Business Set-Up Location Recommendation

Thanya Ramanathan

May 28, 2021

1. Introduction
   1. Background

One of the oldest civilizations in the world, India is a mosaic of multicultural experiences. With a rich heritage and myriad attractions, the country is among the most popular tourist destinations in the world. India is home to the finest architectural heritage, serene ghats, spectacular landscapes and the largest tiger reserve. Tourism in India is vital for the country’s economy and is expanding rapidly. The World Tourism and Travel Council calculated that tourism generated ₹16.91 lakh crore (US$240 billion) or 9.2% of India’s GDP in 2018 and supported 42.673 million jobs, 8.1% of its total employment.

* 1. Business Problem

As mentioned earlier, tourism facilitates a large number of employment opportunities. This project analyses the most popular places in India to recommend the best choice for setting up restaurants, hotels, souvenir shops or other businesses that tourism may help flourish. This project will also be helpful for freelancers like translators, artists, etc., to set up stalls.

1. Data Acquisition and Cleaning
   1. Data source

The names of all the states and the districts they contain were obtained from the Wikipedia page <https://en.wikipedia.org/wiki/List_of_districts_in_India>. The webpage was scraped using the BeautifulSoup package. The datasets obtained from this webpage only contain the name of the district, its headquarters, area in square kilometers, density, and population. In order to obtain the respective latitude and longitudes of the districts, Python’s geopy package was used.

* 1. Data Cleaning

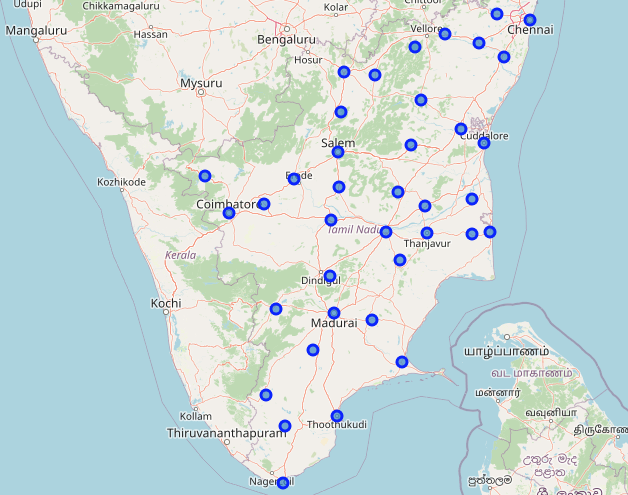
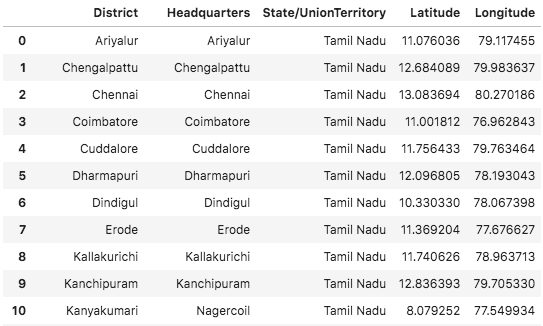
The data collected from the various sources were combined in to one data frame that had the names of the districts, the states they belonged to, their headquarters and their respective latitudes and longitudes. There were many missing values that were dropped. The datasets had many inconsistencies.

The final data frame (Indialatlong.csv) is the combination of several individual data frames that contained the details of districts of a specific state. Though obtained from the same source, some of these data frames had columns that were named differently. Hence, the names of the similar columns in all datasets had to be standardized.

When these datasets were combined, there were repetitions. The duplicates and irrelevant features were dropped. This project can be further developed to include features like population, density and area will be very helpful in determining the perfect location to set up a business but they will not be necessary for this project.

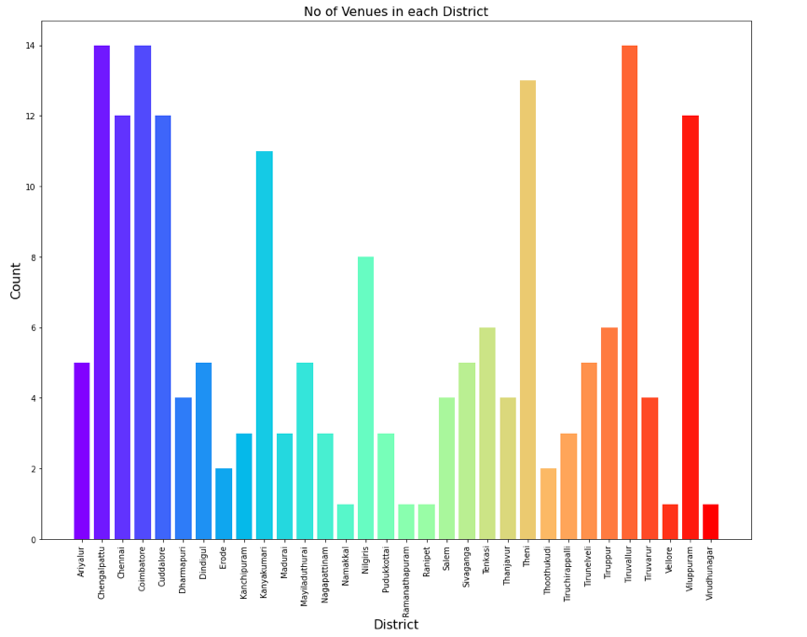
1. Methodology
   1. Data pertaining to the state of choice

The user is required to input the state that he/she wishes to set up his business in. This input is used to obtain the names, latitudes and longitudes of all the districts in that state. These districts are then superimposed on the map of the state using Folium library.



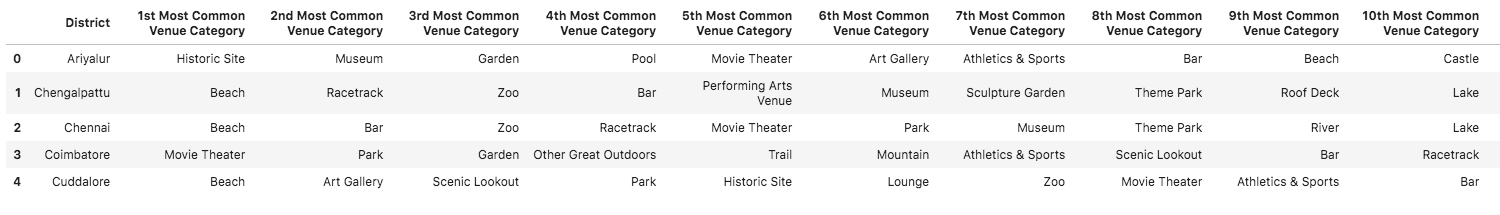
* 1. Foursquare API

Using my Foursquare API, I first looked through all the categories and sub-categories that the API contains. The next step displays all the venues that are present in the 50km radius. The radius can also be set according to the area of the district. Once the venues have been procured, a bar graph with the number of venues in each district is plotted.



* 1. Analyzing Each District

Each of the districts are then analyzed for the number of venues for each category type. This is then stored as a data frame that allows us to predict the most common type of venue in each district. A new data frame (district\_venues\_sorted) is created that specifies the ten most common types of venues in a district.



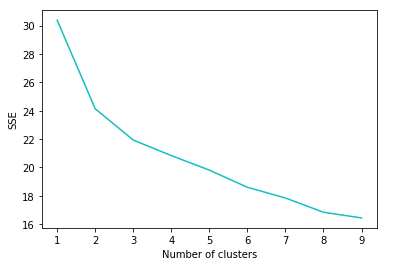
* 1. Nearby Businesses

From this, the user can decide where to start a business in a particular district. In order to help him determine the type of business, a list of all the businesses that have been set up in a radius of 500 meters from the venue is listed out. This is again saved as a data frame (nearby\_business\_sorted) that has the ten most common businesses in that area.



* 1. K-Means Clustering

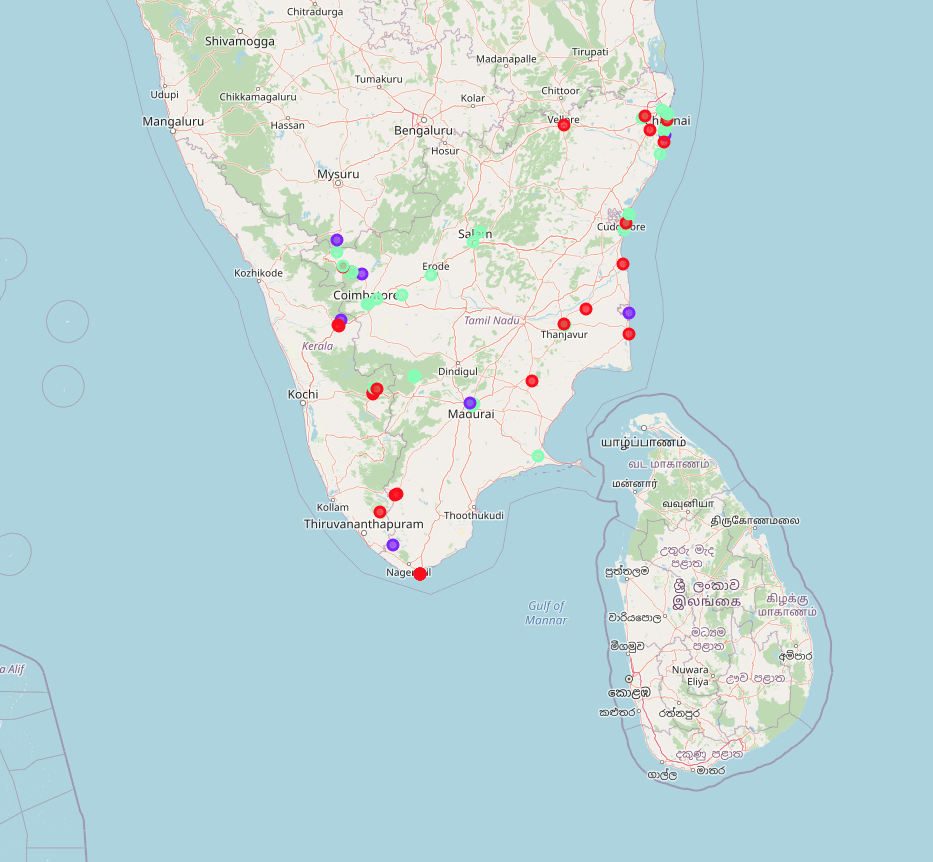
To sort these businesses into clusters, K-Means clustering is chosen. This groups the business into K clusters. In order to obtain the optimum K value, we use the elbow method.



The elbow method runs k-means clustering on the dataset for a range of values for k and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned centre.

1. Results and Discussion

The businesses are grouped in to 3 clusters that are depicted on the state map. The purple, green and red markers represent clusters 1, 2 and 3 respectively. The region close to Chennai has the greatest number of business outlets.



Experiencing the local cuisine will always be an integral part of any travel plan. Undoubtedly, most of these venues are restaurants, cafés and food shops. Hotels for tourists to stay closely follows. Souvenir shops, malls, markets, departmental stores and clothes stores are other types of establishments can be found.

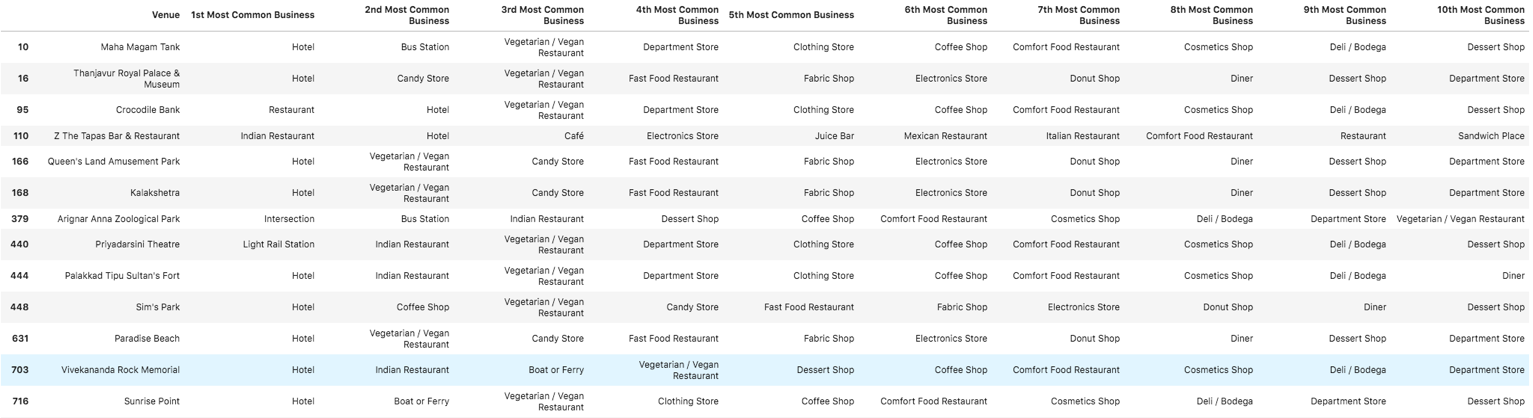
Cluster 1:



Cluster 2:



Cluster 3:



1. Conclusion

K-Means clustering algorithm has helped segregate tourist spots into clusters. This segregation is based on the frequency of visits to famous tourist spots. India has an abundance of these spots and will help anyone looking to start a business to benefit from tourism. Foursquare API covers all the businesses already present; this data can be studied by the user to plan his venture. The user can either decide to come up with a unique business idea to intrigue the tourists or set up something similar to that which is already present in order to offer a variety to the customers.

This project can be developed to include business rates of the nearby businesses, land availability, accessibility. For any business, raw material is important as well. Procuring raw material is not from vendors that are not available in the 500-metre radius might be inconvenient and unwise, especially for food stalls. A list of the available raw materials can be included as well for convenience of the user. Considering more factors that affect a business while developing a model will help in making a more accurate analysis.