RESTAURANT LOCATION RECOMMENDER SYSTEM

(Clustering based approach)

Vijayashankaran R L

May 18, 2020

1. INTRODUCTION

1.1. Background

There are many entrepreneurs out there looking to start a new restaurant or expand their growing business. The first and most important question that pops up in their minds "where do i start a restaurant?" or "which is most suitable neighborhood thats going to profit the business?". While choosing a much less or sparsely populated Neighborhood and much less visited Neighborhood, it might save you on rent, but that is not going to result in profits. Only setting up your restaurant in the crowded and densely populated neighborhood is going to popularize your name through the ears of the city.

1.2. Problem

In simple words **"What is the best location to start a restaurant business in Toronto City, Canada?"**

Have you ever decided to start a restaurant, only to discover that the there is no enough customers to keep the business running with profits? Consider all the important,related and required factor and determine the perfect location to start a resturant in the city of Toronto.

2. DATA ACQUISITION AND CLEANING

2.1. Data Sources

The data required would be

- List of Boroughs and Neighborhoods
- Latitudes and Longitudes of those
- Population
- Population density
- Area size of the neighborhood
- Other venues located nearby to determine the popularity of the neighborhood

These data will be web scrapped from the below links:

- https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods
- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

All the data about the venues, venues category, latitudes and longitudes of the venues are obtained from the Four-sqaure API in the form of **json** data

2.2. Data cleaning

For simplicity of the project, columns of data such as neighborhood maps, Renters, Common Language, Transit commuting, and change in population have been dropped. Population, Area, Density of the neighborhoods are the main columns considered in the data.

Only the venue, venue category, neighborhood and it's cor-ordinates have been considered for the sake of simplicity. Other irrelevant data have been dropped.

Hence final required columns in our dataset would be:

- PostalCode
- Borough
- Neighborhood
- Area
- Population
- Density
- Latitude
- Longitude
- Venue
- Venue Category

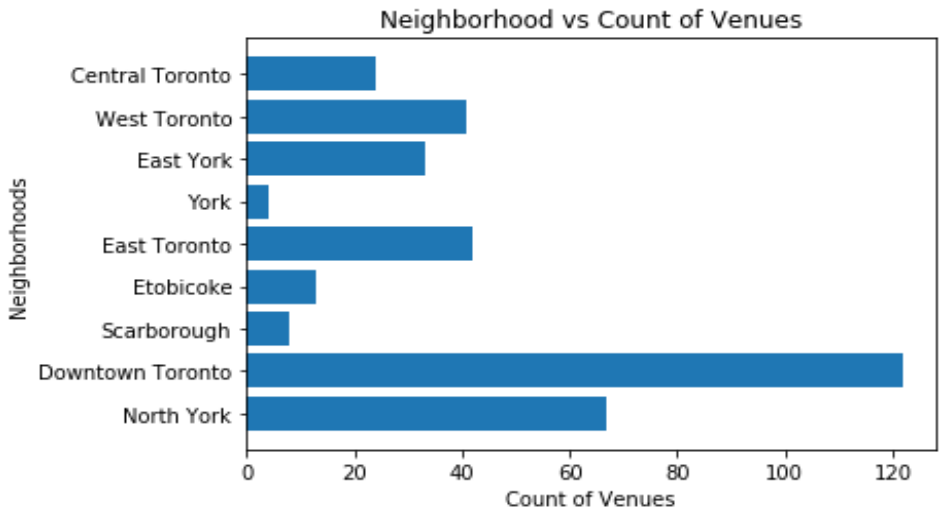
	PostalCode	Borough	Neighborhood	Area	Population	Density	Latitude	Longitude	Venue	Venue Category	Venue Count
0	M3A	North York	Parkwoods	4.960366	26533	5349	43.753259	-79.329656	Variety Store, Brookbanks Park	Food & Drink Shop, Park	2
1	M4A	North York	Victoria Village	4.719546	17047	3612	43.725882	-79.315572	Pizza Nova, Portugril, Tim Hortons, Victoria V...	Pizza Place, Portuguese Restaurant, Coffee Sho...	4
2	M5A	Downtown Toronto	Regent Park, Harbourfront	2.196988	24755	11267	43.654260	-79.360636	GW General, FUEL+, Residence & Conference Cent...	Antique Shop, Coffee Shop, Hotel, Beer Store, ...	45
3	M6A	North York	Lawrence Manor, Lawrence Heights	5.339568	17519	3280	43.718518	-79.464763	Party City, Fairweather, International Clothie...	Miscellaneous Shop, Women's Store, Clothing St...	12
4	M1B	Scarborough	Malvern, Rouge	37.587676	67048	1783	43.806686	-79.194353	Interprovincial Group, Wendy's	Print Shop, Fast Food Restaurant	2

3. EXPLORATORY DATA ANALYSIS

3.1. Calcution of count of venues

The number of venues in a neighborhood can determine the popularity of the neighborhood. That being said, we count from the total number of venues in all the categories in each and every neighborhood since that will determine which neighborhood is the most visited. Thus, logically and simply it may be well suited for our goal, to determine the best suited neighborhood.

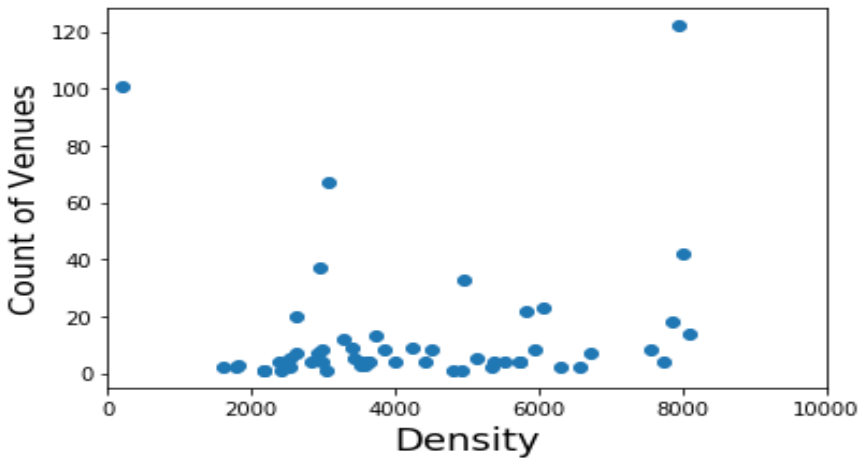
3.2. Plotting neighborhoods and count of venues

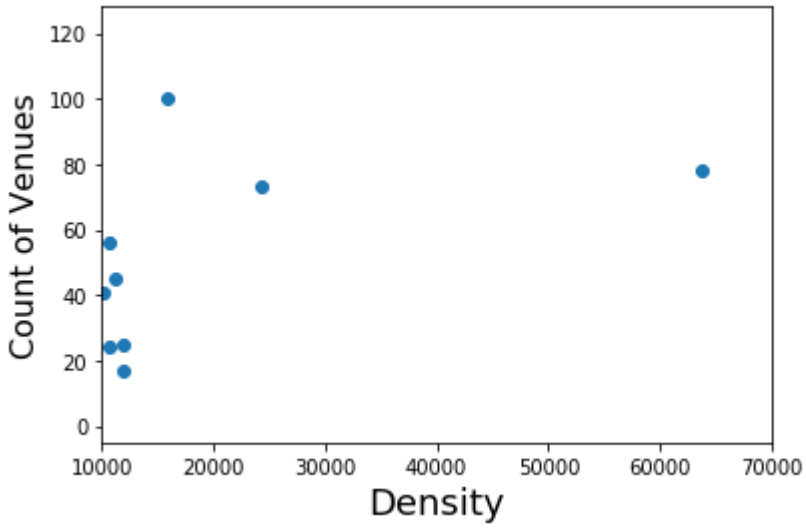


4. MODELLING AND METHODOLOGY

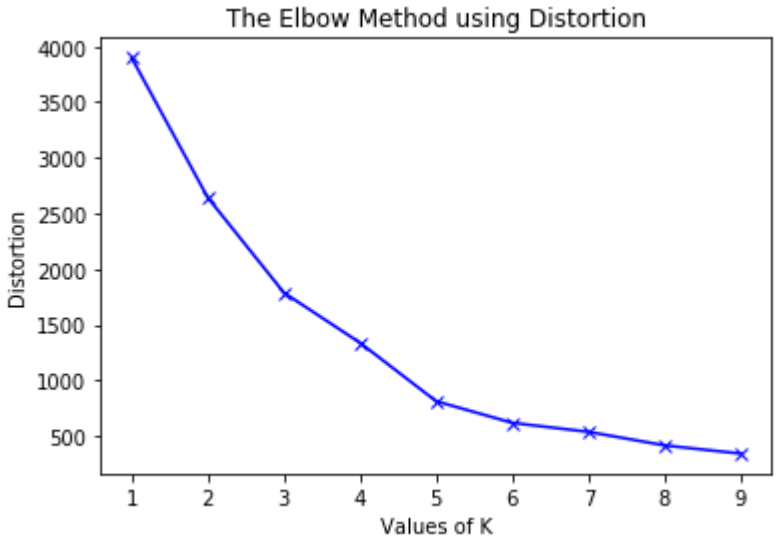
CLUSTERING MODEL

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimila r to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.We choose the density and the number of venues to be main features for our model and use them as input data to the clustering model.





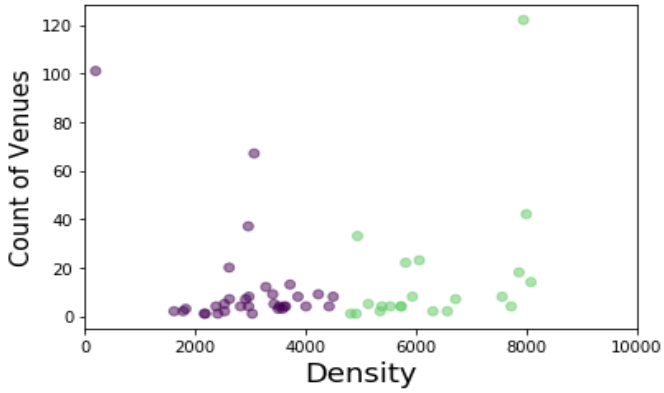
Finding the best K – value (Elbow Method):

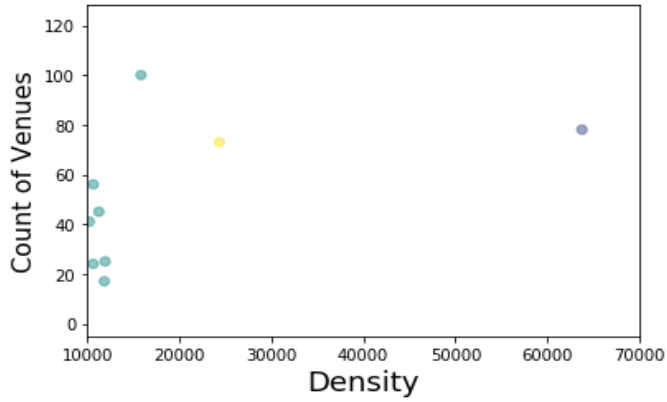


The best value of k is found to be 5.

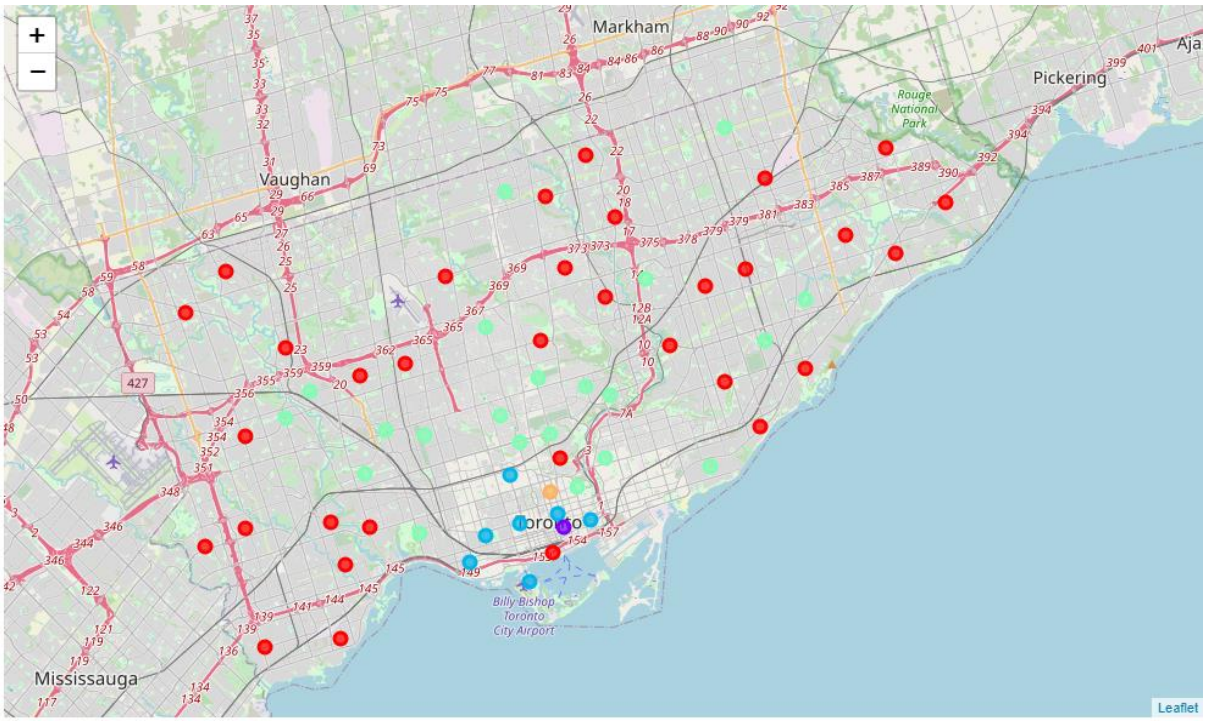
5. RESULTS

After the analysis it was found that the factors are highly co-related with the density and the data was clustered based mainly upon the density feature.





Final plotting of clustered neighborhood on map using folium



- It is found that the most preferred neighborhood is "**St. James Town**" (Violet marker)
- The next most preferred neighborhood is "**Church and Wellesley**" (Yellow marker)
- Followed by blue and green marked locations.
- The Worst neighborhoods are marked as red circles.

6. DISCUSSION SECTION

This project is fairly not very accurate. To increase the accuracy we need more features like Income of the people living in the neighborhood, competetion of restaurants in the neighborhoods (which can be a double edged sword), preferred cuisines and much more. I shall try updating this project with all of them included, but finding the datasets are not possible for now.

7. CONCLUSION

Thus, the most preferred neighborhood to launch a restaurant would be "St. James Town" and "Church and Wellesley". With much more liquid data about frequency of visits we may be able to determine the most accurate results.