Ashlin John
CS 579
Dr. Hood
12.04.2024

Final Project Report

Introduction

The district chosen for this project is Illinois District 03 (IL_D3). Currently, Delia Ramirez is the U.S. House member representing IL_D3. Historically, looking into elected members representing IL_D3, the winning political party has been and still is Democratic[5]. In the upcoming 2024 election for the U.S. House, incumbent Delia Ramirez is running for the Democratic party and John Booras is running for the Republican party[4]. In this report, the winning political party and voter turnover will be predicted.

Project Approach and Data Sourced Explored

For this project, I used a combination of spatial data and demographic datasets to analyze voter turnout trends in IL_D3. The primary datasets included American Community Survey (ACS) data for demographic and socioeconomic variables, election turnout data from Ballotpedia, and the Redistricting Data Hub supplied spatial shapefiles at the precinct, district, and block group levels. These datasets were merged using shared identifiers, such as GEOIDs, to ensure compatibility and consistency. Spatial operations, including clipping and filtering, were performed to focus the analysis exclusively on District 3's boundaries. All data preprocessing and analysis were conducted in RStudio and Excel.

The decision to use these datasets was driven by their relevance to voter turnout analysis. ACS data provided critical insights into population characteristics, such as education and income, while Ballotpedia data offered a historical perspective on turnout trends. The spatial

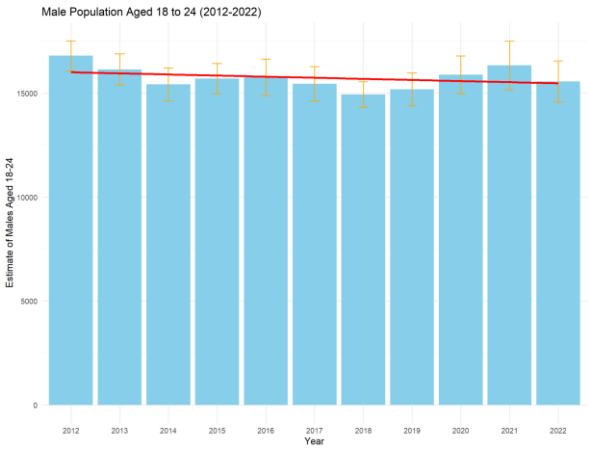
shapefiles allowed for precise mapping and integration of demographic and election data within the district.

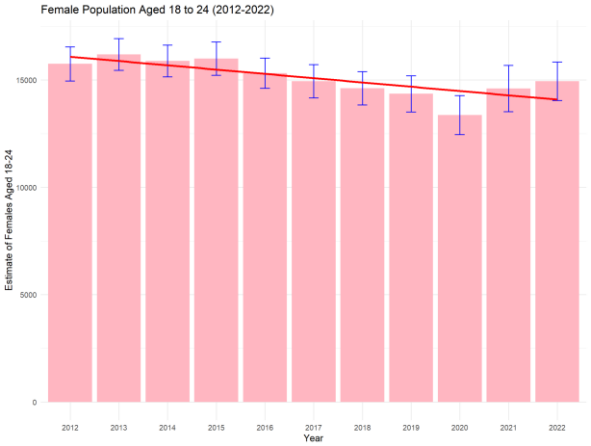
One significant challenge was ensuring the accuracy of historical data for IL_D3, as the district has undergone multiple instances of redistricting. Additionally, Senate-level election data was missing, which limited the scope of certain analyses. These challenges were addressed through preprocessing steps, including validation, and data cleaning, ensuring the data used was as accurate and relevant as possible.

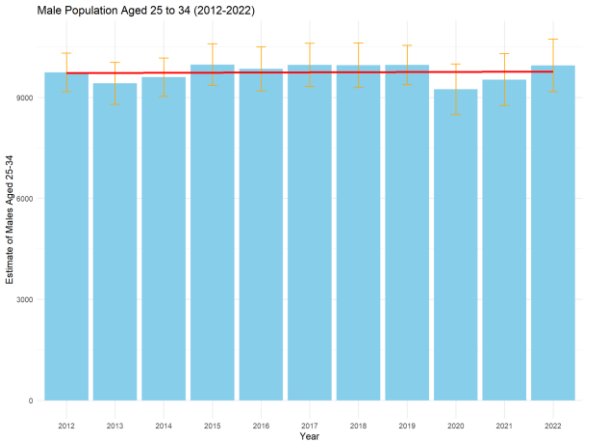
To process this data, I ensured all datasets were aligned to a common Coordinate Reference System (CRS) for accurate spatial operations. The spatial shapefiles were clipped to IL_D3's boundaries to focus the analysis exclusively on the region of interest. Key attributes from the ACS, such as median income and education levels, were linked to block groups using GEOIDs.

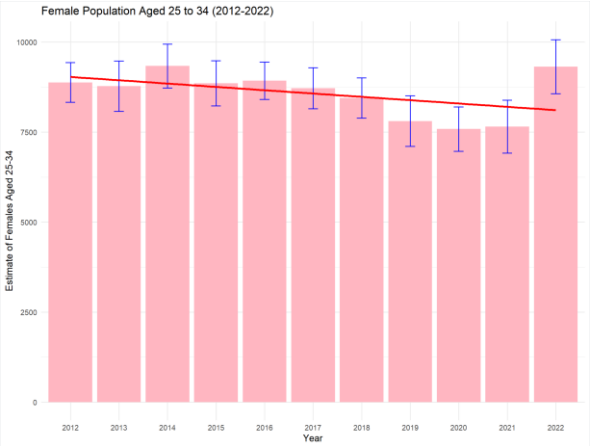
Exploratory Data Analysis

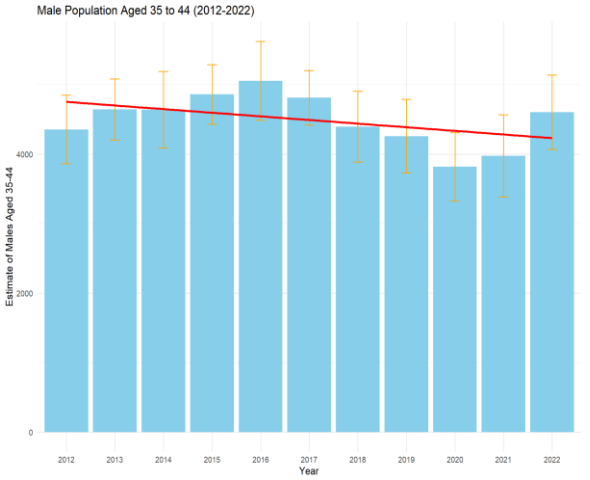
It is important to look at age and gender distribution when it comes to election results because different political groups have different views when it comes to social issues. Moreover, gender and generational differences can cause individuals to have different opinions as well, which can directly reflect their voting behavior. What one age bracket or gender might find appealing in terms of election policies may not be received well in the other age brackets or gender.


Male Population Aged 18 to 24 (2012-2022)
Estimate of Males Aged 18-24
0
5000
10000
15000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year


Female Population Aged 18 to 24 (2012-2022)
Estimate of Females Aged 18-24
0
5000
10000
15000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year


Male Population Aged 25 to 34 (2012-2022)
Estimate of Males Aged 25-34
0
3000
6000
9000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year


Female Population Aged 25 to 34 (2012-2022)
Estimate of Females Aged 25-34
0
2500
5000
7500
10000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year


Male Population Aged 35 to 44 (2012-2022)
Estimate of Males Aged 35-44
0
2000
4000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year

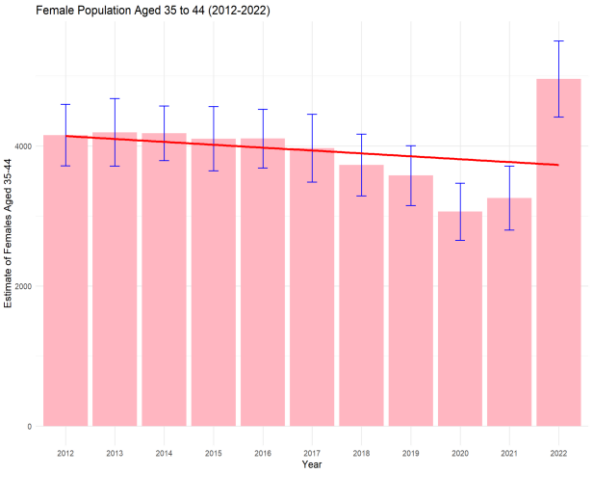

Female Population Aged 35 to 44 (2012-2022)
Estimate of Females Aged 35-44
0
2000
4000
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year

Image 1. Age and Gender Distribution

Economic distribution of wealth is also a strong indicator of voting behavior. According to *Partisanship by family income, home ownership, union membership and veteran status* by Pew Research Center (2024)[3], it was found that population within the lowest few income

brackets and the highest few income brackets lean towards the Democratic Party where as the middle income brackets lean towards the Republican Party. Looking at Image 2, adding up the individuals withing the lower and higher income brackets make up majority of the population when compared to the middle-income brackets. This directly reflects on the voting patterns and election wins over the last few years within IL_D3.

The data for income distribution was derived by calculating the average population that fell under each income bracket. This analysis and data visualization was done using RStudio.

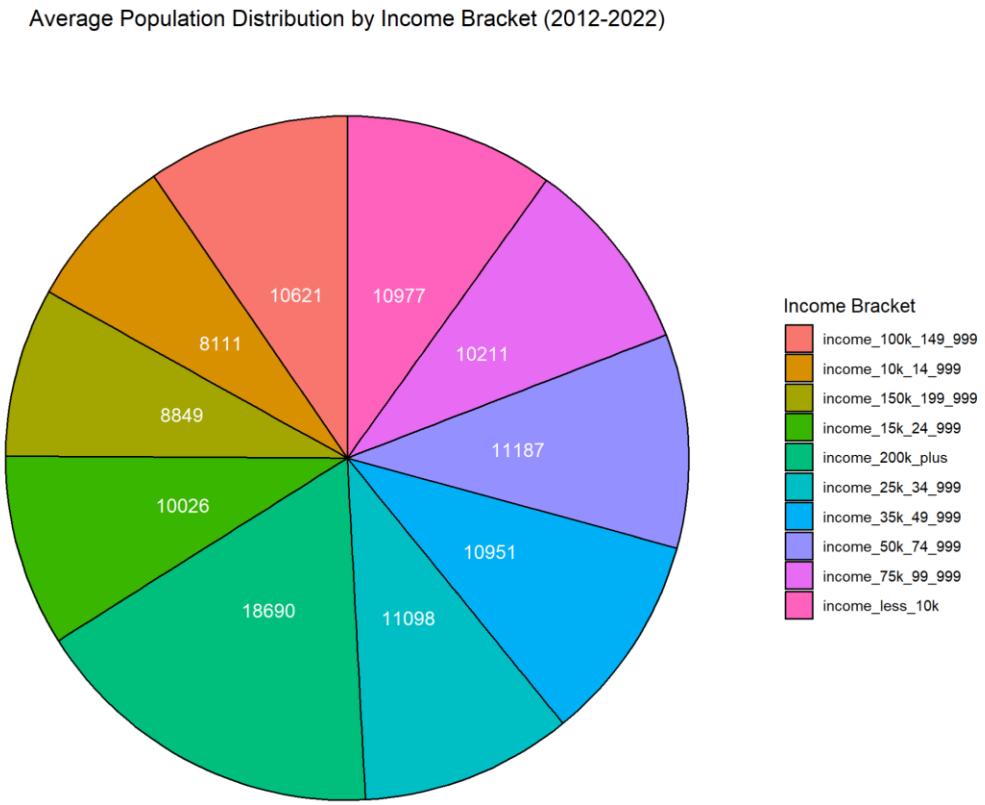


Image 2. Income Distribution

According to *Partisanship by race, ethnicity and education* by Pew Research Center (2024)[4], education level is strongly indicative of political affiliation. In other words, The

higher a person is educated—bachelor's or more—tend to vote democratic where as individuals with no schooling, just high school education or 2 year college degrees tend to vote republican.

By accessing the ACS 1 year data from 2012 to 2022, it can be inferred that the population with bachelor's degree or more make up majority of the population pool[5]. Image 3 shows how more and more people over the years have been getting more educated than the previous years. This analysis and data visualization was done using RStudio.

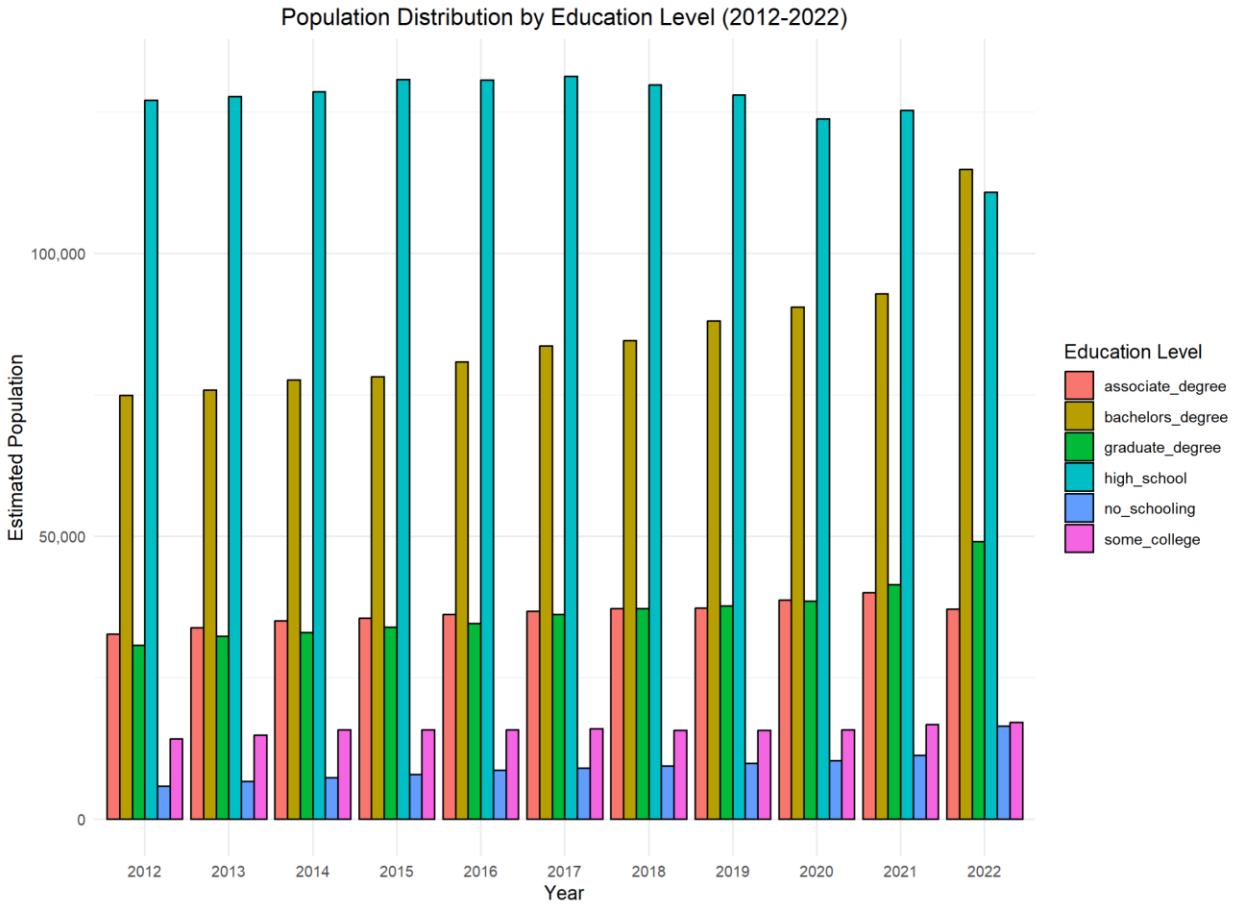


Image 3. Education Distribution

Although all inferences made so far seem correlative, when applied the research findings as well as previous elections' results to the population distribution under different variables, it shows a very strong positive correlation between the election outcome and variable distribution.

Spatial Analysis

Data Processing

For the data processing, I began with validating the spatial datasets. Invalid geometries were identified and resolved using tools in RStudio, such as st_make_valid(), to ensure compatibility with spatial operations. Additionally, I aligned all spatial datasets to a common Coordinate Reference System (CRS) to ensure that all data could be accurately overlaid and analyzed.

To focus the analysis on IL_D3, I clipped block group and precinct-level data to the district's boundaries using st_intersection(). This step ensured that all records in the final dataset were strictly within the district and excluded any overlapping or irrelevant data. Missing or incomplete data points, particularly in voter turnout records, were carefully handled by cross-referencing alternative sources or excluding those records where necessary to maintain dataset integrity.

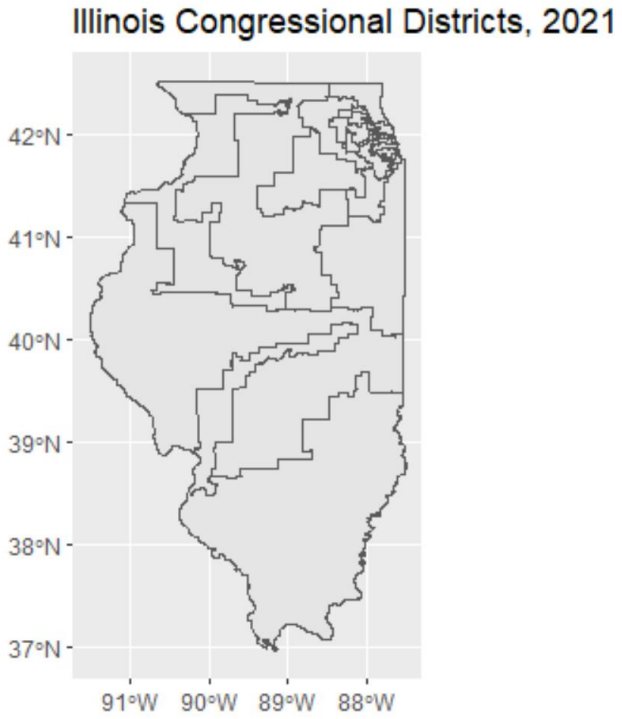
Data Cleaning

During the data cleaning phase, I focused on ensuring the datasets were accurate, complete, and consistent. Invalid or incomplete entries, such as records with missing voter turnout or demographic attributes, were identified and removed to maintain data integrity. Duplicate records, which could have resulted from overlaps in shapefiles during merging, were carefully removed to prevent over-representation of certain areas in the analysis. These data cleaning steps ensured the final dataset was reliable and well-prepared for subsequent modeling and spatial analysis.

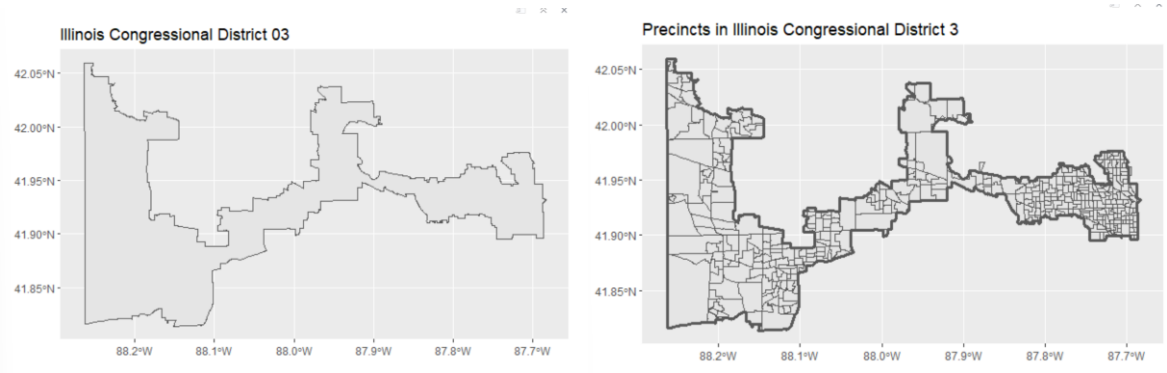
Spatial Analysis

The shapefiles were used to conduct spatial analysis using RStudio. First, I clipped the block group and precinct-level geometries to District 3's boundaries using st_intersection(), ensuring that all spatial data included in the analysis strictly corresponded to the district.

All the districts within Illinois is mapped. IL_D3 is located within the top right corner of the state.



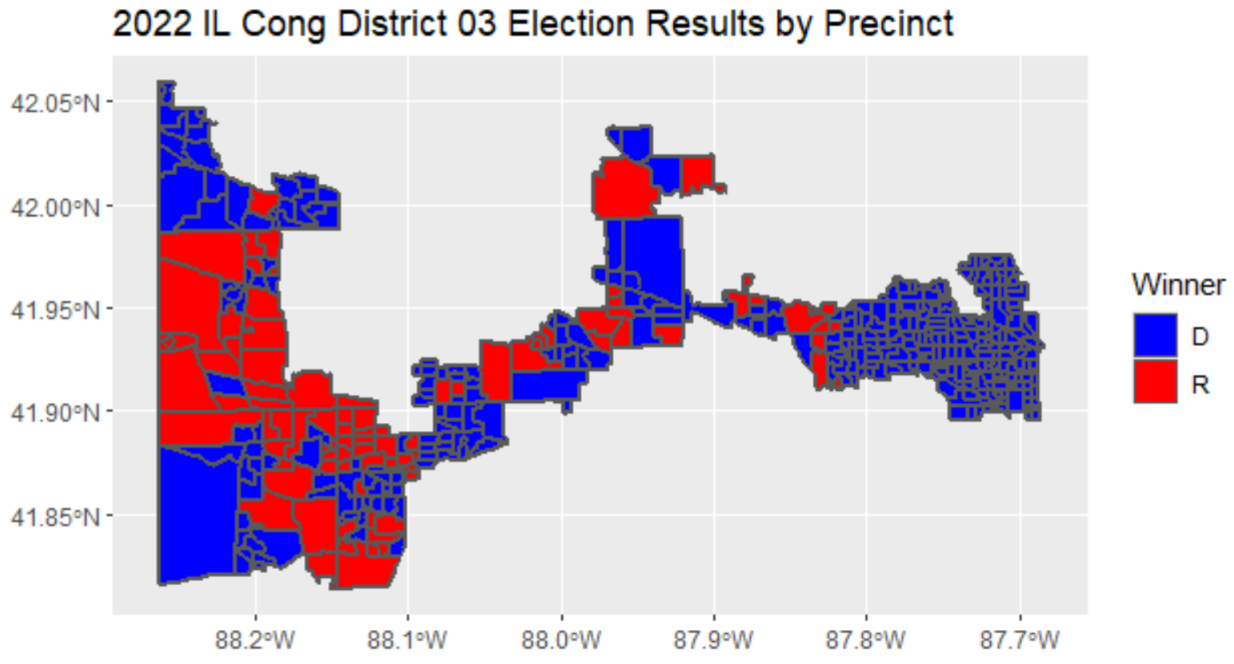
The following is the structure of IL_D3 without and with the precincts added. Historically, since IL_D3 has been redistricted a few times over the period of 2012 to 2022, so only 2022 precinct data is taken into consideration[4].



The following image shows the election results of 2022 congressional election held within different precincts of IL_D3. The results of the election were as follows. The Democratic party won with majority votes.

121764 votes for the Democrat

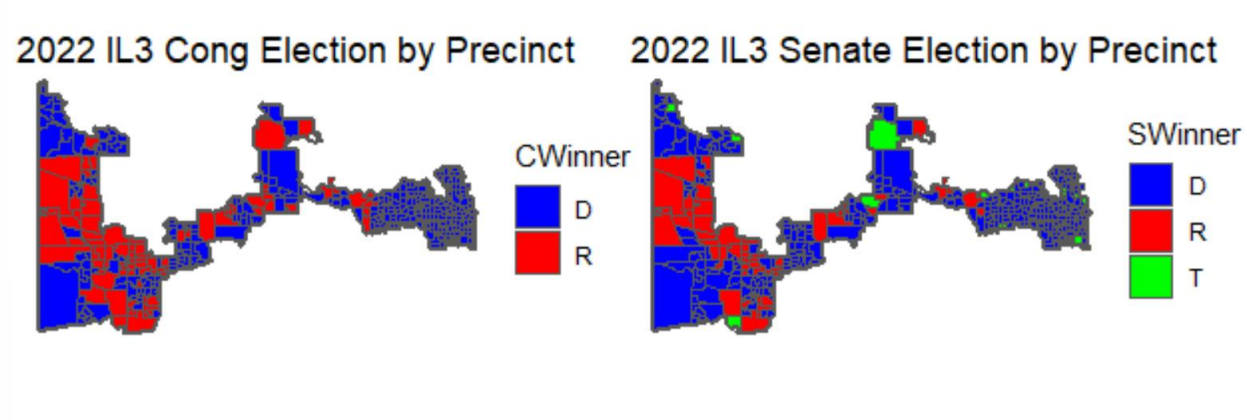
55995 votes for the Republican



Similarly, the Senate election also showed congruent results to U.S. House election. The following are the results of Senate election.

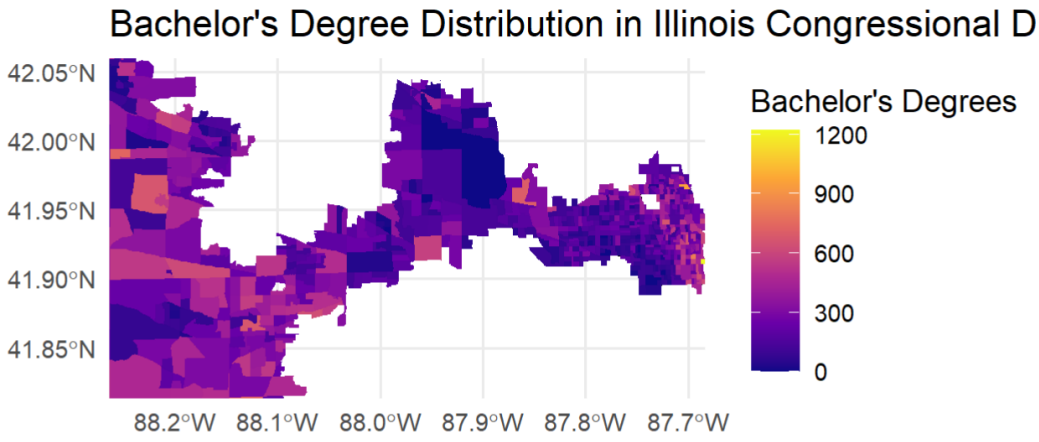
119561 votes for the Democratic Candidate for Senate

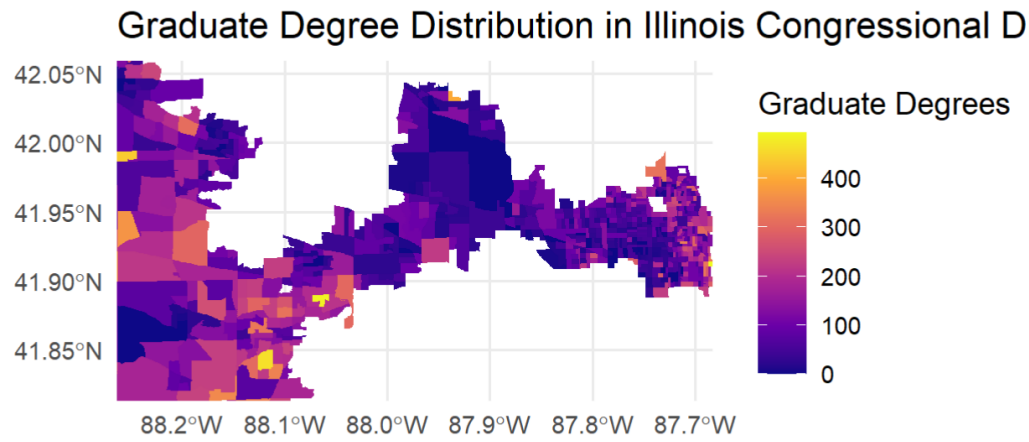
50973 votes for the Republican Candidate for Senate

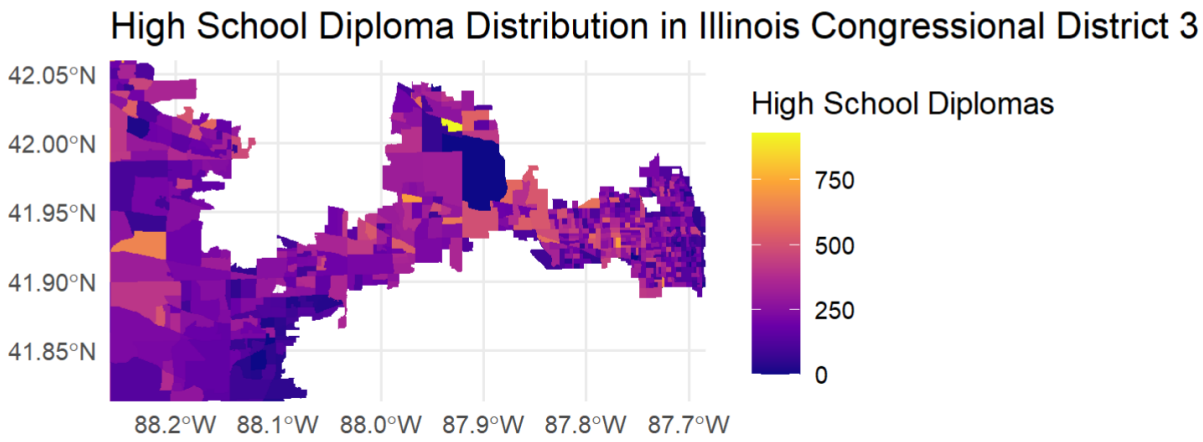


In general, it can be concluded that for both Congressional and Senate Elections, IL_D3 predominantly votes for the Democratic Party as well as has similar voting behavior across precincts.

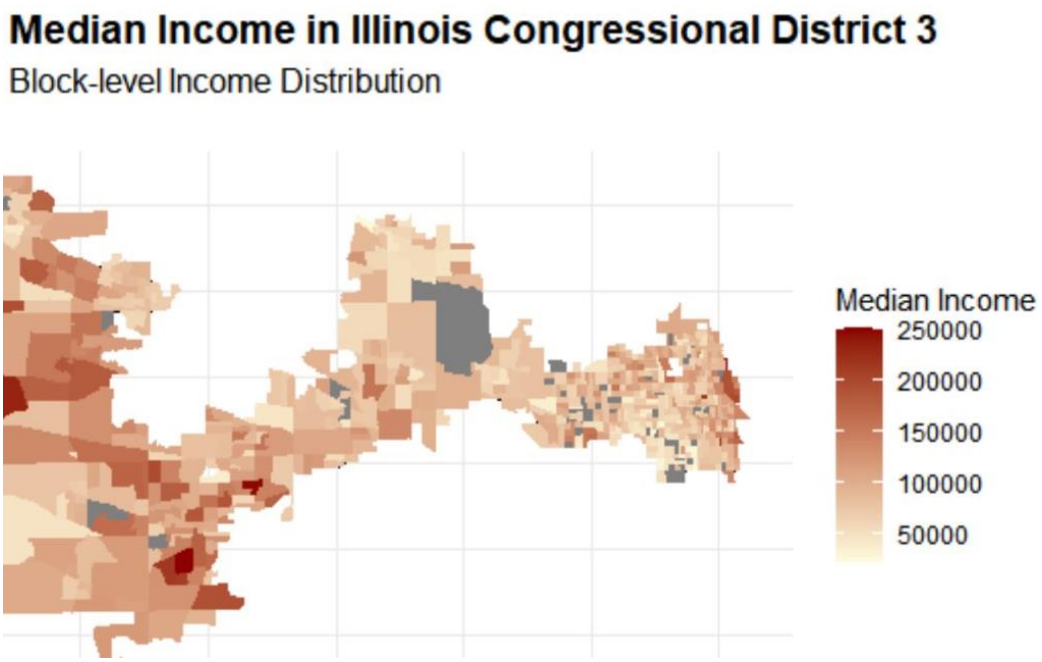
Next, I mapped the education distribution across IL_D3. In the below image, it can be inferred that the individuals with just high school education is higher than individuals with higher education in specific localised regions. These specific localised regions are also similar to the results of the 2022 Congressional Election. When superimposed, regions with lower education is where more Republican party wins have taken place. Although it is not all the regions, but most of them are congruent. Similarly, majority Democratic wins are higher in locales with bachelors or graduate school education is present.







Similarly, I also analyzed population distribution within each income bracket and found that most of the areas with individuals with higher income or lower income voted for the democratic party where as the middle income voted for the republican party.



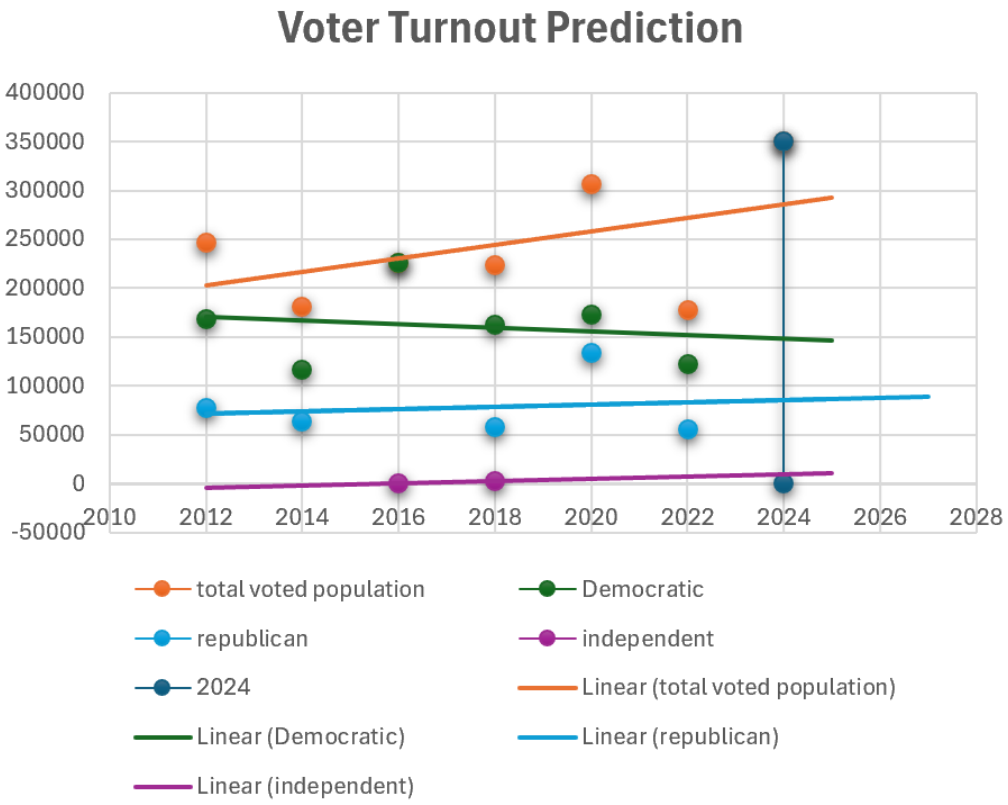
Model Approach

To predict voter turnout and election outcomes for IL_D3, I used historical election turnout data from Ballotpedia.org, covering 2012 to 2022. The dataset included total voter turnout and party-specific turnout for the Democratic, Republican, and Independent parties. Using Excel, I plotted the turnout data over time and applied linear regression to estimate trends. A regression line was fitted for each party's turnout and for the total voting population, enabling projections for the year 2024.

The eligible population number was derived directly from an ACS dataset, specifically from a CSV file listing the total civilian population over the age of 18 for the district. No separate model was created for this calculation; the data was used as-is to provide context for the turnout predictions. Independent party turnout was excluded from the analysis due to its negligible impact and irregular participation in elections.

Voter Turnout Prediction

In order to find out voter turnout, I used Ballotpedia.org's resources. I found the total number of people who voted since 2012 until 2022. I plotted and found the regression line that fits best with the population distribution across Democratic Part voters, Republican Party Voters, Independent Party voters and total eligible voting population.



After finding the linear regression, I looked at what y-value was at year 2024 for each regression line. In year 2024, the population who votes, according to linear regression for each party and the total turnout are as follows. In conclusion, about 38.3% voter turnout is predicted for IL_D3 in the upcoming 2024 election. The winning party at the U.S. House representing IL_D3 will be Democratic Party.

Unfortunately, senate level voting results were unable to be acquired for 2020 as no Congressional District information was available—only Senate District 3 voting results were available. This information was not used for predicting Presidential win because the area covered under both districts are very different. Due to this reason, the Presidential election results are predicted using the predictions of U.S. House member elections. So, Senate election within IL_D3 will have similar outcome (38.3% turnout and democratic party win). This charts with the 2022 election results mentioned.

119561 votes for the Democratic Candidate for Senate

50973 votes for the Republican Candidate for Senate

	Population
Democrat predicted	148593
Republican predicted	80804
Total turnout	230040.4
Turn out percentage =	Total Turnout/Total Population
=	230040.4 / 601334
=	38.25501

In the 2024 Congressional election, the Democratic Party won IL-D3 as predicted. My model estimated that the Democratic Party would receive 148,593 votes, while the actual number of votes was 114,490. Similarly, the Republican Party was predicted to receive 80,804 votes, but the actual number was 65,549 votes. Although the actual vote counts were lower than the predictions, the proportion of Democratic to Republican votes remained consistent with historical results. The actual voter turnout was 29.93%, significantly lower than the predicted turnout of 38.3%.

Conclusion and Discussion

This project successfully predicted voter turnout and election outcomes for IL_D3, with the Democratic Party winning the 2024 Congressional election as projected. However, the actual voter turnout was 29.93%, significantly lower than the predicted 38.3%. This reflects a broader challenge in accurately forecasting voter engagement, particularly in districts with historically low participation rates.

The low turnout in IL_D3 aligns with a national trend in the United States, which has consistently experienced lower voter participation rates compared to many other first- and third-world countries. According to a Pew Research study, the U.S. ranked 31st out of 50 countries in voter turnout for recent elections, despite significant increases in participation over the past decade. Factors contributing to this include systemic barriers to voting, political apathy, and voter fatigue, particularly in non-presidential election years.

In IL_D3, the low turnout is particularly surprising given the district's high educational attainment and income levels—characteristics typically associated with higher voter engagement. This discrepancy could be attributed to the bystander effect, where voters in heavily partisan districts feel their participation is unnecessary due to the perceived inevitability of the outcome. Additionally, logistical barriers and the limited appeal of non-competitive races may further discourage participation.

Future Study

Improving Data Availability: Acquiring more historical data, such as precinct-level Senate election results and accurate voter eligibility statistics, would enhance prediction accuracy and address data gaps.

Incorporating More Variables: Including additional demographic factors like age, employment status, and racial composition could reveal deeper insights into voter behavior.

Using Advanced Models: Experimenting with advanced statistical or machine learning models could capture non-linear relationships and temporal dependencies for more robust predictions.

Understanding Voter Engagement: Analyzing barriers to voter participation in IL_D3 could uncover actionable strategies to address low turnout.

Sources

[1] https://en.wikipedia.org/wiki/Illinois%27s_3rd_congressional_district

[2] https://datausa.io/profile/geo/congressional-district-3-il

[3] https://www.pewresearch.org/politics/2024/04/09/partisanship-by-family-income-home-ownership-union-membership-and-veteran-status/

[4] https://ballotpedia.org/Illinois%27_3rd_Congressional_District_elections,_2012

[5] https://www.census.gov/programs-surveys/acs

[6] https://redistrictingdatahub.org/state/illinois/

[7] https://en.wikipedia.org/wiki/Illinois%27s_3rd_congressional_district#External_links

Other sources

Chat GPT

https://www.electproject.org/election-data/precinct-data

https://dataverse.harvard.edu/file.xhtml?fileId=4773525&version=45.0

https://ufl.maps.arcgis.com/apps/webappviewer/index.html?id=eef1356a00df4e8dafc848482bc81fbd

https://www.graphpad.com/quickcalcs/linear2/

https://www.pewresearch.org/short-reads/2022/11/01/turnout-in-u-s-has-soared-in-recent-elections-but-by-some-measures-still-trails-that-of-many-other-countries/

GITHUB REPOSITORY: https://github.com/jujusjuicebox/CS579-Final-Project-Codes