COLLEGE OF INFORMATION STUDIES

PROJECT REPORT

INST737: Digging into Data

Analysing 2016 US Presidential Elections

Submitted by:

Janhavi Chitale

Kanishka Ramamoorthy

Akshata Salehittal

Table of Contents

1. ABSTRACT

The ability to foresee the outcome of an action can highly influence important business decisions. Predictive analytics possess the capability to anticipate the likelihood of a result and provide insight into what will happen in the future based on what has happened in the past. Organizations can use predictors to beat their rival markets and gain competitive advantage. Predictive analysis comes into play when the data in hand is so large than human insight alone cannot be applied to predict the outcome. Machine learning algorithms are the preliminary start points of predictive analysis. This is the stage where the model is trained using available past data to be able to predict the outcome of similar data in the future. The models gain knowledge from fed data and accustom themselves to fit unseen data in the future and generate stable outcomes. Machine learning algorithms possess the ability to consume large volumes of data and apply statistical algorithms on them to analyze hidden trends that could have skipped the human eye. [1][2]

2. INTRODUCTION

2.1 Motivation

Predictive analysis can serve as an important factor in several decision making processes in the elections and can significantly sway the results of the elections. Envisioning the possible results of a state can be a major decisive tool in setting political funds. Candidates can determine which states require higher investment in campaigns based on assessment of the possible support they would receive from the states. A candidate who has an extremely low chance of winning in a state would rather invest his money in some other state which would provide him a better chance of winning. Predictions can especially be helpful in states with open primaries which highly influence the final results. The presence of independent voters in these states makes it extremely hard to interpret the winner of those states. Demographics and economic status of citizens in a state can reveal a lot about the results that are likely to be obtained from those states. Hence, candidates can make use of this information to gain competitive advantage and to re-direct their funds and resources in the right direction.[3]

2.2 Objective

As part of the project we thoroughly analyzed the dataset, ran descriptive statistics of the crucial attributes in the dataset and produced effective visual representations, to explain variations in the results of the primary elections based on demographics and socio-economic attitudes of citizens in each county. Visual analysis was designed to verify if the winning trends of the Republican candidates differ from that of the Democratic candidates and to analyze whether the same demographics affect both the parties equally or not. The visuals also aimed at performing a comparison of individual candidates' results and recording similarities in winning patterns of two or more candidates.

Our model has been built to predict the results of Primaries for counties where ballots have not yet been conducted. This can be extended to a larger scale and enable prediction of the final presidential nominee from the Republican & the Democratic parties by aggregating the results of the counties, state-wise.

2.3 Election Process - Primaries

The Primary election is the first stage of the U.S Presidential election process, the results of which are used to decide the final presidential nominee from each party. Primaries are divided into three types

1. Open Primary
2. Closed Primary
3. Semi-Closed Primary

Citizens in the U.S need to register themselves as Republican, Democratic or Independent voters. In states holding Closed Primaries, the registered Republican voters can vote for a Republican candidate, registered Democratic voters can vote for a Democratic candidate and the Independent voters do not have the opportunity to place their votes. Semi-Closed Primary works in a similar fashion with the only difference that independent voters have the choice to vote for either a Republican or Democratic candidate of their choice. States conducting open primary offer much more flexibility to their citizens as any voter can vote in any party's primary.

3. DATASET

We obtained our dataset "2016 US Election Kaggle dataset" from the Kaggle competition online. Our dataset contained 2 main csv files. The "primary_results.csv" contained the primary results for the 28 states in which the primaries have been conducted (as of 6th March 2016 when we downloaded the dataset). Each row contains details of the number of votes each Republican and Democratic candidate had gained in that particular county. There were 11 republican candidates and 3 Democratic candidates contesting in IOWA where the Primaries were conducted first. Through the course of the election candidates from both the parties dropped out and few states had only 5 Republican candidates and 2 Democratic candidates as of 6th March 2016.The "county_facts.csv" file obtained from US census has 51 attributes that contain detailed demographic information about all the counties in the U.S including the ones where Primaries have not yet been conducted. The "county_facts_dictionary.csv" served as a dictionary to explain what each attribute name actually represents. The demographics information were from 2014 and most of the important attributes such as race and education were represented in terms of percentages. Additionally there was a directory containing county shapefiles that aided in the process of visualizing our dataset based on counties.

4. Exploratory Data Analysis

Exploratory data analysis assists in summarising the characteristics of the data in hand with the aid of visual methods. This helps in identifying the characteristics, i.e. Independent Variables which significantly affect the prediction of our outcome variable.

We decided to perform visual analysis for the two parties separately so as to identify any trends or patterns which could aid in building our predictive model.

4.1 Descriptive statistics

4.1.1 Republican Descriptive Statistics

4.1.1.1 Iowa Results

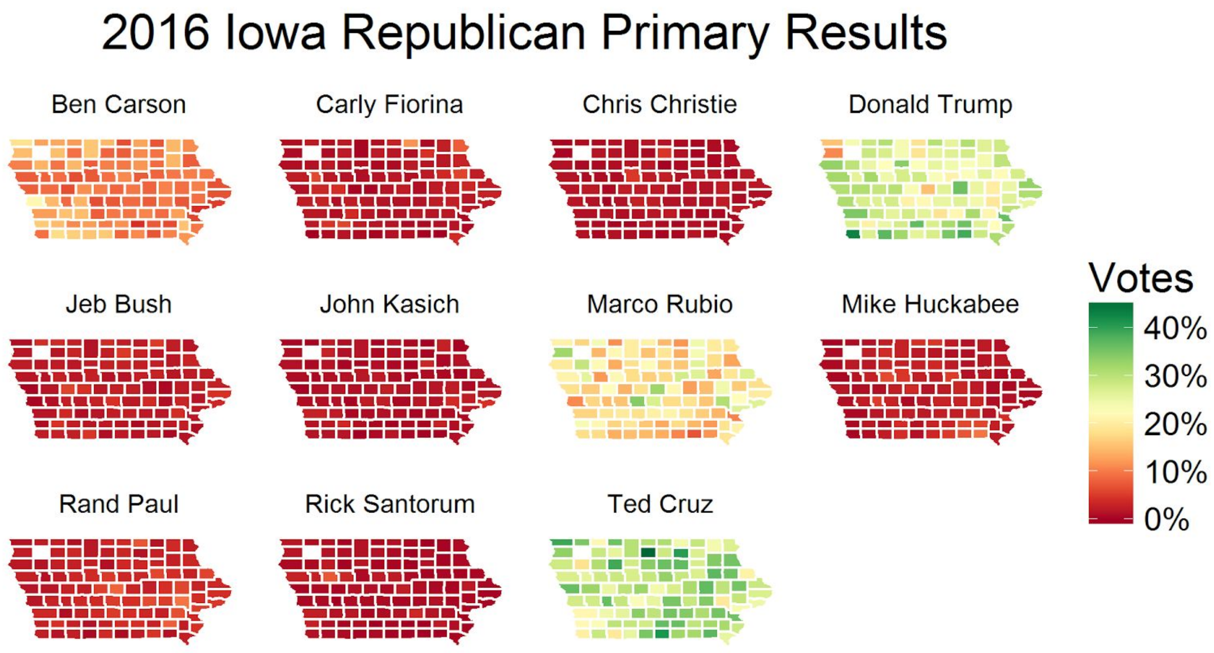


Fig Iowa Republican Results

Iowa was the first of the 50 states in the US to conduct the primary elections. By analysing the primary results in IOWA we gained insight into the total number of Republican candidates that were present at the start of the election process along with the percentage of the votes they received. It was observed that most of the candidates who received about 0-10% votes dropped out of the elections in the later stages.

4.1.1.2 Republican Candidate winners displayed state wise

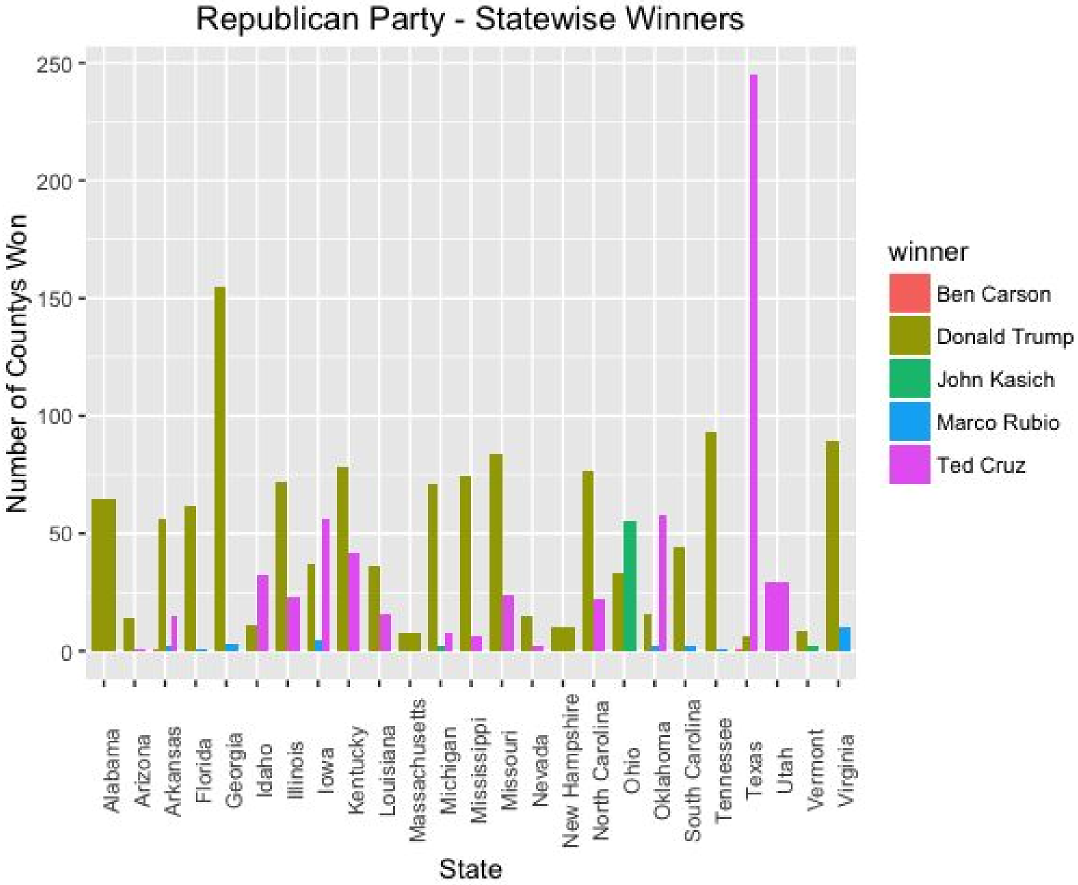


Fig: Republican Candidate winners displayed state wise

The above graph shows the number of counties won by each of the Republican candidates in each state. This helped us identify the candidates with highest support and recognize the main competitors in the Republican party.

4.1.2 Democrat Descriptive Statistics

4.1.2.1 Iowa Results

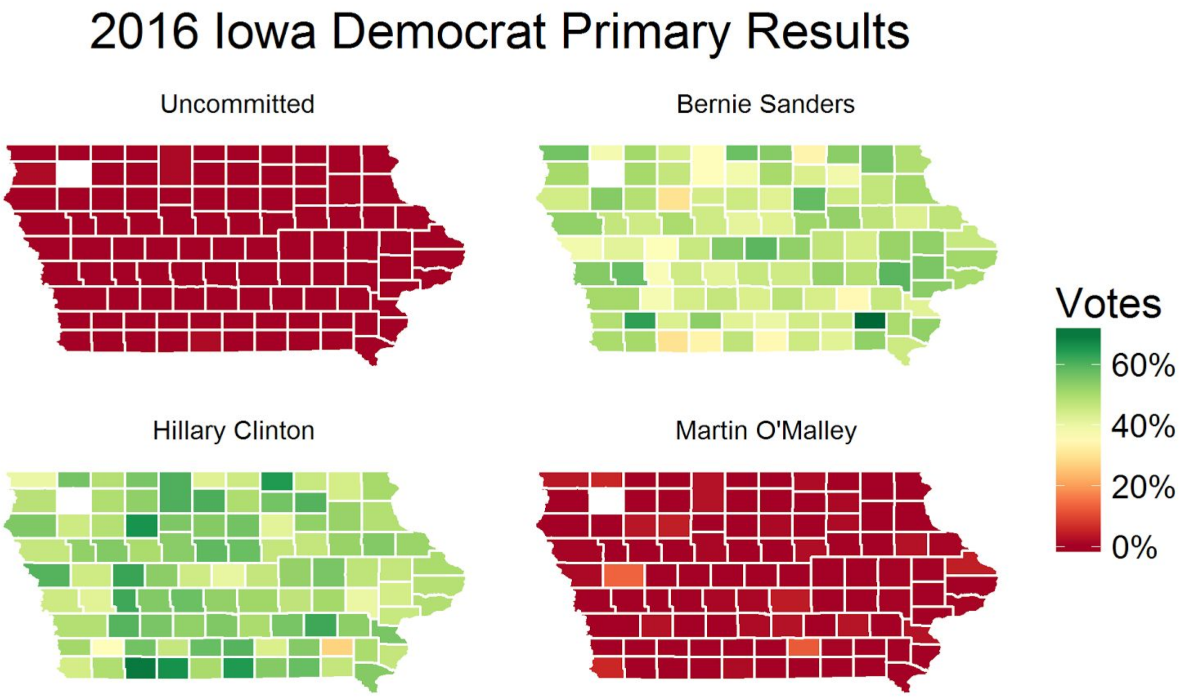


Fig:Iowa Democrat Results

This figure gave insights into the two major competitors from the Democrat party.

4.1.2.2 Democratic Candidate winners displayed state wise

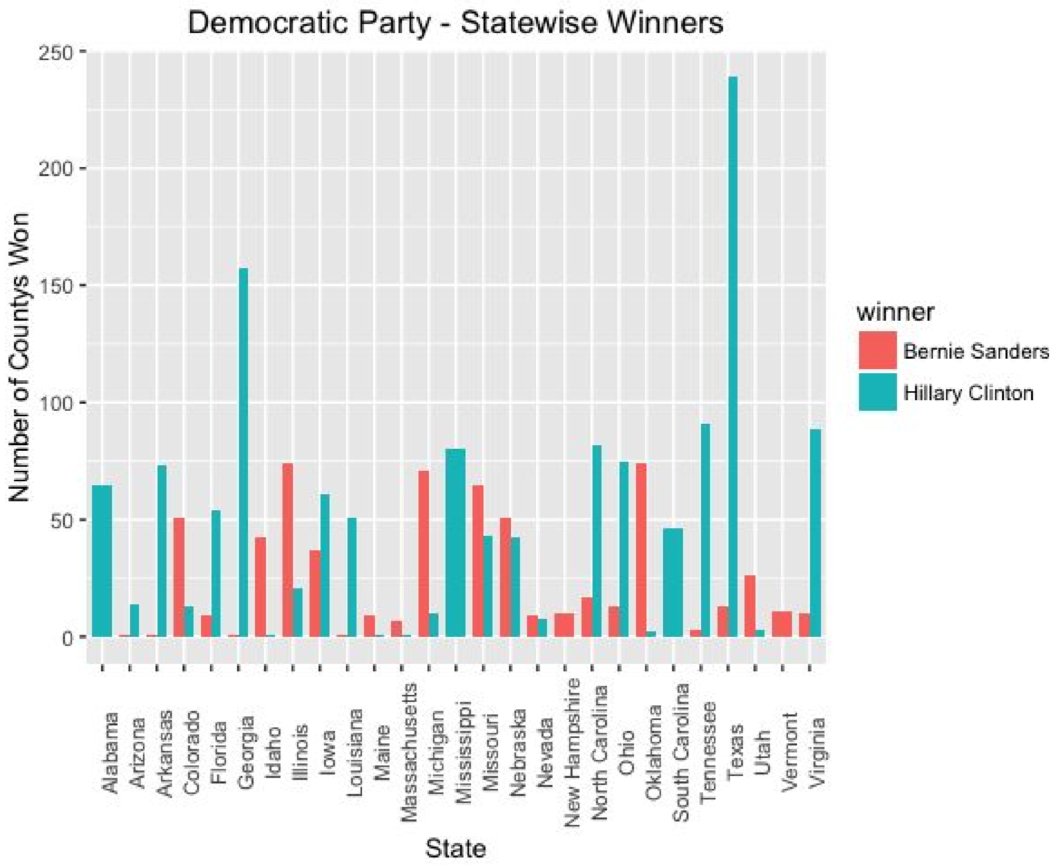


Fig: Democrat Candidate winners displayed state wise

As seen from this graph, the number of counties won by the two candidates state wise do not vary in large numbers except for in a few states, which implies that the chances of a particular candidate winning in a state cannot be determined by this graph and requires further analysis.

4.2 Analyzing Trends between Republicans and Democrats

4.2.1 Republicans

4.2.1.1 County wise winners by African American population and Median Household Income

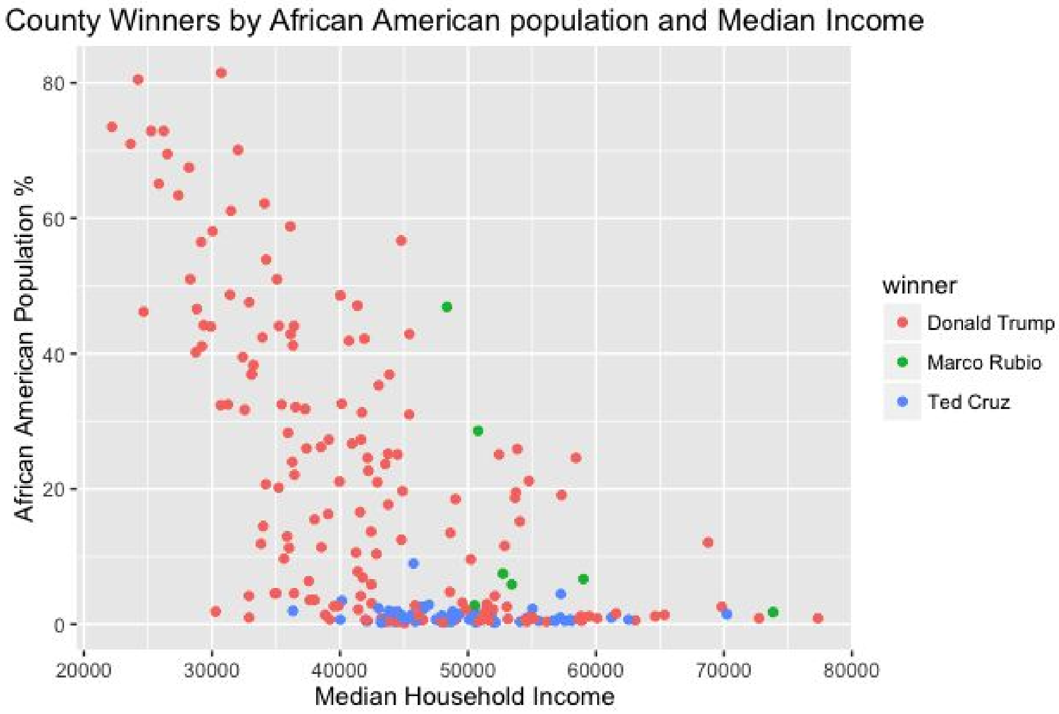


Fig: County wise Republican winners by African American population and Median Household Income

The figure above shows that both Marco Rubio and Ted Cruz are not very popular among the African American population.

4.2.1.2 County wise winners by Poverty level and Education Attainment

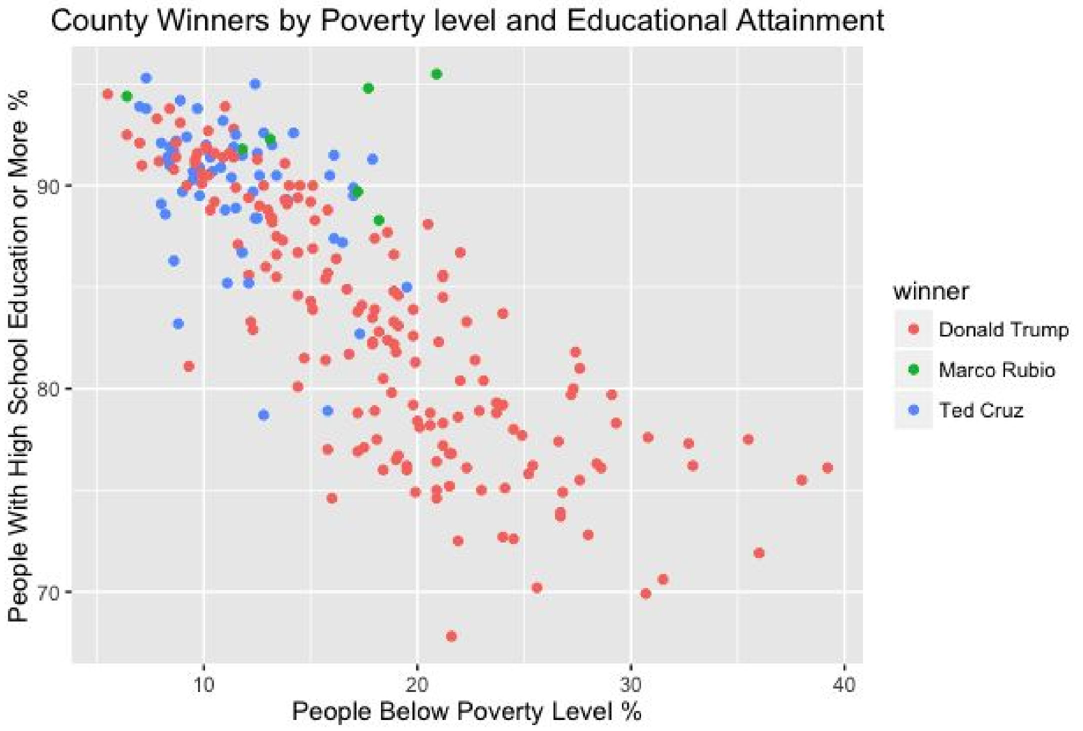


Fig: County wise Republican Candidate winners Poverty level and Education Attainment

This figure reveals that counties with low education levels and higher poverty levels tend to support only Donald Trump.

4.2.2 Democrats

4.2.2.1 County wise winners by African American population and Median Household Income

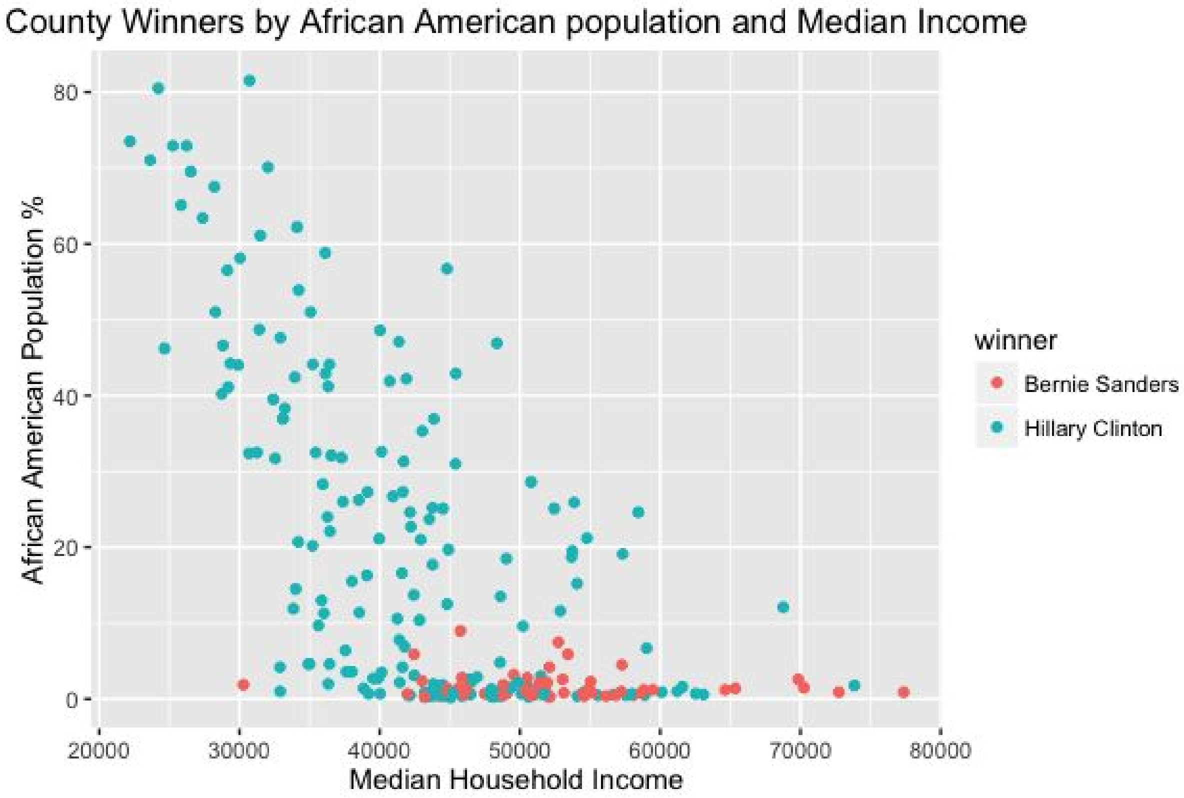


Fig: County wise Democrat winners by African American population and Median Household Income

The graph shows that Bernie Sanders was not popular among the African American population but was popular among people with an average Median household income.

4.2.2.2 County wise winners by Poverty level and Education Attainment

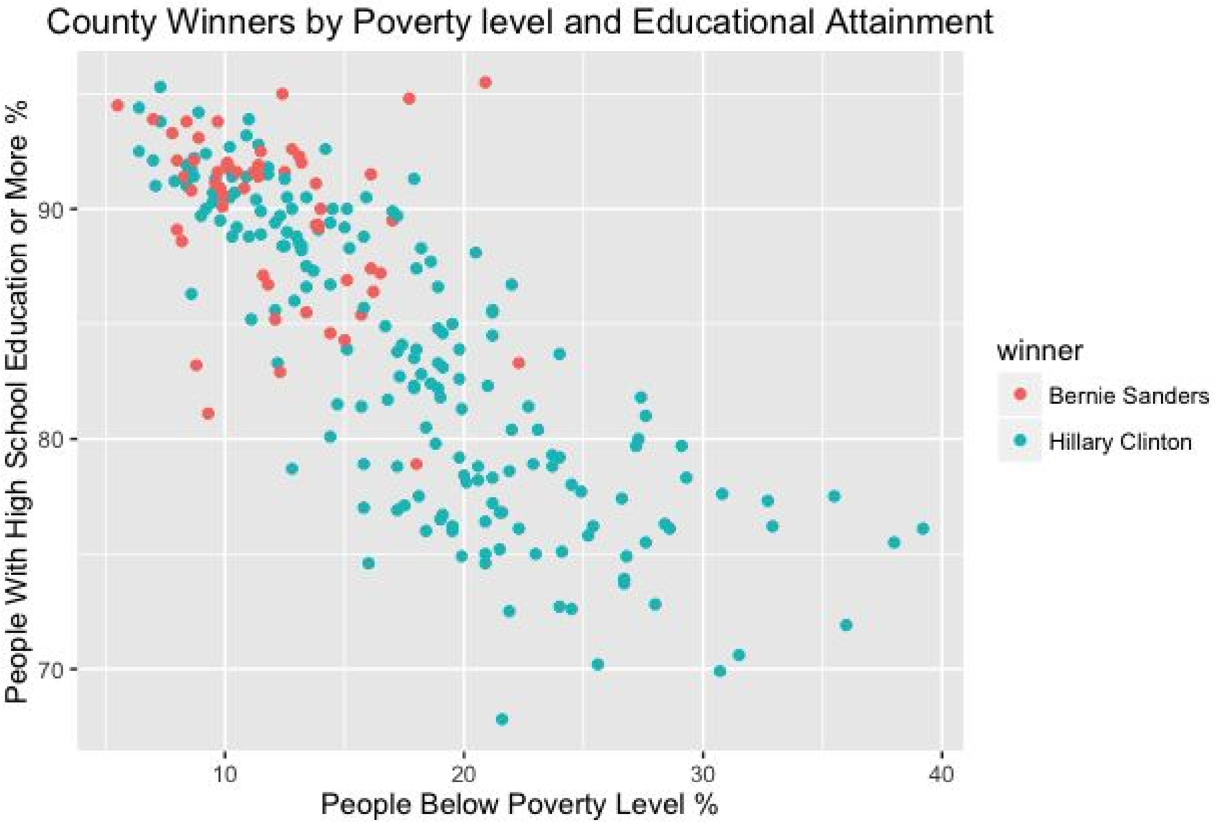


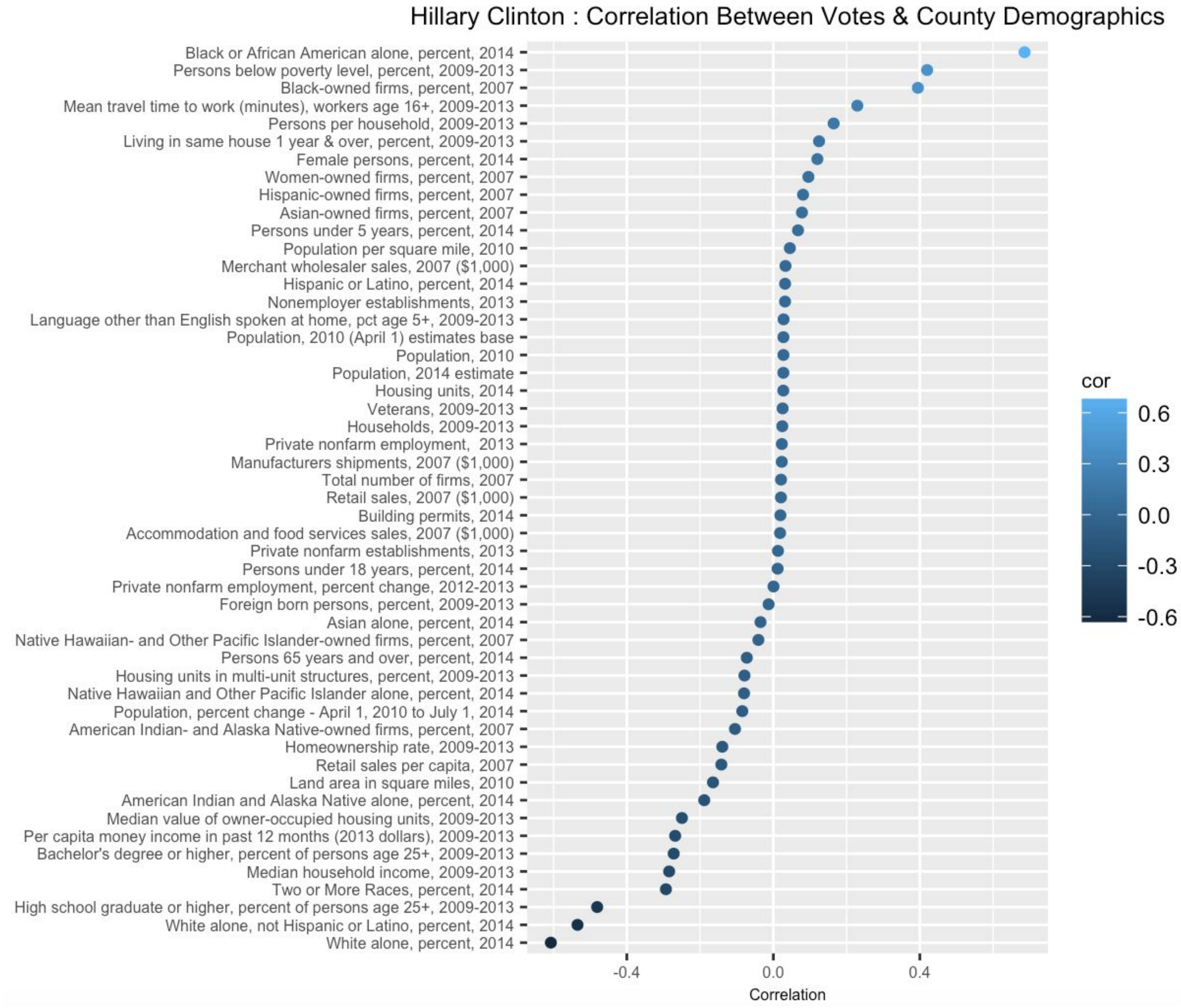
Fig: County-wise Democrat Candidate winners Poverty level and Education Attainment
The above figure reveals that Bernie Sander was only popular among the more educated

4.3 Correlations

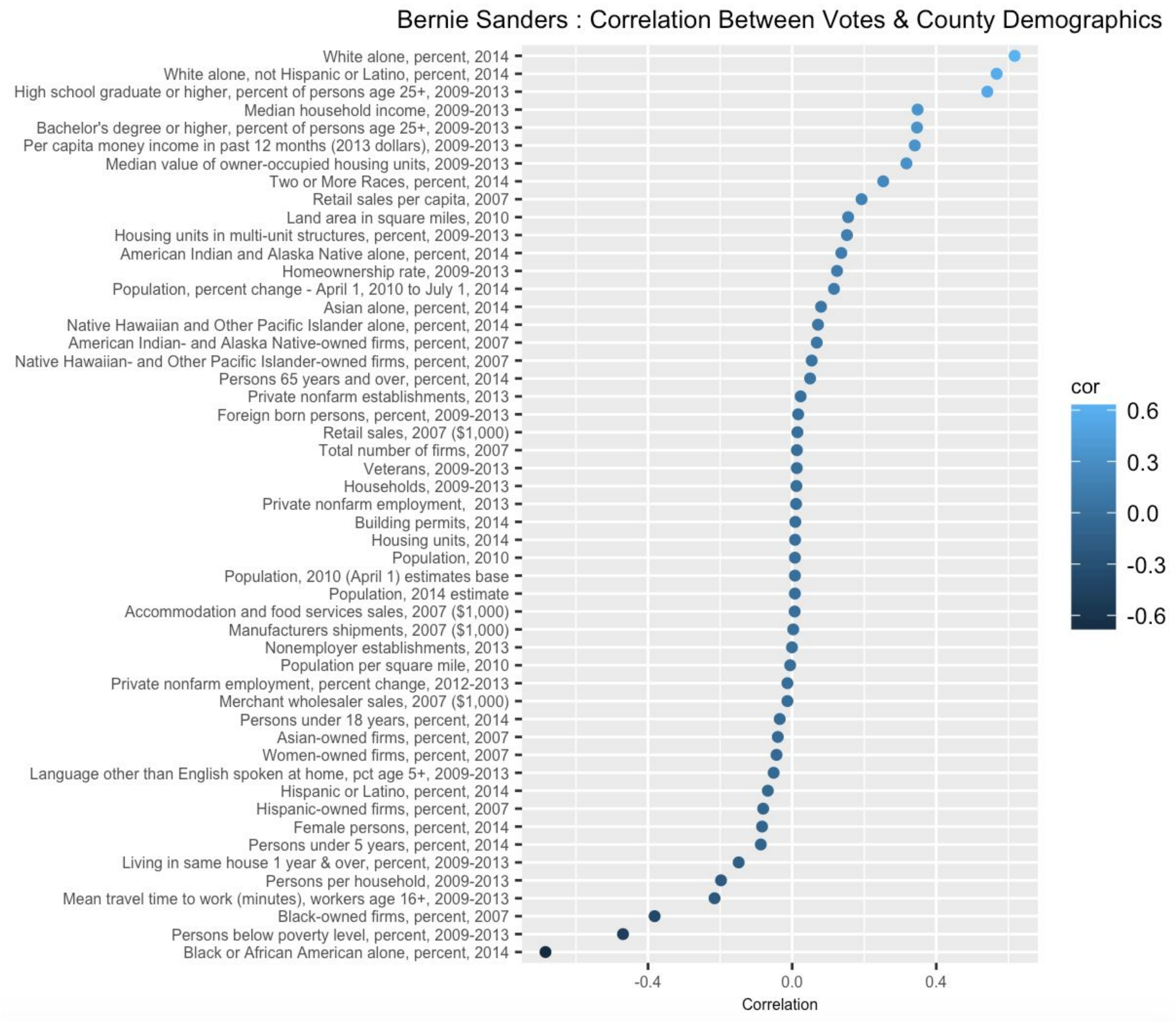
Correlation is a way to determine the relationship between two variables based on their correlation coefficients. By finding correlations between Independent Variables and the Target Variable, we can determine which variables to include in our Models so as to obtain a significant/high accuracy. We ran several correlations between combinations of Independent Variables so as to detect and resolve the problem of multicollinearity. Knowing that different demographics (Independent Variables) affect each candidate differently, we ran correlations between the number of votes won by each candidate and all our Independent Variables to determine which variables to include in our model.

4.3.1 Republicans

4.3.1.1 Hillary Clinton

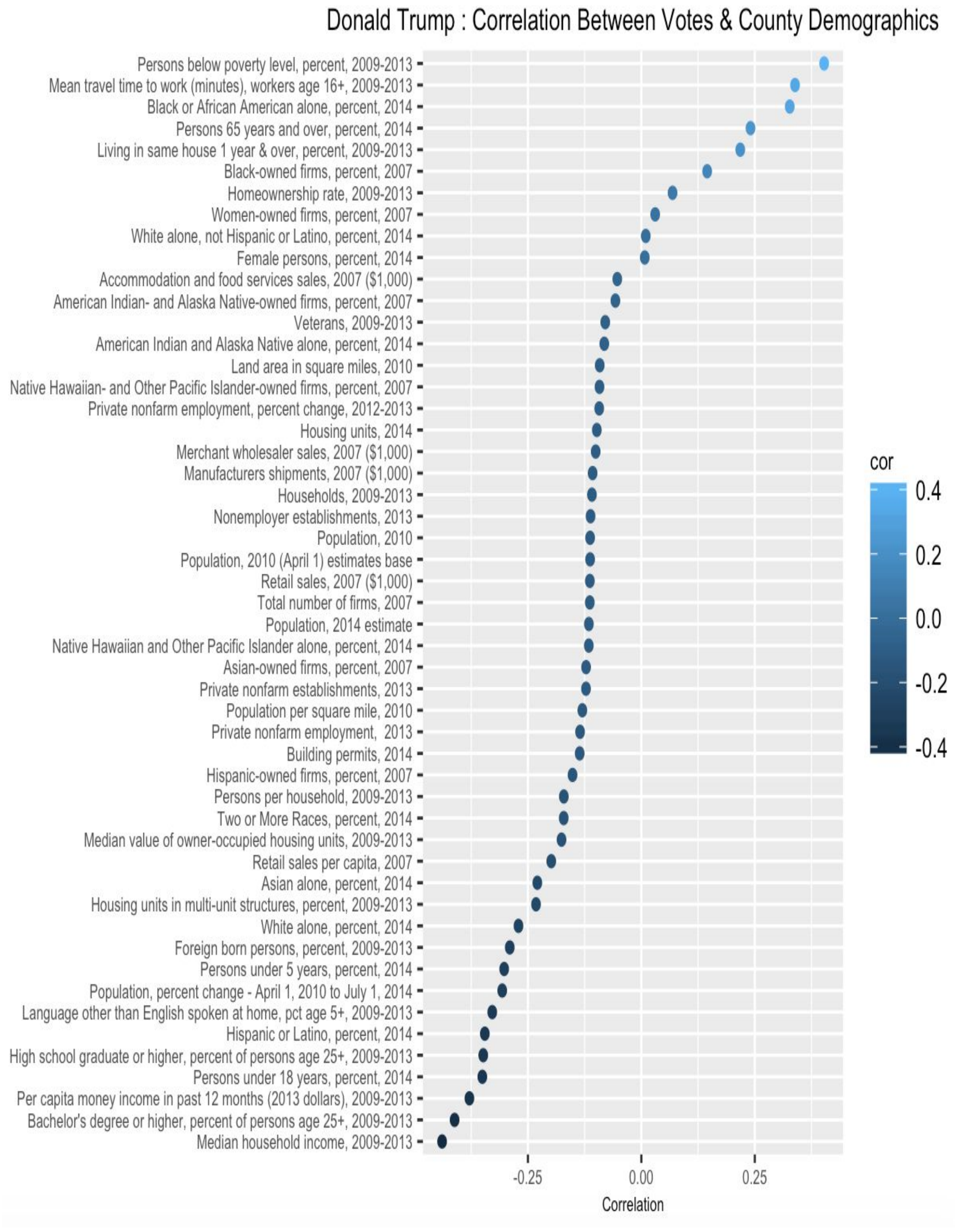


4.3.1.2 Bernie Sanders

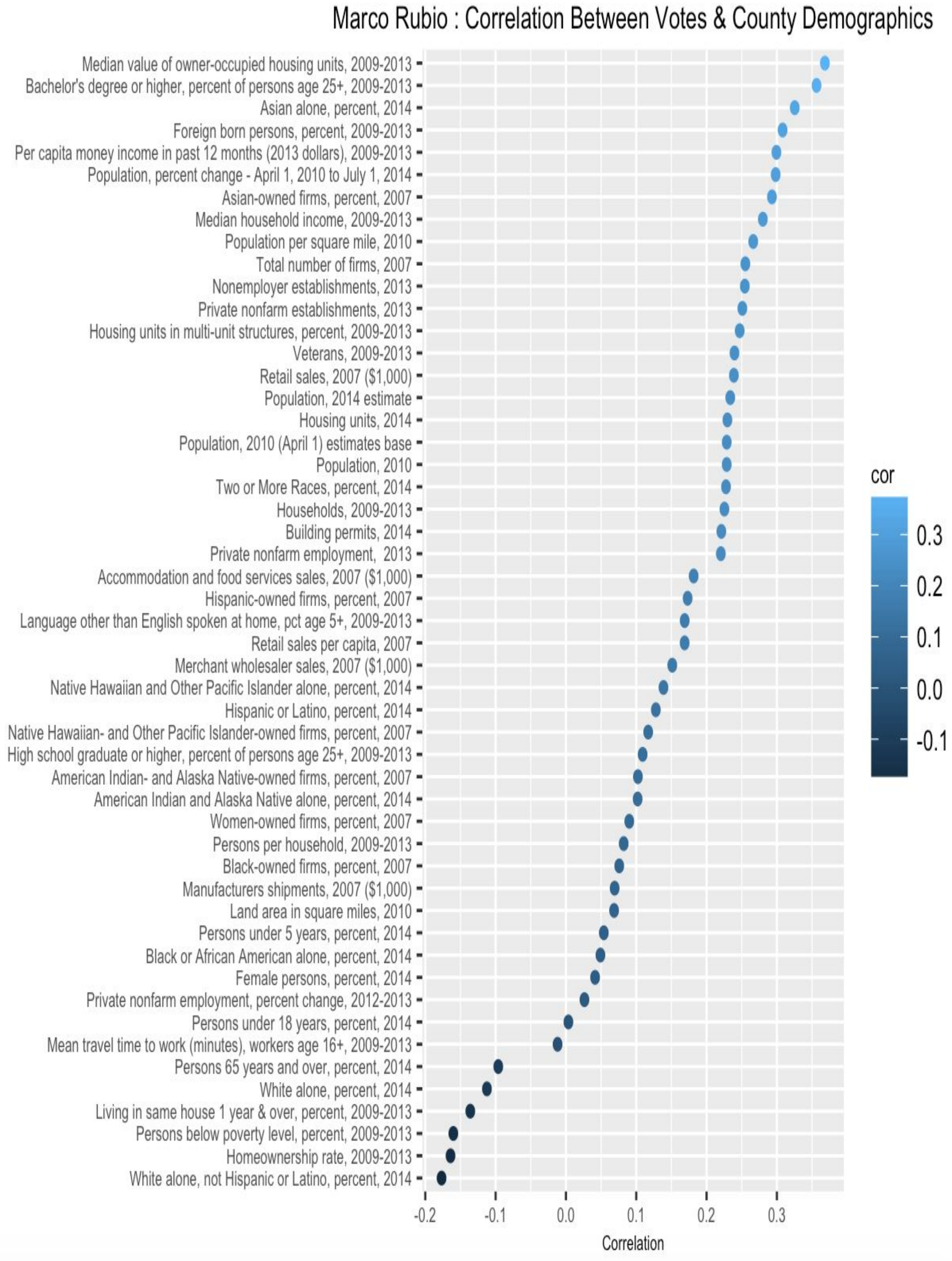


4.3.2 Democrats

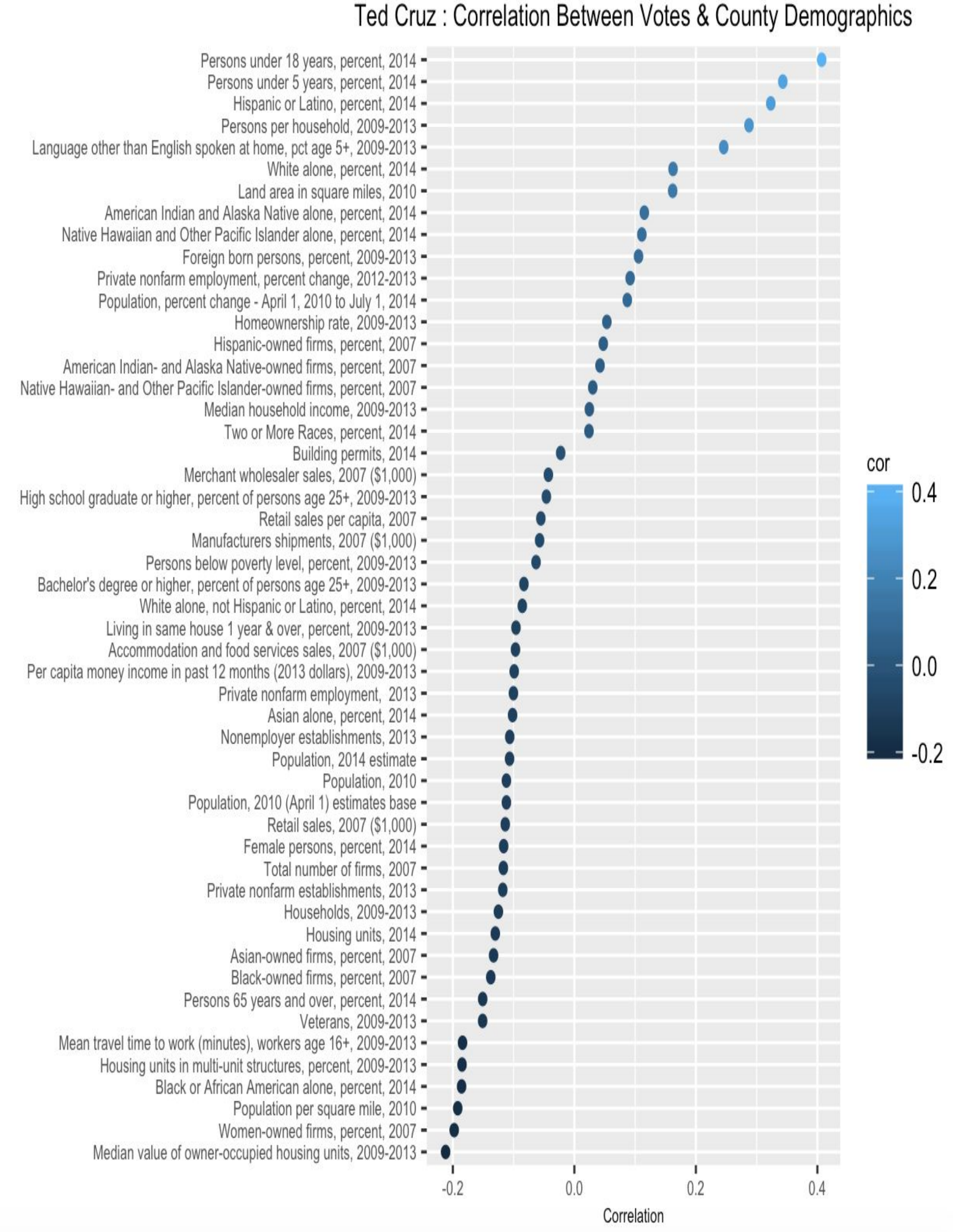
4.3.2.1 Donald Trump



4.3.2.2 Marco Rubio



4.3.2.3 Ted Cruz



5. Implementation

In the United States, the Presidential Election Primaries are conducted to decide the Presidential nominee from the Republican and Democrat parties. Hence, every county has two winners - one from each of the two main parties. Our dataset contained data about the primary results for both Democrats and Republicans in a single file. Since two independent groups of voters determine the winners of the 2 parties, we decided to build two different predictive models. Hence, the first step of the implementation involved dividing the primary results dataset into two data frames; one for Republicans and one for Democrats.

The next step involved the selection of standard methods and algorithms to build the predictive models. Our dataset contained the number of votes and the fraction votes every candidate won in different counties for the 28 states in which primaries had already been conducted. The number of people eligible to vote in the primaries change from state to state depending on the type of primaries that are conducted in that state, the population of the state and the number of registered Democratic and Republican voters. Hence, instead of predicting the number of votes each candidate would win in the primaries, we decided to change our research question into a classification problem and predict the winner of each of the counties based on the demographics. To do so, we first calculated the winners of each of the counties in our train set depending on the fraction votes and added a new attribute "winner". We then merged the county winner file with the demographics of that county. We thus got individual data frames which included the demographics of each county and the winner of that county for both the parties. The train and validation set was formed by splitting these data frames using the standard 80/20 split method. The test set included demographic information of all the counties in the remaining 22 states where Primaries had not been held.

Two algorithms were used for building the classification functions:

5.1 Multinomial Logistic Regression

Since the target variable - ' winner' was a nominal variable with more than 2 levels, multinomial logistic regression was used as one of the algorithms. This algorithm models the logs of the outcomes as a linear combination of the predictor variables. This algorithm was implemented in R using the 'multinom' function from the 'nnet' package. This inbuilt function fits multinomial log-linear models via neural networks and displays the the residual deviance, compared to the full saturated model & the AIC for this fit.[5]

5.2 Support Vector Machine

Support Vector Machines are supervised learning methods which can be used to build a model to classify data into categories when provided with a set of training examples. Hence, SVM can be used as a probabilistic binary linear classifier. The 'ksvm' function from the 'kernlab' package in

R was used to implement this algorithm. The kernlab package allows implementation of Support Vector Machines by applying the kernel trick while using the different kernel methods.
'ksvm' uses the `one-against-one'-approach for multiclass-classification with k classes such that k(k-1)/2 binary classifiers are trained and the appropriate class is found by a voting scheme. Hence, ksvm served as the most effective solution to our predictive problem. The SVM models were implemented using 4 different kernel methods : Radial , Linear, Laplacian & Bessel. [6]

6. Comparative Study

Bases on the results of the visual analysis and the inferences made from the correlations between predictors, 5 best combinations of input variables were selected for each of the 2 parties. Predictive models were built for each of the input combinations using the above mentioned Regression and SVM algorithms. Thus a total of 25 models were constructed for Republican predictions and 25 models for Democratic predictions. Validation sets were used to calculate the accuracy of each of the models and to determine the best predictive model by comparing these accuracies.

6.1 Republican Accuracy Matrix

Predictors	Logistic Regression	SVM (Radial)	SVM (Linear)	SVM (Laplace)	SVM (Bessel)
White alone + White alone(pop above 18yrs)+ African American + Persons below Poverty Level + High School Graduates, (25 yrs+) + Bachelor's degree, (25 yrs+) + Under 18yrs + Per capita money income in past 12 months + Median household income	79.52%	**82.18%**	80.85%	81.11%	80.05%
White alone +African American + Persons below Poverty Level + High School Graduates, (25 yrs+) + Bachelor's degree, (25 yrs+) + Under 18yrs + 65 yrs+ + Per capita money income in past 12 months+ Median household income +Population per Square mile	78.45%	81.64%	80.05%	81.08%	78.19%
White alone + African American+White alone(pop above 18yrs) + Persons below Poverty Level + High School Graduates, (25 yrs+) + Bachelor's degree, (25 yrs+) + 65 yrs+ + Per capita money income in past 12 months	78.19%	81.11%	78.72%	80.85%	80.58%
Hispanic or Latino + Persons below Poverty Level + High School Graduates, (25 yrs+) + Bachelor's degree, (25 yrs+) + Under 18 yrs+ Female persons	76.62%	75.26%	73.13%	75.53%	74.46%

White alone(pop above 18yrs) + African American + White alone + Persons below Poverty Level + Bachelor’s degree, (25 yrs+) + High School Graduates, (25 yrs+) + Per capita money income in past 12 months + Median household income	77.39%	81.38%	78.19%	80.58%	79.52%

6.2 Democratic Accuracy Matrix

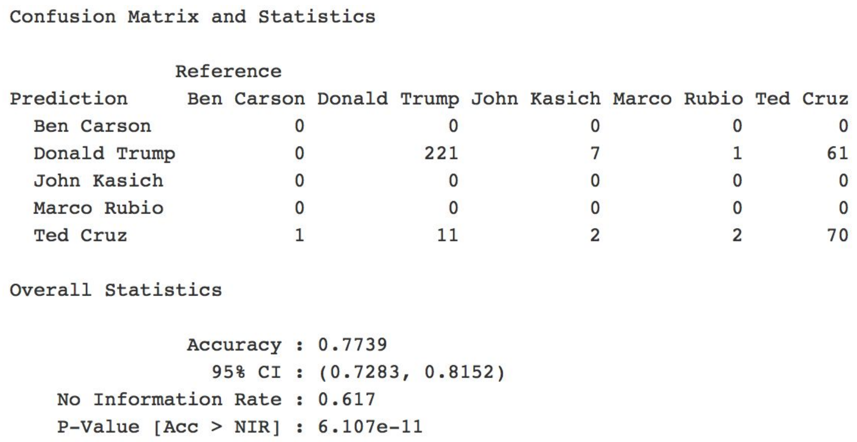
Predictors	Logistic Regression	SVM (Radial)	SVM (Linear)	SVM (Laplace)	SVM (Bessel)
White Alone + 65 yrs + Bachelor’s degree, (25 yrs+) + Median household income	77.92%	76.86%	72.60%	76.06%	75.79%
White alone + Bachelor’s degree, (25 yrs+) + White alone(Not Hispanic or Latino)	77.14%	73.13%	71.80%	72.87%	73.93%
White alone + Bachelor’s degree, (25 yrs+) + 65 yrs	76.62%	73.4%	73.93%	72.34%	74.2%
African American + White alone + 65 yrs + Bachelor’s degree, (25 yrs+) + Median household income	75.88%	80.85%	76.32%	80.05%	81.11%
White alone + White alone(pop above 18yrs)+ African American+High School Graduates, (25 yrs+) + Bachelor’s degree, (25 yrs+) + Under 18yrs	82.07%	82.85%	82.33%	81.29%	82.59%

From the comparative study, the models with the highest accuracy were selected as our final predictor models.

6.3 Final Republican Model

The final republican model was built using the 10 predictor variables mentioned below

1. White alone, percent (2014)
2. African American alone, percent (2014)
3. White alone, not Hispanic or Latino, percent, (2014)
4. Persons below poverty level, percent, (2009-2013)
5. High school graduate or higher, percent of persons age 25+, (2009-2013)
6. Bachelor's degree or higher, percent of persons age 25+, (2009-2013)
7. Per capita money income in past 12 months (2013 dollars), 2009-201
8. Median household income, 2009-2013
9. Female persons, percent, 2013
10. Persons under 18 years, percent, 2013



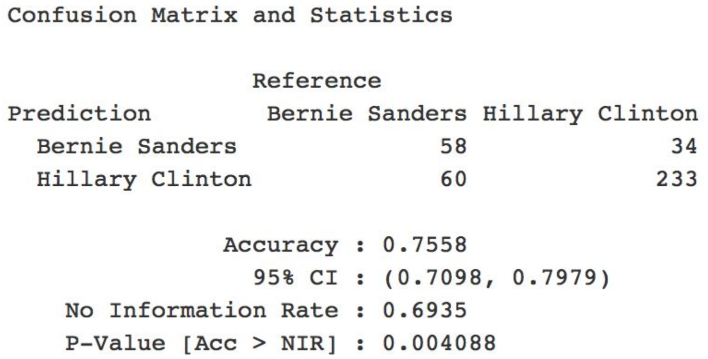
```
Confusion Matrix and Statistics

              Reference
Prediction     Ben Carson Donald Trump John Kasich Marco Rubio Ted Cruz
  Ben Carson            0            0           0           0        0
  Donald Trump          0          221           7           1       61
  John Kasich           0            0           0           0        0
  Marco Rubio           0            0           0           0        0
  Ted Cruz              1           11           2           2       70

Overall Statistics

               Accuracy : 0.7739
                 95% CI : (0.7283, 0.8152)
    No Information Rate : 0.617
    P-Value [Acc > NIR] : 6.107e-11
```

From the confusion matrix, it can be seen that we were able to achieve an accuracy of 77.93%. The confidence interval values suggested that the model accuracy will lie within the range of 0.7183 to 0.8152, 95% of the time. The p value being almost 0 suggests that the model is highly significant.

6.4 Final Democrat Model

The final democratic model was built based on following predictors:

1. White alone, percent (2014)
2. African American alone, percent (2014)
3. White alone, not Hispanic or Latino, percent, (2014)
4. Persons below poverty level, percent, (2009-2013)
5. High school graduate or higher, percent of persons age 25+, (2009-2013)
6. Bachelor's degree or higher, percent of persons age 25+, (2009-2013)
7. Persons under 18 years, percent, 2013



```
Confusion Matrix and Statistics

                 Reference
Prediction        Bernie Sanders Hillary Clinton
  Bernie Sanders              58              34
  Hillary Clinton             60             233

               Accuracy : 0.7558
                 95% CI : (0.7098, 0.7979)
    No Information Rate : 0.6935
    P-Value [Acc > NIR] : 0.004088
```

From the confusion matrix, it can be seen that we were able to achieve an accuracy of 75.93%. The confidence interval values suggested that the model accuracy will lie within the range of 0.7083 to 0.7979, 95% of the time. The p value is less that 0.5, which suggests that the model is significant.

7. Conclusion

Our analysis showed that Race and Education variables were strong predictors of the target variables for both the parties. The best model we obtained was SVM Radial which returned an accuracy of 82.18% for the Republican party and an accuracy of 82.85% for the Democratic party. We went ahead to test our model to predict the party wise winners for the counties of Maryland State, as we did not have its results of the primaries in our data set. On comparing our predicted results with that of the actual primary election results for Maryland, we found that our model predicted the winners of 22 out of 24 counties correctly for Republicans [Accuracy: 92%] and 23 out of 24 counties correctly for Democrats[Accuracy: 96%]. The model was limited to predicting only the winner for the primary elections due to the complex nature of U.S elections which involves delegates to determine the final winner.

8. References

1. S. (n.d.). Machine Learning: What it is and why it matters. Retrieved May 10, 2016, from http://www.sas.com/en_id/insights/analytics/machine-learning.html
2. S. (n.d.). Predictive Analytics: What it is and why it matters. Retrieved May 12, 2016, from http://www.sas.com/en_us/insights/analytics/predictive-analytics.html
3. Shen, G. (2013, January/February). Big data, analytics and elections. Retrieved May 12, 2016, from http://www.analytics-magazine.org/january-february-2013/731-big-data-analytics-and-elections
4. United States presidential primary. (n.d.). Retrieved May 12, 2016, from https://en.wikipedia.org/wiki/United_States_presidential_primary#Types_of_primaries_and_caucuses
5. (2016, February 2). Retrieved May 12, 2016, from https://cran.r-project.org/web/packages/nnet/nnet.pdf
6. _A. (n.d.). Ksvm {kernlab}. Retrieved May 12, 2016, from http://www.inside-r.org/node/63499