IBM Capstone project

Introduction/Business Problem

California is a state known for its population diversity. There are increasingly more Hispanic/Latino immigrants moving into the state. Los Angeles is the biggest city in California and is certainly very attractive for both business/job opportunities and people who enjoy city life. Given the population ethnic composition and economy growth, it could be quite profitable to open a Mexican restaurant, if we can locate the right neighborhood.

When we talk about Mexican restaurants, we are actually referring to a collection of different restaurant styles. Recently 'Modern Mexican' food has come to the attention, you can find more and more start to appear in the vibrant areas. Modern Mexican food focuses more on innovation and fine ingredients, it evolved out of the traditional Mexican food and not so much into the authenticity, which is quite distinguishable at the first time you see the dishes. New York Times had a great review article called "'Modern Mexican' Steps Into the Spotlight" (https://www.nytimes.com/2017/05/16/dining/modern-mexican-food-steps-into-the-spotlight.html) that beautifully addressed the essence of Modern Mexican food style. It is healthier, fancier and distinctive, that's also the reason it is particularly attractive to younger generation who would like to try out new things and can afford the "higher than usual" price.

Therefore we need to identify neighborhoods with high Latino/Hispanic population (for potential customer number), significant portion of young people, and better to have high income level. Additionally, these neighborhoods should already be accustomed to Mexican food as a start.

Data

1. The zip code database is USPS data provided by Los Angeles county website (ZIP_Codes_and_Postal_Cities.csv) https://data.lacounty.gov/GIS-Data/ZIP-Codes-and-Postal-Cities/c3xr-3jw2. It contains the city names, zip codes and their location in coordinates.
2. I will use the coordinate data from above to obtain surrounding venue information from Four Square API, using one hot encoding method to categorize the zip code areas for their venues. I will find out which group of zip codes host more restaurant with Mexican cuisine.
3. I will also get selected characteristics of the total and native populations in LA(from US census), from which I will use the "Total population", "AGE-25 to 44 years", "Median age(years)", "Hispanic or Latino origin", "INDIVIDUALS' INCOME" as parameters to screen the target zip codes.

Methodology:

The zip code information was downloaded from LA county website (originally from USPS), then it is loaded to 'la_zip' as a dataframe. It has 370 entries and 4 useful columns, 'zipcode', 'city', 'latitude', 'longitude'. With the coordinate information from the file, I called Four Square API with credentials, the radius was set to 1000 to obtain more venues. As a result, the 'LA_venues_L' variable received 18,462 venues. By using 'groupby' method, we can see there are 464 different venue categories.

To categorize the zipcode according to the venues, I used one hot encoding to create a dataframe with 18,462 rows, removed the rows with null values. By using 'groupby' function, I summarized the each zip code's venue category list, ranked their top venue categories, made a sorted data frame called 'la_sorted' so you easily perceive the what are the most popular venue types in a given zip code.

Before using KMeans to group the zipcodes, I used 'elbow' method to identify the optimum 'k' value. Here I borrowed the 'KElbowVisualizer' method from 'yellowbrick' package to visualize the 'elbow' point. However, although I tried the k values upto 100, there was not a sharp 'elbow' point can be easily identified from the figure. As a result, I am using k=10 for the later analysis, as it is big enough but not overly big causing too little members in each group.

Then I used KMeans method to categorize the zip codes, displayed them in a folium map with 'rainbow' colors. The map is centered at the mean coordinates of all the zip codes, with zoom=12. From the map we can see that the biggest three groups are group 1, 6 and 9. If we display the top 10 venue categories of them, we will find group 6 zip codes areas have the most Mexican restaurants. Therefore I focused on the group 6 zip codes for the further analysis.

Due to the naming difference in each years' census results, the data had to be cleaned before proceeding to the next step. I used two different function 'subset_57' and ' subset_53' to extract the useful information from year 2011 to 2017.

Based on the notion that more potential customers is positively correlated with potentially more business, I used Seaborn package's stripplot to display the five parameters from year 2017: 'Zipcode',"Total population", "AGE-25 to 44 years", "Median age(years)", "Hispanic or Latino origin", "INDIVIDUALS' INCOME", with sorted "Hispanic or Latino origin" value. The figure shows that the median age is negatively correlated with this value, while others are slightly positively correlated.

I assigned the data with over 75% "Hispanic or Latino origin" to a new data frame, used matplotlib scatterplot to look at the correlation between "AGE-25 to 44 years", "Median age(years)" and "INDIVIDUALS' INCOME" in a 3D plot. The figures shows the percentage of "AGE-25 to 44 years" is a better predictor for the higher income. As you may imagine, higher income customers may find our modern Mexican food more attractive.

When I merged the top zip codes with "Hispanic or Latino origin" and data from above mentioned group 6 from KMeans analysis, I got 22 zip codes from the 2017 data. With the 22 zip codes, I extracted different data from year 2011 to 2017, including 'Total population', 'AGE-25 to 44 years' and "INDIVIDUALS' INCOME". Then I used scatter plot to visualize the change of the values over the years for top 6 zip codes, with population on the y axis, 'AGE-25 to 44 years' in colors, "INDIVIDUALS' INCOME" by the size of the dots. The idea is to identify a zip code area with growing high income, growing population and growing 'AGE-25 to 44 years' subpopulation.

Results:

The result demonstrates 90011 is the best choice among the ones we analyzed. This zip code is in the center of Los Angeles city, it has more than 75% of Latino/Hispanic population, its median age is about 30 and the age 25-44 years young population is over 30%, while the population is increasing over the last a few years. It is worth noting that the average income has also been increasing in the last 7 years.

Discussion:

Our goal is to find a location to open a modern Mexican restaurant and have the best chance to be profitable. In this study I screened for the zip codes with highest Latino/Hispanic population, this will be more likely to ensure there are enough potential customers. By using machine learning I also clustered the zip codes into 10 groups, picked the ones already high in Mexican restaurant, which give us the confidence that the people in the area are used to Mexican food. I also considered age, population and income growth in the areas. Eventually I identified that 90011 is the best area among them. This is because, besides the ethnic composition and local venue types, it is getting more and more young people, their income is increasing as well. The result shows the methods I used served well for my purpose. I consider this approach is helpful in identifying the ideal location for the modern Mexican restaurant, although whether this is a correct model needs to be validated by real life social experiment.

Conclusion:

Los Angeles downtown area, especially 90011 is good for opening a Modern Mexican restaurant. The methodology combination I used here has the potential to be applied to other business types in other cities.