Final Report: Restaurant Placement in Los Angeles

The Problem: Restaurants appear in Los Angeles like shooting stars, appearing suddenly in a blaze, the fading off into the horizon. Many of the restaurants that go out of business in the area are well designed, have an interesting menu, and great staff, but suffer from placement. If a new restaurant is seen in a neighborhood that is too similar to others nearby, the locals are naturally not going to abandon their favorite eateries in lieu of something new, and potentially inferior. This project is designed to assist a prospective restaurant developer in exploring the city using data. This can give us far more information, much more quickly, than exploring a neighborhood on foot.

Foursquare: This project will make extensive use of the Foursquare API to harvest data from the Foursquare database. The venue information in Foursquare is reliable and generally not contained in other databases (county, state, and municipal registration databases are often incorrect about simple classification items like business type and category, I have checked). The crowdsourced nature of the database ensures its general reliability and current data status. For our purposes the "category" information will be our primary metric and standard for restaurant placement.

Analysis of Foursquare data can give us the layout of establishments on a map, telling us which areas would benefit from a new restaurant, potentially of a certain type. Some areas, in the course of this investigation, were found to have a number of commercial establishments nearby, but not enough eateries. This type of situation is ripe for a commercial venture and promises to be more profitable and stable than choosing a location based on happenstance availability.

The audience we are trying to reach with this study is any venture capital firm that would be willing to fund the expansion of the restaurant sector. The idea does not have to have a theme, such as a jungle-themed restaurant, it merely has to be profitable. We are looking to fill a vacuum, not entice people from across the city to sample the restaurant's fare.

Method – K-Means: The data will show us where to place a new restaurant that can meet a need in the neighborhood. If this is replicated ten times over, it will provide the owners and business funding agents a steady, stable source of additional revenue. We will cluster the restaurant types across neighborhoods using the k-means clustering algorithm.

Data Framework: For this developmental process, we will make extensive use of Python, utilizing Jupyter Notebooks to process the data and algorithms.

Python Libraries: Python Libraries provide a ready-made assortment of scripts and algorithms to execute the processes we need to perform on the data. The libraries we will make use of will include:

1. json – library to handle json files
2. geopy – converting addresses into latitude and longitude / geocoding
3. folium – for graphing data
4. BeautifulSoup – to scrape web pages
5. csv – to read and write csv files
6. pandas - to read and create data frames
7. numpy –perform data analysis on frames
8. matplotlib – to graph data and colorize plots on a map
9. sklearn – this library provides us with data analytics algorithms including k-means clustering

Wikipedia: The list of neighborhoods in Los Angeles was scraped using BeautifulSoup from this page: https://en.wikipedia.org/wiki/List_of_districts_and_neighborhoods_in_Los_Angeles. This is the first data set from which we determine the clustering of venues – each venue by category is classified by neighborhood.

ArcGIS: ArcGIS will be used to geolocate all the neighborhoods. This will be the starting point from which we locate the relevant categories of venues around each neighborhood.

Foursquare: The Foursquare API for "Places" will be essential in:

1. Locating venues in each neighborhood from the geolocated coordinates from ArcGIS
2. Classifying each venue according to the categorical information in the Foursquare database

Once all venues are plotted and mapped, pivot tables will be used to classify each category of venue by most frequent in that particular neighborhood.

Once the frequency and quantity of venues in the neighborhoods is established, the data will be analyzed by plotting and pivot for vacuums of restaurants in commercial areas.

Methodology: Several processes were performed to obtain the final data comparison. The method I used was executed as follows:

1. Extract the complete list of neighborhoods of Los Angeles from Wikipedia using BeautifulSoup.
2. Use ArcGIS to geolocate all neighborhoods.
3. Perform a search through the Foursquare API to find all relevant venues marked by neighborhood.
4. Perform a criteria search using the Foursquare API and find the categories for the venues.
5. A pivot function is used to find the top 15 venue types for each neighborhood.
6. All venue geolocations are fed into a k-means cluster.
7. The clusters are analyzed for venue frequency.
8. Any vacuums for dining locations are identified.
9. The final location (neighborhood) was discovered and a recommendation was made.

Results: Arlington Heights was indicated as a good place to place a small diner to maximize profits. Arlighton Heights is a small, diverse neighborhood. It has karaoke bars, grocery stores, a large art gallery, and several convenience stores, but a startling lack of restaurants.

Discussion: The k-means clustering worked surprisingly well, given the data set was so large. Los Angeles is a dynamic city with many restaurants. The algorithm gave the most frequently seen establishments in each area and plotted the clusters appropriately. Some clusters only had a few neighborhood – establishment crosses, which eliminated them from the search, but narrowed down the possibilities and made the overall effort easier.

Conclusion: K-means clustering has a large number of applications and data analysis in real estate and commercial venue analysis is only beginning to show prominence in the business landscape. The application for this process is wide and can be applied to any large-scale urban environment with multiple neighborhoods and a variety of establishments.