Data Exploration of Singapore

Name: yifan yang

INTRODUCTION

This data exploration report focuses on the Airbnb in Singapore. Airbnb is an online marketplace for arranging primarily homestays and travel accommodation. Airbnb becomes a famous way for tourist. I also book house in Airbnb multiple times before, so I am curious about different rooms in Airbnb. The question is "which area is suitable for me to rent if I am a traveler?".

DATA WRANGLING

Foursquare location data is used to show various venues around Singapore. It is a location based app which help user to find the best restaurant or shops in their location. I will use the following request to explore venues https://api.fouraquare.com/v2/venues/explore/....

The result is in the format 'json':

Field	description
Id	A unique identification
name	Name of venue
location	Latitude and longitude
categories	Categories applies to this venue

For the location of Singapore, I will use Geopy. It can return the latitude and longitude.

Next is the data of hotel. The data is collected by Airbnb on 28 August 2019 which list briefly information host house on Airbnb in Singapore (https://www.kaggle.com/jojoker/singapore-airbnb). It is tabular data with 7908 rows * 16 columns. Below is the meaning of each columns.

id - Listing ID

name - Listing Title

host_id - ID of Host

host_name - Name of Host

neighbourhood_group - Borough that contains listing

neighbourhood - Name of neighbourhood that listing is in

latitude - latitude of listing

longitude - longitude of listing

room_type - Type of public space that is being offered

price - price per night, USD

minimum_nights - minimum number of nights required to book listing

number_of_reviews - total number of reviews that listing has accumulated

last_review - date in which listing was last rented

reviews_per_month - total number of reviews divided by the number of months the listing is active

calculated_host_listings_count - amount of listing per host**

availability_365 - number of days per year the listing is active

I read the csv file into python to do data wrangling step. I think host_name is insignificant in our exploration and it is also personal privacy. The columns last_review and calculated_host_listings_count is also irrelevant to this data analysis. As a result, I drop columns ‘host_name, last_review, calculated_host_listings_count’ in this table. Then I check the missing value and result shows below.

```
id                              0
name                            2
host_id                         0
neighbourhood_group             0
neighbourhood                   0
latitude                        0
longitude                       0
room_type                       0
price                           0
minimum_nights                  0
number_of_reviews               0
reviews_per_month            2758
availability_365                0
dtype: int64
```

fig.1

There are 2 null value in ‘name’ column and I think this column is subjective which write by host and I don’t need change it. For ‘reviews_per_month’ column, I can simply append it with 0 for missing value because I assume that the house without reviews data have 0 reviews per month. I save the new data into csv file.

DATA CHECKING

After data wrangling, I read new csv file to check the data. Firstly I plot every location in the map and get fig.2 below. It shows that every location is in Singapore and proves no errors in location.

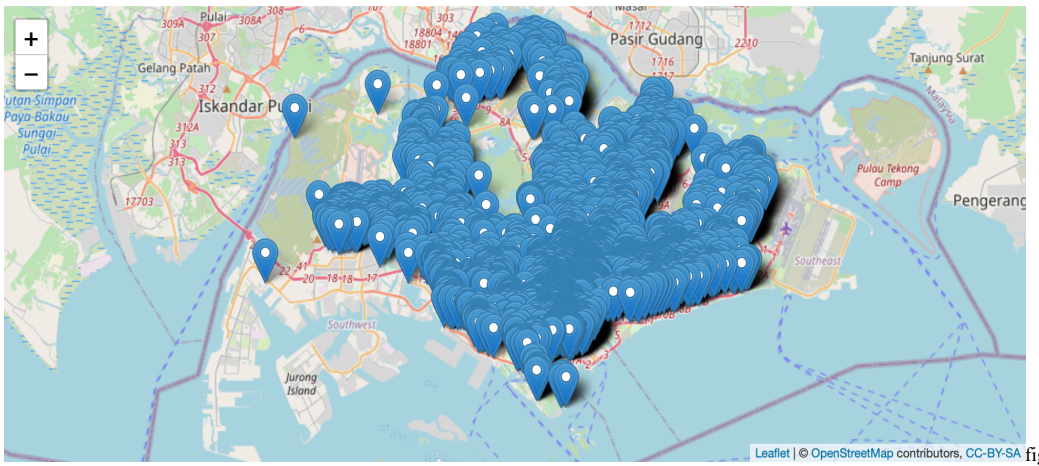


fig.2

Then I plot the graph shows the price distribution group by regions. There are 5 regions in this data and the price range from 0 to 10000. More than half houses are sited in central region. It might exist some outliers which will affect data exploration, so I need to concern about this situation in visualization.

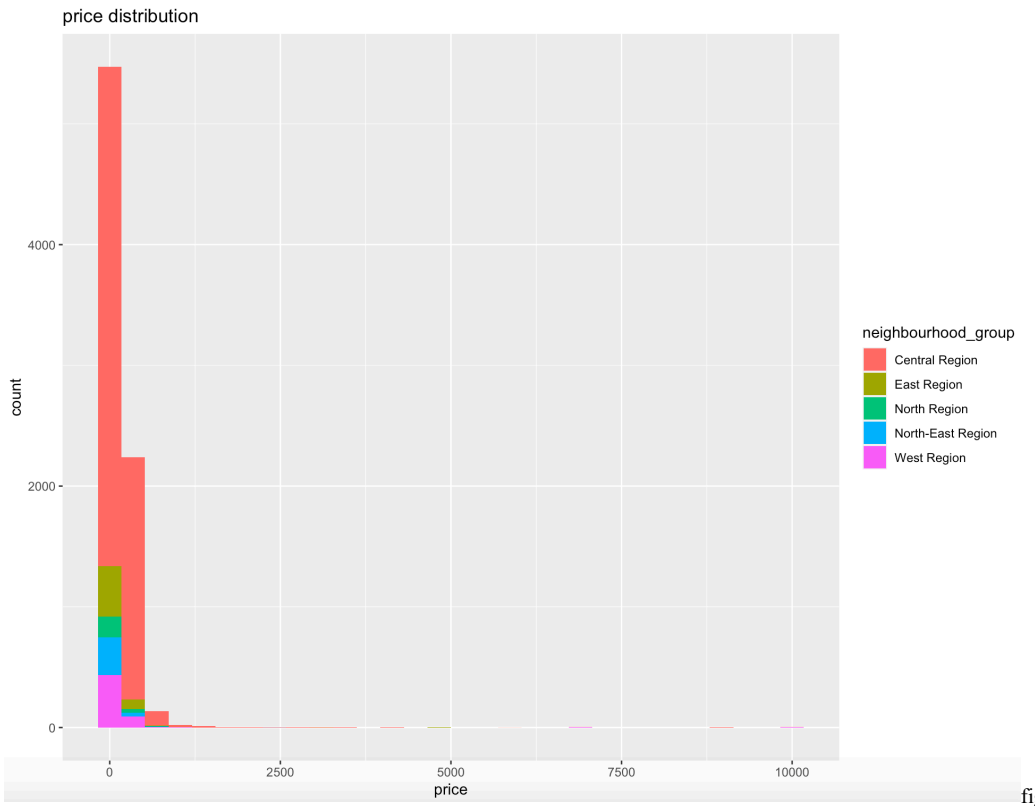


fig.3

Regard to other columns, I choose to check their static status as shown in figure 4.

There are 3 room types and all availability_365 value is less than 365. There is no abnormal value in these columns.

```
          room_type    availability_365 number_of_reviews reviews_per_month
Entire home/apt:4132   Min.   :  0.0    Min.   :  0.00    Min.   : 0.0000
Private room   :3381   1st Qu.: 54.0    1st Qu.:  0.00    1st Qu.: 0.0000
Shared room    : 394   Median :260.0    Median :  2.00    Median : 0.1600
                       Mean   :208.7    Mean   : 12.81    Mean   : 0.6796
                       3rd Qu.:355.0    3rd Qu.: 10.00    3rd Qu.: 0.8500
                       Max.   :365.0    Max.   :323.00    Max.   :13.0000
```

fig.4

DATA EXPLORATION

I will use tableau and R to do data exploration.

- **Price and location**

All rooms in data belongs to 5 different regions. As a result, I will choose to explore the difference between the price of rooms in these regions. Firstly, I put the table into tableau and plot the map below. It is obvious that house in central regions have higher price while the price of house in north region and north-east region is lower compared to others.

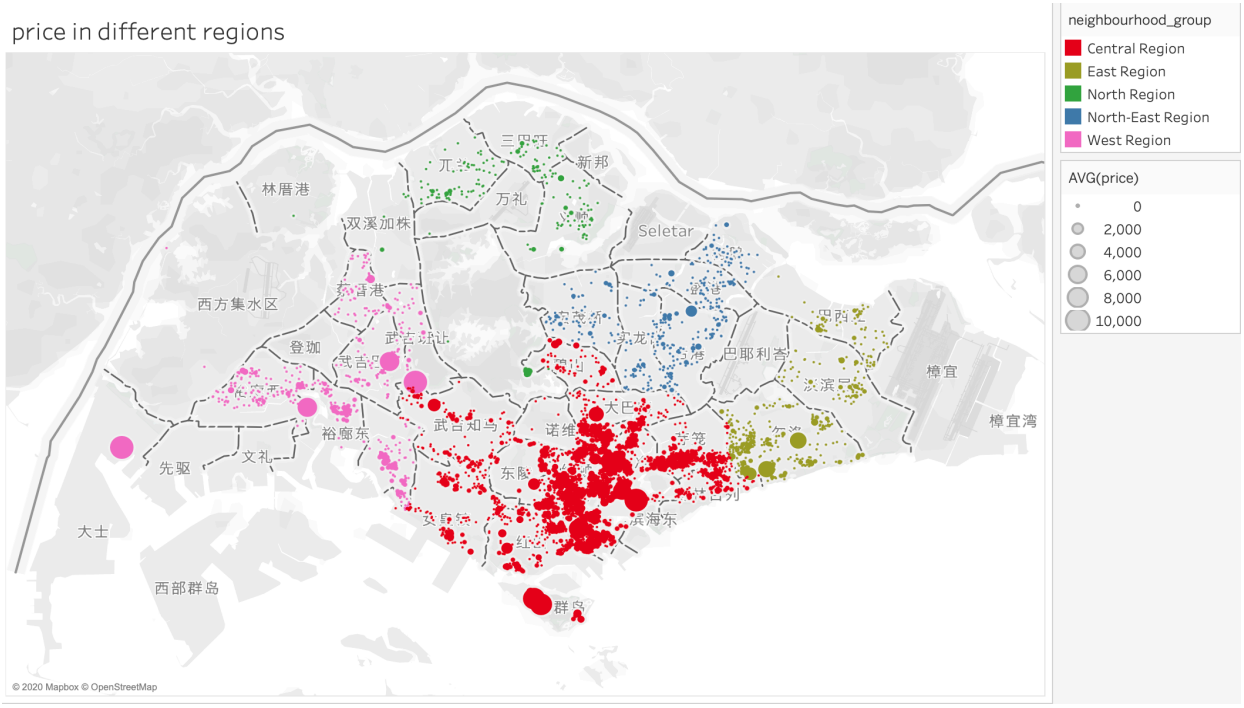


fig.5

To be more clearly to see the actual price mathematical comparation, I also plot the violin plot below (y axes limit to $450 due to large outliers). We can see that the average house price in central region is highest and it has an even distribution. While

in other regions, the average price is lower than $100 and most of house in them are concentrate on the price range $50 to $100.

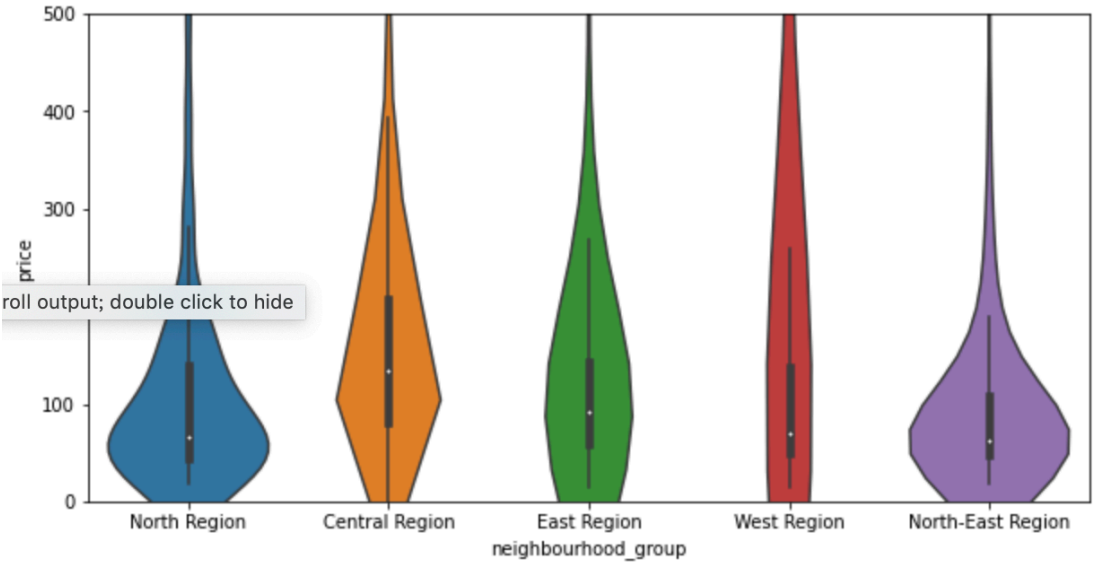


fig.6

Based on two graphs above, it is found that the price of house in central regions is much higher than in other regions. I believe the most important part is that housing price in central regions is too high. Core central regions has the highest caliber and highest priced properties (Popescu, 2019). For example, a two-bedroom, 1,000-square-foot apartment might cost $2.2 million to $2.5 million while it is cost $725,000 to $1.1 million in Outside Central Region. The higher cost of host leads to higher rent price in order to get higher profit. In conclusion, the location of room is closer to business center which has higher housing price, the higher price it has.

Then I compared room type between popular houses and normal houses. The difference between them is small that entire home has larger proportion in popular house than normal. According to figure.8, it is found that customers are more likely to book entire home and private room. The entire home seems to be the first choice when you put your house in airbnb website and try to attract more customers.

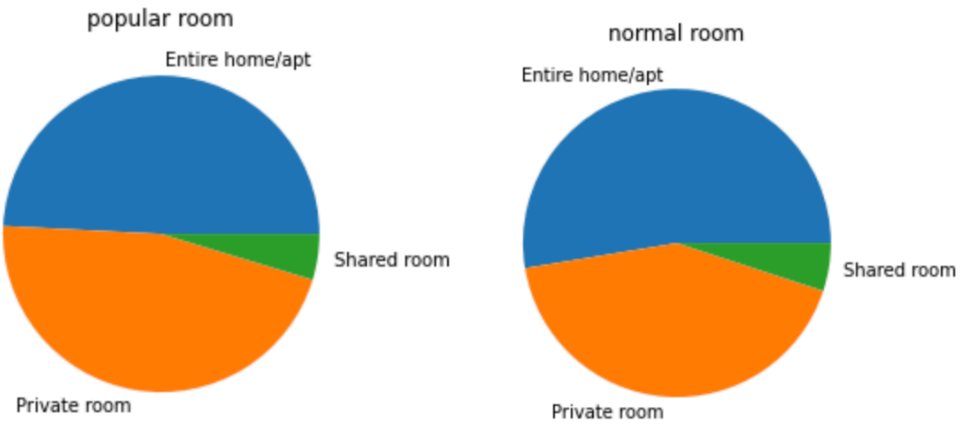


fig.8

After that I want to explore whether location affect the popular rate of house. Pie chart of location distribution of different group is plotted below. Central region ranks 1st in both groups and it is little higher in popular group. East region follows which is 9% in popular group. It shows that central region is still top choice for people in airbnb and most popular house site in here.

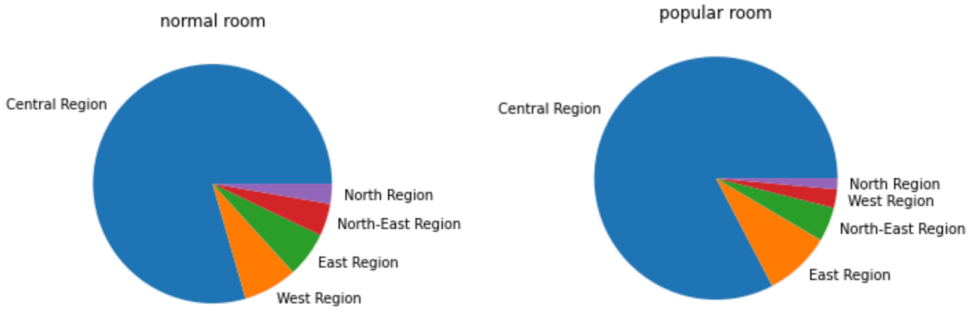
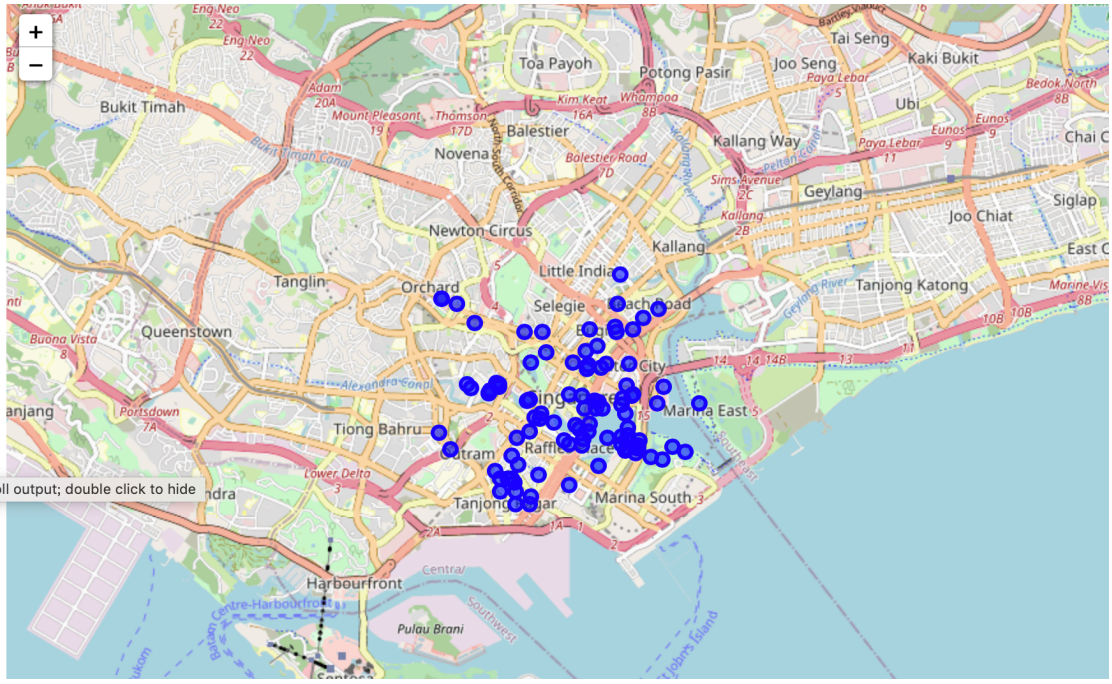


fig.9

Then I check the distribution of venues in Singapore, it shows that most venues are in central region and categories of venues are various including waterfront, theater restaurant and book store. I think it is a huge advantage for traveler if they want to experience more enteriment.

	categories
Hotel	13
Waterfront	5
Japanese Restaurant	3
Park	3
Event Space	3
Movie Theater	2
Bookstore	2
Italian Restaurant	2
Café	2
Concert Hall	2



CONCLUSION

From the data exploration above, I find that central region in Singapore has huge advantages compared to other regions. It has higher housing price and more Airbnb sources. Violin plot and map of price distributions are used to show the higher price in central region which is caused by high housing price and economy central. 4 aspects are used to explain the features of popular house in Airbnb. Among each parts, violin plot, pie chart and word cloud are plotted by RStudio. It is believed that most customers are willing to book entire home in central region with relatively lower price. Comfortable environment and convenient transportation are also attractive features. The central region has more venues than others which is a huge advantage for travel who wants to go to Singapore.

REFERENCE

Popescu, R. (2019). House Hunting in … Singapore. Retrieved 25 April 2020, from https://www.nytimes.com/2019/10/16/realestate/house-hunting-in-singapore.html

Singapore Airbnb (2019), from https://www.kaggle.com/jojoker/singapore-airbnb