Geo-location Clustering Using K-means Algorithm In Python

Outline

- Introduction
- Business Problem and Interest
- Libraries and Modules
- Data
- Methodology
- Data Analysis
- Results
- Discussion
- Conclusion

Introduction

- California's second largest city and the United States' eighth largest, San Diego boasts a citywide population of nearly 1.3 million residents and more than 3 million residents countywide.
- San Diego is renowned for its idyllic climate, 70 miles of pristine beaches and a dazzling array of world-class family attractions including the world-famous San Diego Zoo and San Diego Zoo Safari Park, SeaWorld San Diego, and LEGOLAND California.
- The sunny weather makes San Diego a hot spot for vacationers of all ages from around the world.
- The city is not only about tourism and beach holidays, but also a good destination for plenty of qualified employees and entrepreneurs to start a business.
- In terms of economy, San Diego, having GDP of over $250 million, made it to the top 20 major cities in the United States.
- Business owners often consider San Diego city as a good destination for moving and expanding their business because of housing and low cost of doing business as compared to Los Angeles and San Francisco areas.

The objective of the project is to find a right location for opening up an Italian restaurant in San Diego.

Business Problem and Interest

Business Problem:

- As **San Diego** receives people from all around the world, and people love to try new food, we will try to find an adequate location for opening up an Italian Restaurant in San Diego.
- Finding a proper location for a restaurant is crucial for business success. Hence, to select the right location for the restaurant, we will consider following elements:
  - Know the neighborhood, specifically, who else is doing business in the neighborhood
  - Find a place which is not crowded with similar restaurants in vicinity
  - Accessibility and visibility of the location
  - Population base to know the foot traffic or car traffic in the area to support the business
  - Parking for the customers, and
  - Low crime rate in the area as high crime rates can make potential customers uncomfortable to visit the restaurant due to fears over public safety.

Our objective is to discover a few most promising neighborhoods based on above-mentioned criteria using data science skills, and present them with statistics so that the stakeholders can select the precise location for their restaurant.

Interest:

- Our **target stakeholders** are the **restaurant entrepreneurs** who would be interested in starting a restaurant in San Diego, California

Libraries and Modules

Libraries and Modules	Description	Doc Link	Installation Link
Numpy	Python Library to handle data in a vectorized manner	Doc	Installation
Pandas	Python Library for data analysis	Doc	Installation
Json	Python Library to handle json data	Doc	Installation
Requests	Python Library to handle URL requests	Doc	Installation
Scikit-learn	Python Machine learning library	Doc	Installation
Folium	Python Map rendering library	Doc	Installation
Geopy	Python Library to locate the coordinates of addresses	Doc	Installation
Plotly	Python Plotting library	Doc	Installation
Matplotlib.cm	Python Plotting library for colormap handling	Doc	Installation
Matplotlib.colors	Python Plotting library for colormap visualization	Doc	Installation
Matplotlib.pyplot	Python Plotting library	Doc	Installation
Matplotlib.image	Python library for image plotting	Doc	Installation

Data

1. **San Diego neighborhoods data** has been collected from Wikipedia uisng **BeautifulSoup** library and processed the data in order to use this in this project.
    1. https://en.wikipedia.org/wiki/Template:Neighborhoods_of_San_Diego

2. The **geographical coordinates**− **latitude** and **longitude** of San Diego and other addresses of interest have been obtained using python **geopy** library.

3. Most common **venues** of all the Neighborhoods of San Diego have been collected using **Foursquare API**.

4. The **demographical information** as well as **property facts** data such as population, median home value, median rent, median household income, diversity, cost of Living, commute, parking, walkable to restaurants, and crime and safety for each neighborhood from below websites:
    1. https://www.niche.com/places-to-live/c/san-diego-county-ca/
    2. https://www.trulia.com/CA/San_Diego/

Methodology

- Python **geopy** library has been used to obtain the geographical coordinates of San Diego.
- The **Foursquare API** has been to segment and explore the neighborhoods as well as the latitude and longitude coordinates of each neighborhood. For this, limit is set as 100 and the radius 1000 meter for each neighborhood from their given latitude and longitude information.
- Python **folium** library is used to visualize the map of San Diego with neighborhoods superimposed on top.
- The **explore** function to get the most common venue categories in each neighborhood and then used this feature to group the neighborhoods into clusters with the help of ***K*-means clustering** algorithm.
- For ***K*-means cluster** modeling, the optimal value of the ***k*** is set to **6** (k=6) using the **Elbow Method**.
- Python **folium** library is used to visualize the neighborhoods of San Diego and their emerging clusters.
- The **demographical** as well as **property facts** data about San Diego neighborhoods have been collected and processed to set the overall rating based on these data, and merge them with related clusters of neighborhoods to select the final locations.
- Finally, the **folium** library is used to visualize **final selected neighborhoods** for opening up a **Italian restaurant** based on the criteria mentioned in business problem section.

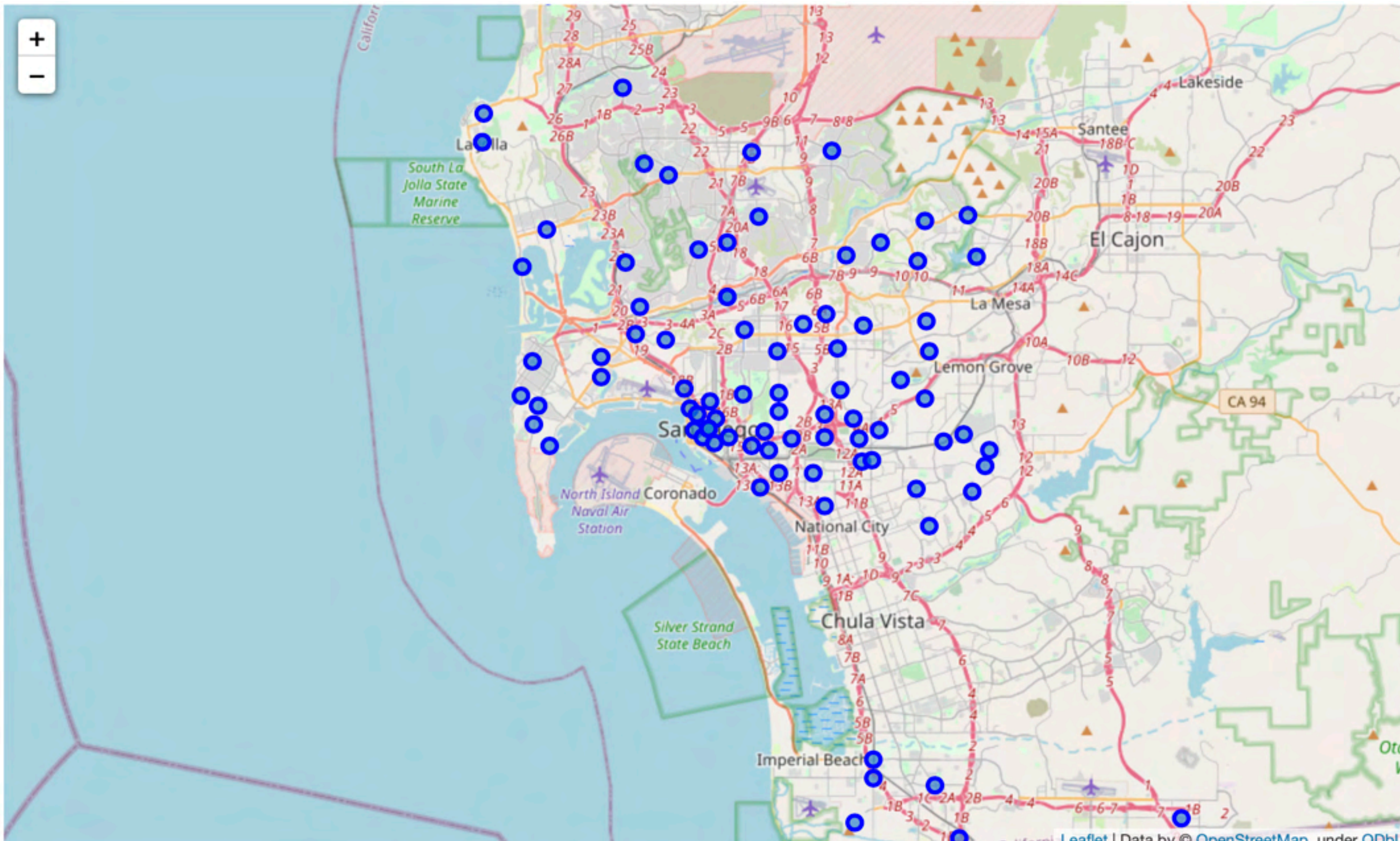
Data Analysis

1. Dataframe containing San Diego neighborhoods and geographical coordinates -- latitude and longitude of each neighborhood

	Neighborhoods	Latitude	Longitude
0	Bay Ho	32.824100	-117.193700
1	Bay Park	32.784638	-117.202605
2	Carmel Valley	32.943434	-117.213979
3	Clairemont	32.819505	-117.182340
4	Del Mar Heights	32.948811	-117.250785

Map of San Diego using Folium

2. Map of San Diego with neighborhoods superimposed on top



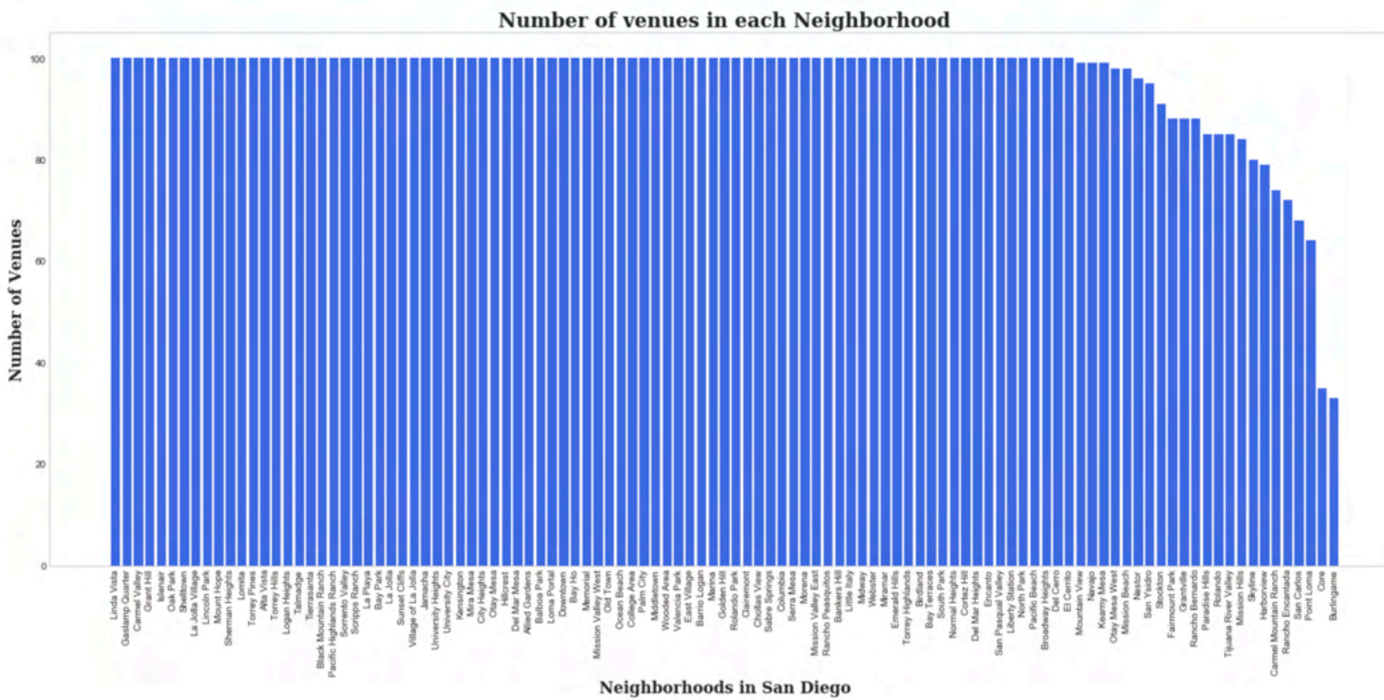
Explore San Diego Neighborhoods

3. Explore and segment San Diego neighborhoods using **Foursquare API**
   - Set the limit as 100 venue and the radius 1000 meter for each neighborhood from their given latitude and longitude value.
   - Number of venues returned by Foursquare : **10283**
   - Number of unique venue category : **510**
   - Below is the dataframe containing venues, venue categories, latitude, and longitude of each neighborhood returned by Foursquare:

	Neighborhoods	Neighborhood_Lat	Neighborhood_Lng	Venue	Venue_Lat	Venue_Lng	Venue_Category
0	Bay Ho	32.8241	-117.1937	Mt. Etna Neighborhood Park	32.822739	-117.191499	Playground
1	Bay Ho	32.8241	-117.1937	John Muir Language Academy	32.823418	-117.193544	Elementary School
2	Bay Ho	32.8241	-117.1937	Hurst Dental Care	32.826343	-117.190943	Dentist's Office
3	Bay Ho	32.8241	-117.1937	QLP Locksmith	32.822519	-117.183990	Locksmith
4	Bay Ho	32.8241	-117.1937	Circle K	32.822577	-117.183807	Convenience Store

Number of Venues in each Neighborhood

4. Bar chart showing number of venues in each neighborhoods



Linda Vista, Serra Mesa, Mira Mesa, Torrey Pines, Hillcrest, Village of La Jolla, and many other neighborhoods have reached the **100** limit of venues. On the other hand, Burlingame and Core have less than **50** venues. Neighborhoods having **100** or more venues have been considered for the rest of the analysis.

Analyze Each Neighborhood

5. Analyze each neighborhood having **100** or more venues
   - Number of neighborhoods having 100 or more venues: **84**
   - No of unique venues: **491**
   - Dataframe displaying the **top 10 venues** for each neighborhood:

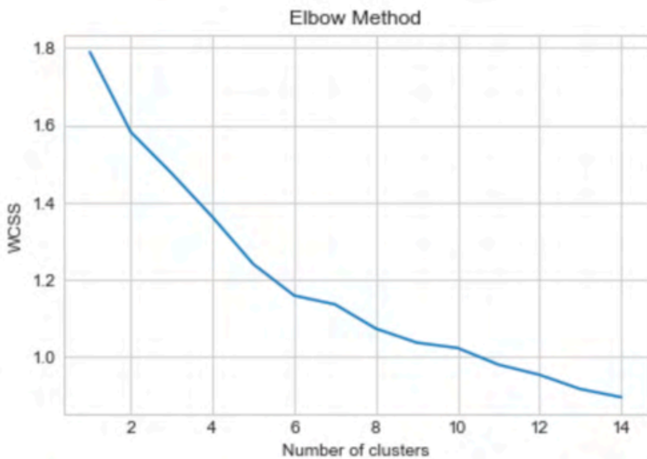
	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allied Gardens	Church	Bank	Office	Gas Station	Salon / Barbershop	Sports Bar	Mexican Restaurant	Liquor Store	Italian Restaurant	Automotive Shop
1	Alta Vista	Playground	Professional & Other Places	General Entertainment	Event Space	Lounge	Gym / Fitness Center	Community Center	Basketball Court	Bridge	Park
2	Balboa Park	Food Truck	Garden	Building	History Museum	Park	Café	Non-Profit	Flower Shop	Art Gallery	Zoo Exhibit
3	Bankers Hill	Office	Coworking Space	Doctor's Office	Laundry Service	Lawyer	Residential Building (Apartment / Condo)	Bank	Massage Studio	Building	Spa
4	Barrio Logan	Automotive Shop	Office	Building	Miscellaneous Shop	Fast Food Restaurant	Tattoo Parlor	Salon / Barbershop	Bus Line	Boat or Ferry	Gas Station

Modeling using *K*-means Clustering

6. Cluster Neighborhoods using ***K*-means**

- Selecting the optimal value for ***k***

The value of ***k*** is set to **6** (***k*=6**) using the **Elbow Method** which gives the value of ***k*** such that the total **within-cluster variation** (or error) is minimum




```
# Identify the elbow point in the wcss curve using KneeLocator()
kl = KneeLocator(range(1, 15), wcss, curve="convex", direction="decreasing")
print("Optimal value of k: {}".format(kl.elbow))
```

```
Optimal value of k: 6
```

- Modeling using ***K*-means Clustering**

```
# set number of clusters
k = 6

# drop the column 'Neighborhood'
sd_cluster = sd_grouped.drop('Neighborhoods', 1)

# Initialize and fit the model
kmeans = KMeans(n_clusters=k, random_state=4).fit(sd_cluster)

# check cluster labels generated for each row in the dataframe
labels = kmeans.labels_

print(labels)
```

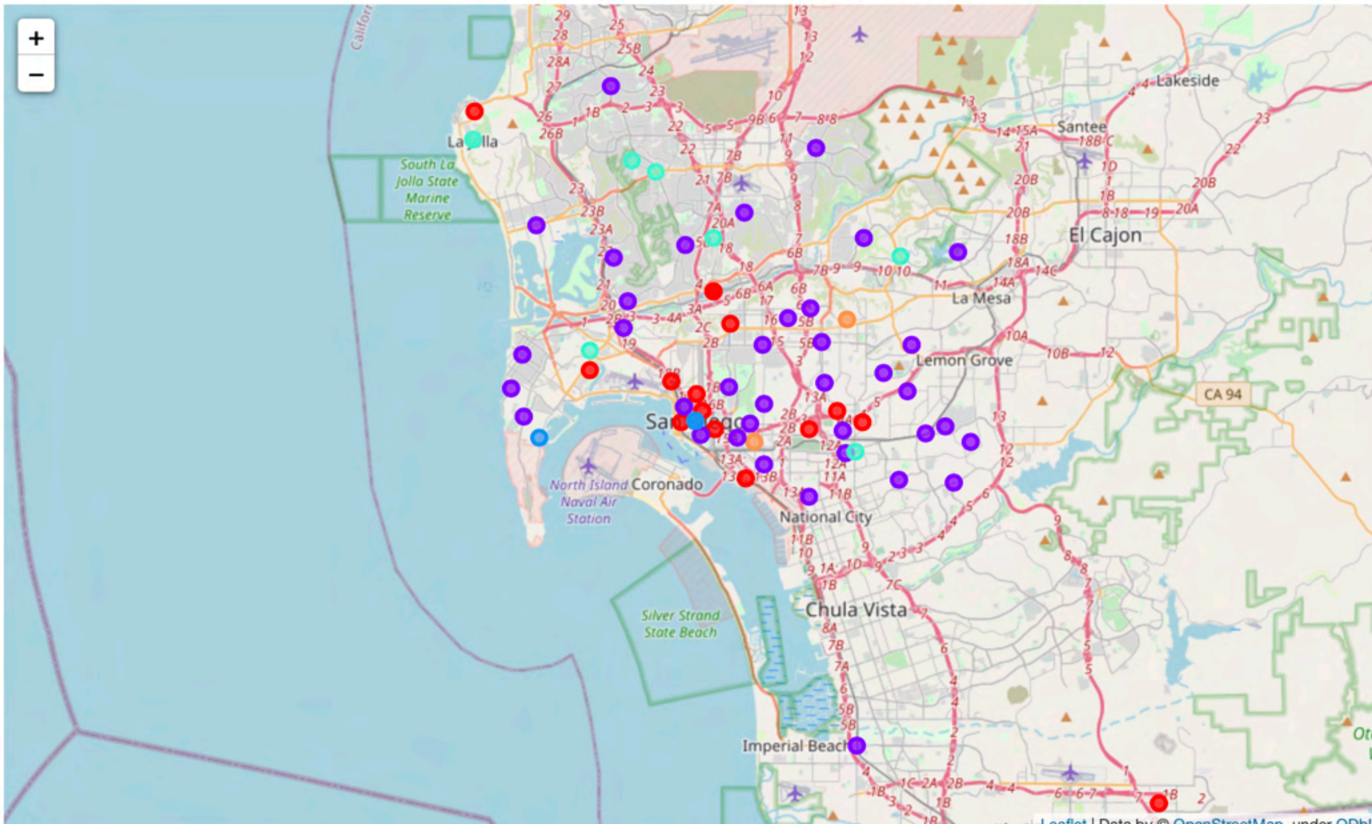
Modeling using *K*-means Clustering

- Dataframe including **cluster labels** as well as the **top 10 venues** for each neighborhood

	Neighborhoods	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10t C
0	Bay Ho	32.824100	-117.193700	3	Dentist's Office	Coffee Shop	Automotive Shop	Church	Government Building	Garden	American Restaurant	Gym / Fitness Center	Farm	Res I (Ap /
1	Bay Park	32.784638	-117.202605	1	Salon / Barbershop	Park	Church	Automotive Shop	Spa	Dentist's Office	Bus Line	Auto Dealership	Gas Station	Art
2	Carmel Valley	32.943434	-117.213979	1	Gym / Fitness Center	Pool	Residential Building (Apartment / Condo)	Trail	Gym	Park	Church	Elementary School	Stables	O
3	Clairemont	32.819505	-117.182340	3	Doctor's Office	Dentist's Office	Bank	Chiropractor	Office	Mobile Phone Shop	Medical Center	ATM	Bakery	
4	Del Mar Heights	32.948811	-117.250785	1	Office	Trail	Residential Building (Apartment / Condo)	Dentist's Office	Elementary School	Salon / Barbershop	Deli / Bodega	Business Service	Ice Cream Shop	I

Modeling using *K*-means Clustering

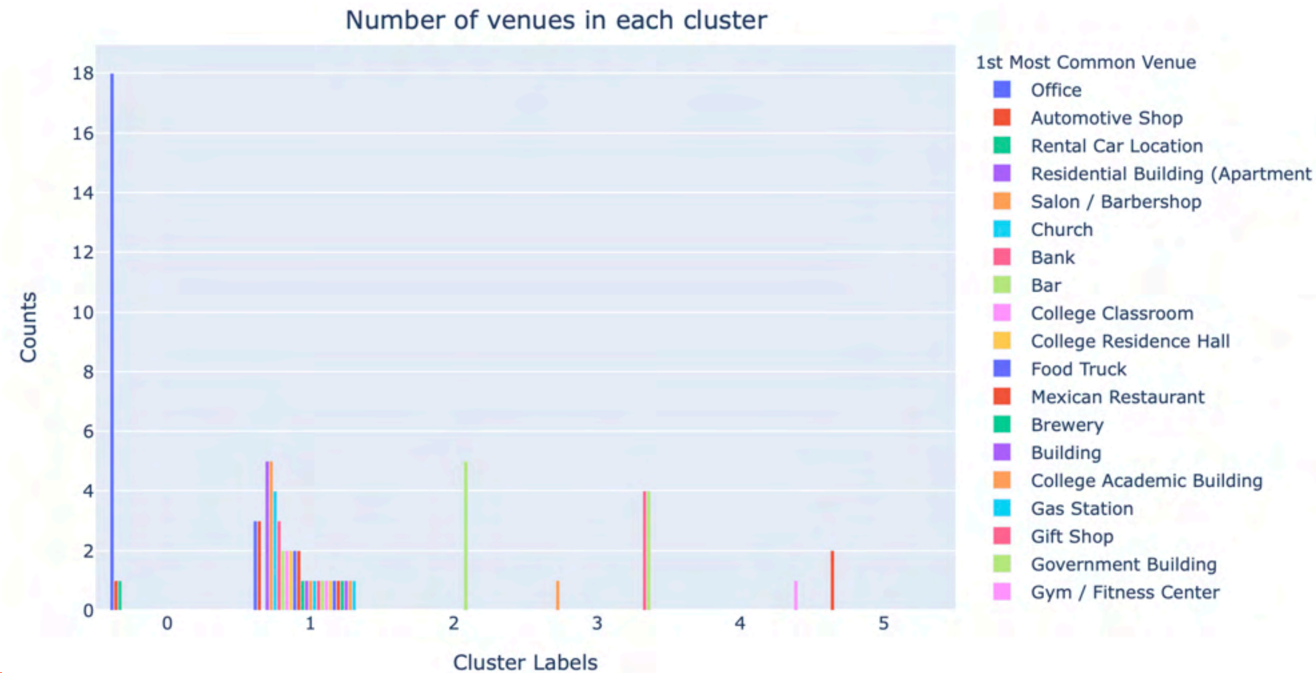
- Visualizing the resulting clusters



Explore Clusters

7. Exploring each Cluster

- Bar chart showing number of 1st common venues in each Cluster



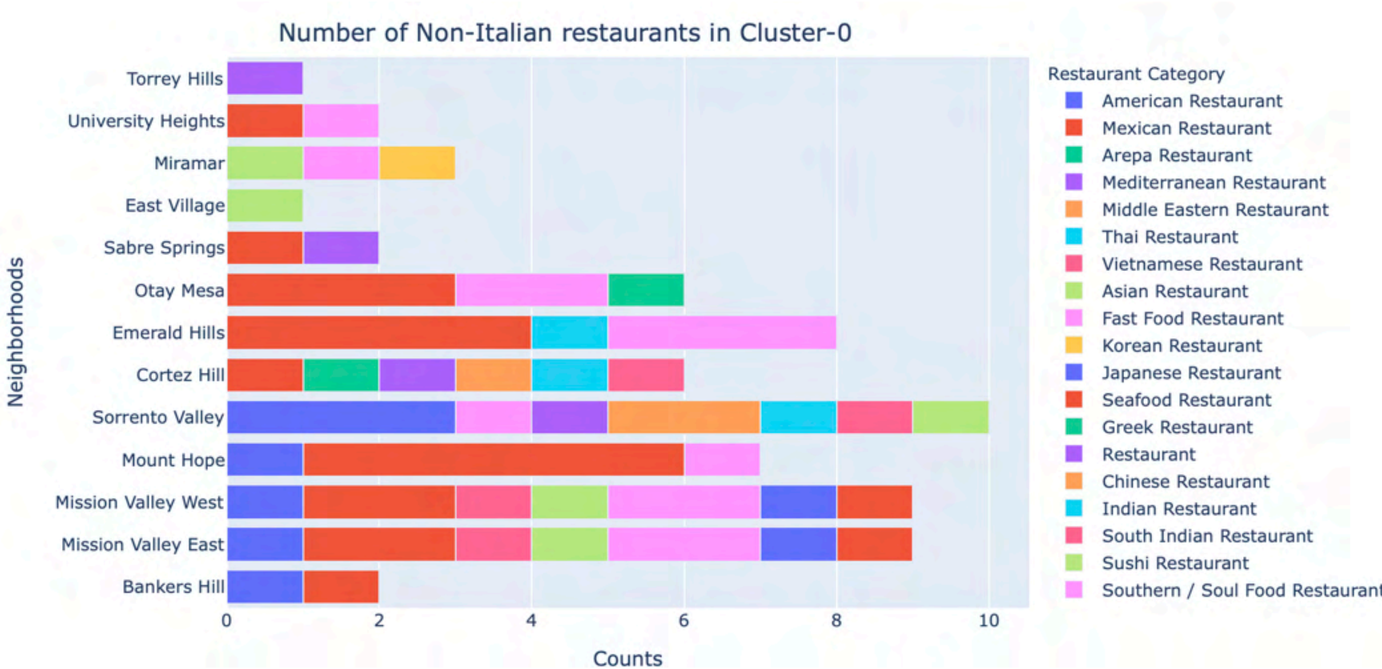
Cluster-0, **Cluster-1**, and **Cluster-3** have been selected for further analysis to get the following information:

- All the neighborhoods
- Number and category of restaurants
- Population base: Foot or car traffic
- Parking
- Demographical information such as population, housing, crime rate etc.

1. Cluster-0: offices and automotive shops
2. Cluster-1: Multiple Venues - residential buildings, college buildings, Salon, office, bank. restaurants
3. Cluster-2: Government buildings
4. Cluster-3: Doctor's and dentist's place
5. Cluster-4: Zoo exhibit
6. Cluster-5: Automotive shops

Cluster-0

- Number of neighborhoods: **20**
- Total number of restaurants: **97**
- Unique restaurants: **22**
- Number of neighborhoods with no Italian restaurants: **13**
- Horizontal stack bar chart displaying number of non-Italian restaurants:

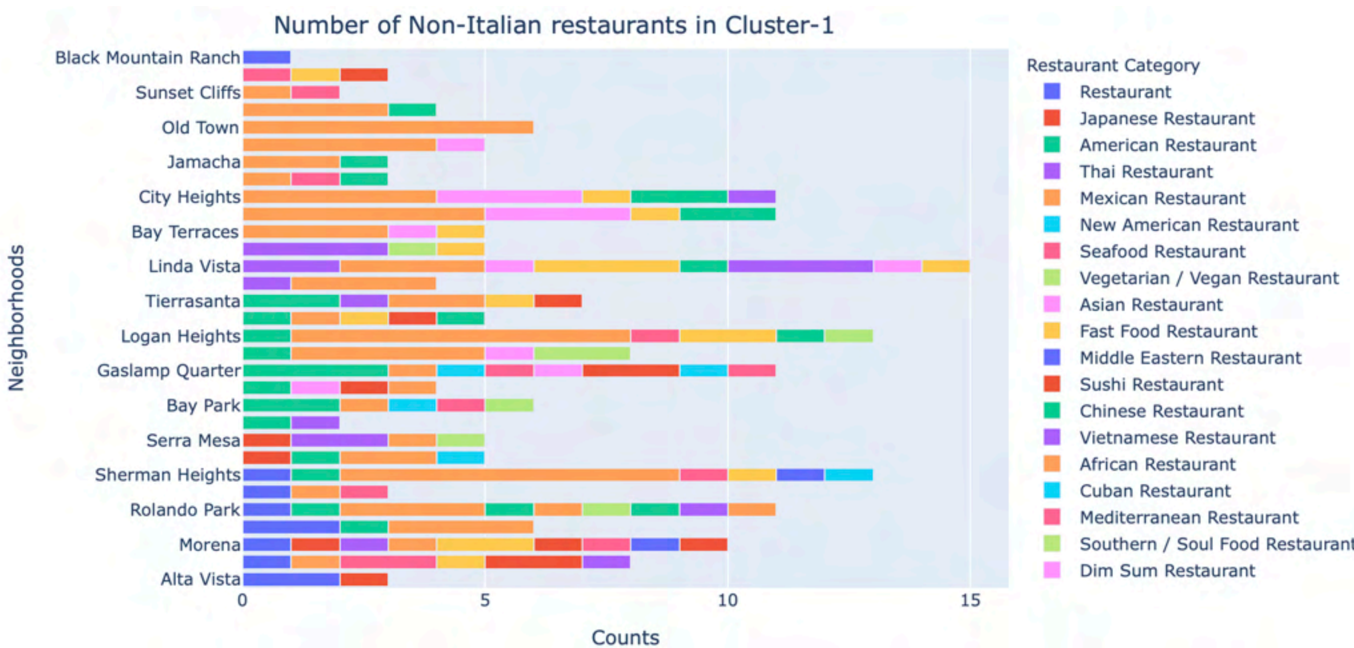


Selected neighborhoods having restaurants of **4 or more** different categories:

1. Sorrento Valley
2. Mission Valley East
3. Mission Valley West
4. Cortez Hill

Cluster-1

- Number of neighborhoods: **47**
- Total number of restaurants: **304**
- Unique restaurants: **33**
- Number of neighborhoods with no Italian restaurants: **31**
- Horizontal stack bar chart displaying number of non-Italian restaurants:

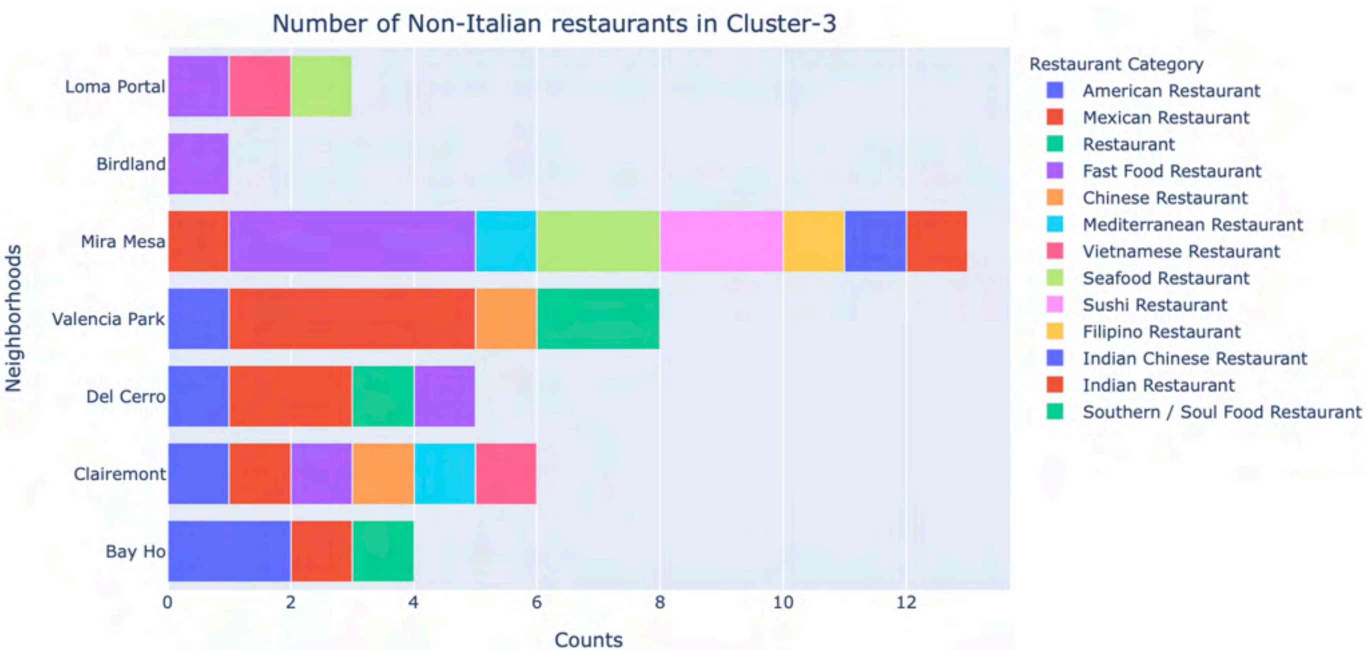


Selected neighborhoods having restaurants of **4 or more** different categories:

1. Linda Vista
2. Logan Heights
3. Sherman Heights
4. Ronaldo Park
5. Gaslamp Quarter
6. City Heights
7. Chollas View
8. Morena
9. Del Mar Heights
10. Islenair
11. Serra Mesa
12. Del Mar Mesa
13. Rancho Penasquitos
14. Tierrasanta

Cluster-3

- Number of neighborhoods: **9**
- Total number of restaurants: **50**
- Unique restaurants: **15**
- Number of neighborhoods with no Italian restaurants: **7**
- Horizontal stack bar chart displaying number of non-Italian restaurants:



Selected neighborhoods having restaurants of **4 or more** different categories:

1. Mira Mesa
2. Valencia Park
3. Clairemont
4. Del Cerro

Merging Demographical Data with Clusters

8. Merging Demographical Data with Clusters
   - Dataframe with demographical data of San Diego

	Neighborhoods	Population	Median Home Value	Median Rent	Median Household Income	Diversity	Housing	Cost of Living	Weather	Commute	Crime and Safety	Walkable to Restaurants	Parking
0	Mission Valley East	31911.0	518367.0	2121.0	90192.0	8.0	4.0	2.0	8.0	7.4	4.6	7.9	4.8
1	Mission Valley West	31911.0	518367.0	2121.0	90192.0	8.0	4.0	2.0	8.0	7.4	4.6	4.8	5.8
2	Linda Vista	38659.0	493701.0	1782.0	66863.0	8.0	3.0	2.0	8.0	8.0	5.0	5.2	6.9

   - Dataframe with a new column **Overall Rating** after summing up the ratings of features − Diversity, Commute, Crime and Safety, Walkable to Restaurants, and Parking:

	Neighborhoods	Population	Median Home Value	Median Rent	Median Household Income	Diversity	Housing	Cost of Living	Weather	Commute	Crime and Safety	Walkable to Restaurants	Parking	Overall Rating
0	Mission Valley East	31911.0	518367.0	2121.0	90192.0	8.0	4.0	2.0	8.0	7.4	4.6	7.9	4.8	32.7
1	Mission Valley West	31911.0	518367.0	2121.0	90192.0	8.0	4.0	2.0	8.0	7.4	4.6	4.8	5.8	30.6
2	Linda Vista	38659.0	493701.0	1782.0	66863.0	8.0	3.0	2.0	8.0	8.0	5.0	5.2	6.9	33.1

Merging Demographical Data with Clusters

- Final dataframe after merging demographical data with clusters:

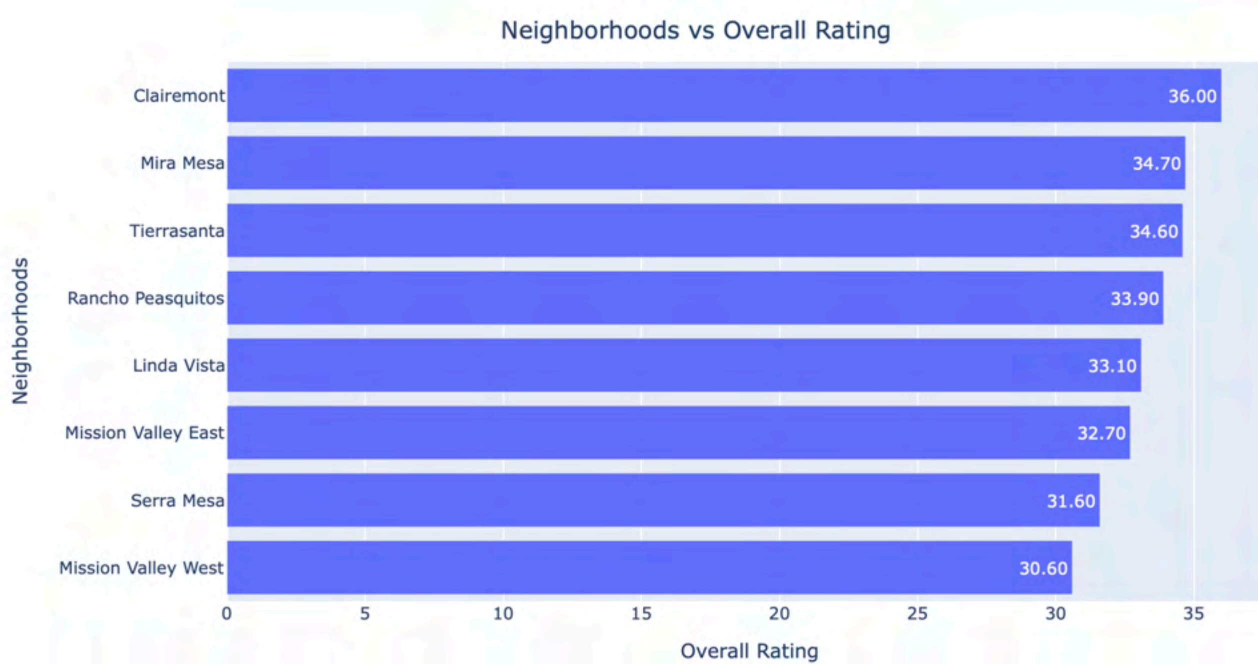
	Neighborhoods	Population	Median Home Value	Median Rent	Median Household Income	Diversity	Housing	Cost of Living	Weather	Commute	Crime and Safety	Walkable to Restaurants	Parking	Overall Rating	Latitude
0	Mission Valley East	31911.0	518367.0	2121.0	90192.0	8.0	4.0	2.0	8.0	7.4	4.6	7.9	4.8	32.7	32.770502
1	Mission Valley West	31911.0	518367.0	2121.0	90192.0	8.0	4.0	2.0	8.0	7.4	4.6	4.8	5.8	30.6	32.770502
2	Linda Vista	38659.0	493701.0	1782.0	66863.0	8.0	3.0	2.0	8.0	8.0	5.0	5.2	6.9	33.1	32.789841
3	Rancho Peasquitos	60519.0	797946.0	2559.0	144186.0	8.0	5.4	2.0	8.0	6.0	6.6	4.5	8.8	33.9	32.957710
4	Serra Mesa	32491.0	547197.0	2190.0	82307.0	8.0	3.4	2.0	8.0	6.6	6.0	4.7	6.3	31.6	32.802899
5	Mira Mesa	87785.0	533150.0	2140.0	101381.0	8.0	4.6	2.0	8.0	7.4	5.4	6.6	7.3	34.7	32.915602
6	Clairemont	89234.0	601041.0	1949.0	88347.0	8.0	3.4	2.0	8.0	7.4	4.5	7.6	8.5	36.0	32.819505
7	Tierrasanta	35542.0	576392.0	2326.0	98759.0	8.0	4.0	2.0	8.0	6.6	6.0	5.7	8.3	34.6	32.829216

Results

A. List of the selected **8** Neighborhoods

1. **Clairemont**
2. **Mira Mesa**
3. **Tierrasanta**
4. **Rancho Penasquitos**
5. **Linda Vista**
6. **Mission Valley East**
7. **Serra Mesa**
8. **Mission Valley West**

B. Bar plot -- Neighborhoods vs. Overall Rating



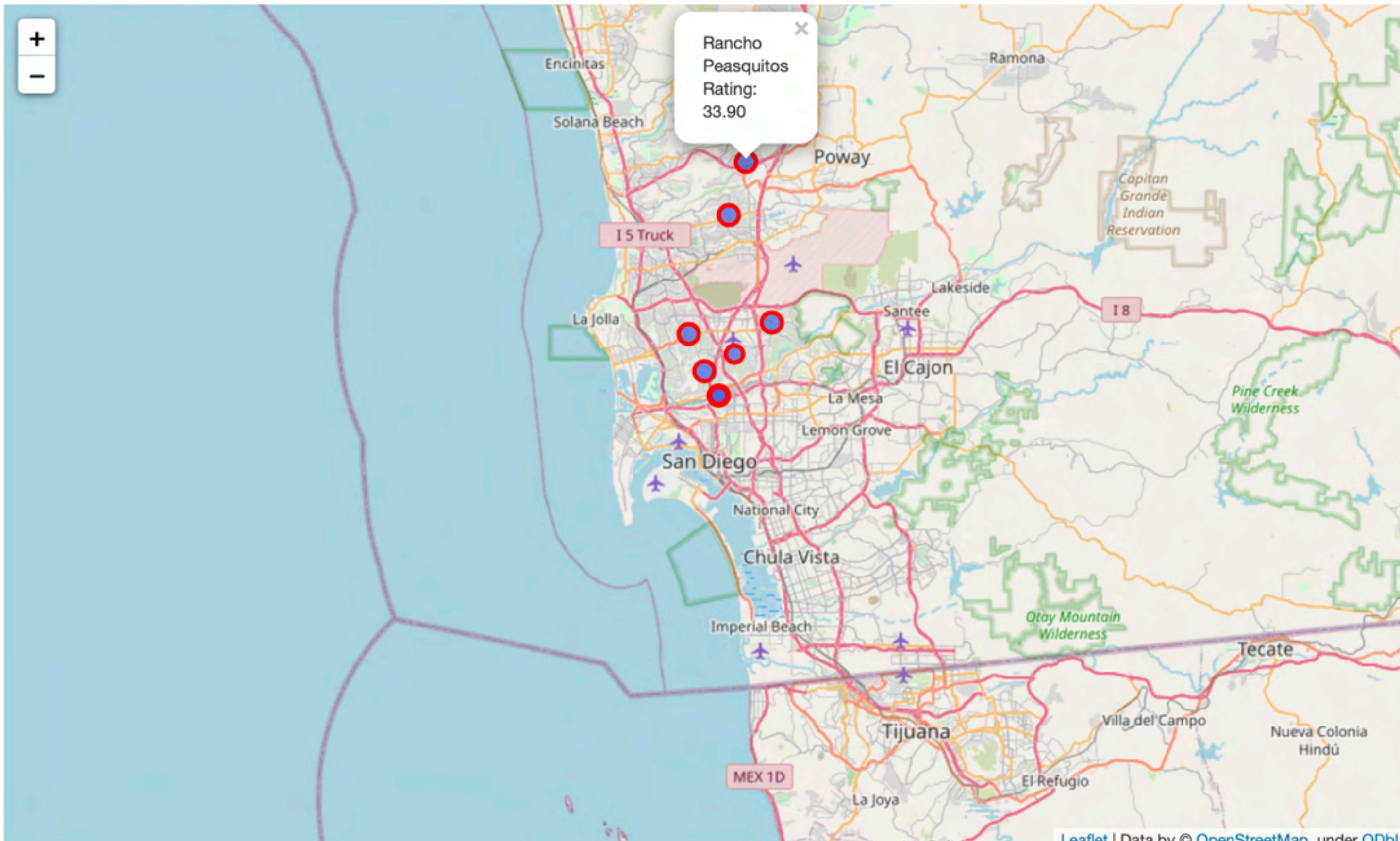
Results

C. Final dataframe combining neighborhoods cluster data and demographical data

	Neighborhoods	Population	Median Home Value	Median Rent	Median Household Income	Diversity	Housing	Cost of Living	Weather	Commute	Crime and Safety	Walkable to Restaurants	Parking	Overall Rating	Latitude
0	Mission Valley East	31911.0	518367.0	2121.0	90192.0	8.0	4.0	2.0	8.0	7.4	4.6	7.9	4.8	32.7	32.770502
1	Mission Valley West	31911.0	518367.0	2121.0	90192.0	8.0	4.0	2.0	8.0	7.4	4.6	4.8	5.8	30.6	32.770502
2	Linda Vista	38659.0	493701.0	1782.0	66863.0	8.0	3.0	2.0	8.0	8.0	5.0	5.2	6.9	33.1	32.789841
3	Rancho Peasquitos	60519.0	797946.0	2559.0	144186.0	8.0	5.4	2.0	8.0	6.0	6.6	4.5	8.8	33.9	32.957710
4	Serra Mesa	32491.0	547197.0	2190.0	82307.0	8.0	3.4	2.0	8.0	6.6	6.0	4.7	6.3	31.6	32.802899
5	Mira Mesa	87785.0	533150.0	2140.0	101381.0	8.0	4.6	2.0	8.0	7.4	5.4	6.6	7.3	34.7	32.915602
6	Clairemont	89234.0	601041.0	1949.0	88347.0	8.0	3.4	2.0	8.0	7.4	4.5	7.6	8.5	36.0	32.819505
7	Tierrasanta	35542.0	576392.0	2326.0	98759.0	8.0	4.0	2.0	8.0	6.6	6.0	5.7	8.3	34.6	32.829216

Results

D. Map of San Diego with the selected 8 neighborhoods superimposed on top



Discussion

- **Exploratory data analysis** (**EDA**) and ***K*-means clustering** algorithm are used in order to discover a few precise locations for opening up an **Italian** restaurant.
- For ***K*-means cluster** modeling, the optimal value of the ***k*** is set to **6** (k=6) using the **Elbow Method**.
- Specific names have been assigned to the **6 clusters** depending on the characteristics and different venues associated with these clusters.
- **Cluster-0**, **Cluster-1**, and **Cluster-3** have been selected for further analysis.
- Following information have been derived using **EDA**:
  - Number of neighborhoods in Cluster-0, Cluster-1, Cluster-3 are **20**, **47**, and **9** respectively.
  - Total number of restaurants in Cluster-0, Cluster-1, Cluster-3 are **97**, **304**, and **50** respectively.
  - Number of unique restaurants category in Cluster-0, Cluster-1, Cluster-3 are **22**, **33**, and **15** respectively
  - Number of neighborhoods with no Italian restaurants in Cluster-0, Cluster-1, Cluster-3 are **13**, **31**, and **7** respectively.
- Finally, following **8** neighborhoods have been selected based:
  1. **Clairemont**
  2. **Mira Mesa**
  3. **Tierrasanta**
  4. **Rancho Penasquitos**
  5. **Linda Vista**
  6. **Mission Valley East**
  7. **Serra Mesa**
  8. **Mission Valley West**

Conclusion

Objective of this project was to discover a few promising neighborhoods of San Diego with having no Italian restaurants in the vicinity so that the stakeholders -- more specifically, the restaurant entrepreneurs can select a optimal location for opening up a new Italian Restaurant.

By using **Foursquare API**, basic **exploratory data analysis**, and ***K*-means clustering** algorithm, some neighborhoods from three selected clusters have been identified. Then, these neighborhoods are explored to find out the locations which satisfy some basic requirements of this project. Finally, **demographical information** of San Diego has been merged with these selected neighborhoods to find more precise locations for an Italian restaurant.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics and locations of these neighborhoods.