Airbnb Project Lab Report*

Keitaro Ogawa[1]

Abstract— **The report aims to develop a model that predicts the price (Y) column based on the numeric variables and the sentiment scores that were calculated. This model gives us a deeper understanding of which variable had greater correlations or significant coefficients to the price of Airbnb than the others. Additionally, this report provides visualizations of various regression models that shows the relationships of components in the data-sets.**

I. INTRODUCTION

A. *Summary of Expectation*

One expectation for the Airbnb lab was the Data Preparation. I would likely to clean and prepare the data by handling missing values, converting data types, and merging the three datasets (calendar.csv, listings.csv, and reviews.csv) using the common identifier, which is the listing id.

Another expectation was the Exploratory Data Analysis (EDA). I would conduct EDA to understand the data better by performing summarization of statistics, distributions of numeric variables, and understanding the sentiment scores. I would explore correlations between variables, identify outliers, and detect any patterns or trends in the data.

Last core expectation was the Model Building and Model Evaluation. I would develop a predictive model to estimate the price of Airbnb listings. Common modeling techniques for this type of regression problem include linear regression, decision trees, random forests, or more advanced methods like gradient boosting. For Model Evaluation, I would evaluate the performance of the predictive model. Common metrics for regression models include Mean Squared Error (MSE), and R-squared (R2).

B. *Summary of the Procedure*

The lab consisted of six parts: Descriptive Analysis, Sentiment Analysis, Data Mining, Linear Regression, and Visualization.

The lab began by performing the descriptive statistics analysis of the calendar, listings, and reviews dataframe, where I showed the minimum, maximum, mean, median, variance, and standard deviation for 28 variables in the listings dataframe.

Next, two sentiment analyses were performed. One was running the analysis on each cell in the comments column and add four new columns to the reviews data set as a result: negativity, neutrality, positivity, compound. The other was writing a loop structure that iterates over each comment in the comments column of the reviews data set, and count the total number of words, total number of positive words, and total number of negative words.

Consequently, Apriori algorithm was applied to calculate the frequent item sets in the listings data set. I looked at property type, room type, accommodates, bathrooms, and bedrooms variables to come up with the top 5 most frequent and top 5 least frequent item sets for MinSupport = 0.1 and 0.2.

Afterwards, I developed a linear regression model using OLS to report the coefficients in the ten variables. These variables came from the numeric columns and sentiment scores. I analyzed which variables had significant coefficients to the price column (Y). Further, I applied PCA function to the whole data set by setting n components to 3 and fit linear regression.

Lastly, I plotted visualizations on all the findings, such as a correlogram that shows the correlations between the numerical variables, a pair plot that shows the relationship between the three principal components, and a table for the linear regression output and PCA output.

C. *Significant Findings*

Some of the key findings are that linear regression using OSL explains a greater proportion of the variance in the price variable, indicating a better fit to the data than PCA model (discussed more later). Based on the coefficients and their associated p-values in linear regression model, "review scores cleanliness" and "review scores rating" have the most significant impact on the price of Airbnb listings.

II. DATA

A. *The Calender Data-frame*

This data set contains 4 columns (listing id, date, available and price). Basically, it indicates the date and the id of the Airbnb and mentions whether each room is available or not. If it is available, it shows the price.

B. *The Listings Data-frame*

This data set contains 95 columns and it contains the overview and basic information of the Airbnb. For example, there are columns, such as name, summary, space, description, and experiences offered. This information is what customers see on the homepage when they want to book specific Airbnbs.

C. *The Reviews Data-frame*

This dataset has 6 columns (listing id, id, date, reviewer id, reviewer name and reviews). It lists every comments from the reviewers on every property along with the date they stayed at and their names. This data frame is a great indicator for those who want to hear personal opinions of the place customers had stayed at.

Last and most importantly, all three data-frames have listing id as a common column, so it is possible to merge those data sets based on the ids.

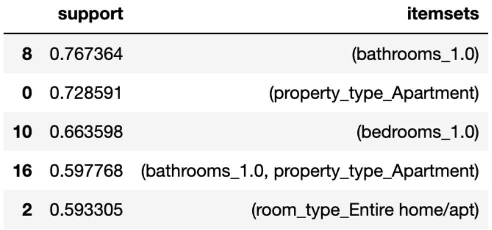
D. Number of Observations and Descriptive Statistics

According to the table for the linear regression output, the number of observations performed was 3585. And descriptive statistics (minimum, maximum, mean, median, variance, and standard deviation) of 28 variables showed host listings count and host total listings count had the exact same statistics. host response rate and host acceptance rate followed similar statistics, but host acceptance rate had very large variance.

III. Results

A. Apriori Algorithm on the Listing Data Frame

As shown in Fig1, the most frequent item set is (bathrooms 1.0) with a support of approximately 76.74 percent. This suggests that a significant proportion of listings have 1 bathroom. The second most frequent item set is (property type Apartment) with a support of approximately 72.86 percent. This indicates that a large number of listings are apartments. The same logic follows for the third, fourth, and fifth item sets.

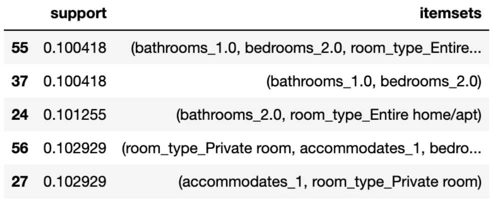


	support	itemsets
8	0.767364	(bathrooms_1.0)
0	0.728591	(property_type_Apartment)
10	0.663598	(bedrooms_1.0)
16	0.597768	(bathrooms_1.0, property_type_Apartment)
2	0.593305	(room_type_Entire home/apt)

Fig. 1. Top 5 Most Frequent Item Sets (MinSup=0.1

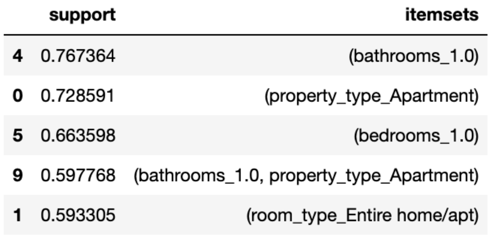
As shown in Fig 2, The least frequent item set has a support of approximately 10.04 percent. It consists of multiple attributes, including (bathrooms 1.0, bedrooms 2.0, room type Entire...). This suggests that listings with these specific combinations of attributes are relatively rare. Another infrequent item set is (bathrooms 1.0, bedrooms 2.0) with a support of approximately 10.04 percent. Listings with 1 bathroom and 2 bedrooms are not very common. The item set (bathrooms 2.0, room type Entire home/apt) has a support of approximately 10.13 percent, indicating that listings with 2 bathrooms and an entire home or apartment rental type are less common.

As shown in Fig3, The most frequent itemset is (bathrooms 1.0) with a support of approximately 76.74 percent. This indicates that a significant proportion of listings have 1 bathroom. The second most frequent itemset is (property type Apartment) with a support of approximately 72.86 percent. This suggests that many listings are apartments.



	support	itemsets
55	0.100418	(bathrooms_1.0, bedrooms_2.0, room_type_Entire...
37	0.100418	(bathrooms_1.0, bedrooms_2.0)
24	0.101255	(bathrooms_2.0, room_type_Entire home/apt)
56	0.102929	(room_type_Private room, accommodates_1, bedro...
27	0.102929	(accommodates_1, room_type_Private room)

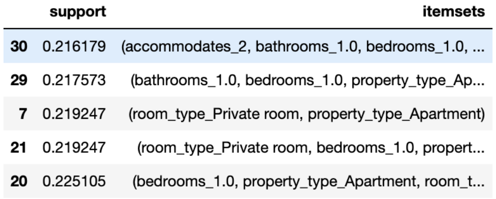
Fig. 2. Top 5 Least Frequent Item Sets (MinSup=0.1



	support	itemsets
4	0.767364	(bathrooms_1.0)
0	0.728591	(property_type_Apartment)
5	0.663598	(bedrooms_1.0)
9	0.597768	(bathrooms_1.0, property_type_Apartment)
1	0.593305	(room_type_Entire home/apt)

Fig. 3. Top 5 Most Frequent Item Sets (MinSup=0.2

As shwon in Fig4, The least frequent itemset has a support of approximately 21.62 percent. It consists of multiple attributes, including (accommodates 2, bathrooms 1.0, bedrooms 1.0, ...). This suggests that listings with these specific combinations of attributes are relatively rare. Another infrequent itemset is (bathrooms 1.0, bedrooms 1.0, property type Ap...) with a support of approximately 21.76 percent. Listings with 1 bathroom, 1 bedroom, and the property type 'Apartment' are less common.



	support	itemsets
30	0.216179	(accommodates_2, bathrooms_1.0, bedrooms_1.0, ...
29	0.217573	(bathrooms_1.0, bedrooms_1.0, property_type_Ap...
7	0.219247	(room_type_Private room, property_type_Apartment)
21	0.219247	(room_type_Private room, bedrooms_1.0, propert...
20	0.225105	(bedrooms_1.0, property_type_Apartment, room_t...

Fig. 4. Top 5 Least Frequent Item Sets (MinSup=0.2

These differences in the item sets indicate that for minSupport = 0.1, when using a lower minimum support threshold (0.1), the frequent itemsets include a wider range of attribute combinations. This means that a larger set of combinations is considered frequent, resulting in more itemsets in the results. For minSupport = 0.2, when using a higher minimum support threshold (0.2), the number of frequent itemsets is reduced. The results include only the most prevalent combinations, making the analysis more focused.

B. Linear Regression R2 Values

After using OLS to fit the linear regression model based on the given variables, the R2 values came out to be 0.0169. This means that only about 1.69 percent of the variance in the price can be explained by the model. In other words, the model does not explain much of the variability in the price, and the predictors used in the model have limited explanatory power.

Further, the R2 values for PCA model was 0.0029. The linear regression model has a higher R² value (0.0169) compared to the PCA model (0.0029). This means that the linear regression model explains a greater proportion of the variance in the price variable, indicating a better fit to the data. Therefore, based on the R values values, the linear regression model did a better job in explaining the relationship between the explanatory variables and the price variable compared to the PCA-based model. Fig 5 visualizes the first three principal components in a pair plot. The plots show less variance in the model as plots are concentrated in left corner.

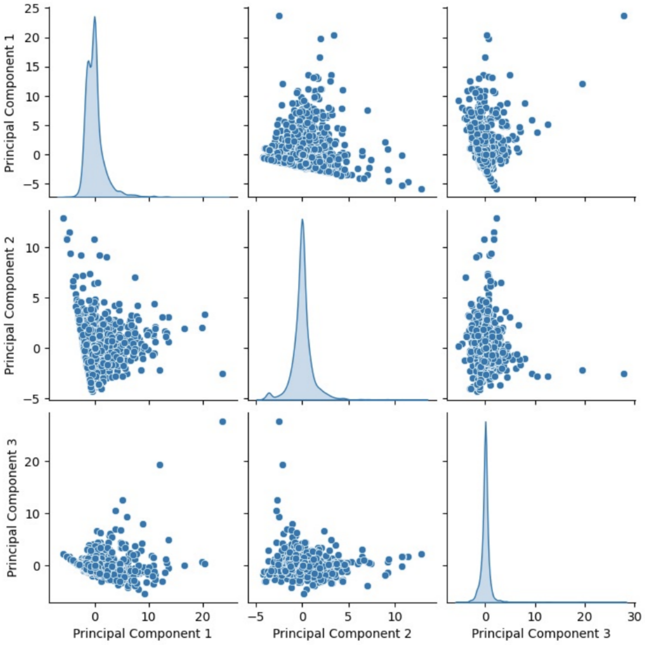


Fig. 5. Relations of Three Principal Components

C. Keys Variables to the Price

Based on the coefficients and their associated p-values in linear regression model, it appears that "review scores cleanliness" and "review scores rating" have the most significant impact on the price of Airbnb listings in this model. These variables are statistically significant, and their coefficients indicate the direction and magnitude of their influence on price.

D. Main Findings and Improvement on Analysis

The main finding was that both of R2 values from the Linear Regression using OLS and PCA did not produce high scores. To improve the model's accuracy, implementing a correlation matrix can enhance it because it helps us identify relationships between variables, including the principal components. We can use this information to guide feature selection and determine which principal components are most strongly correlated with the target variable (e.g., price in our case).

Another approach is applying regularization. Techniques such as L1 (Lasso) or L2 (Ridge) regularization when building regression models helps in preventing overfitting by penalizing large coefficients. This can be especially helpful when we have many features, including principal components. This is effective for Airbnb lab especially because we deal with many variables and we tend to pick ones without knowing how correlated each variable is to the Y-value.

E. Potential Causal Inference

In the analysis of Airbnb data, there are several potential causal inference issues that can arise, including selection bias, simultaneity, and omitted variable bias.

In selection bias, we need to consider inclusion and user Bias. If the datas et includes only a specific subset of Airbnb listings in Boston, the analysis may not generalize well to the entire population of Airbnb listings in the city. Also, if we're using data from reviews, there may be selection bias related to the guests who leave reviews. They may have had particularly positive or negative experiences, and this could skew sentiment analysis or any inferences about the quality of listings.

In simultaneity, there is a possibility of reverse causality. Simultaneity occurs when the cause and effect variables influence each other simultaneously. In the context of Airbnb pricing, reverse causality might occur if the demand for a listing influences its price, but the price also influences demand. For instance, higher demand might lead to higher prices, but higher prices might also reduce demand. Failure to account for this simultaneity could lead to incorrect causal inferences.

In Omitted Variable Bias, we need to note the confounding variables. Omitted variable bias arises when important variables are not included in our model, leading to incorrect inferences about causal relationships. In the context of Airbnb pricing, omitting key factors like local events, economic conditions, or specific property features (e.g., pool, proximity to attractions) could lead to omitted variable bias. These unobserved variables may affect both the price and the predictors in your model, making it difficult to isolate the true causal relationship.