Capstone Project – Predicting the Best Place to Open a Restaurant in Minneapolis MN, USA

1. Introduction:

The project focuses on a scenario to open a restaurant. The vicinity selected here is Minneapolis. Minneapolis is the largest and major city in Minnesota that form the "Twin Cities". Minneapolis is the county seat of Hennepin County. This project predicts the best place to open an Indian Cuisine Restaurant in the largest city Minneapolis. This helps the entrepreneur to decide the location which is suitable for their business.

2. Business Problem:

Opening a Restaurant is always a challenge and it needs lots of research and study about the location. Also, the location which we select must be appropriate based on the cuisine and location which is completely based on the demographics. It is difficult for the entrepreneur to explore the location and select the suitable ones. So, the business problem for the entrepreneur is: Which is the best place for the restaurant to be opened in Minneapolis?

3. Target Audience:

The entrepreneur or the owner of the restaurant who is ready to open an Indian Cuisine Restaurant and looking for a location.

4. Data Description:

Following Data and Clean-up activities will be used for the project:

- For getting the details about the neighborhood data, I have used open source dataset from the below website http://opendata.minneapolismn.gov/datasets. I downloaded CSV file from the open data and cleaned it as per the requirement. Additionally, the data also contains information about the restaurants, food carts, food trucks and

café. This will further help in refining the data and will also give more accuracy on the result. So, the final CSV file contains neighborhoods, restaurants, address, latitude, and longitude data.

- Additionally, geocoder api will be used to get the latitude and longitude for the neighborhoods that does not have a lat /long assigned to it.
- Four square API will also be used further to get more information about the venues around the neighborhood, even though I have the restaurant data this will further help in getting all the latest information, more the data better the results.
- The open data set also contains certain demographic data about each zipcode in Minneapolis, this can be used later for future improvements.

5. Methodology:

Post data collection, removal of invalid content from the data is required, one of those is the NaN. The dropna function was useful to do this and the index was reset using the resetindex function. Final validated data contains Neighborhood, Zip code, Latitude and Longitude.

	ZipCode	Neighborhood	Latitude	Longitude
0	55414.0	Marcy Holmes	44.98097	-93.23602
1	55455.0	University of Minnesota	44.97302	-93.23473
2	55402.0	Downtown West	44.97592	-93.27444
3	55402.0	Downtown West	44.97714	-93.27440
4	55408.0	South Uptown	44.94305	-93.29034
5	55403.0	Downtown West	44.97662	-93.27574
6	55404.0	Whittier	44.95633	-93.27747
7	55410.0	Fulton	44.91273	-93.32741
8	55407.0	Ericsson	44.91756	-93.24690
9	55406.0	Linden Hills	44.93433	-93.31729
10	NaN	South Uptown	44.94743	-93.29732
11	55406.0	Cooper	44.94808	-93.21430

```
df_data_neighborhood.head(12)
```

0]:

	ZipCode	Neighborhood	Latitude	Longitude
0	55414.0	Marcy Holmes	44.98097	-93.23602
1	55455.0	University of Minnesota	44.97302	-93.23473
2	55402.0	Downtown West	44.97592	-93.27444
3	55402.0	Downtown West	44.97714	-93.27440
4	55408.0	South Uptown	44.94305	-93.29034
5	55403.0	Downtown West	44.97662	-93.27574
6	55404.0	Whittier	44.95633	-93.27747
7	55410.0	Fulton	44.91273	-93.32741
8	55407.0	Ericsson	44.91756	-93.24690
9	55406.0	Linden Hills	44.93433	-93.31729
10	55406.0	Cooper	44.94808	-93.21430
11	55417.0	Keewaydin	44.91231	-93.22360

Grouping the data will help us on to know the neighborhoods better from a zipcode standpoint. To do this the helpful Groupby function has been implemented. Grouping is the vital part of the initial step in process.

```
df_final.head(20)
```

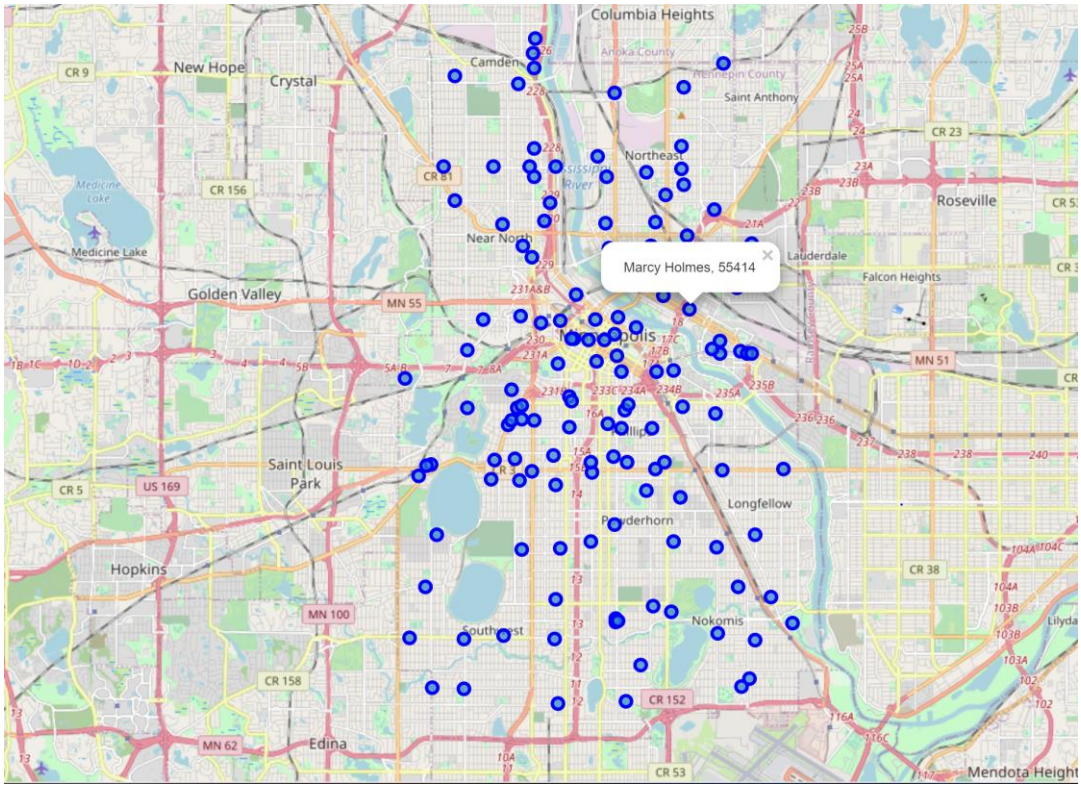
17]:

	Neighborhood	ZipCode	Latitude	Longitude
0	Armatage	55410	44.901680	-93.318473
1	Armatage	55419	44.901263	-93.308797
2	Audubon Park	55418	45.017782	-93.242893
3	Beltrami	55413	44.998570	-93.241360
4	Beltrami	55414	44.991820	-93.243480
5	Bottineau	55418	45.011262	-93.265803
6	Bryant	55409	44.932950	-93.270500
7	Bryn - Mawr	55405	44.973970	-93.308077
8	Bryn - Mawr	55416	44.967940	-93.326950
9	CARAG	55408	0.000000	0.000000
10	Camden Industrial	55412	45.034480	-93.287600
11	Cedar - Isles - Dean	55406	44.949460	-93.318930
12	Cedar - Isles - Dean	55416	44.949242	-93.320354
13	Cedar Riverside	55454	44.969602	-93.245397
14	Central	55407	44.936726	-93.263181
15	Central	55408	44.947700	-93.270149
16	Cleveland	55411	45.013450	-93.314980
17	Columbia Park	55418	45.029143	-93.263213
18	Como	55413	44.990730	-93.224030
19	Como	55414	44.987392	-93.226092

Geocoder API was used for validating the latitude and longitude values which was taken from the opendata(data source). This validation helped us validate and also correct the lat/long information.

5.1. Visualization:

Next step is to visualize which the neighborhood, this would give us a better idea on how the areas are spread across based on the restaurant information that we gathered. Python folium was used to visualize the Minneapolis neighborhood using latitude and longitude values.



5.2. Retrieval of Venues using Four Square:

Next step in the process was to find different venues around the neighborhood specific to our search criteria “Restaurants”. This will help us cluster/explore the data and helps in accuracy as well. Four Square API was used to explore the Minneapolis Neighborhoods with our filter condition. Also, it was limited to 100 venues and the radius within 500 meters was provided. It was further filtered down to just Indian Restaurants in the list of neighborhoods.

Next step is each Minneapolis neighborhood has been analyzed. Then the mean on the frequency of occurrence of each venue category was calculated. This is to prepare data for clustering.

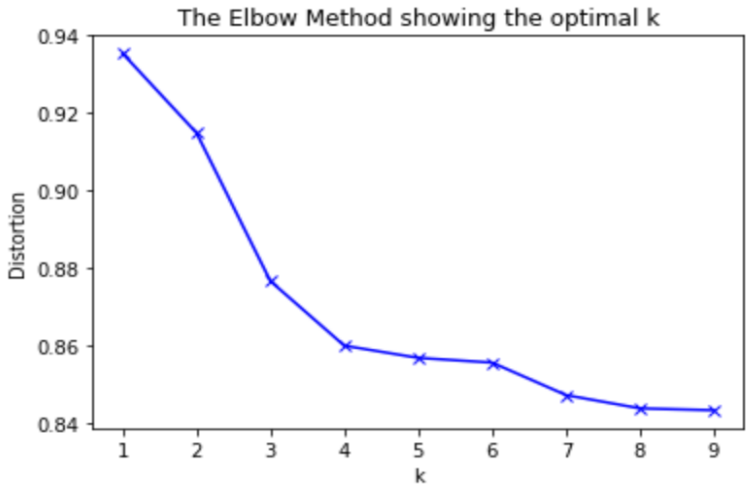
	Neighborhoods	African Restaurant	American Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bistro	Breakfast Spot	Burger Joint	...	Sushi Restaurant	Szechuan Restaurant
0	Armatage	0	0	0	0	0	0	0	0	0	...	0	0
1	Armatage	0	0	0	0	0	0	0	0	0	...	0	0
2	Armatage	0	0	0	0	0	0	0	0	0	...	0	0
3	Armatage	0	0	0	0	0	0	0	0	0	...	0	0
4	Armatage	0	0	0	0	0	0	0	0	0	...	0	0

5 rows × 80 columns

5.3. K Means:

The well-known clustering methodology k-means clustering has been used. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well.

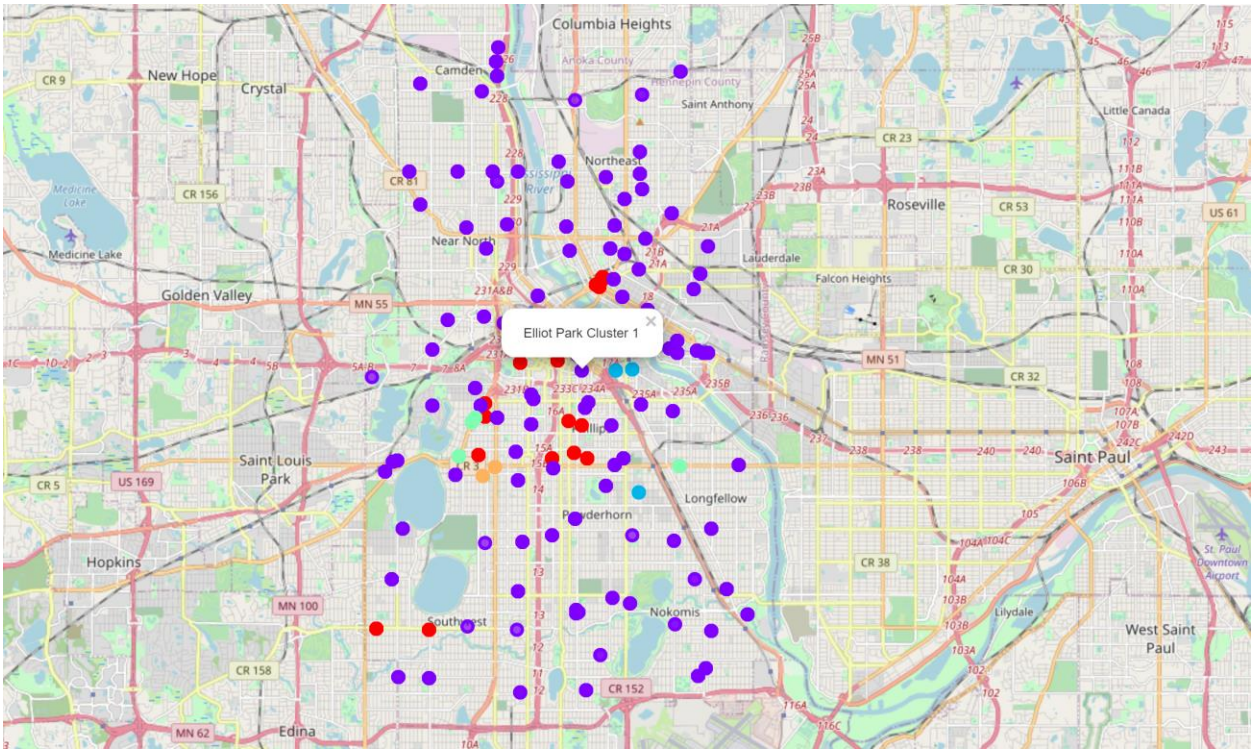
The Elbow method has been used to find the optimal K value. K value has been identified as 5 and this will be used as the number of clusters for the neighborhood. Additionally, this is also based on their frequency of occurrence with respect to the “Indian Restaurants” in the neighborhood.



Based on the elbow method the clustering number was set to 5, K Means algorithm was ran with the set cluster value. The outcome of this clustered Neighborhood gives us more clarity on what type of restaurant each neighborhood has as well as helps us identifying the end goal which is “Indian Restaurant”. Next up sorting of the clustered data based on the Cluster Labels (0 to 4).

5.4. Visualization of Clustered Data:

Initially folium was used to look at the neighborhood, now they have been used to visualize the clustered data.



6. Results:

The results from K-means clustering show that we can categorize Minneapolis neighborhoods into 5 clusters based on how many Indian Restaurants are in each neighborhood:

Cluster 0: Neighborhood has 19 Indian Restaurants.

Cluster 1: Neighborhoods with 1 Indian Restaurant.

Cluster 2: Neighborhood has 3 Indian Restaurants.

Cluster 3: Neighborhood has 10 Indian Restaurants.

Cluster 4: Neighborhood has 3 Indian Restaurants.

	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
47	Marcy Holmes	0.012195	1	44.98947	-93.2513	Gorkha Palace	44.989268	-93.255184	Indian Restaurant

7. Discussion:

Based on the clustered data, the good place to open an Indian Restaurant would be on cluster 1 because there is only one of them. So, the chances of the business picking up would be at a highly relative rate. The Cluster 1 consists of several different neighborhood, as part of this effort we did not zero down to one single neighborhood. This is more of a generalized and has several different options. The main thing for an Indian Restaurant how authentic it is, also the taste and with the location information that we provided this spot will be a hot spot.

8. Conclusion:

Using data retrieval, data cleansing, visualization with maps, clustering using ML algorithm we have concluded that this is an effective way for any sort of business/retailers to start something new with just pure data and analysis.

9. Future Improvements:

In this project we concluded only one cluster to open a restaurant. In the future this can be further filtered/analyzed with demographic data of each neighborhood to find the accurate neighborhood to open a restaurant based on the cuisine. Also, we could have made this a reusable template as well so that anyone in the future who wants to analyze could plugin the variables and run the complete notebook. Due to lack of data and time at this point we stopped at just finding the right cluster.