# Coursera Capstone

## IBM Applied Data Science Capstone

## *Opening a new multiplex in Kolkata, India*

By: Ritesh Manna
Date:February 2020

# Background

A multiplex is a movie theater complex with multiple screens within a single complex. They are usually housed in a specially designed building.

For many people visiting a multiplex is a great way to relax and enjoy themselves during the time of holidays and weekends. It is one of the few destinations where one can watch newly released movies. For retailers large crowd in multiplex provides a good distribution channel to market their products and services.

This project is particularly useful for property developers who are looking to open or invest in new multiplexes around the city of Kolkata. Location is one of the most important decisions that will determine whether the multiplex will be a success or a failure.

# Business Problem

The aim of this project is to analyse and select the best  locations in the city of Kolkata to open a new multiplex. Using data science methodologies and Machine Learning techniques like clustering, this projects tries to answer the question - In the city of Kolkata, India if a property developers is looking for a location to open a new multiplex what would you recommend?

# Data

**To solve the problem, we will need the following data**:

- List of neighbourhoods in Kolkata, India.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to multiplexes. We will use this data to perform clustering on the neighbourhoods.

## Sources of data and methods to extract them:

This url (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Kolkata) contains a list of neighbourhoods in Kolkata, with a total of 199 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use the Foursquare API to get the venue data for those neighbourhoods. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping, working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).