IBM Capstone project

_____

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. The result demonstrates 90011 is the best choice among the ones we analyzed.