DATA SCIENCE
FINAL REPORT

San Diego Restaurant site selection, Foursquare/ ZipCodes.com API's

Summary:

Final assignment for Data Science Capstone Course. This report will document the applied tools and techniques of DS, to a specific problem: selecting a general geographical site to set up a new restaurant, providing insight through DS to an aspiring chef-entrepreneur ("Joe"). Joe knows that DS can be of great help for his venture, and asked for help in analyzing the City of San Diego CA, for this purpose. This report documents the application of different DS techniques in order to obtain and analyze available data, leveraging powerful sources such as Foursquare and Zip-Codes.com. This exercise relies heavily on clustering algorithms (K-Means) to help visualize underlying patterns that are not visible to our senses, complementing Joe's views and not only his "instincts".

Author: **Emet H. Flores Jr.**
Date: **April 08, 2020**
E-mail: **emeterio.flores@gmail.com**
GIT: **https://github.com/Emet-DS**

Background for the assignment

This report is for the Data Science Capstone Final assignment. For this assignment in particular, it's required to leverage the skills and tools learned, and primarily to use location data to explore geographical locations, leveraging the Foursquare location data to explore or compare neighborhoods or cities.

In summary the assignment requires:

- Define a problem that can leverage the methods and tools learned and provide recommendations relevant to the problem.
- Describe the data required, has to leverage Foursquare data.
- Apply relevant tools that will help to understand and provide insight to the problem.
- And finally summarize in a report the relevant conclusions (observing relevant methodology).

With this background in mind, this report will focus on a specific problem of identifying a city's characteristics and provide insight to a restaurant entrepreneur named “Joe”. Joe wants to set up a restaurant in a city that has a great number of competitors, and needs to better understand the city's market to reduce risks and have a better chance of success for his restaurant.

Joe understands that Data Science can provide insight that is objective and support his decision making process, for a better starting point for his venture. The main idea is not to rely only on his instincts and personal views, but complement his process with objective data science insights.

This report will follow a simple structure where:

1. A scope is defined.
2. The method is clarified.
3. The insights are gathered and documented.
4. And conclusion and recommendations given.

In the following pages, the previous steps will be developed and documented.

1. Scope Definition

1.1 Defining a challenging problem.

A good friend (we will call him "Joe") has an interest in starting a restaurant in San Diego. He understands that it's a very competitive market, and also has limited monetary resources to set up the restaurant. Joe understands that Data Science (DS) can be very useful for his ventures, since he can discover underlying patterns that are not visible to our general senses. Considering that context, Joe asked if DS could answer three questions:

- What type of restaurant would be a good bet for San Diego?
- What general location would provide reasonable success for that specific type of restaurant?
- Is it possible to consider a location where there is less competition for that type of restaurant?

These questions are a logical concern for an aspiring Chef-Entrepreneur, and DS should help us to answer them through a consistent application of steps (methodology).

1.2 Defining initial data requirements.

We need to understand which restaurants are popular in San Diego. For this purpose, we can use Foursquare data to obtain a list of popular venues, and filter the ones related to restaurants.

A general ranking of restaurant types for the San Diego area, should provide a good starting point, and enable us to validate Joes preferences for a specific type of restaurant (eg. Fast Food, Italian, etc.) that already has potential customers.

Since we need to find a location that would provide reasonable success, we also require to review restaurant popularity by geographic locations, so obtaining references that can provide geographic information will also be necessary: Neighborhoods, Zip Codes, or other data that can be associated with a geographic reference.

Once we find locations with popular restaurants, we will identify if those locations have similar socio demographic attributes. If we find consistent profiles for the popular restaurant locations, we will use that information to search for those attributes where there are no similar restaurants.

1.3 The essential assumption for the data analysis.

To find locations that share socio demographic attributes of the successful restaurants venues, in locations where there is less (or no) competition.

2. General steps (Method)

2.1 Refine the data requirements

2.1.1 Complementing geographic and socio demographic information for our analysis

First we require a way to focus our analysis in the San Diego Metropolitan area. For this purpose, we found that it's relatively easy to find geographic data through Zip Codes. We scrape available information from the internet to obtain a list of Zip Codes for our geographic area of interest. For this purpose, the following steps were taken:

A. A notebook to scrape data, here is the reference:
   a. https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_1_Initial_Data.ipynb
B. Obtaining a list of 107 Zip Codes

	PostalCode	City
0	91901	Alpine
1	91902	Bonita
2	91905	Boulevard
3	91906	Campo
4	91910	Chula Vista
…	…	…

With a target list of Zip Codes, it was possible to leverage **Zip-Code.com API** to obtain additional data attributes, to identify similarities for the target sites.

This source will provide us with the additional data:

- Number of people for the geographic zone
- Number of businesses
- Total area
- And other demographic attributes that can help us.

For this part of the data, the following steps were taken:

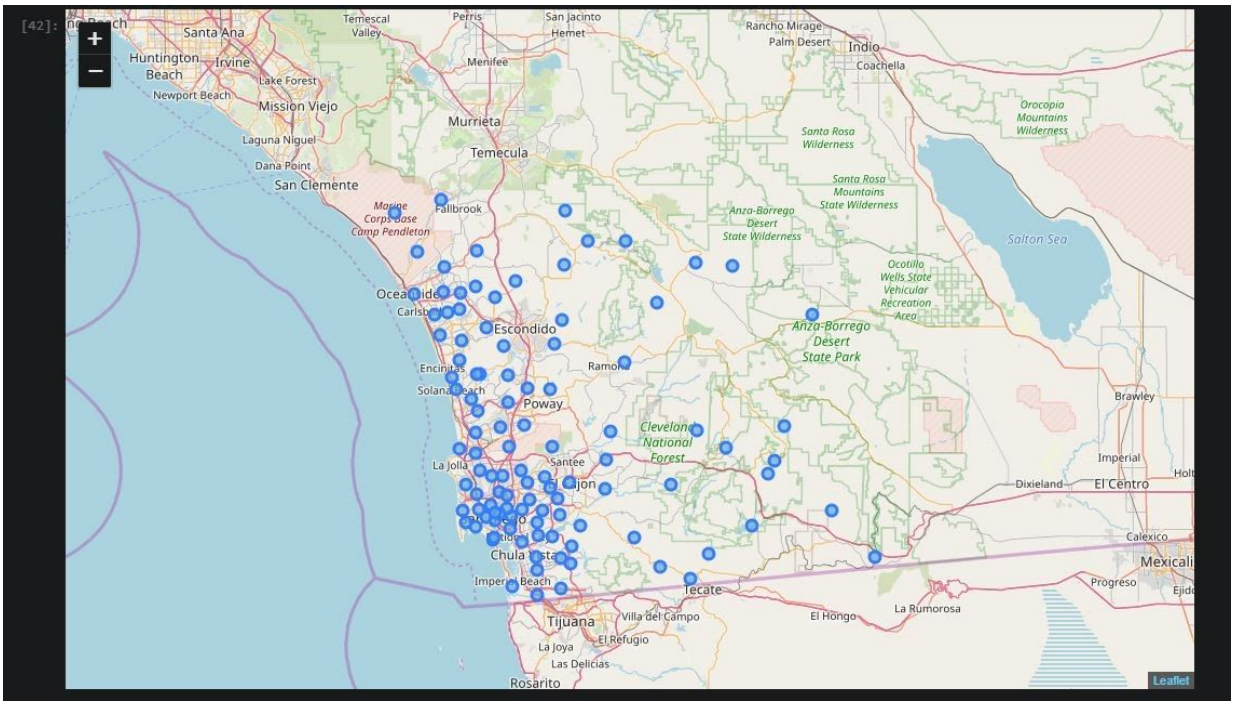
1. Use a notebook to access the **Zip-Code.com API** and obtain the additional data, here is the reference:

    a. https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_0_Initial_Data.ipynb
2. The API provides relevant demographic and geographic data (longitude and latitude), for the analysis:
    a. Here is a the data that the API provides and some preliminary records:

	1	2	3	4
ZipCode	91901	91902	91905	91906
ZipCodePopulation	17403	17653	1700	3627
HouseholdsPerZipcode	6345	5956	632	1254
WhitePop	15466	12379	1335	2828
BlackPop	315	757	76	175
HispanicPop	2644	7326	389	1059
AsianPop	564	2481	25	66
IndianPop	743	272	189	319
HawaiianPop	101	217	20	15
OtherPop	856	2596	133	409
MalePop	8750	8603	965	1959
FemalePop	8653	9050	735	1668
PersonsPerHousehold	2.7	2.96	2.51	2.74
AverageHouseValue	525700	607100	279900	221300
IncomePerHousehold	90397	92759	32819	49919
MedianAge	41.9	43.2	46	37.2
AverageFamilySize	3.1	3.26	3.06	3.23
Latitude	32.789915	32.67855	32.733591	32.702475
Longitude	-116.711202	-117.013671	-116.301864	-116.504617
AreaLand	89.261	8.707	65.814	105.531
AreaWater	0.781	0.297	0.022	0.715
City	ALPINE	BONITA	BOULEVARD	CAMPO
CountyName	SAN DIEGO	SAN DIEGO	SAN DIEGO	SAN DIEGO

With the new dataset, it was possible to visualize the information:

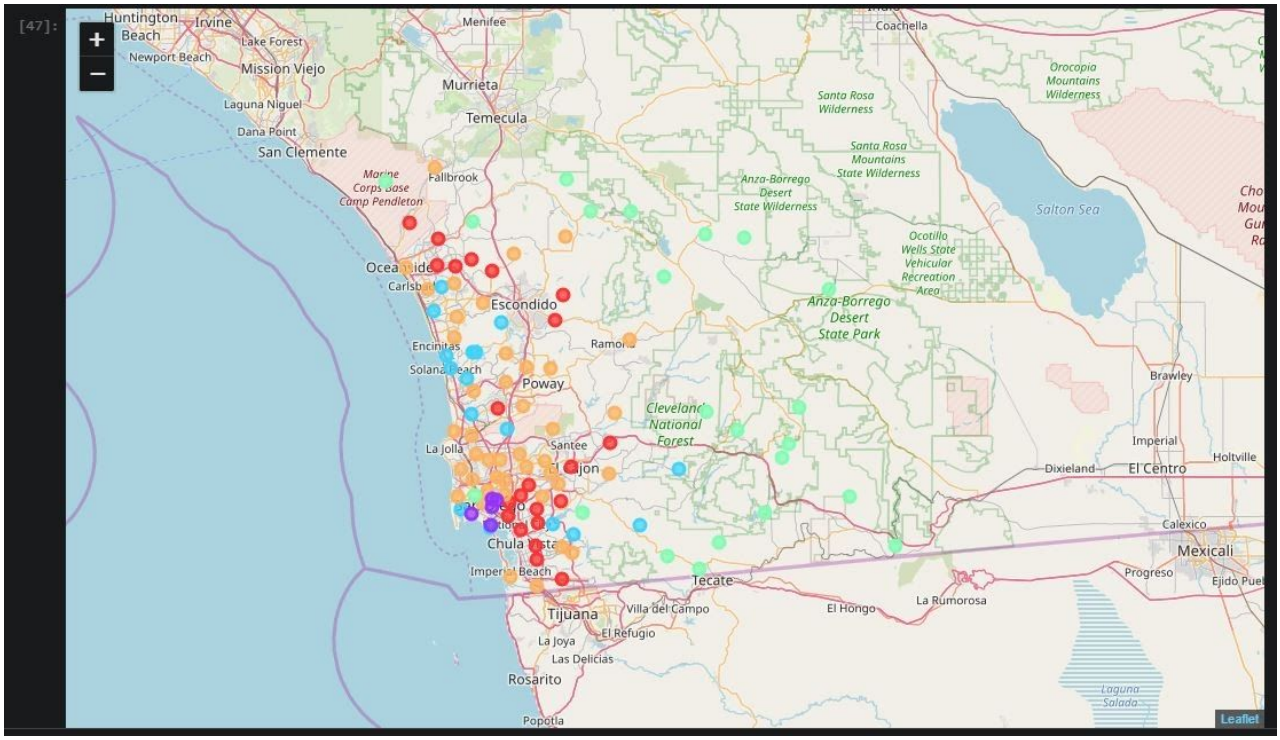
A. In an additional workbook, the data was plotted to have a better grasp of the available information
    a. https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_1_Data_Viz-Copy1.ipynb
B. The initial locations that the data provides can be observed in the following chart:



We can now apply some essential DS techniques to **Zip-Codes.com** API data, in particular:

- Eliminating categorical data
- Normalizing the provided data
- And observing if there are any relevant “Clusters” within the data set using K-Means.

Applying the above steps yielded some initial clusters that provide preliminary insight into the San Diego Metro area, the following Chart maps out these clusters:



The preliminary observations can be summarized as follows:

- ★ Cluster 0 (**Red**): This cluster describes LOWER INCOME households.
- ★ Cluster 1 (**Purple**): This cluster provides no relevant insight, describes government and naval offices.
- ★ Cluster 2 (**Light Blue**): This cluster describes AFFLUENT households.
- ★ Cluster 3 (**Light Green**): This cluster describes low density areas, and are outside our scope of interest.
- ★ Cluster 4 (**Orange**): This cluster is the most common with TYPICAL households.

The data for the Cluster observations supports the general idea. Here follows the relevant cluster tables:

Cluster 0 (Red) - LOWER INCOME households.

We review “Cluster 0” statistics to better understand the cluster.

	count	mean	std	min	25%	50%	75%	max
ZipCodePopulation	22	57,494	13,271	35,125	48,235	56,917	65,342	82,999
HouseholdsPerZipcode	22	17,534	4,461	10,216	14,154	17,458	20,456	26,063
WhitePop	22	32,440	10,121	12,288	25,196	32,076	38,477	50,669
BlackPop	22	5,136	3,633	1,475	2,793	4,290	5,470	16,789
HispanicPop	22	27,074	13,181	10,197	16,998	23,360	35,205	58,816
AsianPop	22	8,137	7,514	1,815	3,481	5,512	11,341	35,353
IndianPop	22	984	201	473	842	1,033	1,128	1,302
HawaiianPop	22	729	302	246	519	728	930	1,292
OtherPop	22	13,519	6,036	5,103	8,463	12,735	17,155	24,733
MalePop	22	28,757	6,654	16,905	24,167	28,441	32,024	41,254
FemalePop	22	28,737	6,748	18,131	24,068	28,502	33,318	43,061
PersonsPerHousehold	22	3.2	0.4	2.6	3.0	3.2	3.4	4.0
AverageHouseValue	22	$ 390,495	$ 53,791	$ 299,400	$ 348,250	$ 369,450	$ 435,575	$ 478,500
IncomePerHousehold	22	$ 56,671	$ 12,394	$ 33,125	$ 48,718	$ 57,793	$ 61,546	$ 92,061
MedianAge	22	32.0	3.4	24.6	30.3	32.9	33.9	38.4
AverageFamilySize	22	3.6	0.3	3.2	3.4	3.7	3.8	4.3

And some raw data for “Cluster 0” to review “AverageHouseValue” and “IncomePerHousehold” data points that describe the cluster.

	ZipCode	ZipCodePopulation	HouseholdsPerZipcode	PersonsPerHousehold	AverageHouseValue	IncomePerHousehold	MedianAge	AreaLand	City
5	91910	75,802	26,063	2.9	$ 424,700	$ 59,371	35.6	12.23	CHULA VISTA
6	91911	82,999	24,622	3.3	$ 361,600	$ 52,274	33.5	11.71	CHULA VISTA
20	91950	60,322	15,869	3.4	$ 345,300	$ 42,942	30.2	7.60	NATIONAL CITY
23	91977	58,368	18,190	3.2	$ 375,400	$ 60,986	33.5	9.67	SPRING VALLEY
35	92020	57,767	19,966	2.8	$ 478,500	$ 52,264	34.7	11.14	EL CAJON
36	92021	65,068	22,649	2.8	$ 363,500	$ 54,154	35.3	29.83	EL CAJON
38	92025	49,978	14,902	3.3	$ 468,800	$ 55,703	31.1	22.04	ESCONDIDO

Cluster 2 (Light Blue): AFFLUENT households.

We review “Cluster 2” statistics to better understand the cluster.

	count	mean	std	min	25%	50%	75%	max
ZipCodePopulation	16	13,106	6,919	1,048	9,307	13,768	17,995	23,575
HouseholdsPerZipcode	16	4,880	2,547	390	3,331	5,382	6,523	9,034
WhitePop	16	10,886	5,953	846	8,537	11,413	15,492	20,030
BlackPop	16	405	465	12	97	248	494	1,815
HispanicPop	16	2,365	2,196	40	656	1,987	2,778	7,326
AsianPop	16	1,282	1,063	50	552	1,000	1,663	4,153
IndianPop	16	192	183	3	90	157	250	743
HawaiianPop	16	82	68	0	32	68	110	217
OtherPop	16	852	781	9	179	780	912	2,596
MalePop	16	6,586	3,605	476	4,616	6,730	8,886	13,532
FemalePop	16	6,520	3,379	572	4,691	7,039	9,152	11,705
PersonsPerHousehold	16	2.6	0.5	1.9	2.3	2.6	2.8	3.7
AverageHouseValue	16	$ 933,063	$ 415,075	$ 525,700	$ 603,400	$ 827,100	$ 1,153,900	$ 2,000,000
IncomePerHousehold	16	$ 109,993	$ 17,790	$ 85,363	$ 99,105	$ 103,781	$ 129,202	$ 141,691
MedianAge	16	41.7	7.5	30.2	38.0	42.2	44.2	59.0
AverageFamilySize	16	3.0	0.3	2.5	2.9	3.0	3.1	3.8

And some raw data for “Cluster 2” to review “AverageHouseValue” and “IncomePerHousehold” data points that describe the cluster.

	ZipCode	ZipCodePopulation	HouseholdsPerZipcode	PersonsPerHousehold	AverageHouseValue	IncomePerHousehold	MedianAge	AreaLand	City
1	91901	17,403	6,345	2.7	$ 525,700	$ 90,397	41.9	89.26	ALPINE
2	91902	17,653	5,956	3.0	$ 607,100	$ 92,759	43.2	8.71	BONITA
8	91914	15,448	4,331	3.6	$ 611,700	$ 131,486	34.2	6.29	CHULA VISTA
15	91935	8,624	2,942	2.9	$ 592,300	$ 98,668	46.7	96.60	JAMUL
28	92007	10,429	4,448	2.3	$ 920,600	$ 105,469	39.9	2.44	CARDIFF BY THE SEA
31	92010	14,382	5,460	2.6	$ 578,400	$ 101,402	39.8	8.08	CARLSBAD

Cluster 4 (Orange): TYPICAL households.

Raw data of “TYPICAL” households in the San Diego Metro area.

	count	mean	std	min	25%	50%	75%	max
ZipCodePopulation	42	37,040	9,848	18,858	27,900	38,703	45,614	53,422
HouseholdsPerZipcode	42	14,455	4,315	6,209	10,988	15,161	17,399	23,349
WhitePop	42	27,976	8,530	13,808	21,415	27,328	34,976	46,690
BlackPop	42	1,703	1,112	227	932	1,395	2,023	4,932
HispanicPop	42	8,132	5,212	2,765	4,342	6,724	10,495	27,436
AsianPop	42	4,905	4,297	723	2,257	2,913	6,054	17,682
IndianPop	42	586	308	252	375	518	721	1,645
HawaiianPop	42	303	133	97	204	281	376	652
OtherPop	42	3,595	2,575	836	1,546	3,099	5,001	10,551
MalePop	42	18,397	4,900	9,446	13,954	19,488	22,603	25,877
FemalePop	42	18,643	5,079	9,411	13,774	19,926	22,765	27,596
PersonsPerHousehold	42	2.6	0.5	1.5	2.2	2.6	2.8	3.9
AverageHouseValue	42	$ 572,345	$ 179,983	$ 326,500	$ 461,775	$ 539,400	$ 647,750	$ 1,276,500
IncomePerHousehold	42	$ 81,270	$ 22,490	$ 38,911	$ 65,508	$ 77,252	$ 88,012	$ 141,926
MedianAge	42	36.6	3.9	29.2	33.7	37.3	39.6	43.0
AverageFamilySize	42	3.1	0.3	2.3	2.9	3.1	3.3	4.2

And some raw data for “Cluster 4” to get a sense of the data points that describe the cluster.

	ZipCode	ZipCodePopulation	HouseholdsPerZipcode	PersonsPerHousehold	AverageHouseValue	IncomePerHousehold	MedianAge	AreaLand	City
7	91913	40,971	12,133	3.4	$ 451,700	$ 87,440	32.0	10.20	CHULA VISTA
9	91915	24,659	7,070	3.5	$ 475,000	$ 106,452	31.8	7.61	CHULA VISTA
13	91932	25,718	9,113	2.8	$ 433,400	$ 49,950	31.5	2.51	IMPERIAL BEACH
16	91941	31,779	12,327	2.6	$ 572,200	$ 79,001	40.7	8.06	LA MESA
17	91942	38,069	16,998	2.2	$ 428,800	$ 56,645	37.5	5.84	LA MESA
18	91945	25,460	8,480	3.0	$ 367,700	$ 60,739	35.0	3.93	LEMON GROVE
29	92008	27,649	11,600	2.3	$ 699,600	$ 81,073	39.3	10.31	CARLSBAD

Cluster Comparison - San Diego demographic profiles.

A side by side comparison is very helpful in understanding the type of demographic profile of the Clusters in this dataset:

	Cluster 0 LOWER INCOME	Cluster 2 AFFLUENT	Cluster 4 TYPICAL
Zip Code Count	22	16	42
	mean	mean	mean
ZipCodePopulation	57,494	13,106	37,040
HouseholdsPerZipcode	17,534	4,880	14,455
WhitePop	32,440	10,886	27,976
BlackPop	5,136	405	1,703
HispanicPop	27,074	2,365	8,132
AsianPop	8,137	1,282	4,905
IndianPop	984	192	586
HawaiianPop	729	82	303
OtherPop	13,519	852	3,595
MalePop	28,757	6,586	18,397
FemalePop	28,737	6,520	18,643
PersonsPerHousehold	3.2	2.6	2.6
AverageHouseValue	$ 390,495	$ 933,063	$ 572,345
IncomePerHousehold	$ 56,671	$ 109,993	$ 81,270
MedianAge	32.0	41.7	36.6
AverageFamilySize	3.6	3.0	3.1

The side by side comparison will be useful later on, since it will facilitate how we define simple assumptions of the population profile, and its affinity to a specific type of restaurant venue. Some initial observations are listed below:

- **TYPICAL** represents approximately 51% of this subset population.
    - White population has a large representation
    - And is in the middle in terms of income
- **LOWER INCOME** represents approx. 42% of this subsets population.
    - Hispanic population is almost the same size of White population.
    - The income is clearly in the lower range for San Diego Metro area
- **AFFLUENT** with approx. 7% of this subset.
    - White population has a large representation.
    - The income is clearly in the upper range for the San Diego Metro area.

These observations will help later on when we reference the demographic information vs. the data from popular venues from Foursquare.

2.1.2 Gather information of popular restaurants for San Diego

In order to gather information of popular restaurants in the San Diego Metro area, we will use the previous information of "target" ZipCode, and use Foursquare to rank popular restaurant venues. For this purpose, the following steps were used:

A. Reuse code from the Labs, and obtain venue data, for this a specific notebook was created:
   a. https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_2_FourSqr_Data.ipynb

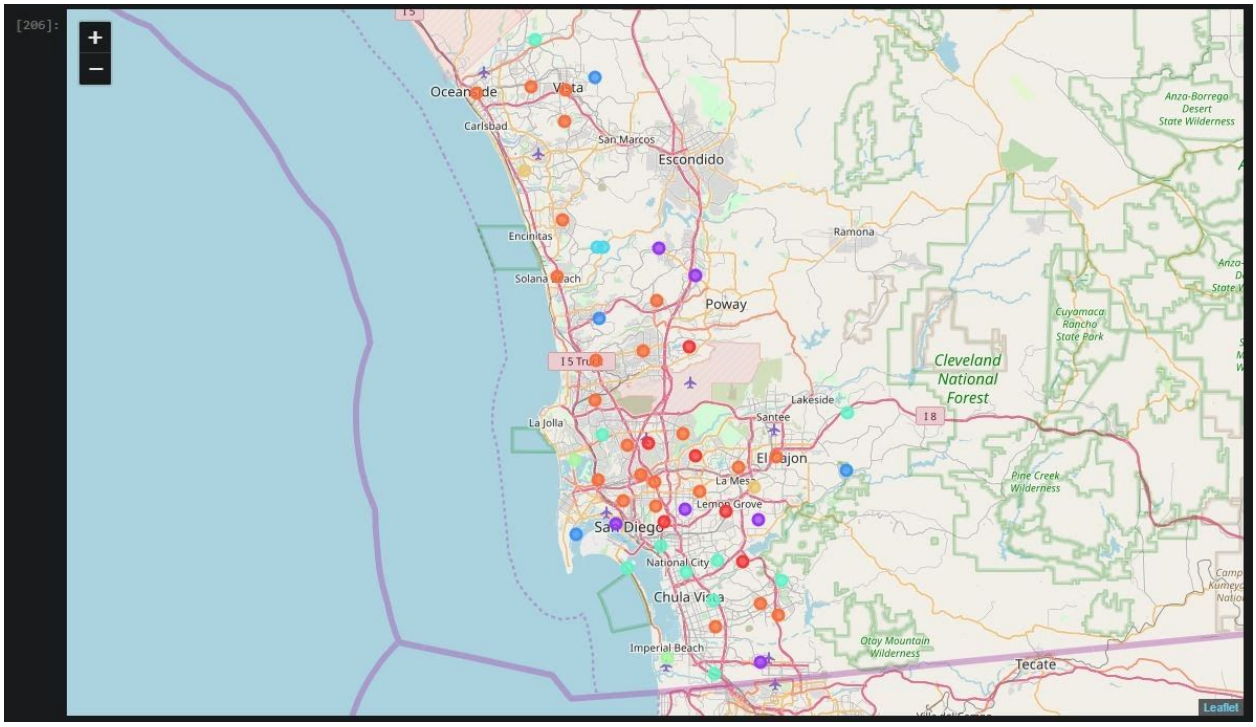
B. The global top 10 Venue Category we obtained for the target zip codes is:
   a. Mexican Restaurant 88 occurrences.
   b. Fast Food Restaurant 49 occurrences.
   c. American Restaurant 31 occurrences.
   d. Chinese Restaurant 28 occurrences.
   e. Sushi Restaurant 27 occurrences.
   f. Seafood Restaurant 23 occurrences.
   g. Italian Restaurant 20 occurrences.
   h. Restaurant 16 occurrences.
   i. Thai Restaurant 14 occurrences.
   j. Vietnamese Restaurant 14 occurrences.

C. A ranking function for each Zip Code was applied, gathering relevant information to determine if there are any underlying patterns through clusters. Here is an example of the rankings:

Index	2	5	6	7	8
ZipCode	91,902	91,910	91,911	91,913	91,914
Latitude	32.67855	32.635694	32.607009	32.632497	32.657754
Longitude	-117.013671	-117.052566	-117.050286	-116.991164	-116.963414
City	BONITA	CHULA VISTA	CHULA VISTA	CHULA VISTA	CHULA VISTA
CountyName	SAN DIEGO	SAN DIEGO	SAN DIEGO	SAN DIEGO	SAN DIEGO
1st Most Common Venue	Mexican Restaurant	Mexican Restaurant	Asian Restaurant	Sushi Restaurant	Fast Food Restaurant
2nd Most Common Venue	Vietnamese Restaurant	Fast Food Restaurant	Filipino Restaurant	Fast Food Restaurant	Mongolian Restaurant
3rd Most Common Venue	Japanese Restaurant	Chinese Restaurant	Mexican Restaurant	Mexican Restaurant	Mexican Restaurant
4th Most Common Venue	Indian Restaurant	Greek Restaurant	Vietnamese Restaurant	Vietnamese Restaurant	Vietnamese Restaurant
5th Most Common Venue	Hawaiian Restaurant	Japanese Restaurant	Indian Restaurant	Hawaiian Restaurant	Caribbean Restaurant

6th Most Common Venue	Halal Restaurant	Italian Restaurant	Hawaiian Restaurant	Halal Restaurant	Chinese Restaurant
7th Most Common Venue	Greek Restaurant	Comfort Food Restaurant	Halal Restaurant	Greek Restaurant	Comfort Food Restaurant
8th Most Common Venue	French Restaurant	Cuban Restaurant	Greek Restaurant	French Restaurant	Cuban Restaurant
9th Most Common Venue	Filipino Restaurant	Eastern European Restaurant	French Restaurant	Filipino Restaurant	Eastern European Restaurant
10th Most Common Venue	Fast Food Restaurant	Caribbean Restaurant	Fast Food Restaurant	Eastern European Restaurant	Italian Restaurant

D. We can now run clustering algorithms on this subset of information, and review further if there are any underlying insight that can be derived:



a. Cluster 0 (**Red**): This cluster describes a combination of: Mexican , Vietnamese , Japanese and Indian .
b. Cluster 1 (**Purple**): This cluster describes: American and Seafood..
c. Cluster 2 (**Blue**): This cluster describes: American, Fast food and Indian.
d. Cluster 3 (**Light Blue**): This small cluster describes a combination: General Restaurant, Vietnamese , Fast food.
e. Cluster 4 (**Green**): This cluster describes: Fast food , Mexican and Chinese.
f. Cluster 5 (**Light Green**): This small 2 zip codes cluster describes: Italian , Fast food and Indian.
g. Cluster 6 (**Yellow**): This small cluster describes: Mexican, Chinese, Vietnamese.
h. Cluster 7 (**Orange**): This cluster describes: Mexican, Sushi, Fast food venues, tending towards a **larger variety of options**.

2.1.3 Refine our understanding of the specific "type" of popular restaurants

A comparison chart for each cluster is useful to better grasp these clusters, here follows a chart for each of the 8 clusters:

Rest. Venue Cluster 0

SUM of ZipCode Population		*2nd Most Common Venue*			
Cluster Labels	*1st Most Common Venue*	Vietnamese Restaurant	Japanese Restaurant	American Restaurant	Grand Total
0	Mexican Restaurant	119,167	26,823	26,317	172,307
0 Total		119,167	26,823	26,317	172,307

Rest. Venue Cluster 1

SUM of ZipCode Population		*2nd Most Common Venue*					
Cluster Labels	*1st Most Common Venue*	American Restaurant	Seafood Restaurant	Vietnamese Restaurant	Chinese Restaurant	Thai Restaurant	Grand Total
1	American Restaurant		79,708		47,490	39,337	166,535
	Seafood Restaurant	37,095		58,368			95,463
	Thai Restaurant	69813					69,813
1 Total		106,908	79,708	58,368	47,490	39,337	331,811

Rest. Venue Cluster 2

SUM of ZipCode Population		*2nd Most Common Venue*		
Cluster Labels	*1st Most Common Venue*	Fast Food Restaurant	Asian Restaurant	Grand Total
2	American Restaurant	139,192	19,330	158,522
2 Total		139,192	19,330	158,522

Rest. Venue Cluster 3

SUM of ZipCode Population		2nd Most Common Venue	
Cluster Labels	1st Most Common Venue	Vietnamese Restaurant	Grand Total
3	Restaurant	10,583	10,583
3 Total		10,583	10,583

Rest. Venue Cluster 4

SUM of ZipCode Population		2nd Most Common Venue				
Cluster Labels	1st Most Common Venue	Mexican Restaurant	Fast Food Restaurant	Chinese Restaurant	Mongolian Restaurant	Grand Total
4	Fast Food Restaurant	199,925		23,575	15,448	238,948
	Mexican Restaurant		192,190			192,190
	Vietnamese Restaurant			35,125		35,125
4 Total		199,925	192,190	58,700	15,448	466,263

Rest. Venue Cluster 5

SUM of ZipCode Population		2nd Most Common Venue	
Cluster Labels	1st Most Common Venue	Fast Food Restaurant	Grand Total
5	Italian Restaurant	71,505	71,505
5 Total		71,505	71,505

Rest. Venue Cluster 6

SUM of ZipCode Population		*2nd Most Common Venue*		
Cluster Labels	*1st Most Common Venue*	Chinese Restaurant	Vietnamese Restaurant	Grand Total
6	Mexican Restaurant	31,779		31,779
	Chinese Restaurant		22,405	22,405
6 Total		31,779	22,405	54,184

Rest. Venue Cluster 7

Special Note: For cluster 7 the format was adjusted to better fit this report, 1st and 2nd Common Venue categories were transposed.

SUM of ZipCode Population		*1st Most Common Venue*							
Cluster Labels	*2nd Most Common Venue*	Fast Food	Sushi	Mexican	Asian	American	Vietnamese	French	Grand Total
7	Mexican	62,728	82,278						145,006
	Fast Food		84,699			59,233			143,932
	Seafood	73,343	49,121						122,464
	Vietnamese	58,560						44,414	102,974
	Filipino				82,999				82,999
	Hawaiian			31,680			45,096		76,776
	Restaurant			57,767					57,767
	Thai	51,536							51,536
	Sushi			43,122					43,122
	American	36,975							36,975
	Japanese			27,404					27,404
	Italian			25,341					25,341
	Indian					4,179			4,179
7 Total		283,142	216,098	185,314	82,999	63,412	45,096	44,414	920,475

Feedback of restaurant type venues with Joe.

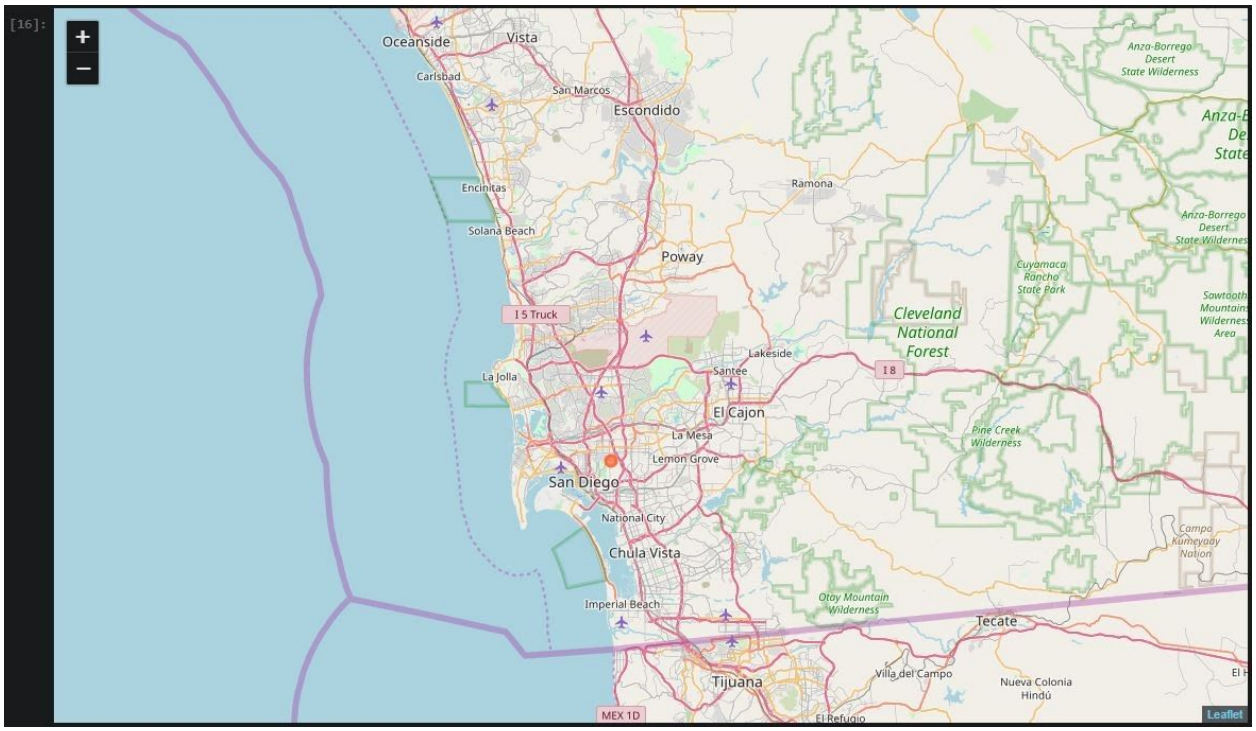
Work session with Joe, and upon some preliminary insight review, Joe understands that his **French cuisine training** is a big factor, and should be taken into consideration, even though the venue type has little representation of this type of venue.

Nonetheless, we know by a simple search, that Zip Code **92104**, has “French Restaurant” as its most popular venue. We also know that Zip Code 92104 fits within the cluster of venues number 7, and is also clustered in the demographic cluster number 4.

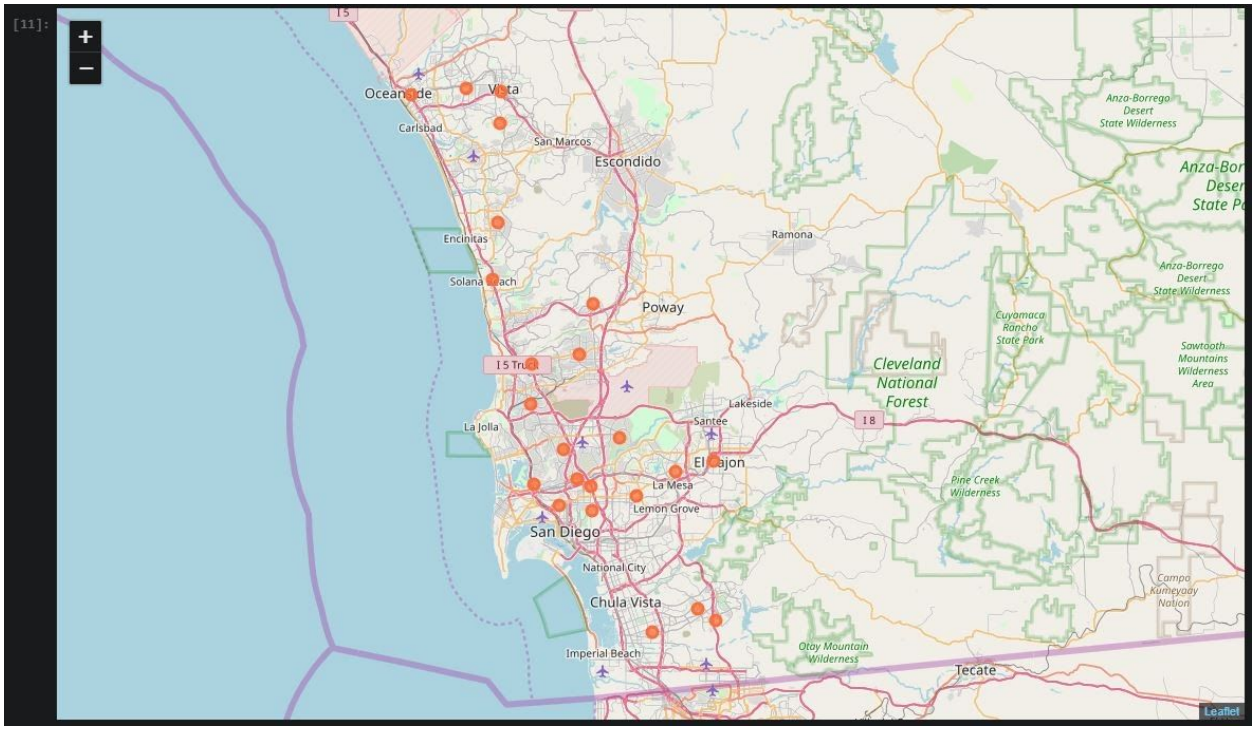
2.2. Apply insight to find ideal geographic target zones

With the preliminary analysis, we can narrow down the effort of selecting a site for Joe’s restaurant. Let's review some of the important observations and visualize the data:

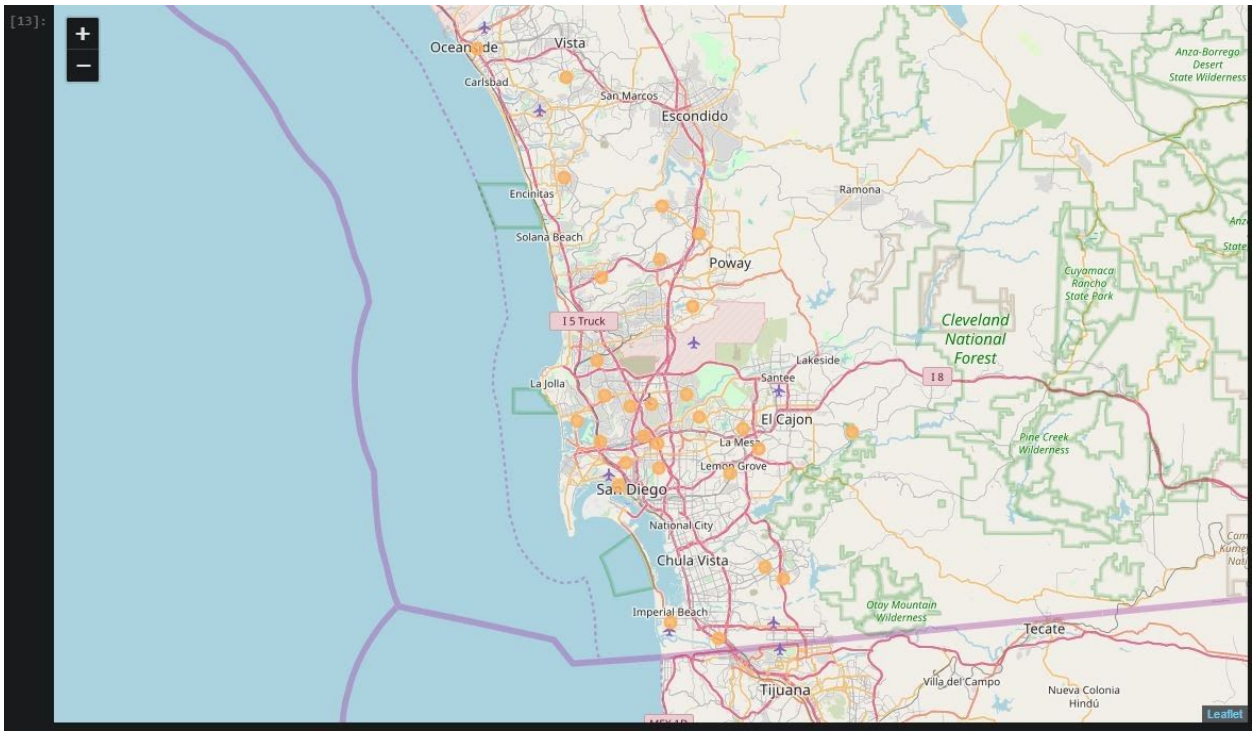
- Identify the location of the successful French Restaurant venue



- Review the cluster that shares characteristics of popular venues that we obtained from Foursquare:

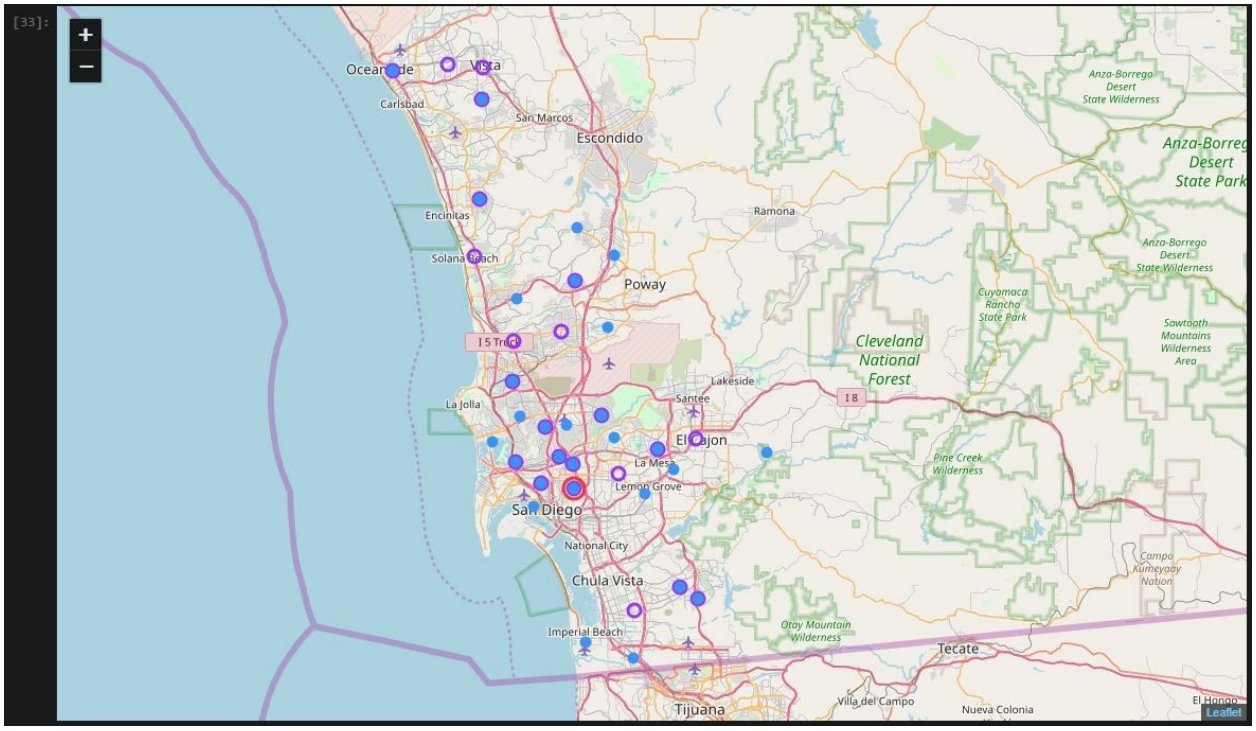


- And also review the locations that share the same demographic profile for the successful French Venue, that we obtained form Zip-Codes.com:



The maps show there is some overlap between the demographic profile, and the cluster of scuseful venues that include the "**French Restauran**t" Type.

Fortunately, we can combine this information into a single map, so we can better visualize the data:

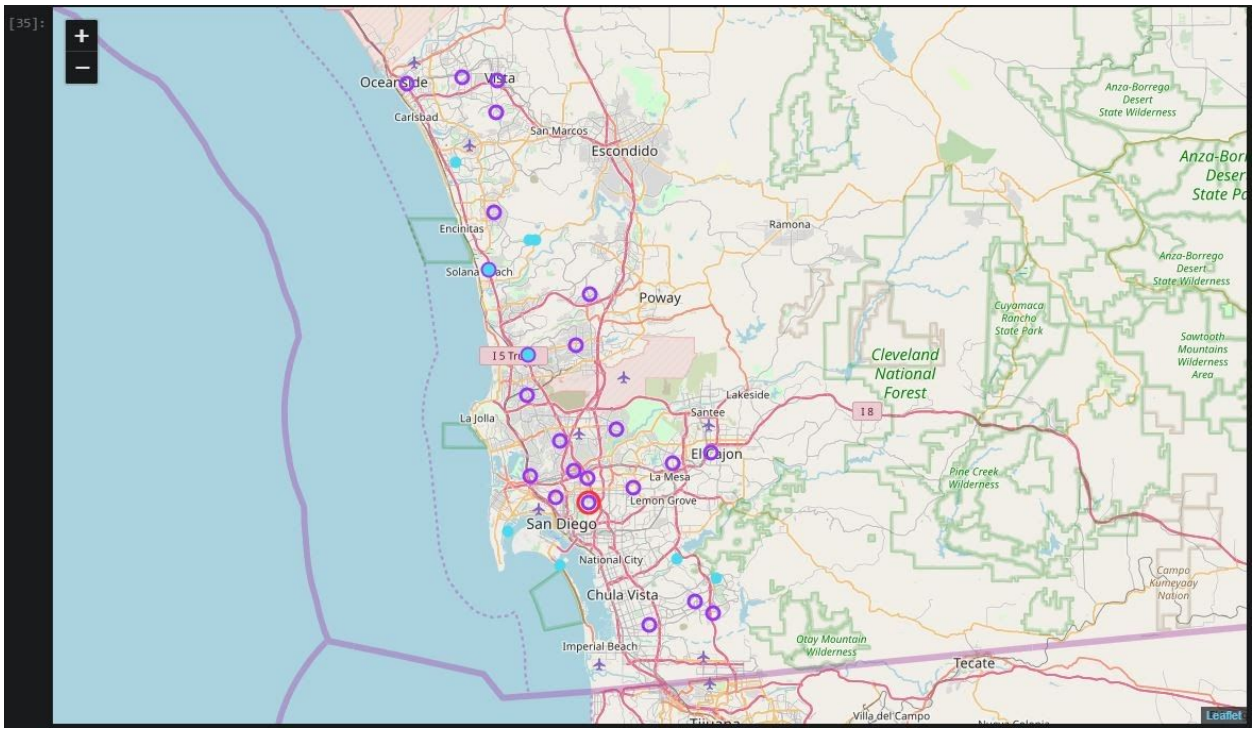


- ★ The Red Circle represents the location of the most popular French Restaurant venue, obtained with the Foursquare data.
- ★ Blue circles represent the demographic cluster that is found in the location of the popular “French Restaurant”.
- ★ Purple circles represent “restaurant venue types” that share characteristics with the popular “French Restaurant”.

Since there is relevant overlap between the classified of the data, there are several competing ideas:

1. Overlapping demographic and popular venue clusters (purple+blue circles), already have confirmed (“TYPICAL”) customers that appreciate diversity in restaurant venues.
2. Popular venue clusters that don't overlap (purple circles only), would only require review for the “AFFLUENT” profile.
3. The rest of the sites with “TYPICAL” customer profiles (blue circles only), have other strong venue preferences.

Further review of the overlap information with the “AFFLUENT” profile, we can visualize the data in the following map:



★ There are two zip codes that overlap (purple + light blue circles): 92075, 92121. These two zip codes would also represent relevant sites.

Its relevant to comment, that we reused most of the techniques learned from various courses and finalizing in this Capstone report, so the above charts and cluster analysis can be reviewed in the following notebook:

- https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_3_Cluster_Info.ipynb

3. Final observations and highlights

3.1 Insight review with Joe

3.1.1 Review of geographic data obtained from Foursquare and Zip-Codes.com data.

Our global review of Foursquare data for top restaurant venue types in San Diego, yielded the following top 10 list:

Restaurant Venue Type	Occurrence
Mexican Restaurant	88
Fast Food Restaurant	49
American Restaurant	31
Chinese Restaurant	28
Sushi Restaurant	27
Seafood Restaurant	23
Italian Restaurant	20
Restaurant	16
Thai Restaurant	14
Vietnamese Restaurant	14

Preliminary review of the data sources, provides some insight into the characteristics of the San Diego Metro area, yielding relevant classifications (“Clusters”) for:

- Popular Restaurant Venue Types, by zip code, from the Foursquare API.
- Demographic Profiles, by zip code, from Zip-Codes.com API.

Clusters for the popular Venue Types are:

SUM of ZipCodePopulation	*Cluster Labels*								
1st Most Common Venue	0	1	2	3	4	5	6	7	Grand Total
Mexican	172,307				192,190		31,779	185,314	581,590
Fast Food					238,948			283,142	522,090
American		166,535	158,522					63,412	388,469
Sushi								216,098	216,098
Seafood		95,463							95,463
Asian								82,999	82,999
Vietnamese					35,125			45,096	80,221
Italian						71,505			71,505
Thai		69,813							69,813
French								44,414	44,414
Chinese							22,405		22,405
Restaurant				10,583					10,583
Grand Total	**172,307**	**331,811**	**158,522**	**10,583**	**466,263**	**71,505**	**54,184**	**920,475**	**2,185,650**

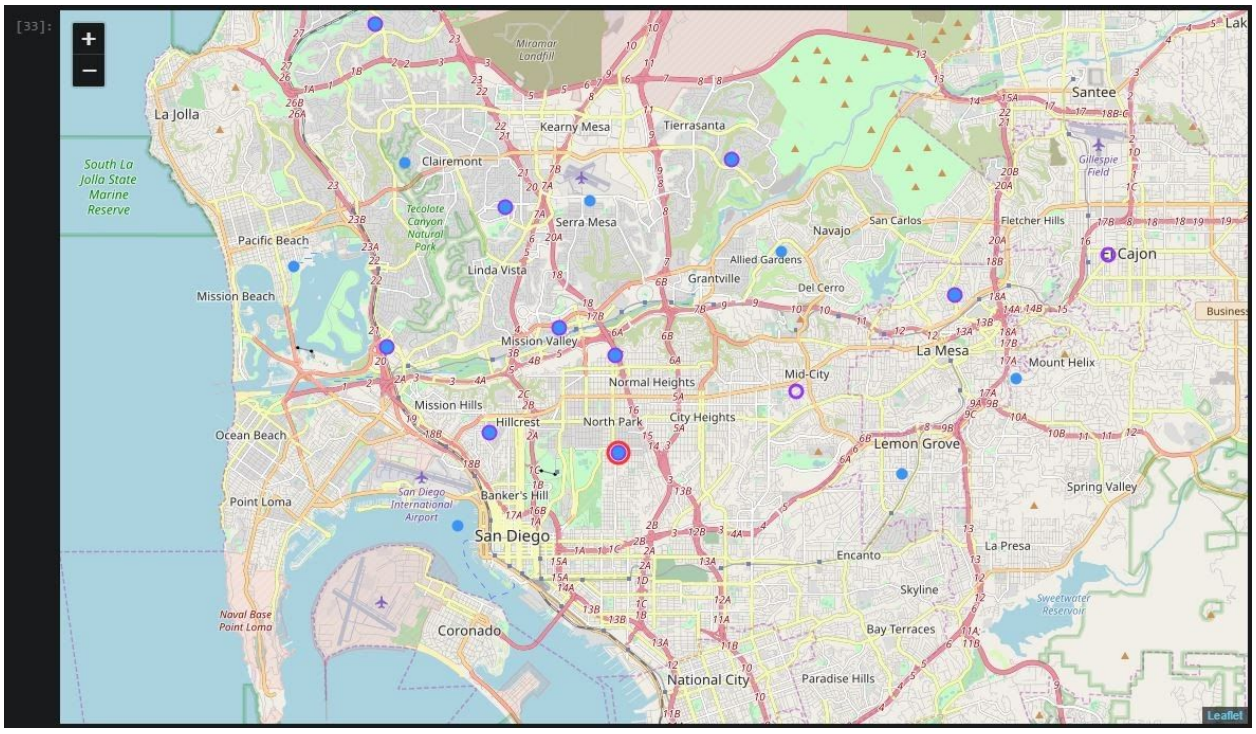
Clusters for the additional demographic data are described in the following comparison chart:

	Cluster 0 LOWER INCOME	**Cluster 2 AFFLUENT**	**Cluster 4 TYPICAL**
Zip Code Count	22	16	42
	mean	mean	mean
ZipCodePopulation	57,494	13,106	37,040
HouseholdsPerZipcode	17,534	4,880	14,455
WhitePop	32,440	10,886	27,976
BlackPop	5,136	405	1,703
HispanicPop	27,074	2,365	8,132
AsianPop	8,137	1,282	4,905
IndianPop	984	192	586
HawaiianPop	729	82	303
OtherPop	13,519	852	3,595
MalePop	28,757	6,586	18,397
FemalePop	28,737	6,520	18,643
PersonsPerHousehold	3.2	2.6	2.6
AverageHouseValue	$ 390,495	$ 933,063	$ 572,345
IncomePerHousehold	$ 56,671	$ 109,993	$ 81,270
MedianAge	32.0	41.7	36.6
AverageFamilySize	3.6	3.0	3.1

Upon review with Joe, new information was provided, and it was necessary to adjust the analysis to take into consideration Joe's experience in "**French cuisiene**" .

3.1.2 Review the selected restaurant type by Joe

Fortunately, the data did contain a location with a highly popular "French Restaurant" venue. It is possible to visualize this location within its classification of popular restaurant venues:



- ★ Red circle represents the most Popular "French Restaurant" Venue location.
- ★ Purple + Blue circles represent the same profile of restaurant type venue and same profile of demographic customers.

The location with the popular "French Restaurant" shares characteristics with two types of classification clusters:

- ★ Cluster 7 of popular venue types from Foursquare
  - ○ This cluster describes a variety of restaurant venues
  - ○ Where variety is part of the profile that defines this classification
- ★ Cluster 4 of demographic profiles from Zip-Codes.com
  - ○ This cluster describes the "TYPICAL" San Diego household
  - ○ With a relevant income level, that enables people to buy and enjoy different types of restaurants.

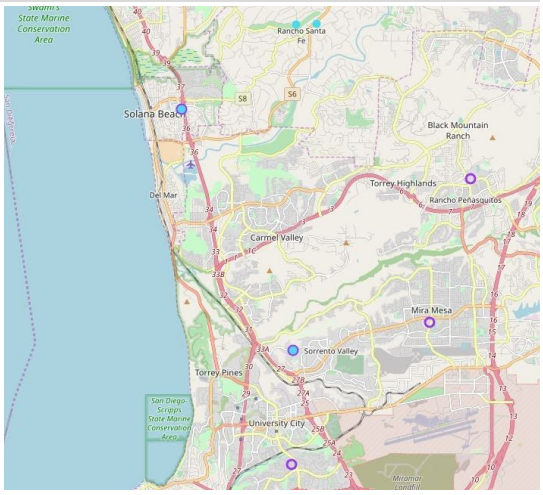
These specific locations where the both profiles overlap, are listed below:

ZipCode	*City*	SUM of ZipCodePopulation	
92129	SAN DIEGO	51,536	
92024	ENCINITAS	49,121	
92111	SAN DIEGO	45,096	
92104	**SAN DIEGO**	**44,414**	**Red Circle**
92122	SAN DIEGO	43,728	
91913	CHULA VISTA	40,971	
92054	OCEANSIDE	40,375	
91942	LA MESA	38,069	
92116	SAN DIEGO	31,680	
92103	SAN DIEGO	31,066	
92124	SAN DIEGO	30,443	
92081	VISTA	27,404	
92110	SAN DIEGO	25,341	
91915	CHULA VISTA	24,659	
92108	SAN DIEGO	18,858	
Grand Total		**542,761**	

An additional list is relevant to the insight, “AFFLUENT” profile is important to includes, since it also has an overlap of restaurant venues, and should also enjoy diversity of restaurant venues:

ZipCode	*City*	SUM of ZipCodePopulation
92075	SOLANA BEACH	12,056
92121	SAN DIEGO	4,179
Grand Total		**16,235**

Observe these two loc. in the map:



3.1.4 Observations and highlights, review with Joe.

Lets review the initial questions asked by Joe to answer through a DS approach:

- What type of restaurant would be a good bet for San Diego?
- What general location would provide reasonable success for that specific type of restaurant?
- Is it possible to consider a location where there is less competition for that type of restaurant?

In order to answer the questions, we defined a method that could provide answers through a Data Science approach:

- Clearly define the problem
- Gather available data
- And apply Data Science tools and or techniques that could provide insight to answer Joes questions.

The preliminary observations we gathered throughout the application of the method, can be summarized as follows:

★ There is available data for the San Diego Metro area through its Zip Codes, that we can use to answer some of these questions through a DS approach.
  - Foursquare API data to determine popular restaurants venues.
  - Zip-Codes.Com API data to gather demographic data to better understand potential customers near popular restaurant venues.

★ Using a clustering algorithm (K-Means), The San Diego Metro area can essentially be classified into three (3) demographic clusters of interest.
  - Cluster 0 - LOWER INCOME households.
  - Cluster 2 - AFFLUENT households.
  - Cluster 4 - TYPICAL households.

★ Using a combination of techniques to rank information from Foursquare data, we can know which types of restaurants are the most popular:
  - **Mexican Restaurants** being twice as popular than **Fast Food Restaurants**, and **American Restaurants** come in as the thirds most popular.
  - And it also provided a way to classify popular restaurant venues by theri zip code, Using a clustering algorithm (K-Means), gathering restaurant venue preferences that people have by zip code (8 particular clusters).

★ The challenge in this case for our friend Joe, was to consider his cuisine expertise - **French Cuisine**.

- ★ A quick search through the Foursquare ranking exercise, provided insight into a very popular French Restaurant Venue, and we know that:
    - This particular venue, shares classification characteristics with:
        - Popular restaurant venue cluster 7 - Mexican, Sushi, Fast food venues, tending towards a larger variety of restaurant options.
        - And shares demographic profiles of TYPICAL households (Cluster 2).
    - There was another overlap of shared characteristics with another demographic profile, AFFLUENT households.

The insight gathered was reviewed with Joes, and the following feedback was noted:

1. Joe now has a better idea of its potential customer base, and that it appreciates having multiple options to enjoy food venues.
2. The overlapping of two demographic household profiles is also useful for Joe, since it serves as a business assumption that can define his restaurant:
    a. Upscale French if it's possible to set up a location in the AFFLUENT zip codes.
    b. Or multiple site options if the concept is catered to the TYPICAL household profile.

4. Conclusions

4.1 Final comments

4.1.1 Answers to Joe's questions

Regarding Joe's questions, a Data Science approach provide some answers:

- What type of restaurant would be a good bet for San Diego?
    - A: Since it has a large variety of competing restaurant venues, it should be answered by understanding Joe's restaurant experience.
    - A: Some restaurant type venues also have more competitors, so differentiating from: Mexican Restaurants, Fast Food or American Restaurants, should be taken into consideration.
- What general location would provide reasonable success for that specific type of restaurant?
    - A: Considering that Joe's experience is in French cuisine, there are two relevant options to consider.

        - For an upscale French restaurant, there are two locations that make sense.
        - Or multiple locations, that cater to a TYPICAL household profile.
- Is it possible to consider a location where there is less competition for that type of restaurant?
    - A: Since there are very limited French Restaurant venues for San Diego, it is possible to select locations that have no other French venues within the same zip code, simply by avoiding: 91915, 92103, **92104**, 92116.
    - A: A good option would be to consider the AFFLUENT zip codes, since they have no French venue presence whatsoever.

4.1.2 Final comments

The application of the defined method provides relevant insight, so we can recommend to Joe that he set up an upscale French Restaurant in zip codes 92075 or 92121.

If an upscale French restaurant is not possible (investment restrictions), then further review of the other locations is recommended, to narrow down the decision with additional data.

4.1.3 Next steps?

Data science can provide new insight that is generally unavailable to how we see and understand the world. Our experiences can forge opinions and views that might be skewed or incorrect, but Data Science can provide new ways to understand our reality.

Applying these tools and techniques must complement our good judgment, and help us make better decisions. This becomes very relevant if these decisions have a significant impact in the investment we make, or the risks we take.

For Joe's particular case, there are additional analysis that would help:

- ❏ Evaluate fixed costs at shop level, for available locations.
- ❏ Evaluate characteristics of potential locations (shop level).
- ❏ Review complementing venues of potential sites (shop level).
- ❏ Rank potential locations.

Finally, there is no analysis or data that can in practice replace "actions". "**Doing"** is a sure way of advancing any enterprise, even if we make mistakes (but learn from them).