Data Science Capstone Project

Finding most suitable place to open a Residency Inn in the Capital city of Telangana, India.

By - Sai Raj Kumar Tellakula

(https://github.com/sairajkumar)

Introduction

Idea Behind Project:

In this Capstone Project, I am assuming a concept of opening a Residency hotel who are exploring for a best location to open a Residency Inn in Twin cities. The idea behind this project is it's a capital city of Telangana which attracts several people for a holiday and has several monument areas to visit. If you have never been here, I recommend to explore this South Indian place and if you are in an idea to open a hotel for business, I invite you to explore this project and you can find useful.

Business proposal

Our project objective is to find a suitable and lovely place surrounded with different popular places across twin cities (Hyderabad and Secunderabad). This can make business lucrative. Then let's discuss the factors for an ideal neighborhood in a holiday. These can be Restaurants, Museums, Theme Parks, Temples, Historic Monuments, Famous Bazaars, Transportation Areas etc.

Using our Data Science Methods, we try to cluster the top areas in the neighborhood and aim to locate a place to answer the question: "What is the most suitable place to open a lucrative Residency Hotel in Twin cities?"

Target Audience:

People who are looking to open a Residency Inn in the capital city of Telangana, IN.

Data Section

What data is needed?

The data needed for this project is

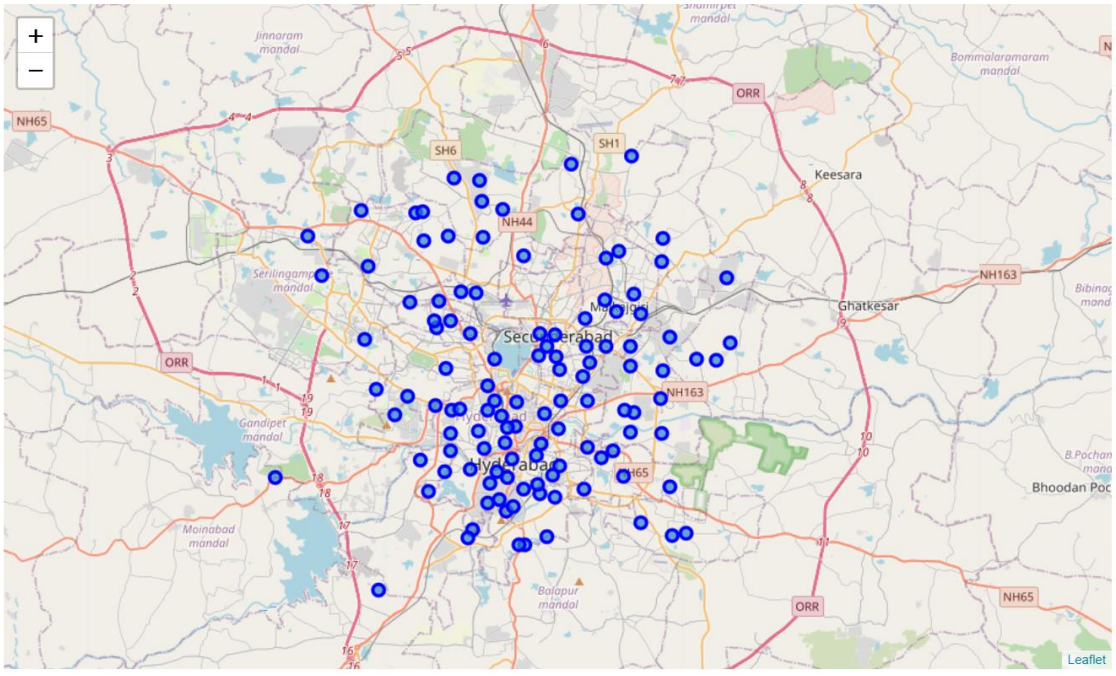
1. Areas across the cities Hyderabad and Secunderabad.
2. Latitude and Longitude values for those areas.
3. All famous venues in the neighborhood of those areas.

Data Extraction

1. Finding required area data of Twin cities from a Wikipedia page.
2. Finding Latitude and Longitude values using Geocoder package.
3. Exploring all famous venues that are related to our project in the neighborhood using Foursquare API.
4. The information extracted from Four Square API are ward name, and nearby venues with respective latitude and longitude values.

Methodology

In this project, our dataset is extracted from a Wikipedia link [1]. This contains the details of zones, circle names and corresponding ward names in Hyderabad District. Use this data and extract the zone and ward names for each zone. Later using the geocoder package in Python, the latitude and longitude values for each ward area are retrieved.



Let's start visualizing the wards on folium map and analyze the areas. After using the foursquare API let's explore the venues around each ward for every zone and extract the data from json file using json_normalize from pandas.

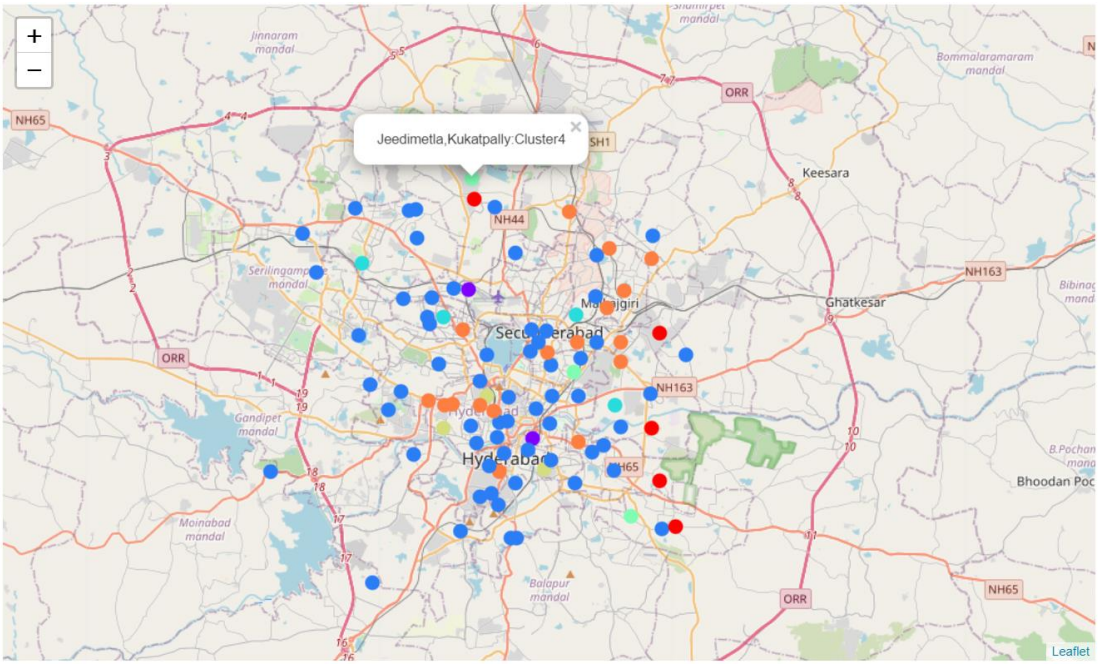
For this project, I focused on different types of popular areas across the district. The data extracted consists of several venues related to food, entertainment and historic places. Analyzing and finding the types of venues and number of different categories and their frequency around the zone, I finally ended to use KMeans clustering algorithm to find clusters of similar venues.

Because KMeans clustering is one of the powerful machine learning algorithms to cluster the geographical data. The venues are not suitable to directly analyze with the KMeans Classifier from sklearn. Firstly, data needs to be transformed into numeric data for analysis. One hot encoding works in this purpose. It converts the categorical to binary(Y/N) formatted data. It increases the number of features of the dataset and depicts a particular category existed or not.

Finally, the data is made fit with KMeans and similar venues are assigned to clusters containing similar venues.

Results

After using the KMeans clustering algorithm with number of clusters as 7. The results showed a discipline analysis between clusters.



The above picture depicts the clusters and their locations on folium map. The cluster of color Blue is the largest cluster of all and comprises the locations which have popular venues of categories such as Shopping Places and Places to visit like parks, historic monuments. Whereas the orange cluster has the locations which have popular venues of categories such as restaurants.

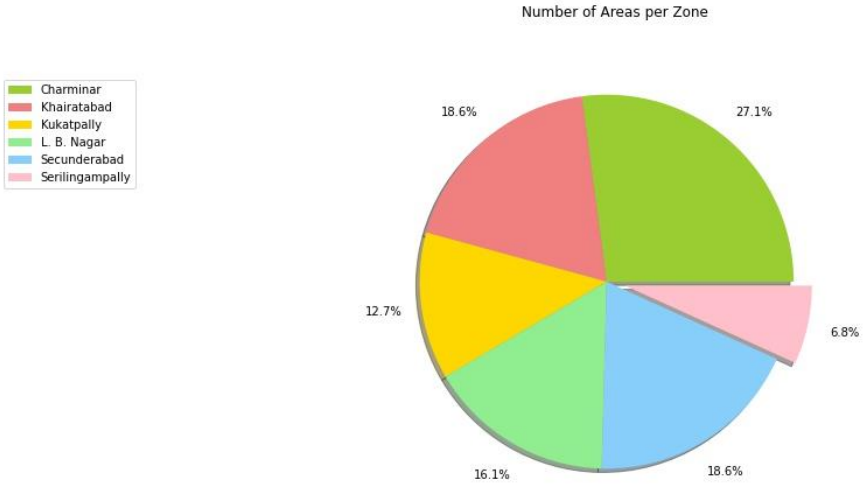
The next most considerable cluster is of color cyan comprises venues such as Snack Places and Food Courts.

Recommendations:

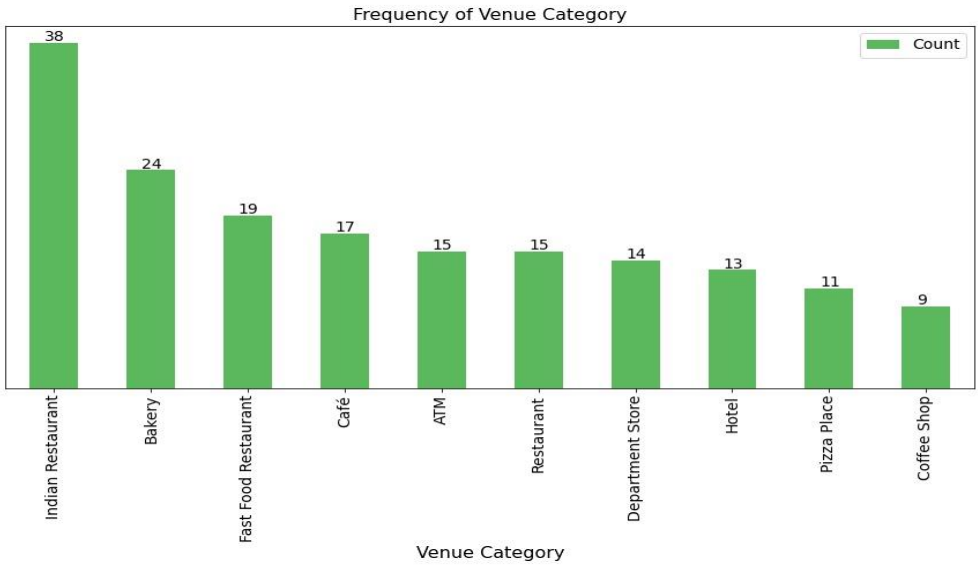
Having a Residency hotel near these specified clusters would boost the business and gives the visitors a great experience to visit the places or venues nearby.

Discussion

During the analysis of this project I found the following things that helped in understanding furthermore.



This pie chart shows how many wards are contributed to every zone. Where the Charminar Zone have more wards and is the largest which can be considered as a factor for finding a place.



Later after exploring all the venues in the wards, I plotted the following bar graph that shows the frequency of top 10 venue categories of all the unique categories found. This shows the people around or people who visit could have interests in Restaurants and café places.

Conclusion

After all the results and analyzing the clusters the places that are surrounded by the clusters of colors Red, Peacock Green and Light Green have the perfect venues that are required for a family of people who comes on a holiday and stays in the Residency.

Then the best recommendation to open a Residency Inn is to be surrounded by these clusters.

References

In order to complete this project, I have used the Wikipedia page link

[1] https://en.wikipedia.org/wiki/Hyderabad_Corporation_zones,_circles_and_wards

and the services of Four-Square API: https://developer.foursquare.com/