# Capstone Project - The Battle of Neighborhoods

## “Finding the best neighborhoods in Downtown San Diego to open a - French Restaurant”

**Presentation by - Ganesh K**

# Business problem statement

Our client "**XYZ French food chain**" is interested in opening a **new French Restaurant** in **Downtown San Diego** neighborhood. For that they have asked us to do research of Downtown San Diego area and suggest probable locations for opening a new French restaurant.

# Introduction

Downtown San Diego is the city center of San Diego, California, the eighth largest city in the United States. Downtown San Diego serves as the cultural and financial center and central business district of San Diego, with more than 4,000 businesses and following neighborhoods

- **Columbia** - the west district of downtown. Located between the Marina and Little Italy, west of Columbia Street
- **Civic Core** - District, the central business district of downtown
- **Cortez Hill** - the northeast district of downtown
- **East Village** - the east district of downtown, which is home to Petco Park and the surrounding Ballpark Village
- **Gaslamp** - Quarter, a two- by ten-block nightlife district in central Downtown
- **Little Italy** - the northwest district of downtown
- **Marina** - the southwest district of downtown, which is home to Seaport Village and Pantoja Park

Downtown San Diego encompasses seven thriving neighborhoods, each with its own unique identity.

# Target Audience

Our client "**XYZ French food chain**" , who is interested in opening a new French Restaurant in Downtown San Diego neighborhood, is the target audience for this project.

# Data set

The data to be used in this project is

- Foursquare data - It is a local search and discovery service which provides information on different types of entertainment, drinking and dining venues. Foursquare has an API that is used to query their database and find information related to the venues, such as location, overall category, reviews and tips. In this project will use Foursquare API to extract the venues details from Downtown San Diego neighborhoods.

- Downtown San Diego has following seven neighborhoods i.e. Civic Core, Columbia, Cortez Hill, Gaslamp, Little Italy, Marina and East Village. After extracting both the neighborhood names and its location coordinates, they are saved in below table(in csv format)

| **Borough** | **Neighborhood** | **Latitude** | **Longitude** | **City** |
|---|---|---|---|---|
| Downtown SanDiego | CIVIC_CORE | 32.7159 | -117.1595 | San Diego |
| Downtown SanDiego | COLUMBIA | 32.7178 | -117.1673 | San Diego |
| Downtown SanDiego | CORTEZ HILL | 32.7214 | -117.1598 | San Diego |
| Downtown SanDiego | GASLAMP | 32.7101 | -117.1601 | San Diego |
| Downtown SanDiego | LITTLE ITALY | 32.7234 | -117.1682 | San Diego |
| Downtown SanDiego | MARINA | 32.7108 | -117.1701 | San Diego |
| Downtown SanDiego | EAST VILLAGE | 32.7137 | -117.1536 | San Diego |

# Methodology

## Data collection and preparation

Using the wiki links, we extract the seven Neighborhoods names of Downtown San Diego. Then extract the corresponding location coordinates i.e. Latitude and Longitude for each of the neighborhood location. This information was saved in a csv file and in our project we read this file and the location info into Pandas dataframe.

[2]:
```
#Read the preprocessed csv file which has Downtown San Diego neighborhood names and its location info
df = pd.read_csv('SD1_LatLong_Dataset.csv')
df
```

[2]:

| | Borough | Neighborhood | Latitude | Longitude | City |
|---|---|---|---|---|---|
| 0 | Downtown SanDiego | CIVIC_CORE | 32.7159 | -117.1595 | San Diego |
| 1 | Downtown SanDiego | COLUMBIA | 32.7178 | -117.1673 | San Diego |
| 2 | Downtown SanDiego | CORTEZ HILL | 32.7214 | -117.1598 | San Diego |
| 3 | Downtown SanDiego | GASLAMP | 32.7101 | -117.1601 | San Diego |
| 4 | Downtown SanDiego | LITTLE ITALY | 32.7234 | -117.1682 | San Diego |
| 5 | Downtown SanDiego | MARINA | 32.7108 | -117.1701 | San Diego |
| 6 | Downtown SanDiego | EAST VILLAGE | 32.7137 | -117.1536 | San Diego |

# Exploratory Data Analysis(EDA)

After extracting the venues data using Foursquare API, lets check the generated dataframe, names of columns and the data types stored in each column

[29]:
```
#check the paramters(columns) in generated dataframe
sd_venues.head()
```

[29]:

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | CIVIC_CORE | 32.7159 | -117.1595 | House of Blues San Diego | 32.716370 | -117.159528 | Music Venue |
| 1 | CIVIC_CORE | 32.7159 | -117.1595 | THE US GRANT, A Luxury Collection Hotel, San D... | 32.716149 | -117.161527 | Hotel |
| 2 | CIVIC_CORE | 32.7159 | -117.1595 | Vin de Syrah Wine Parlor | 32.714723 | -117.160040 | Wine Bar |
| 3 | CIVIC_CORE | 32.7159 | -117.1595 | The Taco Stand Downtown | 32.717749 | -117.158497 | Taco Place |
| 4 | CIVIC_CORE | 32.7159 | -117.1595 | Donut Bar | 32.717619 | -117.158757 | Donut Shop |

[75]:
```
#check the names of columns in dataframe and its type
sd_venues.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 700 entries, 0 to 699
Data columns (total 7 columns):
Neighborhood              700 non-null object
Neighborhood Latitude     700 non-null float64
Neighborhood Longitude    700 non-null float64
Venue                     700 non-null object
Venue Latitude            700 non-null float64
Venue Longitude           700 non-null float64
Venue Category            700 non-null object
dtypes: float64(4), object(3)
memory usage: 38.4+ KB
```

Also lets check the number of venues per neighborhood

[77]:
```
#check the number of venues per neighbourhood
sd_venues['Neighborhood'].value_counts()
```

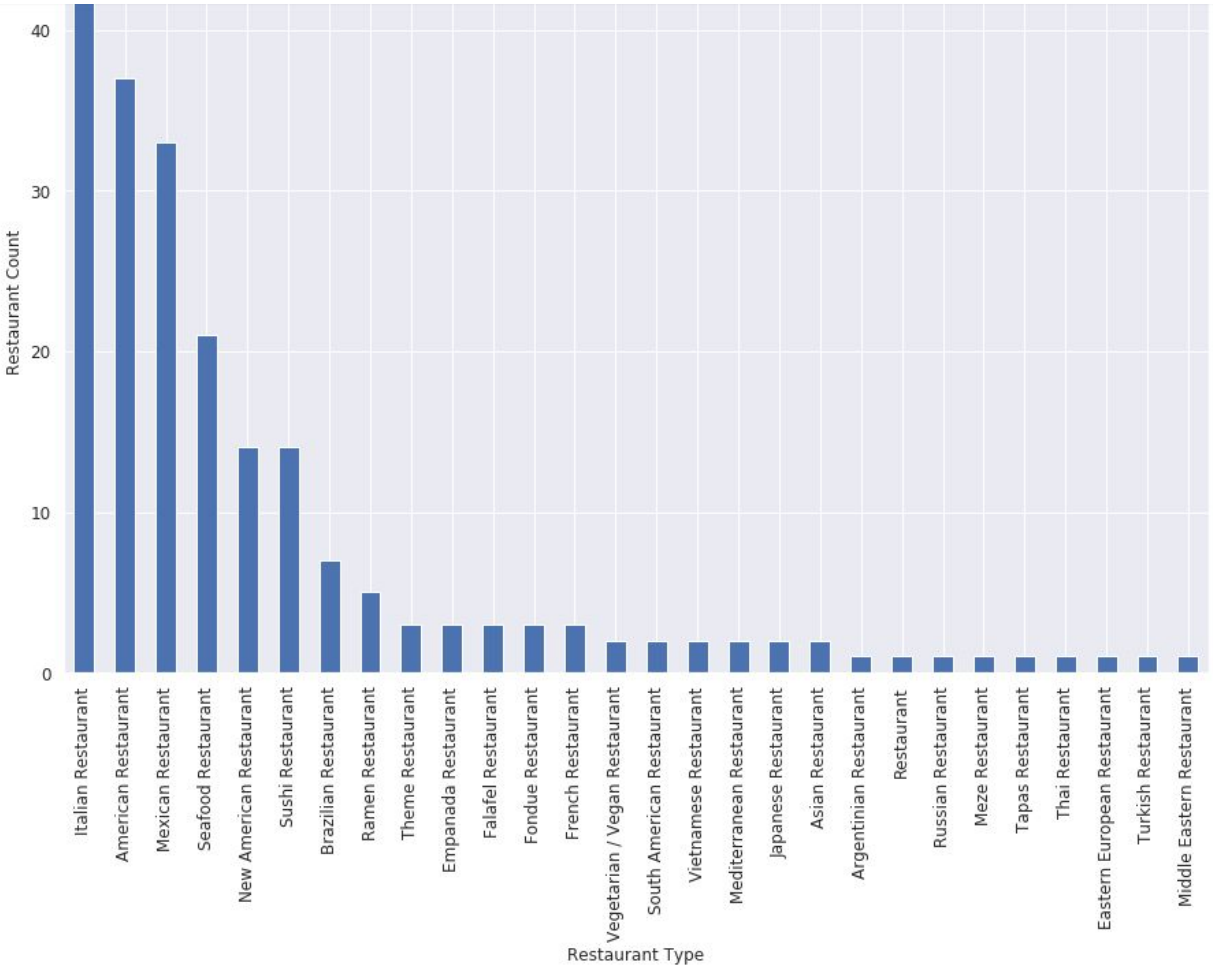
[77]:
```
LITTLE ITALY    100
MARINA          100
EAST VILLAGE    100
CIVIC_CORE      100
GASLAMP         100
COLUMBIA        100
CORTEZ HILL     100
Name: Neighborhood, dtype: int64
```

[78]:
```
#Number of venues per neighborhood
sd_venues.groupby('Neighborhood').count()
```

[78]:

| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| CIVIC_CORE | 100 | 100 | 100 | 100 | 100 | 100 |
| COLUMBIA | 100 | 100 | 100 | 100 | 100 | 100 |
| CORTEZ HILL | 100 | 100 | 100 | 100 | 100 | 100 |
| EAST VILLAGE | 100 | 100 | 100 | 100 | 100 | 100 |
| GASLAMP | 100 | 100 | 100 | 100 | 100 | 100 |
| LITTLE ITALY | 100 | 100 | 100 | 100 | 100 | 100 |
| MARINA | 100 | 100 | 100 | 100 | 100 | 100 |

Lets plot the bar chart of types of Restaurants and its count




Now lets see the French restaurants in neighborhoods

**Now lets analyze Each Neighborhood specifically for French Restaurants**

[83]:
```
# check if the results contain "French Restaurants"
"French Restaurant" in sd_venues['Venue Category'].unique()
```

[83]: True

[84]:
```
#lets get the details of French Restaurant in neighborhoods
sd_venues[sd_venues['Venue Category'] == 'French Restaurant']
```

[84]:

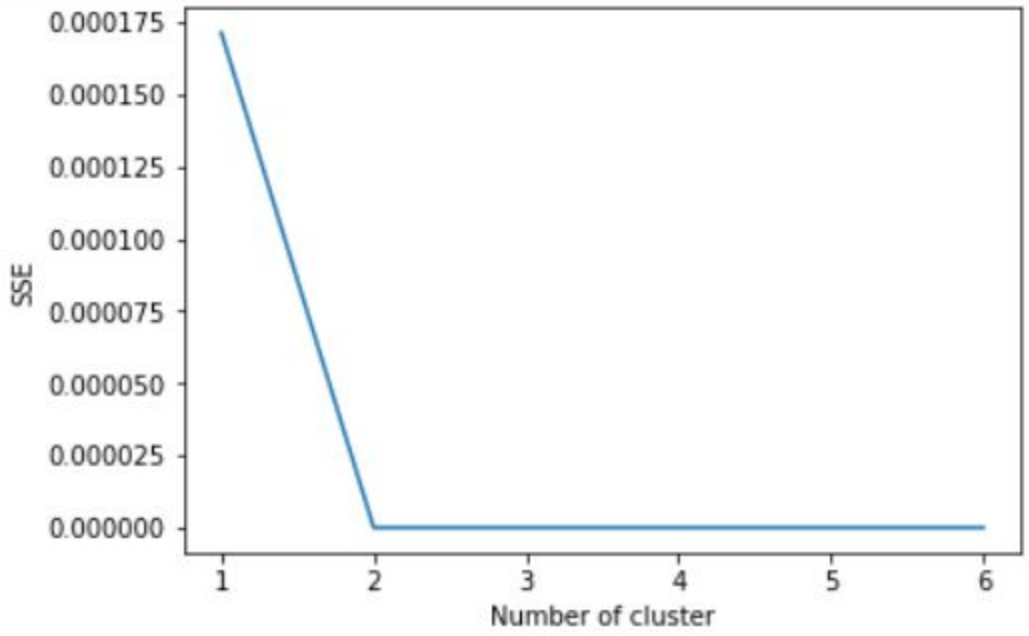
| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| **43** | CIVIC_CORE | 32.7159 | -117.1595 | Le Fontainbleau - The Westgate Hotel | 32.716629 | -117.162721 | French Restaurant |
| **132** | COLUMBIA | 32.7178 | -117.1673 | Le Fontainbleau - The Westgate Hotel | 32.716629 | -117.162721 | French Restaurant |
| **232** | CORTEZ HILL | 32.7214 | -117.1598 | Le Fontainbleau - The Westgate Hotel | 32.716629 | -117.162721 | French Restaurant |

So there are three French Restaurant in neighborhoods CIVIC_CORE, COLUMBIA and CORTEZ HILL

## Machine Learning algorithm for clustering

Now lets use the machine learning algorithm to generate the clusters of neighborhoods which may have the French restaurants. For this will use the K Means clustering algorithm of machine learning. First we need to get the right value of K

```
"""
```




```
As seen in above graph the k value to be used is 2
```

After applying the K Means clustering algorithm with K =2 will get the following result

[101]:
```
# create a new dataframe that includes the cluster as well as the top 10
to_merged = to_french.copy()

# add clustering labels
to_merged["Cluster Labels"] = kmeans.labels_
```

[102]:
```
to_merged.rename(columns={"Neighborhoods": "Neighborhood"}, inplace=True)
to_merged
```

[102]:

| | Neighborhood | French Restaurant | Cluster | Cluster Labels |
|---|---|---|---|---|
| 0 | CIVIC_CORE | 0.01 | 1 | 1 |
| 1 | COLUMBIA | 0.01 | 1 | 1 |
| 2 | CORTEZ HILL | 0.01 | 1 | 1 |
| 3 | EAST VILLAGE | 0.00 | 0 | 0 |
| 4 | GASLAMP | 0.00 | 0 | 0 |
| 5 | LITTLE ITALY | 0.00 | 0 | 0 |
| 6 | MARINA | 0.00 | 0 | 0 |

From above table its clear that Cluster Label 1 contains the Downtown San Diego neighborhoods of CIVIC_CORE, COLUMBIA and CORTEZ HILL which have at least one French Restaurant in it. Whereas Cluster Label 0 contains the Downtown San Diego neighborhoods of MARINA, GASLAMP, LITTLE ITALY and EAST VILLAGE which has no French Restaurant.

Generated cluster info

[108]:
```
#get the neighborhood names per cluster
to_merged.groupby('Cluster Labels')['Neighborhood'].unique()
```

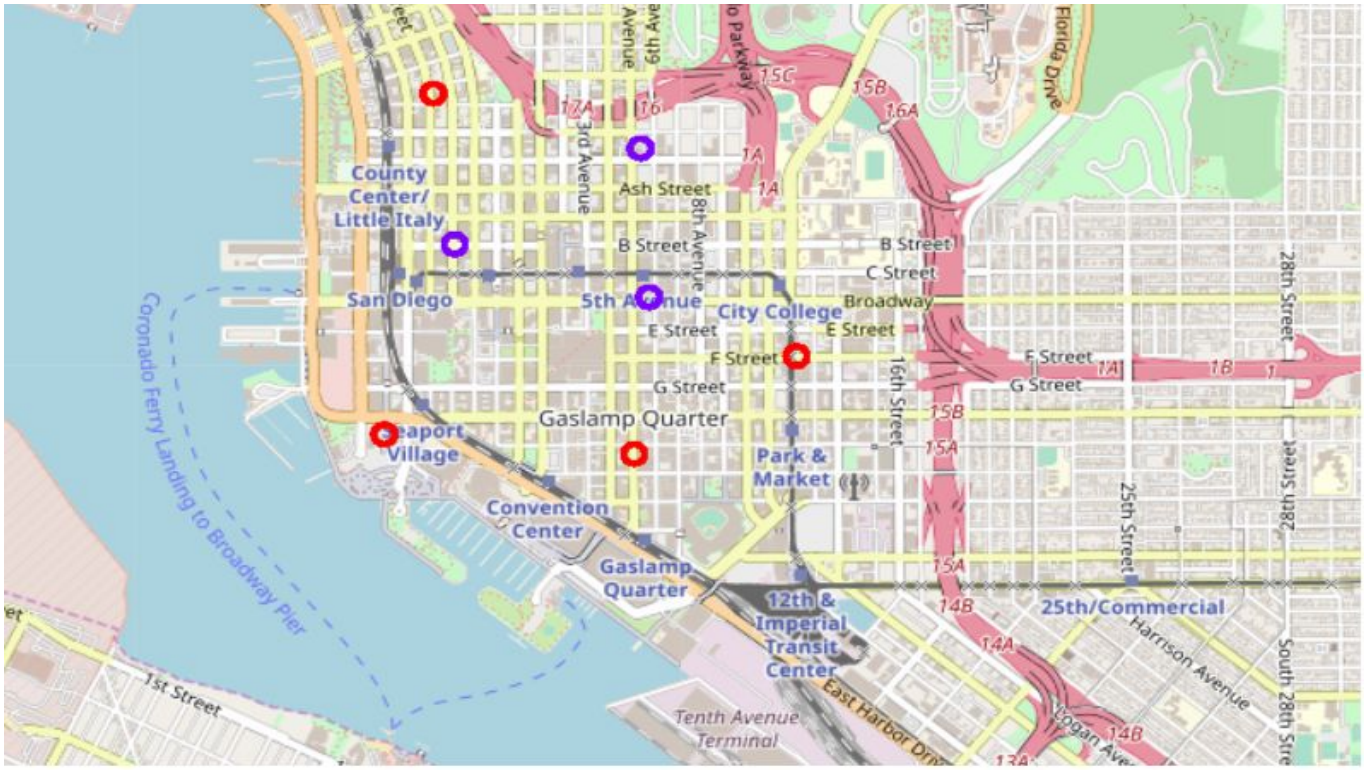
[108]:
```
Cluster Labels
0    [EAST VILLAGE, GASLAMP, LITTLE ITALY, MARINA]
1             [CIVIC_CORE, COLUMBIA, CORTEZ HILL]
Name: Neighborhood, dtype: object
```

[111]:
```
#Lets check the French Restaurant details with cluster info
to_merged[to_merged['Venue Category'] == 'French Restaurant']
```

[111]:

| | Neighborhood | French Restaurant | Cluster | Cluster Labels | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | COLUMBIA | 0.01 | 1 | 1 | 32.7178 | -117.1673 | Le Fontainbleau - The Westgate Hotel | 32.716629 | -117.162721 | French Restaurant |
| 0 | CIVIC_CORE | 0.01 | 1 | 1 | 32.7159 | -117.1595 | Le Fontainbleau - The Westgate Hotel | 32.716629 | -117.162721 | French Restaurant |
| 2 | CORTEZ HILL | 0.01 | 1 | 1 | 32.7214 | -117.1598 | Le Fontainbleau - The Westgate Hotel | 32.716629 | -117.162721 | French Restaurant |

Now lets visualize the generated cluster neighborhoods using folium map

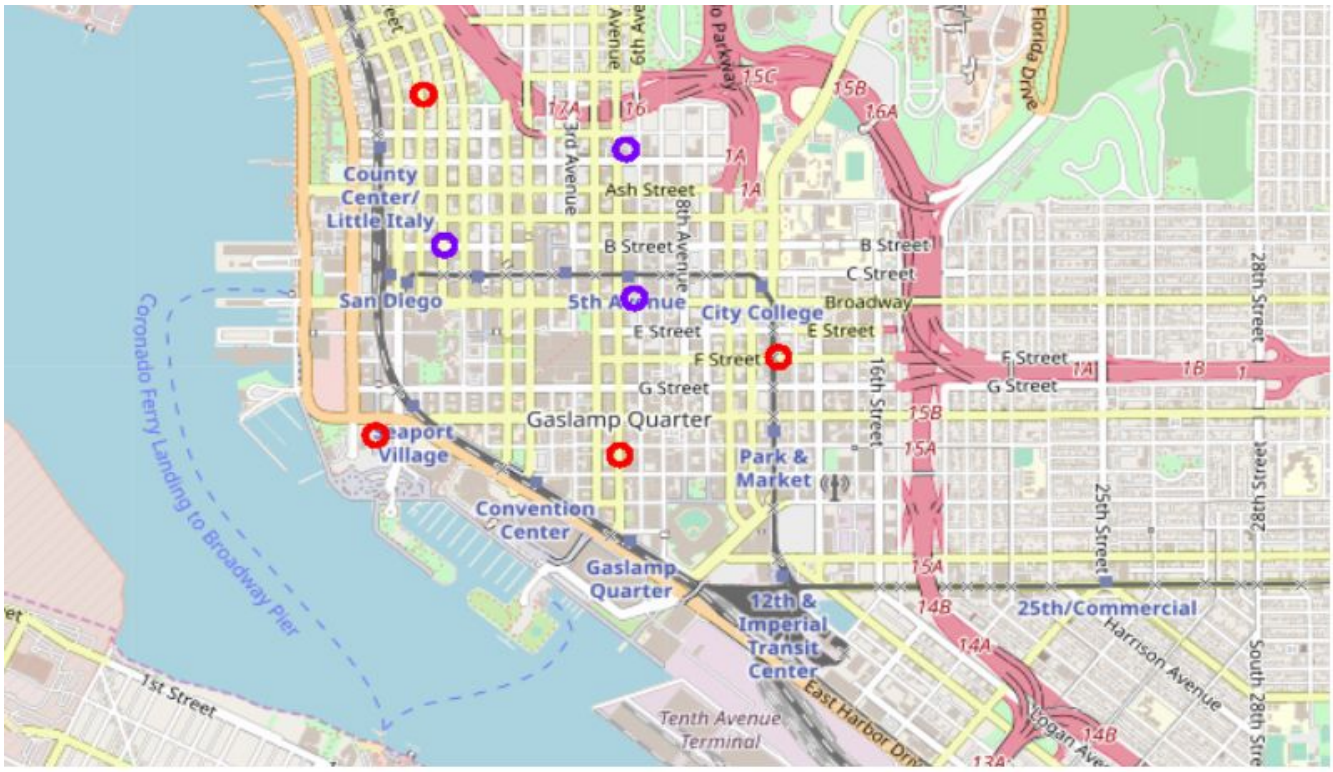


# Results

As per our preprocessed data set there are seven Neighborhoods in Downtown San Diego. Using Foursquare API, we extracted venues in all the neighborhoods. We got total of 700 venues in seven neighborhoods and there are 129 unique venue categories in generated data. There are 28 different types of Restaurants categories and total number of Restaurant venues are 211.

When we further explored the Restaurants data specifically for French Restaurants, we found that there are three French Restaurants in three different neighborhoods i.e. CIVIC_CORE, COLUMBIA and CORTEZ HILL. K Means clustering algorithm further confirms this fact with below clustering details

- **Cluster 1** - contains the Downtown San Diego neighborhoods of **CIVIC_CORE, COLUMBIA and CORTEZ HILL** which have at least one French Restaurant
- **Cluster 0** - contains the Downtown San Diego neighborhoods of **MARINA, GASLAMP, LITTLE ITALY and EAST VILLAGE** which has no French Restaurant



# Discussion

While working on this project I faced following challenges and my observations

- I was looking for a common database of all the major cities in the world with its Borough/Districts and Neighborhoods/localities with zip code and geocodes(Latitude and Longitude). I observed that this info is not freely available and getting the most accurate geocodes at neighborhood level for all venues is a challenge
- Foursquare API is good and contains rich venue data for US and other few countries but when I tried looking for India data, I found that its not complete
- In current project I limited the scope with Foursquare APIs data for Downtown San Diego Neighborhood selection. But along with venues data, even real estate pricing, cost of living, population, earning capabilities/opportunities, public transport, commute time to work, pollution levels, crime rate/safety, local government policies etc can be considered to arrive at proper neighborhood choices

# Conclusion

In this project the given problem statement was, to identify the **Neighborhoods in Downtown San Diego** to open a **new French Restaurant** as enquired by our client **"XYZ French food chain**". We started with data collection of Downtown San Diego Neighborhood names and its geocodes. Then using Foursquare API we extracted the list of venues in each of the neighborhoods. Then we performed the exploratory data analysis to check the type of Restaurant venues in different neighborhoods. After that we performed the EDA, especially for French Restaurants in all the neighborhoods. Then we used the machine learning technique K Means clustering algorithm to create clusters of neighborhoods with availability of French Restaurants. The results of EDA and clustering algorithm shows that there are two distinct clusters in Downtown San Diego i.e.

- **Cluster 1** - contains the Downtown San Diego neighborhoods of **CIVIC_CORE, COLUMBIA and CORTEZ HILL** which have at least one French Restaurant
- **Cluster 0** - contains the Downtown San Diego neighborhoods of **MARINA, GASLAMP, LITTLE ITALY and EAST VILLAGE** which has no French Restaurant

Hence we would like to recommend our client **"XYZ French food chain**" to check the possibility of opening a new French Restaurant in **Cluster 0** i.e. **Downtown San Diego neighborhoods of MARINA, GASLAMP, LITTLE ITALY or EAST VILLAGE** which has no French Restaurant as per Foursquare venue data. The final call of opening a new French Restaurant at above mentioned neighborhoods rests with our client senior management, based on various business considerations.