Capstone Project – Battle of Neighborhoods

City Populations and Nearby Venues

1. Introduction

Los Angeles County is a metropolitan area with more than 10M in population. It's a vibrant hub with a huge variety of businesses and it's intriguing to find out different types of consumer venues that are driving the economy using Foursquare API. I've decided to pin down on coffee shops as the subject for analysis for this capstone project as coffee is an indispensable part of the city's culture. With the knowledge that I acquired from this sequence of IBM data science courses, I decided to further explore the dynamics of coffee shops in comparison with the cities' populations within Los Angeles.

2. Business Problem

Population undoubtedly positively correlates with business activities. By using population and venue data for each city in Los Angeles, if I am planning to open a coffee shop today, which city should I consider setting my coffee shop in?

3. Data

In order to address the problems listed above, we will utilize the following:

- List of neighborhoods in Los Angeles County
- Populations of LA County's neighborhoods
- Latitude and Longitude Coordinates of the neighborhoods (This will allow us to extract the venue data using Foursquare API)
- Venue Data relating to coffee shop

3.1 Source Data

LA County Population Data from Wikipedia:
https://en.wikipedia.org/wiki/List_of_cities_in_Los_Angeles_County,_California

The table contains 88 cities (a.k.a neighborhoods for our analysis purpose) that are located within the Los Angeles County. We will be able to scrap the data using BeautifulSoup package and convert the JSON format data into Pandas Dataframe. The next step will be converting the cities' locations into geo-coordinates (Latitude and Longitude) that are used to feed our searches using Foursquare API. Foursquare API will allow us to query specific location and returns the venues within specified locations. The final analytics step will require using the KMeans package from sklearn to stratify our locations with venue and population data that we clean.

3.2 Methodology

We will first use the BeautifulSoup package to scrape information from Wikipedia and convert the information into useable DataFrame format. The table returns 88 cities with Los Angeles County and we subsequently utilize the geocoder package to translate each city into geo-coordinates (Latitude and Longitude) and append the data into our Dataframe. The initial data cleaning/manipulating step is complete as we have an 88x5 DataFrame with “City”, “Date of Incorporation”, “Population”, “Latitude” and “Longitude” information.

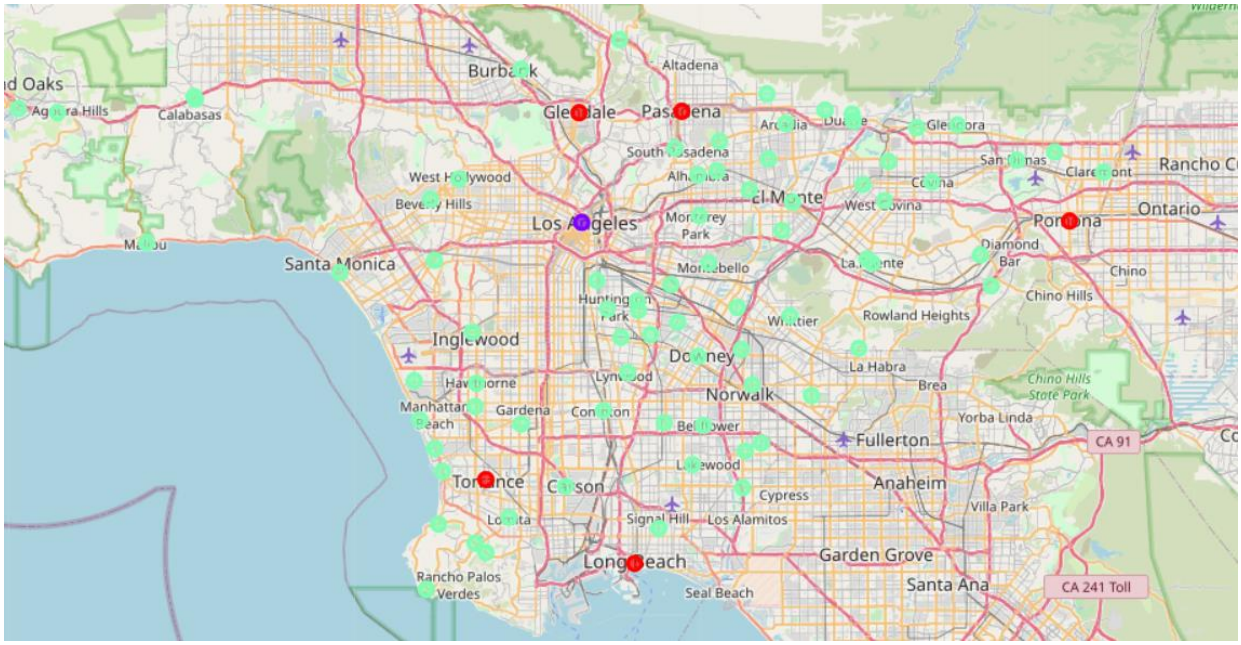
We then utilize the Foursquare API and create a function called “getNearbyVenues” that make calls to the API and return nearby venues by feeding in the latitude and longitude data that are in the previously cleaned dataframe. We set the “limit” to be 100 (it will return up to 100 venues for each location we specified) and “radius” to be 2000 (searching for venues within 2000M radius circle). The function successfully returns a new dataframe with venue information such as “venue name” and “venue category”.

After we have obtained the venue information in the dataframe created by the function above, we further manipulated the dataframe by grouping it with “Neighborhoods” and calling “getDummies” to create a 88x346 dataframe that summarizes venue category values (using mean) for the 88 neighborhoods that are subject to our analysis. Then we add back the population data that’s included in the previous dataframe.

Now after a series of data cleaning and manipulation steps, we are finally ready to call our Machine Learning function “KMeans”. This is a relatively simple step as the data has been cleaned and readily available for clustering. We perform a silhouette test and noted that the silhouette score decreases with an increase in the number of specified clusters. Therefore we elected to choose three clusters as it generates better representation of clusters comparing two (which yields the result of only having one county in a cluster and 87 remaining ones in the other).

The final step performed is the plot the coordinates with identified clusters on a map using the Folium package.

4. Results



Coffee Shop Score: We will now refer to a key indicator for our analysis as the "coffee shop score" that was generated by the "getDummies" dataframe. 0 represents 0 coffee shop exists while 1 represents 100 coffee shops are found within our 100 search results.

Three clusters are identified

a) Cluster 0: Labeled as red in the map above. Coffee shop values between 0.039 and 0.07. The population is more than 130K for each city included in the cluster.
b) Cluster 1: Labeled as purple in the map above. I consider this an outlier cluster because the population is significantly larger than all other cities included in our analysis (LA City has 3.8M population as it's the central area of LA County). The coffee shop value is 0.06, which is in the above-average range since it's a city with high population density.
c) Cluster 2: Labeled as green above. These are relatively smaller cities with population less than 120K. Coffee shop score ranges from 0 to 0.0923.

5. Discussion

The analysis provides good insights into the potential market size (population) and the intensity of competition (coffee shop score) within each city. Using the map to visualize how the clusters behave conveniently helps draw insights such that red clusters represent greater volumes of customers yet fierce competitions are expected. On the opposite spectrum, choosing a green cluster to set up our coffee shop can potentially mean breaking into an untouched coffee market with considerable amount of customer population.

5.1 Limitations

However, this analysis is relatively simple as it neglects many other factors that relate to very low coffee shop scores in cities that have considerable amount of population. Culture factors and the presence of chained coffee shops (e.g. Starbucks) can be some of the elements causing the low coffee shop scores. One particular phenomenon that I've observed is that coffee scores are almost close to '0' in Asian neighborhoods (where I reside). This has to do with cultural inclinations that were not included as part of our analysis. Additional shortfall of our model was that we relied on the search results by Foursquare, which could potentially limited the scope of our searches by not including the most complete and accurate results.

6. Conclusion

It was overall a very fun and practical project that explored a subject that I've always wanted to know – the coffee market dynamics in each neighborhood. The analysis stratified three different clusters for us to consider where it's best to set up a coffee shop in with considerations such as coffee shop density and population sizes. My optimal choice to set up a coffee shop will be "Glendale", which is part of the red cluster that represents larger populations and the coffee score of the city is only 0.04. It's the optimal choice provided by the analysis as the city has the highest population of 200K yet the coffee shop score is relatively low.