Arjun Rao | Jenny Tseng | Kelly Zhang | Grant Zhong

Airbnb Prices in Austin

Project Goals

As the cost of living in the Austin area rises, residents are looking for opportunities for additional income to help cover ends meet. As access to technology expands, Airbnb has provided easy and dependable income for all the better. Those with spare couches, bedrooms, or apartments are turning to Airbnb for that passive income. The goal of our analysis is to enable owners to price their properties competitively with the market, so they can reduce vacant nights and maximize earnings.

We investigated Airbnb properties listed between 2010 and the first half of 2017 for Austin and its surrounding areas Obtained from Kaggle: https://www.kaggle.com/PromptCloudHQ/airbnb-property-data-from-texas. We then developed a model that enables homeowners to price their Airbnb listings appropriately based on features including location, number of bedrooms, title word count, and keywords in the description.

Exploratory Analysis

The dataset we have include the following columns for each listing: average rate per night, bedroom count, city, date of listing (month and year), title, description, latitude, longitude, and URL. We added a title word count column-based and title, description word count and keywords columns based on the description, and zip code column based on latitude and longitude.

Basic Statistics

As can be seen in Figure 1, the mean listing price was $159, with the interquartile range being $50 to $150 and a standard deviation of $279. This indicates that the prices have a long tail of more expensive properties, which heavily influence the average rates. This makes sense as the rate of a mansion will be significantly higher than a one-bedroom apartment. Figure 2 helps with visualizing how many of these high-priced "outliers" we may be dealing with. As they are few in quantity, we followed the standard way of dealing with outliers by removing listings with rates more than two standard deviations above the mean. This only reduced our dataset by 61, yet still resulted in a right-skewed listing rate histogram, as shown in Figure 3.

We also looked at the growth in the number of listings over time. As seen in Figure 4, the volume of listings has been rising since 2009. That said, Figure 5 shows that the largest percentage increase came in 2011, shortly after

Airbnb's founding in 2008. Fun fact, Airbnb also won the "App" award at 2011's SXSW.

Location Analysis

The 1300 listings we have are split amongst 54 zip codes. We analyzed these zip codes in-depth as location is one of the most important predictors we have. As can be seen in Figure 6, the listings in our dataset have a concentration in the central Austin area, as can be expected. That said, most areas appear to have a healthy number of listings, with a few pockets as exceptions.

To examine whether the exceptions would affect our analysis, we sorted the data by the number of listings in each zip code. As can be seen in Figure 7, only 17 zip codes had over 30 listings. Given that we cannot model from a few data points, we decided to limit our analysis to the listings in the 17 zip codes. This furthered reduced our dataset from 1302 to 864. That said, all zip codes left out can refer to their neighboring zip codes for pricing ideas.

For our targeted 17 zip codes, we also did a preliminary pricing analysis by averaging the rates in each zip code. As can be seen in Figure 8, downtown (78701) appears to have the highest average rate across zip codes, followed by Buda (78610) (which, from a quick glance at the data, tends to have more rooms). Pflugerville (78660) and Round Rock (78664) are on the cheaper side, with West Campus (78705) taking the bottom.

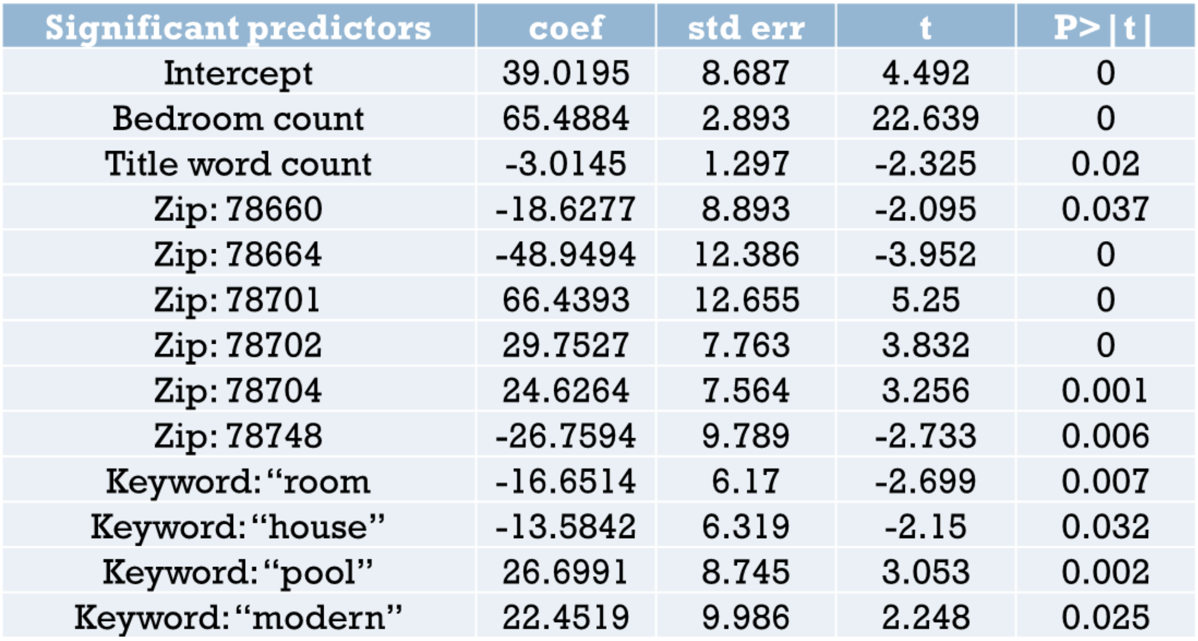
Keyword Analysis

Finally, we conducted text analysis to add additional features for our model. We first ranked all words used in the description by usage frequency. We also calculated and ranked the TF-IDF scores for all words that were mentioned over 30 times. (In both rankings we stopped commonly used English words such as "a", "an", and "the.") From the two lists, we manually picked 34 high-to-medium frequently used words and words with high TF-IDF scores by judging how important their messages are. Figure 9 shows the rankings of the top words by usage and TF-IDF scores; those bolded were chosen to be included in the model.

Solution and Insights

We decided to use regression to predict the average listing rate based on bedroom count, title word count, description word count, the 17 zip codes, and the 34 keywords. A regression model would help us understand which predictors are important and is more interpretable compared to other models.

To prepare for the regression model, we first turned zip codes and keywords into dummy variables and split the dataset 80/20 into training and test sets. We then fit the model based on the training set and predicted on the test sets. The table below shows the predictors that turned out to be significant.



Significant predictors	coef	std err	t	P>|t|
Intercept	39.0195	8.687	4.492	0
Bedroom count	65.4884	2.893	22.639	0
Title word count	-3.0145	1.297	-2.325	0.02
Zip: 78660	-18.6277	8.893	-2.095	0.037
Zip: 78664	-48.9494	12.386	-3.952	0
Zip: 78701	66.4393	12.655	5.25	0
Zip: 78702	29.7527	7.763	3.832	0
Zip: 78704	24.6264	7.564	3.256	0.001
Zip: 78748	-26.7594	9.789	-2.733	0.006
Keyword: “room	-16.6514	6.17	-2.699	0.007
Keyword: “house”	-13.5842	6.319	-2.15	0.032
Keyword: “pool”	26.6991	8.745	3.053	0.002
Keyword: “modern”	22.4519	9.986	2.248	0.025

As expected, the number of rooms is one of the most important predictors on listing rate. Location-wise, only six zip codes are significant. The title word count matters slightly, while most keywords are not relevant to the listing rate.

We performed the regression model again, including only the significant variables. The RMSE was the same for both models ($111), the R square went down slightly (from 0.557 to 0.531), but the Adjusted R went up (from 0.520 to 0.523). As expected, taking out features that are not predictive decreases the complexity of the model, as seen from the increased adjusted R square.

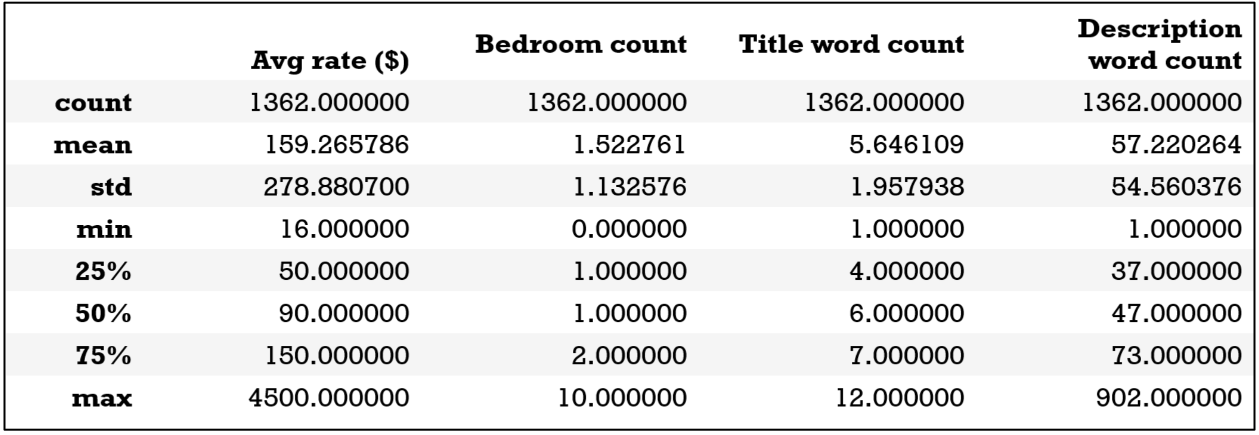
From the model, we concluded that holding everything constant, listings start at $39.05 and increases by $68.49 on average for every additional bedroom. Title world count, while significant, does not matter much. While people don't want to read too much, each additional word only costs about $3. On the location side, Downtown, as seen previously, is the most prime real estate, followed by East Austin. North Austin (78660 and 78664) on the other hand is cheaper. Lastly, having the words "pool" and "modern" in the descriptions, not surprisingly, increase your price, but these are more attributes of a listing, rather than how one describes a listing. The words "room" and "house," for some reason, decrease the price.

As seen from our R square, our model can explain about 50% of the listing prices. The low percentage and the RMSE of $111 indicate that there is room for improvement. First of all, even though we took out the top 61 listings in terms of price, our data is still heavily skewed to the right. As such, our model tends to overpredict listing prices.

Second, and more importantly, we do not have enough information about each listing. We essentially based our predictions off of only location and bedroom count. While we started extracting more property quality with the keywords, they did not take us far enough. Additional data that may be able to help further explain listing rates, and therefore improving prediction accuracies, include more concrete listing attributes, number of photos, and the quality of photos.

Figures

Figure 1. Basic statistics of dataset



	Avg rate ($)	Bedroom count	Title word count	Description word count
count	1362.000000	1362.000000	1362.000000	1362.000000
mean	159.265786	1.522761	5.646109	57.220264
std	278.880700	1.132576	1.957938	54.560376
min	16.000000	0.000000	1.000000	1.000000
25%	50.000000	1.000000	4.000000	37.000000
50%	90.000000	1.000000	6.000000	47.000000
75%	150.000000	2.000000	7.000000	73.000000
max	4500.000000	10.000000	12.000000	902.000000

Figure 2. Listing Rates on File

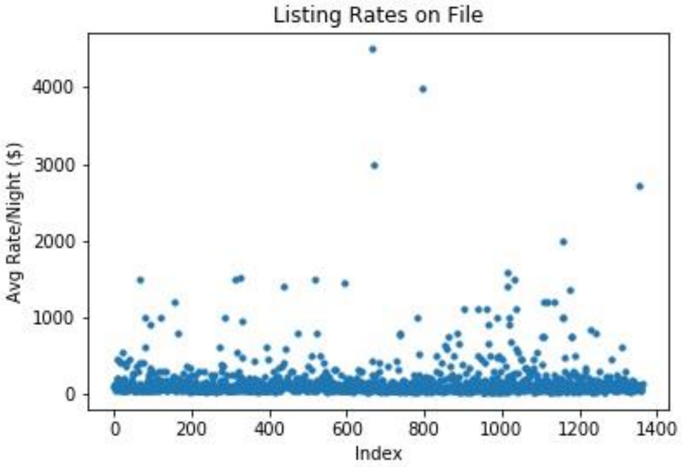


Figure 3. Frequency of Listing Rates

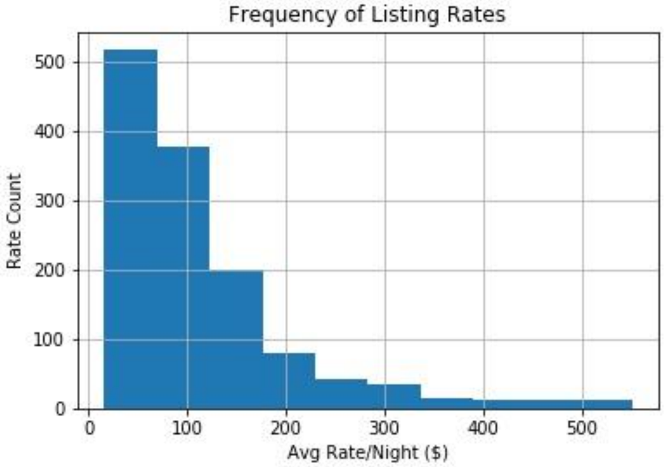


Figure 4. New Airbnb Listings by Year

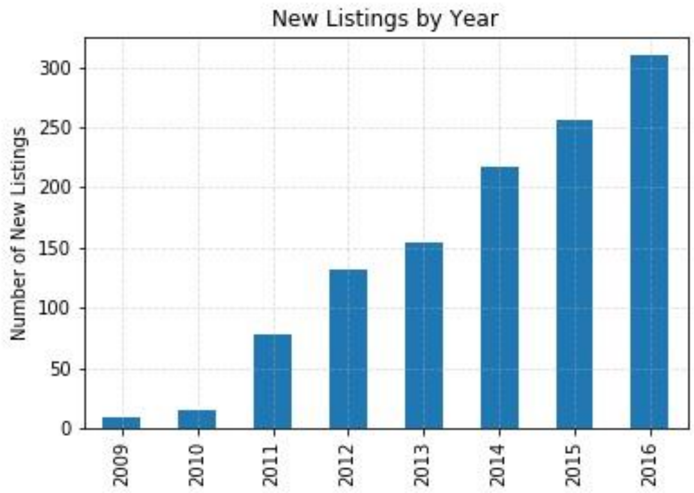


Figure 5. Annual Listing Growth in Percentage

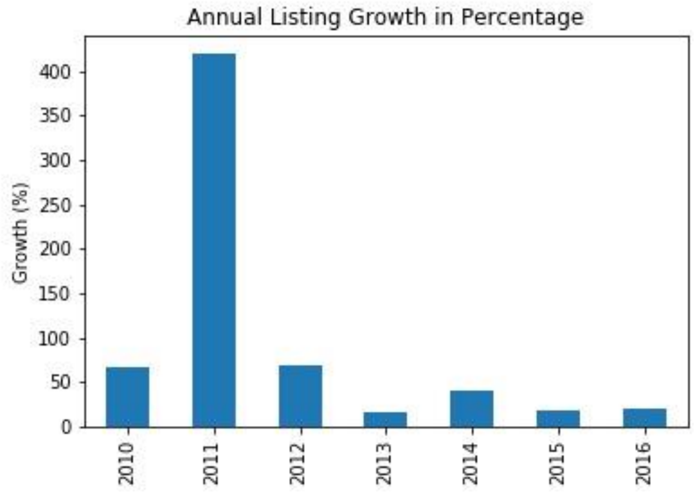


Figure 6. Listing Location Heatmap

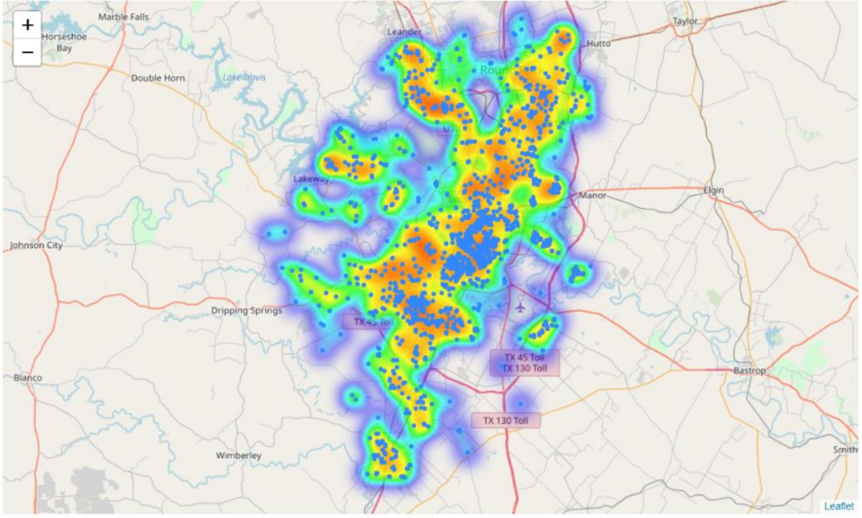


Figure 7. Number of Listings in the Top 25 Zip Codes

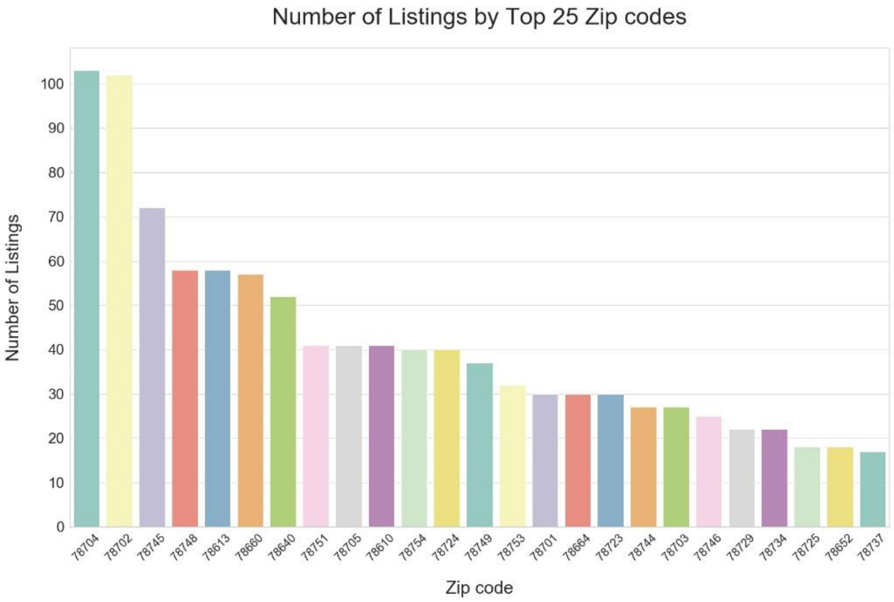


Figure 8. Average Rate per Night in the Top 17 Zip Codes

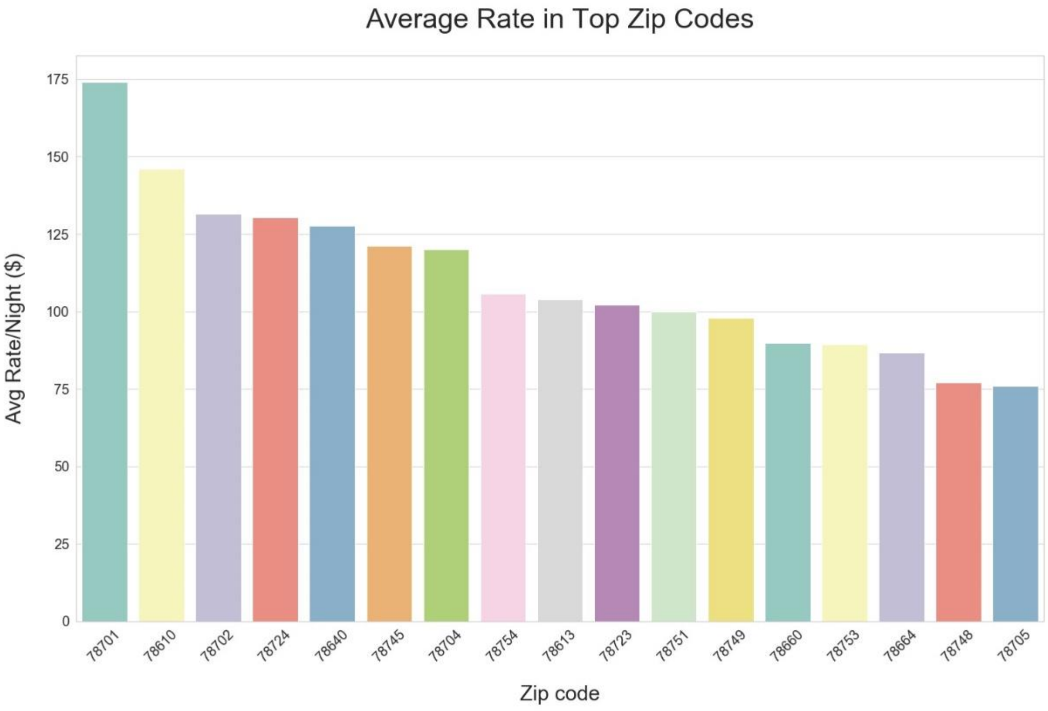
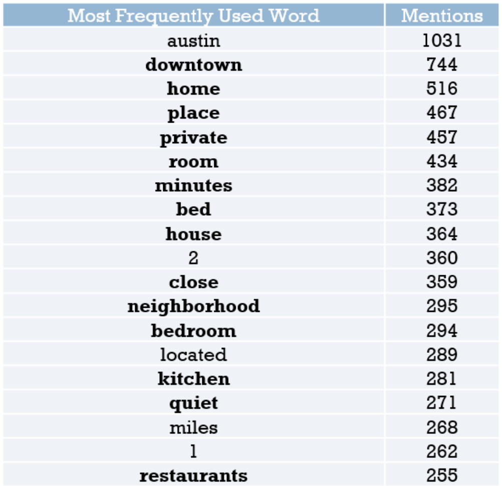
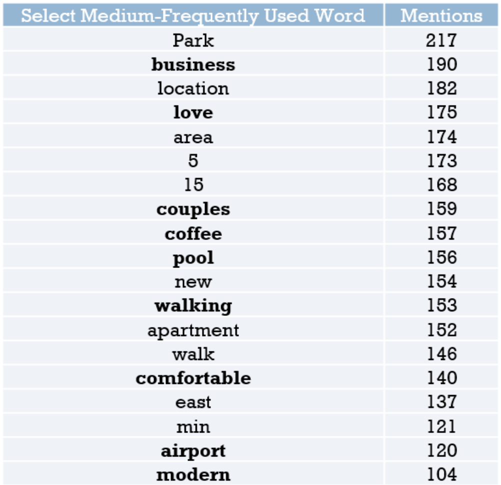


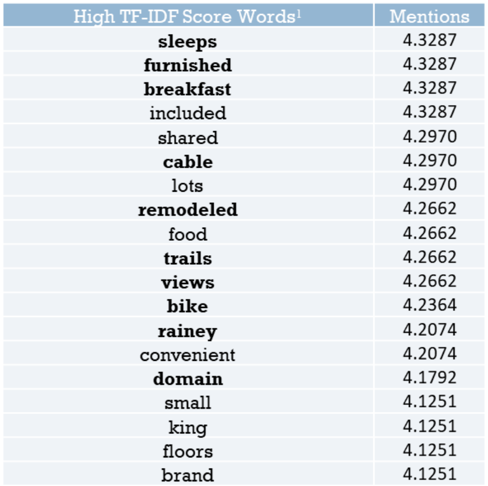
Figure 9. Keywords Analyzed Based on Frequency and TF-IDF Scores



Most Frequently Used Word	Mentions
austin	1031
downtown	744
home	516
place	467
private	457
room	434
minutes	382
bed	373
house	364
2	360
close	359
neighborhood	295
bedroom	294
located	289
kitchen	281
quiet	271
miles	268
1	262
restaurants	255



Select Medium-Frequently Used Word	Mentions
Park	217
business	190
location	182
love	175
area	174
5	173
15	168
couples	159
coffee	157
pool	156
new	154
walking	153
apartment	152
walk	146
comfortable	140
east	137
min	121
airport	120
modern	104



High TF-IDF Score Words[1]	Mentions
sleeps	4.3287
furnished	4.3287
breakfast	4.3287
included	4.3287
shared	4.2970
cable	4.2970
lots	4.2970
remodeled	4.2662
food	4.2662
trails	4.2662
views	4.2662
bike	4.2364
rainey	4.2074
convenient	4.2074
domain	4.1792
small	4.1251
king	4.1251
floors	4.1251
brand	4.1251

Notes: (1) For words mentioned at least 50 times in all descriptions
(2) The bolded words are the selected words for the model