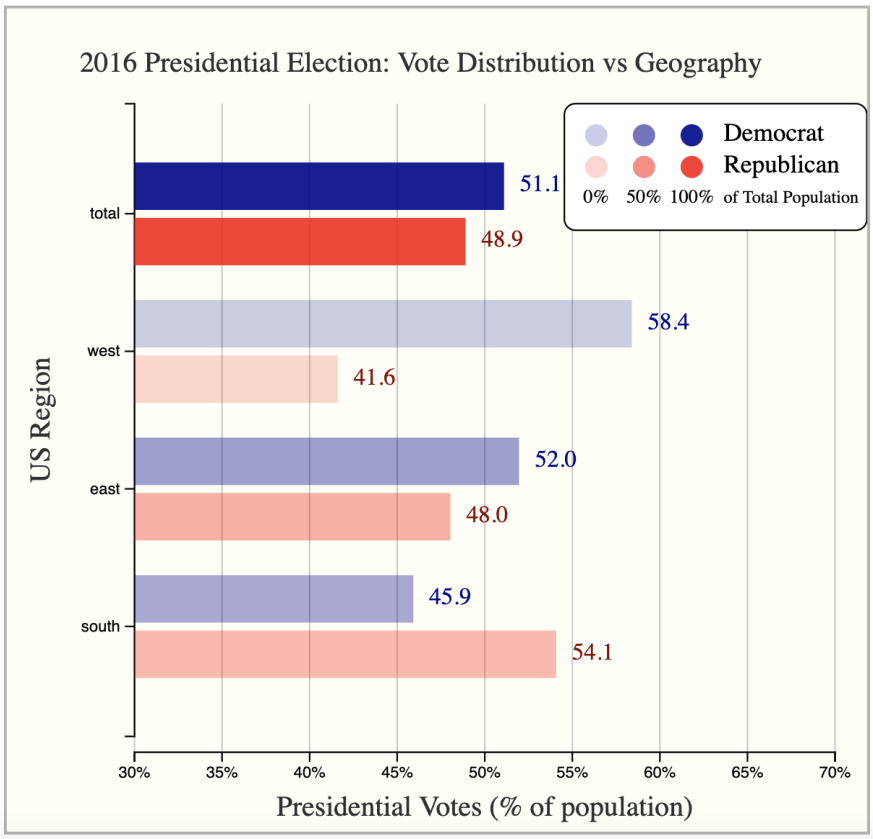
INFO 3300 Project 1

# Exploring County Voting Trends in the 2016 Presidential Election

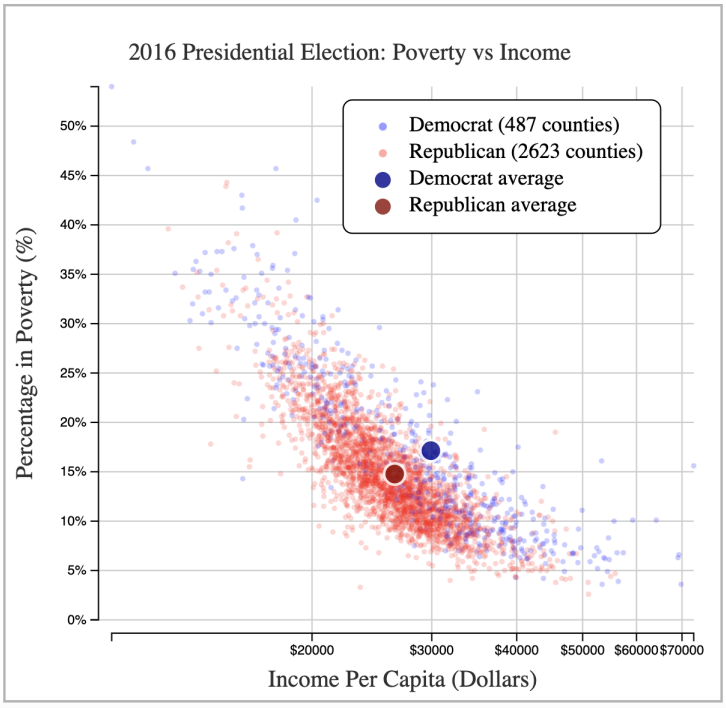
Ian Lee, Amber Zheng, Chris Tedeschi

**A.**

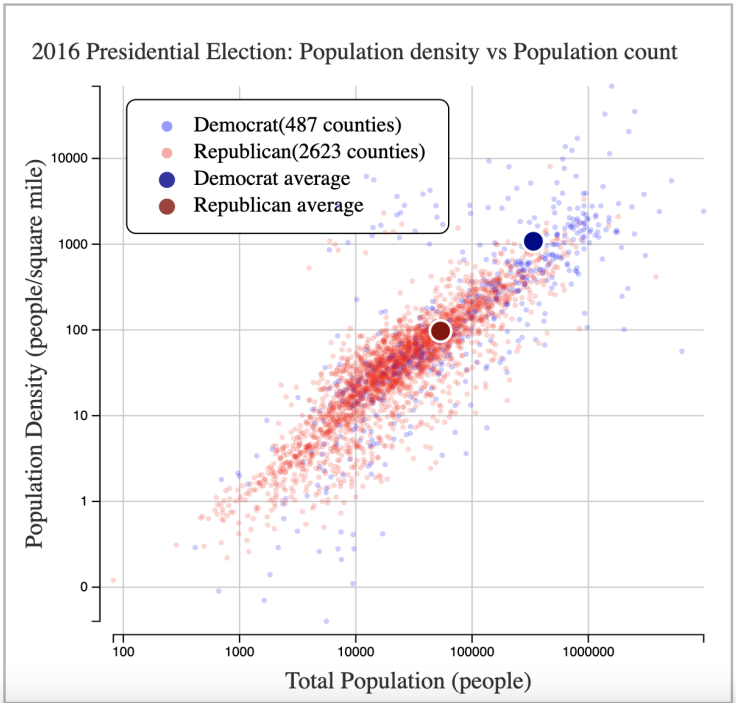
It can be seen that Western states have a higher affinity to the democrats while the Southern states have a higher affinity towards republican. Eastern states have a similar distribution with the total count



It is evident that the democrat supporting states have a higher average poverty rate and income as opposed to their republican counterparts




It is evident that the democrat supporting states have a higher population density and population count as opposed to their republican counterparts. This could mean that democratic counties are more urban and city-like while republican counties are more rural and suburban.



**B.**

We retrieved our first dataset that contains the county-level results for the 2016 presidential elections from the MIT Election Data and Science Lab on the Harvard Dataverse website. The second dataset with statistics on socioeconomic circumstances of the US counties was retrieved from the United States Atlas of Rural America. Both of these datasets have very reliable data because they use government data that was meticulously collected and organized.

The first dataset is named election_data, which contains information from the 2016 election results. It lists all county names from every state in the US, along with the county's vote count for each candidate (Hillary Clinton, Donald Trump, or other) and the total number of votes in the county. The candidates are presented alongside the party they represent (Hillary Clinton: Democrat, Donald Trump: Republican, or other: NA).

The second dataset, which is named final_census_data, includes information about each county's population and financial situations. For each county, the dataset (data retrieved by 2010) lists its median household income, per capita income, poverty percentage, population density, population density, percentages of age and ethnicity group distributions, foreign-born population percentage, homeownership, total population, and unemployment rate. For this project, we used the following variables for creating the charts: county names, per capita income, poverty rate, population density, and total population.

We integrated the two datasets together by joining them on a shared column, the county's FIPS code. A FIPS code is a unique number assigned to each US County. Then, for each county, we inserted three new variables in the latter dataset: Republican vote percentage, Democrat vote percentage, and total votes. The percentage of Republican/Democrat votes is calculated by dividing the Republican/Democrat vote count by the sum of Republican and Democrat vote counts, so third parties are excluded from our results. Since we are analyzing the election turnout for two main political parties in America, eliminating the third parties gives us a cleaner dataset and more evident patterns relating to socioeconomic factors and political stance for each county. We have also checked that no NA values were present in the newly created dataset and filtered out counties that did not have voting data (about 30 total counties).

**C.**

We included three visualizations to highlight various aspects of our data. The first visualization shows the political alignment of the major regions in the United States. Additionally, the percentage of votes for the United States as a whole is shown. This is important because it demonstrates how the combination of the regions comes together to represent a whole. This information is shown in a bar chart where the height of the bars represents the percentage of votes cast for each political party. The exact percentages are printed on top of the bars because it is not possible to determine the exact number from the height of the bars alone. Also, the bars

have varying saturation depending on the number of voters in each region. This helps demonstrate the differences in voting power between these regions. Axis labels and a key were also added to ensure the various aspects of the chart are clearly conveyed.

The marks on this chart are the bars that represent voting percentages. The channels are the aligned length and position, size, color, and saturation. We have included three circles with different saturation (opacity) to indicate the number of people who voted in each geographical area. The information can be identified in the legend, there are also numerical labels indicating the saturation that each represents 0%, 50%, and 100% of the total voter population. It is natural that the bar with the label "total" aligns with the full saturation while the sub-regions are light. This design choice is to allow users to easily identify the number of people in each region. Also, this allows users to easily spot the relative "weight" that each region contributes to the total votes.

The second visualization is a scatter plot of US counties which are mapped according to the county's poverty rate and income per capita. This shows the relationship between these variables and how they affect presidential voting. The points on the plot are partially transparent to help increase visibility at areas of high density where points overlap. A larger circle with a white outline (for visibility) represents the average statistics for Republican-leaning counties and Democratic-leaning counties. A linear scale was used to map the poverty rate, however, a logarithmic scale was used to map income per capita. This is because the relationship between the variables is non-linear, and a logarithmic scale helps spread the data across the entire plot. Axis labels and a key were also added to ensure the various aspects of the chart are clearly conveyed. The key also includes the number of counties that voted majority republican and the number of counties that voted majority democrat.

The third visualization is a second scatter plot of US counties which are mapped according to the county's total population and its population density. This shows the relationship between these variables and how they affect presidential voting. Once again, the points on the plot are partially transparent to help increase visibility at areas of high density where points overlap. They also have a low radius because there are a large number of counties to plot and overlap becomes an issue. A larger circle with a white outline (for visibility) represents the average statistics for Republican-leaning counties and Democratic-leaning counties. A logarithmic scale was used to map both variables. This is because these variables have a huge range of values, yet almost every county except a few fall in a much smaller range. Therefore, a linear scale would not show any relationship between the data and would result in an enormous cluster in the bottom left corner of the plot. Axis labels and a key were added to ensure the various aspects of the chart are clearly conveyed. The key also includes the number of counties that voted majority republican and the number of counties that voted majority democrat.

The marks on these scatterplots are the circles that represent each individual county and the larger circles that represent the averages for Democratic-leaning and Republican-leaning counties. The channels used are aligned horizontal position, aligned vertical position, size, saturation, and color.

**D.**

These visualizations show some surprising trends about how US county factors affect presidential voting results. Below, I outline what each visualization conveys and how this information pertains to trends in the United States.

The first plot demonstrates that there is a clear discrepancy between how different regions of the United States vote. Western voters heavily favor democrats, while Southern voters favor republicans, but not by as wide of a margin. The East is relatively moderate and has very similar voting results to the country as a whole. Conveying this information to the viewer is important so they understand that voting results vary greatly from region to region.

The second plot shows the relationship between a county's income per capita and its poverty rate. It is unsurprising that as a county's income per capita increases, the poverty level decreases. However, one surprising trend is that the counties with the highest income per capita almost all vote for the Democratic candidate. This seems counterintuitive because the wealthiest people have financial incentives to support Republican candidates as they support lower taxes. It is also odd that the average Democratic county has a higher per capita income yet also a higher poverty rate. This is likely because most Democratic counties are major cities as shown in the last visualization. This means that they have a higher rate of impoverished and homeless individuals while also housing many of the highest earners in the country. These outlying earners such as hedge fund managers make so much money that it offsets the average income per capita.

The third plot shows the relationship between a county's total population and its population density. Once again, there is an unsurprising trend that as a county's population increases, so does its population density. However, one thing that the graph shows that is not intuitive is that nearly every one of the counties with the highest population vote overall democratically. These counties are major US cities as mentioned above. This trend is very obvious when looking at the average points on the graph as the Democratic average has a much higher population and population density. This is important to understand because, upon first look at the key, one can see that a vast majority of counties vote Republican overall. Under 16% of the counties that had voting data voted for Democrats, however, these counties have such high populations that over half of the votes cast in the election were for Democrats.

Overall, our goal was to create visualizations that help viewers understand the complicated voting trends that occur in the United States. Each visualization focuses on a different aspect of these trends and shows a unique view of county voting data.

**Data Source:**

Census: https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/

Voting:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

**Team Contributions**

Chris Tedeschi

- Data exploration and creating the dataset (2 hours)
- Helped create the visualization although Ian did a lot of the work (2 hours)
- Helped write and edit this final writeup (2 hours).

Ian Lee

- Coded up the index.html file. Implemented data import, cleaning, exploring, and plotting. I did most of the d3 plots and legends as well as the CSS styling (11 hours)
- A little bit of write-up and clarification of design choices. (1 hour)

Amber Zheng

- Data selection and analysis of the datasets and possible charts (2 hours)
- Helped design and create visualizations for the three charts (3 hours)
- Wrote and refined the project report (3 hours)