**Setting up a new restaurant in New York City - Report**

**The Problem**

The Business Problem that I wish to help solve is basically in which neighborhood of New York city should a cook/business man build his new restaurant, given the fact that Manhattan is already a place filled with many many restaurants of several cultures. Thus, the idea of a new establishment in this city would already be huge challenge to undertake and even more so for the business to thrive.

According to an article made by Nick Hines from the Vinepair website (<https://vinepair.com/booze-news/new-york-restaurants-eat-at-every-on/>) “you can’t walk a New City block without passing a restaurant”. It even states that “80 percent of restaurants fail within five years”, so it would seem very difficult to get a new restaurant business going in this city.

A study of venues in other metropolitan areas around the world will help inform which types of restaurant are less common in New York city, thus improving the possibility of success in a city already filled with so many restaurants.

**The Data**

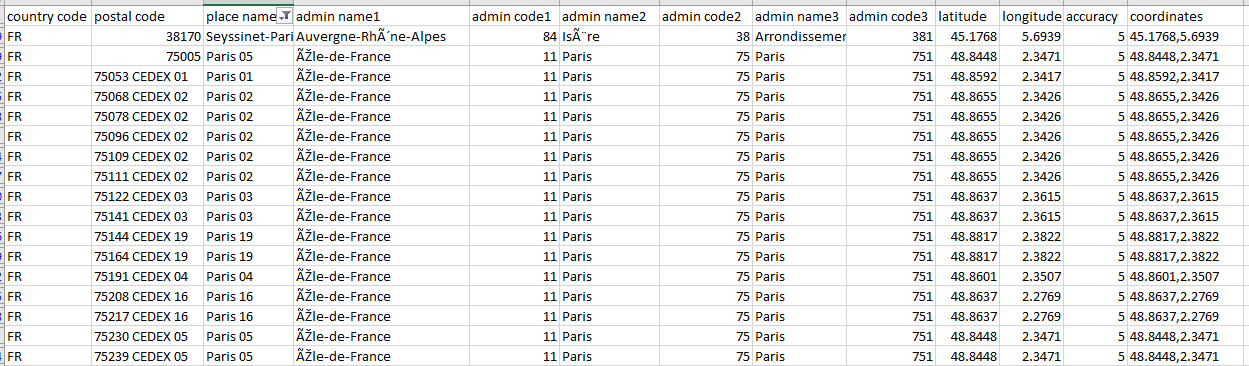
Foursquare API will be the chosen API to collect the data related to the venues for each geographical point.

To gather the information about geographical location (postal code, neighborhood, borough), ***Open Data Soft*** API (https://data.opendatasoft.com/pages/home/) was used, which is very simple to use, by simply writing down the country and city that you wish to research.

The chosen cities were New York City (which was already researched upon in the Laboratory for the Capstone section: https://cocl.us/new\_york\_dataset), Toronto (also researched as a deliverable in this final course: 'https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M'), Paris, Berlin and Porto.

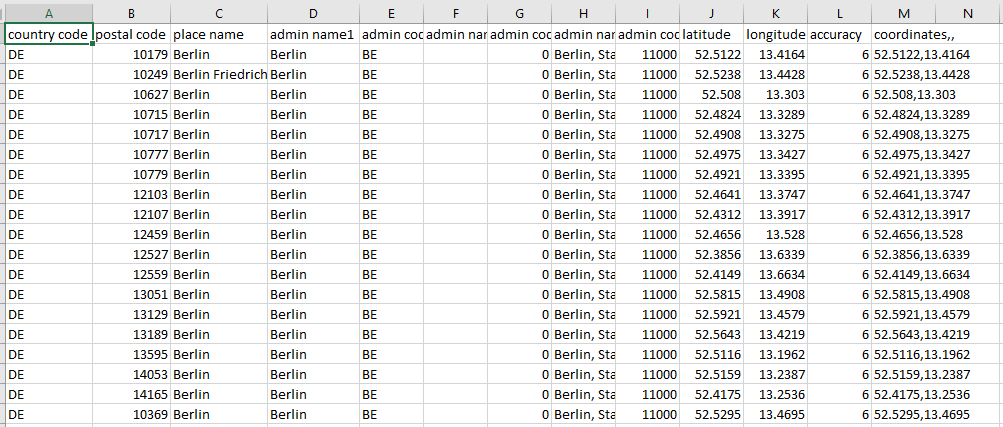
These last three were the ones that were researched through Open Data Soft. Paris and Berlin were chosen, since they definitely are known as multicultural cities with a significant diversity of venues and a considerable number of people living in them. Porto is sort of an outlier, but it’s an emergent city of Portugal, with lots of tourists coming every year and with evolving throughout time with new businesses being implemented.

The information for Toronto and New York is already in the “correct” format but in the case of the other cities the data is categorized in another format, for example, for Paris:



For these cases the Neighborhood and Borough were considered to be the same thing, which is the column “Place Name”, and of course the columns with the postal codes and the latitude and longitude geographical coordinates were kept. All the other columns were not used.

For the case of the Berlin data, the Postal Code was used as the Neighborhood name:



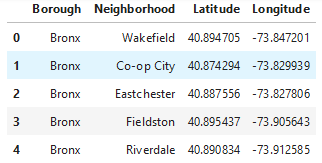
Notice how almost all of the entries have “Berlin” as the place name, so to be able to get the venues for each location, the postal code, which is unique, will be used to distinguish each place from the other.

In the case of Porto’s data, the case is a bit more complicated:

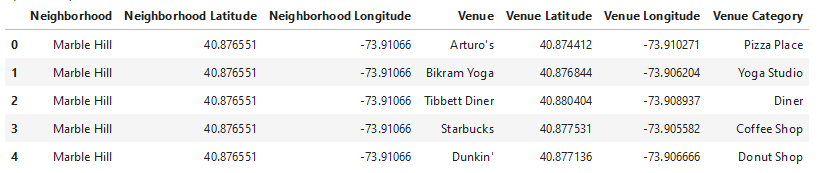


The data found is a lot more scattered throughout the entire Porto district, thus the data is not necessarily centered in the city’s core, but comprises a much larger area. But it will prove to be an interesting observation to compare this particular Portuguese city with metropolitan areas from powerful countries.

The Source data for the NY City locations should have the following format (after interpreting the JSON file):

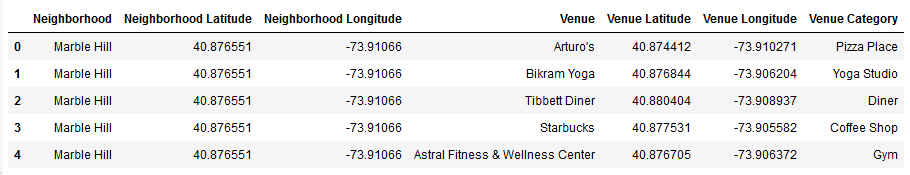


Finally, the data for the venues to be analyzed should have the following format:

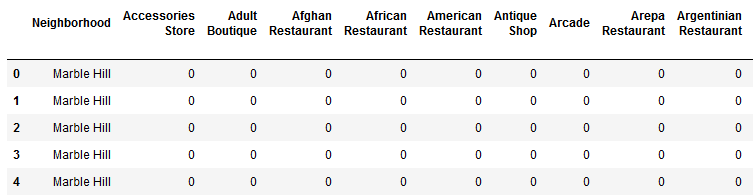


**Methodology**

First and foremost, the venues of each Manhattan location described in the Data section of this report, were retrieved from Foursquare’s API, with the output being a dataframe with following structure:



Then the one-hot coding procedure was done in order to get a matrix with the neighborhoods on the left side and the column names being the type of venue, with 1s marking if the current venue exists for a given location and 0s marking the opposite.



Then the data would be structured in the form that will be essential for this case study and will be used for Machine Learning (ML) algorithms:



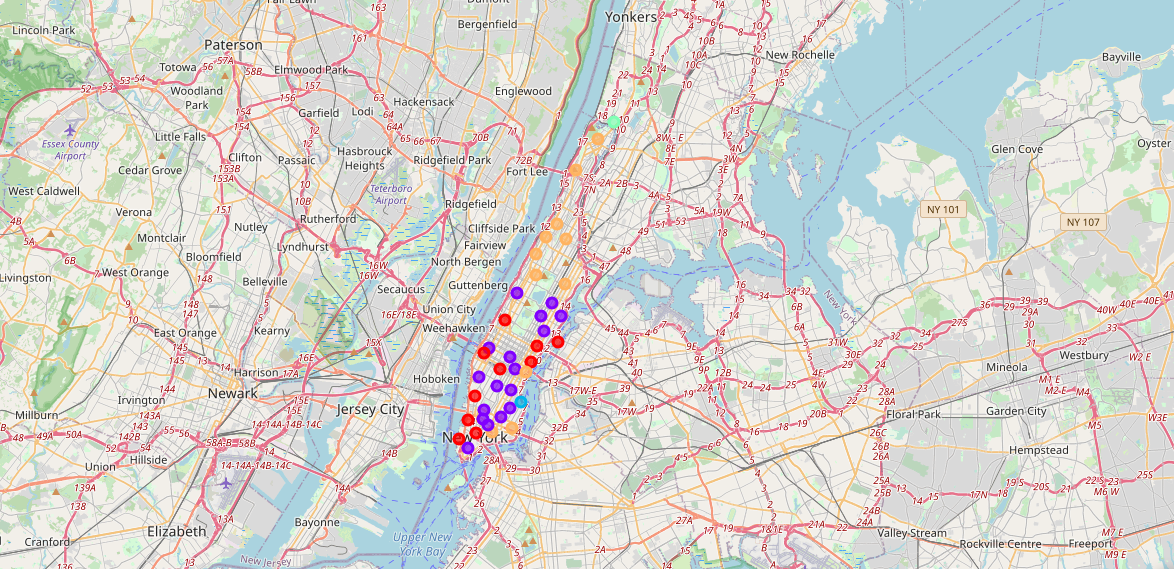
With venues being shown from left to right ordered by their frequency in their corresponding neighborhoods.

Since we don’t have label for our data, in other words, we don’t have our resulting output “y” for the set of features present in the data, an unsupervised method of ML must be used, which Clustering. This algorithm will ease the process of labeling each individual row of data.

Here, a number of 5 clusters was chosen, given the dimension of the data at hand, and by passing the data through the Clustering technique the following results were obtained:



Of course, given the fact that we chose K=5, the obtained Cluster Labels are numbered from 0 to 4 (zero index-based). Then we use Folium package in order to map the obtained labels into their corresponding neighbors throughout New York City, Manhattan.



As you can see, we have obtained 5 separate clusters in 5 different colors, with each of them representing the numbers with the following 10 most common venues:

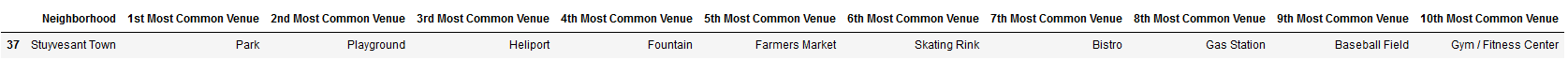
Cluster 1



Cluster 2



Cluster 3



Cluster 4



Cluster 5

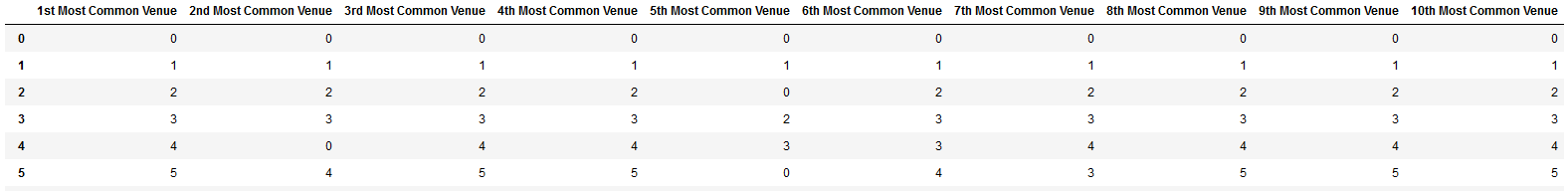


Then the approach was the following:

Since we already have our labels through unsupervised learning we can now proceed to supervised learning. We will use here Classification Techniques learned throughout this course, with our features being, of course, the Columns regarding the most common venues, discarding the Neighborhood column.



Before using this new dataframe as our input variable (y being the cluster labels), we used a method called factorize in order to code each individual venue to a corresponding unique number per column.



Then we trained this data for four separate ML Classification algorithms, such as K Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM) and, finally, Logistic Regression.

In the case of KNN, a relation between train data & test data of 80% and 20%, and the model was trained in order to obtain the K which outputs a greater accuracy, which will be discussed in the results section of this report.

This train-test data split was used also, of course, to train the other ML models and their respective accuracies were measured.

The idea of this trained models is to feed a new dataframe that we’ll create, which features the “ideal cluster label” in which we want to place our restaurant. Then after obtaining the cluster label for our new entry, we would check the neighborhoods with the resulting label and check which entry would be the closest to our new entry.

The test data comprises of one entry which doesn’t feature any restaurant as being at least the most common venue, but with venues that would be great to have right next to the new restaurant.

Test data:



The numbers are according to the factorized feature dataframe.

8 – Park

11 – Clothing Store

11 – Plaza

11 – Coffee Shop

12 – American Restaurant

1 – Spa

20 – Dog Run

7 – Liquor Store

0 – Donut Shop

20 - Thai Restaurant

Again we factorize this data and then use as input for the ML models that we dimensioned.

**Results**

For this new entry, with our models we obtained the following accuracy scores for the 2 main metrics (Jaccard accuracy and F1-score).

|  |  |  |
| --- | --- | --- |
| **ML Techniques Accuracy Scores** | | |
| **Algorithm** | **Jaccard** | **F1-score** |
| KNN | 0.625 | 0.643 |
| Decision Tree | 0.375 | 0.383 |
| SVM | 0.625 | 0.625 |
| Logistic Regression | 0.625 | 0.611 |

As you can see, the KNN algorithm has the best results overall and was the algorithm chosen for obtaining the label for our new entry.

Passing the dataframe through our KNN trained model we obtained the label 0, which corresponds to **Cluster no. 1**:



**Discussion**

Cluster 3, which only has the neighborhood Stuyvesant Town, can be discarded since by just googling location we notice that this is a large private residential development, which are usually not a great ideal place to build your restaurant.



Cluster 4 can also be discarded, but for a different reason. We only have one entry for this particular cluster which, if the model’s chosen label were to be this one, would not be a result giving us great confidence in it.

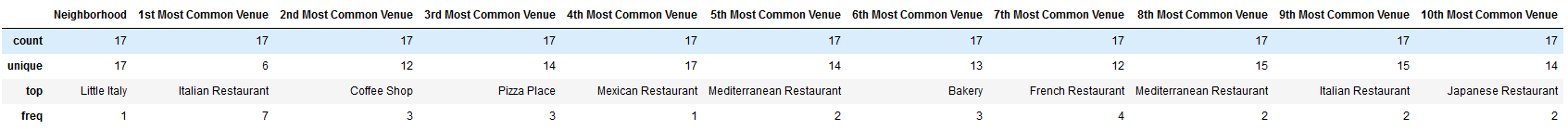
Doing a statistical analysis for Cluster 5 we obtain the following:



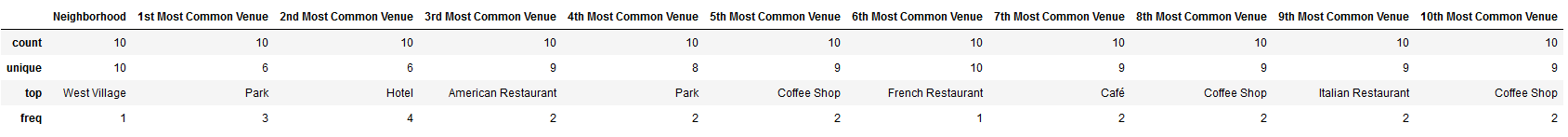
Restaurants are really predominant for this cluster, so it just isn’t a very good candidate.

We are left with cluster 1 and 2.

Cluster 2 has the following statistics:



Which features a great predominance of restaurants with a total number of 20 most common restaurants (not counting the Pizza Places). In contrast with cluster 1, which has the following statistics:



The total number of existing most common restaurants is 5. With the first column (1st Most Common Venue) not even being a restaurant. Even the 2nd one isn’t a restaurant, but a Hotel, which is great for tourists who stay at their hotel and look for a place to eat which is close to the hotel.

Given this statistics, we can state with fairly good confidence that we should place our new restaurant in a neighborhood with Cluster 1.

Looking back at Cluster 1’s dataframe:



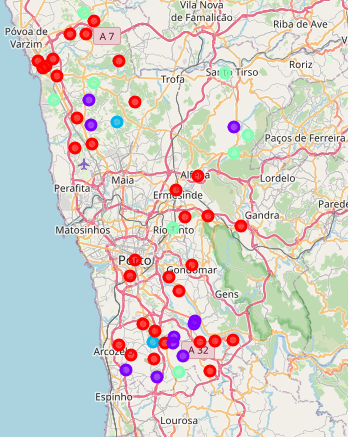
We can already discard neighborhoods Tribeca, West Village, Midtown South, Sutton Place, Turtle Bay and Hudson Yards because they have restaurants at least as their 3rd most common venue. Leaving us with just Roosevelt Island, Lincoln Square, Battery Park City and Civic Center neighborhoods.



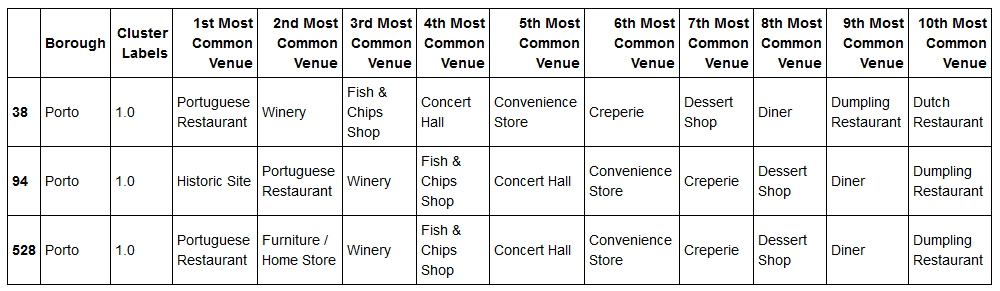
Then in order to choose just one neighborhood, we started to analyze from left to right until we found the first neighborhood with a restaurant as its most common venue, which is Roosevelt Island with a Greek Restaurant as its 6th most common venue.

Again, reading from left to right, we get to the “8th Most Common Venue” column, which has, for both Lincoln Square and Civic Center a restaurant. We eliminate these entries and reach the conclusion that our best choice of a neighborhood is **Battery Park City**, featuring virtually no restaurant as its most common venue, featuring entries such as Park and Plaza which are in the single-entry dataframe that we used as test data.

Now which type of restaurant do we build? For this we looked at several cities like Toronto, Paris and Germany and found that they are somewhat similar in terms of venues, since they’re all multicultural cities (see more in the jupyter notebooks in the links). But I tried to look into my own country, Portugal, regarding a city that I love and with fantastic cuisine, which is Porto.



Clusters such as the one shown in the following dataframe feature Portuguese Restaurants which would certainly be a differentiator in a city such as New York City, which, at least according to the data gathered in this study, seems to lack.



And, being the Portuguese cuisine a very fine cuisine and Battery Park City a neighborhood lacking in restaurants but with places to take a walk and for tourists to spend their nights, it definitely seems like a match made in heaven!

**Conclusion**

There a few limitations with this study. Foursquare’s API doesn’t always return the same results which does not help at all in the analysis, the lack of geographical points to some cities (like Porto for example, in which the entire district of Porto had to be considered and not just the city itself). Also, this approach may not be optimal since the dataframe used as test data was put together from the top of my head, with places that I think should be next to a restaurant.

An American restaurant is in that same dataframe, but I opted for a more realistic approach, since it is difficult to assume that in New York City, a place already filled up to the top with so many restaurants (as discussed in the Introduction section), would have a neighborhood with no restaurants whatsoever. So it is very easy to imply that you’ll probably find an American Restaurant in your neighborhood.

The Thai Restaurant as the last one was chosen since it is usual to have restaurants of foreign cuisine in American neighborhood. However, in this study a Portuguese restaurant was not found!

In pretty much very restaurant-focused data, it is easy to assume these options.

The test data set could be, of course, enhanced in order to feature more entries to check the validity of the model according to the results obtained in this work.

And, also, many more cities could have been researched upon in order to place the new restaurant, however New York City, as it was shown, proved to be a great challenge.