

Recommending location for a new residence

Sergei Merson

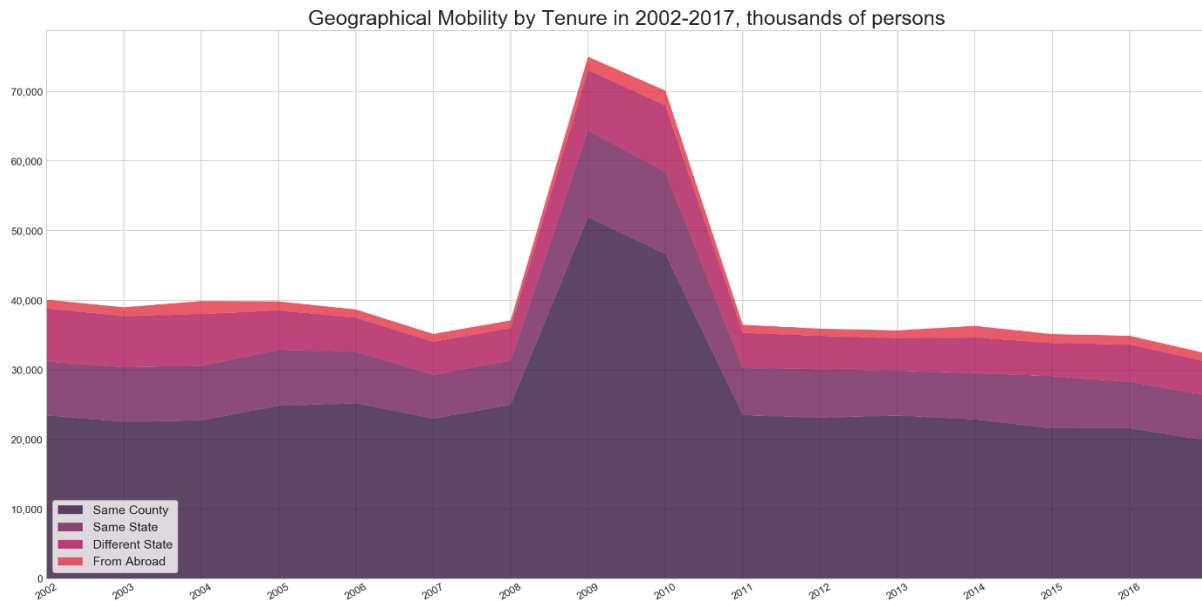
July 09, 2019

1. Introduction

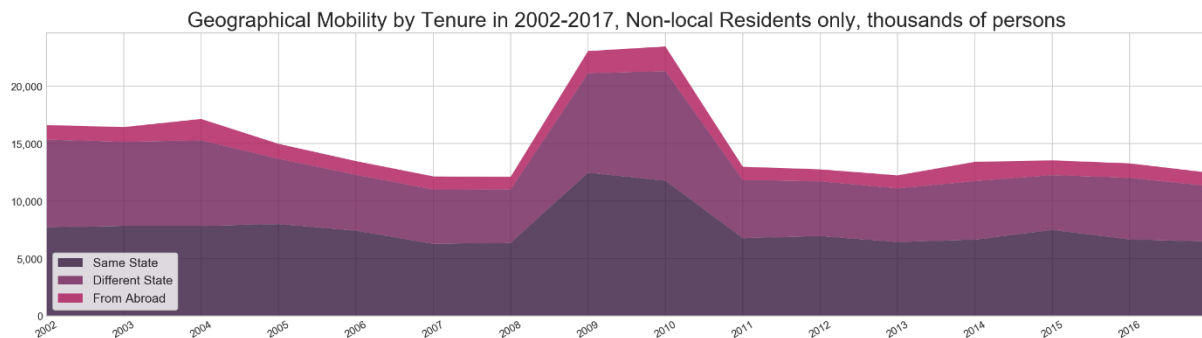
1.1 Background

Every year millions of families are changing their residence location. According to the data provided by the United States Census Bureau, more than 30 million tenures have changed their residents.

Below you can see the graph that shows the dynamic of geographical mobility in the United States:



And here the same data but without resident change within the same County:



1.2 Problem

In many cases when a person must move to another city or even country, she knows nothing about the new location, and it becomes extremely difficult to choose optimal place to search for a new home/flat.

The system developed in this project hopes to help to solve this problem by providing clear description of possible locations according to the person's preferences.

1.3 Target audience

Families which move to another region, real estate agencies, companies that often need to relocate their employees, municipal authorities (that could make simulations and determine what tracts and regions require actions in order to make them attractive to people).

1.4 Positive impact

Right choice of residential location could benefit in many ways:

- Improves total quality of life of new coming residents;
- Reduces transportation costs and efforts, because most of the venues of interest should be not far from home;
- Residents that are more satisfied with their location less tend to change it in the future;
- Stimulates region development in order to satisfy demands and priorities of the citizens etc.

2. Data acquisition and cleaning

2.1 Data sources

The main dataset with statistics of census tracts in New York area could be found on BetaNYC portal (data.beta.nyc). From here we will download three files: csv file with the data, .geojson file with census tract shapes (to plot them on the map) and a .txt description file.

Second source – Foursquare API that would provide us data about different venues in each region.

Last one is imaginary venue preferences of a made-up family that must relocate from Torrington to New York.

2.2 Data cleaning

The Data taken from BetaNYC portal is of very high quality: there is no missing entries, invalid data types in the dataset or duplicated records. The only cleaning that was made is that I dropped records with zero costs because it is unclear if I'm dealing with some special regions with subsidized costs or this is just a typo.

Foursquare API provides results in form of json file with predefined structure, so here too data comes in a relatively clean form.

2.3 Feature selection

The BetaNYC dataset provides information about 27 features of each census tract in New York area, so I kept only information that could be relevant to a person looking for a new home, such as median annual costs by category (taxes, rent, transportation and energy alongside with geolocation. Below you can see first five records on this dataframe:

| ID | County | State | Neighborhood | Latitude | Longitude | Rent | Taxes | Energy | Transportation | Total Costs |
|-------------|---------------|------------|--------------------|-----------|------------|-------|-------------|-------------|----------------|--------------|
| 34017032400 | Hudson County | New Jersey | West New York Town | 40.792844 | -74.013482 | 11532 | 5007.444405 | 1773.280152 | 4145.566539 | 22458.291096 |
| 34017010100 | Hudson County | New Jersey | Bayonne City | 40.691559 | -74.110913 | 10968 | 6146.888195 | 1876.844806 | 5969.286255 | 24961.019256 |
| 34017010200 | Hudson County | New Jersey | Bayonne City | 40.682103 | -74.104573 | 11076 | 6193.436421 | 2167.175106 | 6014.489564 | 25451.101091 |
| 34017010300 | Hudson County | New Jersey | Bayonne City | 40.672439 | -74.081016 | 9792 | 4941.564294 | 2163.964082 | 4798.787758 | 21696.316134 |
| 34017010400 | Hudson County | New Jersey | Bayonne City | 40.670599 | -74.089940 | 10440 | 7838.446154 | 2114.480847 | 7611.970058 | 28004.897059 |

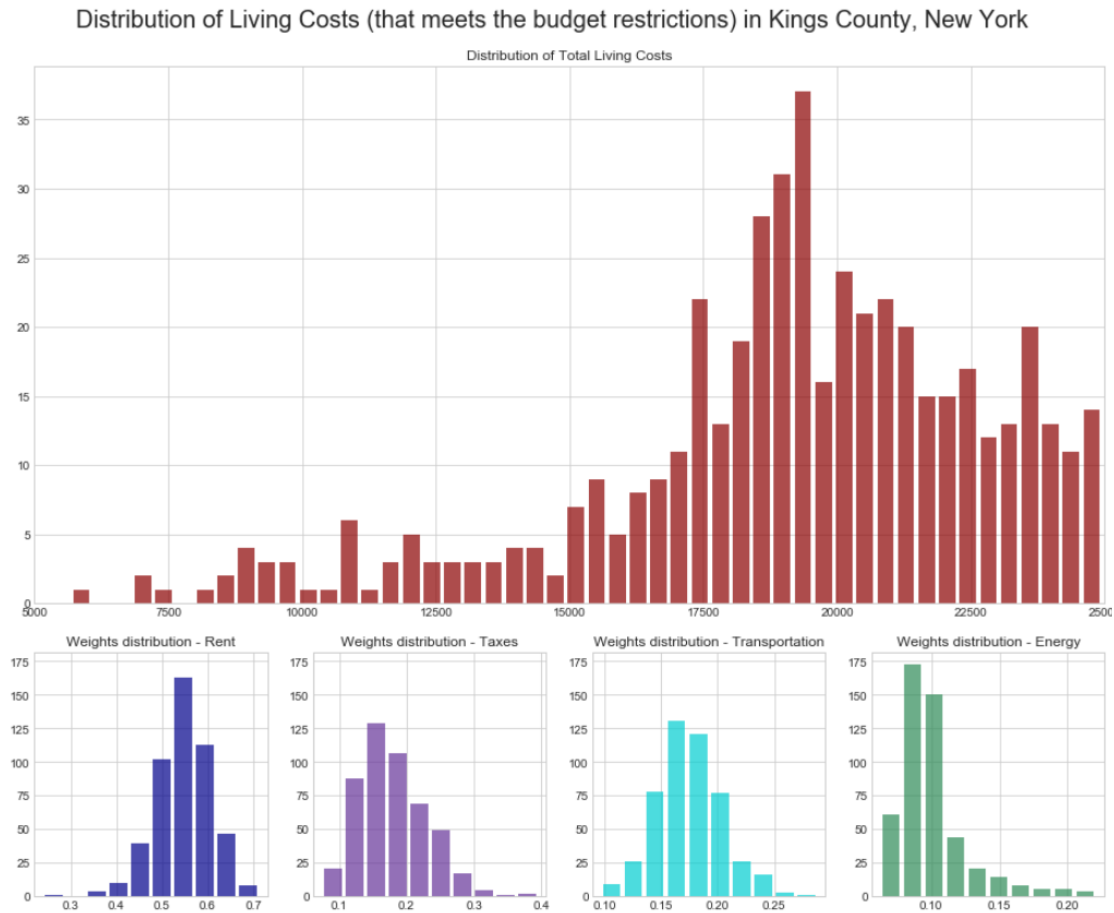
Foursquare API has two types of request, "Regular Calls" and "Premium Calls". Free account has a limitation of 950 regular calls and 50 premium calls per day. In this project I have 488 locations (census tracts) for analysis, so I decided to extract only information about venue names, types and geolocation information:

| | ID | Venue | Latitude | Longitude | Subcategory |
|---|-------------|------------------|-----------|------------|-------------------|
| 0 | 36047031500 | Café Cotton Bean | 40.676133 | -73.950327 | Coffee Shop |
| 1 | 36047031500 | India House | 40.678708 | -73.949651 | Indian Restaurant |
| 2 | 36047031500 | King Tai | 40.676088 | -73.949688 | Cocktail Bar |

3. Exploratory Data Analysis

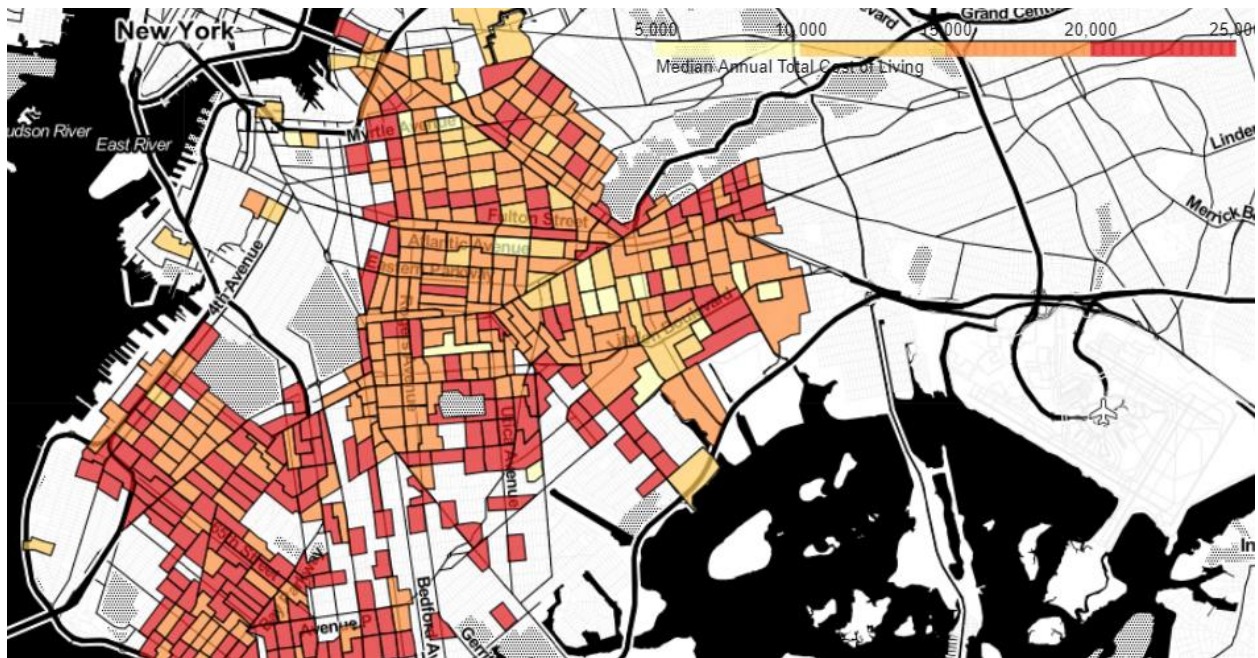
3.1 Census tracts characteristics

After I subset only records that satisfied budget restriction of the target “Customer”, I grouped the locations by total annual cost of living in each tract, and analyzed cost structure by expenditure category:



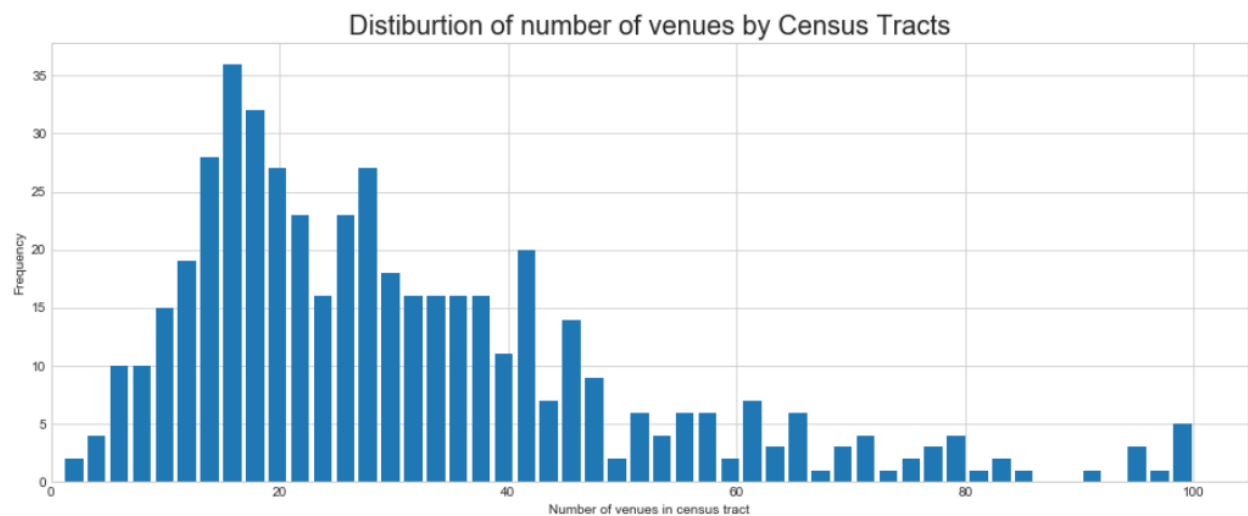
Most of the regions (census tracts) have annual living expenses from \$15,000 to \$25,000 (maximum affordable by budget limitation). In most cases the main expenditure is Rent expenses, that account for more than 50% of total costs.

I also plotted the choropleth map of Kings County, where each tract was colored according with its total annual cost of living:



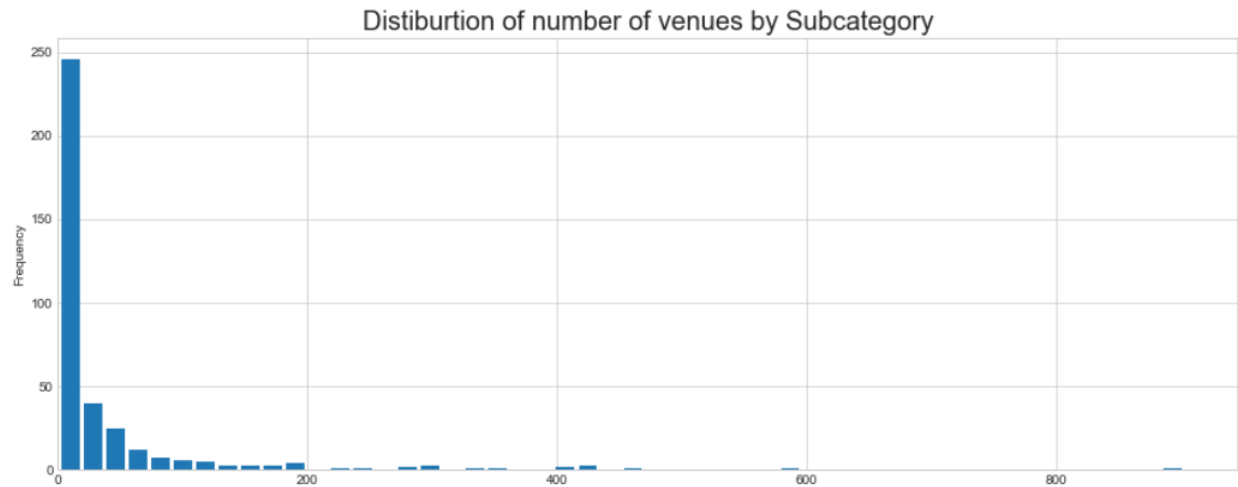
3.2 Venues characteristics

First, I decided to check how venues are distributed by tracts:

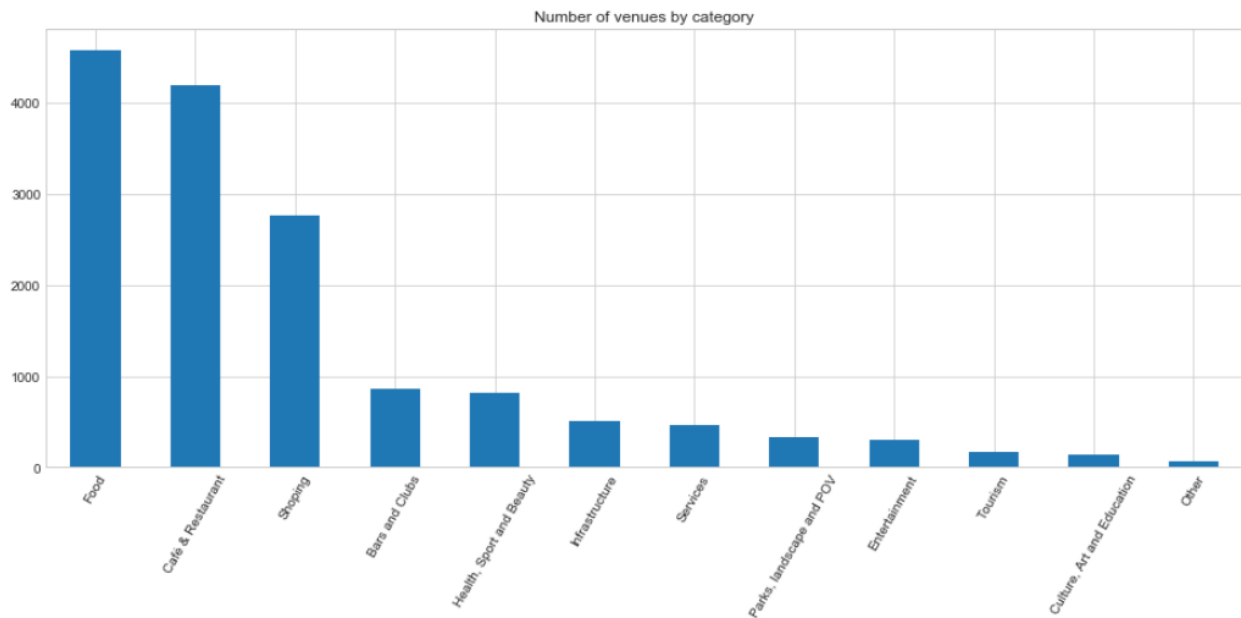


The result seems to be reasonable: main part of locations has about 10-40 venues of different types, alongside with some extremely populated regions with 50-100 venues in each.

Next step – check how often appears each venue category:



Most of the labels appear in the dataset only several times, so I decided to group them into more general categories:



Now it looks much better and suitable for further analysis.

4. Recommendation system

4.1 Choosing recommendation system type

There are two main system types that were explored during the specialization: content-base system and collaborative-filtering system.

In my case we have no information about the similarity between different users, but have data about each region characteristics, “content”, so I chose to implement the system of this type.

Also, I’ve simplified the step of extraction of user preferences: instead of asking her to rate some regions/location that she already knows and then calculate the weights of each characteristic (feature), I’ve “asked” directly what venue types are desirable and what are unwanted. As the result I’ve got such a table:

| | Score |
|--------------------------|-------|
| Café & Restaurant | 6 |
| Infrastructure | 8 |
| Parks, landscape and POV | 6 |
| Pet Care | 5 |
| Health, Sport and Beauty | 8 |
| Bars and Clubs | -8 |
| Tourism | -7 |

4.2 Data aggregation and transformation

Data from BetaNYC portal and data acquired from Foursquare were combined into one table and then transformed into the suitable form, where each column was a venue type, and each row – single census tract.


Instead of one-hot encoding (where we get 0/1 values), we will use pivot table method (because in our case it's reasonable to preserve the number of venues of each category):

| Category | Café & Restaurant | Infrastructure | Parks, landscape and POV | Pet Care | Health, Sport and Beauty | Bars and Clubs | Tourism |
|-------------|-------------------|----------------|--------------------------|----------|--------------------------|----------------|---------|
| ID | | | | | | | |
| 36047000200 | 14 | 0 | 0 | 1 | 1 | 3 | 1 |
| 36047002000 | 4 | 2 | 0 | 0 | 2 | 3 | 0 |
| 36047002200 | 3 | 1 | 1 | 0 | 0 | 1 | 0 |

4.3 Score calculation

Now when I have all the required information in the proper form, I need to multiply these two datasets to get each tract score value:

| Score | Category | Café & Restaurant | Infrastructure | Parks, landscape and POV | Pet Care | Health, Sport and Beauty | Bars and Clubs | Tourism |
|----------------------------|-------------|-------------------|----------------|--------------------------|----------|--------------------------|----------------|---------|
| Café & Restaurant 6 | ID | | | | | | | |
| Infrastructure 8 | 36047000200 | 14 | 0 | | 0 | 1 | 1 | 3 |
| Parks, landscape and POV 6 | 36047002000 | 4 | 2 | | 0 | 0 | 2 | 3 |
| Pet Care 5 | 36047002200 | 3 | 1 | | 1 | 0 | 0 | 1 |
| Health, Sport and Beauty 8 | 36047002300 | 12 | 1 | | 8 | 0 | 9 | 6 |
| Bars and Clubs -8 | 36047002901 | 4 | 2 | | 4 | 0 | 8 | 2 |
| Tourism -7 | | | | | | | | |



| Score | Café & Restaurant | Infrastructure | Parks, landscape and POV | Pet Care | Health, Sport and Beauty | Bars and Clubs | Tourism | Total Costs | Latitude | Longitude |
|-------------|-------------------|----------------|--------------------------|----------|--------------------------|----------------|---------|-------------|--------------|----------------------|
| ID | | | | | | | | | | |
| 36047010600 | 214 | 33 | 0 | 0 | 0 | 3 | 1 | 0 | 18123.882368 | 40.638875 -74.006040 |
| 36047079801 | 199 | 26 | 1 | 1 | 1 | 7 | 4 | 0 | 18803.643250 | 40.660011 -73.958719 |
| 36047029000 | 180 | 26 | 1 | 0 | 0 | 2 | 0 | 0 | 20452.762195 | 40.603174 -73.994348 |
| 36047032700 | 171 | 23 | 1 | 2 | 1 | 5 | 4 | 0 | 17782.949692 | 40.662647 -73.958796 |

To better understand the results, I normalized the Scores to scale from 0 to 100, where score 0 would mean that this location is the worst one for a target customer, and score of 100 - the most desirable place to live:

| Score | Café & Restaurant | Infrastructure | Parks, landscape and POV | Pet Care | Health, Sport and Beauty | Bars and Clubs | Tourism | Total Costs | Latitude | Longitude |
|-------------|-------------------|----------------|--------------------------|----------|--------------------------|----------------|---------|-------------|--------------|----------------------|
| ID | | | | | | | | | | |
| 36047010600 | 100.00 | 33 | 0 | 0 | 0 | 3 | 1 | 0 | 18123.882368 | 40.638875 -74.006040 |
| 36047079801 | 93.88 | 26 | 1 | 1 | 1 | 7 | 4 | 0 | 18803.643250 | 40.660011 -73.958719 |
| 36047029000 | 86.12 | 26 | 1 | 0 | 0 | 2 | 0 | 0 | 20452.762195 | 40.603174 -73.994348 |
| 36047032700 | 82.45 | 23 | 1 | 2 | 1 | 5 | 4 | 0 | 17782.949692 | 40.662647 -73.958796 |
| 36047079601 | 79.18 | 28 | 0 | 1 | 1 | 3 | 5 | 0 | 19334.647301 | 40.656596 -73.958272 |
| 36047123700 | 10.20 | 2 | 3 | 2 | 0 | 0 | 5 | 2 | 23768.295222 | 40.696738 -73.956740 |
| 36047053100 | 9.80 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 21838.451249 | 40.700730 -73.954175 |
| 36047083600 | 8.98 | 0 | 0 | 1 | 0 | 1 | 2 | 1 | 23724.142655 | 40.637103 -73.935127 |
| 36047040900 | 1.22 | 1 | 0 | 1 | 0 | 2 | 7 | 0 | 19337.181729 | 40.691042 -73.904645 |
| 36047039300 | 0.00 | 12 | 0 | 0 | 0 | 1 | 13 | 1 | 20354.618980 | 40.695628 -73.930005 |

5. Results

I decided to check two possible scoring options: Absolute score and Optimal score/cost value.

5.1 Absolute score

If client want to find the best locations possible (within budget limitations of course), I would recommend her the best N locations with the highest score values:

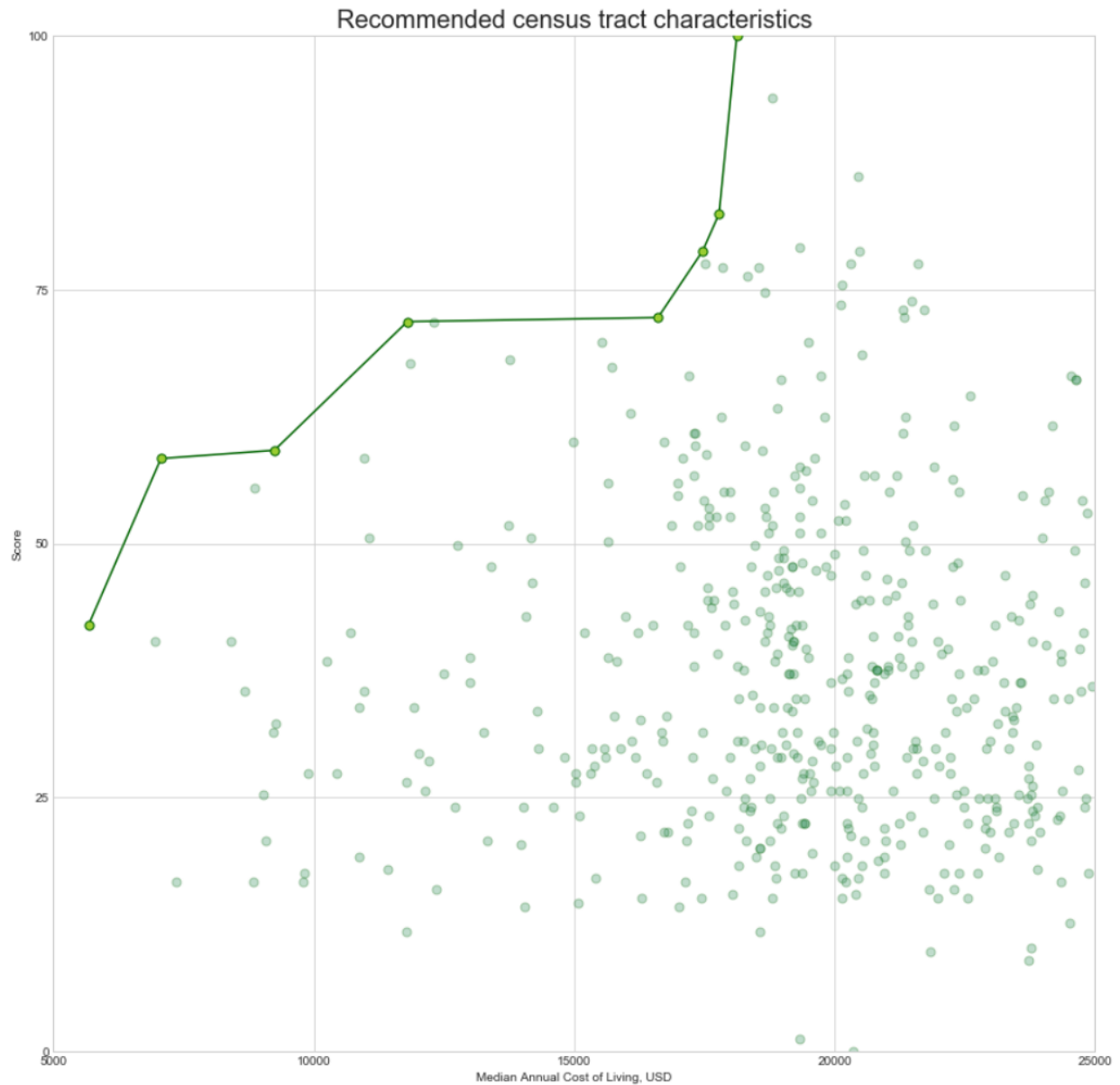
| ID | Score | Café & Restaurant | Infrastructure | Parks, landscape and POV | Pet Care | Health, Sport and Beauty | Bars and Clubs | Tourism | Total Costs | Latitude | Longitude |
|-------------|--------|-------------------|----------------|--------------------------|----------|--------------------------|----------------|---------|--------------|-----------|------------|
| 36047010600 | 100.00 | 33 | 0 | 0 | 0 | 3 | 1 | 0 | 18123.882368 | 40.638875 | -74.006040 |
| 36047079801 | 93.88 | 26 | 1 | 1 | 1 | 7 | 4 | 0 | 18803.643250 | 40.660011 | -73.958719 |
| 36047029000 | 86.12 | 26 | 1 | 0 | 0 | 2 | 0 | 0 | 20452.762195 | 40.603174 | -73.994348 |
| 36047032700 | 82.45 | 23 | 1 | 2 | 1 | 5 | 4 | 0 | 17782.949692 | 40.662647 | -73.958796 |
| 36047079601 | 79.18 | 28 | 0 | 1 | 1 | 3 | 5 | 0 | 19334.647301 | 40.656596 | -73.958272 |

5.2 Optimal score/cost value

The main problem of the first approach is that it doesn't consider the cost of score improvement relative to other existing options. For example, if there exist two locations A and B with score values equal to 100 and 99 relatively, the system would recommend location A, even if the annual total cost of living there is ten times higher than in location B.

To deal with this problem, I selected optimal locations from the dataset, that for the same price or lower provide better score than other options:

| ID | Score | Café & Restaurant | Infrastructure | Parks, landscape and POV | Pet Care | Health, Sport and Beauty | Bars and Clubs | Tourism | Total Costs | Latitude | Longitude |
|-------------|-------|-------------------|----------------|--------------------------|----------|--------------------------|----------------|---------|-------------|----------|-----------|
| 36047035200 | 42.04 | 3 | 0 | 5 | 0 | 5 | 2 | 0 | 5679.75 | 40.5721 | -73.9825 |
| 36047002901 | 58.37 | 4 | 2 | 4 | 0 | 8 | 2 | 0 | 7066.95 | 40.6949 | -73.9781 |
| 36047121000 | 59.18 | 5 | 5 | 2 | 0 | 4 | 0 | 0 | 9244.01 | 40.6701 | -73.8722 |
| 36047002300 | 71.84 | 12 | 1 | 8 | 0 | 9 | 6 | 1 | 11797.5 | 40.6996 | -73.9833 |
| 36047052300 | 72.24 | 34 | 0 | 0 | 1 | 9 | 16 | 1 | 16610.4 | 40.7117 | -73.9588 |
| 36047010400 | 78.78 | 27 | 0 | 0 | 0 | 2 | 2 | 0 | 17467.2 | 40.6366 | -74.0084 |
| 36047032700 | 82.45 | 23 | 1 | 2 | 1 | 5 | 4 | 0 | 17782.9 | 40.6626 | -73.9588 |
| 36047010600 | 100 | 33 | 0 | 0 | 0 | 3 | 1 | 0 | 18123.9 | 40.6389 | -74.006 |



In the graph above points located below the line possibly less attractive than the points on the line.

Why? Because for each of these options exists another one that provides higher score or/and lower annual cost.

Why 'possibly'? Because this is only score representation of computed 'attractiveness' of each recommended location, in the real life it could be the situation that a location with a lower score would be much more interesting for Jack and Jessy.

5.3 Recommendation map

The last step of this project is to show the recommended tracts on the map.



Legend: color: tract score, yellow dots: optimal locations, blue dots: top 20 by score

6. Discussion

Some moments that we want to mention as possible topics for discussion:

- Recommendation system implemented in this project is very "simple" and could be further improved, for example, by extracting the category scores from the list of neighborhoods/tracts with their ratings provided by the customer/user. In fact, it could be difficult to give some digital representation to how a person feels about every single venue type, not to mention how many such types could exist, but much easier to evaluate different districts that the user knows.
- Another way to improve the system - add new features to describe the locations, for example traffic conditions, climate and ecology etc.

- In this project we assumed that the number of desirable/unwanted venues equally important as their scores, for example, that two venues with the score of 4 points have the same attractiveness as one venue with the score of 8 points. Obviously in the real life the situation is slightly more complicated, so the system could be improved by introducing weight coefficients for each additional venue of the same type (for example, $0.5^{(venue\ number - 1)}$)

7. Conclusion

In this Project we've just touched the surface of almost unlimited possibilities provided by opensource data and ML algorithms.

It took us about 20 minutes to find the relevant and high-quality dataset, and less than 40 blocks of Jupyter Notebook to get some fairly good results.

Although the implemented model is very simple, it already helped us to narrow down list of possible locations for a new home from almost 500 to 20 or even 8 options.

References

1. United States Census Bureau, Geographical Mobility by Tenure data:
<https://www.census.gov/data/tables/time-series/demo/geographic-mobility/historic.html>
2. BetaNYC portal, dataset “Median Household Income 2010, Census Tracts”:
<http://data.beta.nyc/dataset/median-household-income-2010-census-tracts>
3. Foursquare API: <https://developer.foursquare.com>