Machine Learning To Find A Neighborhood to Start A New Business

# Overview

You’ve done the initial legwork. You know your target market and the product or service you’re going to sell. But you’re still undecided about where you want to set up shop, and the decision shouldn’t be taken lightly. From street visibility and foot traffic to the cost of commercial real estate, choosing the right space makes a big impact on your bottom line. It really does come down to “location, location, location.”

In this article we’ll use Machine Learning to find an optimal set of neighborhoods to start a new business and all this using freely available data. To be specific we will use K-Means clustering (an unsupervised algorithm) for finding a Mumbai neighborhood to open a new Chinese restaurant.

# What is Machine Learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch of [artificial intelligence](https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html) based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

## Machine Learning Methods

There various methods of machine learning among which the widely used ones are Supervised Learning, Unsupervised Learning, Semi-Supervised Learning & Reinforcement learning a brief overview of these is given below.

### Supervised Learning

Supervised learning algorithms are trained on labeled data i.e. a set of inputs for which output is known. The training results in a model that can then be used to predict the output for new inputs, e.g. using supervised learning a model can be build on a set of insurance policy applications and risk attached to them. This model can then be used to predict risk for new applications.

### Unsupervised Learning

Unsupervised learning algorithms are used on data that have no labels. The algorithm must find out what is lies hidden in the data. The goal is to find a pattern a structure in the data. Unsupervised learning works well on transactional data, e.g. it can used to cluster customer data into various segments of similar features and then customers in same segment can be treated similarly in marketing campaigns.

### Semi-Supervised Learning

Semi-supervised learning is used when there is a small amount of labeled data and large amount of unlabeled data (because unlabeled data is less expensive and easier to acquire). The idea is to first train the model on labeled data, use this to label unlabeled data, train the model again on the complete set of labeled data and the use this new model for prediction. An example application is face recognition on a web cam.

### Reinforcement Learning

Reinforcement learning is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. This type of learning has three primary components: the agent (the learner or decision maker), the environment (everything the agent interacts with) and actions (what the agent can do). The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good policy. So the goal in reinforcement learning is to learn the best policy.

For a more detailed and “simple” coverage of Machine Learning the reader can check this article: <https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/>

Let’s start with finding the optimal set of neighborhoods for opening a new Chinese restaurant in Mumbai city the financial capital of India.

# Business Problem

## Problem Background

**Mumbai**, formerly **Bombay** city, capital of [Maharashtra](https://www.britannica.com/place/Maharashtra) state, southwestern [India](https://www.britannica.com/place/India). It is the country’s financial and commercial center and its principal port on the [Arabian Sea](https://www.britannica.com/place/Arabian-Sea). Located on Maharashtra’s coast, Mumbai is India’s most-populous city, and it is one of the largest and most densely populated urban areas in the world.

Mumbai, long ago the center of India’s cotton textile industry, subsequently developed a highly diversified manufacturing sector that included an increasingly important information technology (IT) component. In addition, the city’s commercial and financial institutions are strong and vigorous, and Mumbai serves as the country’s financial hub

Mumbai's business opportunities, as well as its potential to offer a higher [standard of living](https://en.wikipedia.org/wiki/Standard_of_living_in_India), attract migrants from all over India, making the city a [melting pot](https://en.wikipedia.org/wiki/Melting_pot) of many communities and [cultures](https://en.wikipedia.org/wiki/Culture_of_India).

The market in Mumbai is highly competitive. It is highly developed city and hence cost of doing

business is also one of the highest. Thus, any new business venture or expansion needs to be analyzed

carefully. The insights derived from analysis will give good understanding of the business

environment which will help in strategically targeting the market. This will help in reduction of risk and the Return on Investment will be reasonable.

### Problem Description

Running a restaurant is more than just offering food in exchange for money. One is offering an experience to customers based on items such as the decor, food, and service. As the owner, one must determine the type of experience one wants to give customers. Even after one has decided on the type of restaurant, one needs to make sure there is a market for it, the concept is within budget, and find the right location. **In this report we will be using the data science way to find the best possible location in Mumbai for opening a Chinese restaurant.**

Keeping in line with Mumbai’s cosmopolitan atmosphere the scene for dining and eating out in Mumbai is fabulously all-inclusive. The city has an extensive range of fine restaurants serving Indian, French, Italian, Chinese, Thai, Japanese, Lebanese, Arab and Mexican food. The best restaurants of Mumbai run the gamut from top of the line elegant establishments with the finest selection of wines to family style restaurants to shacks and roadside vendors and stalls. Whatever may be your preference, whoever you are, whatever your wallet size, Mumbai is sure to adapt to your requirements. Thus, one can see that the restaurant market in Mumbai is very competitive and to open a new restaurant and survive in Mumbai it is important to strategically plan the whole process especially selection of location for the new restaurant. Various factors play an important role in the selection of a location for a new restaurant:

1. Visibility

You may have found the cutest, quaintest location for a restaurant in a historic downtown somewhere, but if you’re tucked in a side street with little foot traffic, unless you have amazing marketing tactics, you may have the most Instagram- and Pinterest-worthy space with zero customers. Some measures of visibility are:

* + Foot fall
  + Car traffic

1. Parking Space
2. Space Size
3. Crime Rates

Crime rates are unglamorous considerations, but if you place your restaurant in a crime-laden area, are your target customers going to visit? High crime rates can make potential customers uncomfortable, and if they feel they’ll be mugged walking to their cars, it will only drive away business, no matter how legendary your coq au vin.

1. Surrounding Businesses and Competitor Analysis

You’ll want to do your research surrounding businesses. Are they doing well? Is the area affluent? Is there enough room for your business? Also, you’ll want to know what types of restaurants do well in the area; however, you don’t want to open a pizzeria if there are four in the area. Areas can only support so many of the same type of restaurant. What will distinguish any new restaurant is excellent service and consistently wonderful food.

1. Accessibility

There’s a reason that major restaurant chains are often located near highway exits: It makes them accessible for customers. Certain restaurants can get away with food or service that isn't the best simply because their locations are so accessible, like restaurants near the Eiffel Tower or Coliseum. There is plenty of foot traffic in urbanized areas, and restaurants only need to attract customers from the street into their business.

1. Affordability

Cost is always a bottom-line consideration for any business. If the rent or purchase of the space is more than you’ll bring in each month in profits, that location is not feasible at that time. Safety

1. Population Base

Are there enough people in the area to support your business? There need to be enough people who live in or pass through the area regularly to keep you busy

1. Style of operation

Is your operation going to be formal and elegant? Or kicked-back and casual?

1. Proximity to other businesses and services

Take a look at what other businesses and services are in the vicinity from two key perspectives. First, see if you can benefit from nearby businesses--by the customer traffic they generate.

Second, look at how they'll enrich the quality of your company as a workplace.?

**And the list can just go on….**

### Target Audience

The report would be beneficial to anybody who would like to open a Chinese restaurant in the city of Mumbai. The objective is to find and recommend a neighborhood in Mumbai where a new Chinese restaurant can be started.

### Success Criteria

Based on Mumbai city data that is freely available the recommendation will be based on following factors that have been listed above:

1. Accessibility of the neighborhood.
2. Population Base of the neighborhood.
3. Number of restaurants (all inclusive) in the neighborhood.
4. Number of Chinese restaurants in the neighborhood.

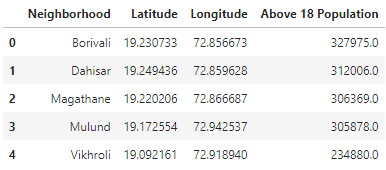
## Data

### Mumbai Neighborhood Data

Mumbai city is divided into regions in various ways based on different administrative requirements like municipal wards, assembly constituencies, etc. However, our requirement for population base is met by the assembly constituency division population. Data for any other division is not available freely. The data for constituency, hereby referred as neighborhood, will be collected as follows:

1. Neighborhood List will be scrapped from <https://en.wikipedia.org/wiki/List_of_constituencies_of_the_Maharashtra_Legislative_Assembly>
2. Neighborhood latitude & longitude data will be obtained from the google API <https://maps.googleapis.com/maps/api/geocode/json?address>

The header of data obtained is as in figure 1.



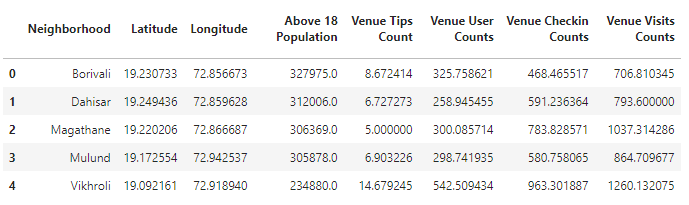
**Figure 1**

### Population Data Base

Total population for each neighborhood is not available freely. However, for each constituency total voters who voted in 2014 and voter turnout is available. Total population above the age 18 years can be calculated from these and this can be used as a proxy for the total population. This data can be obtained from Wikipedia using the link [https://en.wikipedia.org/wiki/<Constituency Name>\_(Vidhan\_Sabha\_constituency)](https://en.wikipedia.org/wiki/%3cConstituency%20Name%3e_(Vidhan_Sabha_constituency)) and using the 2014 election data. For eg. data for Mahim constituency can be scrapped from the link: <https://en.wikipedia.org/wiki/Mahim_(Vidhan_Sabha_constituency)>. See Figure 1 above for the data.

### Accessibility of the neighborhood

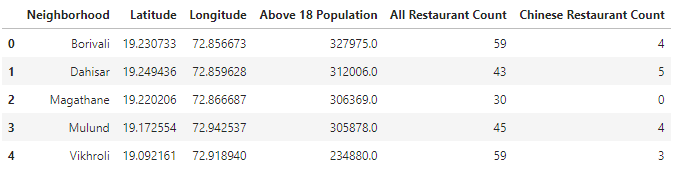
Accessibility of a neighborhood can be determined by the average of visit counts of places of interests in the neighborhood. This data can be obtained from **foursquare.com** using the API <https://api.foursquare.com/v2/venues/> and using the **stats** data from the result. The header of data obtained is as shown in figure 2



**Figure 2**

### Total Number of Restaurants

The total number of restaurants can be obtained from **foursquare.com** using the API <https://api.foursquare.com/v2/venues/explore>? categoryId= 4d4b7105d754a06374d81259 . Total Chinese restaurants can be filtered from this data. The header of data obtained is as shown in figure 3.



**Figure 3**

## Methodology

The methodology for the process is as follows:

1. Date Collection:
   1. Collect data from the sources mentioned in the data section
   2. Process the data (clean, transform, etc.)
   3. Load data into pandas dataframes for exploratory analysis
2. Do exploratory data analysis to get insights into data as to:
   1. Find category wise restaurant total to ascertain how are Chinese restaurants placed against other categories number wise
   2. Plot all restaurant total and Chinese restaurant total neighborhood wise, Chinese restaurant total and visit counts neighborhood wise. This will give an idea of probable location for opening a new Chinese restaurant
3. Aggregate and merge all the dataframes into one dataframe.
4. Use machine learning (train a model on the data using K-means clustering) to get the best neighborhood group for opening Chinese restaurant.

### Date Collection

Data was collected from sources as mentioned in the Data section above.

Mumbai neighborhood data is in html format. The actual data had to extracted by parsing the html pages. For some neighborhood the names were in Marathi, these had to be converted to English names for getting their location and population data e.g. Bandra is mentioned as Vandre so Vandre was changed to Bandra to get its location and population base. For some neighborhood turnout percentage was not available to calculate total above 18 years population, in this case Maharashtra state’s turnout percentage was used. Figure 1 above gives a snapshot of Mumbai’s neighborhood data after parsing, extracting and transformation.

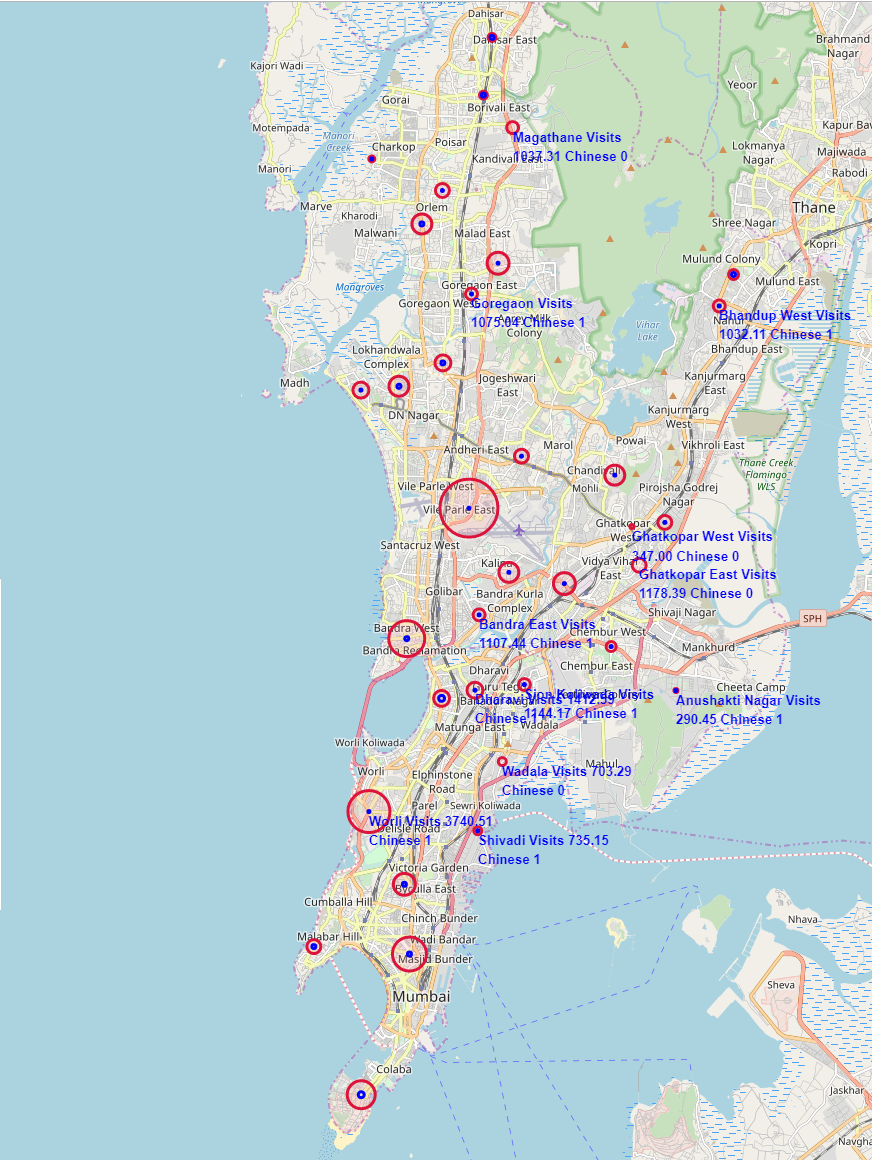
Restaurant data and venue statistics were obtained from foursquare.com. The data obtained was in Json format. The actual data had to extracted by parsing the Json data. The restaurant data was summed neighborhood wise and for venues mean of their statistics were calculated neighborhood wise. Figure 2 above is snapshot of the venue statistics and Figure 3 is a snapshot of restaurant data.

### Exploratory Data Analysis

|  |
| --- |
| Explore Mumbai’s neighborhood on a map Mumbai’s neighborhood data is scrapped from the links mentioned in the data section and stored in dataframe as shown in figure 1. Let’s plot this data on a map as follows:    **Figure 4: Mumbai Neighborhood** Explore Restaurant Counts Figure 5 plots, restaurant category wise, the total number of restaurants in Mumbai city. In the figure one can see that in the case of authentic cuisine restaurants, Chinese restaurants are most popular after Indian restaurants.    **Figure 5: Restaurant Counts** Explore Total Restaurant Count v/s Total Chinese Restaurant Count Let’s explore Neighborhood wise Total Restaurant Count and Chinese Restaurant Count. Neighborhoods with less than 2 Chinese restaurant counts have been explicitly labeled. These neighborhoods could be candidates for opening next Chinese restaurants.    **Figure 6: All Restaurant v/s Chinese Restaurant** |

From the map in figure 5, one can see that Worli (Tot. Rest.39 Chinese 1), Ghatkopar East (Tot. Rest.27 Chinese 0) and Bandra East (Tot. Rest.33 Chinese 1) seem to be good candidates for opening next Chinese restaurant as they have good amount of other kinds of restaurants indicating a good amount of eating-out crowd and have 0 or 1 Chinese restaurants.

#### Explore Chinese Restaurant Count v/s Average Visit Count



**Figure 7: Chinese Restaurant Count v/s Average Visit Count**

From the map in figure 7, one can see that again Worli (Avg. Visit Count: 3740.50, Chinese 1), Ghatkopar East (Avg. Visit Count: 1178.39 Chinese 0) and Bandra East (Avg. Visit Count: 1107.44 Chinese 1) seem to be good candidates for opening next Chinese restaurant as they have good visit counts indicating a good amount of crowd visiting places in the neighborhood and have 0 or 1 Chinese restaurants.

### Clustering of Neighborhoods

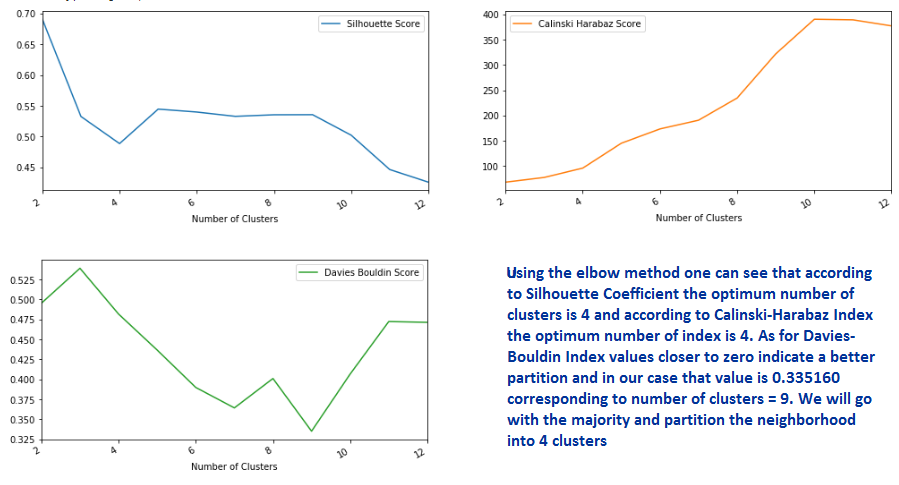
Let’s cluster the neighborhood data created to get the set of neighborhoods that would be ideal for opening the next Chinese restaurant. K-mean clustering algorithm will be used to cluster the neighborhoods. The final data after merging data sets described in the data section is as shown in the figure 8



**Figure 8: Final Data**

#### Determine Number of Clusters

The number of clusters is determined by running the k-means clustering algorithm with k ranging from 1 to 11 and choosing the optimum value using the metrics: Silhouette Coefficient, Calinski-Harabaz Index and Davies-Bouldin Index. These metrics are used to evaluate a model where ground truth labels are not available as in our case since, we don’t have neighborhoods that are pre labeled with rank of choice for opening next Chinese restaurant. For a detail discussion of the aforementioned metrics please refer to <https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation>.



**Figure 9: K-Means Cluster Evaluation To Determine Optimal Number of Clusters Value**

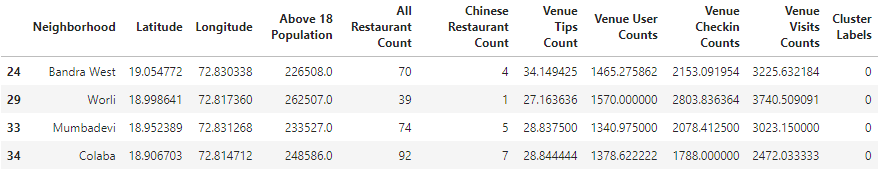
From figure 9, using the elbow method one can see that according to Silhouette Coefficient the optimum number of clusters is 4 and according to Calinski-Harabaz Index the optimum number of index is 4. As for Davies-Bouldin Index values closer to zero indicate a better partition and in our case that value is 0.335160 corresponding to number of clusters = 9. We will go with the majority and partition the neighborhood into 4 clusters

## Results

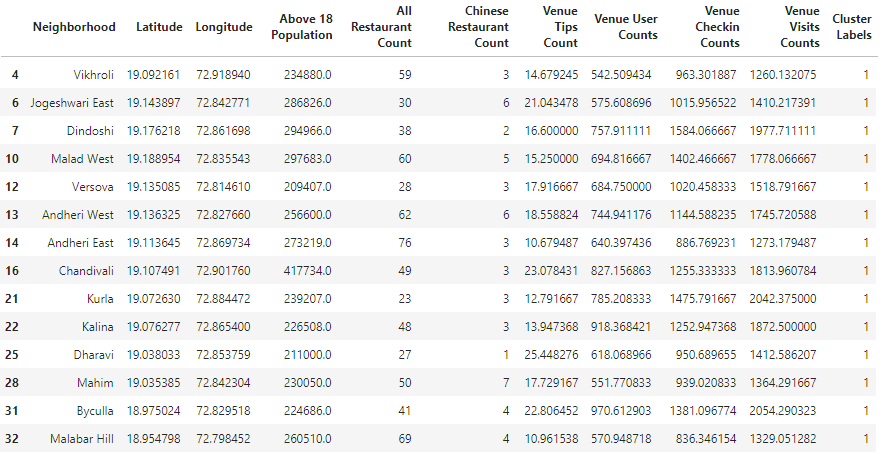
#### The Clusters of Neighborhood

Using Number of clusters = 4, the neighborhoods were partitioned into 4 clusters using the K-Means clustering algorithm. The 4 partitions are as follows:

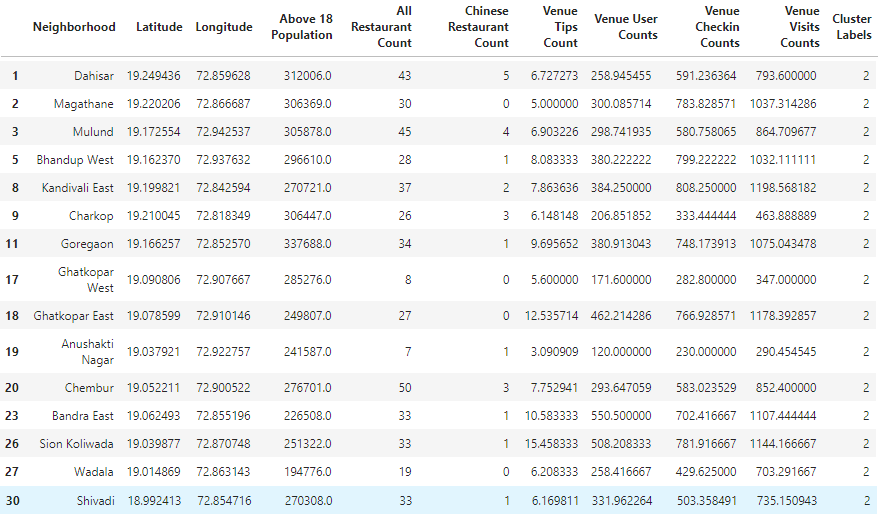
Cluster 1: This cluster contains all neighborhoods with high average count values signifying greater accessibility potential. It contains a neighborhood with high accessibility (average count) and only 1 Chinese restaurant i.e. Worli.



Cluster 2: This cluster contains all neighborhoods with one or more than 1 Chinese restaurants and moderate count values. Not a candidate cluster



Cluster 3: This cluster contains all neighborhoods with no Chinese restaurants and moderate average count values. This cluster is a possible candidate for picking the neighborhood for opening a Chinese restaurant.



Cluster 4: This cluster contains a single neighborhood with 3 Chinese restaurant and highest average counts. Not a candidate cluster



## Discussion

Data Exploration and Clustering both point to similar results for the next neighborhood for opening a new Chinese restaurant. These being

1. Ghatkopar East with no Chinese restaurant and good average counts (12.53 tip counts, 462.21 user counts, 766.89 check-in counts, 1178.35 visit count)
2. Bandra East with just 1 Chinese restaurant and good average counts (10.58 tip counts, 550.30 user counts, 702.05 check-in counts, 1107.08 visit count)
3. Worli, though not part of the same cluster as above two, but with only one Chinese restaurant and high average counts (27.14 tip counts, 1569.89 user counts, 2803.56 check-in counts, 3740.23 visit count) is also a good candidate.

As can be seen from figure 5 the Mumbai city has scope not only for Chinese but other authentic cuisine restaurants. An approach like this report can be taken for finding location for other authentic cuisine restaurants.

## Conclusion

In this article, to solve the problem of finding a neighborhood where a new business can be started, we have demonstrated a Machine Learning approach - right from formulating a business requirement(question), data identification, data collection, data exploration, creating a machine learning model and using the model to solve the business problem.

Mumbai neighborhood data has been collected from Wikipedia and foursquare.com, analyzed and location for Chinese restaurant has been recommended using K-means clustering. Though the data is limited (e.g. population data is about 18+ years only) the results are still good and usable. Three Potential neighborhoods namely **Ghatkopar East, Bandra East** and **Worli** were identified for opening a new Chinese restaurant.

Better data for parameters used for analysis in this report and data for other parameters mentioned in problem section would yield better results.

Mumbai city has good eating out crowd and hence has scope for other authentic cuisine restaurants. Approach like this report can be used for predicting other cuisine restaurants.