Where are you from?

PREDICTING REGIONAL IDENTITY IN THE USA

An exercise in survey analysis and exploration of regional data

Rae'e Yamin

Why?

- Marketing: In order to better sell a product, companies need to learn more about their consumer market. By understanding how consumers view themselves, companies can better tailor their marketing strategies toward them.

- Politics: Politicians would also be curious to learn more about how their constituents in their districts or states view themselves.

- Fun: There are many who are just curious to know which states fit into the South or Midwest.

Data

Data was collected off of the FiveThirtyEight website's GitHub account. They had performed two online surveys asking similar questions but based on two regions of the US: The South and The Midwest.

Both data files were CSV and can be found in the link below:

- FiveThirtyEight GitHub

There were 9 categories of interest in these surveys:

- Written in Responses
- Degree of Identity
- Midwestern?/Southern?
- Zip Code
- Gender
- Age
- Income
- Education
- Census region

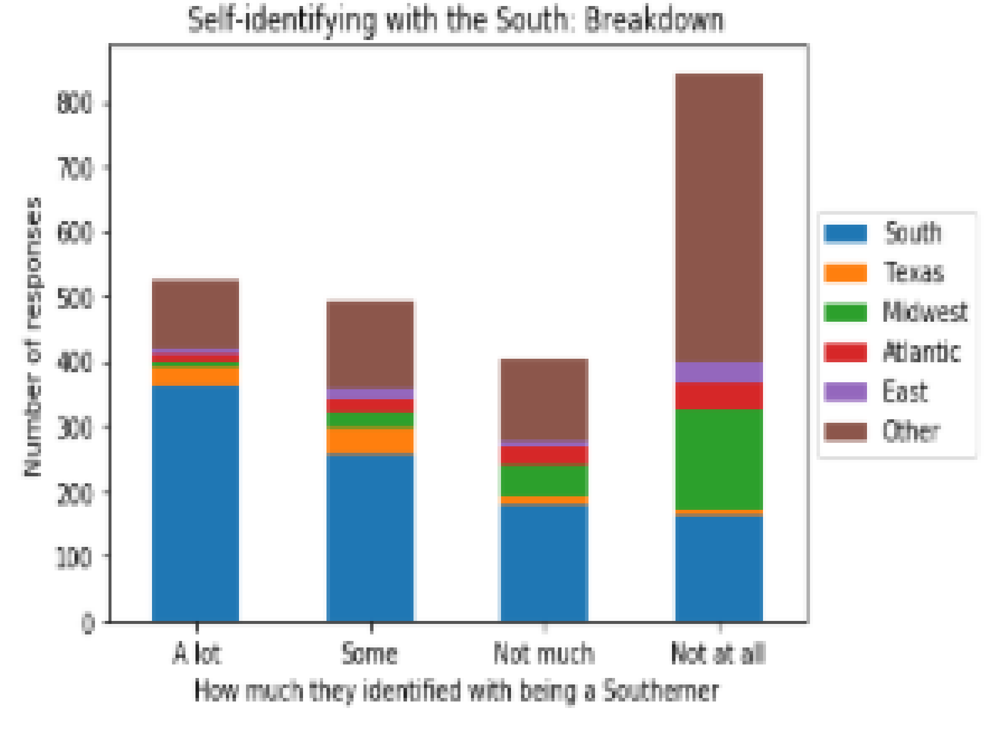
Data Cleaning

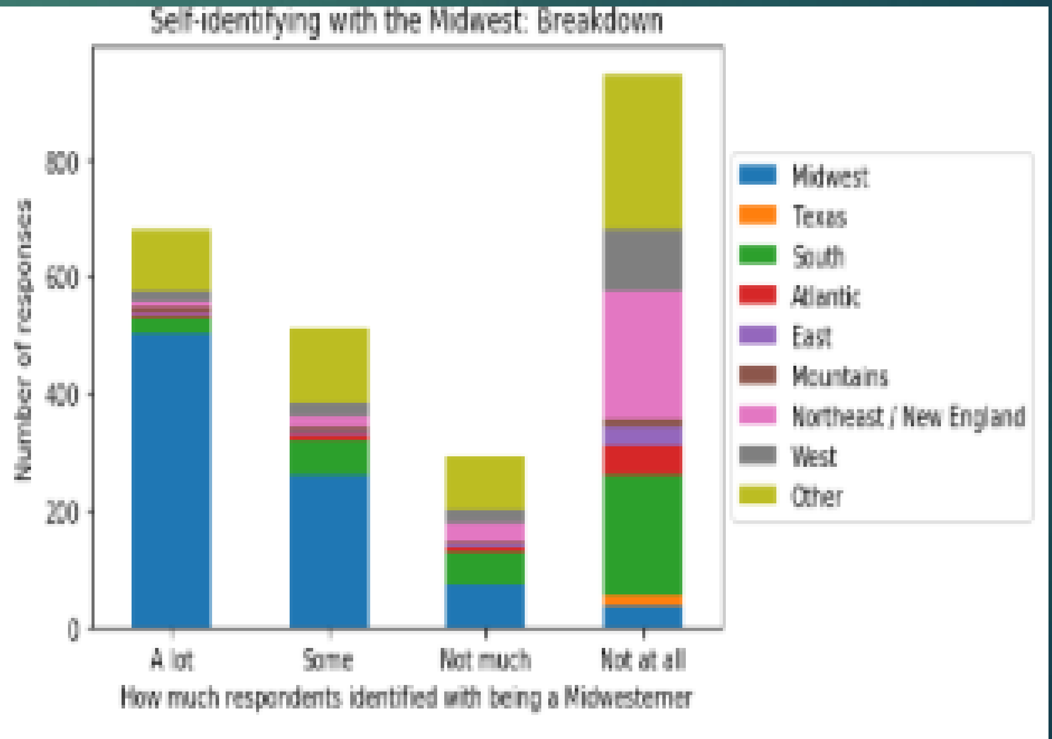
- Problems Encountered:
  - Problem 1: Unnecessary headings.
    - *Solution:* Removal for easier grouping of 9 category columns of interest and rows of individuals' responses.
  - Problem 2: Empty cells.
    - *Solution:* Fill forward on categories that only have a single response, then create a lambda function to collect voted states as single list of all states respondent voted as either Southern or Midwestern.
  - Problem 3: Empty Values.
    - *Solution:* Removal of rows that did not contain values for income or education.

Full Data Cleaning report

EDA

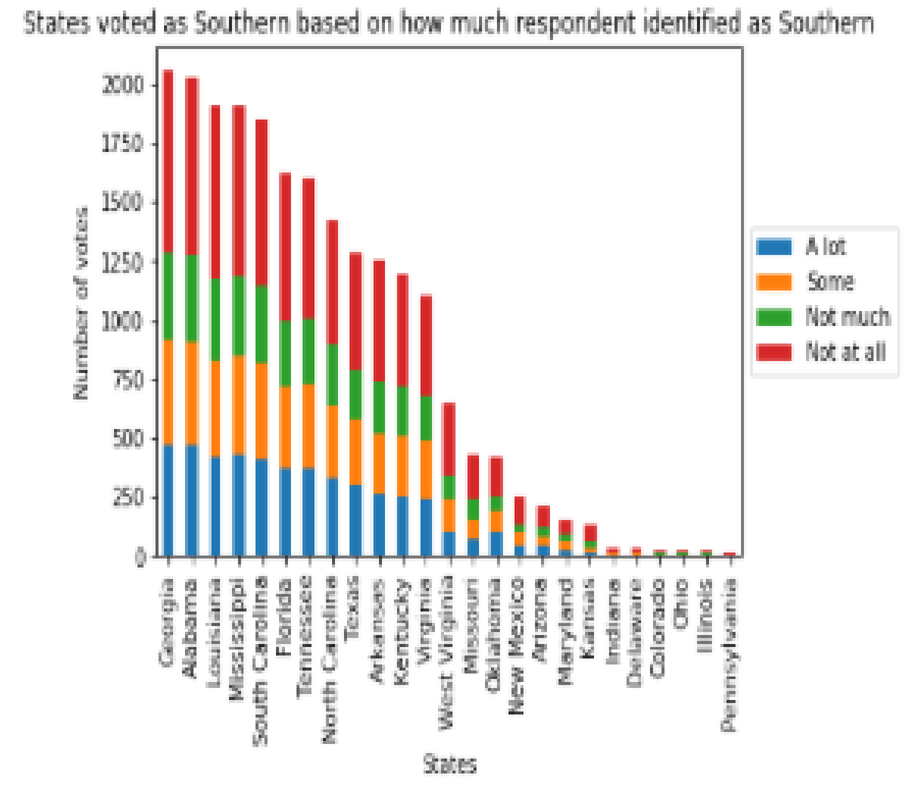
Breakdown on labeled self-identity and written responses:

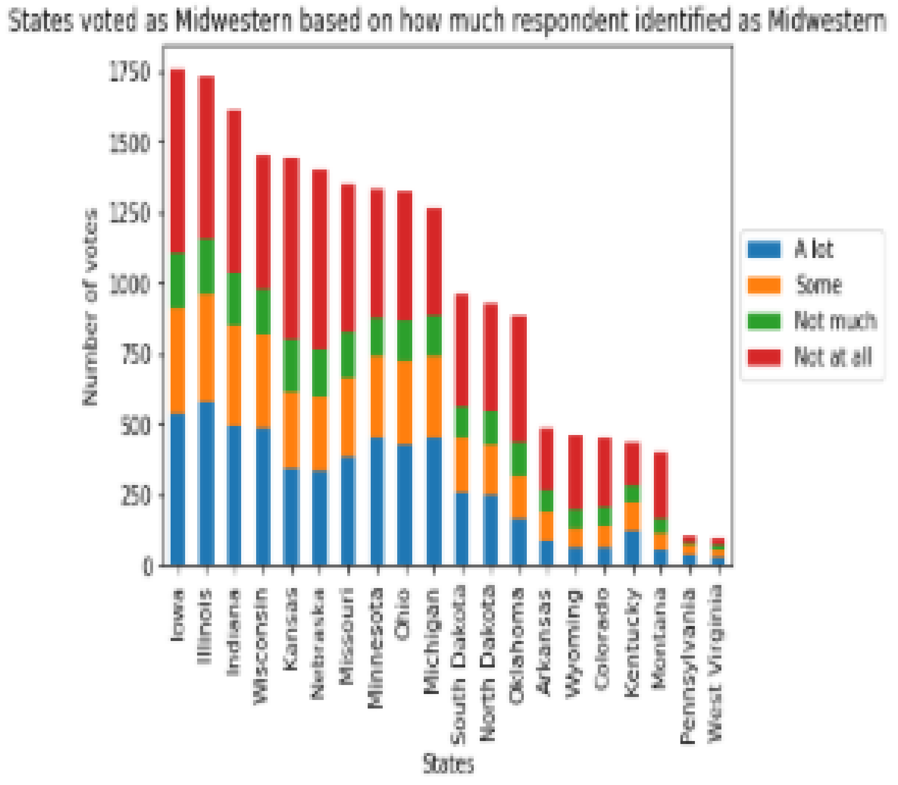




EDA

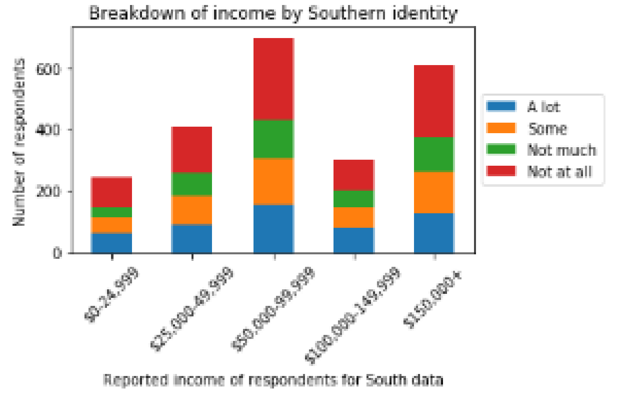
Breakdown of labeled self-identity and states voted:

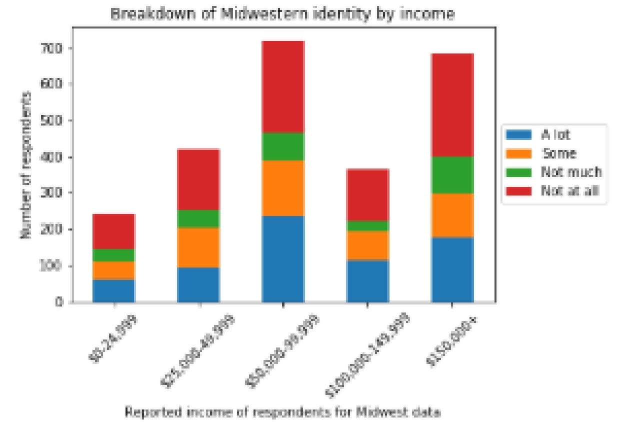

States voted as Southern based on how much respondent identified as Southern


States voted as Midwestern based on how much respondent identified as Midwestern

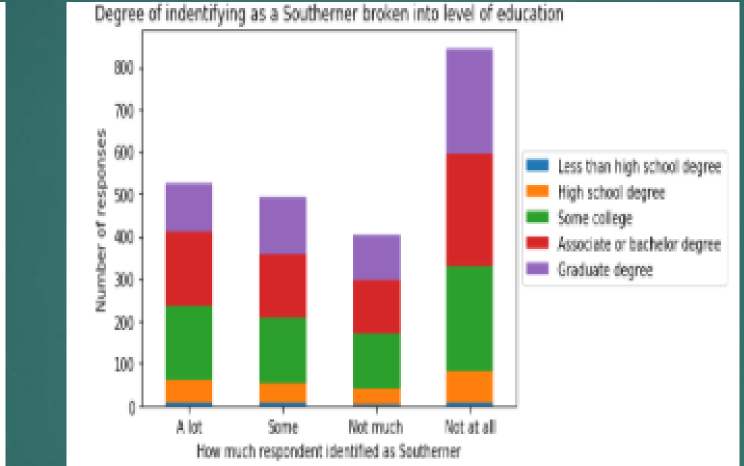
EDA

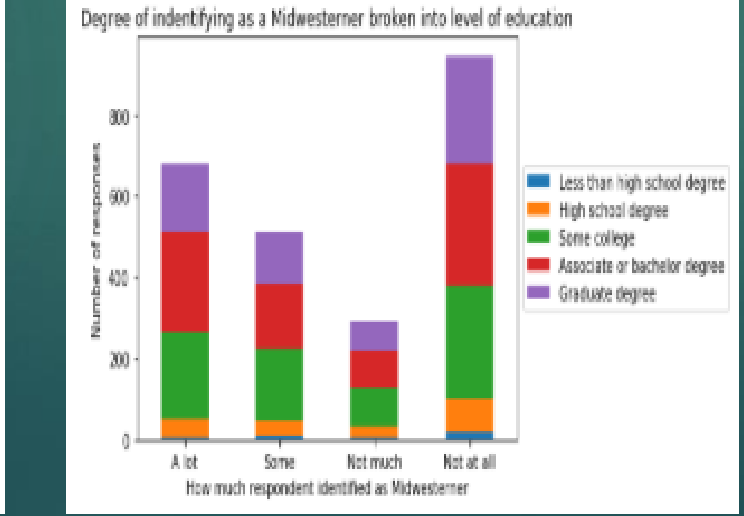
Income:





Education:

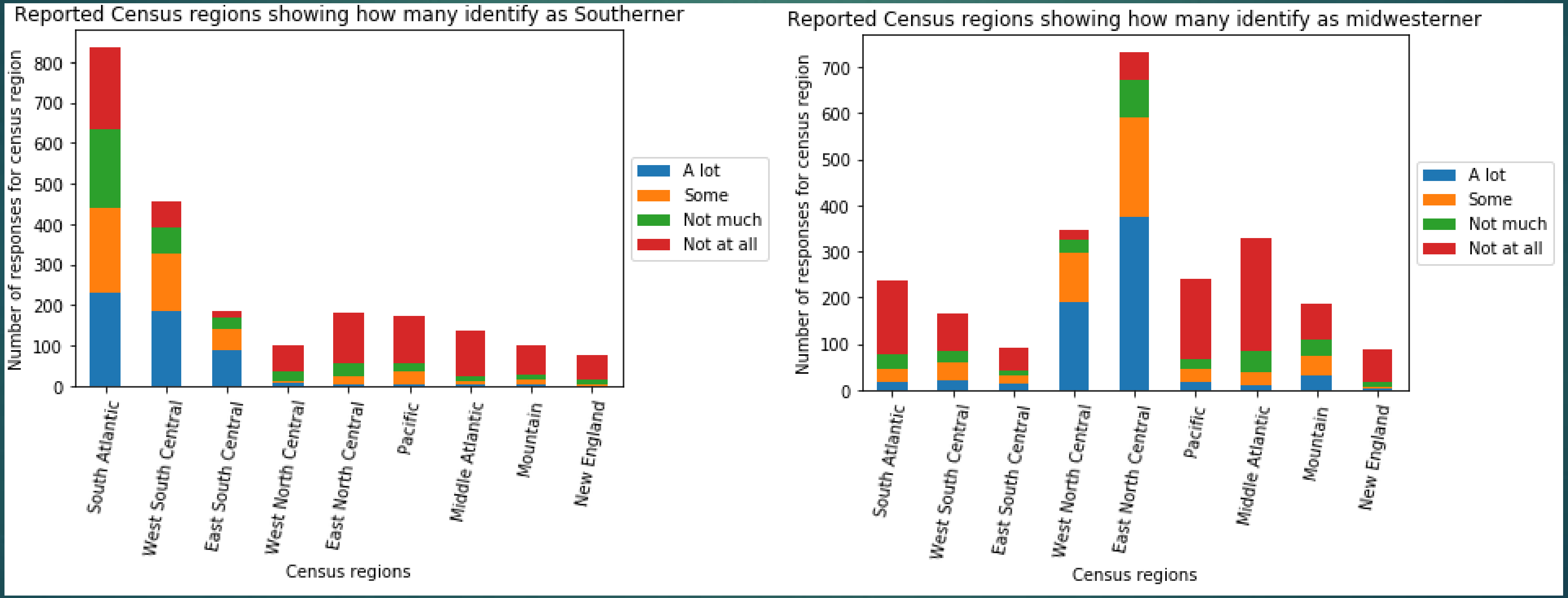




- Reflected more the typical traffic to the website rather than an actual survey of the areas.

EDA

Breakdown of self-identity and census region:



Reported Census regions showing how many identify as Southerner

Reported Census regions showing how many identify as midwesterner

In-Depth Analysis

- Supervised learning was used since the labels for the dataset was already provided with the 4 self-identifying categories: “A lot”, “Some”, “Not much”, “Not at all”.

- Algorithms:
  - SVM and Random Forest classifiers were used because of the categorical nature of the dataset.

- Random Forest classifiers won out for the southern data with 51% accuracy, Midwestern data was unchanged from other model of 61% accuracy.

Future Improvement

- **More detailed questions.** Real data was used from surveys but because the data was from an online website, the types of questions asked and the respondents that answered were clearly more skewed towards the typical visitors to the website. Real businesses would ask more detailed questions and would more directly target different populations in the US.

- **More data.** While over 2000 samples were used for each dataset of the South and Midwest, because of how ambiguous the categorical information was, it was hard to create a predictive model for the categories that were not at the extreme ends of the spectrum of choices.

Thank You!

QUESTIONS?

Rae'e Yamin