Airbnb Lab Report

Xuanyu Shen, Yushu Lyu

***Abstract*— Airbnb, Inc. is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking. In this lab, we are going to discover the patterns hidden in the dataset which is about the airbnb information in Boston.**

I. INTRODUCTION

This lab is mainly consisted of 5 parts: Setting up Dataset and Descriptive Analysis, Sentimental Analysis and Adding New Data, Data Mining, Linear Regression, and Data Visualization. In descriptive analysis, we build a descriptive table that shows the minimum, maximum, mean, median, variance, and standard deviation values for each variable. In sentiment analysis and adding New Data, we take a look at the comments column of the reviews_df dataset, run the sentiment analysis on each cell in the comments column and add four new columns to the dataset as a result. In Data Mining, we use the Apriori algorithm in the mlxtend package of Python, and calculate the frequent itemsets in the "Listings" dataset. We also developed a manual apriori algorithm to compare the results of frequency pairs in the itemset. In lenear regression, we build two OLS models, one is normal, one is with PCA on data features, we compare the two models accuracy and performance based on the parameters shown on the summary. In Data Visualization part, we organized our results build several diagrams and tables to visualize our results.

II. DATA

The raw data comes in as the form of excel data sheets(csv format). There are in total of five files named "calendar", "listings", "reviews", "negative_words" and "positive_words". The "calendar" file has a total of 1308890 rows corresponding to these amount of houses(listings) while some are missing one of the data "price" and were left empty. In the file "listing", there are only 3585 roles indicating only 3585 listings but there are in total of 95 variables about it. This data sheet provides the study with a wider range of information and a better understanding about the houses but in a smaller group. Including descriptive variables of the description of the house; while also containing numerical variables including the price, review scores and fees; also supporting information that allows people to gain a better understanding such as the flexibility of the cancellation policies .The "review" file contains information about the reviews and feedbacks that people gave to the listings. There are in total of 68275 reviews responding to the listings that are in the "listing" file. Apart from the listing id, the reviewer's name and their id comes in along with the date information, and most importantly, is the word description of the review they've had to the listings. These provides the foundation for the analysis in the research when it comes to the analysis of the reviews given in the word format. Lastly, "negative_words" and "positive_words" were two sheets containing words that are in the positive and negative category and are used as tools to perform analysis on the "review" file with extracting, and matching with the information given.

The dataset was then being extracted with more useful information including first, to the "listing" file with essential numerical variables that are useful for further analysis including "host response rate, host acceptance rate. . . . . reviews per month." The maximum, minimum, mean, medium, variance and the standard deviation was calculated for a better understanding to the data and to be aware of potential errors before analysis. Then "Sentiment Analysis" was done to the "reviews" data sheet and utilizing "negative_words" and "positive_words" document, positive and negative simple (percentage of positive and negative words) were added to the dataset along with the average scores of it.

	Variable	Minimum	Maximum	Mean	Median	Variance	Std. Deviation
0	host_response_rate	0.00	100.00	94.989082	100.00	1.566421e+02	1.251568e+01
1	host_acceptance_rate	0.00	100.00	84.173089	94.00	4.741836e+02	2.177576e+01
2	host_listings_count	0.00	749.00	58.902371	2.00	2.927377e+04	1.710958e+02
3	host_total_listings_count	0.00	749.00	58.902371	2.00	2.927377e+04	1.710958e+02
4	accommodates	1.00	16.00	3.041283	2.00	3.163707e+00	1.778681e+00
5	bathrooms	0.00	6.00	1.221647	1.00	2.514189e-01	5.014168e-01
6	bedrooms	0.00	5.00	1.255944	1.00	5.669402e-01	7.529543e-01
7	beds	0.00	16.00	1.609060	1.00	1.023341e+00	1.011603e+00
8	price	1000.00	400000.00	17392.580195	15000.00	2.199604e+08	1.483106e+04
9	weekly_price	8000.00	500000.00	92239.237668	75000.00	4.322444e+09	6.574530e+04
10	monthly_price	50000.00	4000000.00	369209.797297	292500.00	8.400319e+10	2.898330e+05
11	security_deposit	9500.00	450000.00	32469.821162	25000.00	1.080769e+09	3.287505e+04
12	cleaning_fee	500.00	30000.00	6838.014528	5000.00	2.630406e+07	5.128748e+03
13	guests_included	0.00	14.00	1.429847	1.00	1.116487e+00	1.056640e+00
14	extra_people	0.00	20000.00	1088.619247	0.00	3.661522e+06	1.913510e+03
15	minimum_nights	1.00	300.00	3.171269	2.00	7.872827e+01	8.872895e+00
16	maximum_nights	1.00	99999999.00	28725.836820	1125.00	2.788576e+12	1.669903e+06
17	availability_30	0.00	30.00	8.649930	4.00	1.088657e+02	1.043387e+01
18	availability_90	0.00	90.00	38.558159	37.00	1.099164e+03	3.315365e+01
19	availability_365	0.00	365.00	179.346444	179.00	2.019706e+04	1.421164e+02
20	number_of_reviews	0.00	404.00	19.044630	5.00	1.264990e+03	3.556670e+01
21	review_scores_rating	20.00	100.00	91.916667	94.00	9.082026e+01	9.529966e+00
22	review_scores_accuracy	2.00	10.00	9.431571	10.00	8.680547e-01	9.316945e-01
23	review_scores_cleanliness	2.00	10.00	9.258041	10.00	1.366013e+00	1.168766e+00
24	review_scores_checkin	2.00	10.00	9.646293	10.00	5.815821e-01	7.626153e-01
25	review_scores_communication	4.00	10.00	9.646549	10.00	5.407750e-01	7.353741e-01
26	review_scores_value	2.00	10.00	9.168234	9.00	1.021987e+00	1.010934e+00

Fig. 1. Descriptive Table(part)

III. ANALYSIS

A. Setting Up Your Dataset and Descriptive Statistics

Looking at this table there are few things that are quite strange. The first thing is about the price, when comparing the daily, weekly, and monthly price, it seems to have a great discount going from days to weeks but not much to months, and the maximum daily price is about the same as the maximum weekly price which is very strange.The guest included also only had a mean and medium of one, which indicates individuals tend to book "airbnb" rather

than families? There are also few values that have a great difference in mean and medium which is "extra people" and "maximum nights" of "1088.62, 0.00" and "28725.84, 1125.00" which is also quite strange.

B. Data Mining

The first thing we did was select the data features we want. At this time, we chose "property_type", "room_type", "accommodates", "bathrooms", "bedrooms". We calculated the frequent itemsets in the listings dataset and set the minimum support of 0.1 and 0.2 separately. The first technique we used is import the python package "mlxtend" and use the apriori function inside it. After implementation, we first found out the 5 highest and 5 lowest frequent itemsets in the dataset given the minimum support of 0.1. The highest frequent itemsets are (bathrooms 1.0), (Apartment), (bedrooms 1.0), (Apartment, bathrooms 1.0), (Entire home/apt). The 5 lowest frequent itemsets are (Entire home/apt, bathrooms 1.0, bedrooms 2.0), (bathrooms 1.0, bedrooms 2.0), (Entire home/apt, bathrooms 2.0), (bedrooms 1.0, Private room, accommodates 1), (Private room, accommodates 1). Then, we set the minimum support to 0.2, which generates different results. The top 5 frequent itemsets are (bathrooms 1.0), (Apartment), (bedrooms 1.0), (Apartment, bathrooms 1.0), (Entire home/apt). The 5 least frequent itemset are (bedrooms 1.0, Apartment, (bedrooms 1.0, Entire home/apt, Apartment, bat...), (Private room, Apartment), (bedrooms 1.0, Private room, Apartment), (bedrooms 1.0, Entire home/apt, Apartment). According to the algorithm, we can conclude that the top 5 frequent itemsets are the same give the minimum support of 0.1 or 0.2. The reason is that they are all kept during each round of pruning since they have high frequencies. However, if we improve our minimum significantly, this phenomenon may shifts. Moreover, 1 bedroom, 1 bathroom, and apartment have the highest frequency rate, showing that people are more likely live alone or with his/her partner when they choose Airbnb. Also apartment the most convenient and cheapest living place for people who are traveling, so this is the reason why people choose it. In addition, We also developed the algorithm by ourselves. We found that the results we got from ourselves and the result from the apriori function are exactly the same, this makes sure that our algorithm are correct.

C. Linear Regression

We developed a model (Fig .2.) that predicts the price (Y) column based on the numeric variables we had and the sentiment scores we calculated. For this part, we will be focusingvon the following 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. The R squared is 0.49, which shows a relatively fair correlation between the model and the input data. Positivity_simple_mean, negativity_simple_mean have very high positive coefficient,

OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.490
Model:	OLS	**Adj. R-squared (uncentered):**	0.488
Method:	Least Squares	**F-statistic:**	343.3
Date:	Sun, 16 Oct 2022	**Prob (F-statistic):**	0.00
Time:	13:57:24	**Log-Likelihood:**	-23354.
No. Observations:	3585	**AIC:**	4.673e+04
Df Residuals:	3575	**BIC:**	4.679e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	138.6049	6.117	22.658	0.000	126.611	150.599
x2	1.1067	0.468	2.363	0.018	0.188	2.025
x3	-4.2313	4.101	-1.032	0.302	-12.272	3.809
x4	11.6133	3.878	2.994	0.003	4.009	19.218
x5	-7.2552	4.819	-1.506	0.132	-16.703	2.192
x6	-7.2607	5.006	-1.451	0.147	-17.075	2.553
x7	-20.2790	68.184	-0.297	0.766	-153.962	113.404
x8	-8.6657	203.427	-0.043	0.966	-407.511	390.179
x9	146.3839	118.596	1.234	0.217	-86.140	378.907
x10	584.7724	498.235	1.174	0.241	-392.081	1561.626

Omnibus:	4980.968	**Durbin-Watson:**	1.691
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3076159.536
Skew:	7.652	**Prob(JB):**	0.00
Kurtosis:	145.686	**Cond. No.**	1.59e+04

Fig. 2. OLS without PCA Reduction

positivity_mean has relatively high negative coefficient. Among host_response_rate, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_checkin, and review_scores_accuracy has the very high negative value; review_scores_cleanliness has the highest positive value. According to the p value, review_scores_communication,positivity_simple_mean, negativity_simple_mean, positivity_mean, negativity_mean's p_values are too big that we can't reject H0, which means it is not correlated with the dependent variable and these two variables has low significance. Thus, the variables has high significance and effect are: review_scores_rating, review_scores_cleanliness, review_scores_accuracy, review_scores_checkin, host_response_rate. I expected them to be significant, but I also expect review_scores_communication to be significant, but it is not the case in this model. Since effect size is related to the coefficient significance of the variable, so the rank of effect size of the variables is the same as the rank of significance of the variable.

OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.007
Model:	OLS	**Adj. R-squared (uncentered):**	0.006
Method:	Least Squares	**F-statistic:**	8.708
Date:	Sun, 16 Oct 2022	**Prob (F-statistic):**	9.38e-06
Time:	14:03:49	**Log-Likelihood:**	-5073.9
No. Observations:	3585	**AIC:**	1.015e+04
Df Residuals:	3582	**BIC:**	1.017e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.0298	0.007	-4.586	0.000	-0.043	-0.017
x2	-0.0206	0.013	-1.587	0.113	-0.046	0.005
x3	0.0282	0.018	1.605	0.109	-0.006	0.063

Omnibus:	5232.716	**Durbin-Watson:**	1.682
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4097730.397
Skew:	8.371	**Prob(JB):**	0.00
Kurtosis:	167.779	**Cond. No.**	2.70

Fig. 3. OLS with PCA Reduction

In the second OLS Model (Fig. 3.), we first applied standard scalar to normalize the data, then condeucted PCA reduction on the data features to select 3 most significant variables in the q0 variables. We applied OLS model after the reduction. The error values we got are very small compare

to the previous model. However, the R square of the model is very high compare to the previous normal model, which is 0.007. It shows that there is relative no correlation between the model the actual data. Thus, we conclude that the first OLS model is better than the model with PCA reduction.

IV. Result

After different parts of analysis on the Boston Airbnb dataset. We came up with several conclusions. The first finding is that: the price, when comparing the daily, weekly, and monthly price, it seems to have a great discount going from days to weeks but not much to months, and the maximum daily price is about the same as the maximum weekly price which is very strange. The second finding came from the frequency itemset mining that people are more willing to reserve an apartment and they are more likely to choose 1 bedrooms and 1 bathrooms when they choose Airbnb. This may reflects that the user group of Airbnb are individuals or couples. Moreover, the linear regression model shows that review scores of rating, cleanliness are the most important factors for the pricing. Some of ways to improve the analysis are: first, we can use stochastic process to fill an value to NaNs, rather than just drop it. This can improve population while keep the data quality. Second, we can just use the score variables in the dataset to figure out what kind of score rate is crucial to the price. Third, a time series analysis will also help us to understand the trend of price given different score rate at different time. In addition, there exists some typical causal inferences problems: selection bias may happen since these data may exists same user with different orders on Airbnb, which causes bias between the dependent variable and independent variable. There's also simultaneity, since the price will cause different score rate from customers, for example, based on the apartment quality; and the score rate will also cause the price fluctuation: low score rate may cause low price. Last, omitted variable bias is existing when we do PCA reduction.