Applied Data Science Capstone Project -The Battle of Neighborhoods

New Restaurant Site Evaluation Recommendations

Michael Hall
mjhall@ucsd.edu
December, 2018

Contents

BUSINESS PROBLEM

XYZ Restaurant Group has an existing restaurant in Solana Beach, CA . This restaurant has been very successful and XYZ Restaurant Group is looking to open second restaurant somewhere within San Diego County. They believe that the location is one of the key contributors to the success of any restaurant. They want to evaluate additional communities that are closely aligned to the characteristics of Solana Beach. Based on prior experience and research on other successful restaurants, XYZ Restaurant Group would like the have the locations evaluated on the demographics of the community. Some of the characteristics that they would like to be evaluated include:

- Population
- Income
- Nearby venues and attractions
- Age
- Housing ownership

The question that XYZ ultimately wants answered is - "*which of the target communities most closely aligns with characteristics of our current, successful restaurant in Solana Beach?*"

Stakeholders

- XYZ President
- XYZ Marketing Vice President
- XYZ General Manager
- XYZ Board of Directors

Target Audience

- XYZ Board of Directors

DATA

Data was gathered from multiple sources to evaluate the communities. Data on the competing restaurants and nearby venues was obtained from Foursquare. The community data was obtained from San Diego Association of Governments (SANDAG). SANDAG is made up the 18 cities and unincorporated county governments. SANDAG develops annual demographic estimates and long range forecasts in addition to maintaining information from the U.S. Census Bureau. SANDAG provided the 2010 US Census data summarized for each community. SANDAG will also provided the spatial data that describes each of the communities.

Title:	Community Venues
Description:	Venue name, location, category and other firmographic information
Use:	Identify competitors and venues within a community
Format:	JSON
Source:	https://www.foursquare.com

Title:	Community Demographics
Description:	2010 US Census demographics aggregated and summarized by community
Use:	Profile of the population and other demographics of the community
Format:	Excel
Source:	http://datasurfer.sandag.org/

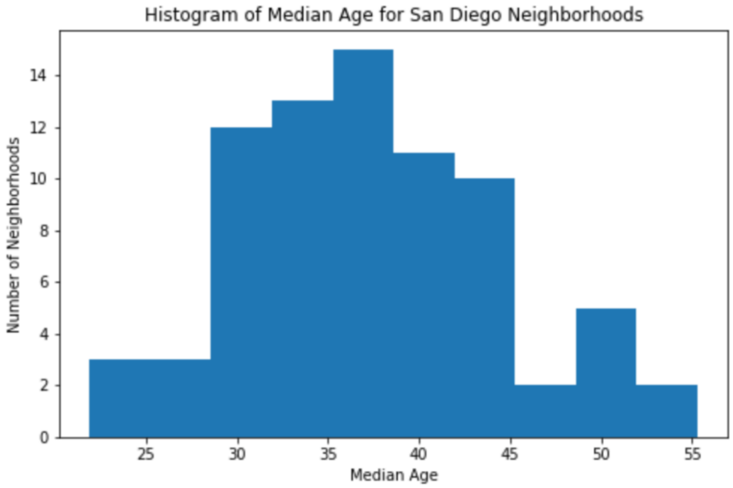
Title:	Community Location
Description:	Boundaries of the communities
Use:	Plotting the data on maps
Format:	Geojson
Source:	http://rdw.sandag.org/

METHODOLOGY

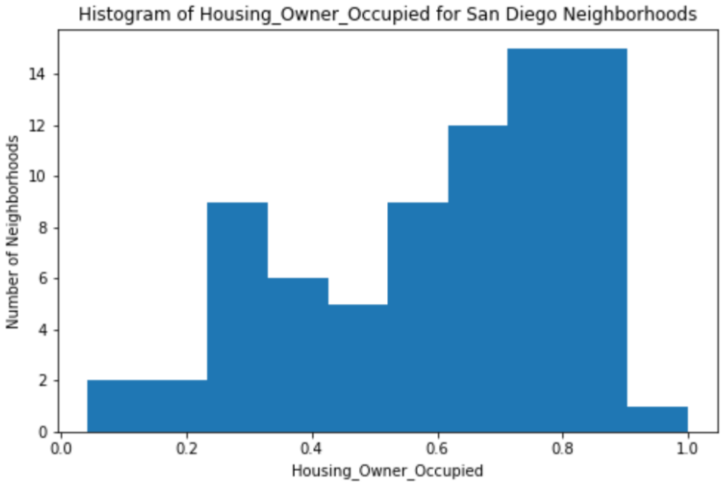
The 2010 census data obtained from San Diego Association of Governments (SANDAG) was used to identify the demographic characteristics of the neighborhoods. This data was already aggregated and summarized for each of the neighborhoods. Additional data preparation was done to filter the various statistics down to just the the few key attributes to be evaluated:

- Median household income
- Median age
- % of home ownership
- Number of households

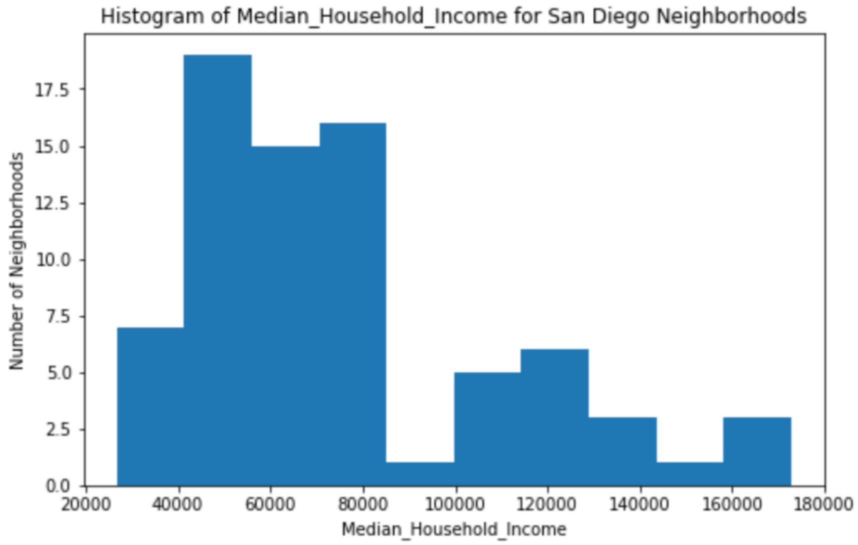
Each of the attributes were analyzed utilizing histograms and were segmented into three categories each based on the distributions



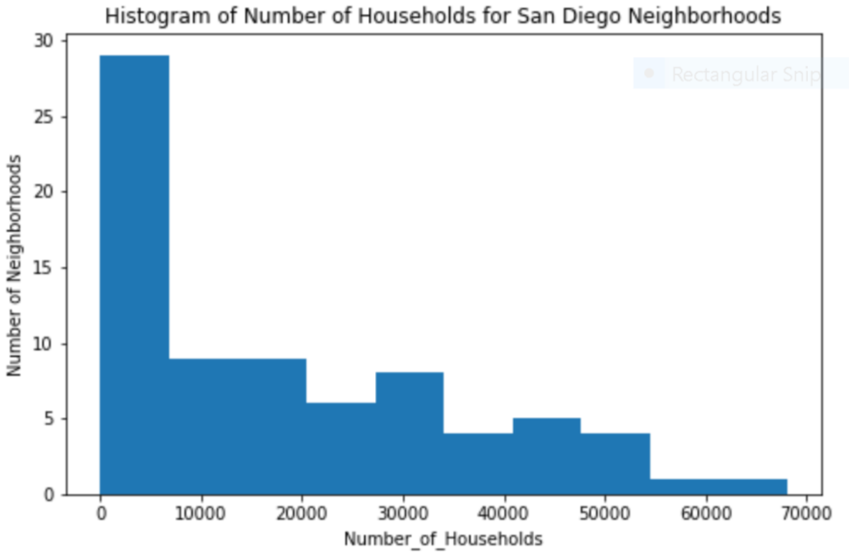
The median age was split into three groups: young < 35, middle 35-45, older > 45



The % of owner occupied housing was segemented by : Rent for less than 30%, Rent or Own for 30-60% and Own for greater than 60%.



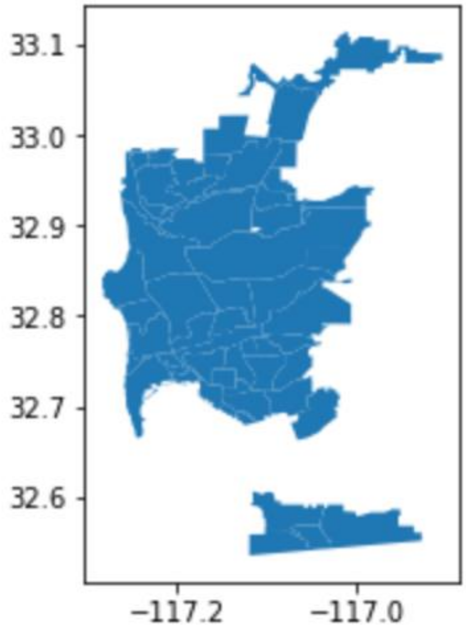
Median household income was segmented into three attributes: Low Income for less than $60,000, Middle Income for $60,000 to $120,000 and High Income for greater than $120,000.

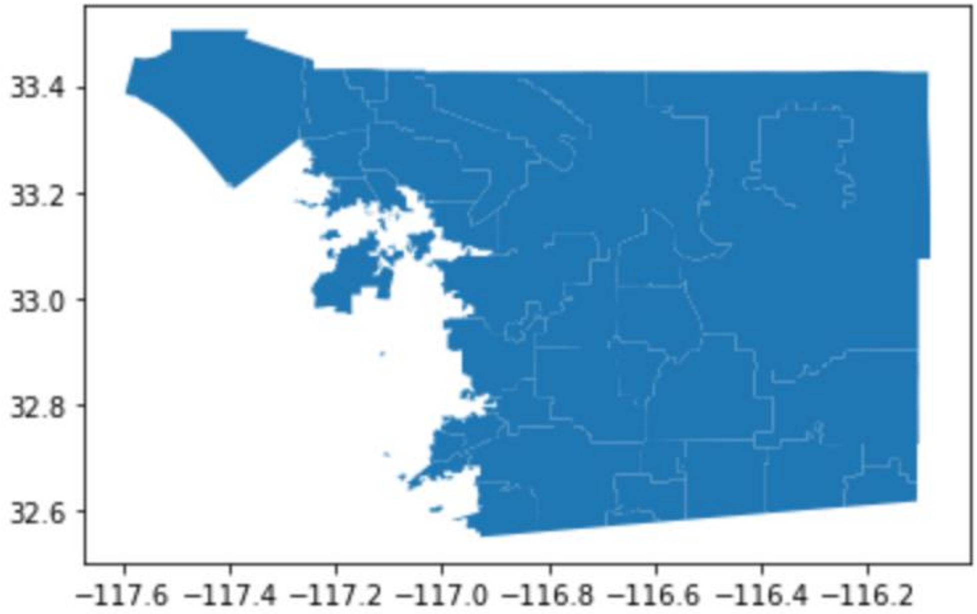


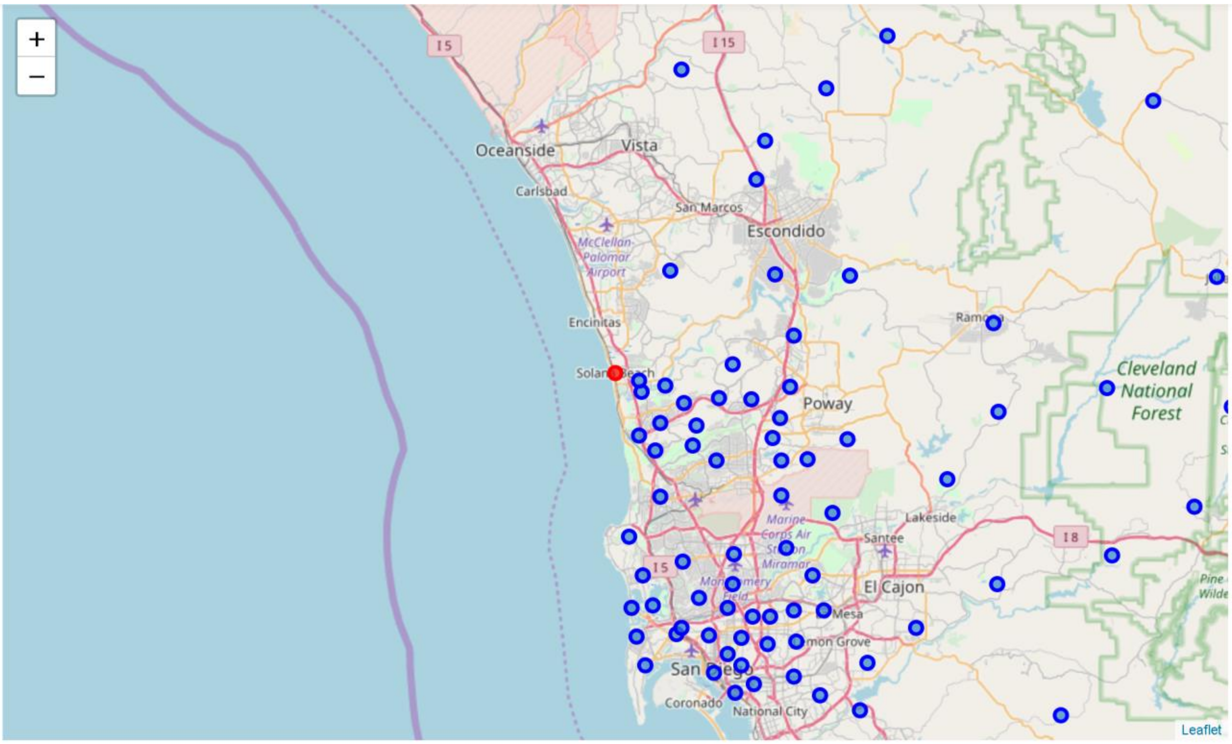
Number of households was segmented into Small for less than 5,000, Middle for 5,000 to 40,000 and High for greater than 40,000.

Each of the transformed variables were converted to binary variables.

The neighborhood boundary data was downloaded from SANDAG utilizing the provided API. The data was provided in GEOJSON format and was provided in two separate datasets. The first included the neighborhoods in the city of San Diego and the others were for the remainder of the county outside of the city. The centroids were calculated for each of the polygons provided. In some cases, multiple polygons were included for non-contiguous boundaries in rural areas. In those instances, just the first polygon and associated centroid were retained for processing. The two resulting datasets were merged and provided the latitude and longitude for each of the neighborhoods.

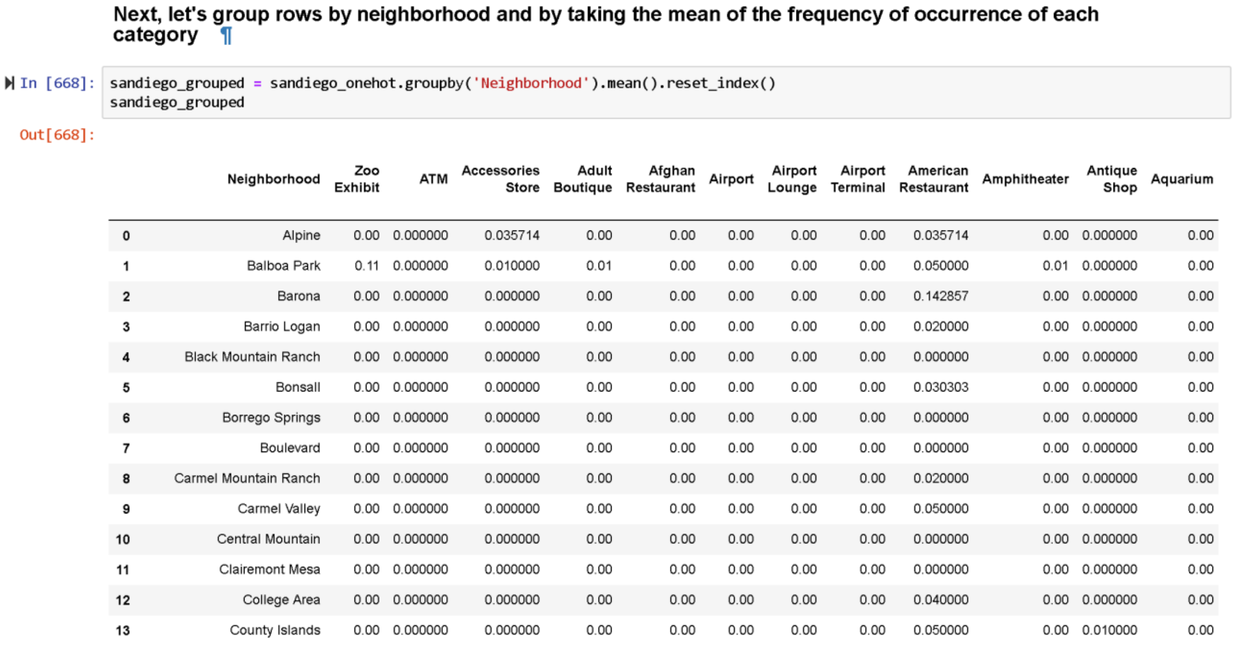






The target existing restaurant is identified by the red dot in Solana Beach. It is located in the Via De La Valle neighborhood.

The resulting neighborhood locations were utilized to obtain up to 100 venues within 2,500 meters of the centroid of the neighborhood. The resulting venues were then aggregated by the category of the venue for each neighborhood. The top 10 categories were identified for each venue.

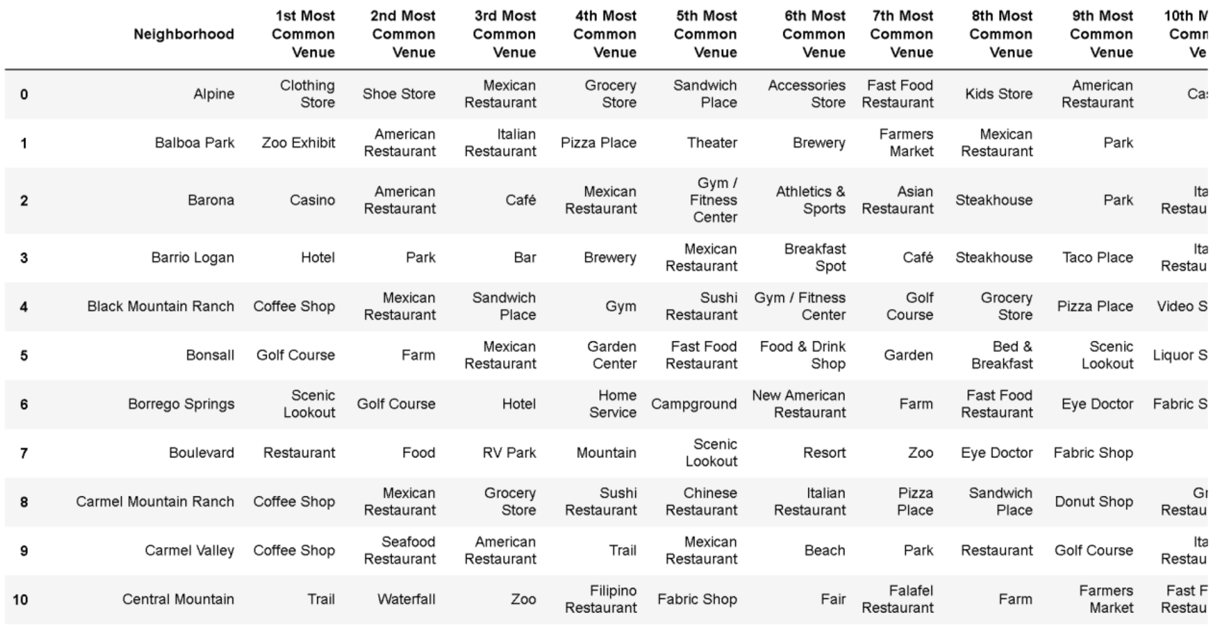


Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category ¶

```
In [668]: sandiego_grouped = sandiego_onehot.groupby('Neighborhood').mean().reset_index()
          sandiego_grouped
```

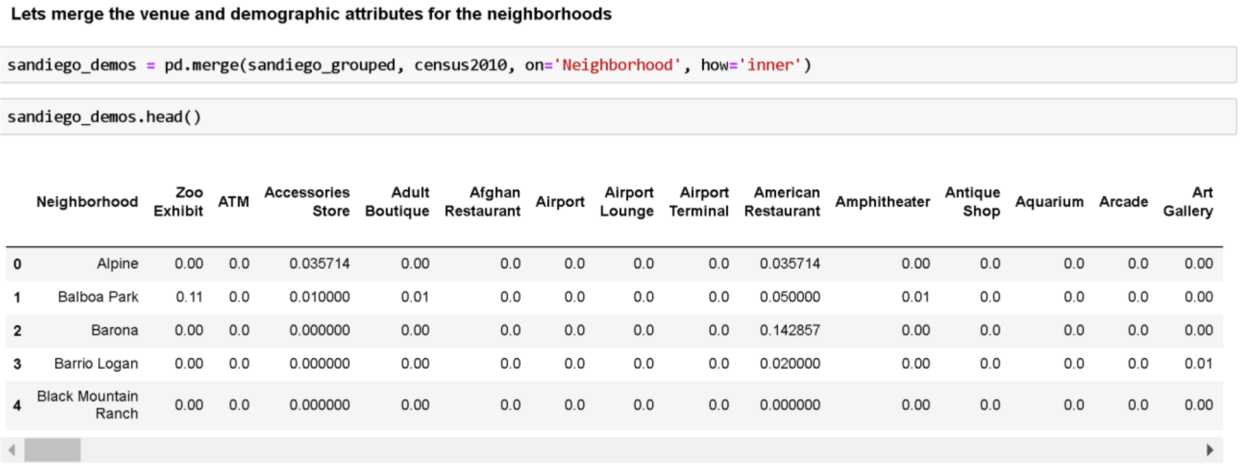
Out[668]:

	Neighborhood	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Lounge	Airport Terminal	American Restaurant	Amphitheater	Antique Shop	Aquarium
0	Alpine	0.00	0.000000	0.035714	0.00	0.00	0.00	0.00	0.00	0.035714	0.00	0.000000	0.00
1	Balboa Park	0.11	0.000000	0.010000	0.01	0.00	0.00	0.00	0.00	0.050000	0.01	0.000000	0.00
2	Barona	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.142857	0.00	0.000000	0.00
3	Barrio Logan	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.020000	0.00	0.000000	0.00
4	Black Mountain Ranch	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00
5	Bonsall	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.030303	0.00	0.000000	0.00
6	Borrego Springs	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00
7	Boulevard	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00
8	Carmel Mountain Ranch	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.020000	0.00	0.000000	0.00
9	Carmel Valley	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.050000	0.00	0.000000	0.00
10	Central Mountain	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00
11	Clairemont Mesa	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00
12	College Area	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.040000	0.00	0.000000	0.00
13	County Islands	0.00	0.000000	0.000000	0.00	0.00	0.00	0.00	0.00	0.050000	0.00	0.010000	0.00



	Neighborhood	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 M Comn Ve
0	Alpine	Clothing Store	Shoe Store	Mexican Restaurant	Grocery Store	Sandwich Place	Accessories Store	Fast Food Restaurant	Kids Store	American Restaurant	Ca:
1	Balboa Park	Zoo Exhibit	American Restaurant	Italian Restaurant	Pizza Place	Theater	Brewery	Farmers Market	Mexican Restaurant	Park	
2	Barona	Casino	American Restaurant	Café	Mexican Restaurant	Gym / Fitness Center	Athletics & Sports	Asian Restaurant	Steakhouse	Park	Ita Restau
3	Barrio Logan	Hotel	Park	Bar	Brewery	Mexican Restaurant	Breakfast Spot	Café	Steakhouse	Taco Place	Ita Restau
4	Black Mountain Ranch	Coffee Shop	Mexican Restaurant	Sandwich Place	Gym	Sushi Restaurant	Gym / Fitness Center	Golf Course	Grocery Store	Pizza Place	Video S
5	Bonsall	Golf Course	Farm	Mexican Restaurant	Garden Center	Fast Food Restaurant	Food & Drink Shop	Garden	Bed & Breakfast	Scenic Lookout	Liquor S
6	Borrego Springs	Scenic Lookout	Golf Course	Hotel	Home Service	Campground	New American Restaurant	Farm	Fast Food Restaurant	Eye Doctor	Fabric S
7	Boulevard	Restaurant	Food	RV Park	Mountain	Scenic Lookout	Resort	Zoo	Eye Doctor	Fabric Shop	
8	Carmel Mountain Ranch	Coffee Shop	Mexican Restaurant	Grocery Store	Sushi Restaurant	Chinese Restaurant	Italian Restaurant	Pizza Place	Sandwich Place	Donut Shop	Gr Restau
9	Carmel Valley	Coffee Shop	Seafood Restaurant	American Restaurant	Trail	Mexican Restaurant	Beach	Park	Restaurant	Golf Course	Ita Restau
10	Central Mountain	Trail	Waterfall	Zoo	Filipino Restaurant	Fabric Shop	Fair	Falafel Restaurant	Farm	Farmers Market	Fast F Restau

With all of the data prepped and normalized, the demographics data and venue data were combined and prepared for completing the cluster analysis



Lets merge the venue and demographic attributes for the neighborhoods

```
sandiego_demos = pd.merge(sandiego_grouped, census2010, on='Neighborhood', how='inner')
```

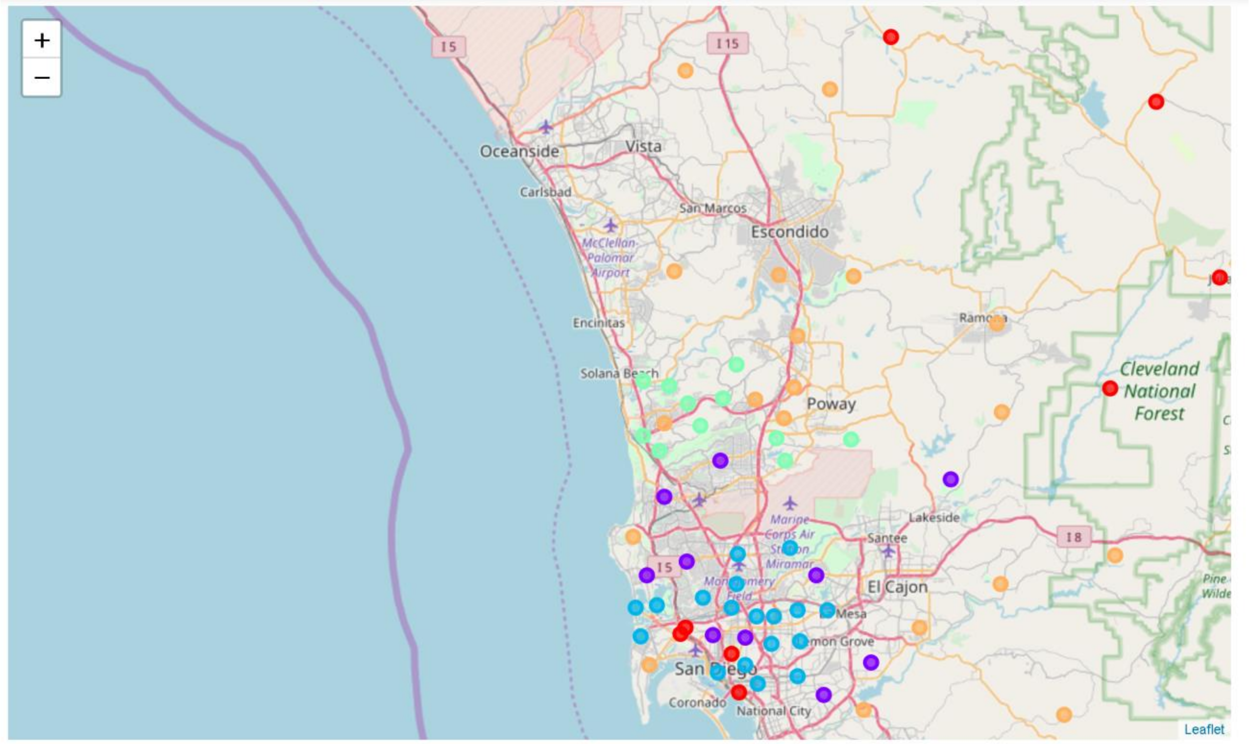
```
sandiego_demos.head()
```

	Neighborhood	Zoo Exhibit	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Lounge	Airport Terminal	American Restaurant	Amphitheater	Antique Shop	Aquarium	Arcade	Art Gallery
0	Alpine	0.00	0.0	0.035714	0.00	0.0	0.0	0.0	0.0	0.035714	0.00	0.0	0.0	0.0	0.00
1	Balboa Park	0.11	0.0	0.010000	0.01	0.0	0.0	0.0	0.0	0.050000	0.01	0.0	0.0	0.0	0.00
2	Barona	0.00	0.0	0.000000	0.00	0.0	0.0	0.0	0.0	0.142857	0.00	0.0	0.0	0.0	0.00
3	Barrio Logan	0.00	0.0	0.000000	0.00	0.0	0.0	0.0	0.0	0.020000	0.00	0.0	0.0	0.0	0.01
4	Black Mountain Ranch	0.00	0.0	0.000000	0.00	0.0	0.0	0.0	0.0	0.000000	0.00	0.0	0.0	0.0	0.00

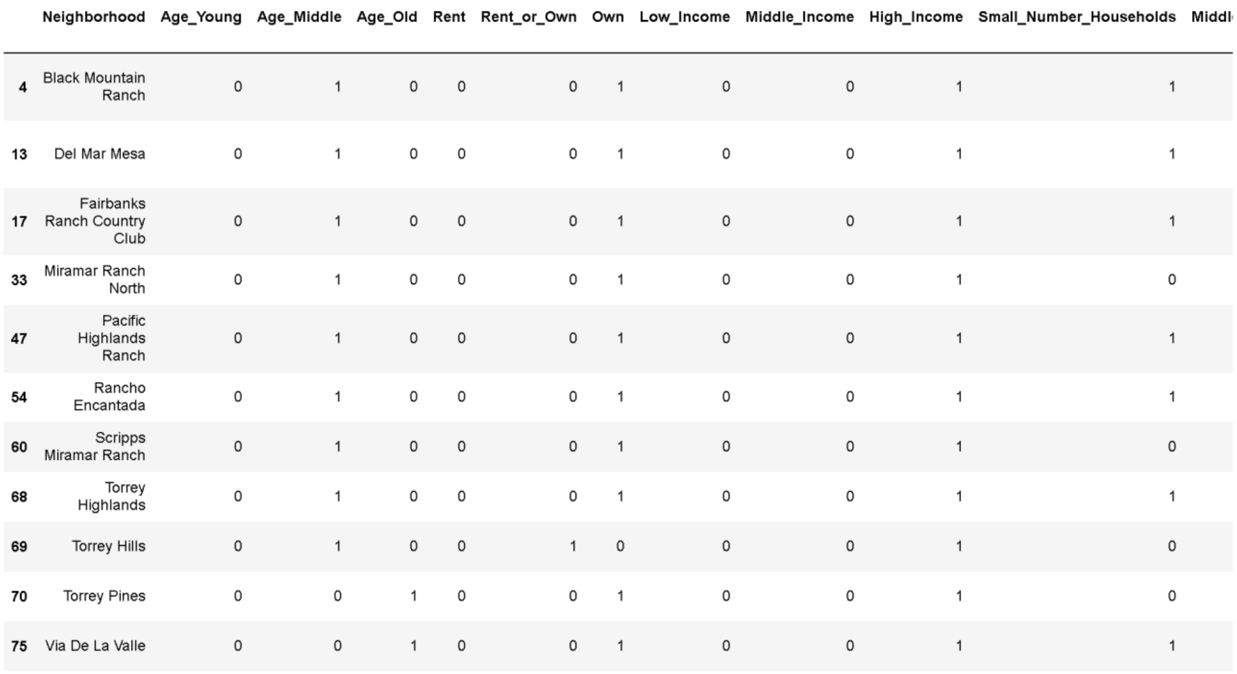
In order to identify the neighborhoods that are similar to the current restaurant location, k-means clustering algorithm was utilized with a target of 5 clusters.

RESULTS

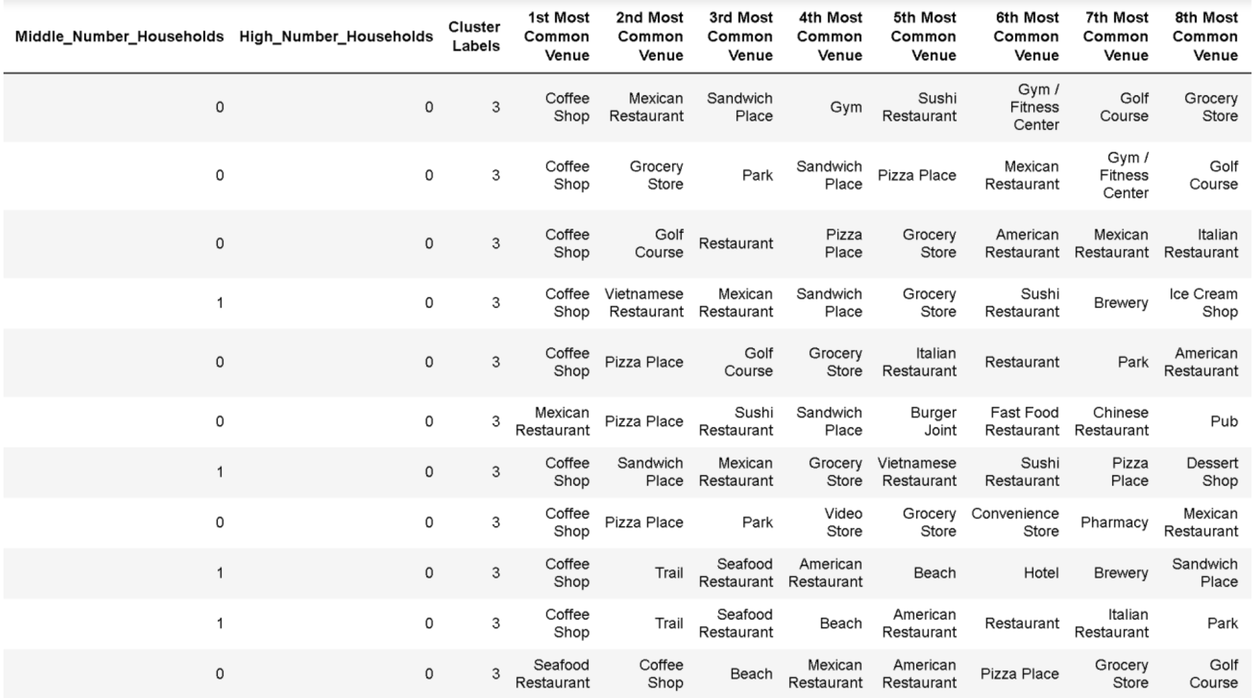
The resulting clusters were plotted to identify any insights.



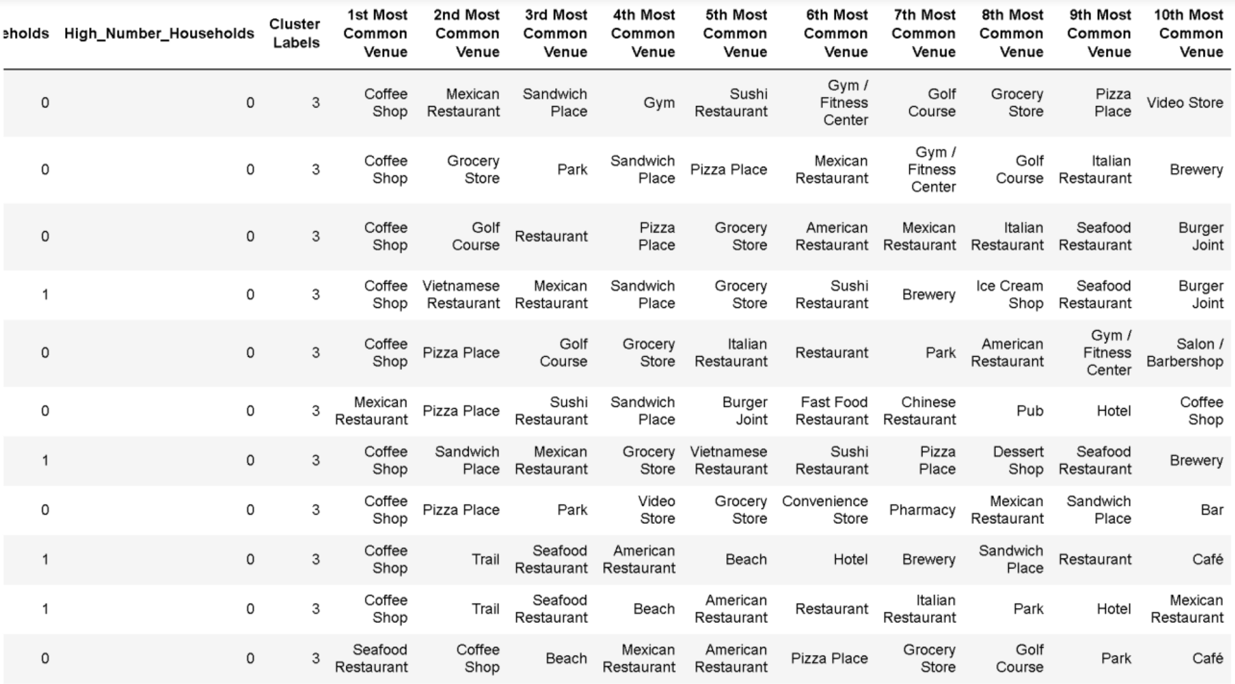
The current restaurant fell within cluster 3. The other neighborhoods within the cluster all contained very similar charactersitics.



	Neighborhood	Age_Young	Age_Middle	Age_Old	Rent	Rent_or_Own	Own	Low_Income	Middle_Income	High_Income	Small_Number_Households	Middl
4	Black Mountain Ranch	0	1	0	0	0	1	0	0	1	1	
13	Del Mar Mesa	0	1	0	0	0	1	0	0	1	1	
17	Fairbanks Ranch Country Club	0	1	0	0	0	1	0	0	1	1	
33	Miramar Ranch North	0	1	0	0	0	1	0	0	1	0	
47	Pacific Highlands Ranch	0	1	0	0	0	1	0	0	1	1	
54	Rancho Encantada	0	1	0	0	0	1	0	0	1	1	
60	Scripps Miramar Ranch	0	1	0	0	0	1	0	0	1	0	
68	Torrey Highlands	0	1	0	0	0	1	0	0	1	1	
69	Torrey Hills	0	1	0	0	1	0	0	0	1	0	
70	Torrey Pines	0	0	1	0	0	1	0	0	1	0	
75	Via De La Valle	0	0	1	0	0	1	0	0	1	1	



Middle_Number_Households	High_Number_Households	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
0	0	3	Coffee Shop	Mexican Restaurant	Sandwich Place	Gym	Sushi Restaurant	Gym / Fitness Center	Golf Course	Grocery Store
0	0	3	Coffee Shop	Grocery Store	Park	Sandwich Place	Pizza Place	Mexican Restaurant	Gym / Fitness Center	Golf Course
0	0	3	Coffee Shop	Golf Course	Restaurant	Pizza Place	Grocery Store	American Restaurant	Mexican Restaurant	Italian Restaurant
1	0	3	Coffee Shop	Vietnamese Restaurant	Mexican Restaurant	Sandwich Place	Grocery Store	Sushi Restaurant	Brewery	Ice Cream Shop
0	0	3	Coffee Shop	Pizza Place	Golf Course	Grocery Store	Italian Restaurant	Restaurant	Park	American Restaurant
0	0	3	Mexican Restaurant	Pizza Place	Sushi Restaurant	Sandwich Place	Burger Joint	Fast Food Restaurant	Chinese Restaurant	Pub
1	0	3	Coffee Shop	Sandwich Place	Mexican Restaurant	Grocery Store	Vietnamese Restaurant	Sushi Restaurant	Pizza Place	Dessert Shop
0	0	3	Coffee Shop	Pizza Place	Park	Video Store	Grocery Store	Convenience Store	Pharmacy	Mexican Restaurant
1	0	3	Coffee Shop	Trail	Seafood Restaurant	American Restaurant	Beach	Hotel	Brewery	Sandwich Place
1	0	3	Coffee Shop	Trail	Seafood Restaurant	Beach	American Restaurant	Restaurant	Italian Restaurant	Park
0	0	3	Seafood Restaurant	Coffee Shop	Beach	Mexican Restaurant	American Restaurant	Pizza Place	Grocery Store	Golf Course



eholds	High_Number_Households	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	10th Most Common Venue
0	0	3	Coffee Shop	Mexican Restaurant	Sandwich Place	Gym	Sushi Restaurant	Gym / Fitness Center	Golf Course	Grocery Store	Pizza Place	Video Store
0	0	3	Coffee Shop	Grocery Store	Park	Sandwich Place	Pizza Place	Mexican Restaurant	Gym / Fitness Center	Golf Course	Italian Restaurant	Brewery
0	0	3	Coffee Shop	Golf Course	Restaurant	Pizza Place	Grocery Store	American Restaurant	Mexican Restaurant	Italian Restaurant	Seafood Restaurant	Burger Joint
1	0	3	Coffee Shop	Vietnamese Restaurant	Mexican Restaurant	Sandwich Place	Grocery Store	Sushi Restaurant	Brewery	Ice Cream Shop	Seafood Restaurant	Burger Joint
0	0	3	Coffee Shop	Pizza Place	Golf Course	Grocery Store	Italian Restaurant	Restaurant	Park	American Restaurant	Gym / Fitness Center	Salon / Barbershop
0	0	3	Mexican Restaurant	Pizza Place	Sushi Restaurant	Sandwich Place	Burger Joint	Fast Food Restaurant	Chinese Restaurant	Pub	Hotel	Coffee Shop
1	0	3	Coffee Shop	Sandwich Place	Mexican Restaurant	Grocery Store	Vietnamese Restaurant	Sushi Restaurant	Pizza Place	Dessert Shop	Seafood Restaurant	Brewery
0	0	3	Coffee Shop	Pizza Place	Park	Video Store	Grocery Store	Convenience Store	Pharmacy	Mexican Restaurant	Sandwich Place	Bar
1	0	3	Coffee Shop	Trail	Seafood Restaurant	American Restaurant	Beach	Hotel	Brewery	Sandwich Place	Restaurant	Café
1	0	3	Coffee Shop	Trail	Seafood Restaurant	Beach	American Restaurant	Restaurant	Italian Restaurant	Park	Hotel	Mexican Restaurant
0	0	3	Seafood Restaurant	Coffee Shop	Beach	Mexican Restaurant	American Restaurant	Pizza Place	Grocery Store	Golf Course	Park	Café

The characteristics of Cluster three are:

- Middle Age
- Own home
- High Income

- Small to medium number of households
- Coffee shops, Mexican restaurants and seafood restaurants are the most common venues

The other neigborhoods identified within the cluster include: Black Mountain Ranch, Del Mar Mesa, Fairbanks Ranch Country Club, Miramar Ranch North, Pacific Highlands Ranch, Rancho Encantanda, Scripps Miramar Ranch, Torrey Highlands, Torrey Hills and Torrey Pines.

The resulting analysis was able to clearly identify a set of neighborhoods that were closely aligned to the neighborhood where the current restaurant is located. In addition, the identified attributes were consistent with the original hypothesis that income, age and home ownership would be contributing factors to a successful location. The one inconsistent result was that the lower population/number of households was identified. This was the result of the current restaurant being in a middle sized population center.

The recommendation is to evaluate potential restaurant sites in the identified neighborhoods: Black Mountain Ranch, Del Mar Mesa, Fairbanks Ranch Country Club, Miramar Ranch North, Pacific Highlands Ranch, Rancho Encantanda, Scripps Miramar Ranch, Torrey Highlands, Torrey Hills or Torrey Pines.

CONCLUSION

The use of k-means clustering proved to be an effective tool for identifying neighborhoods similar to the current restaurant location. Additional factors that could be evaluated in the future include targeting specific locations rather than general neighborhood centroids. In this particular case, certain areas covered broad areas and the results may not have been as accurate versus looking at smaller, more targeted locations. In addition, additional demographic variables, such as number of children and commute times could be utilized.