Full Report

Battle of the Neighborhoods,

30th March 2019

Introduction

"Would you recommend a location in Hong Kong to open a new cinema?"

My boss, the stakeholder wants to **open a new cinema as company's new business**. He explains that 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 **5 minutes** from **public transport facilities** is perfect.

He wants me concentrated on selection of cinema location according to its nearby environment. Cinema facility and rental price is not my concern. He lists out his **top 10 favorite cinemas** in Hong Kong with rating.

I work with my teammates and select **5 possible locations** to build the cinema. Which location should be suggested to the stakeholder?

Data Section

1. Geographic coordinate of Hong Kong cinemas: I need to compare 5 possible locations with current cinemas **in Hong Kong. Therefore, I need to find a list of Hong Kong cinema and cinemas' geographic coordinates. Luckily, I can find the list and coordinates from the website https://hkmovie6.com/cinema .**

2. Geographic coordinates of 5 possible cinema addresses: Geographic coordinates of 5 possible cinemas are required and I can use Google Map API to find this information

3. Favorite cinema list of stakeholder: The favorite cinema list of stakeholder is an important information that I can **use it as profile to select the best location**.

4. Eating, Shopping and Public transportation facility around cinema: The recommended cinema location needs to have many eating and shopping venues nearby. Convenient public transport is also required.
These data can be found by using FourSquare API to find these venues around the location. The radius of exploration distance is set to 500 meters, which is about 5 minutes walking distance.

Methodology

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why. With above data, I can use content-based recommendation technique to resolve the problem.

Combine with FourSquare API which provides how many venues in different category of Hong Kong cinemas, a matrix which captured characteristic of venues nearby cinema are built. Stakeholder's favorite list is the profile to combine with the matrix to become a weighted matrix of favorite cinema.

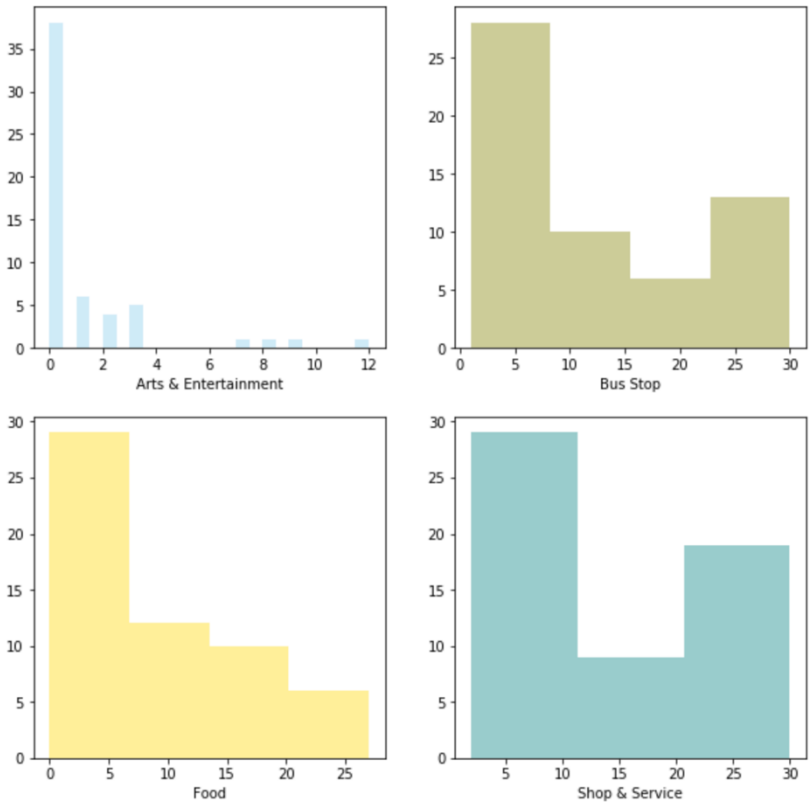
The weighted matrix can be applied on 5 target locations with venues information to generate a ranking result. The the top one on the ranking list can be recommended to the stakeholder.

Before building the matrix, I have to prepare the required data and apply some data analysis.

Data Cleansing and Preparation

Check the cinemas dataset contains any duplicated address: Some "special house" in cinema are separated as a new cinema in www.hkmovie6.com. These records are duplicated in my case and should be corrected. And drop the duplicated cinema records. Cinema '新光戲院大劇場' and '大館' should be considered as cinema in Hong Kong.

Now I can use the FourSquare API to explore nearby venues of Hong Kong cinemas: Cinema really has many 'Bus Stop', 'Food', 'Shop & Service' venues around. However, it is unusual that a cinema has 4 metro stations nearby (within 500 meters). One cinema contains 4 Metro Station around.



The distribution of other variables are quite similar. Now check their **Pearson Correlation:** It seems that 'Bus Stop', 'Shop & Service' and 'Food' category are highly correlated.

Category	Arts & Entertainment	Bus Stop	Food	Metro Station	Shop & Service
Category					
Arts & Entertainment	1.000000	0.494525	0.414387	0.389271	0.506590
Bus Stop	0.494525	1.000000	0.893873	0.563799	0.896388
Food	0.414387	0.893873	1.000000	0.583749	0.872533
Metro Station	0.389271	0.563799	0.583749	1.000000	0.499546
Shop & Service	0.506590	0.896388	0.872533	0.499546	1.000000

Find **P-Value** of the variables

By convention, when the p-value is:

- < 0.001 we say there is strong evidence that the correlation is significant,
- < 0.05; there is moderate evidence that the correlation is significant,
- < 0.1; there is weak evidence that the correlation is significant, and
- is > 0.1; there is no evidence that the correlation is significant

The correlation between 'Bus Stop', 'Food', 'Metro Station' and 'Shop & Service' are statistically significant, and the coefficient of > 0.5 shows that the relationship is positive. Most of Hong Kong cinemas (blue circle) and stakeholder's favorite cinemas (red circle) location are built near main road, and centralized in urban area of Hong Kong. The target locations (yellow circle) of new cinema are not near to main road.

Results

With the boss's profile and the complete list of cinemas and their venues count in hand, I am going to take the weighted average of every lcoation based on the profile and recommend the top location that most satisfy it.

The result is reasonable. Location "L5" has the most number of venues in category "Bus Stop", "Food", "Metro Station" and "Shop & Service".

Location "L5" should be recommeded to the stakeholde

Discussion¶

Number of venues of 5 target locations are actually below the average. I should contact local commercial property agents to find more suitable locations. Moreover, FourSquare is not popular in Hong Kong, the data maybe out-dated or unreliable, the report should gather more data from other location data source such as Google Place API.

The stakeholder's problem is resolved. Stakeholder wants to find the best place to build a new cinema in Hong Kong, and the factors of "best location" is based on the number of venues in eating, shopping, transportation category around the location. Stakeholder also provide his favorite list of cinema to further explain what the "best location" is. Content-based filtering machine learning technique is the most suitable method to resolve the problem. It combines stakeholder's preference and cinema profile to make the recommendation result.

The 5 target locations of new cinema may not be a good choices. As the weighting matrix is developed, I can quickly pick other locations and make the recommendation again.