The Battle of Neighborhoods

1. Introduction

People search for rental listing of any city would be highly interested in the localities with restaurants, cafes, bars, Grocery stores,etc. People might want to know what kind of venuses are nearby a particular rental listing, so they can make informed decisions in selecting the rental place based on the venuses.The information of nearby venuses of various rental listing in the city and their localities in form of map would help people decide which rental place to choose among the many in the city and also which locality to prefer. Also combining the frequency of the venus nearby a rental listing would help people make easy decisions about selecting the rental place. A map of all the rental listing with specific color attributes will be plotted to highlight the rental listing based on nearby venuses. We will classify the various rental lists into different clusters using a Machine Learning Algorithm, the K-means clustering Algorithm. This enables any user to take a quick glance and decide which place to select for renting out.

2. Data

For this project the Foursquare API will be used. A list of available rental place will be obtained with their respective location in longitude and latitude coordinates.

The source of the rental data is Craigslist for rental ad. We will be importing CraigslistHousing package(package designed to get rental ads directly from the Craigslist site) from craigslist library to obtain the relevant rental data. For now we will only be obtaining first one hundred rental posts in the area of 'San Francisco' using the get_result method of CraigslistHousing object.

A Foursquare API GET request is sent in order to acquire the venues that are within a radius of 500m of each rental post. The data is formatted using one hot encoding with the categories of each venue. Then, the venues are grouped by neighborhoods computing the mean of each feature.

The clustering of rental place will be done based on the frequency of venus near a rental place. This would help use to make a final decision of selecting a particular rental place over the other.

3. Feature Extraction

For feature extraction One Hot Encoding was used in terms of categories. Therefore, each feature is a category that belongs to a venue. Each feature becomes binary, this means that 1 means this category is found in the venue and 0 means the opposite. Then, all the venues are grouped by the neighborhoods, computing at the same time the mean. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.

3.2. Unsupervised Learning

For the purpose of doing unsupervised learning to find similarities between neighborhoods, a clustering algorithm is implemented. In this case K-Means is used due to its simplicity and its similarity approach to find patterns.

- K-Means:

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the multi dimensional features.

In the following figure there is a graphical example of how a K-Means algorithm works. As it is possible to see, dispersion is minimized by representing all clustered data into one group or cluster.
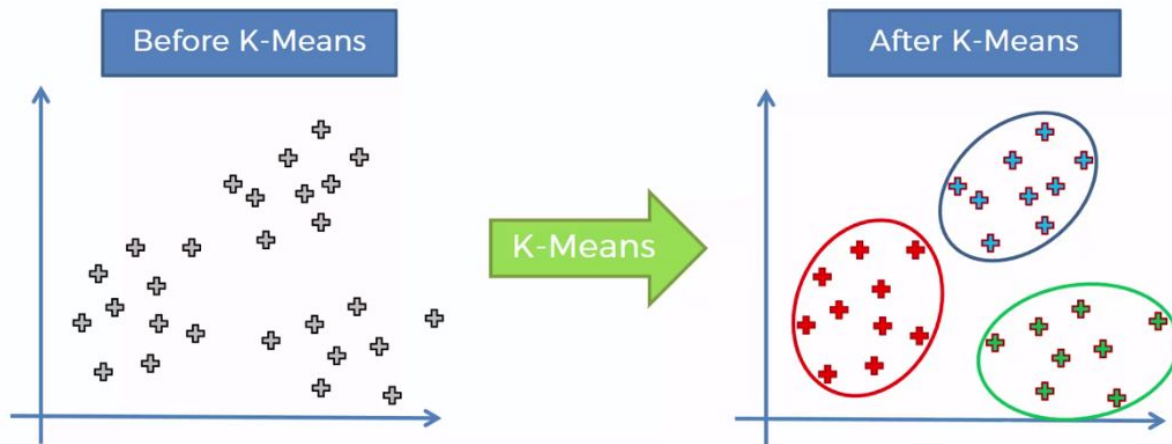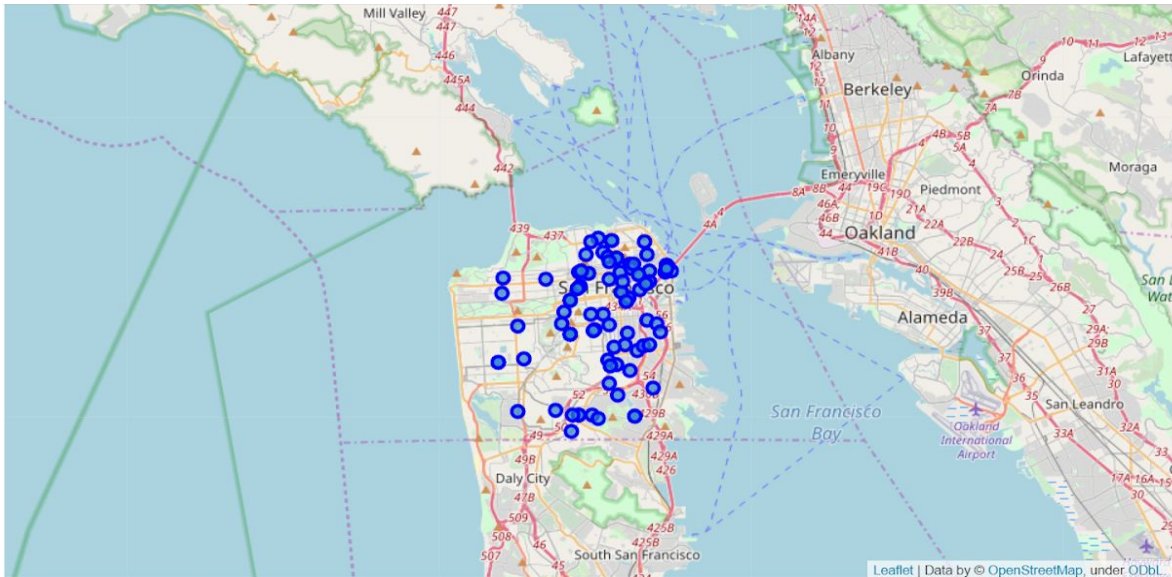

It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm. For this reason, the elbow method is implemented. A chart that compares error vs number of clusters is done and we selected the number of clusters using the elbow method. Then, further analysis of each cluster is done.
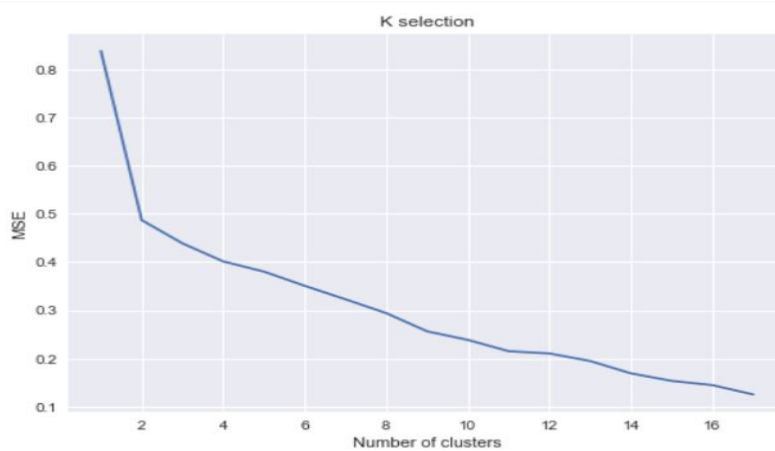
The Battle of Neighborhoods

4. Results

Firstly, data is plotted in a geographical map to get a notion of the rental place location. The following image shows each rental place in San Francisco.


Secondly, the cluster algorithm is implemented. For this purpose, it is necessary to have a prior idea about the number of clusters. Therefore, the mean squared error (MSE) is plotted vs the number of clusters. The number of clusters start with a value of 1 increasing until a value of 15. This chart is shown in the image below.

The Battle of Neighborhoods

As it is expected, the MSE decreases over the number of clusters. The elbow method is implemented in order to select the appropriate number of groups. In this case, it is possible to see that the elbow is found more or less around 5. The MSE found below this number shows little changes rather than big ones. Finally, once the number of clusters is fixed, the clustering algorithm is repeated through samples and each neighborhood is labeled according to the clusters found.

For visualization purposes, the geographical data is again plotted but with different colors. Each color represents the cluster for which each rental place belongs. This image is shown below.
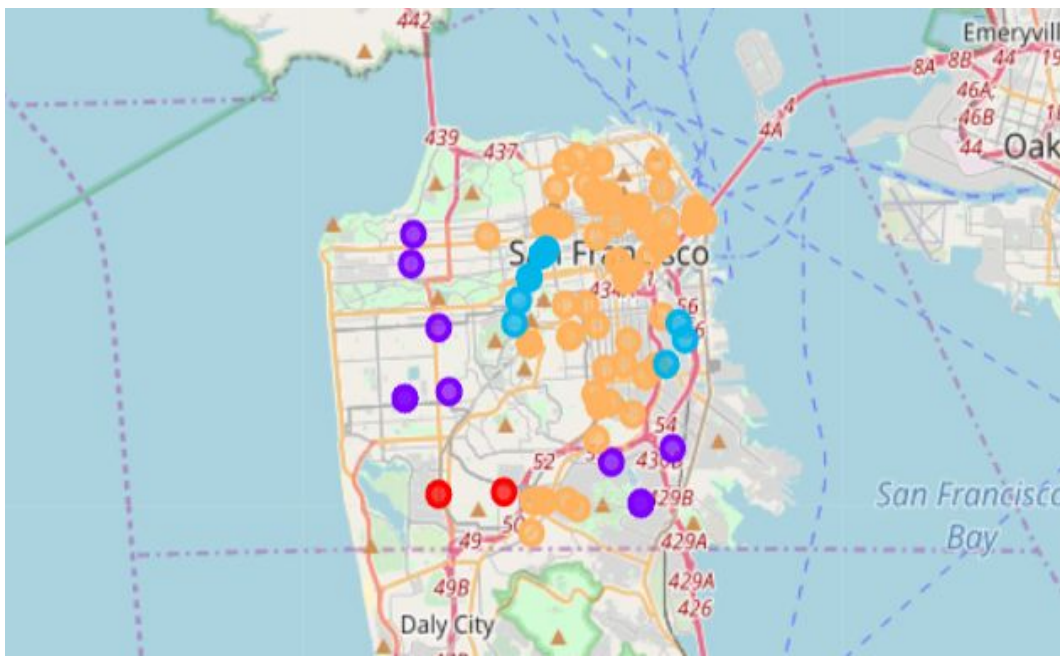
In this image it is evident that cluster algorithm is not segmenting the rental place based on the geolocation, but it is clustered based on the venues nearby each rental place. This cluster gives a fare idea to the user about the venus nearby for all the rental listing. Using such information, users can make a better decision in finalizing a rental place.

The Battle of Neighborhoods

For research purposes, bar charts are employed to find insights within the clusters. The bar chart that is shown below shows the features with higher frequency in the centroids found by the algorithm. In this way we can learn what the algorithm found.
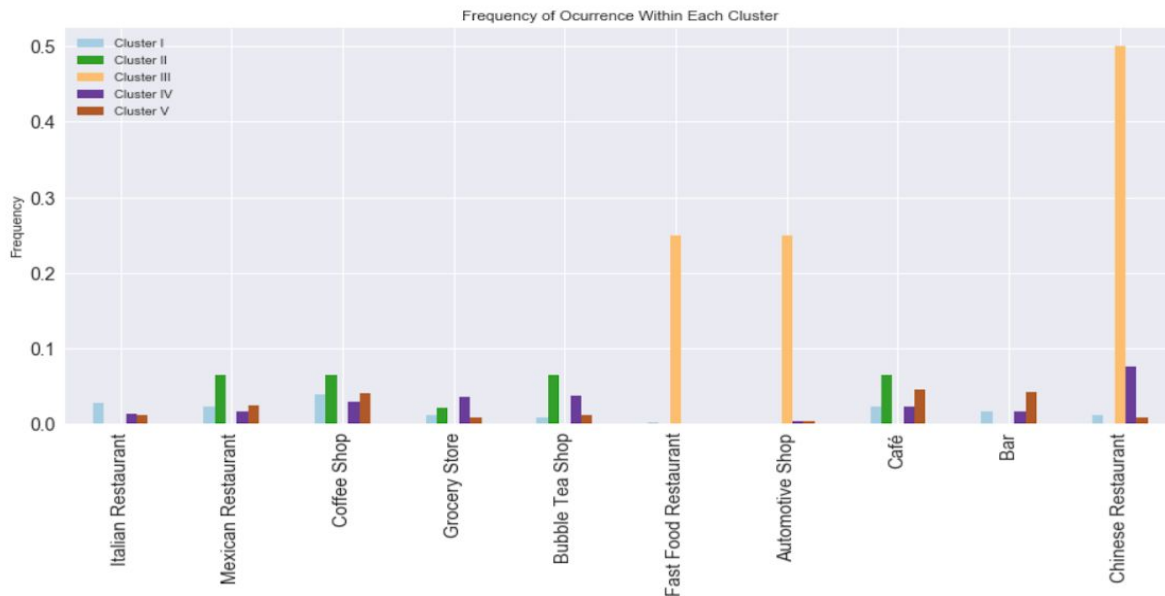
It is possible to see that:

Cluster I focuses on rental place that has venues like **Coffee Shop, Italian Restaurant, Mexican Restaurant, Bar, Grocery Store and Chinese Restaurant**.

Cluster II focuses on rental place that has venuses like **Mexican Restaurant, Coffee Shop, Bubble Tea Shop, Cafe and Grocery Store**.

Cluster III focuses on rental place that has venuses like **Chinese Restaurant, Fast Food Restaurant and Automotive Shop**.

Cluster IV focuses on rental place that has venuses like **Chinese Restaurant, Bubble Tea Shop, Grocery Store, Coffee Shop, Bar, Italian Restaurant and Mexican Restaurant**.

Cluster V focuses on rental place that has venuses like **Coffee Shop, Cafe, Bar, Italian Restaurant, Mexican Restaurant and Chinese Restaurant.**

The Battle of Neighborhoods

5. Discussion

It is worth to note that this work is useful only for those who are looking for rental place in 'San Francisco'. The reason is because there is a limited amount of data we can request using the Foursquare API. Consequently, it will have a greater cost than the Lite version. The clustering algorithm was applied only on the venuses, but we could include price factor as well for further development.

6. Conclusion

In this work a clustering of rental place based on the venuses nearby each is done. This clustering involves the rental place list of San Francisco which was obtained using CraigslistHousing package and venues around the rental places was acquired using the Foursquare API. One Hot Encoding was used to convert the categories of the venues into a feature matrix. Then, all venues are grouped by rental place and at the same time the mean is calculated. Hence, the resulting features used are the frequency of occurrence from each category in a rental place.

The K-Means clustering algorithm was used for finding similarity between all the rental places in the feature matrix. The elbow method was used for selecting the appropriate number of clusters. Hence, the K-cluster selected was 5. Results shown are as follows:

Cluster

- I: Rental place that has venuses like Coffee Shop, Italian Restaurant, Mexican Restaurant, Bar, Grocery Store and Chinese Restaurant.
- II: Rental place that has venuses like Mexican Restaurant, Coffee Shop, Bubble Tea Shop, Cafe and Grocery Store.
- III: Rental place that has venuses like Chinese Restaurant, Fast Food Restaurant and Automotive Shop.
- IV: Rental place that has venuses like Chinese Restaurant, Bubble Tea Shop, Grocery Store, Coffee Shop, Bar, Italian Restaurant and Mexican Restaurant.
- V: Rental place that has venuses like Coffee Shop, Bar, Italian Restaurant, Mexican Restaurant and Chinese Restaurant.

Finally, any user who wants to rent out a place in San Francisco can use this system to get a notion or idea about what kind of venus are nearby each rental place.