CS504 Project Report

Introduction

Voter turnout among members of different groups of Americans varies widely, with Latinos and Asians generally lagging behind other groups. Blacks usually fall in between, with turnout usually ahead of other minorities but behind whites. Low levels of voting matter, because election results are supposed to reflect the preferences of all Americans. In addition, recent trends indicate that Latinos, if they vote at their full potential, have considerable capacity to influence election outcomes. Amplify Latinx is a non-partisan, collaborative movement whose mission is to build Latinx economic and political power by significantly increasing Latinx civic engagement and representation in leadership positions across sectors.

Project Description

The goal of our project is to identify the districts (down to the cities and towns) in Massachusetts where Amplify Latinx and its partners should deploy resources to increase voter engagement, specifically voter registration and voter turnout. In order to do this, we need to identify how many potential voters there are in colored people.

To achieve the goal, we define specific questions:

1. What is the difference between eligible people and registered voters?
2. How many registered voters are actually voted?
3. What is the percentage of each age group people among the voters?

4. What is the difference between the number of people who voted for the winning candidate and the number of people who voted for the candidate with the second-largest number of votes?

Question 1 aims to figure out the maximum potential population of people that can be activated to become voters by our promotion policy.

Question 2 is used to see if there is a low percentage of people who actually voted. If the answer is yes, we can conclude this ward worths more promotion.

Similarly, for Question 3, the young people who will be eligible people soon is considered mostly by us. And these people are deeply influenced by young voters.

Question 4 is to answer the question: if every Lantinx person who was registered had actually voted, would that have flipped the race? If ever Latinx person who was eligible to vote had registered to vote and then voted, would that have flipped the race?

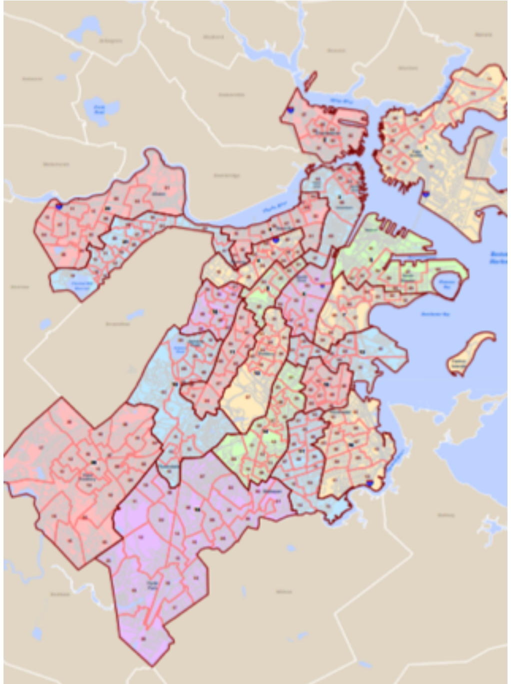


Figure 1. Wards in Boston

In order to refine the demand, we just pay our attention to Latinos in 22 wards of the Great Boston Area in this project, which is showed in the above figure. But the same method can be applied to other race and other districts.

Data Description

1. Vote Builder by NGP VAN

   This source provides us with state election data and voter-specific data for Republican districts in Massachusetts. This data is grouped by district and party affiliation and provides summary statistics for the past four elections for the following attributes: i. voter participation in the past four elections; ii. voter race; iii. voter sex.

2. PD43+

   This source provides Massachusetts election statistics for all elections at the state/federal level since 1970. Our analysis uses the State Senate and House election results in 2017 in all districts. PD43+ provides data on current and past incumbents name, sex, party, vote count, and vote percent for each individual district.

3. OCPF

   This source provides political contribution data including donation amount and donor district for each year since 2009. We use web scraping to get this data. This data can elucidate in which districts any candidate might most successfully raise money and will be used as a feature in our model.

4. US Census data

The Census Bureau collects demographic data about the economy and the people living in the United States from many different sources. Primary sources for additional data are federal, state, and local governments, as well as some commercial entities. We using the data of district information and district vote statistics to generate a population breakdown with respect to certain key features of interest (political affiliation, gender, race, and income) and use them as inputs in our predictive model.

5. Analyze Boston

   Analyze Boston is the City of Boston's open data hub to find facts, figures, and maps related to our lives within the city. They have a series of high-quality and up-to-date datasets and develop a platform that is widely accessible.

6. Data form Amplify Latinx

   We also have some private datasets that are provided by our partner.

Data Processing

Data processing is a critical step to ensure that all of the data collected is of a standardized form such that it can be considered in the aggregate as input to our predictive models. Our goal for the data preprocessing stage was to obtain several aggregate datasets that take into account all of our collected data. This stage can be divided into 3 steps:

1. Fetch the dataset of ages in different congressional districts from US Census Bureau website, and combine them with the dataset from our partner's more detailed dataset to get the age and gender distribution in different CDs;

2. Fetch the map of 10 CD in MA from US Census Bureau website for the past couple of years, convert them into standard GeoJSON format (using ogr2ogr tool) and categorize them by their CD id;

3. Fetch the dataset of candidates and candidate's committees from MA OCPF website, and combine the information with all registered candidate office districts to find the correlation between candidates and districts.

4. Gather and organized ward map data and information for each ward in Great Boston from the website analyze Boston. Linked every registered voter to data in every ward in the database.

Methodology

In order to give suggestions for which area is the best choice to focus voter turnout efforts, we use the values gathered in data preprocessing as input to establish the constraints to use in the constraint satisfaction and optimization algorithm.

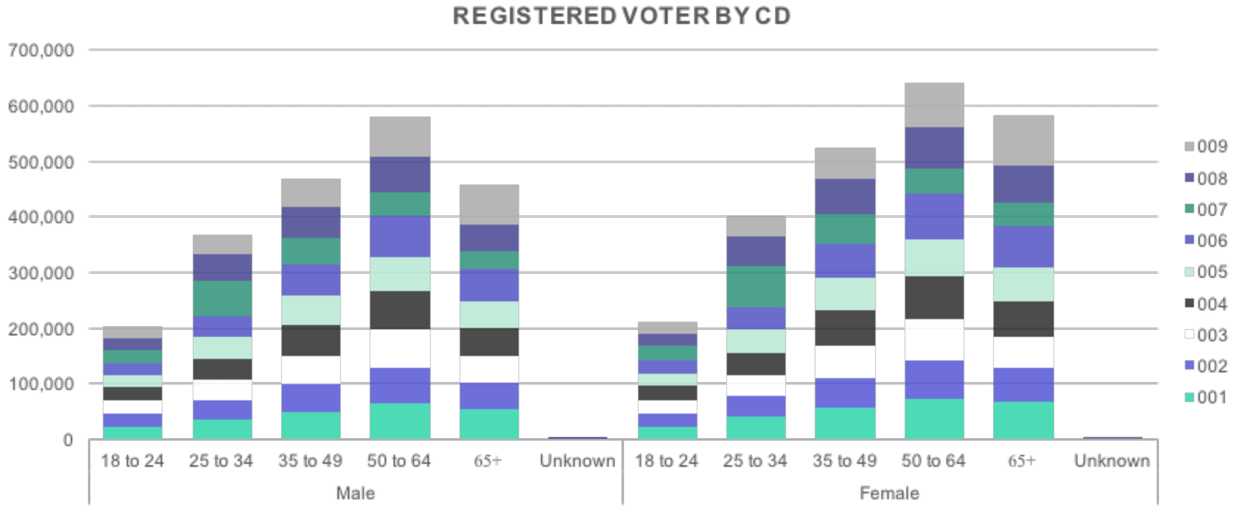


Figure 2. Registered voters number aggregated by congressional district

We tried to find the strength of the relationship between each age group in Boston for non-registered voters and registered voters. In order to solve this problem, we decided to try to calculate the correlation coefficient for each age group. We first fetch the data from a bunch of excel spreadsheets. Then, we aggregate the population data by different age groups in each ward. After that, we calculate the correlation coefficient for different age groups and the total number, as well as the p-value. In this way, we can find out how strongly each age group is connected to the total non-registered/registered group.

Results

We build a web page to visualize voting statistics and election data in Boston City and its 22 wards. As the following screenshot shows, the user can see the differences between registered and eligible voters, the differences between voted and registered voters on both city and ward scale. For the difference between the first two winning candidates, it can be shown on the ward scale. All these data are available for the past 10 years.

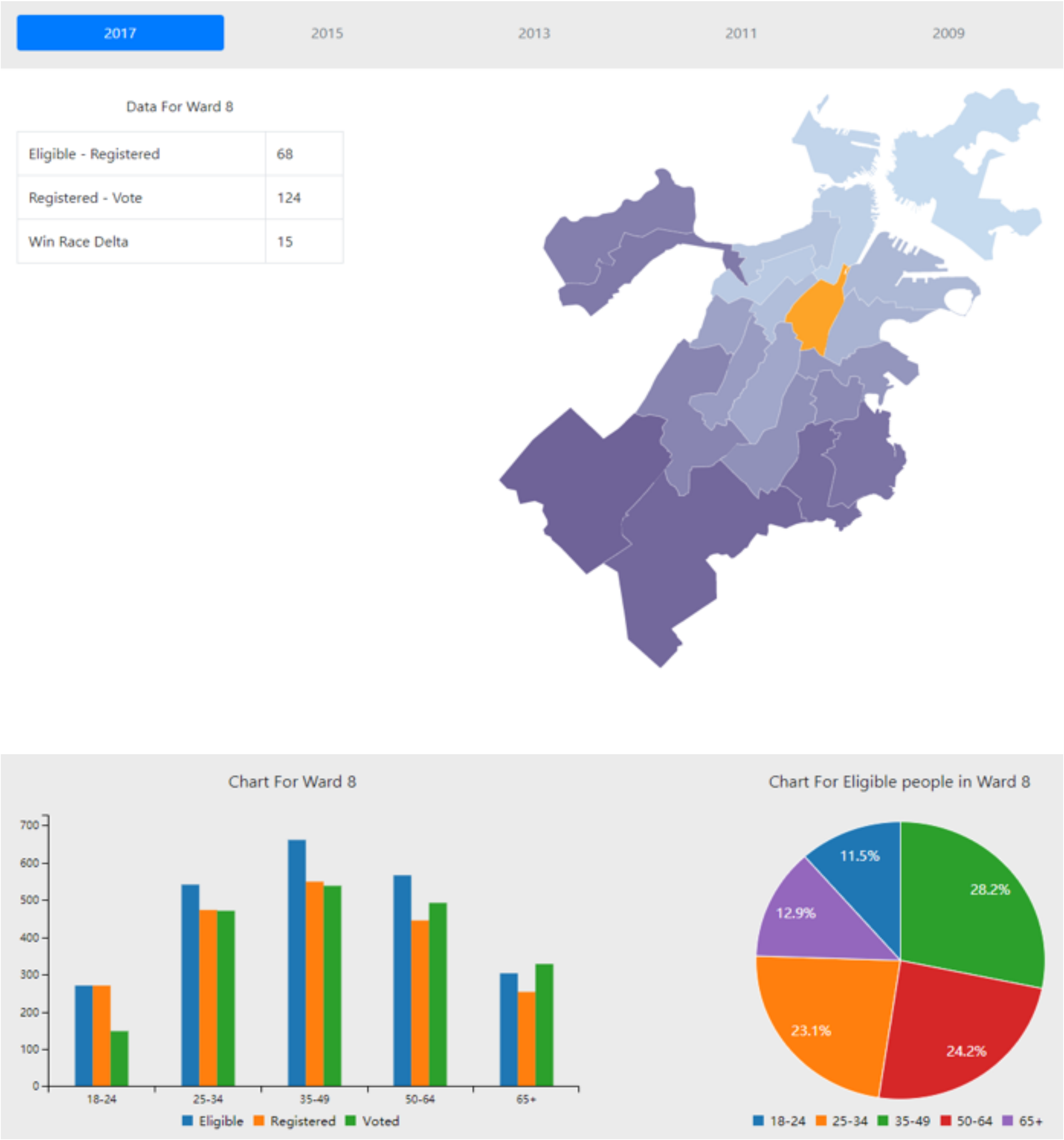
By clicking on of the exact ward on the map, users can check the bar chart and pie chart of different age groups, which helps visualize the proportion.

We also created a couple of Restful API which can fetch the numerical data by ward level. Through these API, user can access:

i. all the election and voting statistic data for any single ward;

ii. the election data and gap of the first two candidates for any single ward;

iii. the voters' distribution data for any single ward.



Data For Ward 8

Eligible - Registered	68
Registered - Vote	124
Win Race Delta	15

Figure 3. Visualization charts for various data of ward 8

Future Work

Due to the time constraints, currently we only focus on the proportion of Hispanic votes in the 22 wards in the Great Boston area. In the future, we hope to extend to all races and all districts of Massachusetts, because a larger amount of data will be more statistically significant, which will give more accurate conclusions.

Additionally, we did not consider the impact of income on voting in this project, which in real life is an important factor. We should take this factor in to account.

Finally, for the election, the appropriate propaganda methods for different age groups are different. If we want to give the best way to improve the votes, we should investigate the appropriate propaganda methods for each age group and make a statistical analysis.