House Prices Prediction using Multiple Linear Regression

Raunak Milind Sathe
School of computing
National College of Ireland
Dublin, Ireland

***Abstract*—The house prices depend on a number of factors. It is important to understand the relationships between the characteristics of the house and the price. In this study Multiple Linear Regression technique has been used to understand this relationship. Multiple transformations and assumption testing were used to improve this understanding. This study addresses an important research question using the House Prices dataset. The purpose of this study is to understand the relationship between house characteristics and house prices in order to predict the house price.**

Keywords—Regression, Transformations, assumption.

I. INTRODUCTION

Buying a house is a sizeable investment and people want maximum value for their money. Calculating the house price can be difficult as a lot of factors need to be considered. The dataset House Prices was used which contained various important factors like *landValue, lotSize,* age etc. By conducting a number of tests, some factors which appeared important turned out to be insignificant. This can be confusing for the customers. This analysis is focused on trying to improve the understanding of the relationship between these characteristics with the prices of the houses using Multiple Linear Regression

A. MULTIPLE LINEAR REGRESSION-

Multiple Linear Regression is a linear approach to estimate the relationship between the Dependent variable and more than one variables, often referred to as the independent variables. The mathematical model of Multiple Linear Regression can be given as –

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

where,

y = The dependent variable projected value

β = *Regression Coefficient*

k = *Total number of independent variables.*

II. DATASET DESCRIPTION

The dataset House Prices contains a total of 15 independent variables and one dependent variable. Further analysis of the dataset was necessary to understand the characteristics of the variables.

A. Variables Description

Dependent variable – In the dataset the variable *Price* is the dependent variable. The dependent variable is a continuous numerical variable.

Independent Variables – The independent variables can be categorized into 3 different categories based on their values.

1. The variables *lotSize, age, landValue, livingArea, pctCollege, fireplaces, bedrooms, bathrooms, rooms* are all continuous numeric variables.
2. The variables *heating, fuel and sewer* are nominal categorical variables with string values.
3. The variables *waterfront, newConstruction* and *centralAir* are dichotomous categorical variables with *yes* and *no* values.

The dataset was analysed and some pre-processing was done as mentioned below.

B. Outlier treatment

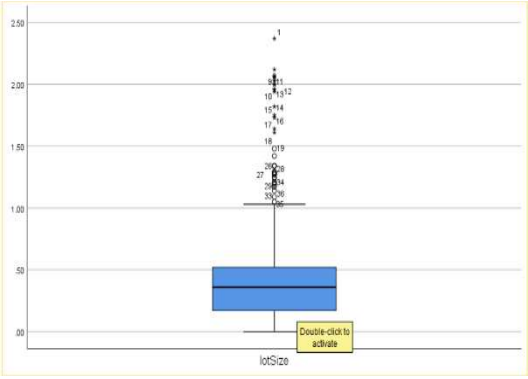
Outliers are values that are found abnormally distanced from the other observations. Outliers can have a significant impact on the mean values, standard deviation etc. Removing outliers can have a significant difference on the normal distribution of variables, decreasing the variance in the data which greatly increases statistical power. There are a few ways to identify and remove outliers from a dataset. In this study two methods have been used.

1. Z-score

Values with Z-score more than +/-3 are generally considered outliers. Hence those values were removed.

2. Boxplots

Box plots are used to graphically distribute the data into groups. They contain an Inter-Quartile range which consists of 50% of the data. Any values that lie 1.5*(Q3-Q1) outside Q3 or Q1 where Q3-Q1 is the Inter-Quartile range are considered outliers. Those values were removed.



As seen in the image below for the variable lotSize, 1.5*(Q3-Q1) is equal to 1.1. Hence the values beyond this limit were removed. The same process was applied to all the variables to remove the outliers.

C. Variable Transformation -

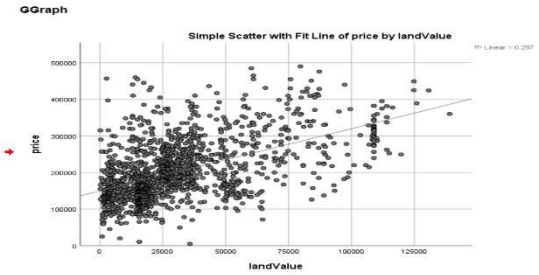
The categorical variables in the dataset were transformed into numeric variables in order to use them in the regression model. The variables *Heating*, *Fuel*, *Sewer* each had 3 categories. They were converted and given values 0,1 and 2. The variables *Waterfront*, *newConstruction*, *centralAir* were dichotomous having variable values as Yes and No. They were converted to 0 and 1.

III. Multiple Linear Regression Assumptions

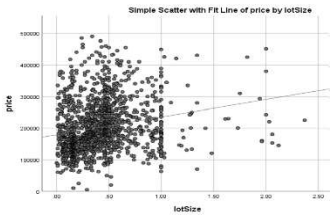
Multiple Linear Regression makes some key assumptions that must not be violated. A violation of these assumptions means that the output cannot be considered valid. Some assumptions can be checked before building the model from the dataset. Some assumptions need to checked after building the model.

Assumption 1 – Linearity Assumption

The first assumption is that the relationship between the dependent variable and the independent variable is linear. This is an important assumption as we want only those variables are linearly related to the dependent variable. These variables will be of the maximum use when predicting the dependent variable.



The image shows the linear relationship between the dependent variable price and a significant variable *landValue*. Similarly, a graph for all the other independent variables was plotted to identify the relationship between them. However, some plots did not show a clear linear relationship.



Although there is some relationship between the variables, it can be concluded that some changes are going to be needed to build a more linear relationship.

Assumption 2 – Multicollinearity

Multicollinearity is said to exist when there is an interaction between the independent variables. This problem causes the coefficients become more susceptible to change for small changes in the model. The solution to this problem as discussed earlier is to usually remove the variable. In order to detect multicollinearity, there are some tests that can be run in the SPSS software.

Variance Inflation Factor/Tolerance

Variance Inflation Factor (VIF) can be calculated by running the collinearity diagnostics test in SPSS. A VIF value above 10 indicates the presence of multicollinearity for sure. Hence, values below 10 are generally accepted. Tolerance is the reciprocal value of VIF. Hence the tolerance values should not be less than 0.1.

Assumption 3 – Normal Distribution of Residuals

In multiple Linear Regression it is assumed that the residuals are normally distributed. To test this assumption a number of tests can be conducted. If they are not normally distributed, this means that the independent variables technically mean different things for different values of the dependent variable.

A. Q-Q plots

A Q-Q plot can be plotted in R that can show us the linearity of the residuals. A straight line is generally expected.

B. Histogram

Histograms can be used for this test. A normal distribution is expected to be observed in the histogram.

Assumption 4 – Homoscedasticity

This assumption states that there is equal variance among the residuals. The presence of any pattern especially a funnel shaped pattern indicates a problem. A violation of this assumption means that the residual values are less precise. Also, the observed P-values are smaller than they should be. The Scale-Location plot can be used to check for a problem in homoscedasticity.

Assumption 5 – Independence of the Errors

This assumption states that the values of the errors are unaffected by errors previously detected. Basically, the observations are independent of each other. The Durbin-Watson statistic is used for this assumption testing. The Durbin-Watson statistic should be between 1 and 3.

Assumption 6 – No influential points

This assumption states that there should be no highly influential points in the dataset. Influential points can greatly affect the slope of the line even changing the direction of the

slope completely. The Cook's distance is used to test the assumption. A point with a Cook's distance value of more than 1 is generally considered to be influential.

IV. Initial Model Building

Multiple Linear Regression model is building a model that has no insignificant variables, a decent R-score and that meets all the assumptions. An initial model was built using just the dependent variable and all the independent variables.

A. Variable Insignificance testing

Variables can become insignificant when they are related to another variable. Hence, the information is already present in the model. *The null hypothesis states that the variable in the dataset has no effect on the dependent variable. A p-value less than 0.05 is statistically insignificant. When the p-value is lower than 0.05 we can reject the null hypothesis and keep the variable in the dataset. When the p-value is greater than 0.05, we fail to reject the null hypothesis. This means that we cannot confirm that a significant difference exists and we do not consider the variable for the regression model.* After regressing every individual variable against the dependent variable, it is observed that all the variables are statistically significant. However, a regression analysis of all the variables against the dependent variable showed that some variables which were significant earlier were not significant anymore. Those variables were removed.

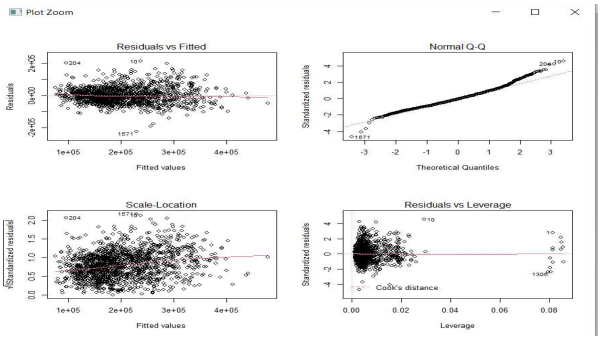
Removing the insignificant terms

The insignificant variables were removed and the model was built that consisted of the following variables.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.281e+04  5.493e+03   2.332 0.019825 *
lotSize          1.599e+04  4.123e+03   3.879 0.000109 ***
age             -2.181e+02  6.784e+01  -3.214 0.001335 **
landvalue        9.304e-01  5.397e-02  17.239  < 2e-16 ***
livingArea       6.079e+01  3.980e+00  15.276  < 2e-16 ***
bathrooms        2.136e+04  2.946e+03   7.251 6.49e-13 ***
rooms            1.583e+03  7.710e+02   2.054 0.040189 *
waterfront       1.398e+05  1.342e+04  10.416  < 2e-16 ***
newConstruction -2.700e+04  6.304e+03  -4.283 1.96e-05 ***
centralAir       1.282e+04  2.754e+03   4.653 3.55e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47850 on 1561 degrees of freedom
Multiple R-squared:  0.6476,    Adjusted R-squared:  0.6455
F-statistic: 318.7 on 9 and 1561 DF,  p-value: < 2.2e-16
```

These regression analysis of these variables against the dependent variable generated the following plot.



V. Initial Model Diagnostics

The model plot as shown above violates a number of key assumptions -

1 – In the Residuals vs fitted plot, which can be considered as a good measure of the validity of the model, the relationship is not exactly linear.

2 – In the Normal Q-Q plot a significant tail can be observed at both ends which maybe shows that the residuals are not exactly normally distributed.

3 – In the Scale Location plot, a funnel shape can be observed which shows that there could be a problem with homoscedasticity.

VI. Transformations and Interactions

There are a number of ways to improve the model. Some methods fit the model better than others.

Transforming the Variables

Transforming the variables can have a big difference on the variables.

A. Taking Square root of dependent variable-

Taking Square root of a variable can improve the linear relationship between two variables. Square root can help overcome problem with homoscedasticity.

B. Taking sqaure of variables-

Taking the squares of the variables can be used to include the effect of the differing variables like age.

C. Taking Interactions of the variables-

Taking interaction values can greatly improve the understanding of the relationship between the variable.

VII. Final Model Building

In order to improve the model, 3 notable differences were added. Square root transformation of the dependent variable Price, square value of all the continuous variables and the interactions between certain independent variables were included. Although, the interaction terms were found significant, they weren't contributing a lot to the final model. They were not included in the final set of variables. Hence, the final set of variables were the Square root of the dependent variable, all the independent variables and the squares of all the independent continuous variables.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.297e+02  1.750e+01  13.126  < 2e-16 ***
lotsize           2.267e+01  1.139e+01   1.990  0.04673 *
I(lotSize^2)     -3.452e+00  7.078e+00  -0.488  0.62585
age              -2.450e-01  8.200e-02  -2.988  0.00285 **
I(landvalue^2)    8.415e-09  5.802e-10  14.503  < 2e-16 ***
livingArea        8.740e-02  1.683e-02   5.192 2.36e-07 ***
I(livingArea^2) -4.835e-06  4.535e-06  -1.066  0.28645
rooms             3.008e+00  3.755e+00   0.801  0.42332
I(rooms^2)       -7.999e-02  2.689e-01  -0.297  0.76614
bathrooms         1.537e+01  1.452e+01   1.058  0.29018
I(bathrooms^2)    2.098e+00  3.636e+00   0.577  0.56407
waterfront        1.406e+02  1.527e+01   9.209  < 2e-16 ***
newConstruction -3.592e+01  7.355e+00  -4.884 1.15e-06 ***
centralAir        1.507e+01  3.168e+00   4.756 2.16e-06 ***
heating           3.015e-01  3.325e+00   0.091  0.92776
fuel              3.314e+00  3.563e+00   0.930  0.35248
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.21 on 1555 degrees of freedom
Multiple R-squared:  0.6241,    Adjusted R-squared:  0.6204
F-statistic: 172.1 on 15 and 1555 DF,  p-value: < 2.2e-16
```

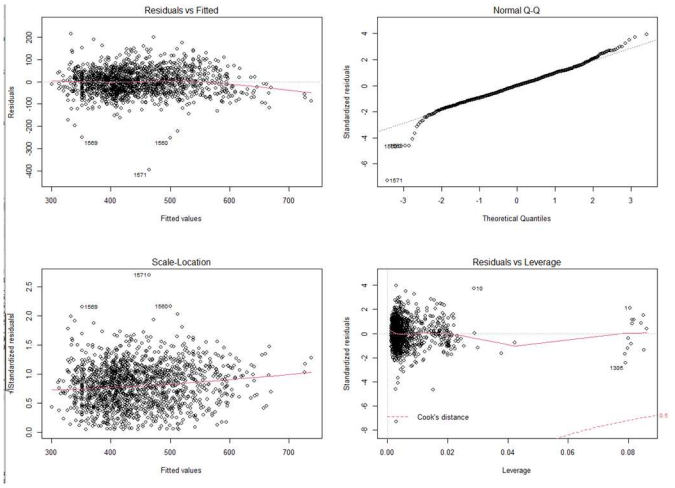
The insignificant variables (denoted by *) were removed one by one. A total of 7 iterations were run where in each step an insignificant variable was removed. Finally, a model containing

all the significant variables was built. The following plots were generated –

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.456e+02  6.084e+00  40.372  < 2e-16 ***
lotSize          1.930e+01  4.673e+00   4.131 3.81e-05 ***
age             -1.935e-01  7.656e-02  -2.527   0.0116 *
I(landvalue^2)   8.272e-09  5.695e-10  14.525  < 2e-16 ***
livingArea       7.532e-02  3.720e-03  20.246  < 2e-16 ***
bathrooms        2.415e+01  3.339e+00   7.233 7.35e-13 ***
waterfront       1.394e+02  1.517e+01   9.186  < 2e-16 ***
newConstruction -3.586e+01  7.254e+00  -4.944 8.48e-07 ***
centralAir       1.542e+01  3.101e+00   4.972 7.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.24 on 1562 degrees of freedom
Multiple R-squared:  0.6219,    Adjusted R-squared:   0.62
F-statistic: 321.2 on 8 and 1562 DF,  p-value: < 2.2e-16
```

The model plots were as follows-



VIII. Final Model Diagnostics

For a better understanding, the model with the same significant variables was replicated in SPSS.

A. Assumption – Checking for multicollinearity

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	245.612	6.084		40.372	.000		
	lotSize	19.302	4.673	.067	4.131	.000	.907	1.102
	age	-.193	.077	-.045	-2.527	.012	.758	1.320
	livingArea	.075	.004	.479	20.246	.000	.432	2.316
	bathrooms	24.149	3.339	.173	7.233	.000	.425	2.353
	waterfront	139.370	15.172	.144	9.186	.000	.991	1.009
	newConstruction	-35.864	7.254	-.085	-4.944	.000	.825	1.213
	centralAir	15.417	3.101	.084	4.972	.000	.838	1.193
	landvaluesquare	8.272E-9	.000	.251	14.525	.000	.808	1.238

a. Dependent Variable: priceSquareRoot

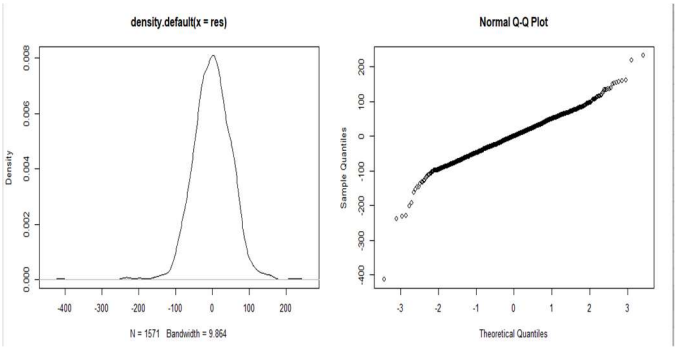
None of the VIF values in the above figure are above 10. Also, there are no tolerance values below 0.1. Hence, the multicollinearity assumption has not been violated.

Correlations

		priceSquareRoot	lotSize	age	livingArea	bathrooms	waterfront	newConstruction	centralAir	landvaluesquare
Pearson Correlation	priceSquareRoot	1.000	.219	-.257	.713	.613	.133	.178	.312	.467
	lotSize	.219	1.000	-.015	.247	.113	.015	-.003	-.075	.067
	age	-.257	-.015	1.000	-.278	-.427	.071	-.250	-.254	-.054
	livingArea	.713	.247	-.278	1.000	.712	-.025	.279	.264	.345
	bathrooms	.613	.113	-.427	.712	1.000	-.022	.241	.321	.271
	waterfront	.133	.015	.071	-.025	-.022	1.000	-.020	-.040	.037
	newConstruction	.178	-.003	-.250	.279	.241	-.020	1.000	.044	.299
	centralAir	.312	-.075	-.254	.264	.321	-.040	.044	1.000	.194
	landvaluesquare	.467	.067	-.054	.345	.271	.037	.299	.194	1.000

All except the coefficient correlation value *for bathrooms: livingArea* are below 0.7 which is generally a good indicator for the presence of multicollinearity. The value is just above 0.7 at 0.712. Hence, the assumption of multicollinearity has not been violated.

B. Assumption – Residual Normality and Linearity

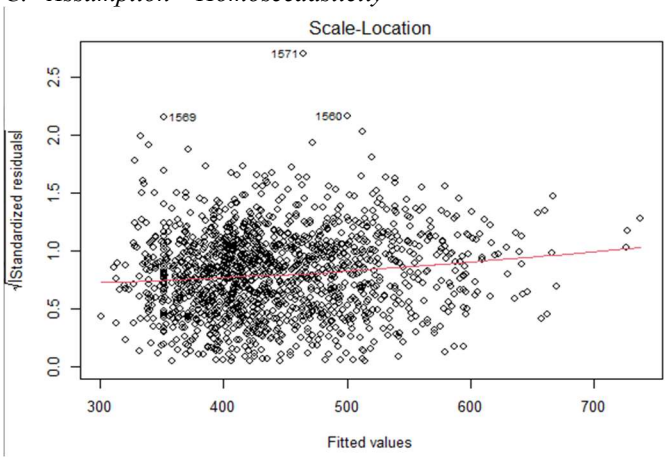


The Density plot shows a relatively good normal distribution. Although the plot is negatively skewed, it can be reasonable to assume that the residuals are normally distributed.

The Q-Q plot shows there is a small tail on both ends. But comparing with the initial plot, the points are relatively close to a straight line. The cases “1560” and “1571” appear to be a problem in this plot.

Hence, this assumption has been satisfied.

C. Assumption – Homoscedasticity



Comparing this plot with the initial model plot, a clear improvement can be seen in this plot. There doesn’t appear to be a funnel shaped pattern here. The residuals look fairly distributed. Although, there seems to some problems with cases

"*1560*", "1509" and "1571", the overall plot can be considered satisfactory. Hence, it can be assumed that this assumption has not been violated.

D. Assumption – Independence of Errors.

Model Summary[b]

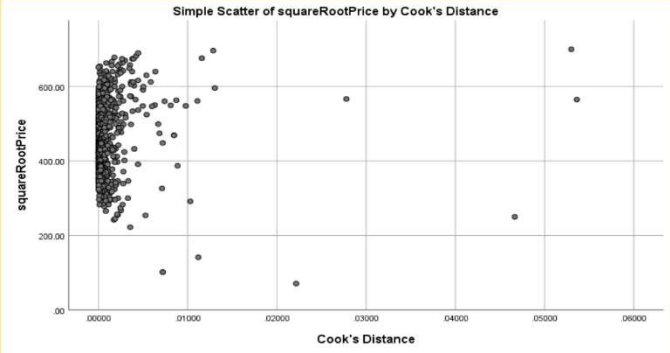
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.789[a]	.622	.620	54.23885	1.071

a. Predictors: (Constant), landvaluesquare, waterfront, lotSize, age, centralAir, newConstruction, livingArea, bathrooms

b. Dependent Variable: priceSquareRoot

The Durbin-Watson value which is used to test this assumption should be between 1 and 3. Although the value should be closer to 2, as the value is above 1 it can be assumed that this assumption has not been violated.

E. Assumption – No Influential Outliers



From the above figure, none of values have a Cook's distance value of more than 1. Hence, this assumption has not been violated.

IX. Model Outcome Understanding

Once all the assumptions have been successfully met, the output of the model can be understood.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	245.612	6.084		40.372	.000		
	lotSize	19.302	4.673	.067	4.131	.000	.907	1.102
	age	-.193	.077	-.045	-2.527	.012	.758	1.320
	livingArea	.075	.004	.479	20.246	.000	.432	2.316
	bathrooms	24.149	3.339	.173	7.233	.000	.425	2.353
	waterfront	139.370	15.172	.144	9.186	.000	.991	1.009
	newConstruction	-35.864	7.254	-.085	-4.944	.000	.825	1.213
	centralAir	15.417	3.101	.084	4.972	.000	.838	1.193
	landvaluesquare	8.272E-9	.000	.251	14.525	.000	.808	1.238

a. Dependent Variable: priceSquareRoot

A. Multiple Regression Equation

The regression equation can be calculated as –

Price = 245.612 + 19.302*(x1) - .193*(x2) + 0.75*(x3) + 24.149*(x4) + 139.370*(x5) – 35.864*(x6) + 15.417*(x7) + 8.272E-9*(x8)

Where -

β_0 (245.612) – Constant

β_1 (19.302) – *lotSize* Coefficient

β_2 (-0.193) – *age* Coefficient

β_3 (0.75) – *livingArea* Coefficient

β_4 (24.149) – *bathrooms* Coefficient

β_5 (139.370) – *Waterfront* Coefficient

β_6 (-35.864) – *newConstruction* Coefficient

β_7 (15.417) – *centralAir* Coefficient

β_8 (8.272E-9) – *landValueSquare* Coefficient

B. Output

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.789[a]	.622	.620	54.23885	1.071

a. Predictors: (Constant), landvaluesquare, waterfront, lotSize, age, centralAir, newConstruction, livingArea, bathrooms

b. Dependent Variable: priceSquareRoot

The Adjusted R Square value is used to determine how good the model is. It is a measure of the variance in the Dependent variable Y that can be explained by the independent variables. An Adjusted R Square value of 0.620 has been obtained in the model summary. This essentially means that the independent variables justify 62% variance in the dependent variable Price.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7558842.012	8	944855.252	321.177	.000[b]
	Residual	4595174.306	1562	2941.853		
	Total	12154016.32	1570			

a. Dependent Variable: priceSquareRoot

b. Predictors: (Constant), landvaluesquare, waterfront, lotSize, age, centralAir, newConstruction, livingArea, bathrooms

From the ANOVA table the F-statistic is 321.177 and the significance level is 0.000, which means that the results are statistically significant at $\alpha = 0.05$.

X. Conclusion

A Multiple Regression analysis was conducted to understand the relationship between different house characteristics and the house prices. The data was transformed, cleaned and outliers were removed. Multiple models were built to finalize a model that best fitted the data and satisfied all the assumptions. There was no collinearity observed among the variables. The residuals were found to be normally distributed and homoscedasticity assumption was satisfied. The adjusted R^2 value of 0.620 showed that 62.0% of the dependent variable variance was explained by the independent variables. Since this value is significant, it can be concluded that the model is a good fit for the data.

XI. References

[1] R for Data Science, Book by Grolemund and Hadley wickham