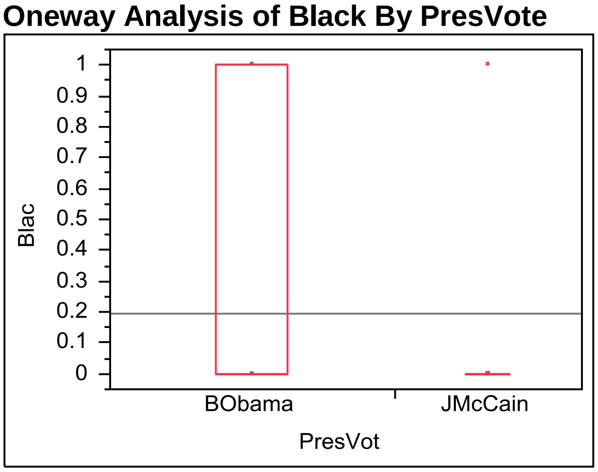
This work was done as a lab assignment for a statistics class. Since every student analyzed the same data, the source and variables are not explained as thoroughly as they would be in more formal report. The class also focused on using graphics to support analysis, rather than on creating very well-designed visualizations.

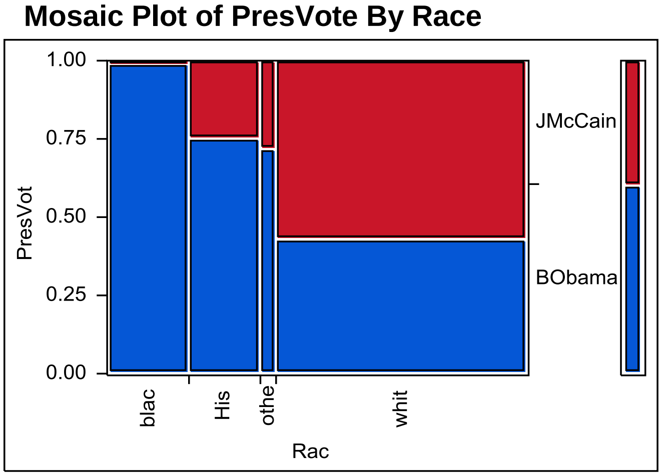
In the 2008 US presidential election, the American National Election Survey conducted pre- and post-election surveys on which candidate a voter selected, party affiliation, and demographic information such as race, gender, income, age, and education. This information can be analyzed to determine what factors are best at forecasting whether a person will vote for a Democratic or Republican candidate.

First, it is worthwhile to graphically explore which factors have visibly wide distinctions in who voted for Barack Obama and who voted for John McCain. One way of beginning this is to look at boxplots of the distribution of a predictor for Obama vs. McCain. Figure 1 shows such a box plot for the indicator variable Black, which has a value of 1 if the respondent identified as black and a value of 0 if she did not. While the median for both candidates is 0, the 75th percentile for Obama is 1, while for McCain it is 0. This indicates that of voters who chose Obama, a greater percentage were Black than that of voters who chose McCain. This observation is confirmed in Figure 2, a mosaic plot of the fraction of voters in four racial categories that chose each candidate. Of Black voters, 99% voted for Obama, while only 43% of white voters did.



Oneway Analysis of Black By PresVote

Figure 1: Of voters for each candidate, the median value of Black is 0, but the 75th percentile of Black is 1 for Obama but 0 for McCain.



Mosaic Plot of PresVote By Race

Figure 2: Of Black voters, 99% voted for Obama, compared to 75% of Hispanic voters, 72% of voters of other races, and 43% of white voters.

A similar trend can be seen in the indicator variable Never Married, which has a value of 1 if a voter has never been married and a value of 0 otherwise. The boxplot looks identical to that for Black and again the median for both candidates is 0 but the 75th percentile for Obama is 1. Figure 3 shows the mosaic plot of the fraction of voters in four categories of marital status that voted for each candidate.

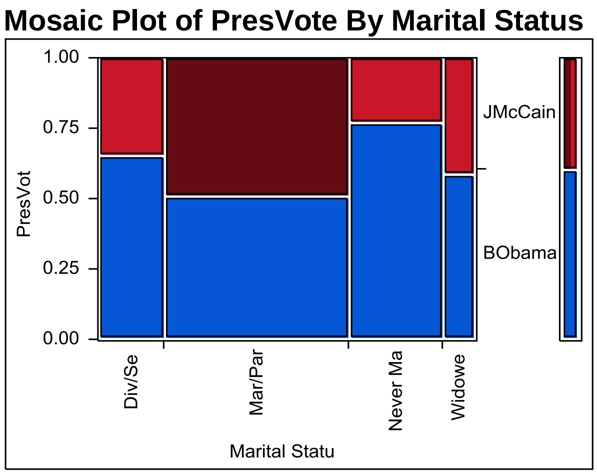


Figure 3: Of never married voters, 77% voted for Obama, compared to 65% of divorced or separated voters, 51% of married or partnered voters, and 58% of widowed voters.

One of the most clear distinctions between voters who chose Obama and voters who chose McCain is in Party ID. Party ID is measured on a scale from -3, strong Democrat, to 3, strong Republican. A value of 0 indicates an independent. Figure 4 shows the boxplot for Party ID, indicating that the median and 75th percentile for voters who chose Obama is 0, while the median for voters who chose McCain is 1. The middle 50 percentiles for the two candidates do not overlap at all.

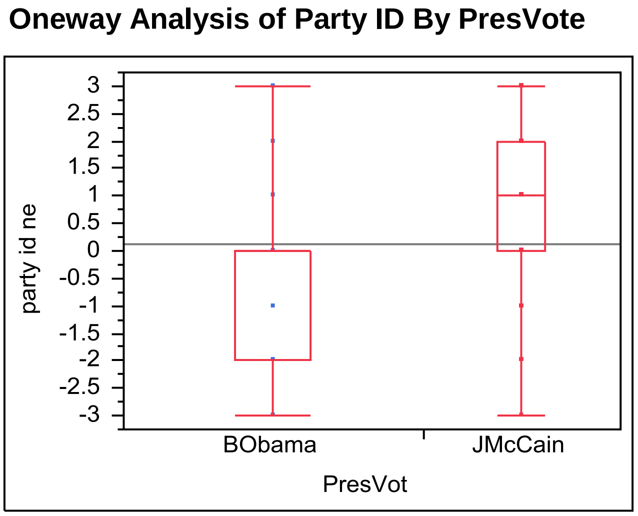


Figure 4: The 25th, 50th, and 75th percentiles respectively are 2, 4, and 4 for Obama and 4, 5, and 6 for McCain

After this exploratory analysis, it is useful to begin to construct models that predict whether a voter chose Obama or McCain based on various demographic predictors. A logistic model is appropriate for this analysis because candidate chosen is a binary categorical variable. Starting with Party ID, the logistic fit (Figure 5) shows how well Party ID alone predicts whether a person chose Obama or McCain. The equation for the natural log of the odds of Obama winning is $\ln(\omega) = 0.756(0.082) - 0.955(0.062)\text{Party ID}$, where the standard error of each regression coefficient is noted in parentheses. Note that all regressions performed in this report are summarized in Table 1.

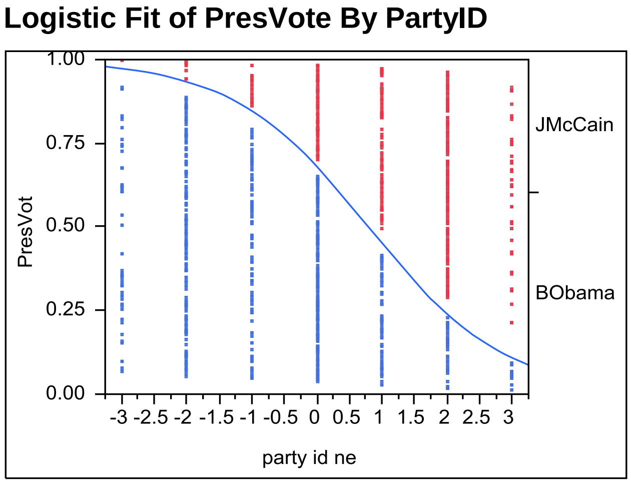


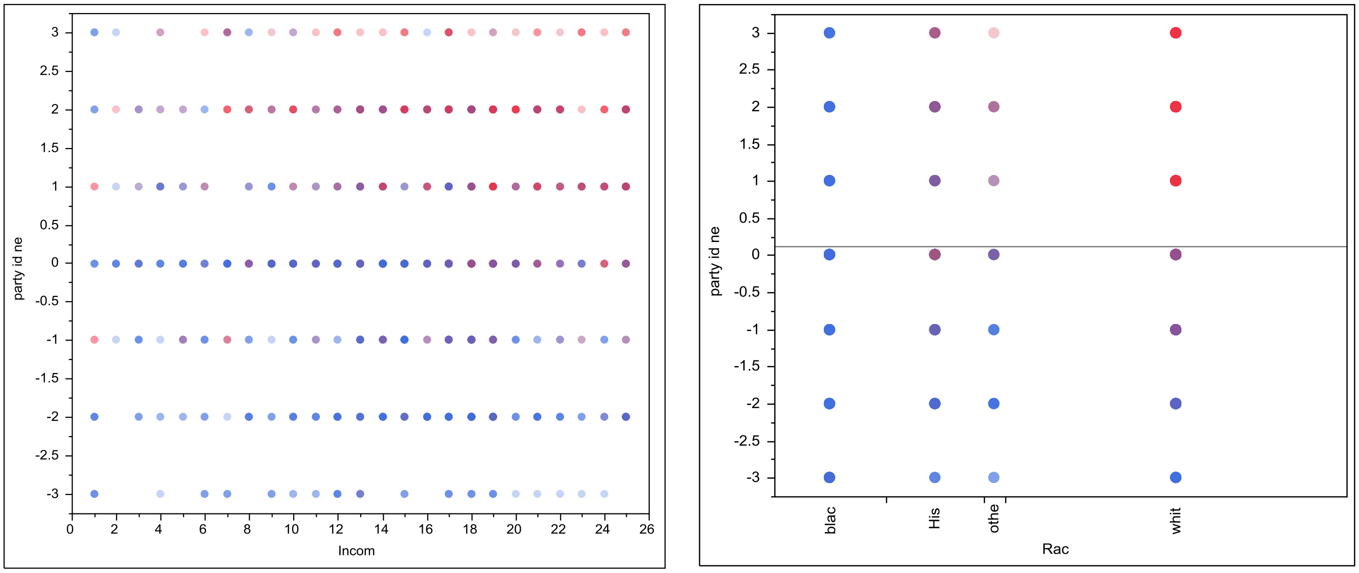
Figure 5: The fraction of voters of each Party ID that voted for Obama, and the logistic fit curve

The error rate for this model, or the number of voters for whom it incorrectly predicted a vote, is 245 out of 1068, or 22.9%. The error rate of the null model – a model that assumes everyone voted for the winner of the election – is the percentage of voters who chose McCain, 39.6%. The addition of the single predictor Party ID improves the model significantly. The further addition of various demographic factors can make it even better. When a logistic regression is run with all the predictors for which data has been provided, the predictors that are statistically significant are Black, Hispanic, Income, and Party ID. The equation for a model with just these predictors is available in Table 1. This model incorrectly predicted the votes of 171 people, so its error rate is 16%.

While this model is good at predicting which candidate a voter will select, it is not always desirable to include Party ID as a factor in the model. A political scientist may instead want to be able to predict outcomes of elections based only on demographic factors regardless of party ID. A few graphics are helpful in illustrating how other predictors are related to Party ID and to which candidate a person voted for. In both Figure 6 and Figure 7, a red dot indicates that a person voted for McCain and a blue dot indicates a vote for Obama. The dots are translucent, so the color gradation indicates the fraction of Obama versus McCain voters at each point. Figure 6 shows that most McCain voters are concentrated on the upper right side of a diagonal blue to red gradation, or that they tend to have both higher incomes and Party IDs. While income is not a perfect substitute for Party ID – high income democrats voted for Obama – it does seem that as income increases the fraction of voters choosing McCain increases, just as it does as Party ID increases. Figure 7 shows that regardless of Party ID, Black voters overwhelmingly voted for Obama. White voters followed a much more strict party line, while Hispanic voters and voters of other races with high Party ID values were more mixed. This indicates race is also a somewhat useful stand-in for Party ID, as White voters are more likely to vote for McCain than voters of other races, particularly Black voters.

Bivariate Fit of Party ID By Income

Oneway Analysis of Party ID By Race



Figures 6 (left) and 7 (right): McCain voters are represented by red dots and Obama voters by blue dots. The dots are translucent to show the relative number of voters for each candidate in a given spot.

Removing Party ID from the predictor space and running a logistic regression with all the remaining variables indicates that Black, Hispanic, Other, Never Married, Age, and Income are all statistically significant predictors. Experimenting with these predictors in various combinations yields the results recorded in Table 1. Error denotes the number of voters the model incorrectly predicted, out of 1068. The values in the first line are the coefficient estimates in the log odds equations. The second line (in italics) denotes the standard error of each coefficient.

Error	Intercept	Party ID	Black	Hisp	Other	Income	Age	Never Married
171	0.988	-1.08	5.23	1.49		-0.059		
245	0.756 *0.082*	-0.955 *0.062*						
305	0.973 *0.407*		4.681 *0.717*	1.297 *0.195*	0.988 *0.374*	-0.049 *0.014*	-0.011 *0.005*	0.441 *0.212*
307	1.458 *0.335*		4.691 *0.716*	1.246 *0.193*	1.025 *0.372*	0.058 *0.0133*	-0.016 *0.004*	
317	0.428 *0.229*		4.890 *0.715*	1.346 *0.715*	1.126 *0.367*		-0.0146 *0.004*	
328	0.588 *0.227*		4.742 *0.716*	1.332 *0.191*	1.146 *0.370*	0.054 *0.013*		
331	-0.290 *0.08*		4.925 *0.715*	1.411 *0.189*	1.225 *0.365*			
423	0.422							

Table 1: Coefficients in the log odds regression equations for various combinations of predictors

In order to best generate predictions about which candidate a person will vote for, a larger model is better because it provides more information. The best model is one that includes all of the statistically significant predictors, which gives an error rate of 305 out of 1068, or 28.6%. Yet even a model with only the three indicator variables for race only has an error rate of 31.0%, indicating that race is highly significant in indicating which candidate a voter choses. The model I suggest selecting for prediction is the fourth in the table, with the predictors Black, Hispanic, Other, Income, and Age. The addition of Never Married to the model does not improve the error rate very much, but the subtraction of any of the predictors increases the error by at least 10 voters.

For this model, switching from a white voter to a Black voter that is otherwise identical multiplies the odds of voting for Obama by 108. Decreasing a voter's age by one year multiplies their odds of voting for Obama by 1.02. The model underestimates the number of Obama voters, as it predicts that 552 people will vote for Obama when in actuality, 645 people did. It is better at predicting Obama voters than McCain voters, however, since of the voters who actually voted for Obama, the model correctly predicted 81%, while of the voters who actually voted for McCain it only correctly predicted 61%.

One main concern with using this study to make predictions about the 2016 election is that it does not take into account the race of the candidates. This analysis reveals that the race of voters is very significant in determining which candidate they voted for, especially because nearly all Black voters selected Obama. However, it does not separate whether voters were inclined to vote for Obama because he was the Democratic candidate or because of other personal or political features, including his race. In the 2008 election, many Black voters may have voted for Obama because he had the potential to become the first Black president, when they might in other circumstances have voted for a Republican candidate. It would have been useful to have asked survey participants what their main reason for selecting the candidate they did was, or about whether they were more likely to vote for candidates of specific races.