Final Project: Data Analysis, Regression MAP535

Honghao YU & Jiayu GAN

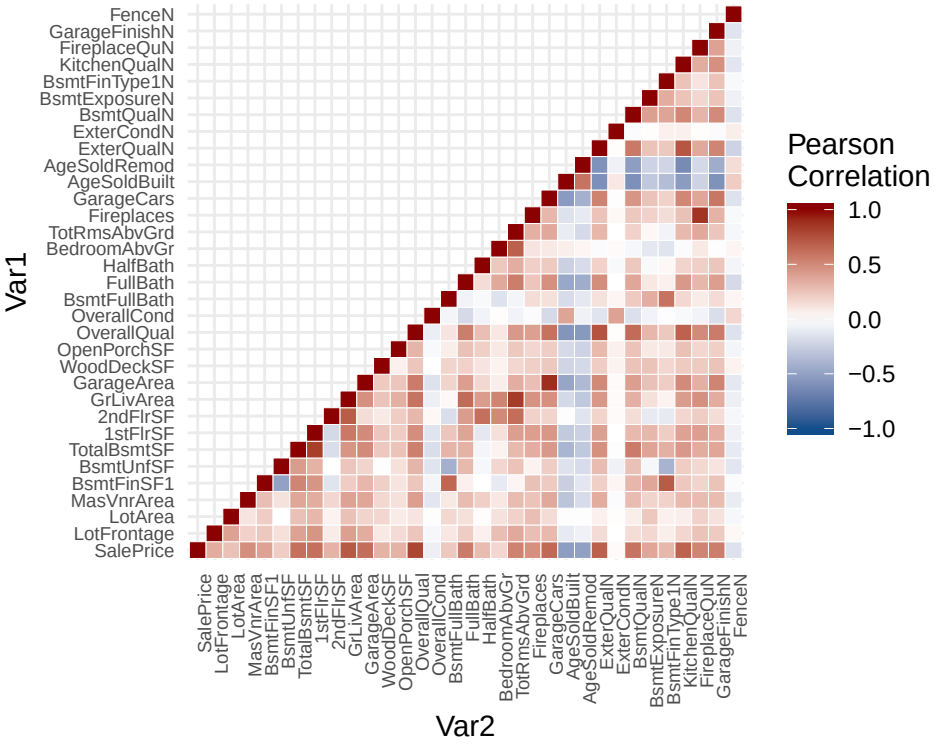
12/20/2019

I. Introduction

In this project we seek to use linear regression model to quantify the relationship between house prices and a comprehensive list of features that describe the houses, as well as make accurate predictions for future sales. Among these features are the detailed conditions of the houses themselves (areas, facilities, style, well-being, ages, etc.) and the characteristics of their surroundings (neighborhood, access to main roads, etc.). We would like to test 1) whether these features are valid for prediction, i.e. whether their coefficients are statistically significantly different from 0, 2) which are the most significant and 3) how they impact house prices respectively revealed by their coefficients.

II. Exploratory Data Analysis

Before looking into the details, we performed a rough filtering to leave out problematic variables. We noticed that some categorical variables with a level "NA" for "No Such Thing" might be taken by R as missing values, thus we replaced "NA" by "Zero" before reading in the data. We then dropped variables with 20% or more missing entries and those with near-zero variance (for both numeric and factor variables) at the frequency cut of 90/10. For the remaining missing values, we performed knn-mean imputation for quantitative variables and most-frequent imputation for factor variables. Two variables were created to indicate the ages of the houses when sold since they were built and since they were last remodeled. We also transformed some supposedly factor variables that described qualities and conditions in an ordered fashion into ordinal numeric variables. These primary modifications being done, we performed a graphical analysis for the numeric variables as below:

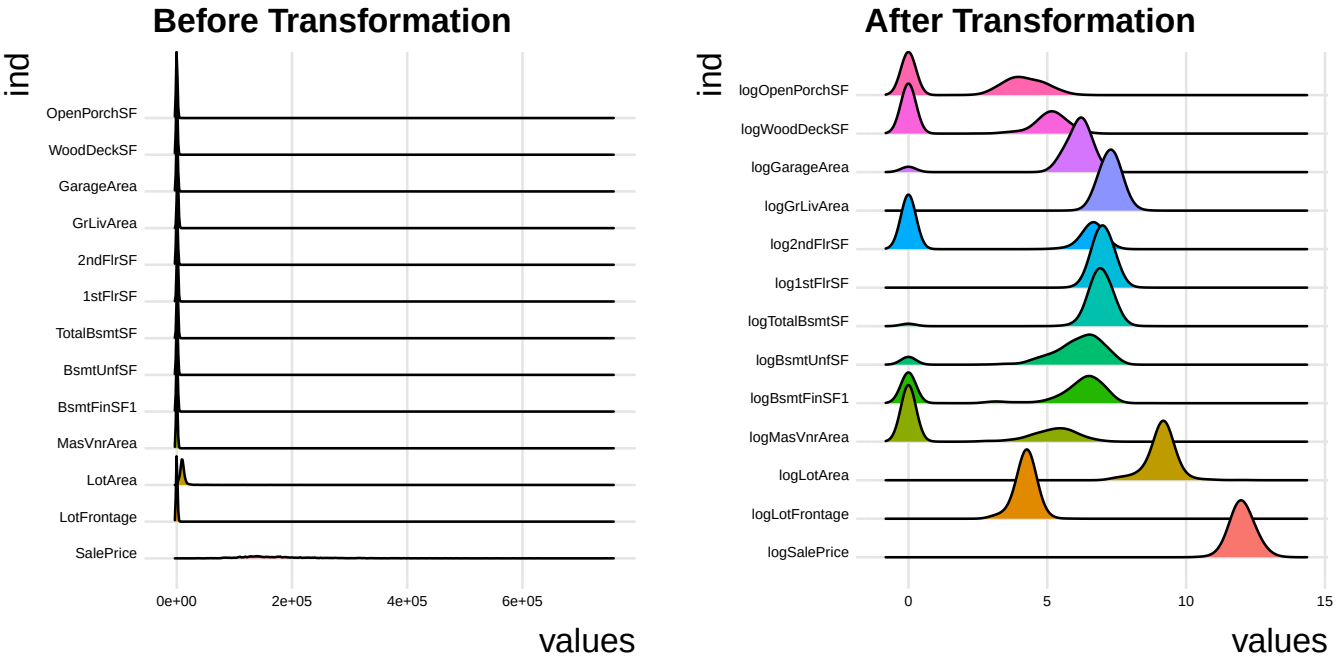


- Figure 1. Heatmap of cross-correlation

whereas for the factor variables, we did ANOVA tests against house prices as below:

```
##                Df    Sum Sq   Mean Sq F value   Pr(>F)
## MoSold         11 6.652e+10 6.047e+09   3.113 0.000371 ***
## MSSubClass     14 2.252e+12 1.608e+11  82.814  < 2e-16 ***
## MSZoning        4 2.282e+11 5.705e+10  29.376  < 2e-16 ***
## LotShape        3 1.631e+11 5.437e+10  27.991  < 2e-16 ***
## LotConfig       4 2.195e+10 5.488e+09   2.826 0.023749 *
## Neighborhood   24 3.040e+12 1.267e+11  65.227  < 2e-16 ***
## HouseStyle      7 1.877e+10 2.681e+09   1.380 0.209539
## RoofStyle       5 1.674e+11 3.347e+10  17.235  < 2e-16 ***
## Exterior1st    14 2.119e+11 1.513e+10   7.792 4.45e-16 ***
## Exterior2nd    14 5.391e+10 3.851e+09   1.983 0.016090 *
## MasVnrType      3 1.485e+11 4.949e+10  25.480 4.95e-16 ***
## Foundation      5 1.019e+11 2.039e+10  10.498 6.69e-10 ***
## HeatingQC       4 6.918e+10 1.730e+10   8.905 4.26e-07 ***
## GarageType      6 5.974e+10 9.957e+09   5.127 3.22e-05 ***
## Residuals    1341 2.605e+12 1.942e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

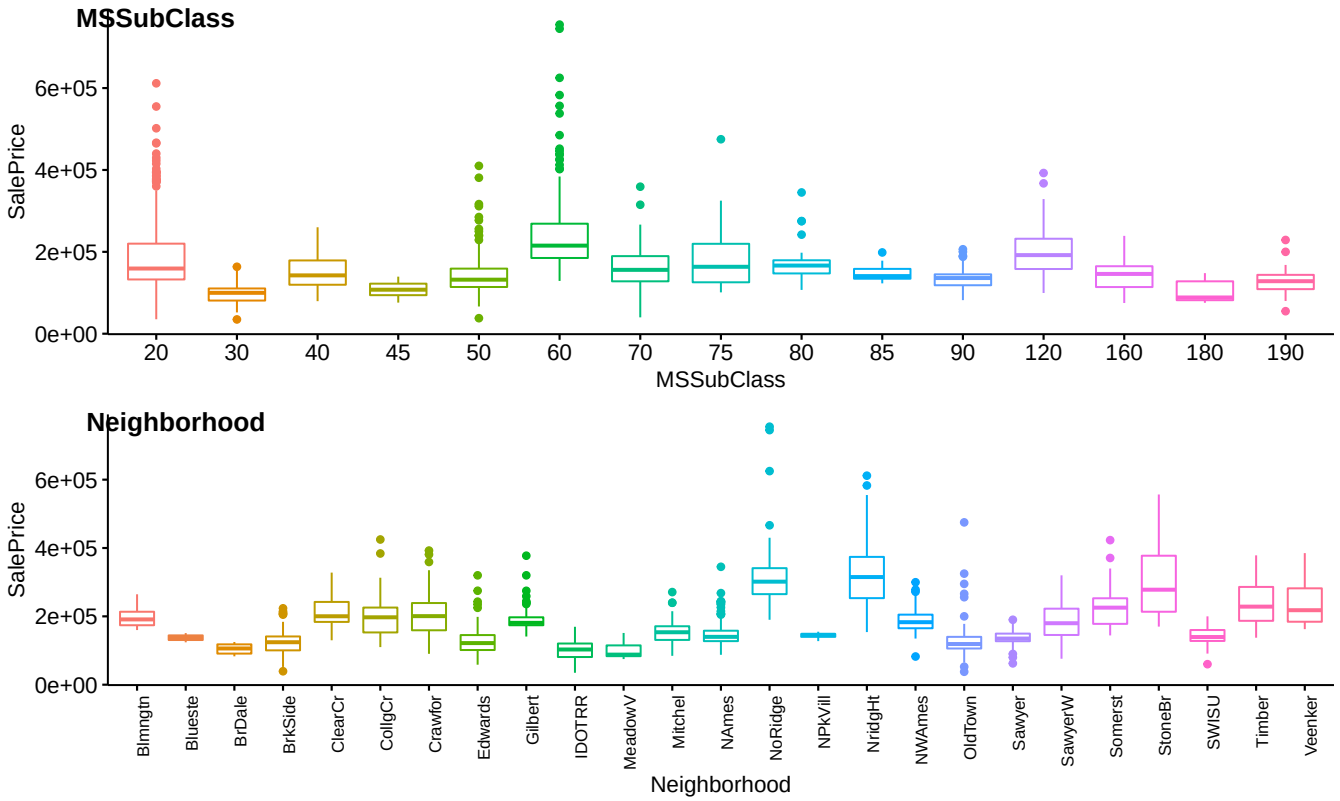
A quick glance would tell us the house sale prices were highly relevant to OverallQual, ExterQualN, KitchenQualN, GrLivArea, TotalBsmtSF, BsmtQualN, 1stFlrSF, GarageArea, GarageCars, etc. (mostly positively correlated except for AgeSoldBuilt and GarageCars) among numeric variables and MSSubClass, Neighborhood, MSZoning, LotShape, MasVnrType, etc. among factor variables. Hence it's reasonable to assume that these variables are valid candidates. We noticed that the distributions of SalePrice and the variables measuring areas were largely skewed, thus we performed a log-transformation (note: since values close to zero will result in aberrant negative values after log-transformation, we did $log(x+1)$ instead of $log(x)$) on them. The comparison of their distributions before and after the transformation is shown as below:



- Figure 2. Histrogram of numeric features

We noticed that there were big clusters of observations at zero value for logOpenPorchSF, logWoodDeckSF, log2nFlrSF, logBsmtFinSF1 and logMasVnrArea due to the lack of corresponding parts in those houses and required attention in the model building phase.

We were also interested in how the sale prices differ across different levels of the factor variables. Here we display the box plots for two of them:



* Figure 3. Boxplot Factors/SalePrice

We can see SalePrice indeed has significant variations across different levels of these two variables. Thus they are reasonable predictors.

III. Modeling and Diagnostics

We started with multiple linear regression model. The relevant features were selected by conducting stepwise feature selection in both directions. We chose the model that yielded the smallest AIC.

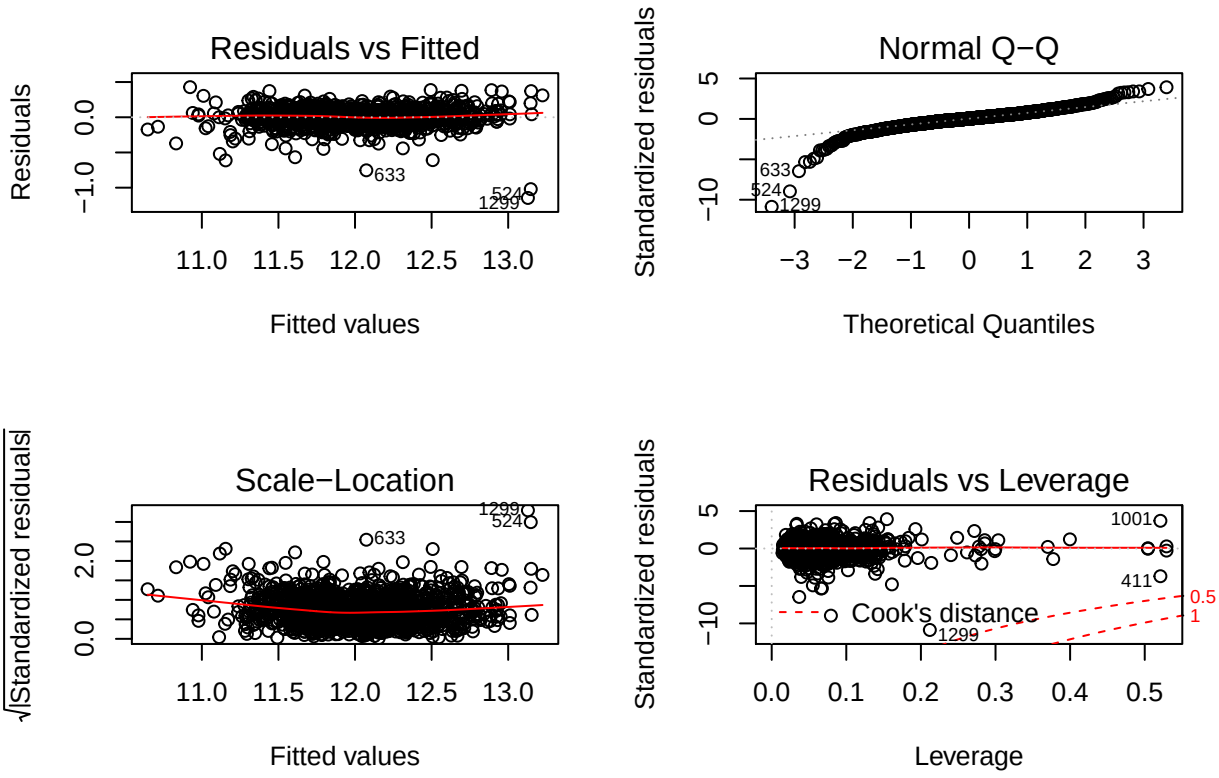
```
##
## Call:
## lm(formula = logSalePrice ~ logLotArea + logBsmtFinSF1 + logBsmtUnfSF +
##     logTotalBsmtSF + logGrLivArea + logWoodDeckSF + OverallQual +
##     OverallCond + BsmtFullBath + FullBath + HalfBath + GarageCars +
##     AgeSoldBuilt + AgeSoldRemod + WithBsmt + BsmtQualN + BsmtExposureN +
##     KitchenQualN + FireplaceQuN + GarageFinishN + WithFence +
##     FenceN + MSSubClass + MSZoning + LotShape + LotConfig + Neighborhood +
##     Exterior1st + MasVnrType + Foundation + HeatingQC, data = train_agg)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.14586 -0.05191  0.00406  0.05940  0.42720
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
##
## Residual standard error: 0.1187 on 1362 degrees of freedom
## Multiple R-squared:  0.9176, Adjusted R-squared:  0.9117
## F-statistic: 156.4 on 97 and 1362 DF,  p-value: < 2.2e-16
```

Note that we didn't print the coefficients here but only provided the model instead.

As mentioned, we have transformed features describing quality and condition into ordinal and left those without

ordinality as they were. The model consisted of 26 features, though with one-hot coding there appeared to be 93 in total. The model took in a factor as long as one modality is significant. For reference, we conducted LASSO regression (with one-hot coding) with penalization on the number of features and obtained another set of 68 features which yielded the optimal RMSE. The resulted feature sets were roughly the same for ordinary linear model and for LASSO except that the irrelevant modelatities were removed in Lasso. The LASSO model was thus more compact.

From the diagnostic plots and related hypothesis tests below we observed immediately that while Postulate 1 and 3 were satisfied, 2 (constant-variance errors) and 4 (gaussian errors) did not hold because of a few atypical points. We thus used hat value, cook's distance and Bonferroni p-value to identify atypical observations and to decide whether to remove these points.

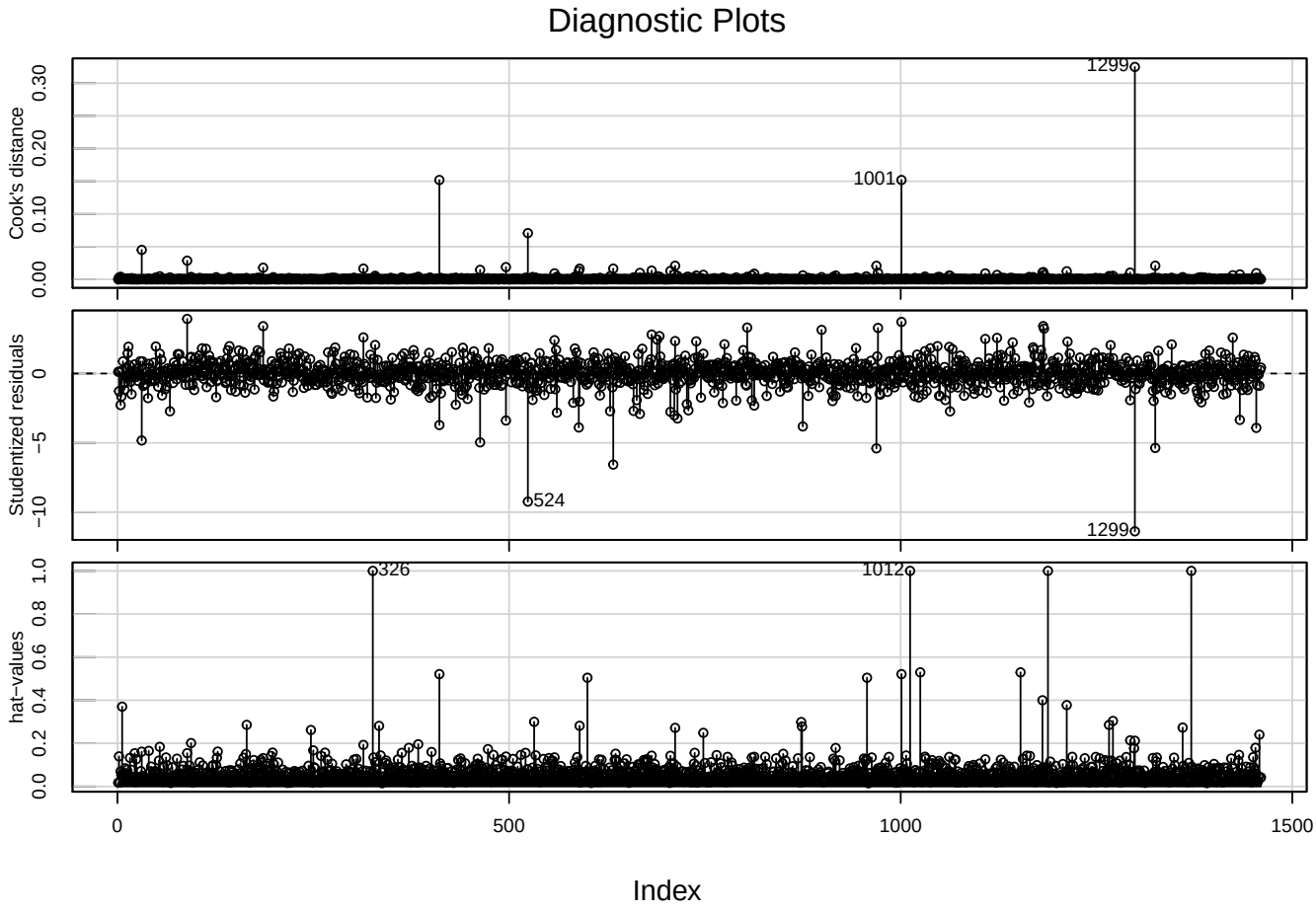


- Figure 4a. Diagnostic plots of our first model(I)

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 42.54637, Df = 1, p = 6.9026e-11
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.04080716      1.918245   0.096
##  Alternative hypothesis: rho != 0
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals((select.variables.both))
## W = 0.88907, p-value < 2.2e-16
```



* Figure 4b. Diagnostic plots of our first model(II)

```
##        rstudent unadjusted p-value Bonferroni p
## 1299 -11.378602         1.0132e-28   1.4752e-25
## 524   -9.237717         9.3745e-20   1.3649e-16
## 633   -6.567612         7.2565e-11   1.0566e-07
## 969   -5.396515         8.0061e-08   1.1657e-04
## 1325  -5.368113         9.3440e-08   1.3605e-04
## 463   -4.964762         7.7486e-07   1.1282e-03
## 31    -4.829736         1.5218e-06   2.2158e-03
```

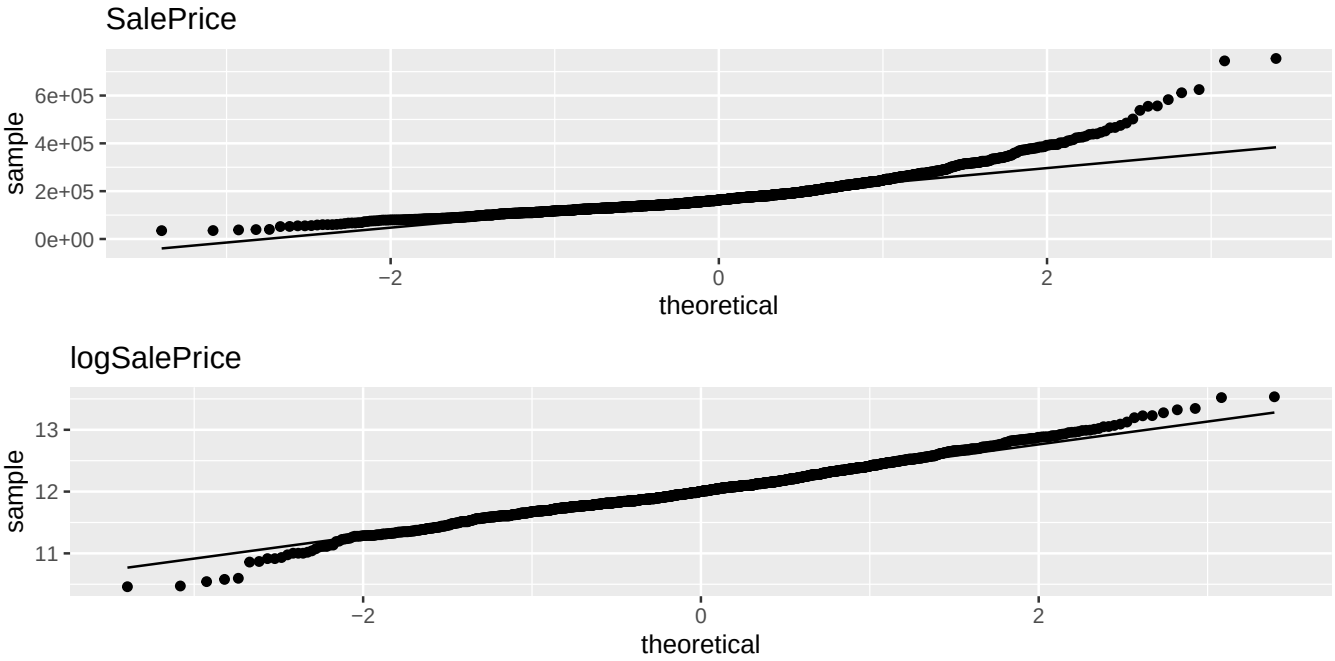
We removed all the outliers with Bonferroni p-value lower than 0.05. Some of them, though being outliers, didn't have strong leverage effect and didn't necessarily need to be removed. Nevertheless, we chose to remove the outliers in a proactive way to make data cleaner. After the removal of these atypical observations we saw that P2 was satisfied and the adjusted R-square raised significantly.

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4666979, Df = 1, p = 0.49451
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals((FModel))
## W = 0.98139, p-value = 1.077e-12
```

However, P4 was still not satisfied even after we removed all the outliers. We tried transforming features in other way or filtering out features more aggressively, but none of them worked. Then we performed normality test on the response variable SalePrice and its logarithmic form and found neither of them to be gaussian. We considered this issue as an innate feature of house prices and decided to ignore it as it wouldn't violate the validity of OLSE

estimates.



* Figure 5. QQ-plot of response variable SalePrice & logSalePrice

V. Final Models

We obtained our final compact model after hand-crafting features, shrinking model and removing atypical observations. We found 25 features in total to be relevant to house prices. Among them were the areas of different parts of a house (logLotArea, logBsmtFinSF1, logTotalBsmtSF, logGrLivArea, GarageCars), the number and conditions of facilities (BsmtFullBath, FullBath, HalfBath, WithBsmt, BsmtQualN, BsmtExposureN, KitchenQualN, FireplaceQuN, HeatingQC), the overall conditions of the house (OverallQual, OverallCond, AgeSoldBuilt, MSSubClass, MSZoning, LotShape, LotConfig, Exterior1st, MasVnrType, Foundation) and types of their neighborhoods. We also removed 19 atypical observations to satisfy the constant-variance error assumption.

```
##
## Call:
## lm(formula = logSalePrice ~ logLotArea + logBsmtFinSF1 + logBsmtUnfSF +
##     logTotalBsmtSF + logGrLivArea + logWoodDeckSF + OverallQual +
##     OverallCond + BsmtFullBath + FullBath + HalfBath + GarageCars +
##     AgeSoldBuilt + AgeSoldRemod + WithBsmt + BsmtQualN + BsmtExposureN +
##     KitchenQualN + FireplaceQuN + GarageFinishN + WithFence +
##     FenceN + MSSubClass + MSZoning + LotShape + LotConfig + Neighborhood +
##     Exterior1st + MasVnrType + Foundation + HeatingQC, data = train_agg[-c(31,
##     326, 411, 463, 496, 524, 589, 633, 813, 875, 969, 1001, 1012,
##     1072, 1188, 1299, 1325, 1371, 1454), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39287 -0.05435  0.00265  0.05722  0.37691
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.1783672  0.1643244  43.684  < 2e-16 ***
## logLotArea           0.0961916  0.0099822   9.636  < 2e-16 ***
## logBsmtFinSF1        0.0075916  0.0013057   5.814 7.60e-09 ***
## logTotalBsmtSF       0.1081499  0.0144662   7.476 1.37e-13 ***
## logGrLivArea         0.3799246  0.0202996  18.716  < 2e-16 ***
## OverallQual          0.0503057  0.0039291  12.804  < 2e-16 ***
```

```
## OverallCond          0.0475887  0.0031999  14.872  < 2e-16 ***
## BsmtFullBath         0.0235100  0.0070417   3.339 0.000865 ***
## HalfBath             0.0328422  0.0082148   3.998 6.74e-05 ***
## GarageCars           0.0405767  0.0054012   7.513 1.05e-13 ***
## AgeSoldBuilt        -0.0022642  0.0003034  -7.462 1.52e-13 ***
## WithBsmt            -0.7067313  0.1015005  -6.963 5.19e-12 ***
## BsmtQualN            0.0279777  0.0071033   3.939 8.61e-05 ***
## BsmtExposureN        0.0190638  0.0033303   5.724 1.28e-08 ***
## KitchenQualN         0.0249098  0.0064093   3.887 0.000107 ***
## FireplaceQuN         0.0083320  0.0019753   4.218 2.63e-05 ***
## MSSubClass90        -0.1060878  0.0176798  -6.000 2.52e-09 ***
## MSSubClass160       -0.1078215  0.0245786  -4.387 1.24e-05 ***
## MSSubClass190       -0.0789667  0.0226805  -3.482 0.000514 ***
## MSZoningFV           0.2501732  0.0523052   4.783 1.92e-06 ***
## MSZoningRH           0.2179833  0.0517677   4.211 2.71e-05 ***
## MSZoningRL           0.2168165  0.0446123   4.860 1.31e-06 ***
## MSZoningRM           0.1833540  0.0422373   4.341 1.52e-05 ***
## NeighborhoodStoneBr  0.1294964  0.0341173   3.796 0.000154 ***
## Exterior1stBrkFace   0.1204533  0.0279448   4.310 1.75e-05 ***
## MasVnrTypeBrkFace    0.0966859  0.0269086   3.593 0.000338 ***
## MasVnrTypeStone      0.1153232  0.0287234   4.015 6.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09829 on 1348 degrees of freedom
## Multiple R-squared:  0.9406, Adjusted R-squared:  0.9366
## F-statistic: 232.1 on 92 and 1348 DF,  p-value: < 2.2e-16
```

Note that due to page limit we only displayed highly significant results (with p-value < 0.001).

With the log model, coefficients should be interpreted in terms of percent change in HousePrice. We know from above that a 1% increase in GrLivArea is associated with a 0.38% increase in HousePrice. Coefficients for other logarithmic variables can be interpreted similarly. Having a basement, however, is associated with a 0.7% decrease. Judging from the coefficients, we can see that these two variables along with MSZoning, NeighborhoodStoneBr, etc. are the key drivers of sale prices.

To measure the prediction accuracy of our model, we also calculated three major metrics of regression after performing 10 fold cross-validation. We can see our final model has quite robust performance.

```
## Linear Regression
##
## 1441 samples
##   31 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 1298, 1298, 1297, 1297, 1297, 1296, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1025963  0.9304904  0.07685548
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

We've also tested our model on the test set provided by Kaggle and got Root Mean Squared Logarithmic Error of 0.12486.

VI. Discussions

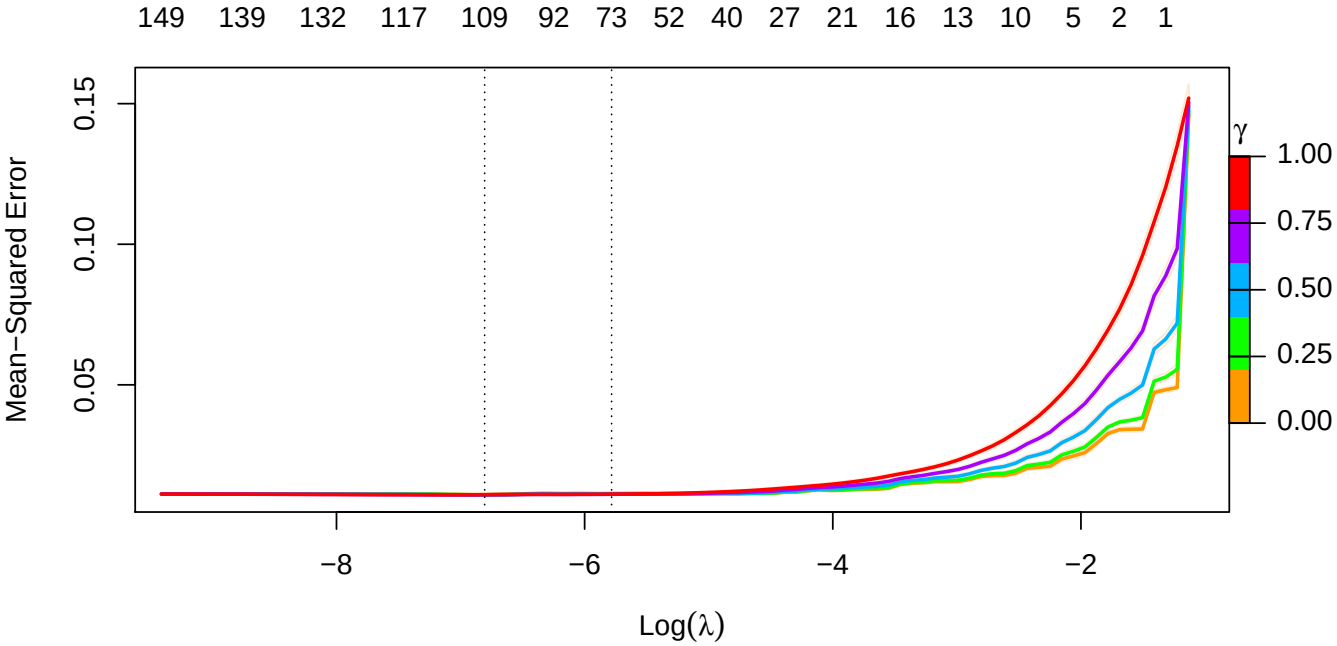
One major issue detected in our model was that the gaussian error assumption did not hold. Although it didn't

falsify our estimates of the coefficients, it indeed invalidated the confidence intervals and p-values. We noticed that it was probably due to the non-gaussianity of the response variable. Perhaps other transformations than logarithmic need to be done to fix this issue, or we should filter observations more aggressively, i.e. to remove more observations until the subset become normally distributed. By doing this we take the risk of overfitting. We've already extended features by hand-crafting or encoding, if we further subsample observations, the full-rank assumption may not hold any more.

We also tried a linear regression model with elastic-net penalty with the help of glmnet package. The elastic-net penalty is a hybrid of L1 and L2 penalties, the objective function with which is given by

$$argmin_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda[(1-\gamma)||\beta||_2^2/2 + \gamma||\beta||_1]$$

. The model with elastic-net penalty is therefore supposed to be more conservative and more compact. Glmnet can automatically fine-tune the hyperparameters such as overall penalty (lambda) and trade-off between L1 and L2 penalties (gamma).



* Figure 6. Plot of cross-validation MSE against penalization lambda

The plot above suggested that a grid search into parameter space (lambda + gamma) was performed and the combination yielding the lowest cross-validation MSE was chosen. The corresponding RMSE is 0.1291, which is significantly higher than that of linear model, but the penalized model was more compact with 51 regressors and should be thus more robust against overfitting.

Given more time, we would like to experiment with kernel-based regression. Feature engineering is a fascinating field and can be greatly beneficial for improving model performance. We can find useful feature interactions by investigating descriptive statistics and then represent them by creating hand-crafted features. But it requires expertise in feature engineering and insights into the real estate industry. The kernel which maps features from its original space into an extended space with higher dimension can be helpful for exploitation of useful feature interactions.