IBM Data Science

Capstone Project

BUSINESS PROBLEM
(Data Collection)

By – Nitin Bisht

Data Collection

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.

Data Collection:

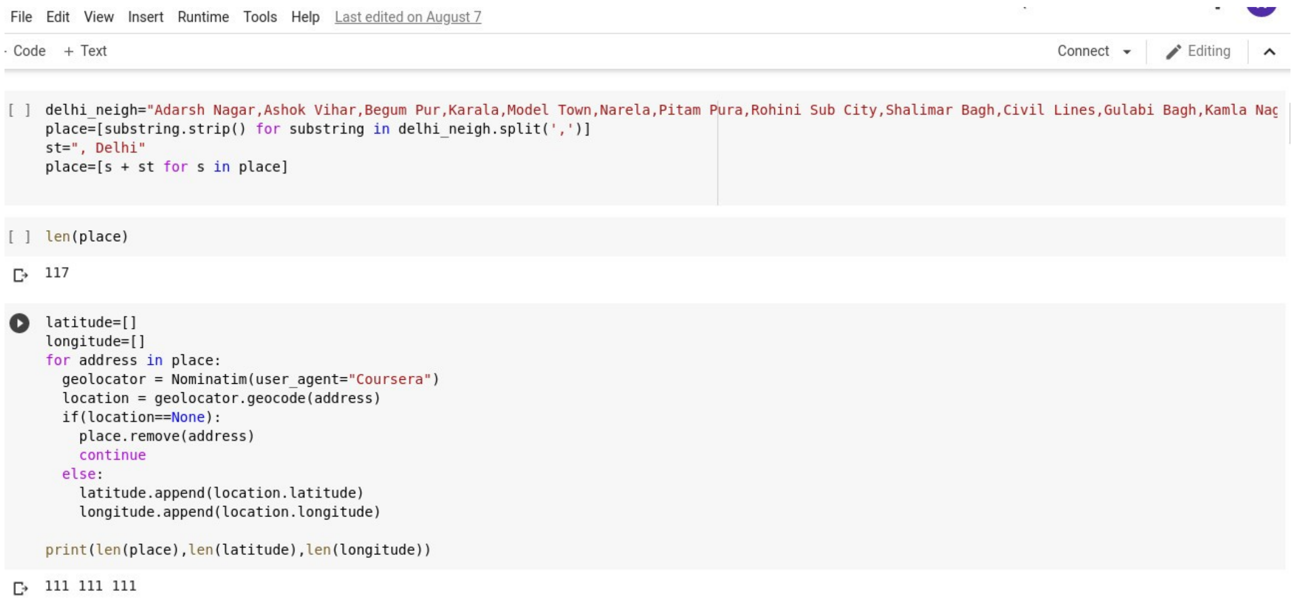
1. So the first step in Data Collection was to get the names and coordinates of the neighbourhoods of Delhi.

Names of the Neighbourhoods of Delhi were found at: "https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi"

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 Nag
    place=[substring.strip() for substring in delhi_neigh.split(',')]
    st=", Delhi"
    place=[s + st for s in place]
```

```
[ ] len(place)
```

```
117
```

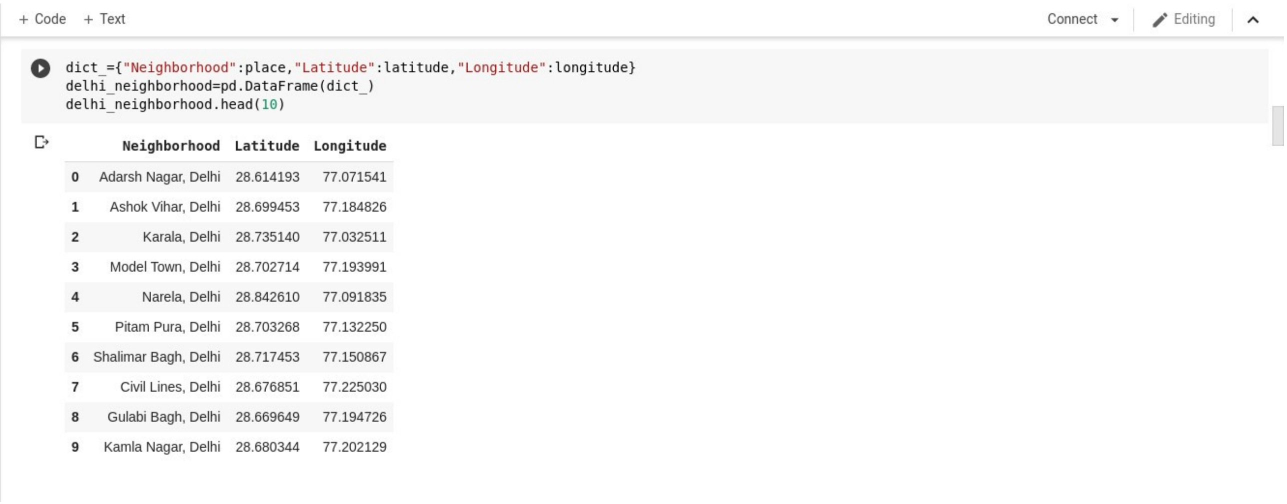
```
latitude=[]
longitude=[]
for address in place:
  geolocator = Nominatim(user_agent="Coursera")
  location = geolocator.geocode(address)
  if(location==None):
    place.remove(address)
    continue
  else:
    latitude.append(location.latitude)
    longitude.append(location.longitude)

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

```
111 111 111
```

I made a dataframe using the names and coordinates of the neighbourhoods.

Below is the screenshot of the jupyter notebook:



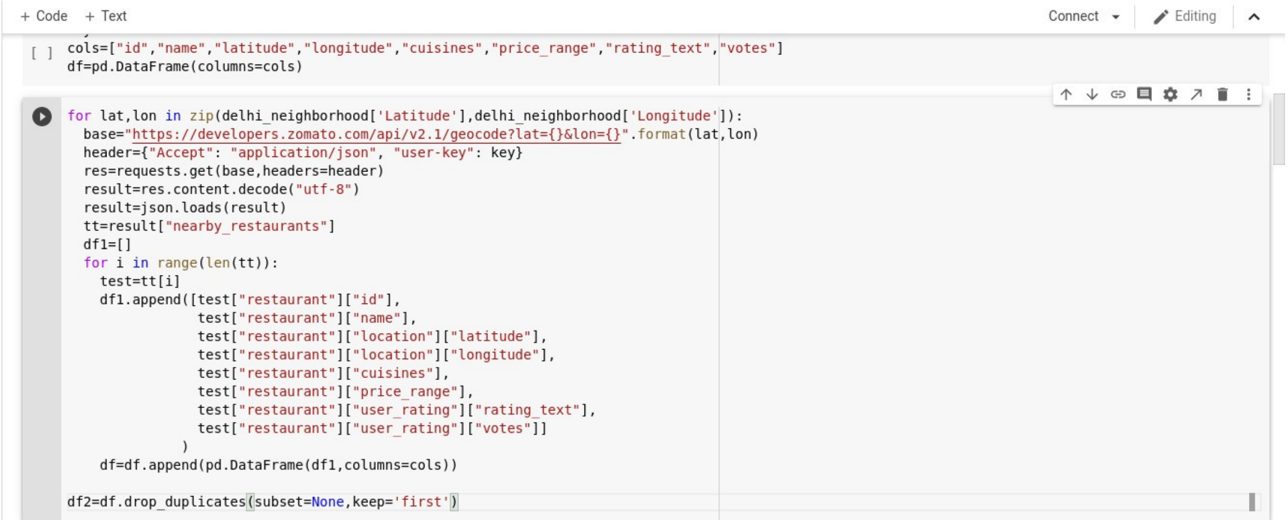
```
dict_={"Neighborhood":place,"Latitude":latitude,"Longitude":longitude}
delhi_neighborhood=pd.DataFrame(dict_)
delhi_neighborhood.head(10)
```

	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. The data I collected include :

- id- Restaurant id
- name- Name of Restaurant
- latitude- Latitudinal coordinate of the Restaurant
- longitude- Longitudinal coordinate of the Restaurant
- cuisines- The type of foods the Restaurant offer
- price_range – 1 represent food for two costs less than Rs. 500,
  2 represent food for two costs more than Rs. 500 but less
  than Rs. 1000, etc.
- rating_text – Rating the Restaurant have(i.e. Very Good, Good, etc.)
- votes – No. Of votes that the Restaurant got it’s rating from

Here is the code to do so:

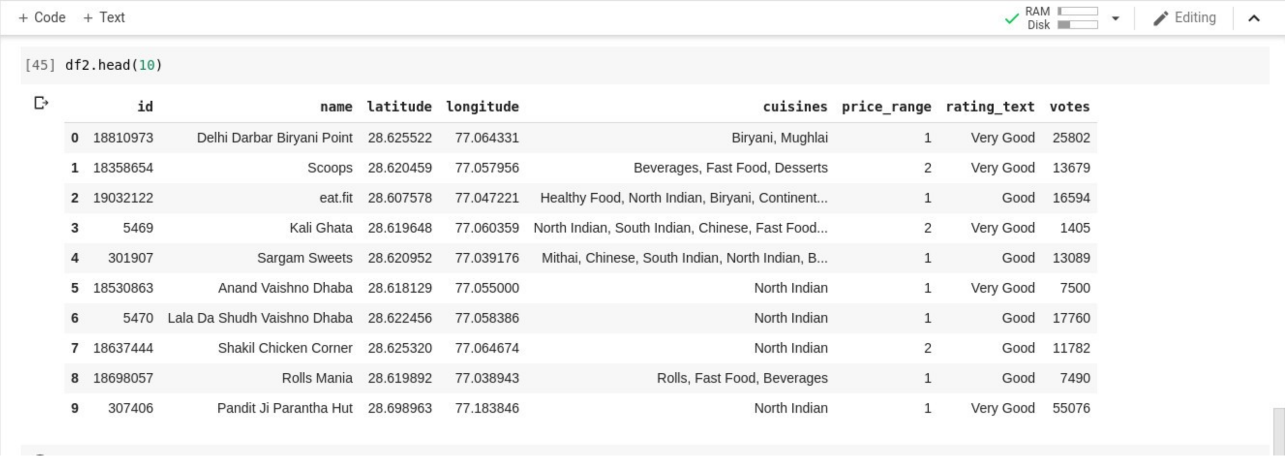


```
cols=["id","name","latitude","longitude","cuisines","price_range","rating_text","votes"]
df=pd.DataFrame(columns=cols)
```

```
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"]["latitude"],
                test["restaurant"]["location"]["longitude"],
                test["restaurant"]["cuisines"],
                test["restaurant"]["price_range"],
                test["restaurant"]["user_rating"]["rating_text"],
                test["restaurant"]["user_rating"]["votes"]]
              )
    df=df.append(pd.DataFrame(df1,columns=cols))

df2=df.drop_duplicates(subset=None,keep='first')
```

Here is a glimpse of the content of the DataFrame:



```
[45] df2.head(10)
```

	id	name	latitude	longitude	cuisines	price_range	rating_text	votes
0	18810973	Delhi Darbar Biryani Point	28.625522	77.064331	Biryani, Mughlai	1	Very Good	25802
1	18358654	Scoops	28.620459	77.057956	Beverages, Fast Food, Desserts	2	Very Good	13679
2	19032122	eat.fit	28.607578	77.047221	Healthy Food, North Indian, Biryani, Continent...	1	Good	16594
3	5469	Kali Ghata	28.619648	77.060359	North Indian, South Indian, Chinese, Fast Food...	2	Very Good	1405
4	301907	Sargam Sweets	28.620952	77.039176	Mithai, Chinese, South Indian, North Indian, B...	1	Good	13089
5	18530863	Anand Vaishno Dhaba	28.618129	77.055000	North Indian	1	Very Good	7500
6	5470	Lala Da Shudh Vaishno Dhaba	28.622456	77.058386	North Indian	1	Good	17760
7	18637444	Shakil Chicken Corner	28.625320	77.064674	North Indian	2	Good	11782
8	18698057	Rolls Mania	28.619892	77.038943	Rolls, Fast Food, Beverages	1	Good	7490
9	307406	Pandit Ji Parantha Hut	28.698963	77.183846	North Indian	1	Very Good	55076

Usage of the Data:

Most of Data from the above DataFrame will not be used.
The usable data includes:

- **Latitude and Longitude:** These values will be used to get an appropriate venue for the New Restaurant to open.
- **Cuisines and Votes:** These values will be combinedly used to get the most popular category of the Restaurant in the city(Delhi). As it’ll be good to open a Restaurant with the Type of food that people like in that City.