# CONTRASTING AND COMPARING THE METROPOLITANS OF INDIA

MANSI PARASHAR

February 24, 2020

## 1.INTRODUCTION

### 1.1 Background

India is the 7th largest country in the world, and the 2nd most populated country.
India is a developing State and has 4 metropolitan cities so to speak of, Delhi which is the capital city, Mumbai(formerly Bombay), Chennai(formerly Madras) and Kolkata(formerly Calcutta).
In recent years, India has seen a tremendous boost in job opportunities and the general quality of life in these metros, leading to a large rate of migration to the aforementioned cities.
These 4 cities belong to different parts of the country, and hence vary significantly in the type of life they offer.

### 1.2 Problem Statement

Aim to analyse the 4 metropolitans of India ; explore the neighborhoods with relevant data and state the final findings in terms of the various venues and amenities provided at each city.
Contrast and compare this data retrieved for the 4 cities on the basis of different clusters formed by the neighborhoods and the services/amenities these clusters have to offer, and how they line with the interest of the stakeholders.
Also, compare the metropolitans on the basis of different unique amenities provided by them and the corresponding quantity.
Finally, identify the different regions in each metropolitan based on shared common venues to locate possible choices for potential residents.

### 1.3 Possible Applications

A family\An individual is looking forward to migrating to one of the metropolitans for better opportunities and for a better quality of life. Which neighborhood would be the best suited for them, and why?
A business owner wants to expand to one of the metros. What are the best regions for them to open up a business to conceivable success?
How similar are metropolitan cities, even when they are located in far corners of the country?
What are the differences between a city and a nearby metropolitan, on the basis of amenities offered?

## 2. DATA ACQUISATION AND PRE PROCESSING

### 2.1 Data Sources

We work on 3 sets of data :-

**2.1.1** Data containing the neighborhoods, pincodes,latitude, longitude and for cities in India.This data is accessed from the official government of India site,https://www.india.gov.in/ which serves as a national portal for accessing different data concerning the country of India.
To access the data set, we need to fill in a form declaring we're using this data for academic purposes.
Unfortunately, all the latitude and longitude values were NaNs.

**2.1.2** Furthermore, we found a data set containing Indian pincodes and their latitude and longitudes from the same site.

**2.1.3** Based on the latitudes and longitudes in our database, we use the FourSquare API to access data about the neighborhood venues required for this study.
The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places.
Additionally, Foursquare allows developers to build audience segments for analysis and measurement.
JSON is the preferred response format which needs to be converted into the required dataframe
Accessing the API would require us to make an account on the Developer's Portal on the foursquare website,https://developer.foursquare.com/

### 2.2 Data Cleaning

From the second database, we delete the columns of city, neighborhood and accuracy to only get the pincodes for latitudes and longitudes, for a cleaner merge.

| | pincode | latitude | longitude |
|---|---|---|---|
| 0 | 110001 | 28.6333 | 77.2167 |
| 1 | 110002 | 28.6333 | 77.2500 |
| 2 | 110003 | 28.6500 | 77.2167 |
| 3 | 110004 | 28.6500 | 77.2167 |
| 4 | 110005 | 28.6500 | 77.2000 |

After merging on pincode, we get the following accurate dataset. We group the neighborhoods belonging to the same pincode and region.

Out[5]:

| | pincode | latitude | longitude | regionname | neighborhood |
|---|---|---|---|---|---|
| 0 | 110001 | 28.6333 | 77.2167 | Delhi | Baroda House S.O,Bengali Market S.O,Bhagat Sin... |
| 1 | 110002 | 28.6333 | 77.2500 | Delhi | A.G.C.R. S.O,Ajmeri Gate Extn. S.O,Darya Ganj ... |
| 2 | 110003 | 28.6500 | 77.2167 | Delhi | Delhi High Court Extension Counter S.O,Delhi H... |
| 3 | 110004 | 28.6500 | 77.2167 | Delhi | Rashtrapati Bhawan S.O |
| 4 | 110005 | 28.6500 | 77.2000 | Delhi | Anand Parbat Indl. Area S.O,Anand Parbat S.O,B... |

Because it is location data, there are no such outliers ; also, as the datasets have been merged on pincode which is a unique value for each region, the dataset is void of NaNs.

The foursquare data is returned as a json file. All the information is in the “items” key.
We extract the categories of the venue
We clean the json and convert it into a pandas dataframe

### 2.3 Feature Selection

From the dataset containing pincodes of all of India, we select the relevant ones;
That is, Pincodes corresponding to the cities :-

| **City** | **number of rows** |
|---|---|
| Mumbai | 197 |
| Delhi | 79 |
| Kolkata | 180 |
| Chennai | 141 |

Which corresponds to the size of the respective cities.

The features we’ve kept are
Pincode , Latitude, Longitude,City,Neighborhood.

Which are the only ones relevant to get the necessary foursquare data for exploratory analysis of neighborhoods in a city.