Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables:

Highlight of the paper:

Removed features with more than forty percent missing or null values (PoolQC", "MiscFeature", "Alley", "Fence", "FireplaceQu).

In variables that have less than forty percent of missing/null values these values are replaced by their mode in categorical variables and by their mean in continuous variables.

The variable "SalePrice" is skewed to the left, because of this a log transformation is necessary to make the variable normally distributed.

Page 10 error, we should keep high correlation variables

Dropped explanatory variables highly correlated with other

Removed near zero variance variables

Final variables:

1) GrLivArea,
2) Neighborhood,
3) OverallQual,
4) MSSubClass,
5) TotalBsmtSF,
6) YearRemodAdd,
7) ExterQual,
8) BsmtCond,
9) FullBath,
10) GarageQual,
11) LotArea,
12) KitchenQual.

Highly significant F-statistic and R squared.

Dropped FullBath predictor for p-value too high

Tested Homoscedasticity with Breusch Pagan and corrected with Box Cox approach

(Cross-Validation non la ho considerata)

Tested final model (Original model - FullBath)

Conclusion with summary of the procedure