2016 Election - Voter Demographics

Johnny Yu, Eric Cha, Amy Kuang

Introduction

For this project, we will divide the US into four regions and identify the PC values that are strongly correlated with demographic variables and use Principle Component Analysis to reduce the dimensionality of the large data set 'census'. By transforming a large set of variables into a smaller one, we are able to extract the important information from the large set. Slicing the US into four regions, we find that different regions had different PC values and variables. We were able to determine 3 different PC variables and link them to the people who were disadvantaged.

Materials and methods

Datasets

We had two primary datasets from 2016: census and election data. The census data is composed of attributes such as gender, ethnicity, transportion means, employment information and status, and income, and population count all sorted by county and state. The raw election dataset is composed of attributes such as votes for candidates, by states, county, and fips.

To preprocess our data, we took the census and election data, filtered out NA counties, and transformed some columns into more mungable data types. The state column was cleaned, and joined with the fips codes. The winner for each state was identified, and then the winner for each county through grouping and left joins. Then, the census data was cleaned by transforming all percentage variables into the same format, and combining races with a small number of people into the "Minority" column. Lastly, we performed a left join on the election and census data, creating one complete dataset.

county	fips	candidate	state	votes	total	pct	Women
autauga	1001	Donald Trump	alabama	18172	24759	0.734	51.57
autauga	1001	Hillary Clinton	alabama	5936	24759	0.2398	51.57
baldwin	1003	Donald Trump	alabama	72883	94261	0.7732	51.15
baldwin	1003	Hillary Clinton	alabama	18458	94261	0.1958	51.15
barbour	1005	Donald Trump	alabama	5454	10436	0.5226	46.17
barbour	1005	Hillary Clinton	alabama	4871	10436	0.4667	46.17
bibb	1007	Donald Trump	alabama	6738	8753	0.7698	46.59
bibb	1007	Hillary Clinton	alabama	1874	8753	0.2141	46.59

Methods

For our first task, we will determine the most prevalent voter demographics for 2016 election candidates of each region. To do so, we will utilize PCA (principal component analysis) to show the prevalent voter attributes for 4 regions of the US: South, West, NorthEast, and Midwest. Using PCA, we will observe what attributes (aka. as variables) are prevalent in voters by region. We then look at which variables contribute to a high positive PC value and variables that have strong negative correlations to a highly positive PC value. With these considerations, we will

conclude what the principal components in our PCA represent. By knowing what our principal components represent, we can then piece together an understanding of what the demographics are like for each US region.

For our second task, we will determine the most prevalent voter demographics for the 2016 election (in general - popular vote data) by utilizing PCA. By using PCA, we will attempt to show the most prominent type of voter backgrounds of the general election. Using PCA, we will observe what attributes (aka. as variables) are prevalent in voters. We then look at which variables contribute to a high positive PC value and variables that have strong negative correlations to a highly positive PC value. With these considerations, we will conclude what the principal components in our PCA represent. By knowing what our principal components represent, we can then piece together an understanding of what the demographics are like.

Results

Voter Demographics for 4 Regions of the US:

First, we made 3 loading plots for each of the 4 regions of the US.

```
## [1] "South PC 1 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
## [1] "South PC 2 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
## [1] "South PC 3 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
```

According to our South region loading plot for PC1, a positive PC1 value is strongly positively correlated with variables Income, IncomeErr, IncomePerCap, IncomePerCapErr, Professional, White. Further, with strong negative correlation to variables ChildPoverty, Poverty, Service, Unemployment, we conclude that PC1 represents the White middle-class.

According to our South region loading plot for PC2, a positive PC2 value is strongly positively correlated with variables FamilyWork, Income, IncomeErr, IncomePerCap, IncomePerCapErr, OtherTransp, Professionals, SelfEmployed, Transit, Walk, and WorkAtHome. However, there is a strong negative correlation to Citizen, Employed, Men, and Women. We conclude that PC2 represents rural/small town communities.

According to our South region loading plot for PC3, a positive PC3 value is strongly positively correlated with variables Asian, Black, Employed, Hispanic, Men, OtherTransp, Poverty, Professional, Service, Transit, Walk, Women. However, with a strong negative correlation to Drive, Production, and White, we conclude that PC3 represents working class minority groups such as women and people of color.

Based on our loading plots on PC1, PC2, and PC3 and analysis above, we summarize that the voter demographics South region of the US are: White middle-class, rural/small town communities, and working class minority groups such as women and people of color.

```
## [1] "West PC 1 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
## [1] "West PC 2 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
## [1] "West PC 3 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
```

According to our West region loading plot for PC1, a positive PC1 value is strongly positively correlated with variables ChildPoverty, Construction, Hispanic, Poverty, Production, Service, and Unemployment. Additionally, there is a strong negative correlation to variables Income, IncomePerCap, and Professional. We then conclude that PC1 represents lower-socioeconomic class Hispanics.

According to our West region loading plot for PC2, a positive PC2 value is strongly positively correlated with variables Citizen, Drive, Employed, Men, Office, PrivateWork, and Women. On the contrary, there is a notable negative correlation with variables IncomePerCap, OtherTransp, Self-Employed, Walk, and WorkAtHome. We then conclude that PC2 represents white-collar workers.

According to our West region loading plot for PC3, a positive PC3 value is strongly positively correlated with variables Construction, Drive, Native, PublicWork, and White. Additionally, there is a notable negative correlation with variables Asian, Black, IncomePerCapp, IncomePerCapErr, MeanCommute, OtherTransp, PrivateWork, Transit, and Walk. We then conclude that PC3 possibly represents blue-collar workers – more specifically those working with public works.

Based on our loading plots on PC1, PC2, and PC3 and analysis above, we summarize that the voter demographics West region of the US are: lower socio-economic class Hispanics, white-collar workers, and blue-collar workers.

```
## [1] "North East PC 1 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
## [1] "North East PC 2 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
## [1] "North East PC 3 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
```

According to our NorthEast region loading plot for PC1, a positive PC1 value is strongly positively correlated with variables Black, ChildPoverty, Hispanic, Poverty, Production, Service, and Unemployment. On the contrary, there is a notable negative correlation with variables Income, IncomePerCap, Professional, and White. We then conclude that PC1 represents specifically predominantly Black and Hispanic disadvantaged urban neighborhoods.

According to our NorthEast region loading plot for PC2, a positive PC2 value is strongly positively correlated with variables Asian, IncomeErr, IncomePerCapErr, MeanCommute, Transit, and Walk. On the otherhand, there is a notable negative correlation with variables Construction, Drive, Production, and White. We then conclude that PC2 represents working class Asians.

According to our NorthEast region loading plot for PC3, a positive PC3 value is strongly positively correlated with variables Citizen, Employed, Men, and Women. However, there is a notable negative correlation with variables IncomeErr, IncomePerCap, IncomePerCapErr, Professional, PublicWork, SelfEmployed, White, WorkAtHome. We then conclude that PC3 represents working class Americans.

Based on our loading plots on PC1, PC2, and PC3 and analysis above, we summarize that the voter demographics NorthEast region of the US are: Black/African Americanss and Hispanics from disadvantaged urban neighborhoods, working class Asians, and wworking class Americans (general).

```
## [1] "Midwest PC 1 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
## [1] "Midwest PC 2 .png"
## attr(,"class")
## [1] "knit_image_paths" "knit_asis"
## [1] "Midwest PC 3 .png"
## attr(,"class")
```

```
## [1] "knit_image_paths" "knit_asis"
```

According to our Midwest region loading plot for PC1, a positive PC1 value is strongly positively correlated with variables Black, ChildPoverty, Poverty, Service, Unemployment. On the contrary, there is a notable negative correlation with variables Employed, Income, IncomePerCap, Professional, and White. We then conclude that PC1 represents disadvantaged Black/African Americans.

According to our Midwest region loading plot for PC2, a positive PC2 value is strongly positively correlated with variables Asian, Black, Income, IncomeErr, IncomePerCap, IncomePerCapErr, MeanCommute, OtherTransp, Professional, Transit, Walk, WorkAtHome. On the otherhand, there is a notable negative correlation with variables Carpool, Citizen, Construction, Drive, Employed, Men, Production, White, Women. We then conclude that PC2 represents working middle-class people of color (predominantly Asians and Black/African Americanss) in cities.

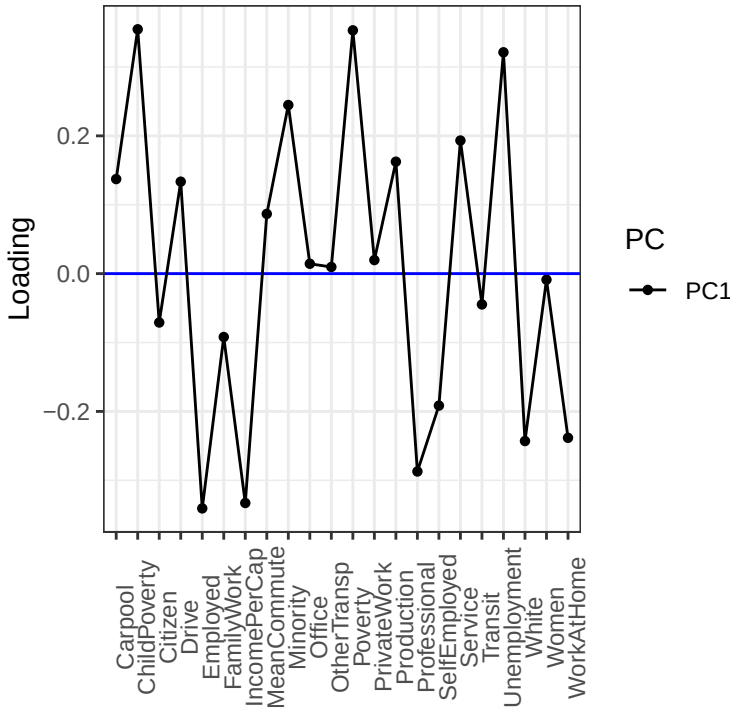
According to our Midwest region loading plot for PC3, a positive PC3 value is strongly positively correlated with variables Citizen, Employed, Men, and Women. However, there is a strong negative correlation with variables Construction, FamilyWork, PublicWork, SelfEmployed, White, and WorkAtHome. We then conclude that PC3 represents working class Americans.

Based on our loading plots on PC1, PC2, and PC3 and analysis above, we summarize that the voter demographics Midwest region of the US are: disadvantaged Black/African Americans, working middle-class people of color (predominantly Asians and Black/African Americans) in cities, and working class Americans (general).

Voter demographics for the 2016 election:

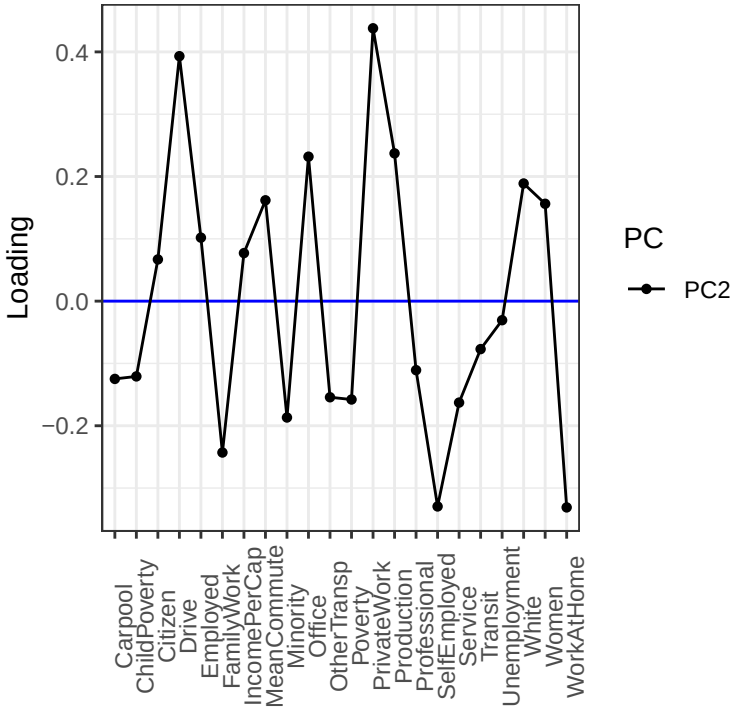
Next, we will determine the most prevalent voter demographics for the 2016 election (in general - popular vote data).

Here, we created loading plots for the first 3 principal components.

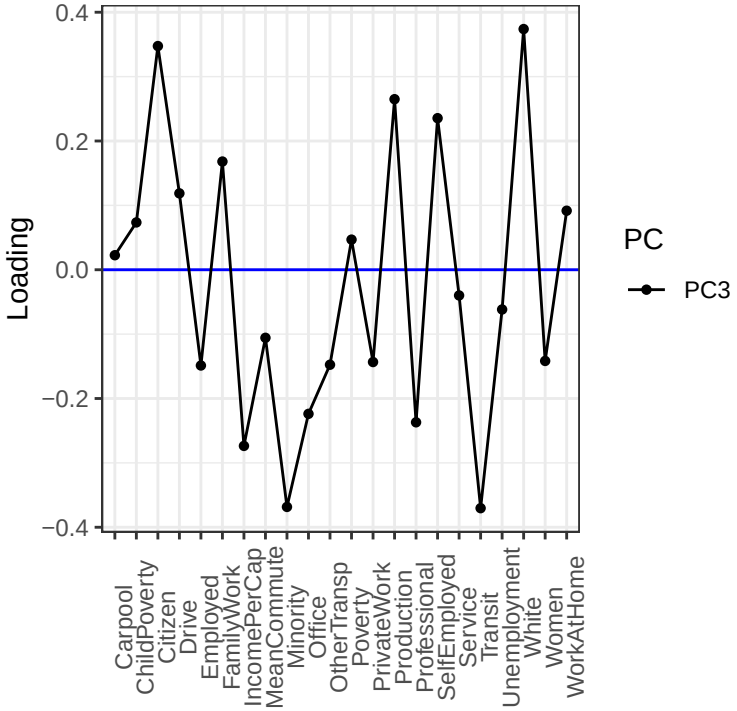


According to our loading plot for PC1, a positive PC1 value is strongly positively correlated with variables ChildPoverty, Minority, Poverty, and Unemployment. However, there is a strong negative correlation with

Employed, IncomePerCap, Professional, White, and WorkAtHome. We conclude that PC1 represents lower-socioeconomic class in the US.



According to our loading plot for PC2, a positive PC2 value is strongly positively correlated with variables Drive, MeanCommute, Office, PrivateWork, White, and Women. There is also a negative correlation to FamilyWork, SelfEmployed, WorkAtHome, and variables of economic status. Overall, we can conclude that these must be middle-class white-collar workers.



According to our loading plot for PC3, a positive PC3 value is strongly positively correlated with variables

Citizen, Production, SelfEmployed, and White. This indicates predominantly individuals that have upper class socioecomic status in the US. With a low correlation to IncomePerCap, MeanCommute, Minority, PrivateWork, Professional, Transit, and Women, we can deduce that PC3 must represent successful entrepreneurs, more specifically wealthy ones.

Based on our loading plots on PC1, PC2, and PC3, we can summarize that the voter demographics for the 2016 election are: lower socio-economic class individuals, middle-class white collar workers, and wealthy and successful entrepreneurs.

Discussion

Our results show key demographic variables that our system is built around. Through our principal component analysis, we found that different regions in the United States yielded different results. Each region had a different positive PC value. Additionally, the minorities were often correlated with ChildPovery, Poverty, and Unemployment; this is to be expected as the American society runs on a white-dominated system (white supremacy). It came as no surprise to see that blacks were disadvantaged in many of the regions. All in all at the end of the day, this analysis shows that we cannot discuss politics without discussing the topics of urbanization, income inequality, and race within our country.