What Predicts Airbnb Prices?

Tongyu Chen, Wendi Chen, Chris Fan, Xingyou (Richard) Song

September 9, 2017

Abstract

In this report, we discuss the various factors that affect Airbnb price, using the given data. Because in the Airbnb hosting service, the host sets a price for the room, ultimately, this becomes a problem about the prediction of human behavior, and which factors determine a host's judgment of a fair price. We were able to show which aspects of an Airbnb order affect the price the most. In doing so, this data may be used to improve pricing for better revenue, as well as give better recommendations for users when searching for new Airbnb locations.

1 Background Information

In AirBnB, hosts may advertise a room for rental, and set the price for themselves, as a form of bidding. Naturally, in rural areas, rooms may be much cheaper than compared to rooms given in metropolitan areas. Furthermore, the likelihood of buying an AirBnB is intuitively affected by the ability to pay for its users. Because the main revenue AirBnB is generated from collecting profit in the form of service fees, it is very important to collect the requests from a buyer, in order to accurately predict the price of the AirBnB requested. Having this predictability power, based on the demographic data from the region, as well as the quality of the AirBnB room, is therefore very important for optimizing both user experience as well as profit.

2 Preliminaries

In this section we discuss the background on the resources given, and justify our reasoning for our investigation. Because zip-codes are well known to have similar prediction patterns, for each zip-code, we aggregated the prices of the AirBnB listings within that zip code.

Because there can be high variance within the prices in each zip code, as a form of data cleaning, we removed zip codes with less than 50 data samples in order to understand aggregate effects. Furthermore, we used the median price in each zip code as an output variable. As shown in 2, one reason for using median is the slight in the correlation value compared to using average.

The most intuitive predictor for AirBnB housing prices was the overall housing price within the area. We verified this claim, as shown below.

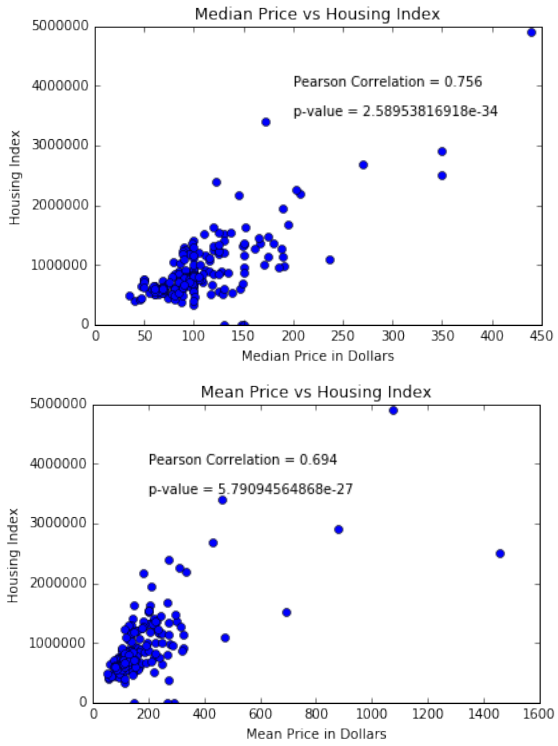


Figure 1: Scatterplot of median zip code prices vs housing index, and mean zip code prices vs housing index. We had 178 points within the plot.

As shown above, the correlation from median aggregation being 0.75 and its p-value being so low already showed a substantial correlation. We then attempted to find, using the various data sources, which other factors would generate good correlations with the prices. Using the data outside of listings.csv, we considered:

- Demographic Data - For each city, do the population and income bracket affect the aggregate prices in the city?
- Which types of venues affect the zip code's prices? Do grocery stores or bars affect the pricing more?

Furthermore, we also analyzed the data inside the listings data, and asked questions such as:

- Does the type of rental (bed type, accommodations, property type, number of bathrooms) affect the cost of the AirBnB?
- Do the review scores affect the cost?

2.1 Constraints with the Data

The listings only consider West Coast AirBnB data, and of the 120 cities, 119 cities were from California, and 1 city (Seattle) was from Washington. The econ_state.csv data was therefore irrelevant, because it only considered the GDP of each state.

Furthermore, many of the cities from the AirBnB listings data were not featured in any of the demographics.csv data. In fact, over 50 of the cities collected from the listings were not present in any of the demographics.csv data, and there were multiple naming mismatches from the cities in the listing.csv data with the demographics.csv data. Therefore, we could only analyze some of the major cities and how income brackets or age groups could affect prices.

In the venues data, only the cities were listed for each venue - the zip codes were absent, and the cities in the venue.csv file also mismatched in exact naming with the cities in the listings.csv

file, which made it difficult to analyze the data within each zip code. A method to circumvent this is to assign a venue to the closest AirBnB listing's zip code based on Longitude/Latitude data, but this took too long.

In the interest of time therefore, our **main focus** was how different factors within an AirBnB listing may affect the price, while we also still provide order statistics and aggregate relationships for general locations.

3 Aggregate Data Relationships

In our aggregate exploratory data analysis, outside of the housing indices, there were no significant relationships to the other data. We examined if there existed any relationships from factors such as the income from a city, as well as the population. Intuitively, AirBnB is generally used by people within the younger age, corresponding to ages (20 - 35). Furthermore, an educated guess would be that the income bracket of a region is also correlated with the ability to pay.

As shown below, these have insignificant affects on the median Airbnb pricing for major cities.

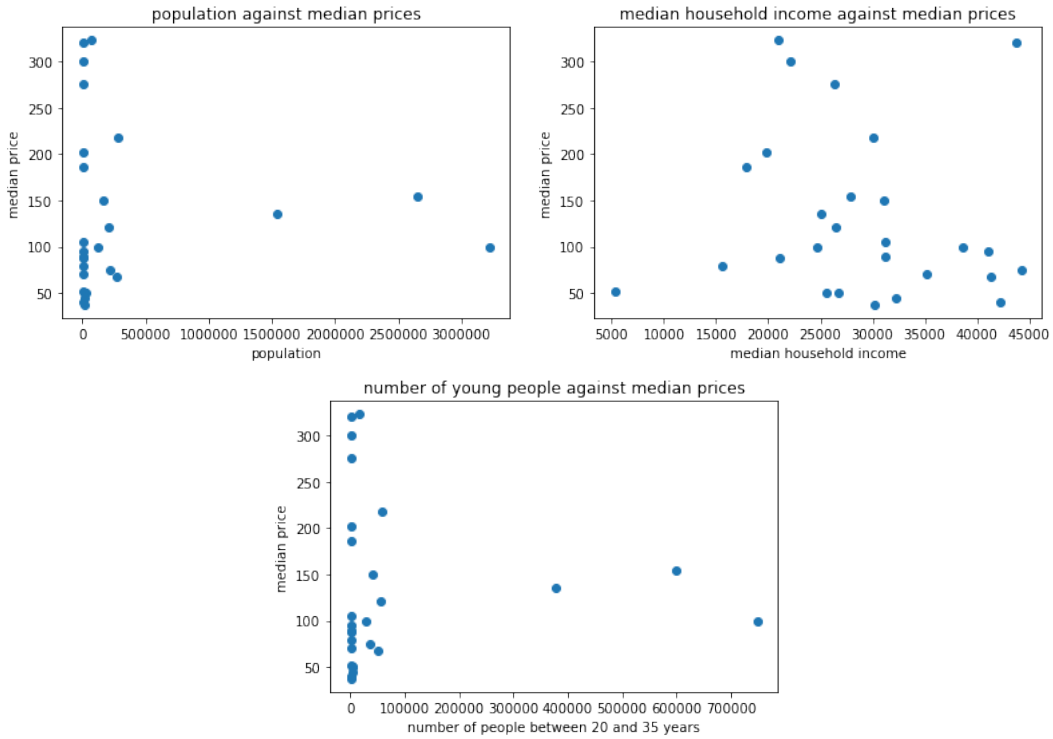


Figure 2: There were 27 cities possible used. (Left) Scatterplot of population vs median AirBnB price for each city as a data point. (Right) Scatterplot of median income vs median AirBnB price for each city as a data point. (Bottom) Scatterplot of population in ages (20-35)

Furthermore, we checked the aggregate data of the venues in each city. Unfortunately, there were only 7 cities that provided venues in the data - these were San Diego, Oakland, Portland, LA, Seattle, SF, and Santa Cruz. As shown below in 3, there was no determinable relationship between number of venues and AirBnB pricing - however, there was a decreasing relationship between transit/total venue ratio to the median price as shown in 3, which suggests AirBnB prices lower when there is more transit, suggesting that people may use AirBnB to stay nights when travel is difficult.

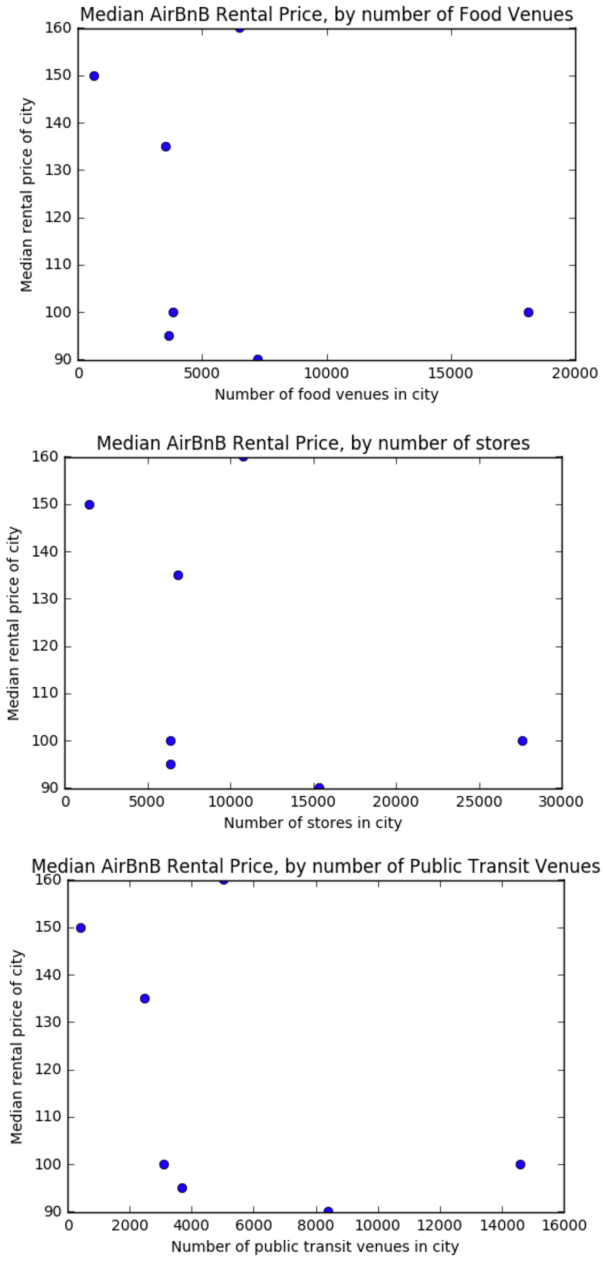


Figure 3: There were 7 cities provided for the Venue data. Each of the graphs above compare the median AirBnB price compared to the aggregate number of venues (stores, food, transits) - Note that the store numbers are also similar - the number of specific types of venues have high correlation with population.

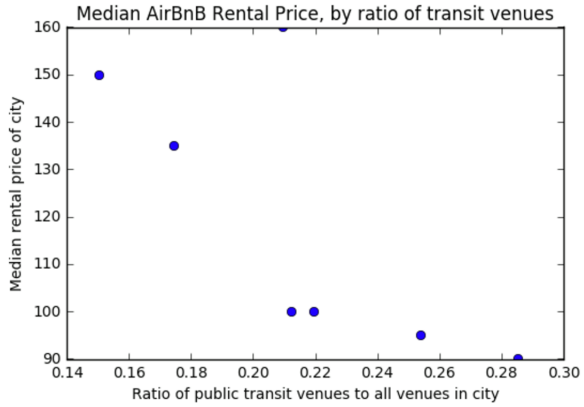


Figure 4: Using the 7 cities, we compared the transit venue/total venue ratio, and compared it to the price. In the absence of more data, however, this is promising and worth investigating.

4 Most Important Factors to a Listing Price from Regression

From the listings data, we performed linear regression, with the response variable as the price of the listing, and the factors from the data in the listings.csv file, as well as the extra factor from the housing index given to each listing based on the common zip code.

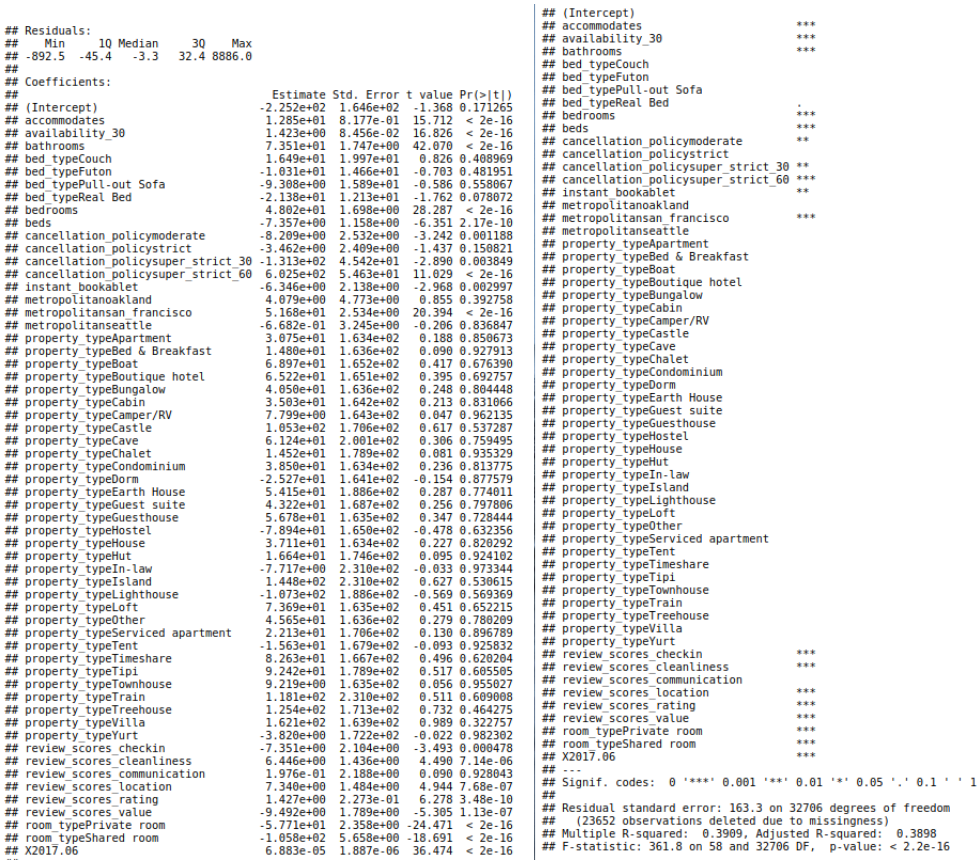
For the technical details on how we applied the regression, there was no regularization because our parameters were far fewer than the number of data points. Furthermore, because there were discrete variables, we used the dummy-variable technique for our regression, assigning 0-1 inputs into the regression, based on the status.

4.1 Most Important Factors

The most important factors we found were (from ***, see in figure caption below for explanation in 4.1): accommodates, availability_30, bathrooms, bedrooms, beds, cancellation_policy_super_strict_60, metropolitan_san_francisco, all review_scores except communication, and X2017.06 (housing index).

An interesting observation for this is that the variable corresponding to whether the AirBnB was located in San Francisco (metropolitan_san_francisco) was very significant. From this result, we can infer that predictability of the price of an AirBnB comes from the accomodations provided, review scores, sharing policies of the room, and the housing index within the area.

What are *not important* are the quality/type of the beds, the property type, the communication review score,



```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -892.5  -45.4   -3.3   32.4 8886.0
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -2.252e+02  1.646e+02  -1.368 0.171265
## accommodates                     1.285e+01  8.177e-01  15.712  < 2e-16
## availability_30                  1.423e+00  8.456e-02  16.826  < 2e-16
## bathrooms                        7.351e+01  1.747e+00  42.070  < 2e-16
## bed_typeCouch                    1.649e+01  1.997e+01   0.826 0.408969
## bed_typeFuton                   -1.031e+01  1.466e+01  -0.703 0.481951
## bed_typePull-out Sofa           -9.308e+00  1.589e+01  -0.586 0.558067
## bed_typeReal Bed                -2.138e+01  1.213e+01  -1.762 0.078072
## bedrooms                         4.802e+01  1.698e+00  28.287  < 2e-16
## beds                            -7.357e+00  1.158e+00  -6.351 2.17e-10
## cancellation_policymoderate     -8.209e+00  2.532e+00  -3.242 0.001188
## cancellation_policystrict       -3.462e+00  2.409e+00  -1.437 0.150821
## cancellation_policysuper_strict_30 -1.313e+02  4.542e+01  -2.890 0.003849
## cancellation_policysuper_strict_60  6.025e+02  5.463e+01  11.029  < 2e-16
## instant_bookablet               -6.346e+00  2.138e+00  -2.968 0.002997
## metropolitanoakland              4.079e+00  4.773e+00   0.855 0.392758
## metropolitansan_francisco        5.168e+01  2.534e+00  20.394  < 2e-16
## metropolitanseattle             -6.682e-01  3.245e+00  -0.206 0.836847
## property_typeApartment           3.075e+01  1.634e+02   0.188 0.850673
## property_typeBed & Breakfast     1.480e+01  1.636e+02   0.090 0.927913
## property_typeBoat                6.897e+01  1.652e+02   0.417 0.676390
## property_typeBoutique hotel      6.522e+01  1.651e+02   0.395 0.692757
## property_typeBungalow            4.050e+01  1.636e+02   0.248 0.804448
## property_typeCabin               3.503e+01  1.642e+02   0.213 0.831066
## property_typeCamper/RV           7.799e+00  1.643e+02   0.047 0.962135
## property_typeCastle              1.053e+02  1.706e+02   0.617 0.537287
## property_typeCave                6.124e+01  2.001e+02   0.306 0.759495
## property_typeChalet              1.452e+01  1.789e+02   0.081 0.935329
## property_typeCondominium         3.850e+01  1.634e+02   0.236 0.813775
## property_typeDorm               -2.527e+01  1.641e+02  -0.154 0.877579
## property_typeEarth House         5.415e+01  1.886e+02   0.287 0.774011
## property_typeGuest suite         4.322e+01  1.687e+02   0.256 0.797806
## property_typeGuesthouse          5.678e+01  1.635e+02   0.347 0.728444
## property_typeHostel             -7.894e+01  1.650e+02  -0.478 0.632356
## property_typeHouse               3.711e+01  1.634e+02   0.227 0.820292
## property_typeHut                 1.664e+01  1.746e+02   0.095 0.924102
## property_typeIn-law             -7.717e+00  2.310e+02  -0.033 0.973344
## property_typeIsland              1.448e+02  2.310e+02   0.627 0.530615
## property_typeLighthouse         -1.073e+02  1.886e+02  -0.569 0.569369
## property_typeLoft                7.369e+01  1.635e+02   0.451 0.652215
## property_typeOther               4.565e+01  1.636e+02   0.279 0.780209
## property_typeServiced apartment  2.213e+01  1.706e+02   0.130 0.896789
## property_typeTent               -1.563e+01  1.679e+02  -0.093 0.925832
## property_typeTimeshare           8.263e+01  1.667e+02   0.496 0.620204
## property_typeTipi                9.242e+01  1.789e+02   0.517 0.605505
## property_typeTownhouse           9.219e+00  1.635e+02   0.056 0.955027
## property_typeTrain               1.181e+02  2.310e+02   0.511 0.609008
## property_typeTreehouse           1.254e+02  1.713e+02   0.732 0.464275
## property_typeVilla               1.621e+02  1.639e+02   0.989 0.322757
## property_typeYurt               -3.820e+00  1.722e+02  -0.022 0.982302
## review_scores_checkin           -7.351e+00  2.104e+00  -3.493 0.000478
## review_scores_cleanliness        6.446e+00  1.436e+00   4.490 7.14e-06
## review_scores_communication      1.976e-01  2.188e+00   0.090 0.928043
## review_scores_location           7.340e+00  1.484e+00   4.944 7.68e-07
## review_scores_rating             1.427e+00  2.273e-01   6.278 3.48e-10
## review_scores_value             -9.492e+00  1.789e+00  -5.305 1.13e-07
## room_typePrivate room           -5.771e+01  2.358e+00 -24.471  < 2e-16
## room_typeShared room            -1.058e+02  5.658e+00 -18.691  < 2e-16
## X2017.06                         6.883e-05  1.887e-06  36.474  < 2e-16
## (Intercept)
## accommodates                     ***
## availability_30                  ***
## bathrooms                        ***
## bed_typeCouch
## bed_typeFuton
## bed_typePull-out Sofa
## bed_typeReal Bed                 .
## bedrooms                         ***
## beds                             ***
## cancellation_policymoderate      **
## cancellation_policystrict
## cancellation_policysuper_strict_30 **
## cancellation_policysuper_strict_60 ***
## instant_bookablet                **
## metropolitanoakland
## metropolitansan_francisco        ***
## metropolitanseattle
## property_typeApartment
## property_typeBed & Breakfast
## property_typeBoat
## property_typeBoutique hotel
## property_typeBungalow
## property_typeCabin
## property_typeCamper/RV
## property_typeCastle
## property_typeCave
## property_typeChalet
## property_typeCondominium
## property_typeDorm
## property_typeEarth House
## property_typeGuest suite
## property_typeGuesthouse
## property_typeHostel
## property_typeHouse
## property_typeHut
## property_typeIn-law
## property_typeIsland
## property_typeLighthouse
## property_typeLoft
## property_typeOther
## property_typeServiced apartment
## property_typeTent
## property_typeTimeshare
## property_typeTipi
## property_typeTownhouse
## property_typeTrain
## property_typeTreehouse
## property_typeVilla
## property_typeYurt
## review_scores_checkin            ***
## review_scores_cleanliness        ***
## review_scores_communication
## review_scores_location           ***
## review_scores_rating             ***
## review_scores_value              ***
## room_typePrivate room            ***
## room_typeShared room             ***
## X2017.06                         ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 163.3 on 32706 degrees of freedom
##   (23652 observations deleted due to missingness)
## Multiple R-squared:  0.3909, Adjusted R-squared:  0.3898
## F-statistic: 361.8 on 58 and 32706 DF,  p-value: < 2.2e-16
```

Figure 5: Linear Regression results from linear regression on the relevant variables found in the listings.csv data, as well as the housing index (X2017.06) from the real estate data. At the right, (*) denotes how significant the variable was, based on the p-value found from regression, with the labelling rule found at the very bottom -e.g. (***) means p-value was less than 0.001.

4.2 Using Only Three-Star Factors

We also performed regression based on only using the factors with (***) to check performance.

Using linear regression after a non-linear power transformation of the price (output variable is price$^{-0.15}$), with only using the significant (***) variables, we obtain the following results with a higher R^2 value of 0.72:

```
Console ~/
Call:
lm(formula = price^(-0.15) ~ accommodates + availability_30 +
    bathrooms + bedrooms + beds + cancellation_policy + instant_bookable +
    metropolitan + review_scores_checkin + review_scores_cleanliness +
    review_scores_location + review_scores_rating + review_scores_value +
    room_type + X2017.06, data = list2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.242146 -0.015687  0.000565  0.016384  0.264002

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      6.015e-01  2.657e-03 226.416  < 2e-16 ***
accommodates                    -5.066e-03  1.329e-04 -38.130  < 2e-16 ***
availability_30                 -4.351e-04  1.380e-05 -31.527  < 2e-16 ***
bathrooms                       -5.593e-03  2.828e-04 -19.779  < 2e-16 ***
bedrooms                        -1.340e-02  2.683e-04 -49.928  < 2e-16 ***
beds                             2.650e-03  1.852e-04  14.309  < 2e-16 ***
cancellation_policymoderate      9.779e-04  4.140e-04   2.362  0.01818 *
cancellation_policystrict       -4.220e-04  3.936e-04  -1.072  0.28371
cancellation_policysuper_strict_30  4.168e-02  7.437e-03   5.604 2.11e-08 ***
cancellation_policysuper_strict_60 -5.104e-02  8.940e-03  -5.709 1.14e-08 ***
instant_bookablet                2.949e-03  3.486e-04   8.459  < 2e-16 ***
metropolitanoakland             -4.015e-03  7.790e-04  -5.154 2.56e-07 ***
metropolitansan_francisco       -2.760e-02  4.079e-04 -67.665  < 2e-16 ***
metropolitanseattle             -1.623e-03  5.293e-04  -3.066  0.00217 **
review_scores_checkin            2.436e-03  3.030e-04   8.041 9.23e-16 ***
review_scores_cleanliness       -2.737e-03  2.346e-04 -11.667  < 2e-16 ***
review_scores_location          -5.803e-03  2.416e-04 -24.016  < 2e-16 ***
review_scores_rating            -5.426e-04  3.621e-05 -14.985  < 2e-16 ***
review_scores_value              5.083e-03  2.914e-04  17.446  < 2e-16 ***
room_typePrivate room            4.178e-02  3.655e-04 114.327  < 2e-16 ***
room_typeShared room             9.598e-02  8.843e-04 108.531  < 2e-16 ***
X2017.06                        -1.757e-08  3.081e-10 -57.037  < 2e-16 ***
---
Signif. codes:  0 ?***?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 0.02673 on 32744 degrees of freedom
  (23650 observations deleted due to missingness)
Multiple R-squared:  0.7247,    Adjusted R-squared:  0.7245
F-statistic:  4104 on 21 and 32744 DF,  p-value: < 2.2e-16

>
```

Figure 6: Linear Regression results from linear regression on the (***) variables found from the original OLS. Note that the R^2 value increases.

5 Conclusion and Future Study

From this study, we provided a satisfactory predictor of the price of a AirBnB offer given its metadata. We also isolated the most important factors, as well as shown that some of the aggregate data at the city level is irrelevant for our analysis, while transit ratio (which implies metropolitan areas) may affect prices. This method can be improved in the following ways:

Due to the interest of time and data availability, we were unable to perform a good cross-validation study. Because the AirBnB listings were only made during the 2017-2018 period, with low sampling within few months, we could not roll the monthly data across the entire time period to check consistency of results. With more data available (e.g. from 2016 or 2015), we would be able to check the consistency of our model, and see if it overfits.

Furthermore, we do not need price to be the only response variable. We may be able to perform PCA on the entire data set to check for dimensionality reduction and important factors - i.e. check the relationship between variables such as e.g. review quality and availability.

Because of a lack of fine-grained data from our given sources (zip-code level data for demographics and venues), we are currently unable to see any real relationships between the venues and the AirBnB general pricings. Hopefully with the zip code, we can determine more information.

We hope that with enough improvement, this model can predict prices better, in order to increase revenue and help users select appropriate AirBnB rentals.

Thank you!