**QMB-6304 Analytical Methods for Business**

**Multiple Regression Project**

(Modelling the Price for Houses based on different Properties)

Members Involved:

Vamsi Bhargav Sattenapalli (U32769556)

Adithya Guntupalli (U22921880)

Himaja Medarametla (U93612401)

Citation: <https://www.kaggle.com/datasets/shivachandel/kc-house-data>

There are many variables in the original dataset. The dataset is edited and transformed according to our project requirements and many of the unwanted columns are removed.

Variables used in the Dataset:

1. Price: Price variable has the continuous data points which are 6 digits and can reach up to 8 digits.
2. Year Built: This dataset contains years from 1900 to 2014.
3. Sqft Living: This is a continuous variable which has values ranging from 100’s of square meters to 1000’s of square meters of living area in a house excluding the parking lot available for that house.
4. Condition: This is a two-level nominal/ordinal variable which has values in either 1 or 0.  
   1 states that the house is in a good condition and   
   0 states that the house is in a bad condition.

The aim of the project is to model the price of the house based on the properties like year built, square feet area and condition of the house.   
Price is a dependent continuous variable,  
Year Built is an independent continuous variable,  
Square Feet area is an independent continuous variable,  
Condition is also an independent of nominal/ordinal data type.

Screenshot of the Dataset after loaded into the R Studio.  
Table

Description automatically generated

Only few data points are shown in the above screenshot. However, going forward screenshots of 100 data points which are randomly generated using the U number are cited.

Starting with the pre-processing of the data into R studio to change the column names, creating sample data based on U number and attaching the dataset into R studios memory for instant access.

Graphical user interface, text, application, email

Description automatically generated

Output:

Graphical user interface, text, application

Description automatically generated  
Text

Description automatically generated

* Plotting the data points against each other to know how the data is distributed/spread.

Text

Description automatically generated  
Chart, scatter chart

Description automatically generated

Chart, histogram

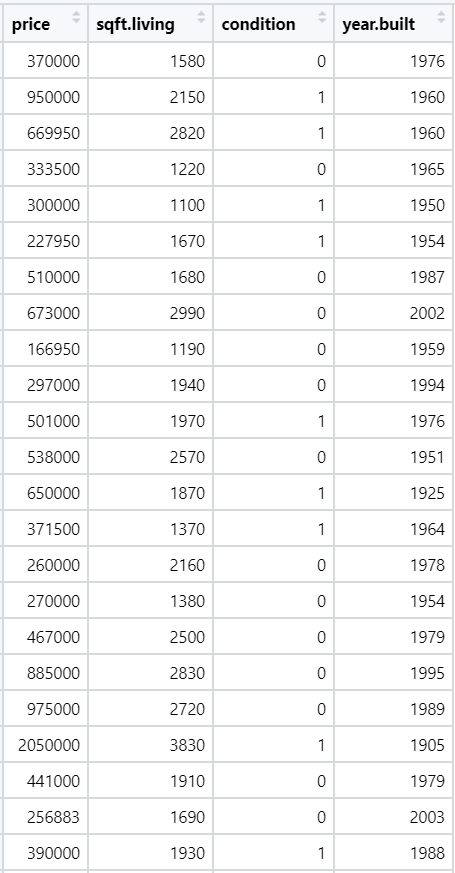
Description automatically generated  
Chart

Description automatically generated  
Chart, histogram

Description automatically generated

* The above three histogram shows that the square feet area of the dataset is not normally distributed and it is skewed to the right; year.built is almost normally distributed.

Table

Description automatically generated

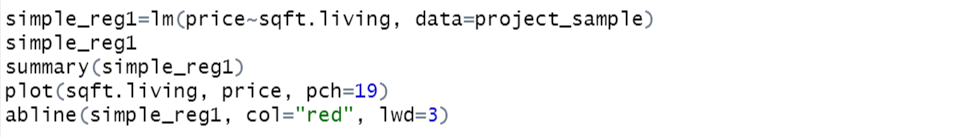
Table

Description automatically generated

Table

Description automatically generated

**Simple Regressions (y, X1), (y, X2), (y, X3):**



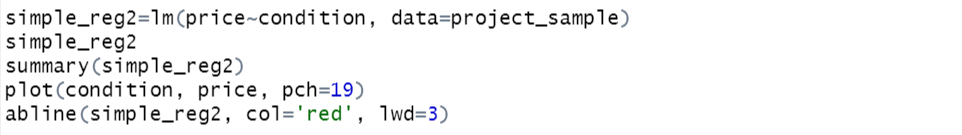
Output:

Text

Description automatically generated  


Chart, scatter chart

Description automatically generated



Output:

Text

Description automatically generated  


A picture containing text

Description automatically generated

Text

Description automatically generated with low confidence

Output:

Text

Description automatically generated  


Chart, scatter chart

Description automatically generated

**simple\_reg1:**

This regression model uses price as dependent variable and sqft.living as independent variable.  
The estimated linear regression model developed using the variables is  
 price=-81263 + 300.9(sqft.living)

The intercept variable in this model is not significant but the sqft.living is highly significant since the p-value is less than 5% i.e. 2e-16

The R2 (Coefficient of Correlation) is 56.16 and  
The Adjusted R2 is 55.72 and  
The Residual Standard error is 230900

Although the coefficient of correlation is above 50% (which can explain the variation in the data) the regression model is not a good fit for the data set since the residual standard error is very high.

The plot also shows that the data points are spread along the regression line not so closely.

**simple\_reg2:**

This regression model uses price as dependent variable and condition as independent variable.  
The estimated linear regression model developed using the variables is  
 price=515487 + 18759(condition)

The intercept variable in this model is significant but the condition is not significant since the p-value is more than 5% i.e. 0.79

The R2 (Coefficient of Correlation) is 0.07 and  
The Adjusted R2 is -0.9 and  
The Residual Standard error is 348600

The regression model is not a good fit for the data set since the coefficient of correlation and the adjusted R2 is zero percent, and the residual standard error is very high.

The plot shows that the data is not linear and when tried to model a linear regression model it does not overlap on the data points.

**simple\_reg3:**

This regression model uses price as dependent variable and year.built as independent variable.  
The estimated linear regression model developed using the variables is  
 price=8068226 - 3836(year.built)

Both the intercept variable in this model and the year.built are significant since the p-value is less than 5% i.e., 0.002 and 0.003 respectively.

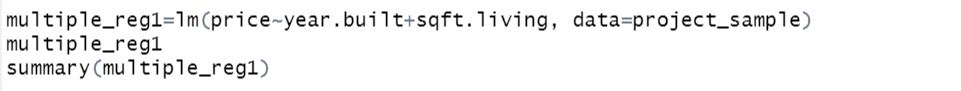
The R2 (Coefficient of Correlation) is 8.19% and  
The Adjusted R2 is 7.2% and  
The Residual Standard error is 334200

The regression model is not a good fit for the data set since the coefficient of correlation and the adjusted R2 is zero percent, and the residual standard error is very high.

The plot shows that the data is linear and when tried to model a linear regression model it does overlap on some of the data points, but it has a negative slope.

**Multiple Regressions: (y, X1, X2), (y, X1, X3), and (y, X2, X3)**

**multiple\_reg1 (y, X1, X2):**



Output:

Text

Description automatically generated

Analysis:  
This regression model uses price as dependent variable and year.built, sqft.living as the independent variables. All the beta coefficients are highly significant to reject the null hypothesis and accept the alternate hypothesis since the p-values are 2.35e-12, 1.72e-12, 2e-16 respectively.

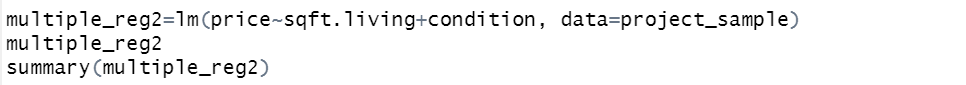
The multiple regression model generated with the above independent variables is   
 price=1.110e+07 -5.717e+03(year.built) + 3.301e+02(sqft.living)

The R2 (Coefficient of Correlation) is 73.82% and  
The Adjusted R2 is 73.29% and  
The Residual Standard error is 179300

This model can be a good fit for the dataset to predict/develop a linear regression since the coefficient of correlation is high enough to explain the variation in the errors. The Residual Standard error is also comparatively low.

With 1 unit change in the independent variables the total change is the product of 1 unit times its beta coefficient value.

**multiple\_reg2 (y, X2, X3):**



Output:

Text

Description automatically generated

Analysis:  
This regression model uses price as dependent variable and sqft.living, condition as the independent variables. Only the beta coefficient sqft.living is highly significant to reject the null hypothesis and accept the alternate hypothesis since the p-values are 0.0665, 2e-16, 0.1777 respectively.

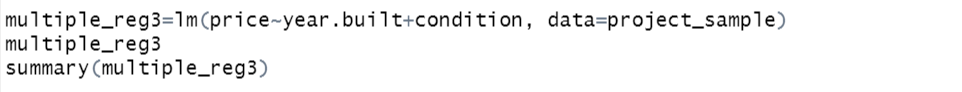
The multiple regression model generated with the above independent variables is  
 price= -122795.34 + 303.94(sqft.living) + 63107.64(condition)

The R2 (Coefficient of Correlation) is 56.98% and  
The Adjusted R2 is 56.09% and  
The Residual Standard error is 229900

The model can be good fit since the coefficient of correlation (R2) is more than 50% which explains atleast half of the variations due to the errors. The residual standard error is relatively high at 229900 and with the addition of condition variable, the linearity has drastically decreased since the condition value are not continuous.

With 1 unit change in the independent variables the total change is the product of 1 unit times its beta coefficient value.

**multiple\_reg3 (y, X1, X2, X3):**



Output:

Text

Description automatically generated with medium confidence

Analysis:  
This regression model uses price as dependent variable and year.built, condition as the independent variables. Only the intercept and first beta coefficient are highly significant to reject the null hypothesis and accept the alternate hypothesis since the p-values are 0.00172, 0.00307, 0.44310 respectively.

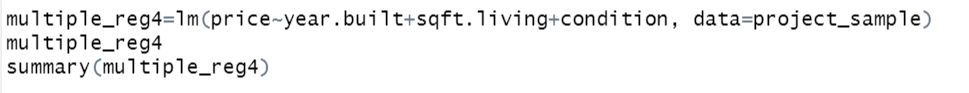
The multiple regression model generated with the above independent variables is  
 price= 8810019 - 4198(year.built) - 55237(condition)

The R2 (Coefficient of Correlation) is 8.75% and  
The Adjusted R2 is 6.86% and  
The Residual Standard error is 334900

This model is not a good fit since the coefficient of correlation (R2) is less than 50% which cannot explain atleast half of the variations due to the errors. The residual standard error is relatively high at 334900 and with the addition of condition variable, the linearity has drastically decreased since the condition value are not continuous.

With 1 unit change in the independent variables the total change is the product of 1 unit times its beta coefficient value.

**multiple\_reg4 (y, X1, X2, X3):**



Output:

Text

Description automatically generated

Analysis:  
This regression model uses price as dependent variable and year.built, sqft.living and condition as the independent variables. Only the intercept, first beta, second beta coefficients are highly significant to reject the null hypothesis and accept the alternate hypothesis since the p-values are 5.22e-12, 3.40e-12, 2e-16, 0.322 respectively.

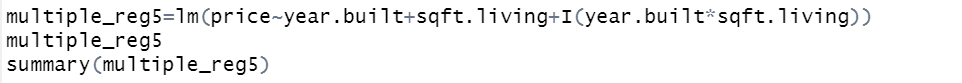
The multiple regression model generated with the above independent variables is  
 price= 1.161e+07 – 5.964e+03(year.built) + 3.295e+02(sqft.living) – 3.829e+04(condition)

The R2 (Coefficient of Correlation) is 74.09% and  
The Adjusted R2 is 73.28% and  
The Residual Standard error is 179400

This model is a very good fit since the coefficient of correlation (R2) is more than 50% which can explain more than half of the variations due to the errors. The residual standard error is relatively low at 179400 and with the addition of condition variable, the linearity has drastically decreased since the condition value are not continuous.

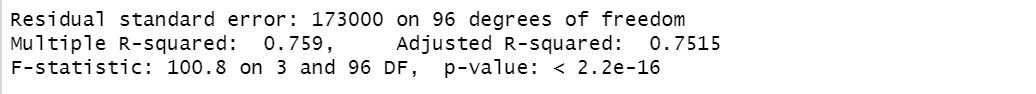
With 1 unit change in the independent variables the total change is the product of 1-unit times its beta coefficient value.

**multiple\_reg5 (y, X1, X2, X1\*X2):**



Output:

Text

Description automatically generated  


Analysis:  
This regression model uses price as dependent variable and year.built, sqft.living and year.built\*sqft.living as the independent variables. Only the second beta and third beta coefficients are highly significant to reject the null hypothesis and accept the alternate hypothesis since the p-values are 0.35113, 0.33616, 0.00226, 0.00496 respectively.

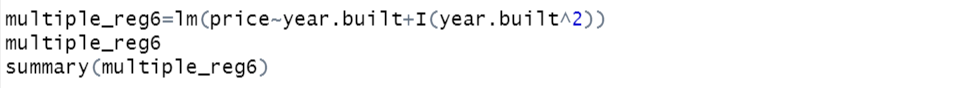
The multiple regression model generated with the above independent variables is  
price= 2.938e+06 – 1.548e+03(year.built) + 3.948e+03(sqft.living) – 1.847(year.built\*sqft.living)

The R2 (Coefficient of Correlation) is 75.9% and  
The Adjusted R2 is 75.15% and  
The Residual Standard error is 173000

This model is a very good fit since the coefficient of correlation (R2) is more than 50% which can explain more than half of the variations due to the errors. The residual standard error is relatively low at 173000

With 1 unit change in the independent variables the total change is the product of 1 unit times its beta coefficient value.

**multiple\_reg6 (y, X1, X1^2):**



Output:

Text

Description automatically generated

Analysis:  
This regression model uses price as dependent variable and year.built, year.built^2 as the independent variables. The intercept, first beta, second beta coefficients are highly significant to reject the null hypothesis and accept the alternate hypothesis since the p-values are 1.47e-05, 1.65e-05, 1.83e-05 respectively.

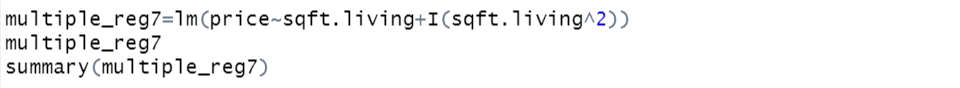
The multiple regression model generated with the above independent variables is  
 price= 6.633e+08 – 6.734e+05(year.built) + 1.710e+2(year.built^2)

The R2 (Coefficient of Correlation) is 24.1% and  
The Adjusted R2 is 22.53% and  
The Residual Standard error is 305400

This model is not a good fit since the coefficient of correlation (R2) is less than 24.1% which cannot explain more than half of the variations due to the errors. The residual standard error is relatively high at 305400.

With 1 unit change in the independent variables the total change is the product of 1-unit times its beta coefficient value.

**multiple\_reg7 (y, X2, X2^2):**



Output:

Text

Description automatically generated

Analysis:  
This regression model uses price as dependent variable and sqft.livin, sqft.living^2 as the independent variables. The intercept, first beta, second beta coefficients are highly significant to reject the null hypothesis and accept the alternate hypothesis since the p-values are 0.000139, 0.024162, 1.28e-06 respectively.

The multiple regression model generated with the above independent variables is  
 price= 4.756e+05 – 2.505e+02(sqft.living) + 1.155e-01(sqft.living^2)

The R2 (Coefficient of Correlation) is 65.62% and  
The Adjusted R2 is 64.91% and  
The Residual Standard error is 205500

This model is a very good fit since the coefficient of correlation (R2) is less than 24.1% which can explain more than half of the variations due to the errors. The residual standard error is relatively high at 205500.

With 1 unit change in the independent variables the total change is the product of 1-unit times its beta coefficient value.

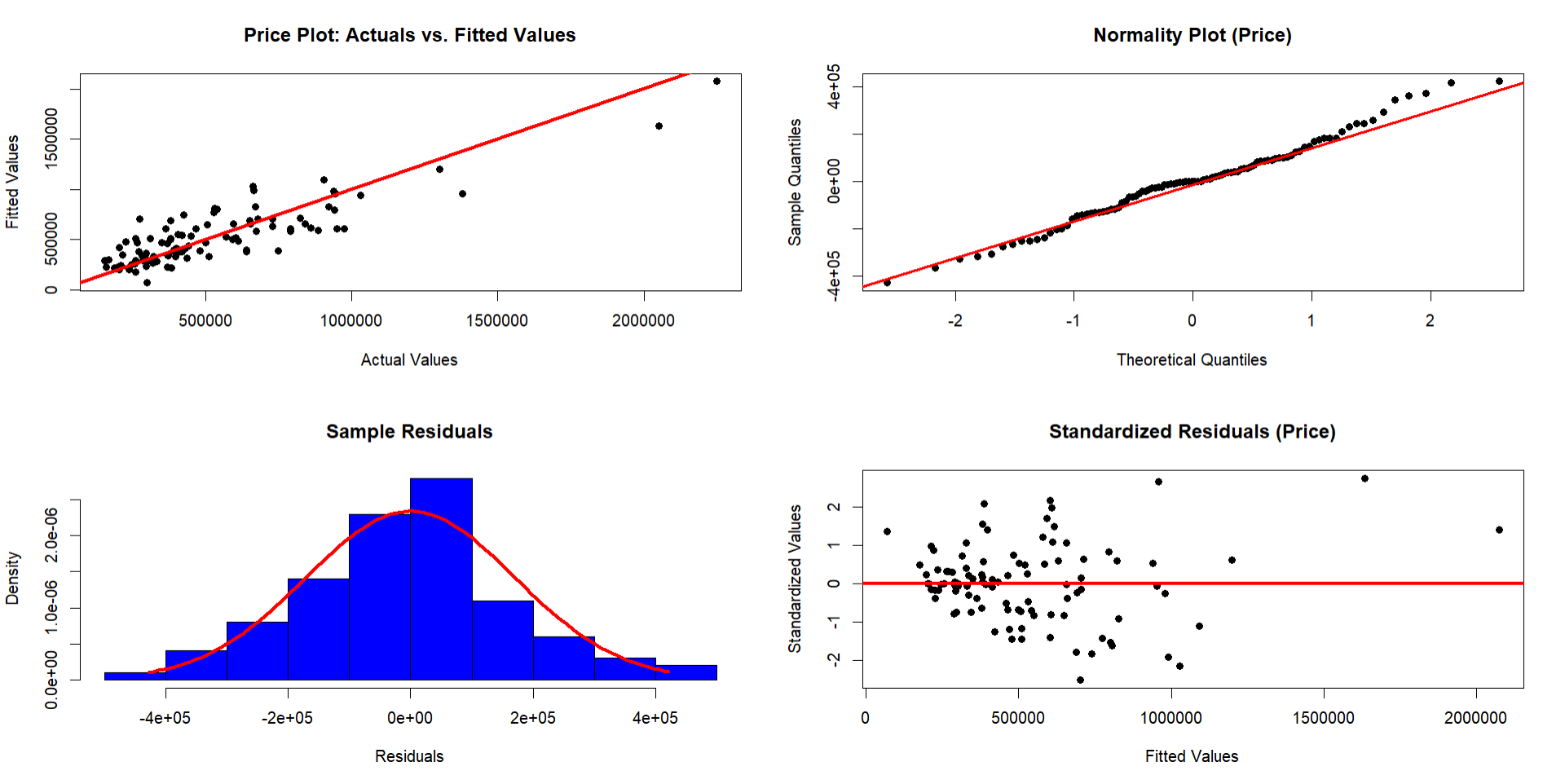
Based on the above regression models, keeping in mind the values of coefficient of correlation, adjusted R2 value, p-values of coefficients and the Residual error value, we conclude that the multiple\_reg5 is the best fit model for the dataset we have taken. The LINE assessment is done below for the same regression model i.e., multiple\_reg5.

**Assessment of L.I.N.E:**

Graphical user interface, text, application

Description automatically generated

Output:



Text

Description automatically generated with medium confidence

Analysis:

Linearity:

When a plot is drawn for original price values to the fitted values of the price the spread shows that the data is linearly spread and there are few outliers. There are few houses which have very high prices. The data points are equally on both sides of the ab line.

Normality:

The QQ-plot and the Histogram drawn on the sample data using the standardized residuals shows that the residuals are normally distributed throughout the ab line. There are few outliers which are due to the high prices which has more sqft.living area.

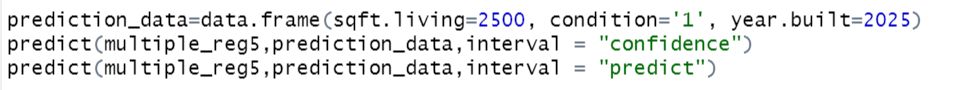
Equality of Variances:

The plot for the equality of variances shows that the data of standardized residuals is Hetero Scedasticity. There are few outliers which are beyond the positive and negative 2 standard deviation range.

Independence of Residuals:

The Durbin Watson value for the simple regression model developed on the initial data before deseasoning is 0.27 which lies farther to 2.5 value which shows that there is positive autocorrelation between the independent variables (year.built and sqft.living). This can be fixed by deseasoning and reseasoning the data.

**Two Types of Predictions (Interval=confidence, predict):**



Output:

Graphical user interface

Description automatically generated with medium confidence

Analysis:

For sqft.living=2500, condition=1 (good), year.built=2025 (future year), when the predict function is run with the developed multiple regression model over the sample data and confidence as interval parameter, the model is 95% confident of getting the price as $321528.4. The upper and lower boundaries are $233620.2 and $409436.6 which are closer to each other showing that the model is accurate and precise since the standard deviation is very low.

When the predict function is run with the same data with predict as the interval, the model predicts that the fitted price value will be 321528.4 and can have range between -32912.74 and 675969.5.

From the above analysis, when year.built variable is added the coefficient becomes negative indicates that the price decreases with years and when sqft.living is added the coefficient becomes positive as area could be the significant parameter in pricing the house. Condition doesn’t play any significant role since houses can be renovated.