Airbnb Lab

Jacob Freiheit

October 2023

1 Abstract

The goal of this lab is to track the relationship between the price to stay at an Airbnb location and a set of numerical explanatory variables using Linear Regression and Principal Component Analysis. In addition to this, sentiment analysis is performed on comments from guests to create measures for both positive and negative sentimentality, also included in the Linear Regression and PCA analyses.

2 Introduction

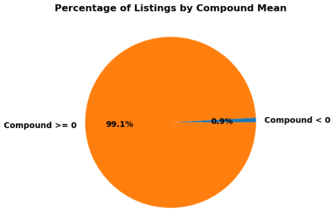
A summary of what you expected and did, and two-three of your most significant findings After performing sentiment analysis and adding sentiment analysis columns, both a package-included and a manually-written apriori algorithm were applied to the dataset to find the most frequent itemsets for property type, room type, accommodates, bathrooms, bedrooms with a minSup of 0.1 and 0.2. It was expected that the highest frequency itemsets would be of single value, which ended up being correct, with bathrooms 1.0 and property type apartment being the top 2. It was interesting however, that the bathrooms 1.0, property type apartment combined itemset was in the top 5 when it came to support, despite not being singular.

Linear Regression and PCA were then used on a set of explanatory variables with the aim of determining their relation to the price of the stay. To work around null values in the dataset, a mean-based imputing method was used. It was expected that the price would be strongly related to the numerical rating average of the location, as well as the positive sentiment of the guests when they stayed there. The Linear Regression R-squared was 0.016, showcasing that while price may be affected by some values, there are others for which it changes very little. While it was expected that the coefficient for positive sentiment would be high, the degree to which positivity (21), as well as cleanliness score (17) changed the price relative to numerical rating (1) wasn't foreseen.

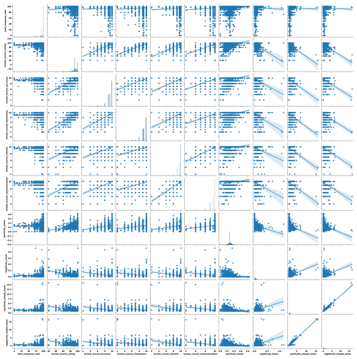
Finally, a PCA function of 3 components was run on the dataset. The PCA function earned a score (R-squared) of 0.003, indicating that very little of the original variance from the original data was preserved by the selected principal components. The values are both low, but it's in this way that it can be determined that Linear Regression was a better fit for this dataset.

3 Data

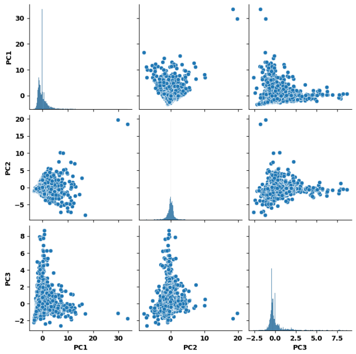
The dataset as a whole was a complete accounting for the information of around 3500 different Airbnb locations.



This pie chart shows the percentage of listings that are above and below the compound mean of 0. As can be seen, the overwhelming majority of listings (99.1 percent) are above that threshold.



This correlogram shows correlations and trend lines for the numeric values of the explanatory variables and the price of the location.



Finally, the Pairplot shown above tracks the correlation for the PCA Linear Regression output to the price of the stay at the location.

4 Results

When finding the most and least frequent itemsets, it was found that both the itemsets at the top as well as their support values remained the same regardless of whether or not the apriori alogrithm was written manually. The top 5 most and least frequent itemsets can be found in the lab pdf included with this report

By far the most significant coefficients are the ones for positivity, cleanliness score, and overall numeric rating. While two of these values, positivity and numeric rating, were expected, cleanliness score, as well as the degree of significance, was not accounted for.

The standard error values found for linear regression are much higher than the ones for PCA. This makes sense, given that individual explanatory variables in the dataset had hundreds of NaN values that needed to be replaced to make the algorithms work.

Given that the R-squared for Linear Regression (0.016) is several degrees higher than the R-squared for PCA (0.003), it can be seen that Linear

Regression showcases more consistent changes for the explanatory variables relative to PCA, albeit with higher standard error values. It's for this reason that Linear Regression does a better job.

The biggest finding that came from this lab was just how much the cleanliness score served as justification for higher pricing. It clearly has been judged as among the most important individual values when it comes to how the owners price their locations.

The variable that most explains the price of the locations however, is the positive sentiment of the comments left by the guests. Owners see the comments, and when they look good, they price higher accordingly.

It would be interesting to test more Regression Models to see what other relations could be revealed. Additionally, the use of a dataset that has less null values would greatly decrease the amount of standard error present in the algorithms used.

There is a large amount of omitted variable bias that exists in this analysis, given that when doing Linear Regression and PCA, hundreds of values needed to be replaced in order for the algorithms to work. Additionally, impressions of the location are going to be affected by the price, as people tend to expect more from a more expensive location. As a result, there is a high chance of simultaneity as well.