Topic: US Election 2020 Analysis
Summary Paper

Introduction

For this analysis, our group was interested in understanding the outcome of the 2020 US Presidential election by answering the following questions:

- *What kind of trends can be seen in governor and presidential elections?*
- *What are the election results versus polling predictions?*
- *How can we predict electoral wins based on median income?*
- *How can we estimate the presidential winning party based on ethnic groups?*
- *What kind of county demographics categories could be used to infer the voter turnout?*

Our main data consisted of US Election 2020 President/Governor votes by county, demographic data, 2019 population estimates and other supporting data. They were collected primarily from Kaggle for the voting data, and the US Census for the demographic data. To facilitate the analysis, our first goal was to generate a data frame containing the voting data, state, county, and their associated information with one row per county. The original data sets have multiple discrepancies in the number of entries, redundant data that we do not need for the desired analysis, and spelling errors. Data cleansing was performed by dropping duplicates, selecting and combining columns, and matching counties names between files before merging them into a single table.

Analysis and Model

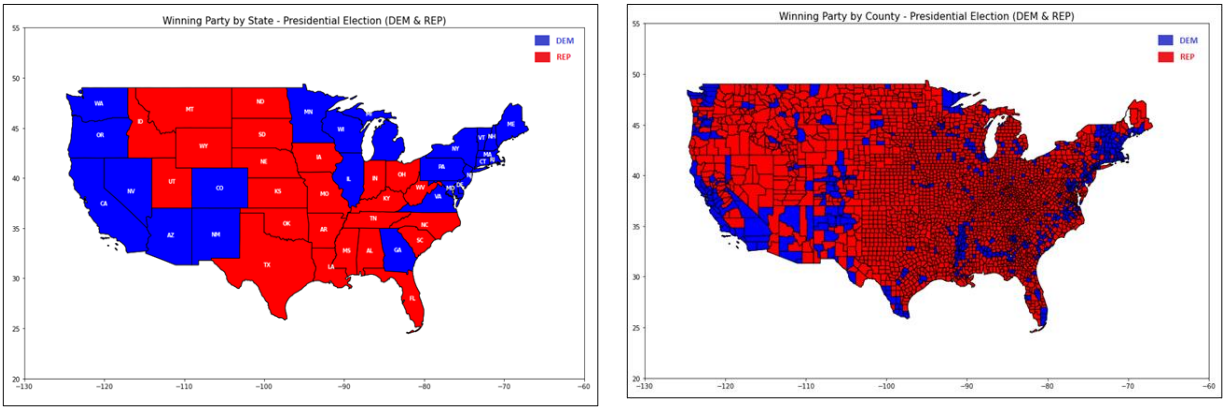


Figure 1: Winning party by state (left) and winning party by county (right)

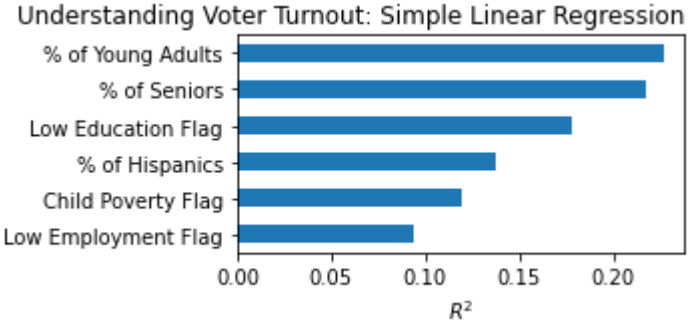
Our first step in data analysis was data visualization. Maps of the winning party by state and winning party by county were plotted up (Figure 1). An interesting observation from the visualizations is that when we look at county level results for both parties, we see that more number of counties voted for the REP party. Even though the number of counties were higher, the more populated counties voted for the REP party.

To answer the second question, we used the 2020 pre-election pollster's survey results from various sources, performed an analysis that compared the popularity of election candidates in

each county and state with the final election result. This comparison aimed to find how accurate the polling outcome was and whether it is a suitable method to predict the election result. Our analysis answers that polling results can only show the chances of accepting a candidate in a state. Still, it cannot fully predict the election outcome since many other significant factors like economic shifts, protests, campaign events, but mainly the electoral college, can influence the final decision.

Overall, there are more populations in democratic states than republic states. There are more difference between democratic and replubican state in White american, asian american, and hispanic american. Young adults group has the most population than mid-age adults and senior adults. As a result, the young adults group could be an important group for election. There is not much difference between male population and female population. From economic data, there is a greater number of counties in low employment in republican states than democratic states.

Two logistic regression models were developed to predict president wins using median income and ethnic groups. The first model, based on median income, predicted more wins than true wins. It was not very good at discriminating between wins and losses as the correlation was weak between the predicted and true class. The model was good at finding wins but not good at discriminating a win from a loss. This means that the model predicted more wins than there were true wins by a notable amount. Using this model would not be reliable. On the whole will predict the correct winner overall but not on the county and state level which is necessary to be accurate enough to accurately predict the outcome of a presidential election. The second model, based on ten different ethnic-gender groups, also showed a weak correlation between resulting regression parameters and the voting results (i.e. ethnic and gender of population are not correlated with the election results). It predicted that the DEM party would lose, based on the data sets. This is consistent with the map shown in Figure 1 where the REP party won in a larger number of counties than the DEM party. This data set is for popular national votes, so it is not enough to determine the election win which is determined based on electoral college votes.



Finally, we created both single and multiple linear regression models to infer the voter turnout in the presidential election and determine the extent to which county demographics could account for the variations in voter turnout. The simple linear regression analysis indicated that of various age, ethnic/language, social, and economic factors analyzed, age was the strongest indicator, with the percentage of seniors being positively associated with voter turnout. The five remaining features shown in this plot were all negatively correlated with the rate of voter turnout in US counties. Independently, each of these factors accounts for approximately 10-20% of the variation in voter turnout between US counties. When combined in a multiple linear regression model, our analysis achieved an adjusted R^2 of 0.49, indicating that the selected demographic variables could account for almost half of the variation between counties. Input variables that are not strongly related to voter turnout are: the rural/urban status of the county as well as most economic drivers of the county (farming, mining, manufacturing, government, and non-specialized).