Finding best zone for recreation

Kaustubh Bawankar

April 16, 2020

1 Introduction

1.1 Background

Traveling is done by everyone now a days, we go to a new city without knowing the best the city can offer .We at times fail to figure out in which location we need to buy the accommodation in order to leverage the best the city can offer.

1.2 Problem

After travel we search the accommodation using booking.com or airbnb but if we are planning to get the best the city can offer like hotes,pubs,bars ,sports etc, nearby we need to chose a location .

1.2 Interest

The solution can be optimized and used to recommend by any travel provider who wishes to give more personalized solution to their customers.

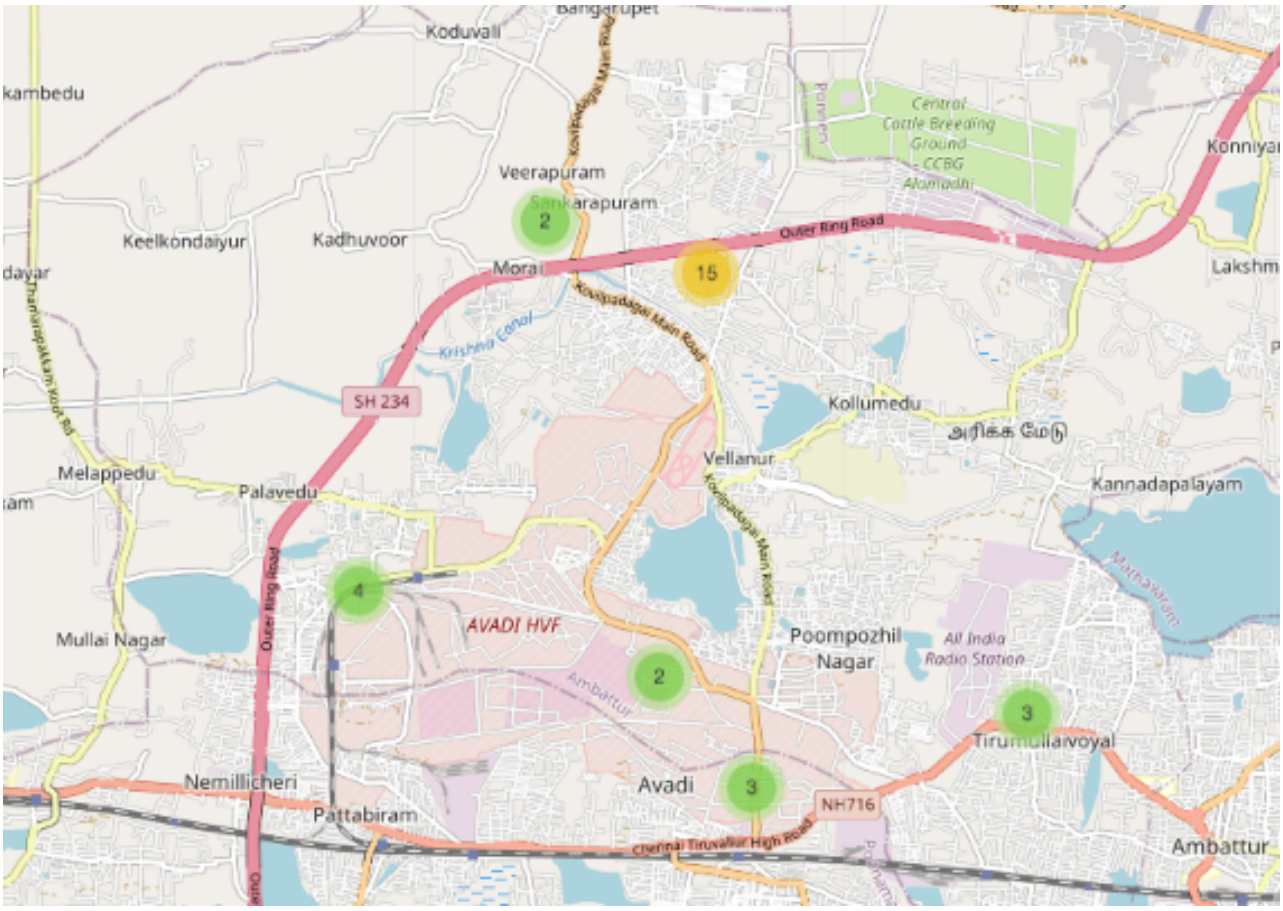
2 Data acquisition and cleaning:

Data has been collected using the foursquare API and solely it's data has been used .
Using the gps coordinates we query the API .However using more payloads we can increase/ decrease the geofence radius.For this query i have added a college Veltech institute.

This is the sample cleaned data from foursquare API :

	id	name	category	lat	lng
0	4ec0a577e5fae16464dd1a90	Vel Tech Engineering College	college engineering building	13.177126	80.098157
1	5098aff4e4b0cba46df4ce11	lords hostel,veltech univesity	university	13.181173	80.101004
2	4ec0ae196c25dfd9820d11b3	Canteen	college cafeteria	13.177580	80.098218
3	4ec0bf5dbe7b04923cda91b3	Gym	college gym	13.175231	80.097885
4	4ec0aed37ee54e4cd3135000	Hostel	college residence hall	13.177162	80.098030

The data when mapped gives the multiple location using follies open source maps:



3 Data analysis:

Now we have the data we need to group by the types and the count and the proximity from the given location and using one label encoding we label the data.

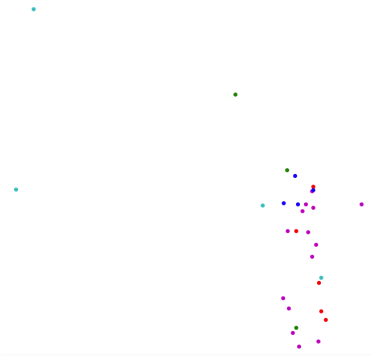
category	
gym	3
concert hall	2
university	1
college library	1
café	1
college administrative building	1
college arts building	1
college basketball court	1
college bookstore	1
college cafeteria	1
college classroom	1
college engineering building	1
college gym	1
college lab	1
college quad	1
school	1
college residence hall	1
college tennis court	1
dog run	1
event space	1
hotel bar	1
indian restaurant	1
juice bar	1
office	1
restaurant	1
bakery	1

	id	name	category	lat	lng
0	4ec0a577e5fae16464dd1a90	Vel Tech Engineering College	8	13.177126	80.098157
1	5098aff4e4b0cba46df4ce11	lords hostel,veltech univesity	25	13.181173	80.101004
2	4ec0ae196c25dfd9820d11b3	Canteen	6	13.177580	80.098218
3	4ec0bf5dbe7b04923cda91b3	Gym	9	13.175231	80.097885
4	4ec0aed37ee54e4cd3135000	Hostel	13	13.177162	80.098030

we then sort the data using the gps coordinates hamming distance

4 Modelling :

Using K means clustering we cluster the data using gps coordinates:



Every color represents a group made my the k means cluster model.
the value of K we have assumed to be 3 for this build but we can enhance it to use the elbow method and count using the unique areas.

```
['bank' 'bed & breakfast' 'café' 'coffee shop' 'department store'
 'fast food restaurant' 'hostel' 'hotel' 'indian restaurant'
 'italian restaurant' 'juice bar' 'movie theater' 'restaurant'
 'snack place' 'spa' 'train station']
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	0	0	0	0	0	0	1	10	2	1	0	0	0	0	0	0
4	0	0	0	2	1	1	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	1	1	3	0	0	0
0	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

here on the y axis is the group id and the x axis is the services count we got the server id from one hot encoding .

We then measure the skewness and get the best cluster:

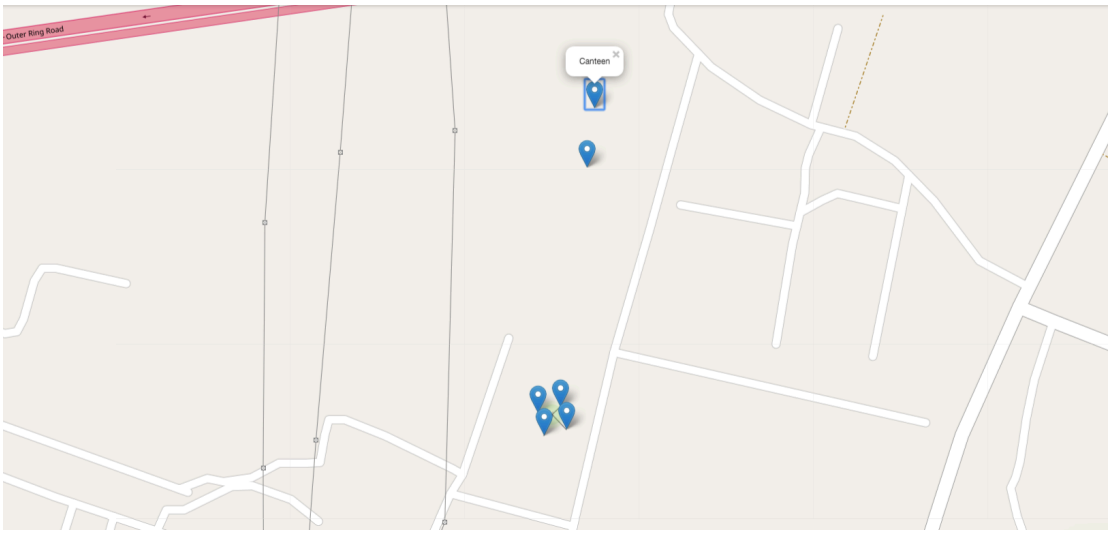
```
3    1.771925
4    2.375384
0    2.375384
1    3.002102
2    3.672686
dtype: float64
```

so as the 3rd cluster as the lest skewness value we show it's details in the follies map.

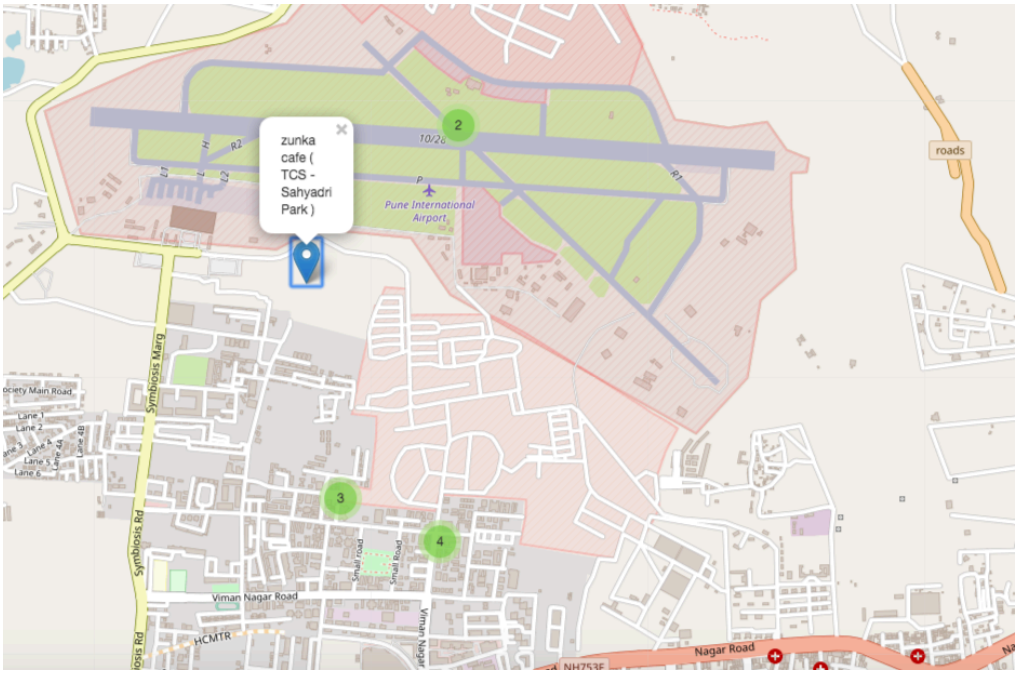
4 Model evaluation :

We have evaluated the model by changing the gps coordinates.

Example for best region near Veltech university is the internal campus of the veltech



Best region near Pune international Airport (Pune) is : <u>Vimannagar</u>



Best region in the suburb in Toronto is near <u>Lake Shore Boulevard</u> :

5 Conclusion

In this build we have using k means algorithm and measured the skewness in each groups to find the optimal region .However the result are purely based on the data on the foursquare.

6 Further Improvement

1) We can automate the value of K using the region groups and the elbow method to get the bigger possible region .
2) Dimension can be increased with multiple tags and rating to get the best region.