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

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

Index

Introduction - Final assignment, Data Science Capstone Course

This presentation will highlight the applied tools and techniques of DS, to a specific problem:

➔ **Selecting a general geographical site to set up a new restaurant.**

We will provide insight through DS to an aspiring chef-entrepreneur and good friend: 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 that purpose, data need to be obtained and analyzed, leveraging powerful sources:**Foursquare, Zip-Codes.com.**

Analysis will rely heavily on clustering algorithms (K-Means) to help visualize underlying patterns, and help Joe to improve his overall knowledge and views of San Diego, and <u>not only rely on his “instincts”.</u>

Background of this assignment

In this presentation, a summary and showcase of the final assignment topics are presented. In essence, how to leverage Data Science ("DS") skills and tools learned, in addition to the use of location data (Foursquare and Zip-Code.com), to explore and compare neighborhoods and or cities.

In brief this presentation will cover:

1. Problem definition that can leverage DS methods and tools learned.
2. Describe the data required, geographical data is essential (Foursquare data).
3. Apply relevant tools that will help to understand and provide insight to the problem, including a structured method.
4. And summarize the relevant conclusions.

All of the above will follow an evolution of steps (**method**), so we can better understand the questions and data needed:

Define Scope > Refine Data Requirements > Apply Insight > Final observations > Recommendations

Method - Define Scope

Define Scope - the problem and questions.

A good friend "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.

So Joe asked some straightforward questions:

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

An important assumptions is considered for this analysis:

Find locations that **share the similar demographic attributes** of the target ("successful") restaurant venue type, in locations where there is less (or no) competition.

Define Scope - The supporting data required

As an essential step, we need to gather or acquire data that can generate insight through the application of DS tools and techniques.

So what basic data requirements can be initially defined by Joes questions?

1	What type of restaurant would be a good bet for San Diego?	A ranking of popular San Diego restaurant venues, a geographical classification will be necessary. Potential customer profile information by **geographical** classification will also be useful.
2	What general location would provide reasonable success for that specific type of restaurant?	Understanding geographically if there are underlying classifications for certain types of **popular restaurants**. Understanding geographically if there are underlying classifications for certain types of **customer profiles**.
3	Is it possible to consider a location where there is less competition for that type of restaurant?	Understanding geographically where competing restaurant venues are located.

Method - Refine Data Requirements

Refine Data Requirements - Geographic data

A quick review of geographic data sources, provided an alternative that compiled useful data for our analysis, and was readily available:

- ❏ **Zip-Codes.com**: This source has detailed demographic information classified by zip code, and has an API that enables us to get the data we need to review the San Diego Metro area.

Nonetheless, a target list of zip codes is required. For this purpose we use some simple techniques to "scrape" simple data from the web and gather a simple zip code list for the San Diego Metro Area.

A jupyter notebook was used to scrape this data, here is the link for those interested in this technique:

https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_1_Initial_Data.ipynb

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

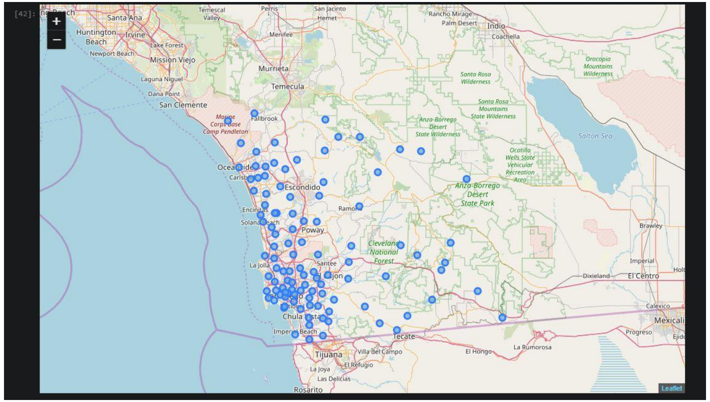
Refine Data Requirements - Geographic data, Zip-Codes.com

With a list of target zip codes, it was posible to get goegraphical information from the **Zip-Codes.com API**. Two important things are provided by this source:

1. Geographical references that will help us later on (**longitude** and **latitude** of each zip code).
2. **Demographic information** of the potential customers, **associated to a geographical reference**.

Sample record =
ZipCode: 91901, ZipCodePopulation: 17403, HouseholdsPerZipcode: 6345, WhitePop: 15466, BlackPop: 315, HispanicPop: 2644, AsianPop: 564, IndianPop: 743, HawaiianPop: 101, OtherPop: 856, MalePop: 8750, FemalePop: 8653, PersonsPerHousehold: 2.7, AverageHouseValue: 525700, IncomePerHousehold: 90397, MedianAge: 41.9, AverageFamilySize: 3.1, **Latitude**: 32.789915, **Longitude**: -116.711202, AreaLand: 89.261, AreaWater: 0.781, City: ALPINE, CountyName: SAN DIEGO.

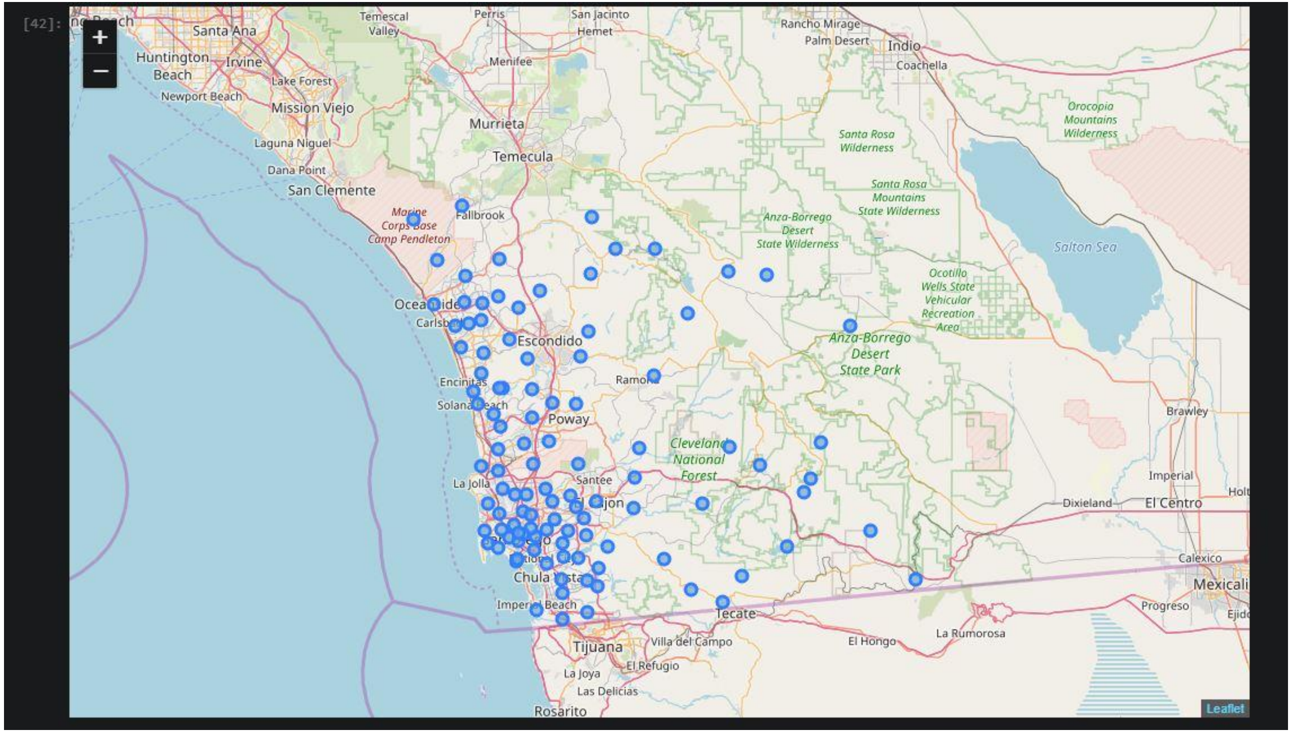
Geographic representation of target zip codes:



Here are some links of the jupyter notebooks that were used:
https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_0_Initial_Data.ipynb
And for the graph:
https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_1_Data_Viz-Copy1.ipynb

Refine Data Requirements - Geographic data, Zip-Codes.com

A better view of the geographical data points gathered from the **Zip-Codes.com API:**



And available data:

ZipCode,

ZipCodePopulation
HouseholdsPerZipcode

WhitePop
BlackPop
HispanicPop
AsianPop
IndianPop
HawaiianPop
OtherPop

MalePop
FemalePop

PersonsPerHousehold

AverageHouseValue
IncomePerHousehold

MedianAge
AverageFamilySize

Latitude
Longitude
AreaLand
AreaWater
City
CountyName

Refine Data Requirements - Geographic data, Zip-Codes.com

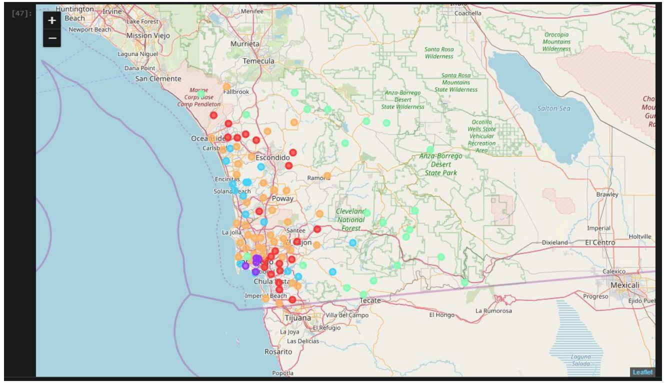
With the avialable Zip-Codes.com data, we can now apply some essential DS techniques, in particular:

- ❏ Eliminating categorical data.
- ❏ Normalizing the data.
- ❏ And determine if there are any relevant "Clusters" within the data set using K-Means.

Preliminary observations can be summarized:

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.

Geographic representation of clusters:



Here is the link of the jupyter notebook that was used for the K-Means and visualization:
https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_1_Data_Viz-Copy1.ipynb

Refine Data Requirements - Geographic data, Zip-Codes.com

A better view of the geographical representation of the clusters from the **Zip-Codes.com** data set:

Useful profiles (“Target zip codes”):

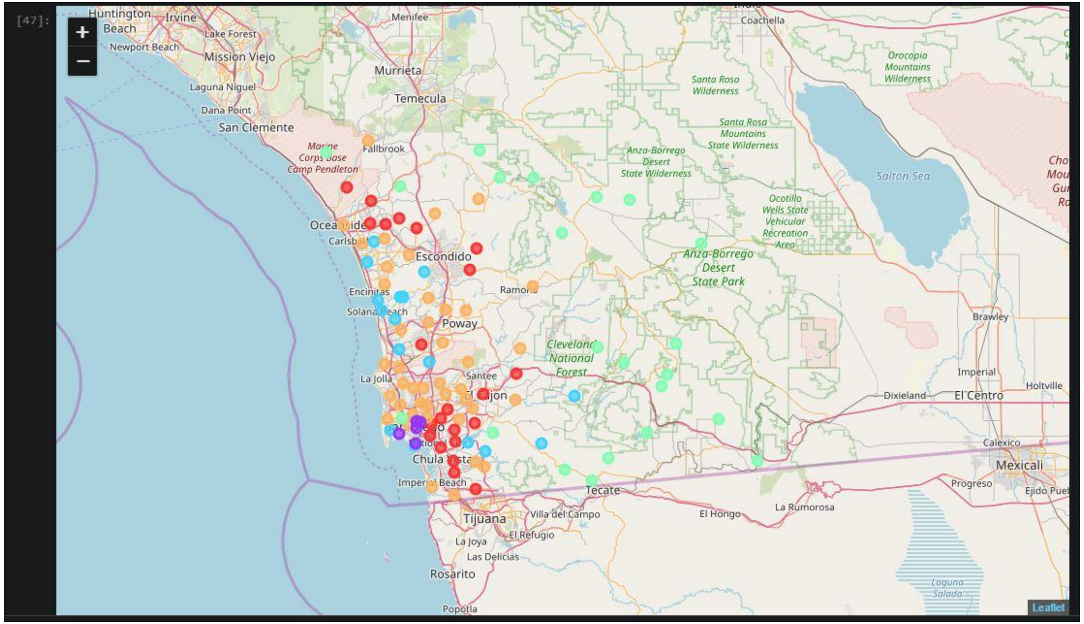
Cluster 0 (**Red**): **LOWER INCOME.**

Cluster 2 (**Light Blue**): **AFFLUENT.**

Cluster 4 (**Orange**): **TYPICAL.**

Not relevant to our scope:

Cluster 1 (**Purple**)
Cluster 3 (**Light Green**)



Here is the link of the jupyter notebook that was used for the K-Means and visualization:
https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_1_Data_Viz-Copy1.ipynb

Refine Data Requirements - Geographic data, Zip-Codes.com

A comparison of the cluster characteristics will also be relevant later on, in order to understand the potential customer profile and give context to where they are located.

- TYPICAL represents approx. 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.

	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

Refine Data Requirements - Restaurant venue data (Foursquare)

It's important to consider very useful data sources with geographic context such as **Foursquare**. It can provide relevant insight for different types of venues. Understanding this, this data source will be used for:

- ❏ Obtaining a dataset that enables us to **rank** restaurant venues for the **target zip codes** of the San Diego Metro area.
- ❏ Review the data to determine if there are relevant classification (clusters) of popular (hence successful) restaurant venues.

The data provided is a good starting point, but we need to filter only the venues we are interested in ("restaurants"). Once filtered, we can list and create a simple top 10 list that gives us an objective idea of the type of restaurants that are popular in San Diego.

Here is the link for those interested in these steps:
https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_2_FourSqr_Data.ipynb

	Type	Occurrence
1	Mexican	88
2	Fast Food	49
3	American	31
4	Chinese	28
5	Sushi	27
6	Seafood	23
7	Italian	20
8	Restaurant	16
9	Thai	14
10	Vietnamese	14

Refine Data Requirements - Restaurant venue data (Foursquare)

The ranking process requires some additional steps to “translate” the Foursquare data, into classifications that we can better understand.

Here are some examples of this “translation”, so we can give context of the popular venue types by zip code.

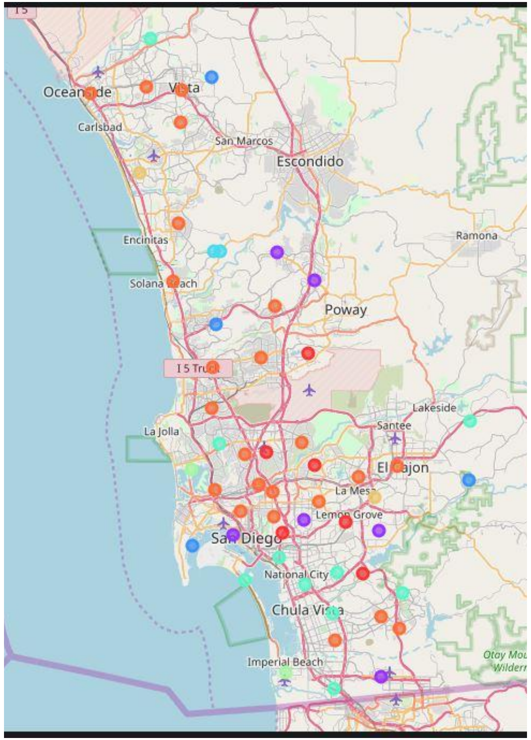
Our analysis will concentrate on the more popular restaurant venues.

Index	2	5
ZipCode	91,902	91,910
Latitude	32.67855	32.635694
Longitude	-117.013671	-117.052566
City	BONITA	CHULA VISTA
CountyName	SAN DIEGO	SAN DIEGO
1st Most Common Venue	Mexican	Mexican
2nd Most Common Venue	Vietnamese	Fast Food
3rd Most Common Venue	Japanese	Chinese
4th Most Common Venue	Indian	Greek
5th Most Common Venue	Hawaiian	Japanese
6th Most Common Venue	Halal	Italian
7th Most Common Venue	Greek	Comfort Food
8th Most Common Venue	French	Cuban
9th Most Common Venue	Filipino	Eastern European
10th Most Common Venue	Fast Food	Caribbean

Refine Data Requirements - Restaurant venue data (Foursquare)

Once applied learned techniques to rank the information, we can apply clustering algorithms (K-Means) to review if there are any significant classifications ("Clusters") for popular restaurant venues:

- ❑ Cluster 0 (Red): Mexican , Vietnamese , Japanese and Indian .
- ❑ Cluster 1 (Purple): American and Seafood.
- ❑ Cluster 2 (Blue): American, Fast food and Indian.
- ❑ Cluster 3 (Light Blue): General Restaurant, Vietnamese , Fast food.
- ❑ Cluster 4 (Green): Fast food , Mexican and Chinese.
- ❑ Cluster 5 (Light Green): Italian , Fast food and Indian.
- ❑ Cluster 6 (Yellow): Mexican, Chinese, Vietnamese.
- ❑ Cluster 7 (Orange): Mexican, Sushi, Fast food venues, **focusing on a larger variety of options**.



The same jupyter notebook includes the steps to visualize these clusters of Foursquare data:
https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_2_FourSqr_Data.ipynb

Method - Apply Insight

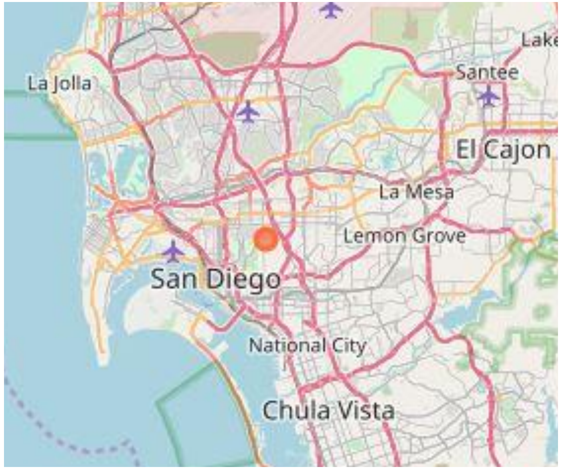
Apply Insight - Joe’s feedback

Upon review of the initial findings, it was very clear to Joe that:

- ★ His **french cuisine training** is a big factor, and should be taken into consideration
- ★ Even though the “**French**” venue type has little representation in the data.

A simple search shed some additional light considering this feedback:

- ★ Zip code 92104, has “French Restaurant” as its most popular venue.
- ★ Zip code 92104 shares characteristics with our classifications (clusters) that are relevant:
  - ○ Popular restaurant venues - Cluster 7
  - ○ Demographic profiles - Cluster 4

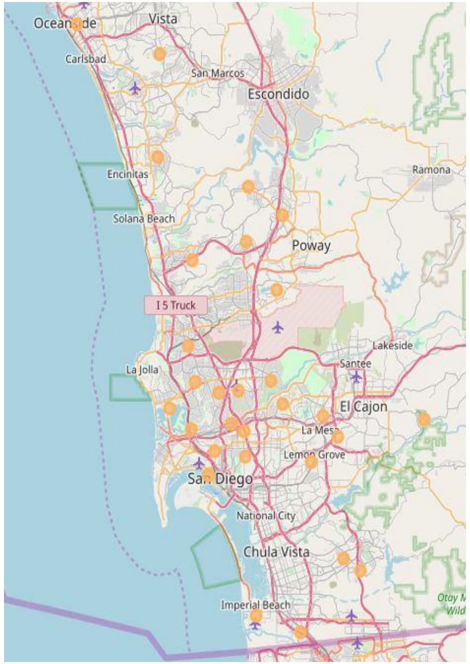


You can find further detail in the link:
https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_3_Cluster_Info.ipynb

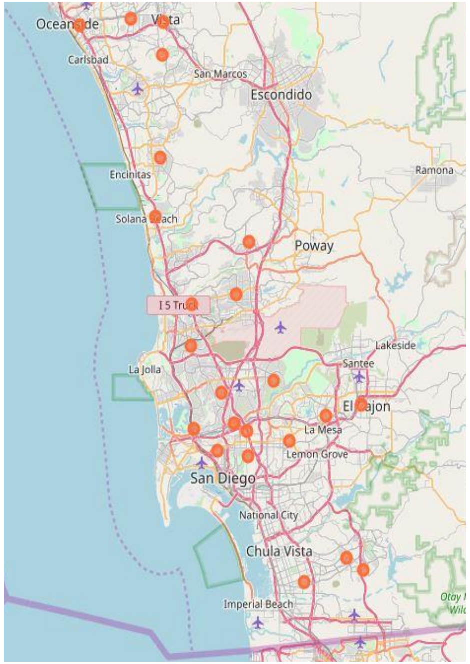
Apply Insight - Shared characteristics

We can also visualize the shared characteristics, and remembering what they represent:

Popular restaurant venues - Cluster 7 - Mexican, Sushi, Fast food, **focusing on a larger variety of options.**



Demographic profiles - Cluster 4 - **TYPICAL** households, with good average income.



Apply Insight - Shared characteristics

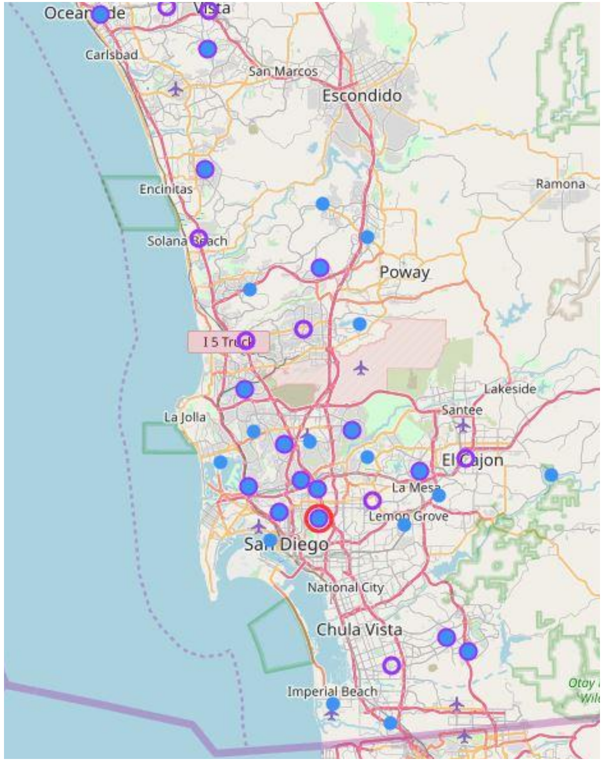
And combine the information so we can see the overlaps:

Red Circle - Location of the most popular French Restaurant venue.

Blue circles - Demographic cluster (4) that share the same characteristics, of the TYPICAL Household.

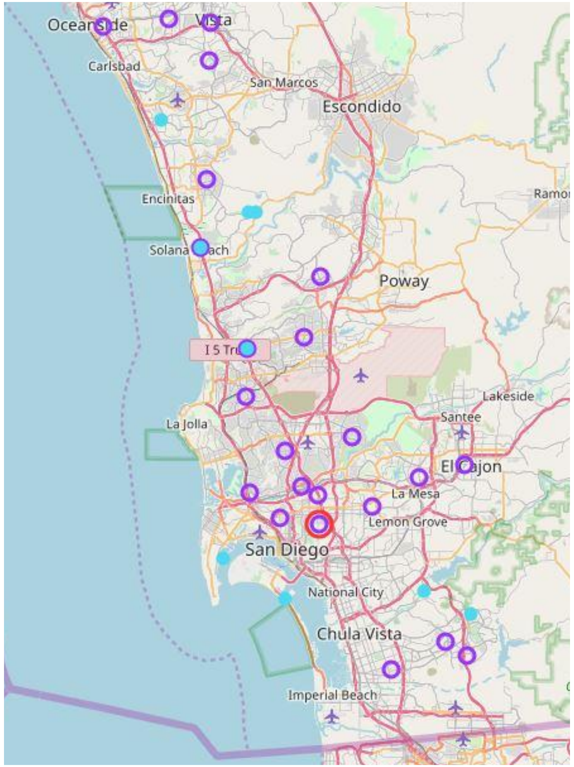
Purple circles - Shared characteristics of the classification cluster for popular restaurants (Cluster 7), where variety is significant.

You can find further detail in the link:
https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_3_Cluster_Info.ipynb



Apply Insight - Shared characteristics

If we remember the demographic profiles, so we can see some that some of the overlaps might fit within other profiles.



Red Circle - Location of the most popular French Restaurant venue.

Light Blue circles - Demographic cluster (2), it does not share share the same characteristics, the AFFLUENT Household is relevant for the site selection, since it enables an “upscale” French restaurant alternative.

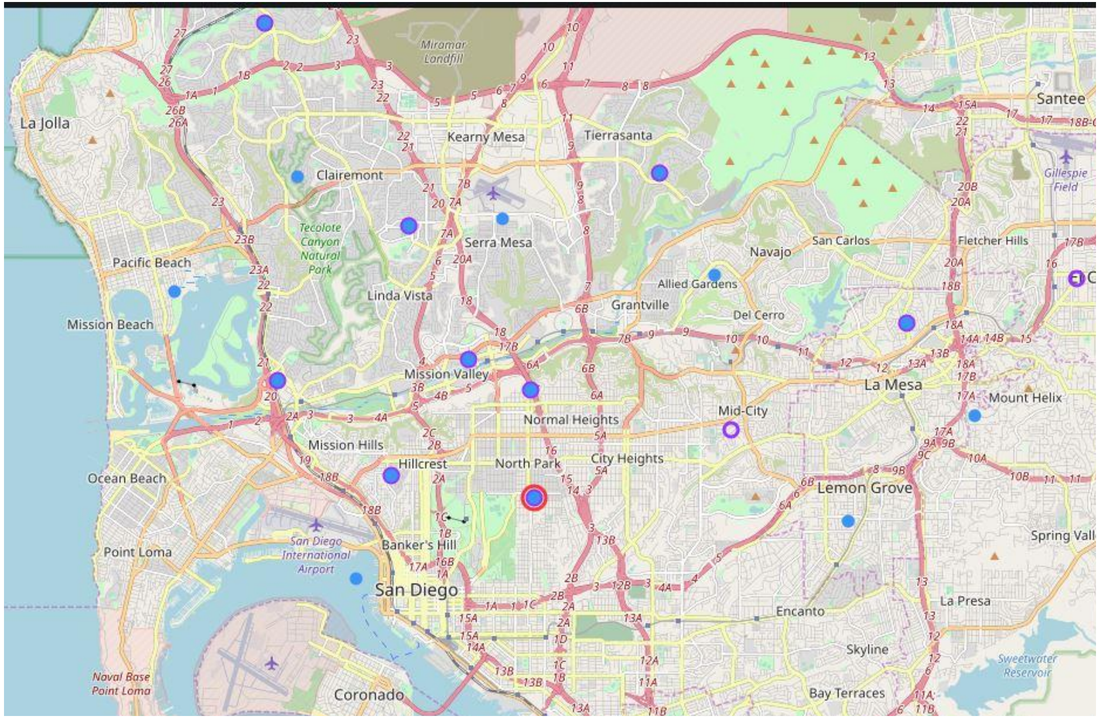
Purple circles - Shared characteristics of the classification cluster for popular restaurants (Cluster 7), where variety is significant.

You can find further detail in the link: https://github.com/Emet-DS/Coursera_Capstone_01/blob/master/W05_Final_Assig_Part02_2_3_Cluster_Info.ipynb

Method - Final observations

Final observations

Understanding the popular "French Venue" reference is important:



Popular venue types from Foursquare (Cluster 7) - The **variety** of restaurant venue options make this classification important.

Demographic profile from Zip-Codes.com (Cluster 4) - The **TYPICAL** San Diego household, with relevant income level, enables people to buy and enjoy different types of restaurants).

Variety and reasonable income level enable a strong customer base location.

Final observations

Locations where shared characteristics overlap:

Zip codes	City	Approx. Population	Note
92129	SAN DIEGO	51,536	
92024	ENCINITAS	49,121	
92111	SAN DIEGO	45,096	
92104	**SAN DIEGO**	**44,414**	**Most Popular French Venue**
92122	SAN DIEGO	43,728	
91913	CHULA VISTA	40,971	
92054	OCEANSIDE	40,375	
91942	LA MESA	38,069	
92116	SAN DIEGO	31,680	**French competitors**
92103	SAN DIEGO	31,066	**French competitors**
92124	SAN DIEGO	30,443	
92081	VISTA	27,404	
92110	SAN DIEGO	25,341	
91915	CHULA VISTA	24,659	**French competitors**
92108	SAN DIEGO	18,858	

Final observations

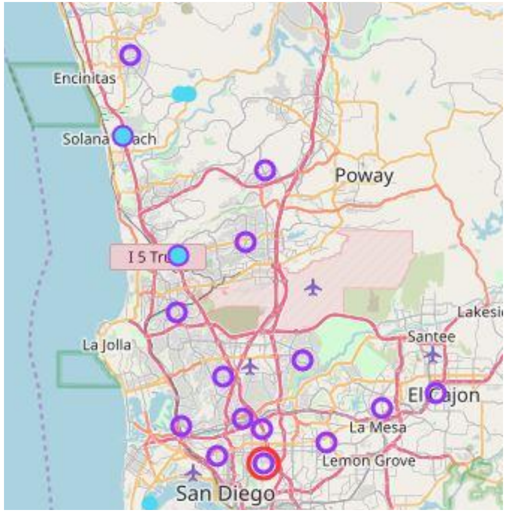
Locations where shared characteristics overlap and there is **no competition**:

Zip codes	City	Approx. Population	Note
92129	SAN DIEGO	51,536	
92024	ENCINITAS	49,121	
92111	SAN DIEGO	45,096	
			Most Popular French competitors
92122	SAN DIEGO	43,728	
91913	CHULA VISTA	40,971	
92054	OCEANSIDE	40,375	
91942	LA MESA	38,069	
			French competitors
			French competitors
92124	SAN DIEGO	30,443	
92081	VISTA	27,404	
92110	SAN DIEGO	25,341	
			French competitors
92108	SAN DIEGO	18,858	

Final observations

Locations for an “upscale” version of French restaurant (no French type competitors):

Zip codes	City	Approx. Population	Note
92075 92121	SOLANA BEACH SAN DIEGO	12,056 4,179	Site to consider if “upscale” is viable



Method - Recommendations

Final recommendations

Without forgetting the important observations (1/2):

★ There was supporting data for the San Diego through its Zip Codes, for the DS analysis. **Foursquare API data** to determine popular restaurants venues, **Zip-Codes.Com API data** to gather demographic data to understand potential customers near popular restaurant venues.

★ Using a clustering algorithm (K-Means), it was viable to classify the San Diego area 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 understand which types and combinations of restaurants venues 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).

Final recommendations

Without forgetting the important observations (2/2):

★ The challenge was to consider Joe’s cuisine expertise: **French Cuisine**.

★ With the Foursquare ranking exercise, it was viable to identify popular French Restaurant Venues:
  - The particular site, 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.

Final recommendations - Conclusions

How does DS anser Joe's questions? (1/2)

- 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.

Final recommendations - Conclusions

How does DS anser Joe's questions? (2/2)

- 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.

The application of the defined method provides insight, we can objectively 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.

What’s next?

After the initial analysis - What's next?

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).