Project report.

Business Problem

The problem will be to find a solution for a client who is searching for a place to start a Chinese Restaurant. The client wants to start the restaurant in Chennai in Tamil Nadu, India. He is looking to start a big Chinese restaurant, so he wants to build it in a main area in the city where there is more footfall. He also doesn’t want a place where there are many already an established Chinese restaurant as it is difficult to pull the customers in such an area. At the same time it should be near to places where it is often crowded such as malls, markets, multiplexes, train and bus stations etc. So the main goal of the project is to find an area in Chennai which has most of the crowded places and also the least number of Chinese type restaurants.

Data section

To help complete this project the location data of Chennai city is needed. The latitudes and longitudes of all the places in Chennai is needed. The latitudes and longitudes of each place in Chennai can be taken from the python pgeocode library. This library returns the location details if given the pincodes of the areas. The pincodes of Chennai city are available in many websites. In this project the pincodes are taken from the following website url by web scraping.
https://www.mapsofindia.com/pincode/india/tamil-nadu/chennai/

After getting the location details, the venues and places in those locations are obtained using the Foursquare API. A Foursquare account is created, the access token and credentials are also obtained which will be used in the project to make requests to the API.

Methodology

Data Analyzing

The data with the pincodes of Chennai city has been taken from this site (https://www.mapsofindia.com/pincode/india/tamil-nadu/chennai/).

Missing data

The data had some missing values. So the rows with the missing data were removed. Then using the pincodes from the website the latitude and longitudes of all the areas are retrieved. The new dataframe is formed with the places, pincodes and their respective latitudes and longitudes. The empty values are dropped from this dataframe also.

Foursquare API

With this dataframe and the foursquare api, the venues in each area is retrieved with the limit of 100 by making requests to the api. A full dataframe is created which has all the retrieved venues with its location for all of the areas. And this is checked for empty values.

Duplicate values.

There was no radius set when getting the venues from the api. So there may be some duplicated values which is not good for analyzing. So they are also removed.

One-Hot encoding.

The dataset was encoded inorder to view the most common places in all areas of Chennai.

This is done because we need to find the amount of restaurants and other common places in each area, so that we can decide on which area to choose. Then the dataframe was grouped and sorted such that now the dataframe shows the top 10 most common venues in each of the areas. Now the data is ready for clustering.

K-means Clustering.

This clustering is used to group data into clusters so that the elements in each cluster don't differ much but differ most from other clusters. The clustering was applied to the dataframe with each area showing the top 10 most common venues.

K-means clustering is used here to group the areas. By this similar areas are grouped which will be easy for us to identify the set of areas which can be ignored and similarly to identify the set of areas which are favourable to our criteria. Now the clusters which are favourable to our criteria can be grouped together.

Results.

In this clustering the last dataframe was clustered into 5 different groups. Each cluster group showed different types of venues in the top most common places.

For example, a cluster showed most of the top 5 common places to be Chinese Restaurants, Fast Food restaurants (as most fast food restaurants follow Chinese cuisines), Cafe etc. But this cluster had no more crowded places like malls, parks, zoos, multiplexes, train stations etc. Even if they had crowded places they were at the bottom of the most common places. So the clusters which had these types of criteria were removed.

On the other hand the clusters which had the most common places to be theatres, parks, bus stations, malls etc. were chosen and they were grouped and made into a new dataframe.

Now this new dataframe was finally analyzed for most crowded places and least number of Chinese and Fast Food restaurants. The area which had the most satisfying result was chosen one to start a Chinese restaurant for the client.
Discussion.

As the clustering was done, it was observed that all the areas in Chennai had the main Indian restaurant as their most common place. When analyzing our type of criteria the 1st most common place was always ignored. Only the place from the second most position was considered. And there was a cluster which was totally ignored because it had mostly Chinese and fast food restaurants within their 3rd most common places. Only one cluster among the 5 clusters that were created had zoos and parks in their most common places.

Conclusion.

In this project I have analyzed the venues in all areas of Chennai city to find a place which will be most suitable to start a Chinese restaurant. At final it was narrowed down to four places which were suitable for this criteria. And it was also observed that all the areas in Chennai have Indian restaurants as the most number of venues.
This project was done only based on the number of certain types of venues in Chennai. This could also be done based on actual footfall of people at times. The trending places in a location was difficult to get from the foursquare api. If this footfall numbers had been used in this project, it would have given more accurate results than using the most common venues.