The Impact of Socioeconomic and Demographic Factors on Electoral Performance.

Bo Wie Tey

Abstract— This paper will focus on analysing the 2016 and 2020 US election results attempting to uncover the relationship between socioeconomic/demographic factors and the Republican electoral performance. Firstly, the extent these factors effect electoral performance is determined by using a combination of scatterplots and OLS. Secondly, regional variation was explored by modelling electoral performance using a regression framework and analysing the resulting residual choropleth map. Lastly, demographic shifts in the Republican voter base was determined by modelling the electoral performance of the 2016 and 2020 elections and comparing coefficient values. This paper aims to emphasise the benefits of iterative computational and visualisation techniques to derive and refine inferences made. All computational and visualisation techniques were conducted using different libraries from Python notably 'sklearn', 'geopandas', 'altair' and 'seaborn'.

1 Problem Statement

During 2016, the world stood in shock as Donald Trump was declared the president of the US. Since then, his presidency has been plagued with controversies making his loss in the recent 2020 elections unsurprising. However, the degree of loss was considerable trailing behind by 7 million in the popular vote and 74 in the electoral votes[1]. The republican loss has largely been credited to an increase turn out of youth voters and many minorities during this election term, seeing higher voter turnout when compared to similar timepoints in the 2016 election[2].

This paper aims to investigate the relationship between the demographic makeup of counties and the way they voted in the 2020 election. This can be achieved by answering the following questions:

1. To what extent is the Republican electoral performance related to socioeconomic and demographic factors of the counties?
2. Can a similar relationship be seen across the country?
3. What were the key demographic shifts in Republican support during the 2020 elections?

Datasets used in this research paper is suitable in answering the research questions as it contains key socioeconomic and demographic features of each county which will be of use in determining the effects of electoral performance. Election results also come from well-known news outlets which provide confidence in its reliability. Measurement on the county level will also help sufficiently capture the heterogeneity of the US population, avoiding the overshadowing of minority classes.

2 State of the Art

The use of data analytics to better understand a political parties core demographic has been an established technique in the political sphere since the early 21st century. The first notable use was during Barack Obama's 2012 campaign in which his team managed to raise over US$ 1 billion using this technology. Using both data and visual analytics, the campaign managed to determine core issues which resonated best with the different voter bases in each state and planned his campaign speeches accordingly during his campaign run across the country[3]. Unsurprisingly a similar strategy was also employed by Cambridge Analytica for Donald Trump's 2016 campaign which displayed the full force of the use of data analytics in elections[4]. By developing a psychometric profile for over a 100 million users the team managed to determine what type of political ad campaign would best suit a voter, e.g. a voter which scores high on the neuroticism scale will tend to respond better to fear mongering adverts. These techniques are not only limited to local political campaigns to better understand and rally their voters but are also used by foreign governments to sway elections. It will not be surprising that moving forward, the use of data will only increase in the political sphere as the 21st century progresses.

The two literatures analysed for this paper employs a similar visual analytics approach. The first studied how the UK's vote to leave the EU varied across the local authorities[5], whereas the second paper studied how the rise of populism varied regional across both the UK and the US[6]. Both papers employed a regression framework to determine the effects of explanatory variables on their respective independent variables followed by the development of a full multivariate model to predict election results. Where these two papers differ are in the methods used for feature selection. The former employed LASSO whereas the latter employed elastic-net which combines the penalization methods of both Ridge and LASSO regression. Furthermore, the first paper also used geographically weighted statistics alongside the spatial distributions of residuals from the global models to investigate variations between the Local Authorities, whereas the latter fitted a separate elastic-net models for each region studied to expose region variation.

In this study, like the papers above a regression framework will also be built to investigate the relationships between the different socio-economic factors and the election outcomes. However, instead of employing either LASSO or Elastic-net for feature selection, a Random Forest model with 600 trees will be built to determine feature importance instead due to its direct interpretability and ability to bypass issues due to collinearity between features. It should be noted that collinearity will still be considered when selecting features for our models with Random Forest acting as an initial filter. As the resolution of our study is much finer than both papers mentioned above, we would expect our results to vary a fair bit as the US population is very heterogenous across different counties.

3 Properties of the Data

5 separate datasets were compiled for the use of this analysis. election results for both 2016 and 2020 were obtained from GitHub[7]. The poster scraped data from 4 different news outlets (Townhall, Fox News, New York Times and Politco) providing confidence in the reliability of this data source. Socio-economic and demographic data were obtained from the US Census Bureau. As both datasets containing election results did not contain county level GEOIDs, a new column was created for each dataset mapping the County, State to their associated GEOIDs from the census data, followed by the merging of all datasets using said GEOIDs.

Election results for each political party was measured as a percentage of the total vote of each county. The elections results were then used to calculate the change in voter percentage share between 2016 and 2020 for both the Republican and Democratic party in each county. This then allowed to calculate the degree of swing using the formula below:

$$swing = \frac{Change_{Republican} - Change_{Democrat}}{2}$$

The rest of the columns in the dataset contained percentage measurements of different socio-economic and demographic factors of each county besides two measures total population and median income. The original data for occupation type and age groups were aggregated into larger classes to reduce the number of features in the datasets. Occupation types were aggregated into white collar (Management, Business, Science, Arts), pink collar (Service, Sales, Office) and blue collar (Construction, Maintenance, Production, Transportation, Material Moving) workers. Age groups were classified based on their associated generations (Silent, Baby Boomer, Gen X, Gen Y, Gen Z). Other features included in the dataset were percentage of people – bachelors or higher, veterans, no health insurance, male, white, black, asian, latino, english speaking and immigrants.

To determine if missing values were present in the dataset, ‘pandas’ inbuilt ‘isna’ was used. It revealed that missing data occurred in the measurements for the number of males (~1%), latinos (~1%) and immigrants (~0.4%). Instead of removing rows which contained missing datapoints which would remove a county completely, the state mean of each feature was calculated and used to replace said missing values.

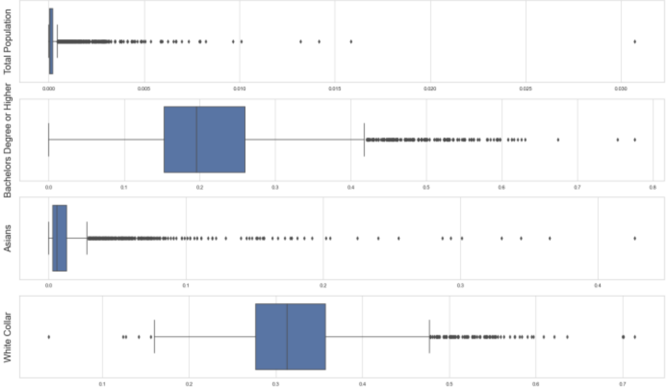


Figure 1. *Boxplots evaluating data quality.*

Boxplots was the preferred method to determine data quality due to its ability to detect erroneous values, outliers and usual distributions present in the data. In the context of this analysis however, values which deviate greatly from the mean were not removed from the dataset. Furthermore, due to the heterogeneity of the US populations, values which deviate far from the mean might represent densely populated city areas which tend to contain measures that deviate greatly from rural and suburban counties e.g. **Figure 1.** The outlier in the far right of the total population boxplot represents Los Angeles County. Left-skewed distributions are also expected in our data as they generally represent minority populations e.g. Asians, white collar workers, people with a bachelor’s degrees or higher.

4 Analysis

4.1 Approach

Analysis pipeline employed in this paper will be discussed in this section in order to demonstrate how an iterative use of both computational and visual analytic techniques can be used to help derive and refine knowledge in order to answer our questions of interest.

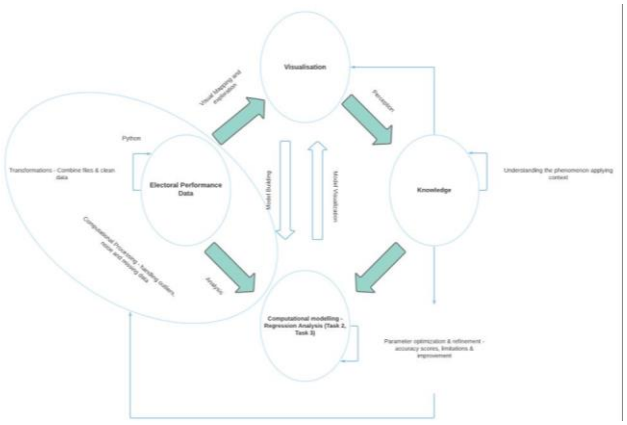

Visualisation
Visual Mapping and exploration
Perception
Python
Transformations - Combine files & clean data
Electoral Performance Data
Model Building
Model Visualization
Knowledge
Understanding the phenomenon applying context
Computational Processing - handling outliers, noise and missing data
Analysis
Computational modelling - Regression Analysis (Task 2, Task 3)
Parameter optimization & refinement - accuracy scores, limitations & improvement

Figure 2. *Analysis Workflow Plan - based on Kein et al’s 2010 work.*

Data cleaning and pre-processing makes up the initial and one of the most crucial steps of our analysis. Firstly, it allows us to pose the relevant datasets to the problem in order to determine if said data is suitable (containing relevant features) in answering the problems using human reasoning supported by metadata information. Secondly, cleaning data will allow us to identify and manipulate faulty data points i.e. outliers, “NaN” values which will improve our ability for visual reasoning and modelling accuracy. This step is constantly revisited throughout the analysis process in order to better fit our data to our analysis question e.g. feature engineering.

Task 1

Ordinary Least Squares will first be used to determine if electoral performance can be explained by the socio-economic factors using the R^2 value. Furthermore, it will provide an initial gauge on which features are important by interpreting the coefficient outputs of the model once all values are scaled. Following this, a Random Forest model containing 600 trees will be used rank features based on their overall importance in the model. Lastly, variance inflation factor will also be used to determine if high levels of multicollinearity exist between the features of our dataset. Scatterplots with an outputted regression line will be used to visually display the degree of correlation between the factors and electoral performance, this will help identify which factors play an important role in both measures of electoral performance. A correlation heatmap will then be used to determine the correlation between all features. Human reasoning will then be used to infer feature importance from these figures which will later be further verified using the empirical evidence from our computational results.

Task 2

Features selected from our prior step will then be used as independent variables for our multivariate regression model. Residuals will then be calculated for both metrics and will be displayed on a choropleth residual map. This will expose regions which are best described by our model and which regions are not. Geographically weighted statistics will then be used to explore the extent of which the factors of interest may vary across the US. Background knowledge and human reasoning will then be applied on both visual outputs to infer why the regions may vary and which possible factors should be included to produce a more robust model for these specific regions.

Task 3

To determine the demographic shift in the republican voter base between 2016 and 2020, a regression model will also be built for the outcomes of the 2020 elections using the same features from the 2016 model. A choropleth residual map will then be used to determine if model quality is like that of the 2016 election. Coefficients are then plotted on a line graph to compare changes. Human reasoning will then be used to reason if the shifts in coefficients are valid.

4.2 Process

Task 1

In order to prepare the data for our analysis, all values were scaled between 0 and 1. Features which were measured as a percentage were divided by a 100. Total population was scaled by dividing the values of each county by the total US population in 2019 (328200000). It should be noted that our analysis will only constitute counties in the mainland US (excluding both Alaska and Hawaii). Median income was scaled using the formula below:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

With the data pre-processing step complete. scatterplots between the socio-economic factors and electoral performance (percent Republican Vote and Swing towards Republican) with a regression line were plotted. The visual output shows that *total population, bachelors or higher*, *white collar workers, median income, number of Asians, Gen Z* and *Gen Y* show a negative correlation between both metrics of electoral performance, with *total population* and *median income* being the highest. Whereas a positive correlation is seen between two factors which are *blue collar workers* and the *silent generation* (**Figure 3**). Results from our two OLS summaries reveal that the socio-economic factors explain around 68.3% and 44% of the variance seen in percent of Republican vote and Swing towards GOP respectively. Furthermore, it also reveals that the factors *Latino*, *male*, *immigrants* as poor explanatory factors for both metrics with extremely low coefficient values, which is in line with their associated scatterplots, with an almost horizontal regression line also indicating little to no correlation with the two electoral metrics. This satisfies the stopping condition – as both scatterplots and R^2 results show that these factors are correlated and can help explain the variation seen in both electoral measures

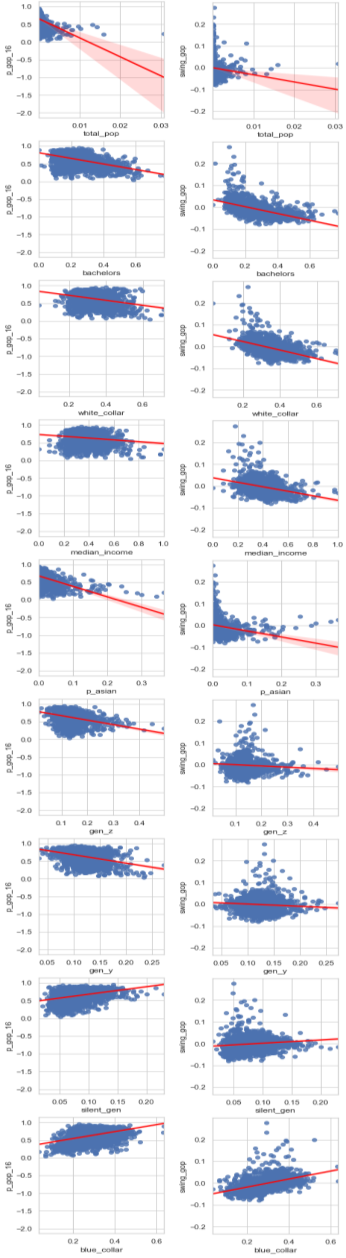


Figure 3. *Scatterplots between socio-economic factors and electoral performance.*

Feature importance output from our Random Forest models also show similar results with both *bachelors or higher* and *total population* appearing in the top 5 most important features for both metrics alongside *blue collar workers*, providing further confirmation for the factors to priorities in our following regression models. Lastly, variance inflation factor and a correlation heatmap was used to determine the collinearity between features. High correlation is seen within the 2 different occupation types (*blue- and white-collar workers*) and between *bachelors*, the factors *white* and *black* are also highly correlated which is also supported by the VIF values for these features with a VIF of > 5. High correlation is also seen between the different generation classes with the highest being between baby boomers and the silent generation (0.73). The factor *Asians* is also seen to moderately correlate with the factor's *total population*, *bachelors* and *white-collar workers* and *blue*-collar workers. VIF was then recalculated after removing the features *blue-collar worker*, *white-collar worker*, *Asians*, *blacks*, *silent generation*, *Gen* Z and Gen *Y* the resulting in all features containing a VIF of < 1.8.

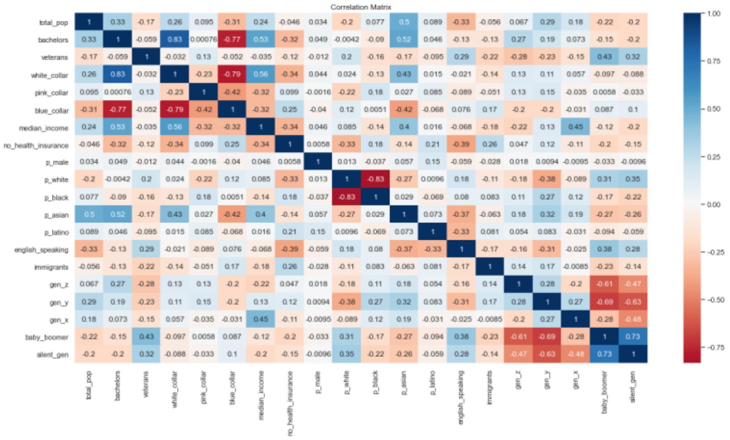


Figure 4. *Correlation heatmap.*

Overall, we can conclude that socioeconomic factors of each county do play a role in the resulting electoral performance of the GOP party, particularly the factors *total population bachelors or higher* being universal factors for both electoral performance metrics, which is in line with the generally assumption of the Republican voter base who mainly live in rural/suburban areas (low *total population*) who are within the lower working class (*no bachelor's degree*).

Task 2

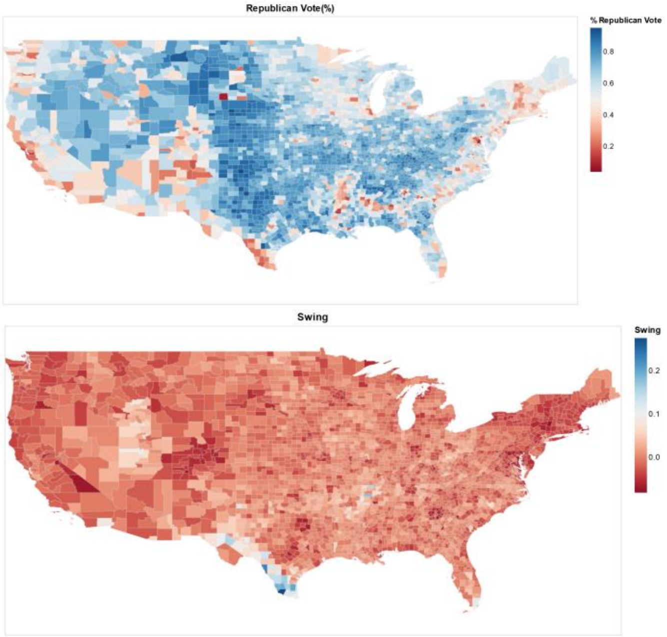


Figure 5. *Choropleth map of electoral performance.*

Before models can be built to uncover potential regional variations that are present, a choropleth map of the US was first plotted to determine if the same factors are driving the two phenomena. **Figure 5** reveals two maps which have extremely different distributions, indicating that the underlying factors that drive these two phenomena are different. Therefore, independent variables used for both our models are as follows:

$$\mathit{Perecent\ Republican\ Vote} = \ \mathit{total\ population} + \mathit{bachelors} + \mathit{white} + \mathit{veterans}$$

$$\mathit{Swing} = \mathit{total\ population} + \mathit{bachelors} + \mathit{english\ speaking} + \mathit{median\ income}$$

Residuals were then calculated and scaled between the values of -1, 1 and plotted on a choropleth map. From the outset we can see that the model predicting percent Republican vote suggest regional variation is present across the country, with the model being able to describe the west, upper mid-west, northeast and southeast fairly well but fails to predict the southwest and lower Midwest indicating that there are other drivers which might be in play(**Figure 6**).

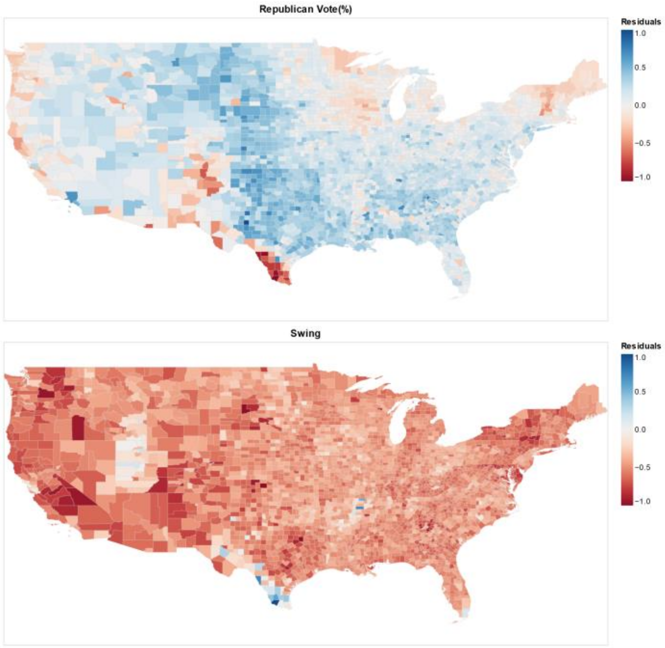


Figure 6. *Residual Choropleth Map*

On the other hand, the choropleth map for swing indicates that the model is severely overpredicting the degree of swing to GOP across the entire United States. This indicates that the features used might not be enough in describing swing. With that in mind, Elastic Net was employed using features which were determined to be low in correlation from the previous task as inputs to predict swing. Resulting coefficients for all features however were 0. This will generally only occur when the input features are highly correlated or that the dependent variable of interest is not predictable by a linear equation. As our input features were already screened for multi collinearity, this indicates that a linear regression framework would not be suitable in predicting swing. Notably however, there are a few regions in the swing map whereby residuals are low (<-0.1) indicating that a linear equation/input features are suitable in describing swing e.g. Box Elder County. Furthermore, across both maps, counties in lower Texas are fairy distinct from the rest of the local regions (**Figure 6**).

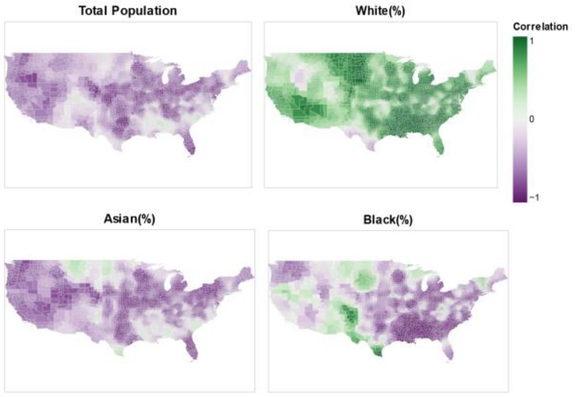


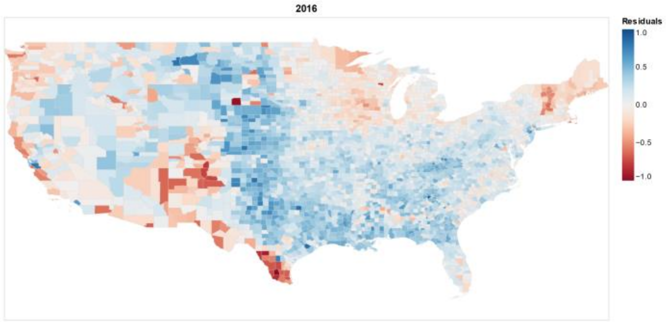
Figure 7. *Geographically weighted correlation maps*

To determine possible factors that could cause the regional variation seen in the percent Republican vote choropleth map, geographically summary statistics were first generated and correlation between the transformed factors and percent Republican vote were then calculated and plotted on a choropleth map. The maps generated suggest that the factors *Black (%)* and *Asian (%)* might be the socioeconomic factors driving the regional variation seen in the Southwest and lower Midwest as these factors have a similar pattern of variation in the regions of the lower Midwest and Southwest. Our maps also reveal that *total population* and *white (%)* are universal predictors of Republican vote with a homogenous choropleth map. The stopping condition is met as results from our visual displays both reinforce the fact the regional variation is present across both measures.

Task 3.

Modelling efforts to describe the degree of swing on the prior task proved to be unsuccessful possibly because the phenomena did not follow a linear model. Therefore, in order to determine the voter base shift of the Republican party, modelling will be done on the 2020 results, and coefficient values of the new model will be compared to the coefficient values from the 2016 model. In order to fully capture the shift in trumps voter base, features used in the models for this task will be modified to include other minority classes of trumps voter base whilst still maintaining as little collinearity between features as possible.

$$\mathit{Republican\ Vote}(\%) \\ = \mathit{Median\ Income} + \mathit{Asian} + \mathit{Black} + \mathit{Latino} + \mathit{Blue\ Collar} + \mathit{Veteran}$$



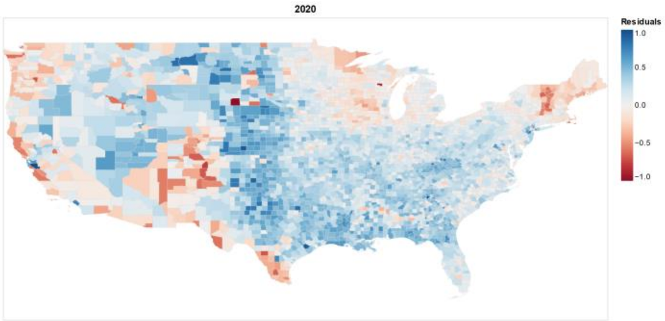


Figure 8. *Residual Maps for the 2016 and 2020 models.*

Before a comparison can be made, model quality must first be assessed. Visually, there is little to no differences detected between the 2016 model from the prior task and the model currently used, with similar patterns of under predictions and overpredictions seen (**Figure 8**). To further verify this, OLS was used to determine the proportion of variation explained by both our models. Only a 5% decrease is observed in the R^2 value indicating that the model quality is still acceptable for deriving inferences from our coefficients. It should be noted that the new features employed better describe the 2020 elections results relative to the 2016 results as the number overpredictions (residual >-0.5) decreases significantly, being able to better describe the Midwest and Southwest better (**Figure 8**). This is also reflected on the R^2 value, being able to describe around 60% of the variation.

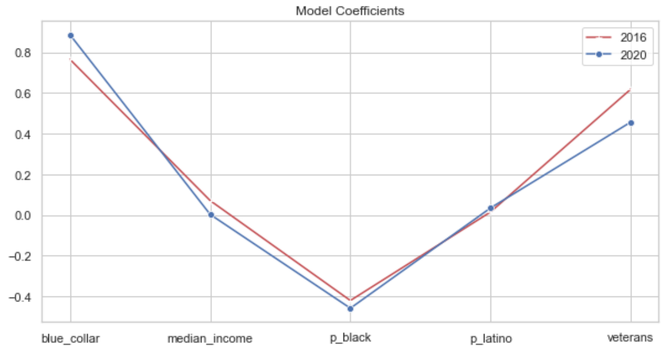


Figure 9. *Model Coefficients*

The *Asian (%)* feature was removed from **Figure 9** as its coefficient values were significantly higher than the other variables making it difficult to interpret the graph. With reference to **Figure 9**, we see a reduction in coefficient values for *median income, Black (%)* and *veterans* alongside *Asian (%)* [2016: -1.81, 2020: -1.83]. As coefficient values are the weights given to each feature for the model a reduction in positive weight generally indicates that it plays a weaker role in model output. Therefore, it can be inferred that counties with a high proportion of black, Asian and veterans which had a high median income saw a decrease in support for the Republican party. The only coefficient that saw an increase between the two time periods were the *blue-collar workers.*

4.3 Results

Socioeconomic factors we included in our analysis plays an important role in determining electoral outcomes as it can account for over 50% of the variance of seen in Republican votes. This is further confirmed by the scatterplots which shows that correlation does occur across multiple variables. The most significant indicator for both percent Republican vote and Swing are the factors *total population* and *bachelors* which makes sense as counties with higher proportions of these two factors tend to be densely populated metropolitan areas which also tend to lean towards the Democrats.

As expected, regional variation is present and strong across the US. This is likely due to the sheer size of the US continent itself as larger countries tend to have less homogenous populations with far greater differences between counties in the rural regions and metropolitan areas. Lastly, the loss of Asian[8], Black[9] and veteran[10] support for the Republican party is also in line with expectations due to the racist and hateful rhetoric's posed by Trump throughout his 4-year presidency that greatly alienated these demographics.

5 Critical reflection

In the cases of such analysis, it must be stated that conclusions of individual voting behaviour cannot be drawn. While relationships are observed in the county level, voting behaviours of groups within the county level cannot be determined. Furthermore, county issues which have affected voting behaviour also cannot be inferred and demographic shifts that might have occurred after 2019 (year in which the census data was collected) cannot be accounted for. However, the main advantage provided by using census data is that it includes every person and every vote providing a wider scope. If determining a relationship between individuals and electoral performance is desired, the use of opinion poll data might be more suitable as a person's vote, demographic and view are linked directly. However, such datasets are expensive to collect in a scale which can expose regional variation.

The choice of county resolution was a key issue in our analysis as it greatly affected the performance of our global models due to the extremes which exist between counties. A finer approach would be to create individual models for larger regions e.g. West, Midwest, Northeast etc or to use state level data if a global US model is desired. The use of multivariate linear regression models for our modelling efforts proved to be useful in determining regional variation and determining shifts in voter base, however, for predicting degree of swing a polynomial model should be considered.

Visualising geographic relationships using a choropleth map provided a clear visual output which made interpretation easy playing a key role in answering the second task. Furthermore, as the direct county shapes were used, variations from specific locations were apparent and easy to interpret. Geographically weighted correlation choropleth maps also helped provide inferences for the regional variations seen by our models. However quantifying effects these variations would have on electoral performance is not direct. The use of human reasoning in our overall analysis helped provide necessary context for our modelling efforts during feature selection. Furthermore, it also played a crucial role when building conclusions helping us link the results from the visual outputs to the computational outputs.

Overall, the techniques used in this analysis is not to restrained to only the domain of voter analysis. This approach is highly generalisable and can be applied to other complex spatial multivariate data e.g. educational availability, impacts of global warming, accessibility to healthcare etc. Future works that could be built on this analysis would include modelling electoral performance for key swing states in the recent 2020 elections and including other socioeconomic variables such as religion, homeownership rate, mode of transport etc.

Table of word counts

Problem statement	243/250
State of the art	493/500
Properties of the data	478/500
Analysis: Approach	523/500
Analysis: Process	1418/1500
Analysis: Results	173/200
Critical reflection	433/500

References

The list below provides examples of formatting references.

[1] Politico.com. 2021. *Live Election Results: The 2020 Presidential Race*. [online] Available at: <https://www.politico.com/2020-election/results/president/> [Accessed 1 January 2021].

[2] Woodward, A., 2021. How Biden Won: The Voters And Demographic Shifts Key To Unseating Trump. [online] The Independent. Available at: <https://www.independent.co.uk/news/world/americas/us-election-2020/how-biden-won-states-2020-beat-trump-b1720878.html> [Accessed 2 January 2021].Yan, Z., 2021. *How Data Analytics Helped Obama Win The 2012 US Presidential Election*. [online] SCMP. Available at: <https://www.scmp.com/yp/learn/college-uni-life/university-programmes/article/3071524/how-data-analytics-helped-obama-win> [Accessed 3 January 2021].

[3] Kaiser, B., 2019. *Targeted*. 1st ed. Harper Collins, pp.100-140.

[4] Beecham, R., Slingsby, A. and Brunsdon, C., 2018. Locally-varying explanations behind the United Kingdom's vote to leave the European Union. *Journal of Spatial Information Science*, [online] (16). Available at: <http://www.josis.org/index.php/josis/article/view/377/201> [Accessed 1 January 2021].

[5] Beecham, R., Williams, N. and Comber, A., 2020. Regionally-structured explanations behind area-level populism: An update to recent ecological analyses. *PLOS ONE*, [online] 15(3), p.e0229974. Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229974#sec006> [Accessed 1 January 2021].

[6] McGovern, T., 2021. *US County Level Election Results 2008-2020*. [online] GitHub. Available at: <https://github.com/tonmcg/US_County_Level_Election_Results_08-20> [Accessed 1 January 2021].

[7] Ali, S., 2021. *False 'Thug' Narratives Have Long Been Used To Discredit Movements*. [online] NBC News. Available at: <https://www.nbcnews.com/news/us-news/not-accident-false-thug-narratives-have-long-been-used-discredit-n1240509> [Accessed 1 January 2021].

[8] Goldberg, J., 2021. *Trump: Americans Who Died In War Are 'Losers' And 'Suckers'*. [online] The Atlantic. Available at: <https://www.theatlantic.com/politics/archive/2020/09/trump-americans-who-died-at-war-are-losers-and-suckers/615997/> [Accessed 1 January 2021].

[9] Yam, K., 2021. *Trump Is 'Legitimizing' Hate Incidents Against Asian Americans: U.N. Experts*. [online] NBC News. Available at: <https://www.nbcnews.com/news/asian-america/u-n-experts-trump-legitimizing-hate-incidents-against-asian-americans-n1243791> [Accessed 1 January 2021].