Predicting House Prices using Multiple Linear Regression

Mohit Kaushal Jain
Dept. of School of Computing
MSc. in Data Analytics
National College of Ireland
Dublin, Ireland
x21191514@student.ncirl.ie

***Abstract*—This research paper discusses about house price prediction using multiple linear regression. The dataset that was analysed on houses at King County, Seattle for year 2014-2015. The steps that were involved in analysis are dataset description, descriptive statistics, preprocessing, Model Review Selection and prediction.**

***Index Terms*—Multiple linear regression, preprocessing, descriptive statistics, preprocessing, model and prediction.**

I. Aim and Objective

The aim of this project is to use descriptive statistics and utilize multiple linear regression model to find how price of the house holds on multiple dependent variables. The reason for evaluating descriptive statistics are to find basic information of the variables in the dataset and to find connection between the variables.

A straight line is used to assess the relationship between a quantitative dependent variable and two or more independent variables in a regression model called multiple linear regression.[1]
The expression for multiple linear regression is:

$$Y = \beta_0 + \beta_1 * b_1 + \beta_2 * b_2 + \beta_3 * b_3 + ... + \beta_p * b_p \quad (1)$$

The above expression is implemented using R programming.

II. Dataset Description and its Variables

The dataset provided was "kc-house-data.csv". This file contains 21,613 rows and 21 columns(also known as feature variables). In this dataset, it contains one single dependent variable known as price and 20 other variables as the independent variables. [2]

The description of of the given dataset is as follows:

- id - House's identification number
- date - the day when the house was sold
- price - The price at which the house was sold
- bedrooms - How many bedrooms are there in the house.
- bathrooms - How many bathrooms are there in the house.
- sqft-living - area of living space in square feet.
- sqft-lot - Area of the lot in square feet.
- floors - How many floors are there in the house.
- waterfront - It signifies whether waterfront is present or not. If the value is "1" then waterfront is present else it is absent.
- view - It signifies how nice the property's view was. It ranges from 0 to 4.
- condition - It signifies the quality of the house. It ranges from 1 to 5.
- grade - Classification based on building quality, which takes into account the kind of materials and level of workmanship utilized. Better-quality (higher grade) structures cost more to construct per unit of measure and are worth more.
- sqft-above - Square feet above ground
- sqft-basement - Square feet below ground
- yr-built - Year when the house was built
- yr-renovated - Year when the house was renovated. '0' if not renovated
- zipcode - Zip code of the house
- lat - latitude of the house
- long - longitude of the house
- sqft-living15 - The mean square footage of the interior living areas in the 15 nearest houses.
- sqft-lot15 - The mean square footage of the lots for the 15 nearest houses

III. Descriptive Statistics

With R programming, there are variety of functions that can evaluate median, mean, mode, standard deviation, variance etc. Descriptive statistics is a process in which one can study and investigate the data. To identify the number of samples 'N' nrow command was used. The min function gives the minimum value of the value of the sample, the max function gives the maximum value of the sample. Apart from the analysis of descriptive statistics, visualizations of appropriate variables were plotted as show in the figures below. [3]

IV. Preprocessing

In this step, the dataset analysis was done by identifying any kind duplication of records or if any kind of data which is missing before the multiple regression model was built. Also certain data types were converted into another data type of the

A data.frame: 14 × 17

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
nbr.val	2.161300e+04	2.161300e+04	2.161300e+04	2.161300e+04	2.161300e+04	2.161300e+04	2.161300e+04	2.161300e+04	2.161300e+04	2.161300e+04
nbr.null	0.000000e+00	1.300000e+01	1.000000e+01	0.000000e+00	0.000000e+00	0.000000e+00	2.145000e+04	1.948900e+04	0.000000e+00	0.000000e+00
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	7.500000e+04	0.000000e+00	0.000000e+00	2.900000e+02	5.200000e+02	1.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
max	7.700000e+06	3.300000e+01	8.000000e+00	1.354000e+04	1.651359e+06	3.500000e+00	1.000000e+00	4.000000e+00	5.000000e+00	1.300000e+01
range	7.625000e+06	3.300000e+01	8.000000e+00	1.325000e+04	1.650839e+06	2.500000e+00	1.000000e+00	4.000000e+00	4.000000e+00	1.200000e+01
sum	1.167293e+10	7.285400e+04	4.570625e+04	4.495287e+07	3.265069e+08	3.229650e+04	1.630000e+02	5.064000e+03	7.368800e+04	1.654880e+05
median	4.500000e+05	3.000000e+00	2.250000e+00	1.910000e+03	7.618000e+03	1.500000e+00	0.000000e+00	0.000000e+00	3.000000e+00	7.000000e+00
mean	5.400881e+05	3.370842e+00	2.114757e+00	2.079900e+03	1.510697e+04	1.494309e+00	7.541757e-03	2.343034e-01	3.409430e+00	7.656873e+00
SE.mean	2.497233e+03	6.326366e-03	5.238720e-03	6.247319e+00	2.817461e+02	3.673054e-03	5.884979e-04	5.212562e-03	4.426414e-03	7.995578e-03
CI.mean.0.95	4.894760e+03	1.240014e-02	1.026828e-02	1.224521e+01	5.522432e+02	7.199457e-03	1.153499e-03	1.021701e-02	8.676097e-03	1.567192e-02
var	1.347824e+11	8.650150e-01	5.931513e-01	8.435337e+05	1.715659e+09	2.915880e-01	7.485226e-03	5.872426e-01	4.234665e-01	1.381703e+00
std.dev	3.671272e+05	9.300618e-01	7.701632e-01	9.184409e+02	4.142051e+04	5.399889e-01	8.651720e-02	7.663176e-01	6.507430e-01	1.175459e+00
coef.var	6.797542e-01	2.759138e-01	3.641851e-01	4.415794e-01	2.741815e+00	3.613636e-01	1.147176e+01	3.270620e+00	1.908657e-01	1.535168e-01

Fig. 1. Descriptive Statistics

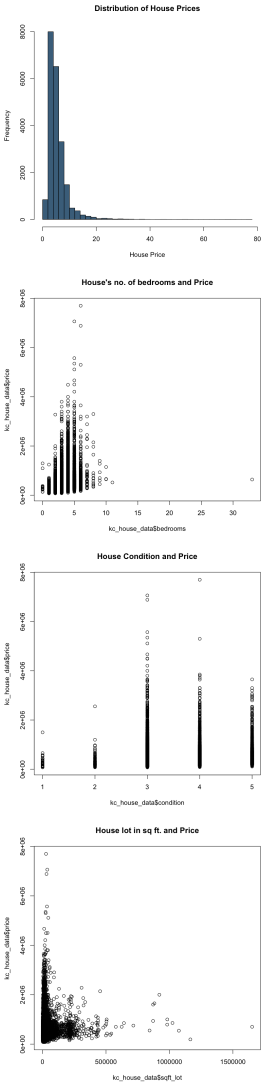

Fig. 2. Visualizations

variables, for example, the waterfront variable of the housing dataset was converted to factors using as.factor() command as it was a categorical variable. A correlation plot was generated to check which variables are to be considered for multiple linear regression analysis and which variables reduces the accuracy of the analysis (as shown in Fig. 1). Certain column variables were subsetted for analysis such as date, id, sqft-above and sqft-basement. To check whether missing values are there in the dataset is.na() command was used that will return a boolean values. To identify the number of duplicated values the duplicated function was used. The price values were having values which were in billion and were expressed in ten thousands by dividing by 1000000 for ease of understanding the data. [4]

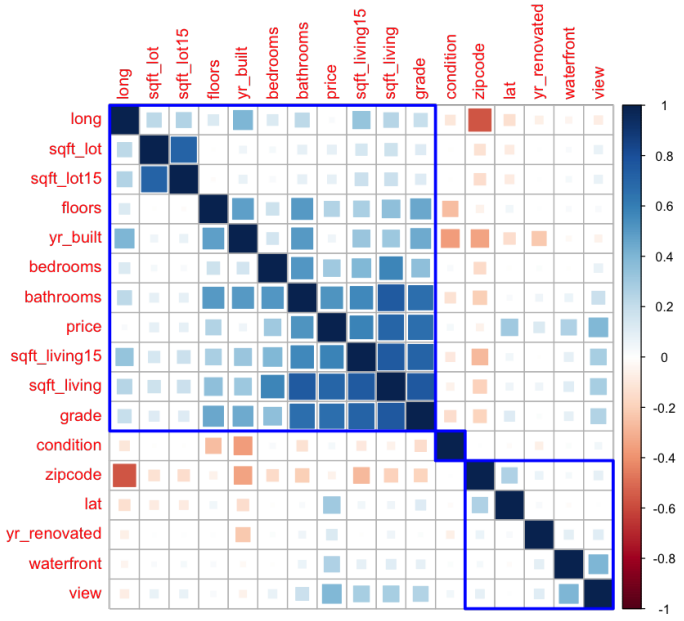

Fig. 3. Correlation Plot

V. Model Analysis And Review

In order to create a linear model lm() function is used. For selecting the best model, 3 different models were analysed.

A. Model 1

For designing this model, the parameters that were considered for multiple linear regression are:

1) sqft-living
2) grade
3) sqft-living15

The summary of the model was found using summary function. The figure below is summary of model 1.

This gives the information about residuals, coefficents, F-statistic and R square values. To further understand the model, the diagnostic plots were plotted. The diagnostic plots depicts the preliminary assumptions of multiple linear regression.

```
Call:
lm(formula = price ~ sqft_living + grade + sqft_living15, data = kc_house_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.359  -1.389  -0.251   1.007  48.398

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.943e+00  1.334e-01 -44.556  < 2e-16 ***
sqft_living    1.784e-03  3.247e-05  54.934  < 2e-16 ***
grade          9.560e-01  2.368e-02  40.370  < 2e-16 ***
sqft_living15  1.581e-04  4.014e-05   3.938 8.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.504 on 21604 degrees of freedom
Multiple R-squared:  0.5348,    Adjusted R-squared:  0.5348
F-statistic:  8280 on 3 and 21604 DF,  p-value: < 2.2e-16
```

Fig. 4. Model 1 Summary

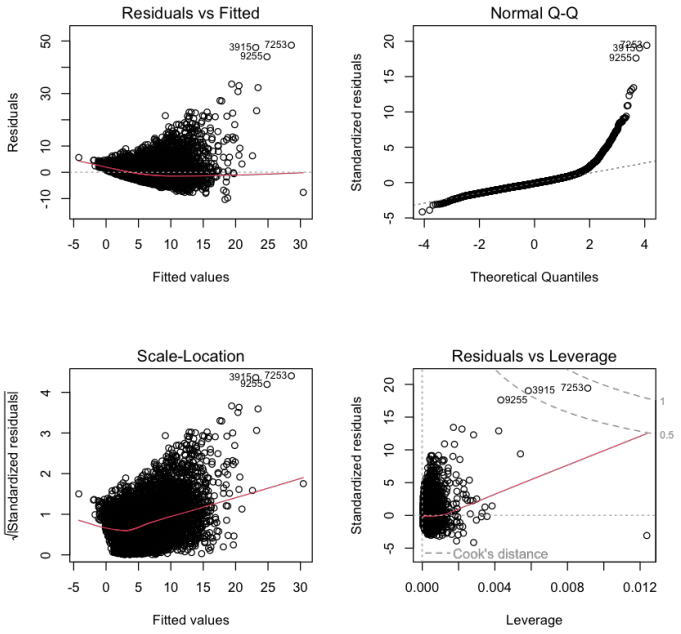

Fig. 5. Model 1 Diagnostic plot

Linearity test :- The Residuals vs Fitted graphs says about linearity. There must not be any relation between residuals and fitted values and there must not be curvature in plot. This model satisfies linearity assumption.

Normal distribution of errors test :- The Normal Q-Q plot says about normal distribution of errors. A probability plot of the standardized residuals against the values that would be anticipated assuming normalcy is shown in the normal Q-Q plot. The points on this graph should lie on a straight line at an inclination of 45-degree if the normalcy assumption has been satisfied. Hence, this model does not fulfills normality assumption.

No influential data points :- The influential data points could be identified using cook's distance. If cook's distance is greater than 1 then it is said to have influential data points. The same can be observed in scale-location graph and residuals vs leverage graph .This model has influential data points.

B. Model 2

For designing this model, the parameters that were considered for multiple linear regression are:

1) bedrooms
2) bathrooms
3) sqft-living
4) floors
5) waterfront
6) view
7) condition
8) grade
9) yr-built
10) yr-renovated
11) sqft-living15
12) sqft-lot15

The summary of model 2 is given in the figure below:

```
Call:
lm(formula = price ~ sqft_living + grade + sqft_living15, data = kc_house_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.359  -1.389  -0.251   1.007  48.398

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.943e+00  1.334e-01 -44.556  < 2e-16 ***
sqft_living    1.784e-03  3.247e-05  54.934  < 2e-16 ***
grade          9.560e-01  2.368e-02  40.370  < 2e-16 ***
sqft_living15  1.581e-04  4.014e-05   3.938 8.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.504 on 21604 degrees of freedom
Multiple R-squared:  0.5348,    Adjusted R-squared:  0.5348
F-statistic:  8280 on 3 and 21604 DF,  p-value: < 2.2e-16
```

Fig. 6. Model 2 Summary and diagnostic plot

The diagnostic plots for model 2 is given below:

Linearity test :- for this case, the model also satisfies linearity condition

Normal distribution of errors test :- The normal Q-Q plot obtained is non-linear and positively skewed, hence normality test fails here.

No influential data points :- There is still influential data points which are in the form of outliers, hence it needs to be removed.

C. Model 3

For designing this model, the parameters that were considered for multiple linear regression are:

1) bedrooms
2) bathrooms
3) floors
4) waterfront
5) view
6) condition
7) grade
8) yr-built

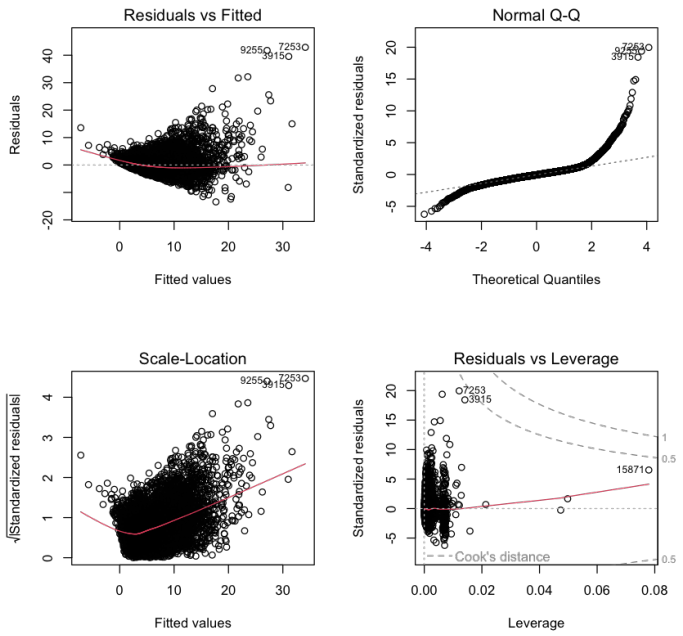

Fig. 7. Model 2 Diagnostic plot

9) yr-renovated
10) sqft-living
11) sqft-lot

The summary of model 3 is given below:

Feature transformation was done in price variable in this case because it the plot of Q-Q plot was highly skewed. The diagnostic plot of model 3 is given above:

Linearity test :- for this case, the model satisfies the linearity condition with slight cuvature in graph.

Normal distribution of errors test :- The normal Q-Q plot obtained is linear and a 45-degree straight line was obtained. The erros are also having constant variance.

No influential data points :- There are very few influential points in the model. [5]

VI. Model Selection and Assumptions

A. Model Selection

Based on the 3 models generated, the model 3 was found to fulfill 3 preliminary assumptions of multiple linear regression. The r-square value of model 3 was 0.645. [6]

B. Assumptions

1) **Durbin Watson Test**:- It is a test to identify autocorrelation between the variables. It ranges from 0 to 4. The durbin watson test for the best model is given below:
It is evident from the results that there is no autocorrelation between the variables and with

```
Call:
lm(formula = log(price) ~ bedrooms + bathrooms + floors + waterfront +
    view + condition + grade + yr_built + yr_renovated + sqft_living +
    sqft_lot, data = kc_house_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81691 -0.21070  0.01755  0.21197  1.38756

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.861e+00  1.990e-01  49.547  < 2e-16 ***
bedrooms     -2.224e-02  2.940e-03  -7.564 4.05e-14 ***
bathrooms     8.363e-02  4.988e-03  16.766  < 2e-16 ***
floors        8.338e-02  4.970e-03  16.778  < 2e-16 ***
waterfront1   3.234e-01  2.704e-02  11.962  < 2e-16 ***
view          5.259e-02  3.230e-03  16.282  < 2e-16 ***
condition     4.319e-02  3.617e-03  11.940  < 2e-16 ***
grade         2.228e-01  3.101e-03  71.844  < 2e-16 ***
yr_built     -5.462e-03  1.022e-04 -53.452  < 2e-16 ***
yr_renovated  9.609e-06  5.681e-06   1.691   0.0908 .
sqft_living   1.661e-04  4.747e-06  34.985  < 2e-16 ***
sqft_lot     -7.941e-08  5.276e-08  -1.505   0.1323
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

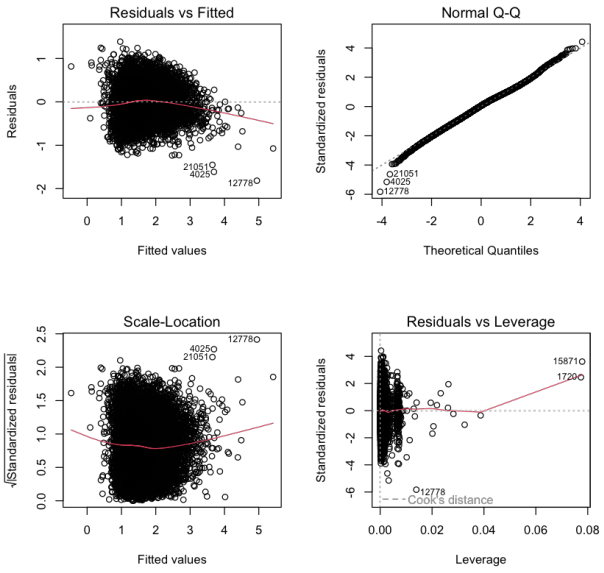

Fig. 8. Model 3 Summary and Diagnostic Plot

```
 lag Autocorrelation D-W Statistic p-value
   1       0.01877143      1.962365   0.008
 Alternative hypothesis: rho != 0
```

Fig. 9. Durbin Watson Test Result

score of 1.96.

2) **Variance Inflation Factor**:- It is also a test to identify correlation between the variables. For analysis if VIF greater than 5 then there is a case of correlation. The VIF plot for the best model is given below:
3) **Non-Constant Variance test**:- It makes a score comparison between the alternative, that the error variance varies with the level of the fitted values, or with a linear combination of predictors, and the hypothesis that the error variance is constant. The NCV test fails in this model. The results for NCV test are as follows:

```
bedrooms    bathrooms       floors   waterfront         view    condition
1.640727     3.238175     1.579444     1.200840     1.343640     1.215785
   grade   yr_built yr_renovated  sqft_living     sqft_lot
2.914401     1.975962     1.141352     4.171170     1.047967
```

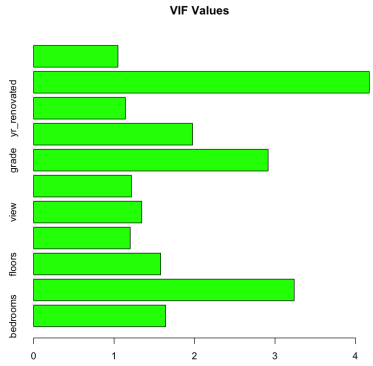

Fig. 10. Variance Inflation Factor and its plot

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.6064468, Df = 1, p = 0.43613
```

Fig. 11. NCV test

VII. Results And Conclusion

After Removing residuals and errors from the model, the fitted model is obtained. The final model also satisfies the Gauss Markov assumption. The figure below shows the fitted model:

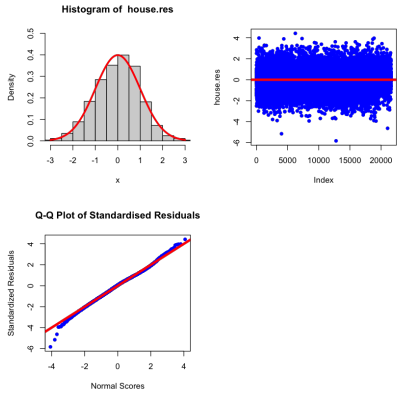

Fig. 12. Fitted Model

The accuracy of the model was evaluated by taking absolute difference between the orignal price values and predicted price values and the difference was divided by the orignal price value. With this approach the accuracy of the model was calculated.

The equation for multiple linear regression for the house price is as follows:

A data.frame: 6 × 2

	pricedt	exp.model3.fitted.values.
	<dbl>	<dbl>
1	221900	321506.9
2	538000	509096.0
3	180000	277054.6
4	604000	436765.0
5	510000	398411.4
6	1225000	1603314.5

0.739105816058778

Fig. 13. Accuracy

price = 9.860 - 0.022 * bedrooms + 0.083 * bathrooms + 0.083 * floors + 0.323 * waterfront + 0.052 * view + 0.043 * condition + 0.223 * grade - 0.005 * yr-built + 0.001 * sqft-living

References

[1] A. Hayes, "Multiple linear regression (MLR) definition, formula, and example," *Investopedia*, 08-Oct-2022. [Online]. Available: https://www.investopedia.com/terms/m/mlr.asp. [Accessed: 19-Nov-2022].

[2] "2014-15 home sales in King County, WA," *GeoDa Data and Lab*. [Online]. Available: https://geodacenter.github.io/data-and-lab/KingCounty-HouseSales2015/. [Accessed: 19-Nov-2022].

[3] "Descriptive statistics in R," *Stats and R*. [Online]. Available: https://statsandr.com/blog/descriptive-statistics-in-r/. [Accessed: 19-Nov-2022].

[4] "Data preprocessing in R," *Section*. [Online]. Available: https://www.section.io/engineering-education/data-preprocessing-in-r/. [Accessed: 19-Nov-2022].

[5] J. C. Watkins, "Organizing and Producing Data," in An Introduction to the Science of Statistics: From Theory to Implementation, Preliminary., pp. 51–56.

[6] Clifford M. Hurvich Chih—Ling Tsai (1990) The Impact of Model Selection on Inference in Linear Regression, The American Statistician, 44:3, 214-217, DOI: 10.1080/00031305.1990.10475722