IDENTIFYING OPPORTUNITIES IN GURGAON FOR OPENING UP A CLINIC

Coursera capstone project

DESCRIPTION OF THE PROBLEM

Gurgaon is among the fastest growing Indian cities and is one of the biggest IT and consulting hubs of India

For its healthcare needs it relies primarily on the major hospitals in Delhi given its close proximity to the capital city

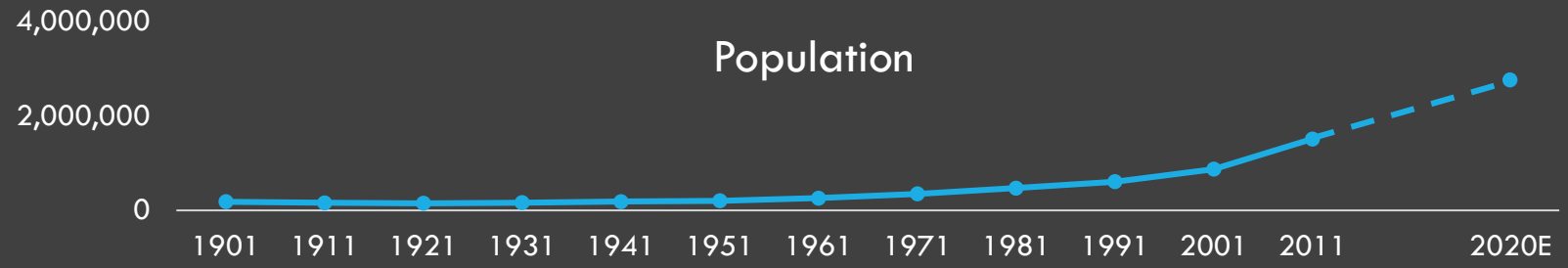
There seems to be a growing need for new hospitals and clinics to be set up within the city to cater to the growing healthcare requirement

A client has approached us to identify an area within the city which is the most underserved when it comes to outpatient clinics

We will **leverage Foursquare** to identify the most optimal location

GURUGRAM DEMOGRAPHICS

In the past one decade, Gurgaon has witnessed its population growing at an alarming pace but the health care facilities and infrastructure have not been updated in keeping with that pace.



Gurgaon, officially Gurugram, is a city located in the northern Indian state of Haryana. It is one of the major satellite cities of Delhi and is part of the National Capital Region of India.

It has the third-highest per-capita income in India and is the site of Indian offices for half of Fortune 500 companies

The city is categorized as very high on the Human Development Index, with an HDI of 0.889 (2017), which is also the highest in India

In March 2019, Gurgaon was named the most polluted city in the world, according to data released by IQ Air Visual and Greenpeace

DESCRIBE THE DATA

There will be three data sources which will be used as part of this assignment:

1. Foursquare API: The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag. Additionally, Foursquare allows developers to build audience segments for analysis and measurement. JSON is the preferred response format.

- Foursquare API will be used in the assignment to generate a list of hospitals and clinics as per communities within Gurugram
- Foursquare "Venues" data provided for a particular location will be used

Source: https://www.programmableweb.com/api/foursquare

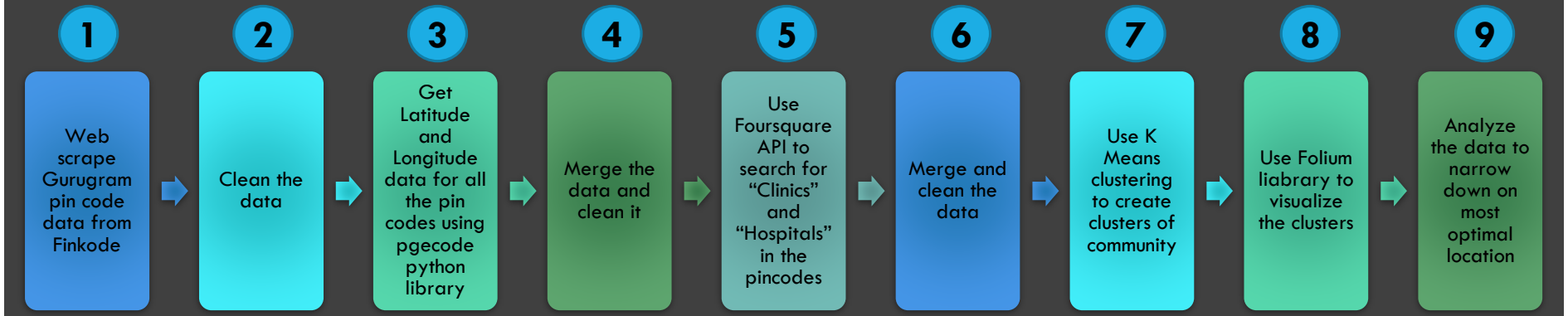
DESCRIBE THE DATA

2. finkode.com: Finkode provides Postal Code data for communities at city level. It will be used to get postal codes for all communities that are within the Gurugram jurisdiction.
   - Web scraping of the data table in the site will be done and Beautiful soup liabrary will be used to extract the relevant data
3. Pgecode: pgeocode is a Python library for high performance off-line querying of GPS coordinates, region name and municipality name from postal codes. Distances between postal codes as well as general distance queries are also supported. The used GeoNames database includes postal codes for 83 countries.
   - This library will be used to generate latitude and longitude for the pin codes captured from finkode.com.

Source: https://pypi.org/project/pgeocode/

METHODOLOGY

The following steps have been carried out as part of the assignment:



Each step was a building block towards arriving at the most optimal area where the Clinic will be setup

All the steps have been detailed out in the following slides

METHODOLOGY

Step 1-3: Get the Pincodes from https://finkode.com/hr/gurgaon.html and clean the Pincode column.

Use pgecode liabrary and get all the location details of the pincodes such as place name, community, coordinates and accuracy

	Area	Pincode
0	Agon	[122104]
1	Air Force	[122105]
2	Alkatel India Ltd	[122225]
3	American Institute	[122210]
4	AO, BSNL GMTD	[122234]

```
df['Pincode'] = df['Pincode'].str[0]
df
```

```
import pgeocode
lis  = df.loc[:,"Pincode"]
lis2 = lis.tolist()
nominatim = pgeocode.Nominatim('IN') #Country code
```

```
df2 = nominatim.query_postal_code(lis2)
df2
```

	postal_code	country code	place_name	state_name	state_code	county_name	county_code	community_name	community_code	latitude	longitude	accuracy
0	122104	IN	Ferozepur Jhirka, Doha, Rawli, Agon, Hirwari, ...	Haryana	10.0	Gurgaon	86.0	Ferozepur Jhirka	NaN	28.3681	76.948750	4.0
1	122105	IN	Rathiwas, Jhamuwas, Baghanki, Mohamadpur Ahir,...	Haryana	10.0	Gurgaon	86.0	Tauru	NaN	28.2620	76.923164	4.0
2	122225	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	122210	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

METHODOLOGY

Step 4-6: Merge the two data frames to create a new one containing only pincodes, area name, community and latitude & longitude

Clean the data by dropping duplicate items from the area column, remove rows containing NAN items in the coordinates columns, sort the values as per pincode and reset the data frame index

Next use Foursquare credentials and create a JSON query to get “venues” data for all the coordinates within a range of 3km of each

This will give the number of hospitals and clinics in each community

Merge the venue data with the original dataframe and clean the data

```
CLIENT_ID = 'SVXBQLDXOARVTQR2OG3JQFHBQK3PAVHOTN4TBLLMT1TC1HIG' # your Foursquare ID
CLIENT_SECRET = 'PCK52X4UC0TBKAAVPITUDTWPCCTVPB0INCUWS2HQA2RA1UM0' # your Foursquare Secret
VERSION = '20180605' # Foursquare API version

LIMIT = 15 # limit of number of venues returned by Foursquare API
radius = 3000 #3Km radius
```

```
# tranforming json file into a pandas dataframe library
from pandas.io.json import json_normalize
search_query_list = ['Clinic','Hospital']
print(type(search_query_list))
venues_count=[]
lat=[]
long=[]
LIMIT = 300
for ind in df6.index:
    latitude = df6['latitude'][ind]
    longitude = df6['longitude'][ind]
    for ind2 in search_query_list:
        search_query=ind2
        url = 'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},{}&
```

	Clinic	Hospital	latitude	Area	Pincode	community_name	longitude
0	2	1	28.418	South City-I	122001	Basai Road	76.9839
1	2	1	28.418	Nathupur	122002	Dlf Qe	76.9839
2	2	1	28.418	Smaspur	122003	Sector -45	76.9839
3	2	1	28.418	Industrial Estate	122007	Industrial Estate	76.9839
4	2	1	28.418	DLF Ph-II	122008	Dlf Ph-ii	76.9839

METHODOLOGY

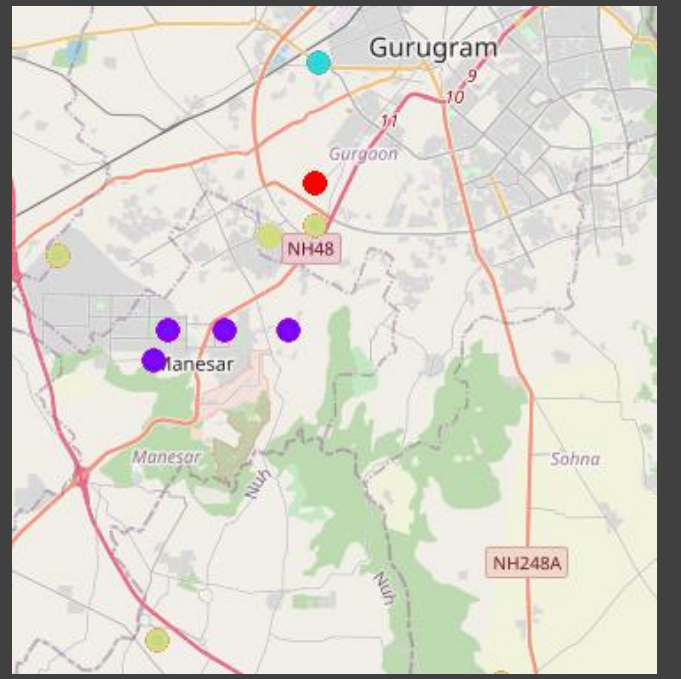
Step 7-9: Import KMeans from sklearn and use it to create 4 different clusters of communities

Go through the merged data frame and analyze the clusters formed using Kmeans to see how similar they are

Use Folium to visualize the various communities in Gurugram on a map and get a sense of how they are scattered

visualize them on the Gurugram map using Folium and see how the communities within each cluster are scattered

From among the clusters select the cluster and community within it where the clinic should be opened



RESULTS

The final resulting data frame consists of 18 unique communities which have distinct pin codes

These communities have been clustered in 4 clusters as seen in the screenshot

Maximum number of communities (7) lie in cluster 3 whereas cluster 2 only has one community

In contrast, Cluster 2 has the most number of clinics (6) and hospitals (4) whereas cluster 3 has the least number of venues

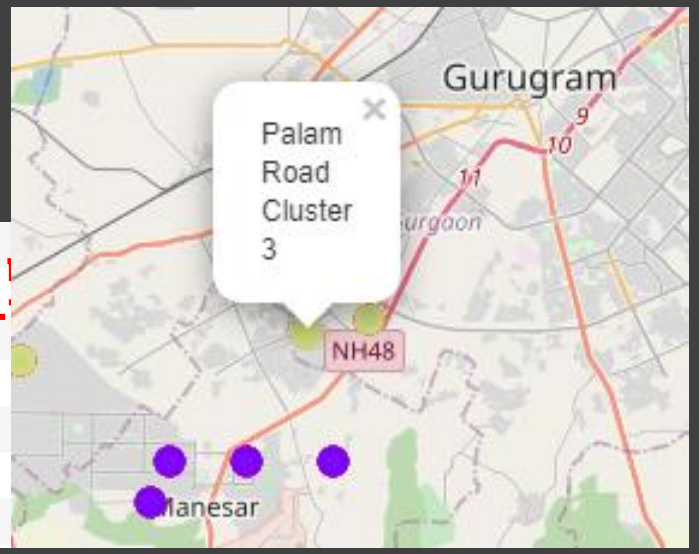
	Cluster Labels	Clinic	Hospital	community_name	Pincode	Area
0	0	2	1	Dlf Qe	122002	Nathupur
1	0	2	1	Sector -45	122003	Smaspur
2	0	2	1	Industrial Estate	122007	Industrial Estate
3	0	2	1	Dlf Ph-ii	122008	DLF Ph-II
4	0	2	1	Basai Road	122001	South City-I
5	0	2	1	Galleria Dlf-iv	122009	Galleria DLF-IV
6	1	0	2	Ferozepur Jhirka	122104	Bewan
7	1	0	2	Bhondsi	122102	Rithoj
8	1	0	2	Nsg Camp Manesar	122051	Nsg Camp Manesar
9	1	0	2	Nagina	122108	Bhadas
10	2	6	4	Railway Road	122006	Railwary Road
11	3	1	0	Industrial Complex Dundahera	122016	Industrial Complex Dundahera
12	3	1	0	Palam Vihar	122017	Carterpuri
13	3	0	0	Badshahpur	122101	Teekli
14	3	0	0	Tauru	122105	Dulawat
15	3	1	0	Narsinghpur	122004	Khandsa
16	3	0	0	Sohna	122103	Bhirauti
17	3	1	0	Dlf Ph-iii	122010	DLF Ph-III

RESULTS

Among the various clusters, the last one, Cluster 3, seems to be the most under-served as it has the least number of clinics and hospitals in its vicinity

Within cluster 3 Palam Vihar community has only one clinic and no hospitals

11	3	1	0	Industrial Complex Dundahera	122016	Industrial Complex Dundahera
12	3	1	0	Palam Vihar	122017	Carterpuri
13	3	0	0	Badshahpur	122101	Teekli
14	3	0	0	Tauru	122105	Dulawat
15	3	1	0	Narsinghpur	122004	Khandsa
16	3	0	0	Sohna	122103	Bhirauti
17	3	1	0	Dlf Ph-iii	122010	DLF Ph-III



Palam Vihar is a good option for the client to open up a clinic within Gurugram

DISCUSSION

1. Gurugram, the millennium city of India has the highest per caita income, fastest economic growth among major cities and ranks fairy high in the HDI index for Indian cities
2. But even basic healthcare facilities are lacking within the city with residents having to rely on the limited health infrastructure available or travel to nearby Delhi hospitals
3. This has created a demand supply gap within the city and there is a good opportunity for new medical establishments to come up within the city
4. The 4 clusters obtained from our analysis, have distinct geographical presence and are also similar in terms of medical facilities available
5. Palam Vihar community within Cluster 3 is close to the main city, and is a thriving locale. It is close to cluster 2 which has the most medical facilities within the city
6. But there seems to be a need for another clinic within the community to support the immediate needs of the residents of that community

CONCLUSION

1. A well planned move towards establishment of an out patient clinic in Palam Vihar can be financially enriching for the client given the latent need for medical facilities
2. Next steps towards this venture are identifying funding requirements, zeroing in on the exact location of establishment and obtaining government approvals