at risk of high drug use.
2. Can we build an accurate prediction model for nationwide drug overdose data based on data of other risk factors?
    a. This question will explore national data, using incarceration rate, unemployment rate, and median income to predict death rates per state.
    b. Successfully completing this question means that we will be able to create a tool that can be used to predict future death rates. This model can be a tool to help mitigate these death rates and prepare in the future for states at risk of increased drug deaths.

Results

We were able to create a predictive model for drug use based on economic data in North Carolina. However, the accuracy of our model was limited by the dearth of economic and drug data available at the county level in North Carolina. We discovered that the economic data we did find was fairly weakly correlated to drug poisoning deaths. There was also a lack of education, mental health, and incarceration data for counties in North Carolina to try to add to our model. If more data is made publicly available for counties in North Carolina in the future, it may be possible to create a more robust and accurate predictive model.

Additionally, we created a prediction model for drug related deaths based on 16 years of statewide data including: population, overdose rate, unemployment rate, median income, prisoner count and the incarceration rate. The model had an average error of 18.13 deaths per year. With an average of 714 deaths related to drug overdoses per state per year our error consists of only 2.5% the number of average deaths. Therefore, this model can successfully estimate the number of drug related deaths based on available state data.

Visualizations are included in the Methods section.

Methods

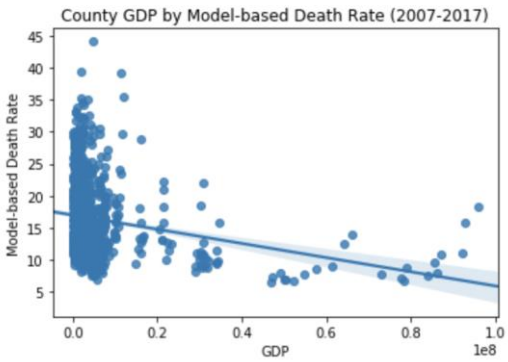
Creation of Predictive Model for North Carolina

We started our initial analysis by merging and wrangling datasets that contained specific information about drug poisoning, median income, unemployment, and GDP data organized by counties in North Carolina. We linked these datasets by merging rows based on shared county codes (FIPS) and years.

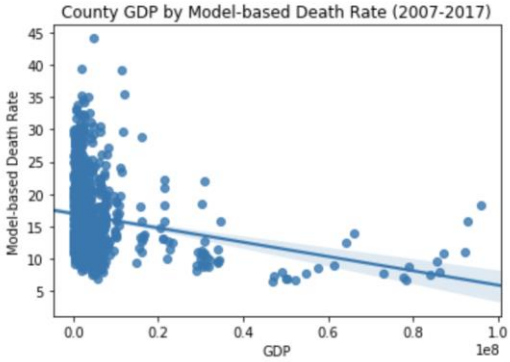
Upon completion of the dataframe, we worked on creating exploratory visualizations to see if they showed any strong correlations between different variables in the data frame. This resulted in several maps showing population, GDP, unemployment rate, and Model-based Death Rate by county for 2017. We hoped that these maps might help us visualize general correlations between drug poisoning death rates and economic factors by county, however, with the results of this analysis, we decided to pivot our methods.

After attempting to create a linear regression model predicting drug poisoning rates in counties in North Carolina using economic data, we discovered that a lot of the economic indicators we were using showed a very weak correlation to drug poisoning rates based on exploratory data analysis using regression plots made with seaborn and r-squared calculations. A few of these plots and their corresponding r-squared values are shown below.

r_squared: 0.046090872491197366



r_squared: 0.046090872491197366



Based on the initial analysis, we questioned whether a linear regression model would be effective using the data we had thus far collected for counties in North Carolina. For our final version, we attempted to find more economic and drug data and also data about incarceration rates, education, and mental health to incorporate in our model. We hoped adding more data would allow us to create a more robust predictive model. However, we were unable to find additional usable data for counties in North Carolina. There was little data available which focused on counties. The data we found for North Carolina counties covered only a short number of years and many of the data sets did not overlap in terms of the years covered. Thus, we decided to create a linear model for counties with the data we had found previously and to create another predictive model at a nationwide state-level using new data. We decided to make the new state-level model because there was much more data for states available and the data covered a larger range of years.

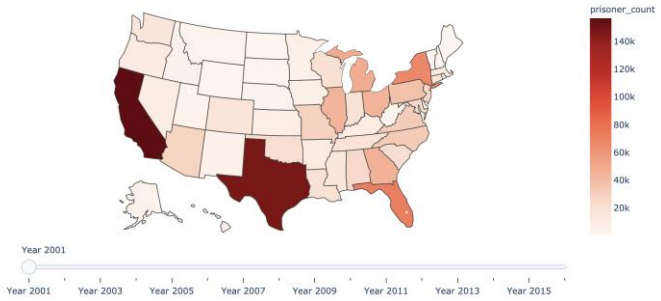
To create our predictive model for counties in North Carolina, we created a linear regression model with sklearn. We then performed backwards elimination on this model to eliminate features that were making it less accurate, which gave us our final model. Using this model, we were able to achieve a mean squared error of 25.69 for the model-based death rate which ranged from 6.42 to 44.07 in our dataset. Thus, as expected, the mean squared error was fairly high for our model and the model was not very effective given the limited data we had access to. We then moved on to the creation of new datasets for state-level data and the creation of a predictive model using these data sets.

Creation of Predictive Model for States

For the state data, we used the CDC's drug poisoning data by state. The incarceration data was accessed from Kaggle, but since it contained only the number of incarcerated people and total state population, we used those columns to create a new column with incarceration rates. We wrangled the data of median income and unemployment rate by state to keep only the information required for our analysis. The median income and unemployment rate datasets were in wide format while the drug overdose rate and incarceration datasets were in long format, so

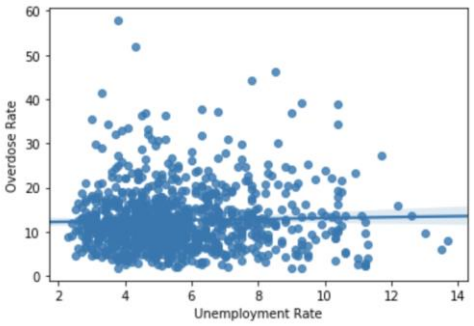
we had to convert the economic data into long format to merge them into one working dataframe with only the information necessary for the predictive modelling.

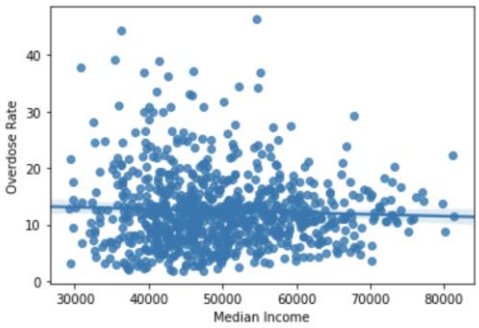
We began working with these newly cleaned and merged datasets through exploratory data analysis. With the Plotly library, we created US maps of different features in our datasets, allowing us to see the statistics in each state with a slider for the year. An example of one of these plots is shown below.

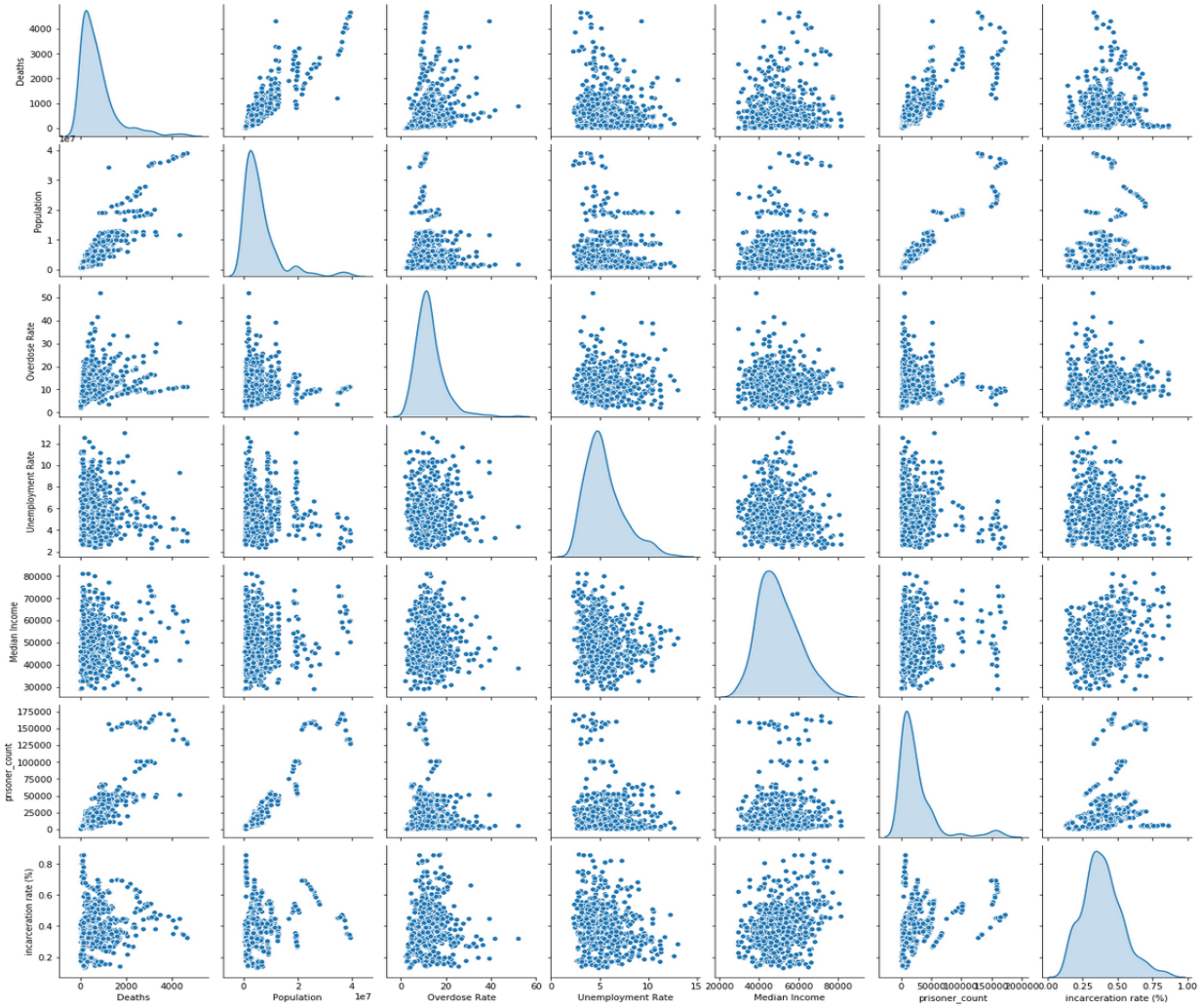


Using these visualizations, we were to gain a sense of how economic conditions, drug use, incarceration, education completion rates, and population have changed over recent decades. Likewise, we made basic notes and predictions about their correlations, and, perhaps more importantly, we identified which of these findings was most notable and thought about ways in which we could learn more.

Next, we started analyzing the data with more concrete methods, the first of which being linear regression. We created regression plots for several variables in our dataset to look for strong correlations. Plots of overdose rates vs. unemployment rate, overdose rates vs. median income and the joint probability distribution for all of the data are shown below.



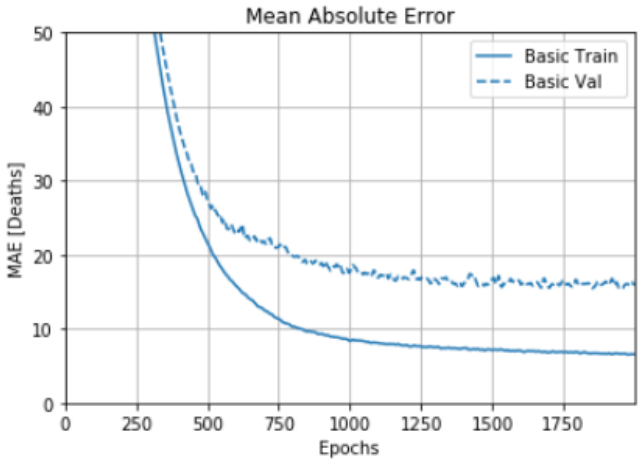




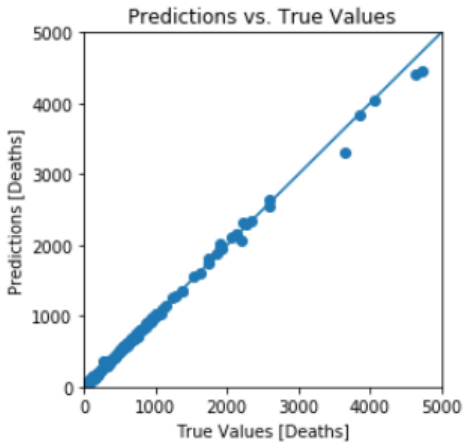
As shown in the plots above, there are not many strong correlations between pairs of variables. Given the great range of our data and the many different figures it included, we were sure that outliers were present and that the data needed to be normalized. Likewise, we felt that more advanced methods were necessary. The next steps we took aimed to address these issues.

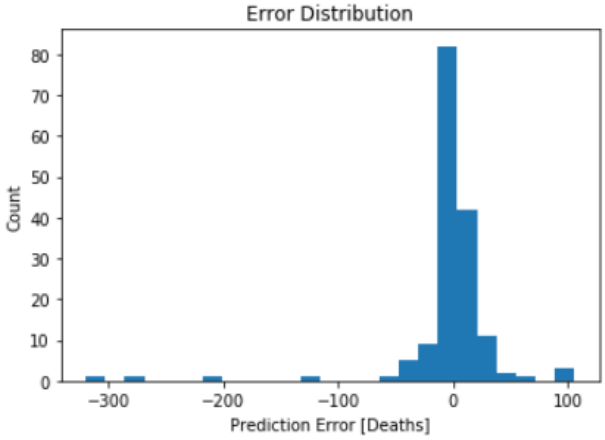
We used a tensorflow model to predict the statewide number of deaths due to overdoses in a given year. This sequential model consisted of two densely connected hidden layers and an output layer that returns a single value representing the predicted number of deaths. To train the model, we split our data up into a training dataset and a testing dataset. The training data was then further divided into a train set and a validation set used by tensorflow. We trained this model for 2000 epochs, achieving an average error of 6.59 Deaths. A plot of the Mean Absolute Error per Epoch is shown below.

After training, we tested our model on the unseen test dataset. We achieved a mean absolute error of 18.13 deaths running the model on the testing set. The plots below show the performance of the test data. The figure on the left shows the predicted number of deaths vs the



actual number of deaths, the plotted line shows a perfect prediction. The figure on the right shows the error distribution of the predictions. This graph clearly shows that the majority of predictions are very accurate but the model also predicted major outliers.





Discussion

With our project, we were able to create a predictive model for North Carolina and the broader US, but we wonder if we would be able to create a much more accurate model if we had more economic, education, or mental health data available.

As explained above, the biggest challenge that we encountered since the prototype has been developing a comprehensive model that would best show some type of relationship with the original North Carolina dataset that we created in the prototyping stage of this project. Due to the weak correlations our initial analysis showed, we decided to perform multiple linear regression using sklearn and see if we could create a more accurate model.

We proved to be able to do so. While the new model we created l is able to predict accurate information, it will probably not be able to predict correctly far into the future. This lack

of scalability is due to limitations on the amount of data and the small time period in which our data is from.

Although our initial explorations of the North Carolina data did not lead to any significant correlations or findings, that does not mean that there is not a strong relationship between any of the variables. Upon the gathering of more robust data from a larger time span, more relationships may arise. Additionally, there are many factors that our initial data may have missed that are risk factors for drug use. Looking for data on factors like level of education and mental health history may prove to be beneficial in data analysis. While mental health data would be incredibly helpful in understanding trends and creating predictive models, that data is difficult to obtain due to privacy laws.

One of the most prominent limitations of our study was the fact that the statistics we focused on (drug use and mental health issues) are commonly underreported. Because these statistics rely on self reporting, the data can be incomplete and inaccurate. As such, this lack of comprehensive data inevitably meant that our models could not be fully accurate; however, we aimed to address these issues by performing data cleansing and normalization.

The primary way in which our results could be improved or extended is by obtaining more data to conduct our analysis. Examples of datasets that would greatly aid in the creation of a more effective predictive model are longitudinal educational attainment and mental health history datasets. According to our research, these two variables are risk factors for drug abuse, therefore might offer interesting data. Educational attainment datasets are readily available, however the data points are reported decadely rather than yearly, which makes it difficult to incorporate into our model. Mental health data on the other hand, is very inconsistent. The datasets we found online were mostly self reported, presenting many issues such as bias and lack of data samples. The more robust data was blocked due to privacy issues. Notably, the Substance Abuse and Mental Health Data Archive (SAMHDA) maintains extensive datasets on drug use, mental health, and economic conditions, all of which is at the county level for the entirety of the United States. To study this data would have likely allowed for fascinating results; however, given confidentiality issues, we were not able to gain access.

Through these efforts, we would have three focuses during our search for data: the first being to find the most accurate and specific data possible, the second being to find data over many years to create a stronger model that accounts for cyclical changes in the economy, and the third being many different datasets that would allow us to select different features in our model to make it as accurate as possible.

Assuming we were able to obtain more county level data, we would like to combine the two scopes of our research efforts – the first being county-level in North Carolina, and the second being state-level nationwide – by exploring county-level data nationwide. This work could allow us to compare the state by state analysis we have already conducted with the

information compiled by county-level analysis of these states, perhaps presenting inconsistencies in the data. And, similarly, we might be able to create a more accurate model, identify regions that are particularly susceptible to dangerous drug use conditions as a result of certain economic conditions, and utilize our methods and results as tools for drug enforcement and healthcare systems.

The initial findings of our project offer a promising look into the future of predictive analytics. We have proven past data can predict outcomes to a certain extent. With more data and complex modeling systems, we may be able to use our foundations to create a comprehensive model that can be used in healthcare to reduce these drug abuse rates in the future.

Link to Git Repo

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

Datasets

Please view the README.md document in the root of the repository linked above to locate our data, analysis, and results.

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/