Predicting US Presidential Elections using demographic data

STEVEN JORDAN DEPAUL UNIVERSITY DSC 540

11/13/2019

Project Goals

1. Predict the percentage difference between Democrat and Republican vote totals in any county
2. Identify the most important demographic features

kaggle Search Competitions Datasets Notebooks Discussion Courses •••

Dataset

66

2012 and 2016 Presidential Elections

Election results with county information on race, income and education

Joel Wilson • updated 3 years ago (Version 2)

Data Kernels (39) Discussion (4) Activity Metadata New Notebook

Usability 6.8 **Tags** politics

Description

These data files contain election results for both the 2012 and 2016 US Presidential Elections, include proportions of votes cast for Romney, Obama (2012) and Trump, Clinton (2016).

The election results were obtained from this Git repository: https://github.com/tonmcg/County_Level_Election_Results_12-16

The county facts data was obtained from another Kaggle election data set: https://www.kaggle.com/benhamner/2016-us-election

AMERICAN FactFinder

United States Census Bureau

Feed

MAIN | COMMUNITY FACTS | GUIDED SEARCH | ADVANCED SEARCH | DOWNLOAD CENTER

As of July 1, 2019 data.census.gov is now the primary way to access Census Bureau data, including the latest releases from the 2018 American Community Survey and 2017 Economic Census and t Census and more. American FactFinder will be decomissioned in 2020.

Read more about the Census Bureau's transition to data.census.gov .

Search - Use the options on the left (topics, geographies, ...) to narrow your search results

Your Selections

Search using...

Program:
American Community Survey

clear all selections and start a new search

load search | save search

Search using the options below:

Topics (age, income, year, dataset, ...)

Geographies (states, counties, places, ...)

Race and Ethnic Groups (race, ancestry, tribe)

Industry Codes (NAICS industry, ...)

EEO Occupation Codes (executives, analysts, ...)

Search Results: 1-25 of 75,531 tables and other products match 'Your Selections'

Refine your search results: topic or table name | state, county or place (optional) | GO

topics | race/ancestry | industries | occupations

Selected: View | Download | Compare | Clear All | Reset Sort

Show results from: All available programs

	ID	Table, File or Document Title	Dataset	About
	S0101	AGE AND SEX	2017 ACS 5-year estimates	
	S0101	AGE AND SEX	2017 ACS 1-year estimates	
	S0102	POPULATION 60 YEARS AND OVER IN THE UNITED STATES	2017 ACS 5-year estimates	
	S0102	POPULATION 60 YEARS AND OVER IN THE UNITED STATES	2017 ACS 1-year estimates	
	S0102PR	POPULATION 60 YEARS AND OVER IN PUERTO RICO	2017 ACS 5-year estimates	
	S0102PR	POPULATION 60 YEARS AND OVER IN PUERTO RICO	2017 ACS 1-year estimates	
	S0103	POPULATION 65 YEARS AND OVER IN THE UNITED STATES	2017 ACS 5-year estimates	
	S0103	POPULATION 65 YEARS AND OVER IN THE UNITED STATES	2017 ACS 1-year estimates	
	S0103PR	POPULATION 65 YEARS AND OVER IN PUERTO RICO	2017 ACS 5-year estimates	
	S0103PR	POPULATION 65 YEARS AND OVER IN PUERTO RICO	2017 ACS 1-year estimates	
	S0501	SELECTED CHARACTERISTICS OF THE NATIVE AND FOREIGN-BORN POPULATIONS	2017 ACS 5-year estimates	

Data Structure

2 Data Sets

- 2012 and 2016 elections – using 2011 and 2015 demographic data

Dependent Variable

- % Margin Between the Dem (+) and GOP (-) vote share

	county_name	diff_2016
fips		
1001	Autauga County	-0.494789
1003	Baldwin County	-0.577862

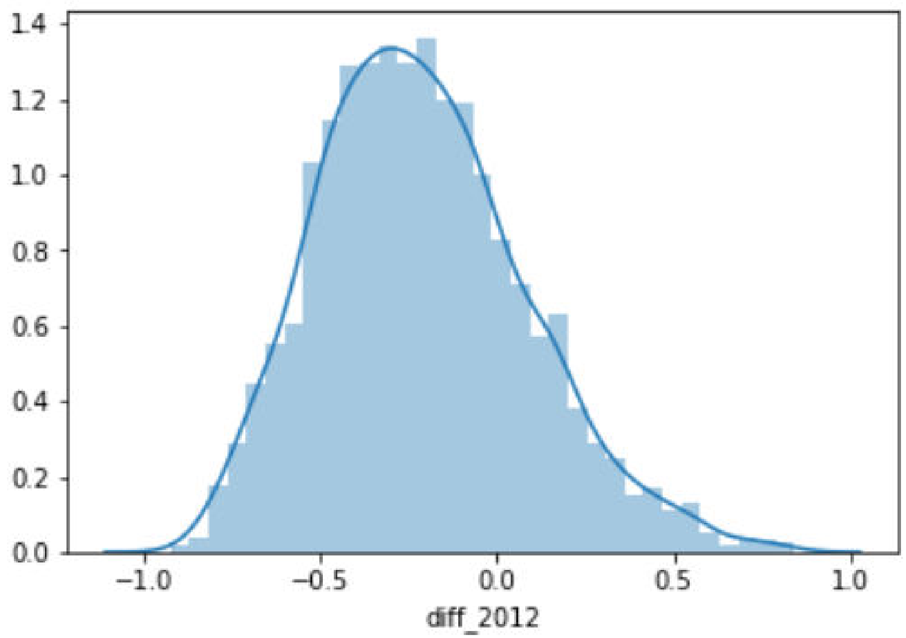
Data Structure

Features

• Population	2010 Census, 2011/2015 Estimates
• Demographics	% White, % Black, % Female, % over 65, etc.
• Housing	# of units, median value, median rent, household size, etc.
• Education	% with High School Degree, % with Bachelor's Degree
• Employment	% in poverty, median income
• Business	# of employers, # of women-owned firms, total payroll, etc.

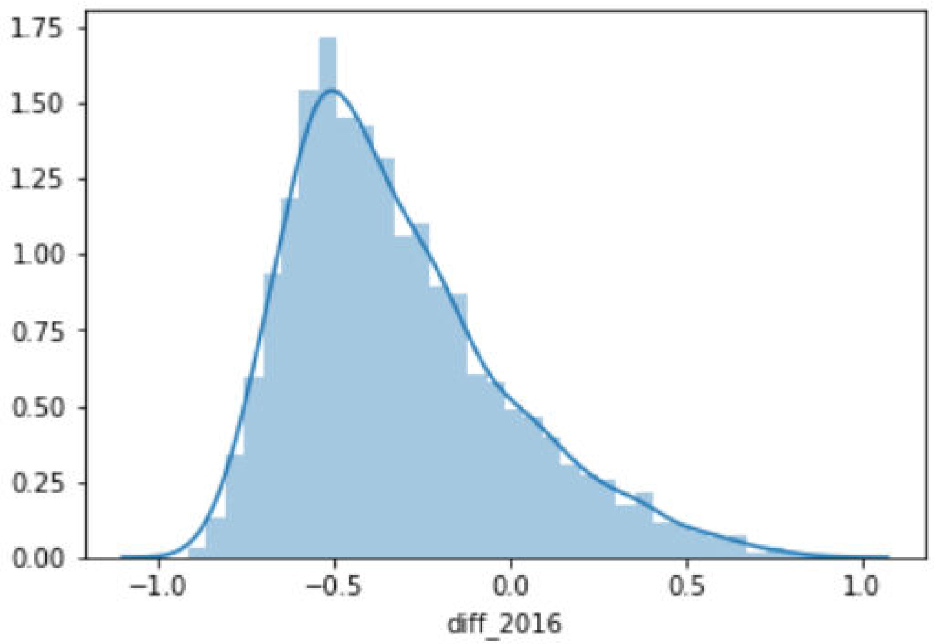
Vote Difference Distribution

Skewness: 0.479841
Kurtosis: 0.120025



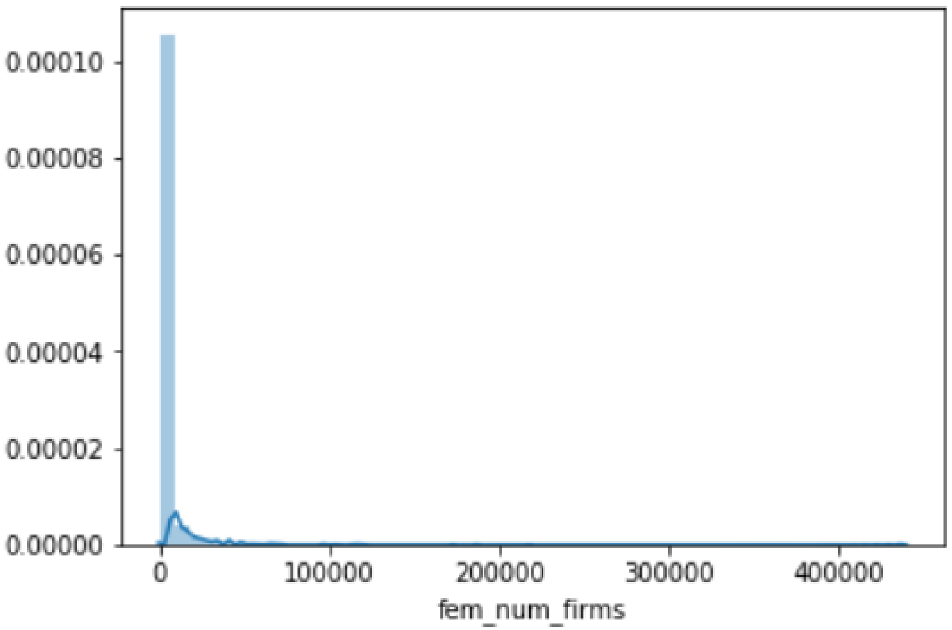
2012 Vote % Distribution

Skewness: 0.896549
Kurtosis: 0.522459

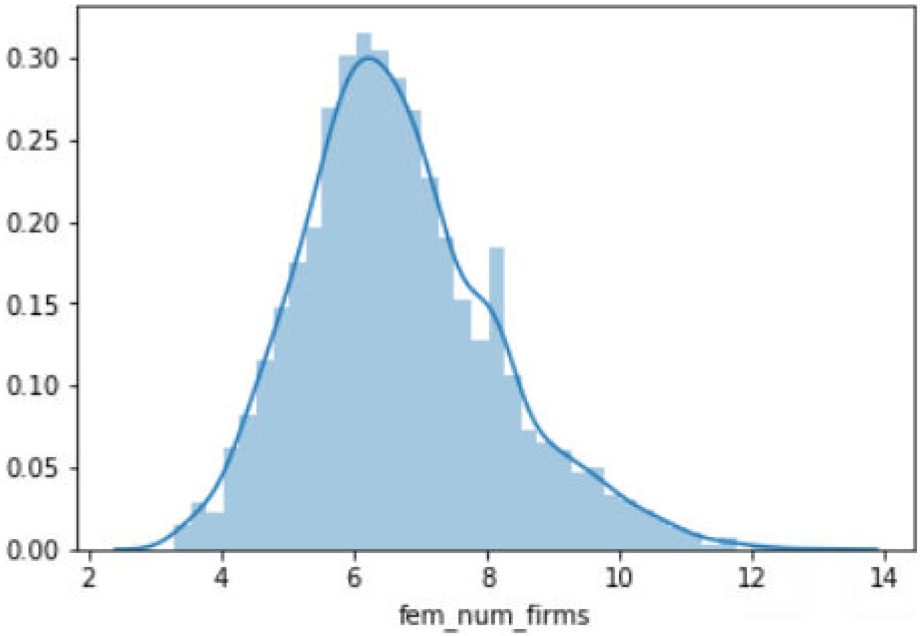


2016 Vote % Distribution

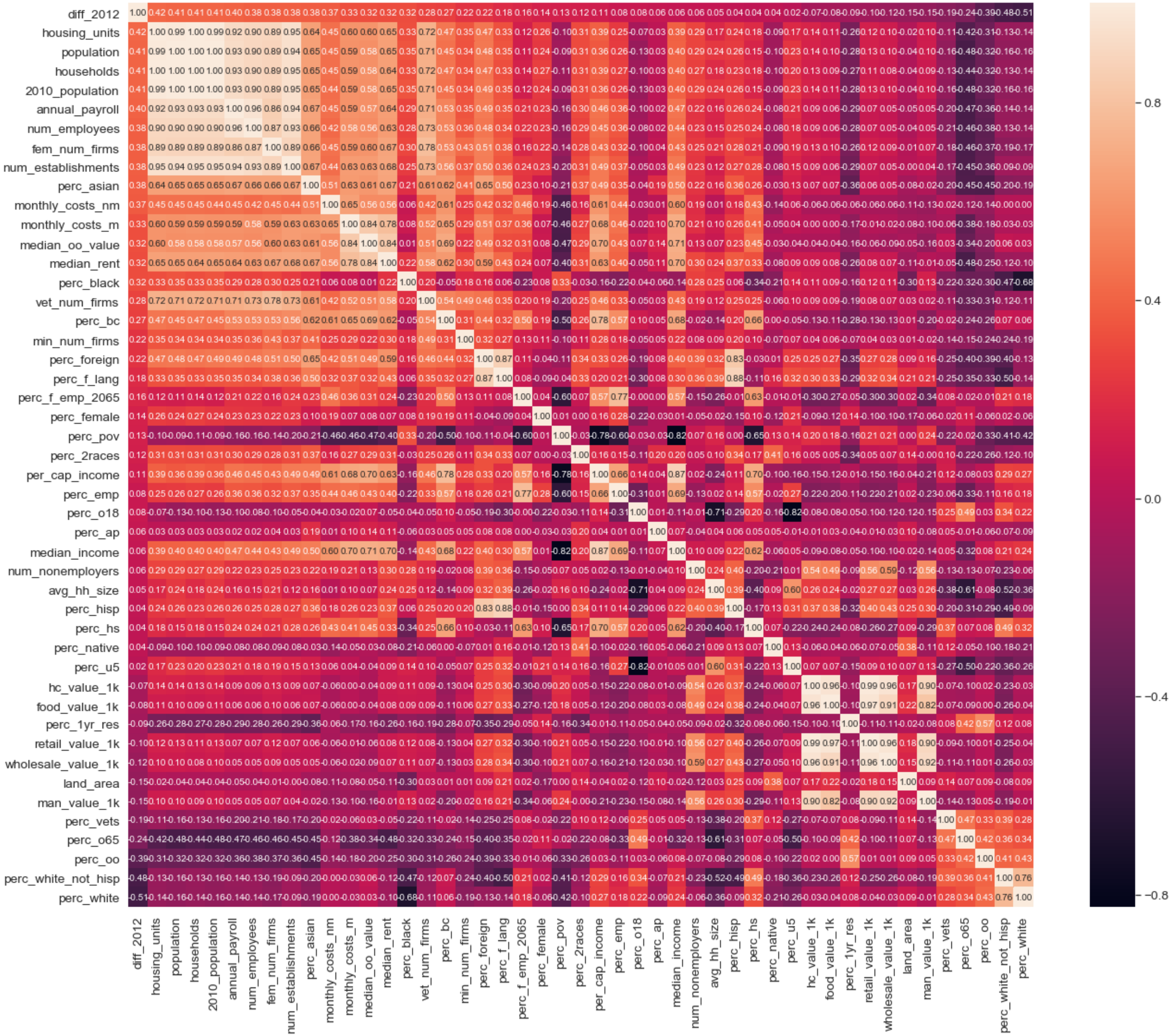
Feature Distributions



2012 Distribution of # of Female-Owned Firms



2012 Distribution of Log # of Female-Owned Firms

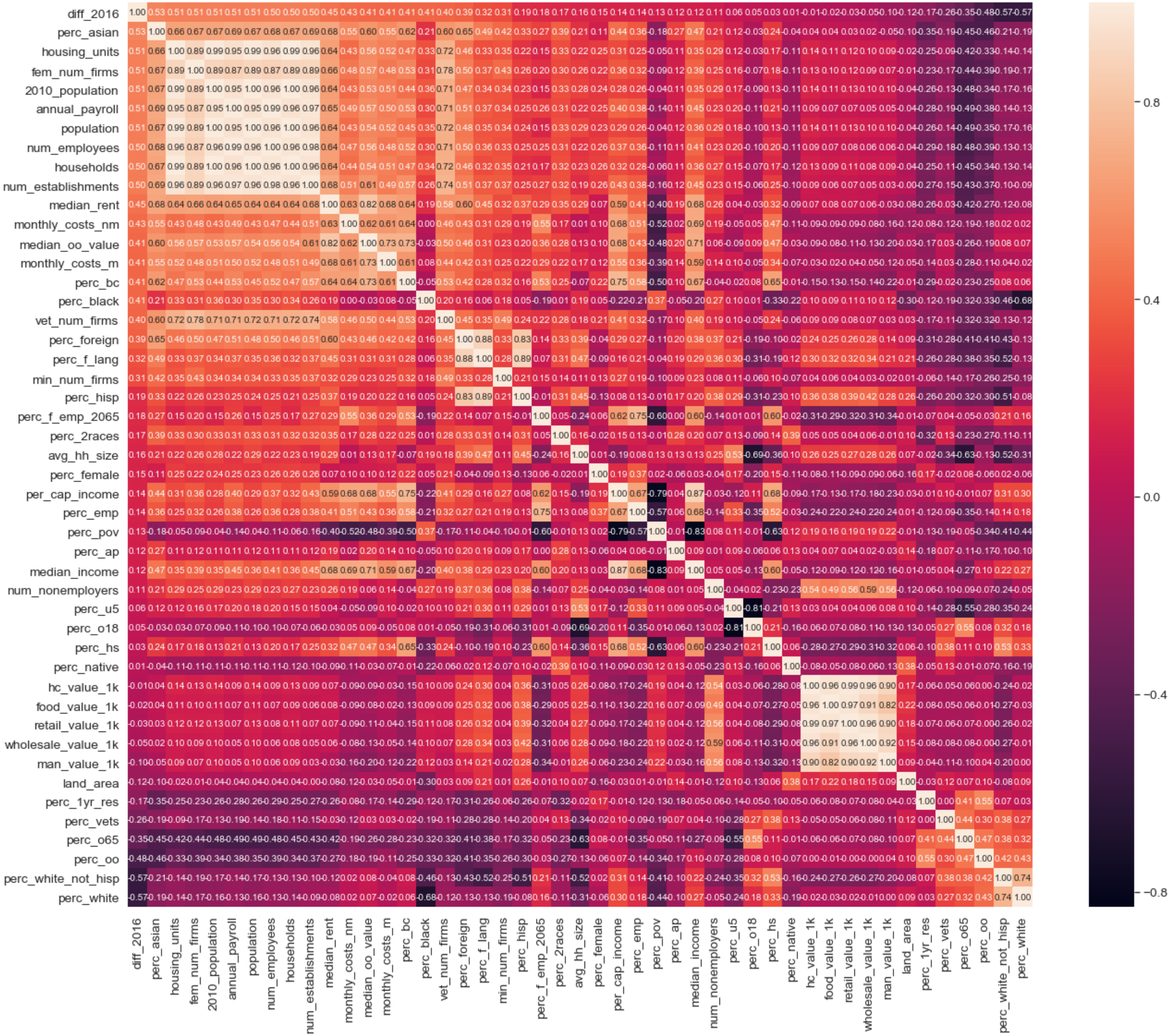


Most Correlated Features – Dem

of Households/Population
# of Businesses/Payroll
Monthly Costs (e.g. Rent)
% Asian
% Bachelor Degrees

Most Correlated Features – GOP

% White
% Homeownership
% Over 65
% Veterans
$ Manufacturing



Most Correlated Features – Dem

% Asian
# Households, Population
# of Female-owned Firms
# of Businesses/Payroll
# of Businesses/Payroll

Most Correlated Features – GOP

% White
% Homeownership
% Over 65
% Veterans
% of non-moving households (1 year)

Models Building

Types of Algorithms Used

- Linear: Lasso, Ridge
- Ensemble: Random Forest, Gradient Boosting
- Support Vector Regression

Model Building Process

- Wrapper-Based Feature Selection
- Randomized Hyper-Parameter Search
- Grid Search (3-fold cross-validation)

Model Generalization

	2012		2016		BOTH	
	MAE	Exp Var	MAE	Exp Var	MAE	Exp Var
2012						
2016						
BOTH						

Lasso (CV)

	2012		2016		BOTH	
	MAE	Exp Var	MAE	Exp Var	MAE	Exp Var
2012	0.15	42%	1.00	15%	0.58	0%
2016	0.22	55%	0.12	73%	0.17	56%
BOTH	0.15	60%	0.12	73%	0.14	67%

Most Important Features

Population

% over 65

% over 18

Median Home Value

Monthly Costs (People w/o mortgage)

Ridge (CV)

	2012		2016		BOTH	
	MAE	Exp Var	MAE	Exp Var	MAE	Exp Var
2012	0.16	48%	0.80	38%	0.48	0%
2016	0.21	56%	0.12	73%	0.17	58%
BOTH	0.15	59%	0.12	73%	0.14	67%

<u>Most Important Features</u>

Unable to determine

Random Forest

	2012		2016		BOTH	
	MAE	Exp Var	MAE	Exp Var	MAE	Exp Var
2012	0.11	75%	0.14	70%	0.13	73%
2016	0.16	52%	0.12	74%	0.14	64%
BOTH	0.14	62%	0.12	75%	0.13	70%

Most Important Features

% White, Not Hispanic

% White

Monthly Costs (People w/o mortgage)

% Black

Population

Gradient Boosting

	2012		2016		BOTH	
	MAE	Exp Var	MAE	Exp Var	MAE	Exp Var
2012	0.10	78%	0.12	72%	0.11	76%
2016	0.16	53%	0.11	76%	0.14	66%
BOTH	0.14	62%	0.12	75%	0.13	70%

Most Important Features

% Black

% White

Monthly Costs (People w/ mortgage)

Population

% over 18

SVM

	2012		2016		BOTH	
	MAE	Exp Var	MAE	Exp Var	MAE	Exp Var
2012	0.12	74%	0.62	0%	0.37	0%
2016	0.22	45%	0.09	83%	0.15	58%
BOTH	0.11	75%	0.09	84%	0.10	80%

Most Important Features

Unable to determine, because the RBF kernel produced the best models

Best Model Selection

SVM

SVM	2012		2016		BOTH	
	MAE	Exp Var	MAE	Exp Var	MAE	Exp Var
2012	0.12	74%	0.62	0%	0.37	0%
2016	0.22	45%	0.09	83%	0.15	58%
BOTH	0.11	75%	0.09	84%	0.10	80%

GB

	2012		2016		BOTH	
	MAE	Exp Var	MAE	Exp Var	MAE	Exp Var
2012	0.10	78%	0.12	72%	0.11	76%
2016	0.16	53%	0.11	76%	0.14	66%
BOTH	0.14	62%	0.12	75%	0.13	70%

Model Selection

SVM produced the highest individual score

Gradient Boosting Models are the most generalizable

Conclusions

- Gradient Boosting Produced the most generalizable models
- The lowest MAE was .10, higher than desired
- The most important features
  - % Black
  - % White
  - Monthly Costs (People w/ mortgage)
  - Population
  - % over 18

Future Work

- Re-examine feature selection, since I am suspicious that the 2012 Gradient Boosting Model was more generalizable than the version trained on 2012 and 2016
- Look at PCA, or another method to group the features since there is still a lot of intercorrelation between the top features
- Test with more years, and with down-ballot initiatives