CS6140 Project

Authors

- Abhay Kasturia [MSCS, fourth semester]
- Nakul Cammasamudram [MSCS, third semester]
- Philip Parker [MSCS, third semester]

Description of problem

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.

Summary of data

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)

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)

One-variable summary statistics

```
summary(boston.data)
```

```
##  host_is_superhost host_identity_verified    zipcode
##  f:3886            f:1919                 02116  : 476
##  t: 932            t:2899                 02130  : 418
##                                           02114  : 314
##                                           02215  : 310
##                                           02118  : 287
##                                           02134  : 285
##                                           (Other):2728
##      property_type            room_type     accommodates
```

```
##  Apartment  :3296   Entire home/apt:3004   Min.   : 1.000
##  House      : 778   Private room   :1758   1st Qu.: 2.000
##  Condominium: 459   Shared room    :  56   Median : 3.000
##  Other      :  84                          Mean   : 3.308
##  Townhouse  :  72                          3rd Qu.: 4.000
##  Loft       :  34                          Max.   :16.000
##  (Other)    :  95
##    bathrooms        bedrooms           beds                bed_type
##  Min.   :0.000   Min.   : 0.000   Min.   : 0.000   Airbed       :  37
##  1st Qu.:1.000   1st Qu.: 1.000   1st Qu.: 1.000   Couch        :   7
##  Median :1.000   Median : 1.000   Median : 1.000   Futon        :  40
##  Mean   :1.247   Mean   : 1.341   Mean   : 1.754   Pull-out Sofa:  24
##  3rd Qu.:1.000   3rd Qu.: 2.000   3rd Qu.: 2.000   Real Bed     :4710
##  Max.   :6.000   Max.   :10.000   Max.   :16.000
##
##      price        guests_included  minimum_nights    number_of_reviews
##  Min.   :   0.0   Min.   : 1.000   Min.   :  1.000   Min.   :  0.00
##  1st Qu.:  80.0   1st Qu.: 1.000   1st Qu.:  1.000   1st Qu.:  1.00
##  Median : 140.0   Median : 1.000   Median :  2.000   Median :  7.00
##  Mean   : 173.1   Mean   : 1.529   Mean   :  3.541   Mean   : 24.79
##  3rd Qu.: 200.0   3rd Qu.: 2.000   3rd Qu.:  3.000   3rd Qu.: 28.00
##  Max.   :4000.0   Max.   :16.000   Max.   :365.000   Max.   :401.00
##
##  instant_bookable is_business_travel_ready      cancellation_policy
##  f:3132           f:4076                   flexible       :1125
##  t:1686           t: 742                   moderate       :1155
##                                            strict         :2494
##                                            super_strict_30:  42
##                                            super_strict_60:   2
##
##
```

Two-variable summary statistics

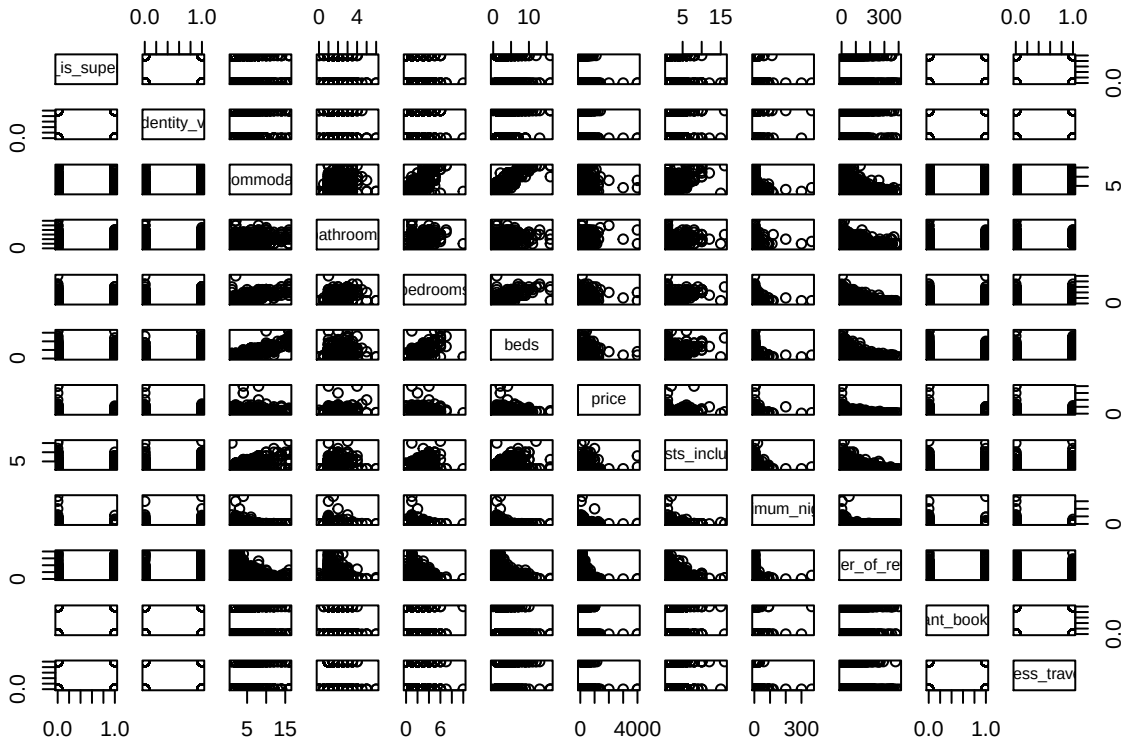
```
two_summary_features <- c("host_is_superhost", "host_identity_verified", "accommodates", "bathrooms", "
boston.two_summary <- boston[ ,two_summary_features, drop=FALSE]
```

```
round(cor(boston.two_summary), digits=2)
```

```
##                          host_is_superhost host_identity_verified
## host_is_superhost                     1.00                    0.17
## host_identity_verified                0.17                    1.00
## accommodates                          0.01                    0.04
## bathrooms                            -0.01                   -0.02
## bedrooms                              0.02                    0.07
## beds                                  0.05                    0.04
## price                                -0.05                   -0.05
## guests_included                       0.11                    0.06
## minimum_nights                       -0.01                    0.00
## number_of_reviews                     0.27                    0.13
## instant_bookable                     -0.05                   -0.10
## is_business_travel_ready              0.20                    0.10
##                          accommodates bathrooms bedrooms  beds price
```

```
## host_is_superhost                0.01     -0.01     0.02  0.05 -0.05
## host_identity_verified           0.04     -0.02     0.07  0.04 -0.05
## accommodates                     1.00      0.42     0.77  0.86  0.42
## bathrooms                        0.42      1.00     0.47  0.39  0.32
## bedrooms                         0.77      0.47     1.00  0.74  0.38
## beds                             0.86      0.39     0.74  1.00  0.35
## price                            0.42      0.32     0.38  0.35  1.00
## guests_included                  0.50      0.14     0.40  0.49  0.21
## minimum_nights                  -0.04      0.02    -0.01 -0.03  0.02
## number_of_reviews                0.00     -0.03    -0.04  0.00 -0.09
## instant_bookable                 0.06     -0.04     0.01  0.04 -0.06
## is_business_travel_ready         0.29      0.06     0.20  0.27  0.13
##                          guests_included minimum_nights number_of_reviews
## host_is_superhost                   0.11          -0.01              0.27
## host_identity_verified              0.06           0.00              0.13
## accommodates                        0.50          -0.04              0.00
## bathrooms                           0.14           0.02             -0.03
## bedrooms                            0.40          -0.01             -0.04
## beds                                0.49          -0.03              0.00
## price                               0.21           0.02             -0.09
## guests_included                     1.00          -0.03              0.09
## minimum_nights                     -0.03           1.00             -0.08
## number_of_reviews                   0.09          -0.08              1.00
## instant_bookable                    0.06          -0.07              0.13
## is_business_travel_ready            0.27          -0.04              0.19
##                          instant_bookable is_business_travel_ready
## host_is_superhost                   -0.05                     0.20
## host_identity_verified              -0.10                     0.10
## accommodates                         0.06                     0.29
## bathrooms                           -0.04                     0.06
## bedrooms                             0.01                     0.20
## beds                                 0.04                     0.27
## price                               -0.06                     0.13
## guests_included                      0.06                     0.27
## minimum_nights                      -0.07                    -0.04
## number_of_reviews                    0.13                     0.19
## instant_bookable                     1.00                     0.12
## is_business_travel_ready             0.12                     1.00
```

```
pairs(boston.two_summary)
```



Findings

Highly correlated predictors

Te

Categorical predictors

The categorical predictors in the dataset are “host_is_superhost”, “host_identity_verified”, “zipcode”, “property_type”, “room_type”, “bed_type”, “instant_bookable”, “is_business_travel_ready”, and “cancellation_policy”.

Missing values

There were 53 rows with missing values. This was a small number, so we removed them.

Outliers

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.