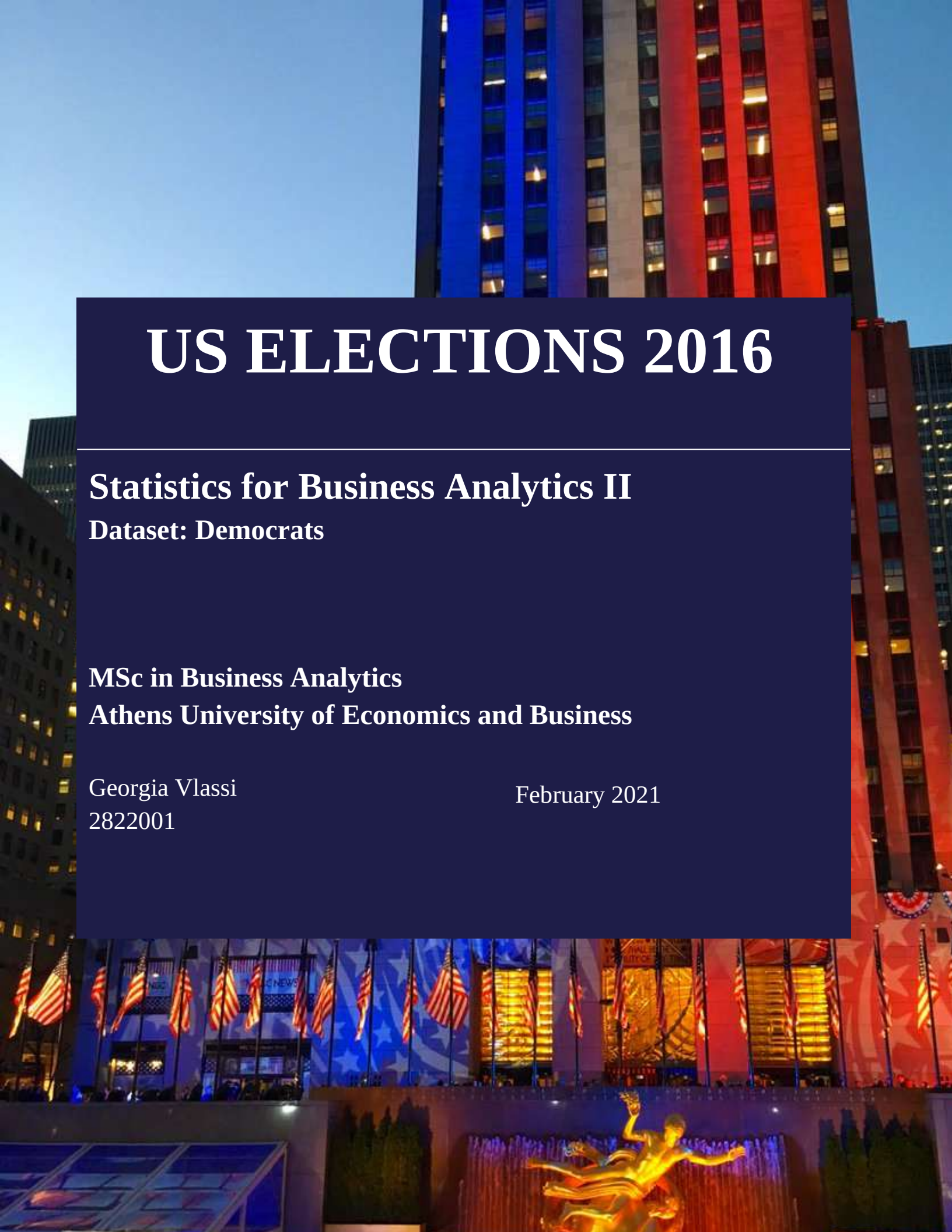

US ELECTIONS 2016

Statistics for Business Analytics II

Dataset: Democrats

MSc in Business Analytics
Athens University of Economics and Business

Georgia Vlassi
2822001

February 2021

Table of Contents

Abstract

The data of this report refer to three spreadsheets, that contain 24611 results - votes of the Democratic and Republican primaries, by US County and each presidential candidate; 3195 demographics of the US Counties that voted in the primaries and the labels of the columns in the county_facts spreadsheet. These datasets contain in summary 64 nominal, continuous and discrete variables, which focus on the votes of each candidate of all parties and the behavior of voters. In order to reduce the number of variables that describe behavior of voters precisely, Lasso and Stepwise procedure were used. The main scope of this report is to identify a good model for describing whether Clinton won over Sanders depend on Democratic voters.

1. Introduction - Description of the Problem

The main scope of this report is to identify a very good, generalized linear model, including the most important variables that precisely describe the behavior of voters regarding the Democratic candidates Hillary Clinton and Bernie Sanders in the US elections of 2016. The dataset is a result of merging three spreadsheets, which contains 58 variables, after removing duplicate variables, as fips, state and county.

Two new discrete variables have been created after pivoting the ’candidate’ variable, which contain the number of Hillary Clinton’s and Bernie Sander’s votes, respectively. There are 4 nominal variables, which contain the state, the state abbreviation, the county and the fip (unique identifier of geographic area) of each county. There are 51 Continuous variables, which describe the voter’s population type and characteristics in proportions, as their origin (e.g., Hispanic, Latino, Asian), language, percentage of age, level of education etc. Finally, a new categorical variable CoS (Clinton over Sanders) has been created, which will be used as the response variable to the generalized linear model.

Prior to the analysis of our model, there are a few transformations that should be implemented. As we import the datasets, we observe that there are duplicate variables after merging, like state and state abbreviation. There are also null values that should be handled. After analysis and data modifications Lasso and Stepwise procedures were performed, to find the 16 most accurate variables that describe Hillary Clinton’s win over Bernie Sanders, using as explanatory variables the socio-economic characteristics of the counties.

After applying logistic regression model, which will be interpreted further in our report ,we end up with 16 attributes. These attributes describe the overall characteristics of the voters, education level, origin, income, language, sex and age.

2. Descriptive analysis

2.1 Data Cleansing

In order to manipulate our data, the R language was used. After importing the three datasets, there were only numeric and character classes of variables.

During Data Cleansing procedure, all null (NA) values must be manipulated. At Tables 1 and 2 there are listed all null values of every variable.

state	state_abbreviation	county	fips	party	candidate	votes	fraction_votes
0	0	0	100	0	0	0	0

Table 1 NAs in votes spreadsheet

fips	area_name	state_abbreviation	PST045214	PST040210	PST120214	POP010210	AGE135214
0	0	52	0	0	0	0	0
AGE295214	AGE775214	SEX255214	RHI125214	RHI225214	RHI325214	RHI425214	RHI525214
0	0	0	0	0	0	0	0
RHI625214	RHI725214	RHI825214	POP715213	POP645213	POP815213	EDU635213	EDU685213
0	0	0	0	0	0	0	0
VET605213	LFE305213	HSG010214	HSG445213	HSG096213	HSG495213	HSD410213	HSD310213
0	0	0	0	0	0	0	0
INC910213	INC110213	PVY020213	BZA010213	BZA110213	BZA115213	NES010213	SBO001207
0	0	0	0	0	0	0	0
SBO315207	SBO115207	SBO215207	SBO515207	SBO415207	SBO015207	MAN450207	WTN220207
0	0	0	0	0	0	0	0
RTN130207	RTN131207	AFN120207	BPS030214	LND110210	POP060210		
0	0	0	0	0	0		

Table 2 NAs in county_facts spreadsheet

2.2 Data Transformation

During Data Transformation procedure variables with null values were handled respectively. The null values of numeric variable ‘fips’ in votes spreadsheet were updated to 0, and the null values of character variable ‘state_abbreviation’ in county_facts spreadsheet were removed, as they were indicating the summary metrics per state.

A new dataset was created, including only Democratic party candidates Hillary Clinton and Bernie Sanders. The new dataset was joined with county_facts dataset through variable ‘fips’. The variable ‘area_name’ of county_facts dataset was renamed ‘county’. Also, duplicate variables like ‘state’ and

'county' were removed. The variable 'party' was removed as it would be a surplus to indicate that H. Clinton and B. Sanders are Democrats. This dataset will be used for the Explanatory Data Analysis. The main scope of analysis is to find if Hillary Clinton won over Bernie Sanders. A dataset was created in order to build a good logistic regression model, which will describe the most accurate characteristics of Hillary Clinton's voters. Before merging the datasets mentioned, the votes dataset was updated by pivoting the votes of two candidates. Instead of variable 'votes', two new variables occurred, which contain the number of votes per candidate. The variable 'fraction_votes' and the ties between the two candidates were removed from this dataset. Our response categorical variable is 'CoS', which is a binary attribute where value 1 indicates that H. Clinton won regarding the votes in a state. The aforementioned data frame will be used for both for the Explanatory Data Analysis and the Descriptive models.

3. Exploratory data analysis

Before implementing different methods to select the most accurate variables for our logistic model, it is important to understand the data of the two candidates.

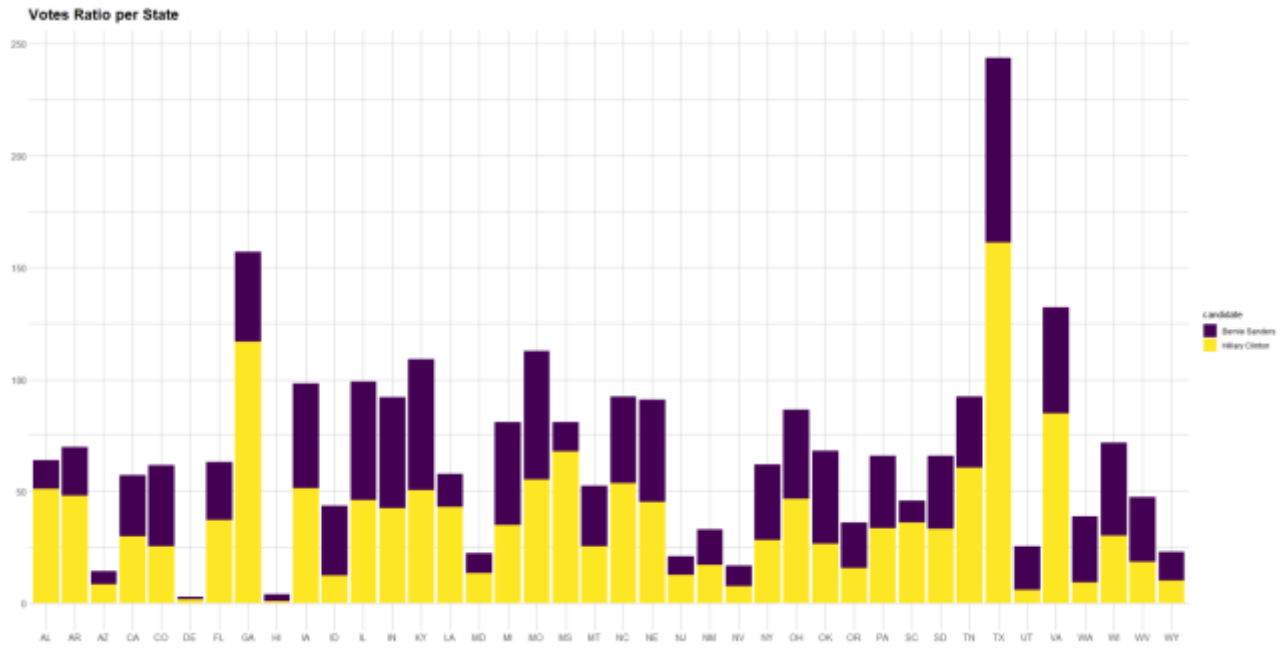

Figure 1 Votes of candidates per state

In Figure 1, a bar plot has been constructed by allocating the fraction of votes per state. We observe that in Alabama, Arkansas, Georgia, Mississippi and Texas there is an overwhelming difference of votes in favor of Hillary Clinton. Hawaii, Utah and Washington voted tremendously in favor of Bernie Sanders.

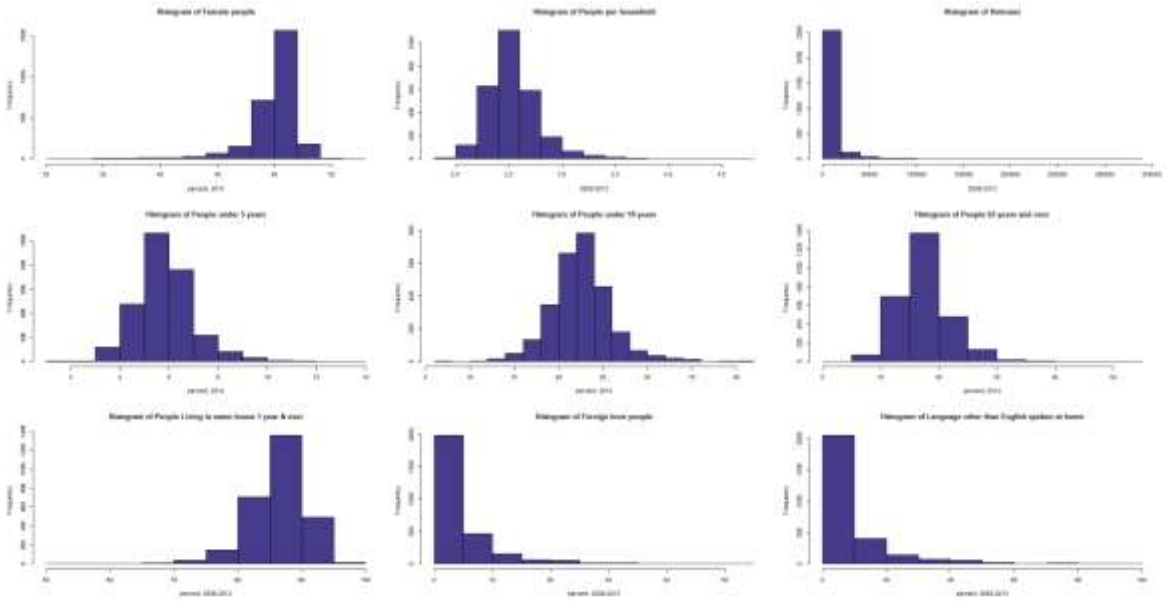

Figure 2 Histograms showing attributes of voters

In Figure 2, different histograms have been constructed to describe the attributes of US voter's population. On these histograms we observe that a percentage of 52% of the voters were women, also the voters live in dwellings with more than 2 people and a very small percentage of them are veterans. In the United States the percentage of people under 18 years of age is almost 25% with the percentage of people over 65 years of age being 20%. In addition, the greater percentage of voters are native people who live in the same house for more than a year and speak mostly English.

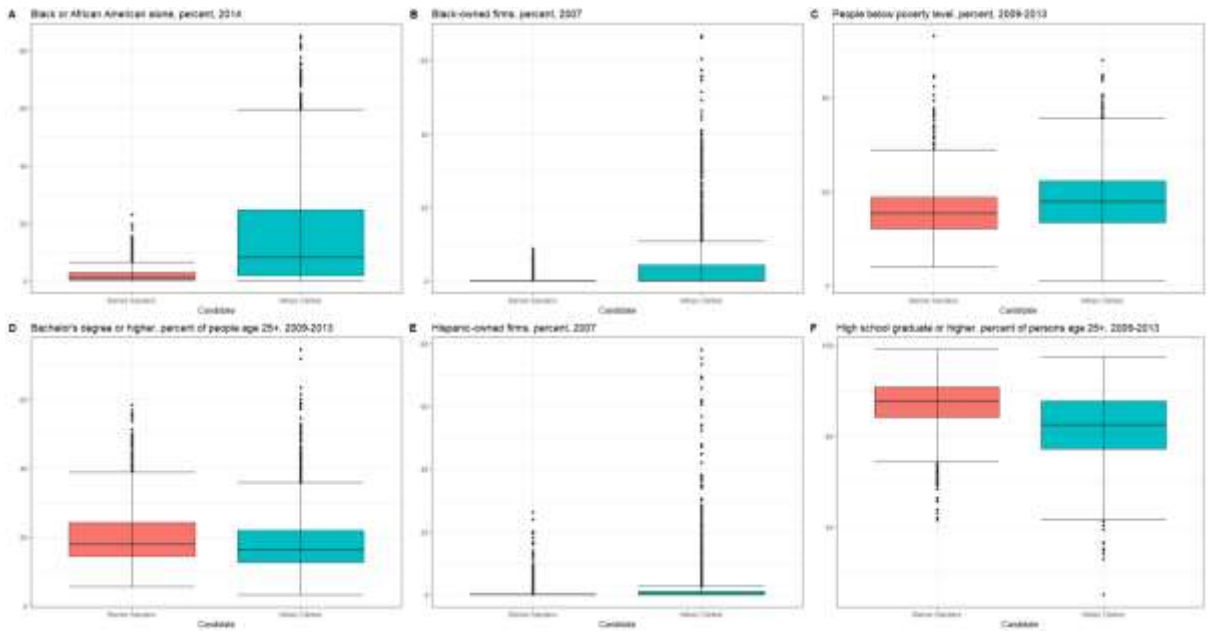

Figure 3 Box Plots showing voters preferences

In Figure 3, box plots were created to show the preferences within different categories of voters. More specifically, it is clear that Hillary Clinton won the counties with a high Black or African population, as also counties with Black or Hispanic population that own a firm. Counties with High school graduated or bachelor's degree population showed a preference in Bernie Sanders. Finally, it is remarkable that counties with population below the poverty level preferred Hillary Clinton over Bernie Sanders.

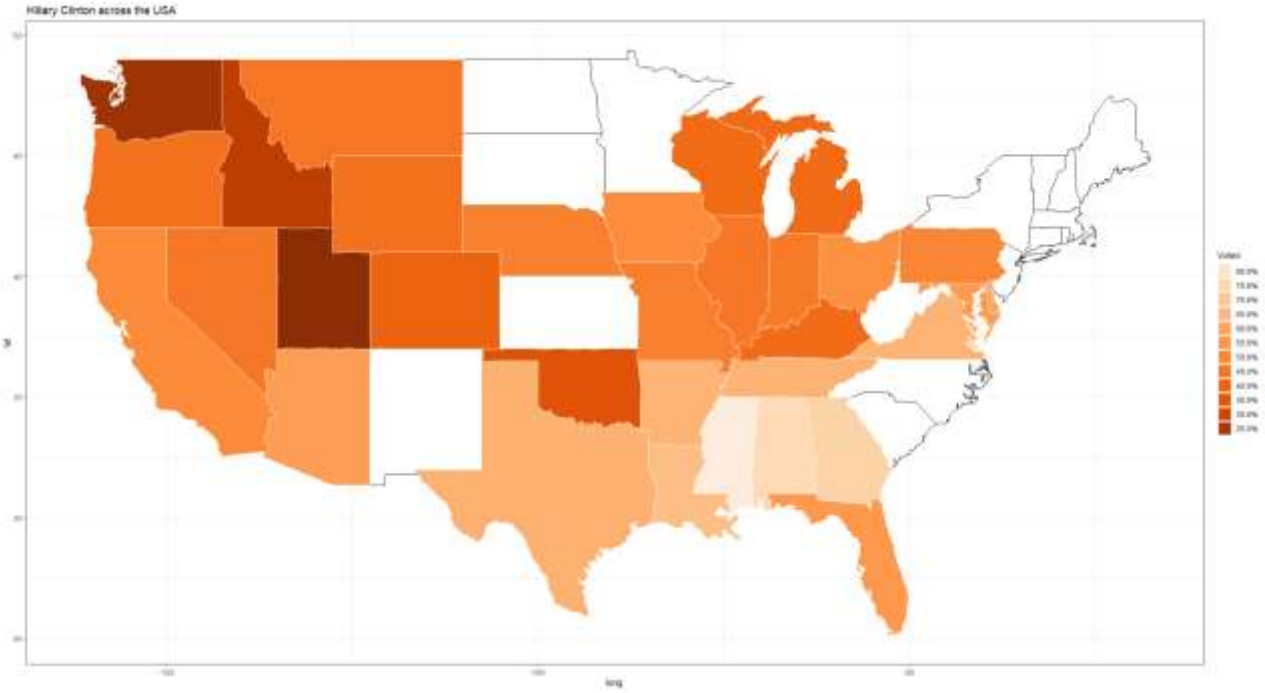

Figure 4 Map with Hillary Clinton's votes across the states

In Figure 4, we observe the performance of Hillary Clinton against Bernie Sanders. In midwest states H. Clinton gained more votes in contrast to Washington, Idaho and Utah where B. Sanders outmatched the competition in Democratic party.

4. Descriptive models

4.1 Select Attributes

Only the numeric variables were used as input in the logistic regression model, using the glm (generalized linear model) function. The LASSO method was used, due to its feature selection technique, that can remove the appropriate variables without much loss of information. It also performs better on large datasets. The covariates with p-value > 0.05 should be later removed from our model. In the summary of a model, the p-value indicates if we should reject or accept the H0 (null) hypothesis, that the specific attribute could be equal to 0. Further analysis and interpretation of coefficients will be

later in our final model. In the beginning, we implement Lasso with the coefficients against the log-lambda value, as shown in Figure 5.

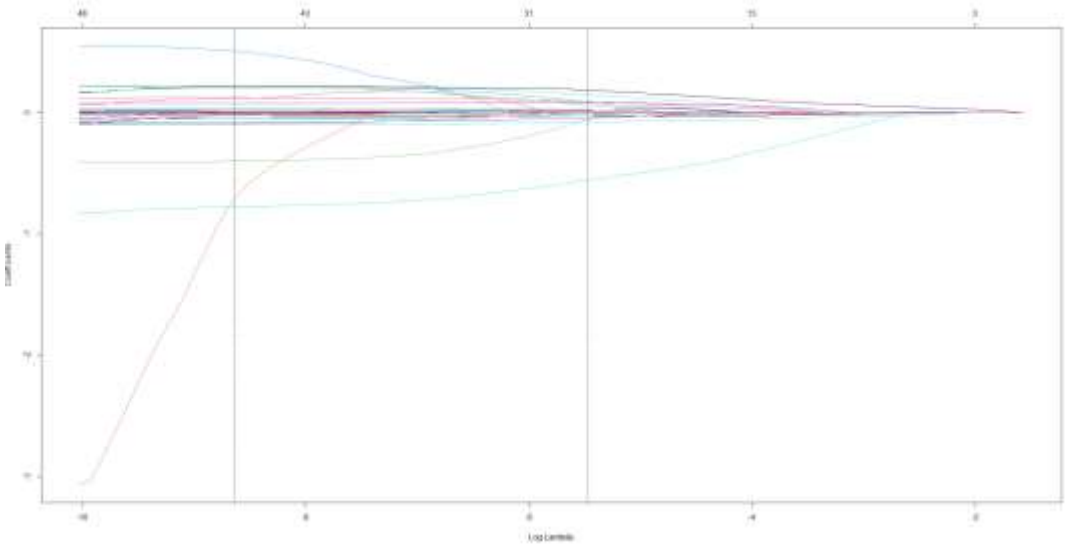

Figure 5 Coefficients against the log-lambda

We used cross validation to find a reasonable value for lambda. To find explicitly the selected optimal values of λ check Figure 6, where the λ at which the minimal MSE was achieved is 0.0042179 and the most regularized model whose mean squared error was within one standard error of the minimal is 0.0001783. From Lasso we chose the coefficients produced from lambda.1se, as it has fewer variables while it does not differ significantly from the model derived from lambda.min in terms of Mean Square Error. The characteristics produced from Lasso are shown at Table 3 and would be used as input to Stepwise procedure.The summary of the model produced from Lasso is dispayed in Table 4.

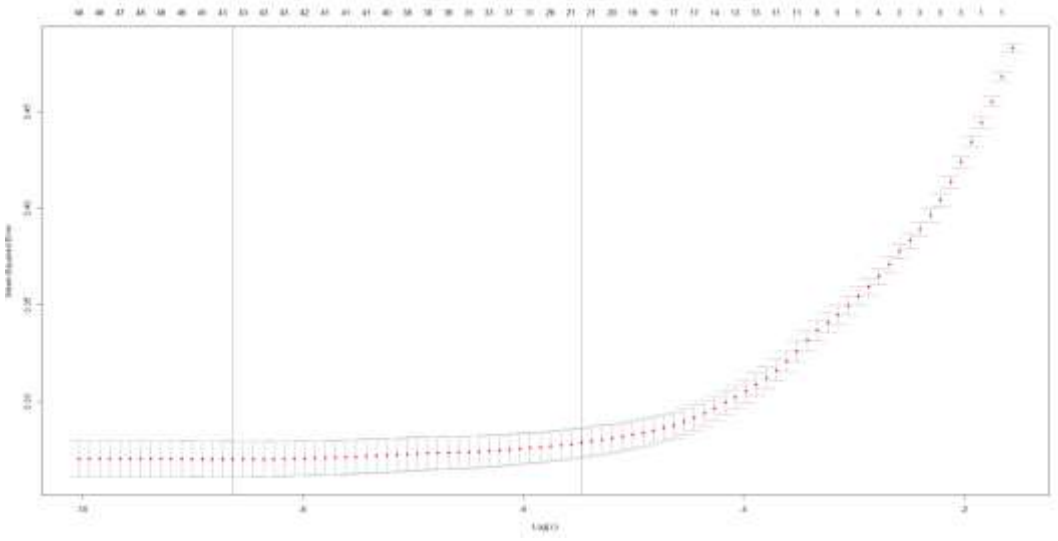

Figure 6 Log-lambda against Mean Square Error

```
 [1] "(Intercept)" "PST120214"   "AGE135214"   "AGE295214"   "AGE775214"   "SEX255214"   "RHI225214"   "RHI325214"   "RHI425214"   "RHI625214"   "RHI825214"
[12] "EDU635213"   "EDU685213"   "VET605213"   "HSG445213"   "HSG096213"   "HSG495213"   "INC910213"   "BZA115213"   "SBO215207"   "SBO415207"   "LND110210"
```

Table 3 Characteristics produced from Lasso

```
Call:
glm(formula = CoS ~ PST120214 + AGE135214 + AGE295214 + AGE775214 +
    SEX255214 + RHI225214 + RHI325214 + RHI425214 + RHI625214 +
    RHI825214 + EDU635213 + EDU685213 + VET605213 + HSG445213 +
    HSG096213 + HSG495213 + INC910213 + BZA115213 + SBO215207 +
    SBO415207 + LND110210, family = binomial(link = "logit"),
    data = elections_num)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9623  -0.7537   0.0227   0.6170   3.5964

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.116e+00  1.896e+00   0.589 0.555947
PST120214    1.279e-01  2.170e-02   5.892 3.82e-09 ***
AGE135214   -3.946e-01  1.060e-01  -3.724 0.000196 ***
AGE295214    2.122e-01  4.190e-02   5.065 4.08e-07 ***
AGE775214    2.047e-01  2.458e-02   8.330  < 2e-16 ***
SEX255214    9.051e-02  3.314e-02   2.731 0.006315 **
RHI225214    2.432e-01  1.893e-02  12.851  < 2e-16 ***
RHI325214   -8.288e-03  1.074e-02  -0.771 0.440491
RHI425214    1.437e-01  6.201e-02   2.317 0.020489 *
RHI625214   -7.249e-01  8.020e-02  -9.039  < 2e-16 ***
RHI825214   -2.593e-02  7.813e-03  -3.319 0.000903 ***
EDU635213   -8.824e-02  1.419e-02  -6.218 5.04e-10 ***
EDU685213   -8.033e-02  1.500e-02  -5.356 8.52e-08 ***
VET605213    8.950e-06  6.856e-06   1.305 0.191745
HSG445213   -5.563e-02  1.319e-02  -4.216 2.48e-05 ***
HSG096213   -4.797e-02  1.406e-02  -3.413 0.000644 ***
HSG495213   -4.398e-06  1.401e-06  -3.139 0.001697 **
INC910213    2.012e-04  2.444e-05   8.233  < 2e-16 ***
BZA115213    1.112e-02  9.823e-03   1.132 0.257515
SBO215207    1.092e-02  6.451e-02   0.169 0.865562
SBO415207    4.708e-02  1.939e-02   2.427 0.015207 *
LND110210   -2.876e-05  4.418e-05  -0.651 0.515040
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3743.8  on 2768  degrees of freedom
Residual deviance: 2221.2  on 2747  degrees of freedom
AIC: 2265.2

Number of Fisher Scoring iterations: 7
```

Table 4 Summary of model after Lasso

The mathematical formulation of model in Table 4 is shown below:

logit(CoS) = 1.116e+00 + PST120214*1.279e-01 - AGE135214*3.946e-01 + AGE295214*2.122e-01 + AGE775214*2.047e-01 + SEX255214*9.051e-02 + RHI225214*2.432e-01 - HI325214*8.288e-03 + RHI425214*1.437e-01 - RHI62521*7.249e-01 - RHI825214*2.593e-02 - EDU635213*8.824e-02 - EDU685213*8.033e-02 + VET605213*8.950e-06 - HSG445213*5.563e-02 - HSG096213*4.797e-02 - HSG495213*4.398e-06 + INC910213*2.012e-04 + BZA115213*1.112e-02 + SBO215207*1.092e-02 + SBO415207*4.708e-02 - LND110210*2.876e-05

4.2 Interpretation of the model

Succeeding the Lasso, Stepwise procedure was conducted with 21 variables as input. Stepwise procedure was selected as an extra method to select variables, which double checks the variables by adding or removing step by step covariates. The AIC is an estimate of a constant plus the relative

distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means that a model is considered to be closer to the truth. The AIC and BIC differ by the way they penalize the number of parameters of a model. More precisely, BIC criterion induces a higher penalization for models with an intricate parametrization in comparison with AIC criterion. The summaries of the models that occurred from AIC and BIC are almost the same, with AIC having one extra covariate, the population of Veterans, which is not statistically significant. As we have a small dataset with almost 3000 rows, the values of AIC and BIC are almost equal. We ended up with a model with 16 covariates, where BIC has its minimum value, as shown in Table 5.

```
Call:
glm(formula = CoS ~ PST120214 + AGE135214 + AGE295214 + AGE775214 +
    SEX255214 + RHI225214 + RHI425214 + RHI625214 + RHI825214 +
    EDU635213 + EDU685213 + HSG445213 + HSG096213 + HSG495213 +
    INC910213 + SBO415207, family = binomial(link = "logit"),
    data = elections_num)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9252  -0.7523   0.0223   0.6201   3.5695

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.707e-01  1.841e+00   0.310 0.756625
PST120214    1.325e-01  2.137e-02   6.200 5.65e-10 ***
AGE135214   -4.108e-01  1.056e-01  -3.890 0.000100 ***
AGE295214    2.157e-01  4.185e-02   5.155 2.54e-07 ***
AGE775214    2.059e-01  2.443e-02   8.429  < 2e-16 ***
SEX255214    9.329e-02  3.237e-02   2.882 0.003958 **
RHI225214    2.529e-01  1.761e-02  14.356  < 2e-16 ***
RHI425214    1.806e-01  4.835e-02   3.736 0.000187 ***
RHI625214   -7.195e-01  7.874e-02  -9.137  < 2e-16 ***
RHI825214   -2.182e-02  6.291e-03  -3.468 0.000524 ***
EDU635213   -9.216e-02  1.344e-02  -6.858 6.96e-12 ***
EDU685213   -8.212e-02  1.489e-02  -5.514 3.51e-08 ***
HSG445213   -5.333e-02  1.304e-02  -4.089 4.34e-05 ***
HSG096213   -4.391e-02  1.380e-02  -3.181 0.001468 **
HSG495213   -4.482e-06  1.389e-06  -3.227 0.001253 **
INC910213    2.073e-04  2.376e-05   8.725  < 2e-16 ***
SBO415207    5.540e-02  1.824e-02   3.038 0.002383 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3743.8  on 2768  degrees of freedom
Residual deviance: 2225.9  on 2752  degrees of freedom
AIC: 2259.9

Number of Fisher Scoring iterations: 7
```

Table 5 Summary of final model using BIC

VIF was used to detect whether multicollinearity exists in our regression model. It measures how much the variance (or standard error) of the estimated regression coefficient is inflated due to collinearity. If VIF (see Table 6) is lower than 10 we accept the coefficient, as it does not disrupt collinearity.

PST120214	AGE135214	AGE295214	AGE775214	SEX255214	RHI225214	RHI425214	RHI625214	RHI825214	EDU635213	EDU685213	HSG445213	HSG096213	HSG495213	INC910213	SBO415207
1.91	5.20	6.57	3.71	1.52	1.71	2.34	1.79	3.06	2.53	5.26	2.77	3.62	3.10	5.21	1.56

Table 6 VIF of covariates

From Table 5 we understand that deviance residuals are a measure of model fit. As the median deviance residual is close to zero (= 0.0223), this means that our model is not biased in one direction. This part of the output shows the distribution of the deviance residuals for individual cases used in the model. The next part of the output shows the coefficients, their standard errors, Wald z-statistic, and the associated p-values. All covariates are statistically significant, as each has p-value < 0.05.

The mathematical formulation of the model in Table 5 is shown below:

logit(CoS) = 5.707e-01 + PST120214*1.325e-01 - AGE135214*4.108e-01 + AGE295214*2.157e-01 + AGE775214*2.059e-01 + SEX255214*9.329e-02 + RHI225214*2.529e-01 + RHI425214*1.806e-01 - RHI625214*7.195e-01 - RHI825214*2.182e-02 - EDU635213*9.216e-02 - EDU685213*8.212e-02 - HSG445213*5.333e-02 - HSG096213*4.391e-02 - HSG495213*4.482e-06 + INC910213*2.073e-04 + SBO415207*5.540e-02

The logistic regression coefficients affect the change in the log odds of the outcome for a one unit increase in the response variable. The odds indicate the estimated probability of voting Hillary Clinton over Bernie Sanders. The signs '-' and '+' before each covariate indicate how much would the increase or decrease of the probability of voting H. Clinton be. For example, for every one unit change in PST120214 (Population, percent change - April 1, 2010 to July 1, 2014), the log odds of voting Hillary Clinton (versus voting B. Sanders) increase by 0.01325, having all other covariates constant. On the other hand, for every single unit change in AGE135214 (People under 5 years, percent, 2014)), the log odds of voting Hillary Clinton (versus voting B. Sanders) decrease by 0.4108, when all other covariates are constant. The positive intercept indicates that when all covariates are zero it is 0.5707 times more possible to vote H. Clinton.

Considering the mathematical formulation of the model that occurred from Lasso and comparing it with the model occurred after Stepwise, we observe that the odds of voting H. Clinton have been less affected. For example, for every single unit change in PST120214 (Population, percent change - April 1, 2010 to July 1, 2014), the log odds of voting Hillary Clinton (versus voting B. Sanders) increase by 0.01279. As a result, we chose as a better model to continue our analysis with, the one that occurred after Stepwise procedure, because it is not over parameterized, has less covariates all being statistically significant at 5% level and the covariates affect more the odds of voting H. Clinton.

Deviance is a measure of goodness of fit for a generalized linear model. The null deviance shows how well the response variable 'CoS' is predicted by a model that only includes the intercept. More specifically, we have a value of 3743.8 on 2768 degrees of freedom. Adding the explanatory independent variables (PST120214, AGE135214, AGE295214, AGE775214, SEX255214, RHI225214, RHI425214, RHI625214, RHI825214, EDU635213, EDU685213, HSG445213, HSG096213, HSG495213, INC910213, SBO415207), decreased the deviance to 2225.9 points on 2752 degrees of freedom, a significant reduction in deviance, which is good as it is close to 1. Another method to count the goodness of fit is by using PseudoR2 statistics, as shown in Table 7. The McFadden's pseudo indicates that this model is 40% better than the model including only the intercept (null model). This was also confirmed by the comparison of the deviances between them, which is great and equal to 1517 points.

McFadden	McFaddenAdj	CoxSnell	Nagelkerke	AldrichNelson	VeallZimmermann	Efron	McKelveyZavoina	Tjur	AIC
0.4054437	0.3963620	0.4219974	0.5692755	0.3540790	0.6159636	0.4464480	0.8516613	0.4453177	2259.9028202
BIC	logLik	logLik0	G2						
2360.6489261	-1112.9514101	-1871.9034236	1517.9040270						

Table 7 PseudoR2 statistics

5. Conclusions and Discussion

The main scope of this analysis was to construct a good, logistic regression model, which will include the most accurate socio-economic characteristics of the counties that affected Hillary Clinton's win. We ended up with a model of 16 variables that are correlated to population's percentage change from 2010 to 2014, percentage of people under 5 years of age, percentage of people under 18 years of age, percentage of people above 65 years of age, female's population percentage, Black/African/Asian population percentages, percentage of people coming from two or more races, percentage of White people without including Hispanic or Latino, High school graduated or bachelors' degree population,

to the capita money income of the last 12 months, Hispanic people who own a firm, the rate of homeownership and finally to the percentage of housing units in multi-unit structures.

Although a good logistic regression model was constructed in terms of goodness of fit, it can be improved by using a bigger dataset, or by transforming our data differently. Last but not least, a different procedure of selecting variables could lead to a better fitted model.