Detached House Sale Price Analysis

AR3031

October 26, 2020

First-time home buyers in Greater Toronto Area are facing extreme high rise in the prices for detached houses due to Covid impact. For analysis on the situation, we are establishing a simple linear model, for buyers to predict the expected sale price of detached, single family homes in Toronto and Mississauga. For this model analysis we are provided with real20.csv data file that was obtained by the Toronto Real Estate Board (TREB) on detached houses in the two neighborhoods.The data frame has five parameters which are:

- ID: property identification
- sold: the actual sale price of the property in millions of Canadian dollars
- list: the last list price of the property in millions of Canadian dollars
- taxes: previous year's property tax in Canadian dollars
- location: M- Mississauga Neighborhood, T- Toronto Neighborhood

I. Exploratory Data Analysis section

From the given csv file 200 randomized data were created in the name of data_fr. To visualize the relationship between the response variable "sales price" and independent variable "list", a single scatter plot has been used. As a house sales' price is set based on the prices that it was last listed therefore listed is the independent variable that would mostly control a house's sold price.

Making subsets:

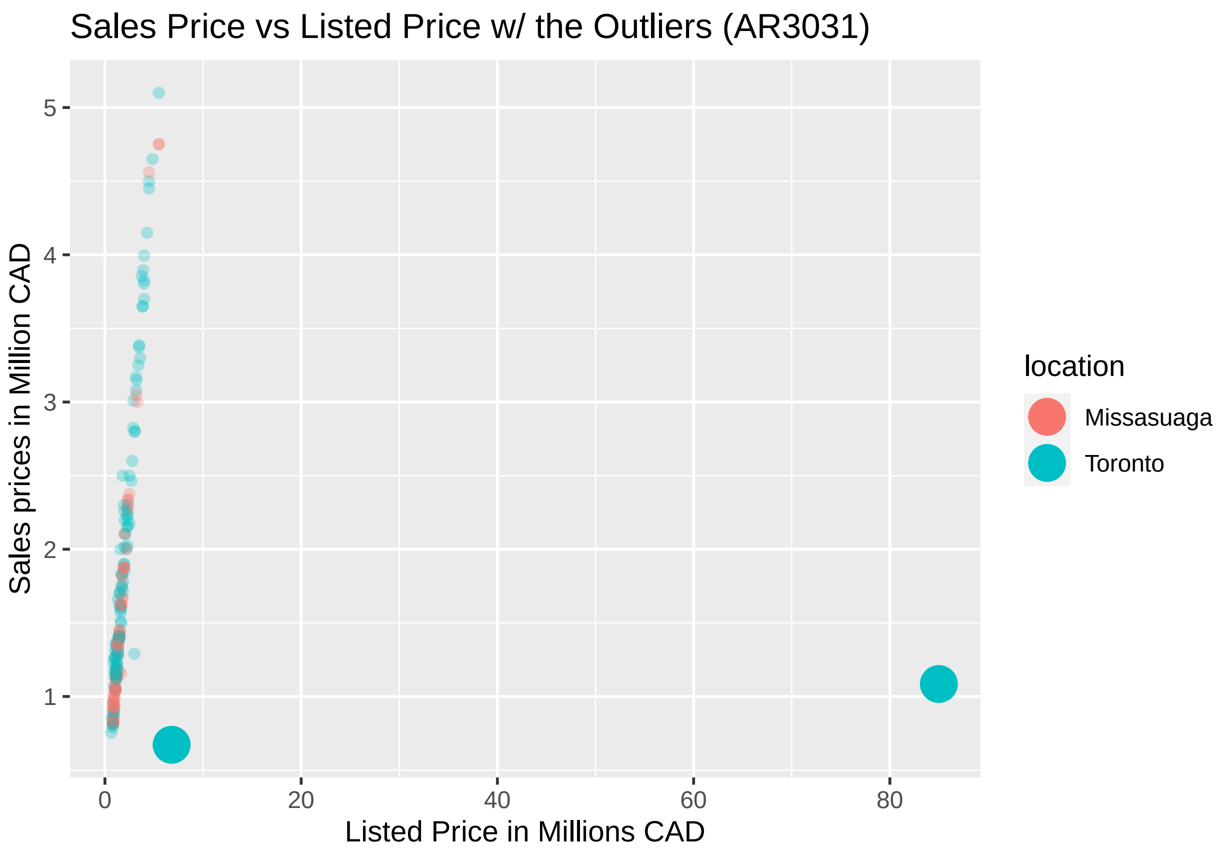


Figure 1. Scatter plot representing the relationship between the Sale Price and Listed price (with the outliers)

The scatterplot shows the positive correlation between the houses' listed prices and their sales price.The blue points represents the houses' sales prices in Toronto, and the orange represents for Mississauga neighborhood.

Evidently the value with the larger data points depicts 2 extreme x values which do not follow the linear trend of the rest of the data. Therefore these 2 extreme values can be predicted as the outliers .

For a better analysis, cook's distance has been used to identify the outliers and later cleaning the data. After eliminating the outliers found from the Cook's distance, a new subset was created. Using ggplot() function drew two scatterplots of the response variable: sale price in relation with (i) list price and then (ii) taxes.

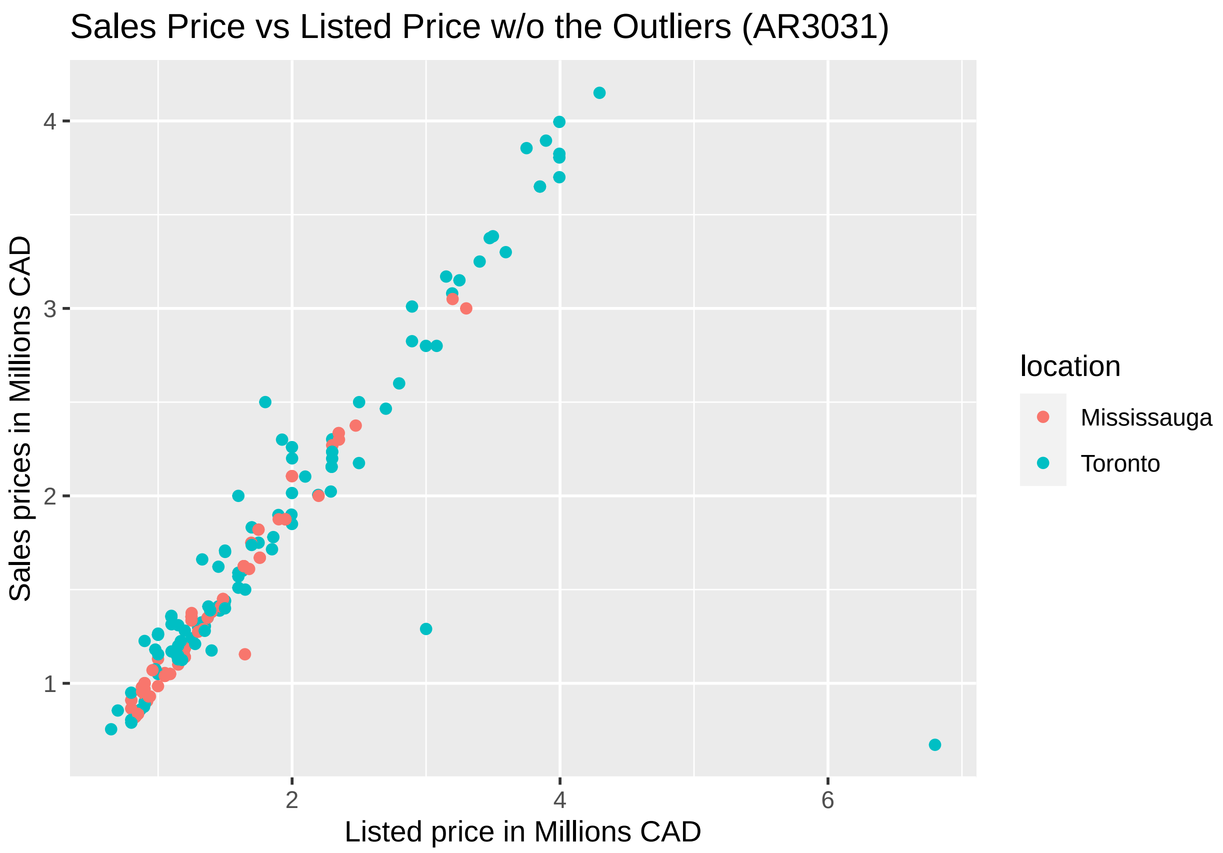


Figure 2. Scatter plot representing the relationship between the Sale Price and Listed price (without the outliers)

After, removing the outliers significantly changed the range of the independent variable (listed price). Thus the slope of the linearity is less steep. Therefore, it can be concluded that the two values are influential points.

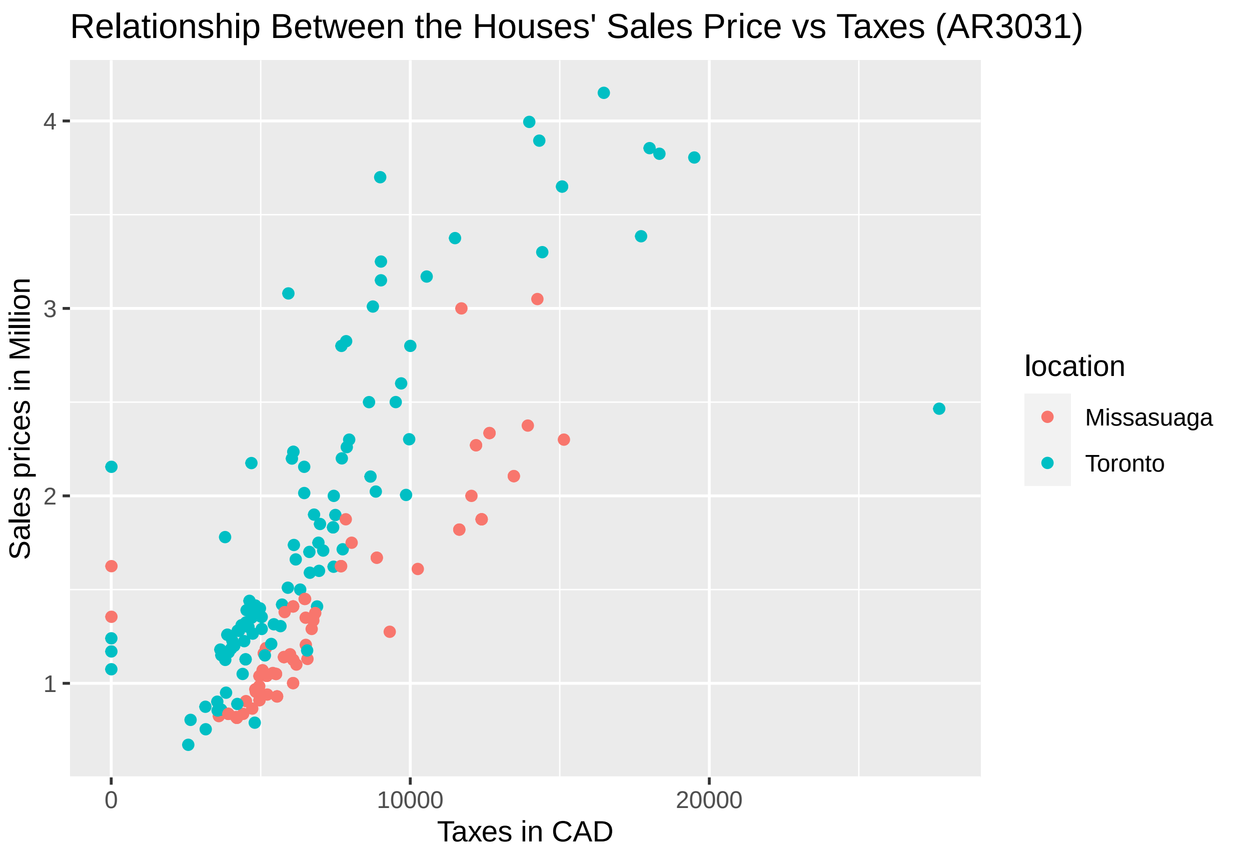


Figure 3. Scatter plot representing the relationship between the Sale Price and Taxes (without the outliers)

Interpretation of Three Models:

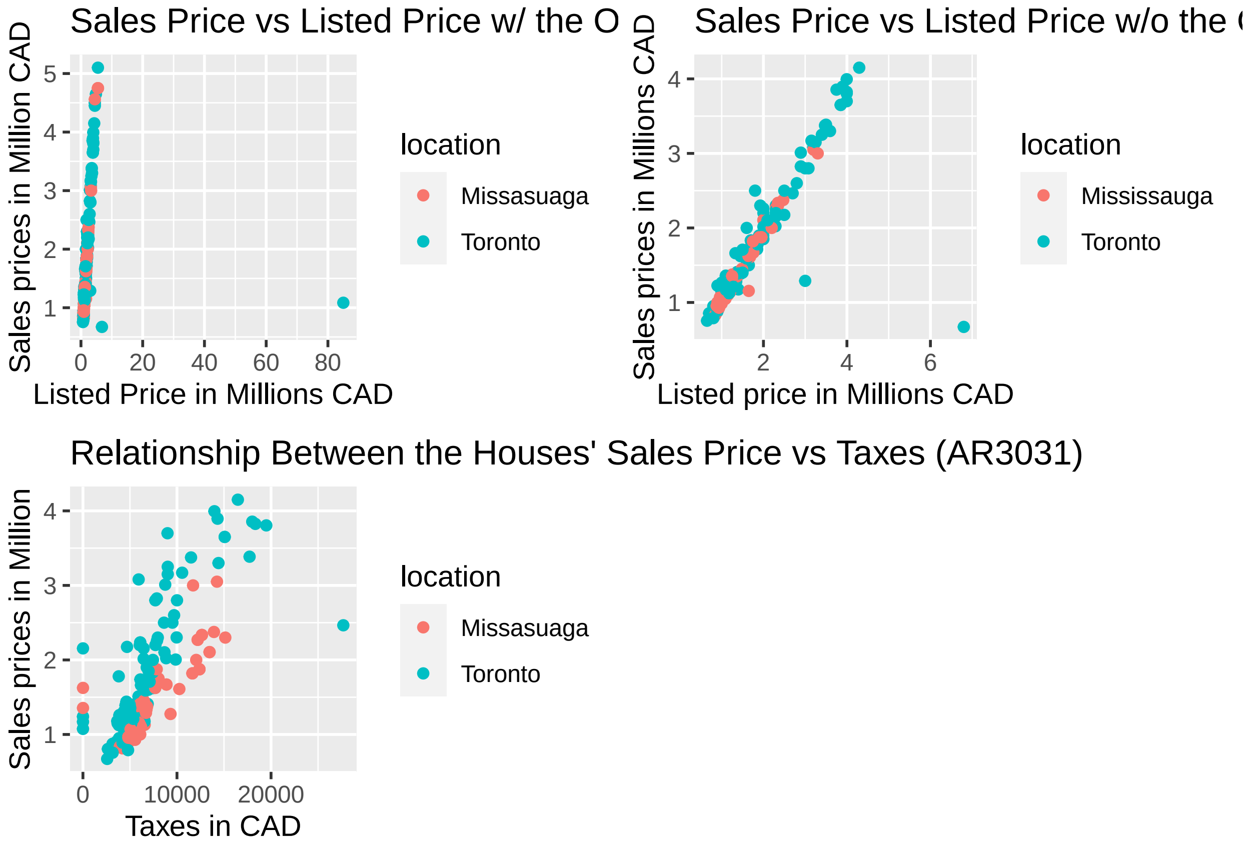


Figure 4: Side by Side representation of the Three plots

- In the first plot the highest range of the listed prices exceeded eighty million CAD and the maximum value in that range had an unrealistic sold price. It's highly unlikely that a house in Toronto that is listed as 80 Million CAD will be sold at around 1.5 Million CAD. Therefore, this is definitely an extreme point.

- In the second plot, the linearity is of the data trend is more visible than the first plot. Even though both the first and the second plots illustrates the same two variables (response variable: Sales price of the house, Explanatory variable: Listed Prices of the House), from the second it can seen than the house prices data in Toronto is more scattered than the Mississauga ones. Therefore, it can be predicted that the variance of the Mississauga sub data set would be less than the Toronto sub data set.

- The third plot shows the relationship between the sales price vs the taxes. Here, the data points are more scattered than the previous plots and there are quite a few outliers for both Toronto and Mississauga houses.Even though there is a hint of linearity for both of the sub data sets, still the scatterplot follows a somewhat clustered pattern. Therefore, can be predicted to have higher residual values.

II. Methods and Model

```
##                          R2 est intercept est slope est var of e       p-val
## SLR summary (All) 0.0079        0.7289    0.0327      22.3218 1.6762e-60
## SLR summary (Tor) 0.6360        0.6660    0.0486      13.7095 3.7952e-07
## SLR summary (Mis) 0.9755        0.9254    0.0161      57.4830 2.9260e-04
```

```
##               Upper Bound Lower Bound
## 95% CI (All)      0.6645      0.7623
## 95% CI (Tor)      0.7933      0.8933
## 95% CI (Mis)      0.5697      0.9574
```

Table: Summary of the Three Linear Models (for all data set, for Toronto Houses, for Mississauga House) and Their Confidence Intervals

In comparison with the 5% significant level, the very small P-values supports the validity of our analysis. Therefore, we have strong evidence to reject the null hypothesis of the slope of all models being 0.From the table above the three regression models are:

1. Simple Linear Regression (SLR) summary of All 200 data from the initial data frame: y = 0.0327x +0.7289
2. SLR summary of the house sales vs listed price data in the Toronto neighborhood: y = 0.0486x +0.6660
3. SLR summary of the house sales vs listed price data in the Mississauga neighborhood: y = 0.0161x +0.9254

The models shows that Toronto SLR model has the highest slope (0.0486). It is approximately 4% higher than the slope of the Mississauga model. Moreover the high slope value in the 1st model (depicting all data) is because the Toronto values may possibly have a higher influence in our data set. This can be due to more data collected for the Toronto neighborhood than the other which can be an indication a possible bias in our data set.

The higher slope in Toronto Model shows that the sales price of detached houses in Toronto increases higher, with a given listed price in million that Mississauga. Therefore, the area is factor in terms of the buyer's affordability to buy a detached house.

The R^2 values for the three models show that based on the location the correlation between the sales price and listed price of a house can be different. A high R^2 value (0.9755) of the Mississauga model shows that the correlation between the listed price and sales price is very strong in that neighborhood.This value is consistent with the inference made in part I from the 2nd plot (sales prices vs listed price without the outliers).

The smaller R^2 values (0.6360) for the Toronto model shows that the correlation between the two variables are not as strong but there are moderate positive correlation between the variables. This is found as we have seen in part I that data sets for Toronto are more scattered.

Finally the R^2 of the all data seems a bit unusually small, which can be due to the bias in our data as prevously mentioned. As we have seen that our total data frame is highly influenced by the Toronto sub data sets and due to that reason, no correlation between the two variables has appeared in this case.

A pooled two-sample t-test can be used to determine if there is a statistically significant difference between the slopes of the SLR models of the two neighborhoods. However, for such a test, the Toronto and Mississauga samples must meet some conditions.

Firstly, the two populations have to be independent. As no detailed information is provided regarding how Toronto Real Estate Board (TREB) has collected the data on detached houses in two separate neighborhoods or what factors have influenced the data collection, therefore we can not certainly say that the two populations are independent of each other.

The second condition is the variances of the two population variances must be equal which in this case evidently has not met. As the plots seen in the part I, the Toronto sample was more scattered than the Mississauga one. Therefore, it can be inferred that their variances are most likely different. Thus in my opinion, a pooled two-sample t-test is not appropriate in this case.

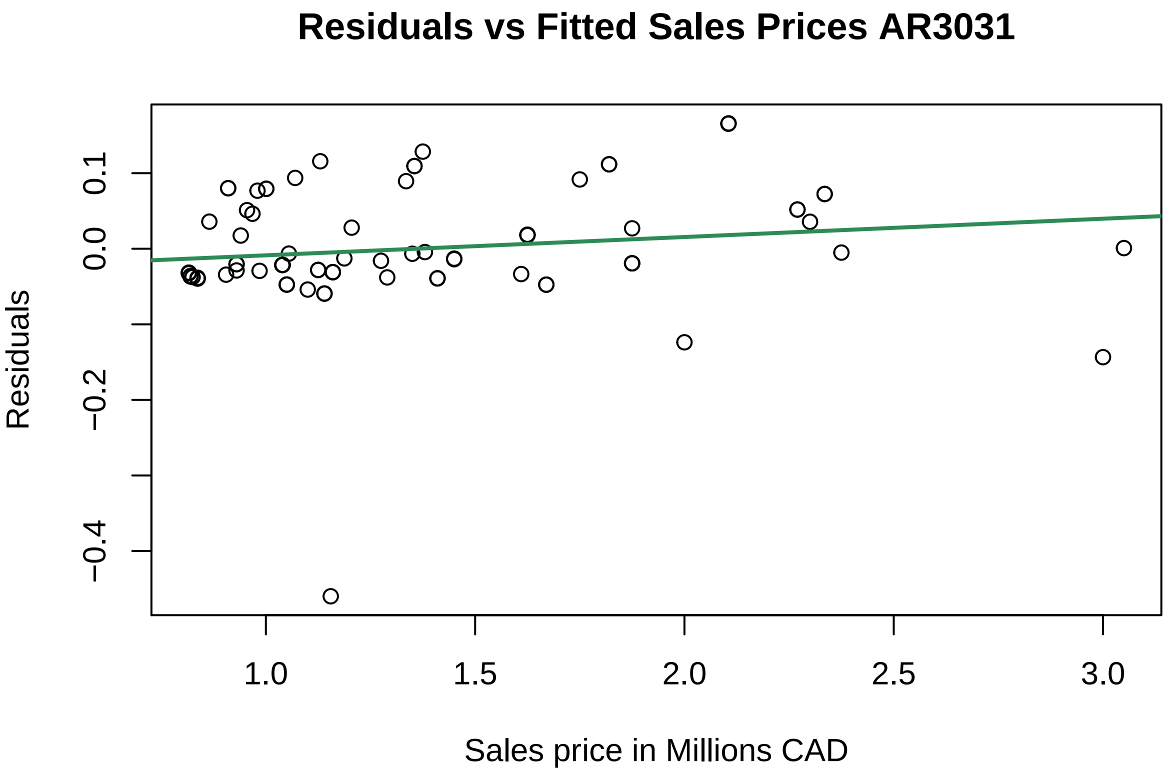
III. Discussion and Limitations section.

The goal of this analysis is to provide the future house buyers with a reliable correlation between a sales price of house with its last listed price through a “House sales price prediction model”. In part II, as the SLR for Mississauga has the highest Rˆ2 value which represents the largest positive correlation. Therefore I believe, this model gives the best prediction of the house sales price for a given listed price, that can concluded for a larger population. Moreover, the low p value (2.9260e-04) also proves the validity of the selected model.

Violations of the normal error SLR assumptions for model II

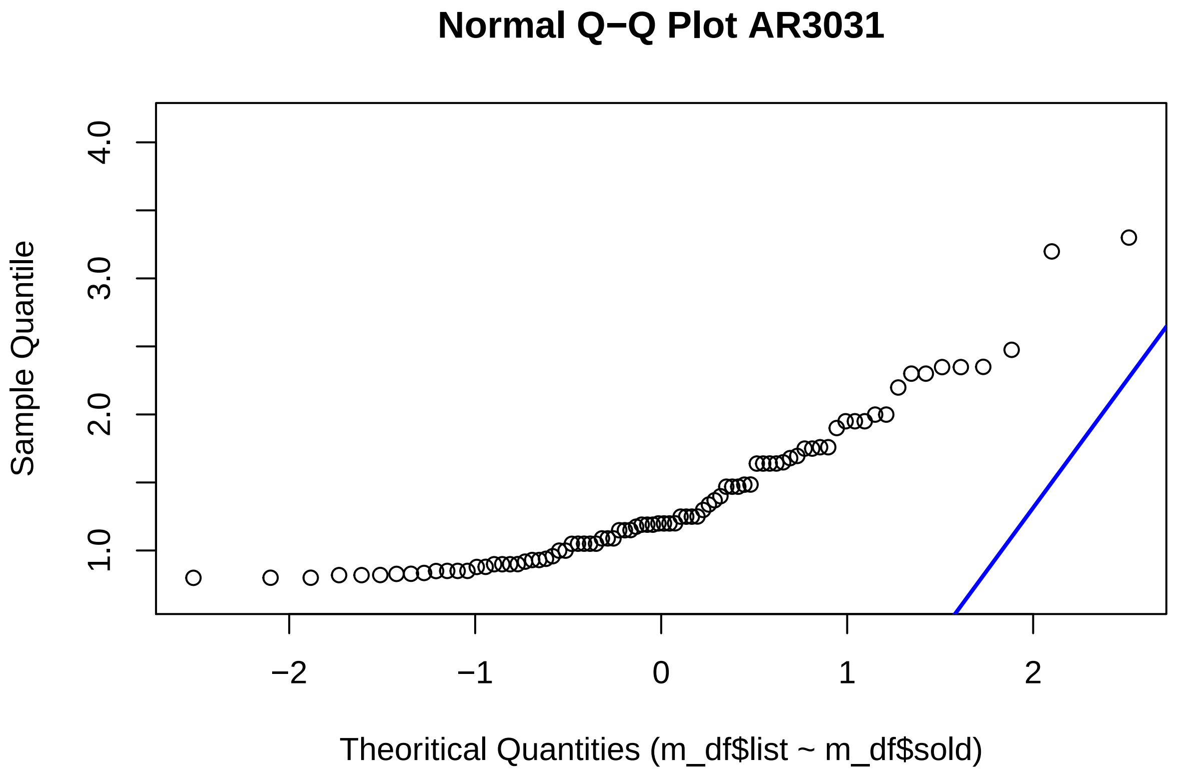
The four normal error SLR assumptions for the selected model would be: 1. A simple linear model is appropriate, which the plots and the table values have shown that the selected model meets. 2. There is not enough information provided to determine if the errors are correlated.

3. The errors have constant variance. For this a residuals vs fitted values will be plotted.



The plot depicts that the errors have constant variance, as the residuals are scattered randomly around zero without following any pattern.

4. The errors are Normally distributed.



From the plot, it is evident that the model violates the assumption of the errors being normally distributed. The points are spread in an non-linear, exponential pattern. The offset between the line and the points suggests that the mean of the model data is not zero.

The plot or lot size of the house and the number of rooms in the house can be the two potential numeric predictors that could be used to fit a multiple linear regression for sale price. For both of these predictors, the larger the number is the higher the price would be.