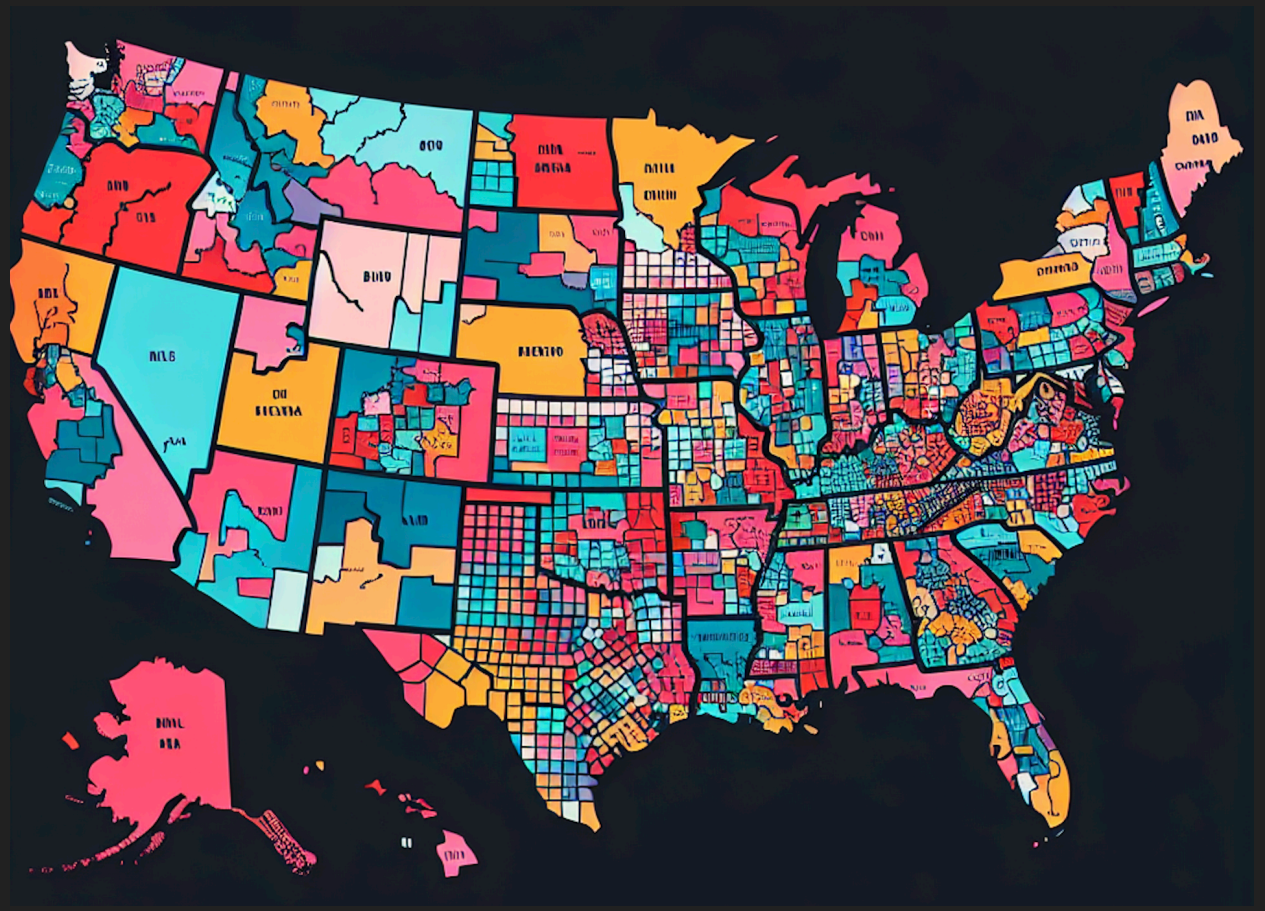
If you’re single, you’re probably a Democrat…

and other insights on US demographics and voting behavior



Arvind Suresh

Fernando Liu Lopez

Project Description

Voting behaviors depend on news and events leading up to the election; these are often unpredictable and undermine the accuracy of election forecasts.

Well-known - certain demographic characteristics are strong predictors of voting tendencies (e.g., rural areas tend to vote Republican), and have been for several years.

Given these persisting correlations, we propose the following thesis.

Thesis: Demographic data is a strong predictor of voting inclination across large timescales.

Data Collection

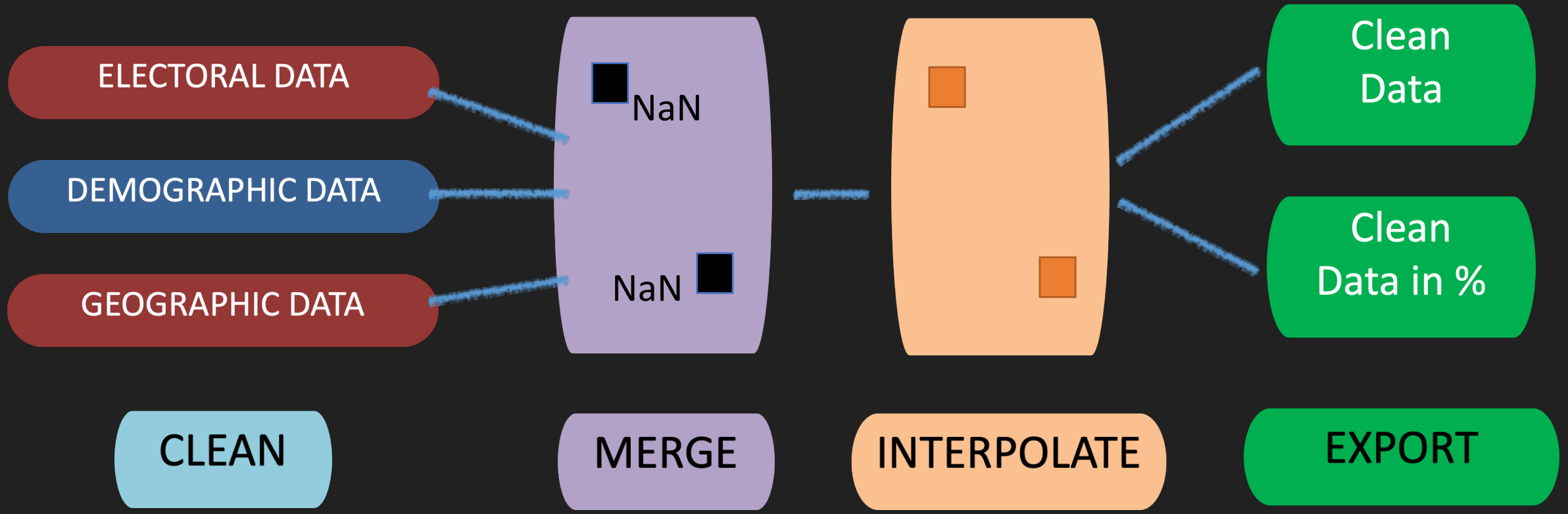
4 election years * 3104 counties

Electoral Data:

MIT ELECTION DATA + SCIENCE LAB

Demographic Data:
(original source is the ACS 5-year surveys, each centered at the respective election year)

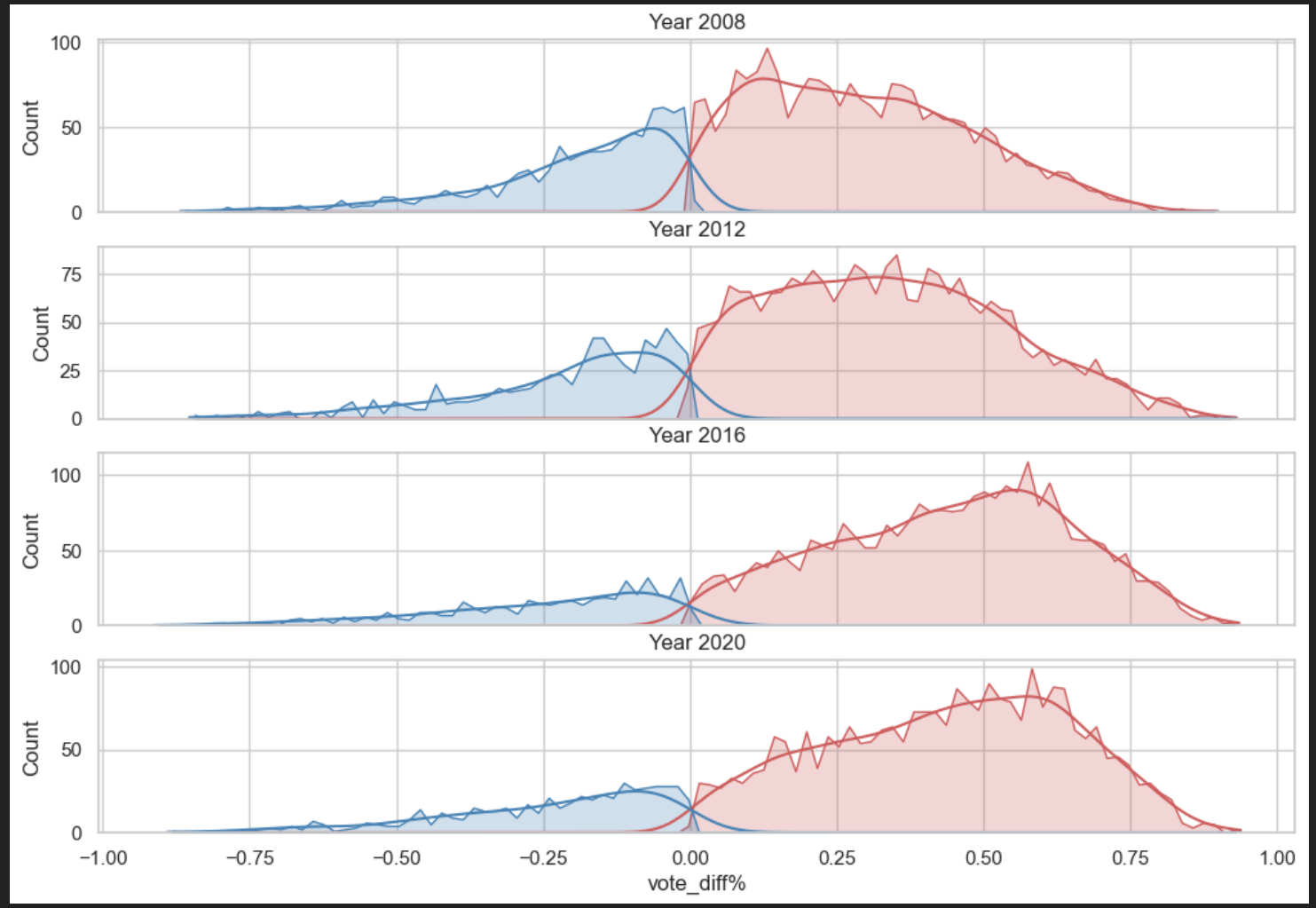
Preprocessing



Features

Income	% of people by income, labor force, etc.
Marital	% of people single/married/divorced
Education	% of people at different educational levels
Gender	% of men/women
Race	% of black/white/hispanic people
Density	Population density, household density

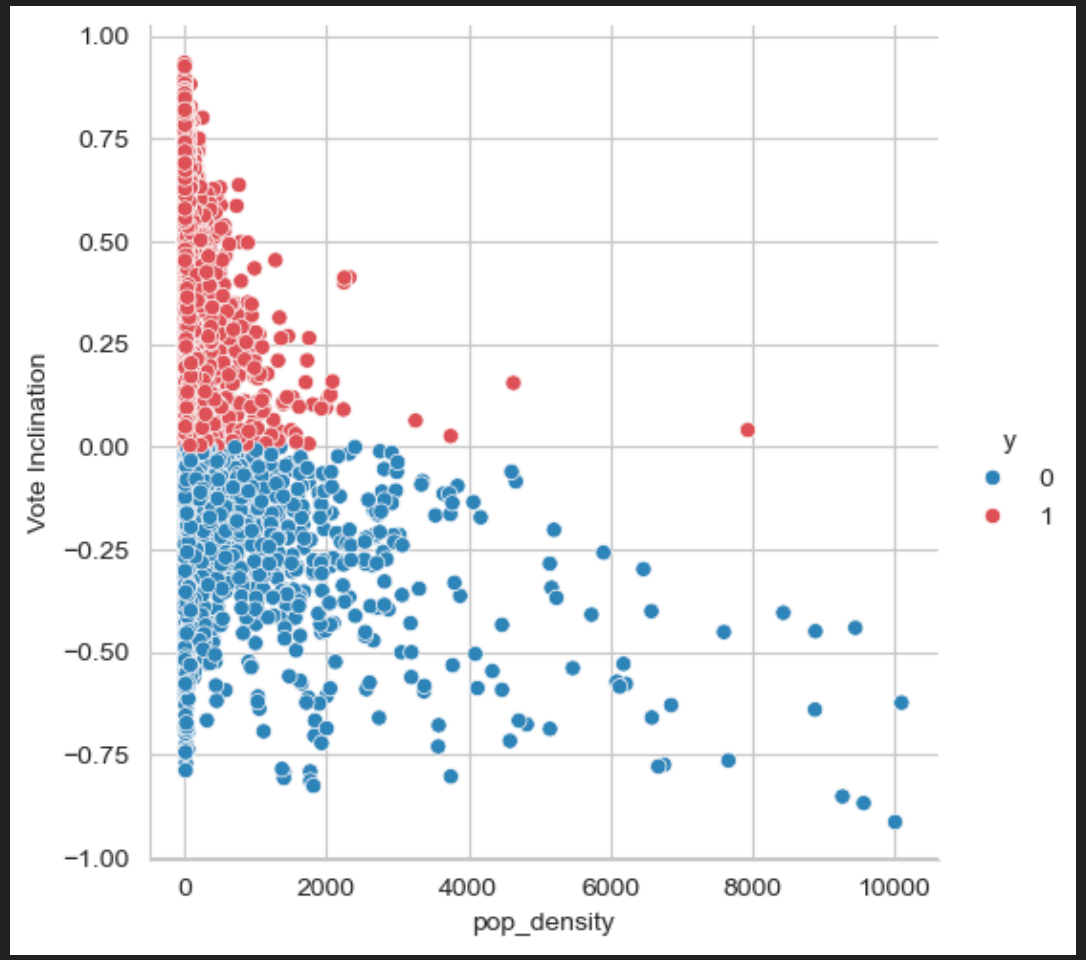
Target variable: '*voting inclination*' as the difference between the (percentage of) republican and democrat votes.



Caption

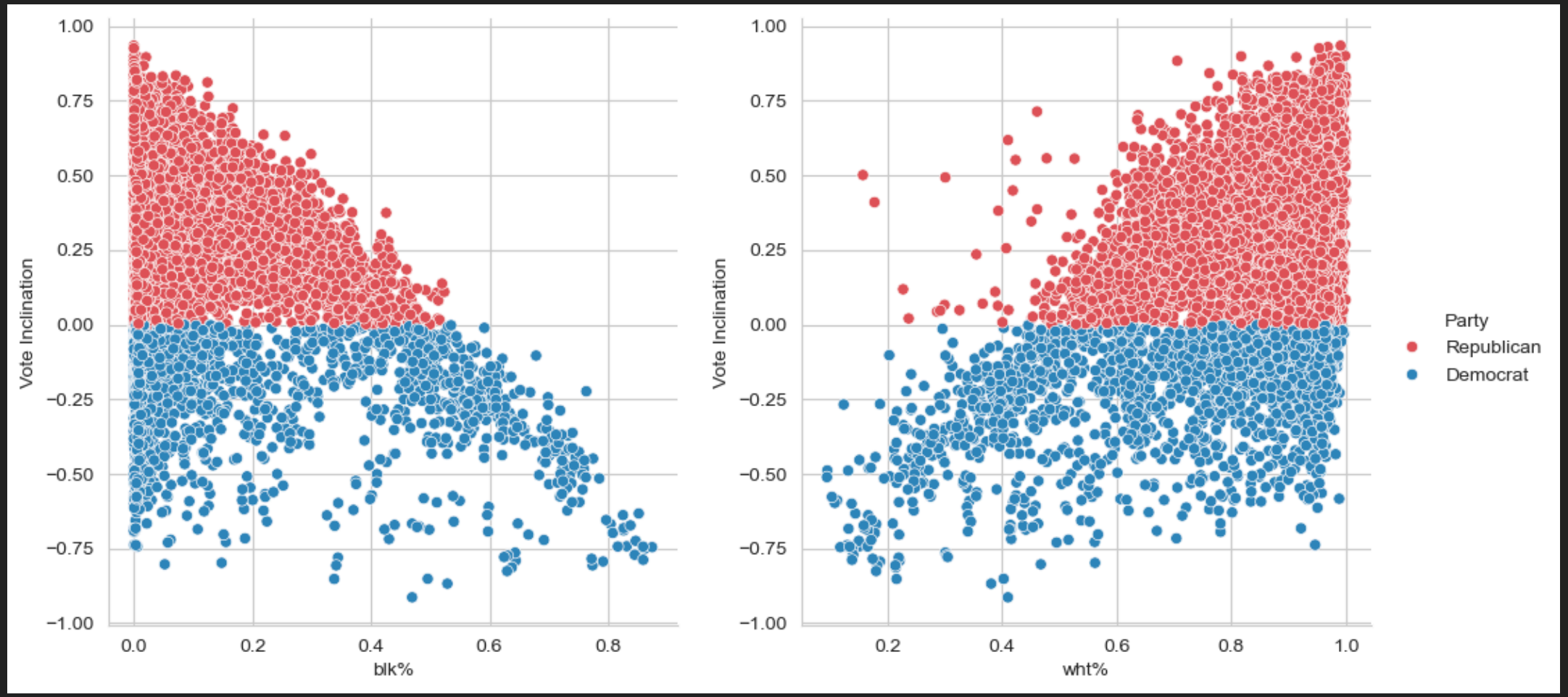
Exploratory Data Analysis

Vote Inclination



Population Density

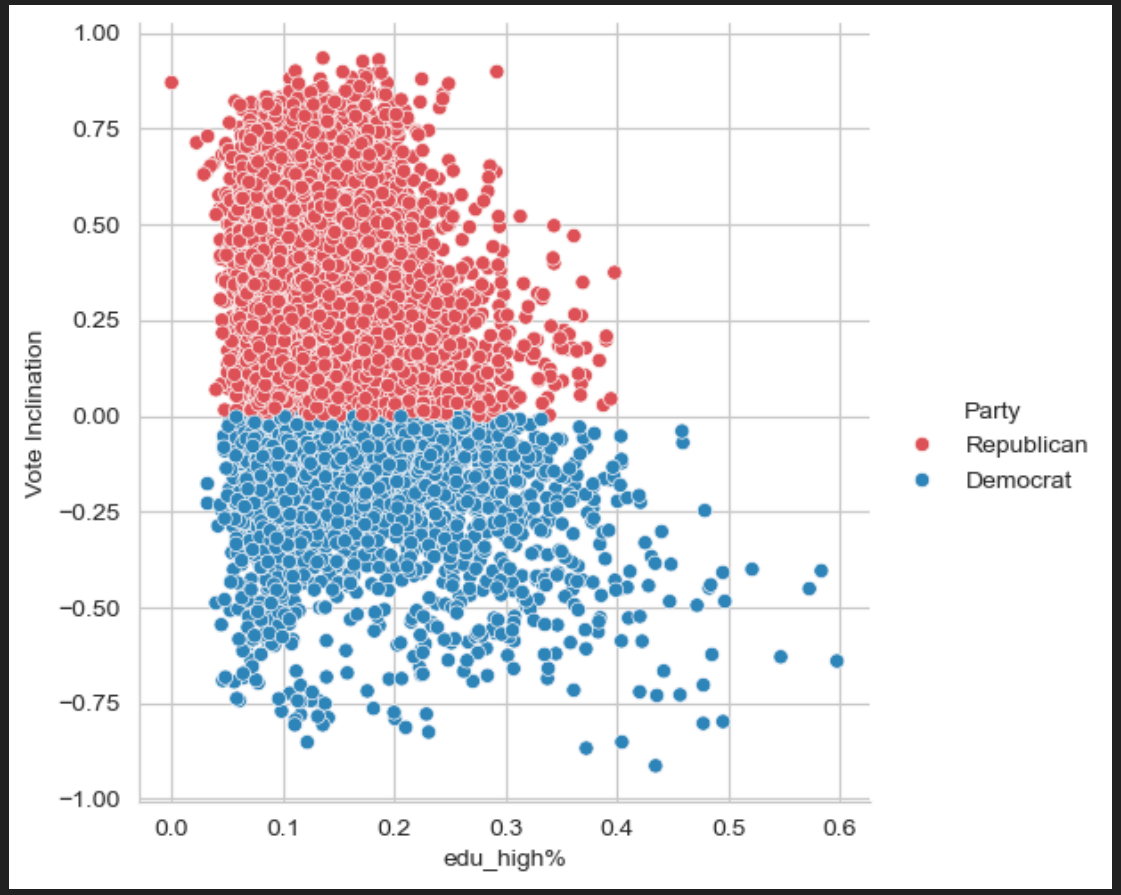
Vote Inclination



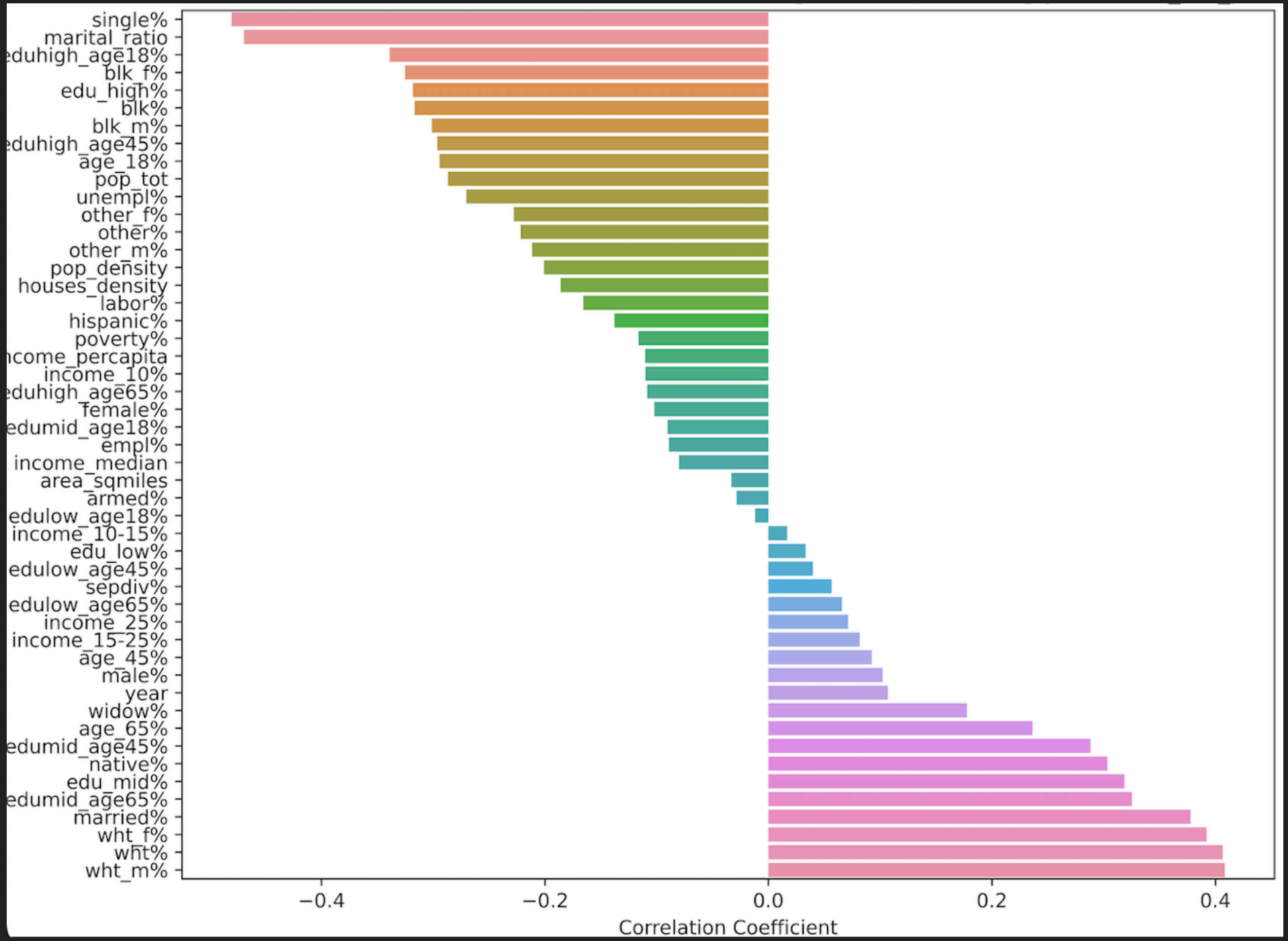
% Black Population

% White Population

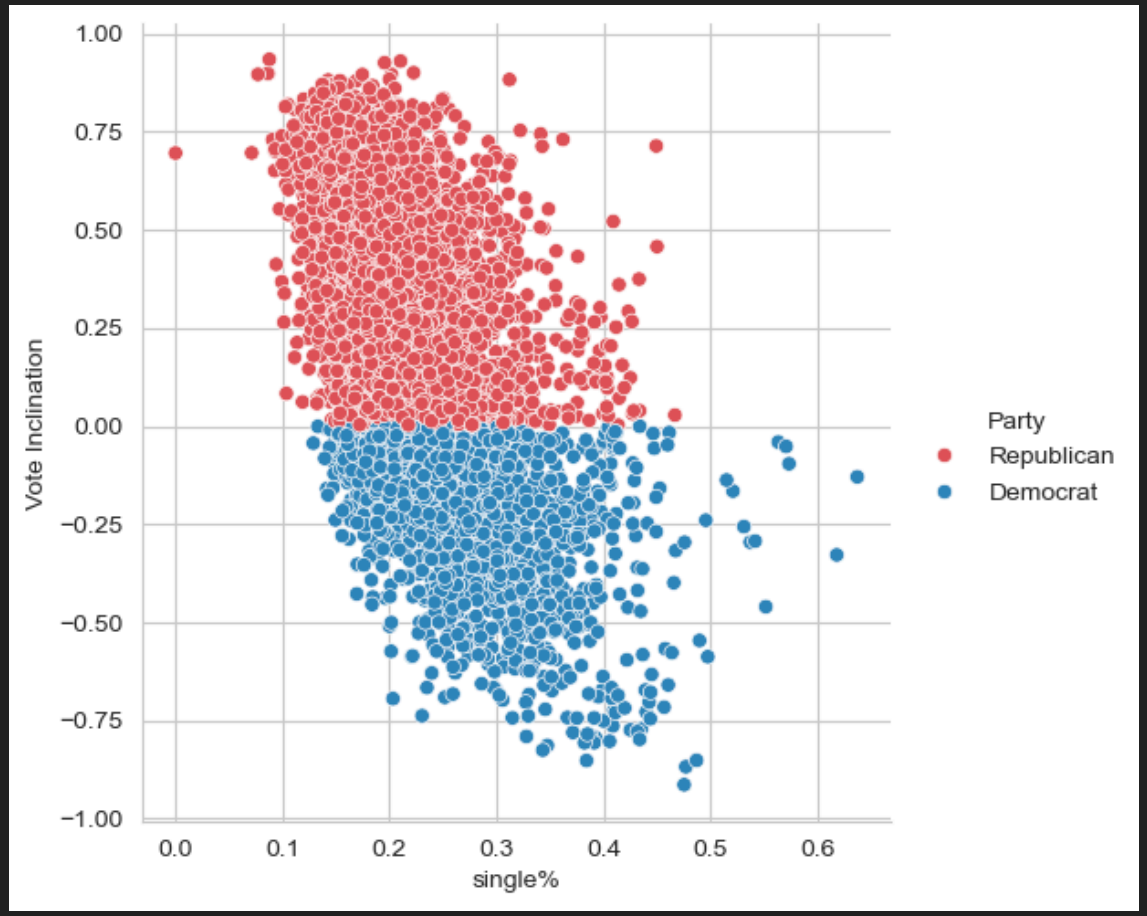
Vote Inclination



% Pop. With High Education


single%
marital_ratio
eduhigh_age18%
blk_f%
edu_high%
blk%
blk_m%
eduhigh_age45%
age_18%
pop_tot
unempl%
other_f%
other%
other_m%
pop_density
houses_density
labor%
hispanic%
poverty%
ncome_percapita
income_10%
eduhigh_age65%
female%
edumid_age18%
empl%
income_median
area_sqmiles
armed%
edulow_age18%
income_10-15%
edu_low%
edulow_age45%
sepdiv%
edulow_age65%
income_25%
income_15-25%
age_45%
male%
year
widow%
age_65%
edumid_age45%
native%
edu_mid%
edumid_age65%
married%
wht_f%
wht%
wht_m%
−0.4
−0.2
0.0
0.2
0.4
Correlation Coefficient

Vote Inclination



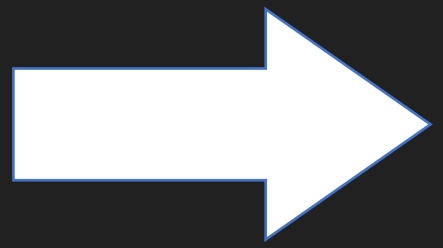
% Single

Models

Ridge Regression

K-Nearest Neighbors

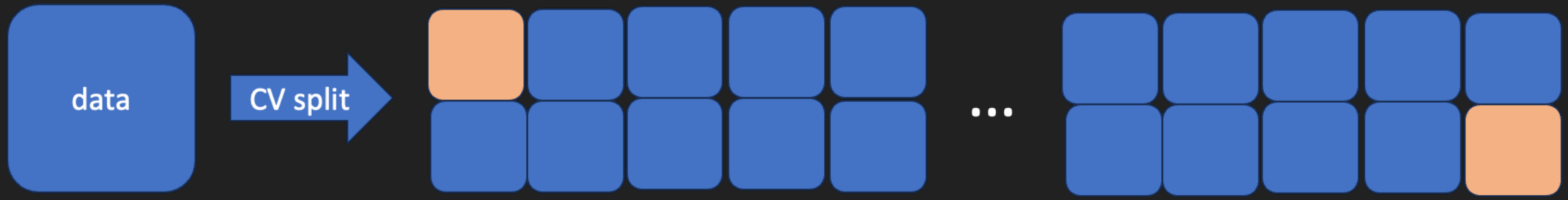
Random Forest Regression



Weighted Average

Tuning and Cross Validation

(Perform 10-fold cross validation) * 5



Total: 50 training/holdout splits

Performance Evaluation

Training Year

Accuracy on Year

	2008	2012	2016	2020
2008	0.855	0.83	0.808	0.783
2012	0.89	0.892	0.88	0.849
2016	0.904	0.939	0.942	0.93
2020	0.877	0.936	0.944	0.940

Results

- Accuracy is better when predicting forward in time.
- In almost all cases, the test accuracy was on par with (or better than) the cross validation accuracy.
- The % single population tended to be rated higher in feature importance when training on a larger sample of years.
- The % unemployed population also turned out to be an important feature in many of the models.