**Capstone Project - The Battle of Neighborhoods (Week 2)**

**I Introduction**

**I***nvestment is in general a complex and challenging task.One of the most difficulty is to collect right data.Today there are certainly lots of data sources thanks to internet data centers developpement ,data processing technologies which make it possible to collect, process and store data in real time.Although these technologies are present in our day life and more often availables as open source tools,they still need to be carefully used in other to obtain right data.*

# 1.1 Problem

**T**he problem as set consists collecting ,porcessing and analyzing the cities **NEW YORK CITY, Toronto And Paris**. More precisely ,the problem is about Collecting informations about all the venues of different neighborhood of all borough in each of these cities in order to find the similarities between them base on data collected using data anlysis tools ,here is python and external API such as foursquar.The result obtained results will be use to suggest a suitable location for investment .

## 1.2 interest

**Evidently,interested are either poeple interested in investing in one of these tree cities or somme one willing to know a litle bit about kaind of venues in these cities.**

## 1.3 Data socuces

### 1.3.1 problem 1 Cities anlysis

**1.3.1.1-NYC(new york city) all data where provided link\_to\_data\_set:** <https://cocl.us/new_york_dataset>

From this link data are downloaded using python api pandas ,process and transform into usable data format. The foursquare API is then used to generate the final data set including geografical coordinate of each neighbouhood.Thinal data set in the generated (Data frame including geographical coordinates of each neighborhood

**1.3.1.2 -Toronto links To dataset:**  <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> -<http://cocl.us/Geospatial_data>\*\*

From this link data are downloaded usind python api pandas ,process and transform into usable data format. The foursquare API is then used to generate the final data set including geografical coordinate of each neighbouhood.Thinal data set in the generated (Data frame including geographical coordinates of each neighborhood.

**1.3.1.3 Paris links to data sets:** <https://fr.wikipedia.org/wiki/Liste_des_quartiers_administratifs_de_Paris> [<https://opendata.paris.fr/explore/dataset/quartier_paris/download/?format=csv&timezone=Europe/Berlin&use_labels_for_header=true&csv_separator=%3B>

From this link data are download ,loded usind python api pandas ,process and transform into usable data format. The foursquare API is then used to generate the final data set including geografical coordinate of each neighbouhood.Thinal data set in the generated (Data frame including geographical coordinates of each neighborhood

### 1.3.2 Problem 2 (venue sugestion for ristaurant investment)

data set used here is the results of the data collected and transform (using python and foursquere api) into strutured and usable data format in the previous section

# II-Data collection , processing and anlysis for each city

## II-1 New York City.

**New York City** (**NYC**), also known as the **City** of **New York** or simply **New York** (**NY**), is the most populous **city** in the United States with polpulation of 8398748 inhabitants.This is the first city on which ours first analisys wiil be focussed.

**AS** previously said,data are collected from links provided,the programming language used is python on IMD cloud,more precisely IBM service.Thanks to these tools we have colletted and tranformed data from unusable data forrmat to structured data frame.Below is the scrreshut of the final data structure where in the columns we have from left to right exactly :**Borough's name ,Neighborhood's name ,Latitude and Longitud**e of the overall newyork city.

Newyork city data set with geogaphical coordinates

*Une image contenant capture d’écran, intérieur, mur

Description générée automatiquement*

The total number of Neighborhood in newyork city is : **103**

Using External fousquare API and further processing data are transformed in the usable the final data structure wifhi can be seen in the following.

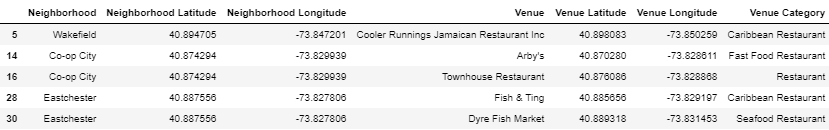
Table including venue categorie

Une image contenant capture d’écran

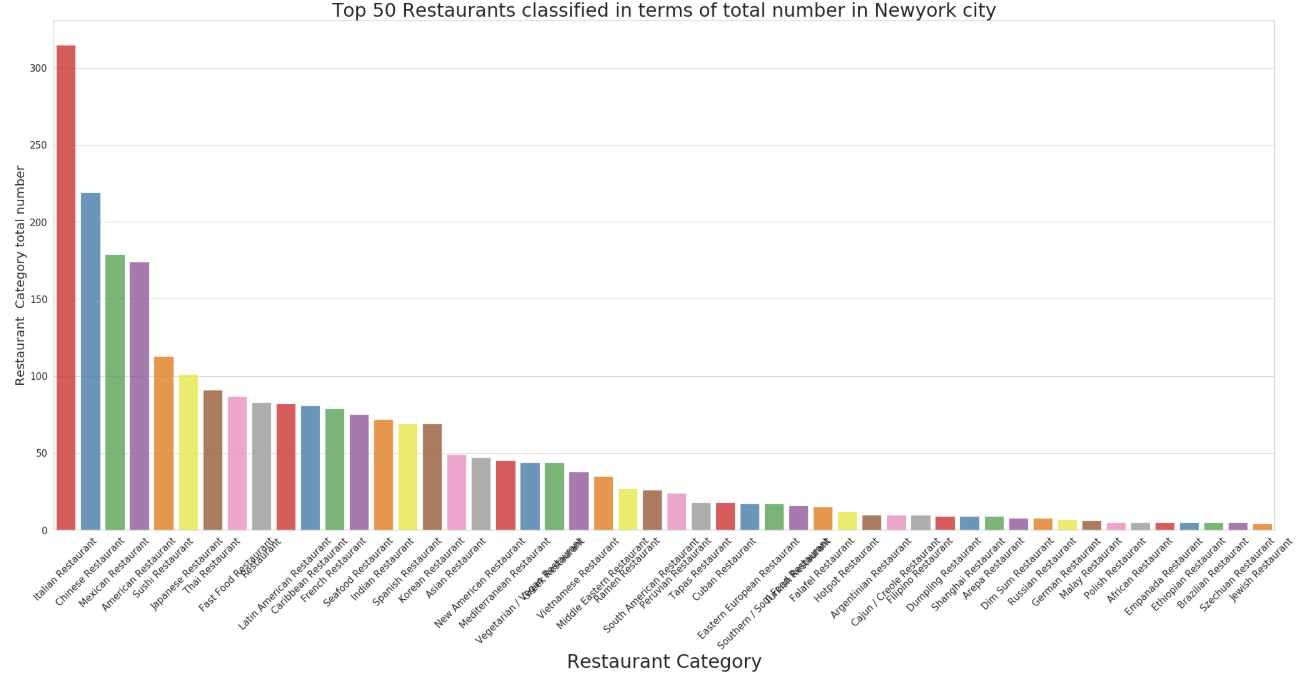
Description générée automatiquementThe total number of venues in newyork city is : 1.22 per 100 inhabitants.

In terms of venues diversity,there is 429 venues categories.

As we are interested in this analisys in investing on a Ristaurant ,we need to extract from the dataframe only rows coresponding to restaurant venues.final data in the folowwing is shown the first 5 element of the final table.

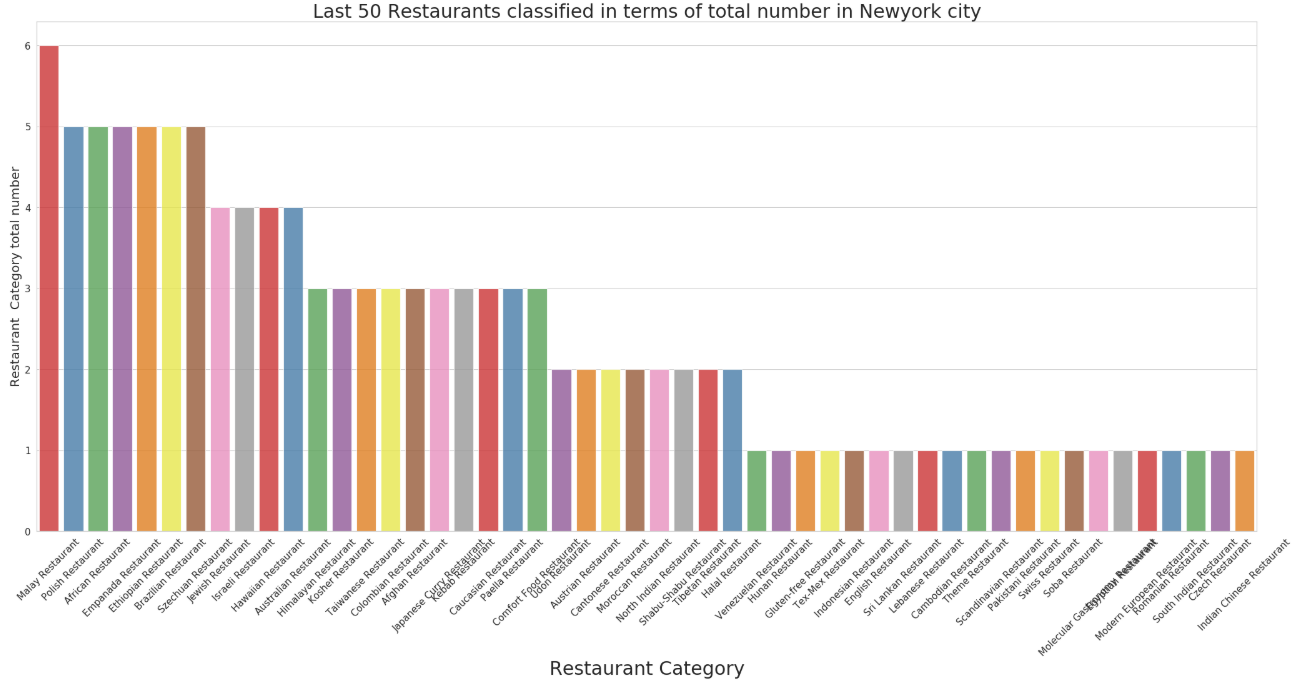


The above data frame is the one we wiil be using for this project.

Herlet plot the top 50 and the 50 last restaurants in terms of number of restaurant in newyork city .

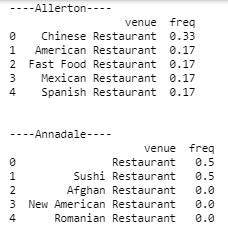
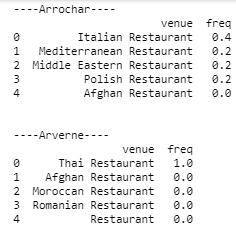
Top 50 of restaurants in NYC 1

With **315** restaurants in NYC,Italian restautrant in terms of number is the most respresented ,folowed Chinese,Mexican and American Restaurant with total number respectively 219,179,174and 113. Unexpectedly,the Italian restaurant comes first in USA .



This last graph shows that,in NYC,they are lots of ristaurants categoris with only one building such as Indian chinese restaurant,South Indian Restaurant,and so on.

Let look a litle bit deeper into the data set .Here below are the top fives venues for 6 neighbohoods in Newyork city.



In this project,our anlysis will be totally decicated to Restaurant,thus they will be used in the first time to seek the similarity and dissimilarity between Newyork city ,Toronto and paris and secondely with deeper analisys will be used to sugest the most appropriate place if not an acceptable place to invest in restautant event which Restaurant category may the best choise according to our analisys.

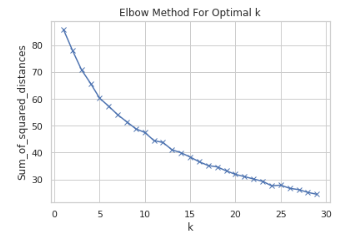
Let create a new data frame with the 10 most comon resturants as following.

In the next section,we are going to use this data to group the data set into différen into cluster.

**5.Clustering (K-mean) :**

**5.1** .Parameter turning .

The first step of our clustring is to choose the aproporiate value of k(number of cluster).

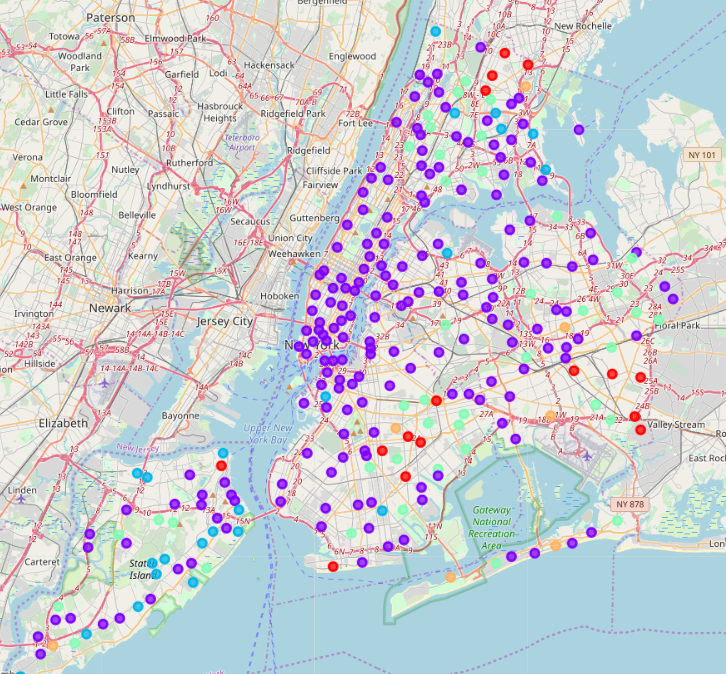


best k for the kmean cluster 1

In the plot above the elbow is at k=5 indicating the optimal k for this dataset is 5

The overall clusters ar shown in the following map of newyork city :

best k for the kmeans cluster 2





NYC cluster 1

NYC cluster 2





NYC cluster 3



NYC cluster 5

NYC cluster 4

**II-2 Toronto**

The City of **Toronto** **I**s the Canada's largest city, the fourth largest in North America with a population of 6196731 inhabitants.Data for this analysis where obtain using web crawling python API and secondely processed used the same processing méthodes decribed in the previous city analysis(NYC).