Team name: Dataholic

CSL Data Analytics Case Competition **| Airbnb**

Tony Tang | Justin Chiu | Chanho Park | Jacob Chiu

Oct 2019, Hong Kong

Executive Summary

Company Overview

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.

Data Preparation & Data Understanding

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.

Analysis Findings

#1 Basic Information
#2 Host information
#3 Property Types
#4 New Listings over Time
#5 Demand and Price Findings
#6 Accuracy of Review
#7 Negative Review
#8 Feature Importance table for price forecasting model

Recommendations

#1 Geographical Distribution of Listings
#2 Monetary Incentive for Reviewers
#3 Personalised Rankings on Search Page
#4 Disclaimer for Guests Determining if Listing is Overpriced or underpriced

Company Overview

Airbnb, Inc. is one of the world's largest marketplaces for unique, authentic places to stay and things to do, offering over ***7 million accommodations*** and ***40,000 handcrafted activities***, *all powered by local hosts*. With more than half a billion guest arrivals to date, and accessible in 62 languages across 191 countries and regions, Airbnb promotes people-to-people connection, community and trust around the world.

Company Background

Founded in 2008 and **Headquartered** in San Francisco, California, United States
Total number of Employees: 26,969

Comparative advantages:

Low overhead: The company doesn't have to worry about the high turnover rates of bellhops and front desk clerks like hotels do.

Lower prices: The Airbnb price point is especially advantageous for targeting budget-conscious millennials.

Convenience: Airbnb successfully delivers solutions for a constrained travel budget, saving time or a preference for a given location.

Income for locals: More and more of us welcome the opportunity for extra income as home owners to rent out space for a fee. Airbnb offers that.

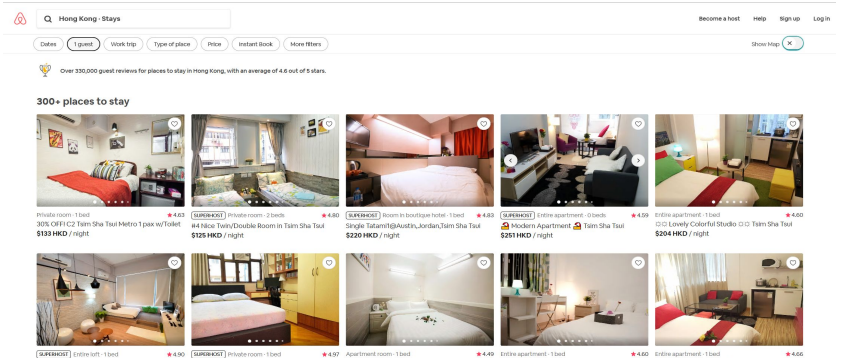
Data Sources

Web Scraping

- Python Libraries of Scrapy and JSON were used to scrape information provided on the Airbnb website.

Scrapy

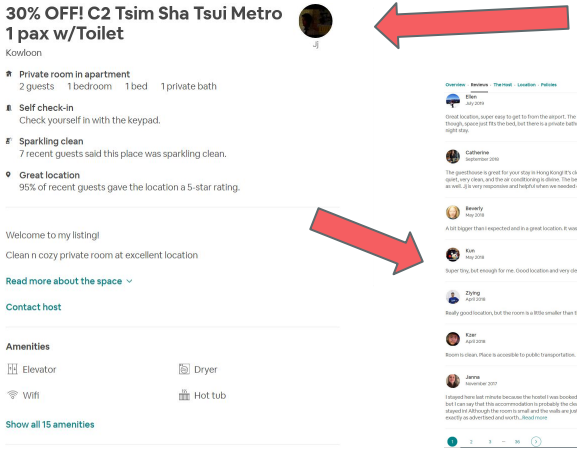
- Airbnb has a search page API that contains all basic information of listings currently displayed on the page
- Information included: Listing ID, Latitude and Longitude, Price, etc
- The Infinity Scroll structure infers that the Website is fetching data from a new API everytime it reaches the bottom
- A for loop is used to iterate through all the search page APIs



- Details for each listing is only available on the individual web page of each listing; which is stored in another API

-The API is identifiable by the individual listing ID

- Information such as the overview by the host, amenities provided, ratings provided by guests, etc



-The reviews of each listing are again stored into another API, identifiable by listing ID

- By changing the 'limit' component and 'offset' component to 1000 and 0 respectively, all the reviews will be displayed on the API accordingly

- Information such as the review comment, review rating, review date, reviewer ID is obtainable

Online Database: Inside Airbnb

Historical data of listings in Hong Kong can be obtained from http://insideairbnb.com/get-the-data.html

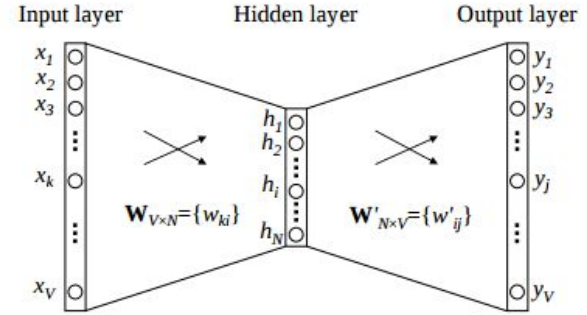
Data Understanding

	Review Dataset (300,000 x 19)	Listing Dataset (12,000 x 13)
Original Features	listing_id id date reviewer_id reviewer_name comments	host_id neighbourhood_group latitude, longitude room_type, minimum_nights, number_of_reviews, reviews_per_month, availability_365
Added Features	english keyword negative Words positive Words Sentiment Neighbourhood bad, small, dirty, unfriend, far, inconvenient, disappoint,	region_id proportion Total number of reviews negative bag of words

Methodology - Sentiment Analysis

Technique and Model

- Word embedding to extract negative/positive keywords
- Pre-trained word2vec model by Google



Model Description

- Assumption of word2vec model is that when given a sentence, a particular word is correlated with the neighboring words. Training data is constructed by pairing a word with its neighboring word.
- When given input word, model predicts the next word
- When input word is given the hidden layer state is what we refer as ‘word embedding’; these states implies the information about the correlation with other words.
- For our analysis, we used pre-trained word2vec model by Google trained on wikipedia corpus.

Q: Why this technique instead of other alternatives?

A: Most of the reviews were positively skewed; only around **5%** of people leave negative reviews on Airbnb

Example of this Technique:

“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”

Positive: ['convenient', 'helpful'],
Negative: ['small', ‘noisy’,'however']

Processing

1. Compare the number of positive vs negative keywords to determine whether review is positive or negative
2. For negative reviews: look into the average embeddings of the keywords and find the cosine similarity between categories of the negative keywords - eg. ‘far’, ‘inconvenient’, ‘unfriendly’ and ‘dirty’
3. Based on the cosine similarity, further label the negative sentiment.

Methodology - Price Modeling

Objective of Price Modeling

- Determine if listings are overpriced or underpriced

Examples of Factors Influencing Price of a Listing

- Location
- Number of Rooms
- Whether it is the Entire Home/Apartment
- Different Amenities and Facilities

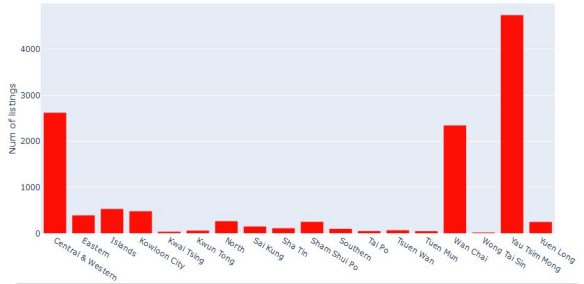
- Variation in price is noise
- This model is used to identify a pattern and ignore the noise

Further Processing

- Data is split into training and test sets
- Applied forward selection process: select one, feature, running the model, checking performance, and repeating the test
- 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

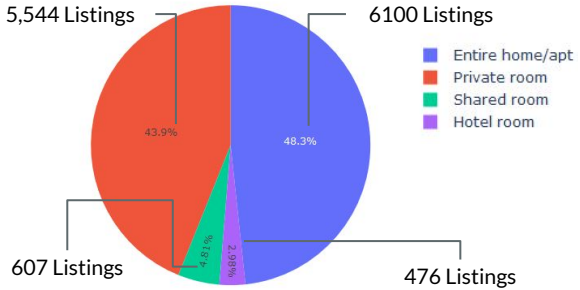
Analysis - Basic Findings

Listings by Room Types in Districts



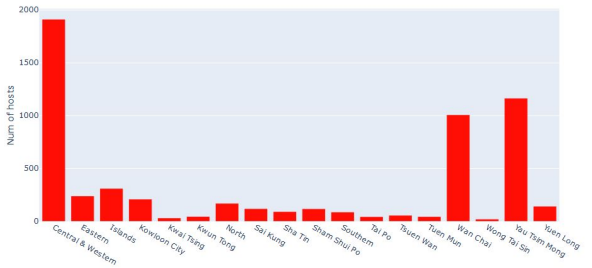
Key Takeaways: 9,712 or 76.91% out of 12,627 total listings in Yau Tsim Wong (4,740 or 37.54%), Central & Western (2,624 or 20.78%), and Wan Chai (2,348 or 18.6%)

Property Types Distribution



Key Takeaways: Entire Rooms/Apartments and Private Rooms take up 92.2% of 12,627 properties listed on Airbnb.

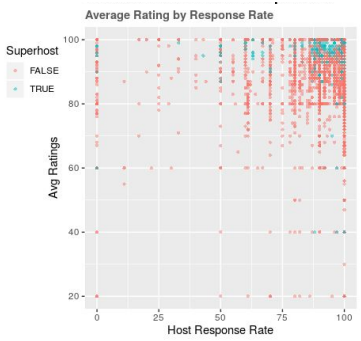
Hosts by district



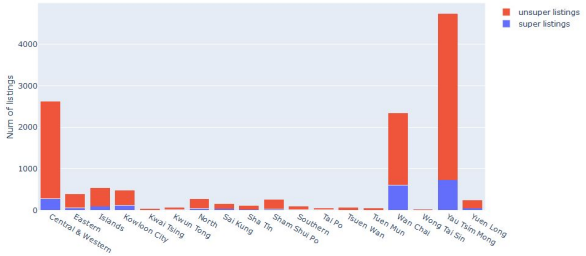
Key Takeaways: The total number of listing hosts amount to 5,490, with 4,117 (74.99) being single listing hosts and 1,373 (25.01%) with multiple listings.

Superhost Ratings and Response Rate

Key Takeaways: There are existing 'Superhosts' with less than 75% response rate and 80% Average Rating; violating Airbnb's criteria of 90% rate criteria and 4.8 Overall Rating (96% Average Rating)



Distribution of Super Listings by District



Key Takeaways: There are a total of 692 'Superhosts' up til today in Hong Kong with 2174 listings; inferring a total of 17.2% listings are 'Super' Listings whilst 12.6% of hosts are 'Superhosts'

New and Total Listings Over Time

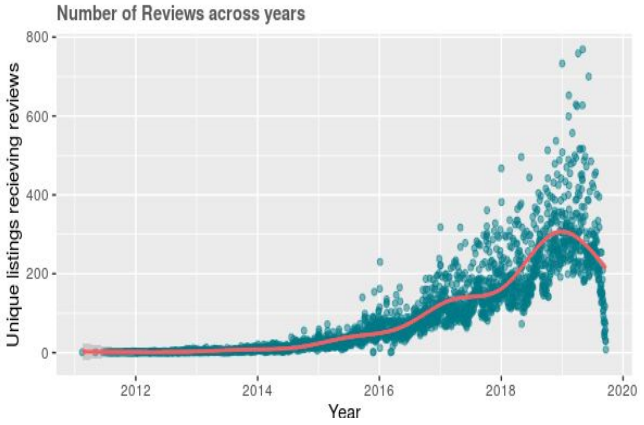
Key Takeaways: Between 2010 - 2015, the number of total listings were and new listings almost increased at an exponential rate and linear rate respectively. The growth in total listings slowed down since 2015

Analysis - Demand and Pricing

Estimation of Demand for listings in Hong Kong

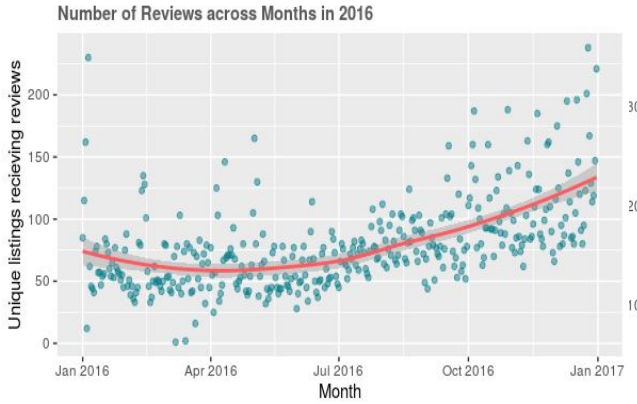
Information on the number of bookings made are unavailable on Airbnb due to Privacy Reasons. According to Airbnb, roughly 50% of guests leave reviews on their hosts/listings. We will be estimating the demand of listings with the number of reviews.

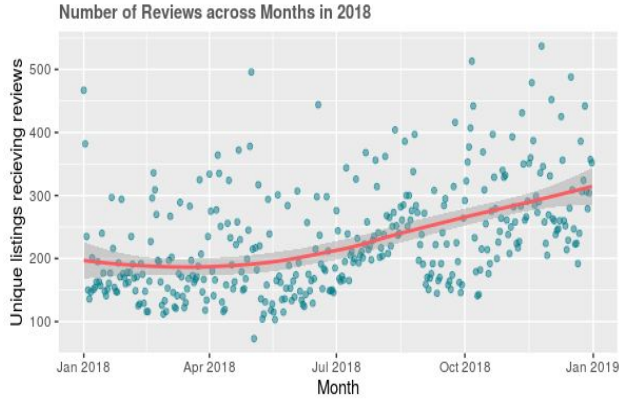
Number of Reviews Across Time



Key Takeaways: Number of unique listings receiving reviews have almost increased exponentially over the years. However, We can observe a sharp decrease in demand starting from mid 2019. This phenomenon could be attributed to the recent events in Hong Kong, which has seen a huge decrease in number of tourists coming to Hong Kong as well as a historical low point of hotel room rental prices.

Seasonality in Demand Across Time - 2016, 2017, 2018





The dots on each graph represent the number of reviews for an individual listing written on a particular month, reflecting the demand in that month.

Key Takeaways: The graphs signify a seasonal pattern in demand for listings in Hong Kong. There are peaks and drops in demand in similar months across three years. The peak period starts in July and ends in January every year.

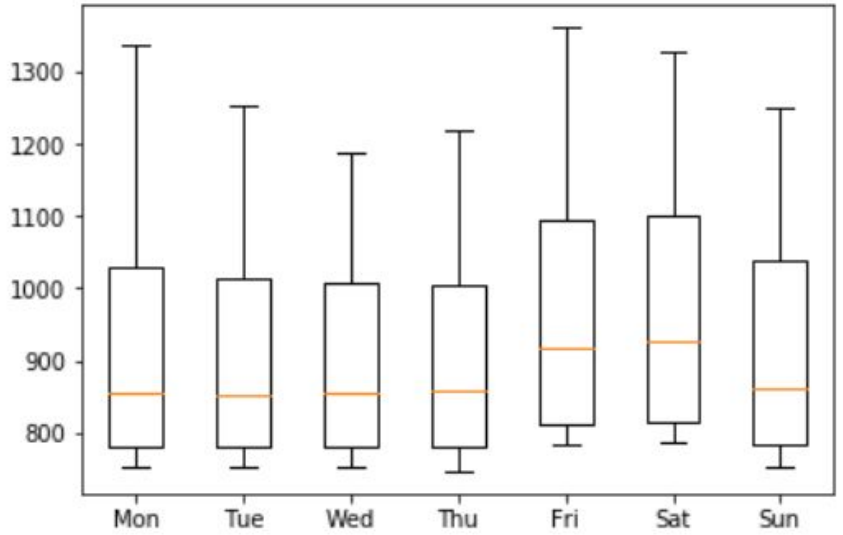
Analysis - Demand and Pricing

Seasonality in Price

Key Takeaways: 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.

Listing Prices in day of the week

Two sets of points can be observed from the graph of seasonality in price, which indicates that average prices on certain days were higher than other days. A box plot of average prices on days of the week will be used to further examine this phenomenon.



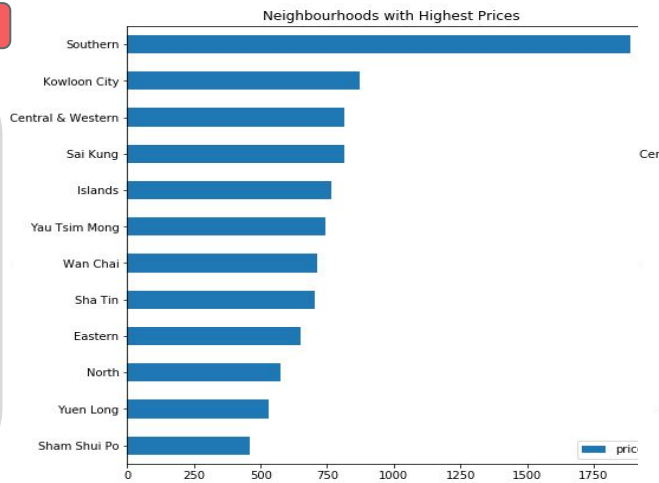
Key Takeaways: Average prices on Fridays and Saturdays are comparatively more expensive than that of other days of the week. This could be explained by the higher demand of accommodation of guests on weekends.

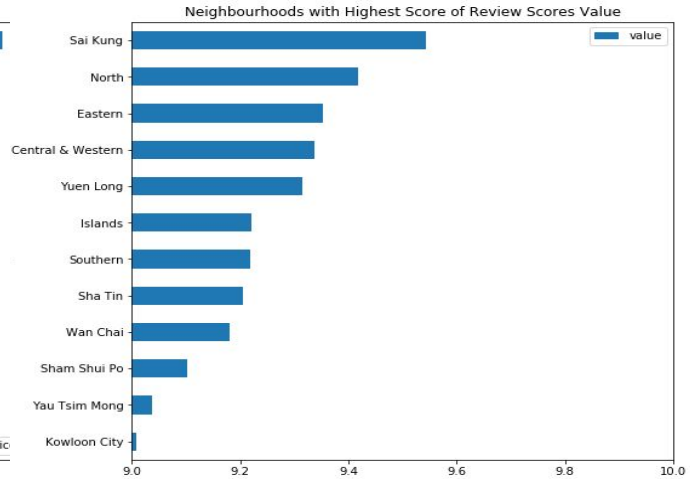
Analysis - Indicators of Price

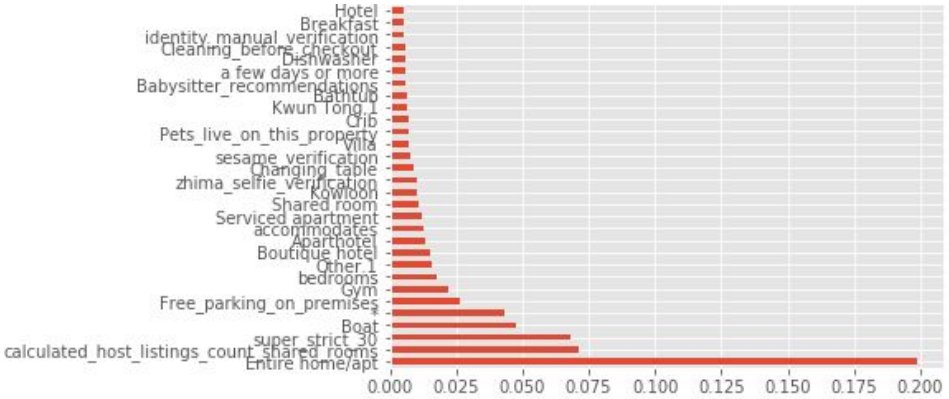
Prices and Value By Neighbourhood

'Value' is the pay off between the cost paid and the benefit received from the experience in the listing, rated out of 10 by guests of the listing.

Key Takeaways:There is no visible pattern between price and value. Listings in Southern and Kowloon district have the highest average price, but rank 7th and and 12th for value; suggesting an inverse relationship. However, Central & Western and Sai Kung are both ranked within the top 4 for price as well as value, which suggests otherwise. The correlation between the two factors is ambiguous.







Feature Importance for Price Forecasting

The price forecasting model interprets the historical data to investigate which features have a higher influence on the predicted values of price. This figure displays the importance of the top 30 features of a listing towards its price.

Key Takeaways: Whether or not a property is an Entire Home/Apartment has the largest weighting and is the dominating factor toward price by a huge margin. Features like number of rooms, free parking, gym, and number of people accommodated have a relatively higher indicative factor to price.

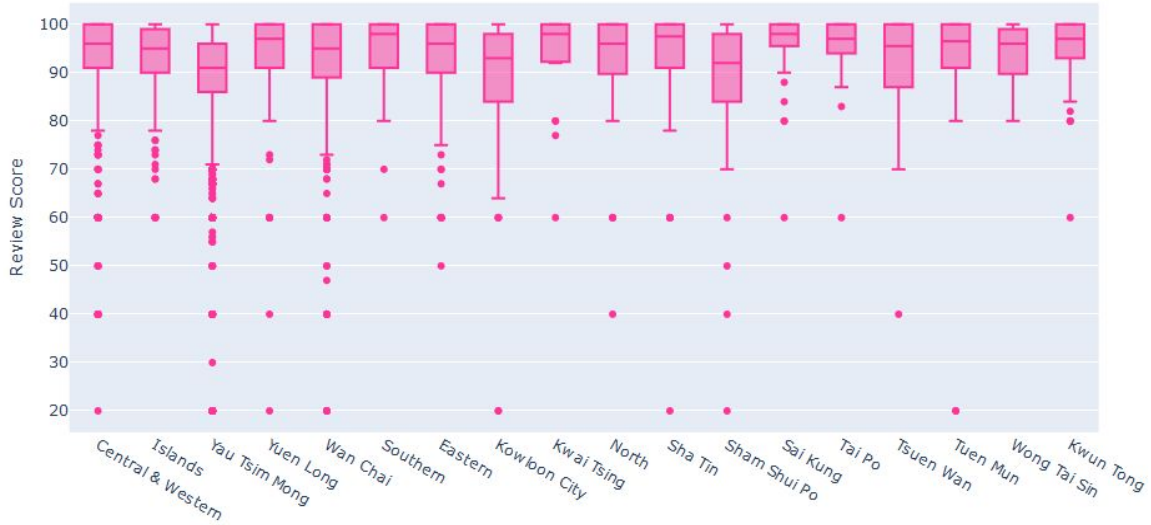
The variable 'Super strict 30' also has a high proportion, referring to the cancellation policy set by the host that sets the booking to be non-refundable upon cancellation within 30 days prior to arranged date.

Analysis - Distribution and Accuracy of Reviews

Section Objective

The accuracy of reviews are crucial for the judgement of potential guests of listings. This section will examine this aspect to improve customer experience and avoid disappointment from the experience.

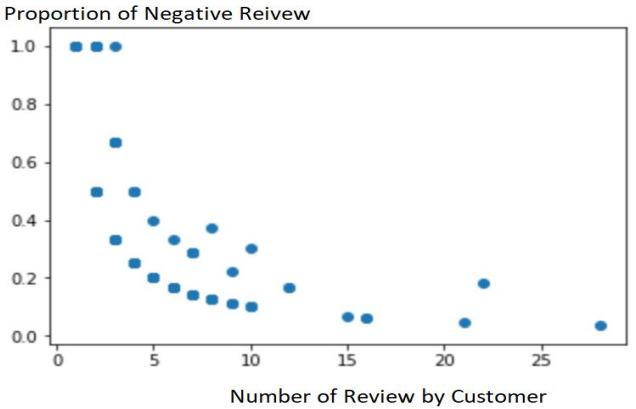
Review Score by District



Key Takeaways: The Review scores by district are all skewed in the high end, mostly scoring between the ranger of 85 to 100. Districts that have the smallest spread tend to be unpopular districts (Sai Kung, Tai Po, Kwun Tong) as compared to the three main districts (Central & Western, Wan Chai, Yau Tsim Wong). Although reviews for the main districts are also concentrated in the high end, there are more outliers with low scores compared to other districts. Our results indicate that roughly 5% out of all reviews are negative, which explains the skewness. .

Negative Reivews by Guests

The graph below exhibits the correlation between number of reviews left (X-Axis) and the proportion of negative sentiment reviews amongst their reviews (Y-axis)



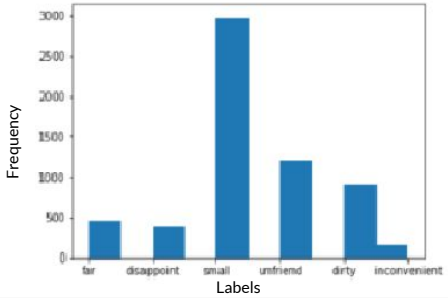
Key Takeaways: An inverse proportion of relationship can be observed between the two variables. This infers that users who have left a higher number of reviews are more likely to leave a positive review.

Guests who had a negative experience tend not to leave comments and reviews, therefore influencing the capacity to assess the overall quality of the listing.

Along with findings for the review scores by district, we can conclude that it is seemingly difficult for customers to make bookings based on reviews given by past users.

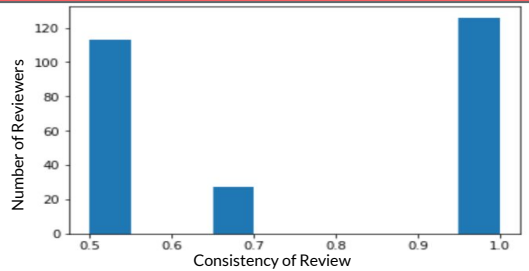
Analysis - Negative Sentiment Reviews

Origins of Negative Sentiment Keyword



Key Takeaways: The keywords or labels identified from eviews in Hong Kong indicate that most guests found listings to be small, along with other problems such as unfriendly hosts or the listing being dirty/ far/ disappointing/inconvenient. The word ‘Bad’ was removed for its generic meaning.

Consistency in Negative Reviews by Reviewers



Key Takeaways: Most people are consistent with negative reviews, signifying that individual guests have particular priorities they look into when staying at a listing. For instance, one guest complaining about dirtiness in one listing is likely to comment the same about another listing.

Labels of Negative Sentiment based on Neighbourhood

A geographical analysis on negative keywords is done via counting such words based on each neighbourhood. For a more detailed analysis, the most crowded districts (Yau Tsim Wong, Wan Chai, Central & Western) are divided into 6 neighbourhoods. From the diagram, each label column (‘bad;. ‘small’ , etc) represent the measure of closeness to the keywords, and the ‘sensitive’ column shows the label of the highest column.

	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

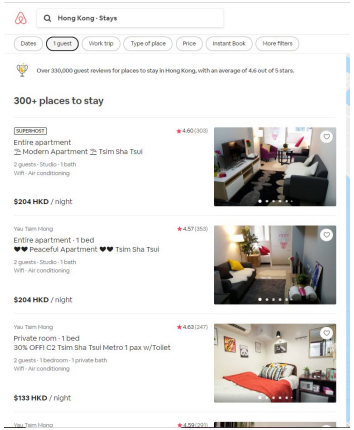
Key Takeaways: The results show that people found the small property size in Hong Kong to be the most common problem among different neighbourhoods, coinciding with the negative keyword histogram. Results like ‘far’ for Yuen Long or ‘small’ for Wan Chai tell us indicative factors that would aid further categorization of listings. For instance, we can help recommend listings to travellers based on personal preferences similar to other travellers.

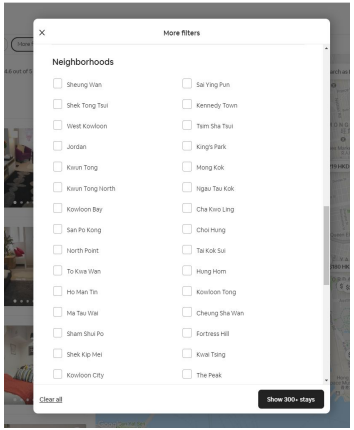
Conclusions

	Interpretation from Analysis
Types of Properties, Districts, Characteristics	There were a total of 12,627 properties listed on Airbnb Hong Kong with 5,490 hosts, most of them falling under the type of Entire Home/Apartment or Private Room. 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 less number of people.
Superhosts: Does It Matter?	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 the not all hosts with high ratings and response rates were superb hosts. Both regular hosts and superhosts have similar cancellation and booking policies. Results also indicate that 'Superhost' listings have higher ratings on average.
Correlation Between Demand and Price	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.
Demand and Pricing Fluctuations of Listings	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.
Indicators of Price	Surprisingly, the relationship between price and value is quite ambiguous with no observable pattern. The Feature Importance Price Forecasting Model indicates several elements of a property that have a comparatively larger weighting towards the listing price(e.g. Entire House/Apartment, Number of Rooms, Free Parking, Cancellation Policy), providing insight for hosts looking to list their property on Airbnb and potential guests looking for accomodation options.
Review Scores and Sentiment	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. Through the analysis of keywords, we infer that there is an individual and regional patterns in the reviews.

Recommendation #1: Geographical Distribution of Listings

Key Issue: Over 75% of listings are concentrated in three districts, including Yau Tsim Mong, Central & Western and Wan Chai. As such, customers are more than likely to select a location from one of three.





Airbnb's Problem: The default search page features a user interactable map with the listings and its price as well as filters that allow users to search for listings in specific neighbourhoods. However, it does not tell the users about how the listings are distributed.

Key Takeaways: It is suggested that Airbnb should not only display the most popular district when a user searches for listings in a city, but also ***additional information*** such as:

- *Number of listings currently listed on each district/ neighbourhood*
- *Average price/ rating*
- *Number of visits in the past year*

User Friendliness to First-time visitors

Increase Customer Understanding

Consumer Confidence in Airbnb

Recommendation #3: Personalised Rankings on Search Page

Key Issue: Customers could not have whole user experience, according to the filters appeared on the search page.

User Preferences Panel	
Cleanliness	Accessibility
Price	Ratings from past users
Ratings of the listings	Daily necessities provided
Amenities	Friendliness of the hosts
% of Negative Reviews	Non-Red Badge

Users will be asked to select 3 most important criteria when considering renting the accommodation.

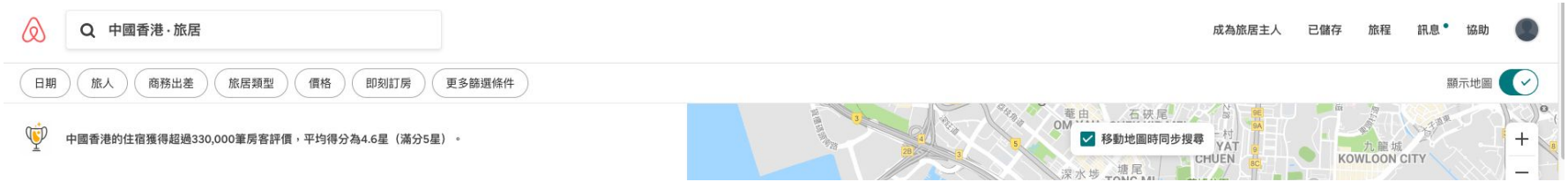
Airbnb will help analyze which listings are much more suitable to the customers by ranking them accordingly.

Aims at reducing searching time and increasing the chance of matching between hosts and users.

Expect to have 20% more transactions made when it comes to the personalised ranking features.

Recommendation #2: Monetary Incentive for Reviewers

Key Issue: Accuracy of review is low, as only half of the users had left reviews on the page, which might overestimate the quality of listings and make it difficult for users to make bookings based on the reviews.



Method #1

Offer a future discount

Key Takeaways: 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.

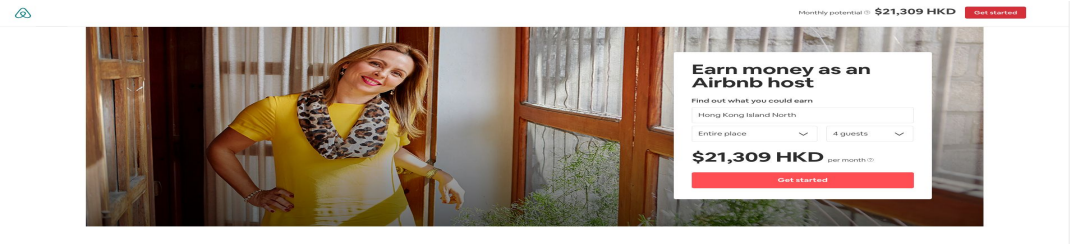
Method #2

Prevent Hosts Incentivizing Users for Positive Reviews

Key Takeaways: Penalties will be given to hosts that offer incentives to their guests for writing reviews to prevent hosts from offering incentives to users for high reviews

Recommendation #4: Price Disclaimers for Airbnb Users

Key Issue: Customer do not know whether or not the property is overpriced or underpriced compared to similar listings



Current Pricing System of Airbnb

- Airbnb offers huge flexibility for hosts to determine the price of their own listings
- The smart pricing system of Airbnb helps hosts 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.

Machine Learning Model to Forecast Price

- Based on the previous data collected, the model can predict prices with an RMSE of 0.908462
- The predicted result is highly accurate; which may provide huge insight for guests when they are selecting from listings to stay in

Our Solution: Disclaimer to Inform Guests

Predicted price range based on filters applied by Guests as well as personal factors selected in Recommendation #3

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.