Final Project Submission

Demographics and Politics

Aaron Moriak
Marquette University
Milwaukee, WI 53233, USA
aaron.moriak@marquette.edu

Bria Powell
Marquette University
Milwaukee, WI 53233, USA
bria.powell@marquette.edu

Dominique Green
Marquette University
Milwaukee, WI 53233, USA
dominique.green@marquette.edu

Tajah Lynch
Marquette University
Milwaukee, WI 53233, USA
tajah.lynch@marquette.edu

Abstract
The purpose of this project is to determine which of the various demographics are useful for future election predictions. To form ideas about the predictive demographic, the 2016 election data is used. It comprises primary election and presidential election results on the county level, as well as the necessary county facts data. The main programing language will be R for statistical calculations and python for maps. Regression and classification analysis will be performed as the measure of predictiveness of each demographic.

Author Keywords
I.5.0: General, I.5.1: Models, I.5.2: Design Methodology, I.5.4: Applications; I.5.5: Implementation, I.6.4: Model Validation and Analysis

Introduction
The objective of this project is to determine which specific demographic is predictive of the presidential vote. This analysis was chosen in response to the controversial president-elect, Donald Trump, as many believe conservative White Republicans controlled this election. First, county maps, scatter plots, and bar plots will be used to determine correlation and distributions of votes based on the demographics in each county. Then, based on relationships of causality, a model will be fit using least-squares estimation. This model will be evaluated using the coefficient of determination, chi-squared values and significance, and normality tests.

Datasets and features
The main dataset used was the *County Facts* dataset which consists of 52 attributes. This dataset consisted of a wide variety of columns and rows, including variables like age, gender, the percentage of the county below poverty levels, and much more. Given the plethora of options, we decided to isolate four main demographics: age cohorts, racial/ethnic identities, income/economic levels, and education levels as these categories often are key identifiers of individuals. The second dataset used was the 2016 presidential election results dataset, which has county names, FIPS codes, leading candidate, 32 candidates, and total votes as columns. Another, data set is for the 2012 election which consists of county information, candidates, precincts, parties & politician ids. A voter turnout dataset was used, which compares 2012 and 2016 election total voters and estimate voting age. We used multiple datasets to combat the lack of recent and/or specific data from larger datasets, but generic *County Facts*. After combining multiple datasets, we were left with unnecessary columns, mostly demographics that we were not interested in, for example, postal codes, persons per household, and the total number of firms. To combat this, we created new data frames as needed

to isolate and perform data analysis testing on the specific demographic.

Data Preprocessing

We used the VIM, tidyverse, and MICE libraries in R to identify missing values and clean the datasets. For the wide votes dataset, the data was melted to reduce the 32 candidate columns to 3, which were candidate, FIPS code, and votes. This dataset and the *County Facts* dataset were subsetted into different data frames with separate demographic variables. For instance, a race data frame was created. For all other datasets, renaming columns and conversion of column types from string to ints or from whole numbers to percentages was performed.

Methods and Tools

For this predictive analysis, the main R libraries (See Appendix) used for the statistical calculations include the general stats library, carets, and rpart for both classification and regression. Libraries used for visualization include plotly and ggplot2. Lastly, we used VIM and MICE to identify missing values and clean the data.

Logistic and Linear Regression

Regression is a statistical method used to determine the relationship between an independent response variable and the dependent regressors. For linear regression, this exact structure holds true. However, logistic regression takes a two-outcome response variable, sometimes categorical. With that said, linear regression will be used to evaluate numerical demographics and logistics for a regression of candidates versus numerical demographics.

Decision Tree

A classification tool which uses association rules to predict an outcome based on the data. It forms new nodes or leaves based on rules and splits based on boolean statements, such as true and false. In this case, for numerical data, it will split on values greater than, less than, or equal to the population, votes, and candidate values. The classes defined are candidates, Hillary Clinton and Donald Trump. The variables include the racial/ethnic groups, income, and education data are also evaluated.

Stacked Bar charts

A standard bar chart visualizing the percentages of support for both Clinton and Trump, using exit poll data.

County maps: Choropleth and regular

The choropleth map displayed the intensity of a numerical value in a given region. For this analysis, a map of U.S. counties and racial demographics are used, as well as candidate support is displayed. For a display of county's wins, a regular county map will be produced. This will provide a visual representation of the relationship between candidate support and the different racial groups.

Results and Analysis

Logistic Regression

From the logistic regression results, we concluded that Asians, Blacks, and Non-Latino Whites had the greatest impact on this election. This conclusion was made based on normal distribution and p-value examination.

The Blacks and Asians had the highest z-values for the racial demographics with Blacks at 5.471 and Asians at 5.440, and both had a normal distribution. These results from Blacks and Asians consistently voting together. From exit poll samples, we concluded that Clinton won counties where Asians and Blacks were the majority about 72.5% of the time. Along with all of this, the p-values for both races was below 0.05, signifying significance of the coefficient estimates. For Non-Latino Whites, the z-value yielded a Z-score of 0.71%, showing that it is within less than one standard deviation of the mean and it is very close to a normal distribution.

Decision Tree

The logistic regression narrowed the racial groups down between the three listed below. So, using the inference that Whites would be the determining racial group based on population, a decision tree was formed with Clinton as one class and Trump as the other. The split values were based on population percentages. On the right, two leaf nodes are shown. The red one shows a prediction which shows that given a sample of a county's population, if Non-Latino Whites are 50% or more of that population, Trump is expected to win that county 91% of the time. The second node shows that when the white population is less than 50%, Clinton support rises to 93%. Note that Non-Latino Whites take up 63.7% of the U.S. population and that Blacks and Asians together take up only 19.5%. The bottom value for each node shows how often the association rules are true and for Trump it is true for 89% of the 3,007 counties and only 3% is it true for Clinton. This would account for Trumps win of office.

Figure 1: Decision tree leaves of the candidate

The choropleth maps below confirm the conclusions of the decision tree. It is clear that wherever the White population is less than 50%, Donald Trump loses the county.

Figure 2: Choropleth of the Candidate support (%)

Figure 3: Choropleth of the White (Non-Latino) population support

The other demographics that we looked at, including median income levels, education attainment, and age which were all inconclusive.

Figure 4: Stacked bar plot of voting percentage by income

Income level p-values were not significant enough to be considered, therefore, income was not a sufficient predictor of the presidential election. In Figure 4, it is illustrated that there was not a clear definitive trend between income levels and presidential elects.

Figure 5: Stacked bar plot of voting percentage by education

Similar to median income levels, when testing for a linear regression model of education, there was not a clear trend. Again, p-values were insignificant and there was a similar divide of presidential votes between the education categories. This yielded education as a demographic that was not indicative of the presidential election.

To see how age affected voting preferences, we looked at a census dataset that described the breakdown of each county's population by race, sex, age, and year. By looking at the age trend of the county (i.e. how skewed the age data is toward either older or younger residents), we can get a general feel for the "age" of the county. Figure 6 plots the relationship between the trend of the county and Clinton's margin of victory. In general, the younger the population is the greater Clinton's margin of victory. However, this correlation is not strong, with a predictive power of only 13%. As a

result, age was not a significant predictor of vote preference.

Figure 6: Scatter plot of Age and presidential votes

Another factor towards a candidate's success is how a demographic turns out to vote. While Clinton performed marginally better among young people, those same young people are less likely to turnout to vote. Figure 7 shows the relationship between a county's "age" and the turnout among its voting age population. In general, the younger the county was the lower its VAP turnout. However, this relationship wasn't strong, having a predictive power of only 22%.

Figure 7: Scatterplot of Age and Voter Turnout

Conclusions

From our findings, Asians, Blacks, and Non-Latino Whites were the most impactful to the 2016 presidential election. Thus, the race demographic was the most predictive of the winning presidential elect. Hillary Clinton won a large majority of the minority vote, however, President Donald Trump won with the support of Non-Latino Whites in counties where electoral votes contributed vastly to his win. Based on our results, in the next presidential election, candidates should focus on communities with high populations of Non-Latino Whites, Blacks, and Asians. This would provide the candidate with the greater number of county wins and electoral votes, leading to an election win.

Contributions

Aaron: Linear regression, age analysis, Data preprocessing

Bria: Logistic regression, stacked bar plots, maps, decision trees and the descriptions of abstract, introductions, preprocessing and methods descriptions, Data preprocessing, Race analysis and results.

Dominique: Education analysis, editing, and proofreading

Tajah: The description of datasets and features, Income analysis, and conclusions.

Acknowledgments

Our team would like to thank both Kaggle and GitHub (see Appendix) for providing public datasets for students to work from. This includes all contributors of the data and publishers as well.

References

1. Abramson, Alana. "Hillary Clinton Officially Wins Popular Vote by Nearly 2.9 Million." *ABC News*, ABC News Network, 22 December 2016, abcnews.go.com/Politics/hillary-clinton-officially-wins-popular-vote-29-million/story?id=44354341.
2. Bansal, Manju "What Data Analysis Tells Us About the U.S. Presidential Election" *MIT Technology Review,* MIT Technology Review, 27 October 2016, https://www.technologyreview.com/s/602742/what-data-analysis-tells-us-about-the-us-presidential-election/
3. Chalabi, Mona. "Who Are the Three-Quarters of Adult Americans Who Didn't Vote for Trump?" *The Guardian*, Guardian News and Media, 18 January 2017, www.theguardian.com/us-news/2017/jan/18/american-non-voters-election-donald-trump
4. File, Thom. *The Diversifying Electorate—Voting Rates by Race and Hispanic Origin in 2012 (and Other Recent Elections)*. U.S. Census Bureau, 2013, *The Diversifying Electorate—Voting Rates by Race and Hispanic Origin in 2012 (and Other Recent Elections)*, www.census.gov/prod/2013pubs/p20-568.pdf.
5. "2016 Fox News Exit Polls." *Fox News*, FOX News Network, www.foxnews.com/politics/elections/2016/exit-polls.
6. Ingraham, Christopher. "About 100 Million People Couldn't Be Bothered to Vote This Year." *The Washington Post*, WP Company, 12 November 2016, www.washingtonpost.com/news/wonk/wp/2016/11/12/about-100-million-people-couldnt-be-bothered-to-vote-this-year/?utm_term=.370f685abcc3
7. Krogstad, Jens Manuel, and Mark Hugo Lopez. "Black Voter Turnout Fell in 2016, Even as a Record Number of Americans Cast Ballots." *Pew Research Center*, 12 May 2017, www.pewresearch.org/fact-tank/2017/05/12/black-voter-turnout-fell-in-2016-even-as-a-record-number-of-americans-cast-ballots/
8. Schramm, Michael, and University of Michigan. "How We Voted - by Age, Education, Race and Sexual Orientation." *USA Today*, Gannett Satellite Information Network, 9 November 2016, college.usatoday.com/2016/11/09/how-we-voted-by-age-education-race-and-sexual-orientation/
9. Tyson, Alec, and Maniam, Shiva. "Behind Trump's victory: Divisions by Race, Gender, Education", *FACTTANK News in the Numbers,* Pew Research Center, 9 November 2016, http://www.pewresearch.org/fact-tank/2016/11/09/behind-trumps-victory-divisions-by-race-gender-education/
10. USA Electoral Votes. *WorldAtlas*, 12 July 2016, www.worldatlas.com/webimage/countrys/namerica/usstates/electorl.htm
11. Wallace, Gregory. "Voter Turnout at 20-Year Low in 2016." *CNN*, Cable News Network, 30 November 2016, www.cn

Appendix

Libraries

- R: car, caret, corrplot, datasets, data.table, dplyr, ggplot2, graphics, knitr, maps, mice, plyr, plotly, reshape, rpart, RColorBrewer, rattle, stats, usmap, VIM, xlsx and xlsxjars
- Python:pandas, numpy, matplotlib, plotly, sklearn, and apyori

GitHub

- https://github.com/bpowell21/Data-Science-Project.git