Correlation between the Presence of Big Box Retailers and a County's Political Leanings

Pouya Mohammadi

Introduction

In recent election cycles, election analysts such as Dave Wasserman have pointed to an interesting metric and its being highly correlated with the results of United States (US) presidential elections. This metric is whether the county contains a Whole Foods or a Cracker Barrel (Wasserman, 2020). This has lead to a flurry of analyses about how the presence of different big box retailers is correlated to that area's political leanings. NBC News, Time Magazine, and The New York Times have all recently reported on similar trends with varying reatilers, and on Twitter after the 2020 election, there was speculation that the presence of a Trader Joe's in a county was becoming a valid indicator for the direction in which a county would vote in the election ("Broke: Joe", 2021).

These trends are important as computational politics becomes of ever-increasing importance in modern political campaigns. In elections that are decided by the slimmest of margins, such as the US presidential election of 2000 that was decided by less than 600 votes, every informational advantage that a campaign has can be of use and drastically influence the politics and future of the United States and the world (Glass, 2018). By understanding the relationships between the presence or absence of certain big box retailers in a county and the political preferences of that county, campaigns may be able to draw insights into the ways that communities' shopping habits correspond with their politics and more effectively and efficiently communicate with potential voters. In addition to providing insight into the politics of communities, this method of using the presence of big box retailers in a county to infer information about politics is computationally efficient and far less expensive than big data methods that rely on storing and analyzing information on an individual basis.

Our case study aims to analyze the presence of a new set of big box retailers in US counties and understand the ways that the presence of these counties corresponds with that county's political leanings. In particular, the stores that we will be focusing on are Trader Joe's, Cinemark, Cabela's, and Nordstrom. These stores were chosen in part because of the fact they were easily accessible data sources, but mostly because of our belief that they will capture information about the different voting blocs in the US electorate incredibly well. Trader Joe's targets singles, couples, and small families living in larger cities or the areas surrounding large cities (Watson, 2014). Cabela's tends to target individuals who tend to be white, conservative males. These include hunters, fishermen, military personnel, and law enforcement (Martin, 2013). Nordstrom's target market is high-end shoppers and the middle class (Bhogaraju, 2015). We suspect that this corresponds with suburban voters who are an important voting bloc. Finally, Cinemark's target market is "midsized markets or suburbs of major cities" (Team, 2020). We believe that the midsized market corresponds to an important voting bloc that is not encapsulated in the target market of the other retailers.

Aside from Trader Joe's, all of these stores have yet to be the subject of a rigorous analysis that relates their presence to political leanings. Trader Joe's was included because we believe that it corresponds to a different voting bloc than the other stores that we chose and the fact that the body of work on the political leanings of "Trader Joe's counties" is new and does not include a methodology similar to ours. Our null hypothesis is that we do not expect there to be a relation between the presence of any of the stores and the political leanings of the counties in which these retailers are located.

Data

We use publicly available data online that can be downloaded for the list of sites in the Data subsection of our References. These include open-source government databases as well as public location listings by private companies. The 2020 presidential election results that we use to measure a county's political leanings are borrowed from the GitHub user @tonmcg, who developed the datasets from information provided by The Guardian, townhall.com, Fox News, Politico, and the New York Times (Tonmcg, 2020). All of the government provided databases are estimations of the current values based on the most US Census which occurred in 2010 and trends in the data since that time. Finally, the location listings are either provided by the stores themselves or reputable third-party sources, such as Fandango. We use the 'geopy' Python library to map these locations to their corresponding geographical county. Any locations whose counties cannot be found by the Python library are hand-encoded. Because Alaska uses electoral districts instead of counties in their voting procedures, we will not consider Alaska in this analysis and remove all data from Alaska in our dataset prior to analysis. While this solution is not ideal, it is consistent with previous literature and prudent since we do not feel comfortable using two different region distinctions in the calculation of our dependent variable (Gomez et al., 2007).

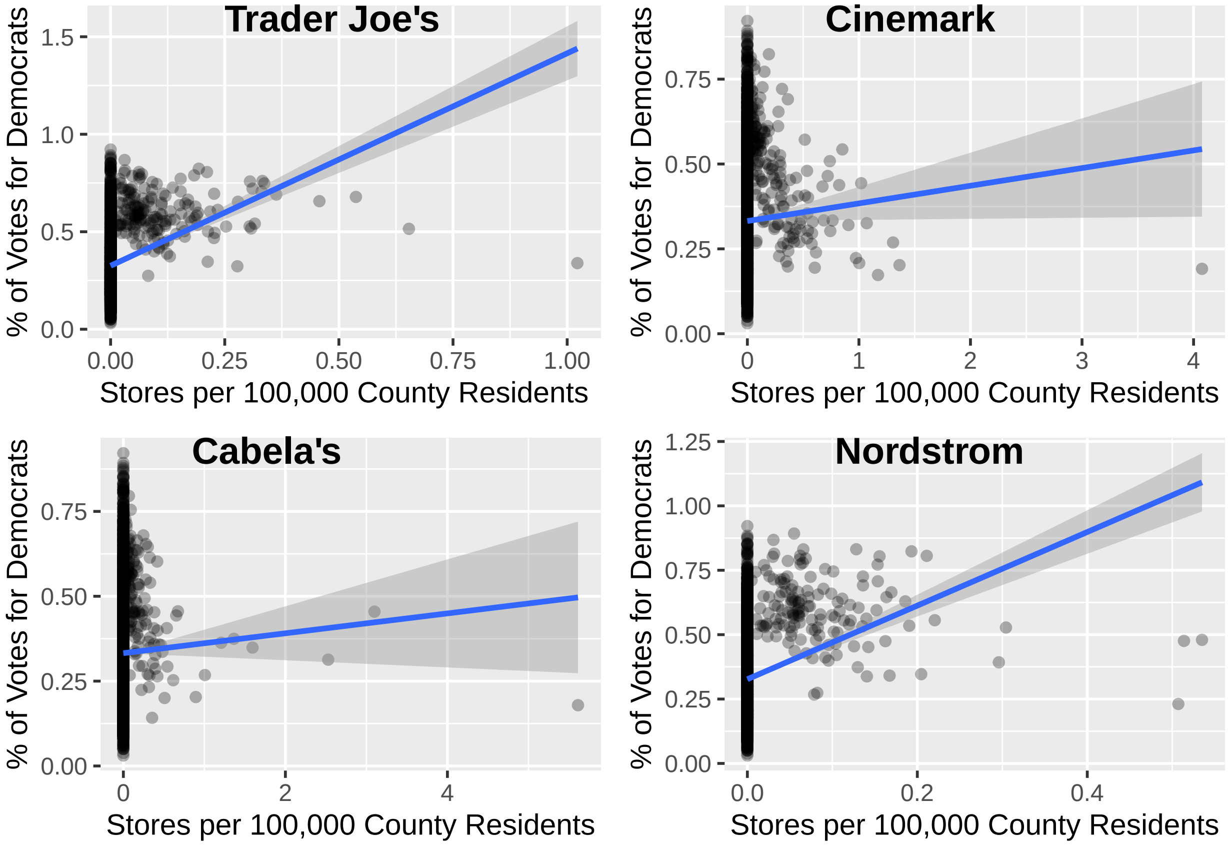
We will use the percentage of votes for the Democratic party in the 2020 presidential election results from a county as our dependent variable, which will be a numeric variable between 0 and 1, to measure the political leanings of a county. For our independent variables, we will have four numeric variables that each correspond to the number of stores per 100,000 residents in a county. Each variable will correspond to one of the 4 stores that we are considering in this analysis. In addition to these variables, we will include indpendent variables that are usually included in analyses of a county's political leanings. These variables are borrowed from Kahane, and they include population totals by race, population totals by gender, education levels, poverty rates, the median income of the county, unemployment rates, and whether the county is urban or rural (Kahane, 2020). We exclude certain variables that we did not have access to, such as religion in our analysis.

Existing Literature

As mentioned previously, plenty of news networks have completed analyses about the ways that big box retailers' locations correspond to political leanings in a county. In particular, NBC News, Time Magazine, and The New York Times have conduced analyses on this phenomenon. The New York Times in their analysis, found that, without controlling for other factors, the presence of a Trader Joe's corresponded with better results for Democrats. The analysis discovered that Democrats won areas that were within five miles of a Trader Joe's by 33 points. Time magazine discovered a similar trend in that Democrats won districts with a Trader Joe's by 30 points in 2014 US congressional races. A different study by Aaron Lee that employs a random forest machine learning model that uses 20 big box retailers discovered that counties that contain a Trader Joe's lean democratic with a feature importance of about 0.11 (Lee, 2020).

Despite all of the attention to these trends, there does not seem to a published work on the topic. We expand on these past analyses by analyzing a new set of big box retailers and their presence's relation to political leanings in a county. We also control for other variables that are customarily used in predicting a county's political leaning, which has yet to be done in a scientific analysis. We gather this list of variables from Kahane, and we explore whether the presence of these big box retailers will provide more significant relationships to a county's political leaning than the standard covariates.

EDA



These graphs show that regardless of the retailer, it appears as if Democrats performed better in counties that contained the retailer more often Republicans. Although, this trend appears to be greater for Trader Joe's and Nordstrom than Cabela's and Cinemark.

Methodology

Motivation

In order to analyze whether the presence of big box retailers corresponds with a county's political leanings, we will utilize a beta regression model using the percentage of votes in the county cast for Democrats during the 2020 president election as our dependent variable. We chose this dependent variable since the percentage of votes that a party receives is the best measure of a county's political leanings in the United States, given that the United States is a largely two-party political system. By using the percentage of votes for Democrats, we can capture information that we would not be able to if we simply used a binary indicator for which party won the county. For instance, a county that Democrats won by 0.2 percentage points would be regarded equally to a county that Democrats won by 50 percentage points, when in reality, one county is a swing district and the other is a staunchly Democratic county. We use the percentage of Democratic votes instead of Republican votes since we believed it would increase interpretability given that our EDA indicated that most counties with one of our four stores will mostly vote for Democrats. However, in our sensitivity analyses in the appendix, we repeat our analysis using the percentage of Republican votes in a county and find that the results are largely the same because of the two party system in the United States.

Our analysis employs a beta regression model. However, other models were considered as well, such as a linear regression model, an OLS model, a poisson model with an offset, and a negative binomial model with an offset. We do not use a linear regression model since our response variable is bounded between 0 and 1, which a linear regression model does not account for. We do not use the OLS model because our dependent variable has values that are close to 0 and as a result, the OLS model could run into the same issue that the linear regression model faces where it does not bound the dependent variable distribution between 0 and 1. Finally, the percentage of votes in a county is a count of the votes for a party divided by the count of total votes in the county, which can be modeled via a model of count data with an offset. In these models, our dependent variable would be the count of votes for a party and the offset would be the total votes in that county. However, we see that a Poisson model is not viable since the mean count of votes for both parties is not equal to the variance of the count of votes for that party, an assumption of the poisson model. We then consider a negative binomial model which does not require that the mean of the dependent count variable is equal to its variance. However, when performing some EDA, which can be found in the Appendix, we found that the count of votes for either party does not fit a negative binomial distribution despite the fact that it is count data with overdispersion. Instead, the beta regression model assumes that the response variable follows a beta distribution, a distribution of values between 0 and 1. Our response variable does the same and has a single mode, which is similar to a beta distribution again. This is further discussed and visualized in Appendix. In general, the beta distribution is incredibly applicable for proportion data, such as the percentage of votes for a party in a county (Ferrari et al., 2004). Since the only other assumption for a beta regression model is that there is a linear relationship between the predictors and the response variable, we feel confident employing a beta regression model, since the percentage of democratic votes in a county falls between 0 and 1, similar to a beta distribution.

We check the assumptions of the beta regression model in the appendix. The beta regression model has the following three assumptions:

1. The response variable, the percentage of votes that Democrats received in a county, follows a beta distribution.
2. There is independence between the observations in our dataset.
3. There is linearity between the predictors and response variable in our model.

Model Formula

$$\begin{aligned} logit(\mu_i) = &\beta_0 + \beta_1 * (\#.of.Trader.Joes.per.100k.residents_i) + \beta_2 * (\#.of.Cabelas.per.100k.residents_i) + \\ &\beta_3 * (\#.of.Cinemark.per.100k.residents_i) + \beta_4 * (\#.of.Nordstrom.per.100k.residents_i) + \\ &\beta_5 * (Percent.White_i) + \beta_6 * (Percent.Black_{.}Population_i) + \beta_7 * (Percent.Hispanic_i) + \\ &\beta_8 * (Percent.Asian_{)} + \beta_9 * (Percent.Male_i) + \beta_{11} * (Percent.with.College.Education_i) + \\ &\beta_{12} * (Percent.in.Poverty_i) + \beta_{13} * (Percent.Unemployed_i) + \beta_{14} * (Urban.County_i = True) \end{aligned}$$

where μ_i is an observation-specific mean for a Beta distribution with mean equal to $Percent.Democratic_i$ and a precision parameter that is modeled as $\phi_i = log(\phi_i)$.

Sensitivity Analyses

We will perform a few sensitivity analyses in which we compare the performance of different models. The first model that we will test is replacing the response variable in our original model with the percentage of republican votes in a county in order to see if there is no significant difference between these two models as we hypothesized by assuming that the two-party system would make third-party votes negligible. We will conduct another sensitivity analysis in which we use a ANOVA test with our original model and a model that does not contain any store data in order to see if there is a significant difference in a traditional model of county affiliation and one that includes the store locations.

Results

	Estimates	Exp(Estimates)	Standard Errors	P-Values
Intercept	-0.953379387277584	0.385436281615812	0.263615831502523	**<0.001**
Trader Joe's per 100k	1.28495272067356	3.61449706219106	0.22270100684162	**<0.001**

	Estimates	Exp(Estimates)	Standard Errors	P-Values
Cabela’s per 100k	0.00157411855201087	1.00157535812695	0.0589165457066467	0.979
Cinemarks per 100k	-0.0898253486069021	0.914090818565614	0.0785621750849679	0.253
Nordstroms per 100k	0.772364600918863	2.16487928180614	0.328527495043392	**0.019**
% White	-1.45316255739235	0.233829617918465	0.12889761412432	**<0.001**
% Black	1.01270577798572	2.7530400614664	0.126321353765895	**<0.001**
% Hispanic	-0.522344524698854	0.593128312564386	0.136371429147572	**<0.001*****
% Asian	4.25335974976539	70.3413451049545	0.417559698157183	**<0.001**
% Male	-1.42561630182102	0.240360283071031	0.405879743018778	**<0.001**
% Educated	0.0277349535344954	1.02812314789915	0.0011056849976536	**<0.001**
% Poverty	-0.00137999900383997	0.998620952756925	0.00241282766993111	0.567
% Unemployed	0.110260504914761	1.11656890426264	0.00724681635963366	**<0.001**
Urban	0.131825805956499	1.14090956230667	0.0199296995383604	**<0.001**
Precision Link	20.6591	937849457		

Discussion

Model Interpretation and Implications

From the above model, we conclude that there is a significant, positive relationship between the number of Trader Joe’s stores per 100,000 residents in a county and that county’s percentage of votes for Democrats. We also see a significant, positive correlation between the number of Nordstrom stores per 100,000 residents in a county and that county’s political preference. However, we do not see significant relationships between a county’s political preference and the number of Cinemarks per 100,000 capita or Cabela’s per 100,000 capita in a county. More specifically, we see that for a 1 store increase in the number of Trader Joe’s stores per 100,000 residents, we see an increase of 1.285 in the log odds for the percentage of the county that voted for Democrats. Then for a 1 store increase in the number of Nordstrom stores per 100,000 residents, we see an increase of 0.772 in the log odds for the percentage of the county that voted for Democrats. Both of these are significant relationships, with p-values of less than 0.001 and 0.019, respectively.

This has many implications for future research in this field. As mentioned, there has been very little publicly available, rigorous, academic research on this topic - examining a correlation between the presence of certain stores in a county and that county’s political preferences. However, the findings of this study indicate that there may be promise in examining this line of research further. In the same way that the governmental agency FEMA uses the “Waffle House Index” to determine the severity of natural disasters, this line of research could do the same for politics and political campaigns. Since these private companies are spending so many resources to determine target demographics and implant stores in areas where those target demographics shop, it would be easier for political campaigns to use the locations of these stores than spend large amounts of resources in order to draw inferences about an electorate that shops similarly to how they vote. If there are certain underlying characteristics in individuals that result in those individuals both shopping at certain stores and possessing a certain political outlook, it would be prudent of political campaigns to simply use the stores as a measure of political affiliation instead of researching the underlying characteristics. This may especially be the case as politics in the United States shift to become increasingly focused on issues around culture.

Strengths and Weaknesses

One of the strengths of this analysis is that it really is one of the first in the field to apply a more rigorous, advanced methodology to analyzing the relationship between the presence of stores in a county and that county’s political influence. Since there has not been any publicly available research published on this topic previously, one of the greatest strengths of this paper is that it is a trailblazer in its field and is investigating a topic of public interest, as evident from the extensive press coverage, in a more scientific manner. In addition, this paper's accessibility is another one of its strengths. Often times, analyses that rely on data from the private sector or location based data are restricted through a paywall or require special permissions to access. However, all of the data that we are using is publicly available. In addition, the methods that we are using are easily interpretable as opposed to some previous analyses that have used methods such as a random forest model, which is a black-box algorithm.

While this analysis has many strengths, it also has a few weaknesses. Despite the fact that the analysis is incredibly accessible and only uses publicly available data, the data collection process for adding a new store is quite cumbersome. If future researchers want to add additional stores to this analysis, as we suggest below as future directions for this research, there is a great deal of web scraping and data wrangling that is required to add these additional stores that may be quite work-intensive and time-consuming. However, one of the most glaring weaknesses in this analysis is the fact that this analysis's larger framework assumes that there are only 2 political preferences in the United States and that if individuals in a county are not voting for Democrats, that they are voting for Republicans. While no inferences are made throughout the paper that explicitly state that individuals who do not vote for Democrats must vote for Republicans, the model framework that we design and our choice of a dependent variable is based on the fact that there is a 2-party system in the United States. Because of this, we employ our dependent variable as the percentage of votes in a county for Democrats and perform a sensitivity analysis to ensure that similar results would ensure from changing our dependent variable to the percentage of votes for Republicans in a county. However, we do not account for 3rd party voters in either our model framework or our sensitivity analysis, but advise future research to explore this in future works.

Future Directions

Given that this study is the first rigorous, academic study of its kind, there are many future directions in which future research could explore this issue. However, there are 3 future directions that we think are the most promising to explore and that we would advise future researchers to explore.

1. As mentioned in the weaknesses, the model framework that we design and our choice of a dependent variable is based on the fact that there is a 2-party system in the United States. It does not account for 3rd party voters, and we would advise future researchers to explore this problem using a framework that accounts for 3rd party voters.
2. We believe that there may be some underlying spatial variables that may not be accounted for in our current model. Incorporating spatial characteristics would improve the robustness of this analysis and perhaps also provide better results. Depending on the sort of spatial model and the potential confounding variables, utilizing a spatial model could help account for confounding variables that were not considered in our initial analysis.
3. Finally, we would advise future researchers to explore more stores than are present in our analysis. While we chose these stores due to our belief that these 4 stores are representative of different voting blocs in the United States that are representative of most of the major voting blocs, this may not be true, and we would encourage other researchers to explore other stores that they believe may represent different groups of voters in order to help advance this area of research.

Appendix

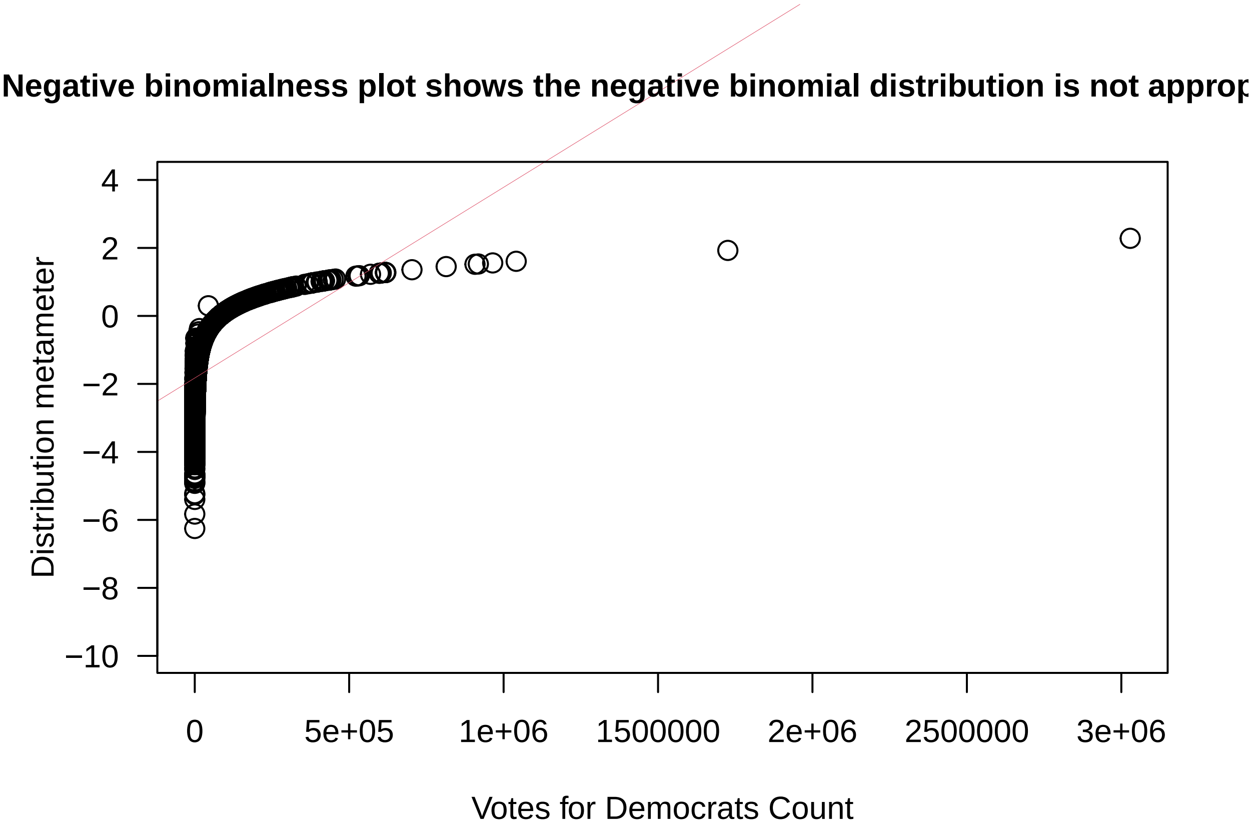
Model Motivation

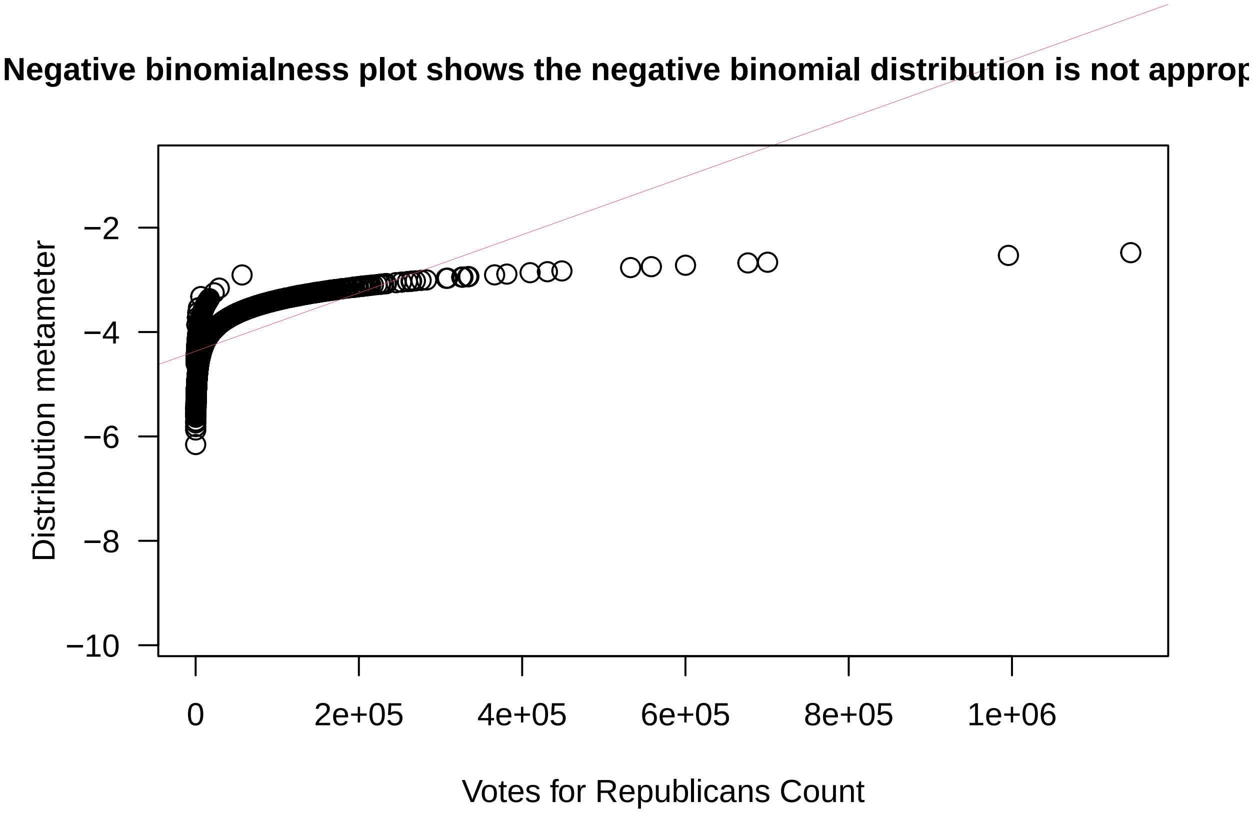
We see from the plots below that the mean and variance of the votes for each party are not equivalent indicating that while the data is count data, it does not follow the poisson distribution and a poisson regression cannot be used here.

	Democratic Vote Count	Republican Vote Count
mean	26064.1	23784.8

	Democratic Vote Count	Republican Vote Count
variance	9509979014.6	2953216930.8
ratio	364868.4	124164.0

From the plots below, we see that neither the votes for Democrats in a county nor the votes for Republicans in a county follows a negative binomial distribution. If a negative binomial model was appropriate, there would be a linear relationship between the distribution metameter and the count of votes for a respective party. However, we see that there is a logarithmic relationship between these variables in our plots below.





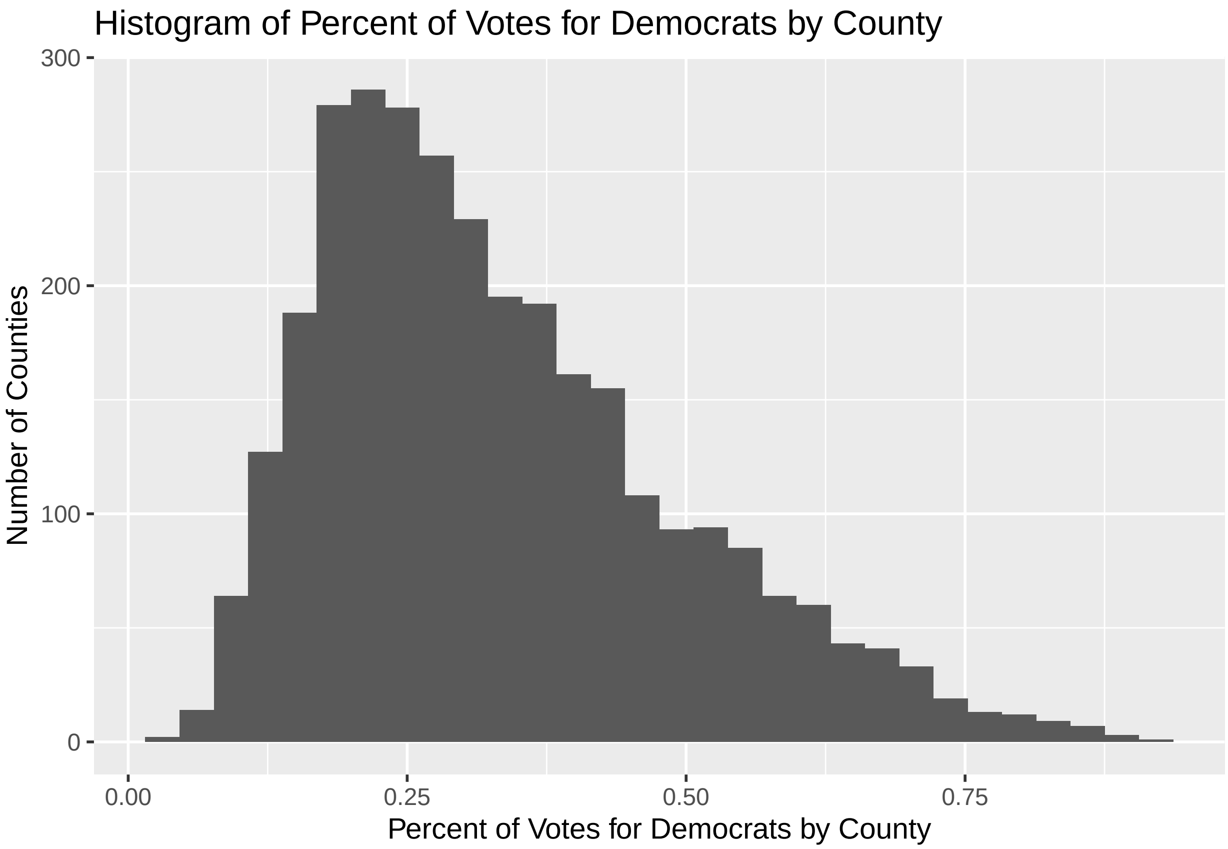
Negative binomialness plot shows the negative binomial distribution is not approp

Model Assumptions and Diagnostics

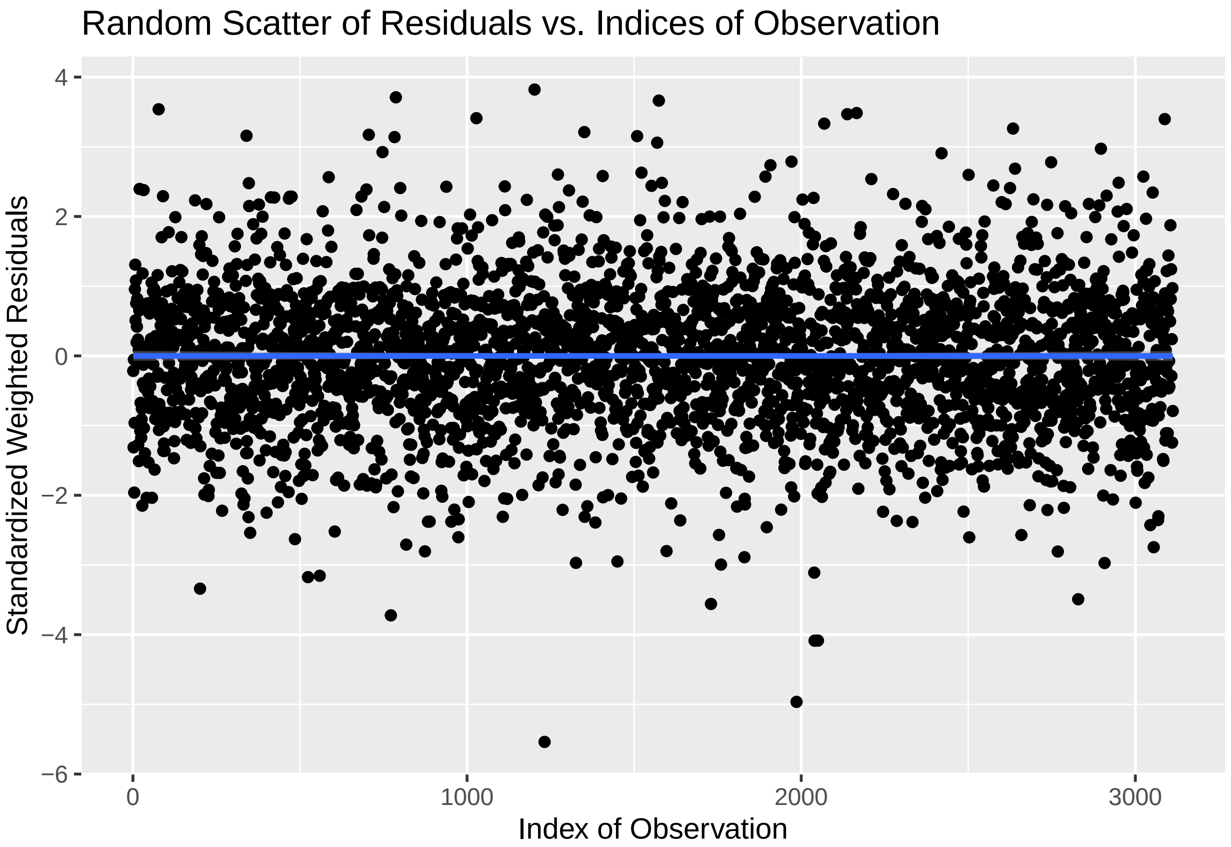
We will check the three assumptions for the beta regression model below. The three assumptions are as follows:

1. The response variable, the percentage of votes that Democrats received in a county, follows a beta distribution.
2. There is independence between the observations in our dataset.
3. There is linearity between the predictors and response variable in our model.

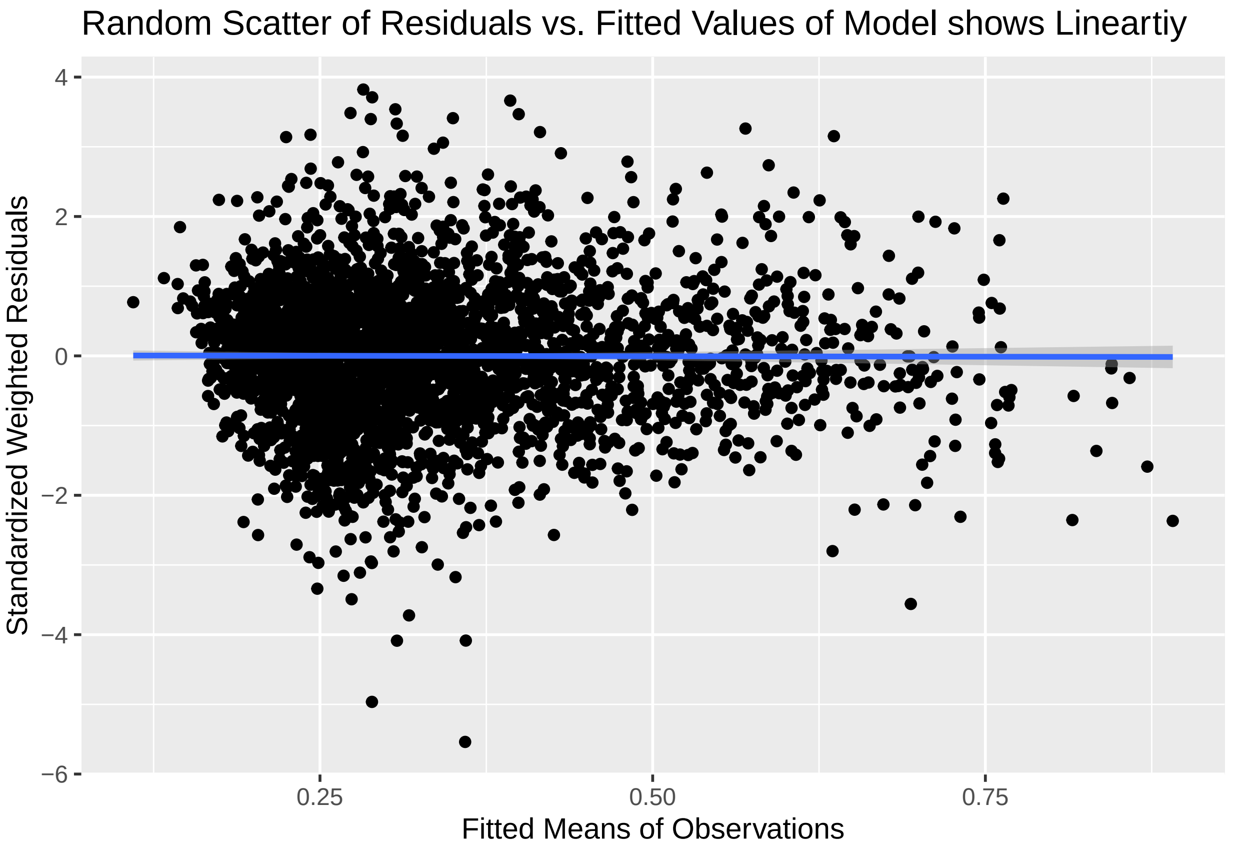
Below, we see that both the percentage of votes that Democrats received in a county have a range between 0 and 1, with a single mode. Since these characteristics are the same as the characteristics of a beta distribution, we feel confident modeling the percentage of votes in a county for Democrats as a beta distribution.



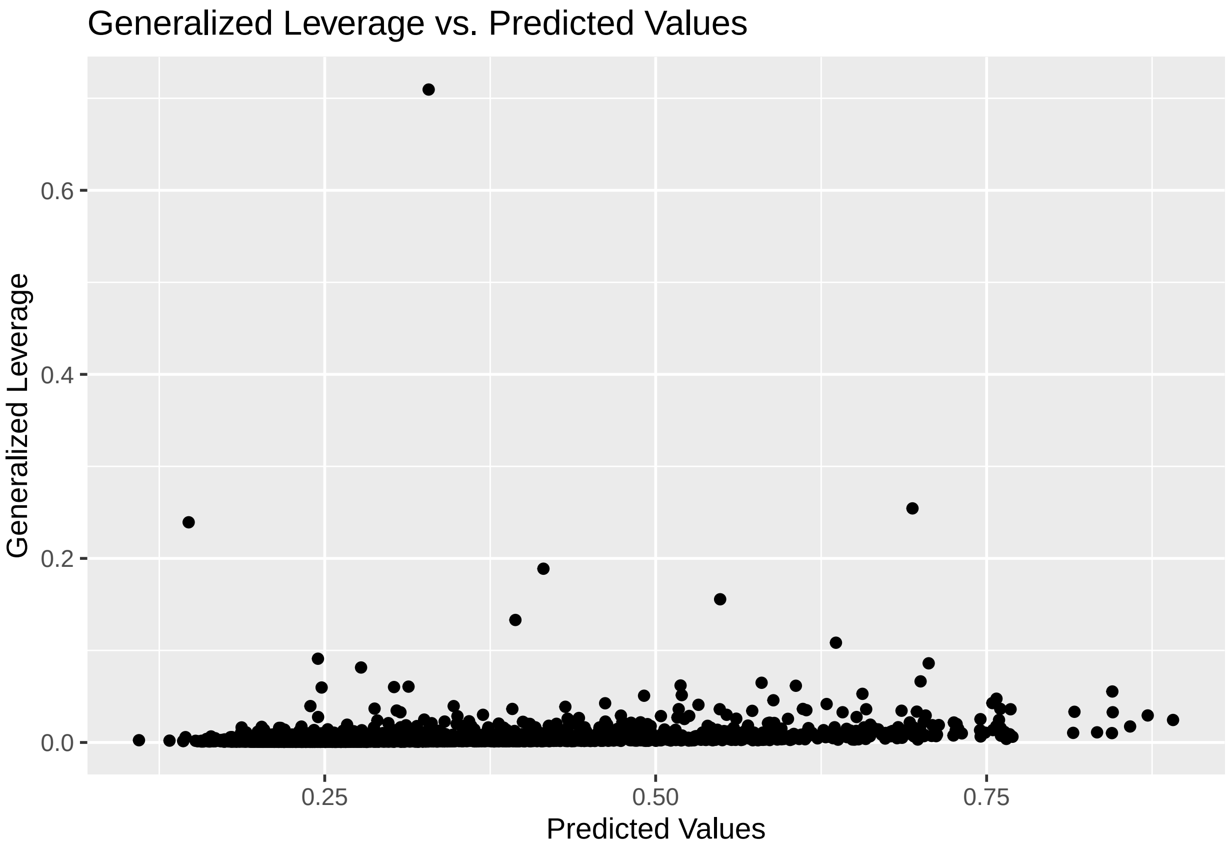
Next, we analyze the assumption that our observations are independent of each other. While there are some concerns here regarding individuals in different counties of a similar makeup being exposed to the same media and neighboring counties being in similar communities that affect the politics of the county, we feel confident that these effects are negligible and that we include the proper covariates in our model to account for any confounding variables. Plotting the residuals below, we see that there is a random scatter of residuals and this helps us feel more confident that there is independence between our observations.

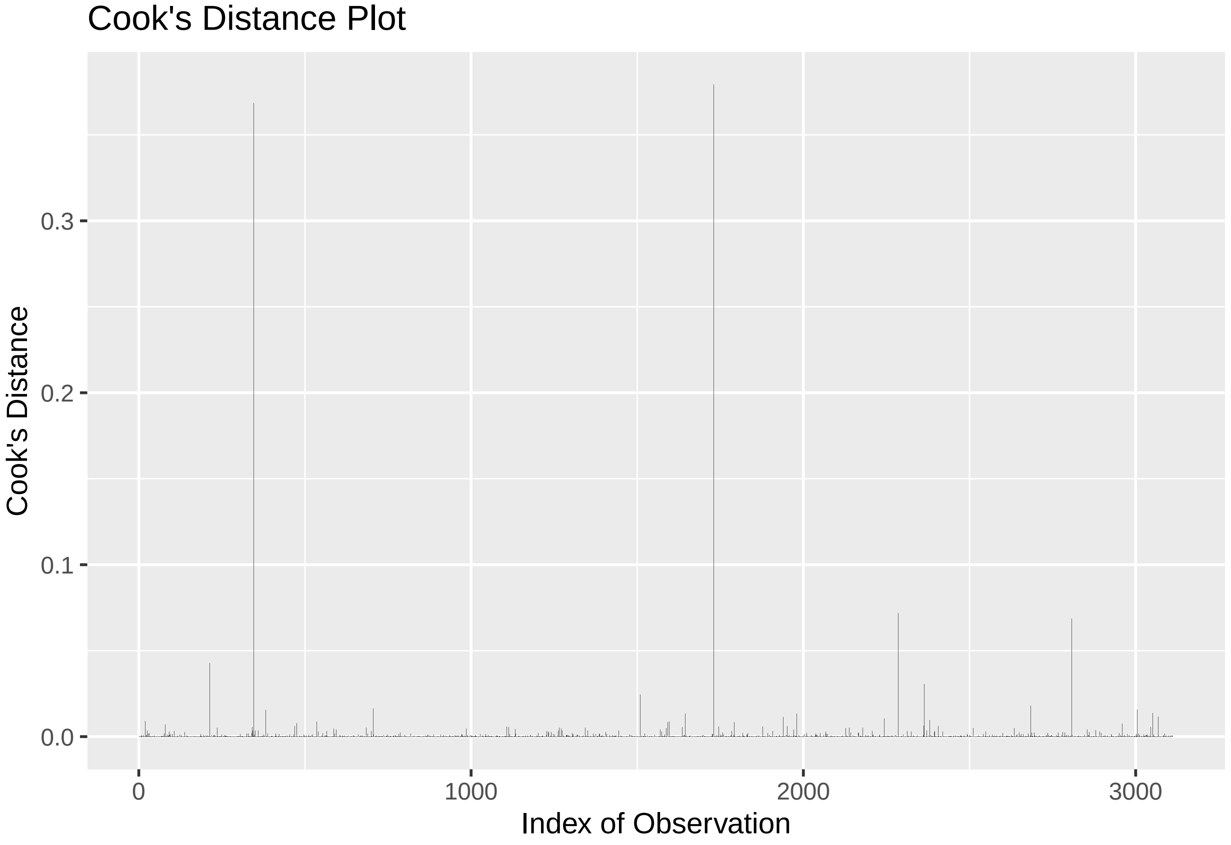


We will next investigate whether the linearity assumption is met for our model by plotting the fitted means for our observed data and the residuals of our data. Seeing that there is a random scatter around the y-axis, we feel confident that the linearity assumption is met.



Next, we will check for potentially influential points in our model. We plot both the leverage and the Cook's distance values of our model, and we see that we have a few observations that have a high leverage. However, looking at the Cook's distance for these values, we see that there are no points with a high Cook's distance. As a result, we feel confident that there are no influential points in our model.


Generalized Leverage vs. Predicted Values
Generalized Leverage
0.6
0.4
0.2
0.0
0.25
0.50
0.75
Predicted Values


Cook's Distance Plot
Cook's Distance
0.3
0.2
0.1
0.0
0
1000
2000
3000
Index of Observation

References

Bhogaraju, Sirisha. “Nordstrom’s Target Customers.” Market Realist. Market Realist, February 18, 2015. https://marketrealist.com/2015/02/nordstroms-target-customers/.

Ferrari, Silvia, and Francisco Cribari-Neto. “Beta Regression for Modelling Rates and Proportions.” Journal of Applied Statistics 31, no. 7 (2004): 799–815. https://doi.org/10.1080/0266476042000214501.

Glass, A. (2018, December 12). Bush declared Electoral victor Over Gore, Dec. 12, 2000. Retrieved March 03, 2021, from https://www.politico.com/story/2018/12/12/scotus-declares-bush-electoral-victor-dec-12-2000-1054202

Gomez, Brad T., Thomas G. Hansford, and George A. Krause. “The Republicans Should Pray for Rain: Weather, Turnout, and Voting in U.S. Presidential Elections.” The Journal of Politics 69, no. 3 (2007): 649–63. https://doi.org/10.1111/j.1468-2508.2007.00565.x.

Kahane, L. H. (2020). Determinants of County-Level voting patterns in the 2012 and 2016 presidential elections. Applied Economics, 52(33), 3574-3587. doi:10.1080/00036846.2020.1713985

Lee, A. (2020, September 29). Trader Joe’s Democrats and Walmart Republicans. Retrieved March 04, 2021, from https://towardsdatascience.com/are-you-a-trader-joes-democrat-or-a-walmart-republican-a7b156131435

M. (2021, February 16). Broke: Joe Biden did so well in counties with a Trader Joe’s because the audience for Trader Joe’s is composed of favorable Democratic demographics Woke: Joe Biden did so well in counties with a Trader Joe’s because voters thought he was Trader Joe. Retrieved March 03, 2021, from https://twitter.com/maxtmcc/status/1361504297477890050

Martin, Cindy. “Cabela’s.” CE Martin, November 10, 2013. http://cemartin.weebly.com/uploads/1/8/9/1/18911943/cabelas_marketing_assignment.docx#:~:text=Cabela’s%20target%20markets%20include%20avid,law%20enforcement%

Team, MBA Skool. “Cinemark SWOT Analysis: Top Cinemark Competitors, STP & USP: Detailed SWOT Analysis of Brands.” MBA Skool-Study.Learn.Share. MBA Skool, April 12, 2020. https://www.mbaskool.com/brandguide/media-and-entertainment/15016-cinemark.html.

Wasserman, D. (2020, December 08). Fact: Biden won the presidency Winning 85% of counties with a Whole foods and 32% of counties with a Cracker barrel - the widest gap ever. Retrieved March 03, 2021, from https://twitter.com/Redistrict/status/1336342894630858755

Watson, Elaine. “Quirky, Cult-like, Aspirational, but Affordable: The Rise and Rise of Trader Joe’s.” foodnavigator. William Reed Business Media Ltd., April 15, 2014. https://www.foodnavigator-usa.com/Article/2014/04/15/Quirky-cult-like-aspirational-affordable-The-rise-of-Trader-Joe-s#:~:text=Trader%20Joe’s%20targets%20singles%2C%20couples%2C%20and%20small%20families%20%E2%80%8B&text=less

Data

Trader-Joes-Stores.pdf. (2020). Retrieved March 03, 2021, from https://www.traderjoes.com/pdf/Trader-Joes-Stores.pdf

Nordstrom Store Addresses. (2020, June 1). Retrieved March 03, 2021, from http://nordstromsupplier.com/Content/sc_manual/Store_Address_List.pdf

Cinemark Movie Theater Locations. (2021). Retrieved March 04, 2021, from https://www.fandango.com/movie-theaters/cinemark

All Cabela’s Locations: Sporting goods & outdoor stores. (2021). Retrieved March 04, 2021, from https://stores.cabelas.com/

Small Area Income and Poverty Estimates (SAIPE). (n.d.). Retrieved March 04, 2021, from https://www.census.gov/data-tools/demo/saipe/#/?map_geoSelector=mhi_c&s_measures=mhi_snc&s_year=2019

Economic Research Service - Download data. (2021, February 24). Retrieved March 04, 2021, from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

Tonmcg. (2020). Us_county_level_election_results_08-20. Retrieved March 03, 2021, from https://github.com/tonmcg/US_County_Level_Election_Results_08-20