The Great American Political Migration Visualised

Emory Swanger, Christian Haynes, Omar Iqbal, Justin Langston, and Zachary Ables

Abstract— **After collecting data from the 1976 to 2016, several charts, graphs, and maps, some interactive, were implemented using python and HTML. Using these visualizations, questions about election trends and patterns were questioned and certified or rejected.**

I. OBJECTIVE

The goal of the project was to find a large data set, find way to analyze it, and present the findings in an aesthetically pleasing manner. Ultimately the topic for the group was decided and agreed upon by all members: The geographical shift in American politics. This idea spawned from a collective question the group shared, "What does the shift in American politics look like?" To answer the question, data had to be collected which could be presented on a map that. This would show how some areas of the country shifted to the right or left. Naturally, this led to several other hypotheses:

- Hypothesis 1 - States with larger overall populations are more likely to vote for the Democratic party.
- Hypothesis 2 - States with a greater population density are more likely to vote for the Democratic party.
- Hypothesis 3 - Republicans are more likely to win states with the most land area.
- Hypothesis 4 - The Republican and Democratic parties generally hold a duopoly on American politics in contrast to the combined total of other parties.
- Hypothesis 5 - Most states are loyal to a single party and rarely flip.

A linear regression model would also be implemented to try and find what feature is most correlated with winning party in a presidential election.

II. DATA

Before any work can be done, a large data set had to utilized or created. The former option was taken because it is more practical. Using Google's dataset search, presidential election data was harvested from the Harvard Dataverse database. The file obtained was a CSV format with a total of 14 features and 3740 entries. This data covered presidential elections from 1976 to 2016, a total of eleven elections. The features include the year, state name, state initials, state_fip, state_cen, state_io, office, candidate name, party, write in, votes for the candidate, total votes for the state, version, and notes. The only features kept were the names of the states, their initials, the election year, the names of the candidates, their associated party, number of votes for both the candidate and the total votes of a state for the given year. All other features were discarded. Population density and state area were manually searched for and entered into their own CSV files. To validate the data, several google searches were made and the results were compared to our data. All of which were within a five percent margin of error.

III. MODELS AND ALGORITHMS

After modify and formatting the data, the data was read in using the python Pandas library. Next, each of the CSV files were used to create several types of graphs and charts: line graphs, scatter charts, bar graphs, pie charts, and maps of the USA. Some of these were made to be interactive; using a slider widget, the map would change results based on the year selected by the slider. Afterwards, the linear regression model would be used to find what feature or set of features could best find correlation between the data and the various election outcomes.

IV. RESULTS

After countless hours of research, coding, and hard work, the graphs, charts and maps were finally generated. These visualizations displayed the total votes of the winners and runners up by state and nationally, the voting distribution amongst the parties, and the number elections won by each party for each state. By analyzing these displays we were able to reject or fail to reject several of our original hypotheses.

A. Hypothesis 1

A sample of the eight most populated states in the union was taken: California, New York, Texas, Georgia, North Carolina, Michigan, Pennsylvania, and Illinois. A large population is defined as having a population of approximately or over ten million. Analysis of both state leaning maps reveal that the aforementioned states are evenly split into Republican or Democrat-favored states. Thus, this hypothesis is rejected.

B. Hypothesis 2

A sample of the eight most population dense states was taken: New Jersey, Rhode Island, Massachusetts, Connecticut, Maryland, New York, Florida, and Washington D.C[1]. A dense population is defined as having an average density of three hundred or more persons per square mile. Analysis of density vs election results strip chart reveal that a majority of the aforementioned states are carried by the Democratic party. Thus, this hypothesis is not rejected.

[1] D.C is not a state but is included because the territory participates in the presidential elections.

C. Hypothesis 3

A sample of the eight largest states in the union was taken: Alaska, Texas, California, Montana, New Mexico, Arizona, Nevada, and Colorado. Analysis of the Land Area vs Overall State Lean strip chart reveals that six of the eight largest states lean toward Republican Presidential candidates. Although, there is a 0.75 correlation, this hypothesis and cannot be strongly suggested.

D. Hypothesis 4

By analyzing the Vote Distribution pie charts, this hypothesis is strongly suggested. In every year, the combined total of other parties did not come close to the number of votes for either the Republican or Democratic parties. Surprisingly, the number of votes for third parties were still larger than expected, most notably 20.0% and 10.5% in 1992 and 1996, respectively.

E. Hypothesis 5

To test the hypothesis, the Total Elections Won by Party vs State chart and the State Leaning by Elections Won map were used. The chart revealed that seven out of the fifty-one voting territories had a six to five ratio of wins for either Democrats or Republicans. On the other had, many other states had a clear party favorite: eleven states have a 11-1 score, ten states have a score of 10-2, and three states have a score of 9-3. The rest could be considered moderately to barely loyal to a single party. With a total of twenty four states that can be considered fiercely loyal to one of the two major parties, this hypothesis can be rejected.

V. Problems Encountered

The most common issue suffered amongst the coding team was installing and importing the necessary libraries. The remedy was found in several blog posts and documentations for the libraries. The next difficulty was implementing the slider code. Originally the code for the slider was from a different project, one that was used to map crime in America. Modifying portions of it was not easy but it was eventually accomplished. There were also issues with using Git hub and opening HTML documents. In regards to the linear Regression, we could not find any strong indicators of whether or not a particular factor had a strong influence on the results of the election. The way the data was separated, using a linear regression was not a sufficient method of machine learning to determine an influential factor.

VI. Future Works

Our maps are binary by design so there is not much room for nuance. In future works, the team plans on learning more about the mapping code. This way, we can give different shades of red or blue to illustrate how some states are much more of a tossup in elections. Once that is accomplished, other elections can be covered. These include other domestic elections such as primaries, congressional, state, or county. If election maps are available for foreign countries, they are also potential projects as well. The team may also explore other charts that can visualize different aspects of elections. Given all the data and mapping data, using machine learning algorithms to predict the results of future elections maybe within grasp. This will require additional machine learning libraries such as sklearn or pytorch, along with additional data such as the final electoral college votes for each state in previous elections.

VII. Organization Chart and Timeline

Most of the real work began in October. At this point, the raw data was found. By the end of the month, the data had been processed and refined. Starting in November, the code for graphing began and was completed by the end of the month. The paper was also started by the second week of the month. The final presentation of the assignment is expected by December 8, 2020.

TABLE I
Responsibilities Table

Name	Responsibility
Christian Haynes	Coding foundation and initial graphing
Omar Iqbal	Worked on the linear regression model
Emory Swanger	Wrote the paper and analyzed the graphs
Justin Langston	Advanced the graphing code
Zachary Ables	Data collection/formatting and linear regression

Acknowledgment

We would like to thank Dr. Mockus for providing us with the skills needed to complete the project as well as the creators of libraries and code we used. Those tools were absolutely crucial for the success of this project.

References

[1] "U.S. President 1976–2016", https://dataverse.harvard.edu/dataset.xhtml?persistentId=d

[2] "Land Area and Persons Per Square Mile", https://www.census.gov/quickfacts/fact/note/US/LND110210: :text=Density%2520is%2

[3] "Step-by-step: How to plot a map with slider to represent time evolution of murder rate in the US using Plotly", https://amaral.northwestern.edu/blog/step-step-how-plot-map-slider-represent-time-evolu