IBM APPLIED

DATA SCIENCE CAPSTONE PROJECT

On

“Clustering the Neighbourhoods of Moscow to identify potential areas to establish a new Japanese Restaurant Business using K-means Clustering Machine Learning Algorithm”

Submitted By :

Saket Misra

Shaheed Sukhdev College of Business Studies

University of Delhi

INDEX

Introduction

According to various articles and reports , In the recent times , several Foreign cuisines have found a way into the favoured taste of the Russians. Amongst top of those , being Japanese with particularly sushi being one of the top favourite dishes on the chart. This project aims to find out potential areas where a Japanese restaurant would be ideal to setup, so for it to flourish and make a name in the local restaurant business in Russia. The capital city, Moscow was chosen to start with , for the self explanatory and obvious reasons. The city of Moscow has 121 neighbourhoods to its name. A fairly large geographical span of around 2511 sq.kms and being the most populous city of Russia, with approximately 15.1 million residents within the city limits 17 million within the urban area and 20 million within the metropolitan area. Moscow is one of Russia's federal cities.Also it being the capital and the above factors , makes it the ideal place to start with , when identifying a good location for exploring business opportunities. This project analyses the various clusters of neighbourhoods of Moscow , and tries figuring out the ideal and a feasible cluster where a restaurant could be setup considering a major factor of competition of the other restaurants in the business.

Business Problem

The Business Problem pertaining to the project is given a Business group possibly an experienced restaurant chain owner or a business group , venturing into the restaurant business with being new to having a few years of experience in the same , which areas (Neighbourhoods) of Moscow could prove them to be good potential hotspots for establishing their restaurant setup , keeping in the mind the heavy competition of other similar or different businesses already established across the area , by recommending them the ideal cluster(s) in the city for them to explore and venture into for setting up their stand-alone restaurant and/or restaurant chain in the future.

Data

The data required for the purpose of the study in this project primarily revolves around the below mentioned particulars :

1. The data containing the names of neighbourhoods of Moscow , since our project is confined to the capital city of Moscow , where potential clusters have to be looked for setting up restaurant business
2. The Co-ordinates of the corresponding neighbourhoods , which acts as a facilitator for the future data requirements , as mentioned in the next point
3. The Venue specific data ie , the various venues , which lie in the proximity of the neighbourhood coordinates that is provided by the Coordinates data as mentioned above. The venue data includes basic information like venue name , category , coordinates ,tips , pictures etc
4. From the above data , only the data pertaining to our needs , i.e., restaurants is collected by cleaning the unnecessary information

The neighbourhood data is web scrapped from https://en.wikipedia.org/wiki/Category:Districts_of_Moscow by using the requests and beautiful soup

libraries of python. After gathering all the neighbourhood names , corresponding neighbourhood co-ordinates (ie latitudes and longitudes) are collected by making use of the geopy.geocoders library of python which helps in returning the coordinates of a location on the basis of an address passed as a string. Once we have the dataframe consisting the names of the neighbourhood along with the coordinates , venue data is gathered in the form of Json files , by making calls to the foursquare API , a location data provider which is used by more than 125000+ developers across the world , and contains a database of more than 105 million + locations.

Methodology

The first and the foremost step is gathering a list of neighbourhoods in Moscow , which was obtained from https://en.wikipedia.org/wiki/Category:Districts_of_Moscow . The following link contains a list of 120 neighbourhoods of Moscow. The information is obtained by making use of the requests and beautiful soup libraries of python. Once all the neighbourhood names are collected , we find out the corresponding latitude and longitudes of those neighbourhoods , we make use of the geopy.geocoders library for the same , where we pass the neighbourhood name as an address to the library methods , and get the corresponding coordinates after iterating over all the 120 neighbourhoods. The latitudes and longitudes gained thus help us in our next step , and that is to extract relevant venue information within the close proximity of the neighbourhood coordinates we just collected. This is done with the help of making iterative calls to the foursquare API , a location data provider that takes in a url , in the form of a string which contains various parameters such as the client secret , ID , version and other information depending upon the requirements. In our case we need the venue information , and hence we are particularly interested in the venue data , and pass the neighbourhood coordinates , our credentials ,

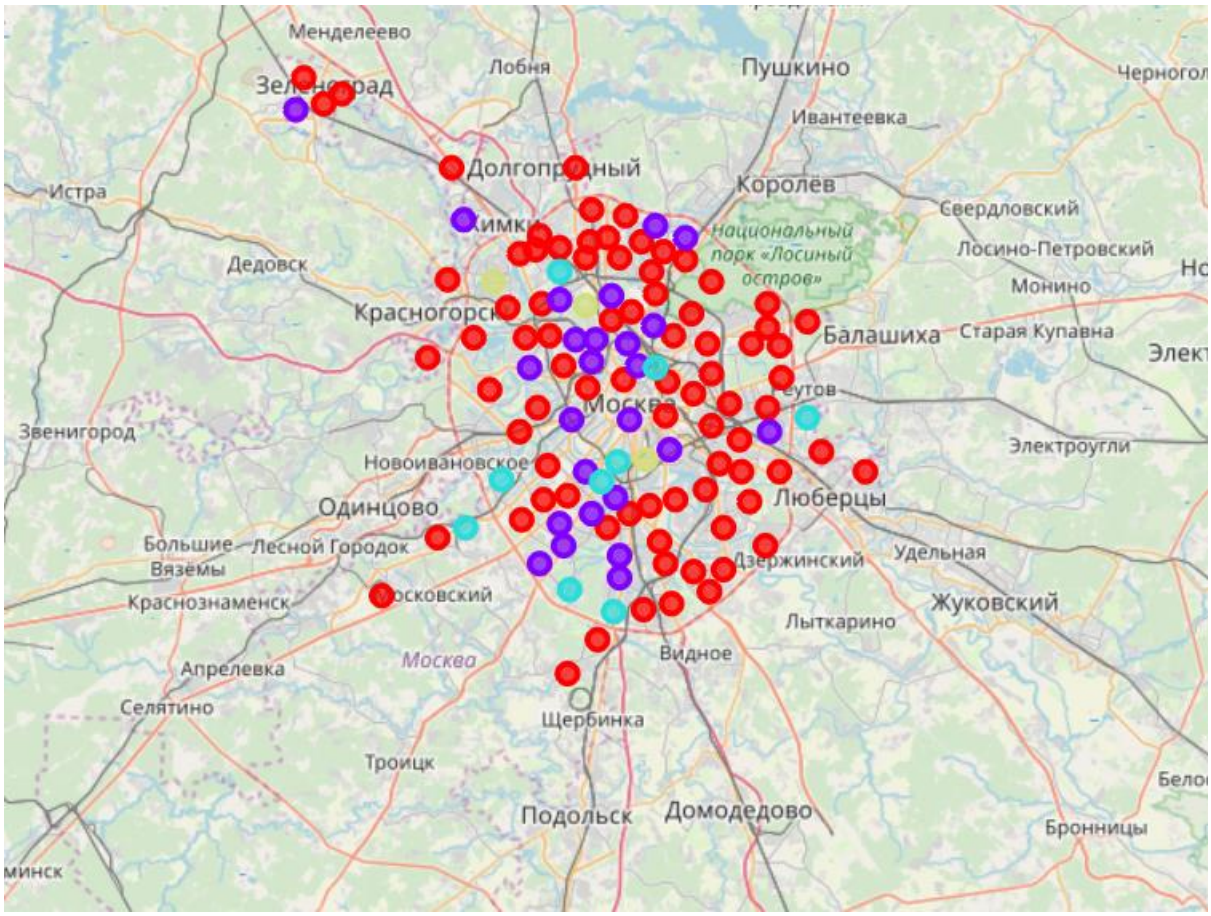
under the ‘explore’ functionality of the foursquare api functionality , some more important parameters like radius where radius specifies the geographical span of the venue data , and limit that confines to a restricted number of venue data we receive. The venue data we received is in the form of a json file that contains information like venue category , name , pictures , tips , venue latitudes and longitudes. We are interested into venue category , and in particular one category that is Japanese restaurants , but before we modify our data into our needs , we first make use of the folium library of python to generate a map of Moscow , and superimpose those neighbourhoods on the map , afterall we have our neighbourhoods coordinates handy. After this , we group our data according to the neighbourhoods , ie the neighbourhoods are indexes to the various venues (Japanese Restaurants) in our case , we apply one hot encoding to the dataframe , so as to be able to work with categorical data (venue category in this case) , and then get the aggreagate mean which lies between 0 and 1 for each category , after this , we extract the column that only contains the aggregate means for the category ‘Japanese Restaurant’ , and so we have our data ready in the form of two columns , one providing us the neighbourhood name , and other the corresponding aggregate mean for Japanese

restaurant. This data is then used to assign clusters to each data point in our data frame , by making use of the k-means clustering machine learning algorithm. The k means is an unsupervised machine learning algorithm , meaning we deal with unlabelled data , which is true in our case. In the k-means clustering algorithm we first assign a k value , which actually corresponds to k different centroids spread across our data points , if visualized on a scatter plot . These k centroids are initially placed in a random fashion , and then their dissimilarity with each corresponding data point is calculated , the dissimilarity (vice versa lesser similarity) can be calculated using various methods like Euclidean distance , cosine similarity etc. Then every data point is assigned a cluster depending upon the minimum dissimilarity between the data points and the centroids. Here every centroid actually is a representative of a cluster , and at the end of the cycle , it’s the dissimilarity between the centroid and data points that determines the membership of a data point to the cluster. Once an iteration completes its time to make this better ie Kmeans is an iterative mechanism , and takes a certain number of iterations to get the optimum results . The centroids are relocated to form better clusters with starting over the process, The relocated location of the centroid is actually the means

of the corresponding data points that were allocated to that particular cluster in the previous iteration , and hence the name “K means” the above process is very simple to achieve in python with the help of kmeans library , where we just need to specify the number of clusters , our dataset on which we fit our model and the iterations that we want to make in order to get the best clusters. In return we get a list of labels to our data points , in our case we decided to go with 4 clusters , and hence we received the data points along with labels ranging from 0 to 3. Now before we are ready to visualise our results , we just need to plug in the original information of latitude and longitude of each neighbourhood in order to generate the map , with different cluster labels being assigned different colors for the circle markers to be able to better visualise the clusters. After this the number of cluster labels , assigned to each cluster was calculated , and conclusions were made on the same.

Results

There were 4 clusters , Cluster 0 , 1 , 2 and 3 observed having different magnitudes in terms of labels , and geographical span , as can be seen in the below map , generated using the folium library of python



As it can be clearly seen that the red cluster dominates the rest of the clusters , and covers the boundary neighbourhoods while the purple cluster in localised to central Moscow , the yellow and blue cluster forms a part of the outskirts of the city , and cover the least number of neighbourhoods

Cluster 0 and Cluster 3 (yellow and blue) cover 6 and 17 neighbourhoods respectively which cover the least number of neighbourhoods , while cluster 1 and 3 predominate the neighbourhoods of Moscow.

Discussions

As the map is evident , and also when calculated quantitatively cluster 0 has least number of neighbourhoods assigned to it , hence an argument can be made here that cluster 0 has little to no competition in terms of setting up a Japanese restaurant , as opposed to other clusters that are heavily dense , This can also be viewed that the central parts of the cities may have established restaurant outlets , leaving no room and /or rising property rates at the central Moscow , pushing the other businessmen to look out for potential in the outskirts of the city. Here cluster 3 also provides a potential answer to our business problem , as cluster 3 can provide the right trade off between opting for a heavily densed cluster with huge population but cut-throat competition and completely outskirts with minimal to no competition but with a lesser population attraction.

Limitations and future scope

In our project ,we remained restricted to one factor that is the frequency of Japanese restaurants in neighbourhoods , to determine the best cluster for setting up a Japanese restaurant , but there could have been several other potential factors ,such as interests of the people , tips that they give regarding the restaurants , maybe specific locations of Moscow having more interest in Japanese food outlets , or a geographical span that could be more pleasing to setup the restaurant than the others. These factors can be looked upon and can be suitable extensions to this project that can act as an indicator to provide even more refined solution to our business problem.

Conclusions

After closely analysing the data , we recommend cluster 0 and 3 to be suitable prospects for setting up a Japanese restaurant given that they offer minimal competition and provide a viable environment for a new restaurant business to flourish in the corresponding labelled neighbourhoods

The other clusters have high concentration of Japanese restaurants and hence , would pose difficulties in terms of expansion and profitability of setting up the

restaurant businesses on the neighbourhoods corresponding to those clusters.

References :

- Foursquare Developer documentation
  https://developer.foursquare.com/docs
- Beautiful soup library documentation
  https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- List of neighbourhoods of Moscow
  https://en.wikipedia.org/wiki/Category:Districts_of_Moscow
- Foreign Cuisine interests of Russian people
  https://healthyhappyhelping.wordpress.com/2015/09/05/7-foreign-foods-russians-just-love/