Agenda

1. **Data Preparation**
2. **Market Insight**
3. **Price Predicting Model**
4. **Further Discussion: Advanced Machine Learning Methods and Next Steps**

Data Preparation

Data Imputation

- Loss of values mostly in Review_score_rating, Room_type and Zipcode. If we eliminate all NA, data loss would be 25% of all observations. Therefore we have to resort to data imputation
- Replace NA values by Median (Review score, Bedrooms, Bathrooms), and Unknown (Room_type , City and Zipcode)

Re-define City Values

- Goal: assign city values back to only 5 cities: Boston, Seattle, New York City, San Francisco and Los Angeles
- Using the Zipcode to re-define city values, given that each city has a unique range of Zipcode
- Keep all 5 relevant cities in the final data set for further analysis

Filter invalid values

- Filter NA values: After data imputation and re-define value, the loss of observations reduce from nearly 10,000 observations to around 200 observations as final
- Filter invalid value: filter out all beds value equal 0, as there still exists bed-type regardless of no beds
- Filter one-factor value: Country and Balcony as they have no impact on the regression

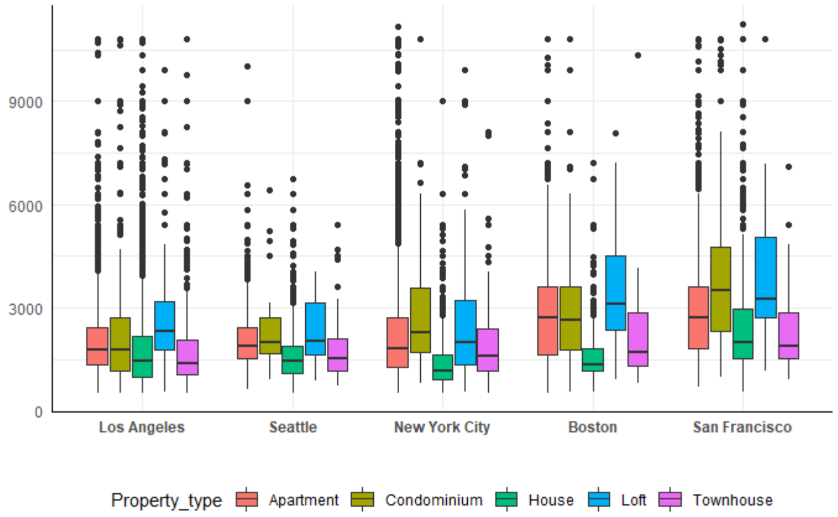
Modify variables type

- Mutate binary and categorical variables as factors (City, Company, Neighborhood, Room_type, Property_type Bed-type and binary variables)
- Add Region (West and East) variable to further observe the price differences in geography

Agenda

1. Data Preparation
2. Market Insight
3. Price Predicting Model
4. Further Discussion: Advanced Machine Learning Methods and Next Steps

New York City is dominant in terms of volume and variety of property type, but San Francisco is the most expensive region



Property Types across Five Cities – All

- New York City and Los Angeles are the top two cities that have most of property for lease (~15,000 objects per city). These two are metropolitan cities leading the economic trends.
- Apartment is the most dominant rental objects in every city. However, the apartment accounts for a bigger percentage in East Region than in West Region. In West Region, people also have the higher tendency to go for House than in the East Region. This is might due to the limited land resources in the East given the high density of population, whereas property are is much larger in the West

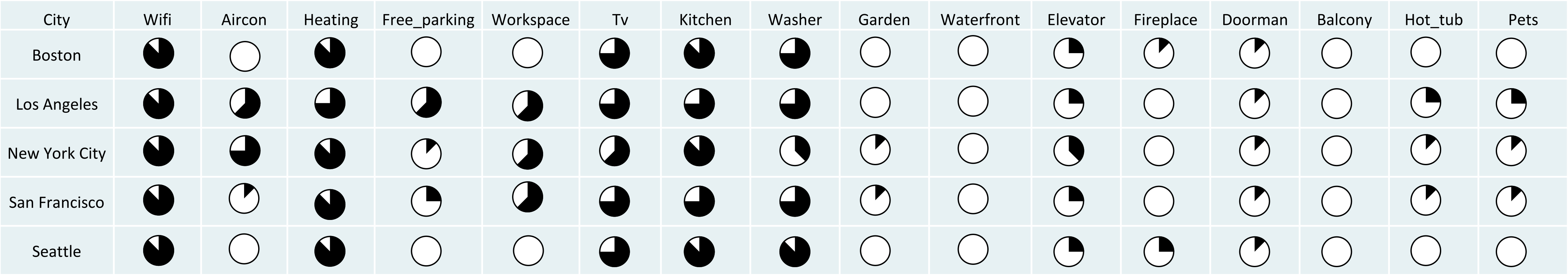
Property Types across Five Cities – Excluding Apartment and House

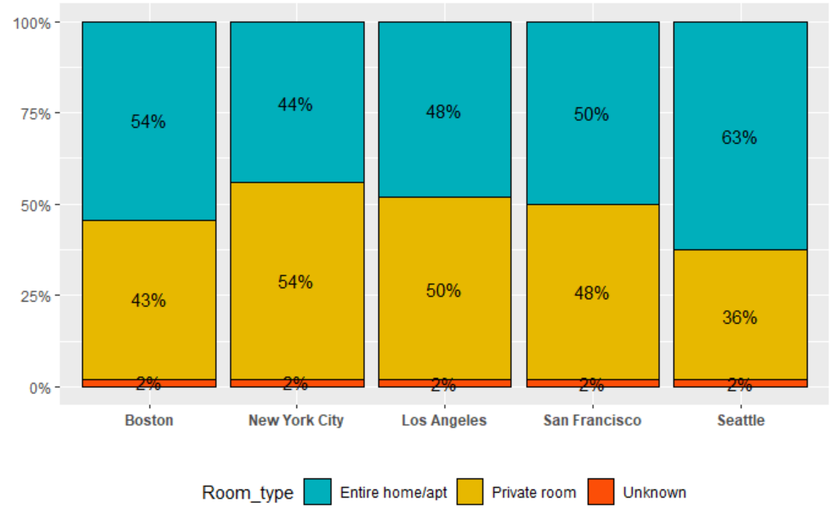
- After apartment and house, condominium, townhouse, loft are the three next popular types with presence in all cities
- In general, the West property market has a more diversified products than the East. Specifically, Los Angeles stands out with some products seems to be exclusive here: Bungalow, Villa and Guesthouse. The tourist attractions and nice weathers are potential attributors to the diversified and unique property market here.

Price of Property across Five Cities

- Looking at the top five property, the price range is quite consistent across five countries with Loft and Condominium have the highest price whereas House in a lowest price range
- Unexpectedly, San Francisco has the highest property type among five cities, given its position in the West. Limited availability of housing and high density of large corporations (Silicon Valley) might be the reason why price is so high here
- Other than San Francisco, property market in the West (LA and Seattle are much cheaper than the East (NYC and Boston)

Due to the differences in using purposes, facilities are significantly different among cities





Property Facilities across Five Cities

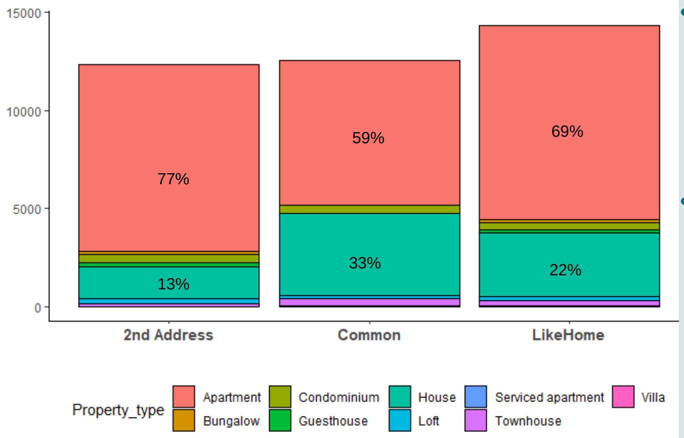
- There are facilities that are considered as essential and present at least 50% across price range at all five cities, including: Wi-Fi, Heating, Tv, Kitchen, Washer
- However, there are some other facilities that have no presence at all at some cities. Specifically at Boston and Seattle, no matter of the price range, no property allows pets and being equipped with Aircon, Free parking, Workspace, Garden, and Hot tub . In general, the property at metropolitan cities such as NYC, LA and San Francisco have more built-in facilities than those in Boston and Seattle. This is might due to the characteristics of housing market at big cities that demand more convenient at hand for the renters.
- It's possible that property at metropolitan cities are hired for the working purpose, as more than 50% of property here have aircon, free parking space and workspace

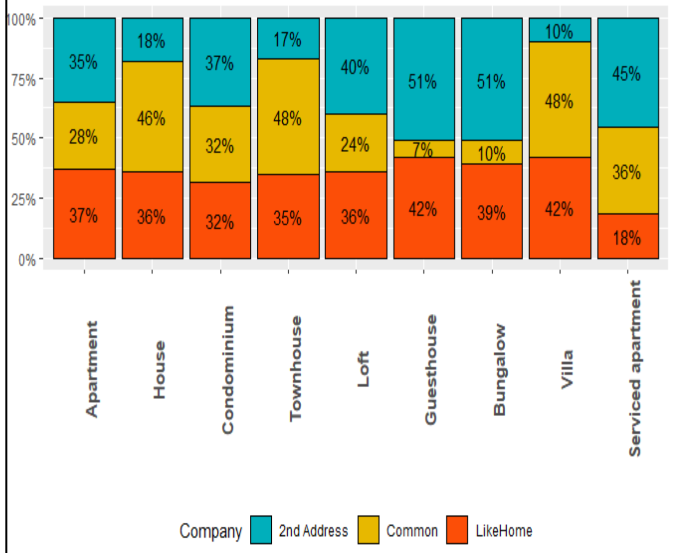
Room Types across Five Cities

- It seems that the more metropolitan the cities are, the higher the tendency that people rent a single/private room than an entire home. Firstly, it might due to the limited land area and high renting fee in bigger cities. Secondly, young people coming to big cities like NYC or LA to search for jobs will go for a single room, unlike other areas that are more family-oriented.

LikeHome is leading in terms of property volume, but 2nd Address is the winner in terms monthly revenue

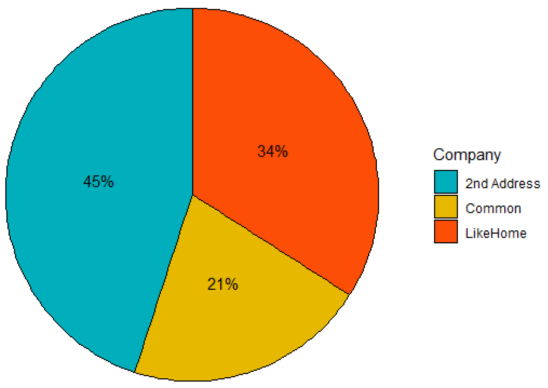
Volume Sales Across Companies

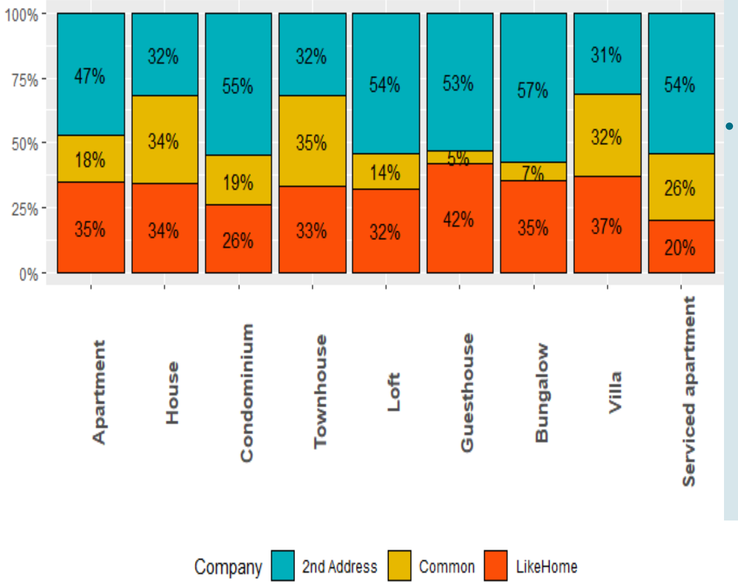




- In terms of volume for rent within each company, LikeHome leads the market (~14,300 units). Common slightly exceed 2nd Address
- Three companies are all strongly focus on Apartment and Houses, but the share between these two types are quite different. While 2nd Address have most volumes in Apt, Common have divided its strategy focus in both House and Apt
- In terms of volume share, Common is dominant in House, Townhouse and Villa, relatively in alignment with its strategy of focusing in housing type. 2nd Address has quite a strong presence in property that usually used for leisure purposes (Bungalow, Loft, Guesthouse). Whereas LikeHome keeps a quite consistent position in all property types (1st or 2nd on the market)

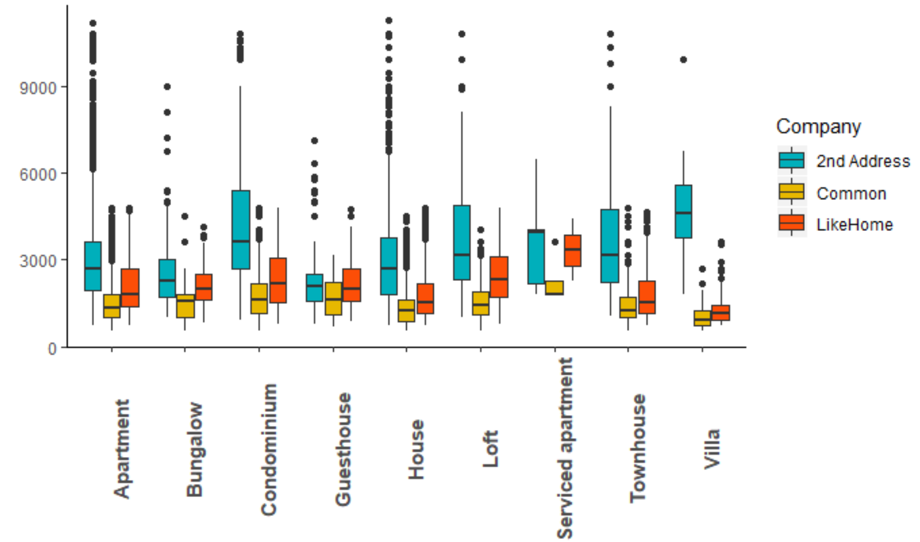
Monthly Revenue Across Companies

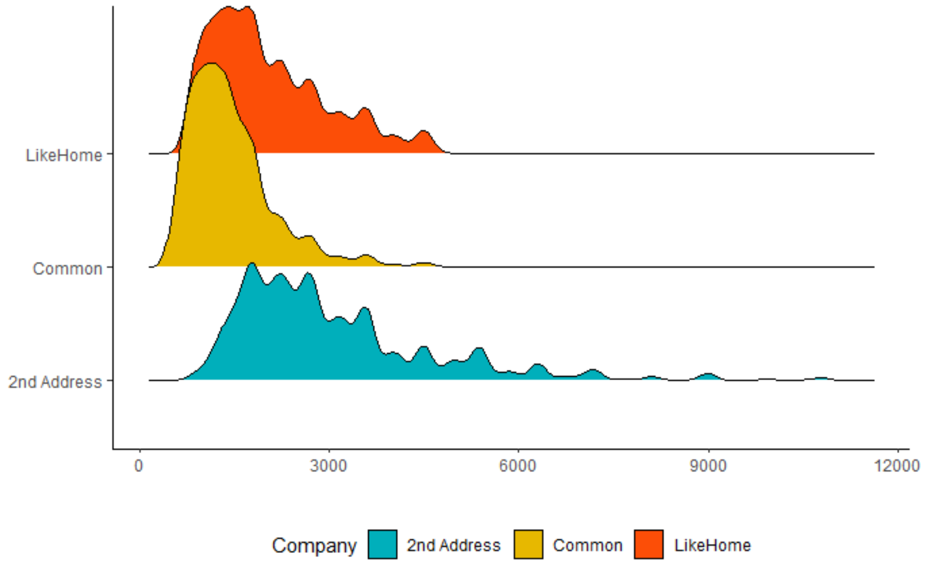




- Even though,2nd Address has the lowest volume for rent, their monthly revenue turns out to be the largest on the market (45%), following is LikeHome and lastly is Common
- Revenue share of 2nd Address always exceed its volume share in the market, which predicts that the price position of 2nd Address is significantly higher than Common and LikeHome
- 2nd Address has dominant position in Loft, Guesthouse and Bungalow. It seems that besides residential and business property, it also focus on property that serve for travelling purposes

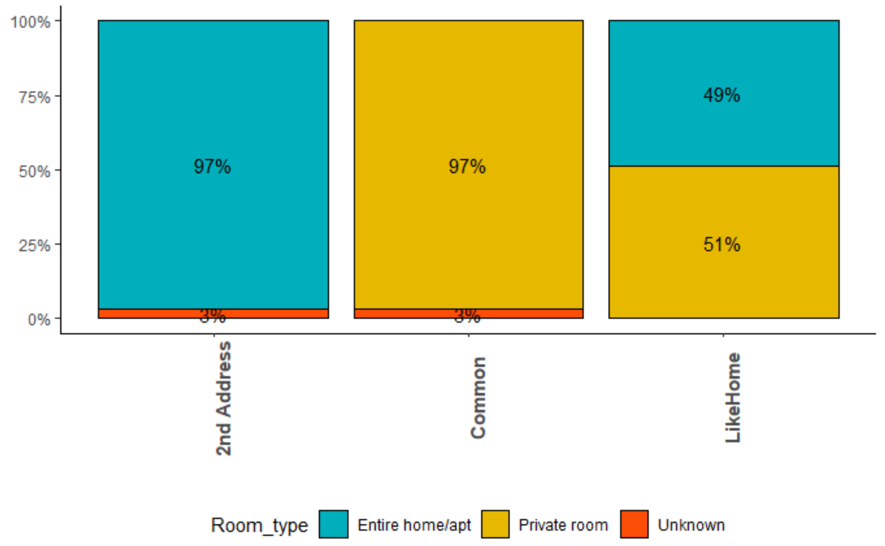
Pricing Position and Property Characters are the key differences between companies

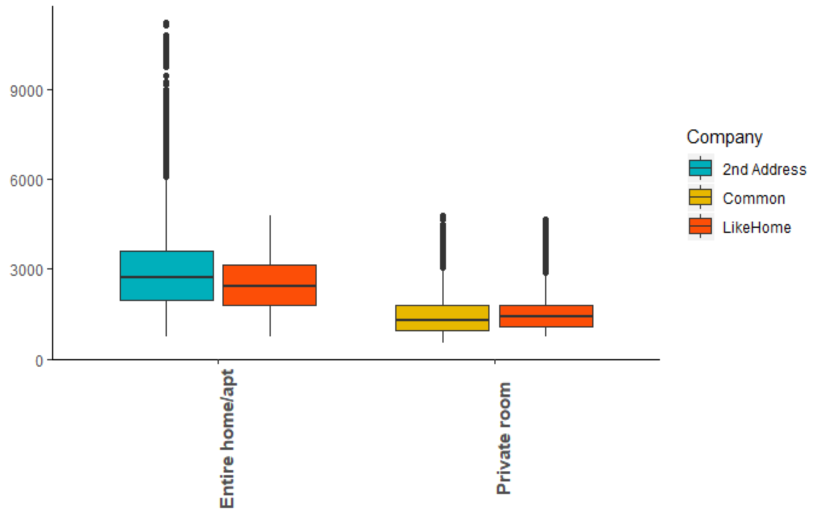




Price Range across Companies

- Price position of 2nd Address is significantly higher than those of LikeHome and Common across all property type. It seems that each company has its own price focus: 2nd Address – High price, LikeHome – Medium, Common – Low
- One factor that can be used to explain the price difference among companies is Room type: 2nd Address seems to solely focus on the Entire home/app, Common only rents property with Private room, whereas LikeHome sells the merge of those two. (Note that Price of Entire home is higher than price of a Private room)
- Second factor is the price position of the company itself. Price of 2nd Address is significantly higher not solely due to its focus on Entire Home/Apt but also in this type, it also has higher price range than those that also has presence in the Likehome product's porfolio

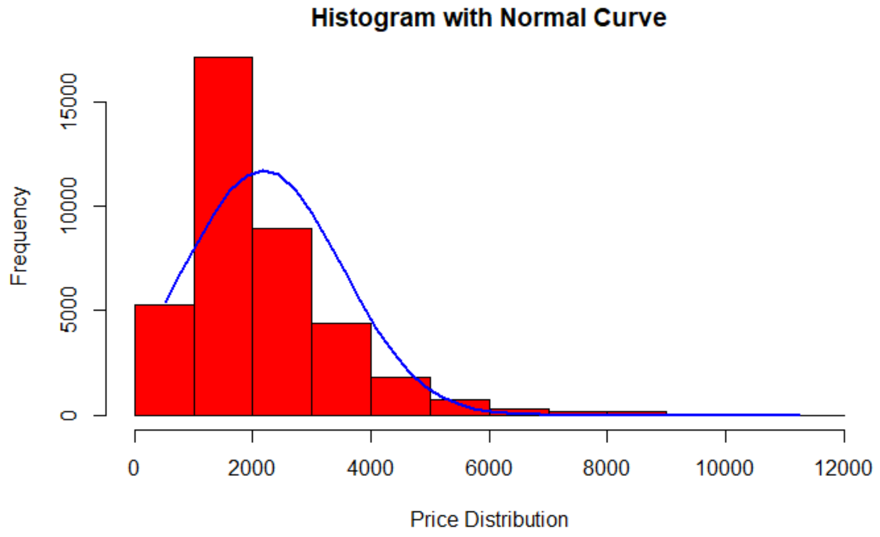


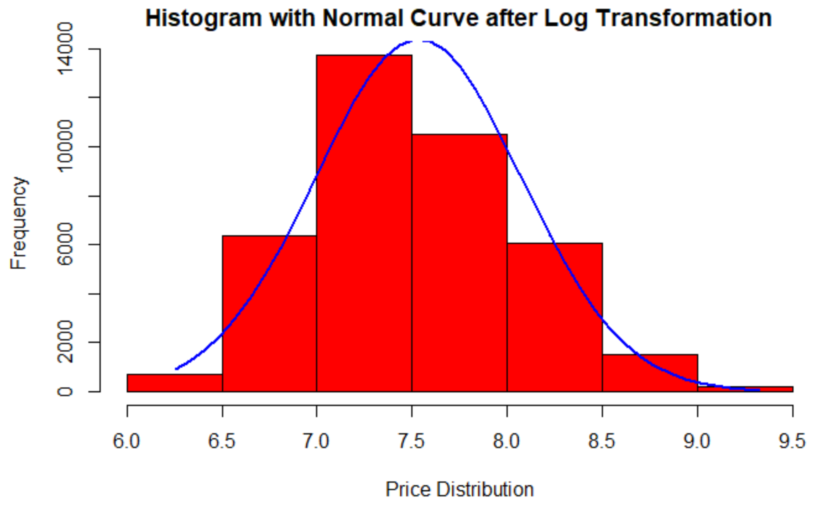


Agenda

1. **Data Preparation**
2. **Market Insight**
3. **Price Predicting Model**
4. **Further Discussion: Advanced Machine Learning Methods and Next Steps**

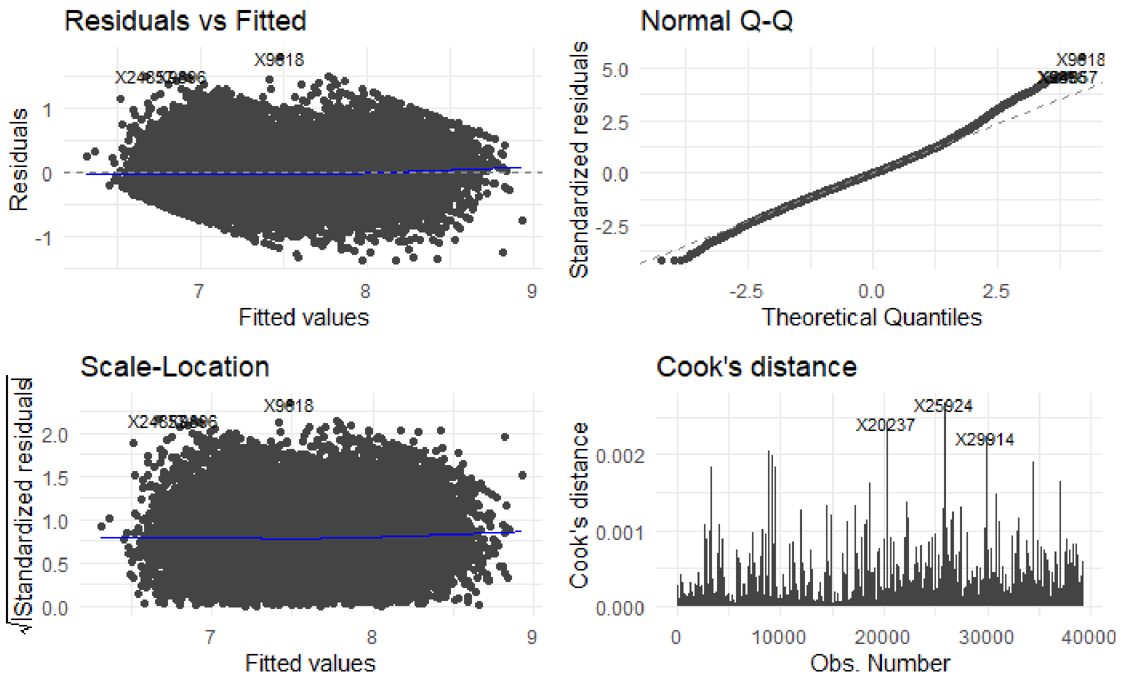
Price Prediction – Simple Linear Regression Model





Log Transformation for Price

The distribution of price is skewed to the right, does not follow the normal distribution under linear regression. Therefore, we will take the log transformation of price before running the linear regression model



1. Linear Relationship

- Equally spread residuals around horizontal line without distinct patterns
- Residuals and Fitted plot confirms a linear relationships between predictors and outcome variable

2. Homoscedascity

- The residuals appear randomly spread around horizontal line => no heteroscedascity

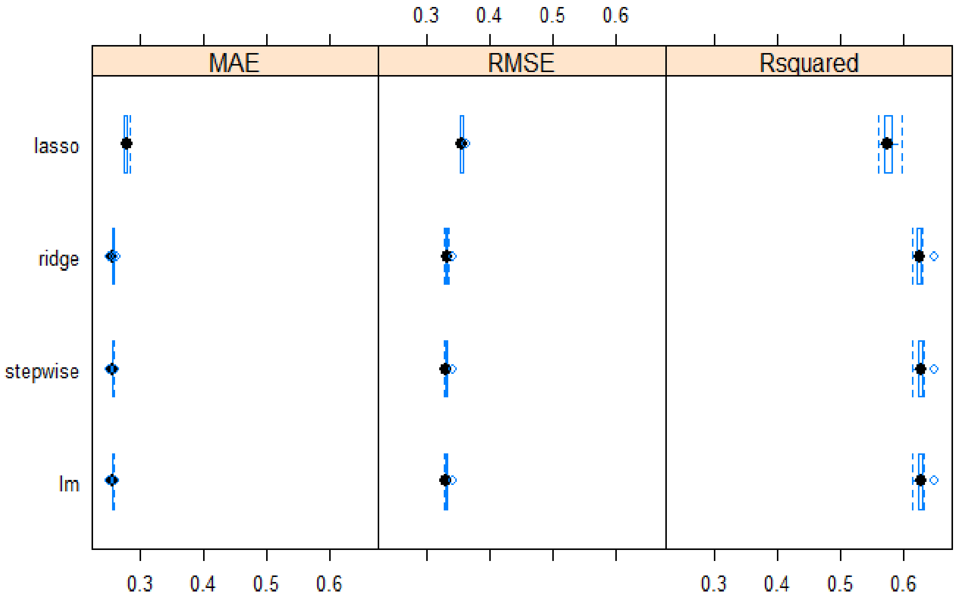
3. Normal Distribution

- The normality assumption is violated for expensive and cheap houses and (normal Q-Q plot not displaying a straight line)

4.Cook's Distance

- Points X20237, X25924, and X29914 are outliers and exert a high influence on our parameter estimates
- After exploration, these points all have imputed Review_score_ratings, the median might not truly reflect the value of these properties.

Price Prediction – Stepwise, Lasso and Ridge Regression Model



MAE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
lm	0.2504709	0.2551092	0.2555595	0.2557393	0.2560493	0.2609123	0
lasso	0.2750213	0.2758168	0.2787856	0.2780817	0.2791334	0.2833075	0
stepwise	0.2504709	0.2551092	0.2555595	0.2557393	0.2560493	0.2609123	0
ridge	0.2515393	0.2568258	0.2569252	0.2570964	0.2572781	0.2619749	0

RMSE

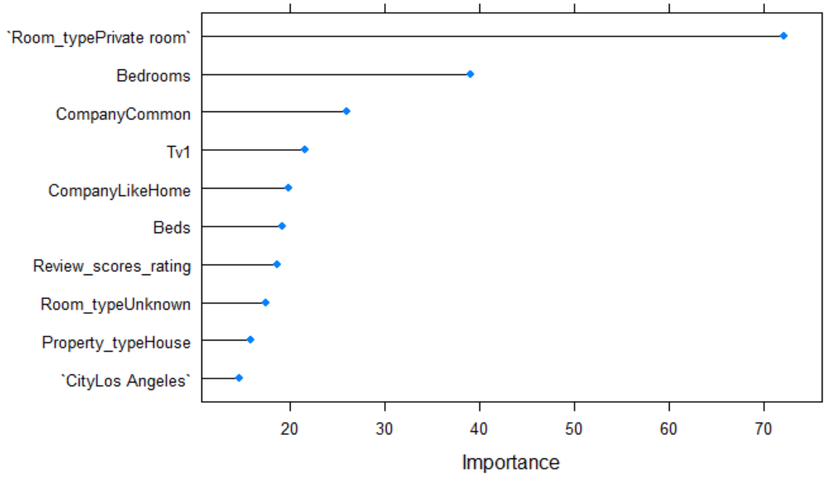
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
lm	0.3260238	0.3292119	0.3301307	0.3306095	0.3311505	0.3397802	0
lasso	0.3515480	0.3528666	0.3536619	0.3550081	0.3562291	0.3628428	0
stepwise	0.3260238	0.3292119	0.3301307	0.3306095	0.3311505	0.3397802	0
ridge	0.3275635	0.3302554	0.3310063	0.3314594	0.3319273	0.3403756	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
lm	0.6148574	0.6239323	0.6276510	0.6277758	0.6301484	0.6497860	0
lasso	0.5606751	0.5705510	0.5741482	0.5759351	0.5801575	0.5971389	0
stepwise	0.6148574	0.6239323	0.6276510	0.6277758	0.6301484	0.6497860	0
ridge	0.6141506	0.6226825	0.6262996	0.6263904	0.6284588	0.6478244	0

Performance Metrics

- R2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the outcome, so it's not a good metrics to compare models
- As we're comparing three models that resolves the same problems, it's relevant to consider between MAE and RSME
- Since our models have lots of outliners, it's better to use RSME to choose the best one as RSME penalize large errors, therefore is better in terms of reflecting performance when dealing with large error values.
- Comparing among four regression methods, based on the performance metrics as well as the efforts and time taken to run the regression, linear regression appears to be the best model that is not only has the lowest RSME, but also lowest MAE and highest Squared
- Stepwise Regression appears to be a quite potential model as it results the similar statistics to linear regression, but considering the large dataset with multiple predictors, using this regression can be complicated and time-consuming



Important Features

- According to the linear regression, Room_type has the most impact on the pricing of properties, especially the Entire Rooms/Apartment type. Further (later) analysis also shows that properties that have Entire room together with Real Bed in Bed types will enjoy a premium in pricing. It seems like tenants place high importance on features that offers comfortability, therefore these feature's existence will drive up the price.
- Other following important features are company 2nd Address, as they follow a high segment properties, they charge quite a high price compared to LikeHome and Common.

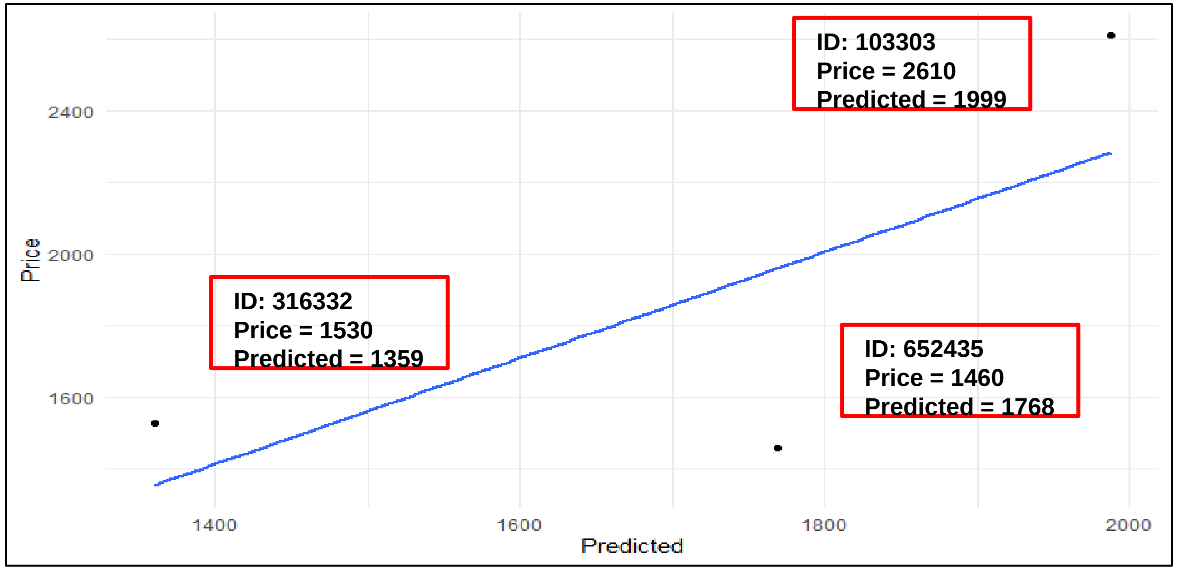
Price Prediction – Potential Features and Model Implications

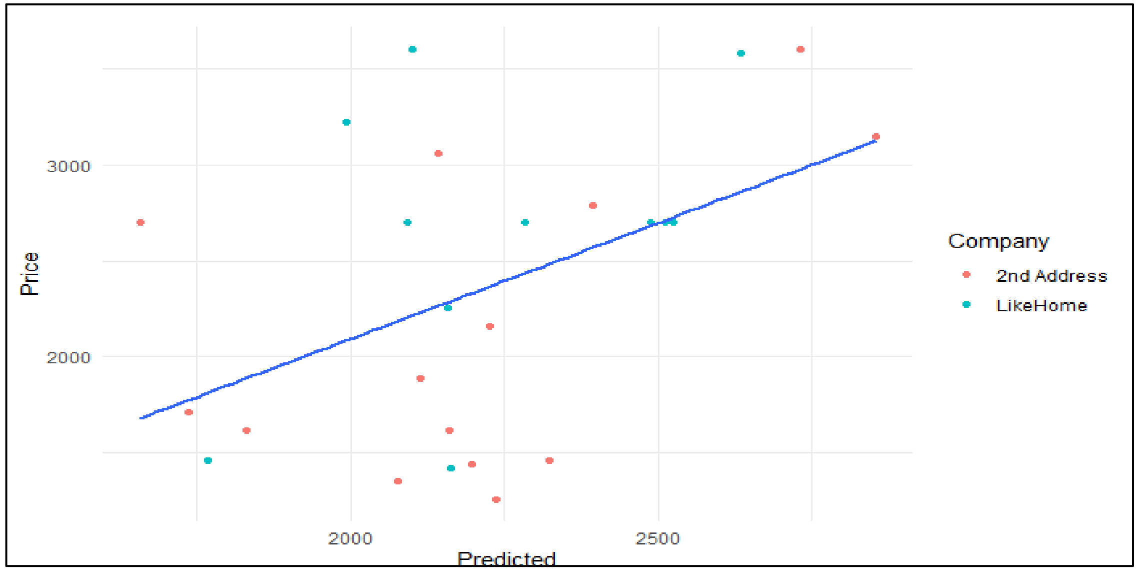
Model Implications

- For product ID 1003303, Price is higher than model predicted, however this does not reflect that the property is overpriced. Because, according to observations of the predictors, this property has both Entire Room/Apt and Real Bed features, that make the product in a premium price range.
- For product ID 316332, the price is higher than model predicted, but again this might due to the median imputation of review score rating, which does not reflect the true condition of this property. Therefore, the stated price might be reasonable because of high review score rating
- For product ID 652435, even though according to the model, this property is underpriced, however, looking at its competitor 2nd Address that offers the same product (location, room type, bed type neighborhood, they also constantly offer underpriced products, therefore this product is also underpriced to be competitive

Other Potential Features to Predict Rental Prices

- Features reflects property's conditions: Year built, Renovation year, Property Size,
- Features reflect property's facilities: Garage, Pool, Garage Size, Pool Size
- Features reflects the neighborhood surroundings: Density, Type of road access, Nearby Attraction Area





Agenda

1. **Data Preparation**
2. **Market Insight**
3. **Price Predicting Model**
4. **Further Discussion: Advanced Machine Learning Methods and Next Steps**

Further Discussions on Models and Insights

Advanced Machine Learning Applications

- It’s undoubtedly beneficial to apply advanced machine learning methods such as Gradient Boosting or Neural Network to predict price of properties. As discussed earlier, Simple Linear Regression stands out as the best model with lowest RSME, however, it’s predicting power still has some limitations that cannot help to capture all the effects of independent variables on price,

- For example, the interactions between variables (the effect on one variable on price is dependent on the value of the variable). In the graph between Price and Predicted (Annex), the red rectangular consists of points that have similar price regardless of facilities features except for both presence of Entire Home/Apt and Real Bed. This means that the interaction between these variables can significantly impact on pricing, which has not been captured in the linear regression. Therefore, if we use advanced machine learning methods, for example, the tree-based model, this interaction can be reflected in the price to be more accurate

- However, we also have considered the effort needed to apply the advanced method into analyzing our data set. It’s very computational costly and time-consuming considering nearly 10,000 observations, especially we might need to extend the data to include more information, for example more countries or more predictors.

Further Business Insights and Recommendations

- Untapped markets and products (Annex): Currently there are some markets have been left blank or not fulfill by the LikeHome itself, the manager can consider to expand its product portfolio. For example, no Guesthouse, Service Apartment and Villa in Seattle: LikeHome can be the pioneer in these markets

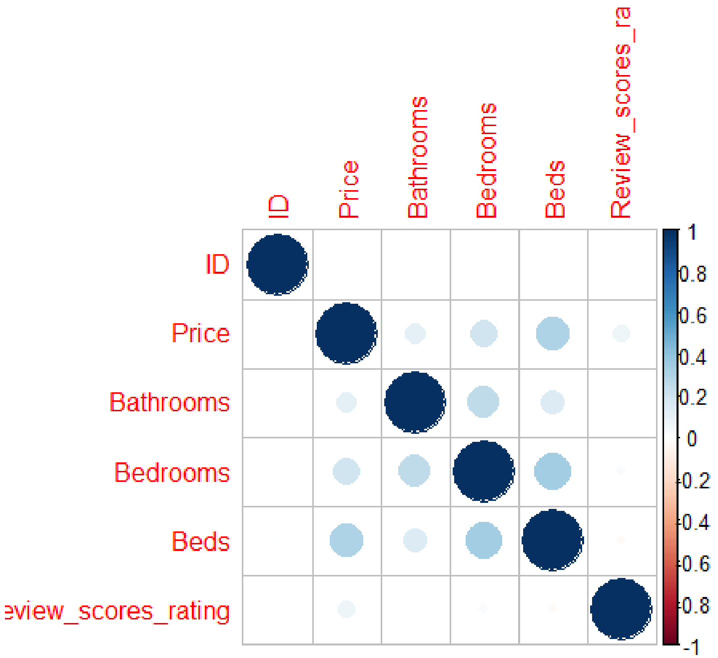
- Important features that impacts on price of properties: There are some facilities that have significant impacts on price, whereas others show no impacts at all. Therefore, it’s possible for LikeHome to focus searching, prioritizing and making decisions on properties that have those important features, for example Room type and Bed type, and potential neglect less important ones.

- Review scores rating is the factors that be highly influenced by customers’ experience and have high impact on price as well. Therefore, it’s important that LikeHome should focus on enhance customer experience, not only through the quality of properties, but also through the whole experience of renting properties with the company (customers care, resolve customer complaint, etc)

- High potentiality in the application of advanced machine learning techniques in price predictions to win the markets. However company need to consider the trade off between accuracy and costly effort related to the implementation

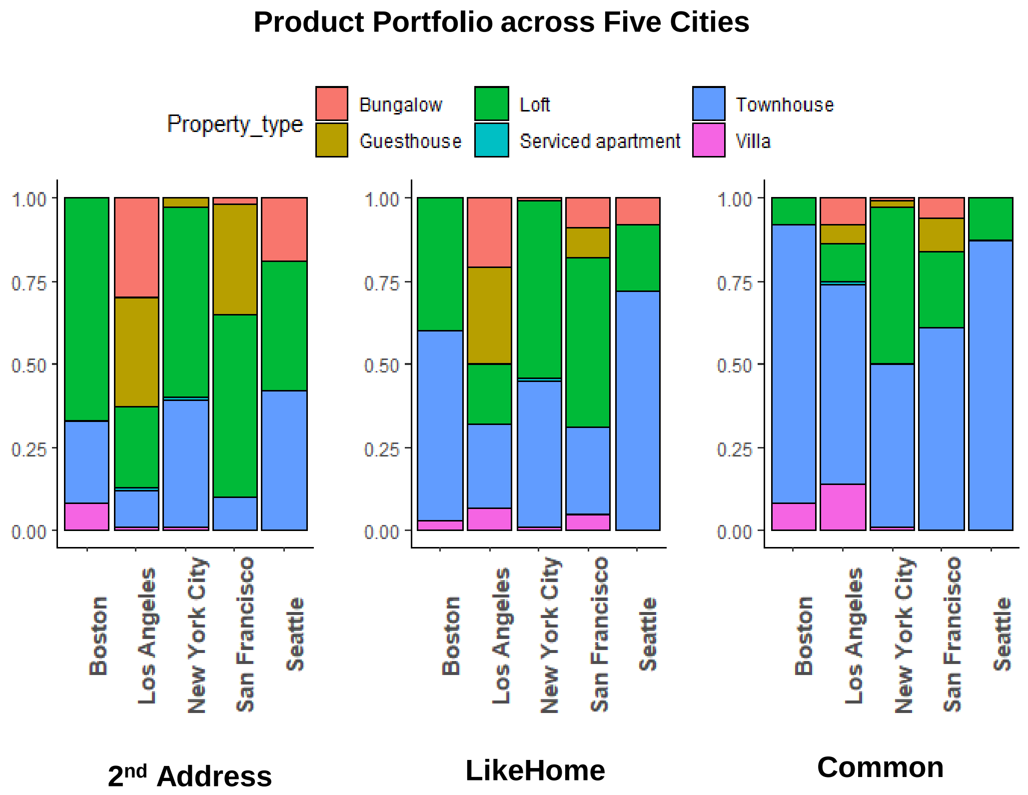
Annex

Correlations between continuous variables



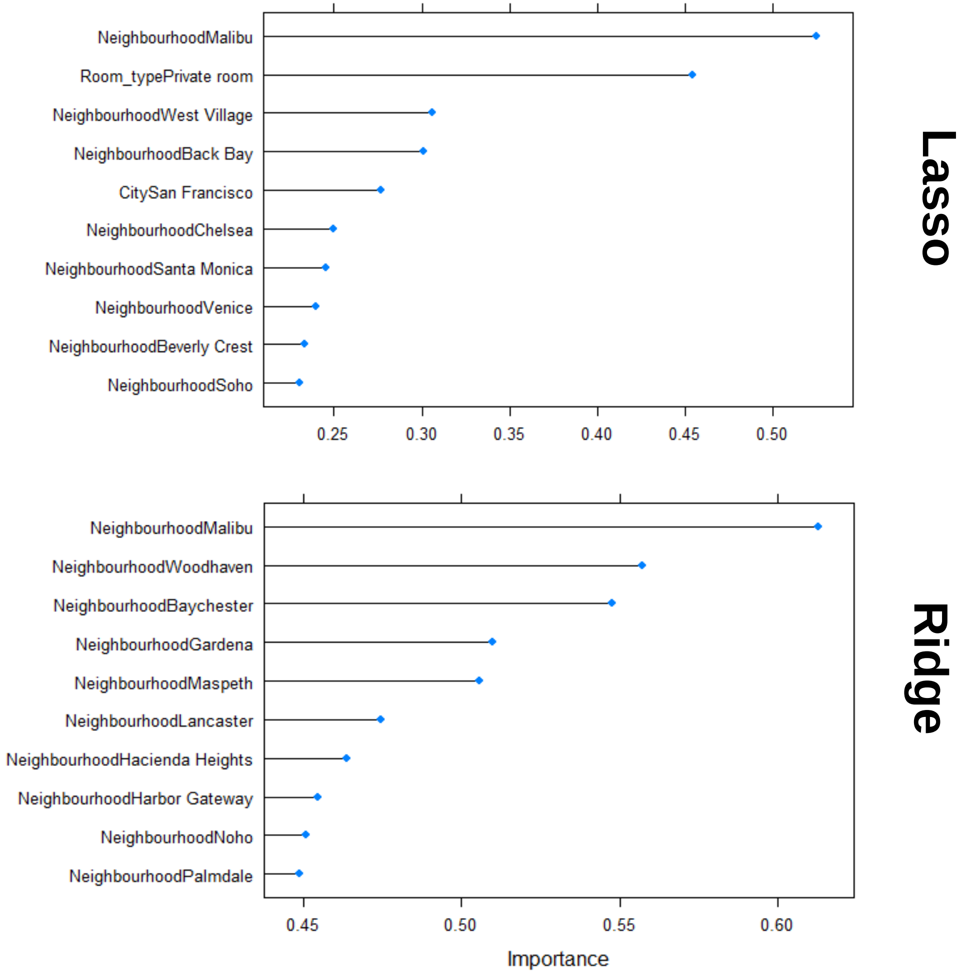
There is no significant relationship between continuous variables, therefore we can keep all of them in the regression model to predict price

Analysis on Product Portfolio



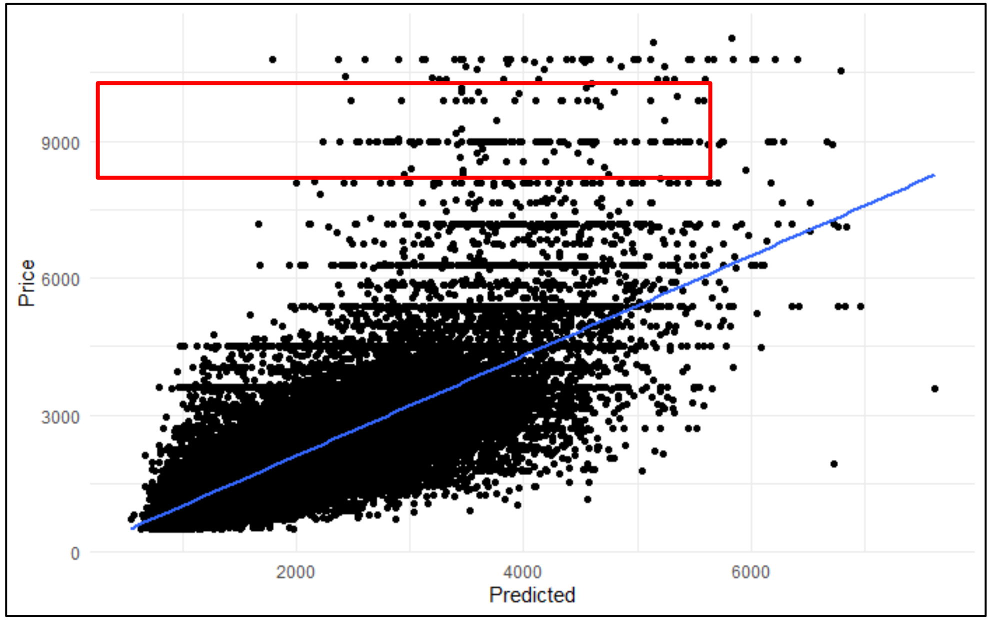
Product portfolio shows that there are untapped markets that firms can exploit and take dominant positions at first

Important Features according to Lasso and Ridge



Lasso and Ridge place more importance on the impact of Neighborhood position on price

Graph plotting Price and Predicted (Linear Regression)



The red rectangular consists of points that have similar price regardless of facilities features except for both presence of Entire Home/Apt and Real Bed. This means that the interaction between these variables can significantly impact on pricing, which has not been captured in the linear regression