CSC 240 Lab Report: Airbnb

Jiahang Wu and Qihao Yun

***Abstract*— This report is the product of a lab assignment from CSC 240, Fall 2021. The work prepared and analyzed an Airbnb dataset.**

I. Introduction

This report analyzes and explores an Airbnb dataset. A key goal is to prepare the data for appropriate frequent pattern mining and a light exploration on the relationship between listing price and its features. The report finds that sentiment within written reviews contributes most significantly in predicting the price of a listing. The report also confirms the prior understanding of Airbnb as a service which predominantly offers single person housings.

II. Data

A. Dataset

The Airbnb dataset chiefly contains a list of listings (3585x95) and a list of reviews (68275x12) corresponding to part of the listing. This report focuses on a selected range of features only.

From listing dataset, a selection of numerical variables relevant to our interest are processed and generated using a simple script. Some variables (e.g. price) has $ sign which needs to be removed before the statistics can be generated. The descriptive statistics of the listing.csv dataset is included in table 1 on page 2.

B. Preprocessing

To aggregate the reviews dataset to our listing analysis, the report performs two different NLP analyses on the dataset to generate a series of emotional values for the listings. The first approach is to use the nltk package which creates composite scores from the review texts; alternatively, we've also created a simple measurement by counting the percentage of negative or positive words within the reviews.

A series of features are appended to the listing dataset. The variables added are: negativity_mean, neutrality_mean, positivity_mean, compound_mean, positivity_simple_mean, and negativity_simple_mean.

From this analysis, we find that the vast majority of the reviews are positive (2734 positive vs 25 negative, based on composite score of nltk analysis).

Because of the complexity of the dataset, the report mines the frequent item sets of the key features to better describe the dataset as a whole. The analysis is done using the apriori algorithm defined in the mlxtend package as well as a self-defined apriori algorithm.

We've conducted the algorithm with two different thresholds of minimum Support of the itemset. The two parameter generated 70 and 31 itemsets for 0.1 and 0.2 minimum support respectively.

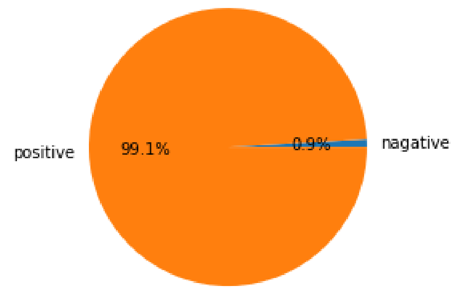

Fig. 1. Positive vs. Negative Reviews

The most frequent (or strongly supported) itemsets are not impacted by the parameter, naturally. We lists the top five itemsets below.

TABLE II
Five Most Supported Itemsets

Support	Itemset	Cardinality
0.767364	1 Bathroom	1
0.728591	Apartment	1
0.663598	1 Bed	1
0.597768	Apartment, 1 Bathroom	2
0.593305	Entire Home OR Apartment	1

It is unsurprising that the most frequent itemsets for the top 15 itemsets (top 5 shown above) more or less describes the same type of listing: single-person independent housing. This aligns with our understanding of Airbnb as a service, which provides predominantly single person temporary housing. The overall R2 value is 0.706.

TABLE I

DESCRIPTIVE STATISTICS OF VARIABLES FROM LISTING DATASET

Variables	Minimum	Maximum	Mean	Median	Variance	Std. Deviation
host_response_rate	0.0	100.0	94.99	100.0	156.69	12.52
host_acceptance_rate	0.0	100.0	84.17	94.0	474.34	21.78
host_listings_count	0	749	58.9	2.0	29281.94	171.12
host_total_listings_count	0	749	58.9	2.0	29281.94	171.12
accommodates	1	16	3.04	2.0	3.16	1.78
bathrooms	0.0	6.0	1.22	1.0	0.25	0.5
bedrooms	0.0	5.0	1.26	1.0	0.57	0.75
beds	0.0	16.0	1.61	1.0	1.02	1.01
price	10.0	4000.0	173.93	150.0	22002.18	148.33
weekly_price	80.0	5000.0	922.39	750.0	432729.54	657.82
monthly_price	500.0	40000.0	3692.1	2925.0	8409789.65	2899.96
security_deposit	95.0	4500.0	324.7	250.0	108157.5	328.87
cleaning_fee	5.0	300.0	68.38	50.0	2631.47	51.3
guests_included	0	14	1.43	1.0	1.12	1.06
extra_people	0.0	200.0	10.89	0.0	366.25	19.14
minimum_nights	1	300	3.17	2.0	78.75	8.87
maximum_nights	1	99999999	28725.84	1125.0	2789354050349.61	1670135.94
availability_30	0	30	8.65	4.0	108.9	10.44
availability_90	0	90	38.56	37.0	1099.47	33.16
availability_365	0	365	179.35	179.0	20202.69	142.14
number_of_reviews	0	404	19.04	5.0	1265.34	35.57
review_scores_rating	20.0	100.0	91.92	94.0	90.85	9.53
review_scores_accuracy	2.0	10.0	9.43	10.0	0.87	0.93
review_scores_cleanliness	2.0	10.0	9.26	10.0	1.37	1.17
review_scores_checkin	2.0	10.0	9.65	10.0	0.58	0.76
review_scores_communication	4.0	10.0	9.65	10.0	0.54	0.74
review_scores_value	2.0	10.0	9.17	9.0	1.02	1.01
reviews_per_month	0.01	19.15	1.97	1.17	4.5	2.12

III. RESULTS

Building on the preparation above, the report performed a multivariate linear regression between a series of variables and the price of the listings. See the summary below for details of the regression model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):          0.706
Model:                            OLS   Adj. R-squared (uncentered):     0.705
Method:                 Least Squares   F-statistic:                     607.5
Date:                Sun, 10 Oct 2021   Prob (F-statistic):               0.00
Time:                        12:44:49   Log-Likelihood:                -15534.
No. Observations:                2543   AIC:                         3.109e+04
Df Residuals:                    2533   BIC:                         3.115e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
              coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          0.2929      0.184      1.594      0.111      -0.067       0.653
x2          1.7684      0.470      3.765      0.000       0.847       2.689
x3        -11.0310      3.443     -3.204      0.001     -17.782      -4.279
x4         17.3160      3.037      5.702      0.000      11.361      23.271
x5         -9.1260      4.005     -2.279      0.023     -16.980      -1.272
x6         -0.8331      4.187     -0.199      0.842      -9.044       7.378
x7        107.6440    126.528      0.851      0.395    -140.465     355.753
x8         26.1721     35.352      0.740      0.459     -43.149      95.494
x9         11.4848     93.372      0.123      0.902    -171.609     194.578
x10      1227.0619    451.050      2.720      0.007     342.597    2111.527
==============================================================================
Omnibus:                     1200.516   Durbin-Watson:                   1.518
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9845.002
Skew:                           2.059   Prob(JB):                         0.00
Kurtosis:                      11.715   Cond. No.                     2.83e+04
==============================================================================
```

Fig. 2. Summary of the Regression

TABLE III

PREDICATIVE VARIABLES COEFFICIENTS OF LINEAR REGRESSION OF PRICE (Y)

Variable	Coefficient
host response rate	0.2929
review rating	1.7684
description accuracy	-11.0310
cleanliness	17.3160
checkin	-9.1260
communication	-0.8331
positivity mean	107.6440
negativity mean	26.1721
positivity simple mean	11.4848
negativity simple mean	1227.0619

Negativity (composite and simple) appear to be the most significant variables. We expected cleanliness to be one of the significant variable, but its actual effect is weak compared to negativity in the written reviews. This is somewhat surprising but perhaps not entirely odd as most listings - as we discovered - are for single person and hence relatively low priced. We also understand as a prior that Airbnb users are generally seeking for price competitiveness from Airbnb (otherwise they would have chosen to stay at hotels instead). It seems that high price and negative reviews are very strongly correlated, and in this case, our simple approach for review negativity appears to be more relevant than the one provided in nltk package.

The initial pairwise plot between the 10 variables is good

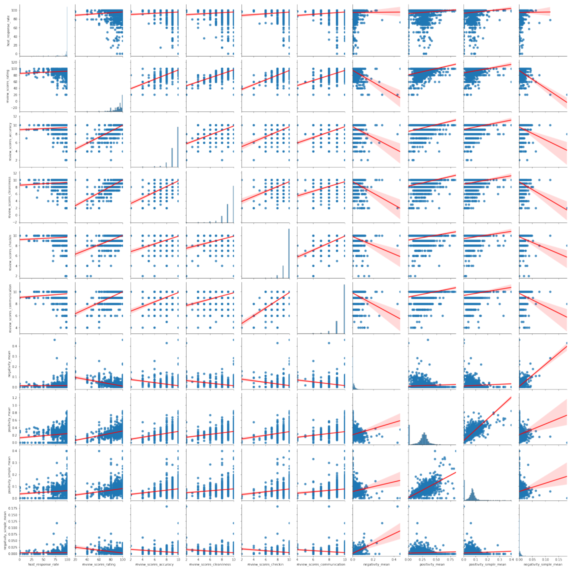

Fig. 3. Unreadable Initial Plot

but needs further refinement. For the most part, the variables Checkin, Cleaniness, Communication, response rate, and Accuracy behave similarly and are mostly discretely leveled. (See Figure 3) They are moderately correlated between each other and the very strongly correlated with the overall score. Thus, they are removed in the second pair plot of the variables for readability. See the plot below.

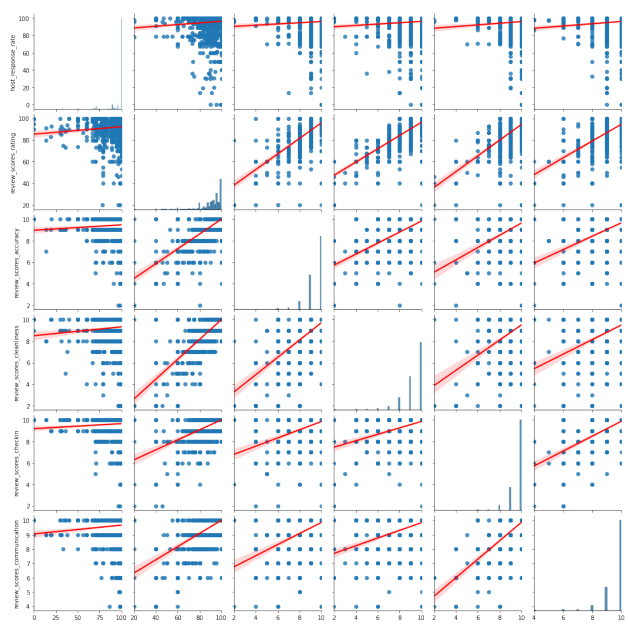

Fig. 4. Pairwise Plot of Selected x

The new plot's variables are in the order (left to right and top to down) of 'host response rate', 'review scores rating', 'negativity mean', 'positivity mean', 'positivity simple mean', 'negativity simple mean'.

As the trendline indicate, this verifies the observation that positivity is much less significant than negativity in review

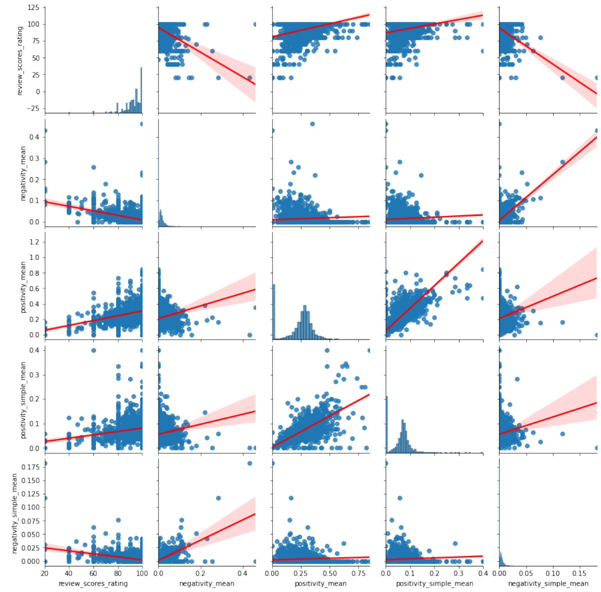

Fig. 5. New Pairwise Plot of Variables of X

for the overall score. Practically, simple_negativity appears to have the strongest correlation with overall score.

This observation also indicates there may be three principal components contributing to variation in price. We propose that the three components may be: overall score (which also represents the 4 ratings to some extent), response rate, and the results of the sentiment analysis. PCA is conducted on the standardized dataset to investigate further.

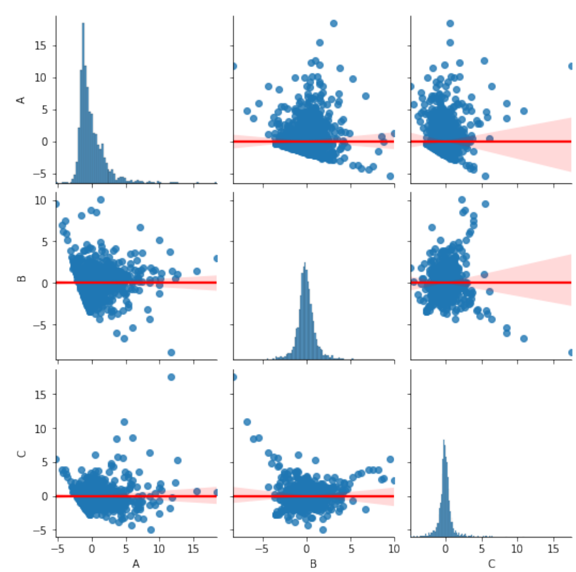

Fig. 6. Pairwise Plot of Pincipal Components

As expected, the three principal components discovered by the sklearn PCA model are practically independent from each other. The analysis also finds that one compoennt contributes significantly to the dataset's variation. See table:

TABLE IV

EXPLAINED VARIATION OF PCA

Component	Explained Variation
A	0.604
B	0.231
C	0.165

The transformed dataset (with principal components as X) are again used for a linear regression model for price (y).

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                  0.001
Model:                            OLS   Adj. R-squared (uncentered):            -0.001
Method:                 Least Squares   F-statistic:                            0.6343
Date:                Sun, 10 Oct 2021   Prob (F-statistic):                      0.593
Time:                        20:32:53   Log-Likelihood:                        -13703.
No. Observations:                2041   AIC:                                 2.741e+04
Df Residuals:                    2038   BIC:                                 2.743e+04
Df Model:                           3
Covariance Type:            nonrobust
=======================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
x1            -3.0107      2.223     -1.354      0.176      -7.371       1.350
x2             0.0387      3.596      0.011      0.991      -7.013       7.090
x3             1.1194      4.254      0.263      0.792      -7.223       9.462
=======================================================================================
Omnibus:                     1010.022   Durbin-Watson:                       0.584
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                 9597.444
Skew:                           2.116   Prob(JB):                             0.00
Kurtosis:                      12.744   Cond. No.                             1.91
=======================================================================================
```

Fig. 7. Summary of the Regression

TABLE V

PCA COEFFICIENT IN REGREESION MODEL

Component	Coefficient
A	-3.0107
B	0.0387
C	1.1194

This largely aligns with our previous correlation analysis and the un-tranformed regression model. The observation of the principal components' contribution certainly confirms our previous model's results: namely that the sentiment analysis indeed contribute the most to predicting price.

A. Discussion

This report first finds that the the vast majority of the Airbnb listings are single person housing that generally receive positive (whenever there are) reviews. The report finds that the sentiment within written review text are the most significant predictor of price (i.e. people are less satisfied when the listing is more expensive), over than other previously assumed features of the listings. Namely, simple negativity (relative amount of negative word in a review) contribute most significantly in our linear regression models.

This analysis remains to be improved in a number of ways. Firstly, the analyzed dataset is quite small, with less than 3000 listings. The report would certainly benefit from having more data in a specific area or more data overall to make basis for argumentative claims or conclusions. There are some features of the dataset that appear poorly measured or generated; fortunately, this report has not focused on those attributes of the listing, but if their quality is sufficient, the report would be able to include them into the regression model and analysis too. This is a case for potential omitted variable bias.

The analysis we conducted is far from a real causal inference. We've discovered the strong relationship between linguistic sentiment (namely, number of explicit negative words) with the price of a listing, but it remains very inconclusive to say whether one causes the other. The selected feartures used in regression are all, from a non-data perspective, inter-correlated. A person may naturally assume that these ratings measure the same thing with certain simultaneity. While our analysis finds otherwise, we may not be free to claim so for the general audience, so more investigation may be needed to confirm this finding.