IBM Coursera Data Science Specialization

Capstone project

Customizing online real estate search engine results according to a personal profile

M. Laver – January 2020

1. Introduction

Looking for a home in a new city can be a challenging task. Most of us don't know the ins and outs of a city before we decide to move there, so choosing the right home will usually require an extensive research. This kind of research will likely include looking into the various neighborhoods of a city, getting information about crime rates, demographics, education etc.
Only after completing this research you can go to one of the many real estate search engines and look for the house of your dreams in a neighborhood that fits your needs.

There is plenty of information available online about various neighborhoods in large cities around the world, however, not everyone has the time or knowledge to perform such a comprehensible research. Furthermore, most of the information is based on people's opinions and not on actual data.

What if instead of exploring each neighborhood, gathering information from various sites and datasets, comparing crime rates, commute times, nightlife and many more, a person can go to a real estate search engine, create a profile of himself and get listings in neighborhoods that will fit his needs.

The objective of this report is to demonstrate how using data science can help users get the best real estate listings results without spending a lot of time on research. By filling out a short questionnaire with information such as marital status, number of children, age, hobbies etc., we can find the user the neighborhoods that will fit him best.
In addition to being helpful for people who don’t have the time or ability to research this themselves, it will also be beneficial to developers of online real estate search engines that can use this approach to add an extra feature that will set them apart from the competition.

2. Data

2.1. Data Sources

This kind of search engine can potentially work for any major city.
For the purpose of this demonstration I chose to use the City of Los Angeles, CA. Los Angeles, or L.A. is the most populous city in California, and the second most populous city in the United States after New York City. It is the cultural, financial and commercial center of California, known for its Mediterranean climate, ethnic diversity, the entertainment industry and its sprawling metropolis. Home to more than 10 million people, Los Angeles County is comprised of nearly 300 communities that vary greatly from one another. For someone who is not familiar with the area choosing the right neighborhood can be a very daunting task.

For this project I sourced information about Los Angeles neighborhoods from two sources –

1. The Neighborhood Data for Social Change (NDSC).
   This platform is a project of the USC Price Center for Social Innovation, a free publicly available online resource for civic actors to learn about their neighborhood.
2. Foursquare API.

I chose to approach this problem in two stages. In the first stage I collected information about each neighborhood in L.A. county. First, I gathered the coordinates of each neighborhood, then I added information about the number of people in each neighborhood, the distribution of ages, the average number of people per household, marital status, education status, unemployment rate, crime rates, median rent and average commute time. These features will help me to characterize each community according to the people who live in it.
This information was used to cluster neighborhoods to six clusters. This will allow me to narrow down the search to the cluster that best fits a person's status.
The data for this stage was obtained from the 'The Neighborhood Data for Social Change (NDSC)' portal.

In the second stage, I will focus only on the cluster that was chosen in the previous part. I will use information from the Foursquare API to cluster the neighborhoods in the original cluster, into subcategories. The Foursquare places API allows us to map thousands of venues which will help us find the best neighborhoods according to a person's areas of interest. For instance, if someone enjoys spending time outdoors, I can choose a neighborhood with many parks.

2.1. **Data Preparation (stage I)**

Neighborhood latitude and longitude were obtained from the NDSC portal.

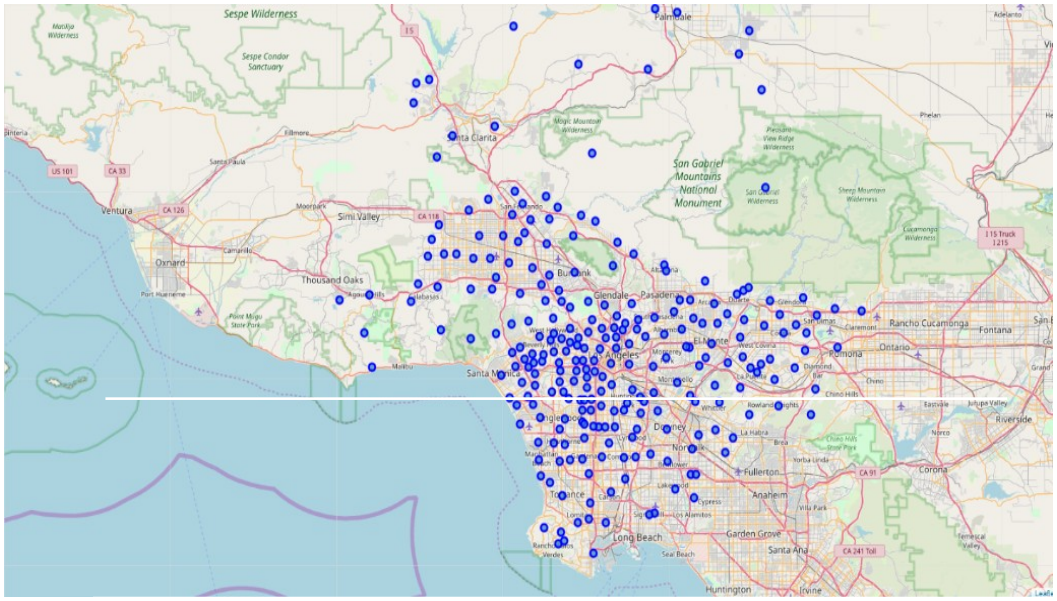

Figure 1 Map of L.A. Neighborhoods

The following variables were also obtained from the NDSC portal for each neighborhood-

1. Distribution of ages (population under 18, population between 18-24, population between 25-34, population between 35-44, population between 45-54, population between 55-64 and population above 65).
2. Total number of populations.
3. Marital Status (Divorced\Separated, Married, Never married and widowed).
4. Average household size.
5. Education (College graduate rate, less than high school and young adults enrolled in school).
6. Unemployment rate.
7. Crime rate (Property crimes per 100 people and Violent crime per 1000 people).

The data was combined into one data frame where each row represents a neighborhood.
Six rows with missing values were removed from the data frame. The final data frame included 220 neighborhoods and 19 features. The first five rows of the data frame can be seen in figure 2 below.

	Neighborhood	Latitude	Longitude	Population Ages 18-24	Population Ages 25-34	Population Ages 35-44	Population Ages 45-54	Population Ages 55-64	Population Ages 65 & Older	Population Under Age 18	Population	Divorced/Separated Population	Married Population	Never-Married Population	Widowed Population	Average Household Size	College Graduation Rate	Less than High School	Young Adults Enrolled in School	Unemployment Rate	Property Crimes Count Per 1000 People	Violent Crimes Count Per 1000 People
0	Acton	34.497355	-118.169810	11.169902	8.469121	10.291287	18.860127	20.441956	14.097019	16.670589	3885.5	8.500000	56.0	30.500000	5.000000	2.855	24.760050	10.885392	42.990420	11.454821	12.491046	1.647143
1	Adams-Normandie	34.031461	-118.300208	18.061678	15.167137	12.517903	12.066921	11.286173	8.616246	22.283943	3506.4	11.600000	34.2	54.400000	4.000000	3.250	17.561293	38.014641	54.191325	13.797029	20.134766	5.912570
2	Agoura Hills	34.146736	-118.759885	7.354860	7.939531	13.327865	19.588894	15.061215	13.090613	23.637022	6286.0	10.666667	57.0	30.666667	3.333333	2.820	54.501554	4.713622	48.622389	4.169422	13.465794	0.619836
3	Agua Dulce	34.504927	-118.317104	7.206238	10.943802	6.802904	18.822264	21.806938	16.617371	17.800484	3719.0	9.000000	68.0	21.000000	4.000000	2.960	26.909842	5.918789	46.416382	4.942166	5.646679	0.537779
4	Alondra Park	33.889617	-118.335156	7.542236	13.515688	13.334674	15.567176	10.496794	18.986324	20.555109	4972.0	14.000000	44.0	37.000000	8.000000	3.120	34.261917	9.276277	38.104839	9.914321	23.129526	3.620274

Figure 2 A data frame representing each neighborhood and its features

2.2. **Data Preparation (stage II)**

100 of the most popular venues within a 500 m radius of the center of each neighborhood were gathered using the foursquare API. The 10 most common venues in each neighborhood were collected in a data frame.

	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 Most Common Venue
0	Agoura Hills	Fast Food Restaurant	Sushi Restaurant	Chinese Restaurant	Breakfast Spot	American Restaurant	Hotel	Lounge	Mexican Restaurant	Multiplex	Restaurant
1	Arlington Heights	Donut Shop	Korean Restaurant	Mexican Restaurant	Vegetarian / Vegan Restaurant	Karaoke Bar	Latin American Restaurant	Sushi Restaurant	Nightclub	Dance Studio	Music Venue
2	Bellflower	Sandwich Place	Pizza Place	Fast Food Restaurant	Video Game Store	Mexican Restaurant	Chinese Restaurant	Clothing Store	Burger Joint	Food Truck	Food
3	Beverlywood	Boutique	Food Truck	Food	Hotel	Park	Yoga Studio	Farmers Market	French Restaurant	Filipino Restaurant	Fast Food Restaurant
4	Canoga Park	Mexican Restaurant	Sports Bar	Restaurant	Sushi Restaurant	Liquor Store	Ice Cream Shop	Donburi Restaurant	Dog Run	Donut Shop	Dumpling Restaurant

Figure 3 Most Common Venues

3. Methodology

3.1. **Clustering neighborhoods by demographics**

In the first stage all neighborhoods were clustered into 6 groups according to the demographic characteristics of the neighborhood. The number of clusters was chosen using the "elbow method". The clustering algorithm that was used is k-means.
Box plots were used to visualize the differences between clusters.

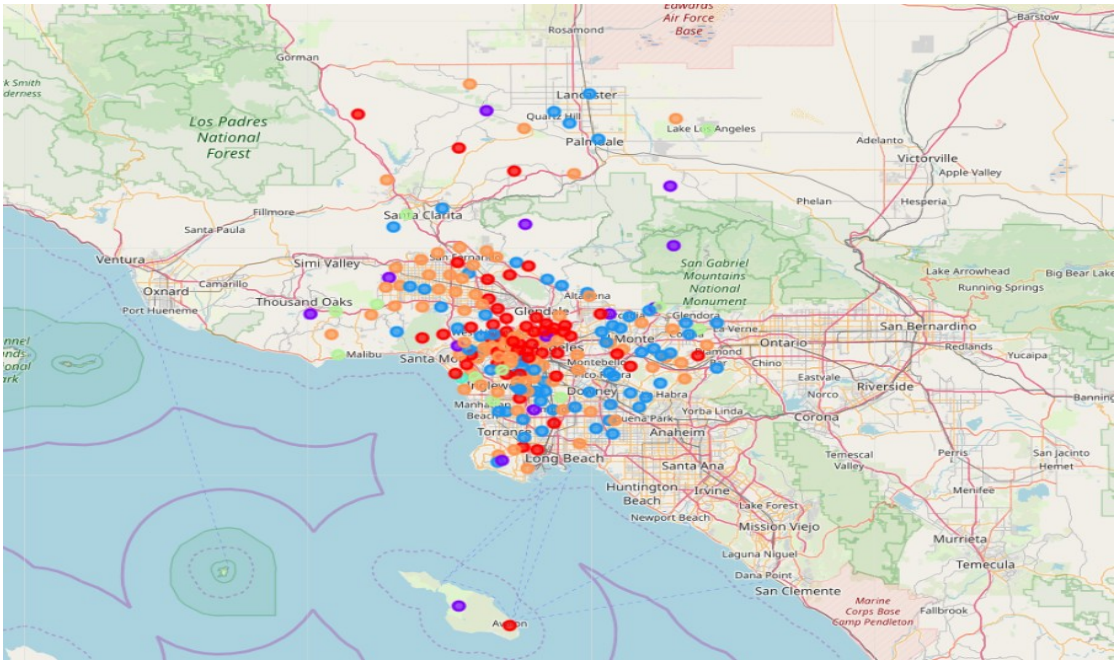

Figure 4 L.A neighborhoods divided into clusters

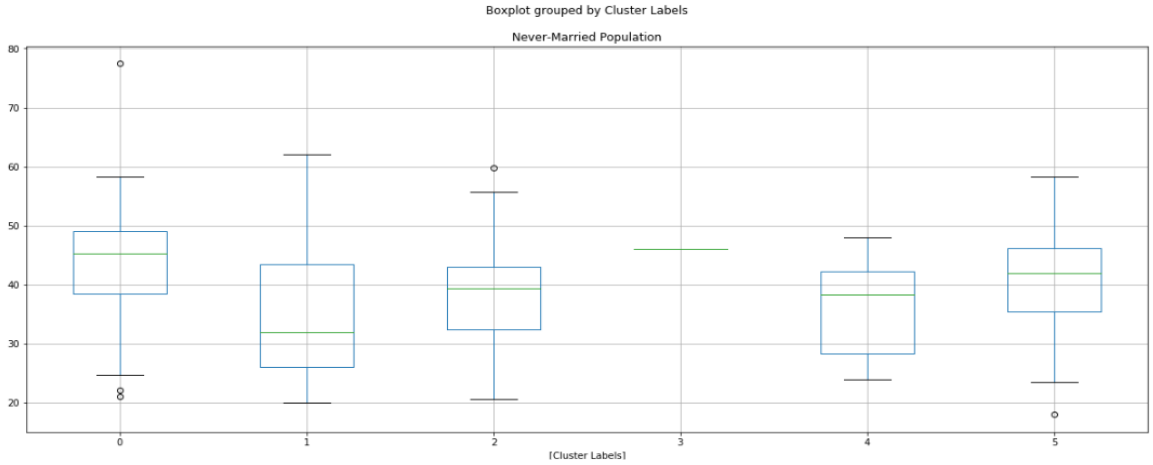

Figure 5 Never Married Population by Cluster

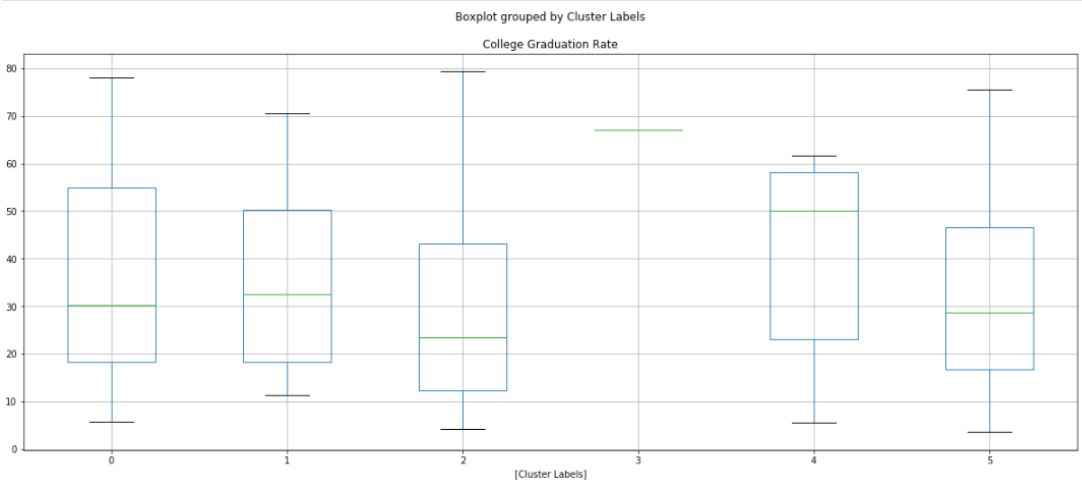

Figure 6 College Graduate Rate by Cluster

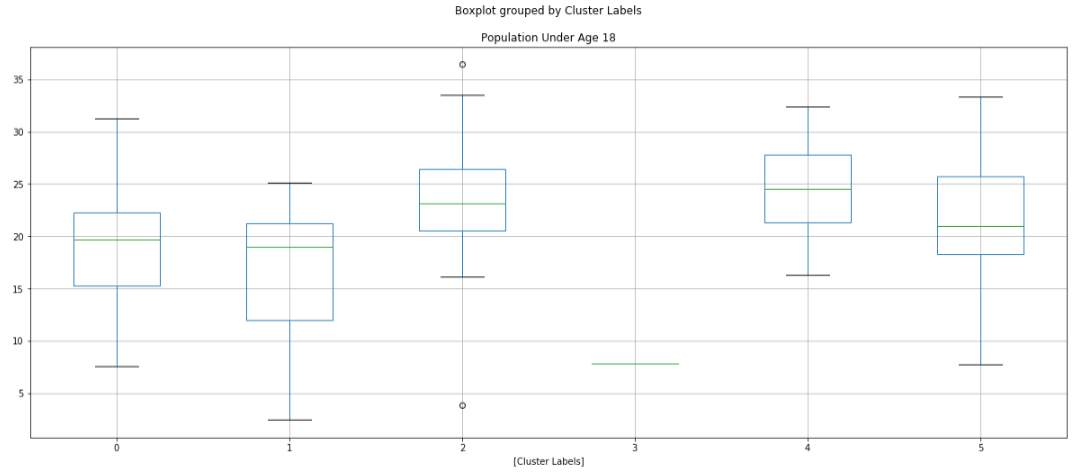

Figure 7 Population Under 18 by Cluster

Figure 8 below shows the mean values of each feature per cluster. Using this table, we can match a user's personal status to the appropriate cluster by finding the cluster were the majority of the population is similar to him.

	0	1	2	3	4	5
Latitude	34.079499	34.127273	34.065411	33.975324	34.129191	34.098553
Longitude	-118.303464	-118.291647	-118.206863	-118.447869	-118.340102	-118.314324
Population Ages 18-24	9.807841	8.321854	9.721387	7.963607	8.294216	10.185861
Population Ages 25-34	17.276146	10.794044	13.695357	24.493969	12.275958	15.310315
Population Ages 35-44	14.375307	11.559610	13.229671	19.883459	13.420844	13.639576
Population Ages 45-54	13.896399	17.492178	14.456244	11.827847	15.538783	14.239887
Population Ages 55-64	12.197921	16.838039	11.954032	15.508076	13.195457	11.844661
Population Ages 65 & Older	13.297869	18.355859	13.371896	12.471888	12.921704	12.937880
Population Under Age 18	19.148517	16.638415	23.571412	7.851155	24.353037	21.841818
Population	3454.588881	1473.192308	5028.554544	9782.000000	6251.111570	4133.566198
Divorced/Separated Population	11.053799	12.365385	10.614980	23.000000	11.743802	10.668872
Married Population	42.945970	49.846154	47.376909	31.000000	48.727273	45.914683
Never-Married Population	43.610323	34.711538	38.939098	46.000000	36.545455	40.855211
Widowed Population	4.965751	5.769231	5.386212	4.000000	4.993113	4.921870
Average Household Size	2.805217	2.687115	3.345190	1.750000	3.089036	3.130108
College Graduation Rate	35.701531	35.090902	29.402887	67.062965	39.261013	32.336082
Less than High School	20.493533	13.793076	21.381896	1.662016	13.625856	21.816550
Young Adults Enrolled in School	46.075832	45.101715	47.986008	53.046597	50.957504	47.884732
Unemployment Rate	8.012590	8.715969	7.668723	5.162283	7.871643	7.658427
Property Crimes Count Per 1000 People	27.717798	74.638756	19.496242	40.686977	16.258457	25.072684
Violent Crimes Count Per 1000 People	6.383143	8.512085	4.681414	4.498058	2.545445	5.581890
Cluster Labels	0.000000	1.000000	2.000000	3.000000	4.000000	5.000000

Figure 8 Mean Values of each Feature per Cluster

3.2 **Clustering neighborhoods by venues**

After selecting the neighborhoods in the clusters that fit the users personal status (age, education, etc.), I used the Foursquare API to get the top 100 venues in each category. I used the same clustering algorithm as before to cluster neighborhoods according to the top venues' categories.

Figure 9 L.A Neighborhoods clustered by Venues

In Figure 10 we can see the top five venue categories in each cluster.

```
----0----
                 venue  freq
0                 Café   1.0
1  American Restaurant   0.0
2           Restaurant   0.0
3         Noodle House   0.0
4                 Park   0.0


----1----
                 venue     freq
0                 Park  3.05238
1   Mexican Restaurant  2.45952
2    Convenience Store  0.67619
3       Clothing Store  0.50000
4  Fried Chicken Joint  0.47619


----2----
                   venue  freq
0                   Park   7.0
1    American Restaurant   0.0
2            Pizza Place   0.0
3          Noodle House    0.0
4  Performing Arts Venue   0.0


----3----
                   venue     freq
0    Mexican Restaurant  4.46714
1     Convenience Store  3.73124
2           Pizza Place  3.26215
3  Fast Food Restaurant  3.18150
4        Sandwich Place  3.14196


----4----
            venue     freq
0           Trail  3.86667
1  Scenic Lookout  1.40000
2            Park  0.61667
3       Racetrack  0.50000
4          Garden  0.25000
```

Figure 10 Top Five Venue Categories in each Cluster

4. Results and Discussion

Neighborhoods were divided into clusters. In the first stage the clusters were formed using demographic data of each neighborhood. Mean values of each cluster were calculated. This data will allow us to match a user to the cluster that best fits his personal status.

For instance, let's say that our user is a 45 years old divorced woman with three young kids and a college degree. We would like to find her a neighborhood with a high number of children, a large divorced population and a large population of people with a college degree. Looking at figure 8 we can see that clusters 3 and 4 match that description.
If our user was a single man who recently graduated from college, we can see that clusters 0,3 and 5 will fit him.

At this point we can continue to cluster the neighborhoods in the chosen clusters into new clusters according to their venues. Now let's assume that our user also wrote that he enjoys spending time outdoors and hiking. Looking at the top 5 venue categories in each cluster (figure 10) we can see that the top venues in cluster 4 are trails, parks, gardens and scenic lookouts so we can recommend to him the neighborhoods in this cluster.

5. Conclusion

In this report I clustered L.A communities based on their demographic characteristics and the types of venues they have. Looking at each cluster I was able to recommend to a user the neighborhoods that fit his needs. I have shown that using simple methods it is possible to save the effort that goes into researching a new city.
Online real estate search engines can use this method to customize the property lists they show each user.