**Fast Food Restaurant**

Kavin Shah

July 2020

**Introduction**

**Description of the Problem and Context:**

A group of investors started a small Fast food restaurant chain in Alabama. One of the decisions made during last business strategy company meeting was to confirm a plan to establish and deploy new stores on the west coast via launching stores in Los Angeles. As a consequence, a first pilot project is launched, and Raymond, project leader was nominated with first objective to launch 10 new stores in Los Angeles, CA. The success of this initiative is crucial for the next phases of the project and for the development of additional stores. Among Raymond’s key tasks, he needs to work with a real estate consultant to identify available venues in Los Angeles and close relevant deals as soon as possible to deploy the stores. After a first discussion with Raymond’s preferred real estate broker in LA, he realized that the criteria he defined for finding the stores are not accurate enough and his broker shared additional information. Here are a few problems shared: A) He may spend too much time finding the right places in such a large city like LA, so he needs to identify some preferred zones within Los Angeles to focus his search. B) He may not be the best person to define priorities and most relevant areas for the Fast food restaurant chain criteria, so he needs more views with targeted areas. As a consequence, Raymond contacted me and asked for some help on where I would recommend they should open the first stores in LA. Raymond and I discussed and we came to a conclusion that the problem could be solved with defining a list of preferred areas in Los Angeles issued from classifying neighborhoods based on exploring existing venues and most frequent categories of venues in each candidate zone. This way, we can identify similar neighborhoods, gather them within several clusters and choose the right cluster of areas within Los Angeles to focus on. Such output will serve as a view for Raymond and a list of target zones for Raymond’s real estate broker.

We will use the Foursquare location data to explore neighborhoods of LA, specifically categories of venues, in a similar way we did with Toronto area in the previous lab and assignment. We plan to use unsupervised machine learning method for classification, like k-means algorithm. This topic will be developed further in the next Methodology section of this report. So, we focus for now on unsupervised clustering method based on venues around candidate locations, and within each zip code in Los Angeles, CA. Additional tools for solving the problem: In addition to Foursquare API we already mentioned, we will use Jupyter notebook for all the coding and explanations of our method, process and computations. Coding will be done in python 3, and leveraging usual libraries: NumPy and SciPy for scientific computing, Pandas for data extracting, cleaning and analysis, Matplotlib and Folium for figures, plots, maps and visualization, Scikit-learn for machine learning, in particular we plan to use clustering kmeans algorithm.

In this report, we mainly present explanations, notes and comments as described above, but we don't insert actual code. Another note book is provided as part of the assignment and gathers all the python code, dataframes, plots and maps involved in the capstone project. The objective is to build clusters to partitions Los Angeles, CA in similar areas and identify the most suitable areas for launching new stores for Fast food restaurants. Once clusters are created, we will review the clusters and identify similarities within a given cluster and similarities between two different clusters.

**Context**

This is a list of over 10,000 fast food restaurants provided by [Datafiniti's Business Database](https://datafiniti.co/products/business-data/" \t "_blank). The dataset includes the restaurant's address, city, latitude and longitude coordinates, name, and more.

**Dataset acquisition and Cleaning.**

Data was acquired from [Datafiniti's Business Database](https://datafiniti.co/products/business-data/). Apart from that I also took dataset for state codes and the population dataset of the states to get further insights in the relation of location and the restaurant.

From the perspective of Null values there weren’t any row with null values. The main problem the dataset had was mention of same restaurant in different forms. For ex. Mc Donalds was also written had mcdonalds or McDonald’s. So, I used NLTK's edit distance function to find the edit distance between those names Edit Distance: The distance between the source string and the target string is the minimum number of edit operations (deletions, insertions, or substitutions) required to transform the source into the target. The lower the distance, the more similar the two strings.

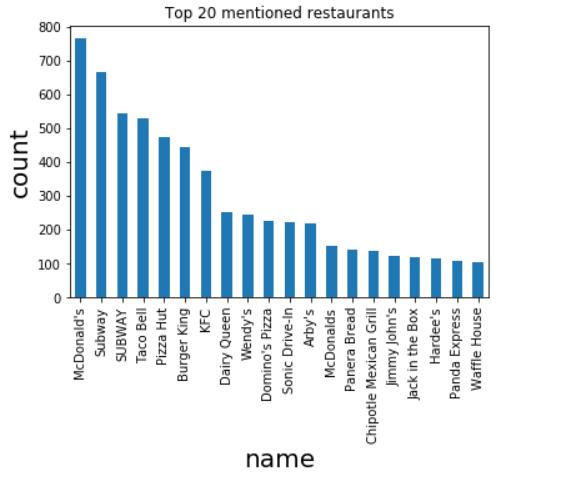
This helped correct one big analysis. Before McD was the largest food chain, after applying nltk Subway became the largest chain.

**Feature Selection**

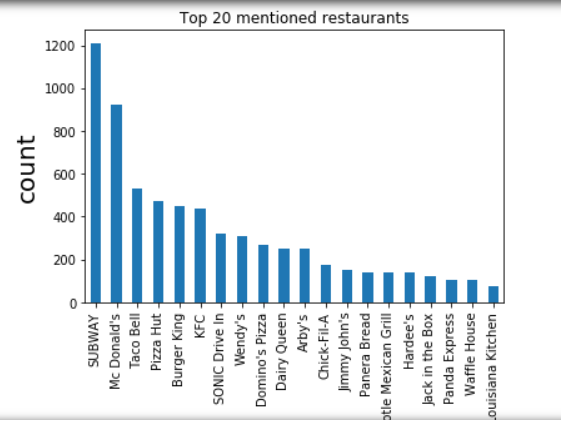
After cleaning the shape of the dataset was 10000 by 16. Upon further examining, it was clear that some of the features were not needed such as id, keys, sourceURLs, and websites which are unique identifier and would not be required for visualization or ML purposes.

**EDA**

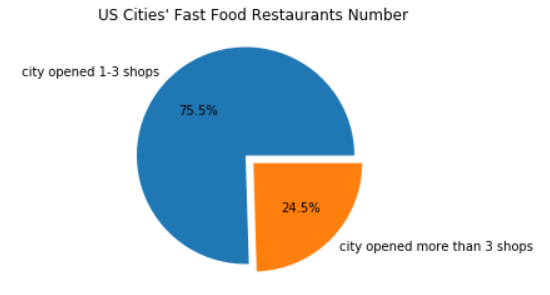
top 20 restaurants recorded by count total.



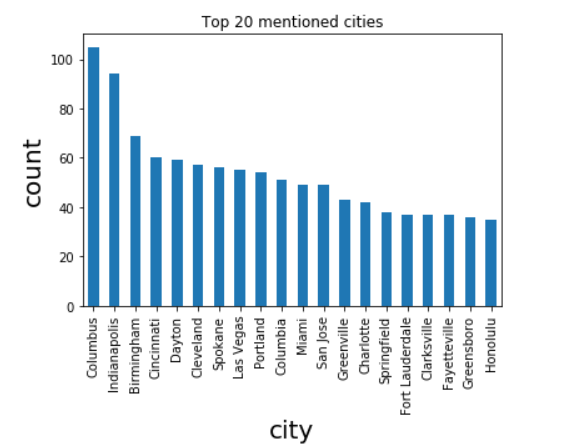
Corrected Top 20



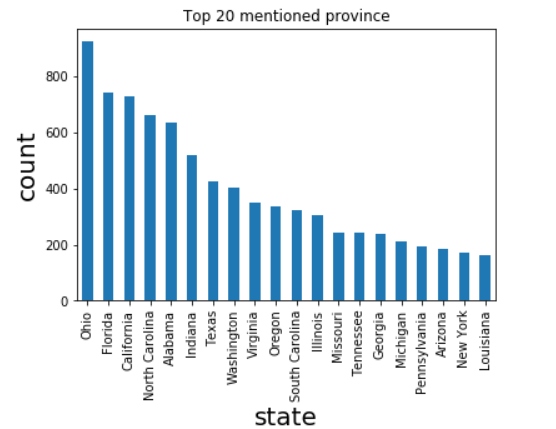
Here, we can see the % for restaurants where branches of restaurant is between 1-3 and more than 3.

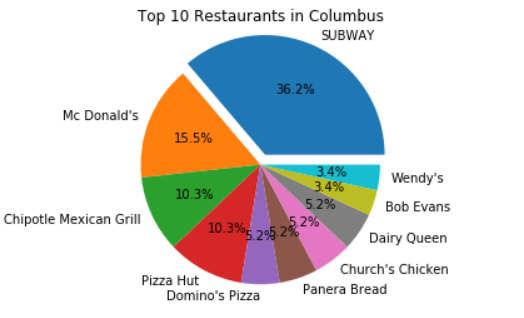


top 20 fast food populated cities recorded by count total. Cities with the most number of fast-food restaurants.

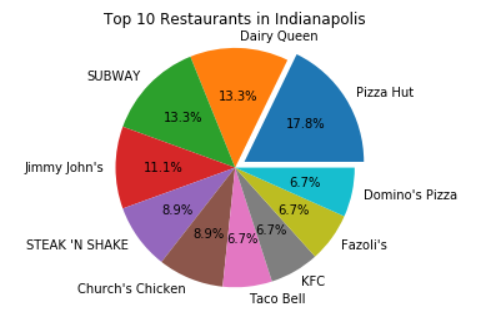


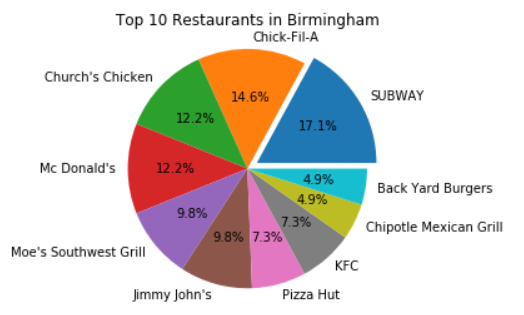
Top 20 fast food populated states recorded by count total.



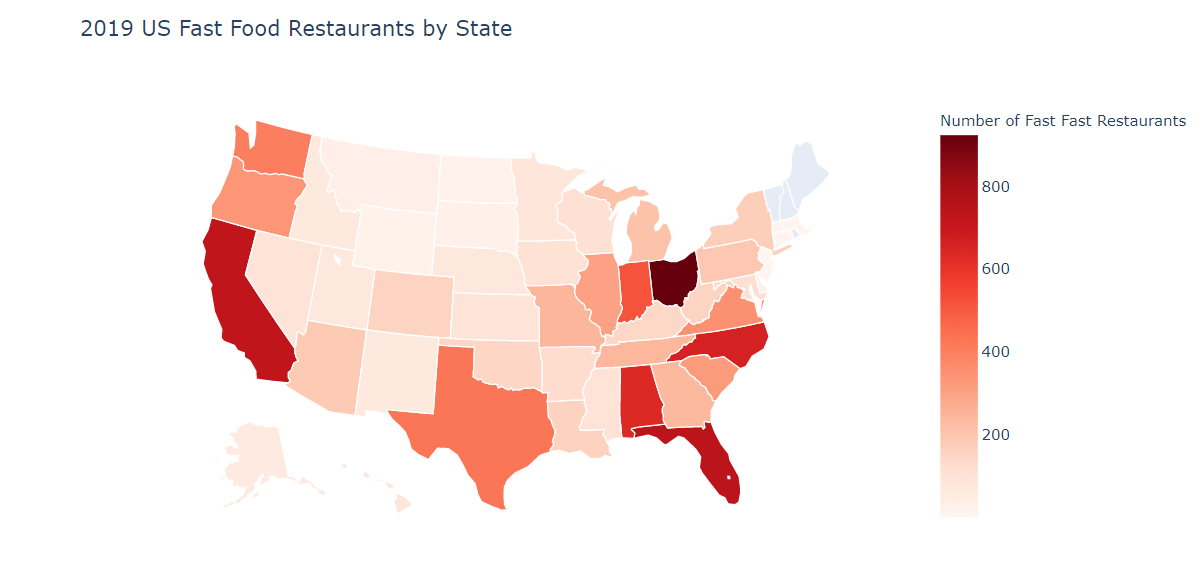


This is the breakdown of fast food restaurants for the 3 top cities with the highest number of fast food restaurants.

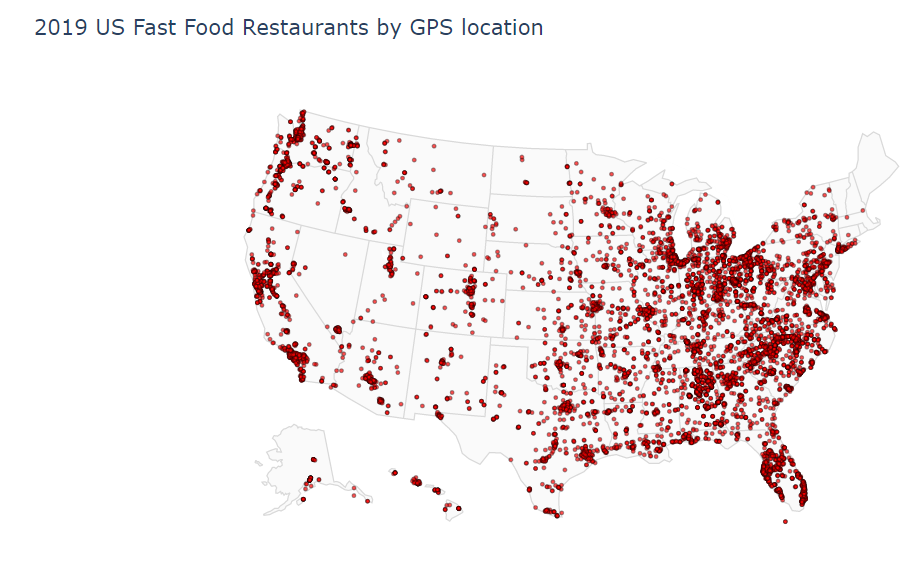




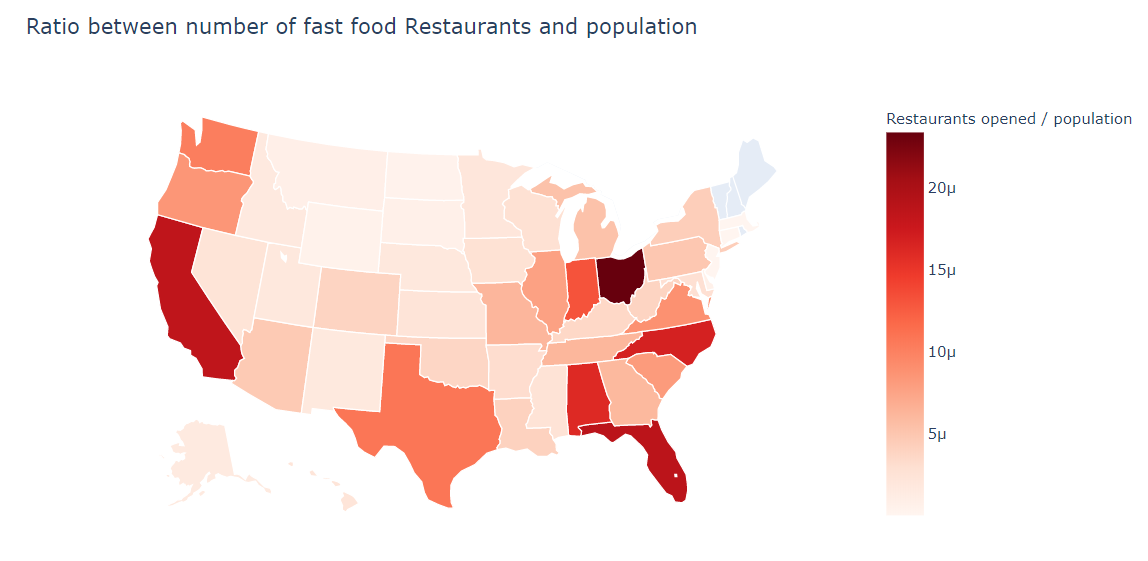
We can see that CA, OH, FL have darker red color indicating more restaurants.



Here, we see that Coastal parts of the country have more fast food restaurants than middle part.



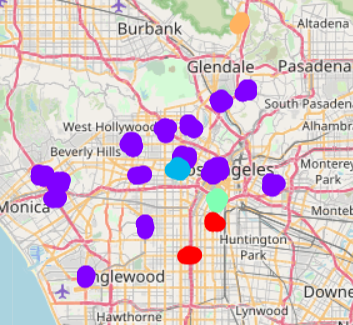
Keeping in mind the above map, we can see here that higher the population, more the restaurants, which makes sense.



**Machine Learning.**

**Clustering.**

Here, we see there are different clusters forming. We had taken the number of clusters as 5 (k=5). Most of the clusters belong to cluster 1(purple). Then next is cluster 0 (red). Cluster 2(sky blue), Cluster 3(neon green) and Cluster 4(Yellow) has only 1 Cluster each. To understand the clusters better we have also taken the top 3 most common venue returned by Foursquare to get a better idea of the neighborhood.

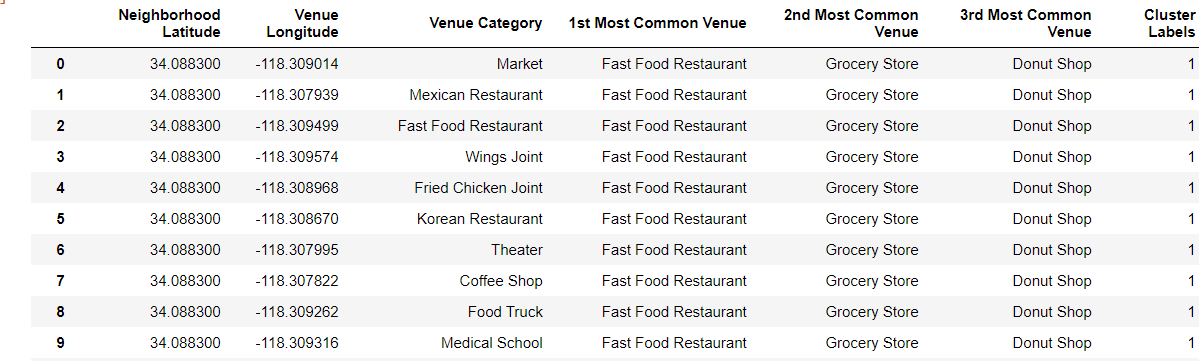


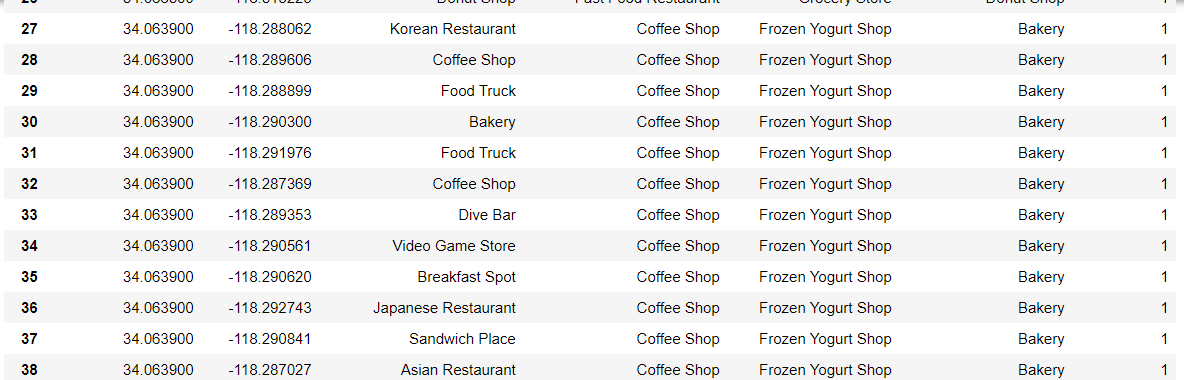
**Results**

This is a snippet of cluster 0(red). Here, we see that in this neighborhood fast food restaurants are the most visited venue along with Pizza place and Chinese Restaurant. This neighborhood is definitely a place that can be considered for opening the restaurant.



This is a snippet of cluster 1(purple). There are many Clusters for Cluster 1. We have taken two places categorized as Cluster1. Here, we see that in one place fast food restaurants are the most visited venue and in other Coffee Shop is the most visited venue but knowing that most of the clusters are classified as Cluster 1, we can assume that there is high foot-fall in Cluster 1 places. Even this Cluster can be a suitable neighborhood to open a fast food restaurant.

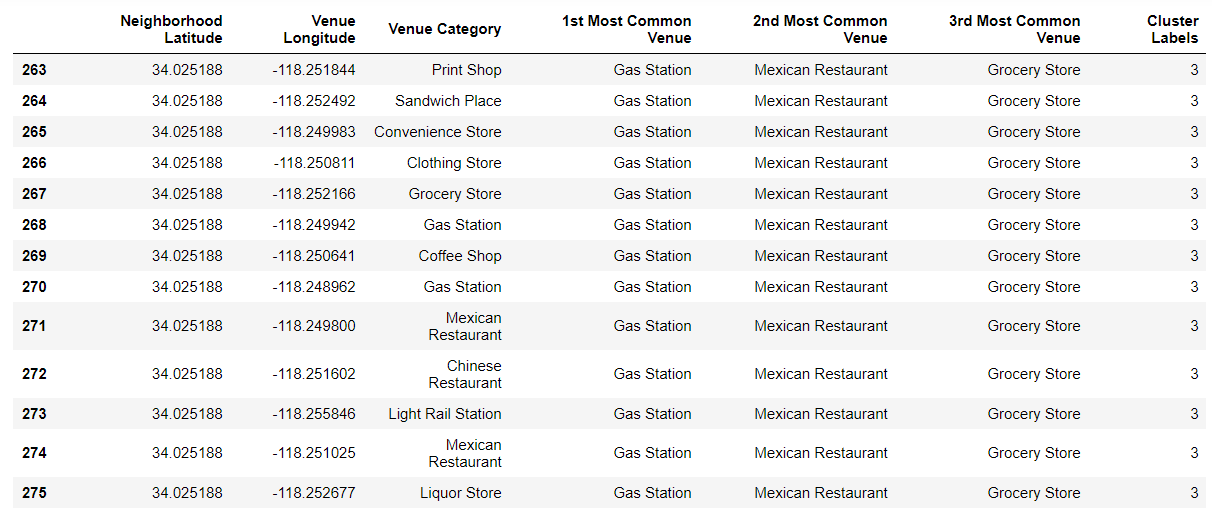




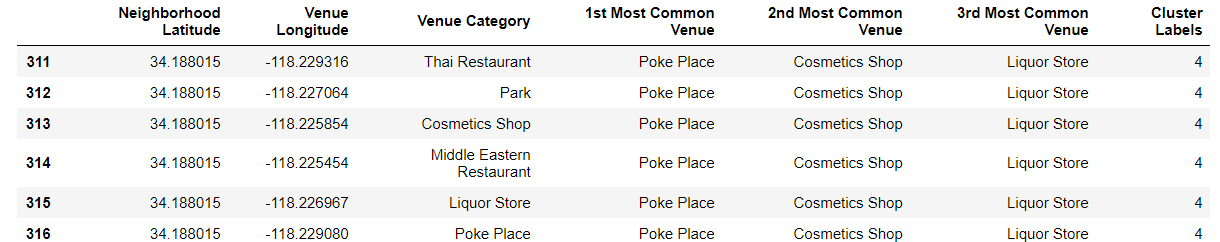
This is the snippet of Cluster 2(sky blue). Here, we see that a very particular Restaurant is the most visited and the fact that the 3rd most visited place is the Hotel, we can say that the foot-fall might be less or there are few restaurants in the vicinity surrounding that hotel and area.



This is a snippet of Cluster 3(neon green). This Cluster doesn’t seem to have the attraction for the fast food restaurants and wouldn’t be a recommended neighborhood.



This is a snippet of Cluster 4(yellow). Like, Cluster 3 even this neighborhood does not seem to be a good place for opening a fast food restaurant.



**Discussion**

As observations noted from the map in the Results section, most of the fast food restaurants are in the Clusters 0 and 1 and moderate number in Cluster 2. On the other hand, Clusters 3 and 4 has very low number of fast food restaurants in the neighborhoods. This can be seen in two ways, firstly, as a great opportunity and high potential areas to open new fast food restaurants as there is very little to no competition. Meanwhile, on the flip side there is a chance of missing on high footfall as suggested in Clusters 0 and 1.

**Limitations and Suggestions**

For Future Research In this project, we only consider one factor i.e. frequency of occurrence of fast food restaurants, there are other factors such as population and income of neighborhoods that could influence the location decision of a new fast food restaurant. However, to the best knowledge of this researcher such data are not available to the neighborhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new fast food restaurant.

**Conclusion**

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. Raymond regarding the best locations to open a new fast food restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 and 1 are the most preferred locations to open a new fast food restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new fast food restaurant.