**Capstone Project: The Battle of Neighborhoods**

**Introduction/Business Problem**

On Monday morning, the guy at the travel agency told me “Computer says, uh!” A fancy statement actually meaning “I don’t know what you are talking about, but I can offer you a 3 weeks tour as a good start. Good damn it! well without knowing it, that’s where my AI journey begins… Sometimes in life, opportunities pop-up in front of you and you face you yourself with a “should I go ahead and take it” moment. Indeed, starting a new life in Manhattan is by far quite a new story for a “didn’t know there were other cities than Paris on earth” type of guy. Manhattan is one of the 5 boroughs of New York City with over 1,6 M of inhabitants, a bit smaller than Paris “Big Picture” which stands at more than 2,1M then – I knew I would have my revenge, but represents only 1/5 of NYC- OK I give up…

Paris on the other hand is split into 20 boroughs and I do love the charming environment of the one I live in (12th but don’t tell anyone) and as my new life in NYC was about to start I was looking for someone who can help me finding a little Paris in Manhattan as I am about to stay for a while… well you know how it ended, and for sure, I haven’t met the future Dr Spence Reid (Criminal Minds). Therefore, I quickly realize that it was a more DIY job posing the question of finding a place likewise mine in that other environment or in other words how I can find a neighborhood in Manhattan that is similar to mine. Furthermore this could be also expanded to neighborhood mapping helping other Paris folks finding their dream land in Manhattan - So welcome to manathan.predict(myParis) journey.

**Data Description**

Because the goal is to find similarities, we should define what similarity is and it should be understood as areas sharing characteristics considered as being equivalent. For that purpose, data will be used as measurement units. In this context data required can be divided into 2 categories: Geo localization data and characteristics data of the concerned areas.

Geo localization data: (neighborhoods, latitude, longitude)

* Regarding NYC and particularly Manhattan, data file from course 3 is readily available via <https://cocl.us/new_york_dataset> .
* Concerning Paris, geo localization data is available via an xls files from <https://opendata.paris.fr/explore/embed/dataset/arrondissements/table/> .

Data regarding characteristics is available via Foursquare. In each area characteristics of top spots can be retrieved.

**Methodology**

First of all, geo localization data related to cities has been loaded and displayed. The overall target is to get neighborhoods, latitude and longitude for both cities. Data has been displayed and tested against null, missing and incoherent value. Target dataframe will be built to form 3 columns each with the information stated above.

Secondly, categories data has been retrieved from foursquare for both areas. It consisted of the top visited spot in a given area. This information takes the form of venue categories and are not necessarily the same for both dataset. Therefore, it is important to align both categories in order to make relevant matching based on identical categories and, structured dataframes have been built for that purpose.

Once the data prepossessing is completed, AI algorithm is needed to identify areas that are similar to each other. For that purpose, K-nearest neighbor has been used. It is widely used and is particularly suitable when clustering is required. Once the Manhattan neighborhoods has been clusterized, Paris data will be mapped to Manhattan through KKN prediction.

**Results**

Manhattan has been divided into 5 clusters as per the below. Paris neighborhoods fall into 4 of the 5 clusters with a large majority on the 3rd Manhattan cluster. The 12e neighborhood I live in lays in the 2nd cluster. The table below summarize the findings and merges results.

1. Cluster 0



1. Cluster 1



1. Cluster 2

A screenshot of a cell phone

Description automatically generated



1. Cluster 3

A screenshot of a cell phone

Description automatically generated

1. Cluster 4



A close up of a piece of paper

Description automatically generated

**Discussion**

One important element to be seen is that Paris data is highly dominated by French and Italian restaurant.

Paris shows in general strong commitment to cluster 3 with 75% (15/20) of the neighborhoods laying in that cluster. That makes sense since best spot are among Italian restaurant, bakery and wine bar/shop venues and this tends to reflect quite well overall Parisian area.

Cluster 1 which maps 3 Paris neighborhoods as a high proportion of Park and Hotel has less restaurant as top visited spot.

For what concercing my area which lays at Paris boder (12-Reuilly), it is indeed less crowed in restaurant with more grab and go place and it make sense to fall into cluster 2.

Cluster 0 and 4 are very small and are less significant in meaning.

To further expand this analysis to other research it would have been interesting to consider for the participant to answer a series of questions about his preferences and starting an analysis following the methodology explained in this document.

**Conclusion**

This assignment had the goal to establish a mapping between Manhattan and Paris neighborhoods and finding the one that matches mine (12e). We have started by identify relevant data namely geo localization data and top spot associated with those. After preparing the data and defining final output format, we have processed them via KNN algorithm in order to find consistent group of clusters and 5 have emerged. Paris neighborhood data was then compared through KNN prediction to find best fit with Manhattan clusters and mine fall into cluster 2. Just need to choose now amongst neighborhood in that cluster.