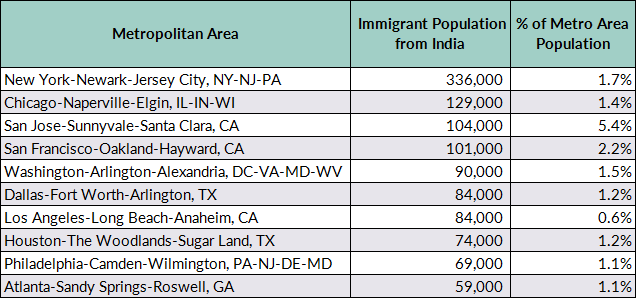
**ANALYZING NEW YORK CITY (NYC) NEIGHBORHOODS**

**TO OPEN AN INDIAN RESTAURANT**

**1.Introduction Section:**

The **City of New York**, is the [most populous city](https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population) in the [United States](https://en.wikipedia.org/wiki/United_States).  The five boroughs – [Brooklyn](https://en.wikipedia.org/wiki/Brooklyn), [Queens](https://en.wikipedia.org/wiki/Queens), [Manhattan](https://en.wikipedia.org/wiki/Manhattan), [The Bronx](https://en.wikipedia.org/wiki/The_Bronx), and [Staten Island](https://en.wikipedia.org/wiki/Staten_Island) – were consolidated into a single city in 1898. The city and its metropolitan area constitute the premier gateway for legal [immigration to the United States](https://en.wikipedia.org/wiki/Immigration_to_the_United_States). According to the [2010 United States Census](https://en.wikipedia.org/wiki/2010_United_States_Census), the Asian Indian population in the United States grew from almost 1,678,765 in 2000 (0.6% of U.S. population) to 2,843,391 in 2010 (0.9% of U.S. population), a [growth rate of 69.37%](https://en.wikipedia.org/wiki/Demographics_of_Asian_Americans).**Indians in the New York City metropolitan region** constitute one of the largest and fastest growing ethnicities in the [New York City metropolitan area](https://en.wikipedia.org/wiki/New_York_City_metropolitan_area) of the United States. The New York City region is home to the largest [Indian American population among metropolitan areas](https://en.wikipedia.org/wiki/Indian_American#Demographics) by a significant margin, enumerating 711,174 uniracial individuals by the 2013-2017 [U.S. Census](https://en.wikipedia.org/wiki/U.S._Census_Bureau) [American Community Survey](https://en.wikipedia.org/wiki/American_Community_Survey) estimates.

**Table 1. Top Concentrations by Metropolitan Area for Immigrants from India, 2011-15**



Source: MPI tabulation of data from the U.S. Census Bureau pooled 2011-15 ACS.

The above table information clearly shows that New York City has greater number of Indians than any other city in US, so it is a good idea to start the restaurant here. Of course, food and service are important to the success of a restaurant, but the location can be just as crucial. There need to be enough people who live in or pass through the area regularly to keep our business busy. This shows how important it is to start a restaurant in an area which is easily accessible and where you can attract your targeted audience easily. We analyse the neighborhoods in New York City to identify the most profitable area.

**2. Data Science Methodology:**

**2.1 Business Understanding:**

In this study, we are going to open an Indian Restaurant in New York, US. When looking for a restaurant location, consider who else is doing business in the neighborhood’s. Are there already half a dozen restaurants with the same concept as yours? Is the area busy or full of empty storefronts? Successful businesses attract other successful businesses.

**2.1.1 Target an audience:**

Why people would care about your problem or how your product can benefit them?

*No product without a customer. There is no film without a viewer.*

**This study will help and give more information to:**

* Indian peoples who wants to know about the neighborhood Indian Restaurants.
* Business personnel who wants to invest or open an Indian Restaurant in New York City.
* Data scientists who wants to analyse the neighborhoods of New York City.

**2.2. Data Requirements and Data Collection:**

i) New York City has a total of 5 boroughs and 306 neighborhood’s. In order to segment the neighborhood’s and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.Use https://cocl.us/new\_york\_dataset to get all the information about the neighborhoods present in New York City.

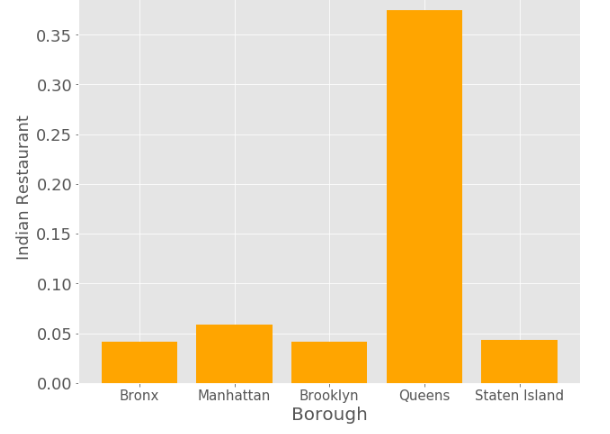
ii) To get the information about New York City: <https://en.wikipedia.org/wiki/New_York_City>

iii) To get the information about densely populated area with Indians: <https://en.wikipedia.org/wiki/Indians_in_the_New_York_City_metropolitan_region#Demographics>

iv) The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. To search for a specific type of venues, to explore a particular venue, to explore a Foursquare user, to explore a geographical location, and to get trending venues around a location in New York City we use (<https://developer.foursquare.com/docs)>

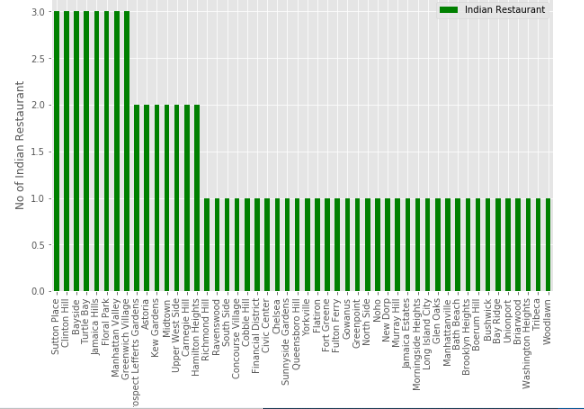
**2.3. Exploratory Data Analysis:**

 Exploratory data analysis (EDA) is an approach to [analysing](https://en.wikipedia.org/wiki/Data_analysis) [data sets](https://en.wikipedia.org/wiki/Data_set) to summarize their main characteristics, often with visual methods. A [statistical model](https://en.wikipedia.org/wiki/Statistical_model) can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Here I used one of the typical analysis technique called Bar Chart to visualize the relationship between 5 boroughs and neighborhoods of New York City(NYC) with respect to Indian Restaurant.



**Relationship between borough and Indian Restaurant**

The above table clearly shows that, among 5 boroughs Queens has the highest number of Indian Restaurant, and by the order Manhattan, Brooklyn, Staten Island, and Bronx. From this insights, we move on to the further analysis of neighborhoods.

****

**Relationship between Neighborhood and Indian Restaurant**

This statistical bar graph shows the number of Indian Restaurant with respect to the neighborhoods of the New York City (NYC) boroughs.

**2.4 Classification Model:**

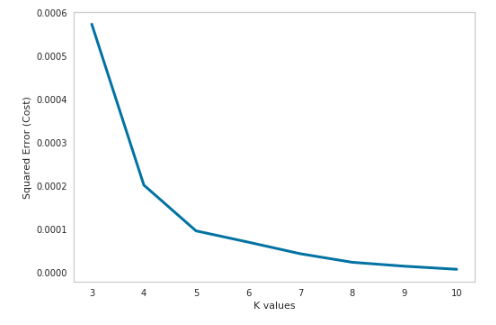
**2.4.1 Cluster Analysis:**

* It is a data mining technique used to place the data elements into their related groups.
* Clustering is the process of partitioning the data (or objects) into the same class, the data in one class is more similar to each other than to those in other cluster.
* The process of partitioning data objects into subclasses is called as cluster.
* A cluster consists of data object with high inter similarity and low intra similarity.
* The quality of cluster depends on the method used.
* Clustering is also called as data segmentation, because it partitions large data sets into groups according to their similarity.

Here, I used **k-means clustering** method of [vector quantization](https://en.wikipedia.org/wiki/Vector_quantization), that is popular for [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) in [data mining](https://en.wikipedia.org/wiki/Data_mining). *k*-means clustering aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) *n* observations into *k* clusters in which each observation belongs to the [cluster](https://en.wikipedia.org/wiki/Cluster_(statistics)) with the nearest [mean](https://en.wikipedia.org/wiki/Mean), serving as a prototype of the cluster. This results in a partitioning of the data space into [Voronoi cells](https://en.wikipedia.org/wiki/Voronoi_cell). *k*-Means minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult [Weber problem](https://en.wikipedia.org/wiki/Weber_problem): the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. Better Euclidean solutions can for example be found using [k-medians](https://en.wikipedia.org/wiki/K-medians_clustering) and [k-medoids](https://en.wikipedia.org/wiki/K-medoids).

The algorithm has a loose relationship to the [*k*-nearest neighbor classifier](https://en.wikipedia.org/wiki/K-nearest_neighbor), a popular [machine learning](https://en.wikipedia.org/wiki/Machine_learning) technique for classification that is often confused with *k*-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by *k*-means classifies new data into the existing clusters. This is known as [nearest centroid classifier](https://en.wikipedia.org/wiki/Nearest_centroid_classifier) or [Rocchio algorithm](https://en.wikipedia.org/wiki/Rocchio_algorithm).

The below graph shows the squared error cost with respect to the k values.



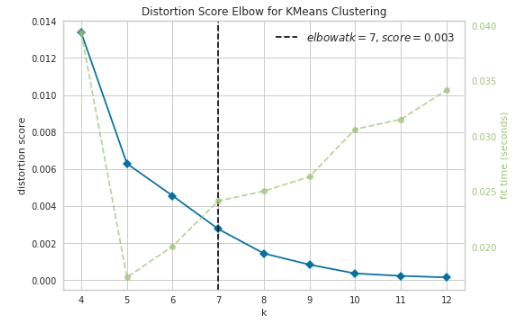
**Relationship between K values and Squared Error(cost)**

## **2.4.2 K Elbow Visualizer:**

The K-Elbow Visualizer implements the “elbow” method of selecting the optimal number of clusters for K-means clustering. K-means is a simple unsupervised machine learning algorithm that groups data into a specified number (k) of clusters. Because the user must specify in advance what k to choose, the algorithm is somewhat naive – it assigns all members to k clusters even if that is not the right k for the dataset.

The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the **distortion**(mean sum of squared distances to centers) score is computed, the sum of square distances from each point to its assigned center. Other metrics can also be used such as the **silhouette** score, the mean silhouette coefficient for all samples or the **calinski\_harabasz** score, which computes the ratio of dispersion between and within clusters.

When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the “elbow” (the point of inflection on the curve) is the best value of k. The “arm” can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.



From the above visualizer it is clear that **7** is the best value.

**2.4.3 Examine Cluster:**

In this we examine each cluster and determine the discriminating categories that distinguish each cluster. In this model we have 7 clusters from 0 to 1, each represent the 7 different categories that will help to end our process with insights.

## **3. Results:**

In the result section we can document all the findings from above clustering & visualization of the data’s. In this project, we started the business problem with identifying a good neighborhood to open a new Indian restaurant, we looked into all the neighborhoods in New York City, analysed the Indian population in each neighborhood & spread of Indian restaurants in those neighborhoods to come to conclusion about which neighborhood would be a better spot for opening a new Indian restaurant.

With the help of clusters examine & Bar chart looks like Queens, Manhattan, Brooklyn are already densely populated with Indian restaurants. So it is better idea to leave those boroughs out and consider only Bronx and Staten Island for the new Indian restaurant's location.

After careful consideration it is a good idea to open a new Indian restaurant in Bronx borough since it has high number of Indian population compared to Staten Island which gives a higher number of customers possibility and lower competition since very less Indian restaurants in the Bronx neighborhoods.

So, Bronx borough is a perfect place for starting a good and quality Indian Restaurant.

## **4. Future Improvement:**

i) The population density of Indians is based on the 2014 Americans Census. Thus population distribution or density would have changed by 2019.

ii) We made the analysis only based on the 5 boroughs population, i.e., it is not based on the neighborhoods density of Indian population. So, if we have that too definitely we will give much more exact report to open an Indian Restaurant.

iii) Since population distribution of Indian crowd in each neighborhood & number of Indian restaurants are the major feature in this analysis and it is not fully up-to date data, this analysis is definitely not far from being conclusory & it has lot of areas where it can be improved. However, it certainly provides us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly. Furthermore, this may also potentially vary depending on the type of clustering techniques that we use to examine the data.

## **5. Conclusion:**

i)This study will help and give more information to:

• Indian peoples who wants to know about the neighbourhood Indian Restaurants.

• Business personnel who wants to invest or open an Indian Restaurant in New York City.

ii) Similar to this analysis, we can choose any scenario such as analysing which venues are popularly visited by the people, their rating, and some of the other popular neighborhoods that surrounds one place etc.,

## **6. Appendix:**

i) Data’s used in this projects are scraped from the Wikipedia pages.

ii) Used many phython libraries such as pandas, numpy, random to fetch the data, to manipulate the contents.

iii) Used visualization libraries such as matplotlib, seaborn to visualize the data in various plot forms.

iv) Used folium library to visualize an interactive maps which helped to got exact analysis of the background data's.

v) Used Foursquare API to explore the venues in neighborhoods.

vi) Used geolocator to get the latitude and longitude of the given location.