Clustering Toronto’s neighborhoods to find out where to open a restaurant.

Final Assignment

Korbinian Nitz

Table of Contents

1 Introduction

Toronto is one of the most famous and trendy cities in Canada. Within almost 3 Million people living in the city center there is also a big chance of having a good running restaurant. But the question is where should I open a restaurant? Where do I know which area is trendy, which has already a lot of restaurants and which areas are more wealthy then others? Yes, maybe the Toronto people know all those things but for a outsider who never been to Toronto, it can be really challenging to find the best neighborhood for your restaurant.

This guide should help people to find the right decision by clustering all neighborhoods based on their similarities into 5 cluster.

2 Data Aquisition And Cleaning

Most data we need for Toronto, e.g. the income, population and population change in each neighborhood, can be found on https://open.toronto.ca/. This portal give access to open source datasets regarding Toronto and related topics. In Addition, a geojson data is also found on the website in order to get all latitude and longitude coordinates of each neighborhood. Lastly, we use the Foursquare API to find the amount of restaurants in each neighborhood / cluster.

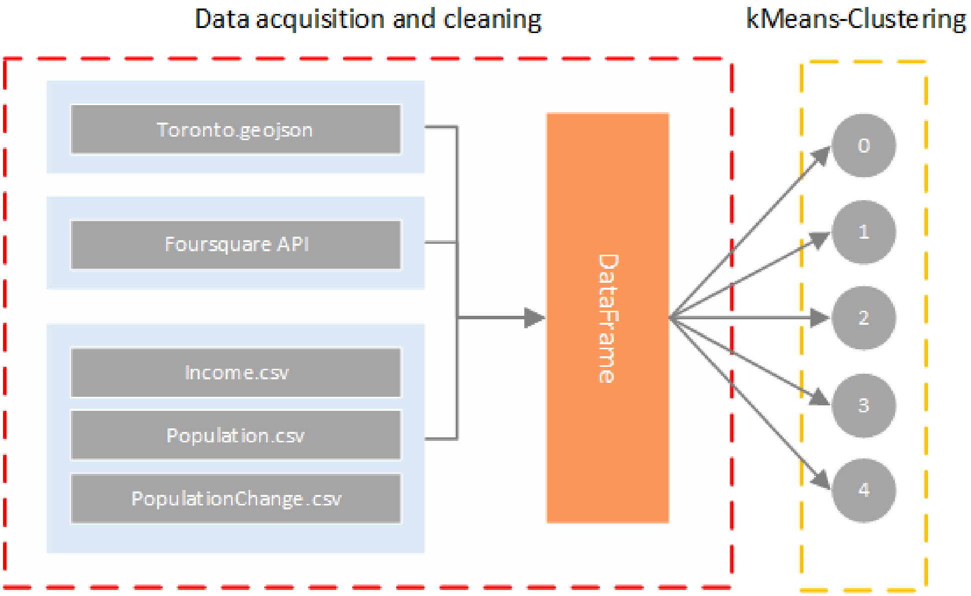

Figure 1: Methodology of the guide.

In figure 1, you can observe the methodology of the concept. First we have do the data acquisition and cleaning which means that we are going to upload all needed inputs into our notebook. After that we will drop all unnecessary columns of each dataset and merge them to a final dataframe which then is be used to cluster the neighborhoods into 5 clusters based on the kMeans cluster algorithm.

First the csv-datasets are cleaned and merged together. The result is shown in table 1 as an example for the first 4 rows.

Table 1: merged csv-datasets / first 4 rows.

Neighborhood	Young [%]	Middle Age [%]	Old [%]	Total Population	Change in Population [%]	Income / Household [$]	Restaurants
West Humber-Clarville	45.86	33.55	20.59	32880	-2.32	63977	2
Mount Olive-Silverstone-Jamestown	51.86	32.80	15.38	33090	0.51	49601	14
Thistletown-Beaumond Heights	43.14	33.41	23.45	10235	2.2	54910	19
Rexadale-Kipling	41.30	35.60	23.09	10350	0.39	53779	4

As you can see there are already differences in each neighborhood which will be explorated later in the assignment.

In order to connect the Neighborhoods with the Foursquare API, the neighborhood has to get their latitude and longitude coordinates. So then we can look up all restaurants in each neighborhood within a radius of 500 by a limit of 100 restaurants. The limit is necessary to reduce the computerial work. The latitude and longitude are uploaded for each neighborhood through the geojson data.

3 Exploratory Data Analysis

First it is interesting to see the demography of Toronto. Therefore we use the binning method to get three category 'young', 'middle aged' and 'old' as we can see it in the table 1. Table 2 shows the conditions which are used to bin the groups into three categories. Young people are defined as people who are younger than 30 years. Middle age people are between 30 and 60 years old and old people are older than 60 years. Within this binning we want to find out if a neighborhood is more likely an area where more younger or older people lives which is interesting for the clusters later.

Table 2: Condition for binning the age groups.

Category	Condition
Young	$< 30\ years$
Middle Age	$30\ years \leq x\ < 60\ years$
Old	$x \geq 60\ years$

If you explore the data firstly for whole Toronto, we can see that most of the neighborhoods have a young population as you can see in figure 2. More than 40 percent of the Toronto population are grouped into the young category, 37 % are middle aged and 23 % are old people.

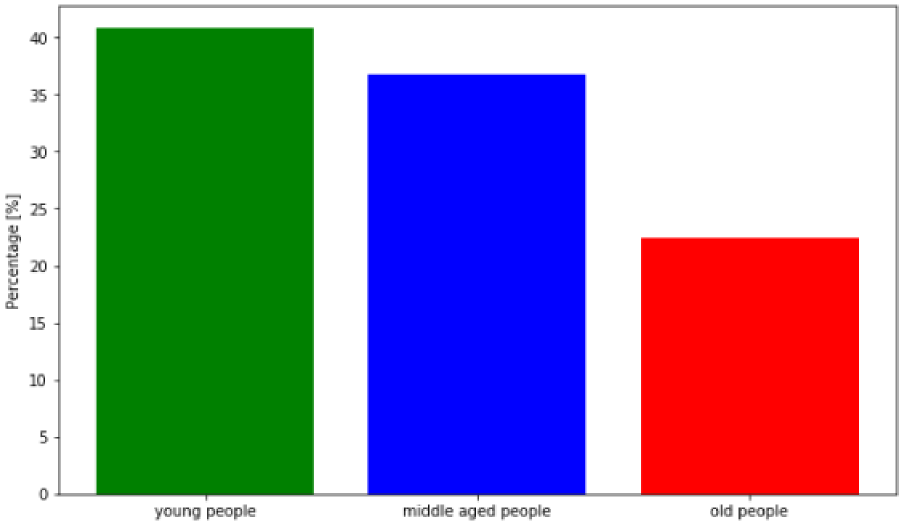

Figure 2: Demography of Toronto.

It shows that Toronto has a population of 80 % which are younger than 60. That is a good indicator for opening a restaurant because mostly younger and middle aged people prefers to go to a restaurant. Of course older people do to but are maybe not as open for new restaurants as younger people.

The next step is to analysis the neighborhoods in more detail. Therefore we create again 3 bins to find out if the neighborhoods in Toronto are trendy or not and if the neighborhoods are mostly rich regarding the Income per household after taxes.

First we want to find out the income of each neighborhood. For this binning we use following condition as shown in table 3. To determine the borders we use lower and upper quartile of the income values.

Table 3: Conditions for income binning.

Category	Condition
Poor	< 48921 \$
Avg	48921 \$ $\leq x < 64768$ \$
Rich	$x \geq 64768$ \$

We can see that poor people are defined as people who has an income per household below 48921 $ per year. In comparison, rich people earn more than 64768 $ per year and average people between those values. In figure 3, we can see the results of analysing the neighborhoods of Toronto. Mostly average people are living in the neighborhoods. Also most in 70 neighborhoods the income of the people are average. That's the half of all neighborhoods. The rest are poor or rich neighborhoods in count of the income of the people.

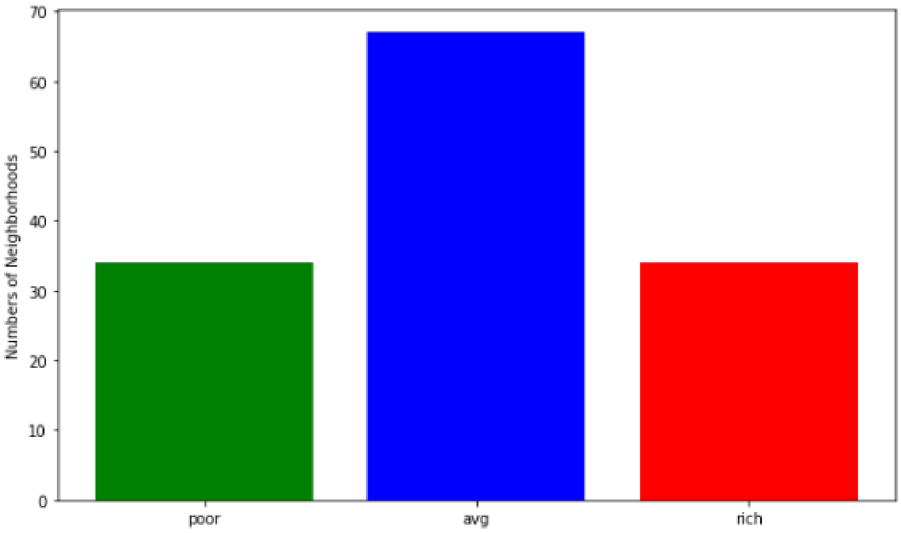

Figure 3: Income in neighborhoods of Toronto.

As it looks like Toronto is a rich city. Only 32 neighborhoods are poor which means that ¾ of the neighborhoods are average or rich which is good for opening a restaurant in Toronto.

Lastely, we bin the population change within the neighborhoods to find out if a lot of people are moving into the neighborhoods or not. So we can determine if a neighborhood is trendy or not. This should also be an indicator it is worth to open a restaurant in a neighborhood or not. If a lot of people are leaving the neighborhood it might be not a good idea to open a restaurant there. So let bin the population change into 3 groups again so we can get a better overview of Tornoto.

Table 4: conditions for trendy areas.

Category	Condition
Not trendy	$< 1.4\ \%$
Lightly trendy	$1.4\ \% \leq x\ < 4\ \%$
Trendy	$x \geq 4\%$

Again we use the lower and upper quartile to determine the borders. As you can see in table 4 we describe a neighborhood as not trendy if we have population change below 1.4 %. If the population change is bigger than 4 % it is called trendy and for values between those condition it is lightly trendy. The results are again plotted in figure 4.

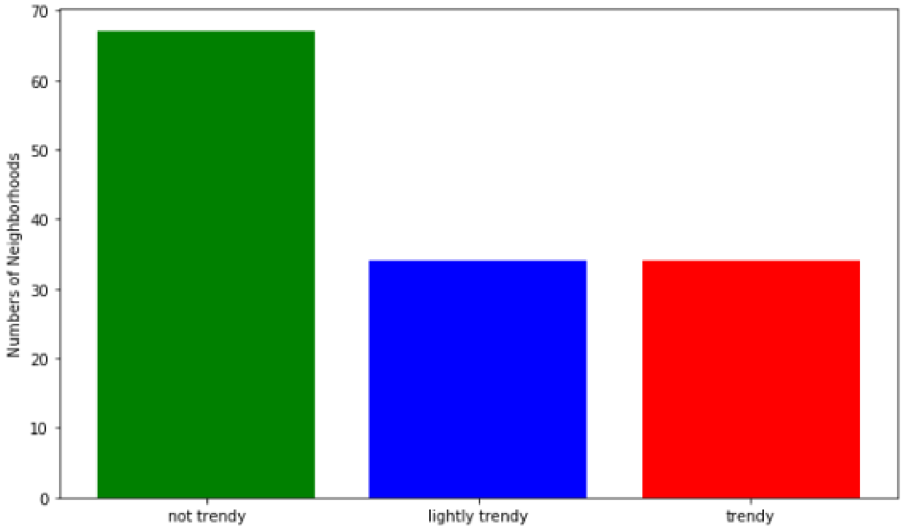

Figure 4: Area types by population change.

As you can see, almost have of the areas in Toronto are not trendy which means less than 1.4 % change in population within the last 4 years. Therefore it is important you cluster neighborhoods to find out where to open a restaurant since its better to open a restaurant in a trendy area than in a not trendy area since more people are moving into those areas the next years.

The we doing the same for the restaurants and bin them in 3 groups again. The amount of restaurants are counted and uploaded via Foursquare API. The conditions are shown in table 5. The condition shows that a neighborhood has less restaurants if the area has less than 4 restaurants in a radius of 500 meters. Vice Versa a neighborhood has a lot of restaurants when there are more than 19 restaurants. This again can also help you to find the neighborhood you prefer to open a restaurant regarding if you want to open a restaurant in a popular food area or more likely want to open a restaurant in a neighborhood with less restaurants in order to gain more costumer throughout less competition.

Table 5: Condition to determine if an area has a lot of restaurants or not.

Category	Condition
Less	< 4
Medium	$4 \leq x \ < 19$
A lot	$x \geq 19$

Lets use those information of each neighborhood to cluster them based on their similarities so an entrepreneur can have an easier overview of neighborhoods which prefers his interest.

4 Cluster Modeling

Before we can cluster the neighborhoods into 5 clusters the values of the dataframes are converted into dummy variable. After the kMean cluster algorithm is use to find the similarities of the neighborhoods. The results are shown in the Toronto map in figure 5. Each colored marker is a neighborhood within a cluster. In general we have 5 clusters.

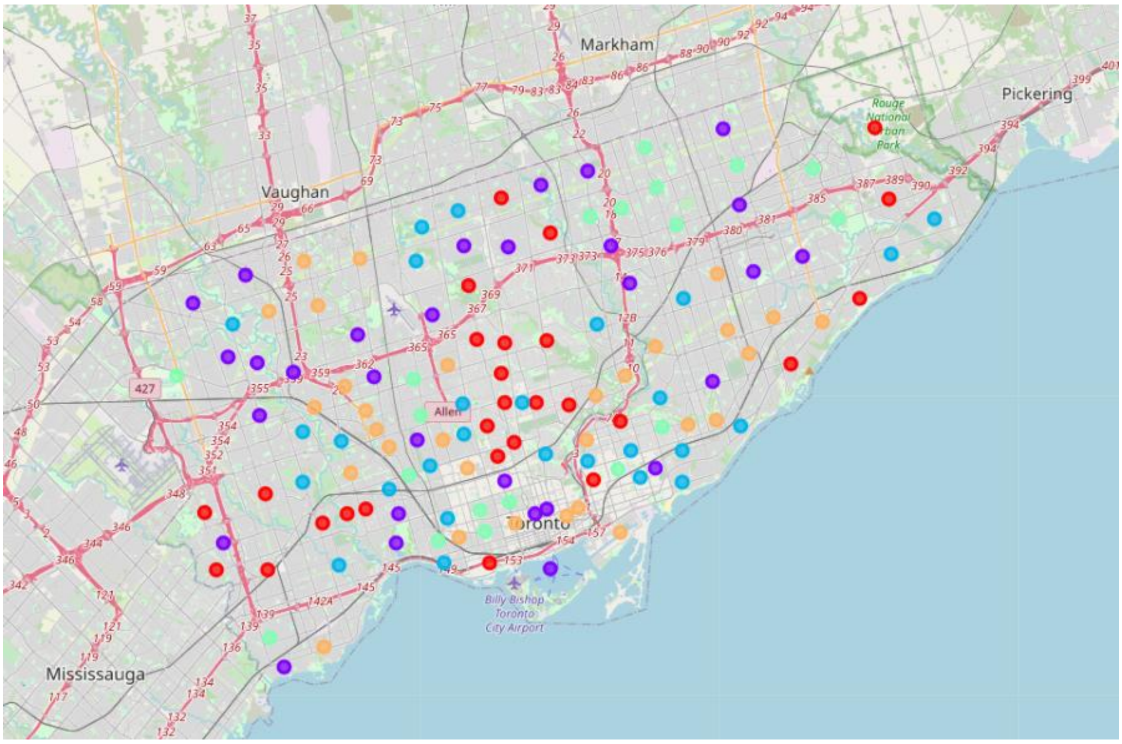

Figure 5: Map of Toronto with marked neighborhood depending of the cluster (red = Cluster 0, purple = Cluster 1, blue = Cluster 2, light green = Cluster 3, orange = Cluster 4.

Each cluster is defined as you can see in table 6. You can observe that cluster 0 is are rich areas which are mostly not trendy or super trendy. It already has an average amount of restaurants. Cluster 1 is a super trendy area with a good mix of poor, average and rich people. But they already a lot of restaurants in their neighborhoods.

Table 6: Category of each cluster.

Category	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Poor	0	7	18	0	9
Average	0	17	0	21	18
Rich	27	10	0	0	3
Not trendy	16	0	18	21	0
Lightly trendy	1	0	0	0	30
Trendy	10	34	0	0	0
Less venues	10	11	7	0	10
Medium venues	10	12	10	16	9
A lot venues	7	11	1	5	11

Cluster 2 are neighborhoods where only poor people live. They are not trendy and have a medium amount of restaurants. Cluster 3 are neighborhoods for the average population but they are not trendy and have an medium amount of restaurants. Cluster 4 are neighborhoods with a good mix of poor and average people. Only a few rich neighborhoods are included. All neighborhoods are lightly trendy and could bring a big opportunity for the future. Indead, they already have a lot of restaurants within the neighborhoods.

5 Conclusions

After we cleaned and merged all datasets we could find out that Toronto is a wealthy city with a lot of young and middle aged people. After we clustered all neighborhoods with the kMean cluster algorithm based on their similarities we could observe only 2, maybe 3 clusters are interesting for opening a restaurant. Cluster 0 shows trendy areas where only rich people live and already have a lot of restaurants which means a big opportunity for opening a restaurants. As it looks people in that cluster likes to go to restaurant. In addition, Cluster 1 are super trendy areas with a good mix of all different population types and a lot of restaurants. Here we also have a big opportunity to open a restaurant. Cluster 4 is maybe also interesting since all neighborhoods are lightly trendy and could become a new hotspot. Neverless, cluster 2 and 3 are not recommend for opening a restaurant since only poor people, not trendy areas and less restaurants are based in. Neverless, still maybe a good idea for a cheap pizza place.