CS 216 Project Proposal

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 regions in America. We will quantify economic health using data about GDP per capita, unemployment, and income. Additionally, we will quantify drug abuse using data about overdoses, drug-related arrests, and surveys of drug use. 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. Below we detail more specific research questions that offer different focuses for our analysis.

Research questions

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.
    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.
2. How do the people in different types of regions – for example, rural and urban – change their use of drugs as economic conditions change?
    a. The analysis of different regions would allow us to assess which regional qualities make a region more susceptible to increases in drug use as economic conditions vary.
    b. Likewise, we can view how different areas around the country (like Northeast and Southeast) are affected differently and consider the cultural dissimilarities present that may affect these differences.

Data Sources and Collection

In order to carry out our study, we will primarily use three datasets.

1. Products - Vital Statistics Rapid Release - Provisional Drug Overdose Data
    a. To assess regional drug overdose rates, we will be using a CDC database that provides information on provisional drug overdoses. This database provides information on the reported and predicted counts of deaths due to overdoses by state, a mapping of the percent change in deaths due to overdoses in the United States, and a report of specific drugs that led to drug overdoses that occurred nationally and in select areas.
    b. For the purpose of this project and the research questions listed above, we will be thoroughly exploring the dataset that pertains to the reported and predicted counts of deaths per state, since that most directly lends itself to our research goals. The CSV file includes monthly counts of drug overdose deaths in each state from 2015 to 2019. In the event that we use this data to predict future outcomes, an analysis of the mapping of the percent change in deaths within the United States will be incredibly helpful.
    c. We chose this database because the CDC is a reputable organization that takes measures to ensure quality control, such as including footnotes if the data has been underreported. This gives the source increased credibility. Furthermore, this data is updated monthly, ensuring that we are working with the most up to date and trusted data. We will use monthly data from these databases from 2015 to 2019 to answer our research questions, filtering out any missing data or data marked as poor quality.
2. State unemployment rates over the last 10 years, seasonally adjusted
    a. This data comes from the U.S. Bureau of Labor Statistics, which provides the data for the monthly unemployment rate by state over the last 10 years. The unemployment rates are visualized through an interactive map of the United States, which will be helpful as we are focusing our data by state. Defining the meaning of “region” to refer to states in both this dataset and the above allows for consistent data aggregation and analysis.
    b. The interactive map also accounts for seasonal change, which is a confounding factor that may come into play in our research. We are able to obtain unemployment rates that are seasonally adjusted, which allows for more concise analysis.

3. Useful Stats: Per capita GDP by state (2008-2017)
    a. This set provides the data for per capita GDP by state. This information comes from the Bureau of Economic Analysis (BEA) and their 2017 GDP predictions per state. This data is extremely useful as it acts as a metric detailing each state's economic performance in relation to others over time.
    b. In this dataset, the information is visualized through various interactive maps, detailing information on the changes over a 1, 5, and 10 year basis. The nonprofit organization SSTI has aggregated this information into a spreadsheet that we plan to explore in our research.

Methods and Evaluations

To approach this problem, we will look into the relationships between past economic and drug abuse data and use the connections seen there to predict how severe drug abuse might be in the future in a given region. We plan on using python, pandas, and numpy to clean and link the data from the sources above. We will then use a supervised machine learning approach to predict the risk of drug overdoses based on economic data. To do this, we will use pythons sciKit-Learn package or tensorflow.

The success of this model will be evaluated by excluding a testing dataset from the training dataset, then using the model to predict the number of drug overdoses seen in the testing dataset. We will come up with a metric to determine the correctness of the predicted overdoses vs the actual overdoses.

While we have worked to find credible data and have created strategies that account for confounding variables, there are limitations to our research. Economic data is readily available, but the extent to which it is valid is something that is out of our control. Additionally, drug abuse is commonly underreported in many areas. However, by understanding the limitations of the data at hand, we can mitigate the effects of them.