**Data Science for Business Case: 2008 Democratic Primaries – Clinton vs. Obama**
**Section B Team 9: Alper Sayiner, Annabelle Nguyen, Vatsal Sanghavi, and Sophia Rowland**

1. Pick two (or more) variables and attempt to show a relation between them via visualization. As discussed before, this requires one to formulate a question, and to communicate clearly a conclusion based on data visualization (specify the why, what, how). (Note that in this question it is not required that the relationship displayed relates to the election.)

Does the age of the population affect the Social Security budget of a state? According to Figure 1, there is a positive relationship between the percentage of population at age 65 and above and the Social Security Rate. Because senior citizens receive Social Security benefits, the data reflects the reality that as the population ages, the government has to spend more on Social Security. Moreover, as the percentage of population at age 65 and above increases, the variance in Social Security rate also increases, suggesting a variation in the allocation of Social Security budget to each state.

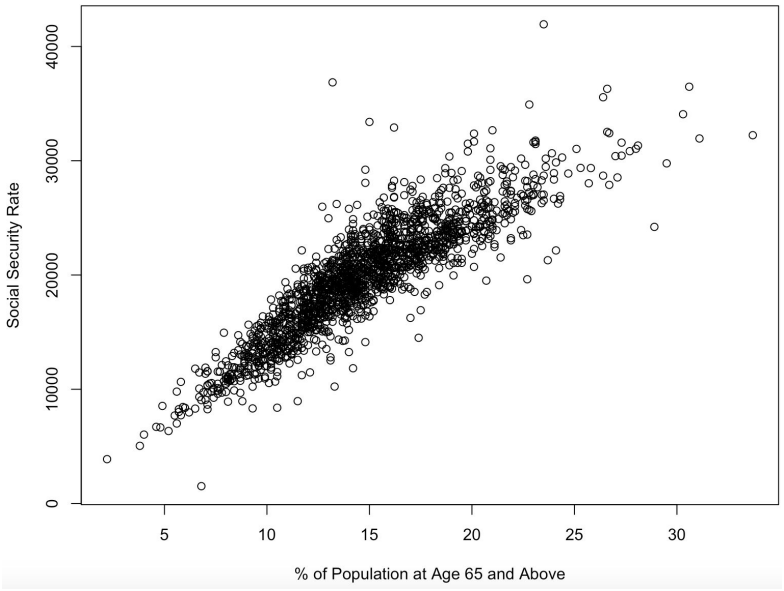


*Figure 1*

Does an individual's education play a role in determining income? Figures 2 and 3 show two interesting facts about the demographics of the US. First, for the majority of the states, the percentage of the population holding a Bachelor's degree is < 20%, showing that the average education level in the US is high school level. Second, the relationship between the percentage of the population holding a high school degree and average income is weak and positive, while the percentage of the population holding a Bachelor's degree and average income is slightly stronger and positive. Therefore, it is worth investigating whether higher education increases average income, but that is a question for a different analysis.

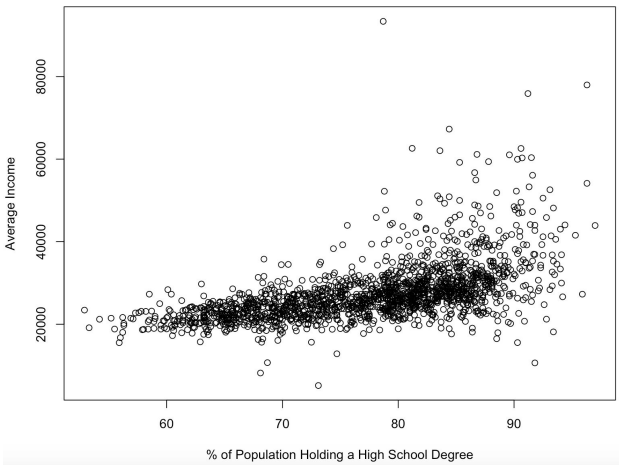


*Figure 2*

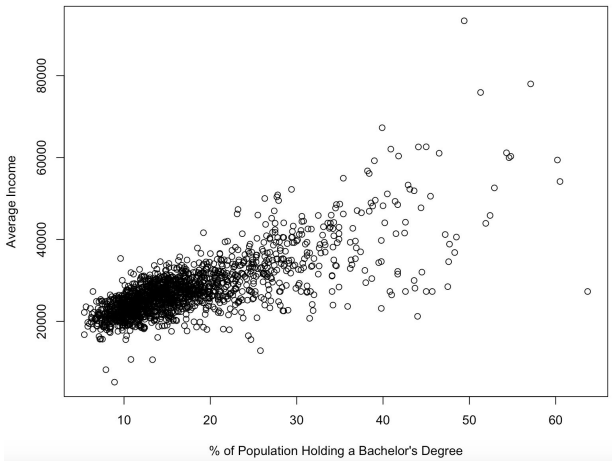


*Figure 3*

2. Provide a model to predict the winning spread of Obama over Clinton measured as percentage of the total vote. Describe clearly the core task, briefly discuss all the models you compared, state which metric is being used to evaluate performance, and how did you chose a final model. Apply and report a K-fold cross validation to evaluate the performance of your chosen model. Based on your final model, predict the winning spread percentages for the test sample (provide the R code that generate your predictions).

Since our core task is to predict the percentage of the total votes that Obama wins over Hillary, we created a linear predictive model. To determine which variables to include in our linear regression model, we first created a correlation matrix to see which variables correlated with the winning spread of Obama (see figure 4). Using the correlation matrix and some thought, we compiled a list of variables to include in our first linear model. Next, we created a data frame of the interaction effects within our set of variables. We ran a regression on this data frame to determine at a glance which interactions seem most significant and which interactions made the most sense. Using these variables, we created a second model that included several interactions.

To evaluate the performance of these models, we used OOS $R^2$ values. Our first linear model without interactions yielded the highest OOS $R^2$ when we applied 10-fold cross validation, and therefore we chose this model (Figure 5). We created a linear regression from this model (Figure 6) and then used it to predict the spread of Obama_margin_percent. Our regression predicted a range of values from -141.233 to 98.963 with an average of -16.324.

3. In order to explore the data, apply one unsupervised learning tool (e.g., k-means, principal component analysis), interpret and communicate briefly the output (e.g., clusters, latent features), and attempt to obtain insights.

First, we looked at how different demographic groups voted for Obama. We standardize the X and Y axes to account for different units. Normalization is done by subtracting the mean

from each data point and dividing by the standard deviation. “Obama” in the X axis in the graphs represents the standardized number of votes that Obama received. Value in the Y axis in the graphs represents the standardized percentage of population of the corresponding groups.

*Figure 7:* There are 2 groups in the plot of standardized number of votes from the black population that Obama received. The first group contains a relatively low black population and Obama received a high number of votes from this group. The second group contains a relatively high black population, and Obama received a smaller number of votes. Therefore, Obama was more successful in gathering votes in areas where the black population is relatively low.

*Figure 8:* There are 2 groups in the plot of standardized number of votes from the white population that Obama received. The first group contains a relatively low white population and Obama received a high number of votes from this group. The second group contains a relatively high white population, and Obama received a smaller number of votes. Therefore, Obama was more successful in gathering votes in areas where the white population is relatively low.

*Figure 9*: There are 2 groups in the plot of standardized number of votes from the Asian population that Obama received. The first group contains a relatively low Asian population and Obama received a low number of votes from this group. The second group contains a relatively high Asian population, and Obama received a higher number of votes. Therefore, Obama was more successful in gathering votes in areas where the Asian population is relatively high.

*Figure 10*: There are 2 groups in the plot of standardized number of votes from the Hispanic population that Obama received. The first group contains a relatively low Hispanic population and the second group contains a relatively high Hispanic population. In both of these groups, Obama received roughly the same number of votes. Therefore, the size of the Hispanic population does not affect Obama’s success in gathering votes.

*Figure 11*: There are 2 groups in the plot of standardized number of votes from the high school graduate population that Obama received. The first group contains a relatively low number of votes and the second group contains a relatively high number of votes. In both of these groups, the population of high school graduate is similar. Because the cluster of high number of votes is a lot bigger than that of low number of votes, the Obama campaign team should focus on converting this group from non-supporters to supporters.

*Figure 12*: There are 2 groups in the plot of standardized number of votes from Bachelor’s degree holder population that Obama received. The first group contains a relatively low

number of Bachelor's degree holder and Obama received a relatively lower number of votes. The second group contains a relatively higher number of Bachelor's degree holder and Obama receives a higher number of votes. Therefore, Obama was more successful in gathering votes from the population of Bachelor's degree holders.

*Figure 13*: There are 2 groups in the plot of standardized number of votes from the population under the poverty line that Obama received. The first group contains a relatively low number of people living below the poverty line and Obama received a relatively higher number of votes. The second group contains a relatively higher number of people living below the poverty line and Obama receives a higher number of votes. Therefore, Obama was more successful in gathering votes from the areas with a small portion of the population living below the poverty line. In other words, Obama was more successful in gathering votes from wealthy areas.

*Figure 14*: There are 2 groups in the plot of standardized number of votes based on median income of the county. The first group contains a relatively low median income population and Obama received a relatively lower number of votes. The second group contains a relatively higher median income population and Obama receives a higher number of votes. Therefore, Obama was more successful in gathering votes from the areas with higher median income, which confirms the insight generated from *Figure 13*.

4. Several sources have been reporting that the demographic composition of the US is changing which can definitely impact how campaigns will be run. In many states, the Hispanic population is growing at a faster pace than others. Looking ahead, provide an estimate for what would have been the average impact on the winning spread for Obama over Clinton (measured in percentage of total voters) had the Hispanic demographic been 5% larger? What if the Black demographic was 5% larger? Be careful to isolate the impact of the specific demographic change alone. (This question would be too open ended. In the R starter script we provide a "simple" model with 1771 variables to be used for which we will assume that the Conditional Independence Assumption (CIA) holds.)

We are assuming a 5% point increase in the percentage of the hispanic and black populations, thus we modified the columns for the percentage of black and hispanic people in the population by increasing them by 5 percentage points. Then, we ran the linear regression model we built in Question 2. For the effect of the increase in black population of we used the model as is. For hispanic, we added the hispanic variable to the model. According to the results, if the hispanic population was 5% higher, the impact of the percentage hispanic population would be -0.102. On the other hand, if the black population was 5% higher, the impact of the percentage hispanic population would be 1.713.

5. Choose one candidate. What kind of advice (based on data analytics) would you provide to your candidate? For example, which voter segment to target with their campaign messages

and why? Or, how to allocate resources (budget and volunteer time) across regions and why? How would you communicate such insights?

Based on the outcome of the regression model and the insights mentioned in Question 3, assuming that the Obama campaign team did not change their stance on the important issues raised on the campaign trail, we would recommend the following strategies to strengthen Obama's current supporter base.

- The Obama campaign team should target counties and states with a low proportion of white population. The reasons are the negative and significant coefficient for the white population, and the insights from the unsupervised learning model that Obama was more successful in gathering votes in areas where the white population is relatively low.
- The Obama campaign team should target counties and states with a high proportion of Asian population. The reason is the insights from the unsupervised learning model that Obama was more successful in gathering votes in areas where the Asian population is relatively high.
- The Obama campaign team should target counties and states with a low proportion of population below the poverty line. The reasons are the negative and significant coefficient for the poverty population, and the insights from the unsupervised learning model that Obama was more successful in gathering votes from wealthy areas.

**APPENDIX**

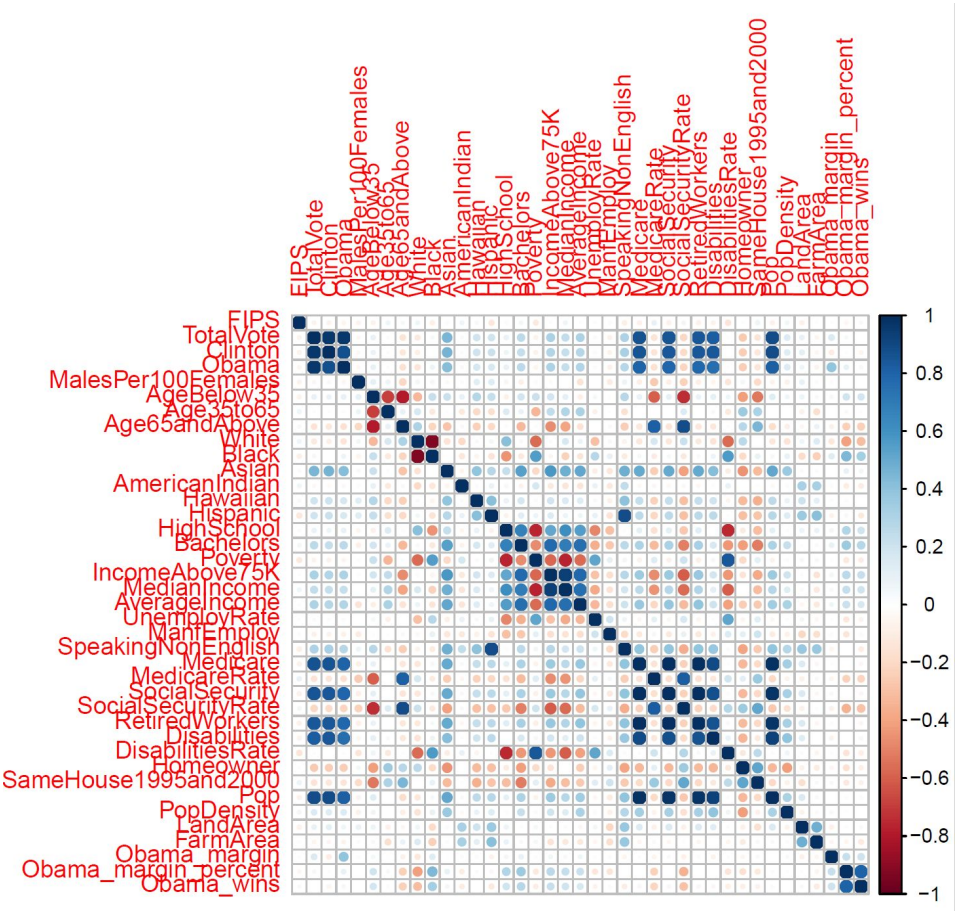


*Figure 4*

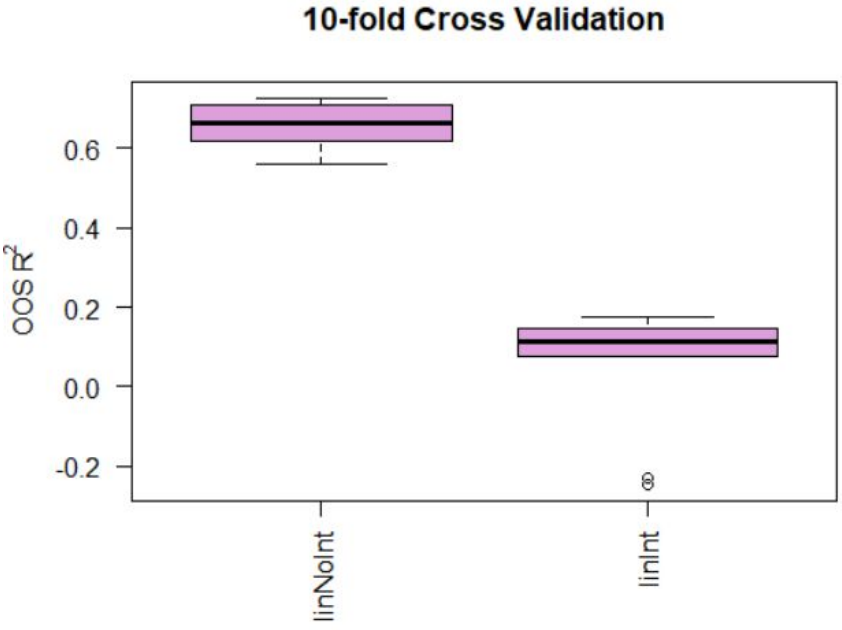


*Figure 5*

```
Call:
glm(formula = Obama_margin_percent ~ AgeBelow35 + Age65andAbove +
    White + Black + HighSchool + Bachelors + Poverty + IncomeAbove75K +
    MedianIncome + AverageIncome + MedicareRate + SocialSecurityRate +
    ManfEmploy + DisabilitiesRate + UnemployRate + Homeowner +
    SameHouse1995and2000 + Region, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-86.539  -10.818   -0.313   11.694   67.545

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -5.194e+00  2.286e+01  -0.227 0.820297
AgeBelow35            1.862e-01  1.643e-01   1.134 0.257017
Age65andAbove        -1.157e+00  3.380e-01  -3.422 0.000635 ***
White                -2.835e-01  8.309e-02  -3.412 0.000661 ***
Black                 1.713e+00  7.802e-02  21.955  < 2e-16 ***
HighSchool            1.417e-01  1.193e-01   1.188 0.235146
Bachelors             1.487e+00  1.264e-01  11.759  < 2e-16 ***
Poverty              -2.297e+00  2.845e-01  -8.072 1.28e-15 ***
IncomeAbove75K       -8.595e-01  2.312e-01  -3.717 0.000208 ***
MedianIncome         -3.649e-04  1.905e-04  -1.915 0.055649 .
AverageIncome        -1.979e-05  1.268e-04  -0.156 0.875960
MedicareRate          7.672e-04  1.633e-04   4.698 2.84e-06 ***
SocialSecurityRate   -5.316e-04  3.102e-04  -1.714 0.086716 .
ManfEmploy           -8.665e-02  6.843e-02  -1.266 0.205598
DisabilitiesRate     -3.014e-03  6.578e-04  -4.583 4.93e-06 ***
UnemployRate         -8.866e-01  3.474e-01  -2.552 0.010792 *
Homeowner             6.559e-01  9.262e-02   7.082 2.07e-12 ***
SameHouse1995and2000  1.919e-01  9.789e-02   1.961 0.050092 .
RegionNortheast      -1.947e+01  1.979e+00  -9.837  < 2e-16 ***
RegionSouth          -2.547e+01  1.525e+00 -16.699  < 2e-16 ***
RegionWest            9.928e+00  1.528e+00   6.497 1.07e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 306.5084)

    Null deviance: 1595745  on 1736  degrees of freedom
Residual deviance:  525968  on 1716  degrees of freedom
AIC: 14897
```

*Figure 6*

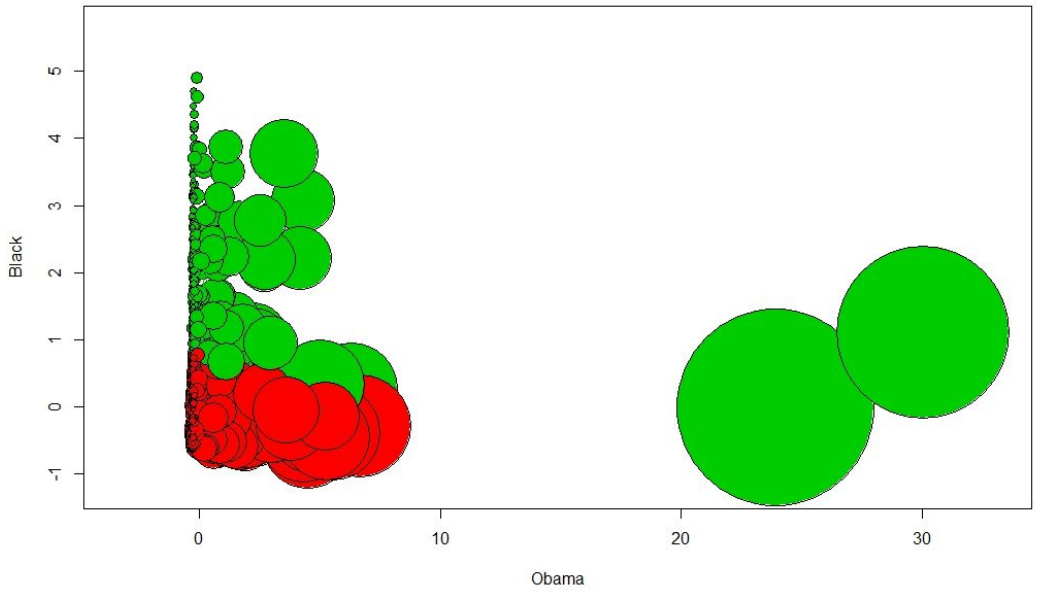


*Figure 7*

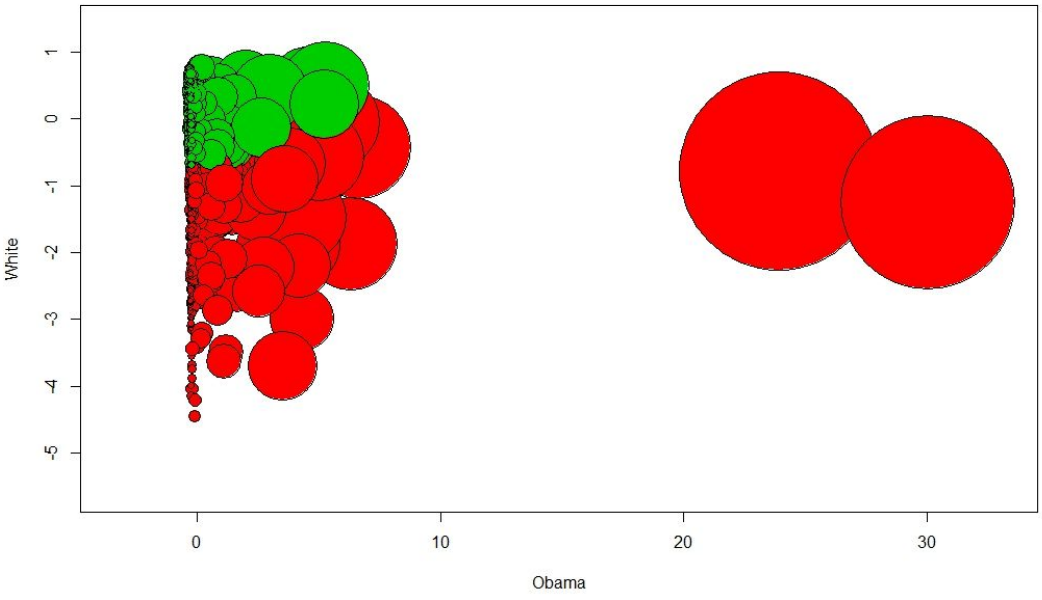


*Figure 8*

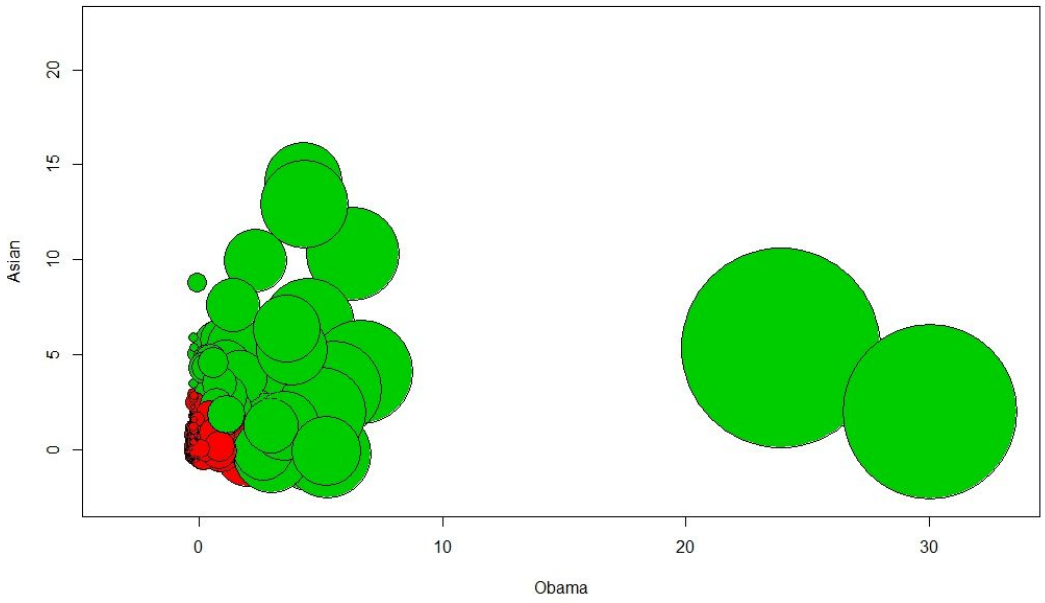


*Figure 9*

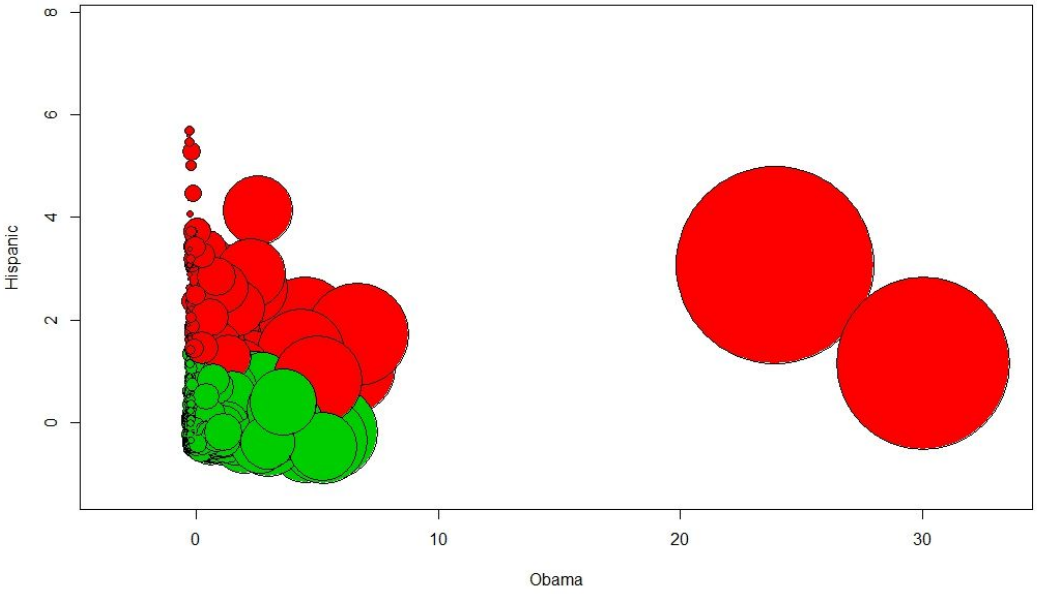


*Figure 10*

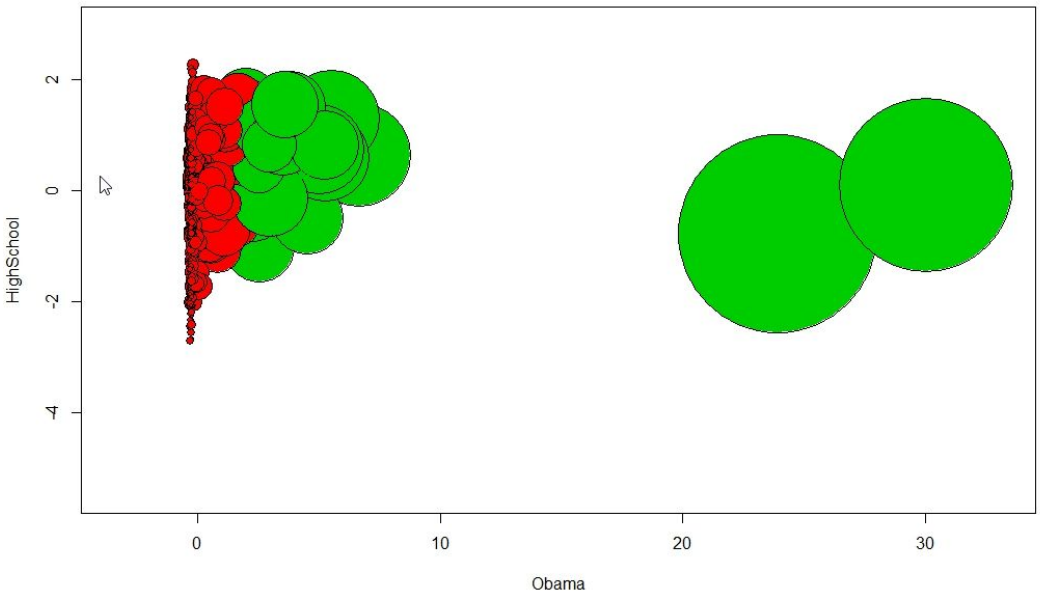


*Figure 11*

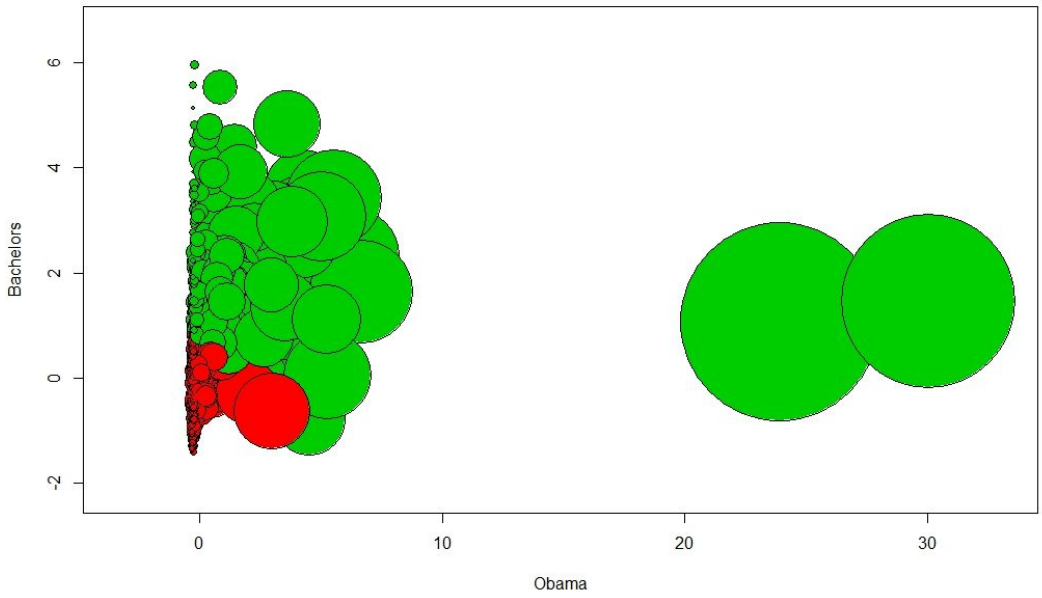


*Figure 12*

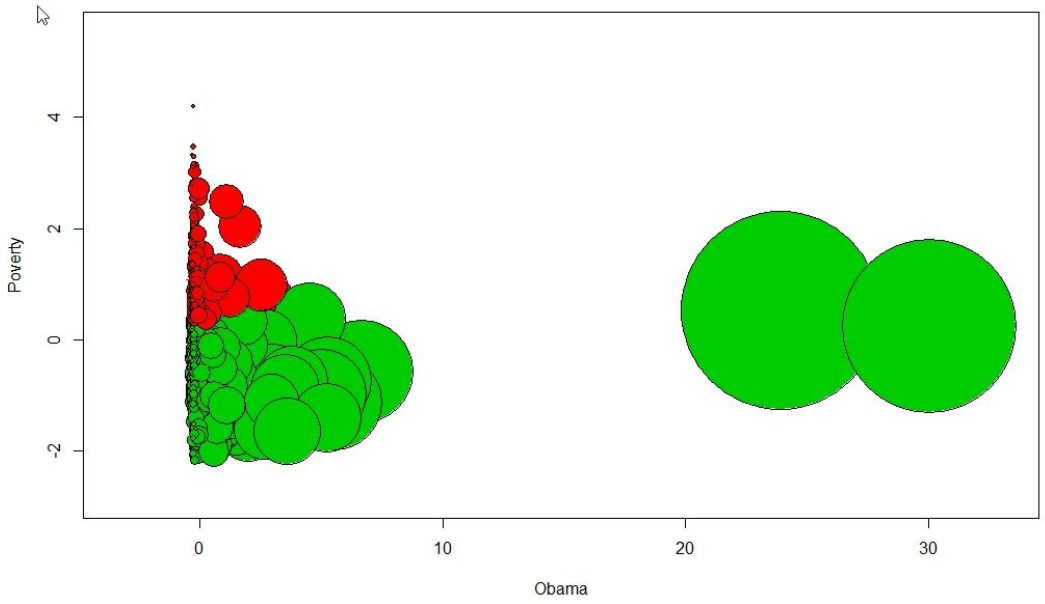


*Figure 13*

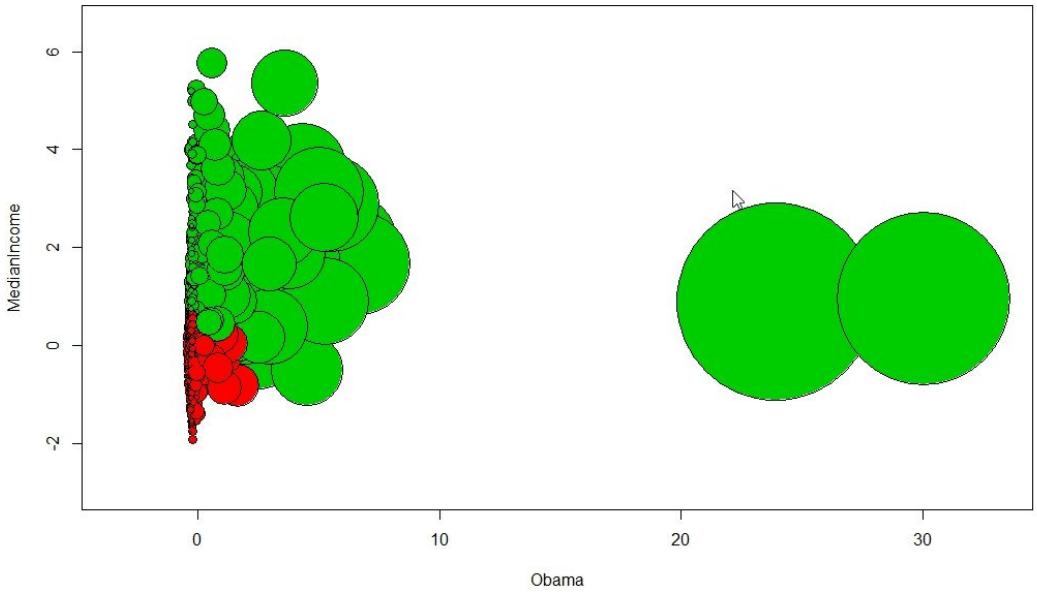


*Figure 14*