**The U.S. Airbnb Rental House Price Prediction for 2020**

Samip Thapa ([s\_thapa2@uncg.edu](mailto:s_thapa2@uncg.edu))

Divya Sanga ([dksanga@uncg.edu](mailto:dksanga@uncg.edu))

University of North Carolina at Greensboro

IAF-603 Preparing Data for Analytics

Dr. Austin Lawson

December 07, 2020

**Abstract**

This project aims to predict the rental house prices of different regions in the United States. The Airbnb dataset consists of 226030 observations and 17 variables. During the data preparation and data cleaning process we will remove unnecessary variables and will handle the missing values using replacement or deletion methods. Additionally, outliers will be assessed and replaced/removed accordingly. Data storage methods will be carried out through the creation of a SQL database in python. Through queries and visualization on this database, we will assess potential questions about the data.To predict the rental house prices, we will use a linear regression model, backward feature selection, and finally evaluate the model using the results of root mean squared error (RMSE).

**Keywords:** Modeling the data, Querying, Linear regression, Backward feature selection, Root mean squared error (RMSE).

**INTRODUCTION**

Airbnb (AirBedandBreakfast) is an online marketplace that lets people rent out their properties or spare rooms to guests. The company is based in San Francisco, California, United States. Since its foundation in 2008 by Brian Chesky, Joe Gebbia, and Nate Blecharczyk, it has become one of the topmost used accommodation services throughout the United States and around the globe. Airbnb is a cheaper alternative to hotel room bookings because these are residential properties and homeowners are not required to register or pay sales taxes unlike hotels. Over the past few years, it has listed over a staggering 800,000 properties in 34,000 cities across 90 different countries. The tourists who need to use Airbnb services must pay some rental fees according to the type and features of the accommodation needed.

This project aims to predict the U.S. Airbnb’s house rental price per night in 28 cities within the United States based on various factors. The prediction of the rental prices would be beneficial for many tourists and other potential users of this service. One of the benefits include early financial planning as they would have the opportunity to have some idea about the rental prices in various cities within the United States.

**Description of Dataset**

The Airbnb dataset was compiled on 20th October 2020 by [Kritik Seth](https://www.kaggle.com/kritikseth), who compiled the data of the U.S. cities’ data from the Inside Airbnb platform. The dataset consists of 226030 observations and 17 variables (features) that includes host id, hostname, listing id, listing name, latitude and longitude of listing, the neighborhood, price, room type, minimum number of nights, number of reviews, last review date, reviews per month, availability, host listings, and city. The Airbnb listings of 28 different cities were given in the dataset. In this study, price will be treated as the dependent variable and all other variables will be treated as independent when performing regression analysis. Table 1 below shows a list of the dataset variable names used in this study with their description and the variable types.

|  |  |  |
| --- | --- | --- |
| Dataset Variable Names | Variable Description | Variable type |
| id | unique listing id | Numerical |
| name | name of listing | Categorical |
| host\_id | unique host Id | Numerical |
| host\_name | name of host | Categorical |
| neighbourhood\_group | group in which the neighbourhood lies | Categorical |
| neighbourhood | name of the neighbourhood | Categorical |
| latitude | latitude of listing | Numerical |
| longitude | longitude of listing | Numerical |
| room\_type | room type | Categorical |
| price | price of listing per night | Numerical |
| minimum\_nights | minimum no. of nights required to book. | Numerical |
| number\_of\_reviews | total number of reviews on listing | Numerical |
| last\_review | date on which listing received its last review | Ordinal |
| reviews\_per\_month | average reviews per month on listing | Numerical |
| calculated\_host\_listings\_count | total number of listings by host | Numerical |
| availability\_365 | number of days in year the listing is available for rent | Numerical |
| city | region of the listing | Categorical |

**Table 1**: Table showing the different variable names used in this study with their description and their types

**Research Questions**

Based on this data set we are going to address some of these research questions that includes:

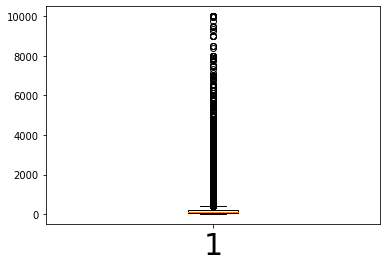
1. Is there any relationship between latitude and price?
2. Is there any relationship between number of reviews and price?
3. How does the price per night differ by room type?
4. What is the maximum and minimum price per night?
5. How many total listings are there for the cities with a minimum number of nights between 14 and 21 days and what are these cities?
6. How are the listings whose number of days in a year greater than 180 days that is available for rent distributed and which cities fulfill this criterion?
7. How are the number of listings in different cities distributed?

**Methods**

The raw data of the U.S. Airbnb was first uploaded into Google Colab where all of the analysis for this study was performed. By looking at the dataset we have identified that there are some missing values in the columns. First of all, all the irrelevant/unnecessary variables for this study were dropped. These irrelevant variables were identified to be id, host\_id, neighbourhood\_group, host\_name, last\_review. Next, there were missing values for the variable reviews\_per\_month which were replaced with its mean since the distribution of this variable was not heavily skewed. Furthermore, there were some observations where there were missing name values. The rows with missing names were also dropped. Additionally, there were some observations with price per night less than $10 and more than $10000 which were removed since Airbnb only allows listing of prices per night between $10 and $10000. After that, outliers were assessed for the numerical variables that were left after cleaning the missing values. For these variables, the outliers were replaced by its median values. “It is advised to not use mean values as they are affected by outliers” (Singh, 2019). Thereafter, the cleaned-up data was stored in a SQL database with now containing 225645 observations and 12 variables.

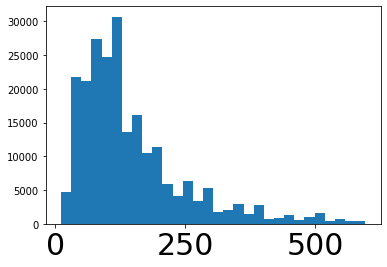
**Analysis**

The data analysis steps consisted of first proposing the research questions. Secondly, the data was gathered from credible sources and extracted for analysis maintaining the ethical considerations of data extraction. Thereafter, data preparation was carried out by cleaning up messy and missing data followed by handling outliers which were assessed through visualization techniques such as box plots for each numerical variable. An example of outlier detection carried out on price variable through visualization technique is shown in figure 1 below.



**Figure 1: Identification of outliers for price variable**

Following the data cleaning and transformation processes, the data was stored in a SQL database where queries were performed using SQL commands such MIN, MAX, COUNT, DISTINCT, GROUP BY, ORDER BY, etc to make initial discoveries and come up with answers to the proposed questions. Additionally, data visualization methods were performed from these queries making the use of graphs such as histograms, box plots, bar plots, scatterplots, and cat plots. An example of a visualization method performed from the query results for the price variable is shown in figure 2 below.



**Figure 2: Histogram plot for price after the data cleaning**

To predict the price per night, the variable room\_type went through one hot code encoding that led to four features being added (room\_type\_Entire home/apt, room\_type\_Hotel room, room\_type\_Private room, and room\_type\_Shared room) to the data. The variable city went through label encoding. Encodings were performed on these two categorical variables as they were deemed to be useful predictors for price. Next, a Spark dataframe was created from the processed data. Then, the features name and neighborhood were dropped as they were considered not useful explanatory variables to predict price. The data was then split into 80% training and 20% testing data. A backward feature selection was carried out to select the best model. And, the scoring metric used for backward feature selection was mean square error (MSE). From the best selected features, a linear regression model was carried out and the evaluation metrics RMSE and were calculated.

**Results**

In assessing whether there was a relationship between latitude and price per night, it was found that there was not a clear relationship between those two variables. Similarly, there was not a clear relationship between number of reviews and price per night. Looking at the room type and price, the price per night for hotels were higher compared to other room types and the price per night for shared rooms were lower compared to other room types as expected. The maximum price per night for a listing was found to be $597 whereas the minimum price per night was found to be $11. New York City had the highest number of listings and Pacific Grove had the lowest number of listings with a minimum number of nights between 14 and 21 days. All of the 28 cities had at least one listing fulfilling this criterion. Salem and Pacific Grove cities had a lower number of listings whose availability of rent was greater than180 days compared to others. All of the 28 cities had at least one listing with availability of rent greater than 180 days. New York City was found to have the highest number of Airbnb listings with a count of 45701 listings and Pacific Grove had the lowest number of listings with a count of 177 listings.

From the backward feature selection method, it was found that all of the 12 explanatory variables used for predicting price provided the best MSE score of 8793.54. The value from the linear regression model was found to be 0.194. This shows that the model used explains 19.4% of the variance in predicting price around its mean. Additionally, the RMSE value obtained from fitting the linear regression model was 93.734.

**Conclusion**

This project predicted the U.S. Airbnb’s house rental price per night of 28 cities within the United States based on 12 explanatory variables. Additionally, other questions about the dataset were proposed that included assessing relationship between latitude and price, relationship between number of reviews and price, price difference according to room type, maximum and minimum price per night, distribution of number of listing having availability greater than 180 days, counts of listing with minimum number of nights between 14 and 21 days, and the distribution of total counts of listings. For these purposes data cleaning, transformation, and integration measures were performed.

New York City offered the most Airbnb listing with various price ranges and Pacific Grove offered the least. The value was found to be low and the RMSE value was found to be quite high from the regression analysis. Thus, other confounding variables such as number of bedrooms, number of bathrooms, number of beds, number of guests rental can accommodate, etc could have been other important predictors for prices that were not available in the dataset.

**References**

Link of the dataset:<https://www.kaggle.com/kritikseth/us-airbnb-open-data>

Singh, D. (2019, October 22). Deepika Singh. Retrieved December 05, 2020, from <https://www.pluralsight.com/guides/cleaning-up-data-from-outliers>