**Exploratory Data Analysis:**

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

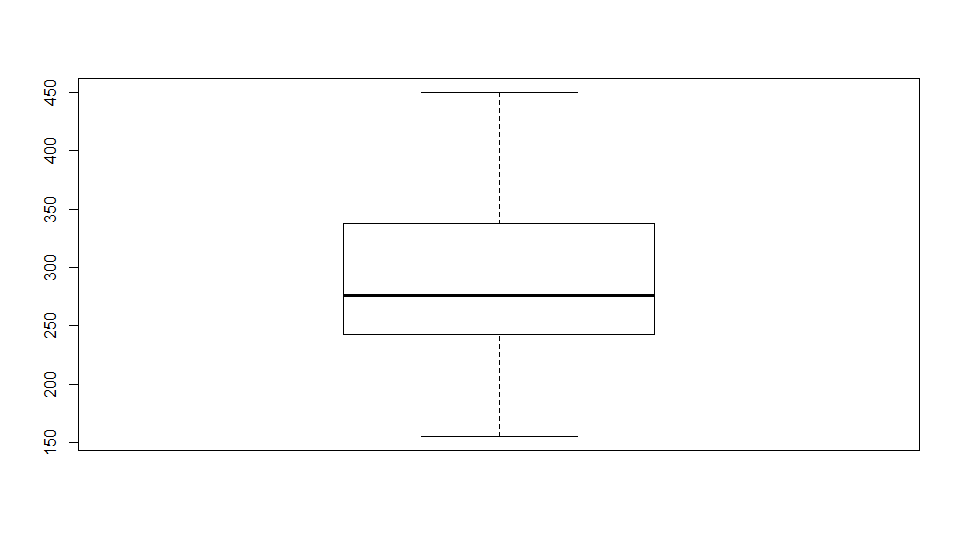
Answer:

# Boxplot

From the boxplot we can see that the median of house prices lies around 275.

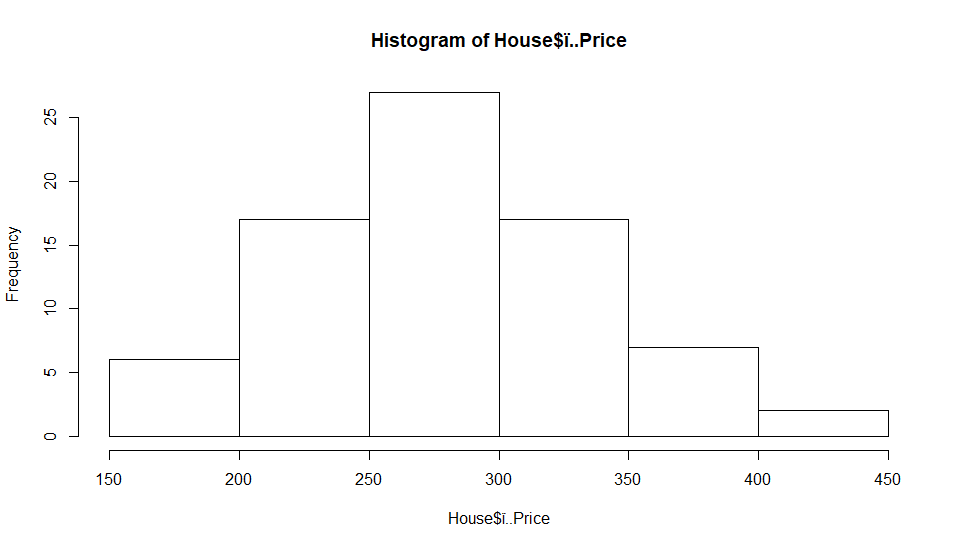
Minimum house price and maximum house price range from 150 to 450.

First Quantile lies at 250 and Third Quantile lies at 350



# Histogram

The Distribution seems to be positively skewed. We can see that tail tapers on the right hand side of x-axis.



# Summary

* Mean of the house price distribution is around 285.8
* Median of the house price distribution is around 276.0
* Mean and Median are bit apart indicating potential outlier presence.
* First Quantile which forms 25% of the house prices lies under 242.8
* Median which is at 276.0 means 50% of the houses are under the price 276.0
* Third Quantile which forms 75% of the house price data are under 336.8

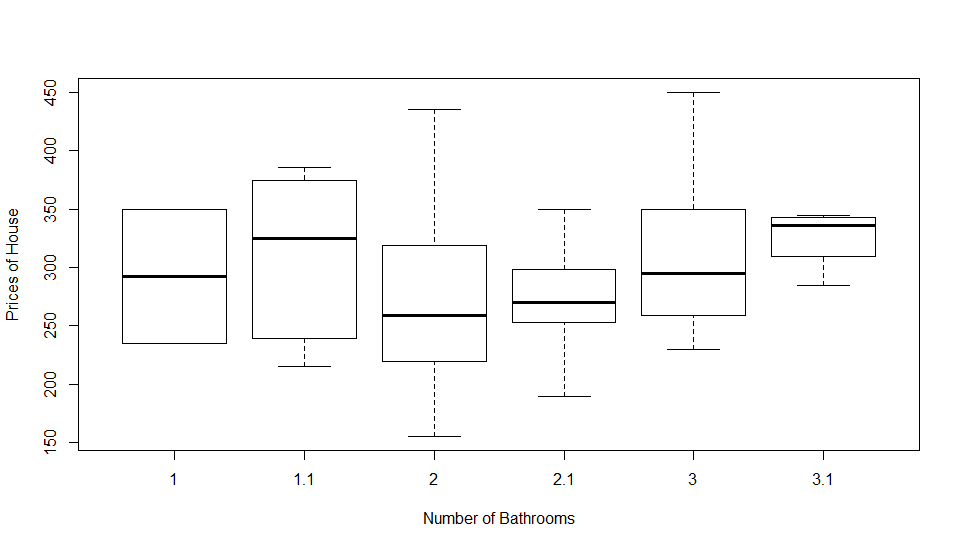
1. Convert all the categorical variables to factors. Using the summary and

a boxplot describes how sales prices vary with respect to the number of

bedrooms, bathrooms, garage size and school.

Answer:

**Bathrooms**



Boxplot comments:

Here we can observe that the median for the price of house is increasing for 1 and 1.1 bathrooms, but there is sudden decrease for 2 bathrooms and again a gradual increase is seen.

Summary:

# House\_1$Bath: 1

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 235.0 263.8 292.5 292.5 321.2 350.0

# ------------------------------------------------------------------------------------------

# House\_1$Bath: 1.1

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 215.0 239.5 325.0 307.9 374.5 385.5

# ------------------------------------------------------------------------------------------

# House\_1$Bath: 2

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 155.5 219.2 259.4 271.3 321.8 435.0

# ------------------------------------------------------------------------------------------

# House\_1$Bath: 2.1

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 189.5 254.8 269.9 274.5 297.7 349.5

# ------------------------------------------------------------------------------------------

# House\_1$Bath: 3

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 230.0 259.0 295.0 307.8 349.5 450.0

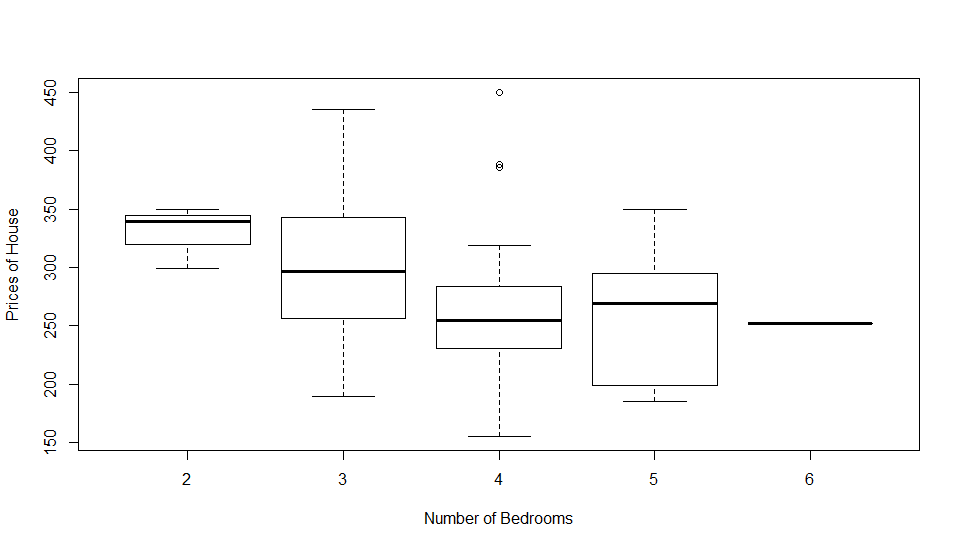
# ------------------------------------------------------------------------------------------

# House\_1$Bath: 3.1

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 285.0 309.4 336.0 324.2 342.5 345.0

Bedrooms:



Boxplot Comments:

* Here we can see that there are few outliers for Prices of the house in the case where the number of bedrooms are 4.
* There is just one sample for number of bedrooms is 6 in the list of prices of houses.
* Median house price is highest for 2 bedroom house and is decreasing till 4 Bedrooms and increase is seen for 5 bedrooms.

Summary:

# House$Bed: 2

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 299.0 319.4 339.9 329.6 344.9 350.0

# ------------------------------------------------------------------------------------------

# House$Bed: 3

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 189.5 256.2 297.0 297.3 342.5 435.0

# ------------------------------------------------------------------------------------------

# House$Bed: 4

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 155.5 231.5 254.4 266.6 283.5 450.0

# ------------------------------------------------------------------------------------------

# House$Bed: 5

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 185.0 199.0 269.0 259.5 295.0 349.5

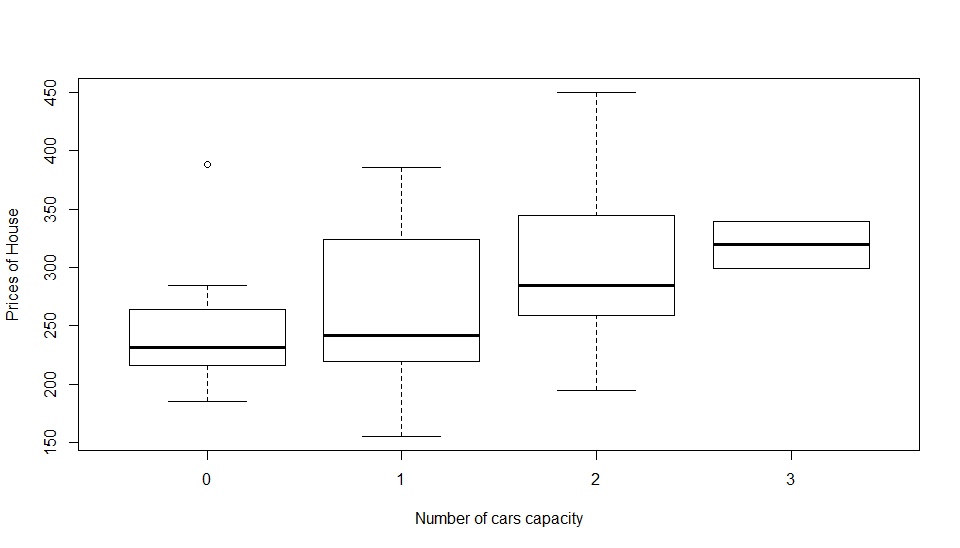
# ------------------------------------------------------------------------------------------

# House$Bed: 6

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 252.5 252.5 252.5 252.5 252.5 252.5

Garage:



Boxplot comments:

* From the boxplot we can see that there is one outlier for the sample set of houses where the garage value is 0 and it is priced way too much.
* This data is arbitrary and can be excluded.
* Median price of houses is increasing as per the capacity of Garage.

Summary:

# House$Garage: 0

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 185.0 216.0 232.0 246.9 264.4 388.0

# ------------------------------------------------------------------------------------------

# House$Garage: 1

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 155.5 220.0 242.0 260.6 324.5 385.5

# ------------------------------------------------------------------------------------------

# House$Garage: 2

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 195.0 259.0 285.0 299.6 343.8 450.0

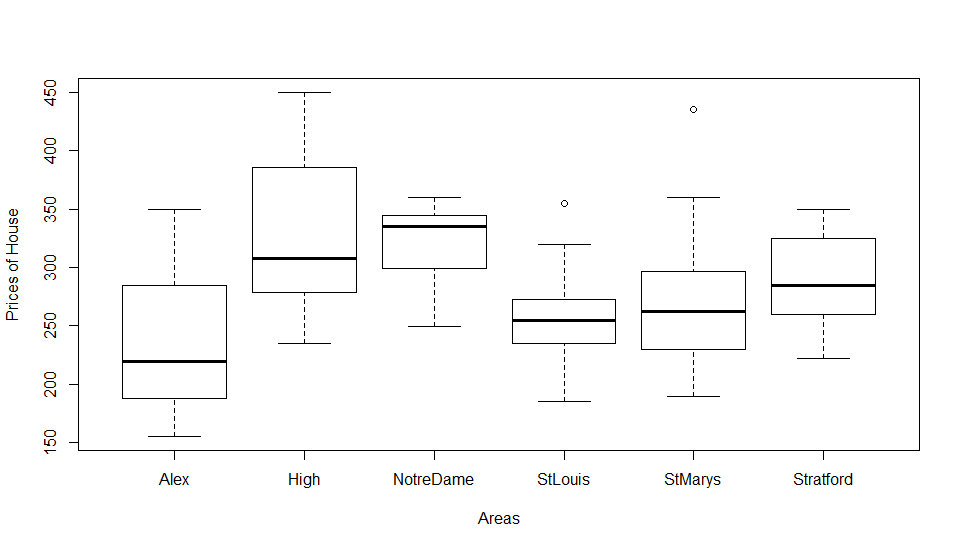
# ------------------------------------------------------------------------------------------

# House$Garage: 3

# Min. 1st Qu. Median Mean 3rd Qu. Max.

# 299.0 309.2 319.4 319.4 329.7 339.9

School:

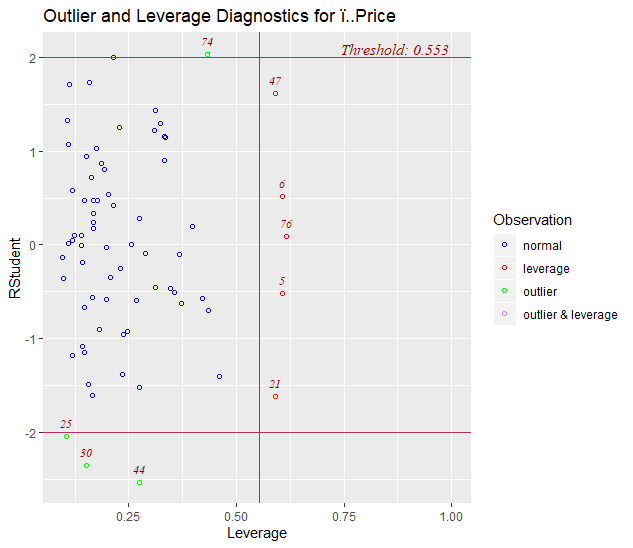


Boxplot comments:

* There is one outlier each in the case of house prices in the Area StLouis and StMarys.
* House near Alex has lowest price median
* House near NotreDam has highest price median

Summary:

|  |
| --- |
| House$School: Alex  Min. 1st Qu. Median Mean 3rd Qu. Max.  155.5 187.8 220.0 241.8 285.0 350.0  ---------------------------------------------------  House$School: High  Min. 1st Qu. Median Mean 3rd Qu. Max.  235.0 279.2 307.5 327.1 385.6 450.0  ---------------------------------------------------  House$School: NotreDame  Min. 1st Qu. Median Mean 3rd Qu. Max.  249.9 304.0 334.9 319.1 345.0 359.9  ---------------------------------------------------  House$School: StLouis  Min. 1st Qu. Median Mean 3rd Qu. Max.  185.0 235.4 255.0 257.4 272.4 355.0  ---------------------------------------------------  House$School: StMarys  Min. 1st Qu. Median Mean 3rd Qu. Max.  189.5 231.6 262.0 269.8 296.5 435.0  ---------------------------------------------------  House$School: Stratford  Min. 1st Qu. Median Mean 3rd Qu. Max.  222.5 266.2 285.0 287.8 315.0 349.5 |
|  |
| |  | | --- | |  |   3. Using the summary, correlation and the pairs plots discuss the relation-  ship between the response sales price and each of the numeric predictor  variables.  Answer:  Pairs between Price and year    Here we can observe that House prices for the house size 2 is well dispersed ranging from 200 to 450 and the house size 1.5 pricing from 150 to 450.  Pairs between Price and Size     * Here we can see that similar pattern of increase in density of the plot in the range of price 200 to 400 over the years 1960 to 2000. * Most number of mid-range priced house where sold most from 1960 to 2000.   **Pairs(House$Price~House$Size+House$Year)**    Most number of mid-range priced house where sold most from 1960 to 2000 and where possibly sized around 2.0.  **Correlation**  The Correlation between the response sales price and each of the numeric predictor variables is weak and less.  **Regression Model**  Question1: 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.  Answer:  **Y= β0+ β1Lot+ β2Size+ β3Year+ β4Bath1.1+ β5Bath2+ β6Bath2.1+ β7Bath3+ β8Bath3.1+ β9Bed3+ β10Bed4+ β11Bed5+ β12Bed6+ β12Garage1+ β14Garage2+ β15Garage3+β16SchoolHigh+β17SchoolNotreDame+β18SchoolStLouis+β19SchoolStMarys+ β20SchoolStratford**  Question 2: Interpret the estimate of the intercept term Beta\_0.  Answer: The estimate of the intercept term **Beta0 is 376.1015959**  Question 3: Interpret the estimate of BetaSize, the parameter associated with floor size (Size).  Answer: Estimate of **BetaSize is 59.4503306**  Question 4: Interpret the estimate of Beta Bath1.1 the parameter associated with one and  a half bathroom.  Answer: Estimate of **Beta\_Bath1.1 is 135.89830**  Question 5: Discuss and interpret the effect the predictor variable bed on the expected  value of the house prices.  Answer:   * The predictor variable beds shows that median value of house prices for 2 bedrooms is nearing 350, median house price value for 3-bedroom houses is decreasing when compared to 2-bedroom houses. * Median house price value for 4-bedroom houses again decreases when compared to 3-bedroom houses. * Median house price value for 5-bedroom houses increases a little bit compared to median value of 4-bedroom houses. * House with 6 bedrooms is just one sample and this value hardly contributes anything.   Question 6: List the predictor variables that are significantly contributing to the  expected value of the house prices.  Answer:  1) Lot predictor variable is significant (\*\*)  2) Bath1.1 is the significant (\*\*) predictor variable is significant among beds.  3) Bed3, Bed4, Bed5, Bed6 is a significant (\*\*) predictor variable  4) SchoolHigh (\*\*) is a significant predictor variable  5) SchoolNotreDame (\*) is a significant predictor variable  6) Garage3(\*) is a significant predictor variable  7) Size (\*) is a significant predictor variable  8) Bath3(.) has least significance among the significant variables  Question 7: For each predictor variable what is the value that will lead to the largest  expected value of the house prices.  Answer:  Generic: **For Numeric predictors its Beta0+ Beta of respective predictive variables multiplied by the Largest value of that predictive variable.**  Largest value House price for predictor variable Lot: 458.6472496  Largest value House price for predictor variable Size: 431.1291106  Largest value House price for predictor variable Year: 395.915725  Largest value House price for predictor variable Bath1.1: 511.9999  Largest value House price for predictor variable Bed3: 308.214  Largest value House price for predictor variable Garage2: 409.5886  Largest value House price for predictor variable SchoolHigh: 489.379  Question 8: For each predictor variable what is the value that will lead to the lowest  expected value of the house prices.  Generic: **For Numeric predictors its Beta0+ Beta of respective predictive variables multiplied by the Lowest value of that predictive variable.**  Lowest value House price for predictor variable Lot: 340.9461  Lowest value House price for predictor variable Size: 344.688656  Lowest value House price for predictor variable Year: 342.1429  Lowest value House price for predictor variable Bath1.1: 240.2033  Lowest value House price for predictor variable Bed6: 121.0805  Lowest value House price for predictor variable Garage2: 166.1978  Lowest value House price for predictor variable SchoolStLouis:367.0649  Question 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.    Answer: Having a look at the residuals and plot, the model seems to be good model.  Because of the outliers the model seems to be tapering off on the right side of x-axis.  From the residual summary we can state that   * Minimum residual is at -87.601 * 25% of the residuals are below -21.429 * 50% of the residuals are below 0.173 * 75% of the residuals are below 24.248 * Maximum residual is valued at 72.581   Question 10: Interpret the Adjusted R-squared value.  Answer:  The Adjusted R-square value is 50% gives value of the influence of all the predictive variables involved in the model.  Question 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.  Answer: F test gives the measure of significance for the overall model.  Hypothesis being tested are:   * Null Hypothesis where all Beta's values are equal to zero and Alternate hypothesis where at least one of the Beta value is non zero. * F test statistics is 4.942 on 20 Beta values and for 55 degrees of freedom. * P-value is telling that probability of all the Beta being equal to zero, and 1.265e-06 is less than 0.05 hence we reject the Null hypothesis.   **Anova:**  Question 1: 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: ï..Price  # Df Sum Sq Mean Sq F value Pr(>F)  # Lot 1 16284 16284.4 9.1767 0.003729 \*\*  # Size 1 10026 10025.7 5.6498 0.020964 \*  # Year 1 4741 4740.6 2.6715 0.107872  # Bath 5 37939 7587.9 4.2760 0.002345 \*\*  # Bed 4 20200 5049.9 2.8458 0.032393 \*  # Garage 3 16101 5367.1 3.0245 0.037179 \*  # School 5 70112 14022.4 7.9020 1.153e-05 \*\*\*  # Residuals 55 97599 1774.5  # ---  # Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Answer:  From the anova table it is clear that, there is at least one predictor variable which is significant, hence rejecting the null hypothesis. (The p value of f statistics is less than 0.05 hence rejecting the null hypothesis)  Question 2: Which predictor variable does the type 1 anova table suggest you should  remove the regression analysis.  Answer:  Year is the predictor variable that is least significant as the p value is 0.10( that is 10% chances of value approximating to zero)  Question 3: Compute a type 2 anova table comparing the full model with all predictor  variables to 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.  Answer:  RSS is 97599 and 102402.  F value: 2.7064  P-value: 0.1057  Conclusion:  The residual does not drop significantly after removing the Year as the predictor variable. Hence it can be considered as the non-significant parameter.  **Diagnostics**  Question 1: 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? |
|  |
| Linearity assumptions interpretation  For Added variable plots for each predictive variable for Response variable:House Price   * For Lot variable the best suited values are from -2 to 1 and the slope of the model is positive (Beta value is 11.7701) * For Size, model is best suited for values from -0.2 to 0.2 and the Slope of the model is positive, the outlier point 76 is influencing the slop (Beta value is 59.4503) * For Year, model is best suited for the values from -20 to 20 and the Slope of the model is slightly positive. (Beta value is 0.5567) * For Bath1.1, model is best suited for the values around 0.0 and the Slope of the model is Highly positive.(Beta value is 135.8983) * Amongst the outliers for Bath1.1 point 37 is influencing the Slope inclination most. * For Bath2, model is best suited for the values around 0.0 and again the Slope is positive and is most effected by the outlier point 37.(Beta value = 73.9317) * For Bath2.1, model is best suited for the values around 0.0 and again the Slope is positive and is mostly effected by point 37.(Beta value = 76.9433) * For Bath3, model is best suited for the values around 0.0 and again the Slope is positive and is mostly effected by the outlier point 37.(Beta value = 98.0694) * For Bath3.1, model is best suited for the values around 0.0 and again the Slope is positive and is mostly effected by the outlier point 37.(Beta value = 85.8037) * For Bed3, model is best suited for values around 0.0 and the slope is negative (-228.1052) and is mostly effected by 4 and 37 numbered outliers. * For Bed4, model is best suited for values around 0.0 and the slope is negative (-238.2609) and is mostly effected by 4 and 37 numbered outliers. * For Bed5, model is best suited for values around 0.0 and the slope is negative (-237.6155) and is mostly effected by 4 and 37 numbered outliers. * For Bed6, model is best suited for values around 0.0 and the slope is negative (-255.0211) and is mostly effected by 4 and 35 numbered outliers. * For Garage1, model is best suited for values ranging from -0.2 to 0.1 and slope is slightly negative (-10.9191) * For Garage2, model is best suited for values ranging from -0.2 to 0.4 and the slope is slightly positive (18.2435) * For Garage3, model is best suited for values around 0.0 and the slope is negative and is highly influenced by outlier points 4 and 37.(-209.9038) * For SchoolHigh, model is best suited for values around 0.0 and the slope is positive(113.2774) and probably is influenced by outlier points 6 and 5. * For SchoolNotreDame, model is best suited for values around 0.0 and the slope is positive(80.9317) and probably is influenced by outliers points 6 and 5. * For SchoolStLouis, model is best suited for values around 0.0 and the slope is positive(9.0367) and probably is influenced by outliers points 6 and 5. * For SchoolStMarys, model is best suited for values around 0.0 and the slope is positive(27.3408) and probably is influenced by outliers points 6 and 5. * For SchoolStratford, model is best suited for values around 0.0 and the slope is positive(31.9254) and probably is influenced by outliers points 6 and 5.     From Coefficient plus residual plot   |  |  | | --- | --- | | Variable | Interpretation | | Lot | Smooth fit line appears to be curved and doesn’t fit exactly the linear best line,  The line fits well from -2 to 1 that it works well for the higher lot sizes | | Size | Smooth fit appears to be a proper fit on the linear curve, the linear relationship are approximately linear | | Year | There is no much difference between the smooth line and the linear best fit line  There are fewer points from -20 to -60 and majority of the points are in the region of -20 to 20 that is the model fits well for the latest year | | Bath | There is a difference in the median values of House Sale Price with different kind of bathrooms | | Bed | There is a difference in the median values of House Sale Price with different kind of Bedrooms | | Garage | Here we can see that there is slight difference in the median values of House Sale Price with 0-2 car capacity and it drops suddenly for Garage with 3 car capacity | | School | For School as well there is a difference in the median values of House Sale Price |  * With the Non-linearity the approximation of the model decreases resulting in the biased and inconsistent estimate * Transformation techniques are used to transform the non-linearity in the data, some of the transformation techniques are taking log, square root of the dependent variables. Also, polynomial and spline techniques are used   Question2: 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).  Answer: H0 = no first order autocorrelation. H1 = first order correlation exists.  The hypothesis of no autocorrelation is rejected and because from the value obtained, we can say that the observations cannot be classified as independent.  Durbin Watson test statistics is 1.614157 but the p-value is 0.048 which is less than 0.05. Means it has positive auto-correlation.  Conclusion:   * The residuals have structural dependencies * The different outliers are causing bias or inefficiency. This can be corrected by treating the outliers. * Because of the Structural dependencies we might not get constant variance. * This can be better dealt with another type of analysis which is Time series Analysis.   What are the two common violations of the random/i.i.d. sample assumption?   * Repeated observations and multiple observations occurrence * No bias with appropriate alternative assumptions (structured dependence) * This can also cause non constant variance. * Outliers from different distributions can cause inefficiency.   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?   * When fitted the regression model into the time series data, the dependencies of the structure, in form of the Autocorrelation can be found in the residuals. * There are different time series versions regression analysis which are suitable for this type of analysis.   Question 3: 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.     * VIF is around 1.5 means that there is no correlation among the jth predictor and remaining predictor variables and hence the Beta value is also not inflated. * Multicollinearity would have no effect if it is just for predicting the response variable, on the regression model, but if for interpretation then Beta values of multicollinearity must be taken care. * To improve the multicollinearity, we can remove the highly correlated predictor from the model. * Choose the highly correlated variables and the one which is to be dropped and perform anova test and compare with all variables in the model and with just one predictor variable removed, the value of F-test p-value gives you confirmation to go ahead dropping the predictor variable. Or use Partial Least Squares Regression(PLS), Principal Components Analysis, Ridge Regression. The is use a method that cuts the number of predictors to a smaller set of uncorrelated components.   Question 4: Check the zero conditional mean and homoscedasticity assumption by interpreting the studentized residuals vs fitted values plots and the studentized residuals vs 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.    In the studentized residual vs fitted value plots,   * We can observe the zero conditionality since all the dots are lined up against the zero and the band which they lie around shows that they have constant variance.   In studentized residuals vs predictor variable   * All the plots show homoscedasticity, there is a constant distribution of the variance across zero. * In the box plot the median is almost same for various categories hence it shows homoscedasticity. * Effect of Heteroscedasticity * Standard errors are biased/distorted * Correct them by using Weighted Least Squares   Question 5: 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.  Answer:  Histogram distribution is bell-shaped, Normal distribution and is peaking around zero and is relatively symmetric.    From Quantile -quantile plot we can see that the errors from our dataset are tracking the red line and we can say   * that the errors are normally distributed. * Effects of Non-normality on regression model, the critical values of F-test and T-test are wrong. * This can be diagnosed by checking studentized residuals and qq-plots.   How to correct or improve   * By doing transformation of response or predictive variables * Creating interaction models or use different model as the current used model is not flexible enough.   **Leverage, Influence and Outliers:**  Question1: 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.     * Leverage point is the one with an unusual X-value. * The measure of leverage is the amount of change in the prediction values for any out of the trend unit in Y direction (influential point). * Leverage always values between 0 to 1. * When a point has leverage 1, the line will follow the point and a point with leverage 0 has no effect on the regression line. * All the predictive variables plots show the existence of leverage points     Question 2: 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 is the one if removed from the data would significantly change the fit. * An influential point may either be an outlier or have large leverage or both, but it will tend to have at least one of those properties * High leverages cases are potentially influential and should be examined for their influence.   Question 3:  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 |



Answer: An outlier is an observation, where the response does not correspond to the model fitted to the bulk of the data.

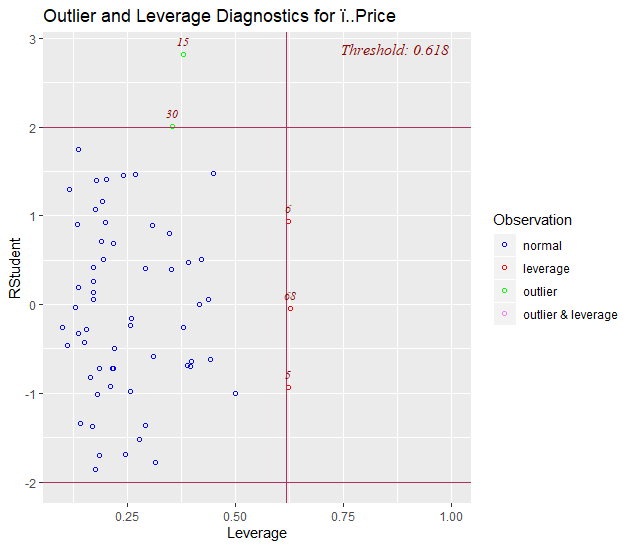
Effect of Outlier:

* Outliers might affect the estimation of the regression coefficient

Methods to deal with the outliers:

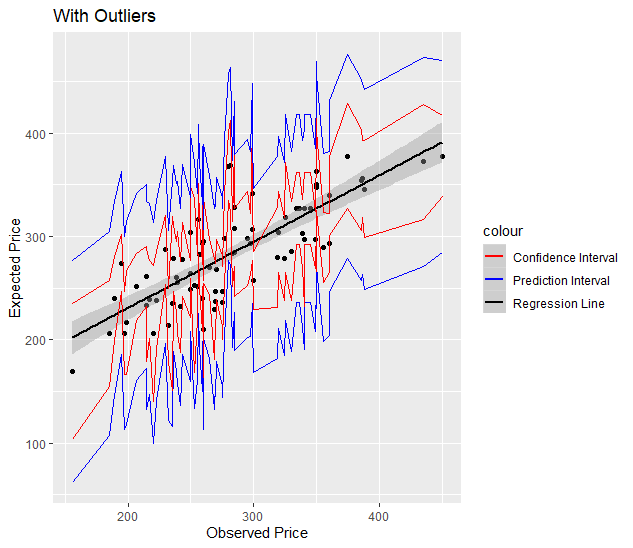
* Exclude the outlier, see its influence. Perhaps present analysis with and without the outlier.

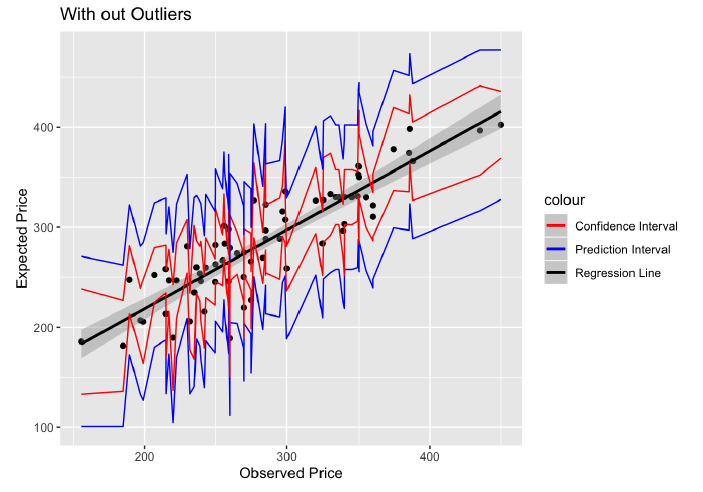
Post the removal of Outliers:



**Expected Value, CI and PI:**

1. 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.





Observing the plots, we can conclude that the model without outliers is providing the better estimates of the House prices compared to the model with outliers