Location Determination for Opening a Restaurant in Toronto

Amit Kumar

December 25, 2020

1. Introduction

1.1 Background

Toronto is the capital city of the Canadian province of Ontario. It is the most populous city in Canada and the fourth most populous city in North America. The city is the anchor of the Golden Horseshoe, surrounding the western end of Lake Ontario, while the Greater Toronto Area proper had a 2016 population of 6,417,516. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. There are people who has a different interest in destination like park, restaurant and other time spending destination.

1.2 Problem

Data that might provide the restaurant in Toronto, but we can determine the people interest in that. The business person need to open a restaurant in Toronto. We need to find the best location for opening the restaurant and will provide valid explanation that will eventually support the stakeholders to get the clear understanding where to invest.

1.3 Interest

Stakeholders are interested to open restaurant in Toronto city where his/her business will grow.

2. Data acquisition and cleaning

2.1 Data sources

We are using Wikipedia page to scrap the Postal Code, Neighborhood and Borough. We have the Geojson.json file which will we use to append location data like latitude and longitude. We are also using FOURSQUARE API to fetch the nearby location of the restaurant. Based on that we will try to look into the people interest which is going to actual help stakeholders to do the decision.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing and inappropriate values from Wikipedia. We remove those anomalies in our data frame.

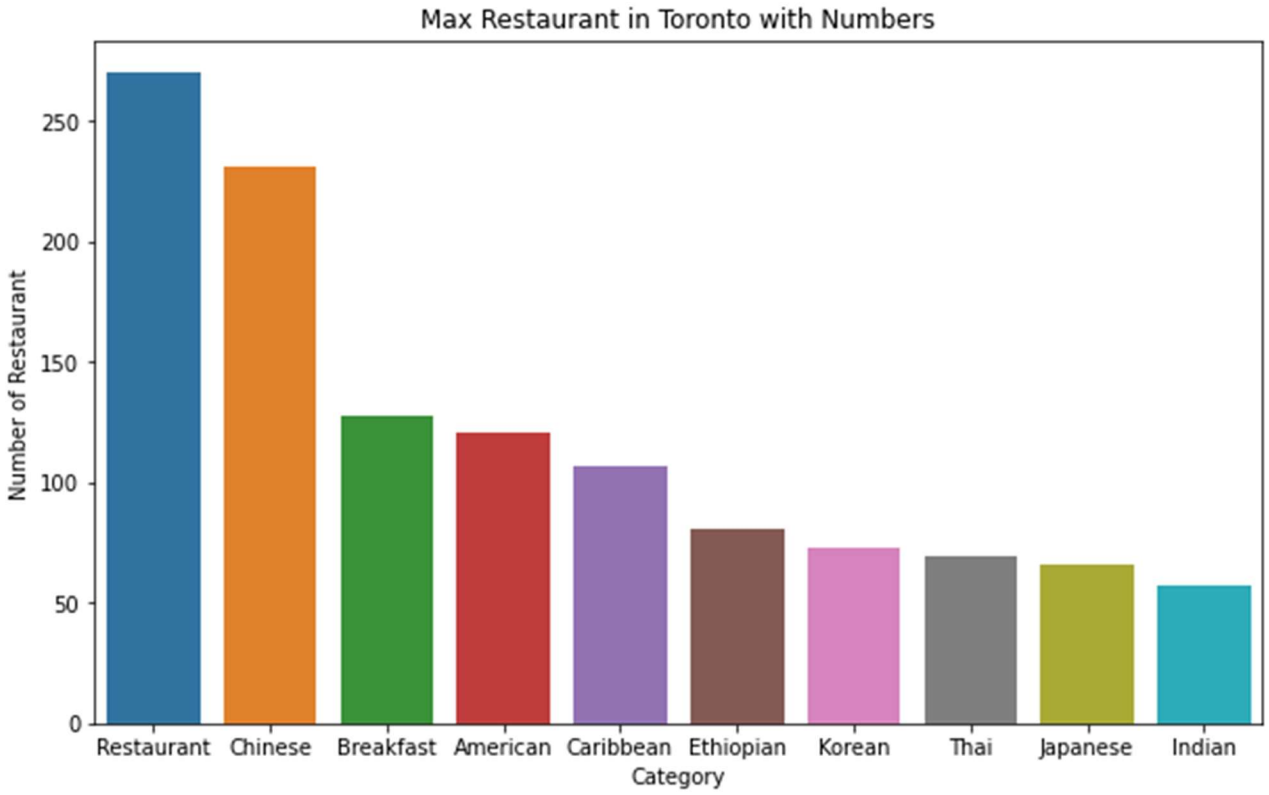
2.3 Feature selection/Engineering

After data cleaning, we use the geojson file to add location data(latitude and Longitude) . Upon adding the location data we are going to fetch nearby destination for every Borugh and will append in the dataframe.

3. Exploratory Data Analysis

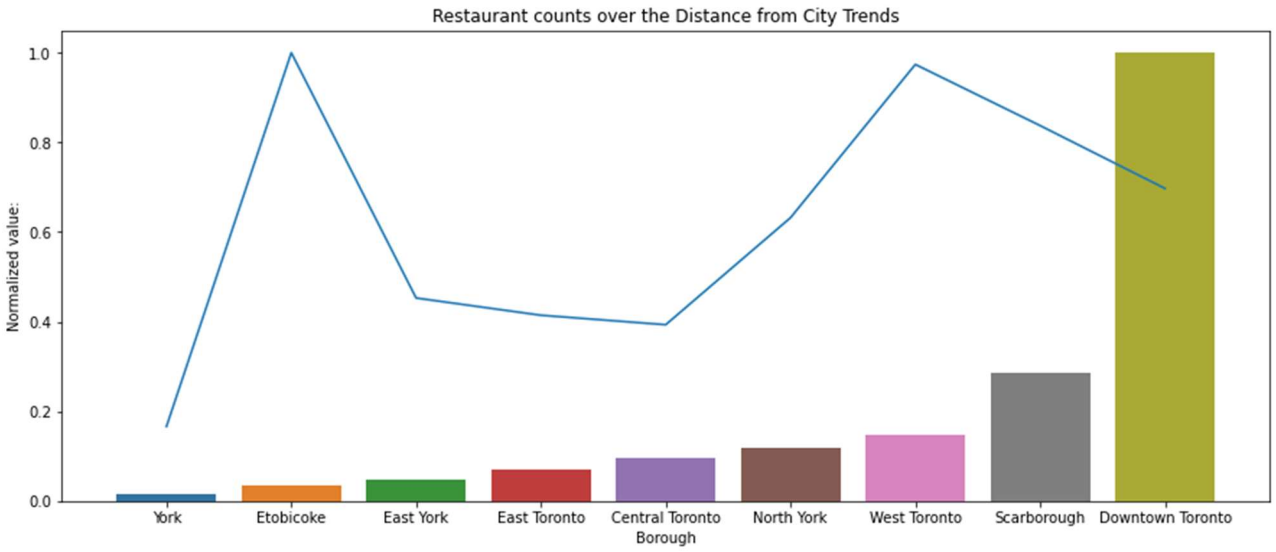
3.1 Relationship over the most typed Restaurant

After plotting the graph between the borough we came to know that general Restaurnt, Chinese and Breakfast item dining, and other category has more effect in Toronto.



3.2 Relationship between Distance and Restaurant Count

It is seen that the Location nearby city center has more no of restaurant but we also came to know that even with low distance with the city center, there are some places which has a less no of restaurant and those location can be the choice for the stakeholders for opening the restaurant. Let see the relationship between the distance from the city center and the count of restaurant in the available Borough, Before the at we have normalize both the data so that we could able to view the data on line graph which will not cause any misunderstanding.



3.3 Relationship between People interest

We have captured those cities where's people interest is medium like people which are average in terms of going to restaurant place. Out of rate of interest from 1-10, we capture those interest which belong to 4-10.

4. Deciding Clustering

There are too many clustering algorithms available that can be used to cluster the location but we have used KMeans for clustering the location

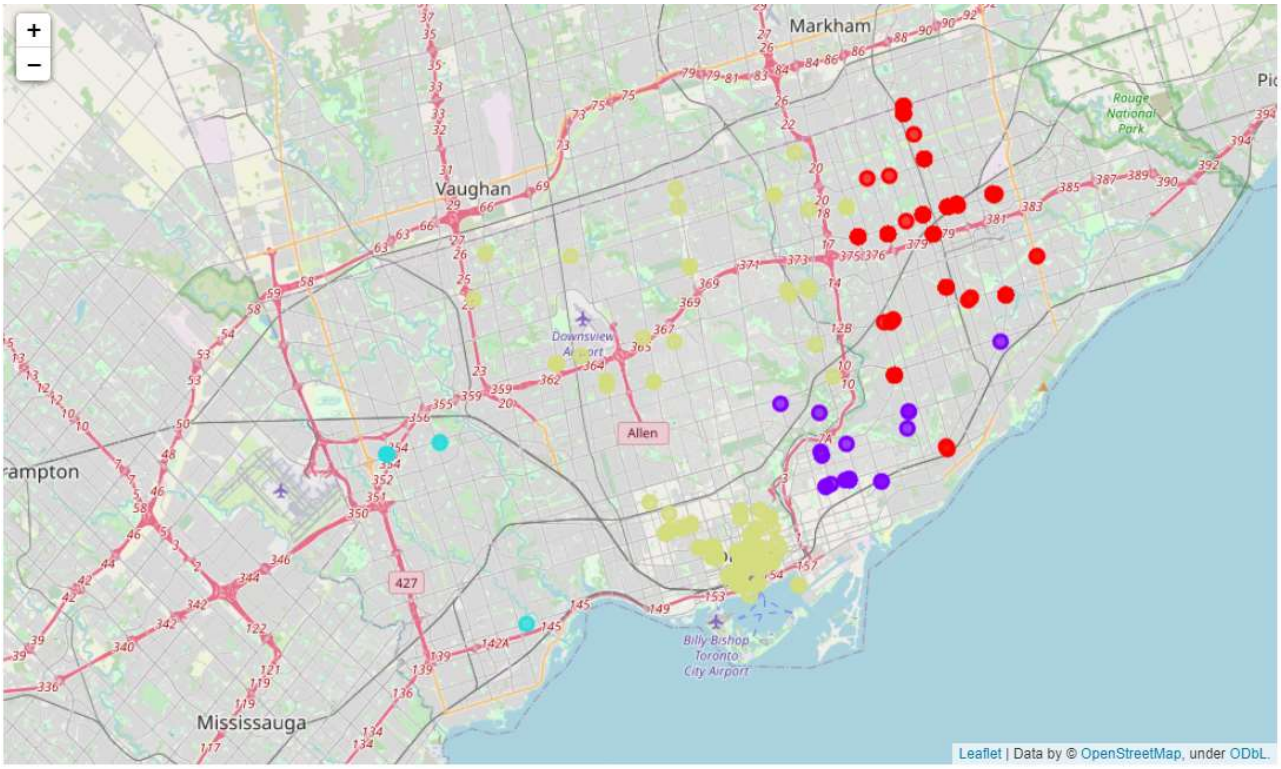
4.1 KMeans

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

4.2 Solution to the problems

We are clustering based on the Borough, after understanding the people interest that will provide the stakeholders the appropriate details from which they will able to take any decision.

Our analysis shows that although there is a great number of restaurants in Downtown Toronto, there are pockets of low restaurant density fairly far to city center. Highest concentration of restaurants was detected at city center, so we focused our attention to areas which are far away from city center. Another borough was identified as potentially interesting (East York, Scarborough), these are also good in context of people interest and less in counts in the area



5. Conclusions

Purpose of this project was to identify best areas with low number of restaurants and high people interest in Restaurant. By calculating restaurant density distribution from Foursquare data we have first identified general boroughs that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decission on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.