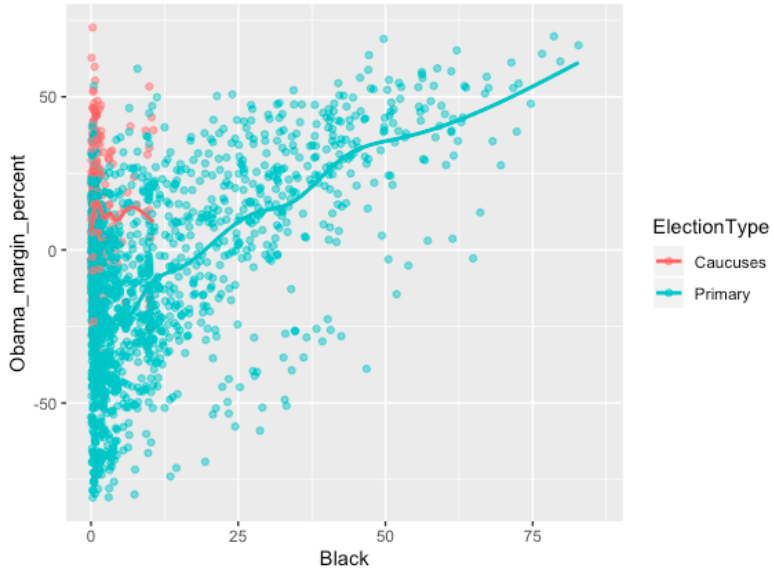
2008 Democratic Primaries (Clinton vs. Obama)

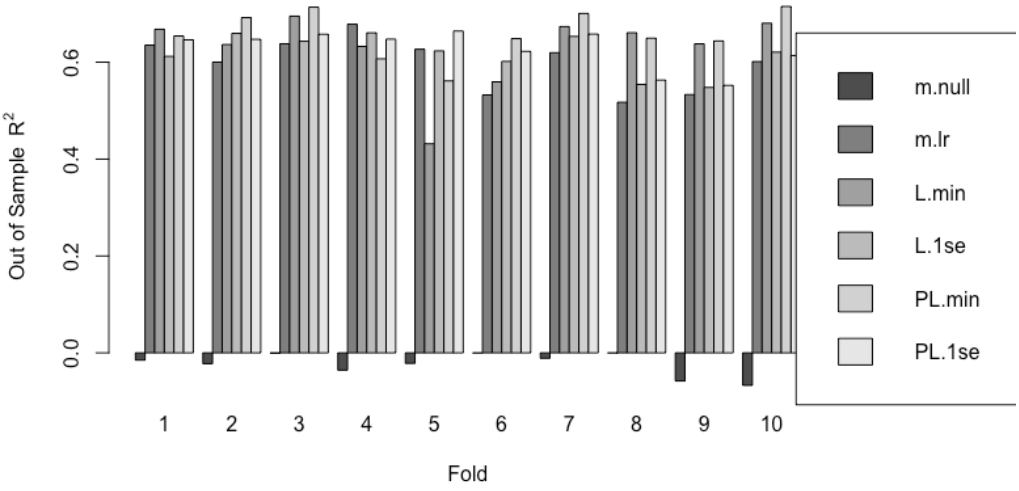
1. As a black presidential nominee, Barack Obama represents the largest minority of the country who are normally underrepresented in politics. We hypothesized whether or not a county's black population proportion had an effect on Obama's polling results.



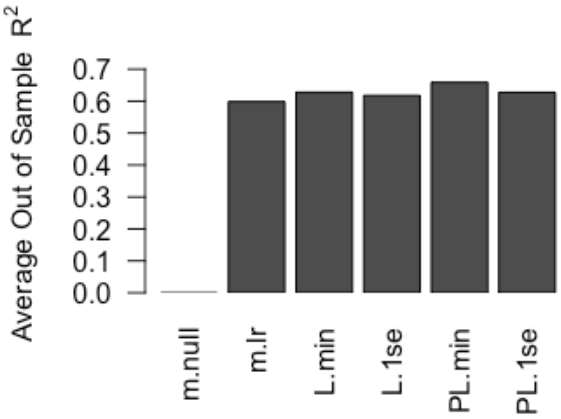
The plot clearly illustrates an upward trend of Obama favorability as the percentage of a county's black population increases. Obama especially won over counties where the majority of the population is black. We also added the dimension of 'ElectionType' to see if that played any significance. Obama received a majority of the votes in the Caucuses, as the trendline remains above 0. However, the caucus counties all appear to have black population proportions of less than 12.5%.

As a government built on the concept of representation, it's safe to say that America's black population leans favorably towards Obama, a black Presidential candidate. However, it appears that they are not as represented during the Caucuses.

2. The core objective is to create a model that can estimate Obama's winning spread (vote margin as a percentage of the total vote) given a county's demographic and voting data. The models in consideration were simple linear regression, regression with interactions using lasso, and linear regression with interactions using post-lasso. For both lasso and post lasso, we looked at two data-driven choices: the minimum and 1se choices. A null model was created as a baseline comparison as well. The out of sample R-squared for each model will be used to evaluate model performance. The results of the K-fold cross validation are displayed below:



Averaging across all folds, it appears that the post-lasso minimum choice merits the highest out of sample R-squared. Though the models seem to hover around the 0.6 mark, all the lasso and post lasso models perform just a little better than simple linear regression.



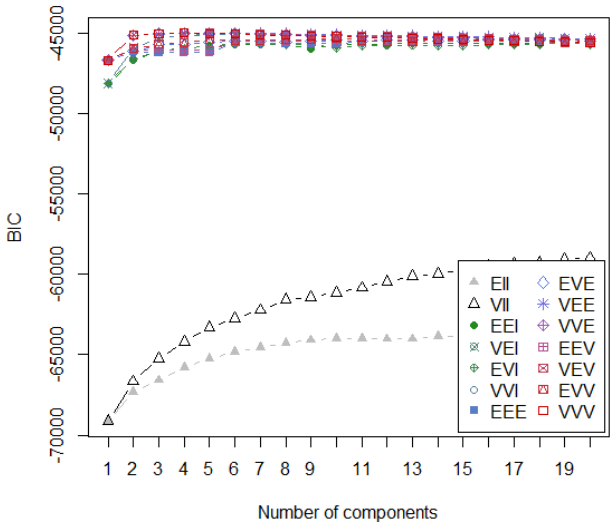
Predicting winning spread for the test sample data can be done with the following code:

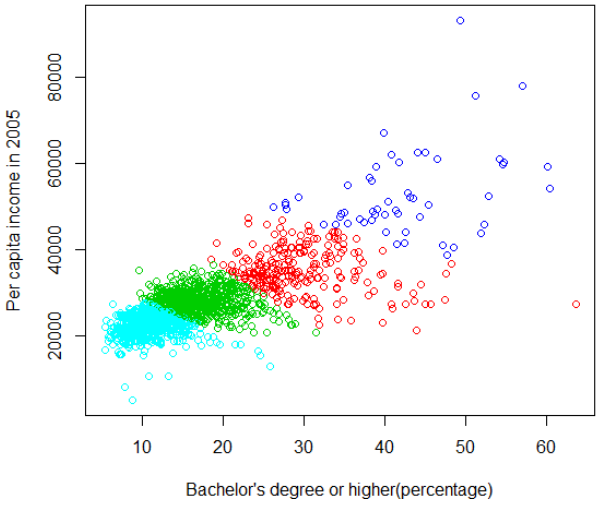
```
predlassomin<-
predict(rmin,newdata=data.frame(newMx),type="response")
```

The predicted values for Obama's winning spread are in the csv file attached to the report.

3. We used the mclust algorithm to determine the ideal number of clusters. Testing the number of clusters from 1 to 20, the algorithm recommends four as the most ideal number of clusters for this set of data.

   Having decided the number of clusters as four, we conducted k-means clustering to the sub dataset. According to the clustering, the data divided the data in terms of average income and percentage of Bachelor's degree.





Cluster	AverageIncome	Education
Teal(1)	22044.95(Low)	12.07962(Low)
Green(4)	27907.39(Medium Low)	16.94871(Medium Low)
Red(3)	36526.68(Medium High)	26.95462(Medium High)
Blue(2)	53941.24(High)	40.08039(High)

Interesting insights can be drawn from the clustering. Looking into different clusters, we are interested in which cluster supports Obama the most.

Teal: 0.6403509

Blue: 0.6078431

Red: 0.4987113
Green: 0.3812317

The result indicates that the cluster supporting Obama most is either a combination of low income and low education or high income and high education level, and the middle-level shows no strong support to Obama.

Teal: 9.557897
Blue: 9.417647
Red: 8.035573
Green: 15.15162

In the teal, blue, and red clusters, a similar percentage of the African-American people exist, while the green cluster shows that those with medium-low income and medium-low education level shows a higher proportion African-American people. It's interesting to see that the group with the highest percentage of African-American people support Obama the least which goes against our initial visualization. Perhaps when interacted with other variables, race itself is not as significant to winning spread. However, this would require further investigation.

4. **Hispanic Demographic:** Since 'county' is a factor variable which has over 1,000 attributes, we plan to use double selection model that hopefully can handle the large set of controls. Upon running the double selection model, the hispanic demographic appears to have no relation to the winning spread since the coefficient is not statistically significant with a p-value of 0.36. Because of this, we conclude that 5% increase in Hispanic demographic will not have a significant impact the winning spread.

   The simple model also suggests the hispanic demographic is statistically insignificant to the winning spread with an uncritical p-value of 0.1.

   Simple model

Estimate	Std. Error	t value	Pr(>|t|)
-0.11511237	0.06972701	-1.65090065	0.09893984

   Double selection

Estimate	Std. Error	t value	Pr(>|t|)
-0.07783028	0.08551319	-0.91015530	0.36286792

   Black Demographic: The double selection model suggested that the black demographic has a positive relation of a 1.86 percent winning spread increase per percentage of black population increase with an extremely small p-value less than 0.001. With a 5% increase of black demographic, the total winning spread increase is about 9.3%.

Using the simple linear regression model, it also revealed a positive relationship between the black demographic makeup and the winning spread of Obama. A 5% increase of black demographic will cause 4.3% increase of Obama's winning margin.

Simple model

Estimate	Std. Error	t value	Pr(>|t|)
8.631140e-01	4.204329e-02	2.052917e+01	5.123968e-84

Double selection:

Estimate	Std. Error	t value	Pr(>|t|)
1.862704e+00	1.336698e-01	1.393512e+01	6.633493e-42

5. For Barack Obama, it would be exceptionally helpful to identify which counties lean favorably for, moderately for, or away from your candidacy. Ideally, it would be optimal to focus more on winning over the moderate counties while keeping in touch with counties already in his favor. However, Obama's campaign team should further investigate any trends in demographic makeup and voter details of these on-the-fence counties. The campaign team could then customize their messages to reflect the people of such counties. These counties can be segmented using a cluster analysis to identify distinctions. It is crucial to ensure clear communication on both ends in order to effectively impact the voters' decisions: the people need to hear from Obama and vice versa.

   Obama's campaign team should identify the counties that fall into the previously aforementioned clusters from our k-means clustering analysis. More clustering analyses can be done to identify additional variables that could significantly explain winning spread. Since the weakest cluster, green, contained the highest percentage of black population per county, this could be an ideal cluster to target for Obama since he has proven to win over these types of counties before. However, he should adjust his message for these 'green cluster' counties to account for their general lifestyle levels.

   These insights should be communicated through clear visualizations identifying target counties with a predicted increase in winning spread for said counties.