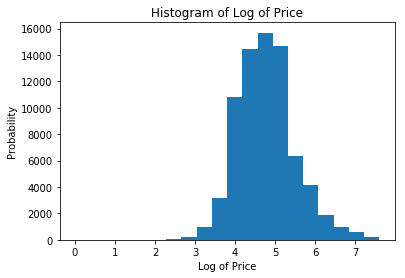
**Data Details**

* **Data source:** Kaggle Dataset (<https://www.kaggle.com/rudymizrahi/airbnb-listings-in-major-us-cities-deloitte-ml>), originally used for a machine learning competition. We only used the training dataset which contains the target variable.
* **Number of observations:** 74,111
* **Description of variables:** There are 29 columns in the original dataset. The first column is id, the second column is the target variable *log\_price*, and the rest 27 variables are features for the listing. The types of data include integer, float, object and Boolean. The dataset contains both categorical and numerical variables. Details can be found in Table 1 below.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Notice** |
| property\_type | Object; categorical |  |
| room\_type | Object; categorical |  |
| amenities | Object | Contains NA |
| accommodates | Integer; numerical |  |
| bathrooms | Float; numeric |  |
| bed\_type | Object; categorical |  |
| cancellation\_policy | Object; categorical |  |
| cleaning\_fee | Boolean; binary |  |
| city | Object; categorical |  |
| description | Object; string |  |
| first\_review | Object; date | Contains NA |
| host\_has\_profile\_pic | Object; binary | Contains NA |
| host\_identity\_verified | Object; binary | Contains NA |
| host\_response\_rate | Object; string | Contains NA; can be transformed to float |
| host\_since | Object; date | Contains NA |
| instant\_bookable | Object; binary | Can be transformed to boolean |
| last\_review | Object; date | Contains NA |
| latitude | Float; numerical |  |
| longitude | Float; numerical |  |
| name | Object; string |  |
| neighbourhood | Object; categorical | Contains NA |
| number\_of\_reviews | Integer; numerical |  |
| review\_scores\_rating | Float; numerical | Contains NA |
| thumbnail\_url | Object; string | Contains NA |
| zipcode | Object; string | Contains NA |
| bedrooms | Float; numerical | Contains NA |
| beds | Float; numerical | Contains NA |

Table 1

* **Distribution of the target variable:**



count 74111

mean 4.782069

std 0.717394

min 0.000000

25% 4.317488

50% 4.709530

75% 5.220356

max 7.600402