Final Airbnb Regression Analysis

Anisha Prashanth, Anish Umredkar, Brian Chen

2022-11-30

I. Introduction

When picking an Airbnb, we prioritize some things and compromise on others. We are going to look at Airbnb data which consists of many possible factors that people consider when selecting an Airbnb. Using regression analysis, we will determine if and how they are related to each other. In other words, we are looking for possible correlations such as whether a closer distance to the city is associated with a higher listing price.

II. Problem/Goal Statement

Knowing that a reasonable price is one of the biggest priorities when choosing an Airbnb, we want to find out what factors significantly affect an Airbnb's price in San Francisco. Some variables that we examined are the distance between the Airbnb and the center of San Francisco, the accommodations of the Airbnb, the review scores of the Airbnb, and many other variables.

III. Data Description

This data was found from Inside Airbnb, a website that provides Airbnb data from various cities. The original data consisted of four separate files including the calendar dates of listings, the location of the property, and finally the accommodations of the listing (number of bedrooms and bathrooms). There were a total of 6,000 listings and over 1,000,000 observations in the calendar file. There were approximately 75 possible variables in the listing file, however, we narrowed our analysis down to only a few variables. We combined data from these various data sets and cleaned this data to ensure all listings had complete information (the cleaning process is shown in the appendix). In the end, we had 11 independent variables with the hopes of determining how they affected the price of a listing.

Our primary dependent (y) variable was price. Here are our independent (x) variables:

$$\hat{price} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_i * x_i + ... + \beta_{11} * x_{11}$$

This is a basic model we want to start our analysis with. As we go through different regression techniques, we want to eliminate non-significant independent variables and revise the model.

	Name	Description
x_1	Host_is_superhost	Dummy variable (Qualitative). 1 if true, 0 if false.
x_2	Accommodates	Quantitative.

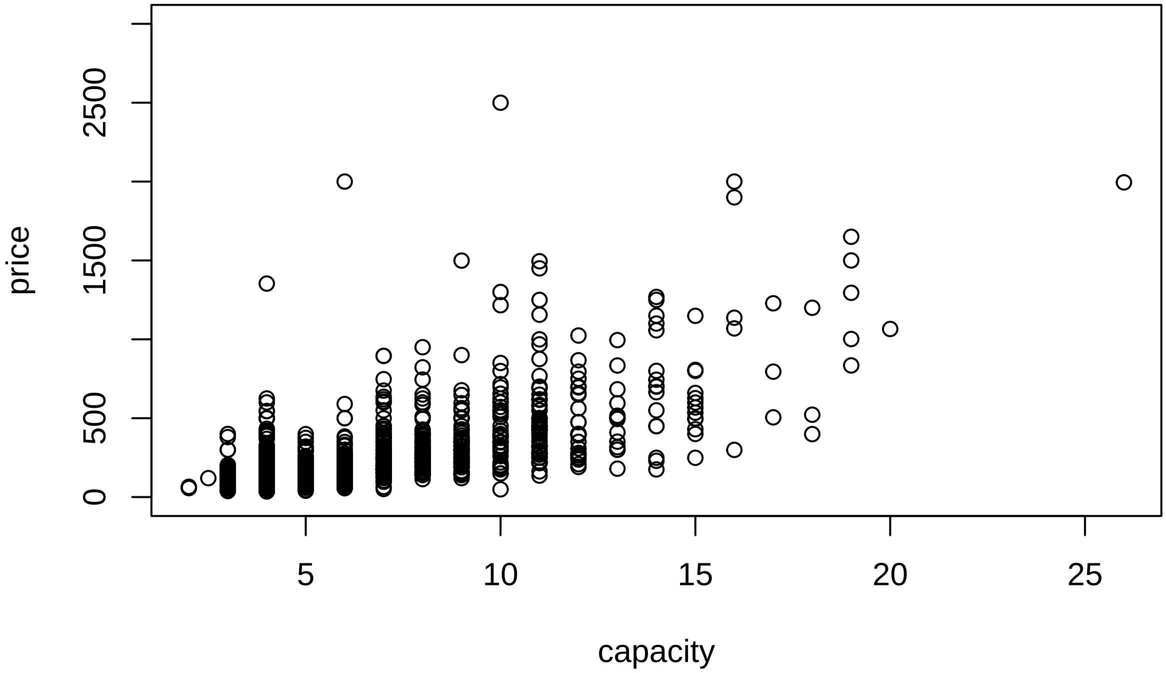
	Name	Description
x_3	Bathrooms	Quantitative.
x_4	Bedrooms	Quantitative.
x_5	Review_scores_rating	Quantitative. Ratings are between 1-5. The data gives the average rating for each listing.
x_6	Host_identity_verified	Dummy variable (Qualitative). 1 if true, 0 if false.
x_7	Review_scores_communication	Quantitative. Ratings are between 1-5. The data gives the average rating for each listing.
x_8	Number_of_reviews	Quantitative.
x_9	Amentites_score	Quantitative. Number of amenities.
x_{10}	Distance_city	Quantitative. Distance from a central point in the city. Determined using the neighborhood of the listing.
x_{11}	Peak	Dummy variable (Qualitative). Whether the listing date was during a peak time (new years, long weekends, etc.) or not. 1 if true, 0 if false.

IV. Regression Analysis

I. Plots of variables

Here, we grouped together our independent variables to form four graphs. In this section, we are exploring these four categories and looking to gain a basic intuition about how they behave relative to price or independently.

1) Capacity (accommodates, bedrooms, bathrooms)



There is a moderate positive correlation apparent in this scatter plot. We can infer that as the capacity is higher, the price is higher. This is consistent with our intuition.