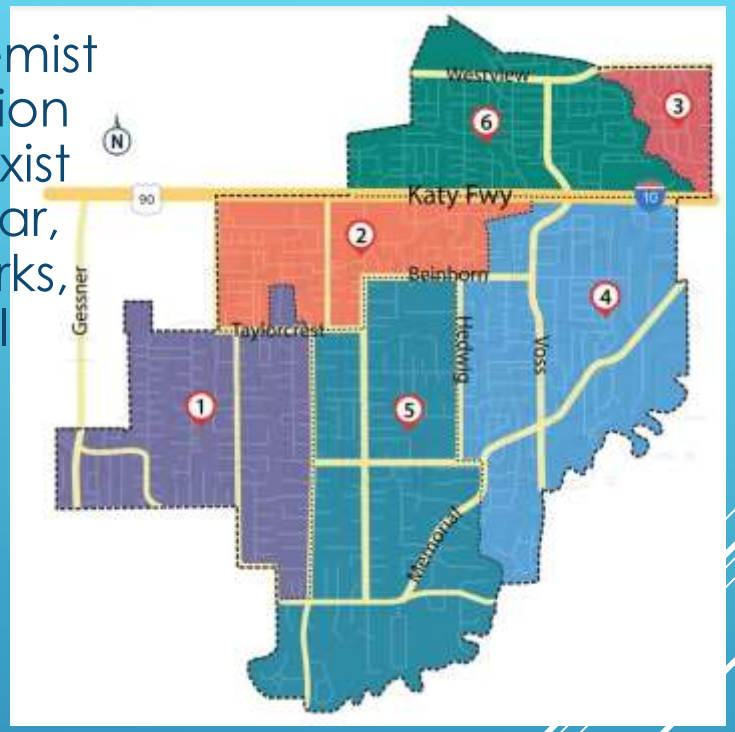
LOCATION ANAYLITICS

Amresh Nikam

A DESCRIPTION OF THE PROBLEM

- 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?



PROBLEM STATEMENT

- Find out neighbourhood with in Pune district which having upscale residency, high number of health care centres and venues categories.

DATA ACQUISITION

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

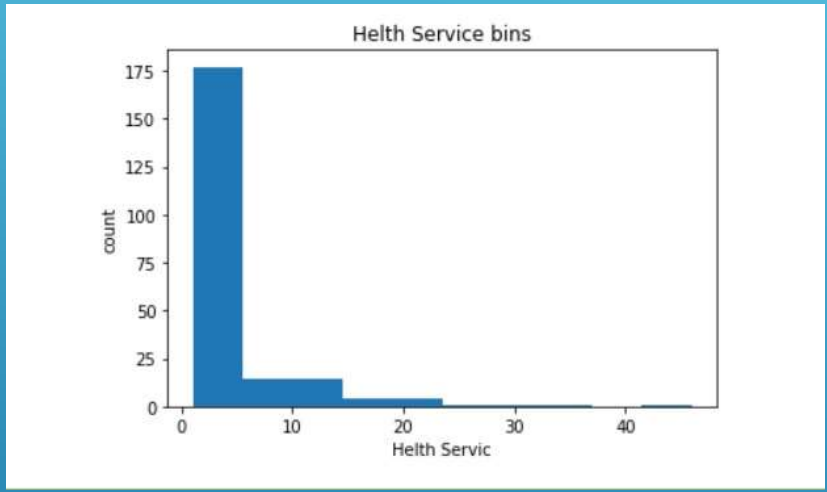
DATA CLEANING

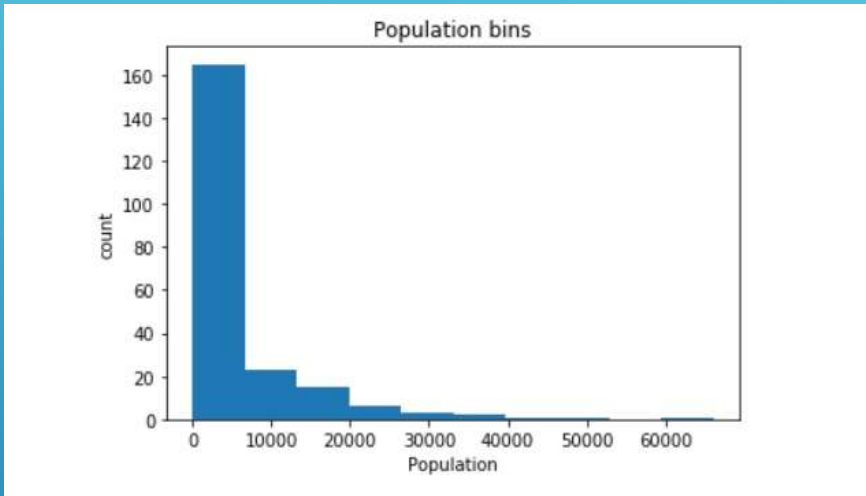
- **Checking the data**
  - Geographical location data, we plot marker on the map and saw that few places location was not correct they are in outside Pune district.
  - Population data were collected from different government site. To compile them it was very time consuming.
  - In compile Population data, the figure of population in thousand separate commas, some in fractional notation.
  - 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.
  - Similarly, in Health care centre position data the number of health care centres in vales, percentages.
  - Also, some values are missing in Health care centre position data

DATA CLEANING (CONT…)

- **Tidying the data**
  - 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.

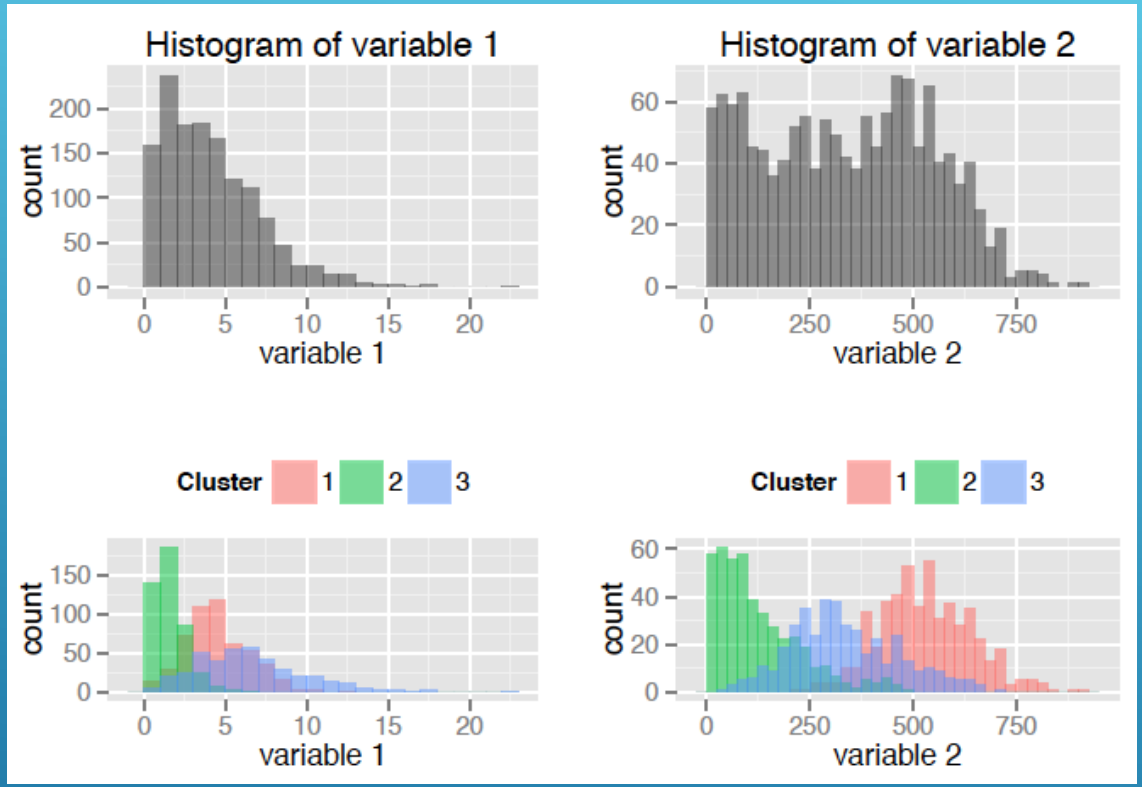
- **Data Conversion**

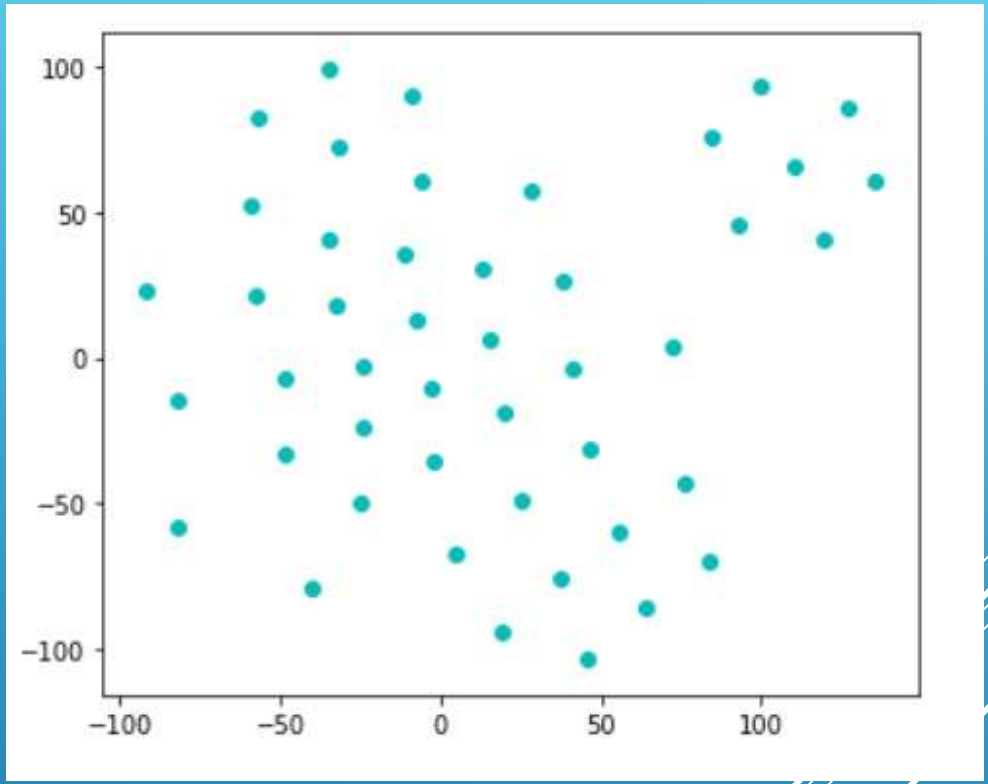




EXPLORATORY ANALYSIS

- Variable clustering

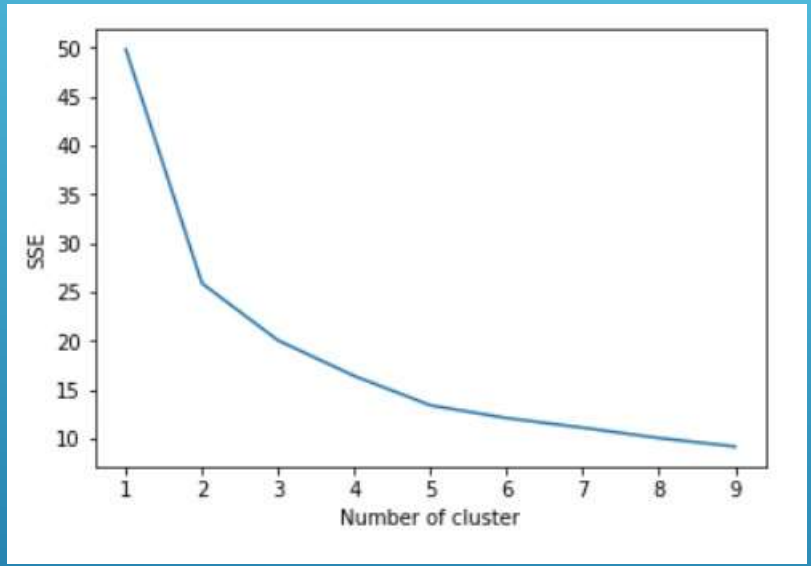


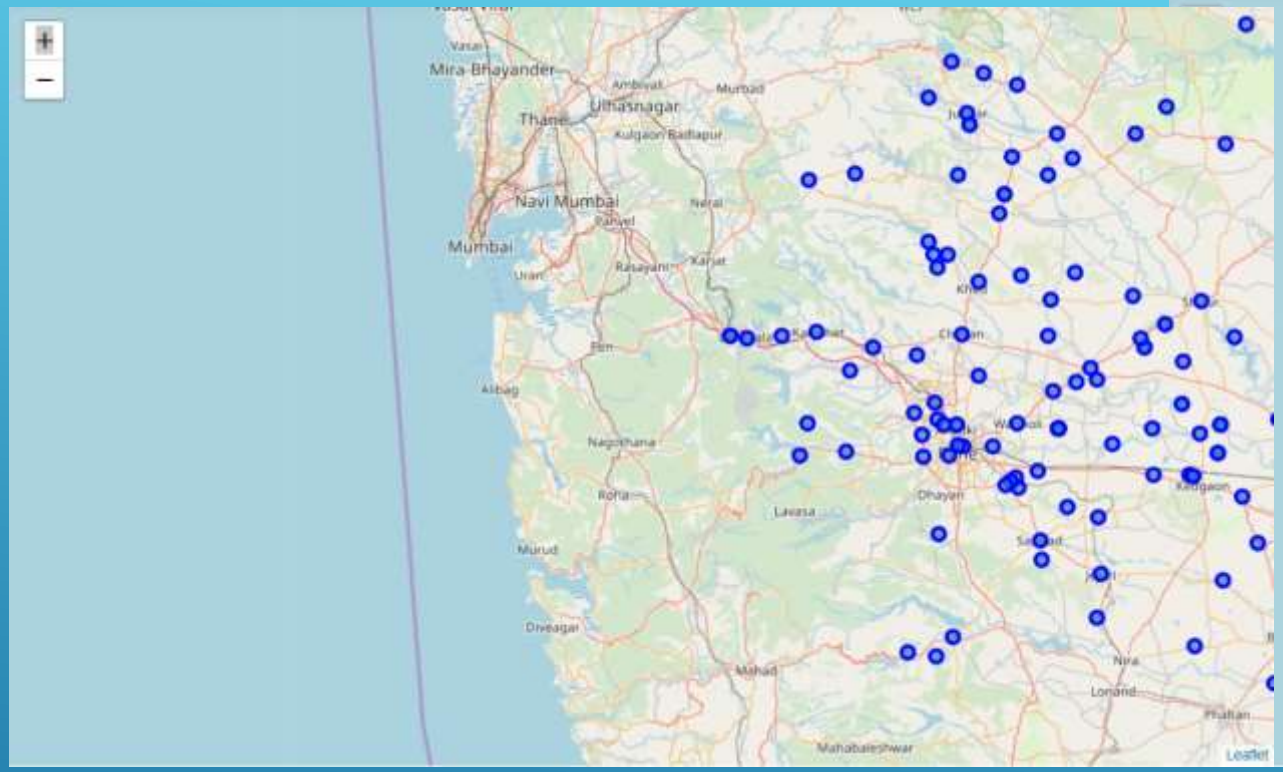


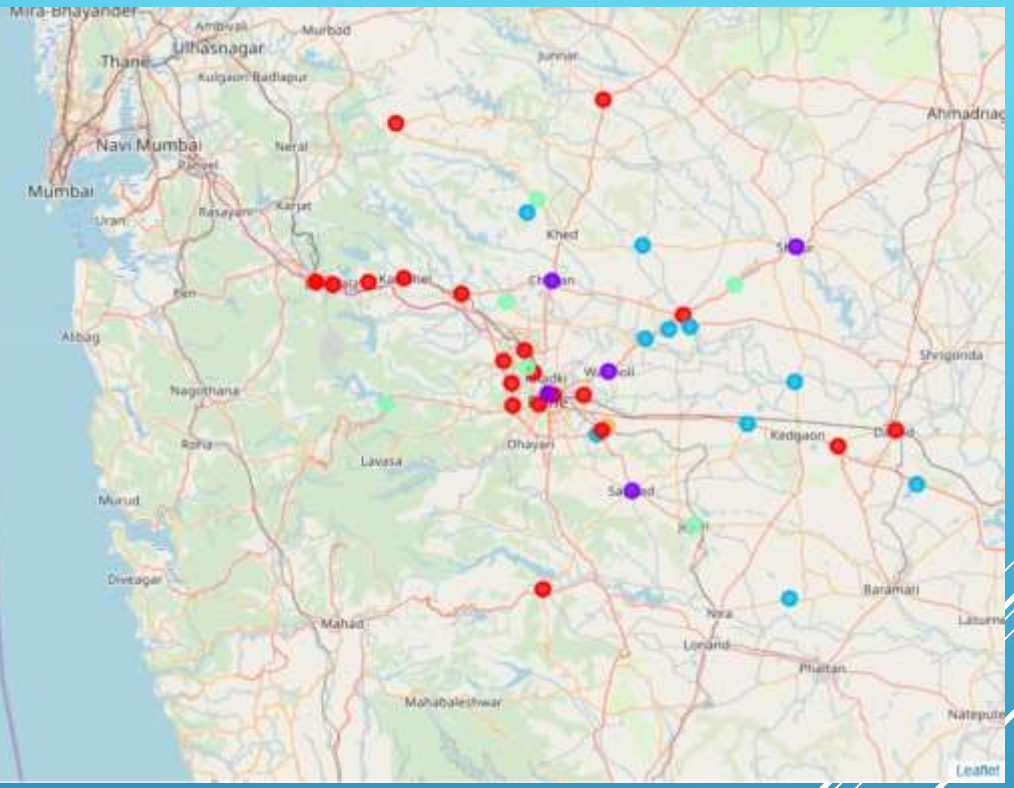
EXPLORATORY ANALYSIS

CLUSTERING

- Getting value of k







CLUSTERING (CONT…)

	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.

RESULT AND CONCLUSION

REFERENCE

- 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