The Search for the Second Manhattan

IBM Capstone Project

Alexandre Vessereau

Problem description

- A man named Jim wants to move out elsewhere
- He's planning to move out to:
  - a European city (Berlin, Paris or Rome)
  - Asian city (Shanghai, China, or Tokyo, Japan)
  - A South American city (Sao Paulo, Brazil).
- Problem is, he would like to go to one of the boroughs of the cities mentioned that is the most similar to Manhattan.
- He wants his future borough to have **a density and venues similar to Manhattan.**

Data Sources and cleaning

- We scraped relevant data (boroughs data) from the cities' Wikipedia pages.
- We cleaned the datasets and computed borough density if it was lacking.
- We used the GeoPy API to fetch each borough coordinates (latitude and longitude)
- We formatted each dataset similarly and then grouped them under a single dataset
- In total, the dataset had 142 rows for 7 columns.

Adding boroughs' venues

- We used the FourSquare API to fetch the most popular venues for each borough
- We turned each venue's categories into a dummy variable
- We then grouped venues by borough by averaging them
- A dataset containing only boroughs' density and averaged venues categories was then created
- Data was then normalized using a MinMax approach

	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

Methodology

Clustering boroughs

- Two clustering approaches were considered to cluster similar boroughs
- Agglomerative and K-Means clustering
- We used those two methods in order to compare the results

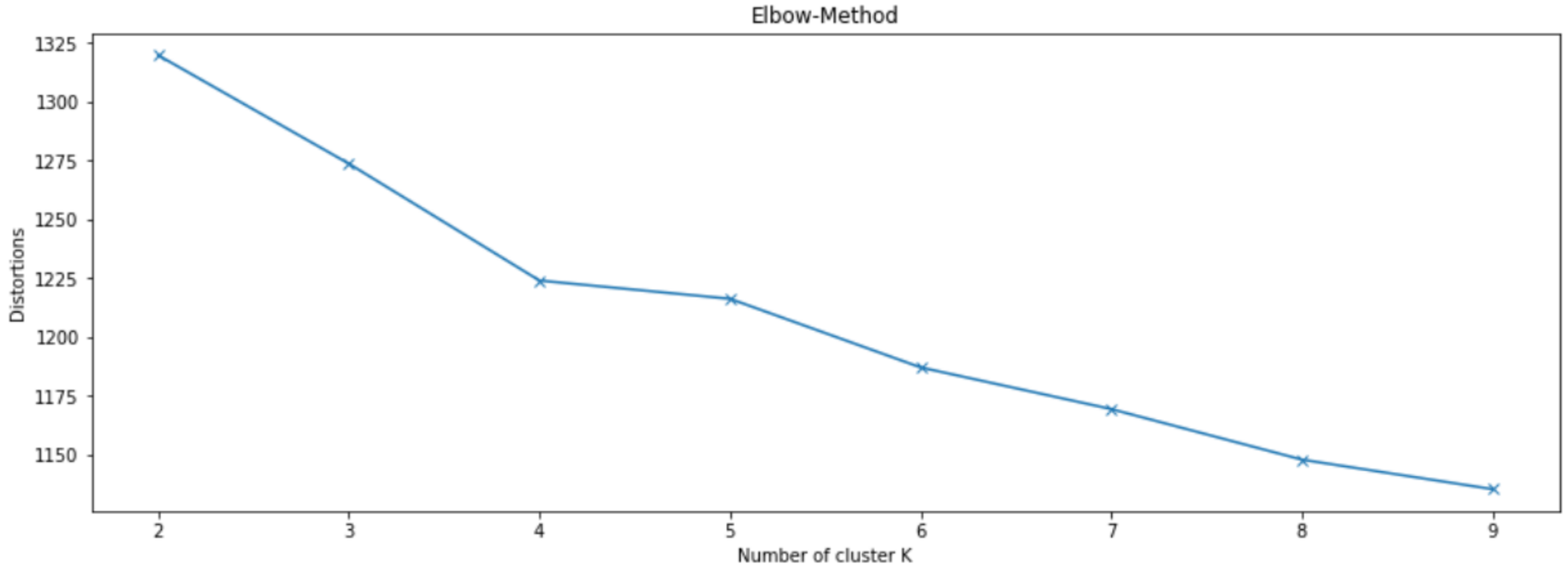
Agglomerative clustering

- In Agglomerative clustering, there is no need to pre-define the number of clusters
- Clusters were evaluated using the Euclidean distance and a complete linkage method.
- A dendrogram was produced to visualize the clusters

K-Means clustering

- For K-Means clustering, we need to find the correct number of cluster
- Two methods were used to do so: the elbow and the silhouette method
- Using those two metrics, we see that the correct number of clusters appears to be 4.

Elbow Method

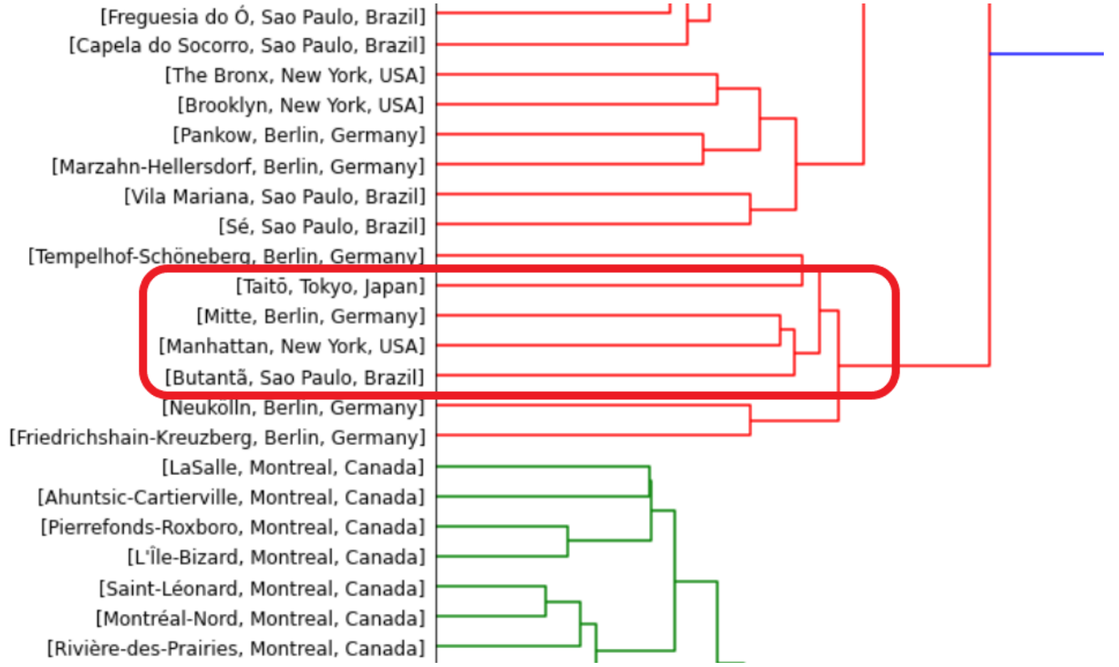


Silhouette Method

Results

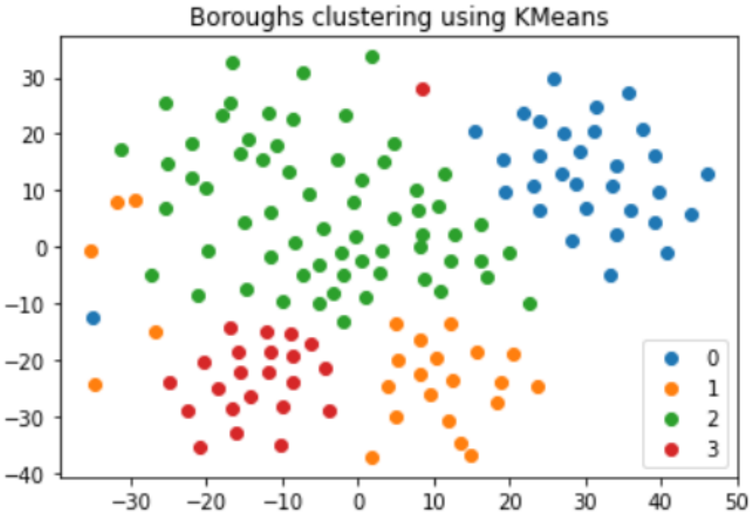
Agglomerative clustering results

- Our initial questioning was to find a borough similar to Manhattan.
- Based on the results of the agglomerative clustering on the dendrogram plot, we can say that three boroughs are similar to Manhattan:
  - Mitte in Berlin, Germany,
  - Butantã, in São Paulo, Brazil
  - Taito, in Tokyo, Japan



K-Means clustering results

- Regarding K-Means, we needed to transform the number of dimensions, as it would be impossible to plot otherwise
- We used a TSNE approach (t-distributed stochastic neighbour embedding) to do so
- We also wrote a function that returns similar boroughs to the one in input (i.e.: belonging to the same cluster)



K-Means clustering results

- Entering the boroughs' obtained in the Agglomerative clustering yields the following results
- We see that only Mitte, in Berlin, 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	Bāoshān 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	Bāoshān 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

Discussion & Conclusion

The Second Manhattan?

- In this report, the goal was to find similar boroughs to Manhattan, sharing similar venues and density for a man named Jim
- Using an agglomerative and K-Means clustering approach, we see that Mitte, in Germany, could be a good choice for Jim

Insights of the project

- The findings of this report are not limited to Jim
- A model that recommends similar (or dissimilar) boroughs to a specified one could be of interest for any individual/families looking to relocate in a different city, to ease their choice.

Limitations and further avenues

- This project is not without limitations
- Using more census data (average education attainment of a borough) could improve the clustering process
- Similarly, using more precise subdivisions (e.g.: districts) could propose a more fine-tune recommendation system
- As such, this project offers interesting avenues to pursue

Thank you !

Manhattan, New York

Mitte, Berlin