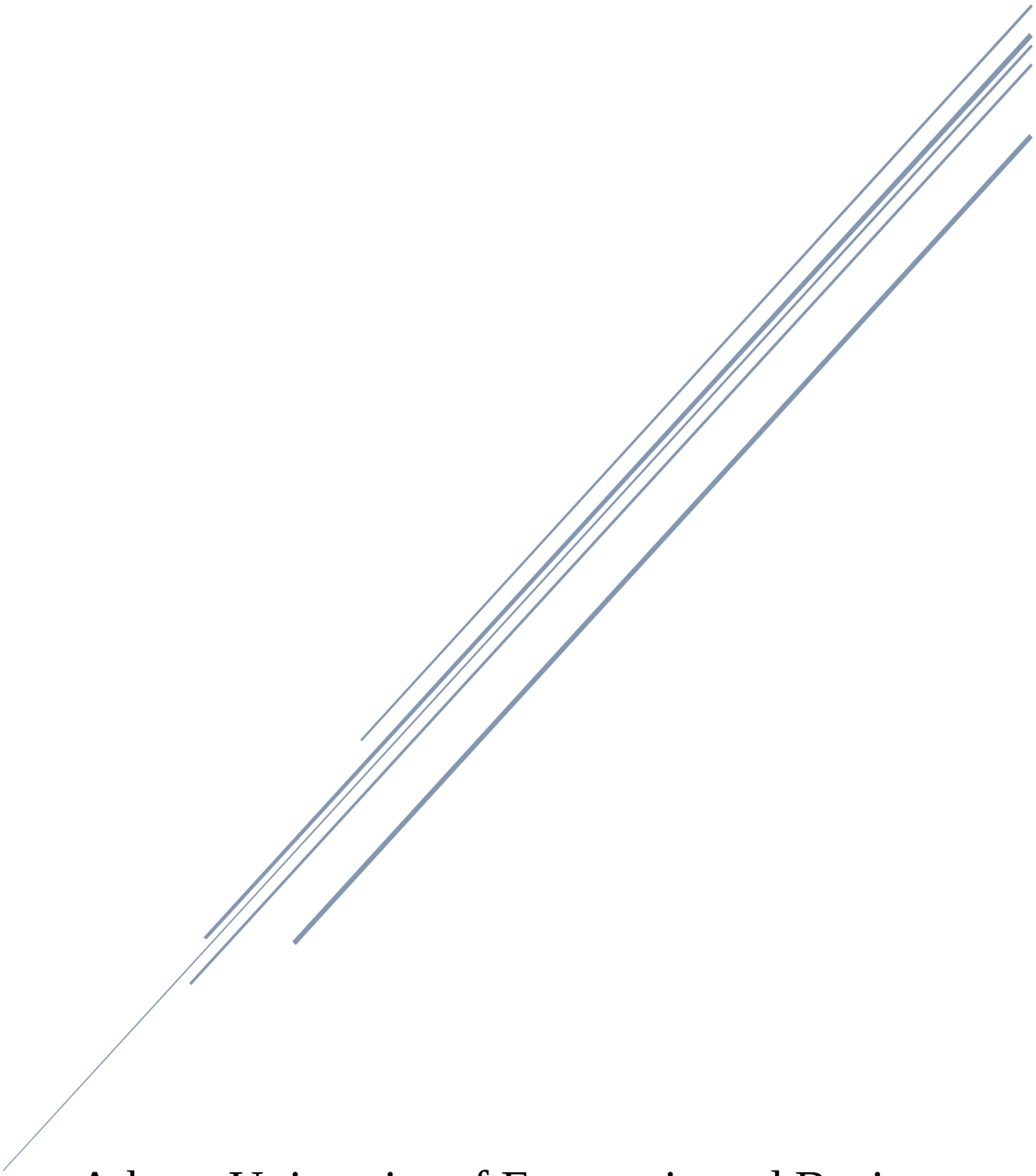
STATISTICS FOR BUSINESS ANALYTICS II - Project I 2020-2021

Vretteas Stylianos
Student Id: p2822003
Professor: Karlis Dimitrios



Athens University of Economic and Business
MSc In Business Analytics

Table of contents

Section I - Introduction

Project I of the MSc BA Statistics II is based on the 2016 Democratic Party presidential primaries, the case is to build a reasonable model using as a response whether Hillary Clinton won over Bernie Sanders using as exploratory variables socioeconomic characteristics of the counties. Our primary focus will be understanding as better as possible the voter's behavior and not prediction.

The provided project data consists of two combined excel sheets, which describe the votes and the socioeconomic characteristics of US counties respectively. Alongside this given information there is also a dictionary in order to do the mapping for the socioeconomic characters which will serve as our variables. The original data consists of the following, for the Votes datasheet there are 24611 observations and 8 variables, while in the county_facts datasheet there are 3195 observations and 54 variables, the dictionary as already said is the mapping of these 54 variables. Variables of the county_facts datasheet are splinted into groups and can be distinguished easily through their first three letters which serve as a code, these variable groups describe measures, metrics and other characteristics of the counties such as income, household information, age, race and gender.

The first task of this project is to walk our way through the two given datasheets, do the proper cleaning and merge them into one data frame before we proceed further into variable selection and modelling.

Our second goal will be to plot and observe several groups of variables trying to find patterns regarding the behavior of the voters towards Hillary Clinton and Bernie Sanders by state, age, gender and other characteristics.

After finishing with the above short exploratory analysis, we will implement variable selection algorithms in order to keep the most important of these characteristics and try to construct a reasonable model in order to understand the voter's behavior.

Finally, at the end through statistical methods we are going to evaluate our model and see if it is enough "good" to explain the data and furthermore the characteristics that forced the outcome of the elections.

Section II – Data cleaning

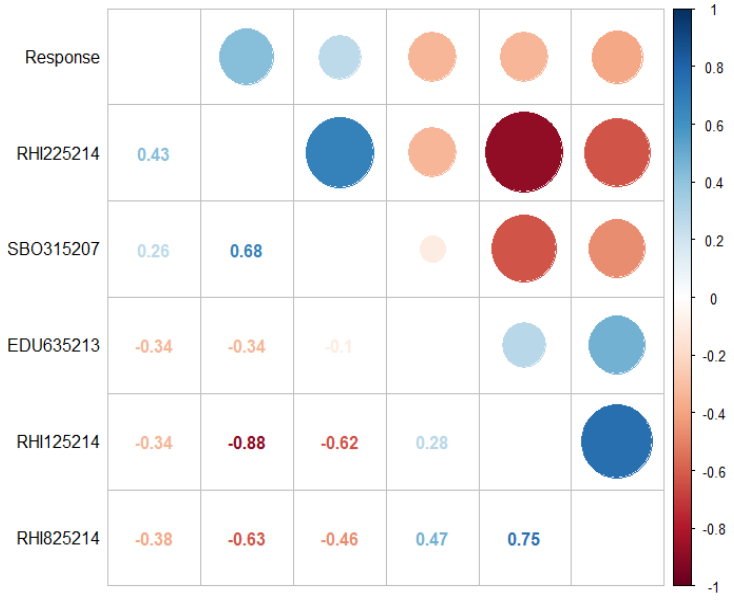
For the implement of this procedure Microsoft Excel used as first tool in order to transform and have a first look of the data files. Each sheet was converted into csv file in order to be loaded separately into R into the votes, county_facts and dictionary objects respectively. After this all procedures were made with R data cleaning techniques and especially with the use of the dplyr package. First of all, the data was filtered into the Democratic party only as the Republican party was out of our interest. The filtered Democratic party votes dataset consists of 8959 observations and 8 columns, the candidates of the Democratic party were Hillary Clinton, Bernie Sanders, Martin O'Malley and there were two more types of candidate choice the Uncommitted and the No preference choices. First problem of this dataset was the missing twenty (20) fips values of the New Hampshire State. Fips stands for Federal Information Processing Standard Publication , it is unique identification number of each county. These values were easily imputed with their actual values as the information was already included into the county_facts datasheet. Next in order to create our response variable and merge the two datasheets, we implement the pivot technique into the votes objects, the thought is to create two columns with the sum of the total votes of each of the two candidates of interest Hillary Clinton and Bernie Sanders by fips, so we create our first pivot and exclude the other three candidate choices. In order to create our response value, a short if statement was implemented and provided the value “1” for Hillary Clinton and the value “0” for Bernie Sanders as the election result is a “winner takes all” game, whom has the most votes wins. At this point, we anticipate the problem of the tie result which occurred 166 times, if there is a tie in real circumstances a county council is held and the outcome is agreed internally. For this project reasons, in situations where the tie was zero – zero votes these observations were dropped (87 observations), for the actual tie situation I decided to impute the value of the winner of the state (79 observations, this is why a second pivot was created in order to act as a reference table).

Finally, merging the two datasheets by fips column we end up with a data frame of 2802 rows and 60 variables which fifty-one (51) of them are exploratory and one is the response variable. The other seven (8) are columns providing information about the State, state abbreviation, sum of the total votes and will serve later as arguments for plotting reasons.

Section III – Exploratory Data Analysis

In this section, first I implement a correlation matrix to have a first look of the associations between the variables with the created response variable. Since

Figure 1



The correlation matrix suggests that variables coded RHI225214 and SBO315207 which stand for Black or African American alone, percent, 2014 and Black-owned firms, percent, 2007 seem to have the more positive correlation with the Response variable, this is reasonable since the dataset consists only of Democratic voters and the African American community have traditionally better ties with the Democratic party.

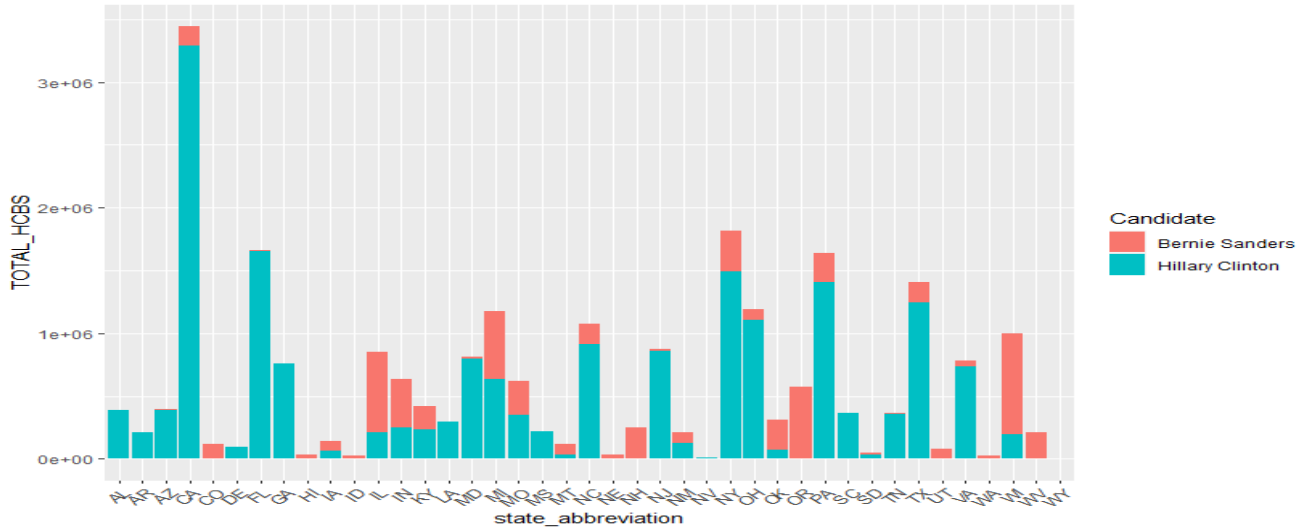
The negative correlated variables are the RHI825214 - White alone, not Hispanic or Latino, percent, 2014 and the RHI125214 - White alone, percent, 2014 and the EDU635213 - High school graduate or higher, percent of persons age 25+, 2009-2013. This indicates that white people tend and the more highly educated to vote opposite sides.

These assumptions should be examined furthermore because since this is a dataset of Democratic party voters and we have to dig deep further in order to find internal differences.

Furthermore, we are going to see a few graphs which provide general information about the election's outcome by State, income, white and black alone voters over the two candidates, age distribution and female voters. I chose these variables and these graphs as I believe they cover a large part of the socioeconomic characteristics of the counties.

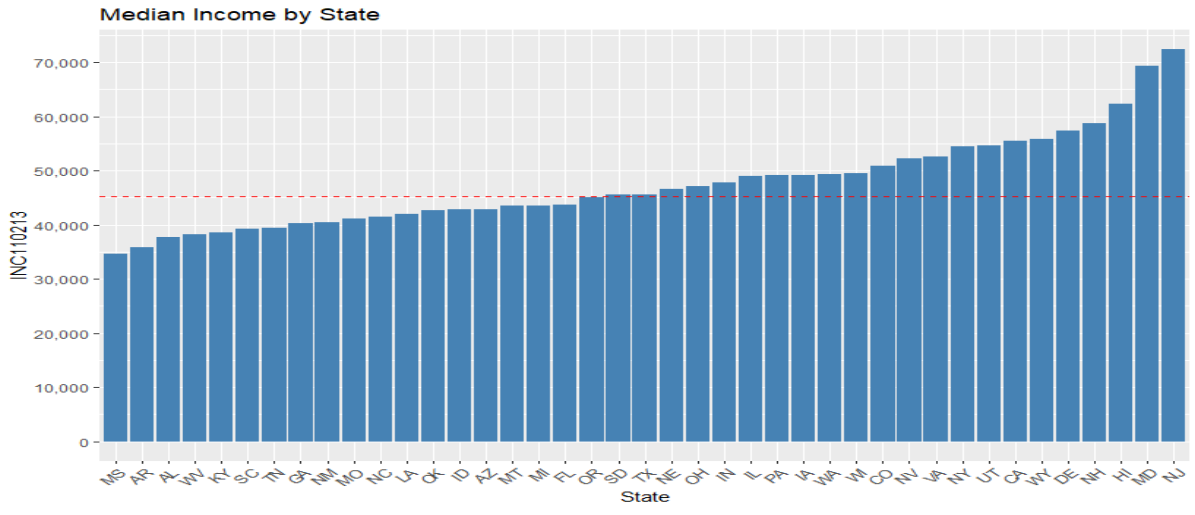
In the below stacked grouped graph, it is presented the total amount of the votes per state. The fill of the bar is the allocation of the votes into the two candidates. It is crystal clear that Hillary Clinton is the absolute winner over Bernie Sanders in the majority of states.

Figure 2



In figure 3, the visualization shows the median income of each state, where New Jersey, Maryland and Hawaii are the "richest" states in contrast of Mississipi, Arkansas and Alabama were are strongly poorer than the rest of the country. The red dotted line is the mean amount of median income.

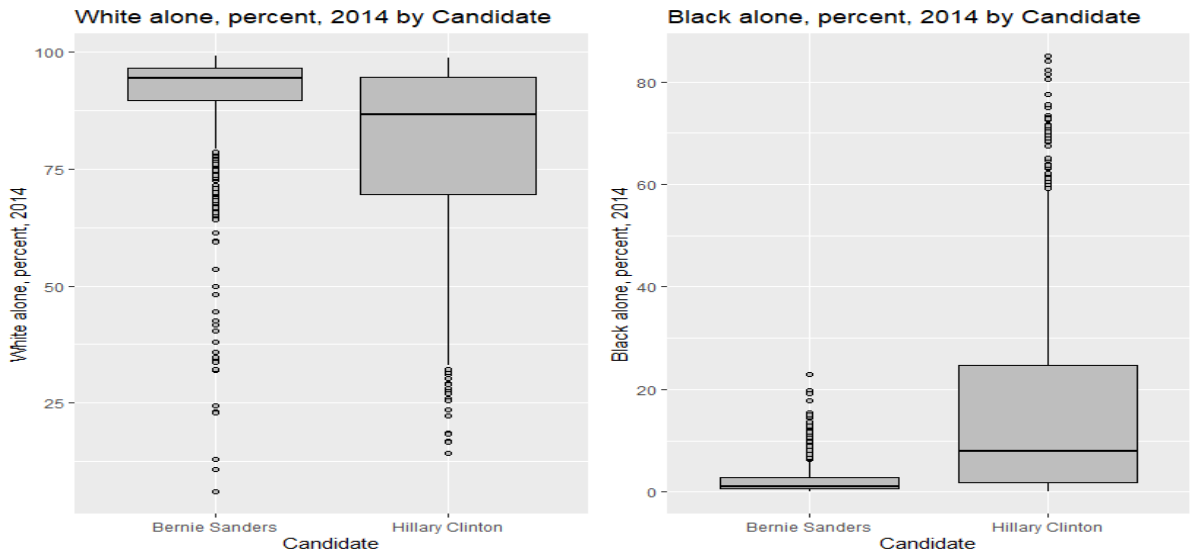
Figure 3



Combining the two above visualizations we can see that Hillary Clinton won in both the poor and the rich States so we cannot indicate that the income factor had a crucial impact in the outcome of the elections.

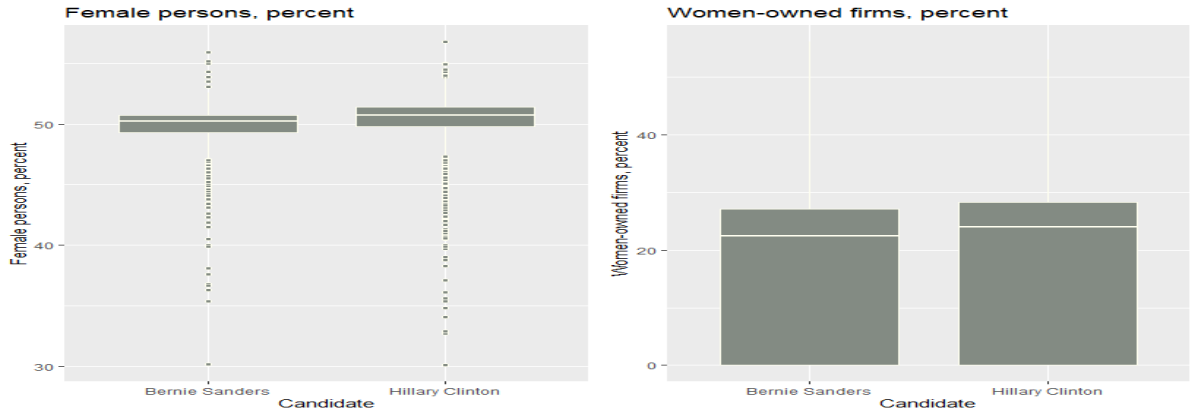
In figure 4, we plot the percent of the white alone voters per candidate next to the black alone voters per candidate. We can say that in both cases there is a large difference between the proportion of the white and black alone voters. We can say that in the Democratic Party Primary Elections of 2016 the percent of the white voters was bigger than the black voters.

Figure 4



Next, we plot the two variables that provide information about the women. Females persons percent and Women – owned firms' percent. We can see that the percent of female voters is quite the same in both Clinton and Sanders voters with low variability. Regarding the women owned firms the variability is large and the percent is equal. We can say that the female voters behaved the same in

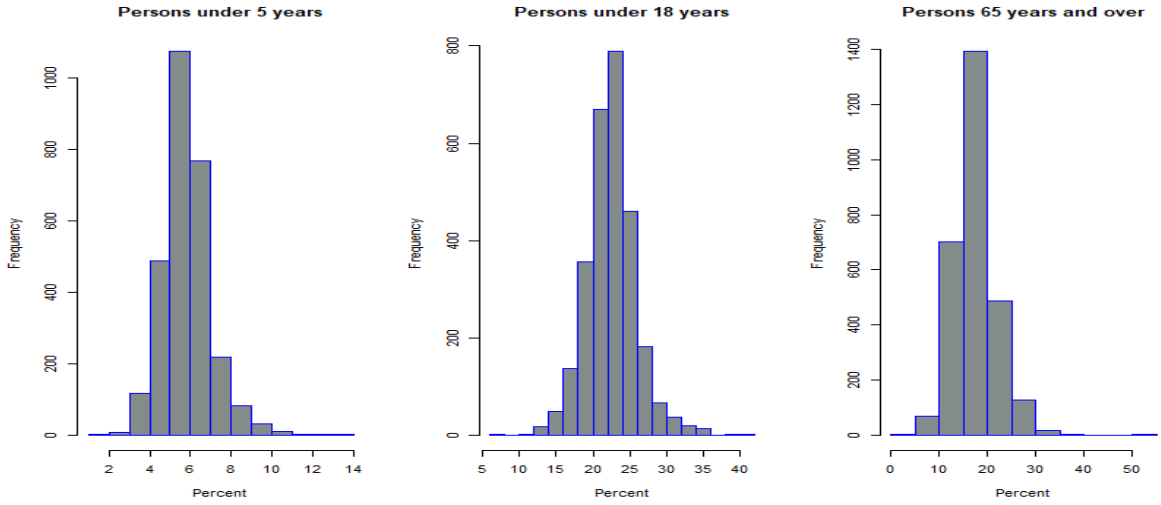
Figure 5



Next in our analysis, we present three histograms regarding the Age percent variables. We can say that in the most territories the mean percentage for persons below five (5) years is 6

percent, for persons under eighteen (18) years is approximately 22 percent and for the age group of the persons above sixty-five years old (65) the mean is about 20 percent.

Figure 6



Section IV – Model building & Evaluation

Before we proceed into the model building section, we have to make sure our data is ready for the variable selection algorithms of Lasso and Step AIC & BIC, we will use these techniques in order to find the most appropriate variables for our model. Each method has its pros and cons which I will describe shortly below.

First of all, we have to ensure that all our variables are of "numeric" class, in case we want to import a categorical variable into the model, we have to transform it into dummies and then use the algorithms. I decide not to use the categorical variable State because I want the model to have as covariates just the socioeconomic variables of the states

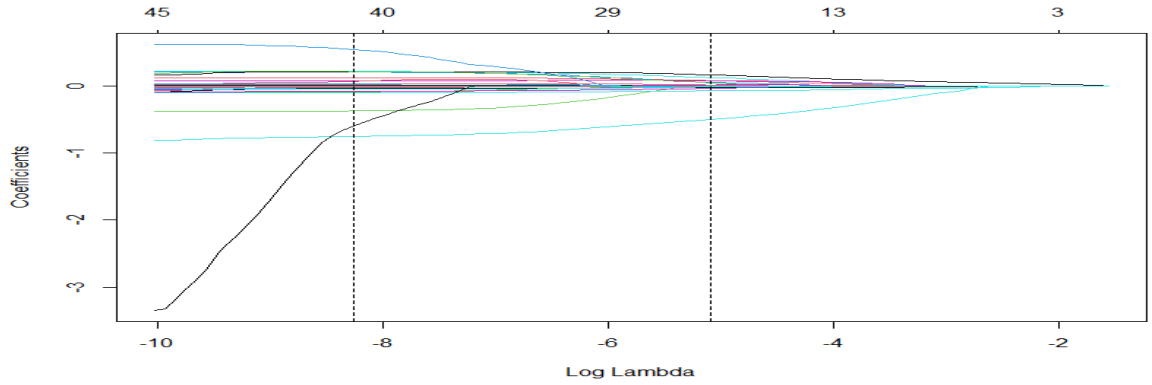
Since the dataset consists of too many variables, I decide to implement first the Lasso regulation technique in order to estimate the general linear model, Lasso includes a penalty term L1 which it reduces the maximum likelihood and shrinks the "biased" coefficients ones into zero through the regulization parameter of λ. As λ increases, the number of nonzero components of β decreases.

Its pros are that it deals heavy with the multicollinearity of the data and keeps the most important variables leading into a more parsimonious model with fewer variables. Multicollinearity problem is where an exploratory independent variable is strongly related

with other independent variables. Later we can further examine these variables importance through their p-values.

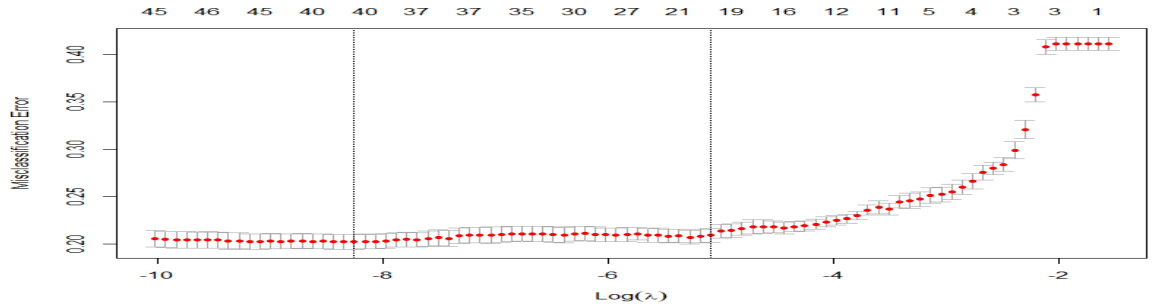
We can see in the below graphs that the optimal Lasso λ parameter decide to keep around 20 variables of the original dataset.

Figure 7



We choose this value of λ over the minimum λ because we have the same amount of shrinkage penalty for the same level of misclassification error.

Figure 8



For continuing the project, I decide to create a new data frame the lasso_data data frame which includes only these selected variables and try to construct and optimize the general linear models only with these variables. I will optimize as much as I can the model from the lasso proposed variable and then I will implement BIC step algorithms in order to do further data selection. I choose BIC because it is used further for interpretation rather than prediction like AIC.

Before I further proceed into the model building section, I provide a short theoretical background briefing for the logistic model. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by

estimating probabilities using a logistic function. First our dependent variable is binary ("success or loss ") in our case Hillary Clinton won over Bernie Sanders and second the predicted values are probabilities and not values. So, in its simple and theoretical form our model will be like the below.

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i$$

At this point we can construct our first model, "model1" which is the full model with all the selected variables. We will explain first the summary of model2 alongside other goodness of fit measures and then the bic_model. Below the summary of model2

Table 1 – model1

```
~/01 Business Analytics - AUEB/03 Statistics 2/project/
> summary(model1)

Call:
glm(formula = Response ~ ., family = binomial(), data = lasso_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8294  -0.7653   0.0238   0.6302   3.5313

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.648e+00  1.697e+00  -1.560 0.118808
PST120214    1.189e-01  2.088e-02   5.697 1.22e-08 ***
AGE295214    1.025e-01  2.968e-02   3.454 0.000553 ***
AGE775214    1.968e-01  2.432e-02   8.095 5.73e-16 ***
SEX255214    9.681e-02  3.296e-02   2.937 0.003316 **
RHI225214    2.373e-01  1.867e-02  12.708  < 2e-16 ***
RHI325214   -1.095e-02  1.070e-02  -1.023 0.306139
RHI425214    1.615e-01  6.138e-02   2.630 0.008526 **
RHI625214   -6.512e-01  7.425e-02  -8.771  < 2e-16 ***
RHI825214   -3.024e-02  7.836e-03  -3.859 0.000114 ***
EDU635213   -8.805e-02  1.394e-02  -6.316 2.68e-10 ***
EDU685213   -7.465e-02  1.474e-02  -5.063 4.12e-07 ***
VET605213    6.881e-06  6.433e-06   1.070 0.284782
HSG096213   -2.700e-02  1.124e-02  -2.402 0.016323 *
HSG495213   -3.938e-06  1.364e-06  -2.888 0.003878 **
INC910213    1.812e-04  2.330e-05   7.775 7.56e-15 ***
BZA115213    9.010e-03  9.541e-03   0.944 0.345030
SBO215207   -4.095e-03  6.517e-02  -0.063 0.949894
SBO415207    3.997e-02  1.943e-02   2.057 0.039702 *
LND110210   -1.945e-05  4.291e-05  -0.453 0.650347
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3796.1  on 2801  degrees of freedom
Residual deviance: 2277.9  on 2782  degrees of freedom
AIC: 2317.9

Number of Fisher Scoring iterations: 7
```

From the first look the model1 seems that it can be optimized further.

There are several variables that they are not statistically significant since their p-values are over the limit of 0.05 and can be considered as zero including the intercept too. I decide to remove those not statistically significant variables and fit the general linear model again into object model2. Summary of the model2 is presented below.

Table 2

```
~/01 Business Analytics - AUEB/03 Statistics 2/project/
> model2 <- glm(Response~. -RHI325214 -VET605213 -BZA115213 -SBO215207 -LND110210,
+               data = lasso_data, family = binomial())
> summary(model2)

Call:
glm(formula = Response ~ . - RHI325214 - VET605213 - BZA115213 -
    SBO215207 - LND110210, family = binomial(), data = lasso_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8600  -0.7679   0.0221   0.6481   3.5115

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.931e+00  1.659e+00  -1.767 0.077215 .
PST120214    1.216e-01  2.057e-02   5.912 3.39e-09 ***
AGE295214    1.013e-01  2.963e-02   3.419 0.000629 ***
AGE775214    1.983e-01  2.419e-02   8.195 2.50e-16 ***
SEX255214    9.634e-02  3.218e-02   2.994 0.002752 **
RHI225214    2.466e-01  1.727e-02  14.282  < 2e-16 ***
RHI425214    1.861e-01  4.851e-02   3.835 0.000125 ***
RHI625214   -6.528e-01  7.315e-02  -8.924  < 2e-16 ***
RHI825214   -2.501e-02  6.151e-03  -4.067 4.77e-05 ***
EDU635213   -9.230e-02  1.325e-02  -6.967 3.23e-12 ***
EDU685213   -7.585e-02  1.464e-02  -5.181 2.20e-07 ***
HSG096213   -2.524e-02  1.114e-02  -2.266 0.023438 *
HSG495213   -3.998e-06  1.353e-06  -2.956 0.003118 **
INC910213    1.885e-04  2.264e-05   8.325  < 2e-16 ***
SBO415207    4.938e-02  1.805e-02   2.736 0.006214 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3796.1  on 2801  degrees of freedom
Residual deviance: 2281.6  on 2787  degrees of freedom
AIC: 2311.6

Number of Fisher Scoring iterations: 7
```

The second attempt with model2 seems better. All coefficients can be assumed that they are not zero since the p-values are smaller than the confidence level of 0.05 and we can say that they are statistically significant. Even the Intercept can be assumed statistically significant in the confidence level of 0.10 which is better than the previous version of the model.

At this point, I have to check further for multicollinearity problems between the covariates via the VIF test (Vif test results for model2 below).

Table 3

```
> vif(model2) # no multicolinearity problems
PST120214 AGE295214 AGE775214 SEX255214 RHI225214 RHI425214 RHI625214 RHI825214 EDU635213 EDU685213 HSG096213 HSG495213
 1.824586  3.404760  3.752798  1.510027  1.675106  2.304204  1.666050  2.925996  2.484183  5.159789  2.398785  2.952433
INC910213 SBO415207
 4.829796  1.548339
>
```

All results are below ten so we can be sure that there is no multicollinearity problem with the variables of the model2.

For checking how well the model fits the data (which is our primary target) we have not the same methods as we already taught in the linear regression.

In linear models we do not have an R2 metric and one of the most major differences is the interpretation of our residuals. In logistic regression models we do not have the Pearson residuals where the interpretation was the distance of the observation from the fitted line of the regression. We have Deviance residuals which are defined by Deviance a goodness of fit measure whose idea is to generalize the idea of using the OLS to cases where model fitting is achieved by maximum likelihood.

The interpretation of the Deviance residuals is as follows. If the proposed model has a good fit, the deviance will be small. If the proposed model has a bad fit, the deviance will be high. In our case the median deviance is close to zero 0.0221 so we can say that our model is not biased into any direction.

Furthermore, we compare the fitted model with the null model (a model with just the intercept) in order to justify the selection of the variables. The results of this comparison for model2 are presented below.

Table 4

```
> with(model2, null.deviance - deviance)
[1] 1514.518
> with(model2, df.null - df.residual)
[1] 14
> with(model2, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
[1] 3.533542e-315
```

The chi-square of 1514.518 with 14 degrees of freedom and an associated p-value of less than 0.001 tells us that our model as a whole fit significantly better than the null model.

Regarding the interpretation of the coefficients, we have to exponentiate them and interpret them as odds – ratio. Below is the table for this output alongside the confidence intervals for each.

Table 5

```
> exp(cbind(OR = coef(model2), confint(model2)))
Waiting for profiling to be done...
                    OR       2.5 %    97.5 %
(Intercept) 0.05335071 0.002035223 1.3658662
PST120214   1.12930347 1.085235844 1.1764402
AGE295214   1.10659901 1.044247947 1.1729294
AGE775214   1.21928843 1.163374093 1.2791500
SEX255214   1.10113872 1.034125309 1.1732148
RHI225214   1.27967178 1.238426496 1.3251881
RHI425214   1.20450789 1.102125778 1.3312515
RHI625214   0.52057874 0.448863299 0.5979955
RHI825214   0.97529685 0.963433770 0.9869845
EDU635213   0.91182970 0.888267697 0.9356458
EDU685213   0.92695955 0.900491422 0.9537013
HSG096213   0.97507819 0.953992413 0.9965687
HSG495213   0.99999600 0.999993331 0.9999986
INC910213   1.00018852 1.000144845 1.0002336
SBO415207   1.05061928 1.015929435 1.0905679
>
```

We can say that for a one unit increase in each variable, the odds of "success" in our case Hillary Clinton win increases by the value of the OR column.

Last measure of fit, in order to see how the model explains the data is the pseudoR2 metric.

Most notable is McFadden's R2 which is defined as 1−[ln(LM)/ln(L0)] where ln(LM) is the log likelihood value for the fitted model and ln(L0) is the log likelihood for the null model with only an intercept as a predictor. It ranges from 0 to 1 with values closer to zero indicate that the model has no predictive power.

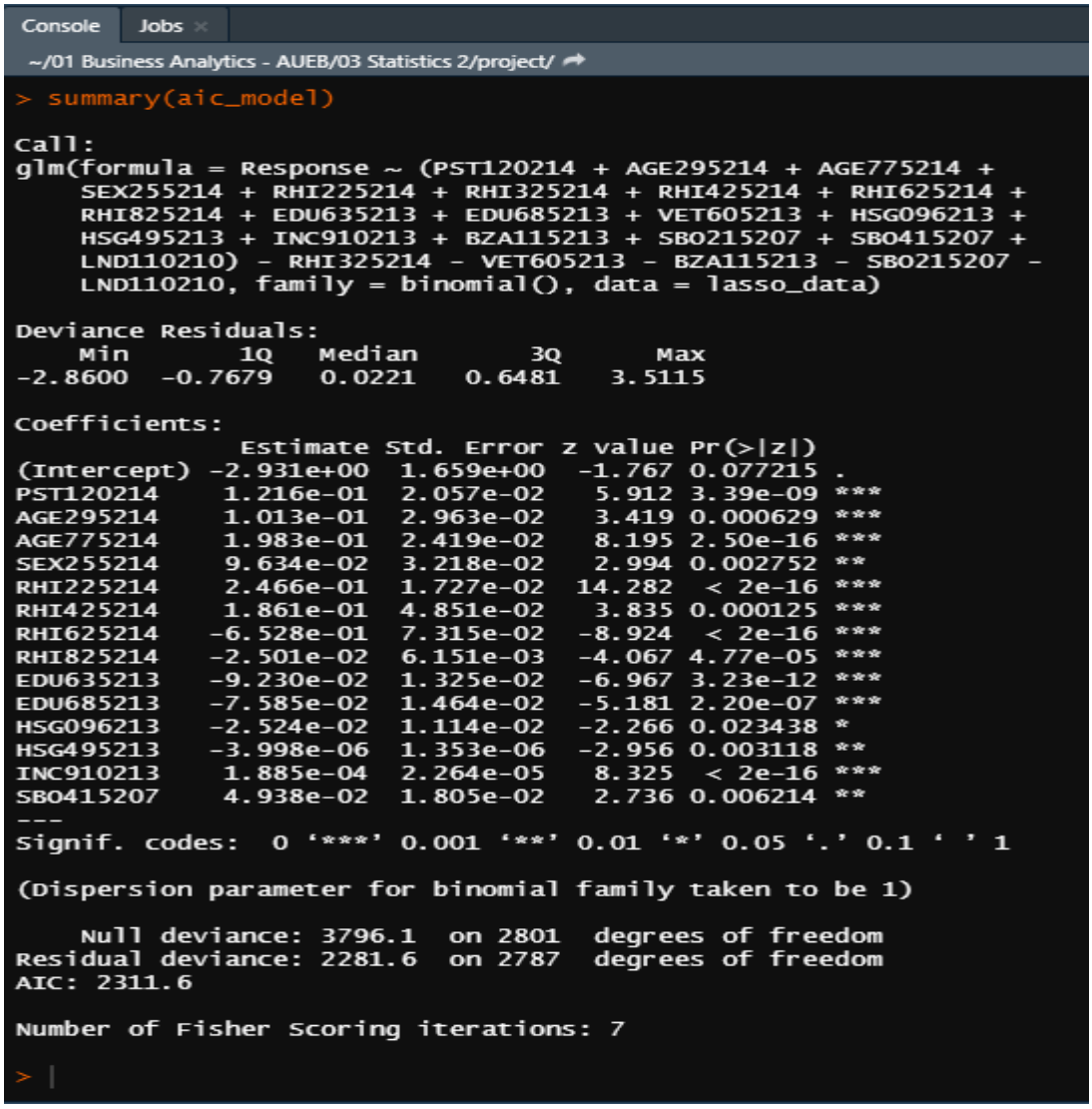
Table 6

```
> pR2(model2)
fitting null model for pseudo-r2
         llh       llhNull            G2      McFadden          r2ML          r2CU
-1140.8072111 -1898.0661390  1514.5178558     0.3989634     0.4175507     0.5627372
>
```

Model2 has a rate of 0.3989, McFadden state that a pseudo R2 ranging from 0.2 to 0.4 indicates very good model fit, so in terms of pseudo R2 our model is very good at explaining the data. That is all for model2

Next we are going to implement step selection algorithms of AIC and BIC in order to optimize furthermore the model. AIC algorithm produces the same output with model2.

Output below.

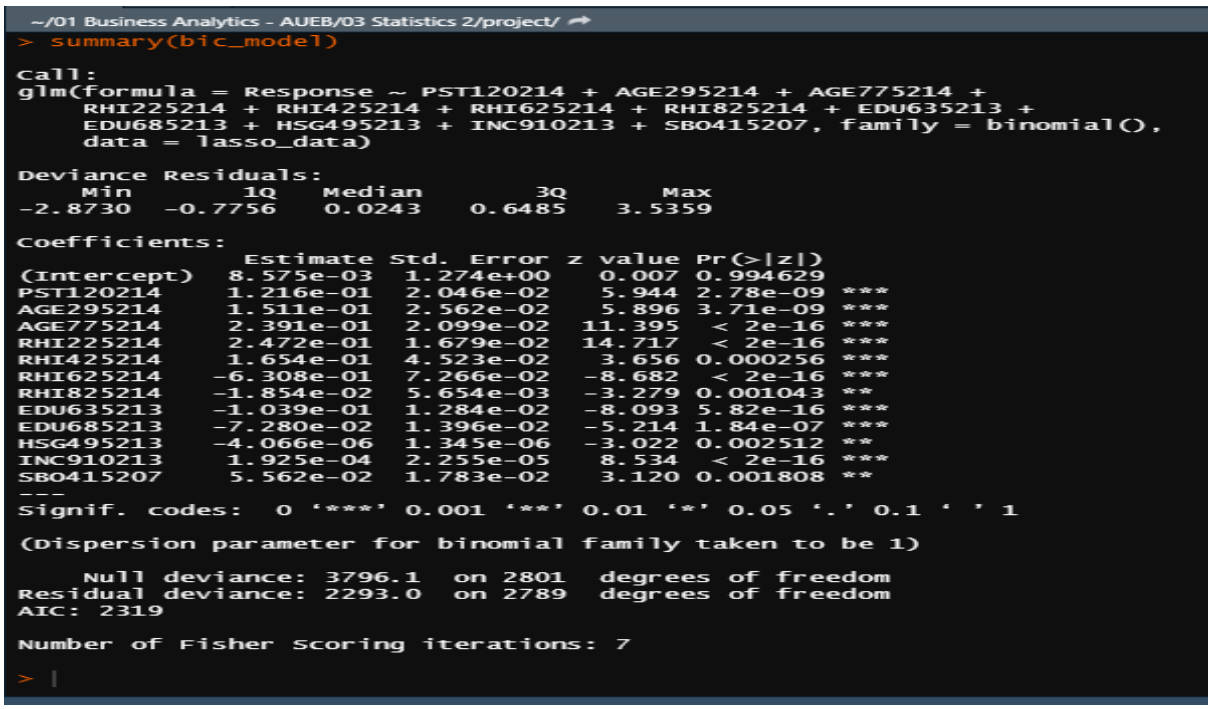


Since this model is same with model2 I will not describe it further.

As I told previously, I decide to implement the BIC step algorithm for further variable selection. I use the BIC for increased interpretation abilities. I name this model as bic_model. And expect this to be a more parsimonious model than the other two since BIC method has more penalty.

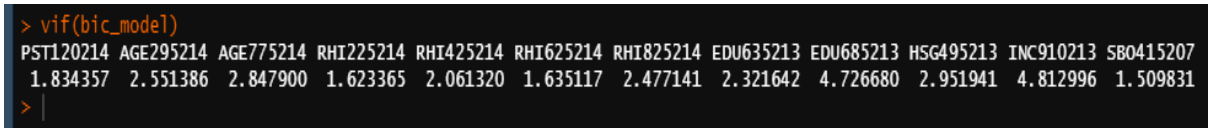
Below its summary

Table 7



```
~/01 Business Analytics - AUEB/03 Statistics 2/project/
> summary(bic_model)

Call:
glm(formula = Response ~ PST120214 + AGE295214 + AGE775214 +
    RHI225214 + RHI425214 + RHI625214 + RHI825214 + EDU635213 +
    EDU685213 + HSG495213 + INC910213 + SBO415207, family = binomial(),
    data = lasso_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8730  -0.7756   0.0243   0.6485   3.5359

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.575e-03  1.274e+00   0.007 0.994629
PST120214    1.216e-01  2.046e-02   5.944 2.78e-09 ***
AGE295214    1.511e-01  2.562e-02   5.896 3.71e-09 ***
AGE775214    2.391e-01  2.099e-02  11.395  < 2e-16 ***
RHI225214    2.472e-01  1.679e-02  14.717  < 2e-16 ***
RHI425214    1.654e-01  4.523e-02   3.656 0.000256 ***
RHI625214   -6.308e-01  7.266e-02  -8.682  < 2e-16 ***
RHI825214   -1.854e-02  5.654e-03  -3.279 0.001043 **
EDU635213   -1.039e-01  1.284e-02  -8.093 5.82e-16 ***
EDU685213   -7.280e-02  1.396e-02  -5.214 1.84e-07 ***
HSG495213   -4.066e-06  1.345e-06  -3.022 0.002512 **
INC910213    1.925e-04  2.255e-05   8.534  < 2e-16 ***
SBO415207    5.562e-02  1.783e-02   3.120 0.001808 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3796.1  on 2801  degrees of freedom
Residual deviance: 2293.0  on 2789  degrees of freedom
AIC: 2319

Number of Fisher Scoring iterations: 7

>
```

Table 8



```
> vif(bic_model)
PST120214 AGE295214 AGE775214 RHI225214 RHI425214 RHI625214 RHI825214 EDU635213 EDU685213 HSG495213 INC910213 SBO415207
 1.834357  2.551386  2.847900  1.623365  2.061320  1.635117  2.477141  2.321642  4.726680  2.951941  4.812996  1.509831
>
```

First, I check again for multicollinearity, all results are below ten so I can say that there is no multicollinearity problem.

We see that the BIC algorithm removed two variables from the model2. The good news is that all covariates are statistically significant since the p-values for all tests are less than 0.05 except the Intercept. In case of linear regression this would be a major problem but, in our case, we fit a general linear model of logistic regression and as we try to estimate probabilities not values the intercept has no practical meaning.

Furthermore to this since the p-value of the intercept is greater than 0.05 we can assume that bo = 0 or very close to 0, this means that our model can be written as log(pi/1-pi) = bo + b1 xi . we have already assumed bo = 0 so the equation is log(pi/1-pi) = b1xi . At the end we conclude that pi = ½ so there we cannot classify it into a response based on its probability.

That is why I decide not to drop the intercept and keep it into the bic_model.

Regarding the deviance residuals, the deviance median is again very close to zero so we can say that bic_model is unbiased towards both directions of the outcomes.

Table 9

```
> with(bic_model, null.deviance - deviance) # 1514.518
[1] 1503.126
>
> with(bic_model, df.null - df.residual) # 14
[1] 12
>
> with(bic_model, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = FALSE))
[1] 8.015522e-315
>
```

We compare again with the null model in order to justify the selection of our variables we can assume from the above that the chi-square of 1514.518 with 14 degrees of freedom and an associated p-value of less than 0.005 tells us that our model as a whole fit significantly better than the null model.

Table 10

```
> exp(cbind(OR = coef(bic_model), confint(bic_model)))
Waiting for profiling to be done...
                    OR      2.5 %     97.5 %
(Intercept) 1.0086115 0.08383126 12.4144540
PST120214   1.1293278 1.08549823  1.1762196
AGE295214   1.1631041 1.10626036  1.2232618
AGE775214   1.2701596 1.21961313  1.3242442
RHI225214   1.2803924 1.24025519  1.3246797
RHI425214   1.1798279 1.08618065  1.2955952
RHI625214   0.5321601 0.45935784  0.6107748
RHI825214   0.9816332 0.97068374  0.9924736
EDU635213   0.9012832 0.87868547  0.9240799
EDU685213   0.9297894 0.90445028  0.9553508
HSG495213   0.9999959 0.99999328  0.9999986
INC910213   1.0001925 1.00014896  1.0002374
SBO415207   1.0572006 1.02266181  1.0968931
>
```

We can say again that for a one unit increase in each variable, the odds of "success" in our case Hillary Clinton win increases by the value of the OR column.

Last about bic_model the McFaden R2 too is very close to 0.4 so we can say that we have a very good fit of the data.

Table 11

```
> pR2(bic_model)
fitting null model for pseudo-r2
         llh      llhNull           G2      McFadden          r2ML          r2CU
-1146.5032659 -1898.0661390  1503.1257461     0.3959624     0.4151778     0.5595392
>
```

Both models are quite similar and have minor differences. In order to choose between them I implement the likelihood ratio test, to further test the goodness of fit of the two competing statistical models based on the ratio of their likelihoods. The best model is the one that makes the data most likely, or maximizes the likelihood function, f

Summary of the test is presented below.

Table 12

```
~/01 Business Analytics - AUEB/03 Statistics 2/project/
> anova(bic_model, model2, test = "LR") # for comparison between the two
Analysis of Deviance Table

Model 1: Response ~ PST120214 + AGE295214 + AGE775214 + RHI225214 + RHI425214 +
    RHI625214 + RHI825214 + EDU635213 + EDU685213 + HSG495213 +
    INC910213 + SBO415207
Model 2: Response ~ (PST120214 + AGE295214 + AGE775214 + SEX255214 + RHI225214 +
    RHI325214 + RHI425214 + RHI625214 + RHI825214 + EDU635213 +
    EDU685213 + VET605213 + HSG096213 + HSG495213 + INC910213 +
    BZA115213 + SBO215207 + SBO415207 + LND110210) - RHI325214 -
    VET605213 - BZA115213 - SBO215207 - LND110210
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      2789     2293.0
2      2787     2281.6  2   11.392 0.003359 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The null hypothesis is that the smaller model is the “best” model;

Since we the test provides p-value less than 0.05 we can reject the null hypothesis and assume that the larger model has a statistically significant improvement in terms of goodness of fit.

Model2 is the model of choice for best describing the 2016 Primary Democratic Party Elections. We see that the model consists of variables that describe age, education, racial and housing characteristics. It is also important to notice that one of the two variables that describe the female population characteristic was chosen to be included in the final model. I refer to the SBO415207 variable which is Female persons, percent, 2014.

Total variables of the model alongside the mapping of the full name of the variables is presented below.

Table 13

```
> model2$coefficients
  (Intercept)     PST120214     AGE295214     AGE775214     SEX255214     RHI225214     RHI425214     RHI625214     RHI825214
-2.930868e+00  1.216010e-01  1.012914e-01  1.982674e-01  9.634484e-02  2.466036e-01  1.860711e-01 -6.528141e-01 -2.501339e-02
    EDU635213     EDU685213     HSG096213     HSG495213     INC910213     SBO415207
-9.230204e-02 -7.584535e-02 -2.523762e-02 -3.998370e-06  1.885053e-04  4.937978e-02
> C
```

Table 14

column_name	description
AGE295214	Persons under 18 years, percent, 2014
AGE775214	Persons 65 years and over, percent, 2014
EDU635213	High school graduate or higher, percent of persons age 25+, 2009-2013
EDU685213	Bachelor's degree or higher, percent of persons age 25+, 2009-2013
HSG096213	Housing units in multi-unit structures, percent, 2009-2013
HSG495213	Median value of owner-occupied housing units, 2009-2013
INC910213	Per capita money income in past 12 months (2013 dollars), 2009-2013
PST120214	Population, percent change - April 1, 2010 to July 1, 2014
RHI225214	Black or African American alone, percent, 2014
RHI425214	Asian alone, percent, 2014
RHI625214	Two or More Races, percent, 2014
RHI825214	White alone, not Hispanic or Latino, percent, 2014
SBO415207	Hispanic-owned firms, percent, 2007
SEX255214	Female persons, percent, 2014

Section - Conclusion

Finally, through testing we decided to keep the more complex one. We can say that the two models were very close.to each other and the variables chosen were very similar in both cases. Among the statistically significant selected variables most of them describe racial characteristics so we can say that the votes of each community, White alone, Black, and Hispanic, was a major factor in describing the outcome that.

Finally, for the further improvement of this project, I believe that more data could be very helpful in order to have a better model.