Airbnb and the Real Estate Market

David Kohler, Karen Ou, Siyi Xu, Caroline Breaux, Maahin Puri

Duke University – COMPSCI 216

Introduction

- The rise of Airbnb since its founding in 2008 has fundamentally transformed the hospitality industry and the way in which people around the world are traveling.
- Its exponential growth in recent years relies heavily on its innovative digital solution to the traditional practices of home-renting.
- Curious about how the listing prices are determined, our team set out to develop a price prediction model that would not only shed light on which factors may have influenced listing prices in different geographical locations, but also provide insight into the market potential of a "new", or test, city if Airbnb were to expand into a city in which it had minimal host network.
- Furthermore, we brought in Zillow's real estate data to see if there was any correlation between Airbnb listing prices and home values across different cities.
- The price prediction model would not only provide insight into what new regions Airbnb should target but also assist new hosts in giving them a sense of what price they should charge for a new listing.

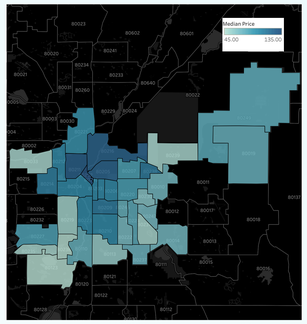
Data and Methodology

- Our group gathered raw data from Inside Airbnb and Zillow research to develop a model for predicting Airbnb listing prices.
- The Airbnb data contained variables such as neighborhood, room type and amenities for each listing, and the Zillow data set contained median home listing prices and home value indices in different areas based on zip code.
- Since nearby real estate value was a significant component, we obtained housing index per zip code from the Zillow data set and normalized it to incorporate it in our model.
- Out of the 106 variables in the original Airbnb dataset, 16 variables were chosen as potential price influencing factors to be used in the prediction model. After data cleaning and feature engineering, we included 42 features in regression.
- For the prediction model, we used various regression algorithms including linear regression, Ridge regression, Lasso regression, decision trees, random forest, and support vector regression (SVR) to see which model achieves the best accuracy on test data.

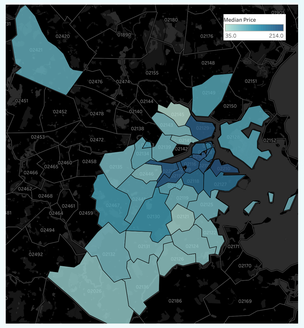
	R Squared	MSE
Linear Regression	0.56	0.31
Ridge Regression	0.56	0.31
Lasso Regression	0.54	0.32
Random Forest	0.52	0.34
SVR	0.53	0.33
Decision Tree	0.53	0.33

Visualizations

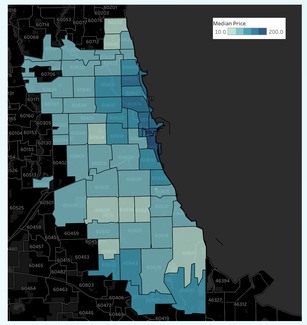
Median Airbnb Listing Price in Denver By Zip code



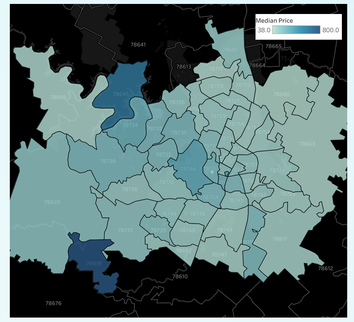
Median Airbnb Listing Price in Boston By Zip code



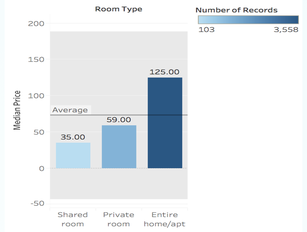
Median Airbnb Listing Price in Chicago By Zip code



Median Airbnb Listing Price in Austin By Zip code

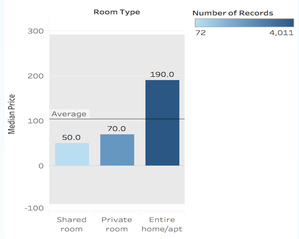


Median Price & Number of listings by Room Type (Denver)



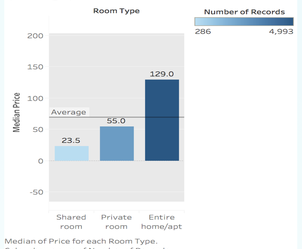
Median Price for each Room Type.
Color shows sum of Number of Records.

Median Price & Number of listings by Room Type (Boston)



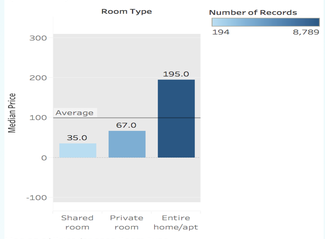
Median Price for each Room Type.
Color shows sum of Number of Records.

Median Price & Number of listings by Room Type (Chicago)



Median Price for each Room Type.
Color shows sum of Number of Records.

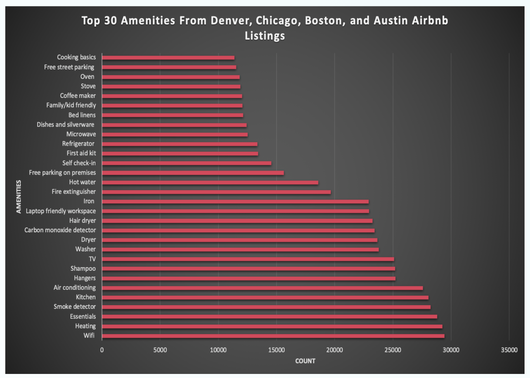
Median Price & Number of listings by Room Type (Austin)



Median Price for each Room Type.
Color shows sum of Number of Records.

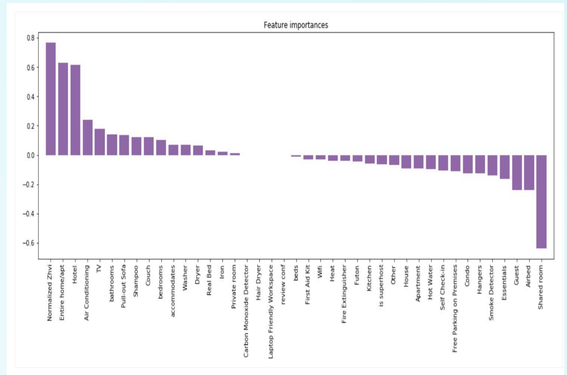
Word Cloud of Reviews

Graph for Amenities Count



Results

- Using k-fold cross-validation, linear and ridge regression models produced the most optimal results on the test datasets with an R-squared value and mean squared error (MSE) of 0.56 and 0.31, respectively.
- The most significant positive contributor to price was the normalized Zillow Home Value Index, and the most significant negative contributor was whether the listing was a shared room.



$$\begin{aligned}\text{logprice} = 4.05 &+ (-0.066725) \times \text{is.superhost} + 0.068651 \times \text{accommodates} \\ &+ 0.139422 \times \text{bathrooms} + 0.100926 \times \text{bedrooms} + (-0.009845) \times \text{beds} \\ &+ 0.625996 \times \text{Entire home apt} + 0.007420 \times \text{Private room} \\ &+ (-0.633416) \times \text{Shared room} + (-0.224364) \times \text{Airbed} + 0.097956 \times \text{Couch} \\ &+ (-0.036596) \times \text{Couch} + 0.127272 \times \text{Pull out Sofa} + 0.035733 \times \text{Real Bed} \\ &+ (-0.087978) \times \text{Apartment} + (-0.121576) \times \text{Condo} + (-0.234617) \times \text{Guest} \\ &+ 0.598255 \times \text{Hotel} + (-0.088035) \times \text{House} + (-0.066049) \times \text{Other} \\ &+ 0.764261 \times \text{Normalized Zhvi} + (-0.029826) \times \text{Wifi} + (-0.038886) \times \text{Heat} \\ &+ (-0.161803) \times \text{Essentials} + (-0.137893) \times \text{Smoke Detector} \\ &+ (-0.059792) \times \text{Kitchen} + 0.238660 \times \text{Air Conditioning} + (-0.126428) \times \text{Hanger} \\ &+ 0.121606 \times \text{Shampoo} + 0.179469 \times \text{TV} + 0.066489 \times \text{Washer} \\ &+ 0.063873 \times \text{Dryer} + (-0.001695) \times \text{Laptop Friendly Workspace} \\ &+ 0.023393 \times \text{Iron} + (-0.040146) \times \text{Fire Extinguisher} + (-0.099524) \times \text{Hot Water} \\ &+ (-0.112362) \times \text{Free Parking on Premises} + (-0.105954) \times \text{Self Check in} \\ &+ (-0.029816) \times \text{First Aid Kit} + (-0.002957) \times \text{review conf}\end{aligned}$$

Conclusion

We found it surprising that the linear regression model outperforms other more sophisticated models in terms of prediction accuracy. The fact that the normalized Zillow Home Value Index has the largest, positive coefficient in the linear regression model indicates that the real estate value largely influences the Airbnb listing price. Also, contrary to our hypothesis, being a superhost is negatively correlated with the Airbnb listing price. We also created a script that can be ran by our consumer to predict how much they should charge per night for their listing. To the right is part of what the user sees and is prompted:

```
>> python new_listing.py
Are you a superhost? (Y/N)
n
Enter number of people your listing accommodates:
4
Enter number of bathrooms in your listing:
2
Enter number of bedrooms in your listing:
2
Enter number of beds in your listing:
2
Enter number corresponding to room type of listing:
1. Entire home/apt
2. Private room
3. Shared room
1
Enter Zipcode of listing:
78654
Enter City of listing:
Horseshoe Bay
Enter number corresponding to bed type of listing:
.
.
.
_____
Using Decision Tree Model: $ 332.0986236602581 per night
Using Random Forest Model: $ 217.15587457726899 per night
Using Linear Regression Model: $ 258.5891285890453 per night
Using Ridge Regression Model: $ 258.31528399984787 per night
Using Lasso Regression Model: $ 262.69312240990837 per night
```