Emily Macris, Hiba Khan, Jesus Zarate, Kiduse Gezehagne, Rix Prakash, Shaveen Saadee, Vaibhav Jha
DS 3001 Foundations of Machine Learning

U.S. Presidential Election 2024: A Predictive Model

I. Introduction

Our group is interested in building a machine learning model that will effectively predict the outcome of the 2024 United States presidential election. Understanding that the election outcome will likely boil down to candidates' performances in seven 'battleground' states, our group decided to focus our research on these jurisdictions. These states are: Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin. Media sources across the political spectrum agree that the 2024 U.S. election outcome will likely come down to these seven states. This report outlines the extensive process our group has taken to both conceptualize and operationalize our project plan.

II. Data Collection

Our group began the data collection process by individually assessing data availability online. We understood that many political, social, and economic measures were likely available through non-profits, the US Census, and other organizations. Confident in the availability of data, we attempted to survey the units of analysis to consider, the years of data available, and the uniformity of data sources across all seven states. These factors were influential in how we would craft our research design.

Following individual research, our group met for a brainstorming session. Discussing our individual findings and thoughts on how to approach the task at hand, we began to define our research project in line with the limitations that we identified. This led us to make concrete observations in three facets of the data collection process:

1) Unit of Analysis: Beginning with a concept of constructing an election-predicting model but with no concrete basis, our group asked one preliminary question: what would "one observation" of our data entail? We explored multiple options from the individual to the state. Following discussion, we agreed on a county-level unit of analysis. We found that most existing data was collected at the county level. In addition, we felt that a county-level analysis would allow us to build the most representative predictive model. We recognized that counties within a state embody different voting practices based on factors that matter uniquely to them. With the Electoral College being the basis of the U.S. election, it was imperative to get a granular understanding of states' voting trends in order to effectively predict the 2024 presidential outcome.
2) Time: The next consideration our group made was how time would affect our analysis. What years could we collect data from? What election years would we train our model on to predict 2024? Would we pool the data to control for time or try to account for time in building a time series model? Our group found most predictor data to be available from the early 2000's to present. Moreover, our group found outcome data from as early as the mid-20th century. After discussion, we decided to limit the scope of our data from 2012 to 2024. This would allow us to use the outcomes of three past presidential elections (2012, 2016, and 2020) to predict the outcome of the 2024 election. As outlined below in

the Challenges section of this report, though, our group is still working on the specifics for many of these questions.

3) Research Methods: Our group understood that our model would largely be built around secondary data collection methods. We aimed to use a mix of macro-level statistics (i.e., demographics, economic indicators) and poll surveys as potential predictors of the model. After further research, though, we realized that the use of poll surveys as indicators of public sentiment was impractical. Polls largely presided at the state-level, with limited data for our intended unit of analysis – the county. Moreover, polls were lacking in uniformity over states and time; there was not consistent polling done over all states nor years of interest. This led us to shift our focus to macro-level statistics as the basis of feature selection.

Feature Selection

Following this acknowledgment of limitations, our group began feature selection. Together, we brainstormed different factors we thought were influential in predicting the outcome of a presidential election. This process was designed to be more idealistic in nature; we asked, assuming the data were available, would we want to collect it? Figure 1 tabulates all the features we thought were ideal in collecting.

Variables to consider		
Race	Education level	Poverty
Gender	Foreign-born civilians	WIC/SNAP benefit usage
Economic status	Businesses present	Law enforcement spending
Age	Crime	Job market

Figure 1. A list of variables brainstormed from our group discussion.

Divvying up the variables among the seven group members, we set out to collect data for as many of these variables as possible. Each group member sought to collect data for all states in an effort to keep the data sources uniform across states.

Figure 2 shows the variables that our group was successful in collecting, with the variable type, a description, and an example of what the variable entails.

Variable	**Variable type**	**Description**	**Example**
Party	Categorical	County outcome for election	Democratic Party
Age x Sex	Categorical	Six age brackets by	Males 35-44

		sex – proportion of total county population	Percentage
Race	Categorical	Proportion of six races by total county population	Black or African American Percentage
Education	Categorical	Proportion of highest-level of schooling by total county population	Regular high school diploma Percentage
Economic status	Categorical	Proportion of civilians falling within six income brackets	$100,000-$149,999 Percentage
Labor force	Numeric	Proportion of those in labor force by total county population	In labor force Percentage
Poverty	Numeric	Proportion of those in poverty by total population	Percent in poverty

Figure 2. List of attained predictors following data collection.

Our group collected predictor data consisting of four categorical variables and two numeric variables. The data span the three election years of interest. The outcome variable is categorical, reflecting the party outcome for each county within each election year. The data sources were the U.S. Census Bureau, IPUMS, and Professor Terry Johnson's past voting data available within the course GitHub. After a follow-up meeting discussing our data collection findings and experience, we concluded on the use of these seven variables.

Hypotheses

Finalizing the preliminary data collection process, our group reflected on how these data would be influential in predicting our intended outcome: the 2024 presidential election. Prior to beginning the data wrangling process, our group proposed twelve hypotheses indicating how we anticipate the predictors to influence the outcome.

Hypothesis #	**Description**
1	The younger a population, the more likely a population is to vote Democrat.

2	The older a population, the more likely a population is to vote Republican.
3	The larger the proportion of females in a population, the more likely the population is to vote Democrat.
4	The larger the proportion of males in a population, the more likely the population is to vote Republican.
5	The larger the proportion of Whites in a population, the more likely the population is to vote Republican.
6	The larger the proportion of non-Whites in a population, the more likely the population is to vote Democrat.
7	The larger the proportion of college graduates in a population, the more likely the population is to vote Democrat.
8	The larger the composition of lower income civilians in a population, the more likely the population is to vote Democrat.
9	The larger the composition of higher income civilians in a population, the more likely the population is to vote Republican.
10	The larger the labor force proportion in a population, the more likely the population is to vote Republican.
11	The smaller the labor force proportion in a population, the more likely the population is to vote Democrat.
12	The larger the proportion of a population is in poverty, the more likely the population is to vote Democrat.
13	The smaller the proportion of a population is in poverty, the more likely the population is to vote Democrat.

The six predictor variables of interest are fundamental for any population. Unlike industry-specific or age-specific metrics (e.g., presence of manufacturing companies, social media usage), we feel that these six variables are instrumental in predicting the outcome of a presidential election. We are confident that each of these variables will provide constructive information for the model.

Data Wrangling

All demographic variables excluding party and poverty were collected within one file, separated by year. There were a total of three demographic data files (by year), one party data file, and seven poverty data files (by state). The goal was to construct three data frames for each state by year.

To clean the demographic data, the three files were filtered by state in Python. The result was three data frames for each state, one for each year of interest (e.g., Nevada 2012, Nevada 2016, and Nevada 2020). To clean the party data, the single file was filtered by state. The data frame was then further filtered by year, resulting in three data frames for each state, one for each year of interest. Lastly, to clean the poverty data, the seven state files were filtered by year. This also resulted in three data frames for each state, one for each year of interest.

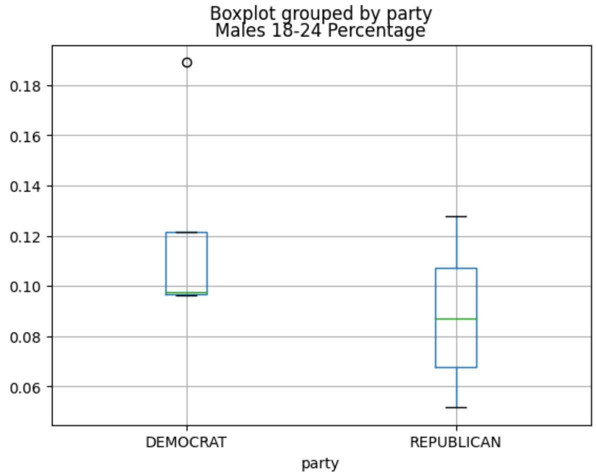
Further data wrangling was necessary in the demographic data frames. The raw data collected consisted of magnitudes of the variables of interest, including race, education, and economic status. To control for population differences, we calculated proportions of these measures by dividing the magnitudes by total population of the respective county. This resulted in new tabulations (columns) with data that we would later use for exploratory data analysis.

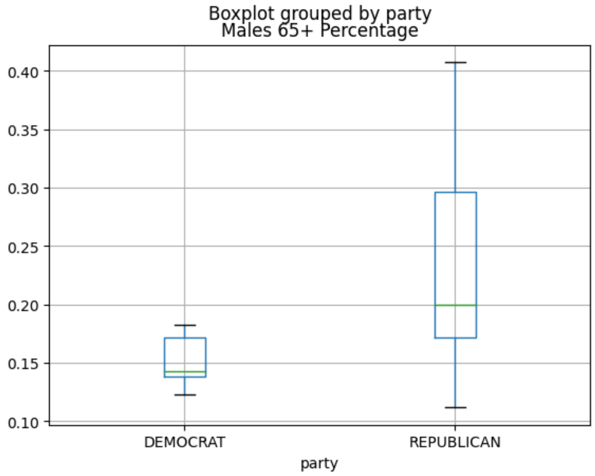
After merging the individual variables' data frames together, limited data wrangling was necessary. We had to drop state/national averages that came with many of the data sets to maintain county-level observations. There was very limited missing data. There was no need for removing a significant number of observations nor imputing values.

III. Exploratory Data Analysis

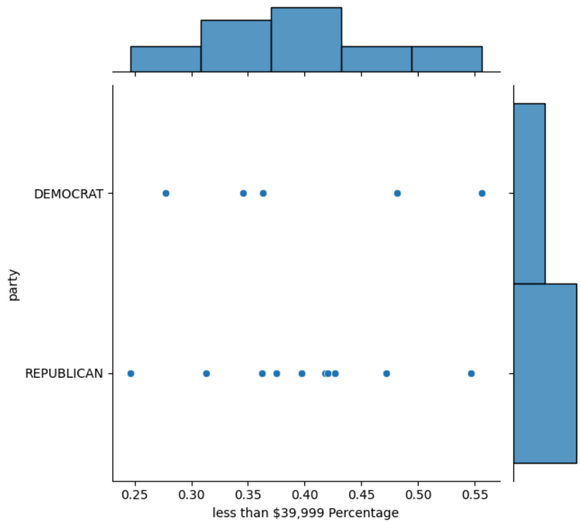
Arizona

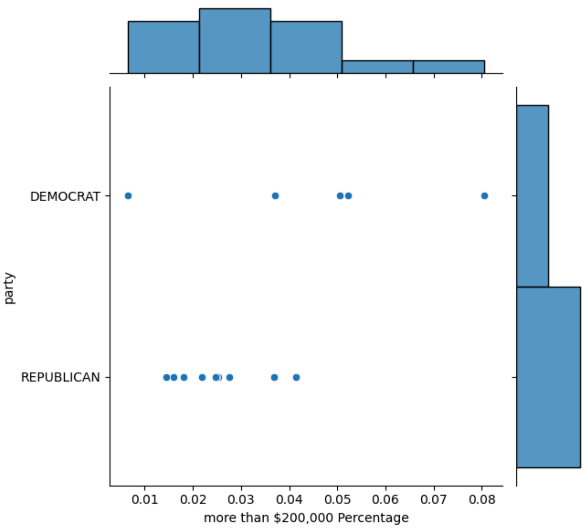
Arizona, known for its riveting sunsets and warm hospitality, is located in the Southwest. With a population of 7.3 million people, the state comprises fifteen counties, with a mix of historically Democratic and Republican-leaning areas. We emphasize that the EDA may not be particularly effective in representing the Democratic party as four counties are used to form a plot of the various quartiles.



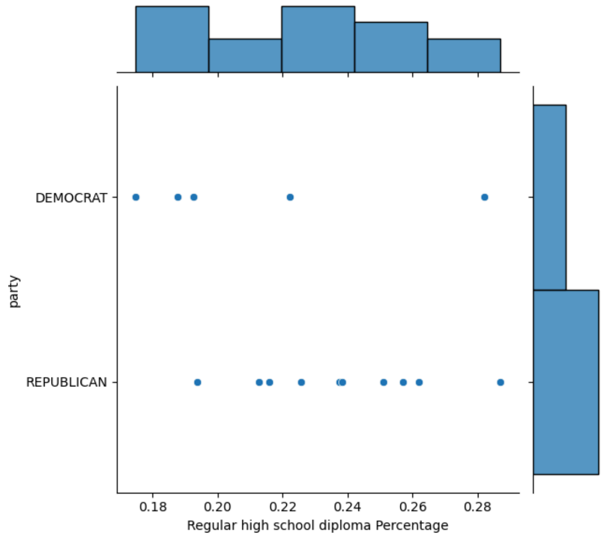


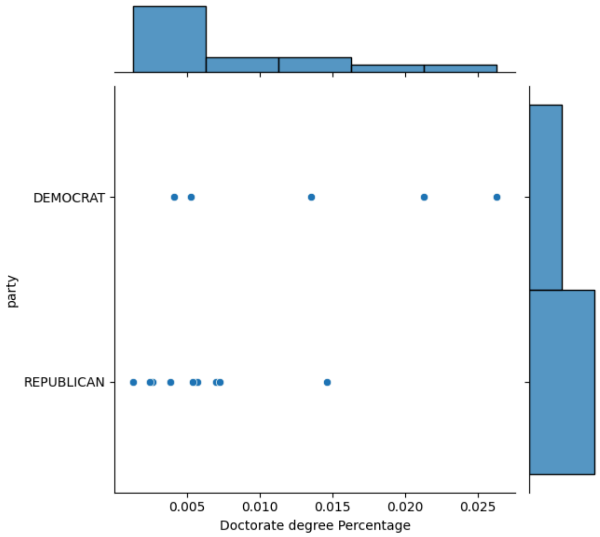
Above are two boxplots of age by party outcome in 2020. The plot on the left shows how Democratic voting counties had a greater proportion of males 18-24 than Republican counties. Conversely, the plot on the right shows how Republican counties in 2020 had higher proportions of males aged 65+ civilians than Democratic counties. This relationship coincides with Hypotheses 2 in the Hypotheses section.





Above are two joint plots depicting the relationship between income bracket and the party outcome in 2020. The plot on the right shows a relationship between counties that have a larger proportion of wealthier civilians and the party, counties with a greater proportion of wealthier civilians tend to vote democrat. However, the left plot does not demonstrate a significant relationship between counties with a greater proportion of lower income civilians and the party outcome. Counties with a higher proportion of civilians earning over $200,000 annually vote both Republican and Democrat. Thus, our preliminary analysis does not provide conclusive results for Hypothesis 8, as counties in Arizona with a greater proportion of wealthier civilians instead tended to vote Democrat. This is in accordance with Hypothesis 9.

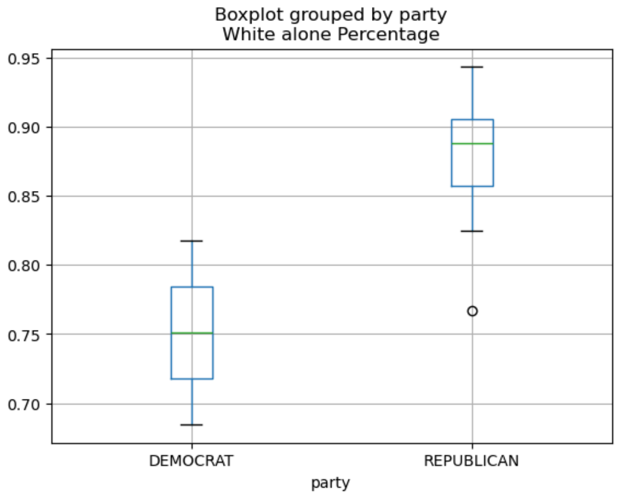


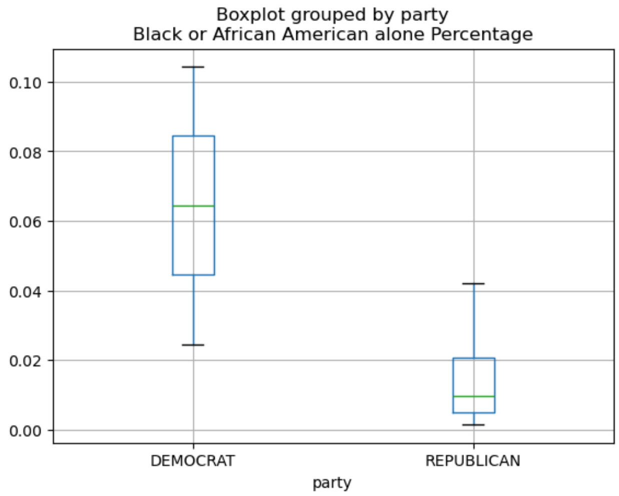


Above are two joint plots depicting the relationship between the highest level of education attained and the party outcome in 2020. With the plot on the left, we can infer that counties with a greater proportion of civilians that have a high school diploma as their highest educational attainment are more likely to vote Republican. Conversely, the joint plot on the right shows a loose but still noteworthy relationship between a higher percentage of civilians holding doctorate degrees and voting Democrat, supporting Hypothesis 7.

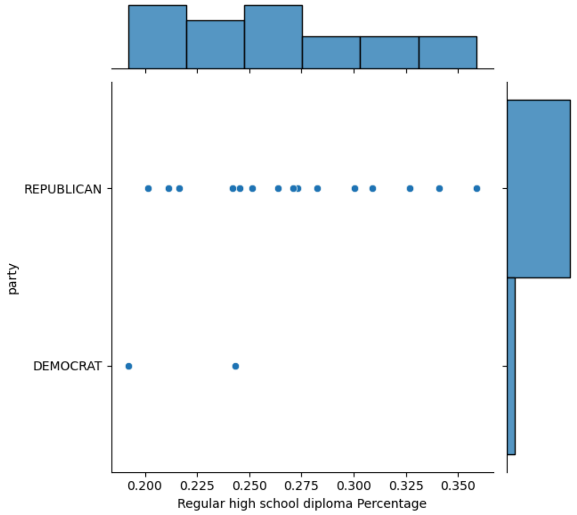
Nevada

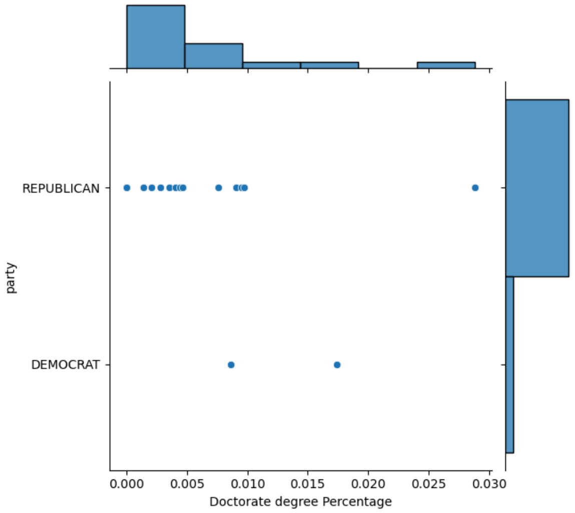
Known for its nightlife and silver, Nevada is the seventh largest state (by area) located in Western America, bordering California and Arizona. With a population of 3.2 million people, the state comprises seventeen counties with two historically Democratic counties and fifteen historically Republican states. For data analysis purposes, it is important to note that the boxplots are not particularly informative for the Democratic party, as two counties are used to form a plot of the various quartiles.



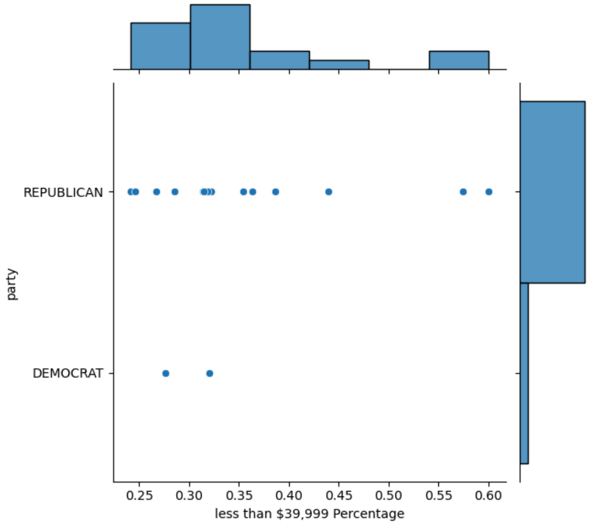


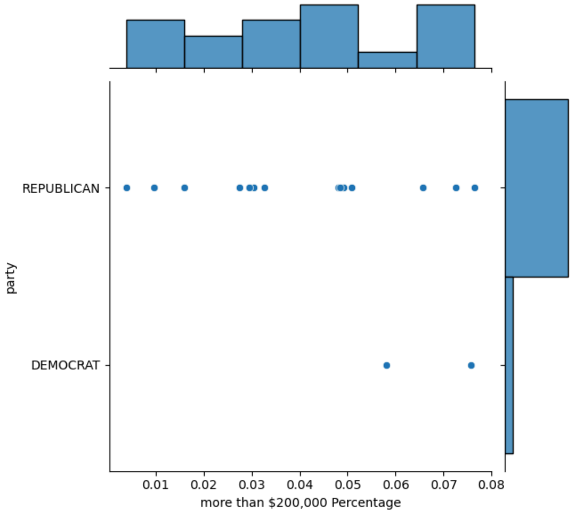
Above are two boxplots of race by party outcome in 2016. The plot on the left shows how counties with higher proportions of White civilians tended to vote Republican. Conversely, the plot on the right shows how counties with higher proportions of Black civilians tended to vote Democrat. This relationship coincides with Hypotheses 5 and 6 in the Hypotheses section.



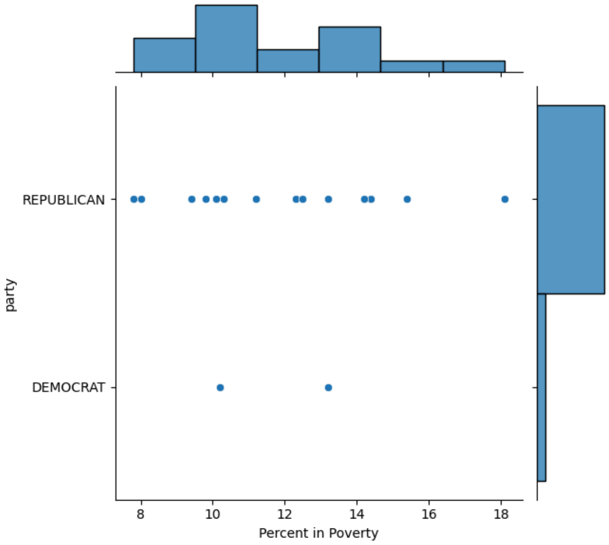


Above are two joint plots depicting the relationship between highest level of education attainment and the party outcome in 2020. As in the case of race, the Democratic party has two counties. With the plot on the left, it can be loosely inferred that the greater the proportion of civilians that have a high school diploma as their highest educational attainment, the more likely the county is to vote Republican. Conversely, the joint plot on the right shows a loose relationship between a higher percentage of civilians having a doctorate degree with voting Democrat. This would align with Hypothesis 7.





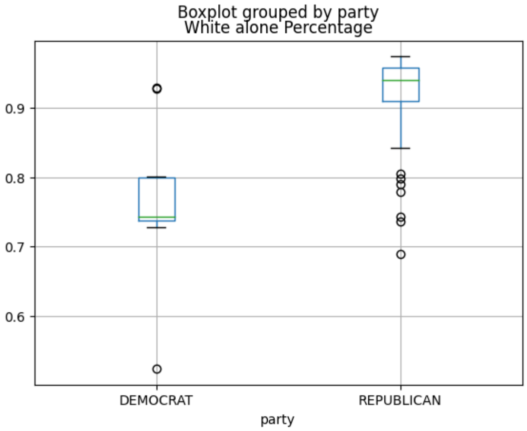
Above are two joint plots depicting the relationship between income bracket and the party outcome in 2020. The plot on the left shows a relationship between counties that tend to comprise more lower income civilians and Republican party votership. This is not in line with Hypothesis 8. Conversely, the plot on the right lacks demonstration of a significant relationship. Counties of higher proportions of civilians that earn over $200,000 vote both Republican and Democrat. The relationship is not clear.

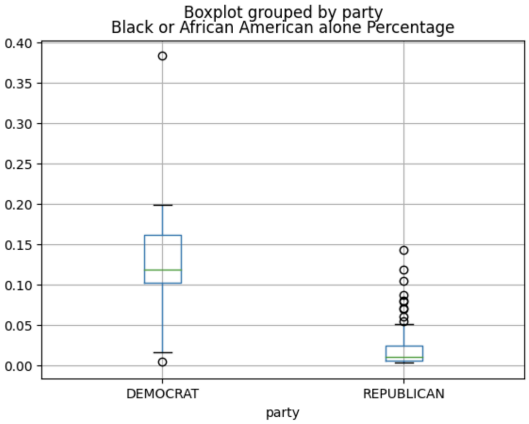


Lastly, the joint plot on the left shows the relationship between the proportion of civilians in poverty in a county and the party outcome in 2020. The relationship is unclear.
As the Nevadan exploratory data analysis continues, it is imperative to account for a proper sample size. The lack of sufficient Democratic representation may result in unrepresentative results for the 2024 presidential election prediction.

Michigan

Michigan, located in the Midwest, has a population of nearly 10 million people and comprises 83 counties. Historically, Michigan has been a key swing state in presidential elections, owing largely to its industrial past and economic transition. The state has a diverse demographic composition, including a significant proportion of White, African American, and Hispanic populations, which often influences voter behavior.



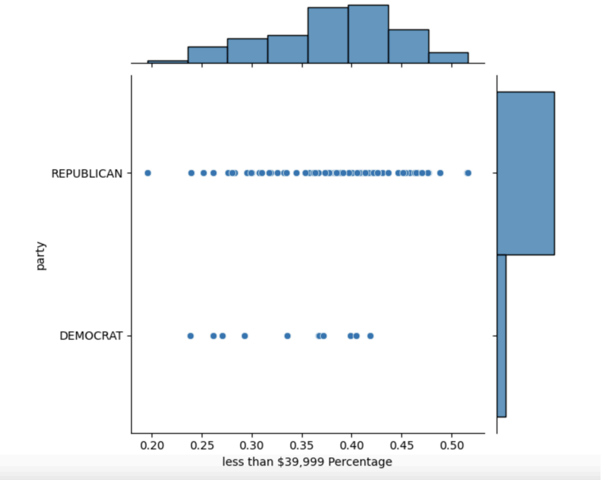


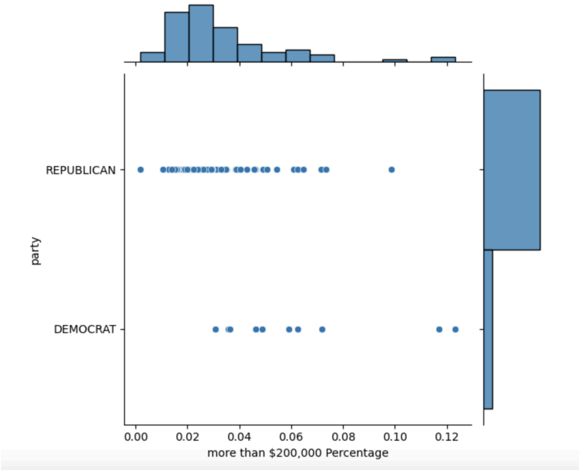
The above boxplot (*left*) shows the relationship between the percentage of the population that identifies as "White alone" and the party outcome in 2016. Counties with higher proportions of

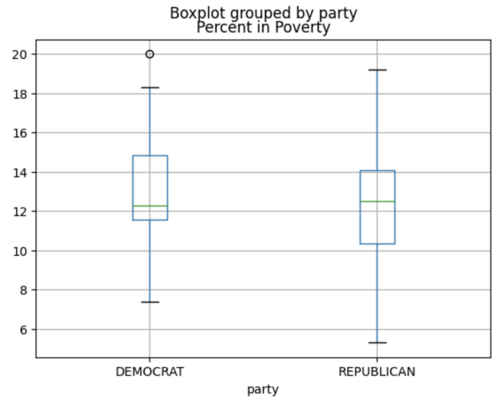
White residents tend to vote Republican. This observation aligns with the trend often seen in Midwest states, where rural, predominantly White counties tend to vote Republican, while more diverse, urban counties tend to vote Democrat.

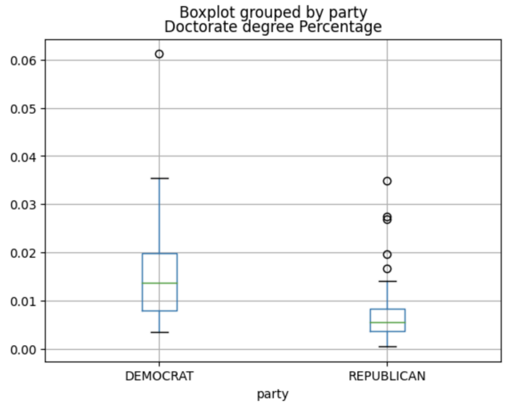
The above boxplot (*right*) for the "Black or African American alone Percentage" in Michigan (2016) illustrates a racial divide in voting behavior. Counties with a significantly higher proportion of Black residents tend to vote Democrat. In contrast, counties with much smaller Black populations tend to swing right.This pattern aligns with broader national trends where African American voters tend to support the Democratic Party, suggesting racial composition, specifically the percentage of Black residents, is a strong predictor of voting outcomes in Michigan.

The below plot (*left)* shows that in Michigan, there is no significant difference in voting patterns between counties with a higher percentage of individuals earning less than $39,999. The below plot (*right)* illustrates no difference in party outcomes with counties with a higher percentage of individuals earning over $200,000. Further statistical analysis and research is required to investigate this relationship.





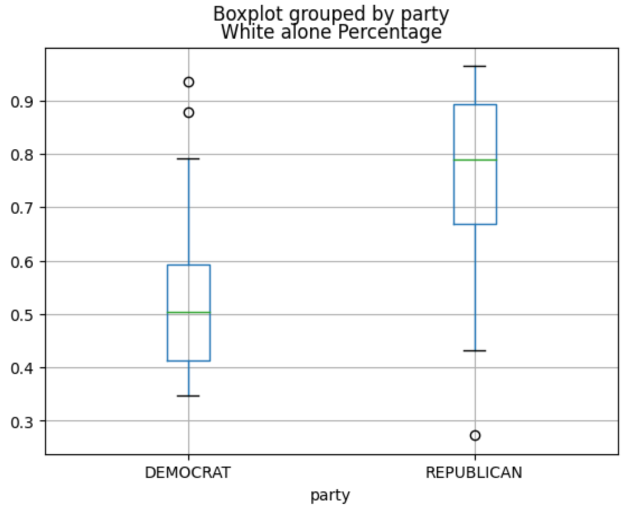


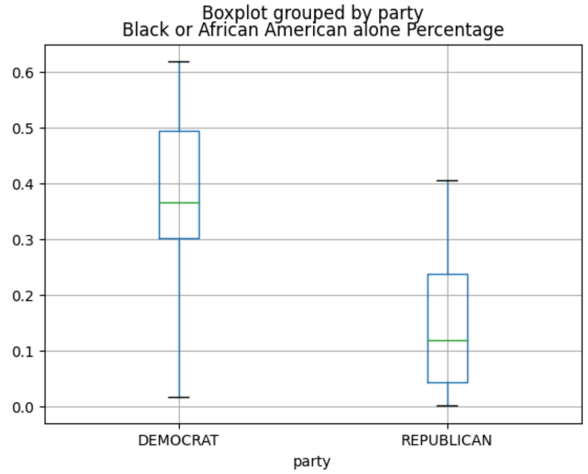


The above plot (*left)* illustrates that in Michigan, the median percentage of individuals in poverty is slightly higher for Republicans, although the difference compared to Democrats appears to be negligible. The above plot (*right*) indicates that in Michigan, counties with a higher percentage of individuals with doctoral degrees tend to vote for Democratic over Republican candidates, although the results are somewhat varied.

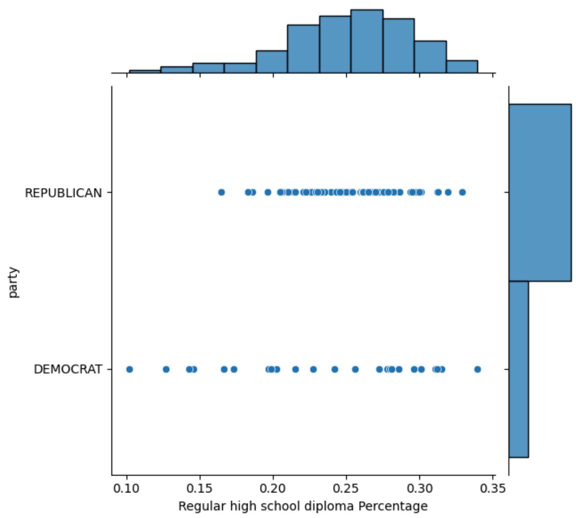
North Carolina

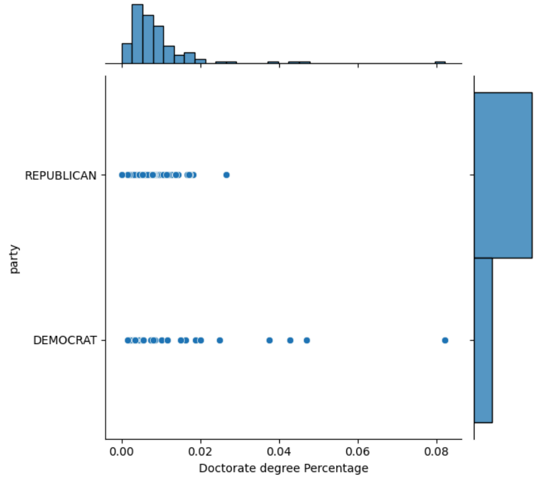
North Carolina is a southeastern state along the Atlantic Coast with a population of 10.4 million people and consists of 100 counties. Historically, North Carolina has been a battleground state in elections, fluctuating between Democratic and Republican.



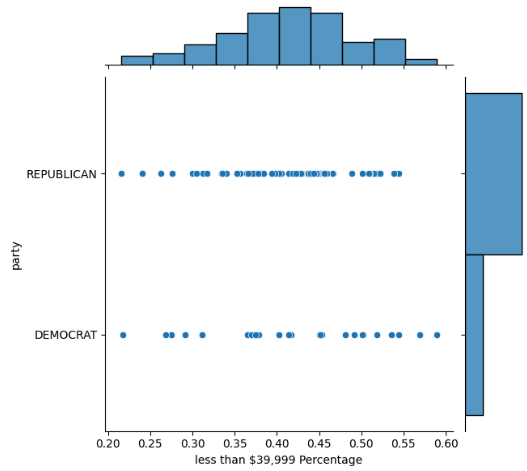


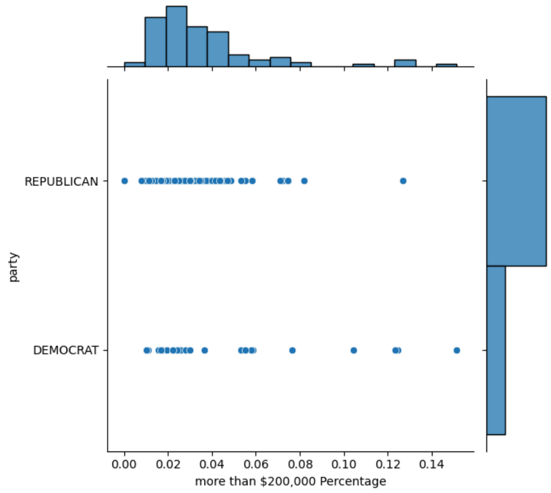
Pictured above are two boxplot graphs for the 2020 presidential election, displaying the relationship between race and party. The plot on the left shows that Democratic-voting counties tend to have a lower median white-alone percentage, whereas Republican counties have a higher concentration of white-alone populations. Similarly, the second box plot depicts that Democratic counties show a higher median of Black residents, while Republican counties exhibit much lower proportions of Black residents. This aligns with hypotheses 5 and 6.



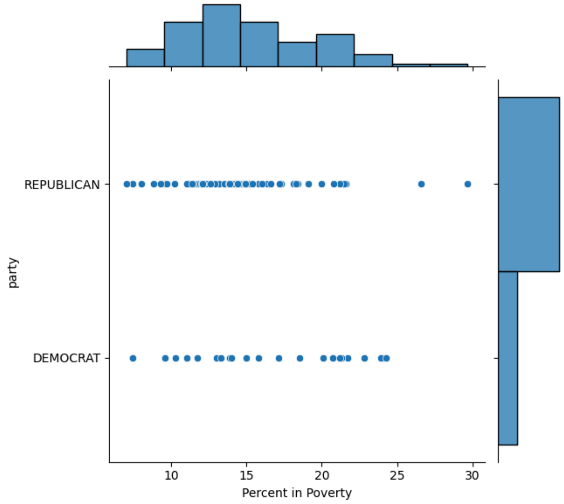


The two joint-plots above showcase the 2020 presidential election results, primarily focusing on the relationship between educational attainment level and party voting patterns. The plot on the left showcases how a higher percentage of Republicans holding only a high school diploma tend to vote Republican. The plot on the right demonstrates that counties with a greater percentage of people holding doctorate degrees tend to vote Democrat. This suggests higher education levels are associated with Democratic voting patterns, aligning with Hypothesis 7.





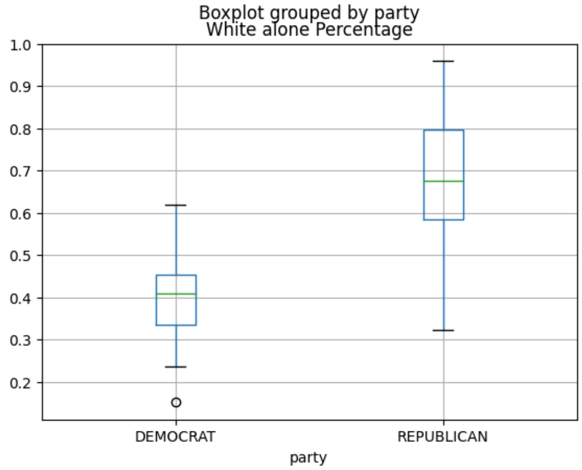
The two joint-plots display the 2020 presidential election results with a focus on the relationship between income brackets and voting patterns. The graph on the left showcases that both Democratic and Republican counties exhibit a wide range of lower-income populations, and there is no clustering or trend that would suggest a clear difference in income distribution between both parties. Similarly, the graph on the right does not have a clear concentration. As a result, these plots neither prove nor disprove Hypothesis 8.

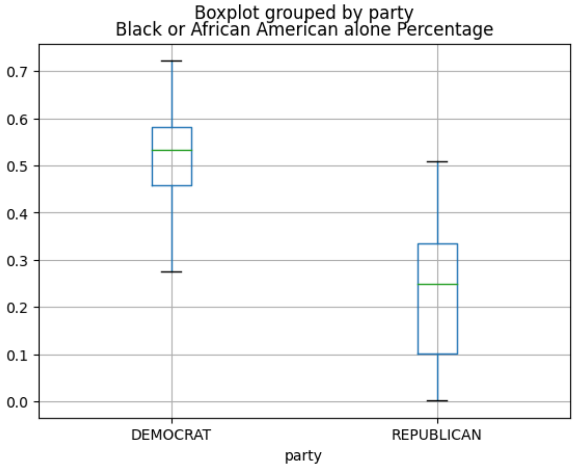


The joint plot above displays the 2020 presidential election results with a focus on the relationship between poverty and party voting patterns. The graph indicates that counties with a higher percentage of poverty tend to lean Democratic, whereas counties with a lower percentage of poverty tend to lean Republican. However, the correlation is not particularly strong. Overall, this graph supports hypothesis 11 but the distribution showcases that the data does not perfectly align with the hypothesis.

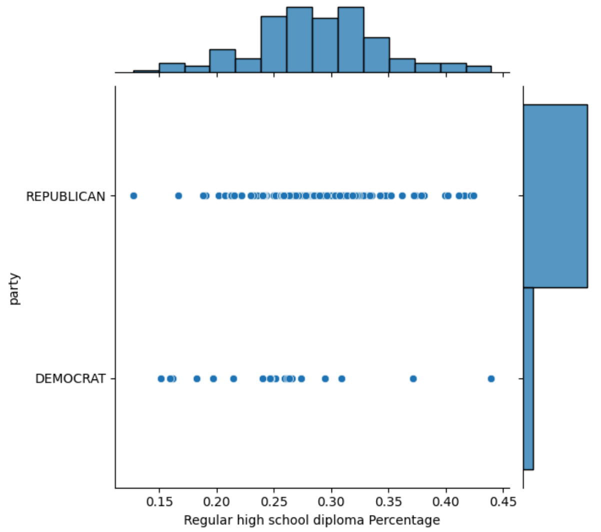
Georgia

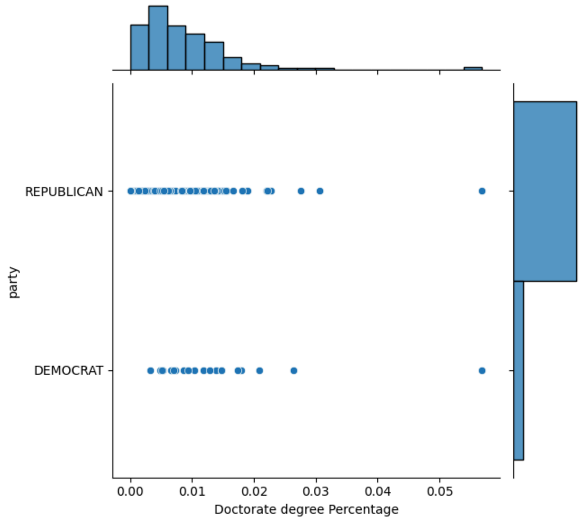
Georgia is a southern state along the East Coast with a population of 3.8 million people and consists of 159 counties. For many years Georgia was a consistently red state but in recent elections, such as the 2020 presidential election, Georgia has become a purple swing state.



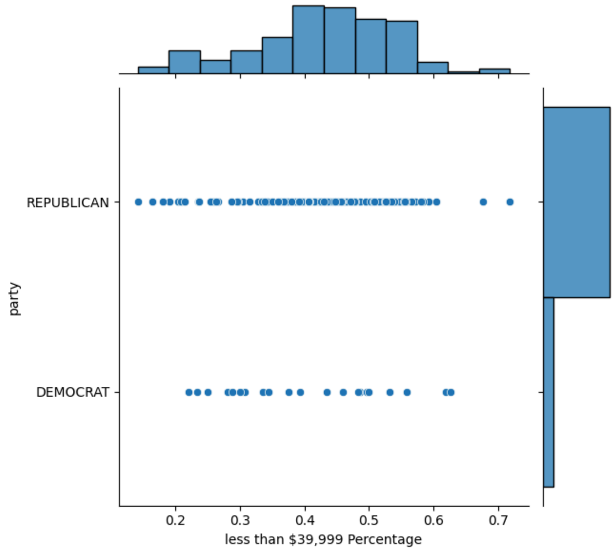


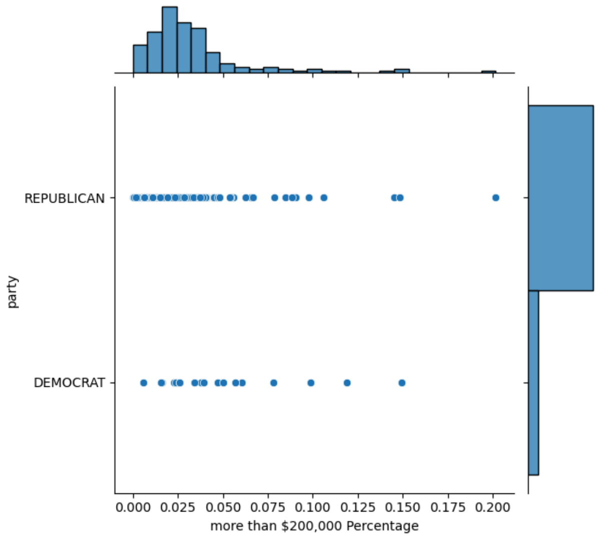
Above are two box plot graphs for the 2020 presidential election election, displaying the relationship between race and party. The graphs demonstrate how Republican counties had higher proportions of White civilians than Democratic counties. Conversely, Democratic counties had higher proportions of Black civilians than Republican counties. These findings support Hypotheses 5 and 6.



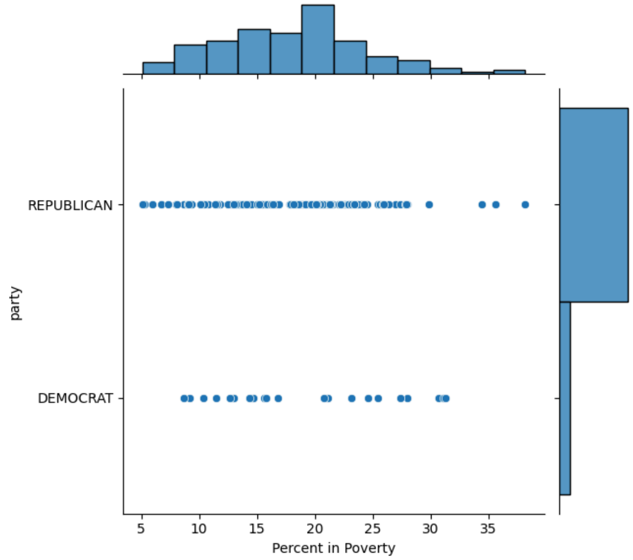


The two joint plots above display the 2020 presidential election results with a focus on the relationship between education and party voting patterns. Based on the graph on the left, it can be inferred that counties with higher proportions of civilians whose highest level of education is a high school diploma tend to vote Republican, while the graph on the right shows that counties with higher proportions of civilians who have earned a doctorate degree tend to vote Democrat. These results support Hypothesis 7.





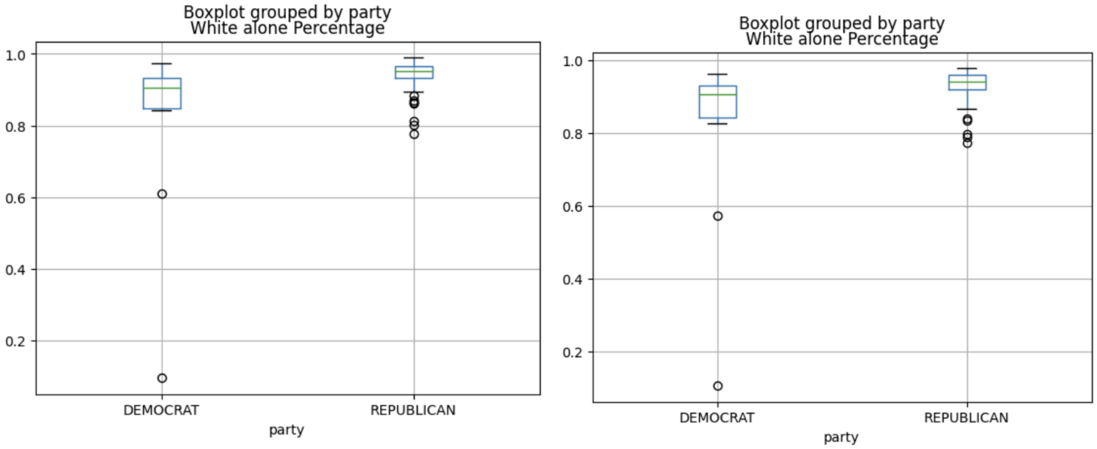
The two joint plots above display the 2020 presidential election results with a focus on the relationship between income bracket and party voting patterns. Neither of the two graphs show a significant correlation between income and party outcome.



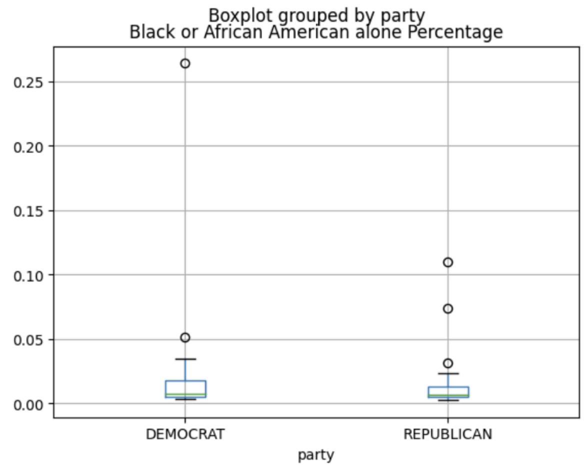
The joint plot above displays the 2020 presidential election results with a focus on the relationship between proportion of civilians in poverty and party voting patterns. However, the findings show there is not a clear correlation between party outcomes and poverty.

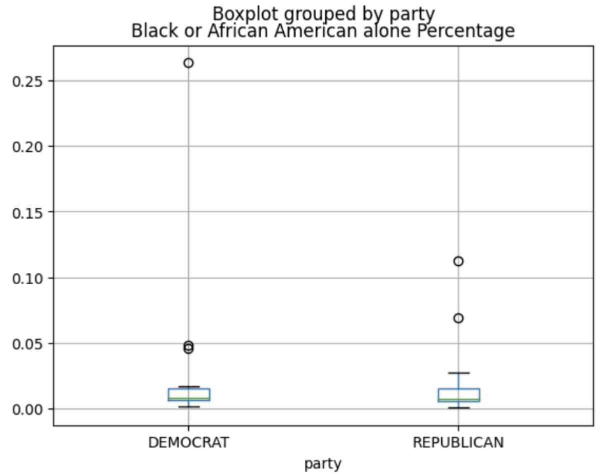
Wisconsin

In Wisconsin, voting patterns among racial groups show a clear divide, particularly between White and Black voters. The "White alone" population in the state tends to lean Democratic, although the margin is not overwhelming, with many White voters supporting the Republican Party as well. Factors such as education level, urban versus rural residency, and income levels often play a significant role in determining political preferences among White voters.

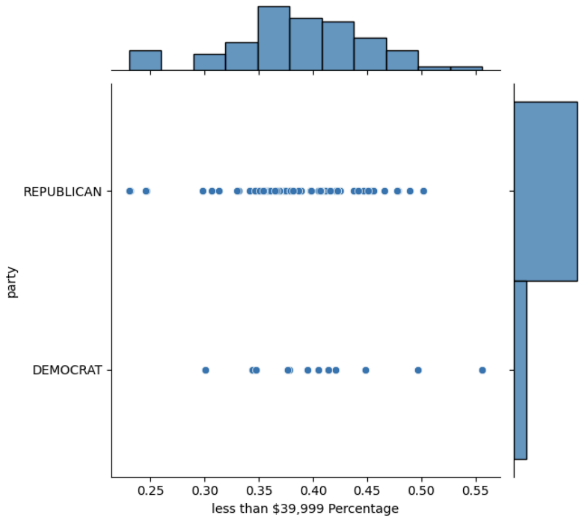


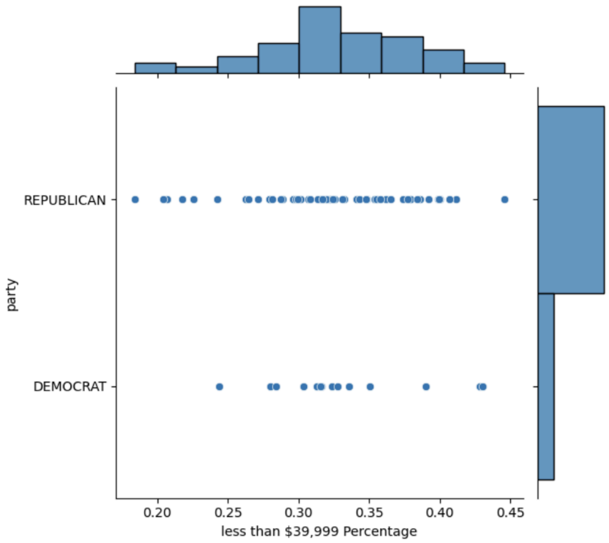
In contrast, the Black population in Wisconsin votes overwhelmingly Democratic both in 2016 and 2020. Historically, the Democratic Party has been seen as more aligned with issues of civil rights, social justice, and policies that address economic disparities, which may explain the strong Democratic loyalty among Black voters in the state. Despite making up a smaller percentage of the overall population, Black voters often play a crucial role in urban areas such as Milwaukee, where their support for Democratic candidates can be pivotal in close elections.





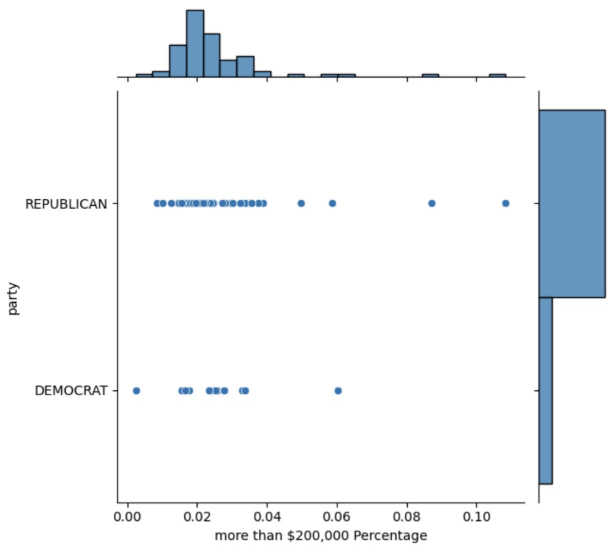
The voter data from Wisconsin in 2016 and 2020 offers valuable insights into the relationship between income levels and party affiliation, shedding light on the state's political landscape. The graphs focus on two notable income brackets: those earning less than $39,999 and those earning more than $200,000, illustrating the voter distribution for both Republican and Democratic supporters. One of the most striking features of this data is the clear income divide between the two parties. Republicans demonstrate a strong presence across both income brackets, with a particularly notable concentration in the lower-income category. In fact, Republican support in the less-than-$39,999 bracket hovers around 40-45% in both elections, suggesting a consistent and significant base among lower-income voters. This trend may reflect the resonance of Republican messaging on issues such as job creation, traditional values, or anti-establishment sentiments with this economic demographic.

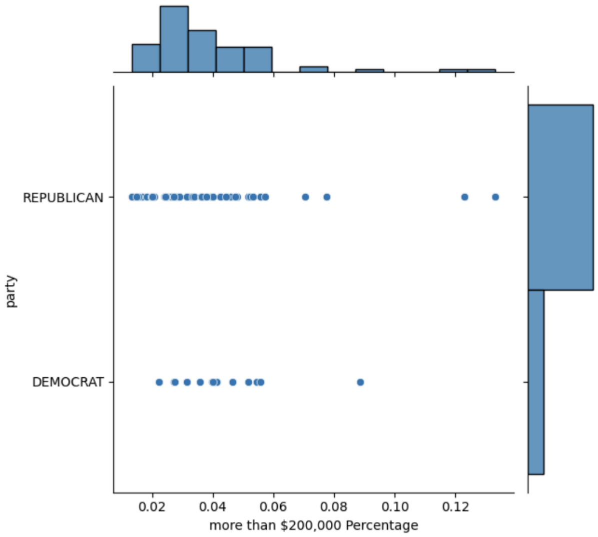




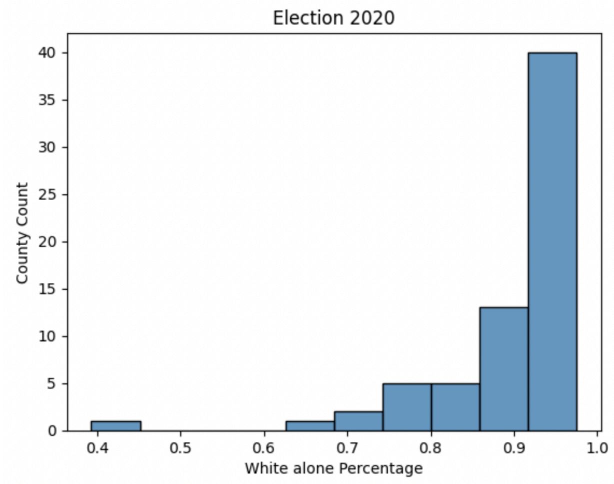
In contrast, the Democratic Party appears to have fewer voters in both income brackets, particularly in the lower-income group. This raises important questions about the party's ability to connect with and mobilize lower-income voters in Wisconsin—a demographic traditionally associated with left-leaning policies. The scarcity of Democratic voters in this bracket may be due to several factors, including challenges in voter turnout, messaging, or shifting party appeal. The persistence of this pattern across both the 2016 and 2020 elections suggests it reflects a structural, long-term issue in Wisconsin's political dynamics rather than a temporary fluctuation.

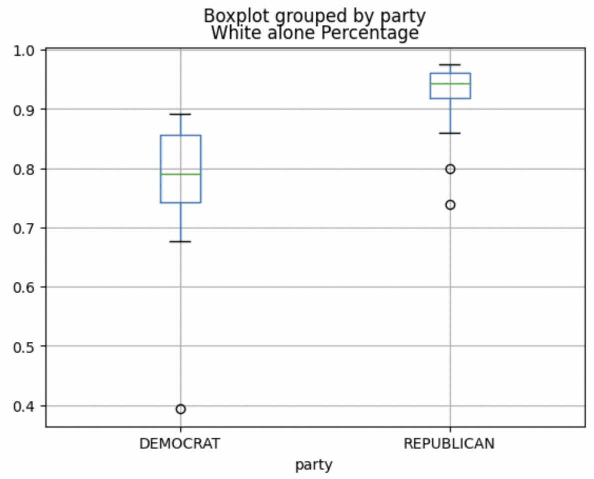
Looking at the higher-income bracket (above $200,000), a different but equally revealing picture emerges. While both parties show fewer voters in this category—a natural outcome given the smaller population size—Republicans again demonstrate a stronger presence. This suggests a fascinating economic polarization within the Republican Party in Wisconsin, as they manage to appeal to both ends of the income spectrum. By addressing diverse concerns such as tax policies that benefit higher-income individuals alongside economic opportunities for lower-income groups, Republicans have crafted a message that resonates with a broad range of voters. The consistency of these patterns between 2016 and 2020 is particularly significant. Despite major national events—such as the Trump presidency, economic shifts, and the COVID-19 pandemic—voting behavior in Wisconsin remained largely stable. This stability points to deeply entrenched political affiliations that appear resistant to short-term events or policy changes. These trends hold considerable implications for future policy-making and political strategies in Wisconsin. For Republicans, there may be pressure to maintain a delicate balance between catering to the differing needs of both lower-income and higher-income supporters. Democrats, on the other hand, face the critical challenge of broadening their appeal among lower-income voters without alienating their current base.

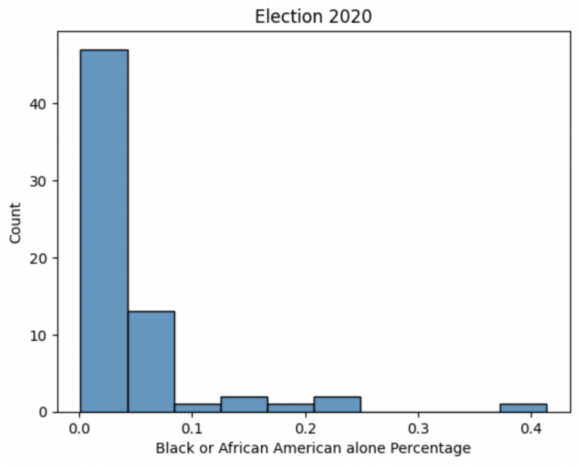


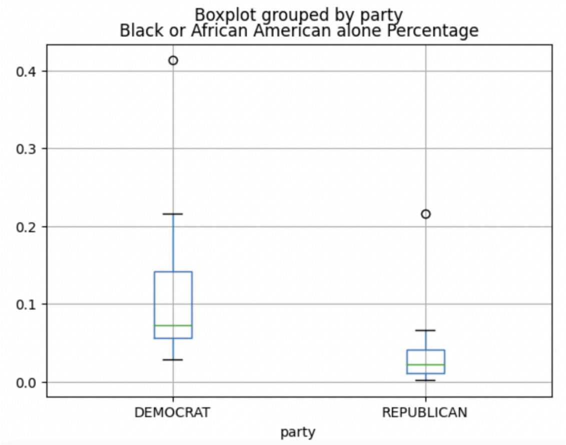


Pennsylvania

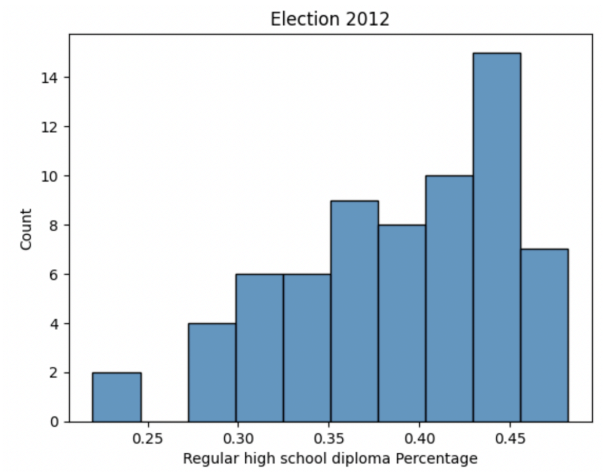


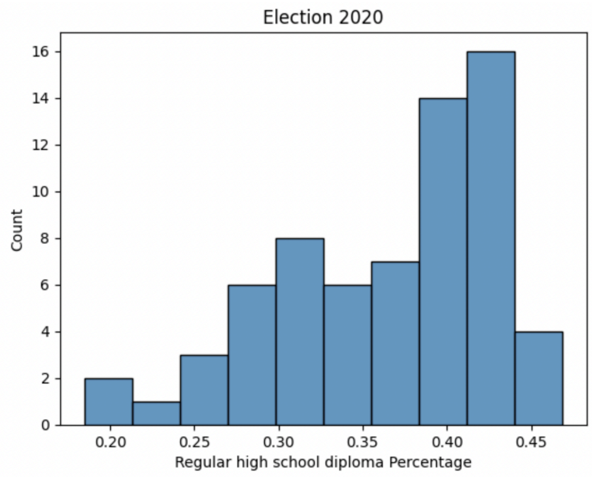


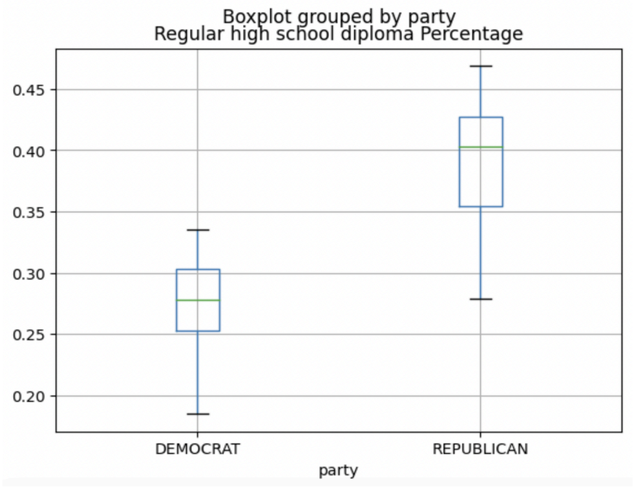


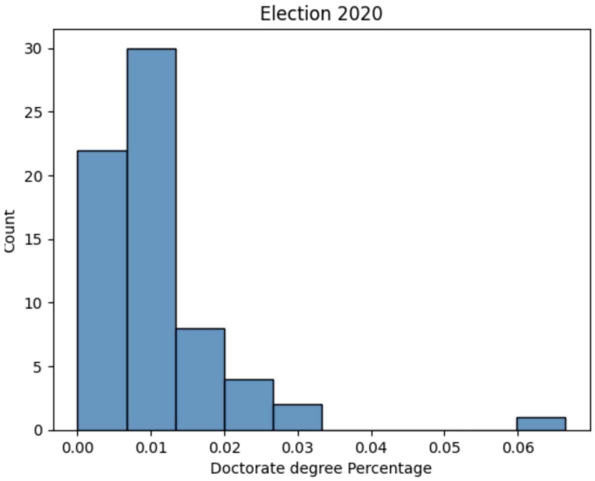


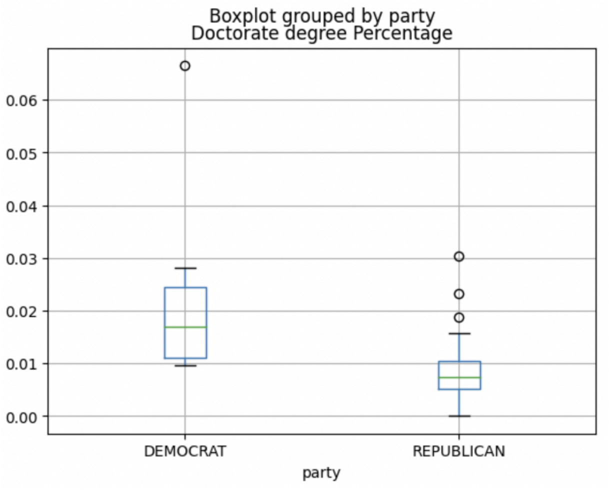
In Pennsylvania, voting patterns among white voters have remained consistent across the last three presidential elections, with a clear tendency to lean Republican in all counties. In contrast, Black voters show a stronger Democratic preference. However, there is also a significant disparity in voter participation, with white voters making up a larger share of voting individuals in Pennsylvania, reflecting their higher percentage of the state's population. While there are more white voters on average, this aligns with a greater white population making up the demographic composition of Pennsylvania. The lower voter turnout among Black voters adds to the disparity in political influence between racial groups. These disparities offer a broader context for interpreting the data.



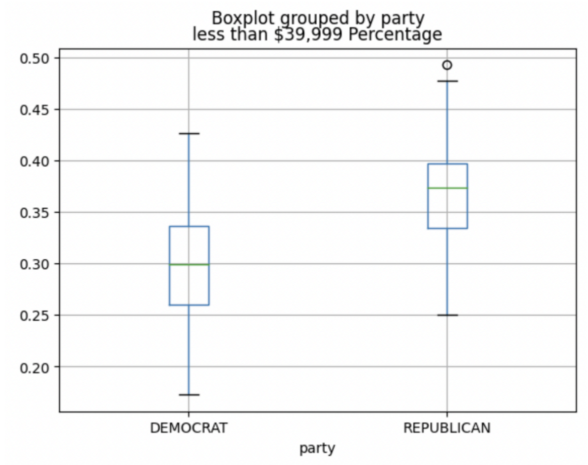


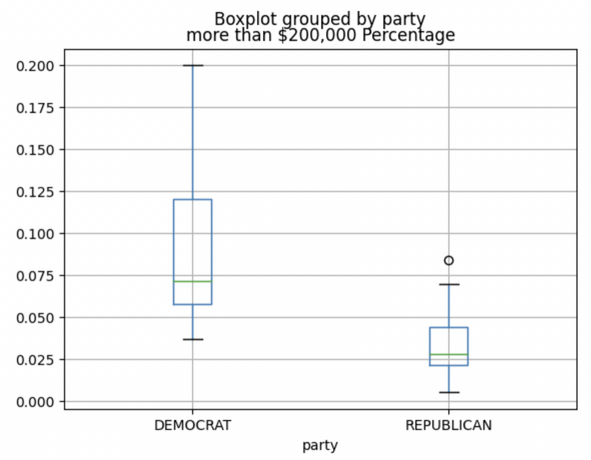






In terms of voting patterns based on levels of education, there seems to be a declining trend among those who have a regular high school diploma voting in presidential elections. Greater voter turnout was present in the 2012 presidential election, but has since decreased with each passing election. Additionally, those who hold degrees beyond a high school diploma, specifically doctorate, lean more Democratic on average. Meanwhile, those who hold only a high school diploma heavily lean Republican. This contrast highlights a growing educational divide in voting behavior, with more educated voters aligning with progressive policies and platforms, while voters who haven't pursued higher forms of education trend more conservative.





These trends also align with voters' annual income. Those with greater incomes, who earn more than $200,000 a year, vote more Democratic, while those who earn less than $40,000 vote more Republican. This reflects a broader narrative where the working class, particularly those with lower levels of education and annual income, often supports Republican candidates despite this seeming to conflict with their economic self-interest.

IV. Challenges

As stated earlier in our data collection section, we encountered a problem in the data due to each county recording counts of individuals for each of the different variables, rather than the proportions. To eliminate the weights based on the population, we had to convert these counts into percentages relative to the population in each county for a view that can be analyzed across each county.

Another challenge we encountered was with the years of the data we collected. We collected data from 2012, 2016, and 2020 but it is important to note that two of these years can classify as an anomaly. In 2012, the economy and market was rebounding from the recession which could represent higher economic variables. Similarly, in 2020, the economy was in the midst of a global pandemic which led to drastic abnormalities in the data. Being said, we are still in the process of deciding how to proceed with these years and if we should take these into account. Any year prior to 2010 may be too old and not a good representation of today's voting as well.