Coursera Capstone Project

IBM Applied DataScience Capstone

Opening a restaurant in Hyderabad, Telangana

Yahwanth Gundela

June 2020

Introduction

Undoubtedly, Food Diversity is an integral part of an ethnically diverse metropolis. The idea of this project is to categorically segment the neighbourhoods of Hyderabad City into major clusters and examine their cuisines. A desirable intention is to investigate the neighbourhood cluster's food habits and taste. Further examination might reveal if food has any relationship with the diversity of a neighbourhood. This project will help to understand the variety of a neighbourhood by leveraging venue data from Foursquare's 'Places API' and 'k-means clustering' unsupervised machine learning algorithm. Exploratory Data Analysis (EDA) will help to discover further about the culture and diversity of the neighbourhood.

Business Problem

This quantifiable analysis can be used to understand the distribution of different cultures and cuisines over Hyderabad City. Also, it can be utilized by a new food vendor who is willing to open his or her Restaurant or by a government authority to examine and study their city's cultural diversity better. Data To solve the following 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 coordinates of those neighbourhoods. This is required to plot the map and also to get the venue data.
- Venue data, particularly data related to Restaurants. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them

This Wikipedia page contains a list of neighbourhoods in Hyderabad, with a total of 220 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and Beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use 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 Restaurant 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). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Hyderabad. Fortunately, the list is available in the Wikipedia page. We will do web scraping using Python requests and Beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform 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 top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order 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 we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the Shopping Center data, we will filter the Shopping Mall as venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, 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 is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 8 clusters based upon the most common venues in its vicinity. The results will allow us to identify if it is possible to rename the neighbourhoods based upon the categories of venues in and around them.

Following could be the name of the clusters segmented and curated by k-Means unsupervised machine learning algorithm:

Results

- The above bar graph visualizes the various categories of restaurants for all the neighbourhoods.
- Shilloute method has been used to find optimum clusters and the optimum clusters for this project is 8.
- Cluster 0 — Indian American. Here, we can clearly see that the cluster is dominated by American Indian places considering that cafés, wings joints, BBQ joints come under the American cuisine and Indian Restaurant being in the first most common venue in the cluster.

- Cluster 1 — Café. Here, we can see that cafés dominate the cluster. We can also observe that this is the most diverse cluster consisting of almost all kinds of cuisines.
- Cluster 2 — Pizza. Here, we can observe that Pizza occupies the top spot in this cluster. We can also identify that it consists of all kinds of small eateries and fast food stalls.
- Cluster 3 — Fast Food. Here, we can see that it is a fast-food dominated cluster and is one of the smallest clusters of all.
- Cluster 4 — Mix of Cuisines. Here, we see that the cluster contains different types of cuisines such as Asian, French and Chinese. This shows us the cultural dominance of the above three in those areas. This is also the second smallest cluster of all.
- Cluster 5 — Snack places. Here, the cluster consists of different snacking places such as wings joint, chaat places and fondue places etc. It too is one of the smallest cluster among all.
- Cluster 6 — Indian Restaurant. Here, we can observe that the Indian Restaurant is the most common venue with 96 being its count value. It is the largest and most dense cluster of all.
- Cluster 7 — Vegetarian/ Vegan Restaurant.

Observations

Here, we see that the vegetarian Restaurant dominate all other cuisines conveying us the different pattern of diet compared to all the others' in the other clusters. It is also one of the smallest clusters.

Hence, by looking at the results above, we can confidently state that Hyderabad city is a diverse place with multiple cuisines all blended within it and people inhabiting it are brimming with the rich, diverse culture brought about from their native lands to this city.

Limitations and Future Suggestions

The results of this project can be improved and made more inquisitive by using a current Hyderabad City's dataset along with API platforms which are more interested in Food Venues (like Yelp, etc.) The scope of this project can be expanded further to understand the dynamics of each neighbourhood and suggest a new vendor a profitable location to open his or her food place. Also, a government authority can utilize it to examine and study their city's culture diversity better.