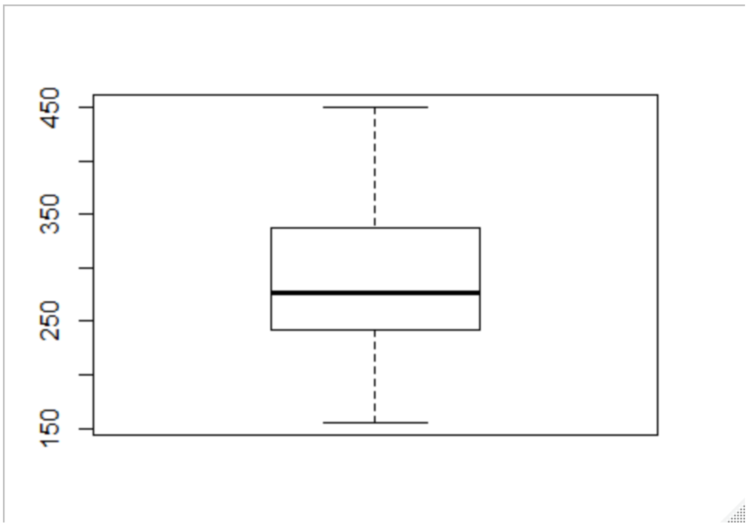
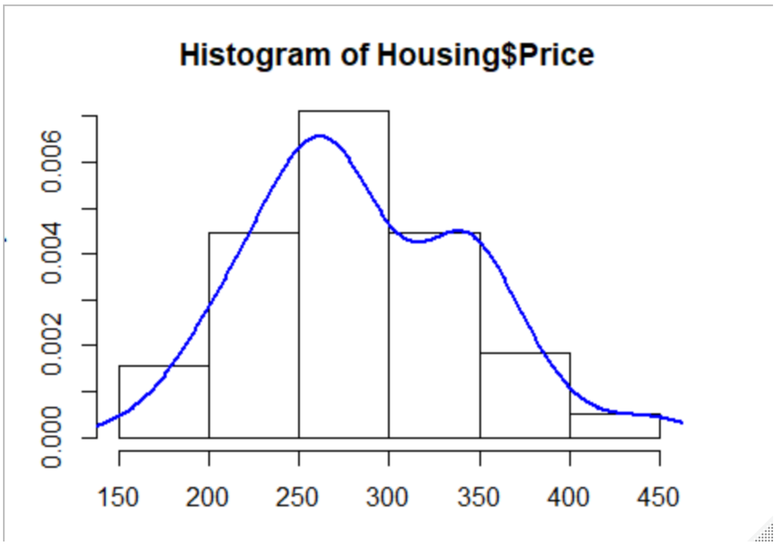
Exploratory Data Analysis:

In order to import correctly I had to use R Studio to remove 'ï»¿' from the beginning of the 'Price' column. R Code has also been included at the end as a safety measure

Q1. Using a boxplot, histogram and summary. Describe the distribution of the sales price of the houses.



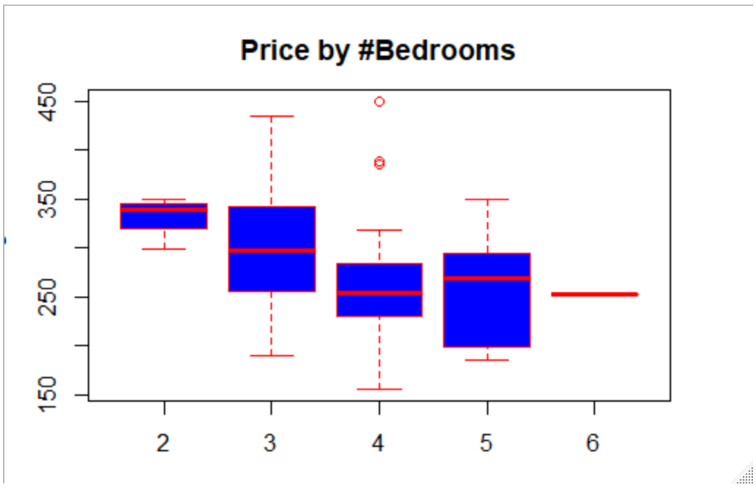


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
155.5	242.8	276.0	285.8	336.8	450.0

The range of house price is $155,500 to $450,000. 50% of house prices fall between $242,800 and $336,800. None of the data appears to be above or below the max/min, as in there are no properties above the $450K mark (1.5 * IQR), i.e no outliers. From the histogram there appears to be a lot of houses just above the $250K dipping at $300K and rising slightly again at ~$350K.

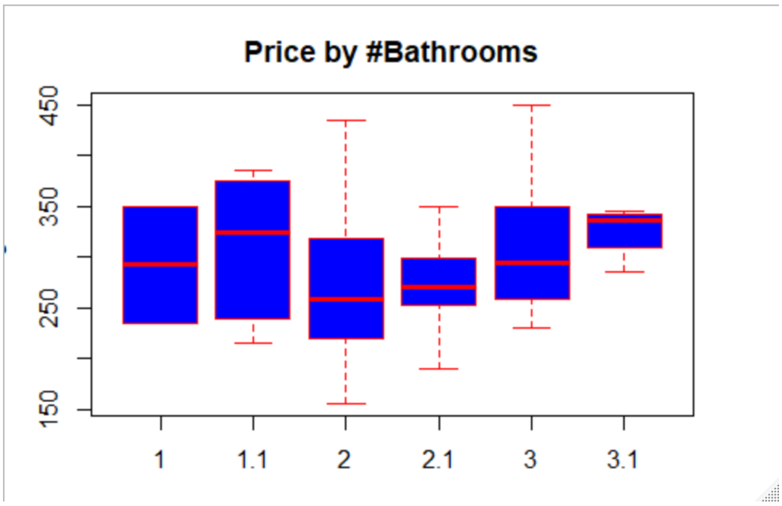
Q2. Convert all the categorical variables to factors. Using the summary and a boxplot describe how sales prices vary with respect to the number of bedrooms, bathrooms, garage size and school.

Since other attributes are already in a numeric representation (i.e. lot size) I have only converted the categorical data School into factors. For comparison I used boxplots grouped by each secondary category.



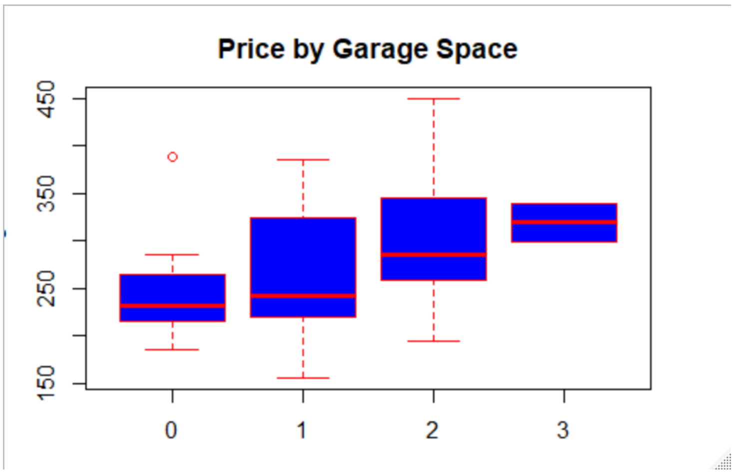

```
$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  299.0   319.4   339.9   329.6   344.9   350.0

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  189.5   256.2   297.0   297.3   342.5   435.0

$`4`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  155.5   231.5   254.4   266.6   283.5   450.0

$`5`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  185.0   199.0   269.0   259.5   295.0   349.5

$`6`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  252.5   252.5   252.5   252.5   252.5   252.5
```

By number of bedrooms, it would appear that price dips until 4 bedrooms and recovers just as quickly. 2 Bedrooms appears to be the highest on average while 4 bedrooms (one of the cheapest on average) has the highest priced home, an outlier. There only appears to be 1 house with 6 bedrooms, this appears to fall in closely with the median for 4 and 5 bedroom homes.



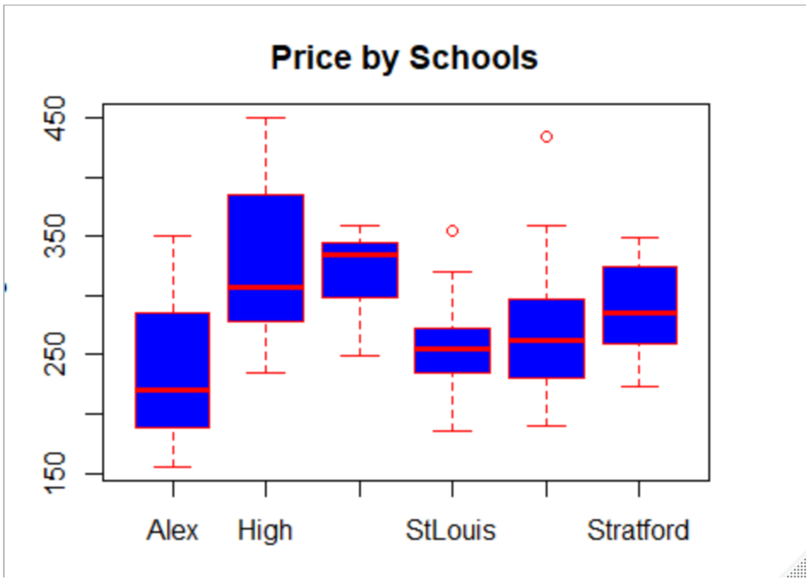

```
$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  235.0   263.8   292.5   292.5   321.2   350.0 

$`1.1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  215.0   239.5   325.0   307.9   374.5   385.5 

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  155.5   220.0   259.0   270.7   319.0   435.0 

$`2.1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  189.5   254.8   269.9   274.5   297.7   349.5 

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  230.0   259.0   295.0   307.8   349.5   450.0 

$`3.1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  285.0   309.4   336.0   324.2   342.5   345.0 
```

Surprisingly homes that had a half-bathroom extra tended to increase the price on average, contradicting the belief that half-bathrooms are not as valued as full bathrooms. The more pricier houses tend to have 1 and a half bathrooms, or 3 bathrooms while 2/2.5 bathroom houses tend to cost less.




```
$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  185.0   216.0   232.0   246.9   264.4   388.0 

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  155.5   220.0   242.0   260.6   324.5   385.5 

$`2`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  195.0   259.0   285.0   299.6   343.8   450.0 

$`3`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  299.0   309.2   319.4   319.4   329.7   339.9 
```

Not surprisingly, the number of car spaces increases the house price. Oddly, there is an outlier with 0 car spaces that is unusually high. Having 1-2 spaces seems the most common, with a similar range but a higher price.



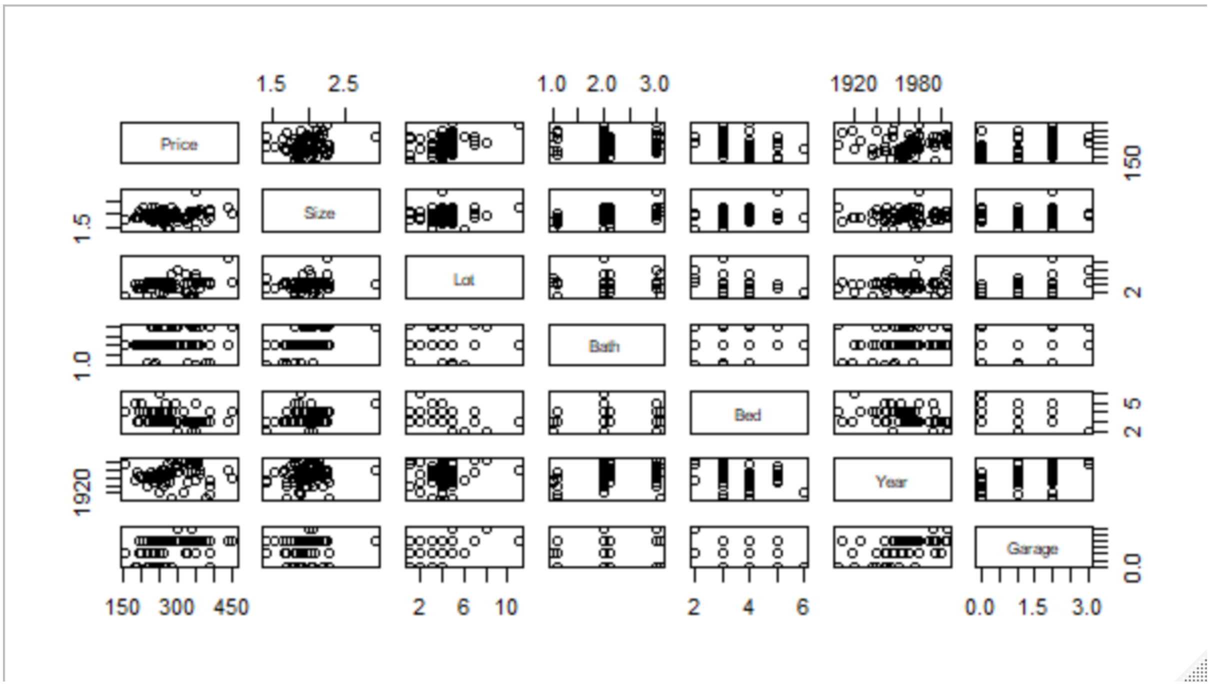

```
$Alex
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  155.5   187.8   220.0   241.8   285.0   350.0 

$High
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  235.0   279.2   307.5   327.1   385.6   450.0 
```

```
$NotreDame
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  249.9   304.0   334.9   319.1   345.0   359.9

$StLouis
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  185.0   235.4   255.0   257.4   272.4   355.0

$StMarys
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  189.5   231.6   262.0   269.8   296.5   435.0

$Stratford
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  222.5   266.2   285.0   287.8   315.0   349.5
```

Comparing the price of homes to schools, it would appear that 2 (High) has the largest impact on school price. While 5 (Stratfor) does has a highly priced outlier, as does 4 (St. Marys), their median is not as high. However, 3 (Notre Dame) does have the largest median price increase.

Q3. Using the summary, correlation and the pairs plots discuss the relationship between the response sales price and each of the numeric predictor variables.

From the correlation matrix we can see that price and garage are the most positively correlated while bed and price are the most negatively correlated. Meaning that as the number of beds increase, the price is more likely to decrease.



Regression Model

Q1. Fit a multiple linear regression model to the data with sales price as the response and size, lot, bath, bed, year, garage and school as the predictor variables. Write down the equation for this model.

```
Call:
lm(formula = Housing$Price ~ Housing$Size + Housing$Lot.Type.f +
    Housing$Bath.Type.f + Housing$Bed + Housing$Year + Housing$Garage +
    Housing$School.Type.f, data = Housing)

Residuals:
    Min      1Q  Median      3Q     Max
-83.626 -18.966   1.722  21.676  70.213

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -855.6150   677.4376  -1.263 0.212111
Housing$Size               41.9278    28.3780   1.477 0.145467
Housing$Lot.Type.f2        24.8812    37.6375   0.661 0.511428
Housing$Lot.Type.f3         5.0597    31.9228   0.158 0.874666
Housing$Lot.Type.f4         9.3078    30.9623   0.301 0.764882
Housing$Lot.Type.f5        35.4518    32.2761   1.098 0.276999
Housing$Lot.Type.f6       279.1165    68.8729   4.053 0.000167 ***
Housing$Lot.Type.f7        38.3898    33.9690   1.130 0.263506
Housing$Lot.Type.f8       -27.4835    55.0113  -0.500 0.619426
Housing$Lot.Type.f11      174.5166    51.9887   3.357 0.001465 **
Housing$Bath.Type.f1.1    142.8145    46.8846   3.046 0.003608 **
Housing$Bath.Type.f2       92.7954    43.6970   2.124 0.038389 *
Housing$Bath.Type.f2.1     94.9218    44.9945   2.110 0.039630 *
Housing$Bath.Type.f3      135.6652    46.0876   2.944 0.004806 **
Housing$Bath.Type.f3.1    107.5644    51.0539   2.107 0.039878 *
Housing$Bed               -11.3629     8.8708  -1.281 0.205794
Housing$Year                0.4473     0.3369   1.328 0.189962
```

```
Housing$Garage                   13.8429    8.4199   1.644 0.106084
Housing$School.Type.fHigh       132.2855   36.0237   3.672 0.000560 ***
Housing$School.Type.fNotreDame   99.8114   34.2670   2.913 0.005234 **
Housing$School.Type.fStLouis     47.1872   34.3808   1.372 0.175692
Housing$School.Type.fStMarys     50.0332   34.0152   1.471 0.147229
Housing$School.Type.fStratford   72.2833   38.7528   1.865 0.067688 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.94 on 53 degrees of freedom
Multiple R-squared:  0.6903,   Adjusted R-squared:  0.5617
F-statistic: 5.369 on 22 and 53 DF,  p-value: 2.82e-07
```

The R code is:

lm(Price ~ Size + Lot + Bath + Bed + Year + Garage + School, data=Housing)

The maths formula would be:

Yi = $\beta_0 + \beta_1 X_{i,1}$ + … + $\beta_{p-1} X_{i,p-1}$

Alternatively:

Price = -855.62 + $\beta_1(Size) + \beta_2(Lot) + \beta_3(Bath) + \beta_4(Bed) + \beta_5(Year) + \beta_6(Garage) + \beta_7(School)$

Q2. Interpret the estimate of the intercept term B0.

The estimated average house price is $-855,615 when Size, Lot, Bath, Bed, Year, Garage, and school are all at 0. However the P-Value is large (> 0.05) but since it is the intercept it could be ignored. For variables they should be considered if removal is better as it affects the models performance. A house with none of the attributes listed would not get a sale price, you have nothing to sell and therefore no meaning should be attached to this intercept value of the constant.

Q3. Interpret the estimate of BSize the parameter associated with floor size (Size).

A one unit increase to Size has an increase of $41,927 (since cost is divided by 100K) to the cost of a house. Since the P-Value for Size is 0.178 > 0.05 we fail to reject the null hypothesis. That is, we do not have enough evidence to support that Size significantly affects the Price of a house.

Q4. Interpret the estimate of BBath1.1 the parameter associated with one and a half bathrooms.

A one unit increase in Bath1.1 would result in $142,815 increase to the property price compared to a 1 bathroom house. Since this is a categorical feature, we have to remember that the first entry is taken as reference i.e. Bathroom 1 is the reference and all other bathroom sizes are compared to it, meaning that value added is in comparison to a 1 Bathroom and not 0 bathrooms. As the P value is 0.003605 < 0.05 it is considered significant. In other words, we have enough evidence to support that Bath1.1 significantly affects the price of a house and can reject the null hypothesis.

Q5. Discuss and interpret the effect the predictor variable bed on the expected value of the house prices.

- From an adjustment made after submission –

It appears that 2 bedroom houses add the most value to the property, for additional beds it appears to reduce the cost, 6 bedrooms has ~double the loss but this may not be accurate market analysis as there is only 1 entry with 6 bedrooms.

```
Housing$Bed.Type.f3              -36.7818     44.1971   -0.832 0.409242
Housing$Bed.Type.f4              -49.5788     45.6583   -1.086 0.282746
Housing$Bed.Type.f5              -46.7850     49.9368   -0.937 0.353322
Housing$Bed.Type.f6              -90.6022     68.9303   -1.314 0.194708
```

Q6. List the predictor variables that are significantly contributing to the expected value of the house prices

```
Variable           Estimate Std. Error t value Pr(>|t|)
Lot 6              279.1165    68.8729   4.053 0.000167 ***
Lot 11             174.5166    51.9887   3.357 0.001465 **
Bath 1.1           142.8145    46.8846   3.046 0.003608 **
Bath 2              92.7954    43.6970   2.124 0.038389 *
Bath 2.1            94.9218    44.9945   2.110 0.039630 *
Bath 3             135.6652    46.0876   2.944 0.004806 **
Bath 3.1           107.5644    51.0539   2.107 0.039878 *
School High        132.2855    36.0237   3.672 0.000560 ***
School NotreDame    99.8114    34.2670   2.913 0.005234 **
School Stratford    72.2833    38.7528   1.865 0.067688 .
```

These variables are the ones that contribute the most to the value of the house prices. Lot 6 contributes the most to expected value of the house prices. Lot 11 also contributes a lot, but considering there is only 1 data entry and it is priced above average its representation could be skewed. Bathroom 1.1 contributes the most from the bathroom category, with bathroom 3 being next. Schools are the next biggest contributors, Stratford is above the 0.05 value, but depending on how strict we need to be with our model I have decided to include it.

Q7. For each predictor variable what is the value that will lead to the largest expected value of the house prices.

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -855.6150   677.4376  -1.263 0.212111
Size                41.9278    28.3780   1.477 0.145467
Lot 6              279.1165    68.8729   4.053 0.000167 ***
Bath 1.1           142.8145    46.8846   3.046 0.003608 **
Bed                -11.3629     8.8708  -1.281 0.205794
Year                 0.4473     0.3369   1.328 0.189962
Garage              13.8429     8.4199   1.644 0.106084
School High        132.2855    36.0237   3.672 0.000560 ***
```

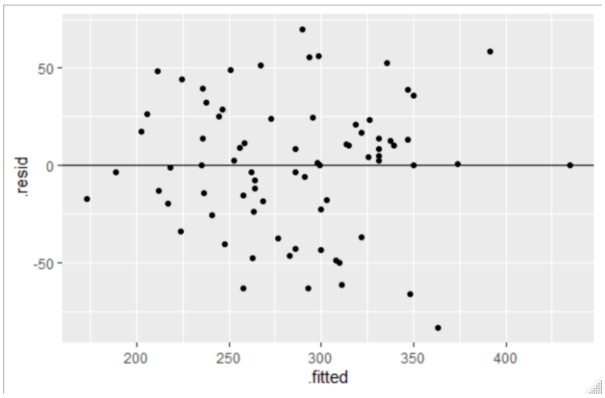
For Size, a one unit increase would increase house price by $41,928 (in the range of [1.44, 2.896]. Lot of category 6 increases the house price by $279,117 in comparison to lot 1. Then, a 1.5 Bathrooms would add $142,815 in price in comparison to having only 1 bathroom. Having a 3 bedroom house would reduce the price of a house by -$66,276 in comparison with a 2 bedroom house, therefore a 2 bedroom house would be better. For the year, the more modern the better the increase ($447 per year from 1905 to 2005). Having a garage with 2 car spaces would add $18,851 to your house in comparison to having no garage/car space. Finally, owning a house near the school the High School would increase its value by $132,286 compared to living next to Alexandra.

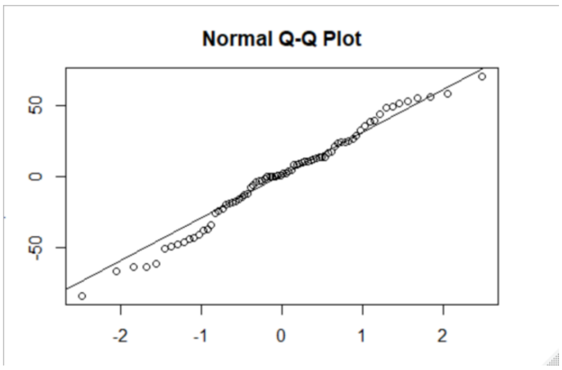
Q8. For each predictor variable what is the value that will lead to the lowest expected value of the house prices.

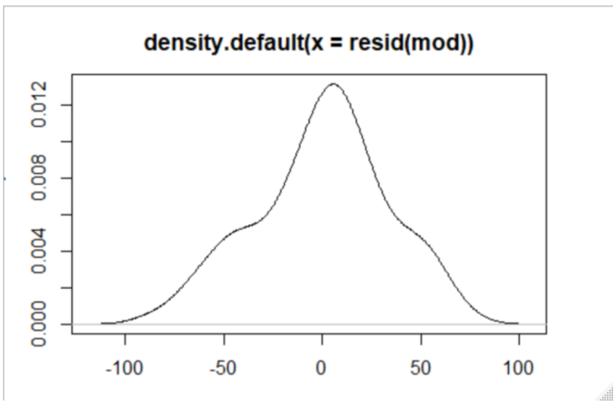
```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -855.6150   677.4376  -1.263 0.212111
Size               41.9278    28.3780   1.477 0.145467
Lot 8             -27.4835    55.0113  -0.500 0.619426
Bath 1
Bed               -11.3629     8.8708  -1.281 0.205794
Year                0.4473     0.3369   1.328 0.189962
Garage             13.8429     8.4199   1.644 0.106084
School Alexandra
```

For the lowest priced house, it would require a relatively small Size, the Lot category should be in 8 as it would reduce the house price by -$27,484 compared to Lot 1. Bathrooms should only have 1 as it will be the lowest price (any other size increases the house price). The number of bedrooms should be 6 as this reduces the house price the most in comparison to having only 2 beds. The year should be as old as possible, 1905. Any year after that adds a value of $447 to the house. Having no car space keeps the house price at its lowest, in comparison to having even just 1 car space. And finally, being located near Alexandra school would have the lowest priced houses than any other school.

9. By looking at the information about the residuals in the summary and by plotting the residuals do you think this is a good model of the expected value of the house prices.







Since the graph of residuals on the left shows homoskedacity, i.e. the variacne from 0 is of ~equal distance in the plus and minus, and the left graph has little variance from the line shows that the model is good. The 3rd graph shows density around the 0 mark which would such a normal distribution.

10. Interpret the Adjusted R-squared value

```
Multiple R-squared:  0.6903,	Adjusted R-squared:  0.5617
```

The R-squared value explains the variance in Y by the model, 0 would mean that none of the variance in Y is explained and 1 would mean all the variance is explained by the model. The adjusted R-square takes into account the number of variables used and is penalised for each one added. While the R-Squared value is at a decent level, 69.03% the gap adjusted on the adjusted r-sqaured of 56.17% would suggest that our model has some random/unnecessary varaibles included in it. The Adjusted R-Squared value suggests that 56.17% of variance in Y is explained by the model.

11. Interpret the F-statistic in the output in the summary of the regression model. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

```
F-statistic: 5.369 on 22 and 53 DF,  p-value: 2.82e-07
```

The Hypothesis is: $H_0: \beta_1$ = 0 vs. H_a: β1 ≠ 0

The test statistic is: 5.39

The P=Value is: 0.000000282 (2.82e-07)

The test statistic is 5.39 and the probability of getting that in an F-distribution consisting of 22 and 53 df (degrees of freedom) is 2.82e-07, since it is less than 0.05 we reject the null hypothesis and conclude that at least one of the variables is non-zero.

ANOVA

Q1 Compute the type 1 ANOVA table. Interpret the output. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

```
Analysis of Variance Table
Response: Housing$Price
                      Df Sum Sq Mean Sq F value    Pr(>F)    
Housing$Size           1  11078 11077.7  6.9438 0.0110053 *  
Housing$Lot.Type.f     8  45378  5672.2  3.5555 0.0022973 ** 
Housing$Bath.Type.f    5  41999  8399.8  5.2652 0.0005348 ***
Housing$Bed            1  21831 21830.9 13.6842 0.0005150 ***
Housing$Year           1   2517  2516.6  1.5775 0.2146302    
Housing$Garage         1   6500  6500.2  4.0745 0.0486087 *  
Housing$School.Type.f  5  59147 11829.4  7.4150  2.47e-05 ***
Residuals             53  84553  1595.3                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA table, we can see the variation in Y explained by the addition of each previous variable. For example, 11,078 of variation in Y is explained by Size when there are no other variables. However, 6,500 of variation in Y is explained by Garage given that Year, Bed, Bath, Lot, and Size are already in the model. Then, 84,553 of the variation in Y is NOT explained by its relationship with all the variables included. The Type I ANOVA table suggests removing 'Year' as it is the only variable that has the least significance.

The Hypothesis is: H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6\ \beta_7 = 0$ vs. H_a: at least one $\beta_k \neq 0$

Given that $F_{0.05,22,53}$ = 0.5256153

With the F-test:

SSR = 11078 + 45378 + 41999 + 21831 + 2517 + 6500 + 59147 = 188449.41

MSR = (SSR/7) = 26921.42

SSE = 84553.06

MSE = SSE/53 = 1595.34

F = 16.87498

Since 16.875 > 0.525 we reject H_0. Which indicates that at least one of the variables are significantly related to Y.

Q2. Which predictor variable does the type 1 ANOVA table suggest you should remove the regression analysis.

```
Analysis of Variance Table
Response: Price
              Df Sum Sq Mean Sq F value    Pr(>F)
Year           1   2517  2516.6  1.5775 0.2146302
Based on the ANOVA type 1 table, the variable 'Year' should be removed as
it has a p-value greater than 0.05 and is not significant.
```

Q3. Compute a type 2 ANOVA table comparing the full model with all predictor variables to the the reduced model with the suggested predictor variable identified in the previous question removed. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem

```
Anova Table (Type II tests) FOR REDUCED MODEL

Response: Price
               Sum Sq Df F value    Pr(>F)
Size             3581  1  2.2131 0.1426584
Lot             54807  8  4.2345 0.0005401 ***
Bath            26817  5  3.3151 0.0110497 *
Bed              4541  1  2.8068 0.0996474 .
Garage           9835  1  6.0787 0.0168910 *
School          56397  5  6.9718   4.35e-05 ***
Residuals       87365 54

Anova Table (Type II tests) FOR NORMAL MODEL
               Sum Sq Df F value    Pr(>F)
Size             3483  1  2.1829 0.1454666
Lot             56661  8  4.4395 0.0003676 ***
Bath            29071  5  3.6445 0.0065922 **
```

```
Bed            2618  1  1.6408 0.2057936
Year           2812  1  1.7628 0.1899620
Garage         4312  1  2.7029 0.1060842
School        59147  5   7.4150  2.47e-05 ***
Residuals     84553 53
```

Hypothesis $H_0: \beta_k = 0$ vs. $H_a: \beta_k \neq 0$

F* = 2.655

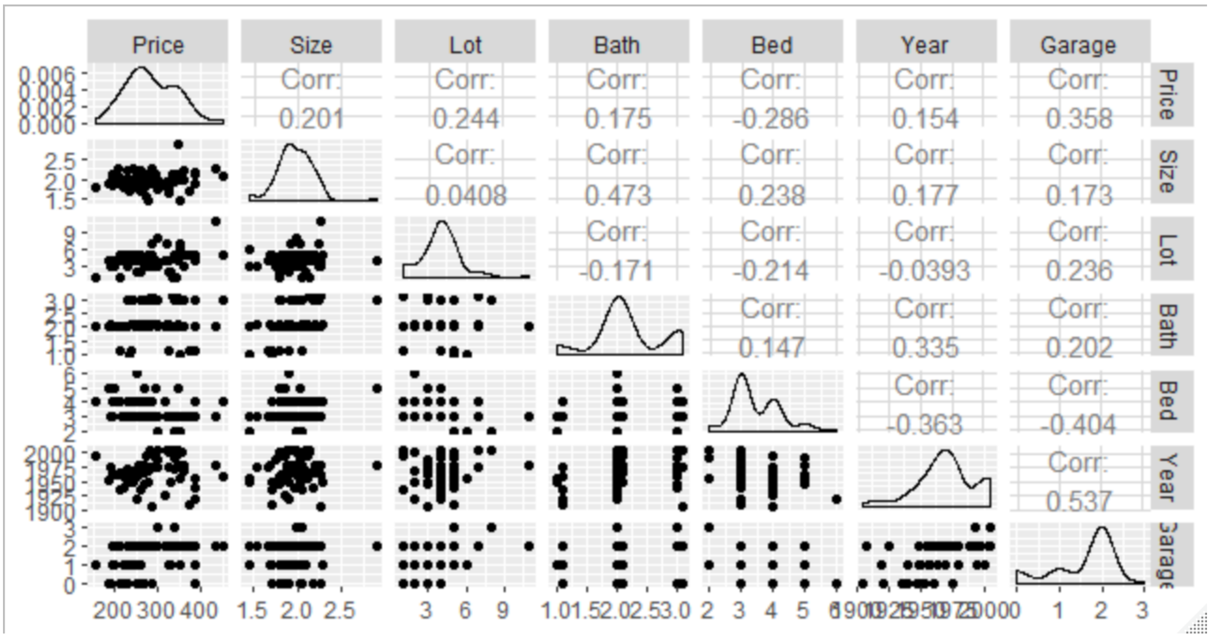
F0.05,21,54 = 0.5188

Since 2.655 > 0.5188 we can reject H_0 i.e that keeping the variable 'Year' does not significantly improve the model and can be removed. This can also be achieved using *anova(mod, red_mod)* as it will perform an F test comparison of them both. This gave me a p-value of 0.19 and since 0.19 > 0.05 it is recommended to reject the second model).

Diagnostics

Since the 'Year' variable was not significant, and by recommendation of the attached video I have removed it and will be using 'red_mod' (model without 'Year') going forward.

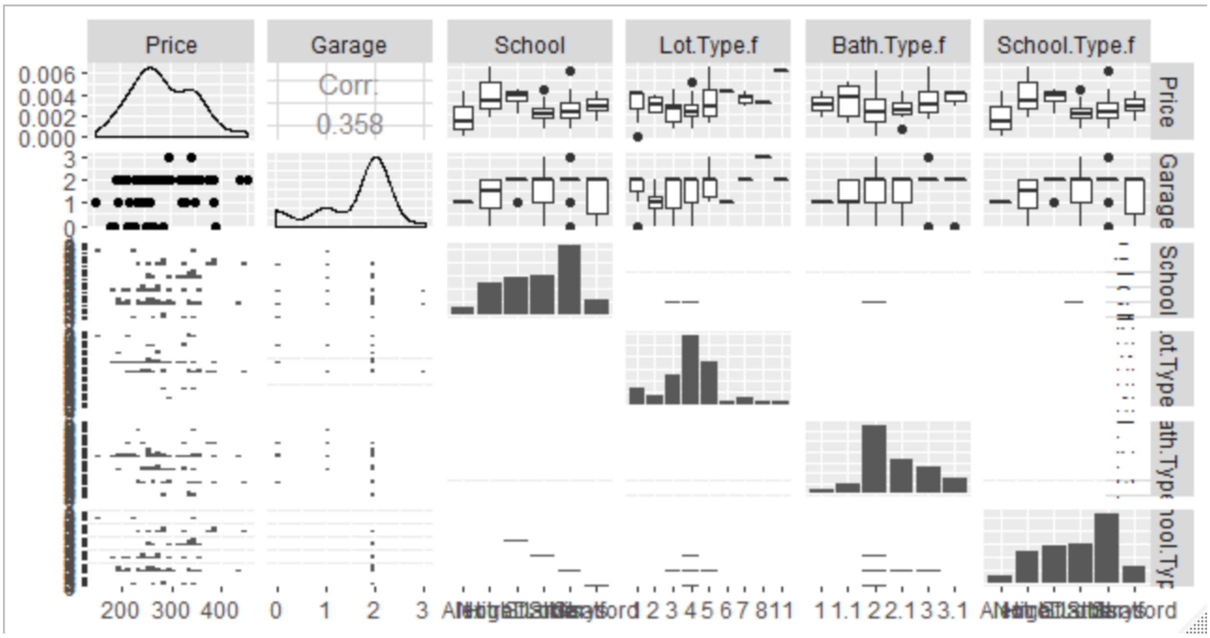
Q1 Check the linearity assumption by interpreting the added variable plots and component-plus-residual plots. What effect would non-linearity have on the regression model and how might you correct or improve the model in the presence of non-linearity?



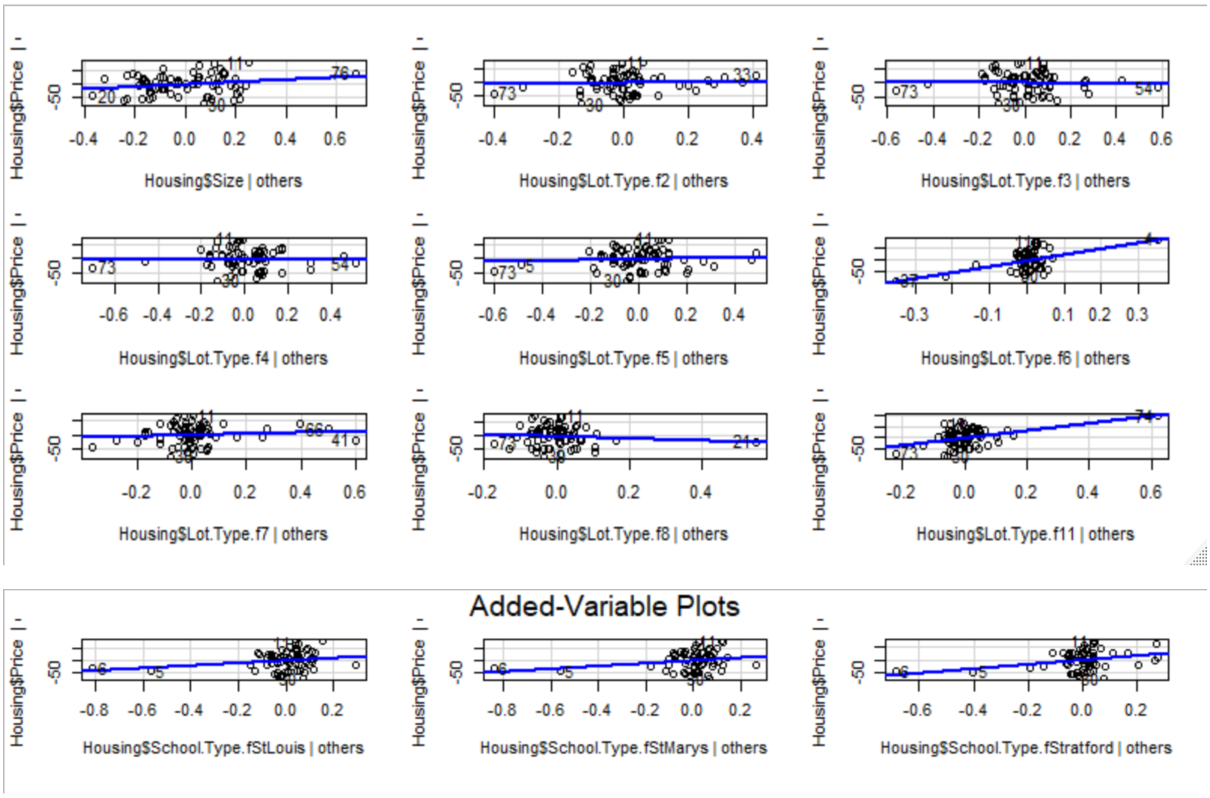
Predictor Variable	Type	Comments

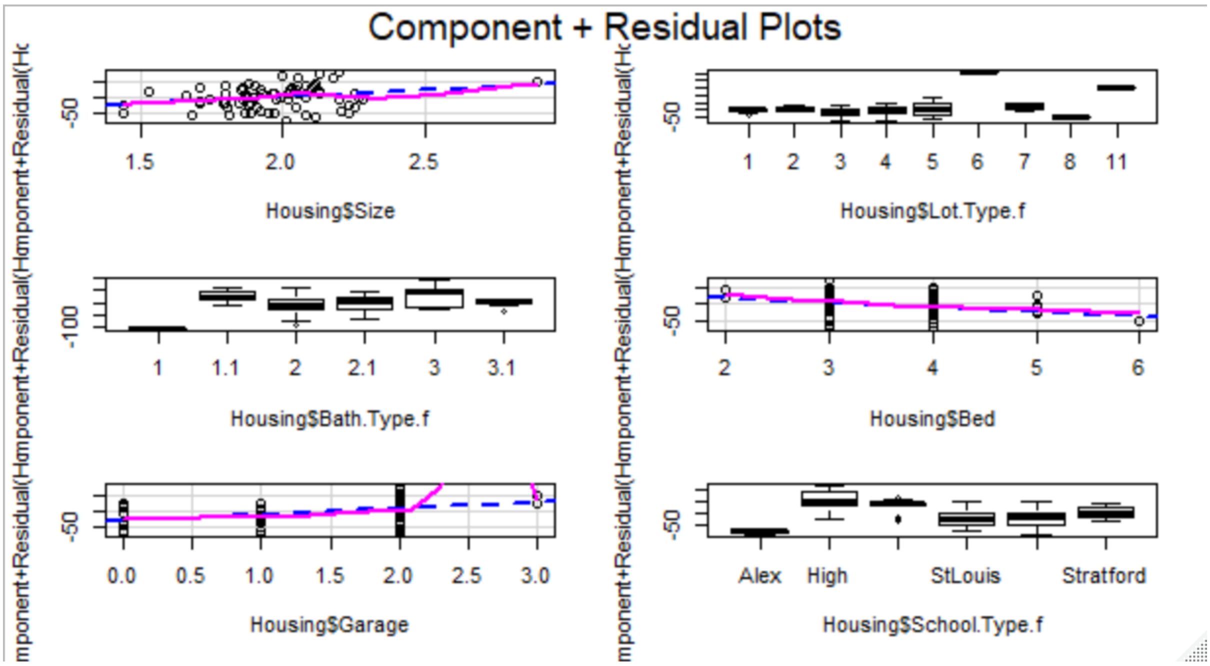
Size	Linear	Linear, but does not appear to have a positive or negative correlation so the line may just go straight across
Lot	Linear	Perhaps slightly positively correlation
Bath	Linear	Categorical but appears linear
Bed	Linear	Perhaps negatively correlated as there appears to be more located in the lower right and upper left.
Year	Non-Linear	The year appears to rise and fall in an inverted U shape, suggesting possible non-linear relationship. This has been removed in the updated model
Garage		Perhaps positively correlated as there appears to be heavier population in an positive slope moreso than if a negative slope was on it.

The remaining variables I had plotted but they appeared odd and hard to interpret.



From the Added variable plot, it would appear that they all have a positive slope except for beds, when looking at the reduced model ('Year' removed).





From the Component-plus-residual plot graphs , Size (+), Bed (-) have a linear relationship but Garage appears to be non-linear as the purple/pink line deviates from the blue linear line. It can create random outcomes in the model that cannot be explained. To correct this we can use transformations , or polynomials.

Q2. Check the random/i.i.d. sample assumption by carefully reading the data description and computing the Durbin Watson test (state the hypothesis of the test, the test statistic and p-value and the conclusion in the context of the problem). What are the two common violations of the random/i.i.d. sample assumption? What effect would dependant samples have on the

REGRESSION MODEL AND HOW MIGHT YOU CORRECT OR IMPROVE THE MODEL IN THE PRESENCE OF DEPENDANT SAMPLES?

```
lag Autocorrelation D-W Statistic p-value
   1       0.1760459      1.612759   0.038
 Alternative hypothesis: rho != 0
```

Hypothesis H_0: There is no correlation among residuals (independent)

vs.

H_a: The residuals are autocorrelated.

Since the test showed a Statistic of 1.6 we can say that it is indicative of positive autocorrelation, but the p-value is < 0.05 so the null hypothesis is rejected, therefore the observations cannot be classed as independent.

The 2 most common violations are repeated measures (the same object is measured on different dates, so medication dose over a month period), and multiple measures (analysing a test when a student performs well on one section, and expect them to perform well on another).

The dependencies would create bias and make the model inefficient as it would become more difficult to interpret. To improve the model, we can use mixed effect models if the data is normally distributed, or log transformations if it isn't, but that data should have repeated measures. If there is dependence then we could look at time series analysis to improve the model.

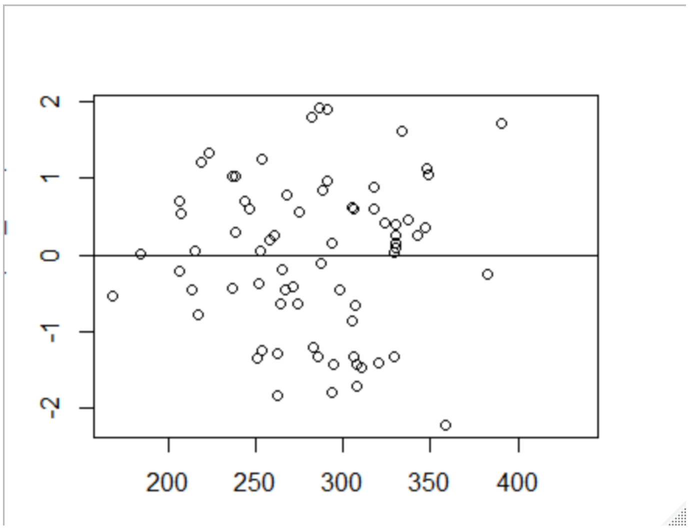
Q3. CHECK THE COLLINEARITY ASSUMPTION BY INTERPRETING THE CORRELATION AND VARIANCE INFLATION FACTORS. WHAT EFFECT WOULD MULTICOLLINEARITY HAVE ON THE REGRESSION MODEL AND HOW MIGHT YOU CORRECT OR IMPROVE THE MODEL IN THE PRESENCE OF MULTICOLLINEARITY.


```
            GVIF Df GVIF^(1/(2*Df))
Size     1.707886  1        1.306861
Lot     25.024713  8        1.222920
Bath    14.697769  5        1.308354
Bed      1.874190  1        1.369011
Garage   1.599886  1        1.264866
School   6.824662  5        1.211736
```

The Variance Inflation Factor using $GVIF^{(\frac{1}{2}*df)}$ does not appear to show multicollinearity as no value is too high (~4 or higher, using rule of thumb). Since we have many degrees of freedom (df) in some variables the Generalized Variance Inflation Factor formula: $GVIF^{(\frac{1}{2}*df)}$ is used instead. A strong correlation between 2 variables X_j and X_k , then $\hat{\beta}$ becomes unstable. The estimate of β_j will then depend heavily on the other variables in the model. Therefore if the variables are correlated we can't interpret the coefficients like before. To improve the model with multicollinearity we could try removing the predictors that are highly correlated, or use Partial Least Squares Regression (PLS), Principal

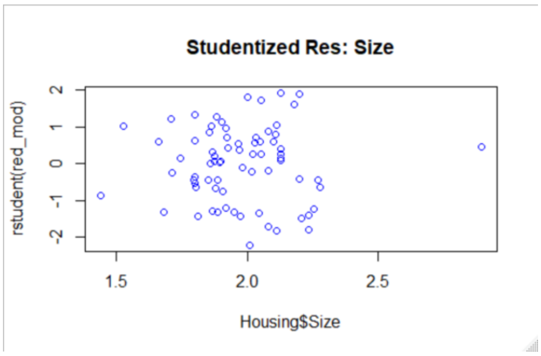
Component Analysis (PCA), or Ridge Regression. Any of these methods achieve improvement by using a subset of variables.

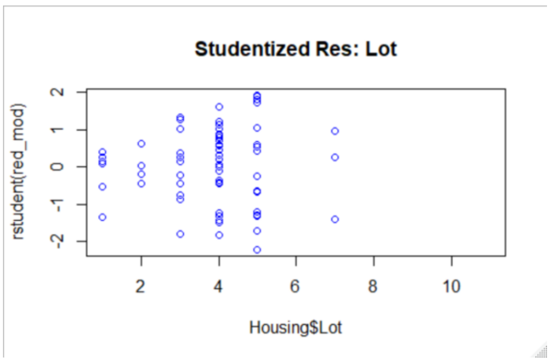
Q4. CHECK THE ZERO CONDITIONAL MEAN AND HOMOSCEDASTICITY ASSUMPTION BY INTERPRETING THE STUDENTIZED RESIDUALS VRS FITTED VALUES PLOTS AND THE STUDENTIZED RESIDUALS VRS PREDICTOR VARIABLE PLOTS. WHAT EFFECT WOULD HETEROSCEDASTICITY HAVE ON THE REGRESSION MODEL AND HOW MIGHT YOU CORRECT OR IMPROVE THE MODEL IN THE PRESENCE OF HETEROSCEDASTICITY.

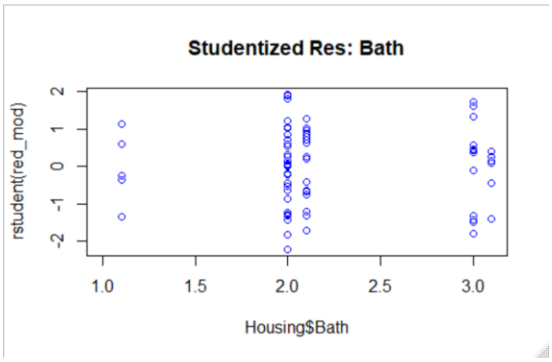


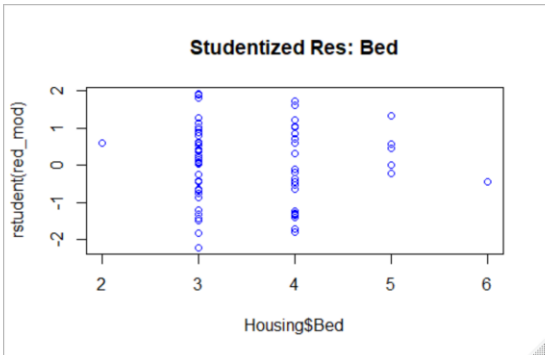
FITTED VS STUDENTIZED RESIDUALS

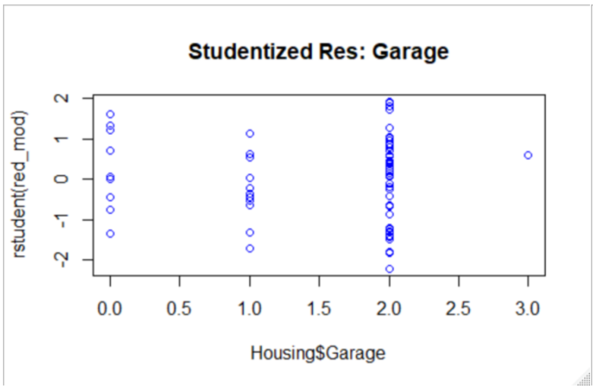
From the plot on the left we can see that there is homoskedasticity as there is a symmetric spread from the 0 line in the positive and negative direction. There is a little less spread on the right which could indicate outliers.

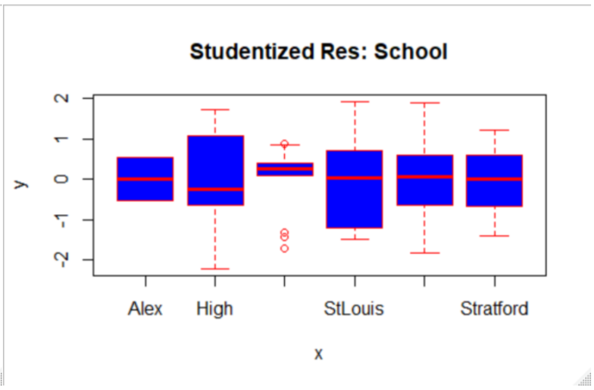








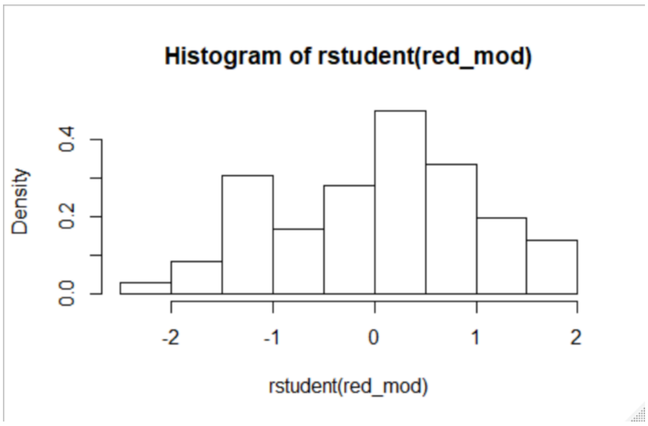




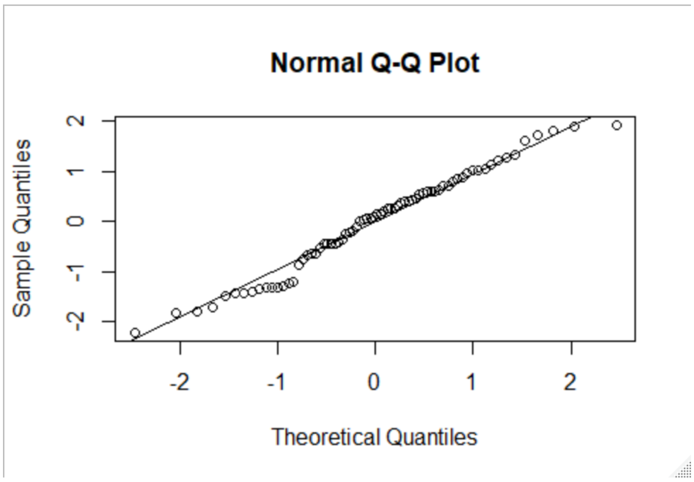
Some of these are categorical and cannot be interpreted correctly, only numeric values should be considered

Heteroskedasticity would cause the standard errors to become biased, which could affect the hypothesis tests. To correct this we can use Weighted Least Squares which adds a weight that are inversely proportionate to the variability of the data, i.e. sparse data is weighted differently to heavily concentrated data. This helps the model become more homoscedastic, but may still have outliers.

Q5. Check the Normality assumption by interpreting the histogram and quantile-quantile plot of the studentized residuals. What effect would non-normality have on the regression model and how might you correct or improve the model in the presence of non-normality.



The histogram shows relatively normal distribution with a slight increase in -1 to -1.5. This could suggest outliers



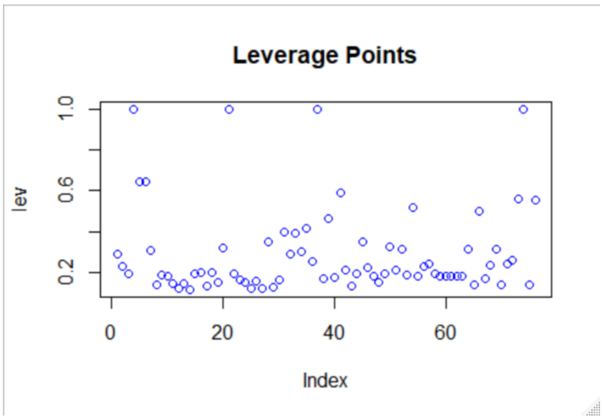
The QQ-Plot shows a very close relationship to the line, with a little deviation around -1, this is expected based from the histogram. However, the points are all close enough to assume normality.

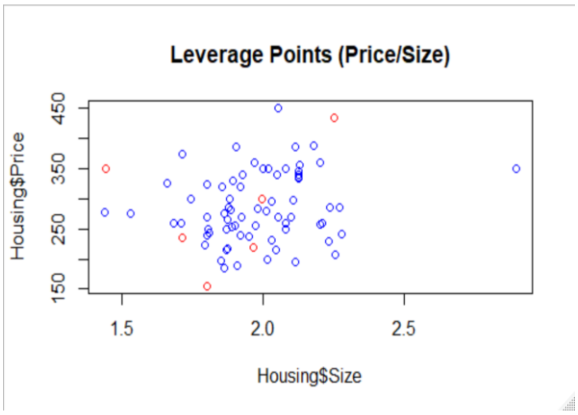
If non-normality was present in our model then it would affect the critical values of the t-test and F-test. To correct this we can perform transformations on response/predictor variables. We can create interaction models, which makes the model more complex. Alternatively, we can use a different model as the current one is not flexible enough for our data.

Leverage, Influence and Outliers

Q1 What is a leverage point? What effect would a leverage point have on the regression model? Use the leverage values and the leverage plots to see if there is any leverage points.

A Leverage point is a data point with an unusual X-value, which can affect the statistics of the model summary (R^2,SSE, etc) but it will have minimum impact on the estimates of regression coefficients. If they are high, they have the potential to affect the fit of the model.

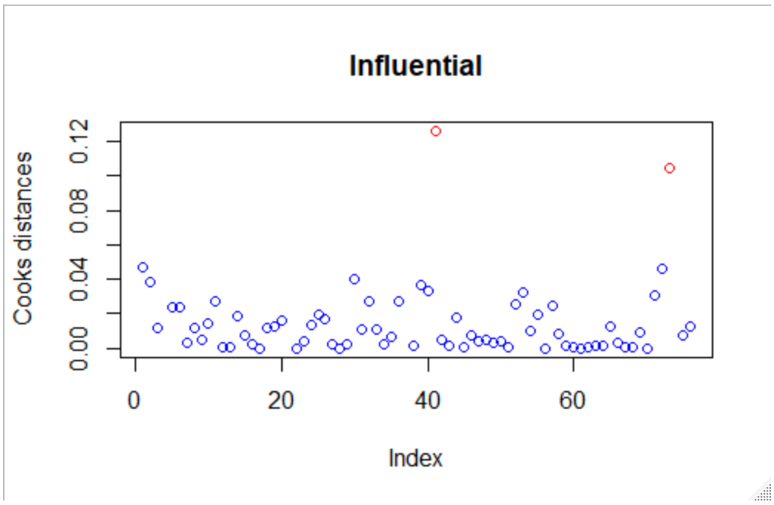





```
   Price  Size
4   350.0 1.442
5   155.5 1.800
6   220.0 1.965
21  299.0 1.994
```

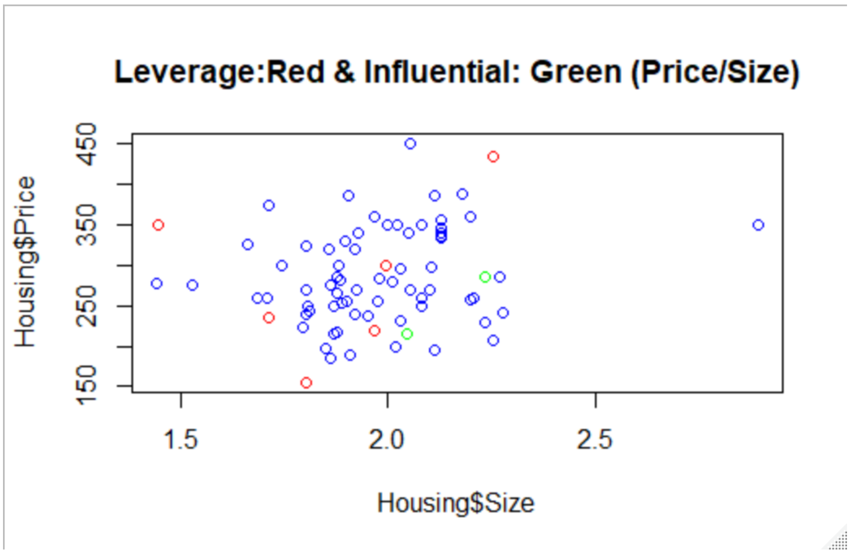
```
37 235.0 1.712
74 435.0 2.253
```

Q2. What is an influential point? What effect would an influential point have on the regression model? Use the influence plot to see if there is any influence points



An Influential Point has an unusual ***Y*** value. It will move the regression model in the direction of its point.

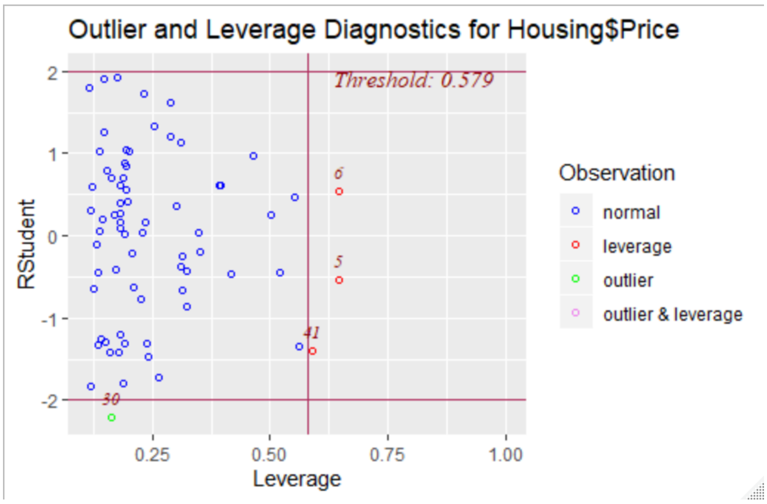
41 and 73 are influential points above 0.08



Q3. What is an outlier? What effect would an outlier have on the regression model? How would you correct for outliers? Use the outlier test and outlier and leverage diagnostics plot to see if there is any outliers. Deal with the outliers if any are identified.

An outlier is an observation (i) within the data where the response does not relate to the what the model fitted for the majority of the data. The impact of an outlier is that it could affect the estimation of the regression coefficients. If it is not a high leverage/influence point we can remove the outlier from the data.

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferroni p
30 -2.213021           0.031225           NA
```



Leverage:

5,6,41

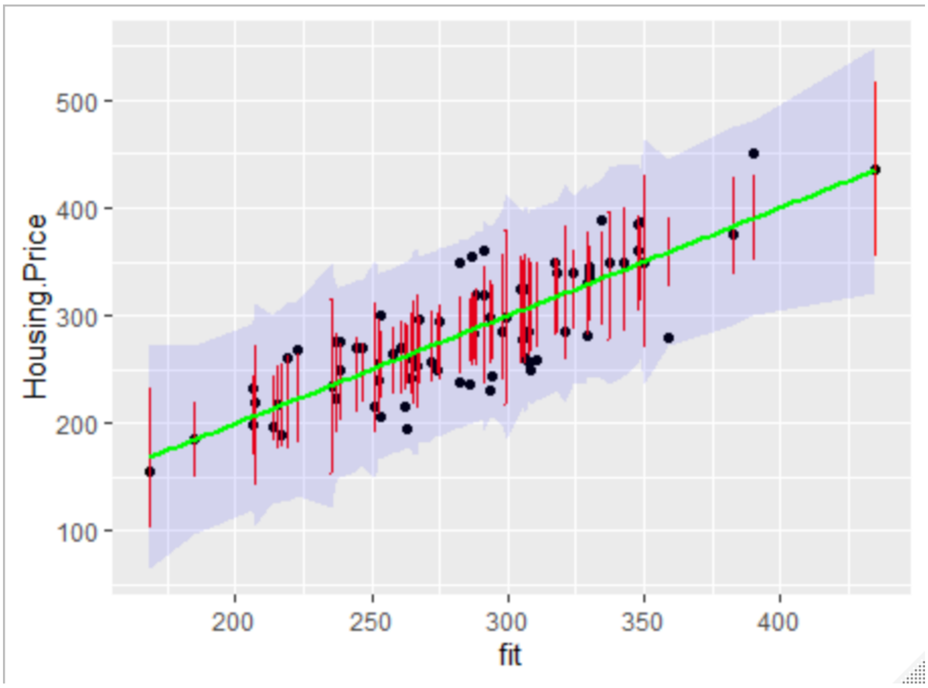
Outlier:

30

Since no studentized residuals had a p-value < 0.05. The graphic does highlight entry 30 as an outlier. The other possibility is to compare the model with and without the model and see which model is better.

Expected Value, CI and PI

Q1. Plot the observed house prices, their expected vale (fitted value), confidence intervals (in red) and prediction intervals (in blue). Looking at this plot is this model providing a good estimate of the house prices.



The model is providing a good measure of predictions, the data falls within the prediction measures and most of the points are within the confidence intervals. The graph shows homoskedasticity

R CODE

```
# Patrick Lowe - 16725829

# open the housing CSV
Housing = read.csv(file.choose())
par(mar=c(3,3,3,3)) # this works best for my machine

######
# Q1 #
######

# Create a boxplot for housing prices
boxplot(Housing$Price)
# Create a Histogram of prices
hist(Housing$Price, freq=FALSE)
lines(density(Housing$Price),lwd=2, col="blue")
# create a summary
```

```
summary(Housing)


######

# Q2 #

######


# Convert Categorical to factors

Housing$Lot.Type.f <- factor(Housing$Lot)

Housing$Bath.Type.f <- factor(Housing$Bath)

Housing$Bed.Type.f <- factor(Housing$Bed)

#Housing$Garage.Type.f <- factor(Housing$Garage) # chosen not to treat as categorical, very well
could be

Housing$School.Type.f <- factor(Housing$School)



#SUMMARY AND A BOXPLOT

#PRICES: BEDROOMS,BATHROOMS,GARAGE,SCHOOL

boxplot(Housing$Price~Housing$Bed,data=Housing,main="Price by
#Bedrooms",col="blue",border="red")

tapply(Housing$Price, Housing$Bed, summary)

boxplot(Housing$Price~Housing$Bath,data=Housing,main="Price by
#Bathrooms",col="blue",border="red")

tapply(Housing$Price, Housing$Bath, summary)

boxplot(Housing$Price~Housing$Garage,data=Housing,main="Price by Garage
Space",col="blue",border="red")

tapply(Housing$Price, Housing$Garage, summary)

boxplot(Housing$Price~Housing$School,data=Housing,main="Price by
Schools",col="blue",border="red")

tapply(Housing$Price, Housing$School, summary)


######

# Q3 #

######
```

```
# Summary, Correlation, Pairs Plots:
#price and each of the numeric predictorvariables.

# Summary
tapply(Housing$Price, Housing$Bed, summary)

# Correlation
# step by step: cor(Housing$Price,Housing$Bed)
# Or graphically:
library(corrplot)
M <- cor(Housing[,1:7])
corrplot.mixed(M)
# Pairs Plots
pairs(Housing[,1:7])

######
# Q4 #
######
# MLR Model
mod = lm(Housing$Price ~ Housing$Size
    + Housing$Lot.Type.f
    + Housing$Bath.Type.f
    + Housing$Bed.Type.f
    + Housing$Year
    + Housing$Garage
    + Housing$School.Type.f,
    data=Housing)
summary(mod)

# Plot Residuals
resid(mod) #List of residuals
```

```
plot(density(resid(mod)))
qqnorm(resid(mod))
qqline(resid(mod))
library(ggplot2)
residualPlot <- ggplot(aes(x=.fitted, y=.resid),data=mod)+geom_point()+geom_hline(yintercept=0)
residualPlot

#########
# Anova #
#########

# COMPUTE THE TYPE 1 ANOVA TABLE. INTERPRET THE OUTPUT
anova(mod)
SSR <- sum(anova(mod)[1:7,2])
MSR <- SSR/7
SSE <- anova(mod)[8,2]
MSE <- anova(mod)[8,3]
test_stat <- MSR/MSE
qf = qf(0.05,22,53)

#Q3. Compute a type 2 anova table comparing the full model with all predictor
#variables to the the reduced model with the suggested predictor variable
#identified in the previous question removed.
# Reduced Model
red_mod = lm(Housing$Price ~ Housing$Size
    + Housing$Lot.Type.f
    + Housing$Bath.Type.f
    + Housing$Bed
    + Housing$Garage
    + Housing$School.Type.f,
    data=Housing)
```

```
Anova(red_mod)
SSR2 <- sum(anova(red_mod)[1:6,2])
MSR2 <- SSR/6
SSE2 <- anova(mod)[7,2]
MSE2 <- anova(mod)[7,3]
test_stat2 <- MSR2/MSE2
qf2 = qf(0.05,21,54)

anova(red_mod, mod)

###############
# Diagnostics #
###############
library("GGally")
library(car)
ggpairs(Housing[,1:7])
ggpairs(Housing[,c(1:1, (ncol(Housing) - 4):ncol(Housing))])
avPlots(red_mod)
crPlots(red_mod)

# random/i.i.d sample
dwt(red_mod)

# multicollinearity
gvif(red_mod)
M <- cor(Housing)
corrplot.mixed(M)

# Zero Conditional Mean & Homoskedasticity
plot(fitted(red_mod),rstudent(red_mod))
abline(h=0)
```

```
plot(Housing$Size,rstudent(red_mod),main="Studentized Res: Size",col="blue")

plot(Housing$Lot,rstudent(red_mod),main="Studentized Res: Lot",col="blue")

plot(Housing$Bath,rstudent(red_mod),main="Studentized Res: Bath",col="blue")

plot(Housing$Bed,rstudent(red_mod),main="Studentized Res: Bed",col="blue")

plot(Housing$Garage,rstudent(red_mod),main="Studentized Res: Garage",col="blue")

plot(Housing$School,rstudent(red_mod),main="Studentized Res: School",col="blue",border="red")


# normality

hist(rstudent(red_mod),freq=FALSE)

lines(density(rstudent(red_mod)), lwd=2, col="blue")

qqnorm(rstudent(red_mod))

qqline(rstudent(red_mod))


# Leverage points

library(olsrr)

lev = hat(model.matrix(red_mod))

plot(lev,main="Leverage Points",col="blue")

# highlighting high lieverage points

plot(Housing$Size,Housing$Price,main="Leverage Points (Price/Size)",col="blue")

points(Housing[4,]$Size,Housing[4,]$Price,col="red")

points(Housing[5,]$Size,Housing[5,]$Price,col="red")

points(Housing[6,]$Size,Housing[6,]$Price,col="red")

points(Housing[21,]$Size,Housing[21,]$Price,col="red")

points(Housing[37,]$Size,Housing[37,]$Price,col="red")

points(Housing[74,]$Size,Housing[74,]$Price,col="red")


# Influential Points

cook = cooks.distance(red_mod)

plot(cook,ylab="Cooks distances")

which(cook>0.08)

plot(cook,ylab="Cooks distances",main="Influential",col="blue")
```

```
points(41,cook[41],col="red")
points(73,cook[73],col="red")

# plot both points
plot(Housing$Size,Housing$Price,main="Leverage:Red & Influential: Green (Price/Size)",col="blue")
points(Housing[4,]$Size,Housing[4,]$Price,col="red")
points(Housing[5,]$Size,Housing[5,]$Price,col="red")
points(Housing[6,]$Size,Housing[6,]$Price,col="red")
points(Housing[21,]$Size,Housing[21,]$Price,col="red")
points(Housing[37,]$Size,Housing[37,]$Price,col="red")
points(Housing[74,]$Size,Housing[74,]$Price,col="red")
points(Housing[41,]$Size,Housing[41,]$Price,col="green")
points(Housing[73,]$Size,Housing[73,]$Price,col="green")

# Outliers
outlierTest(red_mod)
CD=cooks.distance(red_mod)
CD[as.numeric(which(CD> 0.05))]

#PLOT PRICES, FITTED VALUE
new.prices <- data.frame(Housing$Price)
CI <- predict(red_mod, newdata = new.prices, interval = "confidence")
PI <- predict(red_mod, newdata = new.prices, interval = "predict")

plotdata <- data.frame(Housing$Price,CI[,1:3],PI[,2:3])
names(plotdata)[names(plotdata) == "lwr"] <- "CI lwr"
names(plotdata)[names(plotdata) == "upr"] <- "CI upr"
names(plotdata)[names(plotdata) == "lwr.1"] <- "PI lwr"
names(plotdata)[names(plotdata) == "upr.1"] <- "PI upr"

plotdata[,1]=round(plotdata[,1],2)
```

```
plotdata[,2]=round(plotdata[,2],2)
plotdata[,3]=round(plotdata[,3],2)
plotdata[,4]=round(plotdata[,4],2)
plotdata[,5]=round(plotdata[,5],2)
plotdata[,6]=round(plotdata[,6],2)

library(ggplot2)
U <- plotdata$`CI upr`
L <- plotdata$`CI lwr`
U2 <- plotdata$`PI upr`
L2 <- plotdata$`PI lwr`

ggplot(plotdata, aes(x = fit, y = Housing.Price)) +
  geom_point() +
  geom_errorbar(aes(ymax = U, ymin = L),colour = "red") +
  geom_ribbon(aes(ymin = L2, ymax = U2), fill = "blue", alpha = 0.1) +
  geom_smooth(method = "lm", se = FALSE, colour = "green")
```