**Introduction**

Our firm has been hired by the Whole Foods grocery store chain in the states, which was recently acquired by Amazon. They are looking to branch out into the healthy, organic, vegan, fast-casual restaurant business and want us to run 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 shy about spending extra money for organic groceries and foods. My clients would like to carry that business model into their restaurants. The type of food would be very similar to the types of foods they already have in their grocery stores; sushi, pizza, salad bar, cheese bar, sandwiches, etc.

My client's criteria:

1. Lots of foot traffic (population density)
2. High median 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, so looking at neighborhoods only would be too broad of an area to successfully pinpoint the best location. The higher the granularity, the better for our dataset.

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.

**Methodology**

When we took on this project, we knew right away we needed to get our hands on some rather detailed census data for Los Angeles County. We wanted to see what types of people were living where, how much were they making and how populated was the region.

High Median Household Income

Since our clients were looking to open a high-end fast casual location, we knew we needed to first look at zip codes with high median household incomes, which we will refer to as HMHI from now on in this report. So without looking at any other data, we simply looked down the list and sorted by the HMHI per zip code. As it turns out, there are plenty of neighborhoods and zip codes with HMHIs, but that doesn’t mean they make good locations for a fast casual restaurant. The reason for that is because most locations with HMHIs have very sparsely spread out neighbors. What contributes to this is simply that rich people own larger lands and homes and usually there aren’t any apartment buildings around. We knew population had to be a bid part of our algorithm.

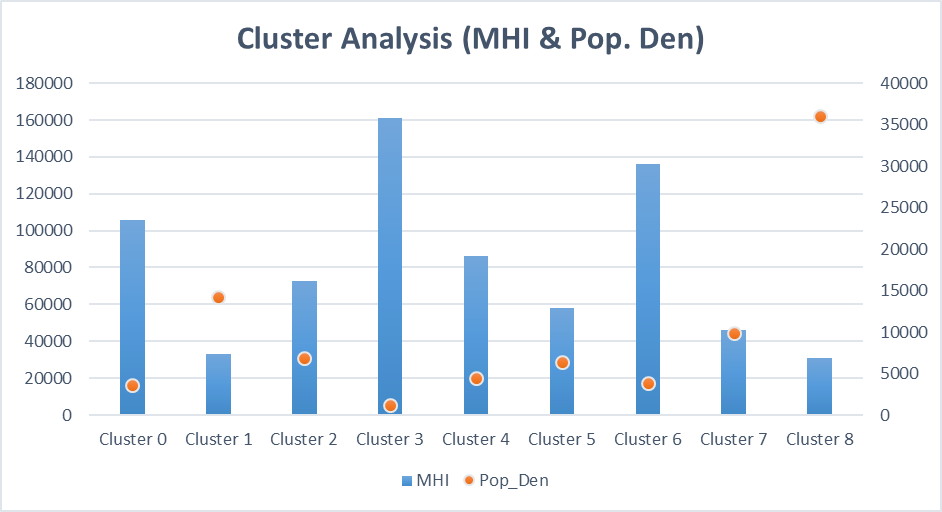
Population

The next thing we looked at was population. Which zip codes have the highest population combined with the highest median household income. Upon manually studying this data, it became evident rather quickly that we needed to look at population density. The reason for this is because you can have a zip code with lots of people living there, but that could be due to the fact that the zip code itself simply covers a lot of square mileage. For some restaurants, that might be completely OK because they don’t really have a need for foot traffic. As long as they have enough parking spots in a decent location, that would be sufficient for them. However, our client’s business is fast-casual and relies heavily on foot traffic. Parking spots are not relevant to them.

Population Density

When looking at the population didn’t yield satisfactory results, we turned to population density. Luckily, the US Census Bureau already calculates population density and we were able to roll into our data frames for analysis. So how do you define population density? It is the number of people per square mile. It normalizes the population data. This would be perfect data for our clients since they want their restaurant to be located in a very densely population location. With this data, we felt comfortable moving forward and comparing the many different locations Los Angeles County has to offer.

With the HMHI and Population Density data loaded in our data frame, we used the KMeans clustering method to find and narrow down for us the most similar zip codes we needed to focus on. There are 241 zip codes within the Los Angeles County and leveraging this method was crucial to finding the perfect location for our clients. After running the KMeams elbow method to determine the optimal number of clusters, we found out that our most optimal number was using 6 clusters. We ran the numbers with 6 clusters, but we weren’t entirely happy with the outcome. We wanted to narrow the zip codes down further and we were glad to take on the extra work/analysis. Therefore, we ran 9 clusters and were very pleased with the results. After analyzing the data by running the ‘.describe’ function for each of the clusters, we discarded the clusters with high mean population density number mixed with very low median household incomes (Clusters 1, 7, 8). Our client would not be interested in these zip codes and it indicates lots of low earners living in apartments. The next set of clusters we looked at were the ones with very high median household incomes, but with a very small population density number (Cluster 3). This signified the neighborhood consisted of large homes with very few folks that lived far from each other. This is not ideal for my clients also as there will likely be almost no foot traffic and not enough people to make for a successful venture. We then turned our focus to two clusters that had the highest mean population density with the highest mean HMHI (Clusters 0 and 6). We will discuss the results later in this report along with any interesting observations we made here. It is now time to discuss our next approach/method with the Foursquare data.





Foursquare Data

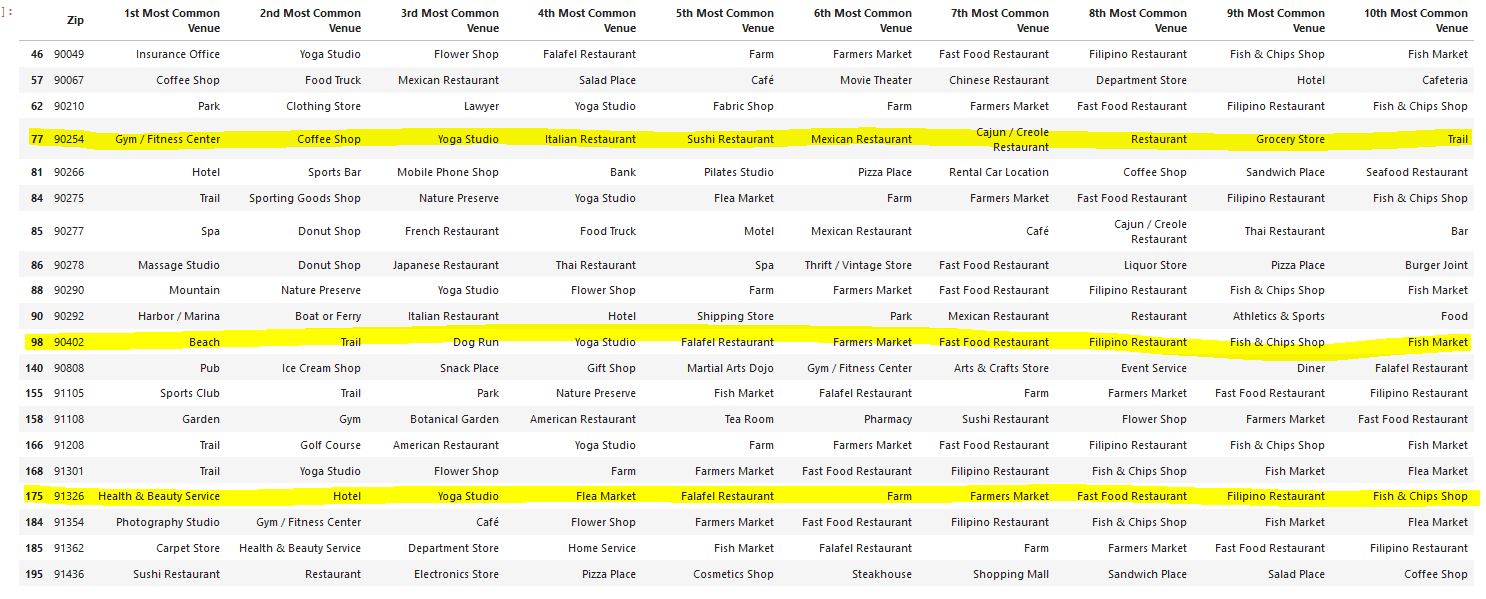
Now that we have chosen our clusters, it was time to analyze this cluster with the Foursquare API. Our clients’ needs were that it had to target a zip code whose population was health conscience. We can simply look at the zip codes with the highest median household income and the highest population density, but it could be a very unhealthy neighborhood. It can be a very wealthy retirement community and being healthy and eating organic is the last thing they care about.

At this stage, we turned our attention to the most common types of venues within these zip codes by using one-hot encoding. We were looking for zip codes with gyms, yoga studios, etc. Additionally, we were also looking out for types of restaurants. If the top 5 out of 10 venues were restaurants or even had maybe a couple organic/vegan restaurants, we’d reconsider that location despite the demographic and health gauge data. The reason being although it might signal that that zip code can sustain such places, we also didn’t want to enter a crowded space. Ideally, we want lots of gyms and not too many restaurants. That would be a slam dunk location for our clients.

**Results**

Between Clusters no. 0 and 6, we got a total of 27 zip codes to compare. After running a one-hot encoding method on those 27 zip codes with the Foursquare API, three zip codes jumped out at us: 90254, 90402 and 91326.

The cities were Hermosa Beach, Santa Monica and Porter Ranch respectively. The reasons these three jumped out at us is because they all showed some level of their population living a healthy lifestyle, which was a very important criteria for our client since they are looking to open up an organic/vegan fast casual restaurant. Santa Monica for example has a beach, trail, dog run and Yoga Studio as their top 4 most common venues. We liked that Porter Ranch has a Flea Market a Yoga Studio and a hotel. You can see a theme here. They all have Yoga studios and not much competition for restaurants. And the few restaurants that show up here aren’t the healthiest or considered organic/vegan. Upon further investigation, our recommendation was to go with a location in Hermosa Beach simply because of its population density. It is through the roof at 13,041 population count vs 3259 for Porter Ranch and 5745 for Santa Monica. Although Santa Monica had a much higher Median Household Income of $134,286 vs $111,187 for Hermosa Beach, the foot traffic that high of a population density could generate would be very enticing to our clients whose no. 1 criteria was ‘high foot traffic’.



Discussion/Observations

Recommendation

Based on our research, we would recommend my clients to open up their first restaurant in Hermosa Beach. It seems to be the most densely populated zip code with the highest median household income who are also very health conscience. We’re sure there are very densely populated areas of Santa Monica as well, and more research needs to be done on foot to further pick a location between Hermosa and Santa Monica. But for now, Hermosa Beach seems like a very encouraging place to start. However, it can be safe to assume, based on the criteria given to us by our clients, they would thrive in most of the beach towns in Los Angeles.

Interesting Observations

1. When we ran a cluster analysis on just the Population Density with the Median Household Income, it clustered a lot of the beaches/beach towns together. This confirmed our suspicion that people would pay almost anything to live on or near the beach. The rich are moving closer and closer to each other, in smaller spaces just to be by the beach. You have your rich neighborhoods with much bigger homes whose population is making just as much as folks living in a condo but by the beach.
2. When we ran a cluster on all 241 zip codes in Los Angeles County, it clustered a lot of trails and farms and farmers markets together. You can tell those were the outskirts of LA County without even looking at the map. We were also able to tell which areas were the poorer zip codes as they contained the higher frequency of fast food restaurants. One such cluster had identical most common venues from no. 1 to no. 10 and they were 18 miles apart. It was very interesting.

Conclusion

In conclusion, our clients had several specific criteria they gave us in order to analyze and suggest the best possible location for their new, organic, and healthy, fast-casual restaurant. We used these criteria as instructions for how we wanted to handle this project. We grabbed the demographics data from the US census bureau, scrubbed it for Median Household Income and Population Density. We found our target clusters, which consisted of 27 zip codes. With those zip codes now in hand, we went ahead and applied the Foursquare API data to it and did some more analysis on venue categories. They wanted not just a wealthy neighborhood who can afford a $10 slice of organic pizza, but also a location with not too much competition and whose population is very health conscience. By analyzing these 27 zip codes we were pleasantly surprised to find many of them had gyms and yoga studios in their top 5 most common venues. This implied they are a very health conscience population. After narrowing it down to our top three, we recommend our clients to further investigate Hermosa Beach, CA. They are wealthy, health conscience group of folks who live very close to one another. Gym is the no.1 most common venue and Italian restaurants are their no. 1 most common restaurants, which is not very healthy. This implies that there is huge potential for a restaurant who puts a lot of emphasis on healthy eating. Thank you for reading.