fINAL REPORT

Capstone-Project Coursera



Guglielmo Sanchini

TAble of  
Contents

INTRODUCTION 1

DATA 2

methodology 3

results 9

discussion 12

conclusion 13

# INTRODUCTION

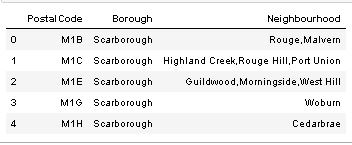
The problem I chose consists in selecting the right location where to open a restaurant in Toronto, in order to make the right choice in terms of neighborhood, type of restaurant, price range etc.

This is interesting because it requires both location data (through the Foursquare API of course) and other kind of data to evaluate how a restaurant is doing.

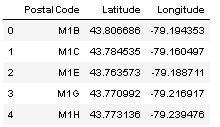
It is of extreme importance for a businessman wanting to step into the food business to find the right spot where to open a restaurant that can generate profits for a long time. 70% of restaurants close after six months, so the problem is real and pretty evident.

# DATA

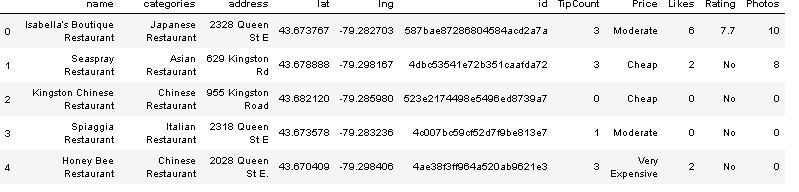
The data I will use consists in a Wikipedia table which lists Canadian postal codes:



Then I will use a csv file downloaded from the Coursera portal which provides latitudes and longitudes of Toronto neighborhoods:



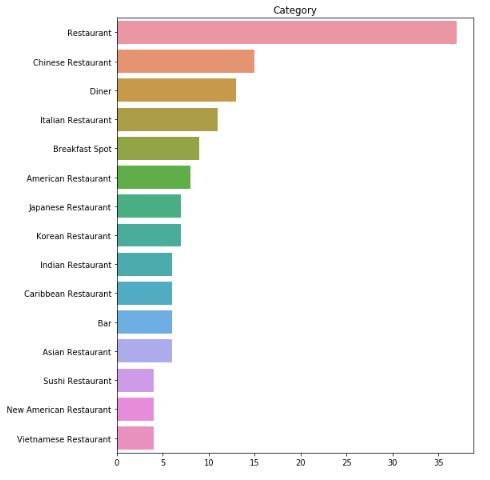
Lastly, I will use the foursquare data to find restaurants in 38 of the neighborhoods of Toronto I selected, and I will keep the information related to the names of restaurants, the count of tips, photos and likes, the rating and the address, as well as the restaurant id and category:

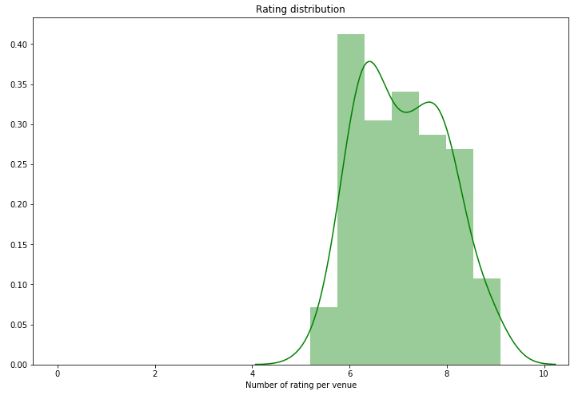
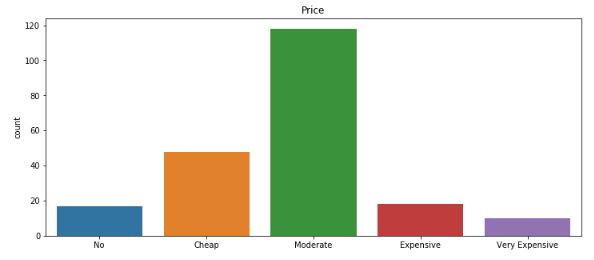


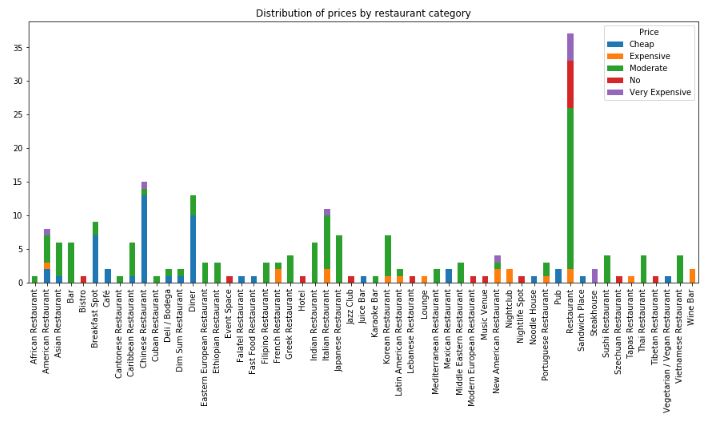
# methodology

First of all, I downloaded the data progressively, since the free account for using the Foursquare API had some strong limitations, especially with regard to the acquisition of tip count, rating and photos data.

Then I concatenated all the partial dataframes I obtained using Foursquare to create a single dataframe with all the data needed for the analysis.

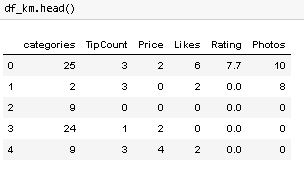
I explored each column, and some relationships between them, for example, how the category of a restaurant affects the price range or the number of tips for restaurant category: 



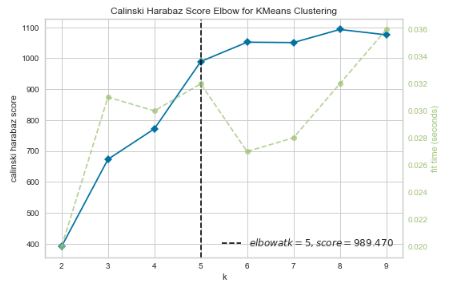
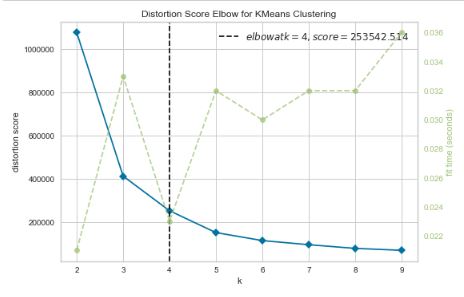


From the above pictures one can see that the majority of the restaurants analyzed had moderate prices, and that, when rating was present, it was in most of the cases above 6.

After this exploratory analysis, I dropped some of the columns that weren’t necessary for the analysis, and I used sklearn’s LabelEncoder to transform the columns of Price and Category into ordinal numbers.

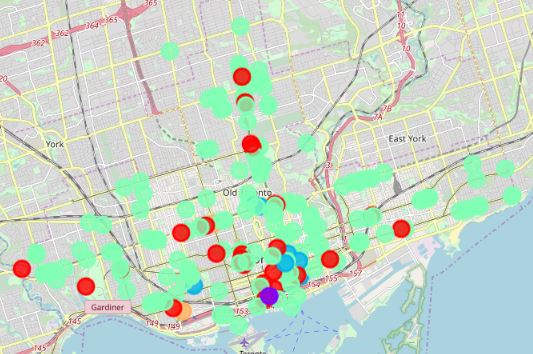


Then, I wanted to apply the usual K-Means algorithm to find clusters of restaurants, so I used the famous elbow method to decide the optimal number of clusters, with two different methods:



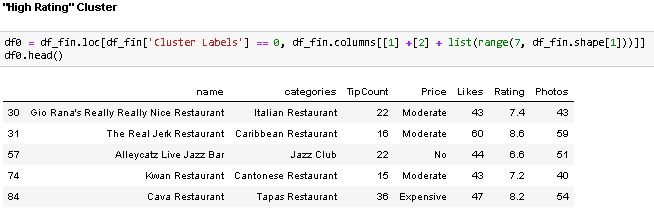
According to the second method I chose to go with 5 clusters. In the next section I am going to present the results.

# results

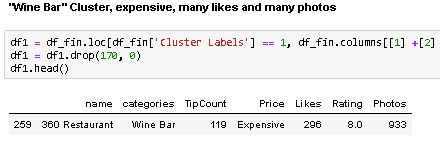
The clusters on the map look like this:

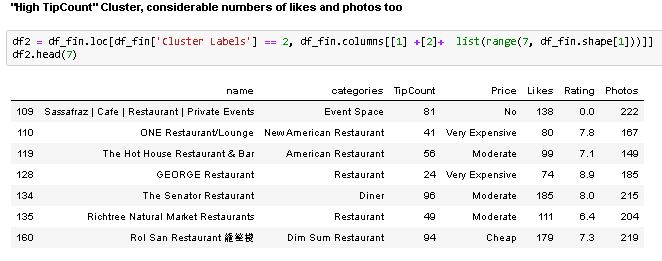
The cluster characterization is as follows:

The HIGH RATING cluster:

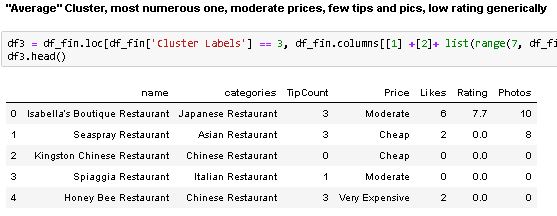


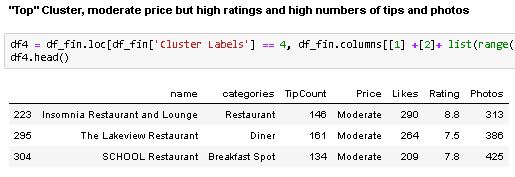
The WINE BAR cluster



The HIGH TIP COUNT cluster

The AVERAGE cluster

The TOP cluster



# discussion

It is clear from the map that the most numerous clusters is not characterized by the location of its restaurant, but from the fact that its members show average features. The other clusters show a more characteristic disposition on the map.

So, where would I open my restaurant in Toronto or where would I recommend someone to open a restaurant in Toronto?

Based on the analyses I conducted, here is my answer: it depends?

We have to be a bit more specific in order to give an appropriate answer: if someone wanted to open a restaurant which will likely produce much social engagement and with a well-off target, then he/she should open a Wine Bar, based on what Cluster1 suggests, the location in this case seems not so relevant, since we are considering a single example.

Instead, for a more general recommendation, I would suggest that the restaurant be opened in the western area of Toronto, for example in Dufferin or Lansdowne, where competition is not too high and there is a prevalence of restaurants belonging to the Average Cluster, so that the possibly to stand out is high.

# conclusion

To conclude this report, I can say that the analysis gave some satisfactory results, but it obviously could have been done very differently and in a much deeper way. One could have used CPLEX of IBM to use decision optimization for example, to select the best area based on some constraints, such as the category of restaurant one wants to open, or the price range he/she intends to choose. Also, not always ratings, likes and pictures are significant for evaluating the overall situation of a restaurant.