Decoding the 2016 U.S. Presidential Election

A Statistical Investigation of Socio-Economic and Demographic Impacts

Zilu Wang
SCHOOL OF MATHEMATICS & STATISTICS
University of Glasgow

1. Introduction

The 2016 United States presidential election marked a pivotal moment in American political history. On November 8, 2016, Republican nominee Donald Trump secured the presidency with 304 electoral votes, compared to Democratic nominee Hillary Clinton's 227 electoral votes. Trump became the first president without prior political or military experience and the fifth president to win the presidency despite losing the popular vote, having received almost 3 million votes less than Hillary Clinton. This unexpected outcome brought renewed attention to the complex interplay between demographic and socio-economic factors and voting patterns in U.S. elections.

This study is motivated by the need to understand the underlying socio-economic and demographic factors that influenced the voting patterns in the 2016 U.S. Presidential election. Despite the extensive analysis that has been conducted, there remains a gap in comprehensively understanding how specific factors correlate with voting behaviours at both the county and state levels. Additionally, developing predictive models based on these factors could be invaluable for political analysts, campaign strategists, and social scientists in forecasting future election outcomes.

Existing research on the 2016 U.S. presidential election has produced mixed results, particularly regarding the impact of demographic changes on voting behaviour. While some studies, such as those by Maggio (2021) and Flaxman (2018), observed a positive correlation between Hispanic population growth and increased support for Trump, others like Hill et al. (2019) found either no significant relationship or a trend favouring Clinton. Similarly, economic factors such as unemployment and income inequality have shown varying levels of influence on voter support for Trump. Education levels, particularly lower education, were consistently linked to Trump support. However, these studies often suffer from limitations such as restricted geographic focus, reliance on aggregate data, and potential biases from self-reported voting behaviour. This study seeks to address these gaps by analysing comprehensive data at both state and county levels, providing a more nuanced understanding of these relationships.

The aim of our study is to investigate the relationship between various socio-economic and demographic factors and voting preferences in U.S. counties and states during the 2016 Presidential election, and to develop models that can forecast election outcomes at different geographic levels. This research is guided by several key questions: Are there state-wide factors associated with a preference for one political party over another? Is there an association between specific socio-economic or demographic factors and the fraction of people voting for a Republican Presidential candidate in a county? Additionally, how well can a model associating socio-economic factors with 2016 election results predict the final outcome of the presidential elections at both state and county levels?

The dataset used in this study combines electoral data from the 2016 U.S. Presidential election with socio-economic and demographic data from the U.S. Census Bureau. The election data was collected from the MIT Election Data and Science Lab, containing the number of votes received by each party at county level. The demographic and socio-economic data encompass over 50 variables, including population, racial composition, education, housing, income, and employment statistics. By merging electoral data with extensive socio-economic and demographic data from reliable sources, this study benefits

from a detailed examination of the factors influencing voting patterns across different geographic levels, enhancing the accuracy and depth of the analysis.

For this study, two key assumptions were made. First, the analysis is restricted to a binary classification of party affiliation, considering only Democratic and Republican candidates, which simplifies the election results by excluding third-party candidates. Second, it is assumed that the socio-economic and demographic characteristics within each county are uniform, meaning all voters in a county are considered to share the same attributes as non-voters. While these assumptions help to streamline the analysis, they introduce limitations. The exclusion of third-party votes may overlook critical nuances in voter behaviour, and the assumption of uniformity between voters and non-voters might mask intra-county variations that could influence voting patterns. These factors should be considered when interpreting the results and drawing conclusions from the findings.

Understanding the socio-economic and demographic correlates of voting behaviour can provide valuable insights for policymaking, campaign strategies, and public discourse on political preferences. This research contributes to the existing body of knowledge by offering a detailed analysis of the 2016 U.S. Presidential election at both the county and state levels. It provides potential explanations for voting patterns and tests the predictive power of socio-economic factors, making significant strides toward resolving conflicting findings in the literature.

This quantitative study employs statistical methods, including descriptive and inferential statistics, correlation analysis, and statistical modelling, to examine the relationships between socio-economic factors and voting patterns. The subsequent chapters will detail the methodology, present the results, discuss the findings, and draw conclusions from the analysis. We will also explore the implications of our findings for future research and potential applications of the developed models in understanding and predicting electoral outcomes.

2. Methodology

This study employs a combination of statistical tests, correlation analysis, and predictive modelling to thoroughly analyse the relationship between socio-economic factors and voting patterns in the 2016 U.S. Presidential election. This methodology section outlines the steps and reasoning behind the approach used in this analysis. The process begins with data preprocessing, where electoral and socio-economic data are cleaned, merged, and prepared for analysis. This is followed by a series of statistical tests to explore the differences in socio-economic factors between Republican-leaning and Democratic-leaning areas. The appropriate effect size measures are also calculated to quantify the magnitude of these differences. Additionally, correlation analysis is conducted to assess the strength and direction of relationships between socio-economic variables and voting outcomes. The predictive modelling approach is then tailored to address the distinct characteristics of voting data at different geographical scales.

2.1 Data Preprocessing

In the data pre-processing phase, several key steps were undertaken to prepare the dataset for analysis. First, electoral data and socio-economic data were read in from corresponding CSV files. Next, the electoral dataset was modified by adding new

columns including the fraction of votes received by each party, the total number of votes cast, and the party that won in each county. The county-level electoral data was aggregated to the state level, summarizing key metrics such as total votes, Republican and Democratic vote counts, and determining the winning party for each state, which resulted in a new state level dataset. For the socio-economic dataset, we separated it into state-level and county-level datasets, then merged the electoral data with the corresponding socio-economic data by matching on state or county identifiers.

Unlike other states, Alaska reports voting data by House District (HD) rather than county equivalents. To ensure consistency in this analysis, an estimation method is employed to allocate votes to county equivalents such as Boroughs and Census Areas (Cinyc, 2019). This approach involves weighting absentee, early, and questioned votes based on differences in election day results observed across HDs that span multiple county equivalents. By accounting for these differences and assuming consistent relative turnout across precincts, we generated estimated vote shares for each county equivalent in Alaska, allowing for a more uniform comparison with other states' electoral results.

2.2 Statistical Tests and Effect Size Measures

To examine the differences in socio-economic factors between Republican-leaning and Democratic-leaning areas, a series of statistical tests were employed. The choice of test for each variable was determined by the distribution of the data, as assessed by the Shapiro-Wilk test for normality (Riffenburgh, 2006).

For normally distributed variables, Welch's t-test, an adaptation of the Student's t-test that does not assume equal variances between groups, was employed. This makes it more appropriate for this analysis where variances may differ between voting outcomes (Welch, 1947). For non-normally distributed variables with similar shapes across voting groups, we applied the Mann-Whitney U test, a non-parametric test that assesses whether one group tends to have larger values than the other (Mann & Whitney, 1947). In cases where the data did not meet the assumptions for either Welch's t-test or the Mann-Whitney U test, we employed permutation tests. These tests make minimal assumptions about the underlying distribution of the data, making them highly flexible and suitable for our diverse dataset (Nichols & Holmes, 2007).

To quantify the magnitude of differences between groups, we used appropriate effect size measures. For Welch's t-test, we calculated Cohen's d, while for the Mann-Whitney U test and permutation tests, we used Cliff's delta. Cohen's d is interpreted using the following thresholds: $|d| < 0.2$ indicates a negligible effect, $0.2 \leq |d| < 0.5$ represents a small effect, $0.5 \leq |d| < 0.8$ suggests a medium effect, and $|d| \geq 0.8$ signifies a large effect (Cohen & Jacob, 1988). For Cliff's delta, the interpretation thresholds are: $|\delta| < 0.147$ for a negligible effect, $0.147 \leq |\delta| < 0.33$ for a small effect, $0.33 \leq |\delta| < 0.474$ for a medium effect, and $|\delta| \geq 0.474$ for a large effect (Romano et al., 2006). These effect size measures allow for standardized interpretation of the differences between groups, facilitating comparisons across different variables.

To control for the increased risk of Type I errors due to multiple testing, we adjusted all p-values from the statistical tests using the Benjamini-Hochberg (BH) method (Li & Barber, 2019). This approach controls the false discovery rate while maintaining reasonable statistical power.

2.3 Correlation Analysis

Kendall's tau was chosen over Pearson's r in this study due to its non-parametric nature, making it more suitable for our data which exhibit strong skewness for most variables (Puka, 2011). Kendall's tau effectively handles non-linear relationships and is more robust to outliers, providing a more accurate reflection of associations in complex socio-economic data. These qualities make it the preferred correlation measure for capturing the relationships between socio-economic factors and voting outcomes in this analysis.

2.4 Predictive Modelling

The predictive modelling approach is strategically designed to address the distinct characteristics of voting data at different geographical scales. At the state level, the analysis is treated as a classification problem due to the "winner-takes-all" principle used in U.S. Presidential elections, where the outcome—whether a state is won by the Democratic or Republican candidate—is the primary concern. Conversely, at the county level, the analysis is approached as a regression problem because the proportion of votes for each party is crucial in determining the final state-wide outcome.

2.4.1 State Level Modelling

At the state level, we utilized two primary classification algorithms: Logistic Regression and Random Forest Classification. Logistic Regression is particularly valuable when the relationship between predictors and the outcome is approximately linear in the logit scale, offering easily interpretable coefficients that represent the log-odds of a state voting for a particular party (Sperandei, 2014). Random Forest Classification, an ensemble learning method, was chosen for its capacity to capture complex, non-linear relationships and interactions among predictors without making strong assumptions about the underlying data distribution (Ho, 1995). Additionally, Random Forest offers built-in feature importance measures, which help identify the most influential socio-economic factors in determining voting outcomes. Feature importance is usually based on the increase in Mean Squared Error (%IncMSE) when a feature is permuted. This is equivalent to the "permutation importance" or "mean decrease accuracy" measure.

2.4.2 County Level Modelling

For county-level prediction, where the response variable is a proportion (fraction of votes for a particular party), we employed Beta Regression and Random Forest Regression. Beta Regression is specifically designed for modelling proportions and is thus well-suited to our study, where the response variable is bounded between 0 and 1 (Ferrari & Cribari-Neto, 2004). This method accounts for the heteroscedasticity and asymmetry often present in proportion data, providing interpretable coefficients that represent the effect of predictors on the mean of the beta distribution. Similar to its classification counterpart, Random Forest Regression was selected for its ability to model complex, non-linear relationships and provide feature importance measures at the county level.

In fitting the beta regression model at the county level, one challenge is multicollinearity due to the large number of demographic and socio-economic variables. To address this issue, complementary approaches were employed. First, we applied principal component regression (PCR), which reduces the dimensionality of the predictor space while retaining most of the variance in the data. This helped to create orthogonal predictors, effectively eliminating multicollinearity. Second, we implemented a feature selection process using stepwise selection with a variance inflation factor (VIF) constraint. In this approach, predictors were added or removed iteratively based on their contribution to the model fit, but only if their inclusion did not cause the VIF of any variable to exceed 5. This strategy allowed us to balance the benefits of dimensionality reduction with the interpretability of individual predictors, while ensuring that the final model was not adversely affected by multicollinearity.

3. Results

3.1 Descriptive Statistics and Statistical Tests

Table 1 *Average of Selected Socio-Economic Indicators by Favouring Party: State and County Level*

Level	Favouring Party	Median House Values ($)	Multi-Structure Housing (%)	Income Per Capita ($)	Bachelor's Degree or Higher (%)	Foreign-born persons (%)	Asian pop. (%)
County	Democratic	$198,993	34.70%	$25,873	33.40%	18.82%	8.25%
County	Republican	$118,009	16.41%	$23,188	23.60%	6.12%	2.06%
State	Democratic	$267,919	32.77%	$31,668	32.80%	18.31%	8.92%
State	Republican	$143,747	21.09%	$25,524	25.80%	8.87%	2.73%

Table 2 *Statistical Test Results for All Variables at State Level (adj. p-value < 0.05)*

Variable	Test Used	Adjusted p-value	Effect Size Measure	Effect Size	Effect Size Interpretation
Median house values ($)	Mann-Whitney U test	2.09×10^{-6}	Cliff's delta	0.91	Large
Bachelor's degree or higher (%)	Mann-Whitney U test	6.79×10^{-6}	Cliff's delta	0.83	Large
Income per capita ($)	Mann-Whitney U test	3.38×10^{-6}	Cliff's delta	0.8	Large
Median household income ($)	Mann-Whitney U test	6.79×10^{-6}	Cliff's delta	0.77	Large
Multi-structure Housing (%)	Permutation test	3.14×10^{-4}	Cliff's delta	0.72	Large
Asian population (%)	Permutation test	1.05×10^{-2}	Cliff's delta	0.71	Large
Foreign-born persons (%)	Mann-Whitney U test	2.76×10^{-4}	Cliff's delta	0.70	Large
Non-English speakers (%)	Permutation test	2.08×10^{-3}	Cliff's delta	0.64	Large
Asian-owned businesses (%)	Permutation test	2.75×10^{-2}	Cliff's delta	0.63	Large
Persons under 18 years (%)	Mann-Whitney U test	2.65×10^{-3}	Cliff's delta	-0.58	Large

Mean travel time to work (min)	Welch's t-test	3.91×10^{-4}	Cohen's d	1.29	Large
Women-owned businesses (%)	Welch's t-test	2.08×10^{-3}	Cohen's d	1.17	Large
Poverty rate (%)	Welch's t-test	1.05×10^{-2}	Cohen's d	-0.92	Large
Hispanic-owned businesses (%)	Mann-Whitney U test	1.82×10^{-2}	Cliff's delta	0.47	Medium
Two or more races (%)	Mann-Whitney U test	2.11×10^{-2}	Cliff's delta	0.46	Medium
Persons under 5 years (%)	Mann-Whitney U test	2.75×10^{-2}	Cliff's delta	-0.43	Medium
Hispanic population (%)	Mann-Whitney U test	3.19×10^{-2}	Cliff's delta	0.42	Medium
Homeownership rate (%)	Permutation test	2.75×10^{-2}	Cliff's delta	-0.32	Small

Table 3 *Statistical Test Results for All Variables at County Level (adj. p-value < 0.05).* Some adjusted p-values are registered as 0 due to extremely low values

Variable	Test Used	Adjusted p-value	Effect Size (Cliff's delta)	Effect Size Interpretation
White alone, not Hispanic or Latino (%)	Mann-Whitney U test	7.69×10^{-113}	-0.63	Large
White alone (%)	Permutation test	9.43×10^{-153}	-0.58	Large
Homeownership rate (%)	Mann-Whitney U test	2.83×10^{-93}	-0.58	Large
Multi-structure housing (%)	Permutation test	0	0.57	Large
Asian population (%)	Permutation test	0	0.53	Large
Accommodation and food services sales ($1,000)	Permutation test	0	0.50	Large
Persons 65 years and over (%)	Mann-Whitney U test	2.88×10^{-67}	-0.49	Large
Foreign born persons (%)	Permutation test	0	0.47	Medium
Women-owned firms (%)	Permutation test	0	0.45	Medium
Private nonfarm employment	Permutation test	0	0.44	Medium
Retail sales ($1,000)	Permutation test	0	0.44	Medium
Bachelor's degree or higher (%)	Permutation test	0	0.43	Medium
Non-employer establishments	Permutation test	0	0.43	Medium
Total number of firms	Permutation test	0	0.43	Medium
Private nonfarm establishments	Permutation test	0	0.43	Medium
Language other than English spoken at home (%)	Permutation test	0	0.43	Medium
Population, 2010	Permutation test	0	0.43	Medium
Black-owned firms (%)	Permutation test	0	0.42	Medium
Population, 2014 estimate	Permutation test	0	0.42	Medium
Housing units	Permutation test	0	0.42	Medium
Asian-owned firms (%)	Permutation test	0	0.41	Medium
Households	Permutation test	0	0.41	Medium
Population per square mile	Permutation test	0	0.40	Medium
Building permits	Permutation test	0	0.40	Medium
Black population (%)	Permutation test	0	0.40	Medium
Hispanic-owned firms (%)	Permutation test	0	0.39	Medium

Median house values ($)	Permutation test	0	0.39	Medium
Veterans	Permutation test	0	0.36	Medium
Two or More Races (%)	Mann-Whitney U test	1.05×10^{-28}	0.31	Small
Female persons (%)	Mann-Whitney U test	7.46×10^{-23}	0.28	Small
Merchant wholesaler sales ($1,000)	Permutation test	0	0.27	Small
Retail sales per capita ($)	Mann-Whitney U test	1.12×10^{-21}	0.27	Small
Hispanic or Latino (%)	Permutation test	0	0.27	Small
Manufacturers' shipments ($1,000)	Permutation test	0	0.27	Small
American Indian- and Alaska Native-owned firms (%)	Permutation test	2.22×10^{-16}	0.27	Small
Population, percent change (%)	Mann-Whitney U test	1.67×10^{-18}	0.25	Small
Native Hawaiian and Other Pacific Islander alone (%)	Mann-Whitney U test	1.03×10^{-20}	0.24	Small
Living in same house 1 year & over (%)	Mann-Whitney U test	6.73×10^{-17}	-0.23	Small
Persons per household	Mann-Whitney U test	3.85×10^{-10}	0.18	Small
Per capita money income ($)	Permutation test	0	0.17	Small
Persons below poverty level (%)	Mann-Whitney U test	1.83×10^{-9}	0.17	Small
Median household income ($)	Mann-Whitney U test	2.24×10^{-6}	0.13	Negligible
Persons under 5 years (%)	Mann-Whitney U test	5.41×10^{-6}	0.13	Negligible
Native Hawaiian- and Other Pacific Islander-owned firms (%)	Mann-Whitney U test	1.24×10^{-45}	0.10	Negligible
Private nonfarm employment, percent change	Mann-Whitney U test	7.13×10^{-4}	0.10	Negligible

The results reveal a notable disparity in the number of statistically significant variables identified at the state and county levels. At the county level (**Table 3**), 45 out of 50 variables were found to be significant (adjusted $p < 0.05$), compared to only 18 out of 50 at the state level (**Table 2**). This discrepancy largely stems from the difference in sample sizes, with the county-level analysis including 3,140 observations compared to just 51 observations at the state level. In large datasets, small differences between groups can become statistically significant, potentially identifying variations that may not be practically meaningful. Conversely, the state-level analysis, with its smaller sample size, tends to focus on variables with more substantial effects, but can also miss minor differences.

Housing values and income levels emerge as significant factors distinguishing between Democratic and Republican-leaning areas. Both at the state and county levels, Democratic regions consistently exhibit higher median house values and income per capita compared to Republican regions. For instance, at the state level, Democratic states have a median house value of $267,919, compared to $143,747 in Republican states (**Table 1**). Mann-Whitney U tests further confirm these differences, with large effect sizes for variables such as median house values (Cliff's delta = 0.91) and income per capita (Cliff's delta = 0.80) at the state level. These results indicate that more affluent areas with higher property values tend to favour Democratic candidates.

Educational attainment is another critical factor in distinguishing voting patterns, with Democratic-leaning areas consistently having higher percentages of population with bachelor's degree or higher. At the state level, Democratic states report 32.80% of population holding a bachelor's degree or higher, compared to 25.80% in Republican

states. The county-level data mirror this trend, with Democratic counties having 33.40% of population with higher education, in contrast to 23.60% in Republican counties. Mann-Whitney U test confirms that educational attainment is a significant differentiator, with a large effect size (Cliff's delta = 0.83) at the state level. This aligns with the narrative that higher education levels are strongly associated with Democratic voting preferences.

Racial and ethnic diversity also plays a crucial role in determining voting preferences. The analysis highlights that Democratic regions are more diverse, with higher percentages of foreign-born persons, Asian populations, and other minority groups. For example, at the state level, Democratic states have an average foreign-born population of 18.31% and an Asian population of 8.92%, compared to 8.87% and 2.73%, respectively, in Republican states. Additionally, Republican areas show a higher percentage of White, non-Hispanic populations, with a significant negative association between White population percentages and Democratic voting preferences at both levels (Cliff's delta = -0.63 at the county level). The county-level data show similar trends, though the effect sizes are somewhat smaller. Mann-Whitney U test and permutation test results indicate large effect sizes for foreign-born persons (Cliff's delta = 0.70) and Asian populations (Cliff's delta = 0.71) at the state level. These results underscore the strong association between ethnic diversity and Democratic voting patterns.

The urban-rural divide is evident in the analysis, particularly in the differences in housing characteristics and population density. Democratic-leaning areas tend to have higher percentages of multi-structure housing, indicative of denser, urban environments. At the state level, Democratic states report 32.77% multi-structure housing, compared to 21.09% in Republican states, with statistical tests showing a large effect size (Cliff's delta = 0.72). At the county level, multi-structure housing remains a significant factor with a large effect size (Cliff's delta = 0.57). Additionally, Democratic areas are generally more populous and urbanized, while Republican regions are characterized by lower population density and more rural settings. These findings align with the broader trend of urban areas favouring Democratic candidates and rural areas leaning Republican.

Economic activities, particularly in sectors like retail, food services, and private employment, also distinguish between Democratic and Republican regions, especially at county level. Democratic-leaning counties show significantly higher accommodation and food service sales, retail sales, and private non-farm employment, with medium to large effect sizes ranging from 0.43 to 0.50. Additionally, the presence of minority-owned businesses, such as Asian and Black-owned firms, is more prominent in Democratic regions, further reflecting the demographic and economic diversity associated with Democratic voting patterns. These economic indicators highlight the differing industrial bases and economic structures that influence political preferences across the U.S.

3.2 Correlation Analysis

Table 4. *Top 10 Socio-economic Features with Highest Kendall's Tau Correlation Coefficient with Republican Vote Share at State Level*

Variable	Correlation Coefficient
Median house values ($)	-0.62
Asian population (%)	-0.59

Foreign born persons (%)	-0.57
Multi-structure housing (%)	-0.55
Bachelor's degree or higher (%)	-0.54
Asian-owned firms (%)	-0.53
Women-owned firms (%)	-0.51
Income per capita ($)	-0.51
Median household income ($)	-0.48
Mean travel time to work (min)	-0.47

Table 5 *Top 10 Socio-economic Features with Highest Kendall's Tau Correlation Coefficient with Republican Vote Share at County Level*

Variable	Correlation Coefficient
Multi-structure housing (%)	-0.40
Asian population (%)	-0.36
Accommodation and food service sales ($1,000)	-0.35
Building permits	-0.34
Private non-farm establishments	-0.34
Private non-farm employments	-0.34
Housing units	-0.33
Black-owned businesses (%)	-0.33
Retail sales ($)	-0.33
Population, 2010	-0.32

Table 6 *Selected Variable Pairs that Exhibit High Correlation Coefficient at State Level ($\tau > 0.5$ or $\tau < -0.5$)*

Variable 1	Variable 2	Correlation Coefficient
Population and Economic Activities		
Population	*Housing units*	0.97
Population	*Total number of firms*	0.91
Population	*Retail sales ($1,000)*	0.95
Population	*Manufacturers' shipments ($1,000)*	0.73
Population	*Wholesale trade ($1,000)*	0.82
Racial/Ethnic Composition and Business Ownership		
Black population (%)	*Black-owned businesses (%)*	0.88
Hispanic population (%)	*Hispanic-owned businesses (%)*	0.66
Asian population (%)	*Asian-owned businesses (%)*	0.77
Education, Income, and Housing Values		
Bachelor's degree or higher	*Median value of housing*	0.63

Bachelor's degree or higher	*Per capita income ($)*	0.67
Bachelor's degree or higher	*Median household income $*	0.65

The correlation analysis using Kendall's Tau provides important insights into the relationships between socio-economic indicators and the share of votes received by the Republican Party during the 2016 U.S. Presidential election at both state and county levels. Overall, the results indicate a consistent pattern where variables associated with wealth, diversity, and urbanization negatively correlate with Republican vote share.

At the state level, the strongest negative correlations with Republican vote share are observed in variables reflecting housing value and demographic diversity. Median house value (τ = -0.62) and Asian population percentage (τ = -0.59) show the strongest negative correlations, indicating that states with higher property values and larger Asian populations tend to favour Democratic candidates. Similarly, the percentage of foreign-born persons (τ = -0.57), multi-structure housing (τ = -0.55), and higher education attainment (τ = -0.54) are also strongly negatively correlated with Republican vote share. These findings align with broader socio-political trends where wealthier, more diverse, and more urbanized areas tend to lean Democratic. Additionally, variables reflecting business ownership, such as Asian-owned firms (τ = -0.53) and women-owned firms (τ = -0.51), also show significant negative correlations with Republican vote share at the state level. These correlations highlight the association between economic diversity and Democratic preferences. Other indicators like income per capita (τ = -0.51) and median household income (τ = -0.48) further reinforce the connection between higher economic status and Democratic voting patterns.

At the county level, while the overall correlations are weaker compared to the state level, similar patterns emerge. The most significant negative correlation is observed with multi-structure housing (τ = -0.40), reflecting the urban-rural divide where urbanized areas are more likely to support Democratic candidates. The percentage of Asian population (τ = -0.36) also shows a notable negative correlation, consistent with state-level findings. Economic indicators such as accommodation and food service sales (-0.35), private non-farm establishments (τ = -0.34), and retail sales (τ = -0.33) further underscore the association between more economically developed areas and Democratic voting preferences.

The correlation analysis also reveals intricate relationships between various socio-economic indicators in the United States. A strong positive correlation exists between population size and various economic activities, suggesting that more populous areas foster diverse and vibrant economies. Concurrently, the data underscores a strong association between an area's racial and ethnic composition and its business ownership patterns, particularly evident in the high correlations between minority populations and minority-owned businesses. Furthermore, the analysis reveals a positive association between educational attainment and economic prosperity, with higher education levels correlating positively with housing values and per capita income.

3.3 State Level Modelling

3.3.1 Logistic Regression

Table 7. *Summary of Final Logistic Regression Model with Lowest AIC*

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-1.685	1.333	-1.264	0.206	
Median house values	-12.070	5.494	-2.197	< 0.05	*
Women-owned firms %	-5.326	2.674	-1.992	< 0.05	*
Persons per household	4.363	2.204	1.98	< 0.05	*

Table 8. *Summary of Final Logistic Regression Model with Highest Accuracy*

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-0.193	0.787	-0.246	0.806	
Bachelor's degree or higher (%)	-3.196	1.836	-1.741	< 0.1	·
Persons under 5 years (%)	2.421	1.143	2.119	< 0.05	*
Asian population (%)	-6.492	2.596	-2.501	< 0.05	*
Population, 2010	2.165	1.134	1.91	< 0.1	·

Table 9. *Model Performance Metrics*

	Accuracy	Precision	Recall	F1-score	AUC
AIC-optimised model	0.941	0.950	0.905	0.927	0.936
Accuracy-optimised model	0.961	0.952	0.952	0.952	0.960

The state-level statistical modelling using Logistic Regression provides valuable insights into the socio-economic factors influencing voting patterns in the 2016 U.S. Presidential election. Two models are presented: one optimized for the lowest Akaike Information Criterion (AIC) and another for the highest accuracy. Both models demonstrate good predictive performance, as evidenced by their high accuracy, precision, recall, F1-scores, and AUC values. The AIC-based model achieves an accuracy of 0.941, while the accuracy-optimized model slightly outperforms it with an accuracy of 0.961. The high AUC values (0.936 and 0.960 respectively) indicate that both models have strong discriminative power in distinguishing between Republican and Democratic-leaning states.

The AIC-based model, which balances model fit and complexity, identifies three significant predictors. Median house values show a strong negative association with Republican voting ($\beta = -12.070$, $p < 0.05$), suggesting that states with higher property values were less likely to support the Republican candidate. The percentage of women-owned firms also demonstrates a negative relationship ($\beta = -5.326$, $p < 0.05$), indicating that states with higher rates of female entrepreneurship tended to lean Democratic. Interestingly, the number of persons per household exhibits a positive association ($\beta = 4.363$, $p < 0.05$), implying that states with larger household sizes were more inclined to vote Republican.

The accuracy-optimized model reveals a slightly different set of predictors. The percentage of the population with a bachelor's degree or higher shows a negative association with Republican voting ($\beta = -3.196$, $p < 0.1$), albeit with lower statistical significance. This aligns with the general observation that higher education levels correlate with Democratic voting tendencies. The percentage of persons under 5 years old has a positive relationship with Republican voting ($\beta = 2.421$, $p < 0.05$), suggesting that

states with younger populations were more likely to support the Republican candidate. Asian population percentage shows a strong negative association ($\beta = -6.492$, $p < 0.05$), indicating that states with larger Asian communities tended to vote Democratic. Lastly, population, 2010 has a positive relationship with Republican voting ($\beta = 2.165$, $p < 0.1$), though with lower statistical significance.

These results underscore the complex interplay of demographic, economic, and social factors in shaping state-level voting patterns. They highlight the importance of housing values, entrepreneurship demographics, household composition, education levels, age distribution, and ethnic diversity in predicting electoral outcomes.

3.3.2 Random Forest

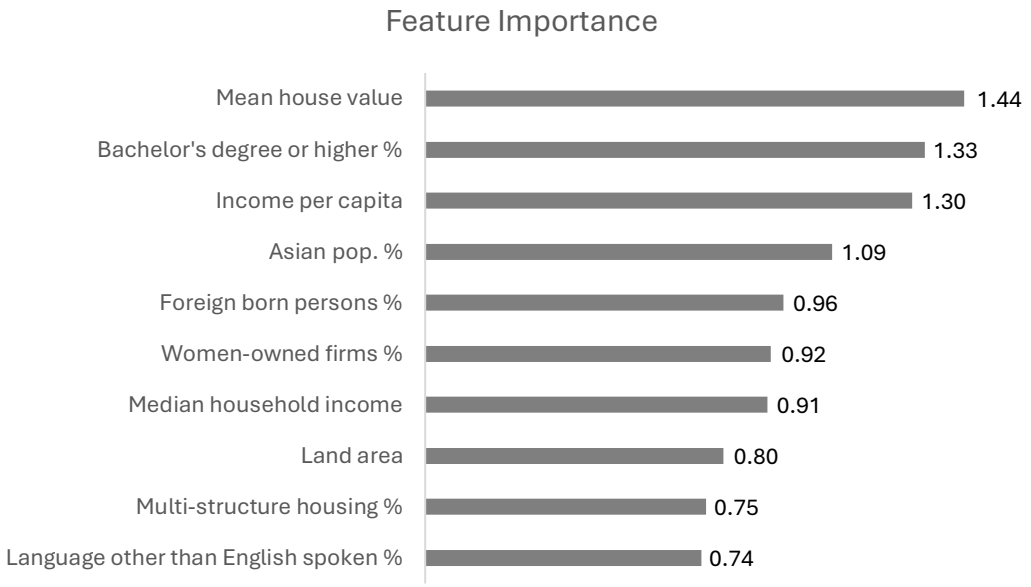

Figure 1. *Top 10 Variables with Highest Feature Importance Scores from Random Forest Classification Model*

Table 10 *Random Forest Classification Model Performance*

Accuracy	Precision	Recall	F1 Score	AUC
0.922	0.947	0.857	0.9	0.912

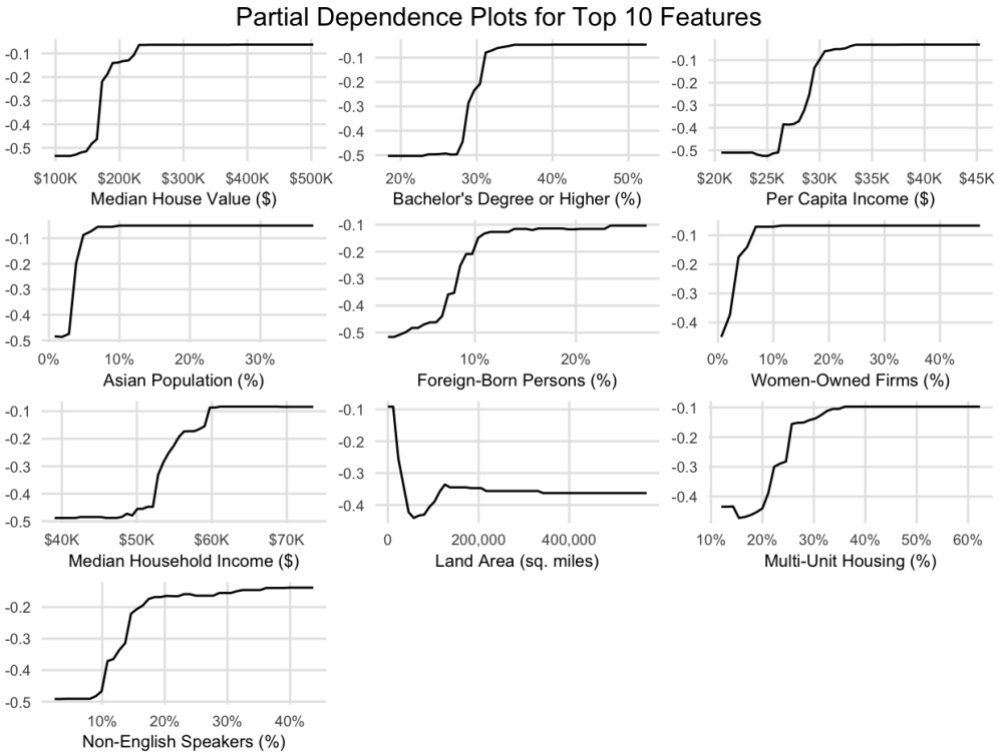

Figure 2. *Partial Dependence Plots for Top 10 Features.* The y-axis represents the relative likelihood of a Republican victory, with lower values indicating a higher probability of a Republican win in the election.

The results of the Random Forest model analysis provide valuable insights into the socio-economic factors influencing voting patterns in the 2016 U.S. Presidential election. The feature importance graph (**Figure 1**) highlights that economic and educational factors were the most significant predictors of political preferences. Among these factors, mean house value stands out as the top predictor, followed closely by the percentage of the population with a bachelor's degree or higher, and per capita income. These findings indicate that higher property values and educational attainment are strongly associated with a preference for Democratic candidates. Furthermore, demographic factors, such as the percentage of the Asian population and foreign-born persons, also rank highly in importance, underscoring the role of diversity in voting behaviour. Additionally, the presence of women-owned firms, median household income, and land area emerge as significant factors, reflecting the broader socio-economic context of these regions.

The partial dependence plots (**Figure 2**) are used to show the relationship between a feature and the predicted outcome while keeping other features constant, which helps to understand the marginal effect of a feature on the prediction. The plot for median house value reveals a sharp decline in the likelihood of a Republican vote as house values increase from approximately $100,000 to $200,000, after which the effect levels off. This suggests that wealthier areas with higher property values are more inclined to favour Democratic candidates. Similarly, the percentage of bachelor's degree holders shows a steep decline in Republican voting likelihood as educational attainment increases from around 20% to 30%, reinforcing the link between higher education levels and Democratic preferences. The influence of per capita income follows a similar trend, with higher income levels leading to a reduced likelihood of Republican voting. The demographic factors, including Asian population percentage and foreign-born persons, exhibit significant effects on voting patterns, with both showing substantial declines in Republican voting likelihood as these percentages increase. This indicates that even small

increases in ethnic diversity can have a pronounced impact on political preferences. Additional factors, such as multi-structure housing and the percentage of non-English speakers, further emphasize the urban-rural divide and the role of diversity in political alignment.

The Random Forest model's performance, as summarized in **Table 8**, confirms its robustness, achieving high accuracy (0.922), precision (0.947), and a notable AUC score of 0.912. These metrics indicate that the model performs well in distinguishing between Republican and Democratic-leaning regions. The consistency of key predictors across both feature importance rankings and partial dependence plots suggests that socio-economic factors such as housing values, educational attainment, and demographic diversity are reliable indicators of voting behaviour in the 2016 election.

3.4 County Level Modelling

3.4.1 Beta Regression with Stepwise Selection

Table 11 *Summary of Final Beta Regression Model by Stepwise Selection*

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	0.555	0.0087	63.776	<0.001	***
Multi-structure Housing %	-0.259	0.0127	-20.386	<0.001	***
White population percentage %	0.350	0.00973	35.933	<0.001	***
Median house value $	-0.201	0.0122	-16.489	<0.001	***
Asian-owned firms %	-0.159	0.0125	-12.794	<0.001	***
Population percent change %	0.117	0.0107	10.981	<0.001	***
Language other than English spoken at home %	-0.045	0.0151	-2.985	<0.01	**
Living in same house 1 year & over %	-0.035	0.0100	-3.534	<0.001	***
High school graduate or higher %	-0.121	0.0121	-9.951	<0.001	***

The final model, selected through a stepwise regression process with a VIF constraint to minimize multicollinearity, identifies several key predictors significantly associated with the fraction of votes received by the Republican Party. The model demonstrates a strong overall fit, as evidenced by a pseudo R-squared value of 0.6456. Additionally, the model performs well on the test set, yielding a mean squared error (MSE) of 0.0085, a mean absolute error (MAE) of 0.0686, and an R-squared of 0.6539. These performance metrics underscore the model’s capability to capture the relationship between the selected socio-economic and demographic variables and voting patterns.

Among the key predictors, the percentage of housing units in multi-unit structures shows a significant negative association with Republican vote share ($\beta = -0.259$). Conversely, the percentage of white population is strongly and positively related to Republican vote share ($\beta = 0.345$). The model also highlights the negative association between median home values and Republican vote share ($\beta = -0.201$), suggesting that wealthier, urbanized areas with higher property values are less likely to vote Republican. The percentage of Hispanic-owned businesses emerges as another significant negative predictor ($\beta = -0.159$), indicating that areas with a higher proportion of Hispanic entrepreneurs tend to vote Democratic.

Other important predictors include population growth, which has a positive effect on Republican vote share (β = 0.117), suggesting that rapidly growing suburban or exurban areas may lean towards Republican. Linguistic diversity, measured by the percentage of residents speaking a language other than English at home, is negatively associated with Republican vote share (β = -0.0451), supporting the notion that more diverse communities are more likely to favour Democratic candidates. Additionally, residential stability and high school population percentage are both negatively associated with Republican vote share.

3.4.2 Beta Regression with Principal Components

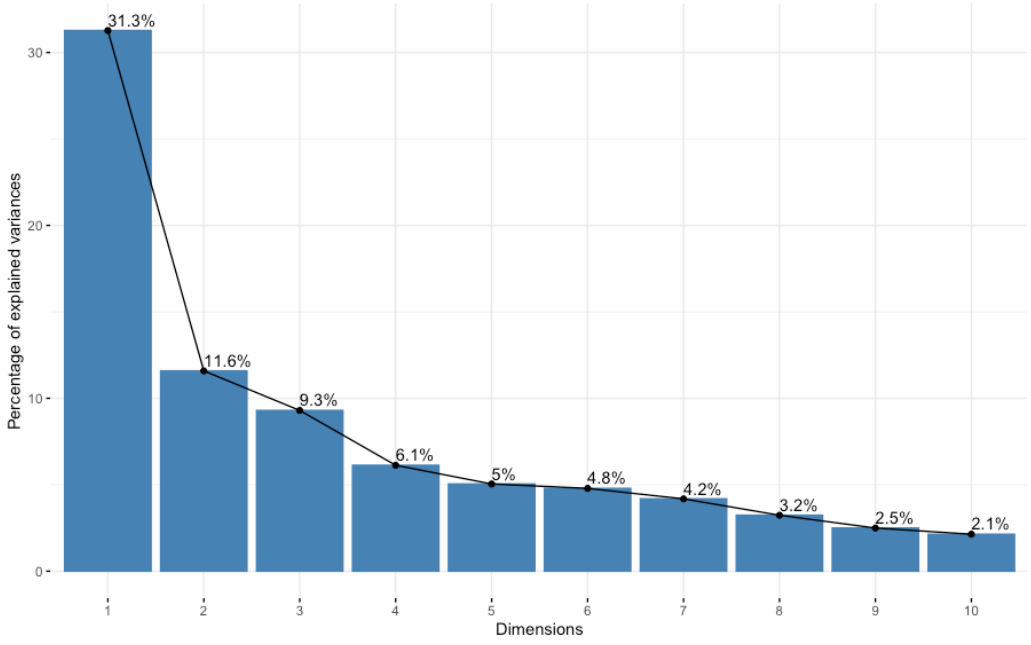

Figure 3 *Scree Plot of Principal Component Analysis (PCA) for County-Level Socio-Economic Data.* The scree plot illustrates the percentage of explained variance by each principal component. The first principal component accounts for 31.3% of the variance. The first 10 components account for over 80% of the variance.

Table 12 *Summary of Beta Regression with Principal Components*

	Estimate	Std. Error	z value	p-value	Significance
(Intercept)	0.5470	0.0095	57.735	< 0.001	***
PC1	-0.0938	0.0025	-37.6	< 0.001	***
PC2	-0.0611	0.0040	-15.376	< 0.001	***
PC3	-0.0785	0.0043	-18.157	< 0.001	***
PC4	0.1446	0.0054	26.73	< 0.001	***
PC5	0.0855	0.0058	14.766	< 0.001	***
PC6	-0.0444	0.0064	-6.897	< 0.001	***
PC7	0.0227	0.0064	3.542	< 0.001	***
PC8	-0.1694	0.0079	-21.371	< 0.001	***
PC9	0.0964	0.0086	11.197	< 0.001	***
PC10	0.0053	0.0109	0.483	0.629	

Table 13 *Top 5 Features with Highest Absolute Loadings for Principal Components 1-4*

Rank	PC1	PC2	PC3	PC4
1	Total housing units (0.2400)	White alone, not Hispanic or Latino % (-0.3377)	Median household income $ (0.2951)	Black population % (-0.3585)
2	Population, 2014 estimate (0.2390)	High school graduate or higher % (-0.2823)	Bachelor's degree or higher % (0.2603)	White population % (0.3468)
3	Retail sales $ (0.2388)	Persons below poverty level % (0.2785)	Per capita income $ (0.2581)	Hispanic or Latino population % (0.3223)
4	Population, 2010 (0.2385)	Persons under 5 yr % (0.2639)	Median housing values $ (0.2528)	Black-owned firms % (-0.2963)
5	Total number of firms (0.2358)	White alone % (-0.2554)	High school graduate or higher % (0.1982)	Non-English speaker % (0.2818)

This model shows good predictive power, as evidenced by a pseudo R-squared of 0.5616 in the training set and an R-squared of 0.6218 in the test set. The mean squared error (MSE) of 0.009264 and mean absolute error (MAE) of 0.07540 on the test set further confirm the model's predictive accuracy. These metrics indicate that the PCR approach effectively captures the underlying patterns in the data while avoiding overfitting.

Nine out of ten principal components included in the model show statistical significance ($p < 0.001$), suggesting that they capture meaningful variations in the data. PC1, primarily captures variables related to population size, economic activity, and urbanization. The top contributing variables in this component include households, housing units, population estimates, retail sales, and private nonfarm establishments. The clustering of these variables suggests that areas with larger populations and higher levels of economic activity are characterized by distinct voting pattern, likely influenced by urbanization and the economic infrastructure.

PC3 also demonstrates a negative association with Republican voting ($\beta = -0.0785$, $p < 0.001$). This component is heavily loaded on income and education variables, with the highest loadings on median household income (0.2951), percentage of population with a bachelor's degree or higher (0.2603), and per capita income (0.2581). The negative coefficient indicates that counties with higher income levels and educational attainment tend to favour Democratic candidates.

PC4 exhibits the strongest positive association with Republican vote share ($\beta = 0.1446$, $p < 0.001$). This component presents a nuanced picture of racial and ethnic composition, with negative loadings on Black or African American population percentage (-0.3585) and positive loadings on White (0.3468) and Hispanic or Latino (0.3224) percentages. The positive coefficient suggests that counties with higher proportions of White and Hispanic residents relative to Black residents tend to lean more Republican.

3.4.3 Random Forest Regression

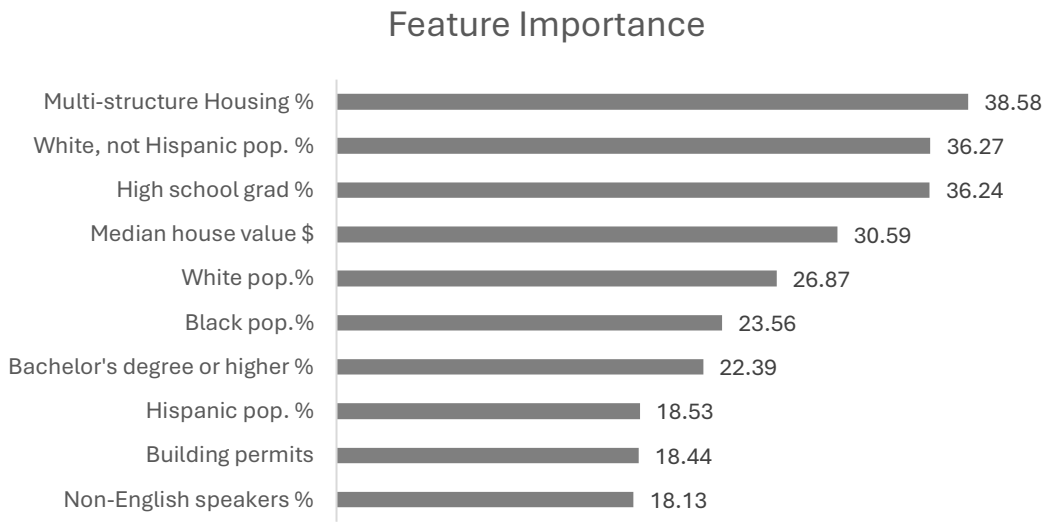

Figure 4 *Top 10 Variables with Highest Feature Importance Scores from Random Forest Regression Model*

The random forest regression model developed to predict the fraction of Republican votes demonstrates strong predictive capability and highlights important socio-economic factors driving voting behaviour in the 2016 U.S. Presidential election. The model's performance metrics are robust, with an R-squared value of 0.7856, indicating that approximately 78.56% of the variance in Republican vote share across counties is explained by the model. Additionally, the model achieves a low Mean Squared Error (MSE) of 0.0055 and a Root Mean Squared Error (RMSE) of 0.0743 on the test set, further underscoring its accuracy and reliability.

The feature importance plots (**Figure 4**) reveal that key variables influencing voting behaviour include housing characteristics, racial demographics, and economic indicators. Specifically, the percentage of housing units in multi-unit structures and the percentage of White non-Hispanic population are the most critical predictors, highlighting the significance of urban-rural divides and racial composition in shaping political preferences. Educational attainment, represented by variables such as the percentage of the population with a bachelor's degree or higher, also plays a prominent role, consistent with the broader narrative that areas with higher education levels tend to lean Democratic. Economic factors, such as median home value and specific industry indicators, further contribute to the model's predictive power, reflecting the complex interplay between socio-economic status and political alignment.

These results emphasize the multifaceted nature of electoral behaviour, where a combination of demographic, educational, and economic factors interact to influence voting patterns. The strong performance of the random forest model in capturing these dynamics provides valuable insights for understanding the 2016 election and offers a robust foundation for future predictive analyses of voting behaviour.

4. Discussion and Conclusion

This study sought to investigate the intricate relationships between socio-economic factors and voting patterns in the 2016 U.S. Presidential election at both county and state levels. Through comprehensive statistical analyses and predictive modelling, we have gained valuable insights into the complex dynamics of U.S. electoral behaviour. Our findings address the three primary research questions and reveal both consistencies and differences in the factors influencing voting preferences across different geographic scales.

This analysis consistently demonstrates that several key socio-economic and demographic factors are strongly associated with voting preferences at the county level. Housing characteristics emerge as a significant predictor, with counties exhibiting higher percentages of multi-unit housing structures and higher median house values tending to favour Democratic candidates. This is evidenced by the strong negative correlation between multi-unit structure housing and Republican vote share and the significant negative coefficient in the beta regression model. Educational attainment also plays a crucial role, as areas with higher percentages of population holding a bachelor's degree or higher are more likely to support Democratic candidates. This relationship is supported by the negative coefficient in the beta regression model for high school graduates or higher and the strong loading on PC3 in the principal component regression.

Racial and ethnic composition proves to be another critical factor in shaping voting preferences at the county level. Counties with greater racial diversity, particularly those with higher percentages of Asian and foreign-born populations, tend to lean Democratic. Conversely, counties with higher percentages of white population are more likely to support Republican candidates. This pattern is reflected in the positive coefficient for white population percentage in the beta regression model and the strong loadings on PC4 in the principal component regression.

Economic indicators further contribute to the complex tapestry of factors influencing voting behaviour. Areas with higher median household incomes, higher per capita incomes, and more diverse economic activities (e.g., higher accommodation and food service sales) tend to favour Democratic candidates. This trend is evident in the negative correlations between these factors and Republican vote share, as well as their strong loadings on PC1 and PC3 in the principal component regression.

At the state level, our analysis reveals that many of the same factors identified at the county level also influence voting preferences, often with more pronounced effects. Housing values emerge as a particularly strong predictor, with states exhibiting higher median house values significantly more likely to support Democratic candidates. This is evidenced by the strong negative correlation and the large effect size in statistical tests. Educational attainment continues to play a crucial role at the state level, with states boasting higher percentages of residents holding a bachelor's degree or higher tending to favour Democratic candidates. This relationship is supported by the large effect size in statistical tests and the negative coefficient in the logistic regression model.

Income levels and racial and ethnic diversity also emerge as significant predictors of voting preferences at the state level. Higher income per capita and median household income are associated with Democratic voting preferences, as shown by the strong negative

correlations with Republican vote share. States with higher percentages of Asian populations, foreign-born persons, and non-English speakers are more likely to support Democratic candidates, as evidenced by the strong negative correlations and large effect sizes in statistical tests. Additionally, the demographics of business ownership play a role, with states having higher percentages of women-owned and minority-owned businesses tending to lean Democratic, as indicated by the negative coefficients in the logistic regression models.

While many of the same factors are significant at both county and state levels, there are notable differences in the strength and manifestation of these relationships. Overall, the associations between socio-economic factors and voting preferences generally appear stronger at the state level compared to the county level. This may be due to the aggregation effect at the state level, which can amplify trends that might be more varied or nuanced at the county level. The granularity of insights provided by county-level analysis offers a more detailed view of local dynamics, capturing variations that might be obscured at the state level. For instance, the influence of specific economic sectors (e.g., accommodation and food services, retail sales) is more apparent in county-level models. The relative importance of demographic factors also differs between levels. At the state level, factors like the percentage of Asian population and foreign-born persons are among the top predictors, whereas at the county level, the percentage of White non-Hispanic population emerges as a key factor. Economic indicators show importance at both levels, but their specific manifestations differ. State-level analysis highlights broader economic measures like median household income, while county-level analysis captures more localized economic dynamics, such as the presence of specific types of businesses.

Our models demonstrate strong predictive power at both state and county levels, indicating that socio-economic factors can indeed be used to predict election results with considerable accuracy. At the state level, the logistic regression models achieved high accuracy (up to 96.1%) and AUC values (up to 0.960), indicating strong discriminative power in distinguishing between Republican and Democratic-leaning states. The random forest classifier similarly showed robust performance with an accuracy of 92.2% and an AUC of 0.912. At the county level, the beta regression model explained approximately 64.56% of the variance in Republican vote share, while the random forest regression model achieved an impressive R-squared value of 0.7856. The principal component regression also showed good predictive power, with an R-squared of 0.6218 on the test set. These results suggest that socio-economic factors can serve as reliable predictors of election outcomes at both geographic levels, with slightly higher accuracy achieved at the state level.

The findings of this study have several important implications for political campaigns, policy development, voter outreach, and media analysis. Political campaigns can use these insights to tailor their messaging and resource allocation based on the socio-economic profiles of different regions. Policymakers can leverage this information to better understand the diverse needs and preferences of different communities, potentially leading to more targeted and effective policies. Organizations focused on increasing voter participation can develop more effective outreach strategies tailored to the socio-economic characteristics of different areas. Furthermore, these results can inform more media coverage of elections, moving beyond simple red state/blue state narratives to consider the complex socio-economic factors influencing voting patterns.

Despite its comprehensive nature, this study has several limitations that should be acknowledged. The study's focus on the 2016 election introduces temporal specificity, and

the relationships observed may not hold for other election cycles. There may be other important variables not captured in the dataset, such as local political histories or specific policy issues. The assumption of uniformity within counties may not reflect intra-county variations that could influence voting patterns. Additionally, the study's focus on a binary classification between Republican and Democratic voting, excluding third-party candidates, may overlook important nuances in voter preferences, especially in closely contested regions.

To address these limitations and further advance our understanding of voting behaviour, future research could incorporate individual-level data, combining aggregate data with individual-level surveys or voting records to provide more precise insights. Conducting longitudinal studies across multiple election cycles could reveal how the relationship between socio-economic factors and voting preferences evolves over time. Exploring finer geographic units, such as precinct-level data, could capture more localized voting patterns and help address the ecological fallacy. Investigating additional variables, including media consumption patterns, social media influence, and specific policy preferences, could enhance the predictive power and explanatory depth of the models.

In conclusion, this study provides a comprehensive analysis of the socio-economic factors influencing voting patterns in the 2016 U.S. Presidential election. By examining these relationships at both county and state levels, we gain understanding of the complex dynamics shaping American electoral politics. While the models demonstrate strong predictive power, they also highlight the multifaceted nature of voting behaviour and the importance of considering diverse factors in political analysis.

5. Reference

Cinyc. (2019, December 27). *Alaska Presidential Results by County Equivalent, 1960-2016*. Cinyc Maps. https://cinycmaps.com/index.php

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Ferrari, S., & Cribari-Neto, F. (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, *31*(7), 799–815. https://doi.org/10.1080/0266476042000214501

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Kahane, L. H. (2020). Determinants of County-Level Voting Patterns In the 2012 and 2016 Presidential Elections. *Applied Economics*, *52*(33), 3574–3587. https://doi.org/10.1080/00036846.2020.1713985

Li, A., & Barber, R. F. (2019). Multiple Testing with the Structure-Adaptive Benjamini–Hochberg Algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *81*(1), 45–74. https://doi.org/10.1111/rssb.12298

Maggio, C. (2021). Demographic change and the 2016 presidential election. *Social Science Research*, *95*, 102459. https://doi.org/10.1016/J.SSRESEARCH.2020.102459

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Nichols, K., & Holmes, A. (2007). Non-parametric procedures. In *Statistical Parametric Mapping* (pp. 253–272). Elsevier. https://doi.org/10.1016/B978-012372560-8/50021-8

OpenAI. (2023). *ChatGPT* (Feb 13 version) [Large language model]. https://chat.openai.com

Puka, L. (2011). Kendall's Tau. In *International Encyclopedia of Statistical Science* (pp. 713–715). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_324

RIFFENBURGH, R. H. (2006). Tests on the Distribution Shape of Continuous Data. In *Statistics in Medicine* (pp. 369–386). Elsevier. https://doi.org/10.1016/B978-012088770-5/50060-5

Romano, J., Kromrey, J., Coraggio, J., Skowronek, J., & Devine, L. (2006). Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices? *Annual Meeting of the Southern Association for Institutional Research*, 1–51.

Smith, G. W., & Young, J. J. (2016). A MULTIVARIATE ANALYSIS OF THE 2016 COUNTY-LEVEL PRESIDENTIAL VOTE AND TURNOUT 1 A Multivariate Analysis of the 2016 County-Level Presidential Vote and Turnout. *Advance*. https://doi.org/DOI: 10.31124/advance.12751913.v1

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 12–18. https://doi.org/10.11613/BM.2014.003

WELCH, B. L. (1947). THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. *Biometrika*, *34*(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28