Title: Median Household Income and 2016 Election Voting Habits in Illinois

Name: Hasan Khan

During presidential elections, people cast their votes for political candidates based off of many factors that may include political party affiliation, education, household income, race, and residential location. Instead of analyzing all elections in the United States, this project will zero in on the 2012 and 2016 election and focus on Illinois to decide whether one's voting habits can be used to predict their median household income.

This final dataset was composed with three datasets: one is based on the election results by county for 2012 and 2016 and was reported by CNN (https://www.cnn.com/election/2016/primaries/counties/ia/Dem) and the other two datasets came from the Department of Agriculture (https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/). One dataset provides information about education level per county. These education levels range between voters who have less than a high school degree to voters who have earned a bachelor's degree or higher. The other dataset has information about unemployment by county, state, and the median household income. Since the education and unemployment datasets both had a "State" column, it was possible to join them. This dataset was then joined with the election dataset and filtered out to only include Illinois. Lastly, some variables had to be adjusted for consistency. Finally, the dataset was split into training and testing datasets for model creation.

The variables that will be used for the data analysis include: the percent of people who voted Democrat in the 2012 election, the percent of people who voted Republican in the 2012 election, the percent of people who voted Democrat in the 2016 election, the percent of people who voted Republican in the 2016 election, the percent of adults with less than a high school diploma, the percent of people with a bachelor's degree or higher, and the median household income. Knowing that voting factors may change over the years, it's important to consider what effect these changes may have on a person's vote and if their household income was affected as a result. While this dataset may only have the median household income of 2018, this analysis will be assuming that income has not been majorly affected since 2018.

Now that the variables involved in this data analysis have been introduced, let's begin discussing the figures generated from the analysis. To start, let's examine **Figure 1. Figure 1** is a scatterplot with two regression lines that is displaying the relationship between the percent of people who voted Democrat in 2012 and the percent of votes in 2016. The blue regression line characterizes the percent of people who voted Democrat in 2016 and the red regression line represents the percent of people who voted Republican in 2016. Each data point represents a specific county's votes in Illinois in the 2016 election. Percentages range between 0.2 to 0.7 on the x-axis and 0.2 to 0.8 on the y-axis. It's not surprising that there is generally a positive correlation between the percent of people who voted Democrat in 2012 and the percent of people who voted Democrat in 2016 and negative correlation with the percent of people who voted Republican in 2016. It can be assumed that these patterns are consistent because some people may not have voted in the election due to candidate.

On the other hand, **Figure 2** is showcasing a scatterplot of two components of PCA on a transformed subset of the data, which uses the percent of people with less than a high school diploma, percent of people with a bachelor's degree or higher, the percent of people who voted Democrat in 2012 and 2016, the percent of people who voted Republican in 2012 and 2016, and the median household income. It uses a color map to differentiate between household incomes. With the wide clustering of points on the scatterplot, it's evident that there is little relationship between both components of the transformed subset of data.

Lastly, **Figure 3** is a horizontal bar plot that uses the coefficients on the y-axis as predictors in a linear regression model to try and predict median household income. With an average cross-validation score of approximately 0.414 on 8 folds, it's clear that this model is a poor choice when trying to compute median household income. Adding other predictors or using a Pipeline did not change the score significantly. Looking to the bars on **Figure 3**, the percent of people who voted for Democrats and Republicans in 2012 have the largest coefficient values (657327 and 54228), which indicates their influence.

To conclude, median household income may be weakly influenced based on the percent of Democratic and Republican votes in the 2012 election. While most of the party votes remained consistent between 2012 and 2016, the percent of Republican votes in 2012 ultimately has the greatest effect in predicting median household income.
I'd like to thank Professor Tyler Harter, Ryan Montsma, and Rohit Menon for their assistance.

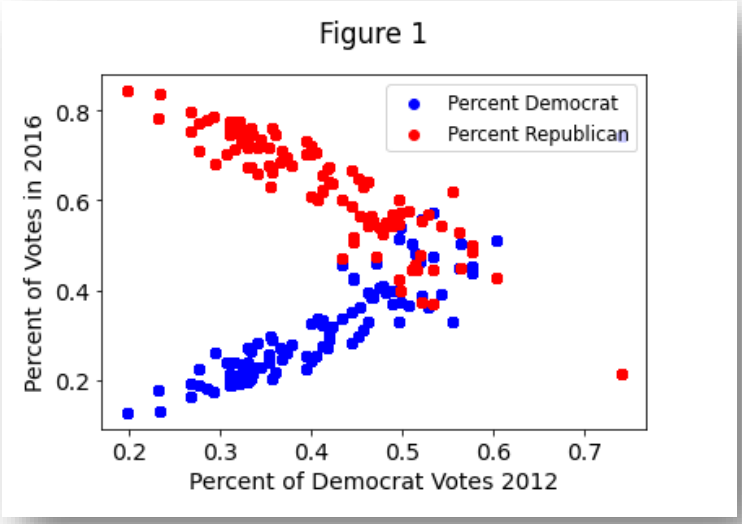
Figure 1
Percent Democrat
Percent Republican
Percent of Votes in 2016
0.2
0.4
0.6
0.8
0.2
0.3
0.4
0.5
0.6
0.7
Percent of Democrat Votes 2012

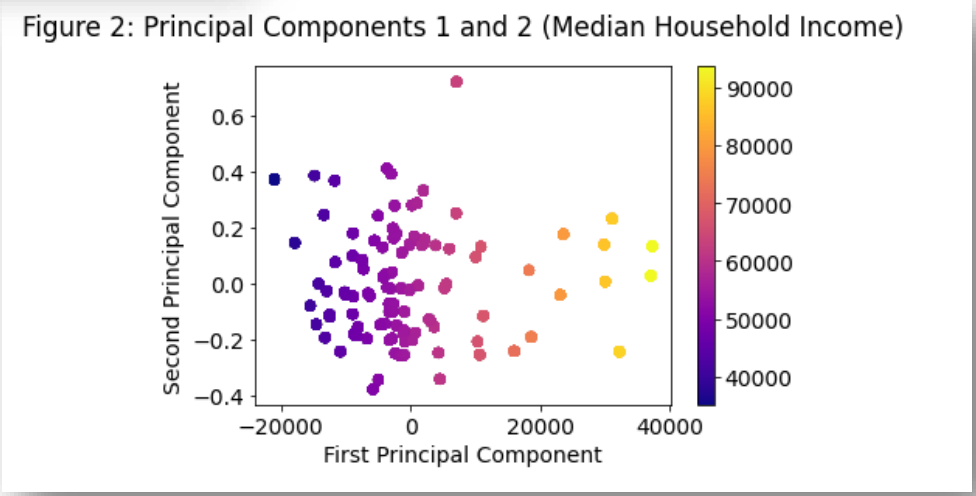
Figure 2: Principal Components 1 and 2 (Median Household Income)
Second Principal Component
0.6
0.4
0.2
0.0
−0.2
−0.4
−20000
0
20000
40000
First Principal Component
90000
80000
70000
60000
50000
40000

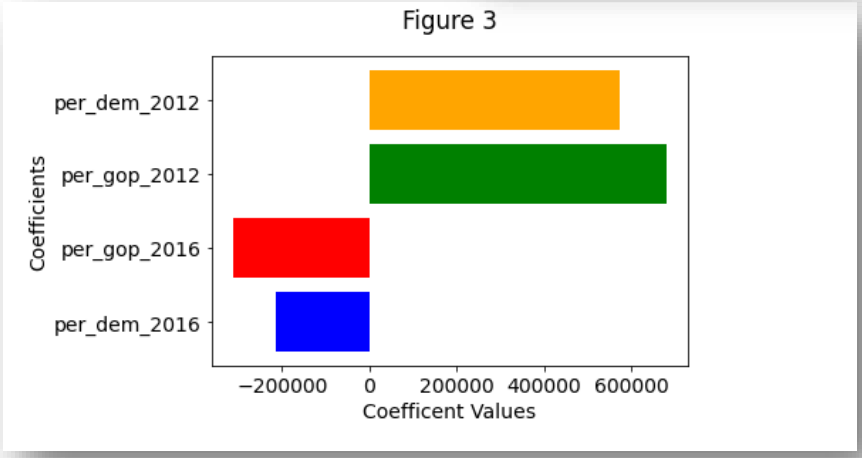
Figure 3
per_dem_2012
per_gop_2012
per_gop_2016
per_dem_2016
Coefficients
−200000
0
200000
400000
600000
Coefficent Values