Analyzing the Neighborhoods in Visakhapatnam for setting up a new Restaurant

IBM APPLIED DATA SCIENCE CAPSTONE

by : Swarna Naga Sri Tejaswi

Introduction

Visakhapatnam also known as Vizag is the proposed executive capital of the Indian state of Andhra Pradesh. It is also the most populated and largest city of Andhra Pradesh. It is the second largest city in the east coast of India after Chennai and also the fourth largest city in South India. It is one of the four smart cities of Andhra Pradesh selected under Smart Cities Mission. With an estimated output of $43.5 billion, the city is the ninth largest contributor to India's overall GDP as of 2016. The city is home to some reputed Central and State educational institutions. The city is a major tourist destination and is particularly known for its beaches, Buddhist sites and natural beauty. It has been nicknamed as the "City of Destiny". The main aim of the project is to study the neighborhoods of Visakhapatnam to determine possible locations for starting a restaurant. This project can be useful for business owners and entrepreneurs who are looking to invest in a restaurant in a smart city like Visakhapatnam. The main objective of this project is to analyze appropriate data and find recommendations for the stakeholders.

Data Collection

The data required for this project is as follows and has been collected from multiple sources. The following data is required for the project:

1) Neighborhood data of Visakhapatnam

2) Geographical Coordinates of Visakhapatnam and all neighborhoods in Visakhapatnam

3) Venue data for neighborhoods in Visakhapatnam

Neighborhood Data

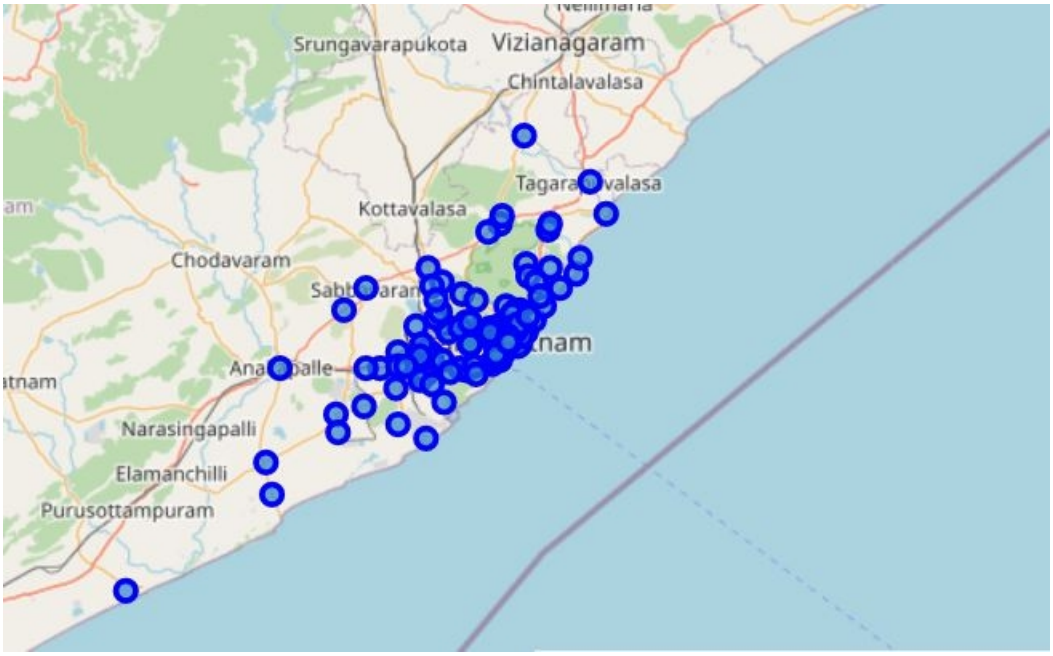
The data of the neighborhoods in Visakhapatnam was scraped from https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Visakhapatnam.We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods. After that, we will use Foursquare API to get the venue data for those neighborhoods. 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).

Target Audience of this Project

This project is specifically useful to the investors looking to open or invest in new restaurants in the city of Visakhapatnam

Map of Visakhapatnam

This map represents all the neighborhoods of Visakhapatnam. Using K-Means clustering technique, the neighborhoods are grouped into different clusters



Methodology

Firstly, the Neighborhood data is extracted from the Wikipedia page https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Visakhapatnam. We will do web scraping using python requests and beautifulsoup packages to extract the list of neighborhoods data. Then 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 the 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 neighborhoods 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 Visakhapatnam. Next, we

will use the Foursquare API to get the top 100 venues that are within a radius of 2000 meters. By using the Foursquare API, we make API calls to Foursquare passing in the geographical coordinates of the neighborhoods. Foursquare will return the venue data in JSON format and we will extract the venue name, category,latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues . Then, we will analyse 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 “Restaurant” data, we will filter the “Restaurant” as venue 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 neighborhoods into 3 clusters based on their frequency of occurrence for “Restaurant”. The results will allow us to identify which neighborhoods have a higher concentration of Restaurants. Based on the occurrence of restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable for opening new restaurants. Therefore, this project recommends investors to capitalize on these findings to open new restaurants in neighborhoods in cluster 0 with little to no competition.

Results

The results from the k-means clustering show that we can categorize the

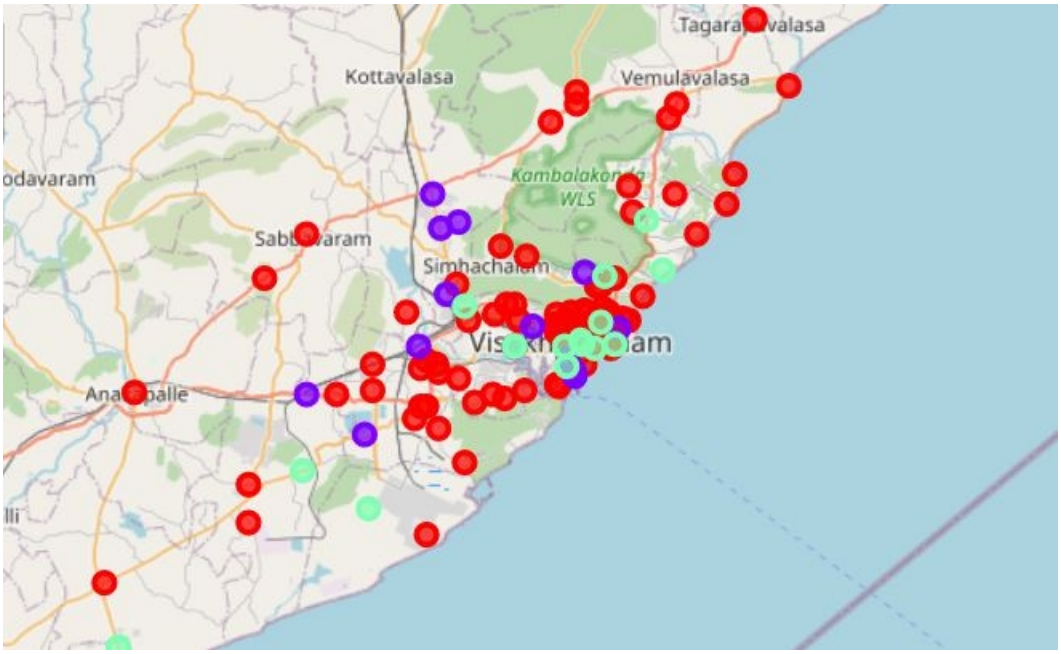
Neighborhoods into 3 clusters based on the frequency of occurrence for "Restaurant":

• Cluster 0: Neighborhoods with a very less or no restaurants

• Cluster 1: Neighborhoods with a high concentration of restaurants

• Cluster 2: Neighborhoods with a moderate concentration of restaurants

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 Green colour.



Discussions

From the observations noted from the map in the Results Section, most of the Restaurants are located in cluster 1, followed by the moderate number of restaurants in cluster 2. This represents a great opportunity to open new restaurants for investors in cluster 0 with little or no competition. It is also observed that more number of restaurants are located in neighborhoods present in the heart of the city, with most of the sub-burbs having very less number of restaurants. The neighborhoods present in cluster 0 are great residential areas with very less or no restaurants.Therefore, this project recommends investors to capitalize on these findings to open new restaurants in neighborhoods of cluster 0 with little competition. Lastly property developers are advised to avoid neighborhoods in cluster 1 which already have high concentration of restaurants and suffer from intense competition.

Limitations

In this project, we only consider one factor i.e frequency of occurrence of restaurants.There are other factors such as population and income of residents that could influence the location decision of a new restaurant.However, to the best knowledge of this researcher, such data is not available to the neighborhood level required by this project.Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new restaurant.In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid service to bypass these limitations and obtain more refined results.

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 machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. investors regarding the best locations to open a new Restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations to open a new restaurant in Visakhapatnam.

References

Data for the neighborhoods retrieved from
https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Visakhapatnam