The Battle of Neighbourhoods
Where To Open Bakery?

1- Business Problem

The Objective of this Capstone Project is to located the best Location to open new Bakery Outlet in City New Delhi, India. Using Data Science and Machine Learning Technique, this Project aim to answer this Business problem. With this help of Machine learning we are able to sorted out the location which is well connected to Airport/Bus Station and other public places. The Solution recommend best neighbourhood to open the Bakery.

2- Data Description

The data set that I have used for solving the problem is:

- A complete list of neighborhoods in New Delhi, India. Source of the data is Wikipedia.org (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delhi))
- Geographical coordinates (latitude and longitude) of those neighborhoods. Source of the data will be FourSquare.
- FourSquare provided Venue data which is related to Bakery. Machine Learning Technique called Clustering will be used for solving the problem.
- Clustering using K-Means clustering algorithm

3- Methodology

Methodology for finding a suitable location for opening a new Bakery in Delhi, India is based on Clustering of venues and places in Neighbourhoods of Delhi. I have grouped the similar venues together on the basis of availability of venues of the different categories. I have used Machine Learning technique called Clustering for the analysis of venues and places of different categories in the Neighbourhoods of Delhi. First step is pulling and preprocessing of the data. In have used web scraping python library for pulling the data from wikipedia pages and then I have performed the preprocessing of the data using python data munging library called pandas. The data received from the wikipedia was in JSON format so I have to use JSON parsing for extracting the different parameters of the data. Second step is all about Exploratory Data Analysis, where I have used few statistical measures and data visualization techniques for understanding the internal structure of the data and the relationship between different parameters of the data. Exploratory Data Analysis has helped me to decide which machine learning technique will be suitable for solving the business problem. I have also used map visualization library (folium) and python library geocoder for fetching the geographical coordinates of different venues and places. I have used Machine Learning technique called Clustering. Clustering unsupervised machine learning where have unlabelled dataset.

3.1 Importing

```
import numpy as np
import pandas as pd

import geocoder # to get coordinates
import requests # library to handle requests
import matplotlib.cm as cm
import matplotlib.colors as colors
from sklearn.cluster import KMeans
import folium # map rendering library
```

Package breakdown:

- *Pandas* : To collect and manipulate data in JSON and HTMl and then data analysis
- *requests* : Handle http requests
- *matplotlib* : Detailing the generated maps
- *folium* : Generating maps of London and Paris
- *sklearn* : To import Kmeans which is the machine learning model that we are using.

3.2- Exploring New Delhi

Neighbourhoods of Delhi. We begin to start collecting and refining the data needed for the our business solution to work.

```
url_delhi= requests.get("https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Delh
i").text
```

Here is a data table consisting of initial 5 neighborhood and their geographical coordinates:

	Neighborhood
0	Ashok Nagar (Delhi)
1	Ashok Vihar
2	Ashram Chowk
3	Babarpur
4	Badarpur, Delhi

```
Total no of neighborhoods: 152
```

3.3- Feature Selection & Engineering

```
df_delhi.shape
 (152, 1)
```

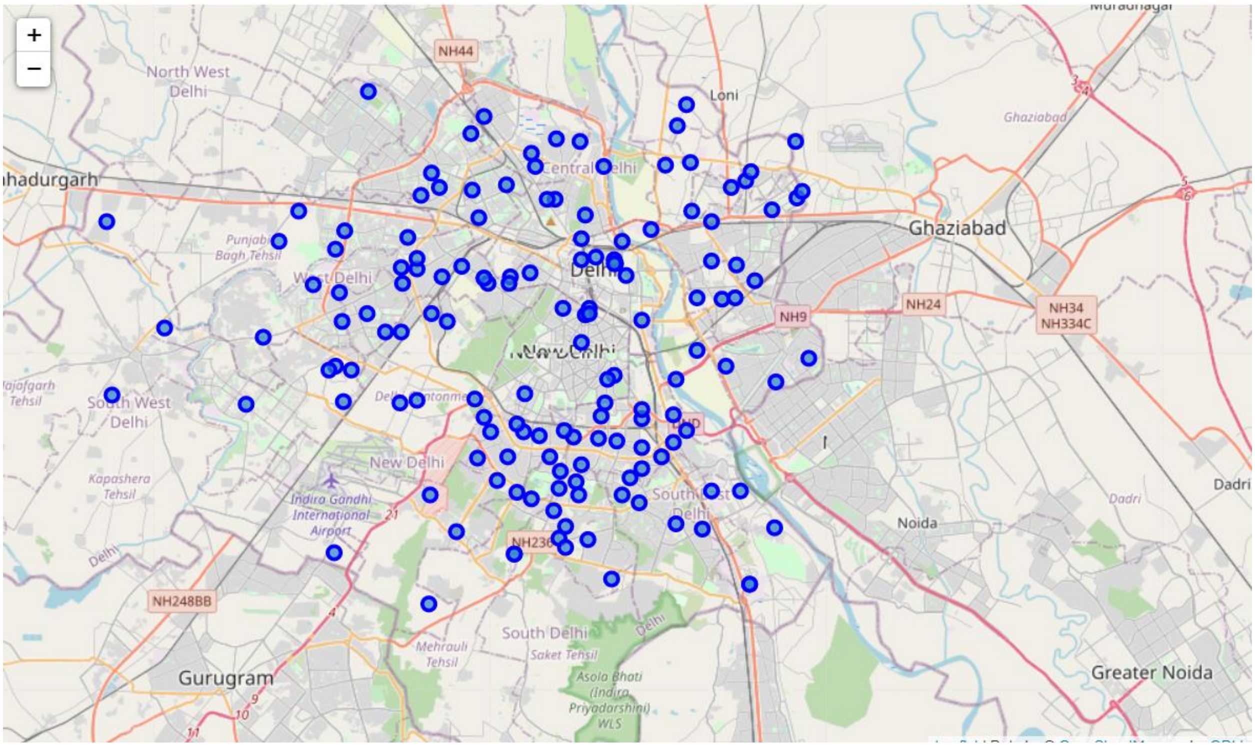
```
df_delhi.dtypes
 Neighborhood
object dtype: object
```

Geolocation of New Delhi Neighborhood

	Neighborhood	Latitude	Longitude
0	Ashok Nagar (Delhi)	28.692230	77.301240
1	Ashok Vihar	28.690370	77.176090
2	Ashram Chowk	28.710598	77.326965
3	Babarpur	28.507380	77.303460
4	Badarpur, Delhi	28.507380	77.303460

After geographical coordinates of all the venues and places in the neighborhood of Delhi, I have plot the places data on the map of Delhi. I have used geocoder and folium python libraries for visualizing the places data on the map:

Map Visualization - Neighborhood of Delhi



```
The coordinates of Delhi are 28.6273928, 77.1716954.
```

The approach taken here is to explore the city, plot the map to show the neighbourhoods being considered and then build our model by clustering all of the similar neighbourhoods together and finally plot the new map with the clustered neighbourhoods. We draw insights and then compare and discuss our findings.

3.4- FourSquare APIs

We will need data about different venues in different neighbourhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

```
neighborhood_latitude = df_delhi.loc[0, 'Latitude'] # neighborhood latitude value
neighborhood_longitude = df_delhi.loc[0, 'Longitude'] # neighborhood longitude value

neighborhood_name = df_delhi.loc[0, 'Neighborhood'] # neighborhood name

print('Latitude and longitude values of {} are {}, {}.'.format(neighborhood_name,
                                                               neighborhood_latitude,
                                                               neighborhood_longitude)
```

```
Latitude and longitude values of Ashok Nagar (Delhi) are 28.69223000000005
2, 77.30124000000006.
```

Pull the nearby venues

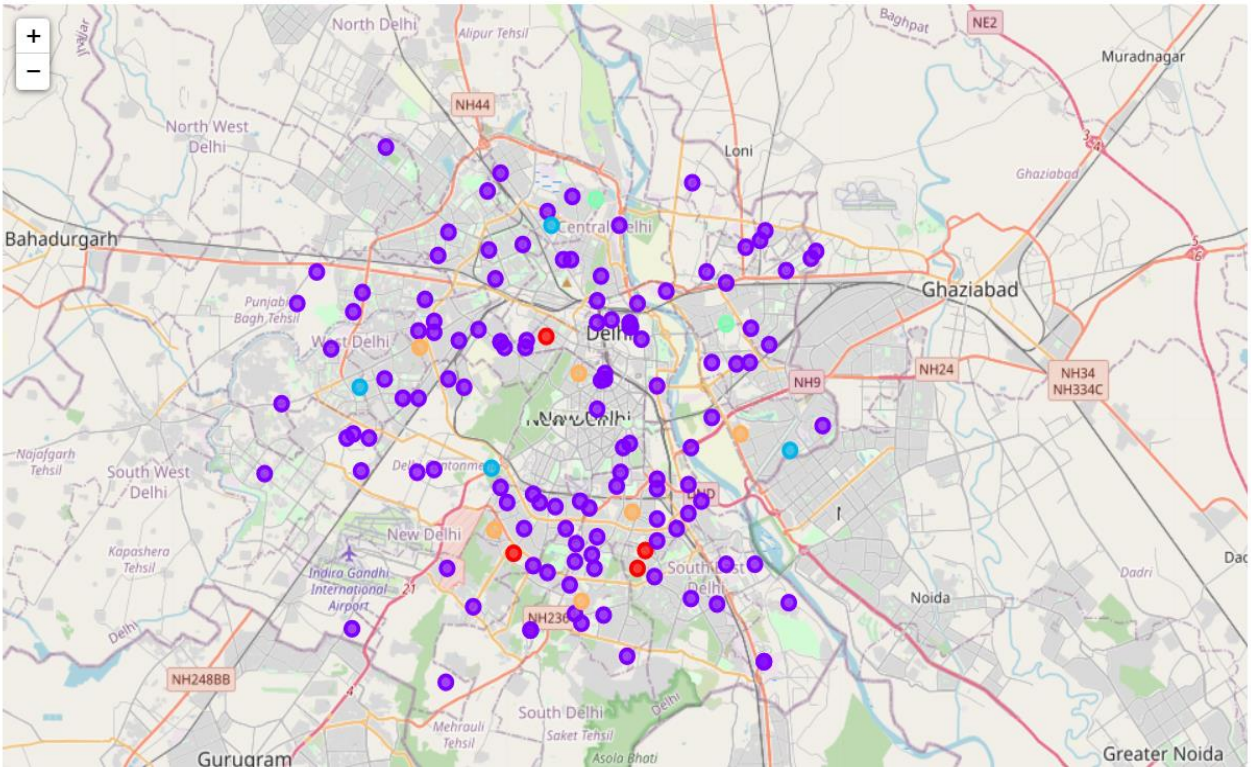
	name	categories	lat	lng
0	Neeraj Kumar Garg	Spa	28.692731	77.298772
1	Axis Bank ATM	ATM	28.696470	77.299910
2	Sutta Chowk	Smoke Shop	28.697897	77.300010
3	My Idea Store	Mobile Phone Shop	28.686405	77.299520
4	Axis Bank ATM	ATM	28.694360	77.308370

3.5- Clustering using K-Means clustering algorithm

I have used KMeans Clustering algorithm for understanding internal complexity of the data and cluster the similar places and venues together. Total number of clusters are 5. Here is the data table after applying KMeans Clustering:

	Neighborhood	Bakery	Cluster Label	Latitude	Longitude
0	Ashok Nagar (Delhi)	0.0	1	28.69223	77.301240
1	Ashok Vihar	0.0	1	28.69037	77.176090
2	Babarpur	0.0	1	28.50738	77.303460
3	Badarpur, Delhi	0.0	1	28.50738	77.303460
4	Bali Nagar	0.0	1	28.65223	77.129411

Map Visualization - Neighborhood of New Delhi Cluster wise



3.6- Analyze the Clusters

First Cluster (Cluster Label - 0)

	Neighborhood	Bakery	Cluster Label	Latitude	Longitude
72	Munirka	0.066667	0	28.55504	77.17132
30	Greater Kailash	0.052632	0	28.54849	77.23667
45	Karol Bagh	0.062500	0	28.65045	77.18873
42	Kailash Colony	0.058824	0	28.55613	77.24060

Second Cluster (Cluster Label – 1)

	Neighborhood	Bakery	Cluster Label	Latitude	Longitude
0	Ashok Nagar (Delhi)	0.000000	1	28.692230	77.301240
90	Pamposh Enclave	0.000000	1	28.544430	77.245650
91	Pandav Nagar	0.000000	1	28.614580	77.275740
92	Paschim Vihar	0.000000	1	28.669330	77.091730
93	Patel Nagar	0.000000	1	28.647830	77.164490
94	Pitam Pura	0.000000	1	28.695900	77.137250
95	Preet Vihar	0.000000	1	28.639030	77.295970
96	Punjabi Bagh	0.000000	1	28.666340	77.125000
97	Raisina Hill	0.000000	1	28.618400	77.215481
98	Rajendra Nagar, Delhi	0.000000	1	28.590750	77.227490

Third Cluster (Cluster Label - 2)

	Neighborhood	Bakery	Cluster Label	Latitude	Longitude
17	Derawal Nagar	0.200000	2	28.699110	77.19105
39	Janakpuri	0.200000	2	28.627910	77.09060
18	Dhaula Kuan	0.166667	2	28.592378	77.15948
134	Vasundhara Enclave	0.250000	2	28.600150	77.31663

Fourth Cluster (Cluster Label - 3)

	Neighborhood	Bakery	Cluster Label	Latitude	Longitude
71	Mukherjee Nagar	0.333333	3	28.71053	77.21440
52	Krishna Nagar, Delhi	0.333333	3	28.65545	77.28336

Fifth Cluster (Cluster Label - 4)

	Neighborhood	Bakery	Cluster Label	Latitude	Longitude
27	Gole Market	0.090909	4	28.634100	77.20569
15	Defence Colony	0.086957	4	28.572980	77.23357
100	Rajouri Garden	0.095238	4	28.645620	77.12209
62	Malviya Nagar (Delhi)	0.090909	4	28.533940	77.20702
64	Mayur Vihar	0.090909	4	28.607714	77.29067
133	Vasant Vihar, Delhi	0.125000	4	28.564940	77.16131

4- Result and Discussion

After careful analysis of all five clusters, It's clear that the Places which are part of Second cluster (Cluster Label - 1) are most suitable for opening a new Bakery.

Second cluster (Cluster Label - 1) has least no of existing Bakery and at the same time places in Cluster Label-1 are well connected to Airport/Railway Station/Bus Station and other popular public places. Cluster Label-1 suggests better business locations to open new Bakery.

So, This analysis suggests to open a new Bakery in the Second cluster (Cluster Label - 1)

5- Conclusion

The purpose of this project was to explore the city of New Delhi and to find the best suitable places to open a new Bakery. We explored the city based on their neighbourhood and there famous venues present in each of the neighbourhoods finally concluding with clustering similar neighbourhoods together.

After careful analysis of all five clusters, It's clear that the Places which are part of Second cluster (Cluster Label - 1) are most suitable for opening a new Bakery. Second cluster has the least no of existing Bakery but It has very good connectivity to other popular public places. This analysis suggests to open a new Bakery in the location mentioned in Second cluster (Cluster Label - 1).