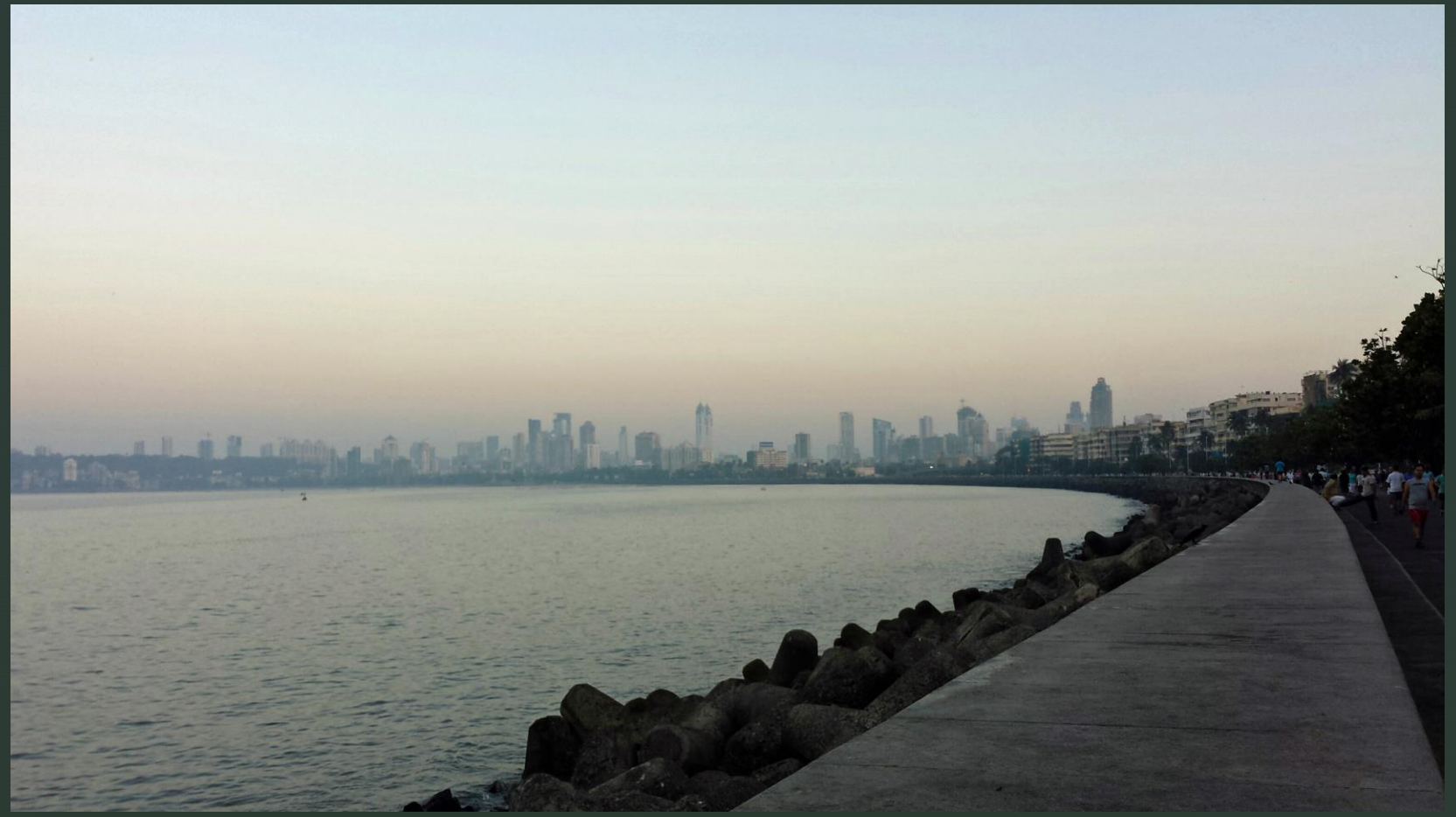


Coursera Capstone Project – Applied Data Science

Clustering India’s Urban Cities

Problem Context – What, Why and Who

ASSIST THE INDIAN SMART CITIES PROGRAM

IDENTIFY CANDIDATE CITIES THAT CAN BENEFIT FROM THE PROGRAM

MINISTRY FOR URBAN DEVELOPMENT IN INDIA

Data Sources

1. **World Cities Database** - The World Cities Database (https://simplemaps.com/data/world-cities) has a list of prominent cities from all countries in the world, along with their geo coordinates.

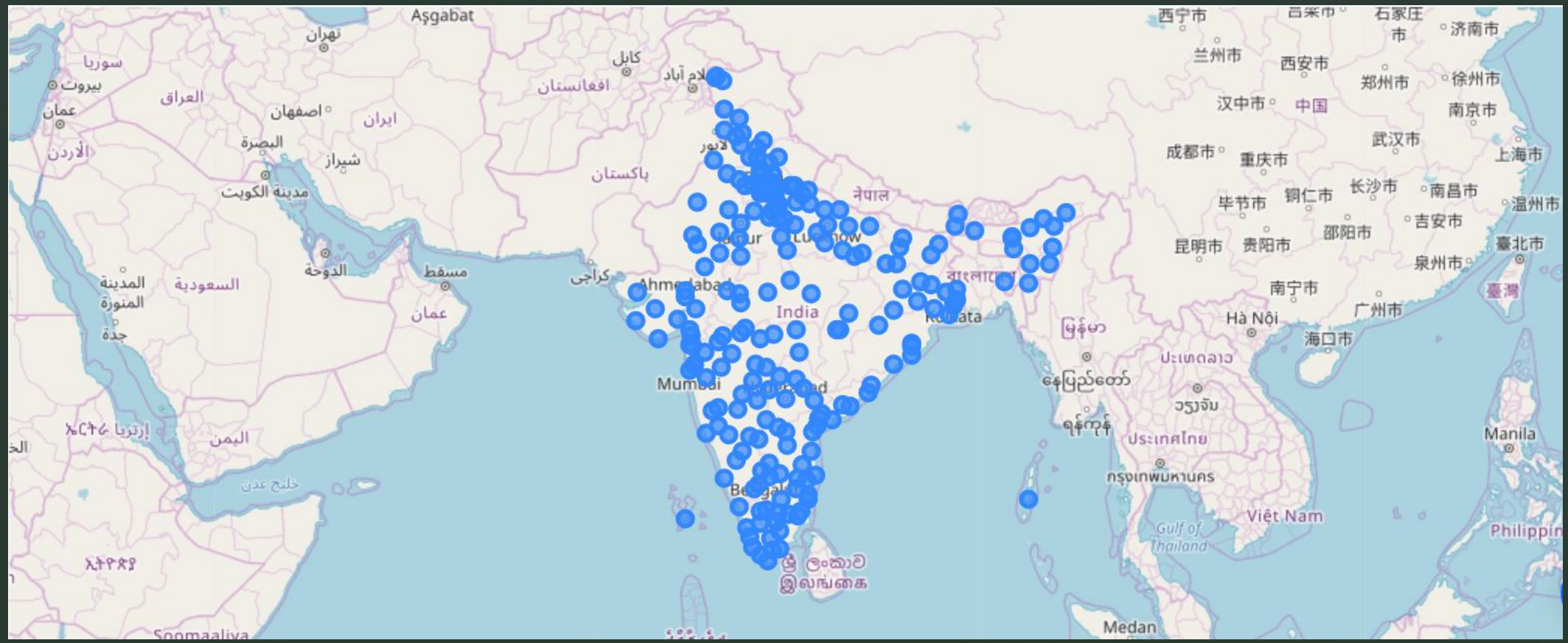
2. **Foursquare** - Foursquare is a social networking mobile app that enables users to 'check-in' and share their location when they visit venues. This project will use the 'Search for Venues' and 'Get Details of Venue' features of the Foursquare API to find out the nature of locations being visited in each city.

Data Cleansing

- Extract Indian cities from World Cities Database
- Remove redundant columns
- Remove cities with under 5 venues
- Consolidate highly common and frequent venues

After data cleansing**, there were 210 cities and 268 unique venue categories**. There were an average of 14.5 unique categories per city, 47 in the city with the most categories and 1 in the least. 30% of cities had 5 or lower unique venue categories.

Cities Prior to Clustering



Feature Set

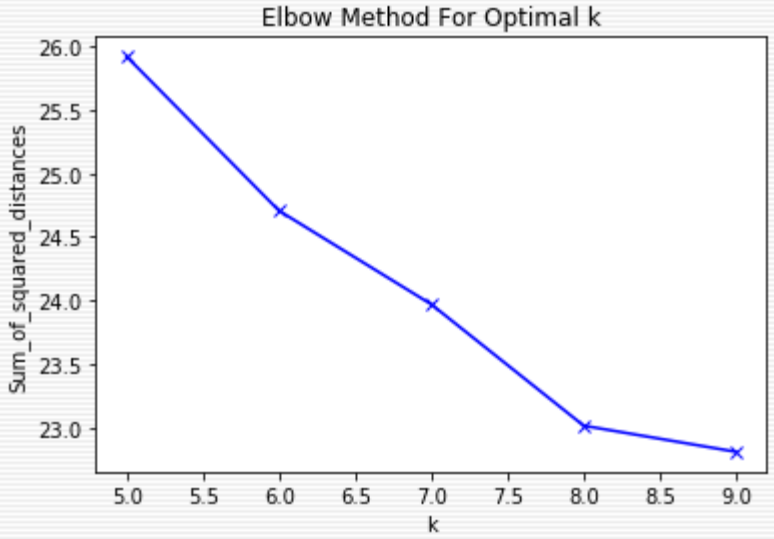
1. **World Cities Database** - The World Cities Database (https://simplemaps.com/data/world-cities) has a list of prominent cities from all countries in the world, along with their geo coordinates.

2. **Foursquare** - Foursquare is a social networking mobile app that enables users to 'check-in' and share their location when they visit venues. This project will use the 'Search for Venues' and 'Get Details of Venue' features of the Foursquare API to find out the nature of locations being visited in each city.

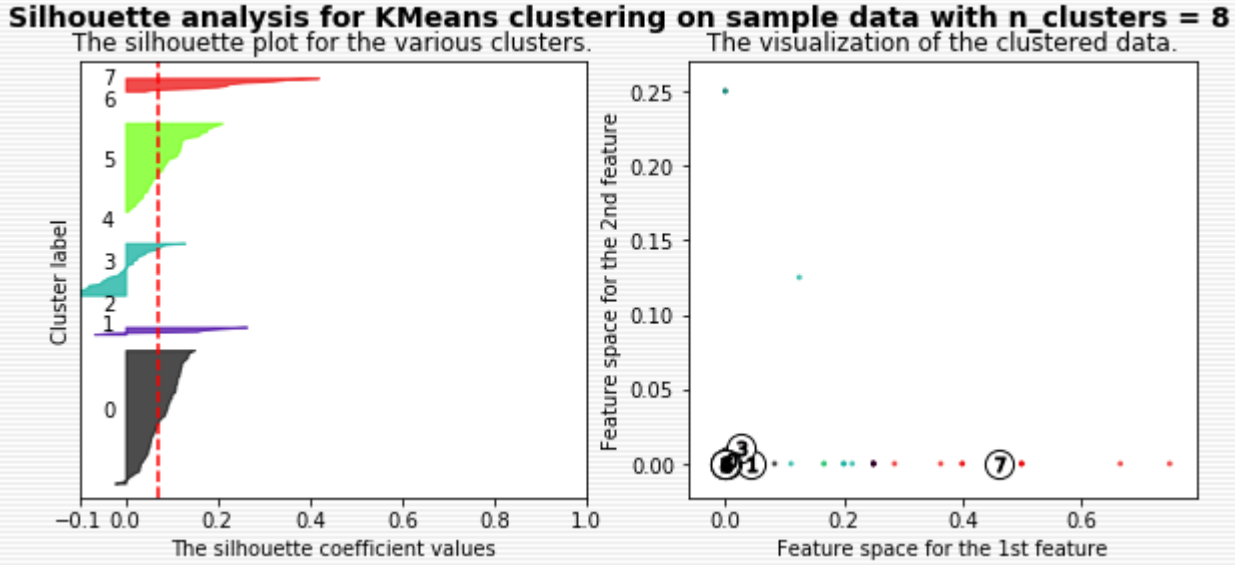
K-Means Clustering

Optimal number of clusters = 8

Elbow Method



Silhouette Method

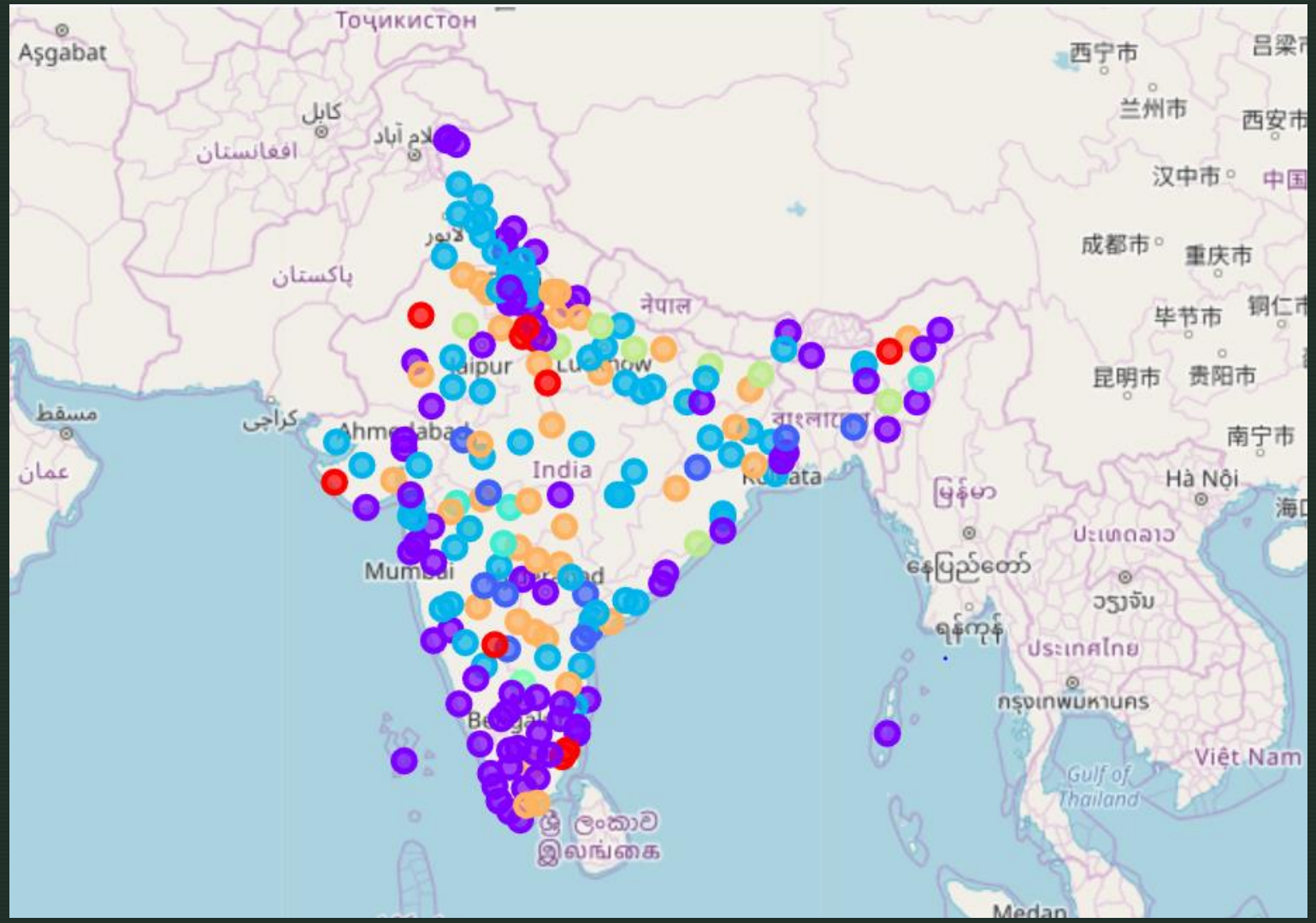


K-Means Clustering

Result

Cluster	Size	Current Smart City
1 – Historic Sites	10	Agra
2 – Urban I	77	Ahmedabad
3 – Railway Junctions	11	Sholapur
4 – Mid-size Towns I	62	Lucknow
5 – Outlier I	4	Kohima
6 – Outlier II	1	--
7 – Mid-size Towns II	9	Muzzafarpur
8 – Mid-size Towns III	37	Gwalior

Cities in Clusters



Limitations

Recommendations for Future Work

Data - Reliance on Foursquare checkins

Additional feature data relating to infrastructure, employment, population, income, industries etc. need to be considered

Lack of Feature Scaling

Feature scaling should be explored to mitigate this, but was not done due to time constraints.

Choice of Algorithm not fully suited to data

A more robust implementation of DBSCAN or other such algorithms could be explored.

Priyanka Chandrasekar

01 July 2019