Coursera Capstone Project

# Weeks 4 & 5

# Korean Restaurants Project

## Dougie Scott

## 06/01/2021

# Contents

Contents

[**Contents 2**](#_Toc61099069)

[**Introduction & Business Problem 3**](#_Toc61099070)

[**The Data and our Approach to a Solution 4**](#_Toc61099071)

[**Suggested Outputs 5**](#_Toc61099072)

[**Methodology 6**](#_Toc61099073)

[**Results 8**](#_Toc61099074)

[Visualisation 8](#_Toc61099075)

[Prevalence of Korean Restaurants in Los Angeles (clustered) 8](#_Toc61099076)

[Prevalence of Non-Restaurant Amenities in Los Angeles (clustered) 9](#_Toc61099077)

[Desirability Scores 9](#_Toc61099078)

[Top 5 Districts based on Desirability Score 9](#_Toc61099079)

[Lowest 5 Districts based on Desirability Score 10](#_Toc61099080)

[Clustering Results 11](#_Toc61099081)

[**Discussion 15**](#_Toc61099082)

[**Conclusion 16**](#_Toc61099083)

# Introduction & Business Problem

We work for a firm of freelance Data Scientists and we have been approached by a firm called “Kor That’s Tasty” (KTC), who operate a chain of Korean Restaurants. They offer a full Korean menu, including all variants of Korean food as well as standard alcoholic and non-alcoholic beverages. They operate both sit-in and delivery services from their restaurants.

For now, KTC are not active in the United States, but they have plans to ambitiously expand into the US. They have identified Los Angeles as the city they would like to start in, given its size and convenience for them in relation to current operations and supply chains.

They have approached our firm of Data Scientists to deliver for them a report on the existing situation with regards restaurants in Los Angeles. They have asked us to return to them with some recommendations on which Districts/Neighbourhoods to open their first restaurant(s), and wish us to keep the following in mind:

* They do not want strong competition from other nearby Korean Restaurants, and would therefore like to set up in an area with few Korean restaurants nearby
* They are not overly concerned if there are other restaurants nearby, provided these are not Korean restaurants
* They want to set up in an area with reasonable population density
* They would like a reasonable level of other non-restaurant amenities nearby, since these can attract footfall into the location of their restaurant.

They would like the deliverable back as a series of maps and other appropriate presentation mechanisms that demonstrate the best locations for them to set up their restaurants. They would like us to segment (cluster) the neighbourhoods based on the best prospects for them, highlighting both good and bad neighbourhoods.

# The Data and our Approach to a Solution

There are some key data that we clearly need in order to begin to work on this problem:

1. **A list of Los Angeles districts, together with their location data**

I have identified a list of Los Angeles districts and zip codes at the following location

<http://www.laalmanac.com/communications/cm02_communities.php>

This will serve as a base for beginning work on Los Angeles. Only those districts listed as belonging to “Los Angeles” will be considered.

1. **The population counts of each Los Angeles district, from census data**

I have identified data from the 2010 Census that can assist us with this. It will be found here:

<http://www.laalmanac.com/population/po24la.php>

We can use this data to determine population densities in each district, and work out how well the population is served by the existing restaurants nearby, as well as whether our client should set up in that district

1. **The Latitude and Longitude of each Zip Code, so that we can input this data to FourSquare**

The data was found here, downloadable as a CSV:

<https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=CA>

1. **Location data of Korean Restaurants throughout Los Angeles**

This is clearly essential, and is data that will be provided by FourSquare, via their API. It is essential that we can gain a measure of neighbourhoods based on how many Korean Restaurants are already present, perhaps in the form of a ratio vs. population size. We will need the Latitude and Longitude of each such restaurant.

1. **Location data on non-restaurant amenities throughout Los Angeles**

This will again be provided by FourSquare. We will need to determine the nearby non-restaurant amenities and find a suitable way to measure this in terms of closeness to the given restaurant. Again, the latitude and longitude will be required, and we must remove restaurant locations from this part of the analysis.

# Suggested Outputs

This is subject to change as I explore the data, but I intend to do the following:

* Obtain all raw data around LA districts, populations, and location data of all amenities throughout the city of Los Angles. This data must be appropriately wrangled.
* Familiarise myself with this data via simple visualisations, and checking its accuracy
* Determine simple measures of neighbourhood fitness such as population density, density of Korean Restaurants, and density of non-restaurant amenities
* Output some simple visualisations for our client based on the measures listed above
* Output a score of “Neighbourhood Suitability” based on the above measures
* Use these “Neighbourhood Suitability” scores to segment the neighbourhoods into ones that are of similar potential, and provide these back to the client

# Methodology

Initially, we cleaned and combined the data from our 3 key sources. We extensively used the pandas library in Python. Specific data sources handles were:

1. [The list of Los Angeles districts](http://www.laalmanac.com/communications/cm02_communities.php)

Main assumptions made here were that those Districts with the specific string “(Los Angeles)” were the ones of interest. In addition, any listed as “(PO Boxes)” were deemed not required, since these were largely duplicates of other entries in the table. In any case, the location of a PO Box is hardly relevant to an investigation around Restaurant locations.

A further major assumption made here was that, because there were sometime multiple zip codes associated with a given District, only the first one listed needed to be used. This seems a reasonable assumption, since the main function of zip code is to yield location data, which in turn is used to get nearby venues for each District.

1. [The population counts of each District](http://www.laalmanac.com/population/po24la.php)

Some cleaning was required in order to make sure that naming of Districts in this table matched with the list of Los Angeles districts. We had issues with, for example, “Bel Air” being listed as “Bel-Air” in the population counts. In any case, this was achieved relatively easily.

1. [Latitude and Longitude of each zip code](https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=CA)

This data was clean and easily joined with our existing data on the field of zip code.

1. FourSquare data on all Amenities for the listed Latitudes/Longitudes of each District

We must assume that FourSquare data is up to date and correct. This data underlies our analysis on Korean Restaurants and Other Amenities within each District. Calls were made to the FourSquare API, and the data appropriately wrangled to provide us with the following critical data:

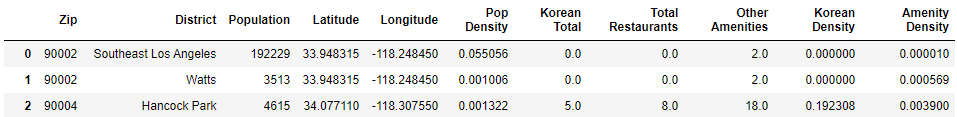
* Total of Korean Restaurants in each District
* Total of non-Restaurant Amenities in each District

Recall that these two attributes are key to determining the suitability of a District, along with Population. We now needed to somehow score each District by creating a function based on the three attributes of **Population**, **Korean Restaurants**, and **Non-Restaurant Amenities**.

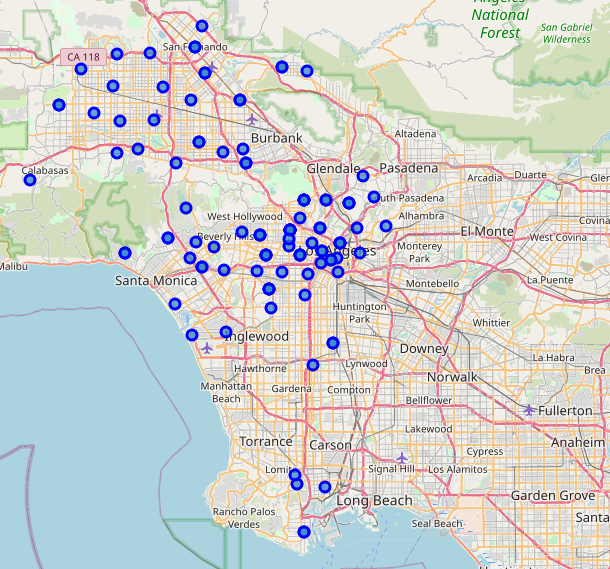
Three additional attributes were created:

1. Population Density – Population of each District divided by Total Population of all Districts
2. Korean Restaurant Density – Total Korean Restaurants dividied by Total Amenities in the District. A core assumption here is that the negative effect of the presence of other Korean Restaurants reduces with increased Amenities in the area
3. Amenity Density – Total Amenities divided by Population of each District. A core assumption here is that more Amenities per head means a busier area of the city, and greater demand for Amenities can be expected.

Here’s a snapshot of how our data looked at this stage:



At this point we decided to visualise our Districts a little better to understand the relevant geography. To begin with, we mapped out our model of Los Angeles using Python’s Folium library, showing the Latitude and Longitude of each District of interest.



Further visualisations showed us some possible Districts that we might avoid, and some that we might think have strong potential. We’ll show these in the Results section. However, we needed a method to evaluate a District based on all of the main criteria: Population, Korean Restaurants, and Other Amenities. For example, a District may have a high Population and no Korean Restaurants, but poor Amenities. We would not necessarily want to recommend such a District.

A method was developed based on the following formula, which calculates a “Desirability Score” for a District based on all three of the main criteria:

Desirability = Population Density \* Amenity Density \* (1 / (1 + Korean Restaurant Density)

This formula sees Desirability increase in direct proportion to Population and Amenities, but inversely to Korean Restaurants. We add 1 to the Korean Restaurant Density to allow for Districts where there are no Korean Restaurants.

The final task was to cluster the Districts into mutually exclusive segments, so that the client could inspect these and prioritise those Districts in the top cluster(s).

Clustering was done using K-means clustering, via the scikit-learn library for Python. K-means clustering is a machine learning algorithm that take an organised set of spatial data points as the input, and then clusters them into groups based on the similarity of data points. Into each cluster will therefore be placed only data points that look very similar.

K-means allows us to determine a number of clusters, and fit the data into those clusters. The machine learning algorithm does this iteratively, essentially defining a centre point (Centroid) for each proposed cluster, then fitting the data points into a cluster based on the closest Centroid. The K-means algorithm will perform this task iteratively, changing the position of the Centroids each time, until the mean difference between each Centroid and its data points is minimised.

We chose to do this in 5 clusters, to give our client organised lists of suitable locations on which they could then conduct further investigations if necessary. Our intention is that the client will prioritise these further investigations based on the cluster and Desirability score.

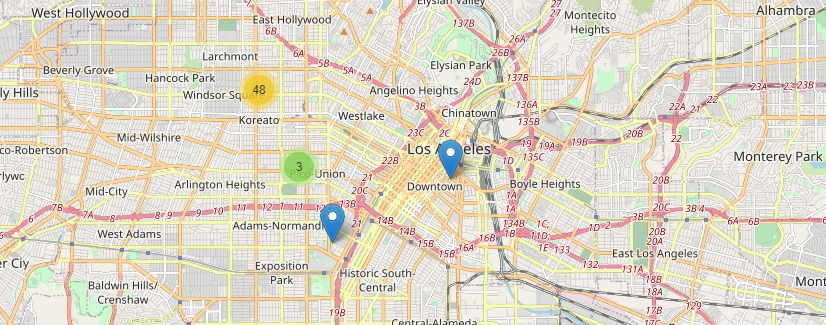
Lastly, we will present the actual clusters in terms of what the cluster represents, and the Districts within each Cluster. This will be provided within the Results section that follows.

# Results

## Visualisation

Let’s first inspect the visualisations, beginning with the locations and prevalence of Korean Restaurants in Los Angeles.

### Prevalence of Korean Restaurants in Los Angeles (clustered)

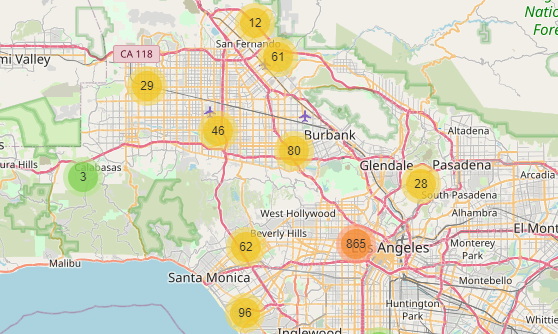


The map above shows the clustering of Korean Restaurants around Los Angeles. It is clear that the west-central part of Los Angeles is extremely well-served by Korean Restaurants, but many of the outlying areas are not served. Also note that downtown Los Angeles is quite underserved too. Based on our client’s requirements, we are unlikely to recommend west-central Los Angeles as an appropriate area to set up in. The competition from other Korean Restaurants is very intense.

We need to remember though that a good Restaurant location must have a good population, and closeness to local Amenities.

We now visualise the Amenities around different Districts within Los Angeles.

### Prevalence of Non-Restaurant Amenities in Los Angeles (clustered)



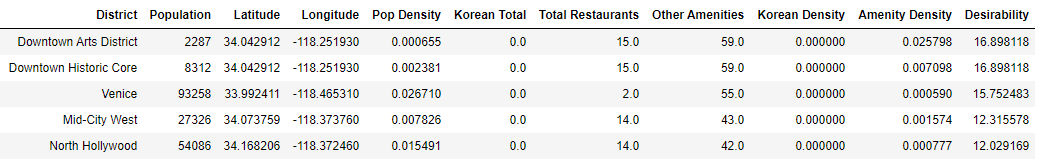
Perhaps unsurprisingly, Amenities are clustered around central Los Angeles. However, we also know that many of the existing Korean Restaurants are clustered in this same area.

Leaving aside the central LA cluster, we note that there are plenty of smaller, but significant, clusters in surrounding Districts. We know that these areas tend not to be served by an existing Korean Restaurant.

## Desirability Scores

Therefore, we can already begin to visualise Districts that have good potential, and those that have poor potential. We mentioned in the Methodology section that we developed a method for calculating a Desirability Score. Let’s take a look at some of the top Desirability scores.

### Top 5 Districts based on Desirability Score

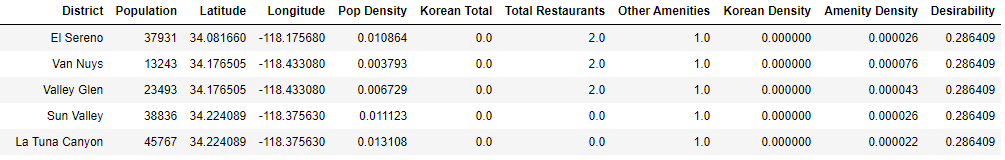


What is immediately noticeable is that the two highest scores are these Downtown locations. Note how they have low population, but extremely high Amenity levels. This suggests that these are locations that receive incredibly high footfall from non-residents. Couple this with the fact that there are no nearby Korean Restaurants and these look like extremely desirable locations. It is clear that many other Restaurants operators think so too, with an already-significant number of Restaurants in the area.

*N.B. It is quite clear that the data is duplicated for Downtown Arts District and Downtown Historic Core. These Districts are clearly side-by-side, or effectively the same, so can likely be served effectively by a single Restaurant.*

For completeness, we will take a look at some of the lowest Desirability scores.

### Lowest 5 Districts based on Desirability Score



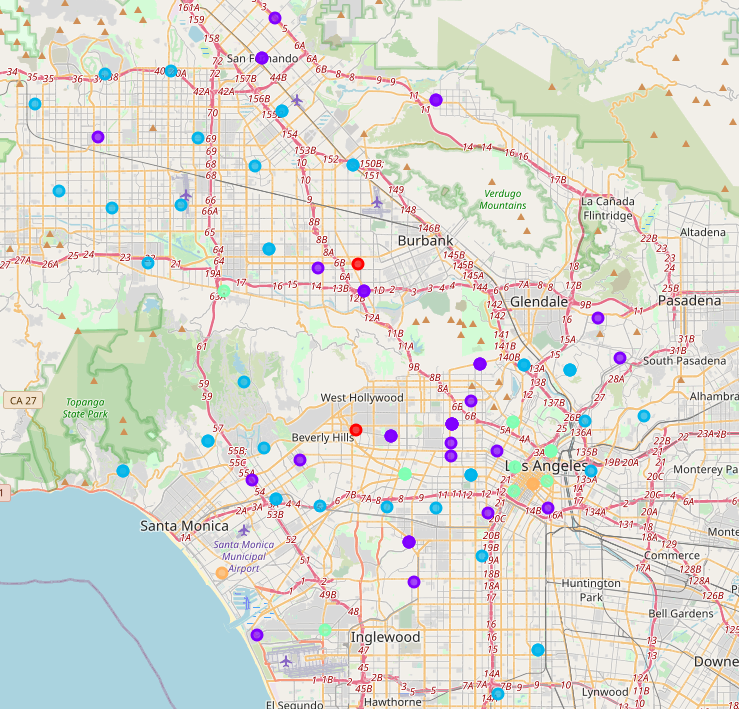
These again appear to be closely related areas, with similar zip codes. Notice the extreme lack of local Amenities. There is reasonable population, but little else of note. It is clear that that other Restaurant operators have also shown little interest in these areas. It is likely that these are primarily residential Districts, with residents tending to leave the District to use Amenities elsewhere.

## Clustering Results

After running the K-means clustering discussed in the methodology section, we have superimposed our clusters onto a map of LA.

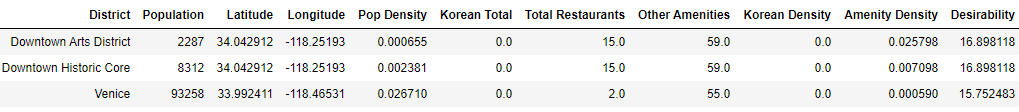
Legend:

|  |  |
| --- | --- |
|  | High Desirability |
|  | Good Desirability |
|  | Moderate Desirability |
|  | Low Desirability |
|  | No Desirability |

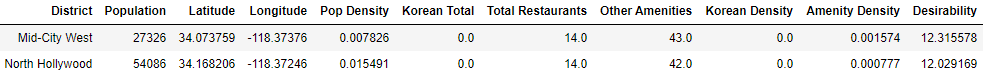


We can see a highly desirable location in Downtown Los Angeles. This is the Downtown Arts District/Historic Core mentioned previously. There is also another desirable area in the Venice area, near to Santa Monica. Let’s display our clusters in closer detail.

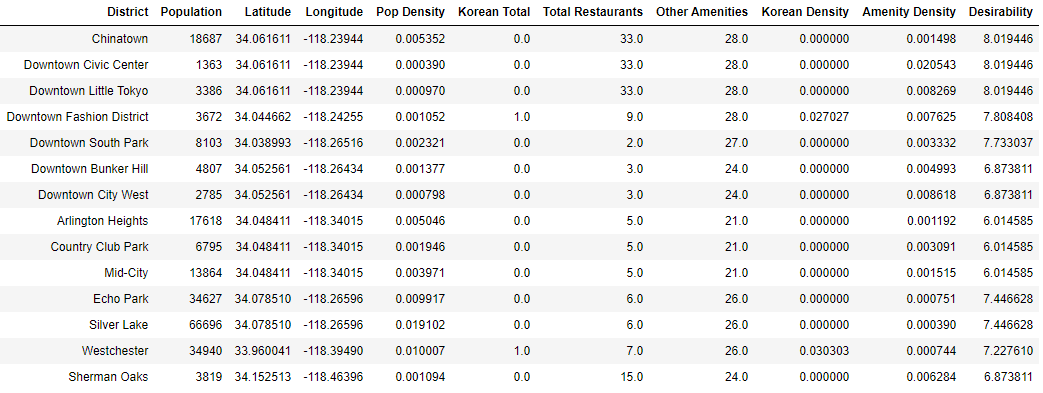
**High Desirability**



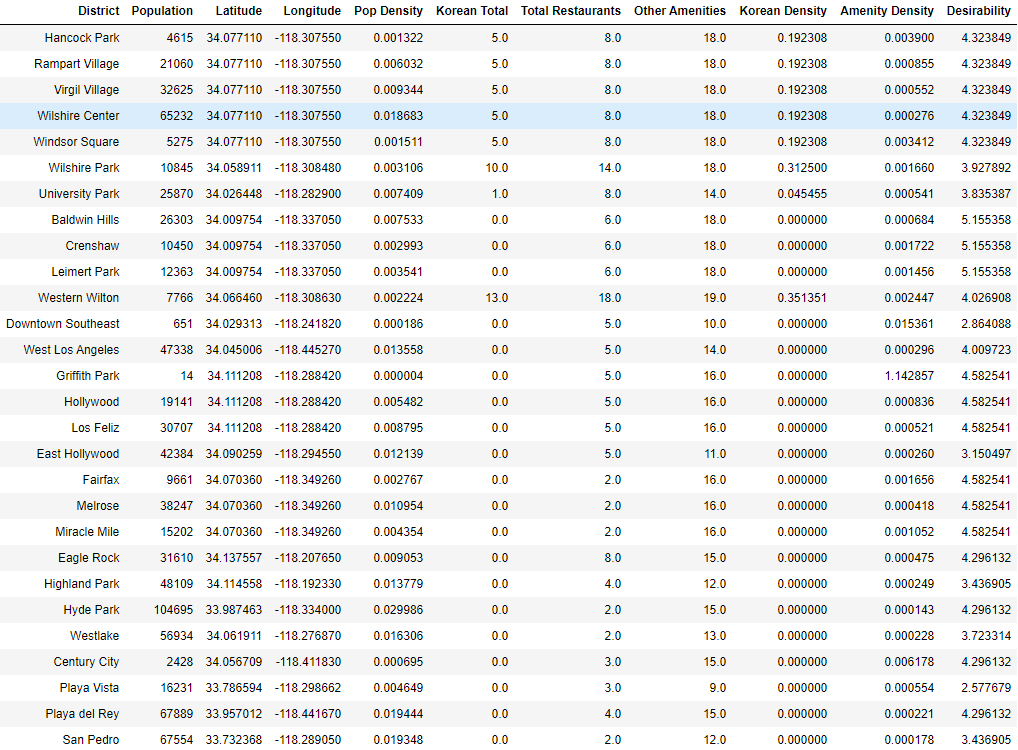
**Good Desirability**



**Moderate Desirability**

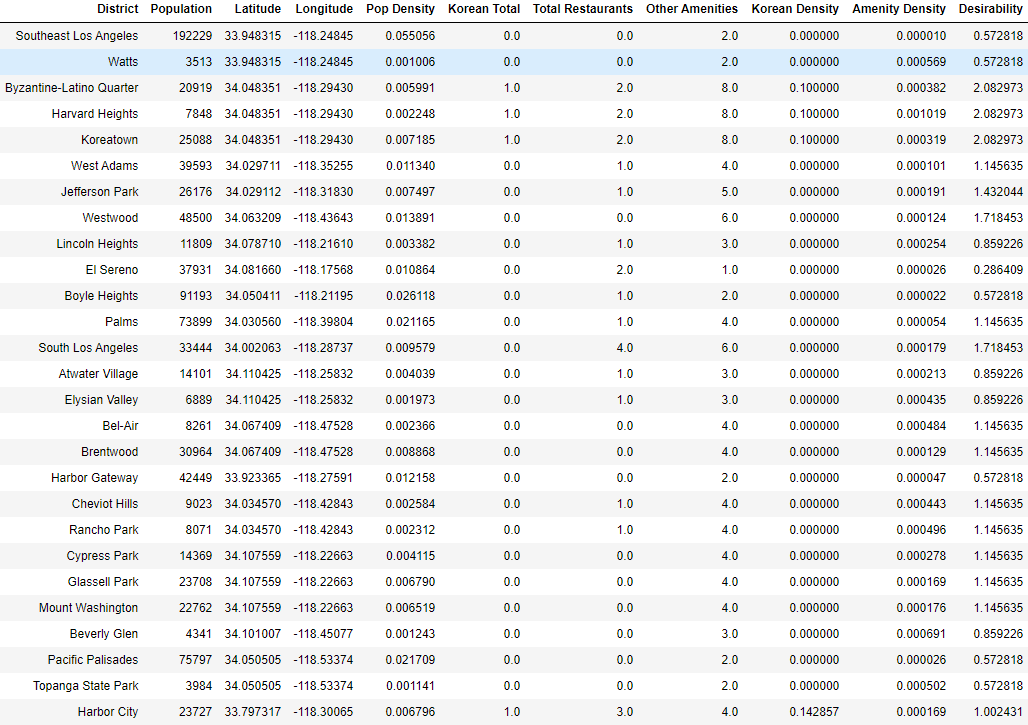


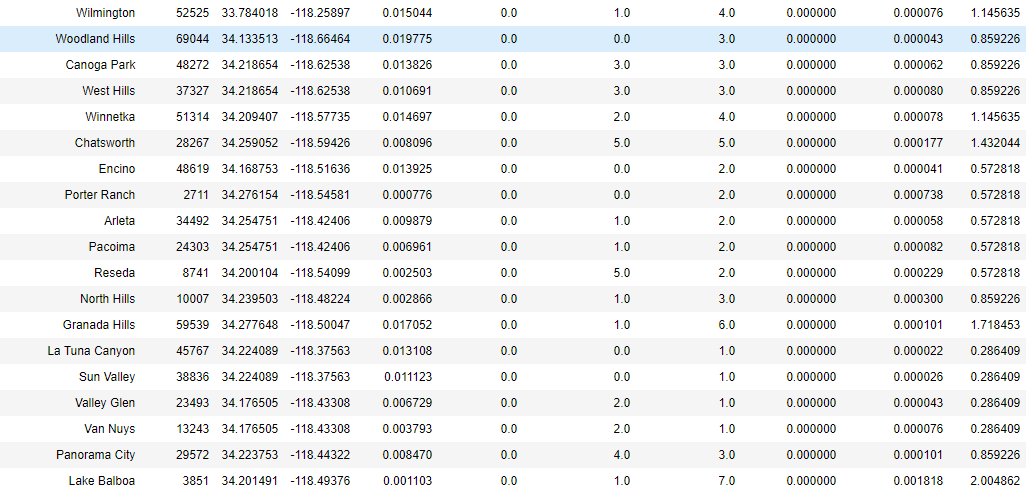
**Low Desirability**





**No Desirability**





# Discussion

We are satisfied that our methodology is sound in terms of evaluating the best locations for a Restaurant, on the basis solely of Population and surrounding Amenities. In reality, the decision to locate a Restaurant would also be based on other factors such as:

* Affluence of the surrounding Neighbourhood
* Demographics and local social attitudes
* Availability/affordability of a suitable premises for the Restaurant
* Ability to obtain an operating license for the Restaurant
* Logistics involved in supplying goods to the Restaurant

We therefore recommend that our client conduct further research along these lines, and are careful to make their decision based on their own industry knowledge.

We also caveat that the FourSquare data provided to us is reliable and up to date. FourSquare is known to be extremely accurate, but it will be worth reviewing again prior to finalising a location for the new Restaurant. We would also recommend checking planning applications to see if any undesirable competition may soon set up in a target location.

We are confident in the Desirability score metric created. Logically, number of Amenities is a strong indicator of probable footfall in an area, and this is born out by other Restaurant operators also choosing to operate in areas that are dense with Amenities. Local Population is also used as an input, since this will also drive trade, but it is our opinion that total footfall in an area is the most important factor. Finally, we used the prevalence of Korean Restaurants to modify the score down, thereby meeting all of our client’s needs.

The clustering undertaken was relatively simple, but sufficient for this exercise. We have provided our client with a number of shortlists. We recommend beginning with the “High Desirability” and “Good Desirability” clusters, with investigation worthwhile on the “Moderate Desirability” cluster as a fallback if no suitable sites are found from the top two clusters. These “Moderate Desirability” locations may also be considered for further expansion later down the line, once our client has become established within Los Angeles.

We do not consider that locations in the “Low Desirability” or “No Desirability” clusters should be pursued by our client. Therefore, significant time and effort can be saved in focusing on the smaller, more desirable clusters.

# Conclusion

We have successfully harnessed a variety of location data in order to create a simplified model of Los Angeles, and paired that with a wealth of data from FourSquare. Appropriate visualisation, together with a balanced scoring of each District has output some strong candidates for our client. Our client will now be able to prioritise further investigation based on a shortlist of locations, and have confidence that the competitive and commercial setting is likely to meet their business goals.

We once again would recommend that our client conduct further investigations as they feel are appropriate. Our analysis here only looks at some elements of selecting a location for a Restaurant, and we invite our client to use their own domain expertise to find the absolute best location for them.