Rix Prakash, Akan Ndem, Nathan Wan, Ali Nilforoush
DS 2002: Data Science Systems
Professor Williamson
5 December 2024

The Demographics that Swing States

Data Selection & Exploration

After the results of the election, the United States saw every single swing state (Arizona, Nevada, Georgia, Michigan, Wisconsin, North Carolina, Pennsylvania) turn red. Historically speaking, after two to three presidential elections, we've seen swing states shift to a dominant stance on a particular party while some other states change into a closer race. For example, back in the 2000 presidential election, the swing states consisted of Florida, Iowa, Minnesota, New Hampshire, New Mexico, Oregon, and Wisconsin which were all within ~3% margin between both parties. Usually swing states are a battle and one party slightly wins by a small margin, however in the 2024 election this wasn't the case with all swing states voting red. With this in mind, we decided to look at the past 3 elections to see if there were any trends in demographics that led to a party change with the current day known swing states.

When we searched for our data, we had to collect two different datasets that we later had to merge. First, we collected demographic data for each county in all 50 states in the United States from Integrated Public Use Microdata Series (IPUMS), which provides census and survey data from around the world. Our variables of interest were the proportion breakdown of age, gender, education attainment, income, race, and labor force participation. These are all percentages of the county total population. Our second dataset contained each county's winning political parties in each state in the United States from the Inter-university Consortium for Political and Social Research (ICPSR). Both datasets had three common identifiers (columns) which allowed us to seamlessly merge the two datasets by their year, state, and county matches.

Our expected insights were to see the counties in the swing states change towards something that was more polarized in political outcomes. For example, we expected to see the swing states in 2020 to have more of a split of republican and democrat counties compared to in 2012 when they may have not been truly identified as a swing state. Overall, we expected to see a shift from 2012 to 2020 in these total demographic data for each county. Throughout our analysis, we are analyzing whether historical stereotypes of demographics hold true in the elections. We are going to analyze the following historical labels:

1. The Republican party consists of older men
2. The Republican party consists of mainly the white racial background
3. The Demographic party consists of more educated individuals
4. The Republican party consists of wealthier individuals

Keep in mind that these expected insights are all based on the assumption that demographics depict political beliefs and outcomes.

ETL Setup

The setup for the ETL pipeline depended on the format of the data and which data would be stored. Since our datasets were structured and tabular, we decided to use MySQL as the database solution. We also decided not to store our analysis graphs in the database or on Google Cloud Storage, focusing on storing the cleaned and transformed dataset. Although it is possible to initialize a local MySQL database using MySQL Workbench and store data locally, we chose to host the database on a MySQL instance in Google Cloud so that all of us had access to the data.

We uploaded a copy of the merged dataset, the output of the ETL pipeline, to Google Cloud Storage and Github. We made sure to consider reproducibility and scalability by allowing for altered inputs within the transformation part of the pipeline. Additionally, we implemented basic security measures within Google Cloud, such as password protection to manage access.

In our next step we decided to clean, normalize and merge the data in python

ETL Implementation

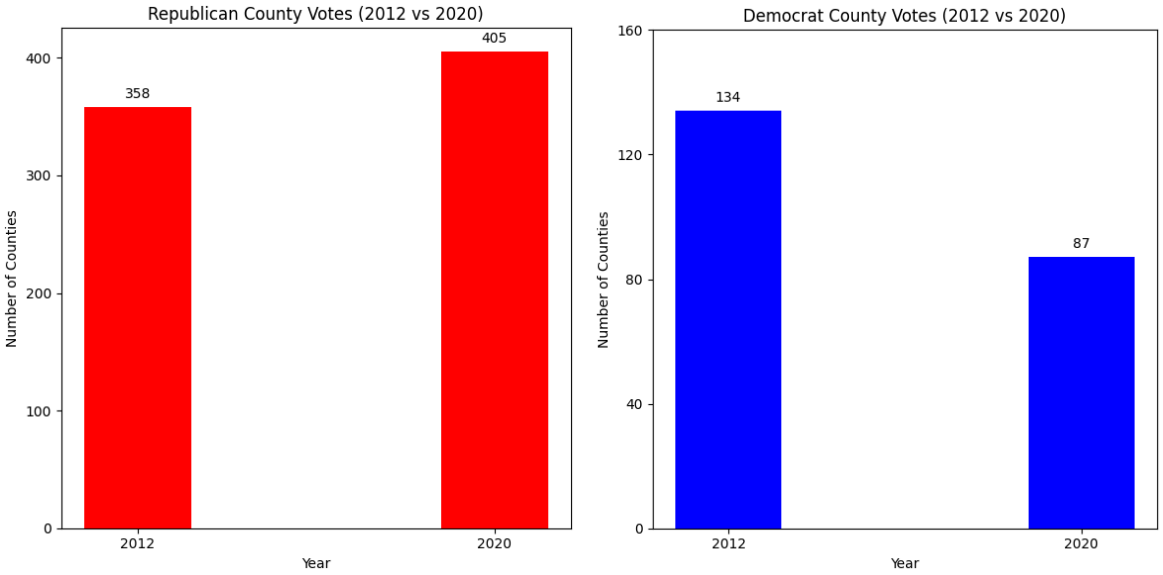
Regarding the ETL pipeline implementation up we had essentially three big stages. In the first stage, we collected 4 datasets with the three being the county demographics across all swing states (Nevada, North Carolina, Wisconsin, Michigan, Pennsylvania, Georgia, Arizona) in each of the 2012, 2016, and 2020 elections then a winning counties dataset containing the county votes and candidate wins for all swing states for all three elections. We then had to load these datasets into the pipeline as data frames and validate the data to prepare the data for cleaning, normalization, and merging.

The next stage of implementation dealt with first cleaning the datasets of unnecessary column entries across all four datasets. After keeping necessary columns, we normalized the county and state names across all four datasets and added a year column to the demographics datasets to prepare for merging. After merging the datasets by year (a merged dataset between demographics of a single election and counties votes of that election), we then merged them into a larger dataset for all 3 elections. This was the final output of the ETL pipeline

The final step of the ETL pipeline was creating a SQL data table and storing the dataset on Google Cloud. We connected our Google Colab runtime to a MySQL instance on Google Cloud by easing incoming IP restrictions and initializing required credential variables for connection. We then created a data table using SQL in python ensuring matching column names and typing of the entries. The final data set was then added to this data table and the number of row entries was verified and came back successful.

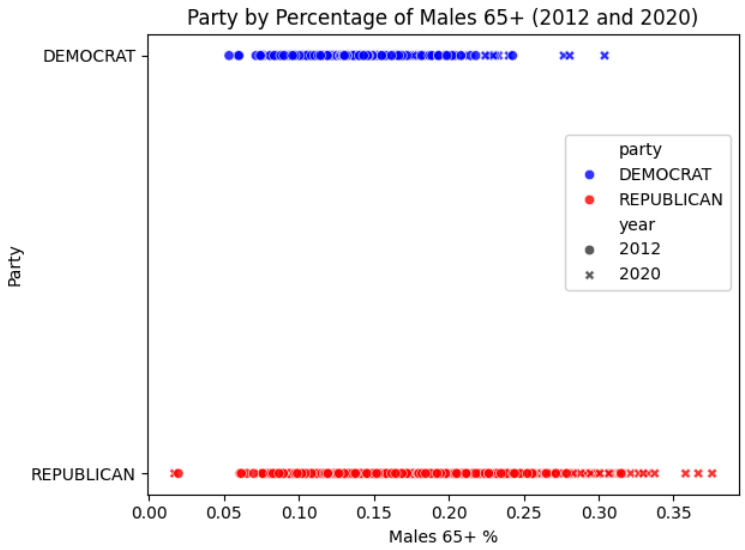
Data Analysis

To begin our analysis on how the current day swing states changed from 2012 to 2020, we decided to count the number of counties that voted democrat and republican in each election. Our findings were consistent with our original insight and hypothesis of seeing a meaningful change in political outcomes, with an increase in roughly 50 counties switching to republican from democratic.



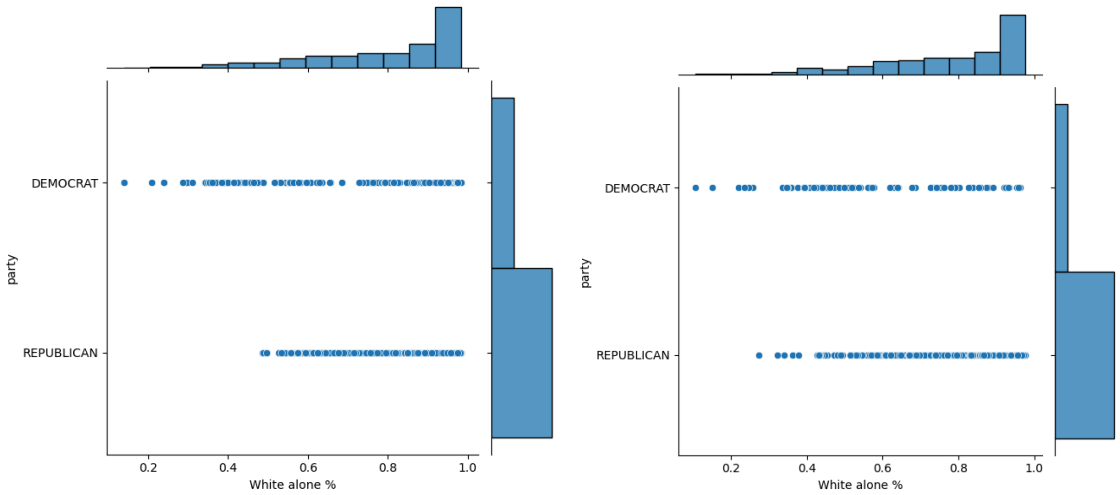
However, it is important to note that a higher number of Republican counties doesn't necessarily translate to more votes, as population density varies significantly. A few populous Democratic counties can outweigh many rural Republican ones which have followed historical trends of rural parts of the US being red and urban cities being blue. This distribution can highlight the urban-rural divide in political preferences and influence campaign strategies.

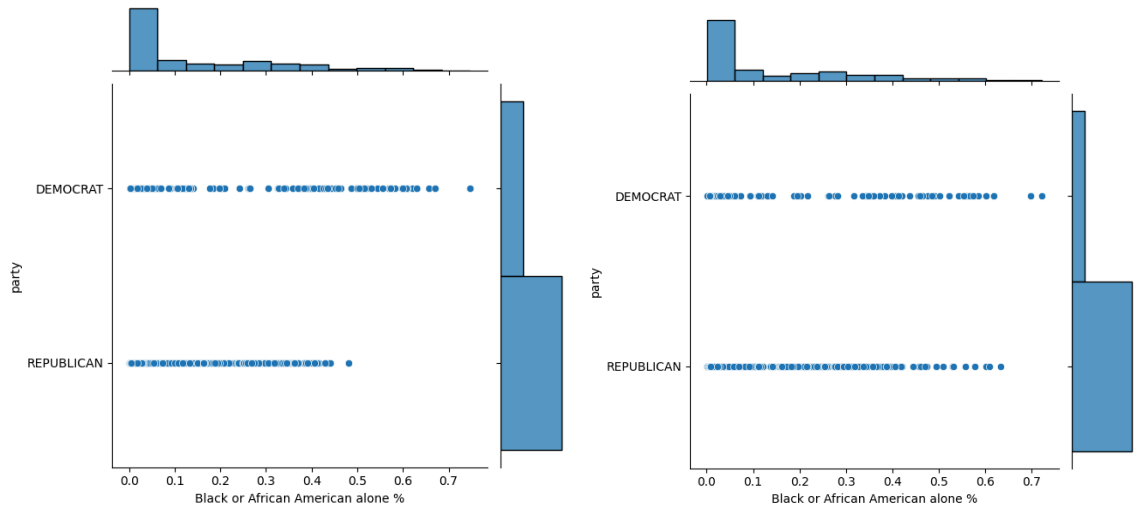
After noting this, we narrowed our demographic variables of interest to focus on age, gender, race, education, and income brackets. When we looked at age across all counties in the swing states, we found that 2020 recorded more Republican counties with a greater male demographic older than 65 compared to 2012. However, the democrat proportion stayed relatively the same, which could mean counties with older men tend to vote republican.



We can see with the graphic the abundance of x (which represents 2020) for the republican party vs the democrat party, indicating that counties with older men tend to vote republican. This stereotype is strong and seems to have more trends moving in this direction.

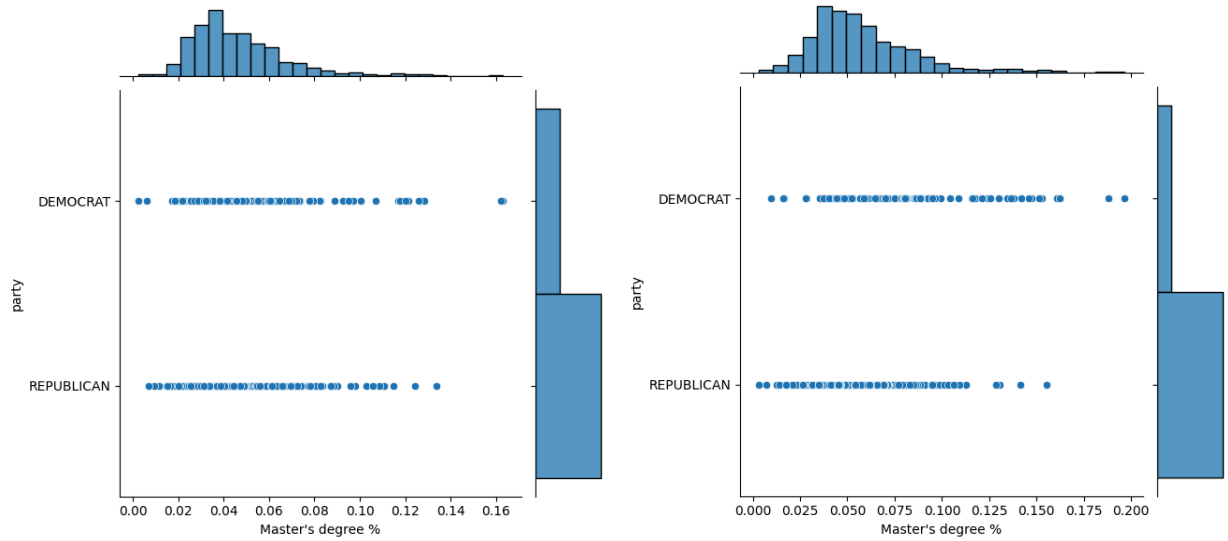
As we move onto race, we had the original insight and stereotype that white folks vote republican and more diverse counties vote democrat. However, what we found here was on the contrary and flipped this narrative.





The visualizations on the left are from 2012 and the visualizations on the right are from 2020. We found that there were more republican counties with a lower number of white racial status, indicating a shift towards more diverse racial counties voting republican. Similarly, republican counties with greater black racial status grew, confirming our previous statement. The Democratic maintained their diverse racial status, but what is meaningful to note here is that the Republican party is becoming more diverse, breaking stigmas and stereotypes set in the past historical elections.

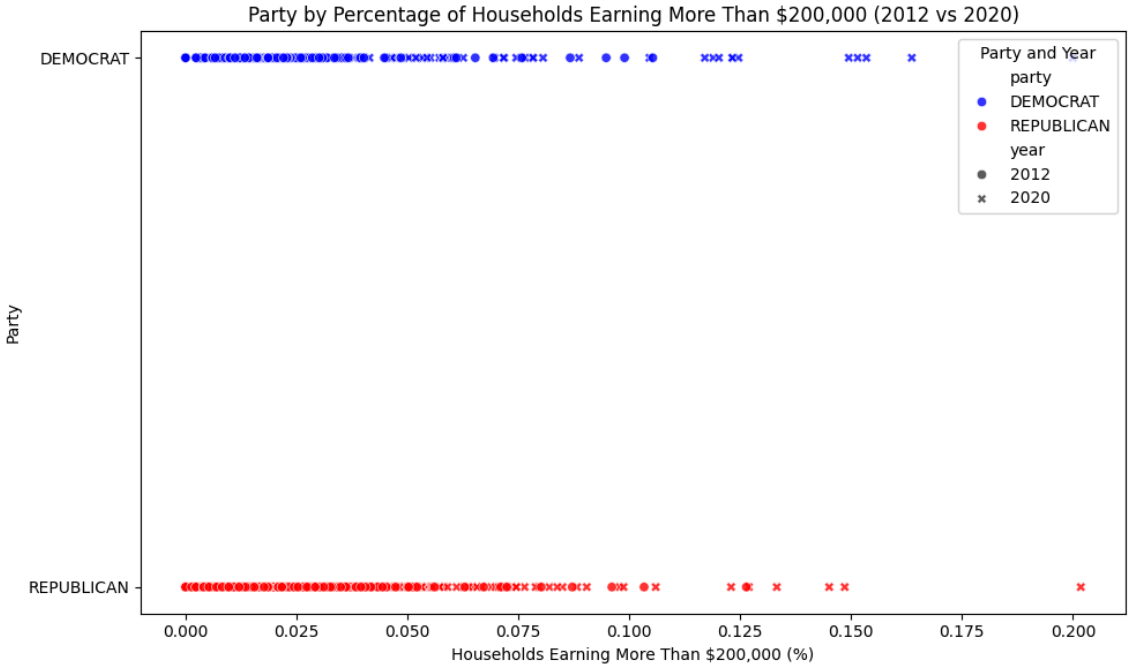
While we talk about stereotypes, it is commonly believed that Democratic counties and folks are more educated than the Republican counterparts. When we dove into this assumption, we found this stereotype being bought into with more data affirming this.



On the left, the 2012 election showed no significant affirmation that having a more educated degree, such as a masters, would affect the political party outcome. However, in 2020, counties with a greater proportion of masters degrees tended to vote Democrat while the counterparts

exhibited roughly the same proportions. This confirmed the original stereotype that more educated individuals and counties tend to vote blue.

Finally, we dove into the income levels to analyze whether the Republican party consisted of weather counties. What we found was that although in 2012 the Republican party had wealthier households, in 2020 we saw more wealthy counties voting Republican.



Despite the Republican party still consisting of wealthy countries, the Democrat party caught up and includes similar wealth status in counties. Thus, the stereotype does not hold true in the swing states, and it is still important to note this shift from 2012 to 2020 with weather counties voting Democrat.

With all these analyses, we can conclude that there was a drastic shift in the swing states from 2012 to 2020 and can expect shifts to occur overtime with the next batch of swing states. If we could run this over again, we would be interested to include data outside of demographics that influence voters such as the economy/prices in the counties, GDP, tax rates, campaigns efforts in the counties, and more.

I. **Cloud Storage & Documentation:** Document the process, including credentials management and access control

Our code starts with setting up authentication by using the Google Cloud service account key using an OS environment variable, which ensures secure access to the Google Cloud project. We then initialize the client which allows us to use operations such as listing buckets and uploading files. We listed the buckets to ensure we are targeting the correct bucket and that it exists. We also list the bucket contents to verify successful uploading. We uploaded the files to

the bucket by using google cloud storage library functions. The credentials and access control are handled using the service account key, which only allows authorized users to perform bucket operations. The bucket permissions were handled during the set up of the bucket to allow reading, writing, and deleting.

II. **Reflection Paper (2-4 pages):** Reflect on the challenges faced during data selection, ETL setup and implementation, analysis, and cloud storage, lessons learned, tech challenges, team coordination, and any improvements for future projects

Our project focused on the relationship between voter demographics and election results data, which is a simple concept, but in practice is not so straightforward to analyze. What exactly we were searching for within the data was not clear at first, and the data being overwhelmingly large in size did not help narrow down our focus. The first conceptual problem we faced was during data selection, where did we want our data to come from? Considering the nature of the U.S. political scene in general, normally candidates tend to show up in swing states, aka states they believe they can potentially win over to secure their victory. Due to this, the group decided to focus on swing state data only, ignoring the other states, as insights on these states would be the most relevant and useful in practice. Since we were left with a sample of only a few states, if we were to attempt to establish relationships between demographics and the winning party in each state, we would only have a few data points to go off of. Therefore it was decided to go one step below the state level and investigate counties.

After gathering up our data like this, the next choice we were faced with was what variables to investigate. To arrive at a decision, we referenced typical stereotypes that people make when talking about a political party, these included things such as “the republican party consists mostly of old, white, men” and things like “people with higher education tend to be democratic”. We thought it would be interesting to investigate these stereotypes, in a sort of fact check, to see if they had any basis. Thus, we derived our demographic variables of interest as race, age, education, and income.

In order to accomplish our analysis, we had to aggregate our data all together, which included not just the related demographic variables and party that won, but also included supplementary files like geojsons that would allow us to make folium maps to give a map-like visual of our data. For this we need a proper ETL pipeline that could take in spreadsheets, jsons, etc. and format them into dataframes to be able to merge them all together. After the data was ingested, cleaning and merging it was our first obstacle, as we first had to identify what columns pertained to what variables, e.g. NAMES was state names in one dataset but names was county names in another. After merging based on county names and ensuring the data was in the correct format and had acceptable values (we wanted proportions in our case not actual counts), we saved the data to a MySQL instance that also was stored in Google Cloud. This involved configuring

credentials and easing ip constrictions in order to store our final dataset post-ETL process in the cloud.

Moving beyond initial stages, and going into the meat of the project, during the actual analysis, we came across the problem of how exactly we wanted to pursue our project goal of investigating the aforementioned political stereotypes. We knew we wanted to see if general trends held based on our research stereotypes and so we came to the conclusion we needed a graph that could plot instances of counties based on their fit to the demographic based on the stereotype and the winning party in the county. In this way, we could see the distribution of the demographic proportions for all the counties for each of the political parties, letting us draw general conclusions on the possible validity of the stereotype. If we saw a higher density of counties for the republican party that had a larger portion of the demographic compared to democratic, we could say the stereotype for that demographic does have some validity for the republican party. For example, we saw many counties that had older men aged 65+ consisting a larger portion of the population vote republican compared to democrats, so the conclusion that older men tend to vote republican seems to have validity.

We learned that sometimes you have to rework your research questions because you realize the analysis needed would require too many resources, and that just because you can research something doesn't mean you should. This was the primary reason we narrowed our focus to swing states, because typically these are the ones politicians are concerned with and hold the most power when deciding election results. Although investigating all 50 states could have been possible, it wasn't very practical due to the larger amount of resources needed in terms of how many graphs and data points we would need to interpret. We also learned that assigning roles is vital for a group project, as a sort of situation where everyone just does what they can doesn't really work out because then no one is responsible for any particular section, which can lead to some oversight.

We believe that our team coordination could have been slightly better if we made a schedule and did more discussion on where and when we could meet up rather than just asking everyone to meet up randomly, and also that being in person was probably better as well to make sure we could hold each other accountable for being present and doing work.

For future research on the topic, we think it would be interesting to incorporate more outside data from different sites that could illuminate not only the presence of these trends in voting patterns but also why. An example could be incorporating economic data, if the economy was doing especially well during one of the voting sessions, perhaps that could explain the voting behavior of the more wealthy individuals for one party. Furthermore, we could expand further and potentially have an algorithm such as random forest identify relevant demographic factors that have high predictive power on the candidate voted for, it would be very interesting to look at.