Final Report

(Prediction model for AirBnb)

IST344-Data Analytics

TaeHoon Kim

OCT 12. 2018

TAEBLE OF CONTENTS

INTRODUCTION

PROBLEM

Recently, AirBnB is rising as a new form of accommodation. Many travelers enjoy staying in AirBnB housing because of relatively cheap price and personalized service. However, in my perspective there is one downside about this platform: the individual host sets the price. Although there are reviews which help us understand if the price is reasonable for the most of time, sometimes when the housing is located in the lonesome area, posted for the first time or has unique features, it is difficult to validate the fair price. This might be damaging to travelers since they can suffer loss from mistaken price.

Therefore, I came up with an idea that what if there is a way to know the sensible price without the reviews. Then I found a AirBnB dataset that contains features such as price, room type, and number of bathrooms. Using regression analysis with the dataset, I am planning to produce the model that predicts the price. Furthermore, I will find which features strongly impacts on the price.

If the model is successfully made and influential features on the price are known, AirBnB users will be able to appreciate this new fascinating platform with higher confidence.

DATA

The size of dataset: about 74,000 rows

Link: https://www.kaggle.com/rudymizrahi/airbnb-listings-in-major-us-cities-deloitte-ml#test.csv

- ⇨ There are two files (test.csv and train.csv) and I will only use the train.csv file since it is the one that contains price attribute.

Since the cost of living in one city is a lot different from the others, I assumed that the difference could hurt my prediction model. For example, rent prices in New York are about 40% higher than in Los Angeles, so I expect that two AirBnb housings in those two areas will be priced differently even though they have the same conditions. Therefore, I will only use the rows located in LA which take about 30% of the whole dataset. Then, I will later figure out how to interconnect the model with objects located in other cities by finding the key features that explain the relationship between the cities.

ETL

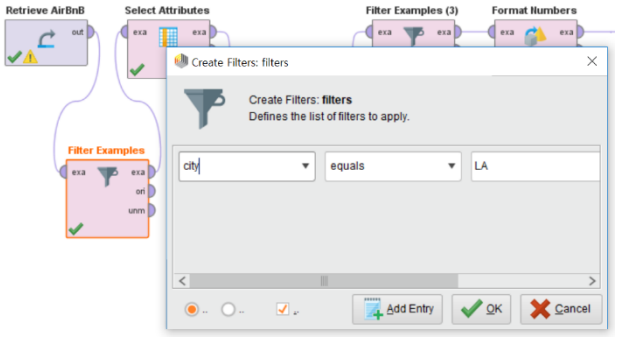
EXTRACTION

Fortunately, the dataset I obtained was formed in CSV file. Therefore, it was not required to be converted into other format so as to extract the data into any applications such as Excel or Rapid Miner which I used for data cleaning.

TRANSPORT

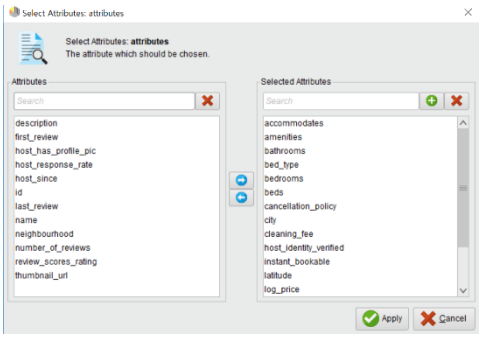
The dataset was not clean enough to be suited for analysis, so I cleaned up the data, using applications, RapidMiner, Excel, and Python. Here are the steps I followed:

1. Since I planned to only use the objects that are belonged to city of 'LA'. I filtered and left the data only located in 'LA'.



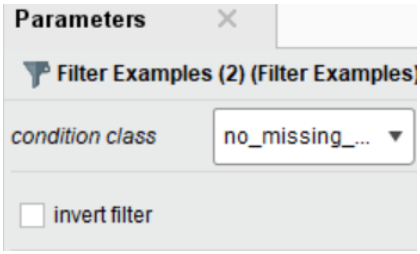
2. In the original dataset, there were 28 attributes. However, a few of them were not necessary for building the prediction model. Therefore, I removed those columns.

   - Related to review: The whole purpose of the analysis is to build a prediction model that forecasts the price without looking at the reviews.
   - Has no impacts: The attributes that are not likely to influence on the price were not needed for the model.

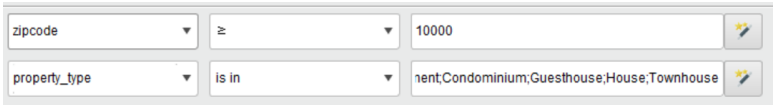


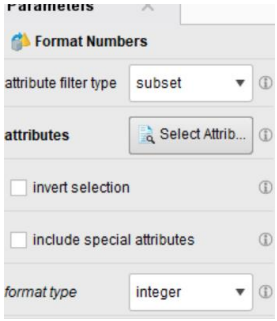
3. A few attributes had outliers, but the amount was not significantly large. Therefore, I considered them as missing completely at random(MCAR) which would result in only minimal loss of information when deleted.

   - The attribute, zipcode, had the largest amount of missing values, but the total was less than 1% of the whole. Therefore, I filtered out every missing values.



4. Some unrealistic values in the attributes such as bathrooms, bedrooms, and zip code, were found, so I either removed them or replaced the values.
   - Zipcode that contained values less than 5 digits.
   - Some of the number of bathrooms and bedrooms had decimal numbers while they were expected to have only integers. Ex) There is no such thing as 1.5 rooms. It is either 1 or 2. Therefore, I rounded off those numbers.





5. In the attribute, property_type, there were a various kinds of values, each of which took less than 1% of the whole data. Therefore, in order to make the data more concise, I left only rows with the 5 largest elements: Apartment, House, Condominium, Townhouse, and Guesthouse.

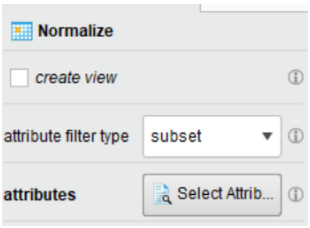
6. One of the essential rule in database is that each cell should contain only one element. However, in the dataset, one attribute called amenities had more than one element in each cell. Those values were placed into different cells.

   - I assumed that not every element was needed to be included. For example, since elements such as Wi-Fi, TV, or air-conditioner were placed almost in any objects, they cannot significantly influence on the prediction model. Therefore, I made dummy variables just for four elements: pool, gym, hot tub, and parking lot, which seem to have high impact when setting the price.

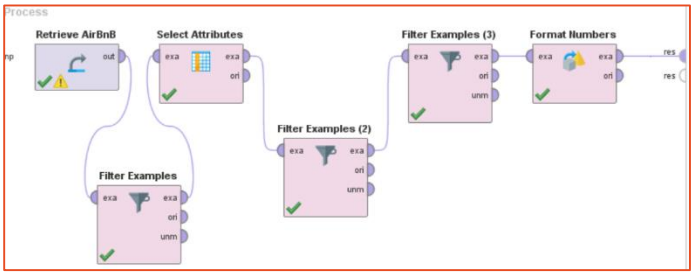
```
if 'Pool' in elements:
    each_result += [1]
else:
    each_result += [0]

if 'Hot tub' in elements:
    each_result += [1]
else:
    each_result += [0]

if 'Free parking on premises' in elements:
    each_result += [1]
else:
    each_result += [0]

if 'Gym' in elements:
    each_result += [1]
else:
    each_result += [0]
```

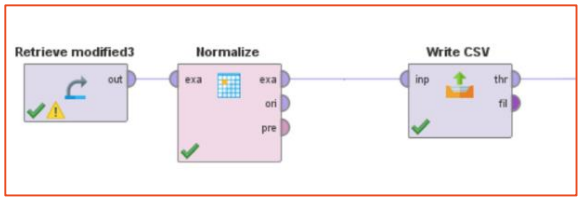
7. Since the unit of each quantitative variable was different, "normalize" function in RapidMiner was used to transform the values to z-score.

   - The quantitative variables:
     - accommodates
     - bedrooms
     - bathrooms
     - beds



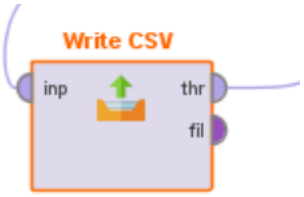
Here is the image of whole process of RapidMiner:





LOAD

Finally, after cleansing up, the dataset was formed back into CSV file and loaded up in Excel. The number of rows decreased to 19,775.



log_price	property_	room_typ	accommo	bed_type	cancellatic	cleaning_f	city	host_iden	instant_bc	latitude	longitude	zipcode	pool	kitchen	parking	gym
4.418841	Apartmen	Entire hon	3	Real Bed	moderate	TRUE	LA	f	t	33.98045	-118.463	90292	1	1	1	
4.787492	Condomir	Entire hon	2	Real Bed	moderate	TRUE	LA	t	f	34.04674	-118.26	90015	1	1	1	
3.583519	House	Private ro	2	Real Bed	moderate	TRUE	LA	f	t	33.99256	-117.896	91748	0	1	1	
5.010635	House	Entire hon	4	Real Bed	strict	TRUE	LA	t	f	33.87586	-118.403	90254	0	1	1	
4.248495	Apartmen	Private ro	2	Real Bed	flexible	TRUE	LA	f	f	33.81323	-118.389	90277	0	1	1	
4.955827	Apartmen	Entire hon	2	Real Bed	strict	TRUE	LA	t	f	33.77853	-118.146	90804	0	0	0	
4.382027	House	Private ro	2	Real Bed	strict	TRUE	LA	t	f	34.16579	-118.444	91401	1	1	1	
4.905275	Apartmen	Entire hon	4	Real Bed	moderate	TRUE	LA	t	t	34.047	-118.267	90015	1	1	1	
4.007333	House	Private ro	2	Real Bed	moderate	TRUE	LA	t	f	34.17536	-118.431	91401	0	1	1	
5.192957	Apartmen	Entire hon	3	Real Bed	flexible	TRUE	LA	t	f	34.05158	-118.243	90028	1	1	1	
3.828641	Apartmen	Private ro	2	Real Bed	strict	TRUE	LA	f	f	34.06408	-118.346	90036	0	1	1	
5.220356	Apartmen	Entire hon	4	Real Bed	strict	TRUE	LA	t	f	33.99751	-118.472	90291	0	1	1	
4.356709	Apartmen	Entire hon	4	Real Bed	flexible	TRUE	LA	f	t	34.08631	-118.271	90026	1	1	1	
4.787492	Apartmen	Shared ro	4	Real Bed	moderate	TRUE	LA	f	f	34.1078	-118.32	90068	1	0	0	
4.317488	Apartmen	Private ro	8	Real Bed	moderate	TRUE	LA	t	t	34.05542	-118.275	90057	0	1	1	

DESCRIPTIVE STATISTICS

TYPE OF VARIABLES:

- Quantitative
  - Log_price: This will be my dependent variable
  - accommodates
  - bathrooms

- bedrooms
- beds
- latitude
- longitude

➢ Categorical

- property_type: (Apartment, House, Condominium, Townhouse, Guesthouse)
- room_type: (Entire home/apt, Private room, Shared room)
- bed_type: (Real Bed, Futon, Pull-out Sofa, Airbed, Couch)
- cancellation_policy: (flexible, moderate, strict, super_strict_30, super_strict_60)
- cleaning_fee: (TURE, FALSE)
- pool: (1, 0)
- kitchen: (1, 0)
- parking: (1, 0)
- gym: (1, 0)
- host_identity_verified: (t, f)
- instant_bookable: (t, f)
- city
- zipcode

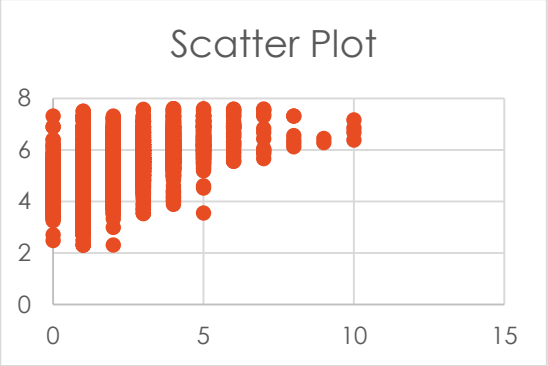
STATISTICAL MEASUREMENT

➢ Quantitative variables:

	bathrooms	bedrooms	beds	accommodates
Mean	1.362	1.355	1.854	3.399
Median	1	1	1	2
Mode	1	1	1	2
Min	0	0	1	1
Max	8	10	16	16
Range	8	10	15	15
Q1	1	1	1	2
Q2	1	1	1	2
Q3	2	2	2	4
IQR	1	1	1	2
Upper Limit	3.5	3.5	3.5	7
Lower Limit	-0.5	-0.5	-0.5	-1
Variance	0.542	0.917	2.007	5.581
Standard Deviation	0.736	0.958	1.417	2.362
Covariance with price	0.2817	0.4324	0.5523	1.1072
Correlation with price	0.5259	0.6207	0.5360	0.6443

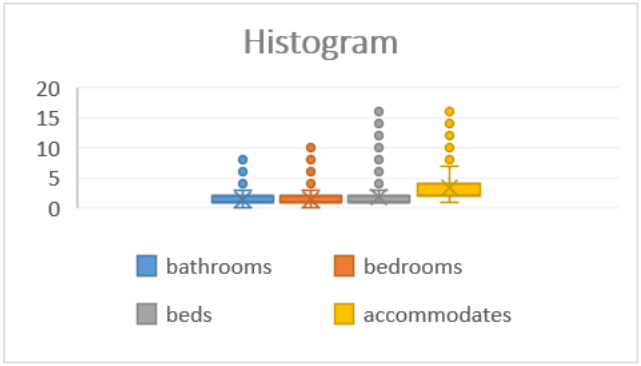
- The numerical figure I am focusing is the correlation between price and the qualitative variables. As shown in the chart above, they all have the positive numbers which indicate the positive linear relationship. This convinces me that those variables will be useful when building a prediction model.

- This is the example of scatter plot:



Y-value represents price and X-value represents the number of bedrooms. The graph depicts that more proportion of dots is located in the higher level of price as the number of bedroom goes up.

- According to the chart above, all of the variables have outliers since their max numbers exceed the upper limit. However, I checked that housings listed in the AirBnb website actually have the values similar to the Max. Therefore, I did not see them as errors and determined to keep those numbers although they seem to be unusual.



➢ Categorical Variables:

pool	frequency	relative frequency	Percent frequency
1	4390	0.222	22.201
0	15384	0.778	77.799
	19774		

hot tub	frequency	relative frequency	Percent frequency
1	3472	0.176	17.558
0	16302	0.824	82.442
	19774		

parking	frequency	relative frequency	Percent frequency
1	13122	0.664	66.360
0	6652	0.336	33.640
	19774		

gym	frequency	relative frequency	Percent frequency
1	2590	0.131	13.098
0	17184	0.869	86.902
	19774		

- I calculated frequency distribution for a few categorical variables to see if there are any particular patterns.

- The variables in the tables above can be considered as extra services provided. I assumed that if the service is less likely to be offered, objects with those attributes might have higher price because of their uniqueness. I found that gym, pool, and hot tub service have relatively low percent frequency. Therefore, I calculated the mean price for each one to see if there is a distinct pattern. Here is the result:

Overall mean price	4.726
mean price with gym	4.719
mean price with pool	4.734
mean price with hot tub	4.995

Unfortunately, there were not much differences between overall mean price and mean price of specific objects. Then I thought that this could happen because most of the prices are located between 3.36 and 5.48. (The high density in the middle may reduce the impact of high price.)

Therefore, I looked at the ratio of having those services based on the price range. Here are the measurements I calculated, using python.

price	pool_number	overall_number	ratio
3.36 or less	19	268	0.070896
4.42 or less	1171	6610	0.177156
5.48 or less	2334	10055	0.232123
6.54 or less	646	2402	0.268943
more than 6.54	220	439	0.501139
	4390	19774	

price	hot tub_number	overall_number	ratio
3.36 or less	14	268	0.052239
4.42 or less	819	6610	0.123903
5.48 or less	1801	10055	0.179115
6.54 or less	638	2402	0.265612
more than 6.54	200	439	0.455581
	3472	19774	

price	gym_number	overall_number	ratio
3.36 or less	16	268	0.059701
4.42 or less	568	6610	0.08593
5.48 or less	1583	10055	0.157434
6.54 or less	361	2402	0.150291
more than 6.54	62	439	0.14123
	2590	19774	

As shown in the table, objects with higher price are more likely to provide those services. Therefore, it can be said that those facilities may be related with the price.

- Next, I compare the average price of objects with and without the variable, cleaning_fees because the variable is directly related to the price. The difference between average prices of the objects with and without the variable was about $32(natural log of 4.793 - 4.481). Given that the overall mean price is 4.72 and standard deviation is 0.727, the result looks minor. However, it is still fair enough to think that housings without cleaning fee would be less expensive if all of the other conditions are alike.

- From the analysis of categorical variables, I was not able to obtain many significantly insightful information. Therefore, I am planning to do other analysis such as the hierarchical clustering so as to find the meaningful patterns during the further study.

DESCRIPTIVE DATA MINING

The application, Analytic Solvers, is able to accept only 10,000rows. However, my data set contained about 19,000 rows. Therefore, using python, I randomly selected 500 objects. (The number was chosen to be 500 because Hierarchical Clustering prefers small data set.) Also, a few

data were converted into dummy variables to be used: "host_identity_verified", "instant_bookable", "cancellation", "property_type", and "room type".

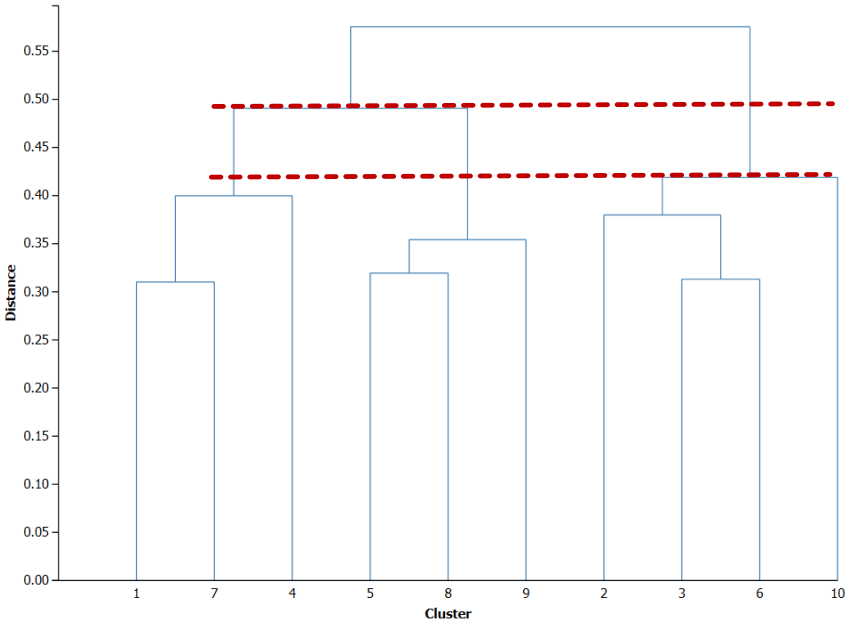
HIERARCHICAL CLUSTERING

- Among the categorical variables in my dataset, I selected 6 of them in order to run the hierarchical clustering, which are "pool", "hot tub", "gym", "cleaning_fee", "parking", and "just room". The first four were the ones that had a result of showing the possibility of having correlation with "price" from the previous descriptive analysis. Then I added two more variables because I assumed that some insight can be made from them.

- The variable, "just room", was derived from "room_type" attribute that contained three different elements, "entire home", "private room", and "shared room". In order to run the clustering, converting into dummy variable was required. Since shared room only took 2% of the whole data, I just aggregated rooms together and made one dummy variable. The value of 1 indicates single room and 0 implies entire home.

- For the similarity measure, matching coefficient was used because matching 0 entries also implies the similarity between the objects. For instance, if an object has a value of 0 for "gym", it means that the object does not have gym. Also, among three clustering methods provided, group average linkage was used in order to prevent damages from outliers.

- Here is the figure of dendrogram:



- The vertical line between two dotted red lines is the longest. This indicates that it is appropriate to divide into 3 clustering since it prevent a loss from clustering the most.

- Results:
  In order to see the characteristic of each cluster, I created a pivot table. However, since the difference between the size of each cluster was huge, another table was created to show the ratio of sum of each variable in each cluster to the size of each cluster.

cluster	size	cleaning_fee	pool	hot tub	parking	gym	just room	Avg price
1	60	59	49	39	52	49	11	5.039706
2	421	336	32	16	259	5	154	4.757153
3	19	0	19	14	16	12	15	4.548839
Ratio								
1	1.000	0.983	0.817	0.650	0.867	0.817	0.183	5.039706
2	1.000	0.798	0.076	0.038	0.615	0.012	0.366	4.757153
3	1.000	0.000	1.000	0.737	0.842	0.632	0.789	4.548839

- Before the clustering, I expected that a cluster with large numbers for the variables, "cleaning_fee", "pool", "hot tub", "parking", and "gym" and small numbers for the variable, "just room", have the highest average price; It seemed to be true with cluster 1 and 2, but 3 However, since cluster 3 took only negligible amount, 4% of the whole, the result could still provide an idea that those variables are related with the price. Therefore, all of those variables, especially "cleaning_fee" and "just room", can be good independent variables to predict the price of AirBnb.

K-MEANS CLUSTERING

- Four quantitative variables in my dataset were used in order to run K-means clustering: "bedrooms", "bathrooms", "beds", and "accommodates". Since each variable had different units, the data was standardized before clustered into 3 groups.

- For the K-means clustering, there is a rule of thumbs that ratio of between-cluster distance to average within-cluster distance should exceed one for every clusters. As shown below in the table, all of the ratios exceeded one.

Inter-Cluster Distances (Between Clusters)

Cluster	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0	3.109348565	4.8986823
Cluster 2	3.109348565	0	1.8393258
Cluster 3	4.898682303	1.839325755	0

Cluster Summary (Within Cluster)

Cluster	Size	Average Distance
Cluster 1	61	2.008358123
Cluster 2	135	1.16246805
Cluster 3	304	0.620712658
Total	500	0.93627936

- Results:

		AVAERAGE				
Cluster	Count	price	bathrooms	bedrooms	beds	accommodates
1	61	5.7281	2.5574	3.1475	4.7049	7.9836
2	135	5.0796	1.4370	1.7037	2.4815	4.6667
3	304	4.4619	1.1053	0.8289	1.1086	2.0987

- The result was similar to what I expected. I estimated that as the cluster has higher average price, the average number of each variable will also be large. As shown in the table above, cluster 1, which has the highest average price, also has the highest average number for all the attributes and cluster 3 with the lowest average price has the lowest average values. Therefore, the result indicates that those three variables are significantly correlated with the price and will be valuable independent variables when producing regression model.

ASSOCIATION RULES

- I created the association rules with 6 categorical variables, which are "hot tub", "gym", "parking", "cleaning_fee", "pool", and "just room".

- The minimum support was 50 transactions (10%) and minimum confidence was 50%. In order to follow the rule of thumbs, the minimum support was first determined to be 20% of total number of transaction. However, only three rules were made from data set. Therefore, 10% was used instead.

- The table below displays top 10 rules, following the descending order of lift-ratio.

Rule ID	A-Support	C-Support	Support	Confidence	Lift-Ratio	Antecedent	Consequent
Rule 6	100	69	53	53.000	3.841	[pool]	[hot tub]
Rule 7	69	100	53	76.812	3.841	[hot tub]	[pool]
Rule10	66	100	50	75.758	3.788	[gym]	[pool]
Rule 9	100	66	50	50.000	3.788	[pool]	[gym]
Rule 11	69	327	61	88.406	1.352	[hot tub]	[parking]

Rule 15	77	327	68	88.312	1.350	[cleaning_fee, pool]	[parking]
Rule 8	100	327	86	86.000	1.315	[pool]	[parking]
Rule 12	66	327	55	83.333	1.274	[gym]	[parking]
Rule 14	100	269	68	68.000	1.264	[pool]	[cleaning_fee, parking]
Rule 4	327	395	269	82.263	1.041	[parking]	[cleaning_fee]

- According to the rules, there is a great possibility to identify that an object would offer "hot tub" when the object provides "pool" and vice versa since the first two rules have lift ratio of 3.841. The next two rules also showed the high association between "gym" and "pool". From the result, it was shown that the three attributes, "pool", "hot tub", and "gyms" are highly related to each other and may have similar characteristics. The outcome proposed that if those three variables are all used for building multiple regression model to predict the price of AirBnb, they may cause multicollinearity. Therefore, if one is planning to use those variables, I would recommend to look into the relationship between the variables first.

LINEAR REGRESSION

The goal of the project is to build a model that predicts the price of AirBnb with the attributes in the dataset as accurate as possible. Therefore, the dependent variable is "log_price". However, since there are too many attributes in the dataset, different subsets of variables were experimented so as to find the most insightful ones for the independent variables.

SIMPLE LINEAR REGRESSION

- As shown in the table below, the variable, 'bedrooms', had the highest correlation. Therefore, 'bedrooms' was selected as independent variable to run the simple linear regression.

Correlation Matrix				
	bathrooms	bedrooms	beds	accommodates
price	0.510	0.622	0.528	0.617

- Here is the result:
  - The coefficient of bedrooms was statistically significant since P-Value was a lot less than 0.05. In other words, the price of AirBnb was related to the number of bedrooms as expected through correlation between two variables.
  - The estimated simple regression equation was made, using the numbers from table below: *Price = 4.152 + 0.468bedrooms*
  - The equation indicates that the price will increase by 0.468 when the number of bedroom increases by 1.

Predictor	Estimate	Standard Error	T-Statistic	P-Value
Intercept	4.15237932	0.04319769	96.1250327	0
bedrooms	0.46792574	0.026387189	17.7330654	6.72E-55

- Conclusion:
  - Although the simple linear regression provided meaningful estimated equation, the value of R2, 0.387, was quite low. The number indicated that 38.70% of the variability in the values of price in the sample can be explained by the linear relationship between the number of bedrooms and price. However, the other 61.30% of the variability in sample price remained unexplained. This result suggested that there may be other factors that would contribute to the price, so the multiple linear regression have been run with other variables.

Metric	Value
Residual DF	498
R2	0.38704796
Adjusted R2	0.38581713
Std. Error Estimate	0.54809691
RSS	149.60429

- Recommendation
  - Looking at the number of bedroom is a good start to find the price of AirBnb, but the variable alone may not be sufficient enough to estimate the precise price of AirBnb. Therefore, when comparing the price of AirBnb, it may be too impetuous to make a decision only with the number of bedroom. There could be other reasons if an object with less bedrooms is more expensive than others with more bedrooms.

MULTIPLE LINEAR REGRESSION

- Several different subsets of independent variables were used in order to run multiple linear regression.

USING ALL VARIABLES IN DATASET

- I assumed that all of the variables in the dataset including both categorical and quantitative were somewhat related to the dependent variable, "log_price". However, the total number of variables was 16 and I was worried that the large amount would cause the overfitting or multicollinearity. Therefore, I used Best Subsets as feature selection method with maximum subset size of 17 in order to select only the most insightful variables.

- Here is the result:
  - As shown in the table below, the highest R2 was about 64%, but there was almost no change of the number from subset4. The subset included 3 variables, "bathrooms", "bedrooms", and "just room". Also, when I looked at the P-value of each variable, those three variables and one called, "parking", had a value a lot less than 0.05. Since "parking" was belonged to subset 5 which was right below the subset 4, I considered those four variables as meaningful ones and used them to run the multiple linear regression once again.

Best Subsets Details

Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2	Probability
Subset 1	1	1942.295494	10241.10011	-6.95789	-6.957886677	0
Subset 2	2	916.9921006	4574.119353	-2.75706	-2.76460353	2.4559E-235
Subset 3	3	111.1017495	120.2900574	0.544799	0.54296706	1.89898E-18
Subset 4	4	91.5942853	14.43179833	0.624724	0.622454155	0.040026347
Subset 5	5	89.57667671	5.276286313	0.63299	0.630024692	0.425989249
Subset 6	6	88.8920367	3.490859404	0.635796	0.632109222	0.668375005
Subset 7	7	88.14914588	1.383359352	0.638839	0.634443776	0.927523982
Subset 8	8	88.00566544	2.590045042	0.639427	0.634297001	0.935570517
Subset 9	9	87.88437155	3.919401858	0.639924	0.634057245	0.938754806
Subset 10	10	87.7364343	5.101445536	0.64053	0.633927678	0.953638311
Subset 11	11	87.60493693	6.374386534	0.641069	0.633728847	0.967176554
Subset 12	12	87.49570538	7.770436945	0.641516	0.633435918	0.978771868
Subset 13	13	87.45758239	9.559651996	0.641673	0.632843264	0.967351884
Subset 14	14	87.38016399	11.13159974	0.64199	0.632413478	0.987776673
Subset 15	15	87.3664508	13.05577846	0.642046	0.631713374	0.972497651
Subset 16	16	87.35658007	15.00120247	0.642087	0.630994147	0.972351908
Subset 17	17	87.35636259	17	0.642087	0.63023108	0

USING THE 4 VARIABLES

- Here is the result:
  - As expected and shown in Table 1 and 2 below, all of their P-value were less than 0.05 and the R2 was about 63%. Also, each of the estimated coefficient was reasonable. Since the dataset did not contain the object with zero price, it was fair that interception had no meaning in this case. (So, no value for intercept was not a problem here.)

  - The coefficient of "bathrooms" was 0.161, which indicated that a price will increase by 0.161 when the number of bathroom increases by one. The result is acceptable because it is common sense that the price gets higher with more bathrooms.

  - The coefficient of "bedrooms" was 0.275, which indicated that the price will increase by 0.275 when one more bedroom was offered. This result was also

reasonable because this meant more room with more money. We could also notice that having one more bedroom is a bit more expensive than one more bathroom.

- The next two coefficients of variable, "parking_0" and "parking_1" showed the different prices between when there is parking lot and when there is not. As expected, if the parking is offered, the price goes up by 0.14(4.489-3.349).

- Lastly, "just room_1" showed the value of -0.691 as the coefficient. The negative coefficient may bring the confusion because it seems like a customer gets paid when the object offers just a room, not an entire house. However, the number actually meant the price difference between when the room type is just a room and when it is an entire house. Since the price was calculated in a base of natural log, antilog of 0.691 might seem to be minimal. Therefore, I ran the model without "parking" and obtain the similar but different coefficient for "just room" in the Table 3. "just room_0" was 4.426 and "just room_1" was 3.717. Although the difference is still about 0.7(4.426-3.717), if those two values were converted before subtraction, their difference would be about $40.

- Two estimated multiple regression equations were made with the numbers in the Table 2: (Put 1 for the value of "just room" if the room type is a room, otherwise 0.)
  i. *In the case that parking = 0,*
     Price = 0.161bathrooms + 0.275bedrooms -0.691just room + 4.349parking
  ii. *In the case that parking = 1,*
     Price = 0.161bathrooms + 0.275bedrooms -0.691just room + 4.489parking

Metric	Value
Residual DF	495
R2	0.63299043
Adjusted R2	0.63002469
Std. Error Estimate	0.42539744
RSS	89.5766767

Table 1 – Regression Summary of 4 Variables

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	0	0	0	0	N/A	N/A
bathrooms	0.16056072	0.089495811	0.231625623	0.036169617	4.43910471	1.11E-05
bedrooms	0.27547346	0.219527892	0.331419031	0.028474389	9.67442907	2.16E-20
parking_0	4.34919656	4.245749845	4.452643279	0.052650856	82.6044792	1.1E-291
parking_1	4.48852123	4.388543767	4.588498699	0.050885126	88.2089051	6.7E-305
just room_0	0	0	0	0	N/A	N/A
just room_1	-0.6911089	-0.772453711	-0.609764109	0.041401734	-16.692753	5.96E-50

Table 2 – Coefficients of 4 Variables

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	0	0	0	0	N/A	N/A
bathrooms	0.16001578	0.088228564	0.231802991	0.036537427	4.37950311	1.45E-05
bedrooms	0.29143956	0.23572733	0.347151799	0.02835577	10.2779632	1.36E-22
just room_0	4.42617374	4.33222431	4.520123166	0.047817296	92.5642836	0
just room_1	3.71663846	3.626116247	3.807160683	0.046072954	80.6685516	2.6E-287

Table 3 – Coefficients of 3 Variables

- Conclusion
  - Although the value of R2 of the multiple linear regression was considerably increased to 63%, there are still 37% of the variability in sample price remained unexplained.
  - The imperfection of my estimated equation was appeared when I experimented the equation with 10 new observations outside the 500 data. Some of the price were accurately predicted while some were not.

Estimate log_price	log_price
5.200029	5.220356
4.785231	5.783825
4.094122	5.105945
5.200029	5.4161
4.233447	3.89182
4.945791	3.401197
4.233447	4.779123
4.233447	4.60517
6.623044	6.543912
6.829253	7.31322

  - This result suggested that there may be other contributing factors to the price that are not in the dataset I obtained. The factors could be things such as "pet allowed", "handicap accessible", and "distance from city center".

- Recommendation
  - When one tries to find the reasonable price for the unknown AirBnb housing, it would be better to look at the value of the four attributes: "bathrooms", "bedrooms", "parking", and "just room". However, since there are some unexplained parts, I would not say that the estimated regression equation will always provide accurate numbers. Therefore, it would be better to use the equation as reference.

PREDICTIVE DATA MINING

Since all of the variables used in this section were not significantly imbalanced, there was no need to do either oversampling or undersampling.

LOGISTIC REGRESSION

- In my dataset, the dependent variable, "log_price" was the continuous variable. Therefore, in order to run the logistic regression, which required to have categorical variable as response, I converted "log_price" into the categorical variable called, "price", using python. The value of 1 for the variable indicated that the price of each object is greater than 4.7 and the value of 0 showed that the price of each object is less than or equal to 4.7. The number, 4.7, was followed by the median of the price.

- I selected 4 variables that were used in the multiple linear regression, "bedrooms", "bathrooms", "just room", and "parking" because I expected that the two regressions see the similar dependent variables as the important ones.

- The cut off value were set to 0.5 and Best Subsets Method was used as the feature selection method. Also, in order to find out that my estimated equation would accurately predict outside the sample data, the data was partitioned into three sets: Training (50%), Validation (30%), and Testing (20%). The training set was large enough to keep the rule of thumbs. (500 * 50% > 6 * 2 outcome categories * 4 variables)

- However, according to the table below, it turned out that the variables, "bathrooms" and "parking" had no meaningful relationship with price in the logistic regression since their P-Value was above 0.05. Therefore, I took out the two variables, and ran the regression once again.

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-1.891716	-2.784811095	-0.998620965	0.150813	0.45566912	17.23505569	3.3E-05
bathrooms	0.28024925	-0.441936216	1.002434707	1.32346	0.368468741	0.578479008	0.446909
bedrooms	1.31633481	0.701484373	1.931185245	3.729726	0.313704967	17.60718564	2.72E-05
parking_0	0	0	0	1	0	N/A	N/A
parking_1	-0.0357767	-0.817630281	0.746076946	0.964856	0.398912235	0.0080435	0.928537
just room_0	3.51387013	2.639446764	4.388293505	33.57797	0.446142571	62.03318542	3.38E-15
just room_1	0	0	0	1	0	N/A	N/A

- Here is the result:
  - As shown in the tables below, the logistic regression provided accurate predictions. Not only the overall accuracy, but also the F1 score was pretty high in all three different partitions. The lowest value was specificity in the testing set, and it was still over 0.7 which indicates the 70% accuracy for the class 1 object. Furthermore, from lift chart and ROC curve, it was shown that the classifier performed very well since the blue lines were above red line.

Confusion Matrix		
Actual\Predicted	0	1
0	101	20
1	17	112

Error Report			
Class	# Cases	# Errors	% Error
0	121	20	16.52893
1	129	17	13.17829
Overall	250	37	14.8

Metrics	
Metric	Value
Accuracy (#correct)	213
Accuracy (%correct)	85.2
Specificity	0.834711
Sensitivity (Recall)	0.868217
Precision	0.848485
F1 score	0.858238
Success Class	1
Success Probability	0.5

Figure 1 - result of training

Confusion Matrix		
Actual\Predicted	0	1
0	57	17
1	13	63

Error Report			
Class	# Cases	# Errors	% Error
0	74	17	22.97297
1	76	13	17.10526
Overall	150	30	20

Metrics	
Metric	Value
Accuracy (#correct)	120
Accuracy (%correct)	80
Specificity	0.77027
Sensitivity (Recall)	0.828947
Precision	0.7875
F1 score	0.807692
Success Class	1
Success Probability	0.5

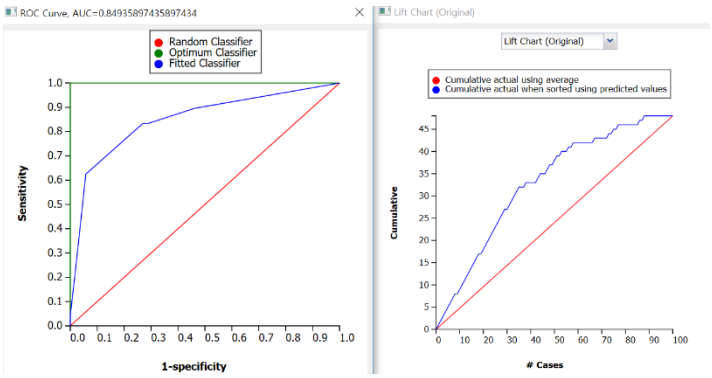
Figure 2 - result of validation

Confusion Matrix		
Actual\Predicted	0	1
0	37	15
1	8	40

Error Report			
Class	# Cases	# Errors	% Error
0	52	15	28.84615
1	48	8	16.66667
Overall	100	23	23

Metrics	
Metric	Value
Accuracy (#correct)	77
Accuracy (%correct)	77
Specificity	0.711538
Sensitivity (Recall)	0.833333
Precision	0.727273
F1 score	0.776699
Success Class	1
Success Probability	0.5

Figure 3 - result of testing



- The coefficient of each variable had insightful values. The coefficient value of 1.399 for "bedrooms" meant that log odds of having price over 4.7 would increase by 1.399 when the number of bedroom increased by 1. Also, the positive coefficient value of just room_0 indicated that having higher price of AirBnb and having entire house as room type are positively related.

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-1.9333296	-2.681008843	-1.185650341	0.144666	0.381476015	25.68486701	4.02E-07
bedrooms	1.39900849	0.842469293	1.955547688	4.051181	0.283953788	24.27422468	8.35E-07
just room_0	3.47287583	2.616953262	4.328798389	32.2293	0.43670321	63.24202124	1.83E-15
just room_1	0	0	0	1	0	N/A	N/A

- Conclusion
  - First of all, the result once again convinced that the number of bedroom and the room type are the important features in order to estimate the price since those two variables appeared as meaningful independent variables both in multiple linear regression and logistic regression.

- This model would have given high accuracy of predicting the price since the classifier only indicated whether the price was over 4.7 while the multiple regression predicted the actual price. However, this model would still be useful when someone wants to know if the price of housing is rationally cheap or expensive.

K-NEAREST NEIGHBORS ESTIMATION

- In order to find the most reasonable number of independent variables, I have tried to run the model with different subsets of variables. First, all 16 variables were selected because I assumed that more amount would increase the predictability although the level of explanation would decrease. Then I ran again the model with 4 variables selected in previous multiple linear regression model, which were "bedrooms", "bathrooms", "just room", and "parking".

- In order to estimate the model's effectiveness when used to data outside the sample, the data was partitioned into three sets: Training (50%), Validation (30%), and Testing (20%).

- Then I compared the results of testing set of each models since they would be the closest outcomes I would get in the real world with the model. I focused on the value of Root Mean Squared Error(RMSE), which provides a measure of how much the predicted value varies from the actual value. (The lower the value, the higher the accuracy.)

- As shown in the tables below, the value of RMSE when all variables were used were slightly lower than when 4 variables were used. However, the difference was insignificant while it lost a lot of explanatory power. Therefore, I assumed that it would be better off to use only 4 variables in this case and looked into the result of model with 4 variables in detail.

Metric	Value
SSE	20.97061
MSE	0.209706
RMSE	0.457937
MAD	0.364558
R2	0.571386

Table 2 – prediction summary of all variables

Metric	Value
SSE	21.110319
MSE	0.2111032
RMSE	0.4594597
MAD	0.3559171
R2	0.5685301

Table 2 – prediction summary of 4 variables

USING THE 4 VARIABLES

- There were two rules of thumbs I needed to consider:

- Rule #1: The amount of observation in training set should be at least 10 times more than the number of variables for the estimation task; Fortunately, the number of 250 training cases exceeded 40 (4 variables * 10).

- Rule #2: In order to judge if a model has good estimation, the value of RMSE should be less than 10% of the average value of the variable being predicted. The average log_price was 4.783 and the RMSE in the test set was 0.459. Therefore, I was convinced that the classifier provided worthy predictions.

- Different number of k was experimented so as to find the optimal k. As shown in the Table 3 below, the most accurate outcome was resulted when k was 10 since it had the lowest value of RMSE.

- Although the RMSE (0.387) in training set was lower than the one (0.459) in testing set, it was still fair to be said that the model would not result in overfitting because RMSE in testing followed Rule #2.

K	RMSE
1	0.463757
2	0.465016
3	0.457089
4	0.457537
5	0.455705
6	0.455705
7	0.451649
8	0.451649
9	0.451649
10	0.45131

Table 3 – Search Log

Metric	Value
SSE	37.36444
MSE	0.149458
RMSE	0.386598
MAD	0.292219
R2	0.70934

Table 4 – Training: Prediction Summary

Metric	Value
SSE	21.110319
MSE	0.2111032
RMSE	0.4594597
MAD	0.3559171
R2	0.5685301

Table 5 – Testing: Prediction Summary

- Conclusion
  - Since the classifier had value of RMSE less than 10% of the average price, it can be said that overall the model provided good estimation. However, as shown in table below, there are also residual with a big difference (1.117). Therefore, one should be aware of it when using this model.

Record ID	log_price	Prediction	Residual
Record 379	5.3936275	4.27709017	1.1165374
Record 101	4.6539604	4.07944883	0.5745115
Record 364	3.8066625	4.27709017	-0.470428
Record 299	4.4998097	4.84212994	-0.34232
Record 58	5.1059455	5.29782542	-0.19188
Record 102	5.6167711	5.8385887	-0.221818

STORY TELLING (DRAFT STORY BOARD)

Who: people who would like to find out if the price of AirBnb is sensible and look for the credible source to solve the problem (predict the price fairly accurately)

What: The prediction models with high accuracy were made now; avoid the postings online that are overpriced by using the models

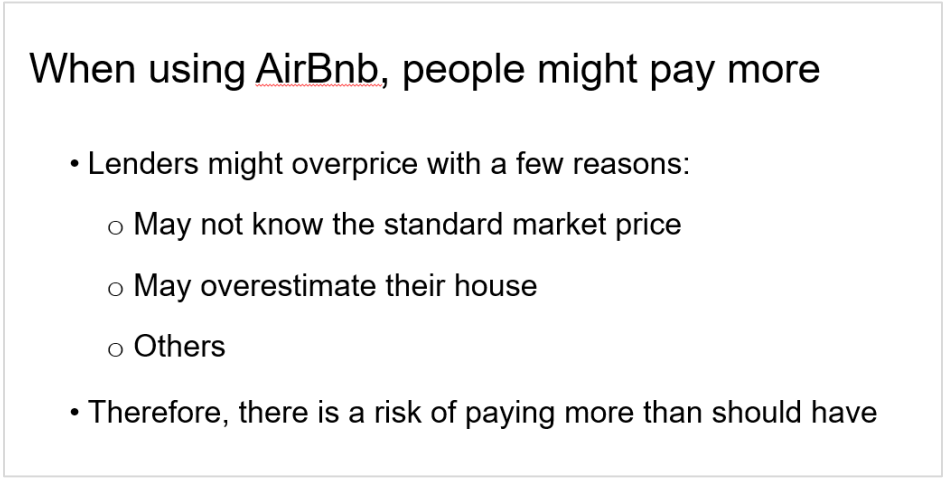
How: Presenting the models how accurate they are and explaining how the models are working with what features

STORYBOARD

1. Issue: People may sometimes pay more than they should have when using AirBnb
2. Demonstrate issue: What if there are no reviews on postings / what other source can people rely on
3. Idea for overcoming the issue, including making prediction models
4. Describe about the models and data used in the models
    a. Showing the statistical description of data (ex. P-value)
    b. Displaying what kinds of models were built
5. Conclusion / Show the accuracy of the models
    a. The regression summary and coefficient table for regression model
    b. The prediction summary table for k-nearest neighbor model
6. Recommendation: Model obtained high accurate, so please use them when selecting the accommodation in AirBnb.

STORY TELLING WITH DATA

1.



2.

There may not be a source to rely on

- People usually rely on reviews.
- There may not be a review on the posting:
  - **New posting**
  - Posting on the **lonesome area**

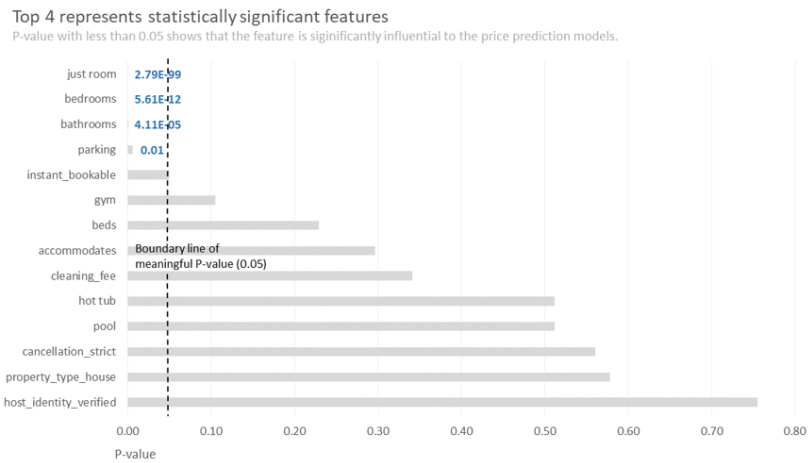
3.

How can we overcome this problem?

- Obtained data set that contains many features of Airbnb located in LA
- Two prediction models were built:
  - **Multiple regression model**
  - **K-nearest neighbor model**

4.

Which features were used?



5. Is multiple regression model reliable?

The model has

63%

more explanatory power than
just guessing every value as mean

6. Is k-nearest neighbor model reliable?

Root mean squared error(RMSE) is

only 9.5%

of average value of the dependent variable, price

7. Consider using the models!

- Both models are not perfect
- But provide **fairly accurate estimation**
- Therefore, **use them** to pay reasonable price.

SUMMARY

The report has been started with the research question that if one could predict the price of AirBnb without looking at the reviews for the housing located in the lonesome area or the new postings. Throughout the project, I was able to experiment the whole process of data analytics while solving the research question. The first step required was to clean up the dataset so that the data could be used for analysis. Otherwise, incorrect or inconsistent data could result in the false conclusions. I did not realize before that there were so many preprocessing works in order to obtain insights from the data. Many things were needed to be considered such as taking out or replacing missing data, standardizing the quantitative variables, and converting categorical data into dummy variables. Then using descriptive analysis, I had to figure out which attributes in the dataset would be meaningful when prediction models were built with them. I found a few quantitative variables that were significantly correlated to my dependent variable, "log_price". Also, frequency table displayed some insightful information from the categorical variables. Next, clustering and association rules revealed meaningful relationships between variables and the risk of multicollinearity. With the information obtained from the descriptive analysis, different kinds of model were built in order to estimate the accurate price of AirBnb. Each models provided quite accurate results although none of them was able to demonstrate the perfection. For the multiple linear regression, it showed the value of 63% of R2, which indicated the 37% of variability in sample price unexplained. Furthermore, K-Nearest Neighbor model showed the value of 0.459 for RMSE. Both models provided good estimation, but still was not enough to be used with full credibility. Therefore, it would be better to use those models as references when predicting the price of AirBnb.