This report:

Describes and predicts the relationship between sale prices and other variables.

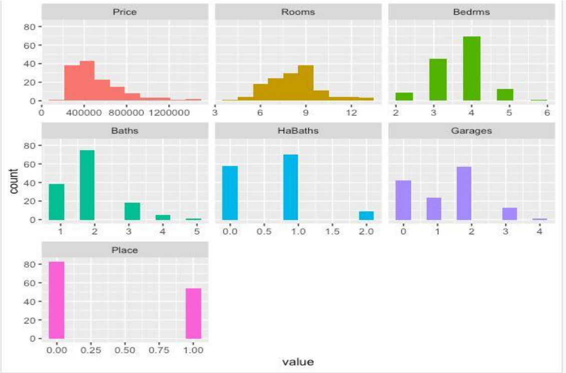
Shows a few steps that how Modelbest comes.

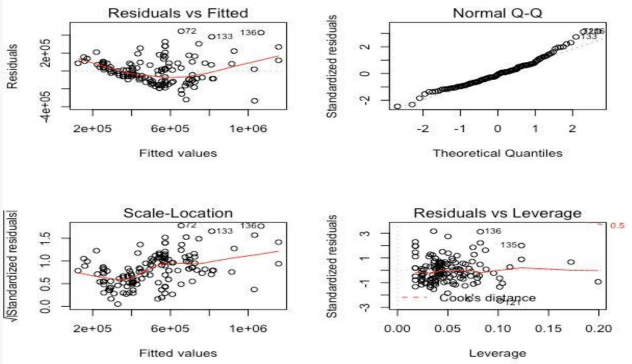
The significance level is 0.05.

As the Place indicator is a qualitative variable, it has been converted to quantitative variable.

1. Overall exploratory analysis

As the plots showed by the left below, most prices are below 800,000 with 6 to 9 rooms, 3 to 5 bedrooms, 1 to 3 baths, 0 or 1 habaths, 1 to 2 garages and more places are from Framington. The first model comes up mind is that trying a model with all variables. The median and the mean of price are 415,000 and 517,512 respectively. This gives a big picture of what the overall price looks like. The mean is much bigger than the median which indicates that the distribution of price is skewed to the right.



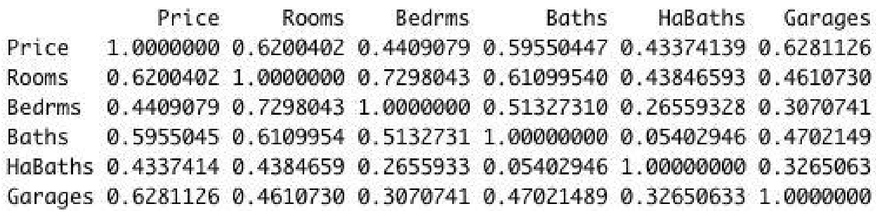


2. Find and omit the outliers

Denote the "ModelOG" as the model included all variables. The plots of ModelOG indicate that the 66th data is an outlier. Take it out from the data set, denote the new model as "Modelnew" ($Price =- 38467 + 17505\times Rooms - 5019\times Bedrms + 102745\times Baths + 103194\times Habathes + 82548\times Garages + 140087\times Place$). From the plots of Modelnew at the upper right, the linearity is not perfectly satisfied and left out in the residuals. Many residuals appear intensively and some spread wider and wider in the Scale-location plot, the red smooth line fluctuates but seems act horizontally, which acts well. The Cook's distance line works well so do not have to omit other data.

3. Interoperation of Fitting

Before fitting the data, it is necessary to determine how the independent variables are related to the price. The correlation provides a quick overview of those relationships. The variables "Garages", "Rooms", "Baths" are the most related variables to the price. However, this does not mean Rooms must pass the significance test in the regression model.

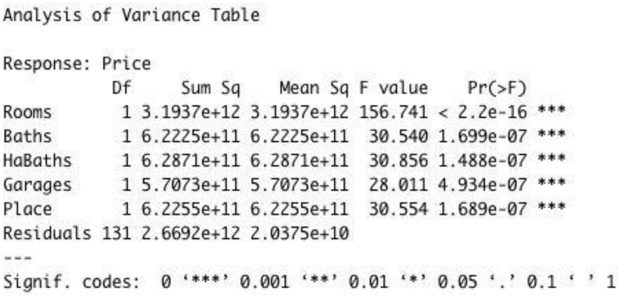


```
            Price     Rooms     Bedrms      Baths     HaBaths    Garages
Price   1.0000000 0.6200402 0.4409079 0.59550447 0.43374139 0.6281126
Rooms   0.6200402 1.0000000 0.7298043 0.61099540 0.43846593 0.4610730
Bedrms  0.4409079 0.7298043 1.0000000 0.51327310 0.26559328 0.3070741
Baths   0.5955045 0.6109954 0.5132731 1.00000000 0.05402946 0.4702149
HaBaths 0.4337414 0.4384659 0.2655933 0.05402946 1.00000000 0.3265063
Garages 0.6281126 0.4610730 0.3070741 0.47021489 0.32650633 1.0000000
```

So far, the Modelnew ($Price =-38467 + 17505\times Rooms - 5019\times Bedrms + 102745\times Baths + 103194\times Habathes + 82548\times Garages + 140087\times Place$) is under consideration.

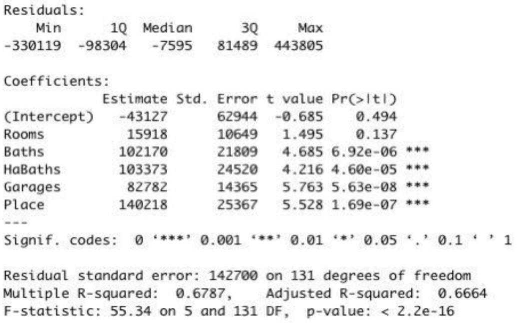
4. *Select the "best" model.*

Summarize the Modelnew to interpret p-value for Bedrms, find that it is much bigger than the significance level. Think about a "ModelFull" without Bedrms. Step the Modelnew and get the ModelFull: $Price = -43127 + 15918\times Rooms + 102170\times Baths + 103373\times Habathes + 82782\times Garages + 140218\times Place$, which proves the previous thought and has a minimum AIC.



```
Analysis of Variance Table

Response: Price
           Df     Sum Sq    Mean Sq F value    Pr(>F)    
Rooms       1 3.1937e+12 3.1937e+12 156.741 < 2.2e-16 ***
Baths       1 6.2225e+11 6.2225e+11  30.540 1.699e-07 ***
HaBaths     1 6.2871e+11 6.2871e+11  30.856 1.488e-07 ***
Garages     1 5.7073e+11 5.7073e+11  28.011 4.934e-07 ***
Place       1 6.2255e+11 6.2255e+11  30.554 1.689e-07 ***
Residuals 131 2.6692e+12 2.0375e+10                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
Residuals:
    Min      1Q  Median      3Q     Max 
-330119  -98304   -7595   81489  443805 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -43127      62944  -0.685    0.494    
Rooms          15918      10649   1.495    0.137    
Baths         102170      21809   4.685 6.92e-06 ***
HaBaths       103373      24520   4.216 4.60e-05 ***
Garages        82782      14365   5.763 5.63e-08 ***
Place         140218      25367   5.528 1.69e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 142700 on 131 degrees of freedom
Multiple R-squared:  0.6787,	Adjusted R-squared:  0.6664 
F-statistic: 55.34 on 5 and 131 DF,  p-value: < 2.2e-16
```

5. *F-test and T-test*

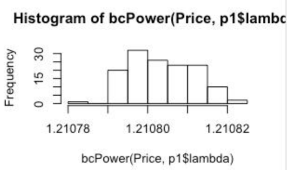
Now check the coefficients by F test with a hypothesis test such that the null hypothesis is all coefficients = 0, the alternative hypothesis ≠ 0. By ANOVA, the code "***" indicates that the significance is extremely small. Consider it as 0. As the p-value is smaller than 0.05, reject the null hypothesis such that the coefficients are not 0 and the alternative hypothesis works.

Continue to the T-test. The null hypothesis H_{0j} is $\beta_j=0$ versus the alternative hypothesis H_{1j} is $\beta_j \neq 0$, j = 1,2,…,p. Access T-test by summarizing ModelFull. In the summary above, Pr(>ltl) = 0.137 indicates that Rooms cannot pass the T-test. Now take Rooms out from out model, where named "Modelbest" as $Frice = 34748 + 120951\times Baths + 120624\times Habathes + 84971\times Garages + 145369\times Place$. All the coefficients are positive means that when the number of variables increases the Price increases.

6. *Conclusion with prediction of price.*

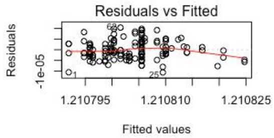
At the least, the Modelbest with yellow marker is the one should be chosen.

After "Box-cox", the histogram of Price becomes normal, (See the right picture)



and the distribution of the residual scattered points has been significantly improved, which basically meets the assumptions.

The "new values" that are chosen to predict the price are medians of each variables in the Modelbest. The predicted price in Framington is 493228.7 and



538053.2 in Natick.