Overview of the Data:

- **Basic Data Dimensions:** The dataset comprises 2,930 observations (or houses) and 82 variables (or features).
- **Target Variable:** `SalePrice` represents the price at which the property was sold, and it's the response variable that we aim to predict or explain.
- **Predictor Variables:** These can be broadly categorized into the following groups:

1. **Identification and Sale Info:**
   - `SID`: A simple serial number.
   - `PID`: Parcel identification number.
   - `SaleType`: Method of sale (e.g., warranty deed, foreclosure).
   - `SaleCondition`: Condition of the sale (e.g., normal, partial).
   - `MoSold`: Month the property was sold.
   - `YrSold`: Year the property was sold.
2. **Property Characteristics:**
   - MS Zoning (Nominal): General zoning classification of the sale
   - Lot Area (Continuous): Lot size in square feet
   - Lot Shape (Ordinal): General shape of the property
   - Land Contour (Nominal): Flatness of the property
   - Utilities (Ordinal): Type of utilities available
   - Lot Config (Nominal): Lot configuration
   - Land Slope (Ordinal): Slope of the property
   - House Style (Nominal): Style of dwelling
   - Year Built (Discrete): Original construction date
   - Roof Style (Nominal): Type of roof
   - Roof Matl (Nominal): Roof material
   - Exterior 1 (Nominal): Exterior covering on the house
   - Exterior 2 (Nominal): Exterior covering on the house (if more than one material)
   - Total Bsmt SF (Continuous): Total square feet of basement area
   - Total Bsmt SF (Continuous): Total square feet of basement area
   - 1st Flr SF (Continuous): Firstfloor square feet
   - 2nd Flr SF (Continuous): Secondfloor square feet
   - Gr Liv Area (Continuous): Above grade living area square feet
   - Yr Sold (Discrete): Year Sold
   - Sale Condition (Nominal): Condition of sale
3. **Sale Property Rating:**
   This considers all the different ratings given to the different parts of the property.
   - Overall Qual (Ordinal): Rates the overall material and finish of the house
   - Overall Cond (Ordinal): Rates the overall condition of the house
4. **Nearby Facilities:**
   This considers facilities available from the property.
   - Lot Frontage (Continuous): Linear feet of street connected to the property

- Street (Nominal): Type of road access to the property
- Alley (Nominal): Type of alley access to the property
- Neighborhood (Nominal): Physical locations within Ames city limits
- Condition 1 (Nominal): Proximity to various conditions
- Condition 2 (Nominal): Proximity to various conditions (if more than one is present)

Define the Sample Population

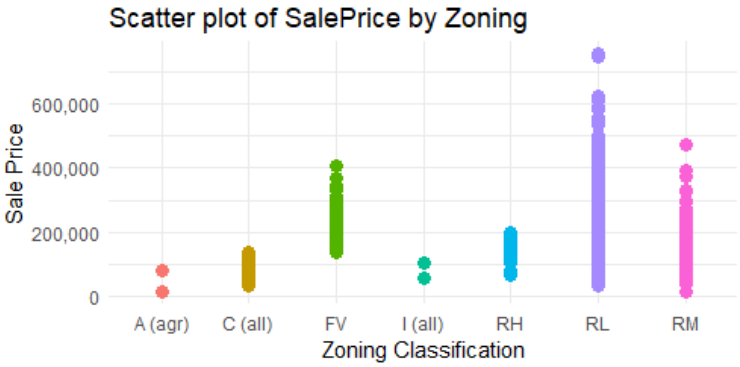
As it says above, we are building regression models for the response variable SalePrice(Y). In order to do this, you need to know the Sample Population. Without this, it is not possible to infer results from the sample to the larger population, it makes the notion of hypothesis testing for population parameter values irrelevant, and it makes the process of determining outliers highly problematic. Frankly, it throws your whole purpose for modeling into chaos.

Defining the Sample Population is actually a very powerful tool for you as the modeler. It gives you license to define what aspects of the data are legitimate for you to work with. You don't have to model ALL of the data you are given in one model. You can break the data up into parts and model them separately. Why would you want to do this? Well, are all properties the same? Would we want to include an apartment building in the same sample as a single family residence? Would we want to include a warehouse or a shopping center in the same sample as a single family residence? Would we want to include condominiums in the same sample as a single family residence? Are there certain kinds of properties that are not like the others? Could one be a derelict property such that it is not like the others? Could one be a mansion such that it is not like the other properties in the data set? You get to define this! In doing this, often many records with extreme scores are eliminated from modeling consideration. Just understand, you get to: a) Define your Target Population and hence the Sample using 'drop conditions'; b) and Create the "waterfall" logic for the drop conditions. If you want to use your conditions from Modeling Assignment #1, that is fine. If you feel you need to make changes, now is the time to do so. Just be sure to include

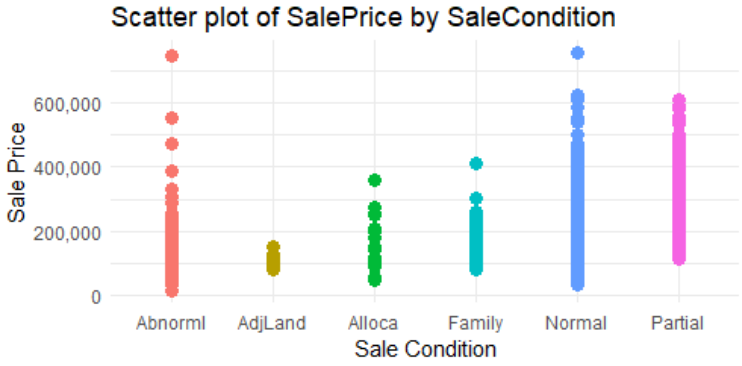
a statement at the beginning of your assignment write-up so that it is clear to any reader what you are excluding from the data set when defining your sample population.

Defining 'Typical' Homes in Ames, Iowa:

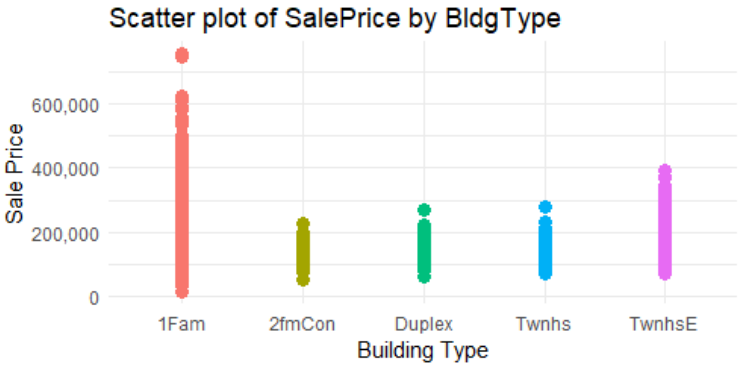
1. Residential Properties: We will retain only those properties that have a `Zoning` value of 'RH', 'RL', or 'RM'. This ensures that we are looking at residential properties only, filtering out commercial and other nonresidential zones.



2. Normal Sale Condition: We will filter out properties where the `SaleCondition` is 'Normal'. This ensures that the dataset represents typical sales and excludes properties that may have been sold between family members or under other unusual circumstances.

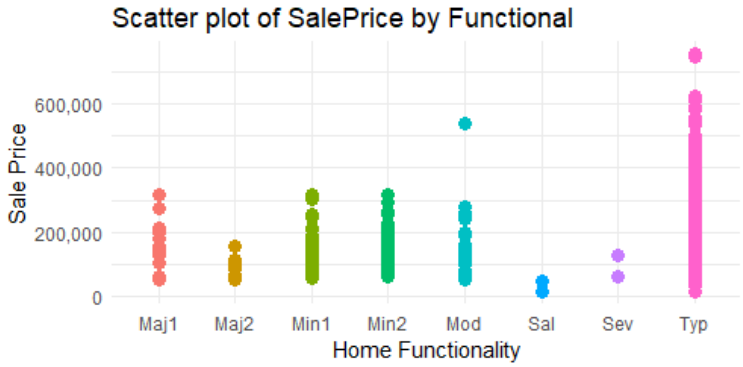


3. Single Family Houses: By selecting homes with `BldgType` as '1Fam', we ensure our dataset represents singlefamily homes. This excludes properties like townhouses, duplexes, and other building types that might not represent the typical homebuyer's preference in Ames.

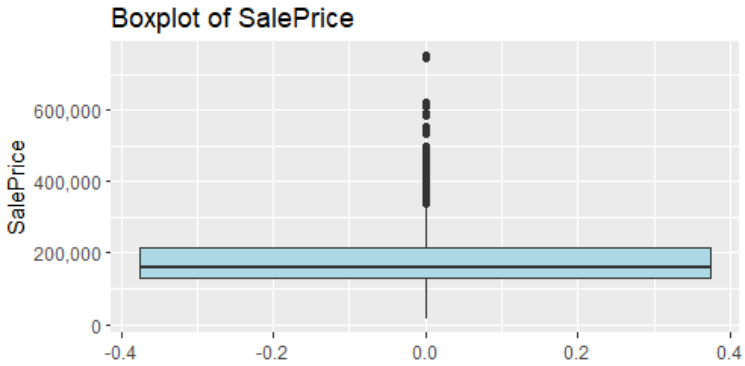


4. Typical Home Functionality: By retaining only homes with a 'Typical' value for the `Functional` variable, we're excluding homes with any kind of damage or any other functionality issues. This

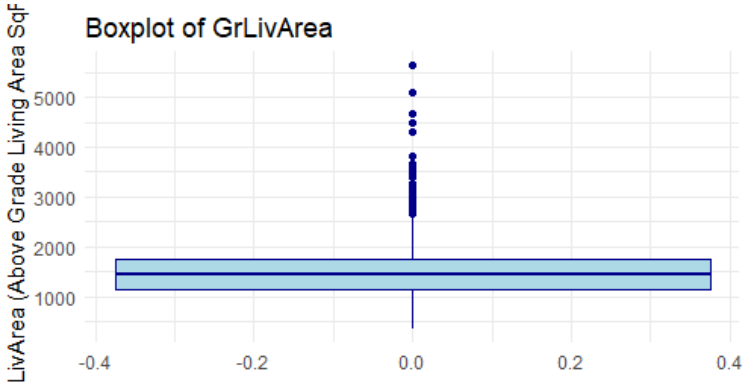
is critical because homes with significant damages or functionality problems may not be representative of a 'typical' home.

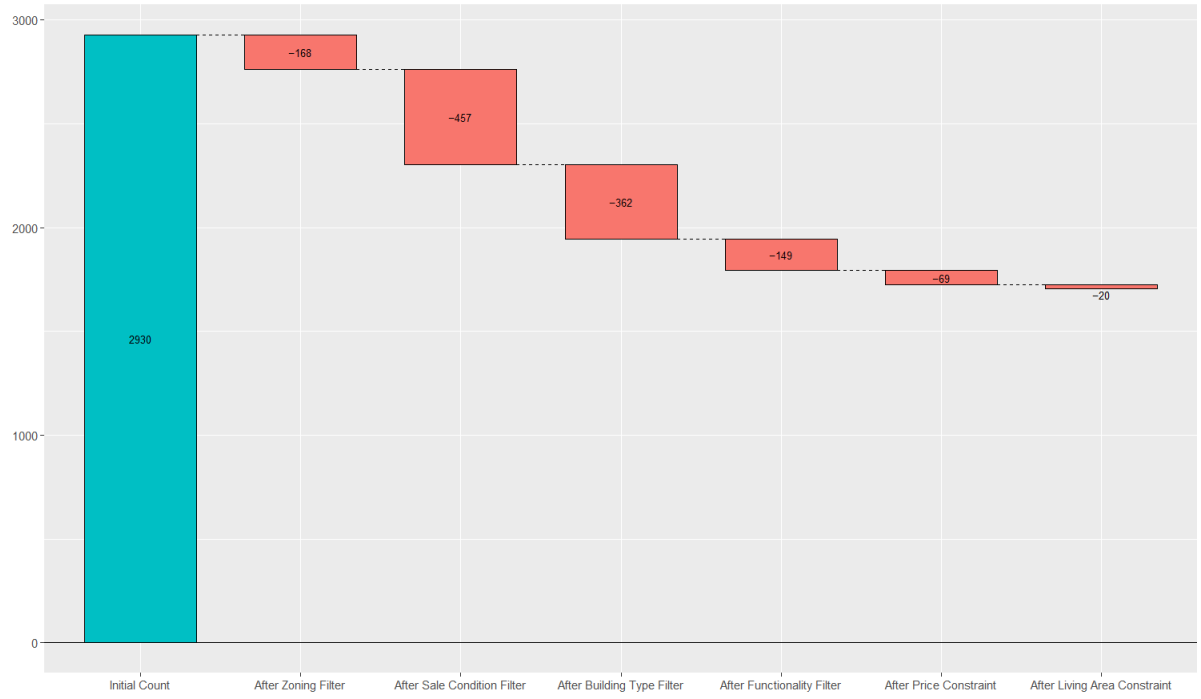


5. Price Constraint: From the provided insight, properties priced above $ 339500 (Q3+1.5*IQR) are considered outliers. By filtering out these properties, we're ensuring that we're modelling prices that are representative of the majority of homes in Ames.



6. Living Area Constraint: By retaining properties with a GrLivArea of 2667(Q3+1.5*IQR) sq. ft. or less, we're further refining our dataset to exclude atypically large homes. Our observation from the box showed that properties above this threshold were unusual compared to the rest.





The sample population (filtered dataframe) has 1705 rows and 82 columns.

Exploratory Data Analysis

Once the Sample Population is clearly well defined, and you've selected only those records and fit the Sample Population definition, you can then continue to perform a detailed Exploratory Data Analysis (EDA). Usually, this is broken up into two parts. The first is data preparation (or data cleaning). Here, you concern yourself with any remaining missing values, extreme scores, and outliers.

- Are there variables with missing values? Should values for these variables be imputed or "fixed"? You can impute values for the missing data points by using a mean or median for the variable. Or, maybe use a decision tree, other contextual information, or models. For variables with large numbers of missing values, you may want to simply eliminate that variable from the dataset. One option is to not do anything. In R, the default way that missing values are handled is to remove the record with a missing value from the

computation, if the variable with the missing value is included in the function. Always keep this fact in mind.

- Do any of the variables have outliers or extreme values? Should these extreme values be replaced? Fix any extreme values that need fixing. Note: This may be something you do in conjunction with the EDA as you find extreme values.

Then, you can turn your attention to understanding the data more deeply. You were exposed to the EDA ideas and traditions in Module 1. For a full blown modeling project, you would want to exam all of the variables in your data set. Some suggestions for things that you could do are:

- Obtain histograms for each continuous variable
- Obtain summary statistics, such as: Means, standard deviations, minimum, maximum, median for all continuous variables
- Are the explanatory variables correlated to the response variable?
- Are the explanatory variables correlated amongst themselves?
- Obtain scatterplots of explanatory variables with the response variable.
- Do you want to create new variables to make the analysis more easily interpretable? For example, you might want to create a variable like PRICE per SQR FOOT. This could be a more meaningful response variable than total home sale price. I'm sure with a little bit of google searching you can find other variables that you would want to compute and potentially use. This is totally voluntary on your part. Not at all required. Do this if you have the interest or think such variables might be of value.

The amount of preparation and EDA is totally up to you. This is always the way it is in practice. No one will ever tell you when you are "done" with data cleaning. From prior experience, up to 90% of one's time modeling data is spent on data cleaning and preparation issues, depending on the type of data one is working with. Just remember: Garbage in → Garbage out! For this assignment, you want to be sure you have a dataset that you are comfortable working with for the remainder of this assignment. All of the statistics and graphs you produce in preparation for modeling are for you. They will be helpful as you go through the following steps, but you do not need to report

anything here. There is nothing that needs to be written about data preparation for this assignment.

1. Missing Values:
    a. There are 338 missing values in the `LotFrontage` column. – dropped the column
    b. The `MasVnrType` and `MasVnrArea` columns each have 9 missing values – dropped the 9 rows.
    c. The `BsmtExposure` column has 2 missing values – dropped the two rows.
    d. The `BsmtFinType2` and `Electrical` columns each have 1 missing value – dropped the row.
    e. The `GarageYrBlt` column has 61 missing values – dropped the column.
2. Outlier Detection: Outliers concerning `SalePrice` have been previously removed.
3. Duplicate Rows: There are no duplicate rows in the dataframe.
4. Consistency Check: YearBuilt < YearRemodel for all records.
5. Other Anomalies: SalePrice > 0 for all records.

- The sample population (filtered dataframe) now has 1691 rows and 80 columns.
- 36 of the 79 predictor variables are continuous, and their correlation with SalePrice.
- The correlation of these 36 continuous predictors with SalePrice is:

Variable	Correlation
OverallQual	0.7920570
GrLivArea	0.7683145
GarageCars	0.6557352
FullBath	0.6511756
GarageArea	0.6102894
YearBuilt	0.6014518
FirstFlrSF	0.5963514
TotRmsAbvGrd	0.5932154
TotalBsmtSF	0.5814266
YearRemodel	0.5227082
Fireplaces	0.4682192
MasVnrArea	0.4496138
BsmtFinSF1	0.3721144
SecondFlrSF	0.3566349
HalfBath	0.3557066
LotArea	0.3333270
OpenPorchSF	0.3307229
BedroomAbvGr	0.3163413
WoodDeckSF	0.3031849
BsmtFullBath	0.2385160
BsmtUnfSF	0.1467313
SubClass	0.1307916
ScreenPorch	0.0639037
ThreeSsnPorch	0.0378414
YrSold	0.0358727
PoolArea	0.0041181
MiscVal	-0.0084463
MoSold	-0.0114247
BsmtFinSF2	-0.0196178
KitchenAbvGr	-0.0390430
SID	-0.0578514
BsmtHalfBath	-0.0586424
LowQualFinSF	-0.0660163
PID	-0.0939847
EnclosedPorch	-0.1434736

OverallCond	-0.1582651

Table 1

- For building regression models, we have selected variables that have a correlation > 0.5 with SalePrice.

Variable	Correlation
OverallQual	0.7920570
GrLivArea	0.7683145
GarageCars	0.6557352
FullBath	0.6511756
GarageArea	0.6102894
YearBuilt	0.6014518
FirstFlrSF	0.5963514
TotRmsAbvGrd	0.5932154
TotalBsmtSF	0.5814266
YearRemodel	0.5227082

Table 2

- Together with these 10 variables and SalePrice, we created a new dataframe, and the dimensions of this new dataframe are 1691 rows and 11 columns.
- The number of outliers in these ten variables is:
  OverallQual: 1
  GrLivArea: 7
  GarageCars: 5
  GarageArea: 8
  YearBuilt: 6
  FirstFlrSF: 27
  TotRmsAbvGrd : 4
  TotalBsmtSF: 66

- Dropped the rows with outliers in OverallQual, GrLivArea, GarageCars, and GarageArea, and YearBuilt. And, columns FirstFlrSF and TotalBsmtSF.
- The new dataframe has 1659 rows and 9 columns.
  Certainly! Below is a brief note on each of the mentioned columns:

1. OverallQual:
   - Description: Rates the overall material and finish of the house.
   - Scale: Typically on a scale from 1 (very poor quality) to 10 (very excellent quality).

2. GrLivArea:
   - Description: Represents the above grade (ground) living area in square feet. It's the amount of living area/usable space excluding the basement.

3. GarageCars:
   - Description: Size of the garage in terms of car capacity. Indicates how many cars can fit into the garage.

4. FullBath:
   - Description: Represents the number of full bathrooms in the house. A full bathroom contains four plumbing fixtures: a toilet, sink, bathtub, and shower.

5. GarageArea:
    - Description: Size of the garage in square feet. It quantifies the space available in the garage.

6. YearBuilt:
    - Description: The original construction date, representing the year when the house was built.

7. TotRmsAbvGrd:
    - Description: Represents the total rooms (excluding bathrooms) above grade. It gives a sense of the spaciousness and functionality of the house.

8. YearRemodel:
    Description: Represents the remodeling date. It's the year when the house was last remodeled. If no remodeling or additions, it might be the same as 'YearBuilt'.

9. SalePrice:
    - Description: The property's sale price in dollars. This is the target variable you might be trying to predict in a regression model.

These columns are crucial in determining the value and desirability of a property. For instance, properties with higher `OverallQual`, larger `GrLivArea`, and recent `YearBuilt` or `YearRemodel` tend to have a higher `SalePrice`. Similarly, features like the number of full bathrooms (`FullBath`)

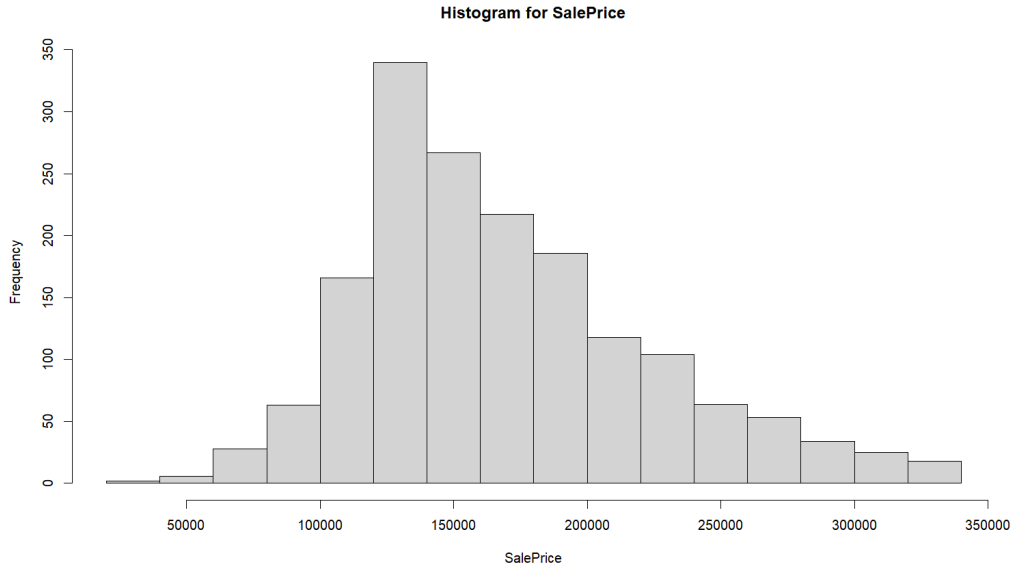
and the capacity of the garage (`GarageCars` and `GarageArea`) significantly influence a home's sale price.

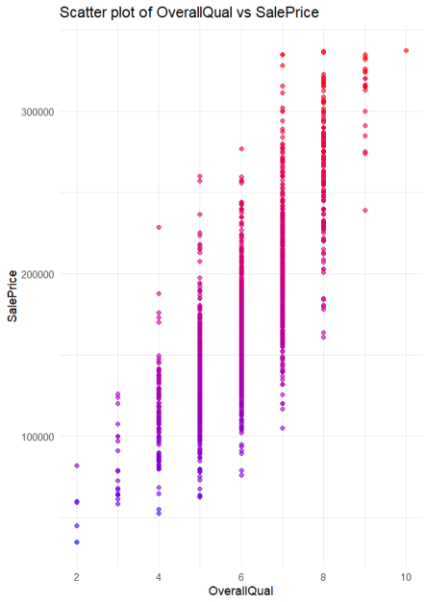
The summary of the data in this dataframe is:

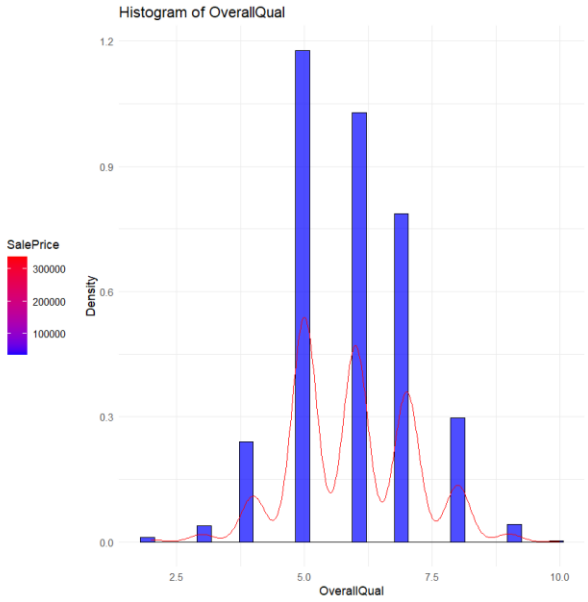
Variable	Mean	Median	Std_Dev	Max	Min
OverallQual	5.915612	6	1.1778478	10	2
GrLivArea	1416.137432	1392	417.3210477	2614	438
GarageCars	1.677517	2	0.6655722	3	0
FullBath	1.471368	1	0.5229302	3	0
GarageArea	445.705244	458	179.1971421	912	0
YearBuilt	1967.249548	1967	28.6754087	2010	1885
TotRmsAbvGrd	6.266426	6	1.2775882	10	3
YearRemodel	1982.030741	1990	20.7162668	2010	1950
SalePrice	169027.877637	158000	53967.8101308	337500	35000

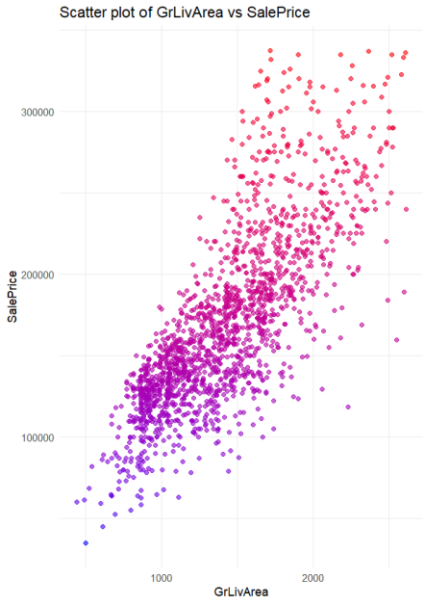
Table 3

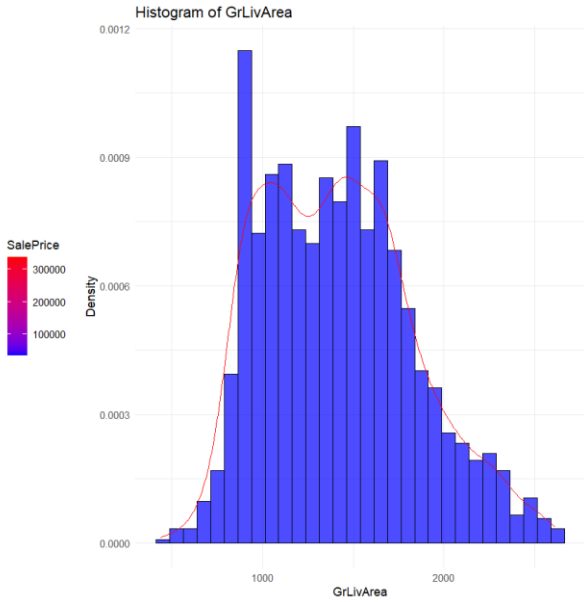
Plots:

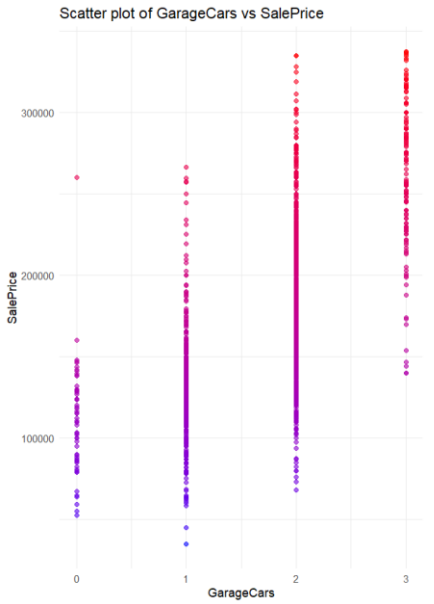


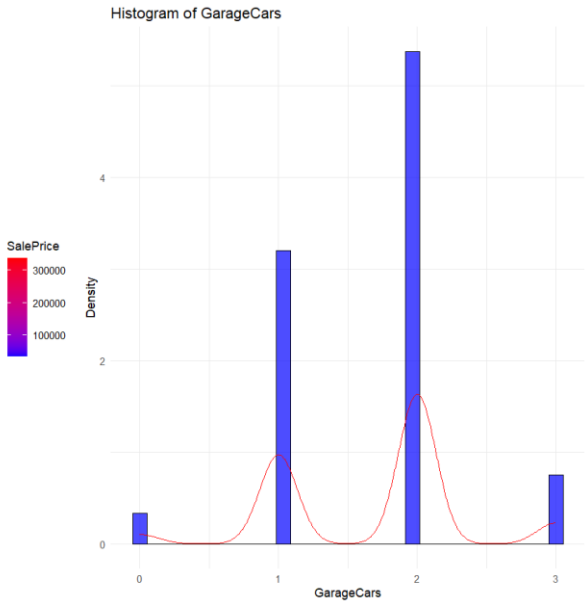

Scatter plot of OverallQual vs SalePrice
SalePrice
300000
200000
100000
OverallQual
2
4
6
8
10

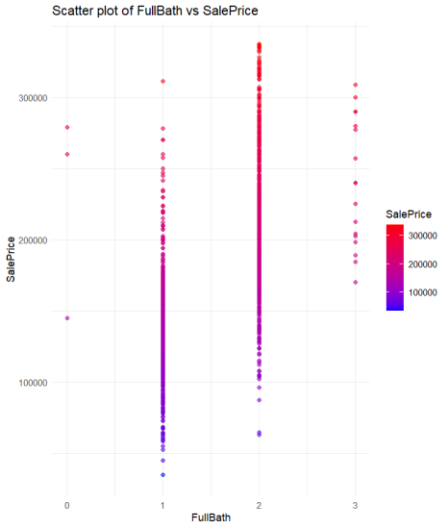

Histogram of OverallQual
Density
1.2
0.9
0.6
0.3
0.0
2.5
5.0
7.5
10.0
OverallQual

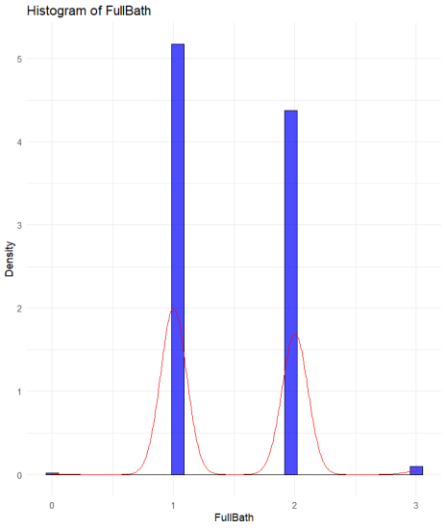

Scatter plot of GrLivArea vs SalePrice
SalePrice
300000
200000
100000
GrLivArea
1000
2000


Histogram of GrLivArea
Density
0.0012
0.0009
0.0006
0.0003
0.0000
1000
2000
GrLivArea

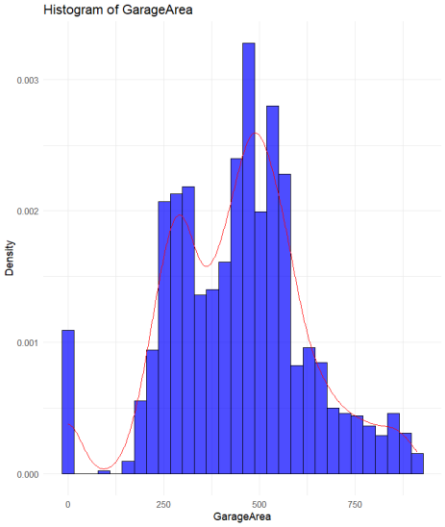

Scatter plot of GarageCars vs SalePrice
SalePrice
300000
200000
100000
GarageCars
0
1
2
3

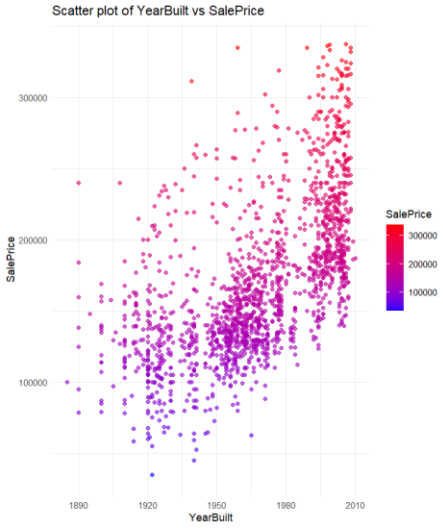

Histogram of GarageCars
Density
4
2
0
0
1
2
3
GarageCars

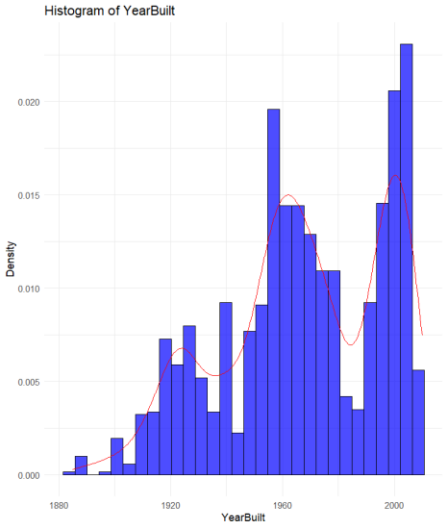

Scatter plot of FullBath vs SalePrice
SalePrice
FullBath
SalePrice
300000
200000
100000
0
1
2
3

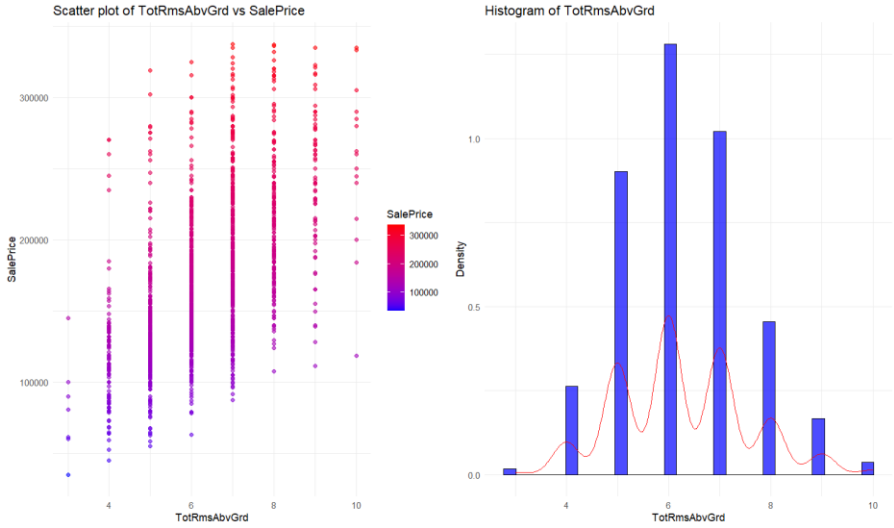

Histogram of FullBath
Density
FullBath
0
1
2
3
4
5

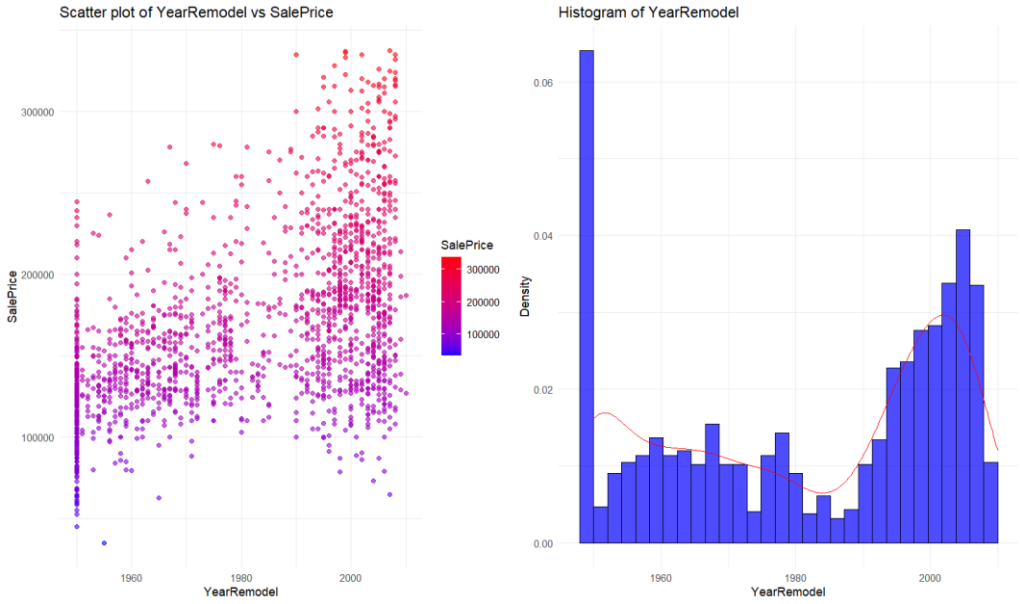

Scatter plot of GarageArea vs SalePrice
SalePrice
GarageArea
SalePrice
300000
200000
100000
0
250
500
750


Histogram of GarageArea
Density
GarageArea
0.000
0.001
0.002
0.003


Scatter plot of YearBuilt vs SalePrice
SalePrice
YearBuilt
SalePrice
300000
200000
100000
1890
1920
1950
1980
2010


Histogram of YearBuilt
Density
YearBuilt
0.000
0.005
0.010
0.015
0.020
1880
1920
1960
2000





1. Let Y = sale price be the dependent or response variable. Select what you consider to be "the best" continuous explanatory variable from the AMES data set to predict Y. Discuss what criteria you used to select this explanatory variable? Fit a simple linear regression model using your

explanatory variable X to predict SALE PRICE(Y). Call this Model 1. To report the results for Model 1, you are to:

a. Make a scatterplot of Y and X, and overlay the regression line on the cloud of data.
b. Report the model in equation form and interpret each coefficient of the model in the context of this problem.
c. Report and interpret the R-squared value in the context of this problem.
d. Report the coefficient and ANOVA Tables.
e. Clearly specify the hypotheses associated with each coefficient of the model, as well as the hypothesis for the overall omnibus model. Conduct and interpret these hypothesis tests.
f. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. Check on these underlying assumptions by plotting:
    - Histogram of the standardized residuals
    - Scatterplot of standardized residuals (Y) by predicted values (Y_hat)

   Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.
g. Check on leverage, influence and outliers. These points can be identified by several statistics such as DFFITS, Cook's Distance, Leverage, and Influence. Discuss any issues or concerns. Describe what course of action should be taken.

Selection of `GrLivArea` as the Best Explanatory Variable:
`GrLivArea` represents the above ground living area in square feet, which is an actual attribute of the property. The size of a house, particularly the livable area, is intuitively an important determinant of its price.
Criteria for Selection:

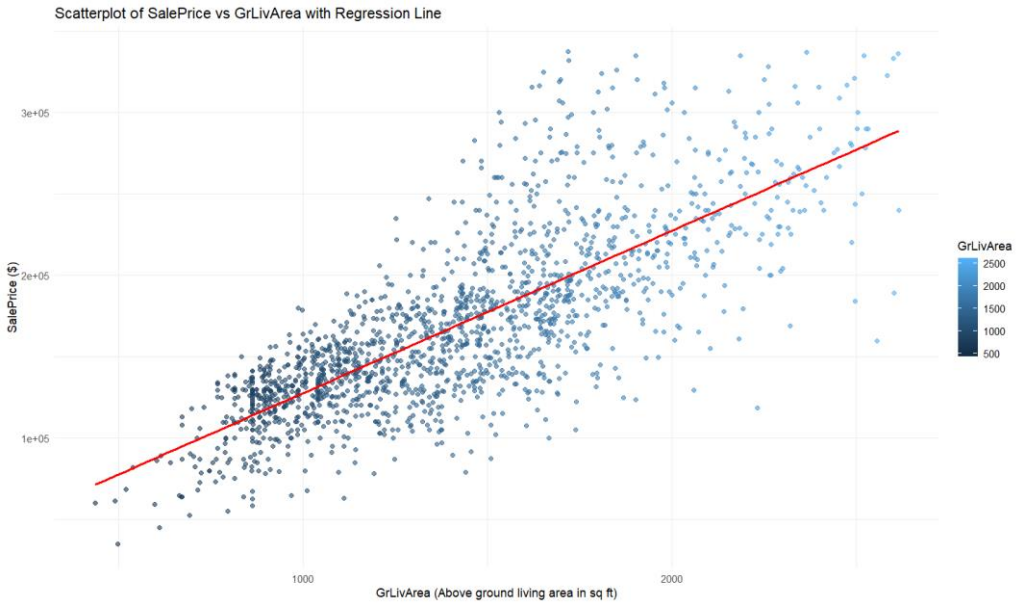
1. Strength of Linear Relationship: As indicated by the correlation coefficient of approximately 0.77, there's a strong linear relationship between `GrLivArea` and `SalePrice`.
2. Relevance: The size of a property (in terms of livable area) is a fundamental attribute that would naturally influence its price.
3. Actual Property Attribute: Unlike subjective scales or ratings, `GrLivArea` provides a direct measurement, making it a tangible and easily interpretable attribute.
4. Simplicity: In the context of simple linear regression, we're interested in using just one explanatory variable. `GrLivArea` stands out due to its high correlation and direct relevance to property valuation.
5. Fitting Model_1: Simple Linear Regression using `GrLivArea` to predict `SalePrice`

```
> Model_1 <- lm(SalePrice ~ GrLivArea, data = new_df)
>
> # Display the summary
> summary(Model_1)
```

```
Call:
lm(formula = SalePrice ~ GrLivArea, data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max
-131845  -20608    -539   17347  138312

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27535.118   2977.610   9.247   <2e-16 ***
GrLivArea      99.915      2.017  49.538   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34270 on 1657 degrees of freedom
Multiple R-squared:  0.5969,   Adjusted R-squared:  0.5967
F-statistic:  2454 on 1 and 1657 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: SalePrice
            Df     Sum Sq    Mean Sq F value    Pr(>F)
GrLivArea    1 2.8826e+12 2.8826e+12    2454 < 2.2e-16 ***
Residuals 1657 1.9464e+12 1.1746e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Equation of the Model:

$$SalePrice = 27535.118 + 99.915 * GrLivArea$$



1. Intercept (`(Intercept)`): The estimated intercept is 27535.118. This implies that when the `GrLivArea` is 0 (which is not practically possible), the predicted `SalePrice` would be 27,535.12.

This value is more of a mathematical result and should be interpreted with caution since having a living area of 0 square feet doesn't have practical relevance.

2. Slope (`GrLivArea`): The estimated slope for `GrLivArea` is 99.92. This means for every additional square foot of above-ground living area, the `SalePrice` is expected to increase by approximately $99.92, holding all else constant.

3. Significance of the Coefficients: Both the intercept and the slope are statistically significant at a very low p-value (`< 2.2e-16`), as indicated by the three asterisks (*). This implies that there's a significant linear relationship between `GrLivArea` and `SalePrice`.

5. R-squared (`Multiple R-squared`): The R-squared value is 0.5969, which means that approximately 59.69% of the variation in `SalePrice` can be explained by `GrLivArea` in this model. This is a fairly strong value for a simple linear regression model, indicating that `GrLivArea` is a good predictor of `SalePrice`.

6. F-statistic: The F-statistic value is 2454, and its associated p-value is extremely small (`< 2.2e-16`). This tests the overall significance of the model, suggesting that the model with `GrLivArea` as the predictor is significantly better at explaining the variation in `SalePrice` compared to a model with no predictor.

7. The ANOVA table tests the hypothesis that the model coefficients are equal to zero. If the p-value (usually labeled "Pr(>F)") is very small, it indicates that at least some of the predictors are statistically significant.

For a simple linear regression model like $Y = \beta 0 + \beta 1 * X + \mathcal{E}$ where Y is the dependent variable (`SalePrice`), and X is the independent variable (`GrLivArea`), the hypotheses associated with each coefficient and for the overall model can be stated as follows:

1. For the Intercept, $\beta 0$:
   Null Hypothesis $H0$: $\beta 0 = 0$
   This hypothesizes that when X is zero, Y is also expected to be zero.
   Alternative Hypothesis Ha: $\beta 0 \neq 0$
   This hypothesizes that when X is zero, Y is not necessarily zero.
2. For the Slope, $\beta 1$ (associated with `GrLivArea`):
   Null Hypothesis $H0$: $\beta 1 = 0$
   This hypothesizes that there is no linear relationship between the dependent and the independent variable.
   Alternative Hypothesis Ha: $\beta 1 \neq 0$
   This hypothesizes that there is a linear relationship between the dependent and the independent variable.
3. For the overall model (Omnibus test):
   Null Hypothesis $H0$: All coefficients (excluding the intercept) in the model are equal to zero.
   Alternative Hypothesis Ha: At least one coefficient in the model is not equal to zero.

1. For the intercept, $\beta 0$: The estimate is 27535.12 with a very small p-value (given as `< 2e-16`, which is practically zero). This means we reject the null hypothesis, and there's statistically significant evidence to suggest that the expected SalePrice is not zero when GrLivArea is zero.

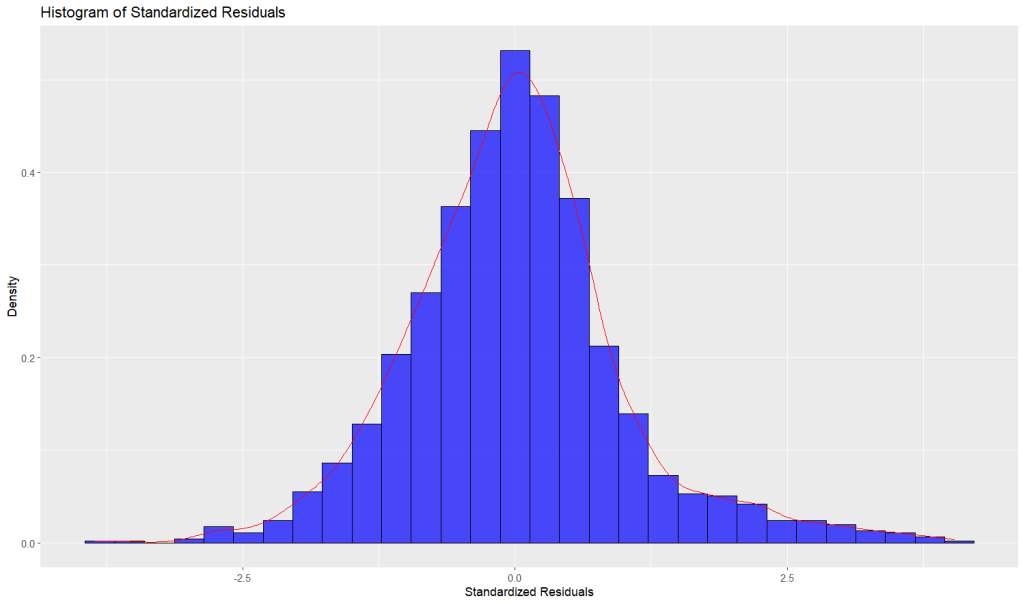
2. For the slope, $\beta 1$: The estimate is 99.92 with a very small p-value (`< 2e-16`). This means we reject the null hypothesis and there's statistically significant evidence to suggest that for every one unit increase in `GrLivArea`, the SalePrice increases by approximately 99.92 units, on average.

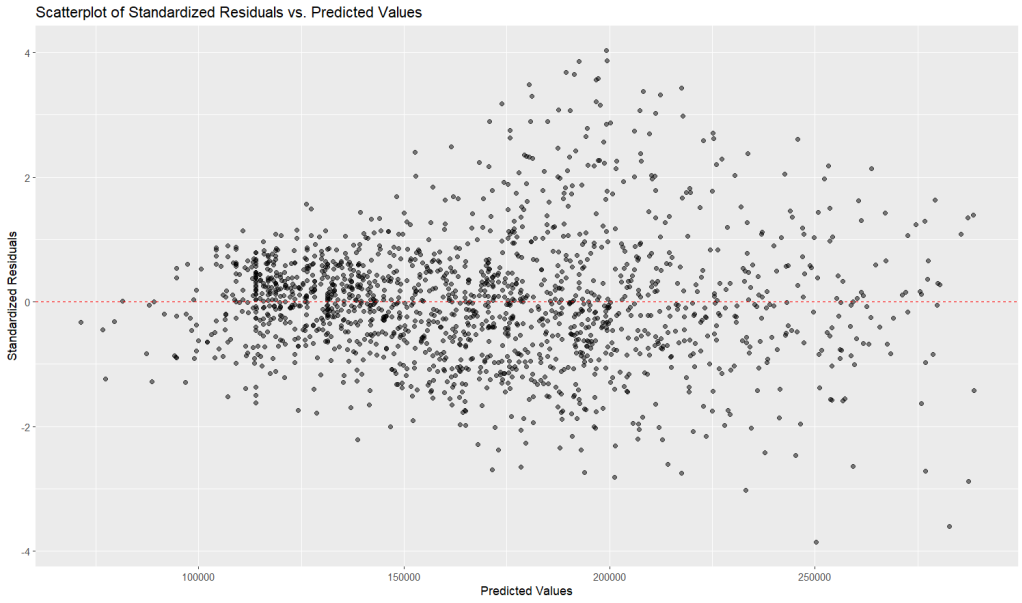
3. For the overall model:

The F-statistic is significantly large, and the p-value for the test is very small (`< 2.2e-16`). This suggests that the model with `GrLivArea` as a predictor fits significantly better than a model with no predictors (i.e., just an intercept). We reject the null hypothesis of the omnibus test, suggesting that the model is statistically significant.

In summary, both the intercept and slope are statistically significant, and the model with `GrLivArea` as a predictor provides a significant fit to the data. `SalePrice` based on `GrLivArea`, though it's essential to be aware of its limitations and potential factors not accounted for in this simple regression framework. `GrLivArea` is indeed a strong and significant predictor of `SalePrice`.

The model suggests that for every one square foot increase in above-ground living area, the sale price of a home increases by approximately 99.92.





1. Histogram of Standardized Residuals: The histogram of the standardized residuals aims to illustrate if the residuals are normally distributed, which is one of the assumptions for linear regression.

- Observation: The histogram, when overlaid with the density curve (in red), suggests that the residuals are approximately normally distributed around the center. However, there's a slight skewness to the right, indicating a potential deviation from perfect normality, especially with some elongated tails on the right.
- Interpretation: This slight right skewness suggests that there are a few observations where the predicted values are underestimating the actual sale price. The elongated tails on the

right indicate potential outliers or influential points that may affect the model's performance.

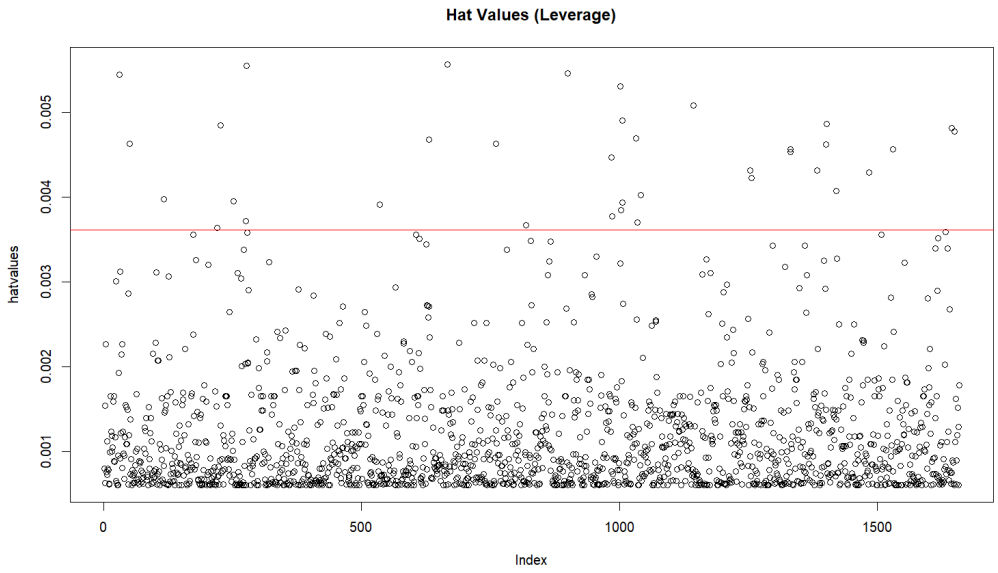
2. Scatterplot of Standardized Residuals vs. Predicted Values: This scatterplot allows for the diagnosis of heteroscedasticity. In a well-behaved regression model, we would anticipate seeing a random scatter of points, with no discernible pattern.

- Observation: Most residuals cluster around the zero line, as expected. However, there seems to be a slight funnel shape, where the spread of residuals increases as the predicted values increase.
- Interpretation: The funnel shape suggests heteroscedasticity in the data, meaning the variability of the residuals isn't constant across all levels of the independent variable. This violates the assumption of homoscedasticity in linear regression. Heteroscedasticity can produce inefficient estimates of the coefficients and can lead to incorrect conclusions about the significance of the predictors.

In summary, while the residuals exhibit approximate normality, there are slight deviations and potential outliers to consider. The presence of heteroscedasticity suggests that a log or It would be worth exploring methods to address heteroscedasticity, such as using transformation on the response variable or considering robust regression techniques. Regression model might not be the perfect model for this data, or some data transformation a different modeling approach might be

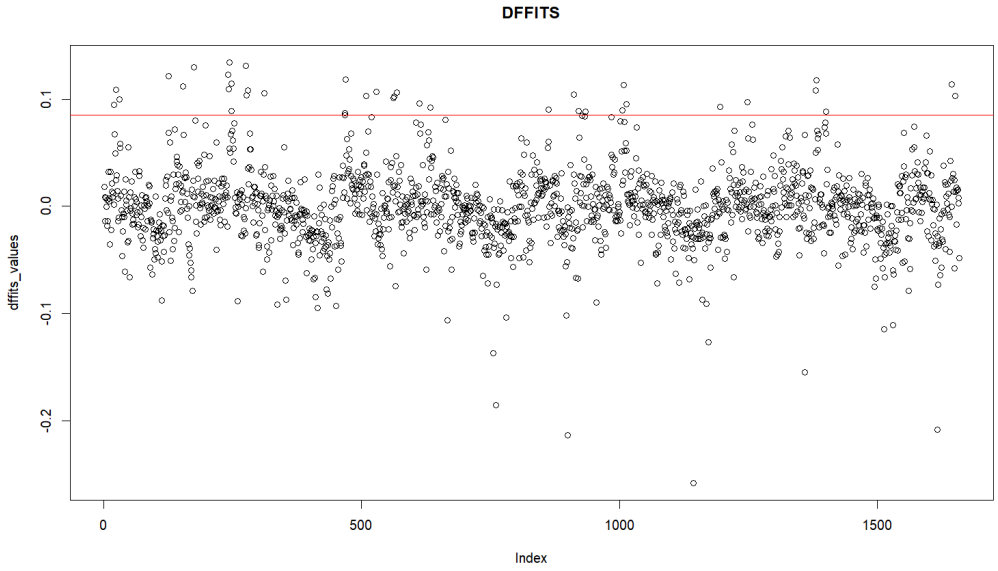
```r
# Leverage
hatvalues <- hatvalues(Model_1)
plot(hatvalues, main = "Hat Values (Leverage)")
abline(h = 2*(2+1)/length(new_df$SalePrice), col = "red")

# DFFITS
dffits_values <- dffits(Model_1)
plot(dffits_values, main = "DFFITS")
abline(h = 2*sqrt((2+1)/length(new_df$SalePrice)), col = "red")

# Cook's Distance
cooks_d <- cooks.distance(Model_1)
plot(cooks_d, main = "Cook's Distance")
abline(h = 1, col = "red")
```

required.



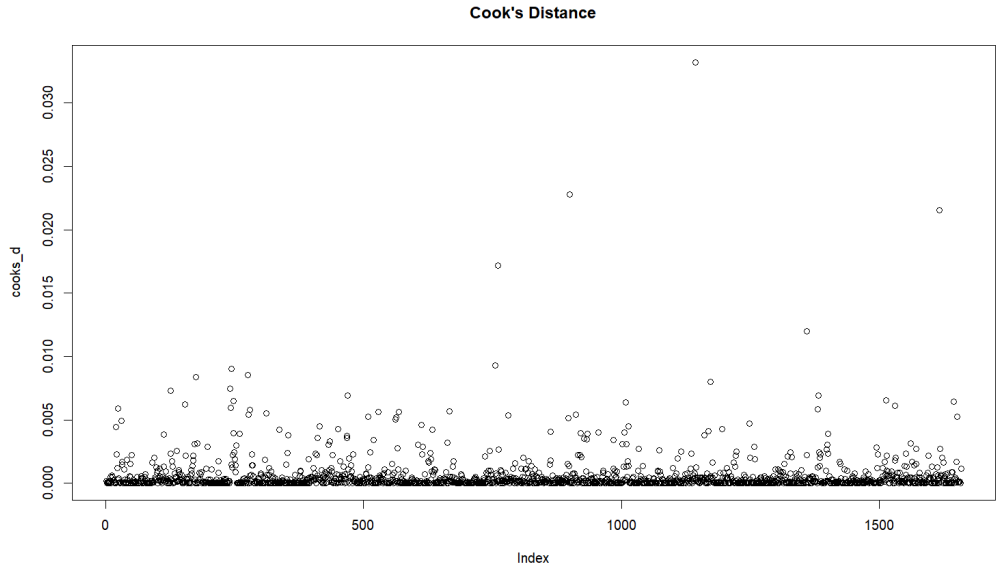
Hat values measure the influence an observation has on its own predicted value, and it helps identify points that might have undue leverage on the regression line. Observations with hat-values greater than $2(p+1)/n$ (where p is the number of predictors and p is the number of observations) are considered to have high leverage. From the plot, it appears that only a few data points exceed the red threshold line, suggesting that they have high leverage. While these points

don't necessarily have a large influence on the regression model, they might potentially have undue leverage because of their extreme values for the predictors.



DFFITS measures the influence of each observation on the predicted value for that observation, essentially telling us how much the predicted value for a given observation would change if that observation were omitted from the model. A rule of thumb is that absolute values greater than $2\sqrt{\frac{p+1}{n}}$ suggest high influence. From the plot, there are a few points that slightly exceed the red

threshold line, suggesting they have some influence on the regression model. However, the majority of the data seems to be within acceptable limits.



Cook's distance measures the influence of each observation on the regression model's fitted values. The plot shows the Cook's distance values for each observation in the dataset. The commonly used threshold for identifying influential points is Cook's distance greater than 1. In this plot, no data point exceeds this threshold, which means there aren't any observations that are particularly influential in the model.

The diagnostic plots suggest that the regression model is relatively robust. While there are a few points of concern in terms of influence and leverage, they are not extreme. It's essential to understand the context of these points — for example, why certain observations have high leverage or influence. If these observations are valid and not a result of data entry errors or other

anomalies, it's often best to keep them in the model and account for their influence in any interpretations or predictions you make.

Issues & Concerns:

1. Few Points of High Influence: In the DFFITS plot, there are a handful of points that slightly exceed the threshold. While they are not highly influential, they can still impact the stability and reliability of the regression model.

2. Potential Leverage Points: The Hat Values plot indicates that there are a few observations with higher-than-average leverage. These points don't necessarily influence the regression outcome, but they have the potential to do so because of their extreme predictor values.

3. General Spread and Density: It's evident from the plots that the bulk of the data has a low influence, low leverage, and is densely packed around low values. However, there are sporadic points with higher values in all three plots.

Course of Action:

1. Investigate High Influence Points:
   - Examine the few points that have a high influence in the DFFITS plot. It's essential to understand why these points are different. Are they valid data points or results from data entry errors, outliers, or anomalies?
   - If these points are errors or outliers that don't represent the population you're interested in, consider removing them to improve the model's reliability.

2. Scrutinize Leverage Points:
   - Assess the observations with high leverage values. High leverage points can be a result of extreme predictor values, which can be either valid data or outliers.
   - Just like with influential points, if these are errors or non-representative outliers, consider removing them. However, if they're valid points, they should be included to maintain the model's integrity.

3. Model Refinement:
   - Depending on the significance and nature of the influential and leverage points, consider refining the regression model. This could involve adding interaction terms, quadratic terms, or considering a different model altogether if the current one isn't appropriate.

4. Additional Diagnostic Checks:
   - Beyond these three plots, consider other diagnostic checks such as residual plots, normal probability plots, or variance inflation factor (VIF) to detect multicollinearity. These can provide a fuller picture of the model's validity and assumptions.

5. Domain Knowledge:
   - Always incorporate domain knowledge. Even if a point is statistically influential or has high leverage, if it's an expected and valid observation based on subject matter expertise, it might be crucial to keep it in the model.

6. Communication:
    - If presenting findings or making decisions based on this regression model, it's essential to communicate the potential concerns about high influence and leverage points so stakeholders are aware of the model's potential limitations.

In summary, while the diagnostic plots suggest the model seems to be fairly robust, the presence of a few influential and leverage points warrants a closer examination. Adjustments should be made based on the nature of these points, additional diagnostic checks, and domain knowledge.

2. For Task 2, you will fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). Call this Model 2. The explanatory variables for Model 2 should be the explanatory variable you had in Model 1, plus the OVERALL QUALITY variable. To report the results for Model 2, you are to:

    a. Report the prediction equation and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here relative to the simple linear regression model in Task 1?
    b. Report and interpret the R-squared value in the context of this problem. Calculate and report the difference in R-squared between Model 2 and Model 1. Interpret this difference.
    c. Report the coefficient and ANOVA Tables.
    d. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the overall omnibus model. Conduct and interpret these hypothesis tests.
    e. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. Check on these underlying assumptions by plotting:
        - Histogram of the standardized residuals
        - Scatterplot of standardized residuals (Y) by predicted values (Y_hat)

        Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.
    f. Check on leverage, influence and outliers, and discuss any issues or concerns.
    g. Based on the information, should you want to retain both variables as predictor variables of Y? Discuss why or why not.

```
> # Display the summary
> summary(Model_2)

Call:
lm(formula = SalePrice ~ GrLivArea + OverallQual, data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max
-101834  -16315    -189   15154  112089

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -52179.334   3370.105  -15.48   <2e-16 ***
GrLivArea       58.930      2.013   29.27   <2e-16 ***
OverallQual  23286.617    713.201   32.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26740 on 1656 degrees of freedom
Multiple R-squared:  0.7548,    Adjusted R-squared:  0.7545
F-statistic:  2549 on 2 and 1656 DF,  p-value: < 2.2e-16

>
>
> # ANOVA table
> anova_table <- anova(Model_2)
> print(anova_table)
Analysis of Variance Table

Response: SalePrice
             Df     Sum Sq    Mean Sq F value    Pr(>F)
GrLivArea     1 2.8826e+12 2.8826e+12  4031.4 < 2.2e-16 ***
OverallQual   1 7.6228e+11 7.6228e+11  1066.1 < 2.2e-16 ***
Residuals  1656 1.1841e+12 7.1503e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a. Prediction Equation and Coefficient Interpretation:

- The prediction equation is:

$$SalePrice = -52{,}179.334 + 58.930 * GrLivArea + 23{,}286.617 * OverallQual$$

- Intercept (-52,179.33): When both `GrLivArea` and `OverallQual` are zero (which might not be practically meaningful), the predicted `SalePrice` would be a negative value of -52,179.33.
- GrLivArea (58.93): For every additional square foot of above-ground living area (`GrLivArea`), holding `OverallQual` constant, the `SalePrice` increases by approximately $58.93.
- OverallQual (23,286.62): For every one unit increase in the overall quality (`OverallQual`), holding `GrLivArea` constant, the `SalePrice` increases by approximately $23,286.62.
- Difference from Task 1: In simple linear regression, we only had one predictor and interpreted its coefficient in isolation. Here, we are interpreting each coefficient while holding the other constant.

b. R-squared Value and Difference:

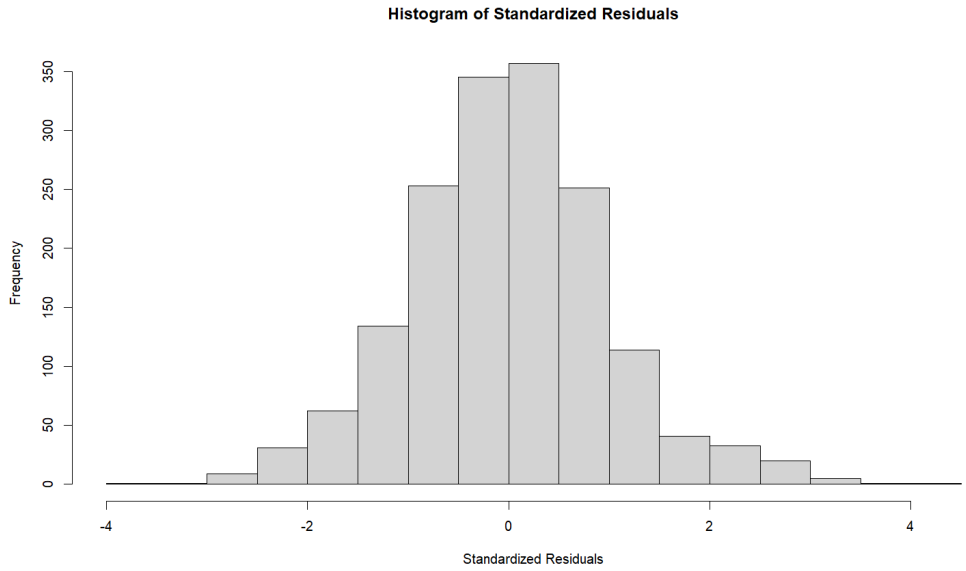
- The R-squared value for Model 2 is 0.7548, meaning 75.48% of the variability in `SalePrice` can be explained by `GrLivArea` and `OverallQual`.

- The R-squared value for Model 1 is 0.5969, meaning 59.69% of the variability in `SalePrice` can be explained by just `GrLivArea`.
- The difference in R-squared between Model 2 and Model 1 is:
- $\triangle R^2 = 0.7548 - 0.5969 = 0.1579$
- Thus, by adding `OverallQual` as a predictor to the model, the R-squared value increased by 15.79%. This indicates that `OverallQual` provides additional explanatory power in predicting `SalePrice` beyond what `GrLivArea` provides alone.

c. Coefficient and ANOVA Tables: Part of the R output

d. Hypotheses and Tests:

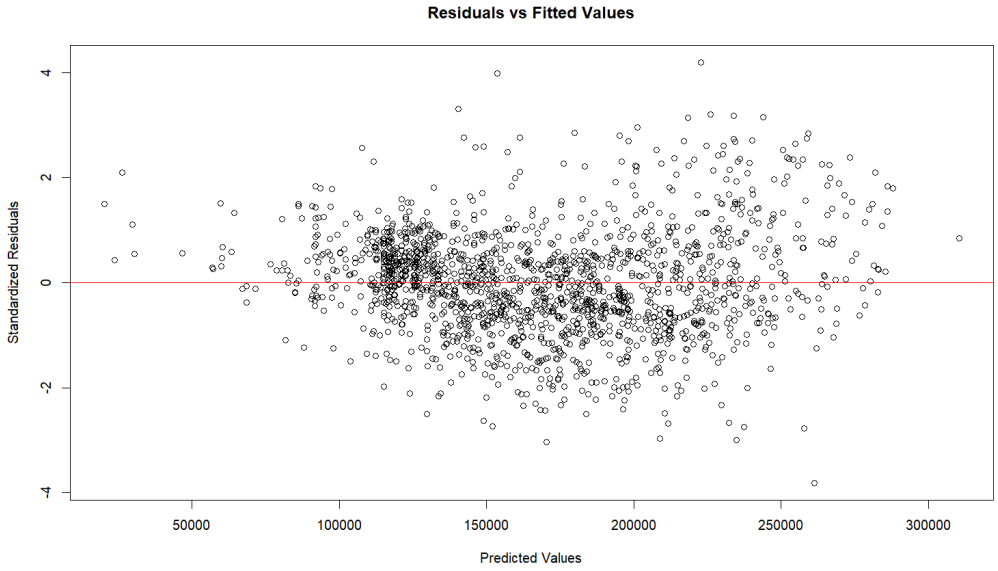
- For each coefficient:
  - Null Hypothesis $H0$: The coefficient is equal to zero (no effect).
  - Alternative Hypothesis Ha: The coefficient is not equal to zero (there is an effect).
- Given the p-values (<2.2e-16) for both `GrLivArea` and `OverallQual`, we reject the null hypotheses for both coefficients, meaning both predictors significantly influence `SalePrice`.
- For the overall model:
  - Null Hypothesis: All predictors' coefficients are equal to zero.
  - Alternative Hypothesis: At least one predictor's coefficient is not equal to zero.
- The F-statistic is 2549, with a p-value of <2.2e-16, allowing us to reject the null hypothesis. This means the model is significant.



Histogram of Standardized Residuals:

- This histogram shows the distribution of the standardized residuals. Ideally, for linear regression assumptions, we'd like the residuals to be approximately normally distributed. This would appear as a bell-shaped (or Gaussian) curve.

- From the histogram, the residuals appear somewhat normally distributed, with a peak in the middle. However, there might be some slight skewness to the left, indicating a potential deviation from perfect normality.

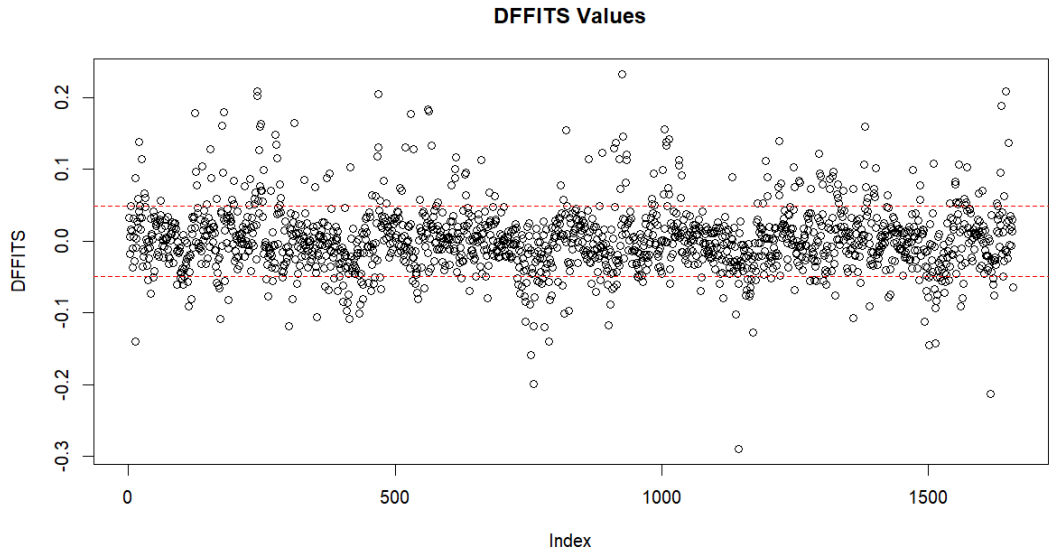


Residuals vs Fitted Values Plot: This plot is used to detect non-linearity and heteroscedasticity.

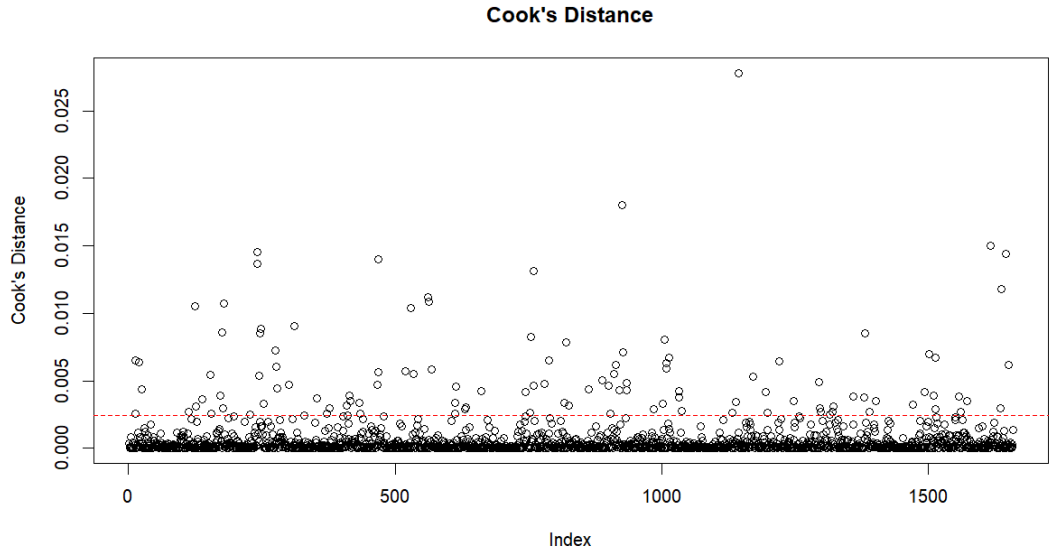
- The red horizontal line at standardized residual value of 0 is the reference line where residuals should be randomly scattered around if the model's assumptions hold.
- Observing the plot, we can see that there's a slight funnel shape where residuals are more spread out for lower and higher predicted values compared to the middle range. This suggests potential heteroscedasticity—meaning the variability of residuals is not constant across all levels of fitted values. In ideal circumstances, we would want to see a random scatter of points around the red line without any discernible pattern.

There's also a concentration of residuals around a particular range of fitted values. This could be due to specific characteristics of the data or perhaps a group of similar observations. This should be investigated further.

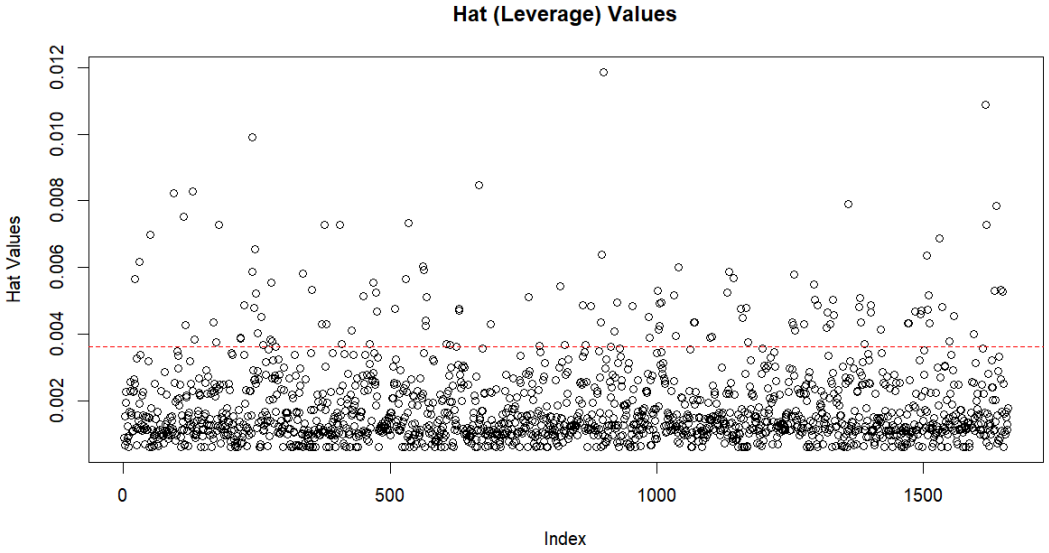
Conclusion: The residuals show a minor deviation from normality as indicated by the histogram. There's evidence of heteroscedasticity in the residuals as seen from the residuals vs. fitted values plot. This can violate the assumptions of linear regression, and in practice, might suggest that a transformation of the response variable or the use of weighted regression might be appropriate.



- Most of the observations are clustered around the zero, but there are a few points that lie outside the bands. These could be influential observations that might be affecting the fitted values.



- Most of the observations are near zero, but there are a few that have a larger Cook's distance, suggesting that they might be influential.



- Most of the data points lie below the threshold, indicating that most of the observations have low leverage.
- There are a few points above the threshold, particularly around the indices 0-500 and slightly beyond 1000. These points may be of interest because they have higher leverage values and could potentially influence the regression model.

g. The inclusion of `OverallQual` in Model_2 significantly improves the model's R-squared value, increasing from 0.5969 in Model_1 to 0.7548 in Model_2. This indicates that adding `OverallQual` as a predictor captures more of the variance in `SalePrice`.

- Both `GrLivArea` and `OverallQual` have highly significant p-values in Model_2, indicating that they are both statistically significant predictors of `SalePrice`.
- The coefficient of `GrLivArea` reduces from 99.915 in Model_1 to 58.930 in Model_2. This reduction suggests some collinearity between `GrLivArea` and `OverallQual`. However, both predictors remain significant in Model_2.
- The residual standard error decreases in Model_2 compared to Model_1, indicating that the model's predictions are more precise with the inclusion of `OverallQual`.
- Conclusion: Given the increase in R-squared, the reduction in residual standard error, and the significant p-values of both predictor variables in Model_2, it would be advisable to retain both `GrLivArea` and `OverallQual` as predictor variables for `SalePrice`.

3. Select any other continuous explanatory variable you wish. Fit a multiple regression model that uses 3 continuous explanatory (X) variables to predict Sale Price (Y). These three variables should be the explanatory variables from Model 2 plus your choice of an additional explanatory variable. Call this Model 3. To report the results for Model 3, you are to:

a. Report Model 3 in equation form and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here to Models 1 and 2?
b. Report and interpret R-squared value in the context of this problem. Calculate the difference in R-squared between Model 3 and Model 2. How would you interpret this difference? Does your variable of choice help to improve the model's explanatory ability?
c. Report the coefficient and ANOVA Tables for Model 3.
d. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret these hypothesis tests.
e. Check on the underlying assumptions. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.
f. Check on leverage, influence and outliers, and discuss any issues or concerns.
g. Based on this information, should you want to retain all three variables as predictor variables of Y? Discuss why or why not.

```
> ### Question 3
> # Fit the model
> Model_3 <- lm(SalePrice ~ GrLivArea+OverallQual+GarageArea, data = new_df)
>
> # Display the summary
> summary(Model_3)

Call:
lm(formula = SalePrice ~ GrLivArea + OverallQual + GarageArea,
    data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max
-109222  -14876    -823   14223  134193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -50720.439   3092.044  -16.40   <2e-16 ***
GrLivArea       53.555      1.871   28.62   <2e-16 ***
OverallQual  19046.294    696.569   27.34   <2e-16 ***
GarageArea      70.084      3.957   17.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24530 on 1655 degrees of freedom
Multiple R-squared:  0.7939,    Adjusted R-squared:  0.7935
F-statistic:  2124 on 3 and 1655 DF,  p-value: < 2.2e-16

> # ANOVA table
> anova_table <- anova(Model_3)
> print(anova_table)
Analysis of Variance Table

Response: SalePrice
             Df     Sum Sq    Mean Sq F value    Pr(>F)
GrLivArea     1 2.8826e+12 2.8826e+12 4792.48 < 2.2e-16 ***
OverallQual   1 7.6228e+11 7.6228e+11 1267.34 < 2.2e-16 ***
GarageArea    1 1.8864e+11 1.8864e+11  313.63 < 2.2e-16 ***
Residuals  1655 9.9545e+11 6.0148e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a. Model_3 in equation form:

$$SalePrice = -50720.439 + 53.555 * GrLivArea\} + 19046.294 * OverallQual\} + 70.084 * GarageArea$$

Interpretation of each coefficient:

- Intercept (-50720.44): When all the predictor variables (GrLivArea, OverallQual, and GarageArea) are zero, the predicted SalePrice is -$50,720.44. It's important to note that this interpretation is not practically meaningful because it's not feasible to have a house with zero living area, zero quality, and zero garage area.
- GrLivArea (53.56): For every one-unit increase in the GrLivArea, holding the OverallQual and GarageArea constant, the SalePrice is expected to increase by $53.56.
- OverallQual (19046.23): For every one-point increase in the OverallQual rating, holding GrLivArea and GarageArea constant, the SalePrice is expected to increase by $19,046.23.
- GarageArea (70.08): For every one-unit increase in the GarageArea, holding the GrLivArea and OverallQual constant, the SalePrice is expected to increase by $70.08.

Differences in coefficient interpretations compared to Models 1 and 2:

- In Model_1, the coefficient for `GrLivArea` was 99.915. In Model_3, it has reduced to 53.555. This indicates that when other variables like `OverallQual` and `GarageArea` are included, the unique effect of `GrLivArea` on the `SalePrice` is less compared to when it's the sole predictor.
- In Model_2, the coefficient for `GrLivArea` was 58.930, and for `OverallQual`, it was 23286.617. In Model_3, these have changed to 53.555 and 19046.294, respectively. This indicates that the addition of `GarageArea` as a predictor has further adjusted the influence of the other two predictors.

In essence, as we add more predictors to the model, the coefficients of the existing predictors can change because the model is trying to account for the shared and unique variance each predictor brings to explaining the dependent variable.

b. R-squared interpretation: The R-squared value for Model_3 is 0.7939. This means that approximately 79.39% of the variance in the SalePrice is explained by the predictors GrLivArea, OverallQual, and GarageArea.

- Difference in R-squared between Model_3 and Model_2:
- $\Delta R^2 = R^2 Model\_3 - R^2 Model\_2 = 0.7939 - 0.7548 = 0.0391$

This difference in R-squared, 0.0391 or 3.91%, represents the additional variance in SalePrice explained by Model_3 compared to Model_2. In other words, by adding the variable GarageArea to the model, we are explaining an additional 3.91% of the variance in SalePrice.

Interpretation: The addition of GarageArea has improved the model's ability to explain the variation in SalePrice. The 3.91% increase in R-squared indicates that GarageArea has provided significant additional explanatory power.

c. Coefficient Table and Anova Tables are a part of the R output.

d. For each coefficient:

- Null Hypothesis $H0$: The coefficient is equal to zero (no effect).
- Alternative Hypothesis Ha: The coefficient is not equal to zero (there is an effect).

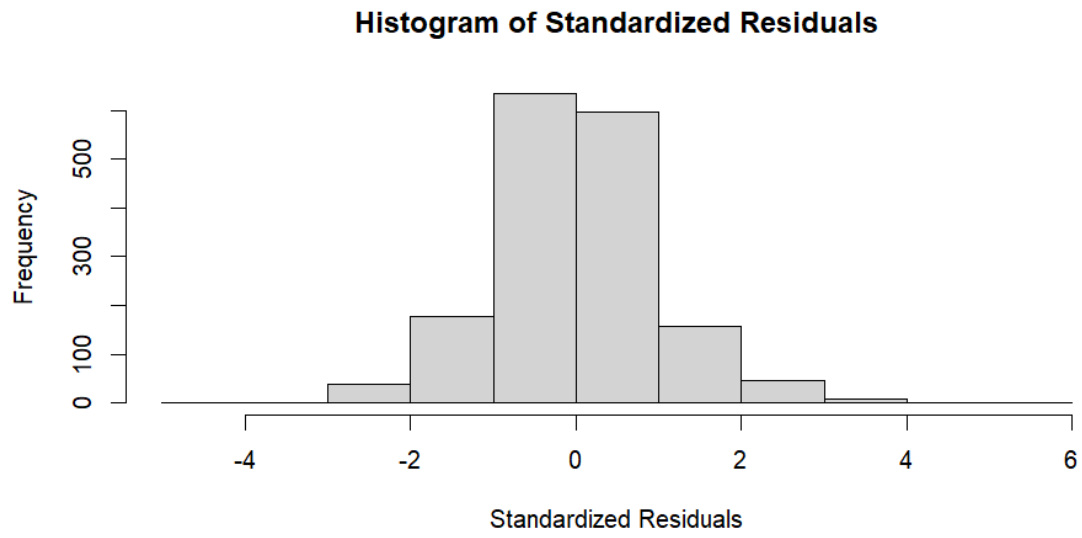
Based on the provided p-values (Pr(>|t|)) for each coefficient, all are significantly different from zero at the conventional 0.05 level (they all have p-values < 2.2e-16). Hence, we reject the null hypothesis for each predictor, indicating that GrLivArea, OverallQual, and GarageArea are all significant predictors of SalePrice.

Omnibus model hypothesis:

- $H0: \beta 1 = \beta 2 = \beta 3 = 0$
- $Ha: Atleast\ one\ \beta i \neq 0$

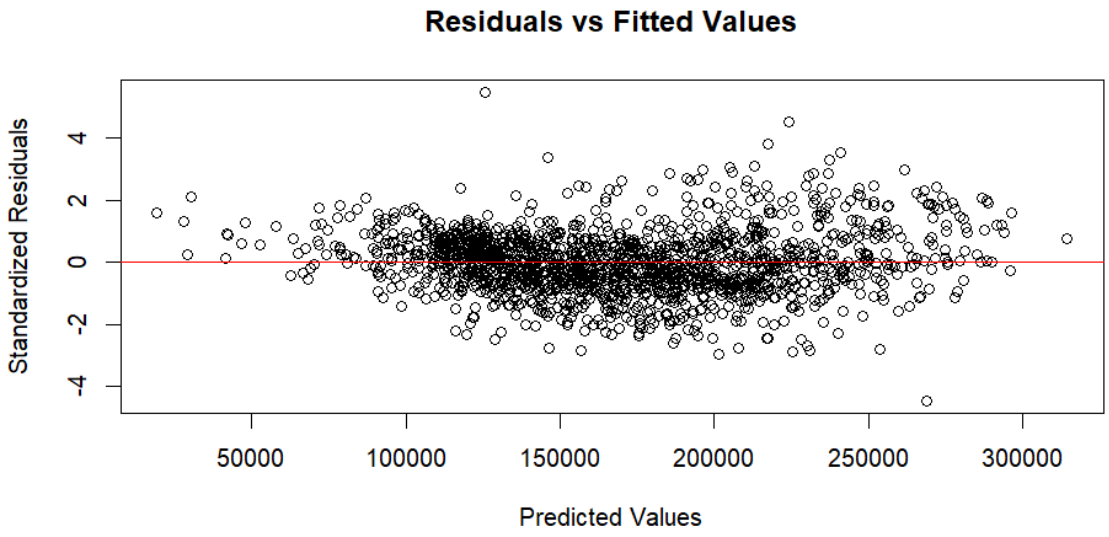
Given the ANOVA table, the F-statistic is extremely large, and the associated p-value is < 2.2e-16, indicating strong evidence against the null hypothesis. Hence, we reject the null hypothesis, suggesting that at least one of the predictors is useful in predicting SalePrice.

e.



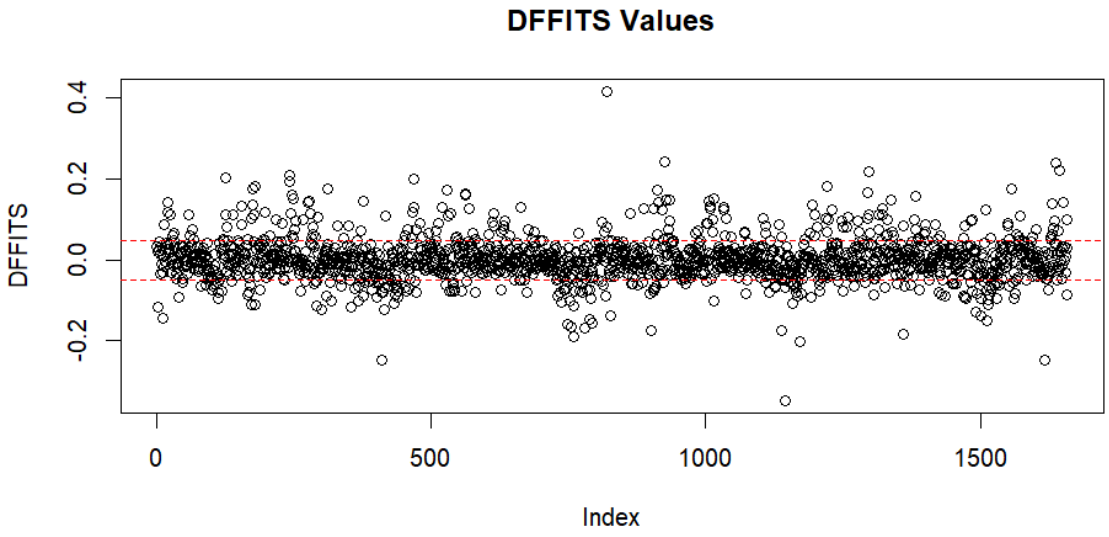
From the plot, the distribution appears to be roughly bell-shaped, centering around zero. This is a good sign. However, it also appears slightly right-skewed, with a noticeable tail to the right (positive values). This suggests there are a few larger positive residuals (outliers) where the model is underpredicting.

The left tail (negative values) seems shorter, suggesting that there are fewer instances where the model overpredicts significantly.

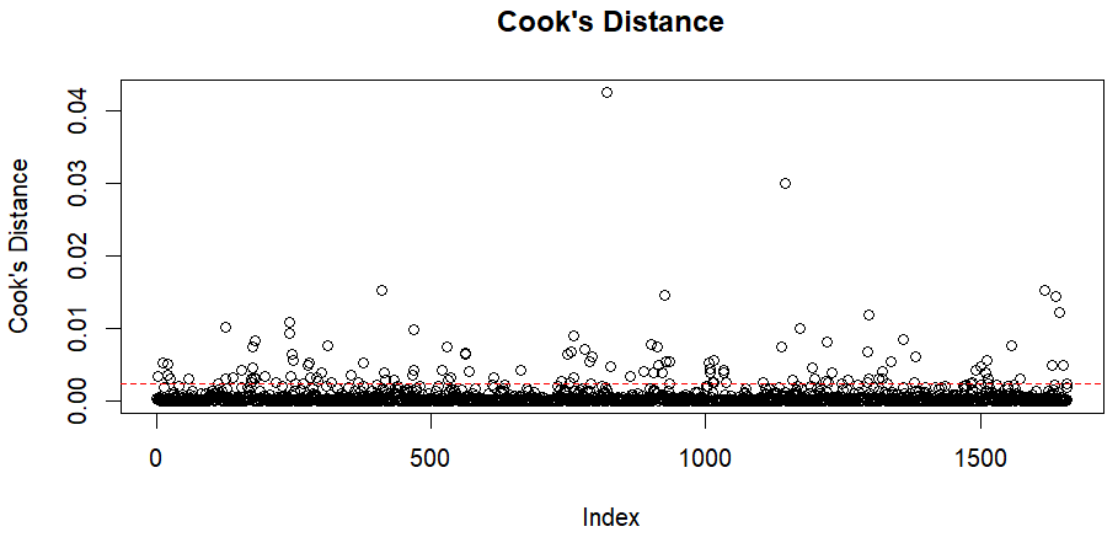


- There is a slight trend visible, especially towards the right side of the plot. As the predicted value increases, we see more pronounced positive residuals. This suggests a hint of heteroscedasticity since the spread of residuals isn't entirely consistent across all predicted values. There's also a slightly higher concentration of negative residuals in the lower predicted values range.
- Some distinct outliers can be observed, especially in the positive residual range, indicating specific observations where the model significantly underpredicts.

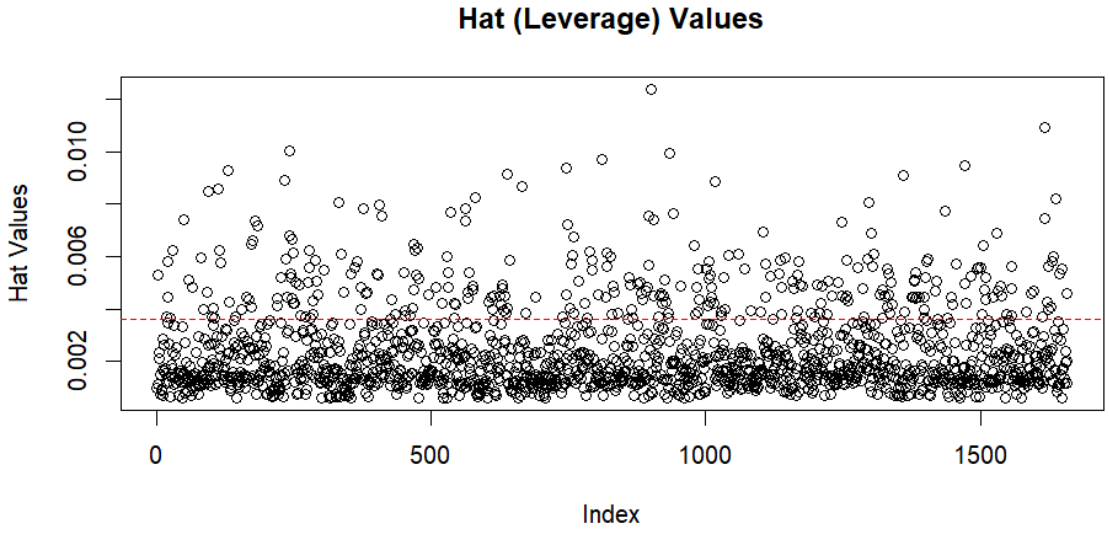
f.



From the plot, most of the DFFITS values are close to zero, which is a good sign. A few scattered points exist with higher positive or negative values, suggesting they might be influential.



Most of the points are hovering around the zero line, indicating they have little influence on the model. However, there are some points above the threshold (red dashed line) that could be considered influential.



In this plot, most of the hat values are well below the red dashed line (threshold). A few scattered points exceed this threshold, indicating they have high leverage.

Conclusions: While most of the data points are not influential and don't have high leverage, there are a few potential outliers that might be worth further investigation.

Observations with high DFFITS, Cook's Distance, or Hat values should be closely examined. It's important to see if they result from data entry errors, are genuine extreme cases, or if they reveal important patterns or structures in the data that the model doesn't currently capture.

Influential observations can unduly impact the regression results. If such observations are errors, they should be corrected or removed. If they are genuine, they might highlight limitations in the model, and one should consider if the model needs refinement.

g.

1. Significance of the Predictors: All three predictor variables (`GrLivArea`, `OverallQual`, and `GarageArea`) in `Model_3` have extremely low p-values (<< 0.05, denoted by ``), indicating they are statistically significant in predicting the response variable `SalePrice`.

2. Comparison between Models: When looking at the adjusted R-squared values (which account for the number of predictors in the model):

- `Model_2` (with only `GrLivArea` and `OverallQual`): Adjusted R-squared is 0.7545.
- `Model_3` (with `GrLivArea`, `OverallQual`, and `GarageArea`): Adjusted R-squared is 0.7935.

- The adjusted R-squared value increases when `GarageArea` is added to the model, suggesting that it provides additional explanatory power beyond what the other two variables offer.

3. Residual Standard Error: The residual standard error for `Model_3` (24530) is smaller than that for `Model_2` (26740). A smaller residual standard error indicates a better fit, suggesting that `Model_3` might be a better model compared to `Model_2`.

Considering the above points:

- All three predictors (`GrLivArea`, `OverallQual`, and `GarageArea`) are statistically significant.
- The inclusion of `GarageArea` increases the adjusted R-squared value, indicating an improvement in the proportion of the variance explained.
- The residual standard error decreases with the inclusion of `GarageArea`, indicating a better fit to the data.

Conclusion: Based on this information, it would be advisable to retain all three variables (`GrLivArea`, `OverallQual`, and `GarageArea`) as predictor variables of `SalePrice`. They collectively provide a better explanatory model for the response variable compared to a model excluding any of them.

4. Refit Model 3 using the Natural Log of SALEPRICE as the response variable. Call this Model 4. This is LOG base e, or LN() on your calculator. You'll have to find the appropriate function using R. Perform an analysis of goodness-of-fit to compare the Natural Log of SALEPRICE model, Model 4, to the original Model 3. Does the transformed model fit better? Provide evidence in your discussion. Discuss if the improvement of model fit justifies the use of the transformed response variable, Log(SALEPRRICE.

```
> # Fit Model_4
> Model_4 <- lm(log_SalePrice ~ GrLivArea + OverallQual + GarageArea, data = new_df)
>
> # Display the summary
> summary(Model_4)

Call:
lm(formula = log_SalePrice ~ GrLivArea + OverallQual + GarageArea,
    data = new_df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70075 -0.07625  0.00656  0.09896  0.73404

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.069e+01  1.847e-02  578.86   <2e-16 ***
GrLivArea   3.158e-04  1.118e-05   28.25   <2e-16 ***
OverallQual 1.123e-01  4.162e-03   26.99   <2e-16 ***
GarageArea  4.100e-04  2.364e-05   17.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1465 on 1655 degrees of freedom
Multiple R-squared:  0.7891,    Adjusted R-squared:  0.7887
F-statistic:  2064 on 3 and 1655 DF,  p-value: < 2.2e-16

> # ANOVA table
> anova_table <- anova(Model_4)
> print(anova_table)
Analysis of Variance Table

Response: log_SalePrice
             Df  Sum Sq Mean Sq F value    Pr(>F)
GrLivArea     1 100.045 100.045 4659.96 < 2.2e-16 ***
OverallQual   1  26.437  26.437 1231.41 < 2.2e-16 ***
GarageArea    1   6.457   6.457  300.76 < 2.2e-16 ***
Residuals  1655  35.531   0.021
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> # Extract Adjusted R-squared for Both Models
> adj_r2_model3 <- summary(Model_3)$adj.r.squared
> adj_r2_model4 <- summary(Model_4)$adj.r.squared
>
> # Extract Residual Standard Error for Both Models
> resid_se_model3 <- summary(Model_3)$sigma
> resid_se_model4 <- summary(Model_4)$sigma
>
> # Compare Models Using Akaike Information Criterion (AIC)
> aic_model3 <- AIC(Model_3)
> aic_model4 <- AIC(Model_4)
>
> # Print Comparison Results
> cat("Adjusted R-squared for Model 3:", adj_r2_model3, "\n")
Adjusted R-squared for Model 3: 0.7934843
> cat("Adjusted R-squared for Model 4:", adj_r2_model4, "\n")
Adjusted R-squared for Model 4: 0.7887126
> cat("Residual Standard Error for Model 3:", resid_se_model3, "\n")
Residual Standard Error for Model 3: 24525.13
> cat("Residual Standard Error for Model 4:", resid_se_model4, "\n")
Residual Standard Error for Model 4: 0.146523
> cat("AIC for Model 3:", aic_model3, "\n")
AIC for Model 3: 38250.56
> cat("AIC for Model 4:", aic_model4, "\n")
AIC for Model 4: -1658.427
```

Comparing Model_3 and Model_4

1. Adjusted R-squared:
   - Model 3: 0.7934843
   - Model 4: 0.7887126

The adjusted R-squared for Model 3 is slightly higher than that of Model 4. This suggests that Model 3 explains a marginally higher proportion of the variance in the response variable compared to Model 4.

2. Residual Standard Error:

- Model 3: 24525.13
- Model 4: 0.146523
-

While it might seem tempting to compare these directly, we must remember that these are not directly comparable because they're in different units (due to the transformation in Model 4). Model 3's residuals are in terms of the SALEPRICE, whereas Model 4's residuals are in terms of the natural logarithm of SALEPRICE.

3. Akaike Information Criterion (AIC):

- Model 3: 38250.56
- Model 4: -1658.427
-

A lower AIC value indicates a better fit, all else being equal. Based on this metric alone, Model 4 (with the natural logarithm of SALEPRICE) is vastly superior to Model 3. This suggests that, while the proportion of variance explained may be similar between the models, Model 4 is more parsimonious (i.e., it achieves a similar fit with fewer parameters or in a more efficient manner).

Discussion: While Model 3 has a slightly higher adjusted R-squared, the drastic improvement in AIC for Model 4 suggests that the transformed model (Model 4) is a better fit. AIC balances the fit of the model with the complexity, and a much lower AIC value for Model 4 suggests it's a more efficient model, despite the slight drop in R-squared.

The improvement in AIC for Model 4 is quite significant and this often justifies the use of a transformed response variable, especially if the goal is to achieve a balance between goodness of fit and model simplicity. Additionally, when modeling prices or other metrics that can vary over a wide range, a log transformation can help in stabilizing variances and making the model more robust.

In conclusion, the evidence suggests that the transformed model (Model 4) is a better choice, given the significant improvement in AIC. The slight drop in adjusted R-squared is outweighed by the benefits the transformation brings, making Model 4 more appropriate for predictions and inferences.

5. For either Model 3 or Model 4, your choice, identify the influential, high leverage, or outlier data points. Remove these data points from the dataset, then refit the model after removing the influential points. How many influential points did you find & remove? When you refitted the model, did the model improve? Comment on whether or not you find the improvement of model fit justifies the potential for the modeler biasing the result by removing potentially legitimate data points.

Approach for removing influential datapoints:

1. Calculate DFFITS, Cook's Distance, and Leverage values.
2. Identify data points that exceed the thresholds.
   - DFFITS: A common threshold is $\pm 2/\sqrt{n}$ where n is the number of observations.
   - Cook's Distance: A common rule of thumb is that if $Cook's\ Distance > \frac{4}{n-k-1}$ (where n is the number of observations and k is the number of predictors), then the observation is influential.
   - Leverage (Hat values): Observations with values greater than $\frac{2p}{n}$ (where p is the number of predictors and n is the number of observations)
3. Remove the points, exceeding the threshold, from the dataset.

```
> # 1. Calculate DFFITS, Cook's Distance, and Hat values
> dffits_values <- dffits(Model_3)
> cooks_d <- cooks.distance(Model_3)
> hat_values <- hatvalues(Model_3)
>
> # 2. Set thresholds
> threshold_dffits <- 2/sqrt(nrow(new_df))
> threshold_cooks <- 4/(nrow(new_df) - length(coef(Model_3)) - 1)
> threshold_hat <- 2*length(coef(Model_3))/nrow(new_df)
>
> # 3. Identify the influential data points for each threshold
> influential_dffits <- which(dffits_values > threshold_dffits)
> influential_cooks <- which(cooks_d > threshold_cooks)
> influential_hat <- which(hat_values > threshold_hat)
>
> # Print the number of influential data points by each threshold
> cat("Number of influential data points by DFFITS:", length(influential_dffits), "\n")
Number of influential data points by DFFITS: 201
> cat("Number of influential data points by Cook's Distance:", length(influential_cooks), "\n")
Number of influential data points by Cook's Distance: 120
> cat("Number of influential data points by Hat values:", length(influential_hat), "\n")
Number of influential data points by Hat values: 157
>
> # 4. Combine the influential points from all three thresholds
> all_influential_points <- unique(c(influential_dffits, influential_cooks, influential_hat))
>
> # Get the total number of unique influential points
> length(all_influential_points)
[1] 318
>
> # 5. Remove those points from the dataset
> new_df_cleaned <- new_df[-all_influential_points, ]
> dim(new_df_cleaned)
[1] 1341   10
>
> # Print the number of observations left after removal
> cat("Number of observations after removal of influential points:", nrow(new_df_cleaned), "\n")
Number of observations after removal of influential points: 1341
>
```

318 influential points were removed to create the Model_3_cleaned.

```
> # 6. Refit the model with the modified dataset
> Model_3_cleaned <- lm(SalePrice ~ GrLivArea + OverallQual + GarageArea, data = new_df_cleaned)
>
> # Display the summary of the refitted model
> summary(Model_3_cleaned)

Call:
lm(formula = SalePrice ~ GrLivArea + OverallQual + GarageArea,
    data = new_df_cleaned)

Residuals:
   Min     1Q Median     3Q    Max
-60920 -11381   1137  12989  46138

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -32681.703   2877.521  -11.36   <2e-16 ***
GrLivArea       49.611      1.733   28.64   <2e-16 ***
OverallQual  15876.020    668.390   23.75   <2e-16 ***
GarageArea      73.463      3.861   19.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17660 on 1337 degrees of freedom
Multiple R-squared:  0.8205,    Adjusted R-squared:  0.8201
F-statistic:  2038 on 3 and 1337 DF,  p-value: < 2.2e-16

>
>
> # ANOVA table
> anova_table <- anova(Model_3_cleaned)
> print(anova_table)
Analysis of Variance Table

Response: SalePrice
             Df     Sum Sq    Mean Sq F value    Pr(>F)
GrLivArea     1 1.5028e+12 1.5028e+12 4819.51 < 2.2e-16 ***
OverallQual   1 2.9058e+11 2.9058e+11  931.91 < 2.2e-16 ***
GarageArea    1 1.1289e+11 1.1289e+11  362.05 < 2.2e-16 ***
Residuals  1337 4.1689e+11 3.1181e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> # Compute AIC for the original model
> aic_original <- AIC(Model_3)
>
> # Compute AIC for the refitted model (after removal of influential points)
> aic_refitted <- AIC(Model_3_cleaned)
>
> # Print the AIC values
> cat("AIC for the original model:", aic_original, "\n")
AIC for the original model: 38250.56
> cat("AIC for the refitted model:", aic_refitted, "\n")
AIC for the refitted model: 30038.75
```

Summary for the model comparison:

Influential Point Analysis:

- DFFITS: Identified 201 influential data points.
- Cook's Distance: Identified 120 influential data points.
- Hat values: Identified 157 influential data points.
- Combining all three methods, a total of 318 unique influential points were identified.

Data After Removing Influential Points:

- The cleaned dataset has 1341 observations left after the removal of the influential points.

Model Comparison:

Original Model:	Refitted Model (after removing influential points):
AIC: 38250.56	AIC: 30038.75
Coefficients: • Intercept: -50720.439 (p < 0.001) • GrLivArea: 53.555 (p < 0.001) • OverallQual: 19046.294 (p < 0.001) • GarageArea: 70.084 (p < 0.001)	Coefficients: • Intercept: -32681.703 (p < 0.001) • GrLivArea: 49.611 (p < 0.001) • OverallQual: 15876.020 (p < 0.001) • GarageArea: 73.463 (p < 0.001)
Fit Statistics: Residual standard error: 24530 (based on 1655 degrees of freedom) • Multiple R-squared: 0.7939 • Adjusted R-squared: 0.7935 • F-statistic: 2124 (p < 0.001)	Fit Statistics: Residual standard error: 17660 (based on 1337 degrees of freedom) • Multiple R-squared: 0.8205 • Adjusted R-squared: 0.8201 • F-statistic: 2038 (p < 0.001)

Analysis:

Adjusted R-squared: The adjusted R-squared value has increased from 0.7935 to 0.8201. An increase in the adjusted R-squared value suggests that the model after removing the influential points can explain more of the variance in the dependent variable (SalePrice). This is a positive sign.

Residual standard error: The residual standard error has decreased from 24530 to 17660. A smaller residual standard error means that the residuals (i.e., the differences between the observed and predicted values) are smaller, implying a better fit to the data.

AIC Value: The AIC value has decreased from 38250.56 in the original model to 30038.75 in the refitted model. A smaller AIC suggests a better model fit, which indicates that the refitted model, after removing influential points, provides a more accurate representation of the underlying data.

Bias Concern: Removing influential points can potentially introduce bias. The decision to remove these points is based on their influence on the model, but one should consider if there's a legitimate reason these points exist in the data. They might be representing rare but possible scenarios or events.

Ethical Considerations: If a model is being used for decision-making, especially in areas with real-world consequences (like finance, healthcare, etc.), it's crucial to ensure the model is not omitting important information by removing these data points.

In summary, while the refitted model has shown improvement in its fit based on the metrics, including a better AIC, the decision to remove influential points should be made with caution. A careful balance between improving model metrics and ensuring the model's integrity and representativeness of the real world is crucial.

6. So far, we have fit a few models to predict SALEPRICE(Y). But, there are many other continuous variables in the data set, with many different possible combinations of variables that could be used in a regression model. You could use theory, or your background knowledge, to select variables for inclusion in a multiple regression model. Many modelers do this. It gives a nice place to start the search process. On the technical side, in this assignment, we know about correlation between variables and have been looking at change in R-squared when a new variable has been added to an existing model to isolate the explanatory contribution of that new variable. We have also been looking at hypothesis tests on the individual coefficients.

   Use the concept of Change in R-squared, plus anything else you wish, to put together a reasonable approach to find a good, comprehensive multiple regression model to predict SALEPRICE(Y). Any of the continuous variables can be considered fair game as explanatory variables. This can feel like an overwhelming task. You don't need to go overboard, or kill yourself, in doing this. We will learn about automated approaches to do this shortly. But, for now, I'd like you to think about how you would do this by hand.

   Use your approach to identify a good multiple regression model to predict SALEPRICE(Y) from the set of continuous explanatory variables available to you in the AMES dataset. For this task you need to:

   a. Explain your approach
   b. Report the model you determined and interpret the coefficients
   c. Report the coefficient and ANOVA tables.
   d. Report goodness of fit
   e. Check on underlying model assumptions.

Mixed Stepwise Regression Approach for the SalePrice Model:

- Our target variable for prediction is `SalePrice`. We consider an initial set of potential predictors: `OverallQual`, `GrLivArea`, `GarageCars`, `FullBath`, `GarageArea`, `YearBuilt`, `FirstFlrSF`, `TotRmsAbvGrd`, `TotalBsmtSF`, and `YearRemodel`.

Approach for this Model:

1. Starting Point: The process begins with two models: a null model (with no predictors) and the full model specified above.

2. Forward Selection: Starting with the null model, predictors are added one by one based on their significance (p-value <= 0.05) and their contribution to the model fit (often using metrics like adjusted R-squared or AIC).

3. Backward Elimination: After adding a predictor, the algorithm evaluates whether the model fit improves by removing any existing predictor. If any predictor's removal results in a better fit, it is excluded.

4. Iteration: The process continues, alternating between adding new predictors and evaluating the necessity of existing ones, until no further improvements to the model fit can be made.

Model Specific Advantages:

- Precision: Given the correlation values provided, predictors like `OverallQual` and `GrLivArea` have strong relationships with `SalePrice`. Using mixed stepwise regression allows us to evaluate the combined effect of these predictors in the presence of others.
- Efficiency: By starting with a comprehensive set of predictors and employing bidirectional elimination, we ensure that only the most pertinent predictors are retained in the final model, reducing potential overfitting.
- Insight: This approach can provide insights into the hierarchy of predictors' importance. For example, while `OverallQual` might be the most correlated with `SalePrice`, its significance in the combined model can be better evaluated using this method.

Conclusion: For the `SalePrice` prediction model, the mixed stepwise regression method offers a balanced and comprehensive approach. It systematically evaluates the contribution of each predictor, ensuring that the final model is both parsimonious and effective. This approach takes advantage of the individual strengths of both forward selection and backward elimination, providing a more nuanced and thorough analysis.

```
> # Initial null model with no predictors
> null_model <- lm(SalePrice ~ 1, data=new_df)
>
> # Perform mixed stepwise regression
> stepwise_model <- step(null_model,
+                        scope=list(lower=null_model, upper=full_model),
+                        direction="both",
+                        trace=1, # to show steps
+                        k=log(nrow(new_df)) # AIC adjustment; for BIC, k=log(nrow(new_df))
+ )
Start:  AIC=36911.06
SalePrice ~ 1

               Df  Sum of Sq        RSS   AIC
+ OverallQual   1 3.1879e+12 1.8936e+12 35249
+ GrLivArea     1 2.9996e+12 2.0819e+12 35410
+ GarageCars    1 2.1850e+12 2.8965e+12 35968
+ FullBath      1 2.1547e+12 2.9268e+12 35986
+ GarageArea    1 1.8926e+12 3.1889e+12 36131
+ YearBuilt     1 1.8382e+12 3.2433e+12 36159
+ FirstFlrSF    1 1.8072e+12 3.2743e+12 36175
+ TotRmsAbvGrd  1 1.7882e+12 3.2933e+12 36185
+ TotalBsmtSF   1 1.7178e+12 3.3637e+12 36221
+ YearRemodel   1 1.3884e+12 3.6931e+12 36379
<none>                       5.0815e+12 36911

Step:  AIC=35249.26
SalePrice ~ OverallQual

               Df  Sum of Sq        RSS   AIC
+ GrLivArea     1 6.2849e+11 1.2651e+12 34575
+ FirstFlrSF    1 4.2933e+11 1.4643e+12 34822
+ TotalBsmtSF   1 3.7251e+11 1.5211e+12 34886
+ GarageArea    1 3.3003e+11 1.5636e+12 34933
+ GarageCars    1 3.2865e+11 1.5650e+12 34934
+ FullBath      1 2.9319e+11 1.6004e+12 34972
+ TotRmsAbvGrd  1 2.5688e+11 1.6367e+12 35010
+ YearBuilt     1 2.5181e+11 1.6418e+12 35015
+ YearRemodel   1 1.3504e+11 1.7586e+12 35132
<none>                       1.8936e+12 35249
- OverallQual   1 3.1879e+12 5.0815e+12 36911
```

```
Step:  AIC=34574.68
SalePrice ~ OverallQual + GrLivArea

               Df  Sum of Sq        RSS   AIC
+ TotalBsmtSF   1 3.0704e+11 9.5807e+11 34112
+ YearBuilt     1 3.0126e+11 9.6385e+11 34122
+ FirstFlrSF    1 2.4305e+11 1.0221e+12 34221
+ GarageArea    1 2.0263e+11 1.0625e+12 34287
+ GarageCars    1 1.7494e+11 1.0902e+12 34330
+ YearRemodel   1 1.1205e+11 1.1531e+12 34425
+ FullBath      1 3.8923e+10 1.2262e+12 34529
+ TotRmsAbvGrd  1 1.7005e+10 1.2481e+12 34559
<none>                      1.2651e+12 34575
- GrLivArea     1 6.2849e+11 1.8936e+12 35249
- OverallQual   1 8.1674e+11 2.0819e+12 35410

Step:  AIC=34112.03
SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF

               Df  Sum of Sq        RSS   AIC
+ YearBuilt     1 1.8468e+11 7.7339e+11 33757
+ YearRemodel   1 1.1842e+11 8.3966e+11 33896
+ GarageArea    1 1.0573e+11 8.5234e+11 33922
+ GarageCars    1 1.0422e+11 8.5385e+11 33925
+ FullBath      1 2.7928e+10 9.3014e+11 34069
+ FirstFlrSF    1 8.2780e+09 9.4980e+11 34105
+ TotRmsAbvGrd  1 5.6800e+09 9.5239e+11 34109
<none>                      9.5807e+11 34112
- TotalBsmtSF   1 3.0704e+11 1.2651e+12 34575
- OverallQual   1 4.7612e+11 1.4342e+12 34787
- GrLivArea     1 5.6302e+11 1.5211e+12 34886


Step:  AIC=33757.35
SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF + YearBuilt

               Df  Sum of Sq        RSS   AIC
+ GarageArea    1 4.8043e+10 7.2535e+11 33656
+ GarageCars    1 3.9036e+10 7.3435e+11 33677
+ YearRemodel   1 3.8805e+10 7.3459e+11 33678
+ FirstFlrSF    1 1.4462e+10 7.5893e+11 33733
+ TotRmsAbvGrd  1 7.1720e+09 7.6622e+11 33749
<none>                      7.7339e+11 33757
+ FullBath      1 1.6910e+07 7.7337e+11 33765
- YearBuilt     1 1.8468e+11 9.5807e+11 34112
- TotalBsmtSF   1 1.9046e+11 9.6385e+11 34122
- OverallQual   1 2.1694e+11 9.9033e+11 34168
- GrLivArea     1 6.1050e+11 1.3839e+12 34734

Step:  AIC=33656.34
SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF + YearBuilt +
    GarageArea

               Df  Sum of Sq        RSS   AIC
+ YearRemodel   1 3.8445e+10 6.8690e+11 33572
+ FirstFlrSF    1 8.7068e+09 7.1664e+11 33643
+ TotRmsAbvGrd  1 5.3409e+09 7.2001e+11 33651
<none>                      7.2535e+11 33656
+ GarageCars    1 1.3859e+09 7.2396e+11 33661
+ FullBath      1 3.5414e+06 7.2534e+11 33664
- GarageArea    1 4.8043e+10 7.7339e+11 33757
- YearBuilt     1 1.2700e+11 8.5234e+11 33922
- TotalBsmtSF   1 1.5098e+11 8.7633e+11 33969
- OverallQual   1 1.8592e+11 9.1126e+11 34035
- GrLivArea     1 5.3325e+11 1.2586e+12 34581
```

```
Step:  AIC=33571.68
SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF + YearBuilt +
    GarageArea + YearRemodel

               Df  Sum of Sq        RSS   AIC
+ FirstFlrSF    1 1.0321e+10 6.7658e+11 33554
+ TotRmsAbvGrd  1 6.4003e+09 6.8050e+11 33563
<none>                       6.8690e+11 33572
+ FullBath      1 1.0674e+09 6.8583e+11 33576
+ GarageCars    1 9.3340e+08 6.8597e+11 33577
- YearRemodel   1 3.8445e+10 7.2535e+11 33656
- GarageArea    1 4.7684e+10 7.3459e+11 33678
- YearBuilt     1 6.7974e+10 7.5488e+11 33724
- OverallQual   1 1.4309e+11 8.2999e+11 33884
- TotalBsmtSF   1 1.6652e+11 8.5343e+11 33931
- GrLivArea     1 5.0894e+11 1.1958e+12 34502

Step:  AIC=33553.51
SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF + YearBuilt +
    GarageArea + YearRemodel + FirstFlrSF

               Df  Sum of Sq        RSS   AIC
+ TotRmsAbvGrd  1 5.2201e+09 6.7136e+11 33548
<none>                       6.7658e+11 33554
+ GarageCars    1 1.1442e+09 6.7544e+11 33558
+ FullBath      1 1.0888e+09 6.7549e+11 33558
- FirstFlrSF    1 1.0321e+10 6.8690e+11 33572
- TotalBsmtSF   1 3.2487e+10 7.0907e+11 33625
- YearRemodel   1 4.0059e+10 7.1664e+11 33643
- GarageArea    1 4.1511e+10 7.1809e+11 33647
- YearBuilt     1 7.1713e+10 7.4829e+11 33716
- OverallQual   1 1.4421e+11 8.2080e+11 33873
- GrLivArea     1 4.4724e+11 1.1238e+12 34404


Step:  AIC=33547.85
SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF + YearBuilt +
    GarageArea + YearRemodel + FirstFlrSF + TotRmsAbvGrd

               Df  Sum of Sq        RSS   AIC
<none>                       6.7136e+11 33548
+ GarageCars    1 1.5473e+09 6.6981e+11 33551
+ FullBath      1 8.2761e+08 6.7053e+11 33553
- TotRmsAbvGrd  1 5.2201e+09 6.7658e+11 33554
- FirstFlrSF    1 9.1406e+09 6.8050e+11 33563
- TotalBsmtSF   1 3.2632e+10 7.0399e+11 33621
- GarageArea    1 4.0166e+10 7.1153e+11 33639
- YearRemodel   1 4.0944e+10 7.1230e+11 33641
- YearBuilt     1 7.2237e+10 7.4360e+11 33713
- OverallQual   1 1.4450e+11 8.1587e+11 33870
- GrLivArea     1 2.5035e+11 9.2171e+11 34076
>
> # Display the final model summary
> summary(stepwise_model)

Call:
lm(formula = SalePrice ~ OverallQual + GrLivArea + TotalBsmtSF +
    YearBuilt + GarageArea + YearRemodel + FirstFlrSF + TotRmsAbvGrd,
    data = new_df)

Residuals:
   Min     1Q Median     3Q    Max
-73244 -12667   -830  10531 109255

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.195e+06  5.686e+04 -21.014  < 2e-16 ***
OverallQual   1.174e+04  6.172e+02  19.027  < 2e-16 ***
GrLivArea     5.749e+01  2.296e+00  25.044  < 2e-16 ***
TotalBsmtSF   2.377e+01  2.629e+00   9.042  < 2e-16 ***
YearBuilt     3.017e+02  2.243e+01  13.453  < 2e-16 ***
GarageArea    3.258e+01  3.248e+00  10.031  < 2e-16 ***
YearRemodel   2.930e+02  2.893e+01  10.128  < 2e-16 ***
FirstFlrSF    1.398e+01  2.922e+00   4.785 1.86e-06 ***
TotRmsAbvGrd -2.365e+03  6.540e+02  -3.616 0.000308 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 19980 on 1682 degrees of freedom
Multiple R-squared:  0.8679,    Adjusted R-squared:  0.8673
F-statistic:  1381 on 8 and 1682 DF,  p-value: < 2.2e-16
```

```
> # ANOVA table
> anova_table <- anova(stepwise_model)
> print(anova_table)
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq  F value    Pr(>F)
OverallQual     1 3.1879e+12 3.1879e+12 7986.827 < 2.2e-16 ***
GrLivArea       1 6.2849e+11 6.2849e+11 1574.591 < 2.2e-16 ***
TotalBsmtSF     1 3.0704e+11 3.0704e+11  769.233 < 2.2e-16 ***
YearBuilt       1 1.8468e+11 1.8468e+11  462.697 < 2.2e-16 ***
GarageArea      1 4.8043e+10 4.8043e+10  120.366 < 2.2e-16 ***
YearRemodel     1 3.8445e+10 3.8445e+10   96.318 < 2.2e-16 ***
FirstFlrSF      1 1.0321e+10 1.0321e+10   25.857 4.087e-07 ***
TotRmsAbvGrd    1 5.2201e+09 5.2201e+09   13.078 0.0003076 ***
Residuals    1682 6.7136e+11 3.9914e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA Table Interpretation:

- OverallQual: The variable is highly significant with a p-value of 2.2e-16, indicating a strong relationship with SalePrice.
- GrLivArea: This variable also has a p-value of 2.2e-16, emphasizing its significance in predicting SalePrice.
- TotalBsmtSF: Significant with a p-value of 2.2e-16.
- YearBuilt: With a p-value of 2.2e-16, it's clear that the year the house was built has a strong association with its SalePrice.
- GarageArea: It is significant with a p-value of 2.2e-16.
- YearRemodel: This variable, indicating when the house was last remodelled, is significant with a p-value of 2.2e-16.
- FirstFlrSF: The first-floor square footage is significant with a p-value of 1.86e-06.
- TotRmsAbvGrd: The total rooms above ground level, excluding bathrooms, is significant with a p-value of 0.000308.

```
> all_predictors <- c("OverallQual", "GrLivArea", "GarageCars", "FullBath",
+                     "GarageArea", "YearBuilt", "FirstFlrSF", "TotRmsAbvGrd",
+                     "TotalBsmtSF", "YearRemodel")
>
> included_predictors <- names(coef(stepwise_model))[-1] # -1 to exclude the intercept
>
> excluded_predictors <- setdiff(all_predictors, included_predictors)
> print(excluded_predictors)
[1] "GarageCars" "FullBath"
```

Exclusion of these variables suggests a few possibilities:

Multicollinearity: These variables might be highly correlated with one or more of the included predictors. In such cases, including both can distort the model's coefficients and make them hard to interpret. For instance, GarageCars (number of cars in the garage) might be highly correlated with GarageArea (size of the garage).

Statistical Significance: The stepwise regression uses p-values (among other metrics) to determine if a predictor should be included. A high p-value for a predictor suggests that it might not be statistically significant in predicting the response variable when other predictors are considered.

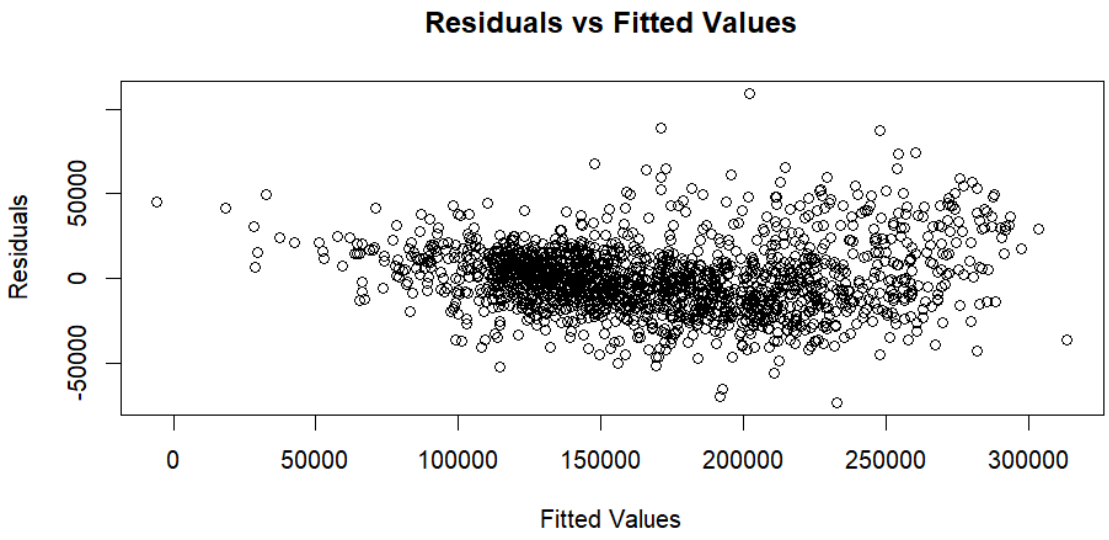
Model Simplicity: Stepwise regression aims to create a parsimonious model – that is, a model that explains the data with the fewest number of predictors. Thus, even if a predictor has some relationship with the response, it might be excluded in favor of a simpler model.

It's worth noting that just because a variable is excluded doesn't mean it's not important in other contexts or datasets. It only means that, given the current data and other included predictors, the stepwise algorithm deemed it redundant or not statistically significant for the prediction of SalePrice.

Goodness of Fit:

- Adjusted R-squared: 0.8673. The adjusted R-squared value provides an indication of the proportion of the variance in the dependent variable (SalePrice) that's explained by the model, after adjusting for the number of predictors. A value of 0.8673 implies that approximately 86.73% of the variability in SalePrice is accounted for by the predictors in the model.
- Residual standard error: 19980 on 1682 degrees of freedom. This metric gives the average difference between the observed values of SalePrice and the values predicted by the model. A lower value suggests a better fit. This is akin to the standard deviation of the residuals.
- F-statistic: 1381 on 8 and 1682 DF. The F-statistic tests whether at least one predictor in the model helps to explain the variability in SalePrice. A high F-statistic value (in this case, 1381) indicates that there's a relationship between at least one of the predictors and the response variable.
- P-value for the overall model: < 2.2e-16. This p-value tests the null hypothesis that all the regression coefficients are equal to zero, meaning none of the predictors contributes to the model. A p-value less than 0.05, and especially as low as < 2.2e-16, provides very strong evidence against this null hypothesis, indicating that the model is statistically significant.

These metrics suggest that the step-wise regression model fits the data quite well. The predictors in the model explain a significant portion of the variability in SalePrice.



The residuals exhibit a cloud-like distribution without a clear pattern, indicating the linearity assumption is likely met.

The variance seems consistent across fitted values, suggesting homoscedasticity. However, there may be minor variations, and it's always good to be cautious about potential heteroscedasticity.

```
> dwtest(stepwise_model)

        Durbin-Watson test

data:  stepwise_model
DW = 1.5905, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The DW statistic value is 1.5905. A DW statistic close to 2 indicates no autocorrelation, whereas a value below 2 suggests positive autocorrelation.

With a p-value < 2.2e-16, we reject the null hypothesis of no autocorrelation. This indicates significant positive autocorrelation in the residuals, which suggests that the model's residuals are not completely independent.

```
> vif(stepwise_model)
 OverallQual    GrLivArea  TotalBsmtSF    YearBuilt   GarageArea  YearRemodel   FirstFlrSF
    2.280893     4.075686     3.387148     1.814639     1.574288     1.519283     3.474543
TotRmsAbvGrd
    3.110400
```

Variance Inflation Factor (VIF):

VIF measures the extent of multicollinearity in regression. A VIF value of 1 suggests no multicollinearity, whereas a value above 10 is typically considered high, suggesting strong multicollinearity.

All VIF values are below the common threshold of 10, suggesting that multicollinearity is not severe in this model. The predictors might have some correlations, but they aren't high enough to distort the model estimates critically.

Interpretation:

The model seems to satisfy the assumptions of linearity and homoscedasticity based on the Residuals vs Fitted plot.

The presence of positive autocorrelation, as indicated by the Durbin-Watson test, may be a concern. It suggests that the model errors (residuals) aren't entirely independent. This might imply that an important predictor or temporal structure is missing from the model, or certain transformations might benefit the model.

Based on VIF values, multicollinearity does not seem to be a significant issue for the predictors in this model. However, always be cautious when interpreting the coefficients and their significance, especially for predictors with higher VIF values.

7. Please write a conclusion / reflection section that, at minimum, addresses the questions:
    - In what ways do variable transformation and outlier deletion impact the modelling process and the results?
    - Are these analytical activities a benefit or do they create additional difficulties?
    - Can you trust statistical hypothesis test results in regression?
    - What do you consider to be next steps in the modelling process?

Conclusion / Reflection:

1. Impact of Variable Transformation and Outlier Deletion on the Modeling Process and Results:
    - Variable Transformation: This often enhances the linearity of relationships between predictors and the response variable. For instance, log-transforming a predictor with a nonlinear relationship can improve the model fit. It can also help stabilize variances and make the model residuals more normally distributed, which is essential for hypothesis testing in regression.
    - Outlier Deletion: Outliers can distort model predictions and make parameter estimates unreliable. By identifying and removing them, we can achieve a more robust model. However, it's crucial to investigate outliers instead of blindly removing them, as they might be indicative of valuable insights or data issues.

2. Benefits vs. Difficulties of Variable Transformation and Outlier Deletion:
    - Benefits: Transforming variables and addressing outliers can lead to a more accurate and robust model. It can improve model assumptions like linearity, homoscedasticity, and normality of residuals. Moreover, addressing outliers can make the model more generalized, reducing the risk of overfitting.
    - Difficulties: Deciding on the correct transformation can be challenging. Moreover, transformations can make interpreting model coefficients less intuitive. For example, interpreting coefficients for log-transformed predictors might require exponentiation to translate changes in terms of percentage. Removing outliers might also lead to loss of vital information if not done thoughtfully.

3. Trusting Statistical Hypothesis Test Results in Regression: While statistical tests offer valuable insights into the significance of predictors, they have limitations:
    - If model assumptions (like linearity, independence, homoscedasticity, and normality of residuals) aren't met, hypothesis test results can be misleading.
    - In the presence of multicollinearity, individual predictor significance can be unreliable.
    - P-values, often used for hypothesis testing, can be influenced by sample size. With large samples, even trivial effects might seem significant. Conversely, with small samples, significant effects might be overlooked. Given these nuances, while hypothesis tests provide guidance, they shouldn't be the sole determinant in evaluating predictor importance.

4. Next Steps in the Modeling Process:

- Model Refinement: Consider interactions between predictors or higher-order terms if there's a theoretical basis.
- Validation: Split the dataset into training and testing sets to validate the model's performance on unseen data and ensure it's not overfitting.
- Diagnostic Checks: Continuously assess model assumptions using residual plots, variance inflation factors, and other diagnostic tools.
- Feature Engineering: Introduce new predictors based on domain knowledge, or consider variable clustering for dimension reduction.
- Alternate Models: Explore other regression techniques like ridge, lasso, or elastic net, especially if multicollinearity is a concern. For non-linear patterns, consider models like decision trees, random forests, or gradient boosting machines.
- In conclusion, the journey of regression modelling is iterative and reflective. Variable transformation and outlier management are tools that can enhance a model but must be used judiciously. Statistical tests, while useful, should be considered within the broader context of the data, domain knowledge, and model diagnostics.