Pricing Optimality

An Airbnb Case Study

STAT 471 Final Project

Grant Cho | Zhangyi Fan | Margaret Zheng

The Opportunity

Why should I Value my House?

airbnb

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

Local hosts underpricing themselves

pricing

L 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?

1 Reply

Housing market is extremely saturated

Adverse selection and asymmetric information results in homeowners constantly dropping their prices

Easy to miss out on profit margins

- Inexperienced entrants will lose out to veterans
- Homeowners 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

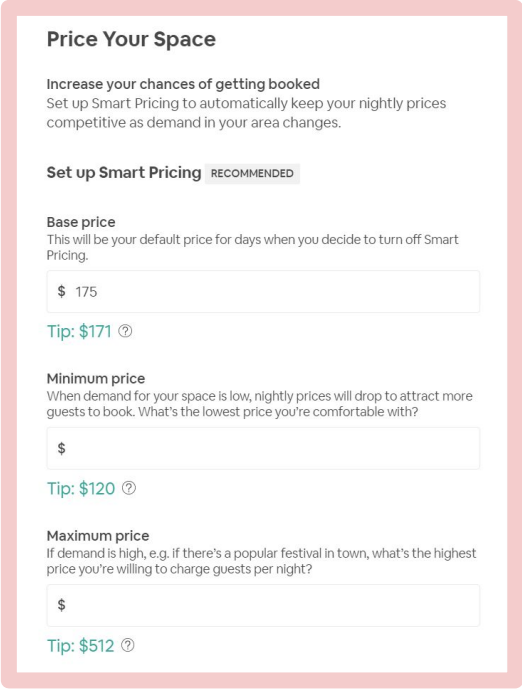
60% of potential housing income: **property, location, and amenities**

40% of housing income: implementing an **effective Airbnb pricing strategy**

Objective 1: Better Approximate Airbnb Prices

Offerings	Issues
Smart Pricing	• Less Ownership Over the Pricing Point • Market-based → No Competitive Edge
Manual Pricing	• Suggested price range is usually extremely wide and imprecise - bringing huge fluctuations to the property value • According to Guesty, *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 to be considered by a customer**
The Challenge	• Can we do better than the offerings Airbnb give right now - **can we provide homeowners with a greater opportunity to best the market?**

Exhibits



Objective 2: Examine Effect of User-defined Features

Competitive Prices

Aside from the pricing point, we are also interested in finding out if there are factors unrelated to the house that can boost your property value.

Challenge: *Can we tweak some of these variables to bring up the competitive pricing of your house?*

Property Description	Microfees	Host Identity	Response Time
Description may affect a customer's valuation of a property	Smaller fees add up to	A verified homeowner's identity results in greater build-up of trust	The speed of the response could affect an opportunities worth

Objective 3: Understand Macro Market Trends

- Customers have started developing a perception that “house prices never fall”
- Housing prices have risen at the third-fastest rate in history over the past several years

- Mortgage rates are increasing
- Majority of home buyers are millennials

The Features

What do people see in my house?

airbnb

Data Overview and Basic Features

Why Airbnb	**Highly Competitive Market with likely correctly priced market value after iterations of testing** **Data Source:** **Dimensions:** 74111 properties and 29 columns | spans across 9 years from 2008 to 2017 in 6 major US cities.

Description	Rental Information	Host information	Location Information
This includes the property type and number of bedrooms	For example: is the rent on a per room basis and/or does it have a cancellation policy?	This includes whether the host is verified or not	Located by city or GPS

Preliminary Feature Engineering

- Description, title length and sentiment (using syuzhet package)
- Title capitalization
- Extracted frequently used words via Text Mining

Dimensionality Reduction: PCA

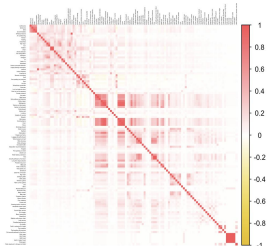
Exploding Dimension

The previous iteration of data preparation left us with **1046 variables**

To achieve a more effective EDA process with fewer variables modeled, we also performed dimension reduction for 3 main facets of our variables, leaving us with **29 final features**

Physical Information (200 + features)

Extracted the top 6 interpretable components (25% of the variations) from Physical features of the house (e.g. # of beds). PCs labeled based on the variables with high loading score



Labeled Components

- Cooking
- Size_and_basics
- Comfort_and_space
- Ease_of_access"
- Bath_and_Toiletries
- Privacy

Description Words (100 + relevant words)

Extracted top 2 interpretable components which accounted for 15% of the variations from words within the property description.
Tabulated the # of descriptive words.

Labeled Components

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

Descriptive Words

Neighborhood (600+ neighborhoods)

Transformed the neighborhood variable into two numerical value from one-hot encoding:

- # of prior listings posted in this area
- Distance of the property to the city center

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)**

Final Data Summary

Final Filtering

Note: From here onwards, we filtered out houses with **fewer than 3 reviews** and placed them into a holdout set. We did this since we don't have any guarantee that these houses are priced at market condition since they are not frequently transacted

Final Data

Location Information

2 dimension reduced variables and 5 dummy variable for the housing city

Physical Information

6 dimension reduced variables

User Based Information

22 variables including amenable information about the house (including sentiment and words in the title & description, fees, host identity and response time)

Next Steps

Final Data Dimension: 46435 x 29

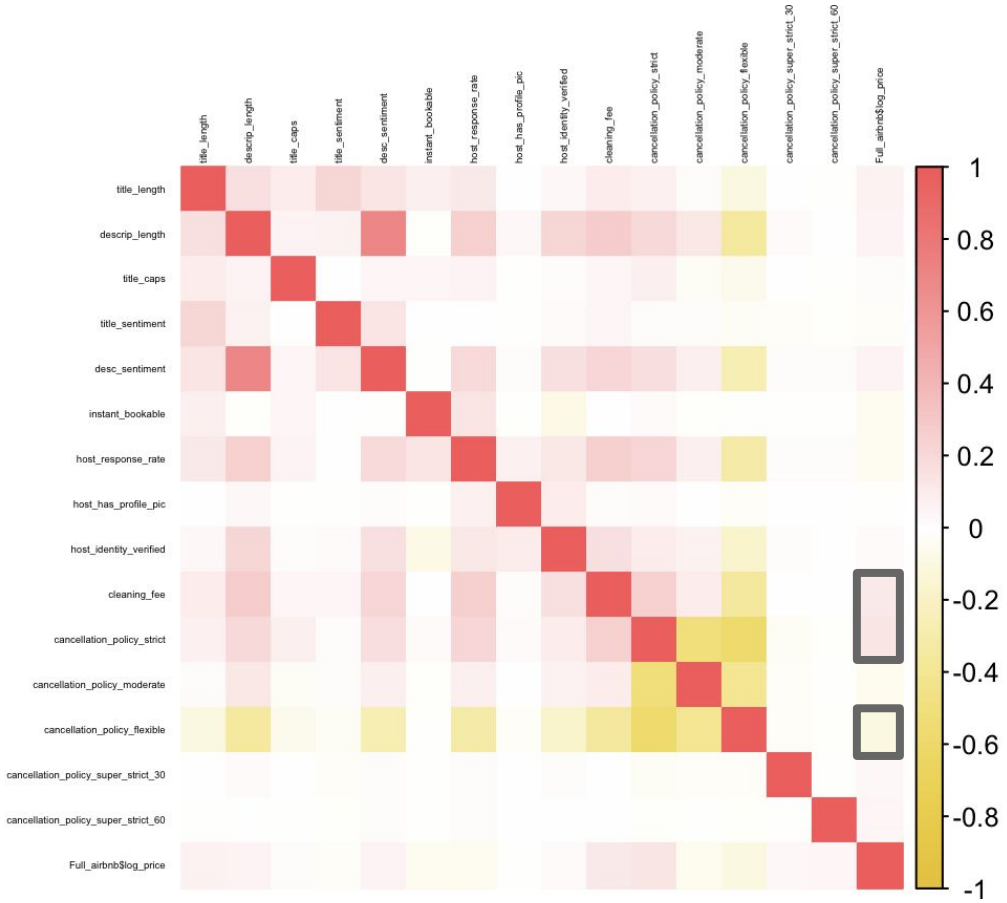
- **EDA**
- **Modelling**
- **Output interpretation**

The Exploration

What relationship do we hypothesize?

airbnb

Correlation Matrix: User-based Features



General Association Trends

- **Cancellation Policies** are mutually negative
- **Description Length is strongly Linked with sentiments** (likely due to repeated use of positive words)
- **Cleaning fee, description length and host response rate** are closely associated (the degree of host attention and engagement)

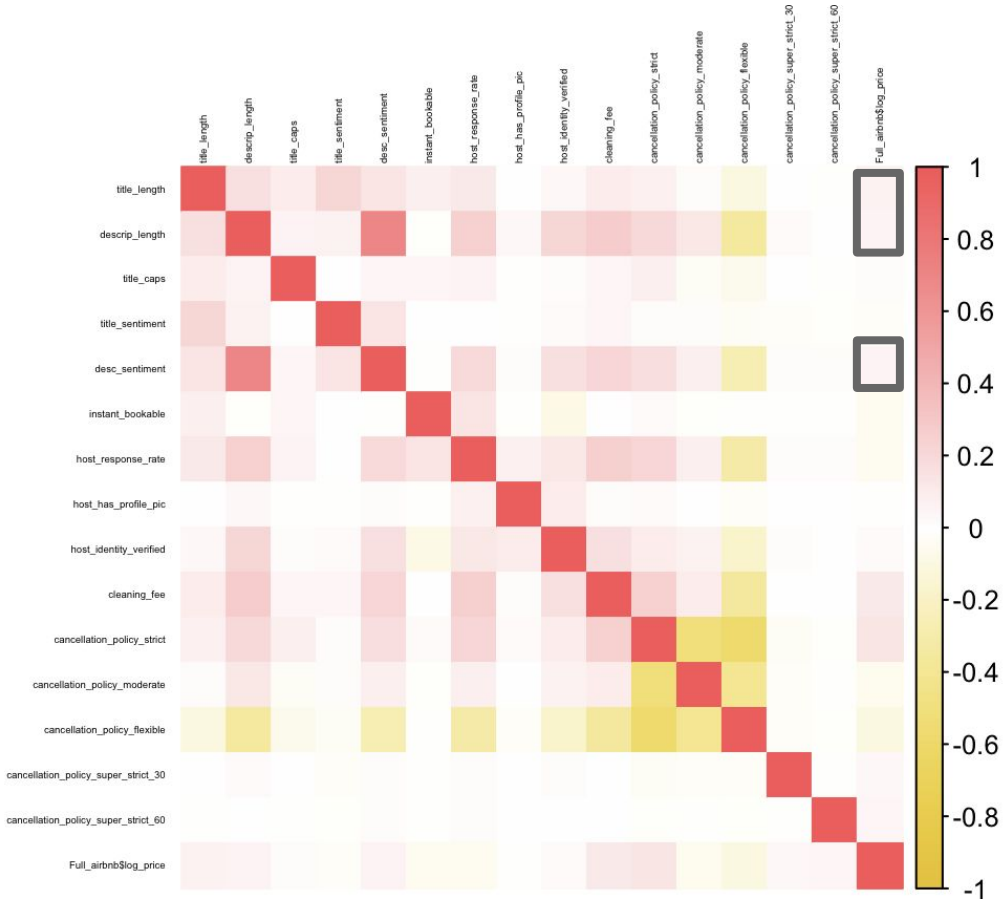
Observation 1

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

Hypothesis 1

There might be some external factors carried over in those factors that serves as a proxy for the "class" of the house. **We hypothesize this trend to diminish as we control for factors like amenities**

Correlation Matrix: User-based Features



General Association Trends

- **Cancellation Policies** are mutually negative
- **Description Length is strongly Linked with sentiments** (likely due to repeated use of positive words)
- **Cleaning fee, description length and host response rate** are closely associated (the degree of host attention and engagement)

Observation 2

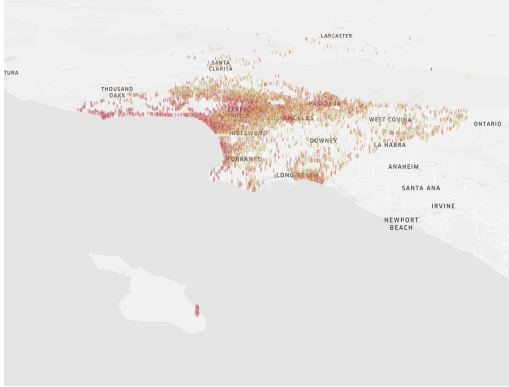
Title length, description length and description sentiment all have rather **positive relation** with the housing price as well

Hypothesis 2

We believe that this signals that the wording attached with the house MAY affect housing prices - **we hypothesize that the effect will be positive due to the power of marketing and glamorization.**

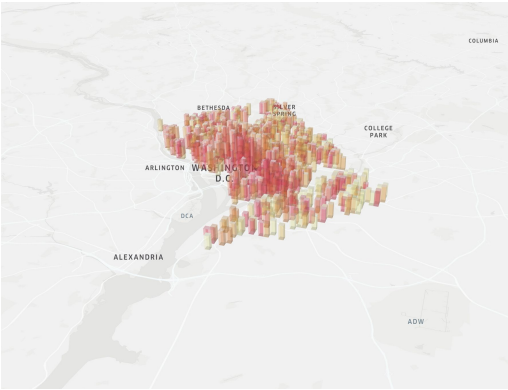
Usage and Pricing of Airbnbs in Different Cities

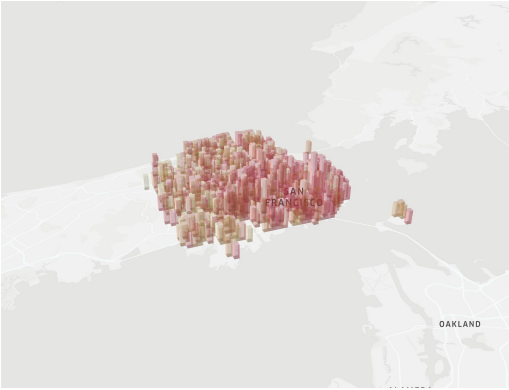
Chicago & LA: Coastal Surge



Usage and Pricing of Airbnbs in Different Cities

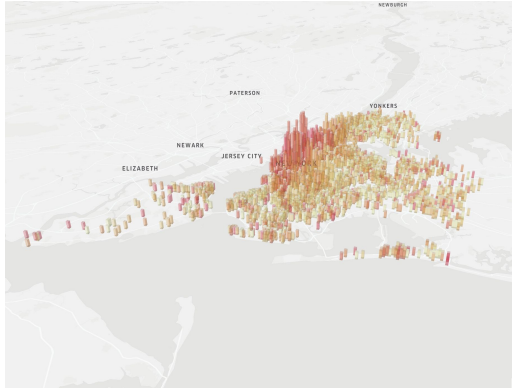
SF & DC: Even Spread and Price

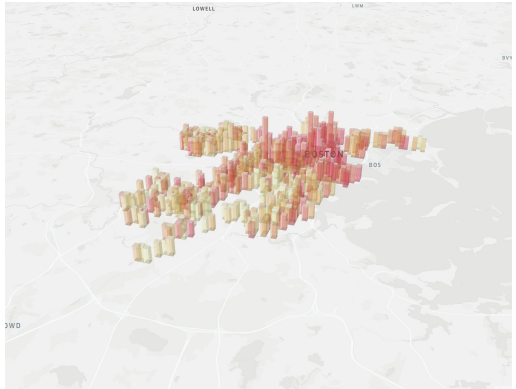




Usage and Pricing of Airbnbs in Different Cities

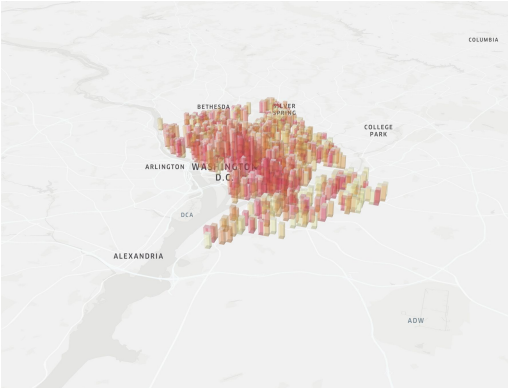
NY & Boston: Dense Central Areas

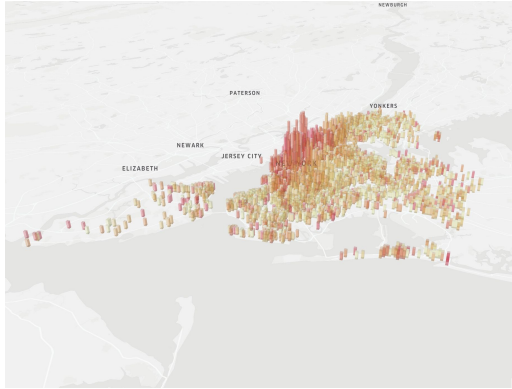


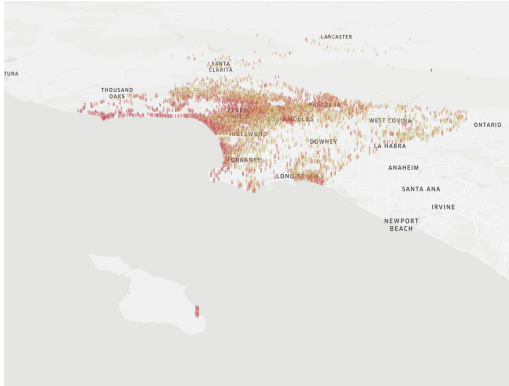


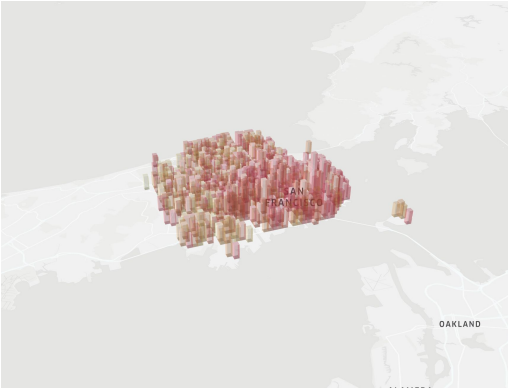
Usage and Pricing of Airbnbs in Different Cities

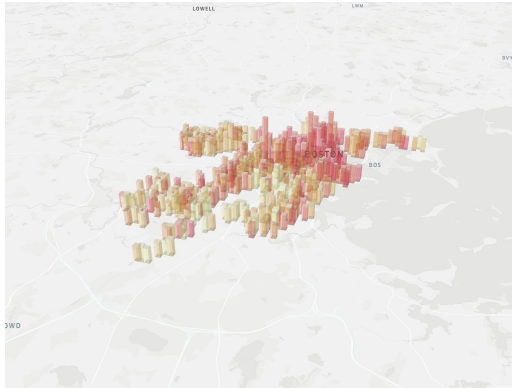
We hypothesize that there exist city wide differences that needs to be mapped out into the model



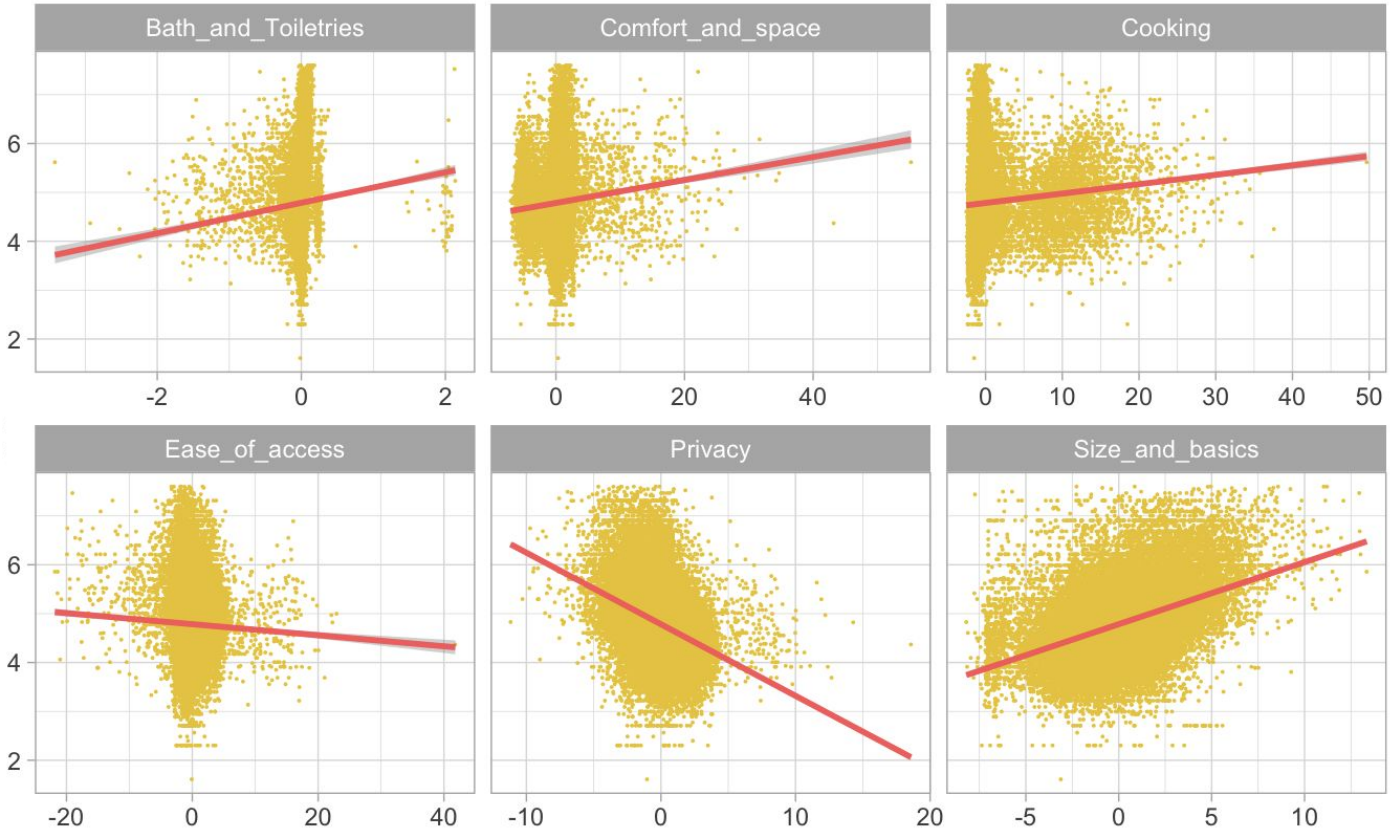






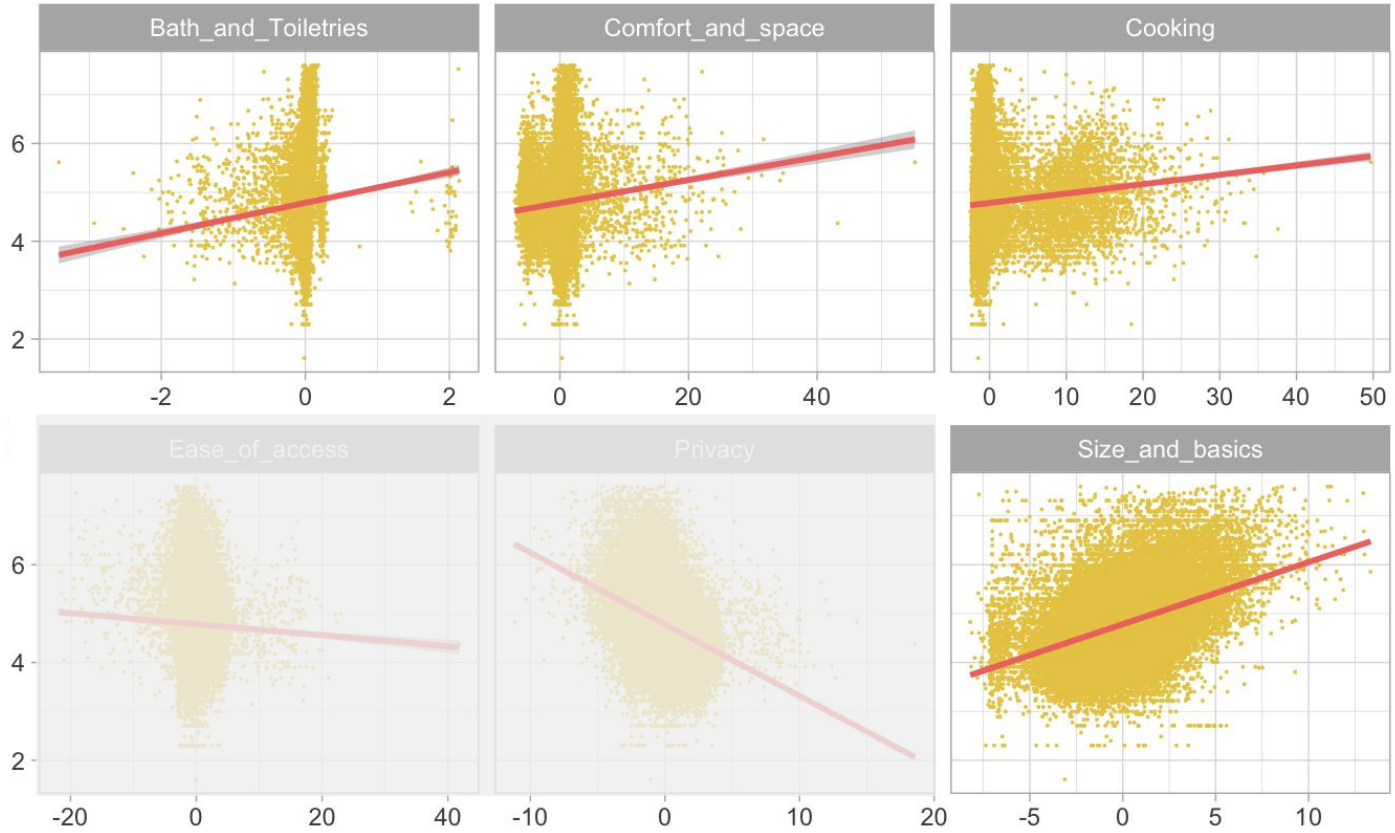


Housing Prices across Physical Principal Components



Housing Prices across Physical Principal Components

Housing Price (Logged) Across Physical PC Index



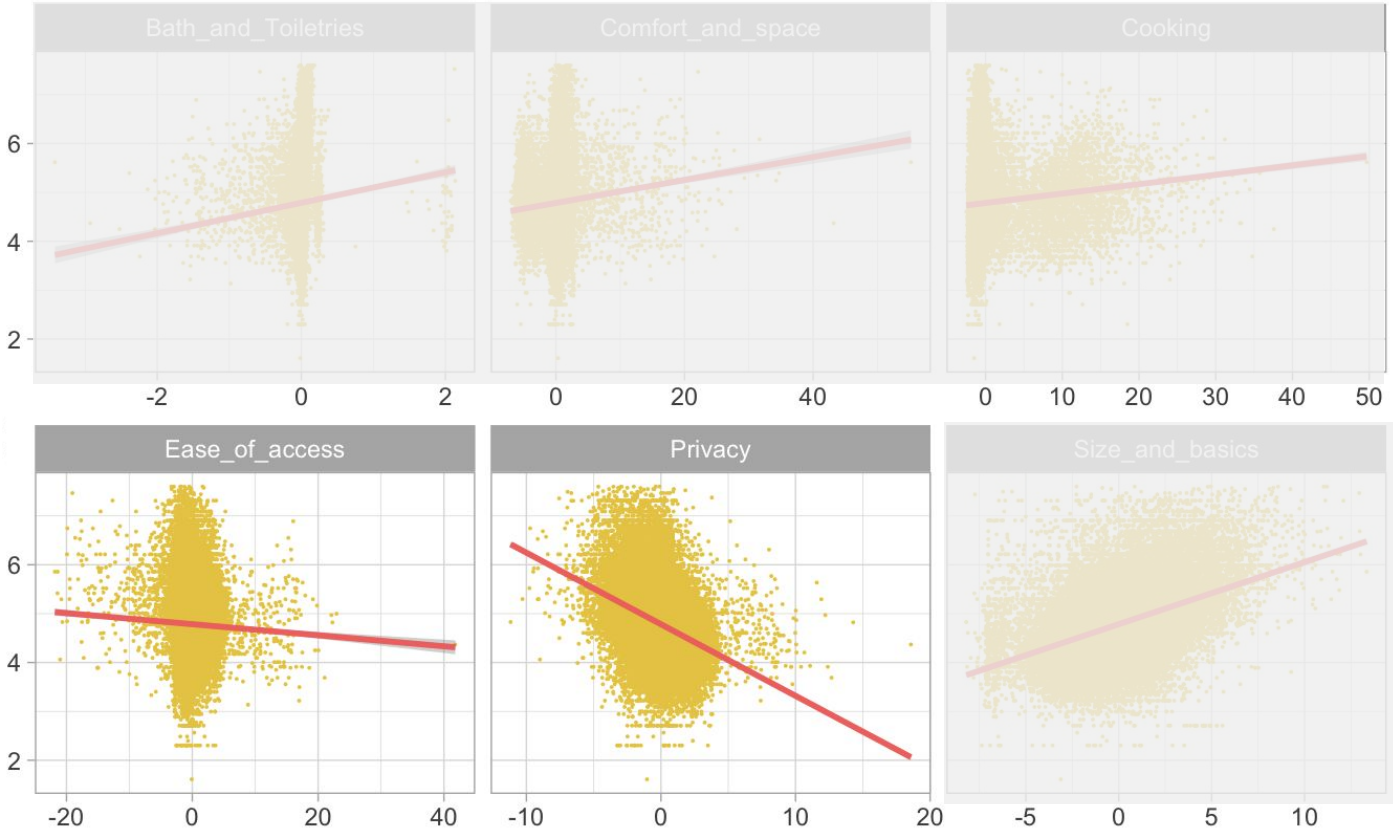
Observations

Bath room, space, cooking and size related physical features appeared rather strongly & positively correlated with housing price

Hypothesis

This seems rather likely, given that all those features seems like rather crucial for any renter **We hypothesize that they will continually be positive in later analysis**

Housing Prices across Physical Principal Components



Observations

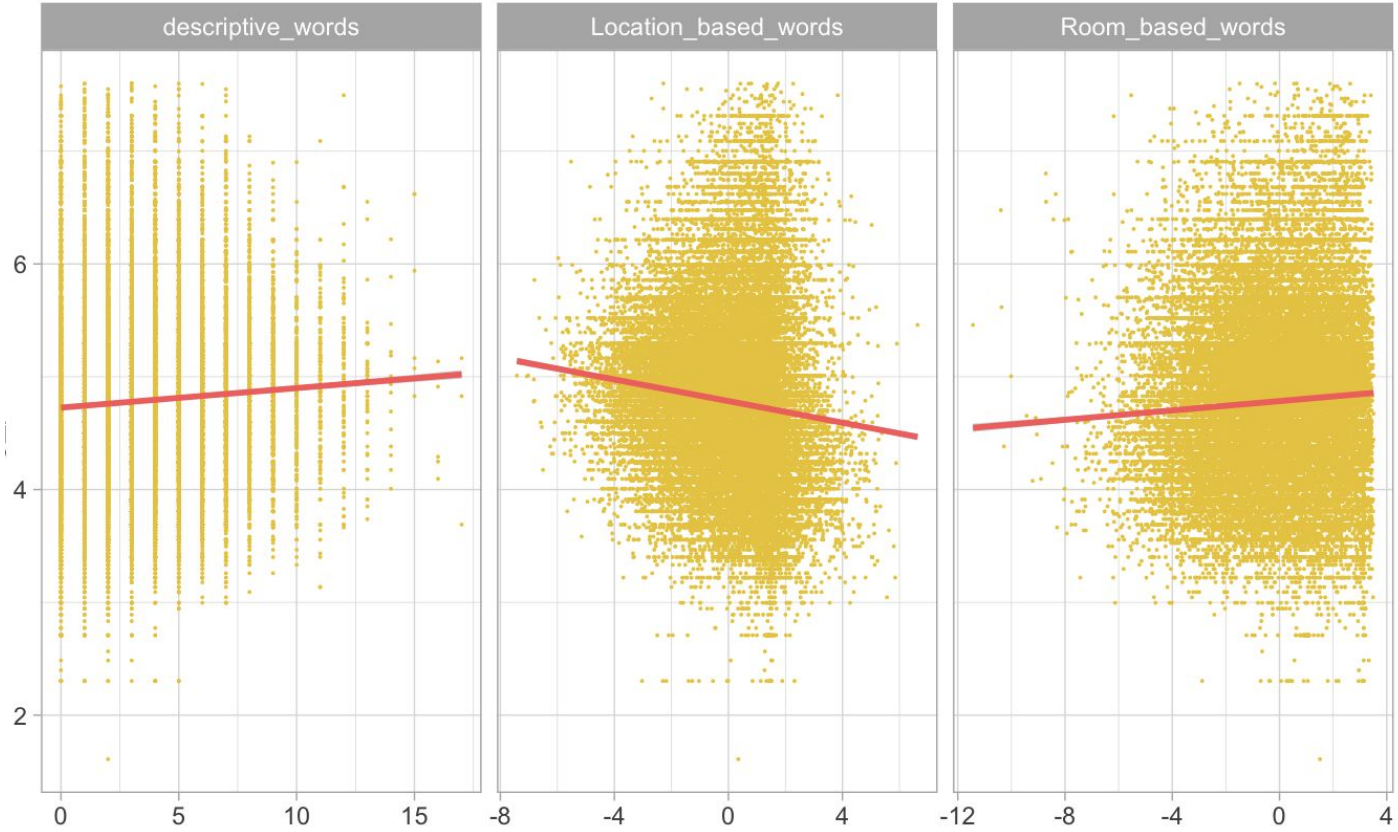
Surprisingly, **privacy and the ease of access related features** seems to be negatively related to the price of interest (Privacy, in particular)

Hypothesis

We hypothesize that **this may be due to the trade off attached with these choices.** (Higher privacy may indicate that this is a shared room or a smaller compartment) causing this trend to be negative. We hypothesize that this may be different post controlling other variable.

Housing Prices across Word Based Index

Housing Price (Logged) Across Word Based Index



Observations

Descriptive Words and Room Based Words both seems to have a positive relationship with the housing price, while **location based words seems to have a negative relationship**

Hypothesis

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, while location based words does not add additional value, but may instead decrease value if the location appear suboptimal

The Model

How much?

airbnb

Model Results & Selection

Models

Relaxed LASSO: adding bias by penalizing coefficients to create a sparser model to better estimate testing errors

Random Forest: generating many uncorrelated decision trees to improve prediction accuracy

Neural Network: algorithms designed to recognize patterns, interpreting sensory data by labeling or clustering raw inputs

XGBoost: 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.

Cross Validation

Used 10 fold Cross Validation to tune the the hyperparameter and loss function:

Relaxed Lasso: Tuned Lambda to perform feature selection

Random Forest: Tuned mtry (14) and number of trees (elbowed at 450) | OOB Prediction Error: 0.483 on Logged Price

Neural Network: Validation set used to evaluate trained model | CV Mean Absolute Error: 104 on Actual Price

XGBoost: Tuned eta (learning rate) and nrounds (number of training iterations) | CV Error: 0.486 on Logged Price

Test Results

The final Evaluation results on the test data are as follows (Evaluated on the mean squared error & the mean absolute error of the actual housing price):

Relaxed Lasso MSE: $188
Relaxed Lasso MAE: $75

Random Forest MSE: $136
Random Forest MAE: $62

Neural Network MSE: $187
Neural Network MAE: $102

XGBoost MSE: $138
XGBoost MAE: $63

RF >= XGBoost > Relaxed Lasso >= Neural Network

Feature Interpretation

Important Variables

The Physical components are most significant (Both Random Forest and XgBoost placed the physical component of the Top 7 in information gain).

Location variable were significant by the selection of Lasso and the information gain criteria of Random Forest , but not in the boosting tree.

Select user based features did appear significant in lasso selection and the ensemble models (sentiments and identities). **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 the conclusion that user based feature does matter**

Specific Feature Findings

Houses with more amenities and perks are pricier to rent

- It's like comparing furnished with unfurnished apartments: the more completed the house, the more valuable homeowners believe their property is

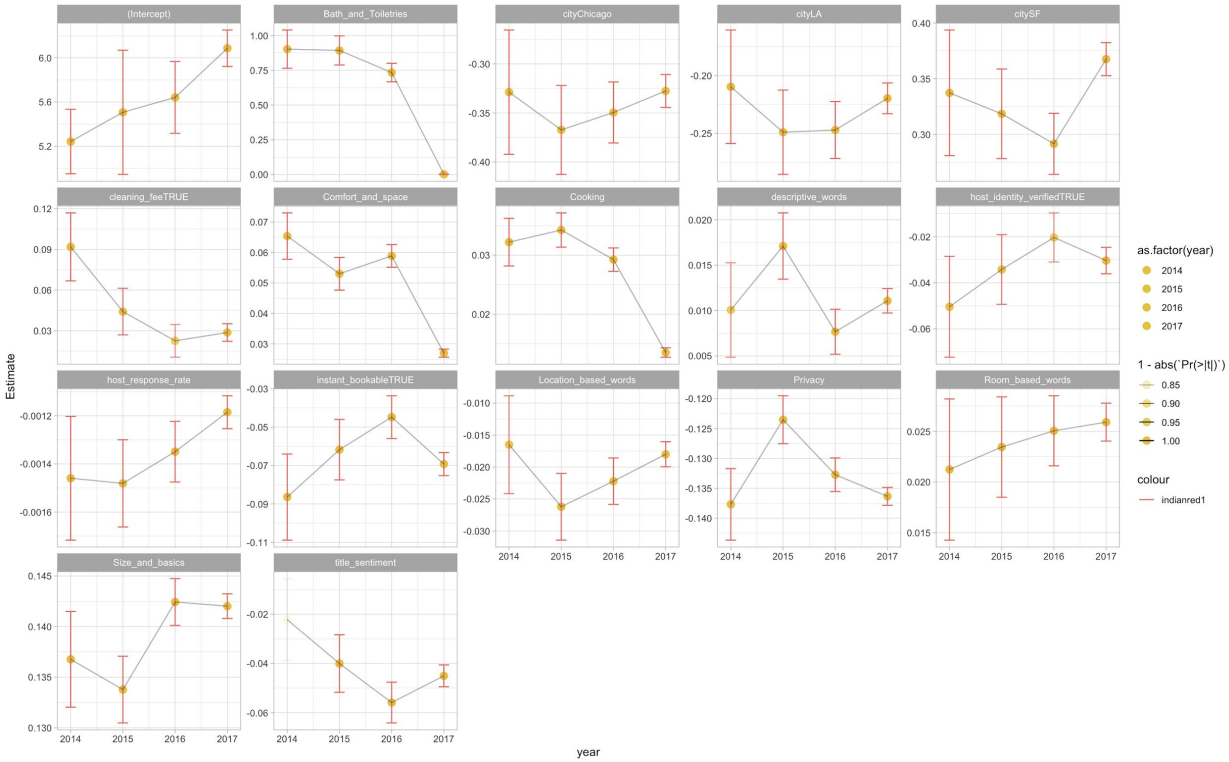
People prefer long title and high description sentiments

- These in general appeals more to public's mindset when making a purchase

3 Most Important Variables: Size and basics, privacy, and bath and toiletries

- Customers prioritize openness and completion of houses so they don't get bored or have to shop

Analyzing Consumer Preferences: 2014-2017



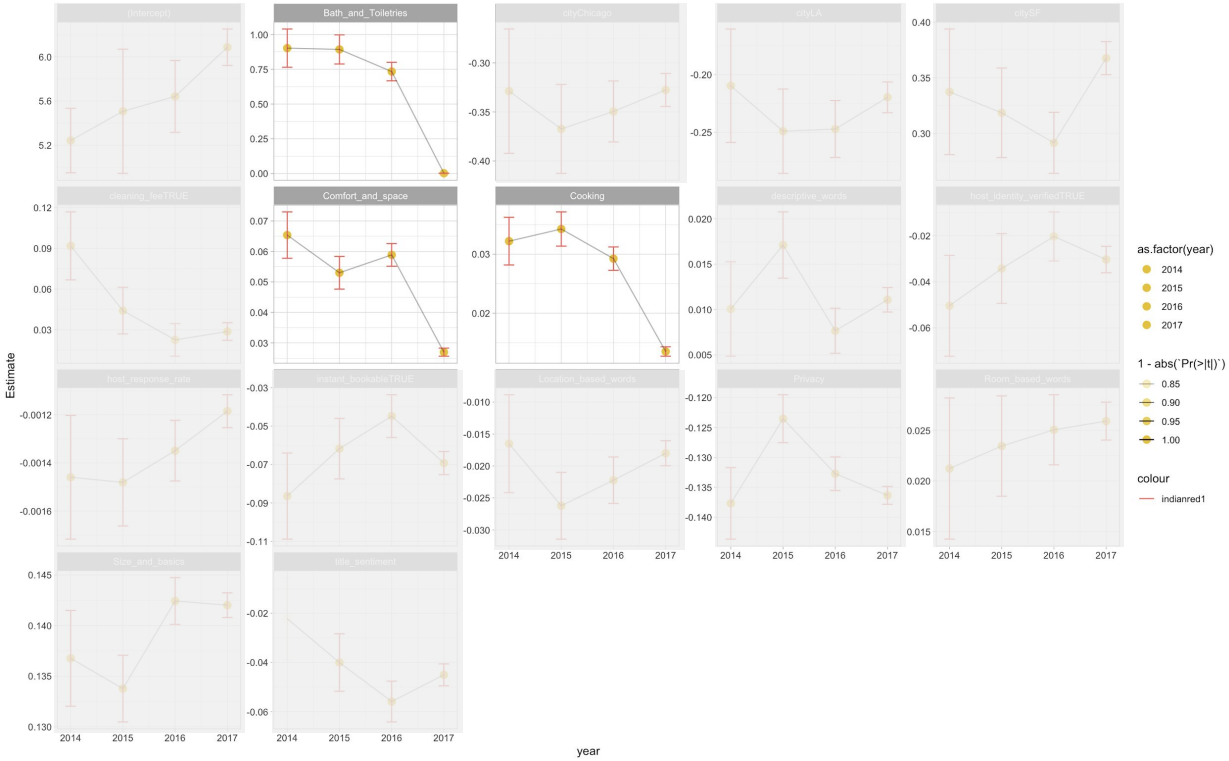
Rationale

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

We are specifically interested to see whether or not consumer preference has changed with regards to particular domains/features of the house

The graphic to the right are the factors that we've compiled which showed a significant effect through the four years.

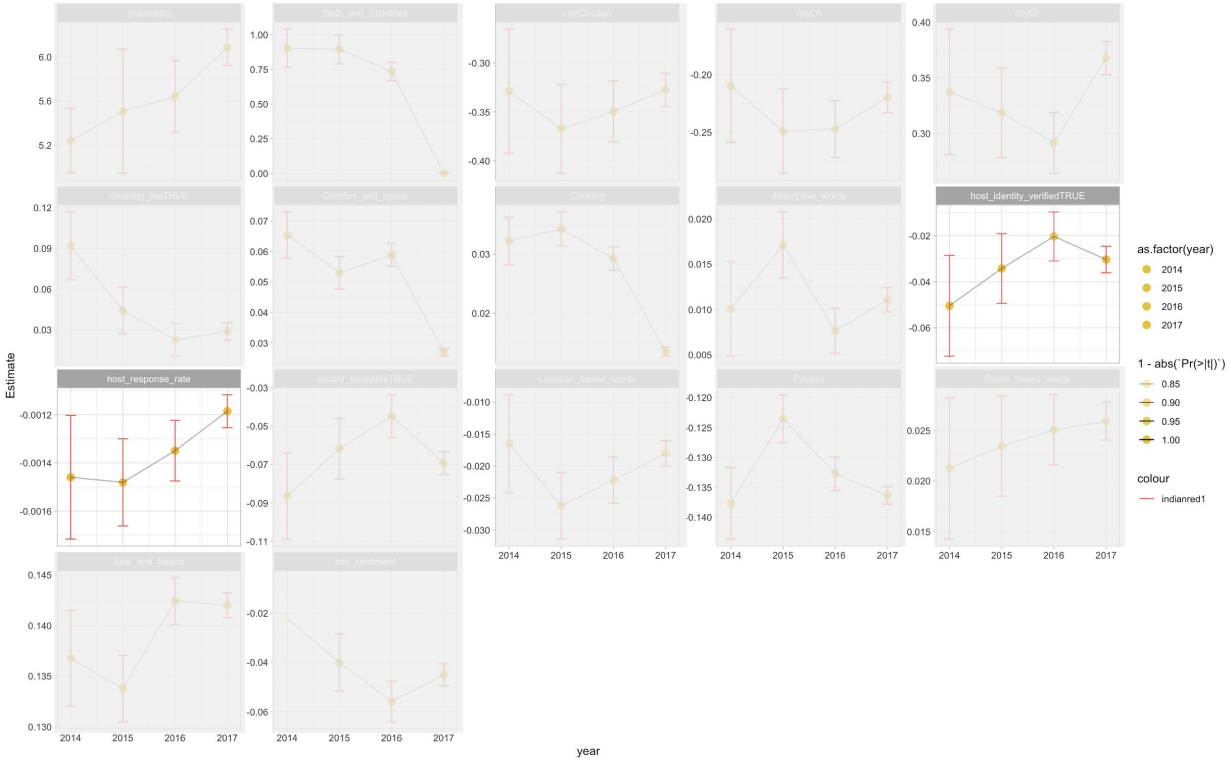
Analyzing Consumer Preferences: 2014-2017



Observation and Interpretation

- Amenities such as **Bathroom and Toiletries, The Comfort and spaciousness, and Cooking** affiliated features have all **gone down in effect** in the past years (particularly in 2017) have actually decreased in effect over the past 4 years
- We believe this may be due to the completeness of these amenities in more and more houses **(diminishing its uniqueness in value)**

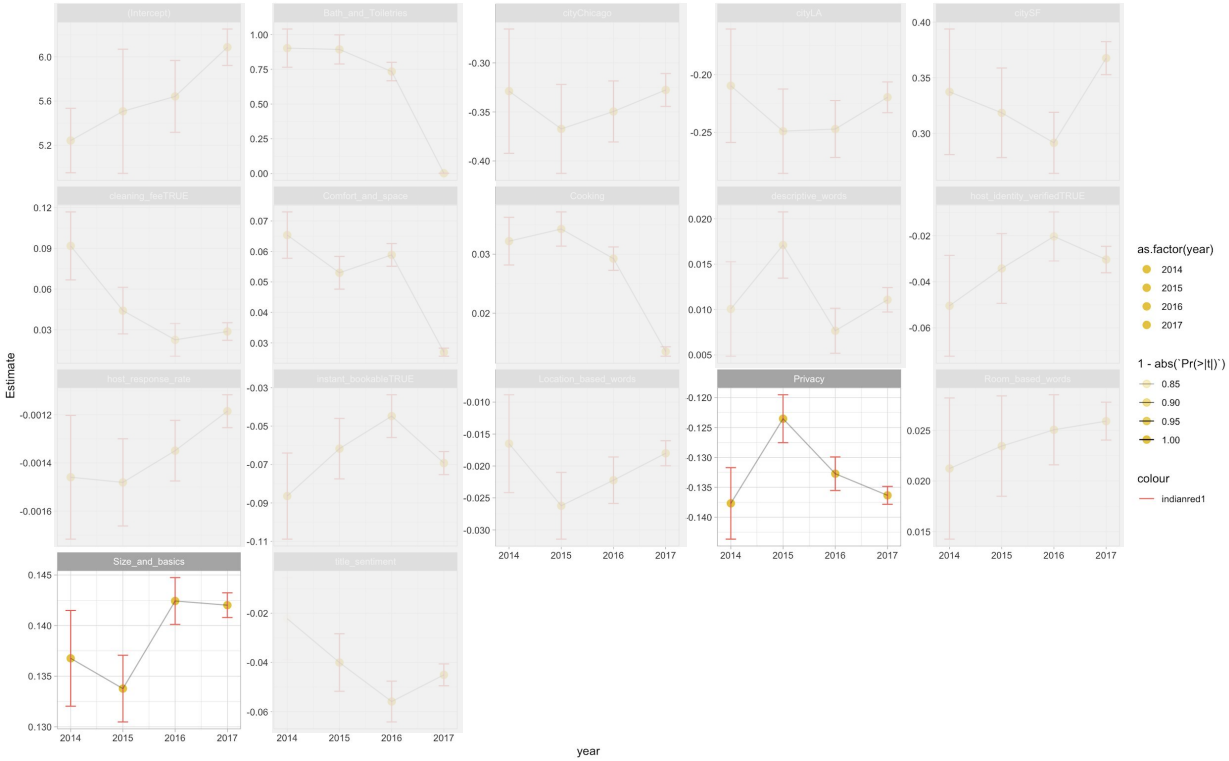
Analyzing Consumer Preferences: 2014-2017



Observation and Interpretation

- **Host identity, host response rate** have both **increase** in its effect in the past years, along with the house being **instantly bookable**
- This could be explained by **the increasingly valued factor of convenience** in the context of electronic services. People want things quick!

Analyzing Consumer Preferences: 2014-2017



Observation and Interpretation

- **Size and basics also increased in its effect**, with a big jump from 2015 to 2016; Parallel to what we observe for people less inclined to prefer Private houses

- The uses of Airbnb **has evidently grown and more frequently used in the different aspects (such as bigger party travels)**, we believe that contributed strongly to this observe trend of an increasing emphasis in house size

The Conclusion

What now?

airbnb

Impact Assessment & Conclusion

We ran our best performing Algorithm (Random Forest) on the Holdout set identified earlier and obtained the results below

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 all seems to be higher than their predicted value by about $20. We believe this may have strongly contributed to these houses not performing well (being rented or traded) on the market. With this algorithm we believe this situation could be remedied and fixed.

Last Note

To end our analysis, we want to offer a couple of advice regarding how to optimize for the value of your house:

1. Offering basic amenities is a must
2. Put time into writing a good description

Some keys features (i.e. size and basics) are inflexible and cannot be changed.

Some Must Haves

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

Thank You!

Any Questions?

airbnb

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
```

Appendix

```{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 
```

Appendix

```
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
```

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