Stats401\_Porject 1

Ruitian Wu; Sissi Wang; Yuzhe Gu

Study on Characteristics Influencing Airbnb Rental Popularity in Shanghai

**Introduction**

Homestay has become a popular choice for vacation rentals instead of hotels in China. Our group is trying to figure out what factors could influence the popularity of the homestays, especially in metropolitans. To answer the main question, we will focus on the data from Airbnb in Shanghai. The mind map of our project is shown in Fig.1.

图形用户界面, 文本

描述已自动生成

**Fig.1** Mind map

**Method**

*Data collection*

As shown in Fig.2 below, the targeted homestay information is within table frames, so we adopted a data-spider software Octoparse to handle it. The scraper follows the crawling logic shown in Fig.2’s right part. As a result, we crawled homestay data with prices from 50¥ to 1500¥ in both Shanghai and Xiamen, including attributes such as “Title”, “Facility”, “Label” etc. We retrieved 4557 and 7271 pieces of raw data items on homestays in Shanghai and Xiamen, respectively.

图片包含 图形用户界面

描述已自动生成图示

描述已自动生成

**Fig.2** Airbnb Website Overview

*Data cleaning*

We cleaned the dataset of Shanghai following the steps shown in Fig.3. The dataset of Xiamen is cleaned following the same steps.

图示

描述已自动生成

**Fig.3** Steps for data cleaning

*Analyze and visualize*

We mainly want to solve five sub-questions to portray the characteristics of homestays that have higher popularity.

For popularity, we care about two things: Firstly, how many people have chosen the homestay before. Unfortunately, we cannot get “number of orders” and turn to “number of comments” instead, which could indirectly reflect the number of orders. Also, to quantify its potential popularity, we use the average grades given by previous customers since they highly influence future customers. Finally, we multiply the log of the number of comments by grades. We use log since some homestays receive fewer comments as they are new. We don’t want time to have too much impact.

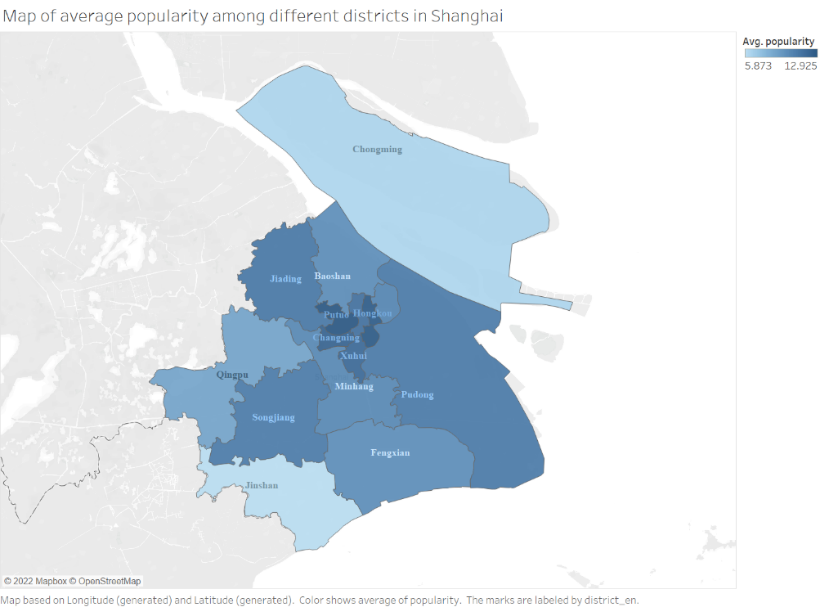
Then let’s introduce the sub-questions that we are solving:

1. The location could impact the popularity of homestays in Shanghai. For example, urban houses are more likely to become popular than rural ones. We will show them on a Shanghai Map to compare all districts.
2. Price is also an essential factor influencing customers’ decisions. However, figuring out what exact price will become most popular is not practical. Instead, we use a Chi-Merge algorithm [1] to segment price distributions into seven most identical price levels and search for the price level with the highest popularity using a **Violin Plot**.
3. Another critical characteristic of Airbnb homestays is the “tag”, showing the service of homestays. We extracted the ten most frequently observed tags in homestays in our project. We want to know how likely popular homestays tend to have these tags, and we compare the occurrence of these tags between Top-200 popular homestays and less popular ones using a **Star Plot**.
4. The description of a certain homestay is considered by customers as well. There are mainly two attributes: the type of a house and its ownership. We want to see what combinations of the two attributes tend to be more popular, and **Sankey Diagram** is a perfect fit.
5. Finally, we want to see how the names of homestays influence their popularity. Usually, the names of homestays contain attractive keywords, like location or decoration style. We want to see what keywords are dominant for the Top-200 popular homestays using **Word Cloud**. Also, to figure out whether the keywords are unique in Shanghai, we compare the Word Could to Xiamen’s.

**Results**

*Map*

Fig.4 is a map comparing the average popularity of rentals among different districts in Shanghai. The luminance of the color represents the average popularity. From the Map, we can find that rentals in Putuo, Huangpu, and Hongkou are the most popular, while rentals in Chongming and Jinshan are less popular. We can conclude that rentals in the center of Shanghai are more popular than rentals in remote areas.



**Fig.4** Map of average popularity among different districts in Shanghai

*Violin Plot*

In Fig.5, homestays in Shanghai are segmented into ordinal variables of 7 levels, ranging from extremely cheap to extremely expensive. The x-axis represents that price level, while the y-axis represents popularity. The width of each violin represents the density distribution of popularity. The luminance of the color represents the price level. Also, we highlighted three lines: the 25% quarter line, median line, and 75% quarter lines. The visualization shows that the overall popularity would be higher for relatively cheap houses or medium prices. As price increases or decreases, the popularity will fall. However, some extremely popular houses are observed in expensive homestays. The “social-media influencer” culture possibly causes the trend in Shanghai.

图表, 图示

描述已自动生成

**Fig.4** Violin Plot of popularity in different price levels

*Star Plot*

Fig.6 is a star plot comparing the commonly observed tags of the TOP-200 popular rentals and those of the rest. It contains ten attributes and two sets of observations. Each attribute represents the percentage of rentals containing the corresponding tag. The tags are the ten most observed tags in the whole dataset. From the Star Plot, we can find that the top-200 rentals have a significantly higher percentage of 自主入住 (check-in by yourself), 超赞房东 (brilliant landlord). It may suggest that customers emphasize the convenience of check-in and the friendliness of the landlord.

*Sankey Diagram*

In Fig.7, the left bar represents common types of homestays. The right bar represents three types of ownership. The width of each bin represents the total popularity of that combination. For the color, the luminance represents different categories in an attribute. We can see that “A whole LOFT” is the most popular in the figure. This sounds reasonable in terms of privacy, aesthetics, and convenience. Also, this is possibly influenced by “social-media influencer” culture since they are suitable places to enjoy city life and take photos.

图表, 雷达图

描述已自动生成

**Fig.6** Star Plot of frequency of 10 most popular tags among the top-200 rentals and the rest

图示

低可信度描述已自动生成

**Fig.7** Sankey Chart between the type of house and ownership

*Word Cloud*

Fig.8 shows the word-cloud comparison of the TOP-200 popular homestay titles in Shanghai and Xiamen, the typical metropolitan modern city and the seaside tourist city. It could be found that popular homestays in both cities will display famous scenic spots in their titles. Also, they both contain keywords suggesting that the homestay location is close to metros or business districts, which is convenient for travelers. Transportation and business districts keywords appear most frequently for homestays in Shanghai. Instead, keywords about tourist attachments show a higher frequency in Xiamen compared to Shanghai. This is also reasonable as it shows the different local colors between modern cities and tourist cities. Tourists living in modern cities care more about the convenience of dining, shopping, and traveling rather than staying in a homestay with good views.

文本, 日程表

描述已自动生成

**Fig.8** Word-Cloud Comparison (SH&XM)

**Discussion**

*Map*

The Map accomplished user tasks: **retrieve the value** of average popularity and **correlate** popularity with the location. For efficiency, the position and color are clear to distinguish, and with the names of the districts, the visualization shows the distribution efficiently.

*Violin Plot*

The Violin Plot accomplished the below tasks: **determine the range** of popularity, **compute derived** **value** like Median, **characterize popularity distributions** in each price level, and **correlate** popularity with the price. For efficiency, the idiom of the violin plot is appropriate for making clear comparisons. Also, the differences in luminance and position clearly illustrated the pattern. However, one possible improvement is highlighting the outlier points to emphasize extremums.

*Star Plot*

The Star Plot accomplished user tasks: **retrieve the value** of frequency of tags and **correlate** popularity with tags. For efficiency, the star plot compares the Top-200 and the rest and the comparison among frequencies of 10 tags efficiently.

*Sankey Diagram*

For the Sankey Plot, it accomplishes the below tasks: **retrieve the value** of popularity, **find the extremum** of popularity, **sort** the attributes and **correlate** popularity with house types and ownerships. For efficiency, the Sankey Diagram idiom is suitable for showing patterns in different attributes combinations. With distinctions in bins’ width and explicit tags, the visualization is also accessible to retrieve information. However, the bar on the two sides could be sorted, and each bin could be attached with values.

*Word Cloud*

The word cloud accomplishes user targets: **compute the derived value** of keyword frequency, **sort** the keywords and **correlate** popularity with hot words in the title. For efficiency, font size and saturation channels allow users to easily catch the most frequent keywords.

**Conclusion**

With the data analysis and visualization, we can give a portrait of popular homestays in metropolitan cities like Shanghai. They usually have downtown locations, moderate prices, people-oriented service, and will suggest convenient transportation and close business districts in titles. Besides, given the prevalent “social-media influencer” culture in shanghai, high-end homestay types like loft houses are also favored by customers.

Reference

[1] Randy Kerber. 1992. ChiMerge: discretization of numeric attributes. In Proceedings of the tenth national

conference on Artificial intelligence (AAAI’ 92). AAAI Press, 123-128