

# IBM Data Science Capstone Project

## **FINAL REPORT**

By – Nitin Bisht

# Introduction

As you may probably have guessed it by now, I'm an Indian and I wanted to build a project on an Indian location, using what I've learnt during the course. But unfortunately there is not a proper amount of data available on Foursquare to build that project. So I searched online for an API on which I can get a decent amount of data to build a project and I found the Zomato API. Many of you might know that Zomato is an Indian restaurant aggregator and food delivery start-up. Zomato provides information, menus and user-reviews of restaurants as well as food delivery options from partner restaurants in selected cities.

Zomato API is somewhat similar to Foursquare API so I didn't face any major problem in dealing with the data I got as a response to my API-calls. With all this data being available on Restaurants of India, I chose the problem that was given in the suggestions of the project.

- ◆ **Problem Statement:** In a city of your choice, if someone is looking to open a restaurant, where would you recommend that they open it?
  - I chose **Delhi** as the city of my choice.
  
- ◆ **Target Audience:** *The target audience of this project is a restaurant group or an individual who wants to open a Restaurant in Delhi, but they/he couldn't decide where to open the restaurant and what cuisine should they serve.*

# Data Collection

## Data Collection:

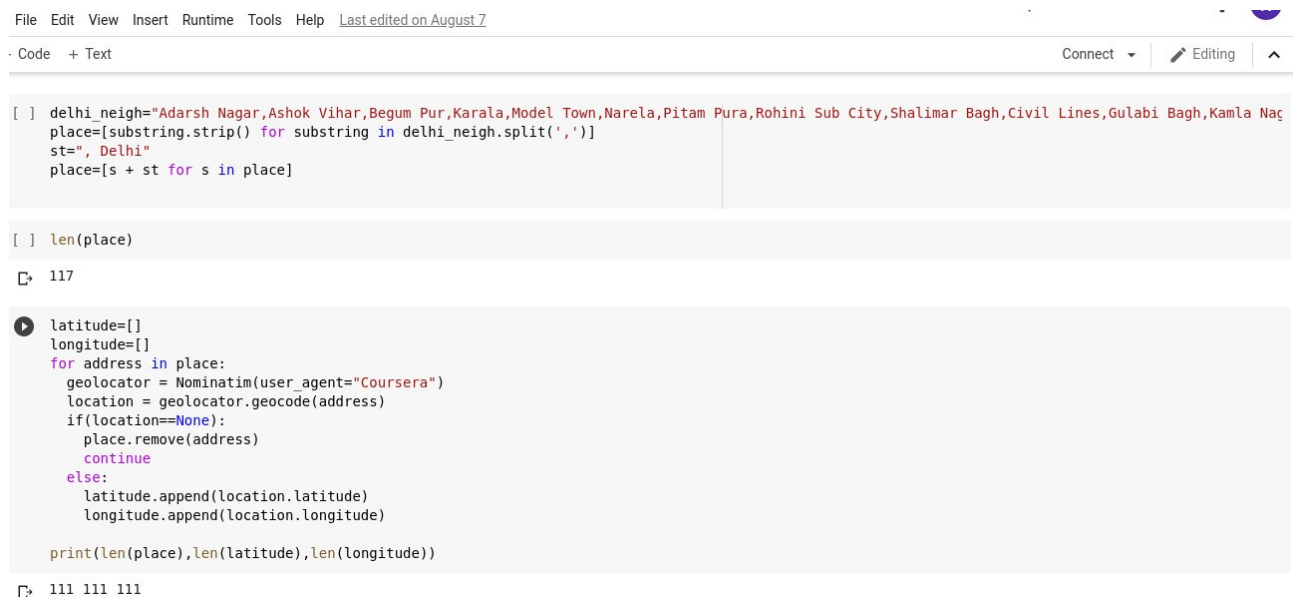
1. So the first step in Data Collection was to get the names and coordinates of the neighbourhoods of Delhi.

Names of the Neighbourhoods of Delhi were found at:  
"[https://en.wikipedia.org/wiki/Neighbourhoods\\_of\\_Delhi](https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi)"

I copied them into a list.

2. Next step was to get the coordinates of the neighbourhoods. I used the Nominatim tool from the geopy.geocoders library to get the latitudinal and longitudinal coordinates of the neighbourhoods.

## Below is the screenshot of the jupyter notebook:



```
File Edit View Insert Runtime Tools Help Last edited on August 7
Code + Text Connect Editing ^

[ ] delhi_neigh="Adarsh Nagar,Ashok Vihar,Begum Pur,Karala,Model Town,Narela,Pitam Pura,Rohini Sub City,Shalimar Bagh,Civil Lines,Gulabi Bagh,Kamla Na
place=[substring.strip() for substring in delhi_neigh.split(',')]
st=", Delhi"
place=[s + st for s in place]

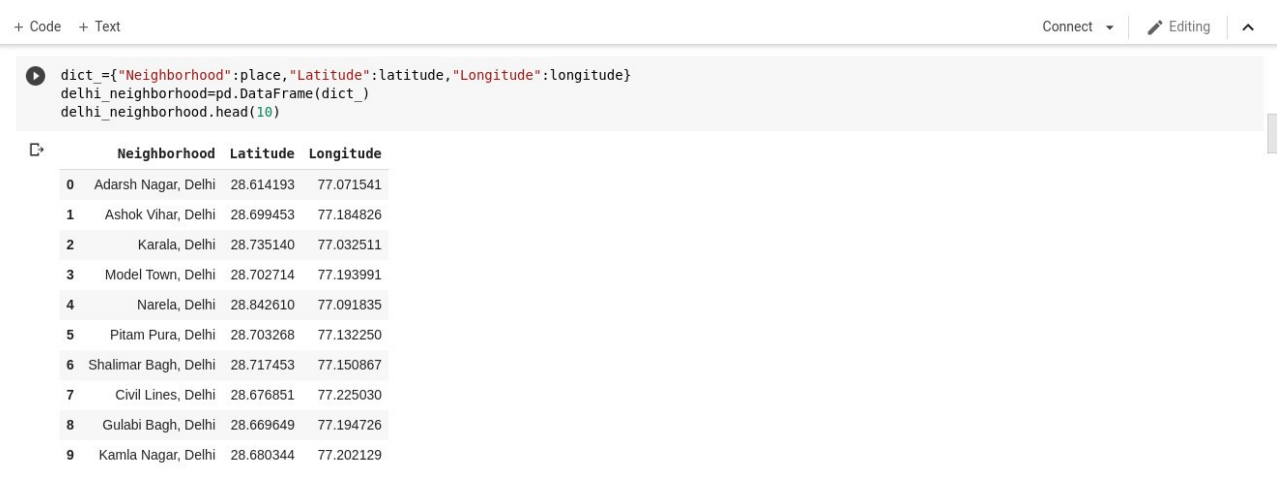
[ ] len(place)
117

latitude=[]
longitude=[]
for address in place:
    geolocator = Nominatim(user_agent="Coursera")
    location = geolocator.geocode(address)
    if(location==None):
        place.remove(address)
        continue
    else:
        latitude.append(location.latitude)
        longitude.append(location.longitude)

print(len(place),len(latitude),len(longitude))
111 111 111
```

I made a data frame using the names and coordinates of the neighbourhoods.

## Below is the screenshot of the jupyter notebook:



```
+ Code + Text Connect Editing ^

dict={"Neighborhood":place,"Latitude":latitude,"Longitude":longitude}
delhi_neighborhood=pd.DataFrame(dict_)
delhi_neighborhood.head(10)

Neighborhood Latitude Longitude
0 Adarsh Nagar, Delhi 28.614193 77.071541
1 Ashok Vihar, Delhi 28.699453 77.184826
2 Karala, Delhi 28.735140 77.032511
3 Model Town, Delhi 28.702714 77.193991
4 Narela, Delhi 28.842610 77.091835
5 Pitam Pura, Delhi 28.703268 77.132250
6 Shalimar Bagh, Delhi 28.717453 77.150867
7 Civil Lines, Delhi 28.676851 77.225030
8 Gulabi Bagh, Delhi 28.669649 77.194726
9 Kamla Nagar, Delhi 28.680344 77.202129
```

3. Next step was to collect data using the Zomato API about various restaurants in the above neighbourhoods. This data was going to be used in solving the problem. The data collected includes:

- id- Restaurant id
- name- Name of Restaurant
- locality- The locality in which the Restaurant was present
- cuisines- The type of foods the Restaurant offer
- votes – No. Of votes that the Restaurant got.

I created a data frame(df) using the collected data.

**Here is the code to do so:**

```
cols=["id","name","locality","cuisines","votes"]
df=pd.DataFrame(columns=cols)

for lat,lon in zip(delhi_neighborhood['Latitude'],delhi_neighborhood['Longitude']):
    base="https://developers.zomato.com/api/v2.1/geocode?lat={}&lon={}".format(lat,lon)
    header={"Accept": "application/json", "user-key": key}
    res=requests.get(base,headers=header)
    result=res.content.decode("utf-8")
    result=json.loads(result)
    tt=result["nearby_restaurants"]
    df1=[]
    for i in range(len(tt)):
        test=tt[i]
        df1.append([test["restaurant"]["id"],
                    test["restaurant"]["name"],
                    test["restaurant"]["location"]["locality"],
                    test["restaurant"]["cuisines"],
                    test["restaurant"]["user_rating"]["votes"]])
    df=df.append(pd.DataFrame(df1,columns=cols))

df=df.drop_duplicates(subset=None,keep='first')
df.id=df.id.astype('int64')
df.votes=df.votes.astype('int64')
df=df.assign(cuisines=df['cuisines'].str.split(',').explode('cuisines'))
df.cuisines=df.cuisines.str.lstrip()
print(df.cuisines.nunique())
print(df.locality.nunique())
df.dtypes
```

**Here is a glimpse of the content of the Data-Frame:**

```
In [10]: df.insert(3,"No Of Res",1)
         df.head()

Out[10]:
```

	id	name	locality	No Of Res	cuisines	votes
0	18810973	Delhi Darbar Biryani Point	Uttam Nagar	1	Biryani	25865
0	18810973	Delhi Darbar Biryani Point	Uttam Nagar	1	Mughlai	25865
1	18358654	Scoops	Uttam Nagar	1	Beverages	13693
1	18358654	Scoops	Uttam Nagar	1	Fast Food	13693
1	18358654	Scoops	Uttam Nagar	1	Desserts	13693

**Usage of the Data:**

The collected data will be used as follows:

- **locality and cuisines:** They will be combined to check how many restaurant offer a specific cuisine in a particular locality.
- **Votes:** Votes will be used to see, what type of cuisine do people like more.(i.e. *Higher the number of votes, higher the number of people who tried the cuisine*).

# Methodology

## Preprocessing the Data:

- As a Restaurant offers a variety of food, there were multiple entries in the cuisines column of each restaurant(check the above screenshot). We need to separate those entries in order to group them by the type of food they offer.
- Before separating the cuisines entries I divided the total votes of the Restaurant by the number of cuisines it offers. I assumed that each cuisine contribute equally in the vote share of a Restaurant. When I split the columns on the basis of cuisines each entry got it's share of votes.

### ▾ Dividing the Votes

The following code is written to divide the votes evenly into cuisines

```
a=df['cuisines'].str.split(',')
b=[len(c) for c in a]
a=pd.Series(b)
df=df.assign(cuisines=df['cuisines'].str.split(',')).explode('cuisines')
df.cuisines=df.cuisines.str.lstrip()
df.votes=df.votes/a
df.votes=df.votes.astype('int64')
df.head()
```

	index	id	name	locality	cuisines	votes
0	0	18810973	Delhi Darbar Biryani Point	Uttam Nagar	Biryani	12936
0	0	18810973	Delhi Darbar Biryani Point	Uttam Nagar	Mughlai	12936
1	1	18358654	Scoops	Uttam Nagar	Beverages	4566
1	1	18358654	Scoops	Uttam Nagar	Fast Food	4566
1	1	18358654	Scoops	Uttam Nagar	Desserts	4566

- Then a column named No Of Res was inserted into the Data Frame to store the number of restaurants that offer a specific cuisine in a particular area. Initially it has a value of 1 for all the rows.

```
[11] df.insert(3,"No Of Res",1)
df.head()
```

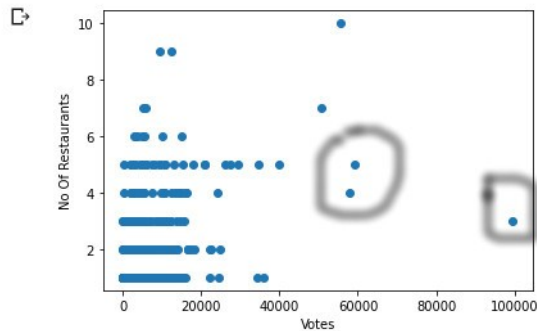
	index	id	name	No Of Res	locality	cuisines	votes
0	0	18810973	Delhi Darbar Biryani Point	1	Uttam Nagar	Biryani	12936
0	0	18810973	Delhi Darbar Biryani Point	1	Uttam Nagar	Mughlai	12936
1	1	18358654	Scoops	1	Uttam Nagar	Beverages	4566
1	1	18358654	Scoops	1	Uttam Nagar	Fast Food	4566
1	1	18358654	Scoops	1	Uttam Nagar	Desserts	4566

- After that a Data Frame(df1) was created using the four columns(i.e. locality, cuisines, No Of Res, and votes) from the first Data Frame(df).
- Then I grouped the Data on the Basis of locality and cuisines. And the values of No Of Restaurants and Votes were aggregated and added on the basis of Groups.

## Visualization and Analysis of Data:

- After the preprocessing of data, a scatter plot was created using the No Of Res and votes columns.

```
plt.scatter(x=df1["votes"],y=df1["No Of Res"])  
plt.xlabel("Votes")  
plt.ylabel("No Of Restaurants")  
plt.show()
```



- As you can see in the above plot, we need the values that are highlighted. because the **No Of Res** values is **Low**(i.e. less competition) and the **votes** value is **High**(i.e. higher amount of customers).

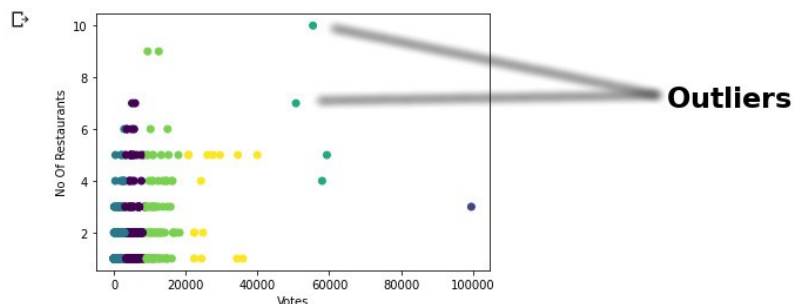
## Machine Learning Used:

- The K-Means clustering approach was used to find the group of the restaurant having low competition and high amount of customers.

The K-Means clustering approach was used because, to solve the given problem we have to find the group of restaurants having similar features(i.e. Low Competition and High amount of Customers).

## Scatter Plot After Clustering:

```
plt.scatter(x=df1["votes"],y=df1["No Of Res"],c=kmeans.labels_)  
plt.xlabel("Votes")  
plt.ylabel("No Of Restaurants")  
plt.show()
```



One limitation of the K-Means algorithm is that it can not detect outliers, and the same thing happened here as well. Two points that we consider as outliers are also present in the cluster we need. So we have to manually eliminate them.

# Result

```
[22] df1.loc[df1['Cluster Label'] == 1]
```

		No Of Res	votes	Cluster Label
locality	cuisines			
Karkardooma	North Indian	3	99563	1

By analysing the data we found that the best locality to open a Restaurant in Delhi is **Karkardooma** as it has a fewer number of Restaurants and a high number of customers. And one should offer the North Indian cuisine. Later on, if he want, he can start selling beverages, Chinese Food and Fast Food as well to grow his target audience because a large number of people like them as well.

One can take a look at the localities and cuisines listed below as well:

```
df1.loc[df1['Cluster Label'] == 3]
```

		No Of Res	votes	Cluster Label
locality	cuisines			
Ashok Vihar Phase 2	North Indian	5	59334	3
IP Extension	North Indian	4	58012	3

## **Discussion**

While Analysing the Data I found that **North Indian cuisine** is the *most popular cuisine* in Delhi. It was due to the facts that Delhi is in the Northern part of India and over the last few decades many people from North India especially from Himachal Pradesh and Uttarakhand have shifted to Delhi for earning livelihood. So the amount of North Indian people who live in Delhi is very high. And it is a fact that North Indian people love North Indian food.

Apart from that people of Delhi love **Chinese Food** and **Fast Food** as well.



# **Conclusion**

- ◆ Being an Indian I wanted to build a project that can be used for solving a business problem in India.
- ◆ I chose a very basic business problem: If someone is looking to open a restaurant in Delhi, where would you recommend that they open it?
- ◆ I chose the Zomato API to collect the data for my project as not much data was available on Foursquare for Indian locations.
- ◆ My analysis was mainly based on 4 values: Locality of the Restaurant, the Cuisines that the Restaurant serves, Number of votes that the Restaurant got, and Number of Restaurants in a particular locality that serves a specific cuisine.
- ◆ While analysing the Data I found that, the locality is best for opening a Restaurant that offers cuisines.
- ◆ I also found that the North Indian cuisine was the most famous cuisine in Delhi followed by Chinese and Fast Food.