Analysis of Locations for a Business

Anil Sasidharan

1. Introduction

1.1 Background

Success of any business, whatever the domain might be, depends on a number of common factors. It is imperative that the owner of the business gets it right on most of them, if not all, to ensure success. Among these factors, the opening of the said business, as an event, carries a huge significance, both in terms of contributing to the success and also in financial management. As a result, this is almost always treated as very important by entrepreneurs. More the information they can acquire that can help towards the successful opening of their business, the better planned they can be. This proves how useful information is to any business, and in turn, data. For this particular project report, I am concentrating on specifically location data. One of the important parts in opening a business is selecting its location. So I will be demonstrating how to achieve a data driven result.

Toulouse, the city where I am currently residing since 2016, has many facets. It is a small and calm typical French city, with 18th century architecture and beautiful cafes and restaurants wherever you go. The different neighborhoods, or 'quartiers' as they are referred to here, each have their own features, along with the different behaviors and lifestyles of the people there. I am using location data from Foursquare to get an insight into these properties so as to obtain good candidate locations for a business, which is – a Games and Recreation center.

1.2 Problem

The aim of this analysis is to show that data analysis can be used in important factors of opening a business. It can uncover certain aspects that tend to be hidden – for example, for a particular public business opening in a particular location, how will the residents of the locality perceive it, what percentage of the population are welcome to its idea, and is it in line with the lifestyles of the residents. These are factors which normally come to light a while after the business has been opened. And there is also a chance of a negative impact. The aim of data analysis is to provide an advantage to business owners in this regard.

1.3 Stakeholders

As noted above, entrepreneurs would be the most interested in this analysis. Existing business owners would like the idea of being a step ahead when opening branches or franchises in unfamiliar regions. Also, this kind of analysis has a broader scope. Gauging public opinions, generating public feedback, comparing growth of businesses, and even measuring economic growth are some of the uses. And so, the stakeholders list will enlarge to include survey companies, reviewing companies, economic departments, government administration and last but not least, social media.

2. Data for the Analysis

2.1 Principle of Use

The data required for this project is classified into two parts – Location data of the different neighborhoods of Toulouse, and Foursquare data for these locations.

Location data will include compiling a comprehensive list of the neighborhoods and the area they cover, and the geographical coordinates. This will be the primary database and a foundation for the analysis. The Foursquare data will generate information on trends in lifestyle and the popularity of different businesses. This data for each location can be added progressively to the primary database to obtain a final version for analysis.

2.2 Source

Data on neighborhoods or quartiers in Toulouse has been taken from the site data.toulouse-metropole.fr. The link for the CSV file given below:

https://data.toulouse-metropole.fr/explore/dataset/recensement-population-2016-grands-quartiers-familles/download/?format=csv&timezone=Europe/Berlin&lang=fr&use_labels_for_header=true&csv_separator=%3B

The file has to be cleaned and prepared to get a final list of neighborhoods and coordinates. The Foursquare data has to be obtained by repeated calls to the Foursquare API. The different API calls will be explained in the Methodology section as we go forward in the analysis. To obtain results with Foursquare API calls, I have my free account with my exclusive credentials.

2.3 Data Preparation and Feature Selection

The initial database that I have downloaded is a census database with population numbers classified under several categories. Thus, it has a lot of features not useful in this analysis.

Firstly, I have cleaned the data by removing all the population numbers and also the features “Index”, “GRD_Quart” and “Dep” numbers. Secondly, I have renamed the features to English as “Coordinates” and “Neighborhood”. Thirdly, I have split the data in label “Coordinates” into “Latitude” and “Longitude” data. After this step, the database is primed for use, to generate interactive maps, as well as generating Foursquare API calls.

3. Methodology

3.1 Principle of Analysis

At this point, I have the primary database with the Toulouse neighborhood data to start my analysis to find optimum candidate locations for the opening of a Games and Recreations Center. Foursquare API calls will be used to retrieve information on each of the neighborhoods which will give an idea on which locations are perfect, and which are not. A sample of this data of neighborhoods of Toulouse, cleaned and prepared for analysis, is shown below.

	Neighborhood	Latitude	Longitude
0	ARNAUD BERNARD	43.607588	1.439794
1	LES CHALETS	43.613401	1.442695
2	MINIMES	43.619162	1.431955
3	SAUZELONG - RANGUEIL	43.579110	1.459158
4	FAOURETTE	43.579090	1.415167

Figure 1. *Sample of dataframe with Toulouse neighborhood data*

The main principle of this analysis is to retrieve information on venues already present and running in the neighborhoods of Toulouse. The factors that will help in my project are the type of venues, their number in a single locality, their popularity and their potential relationship with a Games and Recreation venue to open near them. For example, a Games and Recreational Center is a mostly a weekend or evening activity which can be noisy as well. So a neighborhood which is already accustomed to such venues, and which

enjoys visitors of the like, is more suited to be a candidate location than others. Also the popularity of such venues, marked by 'trending' venues, or by number of check-ins, make a difference. Another factor is the effect of other venues around. For a visitor to a Games and Recreational Center, as it is a physical activity, would require a refreshment or food center nearby. This adds to the appeal of the visitor. Even factors like a nearby metro station or a pub to visit after, adds to the appeal of an entertainment center.

Although there are many factors I can use for my analysis, I am using only the features available free with Foursquare as I am using a free account. These include retrieving venue recommendations, by number, by popularity and by category, which are used in this analysis.

Before I start the analysis, it will be good to visualize the primary database with the neighborhood locations and their coordinates. The map generated of Toulouse with markers denoting the neighborhoods is shown below. The interactive version of this map can be checked out on my notebook for this project which is available on my Github profile – the link is added in the Results section.

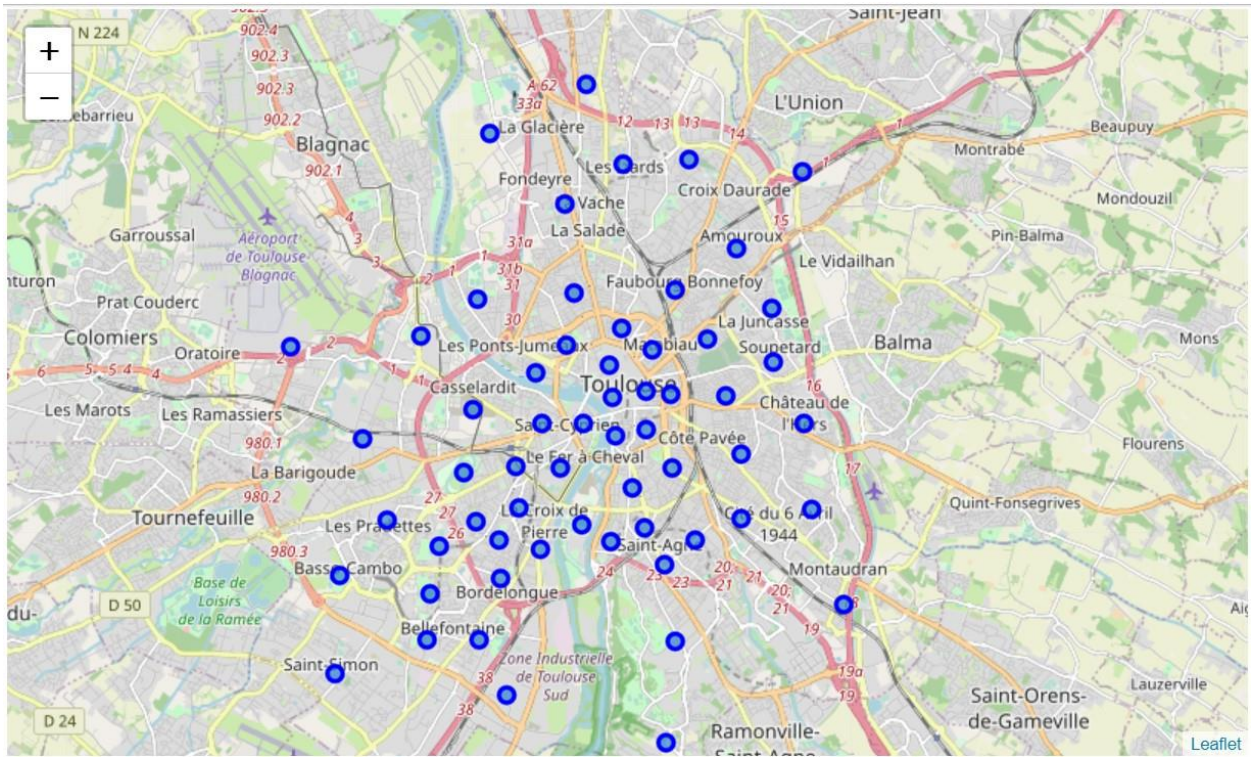


Figure 2. *Map of Toulouse with markers showing neighborhoods*

The analysis is divided into 3 parts:

1. In the first part, I will use the data on the most common types of venues in each neighborhood. More the number of venues of a particular category are present at a location, more the probability of success of that particular type of business is in that neighborhood. From this data, I can use a classification method to identify clusters within the neighborhoods of Toulouse. These clusters can then be marked with characteristics particular to each. This will help in identifying the clusters, and thus the neighborhoods, that are good candidate locations for the business.

2. The second analysis is to use the popularity of venues instead of their commonality. The more popular a venue is, evidently more its success rate is. The data retrieved can be treated in the same way as in the first part. Clusters will be identified and marked to get usable information about the neighborhoods.
3. The final part involves 'key' venue categories. The principle is to identify certain key types of venues whose presence can positively affect the business of a Games and Recreation Center or increase the appeal of it at that particular location. This analysis aims to find localities with highest number of key venues that are identified.

The three analyses have to be combined to get a final result. So I will assign weightages to each analysis and allocate points to each neighborhood according to their performance in the analyses. In this way, a final compiled list of points can be obtained which can ideally generate the best candidate locations for a Games and Recreation Center in Toulouse.

3.2 Analysis 1

As mentioned in the previous subsection, the principle is to retrieve the data for the most common venues for each of the neighborhoods of Toulouse. The data regarding venues is retrieved using Foursquare. The Foursquare platform offers location-based information, namely details regarding venues at a location, number of check-ins, its popularity, trending venues, user reviews, tips, menus among many others. A Foursquare API call is made with the relevant keywords according to the type of search that is required. An API call requires user credentials to perform. Certain calls be performed with a free account whereas many other features can be accessed only by premium users. The analysis in this project is carried out with my Foursquare free account.

The Foursquare API call of 'explore' retrieves venue recommendations for a particular location. The parameters passed to the call include a radius of search and a limit in number of venues retrieved. For the first part of the analysis, I have selected a radius of 300 meters and a limit of 100 venues for the API call. The result includes names of the venues, their location, address, and category, which I have the most interest in. These features of each venue and their corresponding neighborhoods are compiled in a dataframe, sample of which is shown here.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ARNAUD BERNARD	43.607588	1.439794	Breughel l'Ancien	43.609363	1.439101	Pub
1	ARNAUD BERNARD	43.607588	1.439794	Creativ'Yogurt	43.605979	1.441231	Snack Place
2	ARNAUD BERNARD	43.607588	1.439794	George & the Dragon	43.607416	1.439559	Pub
3	ARNAUD BERNARD	43.607588	1.439794	Place Saint-Sernin	43.608278	1.441076	Plaza
4	ARNAUD BERNARD	43.607588	1.439794	Midi Minuit	43.607486	1.439689	French Restaurant

***Figure 3**. Dataframe with all the Venue Recommendations*

From this dataframe, the algorithm uses the Venue Category column to calculate the total number of venues of each category is present in each neighborhood. This in turn gives as output, the most common venues of each neighborhood. I have set my algorithm to select the 5 most common venues of each neighborhood for visualization. This can give a good idea of the characteristics of that location.

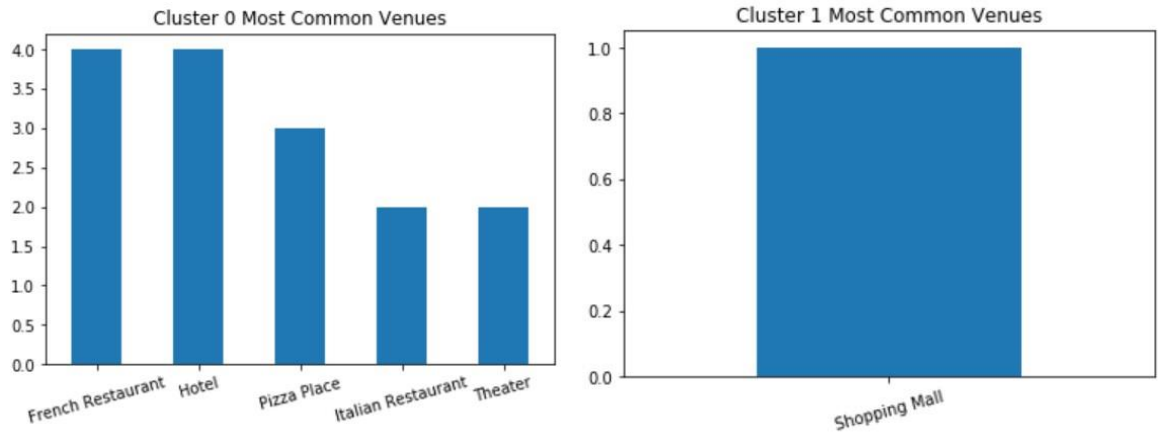
The next step is to classify the neighborhoods. I am using the K-Means clustering method to divide all the neighborhoods into 5 clusters – Clusters 0 to 4. To use the K-Means clustering algorithm, the data regarding venue categories should be numeric and relative to all neighborhoods. So I have used one-hot encoding in my algorithm to obtain a dataframe that can be clustered by K-Means.

After the clustering is performed, the cluster number for each neighborhood is added to generate a final comprehensive dataframe for the first analysis – sample shown below.

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	ARNAUD BERNARD	43.607588	1.439794	0	French Restaurant	Café	Pub	Hostel	Snack Place
1	LES CHALETS	43.613401	1.442695	0	Italian Restaurant	Gym	Hotel	Comedy Club	Sporting Goods Shop
2	MINIMES	43.619162	1.431955	0	Argentinian Restaurant	Pedestrian Plaza	Wine Shop	Farmers Market	Comedy Club
3	SAUZELONG - RANGUEIL	43.579110	1.459158	3	Metro Station	Brewery	Bakery	Tapas Restaurant	Wine Shop
4	FAOURETTE	43.579090	1.415167	3	IT Services	Bakery	Metro Station	Wine Shop	Fast Food Restaurant

Figure 4. *Dataframe with most common venues and Cluster numbers*

This data is used to visualize the most common venues for each cluster. Bar charts seem best suited to this purpose. The bar charts for each cluster are shown below.



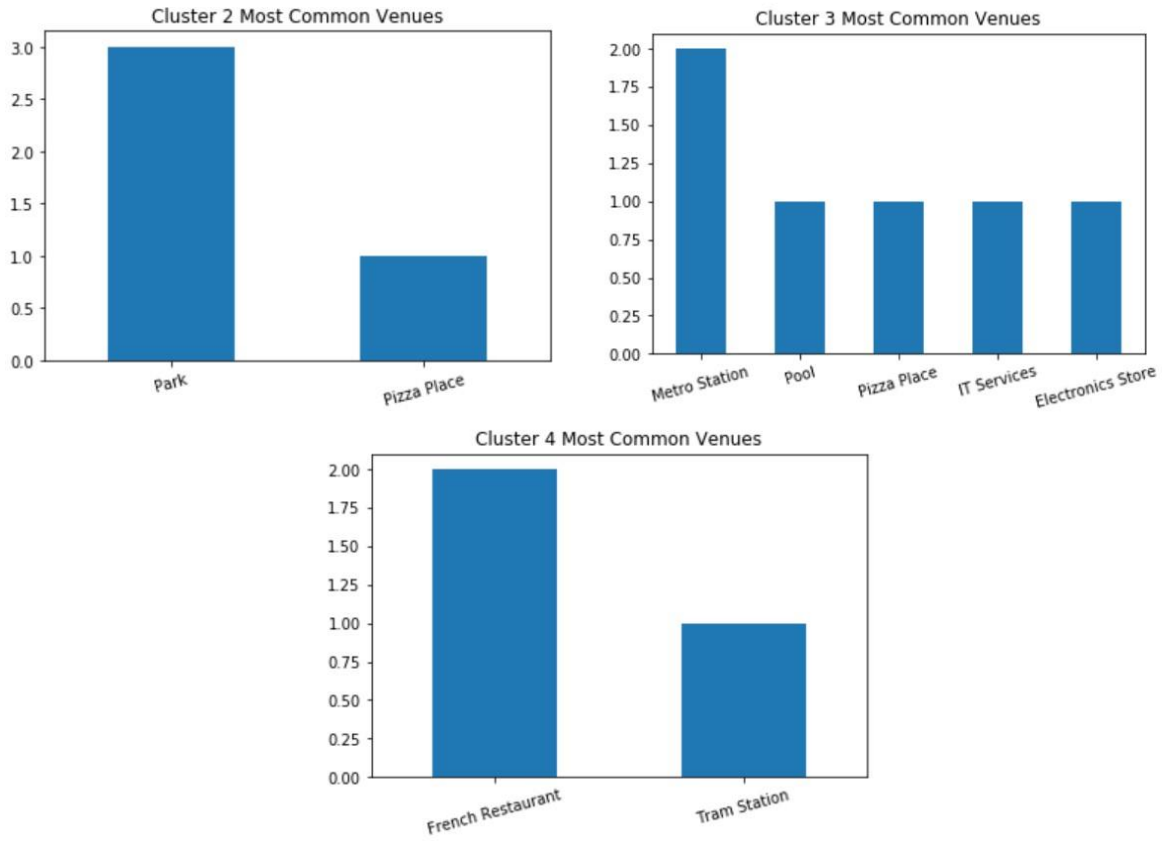


Figure 5. *Most common venues – Clusters 0 to 4 – Analysis 1*

From the above bar charts, I have made the following inferences for each cluster:

- **Cluster 0**: The most common venues for this cluster are restaurants and entertainment places. So these neighborhoods can be viewed as more open to night life and public entertainment, which will be a good factor regarding a Games and Recreation center.
- **Cluster 1**: The most common venue here is Shopping mall. Shopping can be considered mostly a day activity and also is associated with a more calm and peaceful environment. As such, this cluster doesn't really have any supporting information for me.
- **Cluster 2**: The most common venue in this cluster is Park, followed by Pizza place, but with a big difference between the both of them. The features do look a bit similar to Cluster 1, as parks will be more frequented by kids and families for a peaceful get together time. So this cluster can be deemed to be closer to a residential locality than an entertainment zone.
- **Cluster 3**: Many metro stations are part of this cluster, among other stores. It has the very good feature of easy accessibility, which is a big positive. The presence of pools show that it is partly an entertainment cluster too, with even a few pizza places to boast of. This cluster can also definitely be a candidate for a good location.
- **Cluster 4**: Having French restaurants and Tram stations are both positives for this cluster, in terms of accessibility and in preference of location. These neighborhoods also have the characteristics that I am looking for.

On analyzing the above results, I have allocated the following weightages to each cluster:

- Cluster 0 – 5
- Cluster 1 – 1
- Cluster 2 – 2
- Cluster 3 – 3
- Cluster 4 – 4

3.3 Analysis 2

This analysis deals with the popularity of venues. The Foursquare API call for this particular request is also done by 'explore' but with an extra parameter 'sort by popularity'. This is a Boolean, which when set to 1, retrieves the venue recommendations for a particular location in order of their popularity. The radius of search used for this call is set to 300 meters and the limit in number of venues is set to 3. Only the top 3 most popular venues for each neighborhood are of interest, which is sufficient to gauge the interest of the residents. The result is similar including details of the venues including the category. The data is once again compiled in a dataframe, sample of which is shown here.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ARNAUD BERNARD	43.607588	1.439794	Breughel l'Ancien	43.609363	1.439101	Pub
1	ARNAUD BERNARD	43.607588	1.439794	Creativ'Yogurt	43.605979	1.441231	Snack Place
2	ARNAUD BERNARD	43.607588	1.439794	George & the Dragon	43.607416	1.439559	Pub
3	LES CHALETS	43.613401	1.442695	La Comédie de Toulouse	43.611599	1.441049	Comedy Club
4	LES CHALETS	43.613401	1.442695	La Pente Douce	43.611295	1.444354	Restaurant

Figure 6. *Dataframe with all the Most Popular Venues*

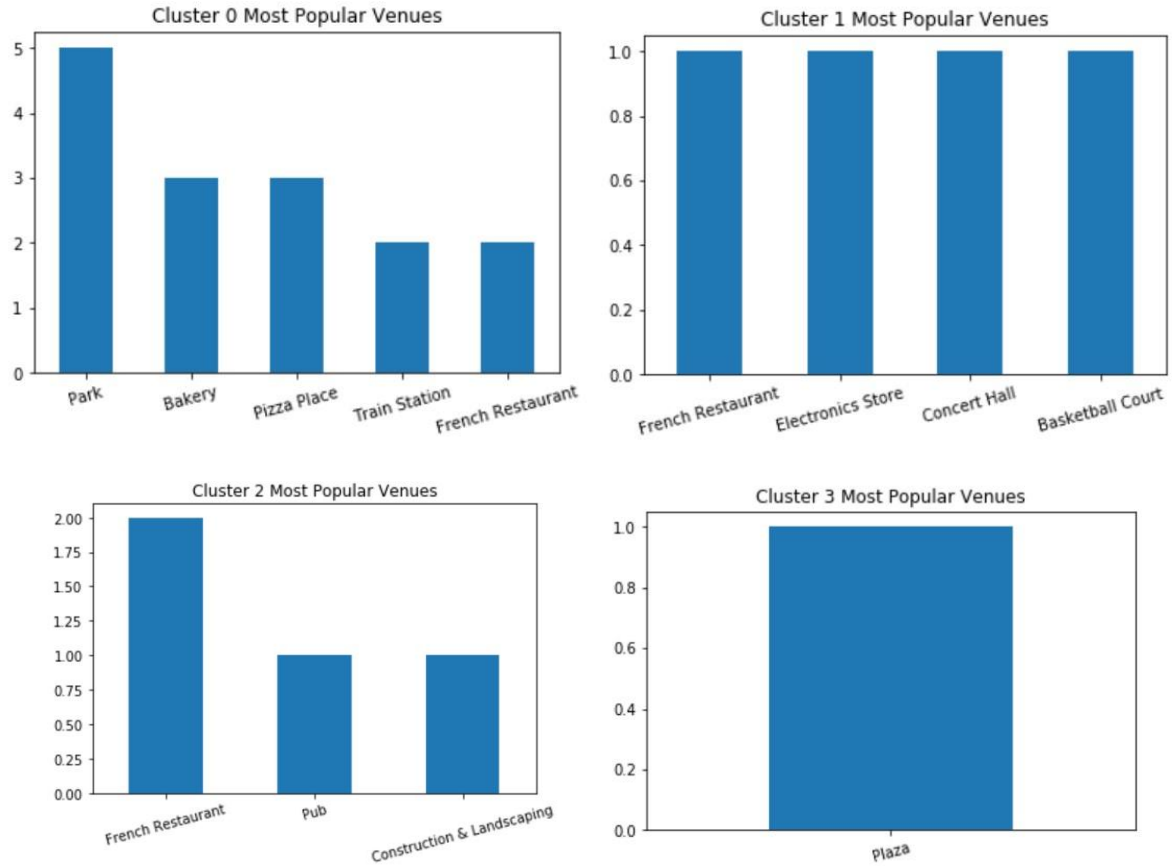
From this dataframe, my algorithm uses the Venue Category column once again to compile the 3 most popular venue categories for each neighborhood. K-Means clustering method will be used again to divide all the neighborhoods into 5 clusters – Clusters 0 to 4. To use the K-Means clustering algorithm, one-hot encoding is used in my algorithm to obtain a dataframe that represents the relation between each neighborhood and all the venue categories numerically.

After the clustering is performed, the cluster number for each neighborhood is appended to the final comprehensive dataframe for the second analysis – sample shown below.

	Cluster Labels	Neighborhood	1st Most Popular Venue	2nd Most Popular Venue	3rd Most Popular Venue
0	0	ARNAUD BERNARD	Pub	Snack Place	Pub
1	4	LES CHALETS	Comedy Club	Restaurant	Bar
2	0	MINIMES	Argentinian Restaurant	Pedestrian Plaza	NaN
3	4	SAUZELONG - RANGUEIL	Metro Station	Bakery	Tapas Restaurant
4	0	FAOURETTE	IT Services	Bakery	Metro Station

Figure 7. *Dataframe with most popular venues and Cluster numbers*

To visualize the clustered results of most popular venues, bar charts are once again used. The bar charts for each cluster are shown below.



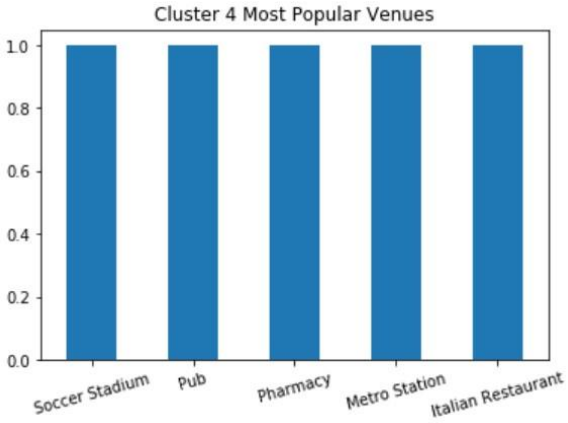


Figure 8. *Most popular venues – Clusters 0 to 4 – Analysis 2*

From the bar charts, these inferences have been made for each cluster:

- **Cluster 0**: The most popular venues for this cluster are parks and bakeries. This information tells that these neighborhoods are probably residential places with a calm and peaceful environment. So these locations are perhaps not ideally suited for a Games and Recreation center.
- **Cluster 1**: With French restaurants and concert halls as popular venues for these neighborhoods, this cluster is entertainment friendly and also food friendly. Both are positives in regards to my analysis and these locations are better candidates.
- **Cluster 2**: Restaurants and pubs are more popular here, meaning that this cluster is open more to nightlife. This feature is quite perfect for a business with peak times generally during late evenings. A popular nightlife spot means that more people are attracted to the location resulting in a larger pool of customers. Thus, these neighborhoods are ideal locations in this regard.
- **Cluster 3**: Plaza is shown as the most popular venue in this cluster. As such, it does not provide much incentive to feature as a candidate location.
- **Cluster 4**: With a soccer stadium, pubs, metro station and restaurants as the venues more frequented in these neighborhoods, I can say it is more open to nightlife as well. The particular aspect regarding stadiums is that on matchdays, it catches everybody's attention. And its bad business for other entertainment centers. Also it is very hard to gauge the interest of residents in the location on rest of the days. So this cluster can be said to be a good location, but with lot of competition as well.

Based on the above results, the following weightages are allocated to each cluster:

- Cluster 0 – 2
- Cluster 1 – 3
- Cluster 2 – 5
- Cluster 3 – 1
- Cluster 4 – 4

3.4 Analysis 3

As discussed earlier, the idea is to use certain 'key' venue categories to generate useful information. For this analysis, I have selected 'Nightlife Spots' as a key venue category. It is possible to identify many key categories in order to perform a very comprehensive analysis, but for now, I have chosen an important one which can produce very relevant results.

The Foursquare API call of 'search' will be used in this analysis. This call returns a result with venues according to search parameters like a search based on query, proximity or location. It also includes search based on category Id, which is of importance to this project. The API call with the 'Nightlife Spots' category Id returns the venues corresponding to this category for each of the neighborhoods. Regarding the parameters, the radius of search is selected as 250 meters for better exclusivity as the distances between neighborhoods in Toulouse are quite small. The limit is set 100 which is a sufficiently high number not likely to be crossed in this search.

The calls for each neighborhood is once again compiled in a dataframe, a sample of which is shown here.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ARNAUD BERNARD	43.607588	1.439794	La Fabrique	43.607444	1.439835	Bar
1	ARNAUD BERNARD	43.607588	1.439794	George & the Dragon	43.607416	1.439559	Pub
2	ARNAUD BERNARD	43.607588	1.439794	Café Des Facs	43.607466	1.439919	Bar
3	ARNAUD BERNARD	43.607588	1.439794	Bichette	43.607560	1.439298	Wine Bar
4	ARNAUD BERNARD	43.607588	1.439794	Campagne	43.606033	1.441295	Wine Bar

Figure 9. *Dataframe of Nightlife spots*

In this analysis, as the data is very straightforward in its information, I will not be using a clustering method to segregate the neighborhoods. Instead, my algorithm compiles the total number of nightlife spots for each neighborhood and selects the 10 neighborhoods with the highest results. These neighborhoods can be visualized in a bar chart as shown below.

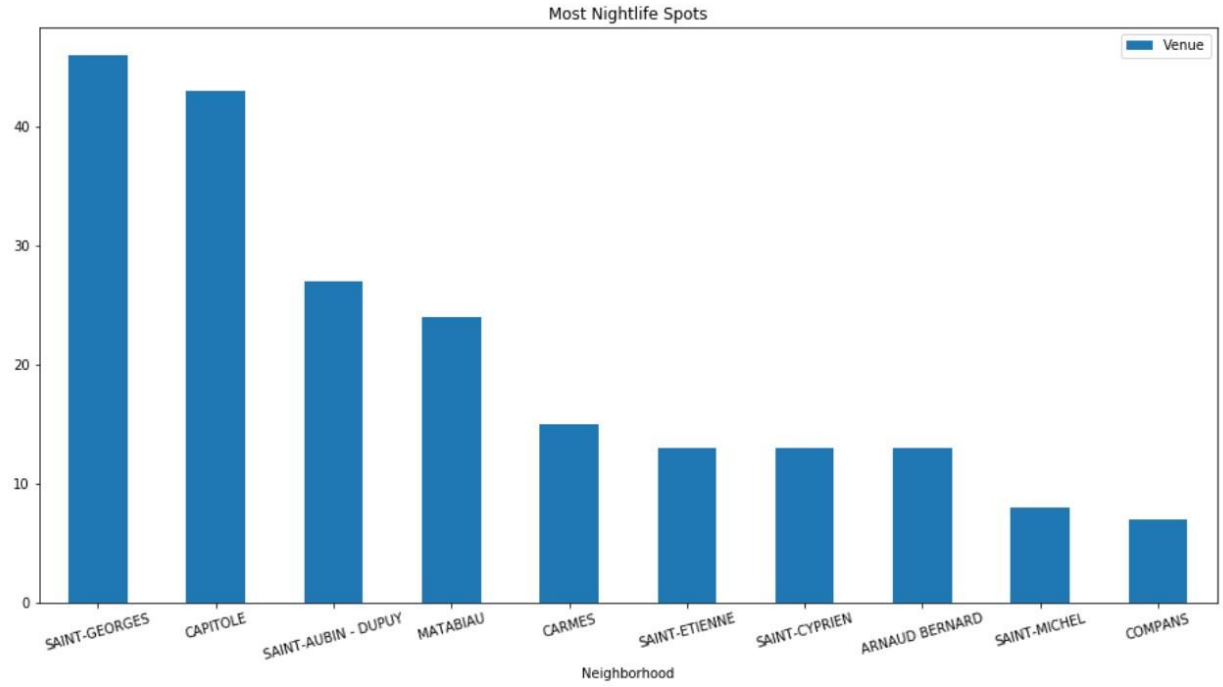


Figure 10. *Most Nightlife spots – Analysis 3*

At the end of the three analyses, I have a set of relevant results which have to be compiled together for the final output. This is discussed in the next section.

4. Results

To obtain the final results of the analysis, I am using a points system to gauge the performance of each neighborhood in each of the three analyses. The points allocated are Points C – for common venues analysis, Points P – for popular venues analysis, and Points N – for nightlife spots analysis.

- **Points C** – The weightages for each cluster of the first analysis have been allocated. Points for each neighborhood is calculated by multiplying the weightage with the total number of the most common venue in the cluster.
- **Points P** – The weightages for each cluster of the second analysis are also allocated. Points for each neighborhood is calculated by multiplying the weightage with the total number of the most popular venue in the cluster.
- **Points N** – The weightage for the third analysis is set as 1. Points for each neighborhood is calculated by multiplying the weightage with the number of nightlife spots for the neighborhood.

The final tally of points for each neighborhood will be the sum of Points C, Points P and Points N. A sample of the dataframe with points tally is shown here.

	Neighborhood	Points C	Points P	Points N	Total Points
0	ARNAUD BERNARD	32	10	13	55
1	LES CHALETS	32	6	2	40
2	MINIMES	32	10	1	43
3	SAUZELONG - RANGUEIL	8	6	2	16
4	FAOURETTE	8	10	0	18

Figure 11. *Points tally for each neighborhood*

The 10 neighborhoods with the most number of total points have been selected as the best candidates for the location of the opening of a Games and Recreation Center. These 10 neighborhoods are visualized below with the points tally.

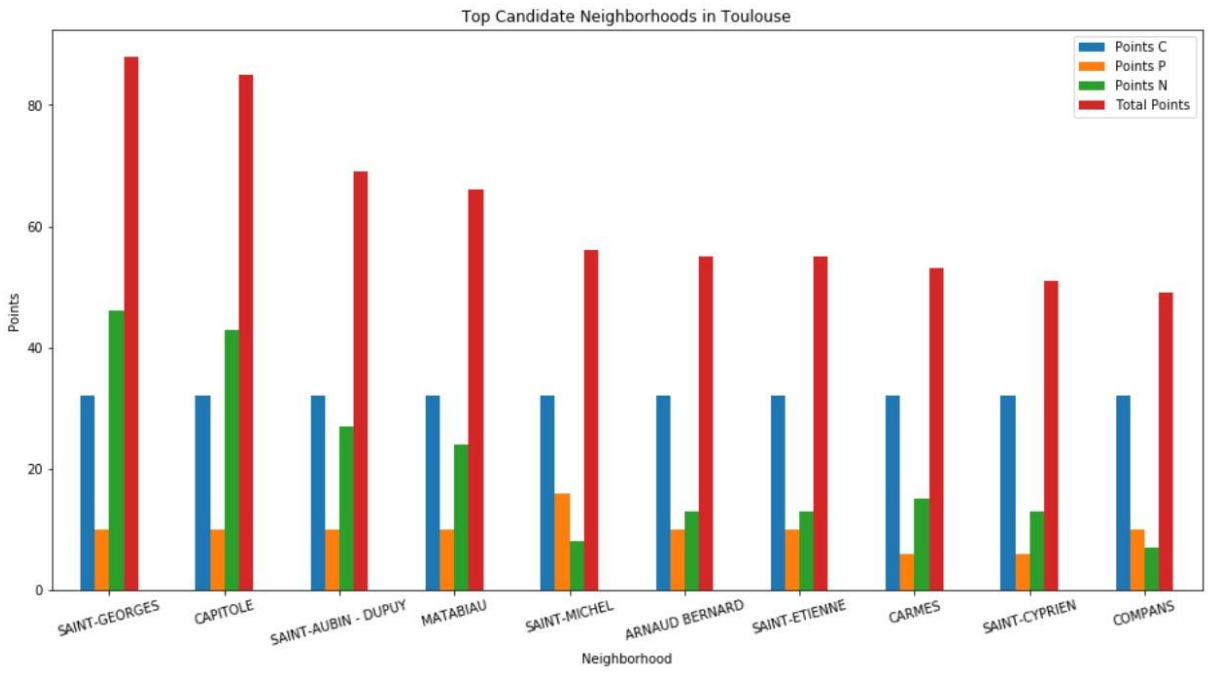


Figure 12. *Top 10 Neighborhoods – Points scored*

Finally, I can once again generate the map of Toulouse with the top candidate neighborhoods marked in Red to best visualize the results of this project.

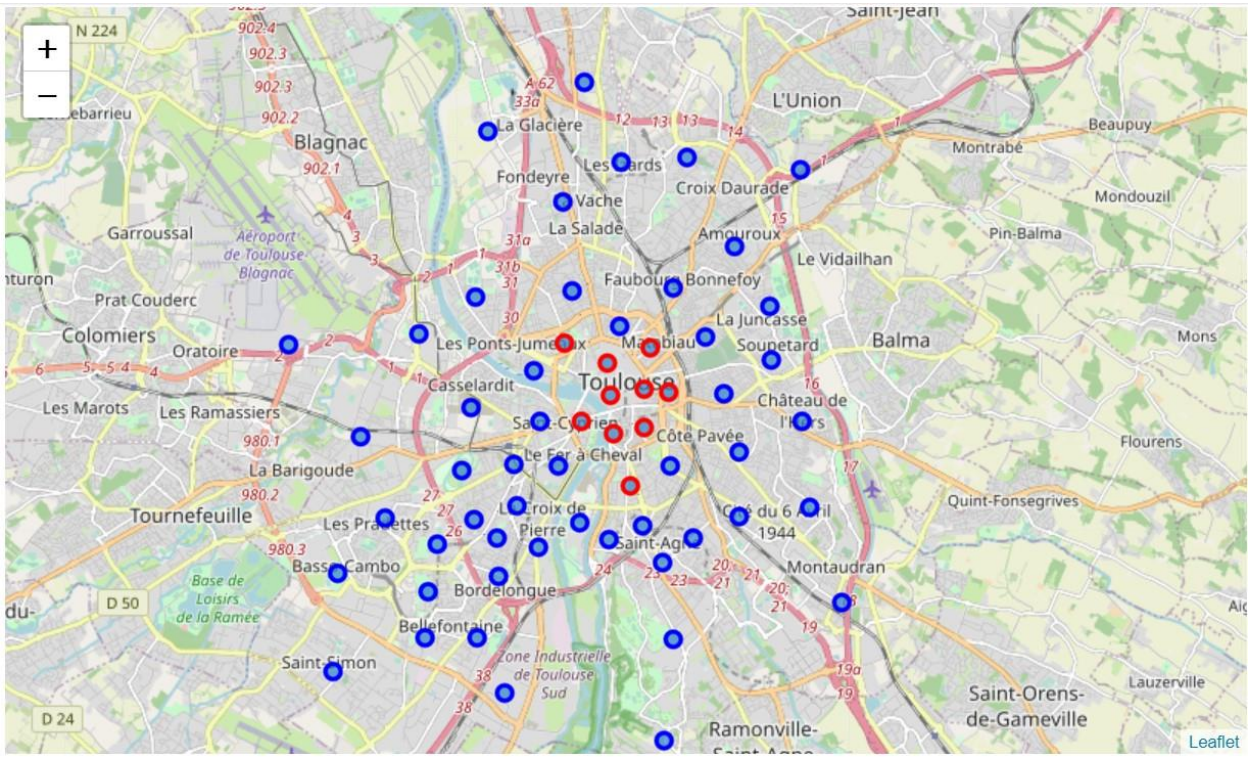


Figure 13. *Map of Toulouse – Top 10 Candidate locations in red*

5. Discussion

On viewing the final results, I can say that the output makes a lot of sense as the best candidate locations selected are very close to the town center. This takes into account many aspects – large customer pool, ease of accessibility, safety, public presence during late nights and weekends and also availability of essential services. It is pleasing to know that these factors have automatically been taken into account without specifying any of them specifically during the analysis.

It is also clear that some neighborhoods have been filtered out even though they seem to be perfect locations from the outset. It shows that the reliance on the information that data provides, along with correctly selected factors during the analysis, is capable of producing the best results. Thus, the final results produced can be trusted with a high level of confidence.

Finally, I can remark that the analysis can always be improved. Other factors that affect the success of a business opening can be used to generate data. The premium features of Foursquare can also be used to generate additional relevant data which can increase the accuracy and dependability of the results. Also many key venue categories can be used to perform a very comprehensive analysis covering all aspects of a location.

6. Conclusion

The aim of this project to find optimum candidate locations for the opening of a Games and Recreations Center have been accomplished and the top 10 candidate neighborhoods have been identified and visualized.

The different parts of the analysis have been performed accurately and well documented. These can be used as bases for more complicated analyses in the future.

Further improvement and additions have been identified and noted for the benefit of data science enthusiasts.

The sections of code, interactive maps and visualizations of all dataframes used in the analysis are part of my notebook on GitHub. The link to it is given here.

https://github.com/anil-sasi/Coursera_Capstone/blob/master/Capstone_Toulouse.ipynb

7. References

1. Foursquare - https://developer.foursquare.com/docs/
2. Toulouse city data - https://data.toulouse-metropole.fr/explore/dataset/
3. Toulouse data Wikimedia - https://commons.wikimedia.org/wiki/Category:Districts_of_Toulouse
4. Python notebook markdown guide - https://www.markdownguide.org/cheat-sheet/
5. Python Pandas package help - https://pandas.pydata.org/pandas-docs/