Predicting Sales Price for Houses at Ames Iowa

As we can tell by the title, in this Kaggle competition, the goal is to predict the sales price of houses at Ames in the test dataset as accurate as possible.

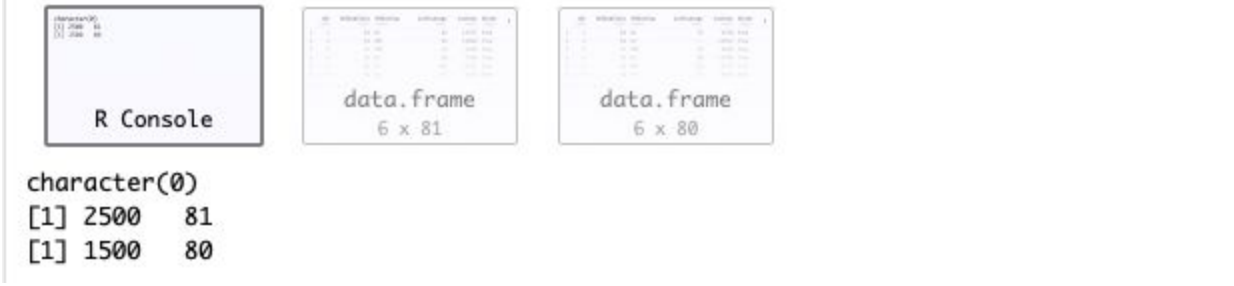
My kaggle name is Yuqing Yang. The R2 of my training model is 0.8723 and the adjusted R2 is 0.8718. And there are 9 predictors in the final model. My Kaggle Rank is 144.

```
Call:
lm(formula = tSalePrice ~ LotArea + OverallQual + Age + GrLivArea +
    FullBath + TotRmsAbvGrd + Fireplaces + GarageCars + BsmtFinSF1,
    data = train2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94774 -0.06624  0.00142  0.07934  0.68181

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.050e+01  1.700e-02 617.698  < 2e-16 ***
LotArea       1.016e-05  9.737e-07  10.435  < 2e-16 ***
OverallQual   1.341e-01  2.856e-03  46.955  < 2e-16 ***
Age          -9.926e-04  1.312e-04  -7.564 5.44e-14 ***
GrLivArea     1.542e-04  1.127e-05  13.681  < 2e-16 ***
FullBath      5.566e-02  7.147e-03   7.787 9.97e-15 ***
TotRmsAbvGrd  1.618e-02  2.966e-03   5.454 5.41e-08 ***
Fireplaces    3.154e-02  5.214e-03   6.050 1.66e-09 ***
GarageCars    7.841e-02  5.248e-03  14.941  < 2e-16 ***
BsmtFinSF1    1.685e-04  7.134e-06  23.617  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1403 on 2490 degrees of freedom
Multiple R-squared:  0.8723,    Adjusted R-squared:  0.8718
F-statistic:  1890 on 9 and 2490 DF,  p-value: < 2.2e-16
```

My goal is to predict the sales price of the houses with the most significant predictors. There are 2500 observations and 81 variables in the training data set. And there are 1500 observations and

80 variables in the test data.

```
head(train)
head(test)
dim(train)
dim(test)
```



First, I take look at how many missing values are there in the training dataset.

```{r}
colSums(is.na(train))
```

Ob	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour
0	0	4	394	0	0	2323	0	0
Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle	OverallQual
2	0	0	0	0	0	0	0	0
OverallCond	YearBuilt	YearRemodAdd	RoofStyle	RoofMatl	Exterior1st	Exterior2nd	MasVnrType	MasVnrArea
0	0	0	0	0	0	0	19	16
ExterQual	ExterCond	Foundation	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1	BsmtFinSF1	BsmtFinType2
0	0	0	71	70	70	68	1	69
BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	Heating	HeatingQC	CentralAir	Electrical	X1stFlrSF	X2ndFlrSF
1	1	1	0	0	0	0	0	0
LowQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	BedroomAbvGr	KitchenAbvGr	KitchenQual
0	0	3	3	0	0	0	0	1
TotRmsAbvGrd	Functional	Fireplaces	FireplaceQu	GarageType	GarageYrBlt	GarageFinish	GarageCars	GarageArea
0	1	0	1225	134	134	134	0	0
GarageQual	GarageCond	PavedDrive	WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch	ScreenPorch	PoolArea
134	134	0	0	0	0	0	0	0
PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
2491	2018	2421	0	0	0	1	0	0

I deleted Alley, PoolQC, Fence, MiscFeature because there are too many missing values. For the numerical variables: LotFrontage, MasVnrArea, I replaced all the missing values by their mean.

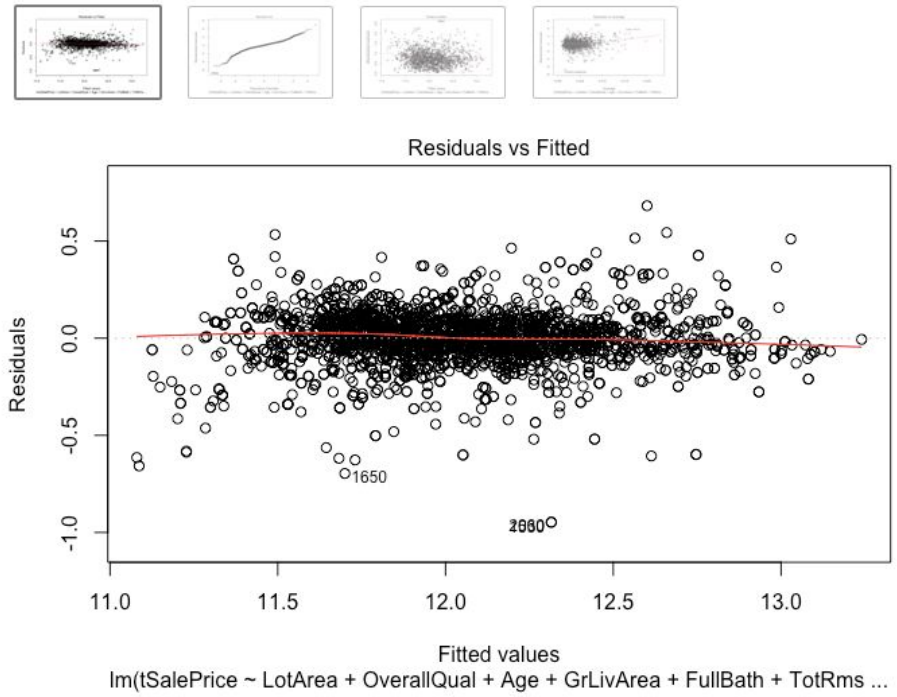
For these numerical variables: GarageYrBlt, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath, BsmtHalfBath. I replaced all the missing values by their median. For categorical variables: MSZoing, Utilities, MasVnrType, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, kitchenQual, GarageCond, GaraQual, BsmtQual, Functional, FireplaceQu, GarageType, GarageFinish, I replaced all the missing values with the most frequent non-missing value. For the test data set, I did the same replacement.

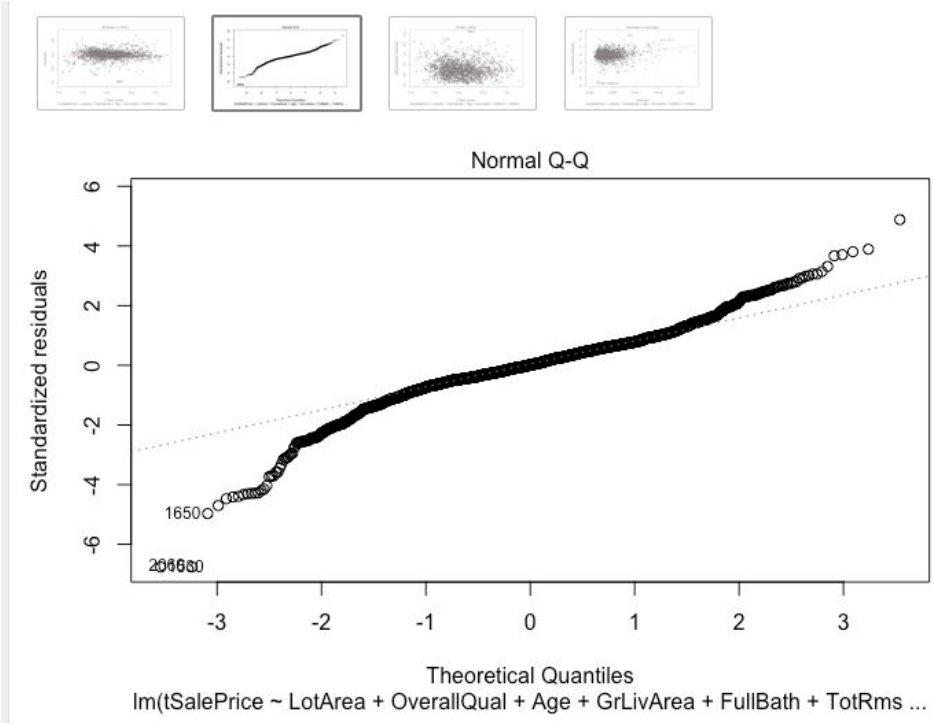
Second, I removed all the outliers in the train dataset and test dataset.
Third, I started to choose good numerical predictors. I deleted some predictors with large p-values and some predictors with large VIF. At the end, LotArea, OverallQual, YearBuilt, YearRemodAdd, GrLivArea, FullBath, TotRmsAbvGrd and GarageArea are picked.

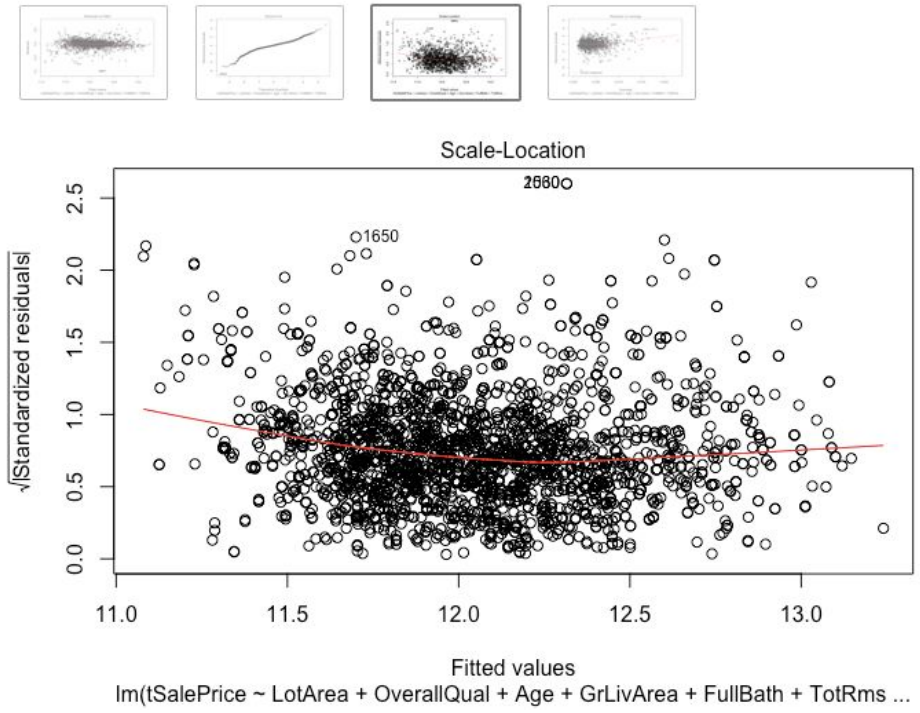
Fourth, in order to select good categorical variables, I looked at their boxplot one by one.
Fifth, I did some transformations. I combined YearBuit and YearRemodAdd into one predictor: Age. I convert predictors FullBath and GarageCars into dummy variables because it will gives more featured insight if we use them as dummy variables. I also take log over the response variables to increase the linearity.

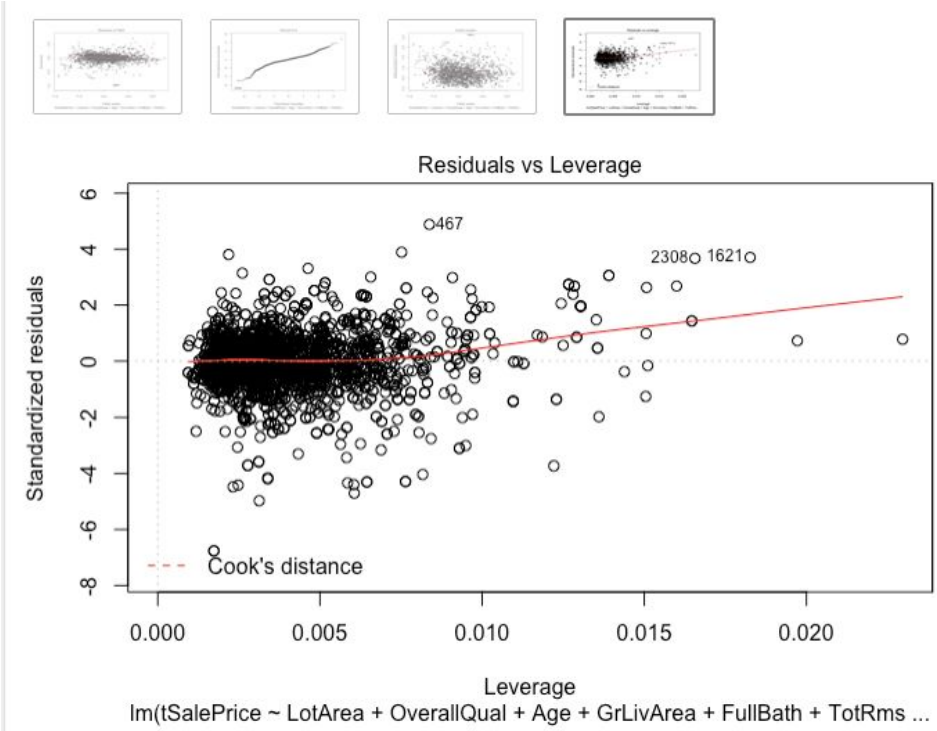
My final model is tSalePrice = 1.064e+01 + LotArea * 2.244e-06 + OverallQual * 1.325e-01 + Age * -1.176e-03 + GrLivArea * 1.474e-04 + FullBath * 4.690e-02 + TotRmsAbvGrd * 9.118e-03 + Fireplaces * 2.669e-02 + GarageCars * 9.445e-02 + BsmtFinSF1 * 1.222e-04.

The diagnostic plots are below.









I think my model is a valid model. But it still have problems with constant variance and normality. My model will be better if I spend more time on it.