**1. Introduction**

"L.A. is a constellation of microclimates and microcosms, a library with dozens of special collections," [wrote Meghan Daum](https://www.latimes.com/opinion/op-ed/la-oe-daum-los-angeles-smell-goeltzenleuchter-sant-20140703-column.html). " Lying on the Southern California Basin, Los Angeles is the biggest and one of the most popular cities on the west coast of the United States. It is the entertainment capital of the world, a cultural mecca having numerous world-class museums and a paradise of sunshine. From tourist spots like the Universal Studio to those job opportunities presented in various industry, Los Angeles is the place to be.

LA is a rich city and the food culture here puts all of it on display. Thanks to the unique natural environment, the golden sunshine, and multicultural demographic, LA has become one of the best places for you to eat in the country. No matter it’s middle east hummus, Mexican birria tacos, or Chinese dim sum, LA had anything and everything for you. The culinary possibility here is only beyond your imagination.

With the flourish of the food industry, it might be a good choice to open up a new restaurant in the LA area. If you have thought about it and are interested in it, this project is the right thing for you. This project is the final delivery of the IBM Data Science Professional Certificate, and its objective is to find out the best potential neighborhood to open up a new restaurant in LA. In this article, I will go through the introduction, methodology, analysis, results, and conclusion section by section, and the detailed code and report can be found at the end of this article.

**2. Data Preparation**

The following data will be necessary for this project and the reason for each one will be explained later:

* List of all neighbourhoods in LA and their coordinates —

<https://usc.data.socrata.com/Los-Angeles/Los-Angeles-Neighborhoods/xegr-9bnh>

* Venue information(restaurant and their category, coordinates within each neighborhood) — Foursquare APIs
* Population, population density, population growth, median rent, median household income — <https://www.niche.com/>
* LA crime data — <https://www.areavibes.com/los+angeles-ca/crime/>

**3. Methodology**

**3.1 Data pre-processing**

A dataframe of all the LA neighborhoods and their coodidates is obtained by using Pandas libiary. The data source is listed

A picture containing table

Description automatically generated

The dataframe is cleaned as below:

Table

Description automatically generated

Then a folium map is created with superimposed makers on it. Each maker represents each neighborhood of Los Angeles. It seems like there are a lot of things we get.

Map

Description automatically generated

**3.2 Exploratary Data Analysis**

**3.2.1 EDA of Ktown**

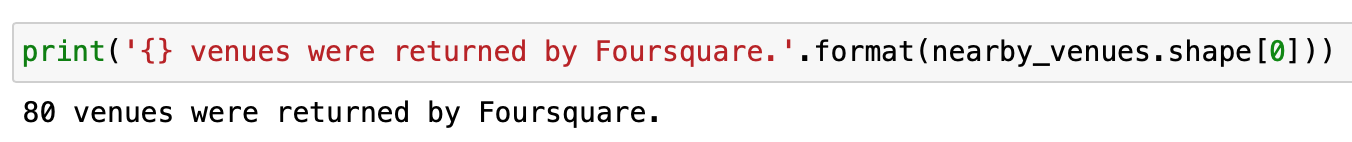
Before we move forward, let’s take some time to explore the Koreatown which is the neighborhood where I currently live in. All the venues information includes the name, category, latitude and longitude is obtained by using the Foursquare API in this case.

First, let’s take a look at what’s really inside Ktown.

Table

Description automatically generated

Then, let’s find out how many venues are within 50o meters radius of Ktown.



Last, let’s find out what kind of venue is the most popular one in Ktown.

Chart, bar chart

Description automatically generated

Unquestionably, korean restaurant is the most popular venue in Ktown.

**3.2.1 EDA of each neighborhood**

As we are searching for the best neighborhood to open up a new restaruant, exploring all the neighborhoods a little bit would be a good choice for us to have a glimpse in mind. Let’s take a look at each neighborhood and all the venues within that specific neighborhood.

Graphical user interface, application

Description automatically generated

Next, let’s check out how many venues are returned by each neighborhood.

Table

Description automatically generated

Then, let’s use one-hot encoding technique to unfold the venue category for each neighborhood, group them by neighborhood and take the mean of each category’s frequency of occurrence.

Table

Description automatically generated

Last, output each neighborhood along with top 5 most popular venues.

A screenshot of text

Description automatically generated

**3.3 Clustering Analysis**

Before we can determine which neighborhood is the one we are looking for, we can use clustering analysis to narrow down our search by clustering them into groups. In our case, we use K-Means clustering technique to do that, and we cluster each neighborhood based on their venue similarities.

The first step in clustering the neighborhood is to determine the optimal value of K for the dataset. This is carried out by using the Elbow method. The elbow figure below show us that the distortion score is lowest when we set K equal to 8. So, as a result, we should group all the neighborhoods into 8 clusters.

Chart, line chart

Description automatically generated

Then, a new dataframe which include the corresponding cluster label and the top 10 venues of each neighborhood is created as below.

A picture containing table

Description automatically generated

Meanwhile, we use another folium map with differently colored makers on it to visulize those neighborhood. Every color in this case represent every cluster. Same color means same cluster, different color means different cluster.

Map

Description automatically generated

Then, several bar charts are generated to show the top 10 venues in each neighborhood, and with food venue highlighted.

*Cluster 0:*

Chart, waterfall chart

Description automatically generated

As we can see, there are only 2 food venues within the top 10 of cluster 0, and only 1 of them is about “Restaurant” which is “Sushi Restaurant”. Meanwhile, the presence of venues that see a high footfall such as home service and grocery store in the list may further indicates that the population density in these neighborhoods are fairly high. All the obervations point to the direction that cluster 0 being nomiated as the cluster to explore further.

*Cluster 1:*

A picture containing chart, text

Description automatically generated

We only have 2 neighborhoods in cluster 1 which are “Irwindale” and “Monterey Park”. From the venue information Foursquare API returned, both of them are not appropriate to open up a restaurant. So, we simply drop cluster 1 right now.

*Cluster 2:*

Chart, waterfall chart

Description automatically generated

As we can see, there are 3 venues which is food-related in cluster 2. They are “Mexican Restaurant”, “Food” and “Food Chicken joint”. But all of them take less than 10% of total which might suggests a not that furious environment in these area. And the presence of “Lake”, “Park” and “Scenic Lookout” on the list also indicates that those neighborhoods are the destinations that people will choose for fun or for vacation and opening up a restaurant in those area may be a good choice. So, we nominate cluster 2 as the cluster to explore further.

*Cluster 3:*

*Chart, bar chart

Description automatically generated*

There are 4 food venues within the top 10 of Cluster 3 which indicates that cluster 3 may not be the proper one for setting up a new restaurant.

C*luster 4:*

Chart, waterfall chart

Description automatically generated

Cluster 4 is even further than cluster 3. There are 7 food venues in cluster 4 with Mexican Reastaurants making up more than 30% of all venues. Which suggests that cluster 4 may not be the one we want to open up a reastaurant in.

C*luster 5:*

Chart, waterfall chart

Description automatically generated

All the neighborhoods in cluster 5 are rural area of Los Angeles, and most of them only have “Construction & Landscaping” venues. Obviously, Cluster 5 is not the one we are looking for.

C*luster 6:*

Chart, waterfall chart

Description automatically generated

Similar to cluster 5, “Park” consisting of nearly all the venues in cluster 6 which indicates that those neighborhood don’t have a high population density, thus not appropriate to open up a restaurant.

C*luster 7:*

Chart, waterfall chart

Description automatically generated

Cluster 7 also got a lot of natural landscapes on list which is similar to cluster 6. But the differences are the presences of “Pharmacy”, “Coffee Shop” and “Garden” in this case are telling us there is population actually living in those neighborhoods of cluster 7. And more importantly, there are no food venues on the top 10 venues list of cluster 7 which suggests a loose environment and even a bright future for opening up a new reastaurants in those neighborhoods. So, we nominate cluster 7 for us to explore further.

**3.4 Candidate Neighborhood**

Based on our discussion, the proper neighborhood for starting a new reastaurant would only come from Cluster 0, Cluster 2 and Cluster 7. So, a list of all the neighborhoods from those clusters are created. Those neighborhoods are our candidates that we want to explore further.

A picture containing table

Description automatically generated

A picture containing table

Description automatically generated

**3.5 Best Neighborhood**

We then start to examine each of our candidate neighborhood one by one. We use U.S. Census Bureau data and data on Niche website to calculate the score for each neighborhood based on the demand for new restaurants and local business cost. Because these are the 2 most imprtant factors which could affect the success of newly opened restaurants.

To analyze the demand, we look at the *population*, *population density*, *population growth* and *median household income*. To access the local business cost, we look at the *crime rate* and *median rent price*.

So, a new dataframe with all the candidates and their corresponding score is generated as below.

**Table

Description automatically generated**

We clean the dataframe, transform the categorical data to numerical, normalize it and sort the neighborhood by its corresponding score. Then, we get the following dataframe:

Table

Description automatically generated

Table

Description automatically generated

At this point, we can easily plot out the top 10 neighborhood for opening up a new restaurant based on their corresponding score.

Chart, bar chart, histogram

Description automatically generated

**4. Results & discussions**

From the figure above we can see that 3 neighborhoods stand out among all others. They are Glendale, Green Valley, and Tujunga which is No.1, No.2, and No.3 on the list respectively. So, if I am planning to open up a new restaurant in the Los Angeles area, those 3 neighborhoods would be my first consideration.

The result of this analysis highlighted the potential neighborhoods where a newly opened restaurant might be favorable. But it solely from the geographical perspective, thus can only be served as a start point in the overall investigation. There are so many other factors and things that should be considered before opening up a new restaurant, such as your customer base, the availability of commercial space, the labor cost, ingredient cost, access to public transportation, and so on.

Besides, the methodology of this analysis could be improved in several ways. First, when we clustered the neighborhoods, we clustered them solely based on their venue information but completely ignored other factors such as the demography, traffic condition, and acreage, which could affect the veracity of the clustering result. Second, when we accessed the score of each neighborhood, we could have to get more industrial or professional data such as payroll costs and growth in labor which may be more useful for our analysis, but those data are kind hard to get on the neighborhood level.

**5. Conclusion**

As a data analyst, I have been always believing in the power of data and convincing myself that there are many real-life problems and scenarios where data can be used to find a solution to it. This project is technically the first project that I’ve ever finished on my own, and it is the final delivery of Coursera’s Applied Data Science Capstone in pursuit of the IBM Data Science Professional Certificate. There are a lot of places that can be improved about this analysis, so if you have any ideas, suggestions, or comments, please let me know.