CS6140 Project Proposal Presentation

Abhay Kasturia, Nakul Cammasamudram, Philip Parker

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

```
host_is_superhost host_identity_verified    zipcode         property_type            room_type
f:3886            f:1919                  02116  : 476   Apartment  :3296   Entire home/apt:3004
t: 932            t:2899                  02130  : 418   House      : 778   Private room   :1758
                                          02114  : 314   Condominium: 459   Shared room    :  56
                                          02215  : 310   Other      :  84
                                          02118  : 287   Townhouse  :  72
                                          02134  : 285   Loft       :  34
                                          (Other):2728   (Other)    :  95
 accommodates      bathrooms        bedrooms          beds                bed_type        price
Min.   : 1.000   Min.   :0.000   Min.   : 0.000   Min.   : 0.000   Airbed       :  37   Min.   :   0.0
1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.: 1.000   Couch        :   7   1st Qu.:  80.0
Median : 3.000   Median :1.000   Median : 1.000   Median : 1.000   Futon        :  40   Median : 140.0
Mean   : 3.308   Mean   :1.247   Mean   : 1.341   Mean   : 1.754   Pull-out Sofa:  24   Mean   : 173.1
3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 2.000   3rd Qu.: 2.000   Real Bed     :4710   3rd Qu.: 200.0
Max.   :16.000   Max.   :6.000   Max.   :10.000   Max.   :16.000                        Max.   :4000.0

guests_included  minimum_nights    number_of_reviews instant_bookable is_business_travel_ready
Min.   : 1.000   Min.   :  1.000   Min.   :  0.00    f:3132           f:4076
1st Qu.: 1.000   1st Qu.:  1.000   1st Qu.:  1.00    t:1686           t: 742
Median : 1.000   Median :  2.000   Median :  7.00
Mean   : 1.529   Mean   :  3.541   Mean   : 24.79
3rd Qu.: 2.000   3rd Qu.:  3.000   3rd Qu.: 28.00
Max.   :16.000   Max.   :365.000   Max.   :401.00

     cancellation_policy
flexible       :1125
moderate       :1155
strict         :2494
super_strict_30:  42
super_strict_60:   2
```

Dataset: initial data analysis (cont.)

Two-variable summary statistics

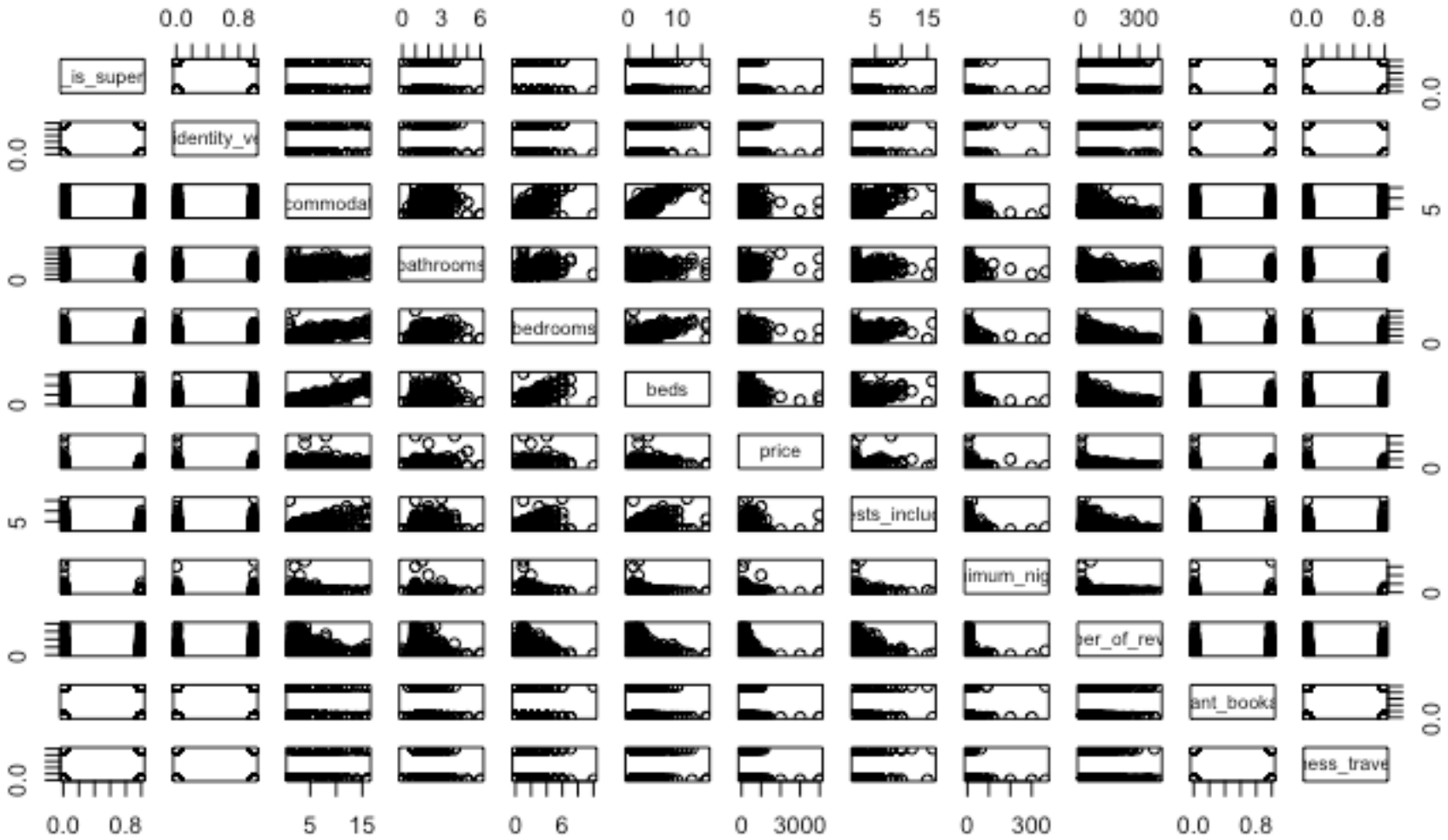

Figure 2:

Dataset: initial data analysis (cont.)

Missing values

There were 53 rows out of ~5000 with missing values, so we removed them.

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.