# Identifying Optimal Location for Chinese Restaurant in Madrid City

August 19th, 2019

## 1. Problem Statement

In 1961 there were only 161 Chinese immigrants in Spain, according to official statistics, and they numbered less than 10,000 in 1995. Nowadays, there are more than 250,000 Chinese immigrants mainly populated in big cities, especially in Madrid. Chinese population has been growing dramatically in Madrid during the past decade and now become the second largest ethnic group. These Chinese immigrants in Madrid have now established themselves in the Spanish market with big purchasing power. Therefore, a new Chinese restaurant or related business entities based on the local population and country of origin will be a good business opportunity.

This project attempted to find the optimal location for a new Chinese restaurant in Madrid city. An analytical approach was used based on data analysis, machine learning algorithms including clustering and finally data visualization tools. Data acquisition and transformation will be performed for the further analysis using machine learning algorithms.

## 2. Methodology

**2.1 Data Acquisition**

The Madrid City Hall website ([www.madrid.es](http://www.madrid.es)) provides valuable data sources in excel format. The data contains the immigration information such as the country of origin, nationality, and total number of population of each group. Foursquare API was accessed through Python environment to obtain the most common venues per neighborhood in Madrid districts, such as the distribution of venues, common places for local residents, and also popular spots for visit.

**2.2 Data Analysis**

As shown in Figure 1, the original data contains lots of information, some of which are unnecessary for our analysis. Therefore, the data was reorganized and cleaned based on the main goal to effectively define our target population, the residential location of major population, and examine the venues and local business entities in the specified area.



Figure 1. Original data acquired from the website.

As the main feature that we need is the country of origin, which indicates where the majority of residents originally came from. By removing unnecessary features and groups of minor population, we can simplify the data significantly. Figure 2 shows the example of the transformed data. As seen from Figure 2, the population of Chinese immigrants are the second largest group in Madrid.



Figure 2. Example of the transformed data.

Foursquare API was accessed through Python environment in order to obtain the locational data as well as the distribution of venues in each neighborhood. Geolocator was used to obtain the latitude and longitude of each neighborhood as shown in Figure 3 below.



Figure 3. Acquired locational coordinates of each neighborhood.

These neighborhoods were then plotted into a map of Madrid city to understand the geographical information of each location as shown in Figure 4.

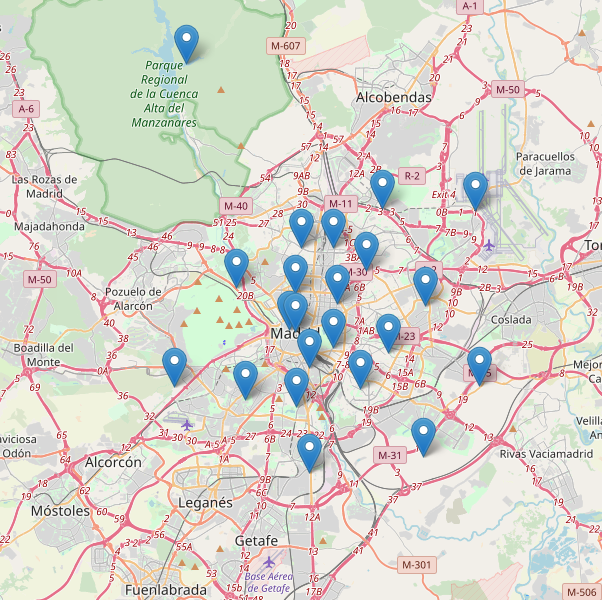


Figure 4. Map of Madrid city with each neighborhood location.

Distribution of nearby venues were obtained with their locational information including latitude and longitude as shown in Figure 5.



Figure 5. Venues of neighborhood with locational information.

Once all the venues were identified, they were categorized based on neighborhood location as shown in Figure 6.





Figure 6. Venue information by neighborhood.

These neighborhoods were then be segmented or divided into a few clusters based on the selected features. In order to find out the ideal K value for K-means clustering, the Elbow method was implemented. This method plots a hypothetical, various number of clusters, among which the rate of decrease slows down dramatically. This point is ideal to be selected as the K value because creating more divisions beyond this point means more difficult to distinguish cluster from each other. As shown in Figure 7, a cluster of 5 was selected based on the sum of squared distances.

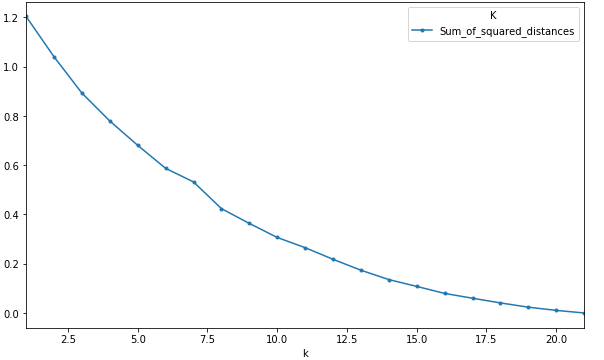


Figure 7. K value selection.

Therefore, based on the number of cluster selection and segmentation of the neighborhoods, the districts were then categorized accordingly as shown in Figure 8.

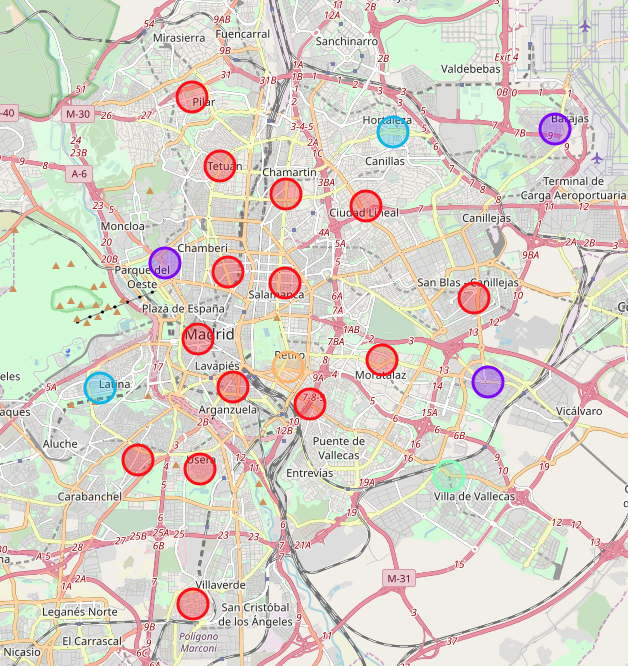


Figure 8. Madrid neighborhood categorized into 5 clusters.

## 3. Results and Discussion

The newly defined five clusters showed diverse population and venue distribution. The first cluster included majority of the neighborhoods with high population from all counties of origin. This cluster included the Chinese immigrants, which showed the most common venues are hotel, restaurant, and Spanish restaurant whereas the Chinese restaurants are the 7th most common venue in this cluster. Therefore, a new Chinese restaurant in this cluster will cover most of the Chinese immigrants. However, high competition among other restaurants are also inevitable according to the venue distribution.

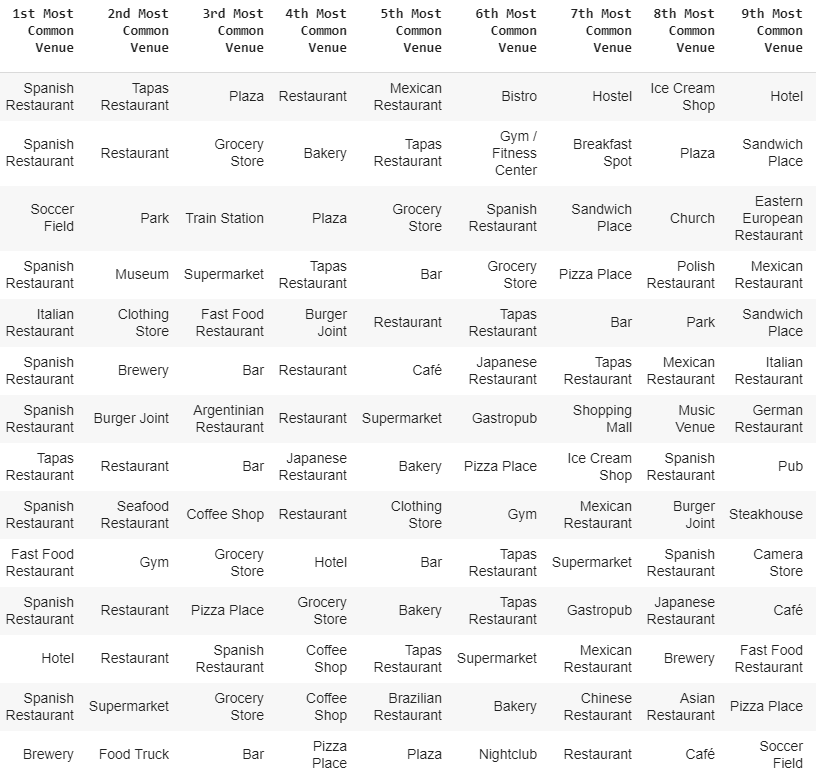
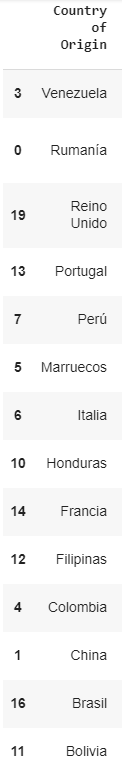


Figure 9. Venue distribution of Cluster 1.

Cluster 2 included three immigrants’ group with most venues in Pizza and Spanish restaurants. No Chinese-related business were identifiable in this cluster. Similarly, cluster 3 and 4 showed most venues from Pizza places, which means very limited room for Chinese-related business.

Interestingly, cluster 4 covers the immigration population of Bulgaria but the most common venue was shown to be Chinese restaurant. Also, the second, fifth, and ninth common venue also showed a close relation with Chinese business entities. This means that although the major population in cluster 3 is not Chinese immigrants, the residents tend to show high interest in Chinese food and drinks (bubble tea), which indicates a high potential for a new business such as a Chinese restaurant in this cluster. Additional study showed that the Usera district in Madrid has a significant growing of immigrant communities of Chinese in recent years. Therefore, Usera district seems to be the ideal location for a new business such as new restaurant for Chinese immigrants.



Figure 10. Venue distribution of Cluster 2.



Figure 11. Venue distribution of Cluster 3.



Figure 12. Venue distribution of Cluster 4.

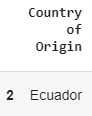


Figure 13. Venue distribution of Cluster 5.

## 4. Conclusions

Cluster 1 shows large number of venues related to Chinese immigrants. However, large number of already-established business entities such as restaurants and markets in this region may lower the possibility of success of new business targeting Chinese immigrants. On the other hand, cluster 4, which is mainly the Usera area in Madrid shows a significant growing of population, especially Chinese. This means a new business that targets this neighborhood will have a much high potential to both satisfy the local residents’ needs as well as succeed from neighborhood competitions.