IBM Data Science Professional Certificate Capstone Project

Mexican Restaurant Location Analysis

Author: Włodzimierz Kuczyński

April 2021

Table of Contents

1. Introduction

The goal of this project will be to find the best location for a Mexican restaurant in San Francisco, CA. As defined by the Investor this will be a restaurant aimed mainly at 30-49 year old clientele with income above 100k$ per annum. One of the main attractions will be the high quality of the food. At first, we assume the main customers will be the inhabitants of the neighbouring areas, with a strong emphasis on clients with Mexican origin.

Firstly I will analyse the demographic data of San Francisco to identify the market area with the optimum clientele base in line with the above assumptions. Secondly I will look into the selected area, examine the competition and pinpoint a more precise location. The analysis will be valuable for anyone planning to open a similar Mexican restaurant in San Francisco.

2. Data

I will use census tracts data obtained from openICPSR for the demographic data analysis of the San Francisco - (National Neighborhood Data Archive (NaNDA): Socioeconomic Status and Demographic Characteristics of Census Tracts, United States, 2008-2017). ICPSR is an Inter-university Consortium for Political and Social Research. Machine learning clustering on the relevant census tracts data will be carried out to identify the preferable neighbourhood. The key data used to cluster the tracts will be:

- the average number of the population with income above 100k USD per year per square mile,
- the average number of the Hispanic population per square mile- the majority of Hispanic population in San Francisco are of Mexican origin,
- the average number of the 30-49 year old population per square mile,

Additionally, the land area of each tract will be taken into consideration.

The geojson census tracts file used for the maps was obtained from Metropolitan Transportation Commission (MTC) website, the transportation planning, coordinating and financing agency for the nine-county San Francisco Bay Area.

Information about the competition in the selected area will be obtained with the Foursquare API. Two key factors will analysed using this data:

- location of other Mexican restaurants - the restaurant should be located as far as possible from the direct competitors,
- location of other restaurants - locating the restaurant next to other restaurants which are not direct competitors could be even beneficial (possibility of taking over their clients).

3. Methodology

- Initially I will convert the proportion of the population with income above 100k USD, Hispanic population and 30-49 year old population to average number of each type of population per square mile.

- My next step will be scaling the data with the minmax scaler in order to use it for the Kmeans algorithm clustering.
- I will evaluate the impact of dividing the tracts into different number of clusters on the quality of the clusters using the yellowbrick library,
- Subsequently, I will assess which cluster would be the most suitable, considering its demographic data and size, for the location of the Mexican restaurant,
- Finally, I will use the data from the Foursquare api in order to pinpoint a more precise location of the restaurant, ensuring it is located as far as possible from the direct competitors

4. Preparation of the demographic data.

I will use the following information about the demographics of San Francisco census tracts in terms of: 1. Persons per square mile, ACS 2013-2017 (popden13_17)

2 .Proportion of people of Hispanic origin, ACS 2013-2017 (phispanic13_17)

3. Proportion of families with income greater than 100K, ACS 2013-2017 (inc_above_100k)

4. Proportion of population 30-49 years of age, ACS 2013-2017 (30-39_pop) 5. Land area of each tract (land_area)

	trctid	popden13_17	phispanic13_17	inc_above_100k	30-49_pop	land_area
0	06075010100	13232.0900	0.070078	0.430274	0.394505	0.299801
1	06075010200	21917.6300	0.030713	0.705506	0.366262	0.199064
2	06075010300	42257.9700	0.092529	0.510823	0.387252	0.103578
3	06075010400	35832.9800	0.117330	0.696329	0.435307	0.129629
4	06075010500	10194.5400	0.106652	0.635531	0.316983	0.263965
...	...	...	...	...	...	...
192	06075980401	0.0000	0.000000	0.000000	0.000000	0.161902
193	06075980501	1277.7750	0.187980	0.312500	0.148338	0.612001
194	06075980600	623.1144	0.173489	0.406250	0.249513	0.823284
195	06075980900	174.7661	0.239669	0.625000	0.632231	1.384708
196	06075990100	0.0000	0.000000	0.000000	0.000000	0.000000

197 rows × 6 columns

Table 1: Sample of the tracts census data used in the project

I converted proportion of persons of Hispanic origin, with income above 100k USD and aged 30-49 to numbers per square mile for each census tract.

	trctid	avg_hisp	avg_inc_above_100k	avg_30-49_pop	land_area
0	06075010100	927.280282	5693.425998	5220.121250	0.299801
1	06075010200	673.152074	15463.014831	8027.589366	0.199064
2	06075010300	3910.093326	21586.323342	16364.464002	0.103578
3	06075010400	4204.300087	24951.551587	15598.339407	0.129629
4	06075010500	1087.266095	6478.947501	3231.491081	0.263965
...	...	...	...	...	...
192	06075980401	0.000000	0.000000	0.000000	0.161902
193	06075980501	240.195549	399.304688	189.542075	0.612001
194	06075980600	108.103664	253.140225	155.474939	0.823284
195	06075980900	41.886091	109.228813	110.492619	1.384708
196	06075990100	0.000000	0.000000	0.000000	0.000000

197 rows × 5 columns

Table 2: Sample of the tracts census data used in the project converted to average population per square mile

	avg_hisp	avg_inc_above_100k	avg_30-49_pop	land_area
count	197.000000	197.000000	197.000000	197.000000
mean	4944.686772	15457.786199	10577.024629	0.237933
std	6420.243833	10152.912164	8075.466483	0.278395
min	0.000000	0.000000	0.000000	0.000000
25%	1450.268496	9008.583807	5873.239622	0.113413
50%	2914.469614	13853.137914	9203.147634	0.158258
75%	5439.174669	20435.035732	13164.907394	0.254182
max	39646.417008	67092.308299	57851.690639	2.358635

Table 3: Key statistical measures of the tracts census data used in the project (tracts population by square mile)

5. Generating the clusters and evaluating their quality.

In the next step I clustered the census tracts using Kmeans algorithm using the above data in order to find the optimal market area for the restaurant. I evaluated the impact of dividing the tracts into different number of clusters on the quality of the clusters using the Yellowbrick library.

I will evaluate what is the optimal number of clusters between the range of 3 and 9 using the Yellowbrick silhouette visualizer.

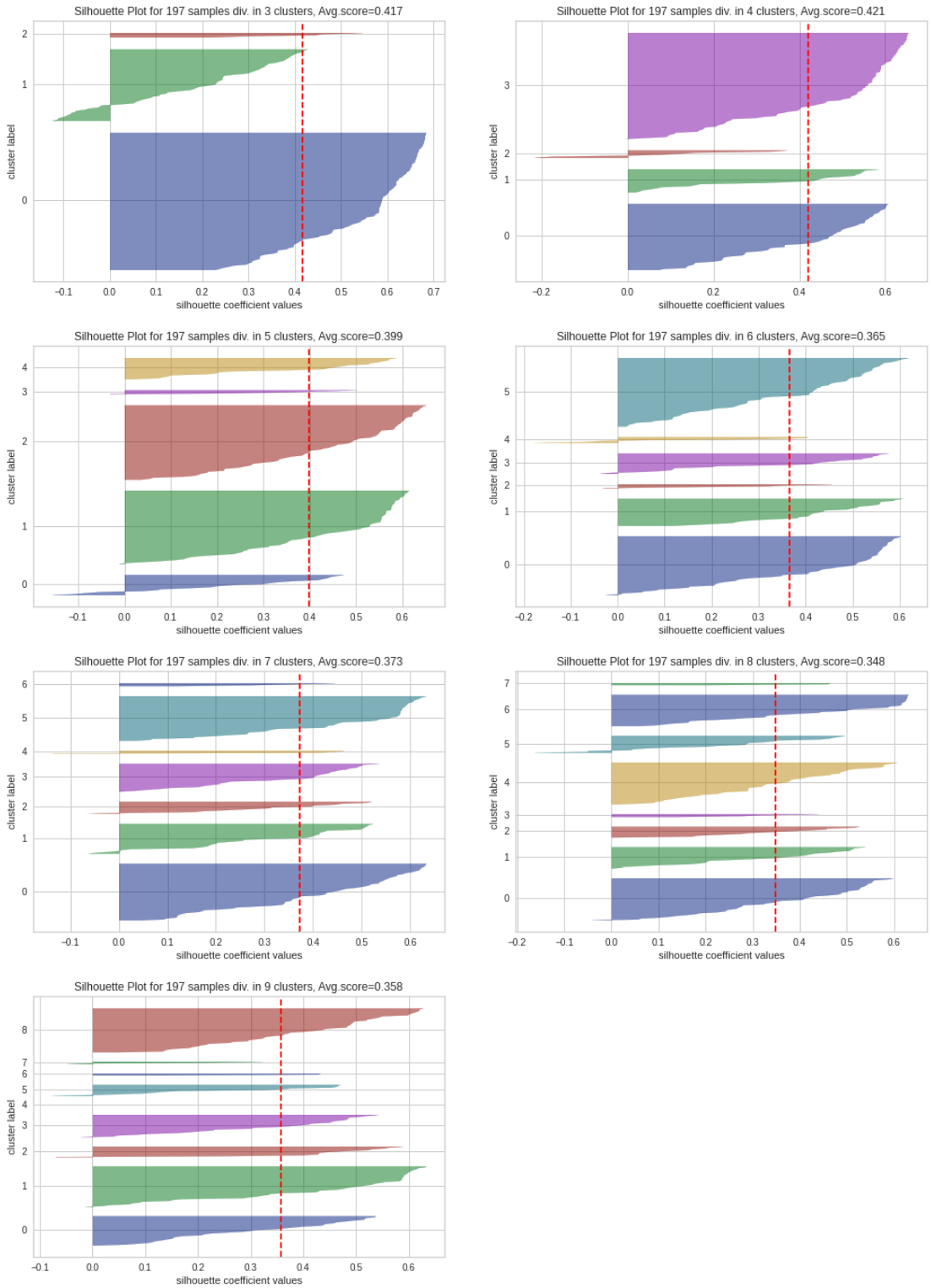

Figure 1: Silhouette plots for different number of clusters

Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are. The value of the silhouette score varies from -1 to 1. If the score is 1, the cluster is dense and well-separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighboring clusters. A negative score indicates that the samples might have got assigned to the wrong clusters. It is also a negative symptom if there are clusters below the average silhouette score. After an initial assessment we can see that most likely we can disregard options with 3, 4 and 5 clusters considering the following:

- cluster 1 of the 3 clusters option has a large number of samples with a negative score
- cluster 2 of the 4 clusters option has some number of samples with a negative score and its score is below the average score
- cluster 0 of the 5 clusters option has a significant number of samples with a negative score

Out of the other options the option with 7 clusters looks most promising as it does not have clusters with a significant number of negative scores or clusters with scores below the average score and it has the highest average score out of the options with more clusters than 5.

We will have a better understanding how well each clustering options suits our requirements when we analyse each of the clusters in terms of their demographics.

6. Analysis of the clusters' demographics and land area

In the table below of various clusters in various clustering options below I ranked by their suitability as far as their demographics are concerned.

The darker the color of each of the parameter of the tract cluster in the table below reflects its suitability for the location of the restaurant.

cluster	avg_hisp	avg_inc_above_100k	avg_30-49_pop	cluster_land_area_sum
4_of_9_clusters	39646.417008	48909.843114	57851.690639	0.035539
6_of_7_clusters	34235.765618	27448.860841	46844.931528	0.130532
3_of_8_clusters	34235.765618	27448.860841	46844.931528	0.130532
2_of_3_clusters	34314.647883	25064.500605	41358.848581	0.226749
2_of_6_clusters	34314.647883	25064.500605	41358.848581	0.226749
3_of_5_clusters	34314.647883	25064.500605	41358.848581	0.226749
2_of_4_clusters	26468.671631	35608.711446	37312.973654	0.334371
7_of_8_clusters	13392.044543	53182.396179	30569.848777	0.107622
6_of_9_clusters	13392.044543	53182.396179	30569.848777	0.107622
4_of_7_clusters	11193.239330	49863.780457	29792.675621	0.174407

Table 4: Tract clusters ranked by their demographic parameters (population by square mile)

Cluster 4 of 9 clusters has the most beneficial demographics but it has also the smallest area which is disadvantageous. Let us plot the above data.

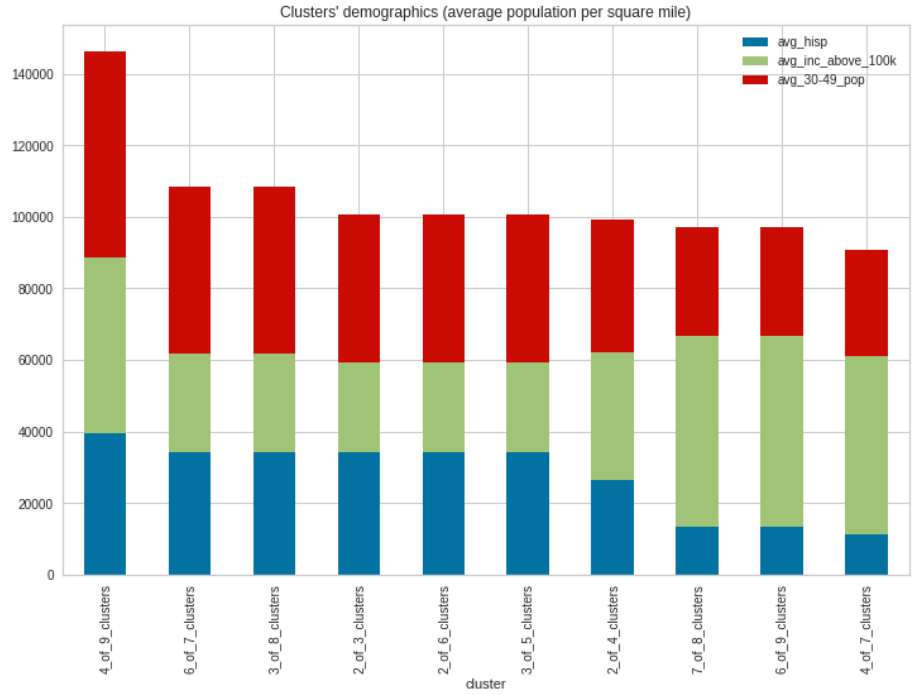

Figure 2: Clusters' demographics (average population per square mile)

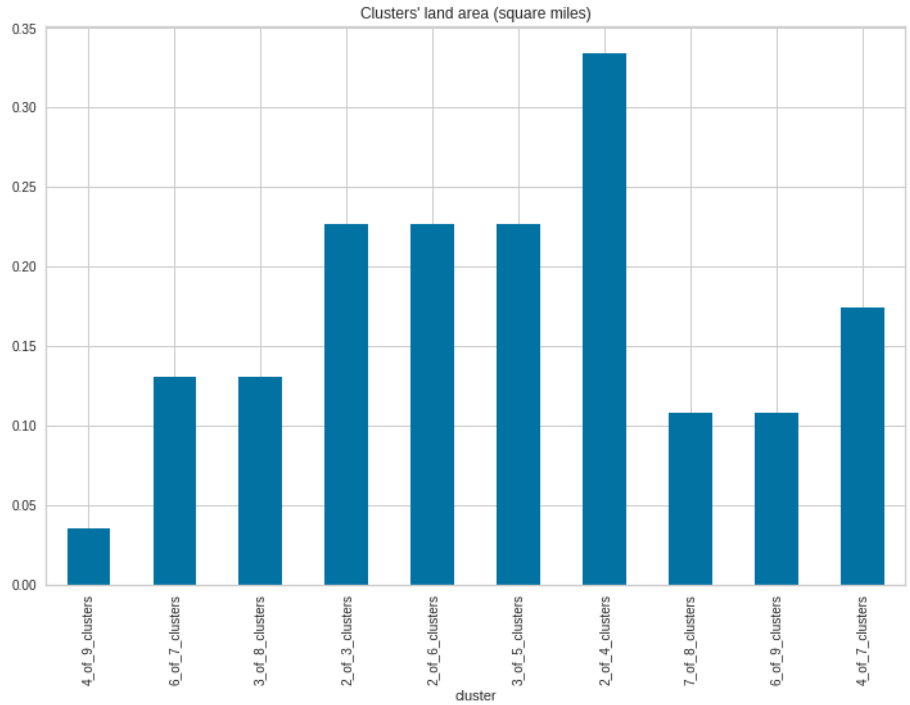

Figure 3: Clusters land area (square miles)

We observe that:

- cluster 4 of 9 has the most beneficial demographics but it has also the smallest area
- cluster 6 of 7 has good demographics particularly the age and ethnicity and significantly larger area then the previous option
- cluster 2 of 4 the largest area but a lower demographics suitability in terms of ethnicity and age of the population

I will visualise these three selected above clustering options on the tracts map of the San Francisco county using the Folium library.

7. Visualisation of the selected clusters on the San Francisco county map.

Plotting the 9 clusters option with the preferred cluster 4 in yellow.

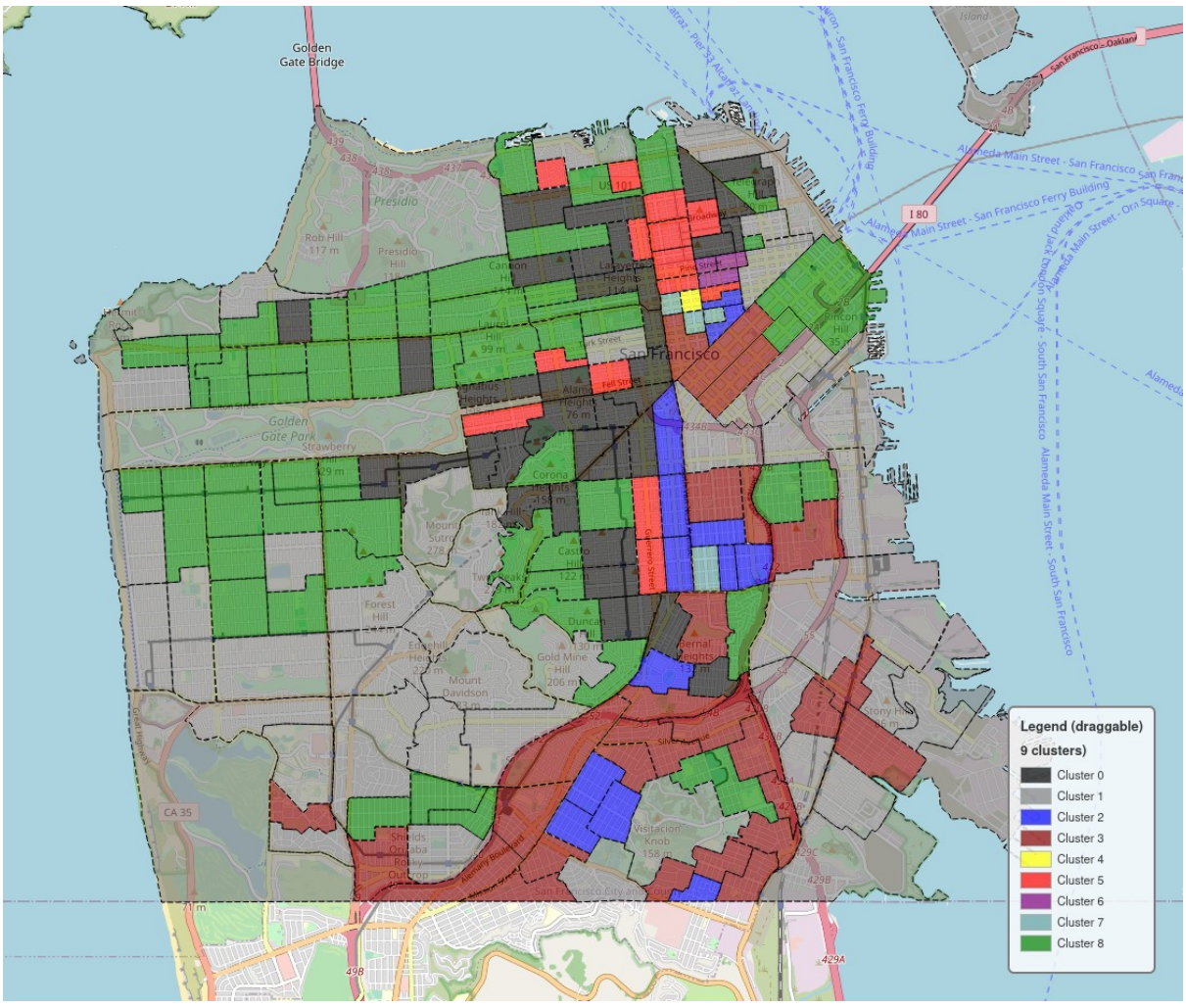

Figure 4: 9 clusters option visualised on the San Francisco county map

We can see above the location of the cluster 4_of_9 with the most preferable demographics.

Plotting the 7 clusters option with the preferred cluster 6 in yellow.

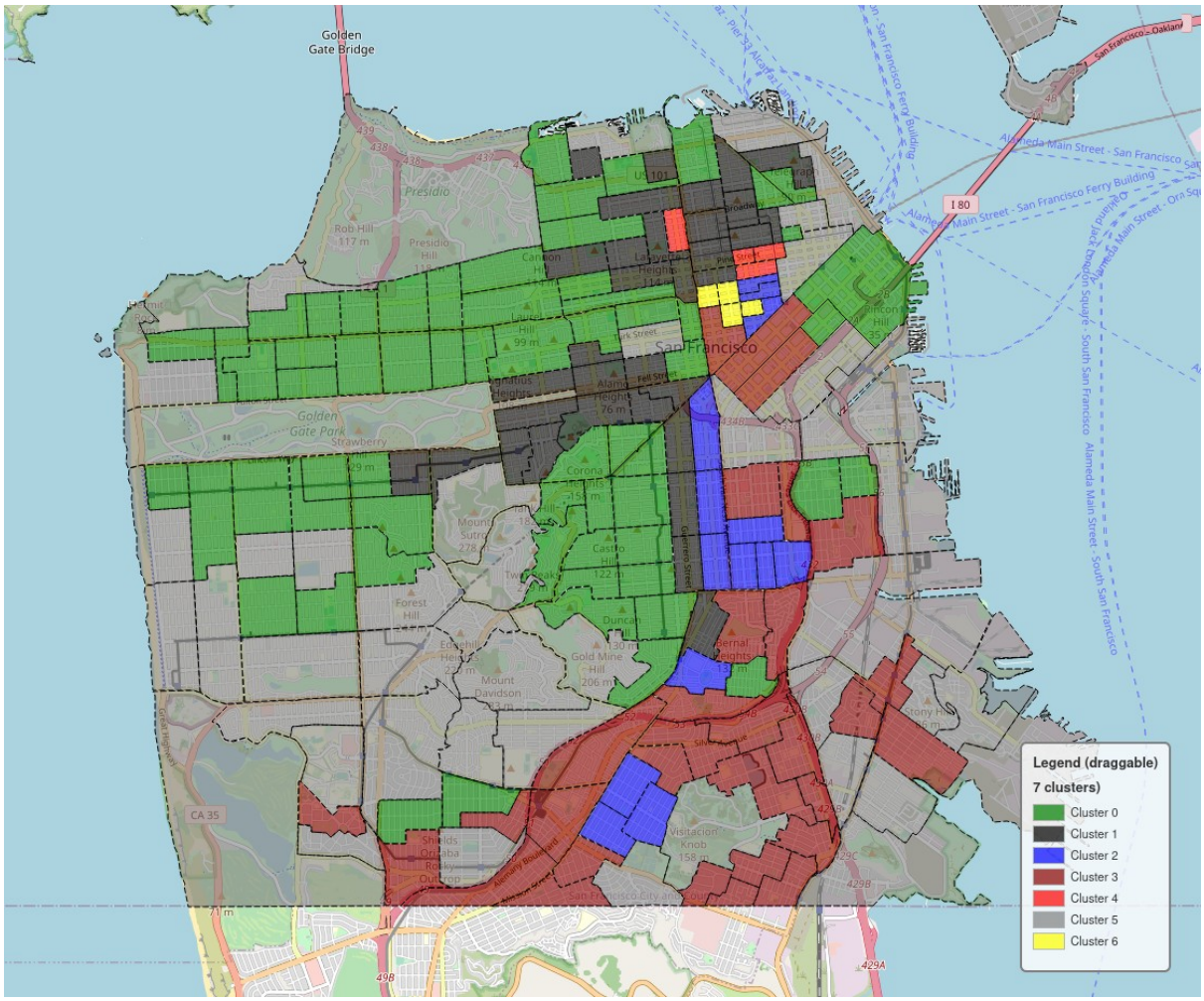

Figure 5: 7 clusters option visualised on the San Francisco county map

We see that that cluster 6 of 7 incorporates cluster 4 of 9 and it is approximately four times larger.

Plotting below the 4 clusters option with the preferred cluster 2 in yellow.

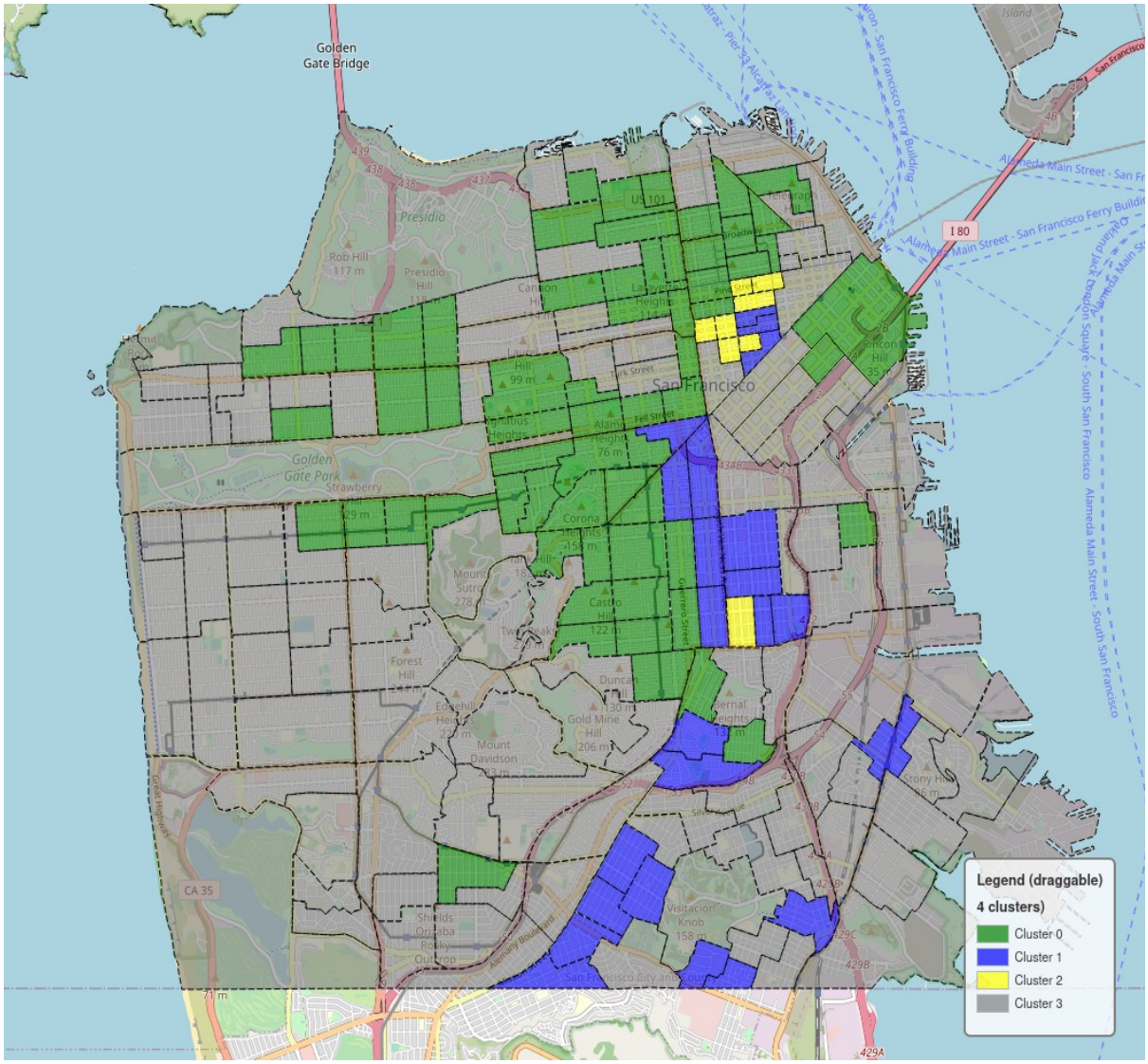

Figure 6: 4 clusters option visualised on the San Francisco county map

Cluster 2 of 4 incorporates both previously analysed clusters and includes a tract which not adjacent to the other tracts of this cluster. Let us remove it from the cluster and analyse the cluster again.

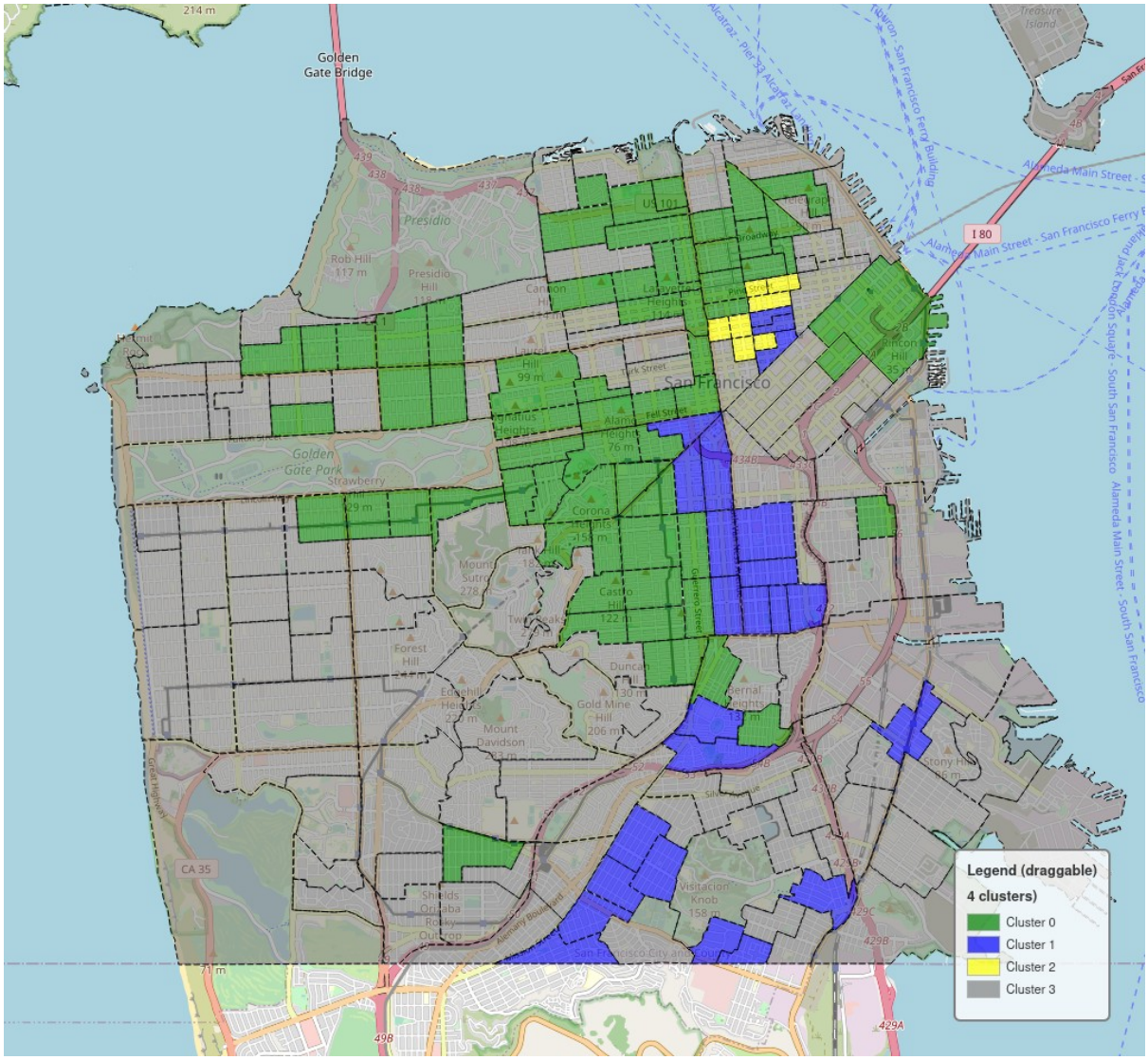

Figure 7: Modified 4 clusters option visualised on the San Francisco county map

cluster	avg_hisp	avg_inc_above_100k	avg_30-49_pop	cluster_land_area_sum
4_of_9_clusters	39646.417008	48909.843114	57851.690639	0.035539
6_of_7_clusters	34235.765618	27448.860841	46844.931528	0.130532
3_of_8_clusters	34235.765618	27448.860841	46844.931528	0.130532
mod_2_of_4_clusters	25302.742300	38477.518843	39869.896063	0.238154
2_of_3_clusters	34314.647883	25064.500605	41358.848581	0.226749
2_of_6_clusters	34314.647883	25064.500605	41358.848581	0.226749
3_of_5_clusters	34314.647883	25064.500605	41358.848581	0.226749
7_of_8_clusters	13392.044543	53182.396179	30569.848777	0.107622
6_of_9_clusters	13392.044543	53182.396179	30569.848777	0.107622
4_of_7_clusters	11193.239330	49863.780457	29792.675621	0.174407

Table 5: Tract clusters ranked by their demographic parameters (population by square mile)

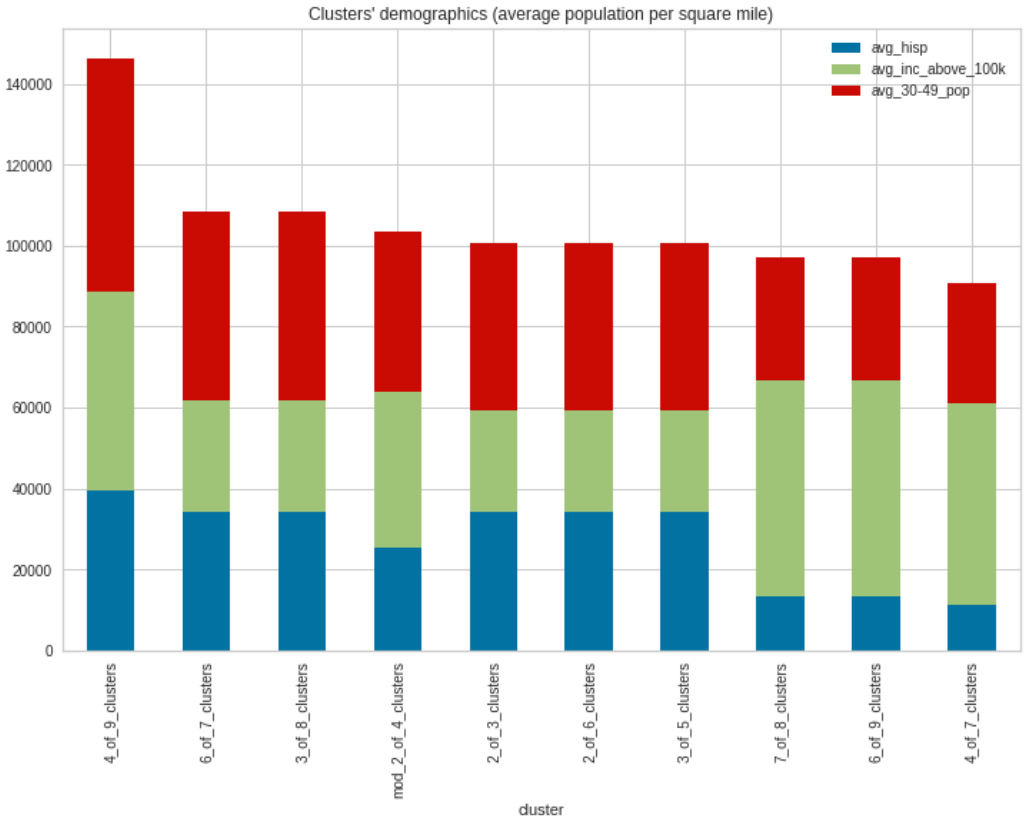

Figure 8: Clusters' demographics (average population per square mile)

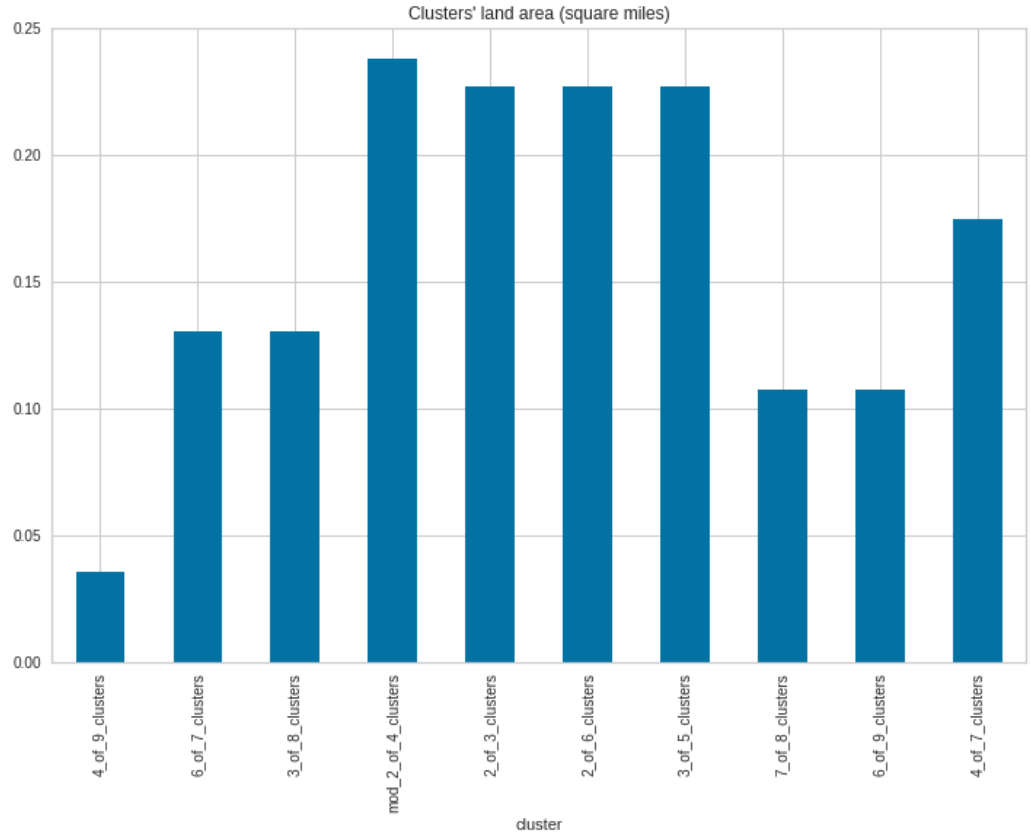

Figure 9: Clusters land area (square miles)

The cluster 2 of 4 after modification significantly improved its demographics, become continuous but of course at the cost of its size although still remaining the largest cluster.

The modified cluster 2 of 4 is a good balance between demographics and size of the cluster. It is the largest and third out of the top ten clusters in terms of the demographic characteristics. It contains both 4 of 9 and 6 of 7 clusters. Looking at the clusters' demographics and the clusters' maps visualizations we observe that:

- the south-western part of cluster 2 of 4 has most preferable demographics in terms of Hispanic population and population aged 30-49
- the north eastern part of cluster 2 of 4 worse parameters as far as Hispanic population and population aged 30-49 are concerned is balanced by a higher population of with income above 100k USD

My initial recommendation would be to locate the Mexican restaurant in the middle of the modified 2 of 4 cluster to fully benefit from the overall beneficial, balanced cluster's demographic data, its size as well as the vicinity to the best in terms of demographic parameters cluster 4 of 9 located to the south-west of the centre of the modified 2 of 4 cluster. But in order to confirm this I need to analyze the locations of the restaurant's direct competitors and other restaurants in cluster 2 of 4 and its adjacent areas.

8. Restaurants in cluster 2 of 4 and its surroundings.

I downloaded a list of all the restaurant's in cluster 2 of 4 and its surroundings using Foursquare API including their location data. Below I present a sample of this data.

	trctid	Tract Latitude	Tract Longitude	Venue Id	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Category ID	color
0	06075011901	37.790473	-122.413914	5c0dd1ad5a2c91002c442f28	Gusto Pinsa Romana	37.789594	-122.413873	Italian Restaurant	4bf58dd8d48988d110941735	blue
1	06075011901	37.790473	-122.413914	44f72e06f964a5204b381fe3	Big 4 Restaurant	37.791490	-122.412393	American Restaurant	4bf58dd8d48988d14e941735	blue
2	06075011901	37.790473	-122.413914	54c2a5e4498ee52a438800f7	Liholiho Yacht Club	37.788332	-122.414605	Hawaiian Restaurant	52e81612bcbc57f1066b79fe	blue
3	06075011901	37.790473	-122.413914	5643ce76498ec3c226f039e3	Del Popolo	37.789807	-122.411347	Pizza Place	4bf58dd8d48988d1ca941735	blue
4	06075011901	37.790473	-122.413914	59cb288d6c08d172265a8e6d	Chisme	37.788467	-122.414802	Mexican Restaurant	4bf58dd8d48988d1c1941735	red
...	...	...	...	...	...	...	...	...	...	...
210	06075012401	37.783051	-122.415789	4a2c0803f964a5200e971fe3	Sam's Diner	37.778375	-122.415426	Diner	4bf58dd8d48988d147941735	blue
211	06075012401	37.783051	-122.415789	5c40381f603d2a002c295ff9	The Pawn Shop	37.781060	-122.408548	Tapas Restaurant	4bf58dd8d48988d1db931735	blue
212	06075012401	37.783051	-122.415789	59a24aa726659b0902fb4d04	Oma Sushi	37.786124	-122.410247	Sushi Restaurant	4bf58dd8d48988d1d2941735	blue
213	06075012502	37.783933	-122.412595	4dfd856d483b96a3aaa9eb34	Box Kitchen	37.781158	-122.406243	American Restaurant	4bf58dd8d48988d14e941735	blue
214	06075012502	37.783933	-122.412595	4af09c87f964a520a1dd21e3	Tín	37.780840	-122.405770	Vietnamese Restaurant	4bf58dd8d48988d14a941735	blue

215 rows × 10 columns

Table 6: Sample of the restaurants data in cluster 2 of 4 and its surroundings.

Below I limited the data to Mexican restaurants, taco and burrito places.

	index	trctid	Tract Latitude	Tract Longitude	Venue Id	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Category ID	color
0	4	06075011901	37.790473	-122.413914	59cb288d6c08d172265a8e6d	Chisme	37.788467	-122.414802	Mexican Restaurant	4bf58dd8d48988d1c1941735	red
1	51	06075011901	37.790473	-122.413914	55ee12f4498e5acd582cd338	El Rincón Yucateco	37.785824	-122.412842	Mexican Restaurant	4bf58dd8d48988d1c1941735	red
2	60	06075011901	37.790473	-122.413914	49fa3c25f964a520dc6d1fe3	Pancho's Salsa Bar & Grill	37.791985	-122.421047	Mexican Restaurant	4bf58dd8d48988d1c1941735	red
3	144	06075012000	37.787965	-122.418527	56b28aa9498ec9031e9f4b0a	Matador	37.788898	-122.411570	Taco Place	4bf58dd8d48988d151941735	purple
4	155	06075012100	37.788807	-122.411888	4a846e6cf964a52094fc1fe3	Taqueria Castillo B2	37.783778	-122.409030	Burrito Place	4bf58dd8d48988d153941735	orange
5	164	06075012100	37.788807	-122.411888	4a7b2e4ff964a52037ea1fe3	Taqueria La Paz	37.783328	-122.413703	Taco Place	4bf58dd8d48988d151941735	purple
6	169	06075012201	37.785846	-122.416353	51b28827498ef0351235c618	El Castillito	37.781961	-122.414737	Mexican Restaurant	4bf58dd8d48988d1c1941735	red
7	172	06075012201	37.785846	-122.416353	49e4f160f964a52074631fe3	Taqueria Cancun	37.781875	-122.410322	Burrito Place	4bf58dd8d48988d153941735	orange
8	185	06075012201	37.785846	-122.416353	4550e681f964a520e43c1fe3	Colibrí Mexican Bistro	37.787109	-122.410533	Mexican Restaurant	4bf58dd8d48988d1c1941735	red
9	190	06075012202	37.785412	-122.419644	4b524b07f964a5205f7527e3	Taqueria El Castillito	37.781569	-122.416897	Burrito Place	4bf58dd8d48988d153941735	orange
10	191	06075012202	37.785412	-122.419644	568ed7e8498ec5e165503bbe	Taqueria Catillito	37.781752	-122.416873	Taco Place	4bf58dd8d48988d151941735	purple

Table 7: Mexican restaurants in cluster 2 of 4 and its surroundings.

There are 11 Mexican restaurants, taco and burrito places in the analysed area. Let us visualize their locations on the map of the area using red markers for Mexican restaurants, purple -for taco places, orange -for burrito places and blue markers for other restaurants.

9. Visualisation of restaurants in cluster 2 of 4 and its vicinity on the map.

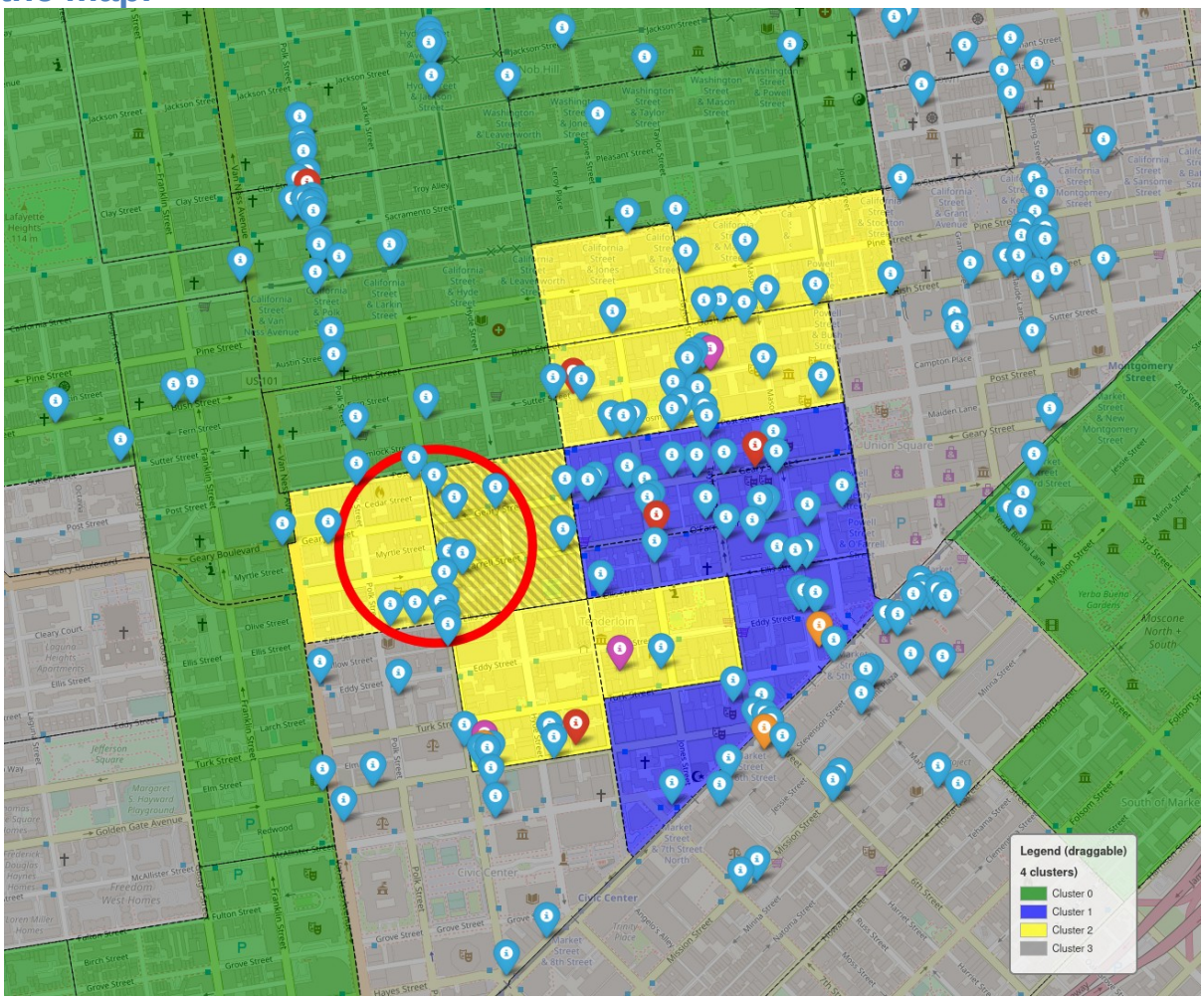

Figure 10: Restaurants in cluster 2 of 4 and its vicinity.

10. Conclusion

After plotting the restaurants on the cluster 2 of 4 map we have a clearer picture. There are two Mexican restaurants within one and two blocks away from the cluster's centre location so this is not the best place for another one. The western part of the cluster is the farthest from the other direct competitors. Taking into account all the previous analyses I propose to locate the Mexican restaurant in area highlighted with the red circle next to the cluster 4 of 9 (marked with the yellow and grey hatch) which is the best in terms of demographic data. This location is within not a far distance away from the rest of the cluster and far enough from the direct competitors. Presence of other restaurants in this area is a sign that there is a substantial footfall with a potential to take over the competitors' clients (certainly this should be analysed more thoroughly).

Of course this just a preliminary demographics and competitors analysis. In order to make a final decision about the location of the restaurant many more factors should by analysed such as:

- visibility

- footfall
- parking
- available venues for rent and their rent rates
- surrounding businesses
- more in-depth analysis of the competitors
- safety and crime rates
- accessibility
- zoning regulations etc.

11. Postscript

After completing the project I downloaded the data again from Foursquare and I found out that exactly in the middle the proposed location for the restaurant a Mexican restaurant has just been created. This means that someone probably made a similar analysis and it is most likely correct. I will gladly observe how the restaurant performs.

Index of Tables

Table of Figures