Module 1 | Assignment: Regression Diagnostics with R | Revised

Trang Tran

CPS, Northeastern University

ALY6015 | Intermediate Analytics

Professor Steve Morin

Apr 26, 2023

Introduction

Ames housing dataset contains information from the Ames Assessorís Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. It consists of 2930 observations of residential homes, with 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers) describing various attributes of the properties, such as the number of bedrooms, bathrooms, and the size of the lot and garage. It presents a challenging regression problem due to a large number of features and the complex relationships between them.

Analysis

1. After performing exploratory data analysis using the [skimr] package, we can see that there are 43 character variables and 39 numeric variables in the summary that has a difference from the data dictionary. Looking deeper into each list of variable types, I realize that several variables (‘Alley’, ‘Fireplace.Qu’, ‘Pool.QC’, ‘Fence’, ‘Misc.Feature’) are having big proportions of missing values (> 20%) so I drop all these columns along with 2 observation identifiers out of this analysis. Besides, three categorical variables (‘MS.SubClass’, ‘Overall.Qual’, ‘Overall.Cond’) are disguised as numeric data. Therefore, I would be converting them into factors.

```
> skim(df) #data types, missing values, min, max, mean, sd, hist
── Data Summary ────────────────────────
                           Values
Name                       df
Number of rows             2930
Number of columns          75
_______________________
Column type frequency:
  character                38
  factor                   3
  numeric                  34
________________________
Group variables            None
```

Figure 1: Histogram of SalePrice

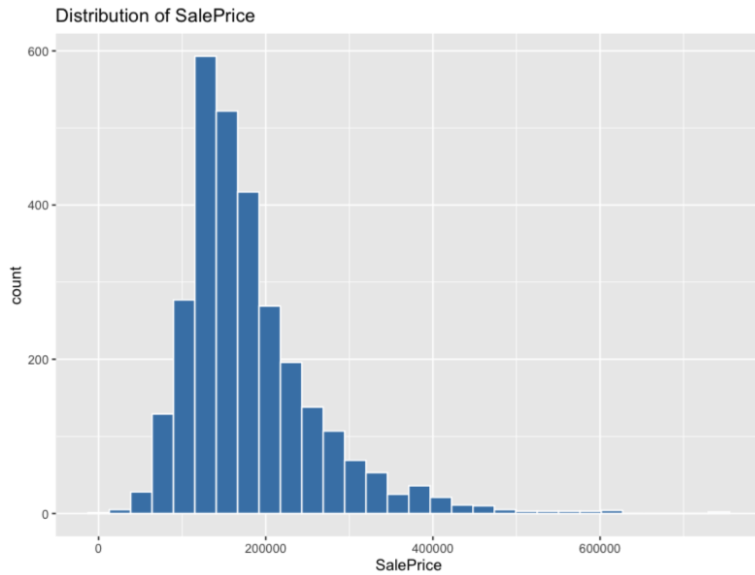


Figure 2: Boxplot of SalePrice by Neighborhood

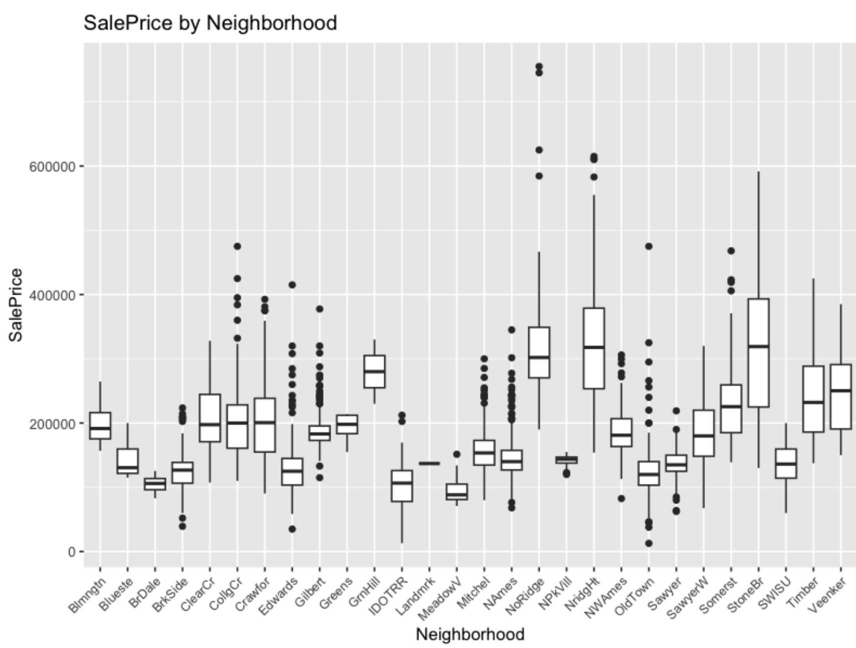
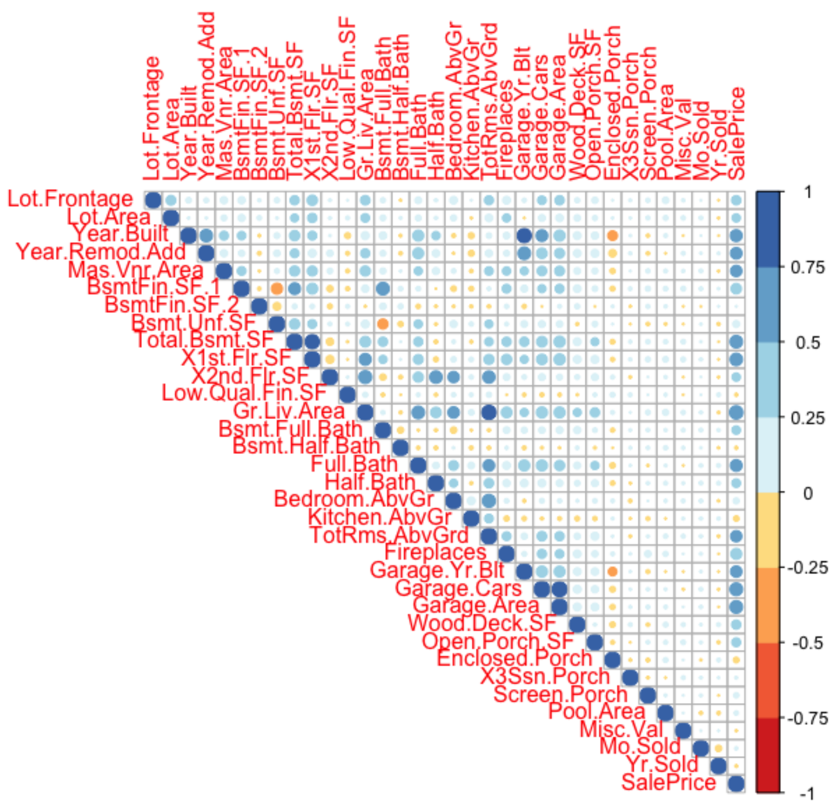


Figure 1 shows a right-skewed distribution of the sale price, with the highest price range from 100,000 to 200,000. The sale price above 400,000 accounts for a very small amount. Figure 2 shows the distribution of sale prices for each neighborhood in the Ames housing dataset. Looking at the plot, we can see that there are some neighborhoods with higher median sale prices than others, such as Northridge Heights and Stone Brook. There are also some neighborhoods with a wider range of sale prices, such as Edwards and Brookside. Overall, the boxplot of the sale price by neighborhood provides a useful visual summary of the variation in sale prices across different neighborhoods in the dataset.

2. In the next step, we prepare the dataset for modeling by imputing missing values with the variable's mean value in all numeric columns. Then we produce a plot of the correlation matrix as follows:

Figure 3: Correlation Plot

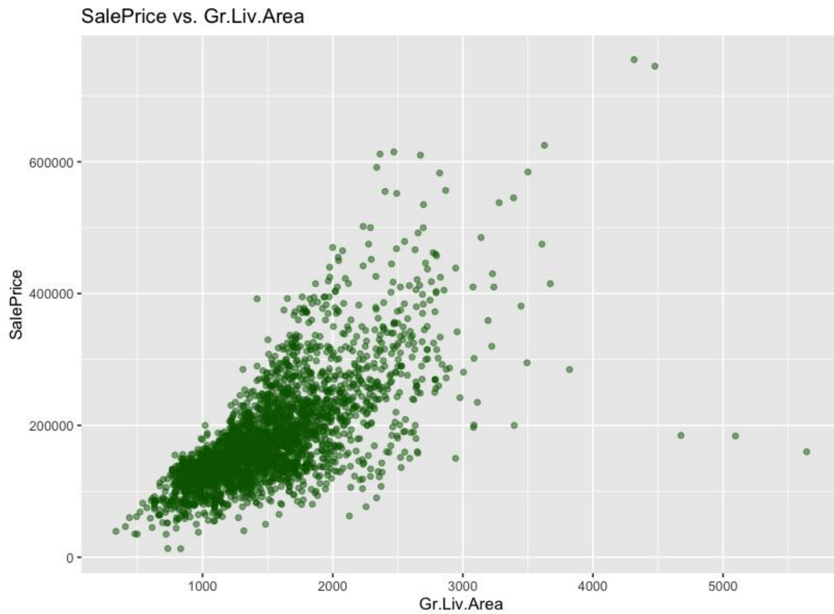


The plot depicts pretty strong positive relationships (>0.5) between ‘SalePrice’ and the variables ‘Gr.Liv.Area’, ‘Garage.Cars’ (Size of garage in car capacity), ‘Garage.Area’, and ‘Total.Bsmt.SF’. Meanwhile, ‘SalePrice’ also has slight negative correlations with variables like ‘Enclosed.Porch’ and ‘Kitchen.AbvGr’. Moreover, we capture that variable

‘Gr.Liv.Area’ has a strong correlation (>0.75) with ‘TotRms.AbvGrd’ (Total rooms above grade (does not include bathrooms).

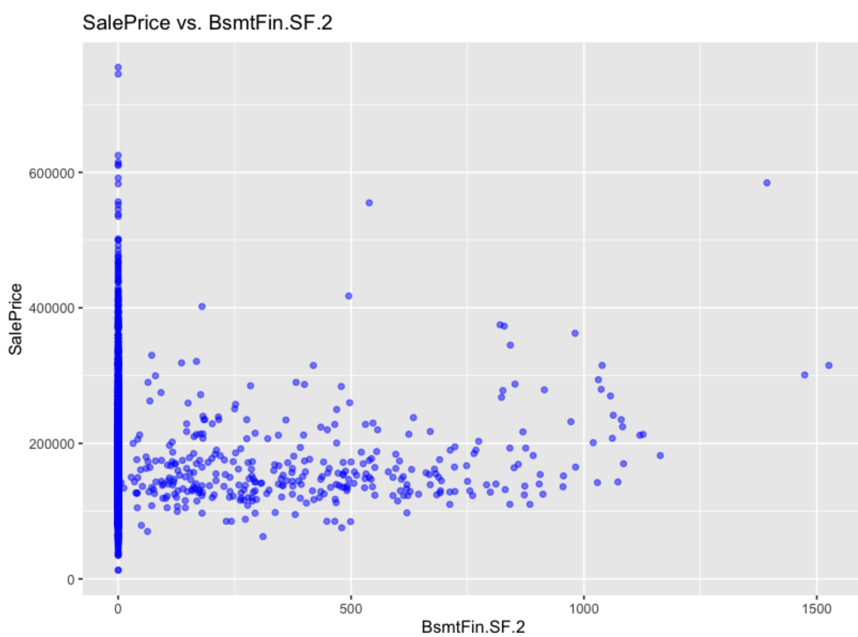
#Scatter plot of SalePrice vs. Gr.Liv.Area (highest correlation = 0.71)

Figure 4: Scatter plot of SalePrice vs. Gr.Liv.Area



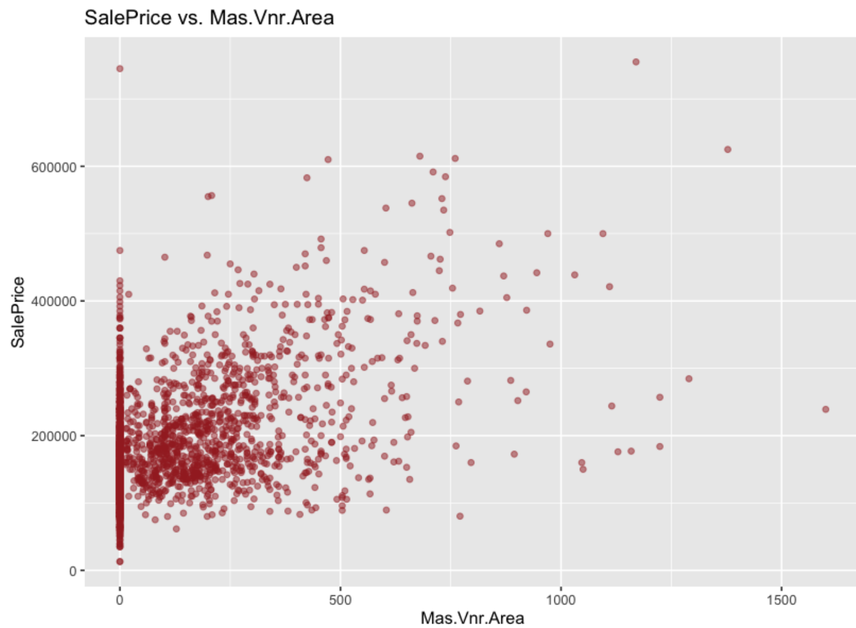
#Scatter plot of SalePrice vs. BsmtFin.SF.2 (lowest correlation = 0.01)

Figure 5: Scatter plot of SalePrice vs. BsmtFin.SF.2



#Scatter plot of SalePrice vs. Mas.Vnr.Area (correlation closest to 0.5)

Figure 6: Scatter plot of SalePrice vs. Mas.Vnr.Area (correlation closest to 0.5)



3. Fit a regression model:

SalePrice = -29593.64 + 68.86 * Gr.Liv.Area + 54.59 * Total.Bsmt.SF + 105.15 *Garage.Area

Figure 7: Regression Model 1

```
Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-681541  -19927     204   19841  266496

Coefficients:
                Estimate Std. Error t value            Pr(>|t|)
(Intercept)   -29593.644   2830.734  -10.45 <0.0000000000000002 ***
Gr.Liv.Area       68.862      1.966   35.02 <0.0000000000000002 ***
Total.Bsmt.SF     54.586      2.257   24.18 <0.0000000000000002 ***
Garage.Area      105.145      4.736   22.20 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45250 on 2926 degrees of freedom
Multiple R-squared:  0.6795,    Adjusted R-squared:  0.6791
F-statistic:  2068 on 3 and 2926 DF,  p-value: < 0.00000000000000022
```

- The symmetry in the residuals quartiles is good according to Figure 7.
- The p-value of each variable shows that we can reject the null hypothesis with 99% confidence.
- The model explains 68% of the variance of the data as the adjusted R-squared is ~ 0.68

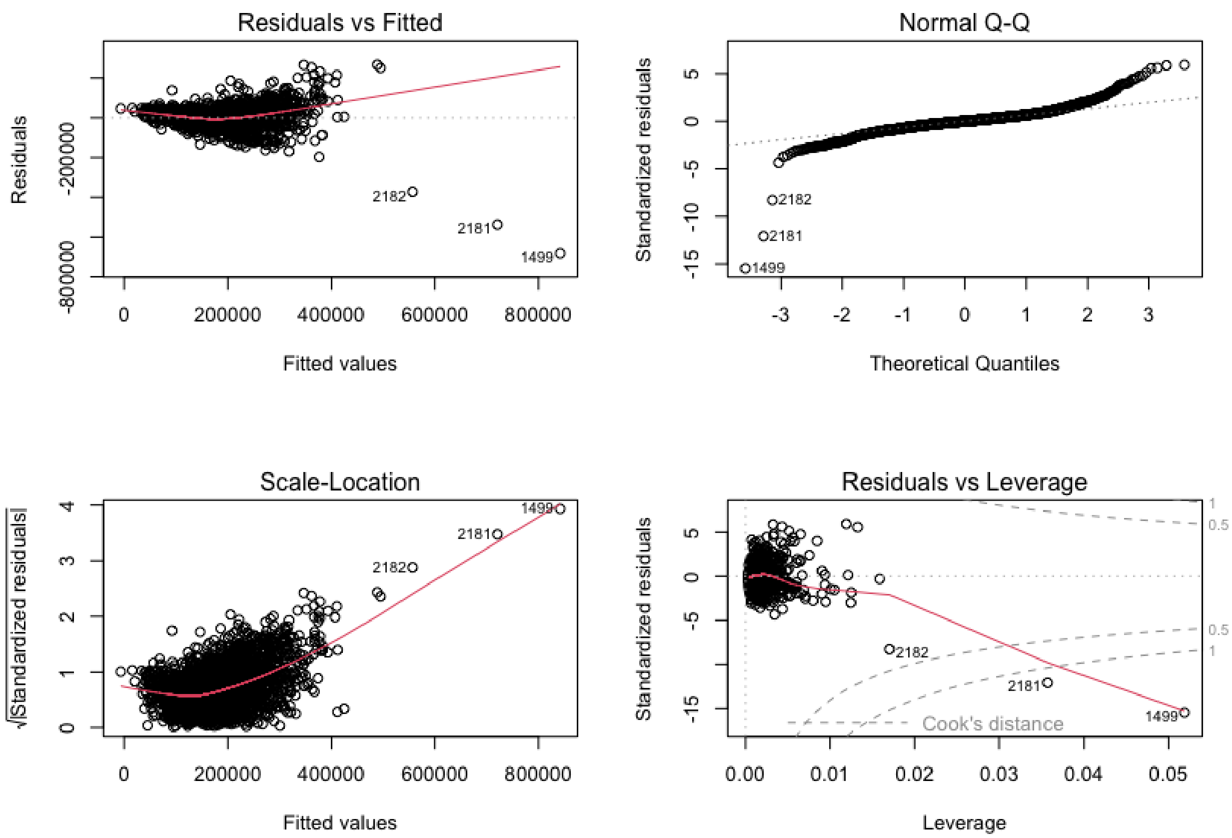


Figure 8: Linear Regression Diagnostic 1

Residuals vs Fitted: This plot shows the relationship between the residuals (difference between actual and predicted values) and the fitted values (predicted values) of the model. The loess line is approximately linear, so a linear may be a good one for this data. The residuals go around the line and the o value, which proves this plot looks good. Besides, we capture some influential points in rows 2182, 2181 and 1499.

Normal Q-Q: This plot shows the distribution of the residuals against a normal distribution. If the residuals follow a normal distribution, the points on the plot will fall on a straight line. This plot looks pretty good.

Scale-Location: This plot shows the square root of the absolute residuals against the fitted values. It is used to check for constant variance (homoscedasticity) of the residuals.

Residuals vs Leverage: This plot shows the leverage (influence of each data point on the model) against the standardized residuals. We can see three influential data points that exceed Cook's distance.

#Check the model above for multicollinearity

The VIF (Variance Inflation Factor) values for the three predictors in the model are all below 5, which is a common threshold for detecting multicollinearity. This suggests that there is no significant multicollinearity among the predictors in the model above. Therefore, no further action is needed to correct for multicollinearity in the model.

Figure 9: Check model for multicollinearity

```
> vif <- print(vif(fit))
  Gr.Liv.Area Total.Bsmt.SF   Garage.Area
     1.413121      1.414133      1.483363
```

#Check for outliers

Based on Figure 8, there are three influential points (rows 2182, 2181, and 1499) in the Residual vs Leverage plot that need to be removed to improve the model. After removing these influential points, I rerun model 1 and plot the 2nd diagnostic as below:

SalePrice = -46157.75+ 75.09* Gr.Liv.Area + 66.05* Total.Bsmt.SF + 96.2*Garage.Area

Figure 11: Rerun the regression model 1

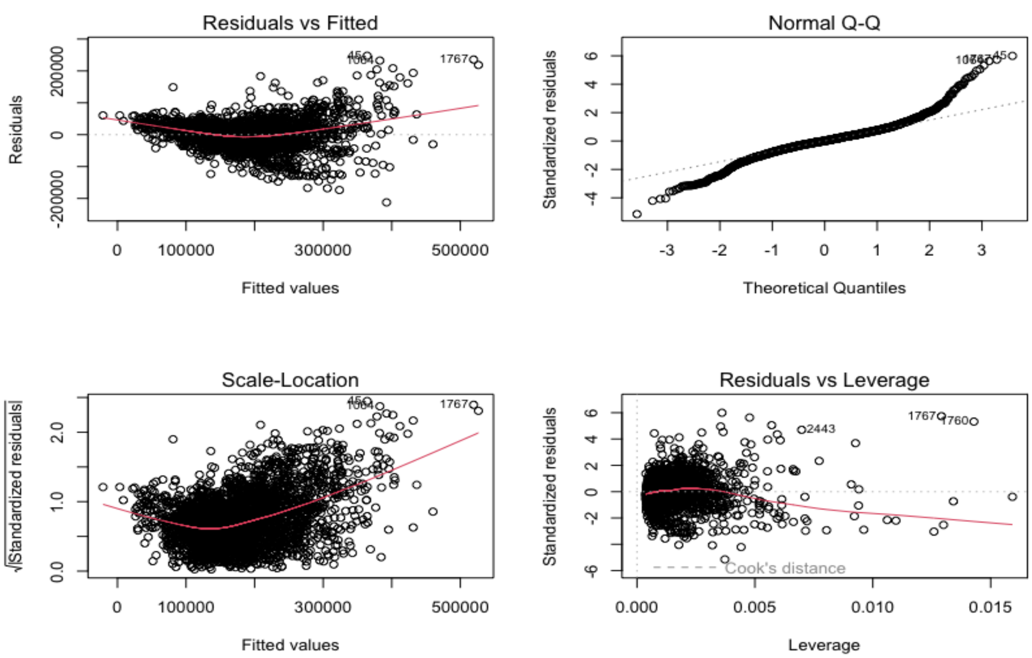
```
Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Total.Bsmt.SF + Garage.Area,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-212448  -20106     801   20798  247518

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)   -46157.752   2677.113  -17.24 <0.0000000000000002 ***
Gr.Liv.Area       75.090      1.817   41.32 <0.0000000000000002 ***
Total.Bsmt.SF     66.051      2.118   31.18 <0.0000000000000002 ***
Garage.Area       96.201      4.345   22.14 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41360 on 2923 degrees of freedom
Multiple R-squared:  0.7325,    Adjusted R-squared:  0.7323
F-statistic:  2668 on 3 and 2923 DF,  p-value: < 0.00000000000000022
```

Figure 10: Linear Regression Diagnostic 2



Looking at the 2nd diagnostic plot (Figure 11) and the summary of the rerun (Figure 10), the newly adjusted R-squared is higher than in the original model (0.73 compared to 0.68), and also other parameters/residuals plots look good, we can conclude that removing influential points has improved the model.

4. Identify the "best" model

 # Perform all-subsets regression with 16 variables

```
> regfit_2
Subset selection object
Call: regsubsets.formula(SalePrice ~ Lot.Area + Year.Built + Mas.Vnr.Area +
    BsmtFin.SF.1 + BsmtFin.SF.2 + Total.Bsmt.SF + Gr.Liv.Area +
    Bsmt.Full.Bath + Full.Bath + Bedroom.AbvGr + Kitchen.AbvGr +
    TotRms.AbvGrd + Garage.Area + Wood.Deck.SF + Open.Porch.SF +
    Pool.Area, data = df)
16 Variables  (and intercept)
```

 # Here, adjusted R2, CP, BIC, RSQ, and RSS tell us that the best model is the one with all 8 predictor variables.

Figure 12: Check for the preferred model

```
> # check the best number of predictor vraibles
> data.frame(Adj.R2 = which.max(fit2_sum$adjr2), CP = which.min(fit2_sum$cp),
+            BIC = which.min(fit2_sum$bic), RSQ = which.max(fit2_sum$rsq),
+            RSS = which.min(fit2_sum$rss))
  Adj.R2 CP BIC RSQ RSS
1      8  8   8   8   8
> # check the preferred model
> plot(regfit_2, scale = "adjr2") #8 variables
> coef(regfit_2, 8)
  (Intercept)     Year.Built   Mas.Vnr.Area   BsmtFin.SF.1  Total.Bsmt.SF    Gr.Liv.Area
-1001427.88262     530.77065       38.57695       21.41148       37.68201       94.05745
 Bedroom.AbvGr  Kitchen.AbvGr    Garage.Area
  -13741.90318   -37272.20930       43.58232
```

**The preferred model is:*

SalePrice = -1001427.88 + 530.77* Year.Built + 38.58* Mas.Vnr.Area + 21.41* BsmtFin.SF.1 + 37.68* Total.Bsmt.SF + 94.06* Gr.Liv.Area - 13741.9* Bedroom.AbvGr + 43.58* Garage.Area - 37272.2* Kitchen.AbvGr

#Below are the summary of the preferred model and the regression diagnostic plots:

The adjusted R-squared in Figure 13 shows a higher percentage (~83%) than the model I built above. It clearly shows the higher effectiveness of this model from this perspective. From the diagnostic plot (Figure 14), we can observe that all residuals look good in general, and no inferential point is shown in the Residuals vs Leverage plot.

Figure 14: Summary of the preferred model

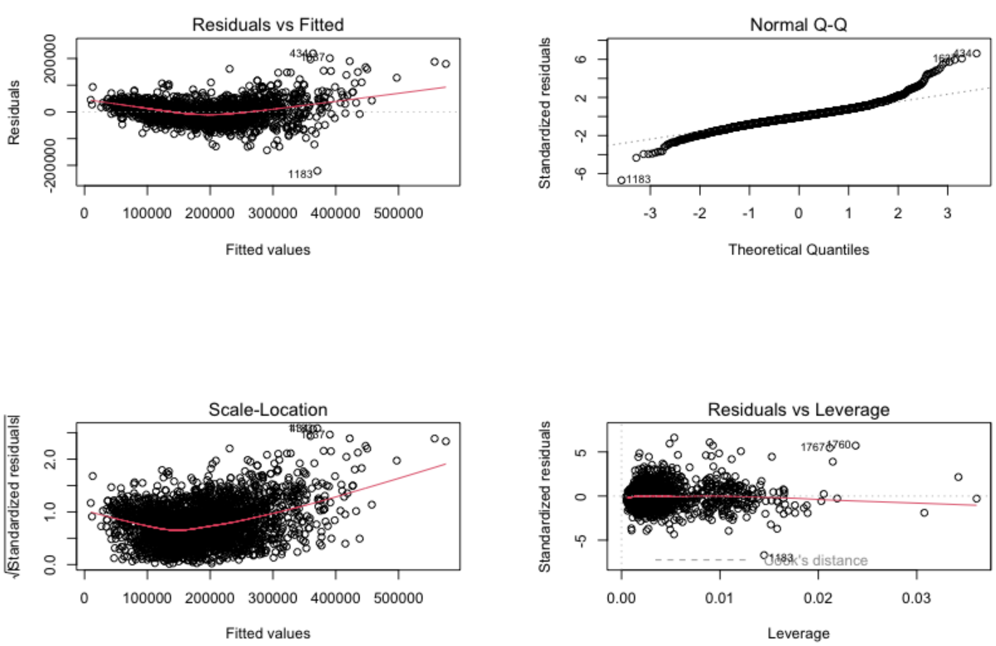
```
Call:
lm(formula = SalePrice ~ Year.Built + Mas.Vnr.Area + BsmtFin.SF.1 +
    Total.Bsmt.SF + Gr.Liv.Area + Bedroom.AbvGr + Kitchen.AbvGr +
    Garage.Area, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-220708  -18168   -1088   16999  219232

Coefficients:
                 Estimate   Std. Error t value             Pr(>|t|)
(Intercept)   -1001427.883    47556.073 -21.058 <0.0000000000000002 ***
Year.Built         530.771       24.292  21.850 <0.0000000000000002 ***
Mas.Vnr.Area        38.577        4.008   9.625 <0.0000000000000002 ***
BsmtFin.SF.1        21.411        1.655  12.941 <0.0000000000000002 ***
Total.Bsmt.SF       37.682        1.935  19.472 <0.0000000000000002 ***
Gr.Liv.Area         94.057        1.836  51.238 <0.0000000000000002 ***
Bedroom.AbvGr   -13741.903      940.420 -14.613 <0.0000000000000002 ***
Kitchen.AbvGr   -37272.209     2985.392 -12.485 <0.0000000000000002 ***
Garage.Area         43.582        3.770  11.561 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33200 on 2918 degrees of freedom
Multiple R-squared:  0.8279,    Adjusted R-squared:  0.8274
F-statistic:  1755 on 8 and 2918 DF,  p-value: < 0.00000000000000022
```

Figure 13: Regression Diagnostic Plots for the preferred model



The preferred model from step 13 and the model from step 12 differ totally in terms of the variables included and the coefficients assigned to each variable. In step 12, my model only includes three variables: Gr.Liv.Area, Total.Bsmt.SF, and Garage.Area. The preferred model includes these three variables and additional variables such as Year.Built, Mas.Vnr.Area, BsmtFin.SF.1, Kitchen.AbvGr, and Bedroom.AbvGr. The coefficients for Gr.Liv.Area, Total.Bsmt.SF and Garage.Area are also different from the previous model. It is likely that the preferred model from step 13 will have a better overall performance in terms of predicting the SalePrice of new data points, as it was selected based on a more rigorous statistical approach and includes a wider range of variables. However, when selecting a final model, I think that it is always important to consider the specific context and goals of the analysis, as well as the interpretability and practicality of the model.

Conclusion

In this dataset, we explored the Ames housing market data and built a linear regression model to predict the sale price of houses based on some features. We started by performing exploratory data analysis, including visualizations such as boxplots and histograms to understand the relationships between variables and the distribution of data. We also preprocessed the data by removing unqualified columns, imputing missing values, and handling categorical variables.

We then built a baseline linear regression model using three quantitative variables (Ground Living Area, Total Basement SF, and Garage Area) and evaluated its performance using various metrics such as coefficients, adjusted R-squared, residuals, and p-values. We also checked the model for assumptions such as linearity, normality, homoscedasticity, and multicollinearity. Moreover, we checked the model for outliers and influential observations and discussed the implications of removing them. Next, we attempted to find the best-fit model by adding variables and performing all-subsets regression methods.

Overall, we found that the model had a reasonable performance in predicting the sale price of houses based on the given features, but there is still room for improvement by considering additional variables and more sophisticated modeling techniques.

References

1. Daniel J. (March 25, 2023). *apply(), lapply(), sapply(), tapply() Function in R with Examples*. Guru99. Retrieved April 19, 2023. https://www.guru99.com/r-apply-sapply-tapply.html#3

2. Science Smith Edu. *Best Subset Selection*, Retrieved April 19, 2023.

http://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-r.html