REGULAR ARTICLE

Demographic indicators of voter shift between 2016 and 2020 presidential elections

Andrew Y Wang*

Correspondence:
wang22a@ncssm.edu
North Carolina School of Science and Mathematics, 1219 Broad St., 27705 Durham, NC
Full list of author information is available at the end of the article
*NCSSM Online Program

Abstract

Political pundits and analysts have long studied the demographic shifts and their impacts on the electorate, and with the results of the 2020 presidential election, this opportunity for analysis comes again. This study analyzes the demographic factors which indicate voter shift between the 2016 and 2020 presidential elections. First, voter shift was calculated based on the differences in winning margins by county for the 2016 and 2020 presidential elections. Then, Principal Component Analysis is applied to the election and demographic data for thousands of counties across the United States to determine the best demographic factors for showing voter shift. This calculation indicates a general Democratic shift in the last four years. Results of PCA show that total population was the most important factor for voter shift. The other three most important being private-sector workers, workers who drive, and voting-age citizens.

Keywords: US election; voter shift; Principal Component Analysis; Demographics; Computational politics; R

1 Introduction

The United States electorate has gone through massive demographic shifts in recent years and has grown far more diverse. Because of this, it is important to maintain an understanding of which demographic factors are the most important in voter shift. 2020 has been an unprecedented year for many reasons. Not only because of the global pandemic and massive amounts of mail-in voting but also because of the record turnout seen, the highest in over a century [10]. Electorate shifts have been subject to much research between past elections due to the insight they provide on the United States population and the trends they show about communities across the United States. The results of the 2020 presidential election give an excellent opportunity to analyze the shifts of the last four years.

In past election cycles, the role of minorities has become more important as their turnout rates have increased and political parties have recognized the importance of targeted messaging towards these groups. Minority groups have commonly tended to favor Democrats, due to the anti-immigration policies commonly seen from Republican candidates [12]. Since the year 2000, the Democratic vote percentage among minorities has risen, accounting for a majority of votes in most minority groups [11]. In 2000, around 65% of the Latino electorate voted Democratic in the presidential election, but by 2012 this number had risen to 73%, representing an overwhelming majority of Latino votes for Democratic candidates [11]. A similar trend was seen for Asian Americans, going from 57% Democratic in 2000 to 73% in 2012, and for

African Americans, going from 90% Democratic in 2000 to 93% in 2012 [11]. Therefore, race is an important demographic factor in voter shift. Based on past trends, as the role and impact of minority populations increases, the votes for Democratic candidates can also be expected to increase.

The urban and rural divide has been another important component in understanding voter shift, as urban areas tend to vote Democratic while rural areas tend to vote Republican. In 2018, polling showed that 64% of voters in urban areas supported Democratic policies, while only 38% of voters in rural areas supported Democratic policies [13]. The partisan difference in rural and urban areas is widening as well. In 2012, Democrats had a 5% margin over Republicans in urban areas, but by 2018 this margin had increased to 17% [13]. This is also reflected in rural areas, in which the Republican margin grew from 29% to 38% between 2012 and 2018 [13]. As counties across America become more urban, it is hypothesized by many political pundits that Republican-dominated states, such as those in the Midwest, will become closely contested as they urbanize [14]. This hypothesis was proven true in the 2020 presidential election in Georgia, where the electoral college votes of the state were sent to the Democratic presidential candidate for the first time in over two decades. Therefore, urban and rural areas are another large component in the demographics of voter shifts.

While much of past research surrounding demographics and voter shift has been focused on particular demographic factors, such as race, this study looks at demographic factors on a broader scale, analyzing over thirty demographic. This study aims to identify the turnout shift between the 2016 and 2020 presidential elections as well as the demographic factors which are suitable indicators for this shift. An initial set of demographic and election data by county was procured and used as a baseline for both shift and demographic analysis. Basic mathematical calculations showed the magnitude of voter shifts across various counties and Principal Component Analysis was performed on the data to yield the most important demographic factors. These data were merged with other datasets to allow for accurate plotting and altered through a variety of methods to provide plots at national and state-level scales. The primary contribution of this study is to provide the top demographic indicators of voter shift. A secondary contribution of this study is to highlight voter shifts of the last four years.

2 Dataset(s) and Preparation

The primary dataset used was drawn from Kaggle, called "Election, COVID, and Demographic Data by County" and was obtained from https://www.kaggle.com/etsc9287/2020-general-election-polls [4]. This dataset combined election data by county with demographic factors from the 2017 American Community Survey 5-year estimate [3, 5]. This dataset initially consisted of 51 columns, describing the voting trends and demographics of each United States county. This dataset was specifically chosen because of its combination and variety of data, as it had combined datasets from various sources and kept formatting consistent for ease of access. Columns 1 - 3 give a number, a name, and a state, in the form of an abbreviation, to the county. Following this, columns 4 - 8 describe the results of the 2016 election, giving the percentage of votes each candidate received, total votes, and votes for each candidate (as a value instead of a percentage). Columns 9 - 13 describe the same things, but for the 2020 election. Following this, columns 14 and 15 give the latitude and longitude of the county. All other columns (16 - 51) describe demographic factors of the county, regarding subjects such as race, employment, income, transportation, and COVID-19. Prior to any alteration, the dataset consisted of 4,868 rows and 51 columns. Below is a table that gives a structural overview of the initial dataset, with the three main column groups outlined.

Rows	Type of data	2016 Election	2020 Election	Demographics
1 - 3110	County	Yes	Yes	Yes
3111 - 4658	Extra Districts	No	Yes	No
4659 - 4689	2020 Alaska election data	No	Yes	No
4696 - 4838	Unassigned values	No	Yes	No
4839 - 4868	Alaska demographic data	No	No	Yes

There are many extra districts and unassigned values in this dataset, which are not useful because they are missing data and not at an accurate scale of analysis. Because they are not counties, they were missing all demographic data, as the American Community Survey only recorded data for counties. The extra districts therefore had to be removed from the dataset, as they would be problematic in analysis and mapping. This cleaning was generalized to simply remove any rows with missing data. However, the data for Alaskan counties were spread out over hundreds of rows due to the name-recording discrepancies at the time of recording. Alaskan counties were split into two groups, one with demographic data and one with 2020 election data. Therefore, the data for Alaskan counties had to be condensed prior to the removal of empty data. Furthermore, presidential election data from 2016 for Alaskan counties was not present in the original dataset and had to be found and added in [6, 7]. Data for Alaskan counties were then combined into a complete dataset, although a few counties were still missing due to naming discrepancies. With the data for most counties present, missing values could then be removed from the data. Removal of rows with missing data cut the dataset from 4,868 rows to 3,039 rows. While there were some counties missing from the end product, 3,039 of 3,110 counties is still enough to run analysis on.

Prior to Principal Component Analysis the demographic data needed to be standardized, as certain columns were recorded as percentages while others were as numbers [3]. This inequality would affect the component analysis and therefore all the percentage data had to be converted to numerical data so that analysis could happen at an even scale. This procedure was completed by multiplying the percentage present by the total population of the county and this provided a set of standardized demographic data. Finally, packages were imported into the model. This model was constructed in the programming language R using the RStudio development environment. Packages used with the model were maps, usmap, gridExtra, and tidyverse. All code used in the construction of the model and any figures in the paper can be found at the end of the paper.

3 Methods and Modeling

In order to analyze the voter shift, margins first had to be calculated based on voting data for both presidential elections. Three new columns were added to the data to represent margins and shifts: "margin2016", "margin2020", and "shift". Columns "margin2016" and "margin2020" are calculations of winning margin in each county by-election. However, these two margin columns do not follow traditional methods of margin computation, as it was necessary for them to not only indicate the degree of the victory, but also the party for whom the victory was for. Instead of taking the absolute value of the difference, as is traditionally done, margins for the two columns were simply the difference between the percentage of Democratic votes and the percentage of Republican votes as $Margin = Democratic - Republican$ [8]. Therefore, margin values ranged from 1 to -1, in which a number closer to 1 indicates a strong Democratic win, while a number closer to -1 indicates a strong Republican win. The "shift" column was then calculated based on these two margin columns, as it would indicate a strengthening or weakening of a party's presence in a county. The "shift" column was represented from a scale of 2 to -2 and calculated as $Shift = margin2020 - margin2016$. A value closer to 2 represented a strong Democratic shift, while a value closer to -2 represented a strong Republican shift. The drawback of the "shift" column is that it cannot portray whether or not a particular county switched parties, but it is valuable for looking at the trends over time. Below is a table that shows the added columns and their meanings.

Added Column	Value Range	Meaning
margin2016	1 to -1	Positive: Democratic win Negative: Republican win
margin2020	1 to -1	Positive: Democratic win Negative: Republican win
shift	2 to -2	Positive: Area got more Democratic Negative: Area got more Republican

To analyze the demographic factors which indicated such voter shifts, I chose to conduct Principal Component Analysis (PCA) on the demographic data. PCA is a dimension reduction technique that emphasizes variation in bringing out the patterns of a dataset. It finds the relationships between columns of variables and produces Principal Components (PC) which reflect the importance of the variables. For each variable (column) in a dataset, a PC is produced. Earlier PCs, such as PC1 or PC2, generally contain the most variance, and by finding the factors which make up those PCs you can find the most important variables [9]. Conducting PCA on the data would yield the demographic factors which contribute the most, thereby providing indicators for electorate shift. The demographic factors with the highest variance would therefore be the most important in determining electoral shift, as they are responsible for the most variation in the data. It was decided that PCA would be conducted on three scales: the United States at large, all states which switched parties between the 2016 and 2020 presidential elections, and each individual state in the aforementioned group. To make this process simpler I created a function which would conduct this analysis, called "customPCA". This function performs PCA on the demographic data for a particular scale, calculates variance, outputs variance as a scaled bar-plot, and creates a graph to show which demographic factors are the most important. I ran seven total PCAs using this function, based on the aforementioned scales.

Finally, a graphical representation of the analyzed data needed to be produced. The usmaps package was used extensively through in this process to create maps of various scales and sizes [1]. However, the existing election/demographic data needed to be combined with a county dataset from the usmaps package in order to plot data correctly. This was because the usmaps plot_usmap function needed to reference Federal Information Processing Standards (FIPS) values for counties in order to plot. Combining the two datasets was much more difficult than initially anticipated, as county names from the usmaps county dataset often had an identifying suffix, such as "county", "parish", or "borough", at the end, and would not directly merge our own data. This meant that the county data containing the FIPS codes had to be cleaned of all suffixes before being matched with the existing data. The end result was a new dataset of the mutual counties, in which each county now has a corresponding FIPS code for plotting. This new dataset had 3031 rows. With this new data, plots could be made to represent all counties in the country. This could also be scaled down, to see certain states or regions of the country. In order to make data trends at the state level clearer, the original dataset was also grouped by state and piped into a new dataset, where cell values were the summed up the values of counties based on state abbreviations. After merging with a corresponding state dataset in the usmaps package, FIPS codes were again matched to observations, and plots could then be made for states.

4 Results

To see the shifts in voter preference between the 2016 and 2020 presidential elections, I used the appended columns of "margin2020", "shift", and "margin2016". The margin and shift data are grouped by state and then put into a map plot, which created the figure below. The darker the shade of blue, the more Democratic a state is. The darker the shade of red, the more Republican a state is.

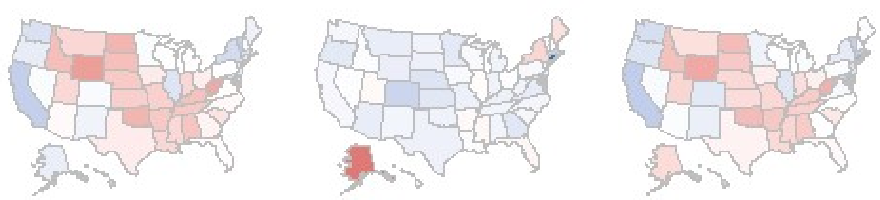

Figure 1 2016 vote margins, voter shift, and 2020 vote margins

Based on the plotting for voter shift, it is evident that most states had a Democratic shift between the 2020 and 2016 presidential elections. The average shift between the two presidential elections on a state scale was 3%. While this statistic disregards the specific population of each state (meaning all states have the same weight), it gives a general idea of the opinions of voters across the United States. This Democratic shift between presidential elections also resulted in five states moving from a Republican win in 2016 to a Democratic win in 2020. These five states were Arizona, Wisconsin, Michigan, Georgia, and Pennsylvania. While the Midwest remains strongly Republican, the flipping of these five states allowed presidential candidate Joe Biden to obtain the needed 270 electoral college votes to become the president-elect. Furthermore, Democrats also made inroads into Southern states, such as North Carolina, Georgia, and Florida, and the flipping of Georgia marked a major milestone for Democratic progress in the South. Based on this data, it seems that many previously strong Republican states may become more competitive for Democrats.

While a majority of states shifted Democratic between the two aforementioned presidential elections, notable exceptions to this statement include states like New York, Alaska, and Maine. A Republican shift in a populous state like New York was rather unexpected, as urban areas tend to have better voting margins for Democrats. The states with the largest Democratic swing were Connecticut and Rhode Island, both with swings of more than 30% in favor of the Democrats. However, most states which shifted Democratic had shifts of between 6-3%. Regarding states which shifted Republican, Alaska is the only state with a Republican shift of more than 10%, the next closest being new York with a shift of 8%. Out of 50 states, only 10 states had a Republican shift, with most of these shifts being around 1-2%. While an overwhelming majority of states may have shifted Democratic, many of the shifts in states were tiny, with 14 states have a shift of less than 2%. Of those fourteen, six states had a shift of less than 1%, those states being Idaho, Nevada, Ohio, Vermont, New Hampshire, and Illinois. However, there does not seem to be a pattern between states with small voter shifts.

Regarding the demographic factors which can indicate such voter shift, Principal Component Analysis (PCA) was conducted on the demographic data and analyzed at various scales for results. Below is a table which shows the top 4 demographic indicators at various scales in the United States, with accompanying indicator explanations.

Table 1 The top four demographic indicators in PC1 at various scales

Scale of analysis	#1 Indicator	#2 Indicator	#3 Indicator	#4 Indicator
United States	Total population	Private sector	Driving	Voting age citizens
All flipped states	Total population	Private sector	Driving	Voting age citizens
Arizona	Total population	Private sector	Driving	Voting age citizens
Wisconsin	Total population	Private sector	Driving	Voting age citizens
Michigan	Total population	Private sector	Driving	Voting age citizens
Georgia	Total population	Private sector	Driving	Voting age citizens
Pennsylvania	Total population	Private sector	Voting age citizens	Driving

- **Total population:** The total population refers to all living people in a county.
- **Private sector:** The private sector refers to the total number of people working in a job which is not controlled by the government.
- **Driving:** This is the total number of people who drive in any vehicle as their main source of transportation to/from work.
- **Voting age citizens:** This is the total number of U.S. citizens over the age of 18.

Based on the PCA results, it seems most areas of the United States share the same indicators for voter shift. This was an unexpected result, as I had hypothesized that there would be more diversity in the demographic factors which could indicate shift, but the results of the PCA analysis have shown that the United States is very homogeneous in this aspect. Excluding Pennsylvania, all flipped states shared the same top demographic indicators. The only difference Pennsylvania had was a swapping of the third and fourth indicators, where Pennsylvania's third indicator was voting-age citizens and the fourth indicator was the number of workers who drive. However, it is notable that these indicators share some similarities. Total population certainly has some influence on all other indicators, as areas with higher populations would also have more workers, more cars, and more voting-age citizens.

While the demographic indicators may have been extremely similar, the percentage compositions of the Principal Components (PCs) varied among scales. Table 2 indicates the percent composition of each PC. Scales at which the first PC has a high percentage indicates the demographic indicators mentioned in Table 1 play more of a role. This is the case in states like Arizona and Michigan, where PC1 makes up more than 99% of the variance. On the other hand, in states like Wisconsin and Georgia, while PC1 percentages still portray the majority of the variance, PC2 plays an increasing role. Wisconsin serves as a major outlier in the PCA calculations, with much much variation resulting from PC2 (42.9%). Looking at the United States as a whole, we can see that PC1 makes up 86.2% of the variance, while PC2 makes up 9.2%. This indicates that our aforementioned demographic indicators (seen in Table 1) cause a majority of the variance across the United States, although they may be weaker in states such as Wisconsin. Demographic indicators which made up PC2 were primarily race, with the number of white people usually being the most important variable in PC2.

Table 2 PCA percentage table

Scale of analysis	PC1 %	PC2 %
United States	86.2%	9.2%
All flipped states	88.2%	8.9%
Arizona	99.1%	0.8%
Wisconsin	52.4%	42.9%
Michigan	99.2%	0.4%
Georgia	70.9%	24%

As seen in the PCA results table, the best demographic indicator for voter shift is total population. This is a very reasonable result, as more urban areas, which have large populations, tend to be more Democratic, while rural areas, which have smaller populations, tend to be more Republican. This is likely due to Democratic policies generally favoring higher governmental services and spending, which corresponds to the integration and close-proximity of services seen in many urban areas. It is easier for those in cities to see results of government spending, and therefore be more inclined to support its increase. On the other hand, in areas with lower population and lower population density, it is harder for the government to spread out services in an adequate manner, and therefore those in lower populated areas could favor Republican policies of tight spending more. In the figure below, voter shift is put side-by-side next to state population. This is the same voter shift map as the one used in Figure 1, but the state population map is new. Darker areas on the state population map represent lower populations, while lighter areas represent a higher population.

The population of a state seems to be inversely proportional to the level of voter shift in that state. As such, states with lower populations tend to have larger voter shifts. A few examples of this relationship include states such as Alaska, Colorado, Connecticut, and Rhode Island. On the other hand, states with large populations, such as California, Florida, and Texas, all had rather small voter shifts. From a statistical viewpoint, this relationship is somewhat logical, as a single person's decision

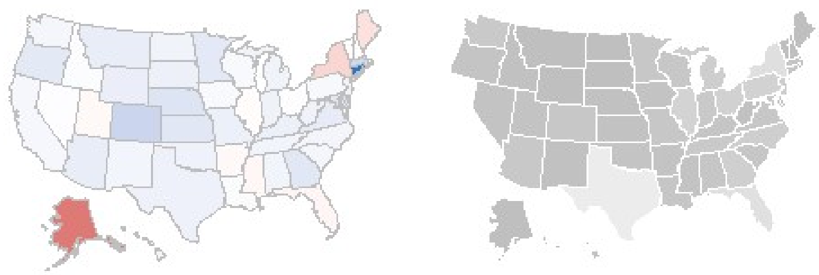

Figure 2 Voter shift in comparison to population (by state)

in a state with low population would have more impact than a single person's decision in a state with high population. As such, an equal amount of people changing sides in both Rhode Island and California would result in vast differences in the voter shift. Below is a 2x2 graph which displays the populations and the voter shifts for Wisconsin and Michigan. While the legend for population is not very clear, the general idea is that darker areas have more population. I would normally keep the color scaling consistent, but in this case, most counties in both states had relatively low population per county. Therefore, keeping the color scaling consistent would've been aesthetically unappealing and make visual comparison challenging.

To further illustrate the impact percentage values of PC1 and PC2 play in the accuracy of determining demographic indicators, I have plotted two flipped states with vastly different PC percentages. In Michigan, PC1 made up 99.2% of the variance, with PC2 making up 0.4% of the variance. On the other hand, in Wisconsin PC1 made up 52.4% of the variance, with PC2 making up 42.9% of the variance. This indicates that the top demographic factors in PC1, which are very accurate in indicating voter shift in Michigan, are far less accurate for Wisconsin. In Figure 3, we can see this difference illustrated. The major population center for Michigan is centered around Detroit and Ann Arbor, with there being considerably lower population in the rest of the state. When comparing the population to the voter shift, it is clear that areas with lower populations experienced higher levels of voter shift. The darkest areas of the voter shift map are around the counties with low population, and most counties in North Michigan, which have low populations, experienced a large voter shift. While the heavily populated areas of Southeast Michigan did experience some voter shift, it is not on the same magnitude as the lower populated counties. Therefore, we can say that population is a good indicator of voter shift in Michigan.

However, this is not the case for Wisconsin. As seen from the PC percentages, PC1 accounts for far less in Wisconsin than it does in Michigan. Milwaukee, one of Wisconsin's major population centers, is also a site of large voter shift. The trends seen between population and voter shift are not reflected as strongly in Wisconsin, with both high and low population counties experiencing large voter shift. That being said, population is likely still the best indicator of voter shift in Wisconsin, but there are many more indicators that play large roles in this calculation, as evidenced by Wisconsin's PC composition. While Wisconsin and Michigan may border each other and both be states which flipped parties, the demographic factors

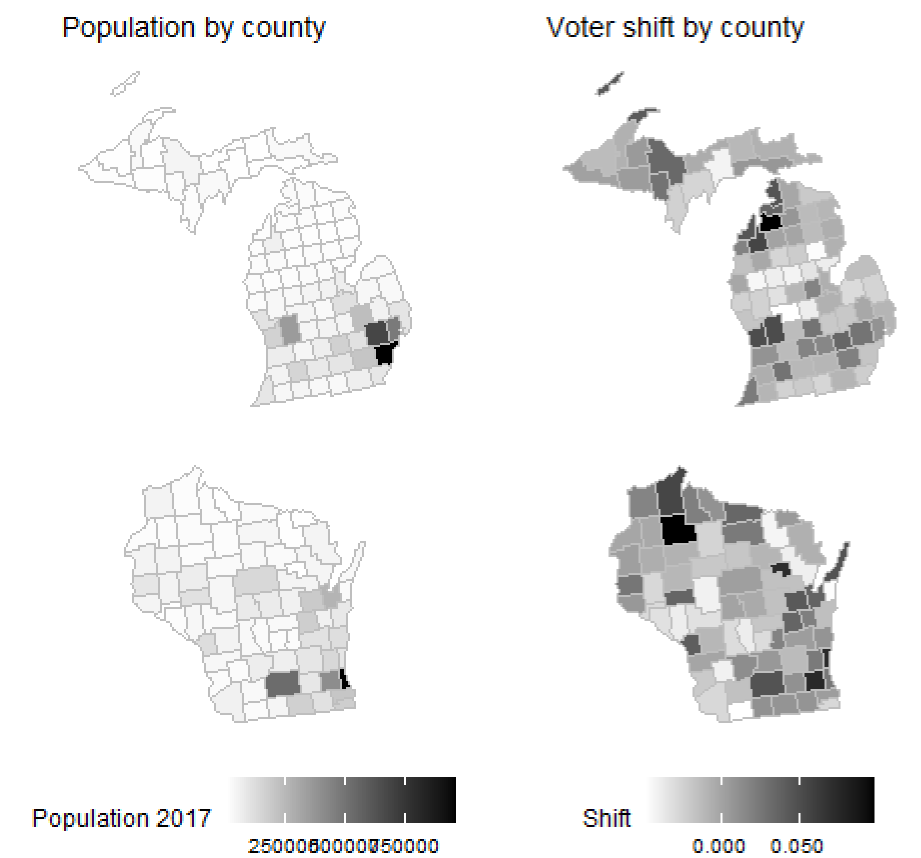

Figure 3 Voter shift in comparison to population (by state)

which indicate voter shift in these states are vastly different. However, based on this comparison the conclusion can be reached that the demographic indicators in table 1 are accurate indicators of voter shift in the United States as a whole.

5 Conclusion

In this paper, I analyzed a set of demographic and election data to determine what kinds of voter shifts occurred and what demographic factors could indicate such a shift between the 2016 and 2020 presidential elections. To find out voter shift, I used basic arithmetic and plotting to visualize areas of large change. These calculations showed that most states became more Democratic over the last four years. This conclusion was also reflected in the 2020 presidential election, as five states flipped from Republican (in 2016) to Democratic (in 2020). To find the demographic factors which could indicate voter shift, I used Principal Component Analysis at various scales and compared the results to voter shift data for accuracy. While indicators may vary by state, the top 4 demographic indicators of voter shift over the entirety of the United States are total population, private-sector workers, driving workers, and the number of voting-age citizens.

Some limitations remain to be addressed in future research. The demographic data was from 2017, therefore various factors, such as the number of employed citizens, are likely inaccurate. Furthermore, there were around 70 missing counties from the data, and complete analysis for states like Alaska and Vermont was not possible because of recording discrepancies in the naming.

The results of this analysis raise interesting questions about the future of voting trends in the United States. The impact of areas like urbanization and nationalization remain large topics, as do their impact on voter shifts in the future. It is entirely possible that the Democratic shift seen between the 2016 and 2020 presidential election was not impacted by demographic factors, but rather voters' opinions of the administration. The application of electorate demographics to voter outreach and engagement will likely become more salient in the future, as they could play a key role in the understanding what voters care about and how to get voters motivated to turn out.

Availability of data and materials
The data for this work was obtained from https://www.kaggle.com/etsc9287/2020-general-election-polls. It was also obtained in part from the R usmaps package.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
This paper is solely the work of the author. All references are included in the bibliography and are cited appropriately.

Funding
Funding for this program is provided by the North Carolina School of Science and Mathematics, the University of North Carolina General Administration, and the General Assembly for the State of North Carolina.

Acknowledgements
The author appreciates the efforts of the American Community Survey, the U.S. Census Bureau, and FiveThirtyEight in the collection of data and of Kaggle user Ethan Schacht for creation of the dataset used. The author also appreciates Mr. Robert Gotwals for his instruction and role in teaching data analysis and modeling.

References
1. Lorenzo, Paolo Di. "2. Mapping the US." Https://Cran.r-Project.org/, 7 Oct. 2020, cran.r-project.org/web/packages/usmap/vignettes/mapping.html.
2. Auguié, Baptiste. "Laying out Multiple Plots on a Page." Https://Cran.r-Project.org, 13 July 2019, cran.r-project.org/web/packages/egg/vignettes/Ecosystem.html.
3. "SELECTED ECONOMIC CHARACTERISTICS." Data.census.gov, United States Census Bureau, 2017, data.census.gov/cedsci/table?tid=ACSDP5Y2017.DP03.
4. Schacht, Ethan. "Election, COVID, and Demographic Data by County." Kaggle, 14 Nov. 2020, www.kaggle.com/etsc9287/2020-general-election-polls.
5. Neutrino, Muon. "US Census Demographic Data." Kaggle.com, Kaggle, 3 Mar. 2019, www.kaggle.com/muonneutrino/us-census-demographic-data.
6. "Alaska Presidential Results by County, 1960-2016: Maps." Thecinyc, 2016, www.thecinyc.com/alaska-pres-results-by-county-equiv.
7. "STATE OF ALASKA - DIVISION OF ELECTIONS VOTERS HISTORY BY AGE REPORT 16GENR-2016 GENERAL ELECTION." Elections.alaska.gov, 7 Feb. 2017, 8:50:26 am, www.elections.alaska.gov/election/2016/General/VoterHistoryByAgeReport.pdf.
8. "Margin-of-Victory (MOV)." Ballotpedia.org, Ballotpedia, ballotpedia.org/Margin-of-victory(MOV).
9. Howlin, Colm. "Principal Component Analysis - A Primer." Realizeit Labs, 22 Oct. 2018, lab.realizeitlearning.com/resource/2018/10/22/Principal-Component-Analysis-A-Primer/.
10. Rabinowitz, Kate. "2020 Turnout Is the Highest in over a Century." The Washington Post, WP Company, 1 Dec. 2020, www.washingtonpost.com/graphics/2020/elections/voter-turnout/.
11. Hudak, John, and Christine Stenglein. "How Demographic Changes Are Transforming U.S. Elections." Brookings, Brookings, 15 Sept. 2016, www.brookings.edu/blog/fixgov/2016/09/13/how-demographic-changes-are-transforming-u-s-elections/.
12. Daniller, Andrew. "Americans' Immigration Policy Priorities." Pew Research Center, Pew Research Center, 30 May 2020, www.pewresearch.org/fact-tank/2019/11/12/americans-immigration-policy-priorities-divisions-between-and-within-the-two-parties/.
13. "How the Rural vs. Urban Mindset Fuels Today's Politics." U.S. News amp; World Report, U.S. News amp; World Report, 14 May 2019, www.usnews.com/news/best-states/articles/2019-05-14/demographic-shifts-in-cities-and-states-bring-political-changes-too.
14. Beyer, Scott. "The Republican Party's Urban Problem." Forbes, Forbes Magazine, 17 May 2016, www.forbes.com/sites/scottbeyer/2016/05/17/the-gops-urban-problem/?sh=7572b2244662.

Additional Files
R Code for this work

```
# Andrew Wang
# November 22
# Final project script
#
# clean up and setup
rm(list=ls()) # clean up any old stuff in R
setwd("C:/Users/hyper/OneDrive/Desktop/Desktop Folders/Programming/R/
    Assignments/Final Project") # go to this folder
#load up myfunctions.R
source("C:/Users/hyper/OneDrive/Desktop/Desktop Folders/Programming/R/
    myfunctions.R")

#library import
library(maps)
library(tidyverse)
library(usmap)
library(gridExtra)

#data import
election <- read.csv("data/dataFresh.csv")
View(election)
dim(election)

#standardize data:
election$Poverty <- round(election$TotalPop * election$Poverty / 100)
election$ChildPoverty <- NULL #no numbers for children
election$Professional <- round(election$TotalPop * election$Professional /
    100)
election$Service <- round(election$TotalPop * election$Service / 100)
election$Office <- round(election$TotalPop * election$Office / 100)
election$Construction <- round(election$TotalPop * election$Construction /
    100)
election$Production <- round(election$TotalPop * election$Production / 100)
election$Drive <- round(election$TotalPop * election$Drive / 100)
election$Carpool <- round(election$TotalPop * election$Carpool / 100)
election$Transit <- round(election$TotalPop * election$Transit / 100)
election$Walk <- round(election$TotalPop * election$Walk / 100)
election$OtherTransp <- round(election$TotalPop * election$OtherTransp /
    100)
election$PrivateWork <- round(election$TotalPop * election$PrivateWork /
    100)
election$PublicWork <- round(election$TotalPop * election$PublicWork / 100)
election$SelfEmployed <- round(election$TotalPop * election$SelfEmployed /
    100)
election$FamilyWork <- round(election$TotalPop * election$FamilyWork / 100)
election$Unemployment <- round(election$TotalPop * election$Unemployment /
    100)

election <- na.omit(election) # we want to get rid of everything without
    numbers

dim(election)

#define party colors for coloring of geographic plots
partyColors <- c("#2E74C0", "#CB454A")
```

```
#margin finding with mutate
election <- mutate(
  election,
  margin2016 = percentage16_Hillary_Clinton - percentage16_Donald_Trump,
  margin2020 = percentage20_Joe_Biden - percentage20_Donald_Trump,
  shift = margin2020 - margin2016
)

#list of flipped states
changed <- c("AZ","WI","MI","GA","PA") #only from R --> D

###PCA analysis for demographic factors``

#subset for demographics
demographics <- election[,16:50]

#PCA analysis function to save code length
customPCA <- function(data, description){
  pca <- prcomp(t(data))
  View(pca$x)
  #screeplot + plot
  plot(pca$x[,1],pca$x[,2])
  #variance calculation
  pcaV <- pca$sdev^2
  pcaV <- round(pcaV/sum(pcaV)*100,1)
  pcaV
  barplot(pcaV)
  #contributing variable plot
  pcaDF <- data.frame(Sample=rownames(pca$x), X=pca$x[,1], Y=pca$x[,2])
  pcaP <- ggplot(data=pcaDF, aes(x=X, y=Y, label=Sample)) +
    geom_text() +
    xlab(paste("PC1 - ", pcaV[1], "%", sep="")) +
    ylab(paste("PC2 - ", pcaV[2], "%", sep="")) +
    ggtitle(paste("Demographic indicators in PCA analysis -", description,
        sep="\n"))
  pcaP
}

customPCA(demographics, "United States")
## Conclusion: Total population, Private work, Drive, Voting age citizens,
    women

#as a group (all changed states)
changedStates <- filter(election, state == "AZ" | state == "WI" | state == "
    MI" | state == "GA" | state == "PA")
changedStates <- changedStates[,c(16:50)]
customPCA(changedStates, "Flipped states")
## Conclusion: Total population, Private work, Drive, Voting age citizens,
    women

#arizona
arizona <- filter(election, state == "AZ")
arizona <- arizona[,c(16:50)]
customPCA(arizona, "Arizona")
## Conclusion: Total population, Private work, Drive, Voting age citizens,
    women

#wisconsin
wisconsin <- filter(election, state == "WI")
```

```
wisconsin <- wisconsin[,c(16:50)]
customPCA(wisconsin, "Wisconsin")
## Conclusion: Total population, Private work, Drive, Voting age citizens,
    women

#michigan
michigan <- filter(election, state == "MA")
michigan <- michigan[,c(16:50)]
customPCA(michigan,"Michgain")
## Conclusion: Total population, Private work, Drive, Voting age citizens,
    women

#georgia
georgia <- filter(election, state == "GA")
georgia <- georgia[,c(16:50)]
customPCA(georgia,"Georgia")
## Conclusion: Total population, Private work, Drive, Voting age citizens,
    women

#pennslyvania
pennslyvania <- filter(election, state == "PA")
pennslyvania <- pennslyvania[,c(16:50)]
customPCA(pennslyvania,"Pennslyvania")
## Conclusion: Total population, Private work, Voting age citizens, Drive,
    women

#adding FIPS numbers to county dataset
localCounty <- data.frame(countypop)
#a bunch of replacement for matching stuff
localCounty$county <- gsub(" county", "", localCounty$county, ignore.case=
    TRUE)
localCounty$county <- gsub(" parish", "", localCounty$county, ignore.case=
    TRUE)
localCounty$county <- gsub(" borough", "", localCounty$county, ignore.case=
    TRUE)
localCounty$county <- gsub(" census area", "", localCounty$county, ignore.
    case=TRUE)
localCounty$state <- localCounty$abbr
localCounty$abbr <- NULL
#join datasets
countyFIPS <- inner_join(localCounty, election, by = c("county","state"))
countyFIPS$pop_2015 <- NULL

###countyFIPS dataset VERY IMPORTANT

#Triple graph format for shift detection (20,shift,16)
main2020 <- plot_usmap(data = countyFIPS, values = "margin2020", color = "
    white") +
  scale_fill_gradient2(low = partyColors[2], mid = "white", high =
      partyColors[1], na.value = "white", name = "Electoral shift", label =
      scales::comma) +
  theme(legend.position = "none")
main2020
mainShift <- plot_usmap(data = countyFIPS, values = "shift", color = "white
    ") +
  scale_fill_gradient2(low = partyColors[2], mid = "white", high =
      partyColors[1], na.value = "white", name = "Electoral shift", label =
      scales::comma) +
  theme(legend.position = "none")
```

```
mainShift
main2016 <- plot_usmap(data = countyFIPS, values = "margin2016", color = "
    white") +
  scale_fill_gradient2(low = partyColors[2], mid = "white", high =
      partyColors[1], na.value = "white", name = "Electoral shift", label =
      scales::comma) +
  theme(legend.position = "none")
main2016


#sample plotting for indicators - by county, whole US
sample <- plot_usmap(data = countyFIPS, values = "TotalPop", color = "gray")
     +
  scale_fill_gradient2(low = "white", mid = "gray", high = "black", na.value
       = "white", name = "Population", label = scales::comma) +
  theme(legend.position = "right") +
  labs(title = "Population by county")
sample

#sample plotting for indicators - by county, single state
singleState <- plot_usmap(data = countyFIPS, values = "TotalPop", include =
    "MI", color = "gray") +
  scale_fill_continuous( low = "white", high = "black", name = "Population
      2017", label = scales::comma ) +
  theme(legend.position = "right") +
  labs(title = "Population by county in Michagin")
singleState

#michgain example
miPop <- plot_usmap(data = countyFIPS, values = "TotalPop", include = "MI",
    color = "gray") +
  scale_fill_continuous( low = "black", high = "white", name = "Population
      2017") +
  theme(legend.position = "none") +
  labs(title = "Population by county")
#miPop
miShift <- plot_usmap(data = countyFIPS, values = "shift", include = "MI",
    color = "gray") +
  scale_fill_continuous( low = "white", high = "black", name = "Shift",
      label = scales::comma ) +
  theme(legend.position = "none") +
  labs(title = "Voter shift by county")
#miShift
grid.arrange(miPop,miShift,nrow=1)

#Wisconsin example
wiPop <- plot_usmap(data = countyFIPS, values = "TotalPop", include = "WI",
    color = "gray") +
  scale_fill_continuous( low = "black", high = "white", name = "Population
      2017") +
  theme(legend.position = "bottom") #+
  #labs(title = "Population by county")
#wiPop
wiShift <- plot_usmap(data = countyFIPS, values = "shift", include = "WI",
    color = "gray") +
  scale_fill_continuous( low = "white", high = "black", name = "Shift",
      label = scales::comma ) +
  theme(legend.position = "bottom") #+
  #labs(title = "Voter shift by county")
```

```
#wiShift
grid.arrange(wiPop,wiShift,nrow=1)

grid.arrange(miPop, miShift, wiPop, wiShift, nrow = 2, ncol = 2)

#sample plotting for indicators - by state, whole US
#pipe for group sorting
stateSet <- election %>%
  group_by(state) %>%
  summarize(
    stateMargin2016 = (sum(votes16_Hillary_Clinton)/sum(total_votes16)) - (
        sum(votes16_Donald_Trump)/sum(total_votes16)),
    stateMargin2020 = (sum(votes20_Joe_Biden)/sum(total_votes20)) - (sum(
        votes20_Donald_Trump)/sum(total_votes20)),
    totalPop = sum(TotalPop)
  )
stateSet$stateShift <- stateSet$stateMargin2020 - stateSet$stateMargin2016
stateSet$abbr <- stateSet$state
stateSet$state <- NULL
#formatting and join stuff for FIPS codes
stateSet <- inner_join(stateSet,statepop)
stateSet$pop_2015 <- NULL
#plot
stateShift <- plot_usmap(data = stateSet, values = "stateShift", color = "
    gray") +
  scale_fill_gradient2(low = partyColors[2], mid = "white", high =
      partyColors[1], na.value = "white") +
  theme(legend.position = "none")
#stateShift
statePop <- plot_usmap(data = stateSet, values = "totalPop", color = "white
    ") +
  scale_fill_gradient2(low = "black", mid = "gray", high = "white", na.value
       = "white") +
  theme(legend.position = "none")
#statePop
grid.arrange(stateShift,statePop,nrow=1)
#plot
state2020 <- plot_usmap(data = stateSet, values = "stateMargin2020", color =
     "gray") +
  scale_fill_gradient2(low = partyColors[2], mid = "white", high =
      partyColors[1], na.value = "white") +
  theme(legend.position = "none")
state2020
#plot
state2016 <- plot_usmap(data = stateSet, values = "stateMargin2016", color =
     "gray") +
  scale_fill_gradient2(low = partyColors[2], mid = "white", high =
      partyColors[1], na.value = "white") +
  theme(legend.position = "none")
state2016
grid.arrange(state2020,stateShift,state2016, nrow = 1)
```