**Analysing districts in Paris for new shopping malls**

Ethan Su

May 25, 2020

1. **Introduction** 
   1. **Background**

Grandeur Limited is one of Asia largest diversified real estate groups. Headquartered in Singapore, Grandeur owns and manages a global portfolio focusing on real estate, infrastructure and private properties. It’s portfolio compromise of integrated developments, shopping malls, lodgings, offices, homes and funds.

Present across more than 100 cities in over 30 countries, the company focuses in Singapore, China and Japan as the core markets while it continues to expand across European countries. The company currently is focusing its resources in developing the portfolio in European countries, especially in France and Italy. Under the current directions of the management, France is the next real estate is the next agenda. The company plans to start its investment in France, Paris and mainly focus in shopping malls, offices and accommodation.

* 1. **Problem**

Based on the investment strategy of Grandeur, the problem would be to identify the appropriate districts in Paris that do not have many similar venue categories that would increase the competition intensity. This project aims to analyse the district and their common venues to estimate the best area to invest a shopping mall.

* 1. **Interest**

Grandeur will be interested in the analysis because it is essential for competition analysis and identifying the complementary services which greatly boost the investment values and Return of investment.

1. **Data Acquisition and Cleaning** 
   1. **Data Sources**

The data sources comprise of two different portions. The first data set comes from Data France that identifies all the Arrodissments in Paris. And for the second data set, it comprises of all the venues within each Arrodissments in Paris which is download from Foursquare datasets. Each dataset will not be sufficient to form a comprehensive analysis without being combined together. Hence, both of these datasets will be joined together forming a entire new data set.

* 1. **Data cleaning**

Data downloaded or scrapped from the sources were cleaned before combining them into one table. In the first glance, the integrity of the data sources is comprehensive. There are no missing values or “null” values. However, there were many unnecessary columns that needs to be removed before both can be joined together.

In the first dataset from DATA France, there are many unnecessary variables such as “NSQAR, CAR.1, SURFACE, PERIMETER and etc.” These variables do not contribute to the analysis in the project and hence deemed as irrelevant and be removed from the dataset. Thus, after data cleaning, the data table comprise of “Neighbourhood, Arrondissment\_num, French\_Name, Latitude and Longitude”. After cleaning, the data set clearly shows that Paris has 20 distinct Arrondissment with their respective neighbourhood in French name together with the coordinates.

Second in line is to process the dataset download from Foursquare API which do not have any multiple entries, repeated data, empty cells and null values. In order to get the data from Foursquare, we need to use the identified neighbourhood French name, longitude, and lati from the previous dataset. Data extracted is in json format and it will be structured into a panda dataframe. In the Json file, it has been normalised to flatten the json file, filtered the column and followed by filtered using the category. In addition, specifically within the category’s column, it will be clean by adding a separation between the words. The clean dataset from the second database will provide us with the name of the shops and which categories it belongs. This will enable us to clearly pinpoint what are the different venues within each Arrondissment.

* 1. **Feature Selection**

After data cleaning, there were 20 Arrondissments identified in Paris and there are 78 venues identified within a radius of 500 meters from 3eme Ardt, and for each Arrondissments there are specific set of venues within their 500meters radius. Given this we are able use these venues as the features that we can select based on Grandeur specification on which they require. Then we will be able to set the districts and analyse which district that is best suited for Grandeur to open their shopping mall. Some features will not be useful for the analysis hence it will be useful to have those removed to provide a comprehensive and accurate analysis.

1. **Data Analysis Methodology**

In order to conduct the exploratory analysis, there will be 7 steps taken to analyse the data based on the data set download over the sources:

1. Geographical Mapping
2. Function Creation using venues for all the neighbourhoods in Paris
3. Identification of the number of venues in each neighbourhood
4. Calculation on the number unique venues categories
5. Grouping rows by neighbourhood and calculate the frequency of occurrence in each category
6. Identify the top 10 most common venues
7. Conversion of data into pandas dataframe

The first step is to clearly distinguish the all the 20 different Arrondissments in Paris. Using the geopy.geocoders library, a map of Paris will be created using the above latitude and longitude values and a circle market will be drawn as shown in the image below (Image 1). From this map layout, we will be able to clearly identify how far each Arrondissments are apart from each other. Also, it can also be served as a proximity marker to ensure that each investment will not be so close to one another. This enables Grandeur to maximise the outreach to different customer group with minimal overlap.

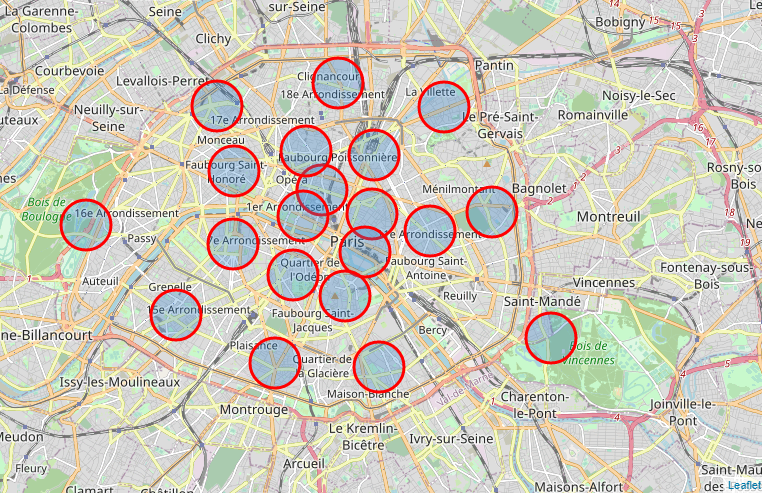


Figure Arrondissments of Paris

Secondly, we need to process the different venues thus, we created a new function called “get\_category\_type” in order to create categories based on the venues which will be later translated to the feature extraction as mention previously. As shown in the Figure below, presenting the function definition of extracting the category type.

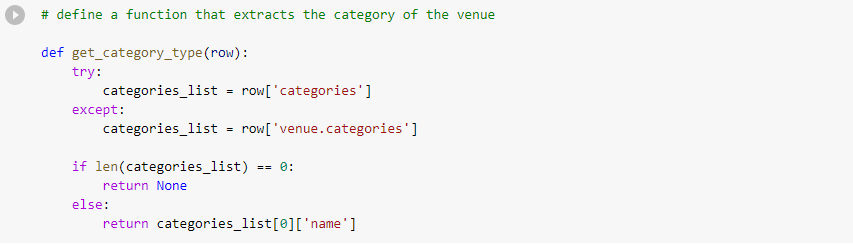
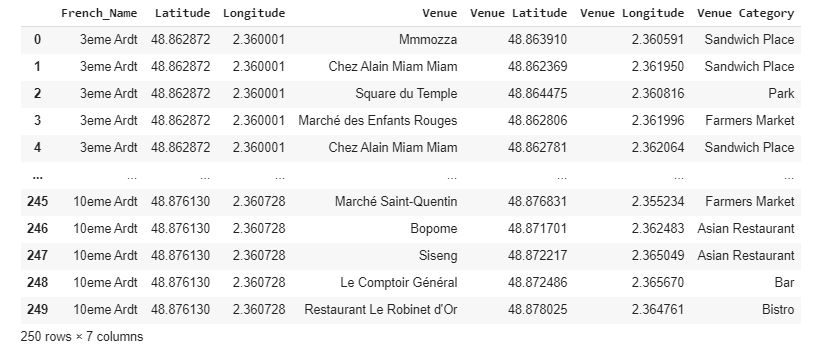


Figure Get\_category\_function

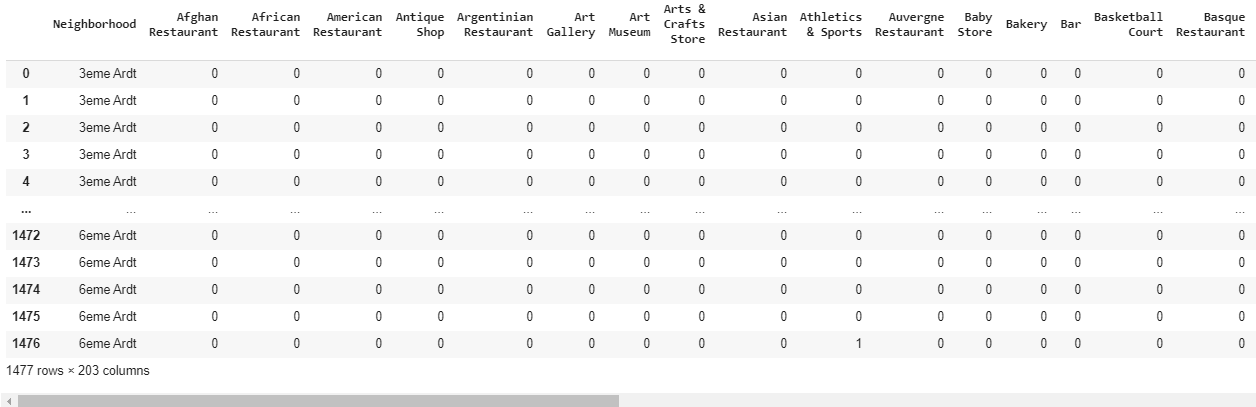
Combined with another function that creates the nearby venue function for all the neighbourhood in Paris, the table generated enables a more in-depth analysis of specific restaurants presented in each Arrondissments as shown in Table 1 below.

Table Detail Analysis of specific venue category



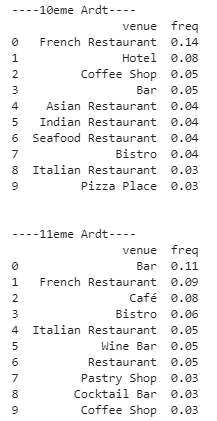
Furthermore, within each neighbourhood, a clearer analysis is conducted to identify the specific venues. In this step, the neighbourhood that have the specific venue will have “1” under the venue column and “0” for non-existence. This will clearly define all the venue down to the details in each neighbourhood as shown in table 2 below.

Table Venue Analysis in Neighbourhood



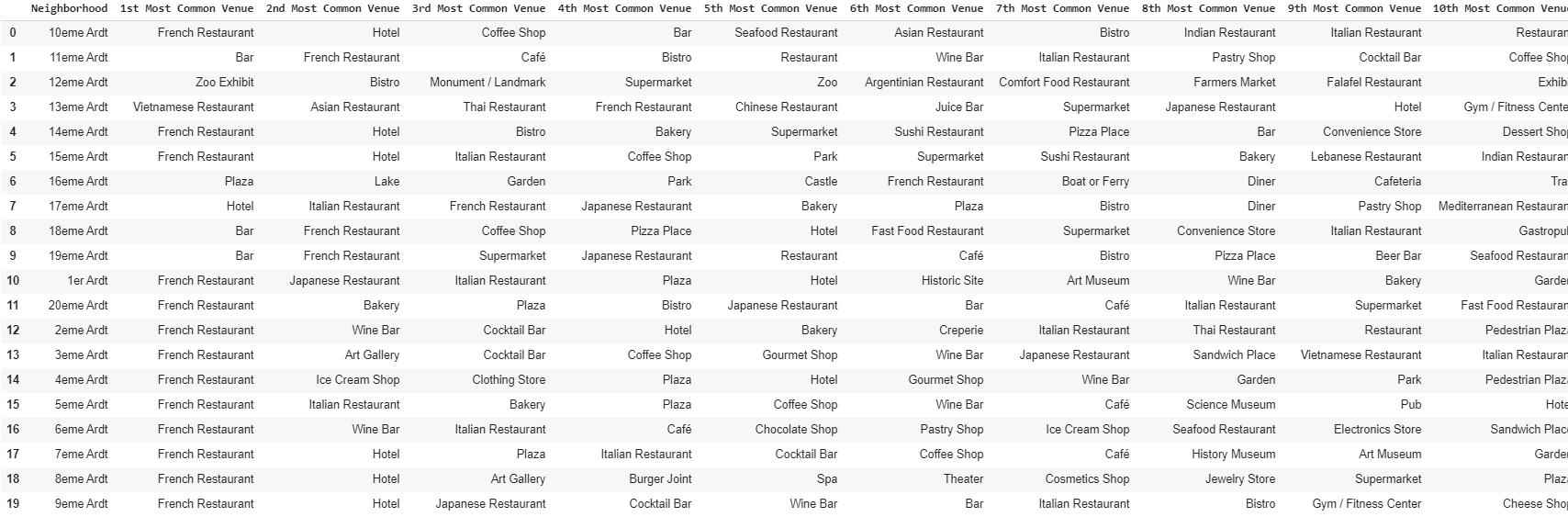
Based on the list of venues, it will be impossible or redundant to include all venues as such the next step will be to normalize the value into percentage form and group all of them together and sort with respect to the frequency level of the top 10 venues in each Arrondissment as shown in the table 3 below.

Table Frequency Analysis on each Neighnourhood



Finally, by compiling all these data together into 1 main table as shown below, we will be able to have a clear overview of the top 10 most common venue in each neighbourhood of which we will be able to bring it forwards for the later analysis to identify the exact neighbourhood for Grandeur investment.

Table Overall Analysis



1. **Analysis**

Based on the initial analysis from Figure 1, we can identify that there are several potential Arrondissement within the central location to build the shopping mall; around 1-9 Arrondissment Temple (3 eme Ardt), Hotel-de-Ville (4eme Ardt) etc. These locations serve as the prime spot because of their human traffic movement and density and proximity to the tourist sites which is ideal for shopping malls to be location. The business types criteria were specified by the client namely that the based on previous trends, 'Ice Cream Shop', 'Cafés' and 'Wine Bars' should be around the location where the ‘Shopping Mall' as it will generate addition value in drawing crowd towards the neighbourhood. Therefore, based on the violin chart below at the frequency of the occurrence throughout the Paris neighbourhoods, isolating the categorical venues. In addition, French restaurant will also be added into the analysis for the client in hope to identify the district with high density of French Restaurant. These would value add to the shopping mall. Hence, the goal of the analysis is to identify the suitable location of which we can maximise the value-added services from surrounding business to complement the shopping mall.

These are the venue categories that the client wants to have an abundant density of for the ideal store location. Therefore, for better visualisation, a violin plot from the seaborn library - the choice of the plot would greatly visualise the frequency distribution datasets, they display a density estimation of the underlying distribution. Hence, based on the results display on the violin plot we will be able to narrow down the specific district for the client.

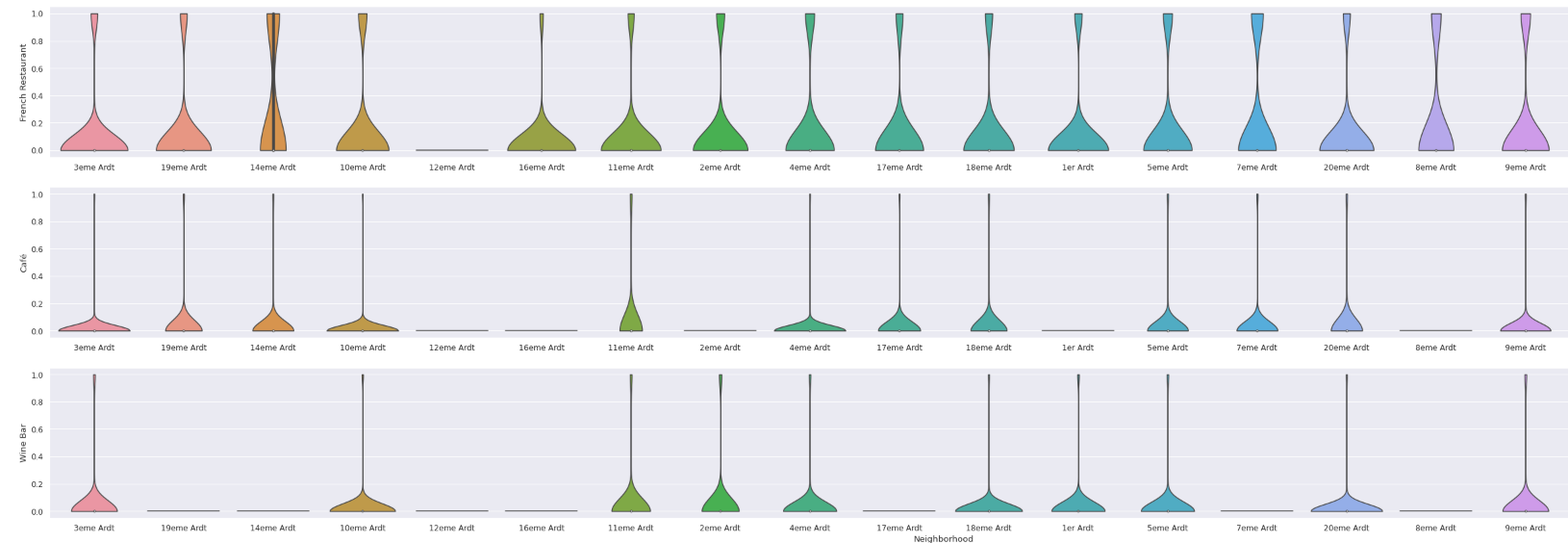


Figure Violin Chart analysis

From the analysis, we identify that majority of the districts in Paris have high density of restaurants within them. And according to the pre-specified criteria by the client, there are 9 neighbourhood identified that are suitable to open new shopping malls and they are as follows:

Neighbourhoods

1. 3eme Ardt

2. 10eme Ardt

3. 14eme Ardt

4. 4eme Ardt

5. 18eme Ardt

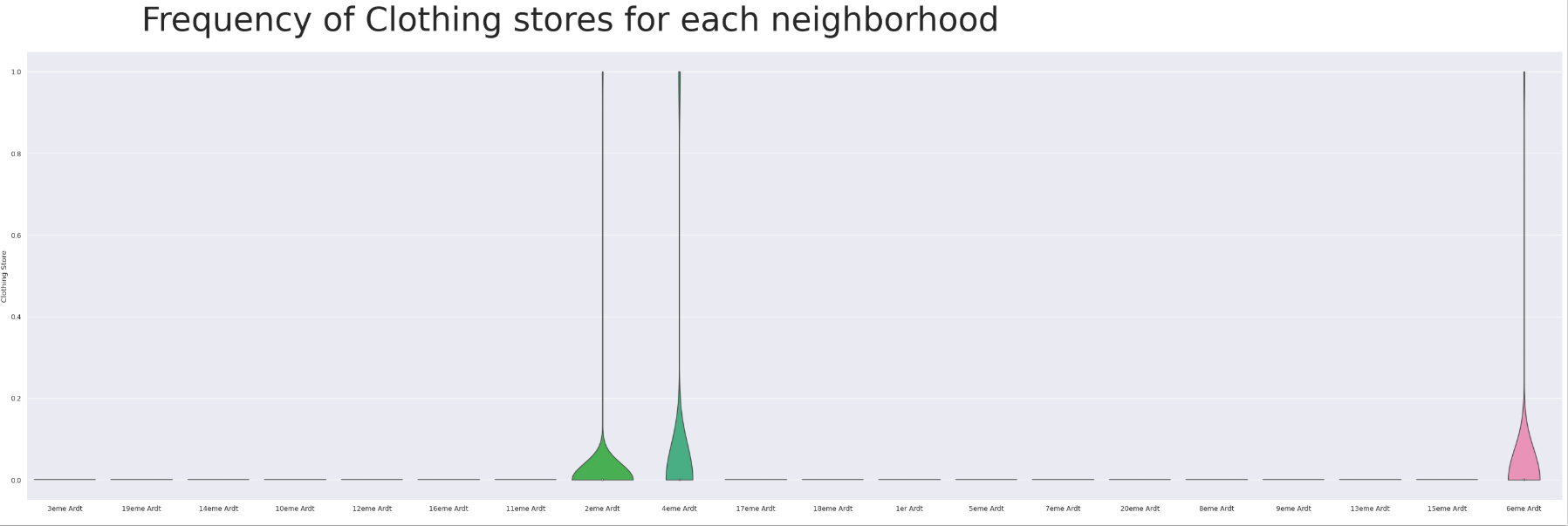
6. 11eme Ardt

7. 15eme Ardt

8. 9eme Ardt

9. 4eme Ardt

However, apart from the criteria specified by Grandeur, more explorative analysis can be down to further rank the neighbourhood accordingly. Based on these 9 neighbourhoods, we will include ‘Clothing\_store” venue category into the analysis then we will be able make inferential analysis from the data using domain knowledge because clothing store is considered as a direct competition to Grandeur shopping mall investment. Therefore, based on our analysis in the frequency of Clothing store in each neighbourhood, 2eme Ardt and 4 eme Ardt present a high frequency as shown in the figure below. This indicates to us that these 2 neighbourhood should be ranked as the last or excluded from the investment plans for Grandeur to minimise the competition intensity which might lead to lower ROI.



1. **Results and Discussion**

After the series of analysis, the results from the inferential analysis using the data together with the domain knowledge in retailing and real estate investment, the list of focus neighbourhoods are reduced to just 3 instead of the previous 9. The reasoning for the 3 choices was that we have tighten the criteria of the client requirement from best to have either “French restaurant”, “Café Bar”, “ wine Bar” and “ice-cream shop” to the criteria of at least 3 out of 4 requirements must be met. The rational for this is to ensure that each shopping mall is able to capture at least 75% of the traffic due to the presence of these shops around it. If we loosen such criteria, the combinational location will not be able to bring out the maximum impact on the profit for Grandeur.

The increase in the criteria to have the inclusion of more category, eventually enable us to narrow down and rank the suggest districts for the new stores and provide better location for the brand.

Therefore, the final 3 prospective neighbourhood that met all the criteria are:

1. 3eme Ardt: Arrondissement 3, Temple
2. 4eme Ardt: Arrondissement 4, Hotel-de-Ville
3. 9eme Ardt: Arrondissement 9, Mairie du

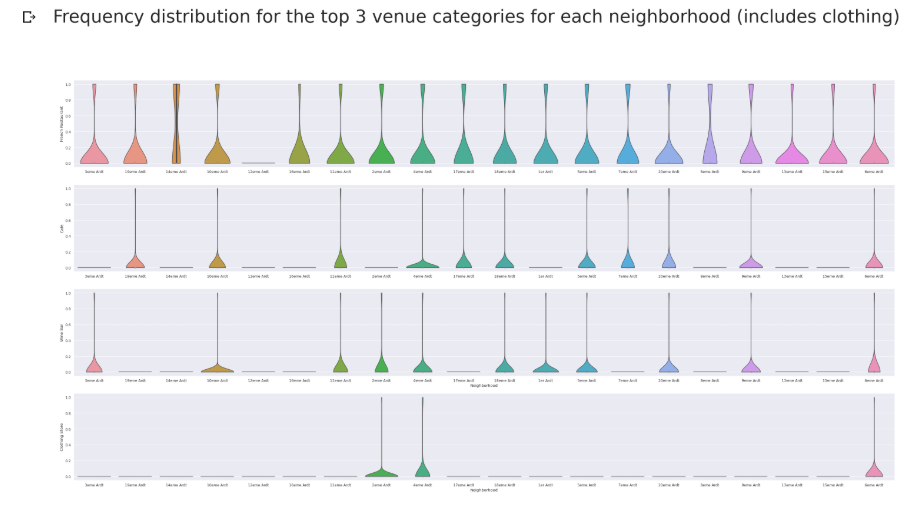
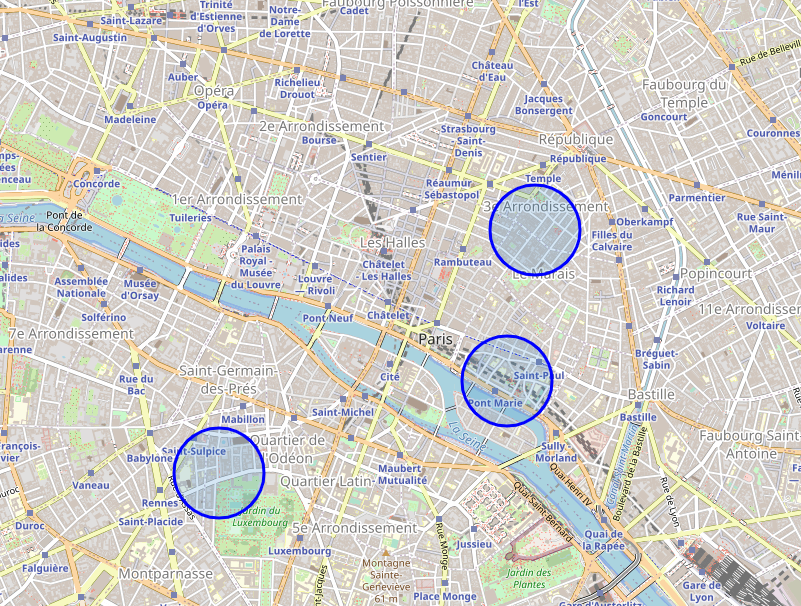


Figure Overall Comparison5

The map below presents the location of the stores that we plan for Grandeur to start their investment policies of investing in real estate in these 3 locations. Coupled with other knowledge, these three locations are rank in accordance to which phase they should be invested in.



Conclusion

In conclusion, there could be other methodologies to conduct the analysis based on the dataset that has been collected. This methodology that has been selected is mainly due to the simplicity in execution and the appropriate visualization of the data analysis. In addition, the output can meet the criteria of Grandeur in identifying their requirement of the venues within each neighbourhood. However, if given more domain knowledge on setting up a shopping mall, more variables could be added into the data for a detailed analysis. The analysis and the results can be further process with the inclusion of human traffic density and competition intensity.

Without the analysis from the dataset, Grandeur might result in opening their shopping mall in a substandard area which will decrease their ROI. Hence a data driven decision will definitely lead to a better solution and strategy planning.