Group 4 Final Project

Drew Lewis
Shane Ulrich
Todd Livergood

Intro/Business Problem

The Democratic National Committee (DNC) has hired our team to diagnose the 2016 presidential election results to **better understand voting patterns** throughout the country. Their objective is to i**dentify any key demographic, educational, and economic attributes** that played a significant role in determining for the likelihood a county votes for one party vs. another. Of **particular concern are those voters in swing counties** that voted for Democrats in 2012 and for a different party in 2016.

Data Preparation

Dataset:

- County-Level 2016, 2012, and 2008 election results
- Census demographic, education and economic data

Data preparation included:

- Cleaning data to ensure consistent data types
- Trimming of white space
- Replacing null values
- Consolidating co-linear variables to avoid inherent correlation between predictors
- Creating dummy values for binary categorical variables

Models

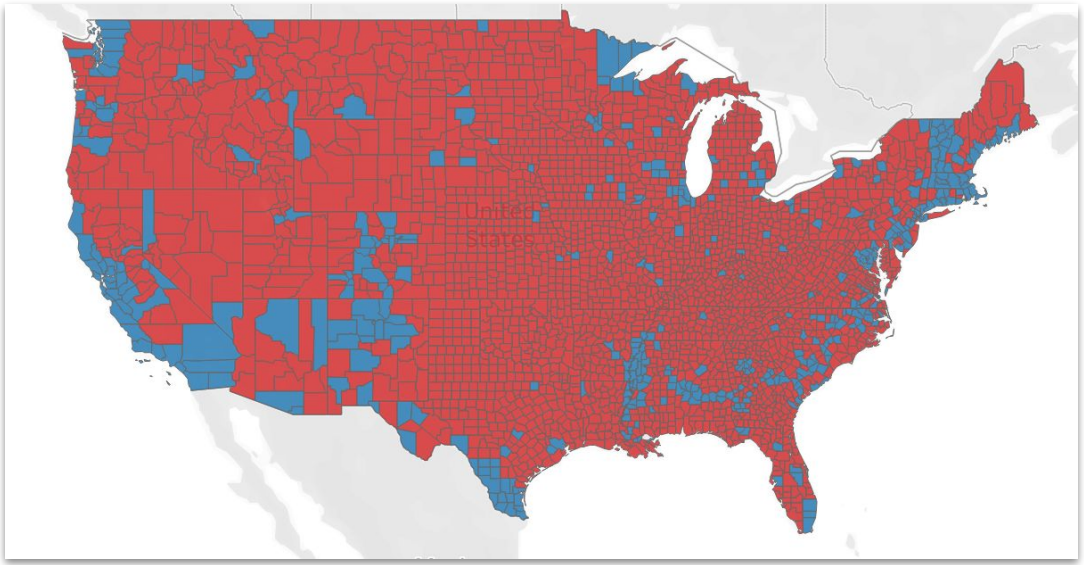
Statistical Models Used:

- Linear regression for determining correlation between predictors and a target variable
- Logistic Regression for determining whether a particular political party wins or loses a county
- Classification Decision Trees with and without Bootstrap sampling
- Random Forest
- Extra Tree

- We'll evaluate model accuracy using the appropriate diagnostic output variables for the respective models (i.e. r-squared, accuracy score, AUC value, etc.)
- We'll also make a determination if these output variables align with expected outcomes and voter behavior to ensure that our model produces results that make intuitive sense.

Exploratory Visualizations

Democrats won 383 fewer counties than they did in the 2008 presidential election and 212 fewer than in 2012.



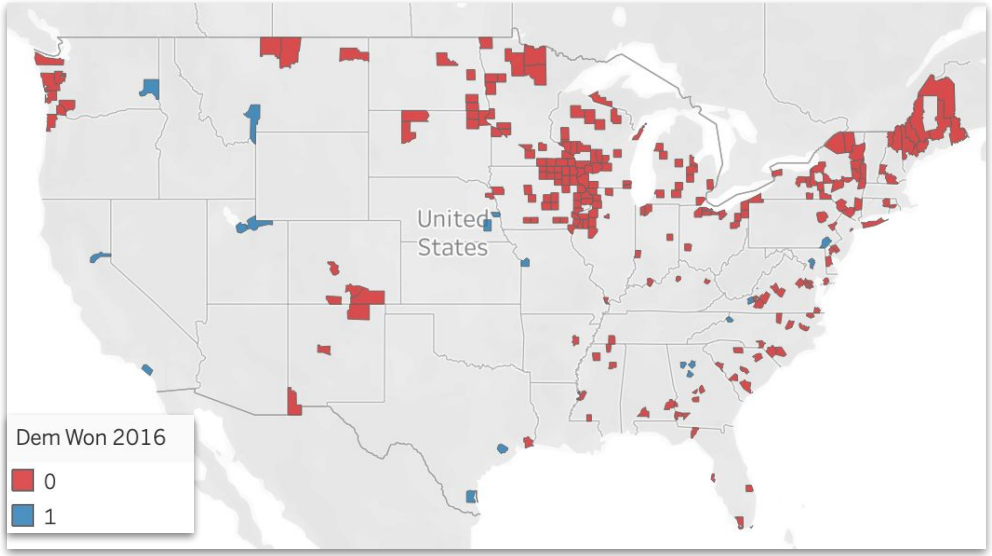
County Totals			
Party	**2008**	**2012**	**2016**
Democrats	870	699	487
Republicans	2,243	2,414	2,626

	Dem Won 2016	
State	0	1
IA	32	6
WI	23	12
NY	23	16
MN	19	9
MI	12	8
IL	12	11
VA	11	39
OH	10	7

62% of the counties that democrats won in 2012 AND lost in 2016 are in 8 states. 6 of the 8 are in the Midwest.

248 counties voted for a different political party in 2016 than in 2012, enough to affect the outcome of the election. Democrats only won 18 of these swing counties; they lost all of the swing counties in the decisive Midwestern states.

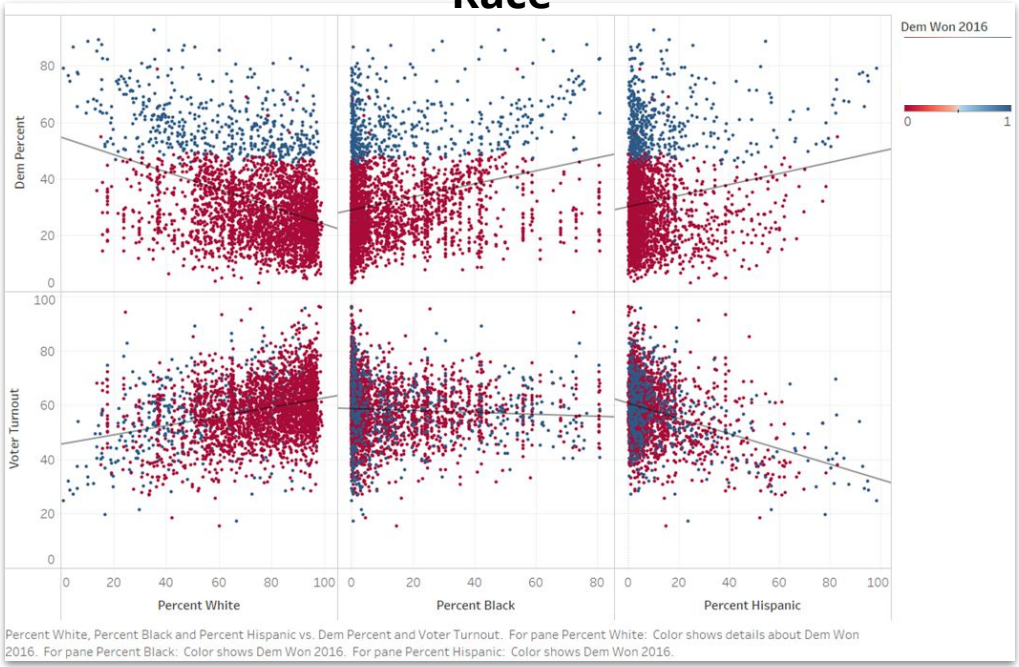
Map of Swing Counties in 2016 Election*



Swing County	Dem Won 2016 0	Dem Won 2016 1
0	2,396	469
1	230	18

*Swing county defined as a county whose 2016 winner was from a different party than the party of the 2012 winner.

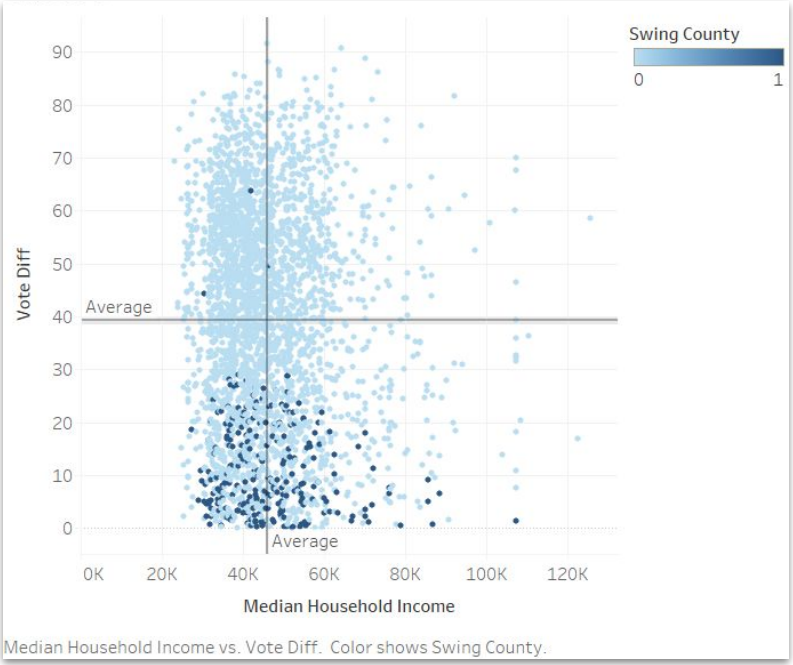
Correlation Matrices: Democratic Percent and Turnout by Race



Percent White, Percent Black and Percent Hispanic vs. Dem Percent and Voter Turnout. For pane Percent White: Color shows details about Dem Won 2016. For pane Percent Black: Color shows Dem Won 2016. For pane Percent Hispanic: Color shows Dem Won 2016.

The scatter plots below show the correlation for all counties between voter turnout or percentage Democratic vote and different races. We observe a **negative correlation between percent white population** and the Democratic percentage of the vote, and a **positive correlation for non-whites**. However, we also observe that voter turnout declines as the non-white percentage of the population increases. This suggests that the **DNC should focus on voter turnout in counties with larger non-white populations** to improve election odds.

Distribution of County Results by Household Income

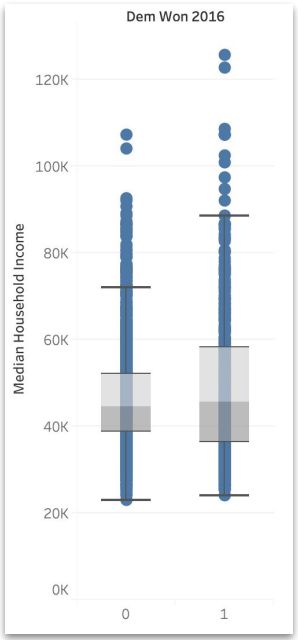


Median Household Income vs. Vote Diff. Color shows Swing County.

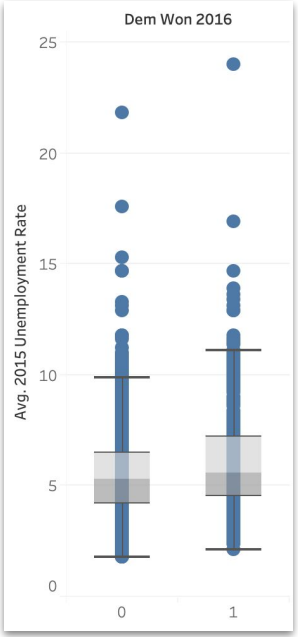
This visualization shows the vote differential between the Democratic and Republican candidates (percentage) vs. median household income for all counties. Surprisingly, the **distribution of median household income is similar for both swing and non-swing counties** despite the narrative that swing counties flipped because of economic distress. Unsurprisingly, the vote differential is generally smaller for swing counties.

County Results vs. Predictors

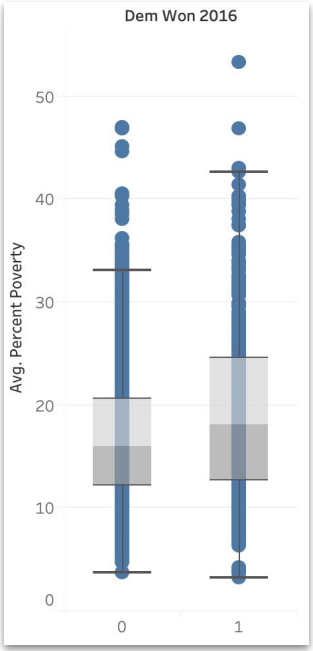
Income



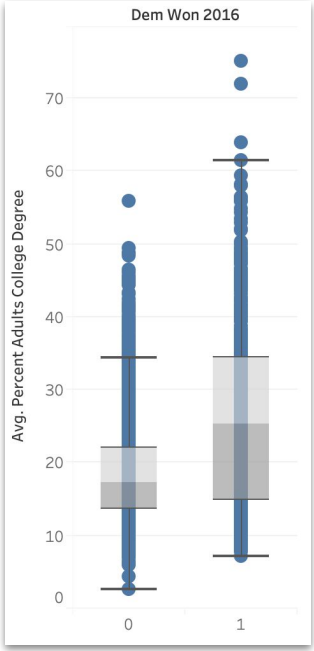
Unemployment



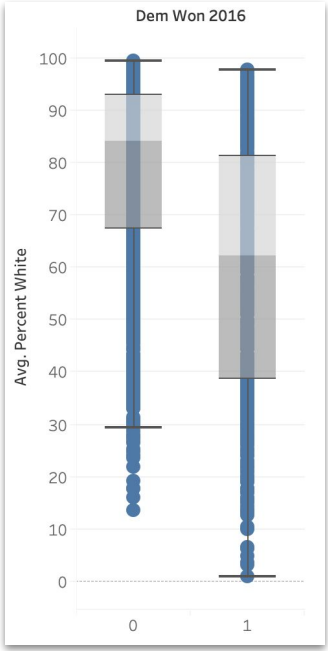
Poverty



College Degrees



White population



Counties that voted for the Democratic candidate in 2016 on average have **higher rates of poverty**, **unemployment and adults with college degrees**; counties that *didn't* vote for the Democratic candidate have a **significantly higher percent of whites**. Median household incomes were similar regardless of election result, though Democratic counties have a broader distribution of incomes.

Predictive Modeling in Python

Leveraging Python, we explored the statistical significance of these variables and the outcome of the 2016 election in an effort to produce the most accurate predictive model. In the process we identified some of the important factors that impact the way a county will vote in a given presidential election.

- Linear regression
- Logistic Regression
- Classification Decision Trees
- Random Forest
- Extra Tree

Predictive Modeling in Python

Response Variable used

-**For classification models**: Yes or No -

Majority of county voted for the Democratic candidate in 2016

-**For regression models:**

% of voters in a county that voted for the Democratic candidate

Predictors Evaluated

-Yes or No: Did majority of county vote for a Democrat in 2012 election

-Yes or No: Did majority of county vote for a Democrat in 2008 election

-Voter Turnout %

-2015 Population

-Population change between 2010 and 2015

-% of county in poverty

-% of adults with college degree

-% white

-County Unemployment Rate

-Median household income

-% of jobs that are white collar

Model 1: Linear Regression

Methodology:

- Split into test/train set using percent of democratic votes as the target variable
- Train the model on the training set and evaluate model performance on the test dataset

Results:

R-squared = 0.65

- Moderately strong r-squared value suggests our model is effective at predicting the variance in the percent of votes a Democratic candidate will receive in a given county. However, this result isn't as strong as subsequent models.

Model 2: Logistic Regression

Methodology: Split into test/train set, run on the train, evaluate model on the test dataset. Use Backwards Elimination method to remove predictors that weren't significant at the 95% confidence level. Then calculated optimal threshold probability in order to classify predictions into yes/no county voted democrat. Evaluate key diagnostics to see how the model is performing.

Predictors that were found to be significant: Democrat won 2012, 2015 population, % poverty, % adults with college degrees, % white, % with white collar jobs.

Model 2: Logistic Regression

Key Diagnostics:

-AUC value was 0.98. Since this was high, that means the model is doing a good job of classifying the dependent variables.

-e^ of coefficients:

const	0.027116
dem_won_2012	244.338368
pop_estimate_2015	1.000002
percent_poverty	1.060157
percent_adults_college_degree	1.147811
percent_white	0.973608
percent_white_collar	0.930944

This shows how a unit change in these variables increases/decreases the odds of a county voting Democrat in 2016. Directionally, these make intuitive sense. For example, the e^ coefficient for percent of adults with college degree is 1.147811. This means that for each unit increase in % of this variable, the odds of a county voting majority Democrat go up by about 15% which makes logical sense and is in line with prior diagnostics.

Model 2: Logistic Regression

Key Diagnostics (cont.):

-Confusion Matrix:

	Predicted No	Predicted Yes
Actual No	460	56
Actual Yes	2	105

-Calculated Accuracy score: 0.907. This means that the model properly classified the test dataset over 90% of the time.

-Final AIC value (669) came down by 6 pts after removing insignificant predictors. This provides further support that removing these predictors was the right decision and improved the model.

-Pseudo R2: 0.69. This is a high value which indicates that a lot of the variance in the outcome is explained by the model.

Conclusion

-The favorable diagnostics indicate that this is a good model for predicting whether the Democratic candidate will win a given county.

-The predictors also make intuitive sense as far as how they impact the response variable. This would be a simple model to explain to others.

Model 3: Decision Tree

Methodology: Split into test/train set, run on the train, evaluate model on the test dataset. Create an initial “full-tree” model to get a baseline. Then create a “pruned-tree” based on evaluation of best hyperparameters to use. Finally, evaluate some diagnostics to see how tree model performs.

Key Diagnostics:
-Best hyperparameters to use for pruned tree: Max Depth = 7, Max Features = 7, Minimum samples per split = 100. (See appendix for visual of tree.)
-Accuracy score of pruned tree after testing on test dataset: 0.92

Conclusion:
-The accuracy score is very good, and even higher than the logistic regression model.
-However the size of the tree even after pruning is quite large; Difficult to interpret the model.

Model 4: Decision Tree with Bootstrap sampling

Methodology: Same methodology as with the original decision tree. However this time we used Bootstrap sampling to see if this would improve the model (with n_samples = 70% of the rows.)

Key Diagnostics:
-Best hyperparameters to use for pruned tree: Max Depth = 6, Max Features = 7, Minimum samples per split = 100. (See appendix for visual of tree.)
-Accuracy score of pruned tree after testing on test dataset: 0.926

Conclusion:
-The accuracy score was slightly higher than the original decision tree model.
-However this didn't seem to improve the model in a significant way; The final pruned tree is still rather large.

Model 5 and 6: Random Forest and Extra Tree Model

Methodology: Split into test/train set, run on the train using number of estimators=500, evaluate model on the test dataset. Finally, evaluate some diagnostics to see how tree model performs including summary of feature importance.

Model 5: Random Forest Model

Key Diagnostics:

-Calculated accuracy score after testing on test dataset: 0.926

-Feature Importance calculations (right):

All Counties

	importance
dem_won_2012	0.387808
pop_estimate_2015	0.107476
percent_white	0.086275
net_population_change_2010-2015	0.084748
dem_won_2008	0.072906
percent_adults_college_degree	0.063127
voter_turnout	0.052921
median_household_income	0.039805
percent_white_collar	0.039632
percent_poverty	0.036874
2015_unemployment_rate	0.028428

Swing Counties

	importance
dem_won_2012	0.229085
net_population_change_2010-2015	0.124021
pop_estimate_2015	0.102256
percent_white	0.094430
percent_adults_college_degree	0.089065
voter_turnout	0.087665
median_household_income	0.065058
percent_white_collar	0.062845
percent_poverty	0.061811
2015_unemployment_rate	0.049999
dem_won_2008	0.033767

Conclusion

- Similar to the decision tree models, the accuracy score is very high.
- The “Important Features” output shows how important the “Democrat won in 2012” variable is to predicting who won in 2016; however, its importance is diminished for swing counties.
- Interestingly, the 5-year population change plays a more important role in predicting swing county results
- Economic variables like household income, poverty, and unemployment aren’t as relevant in determining election results

Model 6: Extra Tree Model

Key Diagnostics:

-Calculated accuracy score after testing on test dataset: 0.928

-Feature Importance calculation:

All Counties

	importance
dem_won_2012	0.446297
dem_won_2008	0.090646
percent_white	0.071528
pop_estimate_2015	0.069511
percent_adults_college_degree	0.062716
net_population_change_2010-2015	0.058446
voter_turnout	0.049069
percent_white_collar	0.042536
median_household_income	0.039344
percent_poverty	0.037681
2015_unemployment_rate	0.032226

Swing Counties

	importance
dem_won_2012	0.251254
pop_estimate_2015	0.098765
percent_white	0.091109
percent_adults_college_degree	0.088686
net_population_change_2010-2015	0.088499
voter_turnout	0.083494
percent_white_collar	0.067040
median_household_income	0.065198
percent_poverty	0.063497
2015_unemployment_rate	0.057521
dem_won_2008	0.044938

Conclusion

-Accuracy score slightly higher than the Random Forest model.

-With the Extra Tree model, the importance of the “Democrat won in 2012” is even higher than the Random Forest model.

- Population size is a more important factor for swing counties; economic variables again rank low in importance for determining election results

Model Summary

Model	Accuracy Score	R^2	Other diagnostics	Comment
Linear Regression	N/A	0.65	N/A	Strong R-Squared suggests does a good job at predicting % of county voting democrat.
Logistic Regression	0.907	0.69	AIC=0.98	With high AIC score, high accuracy score, and reasonable number of features, a good model.
Decision Tree	0.920	N/A	Best Hyperparameters found: Max Depth=7, max_features=7, min samples per split=100	Better accuracy score than logistic model, but harder to interpret the model itself.
Decision Tree w/ Bootstrap Sampling	0.926	N/A	Best Hyperparameters found: Max Depth=6, max_features=7, min samples per split=100	Using Bootstrapping to sample, got an even higher accuracy score than non-Bootstrap sampling. However optimal tree found is twice as large.
Random Forrest	0.926	N/A	N/A	Random Forrest didn't perform better than prior methods above. However was able to view feature importance and draw insights from that.
Extra Tree	0.928	N/A	N/A	Same story as with Random Forrest

Final Conclusions and Learnings

- The most important factor for determining the winner of any given county's presidential vote is the result of the prior election. The percent of the white population also seems to be a highly correlative variable in predicting election outcome (not in favor of the Democrats) regardless of county type (swing vs. non-swing).

- It was surprising that some predictors DID NOT perform very well, such as the Unemployment Rate, Poverty Rate, Median income, and Yes/No whether the county voted democratic in 2008.

  - The lack of significance of these economic metrics in determining election results could inform subsequent analyses focused on non-socioeconomic variables, especially psychographic preferences of voters.
  - The diminished importance of the 2008 results suggests how aberrant that election was and how variable the results can be from one election cycle to another.

- Each of the models seemed to perform very well, with the exception of the linear regression, which performed only moderately well. Based solely on the diagnostics there isn't one model that stood out from the rest. It might be best to stick with the logistic and linear regression models as these are very easy to interpret and explain to others.

- There are some differences in the variable rankings between swing and non-swing counties (for example population change ranks higher for swing counties). Subsequent analysis could include analyzing further segmentations of counties by region, change over time, etc. to understand other salient differences.

Lessons Learned

-It is informative to test a variety of models on the same problem, as different models can reach different/similar conclusions depending on the methodology and variables used.

-The results don't always align with preconceptions; in this case, how variables impact voter behavior.

- An interesting subsequent analysis could explore how these predictors estimate historical elections going back many election cycles; Do these predictors change a lot between elections? Are there significant differences between presidential election years and mid-cycle years?

- Public data are messy and often come from multiple disparate sources, which leads to data inconsistencies and consolidation errors.

- Relative lack of familiarity with Python made it challenging to conduct some of the analysis

- It can be difficult to identify the most salient visualizations and variables for analysis and presentation to ensure the story is compelling

Appendix

Please refer to the submitted HTML and .csv files to see Python code and the data set used for our respective models