IBM Data Science Professional Capstone Project

Opening a Mexican Restaurant in San Diego

By J.B. Paragas

Introduction

For many people, eating at a Mexican Restaurant is a great way to relax, indulge in a bit of culture, and enjoy themselves. For others, a Mexican Restaurant provides a great way to earn a consistent income stream. As with all other business decisions, opening a new restaurant requires a lot of thought and cognizance of several factors. In particular, the location of a restaurant can help determine its success. This project aims to help property developers and entrepreneurs by recommending some cities in San Diego to consider.

Data

This project will focus on cities in San Diego, California as listed on Wikipedia. We will utilize Nominatim, a part of the geopy Python library, to search through OpenStreetMap data to collect geographical coordinates of each city. We will also query the FourSquare API to collect data on venues within a two-kilometer radius of each city.

Methodology

In order to create a list of cities, we needed to create a web-scraping function using the Python library Beautiful Soup. The results were cleaned to only include city names - every output was of the form "City Name, California" until we used the *split* function in Python.

Once the dataframe of cities and corresponding geographical coordinates was generated, the powerful Python library folium was imported in, in order to create a Leaflet map within our notebook.

To leverage FourSquare API data, we needed to define a function to return at most 100 venues within a 2-km radius of each city. The function included an API request URL and a *get* request. We focused most of our energy on a venue's *Category*.

In order to prepare our data for a machine-learning algorithm, we used the Pandas function *get_dummies* to convert categorical variables into dummy/indicator variables. This blew up the dimensionality of our dataframe, made slightly larger by using the *groupby* function. Grouping the new dataframe by city, then calculating the mean number of each venue category, results in over 4000 pieces of data.

Focusing on the top five venue types per city, Mexican Restaurants repeatedly showed up in the results. After analyzing each city and taking the mean of the frequency of each venue type, we filtered out the "Mexican Restaurant" category for each city.

The data was clustered using the machine-learning algorithm k-means clustering. This algorithm identifies a user-defined number of centroids, and allocates each data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simpler, more popular unsupervised learning algorithms and is well-suited for the task at hand. We clustered the cities into 3 clusters based on their mean values of Mexican Restaurants. The results will assist in identifying which cities have higher concentrations of Mexican Restaurants.

Results

Cities were categorized based on how frequently venues were classified as Mexican Restaurants: Cluster 0: Cities with moderate number of restaurants; Cluster 1: Cities with low number of restaurants; Cluster 2: Cities with high number of restaurants

On the map, Cluster 0 cities are marked in red, Cluster 1 cities are marked in purple, and the rest are in a mint green color. Which can be hard to see on the map.

Discussion

Mexican Restaurants are concentrated in clusters 2 and 0. The two Cluster 2 cities, in particular, likely suffer from intense competition due to oversupply and overconcentration of restaurants. Cluster 1 cities present the greater opportunity and high potential areas to open a new Mexican Restaurant. Cluster 0 cities have a higher level of competition among Mexican Restaurants, but may offer a conservative level of risk for a new business.

Conclusion

I recommend property developers to capitalize on the observations above and look into neighborhoods in *Cluster 1* cities to open a Mexican Restaurant with a high potential for success. In particular, the city of **Del Mar** should be explored for high potential neighborhoods. Property developers with unique selling propositions that stand out from their competitors may find better success among their competition by contesting the existing Mexican Restaurants in **Encinitas, Coronado**, and other *Cluster 1* cities. Property developers looking to mitigate risk may find success in *Cluster 0* cities like **Santee, San Diego, and Imperial Beach**. The lower level of risk will mean trading away a potentially higher profit margin. I advise property developers to avoid neighborhoods in *Cluster 2* cities, which already have a high concentration of Mexican Restaurants and suffer from intense competition.