Coursera Capstone

IBM Applied Data Science Capstone

Building a Pizza Place in Hyderabad, India

By: Akshay Agnihotri
January 2021

Introduction

Hyderabad is one of the fastest developing cities of India, with a booming IT industry and various other opportunities. In such a scenario, it attracts lot of people from various other parts of the country, who come for work and also aids the food business. Among millennials and young IT crowd, fast food like Pizza is always a favourite due their busy lifestyles. Pizza, not only caters to the easy and fast-food needs of the locals, but is also a preference of any tourist, who is not sure about local cuisines. Of course, as with any business decision, building a Pizza outlet requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the outlet is one of the most important decisions that will determine how convenient it is for people to order pizza and also due to the competition of other pizza outlets in the area.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Hyderabad, India to build a new Pizza outlet. Using data science methodology and machine learning techniques like clustering, this project aims to solve this and get an answer to the business question: In the city of Hyderabad, India, if a business person or property developer is looking to build a new pizza outlet, where would you recommend that they do so?

Target Audience of this project

This project is particularly useful to individual IT working professionals, property developers, investors looking to buy or invest in new pizza outlets in the capital city of Hyderabad in the state of Telangana. This project is timely as the city's IT sector is booming and attracting more and more young crowd.

Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Hyderabad. This defines the scope of this project which is confined to the city of Hyderabad, the capital city of the state of Telangana in India.
- Latitude and longitude of these neighbourhoods. This is required for plotting the map and also to get the venue data.
- Venue data, particularly data related to Pizza Place. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

- This Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India) contains a list of neighbourhoods in Hyderabad, with a total of 200 neighbourhoods. We will use web scraping techniques for extracting the data from the Wikipedia page, with the help of Python requests and beautiful soup packages.
- Then we will use Python Geocoder package to get the geographical coordinates of the neighbourhoods which will give us the latitude and longitude coordinates of the neighbourhoods.
- After that, we will be using Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Pizza Place category in order to help us to solve the business problem put forward.
- This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

- Firstly, we need a list of neighbourhoods in the city of Hyderabad, India. This is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India). We will do web scraping using Python requests and beautiful soup package to extract the list of neighbourhood's data. This is only a list of names.
- We have to get the geographical coordinates in the form of latitude and longitude in order to use Foursquare API. We will use the Geocoder package to convert address into geographical coordinates in the form of latitude and longitude.
- Now, we will be populating the data into a pandas Data Frame and then visualize the neighbourhoods in a map using Folium package. This helps in a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Hyderabad. Next, we will use Foursquare API to get the venues.
- We need a Foursquare Developer Account to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format

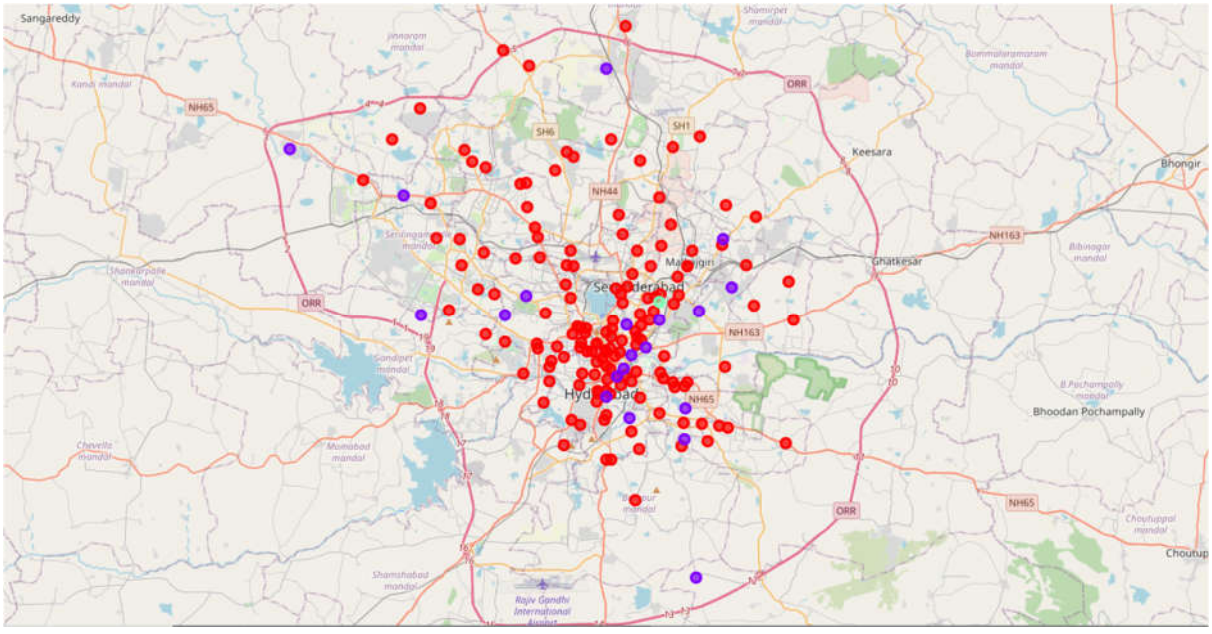
and extract the venue name, venue category, venue latitude and longitude. With the data, we can check the count of venues returned for each neighbourhood and examine how many unique categories can be determined from all the returned venues.

- Then, analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. We are now preparing the data for clustering.
- Since we are analysing the “Pizza Place” data, we will filter the same as venue category for the neighbourhoods.
- Lastly, will perform clustering on the data using k-means clustering. K-means clustering determines k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. This algorithm is optimal to solve the problem for this project.
- We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Pizza Place”. The results will allow us to identify which neighbourhoods have higher concentration of housing development and IT offices
- Based on the occurrence of “Pizza Place” in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to build a new Pizza Outlet

Results

- The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Pizza Place” :
    - Cluster 0: Neighbourhoods with high number of Pizza places
    - Cluster 1: Neighbourhoods with moderate number of Pizza places
    - Cluster 2: Neighbourhoods with low or no concentration of Pizza places

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

- As observed in the Results section, most of the Pizza places are concentrated in the central area of Hyderabad city, with the highest in cluster 0 and moderate in cluster 1. Cluster 2 has no Pizza places in the neighbourhoods. This represents a great opportunity and high potential areas to build a new pizza outlet.
- Meanwhile, pizza places in cluster 0 are facing tough competition, causing some inconvenience to the business. From another perspective, the results also show that the oversupply of pizza places mostly happened in the central area of the city.
- Therefore, this project recommends property developers to capitalize on these findings to build pizza outlet in neighbourhood of cluster 2. Lastly, property developers are advised to avoid neighbourhoods in cluster 0 which have a very high concentration of pizza places.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e., frequency of occurrence of Pizza places, there are other factors such as population and income of residents that could influence the location decision of a new pizza outlet. However, such data is not available to the neighbourhood level required by this project. Future research could estimate such data to be used in the clustering algorithm to find the preferred locations to build new pizza outlet. Also, we use the free Sandbox Tier Account of Foursquare API with limitations of number of API calls and results returned. Future research could make use of paid account to overcome these limitations and obtain more results.

Conclusion

In this project, we started the process by identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters depending on similarities, and then providing recommendations to the relevant stakeholders i.e., property developers and investors regarding the best locations to build a new pizza outlet. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to build new pizza outlet. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding areas in their decisions to build a new pizza outlet.