CSL Data Analytics Case Competition

Whitepaper

Theme:
Draw data insights on the **strengths of current online shopping platforms** and **propose strategies** that can **further enhance customer online shopping experience**

Focused areas: Real Estate, Travel and Tourism

Team name: Dataholic
Team member: Tony Tang | Justin Chiu | Chanho Park | Jacob Chiu

1. Executive Summary

Hong Kong has been one of the hottest markets for Airbnb, recording over 12,000 listings as of October 2019. The Airbnb dataset provides us with a fantastic source to better understand Hong Kong's bustling rental landscape. As of today, there are over 300 places currently listed on Airbnb and over 5000 hosts. Yet, unlicensed rentals on platforms such as Airbnb are still deemed illegal by the Home Affairs Department in Hong Kong. One can perhaps attribute the success of Airbnb in Hong Kong to the high rates charged by the hotels, which are primarily driven by the exorbitant rental prices in the city.

Different approaches of data analytics will be conducted to explore the prices, demand, types of properties, locations and customer sentiments regarding the Airbnb market in Hong Kong. This report will be looking at historical data as well as current data on the Airbnb website. Finally, four recommendations based on the findings and conclusions of our analysis will be provided to Airbnb on how to improve the customer experience of its online platform. These suggestions could also be applied in other cities apart from Hong Kong, which will be able to aid its business on a global scale.

Main analysis done:

1. Demand and Supply: Airbnb Customer Growth vs Listing Prices over time
2. How price and customers sentiment vary with the locations of listings
3. Analyzing Customer Reviews to know understand what customers thinking about
4. Analyzing what it takes to become a superhost
5. Price forecasting model to suggest a reasonable price range for the listings

The implementations suggested to Airbnb (further explained in section 5) will be the following:

1. List popular districts / neighborhoods on search page for listings to bolster the understanding of potential customers towards geographic distribution of listings
2. Incentivise users to leave reviews for listing to decrease positive skewness of review rating and review sentiment
3. Allow users to choose 5 most importance factors on their profile and rank listings on search page of that user based on their personal preference
4. Disclaimers for potential guests that inform them if the listing is overpriced or underpriced based on a feature importance model for price

2. Methodology

CRISP-DM is known as the cross-industry process for data mining. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology.

The major phases of CRISP-DM Framework are –

1. Business Understanding
2. Data Preparation
3. Data Understanding
4. Modeling
5. Evaluation

While the process of evaluation of our suggestions cannot be done at the current stage.

2.1 Business Understanding

Airbnb is an online marketplace connecting travelers with local hosts. On one side, the platform enables people to list their available space and earn extra income in the form of rent. On the other, Airbnb enables travelers to book unique home stays from local hosts, saving them money and giving them a chance to interact with locals. Catering to the on-demand travel industry, Airbnb is present in over 190 countries across the world.

Since its founding in 2008, Airbnb has risen dramatically over the past few years as an online marketplace for arranging homestays in various cities and countries. As more and more travelers opt to use their services, Airbnb has successfully disrupted the traditional hospitality industry; not only for travelers who are looking for a bang for their buck, but also for business travelers resorting to Airbnb as their leading provider of accommodation. In particular, the number of listings in Hong Kong have risen by more than 20% every year since 2010.

2.2 Data Preparation - Web Crawling

Python Libraries of Scrapy and JSON were used to scrape information provided on the Airbnb website. The initial ‘start_requests’ function calls the ‘parse_id’ function to scrape the search page API containing listings of ‘Hong Kong stays’ (see Fig. 2.2.1), which is obtained by the ‘XHR and fetch’ in Network activity (using Google Chrome). Since it is in a JSON file format, the content can be decoded with the ‘loads’ method,

and relevant information can be obtained from the API using the ‘get’ method. Such information includes the listing ID, latitude and longitude of the location, price, and price. A for loop is used to iterate through all the listings and to extract information of each listing. The data is stored in a dictionary with the name ‘items’.

However, some of the information of the listings are only available on the individual URL of the listing (eg. host description and neighborhood description) , which is stored in the individual API page (also in JSON format). This API can be accessed when visiting the individual URL and inspecting Network activity (See Fig. 2.2.2). The URL to this API is identical for all listings, the only difference being the listing ID contained in the URL. Within the for loop in the ‘parse_id’ function, the ‘parse_page’ function is called to scrape the listing API for each listing ID. The data extracted is stored in a dictionary named ‘listing_details’.

The reviews of a listing are stored in another API, also accessible on the individual listing URL with a similar approach as above with the listing API (See Fig. 2.4.3). Again, the URL to the review’s API are in a format distinguished by the listing ID. However, in order to view all reviews on a single JSON file, the ‘limit’ component of the URL is changed to 1000 and the ‘offset’ to 0. The most popular listings in Hong Kong have roughly 300 to 400 reviews, so changing the limit of reviews to 1000 will suffice. The ‘parse_reviews’ function is called inside the for loop of the ‘parse_id’ function, to scrape reviews in the order of listing IDs. A for loop is used to iterate through all the reviews and extracts data unique to each review apart from the review itself (eg. review ID, date of review, reviewer ID, the rating the reviewer gave). The data is stored to the ‘reviews_dict’ dictionary.

The Airbnb search page is an infinite scrolling page, which means that new data is fetched from the next search page API everytime it reaches the bottom. The search page API has an element called ‘has_next_page’, which determines whether or not there is a next API available. An if condition is used where if the ‘has_next_page’ element has a value of ‘True’ on the API, the ‘parse_id’ function is called again to scrape content of listings on the next page. The new API is accessible by altering the ‘items_offset’ and ‘section_offset’ variables in the URL.

3 JSON files were used to organize and categorize the data. The ‘hongkong_listing_overview’ file contains data extracted from the search page API in the ‘parse_id’ function and stored in the ‘items’ dictionary. The data scraped from the individual listing APIs in the ‘parse_page’ function and stored in the ‘listing_details’

dictionary is written into the ‘hongkong_listing_details’ file. Lastly, the ‘hongkong_listing_reviews’ file consists of data regarding each individual review submitted for each listing; this was achieved with the ‘parse_reviews’ function which stored data inside the ‘reviews_dict’ dictionary. The data was scraped on 2 Oct 2019.

Limitations

1.We did not have data for past years and hence could not compare current rental trends with past trends. Hence, there was an assumption made, particularly in the demand and supply section of the report to understand the booking trends.

2.There were certain features such as acceptance_rate, monthly_price and description that either contained missing values or values in the free-text format that was not easy to work on and hence were dropped from our analysis.

2.3 Data Understanding

2.3.1 Description of Data

The dataset comprises of three main tables:

listings - Detailed listings data showing 96 attributes for each of the listings. Some of the attributes used in the analysis are price(continuous), longitude (continuous), latitude (continuous), listing_type (categorical), is_superhost (categorical), neighbourhood (categorical), ratings (continuous) among others.

reviews - Detailed reviews given by the guests with 6 attributes. Key attributes include date (datetime), listing_id (discrete), reviewer_id (discrete) and comment (textual).

calendar - Provides details about booking for the next year by listing. Four attributes in total including listing_id (discrete), date(datetime), available (categorical) and price (continuous).

2.3.2 Assessment of Data Quality

In order to create the necessary visualizations and analysis, we had to perform some imputations and transformations on our dataset. There were no significant inconsistencies or anomalies in the information, but most of the columns / features that we were interested in did not contain data in the correct format and were therefore manipulated to maintain their meanings.

Key Feature Transformations

1. Comment (reviews): We extensively used this feature in our analysis. The dataset featured comments in multiple languages, including Chinese, Spanish, and English, making it difficult to interpret it. We subsetted information to include only the reviews in English and conducted text filtering to exclude popular stop words and phrases that do not contribute significantly to the review's purpose.
2. Price (listings, calendar): The price column contained data in string format with the currency symbol '$' and comma separator ',' attached to it. This column was manipulated to contain integer values for time-series and other analysis.
3. Date (calendar, listings, reviews): The date was contained in mm-dd-yyyy format. It was transformed multiple times during the analysis to obtain weekly, monthly or yearly insights.
4. Rating (listings): There are several ratings that hosts receive including 'location rating', 'cleanliness rating' and 'overall rating'. Values in these columns comprised of percentages, integers, and char string. Data were standardized and transformed to a similar scale.

2.4 Modeling

2.4.1 Sentiment Analysis

In order to carry out the sentiment analysis of the reviews, we make use of word embedding to extract negative/positive keywords from a review. We have made use of pre-trained word2vec model. The reason why we chose to extract keywords instead of using alternatives such as tf-idf was because most of the reviews were positively skewed; there were only around 5 percent of people leaving negative sentiment reviews.

Word2vec is a neural network model which is designed to predict the next word to appear. In the training process, each word is paired with the nearby words (within window size) in the document and is used as input and response. Therefore after being trained in large quantity, the embedding vector representing the word contains information in relation to other words; words that appeared together are closely located and words that do not appear together are distant. This is how each embedding vector of the word can be used to calculate the closeness between certain keywords.

<Word2vec model architecture>
From the diagram, the middle layer of states represents the embedding vector of a word.

Moreover, even for the reviewers were leaving bad reviews, reviews were a mixture of both good and bad sentiments so original models were not giving out good results. An example of such a review is as such:

“His apartment is very small but suited our needs. Convenient lo and Kitty was a helpful host. However, the air conditioning was very noisy so we could not sleep with it on, it was too hot at night. Also, it needs darker curtains because the building opposite is lit up every night”

The extracted negative/positive keywords for the example above respectively were:

['convenient', 'helpful'], ['small', ‘noisy’,'however']

Based on the number of negative and positive keywords, we have constructed a criterion for classifying whether a comment is negative or positive. Then for negative reviews we look into the average embeddings of the keywords and find the cosine similarity between categories of the negative keywords like ‘far’, ‘inconvenient’, ‘unfriendly’ and ‘dirty’; based on the cosine similarity, we further label the negative sentiment.

2.4.2 Price Modelling

To let the Airbnb customer know more about whether the listings they want to rent are overpriced or not, we have decided to build a model to predict the prices of potential listings based on historical data. So, the customers can take it as a reference and see how much the real prices and predicted prices differ.

From the exploratory data analysis, we know that the price depends on the location, number of bedrooms, entire apartment or not, etc. So those factors contribute to the patterns, premium location would typically lead to a higher price. However, all listings in the same area and have the number of bedrooms do not have the exact same price. The variation in price is the noise. Our goal in price modeling is to model the pattern and ignore the noise.

After that, we split our data into training and test sets, choose appropriate features to feed into the models, methods of feature engineering, and the models. We initially started with a forward selection process — selecting one feature, running the model, checking performance, and repeating the steps. I also manually created a few interaction terms, dummied (turning a categorical feature into a boolean matrix), and mapped (scoring/weighting categorical features) a few of the categorical features. Each feature we chose to feed into the model as is or manually engineered was intentional. Most of the reliance was on the feature's strong correlation to price and intuitive assumptions we made about whether or not a feature would have an impact on price.

3. Data Sources

Two major data sources were used in this project. The first source is the database of listings in Hong Kong from (http://insideairbnb.com/get-the-data.html). This contains historical data up till 13 July 2019, including reviews and details of each listing. The second source is the Airbnb website. The new data scraped acts as a continuation of the database compiled by Inside Airbnb.

One item that is not found on the database provided by Inside Airbnb is the individual rating score given by each reviewer on a listing, listed as 'review_rating' on the 'hongkong_listing_review" JSON file.

4. Findings

4.1 Basic Information

There have been 12,627 individual listings (up until 2 October 2019) in Hong Kong posted on Airbnb, spanning over 18 districts (Table 4.1.1). 9,712 or 76.91% of the listings are concentrated in three districts, with 4740(37.54%), 2624(20.78%), and 2348(18.60%) listings located in in Yau Tsim Mong, Central & Western, and Wan Chai respectively (Fig. 4.1.2). Tai Po, Tuen Mun, Kwai Tsing, and Wong Tai Sin consist of the least number of listings, having 57, 55, 43, and 26 listings accordingly.

The total number of listing hosts in Hong Kong amount to 5,490, with 4,117 (74.99%) being single listing hosts and 1,373 (25.01%) owning multiple listings. Likewise, with listings, a majority of hosts are located in Yau Tsim Mong (1,007 or 18.34%), Central & Western (1,910 or 34.79%), and Wan Chai (1,164 or 21.20%) (Table 4.1.3).

4.2 Host information

The highest number of listings owned by a single host is 518, followed by 268 and 143 (Table 4.2.1). Fig. 4.2.2, Fig. 4.2.3, and Fig 4.2.4 show detailed information of the top 3 hosts' listings, including the district and property type.

Airbnb distinguishes normal hosts and hosts with more experience and expertise in accommodating guests with a 'Superhost' status, which shows up as a badge that appears on the page of the listing. Airbnb audits their hosts' status quarterly every year with a set of requirements such as maintaining a 90% response rate or a 4.8 overall rating to determine whether or not a host can earn the Superhost status. There are a total of 692 superhosts based in Hong Kong whom had hosted 2174 listings. This infers that 17.2% of listings in Hong Kong are 'super' listings, while 12.6% of hosts are 'Superhosts'. In terms of districts, Yau Tsim Wong comes up top with 735 'Super' listings and Wan Chai comes in second with 609, whilst Central & Western comes in third with 281 (Fig. 4.2.5).

What makes someone a superhost?

Airbnb gives a small fraction of its reliable hosts the title of "Superhost." This is planned as an incentive program for both the host, Airbnb, and their guests to be a win - win. In the form of higher bookings, the superhost gets more revenue, the customer gets better service, and Airbnb gets satisfied customers.

But what does a Superhost need to be? The page of Airbnb has a set of requirements to be met to become one. Hosting a minimum of 10 stays in a year, maintaining a rate of response above 90%, having at least 80% 5-star reviews,rarely cancelling confirmed reservations, etc.
Here we analyze our dataset to see how the superhosts work on two parameters: "Response rate" and "Score".

A few interesting insights are offered by the scatter plot (Fig. 4.2.6). While most superhosts are in the high-rating: high-response-rate zone, we can also see some superhosts with less than 75% response rates (which violates Airbnb's 90%+ criteria). This is a tiny fraction of the hosts. Nearly all hosts were ranked 80 percent and above in terms of ratings.

Having said that, most Airbnb hosts are located in the high-rating: high-response region, but only a small fraction can be super hosts. So clearly, it takes much more than high ratings & response rates to become a Superhost.

What kind of hosts are reliable?
The result of t-test verifies that the average rating score of the listings owned by “Superhost” is significantly higher than that of other listings. To strengthen distributed trust on the platform, Airbnb also encourages hosts to upload their profile pictures and to complete profile verification. The effectiveness of host verification has been validated by the result of t-test as well, while there’s no significant difference of listing ratings between the hosts with and without a profile picture.

4.3 Property Types

Listings in Hong Kong are categorized into 4 types - namely Entire Home/Apartment, Private Room, Shared Room, and Hotel Rooms. The former two types take up most of the listings, with 6,100(48.3%) Entire Homes/Apartments and 5,544 (43.9%) Private Rooms totalling at roughly 92.2% of all listings. Shared Rooms and Hotel Rooms only amount to 7.79% of the listings, each having 607(4.81%) and 376(2.98%) respectively (Table 4.3.1 and Fig. 4.3.2). Whilst Yau Tsim Wong had almost 3000 Private Rooms listings, both Wan Chai and Central & Western had less than 1000 of the same categories. The three main districts had less than 2000 listings in terms of Entire Home/ Apartment, with other districts mainly composed of Entire Home/Apartments rather than Private Rooms (Fig. 4.3.3).

4.4 New Listings over Time

The number of listings on Airbnb Hong Kong has seen a tremendous growth since 2010. The highest increment rate falls between 2010 and 2015, with the number of total listings almost increasing at an exponential rate. During the same period, the number of new listings per year have also been increasing at a linear rate, with the biggest jump of 140% from 660 to 1589 new listings between 2012 to 2013 (Fig. 4.4.1). However, the new listings per year figure has been steadily decreasing since 2015 where it had reached 2374 new listings, dropping to 1061 in 2019. This is displayed in the logarithmic growth pattern on the number of total listings from 2015 onwards.

The number of new listings in each month in Fig. 4.2.2 portray a similar pattern to that of each year, but with a larger fluctuation rate between each datapoint. The spike in July of 2013 signifies to the number of new listings in that month reaching 584, which corresponds to the huge jump in new listings per year between 2012 and 2013. On average, there is no a certain month consists of the highest number of new listings throughout the year and contributes to the increase in new listings per year to the largest extent.

Given the geographic distribution of the number of listings by district, it is presumed that the growth in the new listings per year of the three main districts (Central & Western, Wan Chai, and Yau Tsim Wong) attribute the most to the overall trend of the new listings in Airbnb Hong Kong, and Fig. 4.4.3 displays the difference between the three over the years. Staring from mid-2013 onwards, the growth in new listings in Yau Tsim Wong has dominated the overall trend and has seen losses on a smaller extent since 2015 compared to the other two.

4.5 Demand and Price Findings

In this section, we will examine the demand for Airbnb listings in Hong Kong. We will look at demand over the years since the launch of Airbnb in 2010 and throughout the months of the year, in order to understand the seasonality. We also wish to explore the relation between price and demand. The question we would like to address is whether the prices of listings correlated with the demand. We will also carry out a more in-depth analysis to understand how prices vary by days of the week.

In figure 4.5.7, looking at the price only, we can see that southern district have a much higher average price compared to all the others.

To study the demand, since we did not have information on the bookings made over the past year, we will use ‘number of reviews’ variable as the market metric for demand. According to Airbnb, about 50 percent of customers review their hosts/listings, so that we will get a good estimate of the demand by analyzing the amount of reviews.

There are several major findings:

1. How popular has Airbnb become in Hong Kong?

   In Figure 4.5.1, the number of unique listings receiving reviews has increased over the years. We can see an almost exponential increase in the number of reviews until mid-2019, which as discussed earlier, indicates an exponential increase in the demand. However, starting from mid-2019, we can observe a shape decrease in the number of reviews (indicator of demand for Hong Kong airbnb listings).

   This phenomenon is probably due to the incidents happened in Hong Kong. Different authorities claimed that the number of tourists coming to Hong Kong has largely reduced and the rental price of hotel rooms has dropped to a historical low point too. These two factors would reduce the demand for Airbnb listings in Hong Kong and it coincides with our findings.

2. Seasonality in demand

   In Figure 4.5.2, 4.5.3 and 4.5.4, we can observe that the number of reviews(demand) also reveals a seasonal pattern. The dots in the above graph reflect the number of reviews written on a particular month, which as per our assumption reflects the demand. There seems to be a consistent pattern in how demand fluctuates across the year, which is reflected in each of the graphs shown above. The demand is the lowest in June and increases until December, when it begins to fall again.

   Every year there are peaks and drop in the demand, indicating that the demand for Airbnb listings in Hong Kong vary with the period of time. Also, we can observe that the peak period starts from July and ends in January every year which is the typical period for people all around the world to travel.

3. How is Airbnb priced across the year?

After observing the patterns in the demand, we are going to investigate whether the prices of the listings following similar patterns. To address the above question, we looked at the daily average prices of the listings across the years using the data from the calendar table.

The average listing prices tend to rise as one progresses throughout the year and peaks in December. The trend is similar to the number of reviews / demands except in the months of November and December, when the number of reviews (indicating demand) is starting to drop. This seems counter-intuitive as with a decrease in demand, one would expect the price to decrease. (See Figure 4.5.5)

4. We can also see two sets of points on the graphs above, which indicate that on certain days average prices were higher than on the other days. Instead, to explain this trend, we must show a box plot of average prices per day of the week. Fridays and Saturdays are, as we can see, more expensive than other days of the week, possibly due to higher demand for accommodation. (See Figure 4.5.6)

Value of listings

Airbnb enables tourists to evaluate listings and their experiences by rating and leaving reviews commenting on cleanliness, accuracy, check-in process, communication with hosts, location, and value. Therefore, “value” refers to the payoff between the cost paid and benefit received. In figure 4.5.8, listings in Sai Kung have the highest value among all districts while those in Kowloon city have the lowest.

4.6 Accuracy of Review

In this section, we will see how accurate the review is in order to allow customers for a better decision on accommodation. In Figure 4.6.1, the review scores by districts are mostly in the range between *85 and 100*. For example, Central Western Islands, Yuen Long, Eastern, Shatin and Tuen Mun are all within the range from *90 to 100*. Moreover, from the correlation between the number of reviews and negative reviews, we could see that users with pessimistic experience tend not to leave comments and reviews on Airbnb web, which could not reflect the overall quality of the listings (see Figure 4.6.2). Furthermore, as mentioned in Section 4.5, only half of the Airbnb users had made reviews after finishing the whole customer experience in Airbnb. Therefore, it could be seen that listings with a low number of reviews tend to overestimate the quality of

listings. With that being the case, it is seemingly difficult for customers to make bookings based on the reviews given by past users.

Box chart of Review Score by district in Airbnb Hong Kong

4.7 Negative Review

In this section we will be discussing the findings by looking deeper into the negative sentiments of the reviews. First of all, we have seen if there is a correlation between the number of reviews by a particular user and proportion of negative reviews.

In figure 4.6.2, it is clear that there is an inverse proportion relationship between the two variables. This means that users with a greater number of total reviews is less likely to be leaving a negative review; this relationship can also be seen as people who travel more being less sensitive to unfavorable environment.

Secondly to look at the consistency in the users' reviews, we have considered users who have left more than one negative review and see if they consistently left negative issues regarding the same issue. The histogram of the consistency is plotted below:
From the histogram (figure 4.7.1), we can see that most of the people are consistent with their negative reviews. For example, a person who complained about dirtiness of one listing is more likely to complain about the dirtiness of another place. So, it is clear that people have particular priorities they look into when staying at a listing.

Next, we looked at the number of labels for the negative comments in figure 4.7.2.

Because label 'bad' has an ambiguous meaning and 'bad' and 'dirty' has a very high cosine similarity, we have chosen to remove bad as a label; the new histogram of the negative labels are changes as the right. From these diagrams we can infer that people found Hong Kong listings to be small and for the other cases they found problems in friendliness and cleanliness in the listings.

Thirdly with the labels of the negative sentiment, we were trying to see in figure 4.7.2 if we can find what users found most problematic based on neighborhood. This is done by counting collecting all the negative keywords based on neighborhood. And for a more detailed analysis, we have divided the most crowded area (Yau Tsim Mong, Wan Chai, Central & Western) into 6 neighborhoods.

From the diagram, each of the column ‘bad’, ‘dirty’, ‘unfriend’, ‘far’, ‘inconvenient’ and ‘disappoint’ represents the measure of closeness to the keywords. With these figures, we label each neighborhood with the label of the highest column. This is possible because each review was all represented in the same way.

We notice that for most of the area people found small size to be the most frequent problem. By looking at the Yuen Long being ‘far’, we can say that the result is somewhat fair. Since we have a measure for determining what people found problematic about the neighborhood, we will try to make use of this to incorporate it into our suggestion.

The sentiments of each user and by location is utilizable in further understanding of customer experience. Through the reviews people leave on the accommodation, it gives us a way to compare users in a more systematic way. For example, based on the preference and priorities a user has, we can find travellers with similar preference and recommend places similar travellers enjoyed.

4.8 Feature Importance table for price forecasting model

Price forecasting modelling in Section 2.4.2 explains how model interpretes the historistic data and which features have a higher influence on the predicted values of price. Figure 4.8.1 exhibits the feature importance of the top 30 variables of the listings towards its price; whether the property is an Entire Home/Apartment has the largest weighting and is the dominating factor towards price. Features such as the ‘number of rooms’, ‘super strict 30’, ‘free parking on premise’, and ‘gym’ take up a relatively smaller weighting compared to the ‘Entire Home/Apartment’ property type. The variable ‘super strict 30’ indicates whether the listing can be cancelled and refunded or rescheduled within the 30 days period prior to the arranged date. This model could be useful for hosts when they are deciding the price for listing the property, as well as guests who are seeking for an estimated price of a certain type of property.

4.9 Cancellation Policy

What about the host’s cancellation policy? Airbnb offers hosts five cancellation choices — flexible, moderate, strict with a grace period, super strict 30 days, and super strict 60 days. The former three policies respectively correspond to a full refund of accommodation fees for a cancellation made 24 hours, five days, and 14 days before check-in time, while the latter two only provide a 50% refund for a cancellation made 30 and 60 days prior.

In figure 4.9.1, The rating scores of hosts who select super strict cancellation policy are, as expected, significantly lower than others, while those selecting 14-day strict cancellation policy are still dependable. In contrast, hosts with moderate cancellation policy have highest rating scores among all the hosts on Airbnb.

Based on the host analysis, the portrait of an ideal host involves complete profile verification, “Superhost” authentication and relatively flexible cancellation policy. Since the category of “strict” has been totally substituted by “strict with grace period” and there are only 63 listings in the dataset with “super strict 30 days” cancellation policy, I removed these two categories from Tukey’s HSD (Honestly Significant Difference) test.

5. Conclusions and Suggestions

5.1 Conclusions

Through this data analysis and visualization project, we gained several interesting insights into the Airbnb rental market. Below we will summarise the answers to the questions that we wished to answer at the beginning of the project:

How do prices and values of listings vary by location? What localities in Hong Kong are rated highly by guests?

Southern district has the most expensive rentals compared to the others. Rentals that are rated highly on the location by the host also have higher prices. However, listings in Sai Kung are the most cost-effective.

How does the demand for Airbnb rentals fluctuate across the year and over years?

The demand (assuming that it can be inferred from the number of reviews) shows a seasonal pattern - demand increases from June to December, then drops slightly in January. In general, the demand for Airbnb listings has been steadily increasing over the years not until the political incidents happened in June 2019. The demand for Airbnb rentals dropped first-ever and sharply right after this period.

Are the demand and prices of the rentals correlated?

Average prices of the rentals increase across the year, which correlates with demand. This observation is assumed in our intuitives and the basic rules in Economics. Moveover, Prices are higher on average on Fridays and Saturdays, compared to the other days of the week.

What are the different types of properties in Hong Kong? Do they vary by neighborhood?

There are more than 20 different types of listings in Hong Kong. The ratio of the type of listings to total numbers varies by district. Yau Tsim Mong and Central tend to have property types that are smaller and can only accommodate a smaller number of people.

What makes a host a Super host?

Ratings and Response rates tend to have a direct correlation with a host being ‘promoted’ to the status of the Super host. However, there are other factors too that makes someone a super host as not all hosts with high ratings and response rates were superb hosts.

Do regular hosts and super hosts have different cancellation and booking policies?

Both have similar cancellation and booking policies.

Are there any common themes that can be identified from the free-text section of the reviews? What aspects of the rental experience do people like and what aspects do they abhor?

There are certain words such as words such as “quiet”, “walkable”, “clean”, “spotless” that are associated with the word “comfortable” demonstrating the importance of environment, location and cleanliness. Words associated with “uncomfortable” include “dirty”, “crowded”,“small”, “stuffy” “cluttered” which indicate that unclean environment and lack of space are the most common complaints.

5.2 Suggestions

5.2.1 Geographical Distribution of Listings:

Since 76.91% of listings in Hong Kong are concentrated in three districts (Yau Tsim Wong, Central & Western, Wan Chai), customers are more than likely to select a location from one of the three. As of now, the default search page features a user interactable map with the listings and its price (Fig. 5.1), as well as filters that allow users to search for listings in specific neighbourhoods (Fig. 5.2). However, it does not tell the users about how the listings are distributed geographically.

It is suggested that Airbnb displays the most popular districts or neighbourhood on the search page when a user searches for listings in a city, as well as the number of listings currently listed on each district/neighborhood. Additional information could also be displayed, such as the average price or rating, number of visits in the past year, etc.

The advantages that arise from including feature are listed as below:

Increase Customer Understanding: Users will be able to expand their knowledge base of how listings in different districts differ from one another. It will assist potential customers to plan their schedule during their stay by providing concrete geographic options for their accommodation needs.

Consumer Confidence in Airbnb: The increased transparency to the variety of listings offered to the user may be able to boost consumer confidence in Airbnb as a reliable source of accommodation. The ability to make informed decisions allows

customers to have a higher perceived control over the transaction, which ultimately leads to them being more comfortable with making the transaction.

User Friendliness to First-time visitors: First-time visitors with limited knowledge about characteristics associated with certain districts and neighbourhoods will be hesitant in making a transaction on Airbnb. By understanding which locations are more popular among other users, they will be able to plan out their stay more effectively versus having no prior knowledge about listings in different locations.

5.2.2 Monetary Incentive for Reviewers

Offer a future discount: In view of low review rates, which could not reflect the accurate reviews for customers, Airbnb could offer discounts on future purchase after customers finished the reviews on their previous bookings. Not only are businesses already adept to determining and offering discounts, but customers are already comfortable with collecting and applying coupons when making online purchases. Either a) by cash or b) by percentage-based discount would be the options for the future discount.

Prevent hosts incentivizing users for positive reviews: Airbnb should explicitly prohibit incentives from users and warns of manual penalties including the dreaded red badge. This policy applies to many review websites. Penalties will be given to hosts that offer incentives to their guests for writing reviews. In all cases, reviews will be removed, and red badge will be added to the hosts for further penalties. The red badge warns users that the listing has not adhered to Airbnb policy and significantly impacts the listing's popularity ranking.

5.2.3 Personalised Rankings on Search Page

Users will be offered to select 3 (among 10) most important factors when considering renting the accommodation. In return, Airbnb will help analyse which listings are much more suitable to the customers by ranking them accordingly. Users will be expected to see the most relevant and particular listings suggested by Airbnb on search page. Personalised rankings would offer users a better customer experience with the reduction of searching time and the increase of the chance matching between hosts and users. We expect that 20% more transactions would be made between hosts and users when it comes to the personalised rankings feature.

5.2.4 Disclaimer for Guests Determining if Listing is Overpriced or underpriced

One of the benefits for hosts to list their property on Airbnb is the flexibility that it provides compared to traditional rental methods; options such as dates of availability or self-determined pricing of the listing are just some of the potential benefits that property owners may find attractive. It is therefore counter-intuitive to force Airbnb hosts to follow a strict pricing method or rubric based on predetermined factors or features of the property. Currently, Airbnb has a smart pricing system that uses an algorithm to determine if a property is expensive or cheap compared to similar listings; but this information is only available to hosts when they are setting the price of a listing.

The machine learning model (explained in Sections 2.4.2 and 4.8) can be used to predict prices with an RMSE of 0.908462 based on the previous data collected where the predicted results are rather accurate. The price forecasting model can be used to provide insights to potential guests when they are searching for listings to stay in. Based on the filters applied by the guest as well as factors selected in Section 5.3, there would be a predicted price range for the type of listing on the search page. On each individual listing, it will tell users if this particular listing is overpriced or underpriced compared to similar listings. These measures will enable users to effectively and efficiently searching for a suitable stay for themselves, as well as reduce the time taken on comparing different listings with similar properties/features.

Appendix

Fig. 2.2.1

Sample of Search Page API

Fig. 2.2.2

Sample of Room Details API of Individual Listing

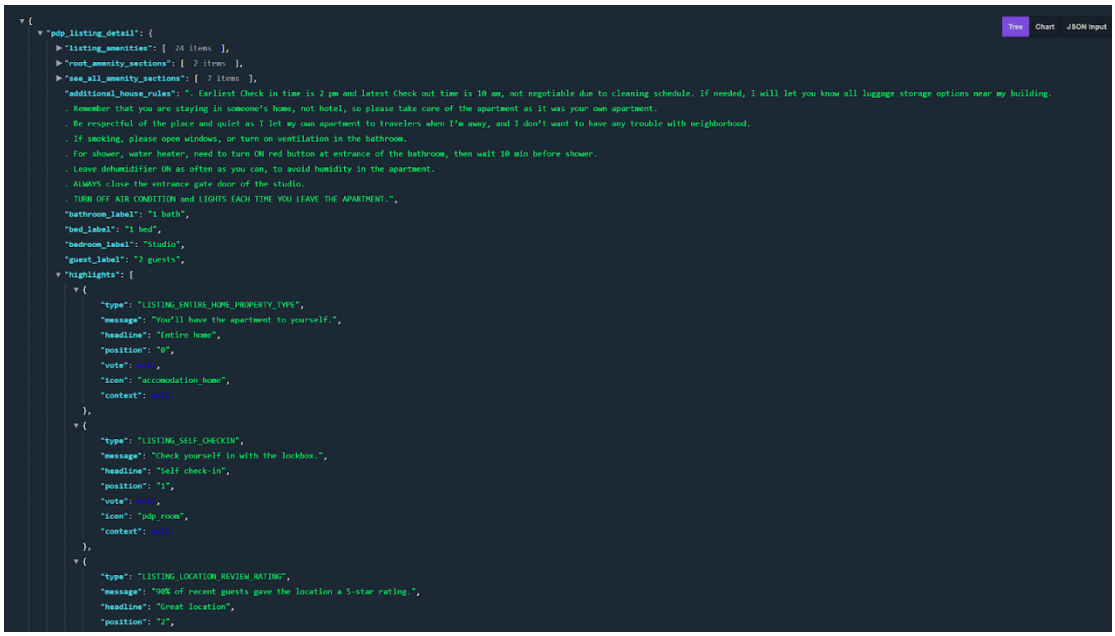
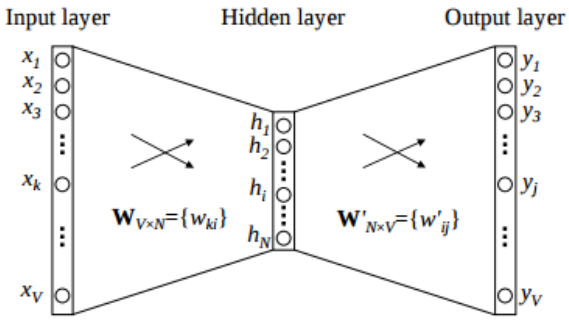


Fig. 2.2.3

Sample of Reviews API of Individual Listing


```
{
  "reviews": [
    {
      "comments": "Had a great time here! Quick responses and great check in information. So accommodating",
      "id": 449090713,
      "language": "en",
      "created_at": "2019-05-06T06:56:10Z",
      "reviewee": {
        "deleted": false,
        "first_name": "Atlas Hostel & Backpackers",
        "host_name": "Atlas Hostel & Backpackers",
        "id": 45970651,
        "picture_url": "https://a0.muscache.com/im/pictures/user/73c7d2b6-7994-4b68-9659-48d7d8ec725c.jpg?aki_policy=profile_x_medium",
        "profile_path": "/users/show/45970651",
        "is_superhost": false
      },
      "reviewer": {
        "deleted": false,
        "first_name": "Alex M",
        "host_name": "Alex M",
        "id": 71621222,
        "picture_url": "https://a0.muscache.com/im/pictures/0123c2a2-7048-4d00-a5b9-79a2eac2eccb.jpg?aki_policy=profile_x_medium",
        "profile_path": "/users/show/71621222",
        "is_superhost": false
      },
      "localized_date": "May 2019",
      "rating": 5
    },
    {
      "comments": "good value, kinda small",
      "id": 417351303,
      "language": "en",
      "created_at": "2019-02-27T05:42:12Z",
      "reviewee": {
        "deleted": false,
        "first_name": "Atlas Hostel & Backpackers",
        "host_name": "Atlas Hostel & Backpackers",
        "id": 45970651,
        "picture_url": "https://a0.muscache.com/im/pictures/user/73c7d2b6-7994-4b68-9659-48d7d8ec725c.jpg?aki_policy=profile_x_medium",
        "profile_path": "/users/show/45970651",
        "is_superhost": false
      },
```


Fig 2.4.1



word2vec model architecture

Table 4.1.1 and Fig. 4.1.2

	district	count
0	Central & Western	2624
1	Eastern	398
2	Islands	538
3	Kowloon City	487
4	Kwai Tsing	43
5	Kwun Tong	68
6	North	272
7	Sai Kung	157
8	Sha Tin	119
9	Sham Shui Po	258
10	Southern	106
11	Tai Po	57
12	Tsuen Wan	75
13	Tuen Mun	55
14	Wan Chai	2348
15	Wong Tai Sin	26
16	Yau Tsim Mong	4740
17	Yuen Long	256

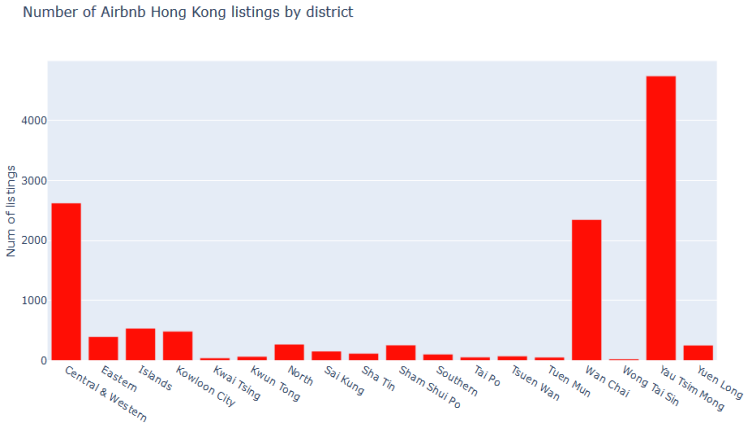
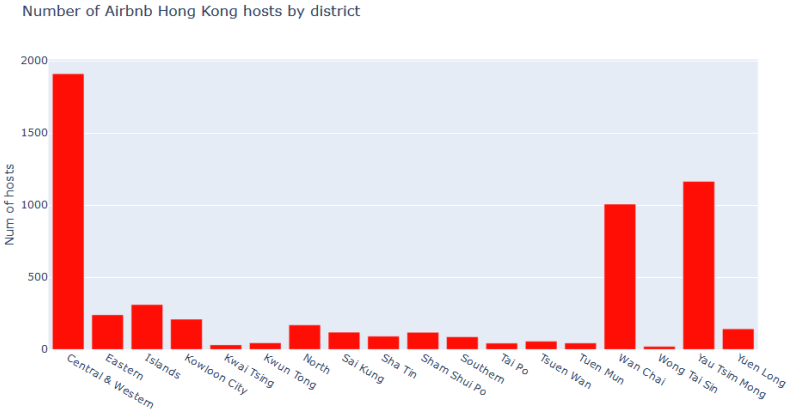


Table 4.1.3 and Fig. 4.1.4

	district	host_count
0	Central & Western	1910
1	Eastern	240
2	Islands	311
3	Kowloon City	210
4	Kwai Tsing	32
5	Kwun Tong	46
6	North	170
7	Sai Kung	120
8	Sha Tin	92
9	Sham Shui Po	119
10	Southern	88
11	Tai Po	44
12	Tsuen Wan	57
13	Tuen Mun	45
14	Wan Chai	1007
15	Wong Tai Sin	21
16	Yau Tsim Mong	1164
17	Yuen Long	143



Number of hosts own one listing only : 4117, Number of hosts own multi listing : 1373

Table 4.2.1

Top 10 hosts in term of number of listings owned

	host_id	number of listings owned
710	7518056	518
3609	97240131	268
3958	122131447	143
4150	138649185	123
476	4584648	111
1231	14861546	102
3197	67709885	95
2898	52473150	72
4333	156409670	67
262	2767794	64

Fig. 4.2.2

Information about top 1 host

neighbourhood_cleansed	id
Central & Western	44
Eastern	10
Kowloon City	75
Sham Shui Po	12
Wan Chai	322
Yau Tsim Mong	55

property_type	id
Apartment	515
Condominium	1
Serviced apartment	2

Fig. 4.2.3

Top 2 host

neighbourhood_cleansed	id
Central & Western	88
Eastern	27
Kowloon City	11
Sham Shui Po	15
Wan Chai	73
Yau Tsim Mong	54

property_type	id
Apartment	268

Fig. 4.2.4

Top 3 host

neighbourhood_cleansed	id
Kowloon City	33
Yau Tsim Mong	110

property_type	id
Apartment	33
Condominium	30
Serviced apartment	80

Superhost

[How to become a superhost?]

Number of superhosts 692 Number of listings from superhosts 2174

Fig. 4.2.5

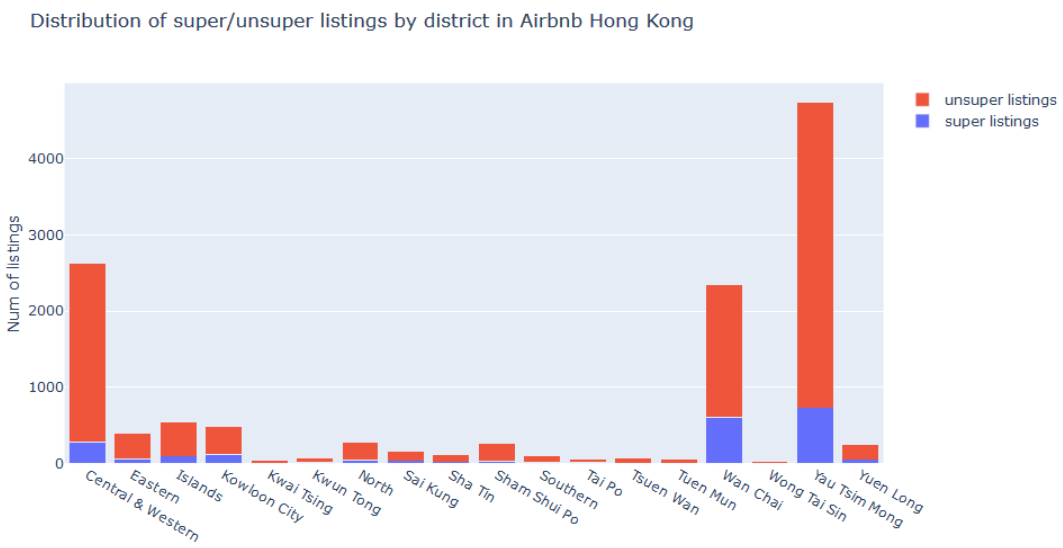


Fig. 4.2.6

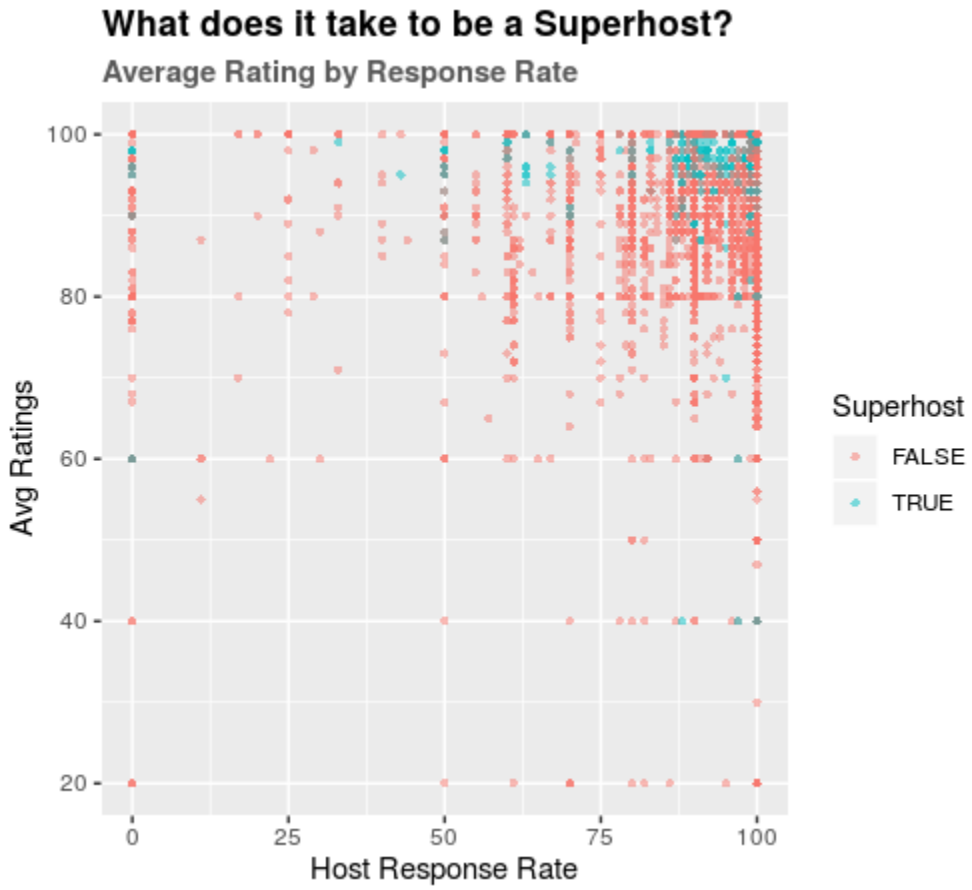


Fig 4.2.7

Fig 4.2.8

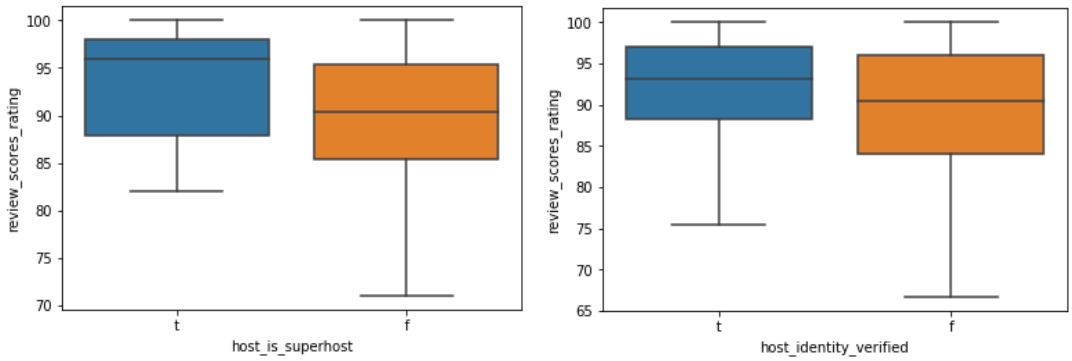
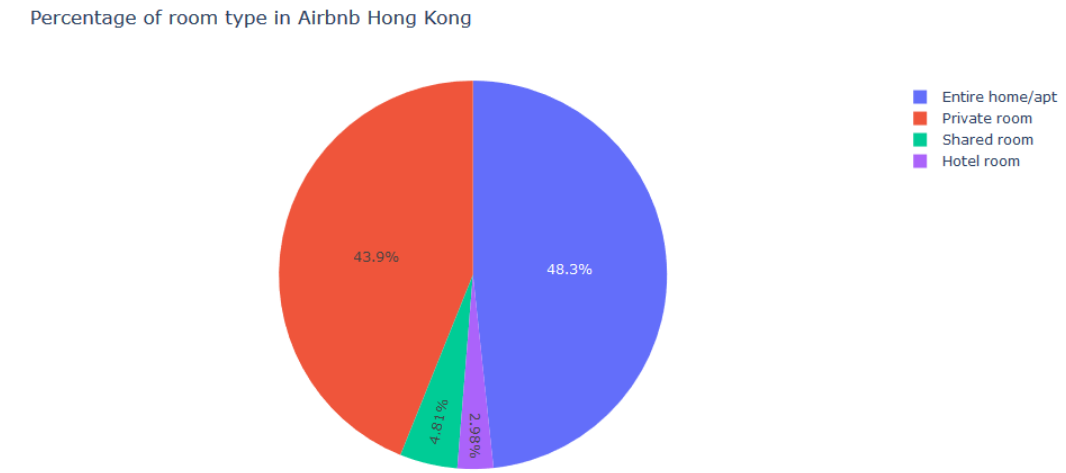
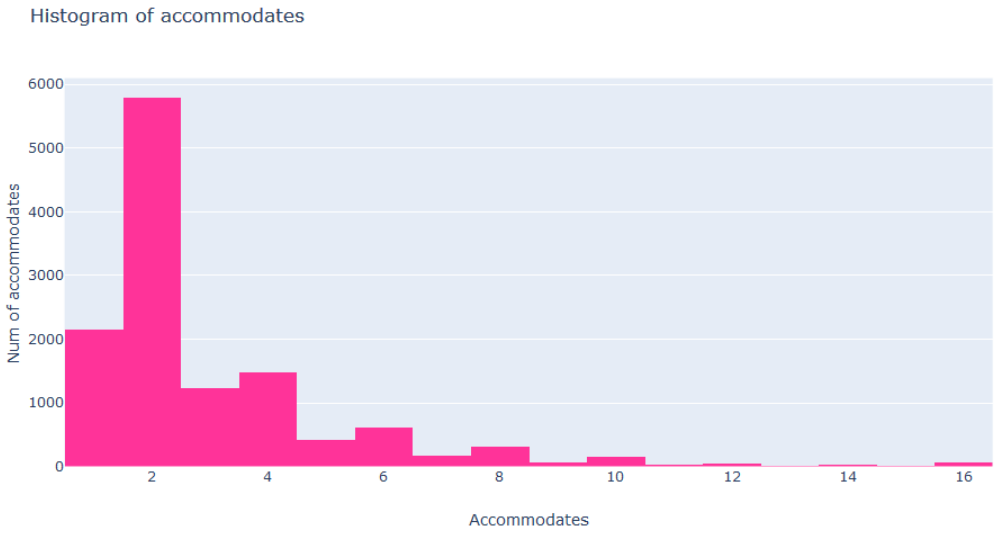


Table 4.3.1

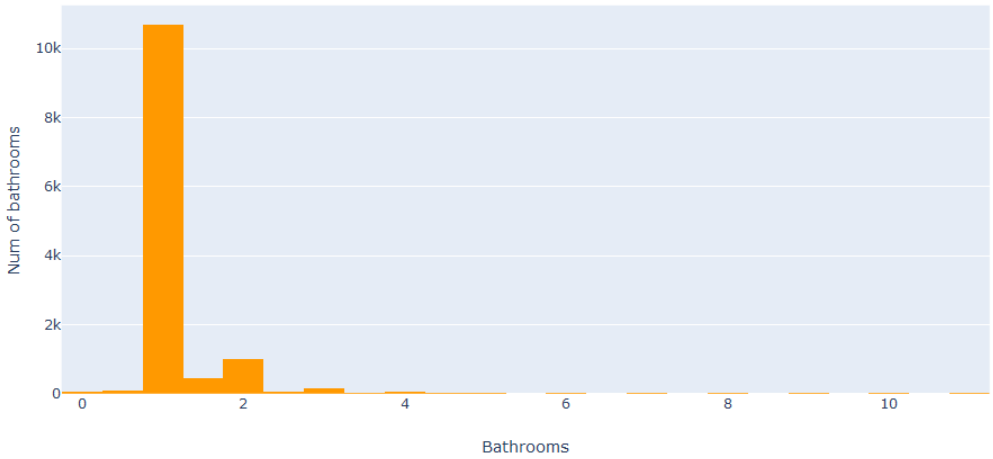
Entire home/apt	6100
Private room	5544
Shared room	607
Hotel room	376

Fig. 4.3.2

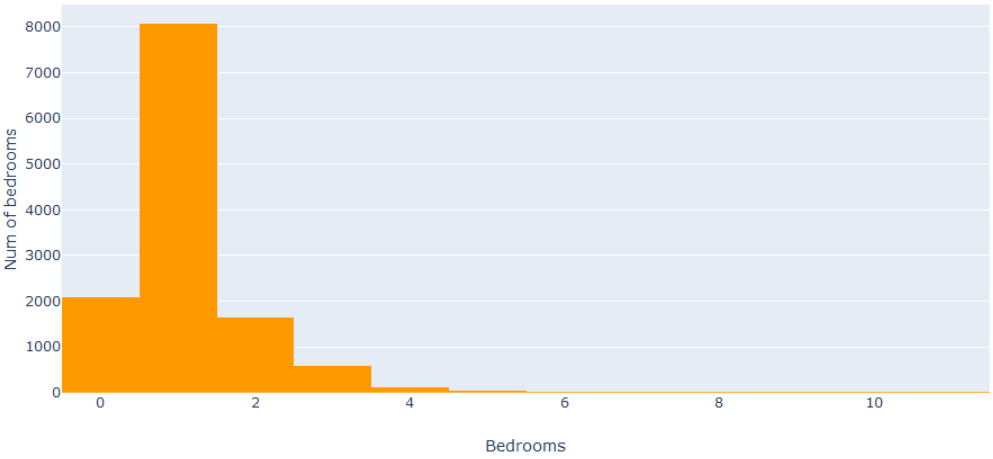

Percentage of room type in Airbnb Hong Kong
Entire home/apt
Private room
Shared room
Hotel room
48.3%
43.9%
4.81%
2.98%


Histogram of accommodates
Num of accommodates
6000
5000
4000
3000
2000
1000
0
2
4
6
8
10
12
14
16
Accommodates

Histogram of bathrooms



Histogram of bedrooms



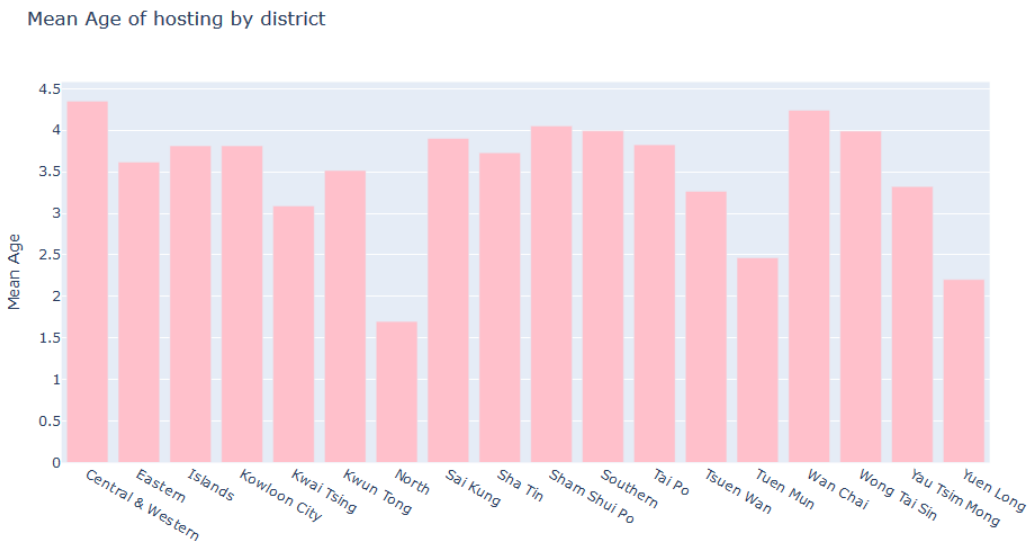

Mean Age of hosting by district
Mean Age
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Central & Western
Eastern
Islands
Kowloon City
Kwai Tsing
Kwun Tong
North
Sai Kung
Sha Tin
Sham Shui Po
Southern
Tai Po
Tsuen Wan
Tuen Mun
Wan Chai
Wong Tai Sin
Yau Tsim Mong
Yuen Long

Fig. 4.3.3

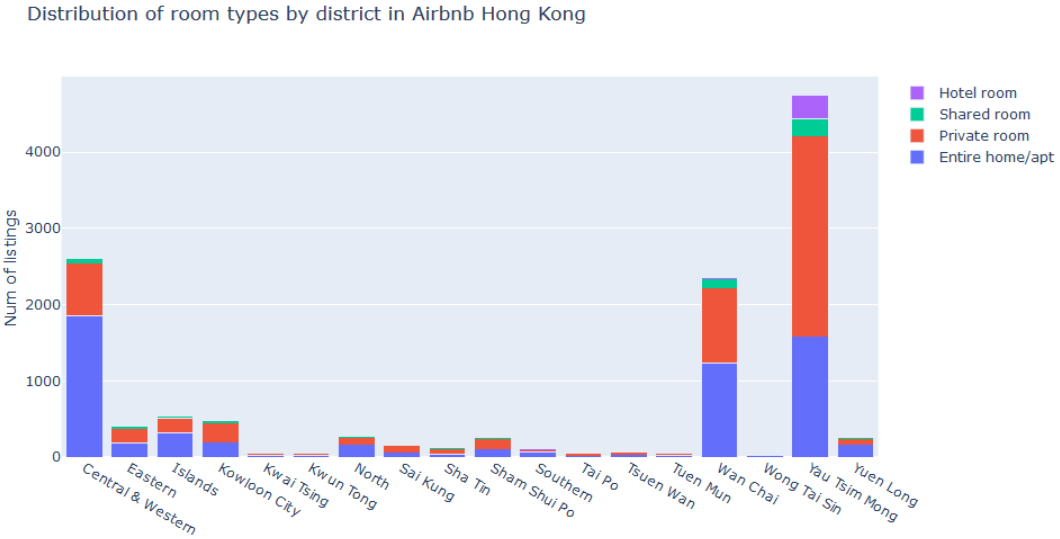

Distribution of room types by district in Airbnb Hong Kong
Num of listings
4000
3000
2000
1000
0
Hotel room
Shared room
Private room
Entire home/apt
Central & Western
Eastern
Islands
Kowloon City
Kwai Tsing
Kwun Tong
North
Sai Kung
Sha Tin
Sham Shui Po
Southern
Tai Po
Tsuen Wan
Tuen Mun
Wan Chai
Wong Tai Sin
Yau Tsim Mong
Yuen Long

Fig. 4.4.1

Fig. 4.4.2

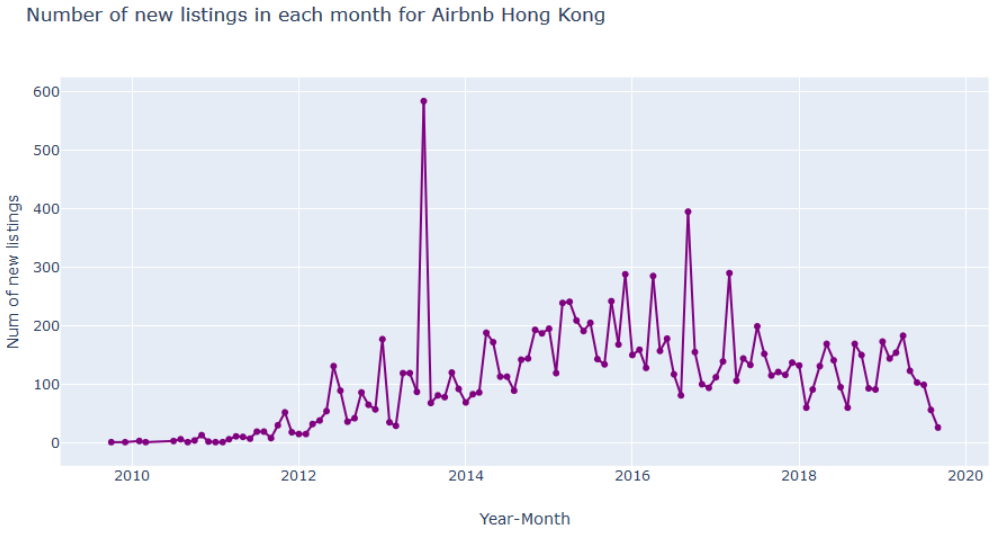


Fig. 4.4.3

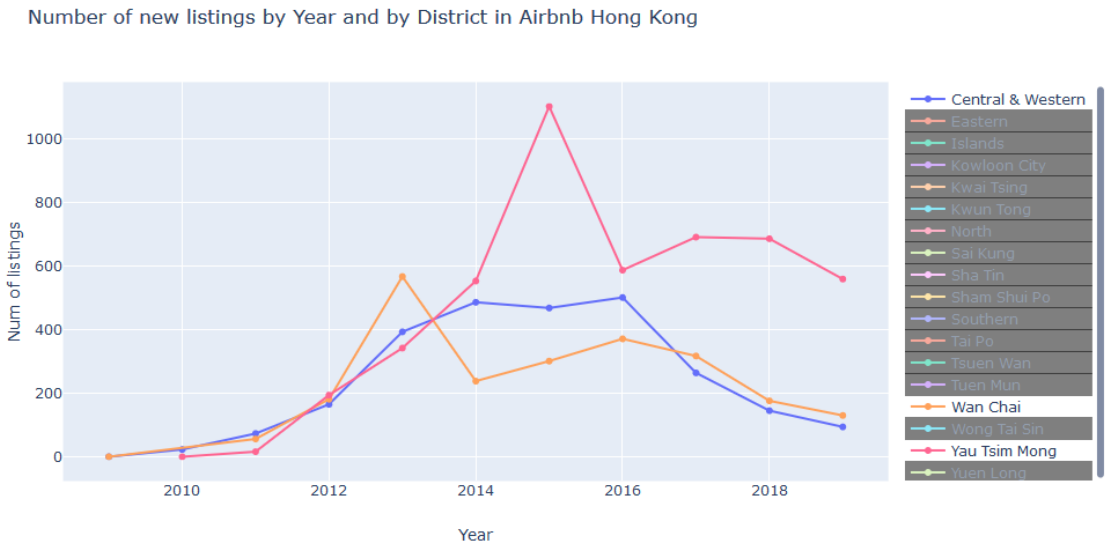


Fig. 4.5.1

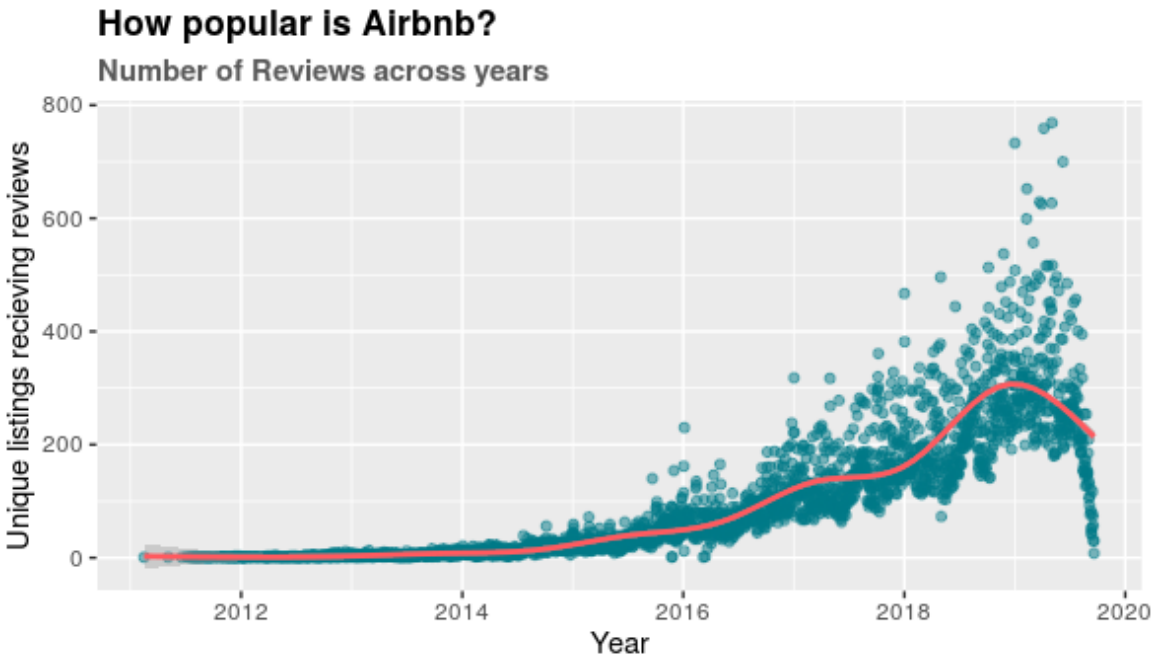


Fig. 4.5.2

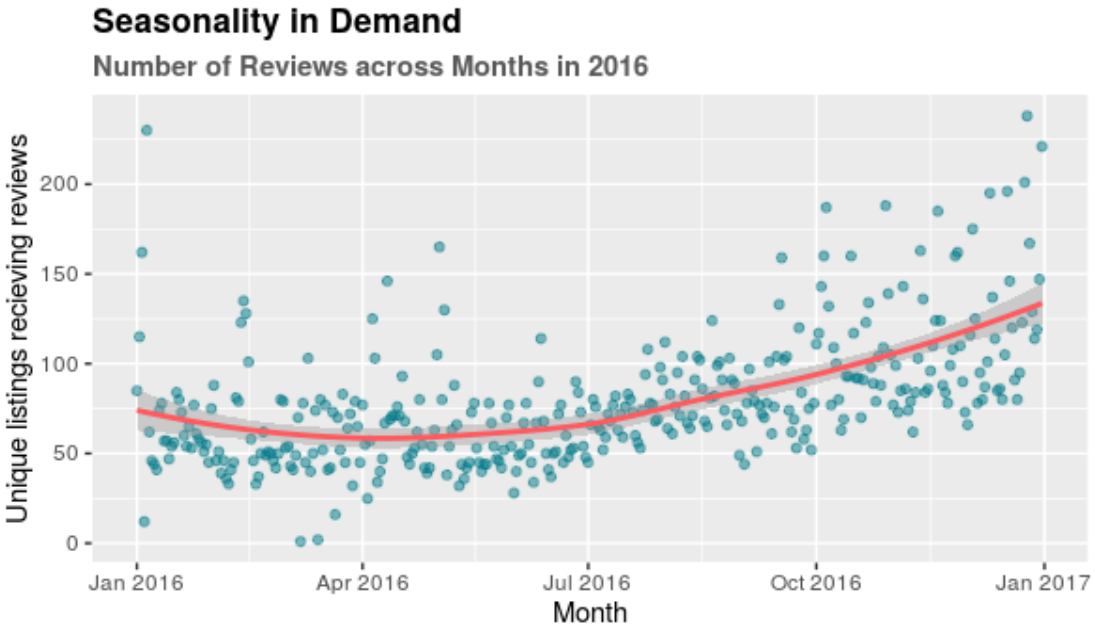


Fig. 4.5.3

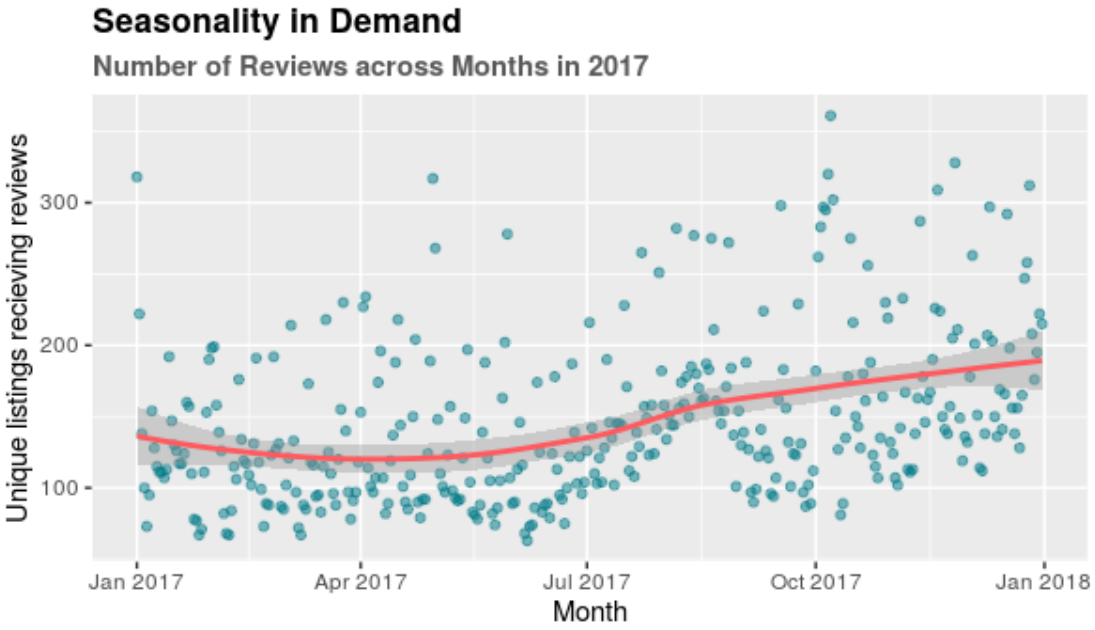


Fig 4.5.4

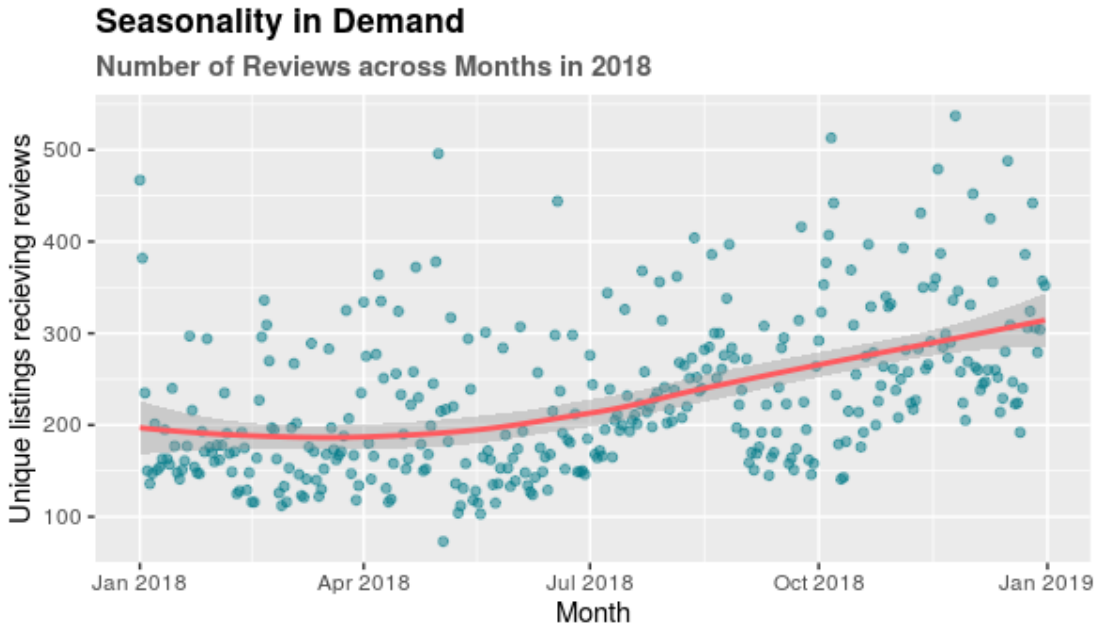


Figure 4.5.5

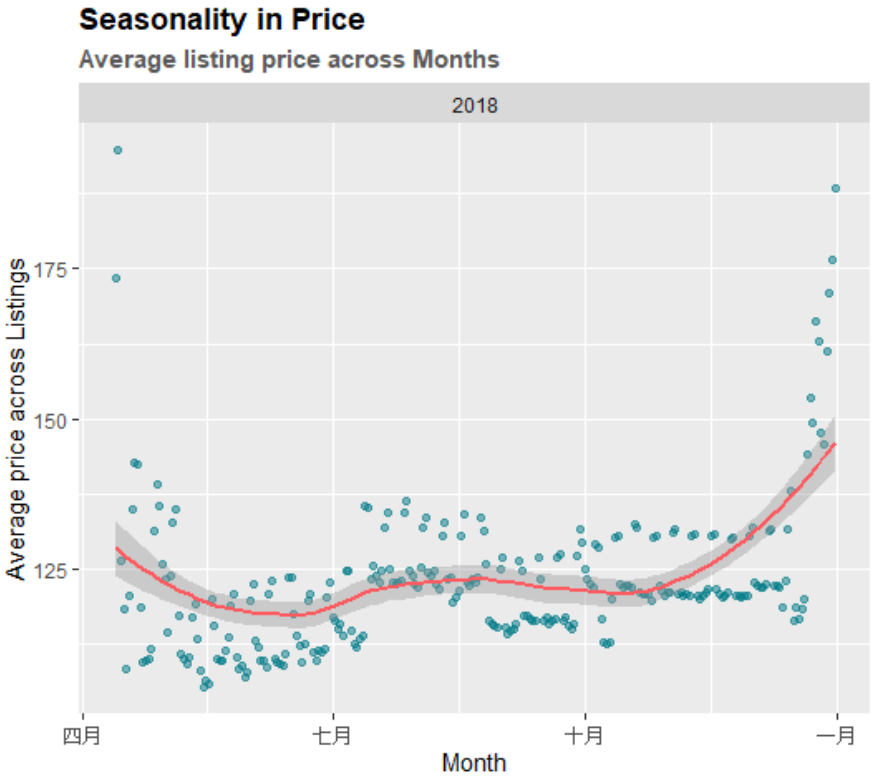


Figure 4.5.6

Listing Prices by Day of the Week

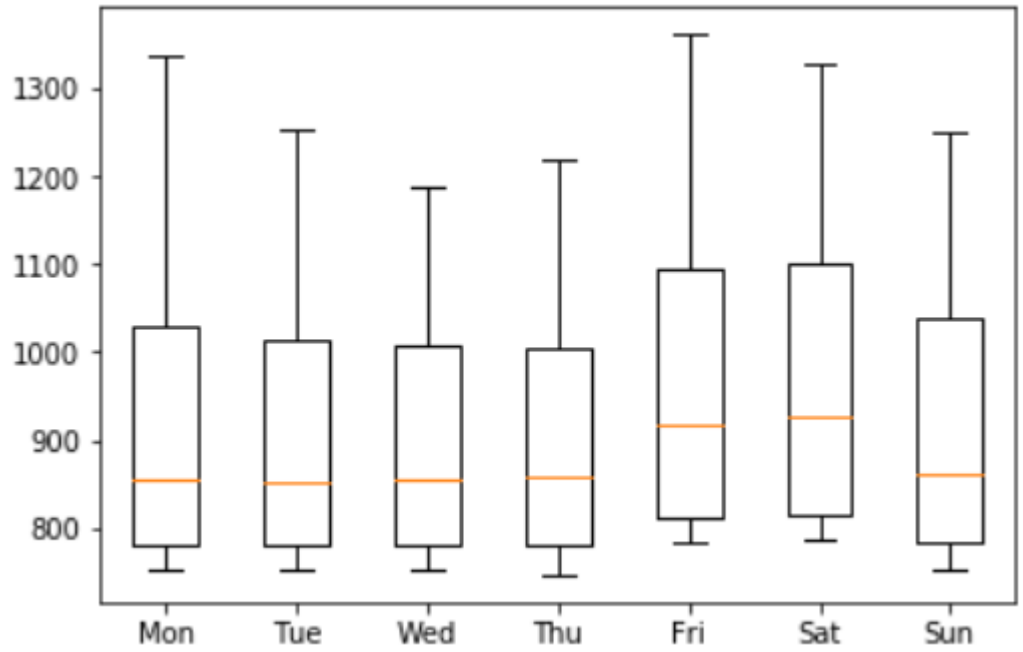


Figure 4.5.7

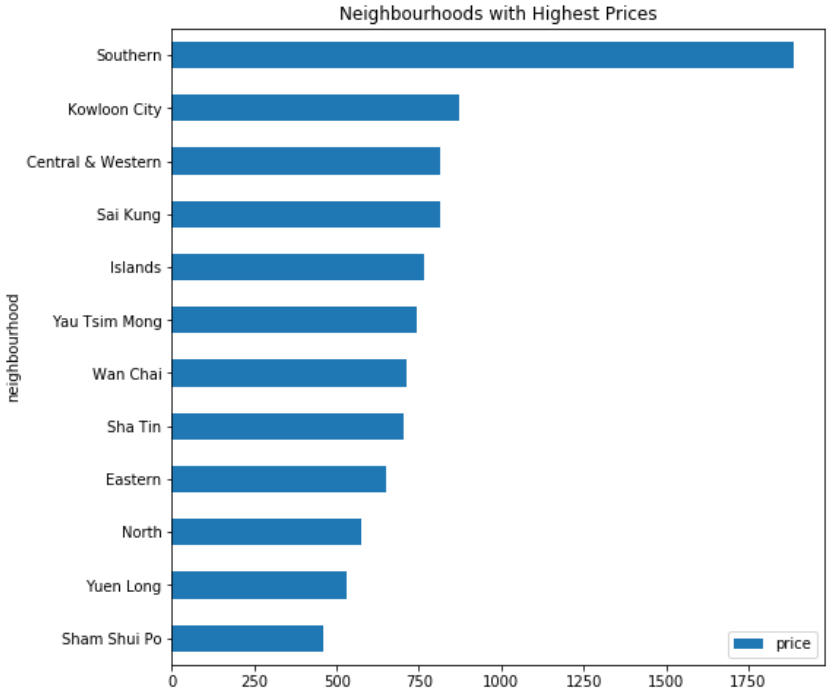


Figure 4.5.8

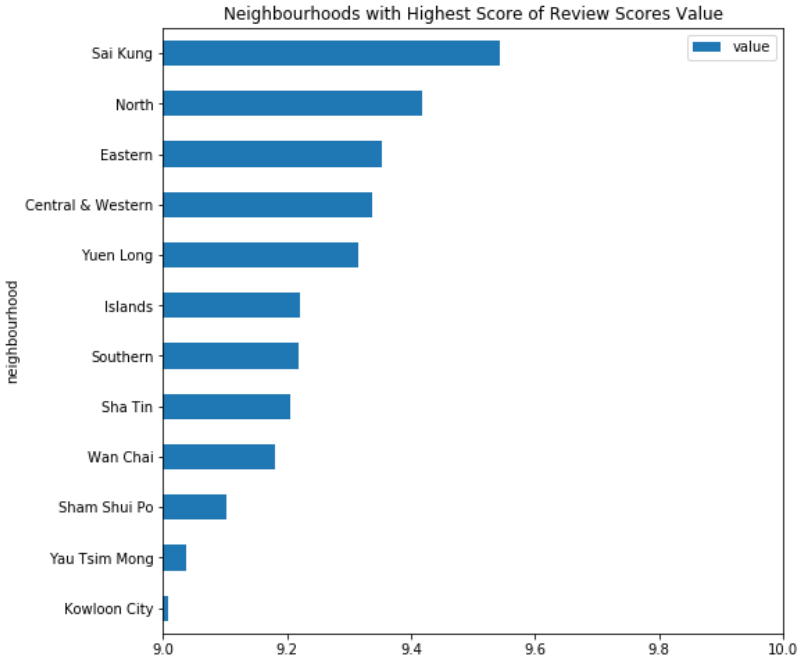


Figure 4.6.1

Box chart of Review Score by district in Airbnb Hong Kong

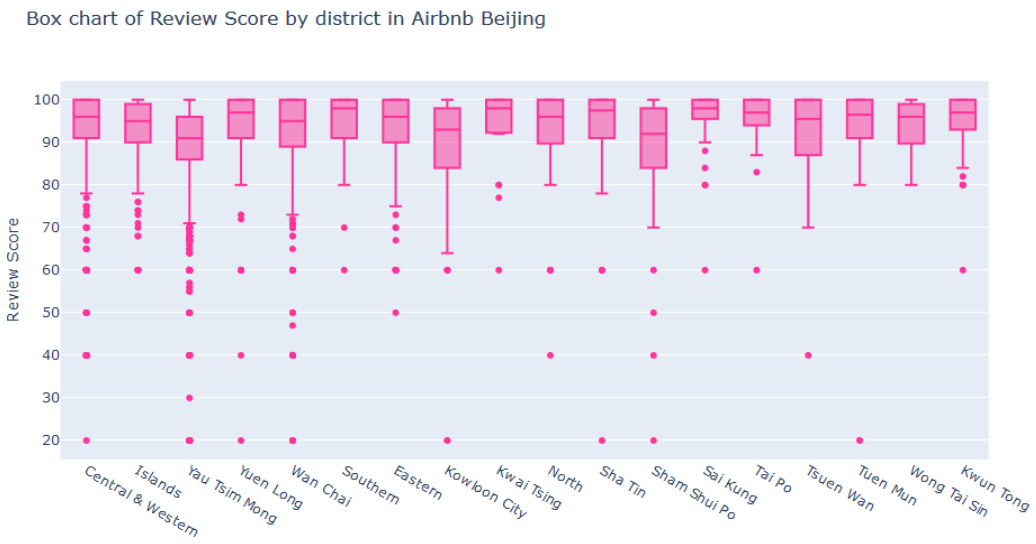
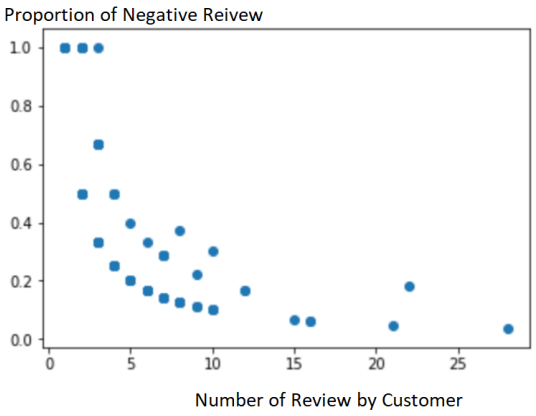
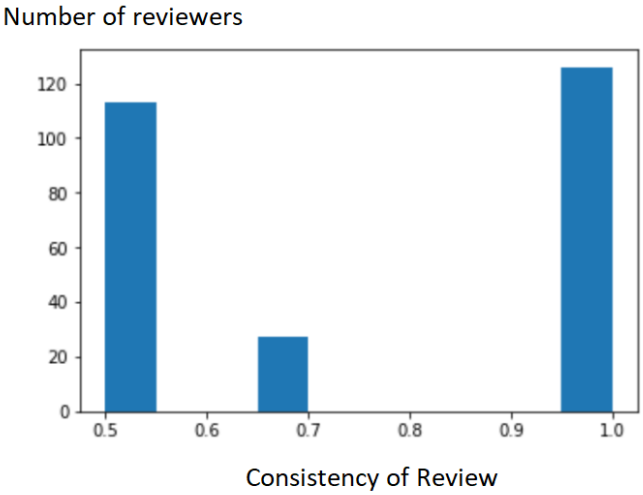


Figure 4.6.2



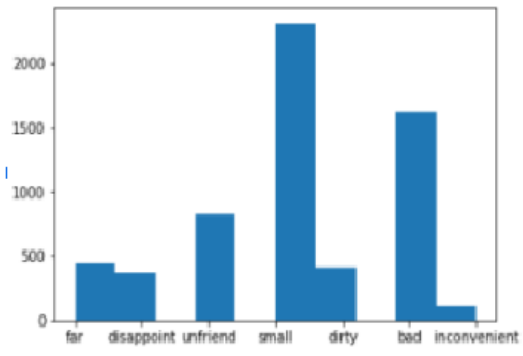
<Correlation between number of reviews left to negative reviews (Y-axis is the proportion of negative reviews)>

Figure 4.7.1

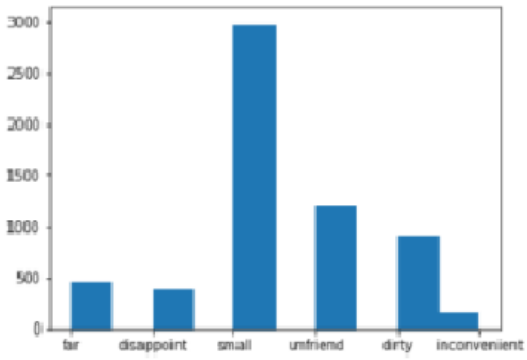


<Consistency in negative reviews by reviewers>

Figure 4.7.2



<Origin of negative sentiment histogram



<Labels without ‘bad’>

Figure 4.7.2

	region_id	proportion	total-reviews	negative-BOW	bad	small	dirty	unfriend	far	inconvenient	disappoint	sensitive
0	2-Central & Western	0.021397	9581	{'far': 26, 'obviously': 4, 'sorry': 4, 'bad':...	0.305428	0.325296	0.291672	0.298382	0.255485	0.276817	0.238910	small
1	-1-Islands	0.021076	8920	{'smaller': 8, 'old': 106, 'however': 24, 'sma...	0.334635	0.327470	0.318186	0.319522	0.273459	0.306585	0.261816	bad
2	5-Central & Western	0.033363	11270	{'however': 53, 'although': 37, 'small': 181, ...	0.228045	0.258640	0.238195	0.252637	0.193155	0.240670	0.193677	small
3	0-Central & Western	0.029519	4370	{'old': 92, 'far': 8, 'dirty': 3, 'small': 75,...	0.317125	0.357576	0.308600	0.315266	0.254342	0.293939	0.251805	small
4	2-Yau Tsim Mong	0.054603	9212	{'old': 258, 'glad': 2, 'small': 350, 'noisy':...	0.211777	0.254372	0.203838	0.209311	0.162996	0.195310	0.158303	small
5	-1-Yuen Long	0.022133	497	{'far': 5, 'bad': 1, 'old': 4, 'small': 1, 'al...	0.728397	0.519595	0.628389	0.600790	0.831397	0.577741	0.559122	far
6	2-Wan Chai	0.033208	9275	{'although': 46, 'noisy': 127, 'small': 204, '...	0.250190	0.298700	0.249577	0.263707	0.201203	0.249258	0.208850	small
7	1-Central & Western	0.010980	1275	{'bad': 2, 'tough': 1, 'noisy': 6, 'tiny': 6, ...	0.554768	0.617330	0.523953	0.560535	0.442066	0.524769	0.425774	small
8	3-Wan Chai	0.033580	6343	{'smaller': 11, 'poor': 7, 'inconvenient': 4, ...	0.275926	0.319103	0.278711	0.286938	0.228258	0.274920	0.220965	small
9	-1-Southern	0.014525	895	{'noisy': 2, 'old': 4, 'bad': 2, 'small': 6, '...	0.640466	0.725530	0.612829	0.619491	0.621765	0.587321	0.563029	small
10	5-Wan Chai	0.048143	2908	{'small': 90, 'bad': 15, 'although': 20, 'howe...	0.308355	0.336326	0.300739	0.303551	0.236533	0.280998	0.234419	small
11	4-Wan Chai	0.018360	1634	{'small': 11, 'unfortunate': 1, 'pleased': 1, ...	0.572282	0.552653	0.576198	0.554907	0.489553	0.529844	0.478621	dirty
12	4-Yau Tsim Mong	0.050848	17621	{'small': 614, 'old': 278, 'noisy': 173, 'bad'...	0.179234	0.218550	0.171151	0.176891	0.135016	0.165578	0.132719	small
13	0-Wan Chai	0.045461	7413	{'grungy': 2, 'tiny': 31, 'dirty': 11, 'howeve...	0.235226	0.299954	0.237758	0.248469	0.188876	0.234395	0.188341	small
14	-1-Eastern	0.028395	2430	{'old': 34, 'satisfied': 3, 'however': 19, 'sa...	0.325246	0.433918	0.312011	0.326884	0.303214	0.308301	0.285901	small
15	0-Yau Tsim Mong	0.045082	8296	{'terrible': 11, 'noisy': 102, 'bad': 52, 'old...	0.231866	0.273213	0.218514	0.230748	0.180960	0.215618	0.177341	small

Fig. 4.8.1

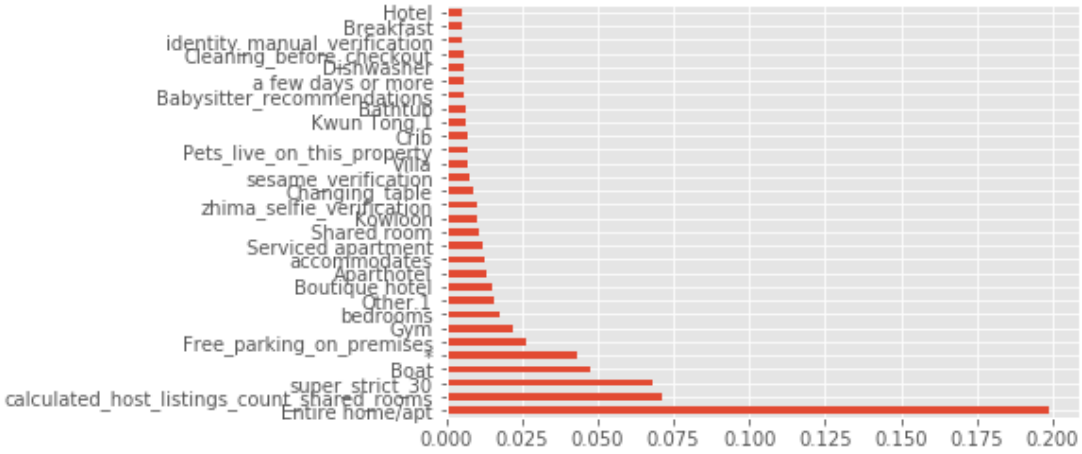


Fig. 4.9.1

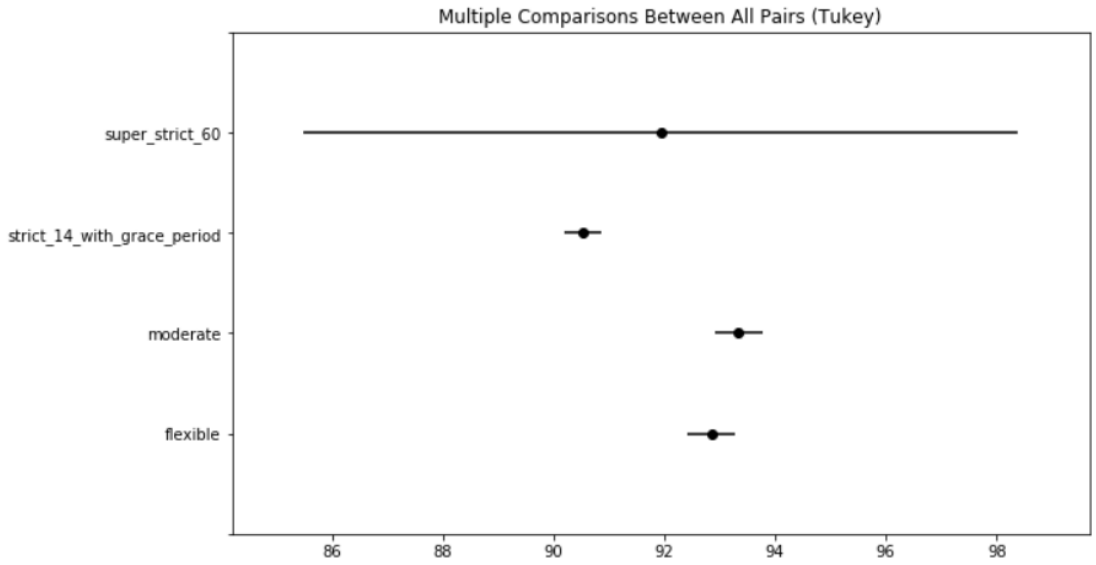


Fig. 5.1

Airbnb Web search page

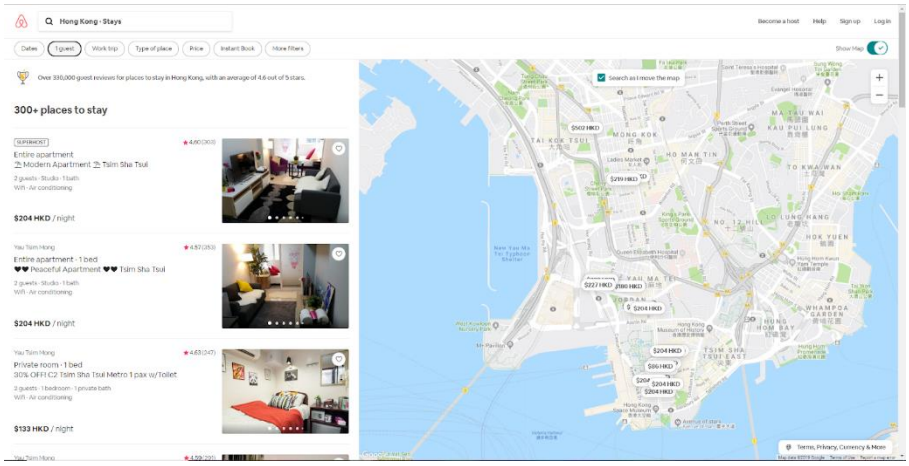
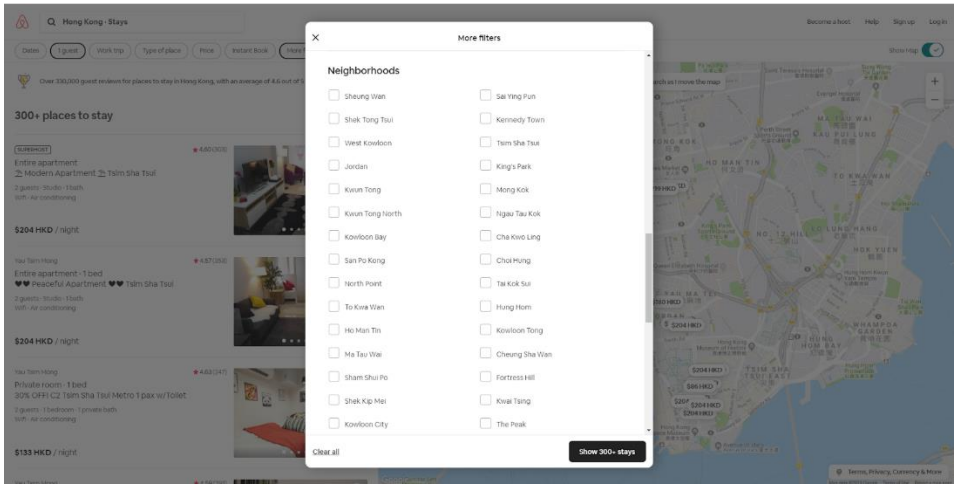
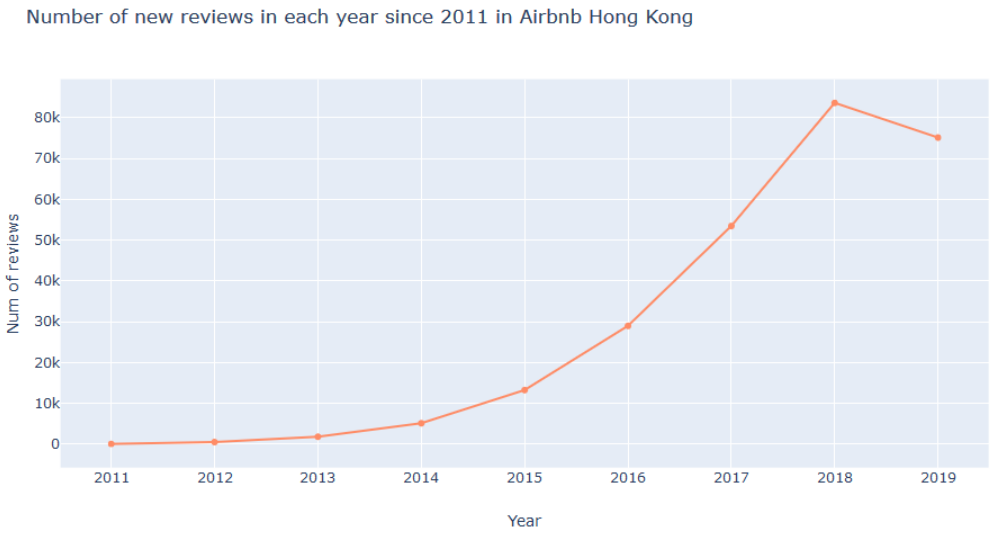


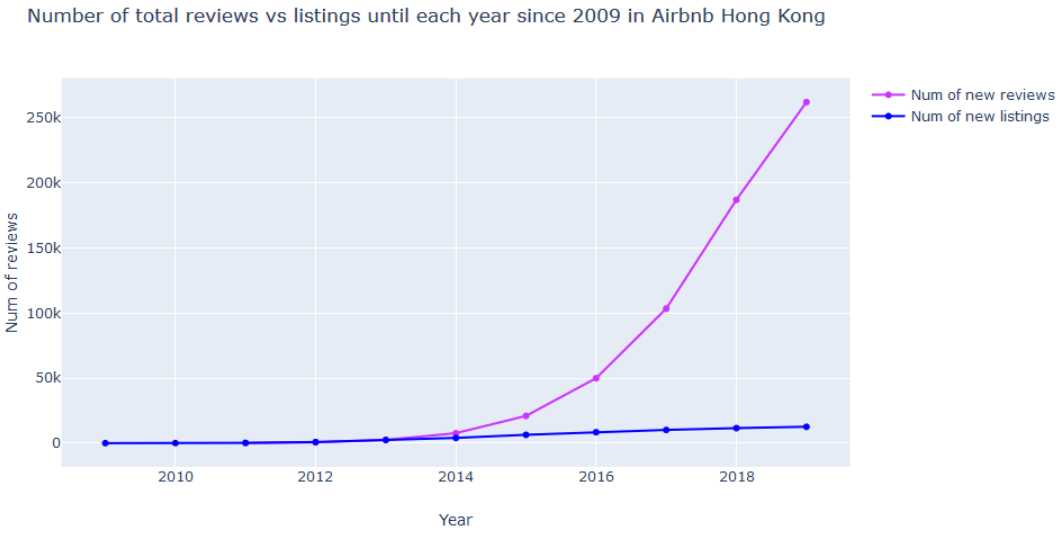
Fig. 5.2 Airbnb Web search Neighbourhood filters



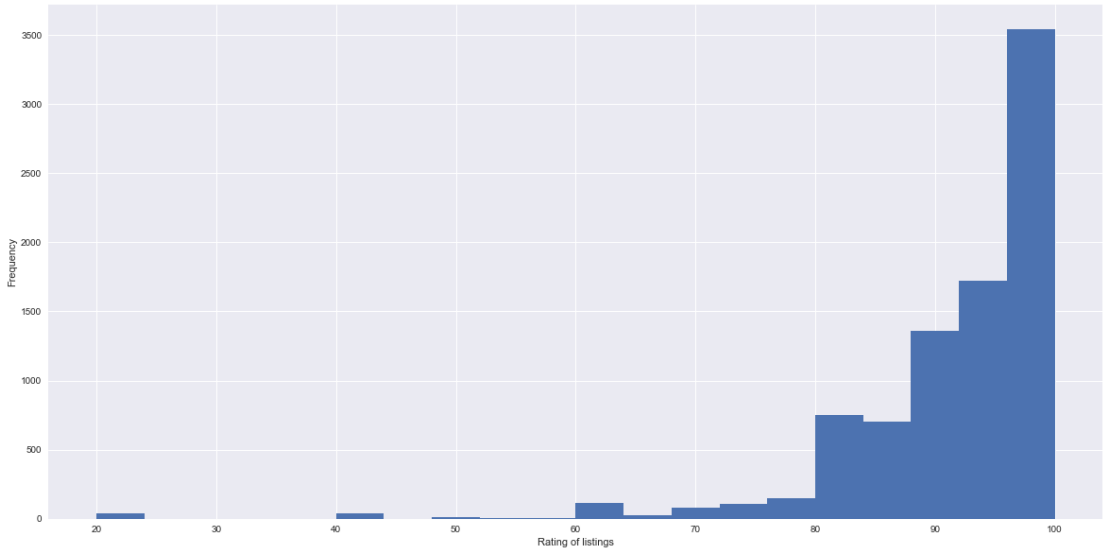
<Origin of negative sentiment based on different neighborhood>

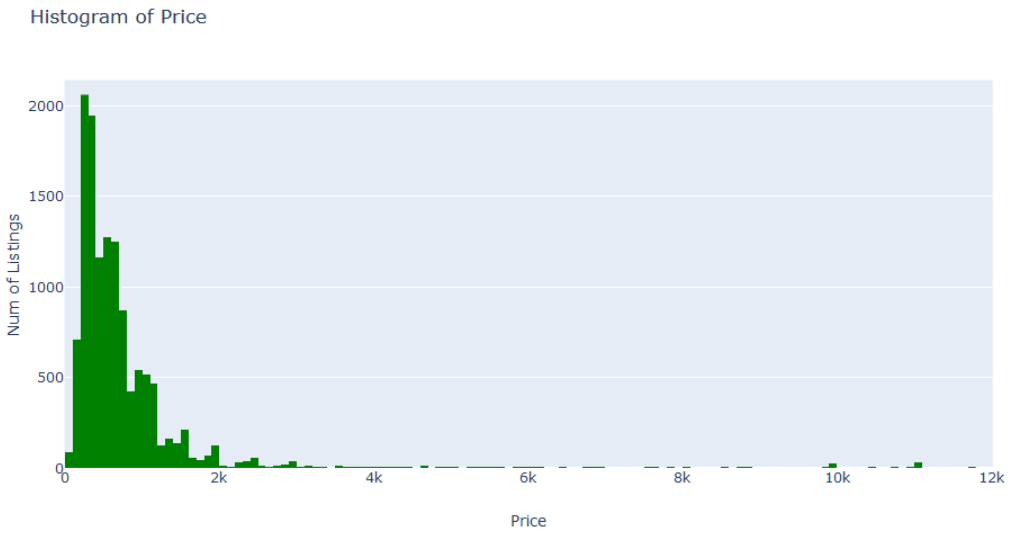
	date_year	count	year_sum
0	2011	91	91
1	2012	552	643
2	2013	1855	2498
3	2014	5160	7658
4	2015	13287	20945
5	2016	29013	49958
6	2017	53471	103429
7	2018	83631	187060
8	2019	75132	262192



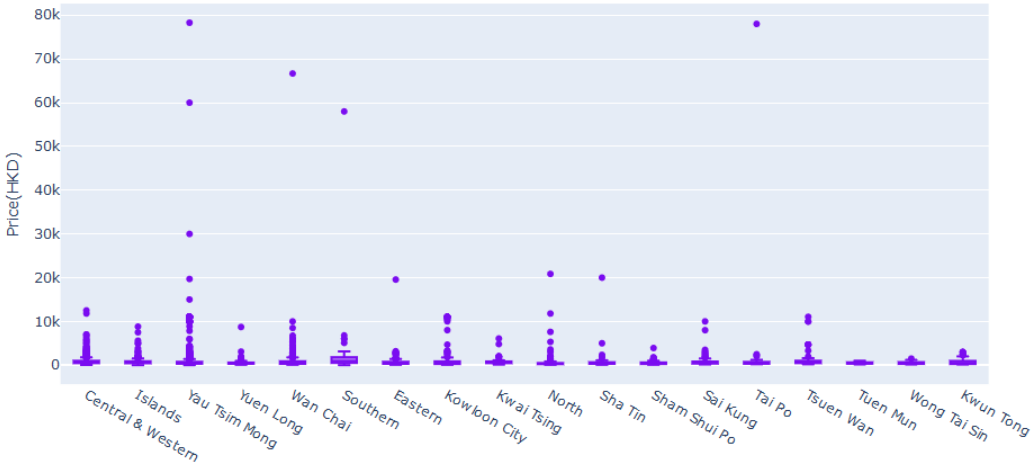


highly skewed distribution of reviews, most reviews towards listings are positives. Extremely few negative reviews.

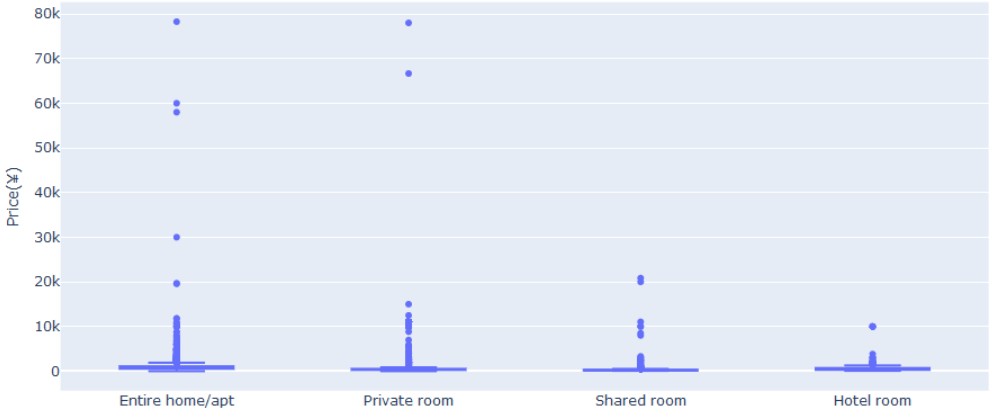

Frequency
Rating of listings
0
500
1000
1500
2000
2500
3000
3500
20
30
40
50
60
70
80
90
100

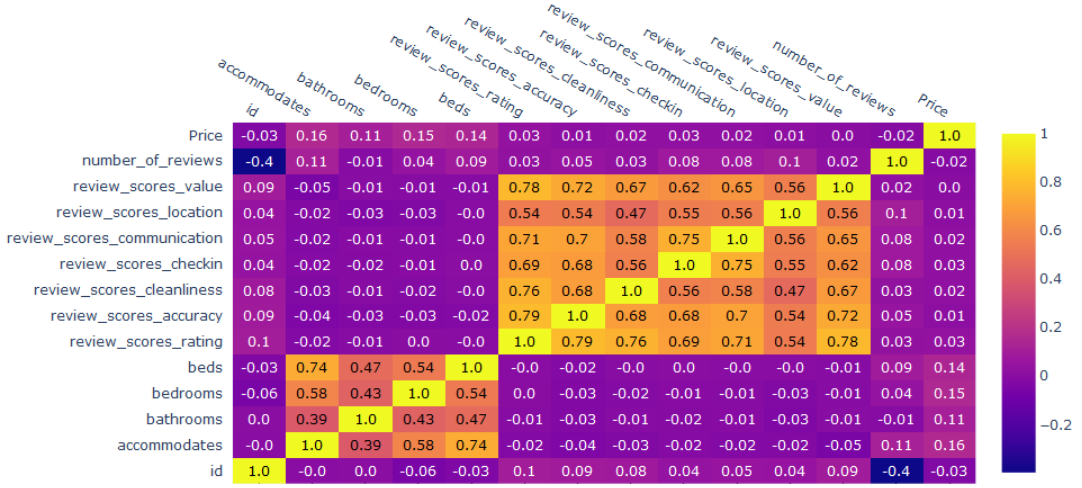

Histogram of Price
Num of Listings
Price
0
500
1000
1500
2000
0
2k
4k
6k
8k
10k
12k

Box chart of room price by district in Airbnb Hong Kong



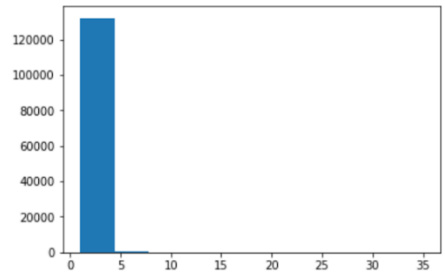
Box chart of room price by room type in Airbnb Hong Kong





Between price and other variables, we have not seen some significant ones.

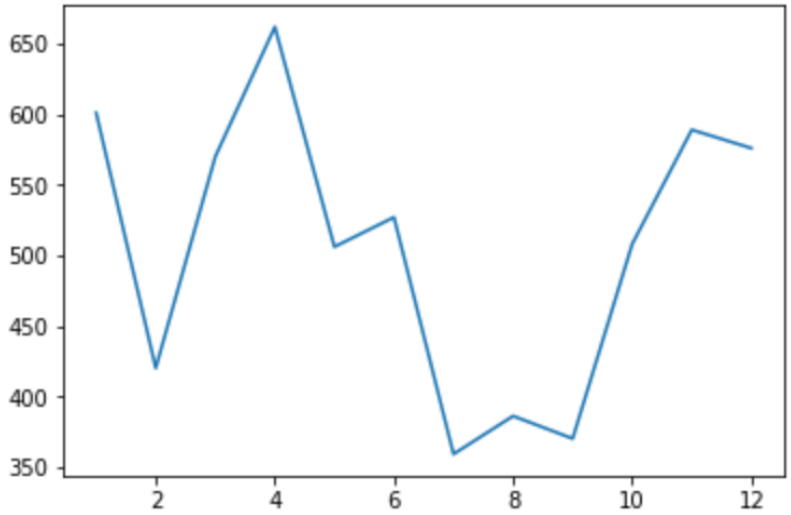
```
Total review number count Counter({1: 120655, 2: 9179, 3: 1722, 4: 557, 5: 212, 6: 114, 7: 59, 8: 32, 9: 22, 13: 10, 10: 10,
12: 9, 14: 7, 11: 5, 16: 4, 21: 2, 18: 2, 35: 1, 34: 1, 28: 1, 22: 1, 17: 1, 15: 1})

Negative review number count Counter({1: 5933, 2: 64, 3: 3, 4: 1})
```



<Histogram of number of reviews per individual user (most of the people leave 1 review)>

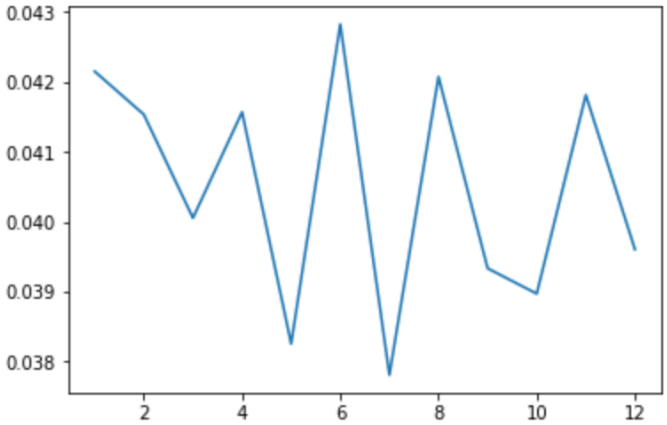
```
Counter({1.0: 5130, 0.5: 586, 0.3333333333333333: 138, 0.25: 62, 0.2: 22, 0.16666666666666666: 14, 0.6666666666666666: 11,
0.125: 8, 0.14285714285714285: 8, 0.2857142857142857: 5, 0.1: 4, 0.1111111111111111: 3, 0.0625: 2, 0.18181818181818182: 1,
0.375: 1, 0.3: 1, 0.2222222222222222: 1, 0.4: 1, 0.03571428571428571: 1, 0.047619047619047616: 1, 0.06666666666666667: 1})
```

<Distribution of the proportion of negative reviews, for example 1.0: 5130 implies there are 5130 users where all their reviews are negatively sentimented>



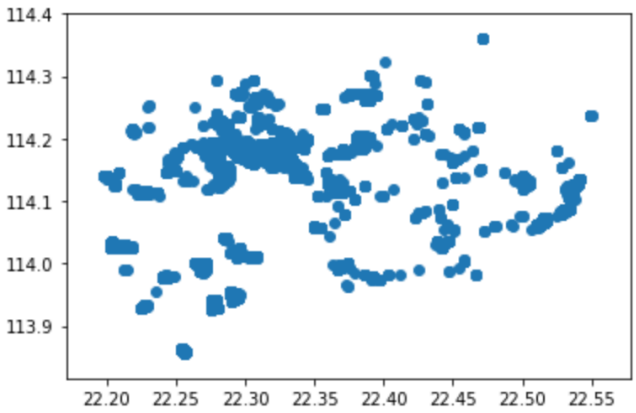
<number of negative review based on time (month)>

```
Out[137]: [<matplotlib.lines.Line2D at 0x2733b7482e8>]
```



<proportion of negative review per month / for example on June around 4.3 percent of reviews all the reviews were negatively sentimental>

```
Out[197]: <matplotlib.collections.PathCollection at 0x27233dc5630>
```

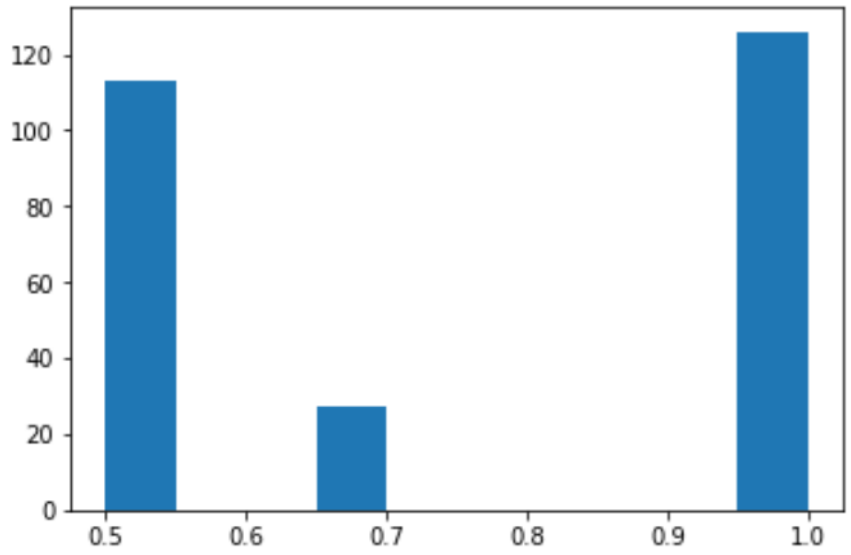


<distribution of negative sentiment listings based on geographical location>

	region_id	proportion	total-reviews	negative-BOW	bad	small	dirty	unfriend	far	inconvenient	disappoint	sensitive
0	2-Central & Western	0.021397	9581	{'far': 26, 'obviously': 4, 'sorry': 4, 'bad':...	0.305428	0.325296	0.291672	0.298382	0.255485	0.276817	0.238910	small
1	-1-Islands	0.021076	8920	{'smaller': 8, 'old': 106, 'however': 24, 'sma...	0.334635	0.327470	0.318186	0.319522	0.273459	0.306585	0.261816	bad
2	5-Central & Western	0.033363	11270	{'however': 53, 'although': 37, 'small': 181, ...	0.228045	0.258640	0.238195	0.252637	0.193155	0.240670	0.193677	small
3	0-Central & Western	0.029519	4370	{'old': 92, 'far': 8, 'dirty': 3, 'small': 75,...	0.317125	0.357576	0.308600	0.315266	0.254342	0.293939	0.251805	small
4	2-Yau Tsim Mong	0.054603	9212	{'old': 258, 'glad': 2, 'small': 350, 'noisy':...	0.211777	0.254372	0.203838	0.209311	0.162996	0.195310	0.158303	small
5	-1-Yuen Long	0.022133	497	{'far': 5, 'bad': 1, 'old': 4, 'small': 1, 'al...	0.728397	0.519595	0.628389	0.600790	0.831397	0.577741	0.559122	far
6	2-Wan Chai	0.033208	9275	{'although': 46, 'noisy': 127, 'small': 204, '...	0.250190	0.298700	0.249577	0.263707	0.201203	0.249258	0.208850	small
7	1-Central & Western	0.010980	1275	{'bad': 2, 'tough': 1, 'noisy': 6, 'tiny': 6, ...	0.554768	0.617330	0.523953	0.560535	0.442066	0.524769	0.425774	small
8	3-Wan Chai	0.033580	6343	{'smaller': 11, 'poor': 7, 'inconvenient': 4, ...	0.275926	0.319103	0.278711	0.286938	0.228258	0.274920	0.220965	small
9	-1-Southern	0.014525	895	{'noisy': 2, 'old': 4, 'bad': 2, 'small': 6, '...	0.640466	0.725530	0.612829	0.619491	0.621765	0.587321	0.563029	small
10	5-Wan Chai	0.048143	2908	{'small': 90, 'bad': 15, 'although': 20, 'howe...	0.308355	0.336326	0.300739	0.303551	0.236533	0.280998	0.234419	small
11	4-Wan Chai	0.018360	1634	{'small': 11, 'unfortunate': 1, 'pleased': 1, ...	0.572282	0.552653	0.576198	0.554907	0.489553	0.529844	0.478621	dirty
12	4-Yau Tsim Mong	0.050848	17621	{'small': 614, 'old': 278, 'noisy': 173, 'bad'...	0.179234	0.218550	0.171151	0.176891	0.135016	0.165578	0.132719	small
13	0-Wan Chai	0.045461	7413	{'grungy': 2, 'tiny': 31, 'dirty': 11, 'howeve...	0.235226	0.299954	0.237758	0.248469	0.188876	0.234395	0.188341	small
14	-1-Eastern	0.028395	2430	{'old': 34, 'satisfied': 3, 'however': 19, 'sa...	0.325246	0.433918	0.312011	0.326884	0.303214	0.308301	0.285901	small
15	0-Yau Tsim Mong	0.045082	8296	{'terrible': 11, 'noisy': 102, 'bad': 52, 'old...	0.231866	0.273213	0.218514	0.230748	0.180960	0.215618	0.177341	small

<origin of negative sentiment analysis according to different neighborhood>
Note that neighbours like Wan Chai and few other regions are further divided into 6 sub-neighbourhood. Hence 4-Wan Chai, 3- Wan Chai represent different area.

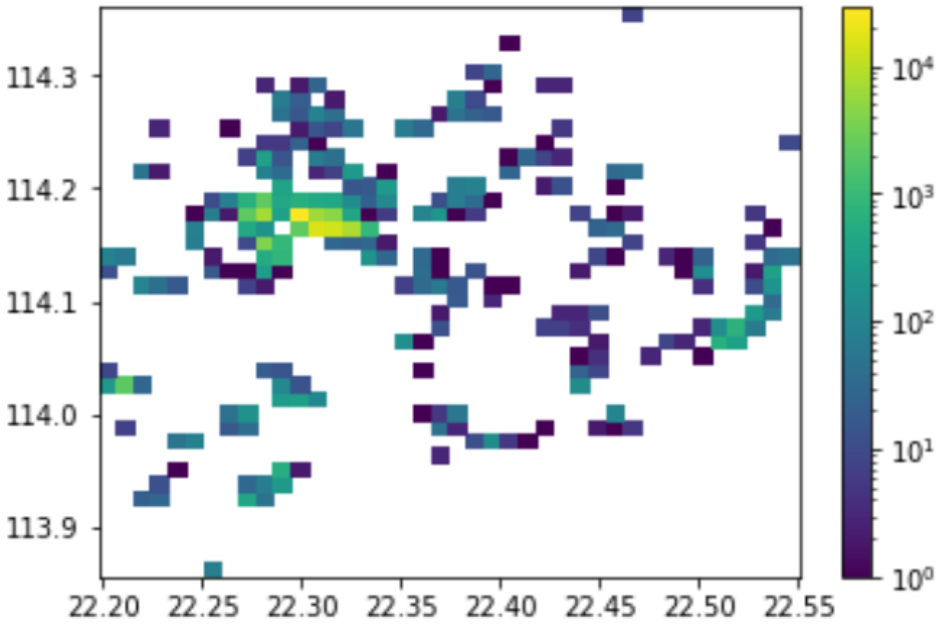
Seems like label ‘bad’ and ‘dirty’ are closely related.



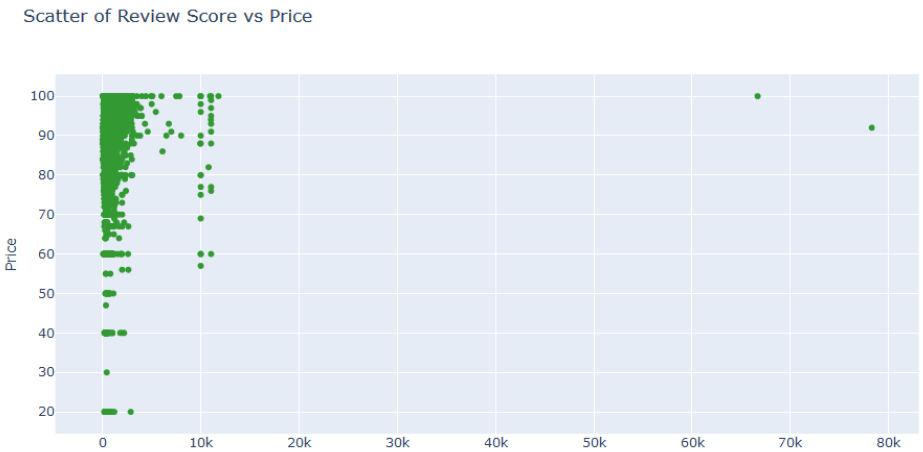
<consistency in the origin of negative review>

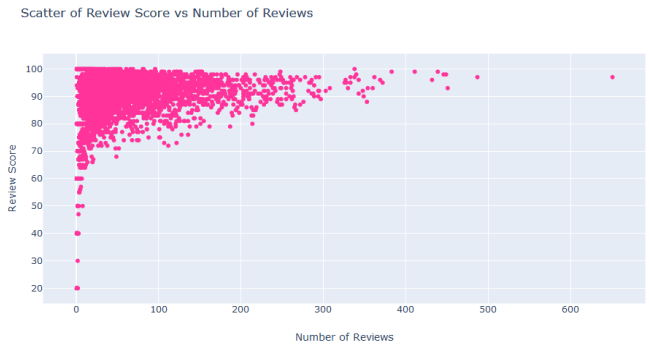
Here origin is one of the following category : dirty,small,unfriendly, far,inconvenient, disappoint

And consistency implies the proportion of the highest category for each user. For example if one person has 4 negative review and 2 has origin of dirty and 1 for unfriendly and 1 for far; then the consistency is 0.5 (2/4)



<heatmap of negative review based on location>




Scatter of Review Score vs Number of Reviews
Review Score
100
90
80
70
60
50
40
30
20
0
100
200
300
400
500
600
Number of Reviews