1.Introduction

1.1 Background

"Is it recommended a location in Hong Kong to open a new cinema?" The stakeholder wants to open a new cinema as company's new business. Watching movie is a part of entertainment. Cinema is better to have many restaurants and shopping places nearby. Transportation is so important that customer can walk to cinema within 5 minutes from public transport facilities. My selection of cinema location should be based on its nearby environment. Cinema facilities and rental price are not my concern. He lists out his top 10 favorite cinemas in Hong Kong with rating. I select 3 possible locations to build the cinema.

1.2 Problem

Data that might contribute to determining the profit of company that might include the size, rental cost of cinema, the number of customers where the location of cinema is. This project aims to predict where a new cinema will be located in the next year based on these data.

1.3 Interest

Obviously, the stakeholder would be very interested in accurate prediction of popular location of cinema, for competitive advantage and business values. Others investor such as property agents and private company may also be interested.

2. Data acquisition and cleaning

2.1 Data sources

Geographic coordinates of Hong Kong cinema and Eating, Shopping and Public transportation facility around cinema can be found on google map and four FourSquare API. The popularity of cinema, however, is hardly determined. To add the criteria of popularity, I ask the list of favourite cinema from stakeholder and calculate popularity in terms of Food, Shop & Service, Bus Stop, Metro Station, Nightlife Spot, Arts & Entertainment.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of many records, because there are many places and cinemas in Hong Kong. I decided to only use the top ten popular venues, because of less popular places were not worthy to consider. There is a problem with the datasets. A place is identified by their coordinates. However, there were many places with the coordinates, which cause their data to mix with each other's. Though it was possible to separate some of them based on further searching on the map, I decided that it was not worth the large effort to do so, because such places only accounted for ~1% of the data.

2.3 Feature selection

After data cleaning, there were many samples and 6 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, there was a feature of cinema facilities, and another feature of rental price he collected. These two features are irrelevant. These features are problematic for two reasons: (1) The popularity factors were duplicated in two features. (2) Rental price was duplicated in multiple features. (Table 1) In order to fix this, I decided not to consider these features. After discarding redundant features, I inspected the correlation of independent variables, and found several pairs that were highly correlated (Pearson correlation coefficient > 0.8). For example, Bus Stop, Food and Metro Station were highly correlated. This makes sense, after all, can save time to go to more places for entertainment. From these highly correlated features, I decide to compare the weighted average of popularity from the samples.

Table 1. Simple feature selection during data cleaning.

Kept features	Dropped features	Reason for dropping features
Arts & Entertainment	Rental Price	Rental price is duplicated in multiple features.
Bus Stop	Cinema Facilities	The popularity factors were duplicated in with Food or Shop & Service.
Food		
Metro Station		
Nightlife Spot		
Shop & Service		

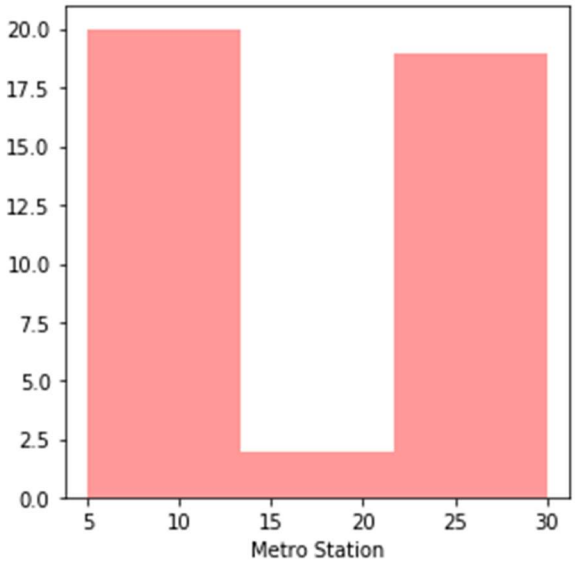
3. Exploratory Data Analysis

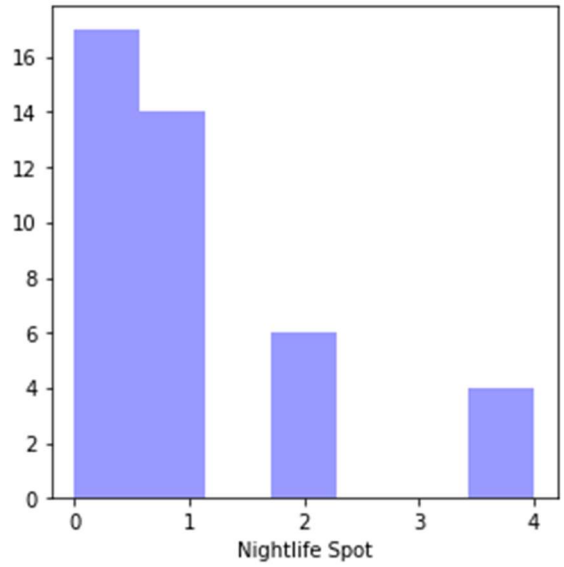
3.1 Calculation of target variable

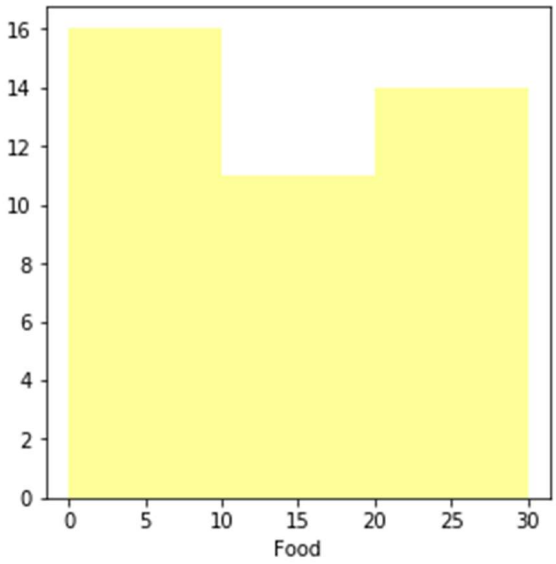
There were no features in the dataset, and had to be created and calculated. I chose to use Food, Shop & Service, Bus Stop, Metro Station, Nightlife Spot, Arts & Entertainment as popularity features. Three big areas in Hong Kong are picked in measurement. To verify if this calculation is consistent with people favourite, I also picked the top three cinemas to compare and found that cinemas in L1 are also the most popular among almost all features. This suggested that L1 is the most recommended place.

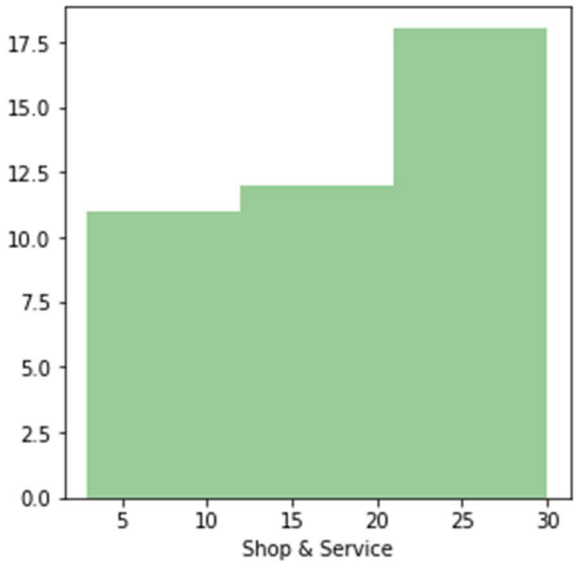
	Location	Name	Latitude	Longitude
0	L1	Kowloon	22.318567	114.179606
1	L2	Hong Kong Island	22.258759	114.191070
2	L3	New Terrtories	22.370424	114.123415

Category	Arts & Entertainment	Bus Stop	Food	Metro Station	Nightlife Spot	Shop & Service
count	41.000000	41.000000	41.000000	41.000000	41.000000	41.000000
mean	4.390244	21.780488	14.097561	18.634146	1.024390	19.853659
std	3.967859	9.188341	10.049390	9.979369	1.214245	9.551338
min	0.000000	5.000000	0.000000	5.000000	0.000000	3.000000
25%	1.000000	12.000000	5.000000	10.000000	0.000000	11.000000
50%	4.000000	30.000000	11.000000	17.000000	1.000000	18.000000
75%	7.000000	30.000000	24.000000	30.000000	1.000000	30.000000
max	14.000000	30.000000	30.000000	30.000000	4.000000	30.000000









Category	Arts & Entertainment	Bus Stop	Food	Metro Station	Nightlife Spot	Shop & Service
Cinema Name						
AMC Pacific Place Hong Kong	13.0	30.0	16.0	30.0	1.0	30.0
Broadway	11.0	30.0	30.0	30.0	0.0	30.0
Broadway Cinema	1.0	12.0	8.0	17.0	0.0	15.0

	Category	Arts & Entertainment	Bus Stop	Food	Metro Station	Nightlife Spot	Shop & Service
0	Arts & Entertainment	strong	strong	strong	strong	no	strong
1	Bus Stop	strong	strong	strong	strong	moderate	strong
2	Food	strong	strong	strong	strong	strong	strong
3	Metro Station	strong	strong	strong	strong	strong	strong
4	Nightlife Spot	no	moderate	strong	strong	strong	moderate
5	Shop & Service	strong	strong	strong	strong	moderate	strong

Category Location	Bus Stop	Food	Metro Station	Nightlife Spot	Shop & Service
L1	1.000000	1.0	0.909091	1.0	0.7
L2	0.000000	0.0	0.000000	0.0	0.0
L3	0.090909	0.0	1.000000	0.0	1.0

In [ ]:

4. Result

Every statistics show that L1 is leading almost all features, it is no need to make further calculations or comparisons and lead to the final decision. L1, Kowloon, is the most recommended place to open a new cinema.

5. Conclusions

In this study, I analyzed the relationship between popularity features and cinema and places. I identified Food, Shop & Service, Bus Stop, Metro Station, Nightlife Spot, Arts & Entertainment among the most important features that affect a cinema location next year. I built Content-Based or Item-

Item recommendation systems figure out people's favourite new cinema location by counting number of nearby venues and ratings given. These models can be very useful in helping company's decision on opening a cinema. For example, it could help identify which place to open, estimate the size of cinema, time plan or buying plan of films, etc.

6. Future directions

I was able to find a recommended place easily. However, there was still significant variance that could not be figured by the models in this study. I think the models could use more improvements on more specific place. For example, Mong Kok and Tsim Sha Tsui are the places in L1, but they cannot be compared in this project. The popular features of these two small places in L1 might be different. More data, especially data of different types, would help improve model performances significantly.