

Capstone Project - The Competing Neighborhoods

(WEEK 5)

- **Recommendation to Open New Indian Hyderabadi Restaurant in New York**

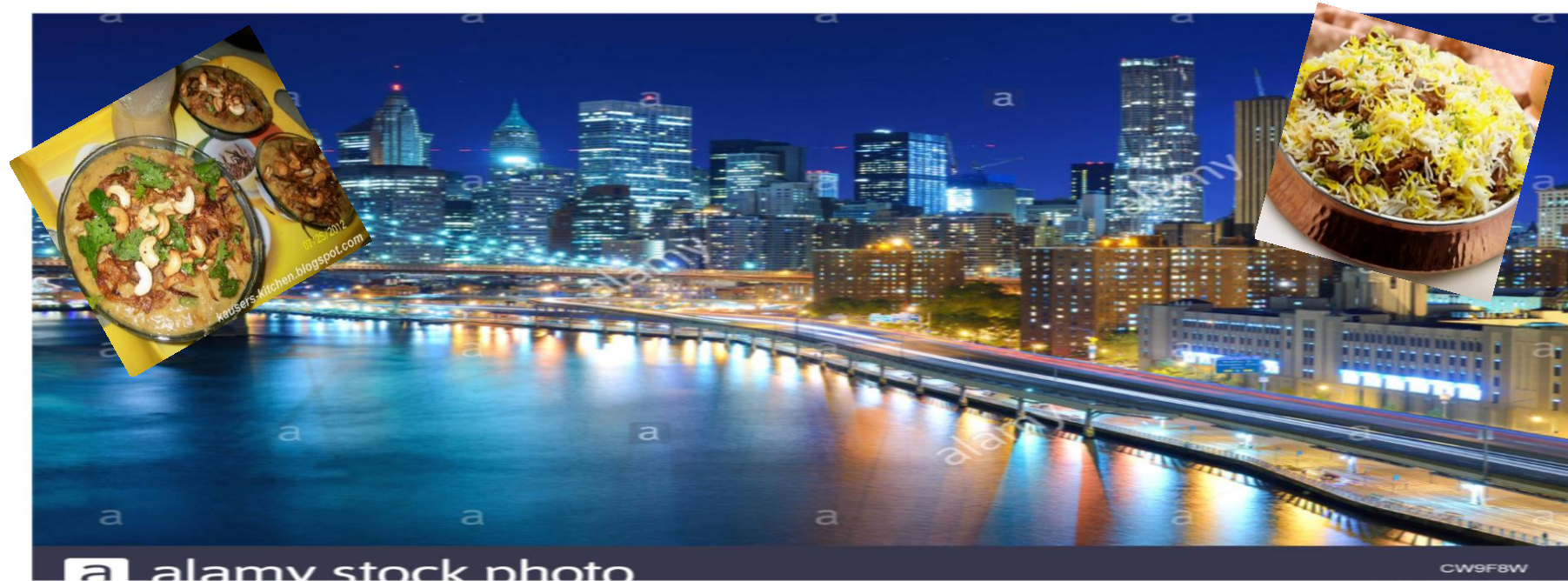


Table of contents

I Introduction: Business Problem

II Data/Dataset

III Methodology

IV Analysis

V Results and Discussion

VI Conclusion



I INTRODUCTION: BUSINESS PROBLEM

- In this capstone project a new restaurant's suitable/profitable/optimal location will be found. Particularly, the stakeholders are interested in **Indian Hyderabdi restaurant in New York, USA**. There are so many restaurants in New York we will choose such a **place/location where there are less restaurants**.
- Also, we also needs to choose such a **location where there are no Indian Hyderabadi restaurants nearby**. Furthermore, we also take care that the **prefer places/locations should be as close as possible to the city center**, along with the first two criteria are met.
- We will recommend the most suitable palces/neighborhoods based above mentioned criteria using the skills we learn during this long journey of Data Science Course. So stakeholders can choose best, suitable, possible area/location as we explain pros and cons each location in clear and easy way.



II DATA/DATASET

- As we defined our problem and based on this, Our decision is influenced by following criteria/factors:
- The number of already existing restaurants(can be of any type) in the area/neighborhood
- If any, how many number of and far to Hyderabadi restaurants.
- From the heart of the city, the neighborhood is how much far away

```
-----  
In [1]: !wget -q -O 'newyork_data.json' https://cocl.us/new_york_dataset  
print('Data downloaded!')
```

Data downloaded!

Load and explore the data

```
In [4]: import json # library to handle JSON files  
  
with open('newyork_data.json') as json_data:  
    newyork_data = json.load(json_data)
```

```
In [5]: newyork_data
```



II DATA/DATASET

- To define our neighborhood/area, we will use regularly spaced grid of locations, centered around heart of the city.
- To extract/generate the required information, below data sources will be required:
- **Foursquare API** to get venue data related to these neighborhoods, number of restaurants & their type and location in every area/neighborhood will be obtained
- To solve this problem, we will need below data:
- List of neighborhoods in New York, USA
- Latitude and Longitude of these neighborhoods
- Venue data related to Indian Hyderabad restaurants. This will help us find neighborhoods that are more suitable to open an Indian Hyderabad Restaurant.



EXTRACTING THE DATA

- The scrapping of New York neighborhoods via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- **Neighborhood Candidates**
- For centroids of our candidate neighborhoods, let's create latitude and longitude coordinates. Around the heart of the New York City, we will create a grid of cells covering our area of interest which is approximately 11x11 kilometers.
- Using specific, well known address and Google Maps geocoding API, let's first find the latitude & longitude of heart of the New York City.



III METHODOLOGY

- First, I need to get the list of neighborhoods in New York, USA. This is possible by extracting the list of neighborhoods from url already available in one of the labs in this course. I did the web scraping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame.
- However, it is only a list of neighborhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder Package but it was not working so I used the CSV file provided by IBM team to match the coordinates of New York neighborhoods.



III METHODOLOGY

- After gathering these coordinates, I visualize the map of New York using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius.
- I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues.
- Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Here, I made a justification to specifically look for “Indian restaurants”.

III METHODOLOGY

```
In [13]: neighborhoods.head()
```

```
Out[13]:
```

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

```
In [14]: print('The dataframe has {} boroughs and {} neighborhoods.'.format(
          len(neighborhoods['Borough'].unique()),
          neighborhoods.shape[0]
        ))
```

The dataframe has 5 boroughs and 306 neighborhoods.

```
In [16]: neighborhoods.rename(columns={'Neighbourhood': 'Neighborhood'}, inplace=True)
```

```
In [17]: neighborhoods.groupby('Borough').count()['Neighborhood']
```

```
Out[17]: Borough
Bronx      52
Brooklyn   70
Manhattan  40
Queens     81
Staten Island 63
Name: Neighborhood, dtype: int64
```



III METHODOLOGY

- Lastly, I performed the clustering method by using k-means clustering. 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.



IV ANALYSIS

In [33]: `newyork_venues.shape`

Out[33]: (10087, 7)

In [34]: `newyork_venues.groupby('Neighborhood').count()`

Out[34]:

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Allerton	31	31	31	31	31	31
Annadale	12	12	12	12	12	12
Arden Heights	4	4	4	4	4	4
Arlington	6	6	6	6	6	6
Arrochar	24	24	24	24	24	24
Arverne	21	21	21	21	21	21
Astoria	99	99	99	99	99	99
Astoria Heights	11	11	11	11	11	11
Auburndale	20	20	20	20	20	20
Bath Beach	48	48	48	48	48	48



IV ANALYSIS

- I have clustered the neighborhoods in New York into 3 clusters based on their frequency of occurrence for “Indian food”.
- Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

V RESULTS AND DISCUSSION

```
In [45]: from sklearn.cluster import KMeans
toclusters = 3

to_clustering = to_indian.drop(["Neighborhoods"], 1)

# run k-means clustering
kmeans = KMeans(n_clusters=toclusters, random_state=1)
kmeans.fit_transform(to_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:20]
```

```
Out[45]: array([0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0],
      dtype=int32)
```

```
In [46]: to_merged = to_indian.copy()

# add clustering labels
to_merged["Cluster Labels"] = kmeans.labels_
```

```
In [49]: to_merged.rename(columns={"Neighborhoods": "Neighborhood"}, inplace=True)
to_merged.head(20)
```

```
Out[49]:
```

	Neighborhood	Indian Restaurant	Cluster Labels
0	Allerton	0.000000	0
1	Annadale	0.000000	0
2	Arden Heights	0.000000	0
3	Arlington	0.000000	0

V RESULTS AND DISCUSSION

```
In [54]: #Cluster 1
to_merged.loc[(to_merged['Cluster Labels'] ==1) & (to_merged['Venue Category'] == 'Indian Restaurant') ]
```

Out[54]:

	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
99	Floral Park	0.375000	1	40.741378	-73.708847	Flavor Of India	40.737425	-73.708540	Indian Restaurant
99	Floral Park	0.375000	1	40.741378	-73.708847	Shahi Darbar	40.737488	-73.710022	Indian Restaurant
141	Jamaica Estates	0.333333	1	40.716805	-73.787227	Dhaka Hajir Biryani	40.720989	-73.786781	Indian Restaurant
99	Floral Park	0.375000	1	40.741378	-73.708847	Namaste Restaurant and Cafe	40.737173	-73.709756	Indian Restaurant

```
In [55]: #Cluster 2
```



VI CONCLUSION

- Most of the Indian restaurants are in cluster 2 which is around Bayside, Astoria, Greenwich Village.
- And lowest in cluster 0 and 3 areas which are in midtown, noho, and north side areas. Also, there are good opportunities to open new "Indian Hyderabad restaurant" in this area.
- So stakeholders can choose this area.

Thank you for review