WHEN PEOPLE WANT TO READ A GOOD BOOK..

Introduction

Being a resident of Pune city, Maharashtra, India, it is very easy to see the huge masses of students entering the city every year. Along with increase in educational institutions, jobs are increasing. The city is expanding in every possible way. For this city, also known as ‘Oxford of the East’, one should think of the very need of books it has. Even if we live in digital era, there are people who want to shop for books and enjoy their free time. This population will be center of the study.

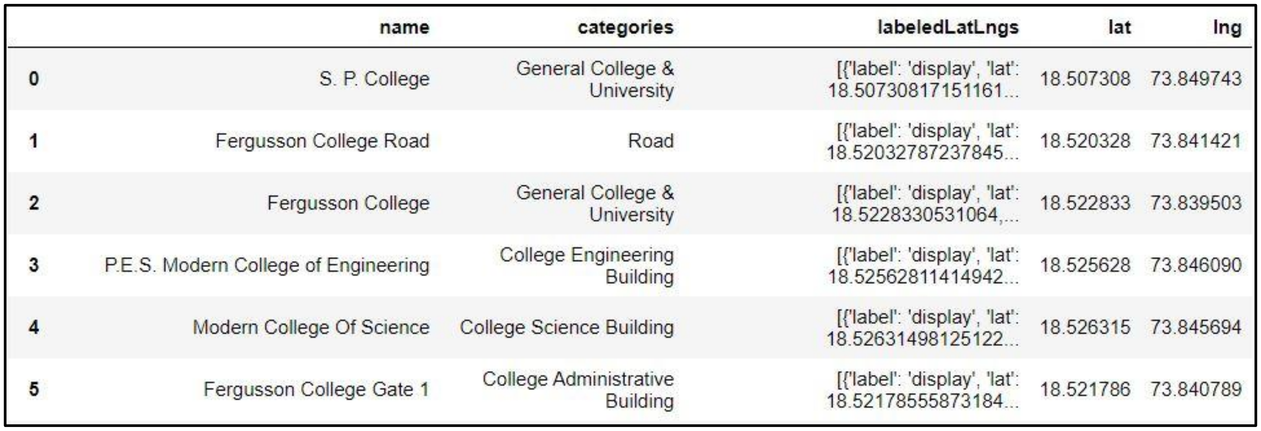
On receiving a complaint from a student, that the number of general bookshops is very low in some areas, I have decided to investigate the situation and find out the areas where setting up a general bookstore or a library will prove helpful to both readers and the owners. The focus will be on the bookshops and libraries which keep books other than only educational ones. The idea is to consider the locations of educational institutions, registered bookstores and libraries and use clustering techniques and local facts for investigation.

Usually, College libraries are not accessible for public, thus, they will be excluded from the list. Considering only bookstores and libraries would serve the purpose, but the abundance of educational institutions along with them might highlight the extent of necessity of bookstores and libraries. This will be a good opportunity to explore the neighborhood of Pune in the perspective of this business problem which will benefit the student community and citizens of Pune in general.

Data

The data of geographical coordinates is needed to carry out the analysis. Queries will be submitted to Foursquare to retrieve locations of Colleges, Institutes, Bookstores and Libraries. The result of queries will be attached together and scanned to check if any duplicate names are present. Folium maps will be useful to visualize locations on map of Pune city. Using latitudes and longitudes and K Means Clustering technique, clusters will be formed and the optimum number of clusters will be the one which has minimum Sum of Squares of Errors, which gives the sum of the squared differences between each observation and its cluster's mean. It will ensure that the clusters are close knit, which in turn, will reduce the geographical distances within the cluster members and their centers.

The data will be filtered to get the following columns in use :



	name	categories	labeledLatLngs	lat	lng
0	S. P. College	General College & University	[{'label': 'display', 'lat': 18.50730817151161...	18.507308	73.849743
1	Fergusson College Road	Road	[{'label': 'display', 'lat': 18.52032787237845...	18.520328	73.841421
2	Fergusson College	General College & University	[{'label': 'display', 'lat': 18.5228330531064,...	18.522833	73.839503
3	P.E.S. Modern College of Engineering	College Engineering Building	[{'label': 'display', 'lat': 18.52562811414942...	18.525628	73.846090
4	Modern College Of Science	College Science Building	[{'label': 'display', 'lat': 18.52631498125122...	18.526315	73.845694
5	Fergusson College Gate 1	College Administrative Building	[{'label': 'display', 'lat': 18.52178555873184...	18.521786	73.840789

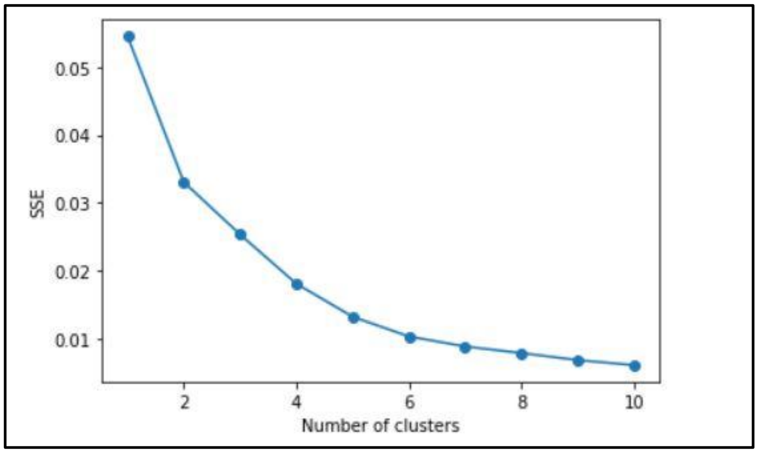
Methodology

Since the focus of the data is longitude and latitudes of the places, exploratory analysis will not be much of a use. Instead, local facts will play important role in the interpretation.

As stated earlier, python libraries Folium and SKLearn will be used for representation and K Means clustering respectively. Pandas and Matplotlib libraries will also play essential role in data processing and plots presentation.

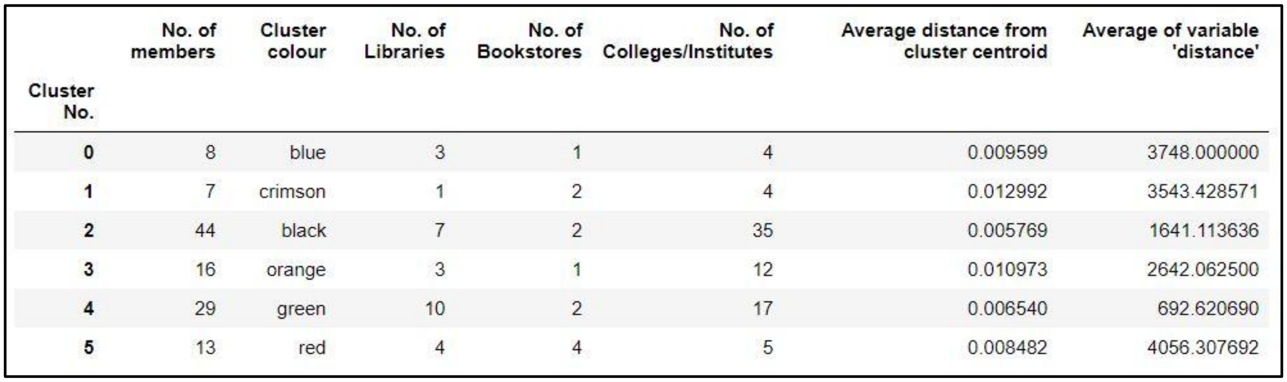
Results and Discussion

The plot obtained for SSE against cluster numbers for K Means clustering was as follows,



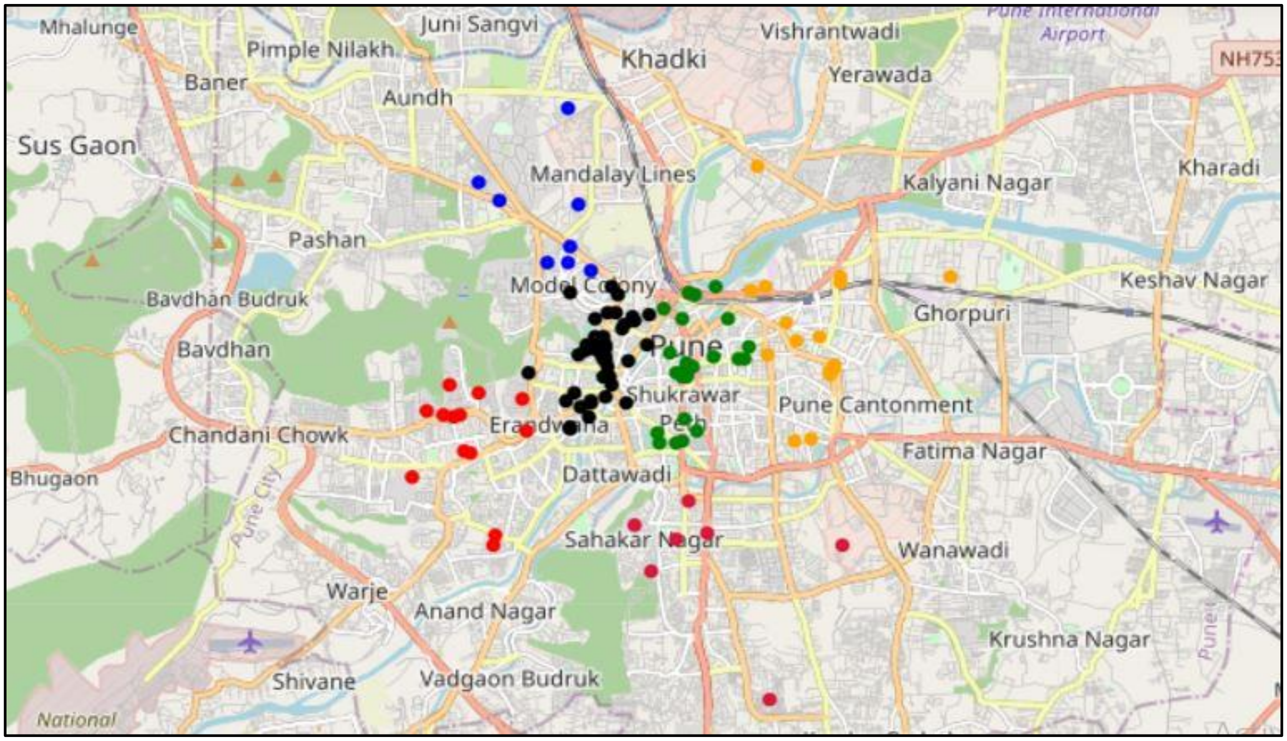
The cluster number with minimum Sum of Squares of Errors (SSE) should be chosen, but it is also important to keep the number as small as possible, otherwise, overfitting will be expected. Thus, using Elbow point from the plot, optimum cluster number was decided to be 6, so that, SSE is less and further increasing the cluster number will not add much of information to the analysis.

The cluster portfolio was as follows,



Cluster No.	No. of members	Cluster colour	No. of Libraries	No. of Bookstores	No. of Colleges/Institutes	Average distance from cluster centroid	Average of variable 'distance'
0	8	blue	3	1	4	0.009599	3748.000000
1	7	crimson	1	2	4	0.012992	3543.428571
2	44	black	7	2	35	0.005769	1641.113636
3	16	orange	3	1	12	0.010973	2642.062500
4	29	green	10	2	17	0.006540	692.620690
5	13	red	4	4	5	0.008482	4056.307692

. The Folium map representing the cluster points was as follows,



Conclusion

Black cluster is the largest cluster with 44 members. It has minimum average distance from cluster centroid.

The second largest cluster is with 29 members, it is close to Black cluster and has second most minimum average distance from cluster centroid.

In these clusters, number of registered bookstores is very low when compared to colleges or institutes. But, the area under Green cluster has well known area called "Appa Balwant Chawk", which has numerous bookstores and shops which provide almost everything related to education.

This place is easily accessible for both the above clusters.

The next large cluster is the one with orange; it has the least number of libraries and registered bookstores. This cluster gives an opportunity for setup of any of them. Another good reason would be that, the cluster has second largest highest distance from cluster centroids, it means, the places are far away from the centroid, thus, setting up new store or library might reduce it for good.

The cluster with red colour has the largest number of registered bookstores and libraries when compared to number of colleges and institutes. But, another fact must be considered that, almost all of them are situated near Maharashtra Institute of Technology, Paud Road. Thus, area near Cummins College of Engineering for Women, Karve Nagar reveals the opportunity for setting up new bookstore or library.

Since area under cluster with blue colour includes campus of Pune University, we must consider it as a special case. It must have a large number of bookstores and libraries even if the University library is considered to be one of the largest libraries in the city. Symbiosis University is also nearby.

Finally, the cluster with crimson colour must be given attention due to its highest average distance from the centroid. Reducing it might help the students and citizens nearby.