Arnav Dey

Is the degree of political polarization in opinions given by Twitter users a good predictor of voter turnout in the 2020 US elections?

Introduction

The purpose of this paper is to understand whether the degree of political polarization in tweets during the 2020 US elections significantly explains variation in voter turnout. We focus on Twitter as the sample for the populations political preferences because it has become a public forum for expression of political opinions in the modern social media age. This is an important result as it acts as a base from which we can design larger studies to understand whether increasing political polarization in US politics has reduced voter turnout. High voter turnout is an important indicator of a healthy democracy and a politically active citizenry which is why understanding causal influences on voter turnout, specifically polarization, is important to ensure civic society and organizations can take steps to mitigate voter demobilization. we will also be evaluating through visualizations whether our data is representative of the populations political preferences.

In this paper, We define political polarization as the emotional intensity of tweets measured on a scale from 1 to -1 using a pre-trained sentiment analysis model. Thus, the degree of political polarization would be those tweets whose sentiment scores extend towards 1 and -1. Voter turnout is simply the percentage of individuals in the voting population who ended up voting in the November 2020 elections. This task will be done using the dataset 'US Election 2020 Tweets' which can be found on Kaggle.

Arnav Dey

We will investigate the effect of polarization on voter turnout using a regression model with voter turnout as our response and sentiment scores as our predictor. Since a base assumption of linear regression is that the model accounts for all possible variables that influence the response, we will have to control for other exogenous variables so that we do not have a bias in our estimate for sentiment scores. To account for this, we will merge variables like income,age, education, marriage status among others that influence voter turnout using US census data and overlaying that at the county level. Then finally, we will use machine learning techniques and run a regression tree to evaluate the overall predictive power of our independent variables which we will compare with our regression model to see whether sentiments are influential or not.

Thus, when pursuing this research question we are making the following assumptions:

Assumptions

- the sentiment analysis of tweets is **representative** of the political preferences of the US population i.e that twitter users are a good sample to use when making inference on the US population
- twitter user tweeting from county is representative of the average resident of the county(i.e group level characteristics like age, income are average).
- Divergence in sentiments is a good measure for polarization.
- Follower count and likes count is a good way to measure the reach of sentiments expressed

result summary

to summarize our results, we found that our data seems to be representative but with some potential biases stemming from problems in our group assumption (twitter user representative of average county resident) due to outliers in merged data and because of a potential left-wing bias in the twitter data. Moreover, we find that sentiments weighted by follower count has a positive influence on voter turnout based on our regression model. However, our regression tree tells us that sentiments have no predictive power.

Maps summary

Our maps depict that there is no significant connection between age, income, and sentiment scores apart from some outliers and a potential anti-trump bias as mentioned before in the results summary.

significance of research question

Overall, this is an important research question because it seeks to study whether a more polarized political climate incentivizes or disincentivizes political participation. This result will have important implications on the health of democracy in this age of increasing polarization.

Literature review

In the book *Red and Blue Nation? by Pietro S Nivola and David W Brady*[1], the authors find that increased negative and polarized political advertising rather than reduce voter turnout, actually increased voter turnout. Moreover, they found that turnout decay coincided with times when the political elite were uncommonly close ideologically. This is interesting as it gives us a hypothesis to work with, namely that a *higher degree of polarization increases voter turnout*. This finding comes from US election data from the 1990s.

[1] Nivola, Pietro S., and David W. Brady. *Red and Blue Nation?: Characteristics and Causes of America's Polarized Politics*. Brookings Institution P, 2006.

In the paper *Quantifying polarization across political groups on key policy issues using sentiment analysis by Dennies Bor,Benjamin Seiyon Lee*[2] the authors found that differences sentiment scores across certain topics is a good proxy for understanding differences in political positions. Moreover, this paper specifically uses the Valence Aware Dictionary for Sentiment Reasoning (VADER) package for calculating sentiment scores on twitter data which is what we will also be using. We can safely assume from this paper that sentiment score divergence is a good way to measure political polarization on twitter.

In the paper *Division Does Not Imply Predictability: Demographics Continue to Reveal Little About Voting and Partisanship by Seo-Young Silvia Kim*[3], the authors state that demographic differences/social sorting has a modest impact on predicting political behaviour and partisanship. What this means is that although the effect is not significant, there is some effect which is what we will look into. This justifies the inclusion of demographic variables in our regression model as they affect political polarization (sentiment scores).

This paper will identify important demographic variables that influence voter turnout and provide empirical results to test the hypothesis that polarization increases voter turnout. It will also give us a base understanding of whether twitter sentiments are a good sample for the populations political sentiments.

In the paper 'The Bush Effect: Polarization, Turnout, and Activism in the 2004 Presidential Election' by Alan Abrahamowitz[4], evidence shows that intense polarization among the electorate

[2] Bor, Dennies, and Benjamin Lee. "Quantifying polarization across political groups on key policy issues using sentiment analysis." *arxiv*, vol. 2, no. 34, 2023, doi.org/10.48550/arxiv.2302.07775. Accessed 12 Apr. 2024.
[3] Kim, Seo-young S., and Jan Zilinsky. "Division Does Not Imply Predictability: Demographics Continue to Reveal Little About Voting and Partisanship." *Political Behavior*, vol. 46, no. 1, 2022, pp. 67-87.
[4] ABRAMOWITZ, ALAN I., and WALTER J. STONE. "The Bush Effect: Polarization, Turnout, and Activism in the 2004 Presidential Election." *Presidential Studies Quarterly*, vol. 36, no. 2, 2006, pp. 141-154.

in relation to George Bush actually increased activism and voter turnout. This supports the mobilization hypothesis we made earlier. Moreover, in the same way this study focused on George W Bush's election we focus on Trump as a focal point for polarization.

Data

The raw dataset ‘US Election 2020 Tweets' consists of tweets from October 15th 2020 to November 8th 2020, with around 1.72 million tweets about Presidential candidate Donald Trump along with certain demographic information like user location, city, state along with data like number of likes, retweets on a post and the number of followers of the posting account. The data was collected using the Twitter API, where the author used ***`statuses_lookup`*** and ***`snsscrape`*** for keywords.

The observations are at the individual level, with each observation representing a specific tweet by a specific twitter account. We will be merging other predictors from other datasets at the group level, in our case at teh county level. This is why we make the assumption that individual twitter users represent the average resident in the county which in our paper means that the twitter user’s demographic variables is a collection of the average of those variables at the county level.

dataset link:https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

Additionally, we use the Valence Aware Dictionary for Sentiment Reasoning (VADER) to calculate sentiment scores (-1 to 1) for all tweets which we put into the column ‘compound score’.

The sentiment scores represent the sentiments of the individual tweets in the data. However, since we want to treat the twitter data sample as representative of the populations political preferences, we must understand the engagement that each tweet gets which tells us the reach the sentiments expressed get. One way to factor this in is to use user follower count to do a weighted calculation of sentiments scores, wherein we multiply follower count of twitter handle with the sentiment score of the tweet. This will give us a more robust picture of the 'reach' of political sentiments expressed in these tweets. We will do the same using 'likes' as a means of weighting sentiments to produce an alternative measure of reach. In doing this we assume:

- the sentiments expressed in the tweet represent the broad sentiments of the followers of the twitter handle
- the number of likes the tweet receives indicates the number of people who resonate with the sentiments of the tweet

Thus, we wish to use both follower count and likes as a 'reach' measure coupled with our existing sentiment score distribution. When computing both, we will normalize follower count and likes count so that we are working within the same scale when multiplying them with sentiment scores.

Summary Statistics

	Likes	Follower count	Compound score	Sentiments weighted by follower	Sentiments weighted by likes
count	20560	20,560	20560	20560	20560
mean	8.395	22,285.12	0.00348	0.000263	0.000047
std	138.3	193,017	0.475	0.016	0.0084
min	0	0	-0.996	-0.312	-0.225
25%	0	220	-0.318	-0.000012	0
50%	0	1,165.5	0	0	0
75%	1	3,070	0.3612	0.000013	0
max	11670	6,488,193	0.9999	0.817	0.95

Interpretation of likes summary table:

based on the interquartile range, 50% data points for number of likes are between 0-2. This is in stark contrast to the mean which is much higher at 11.76. The most striking is that the standard deviation is 211.3 which is more than 10 times bigger than the mean itself. Such a massive standard deviation given our mean and interquartile range implies the presence of extreme outliers, which are severely skewing the data. Based on our max value, it is clear that there is a massive right skew in our likes distribution. this is interesting as it shows that a very small number of posts produce a disproportionately high number of likes from a likes share point of view. *This means that even if a few twitter posts had positive sentiments, the reach of those sentiments is much greater than the majority of posts that have negative sentiments*.

Interpretation of follower summary table

the interquartile range suggests that 50% of all values fall within 187 to 3,313 followers. The mean follower count is 31,588 with the standard deviation being 261,690. This implies that like the previous summary, there are outliers that lead to a large standard deviation and mean, despite most values being within 187 and 3,313. Like before, there is a big right skew implying that there are a few accounts in the dataset that have a disproportionately high share of followers. Again, this will likely have a similar impact on the reach of sentiments as it did with likes wherein a small number of tweets with negative or positive sentiments will have more reach than the majority sentiments expressed in the tweets. *The follower count table does seem relatively more skewed than likes. I say this because the maximum follower count is 291 times the mean follower count whereas the max likes is 138 times the mean likes.*

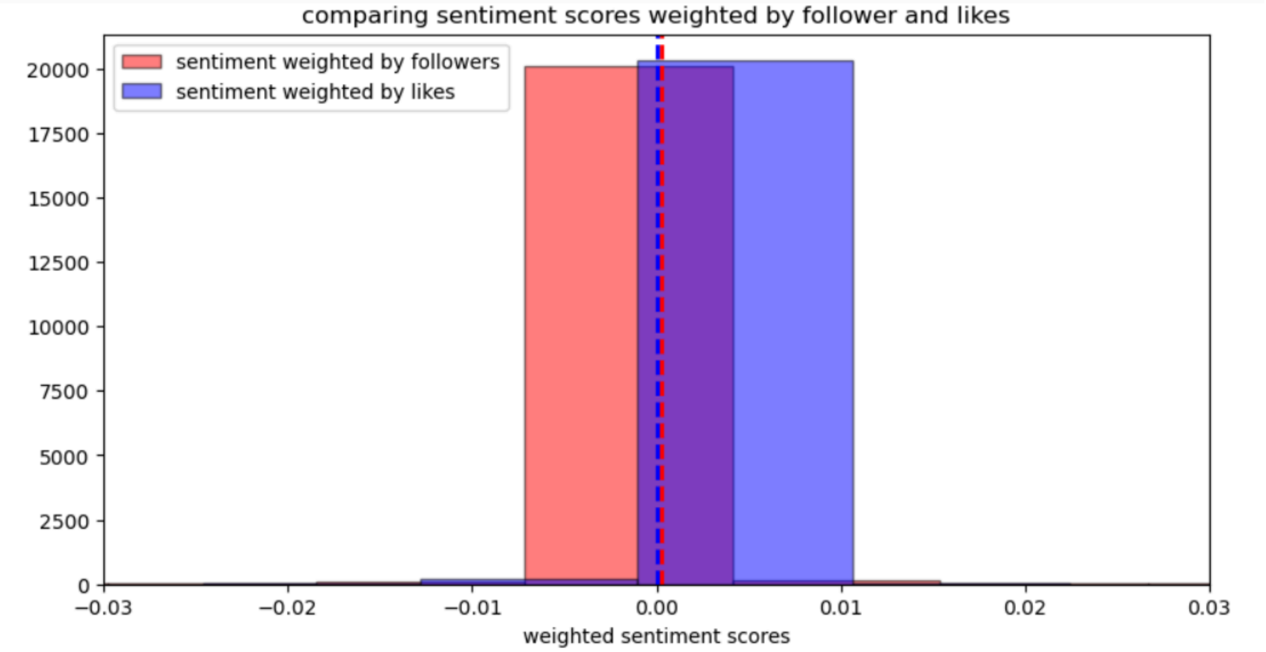
interpretation of sentiment score table

The sentiment data is fairly symmetric with Mean at 0.0035 and Median at 0 being fairly close to each other. this suggests a fairly even split between pro and anti trump tweets. We definitely have extreme views as well since -0.9 and +0.9 are the minimum and maximum respectively. Considering that the distribution is quite normal, the weighted sentiments will give a good picture of the reach of sentiments. This could potentially hint towards the dominant political position in the US.

interpretation of sentiments weighted by followers and likes table

the mean sentiment score weighted by followers is higher than that of likes. The spread of sentiment scores weighted by followers is also more than that of likes. Both of them being positive suggests that there must be a relatively greater reach for positive sentiments (pro-trump) tweets. The difference in means suggests that sentiments have a greater pro-trump reach if we were looking at follower count. However, the sentiments weighted by likes have a higher maximum than that of follower count. This suggests that despite follower count being more skewed, there seems to be a higher number of likes in the extreme position relative to follower count. To conclude on the differences between the two measures, a histogram will have to be computed.

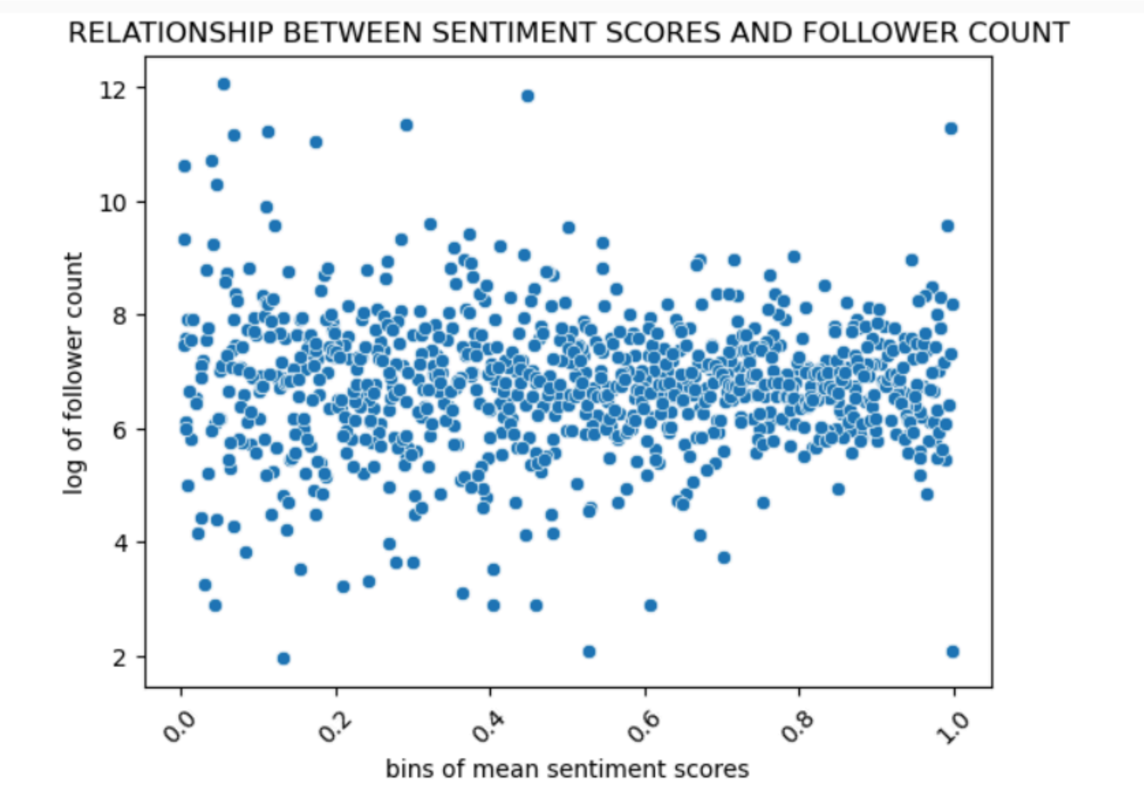
To conclude, *It is clear that we have to look at the small number of highly influential data points as opposed to the majority number of non-influential data points. We can also even out the skewed distributions using a log scale and make interpretations on that basis*.



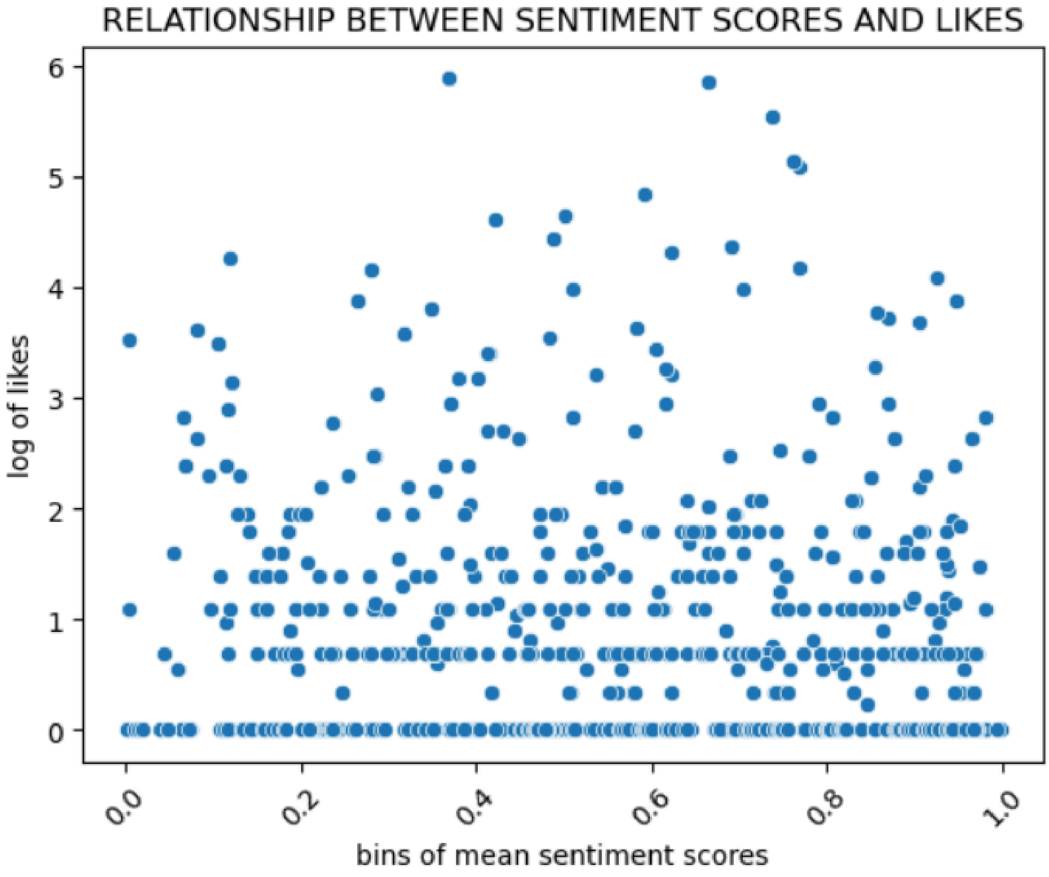
This plot better represents what our summary table has showed us. Sentiments weighted by likes show a better pro-trump reach that the tweets have whereas sentiments weighted by followers have a more anit-trump reach. The means are close to zero, showing that the distribution is still overall centred, with mean sentiment being neutral. this is an interesting results because it seems that more influential accounts from a followers point of view are posting pro-trump content but more posts with high like counts are posting anti-trump content. However, most of the points are not at the extremes, instead being within 0.03 and -0.03. This means that overall the polarization is not massive.

Another very interesting observation is that the the sentiments weighted by followers has a rightward extreme that is higher than all the sentiments weighted by likes. The opposite is true for sentiments weighted by likes. **This is interesting as it suggests that at the extremes the trend changes between sentiments weighted by likes relative to followers**. In other words, the most liked post is likely an anti-trump post eventhough most posts liked have a pro-trump reach. Vice versa, the account with most followers seems to have a pro-trump view eventhough most accounts have a relatively anti-trump view. The reason that the extremes are not very visible is because the distribution of likes and followers is extremely skewed which compresses the data by a lot.

The previous exploration also suggests that there is no clear linear pattern between follower count and sentiments or likes and sentiments. If sentiment scores increase, there is no guarantee that it would increase likes or followers as we have very skewed distributions that can lead to very biased OLS estimates of a linear regression trend. Still, we will see if a linear trend exists at all using scatterplots.



From this visualization we can see that the sentiments of the tweets seem to have a horizontal relationship with log follower count. In other words, sentiments of the tweets do not influence the number of followers an account has. This could mean that the followers of the account follow for reasons that are unrelated to politics. *This definitely inserts doubt into our use of follower count* *as a reach factor as the political sentiments of the twitter user does not necessarily equate to the political sentiments of the follower if the follower is not following for political reasons.*



Even here, it is clear that sentiment scores seem to have a horizontal relationship with likes. Again this brings into question the decision to use likes as a reach factor as it doesn't seem that people are liking the post on the basis of the intensity of the political sentiment expressed. However, the fact that they are liking a post related to Trump must mean that they agree or disagree with the contents of the post at some level i.e it reflects their political preferences.

Arnav Dey

Visualizations

Main message

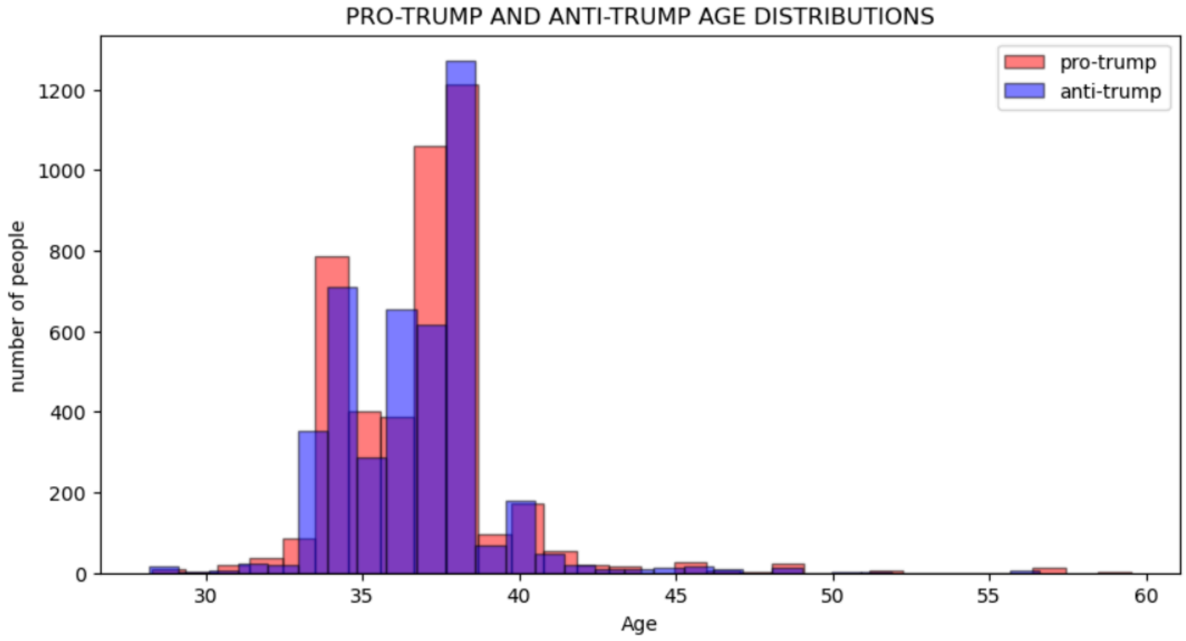
“Our data is mostly representative since there are no big differences between pro-trump and anti-trump twitter users when it comes to age and income. This makes it representative because the academic literature states that demographic trends in the US population is not a good predictor of political position which is the result we seek to find in our data.”

One of the objectives of this paper is to understand how representative our data is. To do this, we will use our academic literature and visualizations to asses the representativeness of our data. To do this, we first merge median income and median age at the county level for all our observations. Now the literature suggests that political partisanship cannot be determined through income and age i.e that there are no real differences in the political positions of people if we group them by income and age.

To visually assess this statement, we will first classify each of our tweets into pro-trump (republican) and anti-trump (democrat). Our classification criterion is the following:

- If the tweet has a positive sentiment (compound score > 0) we classify it as pro-trump
- If the tweet has a negative sentiment score (compound score < 0) we classify it as anti-trump

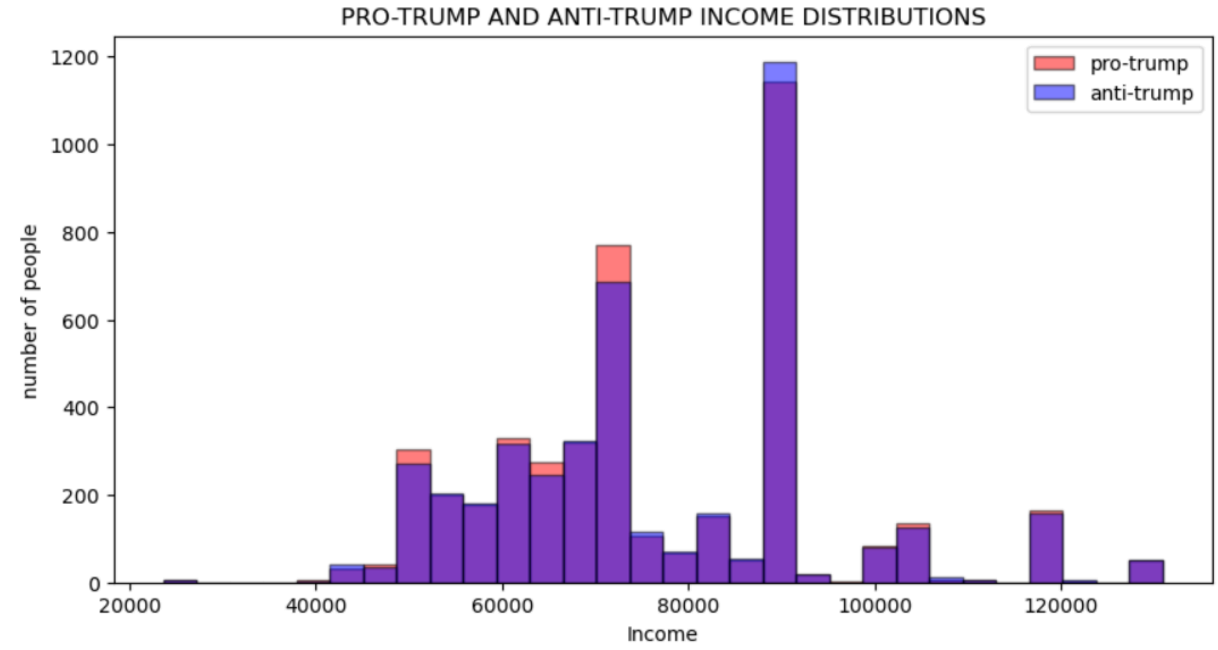
Now, we will plot a histogram of **pro-trump and anti-trump age distributions** to see whether there are any significant differences between the two.



This shows that the pro and anti Trump age distributions are almost the same. This suggests that there are no significant differences age-wise between pro and anti Trump positions *as per our classification criterion*.

There are some odd patterns like age 37 twitter users being relatively more pro-trump than anti-trump, whereas people of the age of 36 are relatively more anti-trump. It seems that around the age of 34 to 35 that we have a similar pattern with relatively more pro-trump twitter users. These are likely random patterns or a result of other variables.

Now, we will plot a histogram of pro-trump and anti-trump income distributions to see whether there are any significant differences between the two.



Income differences between trump and non-trump positions seem to be even less significant. There are slightly more pro-trump people aroud the 70,000 to 80,000 income level with other patterns in other income groups. These are again, likely random patterns. Thus, there are three potential takeaways from this:

reflections

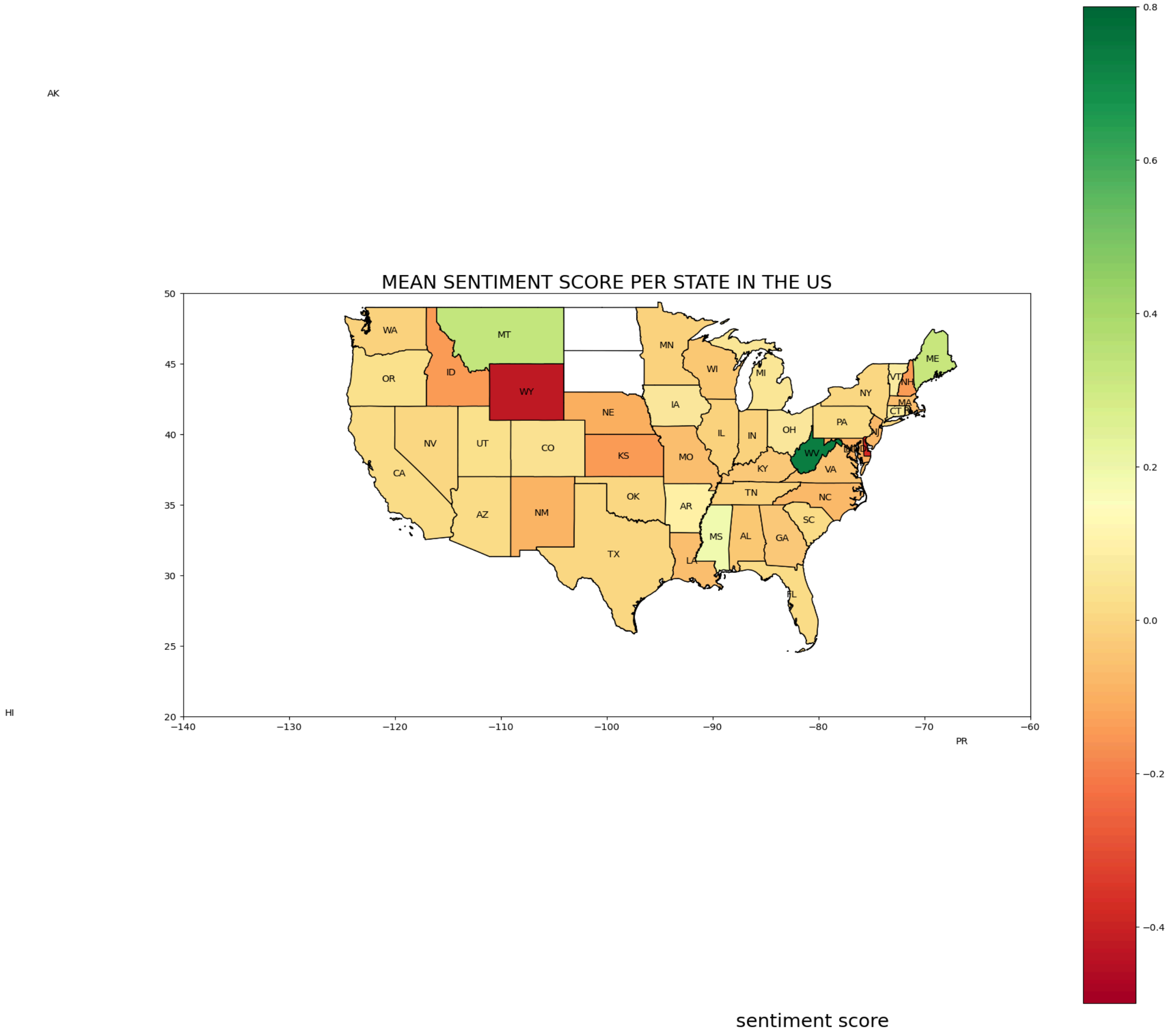
- Our sample of twitter data is representative of the population voter preference as trends do not significantly affect political position. (assuming our classification criterion and group assumptions are correct)

It could be that our classification criterion is incorrect or that our group assumptions are false. This classification criterion will be statistically evaluated later on. There might be some holes in our group assumptions considering the outliers we see in the income and age scatterplot.

Arnav Dey

Maps

In this section, we will look to create three maps for sentiment scores, income, and age. This will be used to measure the level of polarization in our tweets. This is important to understand the distribution of polarized opinions across the US states and how they may affect voter turnout geographically.



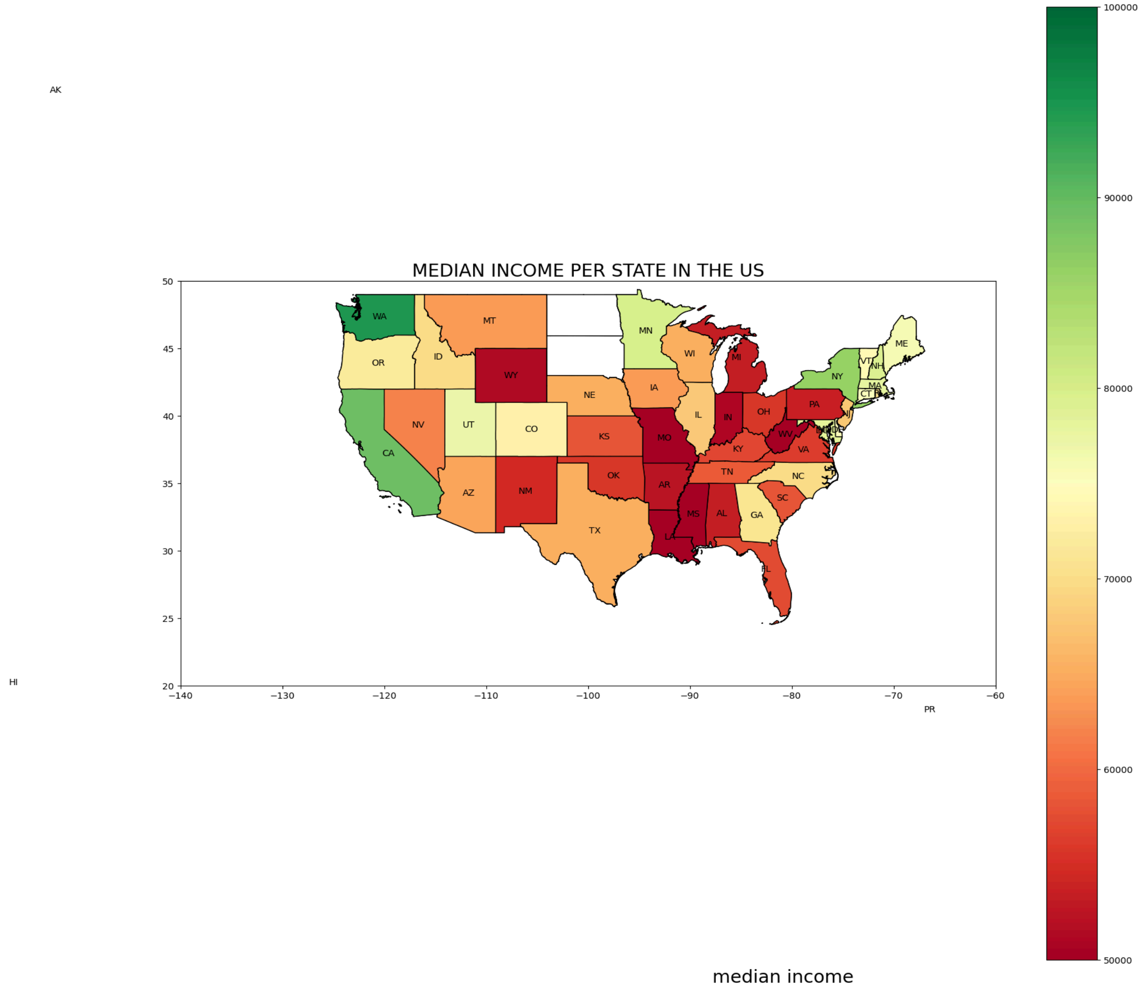
As per the map, tweets from the west and south seem generally neutral about Trump if not slightly negative on average. The state of Missisipi in the south is light green, surrounded by orange states. This suggests that Missisipi is on average a bit more pro-Trump in its tweets. The midwest is seemingly a bit more Anti-trump on average as they have a darker orange colour. The only state with tweets that holds very negative sentiments towards Trump is Wyoming and Delaware. West Virginia is shown on average to be very pro-trump in its tweets. The same goes with Montana and Maine.

These results seem to be consistent with the sentiment score distribution that we plotted before. The interquartile range of our sentiment scores is from -0.3 to 0.3 which means that 50% of the data shows fairly moderate sentiments. If there seems to be more negative sentiment averages this could be a result of negative sentiment outliers in the state specific tweets. Moreover, some states may be over or under represented in terms of how many tweets they have. It could just be that twitter users are genrally more anti-trump or that the population in general is more anti trump in 2020. This also affects the accuracy of their individual mean sentiment score estimates. *These drawbacks will be accounted for in future projects*.

Thus, the presence of potential outliers in an otherwise centred sentiment score distribution may hint at specific twitter users and handles. This could potentially skew our estimate of how sentiments impact voter turnout.

Now we plot a map for our variable income. This is important as it gives us insight into how income differences across US states may affect voter turnout. Since there is evidence that income also influences voter turnout, this will help us understand how voter turnout may vary in relation to income on top of sentiment scores. per state.

Moreover, we can check the correlation between income and sentiments. Based on our visualizations, there are no real differences between pro and anti trump twitter users based on age and income. However, comparing income with sentiment scores directly provides us with more of a scale to work with as opposed to a binary categorical variable.



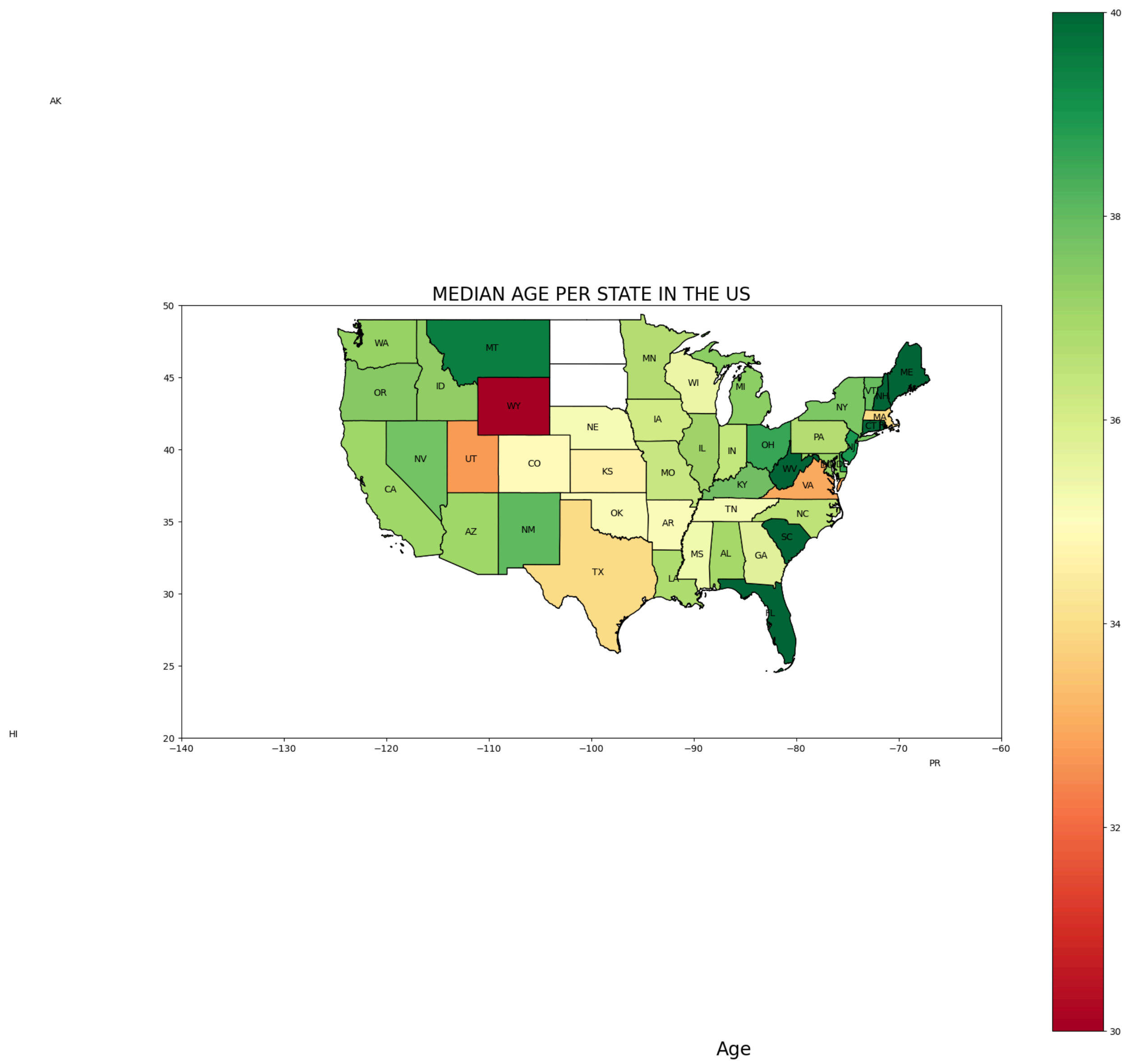
When we observe this map, we see some interesting connections.

Firstly, Wyoming which is the most-anti trump state as per our data is also a state with a very low median income relative to other states. This does not necessarily mean anything as West Virginia which also has a low median income relative to the other states is a highly pro-trump state. Thus, it seems that at the extremes income does not affect polarization in any direct way. The midwest states do have a relatively low median income, also being slightly anti-trump states based on sentiments. However, considering our extreme cases, it does not imply a strong connection.

Thus, we could say that visually we could not determine any connection between income and sentiments which is good as it may suggest low multicollinearity. If at all income affects voter turnout significantly, we would likely see an impact in states like california, washington, Wyoming etc where the median income is on the extreme. However, considering the wealth distribution of states like california and Washington, median income may give us biased estimates. This is because the most valuable tech companies reside in California and Washington with billionaires like Bill gates, Larry page etc residing there.

We may have a slight bias with our twitter data, where our twitter users may be young, relatively low income users. This is because wealthier people may not express extreme sentiments on twitter because they have other channels by which they can express their political views.

We will also look at the map for age, to understand if age correlates with sentiments or income. Moreover, this will give us a preliminary understanding of the potential effect on voter turnout.



Wyoming again presents itself as an odd case with one of the relatively lowest median ages, median incomes being a heavily anti-trump state. West Virginia this time is a state with a relatively high median age, also having a very pro-trump position. This may suggest that states with generally older populations support Trump whereas relatively younger states are against Trump.

Then again this may not be the best conclusion as states like Florida and South Carolina seem to have relatively very old populations but are still somehwat anti-trump. This again brings back the possibility of outliers influencing our estimates. Moreover, the west coast having a relatively old population are also somehwat anti-trump.

One possible explanation is that twitter data consists of mostly young people which is what gives the anti-trump bias. This would explain why states with relatively older populations also seem to be anti-trump. Older people may not be very active or use Twitter for political expression as much as young people.

Now that we have explored sentiment scores and some of our other predictors, we will merge additional predictors and calculate voter turnout from our merged data.

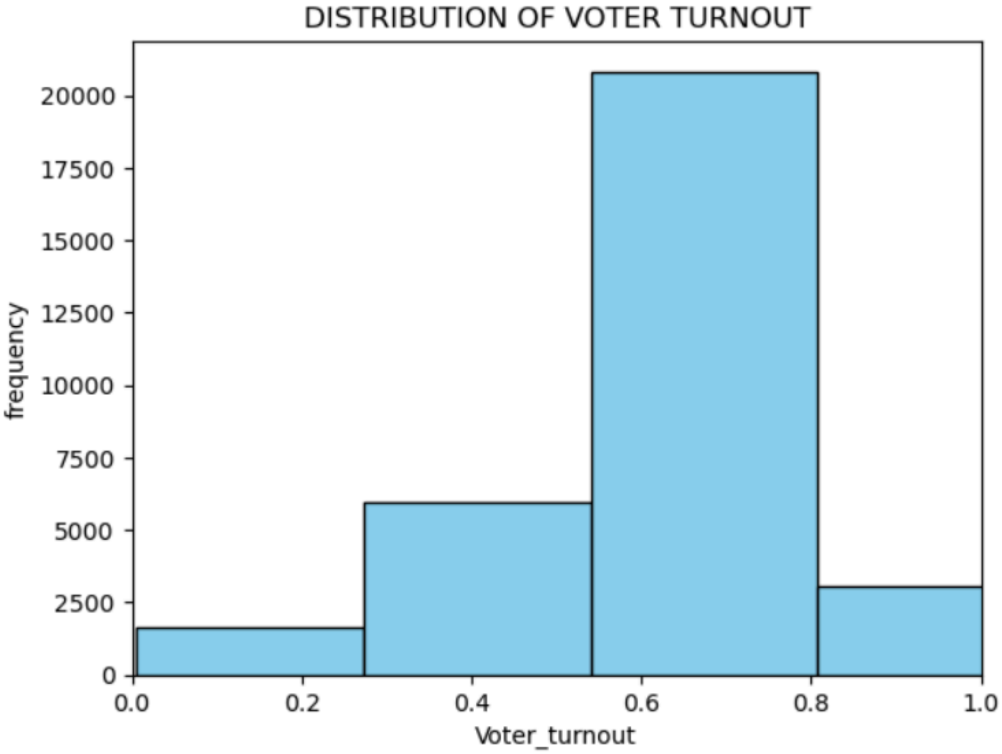
The new dataset we have merged in the 'Election,Covid,and Demographpic data by County' for the 2016 and 2020 elections. We have specifically merged the county statistics, using county names as an index for both datasets.

dataset link: https://www.kaggle.com/datasets/etsc9287/2020-general-election-polls

This data was sourced from FiveThirtyEight and DataWorld as per the data author. These sites have county level data of the variables in the dataset.
The data consists of county-level observations of variables like covid case count, number of residents of different races, total number of votes, voting age citizen, poverty level etc

All these variables can act as potential exogenous controls for our regression model where we check how polarization (through extreme sentiments) influences voter turnout. Like age and income, these variables may also influence voter turnout and so they must be added to prevent problems with endogeneity.

This dataset will also be used to calculate the **voter turnout** variable using the formula: total votes/Voting age citizen.



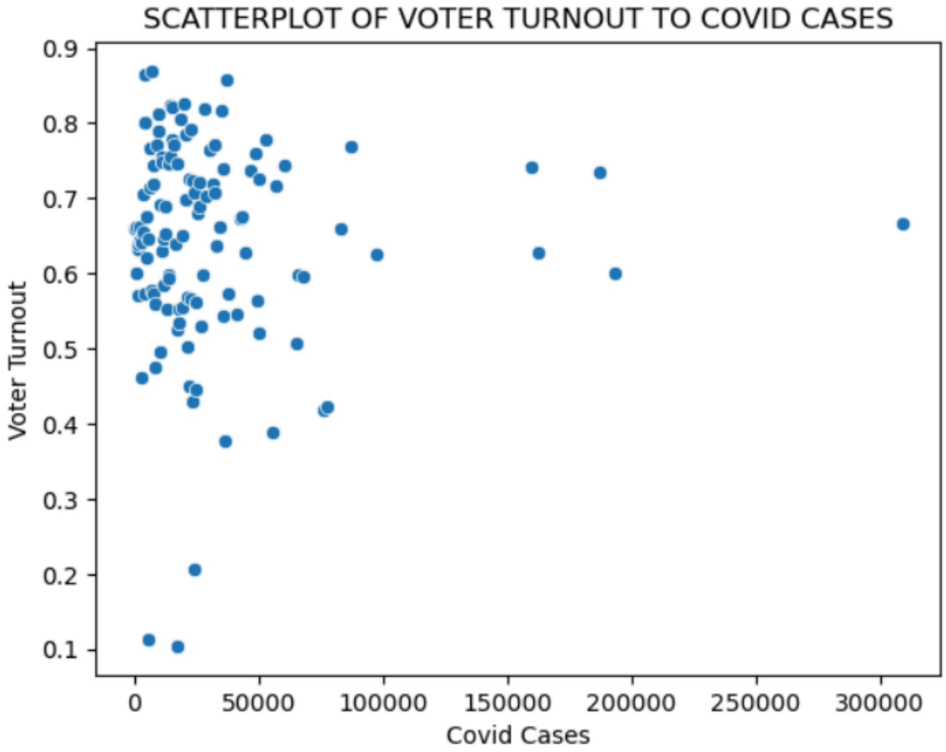
Description

voter turnout on average seems to be high, with the peak at around 0.7 which means a 70% voter turnout rate at the county level. Even historically this is on the higher side of voter turnout rates from the 2000's onwards. (US elections project). This may suggest that polarization may have increased political participation as opposed to reducing at as the book 'Red Nation, Blue Nation' suggests. T

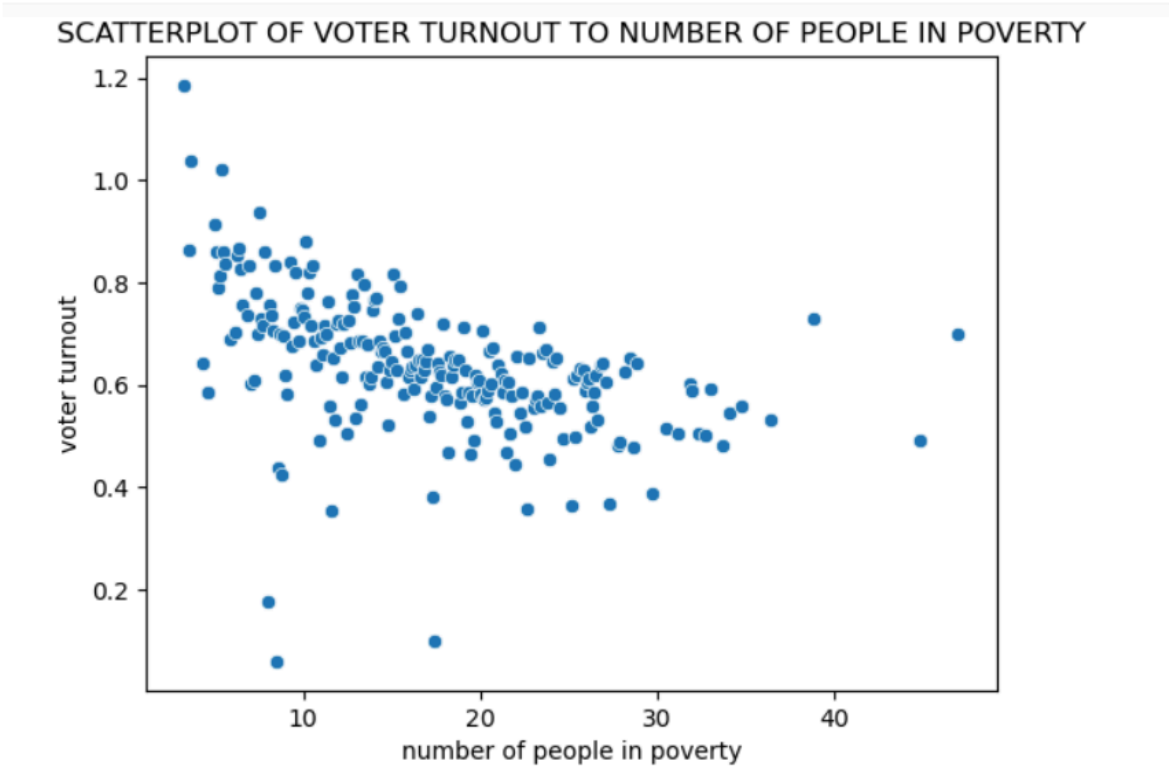
Interpretation

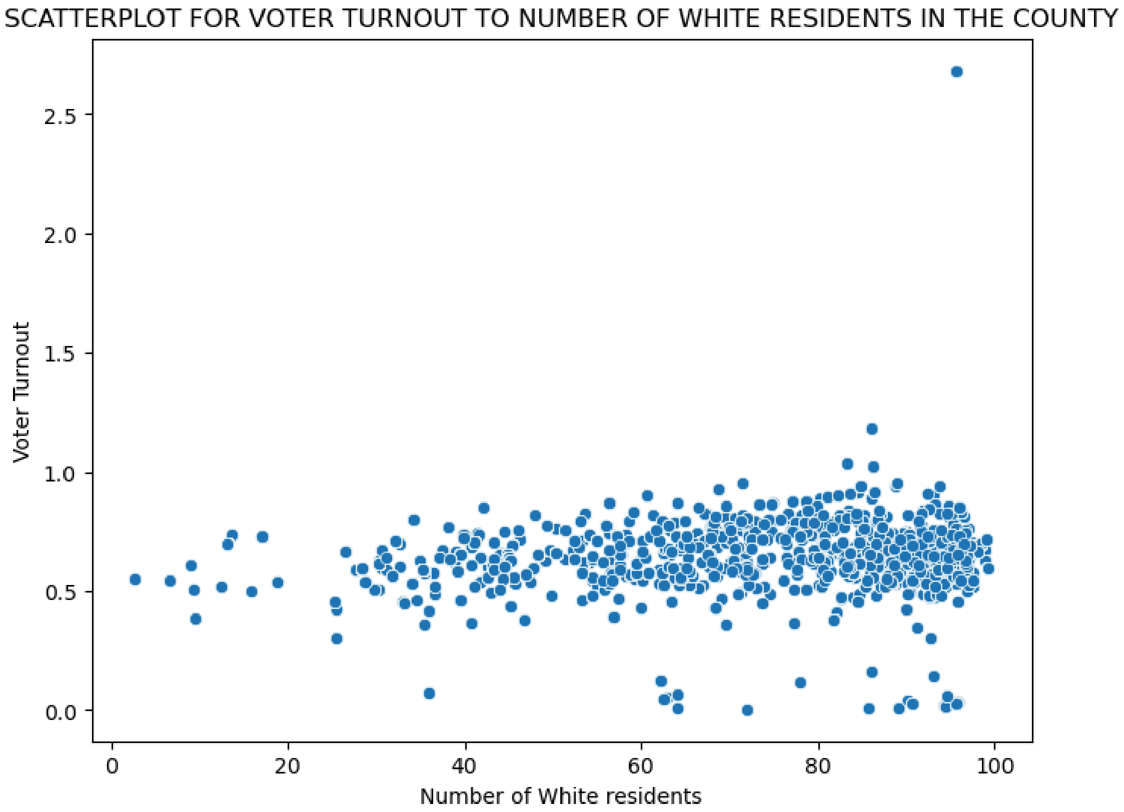
his could also be due to the other exogenous controls we are considering in the model. Covid may be one such factor that increased voter participation as suggested by the research. This could be because more people wished to vote so as to curtail Trump's policies or insure against the perceived negative impact of Biden's proposed policies.

Let us get a sneak peek to see if our additional exogenous controls (poverty, and covid cases) may be the cause of higher turnout rates.



Arnav Dey


SCATTERPLOT OF VOTER TURNOUT TO NUMBER OF PEOPLE IN POVERTY
voter turnout
1.2
1.0
0.8
0.6
0.4
0.2
10
20
30
40
number of people in poverty


SCATTERPLOT FOR VOTER TURNOUT TO NUMBER OF WHITE RESIDENTS IN THE COUNTY
Voter Turnout
2.5
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
Number of White residents

Arnav Dey

interpretation

Based on our scatterplots, it does not seem like our added exogenous controls have a very strong correlation with voter turnout.

number of people in poverty seems to have a negative correlation with voter turnout i.e that the more people that are in poverty, the less they participate in voting.

When it comes to the races, all of them have an approximately linear relationship with voter turnout, but the relationship is not very significant. We have only shown the plot for white residents, but the ones for the other races are very similar. These means voter turnout does not vary between races or that no one race is more politically active than the other.

This suggests that either polarization (which is a measure of twitter sentiment extremity) explains significant variation in voter turnout or that there are other exogenous variables we are not considering. Another potential explanation could be that the twitter data and the counties depicted are not representative i.e that they represent counties that do not represent the norm when it comes to these trends (as we have research evidence of correlation between our exogenous controls and the response)

A causal statement on the variables in use and whether they are part of the true model will be explored when we run the regression model.

The next few steps would be to merge educational level and marriage status as predictors at the county level. Based on the data in the US census bureau, we are merging the number of people who have a high school, bachelors and masters degree at the county level along with the number of married households at the county level.

Now with our host of predictors, we can run our regression model.

Results

The predictors we are going to use are the following:

Median age, median income, weighted sentiments by follower, weighted sentiments by likes, covid cases, poverty, married households, bachelors degrees, graduate degrees, number of White residents, number of Hispanic residents, number of Black residents, number of Asian residents.

Response variable: voter turnout (%)

nature of relationship between response and predictors: As per our scatterplots with voter turnout as the dependent variable, and some of our new predictors (like covid cases, poverty ,races) we see that the relationship is approximately linear. As per our exploration of age and income distributions between pro-trump and anti-trump supporters, the normality of the distribution suggests that its effect on voter turnout is also likely linear. Our distribution of sentiment scores is also very normalized, which implies that its effect is likely linear.

Selection of predictors:

The number of cases can affect voter turnout as a voter may chose not to vote if there is a high case count, citing the risk of infection as a reason to stay at home. According to a study *The COVID-19 pandemic and the 2020 US presidential election by Leonardo Baccini, Abel Brodeur, and Stephen Weymouth*[5], there exists a signficiant positive correlation between covid cases and

[5] Brodeur, Abel, et al. "The COVID-19 Pandemic and US Presidential Elections." 2020.

voter turnout where covid cases *increased* voter mobilization (contrary to our hypothesis) when they ran a 2SLS Instrumental variable regression. This suggests that the nunber of cases can have a potential effect on voter turnout.

poverty also influences voter turnout. the lower the family income relative to the poverty line, the lower the voter turnout. This was established by a Columbia study by Professor Robert Hartley in his report *Unleashing the Power of Poor and Low-Income Americans*[6]. Thus, poverty status does affect voter turnout.

Moreover, according to the MIT Election lab voter turnout is affected by 'income, older age, and being married'. This justifies our inclusion of Age and Income to our data. We will incorporate marriage status later.[7]

In the paper '*Using Experiments to Estimate the Effects of Education on Voter Turnout' by Rachel Milstein and Donal P Green*[8]*,* the findings conclude that there is a well documented strong relationship between education level and voter turnout. Moreover, this relationship has a strong causal grounding as well.

[6] Hartley, Departments of Statistics and Political Science, Columbia University. (2020, August). Unleashing the power of poor and low-income Americans. Poor People's Campaign. https://www.poorpeoplescampaign.org/resource/power-of-poor-voters/

[7] Voter turnout | MIT election lab. (n.d.). MIT Election Lab. https://electionlab.mit.edu/research/voter-turnout

[8] Sondheimer, Rachel M., and Donald P. Green. "Using Experiments to Estimate the Effects of Education on Voter Turnout." *American Journal of Political Science*, vol. 54, no. 1, 2009, pp. 174-189.

Arnav Dey

In '*who votes now, Demographics, issue*s, Inequality and turnout in the United States' By Jan E Leighley and Jonathan Neigler'[9] states that there are significant disparities in voter turnout rates among different racial groups in the United States. Specifically, it points out that African Americans and Hispanics have historically lower voter turnout rates compared to Whites.

Thus, our predictors of interest have been added based on the following literature and the stated relationship between our predictors and voter turnout.

8 models

- full model: with age as categorical and weighted sentiments by follower
- 2: age as median
- 3: compound score not weighted sentiments
- 4: sentiment score weighted by likes
- 5: weighted sentiment by follower interaction with educational variables
- 6: weighed sentiment by follower interaction with race variables
- 7: weighted sentiment score by follower interaction with age categories
- 8: Income inceraction term with weighted sentiments by followers

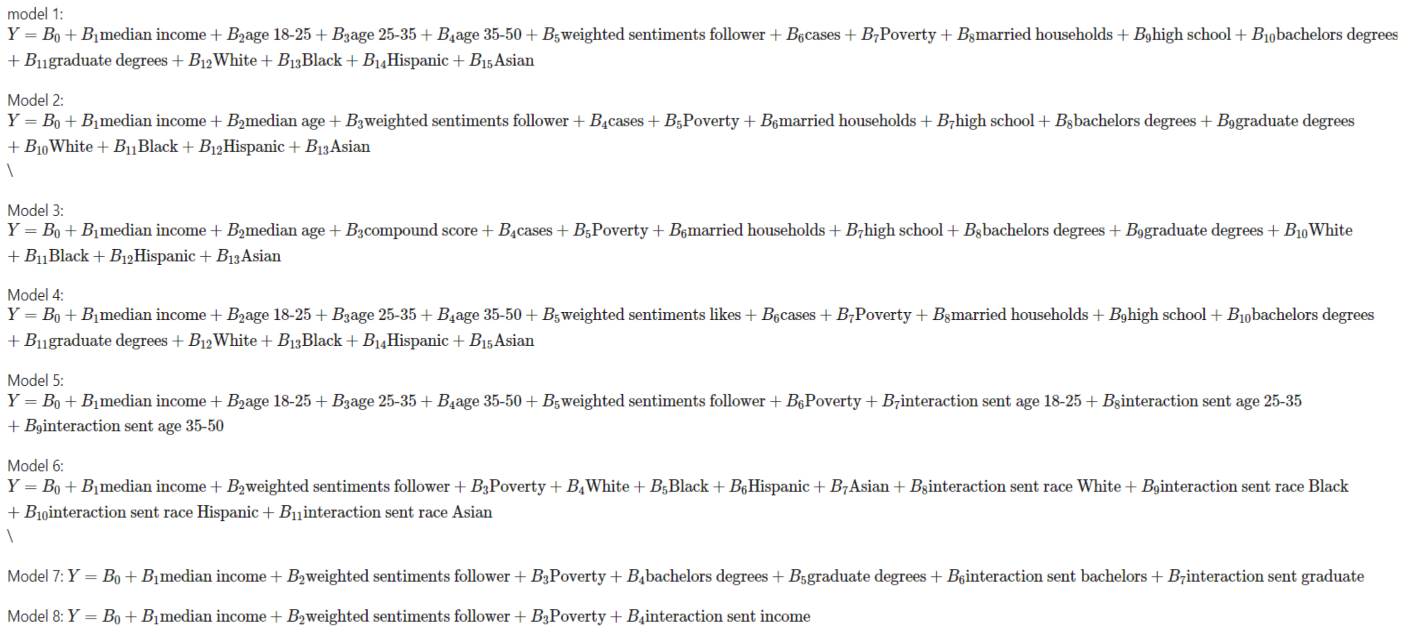
The reason I wish to run these specific regressions are for the purpose of compare and contrast

My full model contains all desirable variables that influence voter turnout.

- model 2 consists of median age as a continous numeric variable as opposed to categories. This is done to evaluate the overall impact of age and whether the categories tell a different story

[9] Leighley, Jan E., and Jonathan Nagler. "Who Votes Now?" 2013

- model 3 consists of regular compound score (sentiment score) as opposed to the weighted sentiments. This is important as it helps us compare how sentiments influence voter turnout depending on our 'reach factors'
- model 4 is another important model which consists of sentiments weighted by likes. Here, we can compare and contrast the coefficient from that in model 3 and model 1. This will grant us an understanding of likes as a 'reach' factor.
- model 5 seeks to investigate whether different age groups change the relationship that sentiment has with voter turnout. In other words, we are seeing how people from different age groups politically engage or disengage when there is sentiment polarization.
- model 6 creates interaction terms between sentiments weighted by followers and race to see if different races change their political engagement differently when sentiment polarization increases.
- model 7 does the same as before but for education levels (bachelors and graduate)
- model 8 interacts sentiments with income to see if people in different income levels engage differently when sentiment polarization increases



model 1:
$Y = B_0 + B_1\text{median income} + B_2\text{age 18-25} + B_3\text{age 25-35} + B_4\text{age 35-50} + B_5\text{weighted sentiments follower} + B_6\text{cases} + B_7\text{Poverty} + B_8\text{married households} + B_9\text{high school} + B_{10}\text{bachelors degrees} + B_{11}\text{graduate degrees} + B_{12}\text{White} + B_{13}\text{Black} + B_{14}\text{Hispanic} + B_{15}\text{Asian}$

Model 2:
$Y = B_0 + B_1\text{median income} + B_2\text{median age} + B_3\text{weighted sentiments follower} + B_4\text{cases} + B_5\text{Poverty} + B_6\text{married households} + B_7\text{high school} + B_8\text{bachelors degrees} + B_9\text{graduate degrees} + B_{10}\text{White} + B_{11}\text{Black} + B_{12}\text{Hispanic} + B_{13}\text{Asian}$
\

Model 3:
$Y = B_0 + B_1\text{median income} + B_2\text{median age} + B_3\text{compound score} + B_4\text{cases} + B_5\text{Poverty} + B_6\text{married households} + B_7\text{high school} + B_8\text{bachelors degrees} + B_9\text{graduate degrees} + B_{10}\text{White} + B_{11}\text{Black} + B_{12}\text{Hispanic} + B_{13}\text{Asian}$

Model 4:
$Y = B_0 + B_1\text{median income} + B_2\text{age 18-25} + B_3\text{age 25-35} + B_4\text{age 35-50} + B_5\text{weighted sentiments likes} + B_6\text{cases} + B_7\text{Poverty} + B_8\text{married households} + B_9\text{high school} + B_{10}\text{bachelors degrees} + B_{11}\text{graduate degrees} + B_{12}\text{White} + B_{13}\text{Black} + B_{14}\text{Hispanic} + B_{15}\text{Asian}$

Model 5:
$Y = B_0 + B_1\text{median income} + B_2\text{age 18-25} + B_3\text{age 25-35} + B_4\text{age 35-50} + B_5\text{weighted sentiments follower} + B_6\text{Poverty} + B_7\text{interaction sent age 18-25} + B_8\text{interaction sent age 25-35} + B_9\text{interaction sent age 35-50}$

Model 6:
$Y = B_0 + B_1\text{median income} + B_2\text{weighted sentiments follower} + B_3\text{Poverty} + B_4\text{White} + B_5\text{Black} + B_6\text{Hispanic} + B_7\text{Asian} + B_8\text{interaction sent race White} + B_9\text{interaction sent race Black} + B_{10}\text{interaction sent race Hispanic} + B_{11}\text{interaction sent race Asian}$
\

Model 7: $Y = B_0 + B_1\text{median income} + B_2\text{weighted sentiments follower} + B_3\text{Poverty} + B_4\text{bachelors degrees} + B_5\text{graduate degrees} + B_6\text{interaction sent bachelors} + B_7\text{interaction sent graduate}$

Model 8: $Y = B_0 + B_1\text{median income} + B_2\text{weighted sentiments follower} + B_3\text{Poverty} + B_4\text{interaction sent income}$

analyzing results

As per model 1, all age groups from around 18-50 negatively contribute to voter turnout. This is an odd result as we would assume that younger people are more likely to vote or are more politically active. Moreover, even adults until the age of 50 seem to not increase county-level voter turnout. Only for ages above 50 do we see that there is a positive contribution to voter turnout as implied in our intercept 1.45. This intercept means that given all else equal, someone above the age of 50 increases the mean voter turnout by 1.45 percentage points. If we look at model 2 (which has only median age) the coefficient for median age is 0.01167 which means that all else held constant, a per unit increase in median age increases mean voter turnout by 0.00167 percentage points. This implies that voters above the age of 50 constitute a large number of voters in the counties.

how it answers the research question

Coming to our most important result. Model 1 shows that the coefficient of weighted sentiments is 0.153. In other words, the mean voter turnout increases by 0.153 points for a per unit change in weighted sentiments(by follower). This implies that the more extreme or strong the sentiment of the opinion is, the greater it contributes to voter turnout. This is line with the mobilization hypothesis that higher polarization actually increases turnout. What is interesting is that if we look at model 3, the coefficient for compound score(sentiment score) is -0.002. This is very interesting as it suggests that any increase in positive(pro-trump) sentiments actually reduces voter turnout whereas the model 1 estimate suggests the opposite. Not only that, the coefficient for weighted sentiments is much larger than that of compound score. This difference can be related back to the formula of weighted sentiments. We are looking at the 'reach' of sentiments expressed using the follower count of the account that posted the tweet. A large number of negative sentiments with accounts with low follower counts do not have as much of an impact. However, given that the weighted sentiments distribution has more sentiment scores in the positive side it implies that even if we have a few pro-trump accounts, they are influential. This is a game changer, as the influence of accounts affects the reach of the political sentiment expressed. This suggests, that very influential accounts may have posted pro-trump rhetoric which contributed to higher voter turnout on average. Now when we look at sentiments weighted by likes in model 4, curiously a per unit increase in sentiment score reduces voter turnout by -0.092 percentage points. This is opposite to sentiments weighted by follower and in line with the effect of compound score. However, sentiments weighted by likes reduces voter turnout by a relatively less margin compared to compound score.

Now we face the reality that when we use followers as a reach factor, voter turnout increases but the opposite is true for likes as a reach factor. So sentiments weighted by followers does fit the mobilization hypothesis but sentiments weighted by likes does not. This may have something to do with the general political position the reach factors suggest. Sentiments weighted by likes are generally more pro-trump and sentiments weighted by follower is generally more anti-trump. This could hint that polarization with extremity of sentiments in the pro-trump side demobilizes voters or makes them less want to vote, whereas polarization with extremity of sentiments in the anti-trump side mobilizes people or makes them want to vote. What we are not sure of, is whether likes or followers is a better reach factor that models population political preferences. However, if we were to consider the fact that Biden won it could suggest that greater anti-trump sentiments or the fear of Trump's reelection was a good motivator for people to vote, then it suggests that sentiments weighted by followers is closer to the populations political preferences.

The remaining models incorporate interaction terms with sentiment scores to evaluate how polarization(weighted sentiment score) affects political engagement(voter turnout) within different age groups, races, and education levels.

Now in model 5 we look to see whether age group changes the relationship between weighted sentiment scores and voter turnout. Based on the coefficients, any increase in sentiments within the 35-50 age group reduces voter turnout relatively more than an increase in sentiments in the 25-35 group does. This implies that people in the 35-50 age group tend to vote lesser when sentiments are more extreme relative to people in the 25-35 age group. However, the p-values are massive which suggests that the coefficients are insignificant. This tells us that there is no real impact that age group has on the relationship between political engagement relative to the polarization of sentiments.

in model 6 we see create interaction terms for race to see how polarization influences political engagement between the races. Based on our coefficients, all the races reduce their political engagement as sentiments increase.

When we look at models 7 and 8, it shows that education level and income does not change the relationship that sentiment scores has with voter turnout, suggesting that people in different income groups and educational levels do not vote differently when sentiments become more polarized.

model performance evaluation

in terms of model performance, most models have similar adjusted R^2 suggesting that the variation they explain is equivalent. However, model 5 has an R^2 of 0.186 which is lower than all the other values which are 0.347 and above. This suggests that sentiments interacting with age group does not explain variation in voter turnout. Even the F-statistics of model 5 is much lower at around 903 compared to 1000 and above for models 1-4. Model 8 has the lowest adjusted R^2 of 0.157. Models 6 and 7 are the same, having lower adjusted R^2 and F-statistic compared to models 1-4. This is because they have less predictors and because the interaction terms are not significant.

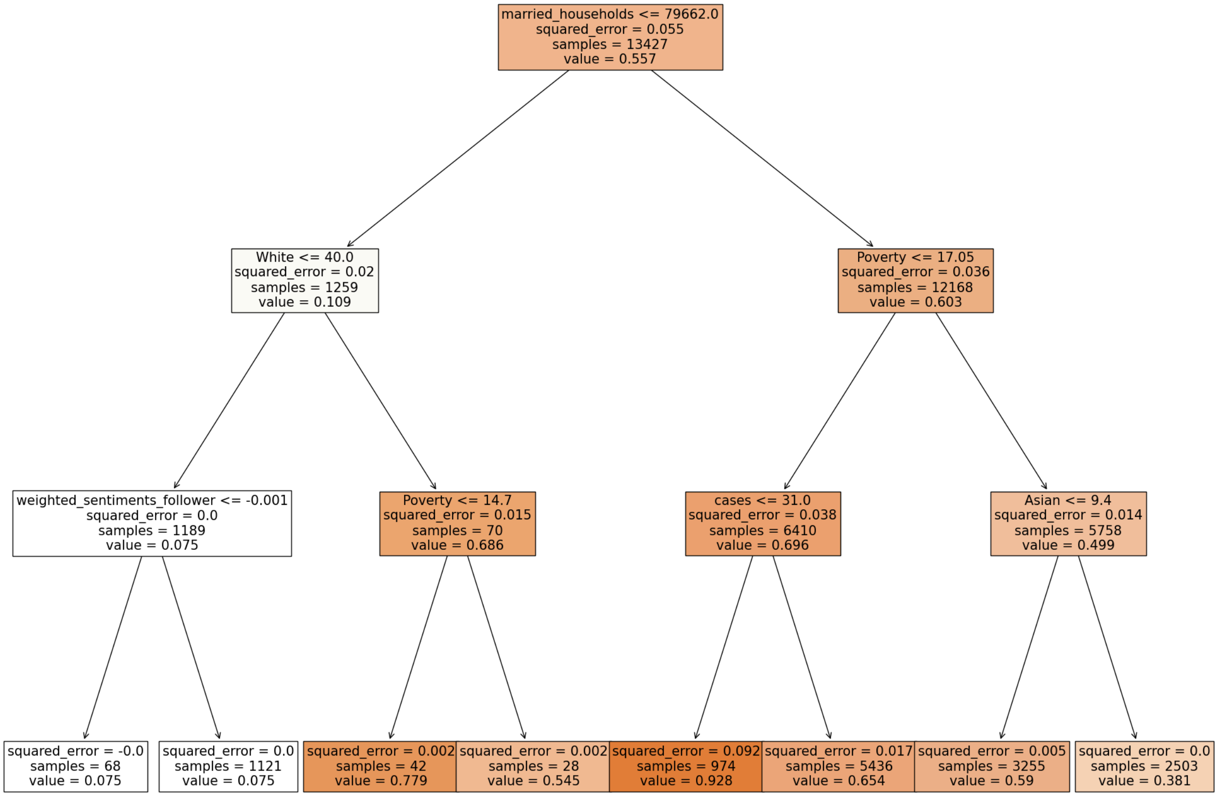
Now, In order to increase the predictive power of our model and identify variables that are important in predicting voter turnout, we will turn to more sohpisticated models in the form of a regression tree model and a random forest model

Objective function:

$$\min_{j,s}\left[\sum_{i:x_{i,j}\leq s,x_i\in Rj}(y_i-\hat{y}_{Rj})^2\right]$$

j is the number of predictors which in our case is 15 and s is the optimal split in predictor Xj. The summation is over all data points where the value of the j'th feature Xj is less than or equal to the split point s. Thus, what we are doing is finding the squared prediction error between the actual response Yi and the mean response in a specific region Rj for predictor Xj. We are then minimizing this prediction error i.e finding the splits that will give regions within the predictor space that minimizes this error.

We will select all the predictors from our full model and see that the regression tree outputs.



The objective with our regression tree will be to identify which branches ultimately lead to the highest mean voter turnout value. This will give us an understanding of what predictors significantly contribute to high voter turnouts and which ones lead to low voter turnouts. Our starting node is marriage status, which is interesting considering that this was not significant in our regression models. Given that the number of people married in the county is less than 79,662 and the number of white people is lower than 40, and the weighted sentiment score is less than -0.001, we get a mean voter turnout of 7.5%. However, even if sentiments are positive (> -0.001) mean voter trunout is still 7.5% suggesting that sentiments don't matter.

Moreover, the threshold of importance is the number of white people. If the number of white people are more than 40, and if the the poverty level is less than 14.7, we have a high mean voter turnout of 77.9%. This suggests, that voter turnout is high among white americans. Moreover, it is high given that poverty levels are low. Even if poverty levels are above 14.7,voter turnout is 50% which is still much higher than counties with a small number of white residents.

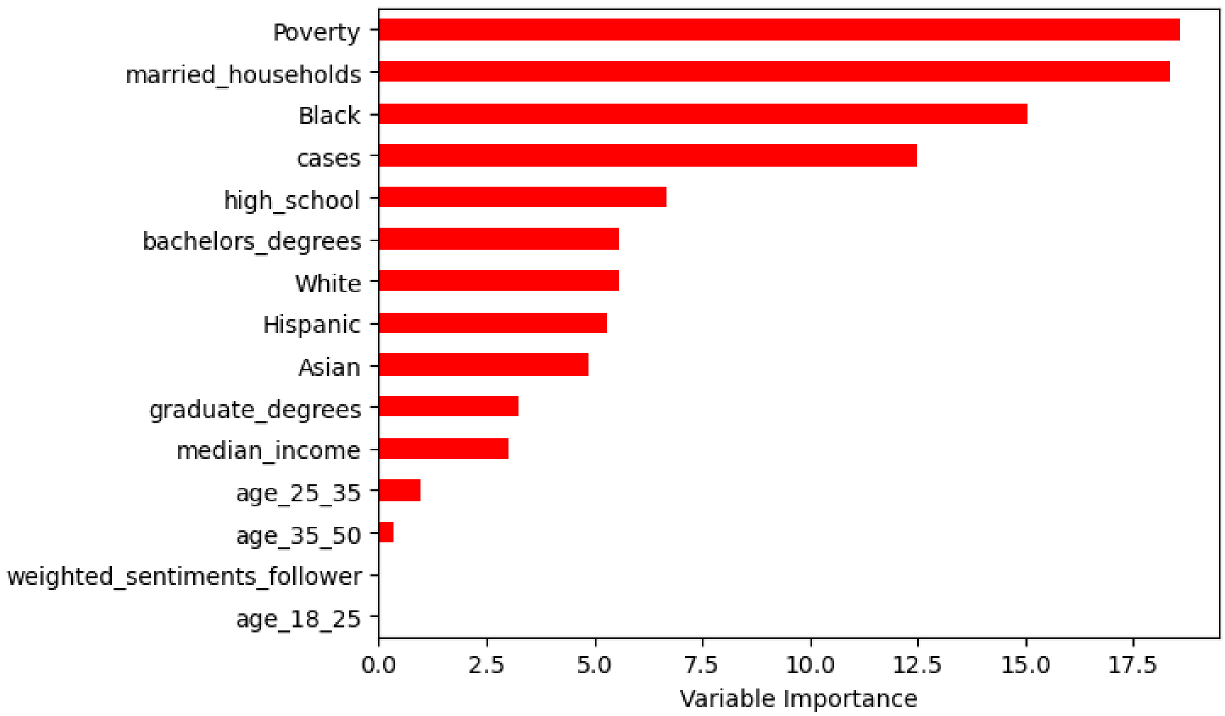
The highest percentage for mean voter turnout observed (around 90%) is when the number of married households is high(>79662) and poverty levels are below 17.05. this suggests that married citizens are more likely to vote. Additionally, a low number of covid cases, less than 31 leads to mean voter turnout of 92.8%. thus, we can say that if the number of covid cases are high, people are discouraged to vote. This is evident through covid case count being above 31 which drops the mean voter turnout to 65.4%.

We can conclude from this tree that marriage status, number of white residents and poverty levels contribute to high mean voter turnout. Sentiment scores or polarization is not significant as per the tree which is in opposition to our regression model where sentiments weighted by follower count had a postive coefficient of +0.053.

The mean squared error is at 0.0146 which is quite good.

Now, we will run the random forest model to see how much the prediction error can be lowered while also plotting an importance matrix to assess the important of our independent variables in terms of their predictive power.

The mean squared error for our random forest model is: 1.36*10^-5 which is far lower than that of our regression tree. To achieve this, the importance of variables is as follows:



Interpretation of importance matrix

From the importance matrix we can see that poverty level, number of married households, and number of black residents are the predictors with the most predictive power. As evident in the regression tree, poverty level and marriage status are very important, but here poverty is the most important predictor for voter turnout. Interestingly, where number of White American residents was an important predictor, number of Black American residents is now an important predictor. Number of White American residents is less important relative to number of black American residents which is not something we saw in the regression tree. Moreover, the covid case count is pretty important as well which does make sense. Two other important predictors not seen in the tree is education level, wherein number of high school students is a more important predictor than the number of bachelors students. This fits into our initial justification for adding education level and covid cases which was that they influence voter turnout.

Age and income are the least important predictors which is surprising. I assumed younger people may want to vote and would be more politically active but it doesn't seem to be the case. Additionally, income level does not affect voter turnout a lot which is a good sign as it implies that there are no economic barriers to voting.

The most important conclusion perhaps is that the weighted sentiments(followers) is not significant at all, echoing the result of the regression tree. Thus, if we were to answer our research question based on this we would have to say that sentiment polarization in twitter has no impact on voter turnout i.e that it is not representative of the populations political preferences.

comparing regression results to regression tree

comparing regression results to regression tree In our regression model, we captured the change in mean voter turnout conditional on our predictors of interest. Specifically when looking at sentiments weighted by followers, we captured that a per unit increase in sentiment scores increase mean voter turnout by 0.3 percentage points.

Our regression tree gave us a different output, focusing less on the specific relationship between our predictors and response, instead picking a set of predictors and regions within the predictor space that best predicted voter turnout i.e gave the lowest mean squared error. Thus, the regression tree gave us insight into the important variables that create categories within our voter turnout data i.e predictors that cause significant variation in voter turnout. Based on this model, weighted sentiments was not significant at all. Another contradiction is that Poverty has a smaller effect than sentiments in the regression model but is an important internal node in the regression tree.

Other curious contradictions include the significance of covid cases and number of residents of a particular race. When we looked at scatterplots comparing covid cases and race of residents, there was no significant relationship. However, the importance matrix then states that these two predictors are important. Since the random forest model did bring a low mean squared error, this suggests that non-paramteric methods may be better to use compared to parametric methods as there is no strong linear relationship between voter turnout and cases/race which means that regression models won't work well.

The reason for this could lie in the objective of each method. Regression models are parametric models which assume a linear functional form to allow for interpretability of the predictors influence on the response. Thus, we can investigate the potential effect twitter sentiments has on voter turnout without considering if sentiments significantly explains variation in voter turnout. Regression trees are non-parametric in that they are focused purely on prediction as opposed to interpretability. Thus, weighted sentiments can be a variable that influences voter turnout but not enough to justify being part of the regression tree. Tree's look purely at predictors that minimize prediction error.

Another way to think about this is through the Least squares estimaton of linear coefficients:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Here we minimize the residual sum of squares relative to our slope and intercept coefficients in our fitted y value. The caveat here, is that the fitted value follows a linear functional form. Thus, this limits the ways in which the OLS algorithm can limit the prediction error.

For regression trees however: $: \min_{j,s} \left[\sum_{i: x_{i,j} \leq s, x_i \in Rj} (y_i - \hat{y}_{Rj})^2 \right]$ h here, the fitted value is based on the mean of the response in that region. The partitions that determine a region are flexible i.e they can be changed, modified using complexity prunning etc. In other words, the tree can find the optimal set of predictors and partitions without any functional constraint as seen in least squares estimation.

this is why the prediction error on the regression tree is much lower. Thus, it seems that under both the regression tree and random forest, our sentiments are not significant predictors. Future studies can explore better ways to measure this polarization so that it is an effective predictor under the regression tree or random forest model.

Conclusion

Coming back to our research question of **'Is the degree of political polarization in opinions given by Twitter users a good predictor of voter turnout in the 2020 US elections?'** We can conclude that based on our linear regression model, sentiments weighted by follower count as a measure of polarization increases voter turnout by 0.304 percentage points under the 5% significance level suggesting that polarization increases voter turnout in line with our mobilization hypothesis.

Based on our regression tree, weighted sentiments by follower count is not a significant predictor at all suggesting that other variables like poverty level, race, covid count, and marriage status are significant predictors of voter turnout.

Thus, based on both model results we can say that twitter sentiment polarization does not explain much variation in voter turnout suggesting that US voters do not significantly consider the polarized political climate in their decision to vote. Therefore, twitter based sentiment polarization is not a good predictor of voter turnout.

We also additionally found that our sample data seems to be representative based on our findings that there are no major differences in age and income between pro-trump and anti-trump twitter users. However, the existence of outliers in age and income data along with a potential anti-trump bias in out twitter data suggests some problems with the representativeness of our data. The abovementioned conclusion that twitter sentiment polarization is not a good predictor of voter turnout may stem from the fact that twitter sentiments are not a good proxy for the populations political preferences or that the group assumptions made are not valid.

This paper is unique in that it seeks to statistically compute the influence of sentiments/polarization on voter turnout while also calculating the predictive power of sentiment scores on variation in voter turnout. Other papers we've come across tend to use public opinion surveys to guage sentiments and do not really build statistical models to evaluate the predictive power of political polarization on voter turnout. The unique result of the paper is using a regression tree and random forest model to identify significant variables that influence voter turnout, and that sentiments is not one of them.

This research can be developed further by creating more sophisticated regression models. One such idea is to create a multilevel model with a random intercept, varying by state and county. This will help us understand statistically how different states and geographies vote differently or how the sentiments of these geographies affect turnout differently. Moreover, finding an instrumental variable for sentiment score will be a very interesting extension as it will give us a better understanding of how sentiments affect voter turnout in the presence of potential endogeneity problems.

For our investigation on the representativeness of the data we could use classifiers like k-nearest neighbours, and other methods to create a good pro-trump anti-trump classification. In this paper we have used an arbitrary classification and so using better classifiers would be better to assess representativeness. Moreover, if state or county representation is skewed then we could use techniques like post-stratification to improve the generalizability of our model.

The scope of this project could be extended to compare tweets as a means of measuring polarization with other data sources like public opinion surveys, polls etc.

References

ABRAMOWITZ, ALAN I., and WALTER J. STONE. "The Bush Effect: Polarization, Turnout, and Activism in the 2004 Presidential Election." *Presidential Studies Quarterly*, vol. 36, no. 2, 2006, pp. 141-154.

Bor, Dennies, and Benjamin Lee. "Quantifying polarization across political groups on key policy issues using sentiment analysis." *arxiv*, vol. 2, no. 34, 2023, doi.org/10.48550/arxiv.2302.07775. Accessed 12 Apr. 2024.

Brodeur, Abel, et al. "The COVID-19 Pandemic and US Presidential Elections." 2020.

Kim, Seo-young S., and Jan Zilinsky. "Division Does Not Imply Predictability: Demographics Continue to Reveal Little About Voting and Partisanship." *Political Behavior*, vol. 46, no. 1, 2022, pp. 67-87.

Leighley, Jan E., and Jonathan Nagler. "Who Votes Now?" 2013.

Nivola, Pietro S., and David W. Brady. *Red and Blue Nation?: Characteristics and Causes of America's Polarized Politics*. Brookings Institution P, 2006.

Sondheimer, Rachel M., and Donald P. Green. "Using Experiments to Estimate the Effects of Education on Voter Turnout." *American Journal of Political Science*, vol. 54, no. 1, 2009, pp. 174-189.

Hartley, Departments of Statistics and Political Science, Columbia University. (2020, August). Unleashing the power of poor and low-income Americans. Poor People's Campaign. https://www.poorpeoplescampaign.org/resource/power-of-poor-voters/

Voter turnout | MIT election lab. (n.d.). MIT Election Lab. https://electionlab.mit.edu/research/voter-turnout