Voter Turnout Study

How Education Affects Voter Turnout

Astha Flynn, George Kolev, Jacinto Lemarroy, Michael Peng, Qianru Ai, Yuxuan Mei

Table of Contents

Problem Statement

Data Cleaning

EDA

Machine Learning

Findings

Challenges

The problem

Problem Statement

This project aims at shedding some light on the issue of disparity between voter turnout and education level. Our team hopes to **determine if there is any disparity in the voter turnout between households who have completed varying degrees of education.**

A few of the questions proposed:

- Does education level have any impact on voting turnout?
- What other factors, such as income disparity impact on voting, could confound education as a factor?
- Does education have a greater effect on General or Primary voter turnouts?
- Is there disparity in turnout between primary and general election for different education levels?

Data Description

Voter Data

- A voter file exists for each state
- Each file contains geographic, demographic, and household information
- Each file contains the history of voting for each registered vote

Education Data

- US Census Bureau
- County level
- Educational Level
  - By age group
  - By race and gender

Data Cleaning

Keep Columns of Interest

- Education features
  Voter turnout

Duplicates

- Dropped duplicates

Data Type Inconsistency

- String format to numeric

Missing Values

- Dropped columns with too much missing data
- Impute categorical missing values with ‘Unknown’
- Impute others with mean

Data Cleaning - continued

Merge Datasets - (lack of unique key)

Merged Voter dataset with Education Attainment dataset from US Census Bureau

- Extract county names using regular expression as primary key
- Extract 1-3 digit FIPS codes using string indexing as secondary key
- Merged on county name and FIPS code

Save to parquet

- Reusability and efficiency

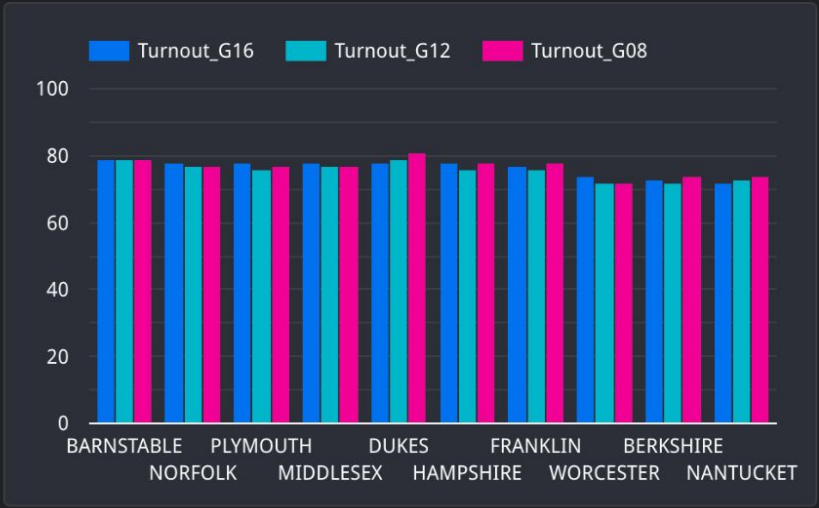
Exploratory Data Analysis

EDA Key Questions

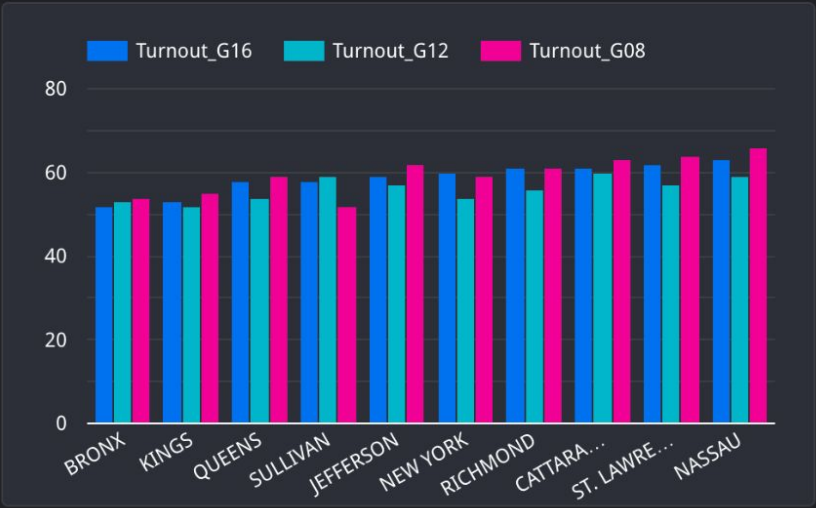
We consider 4 key questions when performing our EDA:

1. What has been the trend in voter turnout over the past few years?
2. Does education level have any impact on voting turnout?
3. Is there disparity in turnout between primary and general election for different education levels?
4. What other factors, such as income disparity impact on voting, could confound education as a factor?

Turnout is fairly consistent in the highest....



...and lowest counties by turnout.



County	Turnout_G08	Turnout_G10	Turnout_G12	Turnout_G14	Turnout_G16	Turnout_G18
HAMPSHIRE	78	57	76	54	78	70
NORFOLK	77	60	77	56	78	69
DUKES	81	63	79	52	78	68
BARNSTABLE	79	65	79	59	79	68
MIDDLESEX	77	59	77	55	78	68
COLUMBIA	69	56	68	49	70	68

County

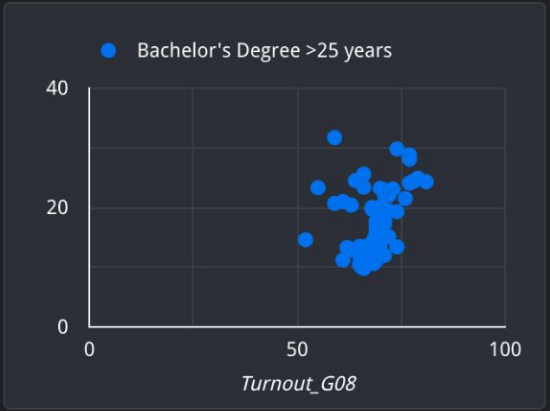
BRONX

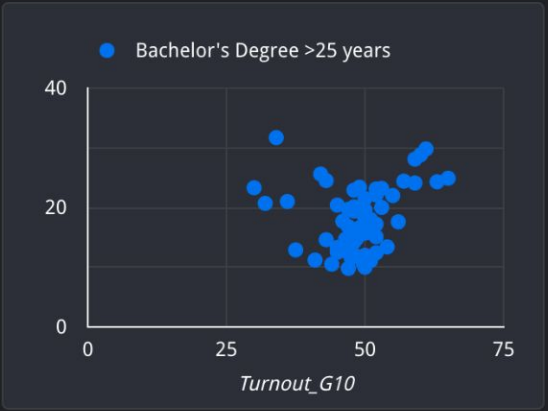
County

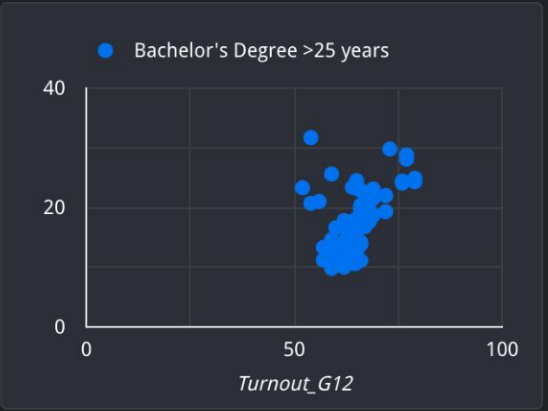
DUKES

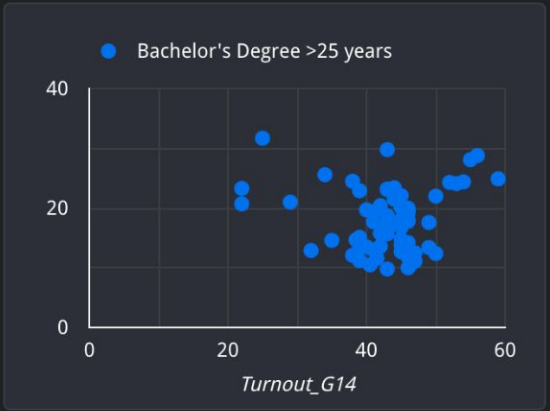
	Turnout_G08	Turnout_G12	Turnout_G16	Turnout_G08	Turnout_G12	Turnout_G16
Presidential elections	54	53	52	81	79	78
Mid-term elections	Turnout_G10	Turnout_G14	Turnout_G18	Turnout_G10	Turnout_G14	Turnout_G18
	26	20	38	63	52	68

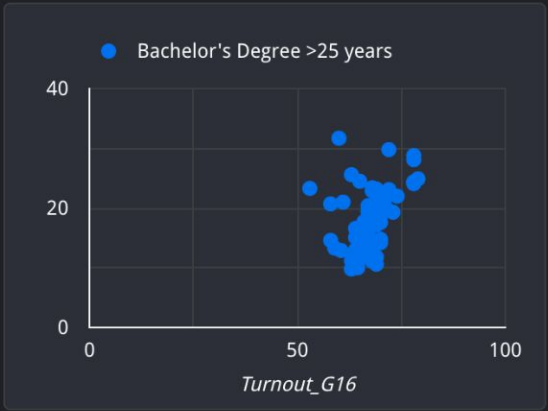
Mapping % population over the age of 25 with a Bachelor's Degree against voter turnout in 2008-2020 elections suggests a **mild positive** correlation between higher education and turnout across all election cycles.

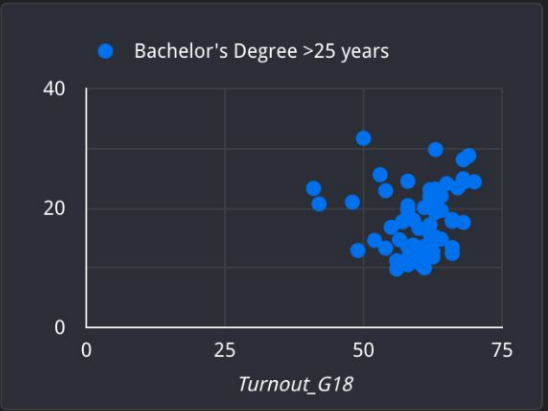




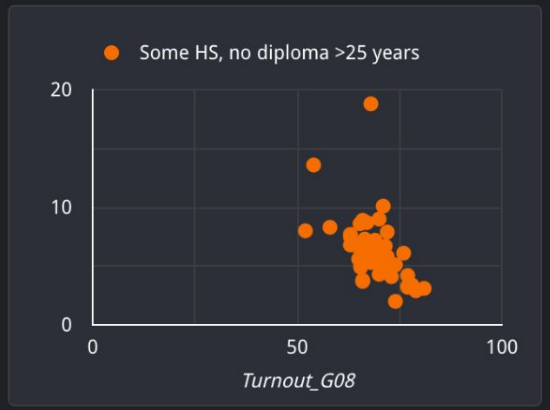


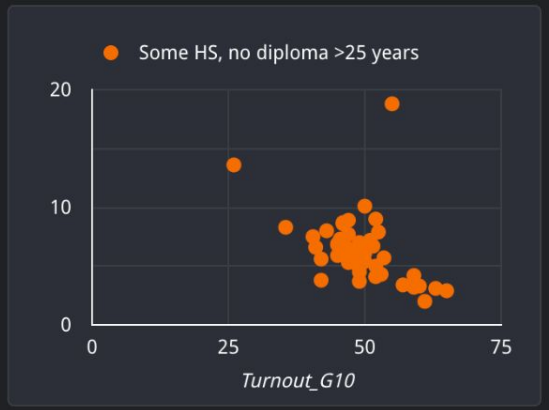


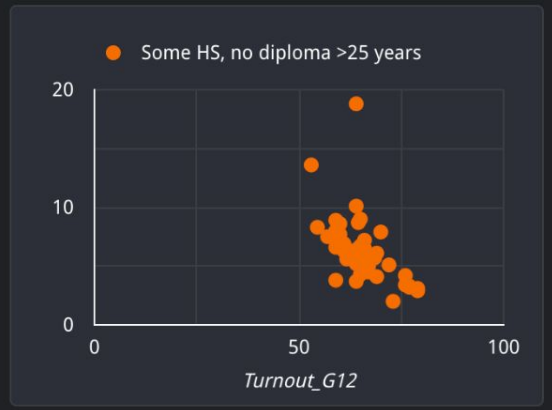


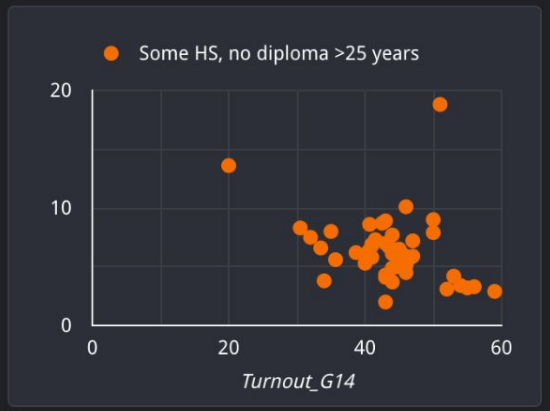


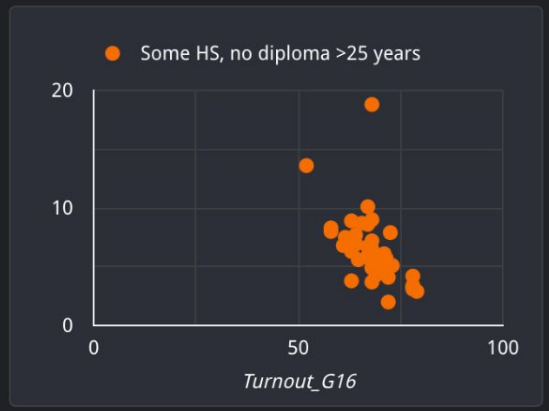
On the other hand, scatterplots of the percentage of the population (>25 years and older) without some high school education, but no diploma, against voter turnout, suggests a **negative** correlation.

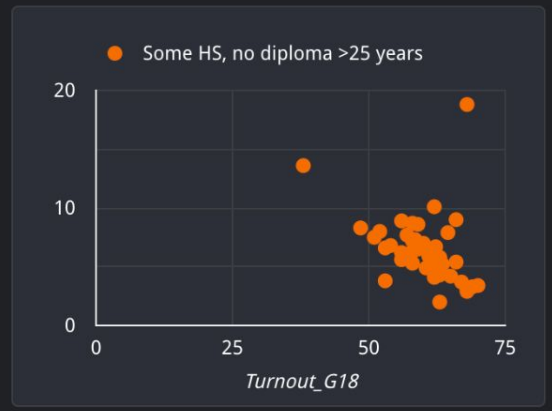




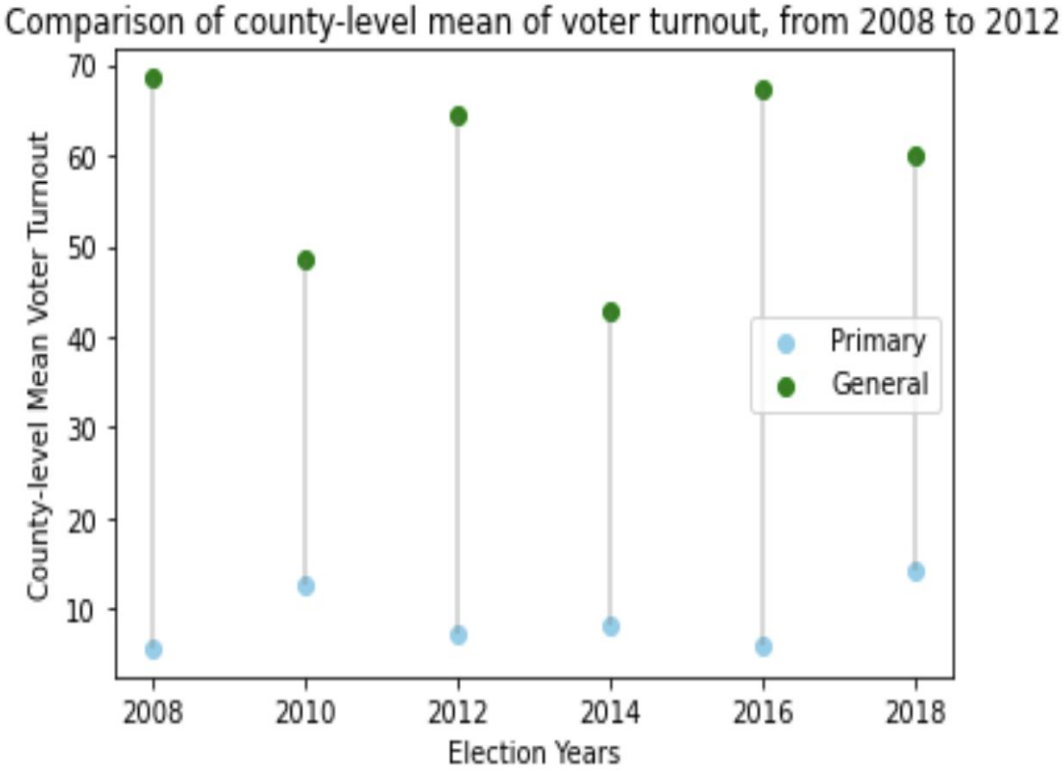






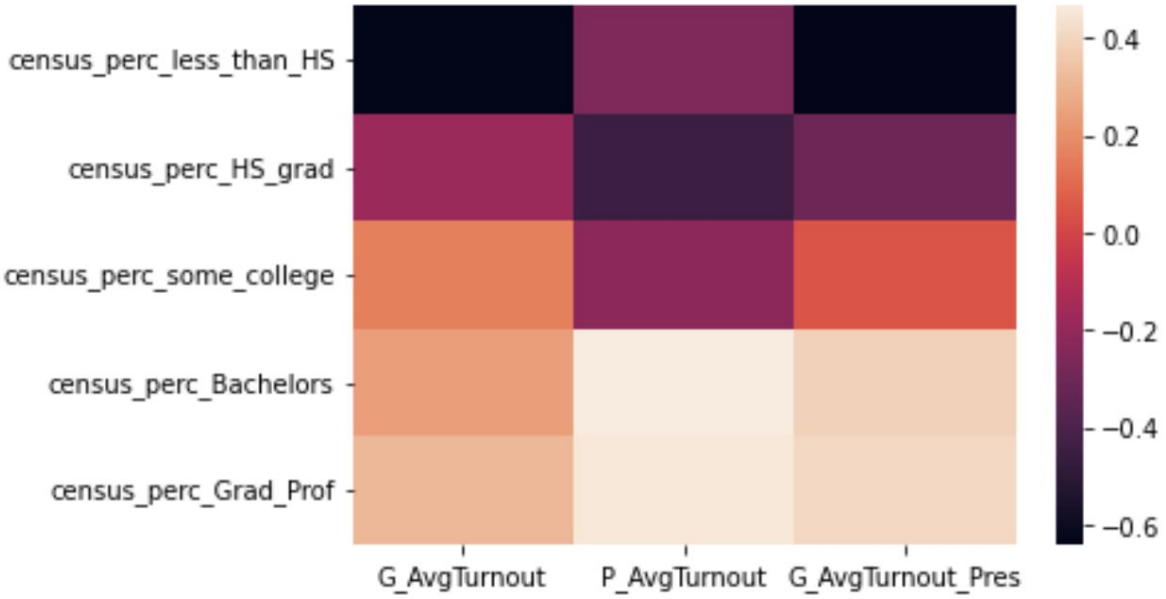


Overall Voter Turnout Trends



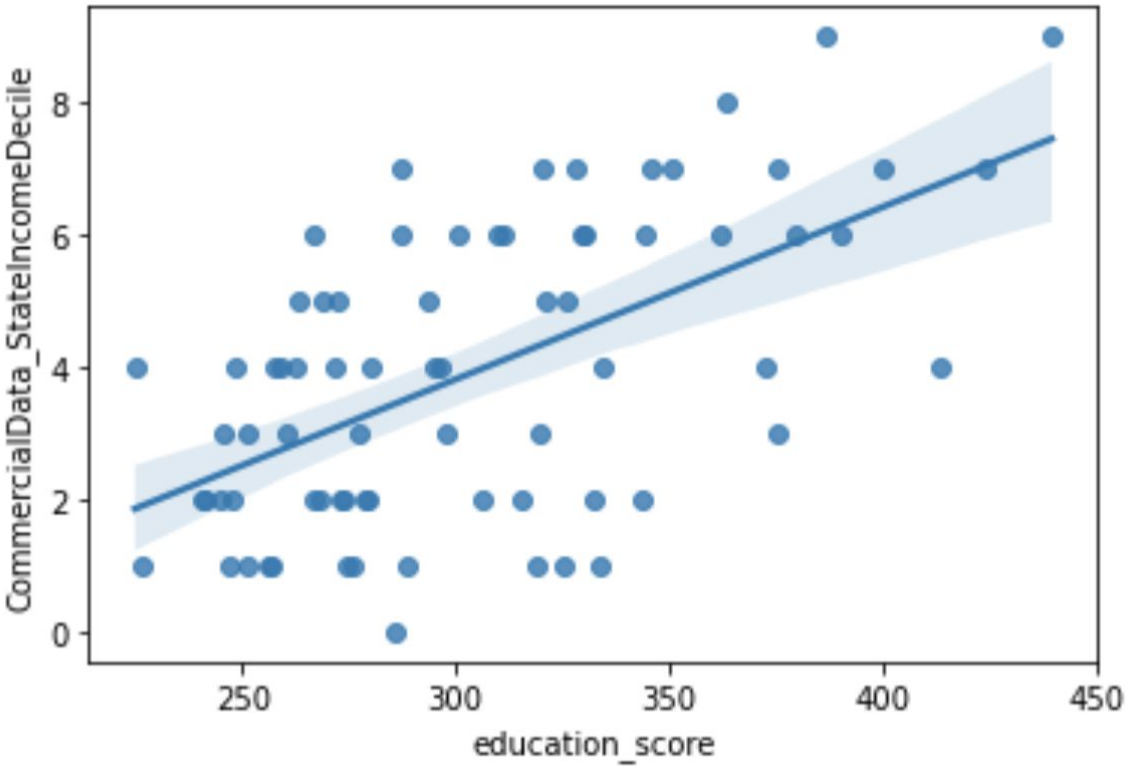
- Voter turnout is significantly higher for general elections than primary elections
- Voter turnout is highest for presidential elections
- Across election types turnout remains relatively constant except for 2018 non-presidential election

Correlation between voter turnout and Education



- Moderate to strong negative correlations for education levels of HS or less than HS and all election types.
- Moderate to strong positive correlations between education levels of Bachelors and Graduate degrees.
- Strongest positive correlations for primary elections and high education levels

Income is a Confounder



- Income and education have strong positive correlation
- Lower income groups had relatively more negative correlation with voter turnout across all election types
- We identified overlap in strong negative correlations for voter turnout between both lower income deciles and lower education levels.
- Thus, income is a potential confounder for education level when attempting to predict its effect on voter turnout.

Machine Learning

Overview

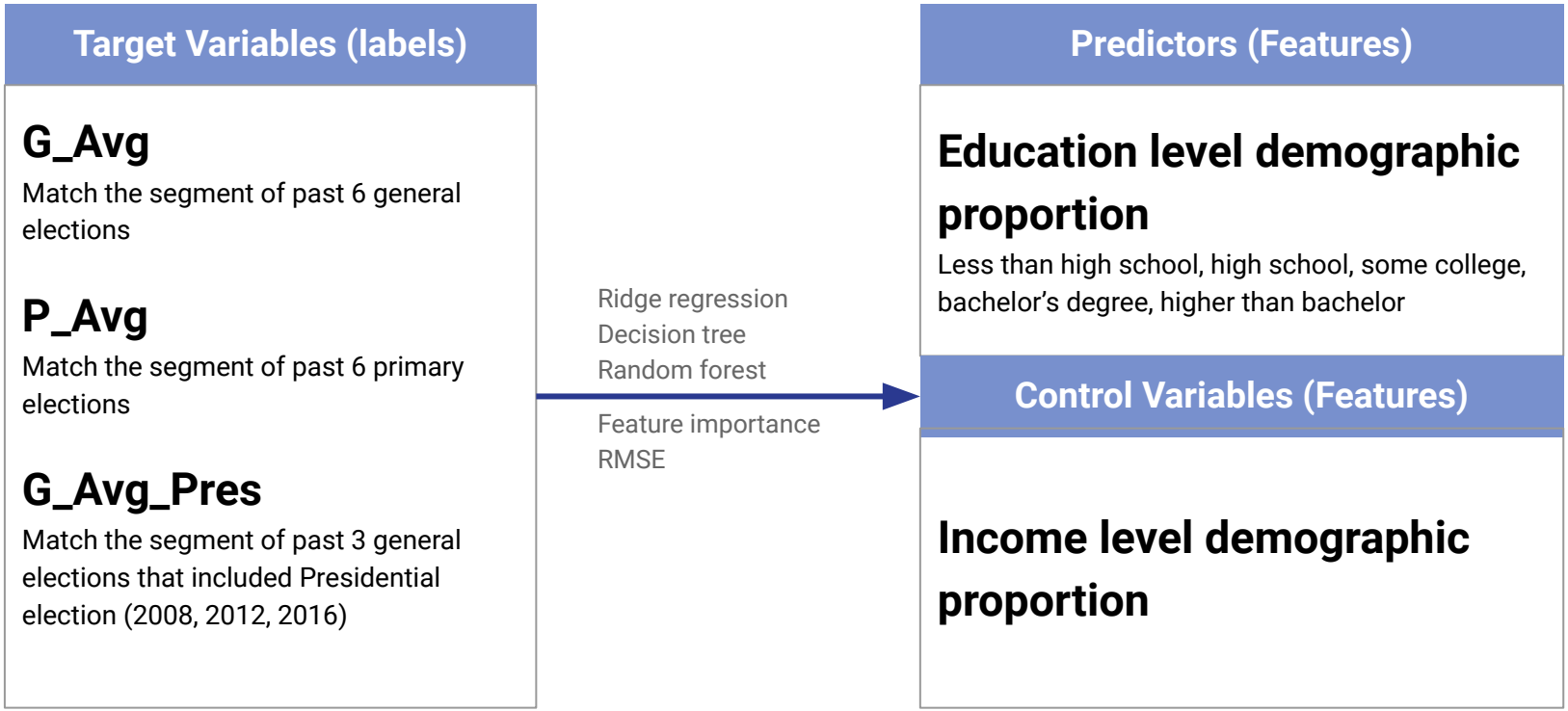
Target: to predict county voter turnout based on features such as education levels.

Motivation: identify the relationship between education and voter turnout.

Our models are focused on two questions:

1. Which variables have higher importance for a county's voter turnout?
2. Which turnout rate is more suitable for prediction by education level?

Overview


Target Variables (labels)
G_Avg
Match the segment of past 6 general elections
P_Avg
Match the segment of past 6 primary elections
G_Avg_Pres
Match the segment of past 3 general elections that included Presidential election (2008, 2012, 2016)
Ridge regression
Decision tree
Random forest
Feature importance
RMSE
Predictors (Features)
Education level demographic proportion
Less than high school, high school, some college, bachelor's degree, higher than bachelor
Control Variables (Features)
Income level demographic proportion

Feature Importance

G_avg

	name	Ridge	Decision Tree	Random Forest	Average
0	census_perc_less_than_HS	-0.688664	0.272283	0.180194	0.380380
1	census_perc_HS_grad	-0.259428	0.006672	0.068940	0.111680
2	census_perc_some_college	-0.841983	0.025636	0.030758	0.299459
3	census_perc_Bachelors	-0.063360	0.000000	0.045002	0.036121
4	census_perc_Grad_Prof	1.739241	0.070289	0.053195	0.620909

P_avg

	name	Ridge	Decision Tree	Random Forest	Average
0	census_perc_less_than_HS	-0.023576	0.569469	0.342963	0.312003
1	census_perc_HS_grad	-1.155375	0.016828	0.130252	0.434152
2	census_perc_some_college	0.017282	0.012821	0.035014	0.021706
3	census_perc_Bachelors	1.405306	0.097491	0.196929	0.566575
4	census_perc_Grad_Prof	0.033312	0.000000	0.061256	0.031523

G_avg (Pres)

	name	Ridge	Decision Tree	Random Forest	Average
0	census_perc_less_than_HS	-1.341839	0.442099	0.380701	0.721546
1	census_perc_HS_grad	-0.447931	0.043164	0.037110	0.176068
2	census_perc_some_college	-0.581186	0.072240	0.106219	0.253215
3	census_perc_Bachelors	0.877985	0.129562	0.085387	0.364312
4	census_perc_Grad_Prof	1.175311	0.007221	0.053949	0.412160

In order to encourage people participate in elections, measures should be adopted to **encourage residents to go to college**.

Model Applicability

Target variable	Avg RMSE for train	Avg RMSE for test
General turnout	1.96	6.04
Primary turnout	2.13	3.44
General turnout (precidential)	1.61	3.90

- it is not very rigorous to draw direct conclusions about which variable is more appropriate to predict with education level.
- The primary and general elections (precidential)may be more worthy of further study.

Limitations

- Limited sample size
- Based on historical data. Future trends may change.

For further exploration

Improve models:

- Larger sample size.
- More gridsearch parameters.

Increase sample size:

- Include more states in our dataset.
- Find data on the demographic distribution of education levels for each county for the past six elections and combine it with our existing voter turnout for each county for the past six elections, so that our sample size becomes 6 * the total number of counties.

Findings

Answers to our Questions

Summary

- Higher overall education levels of counties lead to higher voter turnout.
- Greater impact on primary election turnout.
- Primary election: *bachelors* and *grad,prof* have larger positive impacts, while some *college, less than high school* and *high school* have negative impacts.
- General election: *college, bachelors* and *grad,prof* have positive impacts, while *less than high school* and *high school* have negative impacts.
- All the income levels have negative impacts on general election turnouts, and positive impacts on primary election turnouts.
- Overall income level of counties has similar impacts on both general and primary election turnouts

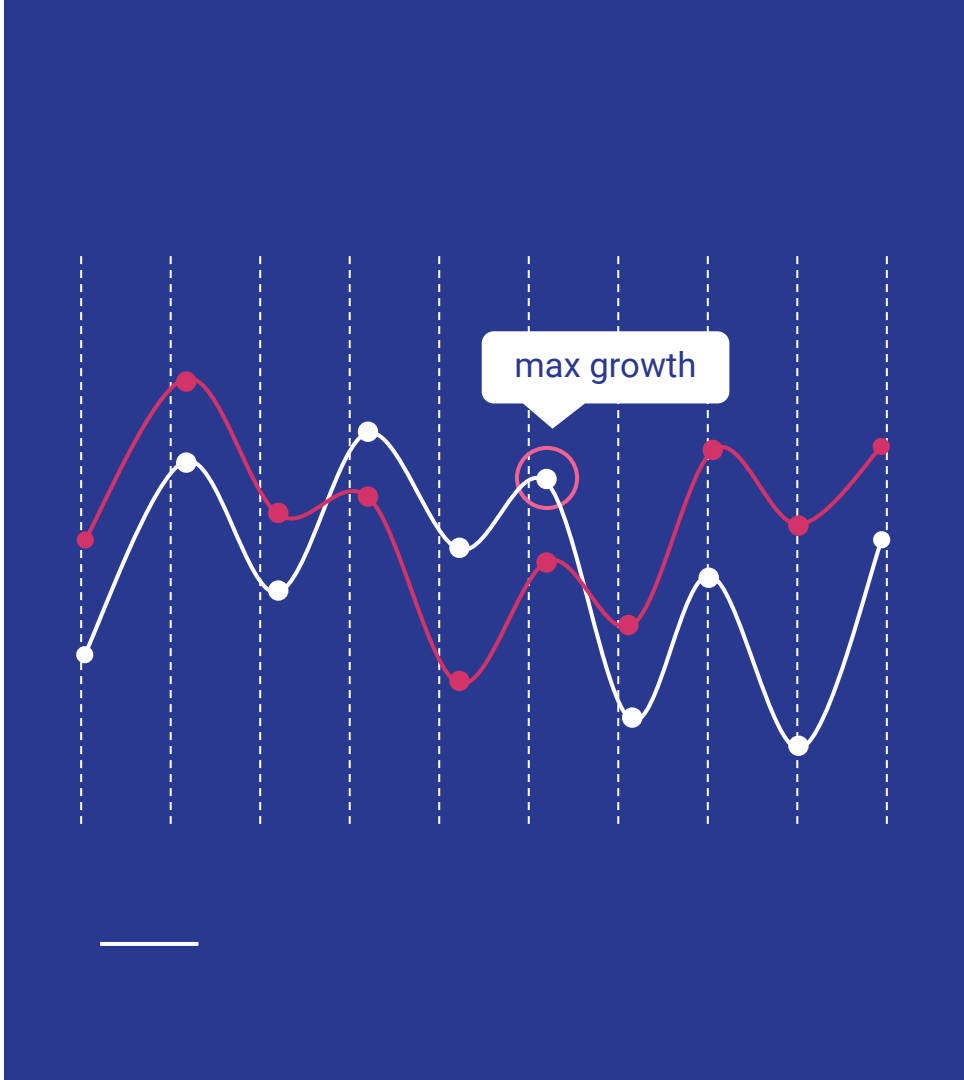
Next Steps

- Investigate on individual-level
- Combine education with other factors, like donating actions, interest in religious inspirational, if the individual is an investor, etc., to find out the approach of the education influencing the election turnout.
- Integrate additional data from recent elections, such as the 2020 general (presidential) elections
- Investigate whether voter turnout behavior and its relation to education level is dependent on state or region (e.g., West Coast, East Coast, Midwest, South)

Challenges faced

- Lack of a primary key to merge voting dataset with education dataset.
- Solution: used regular expression to extract state name information and FIPS code, then use both name and FIPS code as keys to merge.
- Develop a reasonable strategy in handling the missing values in the US Census education dataset.

Thank You



Supplemental

How education score variable was created?

$$\text{Education Score} = 0 * \text{census\_perc\_less\_than\_HS} + 1 * \text{census\_perc\_HS\_grad} + 3 * \text{census\_perc\_some\_college} + 5 * \text{census\_perc\_Bachelors} + 8 * \text{census\_perc\_Grad\_Prof}$$