**Applied Data Science: The Ramen Phenomenon hits LA!**

As part of IBM’s [final capstone project](https://www.coursera.org/learn/applied-data-science-capstone), I’ve attempted to leverage Foursquare’s geospatial data to determine the best locations to start a Ramen shop in Los Angeles. Links to the source code are found at the end of the post.

So without further adieu, let’s dive right in:

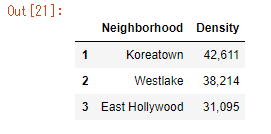
**Part 1: Introduction/Business Problem**Ramen is growing in popularity. [Datassential reports](https://www.getflavor.com/ramen-phenomenon/) a 46% jump in mentions over the last 4 years and predicts this trend will increase by another 40%.

So it is no surprise that Los Angeles, which has [one of the largest Japanese communities](https://en.wikipedia.org/wiki/List_of_U.S._cities_with_large_Japanese-American_populations), is home to many new and old ramen shops.

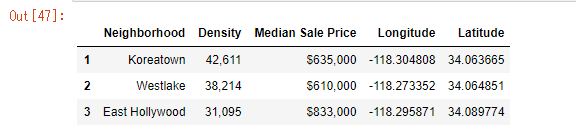
***But with so many neighborhoods in LA, the options are overwhelming. Where would be some good locations to establish a new ramen chain?***

**Part 2: Data Selection**

Fortunately, the [LA Times](http://maps.latimes.com/neighborhoods/population/density/neighborhood/list/)’ data on neighborhoods is very comprehensive. We can scrape the population density of every neighborhood in LA and create a top 5 list.



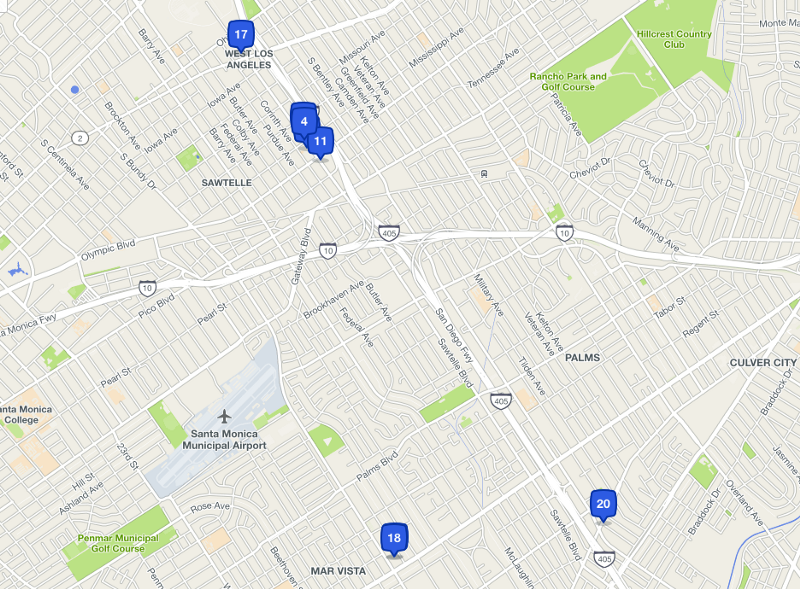
Next, we can add median rental prices from [Property Shark](https://public.tableau.com/profile/property.shark#!/vizhome/LATopNhoods2018/Dashboard1) for our top 5 neighborhoods. Fortunately, this list also contains our geospatial coordinates.



Leveraging [Foursquare’s](http://Foursquare) location API, we then gather detailed information on the price range, popularity, and trends of comparable venues in these neighborhoods

**Part 3: Exploratory Data Analysis**

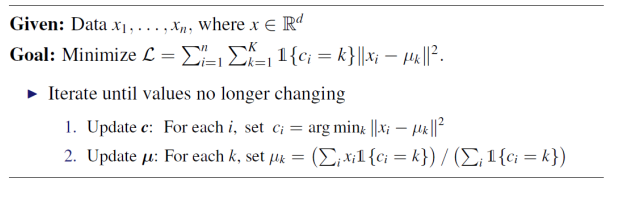
We can see in the folium map the nearby venues for our potential ramen shops



**Part 4: K-Means Clustering the LA Neighborhood data**

Based on our insights from Part 3, we can try clustering our neighborhoods based on the venue categories. For this, we will be using [K-Means clustering](https://blog.easysol.net/machine-learning-algorithms-3/), a form of unsupervised machine learning.

A brief explanation of the K-Means algorithm:

Source: CSMM.102x Machine Learning (Columbia edX course)

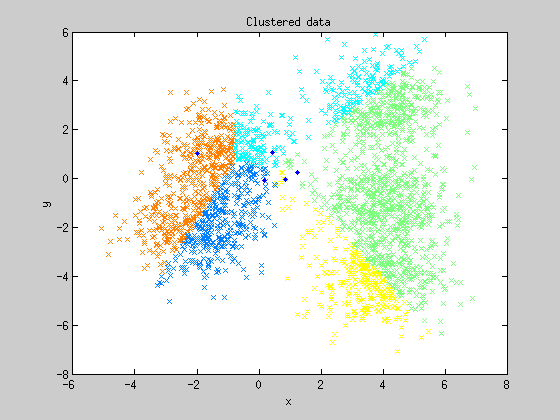
We take a random data set (our given data points x1, x2…) and define *k* centroids. From there, we perform the following 2 tasks *iteratively*:

1. Assign each data point (x) to the closest corresponding centroid (k), using the [standard Euclidean distance](https://en.wikipedia.org/wiki/Euclidean_distance), i.e., the straight-line distance between the data point and the centroid.
2. For each centroid, calculate the mean of the values of all the points belonging to it.

That’s it.

The key here is that at the end of each loop, the mean **becomes the new value of the centroid**. The process reaches convergence when the values of our centroids no longer change, i.e., the cluster assignments no longer change.

We can plot each iteration and see where our centroids shift!

[Source](http://practicalcryptography.com/miscellaneous/machine-learning/yet-another-k-means-tutorial/)

Note this is referred to as “hard” clustering because each data point is assigned to one and only one cluster — **hard-assignment**. This is in contrast to soft-assignment (outside the scope of this project) where each data point is assigned to *every* centroid with different corresponding weights ϕ\_i(k) for each data point.

Applying this process to our data set, our expectation is that, based on the similarities of the venue categories, these districts will be clustered.

**Part 5: Conclusion**

Ultimately it seems like Koreatown is the best place to start a ramen shop due to the proximity and population density!