Kaggle House Price Analysis:

Finding a Correlation and Regression of Leading Contributing Factors

Mohammed Al-Rousan

Robert Lane

Southern Methodist University

Contents

[Q1: House Prices in Ames Iowa 3](#_Toc131854120)

[Q1 Assumptions 3](#_Toc131854121)

[Linear Relationship & Normality 3](#_Toc131854122)

[Linear Relationship 3](#_Toc131854123)

[Multivariate Normality 4](#_Toc131854124)

[No Multicollinearity 4](#_Toc131854125)

[No Autocorrelation 4](#_Toc131854126)

[Homoscedasticity 5](#_Toc131854127)

[Fit Analysis 5](#_Toc131854128)

# Q1: House Prices in Ames Iowa

This question focuses on 3 neighborhoods, marked in the dataset at Names, Edwards, and BrkSide.

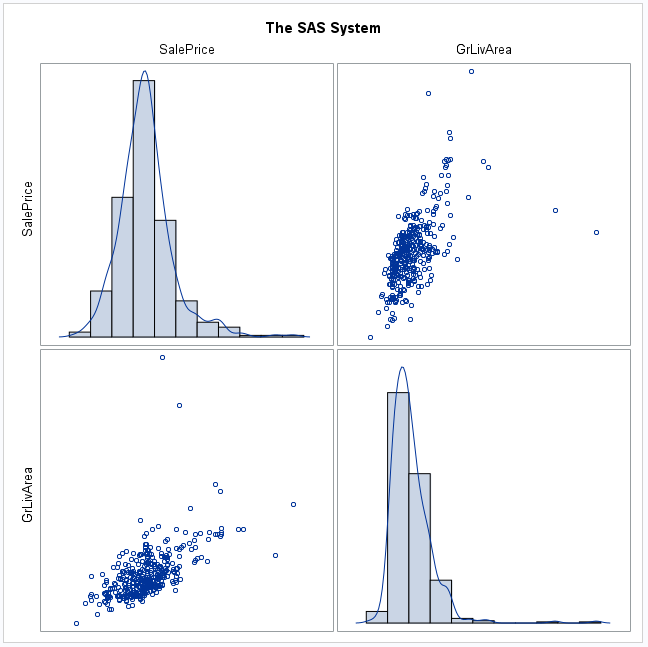
## Q1 Assumptions

### Linear Relationship & Normality

**proc** **sgscatter** data = neighborhoods;

matrix SalePrice GrLivArea / diagonal=(histogram kernel);

**run**;



Fail.

Both the independent and the dependent variables exhibit signs of right-skewness as well as increasing variance. Running a log-log model to attempt to correct for this. Also, coding neighborhoods as dummy variables for later use.

**data** loghood;

set neighborhoods;

logprice = log(SalePrice);

logarea = log(GrLivArea);

BrkSide = (Neighborhood = "BrkSide");

NAmes = (Neighborhood = "NAmes");

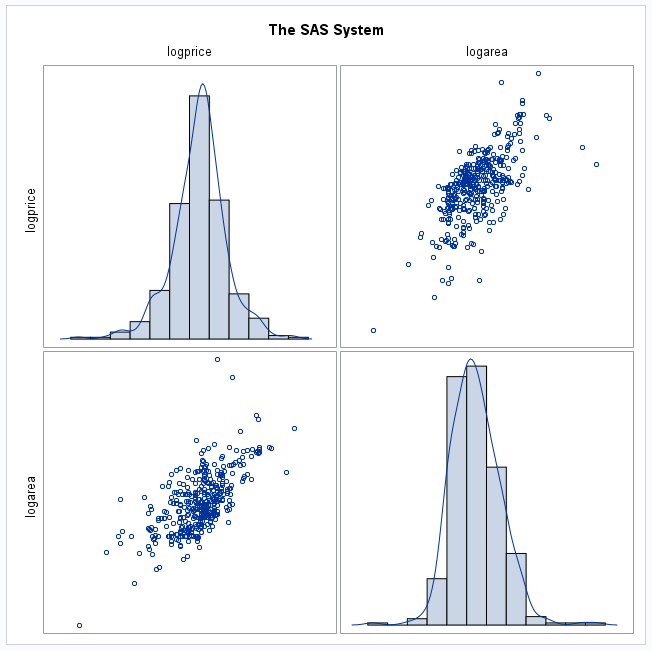
**run**;

### Linear Relationship

**proc** **sgscatter** data = loghood;

matrix logprice logarea / diagonal=(histogram kernel);

**run**;



Pass. The relationship between the log of the area and the log of the price does seem to exhibit a linear relationship.

### Multivariate Normality

Pass. Using the matrix above, it is evident that both inputs have normality.

### No Multicollinearity

Pass. Only one explanatory variable is used here.

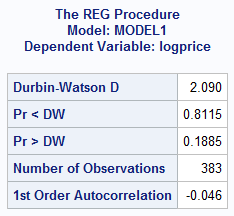
### No Autocorrelation

Running Durbin-Watson test for autocorrelation.

**proc** **reg** data = loghood;

model logprice = logarea BrkSide NAmes / dwProb;

run;



A Durbin-Watson score near 2 indicates that there is almost zero autocorrelation. Pass.

### Homoscedasticity

The variance visually appears to be pretty even at both low and high values, and for each axis. Pass.

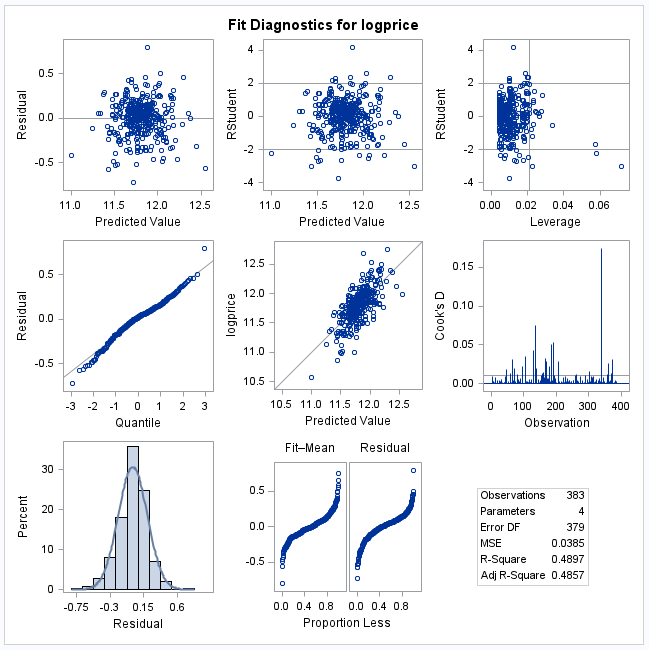
## Fit Analysis

**proc** **reg** data = loghood

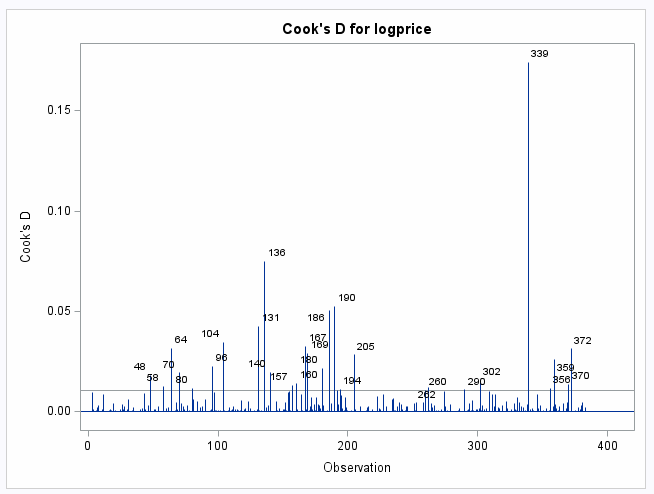
plots = (DiagnosticsPanel ResidualPlot(smooth));

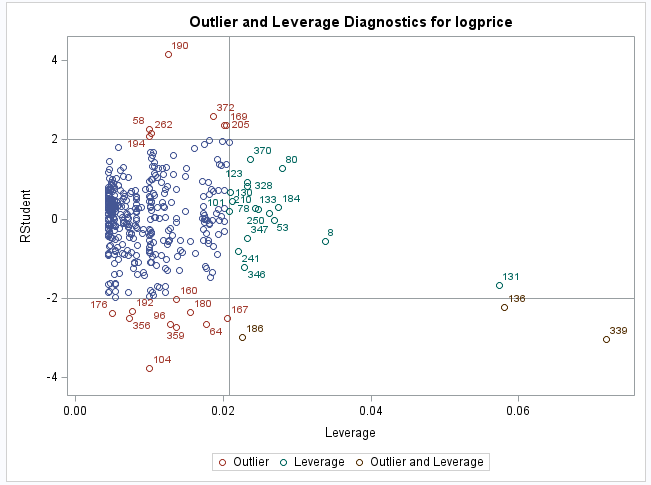
model logprice = logarea BrkSide NAmes;

**quit**;



The histogram and Q-Q plot both indicate that residuals are normally distributed, but the Leverage plot indicates that there are 3 influential outliers which should be checked.





186, 136, and 339 look suspect as influential values.

View just the relevant information for these houses:

**data** temp;

set loghood;

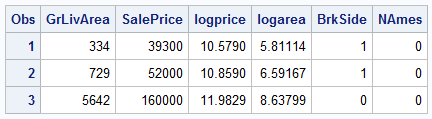
if \_n\_ in (**339**, **186**, **136**);

keep SalePrice GrLivArea logprice logarea BrkSide NAmes;

**run**;

**proc** **print** data=temp(obs=**3**);

**run**;



It seems unreasonable to make much of a prediction for the sale of a livable area of less than 500 square feet. For reference, that would be a living area off less than two typical parking spaces. It further seems unreasonable to expect a house larger than 5,000 square feet to sell at a price of $160,000. Therefore, of these three values, the first and last one (obs = 339 and 136) seem to be unlikely scenarios and may be removed. The middle one, while still an influential outlier, seems at least plausible.

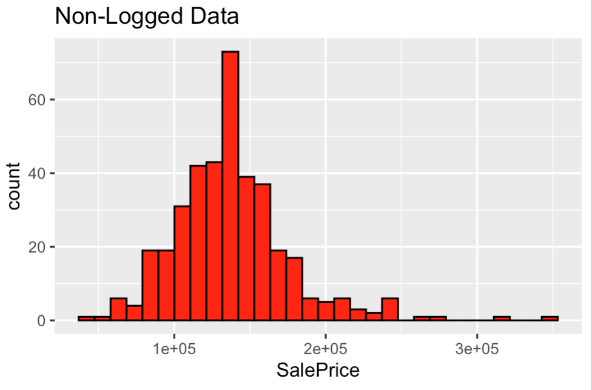
Model selection notes:

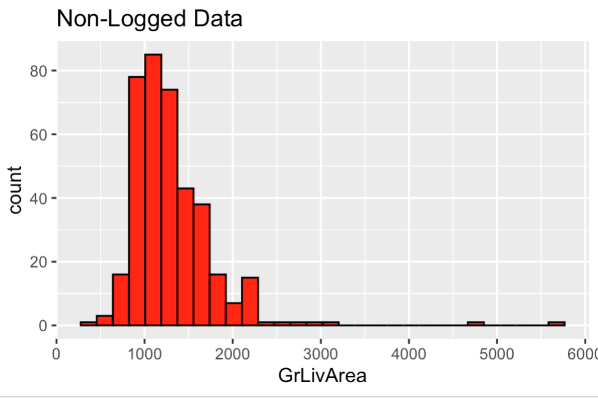
β0+ logarea + β1loglotarea + β1OverallQual + β2OverallCond + β3YearBuilt + β4YearRemodAdd + β5BsmtFinSF1 + β6BsmtFinSF2 + β8FullBath + β9HalfBath + β10BedroomAbvGr + β11KitchenAbvGr + β12TotRmsAbvGrd + β13Fireplaces + β14GarageCars + β15WoodDeckSF + β16OpenPorchSF + β17YrSold + β18logarea\*loglotarea + β19StreetPave + β20LotShapeIR1 + β21LotShapeIR2 + β22LotShapeIR3 + β23LandContourLow + β24LandContourBnk + β25LandContourLvl + β26NeighborhoodBrkSide + β27NeighborhoodSWISU + β28NeighborhoodIDOTRR + β29NeighborhoodEdwards + β30NeighborhoodMeadowV + β31NeighborhoodSawyer + β32NeighborhoodOldTown + β33NeighborhoodCrawfor + β34NeighborhoodSawyerW + β35NeighborhoodNAmes + β36NeighborhoodMitchel + β37NeighborhoodCollgCr + β38NeighborhoodGilbert + β39NeighborhoodNPkVill + β40NeighborhoodBrDale + β41NeighborhoodClearCr + β42NeighborhoodNWAmes + β43NeighborhoodStoneBr + β44NeighborhoodSomerst + β45NeighborhoodTimber + β46NeighborhoodBlmngtn + β47NeighborhoodVeenker + β48NeighborhoodBlueste + β49NeighborhoodNridgHt + β50BldgType1Fam + β51BldgTypeTwnhs + β52BldgTypeTwnhsE + β53BldgType2fmCon + β54HouseStyle1Story + β55HouseStyleSFoyer + β56HouseStyle15Fin + β57HouseStyle15Unf + β58HouseStyleSLvl + β59HouseStyle2Story + β60HouseStyle25Unf + β61RoofStyleFlat + β62RoofMatlCompShg + β63RoofMatlClyTile + β64RoofMatlMetal + β65RoofMatlWdShngl + β66RoofMatlMembran + β67RoofMatlWdShake + β68Exterior1stVinylSd + β69Exterior1stWdSdng + β70Exterior1stAsbShng + β71Exterior1stMetalSd + β72Exterior1stCemntBd + β73Exterior1stWdShing + β74Exterior1stPlywood + β75Exterior1stHdBoard + β76Exterior1stStucco + β77Exterior1stBrkFace + β78Exterior1stBrkComm + β79Exterior1stCBlock + β80FoundationSlab + β81FoundationCBlock + β82FoundationBrkTil + β83FoundationPConc + β84FoundationStone + β85BsmtQualEx + β86BsmtQualTA + β87BsmtQualGd + β88BsmtQualFa + β89BsmtCondTa + β90BsmtCondGd + β91HeatingGasA + β92HeatingGrav + β93HeatingGasW + β94HeatingWall + β95HeatingQCFa + β96HeatingQCTA + β97HeatingQCEx + β98HeatingQCGd + β99CentralAirY + β100GarageType2Types + β101GarageTypeDetchd + β102GarageTypeAttchd + β103GarageTypeCarPort + β104GarageTypeBasment + β105PavedDriveP + β106PavedDriveY

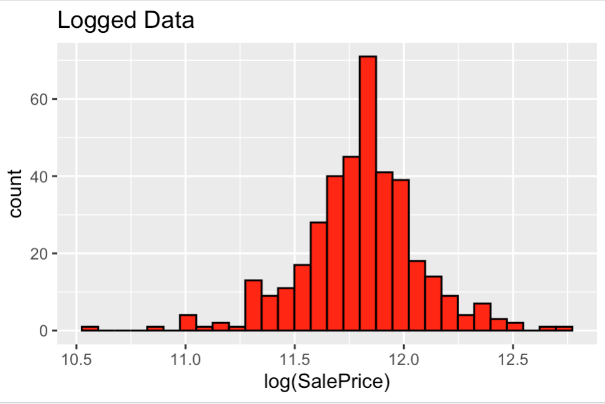
β0+ β1logarea + β2logarea\*loglotarea + β3OverallQual + β4OverallCond + β5YearBuilt + β6YearRemodAdd + β7BsmtFinSF1 + β8BsmtFinSF2 + β9KitchenAbvGr + β10Fireplaces + β11GarageCars + β12NeighborhoodIDOTRR + β13NeighborhoodEdwards + β14NeighborhoodOldTown + β15NeighborhoodCrawfor + β16NeighborhoodStoneBr + β17NeighborhoodSomerst + β18NeighborhoodNridgHt + β19HouseStyle2Story + β20RoofMatlClyTile + β21RoofMatlWdShngl + β22Exterior1stBrkFace + β23FoundationPConc + β24BsmtQualEx + β25HeatingGrav + β26HeatingQCEx

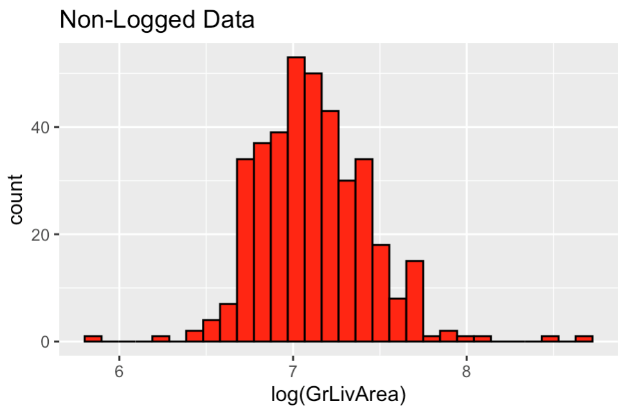
logprice = β0 + β1logarea + β1Names

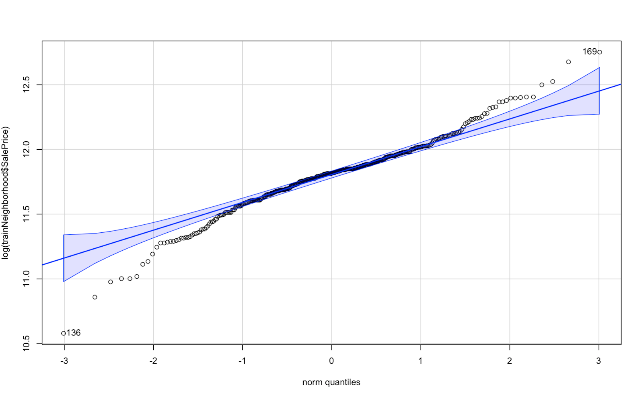
APPENDIX BACKUP

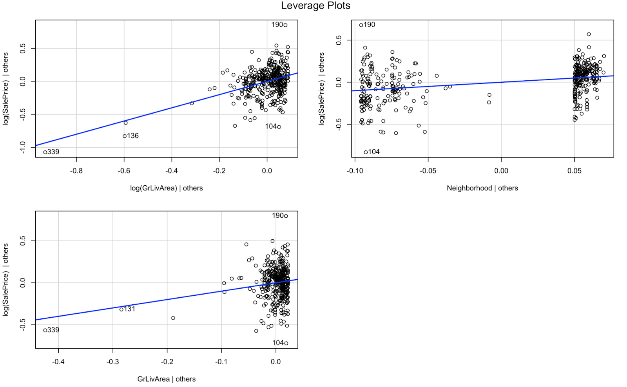
Plot 1  


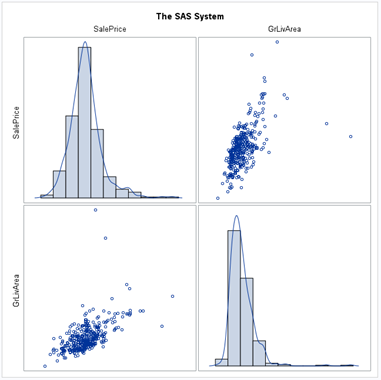
Plot 2  
 

Plot 3  
 

Plot 4   


Plot 5   


Plot 6  


Plot 7  
  
*SalePrice vs. GrLivArea.  Note both the right-skewness of each variable and the increasing variance.*

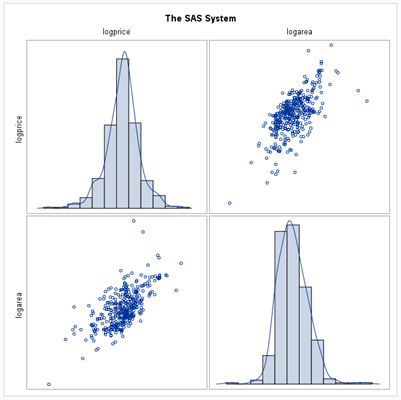
Plot 8  
  
*Log(SalePrice) vs. Log(GrLivArea).  Normality of each variable is increased, and the variance is visually more evenly spread.*

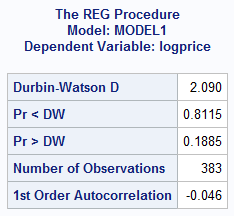
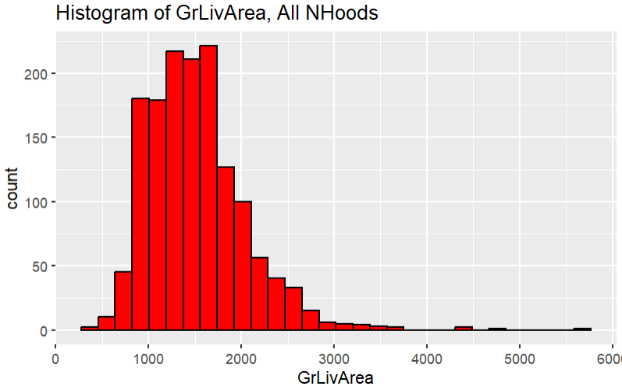
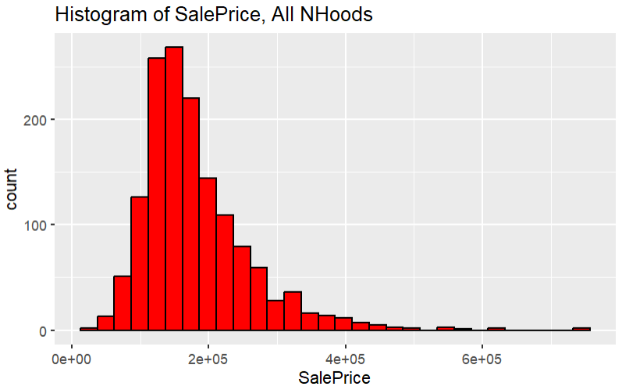
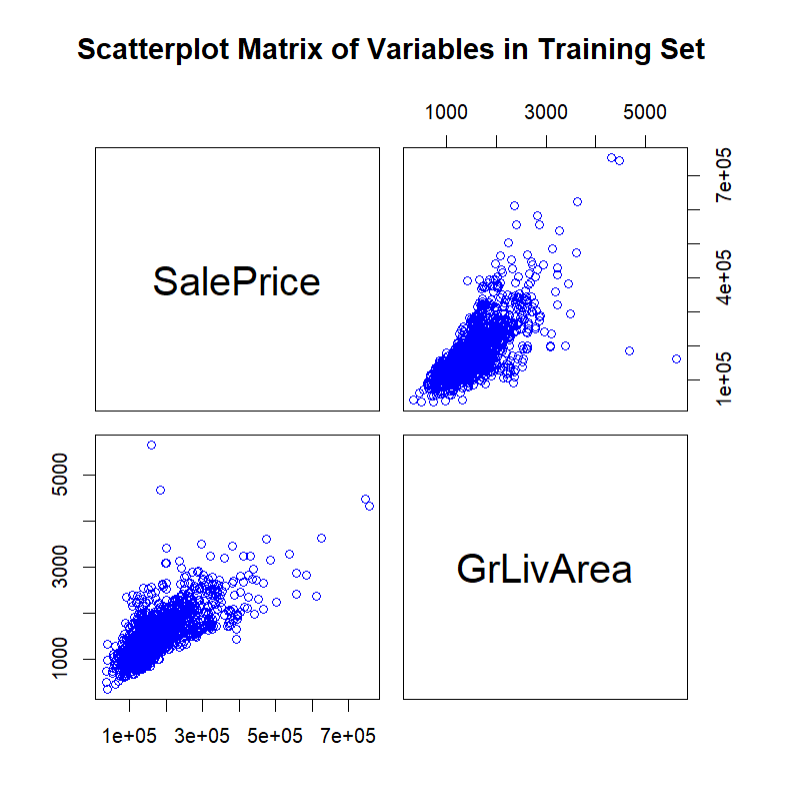
Fig. 9  


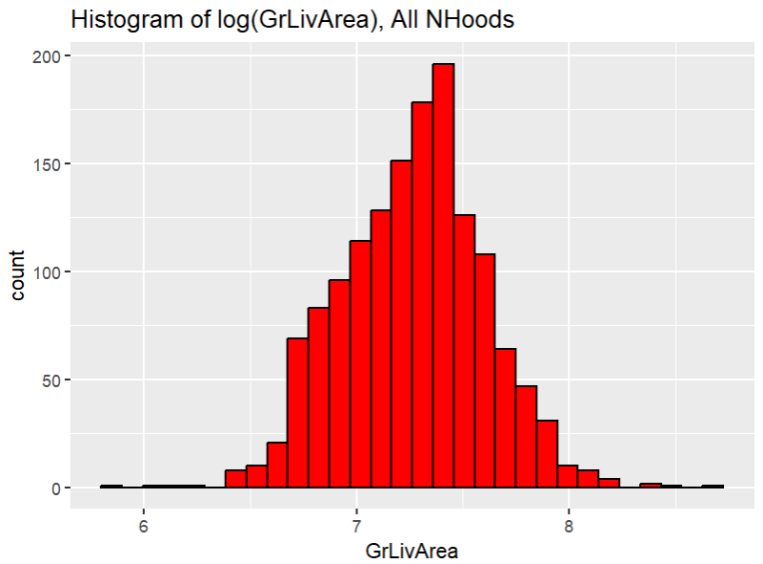
Fig. 10

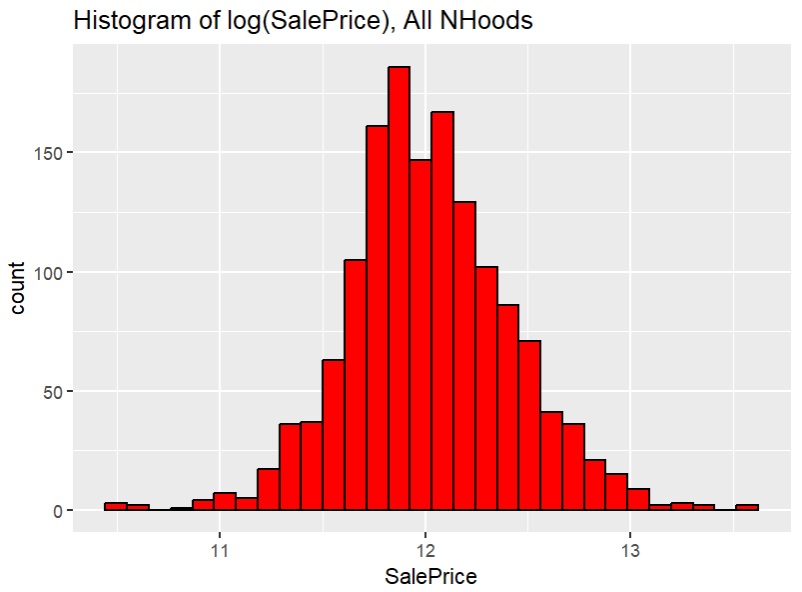
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 0.47464167 | 0.23732083 | 6.5462345 | 0.14161055 |
| **Separate Slopes Model** | 375 | 13.5948865 | 0.03625303 |  |  |
| **Parallel Slopes Model** | 377 | 14.0695281 | 0.0373197 |  |  |

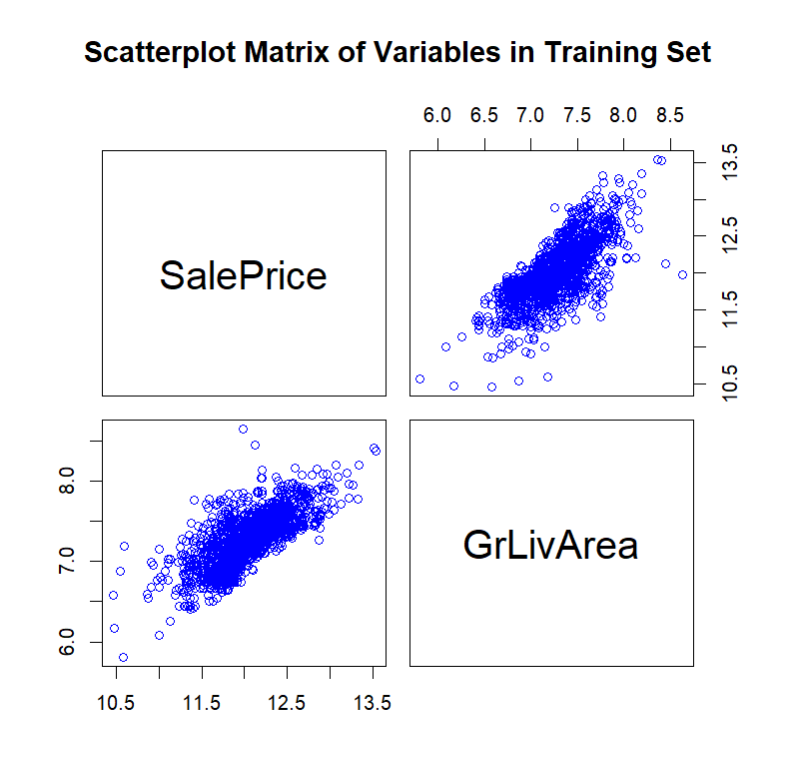
Plot 10  


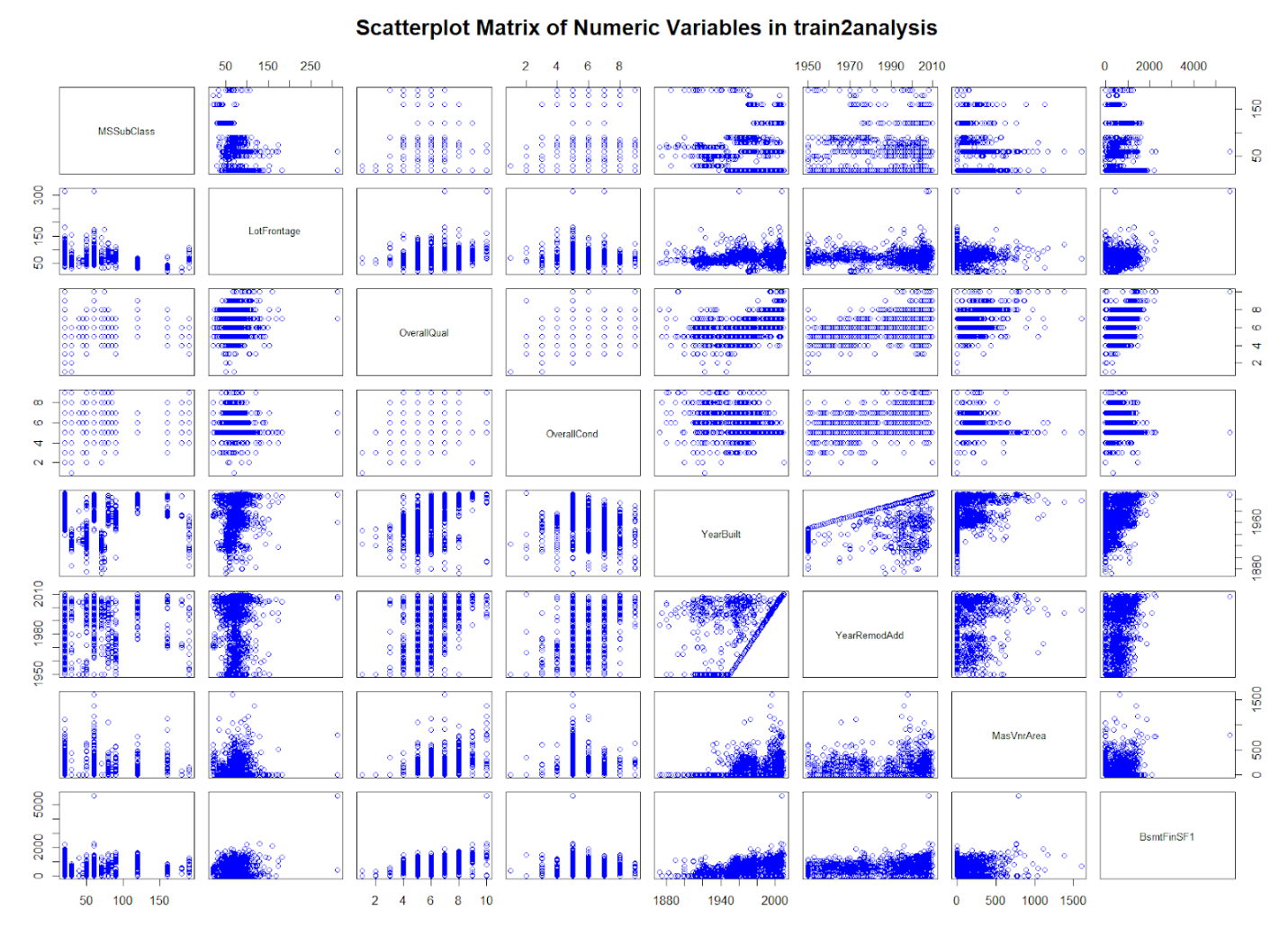
Plot 11  


Plot 12  


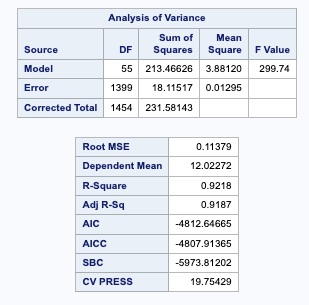
Plot 13  


Plot 14  


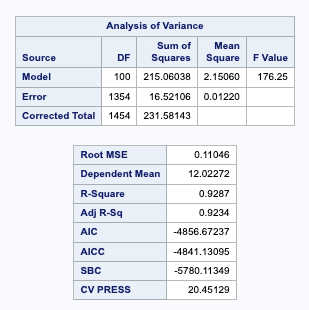
Plot 15  


Plot 16  


Plot 17



Plot 18



Plot 19

