**Pennsylvania Movie Theatres – Capstone Project**

**1. Introduction**

**1.1 Background**

Indian population across Unites States of America is siginifcantly large compared to other Foreign Nation’s population. In the recent decades, Indian population in United states has grown substancially with more than 4.5 million Indian residents in United States as of last census. Culturally Indian’s are more penchant and inclined towards various art forms out of which Cinema is the dominant interest among the Indian communities.

**1.2 Problem**

Indian movies have seen tremendous growth in the screening opportunities across the United States. In the last decade, the movies used to be screened in Major states/cities like NY, NJ, CA, TX etc. In current decade more locations are being explored to screen the Indian movies and average gross revenue generated for popular Indian movies across the country has been increased to 1 Million dollars. Indian residents do show interest in acquiring the distribution rights and exhibiting the movies to screen across various locations in the country. It is essential to know from the distribution stand point to understand the latest Indian population metrics , distribution across the cities and availability of screens to exhibit the movies and fixing the ticket prices. To start with , I am exploring the cities in Pennsylvania and Theatre venues for the Capstone project with the data available.

**1.3 Stakeholders**

Distribution agencies of Indian movies, Local exhibitors and other Indian residents of interest in screening movies in respective locations may be interested in a model that is able to segment the screening venues based on the population density and proximity of the screening locations. With further data available like historical screening locations, ticket prices of Indian movies, the model can be built even more efficiently to understand the location segments which can be targetted for regular screenings.

**2. Methodology & Data**

**2.1 Data Sources**

Details of Indian population based on city and zipcode within the state of Pennsylvania from zipcodeatlas. Results of cities / zipcodes with top 100 percentage of Indian population are considered for the Project.

Foursquare places API to acquire the venues related to the movie theatres across the state of Pennsylvania. Foursquare defines ‘Venue’ as a place of interest. Each foursquare venue is assigned to a category and each category has a category id. The criteria used to retrieve the data will be based on the category id of movies, near by locations as the cities in pennsylania. Radius, search intent, query, latitude, longitude and latitude longitude accuracy, result limit will be used iteratively to achieve the best possible data of movie theatres and multiplex venues based on the location of search. In this project the input paramerts used to access the foursquare api end points are ‘categoryId’,‘near’, ‘intent’, ‘llAcc’,’query’,’radius’,’limit’ and other client access parameters.

**2.1.1 Data Retreival**

Data pertaining to Asian (Indian) population in Pennsylvania by zip code is obtained from zipcodeatlas. The link to the website was read into pandas dataframe. After scraping through the data, the right index location was obtained to get the Indian population data into a tabular format. Columns were added to the dataframe obtained.

Dataset related to the Movie theatre venues across Pennsylvania state was obtained using Foursquare places API. the input paramerts used to access the foursquare api end points are ‘categoryId’,‘near’, ‘intent’, ‘llAcc’,’query’,’radius’,’limit’ and other client access parameters

A function was defined to recursively sends a ‘get’ request to Foursquare for each city that requests a search of movie theatres. While looping through each city from the PA dataset, the function appends each movie theatre venue as a venue entry to a list and, after looping through each city, creates a DataFrame of all of the results. Included for each entry in the dataset are city, venue name, venue category and address.

**2.1.2 Data Cleaning**

The Population dataset has been cleaned by applying the following steps.

* Obtain the city and state separately using split function.
* Drop the unwanted columns like index, rank and location
* Drop the cities with population less than 1000
* Converted the Percentage of Indians to Float

The Theatre Venues dataset has been cleaned by applying the following steps.

* Dropped the venue categories other than ‘Movie Theatres’ and ‘Multiplex’
* Droped the Venue latitude and longitudes
* Removed duplicates on the Venue names as this API looped search will retreive similar venues ‘near’ the input cities.

Data obtained from Four square API was not consistently giving the ‘City’ column of the search criteria for specific locations. As a result the city column cannot be used. This caused improper classification of the venues associated with the cities. To minimize this issue, the latitude longitude accuracy was incorporated in the search criteria.

**2.1.3 Feature Selection**

Initial set of rows obtained for the population data was 100. The data has been cleansed and population less than 1000 was removed and 83 cities/locations were selected for the analysis. The columns selected were city, population, percentage of indians and indian score (calculated based on percentage and population).

The foursquare venue information retreieved 625 records for the Theatre venues within 10000 meters radious for each city from the population matrix. This record set has been minimized to 173 records after removing the venue categories like shopping malls, indie theatres etc. The columns selected are the location, venue catrgory and venue.

**2.1.3 Exploratory Analysis**

Folium maps have been generated on the venue details obtained from Foursquare API to check if the venue details are distributed across the state of Pennsylvania. Ciites with highest count of venue categories and near by cities which were misclassified , not present in the population dataset were identified. The parameters to retrieve results from Foursquare API have been revisited and changed to get the optimal results. Ex:- using radius of 50,000 (maximum) and limit of 50 (maximum) is yielding duplicate venue results which were not tagged to any city (the JSON file does not always return the ‘city’ column for each request. Iterations have been performed to determine the best possible combinations of radius, limit and location accuracy parameters while initiating the search function of foursquare API.

Following questions have been put on the dataset and data was prepared accordingly.

How many venues are misclassified in wrong venue category?

How many venues are missing address details ?

How many venues are repeating in the result dataset with different venue names?

**3. Modeling**

Since the data we are dealing with is based on the parameters like cities, population (Indians), venues and venue categories. We are trying to understand how to distinguish the locations based on the population. A clustering model would be required to accomplish this. The model adopted for this work is K means clustering model.

**3.1 Preparing data for Clustering**

The Foursquare API data has details of venues and theatre counts by categories. This has been grouped and summarized based on the city to obtain the total number of thatres for each city .

The Population dataset has city, population, percentage of indians and indian score details. We have acquired top 100 cities by percentage of zipcode and population. The dataset has duplicate cities (based on zip codes). For example Philadelphia city comes under top 100 indian population ‘across’ various zip codes. To accumulate this data, we need to aggregate the population count and get the sum of population/percentage by city.

The Venue data summarized by theatre counts and category across various cities and population percentage data across various cities has been merged and applied to the scalar function to normalize the data to be fit into the k means clustering model.

Once the final clustered data has been prepared using city, percentage of indians, movie theatres in the city and categories, K means clustering model has been used to fit the data.

**3.2 Cluster Modelling**

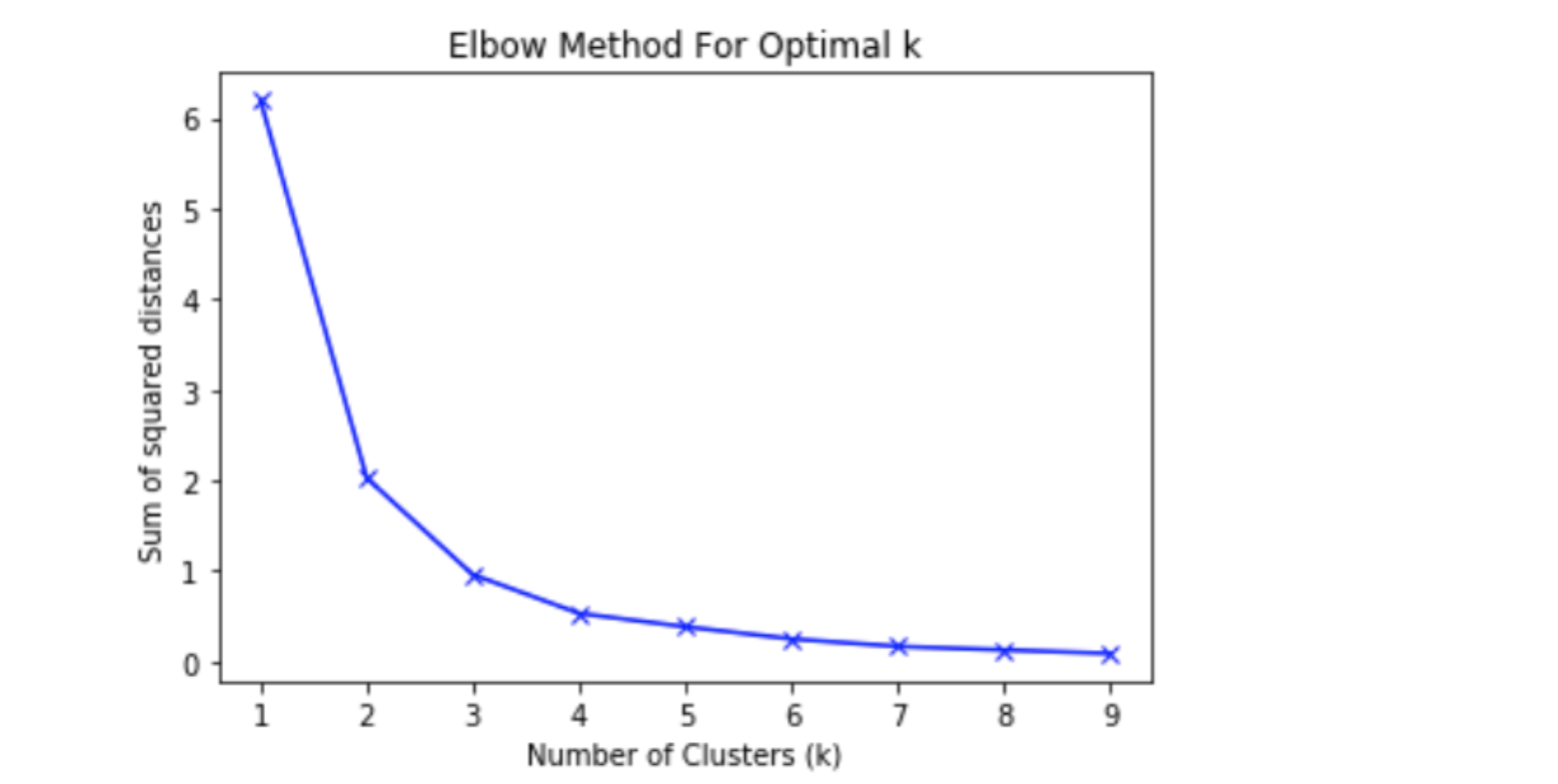
Scikit-learn’s K-Means clustering was used to segment the cities across pennsylvania on the percentage of Indian population in the cities.

Modelling has been applied initially for 4 clusters (Kclusters = 4) to evaluate the clustering of the data.

Queries have been run on the clustered data to understand the clustering across the 4 clusters. Scattered plots have been generated to visualize the results.

The model has been then iterated with number of clusters starting from 1 to 10 and linear plot has been generated for each iteration (cluster label) and sum of the squared distances from the cluster for each iterations.

Elbow method has been used to plot the data between sum of squared distances vs Number of clusters to determine what number of clusters will yeild the optimal results. The data point which shows a break in the reduction of sum of mean squared distances and gradually turns out into a linear relation has been taken into consideration to arrive at the optimum cluster segmentation for the clustered data. In this case the optimal number of clusters is 3.



**4. Model Evaluation**

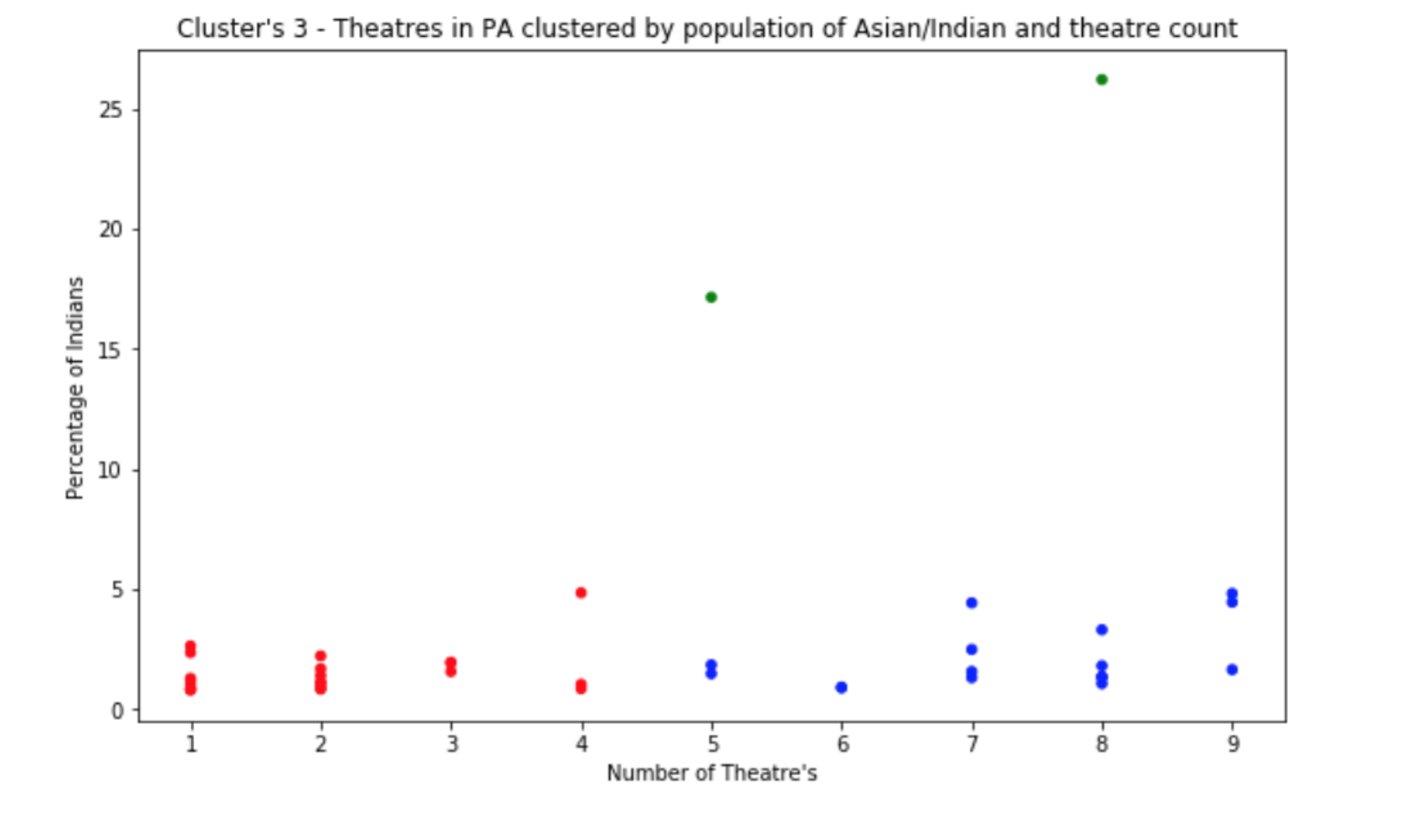
Result of each cluster has been generated by iterating the cluster labels from 0 to 2. Segmentation of the city, population, theatres has been evaluated from the results of each cluster. It is interesting to see that clusters are mainly segmented based on the theatre count. Some clusters are relatively very small and other clusters are of similar size of result set and they apepar to have achieved a profile based on the number of theatres in the near by locations of the city in search.

Cluster 1 – (cluster label 0) – Contains mostly the cities with Low number of theatres (up to 4) and less percentage of Indian population (up to 5%)

Cluster 2 – (cluster label 1) – Contains mostly the cities with medium to high number of theatre counts (6 to 9) and less percentage of Indian population (upto 5%)

Cluster 3 – (cluster label 2) – Contains the cities with Moderate to High number of theatre counts ( 5 to 8) with High percentage of Indian population (17% to 26%). The size of this cluster is small (Philadelphia & Pittsburg)

Scattered plot has been generated on Percentage of Indian Population, Theatres count and Cluster labels to understand the segmentation of each cluster and look at the ‘high’ points in each cluster. The high points at the cluster level are the locations with optimal population percentage and theatre counts at the cluster level.



**5. Conclusion**

In this study, analysis has been performed on the location based data pertaining to the movie theatres for the locations with highest Indian population. Identified the category of the movie venues like Multiplex, Movie Theatres, Indie Theatres, drive in theatres and narrowed down the analysis to the specific categories pertaining to the focus of the research. Pre-processed the clustering data based on the population, city, number of theatres and categories. Applied clustered modeling using K means algorithm and evaluated the results by segmentation of these data points.

### Examining the data points across the clusters provided an idea of the cities by Indian population and number of theatres available near by. The cities were clustered predominantly on the theatre count except label 2 is segmented based on population(high) & theatre count (medium to high). In each segment we can segregate the highest data points (close to percentage of Indians 5) which gives us those cities which have the right proportion of Population by Theatre counts. The cities Philadelphia and Pittsburg are having more Indian population with more number of Theatres available for screening. Other cities like King of Prussia, Bensalem, hatfield and Scranton can be considered for screening opportunities for Indian movies.

**6. Future directions and challenges**

The analysis and model in this study has been mainly focused on Movie theatres available ‘nearby’ to the locations with highest Indian population. Further research can be done on the occupancies of the theatres, ticket pricing by using foursquare venues exploration and recommendations API. However there will always be limitations on getting the right input data ..i.e population , city statistics. With right amount of data pertaining to Indian population spread across various cities, ability to retrieve the accurate data from foursquare or other location based API’s, the research can become much stronger and efficient. Clustering can be done not only on the limited parameters like the movie theatres and categories but also on near by popular venues, foot falls in the theatres for recent movies and ticket pricing. At the end, data availability is the key challenge for data science. This study can be ..not limited to Pennsylvania and extended to other states and across movies to explore the screening opportunities and optimizing the rate of returns for Indian movies screened in United States of America.

**7. References**

<https://zipatlas.com/us/pa/zip-code-comparison/percentage-indian-population.htm>

https://developer.foursquare.com/docs/api/venues/search