RESTAURANT LOCATION FINDER

A CAPSTONE PROJECT REPORT

RINAV SAXENA

DATED - 19082020

CONTENTS

INTRODUCTION

- Good Morning/Afternoon/Evening. This report is about the Capstone project on Coursera.
- Many people are shifting their focus from 9 to 5 jobs to businesses. One such business is the restaurant business and therefore location is the first problem which takes the spotlight.
- One such example would be the growth of restaurants seen particularly where IT parks are coming up. After living 3 years in Chennai (2016-2019), I witnessed an increase in the number of North Indian Restaurants coming up towards the outskirts of Chennai where IT parks were and are dominant.
- Thus, the question becomes relevant here - Which location is optimum to come up with a restaurant?
- In this report, I will be going over one such solution as to how does one go about selecting a restaurant. There was a time when someone had asked me this very question I did not know as to how to go about selecting an area. I had some inputs/ideas as to how would one would decide a location for their restaurant after the decision was made.
- Now, any problem can be solved quickly if its more specific and to the point. It's this very reason for choosing this topic.

INTRODUCTION CONTD.

- So, first let me define the statement that I will be trying to find a solution for – “How to determine where to place a Restaurant based on some data available?”
- The above statement is actually a super-set of the below statements which are more explicit, direct and specific:
    - What type of restaurant does one want to put up?
    - In which locality are they wanting to target? – (Inside the city, Outskirts of the city or somewhere in between)
    - What is the budget and consequently, prices of land?
    - As per requirements what space are they willing to acquire?
    - What is the competition nearby as per selections made? (Top-Most common Venue)
    - Where would they get the raw material from?
    - Population density in each area. (City, Country-Side, In between)

So, how does one go about and Compute the expected solution of the Super-Set statement based on the expected outcomes of the sub-set statements and their weights to the original statement? In this report, I will explore the answers to some statements based on Location Data available on Foursquare and the city of Pune (India) is chosen. For this project I will answer the above Macro Statement using the available location based data (Points 2, 5 and 7)

COLLECTION OF DATA

- This section contains 3 parts.
  - **Importing data onto notebook.** -> Getting pin codes of Pune City with their Lat/Long into a dataframe in the notebook.
  - **Importing data onto a map.** -> plotting these lat/long on a map using folium library.
  - **Fetching Foursquare location data.** -> summarizing a table which contains all buildings (with category) within 1000 m radius with the Pin Code as reference.

COLLECTION OF DATA –

IMPORTING DATA ONTO NOTEBOOK.

- Lets start with the choosing of a reference point for our search. Fixed pin codes for Pune district are chosen. There are around 34 unique pin codes and 151 unique areas in the city of Pune. The link to the raw file extracted is available below:
- https://github.com/sanand0/pincode/blob/master/data/IN.csv
- The csv file was downloaded and is saved on my laptop. When loaded, the table is shown below on Jupyter Notebook converted into a Dataframe.

```
pune_df
```

Out[19]:

	Pin Code	Area Name	State Name	District Name	Taluka Name	Latitude	Longitude	Accuracy
0	411001	Pune	Maharashtra	Pune	Pune City	18.5196	73.8554	4
1	411001	N.W. College	Maharashtra	Pune	Pune City	18.5196	73.8554	3
2	411001	Dr.B.A. Chowk	Maharashtra	Pune	Pune City	18.5196	73.8554	3
3	411001	Pune Cantt East	Maharashtra	Pune	Pune City	18.5196	73.8554	3
4	411001	Pune New Bazar	Maharashtra	Pune	Pune City	18.5196	73.8554	3
...	...	...	...	...	...	...	...	...
146	412307	Manjari BK	Maharashtra	Pune	Haveli	18.5000	73.9771	1
147	412307	Manjari Farm	Maharashtra	Pune	Haveli	18.5000	73.9771	1
148	412308	Phursungi	Maharashtra	Pune	Haveli	18.4361	73.9712	4
149	412308	Vadki	Maharashtra	Pune	Haveli	18.4361	73.9712	3
150	412308	Uruli Devachi	Maharashtra	Pune	Haveli	18.4361	73.9712	4

151 rows × 8 columns

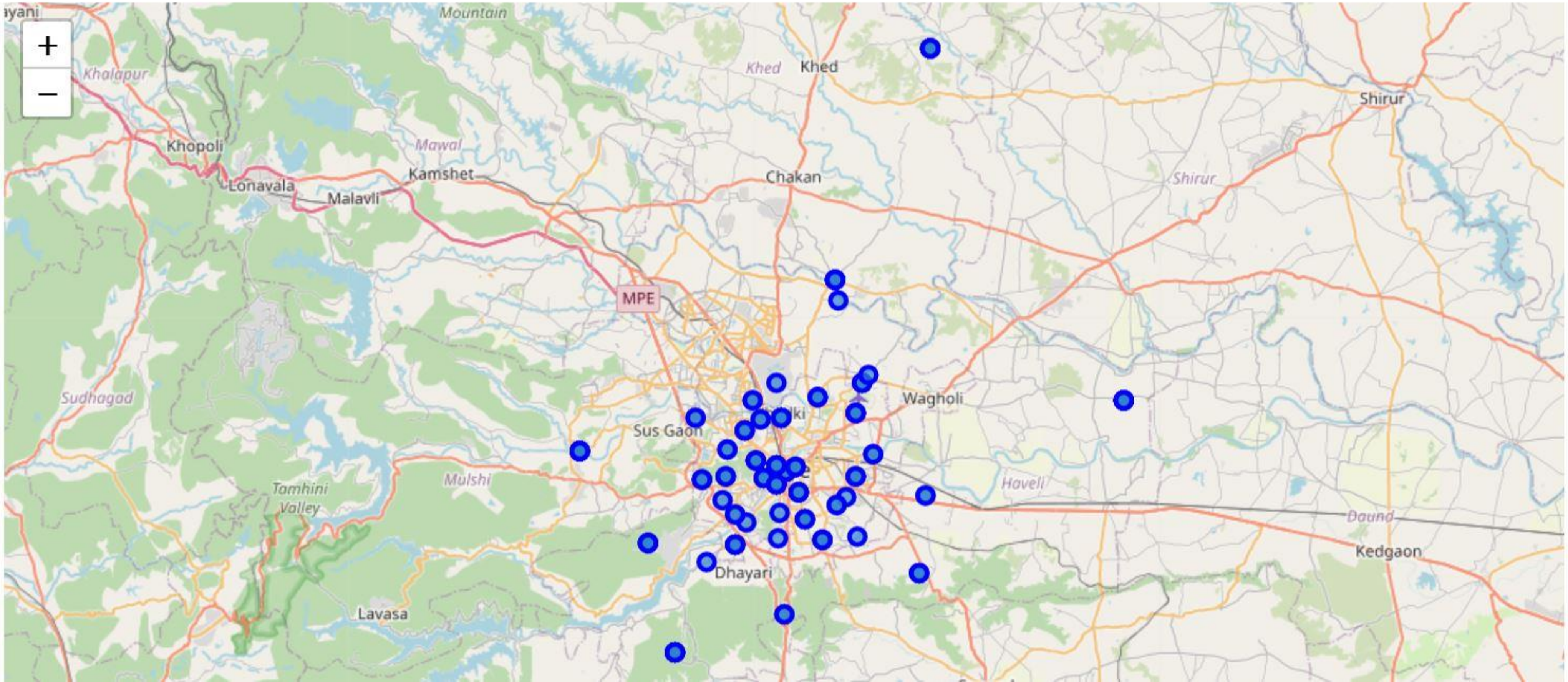
COLLECTION OF DATA –

IMPORTING DATA ONTO A MAP.

- The dataframe pune_df is then plotted on a folium map with Pin Codes as center.

```
map_pun
```

Out[22]:



COLLECTION OF DATA –

FETCHING FOURSQUARE LOCATION DATA.

- Finally, after creating the login to foursquare server, I fetch all buildings within 1000 m radius of the pin code and finally will combine this data with the original pin code lat/long.
- First, a request is given to foursquare server to fetch a list containing all venues nearby.

```
In [23]: import requests

         results = requests.get(url).json()
         results
```


```
            'type': 'general',
            'reasonName': 'globalInteractionReason'}]},
        'venue': {'id': '4f579db56b740547b24e5d3a',
         'name': 'Lal Mahal',
         'location': {'address': 'Corner of Kasba Ganpati Mandir,',
          'crossStreet': 'off Shivaji Road,',
          'lat': 18.518719674058865,
          'lng': 73.85655641555786,
          'labeledLatLngs': [{'label': 'display',
            'lat': 18.518719674058865,
            'lng': 73.85655641555786}],
          'distance': 156,
          'postalCode': '411030',
          'cc': 'IN',
          'city': 'Pune',
          'state': 'Mahārāshtra',
          'country': 'India',
          'formattedAddress': ['Corner of Kasba Ganpati Mandir, (off Shivaji Road,)',
           'Pune 411030',
           'Mahārāshtra',
```

COLLECTION OF DATA – CONTD.

FETCHING FOURSQUARE LOCATION DATA.

- Second, I explore the data with the type of category of building near all pin codes and generate the following table. This is just to get a list of all buildings nearby.

```
nearby_venues
```

```
C:\Users\DELL\anaconda3\lib\site-packages\ipykernel_launcher.py:6: FutureWa
use pandas.json_normalize instead
```

Out[25]:

	name	categories	lat	lng
0	Lal Mahal	Historic Site	18.518720	73.856556
1	Hotel Madhuban	Tea Room	18.519248	73.848688
2	Bhagat Tarachand	Indian Restaurant	18.514332	73.851317
3	Sujata Mastani	Ice Cream Shop	18.511793	73.852145
4	Krishna Juice Bar	Juice Bar	18.523553	73.847651
5	Fish Curry Rice	Seafood Restaurant	18.516415	73.850934
6	New Poona Bakery	Bakery	18.517028	73.854845
7	Raja Dinkar Kelkar museum	History Museum	18.510744	73.854389
8	Café Coffee Day	Coffee Shop	18.523131	73.848347
9	Mohan Ice Cream	Ice Cream Shop	18.520620	73.846070
10	Shaniwar Wada	Historic Site	18.520151	73.855187
11	Shrikrishna Bhuvan	Snack Place	18.513494	73.855074

COLLECTION OF DATA – CONTD.

FETCHING FOURSQUARE LOCATION DATA.

- Finally, collecting all this data and storing it in a final table - pune_venues as shown below:
- This ends the data_collection phase as I will be working on this data to finalize the location of a restaurant.

```
pune_venues
```

(813, 8)

Out[83]:

	Area Name	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Pune	18.5196	73.8554	Lal Mahal	18.518720	73.856556	Historic Site
1	Pune	18.5196	73.8554	Shaniwar Wada	18.520151	73.855187	Historic Site
2	N.W. College	18.5196	73.8554	Lal Mahal	18.518720	73.856556	Historic Site
3	N.W. College	18.5196	73.8554	Shaniwar Wada	18.520151	73.855187	Historic Site
4	Dr.B.A. Chowk	18.5196	73.8554	Lal Mahal	18.518720	73.856556	Historic Site
5	Dr.B.A. Chowk	18.5196	73.8554	Shaniwar Wada	18.520151	73.855187	Historic Site
6	Pune Cantt East	18.5196	73.8554	Lal Mahal	18.518720	73.856556	Historic Site
7	Pune Cantt East	18.5196	73.8554	Shaniwar Wada	18.520151	73.855187	Historic Site
8	Pune New Bazar	18.5196	73.8554	Lal Mahal	18.518720	73.856556	Historic Site
9	Pune New Bazar	18.5196	73.8554	Shaniwar Wada	18.520151	73.855187	Historic Site

METHODOLOGY

- In this section, I highlight as to how it can be decided as to which is the optimum location to deploy a restaurant by using clustering.
- Below, I have created a dictionary which gives a score to a category type based on the following criteria:
    - Lets say, I would like my restaurant to be place near a 'Historical Venue’, 'Gym', 'Park' or a 'theatre' its assigned a score of 1.
    - Its obvious that I would not be looking at places which are saturated with the number of restaurants, so I give a score of -1 for these buildings.
    - For neutral venues, a score of 0 is assigned. For example, an 'Arts & Crafts Store', 'Department Store', 'ATM', etc.

METHODOLOGY – CONTD.

- Then, I assign each Area the score according to the number of buildings and their scores as per the mapping explained in the previous slide and I get the result in a table - final_score_df.

```
In [83]: print(pune_venues.shape)
         pune_venues
         pune_venues['Venue Category'].tolist()

         Score_Dict = {'Venue Category':['Historic Site','Bakery','Arts & Crafts Store','Theater','Badminton Court','Sandwich Place',
         'Home Service','Department Store','ATM','Bistro','Deli / Bodega','Indian Restaurant','Snack Place',
         'Chinese Restaurant','Gym / Fitness Center','Hotel','Restaurant','Garden','Café','Gym','Ice Cream Shop',
         'Bank','Bar','Juice Bar','Stadium','Coffee Shop','Maharashtrian Restaurant','Vegetarian / Vegan Restaurant',
         'Burger Joint','Fast Food Restaurant','Motorcycle Shop','Electronics Store','Shopping Mall','Bus Station',
         'Auto Garage','Clothing Store','Park','Convenience Store','Breakfast Spot','Seafood Restaurant','Flea Market',
         'Hotel Bar','Southern / Soul Food Restaurant','Cheese Shop','Pizza Place','Lounge','Dessert Shop',
         'South Indian Restaurant','Soccer Field','Fried Chicken Joint','Asian Restaurant','Italian Restaurant',
         'Nightclub','IT Services','Grocery Store','Diner','Food Court','Shop & Service','Basketball Court',
         'Wine Shop','Concert Hall','Indie Movie Theater','Farmers Market','Fruit & Vegetable Store',
         'Eastern European Restaurant','Arcade','Dance Studio','American Restaurant','History Museum','Donut Shop',
         'Shoe Store','Liquor Store','Business Service','Mobile Phone Shop','Flower Shop','Punjabi Restaurant','Brewery',
         'Resort','Greek Restaurant','River','Furniture / Home Store'], 'Venue Category Score':[1,-1,0,1,1,-1,0,0,0,-1,-1,-1,-1,-1,1,
                                                                                                  -1,-1,1,-1,1,-1,0,-1,-1,1,-1,
                                                                                                  -1,-1,-1,-1,0,0,1,1,0,0,1,0,-1,-1,
                                                                                                 1,-1,-1,0,-1,-1,-1,-1,1,-1,-1,-1,
                                                                                                 -1,1,0,-1,-1,0,1,-1,1,1,1,1,-1,1,0,
                                                                                                 -1,1,-1,0,-1,1,0,1,-1,0,-1,-1,1,0]}
```

```
final_score_df
```

Out[82]:

	Area Name	Venue Category Score
0	Pune	2
71	Pune	2
49	Dr.B.A. Chowk	2
51	Pune New Bazar	2
52	C D A (O)	2
53	Ghorpuri Bazar	2
54	Pune	2
55	N.W. College	2
56	Dr.B.A. Chowk	2
57	Pune Cantt East	2
58	Pune New Bazar	2

METHODOLOGY – CONTD. – K-MEANS

- By using K-means I selected the cluster count to be 6.
- First I created a table with only the Venue Category Score as a column as shown below.

```
clustering_df.head()
```

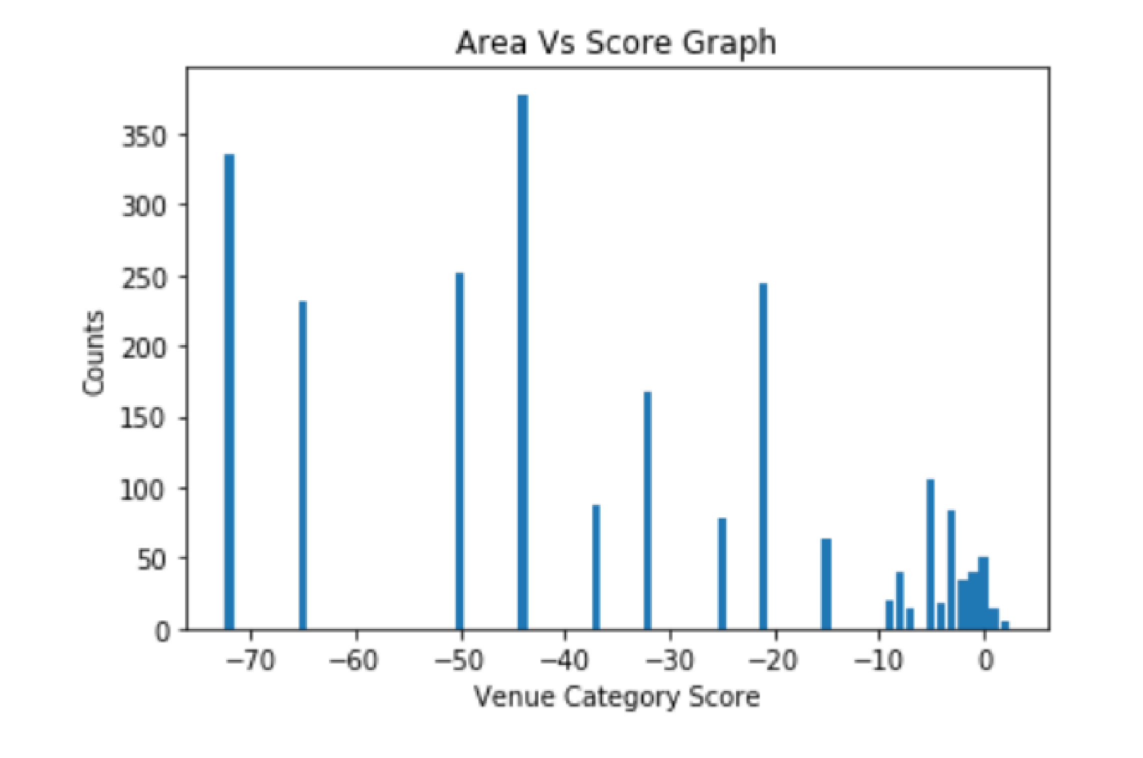
Out[68]:

	Venue Category Score
0	2
1	2
2	2
3	2
4	2

- Next, after running K-Means algorithm, I was able to plot the graph of Score vs No. of Clusters

METHODOLOGY – CONTD.

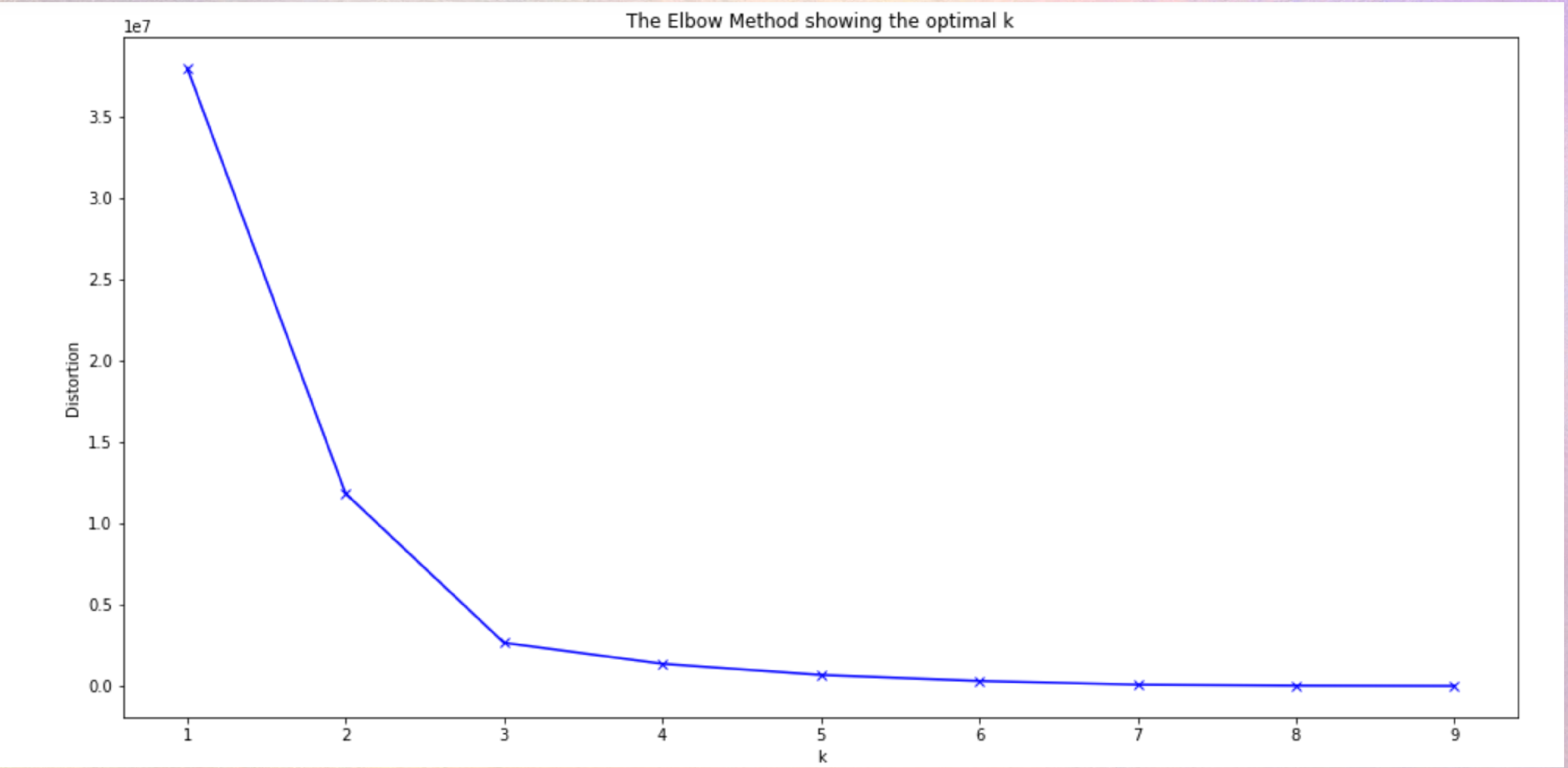
- Lets analyze the distribution of these scores by plotting the counts of these scores.
- As seen below, we can see that very less areas have a positive score, whereas many areas have received scores with < -20.
- When the count distribution of these scores are analyzed, its found that upto -15, there are very few areas that have got these scores.
- The count starts to increase once higher negative numbers are reached.



	Venue Category Score	Cluster	Counts
0	2	4	18
18	1	4	271
289	0	4	129
418	-1	4	127
545	-2	4	196
741	-3	4	102
843	-4	4	1191
2034	-6	4	49
2083	-7	4	291
2374	-8	4	192
2566	-9	4	484
3050	-15	1	441
3491	-16	1	800
4291	-23	1	9583
13874	-24	1	1875
15749	-31	5	9442
25191	-46	2	26136
51327	-51	2	14400
65727	-65	3	17787
83514	-73	0	26896

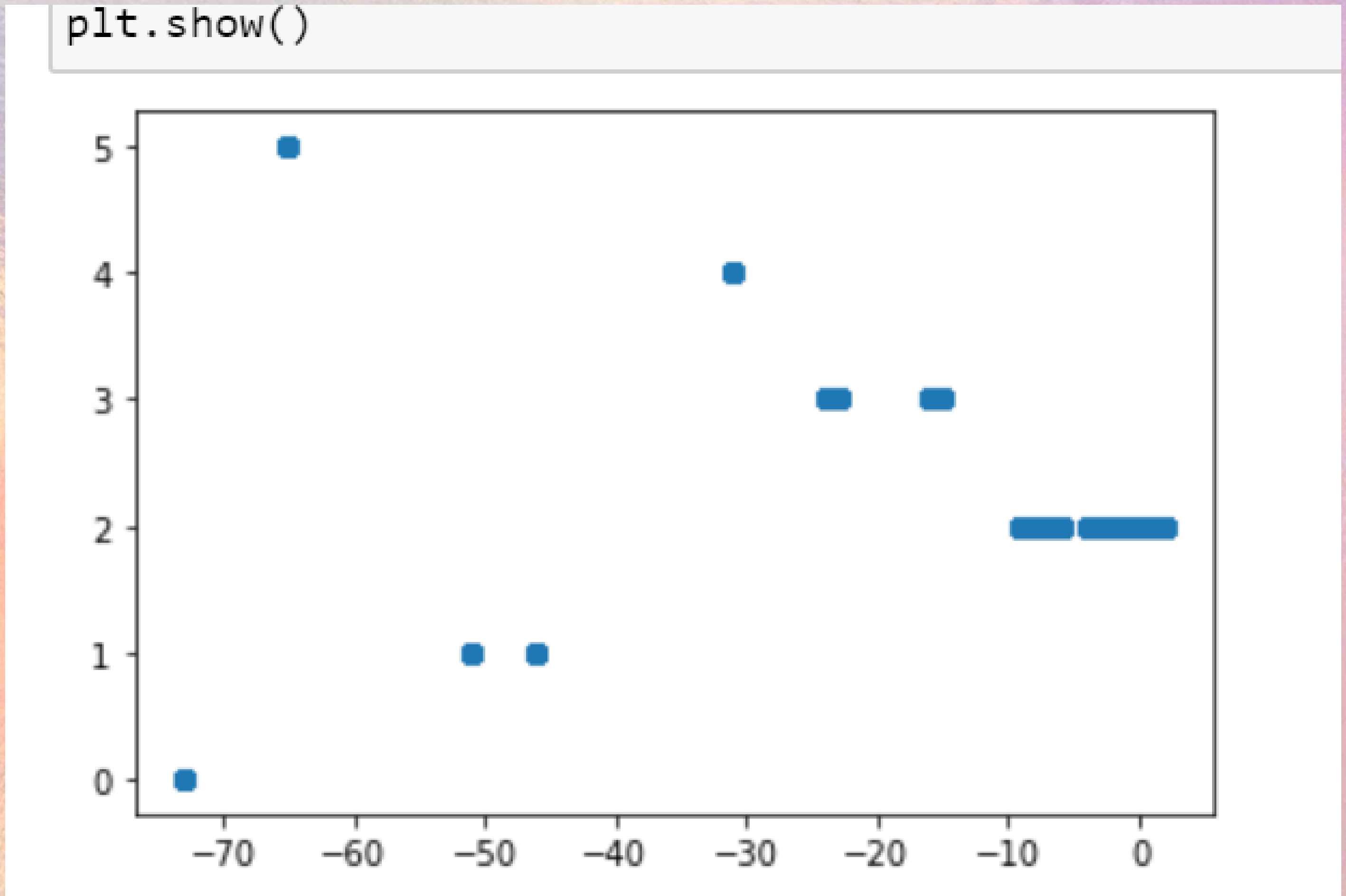
METHODOLOGY – CONTD. – K-MEANS

- As seen below, a value of K with 3 or 4 is the most optimum.



METHODOLOGY – CONTD. – K-MEANS

- Although, when I chose a K of 6, the below distribution came up. As shown, if I selected a lesser K, there would be more values falling under the 1st Cluster which is the optimum cluster for our location.
- A distribution is shown to the right. Thus K = 6 was chosen here.



METHODOLOGY – CONTD.

- Finally, color coding is applied as per the below criteria for K=6 clusters and the table is extended with Lat Long details.
- Color coding is applied as per below criteria and the below table is created after assigning this score to every area:
  - Green for > -10 score.
  - Orange for > -31 score.
  - Cyan for score = -31
  - Blue for > -52 score.
  - Magenta for score = -65.
  - Red for other score.

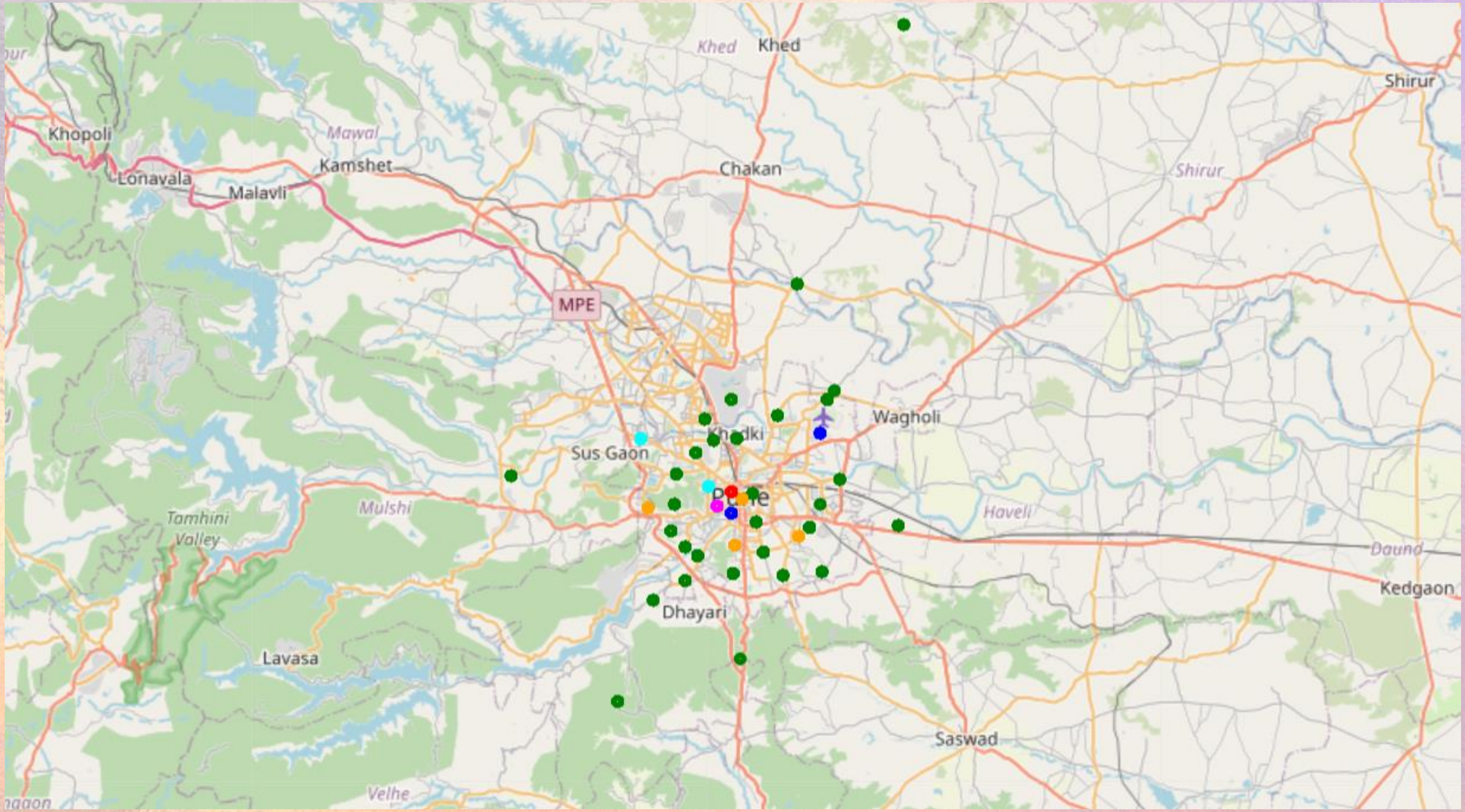
```
final_score_df.head()
```

:[74]:

	Area Name	Venue Category Score	Counts	Area Latitude	Area Longitude	Color_Coding	Pin Code
0	N.C.L. Pune	2	6	18.5379	73.8048	Green	411008
1	N.C.L. Pune	2	6	18.5379	73.8048	Green	411008
2	N.C.L. Pune	2	6	18.5379	73.8048	Green	411008
3	N.C.L. Pune	2	6	18.5379	73.8048	Green	411008
4	N.C.L. Pune	2	6	18.5379	73.8048	Green	411008

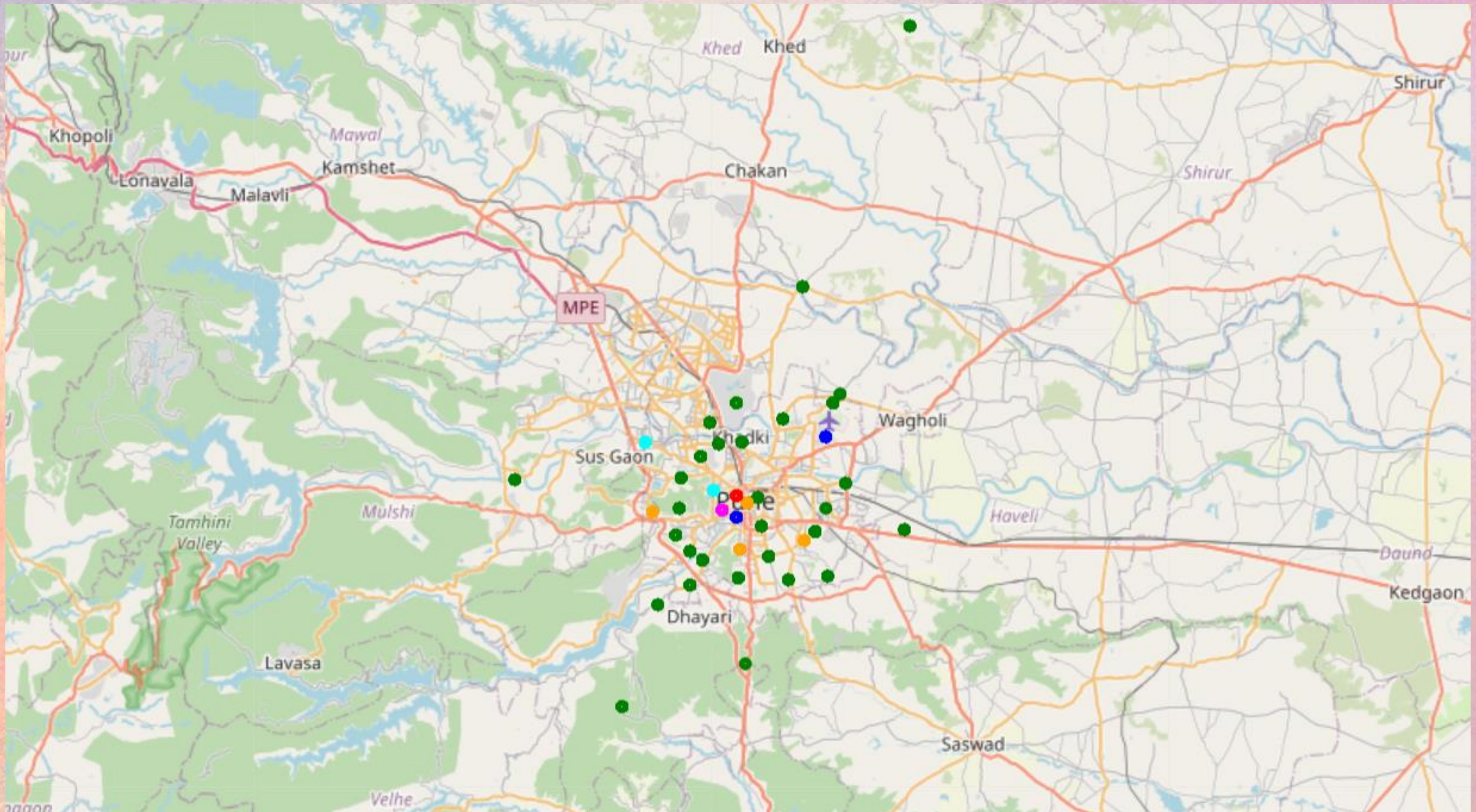
METHODOLOGY – CONTD.

- These are inserted into the map below and finally concludes the methodology section.



RESULTS

- Finally, Lets analyze the map more closely. There are many Green Pin Codes here which signifies that many pin codes have a good score of above -10.
- Next, I categorize these Green Pin Codes as per the following criteria:
  - Count of Venues/Buildings
  - Display the top-most venue category according to frequency.



RESULTS – CONTD.

- A) Count of Venues per pin code.
- First, I filter all the Green Color codes from the table final_score_df.
- Here, I create another table with only the count of venues which is calculated on the number of entries/rows of the pin code. This table is shown below.

```
Count_of_Venues_df.head()
```

Out[93]:

	0
Pin Code	
411001	9583
411002	6
411003	45
411004	17787
411005	26896

RESULTS – CONTD.

- B) Display the top-most venue category as per frequency.
- Next, the venue-category which is the most common in every green pin codes is created in a new table as shown below:

```
Most_Common_Venue_df

Out[82]: Pin Code
         411001        Indian Restaurant
         411002             Home Service
         411003         Department Store
         411004        Indian Restaurant
         411005        Indian Restaurant
         411006             Home Service
         411007               Restaurant
         411008              Bus Station
         411009              Coffee Shop
         411011            Historic Site
         411012             Dessert Shop
         411013            Historic Site
         411014                     Café
         411015        Indian Restaurant
         411016              Coffee Shop
         411020        Indian Restaurant
         411021                     Café
         411022            Shopping Mall
         411024                     Café
         411025            Historic Site
         411028           Ice Cream Shop
         411030        Indian Restaurant
         411031        Indian Restaurant
         411032                      ATM
```

RESULTS – CONTD.

- Finally, these two tables are combined with the final table – Green_Pin_Codes to create the table below:

```
Green_Pin_Codes_df
```

55137	411013	64	Historic Site	Green
55138	411013	64	Historic Site	Green
55139	411013	64	Historic Site	Green
55140	411013	64	Historic Site	Green
55141	411013	64	Historic Site	Green
69542	411015	108	Indian Restaurant	Green
69543	411015	108	Indian Restaurant	Green
69544	411015	108	Indian Restaurant	Green
69545	411015	108	Indian Restaurant	Green
69546	411015	108	Indian Restaurant	Green
69547	411015	108	Indian Restaurant	Green
69548	411015	108	Indian Restaurant	Green
69549	411015	108	Indian Restaurant	Green

RESULTS – CONTD.

- Below are all the categories the Green Pin Codes have which are the top-most common venues.

```
In [157]: Green_Pin_Codes_df['Venue Category'].unique()

Out[157]: array(['Home Service', 'Department Store', 'Restaurant', 'Bus Station',
                 'Coffee Shop', 'Historic Site', 'Dessert Shop',
                 'Indian Restaurant', 'Shopping Mall', 'Café', 'Ice Cream Shop',
                 'ATM', 'Eastern European Restaurant', 'Snack Place',
                 'Breakfast Spot', 'Bakery', 'Fast Food Restaurant',
                 'Business Service', 'Asian Restaurant', 'River'], dtype=object)
```

- Now, in this table I will filter out the pin codes which do not have a restaurant as the most common category and I finally arrive at the **optimum pin codes (12)** where a restaurant location can be advised to deploy.

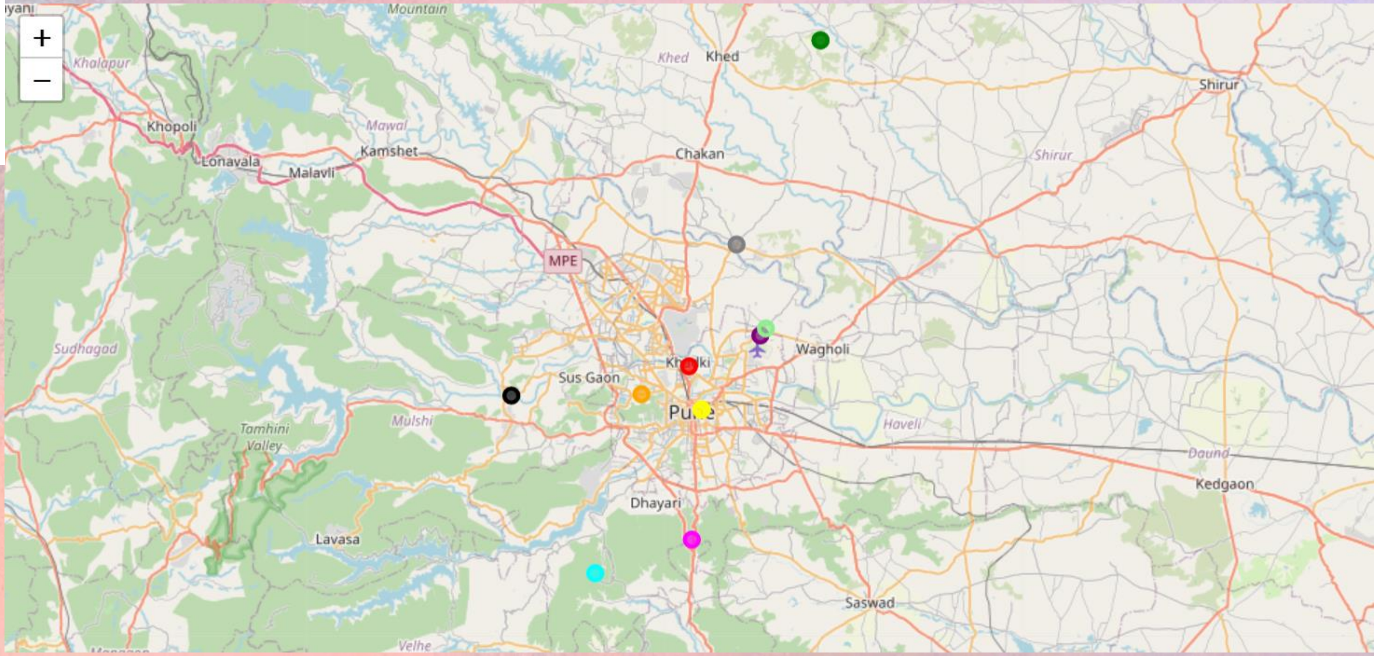
Out[160]:

	Pin Code	Count of Venues	Venue Category	Color_Coding
9583	411002	6	Home Service	Green
9589	411003	45	Department Store	Green
54317	411006	2	Home Service	Green
54394	411008	18	Bus Station	Green
54854	411011	192	Historic Site	Green
55078	411013	64	Historic Site	Green
78320	411025	24	Historic Site	Green
104508	411032	48	ATM	Green
109952	411046	2	Business Service	Green
109954	411047	4	ATM	Green
110251	412105	96	River	Green
110347	412115	13	River	Green

RESULTS – CONTD.

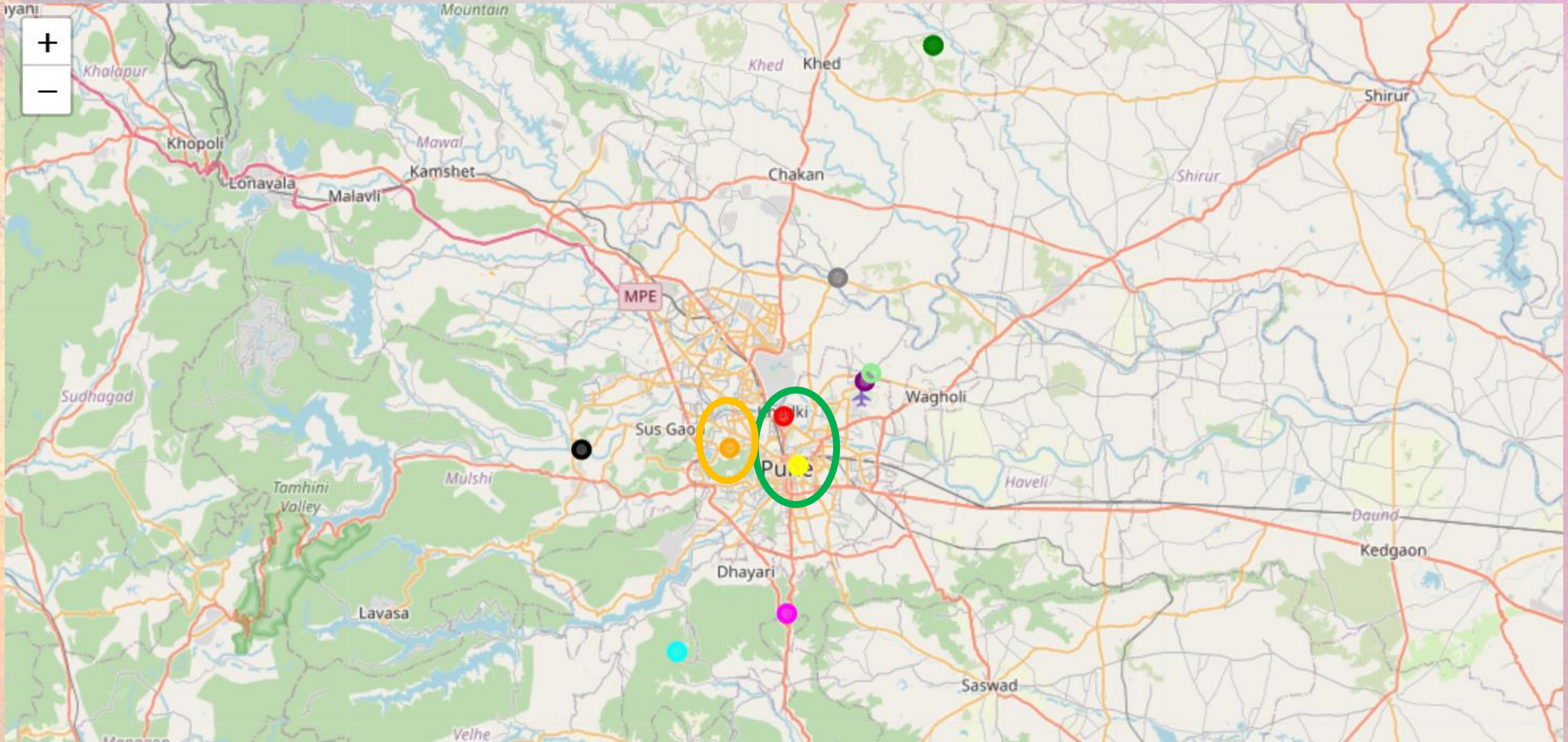
- Lets plot these on the map after assigning these pin codes a unique color.

	Pin Code	Count of Venues	Venue Category	Color_Coding	Area Latitude	Area Longitude
0	411002	6	Home Service	green	18.8671	73.9801
6	411003	45	Department Store	red	18.5639	73.8515
51	411006	2	Home Service	green	18.8671	73.9801
53	411008	18	Bus Station	orange	18.5379	73.8048
71	411011	192	Historic Site	yellow	18.5235	73.8639
263	411013	64	Historic Site	yellow	18.5235	73.8639
327	411025	24	Historic Site	cyan	18.3717	73.7596
351	411032	48	ATM	purple	18.5931	73.9218
399	411046	2	Business Service	magenta	18.4029	73.8537
401	411047	4	ATM	lightgreen	18.5992	73.9270
405	412105	96	River	grey	18.6776	73.8987
501	412115	13	River	black	18.5371	73.6772



RESULTS – CONTD.

- The result can be categorized into two groups.
  - A) A restaurant in the outskirts. – Unmarked in the map below.
  - B) A restaurant in the city – Marked inside the Green Circle.
  - C) A restaurant between the outskirts and the city. (Pin Code inside the orange Circle)



DISCUSSION

- Pune is a very big city distributed across 34 pin codes. There are some pin codes with very high density and some with very less.
- Thus, with a flexible budget I showed as per requirement, which area is suitable to deploy a restaurant.
- Data was fetched from Foursquare API and after making modifications, I was able to extract all venues/buildings falling under each pin code within 1000 m radius. Thus a population density can be categorized based on the count of venues.
- Now, there can be other algorithms used to find out these areas which also will have different results. Also, if the budget is flexible, a restaurant can also be deployed in the competitive/saturated areas. I could not capture this in my report.
- I used K-Means clustering which was suitable to decide where to place a restaurant based on a simple score assigned to every pin code/area. A restaurant venue was give a negative value, a neutral venue was given a value of 0 and a venue where the restaurant could be placed was given a positive value. This mapping can be changed as per requirement and thus will produce different results.
- Optimal pin codes were then extracted using this score.
- Also, I have only done this analysis on 34 pin codes. This analysis can be further specified with more data available. It can be divided into Talukas/Areas.
- After getting the result for optimal pin codes falling in each section of groups – City, Outskirts, In Between, I concluded the report.

CONCLUSION

- As mentioned before, just having a job no longer is the normal today. There is always something on the side that everyone cultivates.
- Restaurant Location selection is a vast and lengthy process. As it can be done based on many factors and which factor has more weightage is subjective.
- Any area can be optimal based on budget and type of restaurant to be deployed. The results can be more specified with the availability of more specific data.
- Thank You for taking the time to review this report.

THANK YOU

RESTAURANT LOCATION FINDER

A CAPSTONE PROJECT REPORT

RINAV SAXENA

DATED - 19082020