Predicting the best location to live in

Aleksandr Plotnikow

1 Introduction

1.1 Background

We are now witnessing the highest levels of movement on record. About 258 million people, or one in every 30, were living outside their country of birth in 2017. Some people move for family reasons, someone goes in search of work, they are those who are looking for happiness or themselves. There are many forced moves due to political, conflict or natural reasons.

To some places in the world there are queues and in the meantime from other places only go away. However, when this moment of change in life comes, there is an indicative goal, it is necessary to start looking for a place. Everyone has different needs and goals and in the meantime one may not know anything about the new place to which one goes. Knowing the initial needs of an individual, it is possible to take and define some initial purpose.

1.2 Approach to the problem

First of all, it is necessary to define the initial variables that interest us in the new place:

1. Location to which the person is located - borough/city/district/country
2. Important to determine the extent to which he is interested in his new place of residence
3. Important facilities that are essential for operation

The above variables allow the extraction of specific data bases. There is still a need to filter these data and extract the necessary information for further analysis. In the following steps, you will see how the initial information was obtained, how it was processed to obtain the key information, then you will see the process leading up to the goal.

1.3 Interest

A significant part of people would like to have an influence on the place where they will settle, regardless of whether it changes, streets, city, country or continent. Prediction of a place according to expectations allows to indicate a place which allows to get access to the maximum number of needs which a given location is able to provide us with.

2 Data acquisition and cleaning

2.1 Data sources

Two sources of data were used to solve the problem. Firstly, a platform for the Foursquare developers, which has provided initial information about the location where the search for the best place will take place. The platform was then used to find the facilities we were interested in. For the purposes of this project, I have used my own example as a person searching for a place in Toulouse City in France. The main attributes that interest me in a location is an access to a metro, gym, and to a gastronomy (sushi and kebab). Apart from data obtained from the above mentioned platform, also I have downloaded the list of city districts from the mayor of the city. This is not a necessary step, but it allows us to become more familiar with the layout of the districts and their number in the city.

2.2 Data cleaning

The data obtained from Foursquare are very raw, so they should follow the sorting and filtration. In the case of a response received to a request for a city, the necessary data, such as:

1. Correct name of the city
2. Name of the province
3. Name the countrie
4. Geographical coordinates

Then the data on the facilities we are interested in were analysed within a certain radius (the range of the Toulouse city was accepted). In order to simplify the initial analysis, the names of the units have been simplified and redundant information has been removed. Information that interested us:

1. Name of the unit
2. Address
3. Geographical coordinates

3 Methodology

3.1 Data visualisation

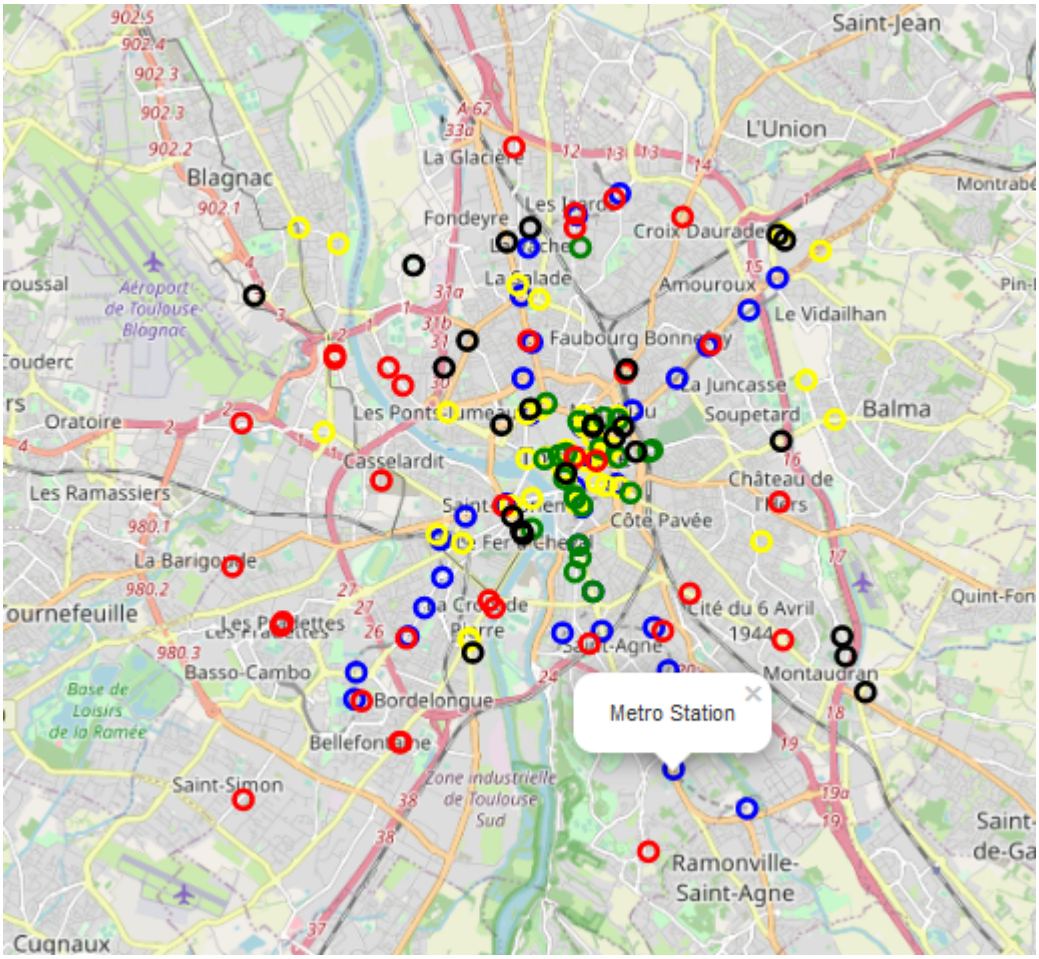

Figure 1: City map with attached points of interest

Fig. 1) gives an initial idea of the information obtained. The locations of the city districts are shown in red, and they are equally spaced in relation to each other. Then with the help of other colours the facilities we are interested in are presented. At first glance it is possible to notice that in the middle part of the map we have a collection of points of interest. This property will be used for data processing.

3.2 Facility Clustering

In view of the previously mentioned fact that Toulouse's districts are situated in a way that is relevant to each other, I have decided to use a cluster analysis. Remembering that districts are equally distributed, there are still a few anomalies connected with the existence of several districts in a large proximity to each other. So only I have used the number of districts as the number of concentration points in an analysis.

Next, the value of the cluster was assigned to each facility and the repetition was presented on the map(Fig. 2)). As was previously predicted, the point of interest is in the centre. However, it is still not clear which district is the best point of interest.

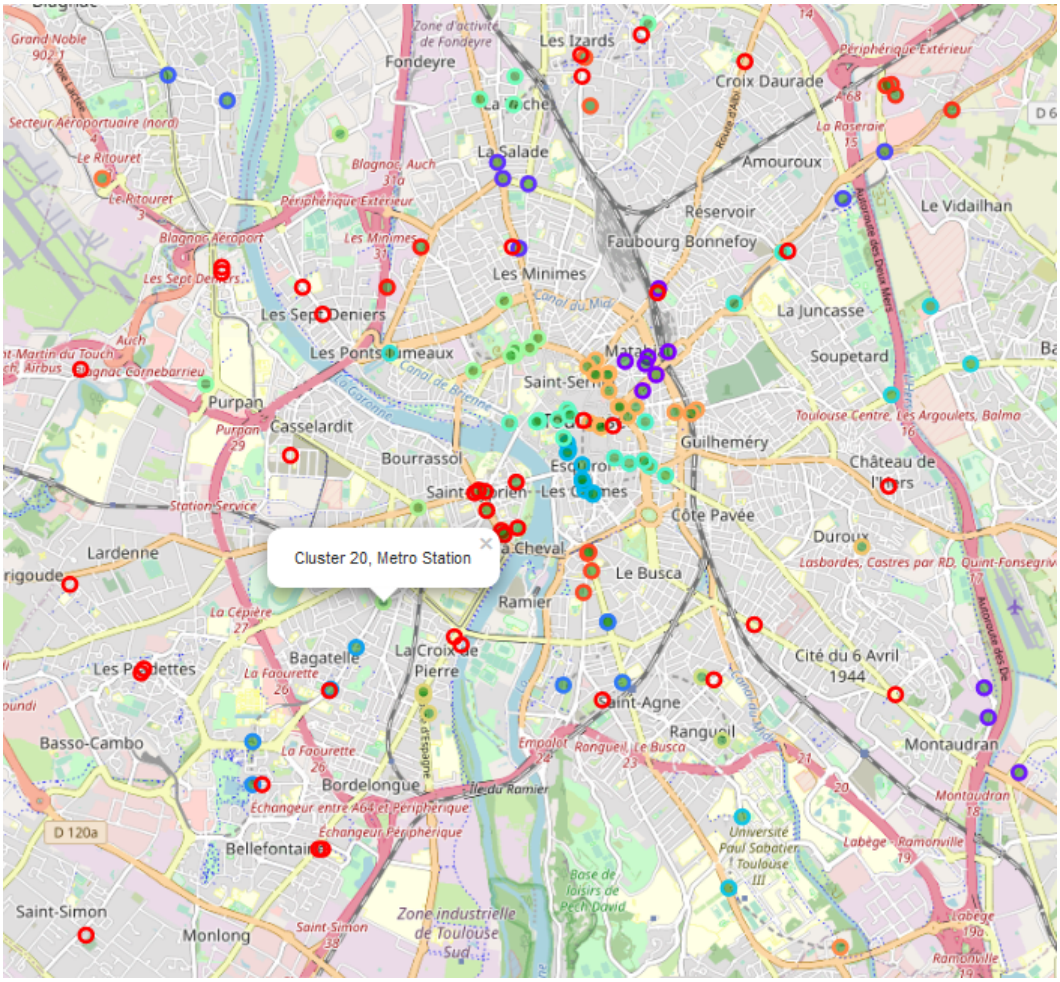

Figure 2: City map with attached clustered points of interest

3.3 Adaptation of the data

Knowing what we need to get, work has started on the data base in order to get a precise answer to the problem. The first stage was the creation of one common database of data with values containing the occurrence of the selected facility in relation to a particular cluster point. Fig. 3 shows us a part of the final data base. Accordingly, we have the number of the cluster, the name of the occurring point of interest and the part of its occurrence and at the end the sum of all occurring conveniences in this area. As the work has been carried out on a limited data base, it can be seen that the 25 cluster has the most facilities, and this can also be deduced from the bar chart (Fig. 4).

	Cluster Labels	Name of Kebab	freq_kebab	Name of Sushi	freq_sushi	Name of Gym	freq_gym	Name of Metro	freq_metro	Total
0	0	Kebab	1.0	Sushi	2.0	Gym	3.0	Metro Station	1.0	7.0
1	1	Kebab	2.0	Sushi	1.0	Gym	3.0	Metro Station	1.0	7.0
2	2	0	0.0	0	0.0	Gym	3.0	0	0.0	3.0
3	3	0	0.0	Sushi	2.0	0	0.0	Metro Station	2.0	4.0
4	4	0	0.0	0	0.0	0	0.0	Metro Station	2.0	2.0
5	5	0	0.0	Sushi	2.0	0	0.0	0	0.0	2.0
6	6	Kebab	1.0	0	0.0	0	0.0	Metro Station	3.0	4.0
7	7	0	0.0	0	0.0	0	0.0	Metro Station	2.0	2.0
8	8	0	0.0	0	0.0	0	0.0	Metro Station	2.0	2.0
9	9	Kebab	4.0	Sushi	3.0	Gym	1.0	Metro Station	2.0	10.0
10	10	0	0.0	0	0.0	0	0.0	Metro Station	2.0	2.0
11	11	0	0.0	Sushi	2.0	Gym	1.0	0	0.0	3.0
12	12	0	0.0	Sushi	1.0	0	0.0	0	0.0	1.0
13	13	0	0.0	0	0.0	0	0.0	Metro Station	2.0	2.0
14	14	Kebab	2.0	Sushi	3.0	0	0.0	Metro Station	1.0	6.0
15	15	Kebab	3.0	Sushi	4.0	0	0.0	0	0.0	7.0
16	16	0	0.0	0	0.0	Gym	2.0	Metro Station	1.0	3.0
17	17	0	0.0	0	0.0	0	0.0	Metro Station	1.0	1.0
18	18	0	0.0	Sushi	1.0	0	0.0	0	0.0	1.0
19	19	Kebab	1.0	Sushi	2.0	Gym	2.0	Metro Station	2.0	7.0
20	20	0	0.0	Sushi	2.0	0	0.0	Metro Station	3.0	5.0
21	21	0	0.0	0	0.0	Gym	1.0	0	0.0	1.0
22	22	0	0.0	0	0.0	0	0.0	Metro Station	2.0	2.0
23	23	0	0.0	Sushi	2.0	Gym	1.0	0	0.0	3.0
24	24	0	0.0	Sushi	1.0	0	0.0	0	0.0	1.0
25	25	Kebab	2.0	Sushi	7.0	Gym	2.0	Metro Station	3.0	14.0
26	26	Kebab	2.0	0	0.0	Gym	1.0	0	0.0	3.0
27	27	0	0.0	0	0.0	0	0.0	Metro Station	1.0	1.0
28	28	0	0.0	0	0.0	Gym	1.0	0	0.0	1.0
29	29	Kebab	1.0	0	0.0	0	0.0	Metro Station	1.0	2.0
30	30	Kebab	3.0	0	0.0	0	0.0	Metro Station	1.0	4.0
31	31	0	0.0	Sushi	2.0	Gym	2.0	0	0.0	4.0
32	32	0	0.0	0	0.0	Gym	2.0	0	0.0	2.0

Figure 3: Final data table

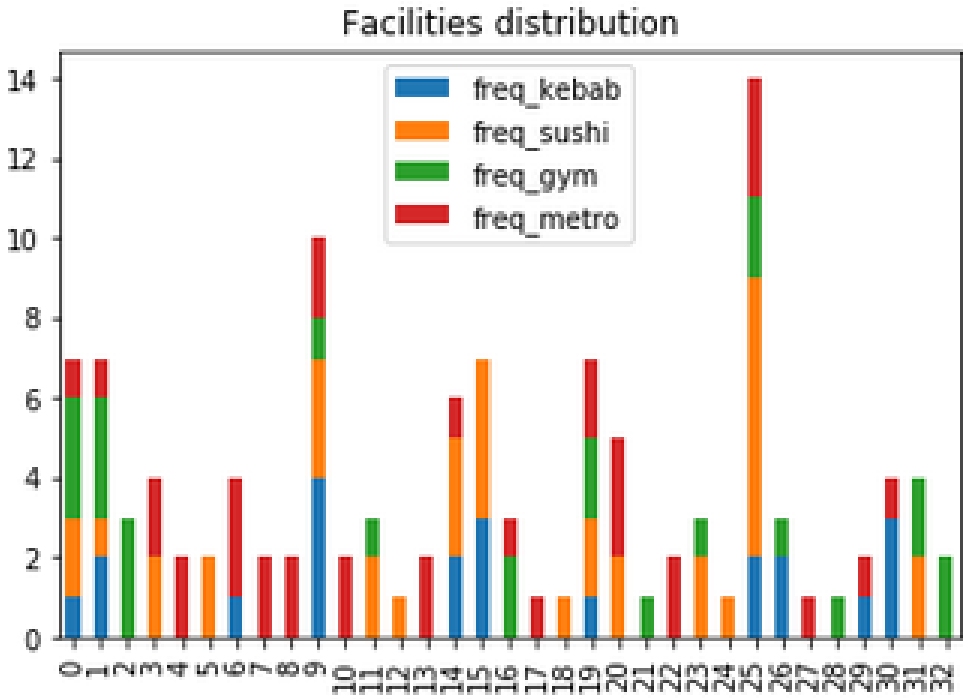

Figure 4: Facilities per cluster bar chart

3.4 Identification of the best location

Knowing the collection of interest to us, we are able to find the point which is in the centre between all objects. Then, for information purposes, we can find out which district is closest to a given point(Fig. 5).

	mairie	libelle	lat	lng
8	MERIEL	RUE PAUL MERIEL	43.603985	1.447996

Figure 5: The closest neighbourhood

As you can see on the map(Fig. 6), the blue marker indicates the point which has the best access to the greatest number of facilities. As predicted earlier it will be in the centre.

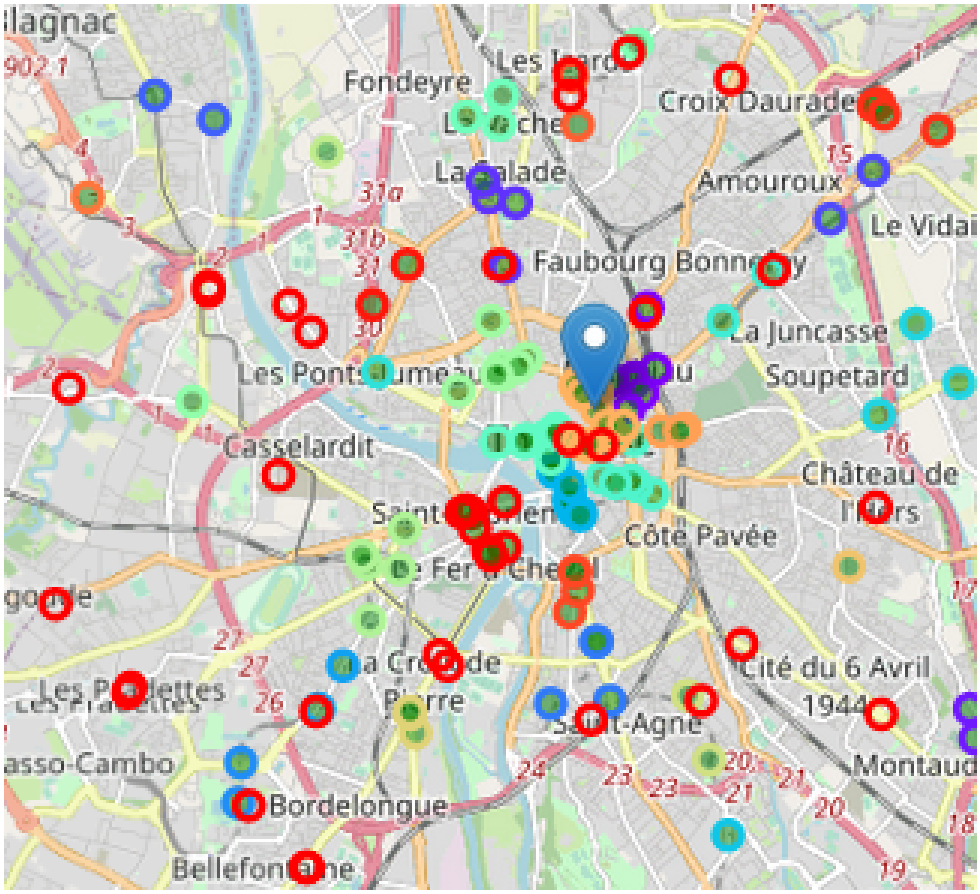

Figure 6: The closest neighbourhood

4 Discussion

It was possible to predict that the obtained result will be in the centre of the map, but I did not take into account a very important parameter, i.e. increase in price with a decreasing distance to the centre of Toulouse. However, it should be remembered that the survey was aimed at indicating the best point in terms of customer demand. However, it is worth mentioning that we still have access to information about the rest of the map. This allows you to see that there are locations with at least one selected facility, so they are also locations that match the demand criterion.

In the future it would be possible to optimise the result more. We work with real estate agencies and their data on the possibilities of purchasing or renting in a given location. This cooperation allowed us to find the most optimal solution for each person, taking into account the current state of the market and their needs.

5 Summary

This paper presents the initial process of creating a program to help find the area that best suits the client. It allowed to find a point in the location where the largest number of selected points of interest are located. As a preliminary analysis of the new location in terms of moving house, it gave the opportunity to see the location of the facilities in the city. As mentioned in the discussion, it is highly likely that this is one of the most expensive locations in the city. However, you can see from the data that there are other locations that meet the search criteria in other neighbourhoods. Foursquare, as a tool for obtaining data, has helped a lot in the process of obtaining information about a particular location. The cluster analysis allowed for the division of the city into sectors, which helped to build the final data base.