# Taking Gender Equity to the Streets

Arezoo Sadeghi, Lasely Liu, Xinyun Cao

**project description:**

Having street names for all the streets in Boston area, the goal of the project is to first classify the streets names into 3 categories: male, female and neutral and then find the streets that are the best candidates for renaming.

Prior to coming up with a renaming approach/algorithm, we needed to define some renaming criteria/factors  based on which we can compare different streets with male names and pick the one that best fits our goal for renaming.

In order to do so it was important to first do some primary analysis on the data set and answer a few primary questions including:

1- How many streets are named after males, females, neutrals?

2- Can we find a way to obtain some implicit information from the given data set about the size of the streets?

3- What is the frequency for each name?

4- What are the most repetitive names?

5- Assuming we have a way for inferring some information about the length of the streets, is it really the case that bigger streets are names after male figures?

6 – For top most repetitive names, can we find a way to know how well they are distributed across Boston area?

7- If we were to divide the Boston area into exclusive (having no intersection) sub areas (using some division factor/ clustering factor) , what are the most repetitive names in each sub area?

8 – How does the size of biggest street classified as female compare to the size of the biggest street classified as male?

9- Are the top repetitive names, also repetitive in each sub area?

Our analysis techniques have provided us with the answer to all these questions and we have defined our renaming criteria using the answers to theses questions which is going to be explained in the next sections.

# Data sets

**original data set:**

The only data set that was available to us in the first place was the street book from the *cityofBoston.gov* website in the format of word document file.

This Street book contains **full names, addresses and all the zip codes** for different segments of all Boston streets but it doesn't explicitly contain any information about GEO locations of the streets.

As mentioned before in order for us to answer the above questions , we needed to obtain GEO locations (latitudes and longitudes) ,size of the streets and gender of the street name.

**For getting latitudes and longitudes**:

One way was to use google GEO API, but due to the limitation of 2500 requests per day for using google API, we found it impractical to use it (for more request we had to pay). Instead since we had all the zip codes associated with each street, we used these zip cods to get latitude and longitude using another online document which included all the U.S. street zip codes and their corresponding latitude and longitude.

**For getting size related information**:

we are using the number of all the zip codes associated with the street names. Looking at the data set, it was sometimes the case where a street had for example 15 (either different or the same) zip codes associated with it (for its different segments). we have given each street a **rank** representing their size using the number of zip codes (number of segments) that was there in the data set for each street. We can interpret this rank as the higher this rank the bigger the street size/length.

**For gender classification**:

we are using one of the most popular name ethnicity and Gender classifier APIs called **NamSor**. This API expects both the first and last name to be part of the request string and that's why we had to clean the street fullname strings before passing it to the API.

Looking at the data set we collected a set of post-fix strings that were commonly used in the street full names including, **Street, Road, Boulevard, highway, Square**, **Alley**,...!

------------------------------------------------------------------------------------------------------------

**Updates about cleaning full street names:** After printing our final results into csv files, we observed that for some of the names like “**Washington**”, the frequency that we are getting is not the same as the real frequency of the name And it’s because some instances of streets with having the string “Washington” either start with some post fixes (“**Mt. Washington Way**”) or end with some prefixes that are not included in our list of words to removed and this way “ **Mt. Washington”** would be a different street name than “ **Washington”** When grouping by street names and calculating names frequency. We fixed the problem with washington in the updated version of code but **it is worth mentioning the accuracy of the frequency values for street names depend on the how inclusive and accurate the list of words to be removed is**.

------------------------------------------------------------------------------------------------------

After cleaning the “full names” depending on the number of words in the remaining string, we pass the first word as the firstname and if there exists more than one word, we pass the second word as the lastname otherwise we pass empty string as the last name.

Judging by the results we are getting from NamSor it is safe to say, the order in which  first name and last name are passed to the API doesn't affect the accuracy since NamSor was able to classify strings like **Harvard** or **Washington** correctly as male (in these cases Harvard and Washington were passed as first names), Even though there are some obviously wrong classified names in the data set too including a couple of neutral names classified as females (**Academics** street is classified as female!)

Even though we are trying to increase our accuracy by making use of 2 other APIs, It is worth mentioning that the number of false positives for “male” class is not significant enough to affect our final results and the number of false positives for female is more than the number of false positives for men.

Most of the wrong classified names are related to females or neutral names. Also given the fact that for renaming streets we are mostly dealing with most repetitive male street names  and their corresponding renaming factors which are going to be introduced in the following section we don't need to be worried about NamSor API false positives

------------------------------------------------------------------------------------------------

**Updates:** we are using gender predictor as secondary API for gender classification. Gender predictor only classifies names into "female" and "name" categories and there's no such category as "unknown" or "neutral". As for our final result for gender, we get the intersection of the results from NamSor and gender predictor meaning if for a given name, both of these 2 APIs vote for the same thing (either male or female) that name is gonna get labeled with the output otherwise they're gonna get labeled as "Neutral". This way all the names that are classified as "unknown" by NamSor are still gonna stay neutral and by doing so  the number of neutral names is gonna increase  (we hope) mostly by decreasing the number of false positives for both male and female (and not by decreasing the number of true Positives)

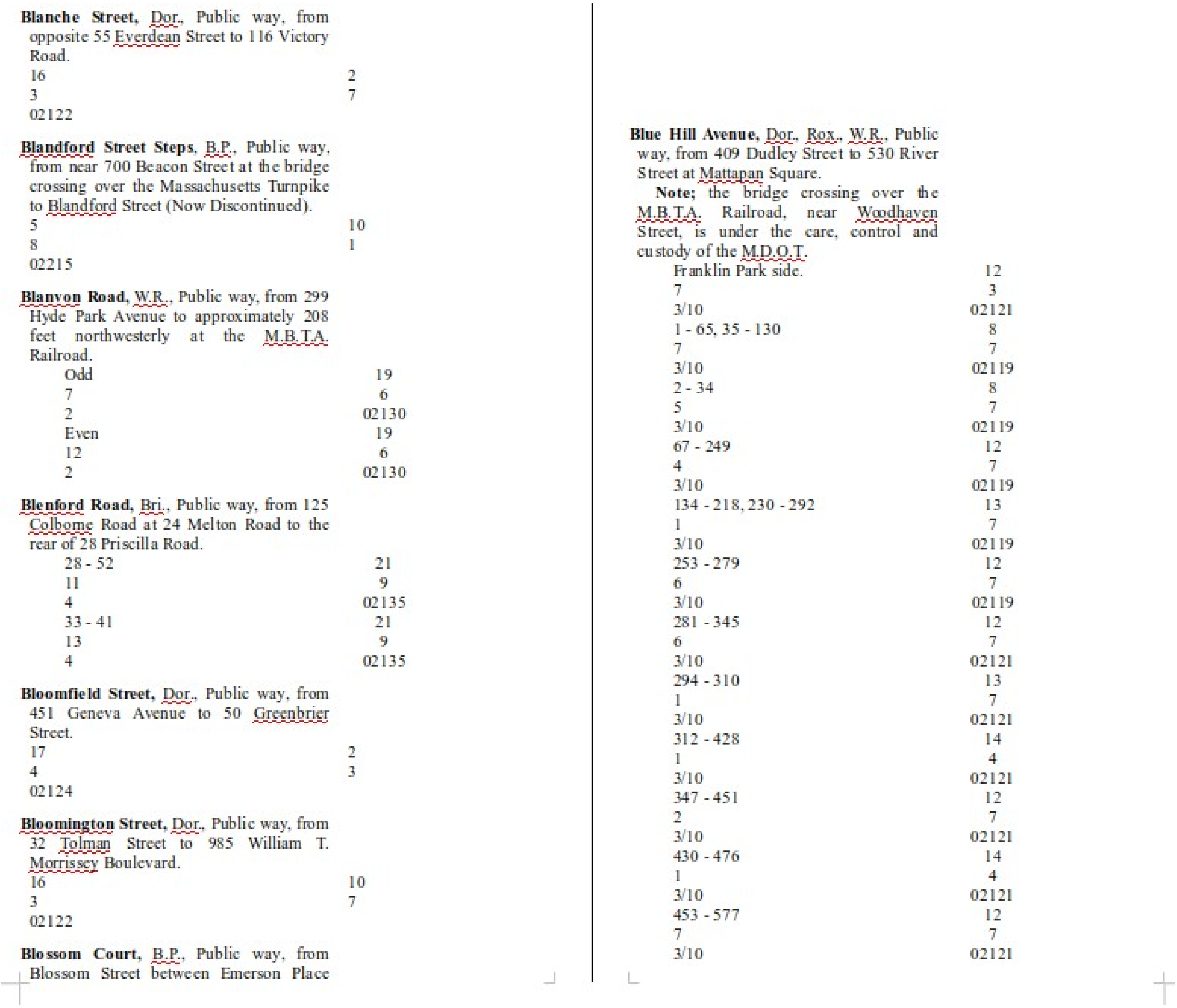
------------------------------------------------------------------------------------------------

Figure 1 is a segment of the original data set for mentioned here for your convenience.

Figure 2 is a segment of our final data frame ready to be analyzed which is sorted using full-name as the key.

Please note that the data frame in figure 2, is already merged with zip code data frame to obtain latitude and longitude and prior to merge, the original data frame obtained from street book document has been processed so that if there is more than one zip code associated with a street, the corresponding row has been separated into new rows congaing each zip code as their value for zip code feature dimension.

So we can't use this data frame for getting the frequency for each name(not all the duplicate records in this data frame count for name duplicate, some of them are there just representing the street segment. For that we have used another data frame before spiting the rows. In general, after getting the main data frame, for doing different types of analysis we have processed the data frame differently. Data frame in figure one is just an example.



**Blanche Street,** Dor., Public way, from opposite 55 Everdean Street to 116 Victory Road.
16 2
3 7
02122

**Blandford Street Steps**, B.P., Public way, from near 700 Beacon Street at the bridge crossing over the Massachusetts Turnpike to Blandford Street (Now Discontinued).
5 10
8 1
02215

**Blanvon Road,** W.R., Public way, from 299 Hyde Park Avenue to approximately 208 feet northwesterly at the M.B.T.A. Railroad.
Odd 19
7 6
2 02130
Even 19
12 6
2 02130

**Blenford Road,** Bri., Public way, from 125 Colbome Road at 24 Melton Road to the rear of 28 Priscilla Road.
28 - 52 21
11 9
4 02135
33 - 41 21
13 9
4 02135

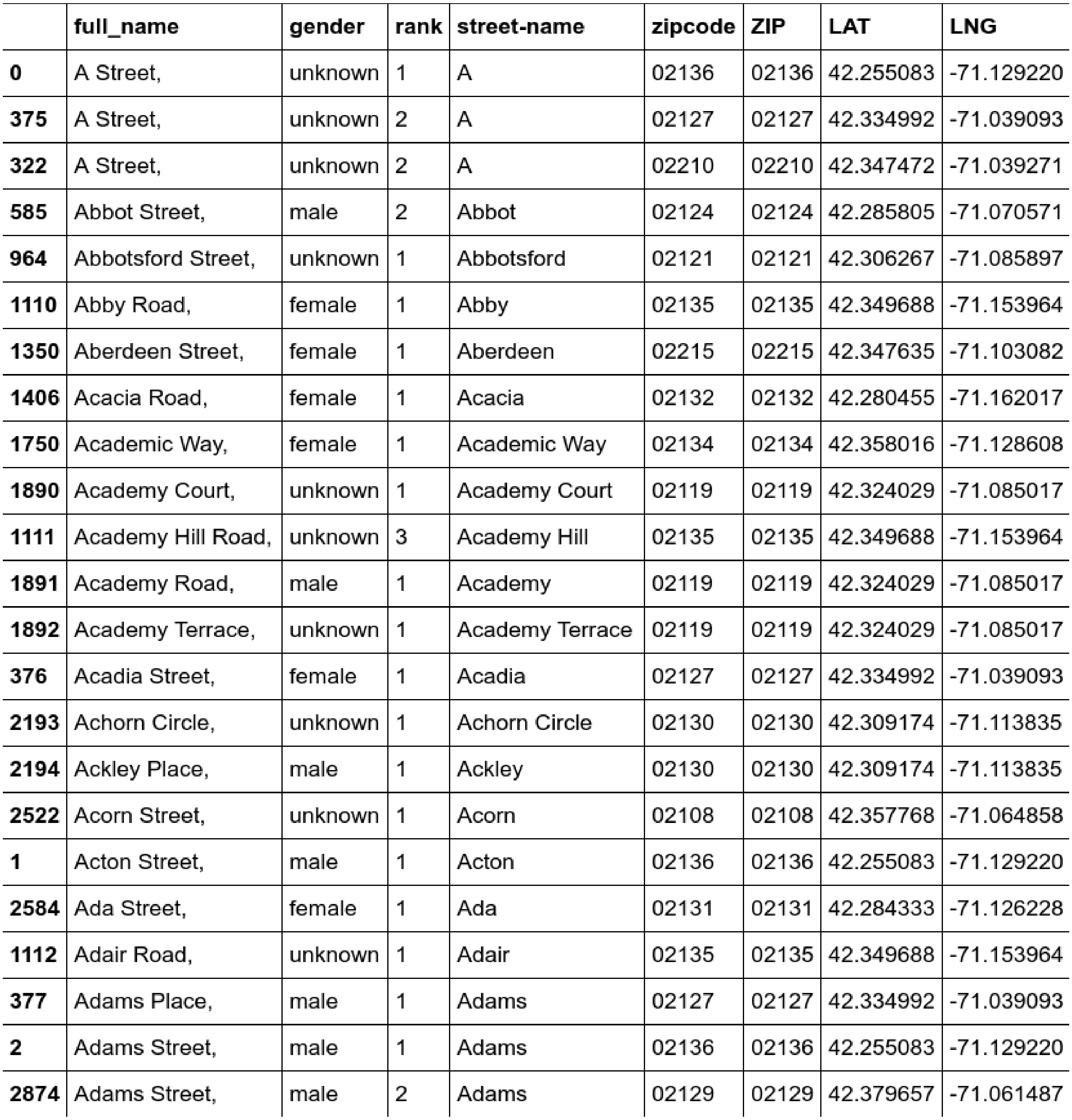
**Bloomfield Street,** Dor., Public way, from 451 Geneva Avenue to 50 Greenbrier Street.
17 2
4 3
02124

**Bloomington Street,** Dor., Public way, from 32 Tolman Street to 985 William T. Morrissey Boulevard.
16 10
3 7
02122

**Blossom Court,** B.P., Public way, from Blossom Street between Emerson Place

**Blue Hill Avenue,** Dor., Rox., W.R., Public way, from 409 Dudley Street to 530 River Street at Mattapan Square.
**Note;** the bridge crossing over the M.B.T.A. Railroad, near Woodhaven Street, is under the care, control and custody of the M.D.O.T.
Franklin Park side. 12
7 3
3/10 02121
1 - 65, 35 - 130 8
7 7
3/10 02119
2 - 34 8
5 7
3/10 02119
67 - 249 12
4 7
3/10 02119
134 - 218, 230 - 292 13
1 7
3/10 02119
253 - 279 12
6 7
3/10 02119
281 - 345 12
6 7
3/10 02121
294 - 310 13
1 7
3/10 02121
312 - 428 14
1 4
3/10 02121
347 - 451 12
2 7
3/10 02121
430 - 476 14
1 4
3/10 02121
453 - 577 12
7 7
3/10 02121

Figure 1: a segment of the street book word document



| | full_name | gender | rank | street-name | zipcode | ZIP | LAT | LNG |
|---|---|---|---|---|---|---|---|---|
| **0** | A Street, | unknown | 1 | A | 02136 | 02136 | 42.255083 | -71.129220 |
| **375** | A Street, | unknown | 2 | A | 02127 | 02127 | 42.334992 | -71.039093 |
| **322** | A Street, | unknown | 2 | A | 02210 | 02210 | 42.347472 | -71.039271 |
| **585** | Abbot Street, | male | 2 | Abbot | 02124 | 02124 | 42.285805 | -71.070571 |
| **964** | Abbotsford Street, | unknown | 1 | Abbotsford | 02121 | 02121 | 42.306267 | -71.085897 |
| **1110** | Abby Road, | female | 1 | Abby | 02135 | 02135 | 42.349688 | -71.153964 |
| **1350** | Aberdeen Street, | female | 1 | Aberdeen | 02215 | 02215 | 42.347635 | -71.103082 |
| **1406** | Acacia Road, | female | 1 | Acacia | 02132 | 02132 | 42.280455 | -71.162017 |
| **1750** | Academic Way, | female | 1 | Academic Way | 02134 | 02134 | 42.358016 | -71.128608 |
| **1890** | Academy Court, | unknown | 1 | Academy Court | 02119 | 02119 | 42.324029 | -71.085017 |
| **1111** | Academy Hill Road, | unknown | 3 | Academy Hill | 02135 | 02135 | 42.349688 | -71.153964 |
| **1891** | Academy Road, | male | 1 | Academy | 02119 | 02119 | 42.324029 | -71.085017 |
| **1892** | Academy Terrace, | unknown | 1 | Academy Terrace | 02119 | 02119 | 42.324029 | -71.085017 |
| **376** | Acadia Street, | female | 1 | Acadia | 02127 | 02127 | 42.334992 | -71.039093 |
| **2193** | Achorn Circle, | unknown | 1 | Achorn Circle | 02130 | 02130 | 42.309174 | -71.113835 |
| **2194** | Ackley Place, | male | 1 | Ackley | 02130 | 02130 | 42.309174 | -71.113835 |
| **2522** | Acorn Street, | unknown | 1 | Acorn | 02108 | 02108 | 42.357768 | -71.064858 |
| **1** | Acton Street, | male | 1 | Acton | 02136 | 02136 | 42.255083 | -71.129220 |
| **2584** | Ada Street, | female | 1 | Ada | 02131 | 02131 | 42.284333 | -71.126228 |
| **1112** | Adair Road, | unknown | 1 | Adair | 02135 | 02135 | 42.349688 | -71.153964 |
| **377** | Adams Place, | male | 1 | Adams | 02127 | 02127 | 42.334992 | -71.039093 |
| **2** | Adams Street, | male | 1 | Adams | 02136 | 02136 | 42.255083 | -71.129220 |
| **2874** | Adams Street, | male | 2 | Adams | 02129 | 02129 | 42.379657 | -71.061487 |

Figure 2: segment of data frame having geo and gender information

**Performance issues parsing the data**:

Even though the data set is not really big, (counting number of distinct street names, there are only about 5000 data entries in the data set), classifying street names was kind of a slow process due to the fact that we had to make API calls to NamSor.

## Renaming criteria and Techniques Used For Renaming Streets

As mentioned before, the main goal of this project is to find best street candidates classified as "male" for renaming and for doing so we need to define some criteria/factors.

The most trivial thing that comes to mind and (of course the one which is mentioned in the project description) is to choose the most repetitive name classified as male. But let's say if two names have the same frequency like (**Harvard** and **Everett** in the data set they both have a frequency of 10) and one of them is well distributed across Boston and the other one is center around a sub-area, it is best to rename the one which is not well distributed .

Having said that, we need to have a way to be able to answer the question of how well a street name is distributed across Boston?

One way to do so is to use Kmeans++ clustering algorithm to cluster the streets based on their GEO locations(latitudes and longitudes) to be able to define **distribution ranks.** This way we can define our renaming criteria As follows and consider a combination of all of them to find the best candidate for renaming:

**1- Global frequency:**

Given a name, frequency of the name in the original data set.

**2- Distribution rank:**

Given a street name and clustering factor, Number of distinct clusters the name appears in. the higher the rank, the better the name is distributed.

**3- Concentration rank:**

Given a street name and clustering factor, maximum of local frequency of that name in each cluster.

**4- Size of the street:**

This factor is implicitly considered in our approach for **some** cases (not all of them) As we have a separate data entry for each street segment with a **distinct** zip code in the data frame mentioned in figure 2. (some times it is the case that a street has multiple segments having the same zip codes, since we care about distinct values for latitude and longitude, we have gotten rid of zip codes with the same value, so this approach doesn’t always reflect the size if the street segments all have the same zip code, so street with a higher distribution rank or concentration rank are not necessarily bigger in size but they could for most of the cases) we can consider size by also considering **ranks** that we talked about in the previous sections. Generally bigger streets tend to have more segment which means they could possibly have higher either distribution rank or concentration values

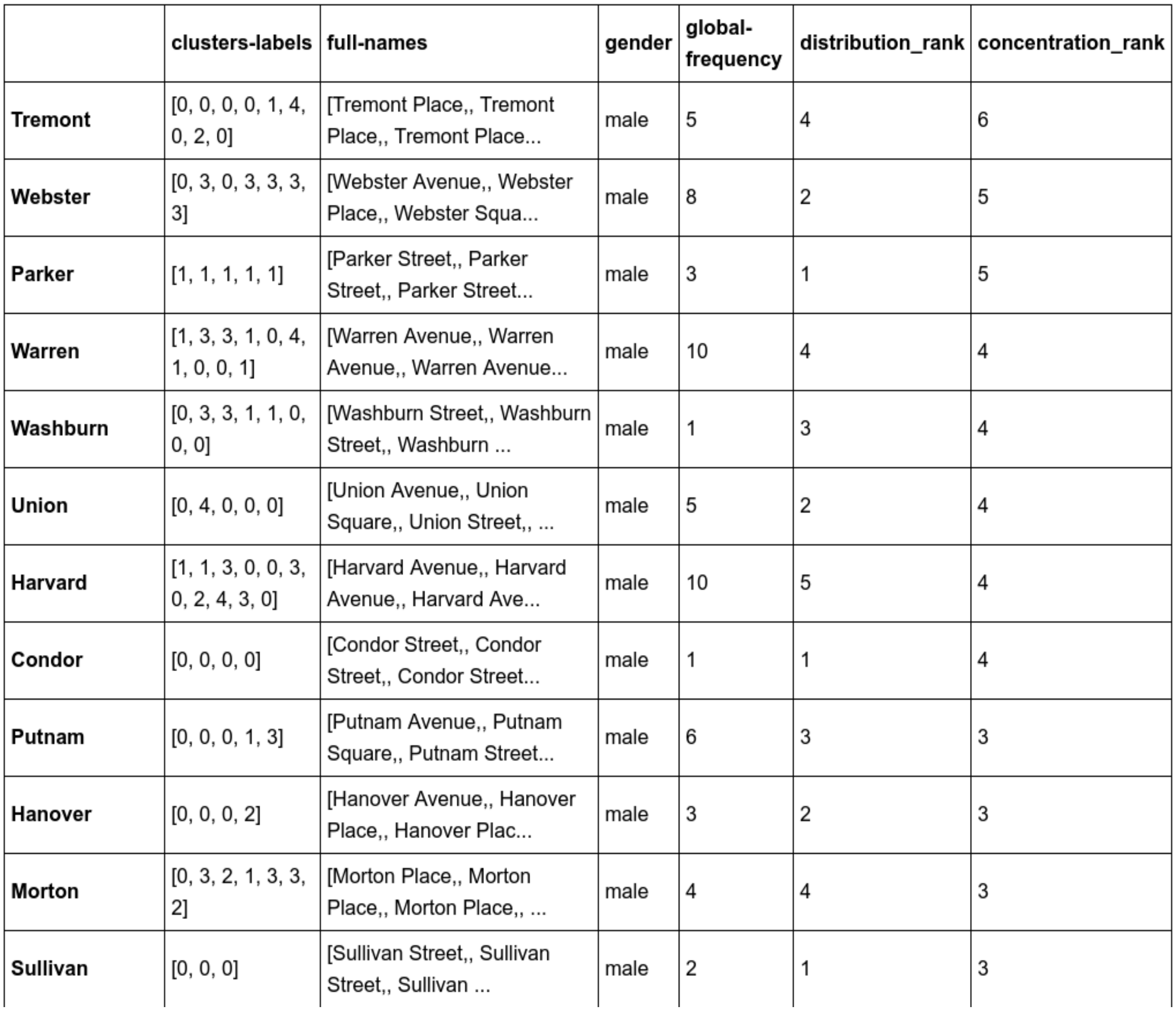
**Limitations of  this approach for getting distribution and concentration rank:**

There are 2 important things to take into account about this approach. Using clustering algorithms **we cannot have overlaps in our clusters** and the **distribution rank and concentration rank both highly depend on the clustering factor/number of clusters**.

Max value of distribution rank is the same as number of clusters.

Figure 3 is a segment of data frame obtained  after clustering male streets based on their GEO locations.

Figure 3 : Data frame including renaming factors



| | clusters-labels | full-names | gender | global-frequency | distribution_rank | concentration_rank |
|---|---|---|---|---|---|---|
| **Tremont** | [0, 0, 0, 0, 1, 4, 0, 2, 0] | [Tremont Place,, Tremont Place,, Tremont Place... | male | 5 | 4 | 6 |
| **Webster** | [0, 3, 0, 3, 3, 3, 3] | [Webster Avenue,, Webster Place,, Webster Squa... | male | 8 | 2 | 5 |
| **Parker** | [1, 1, 1, 1, 1] | [Parker Street,, Parker Street,, Parker Street... | male | 3 | 1 | 5 |
| **Warren** | [1, 3, 3, 1, 0, 4, 1, 0, 0, 1] | [Warren Avenue,, Warren Avenue,, Warren Avenue... | male | 10 | 4 | 4 |
| **Washburn** | [0, 3, 3, 1, 1, 0, 0, 0] | [Washburn Street,, Washburn Street,, Washburn ... | male | 1 | 3 | 4 |
| **Union** | [0, 4, 0, 0, 0] | [Union Avenue,, Union Square,, Union Street,, ... | male | 5 | 2 | 4 |
| **Harvard** | [1, 1, 3, 0, 0, 3, 0, 2, 4, 3, 0] | [Harvard Avenue,, Harvard Avenue,, Harvard Ave... | male | 10 | 5 | 4 |
| **Condor** | [0, 0, 0, 0] | [Condor Street,, Condor Street,, Condor Street... | male | 1 | 1 | 4 |
| **Putnam** | [0, 0, 0, 1, 3] | [Putnam Avenue,, Putnam Square,, Putnam Street... | male | 6 | 3 | 3 |
| **Hanover** | [0, 0, 0, 2] | [Hanover Avenue,, Hanover Place,, Hanover Plac... | male | 3 | 2 | 3 |
| **Morton** | [0, 3, 2, 1, 3, 3, 2] | [Morton Place,, Morton Place,, Morton Place,, ... | male | 4 | 4 | 3 |
| **Sullivan** | [0, 0, 0] | [Sullivan Street,, Sullivan Street,, Sullivan ... | male | 2 | 1 | 3 |

## Choosing clustering factor:

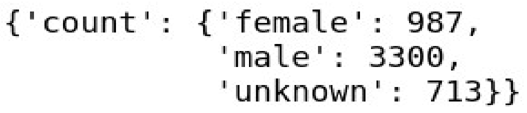
If number of clusters is too big it would lead to high distribution rank values and in the worse case all the streets with a given name will have a distinct clustering label which means, they are all well distributed and value of the distribution rank is going to be the same as the value of their global frequency. So if the number of clusters is too high, effect of distribution rank is going to be the same only considering global frequency and size of street.

On the other hand if number of clusters is too low, concentration rank is going to be really high for most of the street with high global frequency value and bigger size.

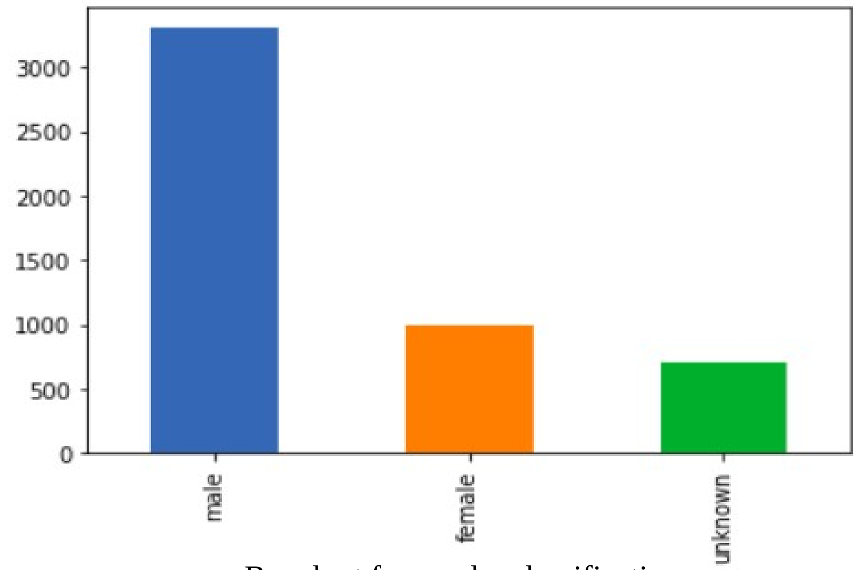
So coming up with reasonable clustering factor is really important for this approach. Running some experiments on our data we concluded that setting our clustering factor to  half of  the max global frequency works the best. **This conclusion is though limited to the experiment we ran.**

**Finally how to decide which street to rename**:
Depending on the number of streets we want to rename and considering the values of renaming factors in the data frame of figure 3, **one way is to get the most repeated names in each cluster (the ones with higher concentration ranks) and comparing the frequency and distribution rank of them to those of top most repetitive name.** Looking our results "**Tremont**" is good candidate to be renamed. It has the highest concentration rank even though its global frequency is 5 which is half of the max value of global frequency and it also has a reasonable distribution rank of 4 considering our clustering factor Which means that even though its not one of the top most repetitive names compared to "**Harvard**" which has global frequency of 10, in certain sub areas there are too many street names Tremont.

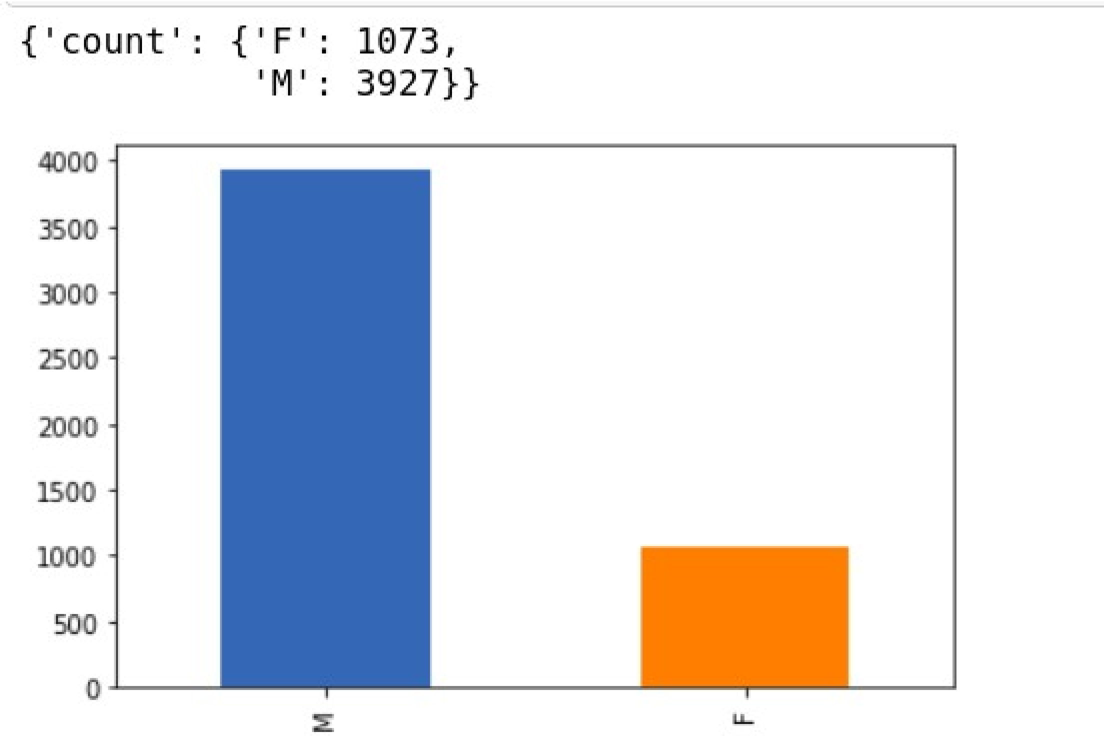
## Results and visualizations



```
{'count': {'female': 987,
           'male': 3300,
           'unknown': 713}}
```

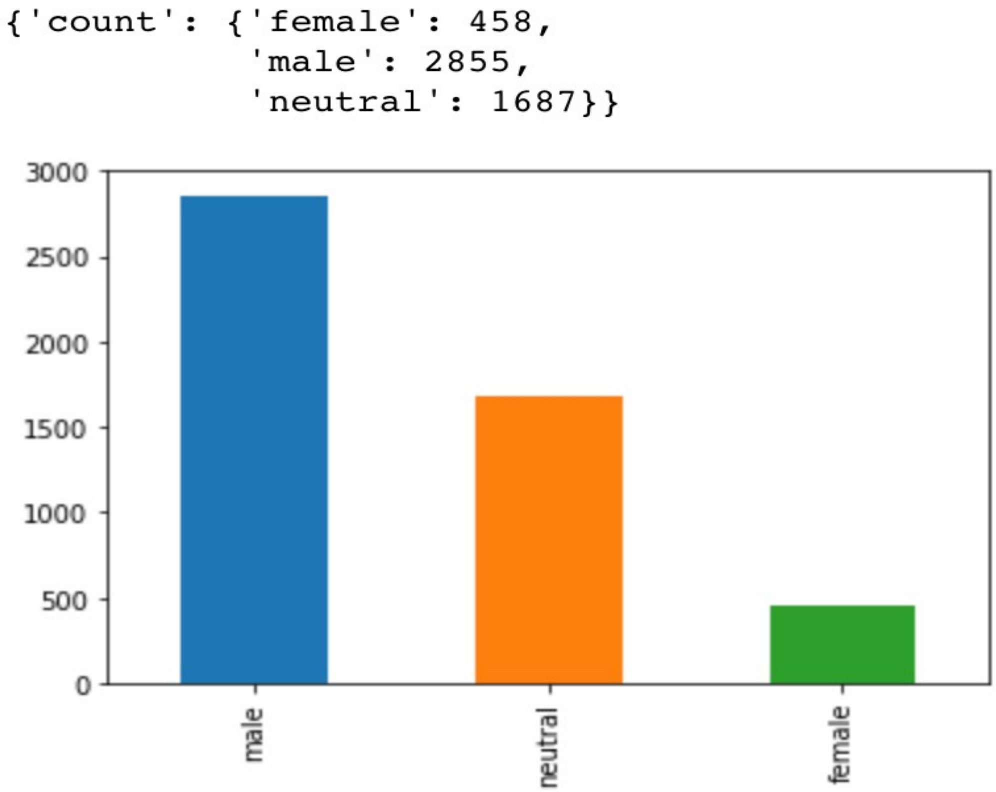


Bar chart for gender classification
Note: unknown means neutral



```
{'count': {'F': 1073,
           'M': 3927}}
```

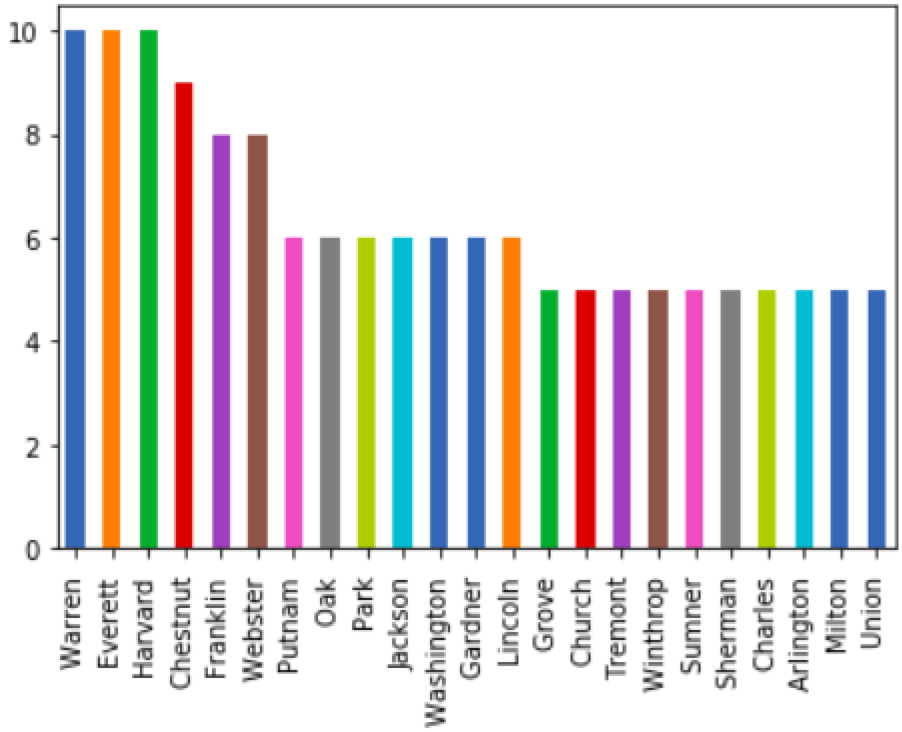
Bar chart for gender classification
( using gender predictor for gender classification)

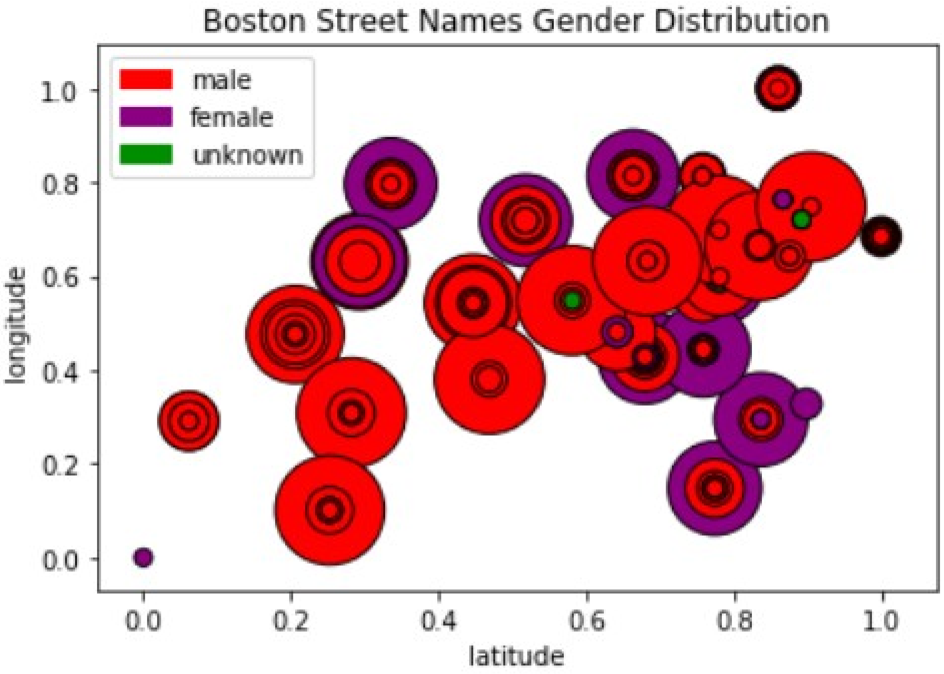


```
{'count': {'female': 458,
           'male': 2855,
           'neutral': 1687}}
```

Bar chart for gender classification
Getting the intersection of results from NamSor and gender predictor

Bar chart for frequency of top most repetitive street name classified as male

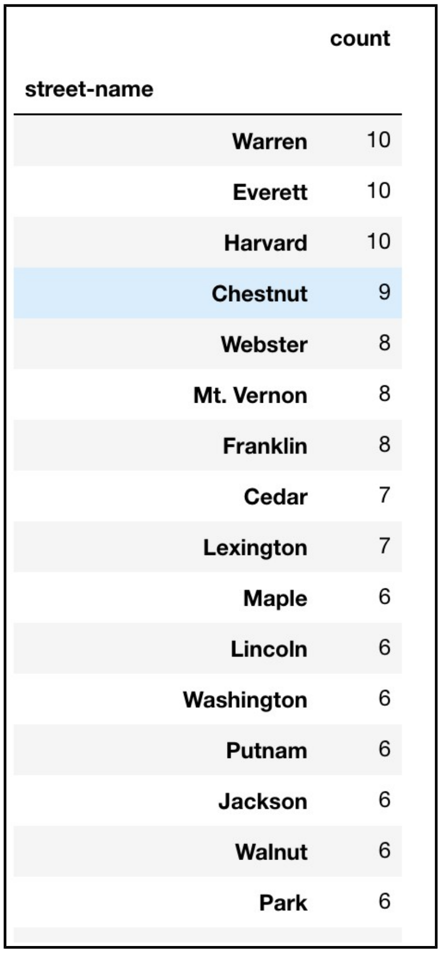




Size of the circles correspond to the size of streets.

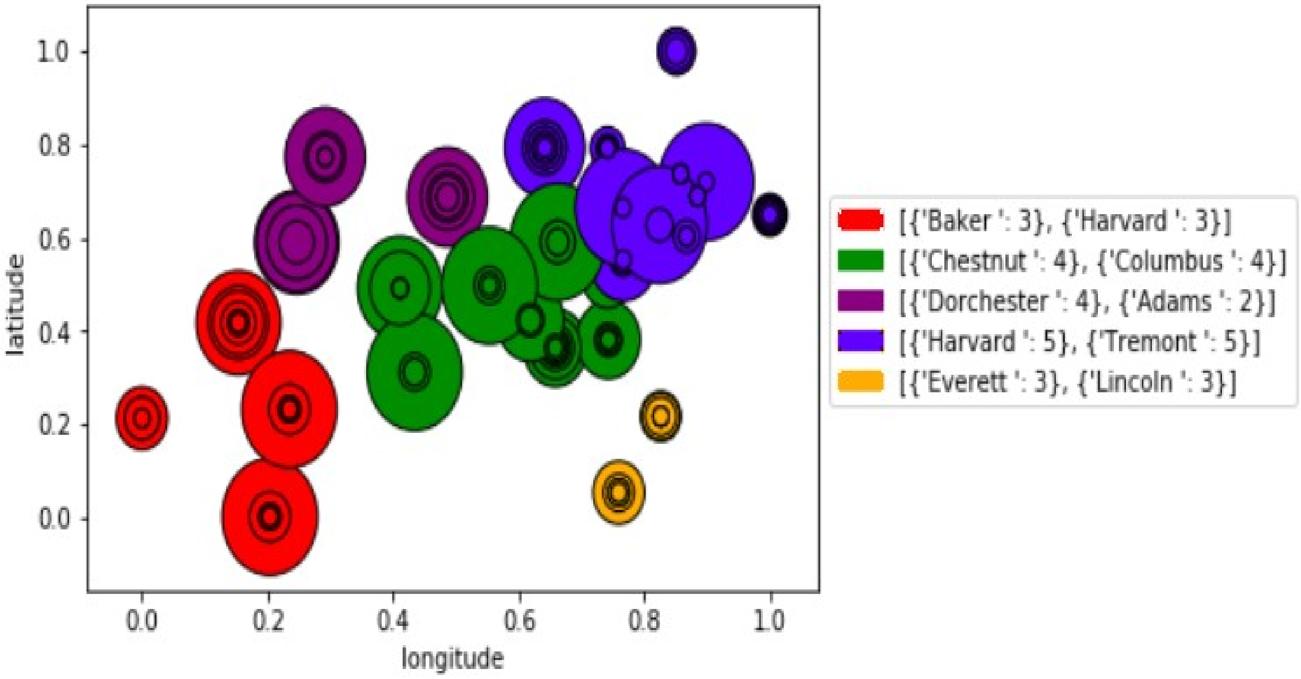
Looking at this we can infer most of bigger street are classified as male.

Note there are so many overlaps and there might be even circles behind these given the fact the latitude and longitude are obtained from zip code and some streets with different name share the same zip code.



| street-name | count |
|---|---|
| Warren | 10 |
| Everett | 10 |
| Harvard | 10 |
| Chestnut | 9 |
| Webster | 8 |
| Mt. Vernon | 8 |
| Franklin | 8 |
| Cedar | 7 |
| Lexington | 7 |
| Maple | 6 |
| Lincoln | 6 |
| Washington | 6 |
| Putnam | 6 |
| Jackson | 6 |
| Walnut | 6 |
| Park | 6 |

Segment of data frame for name frequency



Clustering street names based on GEO locations into 5 clusters
Legend shows top two most repetitive name in each cluster

## Future work and observation:

Renaming streets really depend on what criteria we choose for renaming streets and what matters the most depending on the reason why we want to do this. For example one can give some factors like frequency or size higher importance and in this case distribution ranks don’t really matter as much as those 2 other factors.

Once can  try to use other approaches like K nearest neighbors . Again choice of k is important but this way we have solved the problem of having no overlaps. We can pick random points and count the number of male voters for each point and pick the one that is significantly higher than others and compare our results to the results obtained using clustering approach.