2016 Election: Revisited

STAT 541 Project by Travis Benedict and Jordan Pflum
December 3, 2018

Start

OVERVIEW

2016 Election Background
Brief overview of the 2016 Election and statistical policial leans commonly employed

Data Overview
Describe data sets and predictors contained within

Exploratory Analysis
Initial Exploration of Data

Project Goal
Central guiding question of project

Methods Used
PCA, LDA, exemption of Clustering

Comparison
PCA vs LDA

Conclusion
Concluding thoughts

Q&A

2016 Election Background

Historical Information

Candidates

The Republican ticket was held by businessman **Donald Trump** while the Democratic ticket was held by former Secretary of State **Hillary Clinton**

Result

While Donald Trump **lost the popular vote** to Hillary Clinton by more than 2.8 million votes, he **won 30 states** and a decisive **304 electoral votes** compared to 227, becoming the **47th president**

Message

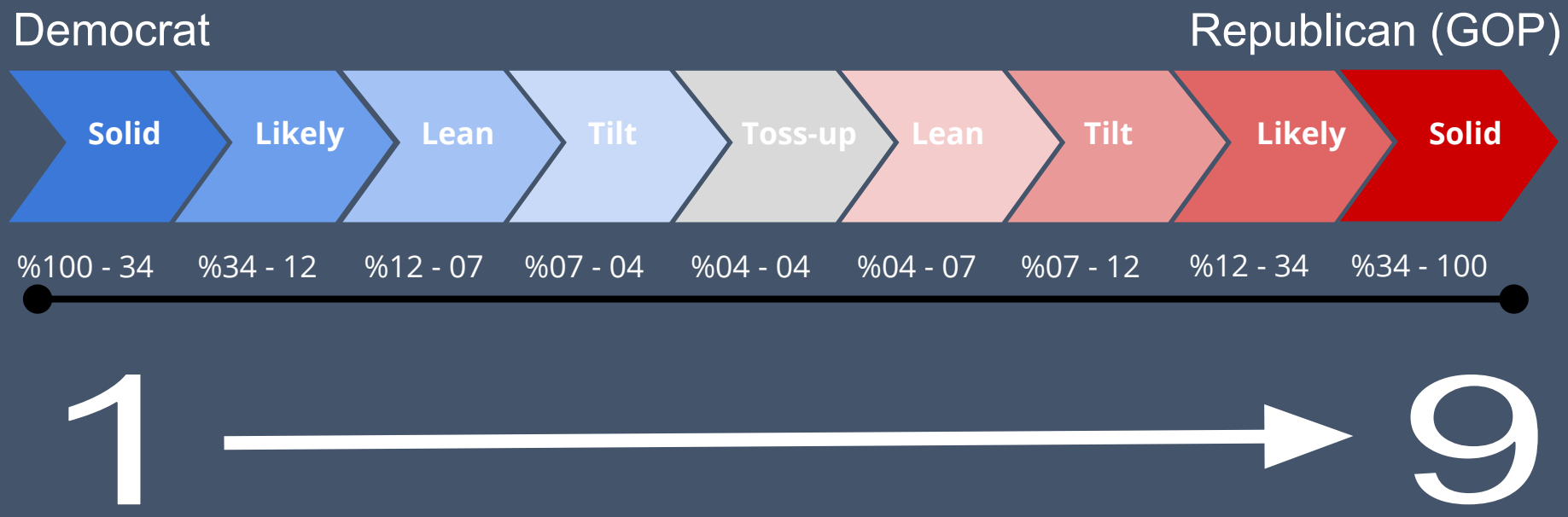
Donald Trump's populist, nationalist campaign, promising to "**Make America Great Again**", starkly contrasted Clinton's expansion and promotion of **racial**, **LGBTQ**, and **women's rights**

Key Demographics

Trump's appeal to **white working-class** voters outside major cities in **pivotal manufacturing states** proved to be a key factor.

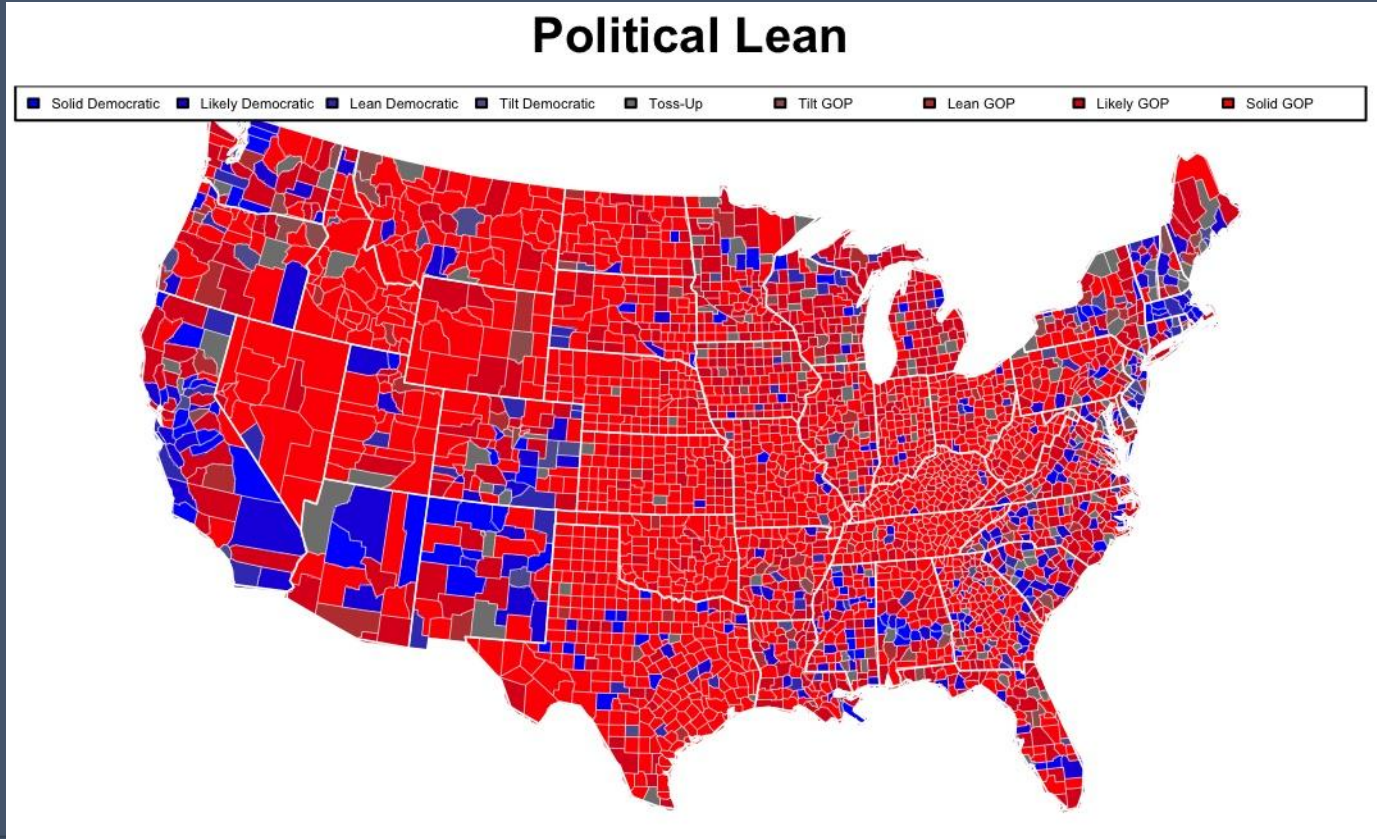
2016 Election Background

Statistical Political Lean



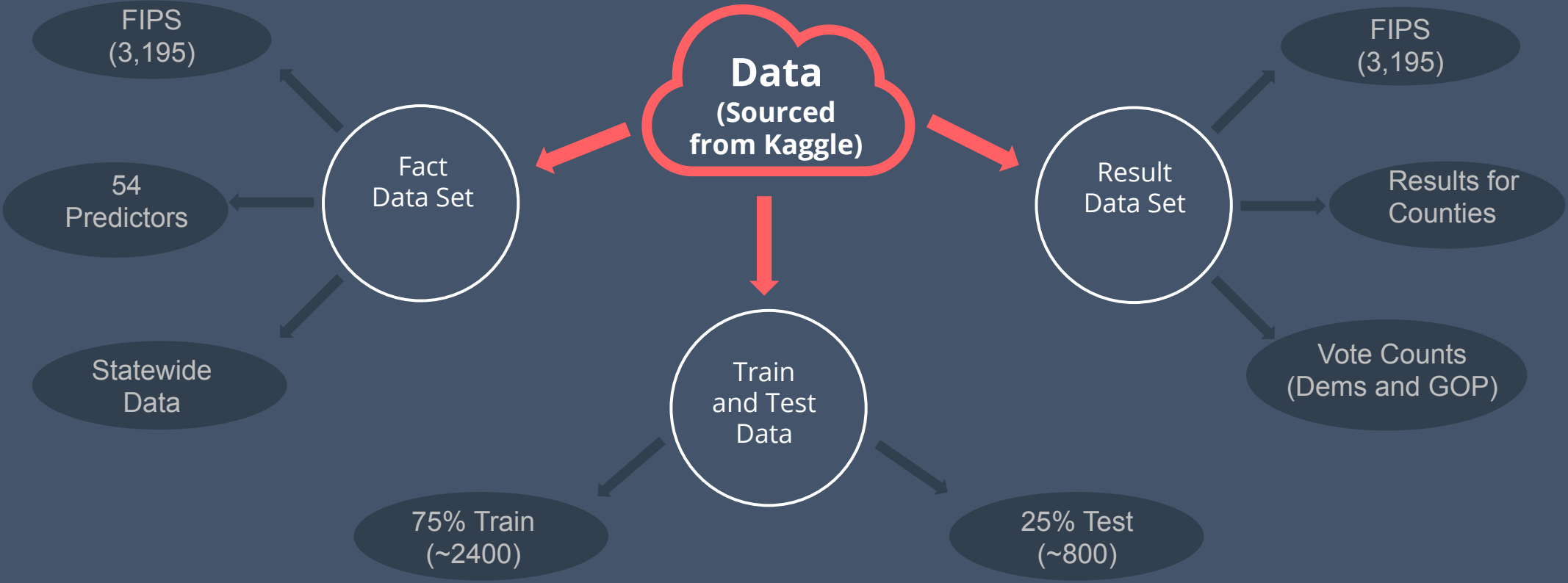
2016 Election Background

Statistical Political Lean Map



Data Overview

Data Sets Description



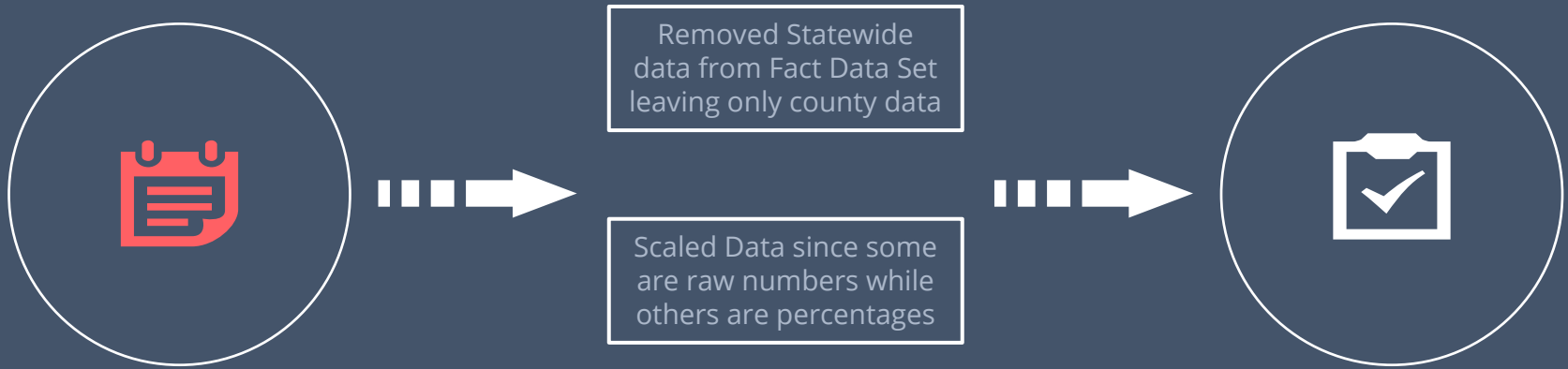
Data Overview

Fact Data Set Predictors (Examples)

Predictor Code	Predictor Description
RHI725214	Hispanic or Latino, percent, 2014
POP815213	Language other than English spoken at home, pct age 5+, 2009-2013
EDU685213	Bachelor's degree or higher, percent of persons age 25+, 2009-2013
INC910213	Per capita money income in past 12 months (2013 dollars), 2009-2013
PVY020213	Persons below poverty level, percent, 2009-2013
PST045214	Population, 2014 estimate
AGE295214	Persons under 18 years, percent, 2014
AGE775214	Persons 65 years and over, percent, 2014
SEX255214	Female persons, percent, 2014
RHI225214	Black or African American alone, percent, 2014
VET605213	Veterans, 2009-2013

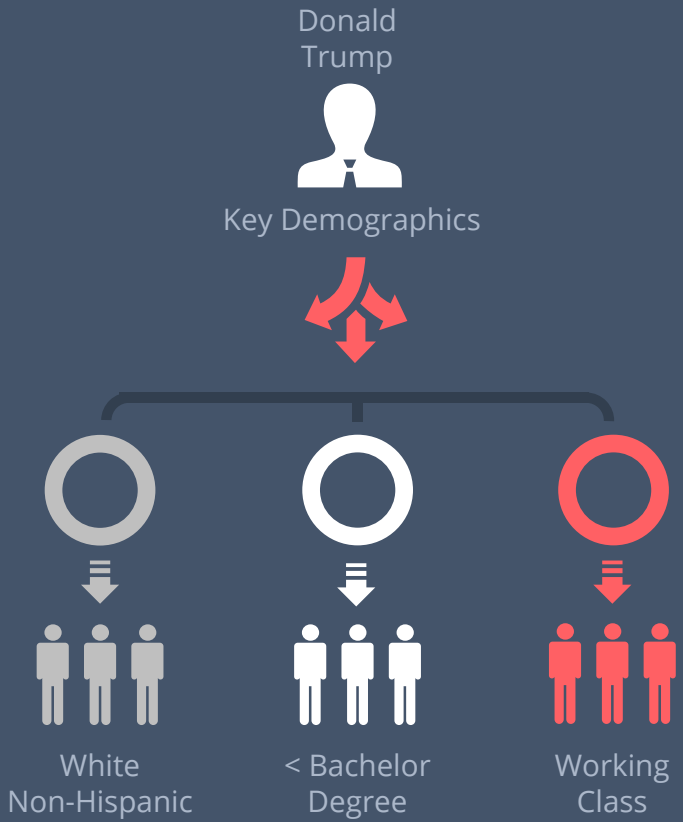
Data Overview

Cleaning the Data



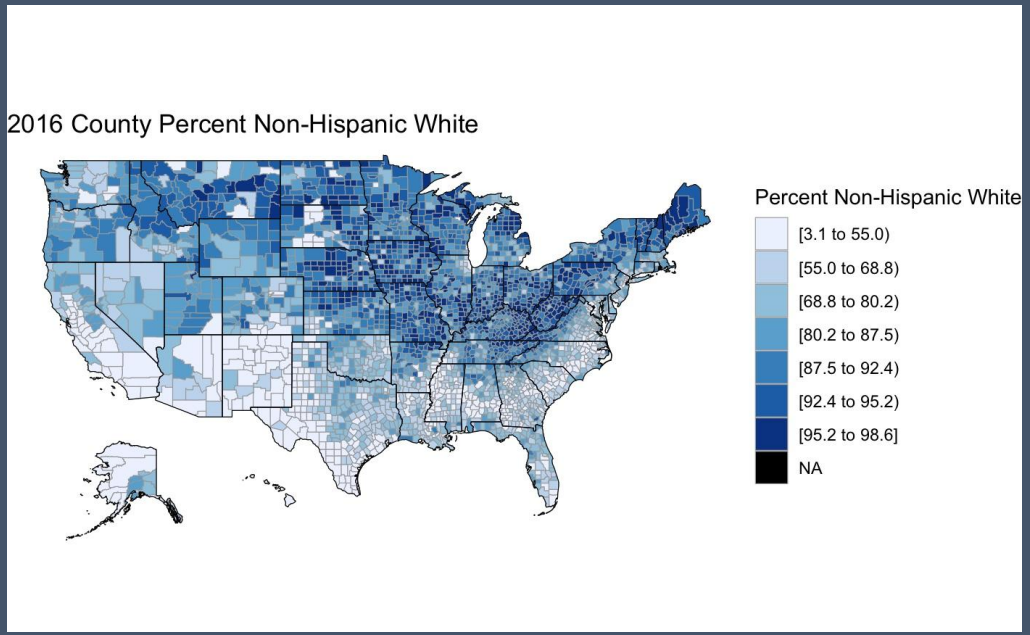
Explanatory Analysis

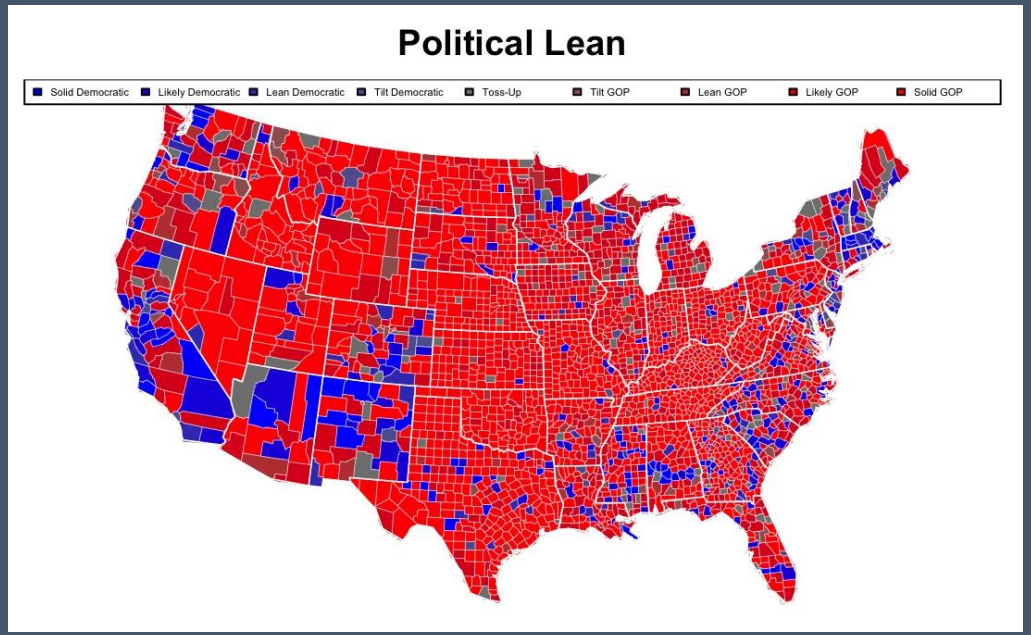
Trump's Base



Exploratory Analysis

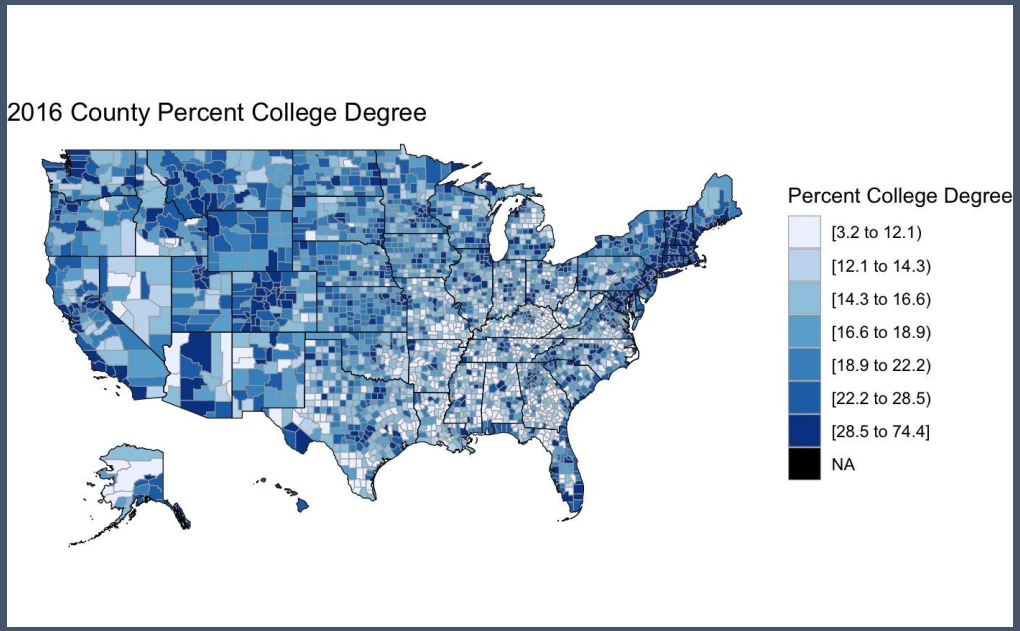
White Non-Hispanic

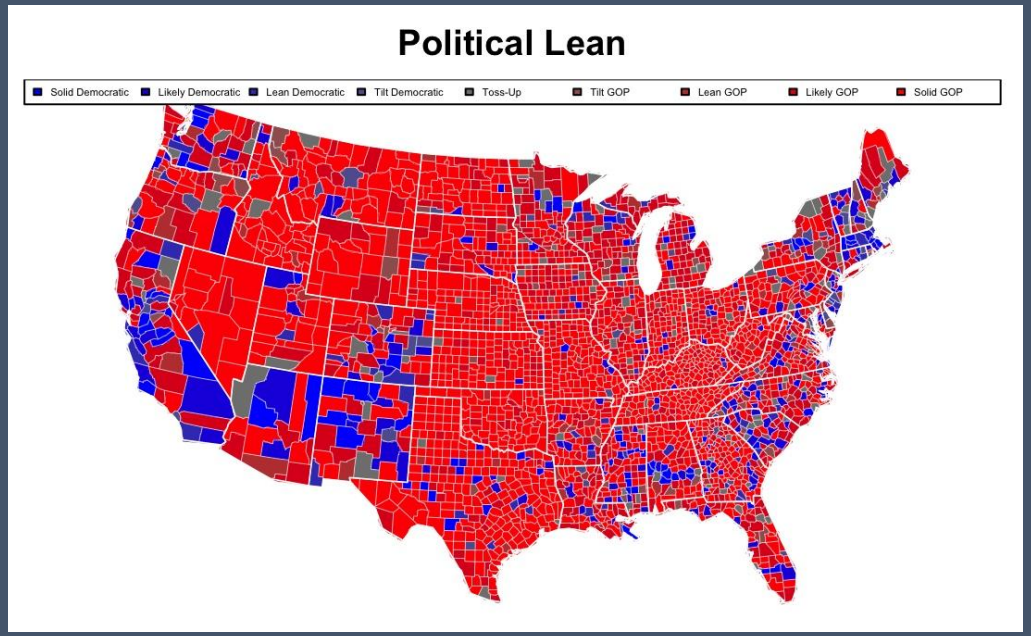




Exploratory Analysis

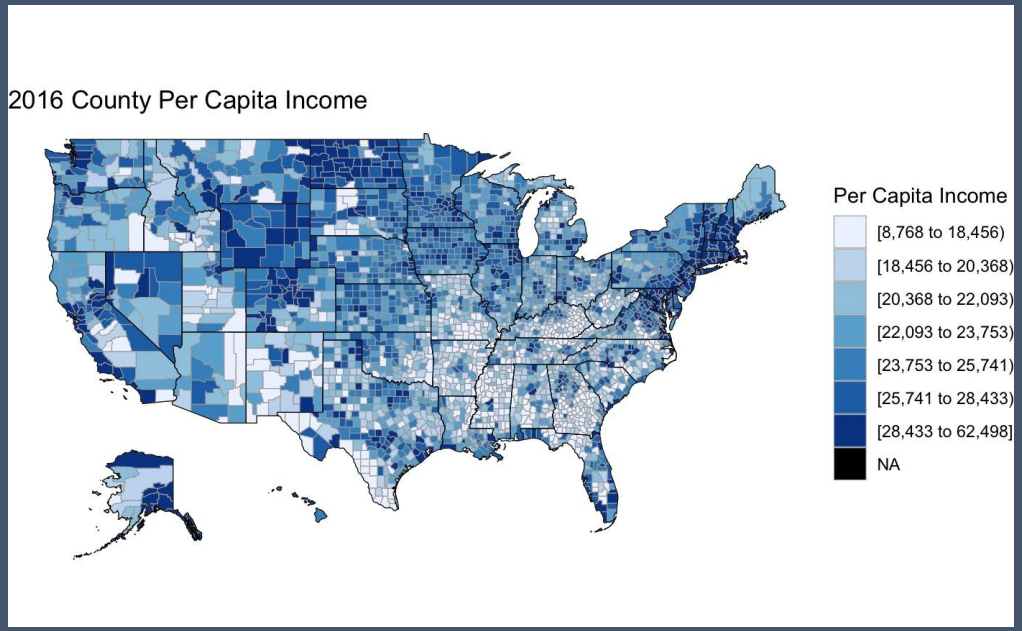
< Bachelor Degree

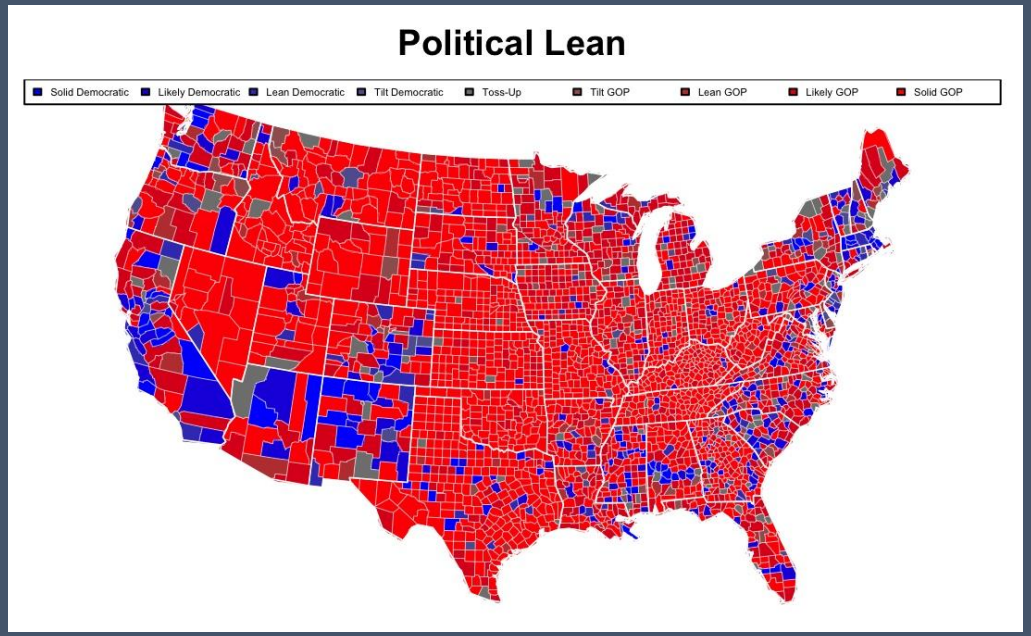




Exploratory Analysis

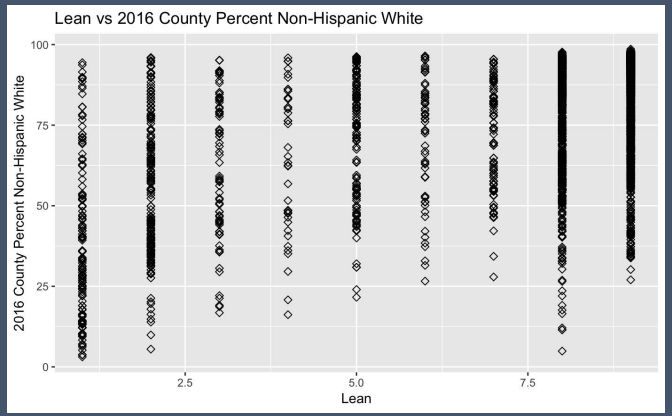
Working Class



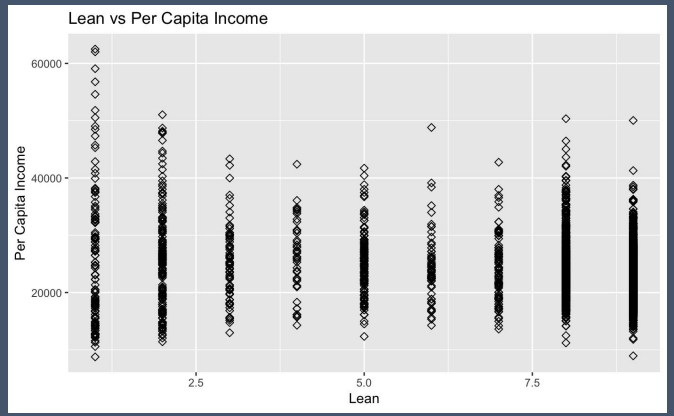


Explanatory Analysis

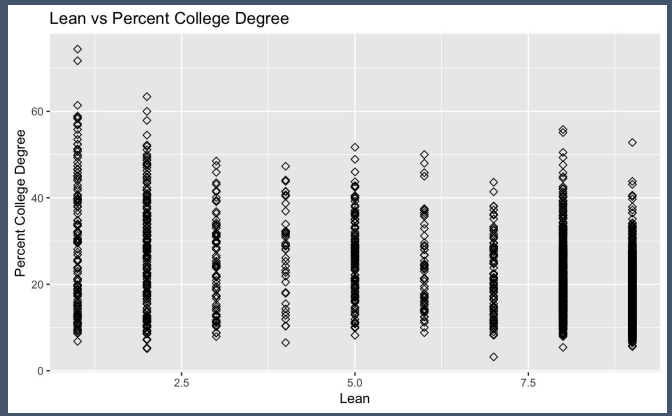
Scatter Plots



Upwards Trend



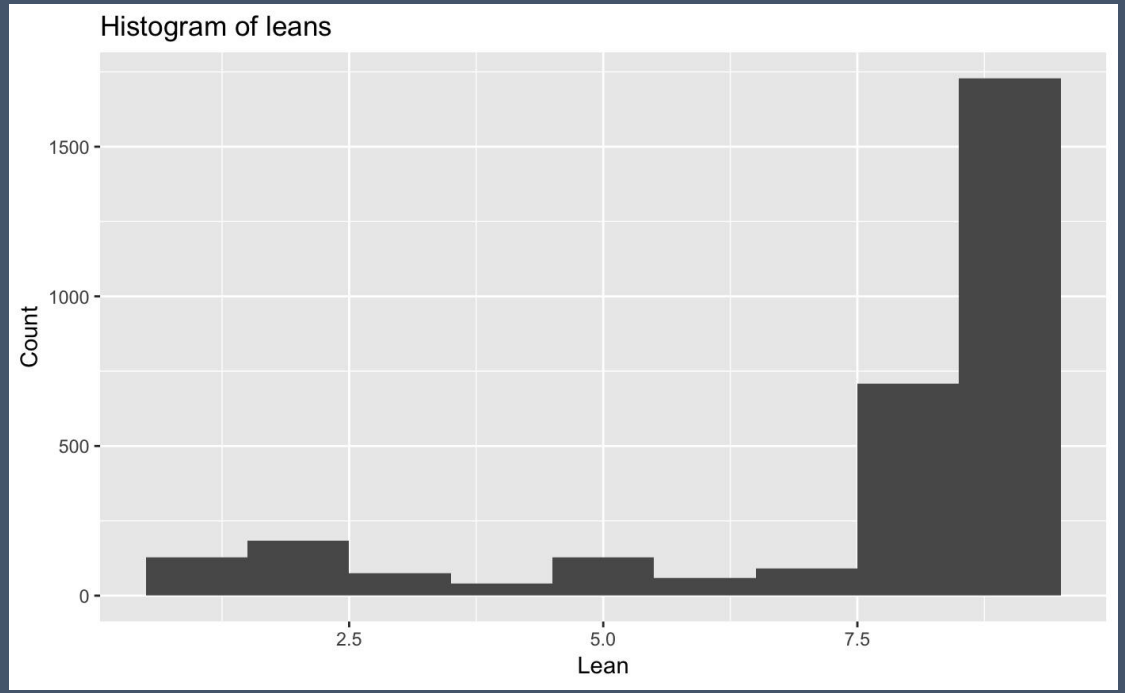
High Variance



High Variance

Explanatory Analysis

Histogram



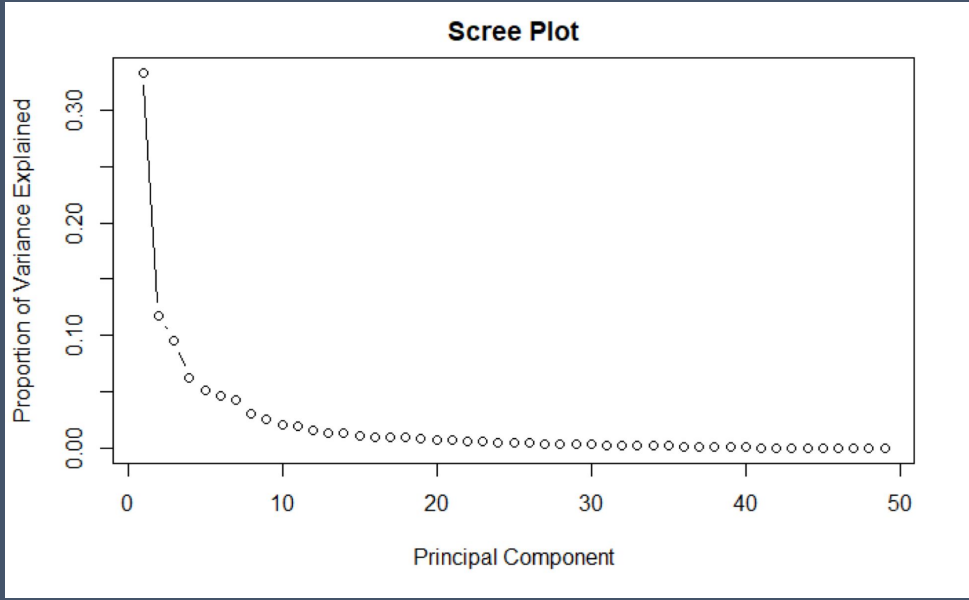
Heavily Skewed GOP

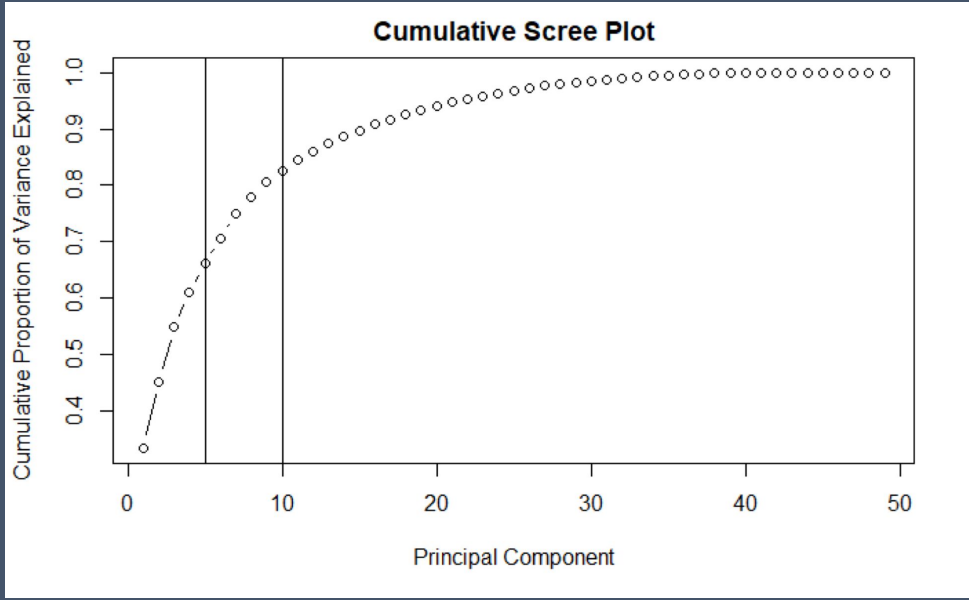
Central Question

How can we reduce the amount of information we have while still gaining insight into voting patterns?

Principal Component Analysis

Variance for each component





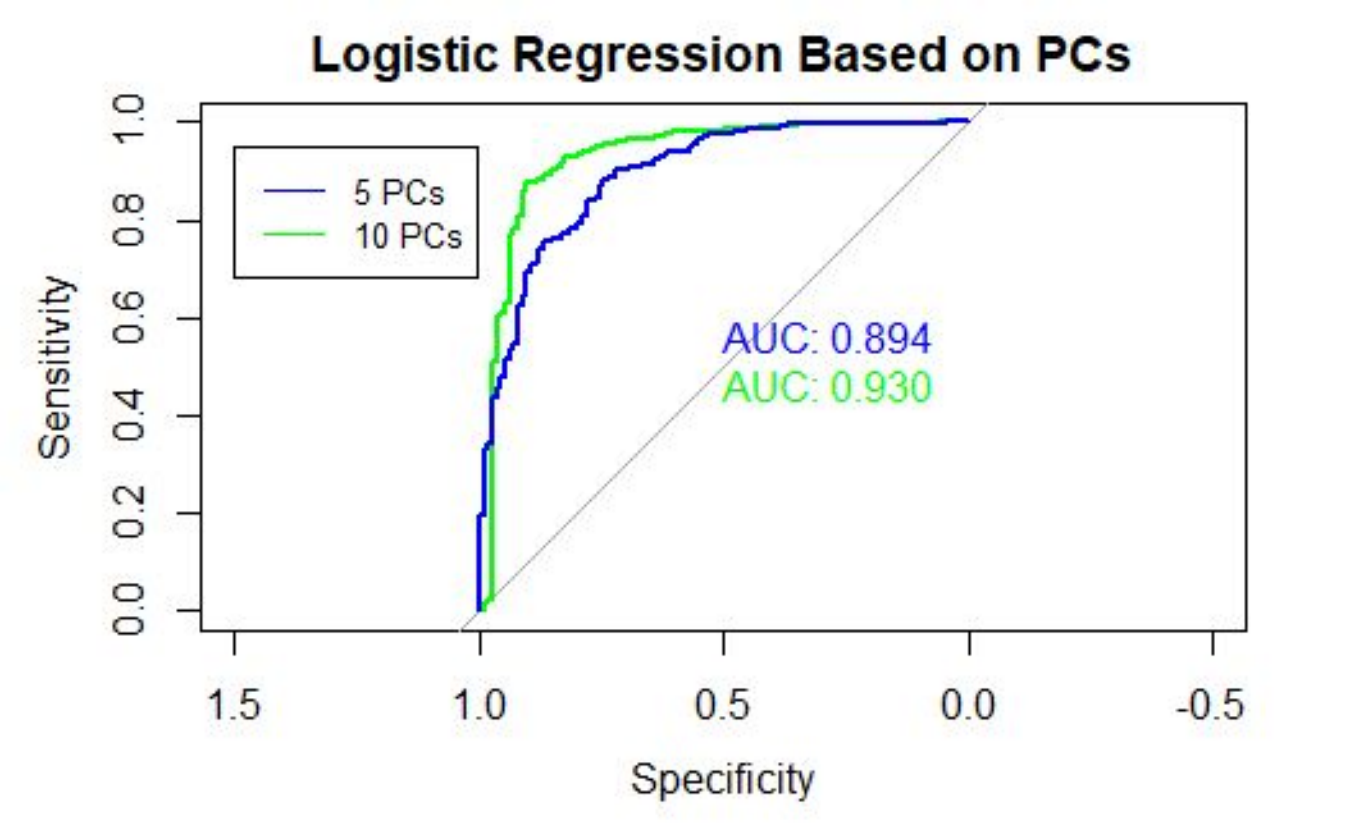
The first 5 and 10 principal components explain about 66% and 80%, respectively. Hardly any variance is added after the 20th principal component.

Principal Component Analysis

Predicting the winning candidate for a county

ROC Curve: Graphical representation of True Positive Rate (Sensitivity) and True Negative Rate (Specificity) of a classifier as the threshold of discrimination changes.

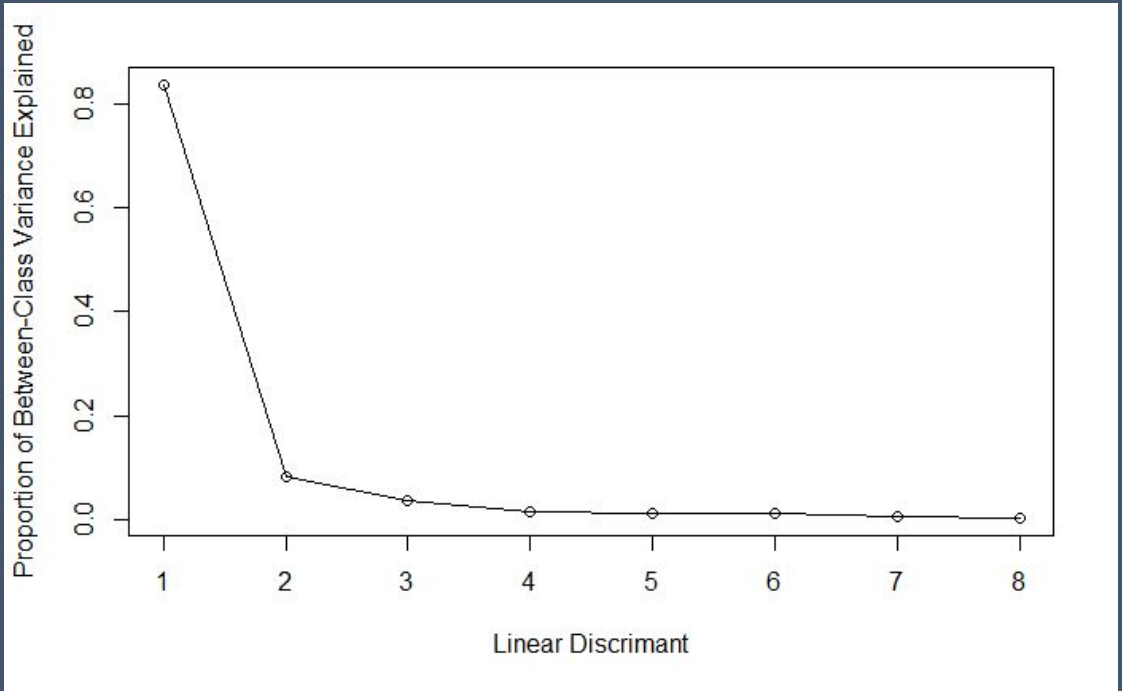
AUC: Area under curve for an ROC curve. “The probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.”



https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Linear Discriminant Analysis

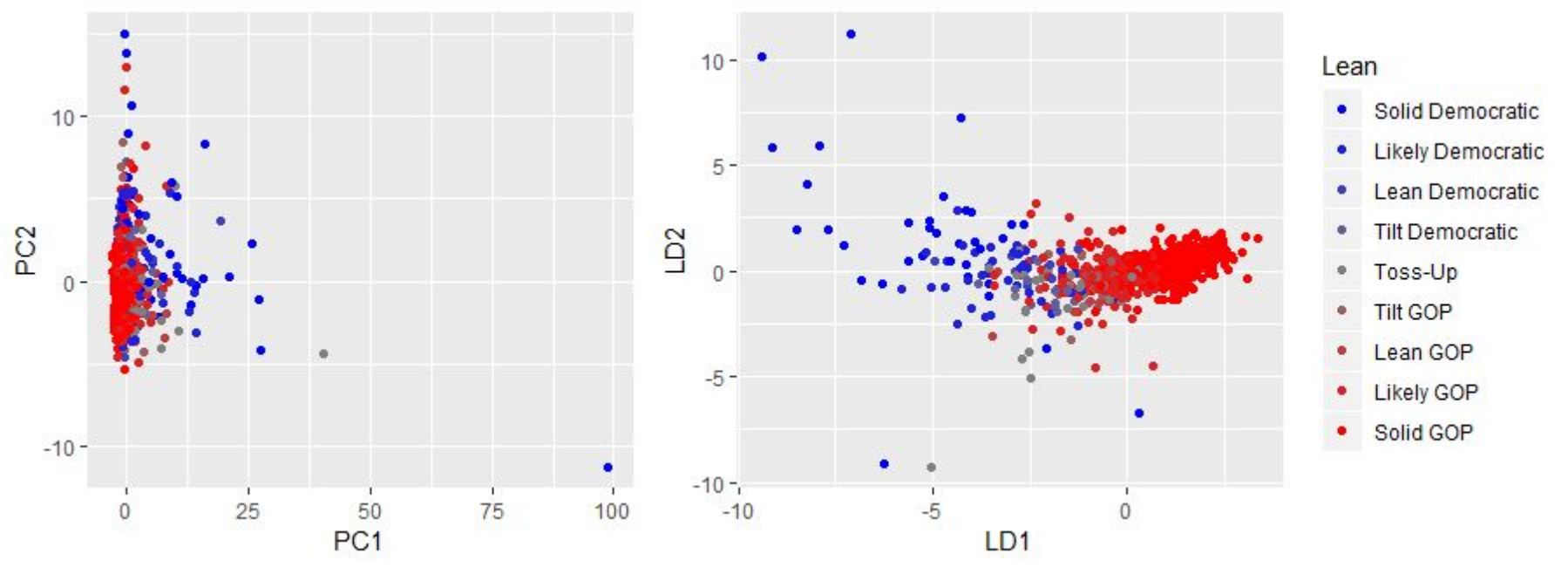
Predicting the partisan lean for a county



	RHI825214	EDU685213	INC110213
Solid Democratic	-1.8756979	1.0598971	0.1032907
Likely Democratic	-1.1179536	0.9251625	0.3508663
Lean Democratic	-0.7659398	0.6504934	0.3245943
Tilt Democratic	-0.6651658	0.9854434	0.5344588
Toss-Up	-0.3236268	0.6922105	0.2212962
Tilt GOP	-0.3033335	0.5090255	0.1632551
Lean GOP	-0.3431409	0.1377018	-0.0546923
Likely GOP	-0.0100408	0.1023369	0.1868314
Solid GOP	0.3799982	-0.3553081	-0.1820781

PCA vs LDA

Predicting the partisan lean for a county

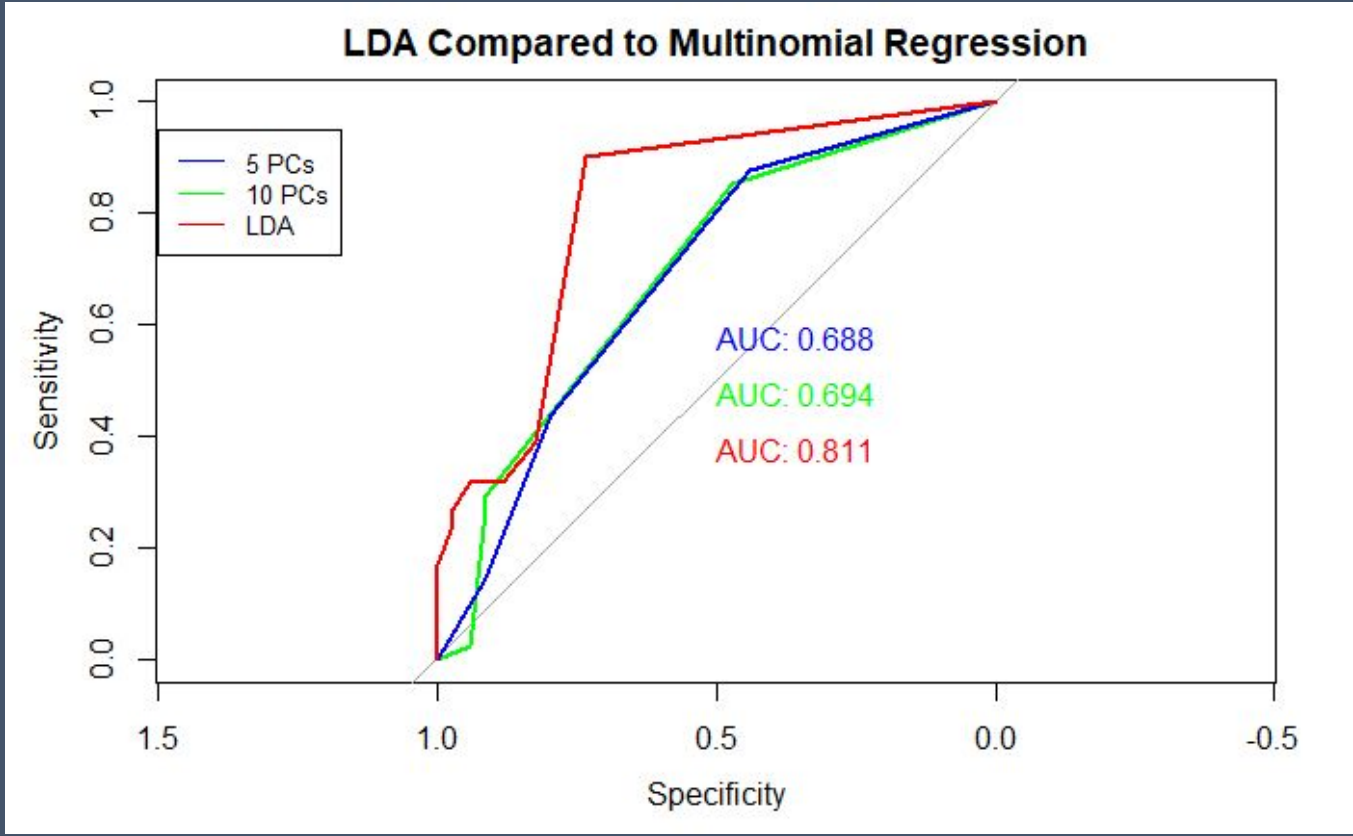


PCA provides some separation between Solid Democratic and Solid Republican but hardly any separation for the other leans. LDA provides more separation between the leans.

PCA vs LDA

Predicting the partisan lean for each county

As expected LDA (supervised) performs much better than PCA (unsupervised) even with a similar number of dimensions.



Clustering

Usefulness in project

Clustering

Issues

1. Based on the EM algorithm used by MClust the quality of clusters significantly decreased after 7 clusters
2. This meant that clusters did not align well with partisan lean basically making them useless

Result

Opted not to use clustering as a method to determine lean of county

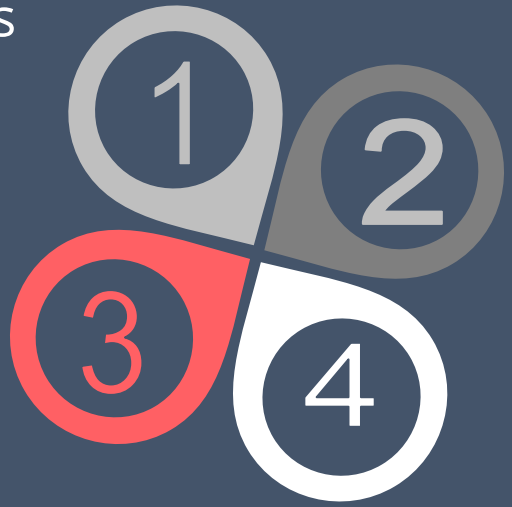
Conclusion

Principal Component Analysis

For predicting just the outcome of elections and not the partisan lean PCA is highly effective. Even just 5 dimensions were enough for a reasonably powerful binary classifier.

Clustering

Our attempts to cluster the data were not very useful. Even when well defined clusters formed they did not correspond well to the partisan leans.



Linear Discriminant Analysis

Linear discriminant analysis was highly effective in predicting partisan leans while also reducing the data down to just 8 dimensions.

Dimension Reduction

Overall this data set has a lot of unnecessary information. Worthwhile predictive power can be gleaned just from a few dimensions once useful projections are applied.

QUESTION TIME!

Have a question? Ask NOW!