Predicting Geopolitics from Food Security and Public Health

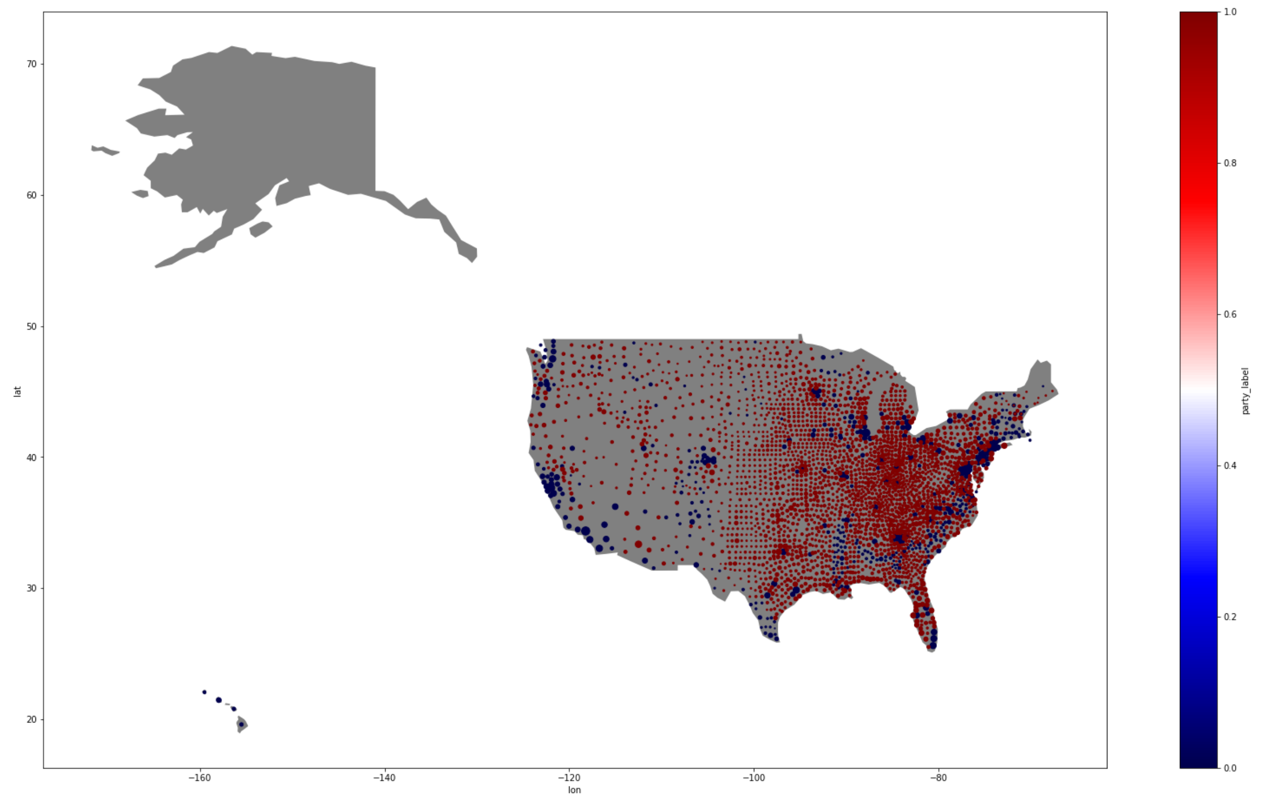
Politics, especially local politics, impact various aspects of life, such as local public health infrastructure and access and proximity to food. We thought it would be interesting to analyze each of these three factors individually (politics, hospitals, and food insecurity) and then see if we could predict the politics of the local community just based on the rating of the hospital system and access to food.

Political Dataset:

Open Elections in the United States Dataset:
This data set is from the OpenElections project and contains primary and general election data broken down by states and counties and includes a master dataset. In our analysis, we used the 2016 presidential election dataset. The columns in the dataset include state, county, office, district, party, candidate, and votes. It is noted that the dataset does not include Alaska because counties work differently in Alaska.

Data cleaning:
Looking at the dataset, the district column values were all NULL so that column was dropped. The office column values were all "president" so the office column was dropped. Next, we constructed the winner_county_df where each row shows the candidate with the most votes in the county. The party column had a variety of values representing the Democratic party and the Republican party. These were standardized so that R represents the Republican party and D represents the Democratic party.

Visualization:
To create the geographical visualization we added the latitude and longitude values for each county and the abbreviation for each state. Each dot in the visualization represents a county and the sizing represents the amount of votes the winning candidate of the county received. Red represents that the Republican candidate won the county and blue represents that the Democratic candidate won the county.



Insights From Map:
This visual may appear to be a little shocking at first; however, given the political context of the United States, it makes sense. Many of the Democratic strongholds are centered around major metropolitan areas which is reflected in the graph. Republican leaning counties are based in more rural areas and thus take up more physical space on the map. Population wise, both parties have similar support, so it is interesting to see the distribution of politics.

Hospital Quality Dataset:
This dataset contains various information about hospital quality across the United States such as overall hospital rating, safety of care, timeliness of care, etc.

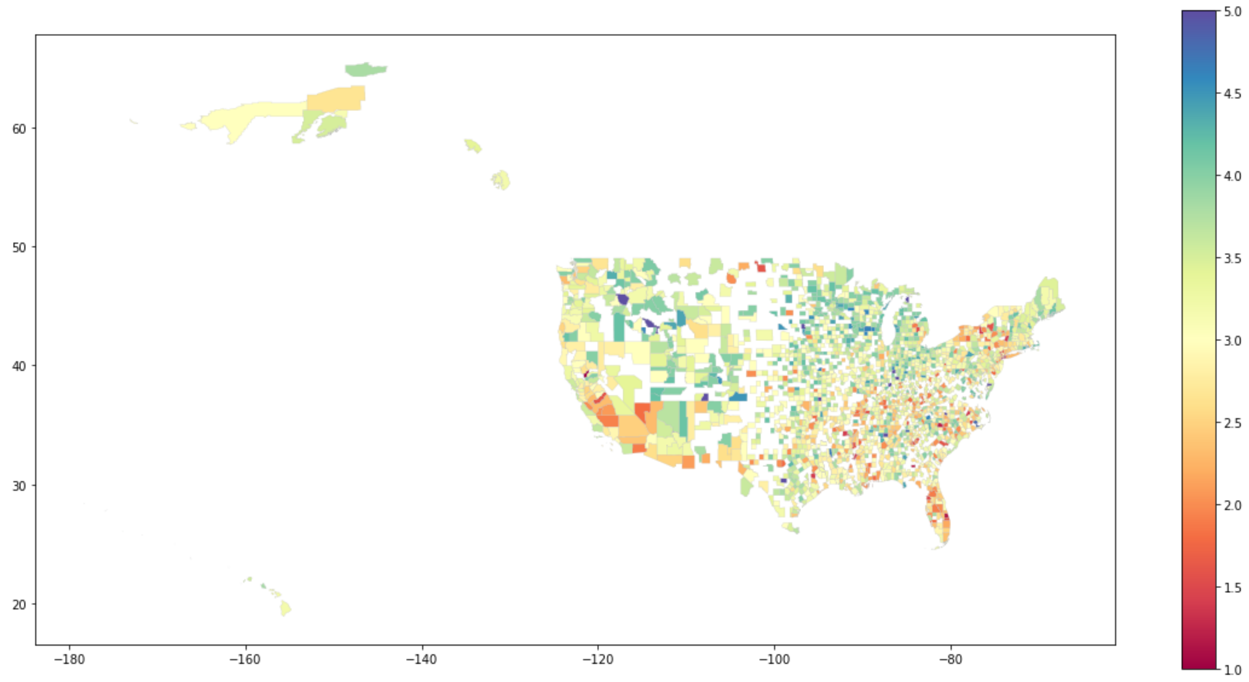
Data Cleaning: To clean up some of the data, we dropped the columns containing the keyword "footnote" because they have a substantial number of NaN values. The column containing "EHR" consists of about 15% NAN values, so we'll encode the NAN values in that column as we would any other value. Additionally, there are only 45 hospitals that do not have a "County Name," so we'll just drop those rows. This does not substantially change our analysis because we are working with more than 25,000 hospitals.

Visualization:

We graphed average hospital overall quality with each of the counties in the United States. Some countries have no hospitals with that data available, and thus they are represented as white on the graph. Red/orange countries have poorer hospital ratings, and green/blue countries have better overall hospital ratings.

Insights From Map:

It is interesting to note that there is no clear discernible pattern in the distribution of hospital ratings across the country. Some loose trends that we noticed are that the northern parts of America (northeast and Pacific Northwest) generally tend to have better quality healthcare. Similarly, those hospital systems down south have more variability as well as lower ratings overall. We were expecting cities to have better healthcare just given proximity to populations and universities. However, it is important to note that these ratings are averages of all the hospitals in an area. The mean, being a non-resistant measure, is skewed by outliers. So a city with a few poorly rated hospitals could have a substantially lower average hospital rating than a rural county with a single but well-rated hospital. Nonetheless, the city could still have better overall access to and quality of healthcare.



Food Deserts in the U.S. Dataset:

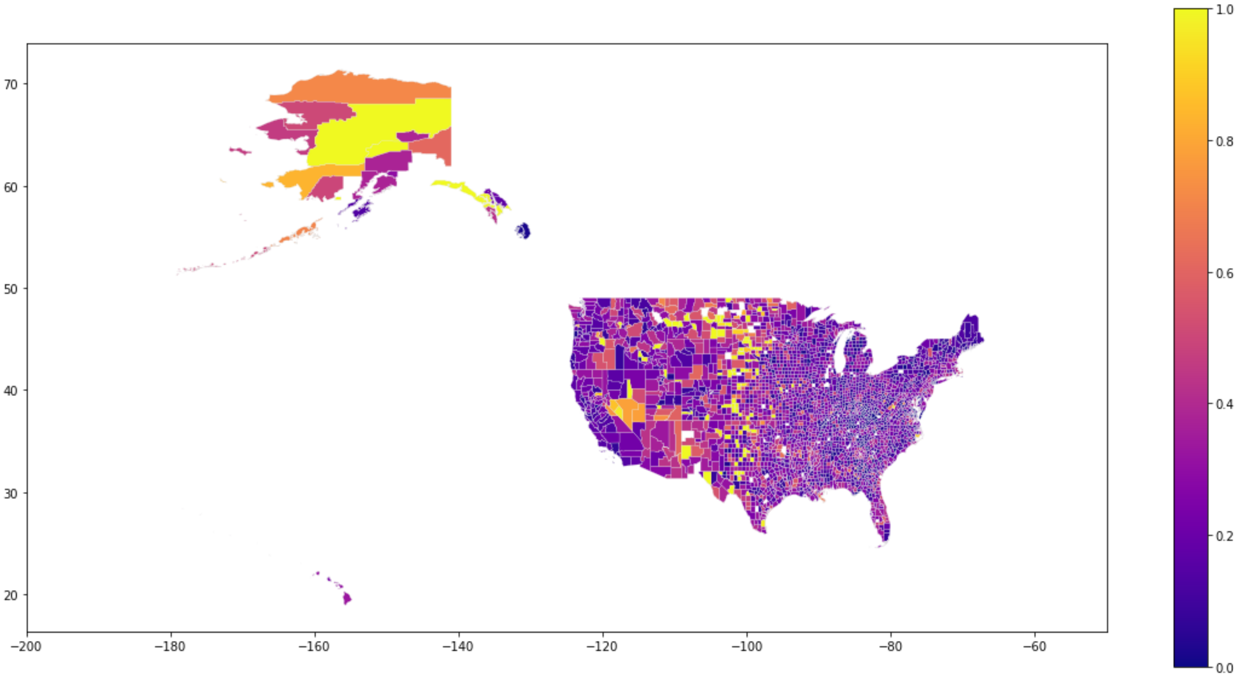
This dataset contains information about different subregions within counties and their access to food in the form of supermarkets along with other demographic information relating to race, ethnicity, and age. There is also information breaking out food insecurity based on certain geographic distances (0.5 or 1 miles for urban areas) and (10 or 20 miles for rural areas), as well as financial information about SNAP benefits and access to vehicles.

Data Cleaning:

This dataset was relatively easy to manipulate and needed no substantial modifications. To aggregate the different subregions within a county, we took the sum for all variables relating to counts which are all of the continuous data. For the data relating to whether or not a country faces food insecurity, we summed up all of the instances of food insecurity within a county and then one hot encoded that information (1 for food insecurity, 0 for not). So, if there was just one instance of food insecurity within a subregion for a given column, it was considered to be food insecurity for the whole county.

Visualization:

Each county is mapped according to the percentage of population facing food insecurity, defined as urban populations not having access to food less than 1 mile away and rural populations not having access to food less than 10 miles away. Blue/purple counties mean there are low instances of food insecurity and orange/yellow counties mean there are higher instances of food insecurity.

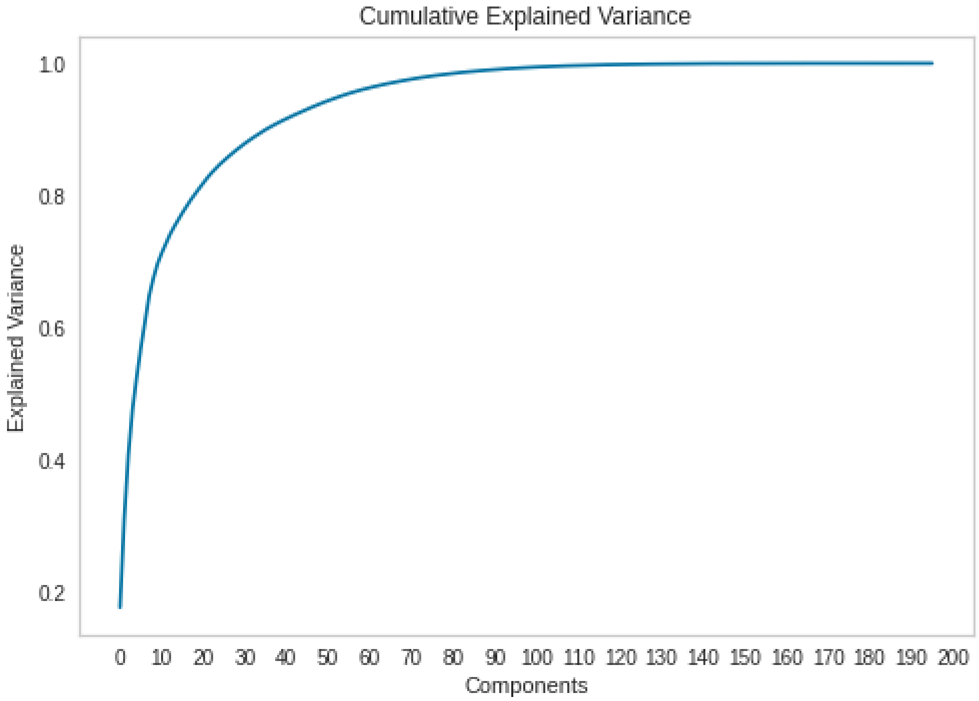


Insight From Map:

The overall trend that we see is that more rural places have greater average rates of food insecurity, which makes sense. Many parts of the midwest as well as Alaska are very remote and there is often only one store for hundreds of miles. Looking at the coastal parts of the US, it is interesting to see that many of the counties are very similar in their levels of access to food. Perhaps, there is more variation within a county itself that needs to be explored further.

Merging Datasets Together

We joined all three datasets based off of the state and county columns. On the joined dataset we ran Principal component analysis (PCA) to reduce the dimensionality of our joined dataset.



Training Models to Predict Election Results

Based on the joined dataset we separated the dataset into features and labels. We tried three different models: logistic regression, neural network and random forest. Without running PCA, the logistic regression resulted in an accuracy of 83.6%. After running PCA at 95% explained variance ratio, the logistic regression rose to 94.1%. For the neural network, we used GridSearch to find the best hyperparameters. Out of the hyperparameters we tested, the best was with hidden layer 1 having 100 nodes, hidden layer 2 having 50 nodes, and hidden layer 3 having 50 nodes with the Relu activation function and a learning rate of .001. This resulted in an accuracy of 93.0%. For our last model, we did a random forest model. Again using GridSearch, we found that the best hyperparameters were a max depth of 70 and 128 estimators. This resulted in an accuracy of 93.3%. These results highlight the power of dimensionality reduction in reducing the noise of the data. Considering the accuracy results, the logistic regression model is the best, but the other model's accuracy results are not far from the logistic regression.

Implications:

Analyzing results like these is always tricky because of the persistence of geography in influencing the model training. Even though we dropped the location information from the datasets when we analyzed them, the relationships between urbanism, food accessibility, and hospital quality are deeply intertwined with social and economic factors that are directly shaped by geography. So yes, our models are clearly able to predict political outcomes based on hospital quality and food accessibility with reasonably high (>90%) accuracy, but there are definitely more factors that are worth considering.

Areas for further analysis:

It was really interesting to see the relationship between hospitals and food insecurity and how that relates to local politics. It would be fascinating to further develop this analysis and look at the education system or local environmental factors and see if predicting election results is possible from that. Moving forward, we would like to expand in several different areas. First, looking at even more specific subregions within counties to further fine tune our analysis and get a more nuanced understanding of local politics. Another area of analysis would be looking internationally, especially in political contexts that are not two party systems. Though it would be a much tricker analysis, it would certainly be a compelling area of research.