BATTLE OF INDIAN CITIES

INTRODUCTION: BUSINESS PROBLEM

In this project we will try to find the optimal location among Indian major cities for the initial launch of any specific CONSUMER GOODS products. This will also be helpful for already established companies who wish to expand their business to other cities of India, and do not have enough data to figure out which newer cities are more likely to welcome their PRODUCT A and which will be the best site for their PRODUCT B and so on.

This project categorizes the major cities with similar characteristics into clusters and helps the companies based on **CPG (Consumer Packed Goods)** and **FMCG (Fast-Moving Consumer Goods)**, to plan for the deployment and distribution of their products and efficiently target different cities corresponding to their different products.

DATA ACQUISITION AND CLEANING

Based on the description of our problem, we will be needing a list of major cities (according to the population and various other characteristics) for this project. The population data, literacy rate, density per sq. km and sex-ratio are some of the main data that we need to group the cities into clusters of similar cities. I got these from the link below:

- **https://www.nriol.com/india-statistics/biggest-cities-india.asp**

The coordinates (latitudes and longitudes) for each of these major cities were required in order to get the nearby venues so that we could group these cities into clusters based on the type of venues in the vicinity of the cities.

- **geopy** is a Python client for several popular **geocoding** web services. **geopy** helped me locate the coordinates of the cities using **geocoders**.

Further data regarding nearby venues was gathered using the ***foursquare API.*** Different cities vary greatly in their size and lifestyle. So, I set the RADIUS of 10 km and limit of 1000 so that I won't miss out anything.

- **https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION, LATITUDE, LONGITUDE, RADIUS, LIMIT)**

FEATURE SELECTION

After acquiring and cleaning the data, we were left with a dataframe consisting of 100 rows and 8 columns. On further analysis of each column's relevance to our project, it was clear that some features were redundant and needed to be removed from the dataframe. For example, the 'Indian States' and the 'Main Language' columns are not really significant to our purpose.