THE BATTLE OF NEIGHBORHOODS – ANALYSIS OF NEIGHBORHOODS SIMILAR TO HOMETOWN FOR JOB RELOCATION

A REPORT

submitted by

KRISHNA SRINIVASAN

for the final project

APPLIED DATA SCIENCE CAPSTONE
(Offered by IBM)

COURSERA

JULY 2020

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER 1. INTRODUCTION

1.1. BACKGROUND

Moving is painful even though it is for the career or job of your dreams. Job relocation or temporary housing in a new location is a huge topic of discussion which is mostly commonly occurring scenario for an individual if not so often. There are so many articles and literatures about this topic where one have to make a life decision or choose for the best option possible. Multiple commercial players are available in market some of them include WHRG Group, SHRM, CapRelo Hellolanding etc. to aid and support individual going through relocation specific phase of their life. There are several factors to consider during relocation which include the new cost of living, weather, company's performance, individual's relationships, a new home or temporary housing and so on. In case of searching of a new home factors like proximity to essential services, convenience of transport, flexibility of agreements, a new lifestyle etc. have to be considered. Apart from personal relocation there are other specific use cases like traveling nurse housing, corporate housing, inter housing etc.

1.2. PROJECT DESCRIPTION

In this project let us focus on a particular case of relocation where an individual would want to relocate from hometown for a new job. Main focus of the current project is restricted to identifying locations near job place that are similar to places in hometown without loss in generality.

CHAPTER 2. DATA DESCRIPTION

2.1. NEIGHBORHOOD

Understanding of the neighborhood of the hometown and job location is essential for solving this problem. First hometown is chosen to be Chennai city, India, where Adyar and Kodambakkam are chosen to be zones of importance. Second the job location is chosen to be Bangalore, India, where zones of interest will be obtained in such a way that they are within 10 Km range from the office location of Whitefield area. Once the zone information of the cities are identified it is equally important to obtain the latitude and longitude geographical coordinates of the zones. After obtaining all relevant information, neighborhood of each zone is explored using Foursquare API. Later the neighborhood information is analyzed for similarities and differences. Finally an appropriate neighborhood in the job location similar to hometown is chosen based on the analysis results.

2.2. DATA SOURCE

Zone information is obtained from Wikipedia pages –

- https://en.wikipedia.org/wiki/List_of_Chennai_Corporation_zones
- https://en.wikipedia.org/wiki/List_of_wards_in_Bangalore

Geographical coordinate information is obtained from –

- https://indiamapia.com/Chennai.html
- https://indiamapia.com/Bangalore.html

Zonal and geographical information for important regions mentioned in this project are not readily available for direct usage. Hence, the date from the above mentioned pages are gathered and combined appropriately to be used in this project.

CHAPTER 3. METHODOLOGY

3.1. DATA EXTRACTION AND CLUSTERING

In this section, first collect the zonal information data from the mentioned website in the data section. Beautiful soup is used to extract data from HTML, which is useful for web scraping. First, inspect the webpage to identify the class attribute for the table of interest. Raise a get request to fetch the raw HTML content and parse the content. Read the table of interest with the identified class attribute and extract all table information.

After obtaining the zonal information, latitude and longitude location data for each zone is obtained using Geopy, Nominatim package which is commonly used for converting address to location information. The location data is visually rendered on a map using Folium.

Each neighborhood location is explored using foursquare API requests and the JSON response from the API request is processed to obtain only relevant information. The Obtained venue data is explored for total number of venues and unique categories.

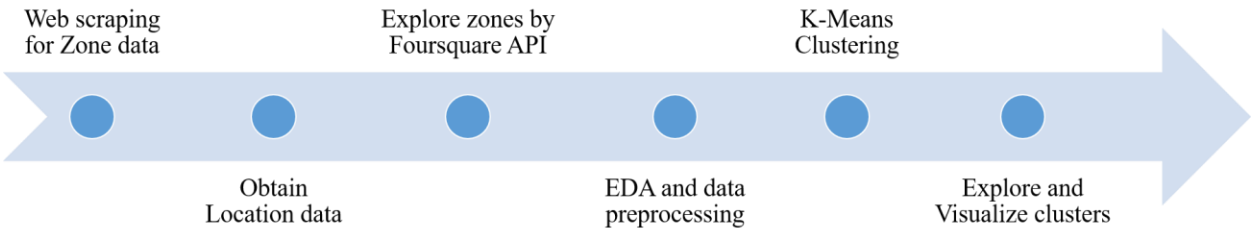

Figure 3.1 Data Extraction and clustering process

One hot coding is done to convert the unique categories to venues to numerical data and the average occurrence of a particular category of venue is obtained. Most

commonly occurring venues in a given neighborhood are identified and a data frame consisting of the neighborhood and its corresponding 10 most commonly occurring venues is prepared.

K-Means clustering algorithms is used to group the venues based on the similarities of the occurrence of venues. In this project, a cluster value of five is chosen to group the neighborhoods into five different categories. The Clustered information is then visually rendered using folium. Individual cluster groups are explored for their similarities and their descriptions. The complete data extraction and clustering process is shown in Figure 3.1.

3.2. IDENTIFYING SIMILAR NEIGHBORHOODS

The above extraction and clustering process is repeated for two main locations i.e. Chennai or hometown location and Bangalore or job location. Neighborhoods of both job and home location are explored first for similarities. Home location neighborhoods are filtered for regions of interest for comparison with job location neighborhoods. Neighborhoods in job location are filtered based on distance from office area which is chosen to be less than 10 Km.

The radial distance between two latitude and longitude location information is calculated using Geopy geodesic distance calculator module.

One important fact to observe is that the distance obtained is radial or perpendicular distance between two latitude and longitude location data and not road wise distance that is noticed commonly on any navigator like google maps.

The filtered location information from hometown and job location are combined in single data frame. The data is then visualized using folium renderer and explore of sufficient information. Already preprocessed data for previous clustering is used

which includes the average occurrence of a particular venue category for each neighborhood.

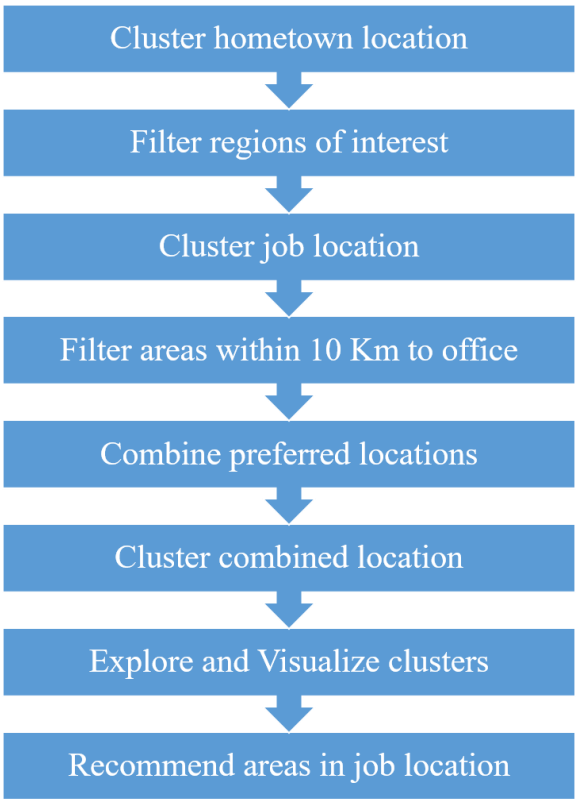

Figure 3.2 Identifying similar neighborhood

The combined filtered data is then clustered using K-Means into five groups. The groups are then visualized and explored individually for similarities and descriptions.

The area in job location, which falls in the same group of the hometown, would become the first choice for selection. In case of multiple areas and groups identified, the most similar hometown location is chosen and the job area in the same group, which is closest to office, are recommended. The process for identifying similar neighborhood is shown in Figure 3.2.

3.3. EXPLORING NEIGHBORHOODS OF CHENNAI AND BANGALORE

3.3.1 Extracting Zonal information

Zones of Chennai and Bangalore are obtained from the website mentioned in data section 2.2 using Beautiful soup. Obtained area and pincode information for Chennai and Bangalore is shown in Figure 3.3 and Figure 3.4 respectively. Total number of Location information include 299 and 295 for Chennai and Bangalore respectively.

	S.No.	Location	Pincode
0	1	Abiramapuram	600018
1	2	Adambakkam	600088
2	3	Adyar	600020
3	4	Agaram	600082
4	5	Alandur	600016

Figure 3.3 Chennai pincode

	S.No.	Location	Pincode
0	0	A F Station Yelahanka	560063
1	1	Adugodi	560030
2	2	Agara	560034
3	3	Agram	560007
4	4	Air Force Hospital	560007

Figure 3.4 Bangalore pincode

3.3.2 Obtain Location data

Latitude and longitude location data for Chennai and Bangalore can be obtained either by scraping through the webpages mentioned in section 2.2 or the zone information can be converted to latitude and longitude using the python module Nominatim from Geopy package. In this project the latter approach is used to obtain location information.

	S.No.	Location	Latitude	Longitude
0	2	Adambakkam	12.982221	80.209121
1	3	Adyar	13.006450	80.257779
2	4	Agaram	13.115860	80.230041
3	5	Alandur	12.994373	80.194284
4	6	Alandur North	12.994373	80.194284

Figure 3.5 Chennai location dataframe

	S.No.	Location	Latitude	Longitude
0	1	Adugodi	12.942847	77.610416
1	2	Agara	12.620112	77.479307
2	5	Amruthahalli	13.066513	77.596624
3	6	Anandnagar	13.033377	77.589523
4	7	Anekal	12.799122	77.680604

Figure 3.6 Bangalore location dataframe

The obtained location information is stored in pandas dataframe as shown in Figure 3.5 and Figure 3.6. A Snap of output from the Geopy Nominatim package during latitude and longitude extraction is shown in Figure 3.7 and Figure 3.8.

```
The geograpical coordinate of Adambakkam, Chennai are 12.9822215, 80.209121.
The geograpical coordinate of Adyar, Chennai are 13.00645, 80.2577791.
The geograpical coordinate of Agaram, Chennai are 13.1158603, 80.2300407.
The geograpical coordinate of Alandur, Chennai are 12.9943729, 80.1942837.
The geograpical coordinate of Alandur North, Chennai are 12.9943729, 80.1942837.
The geograpical coordinate of Ambattur H O, Chennai are 13.1139774, 80.1529202.
The geograpical coordinate of Ambattur West, Chennai are 13.1193746, 80.1507648.
The geograpical coordinate of Anakaputhur, Chennai are 12.9838095, 80.1688264.
The geograpical coordinate of Anna Road H O, Chennai are 13.1207974, 80.0948305.
The geograpical coordinate of Arcot Road, Chennai are 13.0499473, 80.21188.
The geograpical coordinate of Arumbakkam, Chennai are 13.074371, 80.2081312.
The geograpical coordinate of Ashok Nagar, Chennai are 13.0400731, 80.2159247.
The geograpical coordinate of Avadi, Chennai are 13.1254758, 80.09409.
The geograpical coordinate of Ayanavaram, Chennai are 13.0946157, 80.23541.
The geograpical coordinate of Besant Nagar, Chennai are 12.9996907, 80.2682259.
The geograpical coordinate of Broadway, Chennai are 13.0962367, 80.2864133.
The geograpical coordinate of Cathedral, Chennai are 13.0518669, 80.25307675773865.
The geograpical coordinate of Cemetry Road, Chennai are 13.1091962, 80.2861039.
```

Figure 3.7 Obtaining Chennai coordinates

```
The geograpical coordinate of Adugodi, Bangalore are 12.9428472, 77.6104155.
The geograpical coordinate of Agara, Bangalore are 12.6201123, 77.4793074.
The geograpical coordinate of Amruthahalli, Bangalore are 13.0665131, 77.5966238.
The geograpical coordinate of Anandnagar, Bangalore are 13.033377, 77.5895227.
The geograpical coordinate of Anekal, Bangalore are 12.799122, 77.68060437921741.
The geograpical coordinate of Arabic College, Bangalore are 13.0300088, 77.6208657.
The geograpical coordinate of Aranya Bhavan, Bangalore are 13.01061485, 77.57081951019086.
The geograpical coordinate of Ashoknagar, Bangalore are 12.9791198, 77.5912997.
The geograpical coordinate of Attibele, Bangalore are 12.7782588, 77.7712834.
The geograpical coordinate of Attur, Bangalore are 13.0994536, 77.5680336.
The geograpical coordinate of Austin Town, Bangalore are 12.9612736, 77.6152935.
The geograpical coordinate of Avalahalli, Bangalore are 13.0352777, 77.7365655.
The geograpical coordinate of Avenue Road, Bangalore are 12.9684724, 77.5789181.
The geograpical coordinate of Bagalgunte, Bangalore are 13.0566493, 77.5048222.
The geograpical coordinate of Bagalur, Bangalore are 13.1331868, 77.6687093.
```

Figure 3.8 Obtaining Bangalore coordinates

Important point to note is Nominatim package does not provide latitude and longitude coordinates for all the available areas. Python package could provide coordinates for 145 out of 299 areas in Chennai and 184 out of 295 areas in Bangalore. Hence only are areas that have valid coordinates are included for further processing.

3.3.3 Rendering and visualizing location data on folium map

Obtained zonal and coordinate information for areas in Chennai and Bangalore are visualized by rendering it in folium map. Maps of Chennai and Bangalore with neighborhood information is shown in Figure 3.9 and Figure 3.10.

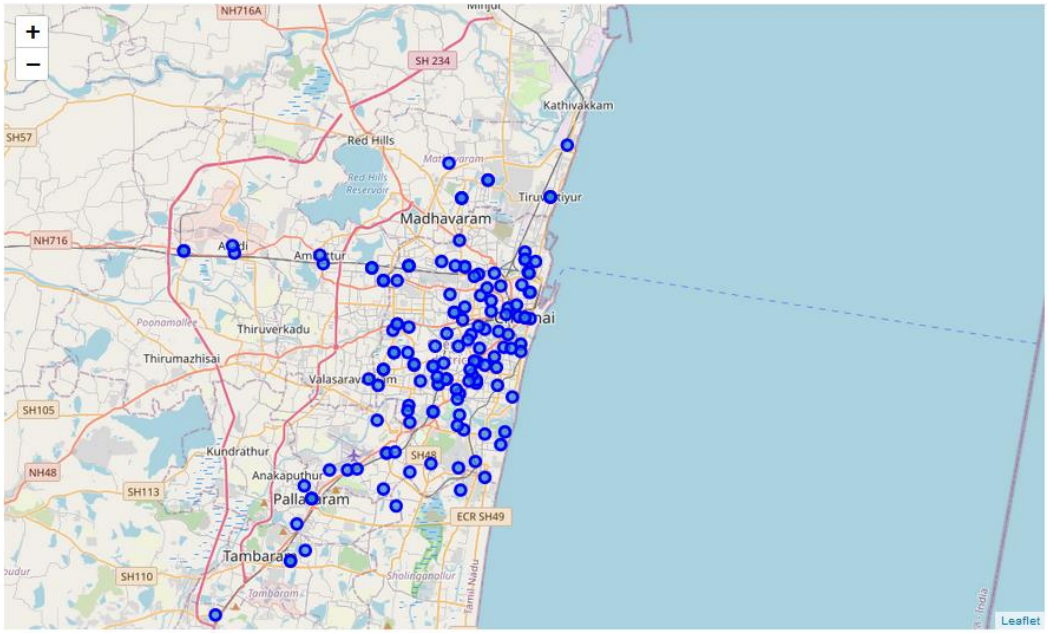

Figure 3.9 Chennai map with neighborhoods

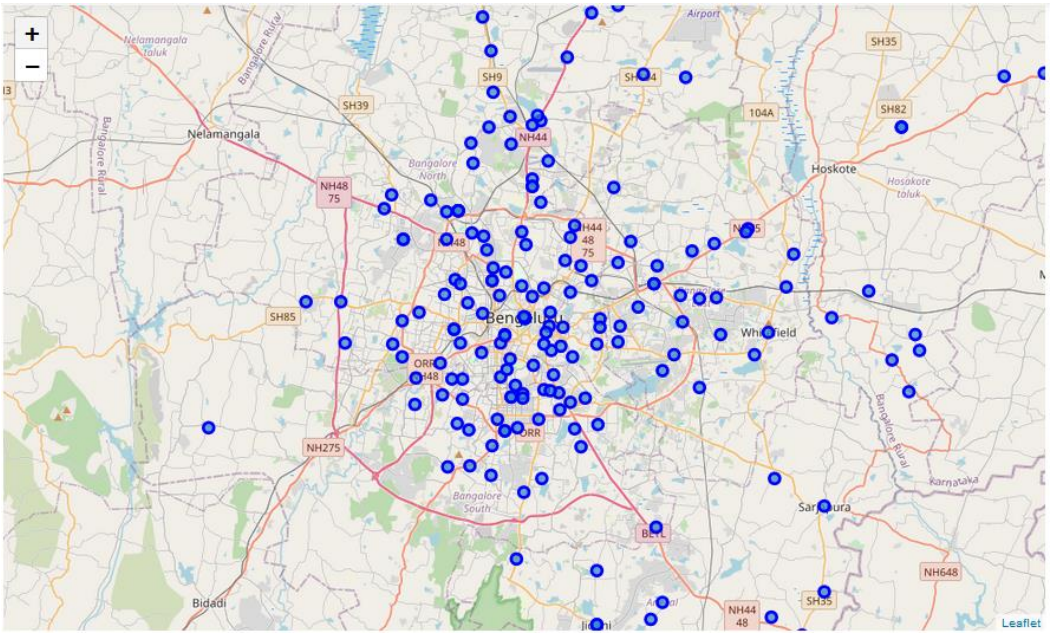

Figure 3.10 Bangalore map with neighborhoods

3.3.4 Explore zones by Foursquare API

Venues near each neighborhood in Chennai and Bangalore are explored using foursquare API through requests module. The obtained JSON file is processed to keep only relevant information and the venue information is stored in pandas dataframe.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adambakkam	12.982221	80.209121	Domino's Pizza	12.984000	80.215000	Pizza Place
1	Adambakkam	12.982221	80.209121	Pizza Empire	12.984231	80.217984	Pizza Place
2	Adambakkam	12.982221	80.209121	Mast Kalandar	12.984715	80.217821	Indian Restaurant
3	Adambakkam	12.982221	80.209121	Sutherland	12.981002	80.205200	IT Services
4	Adambakkam	12.982221	80.209121	Chapathi & Chat Shops	12.982198	80.201599	Indian Restaurant
5	Adambakkam	12.982221	80.209121	simran Aapa Kadai	12.981043	80.218156	Indian Restaurant
6	Adambakkam	12.982221	80.209121	GRT	12.982933	80.217886	Jewelry Store
7	Adambakkam	12.982221	80.209121	Bistro 100	12.982772	80.218000	Modern European Restaurant
8	Adyar	13.006450	80.257779	Bombay Brassiere	13.006961	80.256419	North Indian Restaurant
9	Adyar	13.006450	80.257779	Cool Biz	13.006632	80.257275	Juice Bar

Figure 3.11 Venues near Chennai

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adugodi	12.942847	77.610416	PVR IMAX	12.934595	77.611321	Movie Theater
1	Adugodi	12.942847	77.610416	Lot Like Crêpes	12.936421	77.613284	Creperie
2	Adugodi	12.942847	77.610416	Tommy Hilfiger	12.934552	77.611347	Clothing Store
3	Adugodi	12.942847	77.610416	Zingron - Naga Kitchen	12.936271	77.615051	Indian Restaurant
4	Adugodi	12.942847	77.610416	Koramangala Social	12.935518	77.614097	Lounge
5	Adugodi	12.942847	77.610416	Fenny's Lounge & Kitchen	12.935146	77.613536	Lounge
6	Adugodi	12.942847	77.610416	Dyu Art Cafe	12.937289	77.617591	Financial or Legal Service
7	Adugodi	12.942847	77.610416	XOOX Brewmill	12.935507	77.614982	Brewery
8	Adugodi	12.942847	77.610416	Chai Chowk	12.936329	77.614667	Tea Room
9	Adugodi	12.942847	77.610416	PVR Cinemas	12.934389	77.611184	Multiplex

Figure 3.12 Venues near Bangalore

3.3.5 Exploratory data analysis (EDA)

Preliminary step in data analysis is to summarize main characteristics, gain better understanding of data set, uncover relationship between variables and extract important variables. Obtained venue information from foursquare is then explored for the following information–

- Total number of venues
- Number of venues near each neighborhood
- Unique categories of venues

Total number of venues

Total number of venues near Chennai and Bangalore are around 3368 and 4600 venues respectively as shown in Figure 3.13 and Figure 3.14.

```
print(chennai_venues.shape)
chennai_venues.head(10)
```

```
(3368, 7)
```

Figure 3.13 Number of venues near Chennai

```
print(bangalore_venues.shape)
bangalore_venues.head(10)
```

```
(4600, 7)
```

Figure 3.14 Number of venues near Bangalore

Number of venues near each neighborhood

Venues are group based on their neighborhood and the sum or count of all venues near each neighborhood is obtained as shown in Figure 3.15 and Figure 3.16.

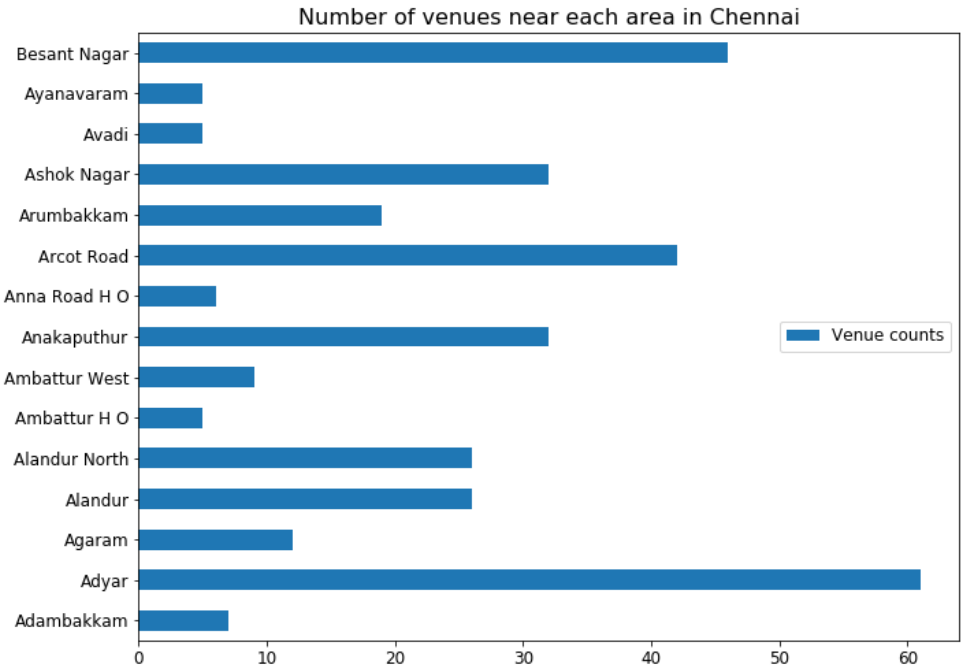

Figure 3.15 Number of venues near each area in Chennai

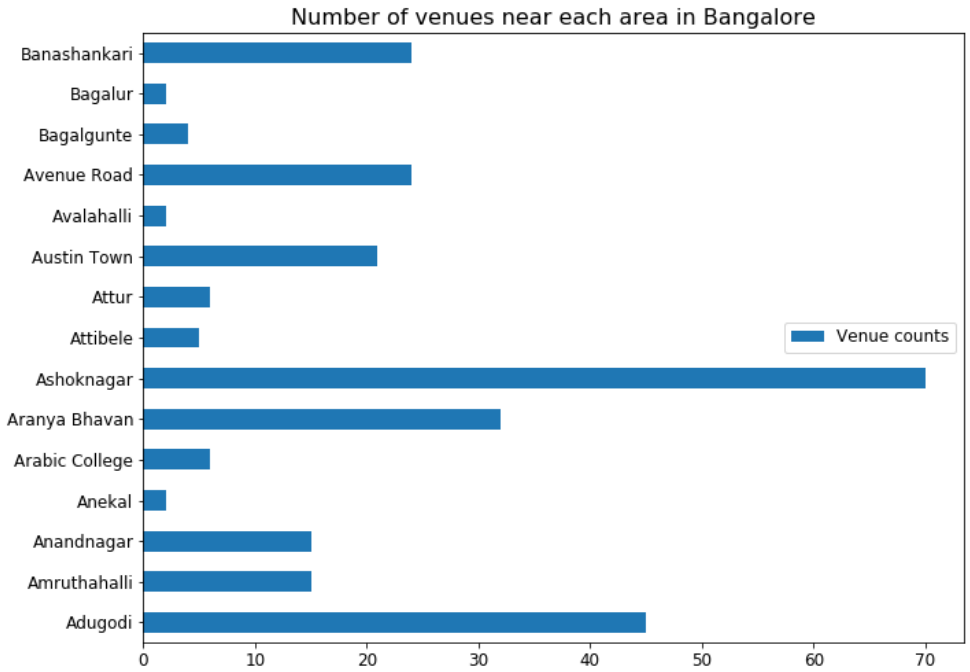

Figure 3.16 Number of venues near each area in Bangalore

Unique categories of venues

Number of unique categories of venues returned by foursquare API is calculated for neighborhoods near Chennai and Bangalore. As show in the Figure 3.17 it is observed that there are 189 and 249 unique categories of venues are present in Chennai and Bangalore respectively.

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

```
There are 189 uniques categories.
```

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

```
There are 249 uniques categories.
```

Figure 3.17 Unique categories of venues near Chennai and Bangalore

3.3.6 Data Wrangling

In this step the data is preprocessed in order to use it for clustering purpose. One hot encoding is used to add dummy variables for each unique category and assign 0 or 1 to each dummy variable. After encoding categorical variables, neighborhoods are

grouped and the mean frequency of occurrence of each venue category is calculated. The grouped data is stored in dataframe that contains ‘n’ rows with each neighborhood and ‘m’ columns with each unique venue category. Grouped data for Chennai and Bangalore are shown in Figure 3.18.

```
chennai_grouped.shape
```

```
(144, 186)
```

```
bangalore_grouped.shape
```

```
(164, 250)
```

Figure 3.18 Chennai and Bangalore grouped data based on neighborhood and venue category

It can be seen that the data is grouped with respect to 144 neighborhoods having the 186 venue categories and 164 neighborhoods having the 249 venue categories containing average occurrence of each category for Chennai and Bangalore respectively. Now the grouped data is sorted to identify top ten most commonly occurred venue category for each neighborhood. Top ten venues for each neighborhood is shown in Figure 3.19 and Figure 3.20.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adambakkam	Indian Restaurant	Pizza Place	Jewelry Store	Modern European Restaurant	IT Services	Hyderabadi Restaurant	Food & Drink Shop	Food	Flower Shop	Flea Market
1	Adyar	Indian Restaurant	Café	Pizza Place	Department Store	Dessert Shop	Fast Food Restaurant	Rock Club	Bakery	North Indian Restaurant	Electronics Store
2	Agaram	Multiplex	Pizza Place	Department Store	Train	Light Rail Station	Shopping Mall	Train Station	Farm	Food & Drink Shop	Food
3	Alandur	Indian Restaurant	Pizza Place	Breakfast Spot	Train Station	Hotel	Café	Arts & Crafts Store	Metro Station	South Indian Restaurant	Fast Food Restaurant
4	Alandur North	Indian Restaurant	Pizza Place	Breakfast Spot	Train Station	Hotel	Café	Arts & Crafts Store	Metro Station	South Indian Restaurant	Fast Food Restaurant

Figure 3.19 Top ten most common venue for each area in Chennai

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adugodi	Indian Restaurant	Café	Lounge	Dessert Shop	Coffee Shop	Bookstore	Multiplex	Donut Shop	Brewery	Salon / Barbershop
1	Amruthahalli	Indian Restaurant	Bus Station	Hotel	Brewery	Chinese Restaurant	Fast Food Restaurant	Bubble Tea Shop	Pizza Place	Ice Cream Shop	History Museum
2	Anandnagar	Indian Restaurant	Pizza Place	Snack Place	Fast Food Restaurant	Motorcycle Shop	Outdoor Supply Store	Coffee Shop	Café	Bus Station	Hotel
3	Anekal	ATM	Donut Shop	Flower Shop	Flea Market	Financial or Legal Service	Field	Fast Food Restaurant	Farmers Market	Event Space	Event Service
4	Arabic College	Pizza Place	Department Store	North Indian Restaurant	Optical Shop	Indian Restaurant	Pool Hall	Dry Cleaner	Dumpling Restaurant	Eastern European Restaurant	Electronics Store

Figure 3.20 Top ten most common venue for each area in Bangalore

Now the data is obtained, explored and prepared, hence the data is ready to be used for clustering and analyzing.

CHAPTER 4. RESULTS

4.1. IDENTIFYING SIMILAR NEIGHBORHOODS

Neighborhoods of both home and job location are clustered and segmented for similarities and differences. The process is clearly explained in section 3.2. Hence we'll go through each step one by one. In this project K-Means clustering algorithm is used to for grouping and number of clusters are chosen to be five for both job and home location clustering. The grouped data frame containing mean frequency of occurrence unique venue category for each neighborhood is used for clustering.

4.2. CLUSTERING HOMETOWN LOCATION

Clustering information is included in the Chennai grouped data and visualized using folium map. Chennai map with clustering information is shown in Figure 4.1.

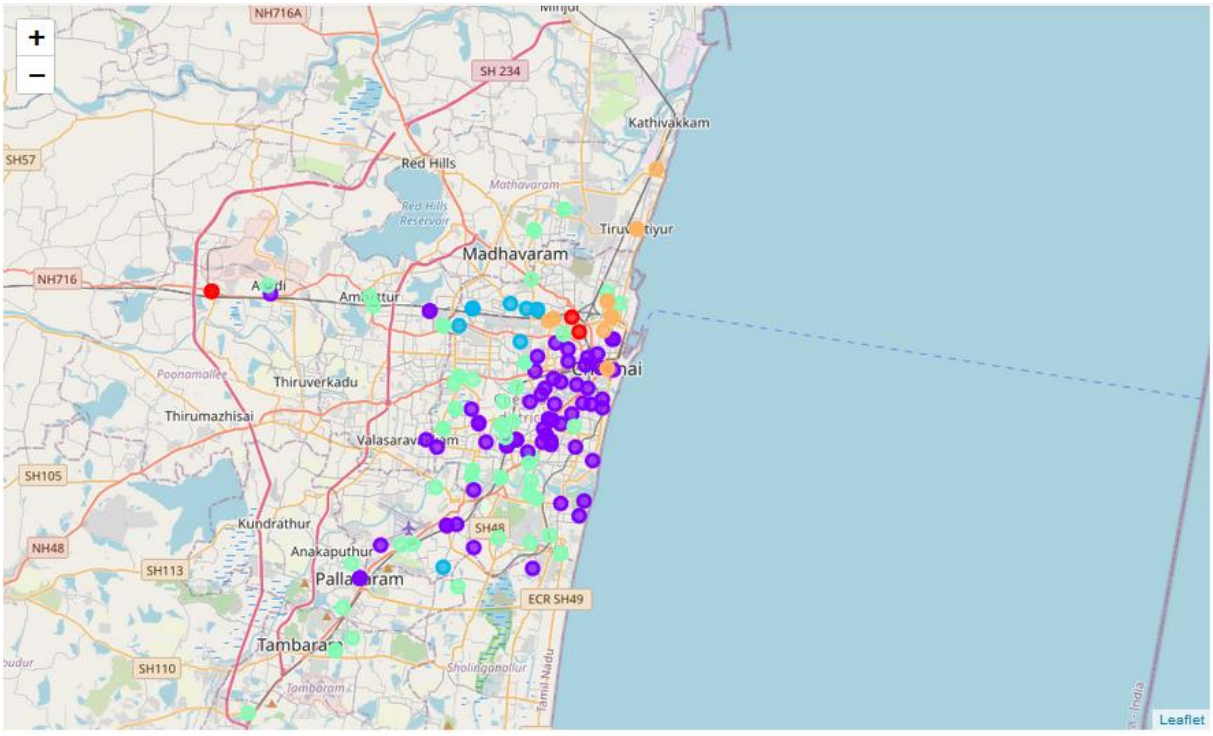

Figure 4.1 Chennai map with clustering information

4.3. FILTER REGION OF INTEREST IN HOME LOCATION

In hometown few locations of interest are chosen for comparison with job location. The filtered locations are shown in Figure 4.2 and Figure 4.3.

	Location	Pincode	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Adyar	600020	4.0	Indian Restaurant	Café	Pizza Place	Dessert Shop	Coffee Shop	North Indian Restaurant	Ice Cream Shop	Rock Club	Fast Food Restaurant	Electronics Store
22	Ashok Nagar	600083	4.0	Indian Restaurant	Pizza Place	Bakery	Snack Place	Diner	Sandwich Place	Fast Food Restaurant	South Indian Restaurant	Restaurant	Café
111	Kodambakkam	600024	3.0	Bakery	Jewelry Store	Juice Bar	Ice Cream Shop	Lounge	Deli / Bodega	Department Store	Restaurant	Chinese Restaurant	Electronics Store
112	Kodambakkam West	600024	3.0	Bakery	Jewelry Store	Juice Bar	Ice Cream Shop	Lounge	Deli / Bodega	Department Store	Restaurant	Chinese Restaurant	Electronics Store
143	Mambalam East	600033	4.0	Indian Restaurant	Clothing Store	Jewelry Store	Snack Place	Women's Store	Park	South Indian Restaurant	Asian Restaurant	Pizza Place	Coffee Shop
144	Mambalam R S	600033	4.0	Indian Restaurant	Clothing Store	Jewelry Store	Snack Place	Women's Store	Park	South Indian Restaurant	Asian Restaurant	Pizza Place	Coffee Shop
279	Velacheri	600042	3.0	Indian Restaurant	Vegetarian / Vegan Restaurant	Pizza Place	Coffee Shop	Sandwich Place	Café	Clothing Store	Chinese Restaurant	Fast Food Restaurant	BBQ Joint
295	West Mambalam	600033	3.0	Jewelry Store	Clothing Store	Indian Restaurant	South Indian Restaurant	Café	Furniture / Home Store	BBQ Joint	Chinese Restaurant	Electronics Store	Coffee Shop

Figure 4.2 Dataframe of filtered locations in Chennai

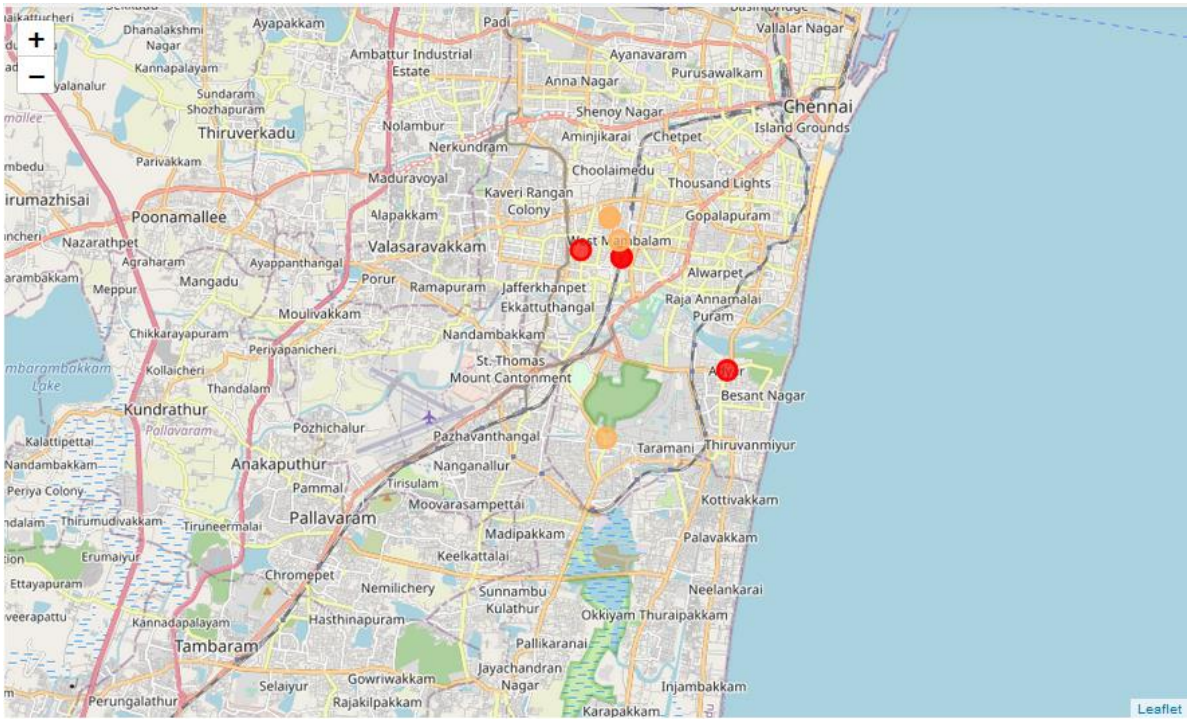

Figure 4.3 Filtered locations on Chennai map

We can see that filtered locations in Chennai fall under two different clusters. For choosing a similar location in Bangalore it is required that locations that fall under either one of the two segmented groups.

4.4. CLUSTERING JOB LOCATION

Bangalore grouped data is clustered and visualized using folium map. Bangalore map with clustering information is shown in Figure 4.4.

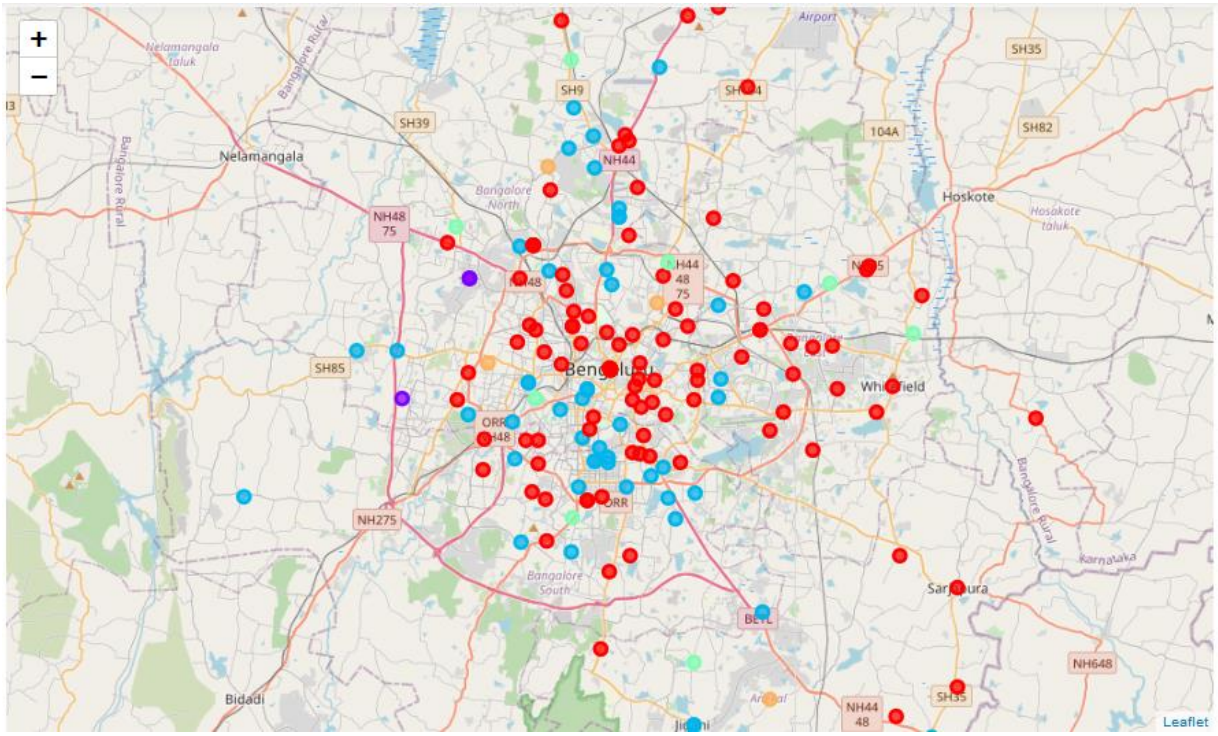

Figure 4.4 Bangalore map with clustering information

4.5. FILTERING AREAS WITHIN 10 KM RADIUS IN JOB LOCATION

In order to filter the regions within 10 Km from office location, it is necessary to calculate distance between two geographical coordinates. The job location is chosen to be Whitefield whose latitude longitude coordinates are shown in Figure 4.5. Now

that we have fixed the job location the radial distance between to geographical coordinates is calculated using Geodesic module from Geopy distance library.

```
wlat, wlon = bangalore[bangalore['Location'].
                       str.contains('Whitefield')][
                        ['Latitude', 'Longitude']].iloc[0,:]
print (wlat, wlon)
```

```
12.9696365 77.7497448
```

Figure 4.5 Geo-coordinates of Whitefield

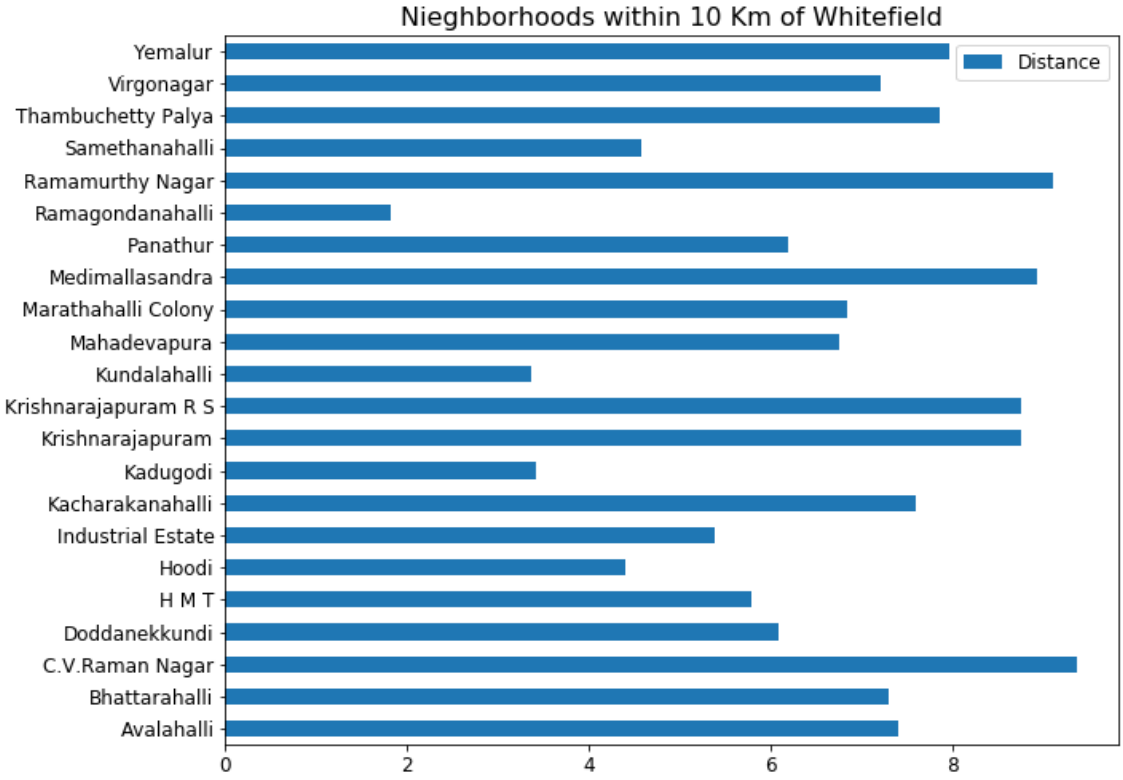

Figure 4.6 Neighborhoods within 10 Km of Whitefield

The distance from Whitefield is calculated for all neighborhoods in Bangalore and locations whose distance value is less than 10 Km are chosen for further calculation. Distances of neighborhoods in Bangalore within 10 Km from Whitefield are shown in Figure 4.6.

4.6. COMBINE PREFFERED LOCATIONS

Now that we have filtered important locations from Chennai and Bangalore, all the important locations are combined and stored in single pandas dataframe. Note that we have the cluster information for each neighborhood based on the respective cities which are not required, hence cluster labels are dropped. Important locations are rendered on a folium map as shown in Figure 4.7.

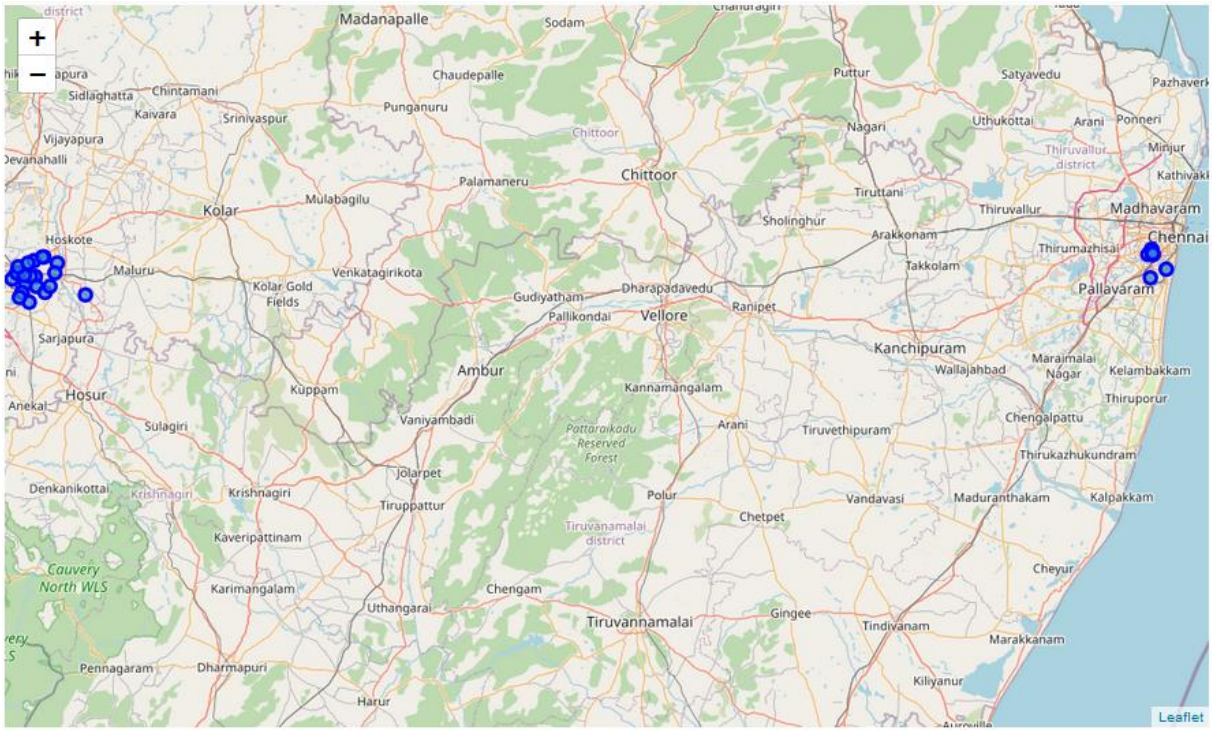

Figure 4.7 Important locations

4.7. CLUSTERING COMBINED LOCATIONS

One hot encoding information for the combined important locations, might include ‘Nan’ values as the unique category of venues in both the cities are not same. Hence, the ‘NaN’ values are replaced by zero and then the data is clustered using K-Means into five similar groups. Important locations with cluster information is shown in Figure 4.8.

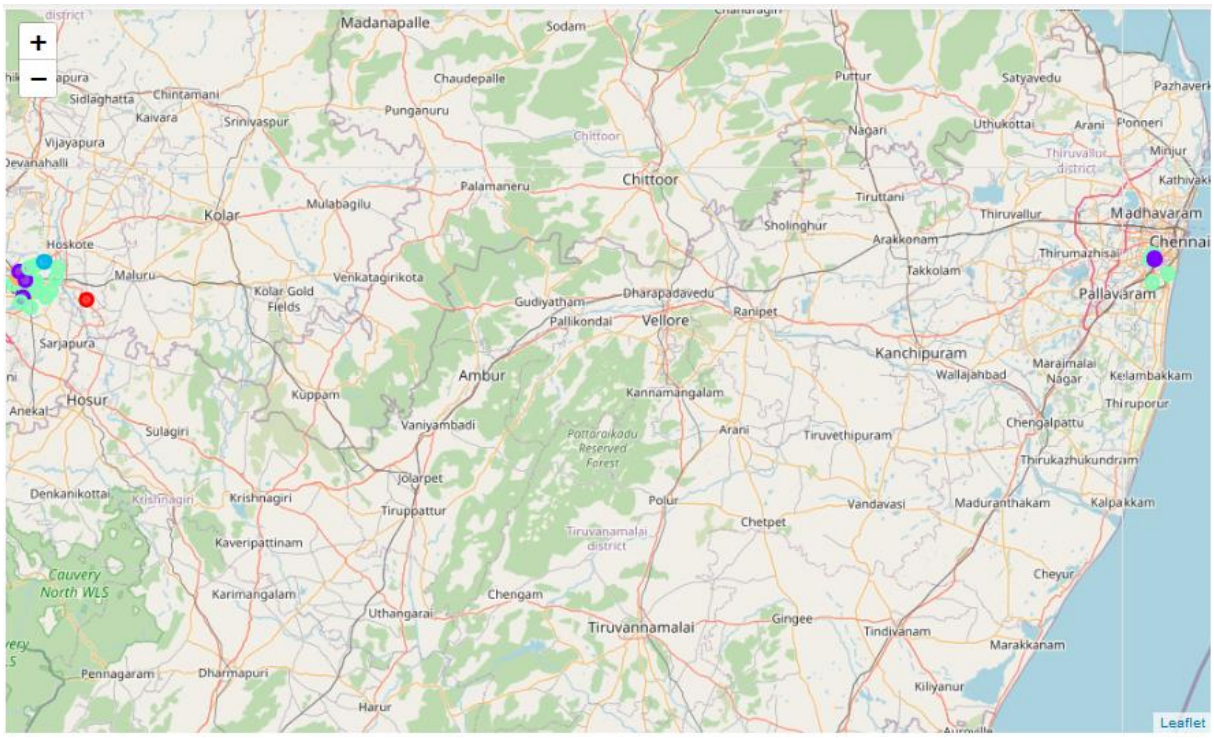

Figure 4.8 Important locations with cluster information

4.8. EXPLORE AND VISUALIZE CLUSTERS

Let us explore the individual cluster groups for important locations. Important location clusters centered on Chennai and Bangalore region shown in Figure 4.9.

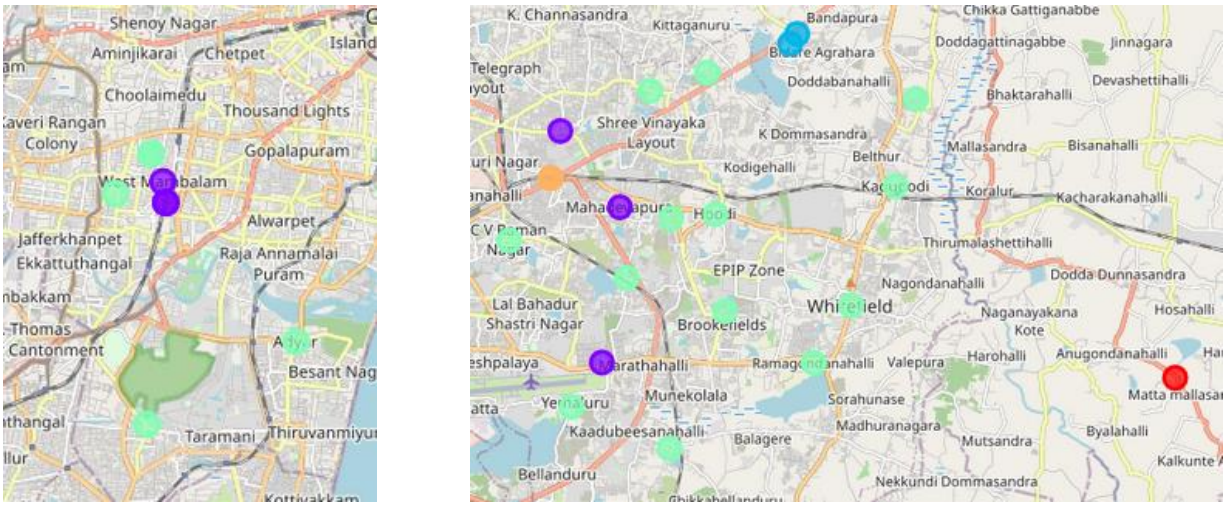

Figure 4.9 Important location cluster Chennai centered (left) and Bangalore centered (right)

	Location
143	Mambalam East
144	Mambalam R S
295	West Mambalam
177	Mahadevapura
185	Marathahalli Colony
234	Ramamurthy Nagar

Cluster = 0

	Location
15	Avalahalli
281	Virgonagar

Cluster = 1

	Location
2	Adyar
22	Ashok Nagar
111	Kodambakkam
112	Kodambakkam West
279	Velacheri
43	Bhattarahalli
53	C.V.Raman Nagar
78	Doddanekkundi
95	H M T
106	Hoodi
119	Industrial Estate
146	Kadugodi
169	Kundalahalli
218	Panathur
232	Ramagondanahalli
263	Thambuchetty Palya
288	Whitefield
295	Yemalur

Cluster = 2

	Location
164	Krishnarajapuram
165	Krishnarajapuram R S

Cluster = 3

	Location
192	Medimallasandra

Cluster = 4

Figure 4.10 Individual cluster groups of important locations

CHAPTER 5. DISCUSSION

5.1. RECOMMENDING AREAS IN JOB LOCATION

Now that we have clustered the important locations in Chennai and Bangalore, it can be observed from individual cluster groups that clusters 1, 3 and 4 does not contain any area from Chennai, in other words, these cluster groups are not similar to areas in Chennai. All the important locations in Chennai are grouped into clusters 0 and 2, hence it is a good choice to select an area in job location that falls in these two clusters. Since there are many locations in cluster groups 0 and 2, in order to reduce the list to make a wiser choice, each neighborhood's distance from Whitefield can be used as a selection criteria.

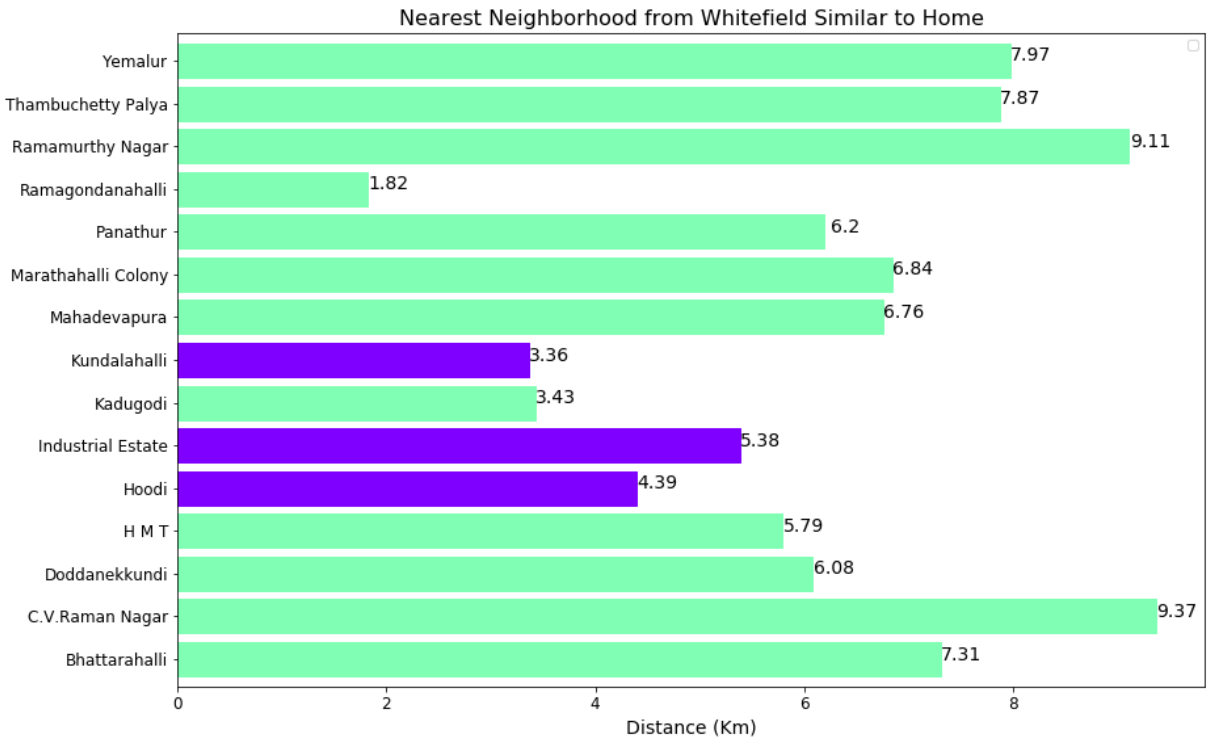

Figure 5.1 Nearest neighborhood from Whitefield similar to home.

Distances of each Bangalore neighborhood from Whitefield, in cluster groups 0 (blue bars) and 2 (green bars) are shown in Figure 5.1. It can be seen from the bar

chart that three neighborhoods Kundhalahalli, Industrial estate and Hoodi fall in cluster group 0, which are similar to West mambalam regions in Chennai. Also, remaining twelve regions including Kadugodi, Ramgondanhalli etc. fall in cluster group 2, which are similar to areas such as Adyar, Velachery and Kodambakkam in Chennai.

Optimal choice would be to recommend the first two closest locations in each cluster group without loss of generality. Hence, the areas in job locations similar to home are Ramgondanhalli, Kundalahalli, Kadugodi and Hoodi which are less than 5 Km from Whitefield location.

CHAPTER 6. CONCLUSION

In this project, a particular case of relocation where an individual would want to relocate from hometown for a new job is chosen and locations near job place that are similar to places in hometown are identified using Machine learning techniques. This is a preliminary work, hence the focus is restricted to identifying similar locations. More features can be included in future such as hosing prices, transportation facilities nearby, weather etc. to make the algorithm more robust in recommending similar locations.

Python modules used in this project include –

1. BeautifulSoup from BS4 - Library to scrape information from web pages
2. Requests from Urllib - Library to open URLs
3. Pandas - Library for data analysis
4. Json - Library to handle JSON files
5. Requests - Library to handle API requests
6. Folium - Map rendering library
7. Nominatim from Geopy Geocoders - Libary to convert address into latitude and longitude values
8. Geodesic from Gropy Distance - Library to calculate geodesic distance between two points
9. Numpy - Library to handle vectorized data
10. Scikitlearn - K-Means clustering algorithm from Machine learning library
11. Matplotlib - Associated plotting modules