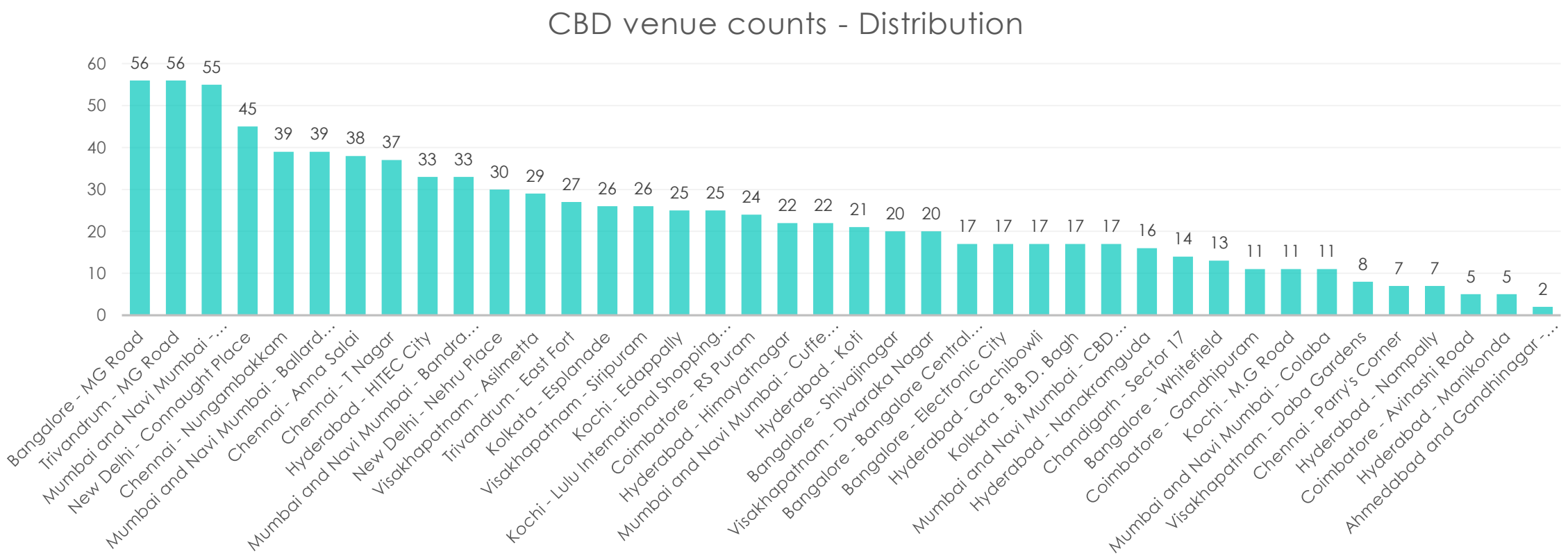
IBM data science – final project

Problem statement

- A Multinational Company X wants to setup its first branch in India. The choice of the place where the company is going to setup will depend on the venues in that area. In order to identify the ideal place, the company is interested to do analysis on the Central Business Districts of India.
- **What are Central Business districts?**
- **Assumptions and Objectives**
  - Identify the CBDs in India
  - Collect the venue data of the CBDs.
  - Perform a clustering analysis among the CBD to identify the similar groups
  - Profile the clusters based on the analysis

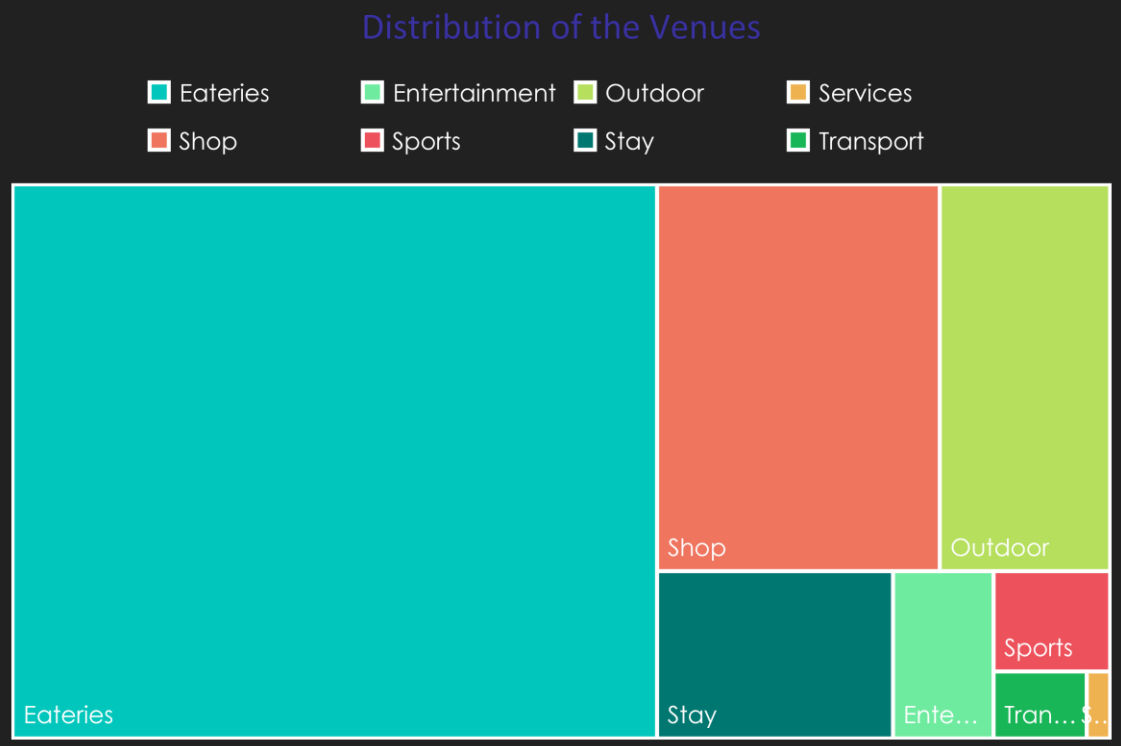
Data Sources and Description:

- The Wikipedia site that lists all the CBDs https://en.wikipedia.org/wiki/List_of_central_business_districts. We can scrape the website and get the information we need.
- There are going to be 2 levels of data. One is a
  - City level data, and another is crime rate, climate, temperature etc.
  - CBD level data which is the venues in that area etc.
- City level data will not be used in the clustering exercise so that the final cluster won't have inherent bias based on the city information. If we use this information, then CBDs in the same city is more likely to be clubbed together.



Distribution of the Venues in each CBD

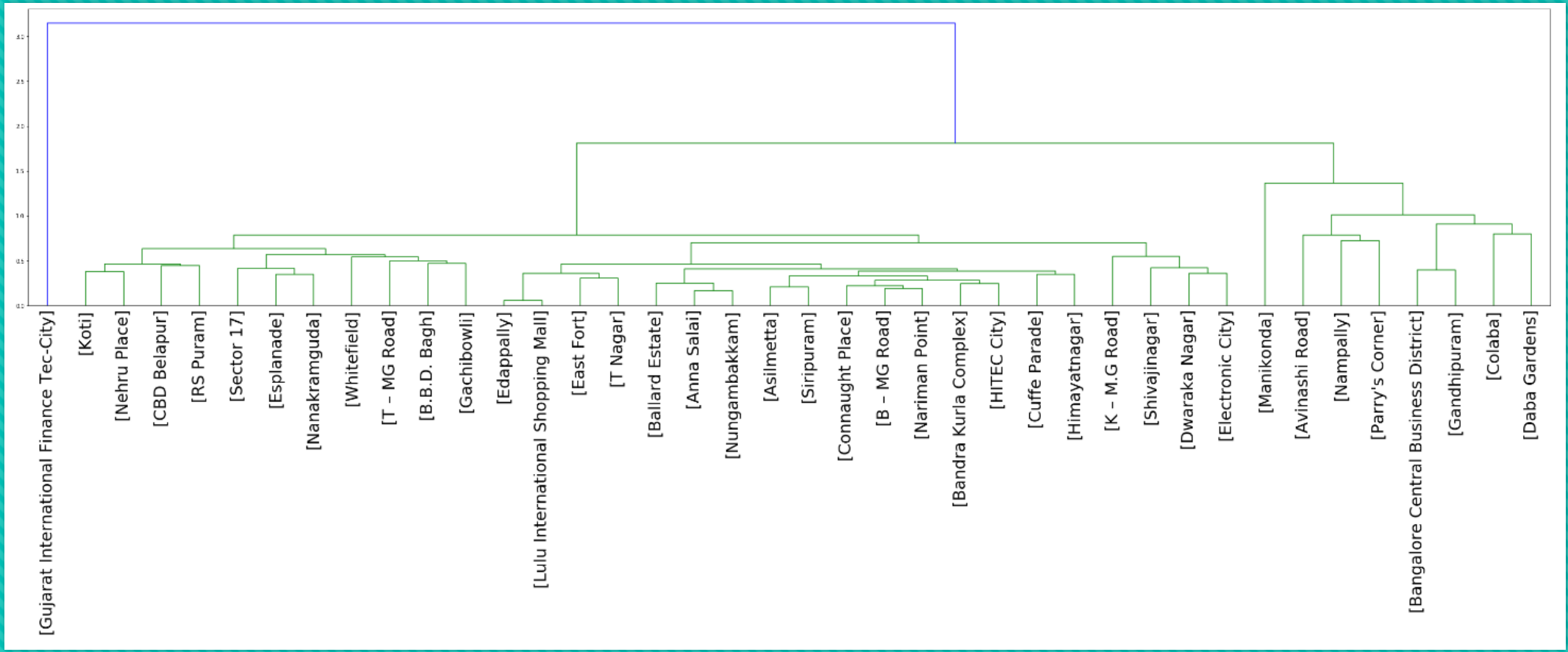
Exploratory analysis



Eateries	Restaurants, Snack bar etc
Entertainment	Move theater, multiplex etc
Outdoor	Beach, Park etc
Services	General Services
Shop	All Stores and shops
Sports	Stadiums and courts
Stay	Hotel, hostel and BnBs
Transport	Bus, train and other transport terminals

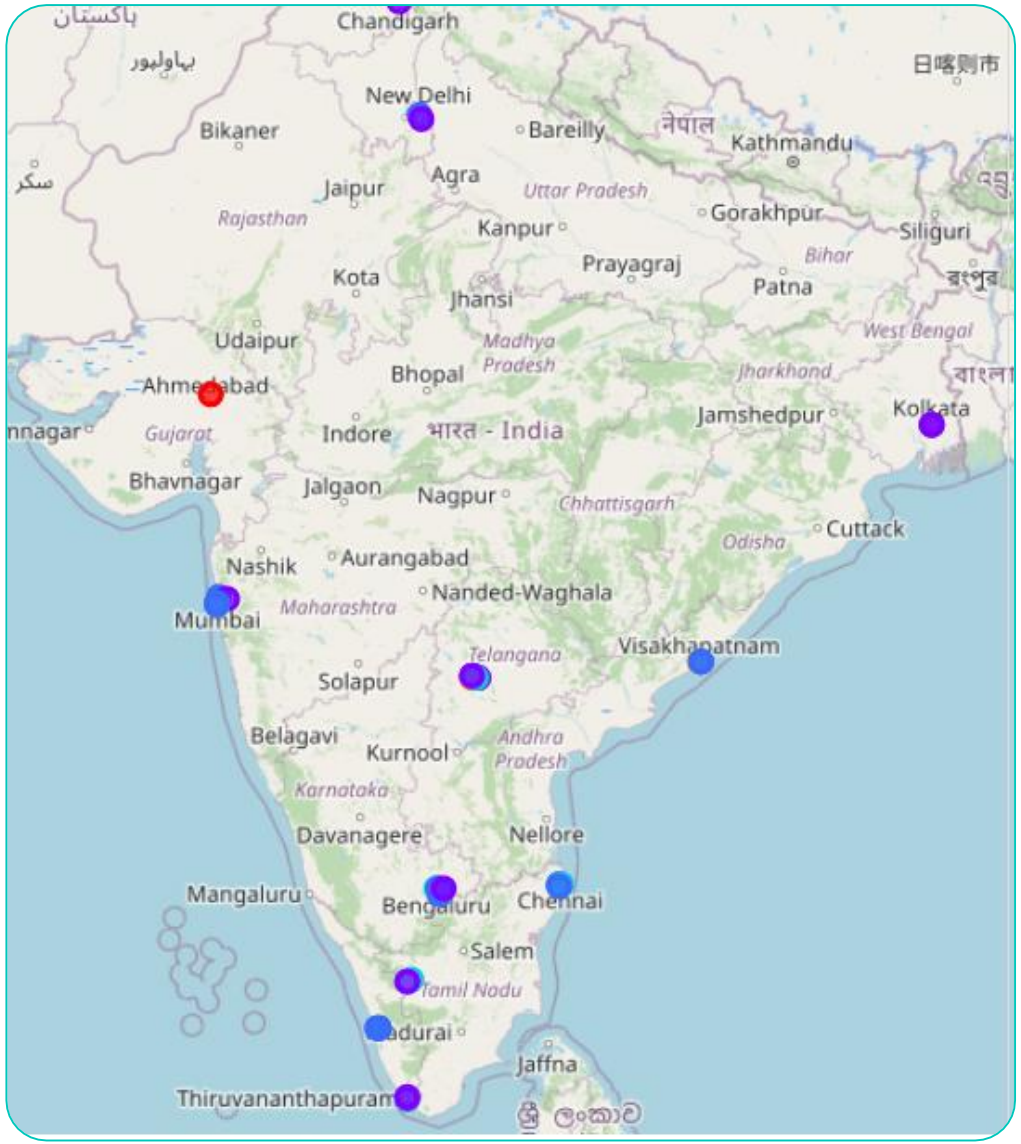
Clustering Algorithm

- Why we chose Agglomerative hierarchical clustering?
- There are two advantages of this algorithm.
  - Euclidean distance is a good measure of the similarity in this case.
  - Client can choose any place that is closest to the ideal branch from the dendrogram.



Final Clusters

Final clusters in the Map of India



Index values

- The interpretation of the clusters are pretty straight forward. We have to see the Venue categories composition with respect to the overall composition. It is called as index numbers. The formula is

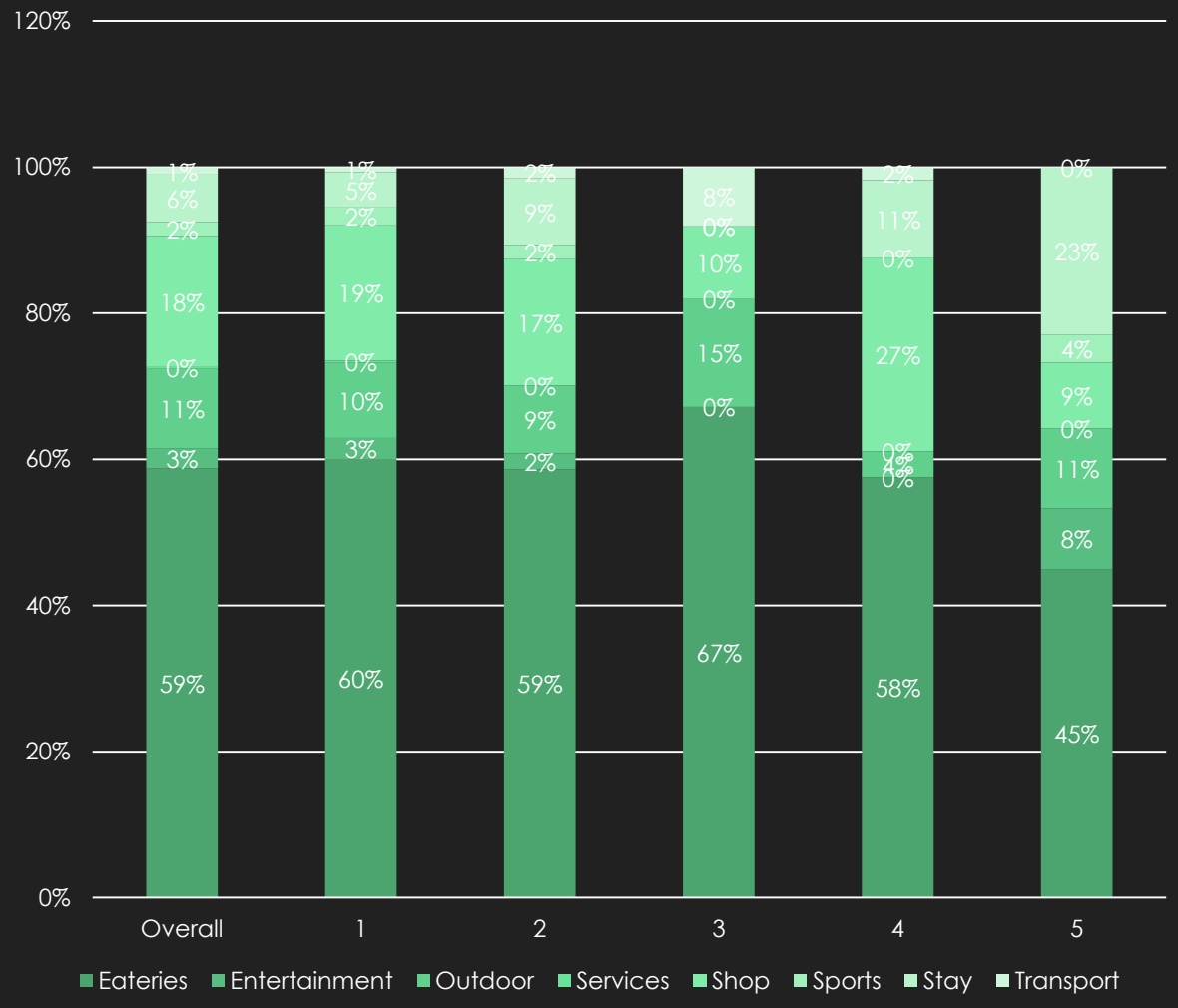
$$Index\ value = \frac{\%\ of\ Venue\ category\ in\ Cluster}{\%\ of\ Venue\ category\ overall} \times 100$$

- For example, for Eateries in cluster 3, the index value is

$$Index\ value_{Eateries\ in\ cluster\ 3} = \frac{67\%}{69\%} \times 100 = 113$$

- Then we can consider a threshold (which is subjective), in this case I have considered 95 and 105 as thresholds. If an index value is less than 95 then it is skewing low, which means the group has less of it and if it is more than 105; it is skewing high and cluster has more of it.

The Distribution of the Groups



Discussion

- All the negative skews, are the opportunity cells. It goes without saying if the company's business involves one of the categories and needs less competition for example, Eateries in Cluster 5 area is an opportunity area. In the above analysis, As we don't know the nature of the company we cannot suggest any cluster as our final solution. In this we will list out the advantages in each cluster and the client and pick the one that fits their profile.
- **Cluster 1:** The CBDs under this cluster are the places with lot of venues. If the company is dependent on all the above groups. Cluster 1 is the ideal place to loot for.
- **Cluster 2:** With high skews in Stay and Transport. If the company wanted a place with high connectivity, this is the cluster to look.
- **Cluster 3:** This is ideal place for companies that requires Eateries, Transport and Outdoor like educational or creative institutions.
- **Cluster 4:** Companies that involves Travels and Tourism. These places are rich in Shops, Stay, Transport. So, it can aid their cause.
- **Cluster 5:** This is probably opportunity CBDs where the company can drive growth.

Conclusion

- The final capstone project was fun.
- As for as the problem statement and results are concerned, it is decent but there are so many places for improvement such as—
  - We can include crime rates in the area which can be used to make the decision on the safety of the employees
  - We can include number of universities in the area, which can be used from growth and recruitment perspective
  - We can do a principle component analysis on the features to get the latent features.
- With the dataset and timeline we had the results as is pretty decent.