The Battle of Neighborhoods

The Battle of Neighborhood is a capstone project from Coursera and IBM. In this project, students are required to perform analysis using at least the Foursquare's service to solve business problems. Brief description of my take on The Battle of Neighborhoods: I will use services provided by Foursquare and other methods to find a suitable location to open a cinema in Paris.

1. Introduction

An investor came to me with an idea of opening a cinema in Paris. He told me that he didn't care about the rental price or the facility. So my job is to recommend to him the most suitable location to open a cinema in Paris.

He explained that in customer point of view, watching movie is a part of whole afternoon or night activities. Cinema should has many restaurants and shopping places nearby. Transportation is also an important factor. Customer can walk to cinema within 3 minutes from public transport facilities such as bus stop and metro station.

2. Data

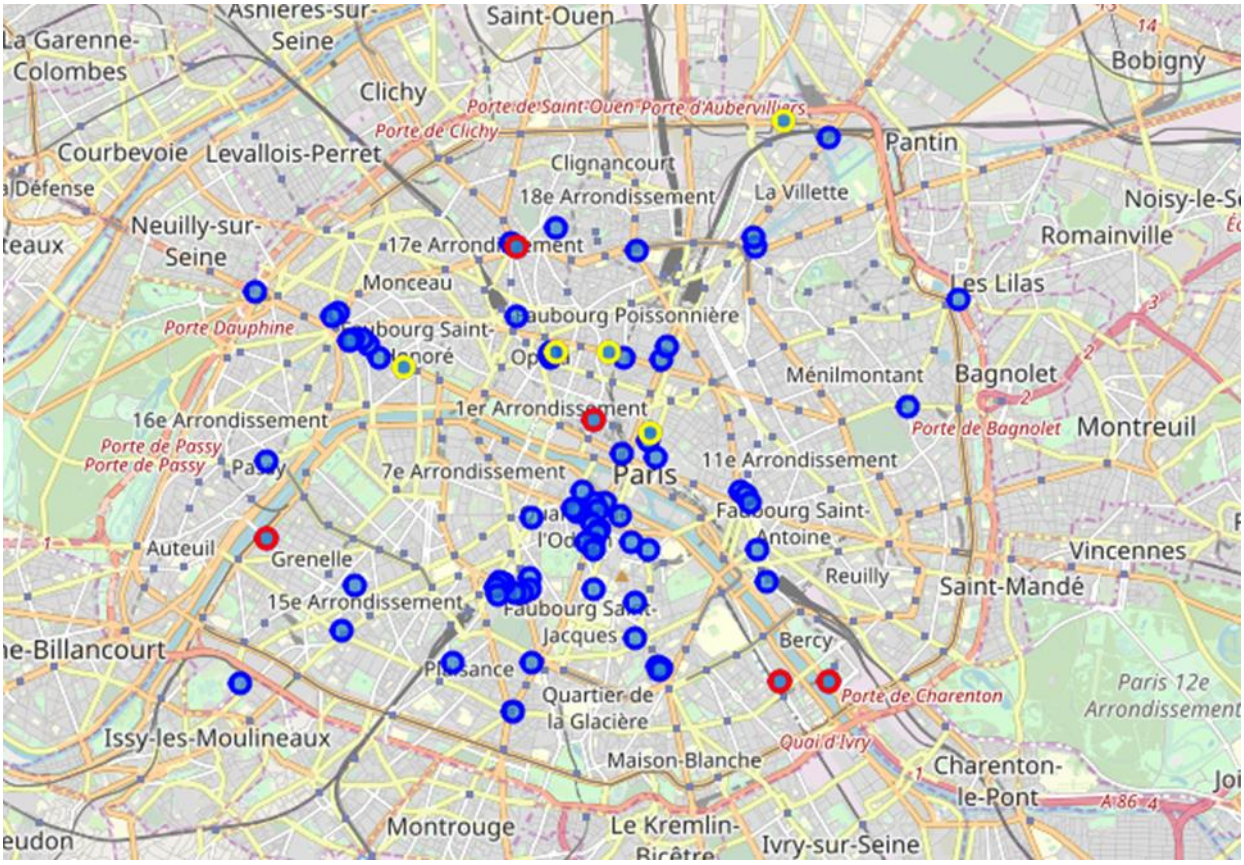
To solve this problem, I need to find the data .The data I'll be using:

- Name, address and geographical coordinates of all cinemas within Paris. Name and addresses will be extracted from this website http://www.allocine.fr. Geographical coordinates will be produced using Bing map API.
- Top 5 most popular cinemas in Paris from the same website and the rating of these cinema done by me and my friend.
- Top 5 possible locations to open a cinema. I will randomly pick 5 addresses from the list of all cinemas in Paris and pretend that those locations don't have any cinema.
- List of venues by categories around each cinema using Foursquare API.

3. Methodology

After defining the source of data, I will perform the collection of data and several cleaning processes. With the data obtained, I drew a map using Folium library. This map shows

all the cinemas within Paris (all circle in blue and red, red circles indicate the most popular cinemas, yellow circles indicate 5 possible locations).



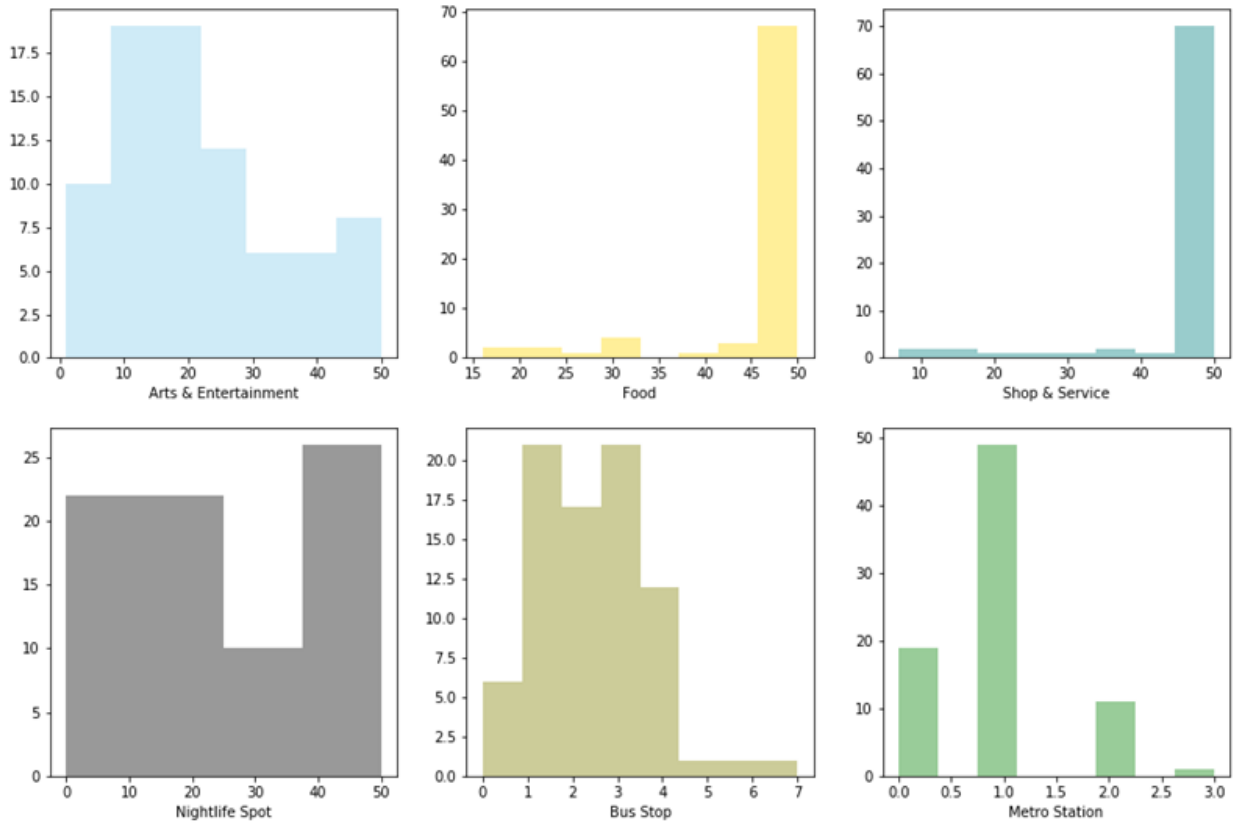
When all the data is ready, I will perform descriptive statistics, creating histogram and perform correlational testing to have a general idea about the data of venues returned by Foursquare.

Descriptive statistics and histogram of the number of venues grouped by category. Here are several observations:

```
In [85]: 1 df_venues_count.describe()
```

Out[85]:

category	Arts & Entertainment	Bus Stop	Food	Metro Station	Nightlife Spot	Shop & Service
count	80.000000	80.000000	80.000000	80.000000	80.000000	80.00000
mean	21.025000	2.300000	46.775000	0.925000	26.262500	46.81250
std	13.196638	1.408833	8.048862	0.651677	16.973355	9.46826
min	1.000000	0.000000	16.000000	0.000000	0.000000	7.00000
25%	11.000000	1.000000	50.000000	1.000000	11.000000	50.00000
50%	18.000000	2.000000	50.000000	1.000000	21.000000	50.00000
75%	28.250000	3.000000	50.000000	1.000000	45.250000	50.00000
max	50.000000	7.000000	50.000000	3.000000	50.000000	50.00000



- "Arts and Entertainment" venues seem to have a normal distribution
- We can see that "Food" and "Shop & Service" has a very similar distribution. This is due to the limit of venues returned by foursquare search and the fact that there are so many food and shop just within 200 meters of a cinema. Almost all cinemas has more than 50 of these venues within 200 meters.
- Most of the Cinema has more than 50 "Nightlife Spot" venues with 200 meters radius.
- The number of "Bus Stop" also seem to have a normal distribution
- Most cinema has only 1 metro station in its 200 meters radius.

After this, I will use machine learning method of content-based recommender to create a profile of venues around the most popular cinemas. With this venues profile, I can predict which possible locations will be most likely to become a successful cinema.

```
In [140]: 1 top_cinemas_profile.sort_values(ascending=False)

Out[140]: Shop & Service          17.339535
          Food                   16.432353
          Nightlife Spot          9.326000
          Arts & Entertainment    7.338776
          Bus Stop                5.028571
          Metro Station           4.066667
          dtype: float64
```

4. Result

After all, we found that among 5 possible locations in Paris ,"50, rue Rambuteau 75003 Paris 3e arrondissement" is most suitable location to open a new cinema with the rating of **0.906**.

```
In [136]: 1 df_final
```

Out[136]:

	location	address	latitude	longitude	rating
0	L1	50, rue Rambuteau 75003 Paris 3e arrondissement	48.86158	2.35229	0.906192
4	L5	24, bd Poissonniere 75009 Paris 9e arrondissement	48.87136	2.34489	0.869479
1	L2	27-33 avenue des Champs-Elysées 75008 Paris 8e...	48.86946	2.30748	0.804068
2	L3	32, bd des Italiens 75009 Paris 9e arrondissement	48.87133	2.33530	0.763140
3	L4	166, boulevard Mac Donald 75019 Paris 19e arro...	48.89899	2.37676	0.104070

5. Discussion

While doing descriptive analysis we found some interesting point:

- 50 is the limit of venues return by search venues method provided by Foursquare. It's remarkable that we can easily find 50 venues of food places, shopping and nightlife spots within 200 meters around a cinema. This can show us how lively Paris is.
- This also show us how popular Foursquare is in Paris since all these venues are collected by check-ins of users. This can be due to the fact that Paris is a big touristic city. People from all around the world come to visit and check-in to share it with their friends.
- If we can somehow remove the limitation of venues returned by foursquare search method, our result can become even more accurate. However, this will certainly become more expensive and I don't have the means to do this.

6. Conclusion

Although some of the data are still partially fictive. We have come up with a way to solve the business problem of finding the most suitable location to open a cinema. In order to apply this in real life, we need to improve several things:

- Increase the radius of search venues since when people go out, they can go for more than 200 meters to find a place to hang out.
- Remove the limitation of search venues returned by Foursquare, doing this remove the cap of our data, hence improve greatly the accuracy.
- This approach still using fiction locations which are randomly chosen from already establish cinemas in Paris.
- The rating of top cinemas in Paris are based totally on subjective preference of my friend and I. Performing a survey to rate the top 5 cinemas is a good way to accurately rate and produce a good profile for top cinemas.