Exploratory Data Analysis:

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

   #Minimum Sale Price of the House is:155

   #Maximum Sale Price of the House is:450

   #Mean is towards the right of median, that represents the distribution is skewed

   #The Distribution is positively skewed from the box plot and histogram

   #25% of the data has the sale price under 242.8

   #50% of the data has the sale price under 276.0

   #75% of the data has the sale price under 336.8

2. **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.**

   Bed:

   #As the number of bedrooms increases the median of the prices decreases
   #House with 2 bedroom has the highest Median sale Price
   #The median of the House Price with 5 beds with respective to the price increases compared to House to 4 bedroom

   Bath:

   #Overall as the number of Bathroom increases from 2, the Price goes on increasing
   #Maximum Sale Price is for the House with 3 bathrooms
   #Minimum Sale price is for the House with 2 Bathrooms
   #The distribution of the Sale Price is negatively skewed for Bath1.1 and Positively skewed for Bath2, Bath3, Bath3.1

   School:

   #House near School NotreDame has the highest median Sale Price
   #House near School Alex has the lowest median Sale Price and Lowest Sale #Price
   #House near High Schools has the Max Sale Price
   increases compared to House to 4 bedrooms

   Garage:

   #Overall as the number of car capacity in the house increases the sale price of the house seems to increasing gradually

#House with 2 car capacity has the maximum sale Price
#House with no car capacity has the lowest median Sale Price
#House with 1 car capacity has the lowest Sale Price

3. **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 pairsPlot and correlation table we can say that the correlation between the price and numerical variables are pretty low.

   Size and Lot has a correlation of 0.2 and 0.24 respectively with the Price and Year has a correlation of 0.15 with the Price

----

Regression Model:

1. **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**

   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

1. Interpret the estimate of the intercept term β0.

   # The Beta0 value is 376.1016 , the chances of Beta0 values approximating to zero is 1.36e-09

   # β0 values gives estimated Sale Price of a house given that Lot, Size and year are zero, considering House with 2 Bedroom, House with 1 bathroom, Garage with 0 car capacity, House near the School Alex

2. **Interpret the estimate of β_{size} the parameter associated with floor size (Size).**

   #βSize=59.4503

   # For every 1 unit increase in the size, the price will increase by 59.4503

   #The chances of Size approximated to 0 is 0.04501(hence it is one of the significant predictor variable that will contribute to the estimation of the predicted value)

3. **Interpret the estimate of $\beta_{Bath1.1}$ the parameter associated with one and a half bathrooms.**

   βBath1.1=135.8983

   #The price of the house will increase from 135.8983when the number of bathroom is increased from 1 bathroom

   #The chances of the price with respective to Bath1.1 approximating to zero is 0.00779 hence it is one of the significant parameter

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

   It can be seen clearly that as the number of bedrooms increases from 2 bedroom to 6bedroom the price goes on decreasing , house with 3 bedroom decreases by 47.4 from 2 bedroom, House with 4 bedroom decreases by 54 from bedroom2, House with 5 bedroom decreases by 48 from house with bedroom 2, and house with 6 bedroom decreases by 118 from House with 2 bedroom,

   Also one factor to notice is House with 6 bedroom has just 1 data so the approximation is inaccurate

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

   The predictor variables that are significantly contributing to the expected value of the house prices are:

   a. Lot
   b. Size
   c. Bath1.1(one and half bathroom)
   d. Bath3 (House with 3 bathrooms)
   e. Bed4
   f. Bed5
   g. Bed6(House with 6 bedrooms)
   h. Garage3
   i. SchoolHigh: Houses with Highschool near by
   j. SchoolNotreDame

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

Predictor Variable	Value/ Level(in case of categorical variable)	Max Price

Lot	7.01316	458.64
Size	0.925605	431.12
Bed	Bed 3	450
Bath	Bath3	450
Garage	Garage2	450
School	SchoolHigh	450
Year	35.59	395.91

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

Predictor Variable	Value/ Level(in case of categorical variable)	Lowest Price
Lot	-2.98684	343.1
Size	-0.530395	344.46
Bed	Bed 4	155.5
Bath	Bath2	155.5
Garage	Garage1	155.5
School	SchoolHigh	Alex
Year	-64.40789	340

8. 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.

   Looking at the residuals and plot, the model seems to be a fairly good model

   The spread of the residuals is around a horizontal line without distinct patterns, this indicates that the model does not have a non-linear relationship
   From the Residual plot it is evident that the residuals are linear, and along the line

9. Interpret the Adjusted R-squared value.

   Adjusted R-square =0.5125

   Adjusted R-Square is used to predict the goodness of the model by considering the number of explanatory variables in the model, here the accuracy is 51.25%, and by adding a new predictive variable unless its significant there will be no major change.

10. 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 measures the significance of the overall model

p value of the F statistic 1.265e-06 is less than 0.05 which says that at least one parameter is significant hence Null Hypothesis is rejected

----

ANOVA:

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.

   From the anova table it is clear that, there is at least one predictor variable which is significant rejecting the null hypothesis. (The p value of f statistics is less than 0.05 hence rejecting the null hypothesis)

2. Which predictor variable doies the type 1 anova table suggest you should remove the regression analysis.

   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)

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.

   The residual does not drop significantly removing the Year as the predictor variable.

   Hence it can be considered as the non-significant parameter

----

Diagnostics:

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?

   From avPlot

Predictor Variable	Interpretation
Lot, Size, Lot,Year	Strong slope is observed, hence definitely it is contributing towards the prediction of the expected value

Bath1.1, Bath2, Bath2.1, Bath3,Bath3.1	Strong slope is observed, it seems like point number 37 is the potential outlier, hence it is evident that it is significantly contributing towards the expected value
Bed3,Bed4, Bed5,Bed6	Strong negative slope is observed, point number 4 and 37 is the potential outlier for Houses with Bed3,4,5 and for House with 6 Bed 4 and 37 is the potential outlier. It is evident it is significantly contributing towards the expected value
Garage1, Garage2, Garage3	There is no strong slope with added Garage1/Garage2 keeping other variables constant, there is a strong negative slope with Garage3, points 3 and 37 seem to be the potential outlier
SchoolHigh, SchoolNotreDam,SchoolStMarys, SchoolStratford	Observes a moderate slope, with points 6 and 5 causing a influence for the best line Large concentration is between -0.1 and 0.2
SchoolStLouis	Observes non-linearity, which in turn can be said that SchoolLouis shows no much effect will be on the Expected value of the price

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

2. 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?

   From the Durbin Watson test:

   D-W statistic=1.614157 and corresponding p value= 0.032, so hypothesis of no-auto correlation is rejected and observations cannot be classified as independent

   Reasons for autocorrelation: Outliers are not treated along the process of building the model can be one of the reason

   Repeated observation n multiple observation of the same individual are the common violations.

   The effects of dependent samples on regression are

   A. Structure Dependency
   B. Causes non-constant variance

C. Outliers from different distributions that causes inefficiency

Use of Mixed effect model improves the model in the presence of the dependent samples

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.

   GVIF value of all the regressors are less than 5 hence they are said to be independent of each other. Since certain parameters has more than 1 degrees of freedom we have considered the GVIF^(1/(2*Df)) value for the interpretation

   Problems of Multicollinearity

   If there exists a strong correlation between the two predictor variables then their Beta becomes unstable, the estimate of the Beta will strongly depend on the other predictors that are included in the model

   If the predictors are correlated then we cannot interpret the regression coefficients

   Improvement:

   Remove highly correlated predictors from the model

   Use Partial Least Square Regression, Principal Component Analysis, Ridge regression.

4. 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.

   In the studentized residual vrs 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.

   studentized residuals vrs 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

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.

   Normal Distribution is observed

   Non-normality- effects:
   Critical values of f and t test can go wrong

   Model can be improved/ Corrected by:
   Using transformation of response/predictor variables, or interaction model, or building a different model

----

Leverage, Influence and Outliers:

1. 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.

   The leverage measures the amount by which the predicted value
   would change if the observation was shifted one unit in the y-direction. The leverage always takes values between 0 and 1. A point with zero leverage has no effect on the regression model. If a point has leverage equal to 1 the line must follow the point perfectly

   Yes there exits leverage point in all the predictive variable plot

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.

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.

   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.

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.

   Yes by the observation we can conclude that the model without outliers is providing the best estimate of the House price compared to the model with the outliers

----

Detailed maximum, minimum and median values of categorical variables

Bed:

a. Bed2:
   #Minimum Price is 299.0
   #Maximum Price is 350
   #Median Price is 339.9

b. Bed3:
   #Minimum Price is 189.5
   #Maximum Price is 435
   #Median Price is 339.9
   #Median Price is 339.9

c. Bed4:
   #Minimum Price is 155.5
   #Maximum Price is 450
   #Median Price is 266.6

d. Bed5:
   #Minimum Price is 185
   #Maximum Price is 349.5
   #Median Price is 269.0

e. Bed6:
   #Minimum Price is 252.5
   #Maximum Price is 252.5
   #Median Price is 252.5

Bath: Overall as the number of Bathroom increases from 2, the Price goes on increasing

a. Bath 1
   #Minimum Price is 235.0
   #Maximum Price is 350.0
   #Median Price is 292.5

b. Bath 1.1
   #Minimum Price is 215.0
   #Maximum Price is 385.5
   #Median Price is 325.0
   #Negatively Skewed

c. Bath 2
   #Minimum Price is 155.5
   #Maximum Price is 435.0
   #Median Price is 259.4 (Mean values are large)
   #Positively Skewed

d. Bath 2.1
   #Minimum Price is 189.5
   #Maximum Price is 349.5
   #Median Price is 269.9
   #Data is distributed equally

e. Bath 3
   #Minimum Price is 230.0
   #Maximum Price is 450
   #Median Price is 295
   #Slightly positive skewed

f. Bath 3.1
   #Minimum Price is 285
   #Maximum Price is 345

Garage:

a. Garage0
   #Minimum value is 185.0
   #Maximum value is 388.0
   #Median value is 232(Difference exist between mean and median)

b. Garage1
   #Minimum value is 155.5
   #Maximum value is 385.5
   #Median value is 242.0(Difference exist between mean and median)

c. Garage2
   #Minimum value is 195.0
   #Maximum value is 450.0
   #Median value is 285.0 (Mean values are large)

d. Garage3
   #Minimum value is 299.0
   #Maximum value is 339.9
   #Median value is 319.4 (Mean values are large)