CASA0013: Foundations of Spatial Data Science

Description, Price and Local difference

-- A Study on the Key Terms for Marketing Airbnb Lists in Different Tourism Areas

Student ID：20088234

Word Count: 1605

1.Executive summary

For audience of potential Airbnb hosts and investors, the research hopes to investigate, in typical tourism hotspot areas of London, which key terms are more frequently used by hosts to market their listings. The research can be divided into three sub-questions: first, how to identify the different tourism hotspot areas in London; second, how to detect and analyse the key terms used in marketing listings; third, what is the potential relation between key terms and the listing price.

For the first question, the Airbnb listings data and POI data are collected to reflect the facility distribution related to tourism activities. Through a series of data cleaning and dimensionality reduction steps, K-means clustering analysis is applied to distinguish the tourism hotspot areas with different tourism characteristics. For the second question, listings within three typical areas in central London are selected to detect their high-frequency terms in the listing description. NLP methods are applied to extract the 2-grams terms and adjective words from description text and calculate their TF-IDF scores. It is found that similar amenity and accessibility-related terms are used in listing descriptions in different tourism areas, while there are significant differences in place names mentioned. Further regression analysis reveals that listings with higher price tend to market occupancy experience with descriptive words while listings with lower price focus more on the amenities and accessibility.

2.Background

As the host's non-formulaic introduction to the property, listing description can be a core embodiment of hosts' marketing strategy. However, for new hosts with little experience, it can be difficult to decide which content to add in their description text. Airbnb and unofficial communites such as AirHost Academy suggest hosts to show unique experience rather than physical space in description (Airbnb,2020; Airhostacademy, 2019). A research in San Francisco also find that hotel-like properties in description may dissuades potential guests (Janssens, Bogaert, and Van den Poel, 2021) The policy and findings above suggest that some unique terms in description may bring advantages to listings, in terms of sales or price. ThoughHowever, there could be regional and local difference in the specific terms used. Taking typical tourism hotspot areas of London as examples, the reseach hope to Identify high-frequency terms and features in the description and analyse their relationship with listing price. The aim is to understand which expressions may become advantages in local Airbnb market of London and provide suggestions for new hosts and investors.

3.Data Analysis

3.1 Data Description

The tourism areas with different characteristics are identified based on the features of Airbnb listing and tourism-related facilities(Table1). The listing data comes from the London 2021-10-10 archive, Inside Airbnb website. Considering the price differences of listings in different room types, the 'Entire home/apt' is selected as the representative. Median listing price, median reviews per month and listing density is summarised in each middle super output area (MSOA) . Tourism-related data comes from OpenStreetMap (OSM), which can be further divided into two themes, including art-culture facilities such as museums, art galleries and theatres and drink-food facilities such as restaurants, cafes and bars, with six items each. Counts rather than density of facilities are summarised for each MSOA, because facilities such as museums may have an impact beyond the local area.

Table1. Preliminary selection of Attributes

Airbnb Listing Attribute	POI Attribute: Food_Drink	POI Attribute: Art_Culture
price	restaurant	memorial
reviews_per_month	cafe	artwork
listing_density	fast_food	monument
	pub	attraction
	bar	museum
	nightclub	theatre

3.2 Dimensionality Reduction

Principal Component Analysis(PCA) is applied to reduce the dimensions of tourism-related features to imrove the clustering accuracy,as in Figure1 and Figure2. One principal component was extracted from six food-drink facilities, accounting for 81.42% of variance, indicating that there is high consistency in the spatial distribution of various service facilities. Three principal components were extracted from art-culture facilities, accounting for 55.15%, 13.93% and 11.45% of variance. Among them, principal component 1 represents the common characteristics of six types of art-culture facilities, principal component 2 highlights the influence of attraction and museum, and principal component 3 highlights the influence of theatre(Table2).

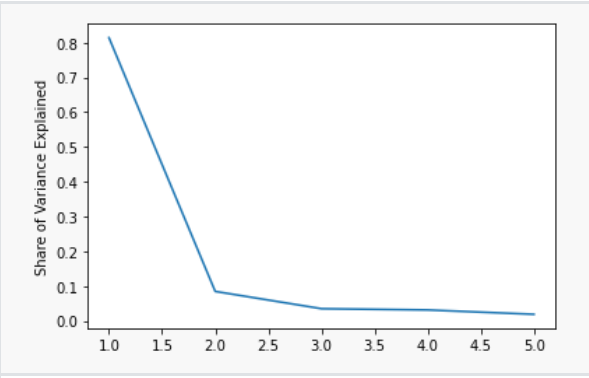

Figure1. Components' share of variance - Food_Drink

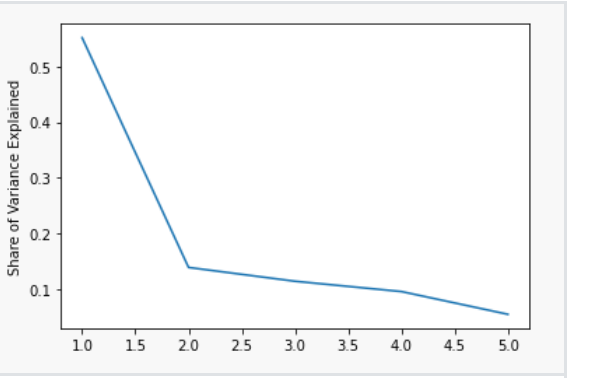

Figure2. Components' share of variance - Art_Culture

Table2. Importance of original attributes in selected componenets

POI_Food_Drink	PC1	POI_Art_Culture	PC1	PC2	PC3
restaurant	0.45	memorial	0.45	0.20	0.21
cafe	0.50	artwork	0.31	0.02	0.23
fast_food	0.42	monument	0.25	0.01	0.26
pub	0.37	attraction	0.43	0.66	0.39
bar	0.41	museum	0.57	0.60	0.18
nightclub	0.25	theatre	0.37	0.40	0.81

3.3 Clustering Analysis

After data cleaning and normalisation, three listing-related factors and four tourism-related factors constitute the factor set participating in clustering analysis(Table3) and K-means method is used. Clustering number of 10 is set for the second-highest silhouette score and the appropriate accuracy relative to the total number of samples.

Table3. Factors participating in clustering analysis

Factors
median_price
median_reviews_per_month
listing_density
fooddrink1
artculture1
artculture2
artculture3

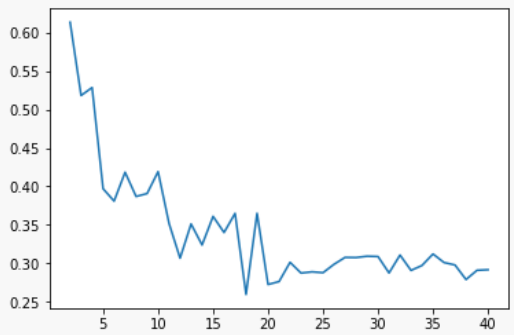

Figure3. Average Silhouette Scores

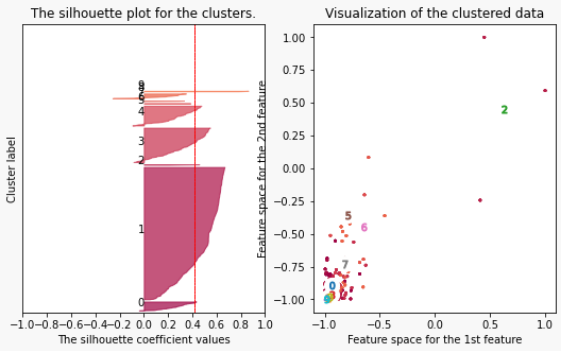

Figure4. Silhouette Results for KMeans clustering with 10 clusters

The count and spatial distribution of different clusters are presented in the Figure3 and Figure4. In general, the central London area has more clustering categories and fewer samples, while the constitution of clusters in suburban areas is relatively single. Further comparison of representative centroids shows that clusters2, clusters5 and clusters6 have outstanding characteristics. Cluster2 includes the main areas of the City and Westminster, which can be regarded as the most important tourist hot spots in London. The standard values of fooddrink1, artculture1 and listing density are the highest for the cluster 2 listings. The standard value of density is the highest. Clusters5 and clusters6 represent sub hot spots for tourism activities, whose main parts are located on the north and south bank of the River Thames, respectively. There are similar properties in cluster5 and cluster6, including the highest standard value of artcultre2, and relatively higher fooddrink1 and artculture1. The difference is that cluster5 is significantly higher than cluster6 in artculture 3, highlighting the characteristics of theatre count.

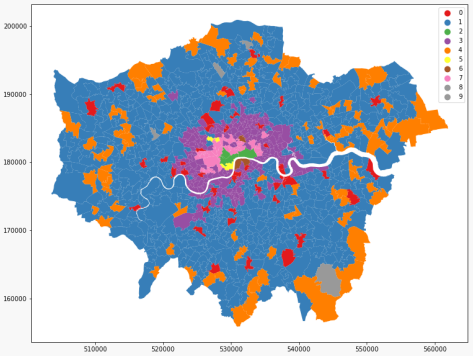

Figure5. Spatial Distribution of 10 Clusters

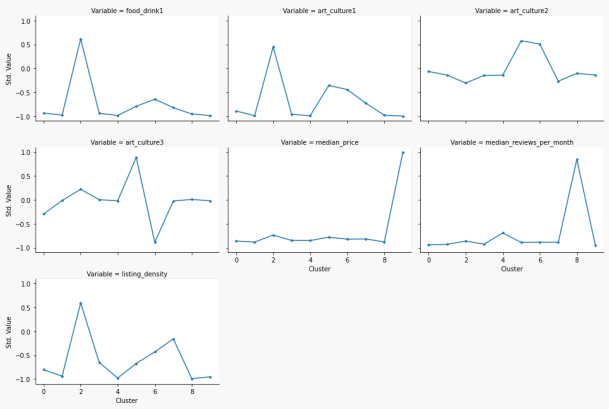

Figure6. representative centroids of Clusters

3.4 Key Term Extraction

Regroup MSOAs belonging to cluster2, cluster5 and cluster6 and adjacent to the canal as group1, group2 and group3. Lists in the corresponding groups are selected as the research object to analyse the characteristics of terms in the list description(Table4). Firstly, 2-grams terms is extracted as the main object of text processing. 2-gram terms maintain more specific information than 1-gram terms. At the same time, compared with terms with 3 or more grams, 2-gram terms have stronger convergence of high-frequency terms. On this basis, calculate TF-IDF scores of 2-

grams terms in each group, and generate word clouds according to TF-IDF scores, as shown in the Figure7. According to the word cloud, terms such as 'minute walk' and 'living room' appear more frequently in the three groups of description texts, indicating that the host prefers to publicise the advantages of listing in room, amenities and accessibility. Different place names also appear with high frequency in different groups, such as 'covent garden' in group1, 'london eye' in group2.

Table4. Selected MSOAs and corresponding listings for key term extraction

New Group	Original Cluster	MSOA Name	Listing Number
Group 1	Cluster 2	'Westminster 013', 'City of London 001', 'Westminster 018'	1437
Group 2	Cluster 6	'Lambeth 036', 'Southwark 002'	444
Group 3	Cluster 5	'Westminster 020'	150

a. Group1 | b. Group2 | c. Group3

Figure7. Word cloud for 2-gram terms from description text

In addition, try using the 'pos_tag' function of the nltk package to extract only adjectives in the text and calculate the TF-IDF score to generate a word cloud. As shown in the Figure8, 'modern' and 'large' appear more frequently in group1 and group2, highlighting the description of room style and scale of facilities and space. Emphasis on 'modern' is weakened in group3, and the high-frequency words are 'available' and 'free'.

a. Group1 | b. Group2 | c. Group3

Figure8. Word cloud for adjective term from description text

3.5 Key Terms Analysis

OLS regression analysis is applied to further understand the association between key terms and listing price in each group. The independent variables include eight non-spatial 2-gram terms and eight adjective terms that are common in all three groups and have the highest TF-IDF scores and four spatial terms in each group(Table5). Terms are converted into binary variables according to TF-IDF score equal to or greater than 0. Listing price is set as the dependent variable. The model information is summarised in Table6. Among the three models, group2 model has better robustness, with adj R2 = 0.607. The coefficients of 'minute walk', 'modern' and 'large' in the model are positive, indicating that the existence of the above term may be associated with higher rent prices. In contrast, terms such as 'living room', 'central London' and 'double bed' are associated with lower rental prices. The fitness of group3 model is weak, with adj R2 = 0.432. However, similarly, 'comfort' and 'free' are associated with higher rental prices. The adj R2 of group1 model is equal to 0.039, where relevant terms are not explanatory of the lease price.

Table5. Terms included in independent variables in regression model

Commn Terms-Adj	Commn Terms-Non Place	Group1-Place	Group2-Place	Group3-Place
modern, large, private, spacious, comfortable, great, free, high	living room, central london, minute walk, fully equip, walk distance, double bed, sofa bed, heart london	covent garden, oxford street, leicester square, piccadilly circus	london eye, london bridge, borough market, tate modern	james park, victoria station, westminster abbey, big ben

Table6. Model fitness for different groups

	Group1	Group2	Group3
R-squared:	0.052	0.625	0.500
Adj. R-squared:	0.039	0.607	0.432
Terms with coef >0	--	minute walk, modern, large, private, comfortable, free	comfortable, free
Terms with coef <0	--	london eye, tate modern, central london, double bed, sofa bed, heart london, great	walk distance, large，great

4.Conclusion

According to the analysis, it is found that:

Firstly, hosts may prefer to publicise the advantages of listing in room amenities and accessibility, and bind listing with nearby scenic spots or landmarks as a marketing strategy. Non-spatial high-frequency terms are very similar in each group and there are significant differences in high-frequency place names.

Secondly, the different frequency of adjectives may imply the general style and form differences of Airbnb listing in each group, and even the deep-seated social background differences. For example, the description in group 2 emphasises more 'modern' and 'private ' than that in group 3. In reality, the south bank area corresponding to group 2 has more new apartments due to the urban renewal, while the buildings in Westminster area corresponding to group 3 may have a longer history.

Finally, from the regression analysis, it can be found that terms related to higher rental prices, such as 'private', 'modern' and 'free', are some more abstract words around experience, while words related to lower prices are mostly specific facilities or direct descriptions of them. Words related to place names are also related to lower prices. This implies that listings with lower price may prefer to take amenity perfection and accessibility as the selling point and highlight the economy of listing. In contrast, listings with higher prices emphasise the occupancy experience for tenants. As what is found by Janssens, Bogaert, and Van den Poel (2021), guests are more likely to choose Airbnb for their unique nature of accommodation. Though listings price may fluctuate frequently due to the market condition, the findings above support that a good focus on ternants' potential experience in description may bring more space for hosts in pricing.

Limitations of this study may include: there is no reliable regression model for group 1 for further comparision; analysis about the relation between airbnb sales and the key terms is missing.

5.References

Airbnb (2020) *Sprucing up your listing description*[Online] Available at: https://www.airbnb.co.uk/resources/hosting-homes/a/sprucing-up-your-listing-description-13 (Accessed: 12 January2022).

Janssens, B., Bogaert, M. and Van den Poel, D. (2021) 'Evaluating the influence of Airbnb listings' descriptions on demand', *International Journal of Hospitality Management*, 99, p. 103071. doi:10.1016/j.ijhm.2021.103071.

AirHostAcademy (2019) *Airbnb Listing Description Samples*[Online] Available at: https://airhostacademy.com/airbnb-listing-description-samples/ (Accessed: 12 January2022).