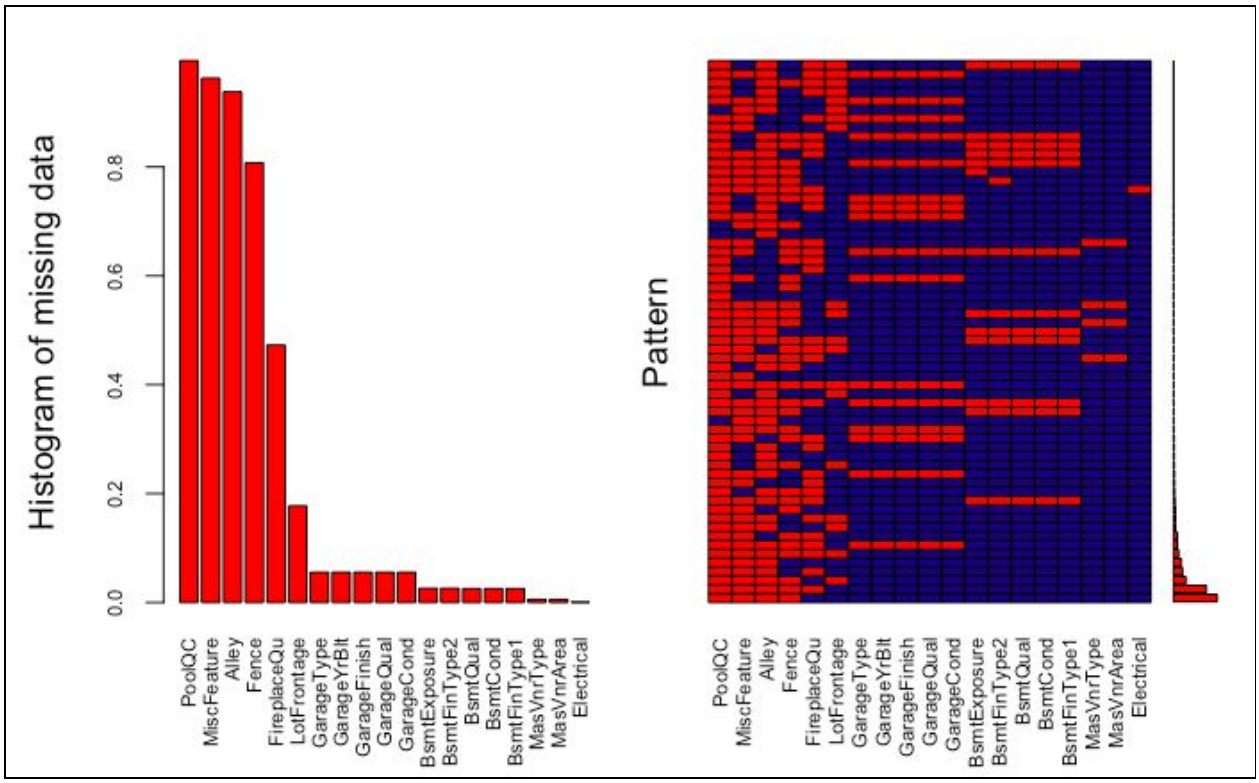
Regression Case Study

Shikhar Gupta, Akshay Tiwari, Prince Grover

Preliminary data treatment:

Missing values and variable selection:

In the raw data we have seen that for many categorical variables 'NA' means that the feature is not available. But while reading the data those entries are treated as null and we can see that in the below chart. Overall there are **19 variables** which have either "feature not available" encoded as NA or there are actually missing values.



We are removing the below variables from the data:

- Pool QC
- Misc Feature
- Alley
- Fence
- FIreplace Qu
- Id (It is just an identifier)
- Utilities (variable has the same value for all houses)

Reason: For most of the houses these features are not available and our raw analyses we found that these variables are not significant. So we have decided to remove these variables beforehand.

Also we can notice that Garage related variables are “not available” simultaneously which implies that the house doesn’t have a garage. Also in our raw analyses we found that “garage” related variables that we have mentioned below have no significant impact so we are removing them from the data:

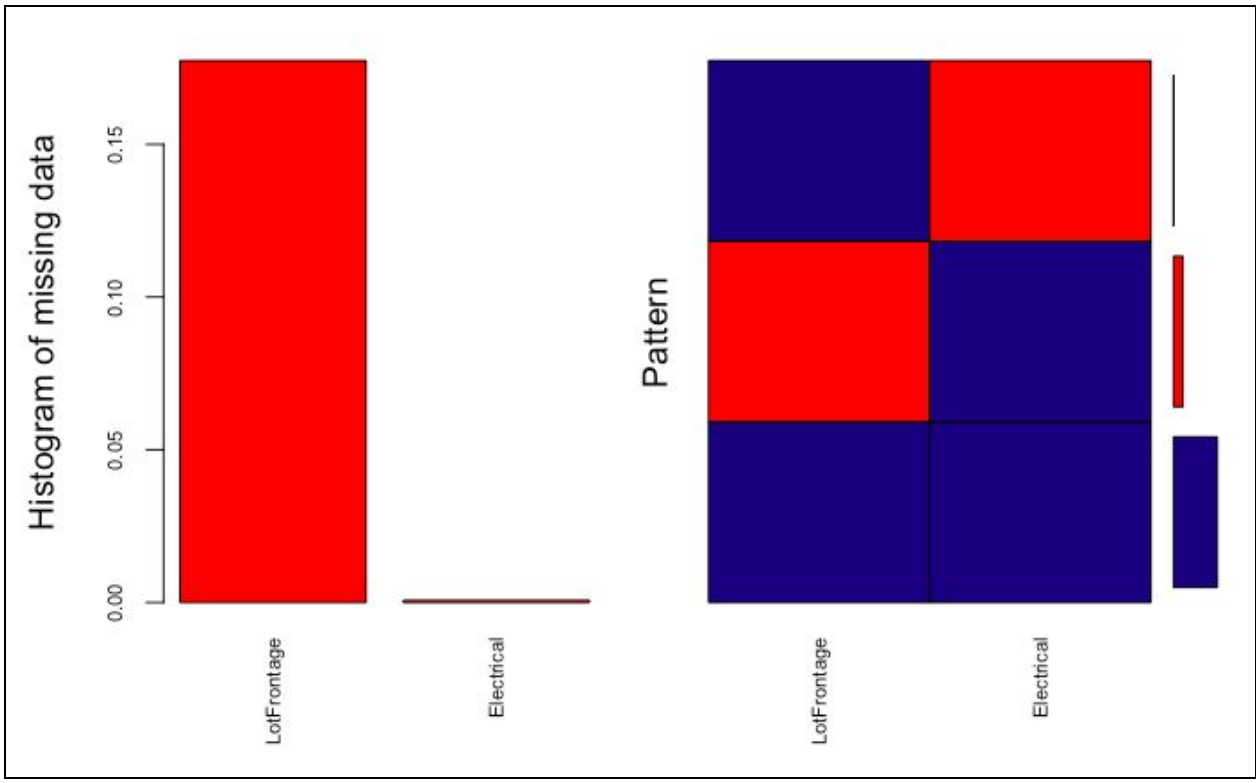
- GarageYrBlt
- GarageArea (Garage cars is a proxy for this variable)
- GarageType
- GarageFinish
- GarageQual
- GarageCond

Treatment of Masonry related variables:
For variable MasVnrArea and MasVnrType, 8 entries have NA. We are imputing these values to 0 and None. Most of the houses don’t have these features and have 0 and None as entries so we are keeping it the same for the 8 missing values.

Treatment of Basement related variables:
For the below mentioned variable, NA means that basement is absent so we have replaced NA with “absent” so that it is not considered as missing values.
Variables: BsmtQual,BsmtCond,BsmtExposure,BsmtFinType1,BsmtFinType2

MoSold variable is also converted from numeric to factor

After the above treatment our missing value chart looks like this:



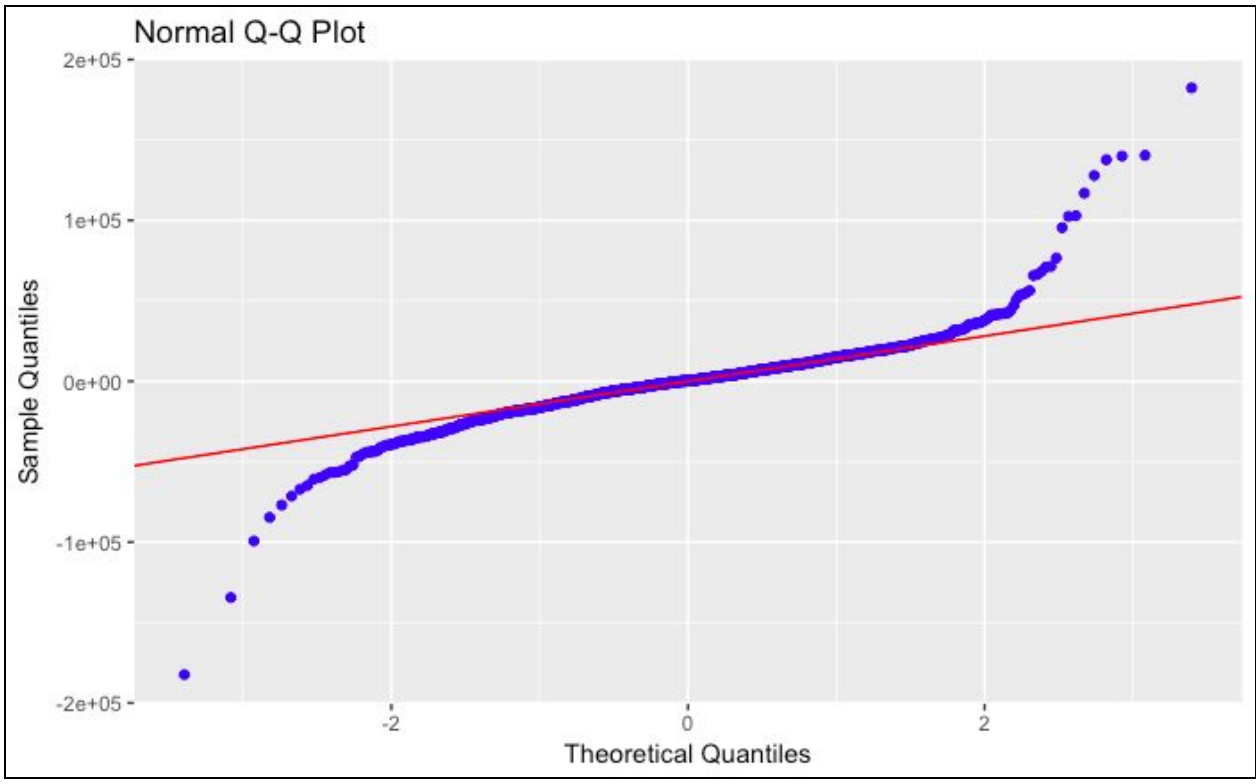
For these 2 variables we’ll use **missForest package** for missing value imputation.

Model testing and variable selection:
After converting dataframe into model matrix we have done a preliminary OLS.

OLS1:
Predictor Variables: 231

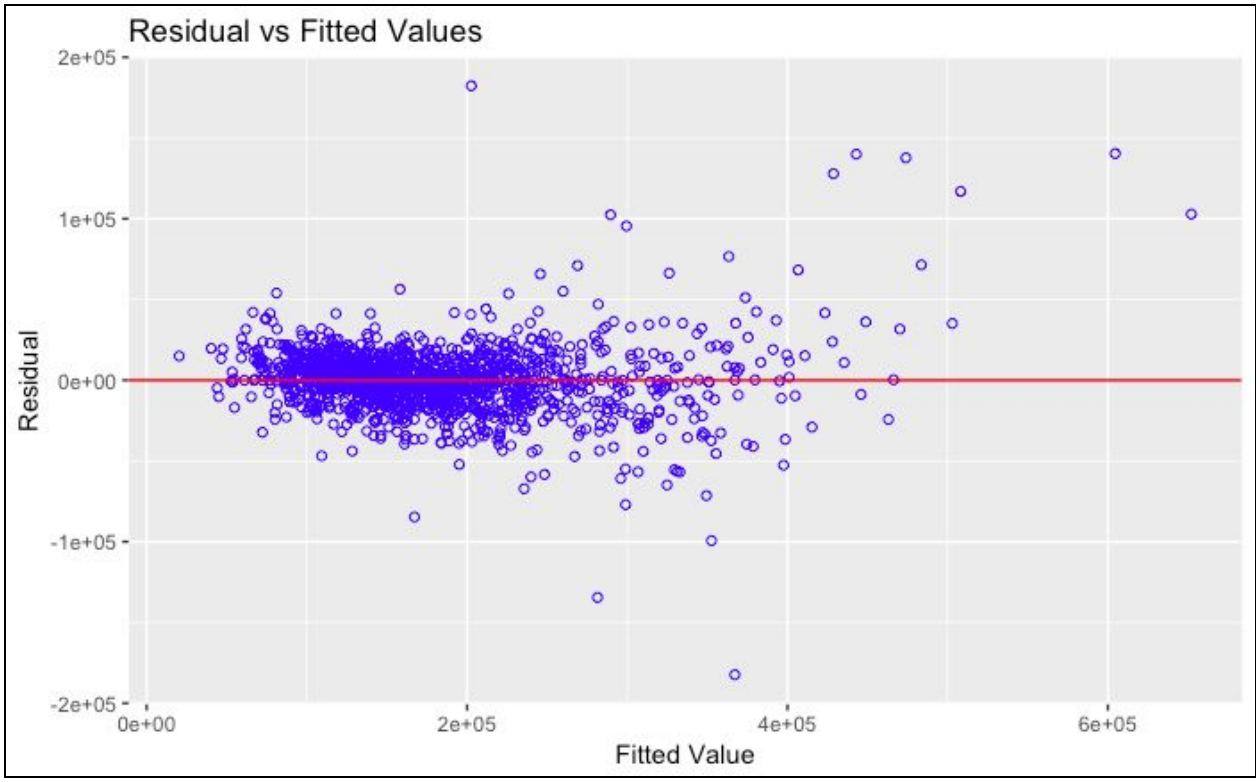
Observation:

- 174 variables are coming out to be insignificant based on p value cut off of .05



Test	Statistic	pvalue
Shapiro-Wilk	0.8555	0.0000
Kolmogorov-Smirnov	0.087	0.0000
Cramer-von Mises	121.7105	0.0119
Anderson-Darling	30.844	0.0000

- We can see that qq plot is diverging from straight line on the extremes which indicate that we need y transformation. Also it is failing the KS test which is a formal indicator of failing of normality assumption

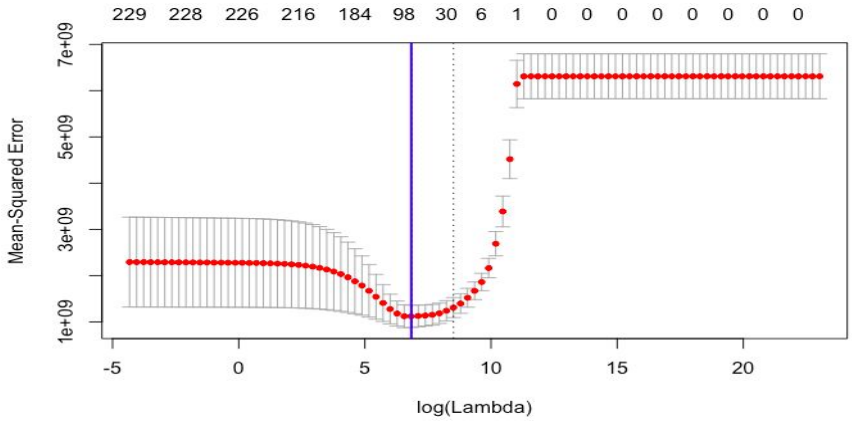


- We can see that for most of the fitted values residuals are having constant variance but for the outlier cases residuals have high variance

So we have to reduce the number of variables and implement response transformation and also treat outliers

---

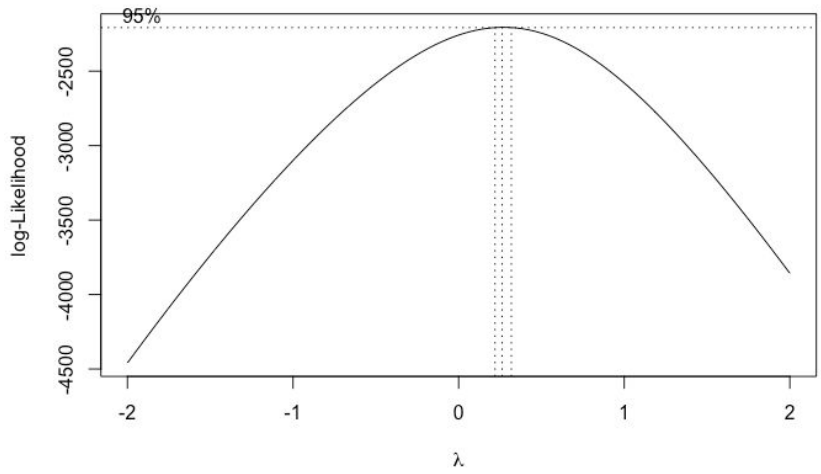
Lasso for variable selection:
Observation:



- Best lambda : 932.6033
- Number of variables reduced to 82 variables

Box Cox for response transformation:

As indicated by the preliminary OLS, we require response transformation. For that we have run Box Cox transformation to get the transformation which will be the best.



Observation:

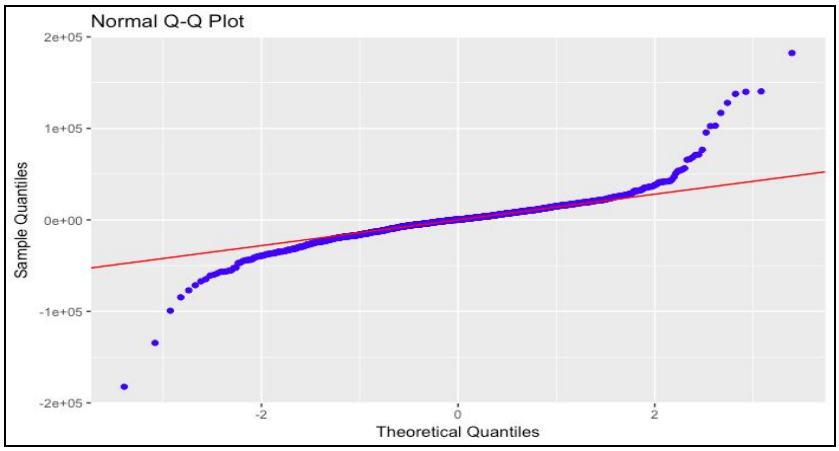
- Lambda value ~.2 which is very close to 0. This indicates that we should use log transformation for response

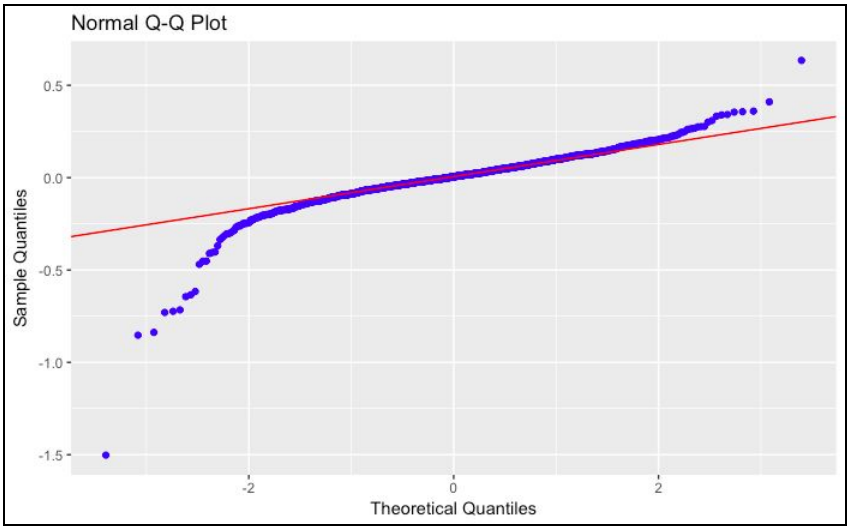
---

OLS2:

Now after transforming response variable we are running least square model.

Observation:





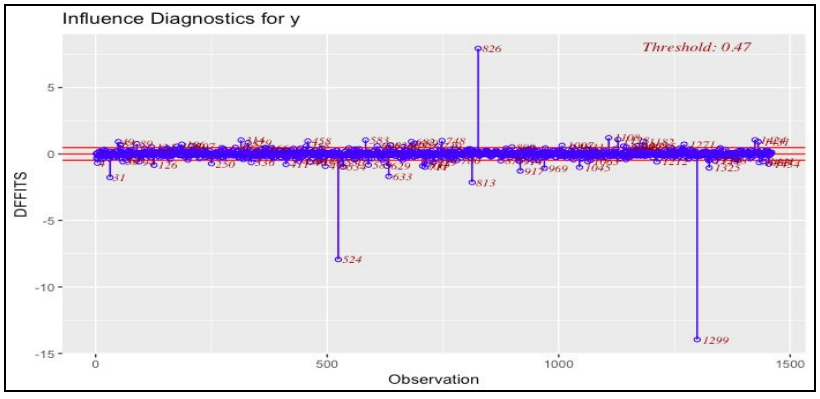
Test	Statistic	pvalue
Shapiro-Wilk	0.8528	0.0000
Kolmogorov-Smirnov	0.0931	0.0000
Cramer-von Mises	393.2536	0.0936
Anderson-Darling	25.0429	0.0000

- We can see that compared to the last OLS this qq plot has converged on the extreme right towards the line but there still exist big deviation from normality on the left extreme. KS test is still failing for normality
- We suspect that this is because of the influential points and we need to remove these observations

---

Influential observations identification:

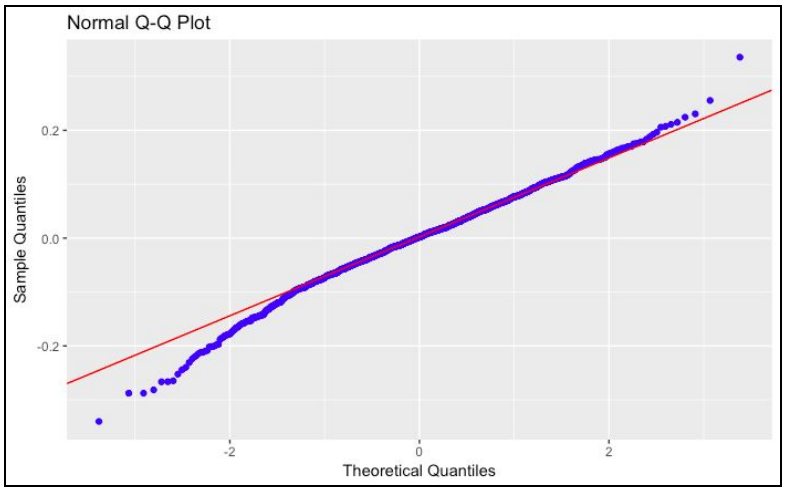
DFFITs:



- There are 74 observations that are influential for the current OLS fit

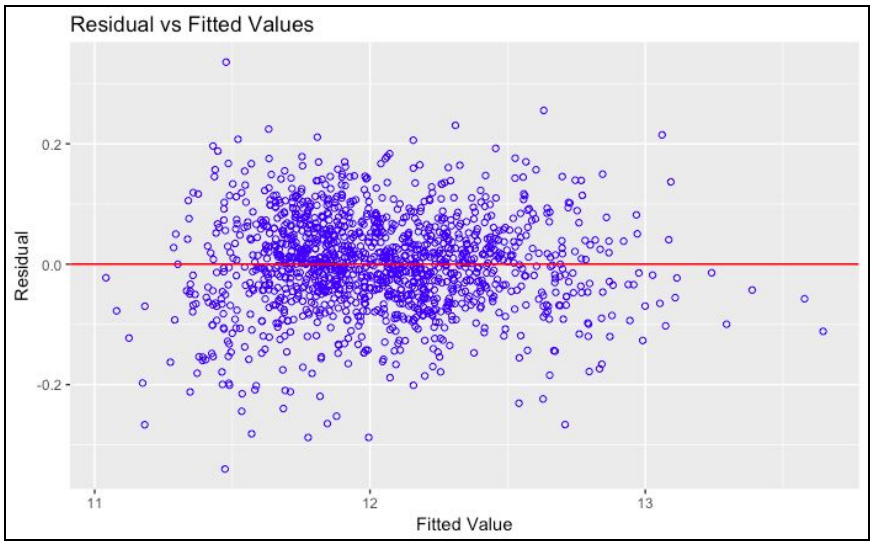
- We are removing those observations from our data

OLS3:
After outlier treatment we have tried OLS again with 82 variables.



Test	Statistic	pvalue
Shapiro-Wilk	0.9917	0.0000
Kolmogorov-Smirnov	0.0299	0.1666
Cramer-von Mises	393.4006	0.0936
Anderson-Darling	2.254	0.0000

- We can see that compared to the last OLS this qq plot has converged a lot on both extremes. This model is also satisfying KS test at a significance level of .05



- We can also see the constant variance in the above plot. This is also formally proven by the below test.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 18.6029    Df = 1     p = 1.609756e-05
```

Backward subset selection:

After backward subset selection number of variables are reduced to 46 variables at significance level of .05.

Variable selection by business logic:

We are removing following variables based on business logic and statistical significance combined:

- **StreetPave :** Street variable has only 6 values as 'Paved' across the dataset. Also, it doesn't appear statistically significant when it's effect is observed across Sale Price.

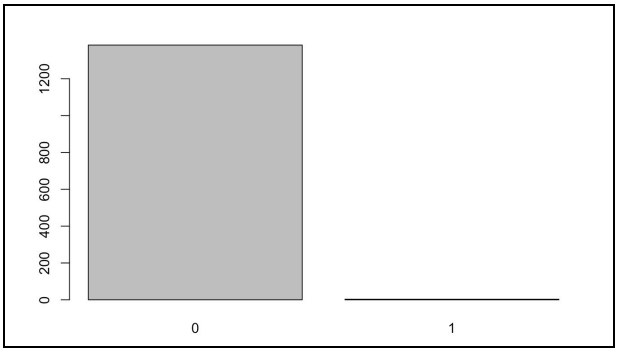
```
Grvl Pave
   6 1454
```

- **LowQualFinSF :** Mostly Single Value across dataset [21:non-zero value, rest:0]

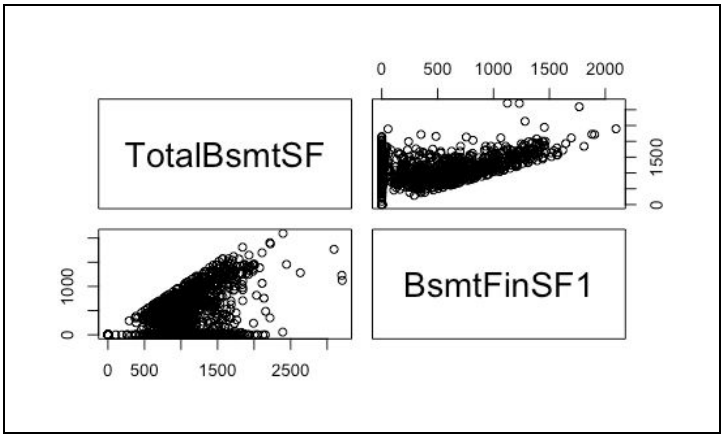
```
> table(data1$LowQualFinSF>0)

FALSE  TRUE
 1434    26
```

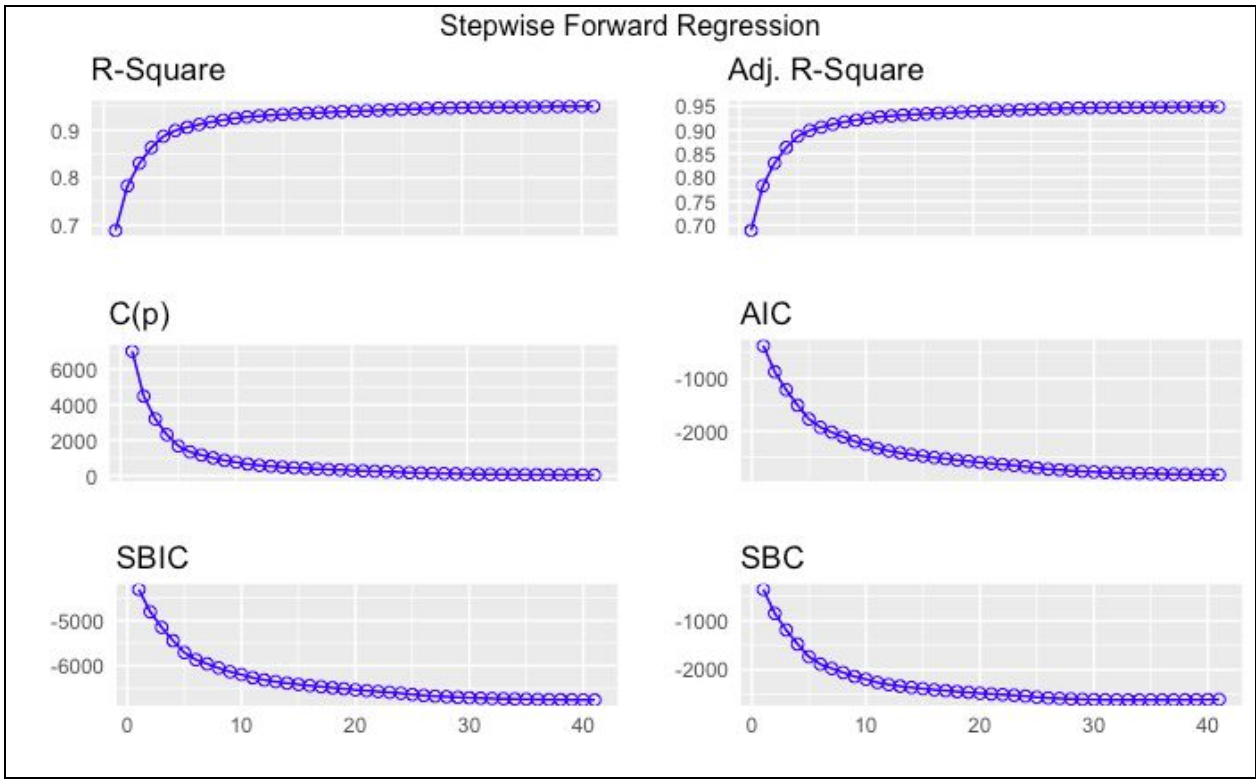
- **BsmtFinSF1 :** Almost linearly correlated with 'TotalBsmtSF'
- **LotShapeIR3:** It has 3 observations with non zero entries



- **BsmtFinType1Unf:** It doesn't seem like an important variable



Forward selection:
By forward selection on the 41 predictors and tracking the movement of adj-R2 we can observe that it saturates after a design matrix of 30 predictors.



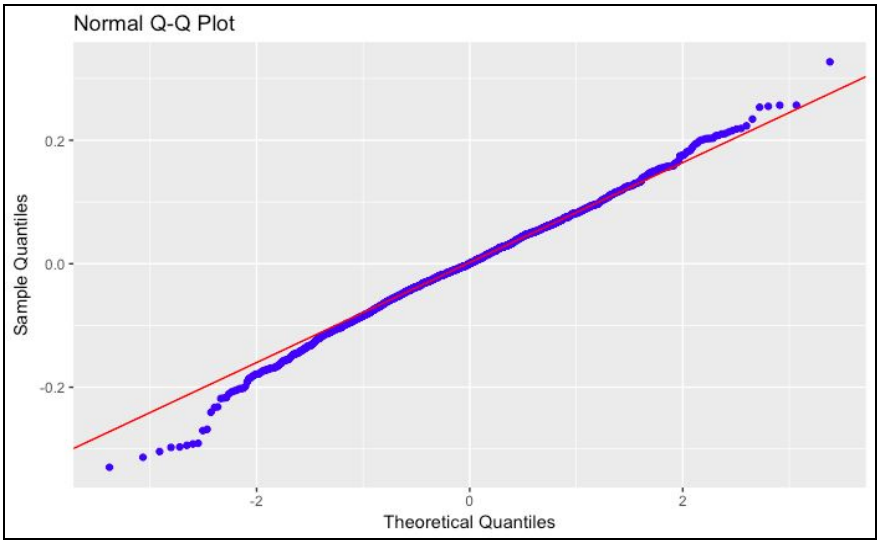
Best subset selection:
Now by using best subset selection we selected the top combination of 20 predictors. From the remaining 21 variables we have kept the ones that seemed relevant. Post that we have developed a OLS model using these 30 variables.

OLS FIT:

Observation:

```
Residual standard error: 0.08829 on 1355 degrees of freedom
Multiple R-squared:  0.9467,    Adjusted R-squared:  0.9455
F-statistic: 801.8 on 30 and 1355 DF,  p-value: < 2.2e-16
```

- **Adjusted R-squared:** .9455

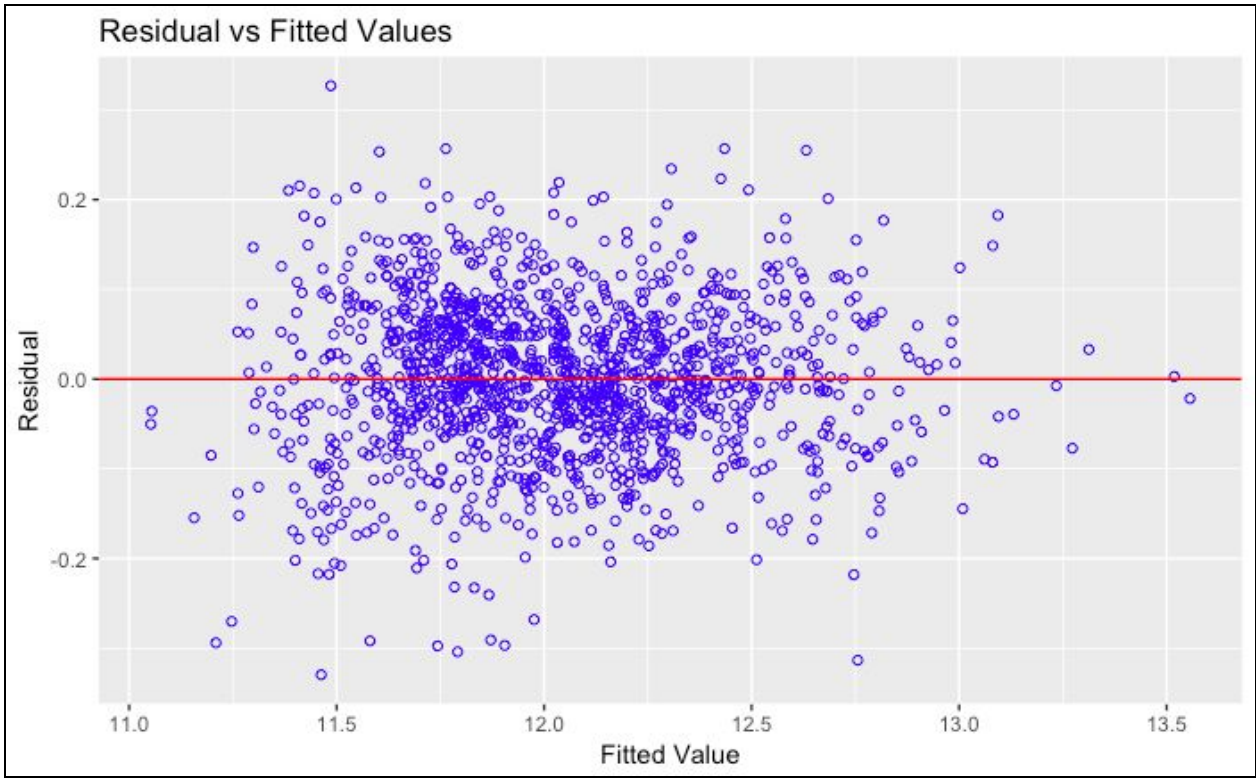


Test	Statistic	pvalue
Shapiro-Wilk	0.9945	1e-04
Kolmogorov-Smirnov	0.0293	0.1841
Cramer-von Mises	386.932	0.0919
Anderson-Darling	1.453	9e-04

- Model is satisfying the normality test

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.115e+00	3.884e-01	8.020	2.27e-15	***
BldgTypeDuplex	-1.059e-01	1.454e-02	-7.281	5.59e-13	***
BsmtFullBath	5.572e-02	5.127e-03	10.867	< 2e-16	***
BsmtQualEx	5.688e-02	1.161e-02	4.900	1.07e-06	***
Condition1Norm	4.403e-02	7.362e-03	5.981	2.84e-09	***
BsmtExposureGd	6.398e-02	9.317e-03	6.867	9.93e-12	***
FunctionalTyp	5.671e-02	1.042e-02	5.443	6.20e-08	***
Exterior1stBrkFace	9.550e-02	1.357e-02	7.036	3.13e-12	***
NeighborhoodStoneBr	1.241e-01	2.012e-02	6.165	9.27e-10	***
NeighborhoodNridgHt	7.693e-02	1.327e-02	5.797	8.38e-09	***
NeighborhoodSomerst	7.329e-02	1.144e-02	6.405	2.06e-10	***
MSZoningRM	-6.207e-02	8.493e-03	-7.308	4.61e-13	***
BldgTypeTwnhs	-1.241e-01	1.553e-02	-7.990	2.86e-15	***
LotArea	4.271e-06	4.785e-07	8.927	< 2e-16	***
NeighborhoodCrawfor	1.317e-01	1.457e-02	9.045	< 2e-16	***
GarageCars	4.964e-02	4.619e-03	10.746	< 2e-16	***
TotalBsmtSF	1.136e-04	7.792e-06	14.581	< 2e-16	***
OverallCond	5.070e-02	2.801e-03	18.098	< 2e-16	***
OverallQual	4.960e-02	3.330e-03	14.895	< 2e-16	***
YearBuilt	3.039e-03	1.611e-04	18.859	< 2e-16	***
GrLivArea	2.800e-04	7.202e-06	38.878	< 2e-16	***
FunctionalSev	-4.098e-01	8.904e-02	-4.603	4.56e-06	***
ScreenPorch	1.850e-04	4.403e-05	4.203	2.81e-05	***
SaleTypeNew	3.131e-02	1.024e-02	3.059	0.002264	**
KitchenQualTA	-2.785e-02	1.052e-02	-2.647	0.008203	**
NeighborhoodEdwards	-5.130e-02	1.021e-02	-5.025	5.71e-07	***
BldgTypeTwnhsE	-5.080e-02	1.032e-02	-4.923	9.58e-07	***
FoundationPConc	2.911e-02	7.504e-03	3.879	0.000110	***
Fireplaces	2.817e-02	4.639e-03	6.073	1.63e-09	***
KitchenQualGd	-3.382e-02	1.007e-02	-3.358	0.000807	***
YearRemodAdd	7.761e-04	1.868e-04	4.154	3.47e-05	***



- Residuals are having non constant variance as can be seen in the plot and formally proved in the NCV test below.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 20.38235    Df = 1     p = 6.341193e-06
```

- Errors are also not correlated as shown by Durbin-Watson test.

```
 lag Autocorrelation D-W Statistic p-value
   1      0.01715884      1.965074    0.51
Alternative hypothesis: rho != 0
```

Prediction for Monty:
Using the above model we have predicted the following range of Sales Price for Morty. So the expected price for the given features of Monty's house is **$156,137.**

fit	lwr	upr
156137.8	131600.7	185249.7

Recommendations for Morty:

- **Basement Quality**
  It is 'Good'. We recommend him to make it "Excellent"
  It can lead to **$ 4961 to $ 11857** sale price for the house.

Calculation - Expected value of coefficient for ***BsmtQualEx = 0.05688***

Therefore, change in price is calculated as ***exp(log(SalePrice)+(0.05688)) - SalePrice***
Similar calculations are used on **2.5% and 97.5% quantiles** of coefficients to calculate 95% CI.

- **Exterior 1st**
  Make Exterior 1st '**Brick Face**' from current 'Vinyl Face'
  It can lead to **$ 10195 to $ 18570** sale price for the house

- **Basement Exposure**
  Have exposure in basement
  It can lead to **$ 6686 to $ 12250** sale price for the house

- **Overall Condition**
  Increasing overall condition by 1 grade can lead to **$ 6611 to $ 8250** increase in house sale price

- **Fireplaces**
  By having a fireplace, sale price can be increase b/w **$ 2753 to $ 5430**

- **Basement Quality**
  Increasing basement quality to **'Excellent'** from good can increase price from **$ 4960 to $ 11848**

In conclusion, Morty can increase sale price from **$ 36166 to $ 68205,** if he does what **we** suggest him.

---

2. **Predictive Modeling:**

For building the predictive model we have done some data treatment:

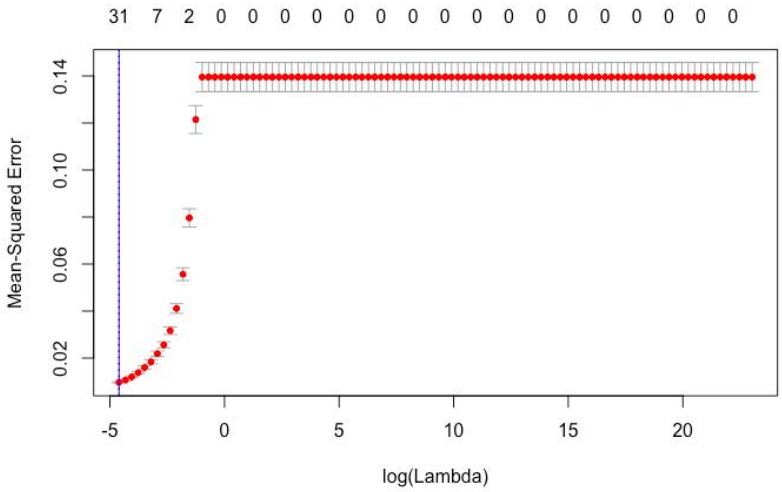
- Created new variables:
  - Yrdiff: yrsold - yrbuilt
  - Bsmtarea = TotalBsmtSF + X1stFlrSF + X2ndFlrSF
- Filter predictors to be 46 variables which we had obtained during the interpretation process after Lasso and backward subset selection
- Ran OLS and removed outlier observations using DFFITS: 74 observations

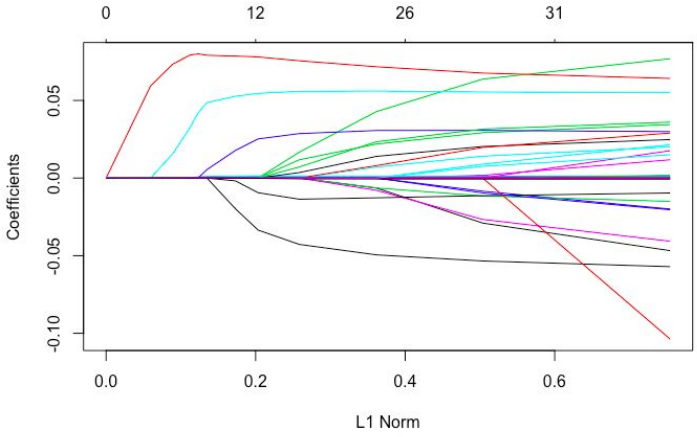
Model1: OLS

MSPE: 219456436

Model 2: Lasso

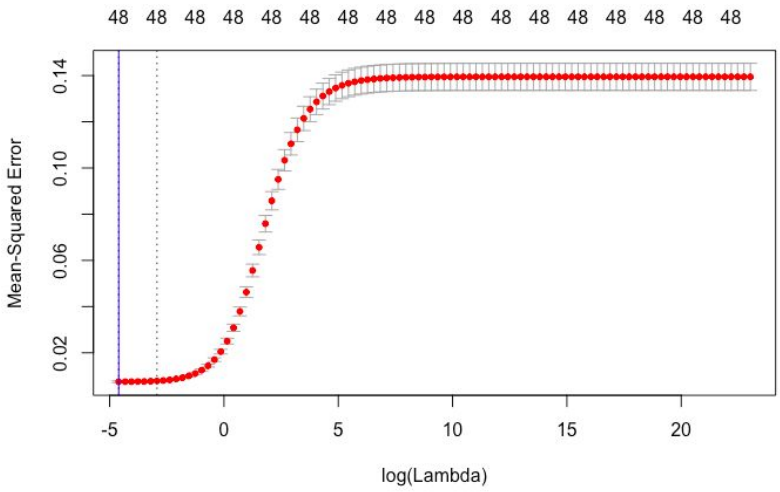
MSPE: 321593710

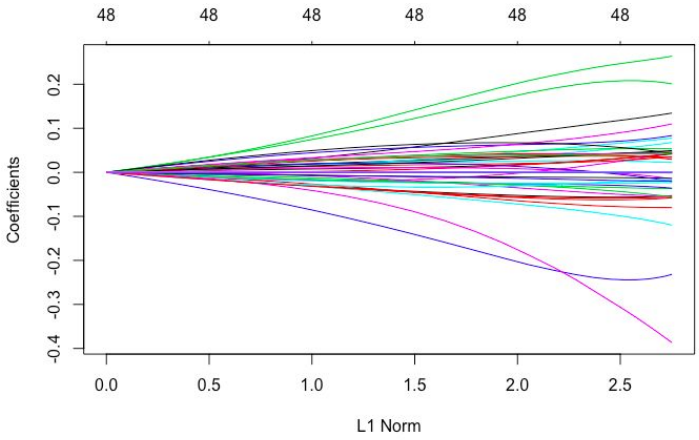




Model 3: Ridge

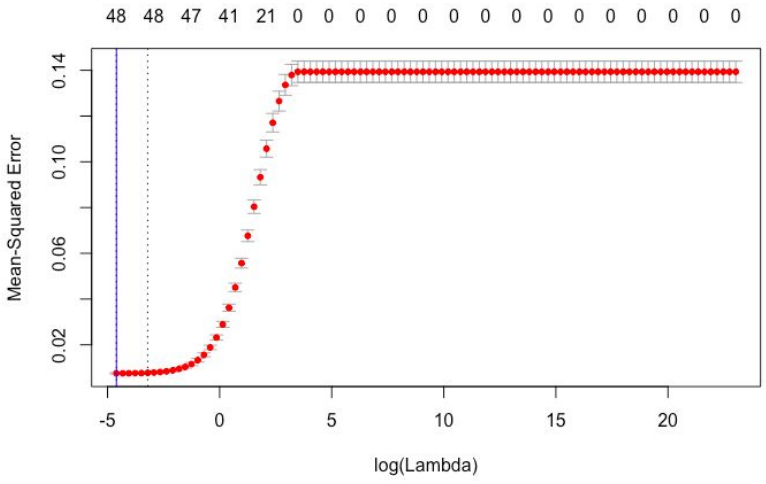
MSPE: 218523464

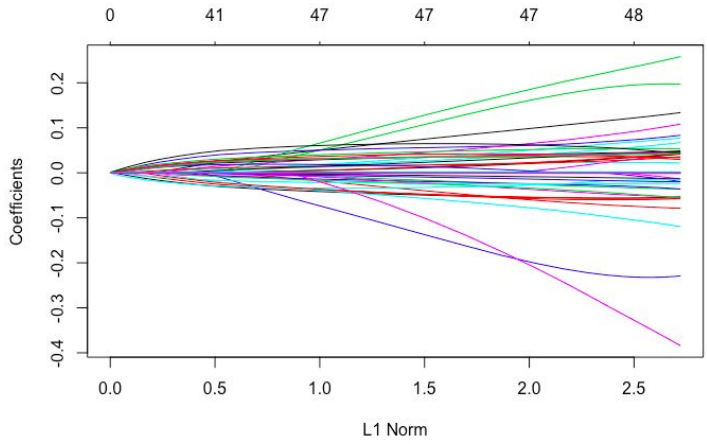




Model 4:Elastic Net

MSPE: 218278873 for alpha = 0.01





Model selection:

MSPE comparison:

Ridge <dbl>	Lasso <dbl>	Elastic.Net <dbl>	ols <dbl>
218523464	321593710	218278873	219456436

From the above comparison we can see that Ridge, OLS and Elastic Net are almost at par in terms of MSPE with Elastic Net (alpha = .01) having the minimum MSPE. All these models have been implemented using 48 predictors.