




Predicting Voter Turnout

By Avi Steiner, Chase Kimball, and Davis Stagliano



Motivation: Persistent Homology

- The study done by Abigail Hickok, Benjamin Jarman, Michael Johnson, Jiajie Luo, and Mason A. Porter, published in [SIAM](#) prompted our inquiry.
- Studied Atlanta, Chicago, Jacksonville, LA, NYC, and Salt Lake City
- Weighed access by time cost to vote
- Prompted our analysis
- We ended up taking a more general approach, looking broadly at census data but including polling access within our features



Chicago, 2016

- Goal 1: Narrow down our scope.
- Chicago
 - Plurality of team members live in Chicago
- 2016 general election
 - General elections have the highest turnout, and the highest time cost
 - 2016 rather than 2020 due to Covid



Stakeholders and KPIs

Potential Stakeholders

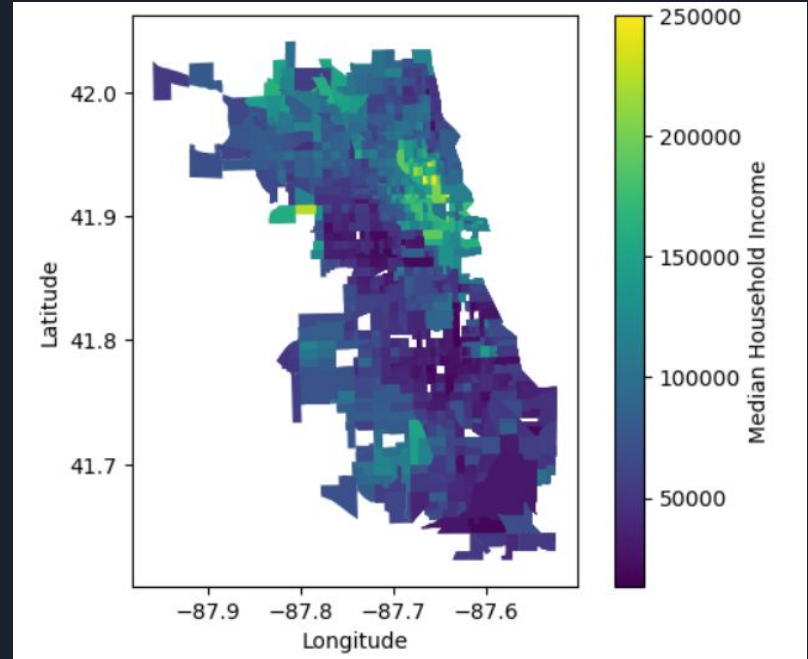
- Election Authorities and Government Agencies
- Policymakers and City officials
- Civil Rights and Advocacy Organizations
- Academic and Research Institutions

Key Performance Indicators (KPIs):

- **Root mean-squared error** for predicted average voter turnout per precinct.
- **F-scores** to identify variables useful in predicting voter turnout
- **Geographic Distribution of Voter Turnout:** Visualization of turnout rates relative to polling site coverage.
- **Polling Site Access:** Measured as average travel time or distance to the nearest polling site.

2010 Census

- Used census API.
- Chosen feature types:
 - Race
 - Transit type
 - Education
 - Employment
 - Income
 - Age





Census data Conversion and Cleaning

- We needed to turn the by census tract data into by precinct data.
- Major Assumption 1) Within a census tract, our populations are evenly distributed.
 - Census tracts are chosen to be extremely small and geographically similar.
 - Necessity. No other reasonable way of distributing the data in a tract
 - Turned problem into one of percents and simple stats.
- Results:
 - By census tract data converted to by precinct
 - Two precincts whose data was unavailable, due to a lack of data in the census.
 - Without a clear way to handle them, we dropped these precincts



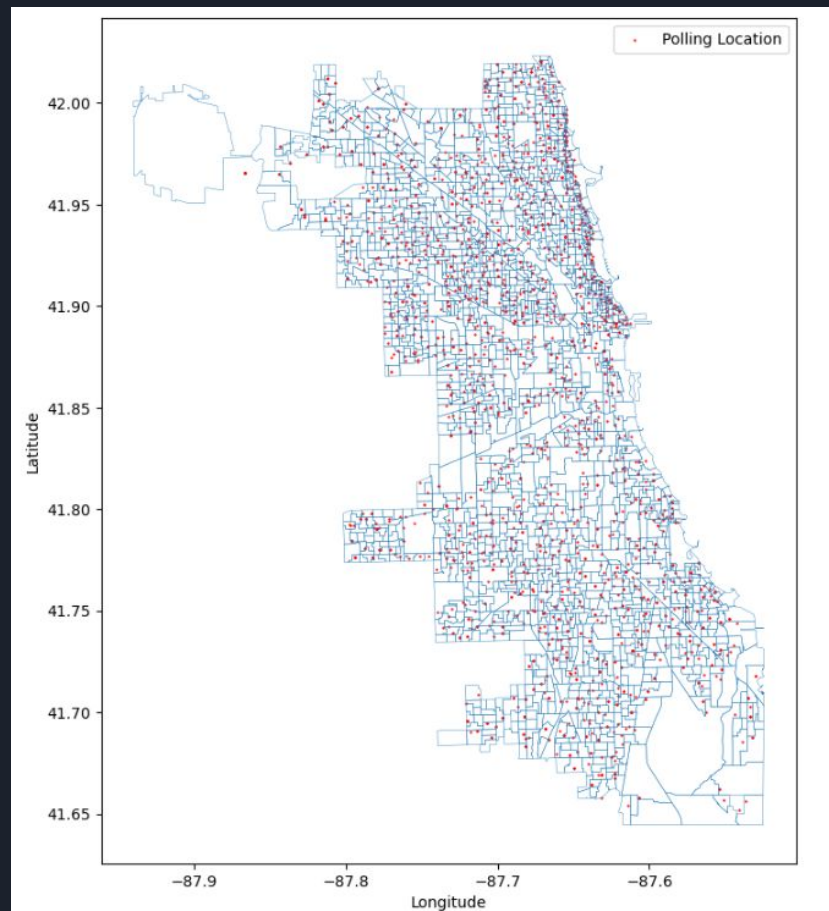
Voter Turnout

- Goal 3) Acquire voter turnout data.
- Voter turnout data for 2016 in Chicago was readily available.
- The Illinois State Board of Elections makes getting that simple.
- Data required minimal cleaning; removing of unnecessary information
 - Restricted to only precincts in Chicago, not the entire state,
 - Dropped columns which were not of use, such as candidate names.
- Some Anomalies:
 - Precinct with only 8% turnout
 - Another with 110%.
 - Able to confirm both from other sources
 - Illinois law allowing for same day registration accounts for over 100% turnout.

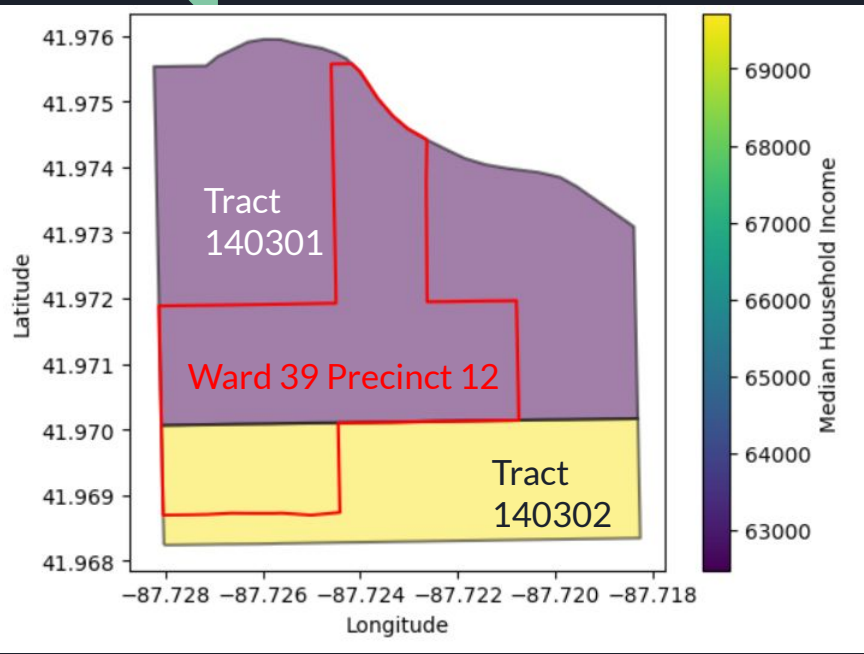
Voting precincts, polling locations, voter turnout

- Pulled geographic data for all 2069 precincts in the City of Chicago under 2012-2022 districting¹
- 1441 polling locations + precinct assignments for the 2016 general election from the Center for Public Integrity²
- Obtained travel times from precinct centers to assigned polling location via Google Maps API for walking, transit, and driving directions
- Precinct-wise voting turnout for 2016 election from the Illinois State Board of Elections³

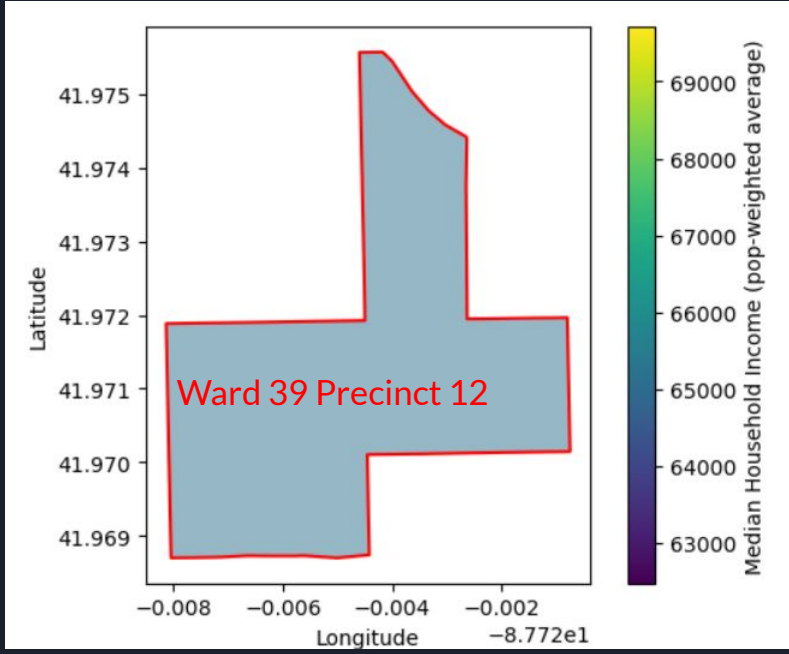
1. <https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Ward-Precincts-2012-2022-luvpq-qeeq>
2. <https://github.com/Public/US-polling-places>
3. <https://www.elections.il.gov/electionoperations/ElectionVoteTotalsPrecinct.aspx?ID=bt7bri46n7l%3d>



Converting Census Tract-wise data to Precinct-wise data



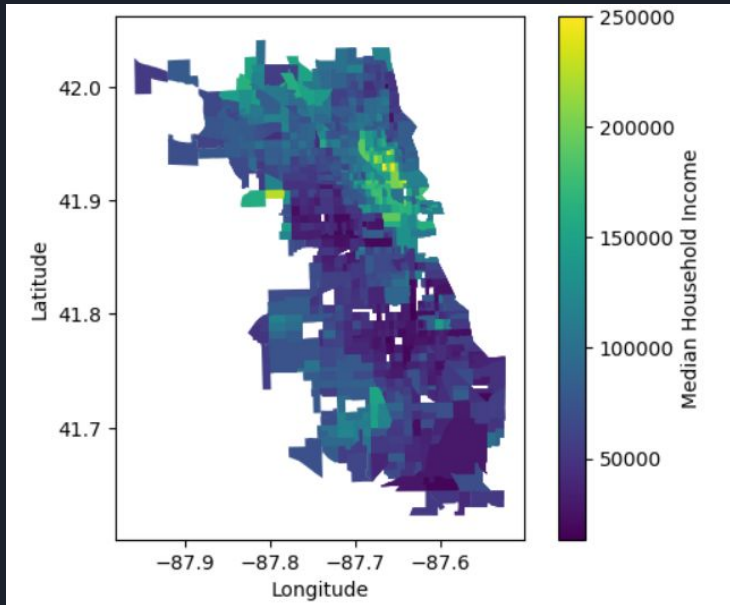
Census Tracts



Precincts

Assume constant density in Census tracts, then take average weighted by population in intersection

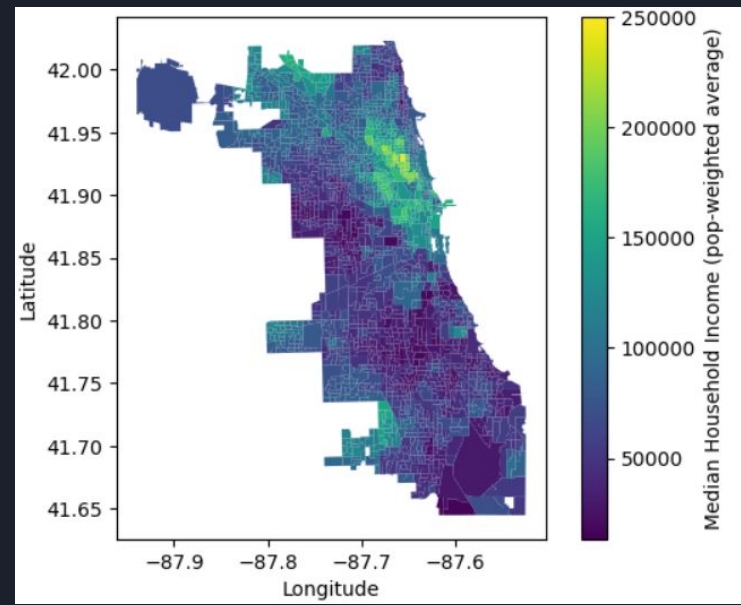
Converting Census Tract-wise data to Precinct-wise data



Census Tracts



Assume constant density in Census tracts, then take average weighted by population in intersection



Precincts



Analysis

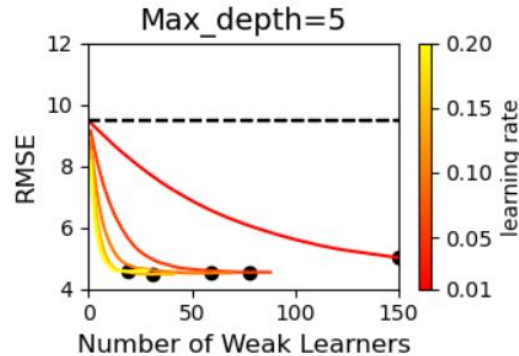
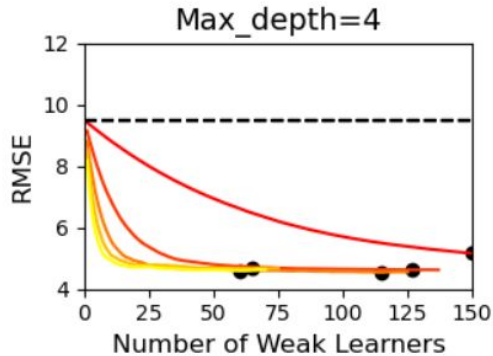
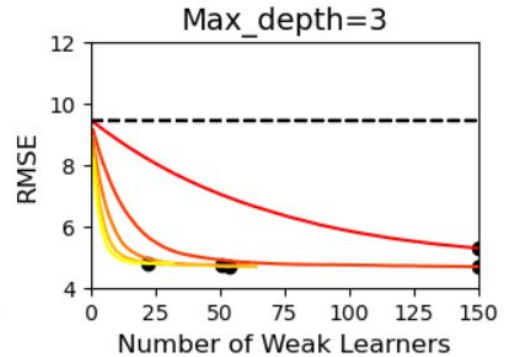
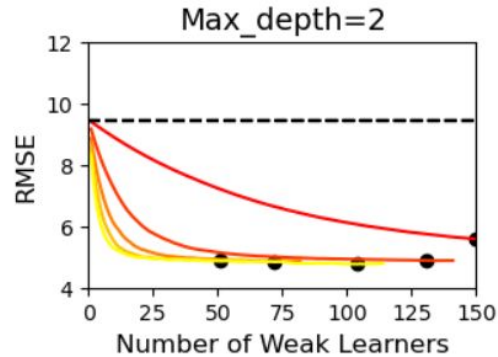
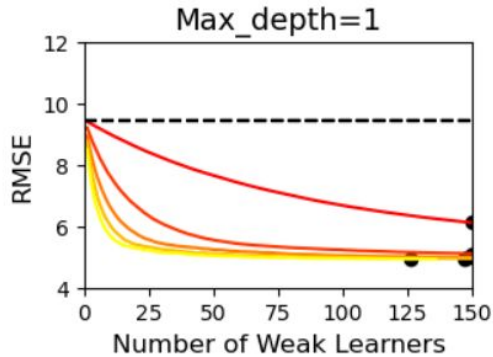
- Goal 4: Analysis
- Target variable: Voter turnout percentage
- Metric: Root mean-squared error
- 31 features related to:
 - Level of education
 - Employment
 - Mode of transit to work
 - Income
 - Race
 - Transit times to polling locations
 - Polling location accessibility
- Performed an 80/20 train test split on our data
- Further 80/20 split of training data into training and validation sets
- Started working on our models:
 - Baseline model
 - Linear regression
 - XGBoost
 - Logistic Regression



Baseline model + Linear Regression

- **Baseline Model**
 - Average voter turnout from all precincts in training data
 - RMSE 9.5% on validation set
- **Multilinear Regression**
 - Using all features, then in subsets (more on feature importance later)
 - RMSE 5.1% on validation set

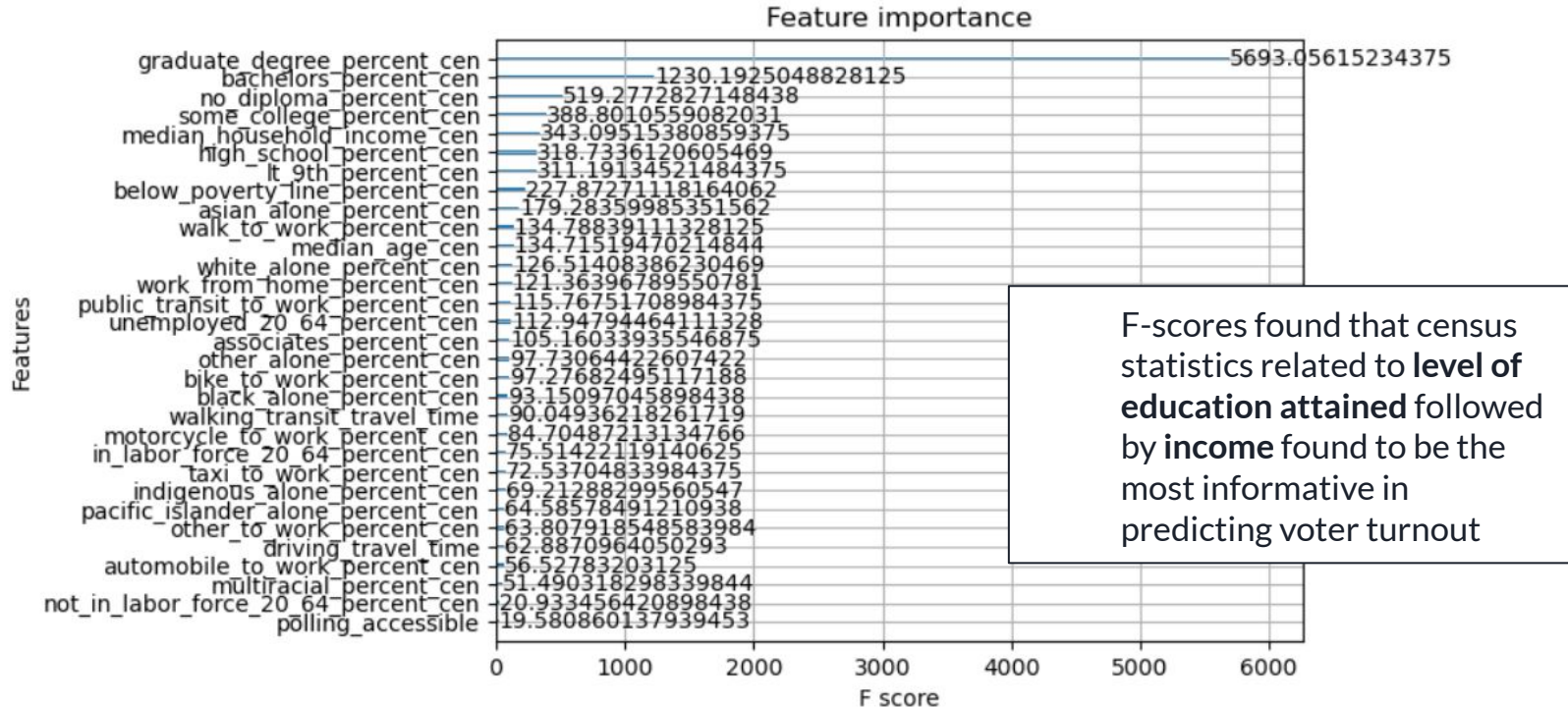
XGBoost and Feature importance



Optimal model:

- max depth = 5 ,
- Learning rate= 0.15
- RMSE of 4.47% against validation set

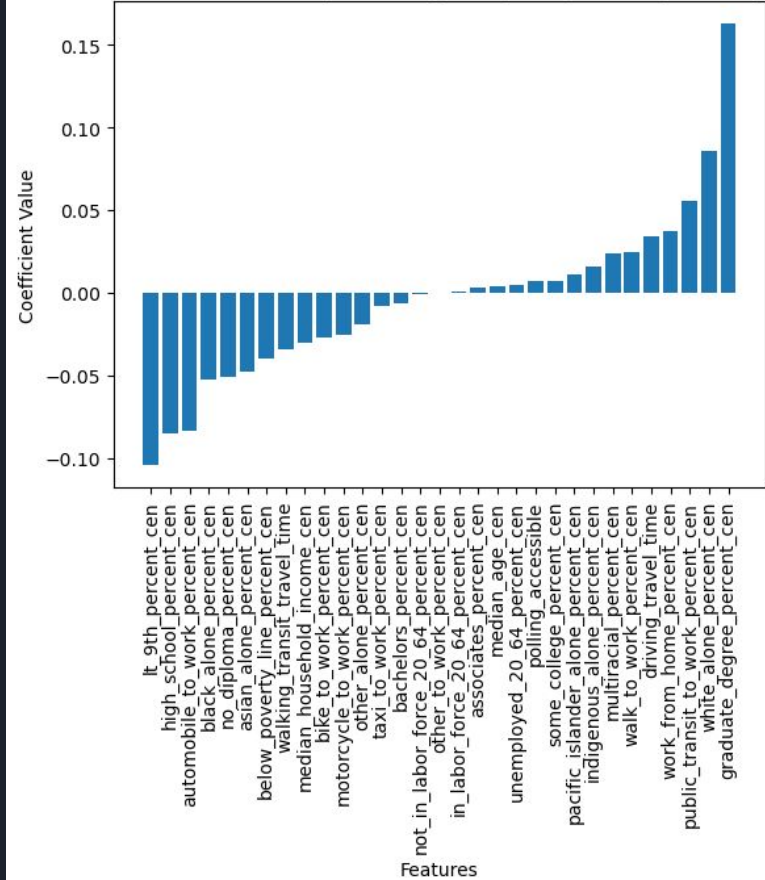
XGBoost and Feature importance



Weighted Logistic Regression

- Performs same as linear.
- Philosophically better.
- Confirms feature importance.
- Also used a “categorical” logistic regression.
 - $\geq 70\%$ vs. $< 70\%$

Feature Coefficients from (Non-Categorical) Weighted Logistic Regression





Conclusions

- Linear and Logistic models performed equally well
 - 45% decrease in RMSE over baseline
 - The logistic model is a better overall fit due to the nature of the data
- XGBoost was almost as good
 - 42% decrease
- Most important features were related to educational attainment
 - Followed by income

Future studies can look into the 2024 election, which occurred after the redistricting process reduced the number of polling locations and precincts.



Limitations, Notes, Problems

- The Covid pandemic has caused lasting changes to how people vote, with more people opting for voting by mail than ever before. As such, the impact of wait times at polling locations has been greatly decreased.
- The census data for a number of tracts for some data fields was unfortunately NaN. We dealt with this by treating NaN as 0. Leaving it as NaN compounded the problem, as in Python, $\text{NaN} + \text{float} = \text{NaN}$, so dozens of precincts ended up with NaN data. Treating it as 0 left us with only two precincts with bad data, which we dropped.
- The persistent homology study builds a model for resource allocation, and finds gaps in that allocation. While it is a fascinating topic, and there is a lot more depth to be explored, we were only interested in some of their topline data, namely the time it takes to vote.



Citations

- Persistent Homology Study: <https://epubs.siam.org/doi/abs/10.1137/22M150410X?journalCode=siread>
- 2016 Voter Turnout: <https://www.elections.il.gov/electionoperations/ElectionVoteTotalsPrecinct.aspx?ID=bt7bri46n7l%3d>
- List/Explanation of ACS codes: <https://api.census.gov/data/2010/acs/acs5/variables.html>
- Additional validation of ACS codes: <https://data.census.gov/table>