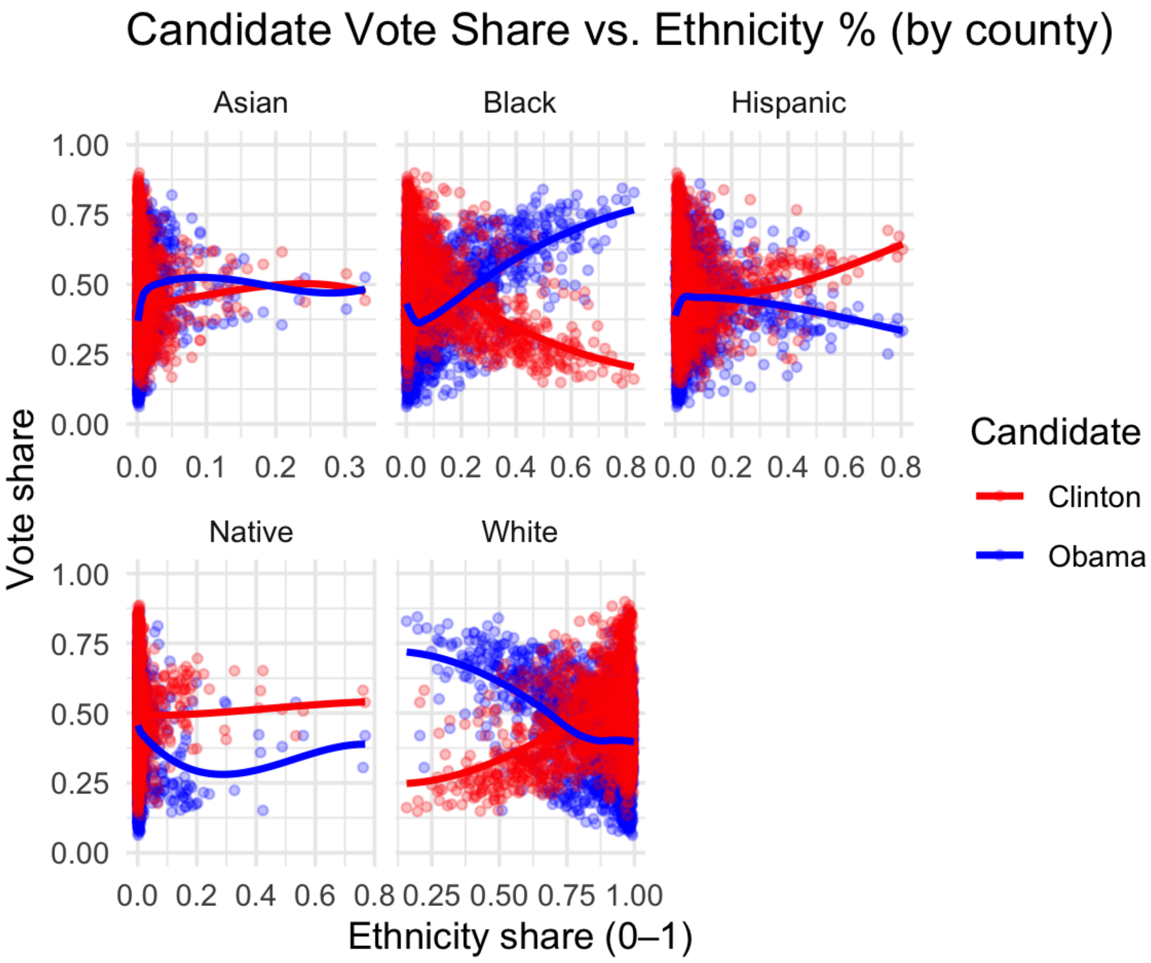
Case Study 3
Section A, Team #11

Ankit Chandekar, Jinghua He, Abigail Miller
Muhammad Memon, Nikki Reddy, Katherine Zhang

1.



The variables we picked to show a relation through visualization are ethnicity and vote share to attempt to find a correlation between whether ethnic composition (percentage per county) is associated with support for a particular candidate (Obama vs. Clinton). We plotted the county share of each ethnicity (scaled from 0-1) against the vote share against each candidate (Obama = blue, Clinton = red) via scatter plot and LOESS. What the visualization shows is a strong positive slope for Obama and a negative slope for Clinton for Black, which means counties with larger Black populations tended to have a higher share of votes for Obama. Meanwhile, there is a positive slope for Clinton and a negative slope for Obama for White, as well as for Hispanic. For both Asian and Native Americans, there is no strong pattern or correlation that can be confidently stated based on the data visualization. The conclusion that can be drawn from this data visualization is that Obama's vote share rises with a higher percentage of Black citizens in a county, while Clinton's rises with a higher percentage of White and Hispanic citizens. Meanwhile, Asian and Native Americans show no strong or consistent slopes and therefore there is no strong correlation between ethnicity share and candidate support. However, while correlations and associations can be seen and drawn between these variables, this visualization does not necessarily show casual effects, and it's worth noting the very small Asian/Native American shares of many counties which produces heteroskedasticity (fan of points near zero).

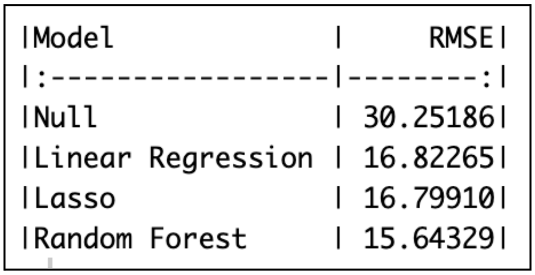
2.

The core task of this analysis is to predict Barack Obama's winning spread over Hillary Clinton across U.S. counties during the 2008 Democratic primaries with the best model. Obama's winning spread is a target variable that reflects how much Obama led Clinton in each county as a share of the total votes. To predict Obama's winning share of votes, demographic and economic indicators such as education, income, unemployment, race, age, etc. are used as predictors. We created 4 different models and compared their performance using k-fold cross-validation and root mean square error (RMSE).

Model Selection:

1. Null Linear Regression: Serves as the baseline model to benchmark all other approaches.
2. Linear Regression: Models a linear relationship between predictors & Obama_margin_percent
    - It provides interpretability which indicates how changes in demographic variables relate to Obama's vote margin.
3. Lasso Regression: Handles multicollinearity and performs automatic variable selection.
    - Many demographic variables are highly correlated.
    - Within each cross-validation loop, an inner cv.glmnet() procedure determines the optimal penalty parameter (λ) that minimizes prediction error.
4. Random Forest: It captures nonlinear effects and complex interactions automatically.
    - It fits hundreds of decision trees, each using random subsets of predictors (mtry = one-third of features).
    - Predictions are aggregated across trees to produce stable, low-variance estimates.
    - It tracks complex interactions like how race composition interacts with education or region.

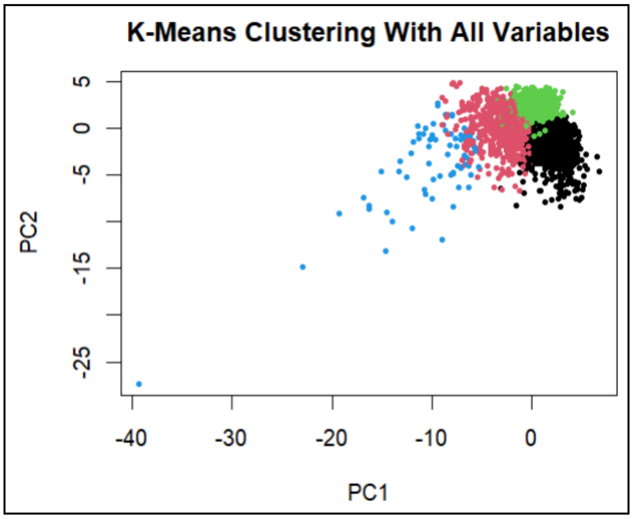
We applied a 16-fold cross-validation to evaluate the performance of each model and got the below RMSE values for each model:

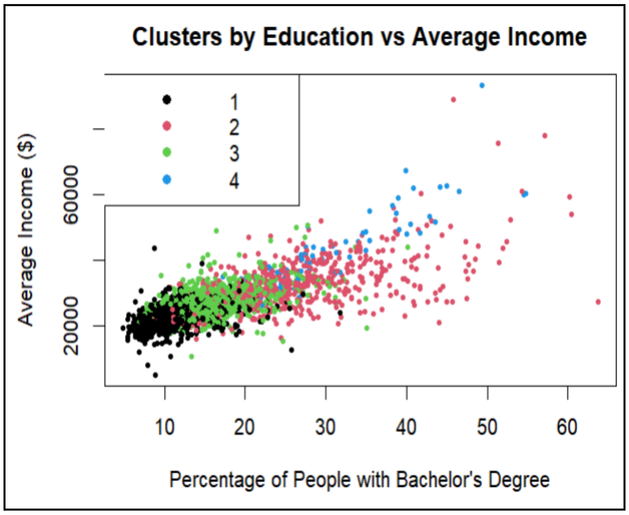


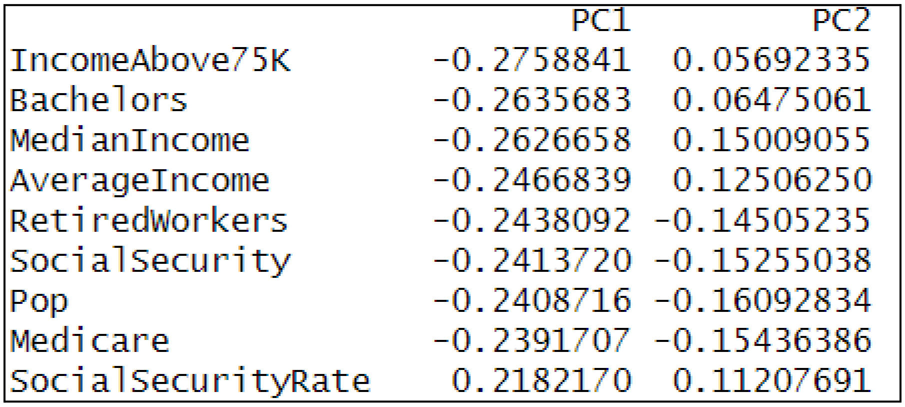
```
|Model            |     RMSE|
|:----------------|--------:|
|Null             | 30.25186|
|Linear Regression | 16.82265|
|Lasso            | 16.79910|
|Random Forest    | 15.64329|
```

Based on this, it is evident that the random forest model is the best choice in predicting Obama's winning spread of votes. The attached R file includes the code used to generate predictions from the model, while the accompanying CSV file provides the resulting predicted y-values.

3.







```
                          PC1          PC2
IncomeAbove75K      -0.2758841  0.05692335
Bachelors           -0.2635683  0.06475061
MedianIncome        -0.2626658  0.15009055
AverageIncome       -0.2466839  0.12506250
RetiredWorkers      -0.2438092 -0.14505235
SocialSecurity      -0.2413720 -0.15255038
Pop                 -0.2408716 -0.16092834
Medicare            -0.2391707 -0.15436386
SocialSecurityRate   0.2182170  0.11207691
```

I applied k-means clustering to the dataset, where k equals 4. The first graph implements PCA to compress the variables into two latent features so that the clusters can be graphed in a two dimensional space. This graph shows that the four clusters show some separation. When observing the weights of variables in the first two principal components, the top contributors are variables relating to socioeconomic status, such as poverty, income, and education. In order to have the clusters be more interpretable, I tried graphing many pairs of variables to see if any had strong correlations. The graph of the average income and percentage of people with bachelors degrees does correlate with the clusters. Cluster 1 is characterized by low education and low income. Cluster 3 is characterized by moderate income and education. Clusters 2 and 4 are highly educated and wealthy people, with cluster 4 ultimately being the most wealthy. This suggests that counties may differ in their responsiveness to campaign messages depending on socioeconomic status like education and income.

4.

Based on the linear models, a 5% increase in a county's Hispanic population will lead to a 0.58% point decrease in Obama’s winning spread and a 5% increase in the Black population will lead to a 4.32% point increase in Obama’s winning spread.

After keeping all the other demographic and geographic factors in control and with a 5% increase in the Hispanic population, there will be a 1.85% point increase in Obama's winning spread and when it comes to 5% increase in the Black population, there will be a 1.17% point decrease in Obama's winning spread.

The model run first is misleading as there was omitted variable bias present in it. The demographic variables were not taken into consideration. These factors include geographic state and potentially urban settings where Obama's support was relatively strong.

The latter model is correct because it includes the omitted variables removing the bias. This way the true isolated effect of each demographic is revealed which explains the drastic difference in the results between both models.

5.

Obama:

Solution: Targeting messages toward counties and states with high concentrations of young voters (under 35 years old), Black voters, and residents holding a bachelor's degree or higher would be most efficient for vote gains.

Why? When analyzing the county-level election results, it is evident that Obama's support was statistically higher in areas with larger percentages of the above demographics. For example, Johnson County, Iowa, where 59.4% of residents were under 35 and 47.6% had bachelor's degrees, gave Obama a significantly higher share of the vote compared to counties with older or less-educated populations

Resource Allocation: Campaign resource allocation (such as budget and volunteer time) should therefore focus on regions where shifts in turnout among these target segments are likely to be decisive. Apply predictive analytics to prioritize canvassing and advertising budgets in areas with the highest projected return on investment (ROI) for increased turnout or persuasion among key segments.

To communicate insights effectively, we can prepare visualizations such as bar charts and correlation graphs that highlight the link between voter demographics (age and education) and vote shares for each candidate.