IBM Applied Data Science Capstone Project

Profiling California Cities

By Mina Abdelmessih

March 2020.

Table of Contents

Introduction

California is the most populous state in the USA. It is home to about 40 million and is the largest state by area. If it were a country on its own, it would be the fifth largest economy in the world. California has an exceptionally diverse geography, climate and culture. There's a place for anyone wishing to move to california. This project aims to profile cities across california. The target audience would be anyone planning on relocating to or starting a business in California.Foursquare venue data together with Zillow housing data will be used for this purpose.

Methodology

Data Collection

For this project, housing data was first acquired from Zillow.com, a major online real estate database. The used dataset includes all cities in california and their latest average median home price as of the time of this analysis. The dataset was in csv format and was fairly clean and easily accessible for analysis.

This was used as a starting point to then pull coordinates for each city using the geocoder module in python. Using the geocoder module was found to be a little challenging as it was not consistently responsive and prone to errors. Therefore, the cities' geographical coordinates were collected over multiple requests then stored into a CSV file titled *"df_ll"*

Using city coordinates, calls were then made to Foursquare places API to provide a count of the venues in each city across 10 major categories:

1. Arts & Entertainment,
2. College & University,
3. Event,
4. Food,
5. Nightlife Spot,
6. Outdoors & Recreation,
7. Professional & Other Places,
8. Residence,
9. Shop & Service,
10. Travel & Transport

Due to the number of API calls necessary for this project being in excess of the free Foursquare account limits per day, this data was collected over multiple iterations and stored into a CSV file titled *"df_venues".*

Data Preparation

Given the limitations on the Foursquare free API developer account, the list of california cities included in this analysis had to be reduced. Only cities with median home prices within the median - 75th percentile range were selected, resulting in a list of almost 270 cities. Also data types of *'City'* and *'Median_Home_Price'* had to be converted to "string" and "float" respectively for the purpose of later analysis.

Upon reviewing the data extracted with geocoder, it was revealed that city coordinates were significantly off of California's. These had to be manually retrieved and entered into the dataset. Finally when all cities' coordinates were identified, these were visualized on a map of california - using a Folium map. The selected set appears to be an adequate representative sample.

Finally, after obtaining venue count per city per category, the dataset below was generated

	City	Median_Home_Price	Latitude	Longitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Del Rey Oaks	652483.0	36.593293	-121.834951	70	76	11	198	62	189	127	24	204	176
1	San Diego	652175.0	32.717421	-117.162771	219	189	15	247	249	224	275	148	235	233
2	Wilton	650138.0	38.420310	-121.206367	3	5	0	37	8	89	74	10	132	35
3	Windsor	644473.0	38.547133	-122.816380	39	59	2	189	84	127	189	30	157	103
4	Garden Grove	643528.0	33.774629	-117.946372	176	177	28	232	239	219	290	138	206	222

The dataset ready for modeling had 14 columns (*'City', 'Median_Home_Price', 'Latitude', 'Longitude', 'Arts & Entertainment', 'College & University', 'Event', 'Food', 'Nightlife Spot', 'Outdoors & Recreation', 'Professional & Other Places', 'Residence', 'Shop & Service', 'Travel & Transport')* and 276 rows.

Modeling

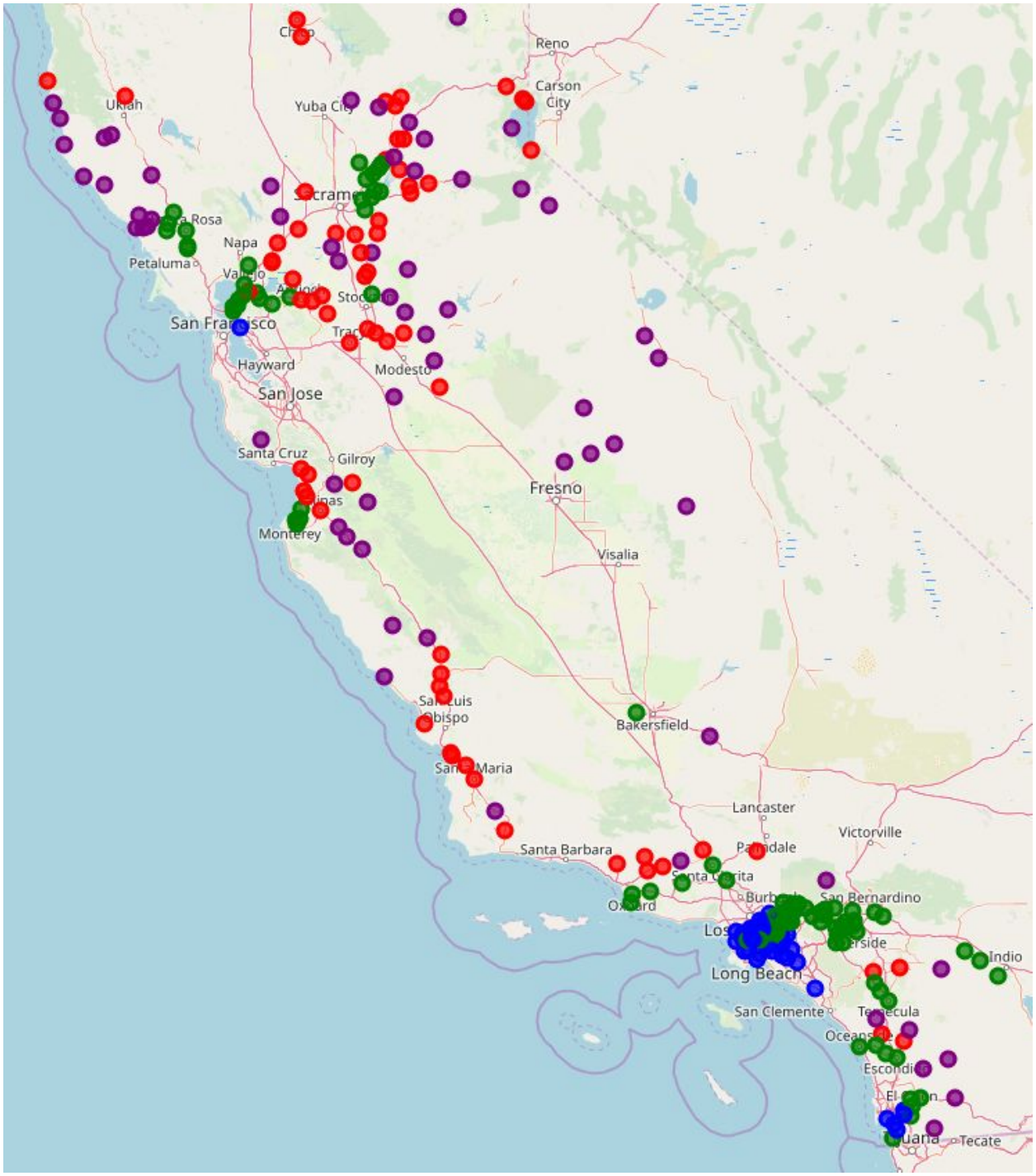
K-means clustering was used as the data modeling approach for this project. The model was initiated with four (4) clusters with the following result:

```
[ ] # prepare the data for modeling
    from sklearn.preprocessing import StandardScaler
    Clus = StandardScaler().fit_transform(df_c)
    Clus_df=pd.DataFrame(Clus)
    Clus_df.head()
```

	0	1	2	3	4	5	6	7	8	9	10
0	0.320084	0.000405	0.453747	0.524542	-0.162664	0.972276	-0.264969	-0.395646	0.809916	1.292639	1.903349
1	3.086027	1.577724	0.934954	1.088445	2.276452	1.470825	1.497237	2.533318	1.284581	2.181714	1.899239
2	-0.923663	-0.990654	-0.869573	-1.328284	-0.867008	-0.452153	-0.896030	-0.726336	-0.292533	-0.906650	1.872057
3	-0.255381	-0.236891	-0.628969	0.420968	0.124290	0.089130	0.473252	-0.253922	0.090262	0.154000	1.796464
4	2.287802	1.410221	2.498878	0.915822	2.146018	1.399604	1.675839	2.297112	0.840539	2.010138	1.783854

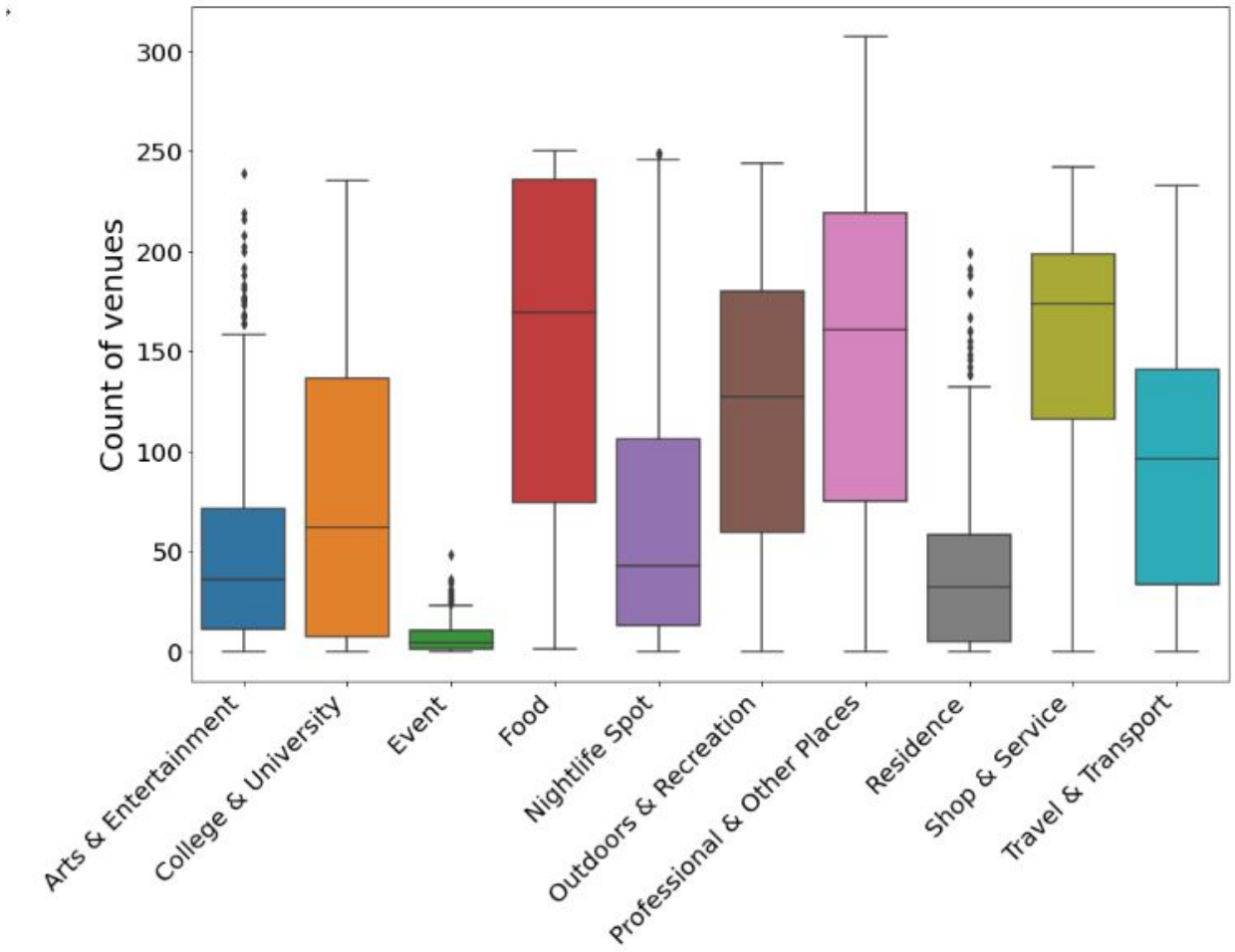
Column headers in the above model represent the foursquare categories in the order identified under Data collection above.

Cluster labels were then inserted into the original data set and visualized on a california map as below


Reno
Carson City
Yuba City
Ukiah
Napa
Petaluma
Hayward
San Jose
Modesto
Santa Cruz
Gilroy
Monterey
Fresno
Visalia
Bakersfield
Lancaster
Victorville
Santa Barbara
San Bernardino
Long Beach
San Clemente
Indio
Tecate

Results

Using the original dataset - before modeling - the box plot below was generated to visualize the venue count per category retrieved from Foursquare

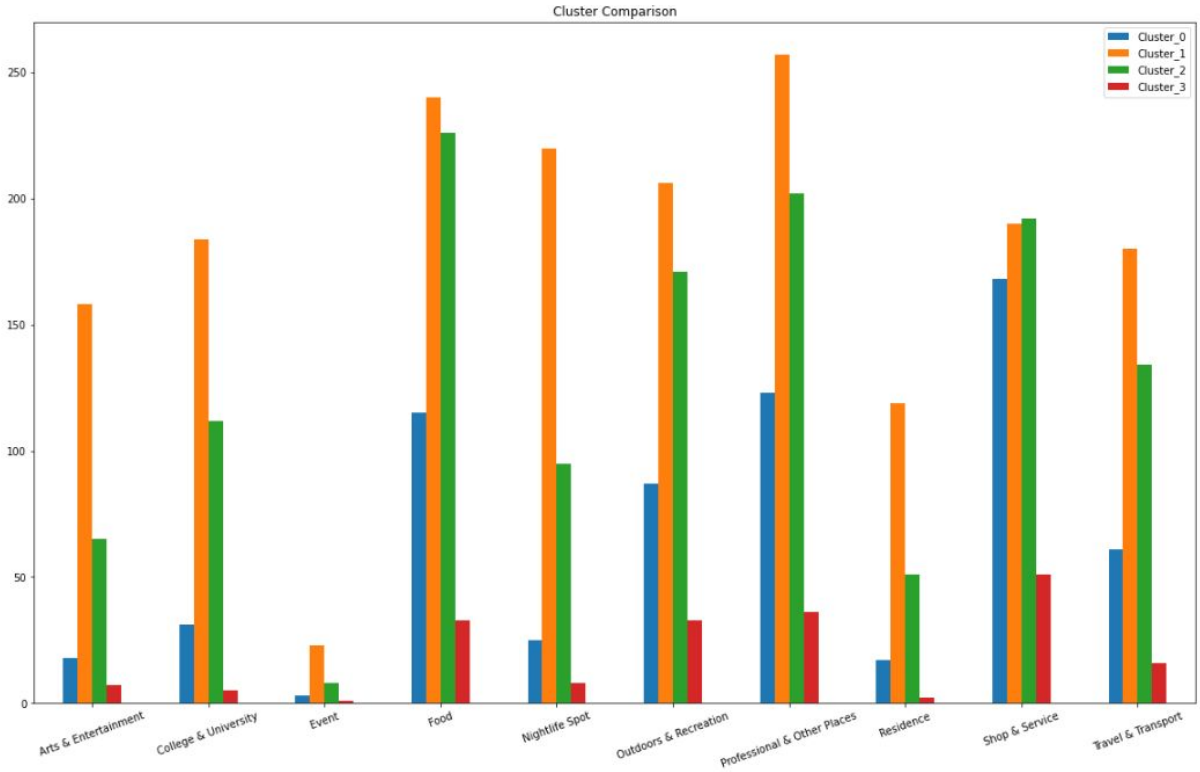


Interesting to note the correlation between “College & University” and “Nightlife”, as well as “Food” with “Professional”

Below is the mean of venue count per cluster. It is clear that the model clustered cities based on the number of venue counts, which is a reflection of urban development and amenities in the cities.

	Category	Cluster_0	Cluster_1	Cluster_2	Cluster_3
0	Arts & Entertainment	18.0	158.0	65.0	7.0
1	College & University	31.0	184.0	112.0	5.0
2	Event	3.0	23.0	8.0	1.0
3	Food	115.0	240.0	226.0	33.0
4	Nightlife Spot	25.0	220.0	95.0	8.0
5	Outdoors & Recreation	87.0	206.0	171.0	33.0
6	Professional & Other Places	123.0	257.0	202.0	36.0
7	Residence	17.0	119.0	51.0	2.0
8	Shop & Service	168.0	190.0	192.0	51.0
9	Travel & Transport	61.0	180.0	134.0	16.0

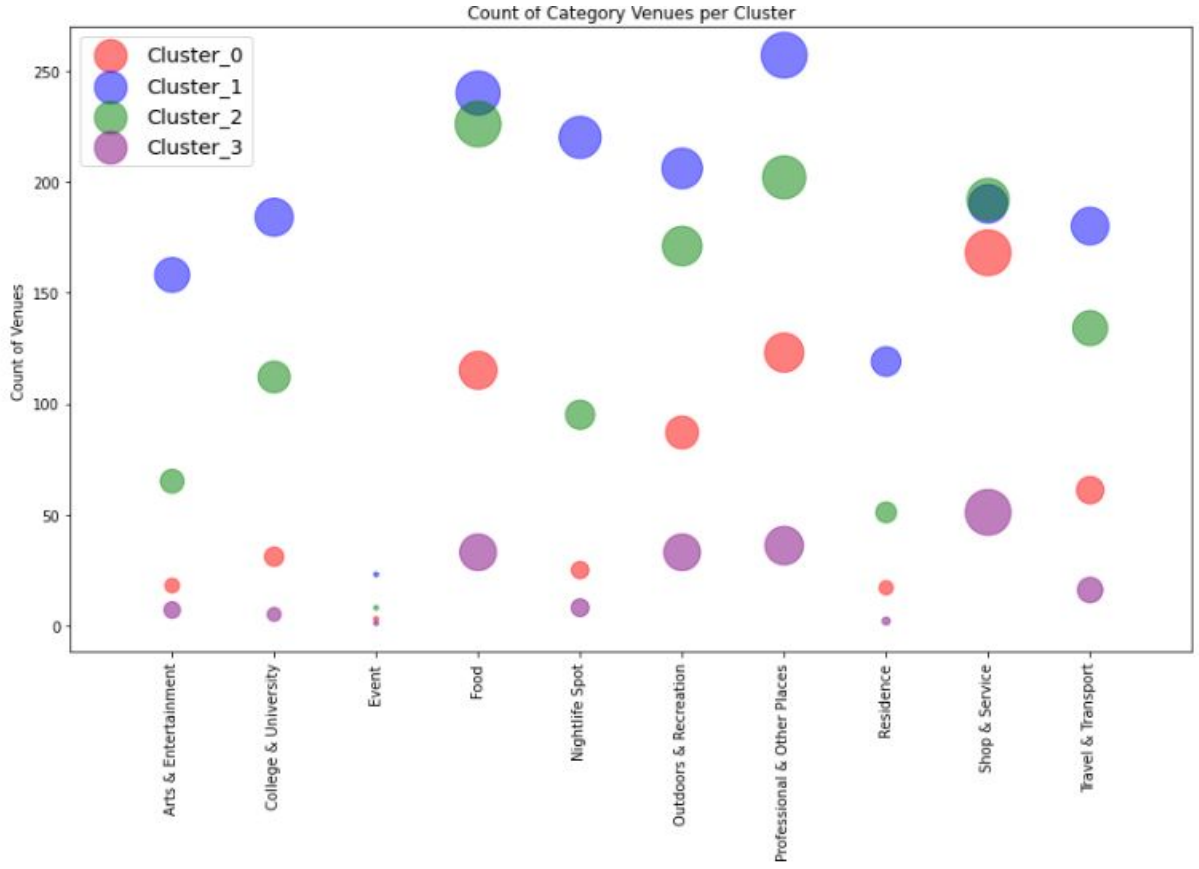
Barchart of the above values for illustration:



Evidently, Clusters 1 and 2 were the highest in terms of venue count across all categories, followed by Cluster 0 then Cluster 3. Reviewing those clusters on the map reveals that:

- Cluster 1 (blue dots on the map snapshot above) is truly the most urbanized and populous cities in California, including cities in west Los Angeles, San Diego and San Francisco. Indeed, these are the typically the largest and most populous cities in California.
- Cluster 2 (green dots on the map)includes cities that are surrounding Cluster 1. They are quite developed but not as Cluster 1. These include East Los Angeles, Sacramento and cities surrounding San Diego and San Francisco.
- Cluster 0 (red dots) appears to be cities surrounding Cluster 2
- Cluster 3 (purple dots) appears to be the least developed cities. They are typically furthest out from metropolitan areas and Urban cities- the “quaint” towns of california.

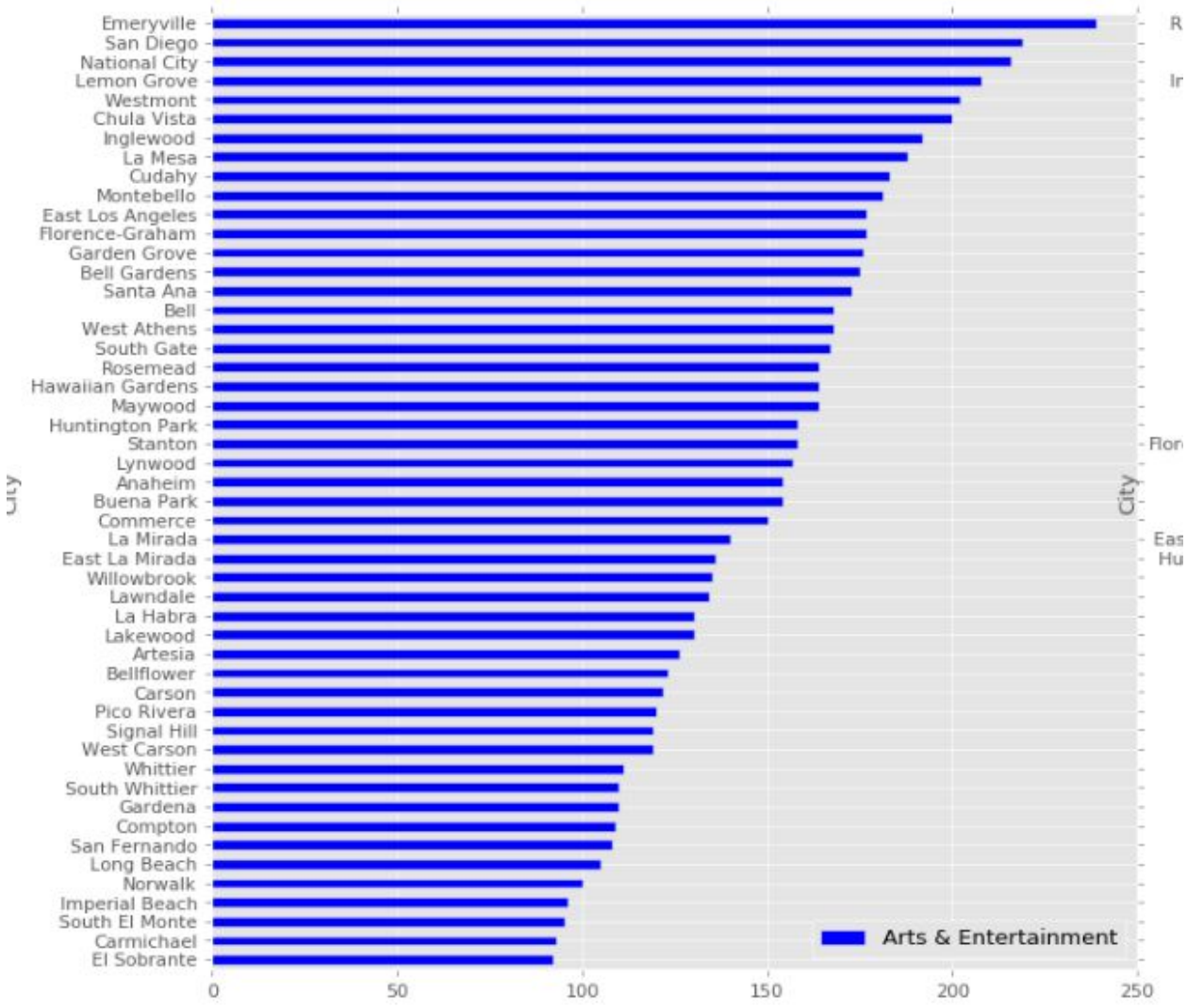
Here’s another way to visualize the above. Note how Clusters 1 and 2 are the highest across categories, followed by Cluster 0 then 3.



Discussion

It seems Cluster 1 would be a nice new home for someone looking to relocate. These cities offer a reasonable compromise between affordability and amenities.

Further, using the data collected from Foursquare we can actually rate the cities across each category to find the best cities in California to live and possibly start a business. For example, below are the best cities in terms of “Arts & Entertainment”, using the count of venues in that category as a measure:



If I were to choose a new City in California to move into, I would choose based on my favorite categories, namely “Arts & Entertainment”, “Food”, “Nightlife” and “Outdoors & Recreation.

Applying the same sort above across these categories, here the cities that consistently showed up as one of the top 30:

- Inglewood
- Westmont
- Lemon Grove
- Emeryville
- San Diego

As it turns out, all top cities were in Cluster 1

City	Cluster
San Diego	1
Lawndale	1
Inglewood	1
Emeryville	1
La Mesa	1
West Carson	1
Montebello	1
Chula Vista	1
West Athens	1
Westmont	1
st Los Angeles	1
Lemon Grove	1
Bell Gardens	1
National City	1

Conclusion

Profiling cities using foursquare proved to be an adequate tool to research and profile cities across california.

Acknowledgements

Credit is due to the Coursera-IBM Data Science for the educational resources and tools followed throughout this project.

Also, credit is due to Stanislav Rogozhin (https://github.com/theptyza) for his idea of using venue count per category.