

CS6140 Project Proposal Presentation

Abhay Kasturia, Nakul Cammasamudram, Philip Parker

##	id	experiences_offered	host_is_superhost	host_id
## 1	20872145	none	f	
## 2	13937181	none	t	
## 3	3575220	none	t	
## 4	17419858	none	f	
## 5	18202496	none	f	
## 6	16987125	none	f	

##	neighbourhood_cleansed	property_type	room_type	accommodates
## 1	Roslindale	Apartment	Private room	
## 2	Roslindale	House	Private room	
## 3	Roslindale	House	Private room	
## 4	Roslindale	Apartment	Private room	
## 5	Roslindale	House	Private room	
## 6	Roslindale	Apartment	Private room	

##	bedrooms	beds
## 1	0	1
## 2	1	1
## 3	1	1
## 4	1	1
## 5	1	1

Overview

Airbnb is a popular online company through which property owners can short-term rent their space to consumers as an alternative to hotels. Using data on Airbnb property listings in Boston, we will develop and test methods for determining the optimal price per night an owner should set for their property.

Dataset: data source

The Inside Airbnb project by Murray Cox has collected public Airbnb listing data for 40+ popular international cities. We will use the Boston dataset containing ~5000 property listings and their features for the Boston area.

(<http://insideairbnb.com/get-the-data.html>)

Dataset: data features

- ▶ *super_host?* (categorical) [yes/no]
- ▶ *verified_host?* (categorical) [yes/no]
- ▶ *zip_code* (categorical) [...]
- ▶ *property_type* (categorical) [House, Apartment, etc.]
- ▶ *room_type* (categorical) [Shared Room, Private Room, etc.]
- ▶ *accomodates* (continuous)
- ▶ *bathrooms* (continuous)
- ▶ *bedrooms* (continuous)
- ▶ *beds* (continuous)
- ▶ *bed_type* (categorical) [Real Bed, Futon, etc.]
- ▶ *minimum_nights* (continuous)
- ▶ *cancelation_policy* (ordered categorical) [Flexible, Moderate, Strict]
- ▶ **price** (continuous)

Dataset: initial data analysis

One-variable summary statistics



X	host_is_superhost	host_identity_verified	zipcode	property_type
Min. : 1	f:3928	f:1936	02116 : 478	Apartment :3331
1st Qu.:1218	t: 942	t:2934	02130 : 419	House : 782
Median :2436			02114 : 314	Condominium: 465
Mean :2436			02215 : 310	Other : 87
3rd Qu.:3653			02118 : 290	Townhouse : 72
Max. :4870			02134 : 285	Loft : 34
			(Other):2774	(Other) : 99

room_type	accommodates	bathrooms	bedrooms	beds	bed_type
Entire home/apt:3031	Min. : 1.000	Min. :0.000	Min. : 0.000	Min. : 0.00	Airbed : 38
Private room :1782	1st Qu.: 2.000	1st Qu.:1.000	1st Qu.: 1.000	1st Qu.: 1.00	Couch : 7
Shared room : 57	Median : 2.000	Median :1.000	Median : 1.000	Median : 1.00	Futon : 40
	Mean : 3.298	Mean :1.246	Mean : 1.337	Mean : 1.75	Pull-out Sofa: 24
	3rd Qu.: 4.000	3rd Qu.:1.000	3rd Qu.: 2.000	3rd Qu.: 2.00	Real Bed :4761
	Max. :16.000	Max. :6.000	Max. :10.000	Max. :16.00	
		NA's :8	NA's :3	NA's :3	

price	guests_included	minimum_nights	number_of_reviews	review_scores_rating	instant_bookable
\$200.00: 192	Min. : 1.000	Min. : 1.000	Min. : 0.00	Min. : 20.00	f:3172
\$150.00: 156	1st Qu.: 1.000	1st Qu.: 1.000	1st Qu.: 1.00	1st Qu.: 90.00	t:1698
\$100.00: 128	Median : 1.000	Median : 2.000	Median : 7.00	Median : 96.00	
\$75.00 : 118	Mean : 1.528	Mean : 3.531	Mean : 24.79	Mean : 93.37	
\$99.00 : 113	3rd Qu.: 2.000	3rd Qu.: 3.000	3rd Qu.: 28.00	3rd Qu.: 99.00	
\$250.00: 111	Max. :16.000	Max. :365.000	Max. :401.00	Max. :100.00	
(Other):4052				NA's :926	

is_business_travel_ready	cancellation_policy	reviews_per_month
f:4125	flexible :1145	Min. : 0.020
t: 745	moderate :1174	1st Qu.: 0.450
	strict :2505	Median : 1.300
	super_strict_30: 44	Mean : 2.205
	super_strict_60: 2	3rd Qu.: 3.248
		Max. :15.100
		NA's :884

Dataset: initial data analysis (cont.)

Two-variable summary statistics

```
##          id          experiences_offered host_is_superhos
## Min.      :    3781 none:4870          f:3928
## 1st Qu.: 7913572          t: 942
## Median :14313760
## Mean     :12936513
## 3rd Qu.:18579076
## Max.     :21231173
##
## host_identity_verified  neighbourhood_cleansed  pro
## f:1936                  Jamaica Plain: 416      Apartme
## t:2934                  Back Bay      : 410      House
##                      Dorchester    : 400      Condomi
##                      Fenway        : 357      Other
##                      South End     : 354      Townhou
##                      Allston       : 297      Loft
##                      (Other)      :2636      (Other)
```

Dataset: initial data analysis (cont.)

Two-variable summary statistics (cont.)

Dataset: initial data analysis (cont.)

Missing values

```
## [1] 14
```

Methods

- ▶ We will perform and compare a variety of regression techniques, including linear regression and kernel methods. The comparison between the results of these methods is straightforward.
- ▶ We will also consider approaching the problem from a classification point of view, dividing the prices into ordered categorical ranges. Doing this will allow us to investigate the use of classification methods such as logistic regression and tree-based approaches.
- ▶ We will research/develop a means of comparing the results of the regression and classification methods.