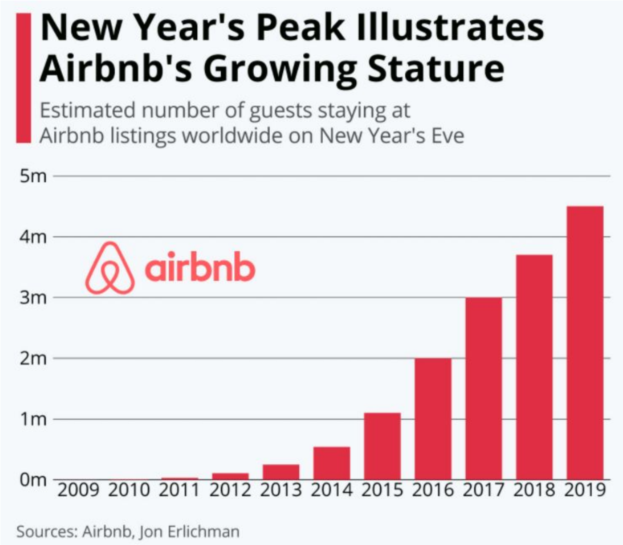
Sentiment Analysis on Airbnb Reviews & Price Prediction

GA DSI-17 Capstone Project

By: Leow Yong Khiang

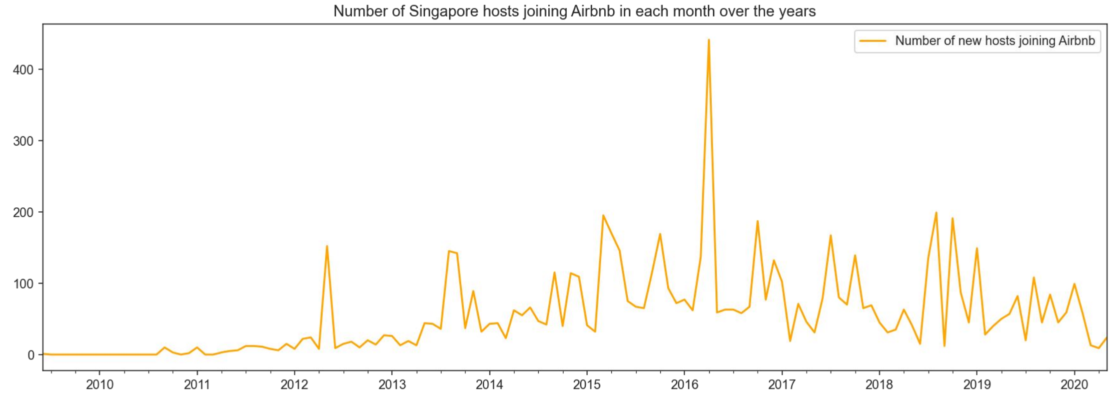
Background of Airbnb



- Online marketplace for people to rent out their properties or rooms to guests
- Increasing popularity over the years; disrupted travel and hospitality industry
- Money-making opportunity for home -owners

SINGAPORE | HOUSING

Average Singapore Airbnb host 'makes about $5,000 a year'



Problem Statement

Challenges faced by Airbnb hosts

- Understanding determinants of customer satisfaction
- Setting an optimal listing price to maximise income

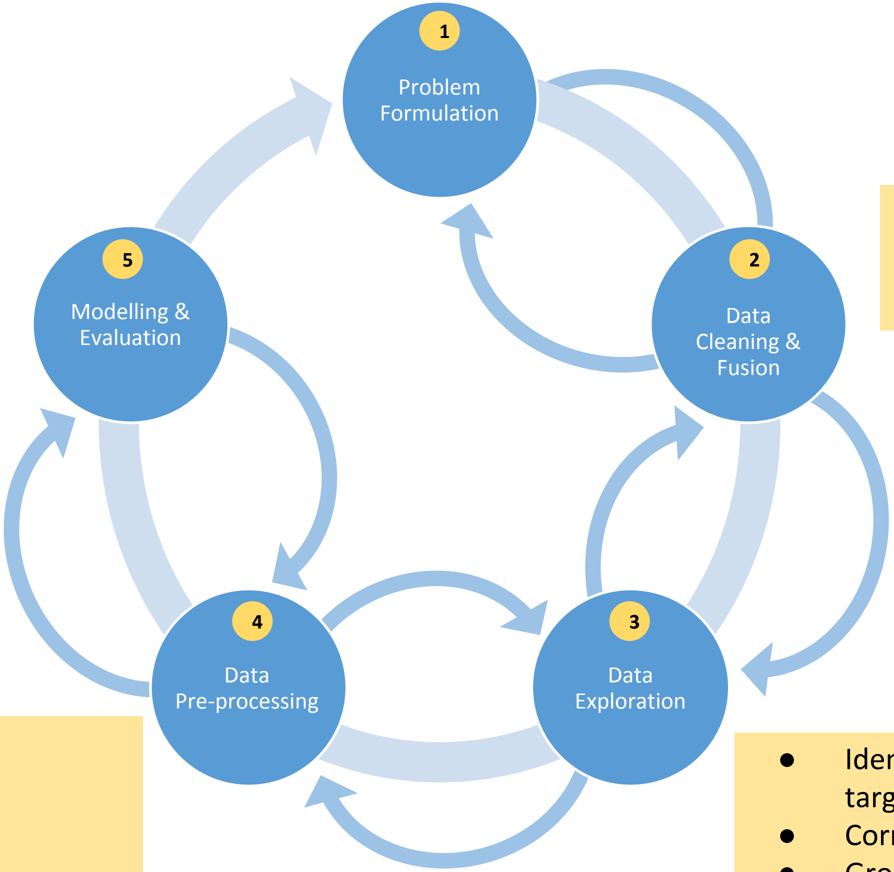
The goal

- Gain insights on sentiments of customer reviews and factors that drive customer satisfaction using Natural Language Processing techniques
- Develop a price prediction model using machine learning techniques
    - Metrics: R^2 and Root Mean Square Error (RMSE)

The Data Set

- Sourced from “Inside Airbnb” website, an investigatory/ watchdog website which scrapes and reports data on Airbnb websites for multiple cities around the world
- Singapore dataset scrapped on 22 June 2020
- Reviews Data
    - 91250 reviews on 4488 unique listings
- Listings Data
    - 7323 listings with 106 attributes

Methodology



- Hyperparameter tuning

- Drop non-English reviews
- Drop/ impute null values
- Remove/ modify outliers

- Stopword removal
- Lemmatize tokens
- Add bigrams
- One-hot encoding
- Transform and scale variables

- Identify influential features on target variable
- Correlation analysis
- Group categorical feature levels
- Feature engineering

Sentiment Analysis on Reviews

Approach

Sentiment Analysis

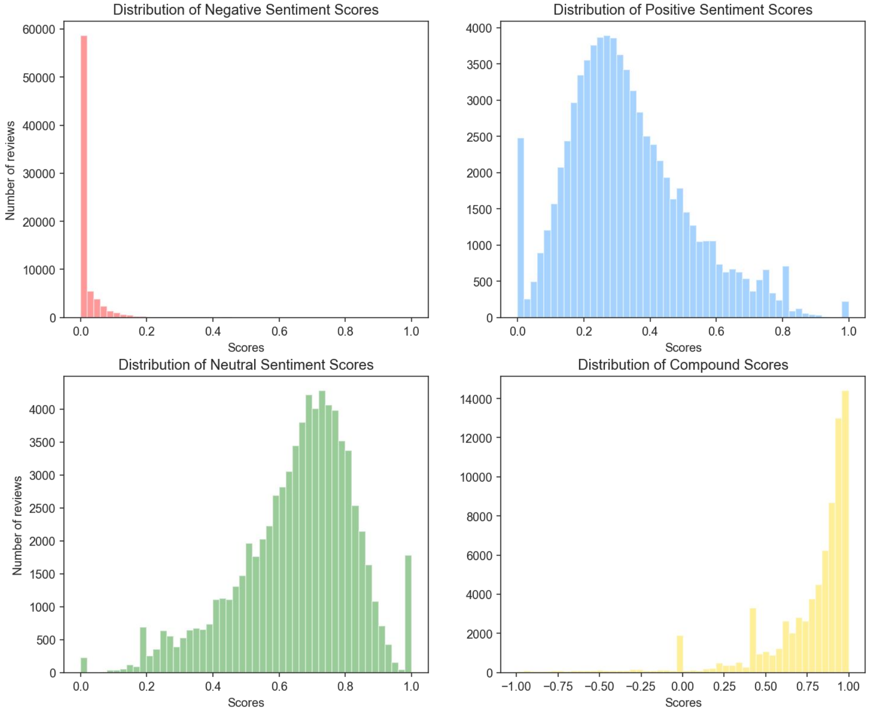
- Classify reviews into positive or negative sentiments
- VADER (Valence Aware Dictionary and sEntiment Reasoner) tool
    - sensitive to both polarity (positive/ negative) and intensity (strength) of emotion
    - able to account for differences in magnitude of sentiment intensity by considering emojis/ emoticons, punctuations and capitalizations found in social media reviews

Topic Modelling

- Extract hidden topics from positive and negative reviews
    - Latent Dirichlet Allocation (LDA) model
- Model selection based on coherence score and interpretability

Findings

- 97% of reviews are overall positive



Positive Comments

Factors driving customer satisfaction

😀 Responsive communication and accurate listing description

😀 Provision of basic amenities e.g. toiletries, washing machine, breakfast

😀 Personalised interaction; friendly and warm hosts

😀 Quiet surroundings for good sleep

😀 Unique and vibrant neighbourhood

😀 Accessible to points of interests e.g. MRT stations, bus-stop, restaurants

😀 Holistic accommodation experience

Most representative documents

Holistic experience

"*What an incredible best-kept-secret hidden gem in Singapore. Photos do not do justice to this gorgeous stylish apartment, sparkling clean, with amazing amenities. No words are good enough to describe Darren's outstanding hospitality. The location is the best you can ever get in Singapore (I lived for 3 years right behind this building and I know what I am talking about!) , and the views from the balcony are breathtaking. Last but not least, if you enjoy durian, you can get the best at a stall right around the corner. Heaven! It was a beyond-exceptional experience. Thank you!*"

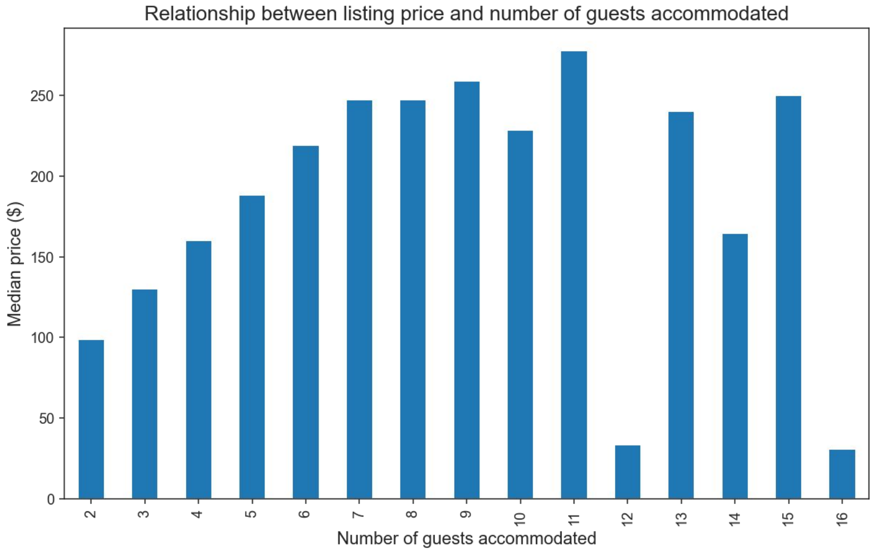
Personalised interaction; friendly and warm hosts

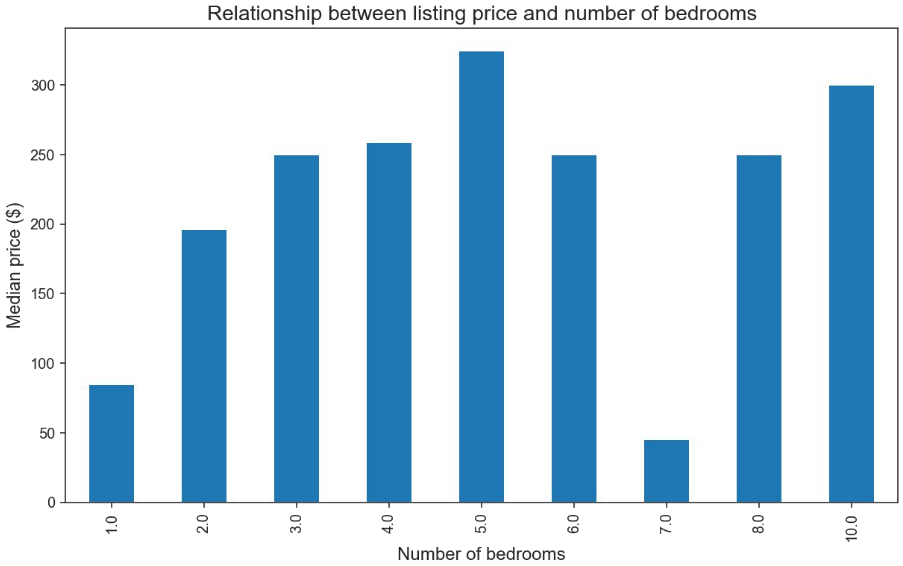
'Fran and Bross were the bet hosts I could have asked for. Before meeting, Fran was extremely prompt and efficient with her contact and emails which made booking extremely easy. She took care to know exactly when I was arriving and also that I knew how to get around, my hosts made me feel most welcome and I really felt pampered when I got to their place. They never disturbed me at any point and always made sure I was taken care of and looked after, I instantly felt like their son and really I have gained some very good friends from my stay at their place. Also, Fran and Bross make for great conversation! I cannot remember the amount of times I spent in deep thought and also in laughter. The place itself was extremely spacious and had everything I needed and pretty much had a bathroom all to myself for the entirety of my stay, Fran also took care of my laundry for me and both Fran and Bross took me around on some evenings which really I did not expect but I just had a great stay with them and they really made my trip! Will never forget them!'

Exploratory Data Analysis

Listings Data

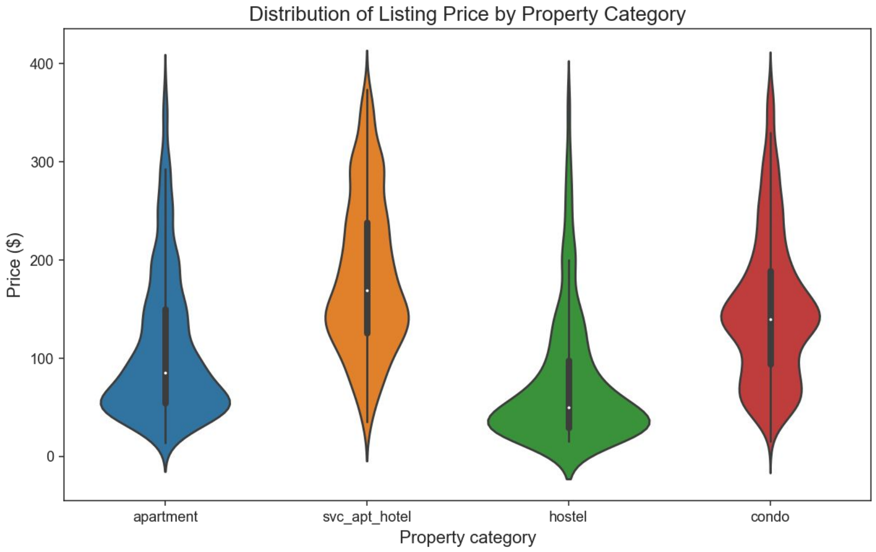
Accommodation Capacity

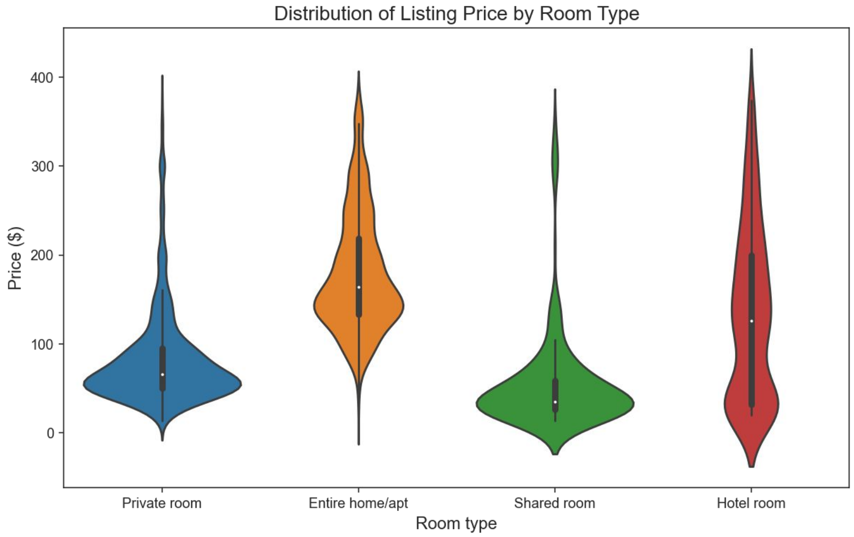




- Price generally increases with accommodation capacity
- Low prices for some listings with high accommodation capacity e.g. 12 and 16
    - Correspond to hostel listings (12 and 16-bed dorms)

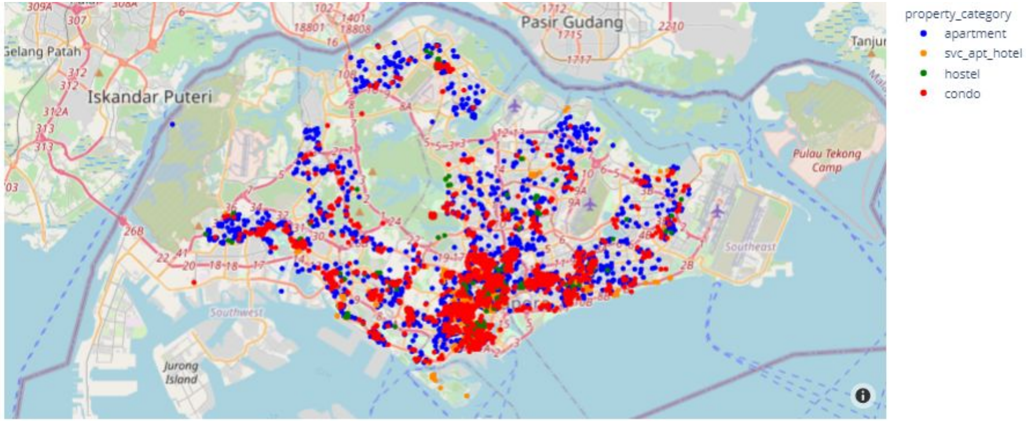
Property and Room Type

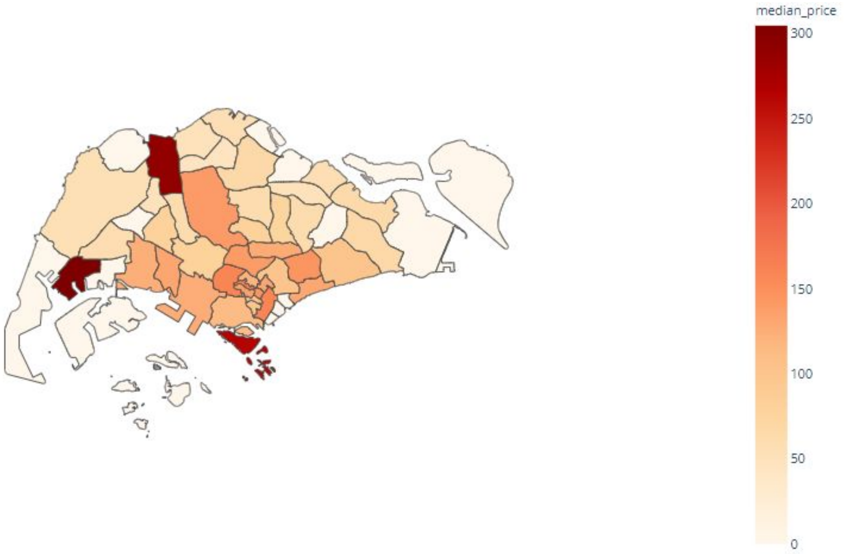




- Higher prices for
  - serviced apartment/ hotel and condominium
  - renting out entire home

Location

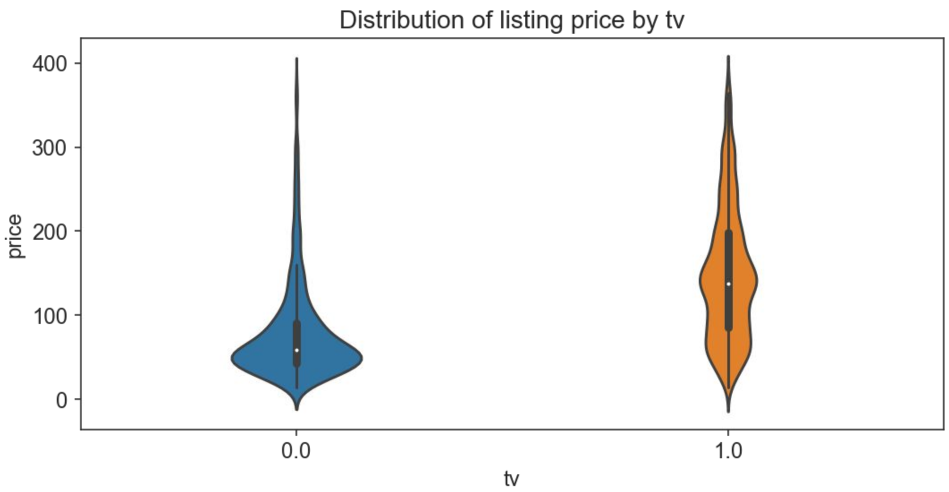




- More than 80% serviced apartments and condos located at city area
- Kallang and Geylang neighbourhoods has highest number of listings (1017 and 797 respectively)

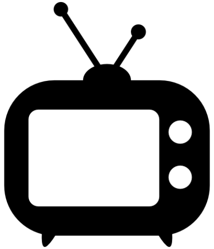
- Higher prices for listings located near the city
- Southern islands and Tanglin generally most expensive

Amenities

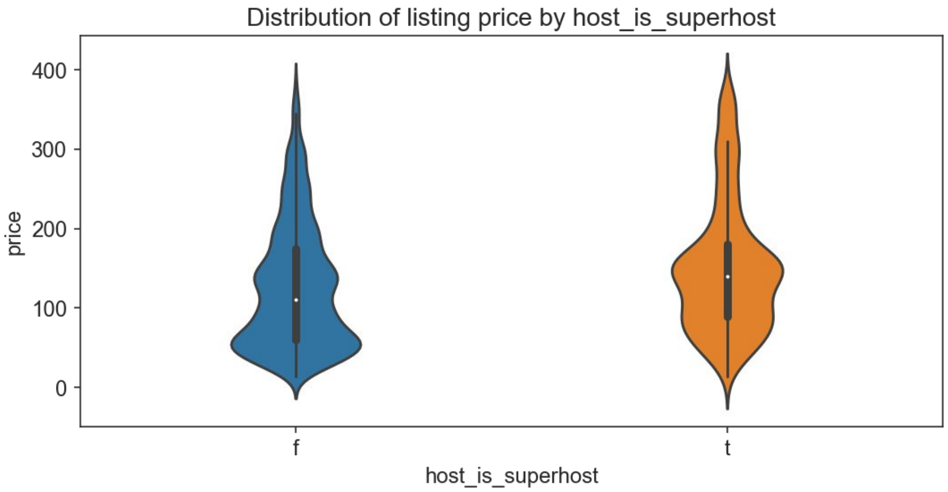


- Price incentive for accommodations with such amenities
  - Air-conditioner (+100%)
  - Gym (+68%)
  - Pets allowed (+60%)
  - Pool (+57%)
  - Bed linen (+25%)
  - Hot-tubs (+25%)
  - Child-friendly (+18%)

- TV amenity is biggest price differentiator
- Median price more than 2x higher for listings with TV (74% of listings provide TV)



Superhosts



- About 16% of hosts are superhosts
  - Host a minimum of 10 stays in a year
  - Respond to guests quickly and maintain a 90% response rate or higher
  - Have at least 80% 5-star reviews or maintain a 4.8 overall rating
  - Honour confirmed reservations (meaning hosts should rarely cancel)

- Pricing power for superhosts
  - 27% higher in listing price

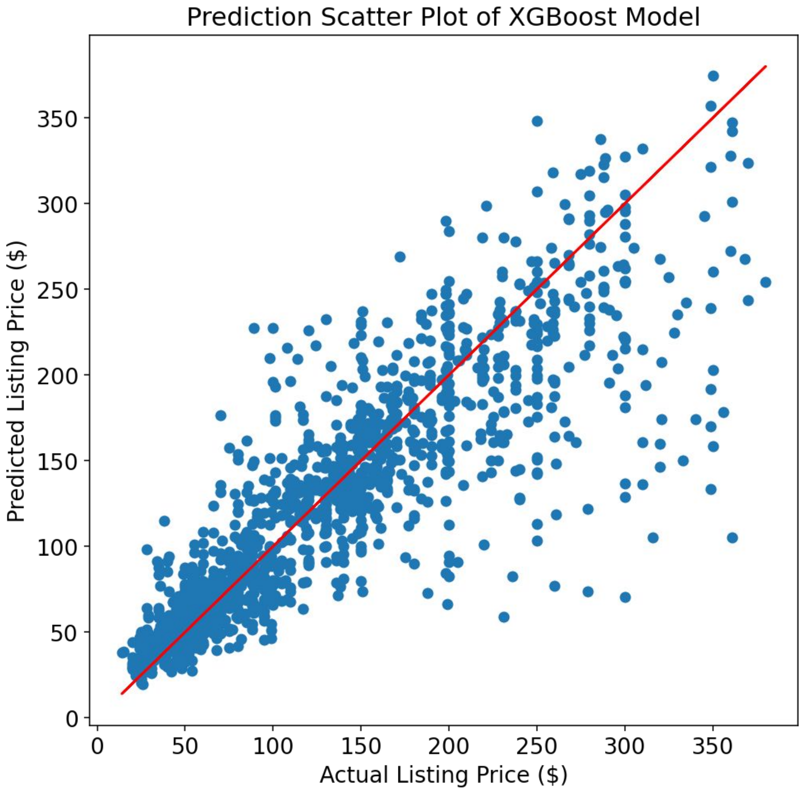
Price Prediction Model

Model Evaluation

Models	R^2 (Train)	R^2 (Test)	R^2 difference	RMSE (Train)	RMSE (Test)	RMSE difference
Linear Regression	0.676	0.682	0.006	50.49	51.88	1.39
ElasticNet	0.676	0.681	0.005	50.35	51.76	1.41
Lasso	0.663	0.664	0.001	50.98	52.37	1.39
Support Vector	0.908	0.824	0.084	26.69	38.84	12.15
AdaBoost	0.636	0.615	0.021	52.35	55.34	2.99
Random Forest	0.753	0.718	0.035	41.76	47.14	5.38
XGBoost	0.844	0.810	0.034	35.35	41.38	6.03

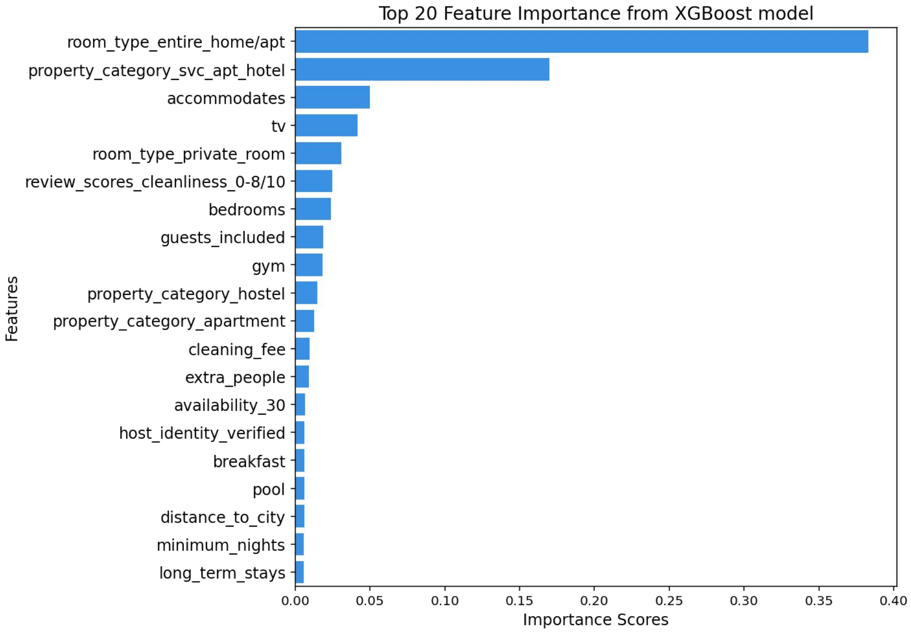
- Selected **XGBoost** regression model
    - High R^2 score and low RMSE against test data
    - Lower variance in R^2 and RMSE between train and test scores compared to Support Vector regressor, so better generalizability to unseen data

XGBoost Model Performance



- Model able to explain up to 81% of the variation in listing prices
- Prediction prices are within $26 of actual listing prices on average
- Model tends to under-estimate listings with higher prices

Feature Selection for Production Model



Models	R^2 (Test)	RMSE (Test)
XGBoost (full 72 features)	0.810	41.38
XGBoost (26 features)	0.801	41.82

- Negligible drop in prediction performance for production model with reduced set of 26 features compared to full model
  - Ease user input requirements with better generalizability

- Majority of features are of relatively low importance
- Selected 26 features for production model

Conclusion and Recommendations

- Insights from analysis on the reviews and listing data can be used by hosts to better target their listings to potential customers or to strategize how to best serve their customers by providing better service quality to boost host’s reputation and get ahead of their competitors
- Price prediction model useful for hosts to better plan their listing price to achieve balance between revenue and occupancy
- Directions for future work:
  - Use more accurate price data based on actual price paid by guests
  - Incorporate listing photo’s image quality as a feature to price prediction model
  - Include other proximity features such as accessibility to supermarkets and restaurants
  - Add in mode granular features specific to the accommodation unit e.g. unit level and views from unit

Thank You!