**Improving Rental Price in the United States**

Chun-Li Hou

**Introduction**

The client is the management in the rental industry. Comparing the United States and China, the people in the United States prefer renting a place than buying it for living. According to online research, it is only 35% of Americans under 35 years old who are willing to buy their own house. On the other hand, the proportion in China is more than 70%. We hereby know that the rental industry is relatively potential in the United States. Thus, how to set a profitable price and maximize profits are important issues to the management in the rental industry.

By analyzing the data, it can help us to clear several questions: (i) what regression model can we use to predict the most accurate rental price? (ii) what property attributes are important and interesting to decide the retail price?

By answering these questions, we can bring the management in the rental industry some interesting insights, which can help them to be profitable.

**Hypothesis**

Question i. What regression model can we use to predict the most accurate rental price?

H0: all coefficients are zero

H1: one of the coefficients is nonzero

Question ii. What property attributes are important and interesting to decide the retail price?

H0: the coefficient beta is zero

H1: the coefficient beta is nonzero

\*For multiple linear regression model & polynomial non-linear regression model use

**Data Description**

The data comes from the machine learning repository in the University of California at Irvine (UCI). The data is about rental price in the United States. It is collected from the web crawling through 12 different rental websites in the United States, such as GoSection8.com, HomeRentals.com, Listanza.com, ListedBuy.com, RealEstateAgent.com, etc. The time point is on December 28, 2019. After cleaning the data with no empty in the column of price, it has 22 columns and 10,000 rows. The data is shown in a glimpse in Figure 1. The attribute information is shown in a table in Figure 2. Since there are still some attributes as a character type that cannot be used in the model, we will proceed with some data cleaning processes more. The detailed processes will be introduced in the next paragraph.

**Data Preparation**

We decide to break some columns for more useful attributes, drop some meaningless columns, and keep some important columns as the original.

Firstly, we take price as the dependent variable. Furthermore, we transform price based on price\_type as monthly in our standard. The column of price\_type has monthly, weekly, and monthly/weekly. After checking, we turn the weekly into monthly due to the simple typing error for the original data. The monthly/weekly is a mistake after making sure by searching on Google. We find the property with a price in weekly. Thus, we take the price and divide it by 7 then multiple it by 30.4, which is the average days per month. In summary, we correct price by standardizing with price\_type.

Secondly, as for the usable independent variables, these are id, bathrooms, bedrooms, and square\_feet, source. Thirdly, we perform some breakdown or transformation to generate extra variables. We break down category into three types, including cat\_apart, cat\_home, and cat\_short.

Then, we extract amenities into special features that are more possible to be concerned when renting a property. The extra generated attributes are clubhouse, fireplace, gated, balcony, sport, gym, pool, parking, laundry, tv, elevator, and wood\_floor. As for the rest features in amenities, they are considered as a standard facility for all kinds. The column of pet\_allowed is factorized.

The column of metro\_area is based on whether the property city is in the top 10 largest metropolitan areas, which refers to the website in reference. The web data is from the United States Bureau of Economic Analysis that estimates the GDP for each United States metropolitan area in 2016. The process also corrects some data with the null value. By using the column of latitude and longitude on Google Map and the column of body with a manual check, we define 77 null values with their city name. The preparation processes conduct in Excel. These non-numeric variables will be set as a factor type to use in the model.

**Data Preprocess**

After data preparation, we can start importing data into R to preprocess data. Firstly, the data in a glimpse is in Figure 3 and the column description table is in Figure 4. After that, we check whether we have any missing data. We decide to drop the observations with missing values. Then, we pick up the near-zero variable to drop due to a lack of variance contributing to the model. Our ready-to-go data has 17 columns and 9,961 rows.

Finally, we remove source and id for no use in the model. Also, we remove the outliers. And, we do data partition to have a train set for modeling fitting and a test set for model evaluating. At the end, the train set has 7,968 rows and the test set has 1,990 rows with 15 columns.

**Exploratory Data Analysis**

Due to having more factor variables than numeric variables, it is not workable to perform the correlation matrix plot. Thus, instead of that, we use ggpairs function to show up all possible demonstrations showing in Figure 5. We can quickly know about the correlation between a pair of variables. For example, the strongest correlation with price is square\_feet. Then, we take a few further steps to see in different plots with interesting variables.

Firstly, we see the distribution of each numeric variable. The distribution for price or square\_feet is more normal distributed after log transformation in Figure 6 and Figure 7. The distribution for bedrooms or bathrooms is skewed but match up with our expectation in Figure 8 and Figure 9. Overall, there is no unexpected sign with these numeric variables.

As for the factor variables, we use two sample t-test to see the price difference. As setting the significant level as p < 0.001, with the two sample t-test, we find that gym, laundry, elevator, and pets\_allowed are statistically the same on price between having or not having it.

Furthermore, we try to combine more dimensions with more variables together to plot out to process datamining. For example, based on the scatterplot in Figure 10, we can clearly see the correlation between bathrooms and bedrooms. Also, price in log and square\_feet in log are correlated with either bedrooms or bathrooms.

**Analysis**

In the first step, we use seven different regression machine learning algorithms with the train set to find out the best regression model. We use the test set to evaluate the different regression models. From Figure 11, we know that the best regression model is the multiple linear regression model (MLR) and polynomial nonlinear regression model (PNLR) due to the lowest root mean square error (RMSE).

We start from the PNLR model. After checking assumptions, we find that the original PNLR model violates some of those. Therefore, we try to fix these violations in some ways. After deleting outliers, transforming the dependent variable, and removing independent variables, the PNLR model fits all of the assumptions, which are outlier, heteroscedasticity, collinearity, and normality. But, the model is more as the MLR model. The results demonstrate in Figure 12 and Figure 13.

This linear regression model is a log-linear regression model. We start to interpret the model. We can see the estimates and the coefficients in Figure 14. Only the estimates of bedrooms, fireplace, and pets\_allowed are not significant. The multiple R squared is 0.332, which means only 33.2% of the dependent variable (price) can be explained by the model inputs (the independent variables). The adjusted R squared is 0.331. We take an estimate as an example of interpreting its coefficient to establish causality. As increasing the property’s square feet by 100 units, the price of the property increases by 4%.

Lastly, we can start answer our business questions and provide insights. The first insight is as considering exercise facility, which kind does positively affect the price? We can see the estimates of sport and gym. We will know that having the outdoor sport facility (sport), such as basketball field or tennis court, the price decreases by 13.51%. On the other hand, having the indoor sport facility (gym), such as gym or fitness center, the price increases by 11.67%. Thus, we know that the indoor facility would bring more profit to the apartment value.

The second insight is to answer what apartment attribute affected the price is the most interesting. We know the location can affect the price easily. Our model also identify this point. The second highest coefficient is elevator, which means whether the apartment has an elevator or not. This is an interesting insight. In detail, having an elevator in the apartment, the price increases by 22.31%. As we know, living in which floors has its pros and cons. Therefore, having an elevator might not be an impact on price. However, from our model result, we can say having an elevator would bring positive impact on the price.

The last one insight is how much should the new releasing apartment set the price? We use this model for real-world validation. We search the crawling websites for a new releasing apartment. The property is 808 square feet, with 1 bedroom and 1 bathroom, locates in Richardson, TX, has gym and parking area, and is pet allowed. The predicted price is $1,225. Thus, with this model, we can provide a reasonable price for a new releasing apartment.

**Summary**

After the analysis, we know the multiple linear regression as the log-linear regression model is our best regression model on this data. We can use this model to help the management in the rental industry to set a profitable price and maximize profits. For example, we can use this model to predict the accurate and reasonable rent for a new releasing apartment. Furthermore, we generate some interesting insights. Those can help the management to have better profits.

In conclusion, the rental industry is thriving regardless of economic fluctuations. The need for a profitable price is important to keep the industry thriving. In this research, we can connect the machine learning model selection and the best selected one model to help to identify the profitable price and to provide interesting insights for making necessary business decisions in the rental industry.

**Appendix**

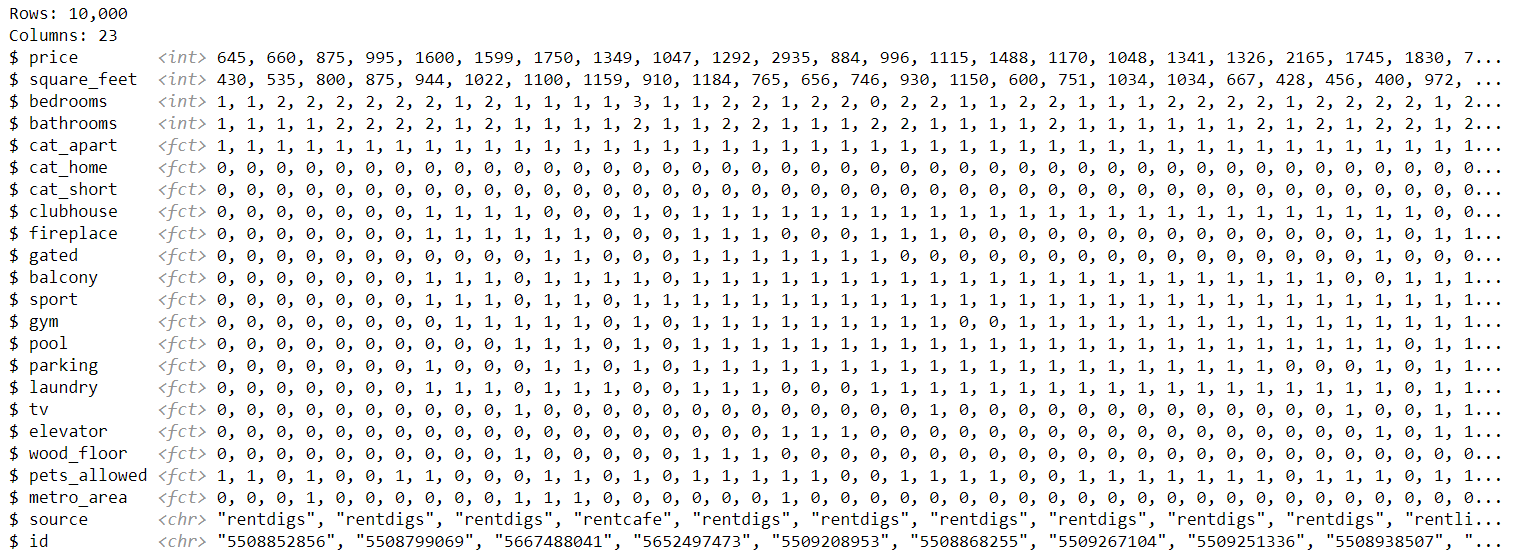
Figure 1. The original data has 22 columns and 10,000 rows.



Figure 2. The original data attribute information shows in descriptive.

|  |  |
| --- | --- |
| Name | Description |
| id | unique identifier of apartment |
| category | category of classified |
| title | title text of apartment |
| body | body text of apartment |
| amenities | like AC, basketball, cable, gym, internet access, pool, refrigerator etc. |
| bathrooms | number of bathrooms |
| bedrooms | number of bedrooms |
| currency | price in current |
| fee | fee |
| has\_photo | photo of apartment |
| pets\_allowed | what pets are allowed dogs/cats etc. |
| price | rental price of apartment |
| price\_display | price converted into display for reader |
| price\_type | price in USD |
| square\_feet | size of the apartment |
| address | where the apartment is located |
| cityname | where the apartment is located |
| state | where the apartment is located |
| latitude | where the apartment is located |
| longitude | where the apartment is located |
| source | origin of classified |
| time | when classified was created |

Figure 3. The prepared data has 23 columns and 10,000 rows.



Figures 4. The prepared data attribute information shows in descriptive.

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| price | int | price of apartment in dollar |
| square\_feet | int | size of apartment |
| bedrooms | int | number of bedroom |
| bathrooms | int | number of bathroom |
| cat\_apart | fct | apartment |
| cat\_home | fct | home |
| cat\_short | fct | short-term |
| clubhouse | fct | clubhouse |
| fireplace | fct | fireplace |
| gated | fct | gated |
| balcony | fct | patio/deck |
| sport | fct | field/court/playground |
| gym | fct | gym |
| pool | fct | pool |
| parking | fct | parking |
| laundry | fct | washer/dryer |
| tv | fct | tv |
| elevator | fct | elevator |
| wood\_floor | fct | wood floor |
| pets\_allowed | fct | pets allowed |
| metro\_area | fct | metropolitan area |
| source | chr | origin website source |
| id | chr | unique identifier |

Figure 5. The overall plot shows the general explanatory data analysis information.

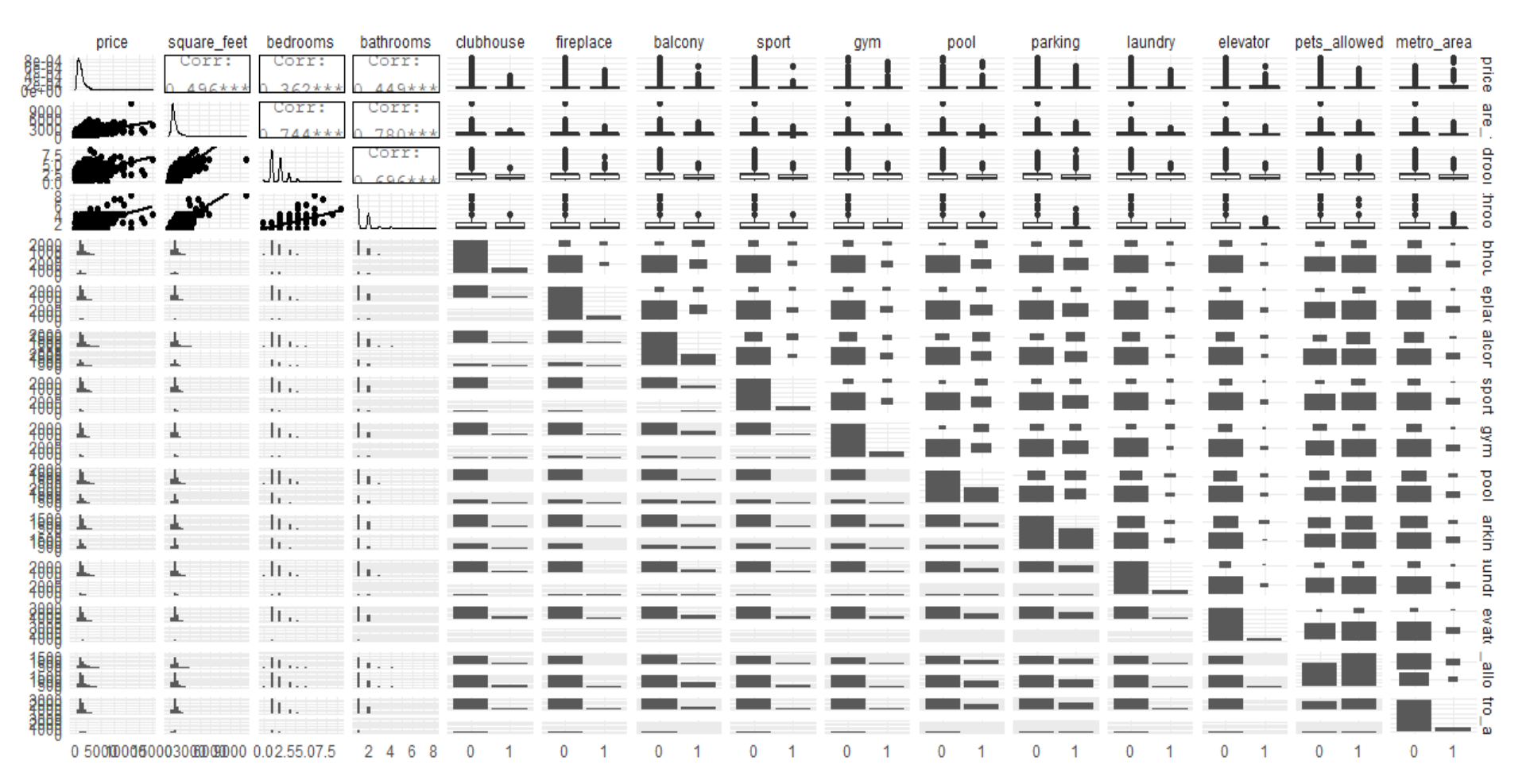


Figure 6. The distribution of price in log-transformation is normal distributed.

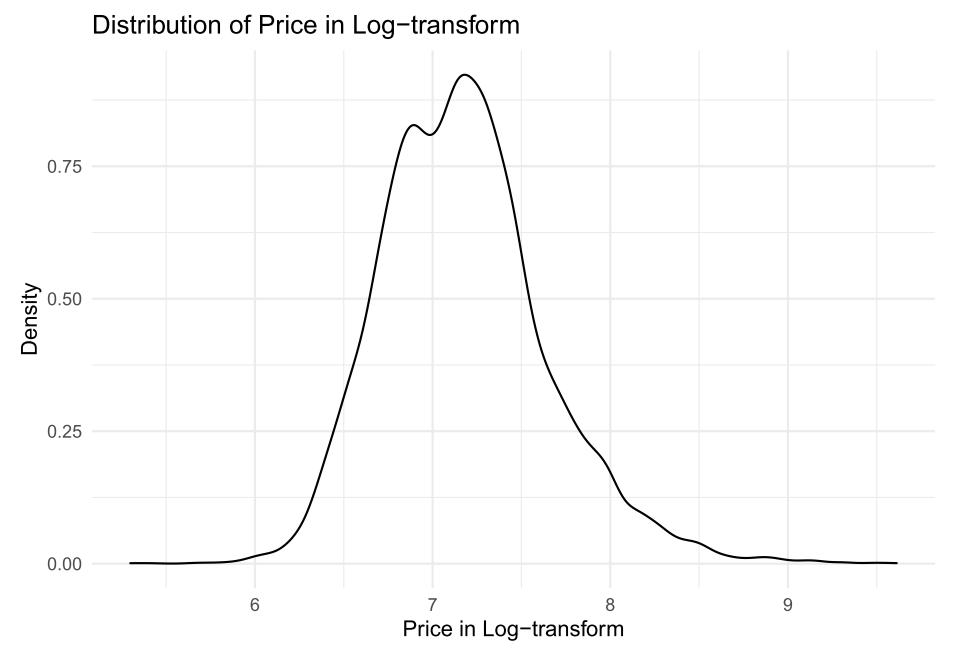


Figure 7. The distribution of square feet in log-transformation is normal distributed.

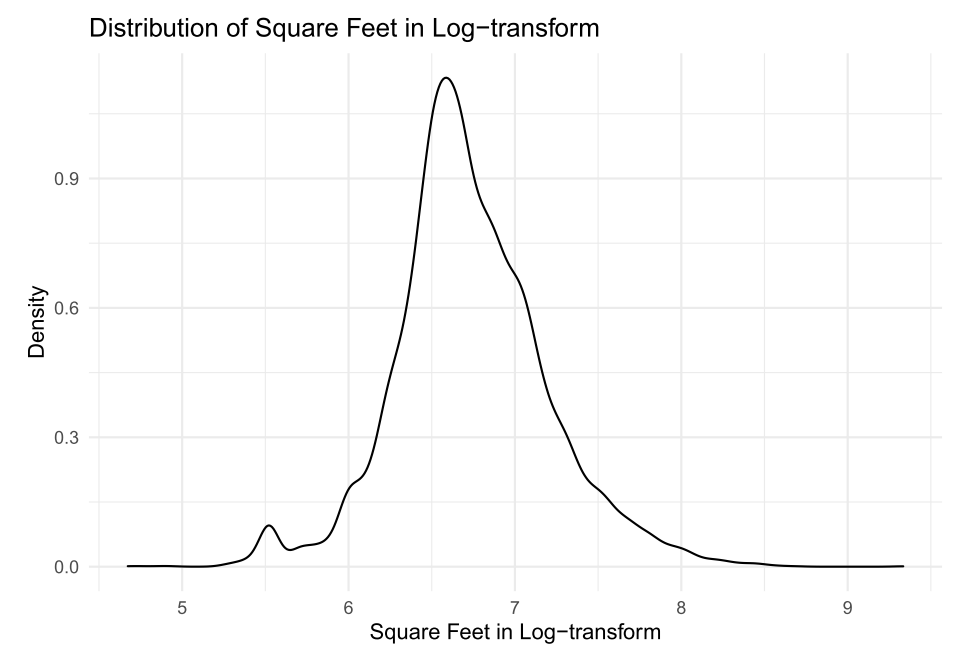


Figure 8. The barplot of number of bedrooms is skewed.

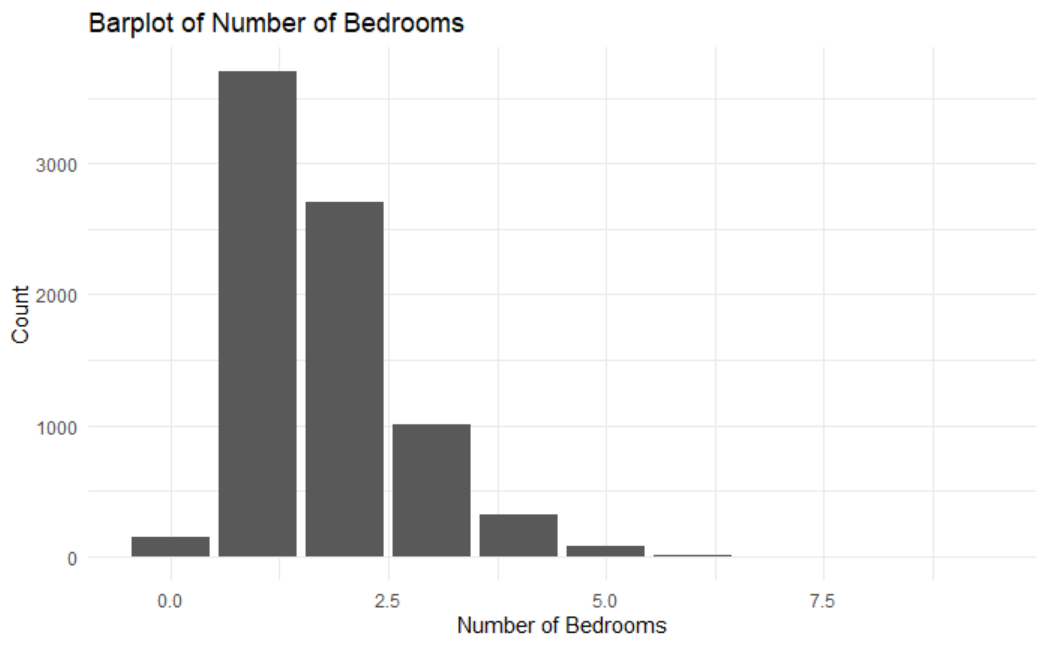


Figure 9. The barplot of number of bathrooms is skewed.

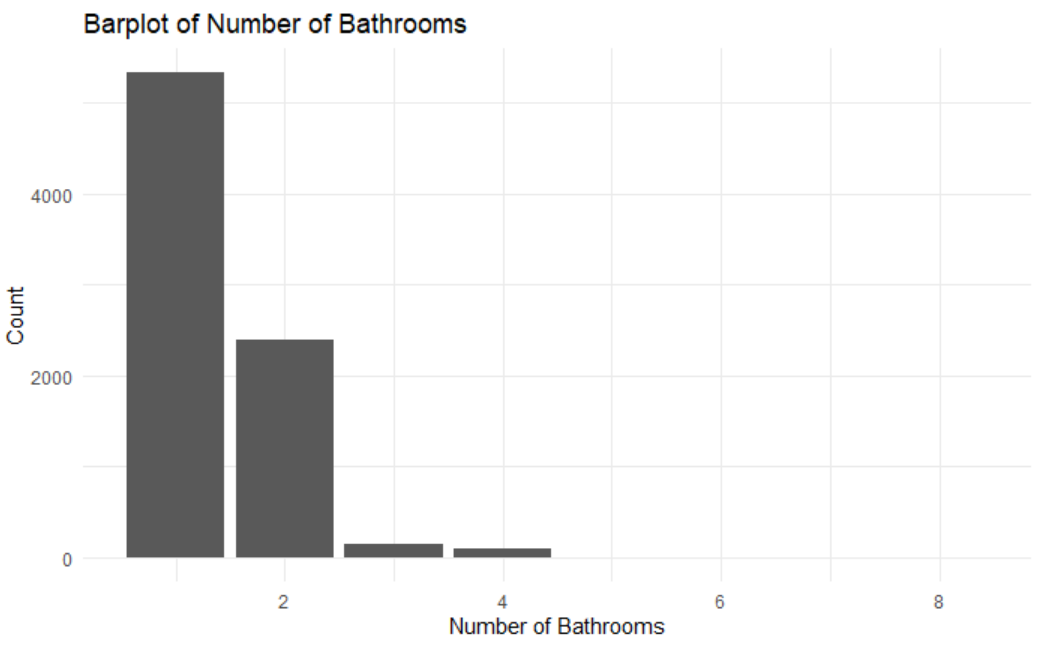


Figure 10. The combination of numeric variables plot shows a correlated relationship.

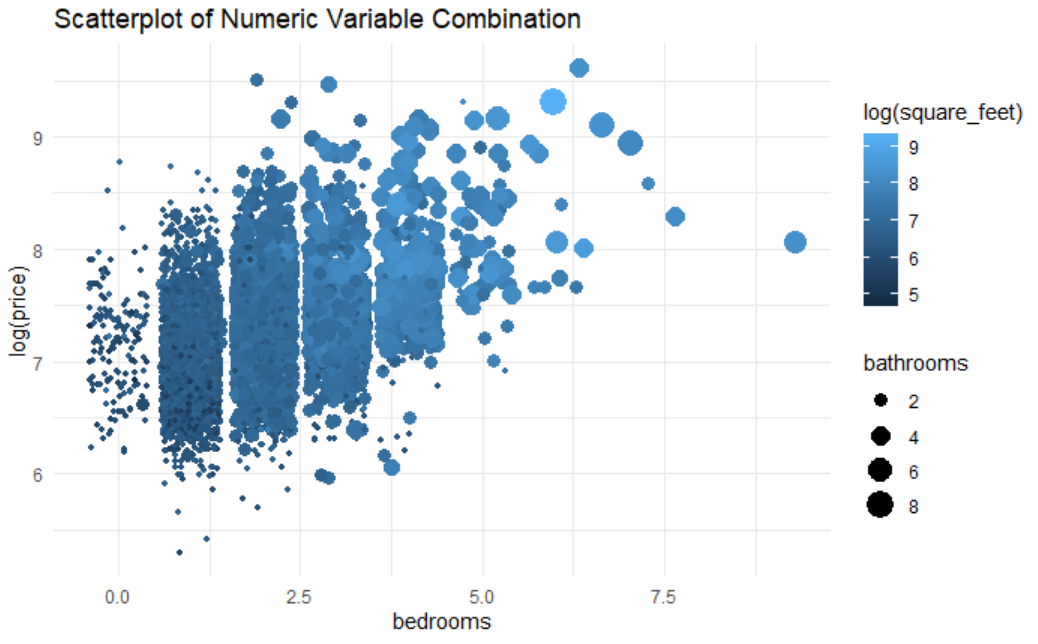


Figure 11. The table shows the result of the model evaluation.

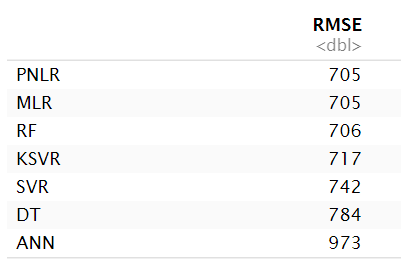


Figure 12. Plots show no issue with outlier, heteroscedasticity, and normality.

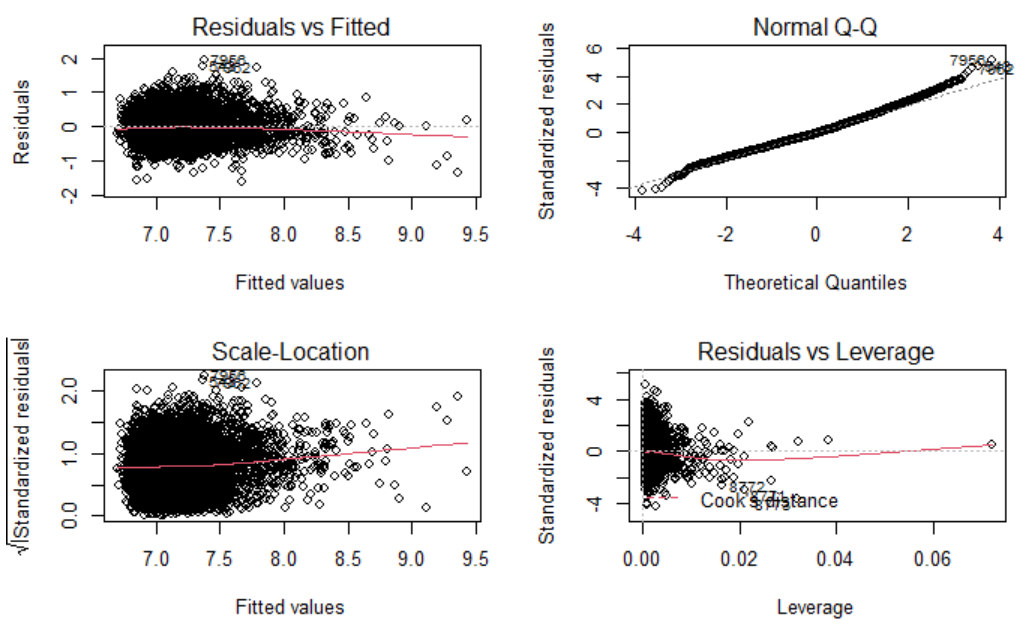


Figure 13. The table indicates no issue with collinearity.

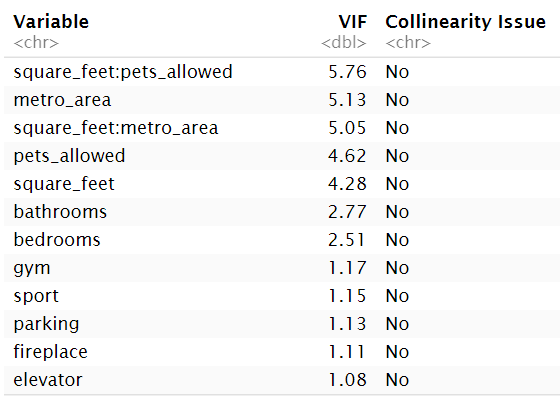
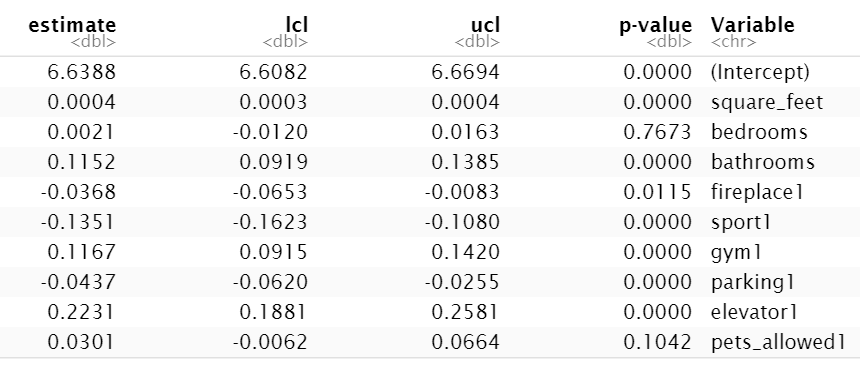


Figure 14. The table shows the log-linear regression estimates and coefficients.



**Reference**

1. UCI Machine Learning Repository – Apartment for Rent Classified Dataset – Fredrick Nilsson – URL (<https://tinyurl.com/y3zzn2au>)
2. Metro Area in the US – Jeff Desjardins – URL (<https://tinyurl.com/y2h5o2r3>)
3. Sublet.com – 1B1B apartment in Richardson TX – URL (<https://tinyurl.com/y59wtrzx>)

**Attachment**

The following pages are the RMD file with code and output in PDF format.