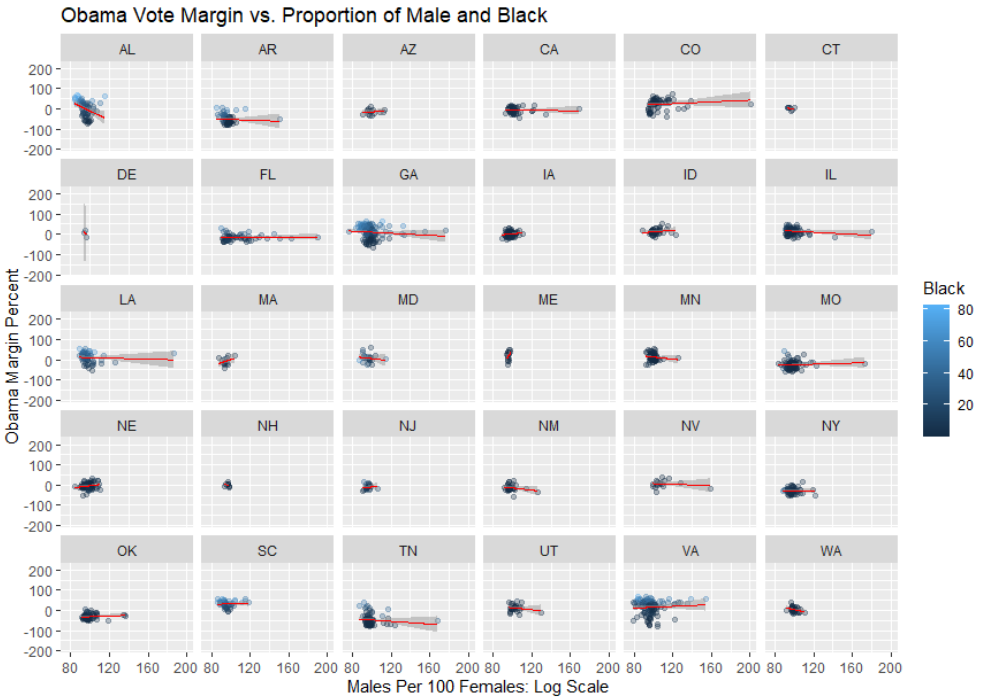
Section: B
Team: 2

Team Members: Nishant Bahl, Tigran Danielyan, Bridge Hu, Saiganesh Musthyala, Rama Sai Sundar Ryali

Question 1:



Question: Is there a linear relationship between Obama margin percent and proportion of males per 100 women? Between Obama margin percent and proportion of Balck population?

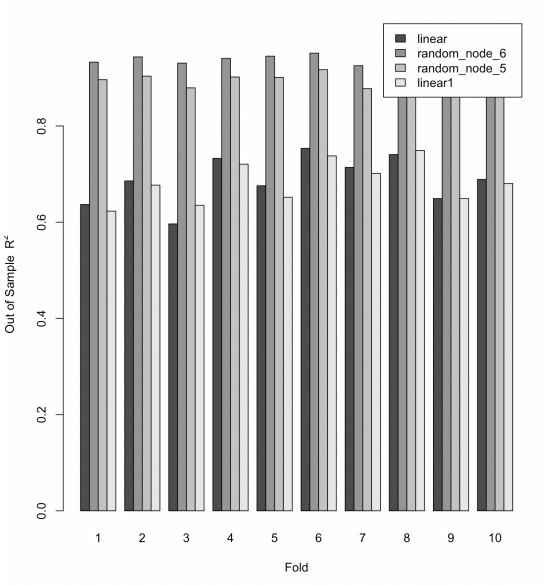
We have used faceting to try to capture state-to-state variations. For Males per 100 females, we have used a log scale. Finally, we have used geom_smoth with a linear model method to visualize the linear relationships. As we can see in many states we observe little to no obvious relationship between Obama margin percent and proportion of males per 100 women. At the same time, in some states, we can observe certain positive relation between Balck population proportion and Obama margin percent (i.e. AL, LA, TN, VA).

Question 2:

Core task: The general task is to predict the winning spread of Obama over Clinton as a percentage of the total vote. We have labeled that as Obama_margin_percent. This margin can be either negative or positive where a positive value indicates that Obama is winning in that county whereas a negative value indicates Clinton is winning.

Approach:

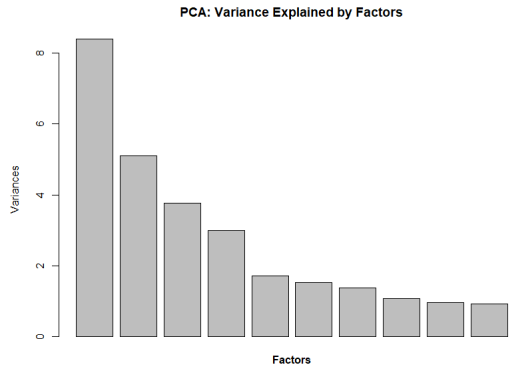
We started with adding dummy variables to explain the region and election type. We ran linear regression with all the variables, followed by linear regression with only significant variables. We then ran a random forest with node size as 5 followed by a random forest with node size 6.



We compared the models using OOS r2 as the evaluation metric since that would give us the best understanding of our accuracy in terms of our final predictions. Based on this evaluation we ran a k fold validation with the number of folds as 10 and found that the random forest model with node size 6 had the best OOS r2 of ~0.94 which made us choose this as our final model for further predictions.

NOTE: The R script has the code to generate the predictions and write them into a CSV. The CSV will also be uploaded while submitting the assignment.

Question 3:



PCA1: Seems to focus on counties with big middle class retired workers.

PCA2: This component is described by high poverty rate, high rate of disabilities, high unemployment and low average income, with a high proportion of black population and low proportion of white population.

PCA3: This component focuses on counties with a young (age below 35) and healthy population.

PCA4: Focuses on counties with an urban population with a high proportion of black people and a low proportion of other Hispanic, white and American Indian populations

Question 4:

(a) impact of changing Hispanic demographic:

We used the DML with a split sampling model as it is one of the most robust causal models. The model suggests that the Hispanic demographic has a statistically significant

impact with a low p-value of 0.0084 (<0.05) and an impact of -0.2328. So a 5% increase in the Hispanic population would impact the Obama winning spread by 5*(-0.232816267)= -1.1641 (it would decrease by 1.16%).

Estimate	Std. Error	t value	Pr(>|t|)
-0.232816267	0.088269097	-2.637573905	0.008424737

(b) impact of changing black demographic:
The DML with split sampling model suggests that the black demographic has a statistically significant impact with a low p-value of 6.488540e-37 (<0.05) and an impact of 1.225082. So a 5% increase in the Black population would impact the Obama winning spread by 5*(+1.225082)= +6.12451 (it would increase by 6.13%).

Estimate	Std. Error	t value	Pr(>|t|)
1.225082e+00	9.427260e-02	1.299510e+01	6.488540e-37

Question 5:

Candidate: Obama

Based on our unsupervised PCA analysis, we found that the group that explains the population the most is the retired middle-class segment, closely followed by the ones in poverty. Which made perfect sense since the nation just entered into a recession when the campaign was ongoing, and these should be the segments that Obama's campaign should target by putting more time and effort into the states that have a larger middle class and higher poverty level.

Our PCA and DML with split sampling model also show a clear relationship between Obama's winning margin and the population of the minority groups. More specifically, we observed a large coefficient (1.225) for an increase in the percentage of the black community with a low standard error. What this implies is that for a 5% increase in black's population proportion, we expect a 6.13% increase in Obama's winning rate, which is quite significant in a presidential campaign. Obama should allocate more of his resources to target the black population.

On the other hand, the coefficient for the Hispanic population is not only small but negative. Obama should move away from focusing on this segment as he would face tremendous opportunity costs and the return could very well be negative for areas with a large Hispanic population.