ANIRUDH SRIDHAR

DATA SCIENCE
CAPSTONE
PROJECT

Chennai Neighbourhood Analysis

Introduction

Background : Street vendors are often those who are unable to get regular jobs in the remunerative formal sector on account of their low level of education and if they start the business they are not able to find the right place for the business and end up being poor.

Business Problem : The condition of hawkers is not good in our society many of them commit suicide as they are unable to pay the debts. They don't have proper knowledge about "Where to set up their business?" , "Which location will get the most of customers and will be ideal for sales?" as they don't have these answers they end up putting their stalls near to large stalls which sell the same thing and end up competing with them.

Interest : "The main objective is to find an ideal place for the hawkers to sell their items so that they could get maximum profit."

Data Acquisition and Cleaning

Neighbourhoods : The data of the neighbourhoods in Kolkata can be extracted out by web scraping using BeautifulSoup library for Python. The neighbourhood data is scraped from a Wikipedia webpage.

Geocoding : The file contents from kolkata.csv is retrieved into a Pandas DataFrame. The latitude and longitude of the neighbourhoods are retrieved using Google Maps Geocoding API. The geometric location values are then stored into the initial dataframe.

Data Cleaning : Outliers are removed and all the values that are not available in the dataframe are also removed.

Venue Data : From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in the required parameters to the Foursquare API and creating another DataFrame to contain all the venue details along with the respective neighbourhood.

Final Dataframe

In [14]:
```
chennai_venues=getNearbyVenues(names=df['Neighbourhood'],
                               latitudes=df['Latitude'],
                               longitudes=df['Longitude']
                              )
chennai_venues.head()
```

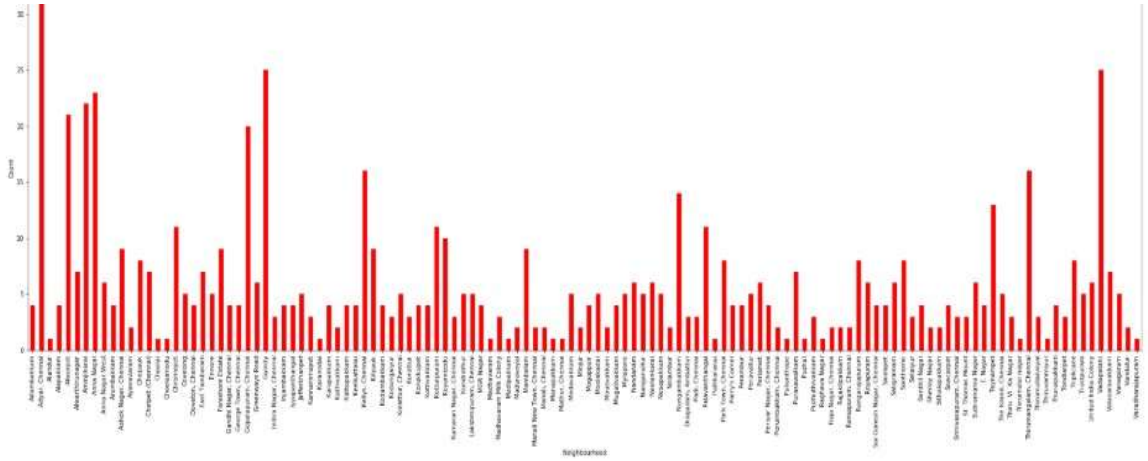
Out[14]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adambakkam	12.982221	80.209121	Return Gifts Online	12.982387	80.208338	Gift Shop
1	Adambakkam	12.982221	80.209121	arun icecream	12.983447	80.207847	Dessert Shop
2	Adambakkam	12.982221	80.209121	Sutherland	12.981002	80.205200	IT Services
3	Adambakkam	12.982221	80.209121	Bistro	12.983193	80.205020	Indian Restaurant
4	Adyar, Chennai	13.006450	80.257779	Bombay Brassiere	13.006961	80.256419	North Indian Restaurant

Methodology

Data Insights :

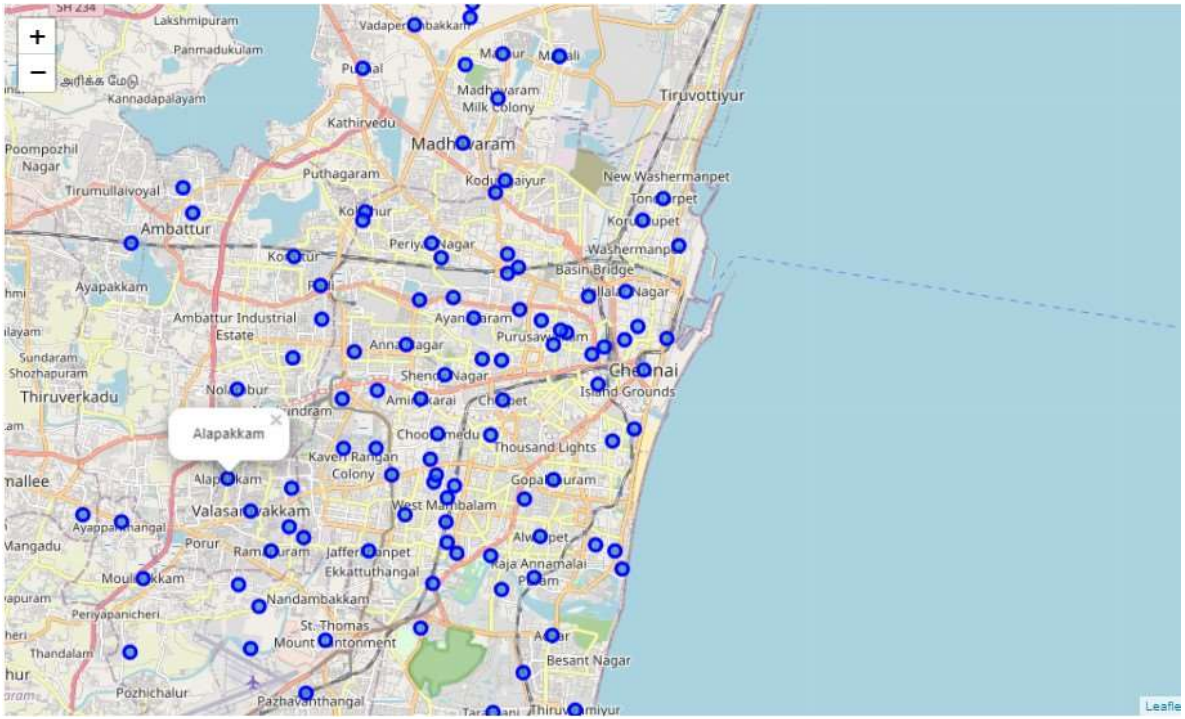
After all venues of each neighbourhood are obtained than a graph is plotted to know which neighbourhood has most venue which shows that Adayar has most of the venues whereas Puzhal has the least.



Methodology

Folium :

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the leaflet.js library. All cluster visualization is done with the help of Folium which in turn generates a Leaflet map made using OpenStreetMap technology.



Methodology

One hot encoding : One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means Clustering Algorithm, all unique items under Venue Category are one-hot encoded.

Top 10 most common venues : Due to high variability in the venues, only the top 10 common venues are selected and a new DataFrame is made, which is used to train the K-means Clustering.

In [21]:
```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)

    return row_categories_sorted.index.values[0:num_top_venues]
```

In [22]:
```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighbourhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
df_venues_sorted = pd.DataFrame(columns=columns)
df_venues_sorted['Neighbourhood'] = chennai_grouped['Neighbourhood']

for ind in np.arange(chennai_grouped.shape[0]):
    df_venues_sorted.iloc[ind, 1:] = return_most_common_venues(chennai_grouped.iloc[ind, :], num_top_venues)

df_venues_sorted.head()
```

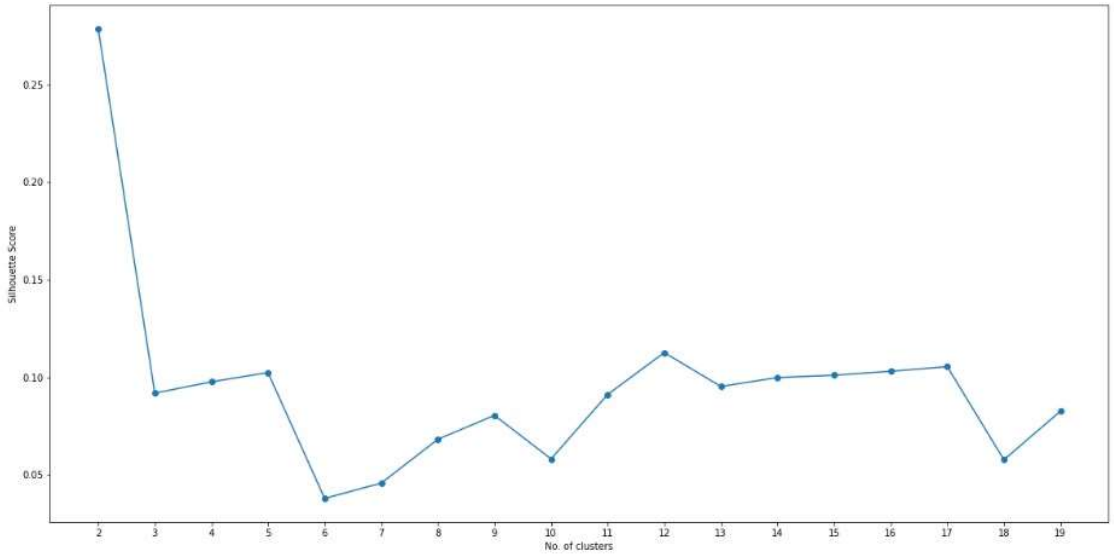
Out[22]:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adambakkam	Indian Restaurant	Restaurant	Dessert Shop	IT Services	Convenience Store	Cosmetics Shop	Cricket Ground	Daycare	Department Store	Fried Chicken Joint
1	Adyar, Chennai	Indian Restaurant	Electronics Store	North Indian Restaurant	Juice Bar	Movie Theater	Café	Bookstore	Snack Place	Fast Food Restaurant	Grocery Store
2	Alandur	Airport Service	Yoga Studio	Farmers Market	Frozen Yogurt Shop	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Event Space
3	Alapakkam	Indian Restaurant	Fast Food Restaurant	ATM	Airport Service	Gift Shop	Daycare	Department Store	Dessert Shop	Diner	Electronics Store
4	Alwarpet	Japanese Restaurant	Coffee Shop	Bakery	Restaurant	Pharmacy	Chinese Restaurant	Fast Food Restaurant	Café	Breakfast Spot	Sandwich Place

Methodology

The optimal number of clusters :

Silhouette Score is a measure of how similar an object is to its cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its cluster and poorly matched to neighbouring clusters. Based on the Silhouette Score of various clusters below 20, the optimal cluster size is determined. Here we have chosen k=12 taking second-highest silhouette value so that the data gets divided into more groups having similar features which were not possible in k=2.



Methodology

K-Means Clustering :

The venue data is then trained using K-means Clustering Algorithm to get the desired clusters to base the analysis on. K-means was chosen as the variables (Venue Categories) are huge, and in such situations, K-means will be computationally faster than other clustering algorithms.

In [42]:
```
kclusters = 12 #taking second highest silhoutte value so that the data gets divided into more groups having similar features whi

# Run k-means clustering
c = chennai_grouped_clustering
kmeans = KMeans(n_clusters = kclusters, init = 'k-means++', random_state = 0).fit(c)
```

In [45]:
```
#df_venues_sorted.drop(['Cluster Labels'],axis=1,inplace=True)
df_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
```

Analysis

Analyse each of the clusters to identify the characteristics of each cluster and the neighborhoods in them.

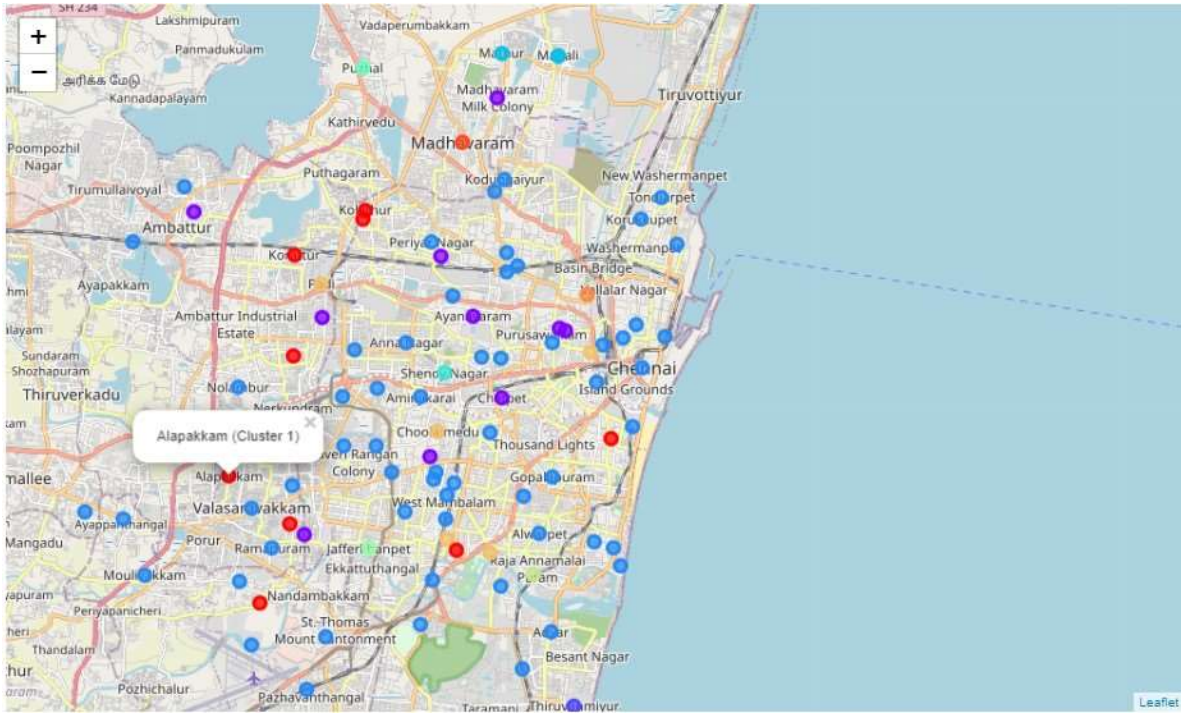
In [46]:
```
chennai_merged = df
chennai_merged = chennai_merged.join(df_venues_sorted.set_index('Neighbourhood'), on='Neighbourhood')
chennai_merged.dropna(inplace = True)
chennai_merged['Cluster Labels'] = chennai_merged['Cluster Labels'].astype(int)
chennai_merged.head()
```

Out[46]:

	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	1(C
0	Adambakkam	12.982221	80.209121	3	Indian Restaurant	Restaurant	Dessert Shop	IT Services	Convenience Store	Cosmetics Shop	Cricket Ground	Daycare	Department Store	
1	Adyar, Chennai	13.006450	80.257779	3	Indian Restaurant	Electronics Store	North Indian Restaurant	Juice Bar	Movie Theater	Café	Bookstore	Snack Place	Fast Food Restaurant	
2	Alandur	13.002822	80.171919	3	Airport Service	Yoga Studio	Farmers Market	Frozen Yogurt Shop	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	
3	Alapakkam	13.049901	80.165435	0	Indian Restaurant	Fast Food Restaurant	ATM	Airport Service	Gift Shop	Daycare	Department Store	Dessert Shop	Diner	El
4	Alwarpet	13.033860	80.254549	3	Japanese Restaurant	Coffee Shop	Bakery	Restaurant	Pharmacy	Chinese Restaurant	Fast Food Restaurant	Café	Breakfast Spot	S

Results

The neighbourhood is divided into n clusters where n is the number of clusters found using the optimal approach. The clustered neighbourhood are visualized using different colours to make them distinguishable



Discussion

After analyzing the various clusters produced by the Machine learning algorithm, Cluster 5, is a good option for the vendors who sell flowers as in these areas there are no flower shops nearby and Event space is the most common venues and mostly in all the events, they need flowers. Hence flower shops near Karanodai, Kundrathur, Manali New Town and many other places in cluster 5 can be profitable.

Cluster 5

```
In [52]: chennai_merged.loc[chennai_merged['Cluster Labels'] == 4, chennai_merged.columns[[0] + np.arange(4, chennai_merged.shape[1]).tol
```

Out[52]:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
42	Karanodai	ATM	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store	Furniture / Home Store
60	Kundrathur	ATM	Pharmacy	Bus Station	Event Space	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store
67	Manali New Town, Chennai	ATM	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store	Furniture / Home Store
68	Manali, Chennai	ATM	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store	Furniture / Home Store
70	Mathur, Chennai	ATM	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store	Furniture / Home Store
90	Pallikaranai	ATM	Other Repair Shop	Cosmetics Shop	Cricket Ground	Daycare	Department Store	Dessert Shop	Diner	Frozen Yogurt Shop	Electronics Store

Discussion

Cluster 9 is also a good option as it contains train station as a common venue and in a train station there is a lot of public movement also there are no flower shops and food stalls nearby so vendors can open their food stall here as most of the people get hungry while reaching the station hence the vendors can earn a large profit. Hence opening shops at Enore, Greenways road, kathivakkam and minjur will be beneficial.

Cluster 9

```
In [56]: chennai_merged.loc[chennai_merged['Cluster Labels'] == 8, chennai_merged.columns[[0] + np.arange(4, chennai_merged.shape[1]).tol
```

Out[56]:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
26	Ennore	Train Station	ATM	Art Gallery	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market
32	Greenways Road	Train Station	Yoga Studio	Rock Club	Dessert Shop	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Event Space
44	Kathivakkam	ATM	Train Station	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store
73	Minjur	Train Station	Scenic Lookout	Yoga Studio	Event Space	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store

Discussion

Cluster 1 is the best place to open a grocery store or a vegetable store as there are lot of Indian restaurants near these areas they will always need vegetables and grocery hence by opening a stall in Alapakkam, Karapakkam, Manapakkam and many more areas near cluster 1 will help street vendors gain more profit.

Cluster 1

```
In [48]: chennai_merged.loc[chennai_merged['Cluster Labels'] == 0, chennai_merged.columns[[0] + np.arange(4, chennai_merged.shape[1]).tol
```

Out[48]:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	Alapakkam	Indian Restaurant	Fast Food Restaurant	ATM	Airport Service	Gift Shop	Daycare	Department Store	Dessert Shop	Diner	Electronics Store
43	Karapakkam	Indian Restaurant	Shopping Plaza	Bakery	Frozen Yogurt Shop	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Event Space
51	Kolathur, Chennai	Indian Restaurant	Electronics Store	Department Store	Cosmetics Shop	Cricket Ground	Daycare	Dessert Shop	Diner	Furniture / Home Store	Event Space
52	Korattur	Indian Restaurant	Pharmacy	Dessert Shop	Campground	Event Space	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market
69	Manapakkam	Indian Restaurant	Juice Bar	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Yoga Studio
71	Medavakkam	Indian Restaurant	Vegetarian / Vegan Restaurant	Electronics Store	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Event Space	Diner
74	Mogappair	Indian Restaurant	Bus Station	Sandwich Place	Convenience Store	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Event Space
106	Rajakilpakkam	Indian Restaurant	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store	Furniture / Home Store
117	Senthil Nagar	Indian Restaurant	Department Store	Electronics Store	Cricket Ground	Daycare	Cosmetics Shop	Dessert Shop	Diner	Furniture / Home Store	Event Space
121	Srinivasapuram, Chennai	Indian Restaurant	Italian Restaurant	Coffee Shop	Resort	Yoga Studio	Event Space	Food Court	Food	Flower Shop	Fast Food Restaurant
124	T. Nagar	Indian Restaurant	Playground	Electronics Store	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Event Space	Diner
128	Thirumalai nagar	Indian Restaurant	Event Space	Fried Chicken Joint	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store	Furniture / Home Store
136	Triplicane	Indian Restaurant	Vegetarian / Vegan Restaurant	Music Store	Event Space	Food Court	Food	Flower Shop	Fast Food Restaurant	Farmers Market	Electronics Store
143	Vanagaram	Indian Restaurant	ATM	Rest Area	Pizza Place	Hotel Bar	Food	IT Services	Convenience Store	Cosmetics Shop	Cricket Ground

Conclusion

This project helps a person get a better understanding of the neighbourhoods with respect to the most common venues in that neighbourhood.

It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a neighbourhood.

Also, the small scale vendors after knowing the right place to do their business will make an immense profit, which will lead to a better lifestyle of the vendors and hence they could help the Indian economy in one way or the other.

Hence the GDP of India will increase and if it goes on increasing than India will become a developed country soon.

The prospect of this project is to make the algorithm better and efficient so that it could analyze the whole country and it could help all the category of people.

Thank You