airbnb

Pricing Optimality

An Airbnb Case Study

STAT 471 Final Project
Cho | Fan | Zheng

I. Abstract

Motivation

“How much is my property worth?” is a recurring question in the space of real estate. Finding competitive pricing for property leasing and sales is essential to both households and businesses in the advent of needs such as relocation or simply seeking an alternative revenue stream as property holders.

The Opportunity

In this paper, our team seeks to **explore a new way to approximate the best price for a property by using predictive tools**. Our objective is to obtain a model that can output an “expected market value” of a property based on some features and fields (physical or non-physical). We believe this generalized approximation can be done by training the model on open market data and pricing points of past openly traded properties. We see the “expected market value” as not only a reference point for the value of the house, but also a point of opportunity for undercutting the market, thereby gaining a competitive advantage over other competitors.

Secondary Objectives

Apart from the aforementioned primary goal of constructing a predictive model to gauge the expected market value of a property, we also carry two secondary goals that we examine in the project.

1. **Influences Across Years**

   We want to examine the factors affecting property prices, specifically regarding how they’ve changed over recent years and across different metropolitan areas. With this goal in mind, we anticipate an opportunity to forecast the future trend of the property industry and offer tailored recommendations to future property owners.

2. **Non-Physical Features**

   We want to determine whether or not contemporary housing prices are affected by non-physical house related features, such as the status of the renter or hosts and qualitative description attached for a listing. With this goal, we seek to better understand current property market in relation to non-traditional brick and mortar factors and construct a nuanced strategy on how to further create value out beyond the physical build of a property.

The Data Source

We obtained our data from the Airbnb housing dataset provided on Kaggle. The dataset selected is a great simulation of the market condition - rational actors, no monopolies, etc. - that is needed for us to make a fair expected value calculation. The dataset contains 74,111 properties and 29 columns. It has a series of relevant information, that caters to our investigative goal:

- **physical features** of the property, including its type and number of bedrooms
- **rental information,** such as whether or not it’s on a per room basis or whether or not it has a cancellation policy
- **host information,** such as whether or not he/she is verified
- **location information** by city and GPS.

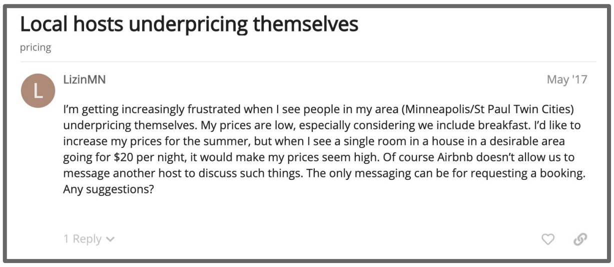
The data also span across 9 years from 2008 to 2017 in 6 major US cities.

Paper Flow

This paper will start off with a section of exploratory data analysis focusing on the inferences of the secondary goals, followed by a process of detailed model construction where we will employ standard modelling and evaluation strategies. The report will conclude with a final interpretation of results and recommendations to relevant stakeholders in the industry.

II. Problem Description

Homeowners aren't able to price their houses at a satisfactory level



Local hosts underpricing themselves

pricing

LizinMN May '17

I'm getting increasingly frustrated when I see people in my area (Minneapolis/St Paul Twin Cities) underpricing themselves. My prices are low, especially considering we include breakfast. I'd like to increase my prices for the summer, but when I see a single room in a house in a desirable area going for $20 per night, it would make my prices seem high. Of course Airbnb doesn't allow us to message another host to discuss such things. The only messaging can be for requesting a booking. Any suggestions?

The housing market is extremely saturated

Two-thirds of adult Americans own houses. Just as Airbnb becomes more popular, so too does the number of listings increases. The consumer will often have their pick of comparable properties to rent.

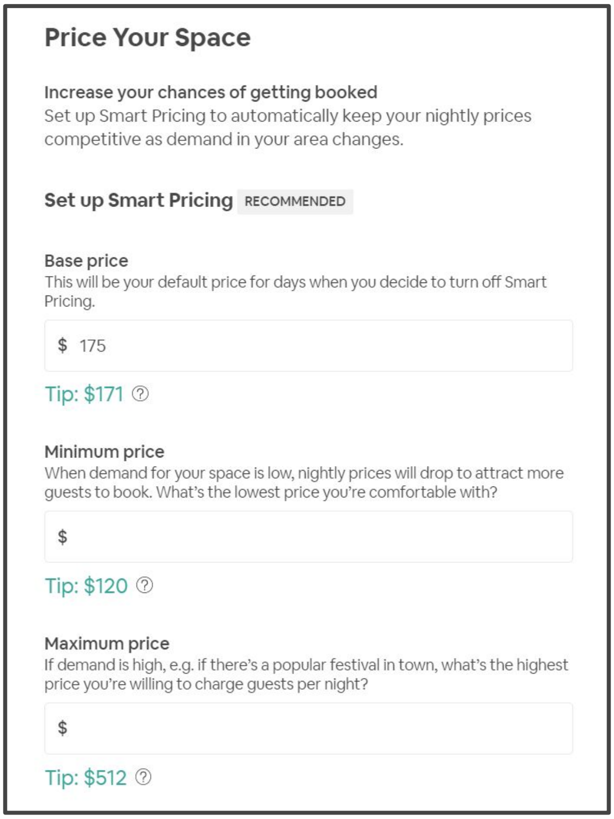
Adverse selection and asymmetric information results in homeowners constantly dropping their prices. In other words, renters may believe that the houses available aren't as good as those that are unavailable, resulting in lower demand and prices in the process.

It is easy to miss out on profit margins if the listing is not optimally priced. With experience, listers would likely gain a better understanding of the market and best pricing strategies. So, inexperienced entrants to the housing market will lose out to veterans as they lack the same background knowledge. Since determining the optimal price for properties can be challenging, property owners will feel less inclined to lease out their homes if the costs begin to outweigh the benefits and if they don't know the worth of their own properties.

Although 60% of potential housing income comes from the type of **property, location, and amenities,** 40% comes from implementing an **effective Airbnb pricing strategy. As such, it's crucial that homeowners understand their property's value.**

There are issues with Airbnb's pricing model that prevent lessors from maximizing property values

Airbnb currently offers two methods for pricing homes; however, both have drawbacks, leaving potential for a better model to be introduced.



Price Your Space

Increase your chances of getting booked

Set up Smart Pricing to automatically keep your nightly prices competitive as demand in your area changes.

Set up Smart Pricing RECOMMENDED

Base price

This will be your default price for days when you decide to turn off Smart Pricing.

$ 175

Tip: $171

Minimum price

When demand for your space is low, nightly prices will drop to attract more guests to book. What's the lowest price you're comfortable with?

$

Tip: $120

Maximum price

If demand is high, e.g. if there's a popular festival in town, what's the highest price you're willing to charge guests per night?

$

Tip: $512

Smart Pricing

Using Airbnb's smart pricing method, property owners can determine their house's value based on the current trends of the housing market. This method has two flaws: firstly selecting the market based price offers no competitive edge to the seller, so it is likely not the best profit margin. Secondly, allowing Airbnb to determine price takes away homeowners' ownership over the Pricing Point.

Manual Pricing

Airbnb offers a manual pricing method, where it recommends a range from which the seller should list their house. However, this suggested range is usually extremely wide and imprecise - sometimes around $500 - imposing huge fluctuations to the property value.

II. Problem Description

According to Guesty, a house renting advice website, *there is an expectation for the host to stay relevant and competitive by comparing similar Airbnb listing in their area*. **However the definition of similarities may be ambiguous as there are multiple facets customers can consider.**

To address this ambiguity, we try to address the following:

1. Can we do better than the offerings Airbnb give right now - **can we provide homeowners with a greater opportunity to best the market?**
2. Are there factors unrelated to the house that can boost a property's value?

Ultimately, can we tweak some of these variables to bring up the competitive pricing of a house?

Macroeconomic trends show big changes in the housing market in the past few years

Customers have started developing a perception that "house prices never fall". Such trend has only become more apparent as house prices continue to increase.

Despite this, mortgage rates are also on the rise, indicating homeowners' willingness to spend and borrow more.

More millennials are entering the housing market, with the generation bringing in an average, but sturdy, household income of around $88,200.

Financial stability is under greater pressure with the current pricing trends. In fact, housing prices have risen at the third-fastest rate in history over the past several years.

In this report, we will also attempt to analyze and verify these macroeconomic trends with respect to Airbnb rental prices.

III. The Dataset

Data Overview and Basic Features

As aforementioned, our dataset was obtained from Kaggle with 74,111 properties and 29 columns. Every row was an actual property listed on Airbnb, with posting dates ranging from 2008 to 2017. All prices included are the last observed price on the website, presuming it will be around the time of the last review.

The dataset was extremely comprehensive, with the four buckets outlined earlier in the Abstract. The specific columns included in each buckets are outlined below:

Description	Rental Information	Host Information	Location Information
Property_type	title_length	host_response_rate	city
room_type	descrip_length	host_has_profile_pic	neighbourhood
bathrooms	title_caps		latitude
bed_type	title_sentiment		longitude
bedrooms	desc_sentiment		
beds	instant_bookable		
accommodates	host_identity_verified		
number_of_reviews	cleaning_fee		
review_scores_rating	cancellation_policy		
	first_review		
	host_since		
	last_review		

Preliminary Feature Engineering

Since we are focused on modelling prices of new houses, we did not include a handful of features, including number of reviews, review score rating, host since, first review and last review. However, on top of the existing features from the raw data source, we performed some preliminary transformations to make the data more comprehensive and model-building friendly:

- Exploded the **amenities** from a string encoded list to **one hot encoded column**
- Used **mean-based-imputation** for the small amount of NAs (in accomodation and # of bedrooms) in continuous variables
- Used **majority-based-imputation** for the small amount of NAs in categorical variables
- Extracted **length of title and description** for every listing
- Extracted **title and description sentiments** (using syuzhet package) for every listing
- Extracted **# of capitalized letters** in the Title
- Extracted **frequently used words** from all of the descriptions (top 100)
- One hot encoded all categorical variables (neighborhoods, cities and etc)

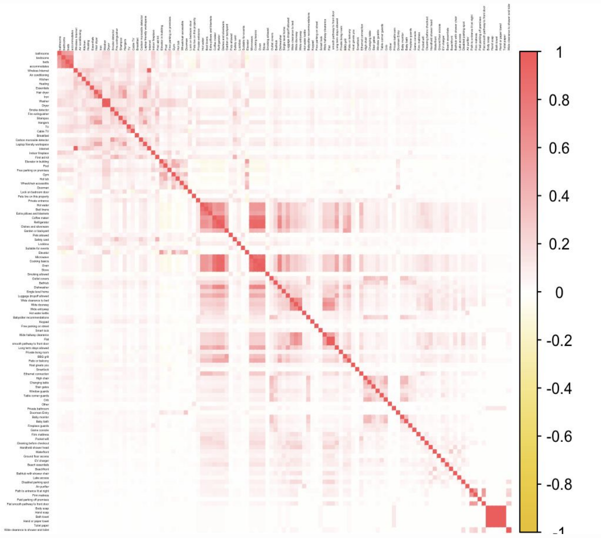
This iteration of data preparation left us with **1046 expanded variables.** This will create serious issues when it comes to making preliminary graphical analysis and hypothesis along with a burdensome modelling process later on.

Thus to achieve a more effective EDA and predictive modelling process, we decided to perform dimension reduction for the main facets of our variables: physical features (200 + variables), words (100 variables), and neighborhoods (600+ variables)

III. The Dataset

Dimension Reduction

Physical Information



We first performed a Principal Component Analysis (PCA) on the physical features of the house (e.g. # of beds and amenities), to reduce unwanted variations. We extracted the top 6 interpretable components (25% of the variations) and labeled these Principal Components based on the variables with the highest loading score.The components nicely aligned into the following 6 interpretable and uncorrelated features:

Labeled Components

- **Cooking** (Influenced by variables such as Refrigerator, Cooking Basics and Stove)
- **Size_and_basics** (Influenced by variables such as accommodates and beds)
- **Comfort_and_space** (Influenced by variables such as smooth pathway to front door and wide doorway)
- **Ease_of_access** (Influenced by variables such as Elevator and Wide Entry Way)
- **Bath_and_Toiletries** (Influenced by variables such as Hand Soap and Bath Towels)
- **Privacy** (Influenced by variables such as Private Rooms and Lock on bedroom door))

Descriptive Words

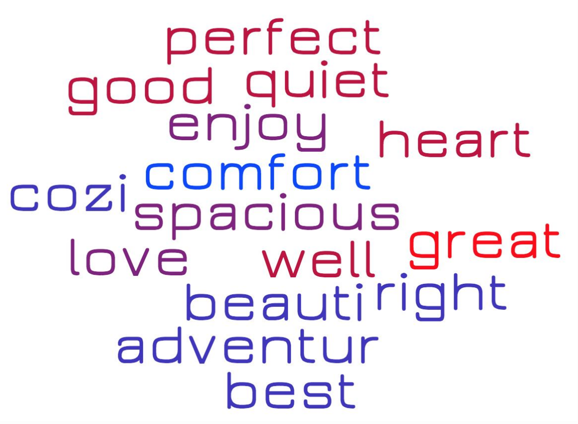
To better condense the linguistic based variables, we also performed a PCA on words, specifically looking for highly concurrent words within the property description. We ended up extracting the top two interpretable components which accounted for 15% of the variations. The components are listed below:

Labeled Components

- Room Based Description (words include: beds, rooms and etc)
- Location Based Description (streets, apart, parks)

We also tabulated some common descriptive words that were among the 100 most frequently appearing set of words. For each listing, we created an index based on the number of descriptive words present in each description. The selected descriptive words are listed below:

Descriptive Words



In doing so, we condensed our semantic based features **from 100 down to only 3**.

III. The Dataset

Neighborhood

To effectively collapse the dimension of neighborhoods, we had to find a way to creatively summarize relevant information contained by the "Neighborhood".

By the end, we came up with 2 ways to numerically transform the previously categorical representation.

The first method was to represent **the neighborhood by its distance to the city center.** This contains information that is more closely related to geography. We presume that city-center is likely to be the area that is most crowded/expensive. Conversely, suburban areas should be presumably less "busy" and "hot".

The specific city center that we have chosen are as follows:

City Based Coordinates
DC 38.8977° N, 77.0365° W **(WHITE HOUSE)**
NYC 40.7527° N, 73.9772° W **(GRAND CENTRAL)**
SF 37.7946° N, 122.3999° W **(FINANCIAL DISTRICT)**
BOS 42.3557° N, 71.0572° W **(DOWNTOWN)**
LA 34.0687° N, 118.3228° W **(CENTRAL LA)**
Chicago 41.8786° N, 87.6251° W **(THE LOOP)**

The second method was to represent neighborhood **in a singular number containing number of previous listing (active listing posted prior to a new posting) in the corresponding neighborhood.** This serves as a proxy for the demand or simply the usage of Airbnb in the area, which is independent from its actual location. We believe this could be a good way to supplement the distance formula above since geographical locations or distance to city center can not be a catchall for popularity of a location in and of itself.

Final Filtering

As we prefaced in the abstract, we chose Airbnb as our dataset since we believe that listings on the site are likely to be priced at market value. However, we don't believe all listings on Airbnb are necessarily priced this way. This assumption exists solely for listings that have 1) been on the market for a while, and 2) been through transactions. As opposed to getting it right from the start, these listings have likely undergone iterations of corrections to land on the price they are at today.

As such, from here onwards, we filtered out houses with **fewer than 3 reviews** and placed them into a holdout set. These listings are filtered due to their low presumed transaction volume. There is no guarantee that these houses are priced at market condition, since they have not been frequently transacted. By excluding these properties, our dataset includes only houses we believe to be priced according to the market.

Final Dataset

Location Information

Two dimension reduced variables that describes the location of the house along with five dummy variable for the housing city.

Physical Information

Six dimension reduced variables that describe the physical nature of the house. These variables include features that are difficult to change in terms of cost and time.

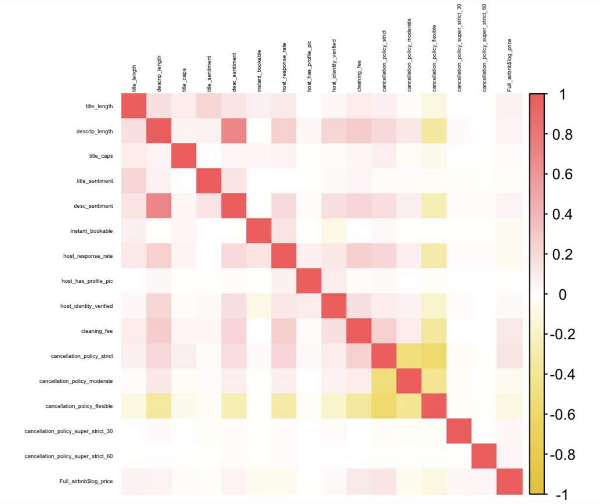
User Based Information

22 variables that include amenable information about the house, including sentiment and words in the title and description, fees, host identity and response time.

IV. EDA

User Based Features

We will start this analysis by exploring some pairwise correlations that exists among user-based features.



Some general associations that can quickly be observed: **Cancellation related Policies** are mutually negative as you can't really have more than one kind of cancellation policy**. Description Length is strongly linked with sentiments,** likely due to repeated use of positive words in the listing description. **Cleaning fee, description length and host response rate** are closely associated - we think this may be a proxy for the degree of host attention and engagement. For example, more engaging or "caring" hosts are likely to be more concerned with areas listed above

We also dove into the relationship of these features and our target variable, arriving at three main observation and hypothesis.

Observation 1

Interestingly, having a **high cleaning fee** and **strict cancellation** policy seem to **positively relate with price,** while having a flexible policy actually negatively correlates with the price

Hypothesis 1: *There may be some external factors carried over in those factors that serve as proxies for the "class" of the house.* ***We hypothesize this trend to diminish as we control for factors like amenities.***

Observation 2

Title length, description length and description sentiment all have rather **positive relations** with the housing price as well

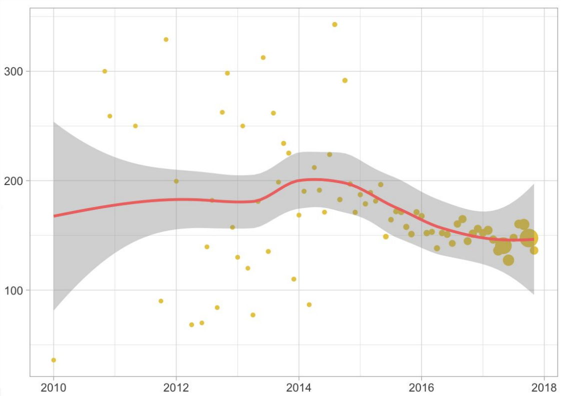
Hypothesis 2: *We believe that this may be a sign that the wording attached with the house may possibly affect housing prices.* ***We hypothesize that the effect will be positive due to the power of marketing.***

Observation 3

Overall, while there seems to some association seen between the user based features and the housing price, **they don't appear to be strongly correlated**

Hypothesis 3: *We believe that this may shed light to our second problem statement.* ***We hypothesize that user based features will not be significant under a linear setting but may be significant in non-linear models, such as ensembles or trees.***

Prices throughout the years

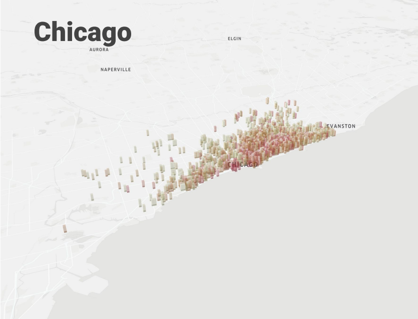


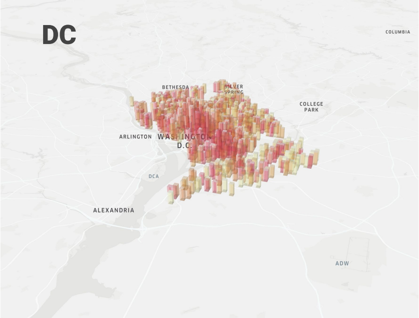
Renting prices appear rather unstable prior to 2015, likely due to the fact that the samples (or Airbnb listings) were significantly smaller by then as well. As prices stabilize, the housing price actually showed a small, **decreasing trend as the years progressed.**

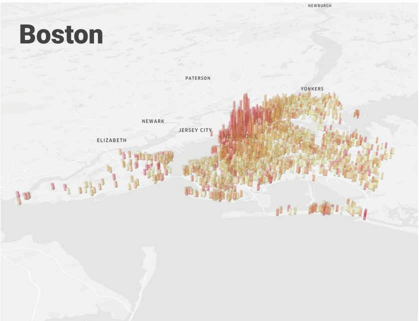
We are interested in diving deeper through the years where the prices were stable in order to better understand whether this change in price is solely affected by **market trends instead of changing consumer preference.** We hypothesize that both may have been at play.

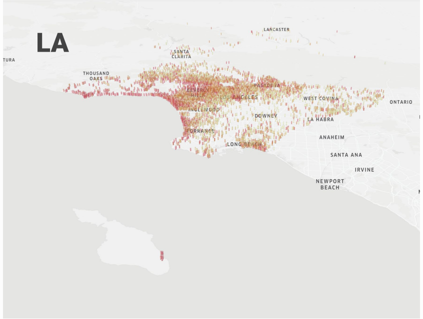
IV. EDA

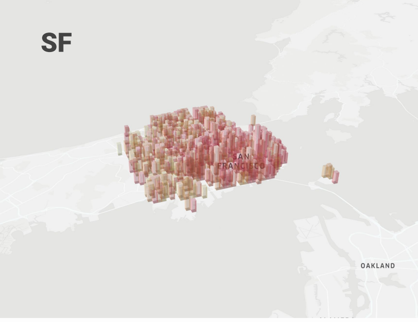
Listings and Geographical Locations

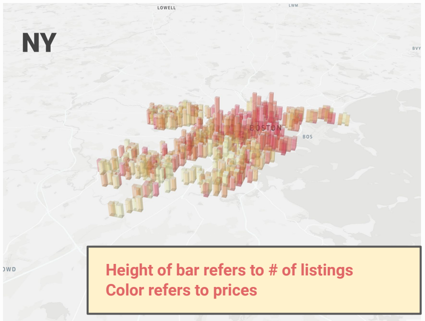












Height of bar refers to # of listings
Color refers to prices

We wanted to further explore the intersection of geographical locations with the overall Airbnb Ecosystem in terms of pricing and listing.

Listing

On the listing side, it appears that while LA has the smallest density of listings, it has the widest reach across the board all over the city. San Francisco and Washington D.C. each has a wide spread of listings across whole areas with not one particular zone standing out tremendously

New York and Boston, on the other hand, are much more concentrated in terms of where the listings tend to be located - it is much denser at the center of these cities and sparse on the outskirts.

Pricing

On the most, pricing partly echoes the trends for listings. For areas like San Francisco and Washington D.C., housing prices are rather evenly distributed across most regions, while it's much more expensive in the central areas of New York and Boston.

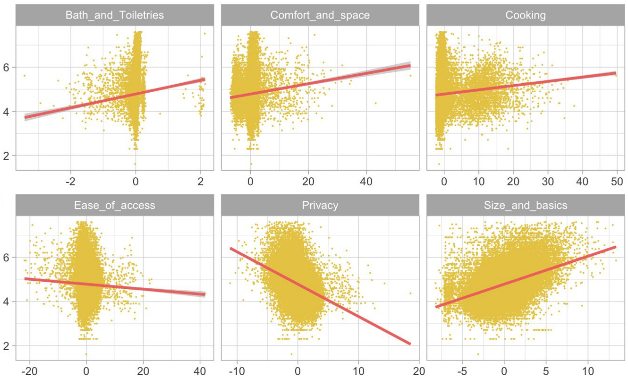
Chicago and Los Angeles, on the other hand, appear to have more coastal listings, where prices seem significantly more expensive compared to those in the inner part of the cities.

Overall, while there does exist city wide similarities for specific city pairings like LA and Chicago (Coastal Zone), SF and Boston(Spread), and, NY and Boston (Dense Central Areas), **we hypothesize that geography does have an impact on Airbnb rental prices.**

IV. EDA (continued)

Physical Principal Components

We then proceeded to explore the relationship of the different physical principal component and prices.



Overall, it seems like there are varying relationships between these components and the actual housing price. In line with that, we formulated two main observation and hypothesis pairing from this relationship

Observation 1

Bathroom, space, cooking and size related physical features appeared rather strongly & positively correlated with housing price, appearing as positive slopes with small confidence interval.

Hypothesis 1: *This seems rather likely, given that all those features would seem rather crucial for any renter and are indeed useful amenities that makes the house more complete and appealing to live in.* ***We hypothesize that the above features will remain positive in later analyses.***

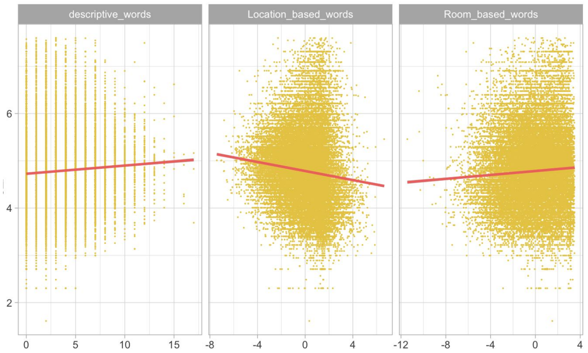
Observation 2

Surprisingly, **ease of access and privacy in particular** seem to be features that negatively relate to the price of interest.

Hypothesis 2: *We hypothesize that* ***this relationship may be due to the trade-off attached to these choices.*** *Perhaps higher privacy may indicate that this is a shared room or a smaller compartment, causing this trend to be negative. We hypothesize that this trend could be different after controlling for other variables.*

Word Based Indexes

Next we went ahead and explored the relationships of the different semantic features we engineered and examined how they related to housing prices



The effects seem less pronounced compared to those of the physical components; however, the directions of the effect are quite interesting. We also formulated two more observation and hypothesis pairings in light of this.

Observation 1

Descriptive Words and Room Based Words both seems to have positive relationships with housing prices.

Hypothesis 1: ***We hypothesize that this trend will persist*** *- our rationale is that descriptive words and room based words are more appealing to the intrinsic needs and wants of the customers.*

Observation 2

Location based words seem to have a negative relationship with housing prices.

Hypothesis 2: *while the trend is quite strange, we believe this trend will also persist. Our rationale is that location based words in general* ***don't add additional value*** *to the house in most cases but* ***may instead significantly decrease value*** *if the location of the property appears suboptimal.*

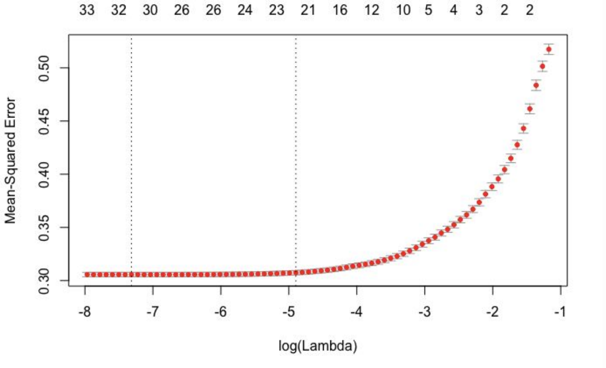
V. Models

Models

We experimented with four higher level machine learning models to develop the best algorithm for price predictions.

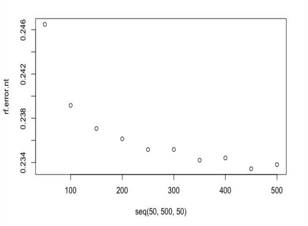
Relaxed LASSO: building on top of a regular regression, LASSO regression works by adding bias (penalizing coefficients) to create a sparser model to better estimate testing errors. However, the implemented bias may reduce the ability to predict. Relaxed lasso resolves this by applying regular regression to the shrunken variables from LASSO to construct the final model.

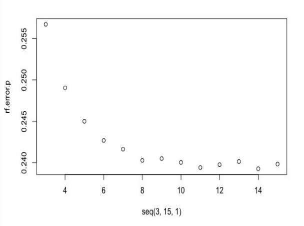
We performed a 10-fold Cross Validation to tune **lambda to perform the feature selection,** and fitted the final linear regression model based on the selected variables.



Random Forest: Random Forest is an algorithm that generates many uncorrelated deep decision trees to preserve low bias while decreasing variance through averages. Random forest has historically performed well in terms of prediction accuracy

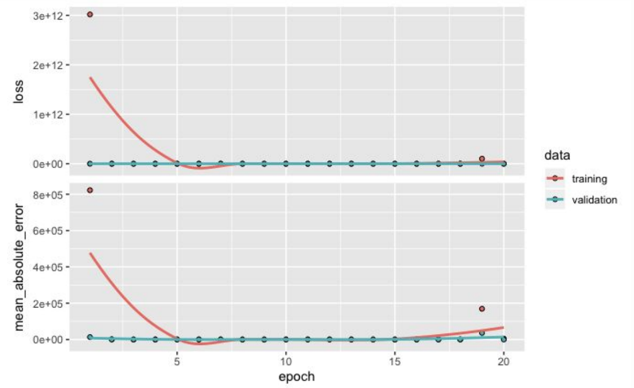
We used 10-fold Cross Validation to tune the mtry and number of trees parameter of RF. The final model has 14 as the mtry parameter and 450 as the ntree parameter. The out of bag, mean squared prediction error was 0.483 on the logged price.





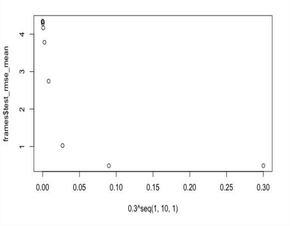
Neural Network: The new and popular algorithms designed to recognize patterns, interpreting sensory data by labeling or clustering raw inputs. Neural network is now the golden standard in the industry in terms of prediction ability.

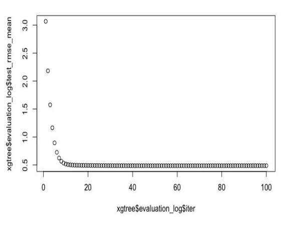
We used a validation set along with training data to tune the final model. The CV Mean Absolute Error was 104 on the actual price of the house.



XGBoost: XGBoost is an implementation of gradient boosting for decision trees algorithm, optimized for speed and performance. A gradual, additive model is generated, with decision trees added to minimize the loss function.

We used 10-fold Cross Validation to tune the nrounds and learning rate parameter of XGBoost. The final model has 14 as the mtry parameter and 450 as the ntree parameter. The XGB has a CV MSE Prediction Error of 0.486 on the logged price





All the models were then evaluated in the final validation set of 5000 observations that was split from the training data at the very start of the process.

V. Results

Test Results

The final evaluation on the test data was evaluated on the mean squared error and the mean absolute error of the actual housing price. The results are:

Model	MAE	MSE
Relaxed Lasso	**$75**	**$188**
Random Forest	**$62**	**$136**
Neural Network	**$102**	**$187**
XGBoost	**$63**	**$138**

From the above, we find that random forest is a more accurate predictor than XGBoost, which is more accurate than relaxed LASSO, which in turn is more accurate than the Neural Network model.

The results in general seems to restrict the price to a rather smaller range than what Airbnb normally suggests. We see this as an improvement, directly resulting from our model.

Furthermore, we wanted to not only more quantitatively investigate the different factors affecting housing price, but also verify our second objective.

Important Variables

The Physical components do appear superior in terms of significance. In fact, both Random Forest and XGBoost ranked the physical component at the top seven in information gain.

Location variable were significant according to LASSO and the information gain criteria of Random Forest. This result not consistent when using boosting tree.

Selected user based features of sentiments and identities did appear significant in LASSO selection and the ensemble models. **Furthermore, we tested the F-statistics for the model with and without user based feature and obtained an F Score of 84, and a p-value of <0.001, leading us to conclude that user based feature does matter.**

Specific Feature Findings

We also found some specific information regarding interesting features that we have observed:

Houses with more amenities and perks are pricier to rent: it's like comparing furnished with unfurnished apartments: the more completed and in someway invested the house, the more valuable homeowners believe their property is.

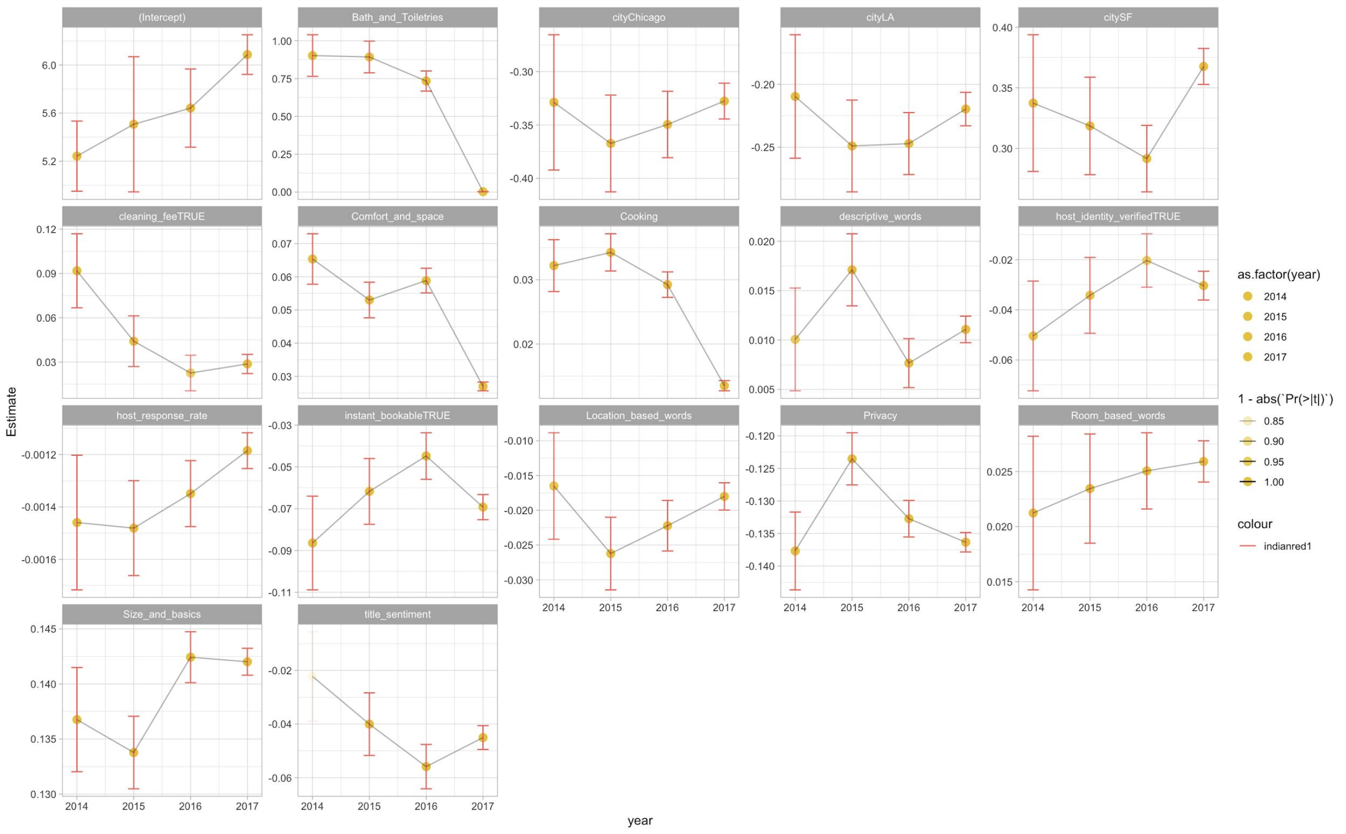
San Francisco appears to be the most expensive city to rent houses in: Housing prices in SF are some of the most expensive in the world. This residual also then resulted in this city having higher rent prices to compensate

People prefer long title and high description sentiments: These in general appeals more to public's mindset when making a purchase speaking to the effect of marketing on consumers

Size and basics, privacy, and bath and toiletries are the 3 most important variables: Customers prioritize openness and completion of houses so they don't get bored or have to shop. Privacy also does seem to be directionally negative, similar to our hypothesis earlier.

Overall, these findings generally adhere to our hypothesis in the exploration slides. Details can be found in the appendix.

VI. Interpretation



Rationale

To answer our third objective, we performed a **regular regression on the data partitioned by the year** when the house was last reviewed.

We are specifically interested in seeing whether or not consumer preferences have changed with regards to particular domains/features of the house.

The graphic above presents factors that we've compiled, which show a significant effect throughout the four years. This analysis led to four main observation; each of which we provided with our own interpretation.

Observation 1

Amenities such as **Bathroom and Toiletries, Comfort and spaciousness, and Cooking** affiliated features have all **gone down in effect** over the past four years, with particular emphasis on 2017.

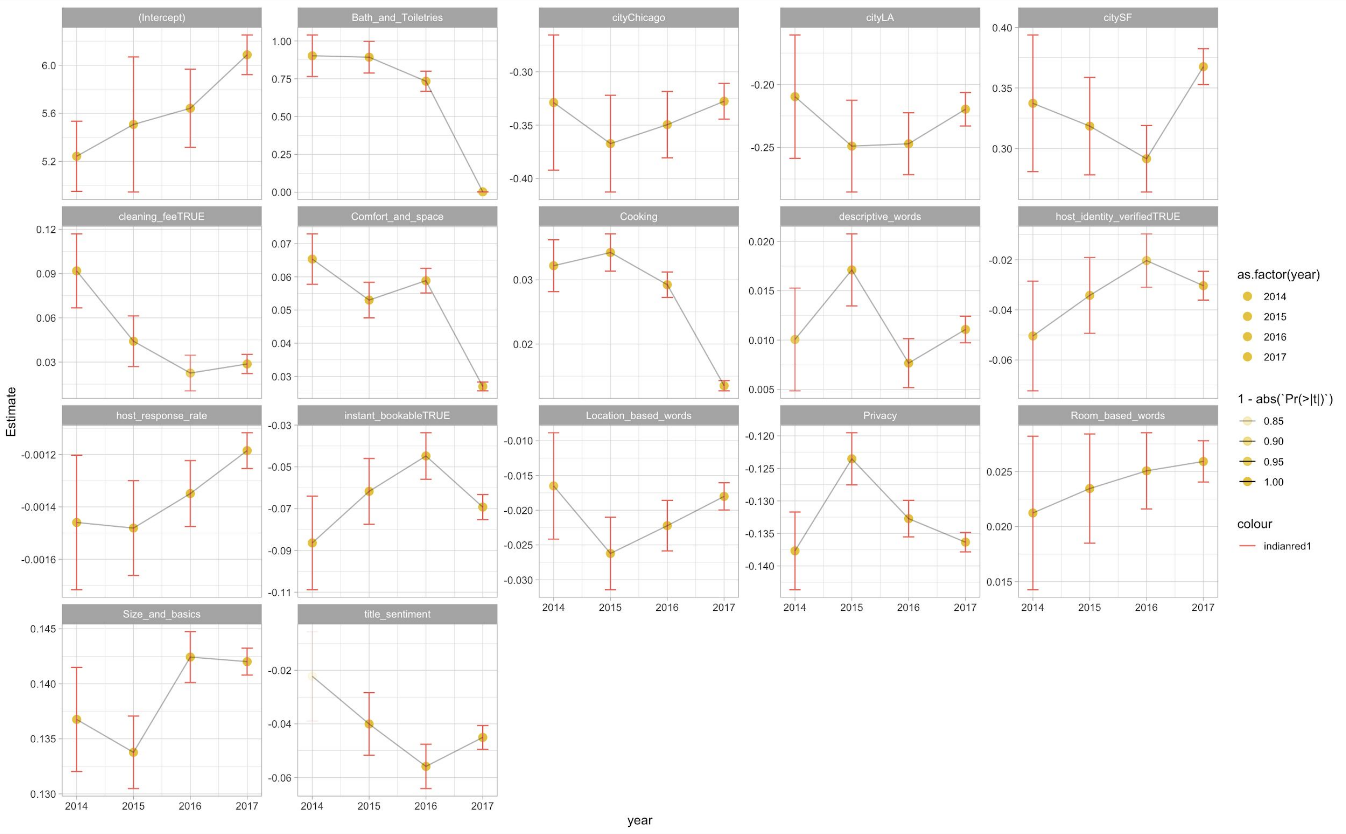
Interpretation 1: We believe this may be due to the completeness of these amenities in more houses, **diminishing its uniqueness in value.**

Observation 2

Chicago, LA, and SF prices have all gone through a parabolic shift, peaking in 2017.

Interpretation 2: We believe this may be explained by the **reversal of rental saturation**, starting with an increase in listing, which pushed down prices, and matures the price is backup as a result of growing supply and demand.

VI. Interpretation



Observation 3

Host identity, host response rate, and bookability have all **increased** in effect over the past years.

Interpretation 3: This could be explained by **the increasingly valued factor of convenience** in the context of electronic services - people want things quickly.

Observation 4

Size and basics also increased in effect, with a big jump from 2015 to 2016 and is parallel to what we observe for people less inclined to opt for private houses.

Interpretation 4: The use of Airbnb **has evidently grown and has been more frequently used in different regards, such as bigger party travels**. We believe diversity in usage contributed strongly to this observed trend, with an increasing emphasis on house size.

Overall

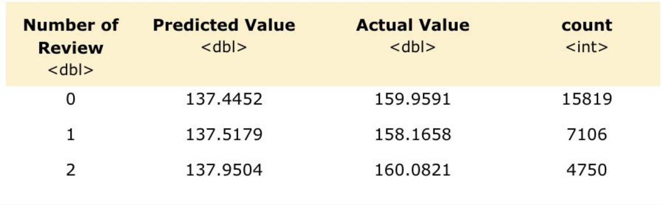
Putting all the observations and interpretation together, we believe that there have been some interesting macro changes in consumer behavior with regards to housing preferences. We think these trends are generally adhering to what we expect the market to be, and the trends will likely to continue in the long term. That is to say that basic amenities are going to start losing more effect as they become more commonplace, while host response rate and the like are likely going to increase in importance as time goes by.

We **don't see physical information losing value** as it still remains the main component of the house. As prefaced, however, **the user decided feature may become just as important as other features, such as amenities.**

VII. Conclusion

Impact Assessment

We will quickly examine how our model could play out in the real world by trying to predict the value of the houses on the holdout set, which includes the set of houses that have fewer than three reviews. The results obtained are as follows:



Number of Review <dbl>	Predicted Value <dbl>	Actual Value <dbl>	count <int>
0	137.4452	159.9591	15819
1	137.5179	158.1658	7106
2	137.9504	160.0821	4750

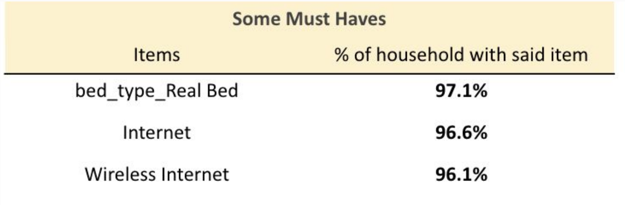
Evidently, for these houses, the actual value for all listed seems to be higher than their predicted value by about $20. We believe this may have strongly attributed to these houses not performing well on the market. With this algorithm, we believe this situation could be remedied and fixed quite easily.

Conclusion.

While there are some key features (i.e. size and basics) that are inflexible and cannot be changed. To conclude our analysis, we want to offer a couple of advice regarding how to optimize for the value of a house:

1. Always offer basic amenities
2. Dedicate time into writing a good description

Because we performed this analysis with PCA, only variables with high variance are chosen. However, we want to note that there are some features that are fundamentally necessary to own in a house, such as:



Some Must Haves	
Items	% of household with said item
bed_type_Real Bed	**97.1%**
Internet	**96.6%**
Wireless Internet	**96.1%**

While these features are not mentioned in the analysis before, we believe they are nonetheless essentials for a listing, simply because it's become so integrated in people's everyday lives.

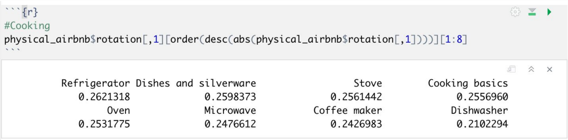
Next Steps and Recommendations:

We believe this report offers a solid general framework on approaching the renting price modelling issue; however, we do think it can be further improved.

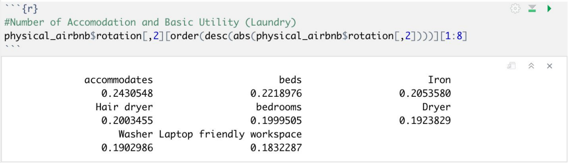
We recommend future studies to look into including more features that more precisely represent geographical locations - such as number of houses nearby, and residential vs commercial - to provide more insight to the state of the district.

We believe that it will also be a good step to analyze what could be done to decompose the image attached with each listing. We believe the quality of the picture is quite essential to how a customer perceives the listing. Including some information on the image could help with both the predictive power and the recommendation that this paper seeks to provide.

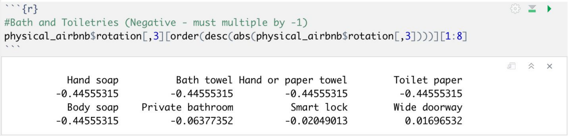
VIII. Appendix



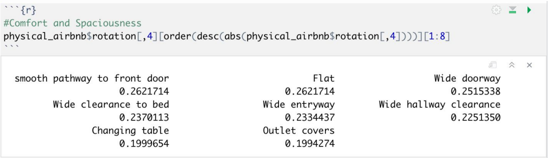
~~~
```{r}
#Cooking
physical_airbnb$rotation[,1][order(desc(abs(physical_airbnb$rotation[,1])))][1:8]
```

        Refrigerator Dishes and silverware                 Stove        Cooking basics
           0.2621318             0.2598373             0.2561442             0.2556960
                Oven             Microwave          Coffee maker            Dishwasher
           0.2531775             0.2476612             0.2426983             0.2102294
~~~



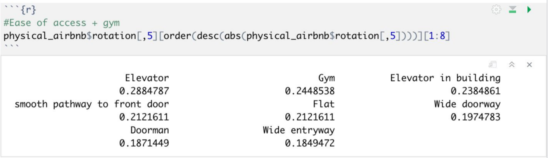
~~~
```{r}
#Number of Accomodation and Basic Utility (Laundry)
physical_airbnb$rotation[,2][order(desc(abs(physical_airbnb$rotation[,2])))][1:8]
```

            accommodates                       beds                       Iron
               0.2430548                  0.2218976                  0.2053580
              Hair dryer                   bedrooms                      Dryer
               0.2003455                  0.1999505                  0.1923829
                  Washer Laptop friendly workspace
               0.1902986                  0.1832287
~~~



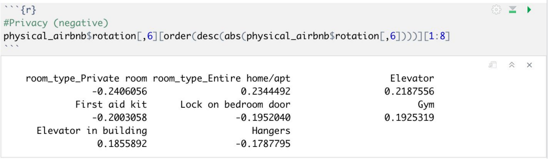
~~~
```{r}
#Bath and Toiletries (Negative - must multiple by -1)
physical_airbnb$rotation[,3][order(desc(abs(physical_airbnb$rotation[,3])))][1:8]
```

          Hand soap          Bath towel Hand or paper towel        Toilet paper
        -0.44555315         -0.44555315         -0.44555315         -0.44555315
          Body soap    Private bathroom          Smart lock        Wide doorway
        -0.44555315         -0.06377352         -0.02049013          0.01696532
~~~



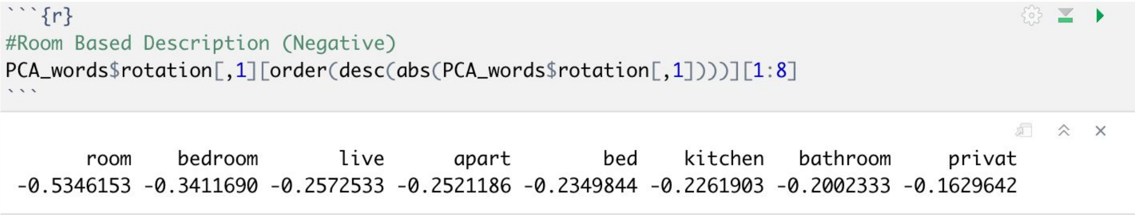
~~~
```{r}
#Comfort and Spaciousness
physical_airbnb$rotation[,4][order(desc(abs(physical_airbnb$rotation[,4])))][1:8]
```

 smooth pathway to front door                         Flat                 Wide doorway
                    0.2621714                    0.2621714                    0.2515338
       Wide clearance to bed                 Wide entryway      Wide hallway clearance
                    0.2370113                    0.2334437                    0.2251350
               Changing table                Outlet covers
                    0.1999654                    0.1994274
~~~



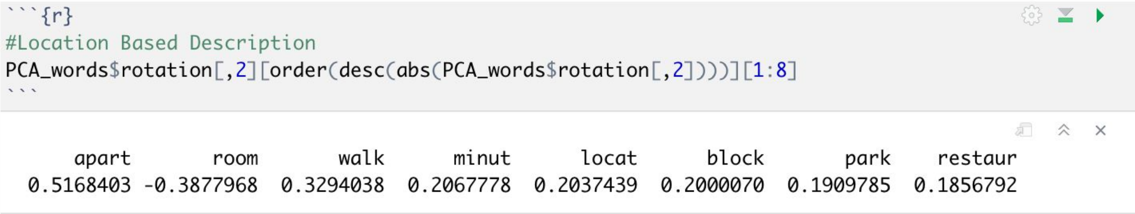
~~~
```{r}
#Ease of access + gym
physical_airbnb$rotation[,5][order(desc(abs(physical_airbnb$rotation[,5])))][1:8]
```

                     Elevator                          Gym       Elevator in building
                    0.2884787                    0.2448538                    0.2384861
 smooth pathway to front door                         Flat                 Wide doorway
                    0.2121611                    0.2121611                    0.1974783
                     Doorman                 Wide entryway
                    0.1871449                    0.1849472
~~~



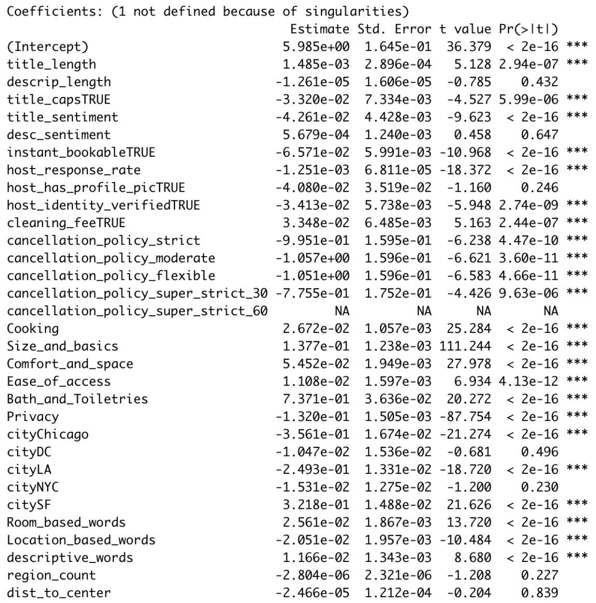
~~~
```{r}
#Privacy (negative)
physical_airbnb$rotation[,6][order(desc(abs(physical_airbnb$rotation[,6])))][1:8]
```

  room_type_Private room room_type_Entire home/apt                  Elevator
              -0.2406056                 0.2344492                 0.2187556
           First aid kit      Lock on bedroom door                       Gym
              -0.2003058                -0.1952040                 0.1925319
    Elevator in building                   Hangers
               0.1855892                -0.1787795
~~~



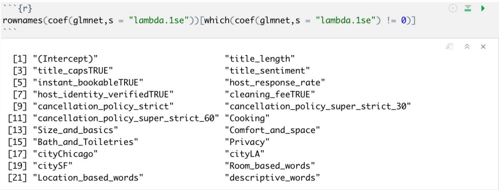
~~~
```{r}
#Room Based Description (Negative)
PCA_words$rotation[,1][order(desc(abs(PCA_words$rotation[,1])))][1:8]
```

      room    bedroom       live      apart        bed    kitchen   bathroom     privat
-0.5346153 -0.3411690 -0.2572533 -0.2521186 -0.2349844 -0.2261903 -0.2002333 -0.1629642
~~~



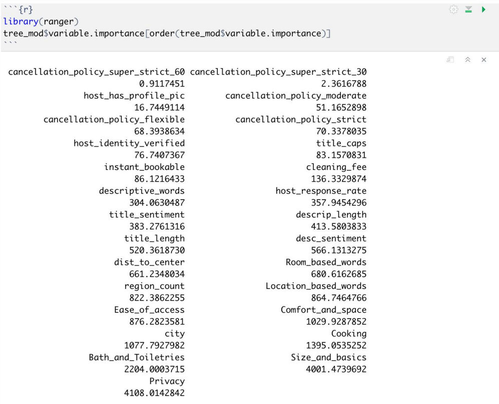
~~~
```{r}
#Location Based Description
PCA_words$rotation[,2][order(desc(abs(PCA_words$rotation[,2])))][1:8]
```

     apart       room       walk      minut      locat      block       park    restaur
 0.5168403 -0.3877968  0.3294038  0.2067778  0.2037439  0.2000070  0.1909785  0.1856792
~~~



```
Coefficients: (1 not defined because of singularities)
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        5.985e+00  1.645e-01  36.379  < 2e-16 ***
title_length                       1.485e-03  2.896e-04   5.128 2.94e-07 ***
descrip_length                    -1.261e-05  1.606e-05  -0.785    0.432    
title_capsTRUE                    -3.320e-02  7.334e-03  -4.527 5.99e-06 ***
title_sentiment                   -4.261e-02  4.428e-03  -9.623  < 2e-16 ***
desc_sentiment                     5.679e-04  1.240e-03   0.458    0.647    
instant_bookableTRUE              -6.571e-02  5.991e-03 -10.968  < 2e-16 ***
host_response_rate                -1.251e-03  6.811e-05 -18.372  < 2e-16 ***
host_has_profile_picTRUE          -4.080e-02  3.519e-02  -1.160    0.246    
host_identity_verifiedTRUE        -3.413e-02  5.738e-03  -5.948 2.74e-09 ***
cleaning_feeTRUE                   3.348e-02  6.485e-03   5.163 2.44e-07 ***
cancellation_policy_strict        -9.951e-01  1.595e-01  -6.238 4.47e-10 ***
cancellation_policy_moderate      -1.057e+00  1.596e-01  -6.621 3.60e-11 ***
cancellation_policy_flexible      -1.051e+00  1.596e-01  -6.583 4.66e-11 ***
cancellation_policy_super_strict_30 -7.755e-01  1.752e-01  -4.426 9.63e-06 ***
cancellation_policy_super_strict_60        NA         NA      NA       NA    
Cooking                            2.672e-02  1.057e-03  25.284  < 2e-16 ***
Size_and_basics                    1.377e-01  1.238e-03 111.244  < 2e-16 ***
Comfort_and_space                  5.452e-02  1.949e-03  27.978  < 2e-16 ***
Ease_of_access                     1.108e-02  1.597e-03   6.934 4.13e-12 ***
Bath_and_Toiletries                7.371e-01  3.636e-02  20.272  < 2e-16 ***
Privacy                           -1.320e-01  1.505e-03 -87.754  < 2e-16 ***
cityChicago                       -3.561e-01  1.674e-02 -21.274  < 2e-16 ***
cityDC                            -1.047e-02  1.536e-02  -0.681    0.496    
cityLA                            -2.493e-01  1.331e-02 -18.720  < 2e-16 ***
cityNYC                           -1.531e-02  1.275e-02  -1.200    0.230    
citySF                             3.218e-01  1.488e-02  21.626  < 2e-16 ***
Room_based_words                   2.561e-02  1.867e-03  13.720  < 2e-16 ***
Location_based_words              -2.051e-02  1.957e-03 -10.484  < 2e-16 ***
descriptive_words                  1.166e-02  1.343e-03   8.680  < 2e-16 ***
region_count                      -2.804e-06  2.321e-06  -1.208    0.227    
dist_to_center                    -2.466e-05  1.212e-04  -0.204    0.839    
```

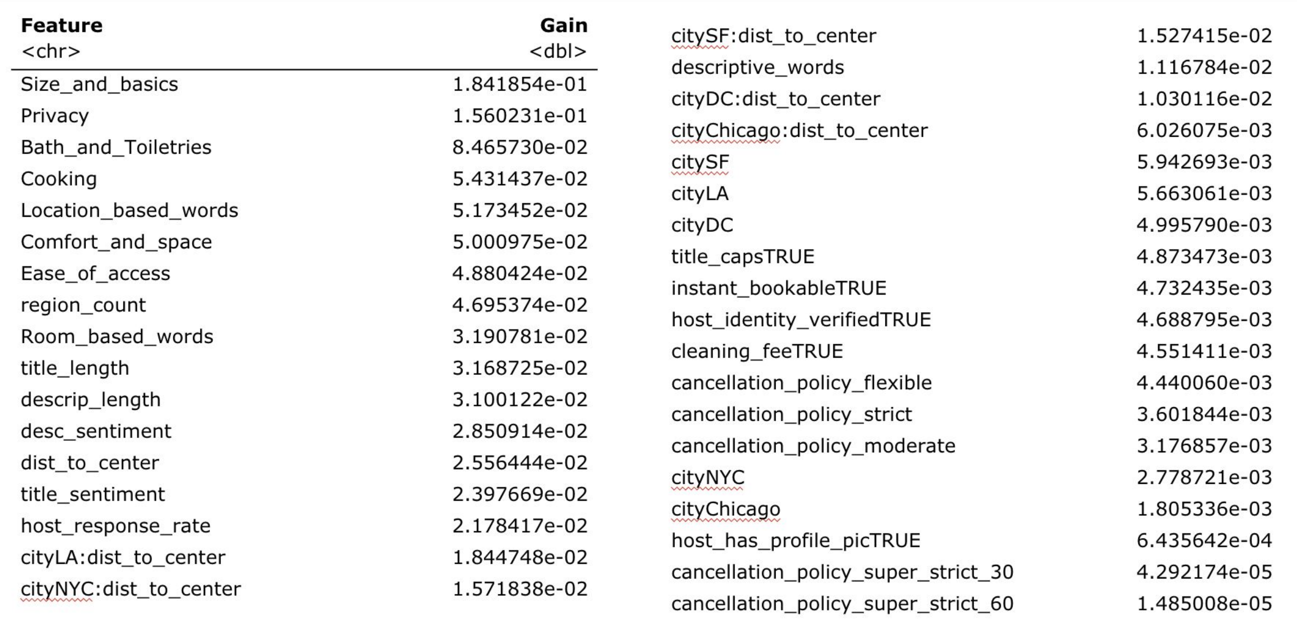


~~~
```{r}
rownames(coef(glmnet,s = "lambda.1se"))[which(coef(glmnet,s = "lambda.1se") != 0)]
```

 [1] "(Intercept)"                         "title_length"                       
 [3] "title_capsTRUE"                      "title_sentiment"                    
 [5] "instant_bookableTRUE"                "host_response_rate"                 
 [7] "host_identity_verifiedTRUE"          "cleaning_feeTRUE"                   
 [9] "cancellation_policy_strict"          "cancellation_policy_super_strict_30"
[11] "cancellation_policy_super_strict_60" "Cooking"                            
[13] "Size_and_basics"                     "Comfort_and_space"                  
[15] "Bath_and_Toiletries"                 "Privacy"                            
[17] "cityChicago"                         "cityLA"                             
[19] "citySF"                              "Room_based_words"                   
[21] "Location_based_words"                "descriptive_words"                  
~~~



~~~
```{r}
library(ranger)
tree_mod$variable.importance[order(tree_mod$variable.importance)]
```

cancellation_policy_super_strict_60 cancellation_policy_super_strict_30 
                          0.9117451                           2.3616788 
               host_has_profile_pic        cancellation_policy_moderate 
                         16.7449114                          51.1652898 
       cancellation_policy_flexible          cancellation_policy_strict 
                         68.3938634                          70.3378035 
             host_identity_verified                          title_caps 
                         76.7407367                          83.1570831 
                   instant_bookable                        cleaning_fee 
                         86.1216433                         136.3329874 
                  descriptive_words                  host_response_rate 
                        304.0630487                         357.9454296 
                    title_sentiment                      descrip_length 
                        383.2761316                         413.5803833 
                       title_length                      desc_sentiment 
                        520.3618730                         566.1313275 
                     dist_to_center                    Room_based_words 
                        661.2348034                         680.6162685 
                       region_count                Location_based_words 
                        822.3862255                         864.7464766 
                     Ease_of_access                   Comfort_and_space 
                        876.2823581                        1029.9287852 
                               city                             Cooking 
                       1077.7927982                        1395.0535252 
                Bath_and_Toiletries                     Size_and_basics 
                       2204.0003715                        4001.4739692 
                            Privacy 
                       4108.0142842 
~~~

VIII. Appendix



Feature <chr>	Gain <dbl>
Size_and_basics	1.841854e-01
Privacy	1.560231e-01
Bath_and_Toiletries	8.465730e-02
Cooking	5.431437e-02
Location_based_words	5.173452e-02
Comfort_and_space	5.000975e-02
Ease_of_access	4.880424e-02
region_count	4.695374e-02
Room_based_words	3.190781e-02
title_length	3.168725e-02
descrip_length	3.100122e-02
desc_sentiment	2.850914e-02
dist_to_center	2.556444e-02
title_sentiment	2.397669e-02
host_response_rate	2.178417e-02
cityLA:dist_to_center	1.844748e-02
cityNYC:dist_to_center	1.571838e-02
citySF:dist_to_center	1.527415e-02
descriptive_words	1.116784e-02
cityDC:dist_to_center	1.030116e-02
cityChicago:dist_to_center	6.026075e-03
citySF	5.942693e-03
cityLA	5.663061e-03
cityDC	4.995790e-03
title_capsTRUE	4.873473e-03
instant_bookableTRUE	4.732435e-03
host_identity_verifiedTRUE	4.688795e-03
cleaning_feeTRUE	4.551411e-03
cancellation_policy_flexible	4.440060e-03
cancellation_policy_strict	3.601844e-03
cancellation_policy_moderate	3.176857e-03
cityNYC	2.778721e-03
cityChicago	1.805336e-03
host_has_profile_picTRUE	6.435642e-04
cancellation_policy_super_strict_30	4.292174e-05
cancellation_policy_super_strict_60	1.485008e-05

IX. Works Cited

"Census of Population and Housing, 2019 [United States]: Summary Tape File 3A." *ICPSR Data Holdings*, October 2019. https://doi.org/10.3886/icpsr09782.v1.

DianaOlick. "Older Millennials Are Driving Home Prices Higher Again." CNBC. CNBC, September 3, 2019. https://www.cnbc.com/2019/09/03/older-millennials-are-driving-home-prices-higher-again.html.

Helsi, and Kkc. "The Largest Non-Official Forum Just for Airbnb Hosts." Airhostsforum.com, May 3, 2017. https://airhostsforum.com/t/local-hosts-underpricing-themselves/13658.