Best Places to open Chinese Restaurants in Pune

This is my final capstone project for IBM Applied Data Science Capstone course in Coursera.

Introduction

For this Capstone project, I am creating a hypothetical scenario for a concept Chinese restaurateur who wants to explore opening an authentic Chinese restaurant in Pune area. The idea behind this project is that there may not be enough Chinese restaurants in Pune and it might present a great opportunity for this entrepreneur who is based in India. With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this entrepreneur and I am designing this project to help her find the most suitable location.

Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Chinese restaurant in Pune, India. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In Pune, if an entrepreneur wants to open a Chinese restaurant, where should they consider opening it?

Target Audience

The entrepreneur who wants to find the location to open authentic Chinese restaurant

Data

To solve this problem, I will need below data:

- List of neighbourhoods in Pune, India.
- Latitude and Longitude of these neighbourhoods.
- Venue data related to Chinese restaurants. This will help us find the neighbourhoods that are most suitable to open a Chinese restaurant.

Extracting the data

- Dataset containing Pune neighbourhoods via geonames.org
- Getting Latitude and Longitude data of these neighbourhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighbourhoods

Methodology

First, I need to get the list of neighbourhoods in Pune, India. This is possible by extracting the list of neighbourhoods from the geonames page
("https://download.geonames.org/export/zip/")

After gathering all this data, I visualized the map of Pune using Folium package to verify whether these are correct coordinates.

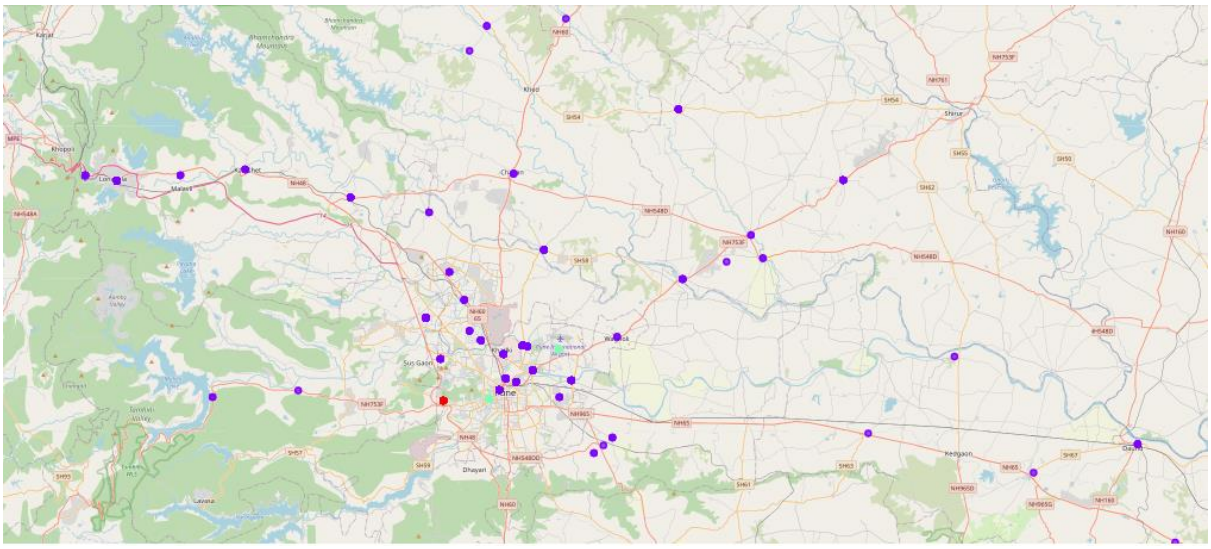
Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighbourhood by grouping the rows by neighbourhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for "Chinese restaurants".

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centriods, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighbourhoods in Pune into 3 clusters based on their frequency of occurrence for "Chinese food". Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Results

Clusters



The results from k-means clustering show that we can categorize Pune neighbourhoods into 3 clusters based on how many Chinese restaurants are in each neighbourhood:

- Cluster 0: Neighbourhoods with little or no Chinese restaurants
- Cluster 1: Neighbourhoods with few Chinese restaurants
- Cluster 2: Neighbourhoods with high number of Chinese restaurants

The results are visualized in the above map with Cluster 0 in purple color, Cluster 1 in light green color and Cluster 2 in red color.

Recommendations

Most of Chinese restaurants are in Cluster 2 which is around Sus, Bavdhan and Armament areas and lowest (close to zero) in Cluster 0 areas which are Lonavala, Narayan Peth, Yerwada areas. Also, there are good opportunities to open near Vadgaon Sheri, Film Institute as the competition seems to be low. Looking at nearby venues, it seems Cluster 1 might be a good location as there are not a lot of Chinese restaurants in these areas. Therefore, this project recommends the entrepreneur to open a Chinese restaurant in these locations with little to no competition.

Limitations and Suggestions for Future Research

In this project, I only take into consideration of one factor: the occurrence / existence of Chinese restaurants in each neighbourhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a new restaurant. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors. In addition, I am relying on the existence of Chinese restaurants only for this project but future research can take into consideration of other

variables such as existence of Asian restaurants, Asian population level in each neighbourhood etc.

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 the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.

References

List of neighbourhoods in Pune: https://download.geonames.org/export/zip/

Foursquare Developer Documentation: https://developer.foursquare.com/docs

All codes for this project can be found here: https://github.com/GB1111/Coursera_Capstone/blob/master/Pune_Chinese_Restaurants_final.ipynb