STAT 306 2024W2 - Project: Report

Group Project Report

Outline

1. Explorative Data Analysis (right skew) + linear regression assumption violation
2. log(Y) transformation
3. model selection (forward)
4. multi-collinearity check
5. Interaction Term
6. outlier, leverage, influence

Data Preparation

The response variable 'SalePrice' column is divided by 1,000, so the unit for response variable is 1 thousand dollar.

Excluded Categories & Why (Reduce from 80 to 14)

Identifiers: `Order`, `PID` – not informative for modeling.

Highly Sparse or Rare Categories: `Misc Feature`, `Pool QC`, `Fence` – too many NAs or uncommon cases.

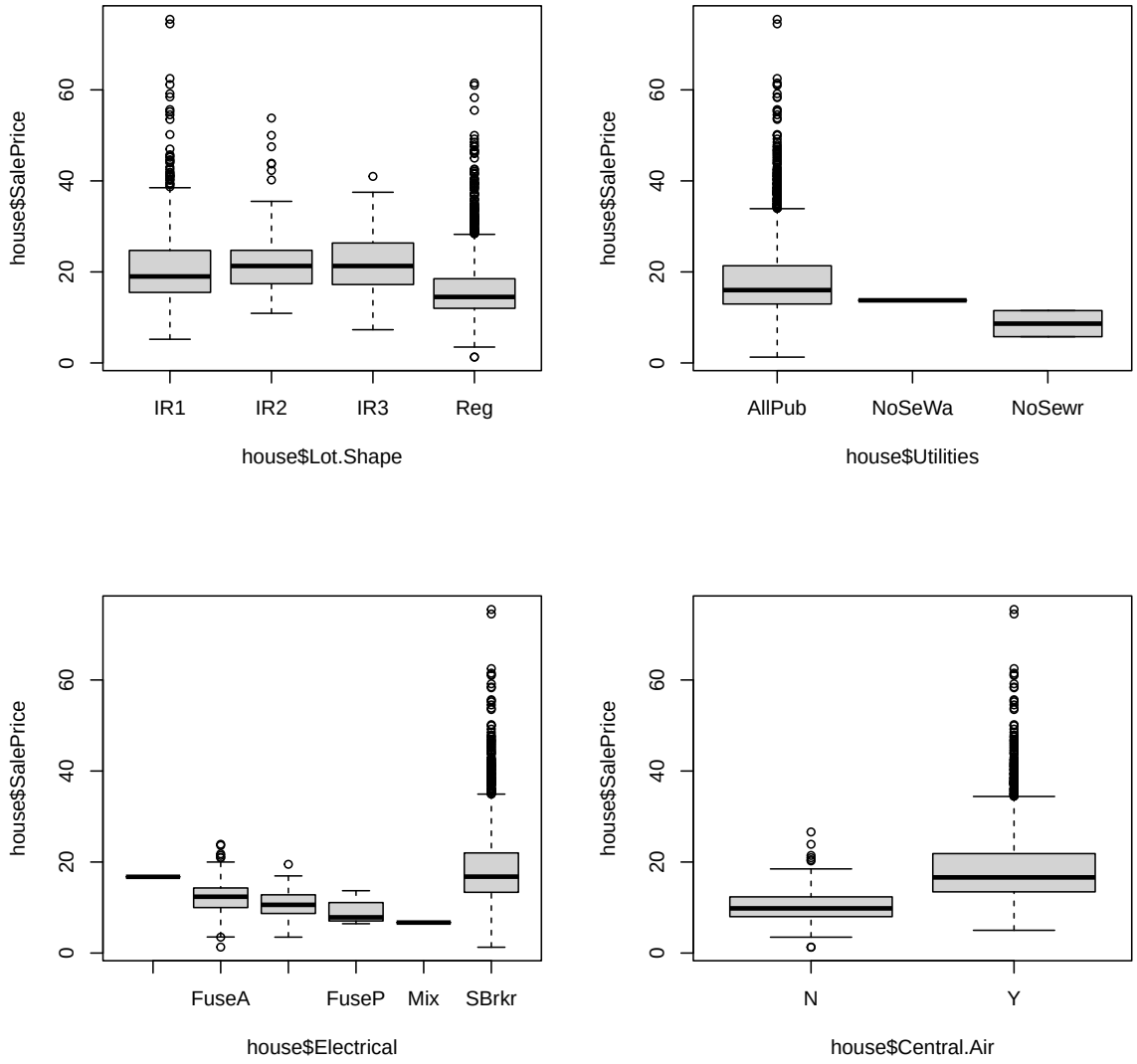
Redundant or Derived Variables: `Year Remod/Add` is often correlated with `Year Built` .

Uncertain Interpretation: `Roof Matl`, `Exterior2`, `Condition2` – often inconsistent or hard to use effectively.

Highly granular: `Neighborhood`, `MS SubClass`, and `Bldg Type` have many levels; including all may lead to overfitting unless consolidated.

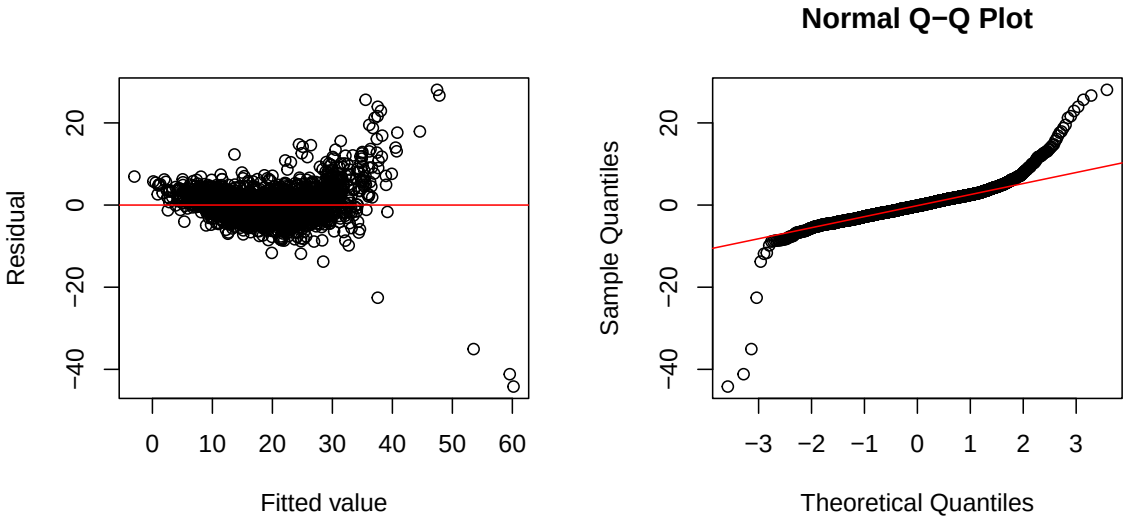
Some quality are too subjuctive evaluation, not good

Explorative Data Analysis



Right skew show on the plot is a signal a do log(Y) transformation + assumption check also

Model Diagnostic (Linear Regression Assumption Check)



verify assumption: residual vs fitted value plot

funnel pattern appear

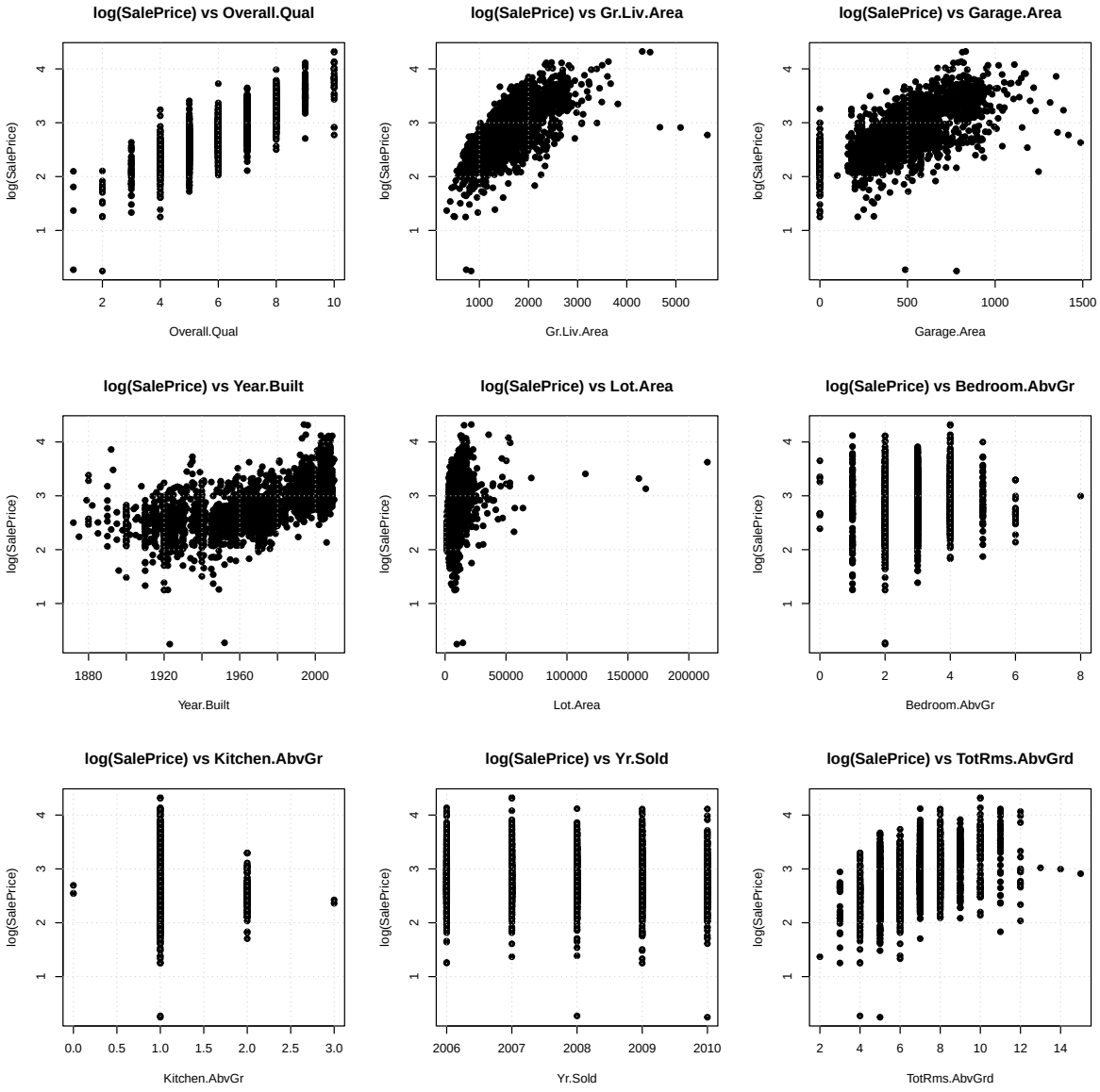
linear regression assumption: homoscedasticity $Var(\Sigma) = \sigma^2$ violated

heavy tail observed -> homoscedasticity $Var(\Sigma) = \sigma^2$ violated

Log(Y) Transformation

take log on response variable Y, log(Y)

create scatterplot between log(Y) and each continuous variables of X



every pair show linear relationship, except for `Year.Built` that still curved.

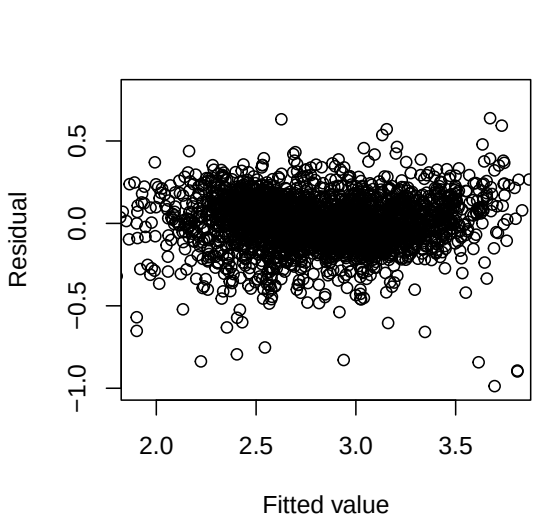
fit linear model again with log(Y) as response variable

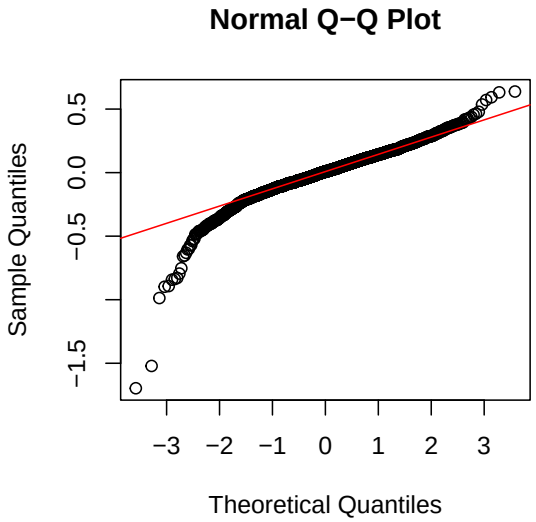
since `Year.Build` shows quadratic relationship, we use

$$(X_{Year.Built} - \bar{X_{Year.Built}})^2$$

looks random -> assumptions aligned! -> conclusion will be valid

now, transformed model become our final model.





Multi-collinearity check

calculate VIF to verify that there is not strong correlation between any continuous variables.

why use VIF? What is the advantage of using VIF? Using thumb theory of 10 as cutoff, since every continuous variable does not over 10, so no multi-collinearity.

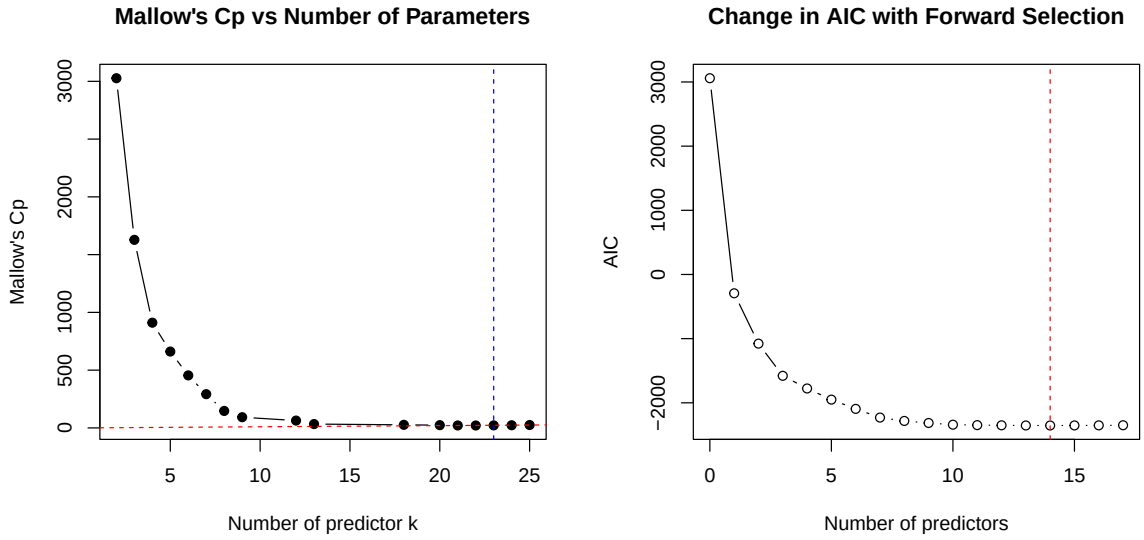
explain what is multi-collinearity -> any correlation among explanatory variables (X)

```
          Overall.Qual             Gr.Liv.Area             Garage.Area
              2.889419                5.419218                1.758177
            Year.Built               Lot.Area            Bedroom.AbvGr
              2.768125                1.149620                2.160246
          Kitchen.AbvGr               Yr.Sold            TotRms.AbvGrd
              1.285590                1.004751                4.695560
  Quatratic_Gr.Liv.Area  Quatratic_Year.Built  Quatratic_Garage.Area
              2.218409                1.778837                1.154445
Quatratic_TotRms.AbvGrd
              1.831163
```

Model Selection

Original dataset contains 80 columns, we use 14 variables within it.

forward selection



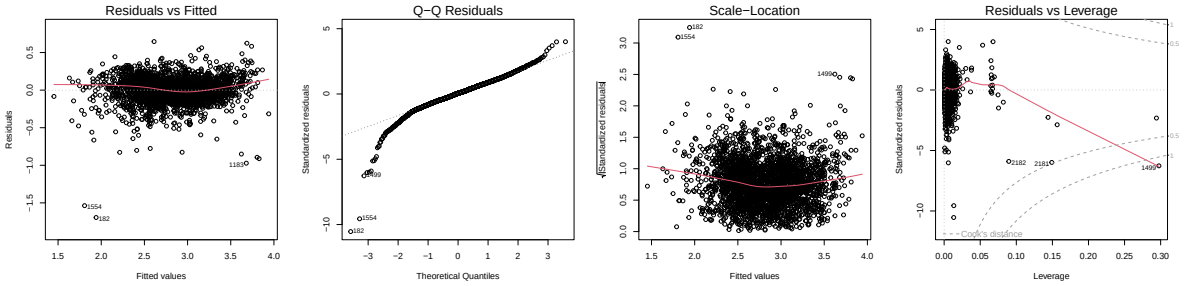
the difference of cp between 14 and 15 is really small, and the AIC shows that we should use 14.

These two plots shows we can use 14 covariates in our final model.

From the plot of Mallow's CP score, the score more close to number of parameter (p+1) means better

From the plot of AIC, the AIC score lower mean better.

Then we decided to use all covariates from forward selection, which covariate that has strongest relationship during each iteration in the greedy algorithm



Interaction Term

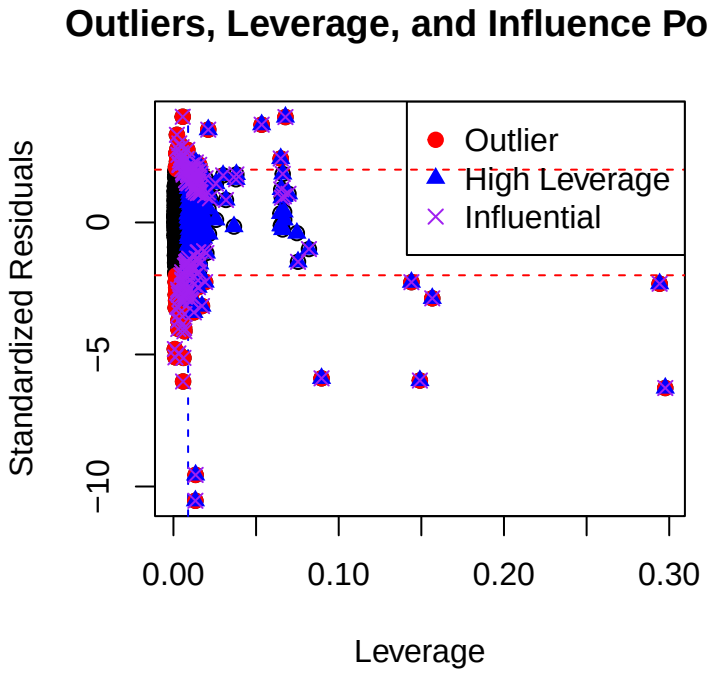
We explored all possible interaction of categorical variables (Central.Air, Lot.Shape, Electrical, Utilities) and numeric variables. The fitness of model that includes these interactions increases a little. We only selected significant interaction term and put them into final model `simpler_interactive_model`.

	number of parameter (p)	adj R^2
additive_model	25	0.8434

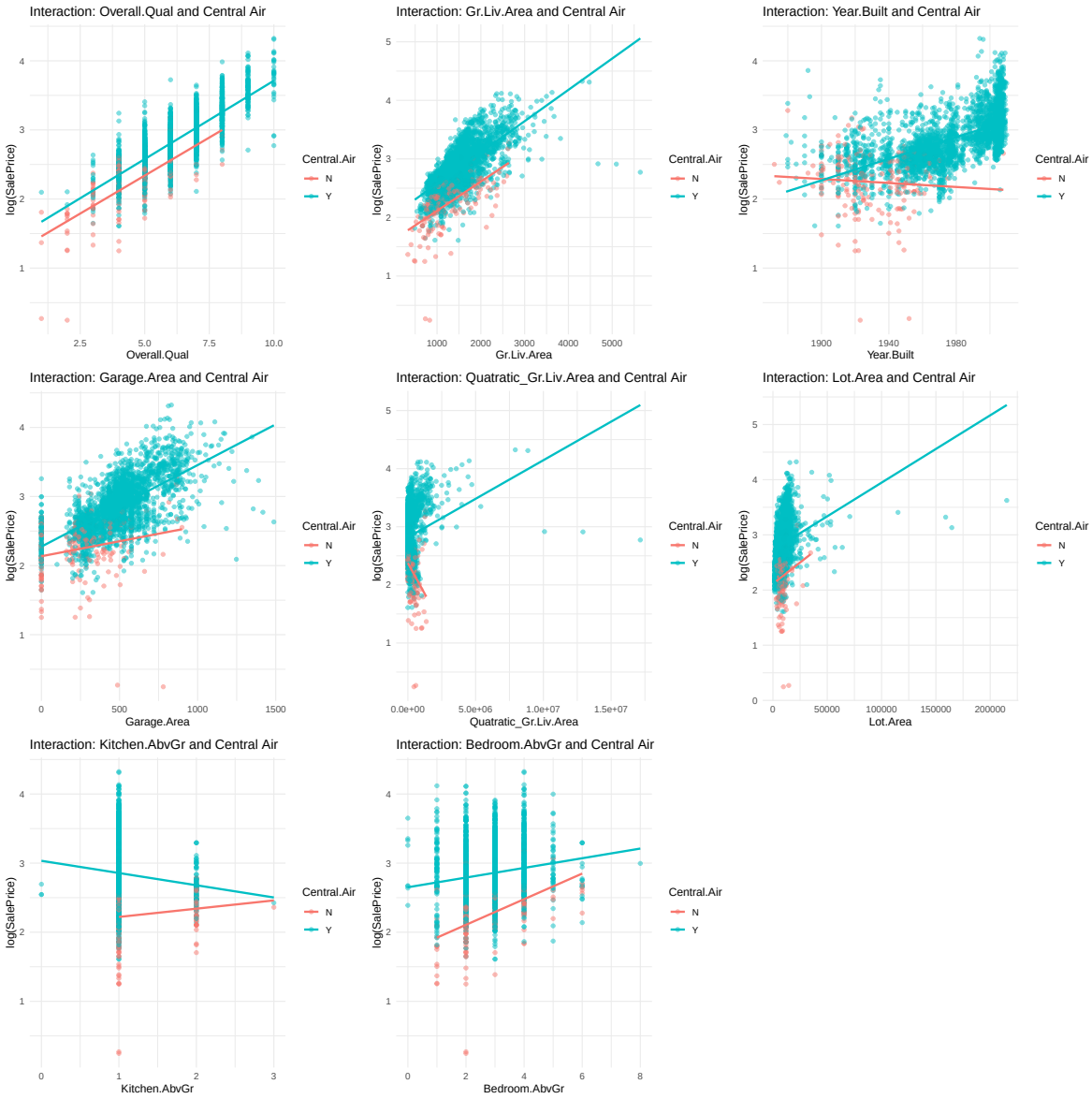
	number of parameter (p)	adj R^2
simpler_additive_model	22	0.8434
final_model	13	0.8423
interactive_model	39	0.8531
simpler_interactive_model	15	0.8443

Outlier, leverage, influence (Limitation)

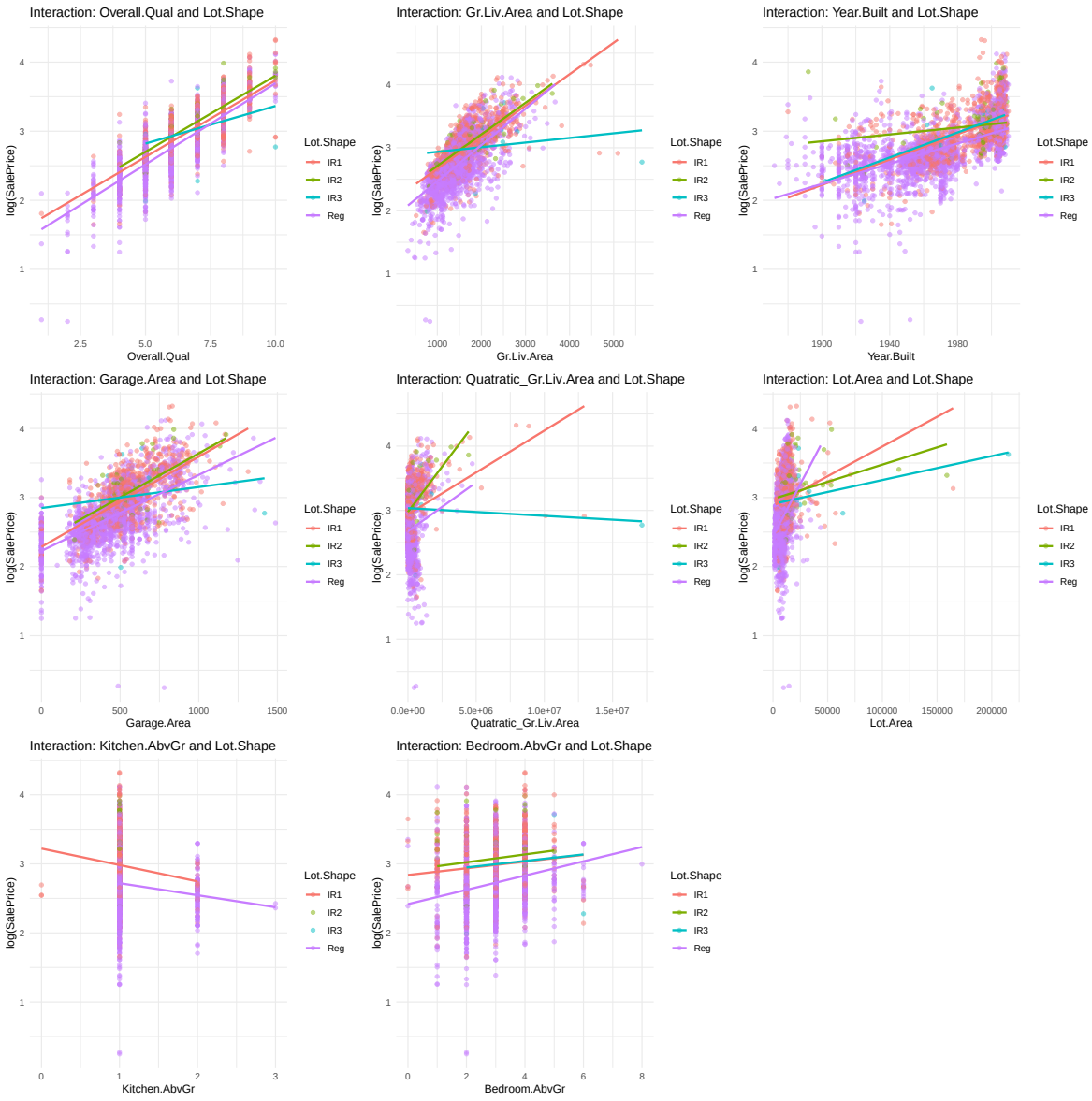
We identify all potential outliers, leverages and evaluated whether they are influential. It seems that there are lot of outliers and evaluated in our dataset, which cause a limitation or weakness.



Appendix Interaction Term




```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```


Interaction: Overall.Qual and Lot.Shape
log(SalePrice)
Overall.Qual
Interaction: Gr.Liv.Area and Lot.Shape
Gr.Liv.Area
Interaction: Year.Built and Lot.Shape
Year.Built
Interaction: Garage.Area and Lot.Shape
Garage.Area
Interaction: Quatratic_Gr.Liv.Area and Lot.Shape
Quatratic_Gr.Liv.Area
Interaction: Lot.Area and Lot.Shape
Lot.Area
Interaction: Kitchen.AbvGr and Lot.Shape
Kitchen.AbvGr
Interaction: Bedroom.AbvGr and Lot.Shape
Bedroom.AbvGr
Lot.Shape
IR1
IR2
IR3
Reg