Applied Data Science Capstone

IBM Data Science Professional Certificate

Best area to live in LA

Jingying Li

January, 2021

Contents

1. Introduction

Moving to a new city is always exciting and exhausting. It means the beginning of a new life. It also means that you need to find a neighborhood to live in and start your new life. I was moving to Los Angeles from an out-of-state a year ago, and the scene of searching for a new neighborhood online was still vivid. Due to my unfamiliarity with the city, I spent a lot of time searching the Internet for the most suitable area to live in Los Angeles. In this project, I intend to collect data from various sources, analyze and cluster neighborhoods of Los Angeles based on safety, convenience and economy in order to provide some useful information to people who have just come to this city or want to find a new place to live.

1.1 Business Problem

In order to make this project more applicative, here we will concretize the problem we are about to solve, and introduce some assumptions and constraints.

For choosing the most suitable area/neighborhood to live, the factors we consider include:

- Community safety: The lower the crime rate per capita, the more suitable the area to live.
- Convenience level: We will measure the number of grocery stores, restaurants and shopping malls in the neighborhood.
- Entertainment activities: We will measure the number of art and entertainment venues, outdoor recreation venues and nightlife spots in the community.
- Rent level: We will tend to live in areas where the average rent is relatively low.

1.2 Target Audience

This project is aimed at singles or families without children who want to relocate to Los Angeles. So educational resources are not considered for the time being. In addition, due to the pandemic, most people are able to work from home. So the impact of commuting time on location selection is not considered here. Last but not least, the audience of this project is not limited to newcomers who have just arrived in LA. It is also instructive for people who already live in LA but want to reconsider their residential communities.

2. Source of Data

In this project, we will fetch or extract data from the following sources:

- List of regions and neighborhoods in Los Angeles: https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_in_Los_Angeles
- Los Angeles Rental Market Trends: https://www.rentcafe.com/average-rent-market-trends/us/ca/los-angeles/
- Violent crime in the City of Los Angeles from December 30, 2019 to June 28, 2020: https://maps.latimes.com/neighborhoods/violent-crime/neighborhood/list/

- Number of Arts & Entertainment, Food, Nightlife Spot, Outdoors & Recreation, and Shop & Service in every neighborhood - **Foursquare API**
- Coordinates of all neighborhoods and venues - **GeoPy Nominatim geocoding**

3. Methodology

In essence, our methodology is to cluster the neighborhoods in Los Angeles based on the crime rate, rent, and the number of different types of venues in each neighborhood to obtain a couple of clusters. In the end, readers can choose a neighborhood from a certain cluster that suits them best according to the factors they value.

4. Analysis and Modeling

4.1 Import Required Libraries

```
import numpy as np # library to handle data in a vectorized manner

import pandas as pd # library for data analsysis
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

import json # library to handle JSON files

!conda install -c conda-forge geopy --yes # uncomment this line if you haven't completed the Foursquare API lab
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

# Matplotlib and associated plotting modules
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors

# import k-means from clustering stage
from sklearn.cluster import KMeans

from bs4 import BeautifulSoup
import re

!pip install folium
#!conda install -c conda-forge folium=0.5.0 --yes # uncomment this line if you haven't completed the Foursquare API lab
import folium # map rendering library

print('Libraries imported.')
```

4.2 Neighborhoods Data

4.2.1 Scrape and clean up the list of neighborhoods

There is no relevant dataset available for the list of neighborhoods in Los Angeles, therefore we need to scrap this from a Wikipedia page.

```
# download data and parse it:
url = requests.get('https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_in_Los_Angeles').text
soup = BeautifulSoup(url, "html.parser")

lis = []
for li in soup.findAll('li'):
    if li.find(href="/wiki/Portal:Los_Angeles"):
        break
    if li.find(href=re.compile("^/wiki/")):
        lis.append(li)
    if li.text=='Pico Robertson[34]': #Pico Robertson is the only item on the list that does not have a hyperlink reference
        lis.append(li)
len(lis)
```

```
200
```

After cleaning up the unnecessary information, we got the following Los Angeles neighborhood list.

```
neigh = []
for i in range(0,len(lis)):
    neigh.append(lis[i].text.strip())

df = pd.DataFrame(neigh)
df.columns = ['Neighborhood']
df.head()
```

	Neighborhood
0	Angelino Heights[1]
1	Angeles Mesa[2]
2	Angelus Vista[2]
3	Arleta[3][1]
4	Arlington Heights[3]

```
df['Neighborhood'] = df.Neighborhood.str.partition('[')[0] #Removes the citation and reference brackets
df['Neighborhood'] = df.Neighborhood.str.partition(',')[0] #Removes the alternatives for 'Bel Air'
df=df[df.Neighborhood!='Baldwin Hills/Crenshaw'] #Removes redundancy as 'Baldwin Hills' and 'Crenshaw' exist already
df=df[df.Neighborhood!='Hollywood Hills West'] #Removes redundancy as it has the same coordinates as 'Hollywood Hills'
df=df[df.Neighborhood!='Brentwood Circle'] #Removes redundancy as it has the same coordinates as 'Brentwood'
df=df[df.Neighborhood!='Wilshire Park'] #Removes redundancy as it has the same coordinates as 'Wilshire Center'
df.reset_index(inplace=True,drop=True)
df.head()
```

	Neighborhood
0	Angelino Heights
1	Angeles Mesa
2	Angelus Vista
3	Arleta
4	Arlington Heights

The next step is to use GeoPy Nominatim geolocator to obtain the longitude and latitude coordinates of each neighborhood. Neighborhoods with missing values and obvious geocoding errors will be removed from the list.

```
# define the data frame columns
column_names = ['Neighborhood', 'Latitude', 'Longitude']

# instantiate the data frame
nhoods = pd.DataFrame(columns=column_names)

# use GeoPy Nominatim geolocator with the user_agent "la_explorer".
geolocator = Nominatim(user_agent="la_explorer",timeout=5)
for i in range(0,len(df)):

    address = df.Neighborhood[i]+', Los Angeles'
    location = geolocator.geocode(address)
    if location == None:
        latitude = 0
        longitude = 0
    else:
        latitude = location.latitude
        longitude = location.longitude

    nhoods = nhoods.append({'Neighborhood': df.Neighborhood[i],
                                              'Latitude': latitude,
                                              'Longitude': longitude}, ignore_index=True)
print("The number of neighborhood before clean up is:",  len(nhoods))
nhoods.head()
```

```
The number of neighborhood before clean up is: 196
```

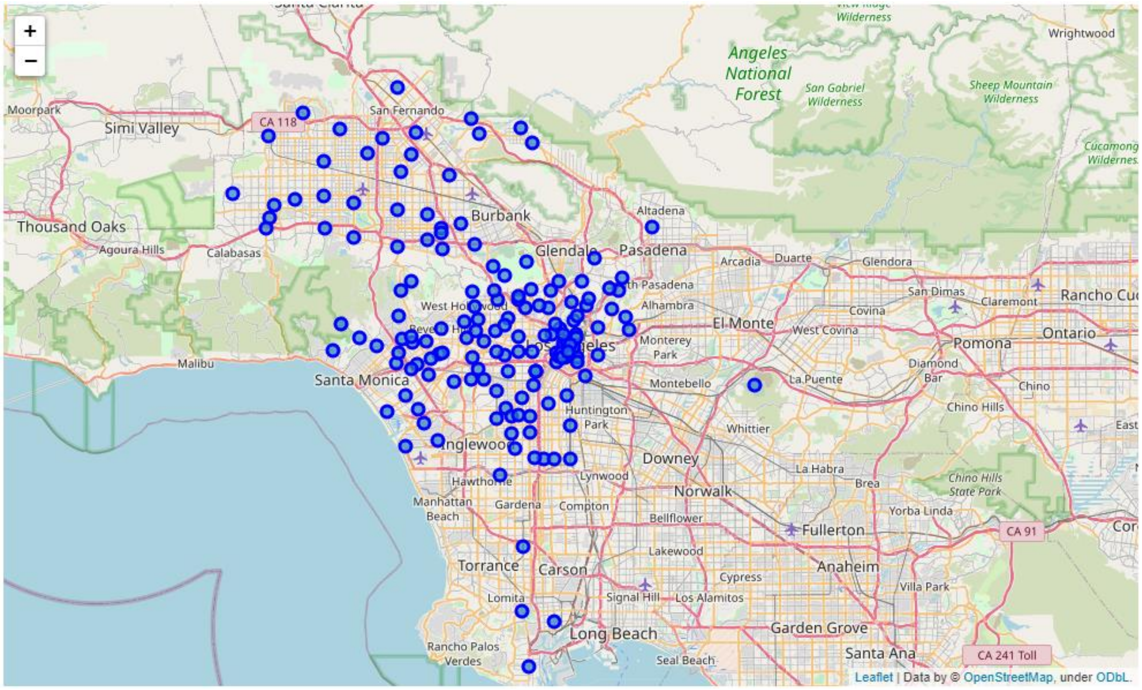
	Neighborhood	Latitude	Longitude
0	Angelino Heights	34.070289	-118.254796
1	Angeles Mesa	33.991402	-118.319520
2	Angelus Vista	-23.403598	-51.965818
3	Arleta	34.241327	-118.432205
4	Arlington Heights	34.043494	-118.321374

```
# clean neighbourhood data by deleting missing values and obvious geocoding errors
nhoods['Latitude']=nhoods['Latitude'].astype(float)
nhoods['Longitude']=nhoods['Longitude'].astype(float)

nhoods=nhoods[(nhoods.Latitude>33.5) & (nhoods.Latitude<34.4) & (nhoods.Longitude<-118)]
nhoods.reset_index(inplace=True,drop=True)
nhoods.head()
print("The number of neighborhood after clean up is:",  len(nhoods))
nhoods
```

```
The number of neighborhood after clean up is: 161
```

4.2.2 Plot LA Neighborhood Map

Geopy library is used to get the latitude and longitude values of Los Angeles. Then a map of Los Angeles is created with the neighborhood superimposed on top.

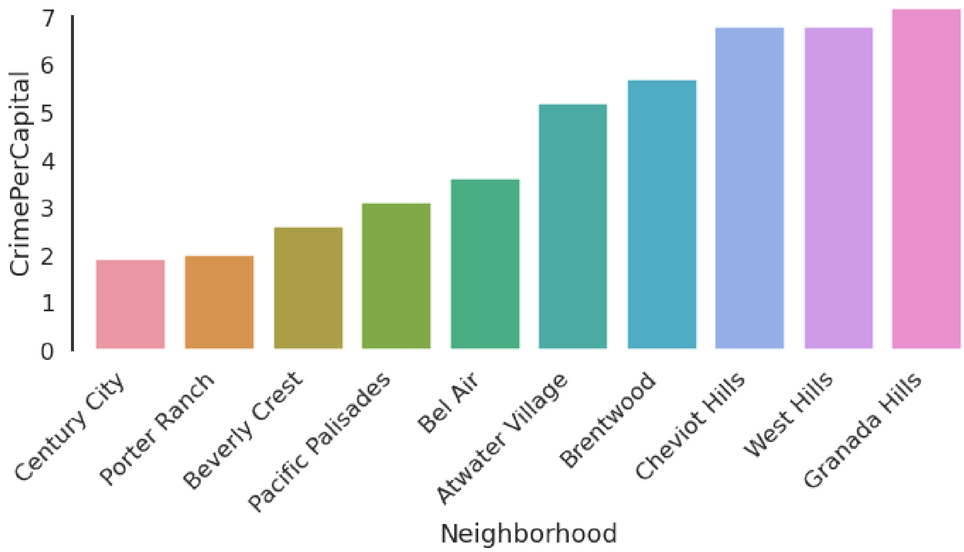


4.3 Crime Data

To analyze the safety of the community, we will use the violent crime rate per 10,000 people from December 30, 2019 to June 28, 2020. Violent crime is defined as homicide, rape, aggravated assault and robbery. The chart below contains both per capita statistics and gross crime counts.

	Neighborhood	CrimePerCapita	CrimeCounts
0	Chesterfield Square	126.9	81
1	Vermont Vista	122.9	306
2	Vermont Knolls	110.4	238
3	Harvard Park	109.3	119
4	Broadway-Manchester	105.4	272

Let's identify and visualize the top 10 neighborhoods with the lowest rates of violent crime per 10,000 people.

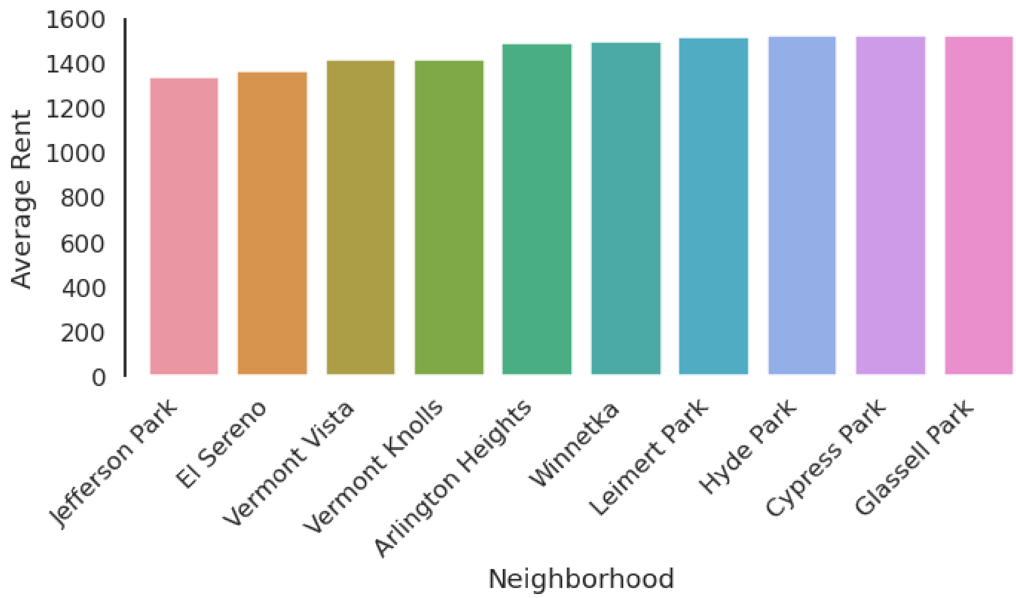


4.4 Rent Data

Rent is also a big factor in deciding where to live, therefore we collected the average monthly rent in different areas of Los Angeles and listed below.

	Neighborhood	Average Rent
0	Adams - Normandie	3595
1	Arleta	1646
2	Arlington Heights	1490
3	Atwater Village	1994
4	Baldwin Hills	2200

Let's identify and visualize the top 10 neighborhoods with the lowest average rent in Los Angeles.



4.5 Venues Data (Foursquare API)

Foursquare API was used to provide information about venues and geolocation.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arleta	34.241327	-118.432205	Back To The Future Filming Location - McFly's ...	34.243429	-118.433655	Historic Site
1	Arleta	34.241327	-118.432205	Edwards Cinema	34.241197	-118.430284	Movie Theater
2	Arleta	34.241327	-118.432205	Canterbury & Kelowna	34.239525	-118.435370	Movie Theater
3	Arlington Heights	34.043494	-118.321374	Underground Museum	34.039758	-118.322934	Art Gallery
4	Arlington Heights	34.043494	-118.321374	Cafe Dabang	34.047407	-118.319082	Café

There are 275 unique categories. For this project, we pay more attention to the general venue category, therefore Foursquare API was leveraged to return a list of general categories and appended to the original venues data.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	General Venue Category
0	Arleta	34.241327	-118.432205	Back To The Future Filming Location - McFly's ...	34.243429	-118.433655	Historic Site	Arts & Entertainment
1	Arleta	34.241327	-118.432205	Edwards Cinema	34.241197	-118.430284	Movie Theater	Arts & Entertainment
2	Arleta	34.241327	-118.432205	Canterbury & Kelowna	34.239525	-118.435370	Movie Theater	Arts & Entertainment
3	Arlington Heights	34.043494	-118.321374	Underground Museum	34.039758	-118.322934	Art Gallery	Arts & Entertainment
4	Arlington Heights	34.043494	-118.321374	Cafe Dabang	34.047407	-118.319082	Café	Food
5	Arlington Heights	34.043494	-118.321374	La Cevicheria	34.047654	-118.322810	Latin American Restaurant	Food
6	Arlington Heights	34.043494	-118.321374	Natraliart Jamaican Restaurant	34.039750	-118.322392	Restaurant	Food
7	Arlington Heights	34.043494	-118.321374	7-Eleven	34.044352	-118.326642	Convenience Store	Shop & Service
8	Arlington Heights	34.043494	-118.321374	Enterprise Rent-A-Car	34.046795	-118.318267	Rental Car Location	Travel & Transport
9	Arlington Heights	34.043494	-118.321374	Winchell's	34.043435	-118.323944	Donut Shop	Food

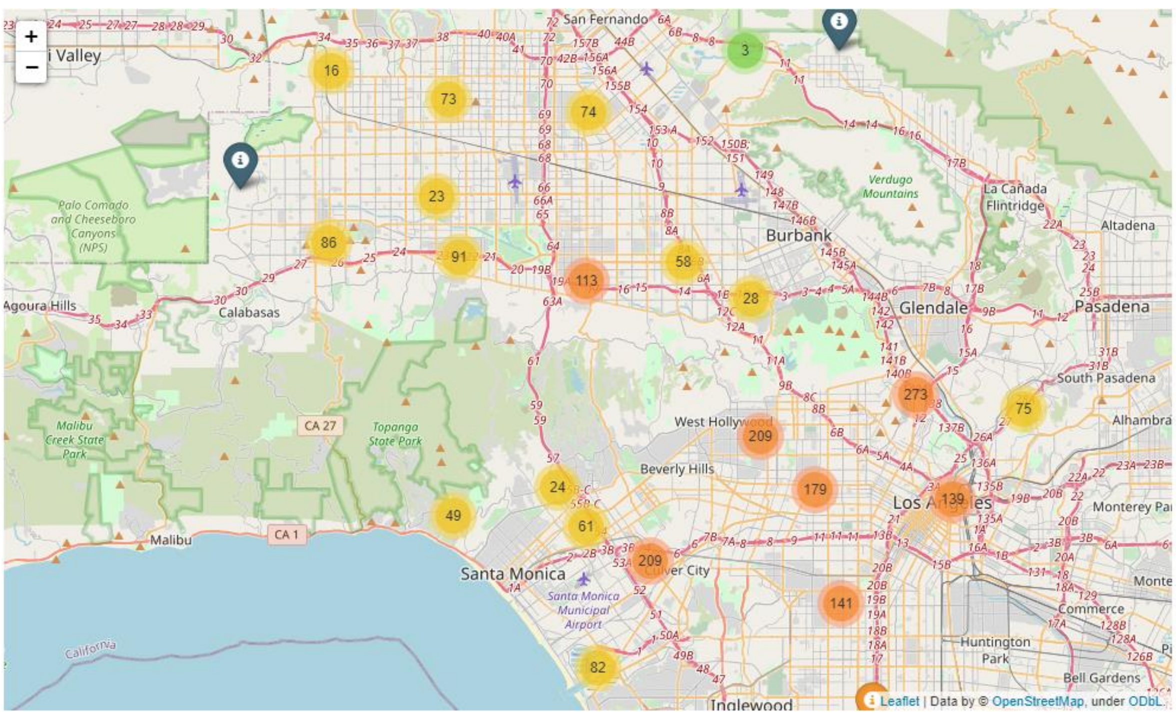
General Venue Categories include the following

- Arts & Entertainment
- Food
- Shop & Service
- Travel & Transport
- Outdoors & Recreation
- Nightlife Spot
- Professional & Other Places
- College & University
- Residence

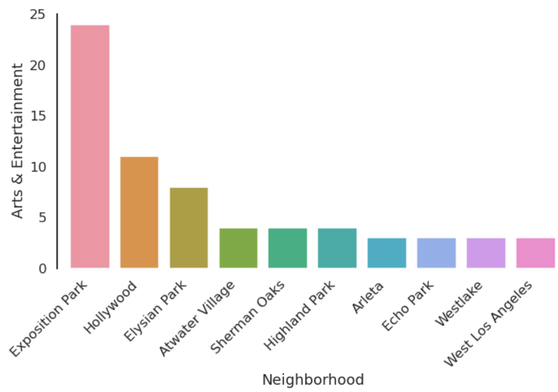
Since in this project, we only take Arts & Entertainment, Food, Shop & Service, Outdoors & Recreation and Nightlife Spot into consideration, so other general categories were dropped.

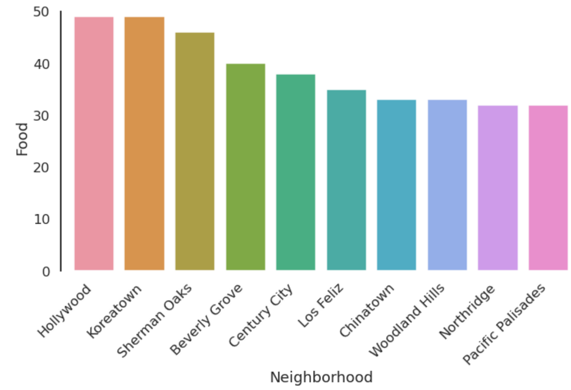
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	General Venue Category
0	Arleta	34.241327	-118.432205	Back To The Future Filming Location - McFly's ...	34.243429	-118.433655	Historic Site	Arts & Entertainment
1	Arleta	34.241327	-118.432205	Edwards Cinema	34.241197	-118.430284	Movie Theater	Arts & Entertainment
2	Arleta	34.241327	-118.432205	Canterbury & Kelowna	34.239525	-118.435370	Movie Theater	Arts & Entertainment
3	Arlington Heights	34.043494	-118.321374	Underground Museum	34.039758	-118.322934	Art Gallery	Arts & Entertainment
4	Arlington Heights	34.043494	-118.321374	Cafe Dabang	34.047407	-118.319082	Café	Food
5	Arlington Heights	34.043494	-118.321374	La Cevicheria	34.047654	-118.322810	Latin American Restaurant	Food
6	Arlington Heights	34.043494	-118.321374	Natraliart Jamaican Restaurant	34.039750	-118.322392	Restaurant	Food
7	Arlington Heights	34.043494	-118.321374	7-Eleven	34.044352	-118.326642	Convenience Store	Shop & Service
9	Arlington Heights	34.043494	-118.321374	Winchell's	34.043435	-118.323944	Donut Shop	Food
10	Arlington Heights	34.043494	-118.321374	Restaurant World	34.040283	-118.323125	Shop & Service	Shop & Service

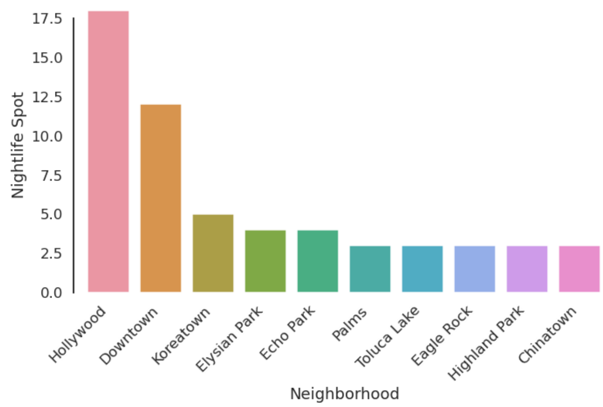
Let's visualize general venue category on a map

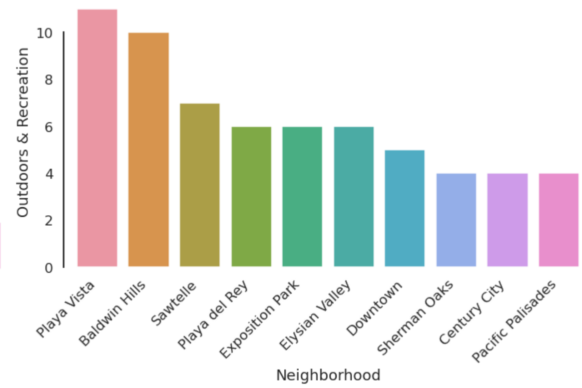


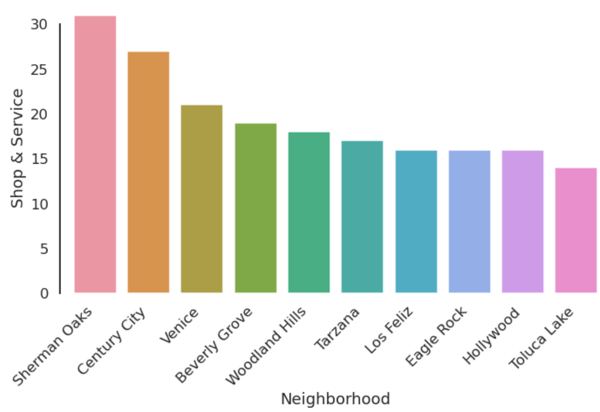
Then let's review the top 10 neighborhoods in each general category.











4.6 Cluster Analysis

We merged venues data frame with the previous crime data and rent data to generate a new one.

	Neighborhood	Latitude	Longitude	CrimePerCapita	CrimeCounts	Average Rent	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service
0	Arleta	34.241327	-118.432205	14.4	47	1646	2	0	0	0	0
1	Arlington Heights	34.043494	-118.321374	28.7	67	1490	3	4	0	0	5
2	Atwater Village	34.116398	-118.256464	5.2	8	1994	4	16	2	2	13
3	Baldwin Hills	34.017616	-118.381694	41.1	132	2200	1	1	0	10	2
4	Bel Air	34.082728	-118.447980	3.6	3	2838	0	2	2	1	1

In addition, we created a new data frame to display the top 1 venue for each neighborhood.

	Neighborhood	1st Most Common Venue
0	Arleta	Arts & Entertainment
1	Arlington Heights	Shop & Service
2	Atwater Village	Food
3	Baldwin Hills	Outdoors & Recreation
4	Bel Air	Nightlife Spot

We would pick up CrimePerCapita, Average Rent and number of each general venue category as input features for K-Means clustering algorithm. But before running the algorithm, don't forget to normalize the dataset. Normalization is a statistical method that helps mathematical-based algorithms to interpret features with different magnitudes and distributions equally. We use StandardScaler() to normalize our dataset.

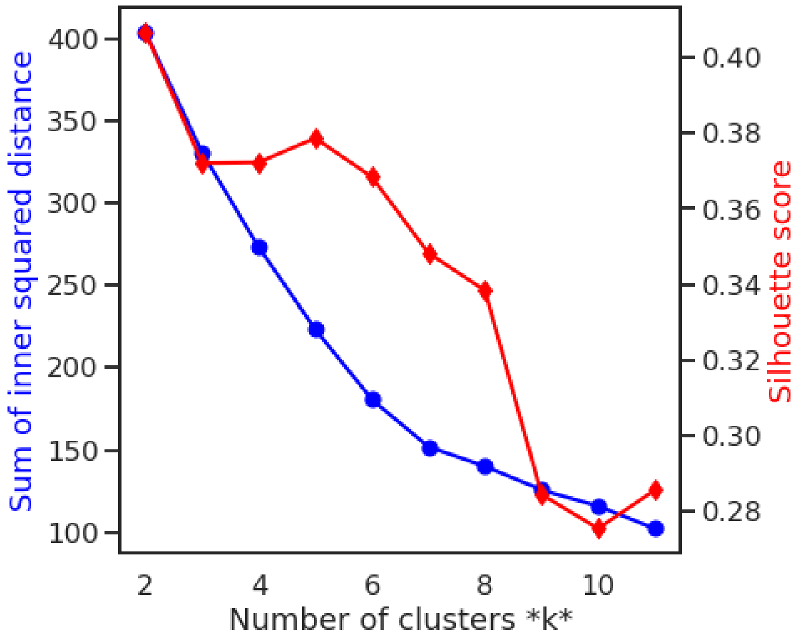
```
la_cluster = la_df.drop(columns=["Neighborhood", "Latitude", "Longitude","CrimeCounts"])

from sklearn.preprocessing import StandardScaler
X = la_cluster.values[:,1:]
X = np.nan_to_num(X)
Clus_dataSet = StandardScaler().fit_transform(X)
```

Another important point for cluster analysis is to determine the optimal value of K. We will use Silhouette Score and Sum of Squared Distance to help us decide.

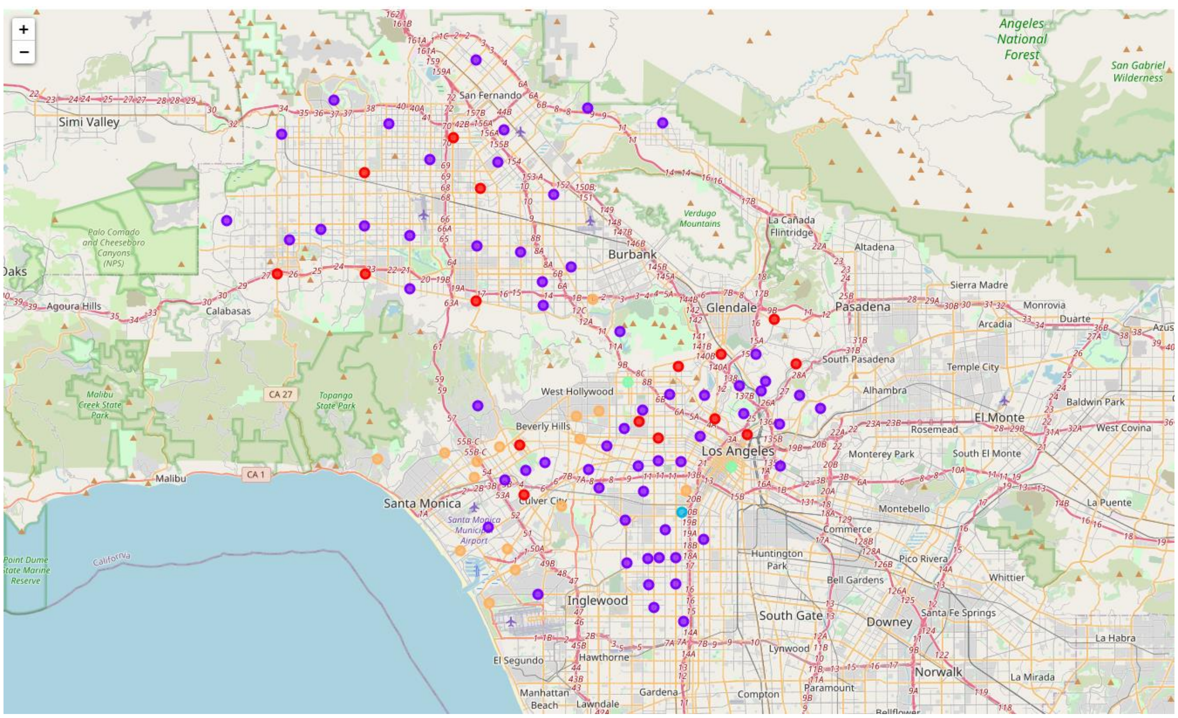
Sum of Squared Distance measures error between data points and their assigned clusters' centroids. The smaller the better.

Silhouette Score focuses on minimizing the sum of squared distance inside the cluster as well, meanwhile, it also tries to maximize the distance between its neighborhoods. A higher Silhouette Score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

From the figure below, we can see that when K=2, the Silhouette Score is the highest, but the Sum of Squared Distance is also very high. We would choose K=5, since this number balances the Silhouette Score and the Sum of Squared Distance.



With all data prepared, we ran K-Means clustering to group the similar neighborhoods into 5 clusters. Let's visualize clustering results with a different color in the map view.



Centroid values can be easily checked by averaging the features in each cluster.

Cluster Labels	CrimePerCapita	Average Rent	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service
0	14.79375	2001.000000	1.750000	29.750000	2.125000	1.937500	14.437500
1	31.14500	1915.566667	0.616667	5.333333	0.316667	0.883333	2.900000
2	47.50000	3562.000000	25.000000	11.000000	1.000000	7.000000	2.000000
3	44.95000	2437.500000	6.500000	36.000000	15.000000	3.500000	11.000000
4	23.94000	3004.000000	1.400000	13.866667	0.533333	3.600000	8.666667

5. Results

K-Means partition neighborhoods into 5 mutually exclusive clusters. The results of clustering are shown below.

- Cluster 1

	Neighborhood	CrimePerCapita	Average Rent	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service	1st Most Common Venue
2	Atwater Village	5.2	1994	4	16	2	2	13	Food
11	Century City	1.9	2593	3	38	0	4	29	Food
15	Chinatown	21.8	2387	2	33	3	2	3	Food
19	Eagle Rock	7.3	1918	2	29	3	1	15	Food
21	Echo Park	19.8	2079	3	26	4	4	6	Food
34	Highland Park	13.1	1808	4	28	4	3	11	Food
39	Koreatown	28.2	1894	2	49	5	1	5	Food
45	Los Feliz	11.4	2014	1	37	3	0	21	Food
50	Mission Hills	14.4	1608	0	18	0	1	14	Food
55	Northridge	15.8	1857	0	32	2	0	11	Food
58	Palms	11.2	2277	0	21	3	2	9	Food
59	Panorama City	23.7	1587	0	14	0	2	13	Food
67	Sherman Oaks	8.7	2018	5	47	2	4	30	Food
74	Tarzana	21.4	1745	2	30	2	3	20	Food
91	Windsor Square	16.1	1951	0	25	0	1	13	Food
93	Woodland Hills	16.7	2286	0	33	1	1	18	Food

- Cluster 2

	Neighborhood	CrimePerCapita	Average Rent	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service	1st Most Common Venue
0	Arleta	14.4	1646	2	0	0	0	0	Arts & Entertainment
1	Arlington Heights	28.7	1490	3	4	0	0	5	Shop & Service
4	Bel Air	3.6	2838	0	2	2	1	1	Nightlife Spot
6	Beverlywood	10.9	2269	0	0	0	1	1	Shop & Service
7	Boyle Heights	28.0	1959	1	15	1	0	10	Food
9	Canoga Park	21.3	1941	0	9	1	1	8	Food
12	Chatsworth	13.2	1901	1	10	0	1	4	Food
13	Chesterfield Square	126.9	1796	0	3	0	0	1	Food
14	Cheviot Hills	6.8	2313	0	0	0	1	0	Outdoors & Recreation
16	Cypress Park	25.8	1524	0	4	0	1	2	Food
20	East Hollywood	29.3	2009	0	16	0	0	3	Food
22	El Sereno	12.6	1368	0	6	0	1	2	Food
23	Elysian Park	15.0	2343	7	3	4	3	2	Arts & Entertainment
24	Elysian Valley	9.0	1971	1	1	0	6	0	Outdoors & Recreation
25	Encino	8.5	1964	0	13	0	3	8	Food
28	Glassell Park	17.3	1524	0	7	0	3	5	Food
29	Gramercy Park	85.0	1796	0	0	0	1	3	Shop & Service
30	Granada Hills	7.2	1909	2	11	1	2	4	Food
31	Hancock Park	21.6	2352	0	1	0	0	0	Food
32	Harvard Heights	36.2	1558	0	9	0	0	5	Food
33	Harvard Park	109.3	1796	0	0	0	1	0	Outdoors & Recreation
36	Hollywood Hills	14.8	2167	0	0	0	2	0	Outdoors & Recreation
37	Hyde Park	47.6	1523	0	4	0	0	1	Food
38	Jefferson Park	35.0	1338	0	4	0	2	0	Food
40	Lake Balboa	17.9	1785	0	9	0	0	4	Food
41	Lake View Terrace	10.2	1797	0	0	0	2	0	Outdoors & Recreation
42	Larchmont	23.9	2046	4	4	0	2	1	Food
43	Leimert Park	69.9	1515	2	15	0	2	8	Food
44	Lincoln Heights	21.3	2251	1	10	0	0	4	Food
46	Manchester Square	83.4	1796	1	4	1	1	0	Food

47	Mar Vista	10.1	2603	0	9	0	1	8	Food
48	Mid-City	27.8	2236	2	2	0	1	2	Shop & Service
49	Mid-Wilshire	24.4	2457	3	16	2	0	7	Food
51	Montecito Heights	9.9	1704	0	0	0	1	0	Outdoors & Recreation
52	Mount Washington	10.3	1682	0	0	0	1	0	Outdoors & Recreation
53	North Hills	15.7	1620	0	0	0	3	0	Outdoors & Recreation
54	North Hollywood	20.4	1962	0	9	2	0	5	Food
57	Pacoima	18.9	1636	0	6	0	0	0	Food
60	Pico-Union	32.2	2758	0	13	0	0	3	Food
63	Porter Ranch	2.0	2003	0	0	0	1	1	Shop & Service
64	Rancho Park	21.9	2313	0	2	2	0	0	Nightlife Spot
65	Reseda	17.6	1686	1	12	1	1	6	Food
68	Silver Lake	11.2	2013	1	7	0	3	2	Food
69	South Park	59.7	1916	0	0	0	1	2	Shop & Service
70	Studio City	13.7	2245	0	6	0	0	2	Food
71	Sunland	14.6	1574	0	0	0	1	0	Outdoors & Recreation
72	Sun Valley	14.3	1592	0	6	0	0	4	Food
73	Sylmar	14.2	1829	0	5	0	0	0	Food
77	Valley Glen	16.1	1780	0	7	1	0	2	Food
78	Valley Village	11.3	2294	2	6	0	0	7	Shop & Service
79	Van Nuys	24.7	1734	0	15	0	0	11	Food
81	Vermont Knolls	110.4	1417	0	1	0	0	2	Shop & Service
82	Vermont-Slauson	99.7	1642	0	3	0	0	3	Shop & Service
83	Vermont Square	69.6	1916	1	1	0	1	3	Shop & Service
84	Vermont Vista	122.9	1417	0	1	0	0	1	Shop & Service
85	West Adams	41.1	2503	0	3	1	0	2	Food
86	Westchester	17.9	2531	0	8	0	0	12	Shop & Service
87	West Hills	6.8	1808	0	0	0	1	0	Outdoors & Recreation
88	Westlake	40.8	2080	2	8	0	0	5	Food
92	Winnetka	13.9	1498	0	10	0	0	2	Food

- Cluster 3

	Neighborhood	CrimePerCapita	Average Rent	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service	1st Most Common Venue
26	Exposition Park	47.5	3562	25	11	1	7	2	Arts & Entertainment

- Cluster 4

	Neighborhood	CrimePerCapita	Average Rent	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service	1st Most Common Venue
18	Downtown	40.8	2534	2	23	12	5	6	Food
35	Hollywood	49.1	2341	11	49	18	2	16	Food

- Cluster 5

	Neighborhood	CrimePerCapita	Average Rent	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service	1st Most Common Venue
3	Baldwin Hills	41.1	2200	1	1	0	10	2	Outdoors & Recreation
5	Beverly Grove	52.1	3505	3	40	0	4	18	Food
8	Brentwood	5.7	2838	2	7	0	2	9	Shop & Service
10	Carthay	46.9	3321	1	15	0	0	3	Food
17	Del Rey	10.6	3252	0	12	1	0	6	Food
27	Fairfax	63.6	2684	2	15	2	0	10	Food
56	Pacific Palisades	3.1	3625	1	32	0	4	12	Food
61	Playa del Rey	14.1	2484	0	9	0	7	4	Food
62	Playa Vista	13.3	3092	1	9	1	10	8	Outdoors & Recreation
66	Sawtelle	12.9	2593	2	8	0	6	12	Shop & Service
75	Toluca Lake	16.4	2288	0	8	3	4	13	Shop & Service
76	University Park	27.0	3938	2	12	0	2	5	Food
80	Venice	34.2	3386	3	18	1	1	21	Shop & Service
89	West Los Angeles	9.6	2593	3	19	0	3	6	Food
90	Westwood	8.5	3261	0	3	0	1	1	Food

Combining the previous centroid values, we can find that each cluster has its own characteristics.

- **Cluster 1**- has the lowest crime rate and medium rent. This cluster has the greatest number of shop & service among 5 clusters. In addition, number of food venues are significantly higher than other venue types.
- **Cluster 2**- has a moderate crime rate and the lowest average rent. However, the number of various venues is the least among 5 clusters, which is not convenient
- **Cluster 3**- has the highest crime rate and the highest average rent. But it also has the most art & entertainment and outdoors & recreation venues. Obviously, it is a good place for art immersion and outdoor relaxation.
- **Cluster 4**- has the second highest crime rate and average rent. The number of food and nightlife spot venues are the most among 5 clusters. It is a good choice for foodies and nightlife lovers.
- **Cluster 5**- has a moderately low crime rate and the second highest rent. The number of food and shop & service in the cluster is significantly more than other venues.

6. Discussion

Based on the above analysis, in my opinion cluster 1 is a cost-effective choice. Not only because I am most concerned about safety when I personally choose the living place, but also because it has second lowest rent among all clusters. Compared to the lowest rent cluster 2, it has more food and shops, which means living in cluster 1 is more convenient.

Of course, if someone simply wants to pursue the lowest rent and doesn't care about venues in the neighborhood, I would recommend them to choose from cluster 2.

Cluster 3 seems to me a paradise for artists and outdoor enthusiasts, even if it has no advantage in crime rate and rent. There is only one neighborhood - Exposition Park in this cluster. A simple search reveal that this area has Los Angeles Memorial Coliseum, Los Angeles Memorial Coliseum, California Science Center, Lucas Museum of Narrative Art, Exposition Park Rose Garden and so on. It is not difficult to understand why it is so special.

If you are a gourmet or like to hang out at night, I would definitely recommend you choose from cluster 4. This category includes Downtown and Hollywood. Living here can definitely satisfy your appetite for delicious food and yearning for bars.

Cluster 5 seems to be inferior to cluster 1 in all aspects (for example, the rent is not as favorable as cluster 1, but the crime rate is higher than cluster 1), however, it has more outdoor & recreation venues than cluster 1. From the map, we can see that the neighborhoods in cluster 5 are mostly located in areas closer to the beach. If you like to enjoy the romance of the beach, I would recommend you to choose from cluster 5.

7. Conclusion

The objective of this project is to find the most livable area in Los Angeles. By acquiring data from different sources, processing and cleaning them into a data frame containing per capita crime rate, average rent, and number of various venues, we were able to apply K-Mean clustering algorithm and finally get 5 clusters.

We analyzed the characteristics of 5 clusters, and on this basis, recommended a suitable neighborhood cluster for target audiences with different needs. I hope that this analysis will be beneficial for people who have just arrived in Los Angeles or who are considering moving to a new place.