Finding the best location for a new Chinese restaurant in Toronto

Ping Yuan

May 29 2020

# Introduction

Opening a new restaurant is a challenge for investors or stake holders. There are so many factors leading to a successful restaurant and, location is one the most important ones. A restaurant can have a very delicious food and warmest atmosphere but, in order to maintain a healthy financial status, it should be placed in the best location to attract the largest number of customers possible. As a resident of Toronto and Chinese food lover, I am going to use data science techniques to help the potential investors or stake holders to find the most suitable neighborhoods for opening a new Chinese restaurant in Toronto, Canada

Toronto is the provincial capital of Ontario and the most populous city in Canada, with a population of over 2.93 million. Toronto is a very diversified city with immigrants from different ethic origins, such as 12.5% Chinese, 12.3% English and 12% Canadian. This diversity is reflected in Toronto's ethnic neighbourhoods, which include Chinatown, Little India, Little Italy and etc.

We are particularly interested in **areas with no or minimum Chinese restaurants in vicinity**. We would also prefer locations which is as lively as possible and **with as many venues as possible.**

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

# Data

Based on definition of our problem, factors that will influence our decision are:

* number of Chinese restaurants in the neighborhood, if any
* number of existing restaurants in the neighborhood (any type of restaurant)
* number of venues in the neighborhood

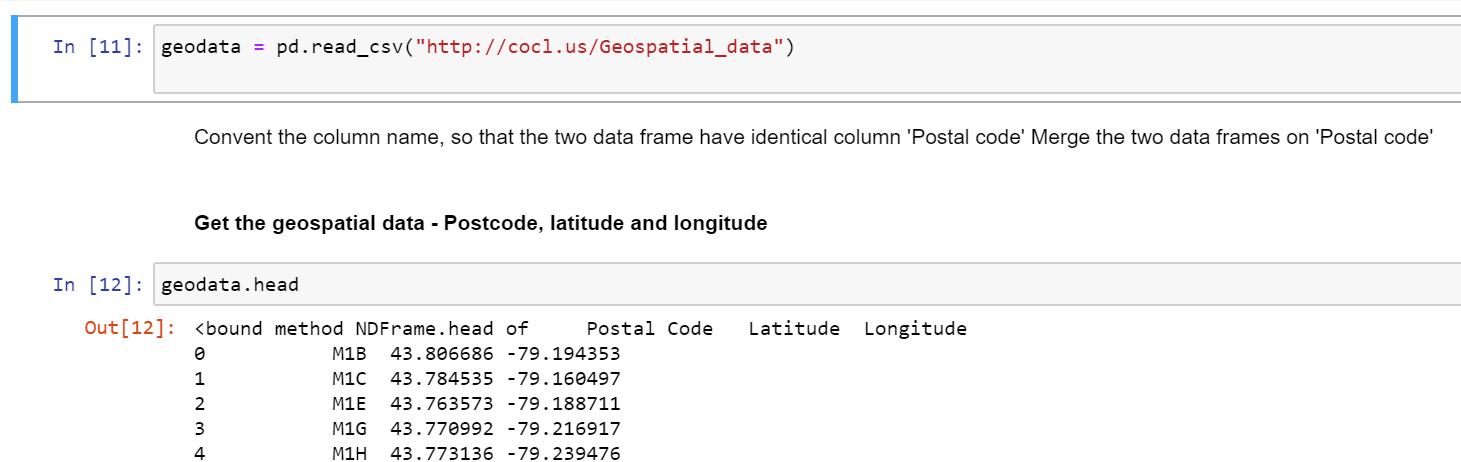
We will need two type of data to start the project.

**One dataset** contains the **neighborhoods and the latitude and longitude coordinates** of each neighborhood in the city of Toronto.

We plan to get this data from Wikipedia.

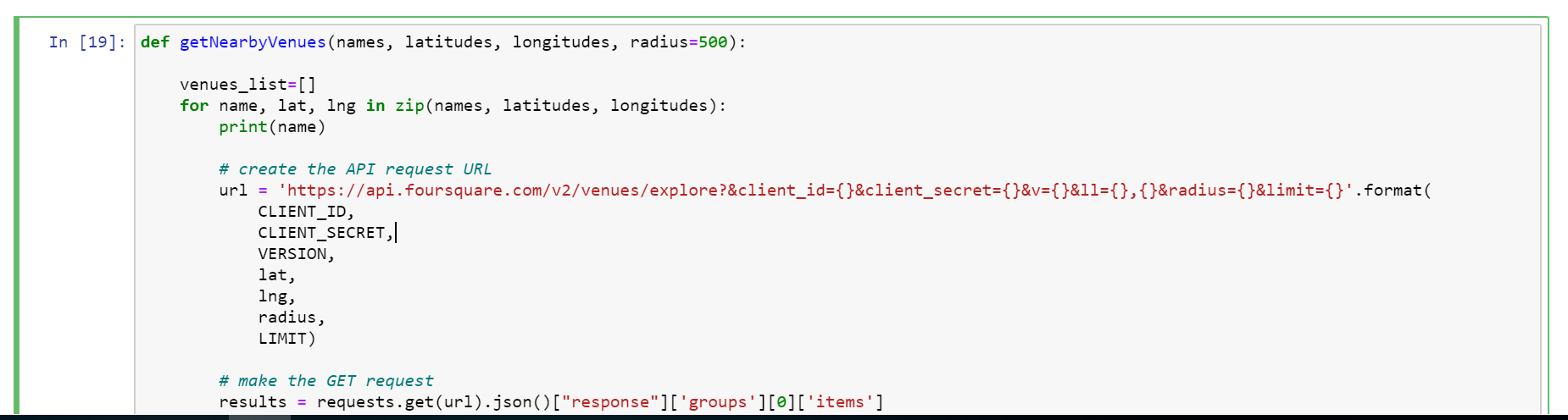


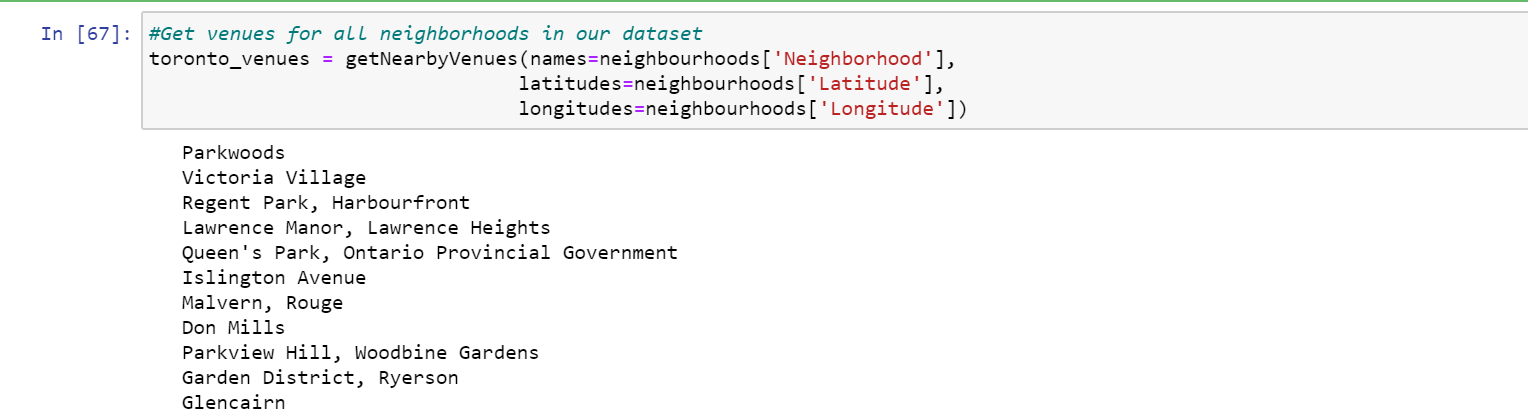
This data frame only has postcode and neighbourhoods. We also need the altitude and longitude coordinates of each neighborhood. We use the data from previous course assignment



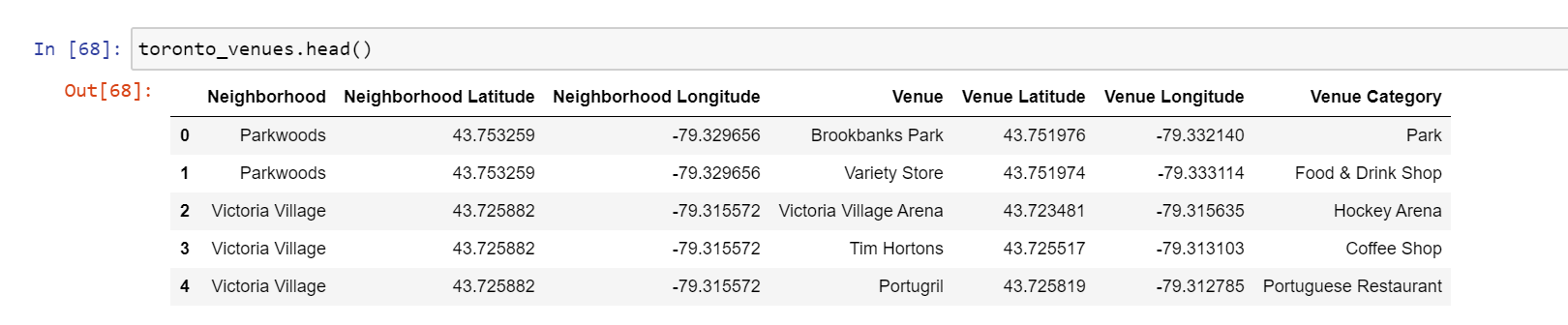
We then clean the two data sets and merge them together for our first set of data .

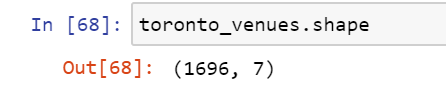
**The second set of data** is **data of Restaurants, Chinese restaurants and different venues and their locations** in the city of Toronto. We will use Foursquare api to obtain this dataset.





We obtain the entire list of venues.





We filter the data set to get the Chinese restaurants to analyze the data



# Methodology

The **methodology** we use is an iterative systematic approach which follows a sequence of steps.

1. Business problem understanding
2. Data gathering and cleansing
3. Data discovery
4. Machine learning
5. Fine tuning and testing
6. Deployment
7. Feedbacks

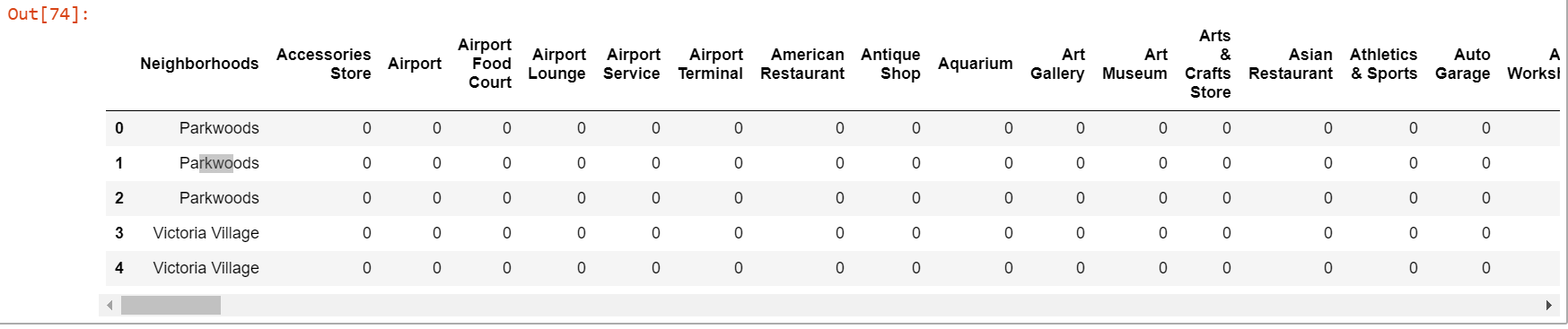
In the previous chapters, we have discussed business problem and initial data preparation.

Next, we will use panda data frame and functions to analyze the data, and group neighbourhoods based on the features.

We use One Hot Encoding to split the column which contains numerical categorical data to many columns depending on the number of categories present in that column. Each column contains “0” or “1” corresponding to which column it has been placed.

It helps us to group the data set based on the features

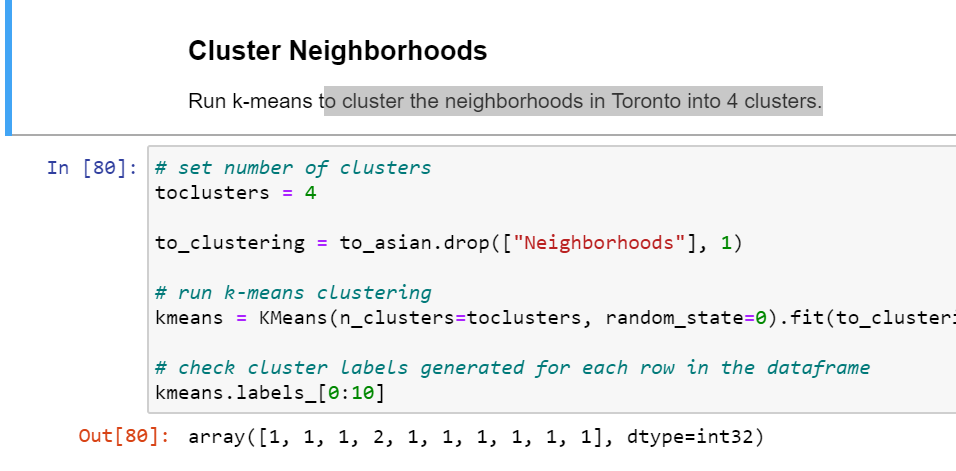




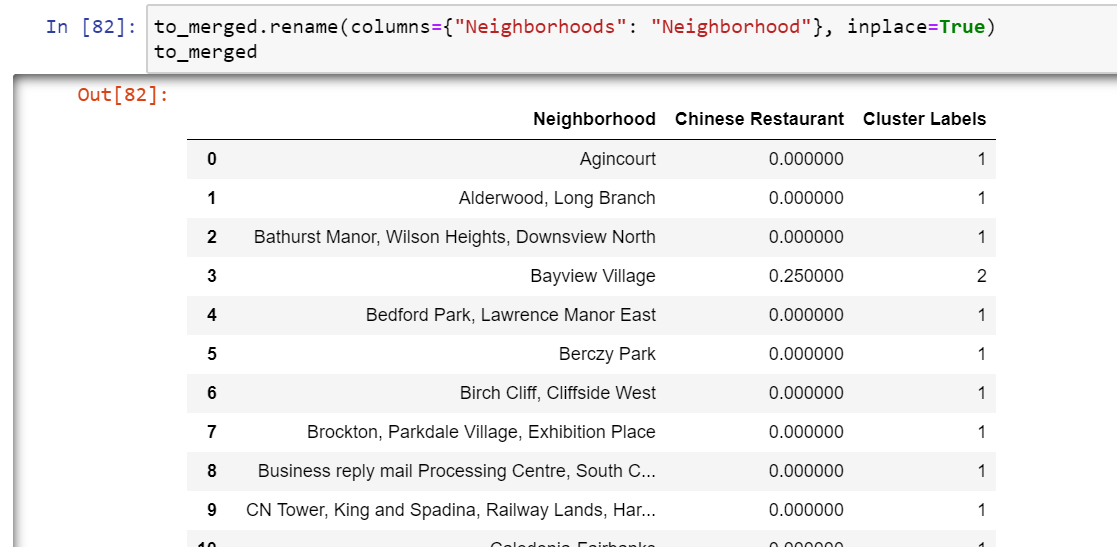
We then get the data frame of neighbourhood with the mean Chinese restaurant percentage over total venues



We use K Mean algorithms to cluster the neighbourhood based on the percentage of Chinese restaurants of each neighbourhood.



We label the clusters



There are 4 clusters:

1 – no Chinese restaurant

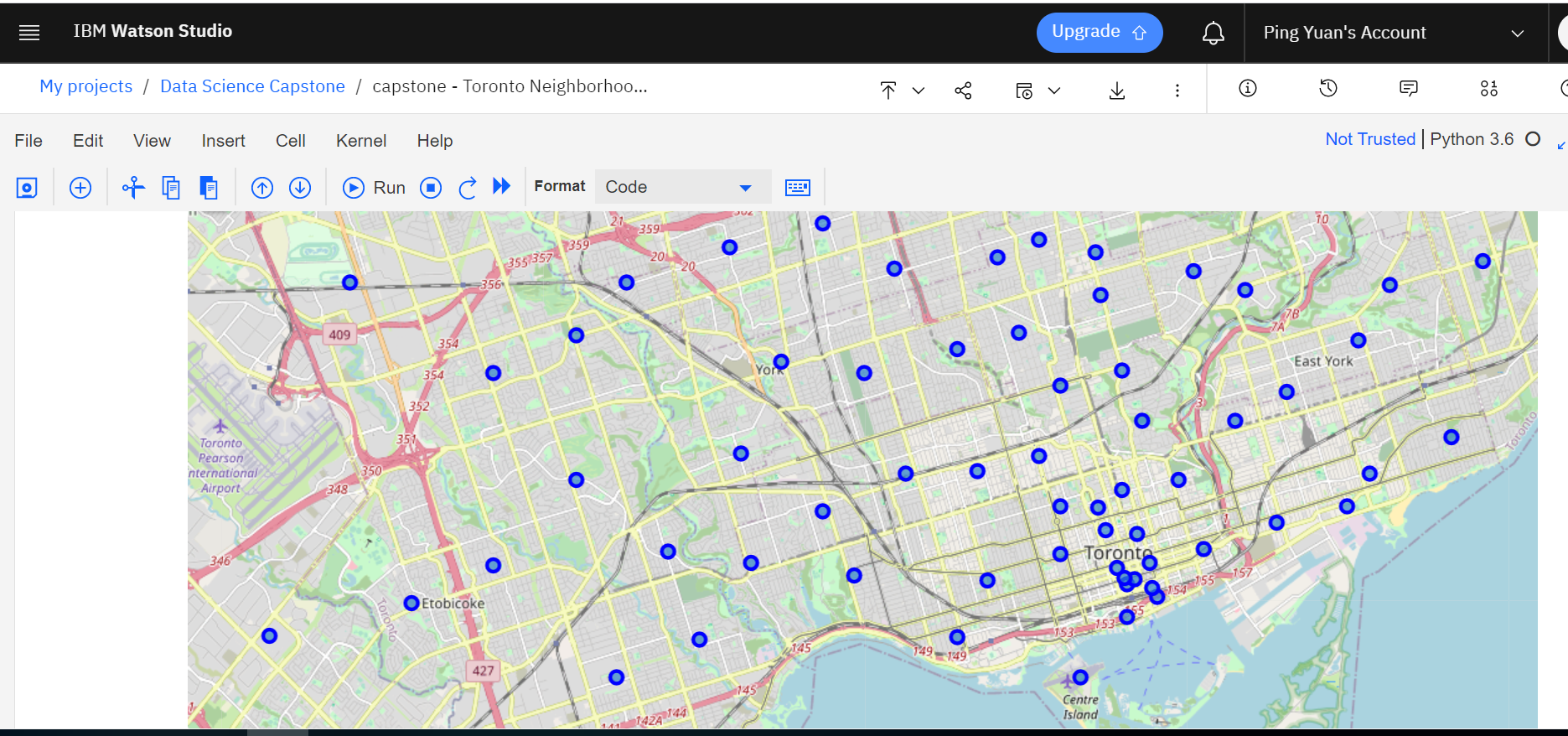
2 – small percentage of total venues are Chinese restaurants

3 - Medium percentage of total venues are Chinese restaurants

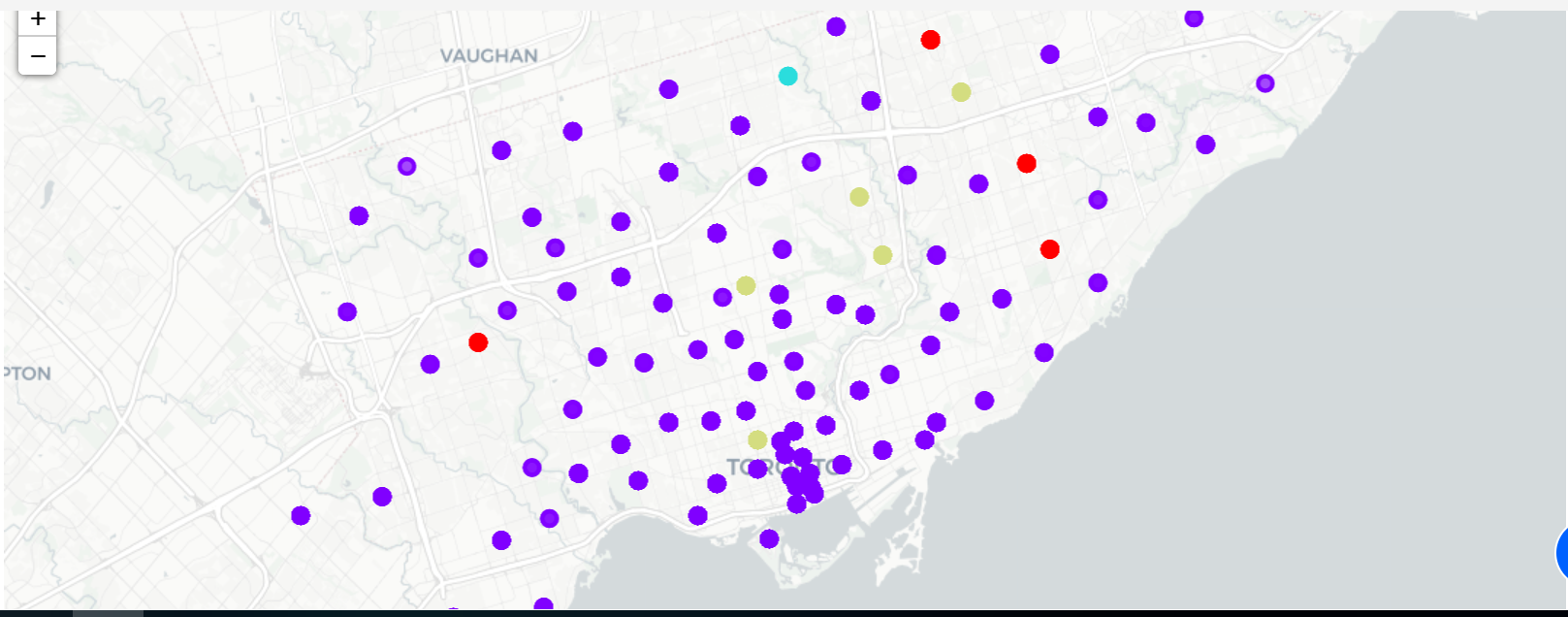
4 - High percentage of total venues are Chinese restaurants

Then, we will use the Folium library to visualize the neighborhoods in Toronto City, the cluster of neighbourhoods we calculated with K mean algorithm and the heatmap with existing Chinese restaurants

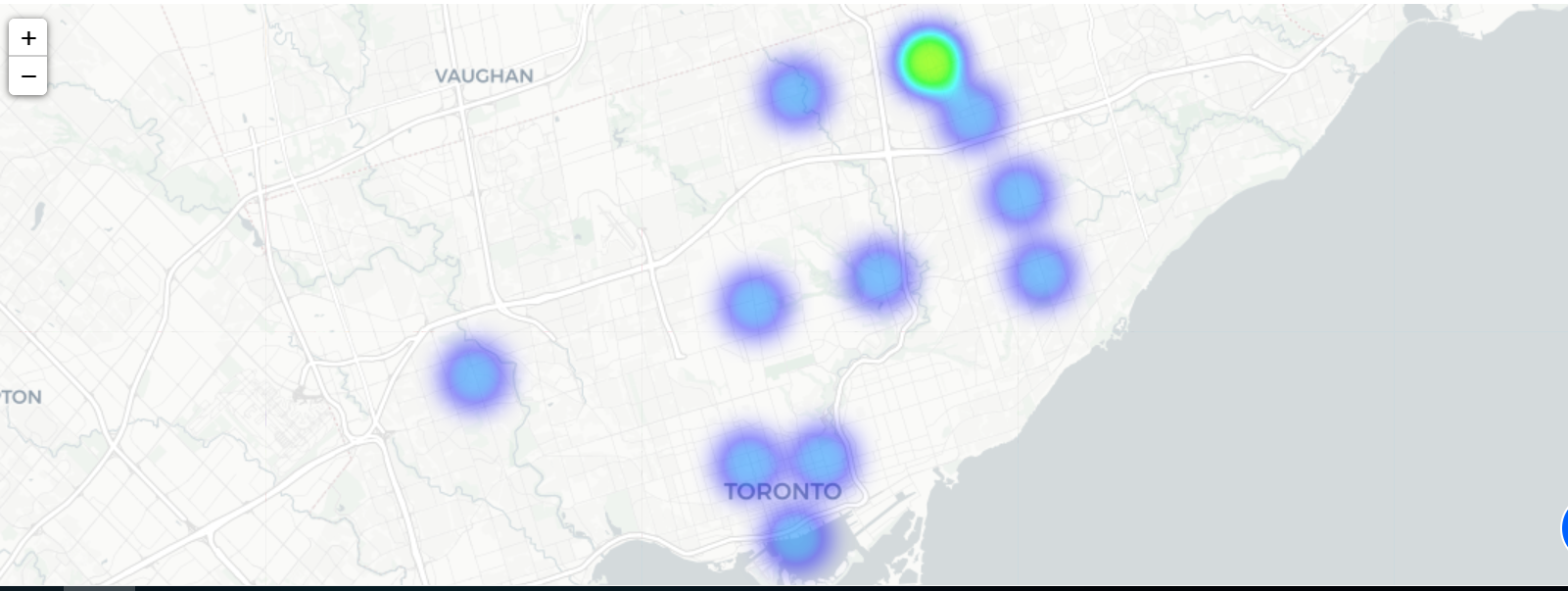
This is the Toronto neighborhoods, shown with Folium library



This is the Cluster result of K Mean Algorithm, showing neighbourhood with lowest percentage of Chinese restaurant in purple



We also visualize the existing Chinese restaurant heatmap in Toronto



Finally, we will present the result to the stake holders and get feedbacks to improve our solution.

# Results

From the K mean cluster algorithm result, we obtain a list of neighbourhoods which have no or minimum Chinese restaurants. These neighbourhood would be ideal candidates for opening new Chinese restaurants. For example:



# Discussion and Conclusion

In this project, we went through the life cycle of business problem definition, data preparation and cleansing, data analysis, machine learning algorithm modeling and visual presentation of study result. With the available public data from Wikipedia and Foursquare API, Python and machine learning algorithms and models, we were able to predict the candidate neighbourhoods and visualize the result on Toronto map in IBM Watson studio.

The study of best Chinese restaurant location was based on Toronto neighbourhood and venue data only. If we could add other data and features, such as neighbourhood income, housing price and etc, the study would be more comprehensive, and the result would be more convincing. There is also room for improvement on experimenting with other machine learning models and algorithms.