Capstone Project — the Battle of Neighbourhoods in Patna

Introduction

Patna capital of Bihar state, northern India. Patna is a riverside city that extends along the south bank of the Ganges. Total population of Patna is more than 6.5 Million. It has total 40 neighborhood. It seems that over last decade its continuously grow as a result it grab 33rd spot in the Ease of living index 2020 released by the Union ministry of housing and urban affairs. The official language of Patna and the one that is most widely spoken is Hindi. However, English is also spoken as a formal language within businesses and government agencies. With its diverse culture, comes diverse food items. There are many restaurants, Food Truck and Cafe's in Patna.

Aim

This project aims to find the best location to open a Café, Chinese Restaurant, Furniture store, American Restaurants in the city of Patna, India, to maximize the profit of the Client and give him best recommendations. The target audience for this project is the people who want to start these above business. We will provide him location where he can make good money.

Data Description

Dataset containing Neighbourhoods of Patna, the dataset holds the names Neighbourhood in Patna extracted from Wikipedia. Here is a link. **https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Patna.**

I used python library geopy.geoder to find the coordinates of neighborhoods in the data frame. After that we visualize using folium. Then used Foursquare API to explore the most common places of a neighborhood in the form of a JSON file.

Methodology

In this section, I will describe the data analysis and how I used the data to yield the results.

I am using BeautifulSoup to scrap the data from the mentioned URL and take the fields which I will be working on. Then, I enabled geopy functions by installing the conda-forge geopy package. I used the nominatim function to add geospatial data to the data frame that is the latitude and the longitude. Using the folium package and my data frame, I then created a map with the forty one city districts on it.

Now, foursquare data comes into play. Then, retrieved the foursquare data for all venues on foursquare with a distance of less than 3000 meters from each center of each city district, as indicated as blue dots in the map above. The result was a list of 75 venues all over Patna City. There are 35 unique categories. There are different type of Restaurant, cafe and Furniture store.

we get Buddha Colony and Serpentine Road returned the highest number of venues i.e. 21. With the help of One hot encoding on the obtained data set i use it find out the 10 most common venue category in each neighborhood.

There are many techniques are available in Data science field. For my project, I am going to use clustering. Particularly, I am going to use K-Means clustering. So, I need to know how many optimal number of cluster could be better for my data. To find the optimal number of clusters we will use silhouette score metric technique. Now we have cleaned data and found all the prerequisite to model our data from the graph it looks cluster 4 could be the better one.so, let's use number of clusters as 4.

Add the cluster labels to the neighborhoods_venues_sorted dataframe. And let’s create a new dataframe kl_merged which has the neighborhood details, cluster labels and the 10 most common venues in that neighborhood. we see in the table are the city districts and their most common venues, and they now have been assigned five different cluster labels from 0 to 3.

We can now use the cluster labels to show the city districts marked with a cluster-specific color on a map, again using folium: As in the map every cluster had given different colors. so there are total 4 cluster of different color cluster 1

color is Purple, Cluster 2 color is sky blue, Cluster 3 color is red and cluster 4 color is Fade yellow.