Airbnb Listing Data Analytics

Hanbo Sun

February 4, 2017

1 Motivation

Even though it was founded no more than 10 years ago, Airbnb has risen as a major competitor to traditional hotels, with the potential to even revolutionize the rental market as well as the way people travel. One of the main reasons people choose Airbnb over traditional hotels is that Airbnb hosts can offer accommodation with great amenities at affordable prices. In this project, I will focus on analyzing a wide range of variables in the dataset and how they affect the listing price.

Amenities is a major selling point for housing. In this dataset, the amenities of the listings are thoroughly listed. We make use of this advantage by dummying this field to a sparse matrix with 0's and 1's that indicate the presence of each feature. With the transformed dataset, I aimed to build models that use the available features to predict the housing price. The models built in this process can serve as a benchmark to estimate the standard price of a unit given its features. Moreover, we can also rank the importance of covariates and see which amenities play a bigger role in the pricing of a unit.

This report also include exploratory data analyses that offer insights of the variability of listing price from a geographical standpoint, as well as using natural language processing to uncover the relationship between listing price and the housing names.

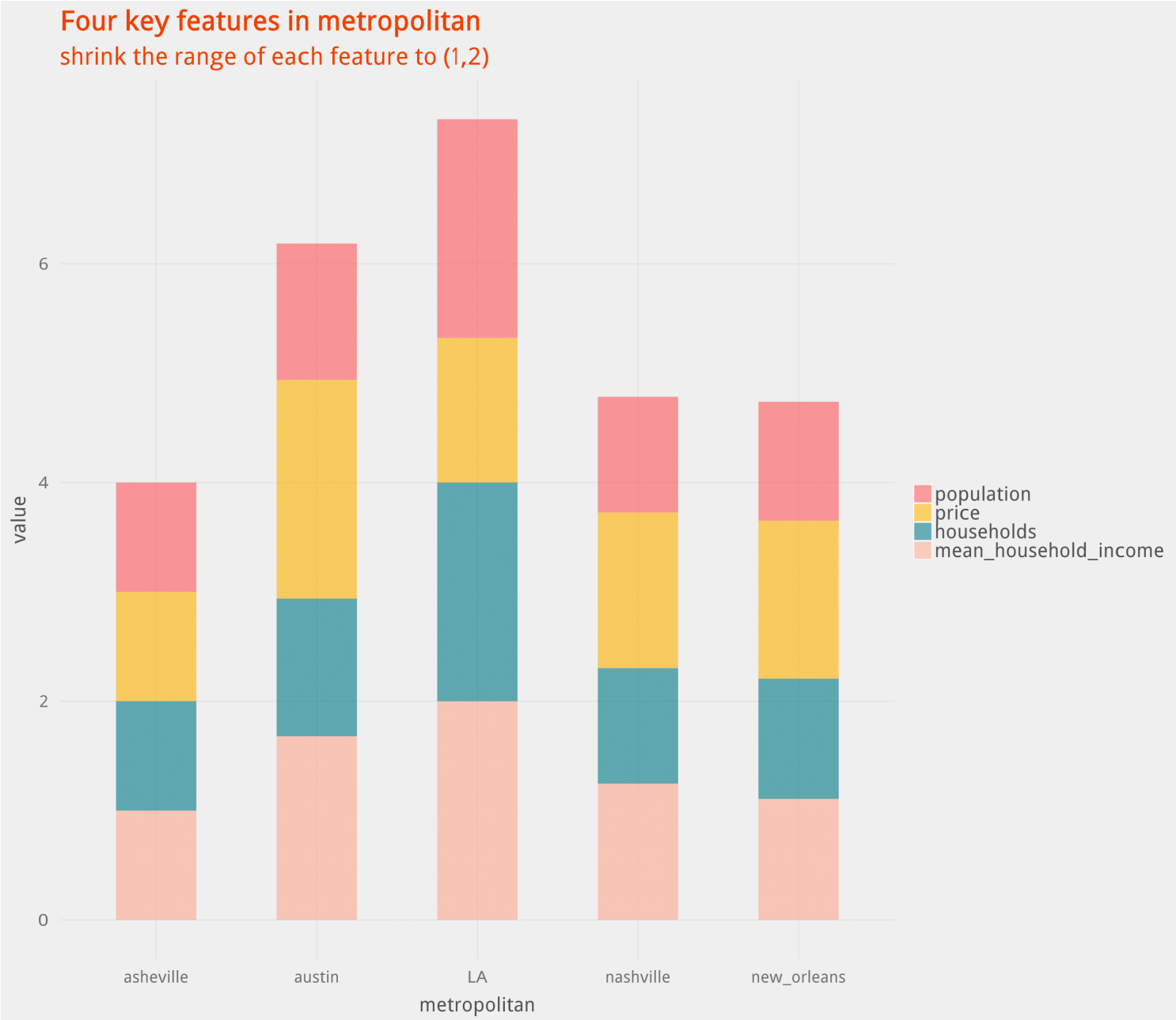
2 Non-Technical Executive Summary

2.1 Data

First I merged the listing and demographic datasets to see the key features in each metropolitan. It's surprised to find that even though LA has the largest average household income, its average price is not the highest. Four key features are scaled to range(1,2) for the sake of visualization and comparison.

Metropolitan	Population	Average Price	Households	Mean Household Income
Asheville	222673	125.36	94258	61160.80
Austin	3011512	293.63	1277711	85036.04
LA	11518350	179.55	4668327	96268.10
Nashville	879279	196.64	350988	69812.71
New Orleans	1223390	200.03	551649	64908.06

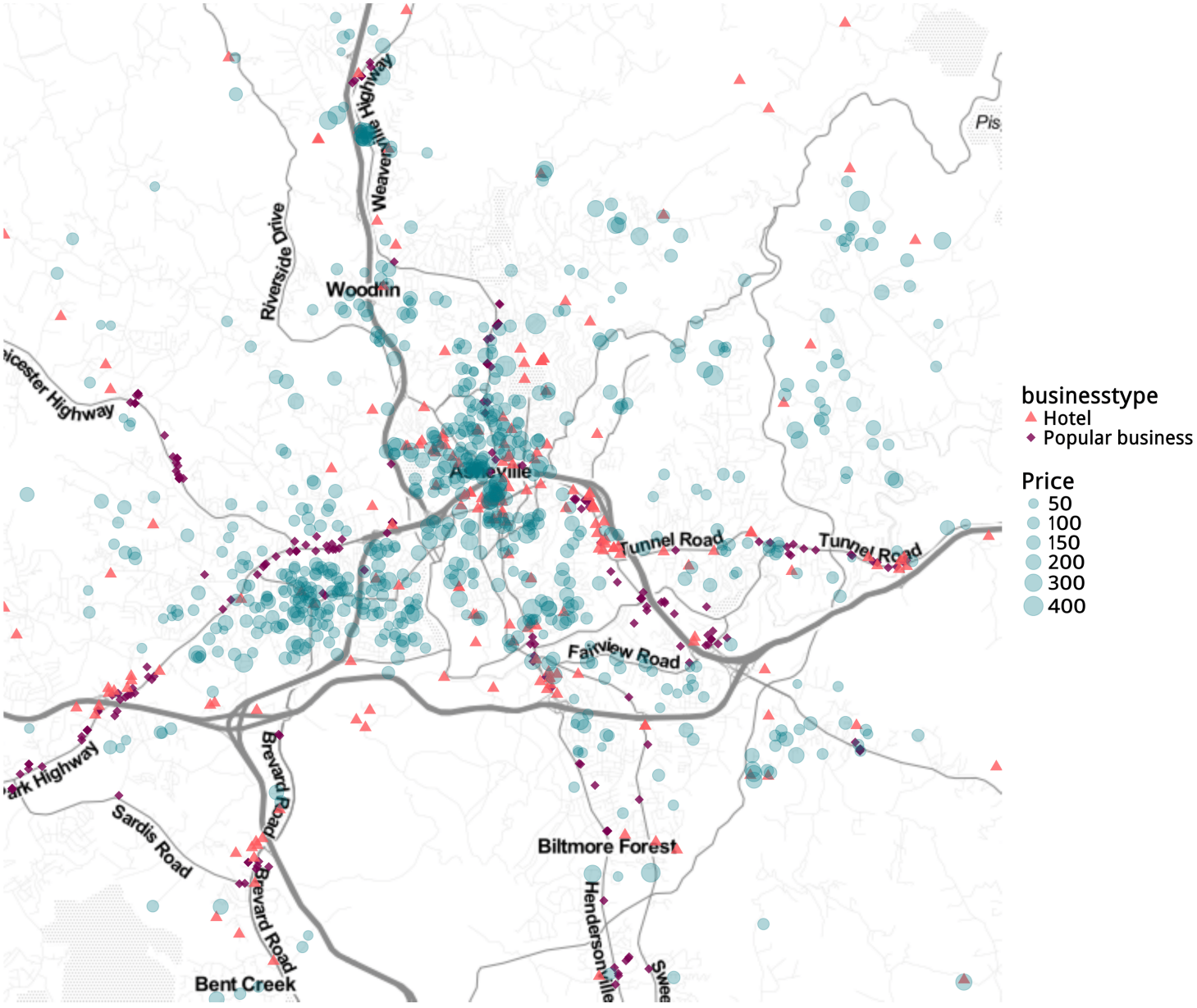
I also included some additional columns from the inside Airbnb website. The data are under the same schema so it can be joined with ease. This additional data give us more information like review counts.



Next, I focused specifically on the metropolitan Asheville and Austin to get a vivid outlook of the distribution of Airbnb listings, some popular business and hotels. The point size represent the price. This is the plot of popular business such as restaurants, gas station as purple square. It is clear to see all the popular businesses are located along with the roads, especially the highways. The density of listings is rather high at the business center than anywhere else. Further, I created the plot showing the locations of hotels in Asheville, because hotels are the main competitors against the Airbnb houses. Notice that some hotels concentration has very sparse Airbnb collections and hotels usually closed to business concentration while Airbnb housing may not.

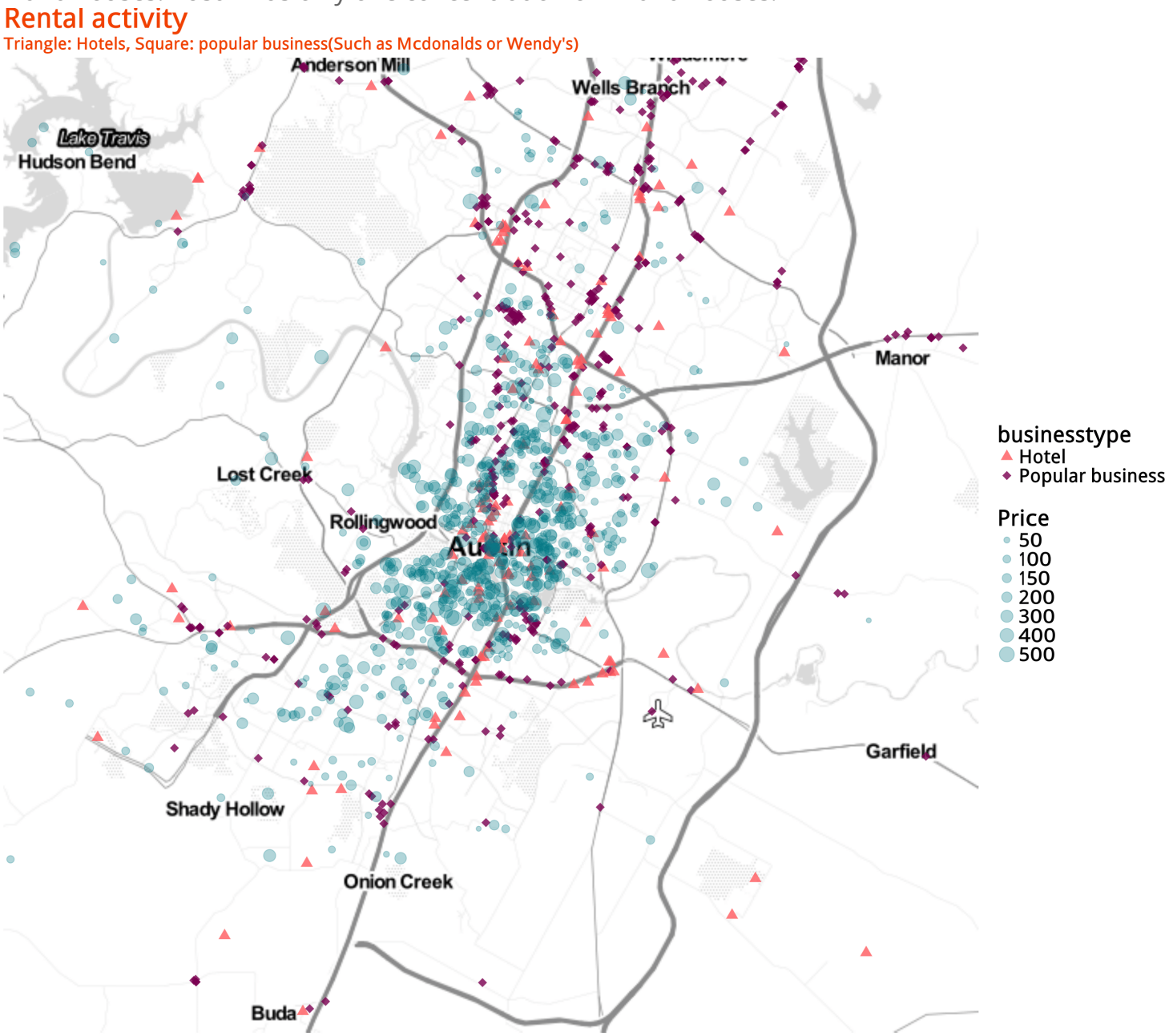
Rental activity

Triangle: Hotels, Square: popular business(Such as Mcdonalds or Wendy's)



However, in Austin the distribution pattern is not consistent with that of Asheville. Nearly all the Airbnb houses are located in the center of this metropolitan while most of the hotels are rather far from the center of the metropolitan. Popular businesses are as usual located along with the roads especially highways. In overall, the price of the Airbnb house is higher if the house is closer to the

center of the metropolitan. Austin is different from Asheville which has two main concentration of Airbnb houses. Austin has only one concentration of Airbnb houses.

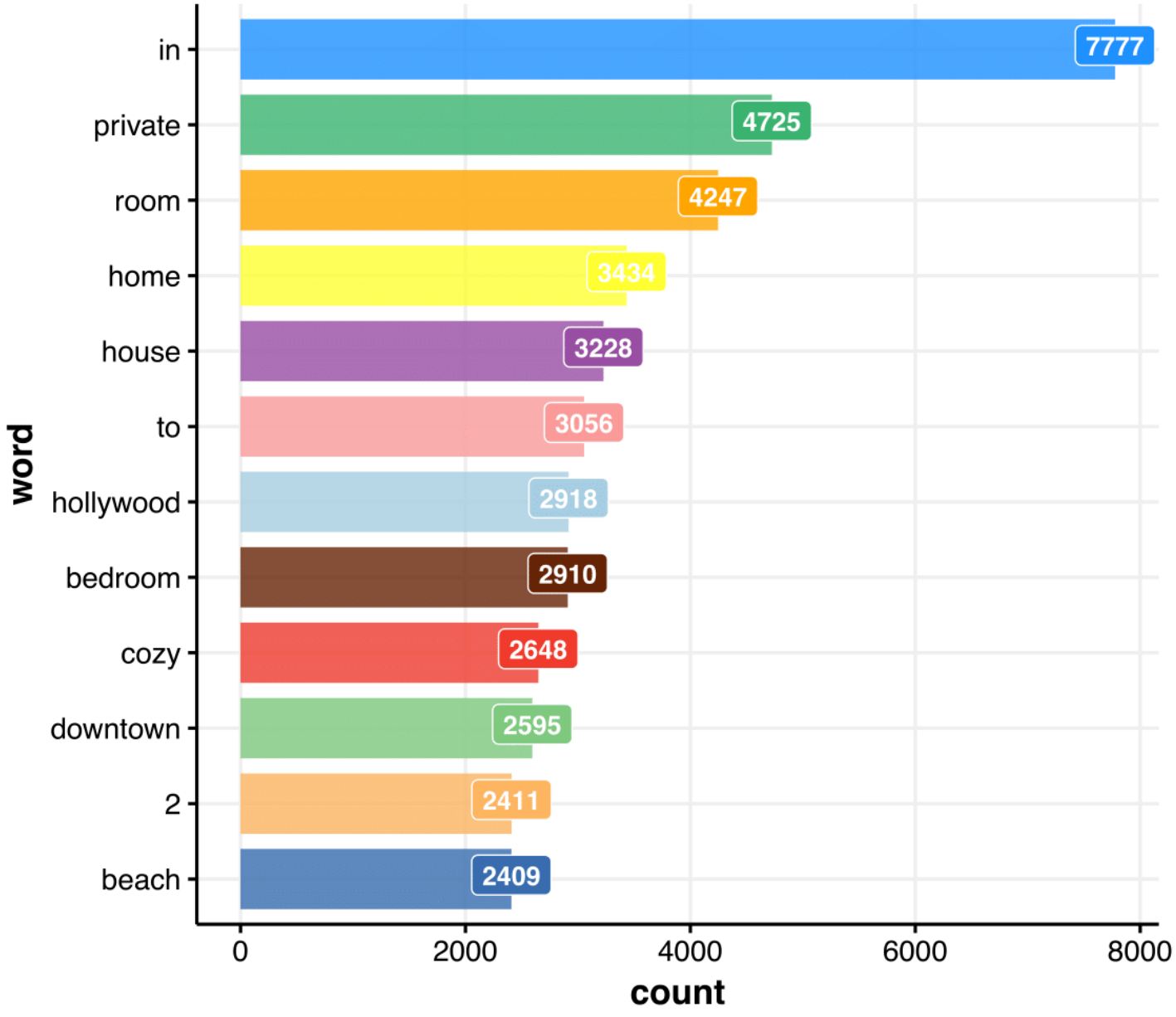


2.2 Analyzing keywords in listing names

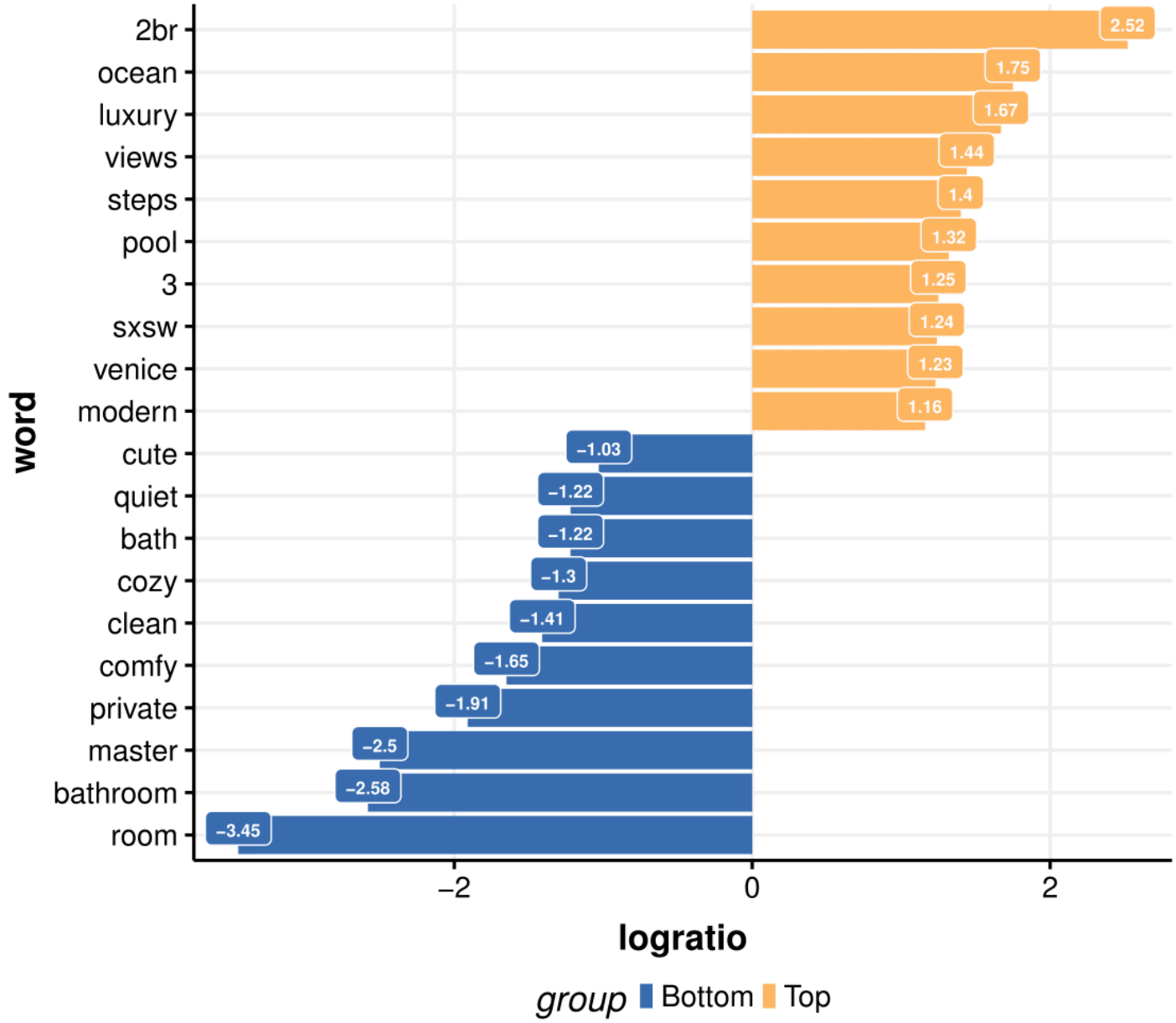
Unlike the usual business naming custom, the names on Airbnb listings are usually rather verbose, as the hosts tend to put as much amenities and appeals of their housing on the names as possible. Names like “Venice Beach and Canals Art House” and “Lovely Private Home - Far East Side” are fairly common format on Airbnb listings. Therefore just the names alone can often provide us with interesting truth.

Here we will look at the most frequent words that appear in listing names. This can give us an idea of what are the hosts’ favorite keywords when the hosts are pitching their houses to Airbnb users. Moreover, since pricing is one of the major focus of this project, I will investigate how those keywords are associated with the pricing.

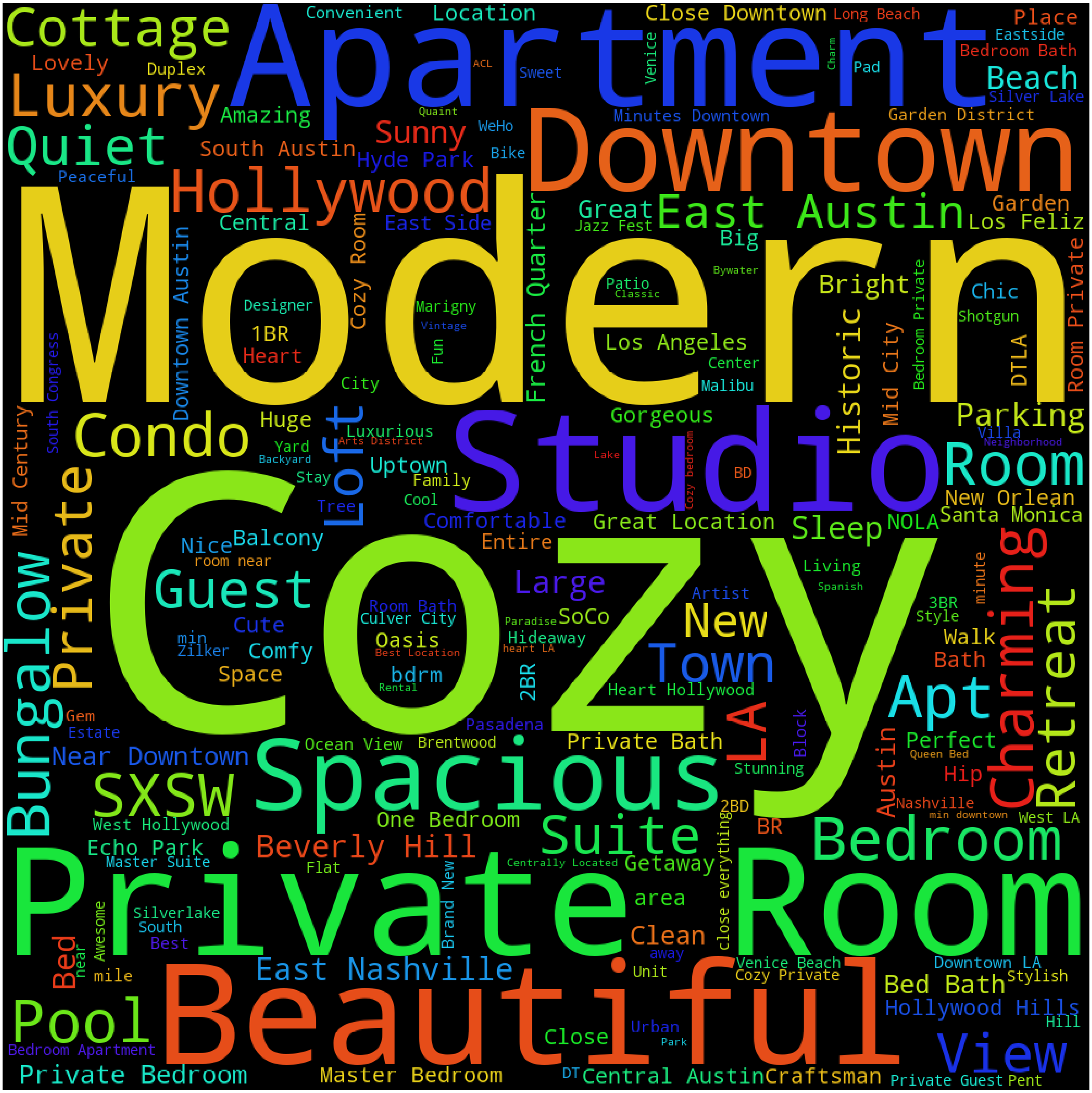
Figure 1 shows the top 12 words used in Airbnb listings within this dataset. Note that uninformative words like “the” are removed from this ranking. From this bar chart, some facts are denoted. Words like “in” and “to” may seem uninformative at first, but these words are usually used in the context like “in downtown area”, “5 min to nearby beach” .Both of them suggest that hosts very often put the location appeal in the listing names. We can also see this from words like “downtown”, “hollywood” and “beach” in this ranking. This choice of names can attract visitors with ease since they save users the trouble of looking up the location on the map. Words “private” and “cozy” are the most used adjectives in listing names. Other popular adjectives include “beautiful”, “lovely”, “spacious”, etc. An interesting thing is that the popularity of “private” seems a bit counter-intuitive, as many Airbnb users would simply select private rooms in the filter options at the start of their searching. It might be certain Airbnb policies that instruct hosts to make sure the private/shared information is clear to the users. If not, it can be a meaningful A/B test problem to investigate whether putting “private” in the names will affect users’ choice.

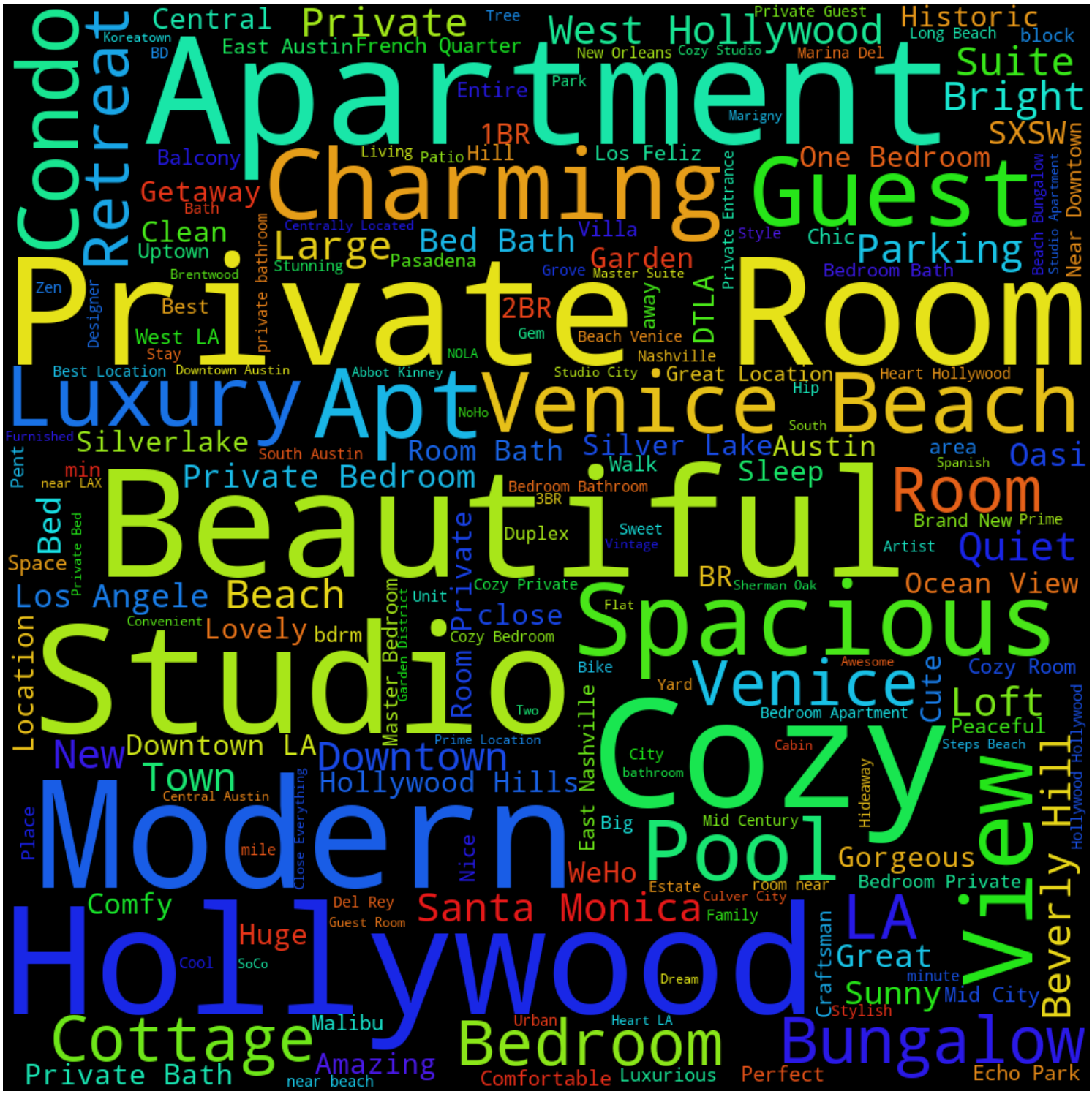


In order to discover the relationship of listing names and price, listings were separated into “Top” and “Bottom” categories based on their price’s position relative to the median in their respective metropolitan area. For each word, I computed the base 2 log ratio of the count it appears in “Top” over the count it appears in “Bottom” category. A positive log ratio indicates that the word is more often associated with listings of higher price while a negative log ratio indicates otherwise. According to Figure 2, generic adjectives like “cute”, “quiet”, “cozy”, “clean” and “comfy” are some of the most commonly used words in lower priced listings. One of the most common word in listing names “private” also belongs to this category. Compared to the popular words in the “Top” category like “ocean”, “luxury”, “pool” and “modern”, it seems that lower priced listings usually don’t have much premium features to boast about so the hosts would more often use words that can attract visitors that look for a comfortable budget stay.



Word clouds are also generated showing word frequencies in popular listings (with a review count higher than 6) and regular listings. From the word cloud of the popular listings, we notice that “modern” and “cozy” are two of the most frequent words. Whereas the common words in other listings are more spread out and no words stand out in particular.


Cottage
Convenient
Location
Apartment
Close Downtown
Long Beach
Place
Eastside
Lovely
Duplex
ACL
Sweet
Venice
Charm
Pad
Bedroom Bath
Luxury
Beach
Silver Lake
Amazing
Quaint
Minutes Downtown
Garden District
Quiet
Sunny
WeHo
South Austin
Hyde Park
Bike
Downtown
Peaceful
Hollywood
Garden
Central
East Side
Great
Jazz Fest
East Austin
Los Feliz
Big
Modern
Downtown Austin
Cozy Room
French Quarter
Bywater
Patio
Classic
Bright
Designer
Marigny
Vintage
Chic
Shotgun
1BR
Heart
Fun
Los Angeles
Center
Historic
Mid City
Bedroom Private
DTLA
Room Private
South Congress
City
Malibu
Mid Century
Condo
Huge
Luxurious
Arts District
Gorgeous
Parking
Villa
Neighborhood
Yard
Backyard
Loft
Studio
Lake
Cozy bedroom
BD
Room
Private
Stay
Uptown
Family
New Orlean
Tree
Cool
Santa Monica
Comfortable
Great Location
Sleep
NOLA
Nice
Balcony
Entire
room near
Cozy
Living
minute
Charming
Guest
Large
Spanish
Artist
Retreat
Bungalow
Room Bath
Culver City
Paradise
SoCo
New
3BR
Style
Cute
min
Zilker
Comfy
Oasis
Best Location
Hideaway
heart LA
Walk
Town
Bath
Space
bdrm
2BR
Rental
Heart Hollywood
Apt
Gem
Estate
Pasadena
LA
Block
Near Downtown
Ocean View
Brentwood
Private Bath
Perfect
Queen Bed
Austin
SXSW
Spacious
Stunning
Hip
2BD
Nashville
min downtown
West LA
West Hollywood
One Bedroom
close everything
BR
Suite
Bedroom
Echo Park
Beverly Hill
Master Suite
Flat
Brand New
Centrally Located
Getaway
Private
Room
Silverlake
South
Awesome
area
Bed
near
Best
Clean
away
Unit
East Nashville
Venice Beach
Cozy Private
Downtown LA
Stylish
mile
Bed Bath
Pool
Beautiful
Hollywood Hills
Hill
Urban
Close
Park
Bedroom Apartment
View
Private Bedroom
Master Bedroom
DT
Central Austin
Craftsman
Private Guest
Pent



3 Technical Executive Summary

3.1 Classification Hotness

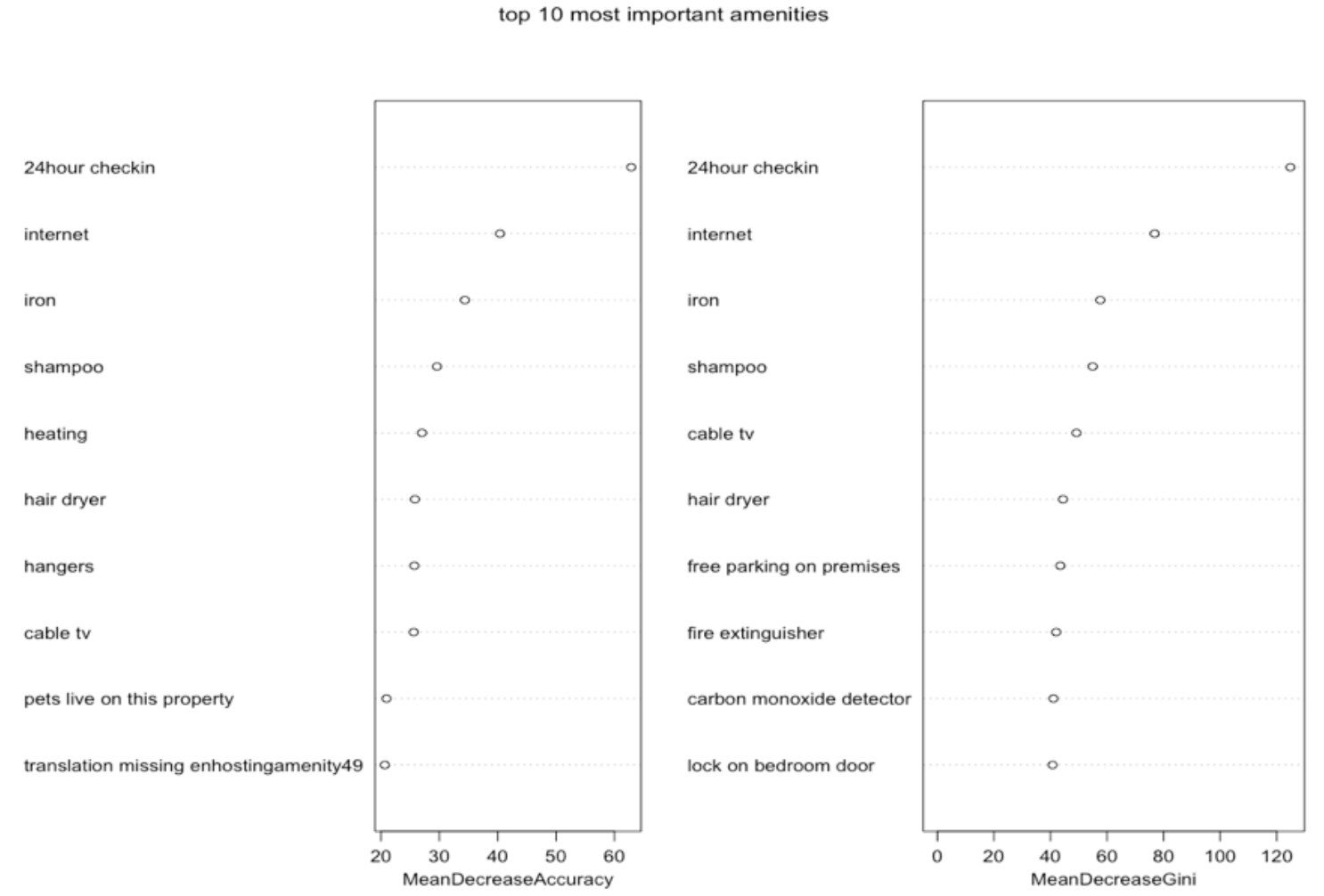
First, let's define evaluation metrics:

True Positive(TP): Number of observations that correctly classified as "Fall" group. True Negative(TN): Number of observations that correctly classified as "Non-Fall" group. False Positive(FP): Number of observations that incorrectly classified as "Fall" group. False Negative(FN): Number of observations

that incorrectly classified as “Non-Fall” group. Sensitivity (SENS) & specificity (SPEC): Sensitivity measures the proportion of “Falls” that are correctly classified while specificity measures the proportion of “Non-falls” that are correctly identified. Positive Predictive Value (PPV) & Negative Predictive Value (NPV): Positive Predicted Value measures the proportion of true “Fall” observations among predicted “Fall” observations. Similarly, Negative Predicted Value measures the proportion of true “Non-fall” observations among predicted “Non-fall” observations.

Methods	Accuracy	Sensitivity	Specificity	ppv	npv	LOR
ALogistic Regression	0.664	0.506	0.825	0.747	0.621	1.577
Random Forest	0.687	0.669	0.705	0.698	0.677	1.578
SVM	0.688	0.651	0.725	0.707	0.671	1.592
AdaBoost	0.681	0.688	0.674	0.683	0.679	1.517
XGBoost	0.685	0.675	0.694	0.692	0.677	1.553


Classification performance
accuracy, sensitivity, specificity
value
0.0
0.2
0.4
0.6
0.8
adative boost
logisticRegression
randomForest
SVM
Xgboost
classifier
variable
accuracy
sensitivity
specificity


top 10 most important amenities
24hour checkin
internet
iron
shampoo
heating
hair dryer
hangers
cable tv
pets live on this property
translation missing enhostingamenity49
20
30
40
50
60
MeanDecreaseAccuracy
24hour checkin
internet
iron
shampoo
cable tv
hair dryer
free parking on premises
fire extinguisher
carbon monoxide detector
lock on bedroom door
0
20
40
60
80
100
120
MeanDecreaseGini

The importance of amenities is in terms of the prediction for the hotness using random forest. The x axis is the mean decrease accuracy (or Gini index), if the variable is remove from candidate variable pool. The larger the drop indicates the more important of the variable. Building on that, the most "important" amenities include "24 hour check in", "internet", "iron", "shampoo" and so forth. ## Prediction Rental Price I would like to provide some insights into both Airbnb and housing providers what are the most important underlying factors to the rental pricing, since by obtaining knowledge of that, it gives housing providers appropriate suggestions on increasing the value of the rentals, on the other hand, it would help greatly for Airbnb to predict their revenue.

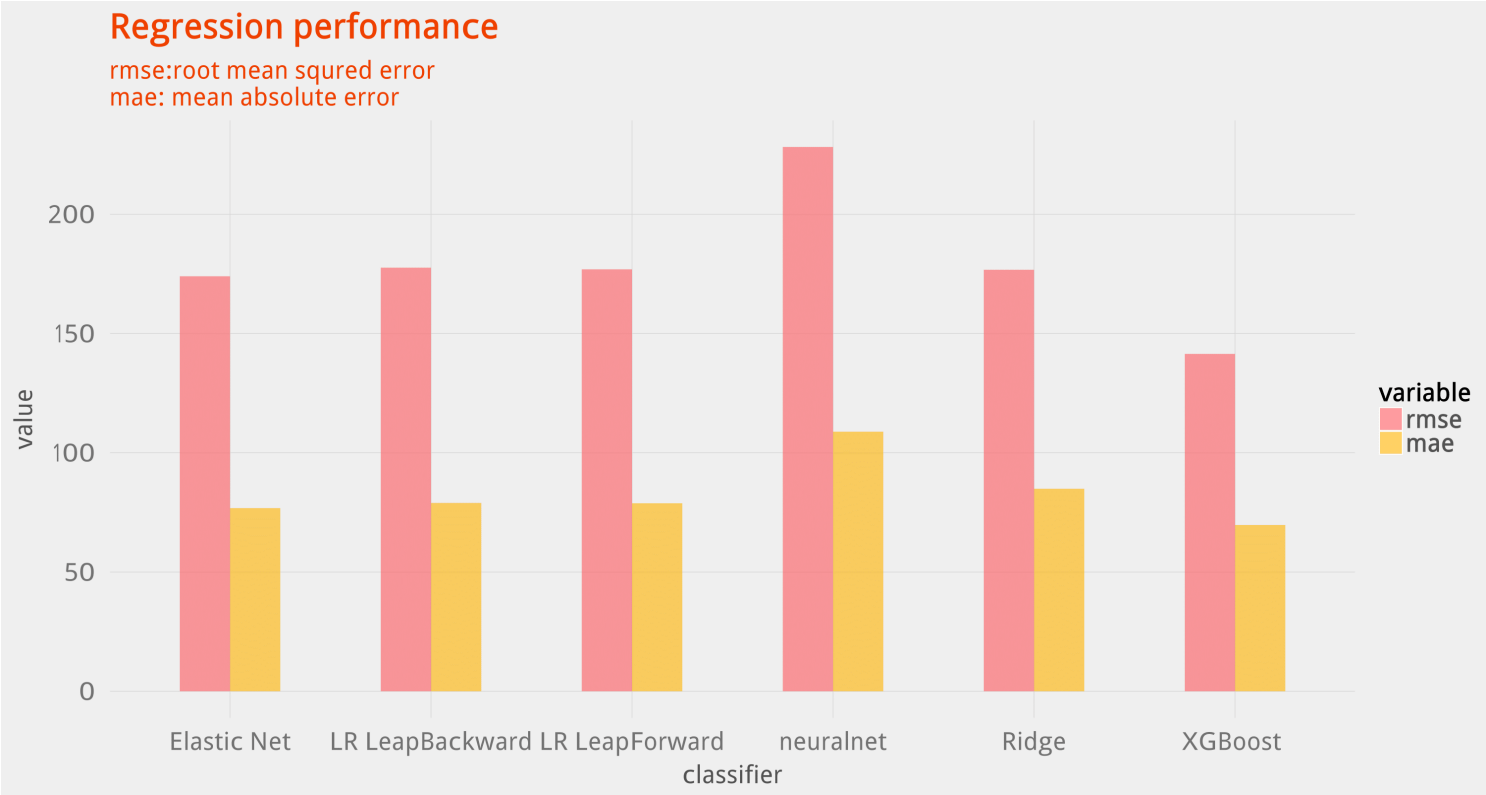
From the given dataset, the pricing of Airbnb rentals is affected in two dimensions, one is the housing's intrinsic features, the other is the external factors. For the intrinsic features, basically it involves a large variety of components, for example the property type, a Boutique hotel is obviously leads to a higher price than a normal apartment. We can find a set of variables representing the rentals' intrinsic value in the Listings dataset. I would like to give specific details on how the amenities variables are processed in the dataset. Amenities variable contains a list of amenities available in the property. All types of amenities are aggregated together and the sparse matrix is created with value one indicating possessed amenities and value zero indicating the house doesn't own this amenity. We ended up obtaining 100 types of amenities' variables, each representing an intrinsic feature about the rentals like whether the housing is equipped with cooking basics or whether there is elevator in the building.

To understand how intrinsic features affect its rentals' pricing, common prediction models were trained between the rentals' price and the amenities. I applied Elastic Net to train the model and conducted five fold cross-validation to find the optimal parameter. The final results gave us a Root Mean Squared Error (RSME) of 191.45 and Mean Absolute Error (MAE) of 99.98. By checking correlation relationship with variable price, I obtained the importance rank of the amenities features as we could see in figure 9, we notice the top five important amenities, among which whether a house has a indoor fireplace is the most important intrinsic feature.

On the other hand, the rentals' pricing is also greatly affected by the external factors, especially the location and the customer's reviews. For any business, location is always one of the key factors. An apartment located in Manhattan is usually more expensive in rental than a house in the countryside. In the Listings dataset, we could include location based variables like zipcode, longitude, latitude and metropolitan to give us more prediction power in the location dimension. Customer's review is also a critical component of Internet sharing economy. Nowadays people tends to choose where to lodge based on the other's reviews. Those rentals and hosts who are hot and highly commented among the Airbnb Community would naturally lead to higher pricing strategy. Based on this reason, I added the review scores from the Listings dataset to improve the prediction results.

In the final model, I included the intrinsic features and external factors as predictors, rentals' price as the response variable. Linear Regression is applied with forward and backward stepwise features selection, respectively. Furthermore, Ridge Regression, Elastic Net, Neural Network and XGBoost

models were trained by tuning the optimal parameters applying five fold cross-validation. The final results are listed below:



By comparison of the models, XGBoost resulted the lowest error in terms of both Root Mean Squared Error(RMSE) and Mean Absolute Error(MAE). The results show that we do approach predictive models which would help housing hosts to increase revenue and help Airbnb to predict the revenue.