

The Relationship between Voting Patterns and County Health Rankings

By: Stephen Lee, Priscilla Siow, Peter Liu

Github Repository: https://github.com/UC-Berkeley-I-School/Project2_Lee_Liu_Siow

Introduction

This exploratory study examines how county-level election outcomes are associated with social factors such as access to healthcare (e.g. insurance coverage and provider availability) and socioeconomic conditions like income and poverty levels.

The analysis begins with data from the MEDSL Dataverse, specifically the "County Presidential Election Returns 2000-2020" dataset maintained by the MIT Election Data + Science Lab. Additional data is sourced from the County Health Rankings & Roadmaps directory, complemented by the U.S. Census Bureau's SAHIE (Small Area Health Insurance Estimates) dataset spanning 2008–2022.

By integrating these datasets, the study aims to uncover new insights into past election outcomes by analyzing a range of metrics at the county level.

Research Question: What is the relationship between voting patterns and county health ranking measures (uninsured rate, healthcare access, and median household income)?

Primary Dataset

Election Data: County Presidential Data (2000-2020)	
Source	https://dataverse.harvard.edu/file.xhtml?fileId=8092583&version=13.0&toolType=PREVIEW
Description	MEDSL contains election results by candidates for each county in the US. MIT Election Data and Science Lab, 2018, "County Presidential Election Returns 2000-2020", https://doi.org/10.7910/DVN/VOQCHQ , Harvard Dataverse, V13, UNF:6:GILITHRWH0LbH2TItBsb2w== [fileUNF]
Size	72617 rows, 12 columns

Field of Interest	Description	Sample Value	Notes
year	Year of Presidential Election	2020	
state	State Name (Full)	ALABAMA	Uppercase
state_po	State Abbreviation	AL	Uppercase
county_name	County Name	AUTAUGA	Uppercase

county_fips	County FIPS code	1001	Up to 5 digits county fips code. The first or first and second digit is the state fips code.
office	Election Type (Presidential Election)	US PRESIDENT	
party	Political Party	DEMOCRAT	DEMOCRAT, REPUBLICAN, OTHER, LIBERTARIAN, GREEN
candidatevotes	Number of Candidate Votes by County	7503	
totalvotes	Total Votes by County	27770	
mode	Voting Mode	TOTAL	Voting Method

Supplemental Datasets

Uninsured / Insured Individuals (Explored by Peter Liu)	
Source	https://www.census.gov/data/datasets/time-series/demo/sahie/estimates-acss.html
Description	United States Census Bureau, 2008 - 2022 Small Area Health Insurance Estimates (SAHIE) using the American Community Survey (ACS). The SAHIE dataset contains various income levels by county.
Size	1070435 rows, 26 columns

Field of Interest	Description	Sample Value	Notes
year	Year of Estimate	2020	2000 - 2020
statefips	Unique FIPS code for each state	01	Two digit state fips. Zero padded.
countyfips	Unique FIPS code for each county within a state	001	Three digit county fips.
state_name	State Name	Alabama	Capitalized
county_name	County Name	Autauga County	Capitalized and ends with the word "County"

PCTUI	Percent uninsured in demographic group for <income category>	10.6	This analysis does not specify an income category. The "All" category was used.
NIPR	Number in demographic group for <income category>	4902	This analysis does not specify an income category. The "All" category was used.

Median Household Income (Explored by Priscilla Siow):	
Source	https://www.census.gov/data/datasets/2020/demo/saipe/2020-state-and-county.html
Description	The files below contain estimates of poverty and income for 2020. There is one data file for each state (or US) with data for ALL the 2020 statistics. Additionally, there is also one file that includes the data for the US and each state and county.
Size	3195 rows, 7 columns (after data cleaning)

Field of Interest	Description	Sample Value
County	County Name	California
Median Household Income	Estimate of median household income	67340
90% CI Lower Bound	90% confidence interval lower bound of estimate of median household income	67251
90% CI Upper Bound	90% confidence interval upper bound of estimate of median household income	67429

Access to Healthcare (Explored by Stephen Lee):	
Source	https://www.countyhealthrankings.org/health-data/county-health-rankings-measures

Description	Health Resources & Services Administration Dataset containing clinician data by Area Health Resource File year, profession, population, and location
Size	3231, 1180 columns

Field of Interest	Description	Sample Value
fips_st_ctny	County Name	1001
phys_nf_prim_care_pc_exc_rsdt_20	Number of Primary Care Physicians	150.0
md_nf_fed_activ_20	Number of MDs Available	722.0

Data Cleaning

- 50515 rows of Election Data containing NA / NaN were dropped
- Include only general election data 2008, 2012, 2016, and 2020
- Some counties do not report total votes. For those counties, totals are calculated by summing up all the voting methods (mode)

Assumptions - Election Data: County Presidential Data

- A number of columns are derived from the source Election dataset. Here are the names and descriptions.

Derived Columns	Descriptions
Votes Democrat / Votes Republican	The data is reorganized (pivoted) by party to create separate columns for Democratic and Republican votes, simplifying analysis.
Win Percentage	The winning percentage is calculated using the formula: $\text{winning percentage} = 100 * \frac{(\text{Votes Republican} - \text{Votes Democrat})}{\text{Votes Total}}$ A positive Win Percentage indicates a Republican win. A negative Win Percentage indicates a Democratic win. The sign indicates which party won the election for the county.
Outcome	Indicates the winning party (DEMOCRAT or REPUBLICAN) based on the vote counts for each party.
Political Affinity	Defined by the winning margin.

	Republican: > 5% Republican Leaning: 3% to 5% Center: -3% to 3% Democrat Leaning: -5% to -3% Democrat: < -5%
--	--

Assumptions - Uninsured / Insured Individuals

- The state and county fips do not match the Election data format. Create a compatible count fips format by combining the two codes. Keep in mind the state fips is padded with 0 if a single digit. The combined county_fips from Election data is not 0 padded. For example, the county_fip 1001 from the Election dataset corresponds to 01 state_fip and 001 county_fips.
- The uninsured rate is based on the population under the age of 65. Populations over 65 are excluded. Note federal medicare coverage for individuals starts at the age of 65, so anyone over the age of 65 has medical insurance.
- Swing States are defined as the states of Michigan, Pennsylvania, and Wisconsin.
- Although the uninsured dataset includes rates at different income levels, this analysis does not consider income level.
- Uninsured categories are quantile based.

Uninsured Category	Definition
Very Low	0 - 0.15 quantile
Low	0.15 - 0.45 quantile
Medium	0.45 - 0.55 quantile
High	0.55 - 0.85 quantile
Very High	0.85 - 1.0 quantile

Assumptions - Median Household Income

- All income is reported income so we are assuming that people are reporting their income accurately and it captures an adequate amount to determine the county level household income

Assumptions - Access to Healthcare

- Total MDs was not as good an indicator of access to healthcare. The number of MDs per county needed to be represented as a percentage of the population to be a more useful basis of comparison
- We are assuming that the rate of doctors living in one county and working in a contiguous county are negligible for this study

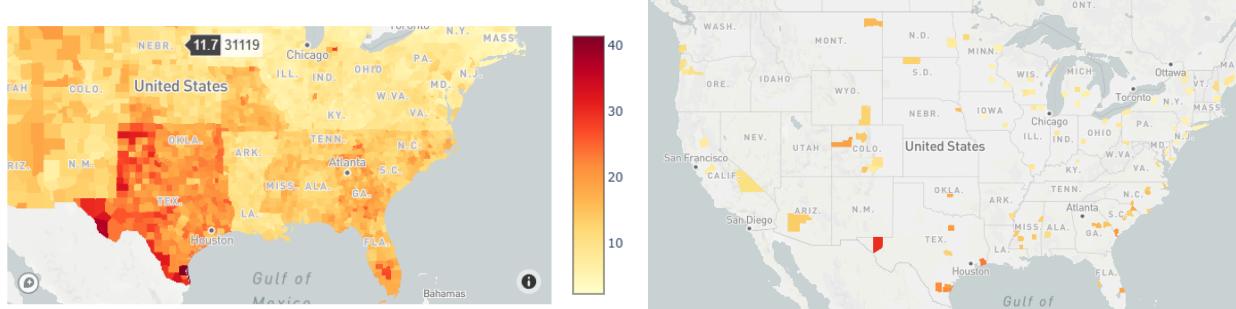
Project Motivations and Variable Selection:

- **Uninsured / Insured Individuals:** Voter perspectives on health insurance reflect a nuanced balance between party ideologies and personal experiences with the healthcare system. These diverging priorities underscore the pivotal role health insurance plays in shaping electoral outcomes.
- **Health Access:** In the election this last November, a large majority of voters said that healthcare is an “extremely or very important” issue in the election (according to Gallup). While everyone thinks it is an important issue, there are contentions on party lines when it comes to making healthcare accessible and affordable. For this part of the study, we are trying to understand the correlation between county voting patterns and access to healthcare professionals.
- **Median Household Income:** During the 2024 election, the economy was frequently brought up on news media outlets as a leading influence in how people voted in the election. With this in mind, we decided to include this measure in our project to study any observable patterns or relationship between income and voting.

Exploratory Questions and Results

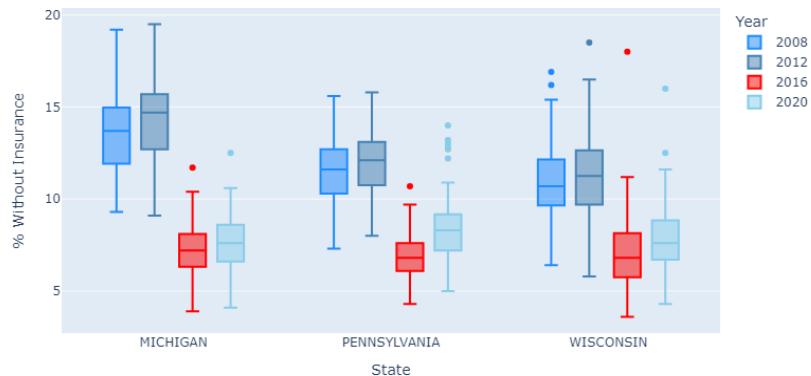
Uninsured/Insured Rate

Question: Where are the counties with a high percentage of individuals without health insurance? What are the percentages for the political center (where the winning margins are low)? Here's an overview of the percentage of uninsured individuals across the United States in 2020. Regions along the Texas border, parts of Florida, and, to some extent, areas in the Southeast show a notably higher percentage of uninsured individuals.



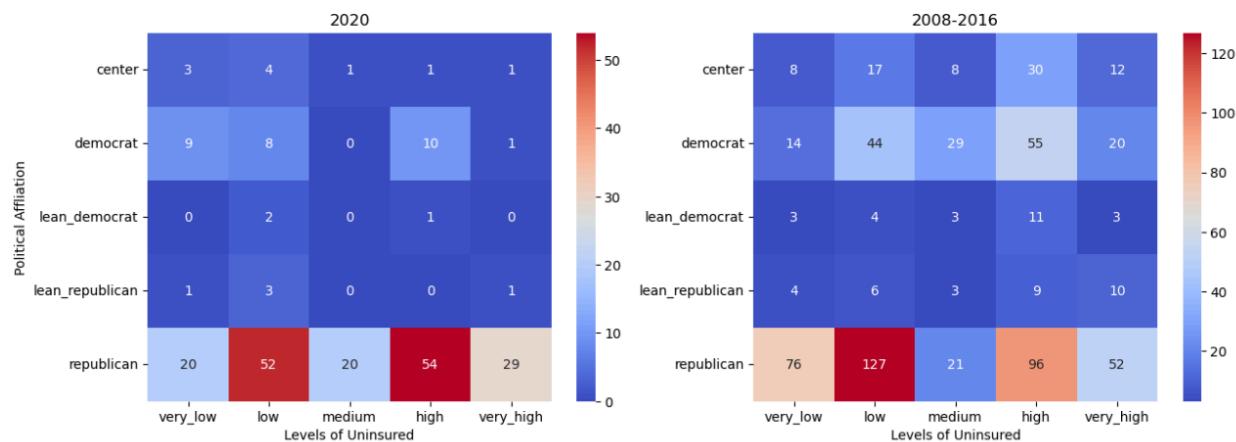
Question: Zooming down to the state level, can we notice any trends in the uninsured rate (percentage of individuals without health insurance) for the past four election cycles? Focusing on the swing states of Michigan, Pennsylvania, and Wisconsin, the percentage of uninsured individuals has decreased in recent years. Notably, this significant drop coincides with the passage of the Affordable Care Act (ACA) in 2013.

Swing States Not Insured 2018 - 2020, ACA (2013)



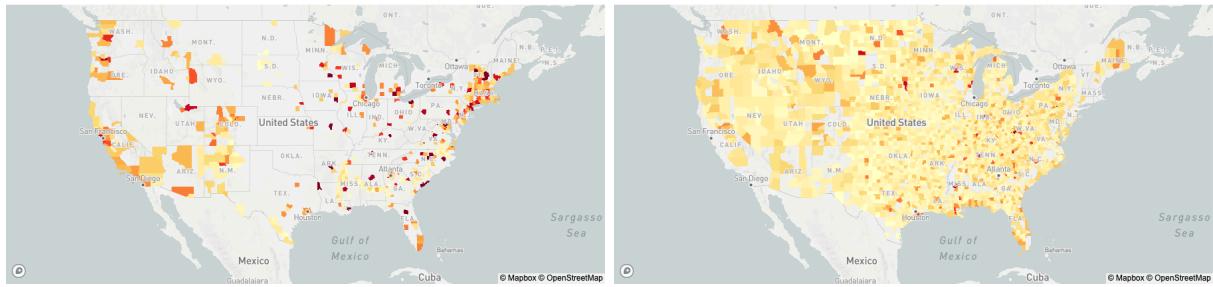
Question: Between 2020 and pre-2020, are there significant shifts in uninsured rate from different political affiliations?

In 2020, most Republican counties saw significantly higher levels of uninsured individuals compared to the 2008-2016 period. Democratic counties experienced a shift as well, though it was less dramatic. There was no noticeable change in the uninsured rates for Lean/Center counties.



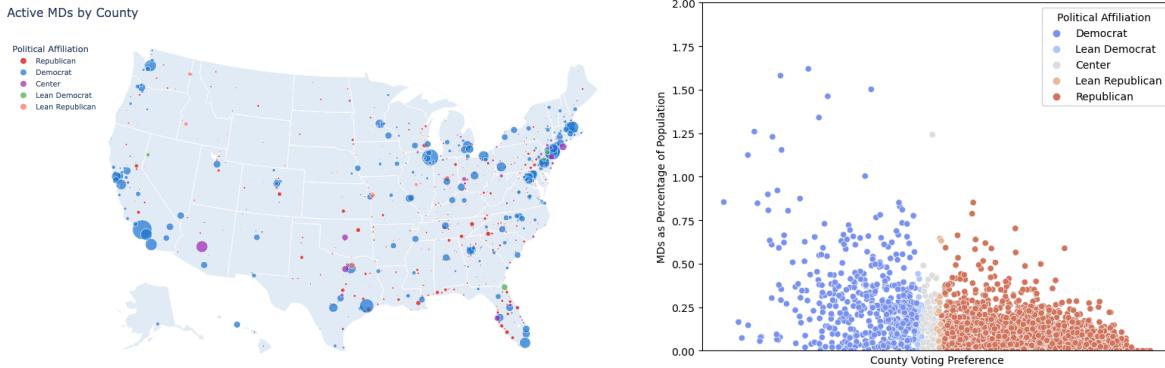
Access to Health Care

Question: What does health access look like for counties with different political leanings? In order to explore this, we can look at counties that voted democratic and counties that voted Republican. It seems that Democratic Counties tend to have a higher percentage of MDs in their population, but it is also worth noting that these counties seem to correspond with large urban areas.



Democratic Counties (on the right) and Republican Counties (on the left)

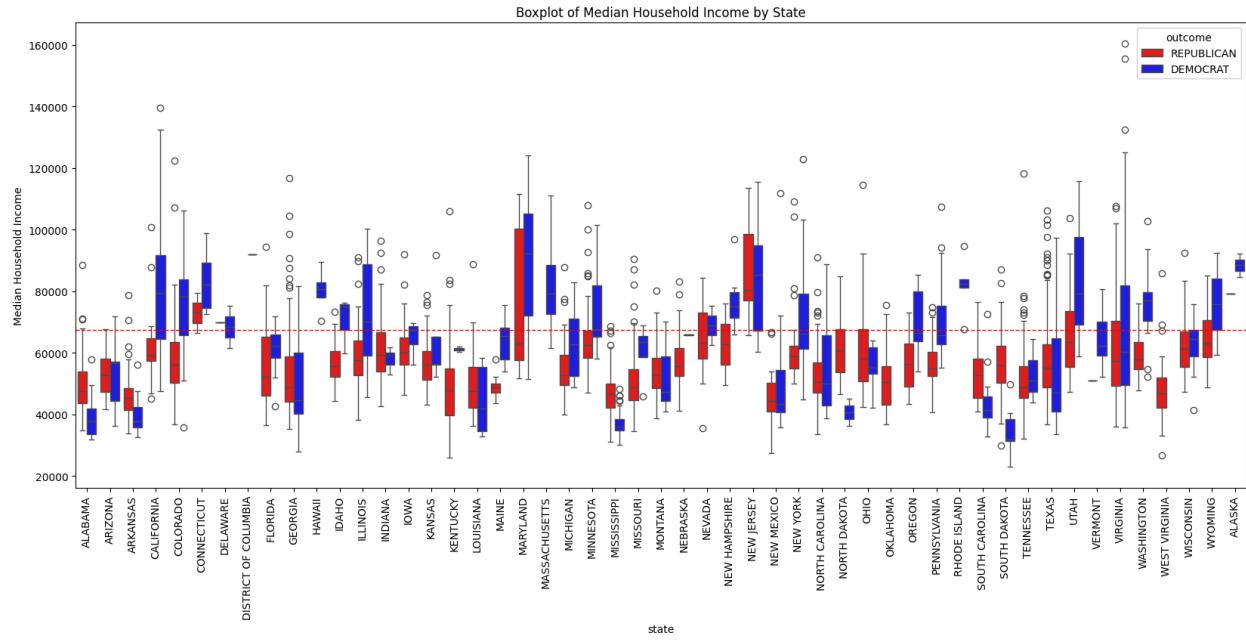
While looking at access to MDs as a percentage of population is important, it is also interesting to document how many MDs are in each county. As expected, the concentrations are around large urban areas (plot on the left).



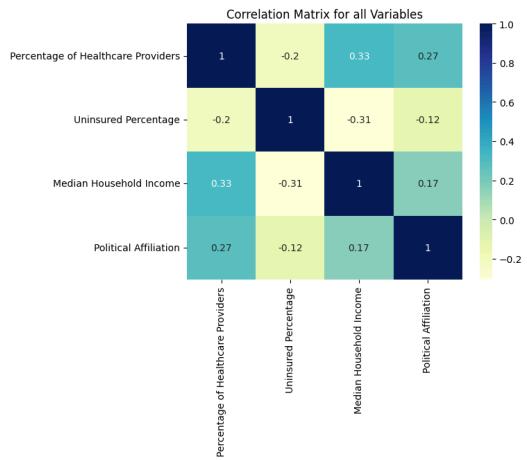
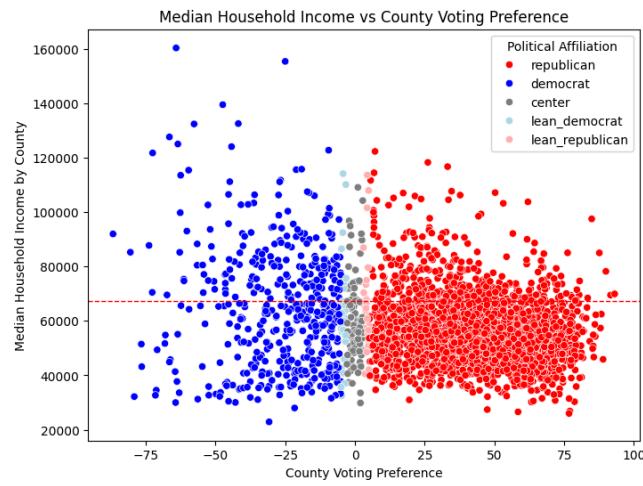
Here's what it looks like mapped on a spectrum (plot on the right). The further left a point is on the plot, the heavier the Democratic lean. Conversely, the further right a point is on the plot, the heavier the Republican lean. It seems that the counties further left on the x-axis tend to have a higher level of access, while it seems that counties tend to decrease in MD percentage as they progress further right on the x-axis. It is also worth noting that there seem to be a much higher concentration of republican counties than democratic ones, which fits expectations given that Democratic counties tend to be more densely populated.

Median Household Income

Question: Are there any patterns between how individuals vote and how much income they earn? To observe this relationship, the boxplot below shows the median household income associated with election results and state. The red horizontal line represents the nation median household income and we can see that most states (like Alabama, Arizona, Arkansas, Mississippi, etc.) fall under the nation median household income of \$67,340.



From an initial look at the boxplots, more republican voting counties in the states are below the nation median household income. Down below, looking at the scatter plot broken down by political party and median income, we can see that the republican counties are extremely dense but don't have as many high earning outliers compared to the democratic side. With the republican counties being more densely populated on the chart, it also means that more republican counties are below the horizontal line that represents the nation median income level—which is also confirmed in the boxplot above.



Correlation between all measures of interest: To calculate correlation, pandas has the corr() function which by default uses the Pearson correlation coefficient to determine correlation between variables. A positive coefficient means a positive

association or correlation to one another. Meanwhile, a negative value indicates a negative association or correlation.

Conclusion

To answer our research question about the relationship between voting patterns and county health rankings, we decided to calculate the correlation between all measures of interest: Percentage of Healthcare Providers, Uninsured Percentage, Median Household Income, and Political Affiliation. With the plot above, we can see that most of the variables are not correlated to one another since most of the values are negative. However, there seems to be a slight positive correlation between Median Household Income and Percentage of Healthcare Providers as well as Median Household Income and Political Affiliation. Percentage of Healthcare Providers is also slightly correlated with Political Affiliation.

As a note, we cannot determine causal relationships through this work, and only intend on calculating correlation to observe any kind of relationship between these variables without identifying any direct causal relationships.

There are countless ways to explore the intersection of election data with health, social, and economic factors. Our analysis has barely scratched the surface, revealing only a small amount of insights hidden within. Further deep dive exploration can uncover even deeper and connected information that can be used to address social challenges.

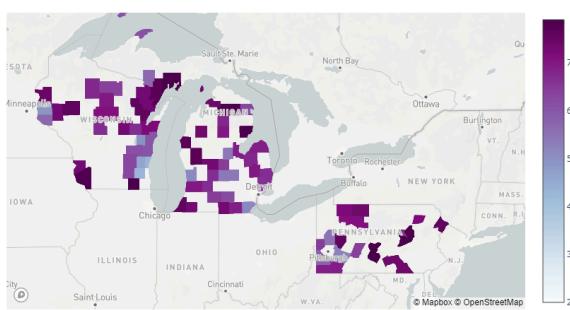
References

1. Amy O'Kruk et el "7 charts and maps show where Harris underperformed and lost the election" cnn.com November 2024,
<https://www.cnn.com/2024/11/27/politics/election-voters-harris-what-matters-dg/index.html>
2. Jeffrey M. Jones, "Healthcare Remains Important U.S. Voting Issue" gallup.com, October 2024,
<https://news.gallup.com/poll/651755/healthcare-remains-important-voting-issue.aspx>
3. Mapbox., "Mapbox", 2023, <https://www.mapbox.com>
4. Plotly Technologies Inc., "Plotly Express", 2024, <https://plotly.com/python/plotly-express/>
5. Plotly Technologies Inc., "Plotly Dash", 2024, <https://dash.plotly.com>

Appendix (Additional Findings)

Question: What counties go against the trend of Republican counties having low levels of uninsured, and on the other side Democrats with high levels of uninsured? The outliers show an interesting pattern when broken down into political affiliation. "Many" Republican counties show a lower level of uninsured even though as a whole, the uninsured rate is higher. Conversely, very few Democrat counties deviate from the generally lower level of uninsured.

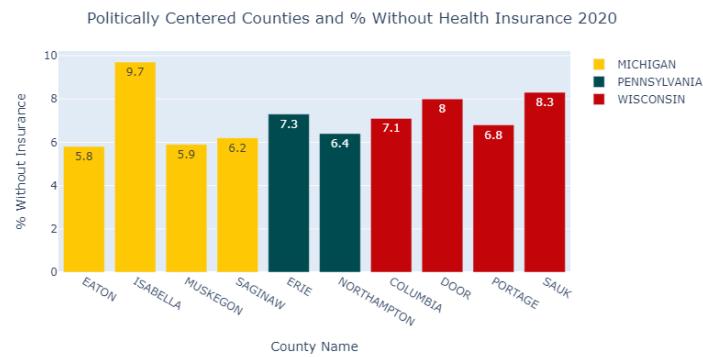
2020 Republican Counties with Low / Medium % Without Insurance



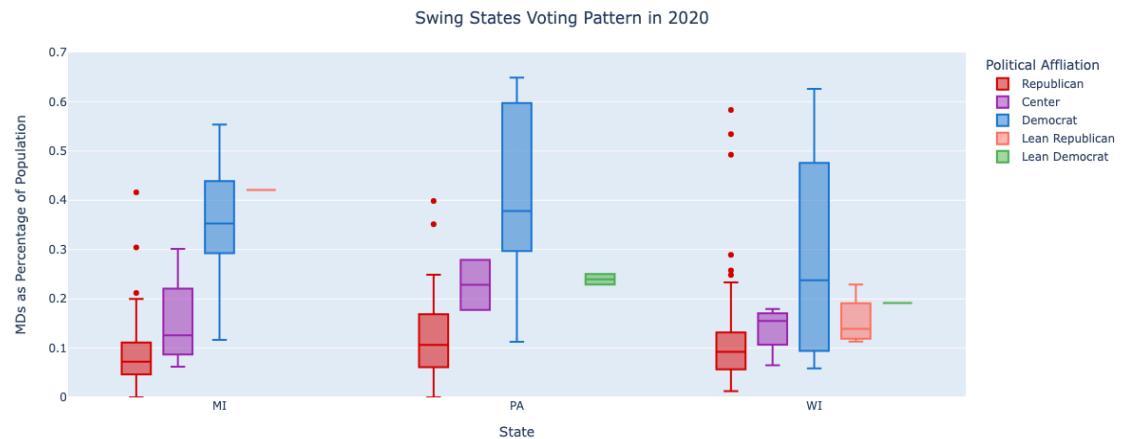
2020 Democrat Counties with High / Very High % Without Insurance



Question: What are the uninsured rates for counties that voted politically centered? Here is a list of center counties from the swing states along with their levels of uninsured residents.



Question: What does the distribution of health care access look like for swing states? In this box plot, it is interesting to note that the Democratic counties have a high range of access to MDs.



When looking at census income data, there are a few states with low income reporting, like Alaska, with only higher earning individuals responding to the census which can misrepresent data for the state. We can see in the map below with counties in darker red with higher income

and lighter yellow with less income. The state of Alaska is mostly unfilled due to low census data.

