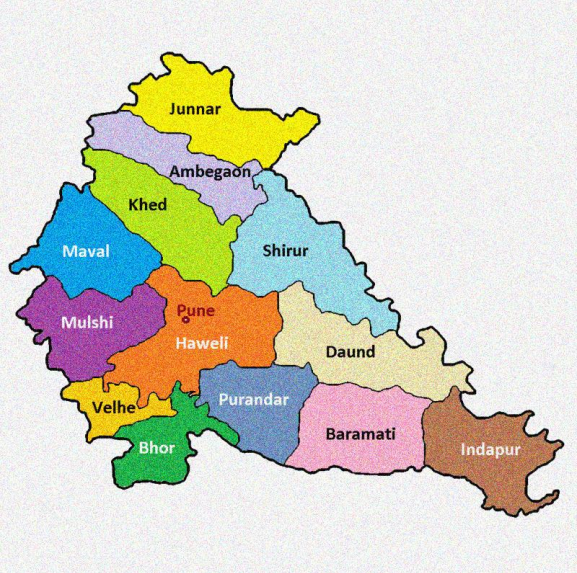


Project Report - Location Analytics

ABSTRACT

Find out neighbourhood with in Pune district which having upscale residency, high number of health care centres and venues that is suitable to open medical shop.

Amresh Nikam

Contain

1. A description of the problem

1.1 Problem Background

Mr. Nilesh live on the Pune City of Maharashtra in India. He is chemist and his medical shop running very well, mainly because of location of shop, all the great amenities and other types of venues that exist in the neighbourhood, such as many hospitals and clinics are near, the area is upscale of residency, the transportation, the area parks, grad schools and so on. Now he wants to open another medical shop on the other side of the city with great profit prospects. However, he afraid and have lot of questions in his mind.

Is shop run well at new place?

Is neighbourhood of new place exactly the same as current neighbourhood?

How can I find place with same neighbourhood?

The location of your business can be an important factor in its success. When choosing a location, carefully assess the right environment for your business.

When starting out, you'll need to decide where you'll conduct business. Unless you're a completely home-based business, this will require buying or leasing a business premises. Each business has different requirements and it's important to consider your business needs and priorities when deciding on the type and location for your business premises.

For example, a location may be ideal for your business because:

- Your suppliers or distributors are nearby
- It's a known centre for the products or services you are providing
- Many of the people who work or live in the area are your business' target audience
- Businesses in the area complement yours (for example, a children's clothing shop could benefit from a childcare centre or toy shop nearby)
- The costs of buying or leasing in the area are affordable and meet the needs of your business
- It's a growing business hub with many opportunities in the near future.

Location analysts can be involved in many areas of business operations. The job may include site research, marketing, market intelligence, property planning and research, retail, facilities, operations, planning and acquisitions. Many private companies use location analysts to decide on new store locations and other business decisions. Employment also exists in the public sector, frequently in local government.

For the above scenario Mr. Nilesh hire a Location analyst for find the location for his business which is same as current neighbourhood.

Location analyst know that, the two things to look for when choosing a pharmacy location. Neither has any relationship to the other, but both must be considered. They are:

1. Adjacent to an upscale residential area
2. Next to a chain store pharmacy

If both of these characteristics are not simultaneously available, then always go with the first: being in or near an upscale neighbourhood. An alternate approach is to be on or close to a highway that connects high-income families to your store.

To find out location based on above characteristic in Pune city is very difficult and time consuming but we have the data of all location with venues then appalling appropriate analysis techniques we can show which are the places that full fill the criteria.

1.2 Problem Statement

Research question is how can efficiently find places in city in minimum time span that have same characteristics.

Can K-means clustering be utilized to localize similar type neighbourhood venues? There is a need of a neighbourhood venues location approach that considers the flexibility, capacity and quantifying the attributes. This approach gives an opportunity of using real numbers from the database website in determining the best location for a new business facility.

2. A description of the data

2.1 Data Acquisition and Description

The data required for the Location analyst to find the similar type of neighbourhood of Pune city are:

- Locations by ZIP code
- Positional coordinates
- Neighbourhoods venues
- Population database.
- Hospitals and clinics location information

Locations by ZIP code:

- This data is required for knowing the different areas in Pune city.

- The location zip code can easily find by googling. The https://www.mapsofindia.com/pincode/india/maharashtra/pune/ web peg having table which contain location name, postal code, state and District.

Positional coordinates

- This data is useful to determine the geographical location when we visualize the places and also require for getting the venues for that places.
- Positional coordinate data of each postal code can be achieve using Geocoder or http://download.geonames.org/export/zip/ web site.

Neighbourhoods venues

- This data is very important to find out what facilities are available in each place.
- To gather this data, the Foursquare API was used. This API provides real-time geographical data of almost anywhere on Earth.

Population database.

- Using this data, we can calculate Population density. So we know the area is upscale resident or not.
- This data we can get from https://www.censusindia2011.com/maharashtra/pune-population.html web site.

Hospitals and clinics location information

- This data we use to find out how many medical facilities are available in each place.
- This data set we get from https://pin-code.org.in/hospitals/listing and www.unipune.ac.in/admin/circular/List%20of%20Hospitals.xls

2.2 Data Cleaning

Checking the data

Most of the curated data sets from the government can have errors in them, and it's vital that we spot these errors before investing too much time in our analysis.

We find following problems in curated data.

1. Geographical location data, we plot marker on the map and saw that few places location was not correct they are in outside Pune district.
2. Population data were collected from different government site. To compile them it was very time consuming.
3. In compile Population data, the figure of population in thousand separate commas, some in fractional notation.

4. The name of place in population data and geographical data are mismatched due to spelling like ‘Maval’ in ne data set and ‘Mawal’ in another.
5. Similarly, in Health care centre position data the number of health care centres in vales, percentages.
6. Also, some values are missing in Health care centre position data

Tidying the data

We've identified several errors in the data set, we need to fix them before we proceed with the analysis. Let's walk through the issues one-by-one.

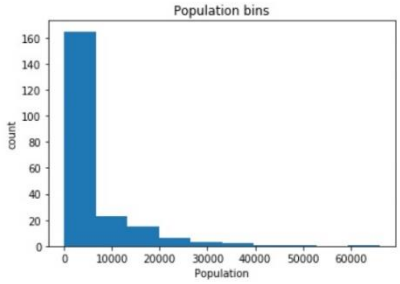
- Plotting markers on map we identify the places whose geographical location was incorrect. Visiting https://www.latlong.net/ web site and manually put up the name of places and corrected them.
- Using Ms-Excel copy the all the values of population and set the cell type of as number and copy back to our file.
- Correcting Name is very difficult. Using SQL query, we find mismatch names of places and manually corrected it.
- The total number of health care centres values in percentages are corrected using simple mathematical formula and converted it into simple number.
- For the missing values of number of health care centres can’t be leave out or dropped. We can’t place mean, mode or medium values. According literature survey, in India there is 7 health care centre per 10,000 people. So, using population value we calculated and filled up missing values.

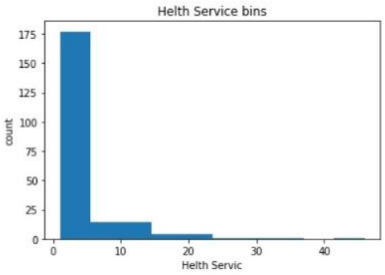
3. Exploratory analysis

3.1 Data Conversion

Now we have corrected data, to proceed to analyse the data. In our data set Population and Health-care are continues numeric value that can not be used directly for clustering because it doesn’t have significance to grouping the cluster. So first we convert them into categorical variable.

To converting them into categorical, we first checkout how they are distributed in overall dataset. The following figure shows the histogram of Population and Health-care attributes.





Distribution of Population Distribution of Health-care

Figure 1: Distribution of Population and Health-care.

From the histogram we decided the Population and Health-care categorize in three categories less, average, good and low, medium, high respectively.

Once converted into categorial again we can't use these attributes directly for clustering because clustering algorithm uses distance criteria to separate out the group so again, we convert this categorial attributes into numeric using creating dummy variables.

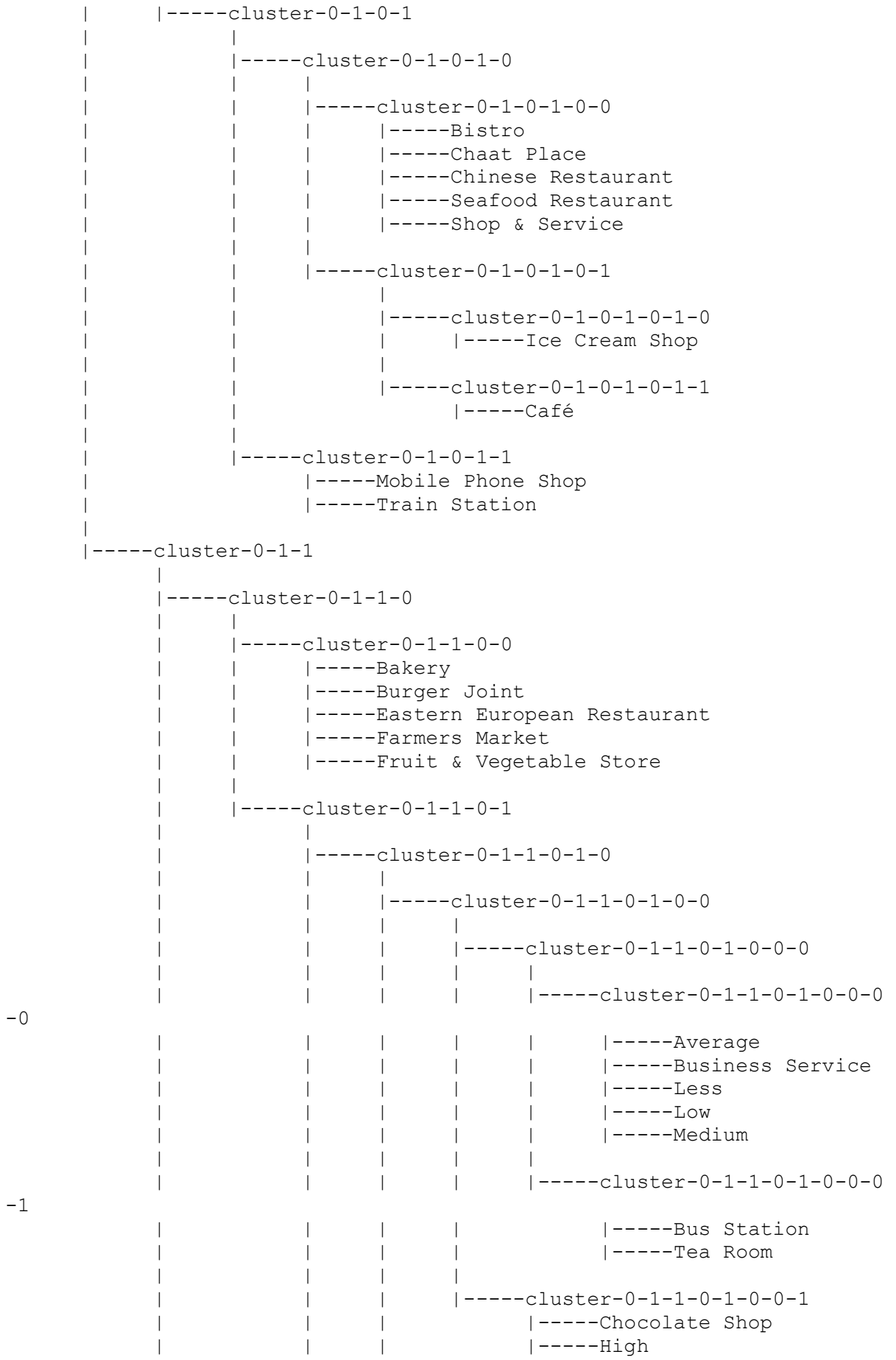
Similarly, the vanes category is also categorical, it also converted into dummy variable.

3.2 Variable clustering

This step is performed to cluster variables capturing similar attributes in data. And choosing only one variable from each variable cluster will not drop the separation drastically compared to considering all variables. Remember, the idea is to take minimum number of variables to justify the separation to make the analysis easier and less time consuming. We decompose the all our features and see how they are falls to separate out the clusters. After decomposition we decide that we kept all features for clustering

```
cluster-0
|
|-----cluster-0-0
|      |
|      |-----cluster-0-0-0
|      |      |
|      |      |-----cluster-0-0-0-0
|      |      |      |-----cluster-0-0-0-0-0
|      |      |      |      |-----Antique Shop
|      |      |      |      |-----Bus Line
|      |      |      |      |-----Fast Food Restaurant
|      |      |      |      |-----Hotel Pool
|      |      |      |      |-----Indian Chinese Restaurant
|      |      |      |      |-----Multicuisine Indian Restaurant
|      |      |      |      |-----Rest Area
|      |      |      |      |-----Trail
|      |      |      |-----cluster-0-0-0-0-1
```

```
|     |     |           |-----Hotel
|     |     |           |-----Lake
|     |     |-----cluster-0-0-0-1
|     |           |-----cluster-0-0-0-1-0
|     |           |     |-----cluster-0-0-0-1-0-0
|     |           |     |     |-----cluster-0-0-0-1-0-0-0
|     |           |     |     |     |-----cluster-0-0-0-1-0-0-0-0
|     |           |     |     |     |     |-----Dessert Shop
|     |           |     |     |     |     |-----Pizza Place
|     |           |     |     |     |
|     |           |     |     |     |-----cluster-0-0-0-1-0-0-0-1
|     |           |     |     |           |-----ATM
|     |           |     |     |           |-----BBQ Joint
|     |           |     |     |
|     |           |     |     |-----cluster-0-0-0-1-0-0-1
|     |           |     |           |-----Coffee Shop
|     |           |     |           |-----Electronics Store
|     |           |     |-----cluster-0-0-0-1-0-1
|     |           |           |-----Garden Center
|     |           |           |-----Music Venue
|     |           |-----cluster-0-0-0-1-1
|     |                 |-----Platform
|     |                 |-----Soccer Field
|     |-----cluster-0-0-1
|           |-----cluster-0-0-1-0
|           |     |-----Farm
|           |     |-----Motel
|           |     |-----Toll Plaza
|           |-----cluster-0-0-1-1
|                 |-----cluster-0-0-1-1-0
|                 |     |-----Forest
|                 |     |-----Restaurant
|                 |     |-----Scenic Lookout
|                 |
|                 |-----cluster-0-0-1-1-1
|                       |-----cluster-0-0-1-1-1-0
|                       |     |-----Breakfast Spot
|                       |     |-----Tennis Court
|                       |-----cluster-0-0-1-1-1-1
|                             |-----Indian Restaurant
|-----cluster-0-1
      |
      |-----cluster-0-1-0
      |     |
      |     |-----cluster-0-1-0-0
      |     |     |
      |     |     |-----cluster-0-1-0-0-0
      |     |     |     |-----Bagel Shop
      |     |     |     |-----Bank
      |     |     |     |-----Deli / Bodega
      |     |     |     |-----Garden
      |     |     |     |-----Gym
      |     |     |     |-----Gym / Fitness Center
      |     |     |     |-----Snack Place
      |     |     |
      |     |     |-----cluster-0-1-0-0-1
      |     |           |-----Asian Restaurant
      |     |
```


|-----cluster-0-1-0-1
|-----cluster-0-1-0-1-0
|-----cluster-0-1-0-1-0-0
|-----Bistro
|-----Chaat Place
|-----Chinese Restaurant
|-----Seafood Restaurant
|-----Shop & Service
|-----cluster-0-1-0-1-0-1
|-----cluster-0-1-0-1-0-1-0
|-----Ice Cream Shop
|-----cluster-0-1-0-1-0-1-1
|-----Café
|-----cluster-0-1-0-1-1
|-----Mobile Phone Shop
|-----Train Station
|-----cluster-0-1-1
|-----cluster-0-1-1-0
|-----cluster-0-1-1-0-0
|-----Bakery
|-----Burger Joint
|-----Eastern European Restaurant
|-----Farmers Market
|-----Fruit & Vegetable Store
|-----cluster-0-1-1-0-1
|-----cluster-0-1-1-0-1-0
|-----cluster-0-1-1-0-1-0-0
|-----cluster-0-1-1-0-1-0-0-0
|-----cluster-0-1-1-0-1-0-0-0
-0
|-----Average
|-----Business Service
|-----Less
|-----Low
|-----Medium
|-----cluster-0-1-1-0-1-0-0-0
-1
|-----Bus Station
|-----Tea Room
|-----cluster-0-1-1-0-1-0-0-1
|-----Chocolate Shop
|-----High

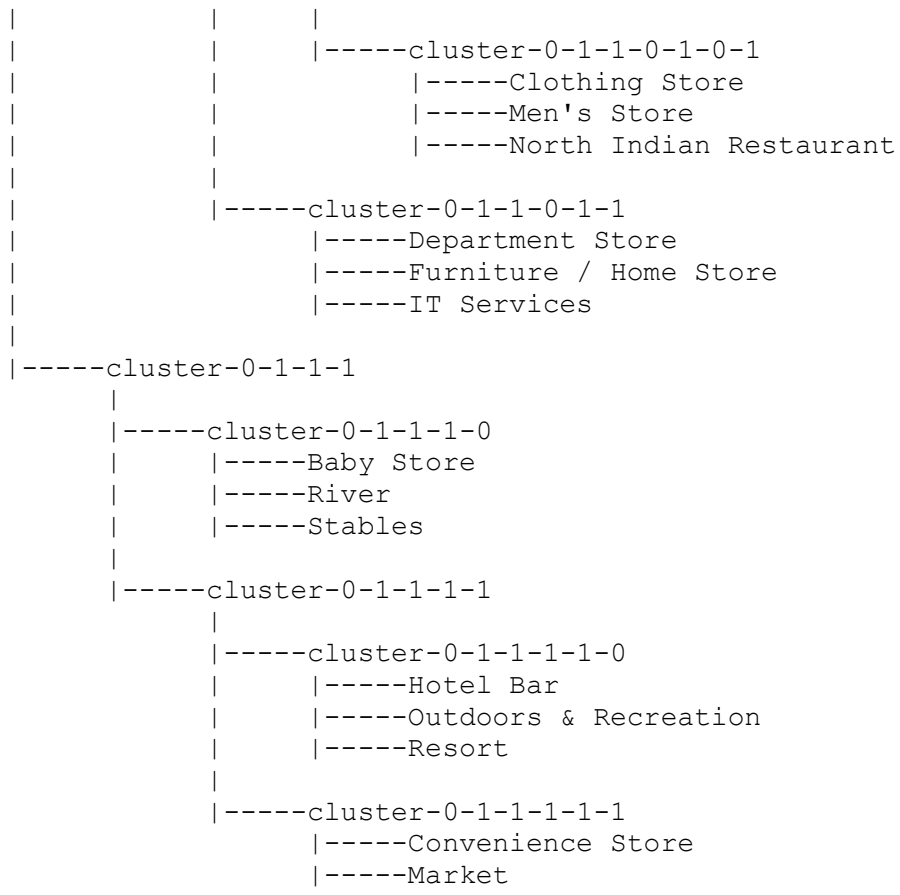


Figure 2: Decomposition of 76 features.

3. Clustering

Gathering all data then cleaning and explorating ready to use data we can use for clustering. The idea was to cluster and determine similar neighbourhoods across Pune city based on the types of venues that they offered, residence density and hospitals, clinics facility.

Now the among K-Means Clustering, Mean-Shift Clustering, Density-Based Spatial Clustering, Agglomerative Hierarchical Clustering which clustering algorithm we choose?

To select the clustering algorithm, we find out the answer of questions listed in below table.

K – means algorithm has all the characteristics that mention in table 1. So, we go with it. K-means is a simple unsupervised machine learning algorithm that groups a dataset into a user-specified number (*k*) of clusters. The algorithm is somewhat naive--it clusters the data into *k* clusters, even if *k* is not the right number of clusters to use.

Questions	Our Answers
Number of variables	76
Input variables type	Quantitative continuous
Should the number of classes be chosen prior to computations?	Mandatory
Results: Class membership	Deterministic
Results: Special features	Profile plot

Table 1 Selection criteria for clustering algorithm

Therefore, when using k-means clustering, users need some way to determine whether they are using the right number of clusters.

One method to validate the number of clusters is the *elbow method*. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of *k* (say, *k* from 1 to 10 in the examples above), and for each value of *k* calculate the sum of squared errors (SSE). Then, plot a line chart of the SSE for each value of *k*. If the line chart looks like an arm, then the "elbow" on the arm is the value of *k* that is the best. The following figure shows the elbow curve for our data set and we selected number of clusters five.

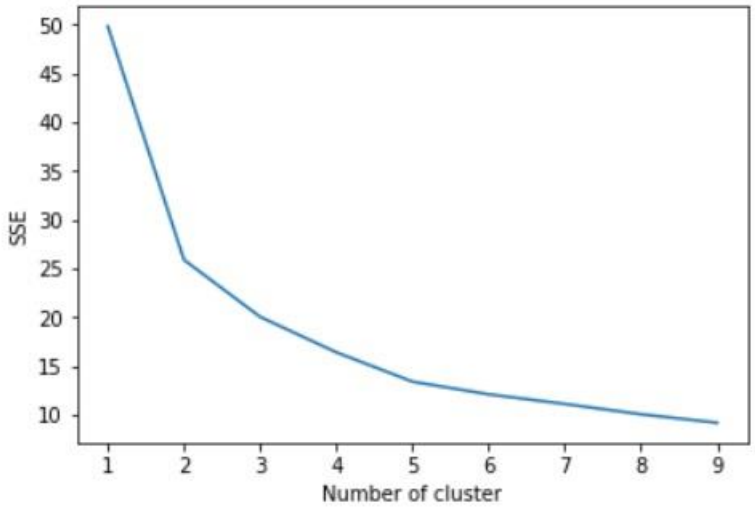


Figure 3: Decomposition of 76 features.

No after applying the k – means we get the five cluster and then we display on map the places have similar characteristics. Following figure shows before and after clustering the places.

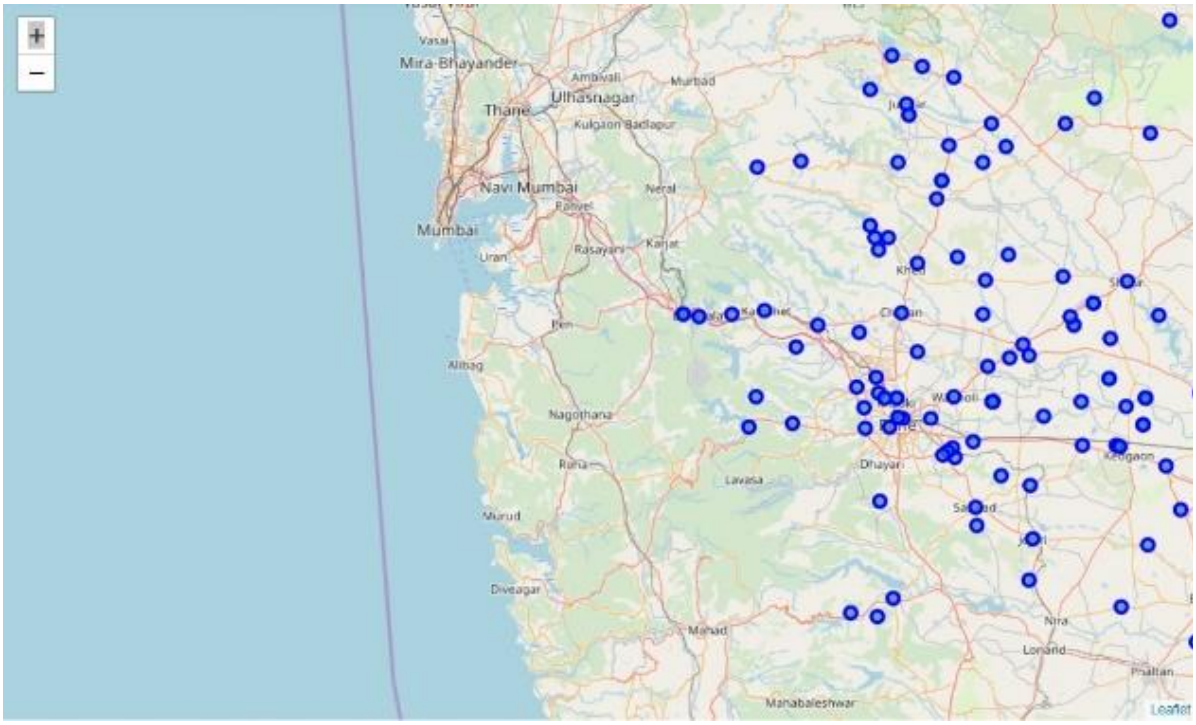


Figure 4: Before clustering all places.

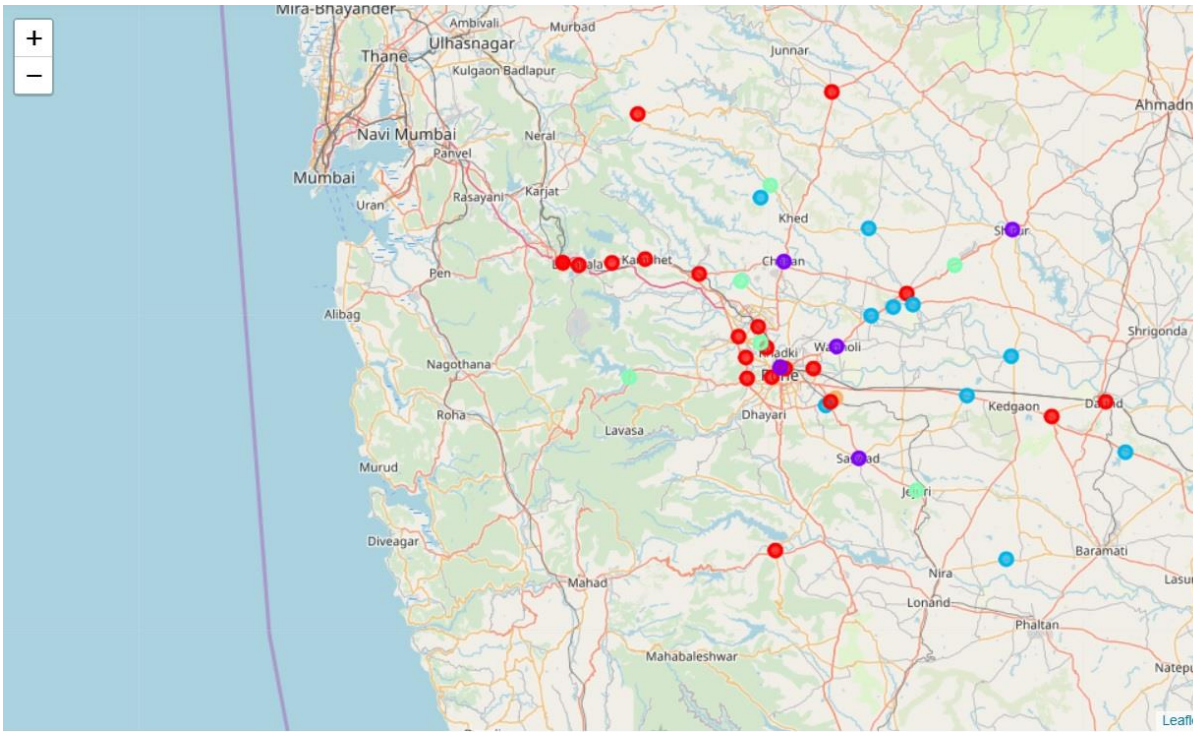


Figure 5: After clustering all places.

5. Result and conclusion

We have got the suitable number of cluster (5) for our clustering k – means algorithm, we used bootstrap method to evaluate the stability of the clustering result. To be specific why we're did this: Often times, clustering algorithms will produce several clusters that represents actual grouping of the data, and then one or two clusters that represents "others". Meaning that they're made up of data points that have no relationship with each other they just don't fit anywhere else. So here we used bootstrap to detect which cluster is considered to be that "other" cluster.
Steps listed in the following:

1. Cluster the original data.
2. Draw a new dataset of the same size as the original by resampling the original dataset with replacement, therefore some data point may show up more than once, while others not at all. Cluster this new data.
3. For every cluster in the original cluster, find the most similar cluster in the new clustering, which is the one with the maximum Jaccard similarity (given two vectors, the Jaccard similarity is the intersect / union, please look it up if it's still unclear).
4. Repeat step 2 and 3 for a user-specified number of bootstrap iterations.

The following table shows the bootstrap result.

	Cluster 1	Cluster 3	Cluster 3	Cluster 4	Cluster 5
The vector of cluster stabilities	0.8834507	0.9643333	0.9555108	0.6520000	0.8190194
The count of how many times each cluster was dissolved	3	1	6	35	23

From the values of bootdissolved (denotes the number of times each cluster "dissolved") and the bootmean value, we can infer that having a low bootmean and high bootdissolved value, cluster 4 has the characteristics of what we've been calling the "other" cluster. Therefore, it is quite likely that it is not an actual cluster, it simply doesn't belong to anywhere else.

Reference

Hyperlinks

- Math formula of the two measures used to determine the suitable k
- Practical Data Science with R Chapter 8 Unsupervised Method
- GitHub - jingmin1987/variable-clustering: A re-creation of SAS varclus
- Getting Started - Foursquare Developer