Taking Gender Equity to the Streets

Maoxuan Zhu, Dongyi He, Wei Jiang, Kaikang Zhu

Goal & Description

Looking through street names in Boston, we recognize that male names and pronouns are often the default and given Boston's long parochial history, we want to make sure our city represents its current demographics (52% women) and know if there are streets that are good candidates for re-naming. In addition, we expect to build an interactive site that allows city and public to evaluate the different options for renaming based on different variables.

To achieve the goal, we have to define some criterions: which streets are suitable to rename; if there are several streets in one district meet the requirement, how would we sort them so we can choose one that is the most suitable to rename.

Initially, there are 5 rules from our partners:

1. The number of other entities containing the same name within a one mile radius e.g. libraries, stores/ restaurant names, statues, parks, schools, landmarks.
2. Streets that are named after women or people of non-white should not be renamed.
3. Streets that are included on major landmarks or routes such as the freedom trail should not be renamed.
4. Streets with large population densities or traffic volume should not be renamed.
5. Streets that were renamed recently, e.g. in last 15 years should not be renamed.

We did some researches and we found that it’s kind of hard to figure out which streets have been renamed in past 15 years so we decided not to consider this factor.

Aside from rules mentioned above, we came up with some more rules:

1. Streets that are named after United States celebrities or Massachusetts celebrities should not be renamed.
2. Streets that are named after universities or colleges should not be renamed.

Before we start our data processing and analysis, we’d better look at existing datasets and do some simple transformations to see what we can know from those data.

1. How could we know if a street should be classified as male, female or neural?(NamSor API)
2. Where can we find transportation data and what information we can get by applying some aggregation operations to it.(Uber movement data)
3. Where can we find boston districts boundary data. If there isn’t any, how would we divide boston into 23 different areas.(Zip code areas boundaries available online)

Previous work conclusion

The streetbook of this project is a legacy dataset inherited from the previous team, which features all streets in Boston with their zip codes, full names, name without last word, and two Gender attributes.

To determine the gender of each street, they trimmed the name of the street to obtain only the probably "human name" part, then use two web service independently to predict the gender by the actual name. The results are stored in Gender and Gender2.

Datasets

1. **Boston Street Book**: Team who worked on this project last year have done the classification process so we can easily know the basic information of each street: Name, Gender, Zip code, etc...
2. **American and Massachusetts famous people**: This dataset comes from https://www.50states.com/bio/mass.htm and We parsed names of famous people to get both last name and first name of them.
3. **Traffic movement from Uber**: Uber movement data can be found at https://movement.uber.com/?lang=en-US and each row indicates both source location and destination location of a ride. But we found that uber locations are not that fine, streets are divided into small blocks instead of maintain name and location of each individual street. And most of streets used to represent their block are named after famous people so they usually get filtered out in early stage.
4. **Boston Colleges and Universities**: This dataset comes from http://bostonopendata-boston.opendata.arcgis.com/datasets/cbf14bb032ef4bd38e20429f71acb61a_2.csv. Basic information such as name, address, city, zip code, year of built are provided. But location of a university or college is given as a point so this makes it really hard to check if a street is within a mile of a university (for instance BU is built alongside charles river).
5. **Public Libraries**: This dataset can be derived from http://bostonopendata-boston.opendata.arcgis.com/datasets/cb00f9248aa6404ab741071ca3806c0e_6.csv. We can get name, district, zip code and geographical information of libraries. Again, geographical information is given as a point. But usually libraries will not be that large so it’s still impossible to filtered out streets within a mile of each library.
6. **Boston Landmarks**: This dataset can be derived from http://bostonopendata-boston.opendata.arcgis.com/datasets/7a7aca614ad740e99b060e0ee787a228_3.csv. For the origin dataset, we fill in the missing values and

select rows that “Petition > 15”. Then we can generate a dataset consists of six columns: Petition, Name of landmarks, Areas_Desi, Address, Neighbourhood, ShapeSTWidth. There is no geographical information involved so we can only filter out streets named after landmarks.

7. **Street GEOjson data**: Street GEOjson data consists of the length and geometry information of each street. Length of a street is an important criterion we applied to sort alternatives of each district.
8. **Zip code area GEOjson data**: Dataset can be derived from https://bostonopendata-boston.opendata.arcgis.com/datasets/53ea466a189b4f43b3dfb7b38fa7f3b6_1. Each row defines a zip code area together with its geographical boundary information. This helps us visualize our candidates of each district. Since we have to draw boundary for each district so we have to map every zip code area to a specific district. Corresponding information can be found at http://archive.boston.com/news/local/articles/2007/04/15/sixfigurezipcodes_city/.

Data Processing & Workflow

First, we collect some datasets from different websites as below:

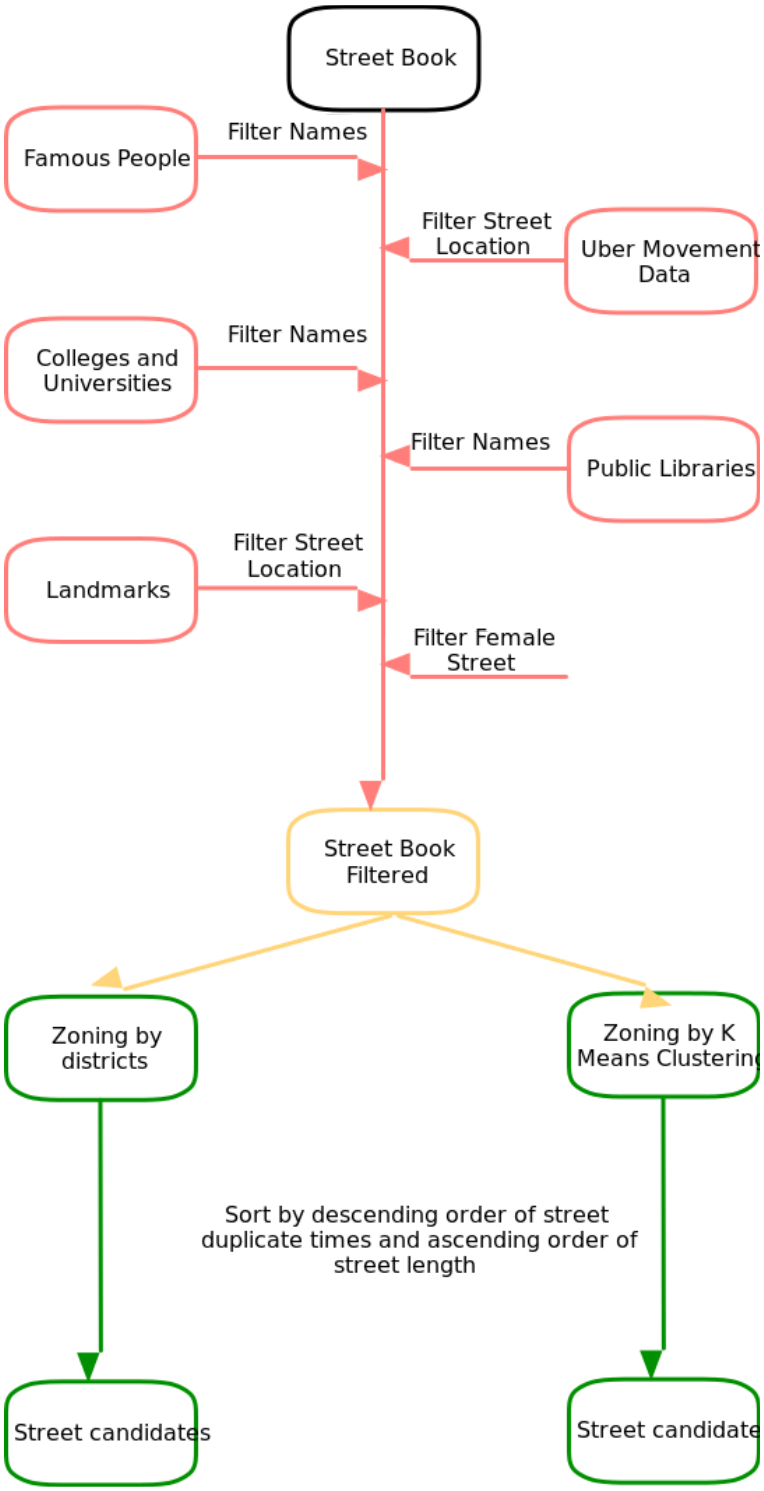
- https://www.50states.com/bio/mass.html
- https://movement.uber.com/?lang=en-US
- http://bostonopendata-boston.opendata.arcgis.com/

Also making use of the dataset obtained from the last team who did this project. Then we use web scraping technology to obtain structural data.

Next, we process the data through selecting certain columns which we think would be useful for us to do the project. Considering we having too many datasets, we choose to decrease the number of datasets. For example, we merge college and universities and landmark into cau_landmark. In this way, we reduce the number of datasets and make the data processing more clearly.

Then we filter the dataset using the renaming criteria. We filter all the streets in street book which have the same name as all the famous people’s names, same name as colleges and universities, same name as public libraries. And we filter streets which have large traffic flow according to our Uber movement data. We also filter all the streets same as the locations in landmarks such as the freedom trail. Finally, we filter all the street names labeled as female names. We get two processed datasets: one dataset containing full name of the streets, the name of street and zip code of the corresponding street. The other dataset is all the streets’ names which have duplicated

for more than one time. And the data is sorted in ascending order. To obtain the location information such as latitude, longitude, we add one more dataset which includes all the geographic information of all streets.



To deal with these dataset, we come out with two solutions. One is zoning by districts, the other is zoning by K-means clustering. For the first solution, we divide all the streets into 26 districts of Boston by their zip codes. Then we project zip code to its corresponding district and sort all the streets in each district by descending order of street duplicate times and ascending order of street length. We pick one candidate which has the largest duplicate times and shortest length for each district. For the second solution, we cluster all the streets based on their locations (longitude, latitude) and also select candidates based on the same criteria as the first solution.

After that, we extend our project with two visualizations. One is an interactive web-based visualization which is built with vue framework obtaining data through RESTful APIs provided by Flask server. It builds word cloud for every district and every cluster. For each word cloud, the bigger the word is and the darker the color is, the more duplicate times this street name has. The other is the candidate map using map display library provided by Mapbox using data from OpenStreetMap. There are four layers in the map: Boston districts (always shown), candidates per-district (optional), streets highlighted by K-Means clusters (optional), candidates per-cluster (optional).

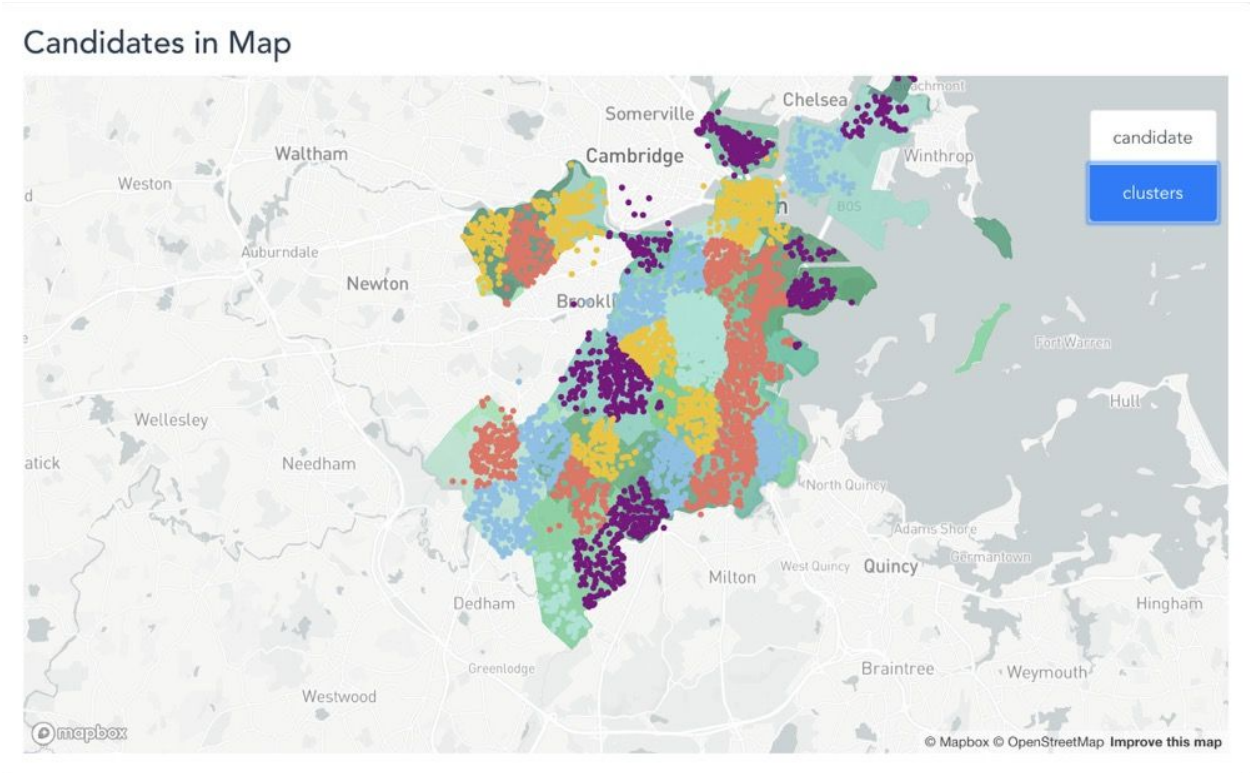
Finally, we provide possible candidates for districts of Boston and all the clusters. We hope this result can provide some advice and help for the government officers.

Result Analysis

1. As is shown in the candidates of entire Boston by neighbourhoods, street names like Reverend and Monsignor appear many times and show darker colors. However, these names don’t become our final candidate street name in every district, because such names are evenly divided into each district, so their repeating times proportion doesn’t outnumber others and won’t be considered as candidates.

Candidates of entire Boston

2. For each district in Boston, if there are several streets with same prefix names, we will compare these streets and pick the final candidate with shortest length.
3. For some districts, as their total area isn't big enough and they basically have no repeating names, we don't provide candidate streets for them.
4. As is shown in the map visualization of Boston districts and clustering, the clustering area doesn't match exactly with the real district distribution, which is reasonable (Boston authority won't divide the districts only from its streets distribution, but from various factors). Therefore, apart from the district candidates, we also provide the clustering candidates as an reference.



Boston districts versus clusters

Candidates of Cluster 0

5. When we look at the street candidate table and word cloud for this district, it's easy to find that this candidate does exist in the word cloud and represents the majority part of repeating times.

Jamaica Plain	Paul Gore Ter	Roxbury Crossing	Davenport St
West Roxbury	March Ter	Financial District / Wharves	Hawley Pl
Fenway / East Fens / Longwood	Westland Ave	Hyde Park	Dell Ter
Beacon Hill	Salt Ln	Roslindale	Chisholm Ter
Charlestown	Cordis Street Ave	West End / Back of the Hill	Lindall Pl
Roxbury / Grove Hall	Beechwood St	Kenmore / Boston University	Overland St
Dorchester / Codman Square	Wilbert Cir	Dorchester / Uphams Corner	Quincefield Pl
North End	Thacher Ct	Markets / Inner Harbor	Fulton St
Back Bay	Warrenton Pl	East Boston	Leverett Ave
Mattapan	Greenfield Rd	Allston	Ashford Ct
South Boston	Woodward Pl	Roxbury	Regent Pl
Chinatown / Tufts-New England Medical Center	Hayward Pl	Brighton	Winship Pl
Dorchester / Fields Corner	Duncan Pl	South End	Pelham Ter
		South Boston / Fort Point	Anchor Way

Candidate per-district

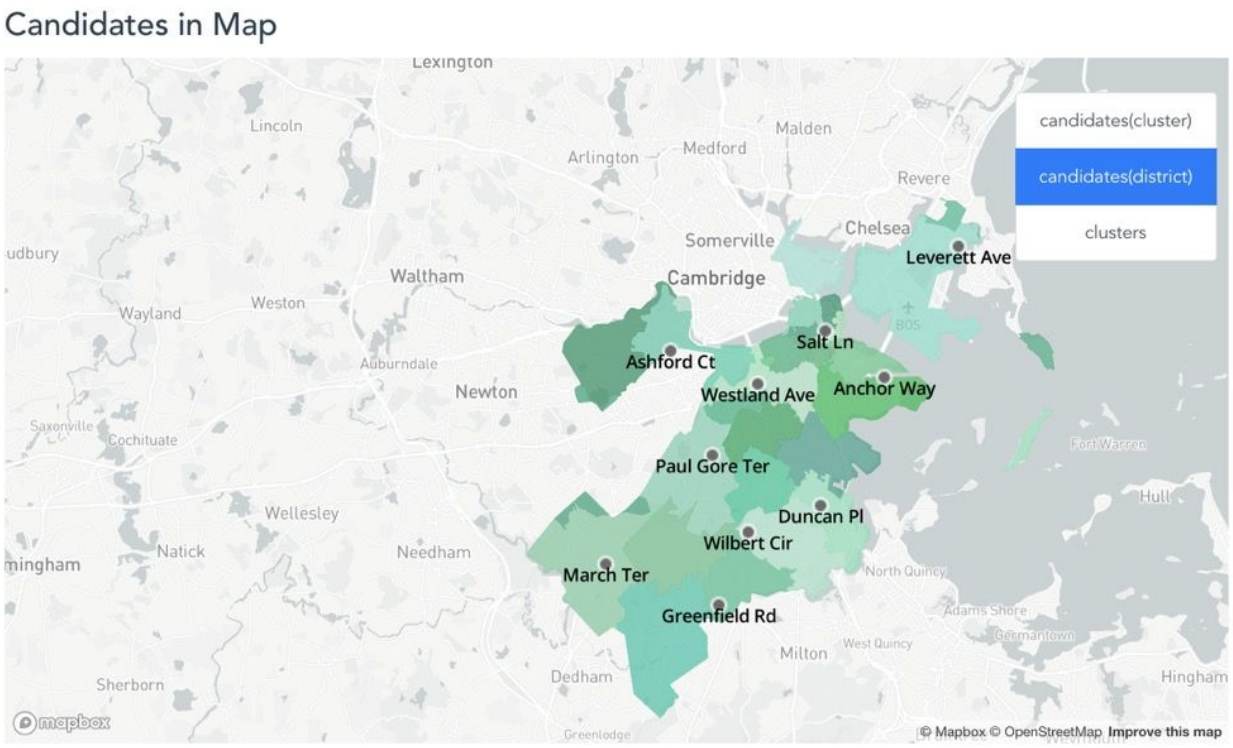
Candidates of West Roxbury

Visualization

We used a webpage to visualize our results. In the webpage, you can see all layers in the map including street clustering, candidates per-district and per-cluster; also we show the word cloud of each district/cluster's street names so that you can have a general idea of the entire program.

As for technical architecture, our Frontend is built with Vue.js framework, which obtains data through RESTful APIs provided by our Flask server.

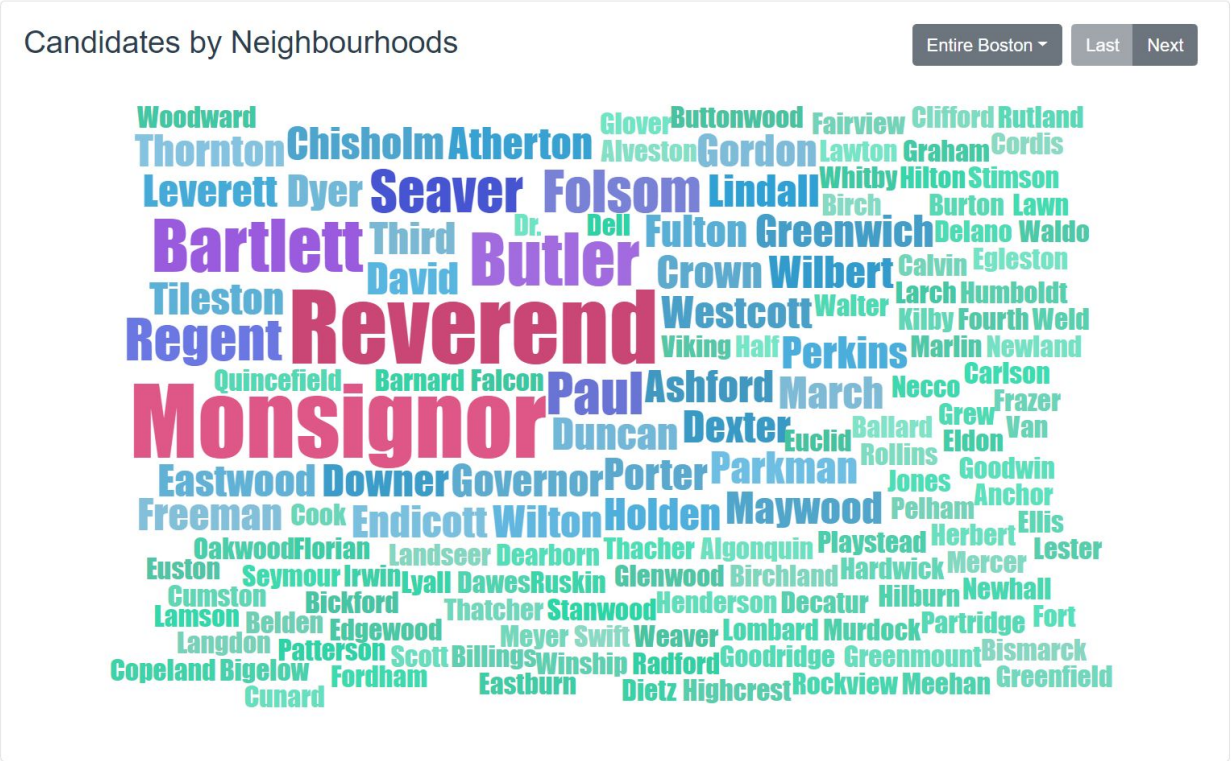
Map:



Map display library is provided by Mapbox using data from OpenStreetMap. We display 4 layers in the map:

- Boston Districts (always shown)
- Candidates Per-District (optional)
- Streets highlighted by K-Means Clusters (optional)
- Candidates Per-Cluster (optional)

Word cloud:



We used the VueWordCloud component to show street names with their sizes by weight. There are multiple word clouds, including word cloud for each district and each cluster. Some buttons is placed on the top-right corner of the card panel for user interaction or control.

Future work

Although we have already provided candidate streets for each district in Boston, we obtain such result dataset based on the descending order of street duplicate times and ascending order of street length. However, we don't set specific weights for these two variables, while one could still introduce other features as measuring principles.
Given the datasets we retrieved for street location, we found that there was still some margin error when GEO data was used in our visualization part. This could be the result of delayed update of geo data and reconstruction of streets in Boston, therefore, one could seek more reliable and effective source to get precise street data.