Airbnb Lab

Matthew Boundy and Jasper Lemberg

***Abstract*— Using the Boston Airbnb Open Data dataset from Airbnb, we were able to analyze listing data and uncover insights related to the data's various patterns and associations. Using this data, we performed a variety of tasks, including running a sentiment analysis, discovering frequent patterns, and fitting multiple ordinary least squares linear regression models.**

I. INTRODUCTION

This lab used Python and the Boston Airbnb Open Data dataset in order to uncover key insights and information about various listings and their reviews. Airbnb accumulated and published the data in 2019. The dataset has a commercial rather than academic focus and is free to download on Kaggle.

Throughout the lab, a number of findings stuck out as especially interesting. First, we used the Apriori algorithm to uncover frequent itemsets. However, the most frequent itemsets only consisted of one item. In fact, the non-one-item-itemset with the largest support was "property_type_Apartment, bathrooms_1.0" (0.597768). Also, while adding principal components to the linear regression model lowered coefficient standard errors across the board, the R-squared value went down to nearly 0, so PCA was not as effective as it seemed.

II. DATA

The dataset actually consists of three separate dataframes: calendar.csv, listings.csv, and reviews.csv. All three dataframes have the "listing_id" feature, which is an integer that displays each listing's unique identification number. "calendar.csv" also includes the following features:

- "date":, which just lists a date that the listing was available or unavailable.
- "available", which is a boolean value that is true if the listing was available on the given date and false otherwise.
- "price", which is a float for the price of the listing per night.

"reviews.csv" includes the following features, excluding the aforementioned "listing_id":

- "id":, which is an integer representing the review's ID.
- "date":, which just lists a date that the listing was available or unavailable.
- "reviewer_id", which is an integer representing the reviewer's ID.
- "reviewer_name", which is a string consisting of the reviewer's or reviewers' name(s).
- "comments", which is a string consisting of the review itself.

"listings.csv" includes ninety_five total features, including the aforementioned "listing_id". These features cover everything from host information to review scores to comment sentiment analyses.

A. *Q1*

The table shown in Fig. 1 features the results of an exploratory data analysis of some of the features found in the listings dataframe. In general, there are some strange variables in this dataset. host_listings_count and host_listings_total_count both have a mean of 58.9, but have a median of 2 which suggests that the data is very skewed to the left. The variance for each of these variables is also 29281.9 which is extremely high for the data having a maximum of 749. Another strange variable is maximum-nights which has a max of 99999999 and a mean of 28725.8 and a median of 1125. The variance is also ridiculous at 2789354050349.2, which makes this variable hard to work with.

Variables	Minimum	Maximum	Mean	Median	Variance	Std. Deviation
host_acceptance_rate	0.0	1.0	0.8417308927424536	0.94	0.047433591547741134	0.21779254245208016
host_listings_count	0	749	58.9023709902371	2.0	29281.939126631325	171.11966317940005
host_total_listings_count	0	749	58.9023709902371	2.0	29281.939126631325	171.11966317940005
accommodates	1	16	3.0412831241283125	2.0	3.164589870990262	1.7789294170905887
bathrooms	0.0	6.0	1.221646597591711	1.0	0.2514892767524199	0.5014870653889488
bedrooms	0.0	5.0	1.255944055944056	1.0	0.5670988217155274	0.7530596402115356
beds	0.0	16.0	1.6090604026845639	1.0	1.0236269770497377	1.011744521630702
price	10.0	4000.0	173.9258019525802	150.0	22002.180877042243	148.33132129473614
weekly_price	80.0	5000.0	922.3923766816143	750.0	432729.5428374428	657.8218169363515
monthly_price	500.0	40000.0	3692.097972972973	2925.0	8409789.653299155	2899.9637331006666
security_deposit	95.0	4500.0	324.6982116244411	250.0	108157.49945988657	328.87307499989504
cleaning_fee	5.0	300.0	68.38014527845036	50.0	2631.4678656228098	51.29783490190214
guests_included	0	14	1.4298465829846583	1.0	1.1167986650727677	1.0567869534928824
extra_people	0.0	200.0	10.886192468619248	0.0	366.25434333906156	19.137772684904103
minimum_nights	1	300	3.1712691771269177	2.0	78.75023457735509	8.87413289157623
maximum_nights	1	99999999	28725.83682008368	1125.0	2789354050349.6133	1670135.937685796
availability_30	0	30	8.649930264993026	4.0	108.89611118375372	10.435329950880984
availability_90	0	90	38.5581589958159	37.0	1099.4710166990667	33.15827222125825
availability_365	0	365	179.34644351464436	179.0	20202.69355947414	142.1361796288128
number_of_reviews	0	404	19.04463040446304	5.0	1265.3428736426627	35.5716582919979
review_scores_rating	20.0	100.0	91.91666666666667	94.0	90.85303139660876	9.531685653472252
review_scores_accuracy	2.0	10.0	9.43157132512672	10.0	0.8683690620966888	0.9318632207017771
review_scores_cleanliness	2.0	10.0	9.25804119985544	10.0	1.3665070800083625	1.1689769373295449
review_scores_checkin	2.0	10.0	9.64629294755877	10.0	0.581792511835173	0.7627532443950488
review_scores_communication	4.0	10.0	9.646548608601373	10.0	0.5409705492451865	0.7355070014929745
review_scores_value	2.0	10.0	9.16823444283647	9.0	1.0223564908002405	1.0111164575854952
reviews_per_month	0.01	19.15	1.970908448214916	1.17	4.496780892959038	2.120561457010628

Fig. 1. Exploratory data analysis for certain features in listings dataframe

III. RESULTS

A. *Q4*

When we set the minimum support to 0.1, the Apriori algorithm uncovers seventy frequent itemsets. The five itemsets with the highest support values are "bathrooms_1.0"

(0.767364), "property_type_Apartment" (0.728591), "bedrooms_1.0" (0.663598), "property_type_Apartment, bathrooms_1.0" (0.597768), and "room_type_Entire home/apt" (0.593305) respectively. The five least frequent itemsets are "bathrooms_1.0, bedrooms_2.0" (0.100418), "room_type_Entire home/apt, bedrooms_2.0, bathrooms_1.0" (0.100418), "bathrooms_2.0, room_type_Entire home/apt" (0.101255), "accommodates_1, room_type_Private room" (0.102929), and "room_type_Private room, bedrooms_1.0, accommodates_1" (0.102929) respectively. This is further revealed in figure 2.

	support	itemsets
0	0.728591	(property_type_Apartment)
1	0.156764	(property_type_House)
2	0.593305	(room_type_Entire home/apt)
3	0.384379	(room_type_Private room)
4	0.122455	(accommodates_1)
...	...	...
65	0.136402	(property_type_Apartment, accommodates_2, bedr...
66	0.194142	(property_type_Apartment, bedrooms_1.0, bathro...
67	0.216179	(property_type_Apartment, accommodates_2, bath...
68	0.190795	(accommodates_2, bedrooms_1.0, bathrooms_1.0, ...
69	0.123013	(bathrooms_1.0, property_type_Apartment, bedro...

Fig. 2. Most and least frequent itemsets found using Apriori algorithm with a minimum support of 0.1

When we set the minimum support to 0.2, the Apriori algorithm uncovers twenty-nine frequent itemsets. The five itemsets with the highest support values remain the same. The five least frequent itemsets are now "property_type_Apartment, accommodates_2, bathrooms_1.0, bedrooms_1.0" (0.216179), "property_type_Apartment, bathrooms_1.0, bedrooms_1.0, room_type_Entire home/apt" (0.217573), "property_type_Apartment, room_type_Private room" (0.219247), "property_type_Apartment, bedrooms_1.0, room_type_Private room" (0.219247), and _Entire home/apt" (0.225105) respectively. This is further revealed in figure 3.

Ultimately, these values make sense. Many listings on Airbnb are smaller apartments, especially in a larger urban area like Boston. As such, it makes sense that the most common itemsets are apartments, listings with only one bedroom, listings with only one bathroom, or a combination of those three. Also, considering the size of these apartments, it would be more likely for a renter to rent the entire small apartment rather than a section of it.

B. Q5

Unfortunately, the Apriori algorithm we coded did not work for anything larger than a 1-item itemset. However, these values are very similar to the values generated by the mlxtend package's Apriori algorithm when min_sup = 0.1, as shown in figure 4.

The same could be said for when min_sup = 0.2, as shown in figure 5

	support	itemsets
0	0.728591	(property_type_Apartment)
1	0.593305	(room_type_Entire home/apt)
2	0.384379	(room_type_Private room)
3	0.413668	(accommodates_2)
4	0.767364	(bathrooms_1.0)
5	0.663598	(bedrooms_1.0)
6	0.492050	(property_type_Apartment, room_type_Entire hom...
7	0.219247	(property_type_Apartment, room_type_Private room)
8	0.290934	(property_type_Apartment, accommodates_2)
9	0.597768	(property_type_Apartment, bathrooms_1.0)
10	0.461646	(property_type_Apartment, bedrooms_1.0)
11	0.446583	(room_type_Entire home/apt, bathrooms_1.0)
12	0.256904	(bedrooms_1.0, room_type_Entire home/apt)
13	0.238494	(accommodates_2, room_type_Private room)
14	0.301813	(bathrooms_1.0, room_type_Private room)
15	0.384379	(bedrooms_1.0, room_type_Private room)
16	0.358996	(accommodates_2, bathrooms_1.0)
17	0.350907	(accommodates_2, bedrooms_1.0)
18	0.567922	(bedrooms_1.0, bathrooms_1.0)
19	0.387727	(property_type_Apartment, room_type_Entire hom...
20	0.225105	(property_type_Apartment, bedrooms_1.0, room_t...
21	0.219247	(property_type_Apartment, bedrooms_1.0, room_t...
22	0.272245	(property_type_Apartment, accommodates_2, bath...
23	0.233752	(property_type_Apartment, accommodates_2, bedr...
24	0.427615	(property_type_Apartment, bedrooms_1.0, bathro...
25	0.247141	(bedrooms_1.0, bathrooms_1.0, room_type_Entire...
26	0.238494	(accommodates_2, bedrooms_1.0, room_type_Priva...
27	0.301813	(bedrooms_1.0, bathrooms_1.0, room_type_Privat...
28	0.297908	(accommodates_2, bedrooms_1.0, bathrooms_1.0)
29	0.217573	(property_type_Apartment, bedrooms_1.0, bathro...
30	0.216179	(property_type_Apartment, accommodates_2, bath...

Fig. 3. Most and least frequent itemsets found using Apriori algorithm with a minimum support of 0.2

```
{'property_type_Apartment': 0.7285913528591352,
 'property_type_House': 0.15676429567642958,
 'room_type_Entire home/apt': 0.5933054393305439,
 'room_type_Private room': 0.38437935843793586,
 'accommodates_1': 0.12245467224546723,
 'accommodates_2': 0.4136680613668061,
 'accommodates_3': 0.1193863319386332,
 'accommodates_4': 0.18131101813110181,
 'bathrooms_1.0': 0.7673640167364016,
 'bathrooms_2.0': 0.13333333333333333,
 'bedrooms_1.0': 0.6635983263598326,
 'bedrooms_2.0': 0.19330543933054392}
```

Fig. 4. Most and least frequent itemsets found using custom Apriori algorithm with a minimum support of 0.1

C. Q6

The coefficients of each variable are 16.4218, 1.4096, -8.6753, 17.3606, -6.6749, 0.6594, 3.3624, 57.5541, 25.4978, and 840.9567. The R-squared value of this OLS model is 0.584 and the adjusted R-squared is 0.583. The coefficients that are statistically significant are review_scores_rating, review_scores_accuracy, and review_scores_cleanliness because their respective p-values (0.015, 0.047, and 0.000) are less than 0.05.

It is not surprising that these variables are significant. Higher-rated and cleaner Airbnbs most likely mean that they are more luxurious and, therefore, cost more. Also, property owners would be less likely to lie about their properties if they are more expensive because the customers are already paying so much, so that variable being significant makes sense as well. However, it is surprising that positivity_mean and positivity_simple_mean are very insignificant because it seems that if a review is more positive, it would incentivize raising the price because there would be more demand for that specific

```
{'property_type_Apartment': 0.7285913528591352,
 'room_type_Entire home/apt': 0.5933054393305439,
 'room_type_Private room': 0.38437935843793586,
 'accommodates_2': 0.4136680613668061,
 'bathrooms_1.0': 0.7673640167364016,
 'bedrooms_1.0': 0.6635983263598326}
```

Fig. 5. Most and least frequent itemsets found using custom Apriori algorithm with a minimum support of 0.2

Airbnb. However, the two variables seem more or less independent, so the significance is negligible. However, because review_score_cleanliness (0.09808746471267853) and review_score_rating (0.07075344894763971) have the largest correlations, they are the most important variables for the model.

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                  price   R-squared (uncentered):                   0.584
Model:                            OLS   Adj. R-squared (uncentered):              0.583
Method:                 Least Squares   F-statistic:                              501.8
Date:                Sun, 10 Oct 2021   Prob (F-statistic):                        0.00
Time:                        11:13:36   Log-Likelihood:                         -22988.
No. Observations:                3585   AIC:                                  4.600e+04
Df Residuals:                    3575   BIC:                                  4.606e+04
Df Model:                          10
Covariance Type:            nonrobust
===============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
host_response_rate           16.4218     19.738      0.832      0.405     -22.277      55.121
review_scores_rating          1.4096      0.579      2.436      0.015       0.275       2.544
review_scores_accuracy       -8.6753      4.363     -1.989      0.047     -17.229      -0.122
review_scores_cleanliness    17.3606      3.788      4.584      0.000       9.934      24.787
review_scores_checkin        -6.6749      5.055     -1.320      0.187     -16.586       3.236
review_scores_communication   0.6594      5.282      0.125      0.901      -9.696      11.015
positivity_mean               3.3624     42.328      0.079      0.937     -79.627      86.352
negativity_mean              57.5541    154.812      0.372      0.710    -245.975     361.083
positivity_simple_mean       25.4978    118.418      0.215      0.830    -206.677     257.672
negativity_simple_mean      840.9567    521.239      1.613      0.107    -181.000    1862.913
==============================================================================
Omnibus:                     5323.683   Durbin-Watson:                   1.699
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4469811.723
Skew:                           8.649   Prob(JB):                         0.00
Kurtosis:                     175.117   Cond. No.                     2.02e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.02e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Fig. 6. Summary of OLS linear regression model

It is also important to note the effect size of each of the components. While the p-values of each component seem satisfactory in explaining a coefficient's effect on a linear regression model, the effect size expands on this by explaining the magnitude of each component's significance. These values are all very low (none are above 0.1 or below -0.1), so the magnitude of the significance is very minimal.

D. Q7

The coefficients of each variable are -3.7654, -0.2058, and -0.4094. The standard error values are 2.045, 3.353, and 3.624 respectively. This is less than all of the standard error values for Q6's coefficients except for review_scores_rating (0.579). However, the R-squared value is 0.001 and the adjusted R-squared value is 0.000. Thus, the model is fitted a lot worse to the data compared to the model from Q6.

E. Q7

For Q8, five plots/tables were generated, the last two of which were taken from Q6 and Q7.

First, a vast majority of the reviews run through the sentiment analysis contained overwhelmingly positive words, as shown by the bar chart in figure 8.

Second, all of the values in Q6's model had low correlation except for review_scores_cleanliness and

```
                            OLS Regression Results
=======================================================================================
Dep. Variable:                  price   R-squared (uncentered):                   0.001
Model:                            OLS   Adj. R-squared (uncentered):              0.000
Method:                 Least Squares   F-statistic:                              1.136
Date:                Sun, 10 Oct 2021   Prob (F-statistic):                       0.333
Time:                        11:49:59   Log-Likelihood:                         -19502.
No. Observations:                2868   AIC:                                  3.901e+04
Df Residuals:                    2865   BIC:                                  3.903e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -3.7654      2.045     -1.842      0.066      -7.774       0.244
x2            -0.2058      3.353     -0.061      0.951      -6.781       6.369
x3            -0.4094      3.624     -0.113      0.910      -7.515       6.696
==============================================================================
Omnibus:                     3325.349   Durbin-Watson:                   0.757
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           843633.098
Skew:                           5.622   Prob(JB):                         0.00
Kurtosis:                      86.266   Cond. No.                         1.77
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Fig. 7. Summary of OLS linear regression model using the three principal components generated in Q7

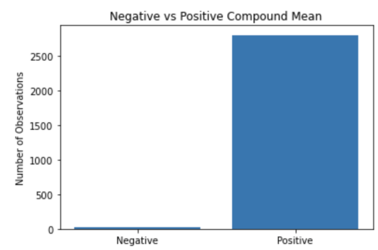

Fig. 8. Positive vs negative compound_ mean

review_scores_accuracy, review_scores_communication and review_scores_checkin, positivity_simple_mean and positivity_mean, and negativity_simple_mean and negativity_mean. This is shown in the "correlogram" in figure 9.

Third, the principal components generated from the PCA in Q7 generate some very definitive cluters, as shown in figure 10.

F. Misc.

In order to improve this model, a number of options could be implemented. First, the NaN values could be replaced throughout all the dataframes with either the column mean or median. This would limit the effect of missing data. Also, more feature engineering techniques such as one-hot encoding could be implemented across all categorical data like it was for the Apriori algorithm. This would help make the data more readily available for modelling. Finally, because a lot of the data is heavily skewed (for example, the compound_ featured in figure 8), certain data could be transformed or standardized so that it appears less skewed.

In terms of causal inference problems, plenty presented themselves throughout the lab. First and foremost, a number of listings do not have reviews, which creates a non-response bias. Also, in terms of selection bias, only a certain clientele would buy larger, nicer Airbnbs. This is a small percentage of the population, so their reviews would not be as diverse or "average" as the reviews for more "middle-of-the-road" listings.

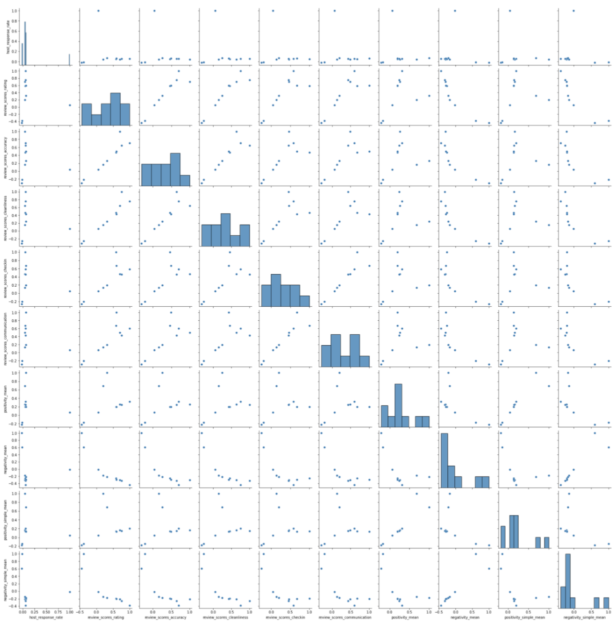

Fig. 9. "Correlogram" of variables used in Q6's OLS linear regression model

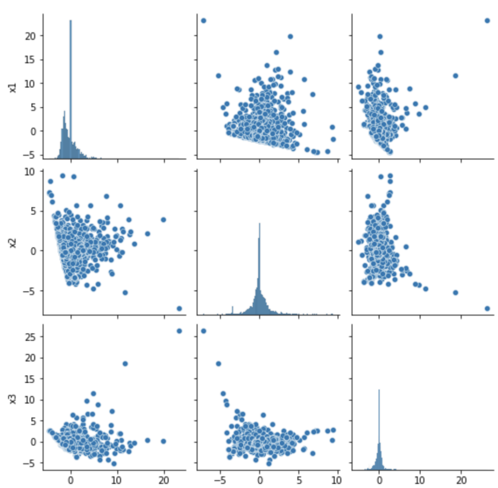

Fig. 10. Pairplot of three principal components generated in Q7