MSDS 6371 Project: House Prices Advanced Regression Techniques

Tim Cabaza
https://github.com/02e2e/02e2e.github.io/blob/e2c5f9d81e3e6da975d6453e5e013f44c8357edd/Final%20Draft%20Final%20Project%20(1).pdf

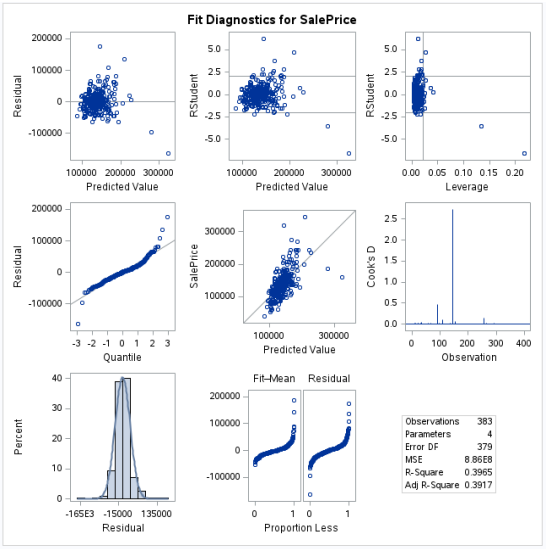
Matthew D. Cusack
https://github.com/Mcusac/Mcusac.github.io/blob/6a0763d5c08f76a161c02f9c455868d64bcf39a4/Final%20Draft%20Final%20Project.pdf

April 10, 2023

Analysis Question 1:

We need to determine the best model for estimating the sales price based on the square footage (per 100 feet) of the living area as well as if the sales price and its relationship to square footage is dependent on which neighborhood the house is located in given three neighborhoods

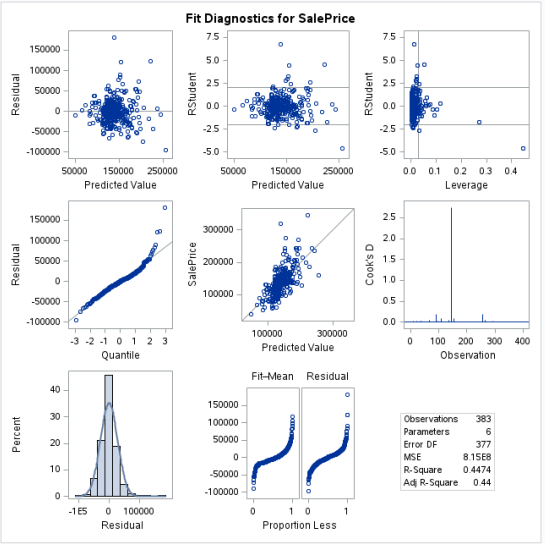
Assumptions Check Without interactions



Evaluate the Matrix one by one:

- Histogram: the residuals look to be normally distributed
- QQ Plot: the points tend to follow a straight line, however, it does have a bit of a tail at the upper end
- Cook's D: One observation has a high Cook's D
- Residual plot: Clustering is present around 0 but there seems to be an even number of points above and below 0 which is to be expected with normality
- Studentized Residual Plot: There are about 8 values that are outside of the lines showing that these are some outliers among the data
- Leverage Plot: Among these outliers, three of them have visual evidence of higher leverage than the others
- Table: the R^2 for the model is low at 0.3965

Assumptions Check With interactions



Evaluate the Matrix one by one:

- Histogram: the residuals look to be normally distributed
- QQ Plot: the points tend to follow a straight line, however, it does have a bit of a tail at the upper end
- Cook's D: One observation has a high Cook's D
- Residual plot: Clustering is present around 0 but there seems to be an even number of points above and below 0 which is to be expected with normality
- Studentized Residual Plot: Several residuals are shown as being outliers among the data points. This will likely need to be followed up by looking at their leverages.
- Leverage Plot: Among these outliers, two of them have visual evidence of higher leverage than the others
- Table: the R^2 for the model is low at 0.45

Comparing Competing Models:

Without Interaction (left) / With Interaction (right)

Root MSE	29759
Dependent Mean	138063
R-Square	0.3965
Adj R-Sq	0.3917
AIC	8279.46157
AICC	8279.62072
SBC	7910.25371
CV PRESS	3.531521E11

Root MSE	28552
Dependent Mean	138063
R-Square	0.4474
Adj R-Sq	0.4400
AIC	8249.72388
AICC	8250.02255
SBC	7888.41209
CV PRESS	3.394779E11

- The adjusted R^2 value of the model without interactions was 0.39 compared to the model with interactions that had an Adj R^2 value of 0.44. The CVPress was slightly lower with the model with interactions at 3.394779E11.

- The model without interaction has a lower Adj R^2 value, meaning that we have evidence

that the model with interactions has better explanatory power.

Model Equations (Allowing For Different Slopes):

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood\} = \beta_0 + \beta_1 GrLivArea + \beta_2 BrkSide + \beta_2 Edwards + \beta_3 NAmes + \beta_4 GrLivAreaBrkSide + \beta_5 GrLivAreaEdwards + \beta_6 GrLivAreaNAmes$$

Brkside:

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = BrkSide\} = \beta_0 + \beta_1 GrLivArea$$

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = BrkSide\} = (19971.5) + (87.2) GrLivArea$$

Edwards:

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = Edwards\} = \beta_0 + \beta_1 * GrLivArea + \beta_2 * Edwards + \beta_5 GrLivAreaEdwards$$

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = Edwards\} = (19971.5) + (87.2) GrLivArea + (68381.6) Edwards - (57.41) GrLivAreaEdwards$$

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = Edwards\} = (19971.5) + (87.2) GrLivArea + (68381.6) Edwards - (57.41) GrLivAreaEdwards$$

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = Edwards\} = (19971.5) + (68381.6) + (87.2 - 57.41) GrLivArea$$

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = Edwards\} = (88353.1) + (29.8) GrLivArea$$

NAmes:

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = NAmes\} = \beta_0 + \beta_1 * GrLivArea + \beta_3 * NAmes + \beta_6 GrLivAreaNAmes$$

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = Edwards\} = (19971.5) + (87.2) GrLivArea + (84704.89) NAmes + (-32.85) GrLivAreaNAmes$$

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = Edwards\} = (19971.5) + (84704.89) + (87.2 - 32.85) GrLivArea$$

$$\mu\{SalePrice \mid GrLivArea,\ Neighborhood = Edwards\} = (104676.4) + (54.35) GrLivArea$$

Full Model Parameter Estimates:

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	19971.51379	B	12351.12493	1.62	0.1067	-4314.21151	44257.23910
GrLivArea100	8716.25326	B	978.19580	8.91	<.0001	6792.84996	10639.65656
Neighborhood Edwards	68381.59099	B	13969.51149	4.90	<.0001	40913.67039	95849.51158
Neighborhood NAmes	54704.88774	B	13882.33364	3.94	<.0001	27408.38290	82001.39258
Neighborhood Brk Side	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho Edwards	-5741.22302	B	1071.76654	-5.36	<.0001	-7848.61227	-3633.83377
GrLivArea*Neighborho NAmes	-3284.66699	B	1081.53753	-3.04	0.0026	-5411.26869	-1158.06528
GrLivArea*Neighborho Brk Side	0.00000	B	.	.	.	.	.

Interpretation w/ Confidence Intervals (with interactions):

- Neighborhood BrkSide is the reference here and will remain as a zero value.
- When all explanatory variables are equal to zero, the y-intercept is equal to 19971.51. We have evidence for this with 95% confidence limits of -4314.21 and 44257.24. This was the only one in this list that was found to have an insignificant p-value of .107.
- When all other explanatory variables are constant, SalePrice increases by 8716.25 units whenever the GrLivArea variable increases by one unit. We have evidence for this with 95% confidence limits of 6792.85 and 10639.66.
- When all other explanatory variables are constant, SalePrice increases by 68391.59 units whenever the Neighborhood Edwards variable increases by one unit. We have evidence for this with 95% confidence limits of 40913.67 and 95849.51.
- When all other explanatory variables are constant, SalePrice increases by 54704.89 units whenever the Neighborhood NAmes variable increases by one unit. We have evidence for this with 95% confidence limits of 27408.38 and 82001.39.

- When all other explanatory variables are constant, SalePrice increases by -5741.22 units whenever the GrLivArea100 and Neighborhood Edwards variables increase by one unit. We have evidence for this with 95% confidence limits of -7848.61 and -3633.83.
- When all other explanatory variables are constant, SalePrice increases by -3284.67 units whenever the GrLivArea100 and Neighborhood NAmes variables increase by one unit. We have evidence for this with 95% confidence limits of -5411.27 and -1158.07.

Reduced Model Parameter Estimates (no interactions):

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	69781.53830	B	5442.399527	12.82	<.0001	59080.45848	80482.61812
GrLivArea100	4576.00645		314.880141	14.53	<.0001	3956.87558	5195.13731
Neighborhood Edwards	-2882.15509	B	4930.632117	-0.58	0.5592	-12576.97584	6812.66566
Neighborhood NAmes	16105.62078	B	4395.351815	3.66	0.0003	7463.29116	24747.95040
Neighborhood BrkSide	0.00000	B	.	.	.	.	.

Interpretation w/ Confidence Intervals:

- Neighborhood BrkSide is the reference here and will remain as a zero value.
- When all explanatory variables are equal to zero, the y-intercept is equal to 69781.54. We have evidence for this with 95% confidence limits of 59080.56 and 80482.62.
- When all other explanatory variables are constant, SalePrice increases by 4576.01 units whenever the GrLivArea variable increases by one unit. We have evidence for this with 95% confidence limits of 3956.88 and 5195.14.
- When all other explanatory variables are constant, SalePrice increases by -2882.16 units whenever the Neighborhood Edwards variable increases by one unit. We have evidence for this with 95% confidence limits of -12576.98 and 6812.67. This was the only one of this list of variables found to be insignificant with a p-value of 0.559.
- When all other explanatory variables are constant, SalePrice increases by 16105.62 units whenever the Neighborhood NAmes variable increases by one unit. We have evidence for this with 95% confidence limits of 7463.29 and 24747.95.

Conclusion:

In conclusion, we found evidence that the model containing the interactions (Adj R^2 = 0.44) was a better fit than the model with no interactions (Adj R^2 = 0.39). Using this model, we compared the differences between the slopes of the three neighborhoods. There is not sufficient evidence to suggest that the lines are not parallel and the interaction terms were not explaining enough variance in the model to be statistically significant. Every one unit increase in living area (100sqft. = 1 unit) holding neighborhood constant, is associated with an increase Sale Price. The neighborhoods themselves have evidence of being good predictors of the SalePrice of the homes, however, this starting value (y-intercept) before other variables are considered was shown to be different between the neighborhoods. From this, we have learned that while the neighborhood itself is a good predictor of SalePrice, the difference between neighborhood SalePrice is likely to stay constant given the same changes in other variables.

Source	DF	SoS	Mean Square	F Value	Pr > f

Model	2	28301438250	70371955	0.0863212105	0.9173177
Error	377	307343083951	815233644.43		
Corrected Total	379	335644522201	885605599.47		

R Shiny: Price v. Living Area Chart (Refer to Appendix)

Analysis Question 2:

We need to produce four models: one from forward selection, one from backward elimination, one from stepwise selection, and an optional custom build. From these models we will generate an adjusted R^2, CV Press, and Kaggle Score for each model and clearly describe which model we feel is best in terms of being able to predict future sale prices of homes in Ames, Iowa.

Data Prep for Model Selection:

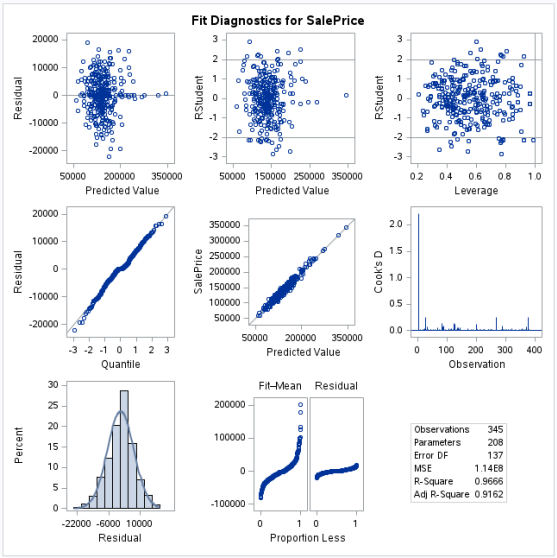
The same data preparation that was used in the first analysis was able to be used for this analysis as well. Except, this analysis used the train.csv data without focusing on just three neighborhoods but instead on all the neighborhoods in the dataset.

Model Selection:

Among the three models that were tested, we determined that the backward elimination model performed better than both the stepwise and the forward selection models. Because of the backward elimination model's performance, we decided to

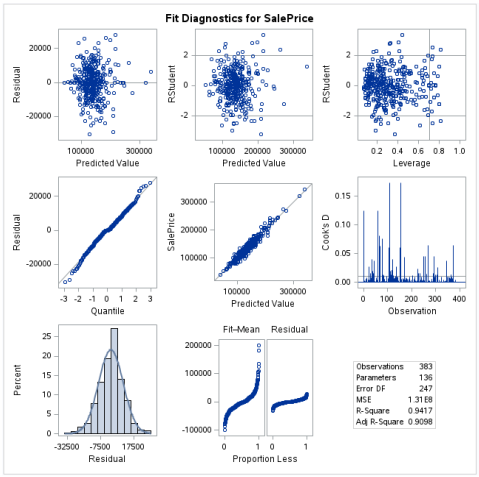
Generate Models:

Backward:

- Residual Plots
  - 

- Matrix Analysis:
    - There seems to be one data point in particularly high CooksD compared to the other observations.
    - When looking at the Rstudentized leverage graph, we do not have any visual evidence for a particularly highly leveraged data point. This is evidence to keep the data points that we possess.
    - Make sure to address each assumption
    - The Residual plot shows a large amount of clustering near the left side of the graph, though most of the points are spread across the y-axis denoting a normal distribution of residuals.
    - The Rstudentized residual plot shows a handful of outliers both above and below the lines given to us. It will be important to note the leverage of any outliers.
    - The Q-Q plot and the Predicted value vs. SalePrice graphs are both very linear and stick closely to the line. This shows evidence for little variance among our data.
    - The histogram plot for the residuals is very normally distributed showing the same description for the data.

Custom:

- Residual Plots
    - 

- Influential point analysis (Cook's D and Leverage)
    - Looking at the Rstudentized leverage plot, we can see quite a few observations with high leverage. We can see that about four of them show visual evidence of being significant residuals with high leverage. These points will likely need to be looked at further.
    - Looking at the Cook's D plot, we see at least a dozen points with high values denoting a high amount of outliers.
    - Make sure to address each assumption
    - The Residual plot shows a large amount of clustering near the left side of the graph, though most of the points are spread across the y-axis denoting a normal distribution of residuals.
    - The Rstudentized residual plot shows a handful of outliers both above and below the lines given to us. It will be important to note the leverage of any outliers.
    - The Q-Q plot is very linear and sticks closely to the line denoting low variance among the data points.

    - The Predicted value vs. SalePrice graphs is quite linear and stick relatively closely to the line aside from a bit of a wider range for the lower values. This shows evidence for little variance among our data though the lower values may need to be looked at more.
    - The histogram plot for the residuals is very normally distributed showing the same description for the data.
- Comparing Competing Models

Predictive Models	Adjusted R2	CV PRESS	Kaggle Score
Forward	0.7509	2.176231E11	2.47285
Backward	0.9172	8.098621E11	2.83563
Stepwise	0.72	2.213974E11	2.47231
CUSTOM	0.9050	3.386708E11	2.82314

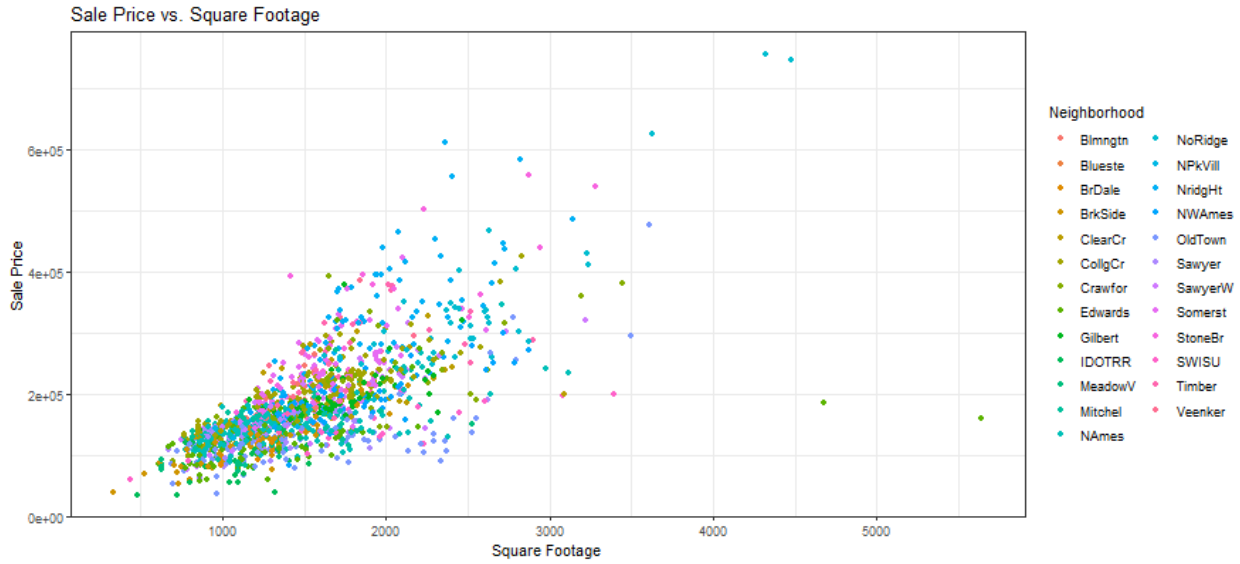
Conclusion:

Based on the Adjusted RSquared and CV press values that we obtained from the backward elimination model, we decided to base our custom model off of this. As can be seen in the table above; while the CV Press was smaller in the custom model than the backward model, so was the Kaggle Score. While this difference and that of the Adjusted R-squared were somewhat small, both of these values give evidence that the Backward model did end up performing better between the two and best among all four. Even comparing values among different seeds yielded quite similar results.

We have enough evidence to say that out of these models, the backward elimination performs best in being able to predict the SalePrice of a given unit based on the selected variables for the model.

Appendix

RShiny



R Code

- #####
- ## dependencies
- library(readr)
- library(ggplot2)
- library(tidyverse)
- library(dplyr)
- library(olsrr)
- #####
- ### ANALYSIS 1
- ## import data
- HouseData <- read_csv("E:/Docs/School/SMU-MSDS/Trimester 1/DS 6371 Statistical Foundations for DS/MSDS_6371_Stat_Foundations/group project/train.csv")
- View(HouseData)
-
- # create new data frame with filtered rows
- HouseData_filtered <- HouseData %>%
- filter(Neighborhood %in% c("NAmes", "Edwards", "BrkSide"))
- View(HouseData_filtered)
-
- # select our columns of interest
- selected_columns<- c("Neighborhood","GrLivArea","SalePrice")
- subset_HouseData_filtered <- subset(HouseData_filtered, select=selected_columns)
-
- subset_HouseData_filtered <- subset_HouseData_filtered %>%
- mutate(GrLivArea = GrLivArea / 100)
-

```
# View(subset_HouseData_filtered)
print(subset_HouseData_filtered)


## Fit Without Interaction
model1 <- lm(SalePrice ~ Neighborhood+GrLivArea, data = subset_HouseData_filtered)
summary(model1)
#Plot
plot(model1)
hist(residuals(model1), main="Histogram of Residuals", xlab="Residuals", freq=FALSE)
# Intervals - confidence and prediction
confint(model1, level = 0.95)
pred_c <- predict(model1,interval="confidence", type=c("response"), level=0.95)
pred_c

## Fit with Interaction
model2 <- lm(SalePrice ~ Neighborhood+GrLivArea + Neighborhood:GrLivArea, data =
subset_HouseData_filtered)
summary(model2)
#Plot
plot(model)
hist(residuals(model2), main="Histogram of Residuals", xlab="Residuals", freq=FALSE)
confint(model2, level = 0.95)
pred_c <- predict(model2,interval="confidence", type=c("response"), level=0.95)
pred_c

#####
### ANALYSIS 2
## import data
HouseData <-
read_csv("/Users/tmc/Desktop/MS_SMU_Admin/2023_Semester_01/DS6371_stats/Unit_1
4_total/_project_/Project/train.csv")

# manipulate data
# columns to replace NA with 0
cols_zero <- c("LotFrontage", "MasVnrType", "GarageYrBlt")

# columns to replace NA with "No"
cols_no <- c("Alley", "MasVnrArea", "BsmtQual", "BsmtCond", "BsmtExposure",
             "BsmtFinType1", "BsmtFinType2", "Electric", "FireplaceQu",
             "GarageType", "GarageFinish", "GarageQual", "GarageCond",
             "PoolQC", "Fence", "MiscFeature")

subset_HouseData_1 <- HouseData %>%
 select(-MiscVal) %>%
 arrange(Neighborhood) %>%
 mutate(GrLivArea = GrLivArea / 100) %>%
 select(Neighborhood, everything()) %>%
 mutate_at(vars(cols_zero), ~if_else(is.na(.), 0, .)) %>% # replace NA with 0
 mutate_at(vars(cols_no), as.character) %>% # convert columns to character
```

- mutate_at(vars(cols_no), ~if_else(is.na(.), "No", .)) %>% # replace NA with "No"
- mutate_if(is.character, as.factor) # change all character columns to factors
-
-
- # modeling
- SalePrice_Fit <- lm(SalePrice ~ ., data = subset_HouseData_1)
- stepwise_model <- ols_step_forward_p(SalePrice_Fit, penter = 0.01, details = TRUE)
- summary(stepwise_model)
-
- #####
- ## Debuggging
- # check number of unique values in each column
- sapply(subset_HouseData_1, function(x) length(unique(x)))
- # Check for missing data in subset_HouseData_1
- sum(is.na(subset_HouseData_1))
- # Check for missing values in data
- missing_values <- is.na(subset_HouseData_1)
- # Count the number of missing values in each variable
- colSums(missing_values)
- #####
- ## modeling
- # Columns that do not work with the Model: Street, Utilities, Condition1,Condition2, RoofMatl, 1stFlrSF, 2ndFlrSF, FireplaceQu, SsnPorch, PoolQC, Fence, MiscFeature
- # Forward Selection
- # install.packages("olsrr")
- library(olsrr)
- SalePrice_Fit <- lm(SalePrice~GrLivArea+SaleCondition+Id+MSSubClass+MSZoning+LotFrontage+LotArea+Alley+LotShape+LandContour+LotConfig+LandSlope+Neighborhood+BldgType+HouseStyle+OverallQual+OverallCond+YearBuilt+YearRemodAdd+RoofStyle+Exterior1st+Exterior2nd+MasVnrType+MasVnrArea+ExterQual+ExterCond+Foundation+BsmtQual+BsmtCond+BsmtExposure+BsmtFinType1+BsmtFinSF1+BsmtFinType2+BsmtFinSF2+BsmtUnfSF+TotalBsmtSF+Heating+HeatingQC+CentralAir+Electrical+LowQualFinSF+BsmtFullBath+BsmtHalfBath+FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+KitchenQual+TotRmsAbvGrd+Functional+Fireplaces+GarageType+GarageYrBlt+GarageFinish+GarageCars+GarageArea+GarageQual+GarageCond+PavedDrive+WoodDeckSF+OpenPorchSF+EnclosedPorch+ScreenPorch+PoolArea+MoSold+YrSold+SaleType, data=subset_HouseData_1)
- stepwise_model<-ols_step_forward_p(SalePrice_Fit,penter=0.01, details=TRUE)
- summary(stepwise_model)
-
- # Backward Selection
- # install.packages("olsrr")
- library(olsrr)
- SalePrice_Fit <- lm(SalePrice~GrLivArea+SaleCondition+Id+MSSubClass+MSZoning+LotFrontage+LotArea+Alley+LotShape+LandContour+LotConfig+LandSlope+Neighborhood+BldgType+HouseStyle+OverallQual+OverallCond+YearBuilt+YearRemodAdd+RoofStyle+Exterior1st+Exterior2nd+MasVnrType+MasVnrArea+ExterQual+ExterCond+Foundation+BsmtQual+BsmtCond+BsmtExposure+BsmtFinType1+BsmtFinSF1+BsmtFinType2+BsmtFinSF2+BsmtUnfSF+TotalBsmtSF+Heating+HeatingQC+CentralAir+Electrical+LowQualFinSF+BsmtFullBath+BsmtHalfBath

+FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+KitchenQual+TotRmsAbvGrd+Function
al+Fireplaces+GarageType+GarageYrBlt+GarageFinish+GarageCars+GarageArea+GarageQua
l+GarageCond+PavedDrive+WoodDeckSF+OpenPorchSF+EnclosedPorch+ScreenPorch+Pool
Area+MoSold+YrSold+SaleType, data=subset_HouseData_1)

- ols_step_backward_p(SalePrice_Fit,penter=0.05, details=TRUE)
-
- # Stepwise Selection
- # install.packages("olsrr")
- library(olsrr)
- SalePrice_Fit <-
  lm(SalePrice~GrLivArea+SaleCondition+Id+MSSubClass+MSZoning+LotFrontage+LotArea+
  Alley+LotShape+LandContour+LotConfig+LandSlope+Neighborhood+BldgType+HouseStyle
  +OverallQual+OverallCond+YearBuilt+YearRemodAdd+RoofStyle+Exterior1st+Exterior2nd+
  MasVnrType+MasVnrArea+ExterQual+ExterCond+Foundation+BsmtQual+BsmtCond+Bsmt
  Exposure+BsmtFinType1+BsmtFinSF1+BsmtFinType2+BsmtFinSF2+BsmtUnfSF+TotalBsmt
  SF+Heating+HeatingQC+CentralAir+Electrical+LowQualFinSF+BsmtFullBath+BsmtHalfBath
  +FullBath+HalfBath+BedroomAbvGr+KitchenAbvGr+KitchenQual+TotRmsAbvGrd+Function
  al+Fireplaces+GarageType+GarageYrBlt+GarageFinish+GarageCars+GarageArea+GarageQua
  l+GarageCond+PavedDrive+WoodDeckSF+OpenPorchSF+EnclosedPorch+ScreenPorch+Pool
  Area+MoSold+YrSold+SaleType, data=subset_HouseData_1)
-
- ols_step_both_p(SalePrice_Fit,penter=0.05, details=TRUE, cvpress=TRUE)
- #####
- ## RShiny
- library(shiny)
- library(ggplot2)
-
- # load data
- HouseData <- read_csv("E:/Docs/School/SMU-MSDS/Trimester 1/DS 6371 Statistical
  Foundations for DS/MSDS_6371_Stat_Foundations/group project/train.csv")
- View(HouseData)
-
- ui <- fluidPage(
- # input widget for selecting neighborhoods
- selectInput(inputId = "neighborhoods",
- label = "Select neighborhoods:",
- choices = unique(HouseData$Neighborhood),
- multiple = TRUE),
-
- # output widget for displaying the scatterplot
- plotOutput(outputId = "scatterplot")
-)
-
- server <- function(input, output) {
- output$scatterplot <- renderPlot({
-   # filter data by selected neighborhoods
-   subset_HouseData <- HouseData %>% filter(Neighborhood %in% input$neighborhoods)
-
-   # create scatterplot
-   ggplot(subset_HouseData, aes(x = GrLivArea, y = SalePrice, color = Neighborhood)) +

```
      geom_point() +
      labs(x = "Square Footage", y = "Sale Price", title = "Sale Price vs. Square Footage") +
      theme_bw()
  })
}
shinyApp(ui = ui, server = server)
```

SAS Code

```
FILENAME REFFILE '/home/u63038496/train.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=TrainData;
    GETNAMES=YES;
RUN;
PROC CONTENTS DATA=TrainData; RUN;

proc print data=TrainData;
run;

/* Filter by Neighborhood Column and Select only columns needed for the model */
data filtered_train;
 set TrainData;
 where Neighborhood in ('NAmes', 'Edwards', 'BrkSide');
 proc sort;
 by Neighborhood;
run;

/* Data Prep: Delete specified columns from filtered_train dataset, note that 1stFlrSF and
2ndFlrSF contains numeric prefix so you need ""n to modify */
data filtered_train;
 set filtered_train;
 drop Street Utilities Condition1 Condition2 RoofMatl "1stFlrSF"n "2ndFlrSF"n FireplaceQu
SsnPorch PoolQC Fence MiscFeature LotFrontage;
run;

data filtered_train;
   set filtered_train;
   GrLivArea = GrLivArea / 100;
run;

proc print data = filtered_train;
run;

/* Relabel Neighborhood as it is a Categorical Variable */
data filtered_train_update;
set filtered_train;
/* Set labeling for Neighborhood */
if Neighborhood = 'NAmes' then number = '1';
```

    - else if Neighborhood = 'Edwards' then number = '2';
    - else if Neighborhood = 'BrkSide' then number = '3';
-
- if Neighborhood = 'NAmes' then symbol = 'N';
    - else if Neighborhood = 'Edwards' then symbol = 'E';
    - else if Neighborhood = 'BrkSide' then symbol = 'B';
-
- /* Plot the scatter plot */
- symbol1 value='N' color=black interpol=none;
- symbol2 value='E' color=green interpol=none;
- symbol3 value='B' color=red interpol=none;
- title 'GrLivArea versus SalePrice by Neighborhood';
-
- /* Plotting our data */
- proc gplot data=filtered_train_update;
- plot GrLivArea*SalePrice = Neighborhood /;
- title 'GrLivArea vs SalePrice by Neighborhood';
- run;
- quit;
-
- /* Plot Explantatory variables versus each other */
- proc sgscatter data=filtered_train_update;
- matrix GrLivArea SalePrice / group=Neighborhood;
- run;
-
- /* Run model on data - predict SalePrice by GrLivArea by Neighborhood */
- /* No interaction here (no bar) this does not allow for different slopes by neighborhood */
- proc glm data=filtered_train_update plots=all;
- class Neighborhood (ref="BrkSide");
- model SalePrice = GrLivArea Neighborhood / solution clparm;
- output out=glm_results p=PredMean r=Resid student=studentresid;
- run;
-
- /* getting the cv press */
- proc glmselect data = filtered_train_update;
- class Neighborhood (ref="BrkSide");
- model SalePrice = GrLivArea Neighborhood / selection=Forward(stop=CV) cvMethod=random(5) stats=adjrsq include = 2 cvdetails=cvpress;
- run;
-
- /* Run model on data - predict SalePrice by GrLivArea by Neighborhood */
- proc glm data=filtered_train_update plots=all;
- class Neighborhood (ref="BrkSide");
- model SalePrice = GrLivArea|Neighborhood / solution clparm;
- output out=glm_results p=PredMean r=Resid student=studentresid;
- run;
-
- proc print data=glm_results;
- title 'Model Results';
- var Neighborhood SalePrice PredMean Resid studentresid;

```
run;

proc glmselect data = filtered_train_update;
class Neighborhood (ref="BrkSide");
model SalePrice = GrLivArea|Neighborhood  / selection=Forward(stop=CV)
cvMethod=random(5) stats=adjrsq include = 4 cvdetails=cvpress;
run;

/* Residual plot from Plotwith Symbols */
proc sgplot data=glm_results;
   scatter x=PredMean y=Resid / markerattrs=(symbol=circle) datalabel=symbol;
   xaxis label='Predicted Values';
   yaxis label='Residuals';
   title "Residual Plot for SalePrice";
run;


data mycritval;
F = quantile('F', 0.975, 2,377);
run;
proc print data=mycritval;
    var F;
run;


/* Next we do the variable selection from week 14 */
/* Forward Selection  */
proc glmselect data=filtered_train_update seed=136239130;
 class MSSubClass MSZoning Alley LotShape LandContour LotConfig LandSlope
Neighborhood BldgType HouseStyle
 OverallQual RoofStyle Exterior1st Exterior2nd MasVnrType ExterQual ExterCond
Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1
 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional
GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType
 SaleCondition OverallCond BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr Fireplaces TotRmsAbvGrd MoSold YrSold GarageCars ;
 model SalePrice = GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape
LandContour LotConfig
 LandSlope Neighborhood BldgType HouseStyle OverallQual OverallCond YearBuilt
YearRemodAdd RoofStyle Exterior1st Exterior2nd
 MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1
BsmtFinType2
 BsmtFinSF2 MasVnrType Foundation BsmtExposure
 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical LowQualFinSF
BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr
 KitchenQual TotRmsAbvGrd Functional Fireplaces GarageType GarageYrBlt GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
 EnclosedPorch ScreenPorch PoolArea MoSold YrSold SaleType/
selection=Forward(stop=CV) cvMethod=random(5) stats=adjrsq;
run;
```

-
-
- /* Stepwise */
- proc glmselect data=filtered_train_update seed=58712578;
- class MSSubClass MSZoning Alley LotShape LandContour LotConfig LandSlope Neighborhood BldgType HouseStyle
- OverallQual RoofStyle Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1
- BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType
- SaleCondition OverallCond BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr Fireplaces TotRmsAbvGrd MoSold YrSold GarageCars ;
- model SalePrice = GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape LandContour LotConfig
- LandSlope Neighborhood BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle Exterior1st Exterior2nd
- MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1 BsmtFinType2
- BsmtFinSF2 MasVnrType Foundation BsmtExposure
- BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical LowQualFinSF BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr
- KitchenQual TotRmsAbvGrd Functional Fireplaces GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
- EnclosedPorch ScreenPorch PoolArea MoSold YrSold SaleType/selection=stepwise(stop=CV) cvMethod=random(5) stats=adjrsq;
- run;
-
-
-
- /* Backward */
- /* Run proc glmselect */
- proc glmselect data=filtered_train_update seed=291253103;
- class MSSubClass MSZoning Alley LotShape LandContour LotConfig LandSlope Neighborhood BldgType HouseStyle
- OverallQual RoofStyle Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1
- BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType
- SaleCondition OverallCond BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr Fireplaces TotRmsAbvGrd MoSold YrSold GarageCars;
- model SalePrice = GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape LandContour LotConfig
- LandSlope Neighborhood BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle Exterior1st Exterior2nd
- MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1 BsmtFinType2
- BsmtFinSF2 MasVnrType Foundation BsmtExposure
- BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical LowQualFinSF BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr

- KitchenQual TotRmsAbvGrd Functional Fireplaces GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
- EnclosedPorch ScreenPorch PoolArea MoSold YrSold SaleType/selection=backward(stop=CV) cvMethod=random(5) stats=adjrsq;
- run;
-
- /* Backward had the Highest R^2 -- proc glm below to get the plots for assumptions check */
- proc glm data=filtered_train_update plots=all;
- class MSSubClass MSZoning Alley LotShape LandContour LotConfig LandSlope Neighborhood BldgType HouseStyle
- OverallQual RoofStyle Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1
- BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional GarageType GarageFinish GarageQual GarageCond PavedDrive SaleType
- SaleCondition OverallCond BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr Fireplaces TotRmsAbvGrd MoSold YrSold GarageCars;
- model SalePrice = GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape LandContour LotConfig
- LandSlope Neighborhood BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle Exterior1st Exterior2nd
- MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1 BsmtFinType2
- BsmtFinSF2 MasVnrType Foundation BsmtExposure
- BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical LowQualFinSF BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr
- KitchenQual TotRmsAbvGrd Functional Fireplaces GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
- EnclosedPorch ScreenPorch PoolArea MoSold YrSold SaleType/ solution clparm;
- output out=glm_results p=PredMean r=Resid student=studentresid;
- run;
-
-
-
- /* Custom ALL */
- /* Removing statistically insignificant columns Exterior2nd, BldgType, ExterQual, ExterCond, BsmtFinType2, BsmtFinSF2
- BsmtUF, BsmtUnfSF, HeatingQC, CentralAir, Electrical, LowQualFinSF,BsmtFullBath, BsmtHalfBath , BedroomAbvGr KitchenAbvGr,GarageType
- GarageFinish GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch MoSold YrSold SaleType
- PoolArea TotalBsmtSF GarageYrBlt */
- /* 01 R^squared=0.93*/
- proc glmselect data=filtered_train_update;
- class MSSubClass MSZoning Alley LotShape LandContour LotConfig LandSlope Neighborhood HouseStyle
- OverallQual RoofStyle Exterior1st Exterior2nd MasVnrType Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1
- BsmtFinType2 Heating KitchenQual Functional SaleType

```
  SaleCondition OverallCond FullBath HalfBath Fireplaces TotRmsAbvGrd MoSold YrSold
GarageCars;
  model SalePrice = GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape
LandContour LotConfig
  LandSlope Neighborhood HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd
RoofStyle Exterior1st
  MasVnrArea BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1
  MasVnrType Foundation BsmtExposure Heating FullBath HalfBath
  KitchenQual TotRmsAbvGrd Functional Fireplaces
GarageCars/selection=backward(stop=CV) cvMethod=random(5) stats=adjrsq;
run;

/* 01 plots for assumptions check   */
proc glm data=filtered_train_update plots=all;
  class MSSubClass MSZoning Alley LotShape LandContour LotConfig LandSlope
Neighborhood HouseStyle
  OverallQual RoofStyle Exterior1st Exterior2nd MasVnrType Foundation BsmtQual
BsmtCond BsmtExposure BsmtFinType1
  BsmtFinType2 Heating KitchenQual Functional SaleType
  SaleCondition OverallCond FullBath HalfBath Fireplaces TotRmsAbvGrd MoSold YrSold
GarageCars;
  model SalePrice = GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape
LandContour LotConfig
  LandSlope Neighborhood HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd
RoofStyle Exterior1st
  MasVnrArea BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1
  MasVnrType Foundation BsmtExposure Heating FullBath HalfBath
  KitchenQual TotRmsAbvGrd Functional Fireplaces GarageCars/ solution clparm;
  output out=glm_results p=PredMean r=Resid student=studentresid;
run;


/* TESTING */
/* Load test data */
FILENAME REFFILE '/home/u63038496/test.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=TestData;
    GETNAMES=YES;
RUN;

data FinalTest;
   set TestData;
   GrLivArea = GrLivArea / 100;
run;

data train2;
set filtered_train_update FinalTest;
run;
```

```
proc print data=final_test; run;

/* Forward */
proc glm data = train2 plots = all;
class OverallQual GrLivArea GarageFinish BsmtCond KitchenQual SaleCondition;
Model SalePrice = OverallQual GrLivArea GarageFinish BsmtCond KitchenQual
SaleCondition / cli solution;
output out  = results p = predict;
run;

/* Backward */
proc glm data = train2 plots = all;
class GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape LandContour
LotConfig
 LandSlope Neighborhood BldgType HouseStyle OverallQual OverallCond YearBuilt
YearRemodAdd RoofStyle Exterior1st
 MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1
BsmtFinType2
 BsmtFinSF2 MasVnrType Foundation BsmtExposure
 BsmtUnfSF Heating HeatingQC CentralAir Electrical LowQualFinSF BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr
 KitchenQual TotRmsAbvGrd Functional Fireplaces GarageType GarageYrBlt GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
 EnclosedPorch ScreenPorch YrSold SaleType;
model SalePrice = GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape
LandContour LotConfig
 LandSlope Neighborhood BldgType HouseStyle OverallQual OverallCond YearBuilt
YearRemodAdd RoofStyle Exterior1st
 MasVnrArea ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1
BsmtFinType2
 BsmtFinSF2 MasVnrType Foundation BsmtExposure
 BsmtUnfSF Heating HeatingQC CentralAir Electrical LowQualFinSF BsmtFullBath
BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr
 KitchenQual TotRmsAbvGrd Functional Fireplaces GarageType GarageYrBlt GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
 EnclosedPorch ScreenPorch YrSold SaleType/ cli solution;
output out  = results p = predict;
run;

/* Stepwise */
proc glm data = train2 plots = all;
class GrLivArea OverallQual BsmtCond KitchenQual GarageFinish;
Model SalePrice = GrLivArea OverallQual BsmtCond KitchenQual GarageFinish / cli solution;
output out  = results p = predict;
run;

/* Custom1 */
proc glm data = train2 plots = all;
```

```
class GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape LandContour
LotConfig
 LandSlope Neighborhood HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd
RoofStyle Exterior1st
 MasVnrArea BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1
 MasVnrType Foundation BsmtExposure Heating FullBath HalfBath
 KitchenQual TotRmsAbvGrd Functional Fireplaces GarageCars;
Model SalePrice = GrLivArea SaleCondition MSSubClass MSZoning LotArea Alley LotShape
LandContour LotConfig
 LandSlope Neighborhood HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd
RoofStyle Exterior1st
 MasVnrArea BsmtQual BsmtCond BsmtFinType1 BsmtFinSF1
 MasVnrType Foundation BsmtExposure Heating FullBath HalfBath
 KitchenQual TotRmsAbvGrd Functional Fireplaces GarageCars / cli solution;
output out  = results p = predict;
run;

data results2;
set results;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 10000;
keep id SalePrice;
where id > 1460
;

proc means data = results2;
var SalePrice;
run;

/*export data to file called data.csv*/
proc export data=results2
   outfile="/home/u63038496/FinalProj/data.csv"
   dbms=csv
   replace;
Run;
```