Growing Voter Engagement in Communities of Color

Carlos Portillo, Jean P. Vazquez, Tallulah Kay, Gabriel Yllescas

Introduction

The United States is a country which was founded on a government elected by the people. It is a concept so integral to what it means to be an American that even the word "American" is often conflated with terms such as freedom of choice. As such, it is very surprising that among all other democratic, well-developed countries, the United States consistently places near the bottom[1] in terms of voter turnout at an average of 60%[2]. However, this value is not very useful by itself and in order to obtain a better idea of how to improve the United States' voter turnout rate, one must also take into account race[3]. In the United States, the racial groups with the highest average voter turnout are non-Hispanic whites and African-Americans with rates between 50-65%. On the other hand, Minorities such as Hispanics and all other racial groups not previously mentioned (i.e. Asian, Native American, Pacific Islander, etc.) consistently have a voter turnout rate between 30-45%. As such, the best way to move towards higher turnout rates must begin with increasing voter turnout among these groups. This is especially so given that Hispanics and other Minorities make up almost 17% of the American population, which translates to 43 million people, 30 million of which aren't voting.

The goal of this project is to create an interactive map of Massachusetts to help guide groups seeking to increase voter engagement towards where they should focus their efforts in order to get more people of color out to vote. Our team, in conjunction with Amplify Latinx and BU Spark!, plans to accomplish this by analyzing voting and demographic data in order to find geopolitical subdivisions (i.e. wards, precincts, etc.) that have many people who can vote or can register to vote but don't end up participating. As our team is responsible for creating only the groundwork for what will be a much larger project in the future, the scope of data analyzed will be limited to city council race results among Latinx voters within Boston Wards. However, the

[1] https://www.pewresearch.org/fact-tank/2018/05/21/u-s-voter-turnout-trails-most-developed-countries/
[2] https://www.fairvote.org/voter_turnout#voter_turnout_101
[3] http://www.electproject.org/home/voter-turnout/demographics

code shall be built with modularity in mind in order to allow for all political races within all the geo-political subdivisions within Massachusetts.

Datasets

For the scope of this project, we are using 6 data sets, three of which are public that we found through our own research, and three private data sets which were provided by Amplify Latinx. The three public data sets are Demographics by Towns, City Council Race Results (2017), and MA Voter Registration Data (2010). Demographics by towns lists all 353 towns in Massachusetts along with their total population and racial demographics. City Council Race Results (2017) contains a breakdown of all the votes cast during the City Council Race of Boston in 2017 per Candidate within each Ward. MA Voter Registration Data (2010) contained a breakdown of registered voters by party affiliation for all 353 towns in Massachusetts. The three private data sets are Registered and Non-Registered Voter Demographics and Massachusetts Early Voting. Registered and Non-Registered Voter Demographics contain the total amount of registered and non-registered (eligible) voters, respectively, for all of Massachusetts' Senate districts, Precincts, and Congressional Districts broken down by Age and Race. Massachusetts Early Voting lists the total amount of registered voters in each town of Massachusetts along with the percentage of those voters who submitted an early voting ballot for the 2016 presidential election.

Methodology

In order to obtain a list of which areas such aforementioned groups should focus on, our team identified 3 distinct phases of data transformations and analyses. The first phase was comprised of parsing through the data sets we would need in the second phase and performing the necessary transformations in order to isolate the data we would need from these data sets. The second phase involves generating all the metrics we would need for the third phase and the final visualization. The third phase consists of designing and implementing a constraint satisfaction and optimization algorithm responsible for generating the list of Wards that we would suggest visiting given the calculated metrics and some user inputs. However, this third phase is only executed at the user's request through a button on the website page that will contain the interactive map of Massachusetts zoomed in and demarcated to show Boston's Wards. Regardless of whether the user submits this request or not, the map will show some of the statistical information generated in phase 2 for a Ward when that specific Ward is hovered over.

For the first phase, since the scope of this project encompasses only Wards in Boston, we had to obtain the total amount of registered and eligible Hispanic people in each Ward from the Registered and Eligible demographics datasets. This was done by first aggregating all precincts by Ward for each data set. Following this, we projected exclusively the results for Hispanics in each Boston Ward into two separate tables for registered and Eligible voters, respectively.

For the second phase, we began by obtaining geographic coordinates for all the Wards in Boston in order to be able to demarcate them on the final map. The coordinates for all the Wards were obtained through a URL request to the boston.gov.data website, which contains a data set listing all Boston Wards as well as their location and size. We also added a script which gathers the coordinates for all the towns in Massachusetts using the Google Geocoding API along with the data contained in the Demographics by Towns dataset. However, this script will not be used in our project and is only present in order to simplify the work done in future iterations of this project. The second set of metrics was gathered using the City Council Race Results data set and calculates both the overall and by Ward difference in votes between the first and second place candidate. This "difference to flip" was calculated by first aggregating the number of voters in each Ward by the candidate they voted for in order to find who came in first and second place overall. Then we obtained the difference between the two highest totals to obtain the "total difference to flip", which refers to the total amount of votes the second person would need in order to flip the race. This was then also done for each ward to generate a list titled "diff by ward" containing the difference between the amount of votes obtained in that ward by the winner and runner-up candidates. The last set of metrics is made up of several values generated from a statistical analysis of the population data for registered and eligible voters. The values generated both for each ward and overall are: total amount of registered voters; total amount of eligible voters; total amount of Hispanic people, both registered and eligible; the proportion of Hispanic people which are eligible; proportion of Hispanic people which are registered; the proportion of the total population (eligible or registered) which are Hispanic; and the average amount of registered and eligible voters in each ward which are Hispanic as well as the standard deviation of this value. These statistics are all generated through several sets of aggregations which first aggregate each of the aforementioned values within each Ward and then across all Wards. The reason why we gather these values are to then use as a constraint for the algorithm in the last phase by choosing which statistic to prioritize in our suggestions for where to focus voter turnout efforts. However, given this project's limited scope, we will not give the option to prioritize anything other than the proportion of total Hispanics.

Lastly, phase three brings together all metrics collected in phase two (aside from the geographic data) and some user inputs to establish the constraints to use in the constraint satisfaction and optimization algorithm. As mentioned previously, the goal of the algorithm is to generate a list of suggested Wards to visit and it does so through a two-step process. The first step consists of identifying the Wards that meet the constraint criteria. While the second step is to then generate an optimal result by minimizing the amount of wards that would need to be visited to flip the entire race. The criteria for the first step are the proportion of Hispanic people in a Ward, regardless of whether they voted or are registered and the particular race and geopolitical subdivision that the algorithm should be run on. Given the scope of this project, this second constraint will be locked to Ward results for the Boston City Council race of 2017. After generating the list of Wards which fit these constraints, the algorithm will then generate the amount of registered voters which could be flipped in each Ward as based on the user input variable dubbed "flip percentage". The "flip percentage" refers to the percentage of registered voters that didn't vote which a particular campaigning group believes they can convince to vote for the runner-up candidate. Then, using this list, the Wards are sorted in descending order by number of registered voters who didn't vote and are added to the final list of wards to visit one by one until the amount "flipped voters" exceeds the amount of votes necessary to flip the entire race.

Toolset and Difficulties

This project was coded entirely in Python with Git as the version-control system, MongoDB as the database, Folium for map visualization, Flask as the web framework, Google's Geocoding API for obtaining geographical coordinates to be used in the map, and a variety of smaller Python modules to generate provenance for all the different files and data. We also used Excel to convert all the files we originally received from xlsx to csv, which we then converted to json using free online software.

As for difficulties encountered during this projects' development, the main issue we faced was the formatting the data sets we received from our partners. After some brief attempts at parsing through our data, it quickly became apparent to our team that several design choices which eased manual readability in Excel would in turn make the files unreadable once converted into .csv files. Things such as multiple levels of column and row indexes as well as empty cells would consistently cause data to be stored with the incorrect keys, thus making it impossible to correctly perform transformations. Although we considered trying to fix this manually and using python, we found that both methods were too time-consuming, although for different reasons. Given the size of our data sets, some of which contained tens of thousands of data points, manually fixing was quickly discarded. However, using Python to fix them would have also been too time-consuming due to how much each file's design varied within it as well as how different every file was from every other file. Meaning that, in order to programmatically re-format every file, we would have to custom make python code each one, each of which would have to cover an extremely large amount of different cases. In order to fix this issue, our team and the project partners agreed to change the scope of the project based around the availability of easily readable files.

Conclusion

With this project, the main problem that we sought to solve was to increase voter turnout by targeting areas with a large Hispanic population. However, even after looking at various statistics which indicate a correlation between race and voter registration/turnout, we first sought to confirm whether this was the case within our dataset. As a result, we performed two additional sets of statistical analyses. The first of which sought to find whether there was some correlation between the amount of Latinx voters in a population with the registration rate and as can be seen in figure 1 below, we found there to be at least a slightly negative correlation.

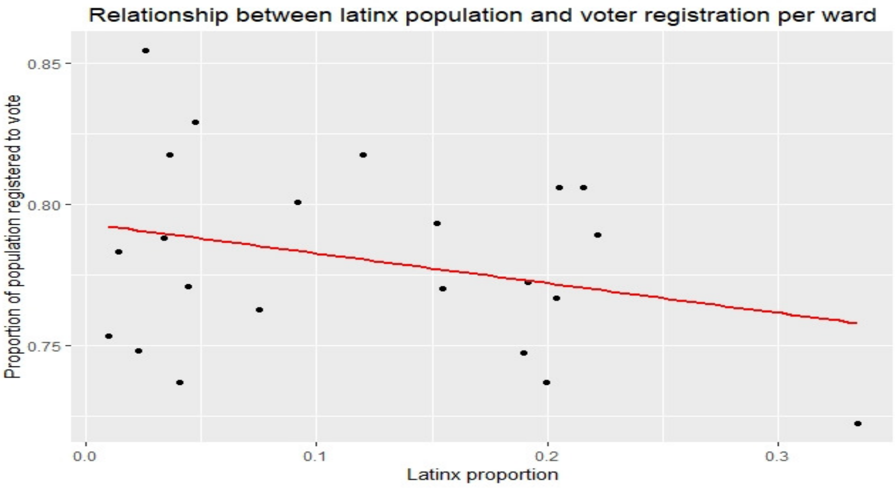


figure 1. Proportion of Latinx population vs. proportion of population registered to vote

The second statistical analysis conducted on the data then sought to gauge the impact the Hispanic population has on voter registration rates throughout Massachusetts as a whole. To do this, we used data from the third public dataset listed above, Demographics by Towns (2010), and voter registration data (2010) to compute 4 correlation coefficients. One for each of the following:

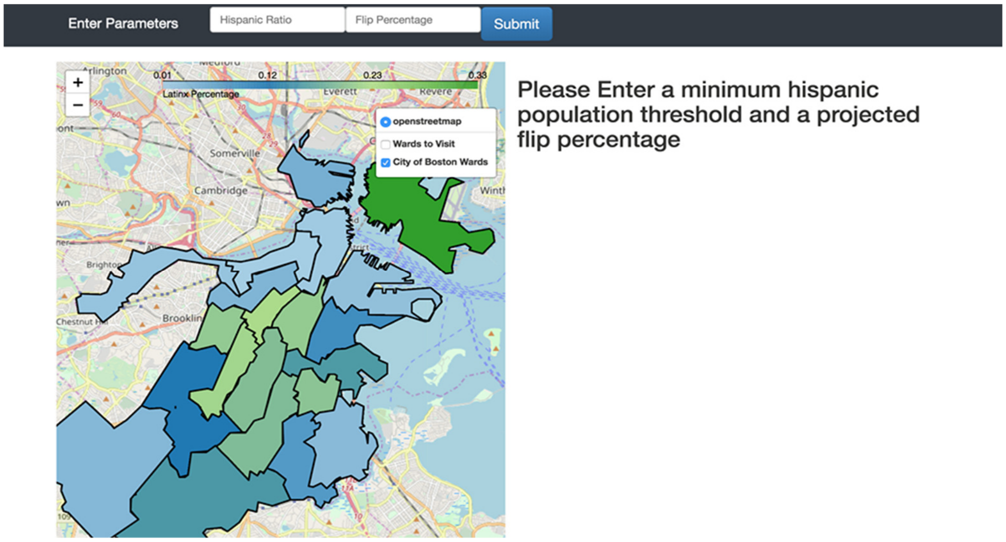
- Hispanic Population vs Overall Registration Rate, Correlation Coefficient = -0.07
- Black Population vs Overall Registration Rate, Correlation Coefficient = -0.05
- White Population vs Overall Registration Rate, Correlation Coefficient = 0.05
- Asian Population vs Overall Registration Rate, Correlation Coefficient = -0.02

We then computed the p-values for each one of them. We found the p-values for each to be as follows:

- Hispanic population and Overall Registration Rate, p-value = 0.21
- Black Population and Overall Registration Rate, p-value = 0.31
- White Population and Overall Registration Rate, p-value = 0.31
- Asian Population and Overall Registration Rate, p-value = 0.66

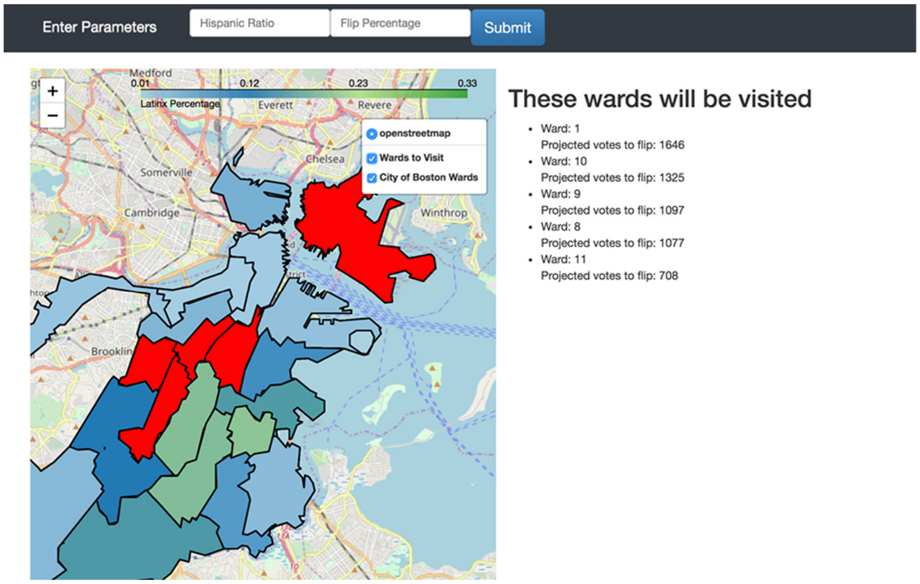
While in the scientific community a p-value is considered statistically significant generally if it is <0.1, we believe the gaps between these p-values, and the fact the Hispanic population size is negatively correlated with overall registration rates, suggests there may be merit to focusing voter engagement efforts on communities with larger hispanic populations.

Following this, we then began working and by using the approach and tools detailed above, created a visualization on a local server as shown in figures 2 and 3.



Please Enter a minimum hispanic population threshold and a projected flip percentage

figure 2. Map Visualization prior to running the algorithm for generating suggestions



These wards will be visited

- Ward: 1
  Projected votes to flip: 1646
- Ward: 10
  Projected votes to flip: 1325
- Ward: 9
  Projected votes to flip: 1097
- Ward: 8
  Projected votes to flip: 1077
- Ward: 11
  Projected votes to flip: 708

figure 3. Map Visualization with suggested Wards to visit

Future Work

Given how limited the scope of this project had to be in order to ensure it could be completed in a timely fashion, the best way to iterate upon this work would be to expand certain functions of the map. The two most useful features that could be added to the map would be the ability to choose which political race the statistics will be generated for and which geopolitical subdivision should be demarcated on the map. This would allow groups to receive campaigning suggestions for any political race that could be either as broad as towns or as specific as precincts. Additionally, giving the user the ability to modify the constraints used to generate suggestions would also greatly aid with voter turnout efforts in the case that a group wants to target different groups with different race or age demographics.