Team 2 Report

Members: Aditi Singhal, Apoorva Shetty, Wenbo Zhu, Zoheb Siddiqui

Research questions:

a) With the data provided we are tying to understand the trend of AIRBNB listing price by checking seasonal effect and variation across 4 years and 12 months for 6 metropolitans.

b) How are Airbnb listing pricing explained by different characteristics and how important is neighbourhood (location).

Analysis a):

To answer the first question, we used the **calendar** table to summarize the monthly median price for each property. Indices of the real estate market (i.e., ZHVI and ZRI from **real_estate** table) were also incorporated in the dataset as control variables. As the Airbnb price and trend can be different across regions, each property is mapped with metropolitan information from **listing** table. A linear regression model was trained to quantify the trend of Airbnb listing price as well as to understand the contribution of different factors. Two time-related variables are included in the model predictors: year and month. While year is used as a numerical variable, month is coded as a categorical factor to account for seasonal effects. The overall distribution of median price is shown in Figure 1(1). Note that this is a right-skewed distribution, which violate the normality assumption of linear regression. Thus, we log-transformed the monthly median price and the result distribution is in Figure 1(2). There is no obvious violation of normality assumption after data transformation.

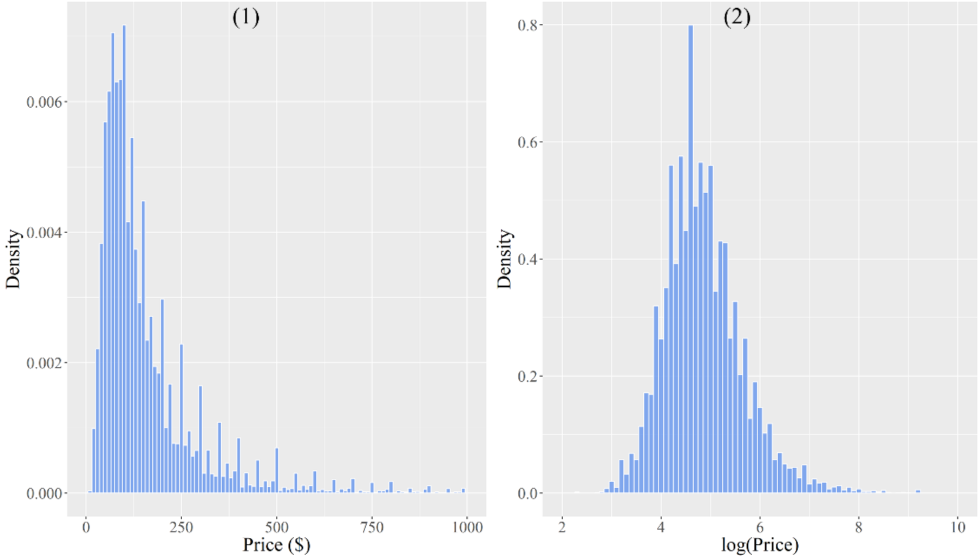


Figure 1. Distribution of Airbnb Listing Price

After trying different input settings and we found that the interaction between year and metropolitan is not very significant, indicating that the price trend over year is similar across regions. Our final model input includes year (numeric), month (factor), metropolitan (factor), ZHVI (numeric), and ZRI (numeric). The regression result is presented in Table 1. In general, there is a significant increasing trend of Airbnb price with the average yearly increase of 2% after controlling all other factors (i.e., season, location, and growth of real estate market). In terms of the seasonal effects, summer months are associated with higher Airbnb listing price (with July being the highest). January (the reference level) appears to have the lowest median price, although is not significantly different from February. The regression result also suggests significant location impacts, as indicated by the significant coefficient estimates for metropolitan factor. Both real estate market indices are positively associated with Airbnb price. The positive coefficient estimate of year also suggest that the increase of Airbnb price is faster than the real estate market.

Table 1. Multiple Linear Regression Results

	Estimate	Std.Error	T- Statistics	p value
(Intercept)	-3.59E+01	1.62E+01	-2.223	0.0262
year	1.98E-02	8.01E-03	2.470	0.0135
month02	4.94E-04	1.06E-02	0.047	0.9629
month03	2.30E-02	1.06E-02	2.175	0.0296
month04	2.04E-02	9.60E-03	2.128	0.0333
month05	4.18E-02	9.07E-03	4.604	4.15E-06
month06	8.85E-02	9.11E-03	9.713	2.00E-16
month07	1.44E-01	1.20E-02	11.969	2.00E-16
month08	1.09E-01	1.20E-02	9.088	2.00E-16
month09	5.42E-02	1.20E-02	4.536	5.74E-06
month10	2.30E-02	1.19E-02	1.926	0.0541
month11	3.90E-02	1.24E-02	3.153	0.0016
month12	4.67E-02	1.23E-02	3.796	0.0001
Metropolitan: Oakland	-2.15E-02	1.05E-02	-2.051	0.0402
Metropolitan: Portland	8.51E-02	9.73E-03	8.749	2.00E-16
Metropolitan: San_Diego	4.76E-01	7.04E-03	67.622	2.00E-16
Metropolitan: San_Francisco	2.66E-01	7.80E-03	34.070	2.00E-16
Metropolitan: Santa_Cruz	5.94E-01	1.25E-02	47.437	2.00E-16
Metropolitan: Seattle	2.59E-01	9.19E-03	28.185	2.00E-16
ZHVI	1.36E-07	2.64E-08	5.157	2.51E-07
ZRI	1.38E-04	8.29E-06	16.698	2.00E-16

Analysis b)

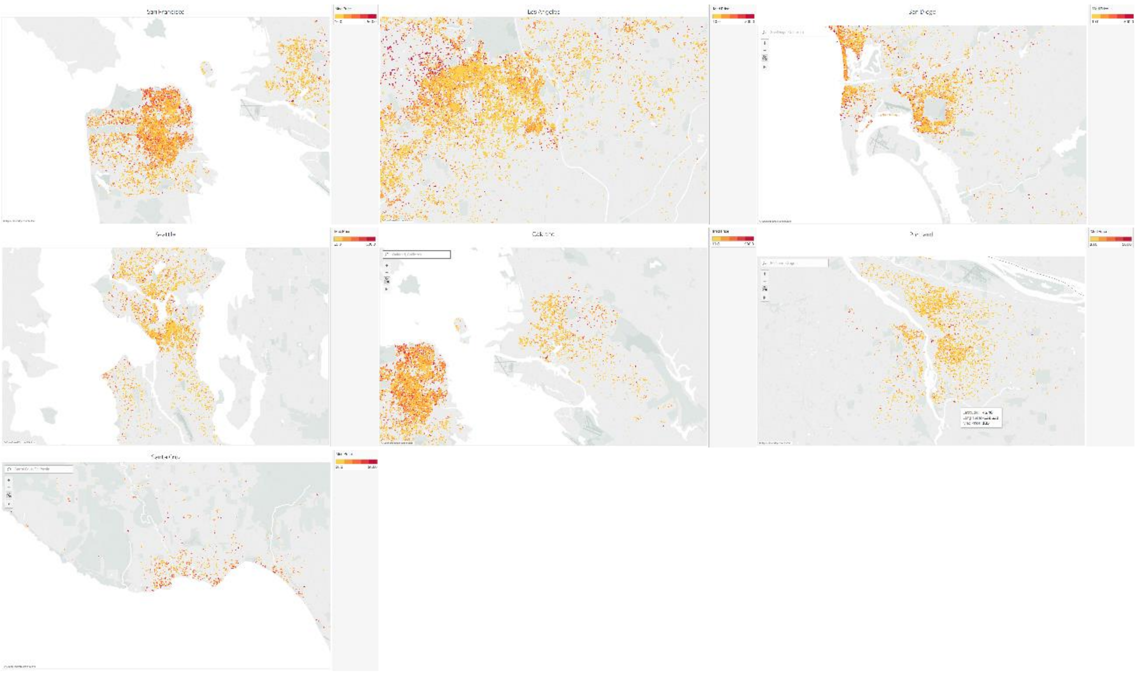


Figure 2. Spatial Distribution of Airbnb Listing Price

Figure 2 shows the spatial distribution of Airbnb listing price. There is a significant neighbourhood pattern behind the data. To answer second question, we used variables in the **listing** table to fit a random forest regressor, and the importance of each variable can be evaluated by the mean decrease in accuracy if using a permutation of the variable. First, we replaced the categorical factors (e.g., bed type, cancellation policy, property_type, etc.) with binary variables with each associated with a specific level of those factors. Then we performed the train-test split. The 10-fold cross validation was utilized to identify the best hyperparameters for the model. The best hyper parameters identified were:

- Bootstrap = True
- Minimum Samples to Split By = 20
- Number of nodes used in the forest = 200

Figure 3 shows the 10 most important variables in the final random forest model with their relative importance. According to the result, the most important factor that drive the pricing of Airbnb property is the number of bathrooms. Additionally, the number of bedrooms and geographic coordinates are also important in explaining the difference of Airbnb listing price. The importance of geographic coordinates indicates strong pattern Airbnb price at different neighbourhood.

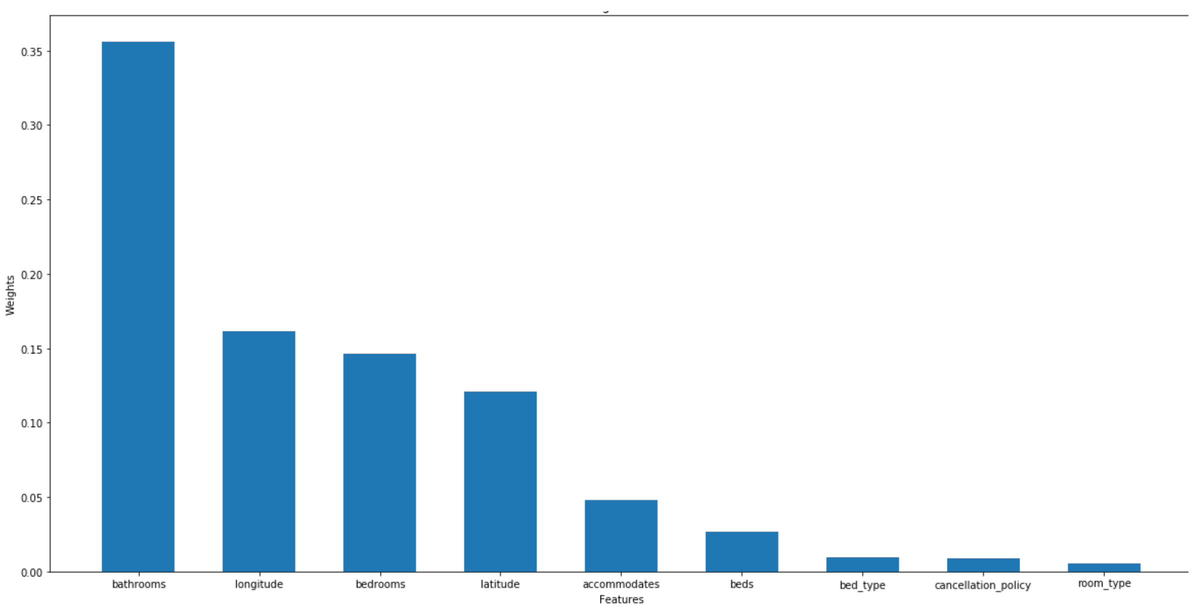
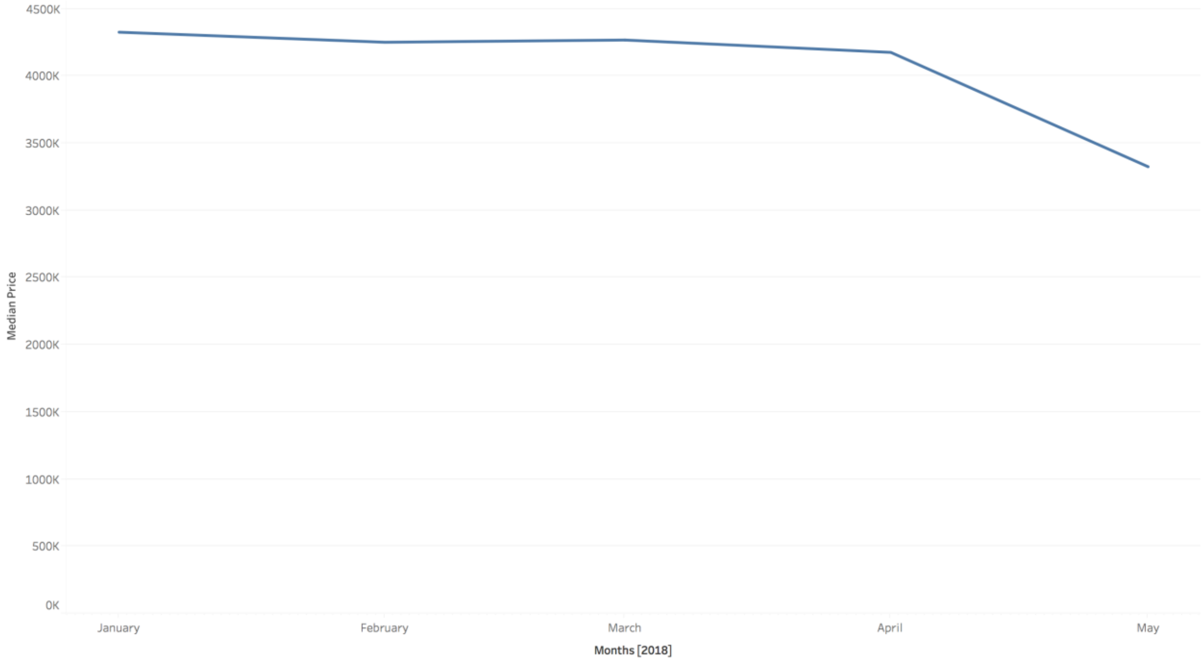


Figure 3. Relative Importance of Variables in Random Forest Model

2018



The trend of sum of Med Price for 2018 Month.