Module 1 Assignment — Regression Diagnostics with R

Sourabh D Khot (ID 002754952)

College of Professional Studies, Northeastern University

ALY 6015: Intermediate Analytics CRN 81176

Professor Behzad Ahmadi

April 17, 2022

Introduction

In this project, my goal is to analyze the ‘AmesHousing.csv’ dataset and build a model to predict the sale price of a house given its several attributes. I will understand the recorded data in terms of numerical variables, and build a balanced model that is neither overfitted nor under fitted.

My questions are how to choose variables most influencing the sale price, evaluate if regression is the correct modeling technique for this dataset, perform diagnostics on a few linear regression models, treat missing values and outliers distorting the model, and decide which features should be selected for getting the best model. My goal is to build a model that is not too complex and still has a high coefficient of determination. I will employ correlation analysis, regression diagnostics, treating outliers, feature selection using all subsets techniques, and finally choose the best model.

Analysis

Exploratory Data Analysis (EDA)

1. Load the Ames housing dataset.

I loaded the 'AmesHousing.csv' dataset with 'UTF-8-BOM' encoding using the read.csv() function into a data frame 'df'. The variable 'Order' is the observation number assigned as row names and subsequently removed from the data frame.

2. Perform Exploratory Data Analysis and use descriptive statistics to describe the data.

The dataset has 2930 observations. 'Saleprice' is the outcome variable, with values ranging from 12.8K to 755K and having a median of 160K. The summary() function is used to analyze descriptive statistics of the remaining 81 variables. Descriptive statistics of five variables of potential interest are shown in Figure 1.

SalePrice	Yr.Sold	Lot.Area	Bedroom.AbvGr
Min. : 12789	Min. :2006	Min. : 1300	Min. :0.000
1st Qu.:129500	1st Qu.:2007	1st Qu.: 7440	1st Qu.:2.000
Median :160000	Median :2008	Median : 9436	Median :3.000
Mean :180796	Mean :2008	Mean : 10148	Mean :2.854
3rd Qu.:213500	3rd Qu.:2009	3rd Qu.: 11555	3rd Qu.:3.000
Max. :755000	Max. :2010	Max. :215245	Max. :8.000

Figure 1. Descriptive Statistics of Variables

Next, I have visually explored the variables and given three plots below in Figure 2. Sale Price has a longer right tail with possible outliers. Like Sale Price, the Lot Area also has a right-tail, but it is much longer.

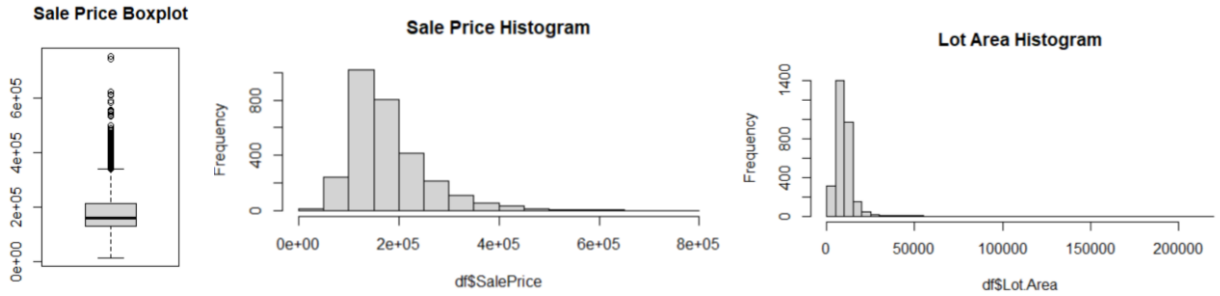

Figure 2. Histogram and Boxplots

I observed some variables having missing values. The continuous variables with missing values in brackets are Lot.Frontage (490), Mas.Vnr.Area (23), BsmtFin.SF.1 (1), BsmtFin.SF.2 (1), Bsmt.Unf.SF (1), Total.Bsmt.SF (1), and Garage.Area (1). Similarly, the discrete variables with missing values are Bsmt.Full.Bath (2), Bsmt.Half.Bath (2), Garage.Yr.Blt (2), and Garage.Cars (1).

3. Prepare the dataset for modeling by imputing missing values with the variable's mean value or any other value that you prefer.

I will create a separate data frame ‘df2’ with all missing values imputed. The continuous variable Lot.Frontage has the highest number of missing values of 490 and is described as ‘Linear feet of street connected to property’. It is closer to being symmetrical (low skewness of 1.5), and hence I will impute it with the mean of the available value using the mean() function. The other six continuous variables are all of area in square feet (length is normally distributed, so area may not be normal) and have more skewness. Hence mean may not be appropriate; I will instead use the median to replace their total of 28 missing values. The four discrete variables indicate count or year, whose mean may not be discrete; hence they will also be imputed using the median.

I will use only the continuous variables for regression analysis, creating a separate data frame ‘df3’ with the 20 continuous variables, including SalePrice.

Relationship between Variables

4. **Use the "cor()" function to produce a correlation matrix of the numeric values.**

A correlation matrix between SalePrice and 19 other continuous variables is created in the first table of Figure 3 below using cor() function. Most variables have a positive Pearson correlation coefficient, which indicates that SalePrice may increase with an increase in the variable. ‘Low Quality Finished Area’, ‘Enclosed Porch Area’ and ‘Miscellaneous Feature Value’ have negative correlation, and their increase may decrease SalePrice.

```
> cor(subset(df3,select=-c(
  SalePrice)), df3$SalePrice)
                      [,1]
Lot.Frontage     0.340751054
Lot.Area         0.266549220
Mas.Vnr.Area     0.502195977
BsmtFin.SF.1     0.432861846
BsmtFin.SF.2     0.006017568
Bsmt.Unf.SF      0.182895481
Total.Bsmt.SF    0.632163853
X1st.Flr.SF      0.621676063
X2nd.Flr.SF      0.269373357
Low.Qual.Fin.SF -0.037659765
Gr.Liv.Area      0.706779921
Garage.Area      0.640381073
Wood.Deck.SF     0.327143174
Open.Porch.SF    0.312950506
Enclosed.Porch  -0.128787442
X3Ssn.Porch      0.032224649
Screen.Porch     0.112151214
Pool.Area        0.068403247
Misc.Val        -0.015691463
```

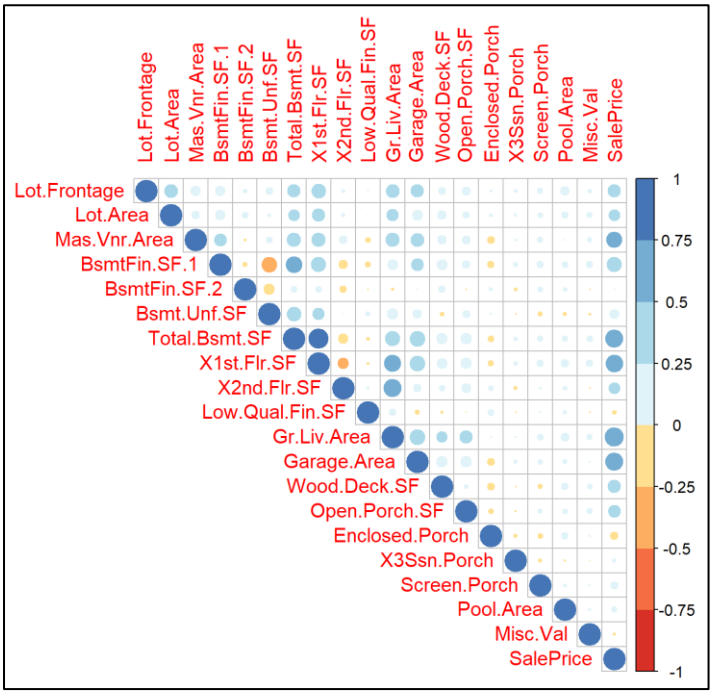

Figure 3. Correlation Matrix (against SalePrice) and Correlation Matrix Plot

5. **Produce a plot of the correlation matrix and explain how to interpret it.**

Plotted using corrplot() function, the second plot of Figure 3 above highlights the correlation matrix plot between all 20 continuous variables, and it helps quickly note variables of interesting correlation visually. Its interpretation can be made in two steps. The circle size between two variables indicates the degree of absolute correlation between them, and the color indicates the sign of correlation. Thus, darker blue shades depict a higher positive correlation, and darker red shades depict a higher negative correlation between the two variables.

With respect to 'Sale Price', we can deduce that 'Above ground living area' (Gr.Liv.Area), 'Garage Area' (Garage.Area) and 'Total Basement Area' (Total.Bsmt.SF) and 'First Floor Area' (X1st.Flr.SF) have the highest correlation (0.6+). Similarly, a correlation of about 0.5 is with the variable 'Masonry Veneer Area' (Mas.Vnr.Area), and the least correlation (<0.01) is with 'Basement Type 2 Finished Area' (BsmtFin.SF.2).

6. **Make a scatter plot for the X continuous variable with the highest correlation with SalePrice. Do the same for the X variable that has the lowest correlation with SalePrice. Finally, make a scatter plot between X and SalePrice with the correlation closest to 0.5. Interpret the scatter plots and describe how the patterns differ.**

I had identified in the previous step that for SalePrice, Gr.Liv.Area has the highest correlation, BsmtFin.SF.2 has the lowest correlation, and Mas.Vnr.Area has a nearly 0.5 correlation. Using the scatterplot() function, I have plotted scatterplots of SalePrice vs. each of these variables in Figure 4 below.

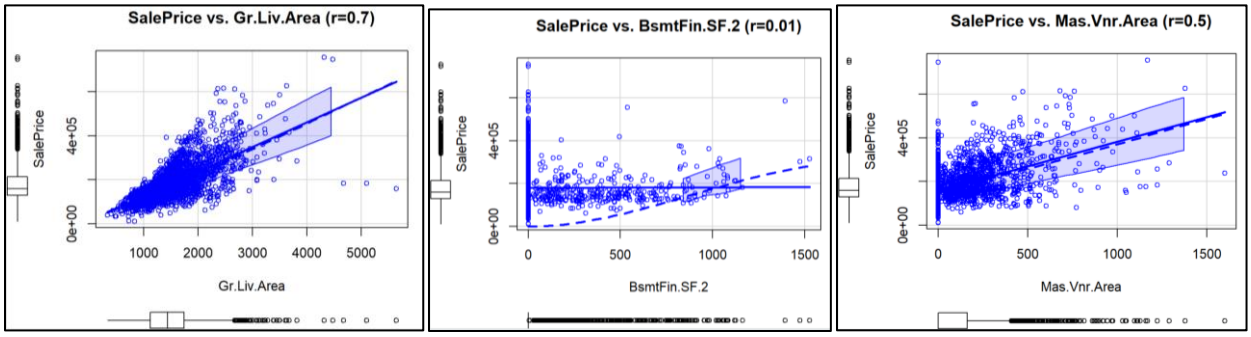

Figure 4. Scatter Plots with Variables of highest, lowest and ~0.5 and correlation

In the first scatter plot of the highest correlation variables, we see that as x_1 (Gr.Liv.Area) increases, y (SalePrice) increases in general, with low variance at lower values of x1. For higher values of x_1, the variance of y increases and needs to be inspected.

In the second plot between variables of nearly zero correlation, for any value of x_2 (BsmtFin.SF.2), the value of y is around the same value with similar variance. We can deduce that x_2 doesn't impact the value of y.

The third scatterplot has 0.5 correlation between variables, and y shows a slight increase with an increase in x_3 (Mas.Vnr.Area), but there is a high variance of y at every value of x_3.

Regression Model

7. Using at least 3 continuous variables, fit a regression model in R.

The top 3 variables having the highest correlation with SalePrice are Gr.Liv.Area, Garage.Area and Total.Bsmt.SF. Using lm() function, I have used these 3 predictor variables and fitted a multiple linear regression model against SalePrice as the outcome variable.

8. Report the model in equation form and interpret each coefficient of the model in the context of this problem.

A summary of the regression model just fitted is shown in Figure 5. The equation form of this regression model can be written as below:

$$\boldsymbol{SalePrice = -2959.6 + 68.9 * Gr.Liv.Area + 105.1 * Garage.Area + 54.6 * Total.Bsmt.SD}$$

```
Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Garage.Area + Total.Bsmt.SF,
    data = df3)

Residuals:
    Min      1Q  Median      3Q     Max
-681560  -19928     205   19844  266499

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -29590.606   2830.632  -10.45   <2e-16 ***
Gr.Liv.Area       68.858      1.966   35.02   <2e-16 ***
Garage.Area      105.133      4.736   22.20   <2e-16 ***
Total.Bsmt.SF     54.595      2.257   24.19   <2e-16 ***
---
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:  0.6792
F-statistic:  2068 on 3 and 2926 DF,  p-value: < 2.2e-16
```

Figure 5. Initial Regression Model Summary

The coefficient of a predictor variable can be interpreted as the coefficient times increase in the outcome variable for every unit increase in the predictor variable, keeping other variables constant. Thus, a square foot increase in ‘Above ground living area’ causes $68.9 increase in the ‘Sale Price’, assuming other variables are constant. Similarly, an only increase in ‘Garage Area’ by one square foot will increase the ‘Sale Price’ by $105.1. However, it should be noted that the range of ‘Garage Area’ should be within the range observed and used to train the model so that the results are realistic. Lastly, increasing one square foot of ‘Total Basement Area’ with no other change will increase ‘Sale Price’ by $54.6.

Regression Diagnostics

9. Use the "plot()" function to plot your regression model. Interpret the four graphs that are produced.

Plot() function is used in R to perform diagnostics as given in Figure 6. The first plot of 'Residuals vs Fitted' is to evaluate the linearity of the relationship. We see that the points are scattered randomly for a limited range of 10,000 and 30,000 for which linearity is good, with extreme observations beyond 50,000. The 'Normal Q-Q' plot of standardized residuals indicates normality of the residuals. We see that the plot goes along the central part's diagonal line but falls apart towards the ends, and hence the residuals are not perfectly normally distributed.

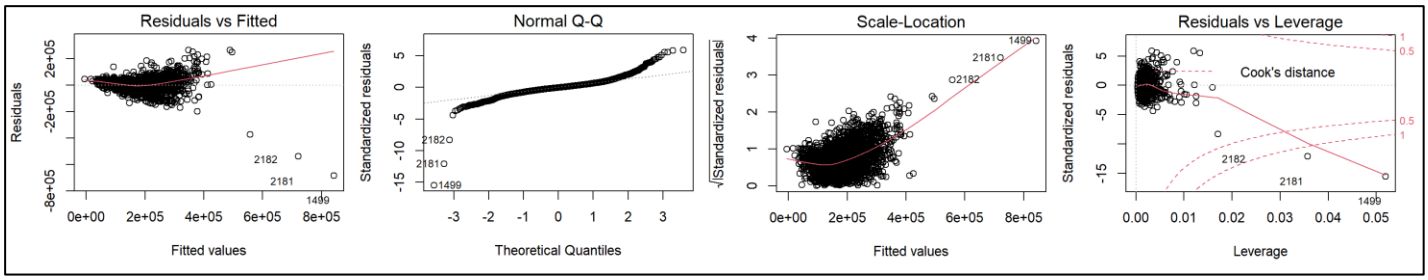

Figure 6. Regression Diagnostics of the Initial Model

The 'Scale-Location' plot points homoscedasticity and a good model should have a constant spread across a horizontal line. Our model does not have good homoscedasticity or constant variance of fitted values after 25000 as the central line starts turning diagonally. We may say that constant variance is present only for fitted values below 20,000.

Lastly, the 'Residuals vs. Leverage' plot highlights influential points. Most of the points are closer to the cook's distance line, except for three points on the lower right. The above plots and my further analysis show that issues in this linear regression model are primarily caused by observation numbers 2182, 2181, 1499, 1761, and 1768. Only these five observations have Gr.Liv.Area greater than 4000 and their removal will improve the model.

10. Check your model for multicollinearity and report your findings. What steps would you take to correct multicollinearity if it exists?

Multicollinearity is the inter-correlation between predictor variables, which is unsuitable for a linear regression model. We will use the vif() function, which calculates Variable Inflation Factors on our model, to check for multicollinearity.

All VIF values are between 1.413 and 1.483, which is much less than 5; hence there is no matter of concern. If there was high multicollinearity (above 5 or 10), we might have to remove some of the highly correlated variables to bring VIF under 5 and stabilize our model.

11. Check your model for outliers and report your findings. Should these observations be removed from the model?

I will detect outliers using outlierTest() function, which employs the Bonferroni Outlier Test. The findings are given in Figure 7.

```
> outlierTest(model = m)
       rstudent unadjusted p-value Bonferroni p
1499 -16.140129         3.3262e-56   9.7459e-53
2181 -12.393261         2.0462e-34   5.9953e-31
2182  -8.387252         7.6354e-17   2.2372e-13
1768   5.959769         2.8279e-09   8.2858e-06
45     5.900948         4.0291e-09   1.1805e-05
1064   5.636304         1.9028e-08   5.5753e-05
1761   5.592800         2.4406e-08   7.1509e-05
433    5.157597         2.6683e-07   7.8180e-04
434    4.806358         1.6146e-06   4.7309e-03
2446   4.776500         1.8716e-06   5.4837e-03
```

Figure 7. Outliers using Bonferroni Outlier Test

The function found 10 outliers with Bonferonni $p < 0.05$. We see that all influential points found in Step 9 / Figure 6, which were deviating from regression assumptions, are detected as outliers. These 10 outliers are just 0.3% of the dataset of 2930 observations and should be removed from the model so that assumptions of linear regression hold good and our model is accurate.

12. Attempt to correct any issues that you have discovered in your model. Did your changes improve the model, why or why not?

I removed the 10 outliers, formed a new data frame ‘df4’, and fitted a new model ‘m2’. The Regression Diagnostics for the cleaned model are shown in Figure 8.

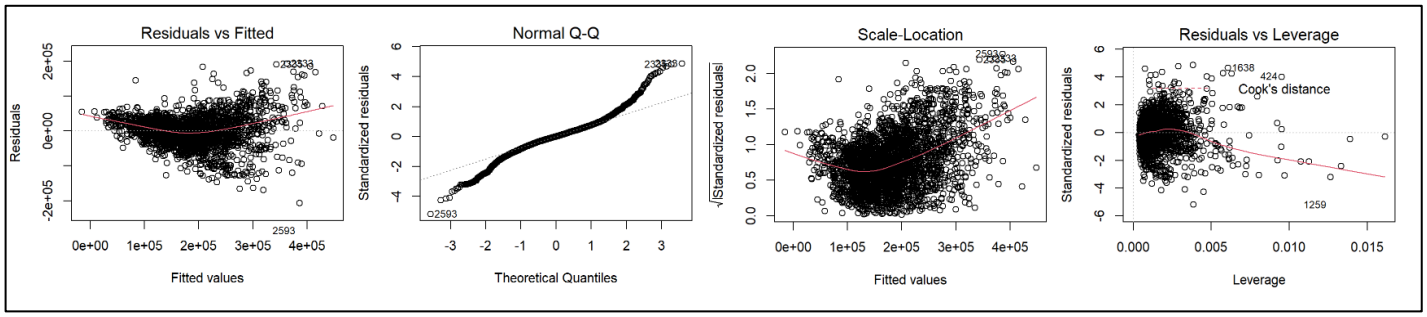

Figure 8. Regression Diagnostics of the Cleaned Model

We observe that the model has improved to a good extent, especially in terms of normality in the Q-Q plot. The outliers removed were common across plots to cause disturbance in the assumptions of linear regression, and hence removing them improved the model.

Feature Selection

13. Use the all subsets regression method to identify the "best" model. State the preferred model in equation form.

For the ‘all subsets regression’, in addition to the three predictor variables, I have considered two more continuous variables of ‘First Floor Area’ (X1st.Flr.SF) and ‘Masonry veneer area’ (Mas.Vnr.Area) since they have correlation above 0.5 as per Figure 3. The results are implemented using the regsubsets () function in Figure 9.

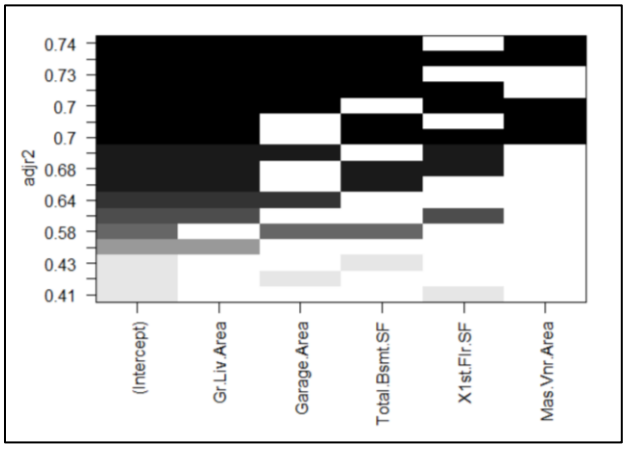

Figure 9. All Subsets Regression on 5 Predictor Variables

We see that Adjusted R^2 is highest when four predictor variables of Gr.Liv.Area, Garage.Area, Total.Bsmt.SF and Mas.Vnr.Area are included in the model for the outcome variable SalePrice. Using these four variables, I built another model 'm3', cleaned the data frame and got the following equation for the regression with the summary given in Figure 9.

$$\boldsymbol{SalePrice = -30230.6 + 67.3 * Gr.Liv.Area + 90.5 * Garage.Area}$$
$$\boldsymbol{+ 59.5 * Total.Bsmt.SF + 58.0 * Mas.Vnr.Area}$$

```
Call:
lm(formula = SalePrice ~ Gr.Liv.Area + Garage.Area + Total.Bsmt.SF +
    Mas.Vnr.Area, data = df4)

Residuals:
    Min      1Q  Median      3Q     Max
-215144  -19391     808   20431  200384

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -30230.593   2667.460  -11.33   <2e-16 ***
Gr.Liv.Area       67.267      1.768   38.05   <2e-16 ***
Garage.Area       90.529      4.140   21.87   <2e-16 ***
Total.Bsmt.SF     58.543      2.040   28.70   <2e-16 ***
Mas.Vnr.Area      58.031      4.672   12.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38900 on 2915 degrees of freedom
Multiple R-squared:  0.7424,    Adjusted R-squared:  0.7421
F-statistic:  2101 on 4 and 2915 DF,  p-value: < 2.2e-16
```

Figure 9. All Subsets Best Model Regression Model Summary

14. Compare the preferred model from step 13 with your model from step 12. How do they differ? Which model do you prefer and why?

My model in step 12 has 3 predictor variables and Adjusted R^2 of 0.68. The model in step 13 using 'all subset best model' technique has 4 predictor variables, so it is more complex and has Adjusted R^2 of 0.74.

However, I have performed regression diagnostics shown in Figure 10 and found that conditions of Linear Regression are significantly improved and much better satisfied by the latter model. Hence, I prefer the 'all subset best' model from step 13.

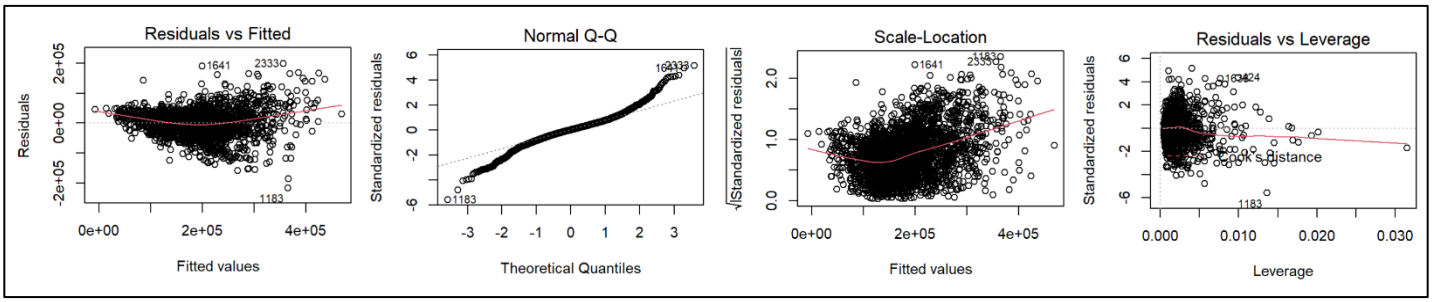

Figure 10. Regression Diagnostics of the 'All Subsets Best Model'

Interpretations & Conclusion

In my project, I found that the house attributes most correlated with Sale Price are 'Above ground living area', 'Garage Area', ' Total Basement Area', and 'First Floor Area' with all having positive correlation above 0.6. A total of 525 missing values (1.5%) among 11 attributes combined were imputed with mean or median and about 0.3% of the outliers were removed as they were distorting the model. The best model well-satisfied the assumptions of regression.

Using the 'All Subsets' feature selection, I am confident that the four features of 'Above ground living area', 'Garage Area', 'Total Basement Area', 'Masonry veneer area' give the right balance of model complexity and neither over nor underfitting. I found the best model to be a multiple regression model represented by the following equation and having an Adjusted R^2 of 0.74.

$$\boldsymbol{SalePrice = -30230.6 + 67.3 * Gr.Liv.Area + 90.5 * Garage.Area}$$
$$\boldsymbol{+ 59.5 * Total.Bsmt.SF + 58.0 * Mas.Vnr.Area}$$

This can be used to predict Sale Price of houses using just the four attributes of the house. It indicates that within the give range of all variables, when other variables are kept constant, a square foot increase in 'Above ground living area' will increase 'Sale Price' by $67.3. Similarly, a unit increase in 'Garage Area' will increase 'Sale Price' by $90.5 and a unit increase in 'Total Basement Area' will increase te price by $59.5. Lastly, an additional 'Masonry veneer area' square foot will increase the price by $58.0.

References

(n.d.). Retrieved from StackOverFlow: https://stackoverflow.com/questions/24568056/rs-read-csv-prepending-1st-column-name-with-junk-text

(n.d.). Retrieved from StackExchange: https://stats.stackexchange.com/questions/5253/how-do-i-get-the-number-of-rows-of-a-data-frame-in-r

(n.d.). Retrieved from Stack Over Flow: https://stackoverflow.com/questions/10085806/extracting-specific-columns-from-a-data-frame

(n.d.). Retrieved from TutorialsPoint.com: https://www.tutorialspoint.com/how-to-replace-na-values-in-columns-of-an-r-data-frame-form-the-mean-of-that-column

(n.d.). Retrieved from ListenData.com: https://www.listendata.com/2015/06/r-keep-drop-columns-from-data-frame.html

(n.d.). Retrieved from StatisticsByJim.com: https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Canvas Module 1 Assignment. (n.d.). Retrieved from https://northeastern.instructure.com/courses/110053/assignments/1345423

Histogram Bins: RPubs.com. (n.d.). Retrieved from https://rpubs.com/rodolfo_mendes/change-number-bins-histogram

Imputation: Medium.com. (n.d.). Retrieved from https://medium.com/analytics-vidhya/feature-engineering-part-1-mean-median-imputation-761043b95379

Linear Models in R: TheAnalysisFactor.com. (n.d.). Retrieved from https://www.theanalysisfactor.com/linear-models-r-diagnosing-regression-model/

Poulson, B. (n.d.). *Learning R: LinkedIn Learning*. Retrieved from https://www.linkedin.com/learning/learning-r-2/importing-data-from-a-spreadsheet

Set Working Directory: R Studio Support. (n.d.). Retrieved from https://support.rstudio.com/hc/en-us/articles/200711843-Working-Directories-and-Workspaces-in-the-RStudio-IDE

Appendix

R Code

```
# Importing libraries
library(psych)       # for EDA describe
library(dplyr)       # for data manipulation
library(car)         # for regression
library(ggplot2)     # for visualization
library(corrplot)    # for correlation
library(RColorBrewer) # for color in correlation
library(RColorBrewer) # for color in correlation
library(leaps)       # for feature selection


# 1. Load dataset ###################################

df = read.csv('Assign1/AmesHousing.csv',fileEncoding = "UTF-8-BOM")
rownames(df) <- df$Order
df <- df %>%
  select(-Order)


# 2. EDA ####################################### 

# view data
glimpse(df)
head(df)

# descriptive statistics
summary(df)
describe(df)  #to analyze skewness

# Histograms, Boxplot
hist(df$SalePrice, main='Sale Price Histogram')
hist(df$Lot.Area, breaks=50, main='Lot Area Histogram')
boxplot(df$SalePrice, main='Sale Price Boxplot')


# 3. Preparing Data ################################

df2 <- df
# Imputing continuous length variable with mean
df2$Lot.Frontage[is.na(df$Lot.Frontage)] = mean(df$Lot.Frontage, na.rm=TRUE)

# Imputing continuous area variables with median
```

```
df2$Mas.Vnr.Area[is.na(df$Mas.Vnr.Area)] = median(df$Mas.Vnr.Area, na.rm=TRUE)
df2$BsmtFin.SF.1[is.na(df$BsmtFin.SF.1)] = median(df$BsmtFin.SF.1, na.rm=TRUE)
df2$BsmtFin.SF.2[is.na(df$BsmtFin.SF.2)] = median(df$BsmtFin.SF.2, na.rm=TRUE)
df2$Bsmt.Unf.SF[is.na(df$Bsmt.Unf.SF)] = median(df$Bsmt.Unf.SF, na.rm=TRUE)
df2$Total.Bsmt.SF[is.na(df$Total.Bsmt.SF)] = median(df$Total.Bsmt.SF, na.rm=TRUE)
df2$Garage.Area[is.na(df$Garage.Area)] = median(df$Garage.Area, na.rm=TRUE)

# Imputing discrete variables with median
df2$Bsmt.Full.Bath[is.na(df$Bsmt.Full.Bath)] = median(df$Bsmt.Full.Bath, na.rm=TRUE)
df2$Bsmt.Half.Bath[is.na(df$Bsmt.Half.Bath)] = median(df$Bsmt.Half.Bath, na.rm=TRUE)
df2$Garage.Yr.Blt[is.na(df$Garage.Yr.Blt)] = median(df$Garage.Yr.Blt, na.rm=TRUE)
df2$Garage.Cars[is.na(df$Garage.Cars)] = median(df$Garage.Cars, na.rm=TRUE)

# Converting response variable to continuous
df2$SalePrice <- as.numeric(df$SalePrice)

# New dataframe with only continuous variables
df3 =
subset(df2,select=c(Lot.Frontage,Lot.Area,Mas.Vnr.Area,BsmtFin.SF.1,BsmtFin.SF.2,Bsmt.Unf
.SF,Total.Bsmt.SF,X1st.Flr.SF,X2nd.Flr.SF,Low.Qual.Fin.SF,Gr.Liv.Area,Garage.Area,Wood.D
eck.SF,Open.Porch.SF,Enclosed.Porch,X3Ssn.Porch,Screen.Porch,Pool.Area,Misc.Val,SalePrice
))

# 4. Correlation Matrix ##############################

cor(subset(df3,select=-c(SalePrice)), df3$SalePrice)

# 5. Correlation Matrix Plot #########################

cors = cor(df3, use='pairwise')
corrplot(cors, type='upper', col=brewer.pal(n=8, name="RdYlBu"))

# 6. Scatter Plots ###################################

scatterplot(SalePrice ~ Gr.Liv.Area, data=df3, main='SalePrice vs. Gr.Liv.Area (r=0.7)')
scatterplot(SalePrice ~ Mas.Vnr.Area, data=df3, main='SalePrice vs. Mas.Vnr.Area (r=0.5)')
scatterplot(SalePrice ~ BsmtFin.SF.2, data=df3, main='SalePrice vs. BsmtFin.SF.2 (r=0.01)')

# 7. Fit Model #######################################

m <- lm(formula = SalePrice~Gr.Liv.Area+Garage.Area+Total.Bsmt.SF, data = df3)

# 8. Report Model ####################################

summary(m)
```

```
# 9. Regression Diagnostics ###########################

par(mfrow=c(2,2))
plot(m)
dev.off()

# 10. Multicollinearity ##############################
vif(m)

# 11. Outliers #######################################
outlierTest(model = m)

# 12. Correcting Issues ##############################

df4 <- df3[-c(1499,2181,2182,1768,45,1064,1761,433,434,2446),]
m2 <- lm(formula = SalePrice~Gr.Liv.Area+Garage.Area+Total.Bsmt.SF, data=df4)
par(mfrow=c(2,2))
plot(m2)
dev.off()
vif(m2)
outlierTest(model = m2)

# 13. All Subsets Regression ##########################

leaps <-
regsubsets(SalePrice~Gr.Liv.Area+Garage.Area+Total.Bsmt.SF+X1st.Flr.SF+Mas.Vnr.Area,
data=df4, nbest=2)
summary(leaps)
plot(leaps,scale="adjr2")

m3 <- lm(formula = SalePrice~Gr.Liv.Area+Garage.Area+Total.Bsmt.SF+Mas.Vnr.Area, data =
df4)
summary(m3)

# 14. Comparing Models ###############################

par(mfrow=c(2,2))
plot(m3)
dev.off()
```