NICOLE SAMRAO

ANALYSIS OF AIRBNB LISTINGS IN U.S. CITIES

PREDICTING LISTING PRICE USING THE ZILLOW HOME VALUE INDEX

Introduction

The principal purpose of this project is to use Airbnb listing data and Zillow home value data to provide marketing and business recommendations to Airbnb. This will be achieved through exploratory data analysis and predictive models for Airbnb listing price using both data sources.

The first part of the project is dedicated to taking a visual look into Airbnb listing data of major U.S. cities to discover trends in listings price, occupation, rating, value etc. The cities used in the analysis range from large cities such as New York City, Los Angeles and San Francisco to smaller, but emerging cities such as Portland, Austin and Nashville. We are looking for insights into listing behavior that could help Airbnb make strategic business decisions. The second part of the project introduces the Zillow data and the process of creating a predictive model for listing behavior that uses both data sets.

The Data

Data Sources

The Airbnb Data was obtained through the third party website, http://insideairbnb.com, which scrapes listing data regularly. All the listing data used in this project is from the June-July timeframe of 2016. Zillow makes their home value index available on their website at: http://www.zillow.com/research/data. The home value data used was from 2016.

Breakdown of Analysis:

1) Exploratory Data Analysis: The Airbnb Data
2) The Zillow Home Value Index and Listing Price
3) Prediction Model: Predicting Price of a Listing

The Airbnb Data

We begin by importing and cleaning the raw Airbnb listing data. The cities used for this exploratory analysis are:

- Austin, TX
- Boston, MA;
- Washington, DC;
- Denver, CO
- Los Angeles, CA
- Nashville, TN
- New York City, NY
- Portland, OR
- San Diego, CA
- Seattle, WA
- San Francisco, CA
- New Orleans, LA

The columns or features selected to be used in the analysis:

ID: This is a unique identifier for each listing

Room Type: Defines the layout type of the listing as either: 'Single Room', 'Shared Room' or 'Entire Home/Apartment'

Accommodates: The number of people that each listing can accommodate.

Bathrooms: The number of bathrooms in each listing

Bedrooms: The number of bedrooms in each listing.

Availability 365: The number of days that a listing is available to be booked in the next 365 days

Number of Reviews: The number of reviews for each listing.

Review Scores, Rating: This is the average overall rating for each listing out of 10.

Review Scores, Accuracy: This is the average accuracy rating for each listing, The user rates the listing based on how accurate the listing was as compared to the posting online.

Review Scores, Value: This is the average value rating for each listing. The user rates the listing based on its value.

Exploratory Data Analysis

City Statistics

After importing the data, we want to compare different cities to look for trends and correlations in the listing data. To begin, we create a DataFrame with a statistical summary for each feature in every city. The get_stats and reorder function are used to create this DataFrame row for each city:

	Number of Guests Accommodated					Number of Bathrooms				
	Count	Max	Mean	Min	STD	Count	Max	Mean	Min	STD
Location										
Austin, TX	5835	16	4.388174807	1	2.689056023	5789	8	1.479616514	0	0.77462954
	Number of Bedrooms					Price				
	Count	Max	Mean	Min	STD	Count	Max	Mean	Min	STD
	5829	10	1.738720192	0	1.136759507	**5612**	999	231.5032074	0	198.997221
	Number of Minimum					Availability to Book in next 365 Days				
	Count	Max	Mean	Min	STD	Count	Max	Mean	Min	STD
	5835	365	2.101799486	1	5.584970697	5835	365	269.7386461	0	119.6367993
	Number of Reviews					Review Scores Rating				
	Count	Max	Mean	Min	STD	Count	Max	Mean	Min	STD
	5835	314	10.79468723	0	25.4055958	3789	100	95.43573502	20	7.260422023
	Review Scores, Accuracy					Review Scores, Value				
	Count	Max	Mean	Min	STD	Count	Max	Mean	Min	STD
	3776	10	9.64565678	2	0.757809456	**3778**	10	9.416093171	2	0.892990611

We now have descriptive statistics for the features in all cities (the figure above displays the row for our first city, Austin, in stacked form). Our analysis begins by filtering, sorting and plotting this statistical data to get an overview of Airbnb listing behavior in U.S. cities.

Population and Popularity: In which cities are there the most Airbnb listings?

Count of Listings

We begin by analyzing the listing count normalized by the population of each metro area. Looking at this normalized data, while it is true that Los Angeles and New York City have a large amount of listings, when divided by the population, we see that the listing count is very dependent on the population. Now there are some more interesting points to extract.

City	Count of Listings	Population	Normalized by Population
New Orleans, LA	4514	1.262888	3574.347052
Austin, TX	5835	2.00086	2916.246014
San Diego, CA	6608	3.299521	2002.71494
New York, NY	39553	20.182305	1959.78606
Los Angeles-Long Beach-Anaheim, CA	26080	13.340068	1955.012523
San Francisco, CA	8619	4.656132	1851.107314
Nashville, TN	3277	1.830345	1790.372853
Portland, OR	3360	2.389228	1406.311997
Seattle, WA	3818	3.73358	1022.611006
Denver, CO	2505	2.81433	890.0875164
Boston, MA	3585	4.774321	750.8921164
Washington, DC	3723	6.097684	610.5596813

New Orleans, Austin and San Diego have the highest normalized listing count. These cities out rank the front-runners of the un-normalized list, LA and NYC, which follow.

But what does this normalized listing count actually indicate? This count tells us the number of listings each city would have disregarding population. In other words, it tells us the number of listings there would be in each city if they all had the same population.

The higher a city ranks in normalized listing count, the more hosts are listing their homes on the Airbnb website compared to others. Thus, a higher normalized listing count indicates that there is a higher supply of listings in those cities. This high supply may also suggest that these are cities where the Airbnb market is strong in terms of host listings.

This would mean that cities with a lower normalized listing count are ones where Airbnb is not as popular from a supply or host side of the service. In cities such as Denver, Boston and Washington DC, Airbnb may want to investigate why the supply of listings in these areas are low using more granular data they possess regarding these areas.

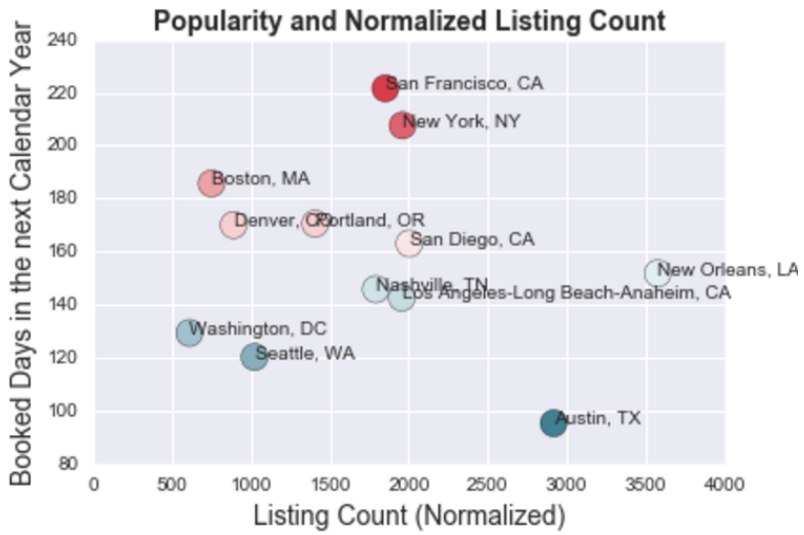
Recommendations and Takeaways:

- Population does not necessarily indicate listing density in a city. By normalizing the data, we get a look at the density, or supply of listings in a city regardless of its population.
- Denver, Washington and Boston are markets with a low supply of listings relative to population. It may be worth looking into why this is and implement a marketing plan to engage more of the population in listing their homes on Airbnb.

Relationship between the 'Supply' of Listings and their 'Demand'

Let's look at what happens when we pair this 'supply' side data with a parameter that estimates demand. 'Demand' of a listing in the context of this data can be estimated by the popularity of a city to visit. This is indicated by the availability of a listing in the next calendar year. Plotting the mean availability (demand indicator) against the normalized listing count (supply indicator) we can explore the relationship.

We take our mean availability metric (which measures the average number of days each listing is available to book in the next calendar year) and subtract 365 to get the number of days the average listing is booked in each city. A higher number indicates that the city is more popular, and therefore has more 'demand'.



There seems to be a general trend of increasing supply of listing, as the city gets more popular, with the exception of Austin and New Orleans. These two cities' high supply of listings is not in response with their demand. This should not be an issue from Airbnb's eyes as there will always be enough places for their users to stay if they chose to travel to New Orleans or Austin. The rest of the cities seem to have a suitable supply of listings given the popularity. This implies that Airbnb is doing a good job encouraging individuals to host their homes as listings given the demand of travel in those cities.

Recommendations:

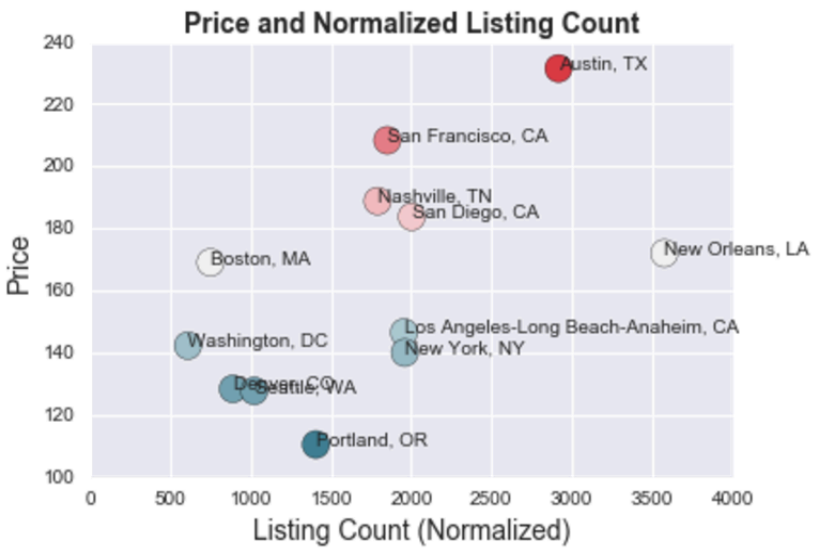
- Airbnb appears to be doing a good job matching the average supply and demand of listings, according to our estimators here. The demand and supply of listings seems to follow a consistent upward trend.
- Exceptions are New Orleans and Austin, where there are far more listings for the population given the demand. Airbnb may want to investigate this further.

Price: Where is it the most expensive and cheapest to stay, on average?

Austin, San Francisco and Nashville are the most expensive cities to stay in an Airbnb, while Denver, Seattle and Portland are the least expensive. These results are interesting, but it is difficult to explain why the cities fall in the order they do. Is it because of supply and demand? This average price metric is more interesting when we pair it with other data sets and explore relationships.

Average Price and Normalized Listing Count

Price and listing count have an obvious correlation. As the normalized number of listings increases, price increases as well. When we run the correlation coefficient between these two variables we get .54, which is an indication of a strong positive relationship. So as the supply of listings increase, it does not necessarily mean that the increased competition due to increased supply drives down price.



Location	Average Listing Price
Austin, TX	231.5032074
San Francisco, CA	208.3764299
Nashville, TN	188.8362519
San Diego, CA	183.7897499
New Orleans, LA	172.0319982
Boston, MA	169.072768
Los Angeles-Long Beach-Anaheim, CA	146.5676098
Washington, DC	142.3232514
New York, NY	140.0578267
Denver, CO	128.3929572
Seattle, WA	127.7477076
Portland, OR	110.5080405

There are places where there is wide dispersion in price for similar listing count and the relationship is not as strong. San Francisco, Nashville, Los Angeles and New York City have the similar normalized listing counts but fall into two distinct price groups. Los Angeles and New York City have lower average prices. It may be worthwhile for Airbnb to explore the reason for this.

It is also worth noting that New Orleans is an outlier yet again, with its high normalized listing count noted above.

Recommendations:

- Look into cities where there is a lot of dispersion in price for same normalized listing count or density of listings, given the population count. Could this be an indication about the demand of listings in those cities?

Popularity Indicators: What cities are the most popular to visit and stay at an Airbnb listing?

There are a few metrics we will use to estimate the average popularity of listings per city. The first metric we will analyze is the average number of days the listings of a city are booked in the next year. The higher the average number of booked days, the more popular a city's listings are. The most popular travel destinations for users are San Francisco, New York and Boston. This result is not surprising because they are large cities. Portland and Denver are fourth and fifth most popular. These cities' listings are characterized by a low number of listings and low price. Users may be drawn to these cities due to their low prices.

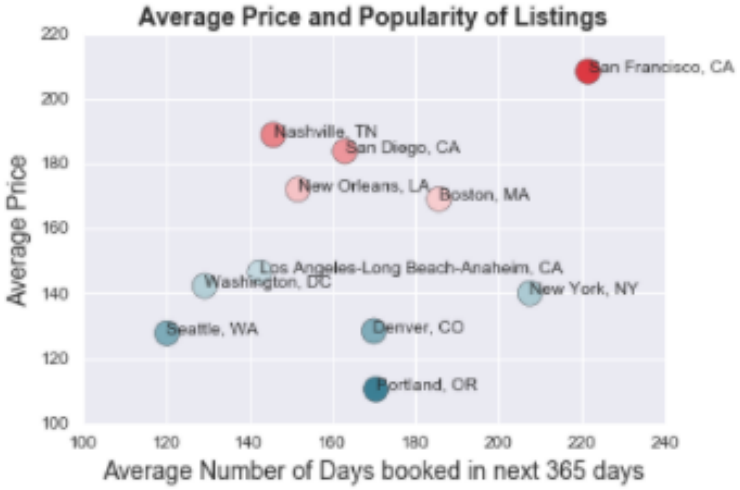
The second metric we use to determine the popularity of listings in a particular city is the mean number of reviews per city. We use the number of reviews as a proxy for estimating how frequently the listing is booked. The reasoning is that more reviews a listing has, the more times it is booked and therefore the more popular it is. Here we are using the average number of reviews for each city's listings. There are similarities in how the cities rank in both metrics. Seattle jumps from the bottom of the

list on availability to the top on average number of reviews. So although Seattle is not a popular place for travelers relative to the other cities included in the data set, it is well reviewed, all on average. This may be an indication of the superior value of Seattle Airbnb listings.

Location	Avg. Availability to Book in next 365 Days
San Francisco, CA	221.5004061
New York, NY	207.5264076
Boston, MA	185.6535565
Portland, OR	170.48125
Denver, CO	169.988024
San Diego, CA	163.0230024
New Orleans, LA	151.7408064
Nashville, TN	145.797986
Los Angeles-Long Beach-Anaheim, CA	142.5786043
Washington, DC	129.2976095
Seattle, WA	120.2273442
Austin, TX	95.2613539

Location	Avg. Number of Reviews
Portland, OR	32.64375
Nashville, TN	25.4436985
New Orleans, LA	24.71953921
Seattle, WA	22.2234154
San Francisco, CA	19.6933519
Boston, MA	19.0446304
Denver, CO	18.54451098
Los Angeles-Long Beach-Anaheim, CA	16.84873466
Washington, DC	15.30674187
New York, NY	14.53174222
San Diego, CA	14.05266344

Next, we plot average availability per city versus its average price to get an idea of how popularity and price are related. Austin is a major outlier with a high price and high availability. This relationship does not fit the trend of the rest of the data, so we drop Austin from our analysis of this relationship. After doing so, there is a slight positive correlation between the price and the popularity of listing with a correlation coefficient of .3405. This coefficient means that, on average, the more popular a listing (higher the number of days booked) the higher the price of the listing. This seems to suggest that cities where it is more popular to travel have high demand, which drives up price, on average.



Recommendations:

- This relationship between price and popularity suggests how prices of listings may fluctuate as cities become 'trendier' and less 'trendier' to travel to. If Airbnb can track booking dates of different cities over time, they can identify cities where booked dates are increasing in the upcoming calendar year. Using this information, they may be able to also identify where prices are going to be driven up by demand.

The Zillow Data

What is the Zillow Home Value Index (ZHVI)?

Unlike commodities and consumer goods, for which we can observe prices in all time periods, we can not observe prices on the same set of homes in all time periods because not all homes are sold in every time period. Zillow has developed a way of approximating the ideal home price index. Instead of actual sale prices on every home, the index is created from estimated sale prices on every home. Because of this, the distribution of actual sale prices for homes sold in a given time period looks very similar to the distribution of estimated sale prices for this same set of homes. But, importantly, Zillow has estimated sale prices not just for the homes that sold, but for all homes even if they did not sell in that time period.

Using this methodology, we now have a comprehensive and robust benchmark of home value trends that can be computed which is immune to the changing mix of properties that sell in different periods of time.

This section was annotated from: *Methodology for determining the ZHVI:* http://www.zillow.com/research/zhvi-methodology-6032/

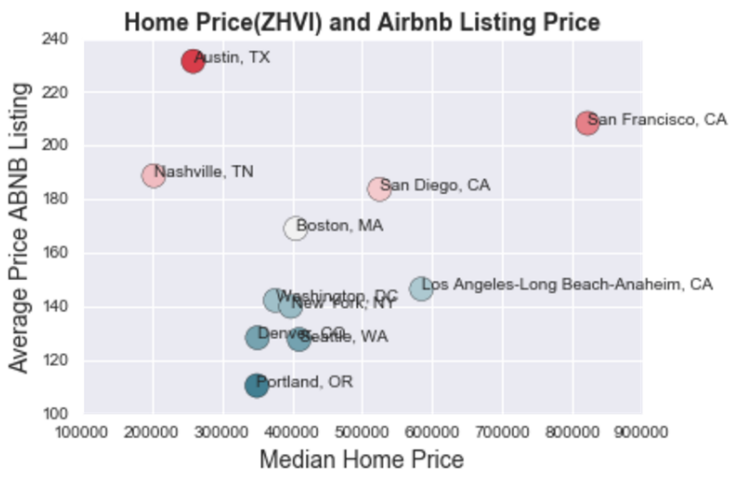
The Data

From the Zillow website we have imported data on the ZHVI or median home price of every city. This data includes the growth rate of the index over various periods of time including: Month over Month, Quarter over Quarter, Year over Year, the 5 year change and the 10 year change.

Location	ZHVI	ZHVI Month over Month	ZHVI Quarter over Quarter	ZHVI Year over Year	ZHVI 5 Year Change	ZHVI 10 Year Change
Austin, TX	258600	0.00271423	0.012133072	0.083822297	0.082968339	0.039407092
Boston, MA	405200	0.007208551	0.019114688	0.059623431	0.053598476	0.010530806
Washington, DC	375900	0.004274646	0.011571582	0.026488258	0.037900477	-0.010914042
Denver, CO	350400	0.007475561	0.023065693	0.101886792	0.102926091	0.040801976
Los Angeles-Long Beach-Anaheim, CA	584700	0.005157298	0.014751822	0.059815117	0.083030829	-0.002961283
Nashville, TN	202400	0.012	0.031600408	0.114537445	0.066052602	0.027839769
New York, NY	398000	0.005558363	0.023136247	0.05738576	0.02941196	-0.011077936
Portland, OR	349500	0.009240543	0.029455081	0.146277468	0.101232824	0.020503599
San Diego, CA	524900	0.004016832	0.015084123	0.064058382	0.0862991	-0.000266327
Seattle, WA	409900	0.008364084	0.023215177	0.122091432	0.093058911	0.011144027
San Francisco, CA	821800	0.005136986	0.015445447	0.053724837	0.113259665	0.01701686

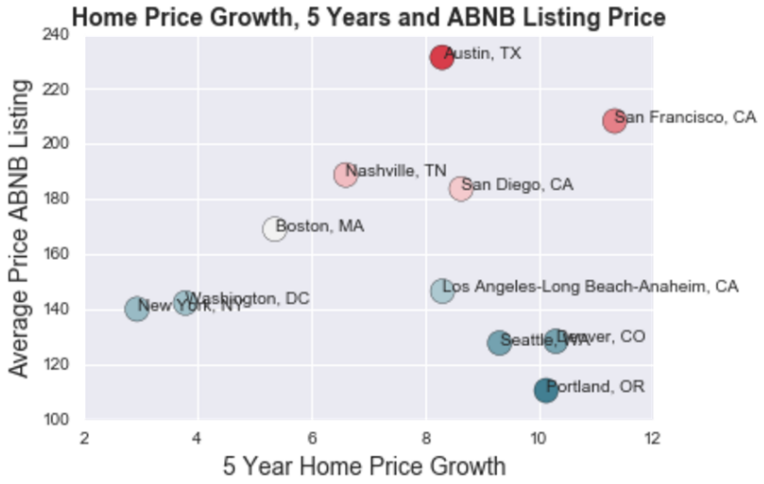
How does the median home price of a city influence the price of an Airbnb listing?

The average price of a listing seems to increase as the median home value increases, as measured by the Zillow Home Value Index. However, Nashville and Austin are two outliers, with low median home values and high average listing prices. By dropping these two outliers we increase our correlation coefficient between ZHVI and average List Price from .11 to .81. This confirms that there is a very strong relationship between the median home value and average list price.



How does the change in home price over five years correlate to the price? Are listings in emerging cities, with high growth, priced at a higher value?

A little over half the data follow the relationship of the average listing price increasing as home value increases. However, LA, Seattle, Denver and Portland have had high growth in home values over the last five years yet have maintained relatively low prices.



Machine Learning: Regression Model for Predicting Price

The purpose of this section is to develop a linear regression model for predicting the price of a listing based on the Airbnb feature data described above and the Zillow Home Value Index. Building this model will also help further investigate the relationship between price and the selected features.

Feature Selection

We will be using the individual listing data from the following set of cities explored above:

Austin, TX; Boston, MA; Washington, DC; Denver, CO; Los Angeles, CA; Nashville, TN; NYC, NY; Portland, OR; San Diego, CA; Seattle, WA and San Francisco, CA

We begin by appending all the listing data from these cities into one table. The next step is adding on the appropriate Zillow home value data for each listing based on the zip code of the listing. We will merge the two data sets using the zip code

information contained in both data sets. That way more granular and accurate home value data will be paired to each listing, as opposed to just using the general metro zip code. Setting the index to the zip code for each data frame and merging on this index, we create a new table such that each listing has its relevant ZHVI, 5 Year ZHVI and 10 Year ZHVI appended to its Airbnb data. This data frame has all of the independent data and dependent data (price) necessary to fit the model.

Our dependent variable is the variable we will be predicting using the model. In our case it is listing price. Our features are the independent variables we will use to predict price. We begin by looking at the relationship between the features we plan on using and listing price to determine that they have a linear relationship. A linear relationship between the dependent variable and features is one of the requirements for building a linear model. If a linear relationship is not present, some features may require a transformation.

We evaluate linearity by plotting the price of all listings against each feature to ensure it has a linear relationship. The only two variables that needed transforming are the number of reviews and availability. Before transforming with a log function price exponentially decreases as number of reviews increases. After taking the log of the number of reviews, a linear relationship emerges between these variables. With Availability, there was an exponential increase in price as the availability increases. Again, taking the log of the availability, a linear relationship emerges between these variables. Also, the sum of all ratings was taken and used as a feature as this variable had a stronger relationship to listing price.

Another requirement in creating a linear model are that independent variables, or features, do not have a strong relationships to one another, as this confounds the linear model. To check for this, we use a correlation matrix. Relationships between the variables are measured between -1 and 1. The higher the absolute number, the stronger the relationship, with the sign signifying the direction of the relationship. As a result of this, bedrooms was removed as feature as it had very strong correlations to bathrooms, accommodates and ZHVI. The final correlation matrix demonstrates no strong correlations between the independent variables.

Based on the evaluation of linearity and correlation, the Airbnb Features being used are:

- Log of Number of Reviews, Log of Availability in the next 365 days, Sum of Ratings (Overall + Value Rating + Accuracy Rating), # of Reviews, Bathrooms

and The Zillow Home Value Data Used:

- ZHVI (Median Home Price), ZHVI 5 Year Change, ZHVI 10 Year Change

We apply the necessary transformations to our merged feature DataFrame described at the beginning of this section and continue to build the model.

Correlation Matrix:

	Zhvi	Number of Reviews	Availability to Book in the next 365 Days	Review Total	ZHVI 10 Year Growth	ZHVI 5 Year Growth	Number of Guests Accommodated	Number of Bathrooms
Zhvi	1	-0.092974507	-0.003923569	-0.039193221	-0.148454157	-0.44560836	0.116747769	0.145473807
Number of Reviews	-0.092974507	1	0.25081461	0.091695089	0.233925209	0.211347632	-0.118742213	-0.164223424
Availability to Book in the next 365 Days	-0.003923569	0.25081461	1	-0.019996261	-0.025103196	-0.04122184	0.041699232	0.004626737
Review Total	-0.039193221	0.091695089	-0.019996261	1	0.099945285	0.085192216	-0.081789585	-0.048089963
ZHVI 10 Year Growth	-0.148454157	0.233925209	-0.025103196	0.099945285	1	0.713633301	-0.104282473	-0.111367897
ZHVI 5 Year Growth	-0.44560836	0.211347632	-0.04122184	0.085192216	0.713633301	1	-0.153528119	-0.175453604
Number of Guests Accommodated	0.116747769	-0.118742213	0.041699232	-0.081789585	-0.104282473	-0.153528119	1	0.624003938
Number of Bathrooms	0.145473807	-0.164223424	0.004626737	-0.048089963	-0.111367897	-0.175453604	0.624003938	1

Linear Model

Using the feature DataFrame with all of the chosen independent variables and our dependent variable, price, we split the data into training and testing sets, with 80% of the data used to train the data and 20% used to test it.

This means we will use 80% of our data to fit the model. The remaining 20% of the data is used to evaluate the performance of the model. It uses the model created by the training data to predict price using the listing features of the testing data. The model then compares this fitted data (or predicted price) to the actual price of each listing to determine the strength of the model. If the model is strong then the difference between the test data's predicted price (using the model) for a listing will be very close or exactly the same as the actual price of the listing.

The R-squared is the coefficient of determination and measures, as a percentage, how well our linear model fits the data, or how close the training and test data are to the fitted regression line. In the context of the Airbnb data, how good the model will do at predicting the price of a listing based on our selected features.

The R-squared for this initial model was 56.9% for the testing data. This coefficient of determination tells us that the features can explain about 56.9 percent of the total variance in the price of a listing. At 56%, a little over half of the variability can be attributed to our model. This is reasonable, but we may be able to achieve better results using other algorithms for linear regression, by changing the number of features or type of cities we include in our model. The test data performed somewhat similarly, which is another good sign for the data. The high MSE of around 6500 is a concern, but something that can be corrected by refitting the model.

The coefficient for each feature describes how much the listing price changes for a one unit increase in said feature net of all others. The 10 Year change and 5 Year change in median home value seem to be the most powerful predictors. They have a

large impact on the price of a listing net of all other features changing. For example, a one percent increase in home value over 5 years increases, on average, the price of a listing by $88 net of all other features.

```
R-squared on Training Data: 0.635793397613
R=squared on Testing Data: 0.569124420483
Mean squared error on the test set:  6528.8317681
Mean squared error on the training set:  4981.23342123
```

Feature	Coefficient
Zhvi	8.83E-05
number_of_reviews	-6.692230095
availability_365	1.926249724
reviewtotal	0.704794579
10Year	-379.3361971
5Year	88.82070895
accommodates	25.29388748
bathrooms	42.51968205

We will use the mean squared error along with the R-squared to compare the accuracy and linear fit to others as we try to improve the model. There is definitely room for improvement, especially with MSE and R-squared.

The residuals above describe the difference in the actual listing price and predicted listing price using the model. The residuals should be distributed randomly above and below the axis tending to cluster towards the middle without any clear patterns. These qualities are indicative of a good linear model, with a normal distribution. Normality in the residuals is important, as it is an assumption in creating a linear model. The model seems to be heteroskedastic, meaning that the residuals get larger as the prediction moves from large to small. The model is better at predicting prices for average priced listings at around 300 dollars a night or lower. For those values, the residuals are evenly and randomly distributed above and below the axis with no pattern, indication of a good model. The heteroskedasticity of the residuals may seem like an issue. However, an analysis of the distribution of listing prices confirms that a majority of listings fall below this $300 benchmark. So, the model does a good job predicting prices for a majority of listings. The define the model as being best used for predicting the price of listings that are an average price. We can therefore disregard the heteroskedastic effect on the residuals.

Improving the Model: Predicting Price for Highly Popular Cities

Location	Percent Availability in the Upcoming Year
San Francisco, CA	39.31%
New York, NY	43.14%
Boston, MA	49.14%
Portland, OR	53.29%
Denver, CO	53.43%
San Diego, CA	55.34%
New Orleans, LA	58.43%
Nashville, TN	60.06%
Los Angeles-Long Beach-Anaheim, CA	60.94%
Washington, DC	64.58%
Seattle, WA	67.06%
Austin, TX	73.90%

We may be able to achieve a better prediction model and higher R-squared if we segment the cities by a strong variable such as Availability. The idea is that we will have a higher fit by putting similar listings together, resulting in a stronger relationship. Recall from the exploratory analysis above that the most popular cities are those that have the smallest percent average availability in the next year (table on left).

We will segment the Highly Popular (Less that 60% availability in the coming year) from the rest of the samples of cities. We might be able to predict price better for more popular cities. Highly Popular Cities used to fit a new model are:

San Francisco, NYC, Boston, Portland and San Diego

This model's coefficient of determination tells us that the linear regression model can explain around 64.13% percentage of the total variance in the price of a listing from a popular city. At 64%, almost two thirds of the variability can be attributed to our features. This is much higher than our original linear model using all cities. This better result proves that by grouping popular cities, we have created a stronger linear model. These city's listing prices react similarly to changes in their features and create a better linear fit.

```
R-squared on Training Data: 0.614727537965
R=squared on Testing Data: 0.641339658715
Mean squared error on the test set:  5739.10422525
Mean squared error on the training set:  5184.12229197
```

Feature	Coefficient
Zhvi	8.38E-05
number_of_reviews	-7.749282016
availability_365	1.554674857
reviewtotal	0.677756394
10Year	-379.6332652
5Year	64.85262089
accommodates	26.0359725
bathrooms	38.64776199

The MSE is also reduced to 5700, another indication that by taking a subset we have created a better model.

The residual plot above indicates that the model seems to be less heteroskcedastic than the previous one. The range of the residuals looks to be tighter around the axis and still displays no pattern, indicative that we have retained normality in this new subset of our data.

This is a good starting point for developing a better model by manipulating the data set, however we may be able to increase the level of fit by using another algorithm to create a linear model.

Random Forest Regression for Popular Cities Model

Next, we will be using the Random Forest Regression model on our popular cities data to see if it results in a better fitting model. The Random Forest model uses an algorithm which 'bootstraps' (taking of many random samples from the training data) to build the ‘nodes’ of a decision or regression tree that models the behavior of the data to create a linear model.

After implementing the Random Forest Regression, fit improves. The model now has an R-squared of 71.2%, much higher than both initial models. The mean square error also drastically reduced to around 3700.

We plot the predicted price of the test data versus actual price to evaluate the model. The main effect apparent in this graph is a funneling effect, suggesting heteroskedasticity. As price increases, the model does not do as good of a job at predicting it. This funneling effect seems to compromise the accuracy of predicted price data where the price per night is above around 400. Before this price, the model does a good job predicting the price of listings. When we look at the distribution of price for the dataset used in creating the model, we see that a majority of the listings are priced below this benchmark. This means that a majority of listings are priced in a range that is able to be accurately predicted by the model. Therefore, it is established that this model is best used for 'average' priced listings and disregard the funneling effect.

Conclusion: Recommendations for Airbnb

The linear regression model reveals some interesting attributes regarding the relationship between price and various attributes of listings, which can be leveraged to help business strategy for Airbnb.

1) There is a big difference in the way the appreciation of home values impacts the price of an Airbnb listing over a 5 year and 10 year timeframe. This trend is present in all the models we looked at, but let us use our ‘Popular Cities’ model as an example here (the second model we created). Net of all other variables, the price of a listing increase by $64 for

every one-point increase in home value over five years. This is in stark contrast to the -$379 variable of the same measure over 10 Years. This suggests that as home values increase over a larger period of time, the price of listing decreases. A theory may be that it becomes more profitable for hosts to list their property as their area increases in value, which spikes supply and drives down price. The 10 Year figure may not be an accurate way to analyze listing value, as the real estate market can change dramatically in this time. Looking at the coefficient, the 5 year may be a more reasonable indicator. Given that most listings are priced around $100 a night, the fact that home value can impact the price by $64 seems like a significant variable to investigate.

Airbnb could utilize more granular listing data from wider range of US cities and develop a model that can use the 5 Year home value index to predict where prices will increase. In these areas it will be more profitable for hosts to open their homes to guests, and Airbnb could gain a larger foothold in those regions.

2) In terms of listing variables, the size of the listing seems to be one of the biggest influences on the price of a listing. The number of people the listing accommodates the bathrooms are some of the largest regression coefficients in the model. This relationship demonstrates how hosts and guests value listings. A way Airbnb can use this data could be as a way to control the average price of listings in an area. If they want to create more listings at lower prices for certain cities, Airbnb could make hosts post listings that accommodate less people. This relationship suggests that it may be a good next step to segment the data based on the size of listings to determine if there are more relationships to explore.

3) The exploratory analysis revealed relationships that can be used by Airbnb to improve their product:
   - Denver, Washington and Boston are markets with a low supply of listings relative to population. It may be worth looking into why this is and implement a marketing plan to engage more of the population in listing their homes on Airbnb.
   - New Orleans and Austin are cities where there are far more listings for the population given the demand. Airbnb may want to investigate this further. The marketing department may want to implement campaign that encourages users to travel to these cities given the high supply of host listings. This can be achieved through an email campaign or by changing website copy to encourage users to look into booking a trip to these cities using the site.
   - The relationship between price and popularity suggests how prices of listings may fluctuate as cities become ‘trendier’ and less ‘trendier’ to travel to. If Airbnb can track booking dates of different cities over time, they can identify cities where booked dates are increasing in the upcoming calendar year. Using this information, they may be able to also identify where prices are going to be driven up by demand.

The model and exploratory analysis conducted in this paper provided an interesting look into the behavior of Airbnb listings for popular travel destinations in the US. However the model and analysis could be greatly improved upon with data from more cities. Improved correlations between variables and clearer relationships could be established with more listing information from a wider variety of cities. The conclusions established from this analysis would also be more significant. Another expansion on this analysis would be to take it beyond major cities and travel destinations in the US. The data used here does not provide a look into the nature of the Airbnb market in midsized or smaller cities. There could be an untapped market for short-term rentals or local travel in these areas that should be explored.

There are also a number of conclusions and recommendations presented in this paper that can be further evaluated using the proprietary information that only the Airbnb may have access to. An interesting addition to this analysis would be to

supplement the data here with user web page behavior for each listing (how many clicks a listing gets, how long it is viewed etc.) Since Airbnb is an online product, the way that users interact with the 'service' online is a very important part of the experience and could reveal insights into the way users value listings in differently depending various characteristics of their trip. These characteristics are entered into the search parameters on the Airbnb web page and include the users: length of stay, preferred room type, location they are travelling to, etc.