**Introduction**

Our firm has been hired by the Whole Foods grocery store chain in the states, which was recently acquired by Amzaon. They are looking to branch out into the healthy, organic, vegan, fast-casual restaurant business and want us to run some analytics to find the best possible location based on their set of criterion. Initial internal studies suggest they need to open up their first location in Los Angeles, CA, often referred to simply as “LA”. But where in LA should they open their first ‘joint’? The first store is very important as it sets the tone for the franchise, so it’s important to get it right. LA is a very large, dynamic and diverse city with a couple hundred neighborhoods. Topographically, we have everything from beautiful beaches to the mountains to the deserts. Socioeconomically, it ranges from the very rich in the world to the very poor. The rich ethnic diversity also means we have some of the best, most diverse restaurants in the world. As you may know, Whole Foods’ clientele consists of health conscience individuals who are not why about spending extra money for organic groceries and foods. They would like to carry that business model into their restaurants. The type of food would be very similar to the type of foods they already have in their grocery stores.

My client's criteria:

1. Lots of foot traffic (population density)
2. High mean household income
3. Health conscientious population
4. Not too much competition. Don’t want to enter a crowded space/market.

**Data**

The data we will be using is going to be a combination of census data and Foursquare data. The census data is a CSV file we downloaded from the US Census Bureau website and the Foursquare data will be obtained via an API.

The census data consists of every single zip code in the LA County with their corresponding population density (population per square mile), median household income and latitude longitude – later to be used in conjunction with the Foursquare data and for mapping the clusters.

The reason we are using zip codes instead of neighborhoods is because it is the finest granularity in LA. You can have two or three zip codes within a given neighborhood.

The reason we are using population density instead of just population is because they want to be in a neighborhood with high foot traffic as that’s what they’re expecting most of their business to come from. Population density is a great way to measure foot traffic.

We will run the KMeans clustering machine learning algorithm to determine which cluster to focus on based on the population density and the mean household income. We will also use this method with the Foursquare data to determine the types of venues within each of the zip codes we analyze after we narrow it down by the census data. For example, if the first 5 most commonly venues in that zip code are organic restaurants, then we will focus on the next best location. Or if we notice a gym in the neighborhood, that’s a good indicator that they the residence of that neighborhood are healthy eaters, which is exactly what my clients are looking for.