The Battle of Neighborhoods | Business Problem

Introduction: Tourism's value is shown all over the world. There is no denying the significance of tourism, from the economic benefits it provides to host communities to the delight it provides to travelers. Tourism's significance may be seen from two perspectives: the tourism business and the visitor. Traditional processing is used by many tourism businesses. A businessperson in this field must be skilled in management, communication, and planning. There are several challenges in managing various facilities for consumers in a typical company model. Every firm nowadays is going digital. In the travel sector, data science may sound strange, but it is the most effective method to build a tourist firm online. To begin, let's define data science. Data science is a field that extracts data and understanding from unstructured and structured data using scientific methods, procedures, algorithms, and systems. Every business collects and creates a large amount of data from a variety of sources.

Problem Identified: The problem identified for this project is to provide the customers/users with the most personalized experience of venues around them for them to visit. This project attempts to give solutions to the business challenge using data science approach and machine learning algorithms such as clustering. Therefore, for my capstone project, I wanted to create something that people could use in their daily lives. This prompted me to create a recommendation system model that might propose local restaurants their location based on other people's restaurant ratings, to potentially enhance the recommendation suggestions that you can see on popular food delivery services.

Objectives: This capstone project aims to fulfil the following objectives:

- We describe accurate segmentation and targeting using fine-grained segmentation of national tourism based on patterns of movements and visits, which substantially increases the incomplete and fragmented data gathered from population data.
- Recommendations to tourists based on their preferences and ranked accordingly
- Popular restaurants around the area

Data Requirements: To build the recommender system, we require population data regarding the city and areas in Bengaluru. The most important data that needs to collected are:

- Geographic coordinates such as latitude and longitude of the areas to find out where they are located
- Population of the area where the venues are located
- Average income of population across the areas.

The following steps must be accomplished for data collection:

- To go to a restaurant's location, we need to know its Latitude and Longitude so that we can point to its coordinates and generate a map with all of the restaurants labelled appropriately.
- A neighborhood's population is a significant element in influencing a restaurant's growth and the number of customers that come in to dine. Logic dictates that the larger the population of an area, the more people will be interested in walking into a restaurant openly, and the smaller the population, the less people will frequent a restaurant. Also, the higher the number of visitors, the higher the restaurant's rating because it is visited by individuals of various tastes. As a result, it is a crucial element.
- A neighborhood's income is just as significant as its population. The wealth of a neighborhood is directly related to income. If residents in an area make more than the national average, it is highly likely that they will spend more, although this is not always the case. As a result, a restaurant's evaluation is proportionate to the neighborhood's revenue.

Data Collection:

Gathering location position is not hard, but that was not accessible on open access websites such as Wikipedia, India's government website, census report websites, and so on after more than two days of searching. So I opted to utilize Google Maps API to get latitude and longitude, however the free account only allowed me to make a limited amount of requests. Firstly, I used beautifulSoup4 to scrape a list of neighbors from Wikipedia[1]. The table headers serve as boroughs, while the data serve as neighborhoods. Bangalore is divided into eight boroughs and 64 neighborhoods. So, I did a Google search and manually looked up each neighborhood's coordinates. Following that, I created the data

[1] https://en.wikipedia.org/wiki/List_of neighbourhoods_in_Bangalore

frame below. Because the information is easily available, determining the population by neighborhood is very simple. In Bangalore, however, this is not the case. For a few areas, it was possible to find population figures. The population of the rest of the neighborhood is estimated and may be wrong, but because this is a demonstration project, the major goal is to get the model to operate. Neighborhood income is freely accessible through Wikipedia page. We next extract all of the columns from this Wikipedia page and convert them to a pandas dataframe using the beautifulsoup4 package, a Python tool that helps you scrape data from web sites. The latitude and longitude of all of the regions in the dataframe are then obtained using Python's GeoPy module. The purpose of using Foursquare is to find the closest venue locations so that we may build a cluster. The Foursquare API makes use of the ability of Foursquare to identify nearby venues within a certain radius (in my instance, 500 meters) as well as matching coordinates, venue location, and names.

Methodology:

To begin, we must get a list of the city of Bengaluru's neighborhoods. Conveniently, the list may be found on Wikipedia. To retrieve the lists of neighborhoods info, we will use web scraping using Python requests and beautifulsoup packages. This is, however, merely a list of names. To utilize the Foursquare API, we ought to obtain geographical position in the form of latitude and longitude. To do so, we'll utilise the fantastic Geocoder library, which allows us to translate addresses into geographical coordinates (latitude and longitude). We'll collect the data, put it into pandas DataFrame, and then use the Folium package to display the neighborhoods on a map. This allows us to do a validation check to ensure that the geographical coordinates data given by Geocoder is plotted accurately in Bengaluru.

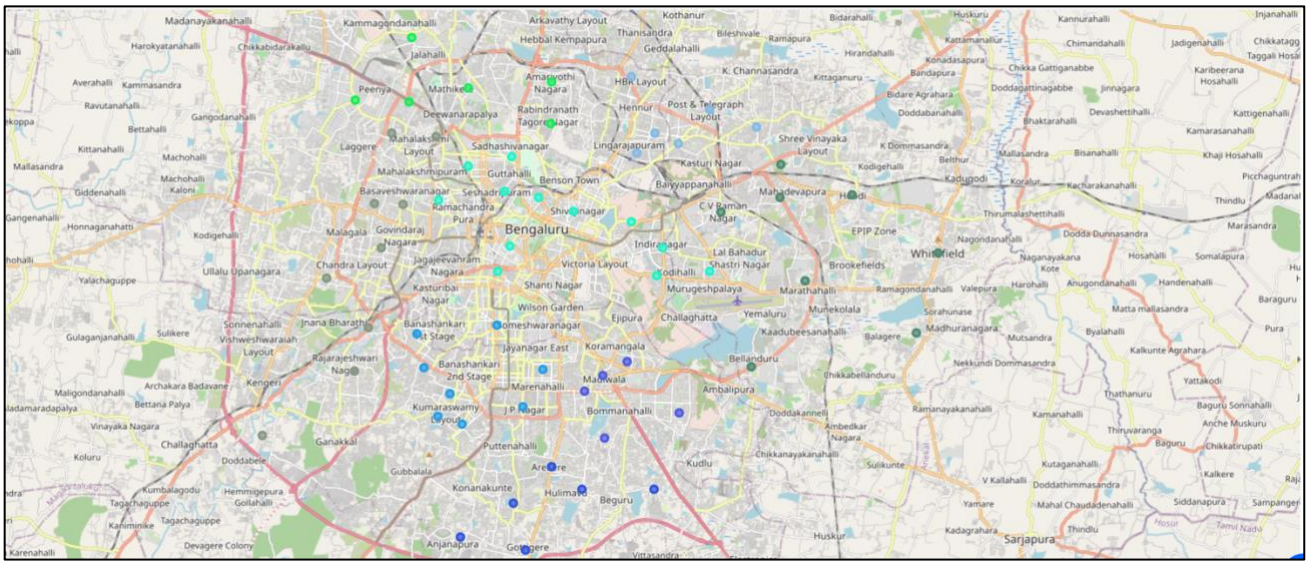
To acquire the Foursquare ID and Foursquare secret key, we must first create a Foursquare Developer Account. In a Python loop, we then make API requests to Foursquare, handing in the geographical coordinates of the neighborhoods. Foursquare will send the venue data in JSON format, from which we will extract the name, category, latitude, and longitude of the venue. We can use the data to see how many venues were returned for each neighborhood and how many distinct categories can be curated from all of the venues that were returned. Then, by grouping the rows by neighborhoods and calculating the mean of the frequency of occurrence of each venue type, we will analyze each neighborhood. We're also prepping the data for clustering by doing so. We will pick the "Restaurant" as a venue type for the neighborhoods because we are studying the " Restaurant " data.

Finally, we will use k-means clustering to do data clustering. The K-means clustering algorithm finds k centroids and then assigns each data point to the closest cluster, keeping the centroids as tiny as feasible. It is one of the most basic and widely used unsupervised machine learning methods, and it is well suited to the task at hand. We'll divide the neighborhoods into three groups depending on the frequency with which "Restaurants" appears. The findings will help us determine which areas have a larger concentration of restaurants and which areas have a lower number of restaurants. It will help us answer the issue of which neighborhoods are most suited to develop new retail malls based on the presence of restaurants in different neighborhoods.

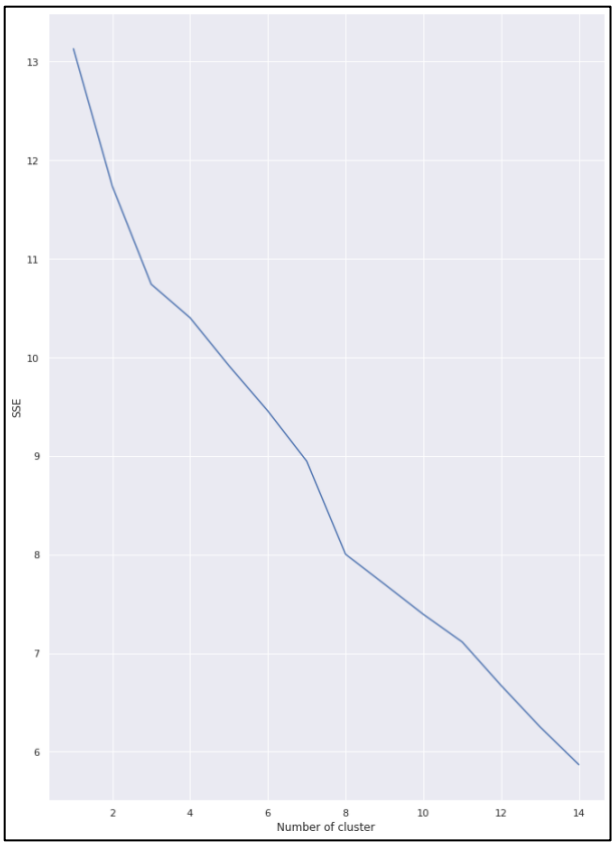
Analysis:

To begin with, depending on the pattern of occurrence for "no. of existing restaurants," the k-means clustering findings suggest that we may group the neighborhoods into three clusters:

- Cluster 1 (cyan): Neighborhoods with a low number of restaurants or none at all.
- Cluster 2 (Sky blue): Communities with a large number of restaurants
- Cluster 3 (Light Green): Areas with a moderate number of restaurants.
- Cluster 4 (Violet): Clusters with maximum number of restaurants



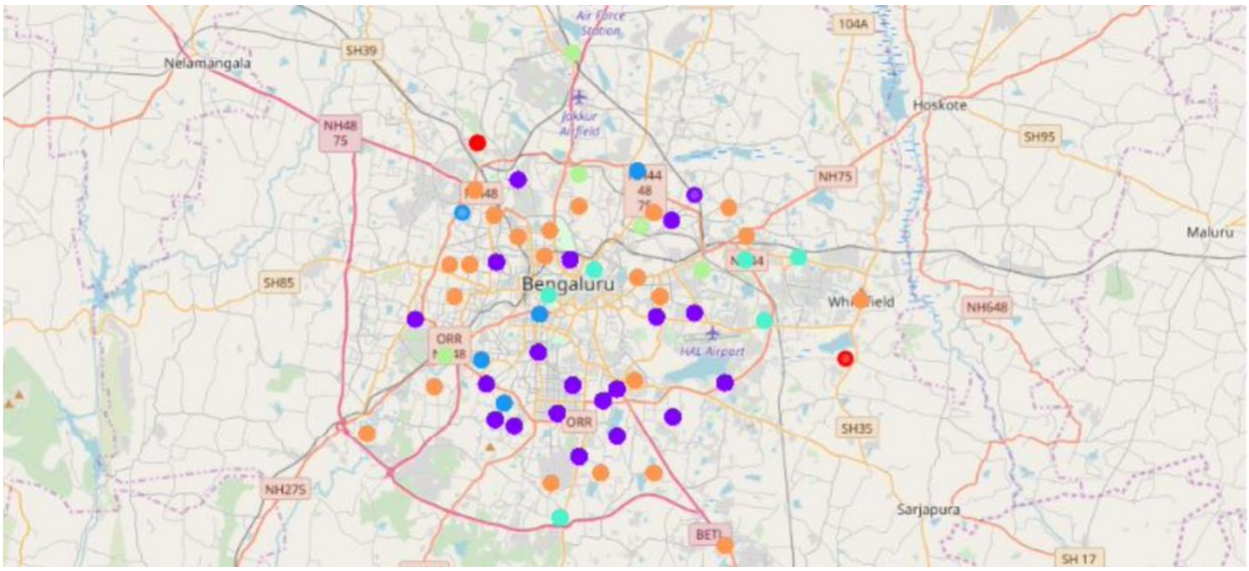
The elbow point for the k-mean cluster produced is 4, and the graph is produced to explore all the values for 4 clusters.



Discussion:

Population and Income were the most significant variables in developing the recommender system. They are the most important component since, according to our research, they have a nonlinear connection. To comprehend this nonlinear link, some inferential analysis was required. When the population of a neighborhood grows, it does not always indicate that the community's average income grows as well. This is true in the majority of situations, although many cases deviate from the norm. Similarly, a neighborhood with fewer residents does not always imply a lower average income. It is conceivable to have a lower population yet a higher income, and vice versa.

Because there was a nonlinear link between income and population, we may conclude that we must constantly use an inferential technique to discover relationships between various variables. Similar neighborhoods must also be thrown into the appropriate cluster during clustering. Another point to consider is that the number of clusters used might yield a wide range of results. Some may be too tight, while others may be too loose. As a result, a number of clusters analysis is required. In the Methodology section, look up elbow graph for reference.



Result

Out[103]:

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Arekere	Venue Category_Indian Restaurant	Venue Category_Sporting Goods Shop	Venue Category_Department Store	[0.22959888840700646]
1	Begur	Venue Category_Supermarket	Venue Category_Café	Venue Category_Mobile Phone Shop	[0.6361321887351776]
2	Bommanahalli	Venue Category_Department Store	Venue Category_Indian Restaurant	Venue Category_South Indian Restaurant	[0.4365669702740494]

The recommender system generates a list of top restaurants as well as the most popular venue item that the user may enjoy. During the model's runtime, a simulation was conducted using “Ulsoor” as the neighborhood, which was then processed via our model to suggest neighborhoods with comparable characteristics to “Ulsoor”.

Limitations and Scope for Future Study:

I attempted to cover the majority of the factors that impact people in this effort. However, there are additional elements that may be considered in order to promote realistic business development, such as infrastructure quality, all of which could have an impact on the preferred placement of the existing restaurants and visitors to the area. However, to the best of my knowledge, such data are not available at the local level that this project requires. Future study might develop a system for estimating such data, which could then be utilized in the clustering process to find the best places to build additional infrastructure. In addition, this project took use of the Foursquare API's free Sandbox Tier Account,

which has restrictions on the amount of API requests and results delivered. To get around these constraints and get better findings, future study might use a premium membership.

Conclusion:

The recommender system uses the Foursquare API to find local venues, taking into account characteristics such as population and wealth. It's a strong data-driven model whose efficiency may suffer as more data is added, but accuracy improves. It will assist consumers in quenching their hunger by making the finest recommendation to meet all of their requirements. In this project, I went through the steps of identifying business problems, specifying the data needed, extracting and preparing the data, visualizing the results, performing machine learning by clustering the data into three clusters based on frequency similarities, and tackling and resolving business problems (mentioned in results).