IBM

Data Science Capstone Project

BUSINESS PROBLEM

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Indian Restaurant in New York, USA. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question : In New York, if an entrepreneur wants to open an Indian Restaurant, where should they consider opening it?

TARGET AUDIENCE

The entrepreneur who wants to find the location to open authentic Indian restaurant.

DATA

To solve this problem, we will need below data:

- List of neighbourhoods in New York, USA
- Latitude and Longitude of these neighbourhoods
- Indian American population in New York
- Venue data related to Indian restaurants. This will help us find the neighbourhoods that are more suitable to open an Indian Restaurant

EXTRACTING THE DATA

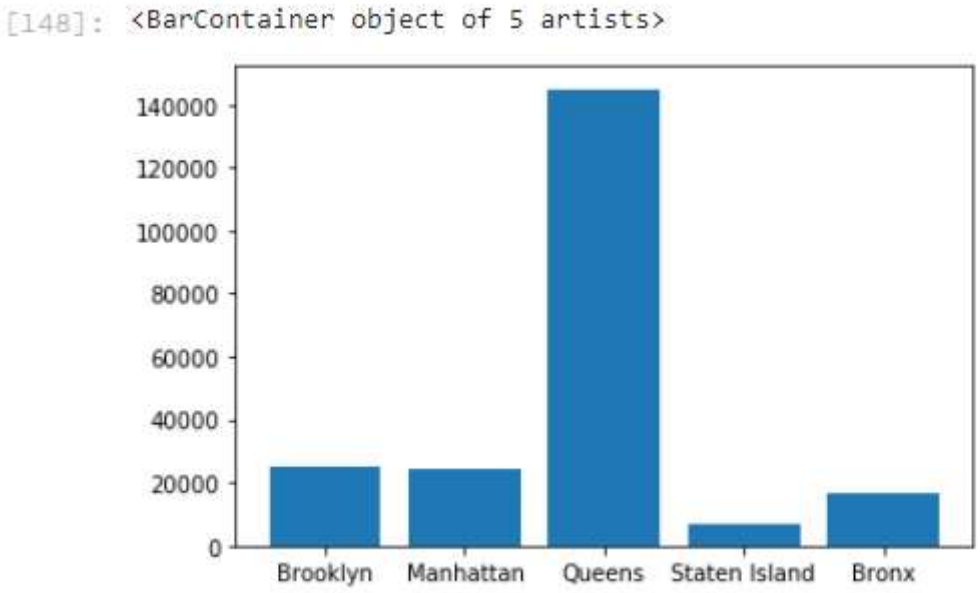
- Neighbourhoods data from Internet.
- Getting Latitude and Longitude data of these neighbourhoods via Geocoder package
- Using Foursquare API to get venue data related to these Neighbourhoods

METHODOLOGY

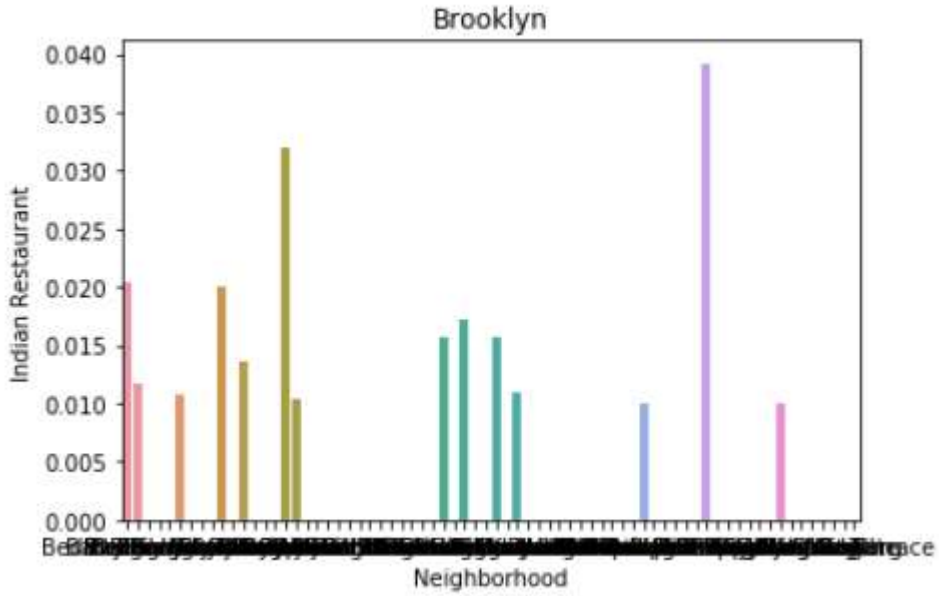
First, we need to get the list of neighbourhoods in New York, USA. The json data was converted into pandas data frame with attributes like Borough, Neighbourhood, Latitude and Longitude. Using Folium the data points were plotted. For each neighbourhood in New York data frame using Foursquare API list of top 100 venues within 500 meters radius collected and stored in another data frame. Then, grouping the rows by neighbourhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Then data frame was filtered to get mean values of “Indian restaurants”. Indian American population dataset was obtained from wikpedia And it is merged with above data frame.

K- means algorithm was applied to filtered data.K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. Filtered data frame was clustered into 3 clusters based on their frequency of occurrence for “Indian food”. Based on the results (the concentration of clusters), recommend the ideal location to open the restaurant was done.

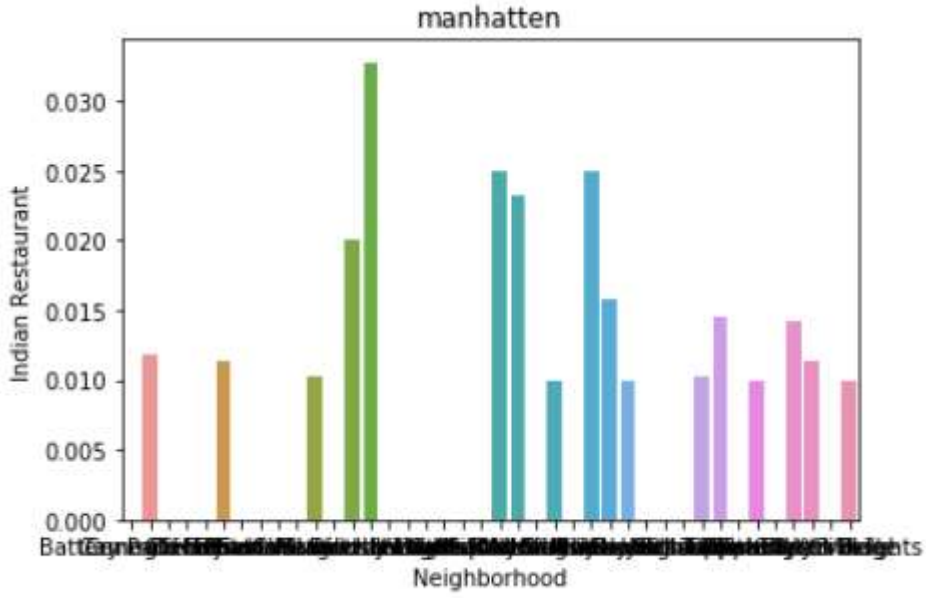
Plot of Borough and Indian American population



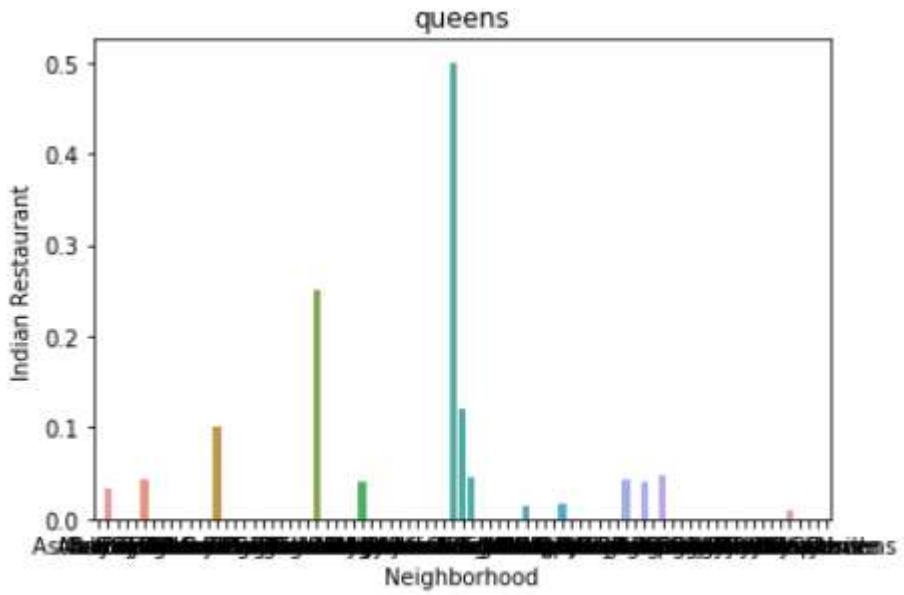
Brooklyn Neighbourhood



Manhatten Neighbourhood

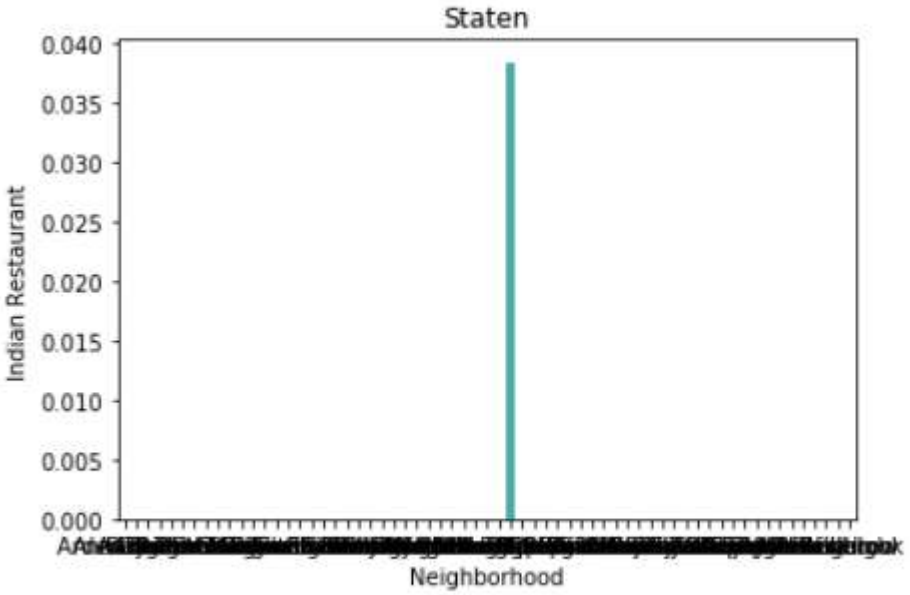


Queens Neighbourhood

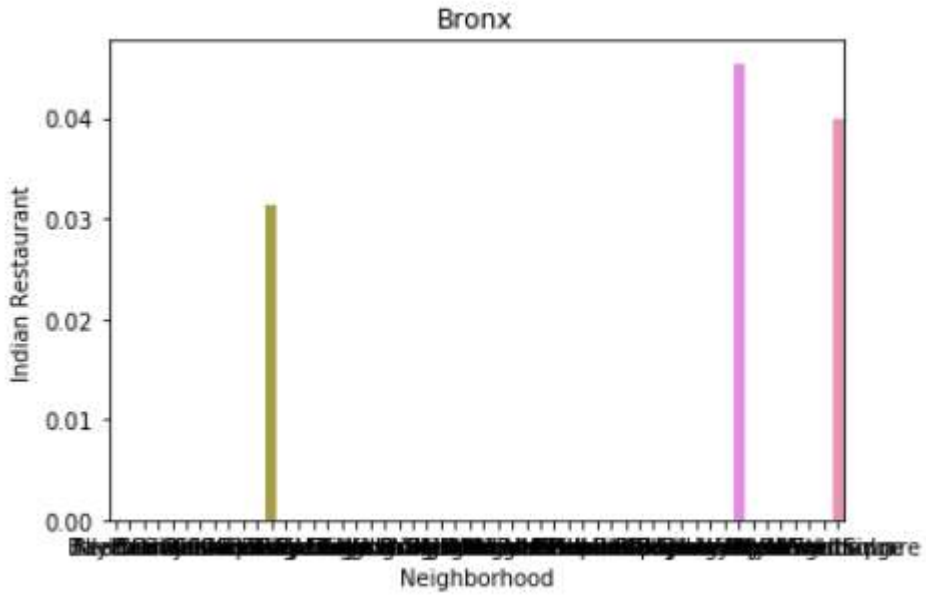


Staten Island Neighbourhood

```
]: Text(0.5, 1.0, 'Staten')
```



Bronx Neighbourhood



RESULT

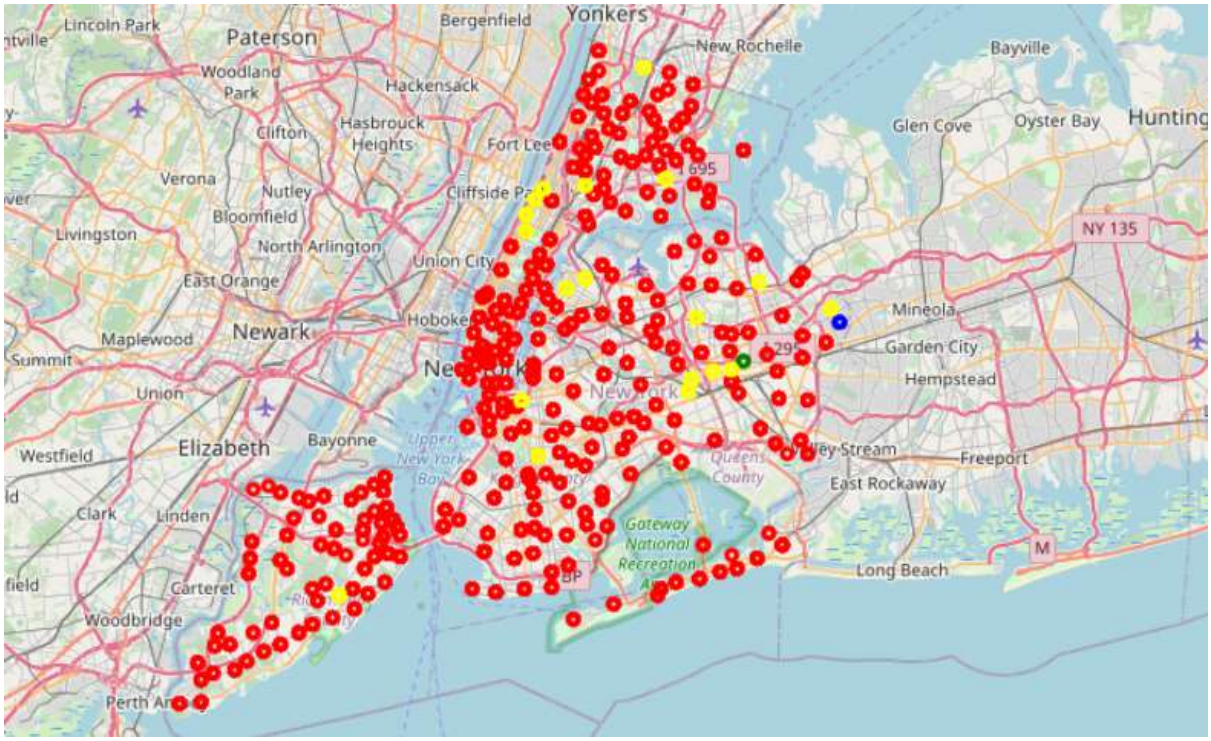


Fig: Clustered data points

The results from k-means clustering show that we can categorize New York neighbourhoods into 4 clusters based on how many Indian restaurants are

in each neighbourhood:

- Cluster 0: Neighbourhoods with More number of Indian restaurants.
- Cluster 1: Neighbourhoods with one Indian restaurants.
- Cluster 2: Neighbourhoods with one Indian restaurants
- Cluster 3: Neighbourhoods with average number of Indian restaurants but greater than cluster2

```
Total datapionts in clusters
Cluster-0: 281
Cluster-1: 1
Cluster-2: 1
Cluster-3: 19
```

Conclusion:

Above clustering was done by using only 100 venues nearby to each city, this number of venues can be increased to get more precise result. Also we can collect data about citizens origin and use it to get good result.