1. INTRODUCTION

1.1 Background

A work- life integration is crucial to improve our physical, mental wellbeing and productive workforce.

Not everyone has the same work-life balance goals, but each one needs to achieve your own state of equilibrium.

Sometimes getting out of your comfort zone is the best way to find this crucial balance. Maybe to move to another city.

Perhaps a workplace relocation encourages to look for a better work-life balance.

Whatever the reasons to move to a new city, we need to consider all the advantages and disadvantages of our relocations first.

The idea is to create a model that can facilitate and automate to check how affordable will be to live there with the salary he/she expects to earn in the new city.

1.2 Problem

To accomplish this, it will be necessary to find out a reliable source of information from where to get the data. Then, it needs to scrape the internet looking for a neighborhood data set for the cities "X' and "Y" and Housing costs data set.

In this case, it has selected two famous cities in the US

City Y: San Francisco

City X: New York City

The person can choose 5 categories of the venues he/she would like to have in the neighborhood of the new city "Y". Let's suppose that he/she likes:

- Parks
- Mall
- Martial Arts Academies
- Yoga Studios
- Tennis centers

1.3 Interest

The audience expected are many professionals who can use this tool to find out which neighborhood to live in a new city.

It also could be interesting for realtors looking for places to live for their customers.

2. DATA

2.1 Data sources

Rental Amount data set: https://www.zillow.com/research/data/

All Postal Codes in the US: http://files.zillowstatic.com/research/public/Zip/Zip_Zri_AllHomesPlusMultifamily_Summary.csv

Neighborhoods by postal code (NBPC): http://www.geonames.org/

City Y: https://www.geonames.org/postalcode-search.html?q=California%2C+San+Francisco&country=US'

City X: https://www.geonames.org/postalcode-search.html?q=New+York%2C+Manhattan&country=US

2.2 Data Preparation and Cleaning

After downloading the information from the Data sets, it has dropped all unnecessary columns from the datasets and rename the columns with the names will use in the notebook, i.e.: Postal Code instead of code.

From the rental data set, we only extracted two columns: Postal Code and Rental Amount

Then, it has joined the Rental Amount to the Neighborhood venues based on the Postal Code

2.3 Feature selection

We will use the “venue category” to match each venue to the characteristics chosen.

This will check for each of the venues located within a radius of 500 Mt from the centroid of the neighborhoods of the city.

Using the FOURSQUARE[1] APIs, it is possible to find out all the venues for each neighborhood “NY” in the city “Y”

3. METHODOLOGY

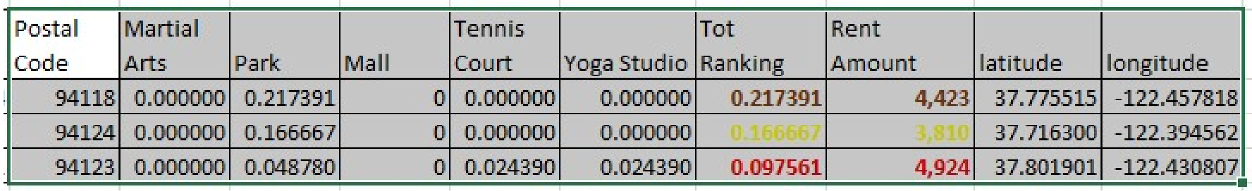
3.1 Exploratory analysis

It is unknown what neighborhood has the venues that a person desire to be available, first it is mandatory to explore all the venues of all the neighborhoods in the city.

Then, it will be considered only the venues whose venue characteristics match at least one of the categories desired by the person.

The neighborhood with greater total ranking and less rental amount is considered the best neighborhood to live. In the example below should be 94118 - Richmond District

[1] FOURSQUARE City guide



Postal Code	Martial Arts	Park	Mall	Tennis Court	Yoga Studio	Tot Ranking	Rent Amount	latitude	longitude
94118	0.000000	0.217391	0	0.000000	0.000000	0.217391	4,423	37.775515	-122.457818
94124	0.000000	0.166667	0	0.000000	0.000000	0.166667	3,810	37.716300	-122.394562
94123	0.000000	0.048780	0	0.024390	0.024390	0.097561	4,924	37.801901	-122.430807

3.2 Platform

It has used Python Notebook in Watson studio[2] and its associated open source partners.

In order to identify the centroids of the neighborhoods scrapped from the data set (NBPC) it has used Nominatim[3] to get the Latitude and Longitude of each one. It has found the following quantity of neighborhoods

City Y: 52

City X: 140

Then it has mapped all neighborhoods for each city using folium[4] in order to have an idea where they are located.

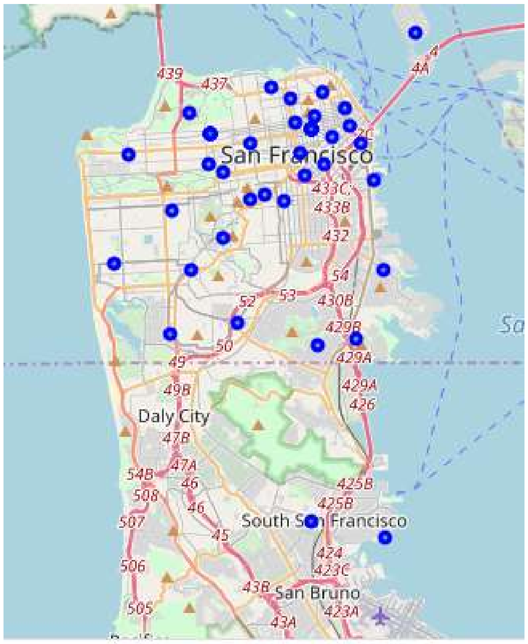


Figure 1: city Y – Neighborhoods of San Francisco

[2] IBM software on the cloud for Data Science
[3] See Geolocator / geopy.geocoders
[4] folium from conda - forge

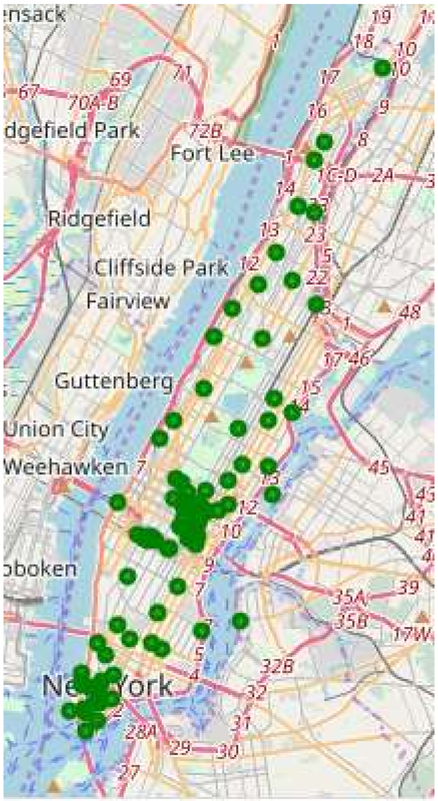


Figure 2: City X – Neighborhoods of New York city

3.2 Modeling

In this case of neighborhood segmentation by its venue categories and considering the low volume of data it has selected the K-Means algorithm to cluster them.

Find all nearby venues within 500 Mt of the centroid of each neighborhood using APIs provided by FOURSQUARE[5]

City Y: 3156 Venues

City X: 8225 Venues

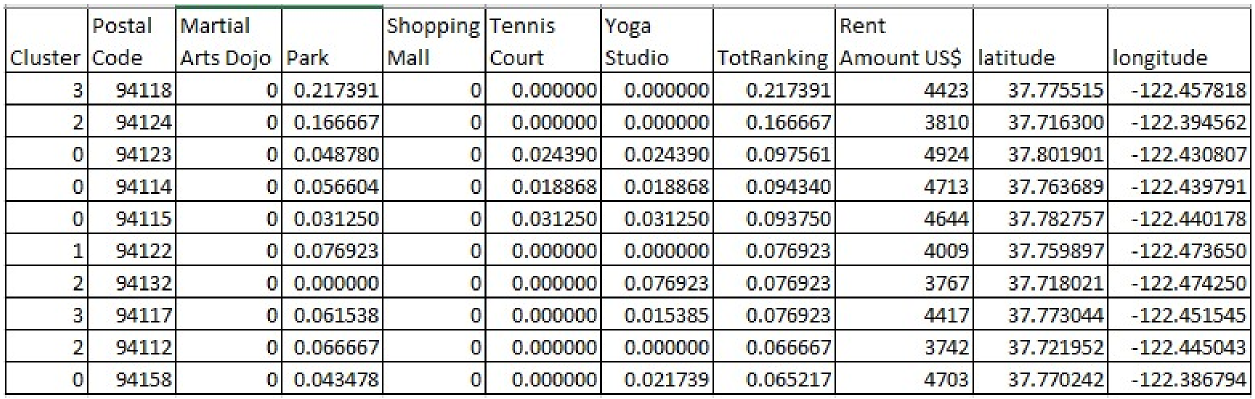
Since the Venue category is a categorical variable it has normalized in order to perform clustering and also has considered the rental amount associated with the postal code.

After filtering the neighborhoods that have venues which match the characteristics desired, i.e: Parks, Mall, Yoga studio, Martial Arts academy and Tennis courts, remain the following:

City Y: 21

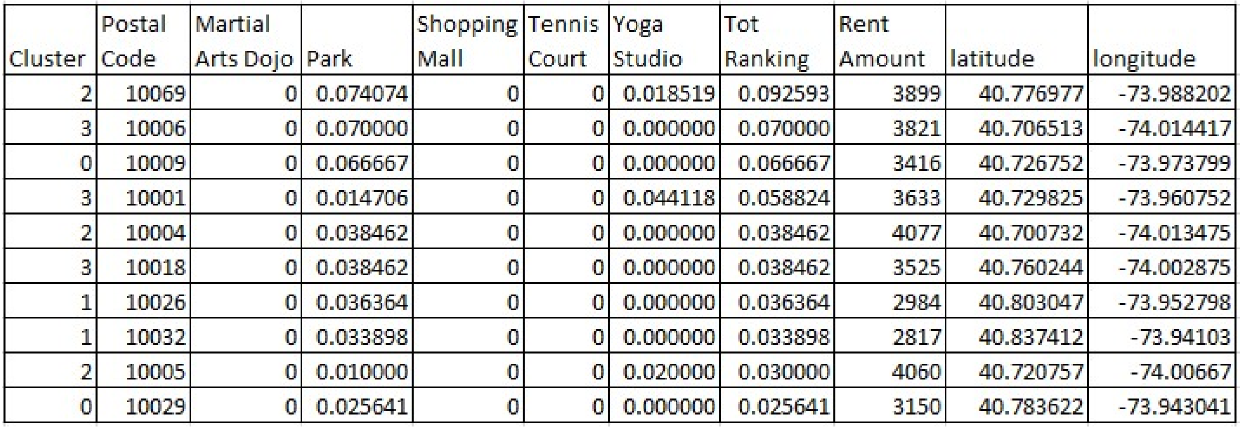
City X: 18

[5] https://foursquare.com/city-guide



Cluster	Postal Code	Martial Arts Dojo	Park	Shopping Mall	Tennis Court	Yoga Studio	TotRanking	Rent Amount US$	latitude	longitude
3	94118	0	0.217391	0	0.000000	0.000000	0.217391	4423	37.775515	-122.457818
2	94124	0	0.166667	0	0.000000	0.000000	0.166667	3810	37.716300	-122.394562
0	94123	0	0.048780	0	0.024390	0.024390	0.097561	4924	37.801901	-122.430807
0	94114	0	0.056604	0	0.018868	0.018868	0.094340	4713	37.763689	-122.439791
0	94115	0	0.031250	0	0.031250	0.031250	0.093750	4644	37.782757	-122.440178
1	94122	0	0.076923	0	0.000000	0.000000	0.076923	4009	37.759897	-122.473650
2	94132	0	0.000000	0	0.000000	0.076923	0.076923	3767	37.718021	-122.474250
3	94117	0	0.061538	0	0.000000	0.015385	0.076923	4417	37.773044	-122.451545
2	94112	0	0.066667	0	0.000000	0.000000	0.066667	3742	37.721952	-122.445043
0	94158	0	0.043478	0	0.000000	0.021739	0.065217	4703	37.770242	-122.386794

Table 1: City Y - Clusters (first 10 rows)

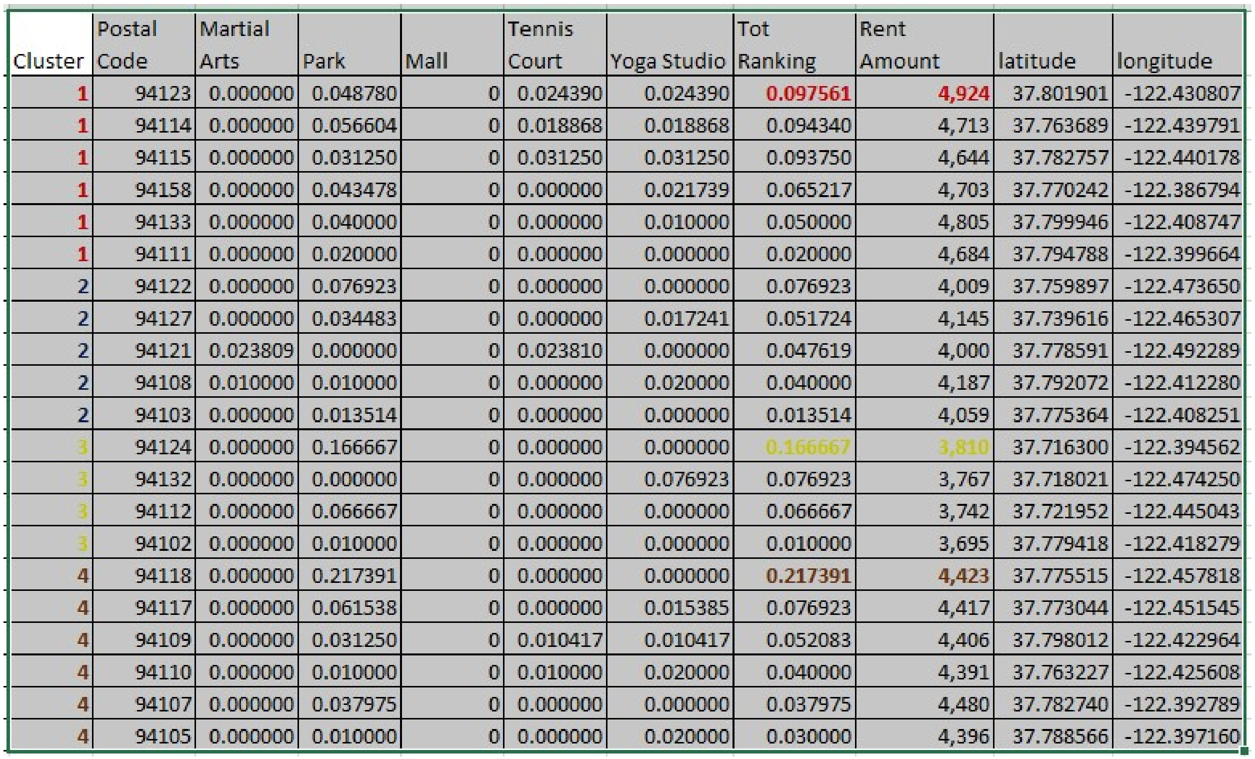


Cluster	Postal Code	Martial Arts Dojo	Park	Shopping Mall	Tennis Court	Yoga Studio	Tot Ranking	Rent Amount	latitude	longitude
2	10069	0	0.074074	0	0	0.018519	0.092593	3899	40.776977	-73.988202
3	10006	0	0.070000	0	0	0.000000	0.070000	3821	40.706513	-74.014417
0	10009	0	0.066667	0	0	0.000000	0.066667	3416	40.726752	-73.973799
3	10001	0	0.014706	0	0	0.044118	0.058824	3633	40.729825	-73.960752
2	10004	0	0.038462	0	0	0.000000	0.038462	4077	40.700732	-74.013475
3	10018	0	0.038462	0	0	0.000000	0.038462	3525	40.760244	-74.002875
1	10026	0	0.036364	0	0	0.000000	0.036364	2984	40.803047	-73.952798
1	10032	0	0.033898	0	0	0.000000	0.033898	2817	40.837412	-73.94103
2	10005	0	0.010000	0	0	0.020000	0.030000	4060	40.720757	-74.00667
0	10029	0	0.025641	0	0	0.000000	0.025641	3150	40.783622	-73.943041

Table 2: City X - Clusters (first 10 rows)

4. RESULTS

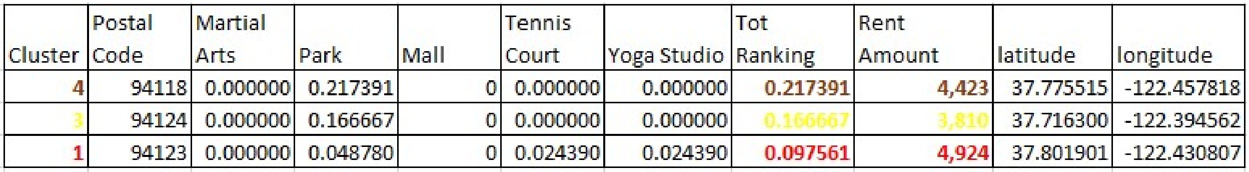
4.1 City Y: San Francisco



Cluster	Postal Code	Martial Arts	Park	Mall	Tennis Court	Yoga Studio	Tot Ranking	Rent Amount	latitude	longitude
1	94123	0.000000	0.048780	0	0.024390	0.024390	**0.097561**	**4,924**	37.801901	-122.430807
1	94114	0.000000	0.056604	0	0.018868	0.018868	0.094340	4,713	37.763689	-122.439791
1	94115	0.000000	0.031250	0	0.031250	0.031250	0.093750	4,644	37.782757	-122.440178
1	94158	0.000000	0.043478	0	0.000000	0.021739	0.065217	4,703	37.770242	-122.386794
1	94133	0.000000	0.040000	0	0.000000	0.010000	0.050000	4,805	37.799946	-122.408747
1	94111	0.000000	0.020000	0	0.000000	0.000000	0.020000	4,684	37.794788	-122.399664
2	94122	0.000000	0.076923	0	0.000000	0.000000	0.076923	4,009	37.759897	-122.473650
2	94127	0.000000	0.034483	0	0.000000	0.017241	0.051724	4,145	37.739616	-122.465307
2	94121	0.023809	0.000000	0	0.023810	0.000000	0.047619	4,000	37.778591	-122.492289
2	94108	0.010000	0.010000	0	0.000000	0.020000	0.040000	4,187	37.792072	-122.412280
2	94103	0.000000	0.013514	0	0.000000	0.000000	0.013514	4,059	37.775364	-122.408251
3	94124	0.000000	0.166667	0	0.000000	0.000000	0.166667	3,810	37.716300	-122.394562
3	94132	0.000000	0.000000	0	0.000000	0.076923	0.076923	3,767	37.718021	-122.474250
3	94112	0.000000	0.066667	0	0.000000	0.000000	0.066667	3,742	37.721952	-122.445043
3	94102	0.000000	0.010000	0	0.000000	0.000000	0.010000	3,695	37.779418	-122.418279
4	94118	0.000000	0.217391	0	0.000000	0.000000	**0.217391**	**4,423**	37.775515	-122.457818
4	94117	0.000000	0.061538	0	0.000000	0.015385	0.076923	4,417	37.773044	-122.451545
4	94109	0.000000	0.031250	0	0.010417	0.010417	0.052083	4,406	37.798012	-122.422964
4	94110	0.000000	0.010000	0	0.010000	0.020000	0.040000	4,391	37.763227	-122.425608
4	94107	0.000000	0.037975	0	0.000000	0.000000	0.037975	4,480	37.782740	-122.392789
4	94105	0.000000	0.010000	0	0.000000	0.020000	0.030000	4,396	37.788566	-122.397160

Table 3: Clusters City Y: San Francisco

Considering the three best rankings for city Y



Cluster	Postal Code	Martial Arts	Park	Mall	Tennis Court	Yoga Studio	Tot Ranking	Rent Amount	latitude	longitude
4	94118	0.000000	0.217391	0	0.000000	0.000000	**0.217391**	**4,423**	37.775515	-122.457818
3	94124	0.000000	0.166667	0	0.000000	0.000000	0.166667	3,810	37.716300	-122.394562
1	94123	0.000000	0.048780	0	0.024390	0.024390	**0.097561**	**4,924**	37.801901	-122.430807

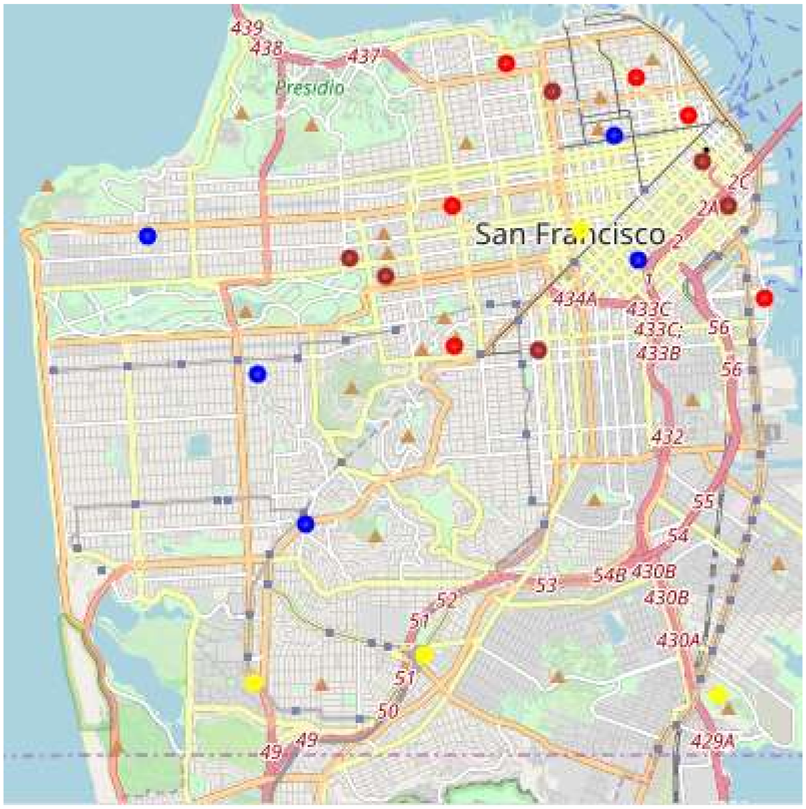
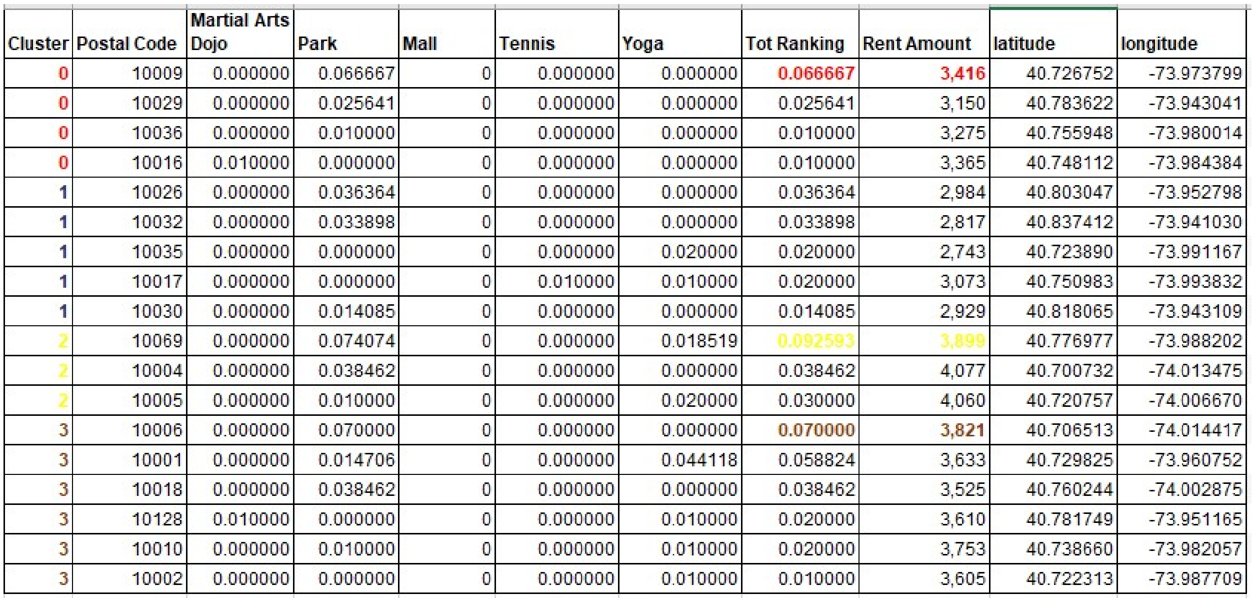


Figure 3: Map of Clusters City Y: San Francisco

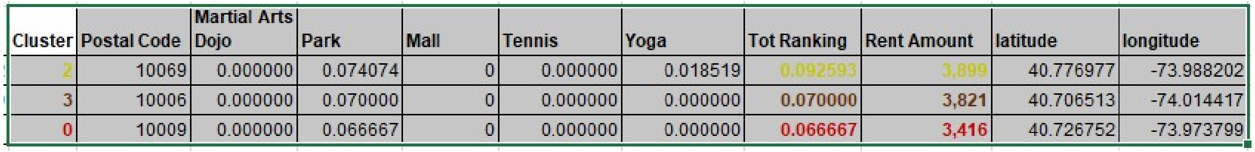
4.2 City X: New York



Cluster	Postal Code	Martial Arts Dojo	Park	Mall	Tennis	Yoga	Tot Ranking	Rent Amount	latitude	longitude
0	10009	0.000000	0.066667	0	0.000000	0.000000	0.066667	3,416	40.726752	-73.973799
0	10029	0.000000	0.025641	0	0.000000	0.000000	0.025641	3,150	40.783622	-73.943041
0	10036	0.000000	0.010000	0	0.000000	0.000000	0.010000	3,275	40.755948	-73.980014
0	10016	0.010000	0.000000	0	0.000000	0.000000	0.010000	3,365	40.748112	-73.984384
1	10026	0.000000	0.036364	0	0.000000	0.000000	0.036364	2,984	40.803047	-73.952798
1	10032	0.000000	0.033898	0	0.000000	0.000000	0.033898	2,817	40.837412	-73.941030
1	10035	0.000000	0.000000	0	0.000000	0.020000	0.020000	2,743	40.723890	-73.991167
1	10017	0.000000	0.000000	0	0.010000	0.010000	0.020000	3,073	40.750983	-73.993832
1	10030	0.000000	0.014085	0	0.000000	0.000000	0.014085	2,929	40.818065	-73.943109
2	10069	0.000000	0.074074	0	0.000000	0.018519	0.092593	3,899	40.776977	-73.988202
2	10004	0.000000	0.038462	0	0.000000	0.000000	0.038462	4,077	40.700732	-74.013475
2	10005	0.000000	0.010000	0	0.000000	0.020000	0.030000	4,060	40.720757	-74.006670
3	10006	0.000000	0.070000	0	0.000000	0.000000	0.070000	3,821	40.706513	-74.014417
3	10001	0.000000	0.014706	0	0.000000	0.044118	0.058824	3,633	40.729825	-73.960752
3	10018	0.000000	0.038462	0	0.000000	0.000000	0.038462	3,525	40.760244	-74.002875
3	10128	0.010000	0.000000	0	0.000000	0.010000	0.020000	3,610	40.781749	-73.951165
3	10010	0.000000	0.010000	0	0.000000	0.010000	0.020000	3,753	40.738660	-73.982057
3	10002	0.000000	0.000000	0	0.000000	0.010000	0.010000	3,605	40.722313	-73.987709

Table 4: Clusters City X: New York

Considering the three best rankings for city X



Cluster	Postal Code	Martial Arts Dojo	Park	Mall	Tennis	Yoga	Tot Ranking	Rent Amount	latitude	longitude
2	10069	0.000000	0.074074	0	0.000000	0.018519	0.092593	3,899	40.776977	-73.988202
3	10006	0.000000	0.070000	0	0.000000	0.000000	0.070000	3,821	40.706513	-74.014417
0	10009	0.000000	0.066667	0	0.000000	0.000000	0.066667	3,416	40.726752	-73.973799

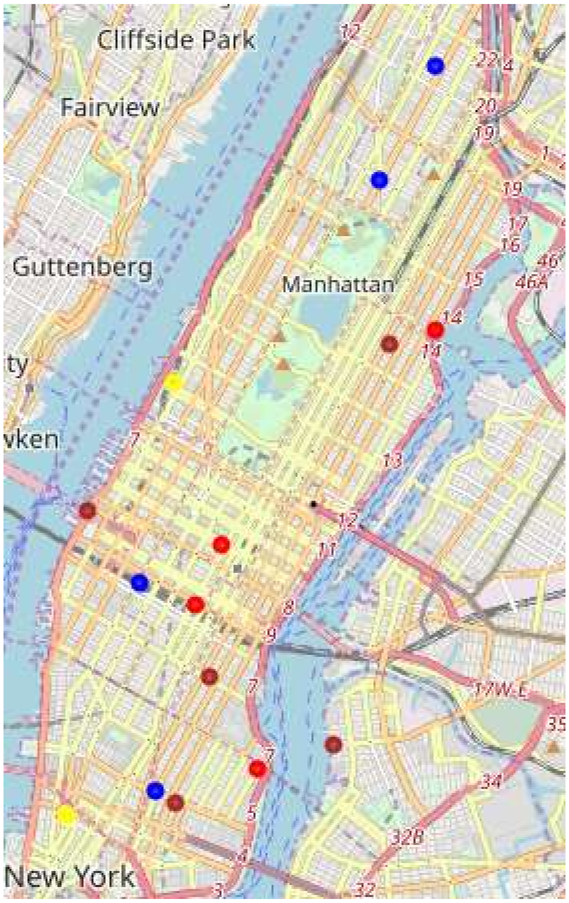


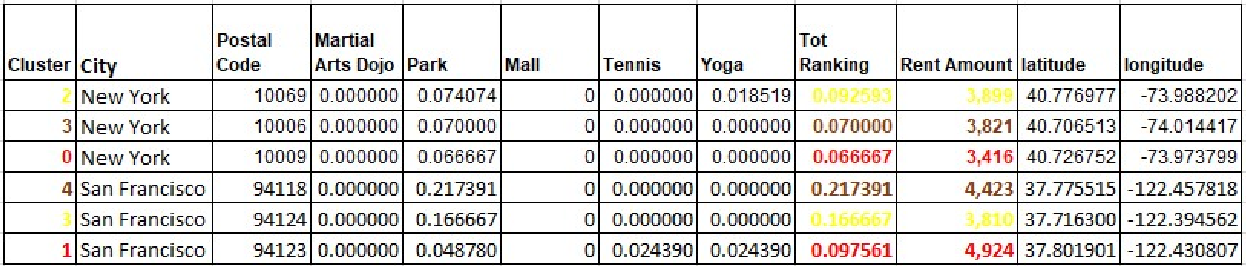
Figure 4: Map of Clusters city X: New York

5. DISCUSSION

Putting together the three best neighborhoods of each city can be seen that San Francisco has better rankings than New York however it appears to be more expensive.

The best one in San Francisco 94118 - Richmond District the rental is about $ 4,423 while New York 10069 – Riverside Park is $ 3,899 (13%)

The cheapest one in San Francisco 94124 – Hunters Point is about $ 3,810 while New York 10009 – East Village is $ 3,416 (12%)



Cluster	City	Postal Code	Martial Arts Dojo	Park	Mall	Tennis	Yoga	Tot Ranking	Rent Amount	latitude	longitude
2	New York	10069	0.000000	0.074074	0	0.000000	0.018519	0.092593	3,899	40.776977	-73.988202
3	New York	10006	0.000000	0.070000	0	0.000000	0.000000	0.070000	3,821	40.706513	-74.014417
0	New York	10009	0.000000	0.066667	0	0.000000	0.000000	0.066667	3,416	40.726752	-73.973799
4	San Francisco	94118	0.000000	0.217391	0	0.000000	0.000000	0.217391	4,423	37.775515	-122.457818
3	San Francisco	94124	0.000000	0.166667	0	0.000000	0.000000	0.166667	3,810	37.716300	-122.394562
1	San Francisco	94123	0.000000	0.048780	0	0.024390	0.024390	0.097561	4,924	37.801901	-122.430807

6. CONCLUSSION

The cost of living in San Francisco appears to be 13-15% more than New York, However, looking at some of the web sites[6] that analyze the cost of living in the US this percentage could be greater about 30%, but they consider other factors like Food & Groceries, Health, etc.

In my opinion, if you consider 20% of increment on the salary moving to San Francisco you can consider a neighborhood with the best ranking on the venue categories.

This tool can help with your findings based on your priorities because it can explore thousands of venues that are almost impossible to do walking the streets of the new city.

[6] www.salary.com ; https://www.bestplaces.net/