Coursera Capstone Project

The Battle of the Neighborhoods

Finding best location to open an ATM in Bangalore, India

IBM DATA SCIENCE PROFESSIONAL SPECIALIZATION BY COURSERA

BY: PADHMA M

TABLE OF CONTENTS

1. INTRODUCTION

Bangalore is one among the fastest growing cities in the world. Bangalore is referred to as the Silicon Valley of India because of its role as the nation's leading Information Technology exporter .It has a population of over ten million, making it a megacity and third most populous city and fifth most populous urban agglomeration in India. Being a demographically diverse city, the needs of the residents are also increasing rapidly. Hence, any new organization or an existing one should keep up with their pace in supplying the needs of the customers.

2. BUSINESS PROBLEM:

Our customer is ABC Bank, which is an International Bank and also a market leader. They have a wide variety of customers all over the country, major of them residing in Metropolitan cities. ABC Bank has received ample amounts of complaints from residents of Bangalore that there aren't sufficient amount of ATM's. Given the extremely large population and the population of the city, our customer wants to identify the best neighborhood area to open more ATM covering the majority of the population. The problem statement will be: **Which neighborhood is most densely populated and has lesser number of ATM's?**

3. DATA:

The data to be used in this project is not readily available. Hence, the data has been obtained from various sources such as

- **Foursquare**, which is a local search-and-discovery mobile app which provides search results for its users. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history.
- **Wikipedia,** which has the details about the neighborhoods in Bangalore.
  https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore
- The geographic coordinates of each location have been obtained through **Geopy,** which makes it easy for Python developers to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders.
- The population data about each neighbourhood has been obtained from:
  https://www.ichangemycity.com/assembly-constituencies/mahalakshmi-layout
  https://www.census2011.co.in/census/district/242-bangalore.html

3.1 DATA DESCRIPTION:

As we have to explore and identify the neighborhoods in the city of Bangalore, the Bangalore neighbourhood data is the crucial data for this project. The data about each neighbourhood is not readily available, hence we have to scrape the Wikipedia page and obtain the data. The data also needs information about coordinates which makes it easier for us to cluster the neighborhoods. In order to obtain the coordinates, we make use of geopy library in Python. We also need information about each neighbourhood which will make it simpler to recognise which area has least number of ATM's. Hence, the details about each venue is obtained through FourSquare API. The population about each neighbourhood will let us know which neighbourhood is more preferable. Hence, the demographic data about each neighbourhood is also obtained.

Our DataFrame containing Population data will be:

	Neighborhood	Population
0	Cantonment area	157683
1	Domlur	69406
2	Indiranagar	85890
3	Jeevanbheemanagar	76502
4	Malleswaram	60697
5	Pete area	47076
6	Sadashivanagar	60794
7	Seshadripuram	45630
8	Shivajinagar	60506
9	Ulsoor	70180
10	Vasanth Nagar	75012
11	Bellandur	80180
12	CV Raman Nagar\t	58815
13	Hoodi	41440
14	Krishnarajapuram	186210
15	Mahadevapura	154223
16	Marathahalli	72489

Our Data Frame containing neighbourhood data will be:

	Neighborhood	Region	Latitude	Longitude
0	Cantonment area	Central	12.995441	77.601429
1	Domlur	Central	12.962467	77.638196
2	Indiranagar	Central	12.979441	77.641689
3	Jeevanbheemanagar	Central	12.968926	77.652705
4	Malleswaram	Central	13.006163	77.567158
5	Pete area	Central	12.970878	77.548101
6	Sadashivanagar	Central	13.007708	77.579589
7	Seshadripuram	Central	12.991462	77.576226
8	Shivajinagar	Central	12.983752	77.604315
9	Ulsoor	Central	12.977879	77.624670
10	Vasanth Nagar	Central	12.988721	77.585169
11	Bellandur	Eastern	12.936121	77.666617
12	CV Raman Nagar	Eastern	12.980266	77.663726
13	Hoodi	Eastern	12.991903	77.716201
14	Krishnarajapuram	Eastern	13.016118	77.703510
15	Mahadevapura	Eastern	12.988679	77.688249
16	Marathahalli	Eastern	12.955257	77.698416
17	Varthur	Eastern	12.940615	77.746994

The neighbourhood data contains region in which the neighbourhood is located and the geographic coordinates (latitude, longitude) of each neighbourhood.

4. METHODOLOGY:

4.1 Data Preprocessing:

The data preprocessing stage involves obtaining and cleaning the dataset and structuring the data into a proper format. The initial stages include:

1. Scraping from the wikipedia page:

```
source_page = requests.get('https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore').text
soup = BeautifulSoup(source_page,'lxml')
```

```
neigh = []
areas = []
locality = ['Central','Eastern','North Eastern','Northern','South Eastern','Southern','Southern suburbs','Western']
for k in range(len(locality)):
    table = soup.find_all('table')[k]
    table_data = table.find_all('td')
    col_len = len(table_data)

    for i in range(0,col_len,3):
        neigh.append(table_data[i].text.strip())
        areas.append(locality[k])
```

```
df_neigh = pd.DataFrame(data=[neigh,areas])
df_neigh = df_neigh.transpose()
df_neigh.columns = ['Neighbourhood','Area']
df_neigh.head()
```

	Neighbourhood	Area
0	Cantonment area	Central
1	Domlur	Central
2	Indiranagar	Central
3	Jeevanbheemanagar	Central
4	Malleswaram	Central

2. Obtain coordinates for each location using geopy library.

```
address = neighborhoods
latitude = []
longitude = []

for i in range(len(neighborhoods)):
    geolocator = Nominatim(user_agent="ny_explorer")
    location = geolocator.geocode(address[i])
    if location:
        latitude.append(location.latitude)
        longitude.append(location.longitude)
    else:
        latitude.append(None)
        longitude.append(None)
```

3. Finding columns with null values and replacing them.

```
null_columns=df_new.columns[df_new.isnull().any()]
print(df_new[df_new.isnull().any(axis=1)][null_columns])
```

```
   Latitude  Longitude
3       NaN        NaN
```

4. Merge the neighborhood dataset with population dataset.

```
df_bang = df_reg.merge(df_pop,on='Neighborhood')
df_bang.head()
```

	Neighborhood	Region	Latitude	Longitude	Population
0	Cantonment area	Central	12.995441	77.601429	157683
1	Domlur	Central	12.962467	77.638196	69406
2	Indiranagar	Central	12.979441	77.641689	85890
3	Jeevanbheemanagar	Central	12.968926	77.652705	76502
4	Malleswaram	Central	13.006163	77.567158	60697

4.2 Exploratory Data Analysis:

Let's begin by exploring the city and finding the count of neighborhoods and regions in the city.

```
# number of neighborhoods and regions in the city
print('The number of neighborhoods in Bangalore is: {}'.format(df_new['Neighborhood'].nunique()))
print('The number of regions in Bangalore is: {}'.format(df_new['Region'].nunique()))
```

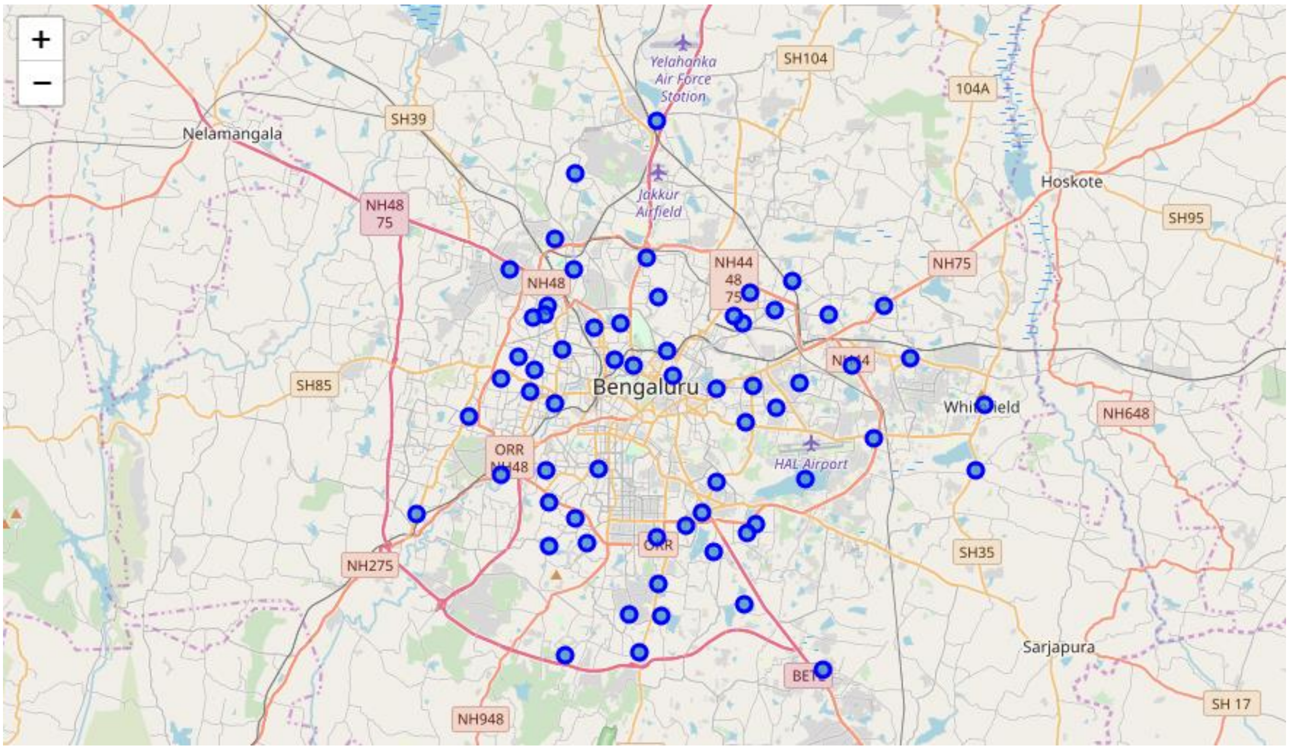
```
The number of neighborhoods in Bangalore is: 65
The number of regions in Bangalore is: 8
```

Now that we know about each neighborhood, lets find the unique venues in each neighborhood.

```
print('There are {} uniques categories.'.format(len(bangalore_venues['Venue Category'].unique())))
```

```
There are 145 uniques categories.
```

Let's visualize the neighborhoods using a folim map:



We then obtain top 5 venues at each neighborhood, which will let us know which neighborhoods lacks in ATM Machine's.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Anjanapura	Pool	Women's Store	Department Store	Eastern European Restaurant	Donut Shop
1	Arekere	Sporting Goods Shop	Indian Restaurant	Ice Cream Shop	Chinese Restaurant	Burger Joint
2	BTM Layout	Indian Restaurant	Chinese Restaurant	Snack Place	Ice Cream Shop	Bakery
3	Banashankari	Café	Indian Restaurant	Clothing Store	Pizza Place	Women's Store
4	Banaswadi	Indian Restaurant	Vegetarian / Vegan Restaurant	Pharmacy	Juice Bar	BBQ Joint
5	Basavanagudi	Indian Restaurant	Sandwich Place	Fast Food Restaurant	Restaurant	Juice Bar
6	Basaveshwaranagar	Fast Food Restaurant	Indian Restaurant	Ice Cream Shop	Juice Bar	Snack Place
7	Begur	Supermarket	ATM	Clothing Store	Bakery	Women's Store
8	Bommanahalli	Indian Restaurant	Department Store	Women's Store	Eastern European Restaurant	Donut Shop
9	CV Raman Nagar	Indian Restaurant	Bakery	Smoke Shop	Café	Coffee Shop
10	Cantonment area	Music Venue	Café	Bakery	Indian Restaurant	Event Space
11	Domlur	Indian Restaurant	Café	Chinese Restaurant	Sports Bar	Sandwich Place
12	Electronic City	Café	Breakfast Spot	Indian Restaurant	Middle Eastern Restaurant	Bus Station

By placing more ATM machine's in densely populated regions, more customers will be satisfied and the complaints will gradually be reduced. Hence, we sort top 15 neighborhoods based on maximum population.

	Neighborhood	Region	Latitude	Longitude	Population
0	Mahalakshmi Layout	Western	13.011894	77.543858	586460
1	Hulimavu	Southern suburbs	12.873457	77.598534	500006
2	Bommanahalli	South Eastern	12.902924	77.622897	224980
3	Horamavu	North Eastern	13.027331	77.660151	196553
4	Electronic City	South Eastern	12.848292	77.674371	186234
5	Krishnarajapuram	Eastern	13.016118	77.703510	186210
6	Cantonment area	Central	12.995441	77.601429	157683
7	Mahadevapura	Eastern	12.988679	77.688249	154223
8	Banashankari	Southern	12.925932	77.545919	150000
9	Hebbal	Northern	13.038218	77.591900	132571
10	Peenya	Northern	13.032942	77.527325	115628
11	BTM Layout	South Eastern	12.915177	77.610282	104500
12	Yelahanka	Northern	13.100698	77.596345	94234
13	Banaswadi	North Eastern	13.014162	77.651854	92284
14	Nagarbhavi	Western	12.965101	77.507863	87548

4.3 Clustering:

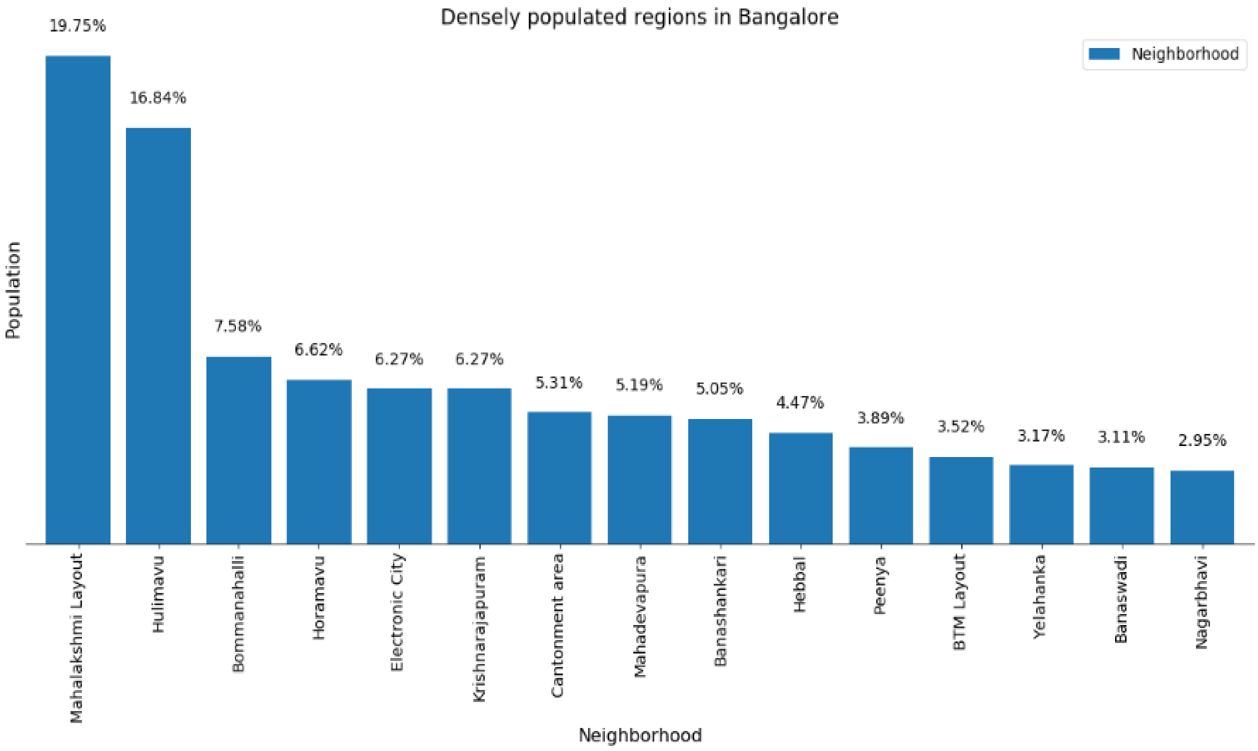
We sort the neighborhoods into five clusters to obtain better insights about each neighborhood and by we will know which cluster of neighborhoods have more population.

```
# set number of clusters
kclusters = 5

bangalore_grouped_clustering = bangalore_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(bangalore_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

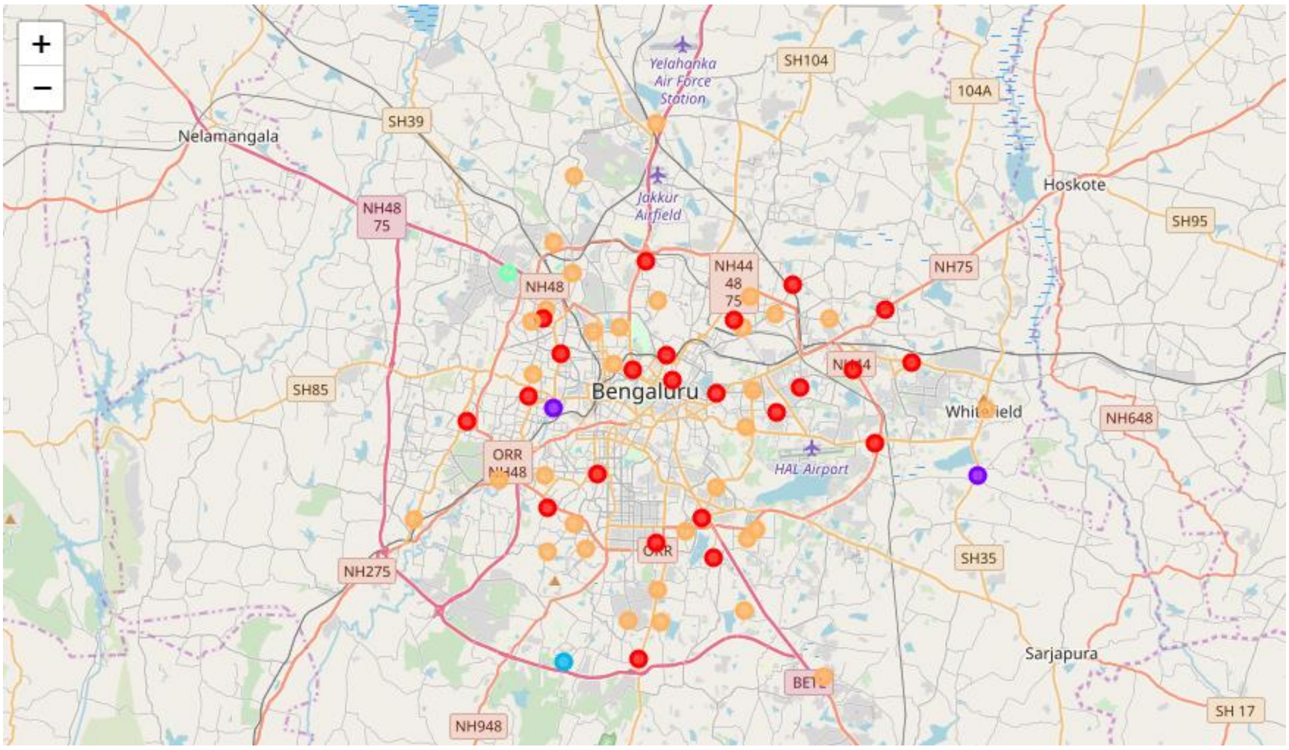
```
array([2, 4, 4, 0, 4, 0, 4, 4, 0, 0], dtype=int32)
```

5. RESULTS AND DISCUSSION

By sorting the neighborhoods based on population and visualizing it in the form of a bar plot, we find that the top 3 locations contribute to the 35% of the total population of the city.



By clustering the neighborhoods, we find that most of the densely populated neighborhoods belong to the first cluster.



6. CONCLUSION

By the population graph and the cluster map, we see that the first cluster contains most of the populous neighborhoods. Hence, by placing more ATM machine's in the first cluster's neighborhoods, the needs of the customers will be successfully met and the complaints will reduce drastically.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Cantonment area	Music Venue	Café	Bakery	Indian Restaurant	Event Space
3	Jeevanbheemanagar	Indian Restaurant	Chinese Restaurant	Café	Dessert Shop	Kerala Restaurant
8	Shivajinagar	Indian Restaurant	Hotel	Jewelry Store	Bus Station	Market
9	Ulsoor	Indian Restaurant	Hotel	Restaurant	Light Rail Station	Dessert Shop
10	Vasanth Nagar	Indian Restaurant	Hotel	Nightclub	Coffee Shop	Japanese Restaurant
12	CV Raman Nagar	Indian Restaurant	Bakery	Smoke Shop	Café	Coffee Shop
13	Hoodi	Indian Restaurant	Furniture / Home Store	Bus Station	Women's Store	Eastern European Restaurant
14	Krishnarajapuram	Bakery	Hotel	Tibetan Restaurant	Indian Restaurant	Women's Store
15	Mahadevapura	Indian Restaurant	Convenience Store	Bus Station	Shopping Mall	Department Store
16	Marathahalli	Indian Restaurant	Clothing Store	Vegetarian / Vegan Restaurant	Ice Cream Shop	Shoe Store
21	Horamavu	Chinese Restaurant	Indian Restaurant	Athletics & Sports	Dessert Shop	Electronics Store
24	Lingarajapuram	Indian Restaurant	Kerala Restaurant	Department Store	Paper / Office Supplies Store	Steakhouse
26	Hebbal	Park	Bar	Sporting Goods Shop	Indian Restaurant	Women's Store
34	Bommanahalli	Indian Restaurant	Department Store	Women's Store	Eastern European Restaurant	Donut Shop
40	Madiwala	Indian Restaurant	Bakery	Fast Food Restaurant	Department Store	Furniture / Home Store

During this work, some of the machine learning techniques, data wrangling with pandas and data visualization techniques were put to use. When the data becomes complex, some higher metrics can be used. Also, given the limitation that the foursquare sandbox account can create only 100 requests and return venues, better insights can be obtained in future works by accessing larger datasets.