BRANDEIS UNIVERSITY
International Business School

Airbnb in Boston

An Exploratory and Predictive Analytics

Presented by Team SXNB

Xiang Li

Shixuan Wan

Shen Xin

Jiawei Zhang

Nov 9, 2019

Airbnb has seen a dramatic growth since its inception in 2008 with the number of rentals listed on its website growing exponentially each year. Airbnb has successfully disrupted the traditional hospitality industry as more and more travelers, not just the ones who go for tourism but also business travelers who resort to Airbnb as their premier accommodation provider.

Boston, capital of Massachusetts, economic center in east coast, is a livable and beautiful city with rich history, attracting millions of tourists all over the world. Would the combination of sharing economy and tourist city create something unusual? Let go and have a look.

In this report, we are going to dig deep into Boston Airbnb data and provide insightful suggestions to Airbnb host, customers and Airbnb. Things we did are included but not limit to data cleaning, pre-processing, information extraction, data visualization, daily price modeling (machine learning), review analytics (natural language processing).

1.Data Exploration

Data source

Data is from insideairbnb.com, they get them by scrapping listing in Boston area from Airbnb website periodically (every month since 2018). Some fields are created and calculated after scraping the data.

Table 1: data source

File	Description	Time period
listings_details	Detailed Listings data for Boston	2015 – 2019
calendar	Detailed Calendar Data for listings in Boston	2019 – 2020
reviews_details	Detailed Review Data for listings in Boston	2009 – 2019

Descriptive Statistics

Table 2: descriptive statistic for some important variables

	min	mean	median	max	s.d.	n
price	0	196.2	149	10000	324.6	8806
minimum nights	1	5.78	2	1000	20.22	8806
number of reviews	0	26.99	4	650	53.83	8806
reviews per month	0.01	1.886	1.08	13.05	1.996	6311
host listing count	1	30.32	4	309	63.04	8806
availability 365	0	138.33	89	365	128.84	8806

Listings

a) **trend**

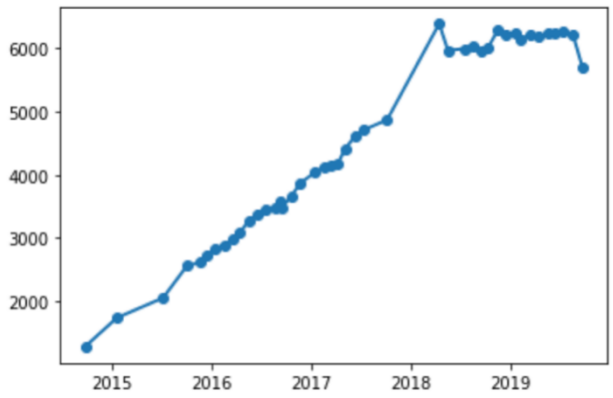


Figure 1: Total listing in Boston of over time

We can see that the number of listings is increasing steadily every year from 2015 – 2018[1]. There is a local maximum at Apr 2018. After Nov 2018, the listing number is at a relatively stable level. Also, we see a temporary fall at Sep 2019.

b) **geographic pattern**

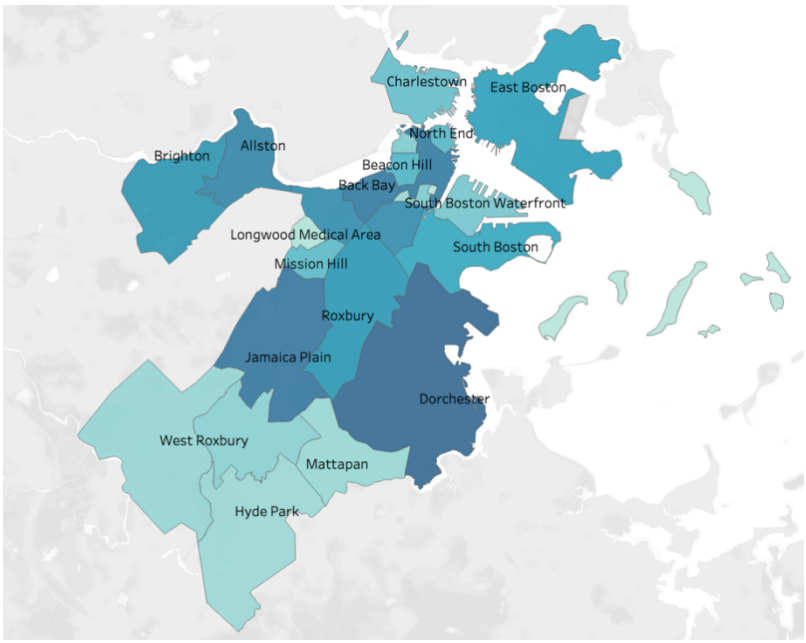


Figure 2: Number of total listing by neighbourhood in 2019

We can tell that large areas generally have high listings, such as Jamaica Plain, Dorchester, East Boston, Roxbury, Allston. Also, listing density of Back Bay and downtown area is large for

It might be more reasonable to standardize the listing by dividing each neighborhood's housing units.

---

[1] The extra data (2015 - 2018) is from insideairbnb.com. http://insideairbnb.com/get-the-data.html

Price

a) **distribution**

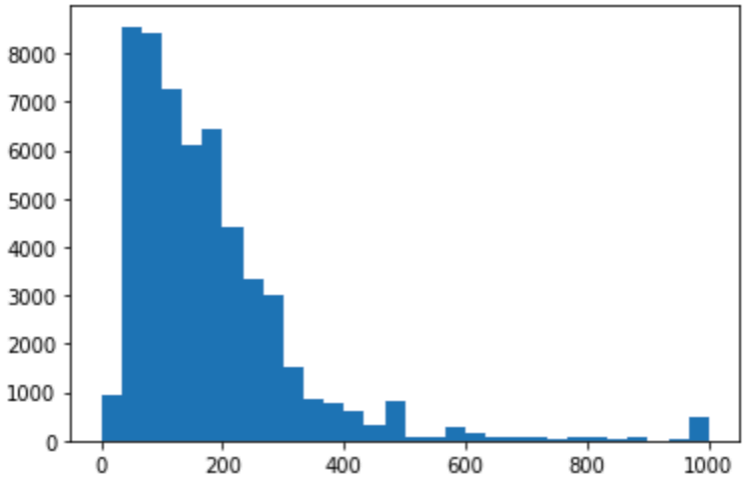
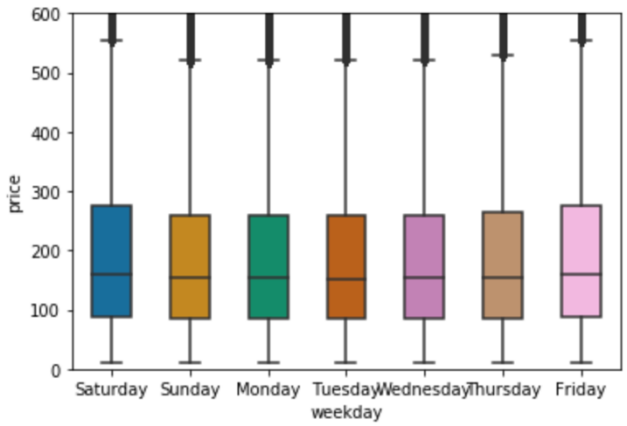


Figure 3: 2019 price histogram

The price in year 2019 is right skewed where 75% of data is below 228, and we can see some outliers exceeding 500.

b) **Price by weekday**



Weekday	Avg Price
Monday	219.29
Tuesday	219.10
Wednesday	219.77
Thursday	221.54
Friday	230.24
Saturday	231.26
Sunday	220.28

Figure 4: Weekday price boxplot (future price data)

We can see that Friday and Saturday have higher price than other days. The reason might be for short-terms travel/stays, people tend to start at Friday and end at Sunday (for preparation of word/study in Monday).

c) Price by date

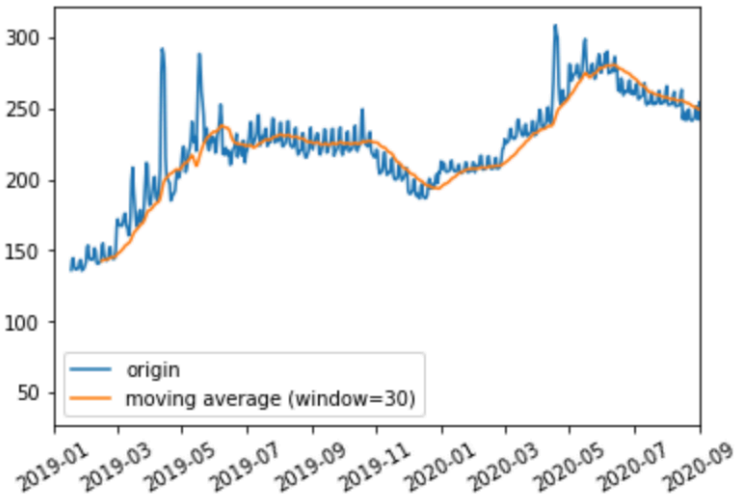


Figure 5: Price time series plot (future price data)

Date during year 2019 would more reliable since the data is scraped at Feb 2019, data in year 2020 would be too far to reference. We can see that Jun to Nov is more expensive than other months. And the price is stable with small variation. The price drops as date departs from these period (Jun to Nov).

d) price by geographic

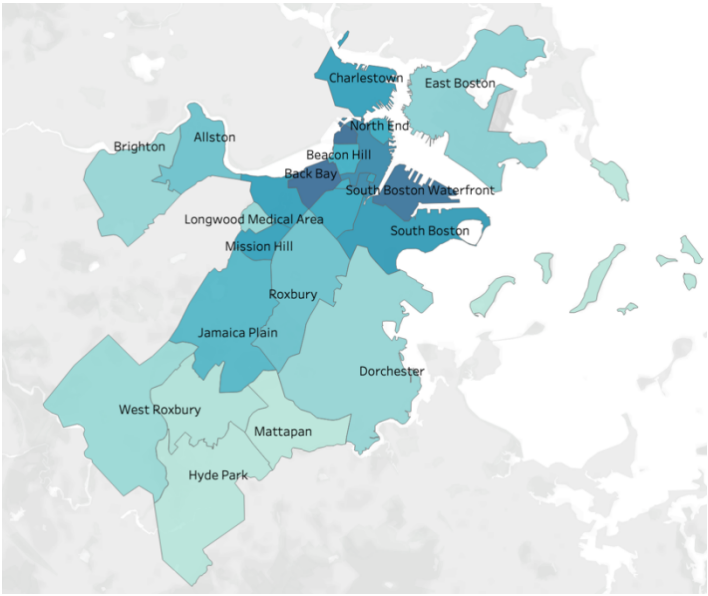


Figure 6: Average listing price by neighbourhood in 2019

We can see that inner harbor area has higher price (Back Bay, West End, Downtown, Chinatown, South Boston). It's central business area and has a lot of places of interest. It's popular with tourists and businesspeople. Generally, the far the neighbourhood is from downtown area, the cheaper the daily price would be.

Occupancy rate

a) **time series (by percentile)**

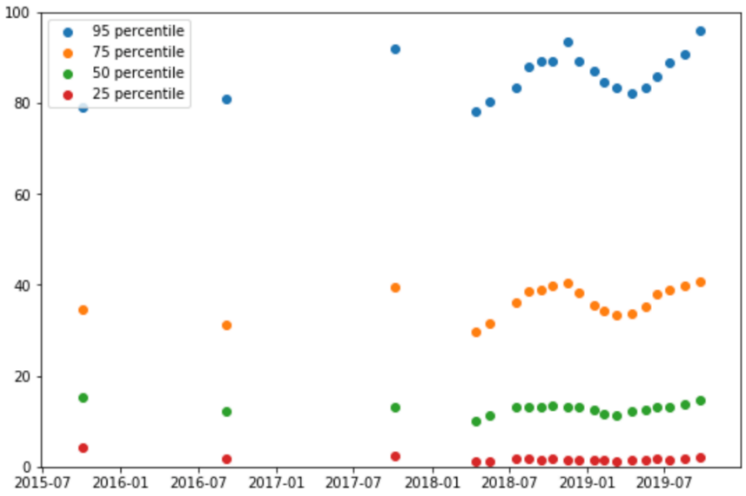


Figure 7: Occupancy rate time series plot with different percentiles

We estimate the occupancy rate based on data given. Details of the model are in the appendix. The single point for year 2015, 2016 and 2017 is because we only have annual data in these years. The occupancy rate has seasonality. Sep to Dec would be season with high demand. It also worth noticing that though some listing might have occupancy rate almost 100%, most of listings'(75%) stats are under 40%.

b) **vertical comparation**

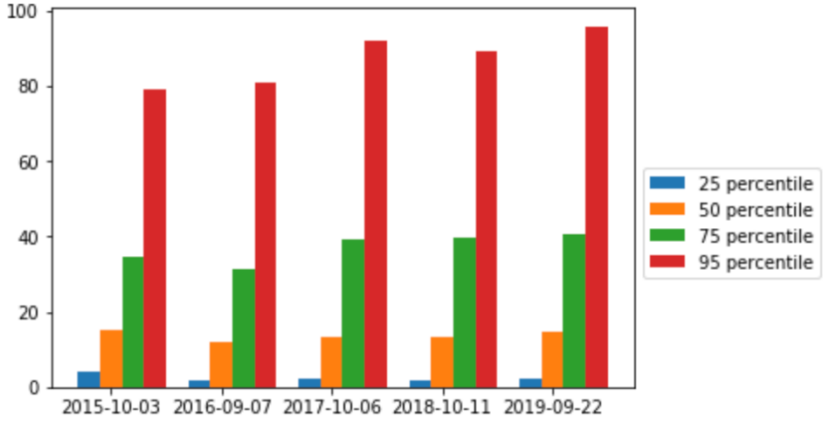


Figure 8: Occupancy rate time series plot with different percentiles

We plot the occupancy rate in similar months for the 5 years (year 2016 and 2019 only have Sep data). We can see there is a small but positive growth of occupancy rate from year 2015 to 2019.

c) by geographic

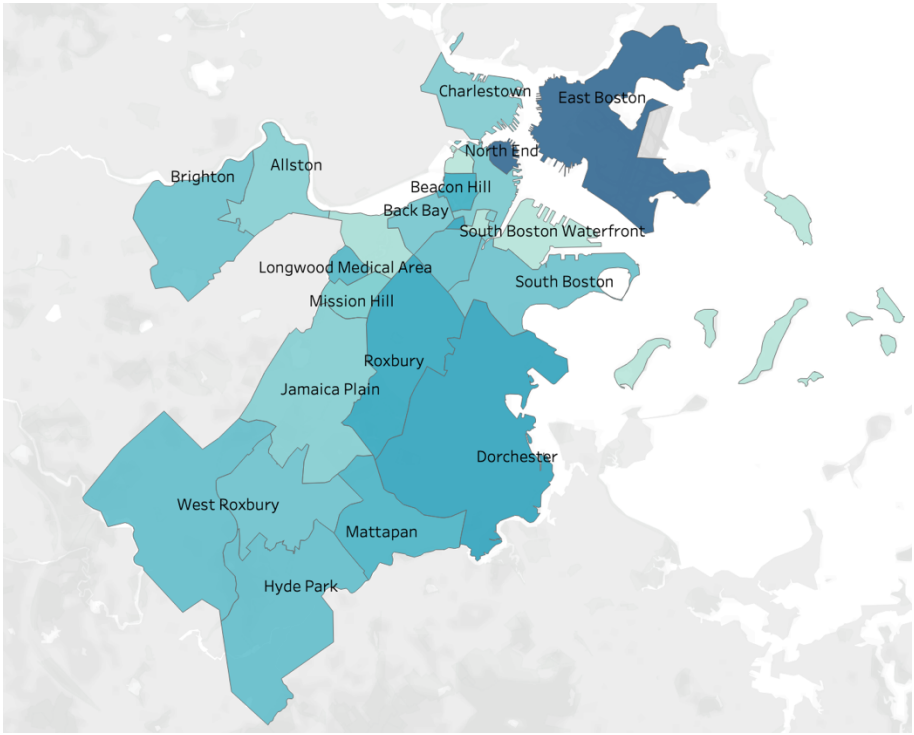


Figure 9: Median occupancy rate by neighbourhood in year 2019

We can see that North End and East Boston has highest occupancy rate. The reason might be they have relatively low price while not far from the downtown area. Fenway, South Boston Waterfront, West End, Mission hill have low occupancy rate.

Reviews

a) Number of reviews

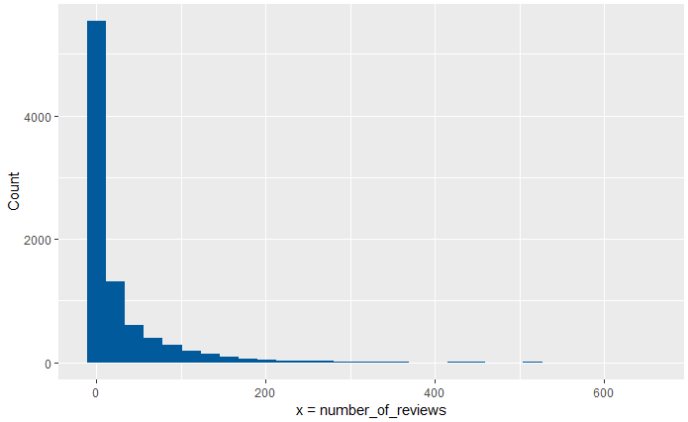


Figure 10: reviews distribution

The distribution of number of reviews shows that customers are quite unlikely to leave reviews for hosts, which may lead to the difficulties for Airbnb, hosts and other customers to evaluate the popularity and quality of the services.

b) Reviews trend through time

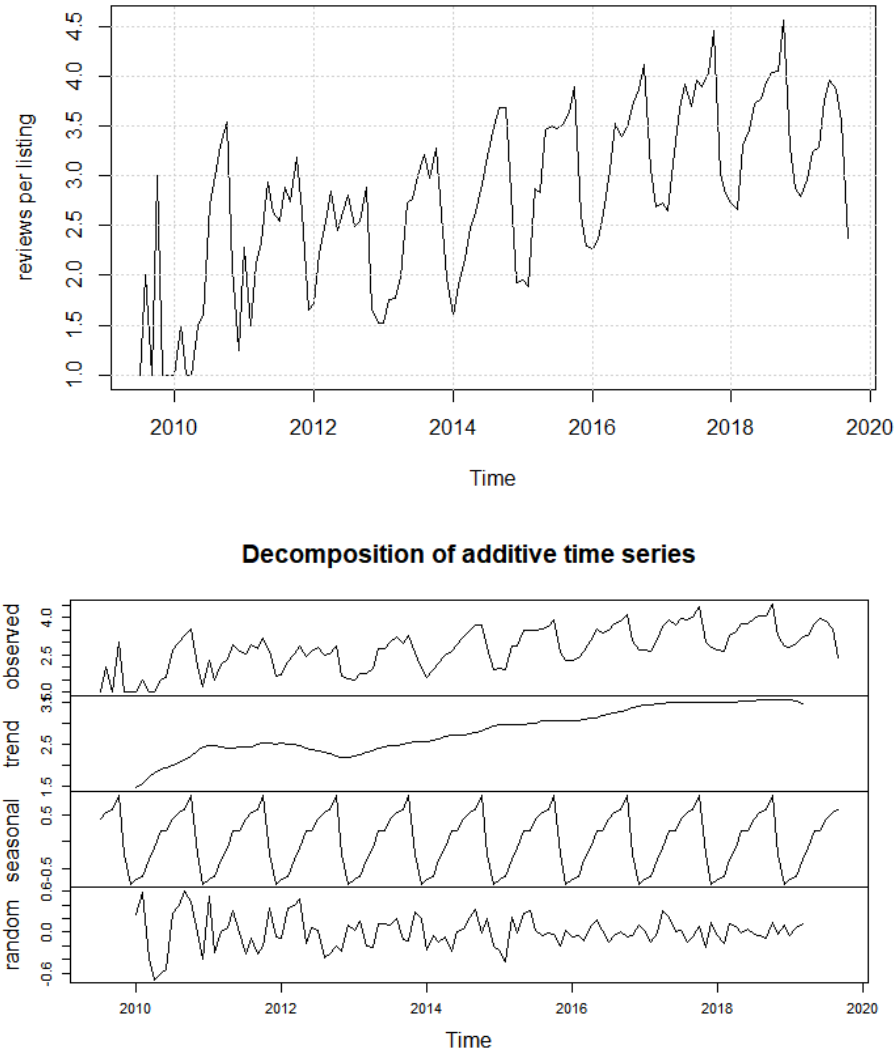


Figure 11: reviews per listing per month time series plot and decomposition

We use reviews per listing (file review_details) from 2009 to 2019 to represent popularity trend of Airbnb in Boston area. We use exponential smoothing state space model to decompose the review time series into 3 parts – trend, seasonality and error part.

We can find that the reviews per month continue to grow (though the growth rate gradually declines to flat) in past 10 years with a seasonal pattern. It shows that the reviews rate reach peaks around July and August and reach bottom around December. At this rate, the popularity of Airbnb will soon stop growing in a few years, implying it may be approaching its market limit. Thus, in order to continue its growth, Airbnb must focus on seeking new growth points.

Others

accommodates

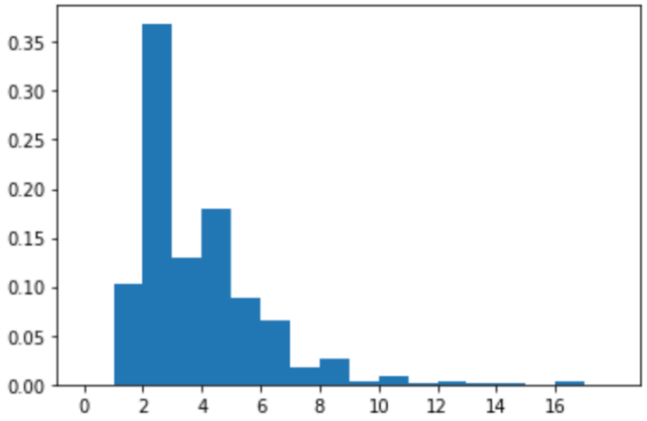


Figure 12: accommodates distribution

The accommodates data shows that most listings tend to host 2-5 customers. But it is worth noticing that there is a significant share of listings which can host more than 5 people, even up to 10 to 30 customers.

Listings count for each host

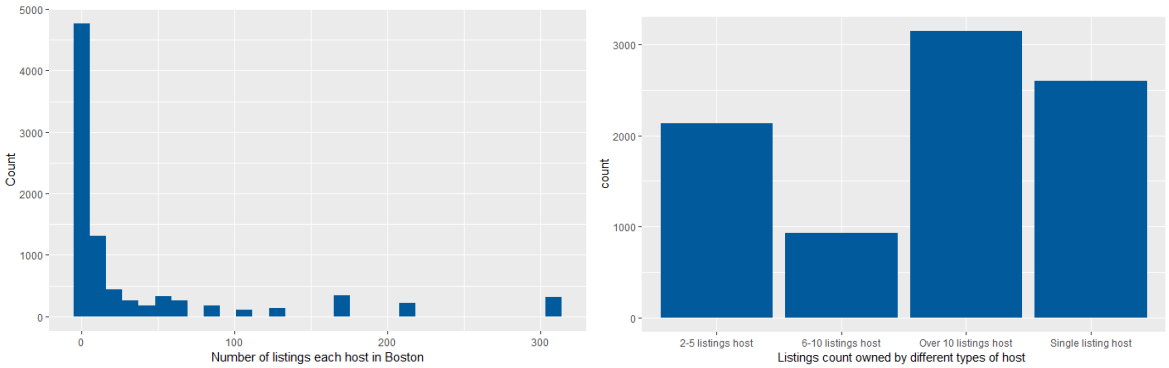


Figure 13: listing count for each host

From the distribution of number of listings owned by each host in Boston area, we can find that though most host only have a listing on Airbnb, there is a significant share of host who owns more than 50 listings, some even have more than 300 listings.

Compared with next graph(which shows that over 2/3 of listings are owned by hosts with multiple listings and many of them are owned by host with more than 10 listings), it shows that the Airbnb market in Boston are mostly controlled by professional landlords who may be using Airbnb as a housing rental platform.

Room type

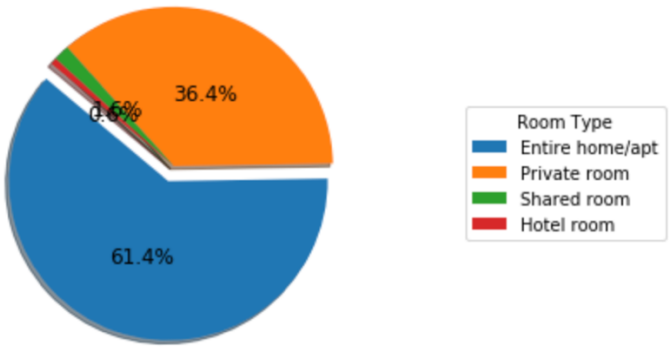


Figure 14: Room type pie chart

Entire rooms (61.4%) and private rooms(36.4%) are most common which account for more than 97% of listings. The distribution of room types shows around 60% of the total listings are rented as entire apartments, which means the hosts may not actually live in the apartment regularly.

2. Data Mining

2.1 Linear Regression

a) origin model

Price~host_response_rate+host_listings_count+accommodates+bathrooms+bedrooms+beds+security_deposit+guests_included+extra_people+minimum_nights+maximum_nights+minimum_minimum_nights+maximum_minimum_nights+minimum_maximum_nights+maximum_maximum_nights+minimum_nights_avg_ntm+maximum_nights_avg_ntm+availability_30+availability_60+availability_90+availability_365+number_of_reviews+number_of_reviews_ltm+review_scores_rating+review_scores_accuracy+review_scores_cleanliness+review_scores_checkin+review_scores_communication+review_scores_location+review_scores_value+calculated_host_listings_count+calculated_host_listings_count_entire_homes+calculated_host_listings_count_private_rooms+calculated_host_listings_count_shared_rooms+reviews_per_month+host_is_superhost+host_has_profile_pic+host_identity_verified+is_location_exact+requires_license+instant_bookable+is_business_travel_ready+require_guest_profile_picture+require_guest_phone_verification+host_response_time+neighbourhood+property_type+room_type+bed_type+cancellation_policy

```
Call:
lm(formula = price ~ ., data = listing_details)

Residuals:
    Min      1Q  Median      3Q     Max
-607.25  -43.96   -3.30   31.67  672.66

Coefficients: (1 not defined because of singularities)
                                              Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                                  15.781253   18.938106   0.833             0.404679
host_response_rate                          -58.724673    9.377192  -6.263   0.0000000003843686 ***
host_listings_count                           0.365590    0.004247  86.074 < 0.0000000000000002 ***
accommodates                                  9.268209    0.591659  15.665 < 0.0000000000000002 ***
bathrooms                                    43.094380    1.242573  34.682 < 0.0000000000000002 ***
bedrooms                                     28.979564    1.007775  28.756 < 0.0000000000000002 ***
beds                                          1.275352    0.848322   1.503             0.132752
security_deposit                              0.020299    0.001876  10.819 < 0.0000000000000002 ***
guests_included                               3.158549    0.454776   6.945   0.0000000000038595 ***
extra_people                                 -0.137697    0.026474  -5.201   0.0000001993696556 ***
minimum_nights                               -0.913638    0.067348 -13.566 < 0.0000000000000002 ***
maximum_nights                               -0.004435    0.008513  -0.521             0.602405
minimum_minimum_nights                        0.219291    0.069060   3.175             0.001498 **
maximum_minimum_nights                       -0.045669    0.070337  -0.649             0.516156
minimum_maximum_nights                       -0.136728    0.012025 -11.370 < 0.0000000000000002 ***
maximum_maximum_nights                        0.193610    0.046375   4.175   0.0000299074960750 ***
minimum_nights_avg_ntm                        0.082133    0.120167   0.683             0.494300
maximum_nights_avg_ntm                       -0.054665    0.052422  -1.043             0.297049
availability_30                               0.123206    0.150112   0.821             0.411789
availability_60                               0.044531    0.147134   0.303             0.762156
availability_90                              -0.269223    0.073849  -3.646             0.000267 ***
availability_365                              0.076422    0.005661  13.501 < 0.0000000000000002 ***
number_of_reviews                            -0.027179    0.013292  -2.045             0.040888 *
number_of_reviews_ltm                        -0.126922    0.049398  -2.569             0.010193 *
review_scores_rating                          1.112234    0.132551   8.391 < 0.0000000000000002 ***
review_scores_accuracy                       -0.781929    1.066527  -0.733             0.463471
review_scores_cleanliness                     6.897006    0.902276   7.644   0.0000000000000217 ***
review_scores_checkin                        -3.955985    0.942750  -4.196   0.0000272246256831 ***
review_scores_communication                 -10.070473    1.011904  -9.952 < 0.0000000000000002 ***
review_scores_location                        7.542406    0.901756   8.364 < 0.0000000000000002 ***
review_scores_value                          -8.461273    0.949986  -8.907 < 0.0000000000000002 ***
calculated_host_listings_count               -4.780711    1.711183  -2.794             0.005213 **
calculated_host_listings_count_entire_homes   3.421935    1.711290   2.000             0.045551 *
calculated_host_listings_count_private_rooms  4.020819    1.715462   2.344             0.019092 *
calculated_host_listings_count_shared_rooms   2.001915    1.773020   1.129             0.258866
reviews_per_month                            -4.998171    0.492421 -10.150 < 0.0000000000000002 ***
host_is_superhost                             9.489905    1.344699   7.057   0.0000000000017377 ***
host_has_profile_pic                         39.137750   13.262436   2.951             0.003170 **
host_identity_verified                       -0.307211    1.190057  -0.258             0.796295
is_location_exact                            -2.333361    1.522449  -1.533             0.125377
requires_license                              1.968178    1.677520   1.173             0.240699
instant_bookable                              3.018420    1.244704   2.425             0.015314 *
is_business_travel_ready                            NA          NA      NA                   NA
require_guest_profile_picture               -32.889535    6.799387  -4.837   0.0000013242736104 ***
require_guest_phone_verification             55.943954    5.015337  11.155 < 0.0000000000000002 ***
host_response_timewithin a day               35.008076    9.399359   3.725             0.000196 ***
```

```
host_response_timewithin a few hours          19.340661   10.092717   1.916             0.055337 .
host_response_timewithin an hour              28.818740   10.183544   2.830             0.004659 **
neighbourhoodBack Bay                        101.745389    2.601413  39.112 < 0.0000000000000002 ***
neighbourhoodBeacon Hill                      69.855357    3.025726  23.087 < 0.0000000000000002 ***
neighbourhoodBrookline                         6.805731   27.241084   0.250             0.802718
neighbourhoodCambridge                         9.810958   28.708583   0.342             0.732547
neighbourhoodCharlestown                      41.865352    3.796033  11.029 < 0.0000000000000002 ***
neighbourhoodChelsea                         -30.343544   49.642709  -0.611             0.541047
neighbourhoodChestnut Hill                   -65.431301   49.643939  -1.318             0.187510
neighbourhoodChinatown                        68.873891    4.292588  16.045 < 0.0000000000000002 ***
neighbourhoodDorchester                      -17.678557    2.399775  -7.367   0.0000000000001797 ***
neighbourhoodDowntown                         88.202184    3.021522  29.191 < 0.0000000000000002 ***
neighbourhoodDowntown Crossing                87.248283    5.301597  16.457 < 0.0000000000000002 ***
neighbourhoodEast Boston                      10.546812    2.730140   3.863             0.000112 ***
neighbourhoodEverett                         219.243175   85.923024   2.552             0.010728 *
neighbourhoodFenway/Kenmore                   67.531061    2.963593  22.787 < 0.0000000000000002 ***
neighbourhoodFinancial District               49.983023    9.183093   5.443   0.0000000528491983 ***
neighbourhoodGovernment Center                81.241816   12.654963   6.420   0.0000000001386809 ***
neighbourhoodHarvard Square                  -62.844448   85.955063  -0.731             0.464705
neighbourhoodHyde Park                       -24.444853    5.313725  -4.600   0.0000042368813014 ***
neighbourhoodJamaica Plain                    -4.165620    2.714985  -1.534             0.124966
neighbourhoodLeather District                217.426460   16.762990  12.971 < 0.0000000000000002 ***
neighbourhoodMattapan                        -27.684655    5.402712  -5.124   0.0000003007686797 ***
neighbourhoodMission Hill                     10.729580    4.005191   2.679             0.007390 **
neighbourhoodNorth End                        35.295942    3.078148  11.467 < 0.0000000000000002 ***
neighbourhoodRevere                            8.120756   85.923575   0.095             0.924704
neighbourhoodRoslindale                      -37.625143    4.216253  -8.924 < 0.0000000000000002 ***
neighbourhoodRoxbury                          14.787712    2.770810   5.337   0.0000000952528671 ***
neighbourhoodSomerville                       -6.018360   22.286532  -0.270             0.787128
neighbourhoodSouth Boston                     55.777512    2.657249  20.991 < 0.0000000000000002 ***
neighbourhoodSouth End                        69.559122    2.655655  26.193 < 0.0000000000000002 ***
neighbourhoodTheater District                 57.721741    5.551143  10.398 < 0.0000000000000002 ***
neighbourhoodWest End                         64.021739    4.428800  14.456 < 0.0000000000000002 ***
neighbourhoodWest Roxbury                    -26.329027    5.050268  -5.213   0.0000001867375133 ***
property_typeBarn                             96.020493   35.181077   2.729             0.006350 **
property_typeBed and breakfast                56.248003    5.306297  10.600 < 0.0000000000000002 ***
property_typeBoat                             46.534329   13.084248   3.557             0.000376 ***
property_typeBoutique hotel                   25.687208   17.477279   1.470             0.141641
property_typeBungalow                         25.568320   49.885799   0.513             0.608279
property_typeChalet                          -80.413786   49.715710  -1.617             0.105788
property_typeCondominium                       0.071451    1.997143   0.036             0.971461
property_typeCottage                         -13.704851   35.139299  -0.390             0.696528
property_typeGuest suite                      -1.404104    3.888663  -0.361             0.718045
property_typeGuesthouse                      -36.105762   15.080446  -2.394             0.016663 *
property_typeHotel                            76.277725   18.579626   4.105   0.0000404675884075 ***
property_typeHouse                            -7.765619    1.781264  -4.360   0.0000130759378622 ***
property_typeHouseboat                        15.382365   38.598426   0.399             0.690248
property_typeLoft                              2.090393    5.308536   0.394             0.693747
property_typeOther                            -4.922437   10.402961  -0.473             0.636091
property_typeResort                           30.546291   50.016891   0.611             0.541390
property_typeServiced apartment               39.672443    4.119170   9.631 < 0.0000000000000002 ***
property_typeTownhouse                        15.492728    3.329452   4.653   0.0000032826360039 ***
room_typeHotel room                          -23.100237   24.149565  -0.957             0.338803
room_typePrivate room                        -57.181388    1.852244 -30.871 < 0.0000000000000002 ***
room_typeShared room                         -79.068101    7.501045 -10.541 < 0.0000000000000002 ***

bed_typeCouch                                 28.317195   29.790591   0.951             0.341845
bed_typeFuton                                 24.549987   13.028790   1.884             0.059536 .
bed_typePull-out Sofa                        -15.182911   13.967703  -1.087             0.277046
bed_typeReal Bed                              -0.427640    7.705582  -0.055             0.955743
cancellation_policymoderate                  -11.348839    1.910117  -5.941   0.0000000028588065 ***
cancellation_policystrict                   -167.529806    7.625737 -21.969 < 0.0000000000000002 ***
cancellation_policystrict_14_with_grace_period  -5.614605   1.838758  -3.053             0.002264 **
cancellation_policysuper_strict_30           -39.856675    5.600265  -7.117   0.0000000000011302 ***
cancellation_policysuper_strict_60           -69.994569   12.360862  -5.663   0.0000000150563052 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85.86 on 27918 degrees of freedom
Multiple R-squared:  0.6478,    Adjusted R-squared:  0.6464
F-statistic: 475.5 on 108 and 27918 DF,  p-value: < 0.00000000000000022
```

b) Backward Selection, Forward Selection and Forward and Backward Selection

To avoid too much information, we will not to show the output. Detailed output is in attached files

c) Check Collinearity

	GVIF	Df	GVIF^(1/(2*Df))
host_response_rate	2.964111	1	1.721659
host_listings_count	1.942696	1	1.393806
accommodates	7.155057	1	2.674894
bathrooms	1.574542	1	1.254808
bedrooms	3.582810	1	1.892831
beds	5.349676	1	2.312937
security_deposit	1.199015	1	1.094995
guests_included	1.859174	1	1.363515
extra_people	1.243592	1	1.115165
minimum_nights	1.811629	1	1.345968
minimum_minimum_nights	1.686367	1	1.298602
availability_90	1.605606	1	1.267125
availability_365	1.703204	1	1.305069
number_of_reviews	2.654595	1	1.629293
number_of_reviews_ltm	5.406316	1	2.325149
review_scores_rating	3.623038	1	1.903428
review_scores_cleanliness	1.929102	1	1.388921
review_scores_checkin	1.788330	1	1.337285
review_scores_communication	2.265745	1	1.505239
review_scores_location	1.666662	1	1.290993
review_scores_value	2.586813	1	1.608357
reviews_per_month	4.081794	1	2.020345
host_is_superhost	1.460880	1	1.208669
host_has_profile_pic	1.022194	1	1.011036
instant_bookable	1.412838	1	1.188629
require_guest_profile_picture	5.031595	1	2.243122
require_guest_phone_verification	3.960995	1	1.990225
host_response_time	3.941404	3	1.256826
neighbourhood	7.096755	32	1.031093
property_type	3.541579	18	1.035751
room_type	2.911227	3	1.194940
bed_type	1.147837	4	1.017384

d) Final Model

Price~host_response_rate+host_listings_count+accommodates+bathrooms+bedrooms+beds+security_deposit+guests_included+extra_people+minimum_nights+minimum_minimum_nights+maximum_nights_avg_ntm+availability_90+availability_365+number_of_reviews+number_of_reviews_ltm+review_scores_rating+review_scores_cleanliness+review_scores_checkin+review_scores_communication+review_scores_location+review_scores_value+reviews_per_month+host_is_superhost+host_has_profile_pic+instant_bookable+require_guest_profile_picture+require_guest_phone_verification+host_response_time+neighbourhood+property_type+room_type+bed_type

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.90450 -0.22084  0.00036  0.21474  2.58651

Coefficients:
                                            Estimate   Std. Error t value             Pr(>|t|)
(Intercept)                              3.975868507  0.080964582  49.106 < 0.0000000000000002 ***
host_response_rate                      -0.234753975  0.040117507  -5.852  0.00000000492124236 ***
host_listings_count                      0.000359712  0.000008491  42.361 < 0.0000000000000002 ***
accommodates                             0.066442147  0.002525978  26.304 < 0.0000000000000002 ***
bathrooms                                0.062168861  0.005304663  11.720 < 0.0000000000000002 ***
bedrooms                                 0.109274203  0.004292536  25.457 < 0.0000000000000002 ***
beds                                    -0.011295820  0.003631215  -3.111             0.001868 **
security_deposit                         0.000051300  0.000007905   6.489  0.00000000008759677 ***
guests_included                          0.012804817  0.001902056   6.732  0.00000000001704796 ***
extra_people                             0.000543531  0.000112501   4.831  0.00000136345797560 ***
minimum_nights                          -0.004243296  0.000285786 -14.848 < 0.0000000000000002 ***
minimum_minimum_nights                   0.000345591  0.000201566   1.715             0.086443 .
availability_90                         -0.000212137  0.000096401  -2.201             0.027775 *
availability_365                         0.000143470  0.000023503   6.104  0.00000000104505219 ***
number_of_reviews                        0.000242568  0.000055765   4.350  0.00001367394593431 ***
number_of_reviews_ltm                   -0.001067426  0.000210739  -5.065  0.00000041065199782 ***
review_scores_rating                     0.004136364  0.000526887   7.851  0.00000000000000429 ***
review_scores_cleanliness                0.037169546  0.003812377   9.750 < 0.0000000000000002 ***
review_scores_checkin                   -0.026871370  0.004018390  -6.687  0.00000000002319308 ***
review_scores_communication             -0.040722574  0.004316839  -9.433 < 0.0000000000000002 ***
review_scores_location                   0.053840622  0.003855081  13.966 < 0.0000000000000002 ***
review_scores_value                     -0.033180765  0.004036594  -8.220 < 0.0000000000000002 ***
reviews_per_month                       -0.025530105  0.002091485 -12.207 < 0.0000000000000002 ***
host_is_superhost                        0.090867458  0.005719566  15.887 < 0.0000000000000002 ***
host_has_profile_pic                     0.145440238  0.056895925   2.556             0.010586 *
instant_bookable                         0.014643187  0.005292350   2.767             0.005664 **
require_guest_profile_picture           -0.113994428  0.029026378  -3.927  0.00008612188785108 ***
require_guest_phone_verification         0.208197237  0.021195771   9.823 < 0.0000000000000002 ***
host_response_timewithin a day           0.161016425  0.040245080   4.001  0.00006326544195390 ***
host_response_timewithin a few hours     0.126956429  0.043158826   2.942             0.003268 **
host_response_timewithin an hour         0.096654817  0.043568671   2.218             0.026532 *
neighbourhoodBack Bay                    0.461417409  0.011020480  41.869 < 0.0000000000000002 ***
neighbourhoodBeacon Hill                 0.305176414  0.012890792  23.674 < 0.0000000000000002 ***
neighbourhoodBrookline                   0.156446094  0.116964835   1.338             0.181055
neighbourhoodCambridge                   0.016549053  0.123269291   0.134             0.893205
neighbourhoodCharlestown                 0.223695499  0.016265263  13.753 < 0.0000000000000002 ***
neighbourhoodChelsea                    -0.127222439  0.213188253  -0.597             0.550672
neighbourhoodChestnut Hill              -0.288952864  0.213182359  -1.355             0.175293
neighbourhoodChinatown                   0.342835739  0.018327370  18.706 < 0.0000000000000002 ***
neighbourhoodDorchester                 -0.119507478  0.010194675 -11.723 < 0.0000000000000002 ***
neighbourhoodDowntown                    0.404776634  0.012841253  31.522 < 0.0000000000000002 ***
neighbourhoodDowntown Crossing           0.438058811  0.022482081  19.485 < 0.0000000000000002 ***
neighbourhoodEast Boston                 0.052432374  0.011689420   4.485  0.00000730481642709 ***
```

```
neighbourhoodEverett                              0.826277868  0.368992232   2.239              0.025145 *
neighbourhoodFenway/Kenmore                       0.315649536  0.012654195  24.944 < 0.0000000000000002 ***
neighbourhoodFinancial District                   0.277303773  0.039389084   7.040  0.00000000000196518 ***
neighbourhoodGovernment Center                    0.531316684  0.054123484   9.817 < 0.0000000000000002 ***
neighbourhoodHarvard Square                      -0.481853755  0.369146028  -1.305              0.191795
neighbourhoodHyde Park                           -0.231317659  0.022786367 -10.152 < 0.0000000000000002 ***
neighbourhoodJamaica Plain                        0.022202632  0.011562768   1.920              0.054845 .
neighbourhoodLeather District                     0.779021793  0.071937251  10.829 < 0.0000000000000002 ***
neighbourhoodMattapan                            -0.283405725  0.023166739 -12.233 < 0.0000000000000002 ***
neighbourhoodMission Hill                         0.118516395  0.016700724   7.096  0.00000000000131025 ***
neighbourhoodNorth End                            0.171362625  0.013114627  13.067 < 0.0000000000000002 ***
neighbourhoodRevere                               0.145978375  0.368911357   0.396              0.692329
neighbourhoodRoslindale                          -0.220847911  0.018076728 -12.217 < 0.0000000000000002 ***
neighbourhoodRoxbury                             -0.038528954  0.011594405  -3.323              0.000891 ***
neighbourhoodSomerville                          -0.179223727  0.095690750  -1.873              0.061086 .
neighbourhoodSouth Boston                         0.305801516  0.011304236  27.052 < 0.0000000000000002 ***
neighbourhoodSouth End                            0.334858268  0.011329742  29.556 < 0.0000000000000002 ***
neighbourhoodTheater District                     0.292142770  0.023779634  12.285 < 0.0000000000000002 ***
neighbourhoodWest End                             0.335951964  0.018614128  18.048 < 0.0000000000000002 ***
neighbourhoodWest Roxbury                        -0.150126690  0.021655715  -6.932  0.00000000000422672 ***
property_typeBarn                                 0.604453022  0.151049865   4.002  0.00006305691634719 ***
property_typeBed and breakfast                    0.526604149  0.022729001  23.169 < 0.0000000000000002 ***
property_typeBoat                                 0.308442377  0.056117113   5.496  0.00000003909991440 ***
property_typeBoutique hotel                       0.414882526  0.074888303   5.540  0.00000003051665854 ***
property_typeBungalow                             0.356578733  0.214155959   1.665              0.095916 .
property_typeChalet                              -0.333050086  0.213462774  -1.560              0.118718
property_typeCondominium                          0.075940116  0.008470755   8.965 < 0.0000000000000002 ***
property_typeCottage                              0.230474806  0.150895741   1.527              0.126678
property_typeGuest suite                         -0.004981872  0.016662014  -0.299              0.764946
property_typeGuesthouse                          -0.276319063  0.064733603  -4.269  0.00001973904033120 ***
property_typeHotel                                0.791823817  0.079675880   9.938 < 0.0000000000000002 ***
property_typeHouse                               -0.033198489  0.007534826  -4.406  0.00001056859467920 ***
property_typeHouseboat                            0.228060649  0.165692923   1.376              0.168707
property_typeLoft                                 0.050365788  0.022753206   2.214              0.026867 *
property_typeOther                               -0.039728961  0.044442120  -0.894              0.371357
property_typeResort                               0.585977945  0.214735651   2.729              0.006360 **
property_typeServiced apartment                   0.133982092  0.017551184   7.634  0.00000000000002352 ***
property_typeTownhouse                            0.051111826  0.013875379   3.684              0.000230 ***
room_typeHotel room                              -0.295683089  0.075711890  -3.905  0.00009430189425161 ***
room_typePrivate room                            -0.532485079  0.007400032 -71.957 < 0.0000000000000002 ***
room_typeShared room                             -1.127953647  0.027979347 -40.314 < 0.0000000000000002 ***
bed_typeCouch                                     0.307666165  0.127813883   2.407              0.016084 *
bed_typeFuton                                     0.306839222  0.055721355   5.507  0.00000003688996899 ***
bed_typePull-out Sofa                             0.155911344  0.059707244   2.611              0.009026 **
bed_typeReal Bed                                  0.162163608  0.032859410   4.935  0.00000080581593042 ***
cancellation_policymoderate                       0.002527303  0.008103249   0.312              0.755128
cancellation_policystrict                        -0.237447834  0.029591926  -8.024  0.00000000000000106 ***
cancellation_policystrict_14_with_grace_period    0.012538441  0.007738701   1.620              0.105195
cancellation_policysuper_strict_30               -0.099550322  0.023524631  -4.232  0.00002326213108271 ***
cancellation_policysuper_strict_60               -0.135509509  0.052973573  -2.558              0.010531 *
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3687 on 27918 degrees of freedom
Multiple R-squared:  0.7337,    Adjusted R-squared:  0.7328
F-statistic: 835.9 on 92 and 27918 DF,  p-value: < 0.00000000000000022
```

In the initial model, we used 50 independent variables, and found some of them are not very significant, so we removed them out. Also, the reason we took the logarithm of price is that we realized the scale of price is a not very similar - a bit of them are higher than $500 while a bit of them are less than $100, so it is better to smooth the scale. The final model includes 38 independent variables, the adjusted R-square is 0.743, which is pretty good, although it does not necessarily mean our linear regression model is a really good model. Since in this part, we focus on the X variables, we care a lot about the assumptions of regression to make sure the coefficients explain causality, but not correlation. More of the assumptions will be addressed next part. However, in the machine learning part, which we done in python, we focus only on the Y variable, the target, the price, so we will not care too much about whether these assumptions are violated, but just tries to minimize the error. We think it is too redundant to explain the meaning of each independent variable. But we combine the result of the feature importances of our two best machine learning methods, and choose the intersects in top 20 important variables. Detailed interpretation will be addressed later. Here are some brief interpretations:

More beds, more bedrooms, more accommodates or more bathrooms will push price up.

Price depends on the location, the style and the furniture of the listing. Certain neighborhood, certain bed type and certain room type will push price up while certain will push price down.

e) Plots Analysis

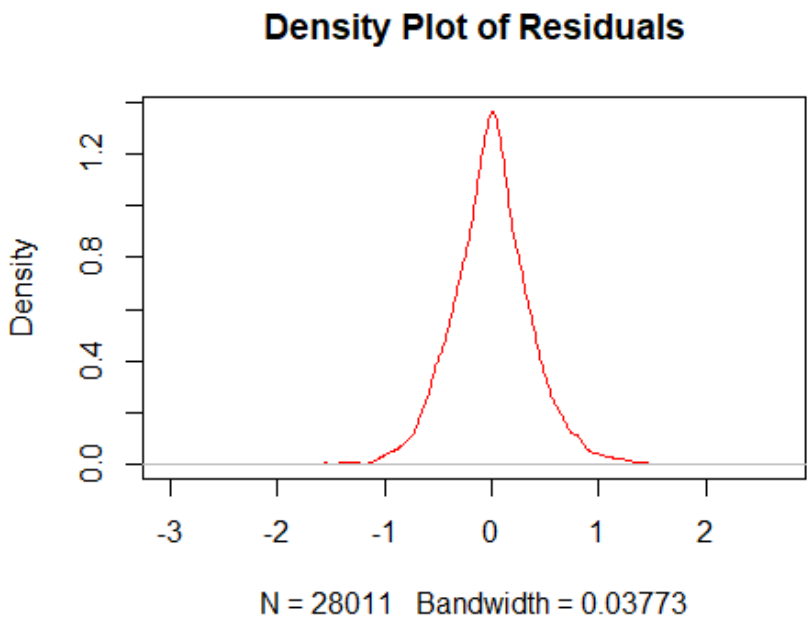


Figure 15: Density Plot of Residuals

This plot shows the residuals of our log-linear regression model. We can see that, the residuals are basically normally distributed, which means our regression is done pretty.

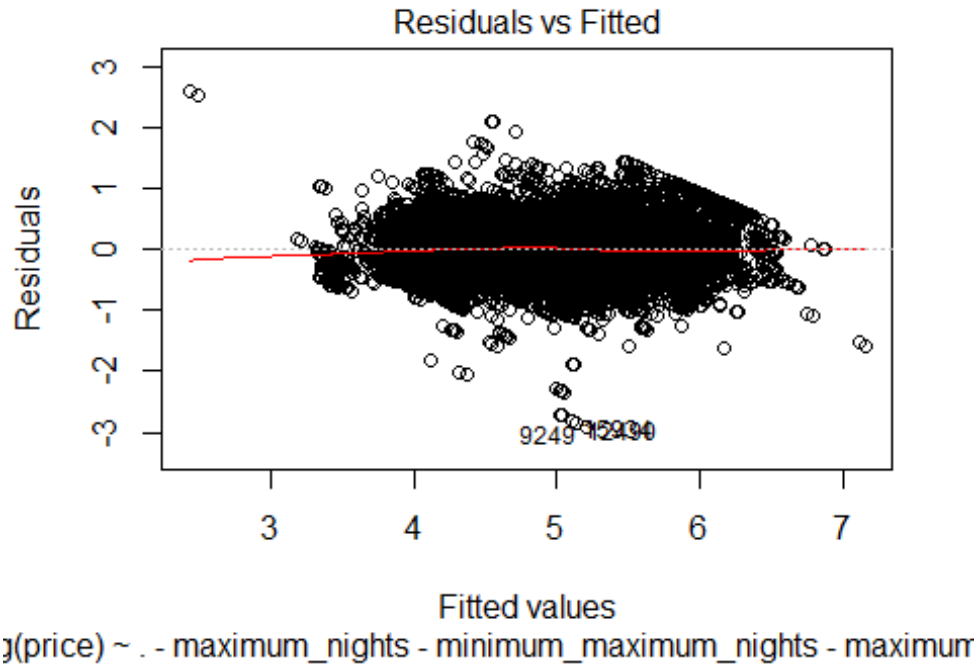


Figure 16: Residuals vs Fitted

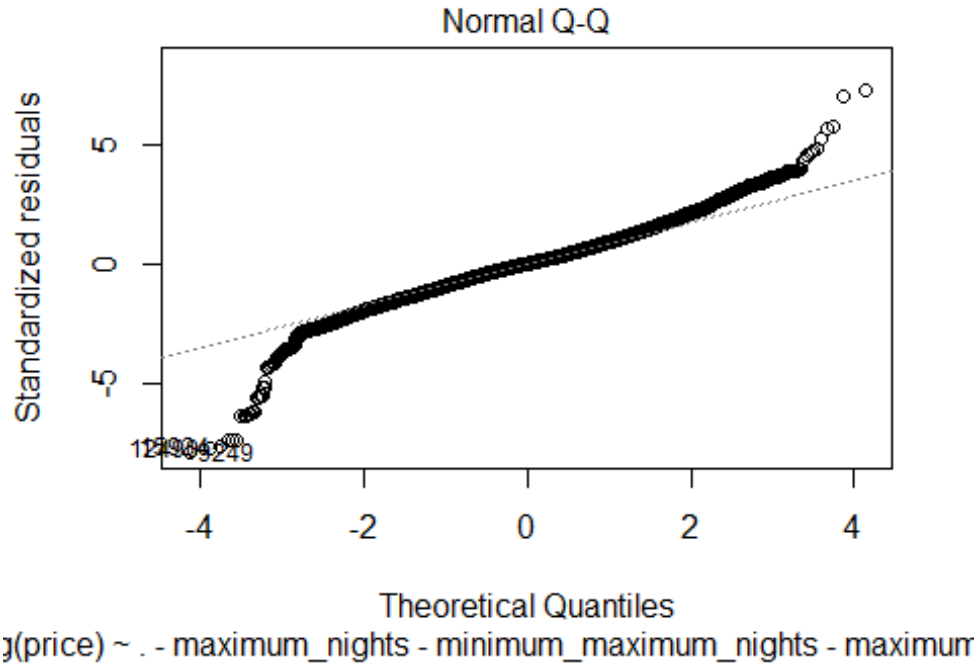


Figure 17: Normal Q-Q Plot

In the residual vs fitted value plot, we can see that most of the error distributed randomly and evenly in two sides of x-axis. The trend (red line) is horizonal, which means that the model specification (log-linear) is correct. There are still some patterns is the upper right and lower middle part, which means we still lose some variables to explain the price well.

In the residual QQ-plot, we can see that most of the residuals track the diagonal line, while some large residuals are larger than standardized, which means that there are some "outliers" in the listing data that are difficult to be explained by our model. This may because we still lack information – we missed some important regressors.

2.2 Machine Learning

For the machine learning part, we use linear regression, lasso regression, ridge regression, K-Neighbor regression, random forest regression, neural nets regression, extreme gradient boosting tree regression(XGBoost), lightgbm regression and catboost regression. The regressors we use is shown below. More detailed information can be seen in the attached jupyter notebook file.

```
col_numeric=['host_response_rate','host_listings_count',
             'accommodates','bathrooms','bedrooms','beds','security_deposit','guests_included',
             'extra_people','minimum_nights','maximum_nights','minimum_minimum_nights',
             'maximum_minimum_nights','minimum_maximum_nights','maximum_maximum_nights',
             'minimum_nights_avg_ntm','maximum_nights_avg_ntm','availability_30','availability_60',
             'availability_90','availability_365','number_of_reviews','number_of_reviews_ltm',
             'review_scores_rating','review_scores_accuracy','review_scores_cleanliness',
             'review_scores_checkin','review_scores_communication','review_scores_location',
             'review_scores_value','calculated_host_listings_count',
             'calculated_host_listings_count_entire_homes','calculated_host_listings_count_private_rooms',
             'calculated_host_listings_count_shared_rooms','reviews_per_month']
col_binary=['host_is_superhost','host_has_profile_pic','host_identity_verified','is_location_exact','requires_license',
            'instant_bookable','is_business_travel_ready','require_guest_profile_picture','require_guest_phone_verification'
           ]
col_factor=['host_response_time','neighbourhood','property_type','room_type','bed_type','cancellation_policy']
```

We split our regressors into three kinds: numeric regressors, binary regressors and factor regressors.

Since the goal of this part was to predict the price, we no longer needed to care too much about the internal relation of the Xs or whether the assumptions of regression are violated. Our goal is to try our best to learn the underlying pattern between regressors and target as much as possible so that we can minimize the predicting error as much as possible. We did know some of the regressors may have little use of predicting price, but we just put them into the model, maybe just as placeholders.

After choosing regressors, we also confronted a problem, that is, the dataset contained too many missing values. At first, we tried to fill them with the mean of their belonging columns or used KNN to fill them with the observation most similarly to them separately, but we found that, if doing so, we would alter nearly half of the observations, which may potentially harm the pattern underlying the dataset. Therefore, we chose to delete observations with missing value. We did concern the problem that if we used all listings that had been reviewed as training data, can the model really be applied to new listings that have no reviews? We will express this problem later.

For the string columns, we used one-hot encoding and label encoding to alter them accordingly.

The summary of the performance is shown in the table below.

Learning Method	Running Time	RMSE (Train)	RMSE (Test)
Linear Regression	1s	86.07	88.09
Ridge Regression	1s	86.16	87.73
Lasso Regression	1s	87.44	86.65
K-Neighbor Regression	30s	71	64.61

Random Forest Regression with Grid Search	5min	77.07	78.82
Neural Nets Regression	2min	63.7	73.83
XGBoost Regression with Grid Search	814min	0.72	27.96
LightGBM Regression with Grid Search	10min	10.79	33.79
CatBoost Regression with Grid Search	5min	17.05	33.11

As the table shows that, the top 3 predicting methods are all boosting methods - xgboost, lightgbm and catboost, all of which take longer time to run yet have rmse less than 35. As our analysis in step 1 shows that, the mean of listing price is $196.2. As a result, their rmse are only about 1/7 of the mean, which proves that all of the three can be very good predicting methods. However, it takes xgboost 814 minutes to do the grid search, which makes it impossible to be a predicting model that needs to be run frequently. For lightgbm and catboost, they share similar testing rmse. However, the running time of lightgbm is about five times that of catboost, although far less than xgboost, and lightgbm is a little bit overfitting than catboost. Also, we can look at their feature importance lists. Left is lgbm's, right is catboost's.

Feature Importance of LightGBM

Features	Importance
reviews_per_month	7418
availability_365	6582
number_of_reviews	6156
availability_90	5286
number_of_reviews_ltm	4980
host_listings_count	4486
availability_60	4141
review_scores_rating	3503
availability_30	3502
calculated_host_listings_count	3391
accommodates	2719
calculated_host_listings_count_entire_homes	2640
extra_people	2514
security_deposit	2507
maximum_nights	2127
minimum_nights_avg_ntm	2108
beds	1787
host_response_rate	1758
guests_included	1651
minimum_maximum_nights	1565

Feature Importance of CatBoost

Features	Importance
host_listings_count	14.024699
room_type	9.575382
bathrooms	9.119572
neighbourhood	8.310773
bedrooms	7.945154
security_deposit	6.041049
extra_people	4.788242
accommodates	4.151853
calculated_host_listings_count	4.053424
reviews_per_month	3.308572
calculated_host_listings_count_entire_homes	3.222434
guests_included	2.723390
availability_365	2.197554
property_type	1.955736
minimum_nights_avg_ntm	1.880276
number_of_reviews	1.654528
beds	1.275350
review_scores_rating	1.270184
cancellation_policy	1.194492
number_of_reviews_ltm	0.905064

We can see that the feature importance list of lgbm is not very useful. It cannot give any instruction to hosts or Airbnb. It may just wrap up all information and find it can predict price in such an arcane way very well. However, feature importance list of catboost tells us a lot. For example, host_listing_count has high feature importance, which because hosts who own lots of listings are professional hosts – they know how to clean and furniture their rooms, how to handle the leasing stuff, how to get along with renters, and most importantly, how to set a good price. It is also obvious and understandable that room type, bathrooms, neighborhood, bedrooms, accommodates, property type and beds will account for high importance. Such conclusions are consistent with

our word cloud analysis of reviews below, which also shows that most customers care a lot about location(neighborhood).

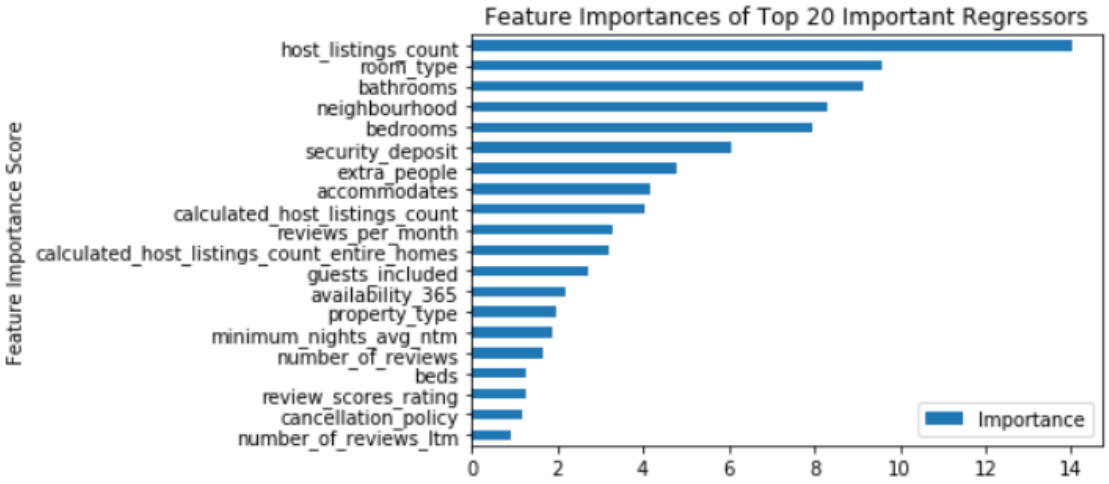


Figure 18: Top 20 Important Regressors in CatBoost model

Figure 19:Word Cloud of Reviews

To summarize, our best model is catboost regression, which is fast, accurate and meaningful. Airbnb can use our model as an internal calculator to know the real-time standard “market price” of a room and recommend that price to hosts. When hosts set unreasonable price, Airbnb can remind them. This prevents the situation that a host set too high price for his room that no one wants to rent and the situation that a host set too low price for his room that lose revenue from happening. It is also applicable to new listings. For new listings with no reviews or review score, our model can treat such regressors as zero and will suggest lower price than it should be. This is consistent with the principle that new products should be set low price in order to attract users to try them. Renters will be attracted by low price and get into renting this new room and make reviews and give scores. With reviews score above than zero, our model will predict higher price and gradually, the price for a room will converge to its market value. As a result, with the “best price calculator”, Airbnb can maximize the opportunity that a listing will be rented so as to increase revenue, improve customer loyalty and attract new users.

4. Insight and Recommendations:

4.1 How is Airbnb really being used in and affecting the neighborhoods?

With the discussion above, we can find that Airbnb is closer to a housing rental platform than a "shared economy" as it advertised. We see that the majority of its market is more or less controlled by professional landlords who do not live in their apartments regularly, and the occupancy rate is relatively high, even in slack seasons. That means the listings of Airbnb is mainly occupied by residents who are nonlocals to the neighborhood, driving out locals who tend to become landlords, which may disrupt natives and local communities and have potential legal risks. The areas near inner harbor is most popular.

4.2 Is there any trend of using Airbnb in Boston over time?

Listings: The listing numbers are growing steadily till 2018, it seems that the expanding seems met some bottlenecks in 2019 such as Boston housing regulation.

From the chart of reviews, we can see that in past 10 years, the popularity of Airbnb is growing gradually, but it has slowed down and even shows signs of decline in recent years. The using of Airbnb reaches peaks around summer holidays and reach the bottom around December. That shows a clear pattern.

Occupancy rate: From the given partial data, we can't see a significant upward trend. The logging and vacation rental market is super competitive, Airbnb cannot gain the market share from the other logging options easily.

4.3 What recommendation you will make to Airbnb hosts and Airbnb?

Suggestions for hosts:

1) Take extra attention on the neighborhood of their rooms. If the location of their room is great, they can set higher price for their rooms. If the location of their room is not so good, they should avoid setting too high price.

2) Make sure the rooms are clean and well prepared. Renters are willing to pay higher price in exchange for clean and convenient place.

3) Avoid using airbed. Using airbed is likely to drop price down.

4) Try to get along well with renters. Good hosts will impress renters and make renters give good reviews, which finally improves the value of the room.

Suggestions for Airbnb:

1) Use a "price calculator" to give real-time price recommendations to hosts. By correctly set price for listings, Airbnb can maximize the possibility that a trade is made and thus the possibility of increasing revenue.

2) Adopt policies that encourages customers to write reviews and give review scores. Currently, lots of customers do not write reviews. Airbnb can, for example, give coupons to customers who write reviews, which can increase number of reviews and ultimately increase average listing price and thus revenue.

3) Provide a playbook about how to prepare a room to hosts. By instructing the hosts how to do well in leasing their rooms, Airbnb can dig the most of value from its listings so that maximize revenue.

4) Find solutions of legal problems. Foreigners who use Airbnb may potential affect natives, causing legal problems. Airbnb should find good solutions in prepare.

Appendix:

1. Minimum occupancy rate model for a listing

We can't get the occupancy rate from the data given directly. So we use the reviews per month, average stay and review rate to estimate the minimum occupancy rate for a listing. If the minimum stay is larger than average stay, we use minimum stay to multiply instead.

$$Occupancy\ rate = \frac{review\ per\ month\ \times \max(avg\ stay,\ minimum\ stay)}{review\ rate}$$

Since the stats' distribution is skewed, we also calculate different percentiles.

We estimated the average stay would be 3.27 days based on the Visitors Bureau's conclusion. They estimate the average overseas visitor (18.5%) spends 8 nights in Boston when visiting while the average visitor from somewhere in the United States spends 2.2 nights (81.5%)[2]. We got the travel demographic statistics from Logan Airport, international passengers 7,583,887 and domestic passengers 33,245,880.

As for the review rate: we choose 80%, a number answered by an Airbnb employee.

2. To keep conciseness of the report, we put the code in a separate document. However, the most clear and easiest way to follow up our logit is to look at our raw code which is also attached in the zip file.

[2] https://www.bizjournals.com/boston/news/2015/09/16/bostons-2015-tourism-season-was-best-in-years.html