Recommending location for a new restaurant in city of Los Angeles

1. Introduction

1.1. Background

Topography: The city of Los Angeles covers a total area of 502.7 square miles (1,302 km2), comprising 468.7 square miles (1,214 km2) of land and 34.0 square miles (88 km2) of water. The city extends for 44 miles (71 km) north-south and for 29 miles (47 km) east-west. The perimeter of the city is 342 miles (550 km). [1]

Demographics: Los Angeles had a population of 3,792,621. The population density was 8,092.3 people per square mile (2,913.0/km2). The age distribution was 874,525 people (23.1%) under 18, 434,478 people (11.5%) from 18 to 24, 1,209,367 people (31.9%) from 25 to 44, 877,555 people (23.1%) from 45 to 64, and 396,696 people (10.5%) who were 65 or older. The median age was 34.1 years. [2]

Economy: The economy of Los Angeles is driven by international trade, entertainment (television, motion pictures, video games, music recording, and production), aerospace, technology, petroleum, fashion, apparel, and tourism. Other significant industries include finance, telecommunications, law, healthcare, and transportation. Los Angeles is the largest manufacturing center in the United States. [3]

Culture: Los Angeles is often billed as the "Creative Capital of the World", because one in every six of its residents works in a creative industry and there are more artists, writers, filmmakers, actors, dancers and musicians living and working in Los Angeles than any other city at any other time in history. There are 841 museums and art galleries in Los Angeles County, more museums per capita than any other city in the U.S. [4]

Los Angeles is a huge city and is densely populated with mostly younger groups of people. With its large area and demographics, ever growing economy with varied industries and flourishing international trade, and creative culture with varied creative industries and museums, makes it a preferred place for any new business to start.

1.2. Business Problem

This project aims at recommending stakeholders an optimum location to set up their new restaurant business. We would consider several factors like location which has younger population groups, more number of venues, distance from the city center before deciding on a location for this new restaurant in this diverse city of Los Angeles.

1.3. Interest

Any stakeholder who is interested in looking for an optimum location for a new restaurant and looking for help to decide on a location which could bring in more business by more number of people visiting the new set up and thus deriving more profits.

2. Data description

2.1. Factors to be considered

Factors that will influence our decision of location of a restaurant would be:

- Population : More number of people, more likely is the footfall to the restaurant. Location which has a considerable population would be considered . Less populated areas would be discarded.
- Venues: Location with more number of venues and less restaurants is assumed to have more footfall for these other venues and thus a restaurant in such an area would also likely have more people visiting it and thus would make it a preferred location for setting up a new restaurant business.
- Distance: Locations which are closer to the city center would be one of the factors which could be considered.
- Age based analysis of the population around the other venues. Youth and younger groups of people are the ones who are more likely to visit the restaurants. Thus location with younger population to be considered.

2.2. Data Sources

Following will be the dataset and data sources used for our analysis and to extract the required information:

- Data of Los Angeles - 2010_Census_Populations_by_Zip_Code.csv [5]
  Population and age based analysis could be done using this data set which has the data relating to these fields for the year 2010.
- Number of restaurants, other venues in the neighborhood, vicinity of the venue, footfall to these venues to decide the popularity of the location will be obtained using the Foursquare API.

3. Methodology

3.1. Exploratory data analysis

3.1.1. Analysis of the Los Angeles population data

The dataframe loaded from the csv file i.e. the 2010 census population data has the following fields - zip code, total population, median age, total males, total household and average household size. There are in total 319 unique zip codes. I have dropped the last 2 columns (total household and average household size) in the data frame as we would not be using those in our analysis.

3.1.2. Analysis of the foursquare results

To retrieve the venue details I used the foursquare API. Used the explore end point feature, to search for the venues within 20000 meters radius. The retrieved results were in json format and few of the required fields were retrieved from it for our analysis and structured it into a pandas dataframe. The following fields for every venue were retrieved from the json file which would facilitate our analysis further - venue name, venue postalCode, venue categories, venue latitude, venue longitude, and venue distance. Total of 99 venues were

3.1.3. Analysis of the merged data

Further, I merged both the dataframes using the common column i.e. Zip Code to get a single dataframe. This helped to get all the information on which the analysis has to be done in a single dataframe and also getting all the features of the particular zip code in a single row, which would ease the process for further analysis.

I further grouped the data by population and sorted it in descending order to filter only the zip codes that have most of the population and discarded the low population areas, which resulted in 91 rows i.e. 91 venues which would be considered for further analysis. There are 57 unique venue categories and 27 unique zip codes at this point of which the location for the restaurant could be chosen.

Data processing in machine learning, we often need to prepare our data in specific ways before feeding into a machine learning model. One of the major problems with machine learning is that a lot of algorithms cannot work directly with categorical data. One such methodology to convert the categorical data to binary vector representation. One such methodology is One-Hot encoding on categorical data. [6]. In our analysis, we used the get_dummies function in

pandas library to perform the one hot encoding on category column. The data frame created after one hot encoding has a column for every category showing the binary value. Next, we grouped rows by zip code and by taking the mean of the frequency of occurrence of each category. This would be the input for a machine learning algorithm we are going to use. We would be more interested in the classification algorithm as we are looking at different factors which would influence or decision. Hence a classification model helps us categorise our data (zip codes) based on these different factors. We could then choose one that is best amongst these categories.

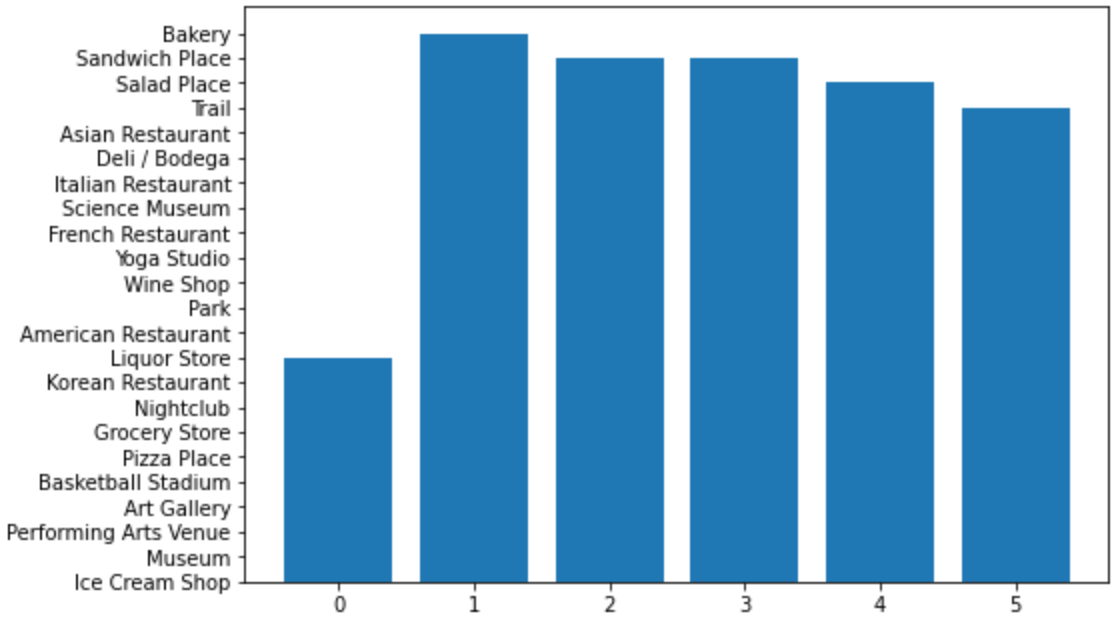
3.2. Machine Learning Algorithm

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training [7]. In this project we are using the K-Means clustering algorithm to cluster the venues into 6 clusters.

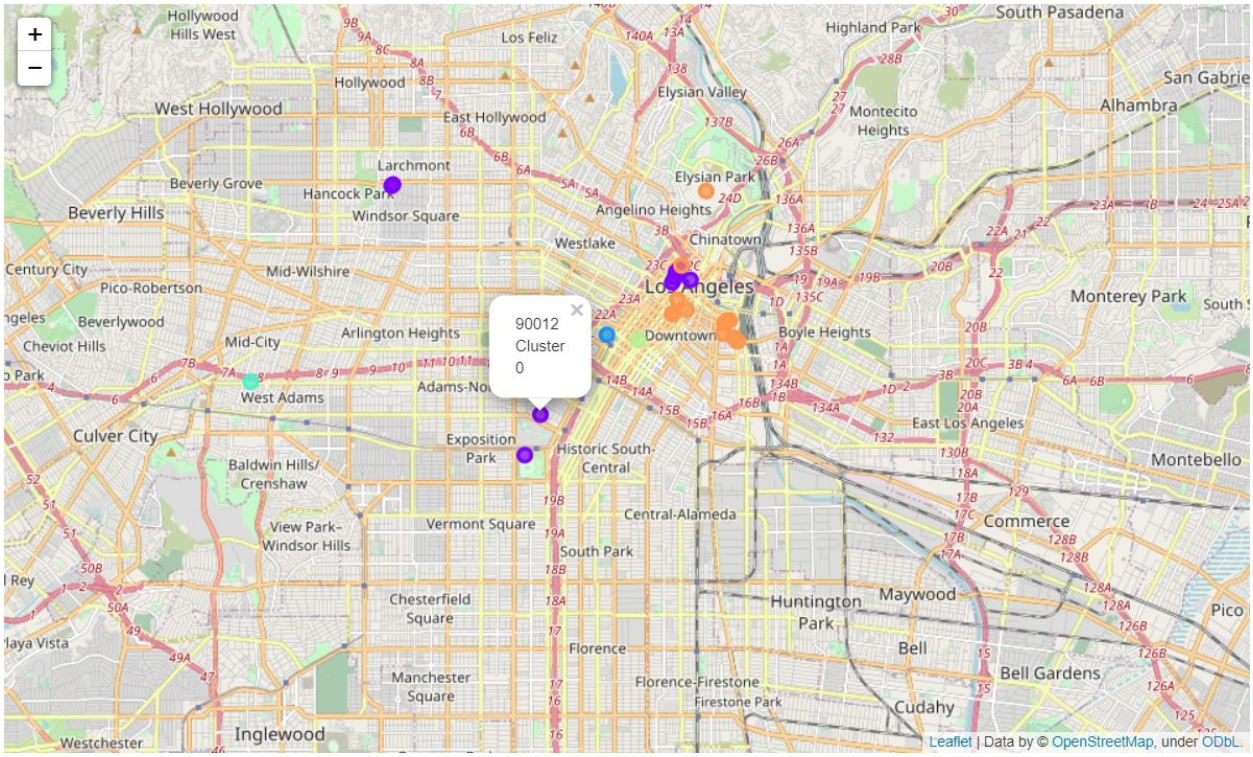
4. Results

The 6 cluster labels k-means algorithm generates for every row in the dataframe are as follows: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 3, 1, 2]). We named them Cluster 1, Cluster 2, Cluster 3, Cluster 4, Cluster 5 and Cluster 6. The cluster labels, latitude and longitude are assigned to the dataframe row. These clusters are then mapped into a map using the visualization library ‘folium’. The number of samples in each cluster varies with one cluster having more number of restaurants than others. We considered the cluster with less number of restaurants but more number of other venues.

The bar plot below shows the first most common venue for every cluster and the map shows different cluster with different color circles.



We can see the 6 clusters categories formed based on the K-means algorithm which are shown in different colours in the map below.



We examined all the clusters categorized by the kmeans algorithm and label each cluster

Cluster 1: Arts and physical activities with a very few eateries mostly fast food and icecream shops
Cluster 2: Bakery only
Cluster 3: Sandwich place
Cluster 4: Sandwich place
Cluster 5: Mostly restaurants of different cuisines and some fast food restaurants and coffee shops
Cluster 6: Trail only

One of the few locations in Cluster 1 or Cluster 6 could be considered for our new restaurant. Based on the analysis done so far, these two clusters have a minimum number of restaurants and also cluster 1 has other venues visited frequently by people. Further analysis could be done based on the distance and population age.

Cluster 1

```
c1 = zip_venues_sorted.loc[zip_venues_sorted['Cluster Labels'] == 0, zip_venues_sorted.columns[[1] + list(range(2, zip_venues
c1
```

	Longitude	Latitude	Zip Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	-118.323704	34.075327	90004	Ice Cream Shop	Sandwich Place	Yoga Studio	Concert Hall	Home Service	Gym	Grocery Store	German Restaurant	Fried Chicken Joint	French Restaurant
1	-118.323535	34.075836	90007	Museum	College Residence Hall	Yoga Studio	Deli / Bodega	Ice Cream Shop	Home Service	Gym	Grocery Store	German Restaurant	Fried Chicken Joint
2	-118.284652	34.025655	90012	Performing Arts Venue	Art Museum	Music Venue	Concert Hall	Baseball Stadium	Theater	Park	Gym	Grocery Store	German Restaurant
3	-118.288829	34.016829	90013	Art Gallery	Bookstore	Ice Cream Shop	Home Service	Market	Mediterranean Restaurant	German Restaurant	Coffee Shop	Yoga Studio	Farmers Market
4	0.000000	0.000000	90015	Basketball Stadium	Fried Chicken Joint	Yoga Studio	Deli / Bodega	Indie Movie Theater	Ice Cream Shop	Home Service	Gym	Grocery Store	German Restaurant
5	-118.245179	34.055034	90016	Pizza Place	Italian Restaurant	Ice Cream Shop	Home Service	Gym	Grocery Store	German Restaurant	Fried Chicken Joint	French Restaurant	Food Truck
6	-118.248886	34.056625	90017	Grocery Store	Yoga Studio	Wine Shop	Indie Movie Theater	Ice Cream Shop	Home Service	Gym	German Restaurant	Fried Chicken Joint	French Restaurant
7	-118.249284	34.055511	90019	Nightclub	Yoga Studio	Italian Restaurant	Ice Cream Shop	Home Service	Gym	Grocery Store	German Restaurant	Fried Chicken Joint	French Restaurant
8	-118.250051	34.054474	90020	Korean Restaurant	Italian Restaurant	Indie Movie Theater	Ice Cream Shop	Home Service	Gym	Grocery Store	German Restaurant	Fried Chicken Joint	French Restaurant
9	-118.248354	34.057133	90023	Liquor Store	Yoga Studio	Italian Restaurant	Ice Cream Shop	Home Service	Gym	Grocery Store	German Restaurant	Fried Chicken Joint	French Restaurant

Cluster 6

Cluster 6

```
c6 = zip_venues_sorted.loc[zip_venues_sorted['Cluster Labels'] == 5, zip_venues_sorted.columns[[1] + list(range(2, zip_venues
c6
```

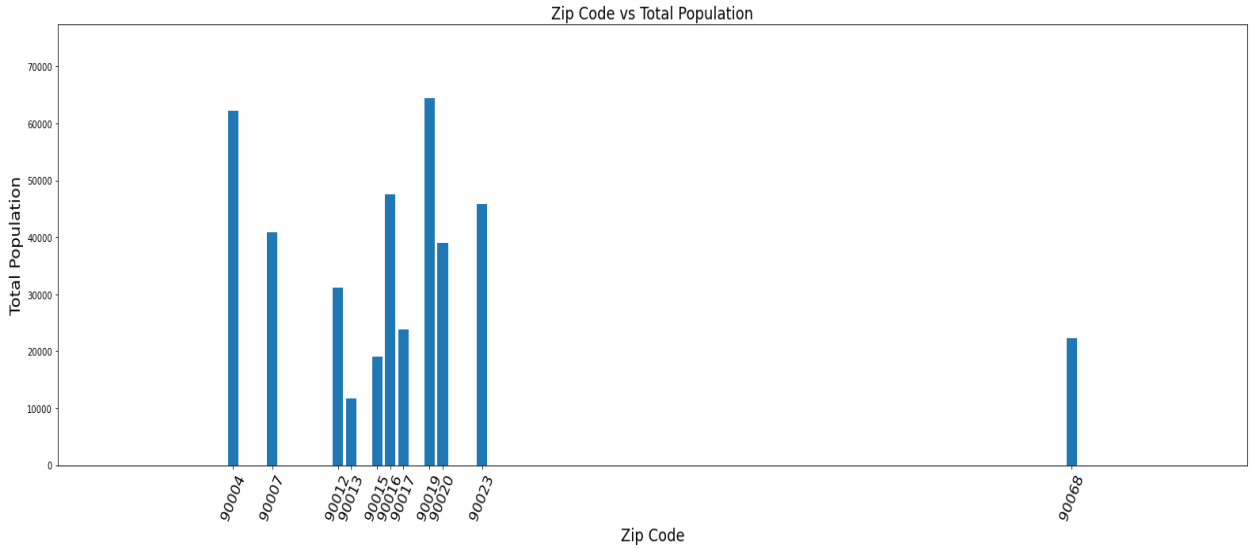
31]:

	Longitude	Latitude	Zip Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
23	0.0	0.0	90068	Trail	Scenic Lookout	Music Venue	Yoga Studio	Concert Hall	Home Service	Gym	Grocery Store	German Restaurant	Fried Chicken Joint

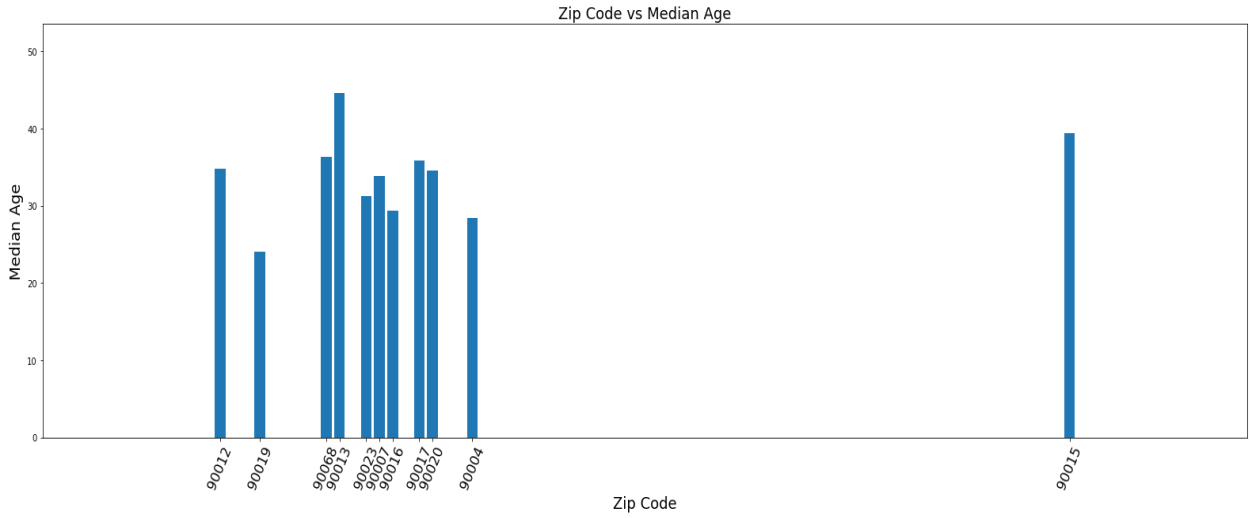
These clusters are merged into one dataframe which gives us 11 zip codes to choose from after further analysis which would be based on location having larger population, younger population, distance from the city center and total number of venues other than the restaurants in the respective zip codes.

5. Discussion

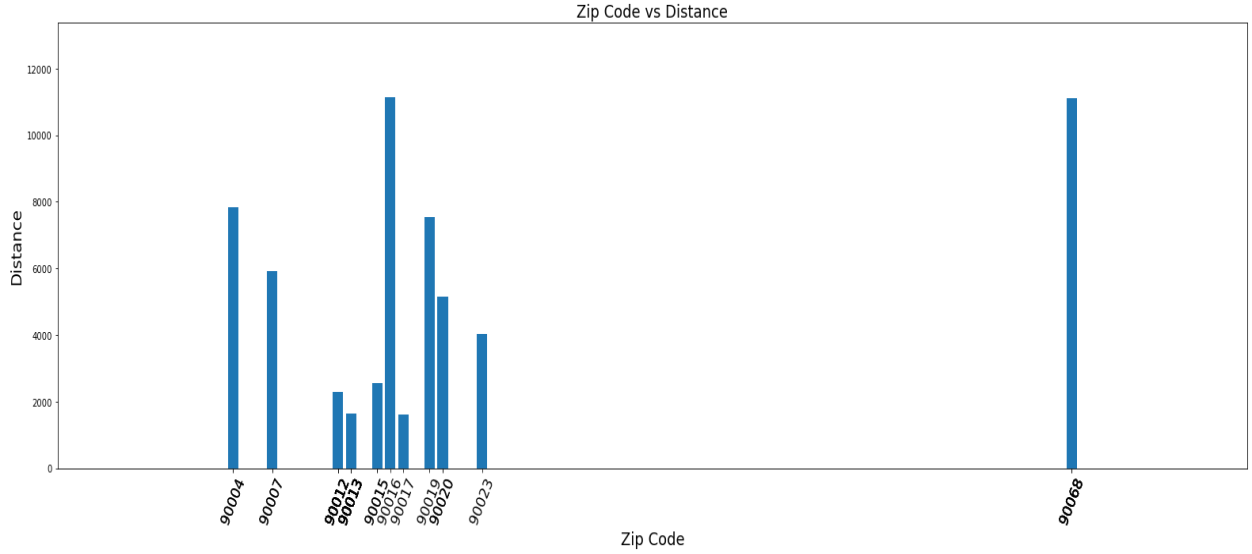
We are sorting the data frame by the total population and the bar plot shows the total population per the zip code selected to see which zip code has a higher population, which could be a factor to decide the location of the restaurant. The plot shows that 90019, 90004, 90016, 90023, 90007 (in descending order) are zip codes which are more populous compared to the other zip codes in the cluster chosen.



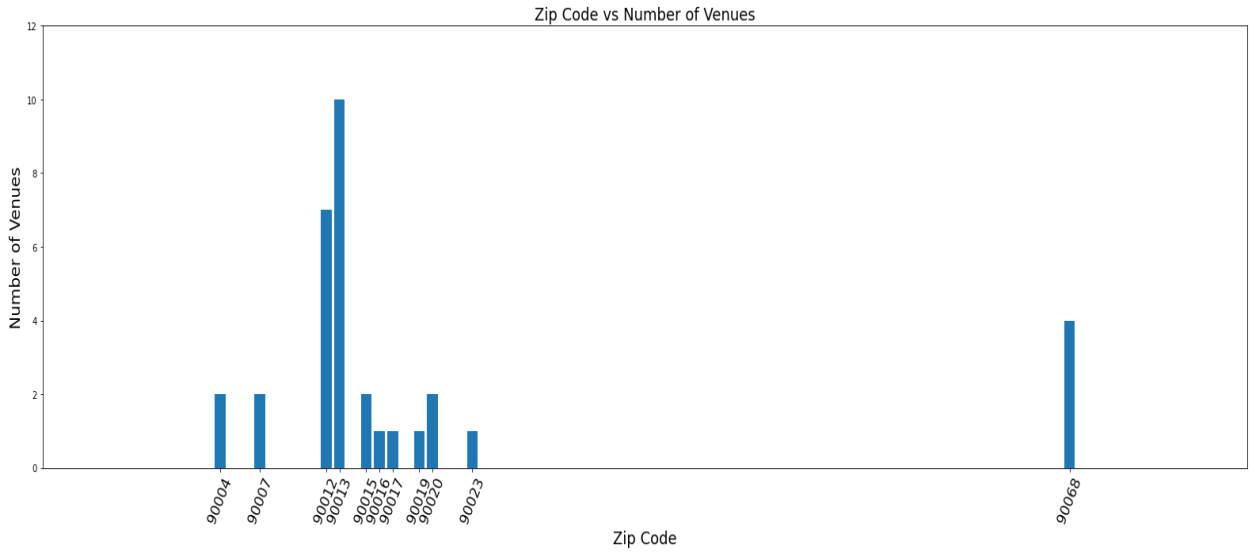
We next plotted a bar chart to show the median age of the population in the clusters chosen. It shows that 90007, 90023, 90017, 90015, 90016 are the zip codes with lesser median age (in ascending order), which means that these are zip codes with younger population compared to all other zip codes in the chosen cluster



Next, sort the dataframe by distance from the city center to see which zip code is closer to the city center, which could be a factor to decide the location of the restaurant. Closer to city center, more footfall. Bar plot for distance from city center shows that 90012, 90013, 90017, 90015, 90023 (in ascending order) are closer to the city center compared to all the other zip codes in the chosen cluster.



Bar Plot to show the number of venues in each zip code in the clusters chosen. It shows that 90013, 90012, 90068, 90004, 90007 (in descending order) have more venues in the respective area compared to all the other zip codes in the chosen cluster



6. Conclusion

We could draw the following conclusions based on the above analysis made on the chosen cluster:

1) Bar plot for population shows that 90019, 90004, 90016, 90023, 90007 (in descending order) are zip codes which are more populous compared to the other zip codes in the cluster chosen.

2) Bar plot for median age shows that 90007, 90023, 90017, 90015, 90016 are the zip codes with lesser median age (in ascending order), which means that these are zip codes with younger population compared to all other zip codes in the chosen cluster

3) Bar plot for distance from city center shows that 90012, 90013, 90017, 90015, 90023 (in ascending order) are closer to the city center compared to all the other zip codes in the chosen cluster.

4) Bar plot for the number of venues shows that 90013, 90012, 90068, 90004, 90007 (in descending order) have more venues in the respective area compared to all the other zip codes in the chosen cluster

Based on these four analysis made above, the stakeholder take one or few or all of the results into consideration while chosing a location for his restaurant.

If we were to take distance and number of venues into consideration, few of the zip codes we could recommend are 90012, 90013.

If we were to take population and median age into consideration, few of the zip codes we could recommend are 90007, 90023, 90016

If we were to take median age and distance into consideration, few of the zip codes we could recommend are 90017, 90015, 90023.

7. References:

1. Topography : Los Angeles - Wikipedia
2. Demographies : Los Angeles - Wikipedia
3. Economy : Los Angeles - Wikipedia
4. Culture : Los Angeles - Wikipedia
5. Data for city of Los Angeles population: 2010_Census_Populations_by_Zip_Code.csv https://catalog.data.gov/dataset/2010-census-populations-by-zip-code/resource/2f420e98-e3f8-4777-9a83-ce1fdd00e7b4
6. One hot encoding using pandas get_dummies function https://towardsdatascience.com/what-is-one-hot-encoding-and-how-to-use-pandas-get-dummies-function-922eb9bd4970
7. K-means clustering algorithm - K-Means Clustering Algorithm - Javatpoint