The Factors that Contribute to Self-Identity:
Region Data Capstone Report

1. Introduction

A large part of a person's identity derives from how they view their environment. Where they are located and how closely they identify with that area can help when trying to better understand them. This type of information is crucial in marketing. One of the key factors for selling any product, be it a physical item or a political person or body is to know the market in which you are selling, thus knowing the consumer's identity. Marketing campaigns can then be tailored for these consumers that are more inline with how they self-identify.

However, within the United States, outside of the state in which they live, citizens of the US have a further category by which they self-identify: region. This regional distinction comes with certain stereotypes that can help others more easily understand how those from that region may think and what their beliefs might be. Yet, defining where the borders to a particular region are located can be difficult, particularly for the regions of the South and Midwest.

How are such areas marked out? Do most people from that region share common attributes? Do their income or education (or their neighbor's) play a factor in how they self-identify with their region? Within this project, we acquire data taken from an online survey and look into what patterns and trends are shared in respondents who closely identify with being from a particular region.

2. Dataset

The dataset was acquired from the FiveThirtyEight github. They are from surveys that were created with similar questions regarding two regions of the US: the South and the Midwest. People could select how much they identify with being a southerner or midwesterner with the same categorical choices ranging from "A lot", "Some", "Not much", and "Not at all". They could also write in their own answers for how they would self-identify. There was also an area for general collection of information such as zip code, income, education, and which region according to the census bureau each respondent was located. Additionally, for each survey, there was a place that respondents could vote for which states they believed were midwestern or southern, with a different list of states for each of the two surveys. The link for the original dataset can be found here.

3. Data Cleaning

The purpose of cleaning the data was to organize the raw csv files that held the data of interest. This was necessary because when viewing the data within the files, each row representing an individual

respondent had many empty cells of NAN that could be condensed. Therefore, the goal became to take the original 53 columns in the midwest dataset and the 59 columns in the south dataset and reduce them each into 9 columns, indexed by the actual respondents ID from the original surveys. These 9 columns would be:

1. Written in Responses
2. Degree of Identity (how much each respondent identified with the region in question)
3. Midwestern?/Southern? (which states the respondent voted as being midwestern or southern)
4. Zip Code
5. Gender
6. Age
7. Income
8. Education
9. Census region

After downloading the raw csv files from the FiveThirtyEight github website and saving them locally, they were opened in a Jupyter Notebook and Pandas was imported to begin exploring the data.

3a. Removing Unnecessary Heading Data

The first issue noticed was that even though there were 9 parts to each survey, there were a number of unnamed columns. It was decided then to begin cleaning by removing the first row, since the second row contained the possible choice responses and could be used as column names. This eventually lead to better organization of the columns later.

3b. Condensing Empty Cells into a Single Column

Many respondents had single responses for certain columns, yet there was a column for each possible response a respondent could give. This meant that much of the information from groups of columns could be condensed into a single column with a single answer the respondent gave, removing any other empty cells. By forward filling the answers along rows to a single column, then pulling that column into a clean dataframe with a new column name, it was possible to extract all the information of interest, taking many rows with empty cells and efficiently compiling them into single columns of interest with all information necessary. This significantly reduced clutter, presenting clearer information for all the responses by the survey taker.

A difficult step came from the list of states that the survey taker could choose from in order to indicate which states they believed belonged to certain regions. Because there were different columns for each state, and the value for that column was the state name, a different method was used that could combine only the states that were selected and leave all the empty cells behind. By using a lambda function that was able to select only the cells in the dataframe that had value, and then placing them all in a list, only the information of interest was extracted and compiled into a single column.

3c. Empty Values

The final choice came from the remaining NaN values. Though many survey takers had filled in most of the reponses, some left areas blank. Two significant areas left blank were the areas of self-reported income and education. Since these areas were ones of interest for what personal factors can determine regional identity, it was decided to drop the rows that did not contain at least one value in those columns. This dropped 348 respondents from the midwest dataset and 267 respondents from the south dataset, leaving still over 2200 respondents from each dataset. This was a minimal loss compared to the original totals of 2778 from the midwest dataset and 2528 from the south dataset.

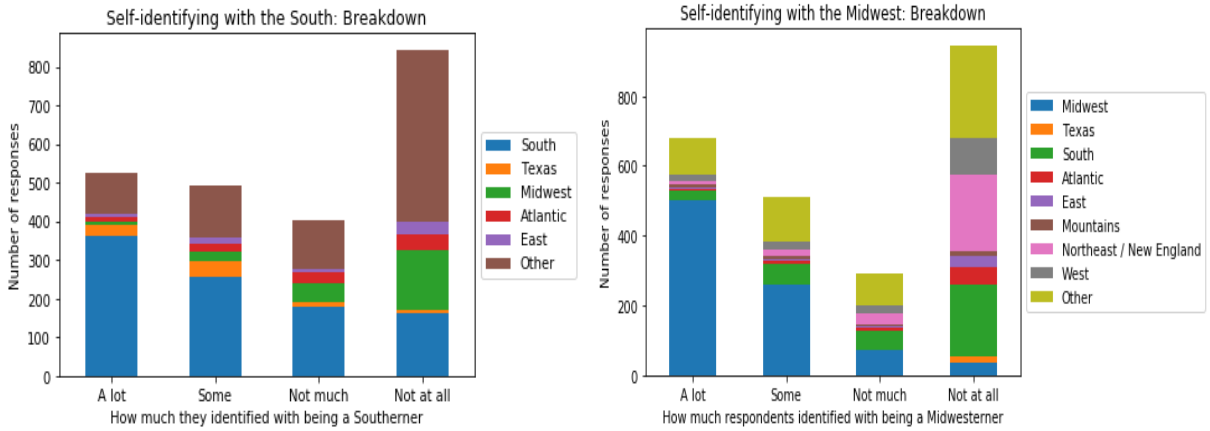
3d. Extra Data

Lastly, the dataframe for the median reported household income for each state was acquired by downloading the excel file locally from the Census Bureau website, then isolating the columns related to the year 2014. This included a column for the standard error for each state. There were no NaN values in this dataset, so there was no further cleaning.

4. Data Story

4.1 Breakdown of Written Responses and Self-Identity

Looking at the data and breaking down the responses of how people identified with being either a southerner or midwesterner, there was an interesting spread of results. Below is the comparison of the responses of those that identified with being either southern or midwestern and the written responses:



Looking at the southern data on the left, if the written response contained the word “south” or “southern”, they were more likely to identify as strongly being a southerner. Similarly, but with a stronger

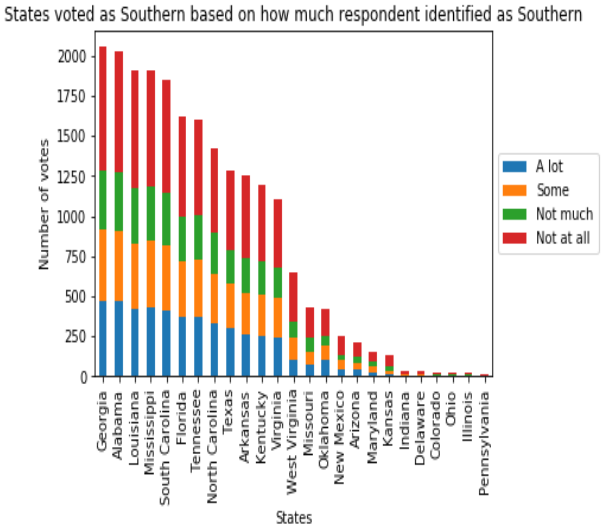
distribution seems to be the use of the word “Midwest” in peoples’ responses when identifying as a midwesterner.

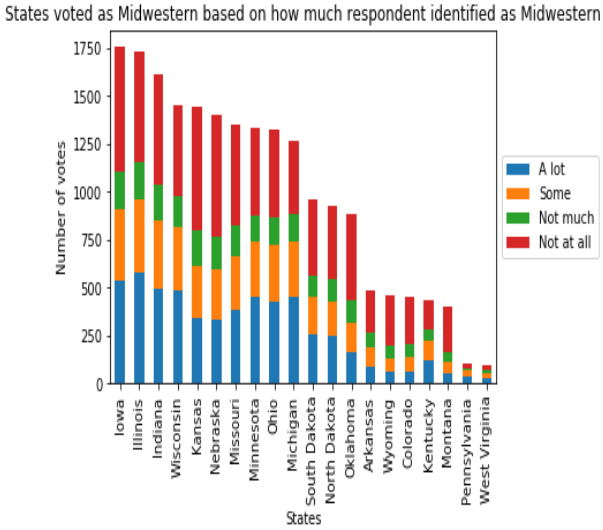
Yet, an interesting result that occurred when looking at this data is that there were some responses that used “Midwest” in their written responses, even though they strongly identified with being southern. Similarly, this was also seen with the midwest data that some respondents used “South” in their written responses even though they strongly identified as being midwestern.

An issue that arose from this data was that there were many responses that were unique for the respondent. These included long sentences explaining their region, or colloquialisms such as “heaven” or “bible country” which did not give a regional location and were only used once out of all the responses. This left the “other” category quite large and hard to breakdown accurately.

4.2 Breakdown of States and Self-identity

The next step was exploring the states that were voted to be in either the Midwest or South based on the respondents’ personal identity. This lead to the results below:



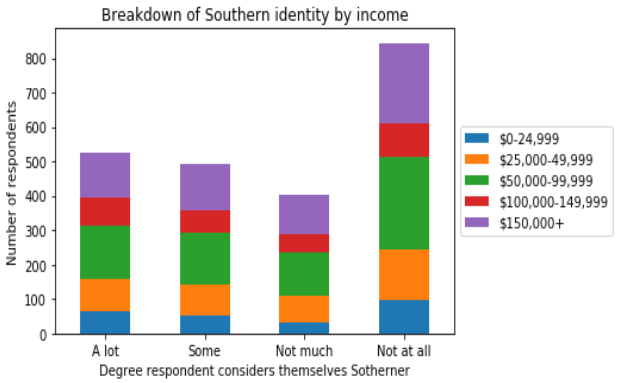


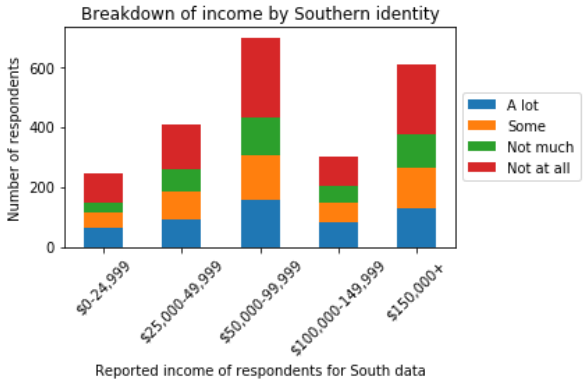
Looking at the data, it is clear that there are a number of states that people identified as being southern or midwestern. Looking at the southern data, Georgia to Virginia clearly had the most votes, with a drop-off at West Virginia. However, we can see that a large amount of the votes for these states also come from those respondents that do not strongly identify as a southerner. Looking at the red bars above, there are plenty of people that do not identify as a southerner that have their own idea of which states are southern.

This is also evident in the midwestern data. Looking just at Kansas and Nebraska, very few votes for those two states came from those that strongly identify as being midwestern. In fact, if we just look at the two respondent groups that identify “A lot” and “Some”, Kansas and Nebraska would actually fall closer to South Dakota in terms of how midwestern they are thought to be. Yet, with the votes from those that do not really identify themselves as midwestern, these two states seem to be closer to Wisconsin, a state that a lot of self-identifying midwesterners agree as being in the Midwest. From these two graphs, we see the influence of how people outside of these regions view these regions and apply their own thoughts and biases on what they believe to be in the region.

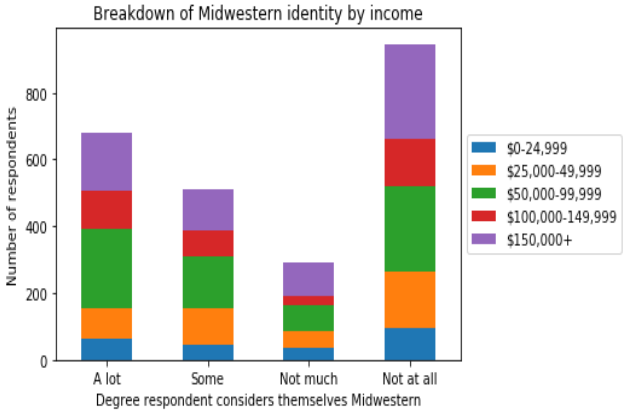
4.3 Income

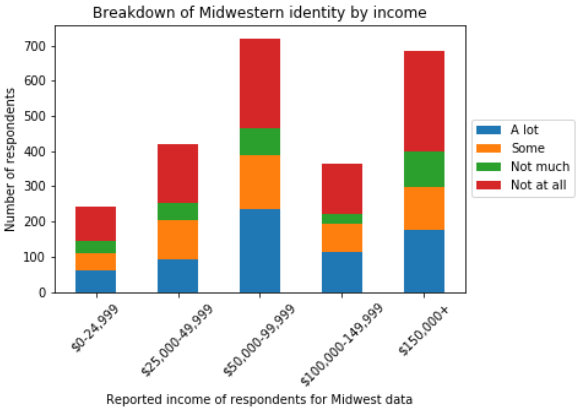
Looking at the breakdown of income for both the southern data,





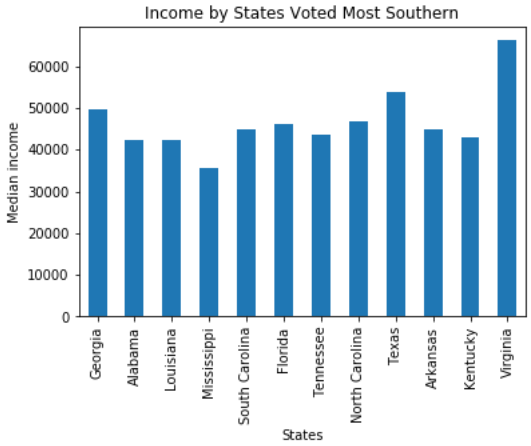
And the midwestern data,

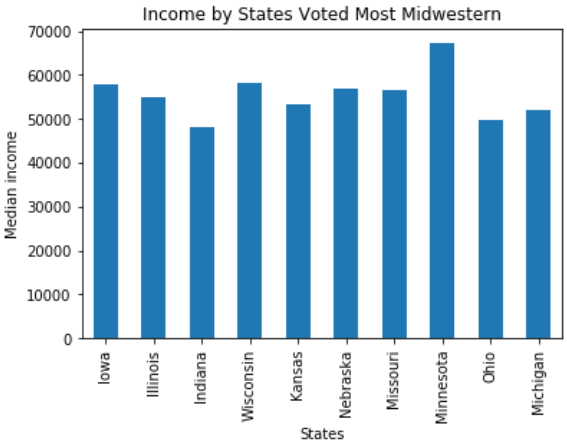




We can see a lot of similar breakdowns. However, because of this, it is actually more likely a description of the type of people this survey reached rather than an accurate description of what the breakdown of income is in those regions. This is because we can see that a large part of those that identified “A lot” with being southern or midwestern had an income of $150,000+ despite the fact that the average income for those states back in 2014 was much lower.

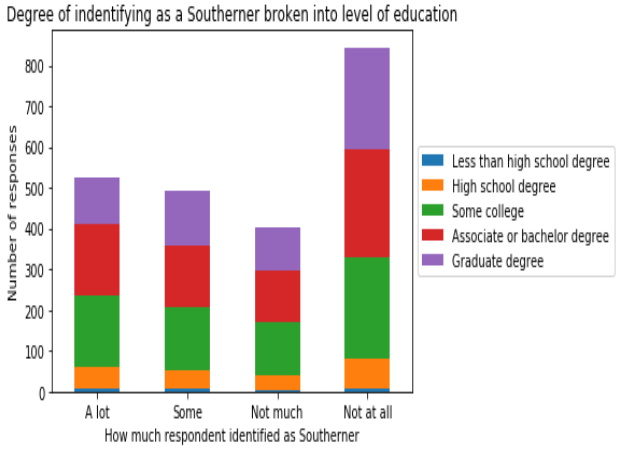
Looking below at the top states that were voted most southern or midwestern, the median income was closer to $50,000 for the south and $60,000 for the midwest. There was a large group of respondents that did fall into those income ranges above. However, because a significant amount of respondents seemed to make over $150K, this raised some doubts as to this survey being an accurate representation of the average population from the South or Midwest.

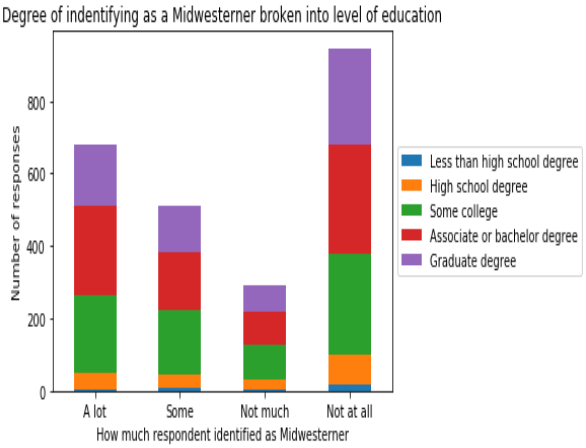




4.4 Education

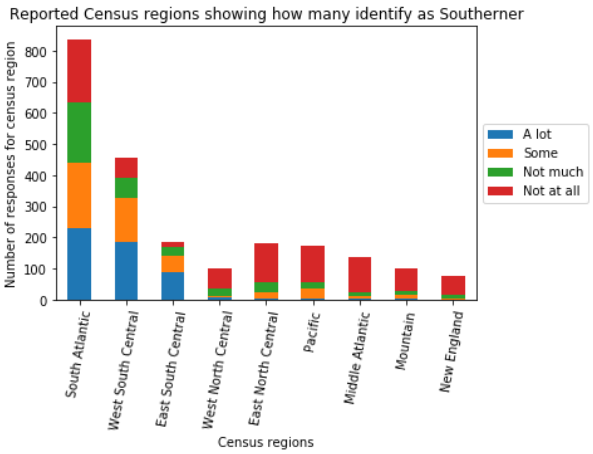
Looking at the graphs representing the breakdown of education, we can see that most of the respondents had at least some level of college education and that proportionally, there was no difference between groups ranging from strongly identifying with their region to not at all. This, again, most likely has to do with the general reach of this survey and that most visitors to the FiveThirtyEight site have at least some college level education.

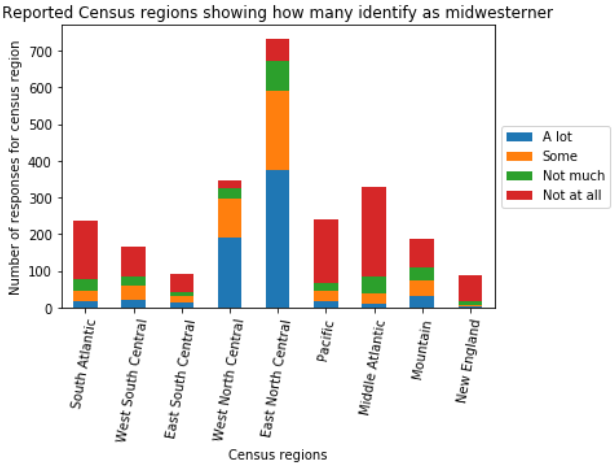




4.5 Census Region

Looking at the Census region responses, there is an interesting breakdown of the respondents based on their actual locations. Based purely on these responses, we can see that respondents that did not at all identify with being a southerner or midwesterner were spread out throughout most of the country, yet those that did identify strongly fell into clear census regions.





It can be clearly seen for the southern dataset that those that identified the most strongly as a southerner fell into the three census regions: South Atlantic, West South Central, and East South Central. Yet, an interesting point of data came from the fact that a lot of those in the South Atlantic region did not identify as a southerner at all, despite the fact that the majority that did had come from this region. Then looking at the other regions for this dataset, we can see where most of those that did not at all identify as a southerner came from many of these different regions.

Looking at midwestern data, we see different interesting results. The two regions that the majority of strongly identifying midwesterners came from were West North Central and East North Central. Yet, there were still several respondents that identified strongly with being a midwesterner that actually fell into very different regions. For example, the same regions that above that identified the most as southern also had individuals that identified "A lot" as being midwestern.

5. Conclusions of EDA

The exploratory data analysis of the regional data of the FiveThirtyEight survey had drawn the following conclusions:

- For the most part, when describing their location in their own words, those that identify strongly with their region will use words that identify with that region specifically (i.e. "South" for southerners and "Midwest" for midwesterners).
- There are clear results for particular states that are believed to be southern or midwestern, though how those breakdowns happen are shown to be distinct between those that identify strongly with those regions than those who do not.
- The income and education distinctions between the fours groups of self-identity were non-existent. If anything, the data for these two categories showed more information about the visitors to the FiveThirtyEight website than any actual breakdown of the midwest or south respondents.

- Lastly, the census region data showed how the breakdown of regions were in terms of the Census Bureau defined regions of the United States. There was a clear breakdown here, with those identifying a lot with south or midwest regions of the US falling into distinct regions.

The following these conclusions, the next step was to perform more in-depth analysis with different machine learning models. Judging from the results above, such models would likely rely heavily on the states respondents voted for and their census region to determine how likely they are to be southern or midwestern.

6. In-depth Analysis with Machine Learning Algorithms

Based on the EDA, the two main categories of interest in order to determine whether someone thought of themselves as a southerner or midwesterner were the states they voted for as being in the South or the Midwest and what was their census region. Since the data was already labeled into four distinct categories, this meant that we had to narrow down which algorithms would be best to use based on supervised learning algorithms. The two best algorithms being Support Vector Classifier and Random Forest Classifier.

The reason behind these two models is that they are able to work with categorical data and can break them down into their labels of interest. The main difference between them is that Support Vector Machines (SVMs) are more affected by the amount of features you add. Therefore, they can work in higher dimensions, but more information does not necessarily make them function better.

Yet, Random Forests are unaffected by additional features that have no influence on the data. Therefore, it makes an excellent kind of safety net to ensure that if somehow there was another feature that influenced the designation of the labels, it could be discovered without overfitting the data. This meant that we could add the categories of education, income, and gender to determine if any of these additional features played a role in self-identifying with a particular region of the US.

Starting with the SVM for classifying with the south and midwest data, the Midwestern?/Southern? category for each of the clean datasets was broken down such that each row had a number of columns equal to the number of states they could have voted for as Southern or Midwestern. Each of the cells in those columns were then given a one or a zero based on whether they were voted for, in order to create a multidimensional array of results for each of the respondents. The census regions for each respondent was broken down similarly into binary responses.

Breaking the data into training and testing data, a grid of parameters were given such that the hyperparameters could be tuned in order to achieve the highest level of accuracy possible for this model. For the South dataset, however, it was only possible to achieve a 49% accuracy result for the SVM model. This is likely due to the fact that with the southern dataset, there was no cleanly defined results. Looking at the distribution of the census regions in the EDA, we can see that there are three regions of the US that people who identify strongly as being southern come from: South Atlantic, West South Central, and East South Central. This means that with no clearly defined line, achieving a higher accuracy for the model was difficult.

By comparison, the midwestern dataset accuracy was 61% for the SVM model. This could be because due to the fact that the census region data was more cleanly split into the two categories of West North Central and East North Central. This would have helped in narrowing down at least two labels of interest for self-identity: “A lot” and “Not at all”. In fact, the other two labels of “Some” and “Not much” were not even given a precision and f1-score in the classification report for the SVM of the midwest dataset.

For the Random Forest Classifier, the results in accuracy were the same for the midwest data, but at least 2% improved accuracy for the south dataset, for a total of 51% accuracy for that data. While the Random Forest model worked slightly better than the SVM for the south dataset, they were pretty similar, despite the fact that the Random Forest had more information. This meant that the extra categories that were added to the Random Forest Classifier were probably not as significant as the first two categories of voted states and census region information.

7. Results

Through looking at these two classifiers, we see that the two most significant categories for determining a person’s self-identity with their region are there physical location and their opinion of other similar states to that location. Yet, even with that information, there is no guarantee that we could create an accurate machine learning model to predict if a person was from that region or not. This is likely because there are a number of different factors that can determine how a respondent might identify with being from a particular region. Population size for the city they live in, local diversity, and annual traditions can help explain why one respondent from a state may identify as being “A lot” southern and why another from the same state may say “Not so much”. There were only nine categories for this particular survey and because it was done on an online platform through a website, the amount of respondents and their different backgrounds may have influenced the data. This meant that though we could analyze the information given by the data collected, we are not likely looking at an accurate representation of people from these regions of the US.

8. Next Steps

For future surveys, I would try to create a campaign to distribute the survey more, perhaps even specifically targeting people from the regions in question. I would also suggest adding questions such as what is the population size of their city, or whether they think other citizens in their city would identify their state as southern or midwestern. By acquiring more information, we would be able to get cleaner data than just from visitors to the FiveThirtyEight website and more information from the respondents as well. Then, we could look at creating a better machine learning model.