Tori Fredell, Kelsey Matsik, Ashleigh Curry, Caroline Wynne, Theresa Trinh
Professor Johnson
DS 3001: Introduction to Machine Learning
May 3, 2024

Voting Project

Summary

In our study, we focused on analyzing the effects of demographic and socioeconomic factors on voting margins in Virginia's presidential elections. We used datasets with information from the years 2000 to 2020 in order to predict future election outcomes based on these variables. Our cleaning process involved removing columns that were irrelevant for the purposes of our models, re-writing column names for clarity, handling null values, and addressing inconsistencies across the different datasets. A significant piece of our data preparation involved merging the cleaned voting data with county-level demographic data, which allowed us to integrate data concerning race, income, and educational attainment with electoral results. Then, we used linear regression models to explore how various predictors influenced the voting margins, defined as the difference in votes between the Democratic and Republican candidates. Our models incorporated variables such as racial demographics, levels of education beyond high school, and income levels. Our analysis showed that these demographic factors are good predictors of voting margins in Virginia; they were effectively used in our models to predict electoral outcomes. Our results were substantiated by an R-squared value of about 0.832, meaning that about 83% of the variability in voting margins could be explained by our selected predictors. While our models established the importance of these factors as predictors, it ultimately did not measure how variations in these demographics correspond with increases or decreases in voter

margins. An additional limitation stems from a high Mean Squared Error (MSE), suggesting the possibility of other unaccounted variables that influence election outcomes. Moving forward, further research could benefit from incorporating additional data across more election cycles and using more complex models to identify patterns more effectively.

Data Cleaning

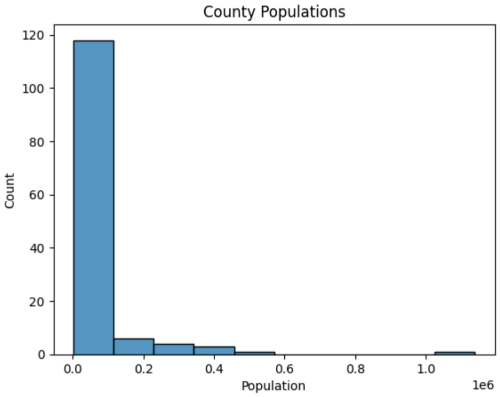
When cleaning the voting data for presidential elections for Virginia from year 2000 to 2020 (voting_VA), the first step we did was look at the overall dataset to see what columns were included. Afterwards, simply scanning the dataset led us to suspect that some columns may not contain any unique values. Columns such as “state” and “state_po” were dropped since the dataset is focused on presidential elections in Virginia; therefore, every row would contain the values “Virginia” and “VA”. The columns signifying “office” and “version” were also dropped due to 0 unique values, as the dataset was looking solely at presidential elections so there was no difference in the type of office each candidate was running for. The next step in the data cleaning process involved renaming columns to provide additional clarification as to what they were indicating as well as simplify them. For example, “county_name” was renamed to “countyName”; “county_fips” to “countyCode”, and “mode” to “voteType”. Additionally, capitalization of values in columns such as “countyName”, “candidate”, “party”, and “voteType” was completed for simplified reading of values. Furthermore, the “Unnamed: 0” column was dropped from the dataset. After further examination of the original dataset, we noticed that values for “voteType” for elections prior to 2020 looked at total votes across all methods of voting, while elections in 2020 distinguished between absentee, election day, and provisional voting when looking at the number of candidate votes. Therefore, we first filtered the dataset by election year to only include values from 2020. Afterwards, we summed up candidate votes across all 3 voting types for each candidate and changed “voteType” to total. This reduced the number of rows for each candidate in the 2020 presidential election from three rows to one row that signified the total number of candidate votes across all voting methods. Total votes remained constant and did not change from the original dataset. Next, we dropped the old rows pertaining to the 2020 election from the original dataset and added the updated rows corresponding to the 2020 election to update the dataset. Now, the dataset looked at the total number of candidate votes for each individual across all voting methods rather than making distinctions between different methods. Therefore, we also decided to drop the “voteType” column from the updated dataset, as it now contained 0 unique values since it reflects information totaled from all voting types. This allowed for simplification of the dataset. Finally, we checked for any missing values in the dataset. There were zero missing values across all variables within the dataset.

For the county_adjacenies dataset, the first step I did in the cleaning process was renaming the columns to clarify what they were indicating and for simplicity's sake. “Population2022” was renamed “Population”, “FIPS” was renamed “CountyCode” and each “N1, N2, N3…” column was renamed “Neighborhood1, Neighborhood2, Neighborhood3…”. Second, I looked at the value counts for each variable to have a better idea of the data types and values. Next, I checked for missing values in which a high percentage of missing values were found for Neighborhoods 2 through 12. To address this issue, I replaced all missing values in these Neighborhood columns with “nan”. I rechecked for missing values and found all columns had 0.00% missing.

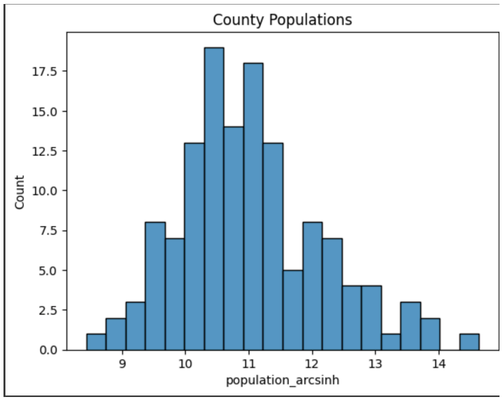
After building an initial model that regressed the margin by which the Democratic party would win (or lose) on year, we decided to add in additional data to our model. We decided to look at demographic data for the two most recent election periods: 2014-2018 and 2016-2020. We extracted variables relating to race, ethnicity, education, and income from these data sets because we thought it would be interesting to see how these variables affect how different Virginia counties vote. To get and clean these data, we first downloaded the Block Groups and Larger Area Estimates data sets for the 2014-2018 and 2016-2020. Because these data consisted of observations for every county in the country, our next step was to filter out only Virginia counties. Then, we concatenated the 2014-2018 and 2016-2020 data together to get one new dataset. Third, we subset variables of interest relating to race, ethnicity, education, and income, as well as GIS join code, county, county code, and year. Fourth, we renamed the columns from their American Community Survey codes to interpretable names. For example, we renamed “AMPWE002” to White, representing the number of white voters in a county. Next, we tried reshaping the dataframe. The data was organized so that each level of a demographic indicator was a column itself, rather than a value. For example, ‘white’, ‘black’, ‘american_indian_and_alaskan_native’, and so on, were variables that made up race. However, the data was structured so that for each county that did not record people of a certain race, there were NaN values. After melting the data to make one race column, it created extra NaN values in other identifier columns, such as County and Year. For this reason, we decided to leave the data in its original format. Our final step involved handling missing (NaN) values. In context, we interpreted each NaN value to be 0, so we replaced all missing values in the dataframe with 0. Though this might not be the perfect approach, it made the most sense given the constraints we were working under.

Visualization and EDA

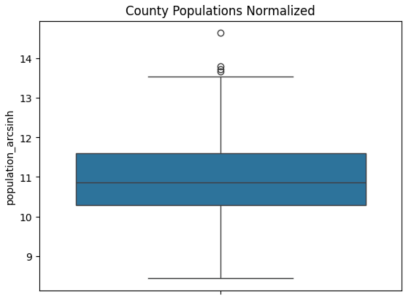
Our Exploratory Data Analysis and Visualization tools centered around normalizing the key variables related to county population and voting record, as well as utilizing Heat Maps. A first glance at the county population, total votes, and candidate votes variables indicated that they would need to be normalized via arcsin function. Those transformations can be seen below in Graphs 1 and 2.



Graph 1: County Populations Pre-Normalization

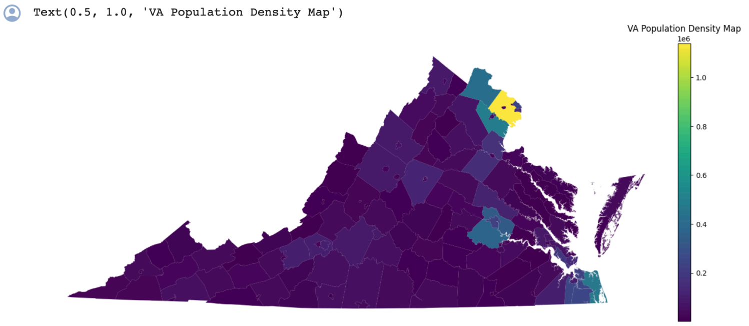


Graph 2: Normalized County Populations



Graph 3: Normalized County Populations

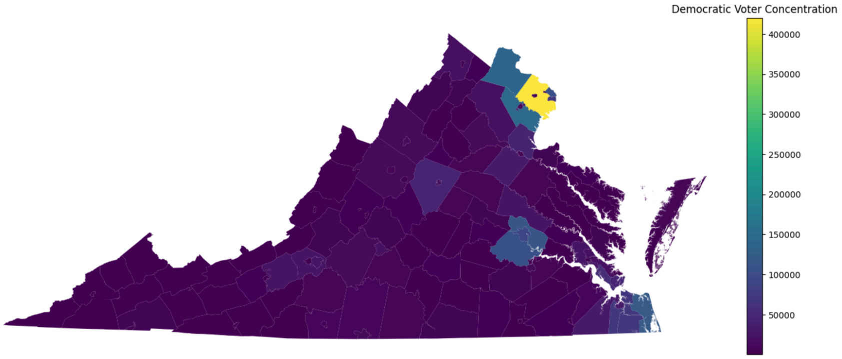
Given that the vast majority of counties fall under a population of 1,000,000 (1e6), normalizing the data creates a visual that can be compared to visualization of other key variables, such as total votes and candidate votes. We can also use other visualization tools to get a more complete picture of what our data signifies. For instance, the box plot and histogram of county populations show us that the vast majority of county populations tend to center around 500,000 (the mean is approximately 652,000), there are a few outliers that have significantly greater populations. The population heat map shows us where those clusters occur, specifically in Northern Virginia, as well as around Richmond.



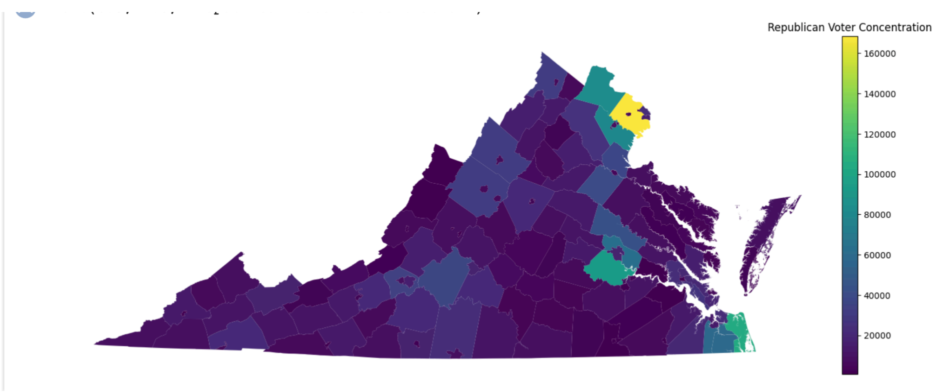
```
Text(0.5, 1.0, 'VA Population Density Map')
```

Graph 3: Heat Map of Population Densities

This visualization allows us to see population centers where a greater electoral body may be mobilized. A greater population represents a greater potential for election influence but only if voters are mobilized. Areas with greater population density also cluster around cities (D.C. and Richmond), which tends to correlate with higher levels of education and income as well. Graphs 4 and 5 show the concentration of partisan votes around cities as well. While higher clusters of each party correlates with population density, especially for Democrats, it’s interesting to note how Republican voter concentration increases as you head further down the state. Richmond has a higher number of Republican votes than Northern Virginia, outside of D.C. does.

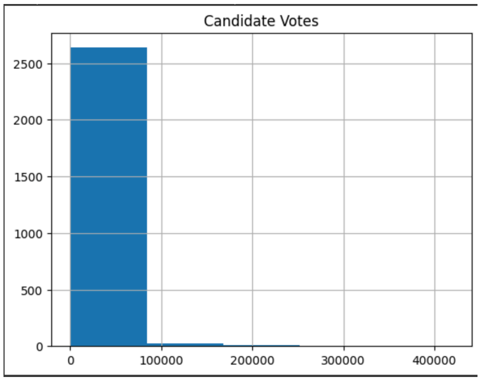


Graph 4: Democratic Voter Concentration

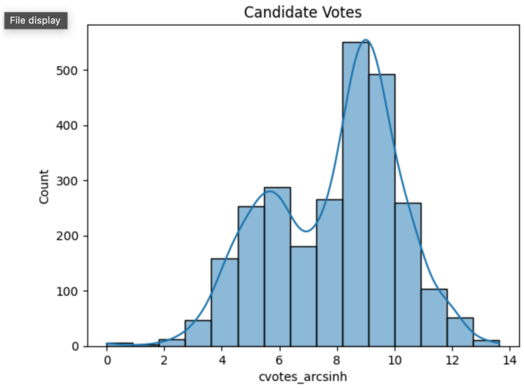


Graph 5: Republican Voter Concentration

Another avenue for EDA involves looking at patterns related to the total number of candidate votes and overall total votes cast within a given year. Normalizing candidate votes (defined as the number of votes given to a specific candidate in a specific county) helps us to get a more complete picture of the trends related to votes cast.

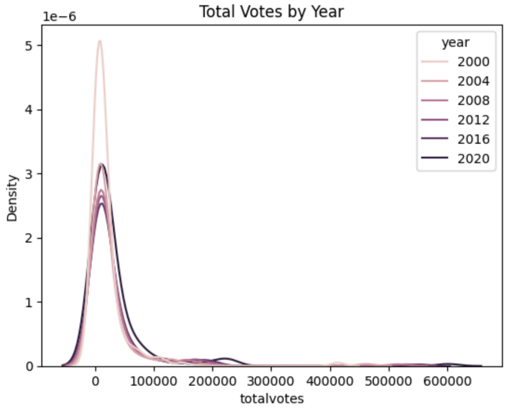


Graph 6: Candidate Votes Per County

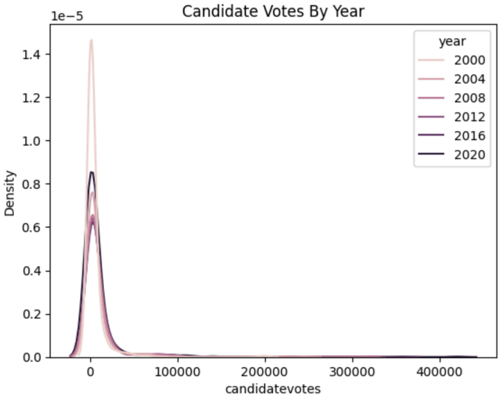


Graph 7: Normalized Candidate Votes Per County

By normalizing Candidate Votes we see that the distribution for the number of votes cast for a candidate is bimodal, with clusters around 5,000 and 8,000 votes respectively (the mean of the variable is 8206.7). We can then use this data to examine how the proportion of votes given to a candidate may change depending on the election year. Particularly relevant for this project are the years 2016 and 2020 given the contention around the election and subsequent loss of Donald Trump. Higher votes cast in those years may indicate higher voter participation that could translate into the 2024 election. This is also relevant given Virginia's role as a swing state. KDEs comparing Total Votes Per Year and Total Candidate Votes Per potentially give insight into the role of tenuous elections in voter turnout. 2000 and 2020 both had the highest levels of total votes cast and the highest number of votes cast for individual candidates. Given that these two elections: Bush v Gore and Biden v Trump were highly contentious, this analysis can help inform our predictions and modeling. Important elections seem to increase overall turnout. We can use predictive modeling to look at what population subsets - geographic location, race, education, income, etc. - have higher turnout rates thereby influencing election results.



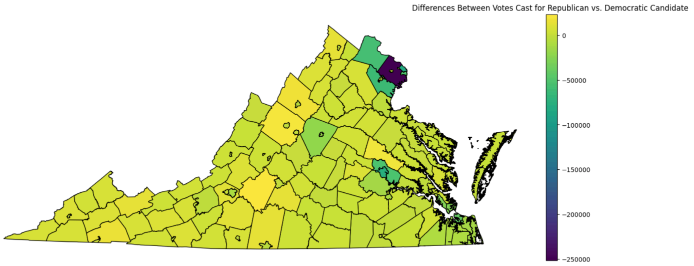
Graph 8: Total Votes Per Year



Graph 9: Candidate Votes Per Year

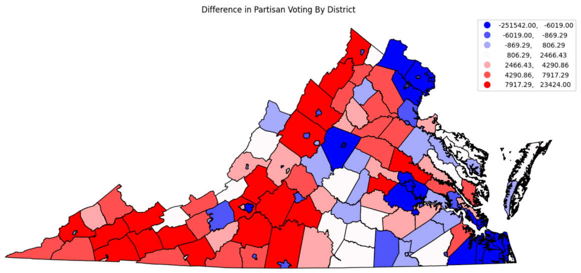
Correlation and Covariance Matrices were also computed for Total Votes Per Year and Candidate Votes for Year. The correlation between candidate votes and year was .05 while that of total votes and year was .07. These are extremely minimal positive correlations which can’t lead to the conclusion that voter turnout has increased consistently over the past 20 years. However, the slight difference between total votes and candidate votes may indicate that the concentration votes given to candidates has not increased proportionally to total votes cast. However, more extensive analysis would be needed to determine if there is correlation between the concentration of votes cast for certain candidates based on the election year (ie: do elections perceived as more contentious result in a higher concentration of votes being cast for the two primary candidates instead of for third party candidates).

The importance of data transformation is also demonstrated when analyzing differences in voter concentrations by district. As noted in Graphs 4 and 5, the distribution of voters tends to cluster around cities, with increasing Republican concentrations as you head South. Graph 10 shows the difference in votes cast for the Republican vs Democratic candidate in 2020 (Trump v. Biden).

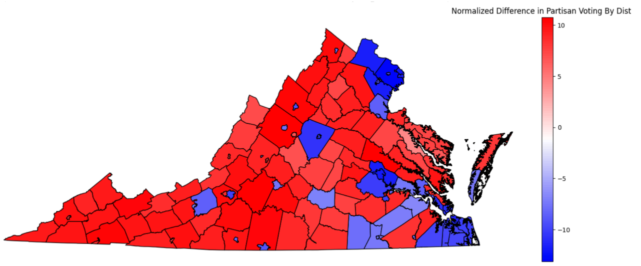

Differences Between Votes Cast for Republican vs. Democratic Candidate
0
−50000
−100000
−150000
−200000
−250000

Graph 10: 2020 Differences in Votes Cast for Republican vs. Democratic Candidates

This visualization demonstrates that while much of Virginia has minimal difference between partisan votes cast, cities again remain more polarized and thus more consequential in elections. Further investigation could look into how this difference between party votes cast has changed during Northern Virginia recent expansion in the past few decades. Given that most voters and most partisan differences are concentrated in specific areas of the state, normalizing the data helps us deal with this long tail. Graphs 11 and 12 show the normalization of partisan differences via the inverse sine function.



Graph 11: Partisan Differences by District



Graph 12: Normalized Partisan Differences By District

This process gives a more complete picture of what districts are more homogeneously “Republican” or “Democrat”. Further analysis could look into changes in voter behavior as the population has increased, or demographic changes related to income, education, and racial make-up have changed. Another avenue for analysis could look at how individual districts have “flipped” over the years, and measure partisan reliability by district. This would be relevant given Virginia’s status as a swing state, and importance in presidential elections.

Results

Our experiments aimed to explore the impact of various demographics on voting margins in Virginia’s presidential elections. The dataset for this study was compiled from county-level voting records for presidential elections as well as demographic data sourced from publicly available census data. We merged the voting data with the demographic data, using countyCode as the key for merging. Then, we developed a linear regression model to predict the margins of victory in Virginia counties; our model included educational attainment (particularly those beyond high school, such as bachelor’s and master’s degrees), racial composition (e.g. percentages of white and black populations), and income levels. The target variable was the voting margin, calculated as the difference in votes between the Democratic and Republican candidates. A positive margin indicates that the Democrat candidate won the election, and a negative margin indicates that the Republican candidate won the election for the state of Virginia.

Using scikit-learn, we trained several linear regression models using one or more of the following predictors: racial demographics (‘white’, ‘black’), higher education (‘high_school_diploma’, ‘associate_degree’, ‘bachelor_degree’, ‘master_degree’, ‘doctorate_degree’), ‘income’, and ‘countyName’. Using ‘countyName’, or just the counties, in our models may seem redundant, but we chose to use it in some just to see how it would affect voter margin. We used an 80-20 train-test split with a random seed to ensure replicability. The model's performance was assessed using the R-squared value. Each model predicted a voter margin for every observation in the test set, so we took the average of that margin to get the predicted voter margin for each of our models. The models that yielded the most interesting results and had the highest R-squared values predicted voter margin on year, race, and ethnicity, and on highest education level attained, year, and income. The former model achieved an R-squared value of 0.832, indicating that about 83% of the variability in voting margin could be explained by the model's predictors. This model predicted the Democratic candidate winning by a margin of 185,091 votes.The second model achieved an R-squared of 0.831, predicting the Democratic candidate winning by a margin of 185,265 votes. These high R-squared values suggest a strong relationship between the demographic predictors and the voting margins.

Conclusion

If we were to take this project further, we would first address improving the model's fit. All of our models had an extremely large MSE value, indicating that the model is not a good fit for the data. The high MSE values could be a result of overfitting, as several of our variables really relate to one variable. For example, 'white', 'black', 'asian', and other racial columns all relate to the race variable. To address this issue, we could figure out a better way to reshape our data, but as explained in the Data Cleaning portion, melting the dataframe in a way that made sense and retained the integrity of the data given the constraints we were working under was extremely difficult and nearly impossible. After doing this, we could use cross-validation techniques to eliminate variables that provide redundant information. We would also consider other ways to handle missing values in the county demographic data for each election year, instead of replacing all missing values with 0. Finally, we could explore interactions between variables when training and fitting our model.

Another step we could take is adding in more data. First, we'd add in the demographic estimate data for all counties in the US and more years and election cycles, such as the last 10 election cycles instead of the last two. Additionally, we could find other datasets that might capture other possible influences on these elections, such as political factors (campaign spending, events, etc.), socioeconomic data (more detailed income data, employment rates, etc.), or urban vs. rural distribution.

Though our model risks overfitting as seen with our high MSE, our model makes a valid attempt at predicting factors that influence election outcomes. It accounts for key demographic indicators such as race, ethnicity, income, and highest education level attained. These variables are critical for predicting how counties will vote in the 2024 election, as existing trends between voting patterns and demographic variables are reinforced and new voting patterns emerge.