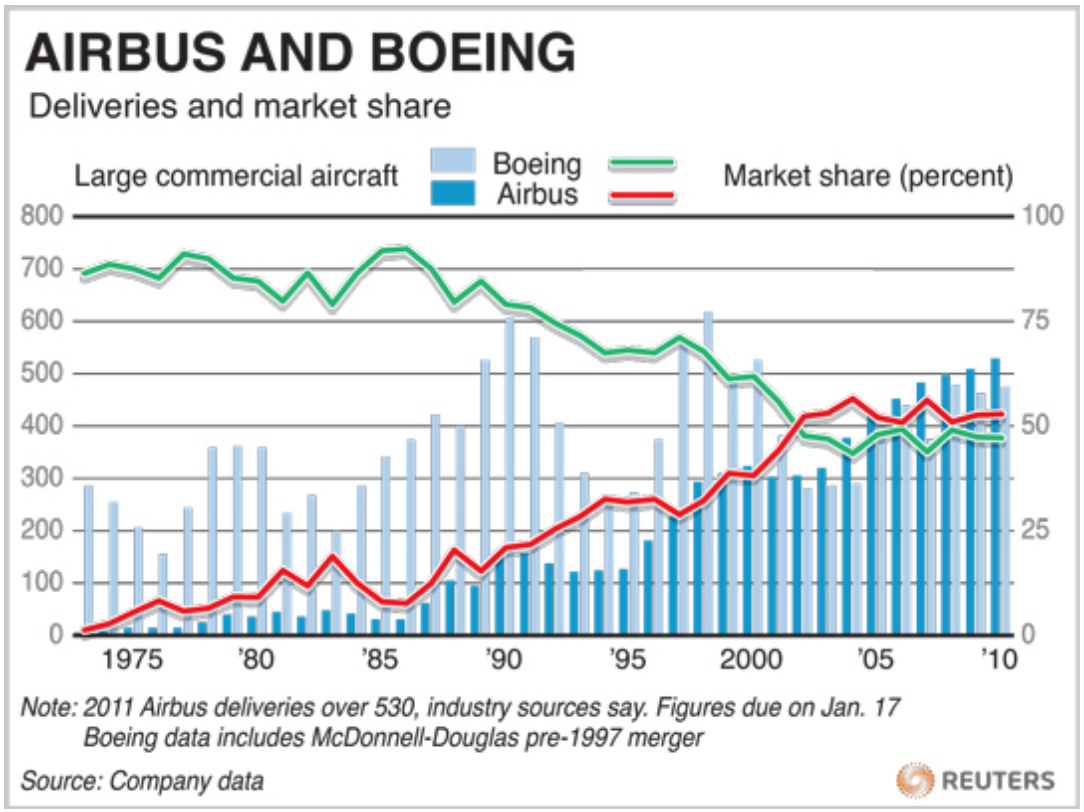
Toulouse Vs Seattle

Stéphane MARTINEZ
April 28th, 2019
Coursera Applied Data Science Capstone Assignment

Introduction

Background

Airbus and Boeing are the two leaders of the commercial aircraft industry. Each one is sharing about 50% of this market.



Both competitors are fighting to be the One, designing and delivering the best aircraft regarding market acceptance. Aeronautics Industry is more than these two companies. A dense network of subcontractors has grown to leverage and support the industry.

Boeing biggest factory is based at Seattle (USA):

Airbus' one in Toulouse (France).

Problem

It will be interesting to compare the two cities, located on two distinct continents, to see if they have similar characteristics, and if the long distance in between can be compensated by the major industry that live there.

Interest

In a concrete way, the two competitors need and attractive city background to retain best talents. Results from this study should also be used to discuss with the respective cities governance some action plan to improve some neighborhood services offer.

Data acquisition and cleaning

Data sources

To consider the problem we can list the data as below:

- First, we can find detailed information on cities neighborhoods from wikipedia, for Seattle (https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Seattle) and for Toulouse (https://fr.wikipedia.org/wiki/Quartiers_de_Toulouse).
- Then we can use GeoPy (https://geopy.readthedocs.io/en/stable/) to localize these neighborhoods.
- The foursquare API (https://developer.foursquare.com/) can be used to gather venues information.

All these information can be processed to analyze the two cities and to see if they sharing to the same structure or not.

Data cleaning

Seattle's and Toulouse neighborhoods from Wikipedia need to be preprocessed to build a clean dataset, including :

- Removing page break
- Removing text marks
- Removing aliases in parentheses
- Adding City's Name
- Separating neighborhood list to distinct dataset rows
- Filtering filtered to discard non geolocalized data

Feature Selection

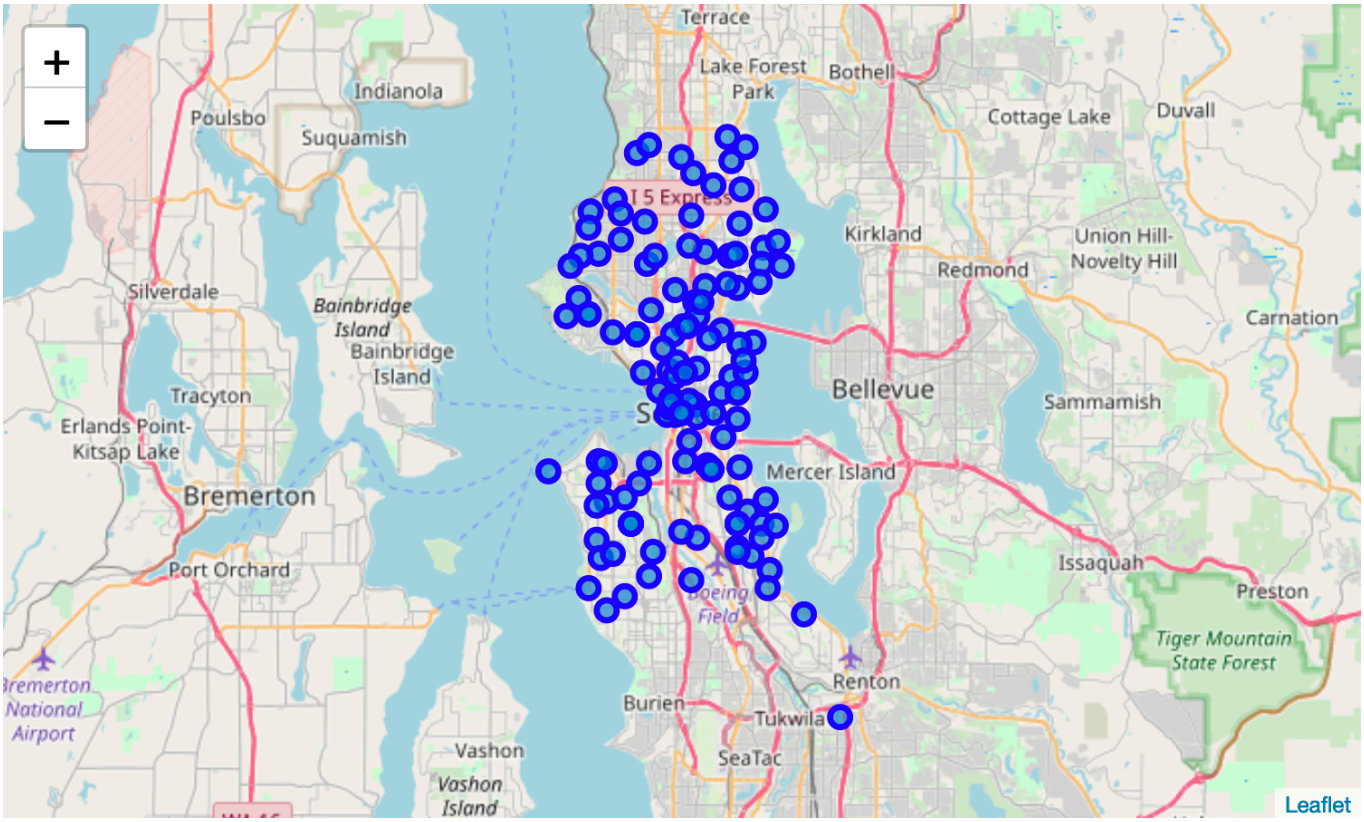
Then, the Foursquare API is used to gather venues for each geolocalized neighborhood. Venues are hot encoded, grouped by neighborhood, and the selected features correspond to the top 10 venues for each neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alaska Junction , Seattle	Pizza Place	Coffee Shop	Spa	Asian Restaurant	Bakery	Lounge	Beer Store	Supermarket	Steakhouse	Furniture / Home Store
1	Blue Ridge , Seattle	Dance Studio	Café	Garden Center	Pool	Zoo Exhibit	Fish Market	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field
2	Central District , Seattle	Coffee Shop	Performing Arts Venue	Chinese Restaurant	Water Park	Gym	BBQ Joint	Fish & Chips Shop	Art Gallery	Vietnamese Restaurant	Ethiopian Restaurant
3	Denny-Blaine, Seattle	Park	Monument / Landmark	Beach	Zoo Exhibit	Fish Market	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field	Filipino Restaurant
4	Lakeridge , Seattle	Park	Pizza Place	Playground	Fish & Chips Shop	Factory	Fair	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Field

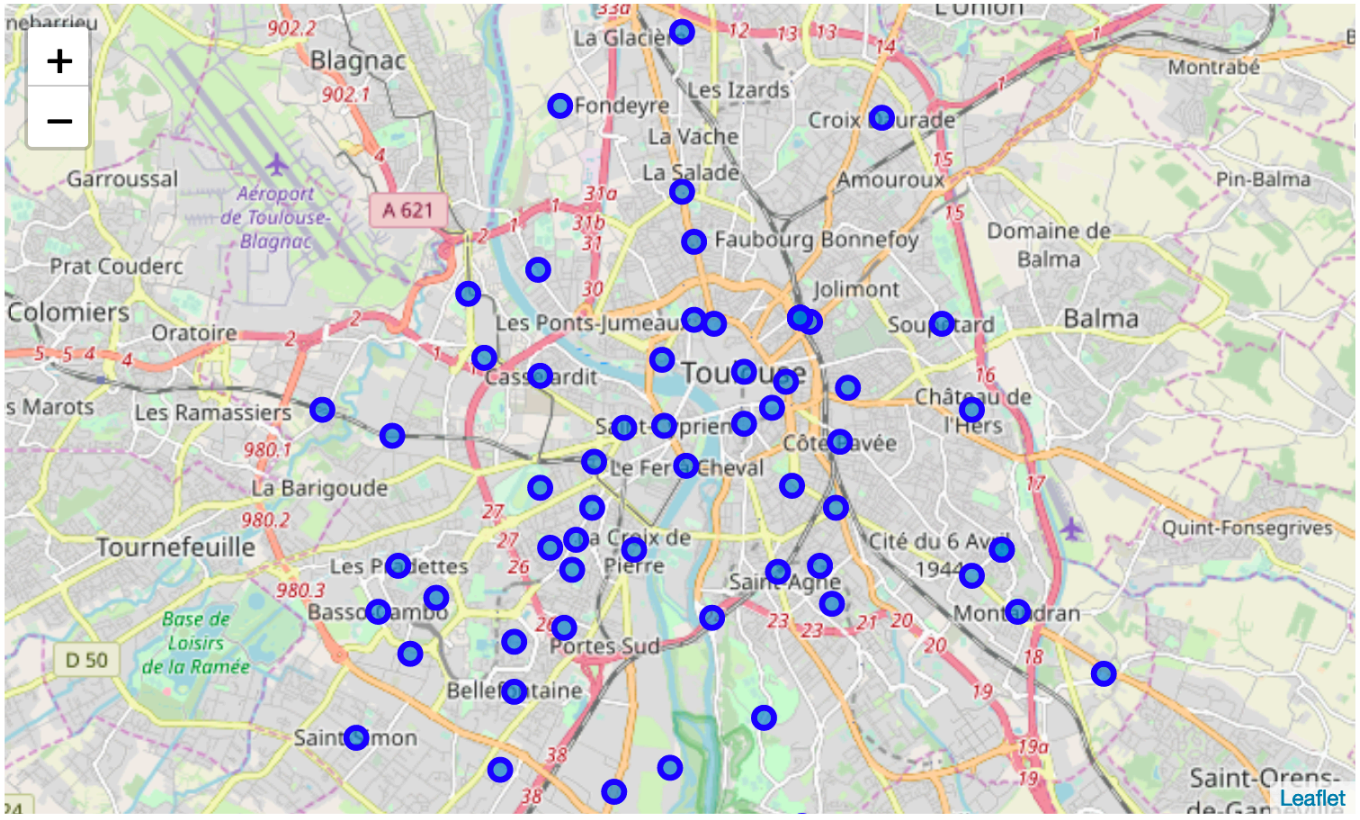
Methodology

Exploratory Analysis

After data cleaning, there were 132 geolocalized neighborhoods in the Seattle's data, and 62 in the Toulouse's data.
First check is to plot neighborhoods on the map of the two cities:



Seattle, Neighborhoods without clustering



Toulouse, Neighborhoods without clustering

To determine whether the two cities are similar or not, neighborhoods are analyzed using two different clustering methods.

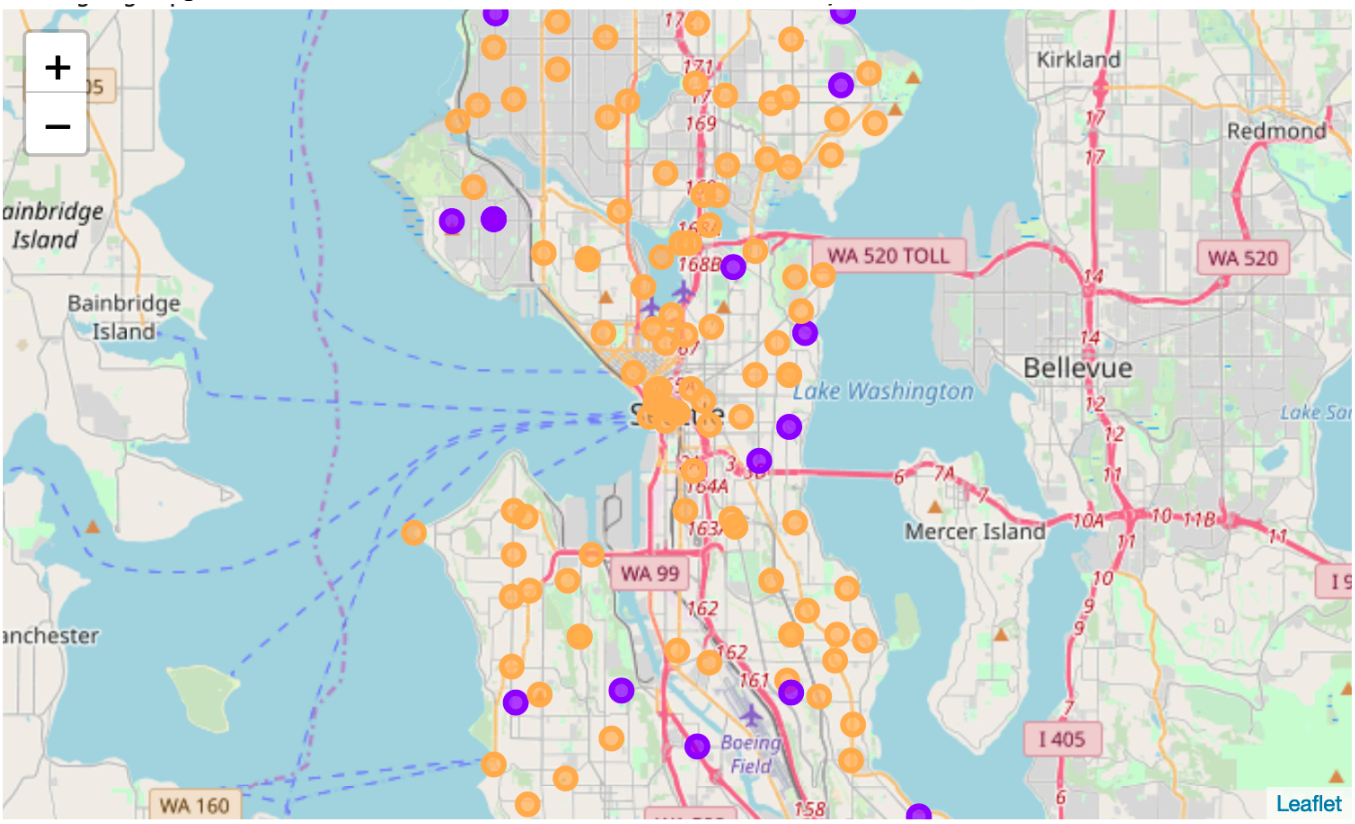
1. The first one is the k-means clustering, using 5 clusters

2. The second one is the Agglomerative Clustering, limited to three levels

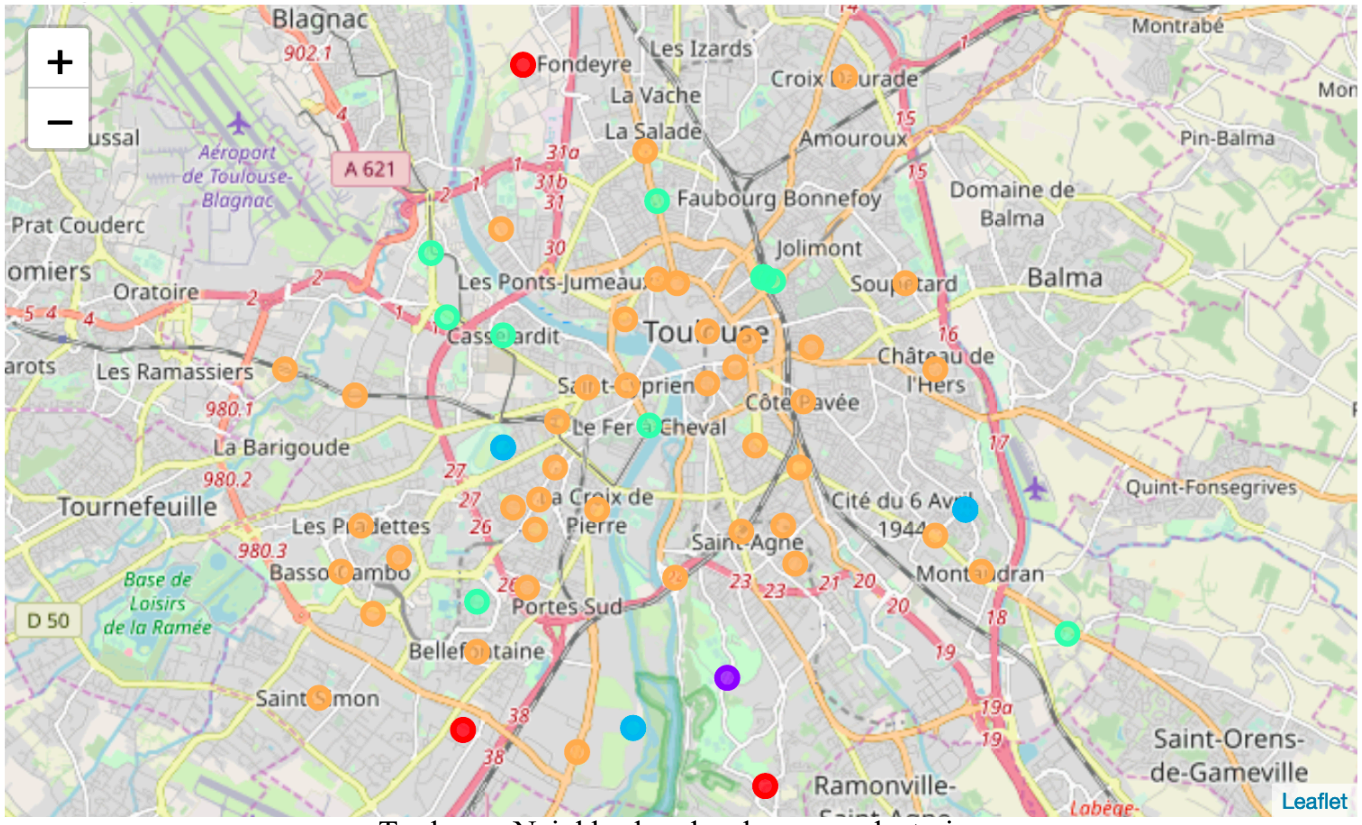
Results

Neighborhood k-means clustering

The clustered neighborhoods can be represented in the cities' maps. Color for a given cluster is the same in the both maps.



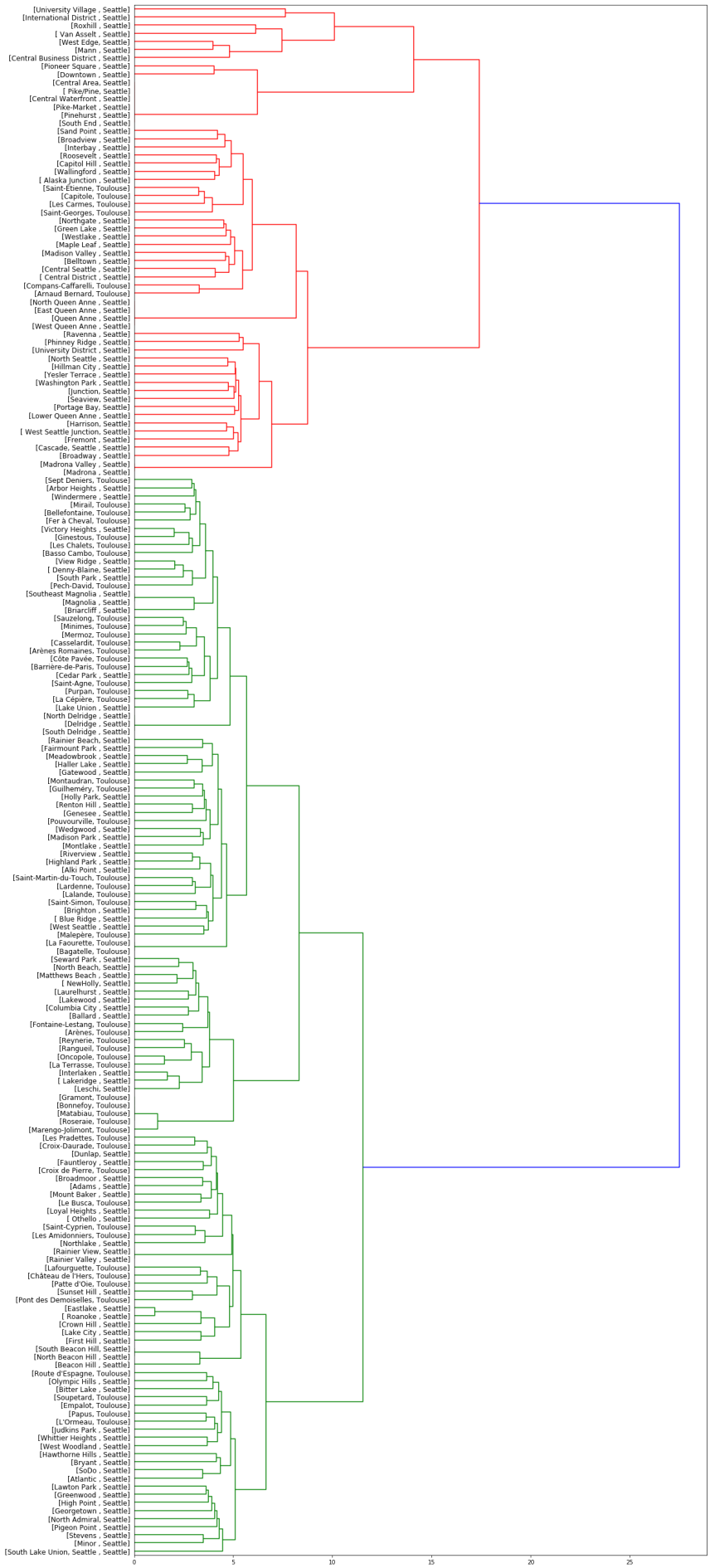
Seattle, Neighborhoods – k-means clustering



Toulouse, Neighborhoods – k-means clustering

Neighborhood Agglomerative clustering

This second approach is taking the problem by the other side, trying not to divide the original pool of neighborhoods, but trying to group neighborhoods based on their similarities.


[University Village , Seattle]
[International District , Seattle]
[Roxhill , Seattle]
[Van Asselt , Seattle]
[West Edge, Seattle]
[Mann , Seattle]
[Central Business District , Seattle]
[Pioneer Square , Seattle]
[Downtown , Seattle]
[Central Area, Seattle]
[Pike/Pine, Seattle]
[Central Waterfront , Seattle]
[Pike-Market , Seattle]
[Pinehurst , Seattle]
[South End , Seattle]
[Sand Point , Seattle]
[Broadview , Seattle]
[Interbay , Seattle]
[Roosevelt , Seattle]
[Capitol Hill , Seattle]
[Wallingford , Seattle]
[Alaska Junction , Seattle]
[Saint-Etienne, Toulouse]
[Capitole, Toulouse]
[Les Carmes, Toulouse]
[Saint-Georges, Toulouse]
[Northgate , Seattle]
[Green Lake , Seattle]
[Westlake , Seattle]
[Maple Leaf , Seattle]
[Madison Valley , Seattle]
[Belltown , Seattle]
[Central Seattle , Seattle]
[Central District , Seattle]
[Compans-Caffarelli, Toulouse]
[Arnaud Bernard, Toulouse]
[North Queen Anne , Seattle]
[East Queen Anne , Seattle]
[Queen Anne , Seattle]
[West Queen Anne , Seattle]
[Ravenna , Seattle]
[Phinney Ridge , Seattle]
[University District , Seattle]
[North Seattle , Seattle]
[Hillman City , Seattle]
[Yesler Terrace , Seattle]
[Washington Park , Seattle]
[Junction, Seattle]
[Seaview, Seattle]
[Portage Bay, Seattle]
[Lower Queen Anne , Seattle]
[Harrison, Seattle]
[West Seattle Junction, Seattle]
[Fremont , Seattle]
[Cascade, Seattle , Seattle]
[Broadway , Seattle]
[Madrona Valley , Seattle]
[Madrona , Seattle]
[Sept Deniers, Toulouse]
[Arbor Heights , Seattle]
[Windermere , Seattle]
[Mirail, Toulouse]
[Bellefontaine, Toulouse]
[Fer à Cheval, Toulouse]
[Victory Heights , Seattle]
[Ginestous, Toulouse]
[Les Chalets, Toulouse]
[Basso Cambo, Toulouse]
[View Ridge , Seattle]
[Denny-Blaine, Seattle]
[South Park , Seattle]
[Pech-David, Toulouse]
[Southeast Magnolia , Seattle]
[Magnolia , Seattle]
[Briarcliff , Seattle]
[Sauzelong, Toulouse]
[Minimes, Toulouse]
[Mermoz, Toulouse]
[Casselardit, Toulouse]
[Arènes Romaines, Toulouse]
[Côte Pavée, Toulouse]
[Barrière-de-Paris, Toulouse]
[Cedar Park , Seattle]
[Saint-Agne, Toulouse]
[Purpan, Toulouse]
[La Cépière, Toulouse]
[Lake Union , Seattle]
[North Delridge , Seattle]
[Delridge , Seattle]
[South Delridge , Seattle]
[Rainier Beach, Seattle]
[Fairmount Park , Seattle]
[Meadowbrook , Seattle]
[Haller Lake , Seattle]
[Gatewood , Seattle]
[Montaudran, Toulouse]
[Guilhemery, Toulouse]
[Holly Park, Seattle]
[Renton Hill , Seattle]
[Genesee , Seattle]
[Pouvourville, Toulouse]
[Wedgwood , Seattle]
[Madison Park , Seattle]
[Montlake , Seattle]
[Riverview , Seattle]
[Highland Park , Seattle]
[Alki Point , Seattle]
[Saint-Martin-du-Touch, Toulouse]
[Lardenne, Toulouse]
[Lalande, Toulouse]
[Saint-Simon, Toulouse]
[Brighton , Seattle]
[Blue Ridge , Seattle]
[West Seattle , Seattle]
[Malepère, Toulouse]
[La Faourette, Toulouse]
[Bagatelle, Toulouse]
[Seward Park , Seattle]
[North Beach, Seattle]
[Matthews Beach , Seattle]
[NewHolly, Seattle]
[Laurelhurst , Seattle]
[Lakewood , Seattle]
[Columbia City , Seattle]
[Ballard , Seattle]
[Fontaine-Lestang, Toulouse]
[Arènes, Toulouse]
[Reynerie, Toulouse]
[Rangueil, Toulouse]
[Oncopole, Toulouse]
[La Terrasse, Toulouse]
[Interlaken , Seattle]
[Lakeridge , Seattle]
[Leschi, Seattle]
[Gramont, Toulouse]
[Bonnefoy, Toulouse]
[Matabiau, Toulouse]
[Roseraie, Toulouse]
[Marengo-Jolimont, Toulouse]
[Les Pradettes, Toulouse]
[Croix-Daurade, Toulouse]
[Dunlap, Seattle]
[Fauntleroy , Seattle]
[Croix de Pierre, Toulouse]
[Broadmoor , Seattle]
[Adams , Seattle]
[Mount Baker , Seattle]
[Le Busca, Toulouse]
[Loyal Heights , Seattle]
[Othello , Seattle]
[Saint-Cyprien, Toulouse]
[Les Amidonniers, Toulouse]
[Northlake , Seattle]
[Rainier View, Seattle]
[Rainier Valley , Seattle]
[Lafourguette, Toulouse]
[Château de l'Hers, Toulouse]
[Patte d'Oie, Toulouse]
[Sunset Hill , Seattle]
[Pont des Demoiselles, Toulouse]
[Eastlake , Seattle]
[Roanoke , Seattle]
[Crown Hill , Seattle]
[Lake City , Seattle]
[First Hill , Seattle]
[South Beacon Hill, Seattle]
[North Beacon Hill , Seattle]
[Beacon Hill , Seattle]
[Route d'Espagne, Toulouse]
[Olympic Hills , Seattle]
[Bitter Lake , Seattle]
[Soupetard, Toulouse]
[Empalot, Toulouse]
[Papus, Toulouse]
[L'Ormeau, Toulouse]
[Judkins Park , Seattle]
[Whittier Heights , Seattle]
[West Woodland , Seattle]
[Hawthorne Hills , Seattle]
[Bryant , Seattle]
[SoDo , Seattle]
[Atlantic , Seattle]
[Lawton Park , Seattle]
[Greenwood , Seattle]
[High Point , Seattle]
[Georgetown , Seattle]
[North Admiral, Seattle]
[Pigeon Point , Seattle]
[Stevens , Seattle]
[Minor , Seattle]
[South Lake Union, Seattle , Seattle]
0
5
10
15
20
25

Discussion

First result we can discuss is that on the k-means clustering part, the 5 clusters don't have similar size, whatever the city:

Cluster	Number of neighborhoods
1	4
2	17
3	4
4	13
5	154

Second interesting result is that for both cities, Seattle and Toronto, the major part of the neighborhoods is part of cluster #5.

Then, the 4 remaining clusters seems to be specific to one city:

- Cluster 1 → Toulouse
- Cluster 2 → Seattle
- Cluster 3 → Toulouse
- Cluster 4 → Toulouse

Looking at the maps, we can also see the two cities have different schemas:

- For Seattle, we can clearly see a core of cluster 5 (orange points), with around, a belt of cluster 2.
- For Toulouse, all the clusters seem to be randomly plotted.

Based on the dendrogram (agglomerative clustering), we can see similar results:

- In the red branch, we can see clearly that the firsts levels or grouping mainly consider neighborhoods from the same city
- In the green one, mix between neighborhoods coming from different cities is more important. And the main part of all the neighborhoods are linked to the green branch

Conclusion

Based on two different machine learning techniques based on neighborhood venues analysis, **we can conclude that Seattle and Toulouse have very similar neighborhoods**, because the greatest part of the neighborhoods are in the same cluster (cluster #5).

But for a few numbers of neighborhoods, we can see specificities linked to the city (Seattle for cluster #2 and Toulouse for clusters #1, #3 and #4). Airbus and Boeing should focus on these ones to determine what is the added value of these neighborhoods. If the added value is

- positive, they could discuss city governance to find ways to improve other neighborhoods venues,
- negative, they could discuss city governance to find ways to improve these neighborhoods venues