The Search for the Second Manhattan

IBM Capstone Project - Alexandre Vessereau
Date: July 23, 2021

1 Introduction

1.1 Background and interest of the project

For individuals who want to move elsewhere, it is sometimes essential for them to get a similar, or drastically different, from the one they are leaving. The criteria may vary for each individual, but essentially, the goal remains the same: to have insights regarding similar locations to a pre-defined one to compare them and make a more informed choice.

1.2 Description of the problem

Let us imagine that a man named Jim is living in Manhattan, and is planning to move out elsewhere. He considers several locations, such as other New York's boroughs, but also other global cities. To this end, he selects three European cities (Berlin, Rome, Paris) that he likes, as well as two Asian (Tokyo and Shanghai) and one South-American (São Paulo) city. One thing is for sure, Jim loves Manhattan and its *population density*, and *all most popular venues* that compose his borough. Thus, would it be possible to have a recommender system that suggests boroughs located in those cities that share, or at least are similar, to a precise borough? In Jim's case, boroughs similar to his Manhattan?

2 Data acquisition and cleaning

2.1 Data sources

To address the problem, we collected data pertaining to the considered cities. As such, data were scraped from the cities' Wikipedia pages. Pages did not necessarily need to be in English: for instance, the English São Paulo page is not displaying the data we need (population of the subprefectures), but the Portuguese page does.

2.2 Data cleaning

All pages contained two important elements: the borough's names and information regarding their population density. This latter information was lacking for some cities, but other kinds of information, such as the borough's population and area, were present. Thus, population density could be computed according to the following formula:

$$Density = \frac{Population}{Area}$$

Additionally, we used GeoPy to fetch the coordinates (Latitude and Longitude) of each borough of the dataset. If this information was lacking regarding a given borough, then it was dropped from the dataset. We then cleaned each city's dataset and formatted them similarly (e.g., same column names, casting the correct types for some columns).

Once all cities were collected and formatted accordingly, they were grouped under a single dataset.

Furthermore, we needed to add the most popular avenues for each borough of our dataset. To this end, we used the FourSquare API, specifying several parameters in our GET request. We queried 100 venues per request, iterating 10 times to have different batches of results, and sorted by popularity to obtain only the most popular venues per request. Importantly, we did not specify a radius, as this parameter could vary significantly according to the borough. When left unspecified, FourSquare automatically set up a radius suggested based on the density of venues in the area.

2.3 Features preparation and selection

With all the venues categories for each borough of our dataset collected, the first step was to transform them into dummy variables. Then, we grouped venues' categories per borough by averaging them, and added back other boroughs' information such as area or density. Once that was done, we selected only relevant features for our clustering tasks: we only kept the different categories scores and the boroughs' density. Since Density is the quotient of the Population of a borough by its area, we dropped those two features, as they would be redundant otherwise. As such, our dataset for our clustering task looked like the one on the Figure 1, and was ready for the clustering tasks.

	Density	ATM	Abruzzo Restaurant	Acai House	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport
0	12660	0.011111	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
1	5547.3	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
2	3123.8	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
3	21030	0.007246	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
4	12411.1	0.000000	0.0	0.0	0.012048	0.0	0.0	0.0	0.000000
...	...	...	...	...	...	...	...	...	...
137	8718.3	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
138	19803	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.011905
139	21615	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
140	9631	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
141	11910	0.010101	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000

142 rows × 549 columns

Figure 1: Sample of the boroughs' features

3 Methodology

3.1 Clustering similar boroughs

We considered two approaches for our clustering task: Agglomerative Clustering and K-Means clustering. Both have their strengths and weaknesses, thus why they are included in order to compare the outcomes. Note however that since Density and venues' categories scores were on a different scale, we needed to normalize them using a MinMax approach before doing any machine learning task.

3.1.1 Agglomerative Clustering

We first clustered our boroughs using an Agglomerative Clustering approach. We did not specify a pre-defined number of clusters to let the algorithm decides by itself, and evaluated the distance between the groups using the Euclidean distance and a Complete linkage method. With the clusters obtained, the corresponding dendrogram was subsequently plotted to visualize similar boroughs.

3.1.2 K-Means Clustering

The approach of the K-Means differs from Agglomerative Clustering in that, when using K-Means, it is generally strongly advised to check for the number of clusters to use beforehand. To achieve this, we used two metrics: the elbow method and the silhouette score. The elbow method helps to visually determine the number of clusters by spotting "where" an elbow appears for an increasing number of clusters. As for the silhouette method, it indicates how similar an item is to its own cluster compared to others, again for an increasing number of clusters. The result for the elbow method can be visualized in Figure 2, whereas the result of Silhouette can be seen in Figure 3.

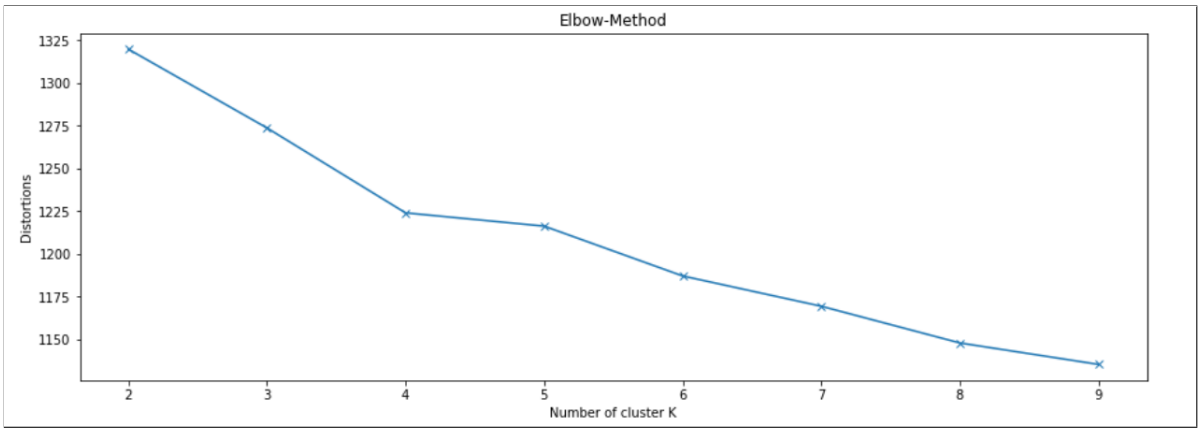

Figure 2: Elbow-Method for the different K of boroughs clustering

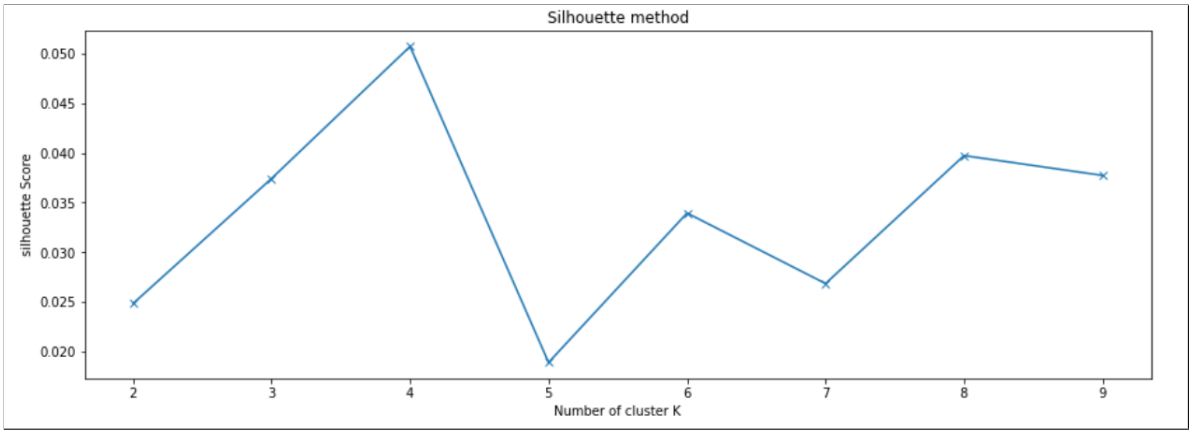

Figure 3: Silhouette scores for the different K of boroughs clustering

As shown by the results of the two plots above, it seems that the right number of clusters for our boroughs' dataset is 4, and as such, we ran our K-Means algorithm with

this number. Note that we had to reduce our dimensions to plot them, as it would be impossible to do otherwise. Accordingly, we used a TSNE (standing for t-distributed stochastic neighbour embedding) to plot our clusters on a 2-D graph.

4 Results

4.1 Agglomerative clustering results

Our results for the Agglomerative clustering can be seen on its corresponding dendrogram plot in Figure 4. As we look at boroughs similar to Manhattan, we see that three boroughs might be of interest: Mitte, in Berlin (Germany), Butantã, in São Paulo (Brazil) and Taito, in Tokyo (Japan). Let us see if those results hold in the K-Means clustering.

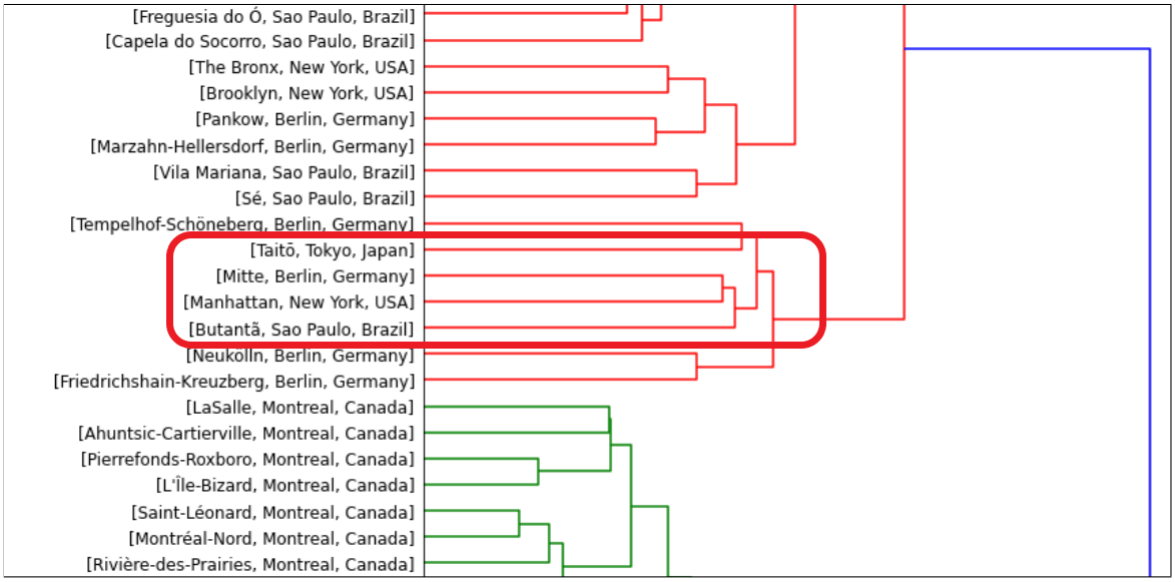

Figure 4: Dendrogram of boroughs' agglomerative clustering

4.2 K-Means clustering results

Results for the K-Means clustering and its subsequent TSNE transformation can be found on its corresponding plot, in Figure 5.

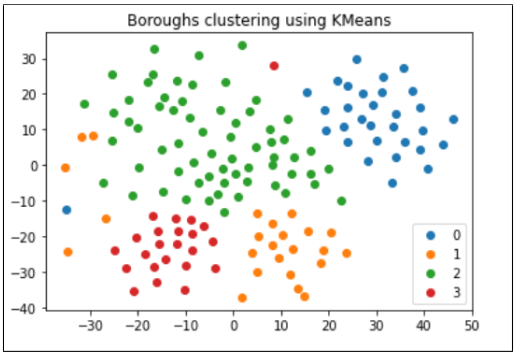

Figure 5: K-Means and TSNE boroughs clustering

However, it is not directly apparent to which clusters the boroughs belonged to. As such, a function that took as input a given borough was written, and was returning the borough's cluster number and the other boroughs also belonging to the cluster. We ran the function with the three boroughs obtained in the Agglomerative Clustering, and displayed their corresponding clusters in Figure 6. As a result, we see that only Mitte in Berlin, Germany belonged to the same cluster as Manhattan.

```
Manhattan belongs to the cluster n° 2
```

	Borough	Population	Density	Area	Latitude	Longitude	country	city	Cluster	1st Most Common Venue	2nd Most Common Venue
0	Ahuntsic-Cartierville	134245	5547.3	24.2	45.541892	-73.680319	Canada	Montreal	2	Café	Pharmacy
1	Anjou	42796	3123.8	13.7	45.604898	-73.546672	Canada	Montreal	2	Restaurant	Fast Food Restaurant
2	Brooklyn	2559903	13957	183.42	40.650104	-73.949582	USA	New York	2	Caribbean Restaurant	Pizza Place
3	Bǎoshān Qū	2042300	7536	270.99	31.406634	121.485158	China	Shangai	2	Coffee Shop	Shopping Mall
4	Charlottenburg-Wilmersdorf	319628	4878	64.72	52.507856	13.263952	Germany	Berlin	2	Café	German Restaurant

```
Boroughs similar to Mitte
Mitte belongs to the cluster n° 2
```

	Borough	Population	Density	Area	Latitude	Longitude	country	city	Cluster	1st Most Common Venue	2nd Most Common Venue
0	Ahuntsic-Cartierville	134245	5547.3	24.2	45.541892	-73.680319	Canada	Montreal	2	Café	Pharmacy
1	Anjou	42796	3123.8	13.7	45.604898	-73.546672	Canada	Montreal	2	Restaurant	Fast Food Restaurant
2	Brooklyn	2559903	13957	183.42	40.650104	-73.949582	USA	New York	2	Caribbean Restaurant	Pizza Place
3	Bǎoshān Qū	2042300	7536	270.99	31.406634	121.485158	China	Shangai	2	Coffee Shop	Shopping Mall
4	Charlottenburg-Wilmersdorf	319628	4878	64.72	52.507856	13.263952	Germany	Berlin	2	Café	German Restaurant

```
Boroughs similar to Butantã, Brazil
Butantã belongs to the cluster n° 0
```

Figure 6: Boroughs' clusters belonging for Manhattan, Mitte and Butantã

5 Discussion

The initial problem that was posed at the beginning was simple. A man named Jim needed to get recommended boroughs similar to Manhattan that were close to both its density and its most popular places. To address this problem, we ran two different clustering approaches, namely Agglomerative and K-Means clustering, in order to compare the outcomes.

Interesting results were found in the two methods. For instance, findings showed that boroughs such as Mitte in Germany, or Butantã, in Brazil, could potentially be similar to Manhattan, and therefore of interest to Jim. More generally, this recommendation system could be of general interest: for any individual, or even families, discovering if it exists similar places to the one they are leaving, in the form of infrastructure or population, could ease their choice of relocation.

Of course, this report is not without any limitations. For instance, the silhouette score was relatively low, thus indicating that clusters were not well separated. As such, our model could incorporate other variables when clustering boroughs: features such as the education or crime level are indeed quite relevant when choosing a future place to live. Furthermore, we focused here only on boroughs, as census data regarding more specific subdivisions (e.g., districts, neighbourhoods) were missing. Using those subdivisions, instead of boroughs, could perhaps create more fine-tuned recommendations in the clustering process.

6 Conclusion

In this report, we sought to create boroughs recommendations according to a pre-defined one, based on their density similarity and their most popular venues' category. We used two clustering approaches to achieve this goal: an Agglomerative and K-Means clustering. Using those clustering, we are now able to recommend similar boroughs to a pre-defined one. Nevertheless, further research is required to better fine-tune this model. Consequently, incorporating more of the boroughs' features and using more precise subdivisions (districts) could be interesting avenues to pursue.