# Airbnb Lab

Xubin Lou and Jiayue Meng

***Abstract*— Through summarizing data about the Airbnb listing and review dataset, descriptive analysis with statistics, Sentiment Analysis and adding new data, data mining, creating linear regression model with OLS, and visualization, the report of the Airbnb Lab tries to figure out the relationship of multifactors on the Airbnb price.**

## I. INTRODUCTION

We expected to find the relation of multifactors to the price of Airbnb. We firstly expect that the comments with positive and negative words will be the main significant effect to the price of Airbnb. Nevertheless, as we do linear regression with positivity mean and negativity mean variables with price, we find the coefficient of them are not significant in 0.05 significant level. Thus, the comments are not that significant as we thought before to affect the price.

## II. DATASET

In the Airbnb Lab, there are three datasets: calendar.csv, listings.csv, and reviews.csv. Here we mainly use listings and reviews datasets to do the data analysis. Firstly, we set up the dataset and descriptive Statistics, process the data before the analysis like transferring to appropriate data type and format to prepare for the EDA analysis. In the EDA analysis, we got the minimum, maximum, mean, median, variance, and standard deviation of each given numerical variables in the listing dataset. Figure 1, 2.

| Variables | Minimum | Maximum | Mean | Median | Variance | Std. Deviation |
|---|---|---|---|---|---|---|
| host_response_rate | 0.0 | 1.0 | 0.9498908156711676 | 1.0 | 0.015664214154569103 | 0.12515675832558584 |
| host_acceptance_rate | 0.0 | 1.0 | 0.8417308927424536 | 0.94 | 0.047418359180513216 | 0.2177575697433116 |
| host_listings_count | 0 | 749 | 58.9023709902371 | 2.0 | 29281.9391266313 | 171.1196631794 |
| host_total_listings_count | 0 | 749 | 58.9023709902371 | 2.0 | 29281.9391266313 | 171.1196631794 |
| accommodates | 1 | 16 | 3.0412831241283125 | 2.0 | 3.164589870990237 | 1.7789294170905818 |
| bathrooms | 0.0 | 6.0 | 1.221646597591711 | 1.0 | 0.2514188513038815 | 0.5014168438573653 |
| bedrooms | 0.0 | 5.0 | 1.255944055944056 | 1.0 | 0.5669401926744584 | 0.752954309818636 |
| beds | 0.0 | 16.0 | 1.6090604026845639 | 1.0 | 1.0207716716744826 | 1.0103324560136047 |
| price | 10.0 | 4000.0 | 173.9258019525802 | 150.0 | 21996.04358809467 | 148.31063208042326 |
| weekly_price | 80.0 | 5000.0 | 922.3923766816143 | 750.0 | 432244.4200315712 | 657.452979331276 |
| monthly_price | 500.0 | 40000.0 | 3692.097972972973 | 2925.0 | 8400319.16945535 | 2898.3304106770415 |
| security_deposit | 95.0 | 4500.0 | 324.6982116244411 | 250.0 | 108076.90519799397 | 328.75052121326587 |
| cleaning_fee | 5.0 | 300.0 | 68.38014527845036 | 50.0 | 2630.405933473648 | 51.287483204712316 |
| guests_included | 0 | 14 | 1.4298465829846583 | 1.0 | 1.1167986650727235 | 1.0567869534928616 |

Fig. 1. EDA1

For the listings dataset, there are many missing data. Something strange is that the differences of variance between the variables are huge. Some data like maximum nights, host listings count, host total listings count, price, weekly price, monthly price have extremely large variance and standard deviation. Thus, these data are more likely to be badly measured compared with the data with small variance like host response rate and host acceptance rate with variance

| | | | | | | |
|---|---|---|---|---|---|---|
| cleaning_fee | 5.0 | 300.0 | 68.38014527845036 | 50.0 | 2630.405933473648 | 51.287483204712316 |
| guests_included | 0 | 14 | 1.4298465829846583 | 1.0 | 1.1167986650727235 | 1.0567869534928616 |
| extra_people | 0.0 | 200.0 | 10.886192468619248 | 0.0 | 366.1521803423143 | 19.135103353321984 |
| minimum_nights | 1 | 300 | 3.1712691771269177 | 2.0 | 78.75023457735604 | 8.874132891576282 |
| maximum_nights | 1 | 99999999 | 28725.83682008368 | 1125.0 | 2789354050349.153 | 1670135.9376856582 |
| availability_30 | 0 | 30 | 8.649930264993026 | 4.0 | 108.89611118375176 | 10.43532995088089 |
| availability_90 | 0 | 90 | 38.5581589958159 | 37.0 | 1099.4710166990435 | 33.1582722212579 |
| availability_365 | 0 | 365 | 179.34644351464436 | 179.0 | 20202.693559474 | 142.1361796288123 |
| number_of_reviews | 0 | 404 | 19.04463040446304 | 5.0 | 1265.3428736426579 | 35.57165829199783 |
| review_scores_rating | 20.0 | 100.0 | 91.91666666666667 | 94.0 | 90.82025613275613 | 9.529966218867521 |
| review_scores_accuracy | 2.0 | 10.0 | 9.43157132512672 | 10.0 | 0.868054663450018 | 0.9316945118707194 |
| review_scores_cleanliness | 2.0 | 10.0 | 9.25804119985544 | 10.0 | 1.3660132212877545 | 1.1687656827986328 |
| review_scores_checkin | 2.0 | 10.0 | 9.64629294755877 | 10.0 | 0.5815820986301907 | 0.762615301859457 |
| review_scores_communication | 4.0 | 10.0 | 9.646548608601373 | 10.0 | 0.5407750412765244 | 0.7353740825433844 |
| review_scores_value | 2.0 | 10.0 | 9.16823444283647 | 9.0 | 1.0219866078440818 | 1.010933532851731 |
| reviews_per_month | 0.01 | 19.15 | 1.970908448214916 | 1.17 | 4.495191362774157 | 2.1201866339485673 |

Fig. 2. EDA2

less than 0.05. An obvious contrast here is the variance of minimum nights is about 78, while the variance of maximum nights 2789354050349 is extremely larger.

Then, we conduct sentiment analysis and adding new data, focus on the review dataset, use sentiment analysis template.py to get the new four columns negativity, neutrality, positivity, and compound and add then to the review dataset. Also, calculate and add positivity simple and negativity simple these two new column to the review dataset to show the proportion of positive words and negative words respectively in each comments. Figure 3

| d | id | date | reviewer_id | reviewer_name | comments | negativity | neutrality | positivity | compound | total_number_words | positivity_simple | negativity_simple |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 4724140 | 2013-05-21 | 4298113 | Olivier | my stay at islams place was really cool good l... | 0.0 | 0.648 | 0.352 | 0.9626 | 47 | 0.127660 | 0.042553 |
| 2 | 4869189 | 2013-05-29 | 6452964 | Charlotte | great location for both airport and city grea... | 0.0 | 0.639 | 0.361 | 0.9061 | 23 | 0.130435 | 0.043478 |
| 2 | 5003196 | 2013-06-06 | 6449554 | Sebastian | we really enjoyed our stay at islams house fro... | 0.0 | 0.767 | 0.233 | 0.9663 | 86 | 0.058140 | 0.046512 |
| 2 | 5150351 | 2013-06-15 | 2215611 | Marine | the room was nice and clean and so were the co... | 0.0 | 0.673 | 0.327 | 0.9267 | 36 | 0.138889 | 0.027778 |
| 2 | 5171140 | 2013-06-16 | 6848427 | Andrew | great location just 5 mins walk from the airpo... | 0.0 | 0.637 | 0.363 | 0.8658 | 22 | 0.136364 | 0.000000 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 3 | 80537457 | 2016-06-18 | 22034145 | Antonio | joe y su mujer son encantadores la habitación ... | 0.0 | 0.946 | 0.054 | 0.34 | 48 | 0.041667 | 0.041667 |
| 3 | 83640094 | 2016-07-03 | 40052513 | Steve | joe was on his way to jamaica to be married on... | 0.014 | 0.822 | 0.164 | 0.9504 | 129 | 0.077519 | 0.062016 |
| 3 | 85797088 | 2016-07-13 | 77129134 | Nick | the room was very clean as were the | 0.0 | 0.784 | 0.216 | 0.9693 | 92 | 0.097826 | 0.032609 |

Fig. 3. comments

Next, we try to find the unique values of listings in the reviews dataset (listings id column). Calculate the average scores for each listing and name them in the following way: negativity mean, neutrality mean, positivity mean, compound mean, positivity simple mean, negativity simple mean. Add these values to the listings dataset as new columns. Figure 4, 5.

| | listing_id | listing_url | scrape_id | last_scraped | name | summary | space | description | experiences_offered | neighbor |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 12147973 | https://www.airbnb.com/rooms/12147973 | 2.020000e+13 | 9/7/16 | Sunny Bungalow in the City | Cozy, sunny, family home. Master bedroom high... | The house has an open and cozy feel at the sam... | Cozy, sunny, family home. Master bedroom high... | none | Ro conveni |
| 1 | 3075044 | https://www.airbnb.com/rooms/3075044 | 2.020000e+13 | 9/7/16 | Charming room in pet friendly apt | Charming and quiet room in a second floor 1910... | Small but cozy and quite room with a full size... | Charming and quiet room in a second floor 1910... | none | Roslindal |
| 2 | 6976 | https://www.airbnb.com/rooms/6976 | 2.020000e+13 | 9/7/16 | Mexican Folk Art Haven in Boston | Come stay with a friendly, middle-aged guy in ... | Come stay with a friendly, middle-aged guy in ... | Come stay with a friendly, middle-aged guy in ... | none | Roslind |
| 3 | 1436513 | https://www.airbnb.com/rooms/1436513 | 2.020000e+13 | 9/7/16 | Spacious Sunny Bedroom Suite in Historic Home | Come experience the comforts of home away from... | Most places you find in Boston are small howev... | Come experience the comforts of home away from... | none | Roslindal neig |
| 4 | 7651065 | https://www.airbnb.com/rooms/7651065 | 2.020000e+13 | 9/7/16 | Come Home to Boston | My comfy, clean and relaxing home is one block... | Clean, attractive, private room, one block fro... | My comfy, clean and relaxing home is one block... | none | I love |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 3580 | 8373729 | https://www.airbnb.com/rooms/8373729 | 2.020000e+13 | 9/7/16 | Big cozy room near T | 5 min walking to Orange Line subway | NaN | 5 min walking to Orange Line | none | |

Fig. 4. mean1

| ulated_host_listings_count | reviews_per_month | negativity_simple_mean | positivity_simple_mean | negativity_mean | positivity_mean | neutrality_mean | compound_mean |
|---|---|---|---|---|---|---|---|
| 1 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | 1.30 | 0.042922 | 0.128897 | 0.014000 | 0.321944 | 0.664056 | 0.837644 |
| 1 | 0.47 | 0.044825 | 0.119082 | 0.010854 | 0.264659 | 0.724488 | 0.905349 |
| 1 | 1.00 | 0.041667 | 0.125000 | 0.000000 | 0.484000 | 0.516000 | 0.950600 |
| 1 | 2.25 | 0.045183 | 0.141893 | 0.017034 | 0.273897 | 0.709069 | 0.780959 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 8 | 0.34 | 0.038744 | 0.097125 | 0.019000 | 0.219500 | 0.761500 | 0.811475 |

Fig. 5. mean2

```
     support                                         itemsets
46  0.100418  (Entire home/apt, 1.0 bathrooms, 2.0 bedrooms)
16  0.100418                   (1.0 bathrooms, 2.0 bedrooms)
30  0.101255                 (Entire home/apt, 2.0 bathrooms)
38  0.102929    (Private room, 1.0 bedrooms, 1 accommodates)
13  0.102929               (Private room, 1 accommodates)
..       ...                                              ...
9   0.593305                              (Entire home/apt)
19  0.597768                      (1.0 bathrooms, Apartment)
2   0.663598                                 (1.0 bedrooms)
8   0.728591                                    (Apartment)
1   0.767364                                (1.0 bathrooms)

[70 rows x 2 columns]
```

Fig. 6. minSup=0.1

```
     support                                           itemsets
29  0.216179  (1.0 bathrooms, 1.0 bedrooms, Apartment, 2 acc...
30  0.217573  (Entire home/apt, 1.0 bathrooms, 1.0 bedrooms,...
28  0.219247            (Private room, 1.0 bedrooms, Apartment)
18  0.219247                          (Private room, Apartment)
27  0.225105         (Entire home/apt, 1.0 bedrooms, Apartment)
25  0.233752         (1.0 bedrooms, Apartment, 2 accommodates)
26  0.238494      (Private room, 1.0 bedrooms, 2 accommodates)
16  0.238494                     (Private room, 2 accommodates)
21  0.247141     (Entire home/apt, 1.0 bathrooms, 1.0 bedrooms)
13  0.256904                      (Entire home/apt, 1.0 bedrooms)
23  0.272245        (1.0 bathrooms, Apartment, 2 accommodates)
15  0.290934                        (Apartment, 2 accommodates)
19  0.297908       (1.0 bathrooms, 1.0 bedrooms, 2 accommodates)
10  0.301813                     (Private room, 1.0 bathrooms)
22  0.301813       (Private room, 1.0 bathrooms, 1.0 bedrooms)
11  0.350907                     (1.0 bedrooms, 2 accommodates)
7   0.358996                    (1.0 bathrooms, 2 accommodates)
14  0.384379                      (Private room, 1.0 bedrooms)
5   0.384379                                    (Private room)
24  0.387727       (Entire home/apt, 1.0 bathrooms, Apartment)
2   0.413668                                 (2 accommodates)
20  0.427615          (1.0 bathrooms, 1.0 bedrooms, Apartment)
9   0.446583                   (Entire home/apt, 1.0 bathrooms)
12  0.461646                          (1.0 bedrooms, Apartment)
17  0.492050                       (Entire home/apt, Apartment)
6   0.567922                      (1.0 bathrooms, 1.0 bedrooms)
4   0.593305                                 (Entire home/apt)
8   0.597768                        (1.0 bathrooms, Apartment)
1   0.663598                                    (1.0 bedrooms)
3   0.728591                                       (Apartment)
0   0.767364                                   (1.0 bathrooms)
```

Fig. 7. minSup=0.2

## III. RESULT

After finishing the data processing, descriptive statistics, conducting sentiment analysis and adding new data, we begin data mining process the further analyze. We look at the following variables in the listings.csv file: property type, room type, accommodates, bathrooms, bedrooms. Using the Apriori algorithm to calculate the frequent itemsets in the listings dataset.

With minSup is 0.1, the top 5 most frequent itemsets are (1.0 bathrooms), (Apartment), (1.0 bedrooms), (Apartment, 1.0 bathrooms), (Entire home/apt). The top 5 least frequent itemsets are (1.0 bathrooms, Entire home/apt, 2.0 bedrooms), (1.0 bathrooms, 2.0 bedrooms), (Entire home/apt, 2.0 bathrooms), (1.0 bedrooms, 1 accommodates, Private room), (1 accommodates, Private room). Figure 6

With minSup is 0.2, the top 5 most frequent itemsets are (1.0 bathrooms), (Apartment), (1.0 bedrooms), (Apartment, 1.0 bathrooms), (Entire home/apt). The top 5 least frequent itemsets are (1.0 bedrooms, Apartment, 1.0 bathrooms, 2 accommodates), (1.0 bedrooms, Apartment, 1.0 bathrooms, Entire home/apt), (1.0 bedrooms, Apartment, Private room), (Apartment, Private room), (1.0 bedrooms, Apartment, Entire home/apt). Figure 7

I found that under the minSup 0.1 and 0.2 have the same top 5 most frequent itemsets.

We do the linear regression model that uses price as the outcome variable (Y) price and all the other variables listed above as explanatory variables (X's) host response rate, review scores rating, review scores accuracy, review scores cleanliness, review scores checkin, review scores communication, positivity mean, negativity mean, positivity simple mean, negativity simple mean. Here, x1=host response rate, x2=review scores rating, x3=review scores accuracy, x4=review scores cleanliness, x5=review scores checkin, x6=review scores communication, x7=positivity mean, x8=negativity mean, x9=positivity simple mean, x10=negativity simple mean. Figure 8

The R-square is 0.05. According to to P-value in the chart and 5 percent confidence interval, we find the P-value of the coefficients of constant, review scores rating, review scores accuracy, review scores cleanliness, review scores checkin, positivity simple mean, and negativity simple mean are less than the the significance level 0.05. Thus, these coefficients are significant. This means the price y can be closely related to these variables. Effect size is a number measuring the strength of the relationship between two variables in a population, or a sample-based estimate of that quantity.

We then regard the x variables as three main groups of coefficients that determine the price (at least in theory): host

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.050
Model:                            OLS   Adj. R-squared:                  0.047
Method:                 Least Squares   F-statistic:                     13.40
Date:                Sun, 10 Oct 2021   Prob (F-statistic):           3.23e-23
Time:                        18:09:19   Log-Likelihood:                -15507.
No. Observations:                2543   AIC:                         3.104e+04
Df Residuals:                    2532   BIC:                         3.110e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        116.1989     39.138      2.969      0.003      39.453     192.945
x1            -3.0095     20.245     -0.149      0.882     -42.707      36.688
x2             1.6745      0.456      3.674      0.000       0.781       2.568
x3           -13.0994      3.464     -3.782      0.000     -19.892      -6.307
x4            16.4786      3.005      5.484      0.000      10.586      22.371
x5           -12.5665      4.088     -3.074      0.002     -20.582      -4.551
x6            -5.4678      4.303     -1.271      0.204     -13.906       2.971
x7           -34.0559     48.655     -0.700      0.484    -129.463      61.351
x8           -31.3040    117.148     -0.267      0.789    -261.020     198.412
x9           186.1752     78.452      2.373      0.018      32.339     340.012
x10          611.9750    108.850      5.622      0.000     398.530     825.420
==============================================================================
Omnibus:                     1242.450   Durbin-Watson:                   1.542
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11140.087
Skew:                           2.116   Prob(JB):                         0.00
Kurtosis:                      12.339   Cond. No.                     5.49e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.49e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Fig. 8. OLS

response rate, review scores, and the results of the sentiment analysis. Conducting PCA to see if we indeed get three categories after the dimensionality reduction process. Here, x1=host response rate, x2=review scores, x3=the results of the sentiment analysis. Figure 9

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.003
Model:                            OLS   Adj. R-squared:                  0.002
Method:                 Least Squares   F-statistic:                     2.271
Date:                Mon, 11 Oct 2021   Prob (F-statistic):             0.0784
Time:                        22:02:27   Log-Likelihood:                -12415.
No. Observations:                2034   AIC:                         2.484e+04
Df Residuals:                    2030   BIC:                         2.486e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        166.7920      2.404     69.387      0.000     162.078     171.506
x1            -2.6512      1.212     -2.188      0.029      -5.027      -0.275
x2            -0.5756      1.843     -0.312      0.755      -4.190       3.038
x3             3.3390      2.404      1.389      0.165      -1.376       8.054
==============================================================================
Omnibus:                      757.241   Durbin-Watson:                   2.044
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3174.870
Skew:                           1.774   Prob(JB):                         0.00
Kurtosis:                       7.987   Cond. No.                         1.98
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig. 9. PCA

As doing PCA, the standard errors of coefficients are generally smaller than before, which means the predictors are performs better in the model. However, the R-square reduces a lot from 0.05 to 0.003, which means this model does not do well in predicting the price.

For visualization, we report the percentage of listings that have a compound mean below zero and those that have a compound mean above zero using a bar chart. Figure 10

Then, we create a "correlogram" that shows the correlations between the numerical variables host response rate, review scores rating, review scores accuracy, review scores cleanliness, review scores checkin, review scores communication, positivity mean, negativity mean, positivity simple mean, negativity simple mean. Figure 11, 12.

Create a pairplot that shows the relationship between the three principal components host response rate, review scores, the results of the sentiment analysis: Figure 13

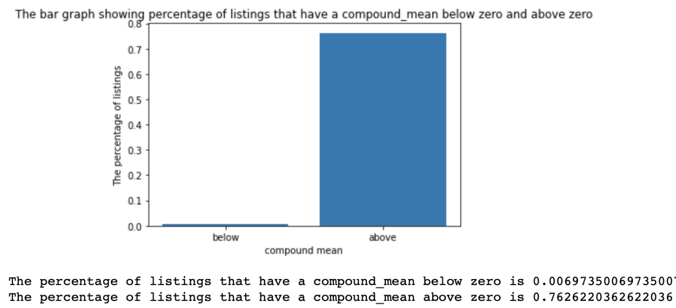


Fig. 10. bar

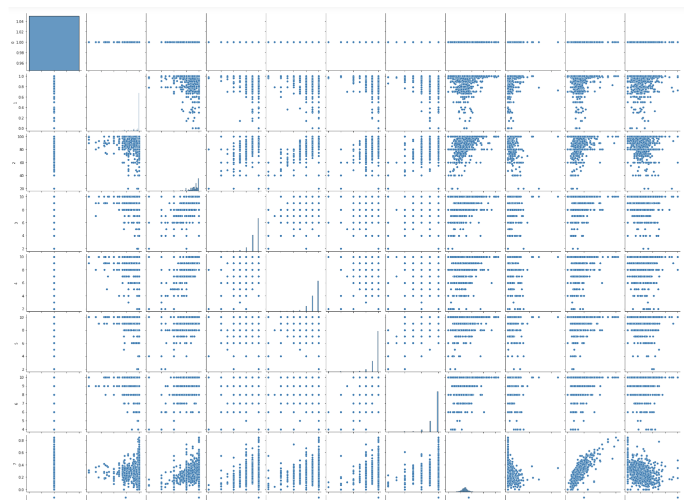

Fig. 11. pairwise1

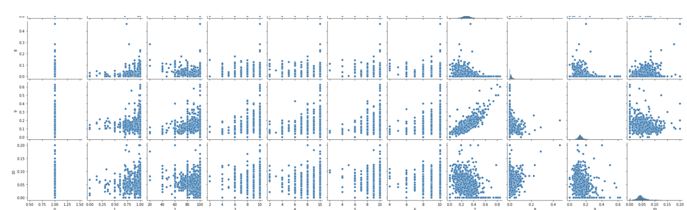

Fig. 12. pairwise2

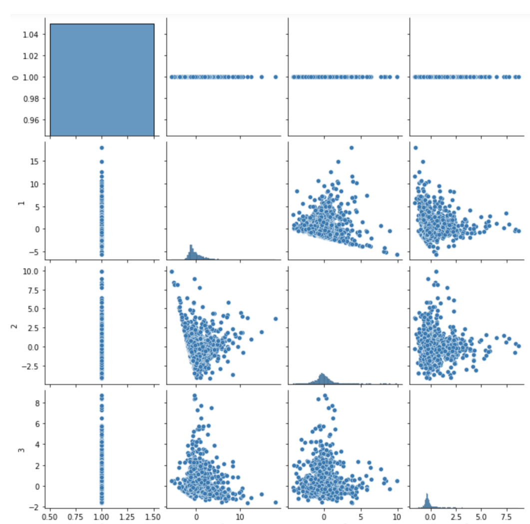

Fig. 13. pairplot

Table for linear regression output (from Q6) that shows the coefficients for different variables, number of observations, degrees of freedom, statistical significance, R2 value is the result in Figure 8 we shown above.

| Variables | value |
|---|---|
| number of observations | 2543 |
| degrees of freedom | 10 |
| F-statistic | 13.4 |
| R^2 value | 0.05 |
| coefficient of constant | 116.1989 |
| coefficient of host_response_rate | -3.0095 |
| coefficient of review_scores_rating | 1.6745 |
| coefficient of review_scores_accuracy | -13.0994 |
| coefficient of review_scores_cleanliness | 16.4786 |
| coefficient of review_scores_checkin | -12.5665 |
| coefficient of review_scores_communication | -5.4678 |
| coefficient of positivity_mean | -34.0559 |
| coefficient of negativity_mean | -31.304 |
| coefficient of positivity_simple_mean | 186.1752 |
| coefficient of negativity_simple_mean | 611.975 |

Fig. 14. table1

The table for linear regression output (from Q7) that shows the coefficients for different variables in three principal components, number of observations, degrees of freedom, statistical significance, R2 value is the result Figure 9 we shown above.

| Variables | value |
|---|---|
| number of observations | 2034 |
| degrees of freedom | 3 |
| F-statistic | 2.271 |
| R^2 value | 0.003 |
| coefficient of constant | 166.792 |
| coefficient of host response rate | -2.6512 |
| coefficient of review scores | -0.5756 |
| coefficient of results of the sentiment analysis | 3.339 |

Fig. 15. table2

In the Airbnb Lab, we find that in data mining, under the minSup 0.1 and 0.2, they have the same top 5 most frequent itemsets (1.0 bathrooms), (Apartment), (1.0 bedrooms), (Apartment, 1.0 bathrooms), (Entire home/apt).

When doing the linear regression model, we find that review scores accuracy, review scores cleanliness, review scores checkin, and negativity simple mean have the p-value 0, which is far less than the significance level 0.05, thus they are the variables seem to explain the price the most.

Some of the ways to improve this analysis are that we can include more variables related in the regression to search for more significant factors to price. And we can make the data more completed without that many missing values.

I did see any typical causal inference problems in this analysis, including selection bias and response bias. In making comments, there always some people who do not want to give feedbacks for some privacy reason. The comments collected are generally from the kinds of people who have strong feeling about the living condition in Airbnb. For those who don't have strong feeling or reactions in living experience, they might won't give the feedback like leave comments spontaneously. Thus, there're selection bias and response bias.