The Battle of the Neighborhoods

Where to open a movie theater in Toulouse (F)

This report is submitted as part of the evaluation for the IBM Data Science Professional Certificate.

Code can be found on the following Github repository: link.

1. Introduction and Business Problem

we imagine a scenario in which an investor wishes to find a suitable place for installing a **new cinema** in the city of Toulouse in France. This person is a foreign investor and it is therefore central that a detailed analysis of the city is provided. If the person would be a local, the importance of identifying similar regions in a city would not be that central as the local person normally already knows more or less which place would be more adequate. In fact, it seems that the location of a cinema is far from being straight forward. Indeed, unlike a cafe place or a restaurant, it is less common and its location certainly belongs to its most important features.

An interesting element in Toulouse is that all movie theaters are located on the east-side of the city center. Therefore, to simply implement a new one in the same areas may be ill-advised. Therefore, we wish to identify a new location for a movie theater that is similar to those locations where we already have existing movie theaters. To do so, we will be using an unsupervised classification method called K-Mean clustering.

Accordingly, we identify the following requirements:
* the cinema should be in an area within walking distance of a metro station.
* the cinema should be in an area that encompasses other relevant businesses (restaurant, bars, etc.).
* the cinema should be in an area that is densely populated.

From the above-mentioned requirements, we identify the following needs in terms of **data**:
* we need to cluster the metro station and find the type of venues in a radius of 500m from the metro station.
* we need to find census data about the number of people living in which area.

2. Data

Based on the requirements mentioned above, the source of the data will be as follow:

* location of **metro station**: we will create a first dataframe that will consist in the name of the metro station, the line as well as its coordinates (longitude and latitude). https://data.toulouse-metropole.fr/explore/dataset/stations-de-metro/export/

* location of **existing movie theater:** using the foursquare API (search endpoint) we will be identifying all existing movie theaters as to later find a suitable place for our new one.

* **census data**: Using this information , we will be able to identify areas that are more densed that other, thus maybe suggesting that some parts of the city are more likely to be suitable for the establishment of a new movie theater: https://data.toulouse-metropole.fr/explore/dataset/-recensement-population-2016-grands-quartiers-population/information/

* **Foursquare API** will also provide us with the required data of the venues surrounding metro stations. We will be using both the **explore** and the **search** endpoint**s.**

* Coordinate of Toulouse will be found by using the Nominatim geocoder.

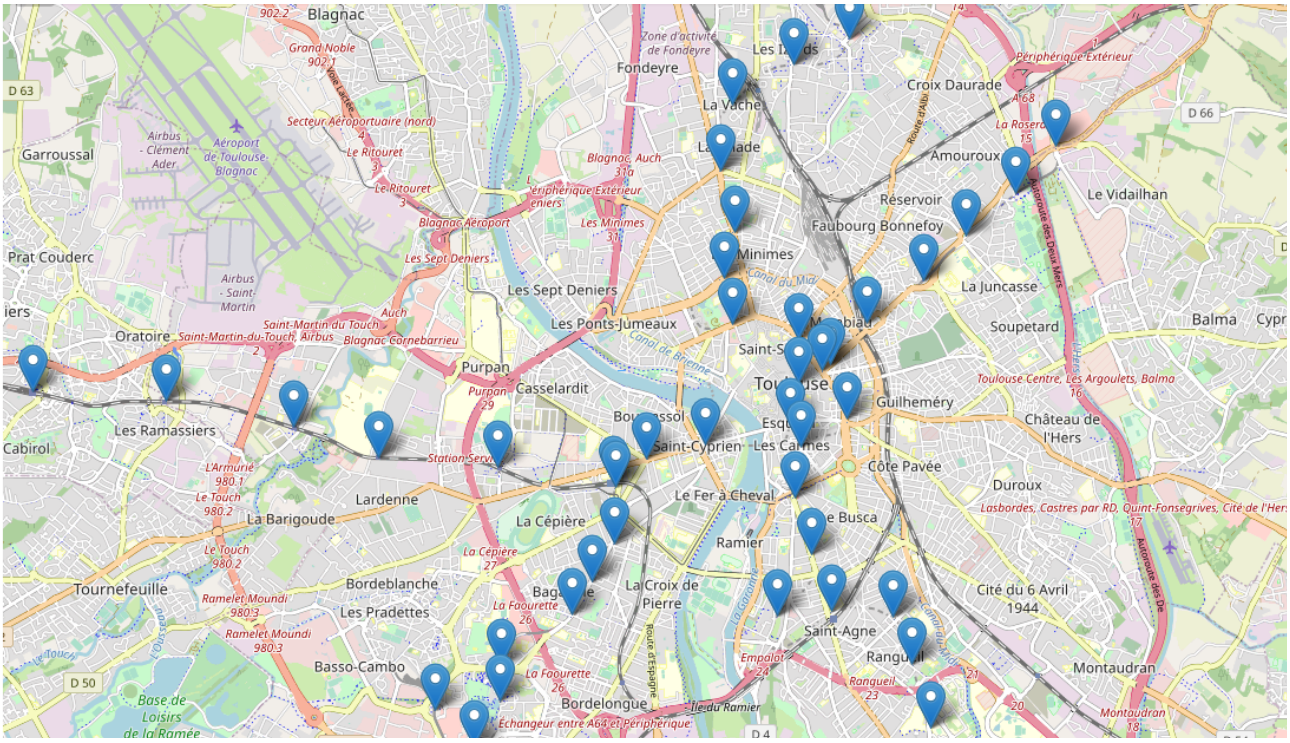
(note that additional census data can be found on the same website)

3. Methodology

3.1 Get Coordinates for all metro stations

We first create a map using FOLIUM and add the location of all metro stations; this information is downloaded from https://data.toulouse-metropole.fr/explore/dataset/stations-de-metro/information/

There are 43 metros stations across 3 different metro lines.



Information about all metro stations are saved in a dataframe looking as follow:

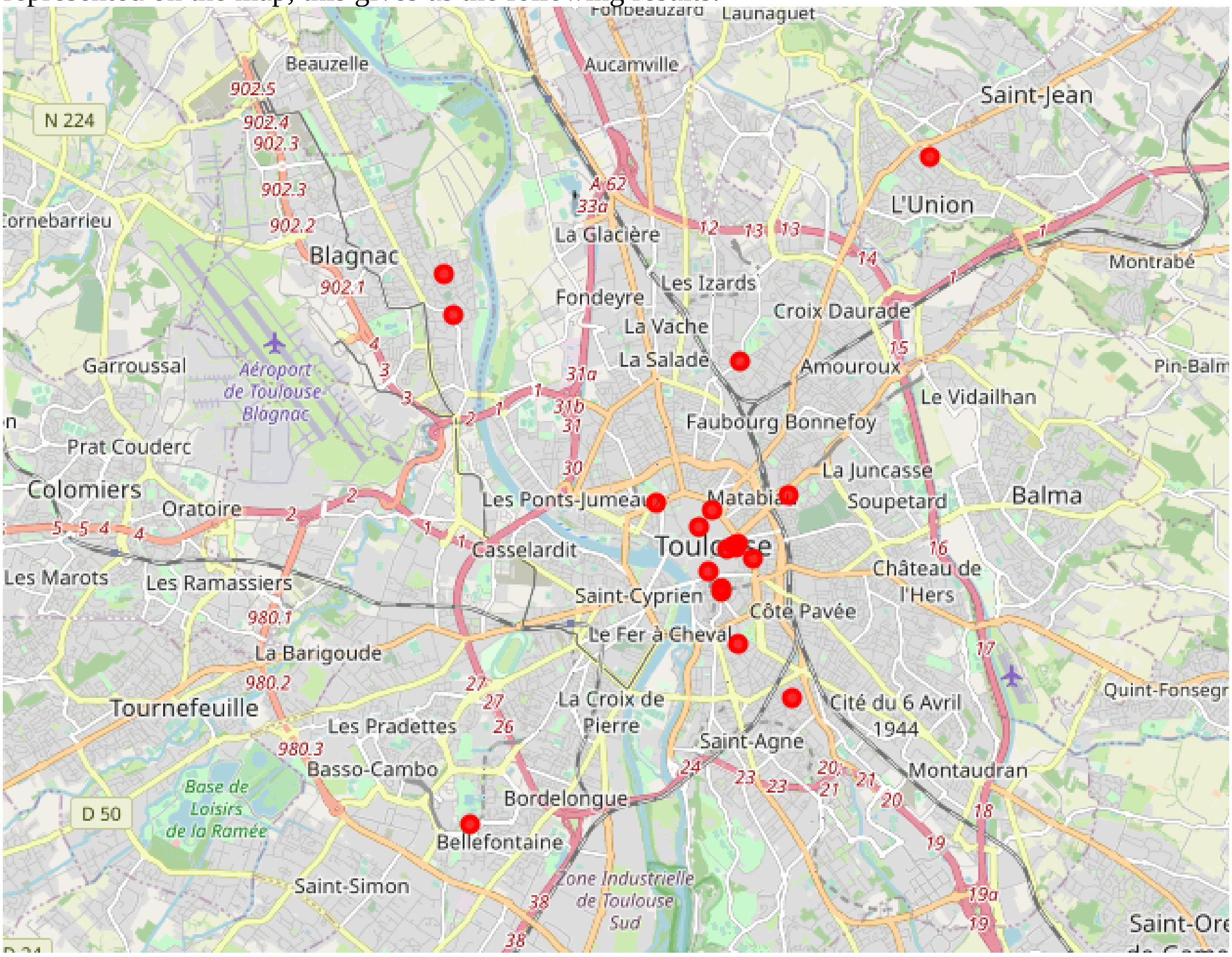
	etat	commune	nom	ligne	couvert	en_service	type	geometry	longitude	latitude
0	Existant	TOULOUSE	BASSO CAMBO	A	Aérien	1993	Metro	POINT (1.39228 43.56996)	1.392278	43.569961
1	Existant	TOULOUSE	REYNERIE	A	Souterrain	1993	Metro	POINT (1.40173 43.57093)	1.401730	43.570927
2	Existant	TOULOUSE	CAPITOLE	A	Souterrain	1993	Metro	POINT (1.44527 43.60435)	1.445273	43.604350
3	Existant	TOULOUSE	ARGOULETS	A	Souterrain	2003	Metro	POINT (1.47674 43.62433)	1.476740	43.624335
4	Existant	TOULOUSE	UNIVERSITE PAUL SABATIER	B	Souterrain	2007	Metro	POINT (1.46310 43.56113)	1.463097	43.561132

3.2 Find all existing movie theaters

Using the Foursquare API with the search endpoint, we then find all existing movie theaters. To do that, we make a get request by passing the url specifying the category ID (i.e. ‘4bf58dd8d48988d17f941735’).
After extracting the information we need, we obtain the following dataframe:

	name	location.lat	location.lng
0	Cinéma Utopia	43.604095	1.446715
1	Gaumont Wilson	43.604251	1.448624
2	La Cinémathèque de Toulouse	43.606907	1.441664
3	Toulouse Centre	43.611280	1.458200
4	Gaumont Wilson - Salle 1	43.598250	1.445708

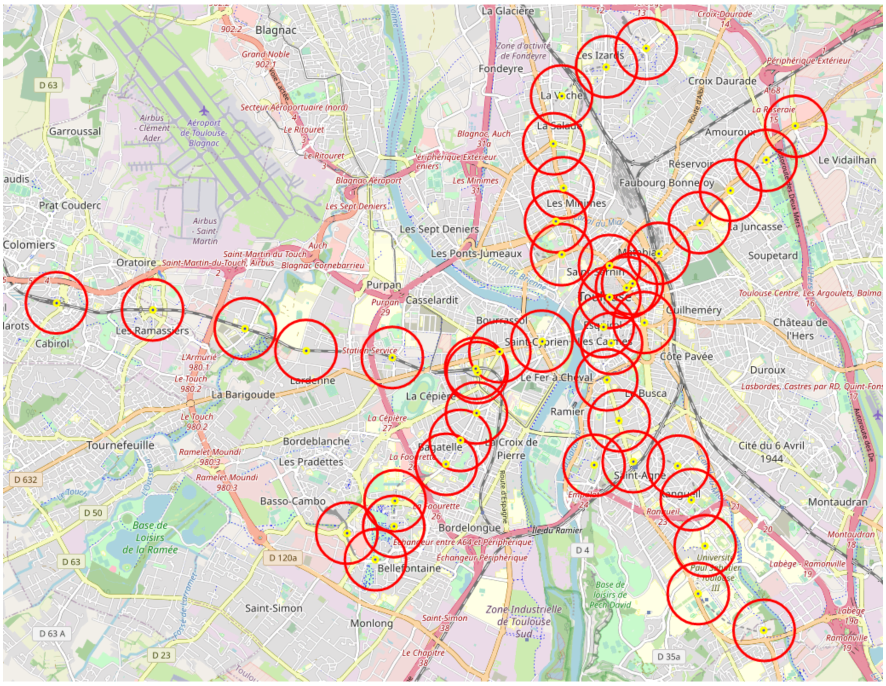
represented on the map, this gives us the following results:



A first observation is worth mentioning. Namely that most of the movie theaters are situated on the east bank of the river in the city center.

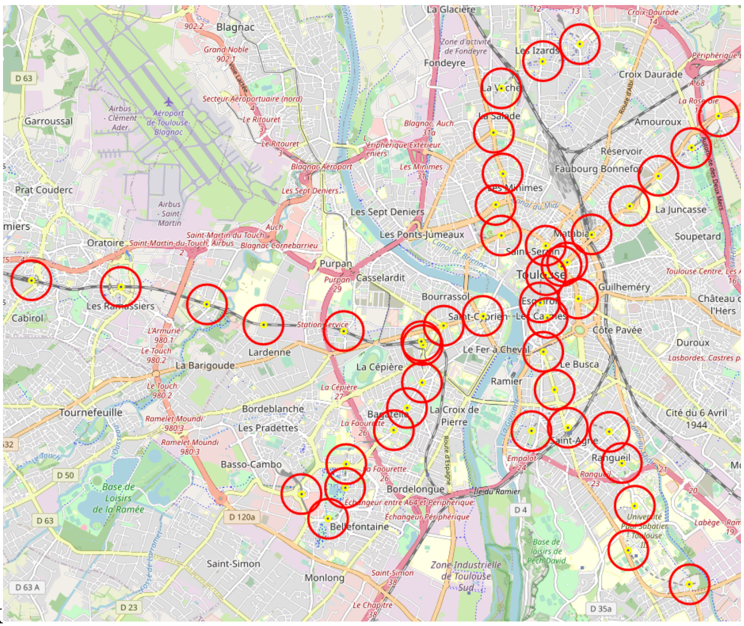
3.3 Getting nearby venues with the Foursquare API

Now we will be using the Foursquare API (explore endpoint) to find venues situated near metro station. And there is a first methodological worry here. In fact, the only thing that we will be able to specify in the request, will be the radius. And the value of the radius will also determine which venues we will get. And if we have overlapping radiuses, then we are likely to have metro stations that belong to the same cluster, of those overlaps. Therefore, we first represent on a map how it would be like to have **radius of 500m**:



Accordingly, we see here that the radius of 500m is problematic since we get way to many overlapping stations (bottom left, center and top).

Therefore, we change the radius and set it to **350m:**



Here the results look more promising. Indeed, we considerably reduce the overlaps. Whilst we still have overlapping radius in the city center, we considerably reduce the number of overlaps outside of the center. And it should be noted that it might not be problematic to have those overlaps in the city center as:

1. some stations are really near to each other so it makes sense that they also land in the same cluster, i.e. they are indeed comparable.

2. a small city center like in Toulouse is normally perceived as "a block" where you don't take the metro to navigate around the city center. So according to our business problem, it should not be problematic to have those overlaps in the center.

Therefore, we consider a radius of 350m as acceptable.

Then, using the **foursquare API** (explore endpoint), we find all venues within a 350m radius from all metro stations. Note that we used the version on December 2019 (the same date as the information for the metro stations were uploaded). After some data manipulation (cleaning), we obtain the following data frame:

	metro_station	latitude	longitude	venue_name	venue_lat	venue_lng	categorie	address
0	BASSO CAMBO	43.569961	1.392278	Buffalo Grill	43.569756	1.390050	Steakhouse	[ZAC de Basso Combo. 4 Impasse Michel Labrouss...
1	BASSO CAMBO	43.569961	1.392278	Métro Basso-Cambo Ⓐ	43.570034	1.392281	Metro Station	[Avenue du Mirail, 31100 Toulouse, France]
2	BASSO CAMBO	43.569961	1.392278	Quick	43.569553	1.390576	Fast Food Restaurant	[7 place E.Bouillères (place E.Bouillères), 31...
3	BASSO CAMBO	43.569961	1.392278	Arrêt Basso Cambo [21]	43.569854	1.392042	Bus Stop	[Toulouse, France]
4	BASSO CAMBO	43.569961	1.392278	McDonald's	43.567409	1.392363	Fast Food Restaurant	[avenue du Mirail, 31100 Toulouse, France]

We have some elements in this dataframe that are "metro station". And this is likely to bias our clustering parts, so we remove those rows.

The data frame now consists of 8 columns: the metro station name to which the venue is the nearest, the latitude and longitude of that metro station, the venue name, its latitude and longitude, its category and address.

This data frame consists in 504 venues – compared to big cities, the number of venues is quite low. Indeed, Toulouse is not so comparable to Paris for instance. Moreover, since we reduced the radius to 350m, the number of returned venues is also lower. However, this was a trade-off we had to make: either increase the radius, with the risk of having overlapping radius but get more venues, or reduce the overlap and get less venues. We choose the second option as it seemed to be a lesser methodological problem.

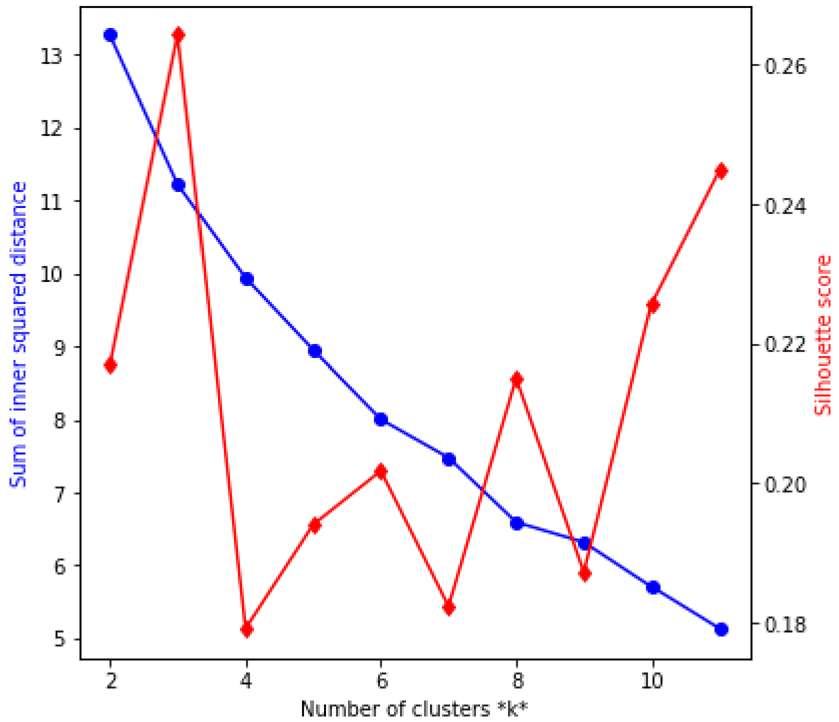
Then we identify the frequency of all category of venues per metro station. In other words, we identify the distribution of the different categories per metro station. Indeed, it is on that basis that we will later cluster those metro stations.

3.4 Clustering with the K-Means algorithm

The K-Means algorithm is an unsupervised classification method – I will not further explain how this method work, since plenty of resources are already available online.

Nonetheless, it is worth explaining why we use the K-Means algorithm here as to answer our business problem. In fact, we now try to identify similar metro station, i.e. metro station that have similar venues in their neighborhoods. Moreover, we will also analyse and discuss the most common venues in the neighborhoods of those metro stations. This will provide us with valuable information regarding our business problem. Indeed, we will be able to better understand why existing movie theaters are located in the same area, and identify if there are other promising regions that would be suitable to implement such a movie theater.

One methodological crux of the K-Mean is that we need to find a suitable number of k (cluster). To do so, we apply the silhouette-score and the sum of inner squared distance:

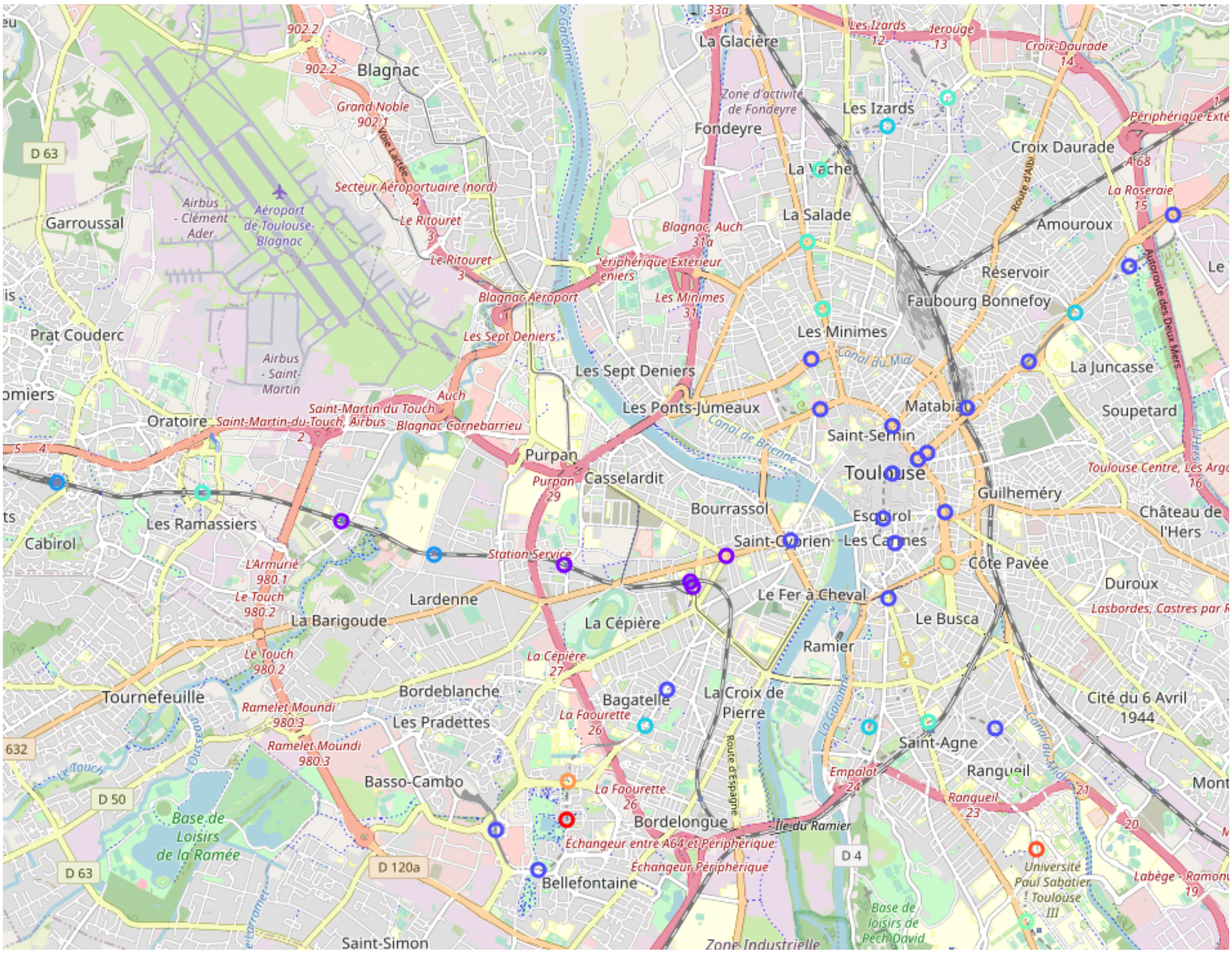


Recall that we aim here at finding a high silhouette score and a low sum of inner squared distance. A prima facie good candidate could be k=3, since the silhouette score is very high. However, with k=3, the sum of inner squared distance is also very high. Therefore, we decide to take k=11.

Then we run the K-Means algorithm with k=11 and obtain the following data frame:

	nom	ligne	couvert	en_service	type	geometry	longitude	latitude	cluster_labels	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 Com V
0	BASSO CAMBO	A	Aérien	1993	Metro	POINT (1.39228 43.56996)	1.392278	43.569961	2.0	Fast Food Restaurant	Construction & Landscaping	Shopping Mall	Bus Stop	Steakhouse	Convenience Store	Cosmetics Shop	Cre
1	REYNERIE	A	Souterrain	1993	Metro	POINT (1.40173 43.57093)	1.401730	43.570927	0.0	Park	Wine Bar	Electronics Store	Convenience Store	Cosmetics Shop	Creperie	Cupcake Shop	D S
2	CAPITOLE	A	Souterrain	1993	Metro	POINT (1.44527 43.60435)	1.445273	43.604350	2.0	French Restaurant	Plaza	Burger Joint	Coffee Shop	Hotel	Restaurant	Tea Room	Ice C
3	ARGOULETS	A	Souterrain	2003	Metro	POINT (1.47674 43.62433)	1.476740	43.624335	2.0	Pool	Gas Station	Park	Wine Bar	Dog Run	Convenience Store	Cosmetics Shop	Cre
4	UNIVERSITE PAUL SABATIER	B	Souterrain	2007	Metro	POINT (1.46310 43.56113)	1.463097	43.561132	6.0	Bus Stop	Wine Bar	Event Space	Cosmetics Shop	Creperie	Cupcake Shop	Dance Studio	Depart

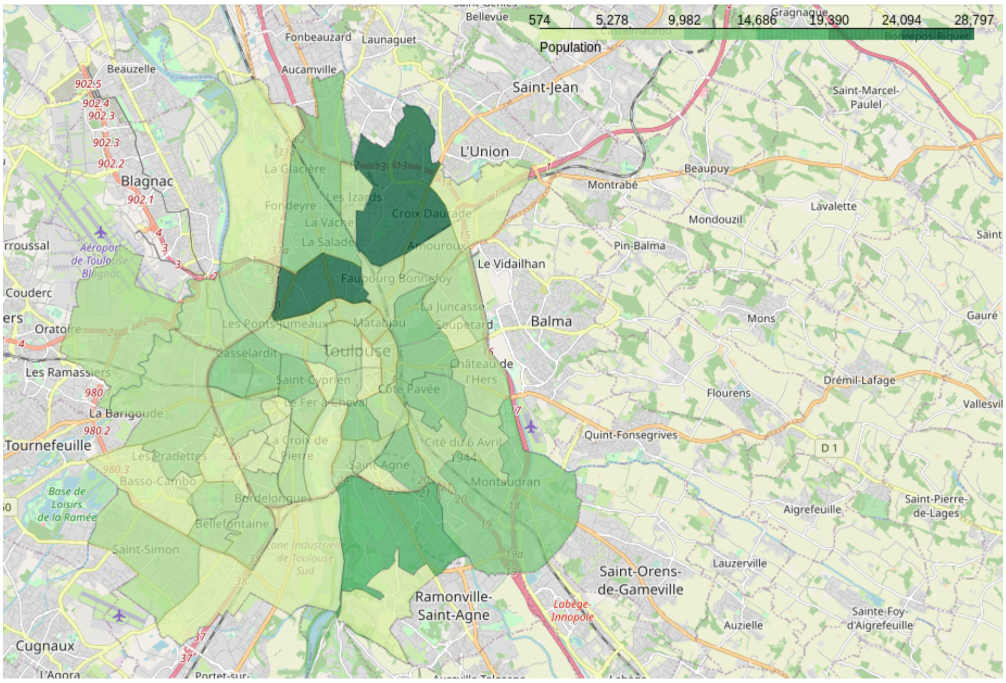
There we see the name of the metro station, its line, coordinates, cluster labels and the 10th most common venues present in the neighborhood of that metro station. We now visualize the results on a map:



We directly identify that movie theaters are mostly present in cluster 2 (blue).

Before turning to the analysis of our results, we also consider census data, i.e. population in each neighborhood. This will also help us to make recommendations later on. To do so, we get census data from the following page:https://data.toulouse-metropole.fr/explore/dataset/-recensement-population-2016-grands-quartiers-population/information/

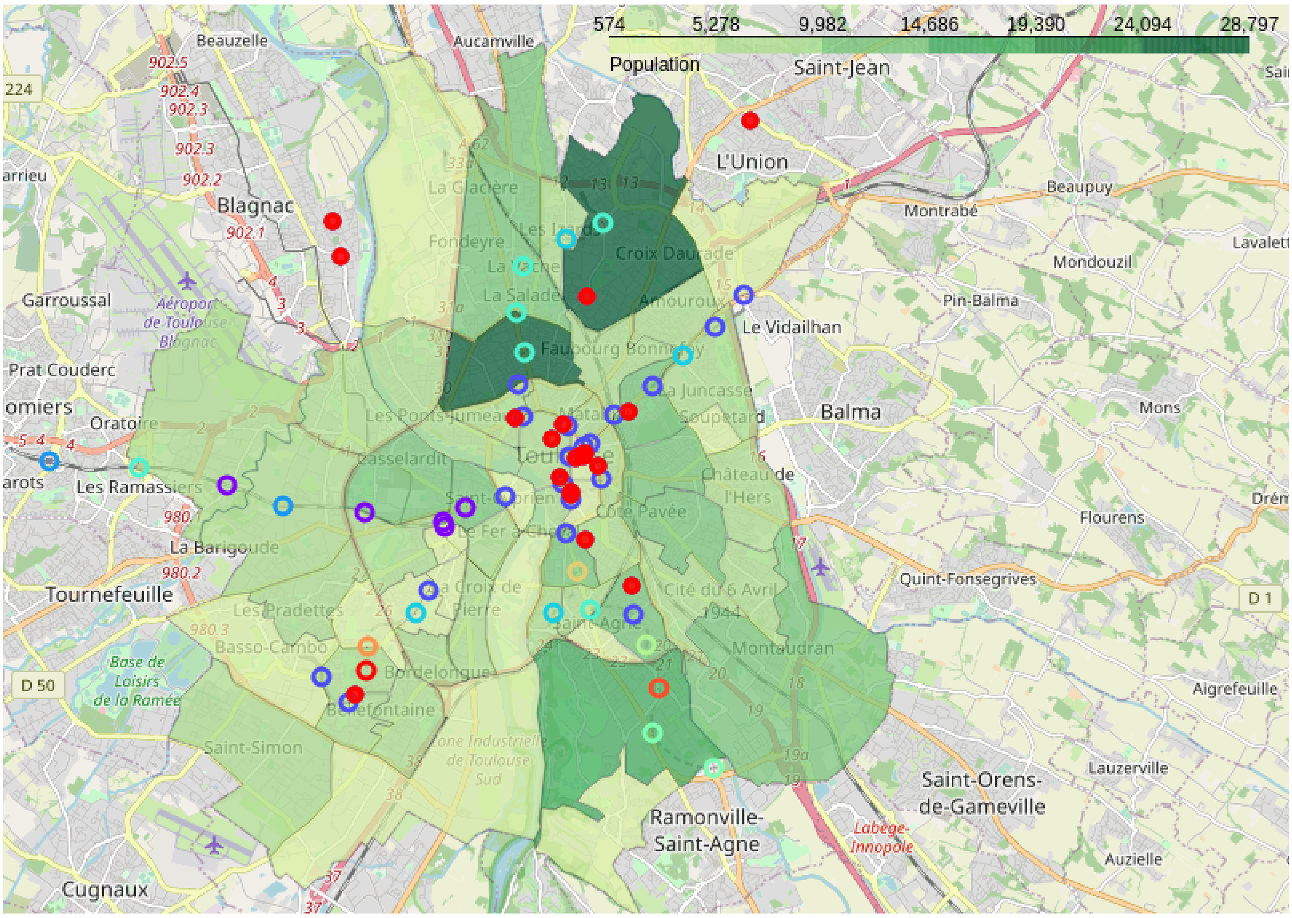
We then represent the results on a choropleth map:



in the next section, we will use this map in conjunction with clusters and existing movie theaters as to identify a suitable place for a new movie theater.

4. Results

On the following map, we see in red the location of existing movie theaters, the concentration of population and the different metro stations and the cluster to which they belong.



As previously mentioned, movie theaters are mostly located in cluster number 2. the categories of venues in cluster number 2 are the following:

	ligne	geometry	longitude	latitude	cluster_labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	A	POINT (1.39228 43.56996)	1.392278	43.569961	2	Fast Food Restaurant	Construction & Landscaping	Shopping Mall	Bus Stop
2	A	POINT (1.44527 43.60435)	1.445273	43.604350	2	French Restaurant	Plaza	Burger Joint	Coffee Shop
3	A	POINT (1.47674 43.62433)	1.476740	43.624335	2	Pool	Gas Station	Park	Wine Bar
6	A	POINT (1.44985 43.60634)	1.449854	43.606345	2	French Restaurant	Pub	Plaza	Bar
7	A	POINT (1.48272 43.62931)	1.482715	43.629314	2	Hobby Shop	Supermarket	French Restaurant	Fast Food Restaurant
9	B	POINT (1.44563 43.59764)	1.445625	43.597642	2	French Restaurant	Bar	Plaza	Café
15	B	POINT (1.44863 43.60580)	1.448634	43.605803	2	French Restaurant	Pub	Plaza	Bar
18	B	POINT (1.44465 43.59232)	1.444649	43.592324	2	Caribbean Restaurant	Restaurant	Tram Station	French Restaurant
19	B	POINT (1.43435 43.61539)	1.434348	43.615390	2	Wine Bar	Indian Restaurant	Event Space	Stadium
20	A	POINT (1.43158 43.59797)	1.431577	43.597970	2	Plaza	Supermarket	Tapas Restaurant	Sandwich Place
25	A	POINT (1.39806 43.56607)	1.398055	43.566072	2	Cosmetics Shop	Shopping Mall	Wine Bar	Electronics Store
27	A	POINT (1.45512 43.61077)	1.455121	43.610771	2	French Restaurant	Hotel	Convenience Store	Train Station
28	A	POINT (1.46334 43.61526)	1.463342	43.615256	2	Hotel	Photography Studio	Supermarket	Bus Station
30	B	POINT (1.45903 43.57973)	1.459032	43.579729	2	Brewery	Bakery	Jewelry Store	Wine Bar

There are two main observations that we can do here. Firstly, existing movie theaters are located in the city center, where the number of metro station is quite high, i.e. a region in the city that is well connected to other parts of the city. Therefore, this belongs to our first criteria for a suitable movie theater: near public transports.

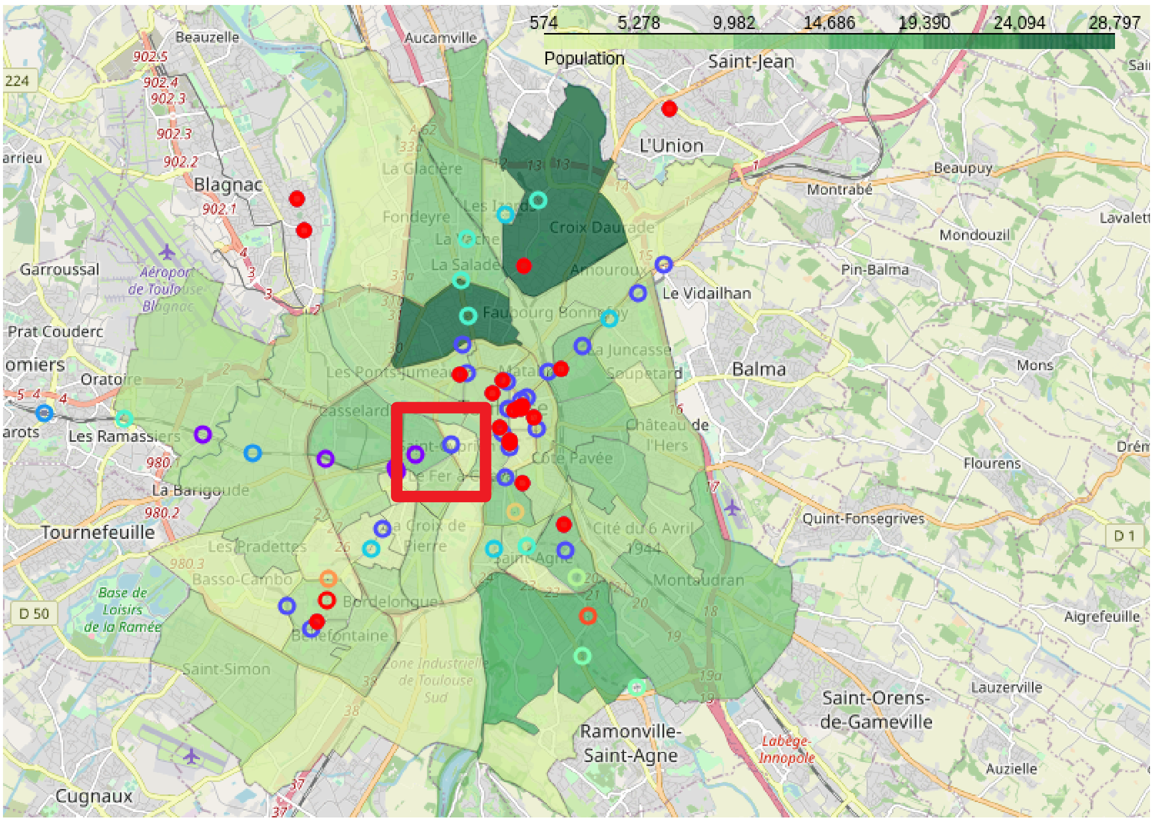
Secondly, existing movie theaters are located in regions where we primarily have restaurants, bars, pubs, brewery etc… In other words, those tend to be located in typical “night life” areas or popular place to go out and hang out.

Thirdly, population density seems not to play a central role as to where a movie theater is. Indeed, we may make the hypothesis that if theaters are located in popular areas for going out and also well connected to other parts of the city, movie theaters need not to be in walking distance from people home. For instance, this hypothesis is likely not to hold in the case of supermarkets where distance is more likely to play a role.

To summarize, we identify the following key elements required for the implementation of a new movie theater: 1) good connection with other parts of the city 2) being located in popular areas with other entertainment possibilities and 3) population density is likely not to play a major role.

5. Discussion

Based on our analysis and results, we identify the following interesting areas for the implementation of a new movie theater: metro station: **Saint-Cyprien Republique.** Indeed, this metro station belongs to the cluster number 2, namely, there are a lot of existing entertainment venues but no movie theaters. Moreover, this metro station is also well situated since it is near the city center – west bank of the city center (see red square on the picture below).



Therefore, our recommendation for the implementation of a new movie theater is in this red squared area as it fulfills both our necessary conditions for the implementation of a movie theater. Of course, the success of a movie theater will not solely depend on its location. The type of movies displayed, the atmosphere, customer services are other variables that will likely play a significant role in a business success of any type. However, to find a good match for those variables is something that will be left for potential investors. But as a location, the best one could be identified.

6. Conclusion

Finding a suitable place for a movie theater is not an easy task. There are already several movie theaters present in Toulouse. Yet, they all seem to be located in one particular place, namely, on the east-side of the city center. The purpose of this analysis was to find another new location that would be suitable for the implementation of a new movie theater, yet in a different area as to lower competition with existing movie theaters. A requirement was that the location for this new movie theater be located near a metro station.

To do so, we first found all metro stations and their surrounding venues. Then, accordingly, we clustered those metro stations based on the “type” (category) of those venues using the Foursquare API and the K-Mean algorithm.

We found two condition for a movie theater to be successful, namely:
- it has to be well-connected to the rest of the city
- it should have other entertainment possibility in its vicinity
- it should not be located in an area where we already have a lot of existing movie theaters as to reduce competition.

We found that the metro station “Saint-Cyprien Republique” was the most suitable location for that new movie theater, as it fulfilled all our criteria. Therefore it is the area near this metro station that we recommend investors for the implementation of a new movie theater.