STA302 Final Project

Khizer Asad, Shirley Ching, Dionysius Indraatmadja, Daleep Singh

1 Introduction

The question we seek to address is: which quantitative or qualitative features of a house can be used to predict its sale price in Ames, Iowa? We categorize relevant variables into 4 groups and develop a model for each grouping and use hypothesis testing on the models to test whether each feature-grouping can predict sale price. Our analysis will show whether select regression models can accurately explain and predict the relationship of the sale price with the features of a house.

2 Exploratory Data Analysis

The dataset used in this regression analysis comes from the Ames Housing dataset: `www.kaggle.com/c/house-prices-advanced-regression-techniques/data`. The 4 categories of features are Premium, Interior, Exterior and Qualitative. The response variable for all the models is SalePrice, is the property's sale price in USD. A log-transformation is applied to SalePrice to reduce its left-skew and after this transformation, the response value is more normally distributed. **When we talk about "sale price", "SalePrice", or the "response variable", we will be referring to the log-transformed variable, not the original SalePrice variable**.

2.1 Premium

This grouping includes the nice-to-have or *premium* features of a house: masonry veneer area (MasVnrArea) in square feet, total basement finished area (BsmtFinSF) in square feet, remodel date (YearRemodAdd) in years, total rooms above grade (TotRmsAbvGrd) not including bathrooms. Note that BsmtFinSF is the sum of BsmtFinSF1 and BsmtFinSF2 (if applicable) from the dataset.

In Figure 1 we observe a linear relationship between SalePrice and each of the predictor variables. We do not see any non-random relationships between pairs of predictor variables. In Table 1 we do not observe any high correlation coefficients between any pair of predictor variables. This suggests no severe multi-collinearity between the predictor variables. We observe that MasVnrArea, BsmtFinSF, and YearRemodAdd have high frequency of values and high variance of the response variable at 0.

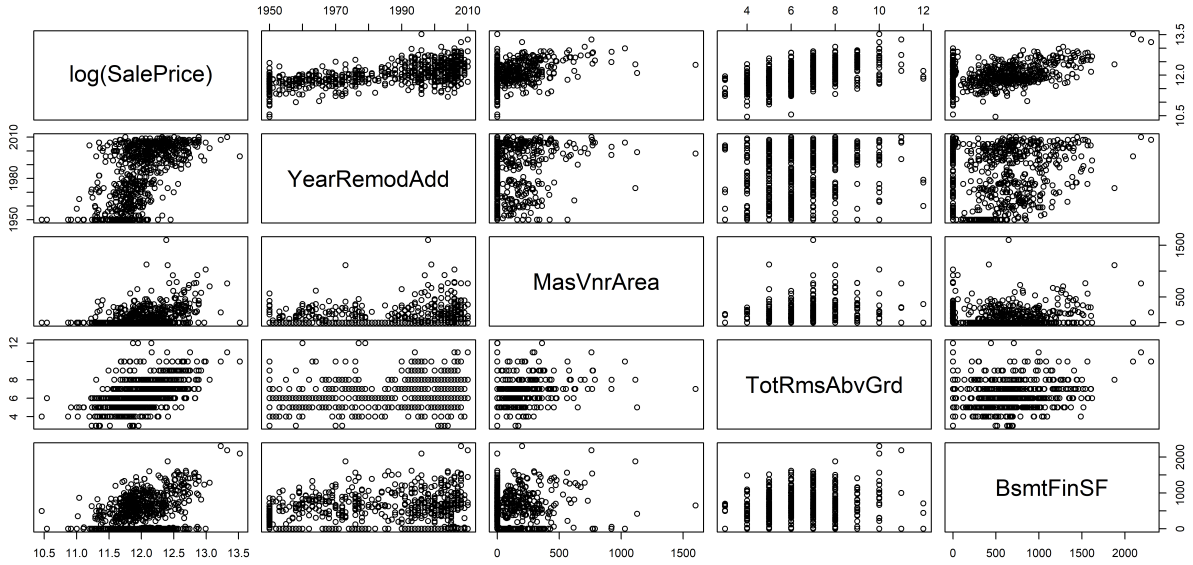

Figure 1: Pairs plot of premium predictor variables and sales price

	ln_SalePrice	YearRemodAdd	MasVnrArea	TotRmsAbvGrd	BsmtFinSF
ln_SalePrice	1.0000000	0.5718152	0.4414690	0.544094407	0.369362937
YearRemodAdd	0.5718152	1.0000000	0.1829164	0.211735448	0.093333103
MasVnrArea	0.4414690	0.1829164	1.0000000	0.272309877	0.206454188
TotRmsAbvGrd	0.5440944	0.2117354	0.2723099	1.000000000	0.009561723
BsmtFinSF	0.3693629	0.0933331	0.2064542	0.009561723	1.000000000

Table 1: Correlation matrix of premium predictor variables and sales price

Since our goal is to predict SalePrice given the values for the predictor variables, we sample the data randomly into two halves: one for **model-building**, and one for **model-validation**.

2.2 Interior

For the interior, we are interested in garage square footage(GarageArea), first floor square footage(X1stFlrSF), second floor square footage(X2ndFlrSF), and total basement square footage (TotalBsmtSF) as the predictor variables. A linear model was initially tested but did not fit the data well. Hence, a quadratic model was attempted which the explorative data is below. Note, the predictor variables include interaction terms. The response variable isthe sales price of a home. Please note that each predictor has been centered about its mean prior to the correlation matrix and pairs plot call, in order to reduce multi-collinearity.

	logSalePrice	frstflr	sndflr	bsmt	grgarea	frstflr_Sqrd	sndflr_Sqrd	bsmt_Sqrd
logSalePrice	1.00000000	0.59698132	0.319300139	0.61213423	0.65088768	0.162090118	0.43285088	0.085568444
frstflr	0.59698132	1.00000000	-0.202646181	0.81952998	0.48978165	0.494929782	0.14398723	0.414021979
sndflr	0.31930014	-0.20264618	1.000000000	-0.17451195	0.13834696	0.005285226	0.67581483	-0.007772648
bsmt	0.61213423	0.81952998	-0.174511950	1.00000000	0.48666546	0.488749839	0.13507920	0.390547872
grgarea	0.65088768	0.48978165	0.138346959	0.48666546	1.00000000	0.204818212	0.28681441	0.182375201
frstflr_Sqrd	0.16209012	0.49492978	0.005285226	0.48874984	0.20481821	1.000000000	0.06517600	0.896365560
sndflr_Sqrd	0.43285088	0.14398723	0.675814831	0.13507920	0.28681441	0.065175998	1.00000000	0.041615736
bsmt_Sqrd	0.08556844	0.41402198	-0.007772648	0.39054787	0.18237520	0.896365560	0.04161574	1.000000000
grgarea_Sqrd	0.01834636	0.15988594	0.023602133	0.18600999	0.10540146	0.315277088	0.11535524	0.327504769
frstflrXsndflr	-0.09886007	0.01353993	0.159529868	0.01561734	-0.03425364	0.206798750	0.42795001	0.265094454
frstflrXbsmt	0.14057759	0.42863573	0.005346333	0.50944203	0.20194809	0.953460797	0.05699645	0.949265319
frstflrXgrgarea	0.11656059	0.37905356	-0.024744909	0.42615737	0.18163875	0.790656378	0.06052203	0.773014783
sndflrXbsmt	-0.11405795	0.01569798	0.150433004	-0.02808201	-0.03573521	0.286984983	0.38808792	0.353033086
sndflrXgrgarea	0.06810274	-0.03367551	0.312411056	-0.03495161	0.03649033	0.072879411	0.52748157	0.098984421
bsmtXgrgarea	0.10953612	0.37310834	-0.022485568	0.41460294	0.18501181	0.790973378	0.04892108	0.855105802

	grgarea_Sqrd	frstflrXsndflr	frstflrXbsmt	frstflrXgrgarea	sndflrXbsmt	sndflrXgrgarea	bsmtXgrgarea
logSalePrice	0.01834636	-0.09886007	0.140577592	0.11656059	-0.11405795	0.06810274	0.10953612
frstflr	0.15988594	0.01353993	0.428635729	0.37905356	0.01569798	-0.03367551	0.37310834
sndflr	0.02360213	0.15952987	0.005346333	-0.02474491	0.15043300	0.31241106	-0.02248557
bsmt	0.18600999	0.01561734	0.509442034	0.42615737	-0.02808201	-0.03495161	0.41460294
grgarea	0.10540146	-0.03425364	0.201948088	0.18163875	-0.03573521	0.03649033	0.18501181
frstflr_Sqrd	0.31527709	0.20679875	0.953460797	0.79065638	0.28698498	0.07287941	0.79097338
sndflr_Sqrd	0.11535524	0.42795001	0.056996451	0.06052203	0.38808792	0.52748157	0.04892108
bsmt_Sqrd	0.32750477	0.26509445	0.949265319	0.77301478	0.35303309	0.09898442	0.85510580
grgarea_Sqrd	1.00000000	0.17510097	0.335170394	0.58707901	0.18152777	0.13172368	0.55875044
frstflrXsndflr	0.17510097	1.00000000	0.247939365	0.28307230	0.85532824	0.52392432	0.27159012
frstflrXbsmt	0.33517039	0.24793936	1.000000000	0.81023980	0.31043670	0.08387865	0.84950383
frstflrXgrgarea	0.58707901	0.28307230	0.810239801	1.00000000	0.30369252	0.03234596	0.92917457
sndflrXbsmt	0.18152777	0.85532824	0.310436704	0.30369252	1.00000000	0.51370287	0.35071221
sndflrXgrgarea	0.13172368	0.52392432	0.083878649	0.03234596	0.51370287	1.00000000	0.04878810
bsmtXgrgarea	0.55875044	0.27159012	0.849503827	0.92917457	0.35071221	0.04878810	1.00000000

Table 2: Correlation matrix of predictor variables.

We can see that there exists some high multi collinearity(> 0.75) between the predictors. For example, bsmt vs frstflr(0.82), along with expected multi-collinearity between the first order terms and the respective quadratic and interaction effects. However, as a whole with the centering around mean applied, most multicollinearity is < 0.75.

The resulting pairs plot is on the next page.

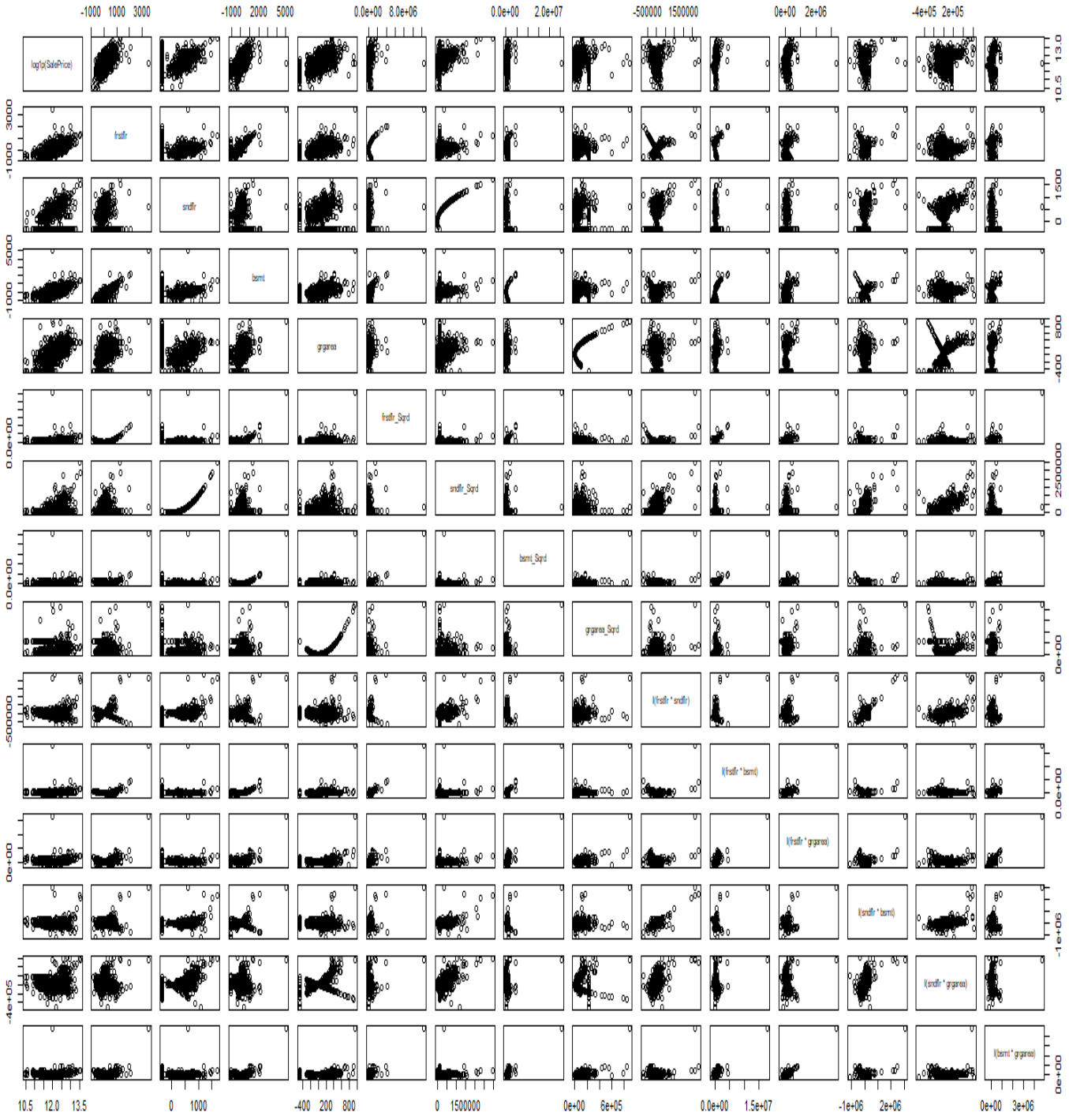

Figure 2: Pairs plot of Interior predictor variables and sales price

2.3 Exterior

The model for exterior category includes predictor variables regarding the area of porches in square ft for houses sold: wood deck (WoodDeckSF), open porch (OpenPorchSF), enclosed porch (EnclosedPorch), three season porch (X3SsnPorch), and screen porch (ScreenPorch).

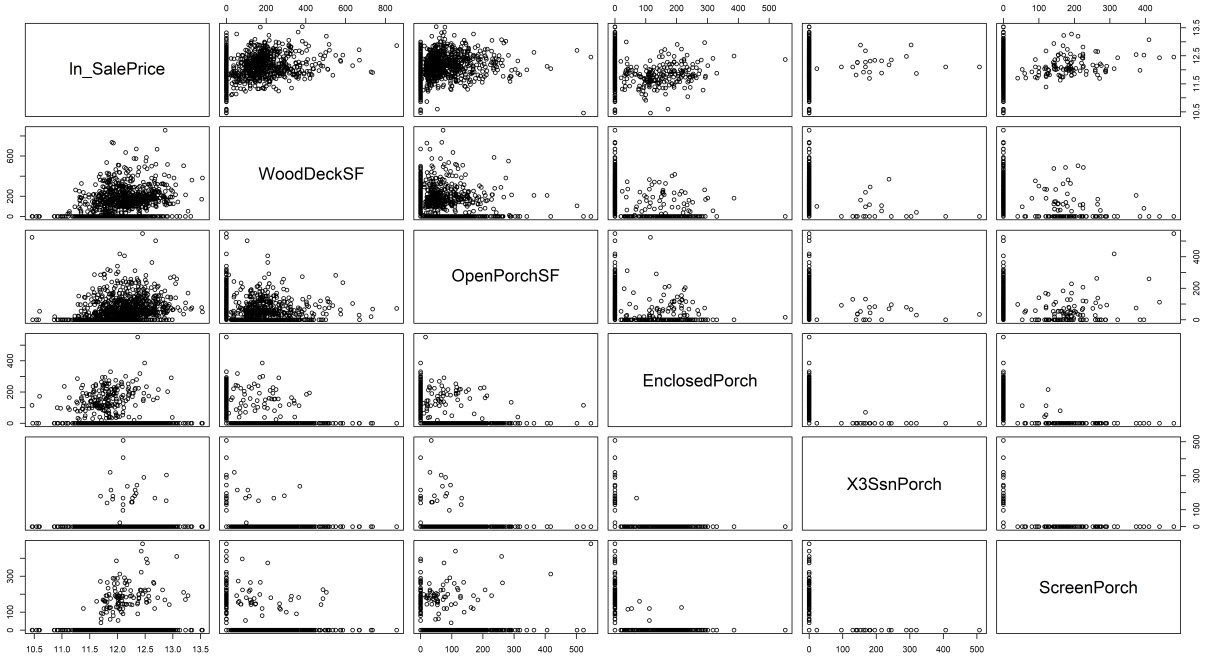

Figure 3: Pairs plot of exterior predictor variables and sales price.

	ln_SalePrice	WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch	ScreenPorch
ln_SalePrice	1.00000000	0.33413507	0.321052972	-0.14905028	0.054900226	0.12120760
WoodDeckSF	0.33413507	1.00000000	0.058660609	-0.12598889	-0.032770634	-0.07418135
OpenPorchSF	0.32105297	0.05866061	1.000000000	-0.09307932	-0.005842499	0.07430394
EnclosedPorch	-0.14905028	-0.12598889	-0.093079318	1.00000000	-0.037305283	-0.08286424
X3SsnPorch	0.05490023	-0.03277063	-0.005842499	-0.03730528	1.000000000	-0.03143585
ScreenPorch	0.12120760	-0.07418135	0.074303944	-0.08286424	-0.031435847	1.00000000

Table 3: Correlation matrix for exterior predictor variables.

From the correlation table (Table 3), we can see that there is no multi-collinearity present. However, the three season porch area (X3SsnPorch) has a near close 0 correlation with the house sale price (SalePrice). This implies we cannot use this predictor variable in our model as the modelling assumption that 'there is a link between the explanatory (X3SsnPorch) and response (SalePrice) variable' would not be valid as required by Gauss-Markov Theorem.

2.4 Qualitative

For modeling the effect of qualitative factors of a house on its price, a combination of allocated codes and indicator variables were used. Variables that had many possible values

used allocated codes, these included the neighborhood (Neighborhood) and exterior material (Exterior) of the house. The allocated codes for neighborhood and exterior material type can be found in the appendix. In addition to these, indicator variables were added to the model for the less diverse qualitative variables in the dataset. These include whether the house has a central air conditioning system (CentralAir) and the type of house (Type). Type is a modified variable from BldgType in the original dataset, which instead of delineating between several different house types (e.g. single detached, townhouse, duplex, etc.) only has two values: detached and attached. The values of the indicator variables are as follows:

- **CentralAir:** $\begin{cases} 0 = No\ Central\ AC \\ 1 = Yes\ Central\ AC \end{cases}$
- **Type:** $\begin{cases} 0 = Attached\ House\ (Townhouse,\ Duplex,\ Double\ Attached) \\ 1 = Single\ Detached\ House \end{cases}$

From (Figure 4) below it is visually apparent that a linear relationship exists between SalePrice and the two explanatory variables, and there is minimal multi-collinearity between Neighborhood and Exterior. The dummy variables are not included in the pairs plot as they only take on values 0 and 1, hence no linear relationship would be visible.

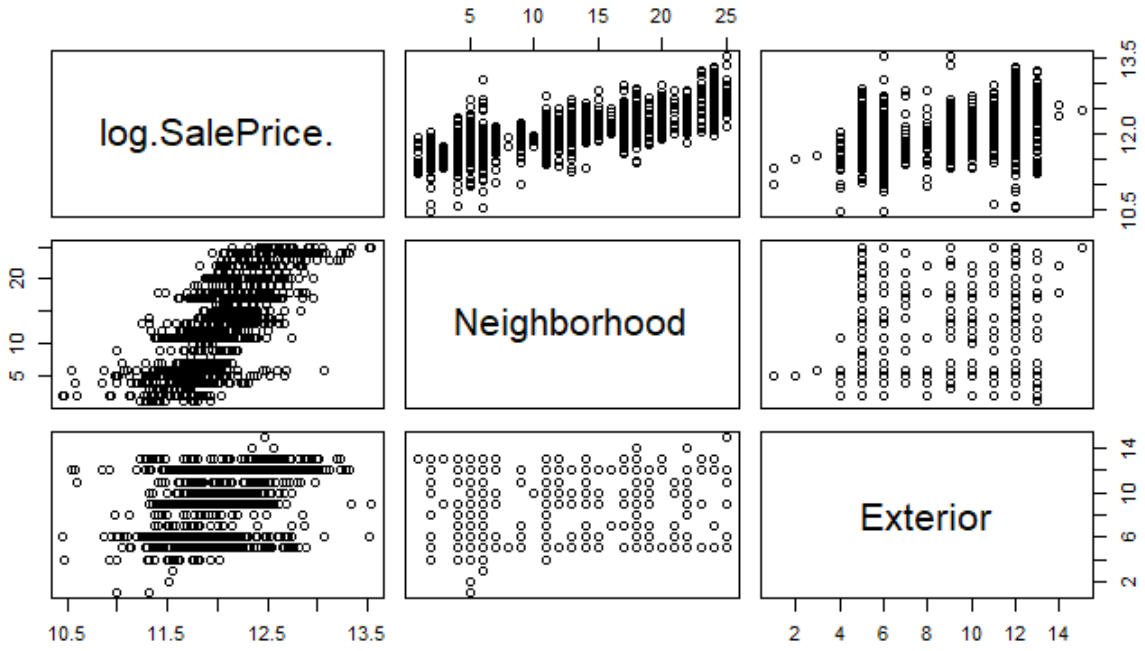

Figure 4: Pairs plot for qualitative predictor variables and sales price

	log.SalePrice.	Neighborhood	Exterior	Type.I	CentralAir.I
log.SalePrice.	1.0000000	0.74115436	0.40564321	0.13246913	0.35160018
Neighborhood	0.7411544	1.00000000	0.39673036	0.03387692	0.27041706
Exterior	0.4056432	0.39673036	1.00000000	-0.01294112	0.23011813
Type.I	0.1324691	0.03387692	-0.01294112	1.00000000	0.08529378
CentralAir.I	0.3516002	0.27041706	0.23011813	0.08529378	1.00000000

Table 4: Correlation matrix for qualitative predictor variables and sales price

From the correlation matrix in (Table 4) it can be seen that there is a strong (0.74) correlation between the Neighborhood and Sale Price, while the correlation between Exterior and Sale Price (0.4) is weaker. The coefficients between CentralAir and Type on SalePrice are low here as they are dummy variables, however a substantial relationship will be seen later.

3 Model Development

We want to see if there is a linear relation between the house's sale price the respective predictor variables within category. We assume that the distribution follows a multivariate linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{1}$$

Alternatively, certain relationships may be better characterized as linear with a log-transformed response variable. In such cases we assume the model is:

$$\log(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

with

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

where $\boldsymbol{Y}$ is the response variable vector for n data points, $\boldsymbol{\beta}$ is the parameter vector, $\boldsymbol{X}$ is the predictor variable matrix for p predictor variables, and $\boldsymbol{\epsilon}$ is the error vector. We assume that the errors $\boldsymbol{\epsilon}$ are pairwise independent, have constant variance, and are normally distributed. We will check that these four Gauss-Markov assumptions are justified for our dataset as follows:

- Plot the response variable against each explanatory variable to qualitatively check for a linear relationship.
- Check that the plot of the residuals of the multivariate linear model against each of the explanatory variables have no trend. This indicates linearity between the true mean of Y and X.
- Plot the residuals against the fitted values. A lack of a pattern indicates independence of errors.

- In the same residuals-fitted plot, check that the variance of the residuals is constant throughout the data. A pattern in the variability of the residuals indicates heteroscedasticity (a sign of non-constant variance of errors).
- Use a normal Q-Q plot to see if the residuals follow a normal distribution.

Hypothesis Test

We construct a two-sided F-test to see if there is a relationship between the chosen predictor variables with the response variable within model. We use the p-value and the standard error to examine the significance of individual predictor variables in the multiple linear regression model. High R^2 and R^2_{adj} values indicate how well the regression surface fits the data.

- **Two-sided test:** $\begin{cases} H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\ H_1 : \text{ at least one of the } \beta_j, k = 1, \ldots, p \text{ is non null} \end{cases}$
- **Test Statistic:** $F^* = \text{MSReg}/\text{MSRes}$.
- **Decision rule:** Reject H_0 if $F^* > F_{1-\alpha;p'-1;n-p'}$, do not reject H_0 if $F^* \leq F_{1-\alpha;p'-1;n-p'}$.

Model Selection, Validation, Diagnostics

3.1 Premium

We apply the hypothesis test with the log-transformed model (Equation 2) on the model-building set and seek to establish a linear relationship between the response variable and the multiple predictor variables. We choose a final model using stepwise backward elimination: after building the linear model with all relevant predictor variables, we eliminate the predictor variable with the lowest t-value if its corresponding p-value exceeds our specified significance level $\alpha = .05$.

3.1.1 Significance of Estimates

In the summary in Table 5 we observe that all predictor variables are significant at the $\alpha = 0.5$ level and we have that $R^2 = 0.6593$ and $R^2_{\text{adj}} = 0.6574$ which indicate that the regression surface fits the data well. We notice that the standard error of each estimate is an order of magnitude smaller than the estimate itself, which indicates that the estimates are relatively precise.

3.1.2 Model Selection

We find the F-statistic to be $F^* = 348.9$ and the critical value $F_{.95;4;721} = 2.38$. Since $F^* > F_{.95;4;721}$, we reject the null hypothesis H_0 and accept the alternative hypothesis H_1. Using `stepAIC()` from the MASS library, we confirm that the final model which minimizes the Akaike Information Criterion ($AIC_p = -2137$) is the one that includes all 4 predictor variables. We include all variables in the model since we have that individual p-values are

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-5.743e+00	8.327e-01	-6.897	1.16e-11
YearRemodAdd	8.541e-03	4.228e-04	20.200	<2e-16
MasVnrArea	3.983e-04	5.105e-05	7.800	2.17e-14
TotRmsAbvGrd	9.912e-02	5.749e-03	17.241	<2e-16
BsmtFinSF	2.586e-04	1.951e-05	13.255	< 2e-16

Table 5: Summary of premium model estimates for the model building dataset.

smaller than the significance level, standard errors are small relative to the estimates, F-test led us to accept the alternative hypothesis, and the AIC_p is minimized.

3.1.3 Model Validation

We use the mean squared prediction error (MSPE) as an indicator for the predictive ability of the model. Note that for our model, we have MSRes = 0.052. We find the predicted values using the coefficients from our model and the predictor variables from the model-validation set. We find that:

$$\text{MSPE} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n^*} = 0.068,$$

where Y_i are the (log-transformed) response variable values from the validation dataset, and $\hat{Y}_i$ are the predicted values. Since MSPE is reasonably close to MSRes, we expect the model to have good predictive ability.

3.1.4 Diagnostics

Improper functional form

We check for an improper functional form by plotting the residuals of the model against each predictor variable, from the model-building set, in Figure 5. We do not see non-random patterns in the residual plots, which indicates that the functional form of the model is appropriate.

Outliers and Influential Points

We find 2 outlying observations (large studentized deleted residuals) in the model with $DFFITS$ values of 0.297 and 0.415. Since the values are less than 1, the observations are not considered influential to the fitted value. We observe later, in the Diagnostics Plots section, that no observations have a high Cook's distance. Therefore, there are no influential outlying observations in the model.

Multi-collinearity

We compute the variance inflation factor (VIF) of the model in Table 6 and observe that each individual VIF is smaller than 10. In addition to the fact that the mean VIF of

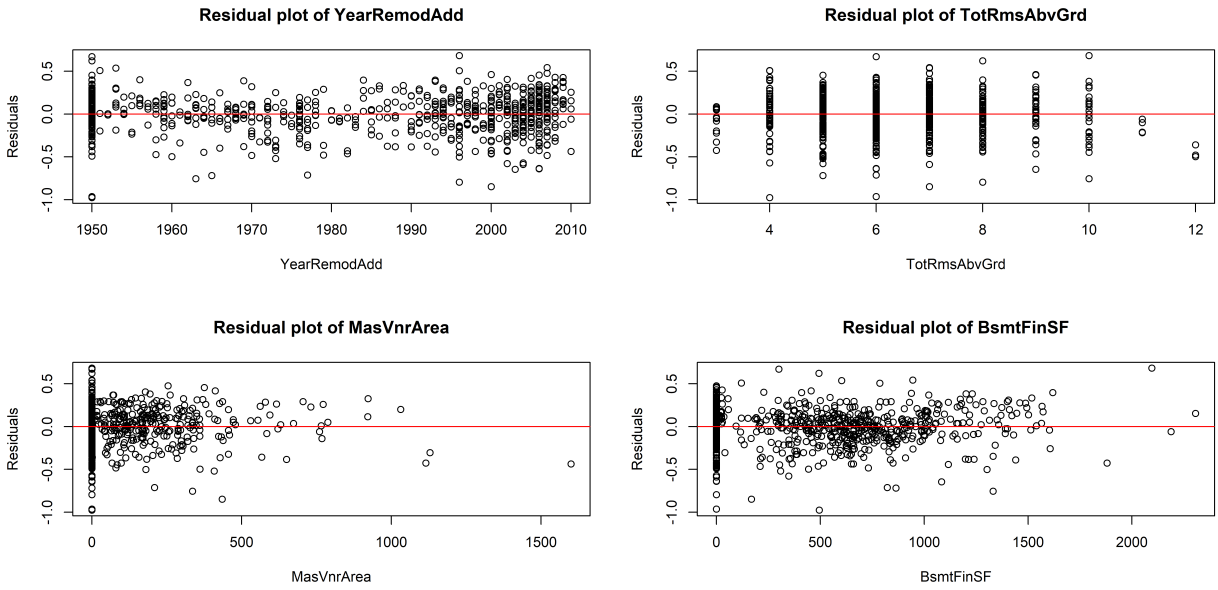

Figure 5: Residuals vs. predictor variables of the premium model.

1.08 is not considerably larger than 1, we conclude that there are no indications of severe multi-collinearity.

	YearRemodAdd	MasVnrArea	TotRmsAbvGrd	BsmtFinSF
VIF	1.067269	1.131248	1.089786	1.051189

Table 6: VIF of the predictor variables.

Diagnostics Plots

We analyze the diagnostics plots displayed in Figure 6 as follows:

- **Checking Linearity**: we observed in the 'Improper functional form' section that the form is appropriate - the linear model is justified.
- **Checking Independence of Errors (Residuals-Fitted)**: we observe equally spread residuals around a horizontal line at $y = 0$ without distinct patterns, so the errors can be assumed to be independent.
- **Checking Homoscedasticity (Scale-Location)**: we observe equally spread points around a horizontal line, so there is no indication of heteroscedasticity.
- **Checking Normality of Errors (Normal Q-Q Plot)**: the residuals follow a straight line except for the left tails, where it diverges significantly. We discuss in detail the effects of this deviation later in the model selection section, but we will treat this assumption as being justified.
- **Checking Influential Outliers (Residuals-Leverage)**: all observations have small Cook's distance, so we do not have high leverage of any outlying observation and no observations are influential.

Therefore, the Gauss-Markov assumptions are justified.

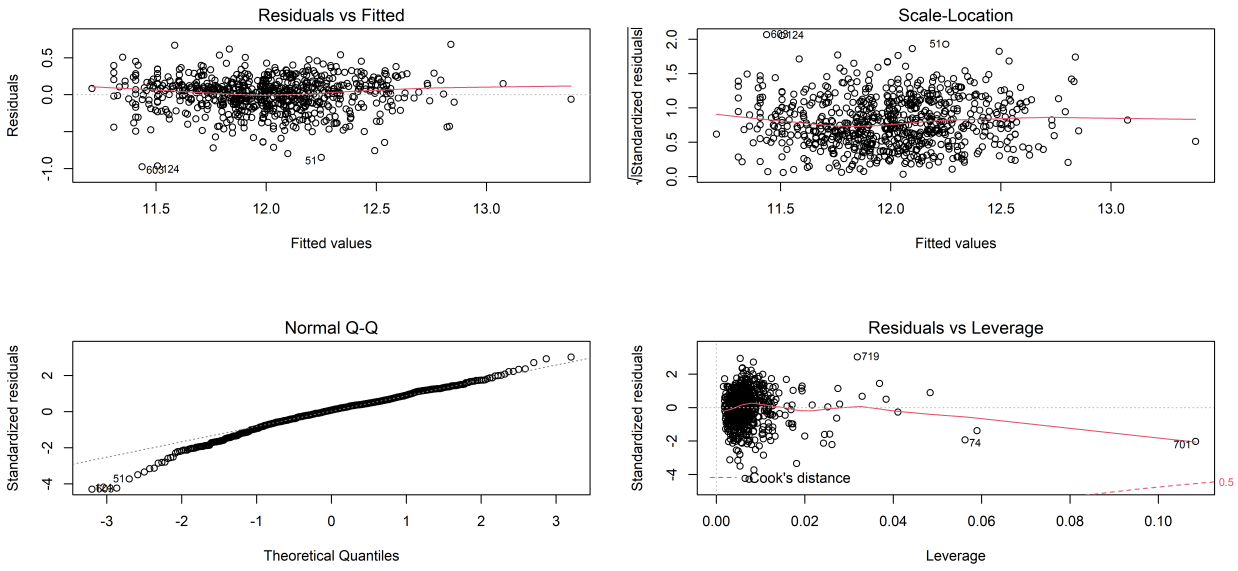

Figure 6: Diagnostics of the premium model.

3.2 Interior

For the Interior model we are interested in a quadratic fit of 4 predictor variables (frstFlrSF, sndFlrSf, BsmtSF, GrgArea) of order 2(interaction effects included) against the log transformation of the sales price of a home. It is also worth noting that 1428 observations were used for the below analysis, as well as the fact that we use backwards elimination to arrive at the final model.

3.2.1 Building the Model

By looking at the summary of the model before backwards elimination (Table 7), we can see that bsmt_sqrd and 3 of the interaction effects are not significant (p-value > 0.05), indicating that those predictor variables are superfluous to the model. Please note that centering about the mean for each predictor has been applied in the analysis below, to reduce multi collinearity.

```
Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)          1.204e+01  1.130e-02 1064.930  < 2e-16 ***
frstflr              3.383e-04  3.534e-05    9.573  < 2e-16 ***
sndflr               3.649e-04  1.964e-05   18.583  < 2e-16 ***
bsmt                 3.529e-04  3.299e-05   10.698  < 2e-16 ***
grgarea              5.090e-04  3.051e-05   16.684  < 2e-16 ***
frstflr_Sqrd        -1.001e-07  5.661e-08   -1.769  0.07716 .
sndflr_Sqrd          8.677e-08  4.287e-08    2.024  0.04312 *
bsmt_Sqrd           -4.316e-08  3.276e-08   -1.318  0.18785
grgarea_Sqrd        -5.980e-07  9.150e-08   -6.535 8.80e-11 ***
I(frstflr * sndflr) -3.434e-07  6.163e-08   -5.571 3.01e-08 ***
I(frstflr * bsmt)   -1.125e-07  7.129e-08   -1.579  0.11459
I(frstflr * grgarea) 1.667e-07  1.402e-07    1.189  0.23446
I(sndflr * bsmt)    -1.643e-09  5.571e-08   -0.029  0.97648
I(sndflr * grgarea)  1.974e-07  6.834e-08    2.889  0.00392 **
I(bsmt * grgarea)    2.997e-07  1.243e-07    2.412  0.01601 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2007 on 1445 degrees of freedom
Multiple R-squared:   0.75,     Adjusted R-squared:  0.7476
F-statistic: 309.7 on 14 and 1445 DF,  p-value: < 2.2e-16
```

Table 7: before backwards elimination

```
Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)          1.204e+01  1.090e-02 1103.672  < 2e-16 ***
frstflr              2.903e-04  2.510e-05   11.564  < 2e-16 ***
sndflr               3.647e-04  1.948e-05   18.719  < 2e-16 ***
bsmt                 3.978e-04  2.290e-05   17.371  < 2e-16 ***
grgarea              5.090e-04  3.034e-05   16.773  < 2e-16 ***
sndflr_Sqrd          8.788e-08  4.264e-08    2.061  0.03948 *
grgarea_Sqrd        -5.574e-07  8.739e-08   -6.378 2.41e-10 ***
I(frstflr * sndflr) -3.309e-07  3.817e-08   -8.669  < 2e-16 ***
I(frstflr * bsmt)   -2.398e-07  2.150e-08  -11.151  < 2e-16 ***
I(sndflr * grgarea)  1.751e-07  6.690e-08    2.617  0.00897 **
I(bsmt * grgarea)    3.808e-07  7.653e-08    4.976 7.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2008 on 1449 degrees of freedom
Multiple R-squared:  0.7491,    Adjusted R-squared:  0.7474
F-statistic: 432.6 on 10 and 1449 DF,  p-value: < 2.2e-16
```

Table 8: Final model after backwards elimination

We can see that in Table 8, after using backwards elimination, most predictors are significant at the 0.001 level, with two being significant at 0.05 and 0.01 levels, respectively.

3.2.2 Model Selection

We can see that in Table 8, we have R^2= 0.7491 and R^2_{adj} = 0.7474, indicating that the model is fitting the data reasonably well. For our F-test, we have a test statistic of $F^*_{10,1449}$ = 432.6 and a critical value of $F_{0.95,10,1449}$ = 2.3842. Since $F^* > F_{0.95,10,1449}$, we reject H_0 and accept H_1. Furthermore, using the AIC criteria, we arrive at a AIC_p value of -4677.368, which was the lowest value out of all possibilities of the model. Note that no predictor variables were removed from the final model compared to the initial AIC input, implying the arrived at model is good.

3.2.3 Diagnostics

Improper Functional Form

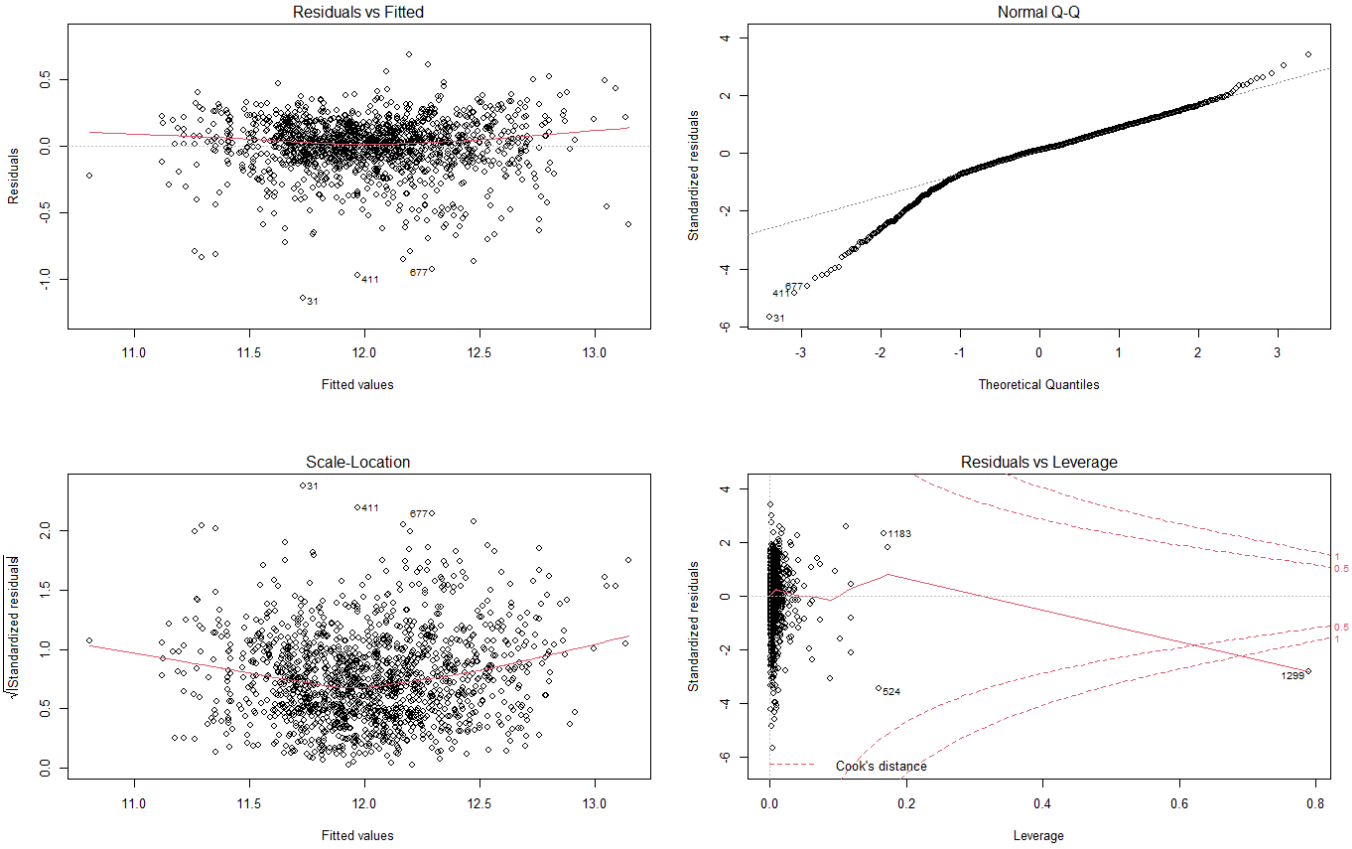

Figure 7: functional forms before outliers and sndflr_sqrd removal

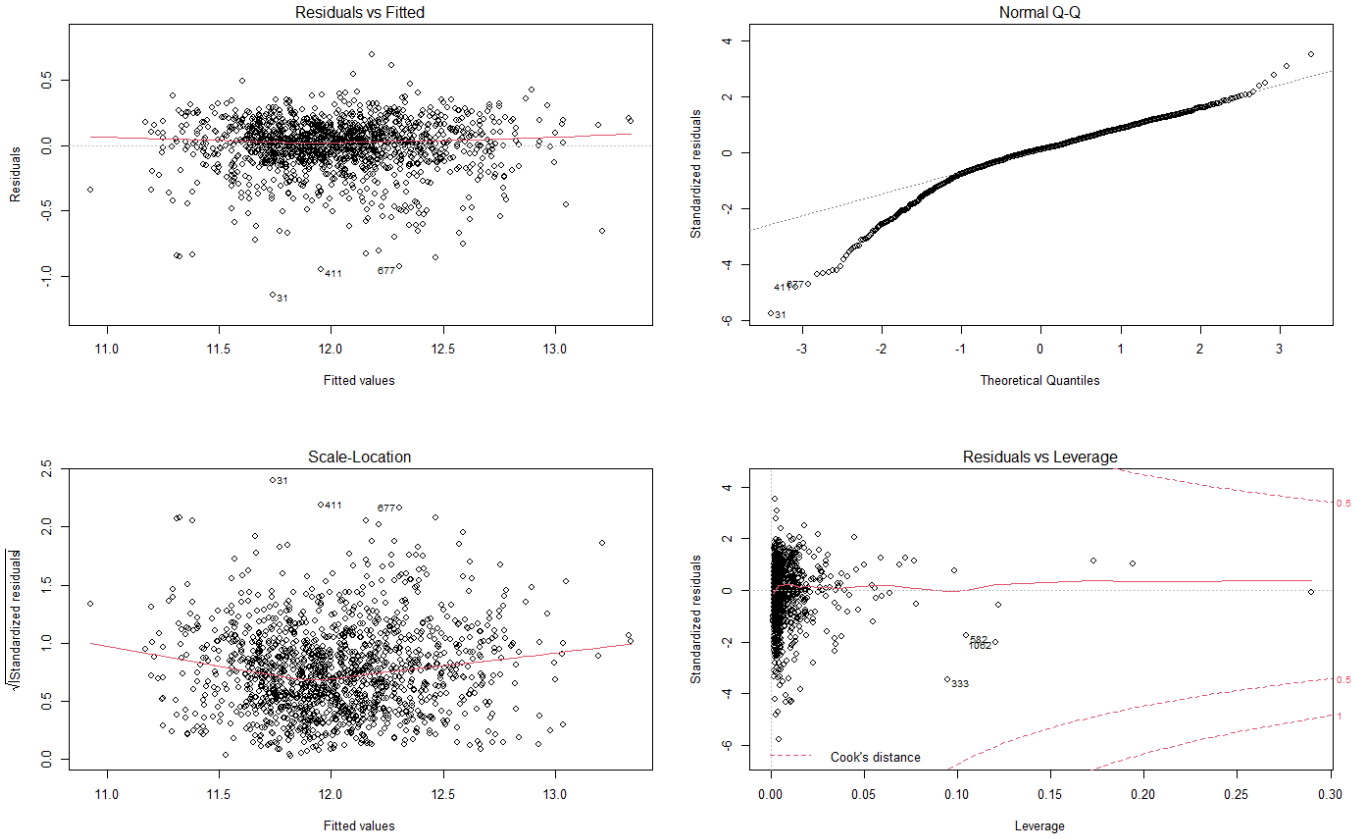

Figure 8: functional forms after outliers and superfluous sndflr_sqrd variable removed

3.2.4 Outliers and Diagnostic Plots

We remove influential outlier (1299) from the data set to find that sndflr_sqrd became insignificant (p-value > 0.05) and to find another influential outlier (524) to come into existence.

After removing the second influential outlier and removing sndflr_sqrd from the regression model, we obtain the same exact summary of the regression model as Table 8, except a stronger F statistic (495) and all predictors significant at the 0.001 level. Furthermore, the AIC was reduced further to -4704.934 from -4677.368. R^2 and R^2_{adj} was increased by 0.1. Overall, the model is improved with the influential outliers and predictor variable removed.

- **Checking curvilinear:** We observe in Figure 8 that the residual vs fitted plot looks random with no discernible patterns with the exception of a slight negative skew. This means that our quadratic model does fit the data.
- **Scale Location**: Looking at the Scale-Location plot in Figure 8, it looks as though non constant variance may exist.
- **Normality of Errors**: The errors are mostly normal, but deviate from normality significantly for residuals < -2.

- **Outliers**: Looking at the Residuals vs Leverage in Figure 8, we can see that no additional influential outliers exist.

3.2.5 Multicollinearity

Using VIC testing, we found the following before the outliers and the sndflr_sqrd predictor variable removed:

```
> VIF <- vif(updated_multi)
> VIF
           frstflr             sndflr               bsmt            grgarea       sndflr_Sqrd       grgarea_Sqrd
          3.409290            2.618574            3.653600            1.523495            3.448936            1.679816
I(frstflr * sndflr)   I(frstflr * bsmt) I(sndflr * grgarea)   I(bsmt * grgarea)
          1.765684            4.829281            1.716885            5.482162
> VIFbar <- mean(vif(updated_multi))
> VIFbar
[1] 3.012772
```

Figure 9: VIC results before outlier and sndflr_sqrd removed

Looking at Figure 9, we can see that no VIC values are greater than 10. We did however, find that the mean of VIC values was equal to 3, indicating that there is some multicollinearity present.

After the removal of the outliers and the predictor variable, we obtain much better numbers:

```
> VIF <- vif(outliers_removed)
> VIF
           frstflr             sndflr               bsmt            grgarea       grgarea_Sqrd I(frstflr * sndflr)
          3.051341            1.273317            3.278912            1.490825            1.562324            1.430575
  I(frstflr * bsmt) I(sndflr * grgarea)   I(bsmt * grgarea)
          1.949334            1.515615            2.210134
> VIFbar <- mean(vif(outliers_removed))
> VIFbar
[1] 1.973597
```

Figure 10: VIC results after outlier and sndflr_sqrd removed

So we can see alot smaller indivudal VIC values than before as well as an overall mean reduction from 3 to 1.97. Thus, a much lower level of multicollinearity is present.

3.2.6 Model Validation

Using an 70/30 split for training and validation, we arrive at very close MSPE and MSE values. indicating that the model has good predictive ability for a split of this ratio.

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{1010} = 0.03980693$$

$$MSPE = \frac{\sum_{i=1}^{n*}(Y_i - \hat{Y}_i)}{438} = 0.03850642$$

3.3 Exterior

The dataset has been split in two evenly. One for model-building set and one for model validation. The number of observations in each is 730.

3.3.1 Significance of Estimates

From the model summary (Table 9), we can see that it is significant. $R^2 = 0.2053$ and $R^2_{\text{adj}} = 0.2009$. The two being close is good, but the regression surface weakly fits the data. Also, unlike the validation set, EnclosedPorch shows only marginal significance in the $\alpha = 0.1$, while the ScreenPorch exhibits significance at $\alpha = 0.05$, and WoodDeckSF, OpenPorchSF are significant at $\alpha = 0.001$. While predictability of a model is important in our investigation, it is not as important as the model estimate themselves being unbiased and satisfying the Gauss-Markov assumptions. Unfortunately, the dataset used in our model is heavily zero-inflated, and to deal with such an ordeal is not covered in the scope of this course. A larger sample size could definitely fix the issue of p-values being too high as we were shown differently in the validation set. The rest of this section will shed light as to how our final Exterior model came to be despite the high p-values from the model-building set along with further analyses.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.8593877  0.0190435 622.752   <2e-16 ***
WoodDeckSF     0.0008885  0.0001010   8.796   <2e-16 ***
OpenPorchSF    0.0016920  0.0001917   8.828   <2e-16 ***
EnclosedPorch -0.0004059  0.0002100  -1.933   0.0536 .
ScreenPorch    0.0005880  0.0002288   2.570   0.0104 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3412 on 725 degrees of freedom
Multiple R-squared:  0.2053,    Adjusted R-squared:  0.2009
F-statistic: 46.82 on 4 and 725 DF,  p-value: < 2.2e-16
```

Table 9: Model-building dataset model summary.

3.3.2 Model Selection

From the model-building dataset regression model summary and F-test results (Table 9), the model shows that it is significant. The F-statistic is $F^* = 46.82$ and the critical value $F_{.95;4;725} = 2.384216$. Since $F^* > F_{.95;4;725}$, we should reject H_0 and accept H_1, that at least one of the $\beta_i \neq 0$ is non-zero.

Also, using backwards elimination on the validation set and AIC_p (Table 10), the final model with the lowest AIC consists of the predictor variables: WoodDeckSF, OpenPorchSF, EnclosedPorch and ScreenPorch.

```
Start:  AIC=-1472.54
valid.ln_SalePrice ~ WoodDeckSF + OpenPorchSF + EnclosedPorch +
    ScreenPorch

               Df Sum of Sq     RSS     AIC
<none>                      95.791 -1472.5
- EnclosedPorch  1    0.7106  96.501 -1469.2
- ScreenPorch    1    2.7022  98.493 -1454.2
- OpenPorchSF    1    9.9734 105.764 -1402.2
- WoodDeckSF     1   13.8993 109.690 -1375.6
> step$anova # display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
valid.ln_SalePrice ~ WoodDeckSF + OpenPorchSF + EnclosedPorch +
    ScreenPorch

Final Model:
valid.ln_SalePrice ~ WoodDeckSF + OpenPorchSF + EnclosedPorch +
    ScreenPorch


  Step Df Deviance Resid. Df Resid. Dev       AIC
1                       725   95.79057 -1472.543
```

Table 10: Model-building dataset model AIC results.

3.3.3 Model Validation

We will be comparing the MSPE (mean squared prediction error) and MSres(mean squared residuals) to test the predictive ability of the model.

$$\text{MSPE} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n^*} = 0.131,$$
$$\text{MSRes} = 0.116$$

We can see that the two values are fairly close to each other, meaning the predictive ability of our model is fairly good.

3.3.4 Diagnostics

Improper functional form

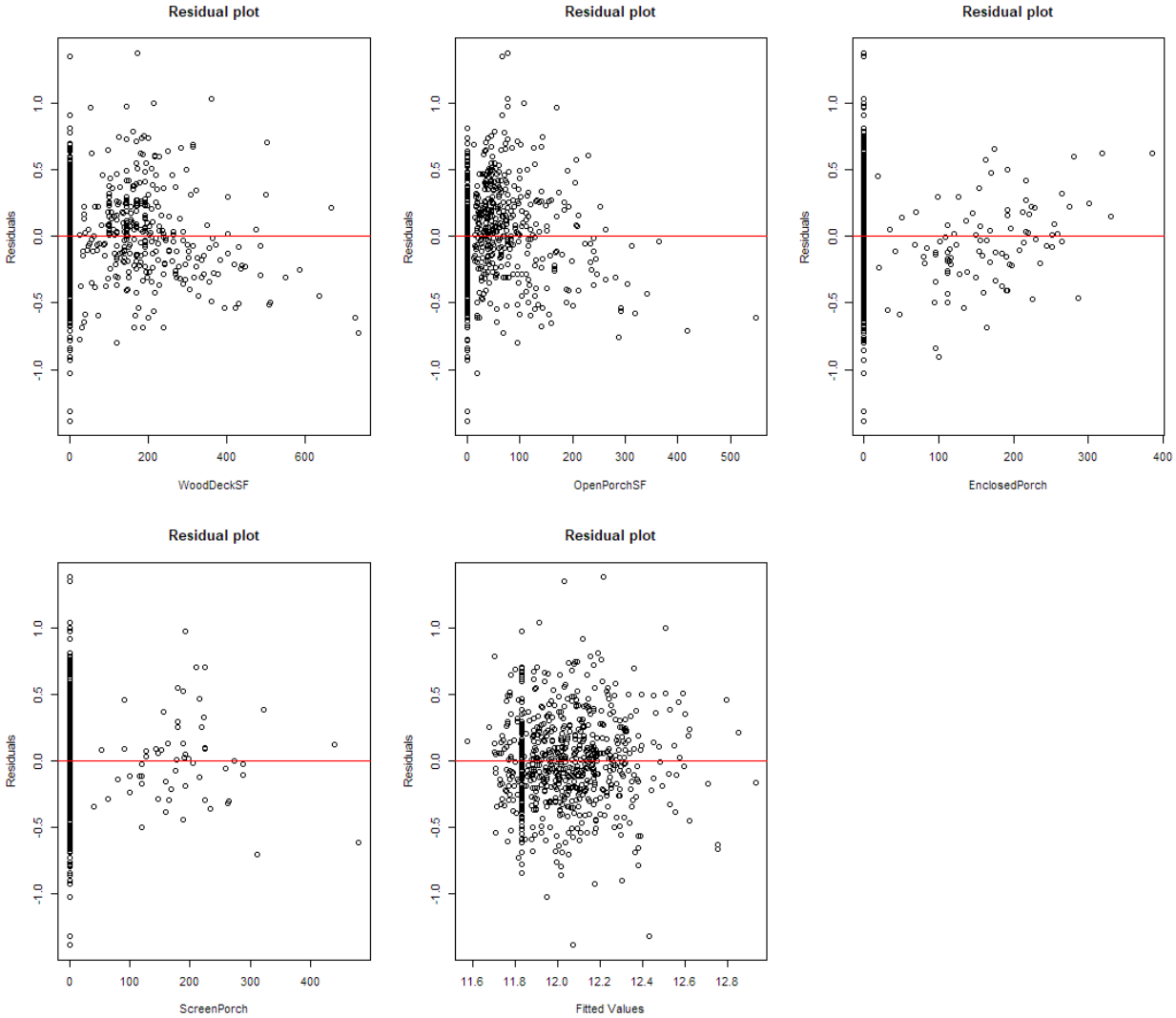

Figure 11: Diagnostic residual plots used to check improper functional form.

Diagnostic residual plots (Figure 11) between the model residuals and predictor variables are used to check for improper functional form. From the residual graphs, we can see that there is no non-randomness, which is a sign of adequate functional form.

Outliers

Testing for outliers in the y-direction, we used studentized deleted residual values. In the y-direction, the outliers were observations: 899, 917 and 1183. Testing for outliers in the

x-direction, we used each observation's Pii (projection matrix) value to compare.

$$P_{\text{ii}} > 2\frac{\sum_{i=1}^{n} P_{\text{ii}}}{n} = 0.01369863.$$

In the x-direction, the outliers were observations: 736, 748, 764, 765, 770, 776, 785, 786, 796, 800, 801, 804, 808, 814, 829, 831, 837, 841, 845, 847, 849, 855, 860, 861, 876, 888, 889, 894, 908, 915, 919, 920, 940, 945, 946, 948, 962, 975, 997, 1014, 1031, 1038, 1045, 1056, 1068, 1069, 1071, 1082, 1095, 1107, 1120, 1140, 1151, 1153, 1155, 1156, 1172, 1185, 1186, 1188, 1194, 1198, 1203, 1211, 1228, 1229, 1249, 1267, 1283, 1293, 1294, 1299, 1302, 1311, 1313, 1314, 1318, 1321, 1327, 1329, 1361, 1370, 1372, 1383, 1387, 1394, 1415, 1420, 1424, 1439, 1440, 1446 and 1460.

Influential points

Cooks distance will be used to decide whether an observation is influential to our model.

From R, $F_{.95;5;725} = 2.226458$. When looped to find whether any of the observations' cooks distance are larger than 50^{th} percentile of $F_{.95;5;725}$, none of them were flagged. This comes to no surprise as the model-buiding dataset has 730 observations, it is a sample big enough that it would be difficult for data points to be highly influential. Since none of the points are influential, they have not been removed from the model despite being outliers.

Multi-collinearity

	WoodDeckSF	OpenPorchSF	EnclosedPorch	ScreenPorch
VIF	1.015562	1.034292	1.027436	1.028064

Table 11: VIF of exterior predictor variables.

The VIF (variance inflation factor) is used to find any multi-collinearity within our model. If the VIF of a predictor variable is larger than 10, there is serious multi-collinearity. Also, the $\overline{VIF}$ value is 1.026338, which is close to 1. A good sign from the two VIF values that there is no severe multi-collinearity.

Diagnostic Plots

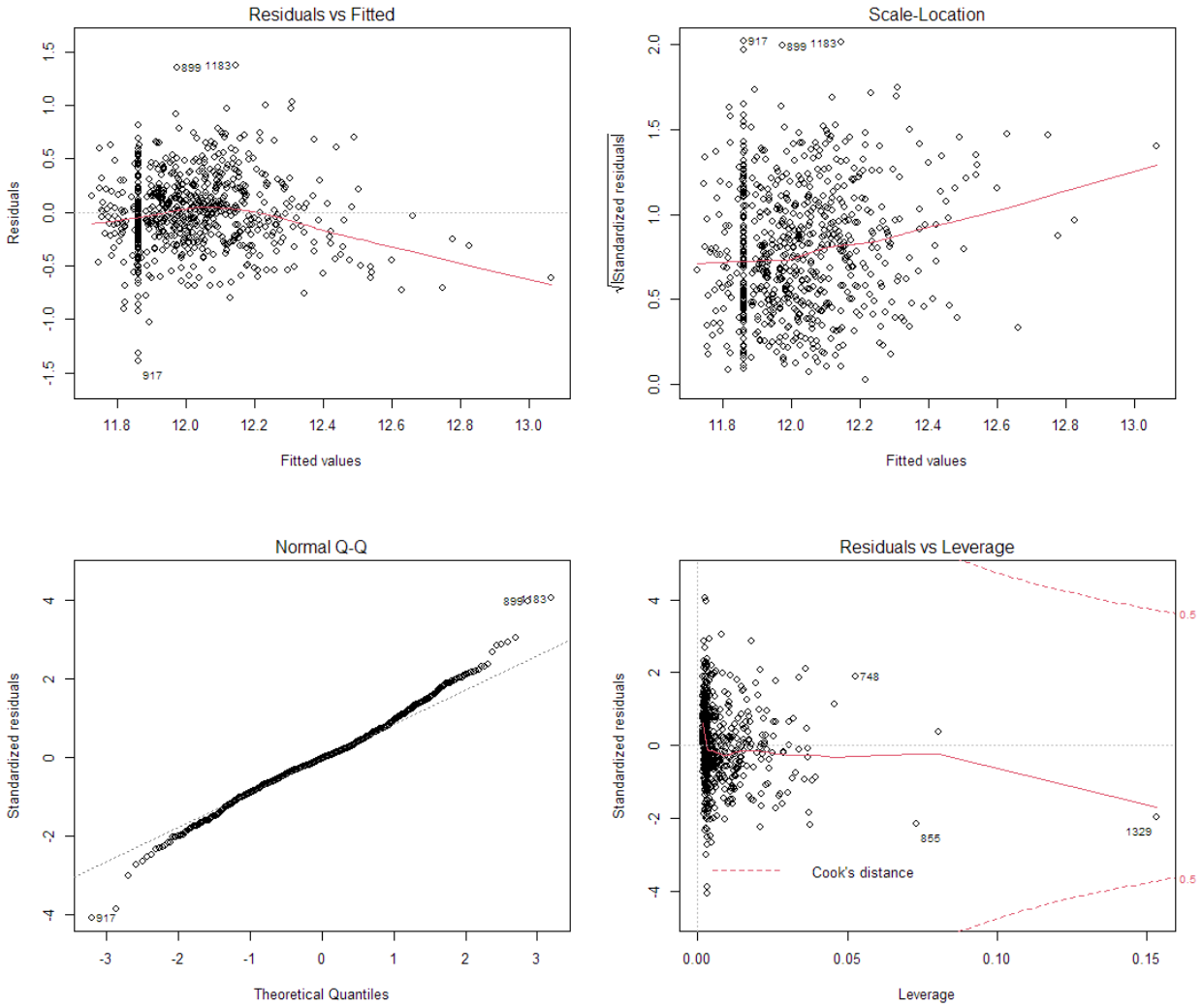

Figure 12: Diagnostic plots of our exterior predictors model used to Gauss-Markov assumptions.

From the diagnostic plots of our model (Figure 12), we will be checking for any violation of the Gauss-Markov assumptions.

- **Checking Linearity:** In the predictor-residual plots (Figure 11), we can see that the points are generally scattered along the horizontal line. This signals there is linearity between SalePrice and each predictor variable. It is safe to say a linear model is appropriate in this case.
- **Checking Independence of Errors:** Back to the residual-fitted plot (Figure 12), the points appear random though it slightly goes below the horizontal line on the right. However, the points still seem non-patterned and not focused on the right, implying there is an independence of errors to the response.
- **Checking Constant Variance of Errors:** Using the Scale-Location, we can see that there is no pattern. And so homoscedasticity is satisfied for the model.

- **Checking Normality of Errors:** Using the QQplot, we check for normality. Non-normal errors is present if they go off line. Most of the points are on top of the line with the exception of some non-influential outliers, which is a sign of normal errors.

Therefore, our model satisfies the Gauss-Markov assumptions.

3.4 Qualitative

For developing the model and hypothesis testing, the data set was randomly split in half for a separate training and testing set each with 730 entries.

3.4.1 Significance of Estimates

```
Call:
lm(formula = log(SalePrice) ~ CentralAir + Type + Neighborhood +
    Exterior)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98480 -0.15232 -0.01674  0.13564  1.06149

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.059224   0.047660 232.043  < 2e-16 ***
CentralAirY   0.221809   0.040268   5.508 5.03e-08 ***
TypeDetached  0.105548   0.025691   4.108 4.44e-05 ***
Neighborhood  0.039589   0.001589  24.917  < 2e-16 ***
Exterior      0.017032   0.003458   4.925 1.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2514 on 725 degrees of freedom
Multiple R-squared:  0.5904,    Adjusted R-squared:  0.5882
F-statistic: 261.3 on 4 and 725 DF,  p-value: < 2.2e-16
```

Table 12: Qualitative variables regression model summary

The summary statistics for the multi-linear regression model in (Table 12) show that all the variables are significant at a p-value < 0.001. In this model CentralAir and Type are binary variables as aforementioned, the Intercept estimate in Table 12 assumes their values are 0 (i.e. a house with no central AC and is attached). And from the summary table CentralAirY and TypeDetached are their values when the variables are each 1, respectively.

Additionally the coefficients of determination $R^2 = 0.5904$ and $R^2_{\text{adj}} = 0.5882$ indicate that the model is a good fit, as the two statistics are close in value and sufficiently large to indicate a correlation. The standard errors for each predictor variable are also sufficiently smaller than their estimates to signify confidence in the regression model.

3.4.2 Model Selection

From the model summary in (Table 12) the F-test statistic is significant with p-value $<$ 2.2e-16. We can corroborate this with the F-test, given that $F^* = 261.3$ and $F_{.95;4;725} = 2.384216$ it can be concluded that $F^* > F_{.95;4;725}$. Hence we reject the null hypothesis H_0 and can claim that that at least one $\beta_j \neq 0$ in this regression.

```
> step <- stepAIC(categ, direction = "both")
Start:  AIC=-2010.89
log(SalePrice) ~ CentralAir + Type + Neighborhood + Exterior

               Df Sum of Sq    RSS     AIC
<none>                      45.819 -2010.9
- Type          1     1.067 46.886 -1996.1
- Exterior      1     1.533 47.352 -1988.9
- CentralAir    1     1.918 47.736 -1983.0
- Neighborhood  1    39.237 85.056 -1561.3
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
log(SalePrice) ~ CentralAir + Type + Neighborhood + Exterior

Final Model:
log(SalePrice) ~ CentralAir + Type + Neighborhood + Exterior


  Step Df Deviance Resid. Df Resid. Dev       AIC
1                        725   45.81887 -2010.894
```

Table 13: Qualitative variables AIC results

Given that the variables in this regression were all fairly discrete, a decision had to be made whether to model them using indicator variables or allocated codes. For the variables with relatively binary, configurations indicators were employed, however this would be quite challenging for the Neighborhood and Exterior variables as they had 25 and 14 configurations respectively. So for these two variables a linear model was used using allocated codes. Initially we attempted to assign the codes by researching the neighborhoods median incomes and the perceived quality of the external materials, however sufficient data was not available to do such. Instead the allocated codes were assigned by ranking the mean sale price value for each configuration in ascending order, as this provided the best linearly correlated model. The legend for the allocated codes for Neighborhood and Exterior material can be found in the appendix.

In the first iteration of the model for qualitative factors of houses effect on sale price, all possible qualitative variables in the data set with correlation to the response variable were included in the regression. These included the current four in addition to: building class (MSSubClass), zoning classification (MSZoning), lot shape (LotShape), lot configuration (LotConfig), roofing (RoofStlye and RoofMatl), heating system type (Heating), electrical system (Electrical), fireplace factors (Fireplaces and FireplaceQual), foundation material (Foundation), swimming pool quality (PoolQC), and the conditions under which the sale was made (SaleCondition). Ultimately the only statistically significant multi-variate model at at least 95 percent was the final model with only the current 4 variables, with non-significant interaction terms.

This is corroborated in (Table 13), which using the stepAIC() function in both directions shows us that this regression model minimized the AIC_p value. Therefore we can conclude that this is the optimal model.

3.4.3 Model Validation

To check the validity of the model, the mean-squared prediction error (MSPE) statistic of our model validation set will be compared to the mean-squared residuals (MSRes) of the model creation set. Using their respective formulae we observe:

$$\text{MSPE} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n^*} = 0.139,$$

$$\text{MSRes} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n - 5} = 0.063,$$

There is an glaring two-fold difference between the MSRes and MSPE values, which suggests that the predictive ability of the model may not be entirely accurate. This discrepancy may stem from the fact that all the predictors in this section are very discrete, and the random sampling to create the separate model building and validation sets may have created a bias in the samples.

3.4.4 Diagnostics

Improper Functional Form

Using (Figure 13) to examine if there is any improper functional form, it can be see that the Neighborhood variable's plot has no patterns and is fairly random. The same can be said for the Exterior however it does have a left skew. For both the indicator variables it is apparent that they are left skewed and have less data points at their 0 values. As previously mentioned the homogeneity of data was one of the main challenges in using the houses qualitative factors as predictors, despite this their estimated $\hat{\beta}$ values were still significant.

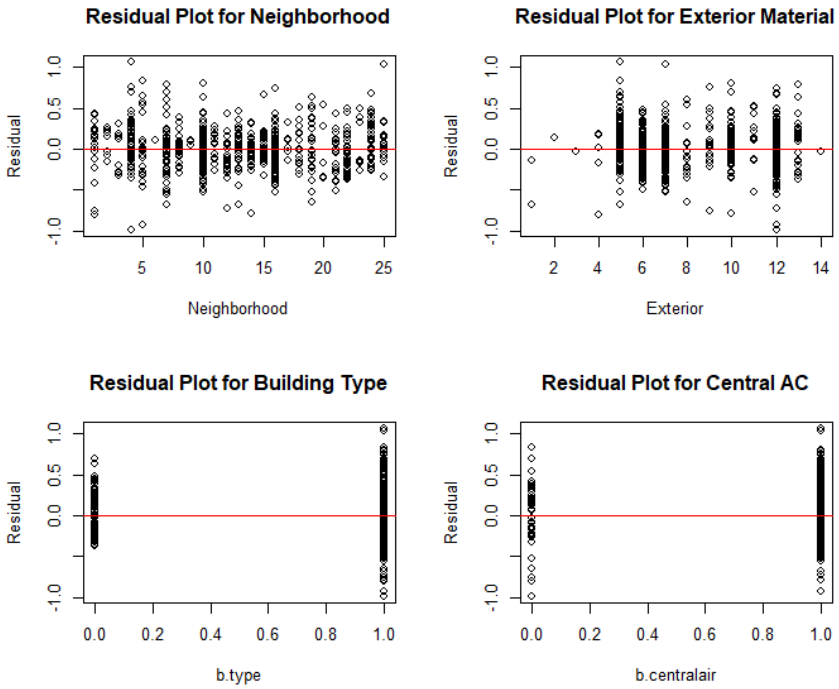

Figure 13: Residual plots for each predictor

Outliers and Influential Observations

Using the studentized deleted residuals to test for outliers on Y, we find that observation 208, 244, and 278 are outliers. Then using the projection matrix we find outliers from X. These include: 21, 46, 58, 71, 72, 80, 101, 115, 140, 141, 150, 152, 153, 159, 162, 177, 187, 208, 216, 217, 220, 225, 236, 243, 249, 263, 269, 279, 281, 294, 295, 299, 305, 310, 314, 325, 326, 332, 342, 345, 347, 348, 360, 375, 379, 407, 422, 432, 438, 473, 504, 510, 518, 521, 511, 570, 571, 575, 605, 609, 610, 620, 625, 628, 641, 643, 660, 669, 671, 678, 704, 710, 715, 724.

However using Cook's distance we find that none of the outliers were influential to the regression, thus they can be kept in the model.

Multi-collinearity

	CentralAir	Type	Neighborhood	Exterior
VIF	1.107103	1.203121	1.019392	1.249986

Table 14: VIF of qualitative predictor variables.

The variance inflation factor (VIF) for each variable seen in (Table 14) shows that the VIF is significantly less than 10 for each predictor. Additionally $\overline{VIF} = 1.144901$ which is relatively close to 1, so we can conclude that there is likely no multi-collinearity effect between the predictors.

Diagnostic Plots

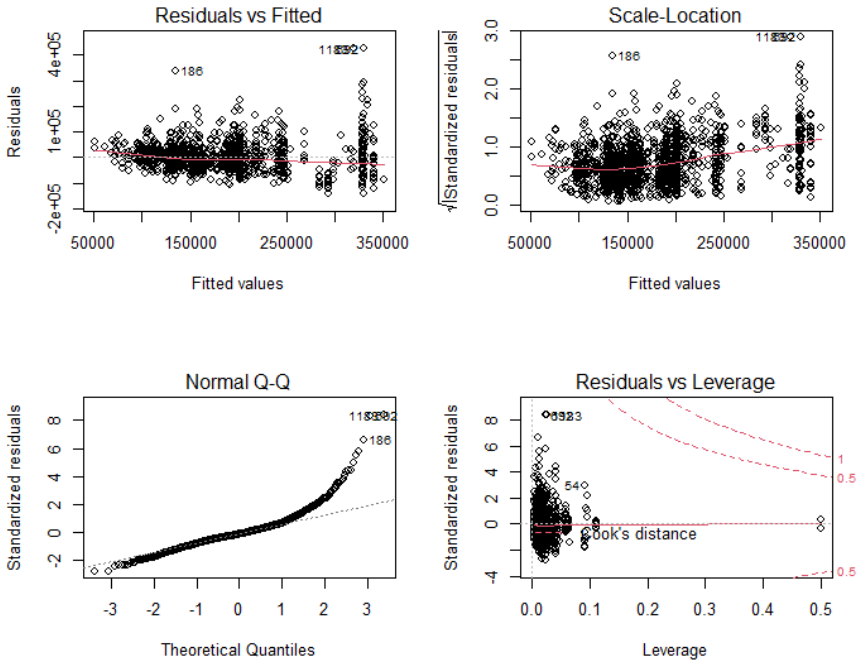

Figure 14: Qualitative regression diagnostic plots

The diagnostic plots in (Figure 14) allow us to check the Gauss-Markov assumptions and validity of the regression:

- **Checking Linearity:** From (Figure 4) and (Figure 13) the linear relationship between SalePrice and the predictors is visible, which satisfies our linearity assumption.
- **Checking Independence of Errors:** The Residual vs. Fitted plot in (Figure 14) shows a random relationship between the residual and fitted values, but with a slight fanning effect. The randomness however let's us satisfy the independence of errors assumption.
- **Checking Constant Variance of Errors:** A similar fanning effect is visible in the Scale-Location (Figure 14), though the regression line is fairly horizontal which satisfies homoscedasticity (i.e. homogeneity of variance).
- **Checking Normality of Errors:** The normality of residuals assumption is violated as seen in the Q-Q plot (Figure 14). The data is heavily right tailed, likely due to the homogeneity of our predictors as was seen in (Figure 13) in the binary variables. This was one of the main drawbacks of using the data set for predictive purposes.
- **Residuals vs Leverage (Figure 14)**: Although there are outliers, the fourth plot shows that there are no observations outside Cook's distance, implying that there are no high leverage points interfering in the regression.

4 Conclusion

4.1 Premium

We observe that a change in the (log-transformed) sale price of properties is directly associated with a change in the masonry veneer area, total basement finished area, remodel date, and the total rooms above grade. We observed that an increase in masonry veneer area tends to drive the sale price up more than an increase in finished basement area. As one would expect, a recent remodelling date of a house is associated with a higher house value. Lastly, the sale price tends to increase with a greater number of rooms above grade in the house.

The only significant technical issue with the model is the non-Gaussian errors, as seen in the normal Q-Q plot of Figure 6. The estimators are still unbiased and minimum variance, but since our dataset is small-medium (<1000 observations), the non-normal errors of our model indicate that the tests of significance and the construction of confidence and prediction intervals may not be valid. However, with the p-values of the predictor variables being much smaller than the significance level $\alpha = 0.05$, the F-statistic being much greater than the critical value, and the lack of issues with other diagnostic tests, the model is still a justified representation of the dataset and a decent predictor of new data. Since the model is a decent predictor of new data, it may be used to estimate the sale price of a house, given the values for the predictor variables.

4.2 Interior

In conclusion we found that our curvlinear model was good after the removal of outliers and non significant predictor variables using backwards elimination. All predictor variables were significant at the 0.001 level, and we had an F statistic of 495.6 when the $F_{1-\alpha=0.95}$ was 2.3842. Thus evidence of atleast 1 significant predictor variable is present. Furthermore, we had a fairly high $R^2 = 0.75$ after the outliers was removed, indicating that the fitted values were close to the actual values. Furthermore, testing of multicollinearity with VIC, all final predictors had VIC values ≤ 3.051341 and a $\overline{VIC} = 1.97$, the mean is a little high but otherwise not bad. With regard to the predictive power, we had very close MSE and MSPE values of 0.03980693 and 0.003850642 values respectively, on a 70/30 training/validation split. Some drawbacks and problems that were encountered were on the data itself. The data deviated from the normal distributed for negative quantiles but remained true to the normal distribution for positive quantiles. Also, the non-linear line in the scale vs location graph indicated a non constant variance, although the line bends at most 0.25, more advanced testing methods are required to get an exact classification of constant versus non-constant variance. Certain problems with non normal errors could perhaps be fixed by a larger data set. Overall, we can see that the model, with some short comings, is decent. This is a good starting place for more research to be done on the relationship of interior SF versus Sale Price.

4.3 Exterior

Converting back from the log-transformation, for every increase in 1.000889sq ft of wood deck, 1.001693sq ft of open porch, 0.9995942sq ft of enclosed porch, and 1.000588sq ft of screen porch, there is an increase of a dollar in house sale price.

From the hypothesis testing for the Exterior predictors model, we can say that the exterior predictors do play a part in the prediction of the house sale price. However, we have to note that the predictor values were heavily zero-inflated. Another issue is the difference in significance for the predictor variables between the model-building model and validation model, that can most likely be solved with a larger sample size. Also, it is unlikely for a house to have the four different types of porches at once. Despite discrepancies between the two models, the predictive ability of the exterior aspect is undeniably present from the diagnostic analyses even if weak.

4.4 Qualitative

To provide a concrete result to the analysis by reversing the log-transformation on sale price, we find that having central air conditioning in a house correlates with an average $16,409 price increase. In addition going from an attached to detached house on average correlates with a $22,156 sale value increase. The neighborhood in which a house is located can also correlate with up to a $185,732 price difference, and the exterior material type a $47,205 difference. Clearly the largest effect within this section is caused by the neighborhood, though all three other predictors also represent significant price changes. Hence we can claim that the qualitative factors of a house can have a substantial affect on its price.

However in the model development section it was seen that the posited regression model unfortunately does not yield an entirely accurate prediction model for determining house sale price. For this part of the analysis the data set was lacking in sample size and heterogeneity to produce meaningful results.

4.5 Final Conclusion

Overall, we have found that the models focusing solely on different aspects of house: Premium, Interior, and Exterior do indeed have predictive ability in the market sale price of a house in Ames, Iowa, answering our question. However the predictive ability of the qualitative variables model is not as potent. The application of this model in other locations or contexts may be limited, as the data lacks observations from other environments.

5 Reference

"House Prices: Advanced Regression Techniques — Kaggle". 2020. Kaggle.Com. https://www.kaggle.c
prices-advanced-regression-techniques/.

Appendix: R code for Premium Model

Dionysius Indraatmadja

Use `setwd` to set the working directory to the extracted folder `house-prices-advanced-regression-techniques`, and set `plotwd` to a folder where the images will be saved.

```
library(car)
library(MASS)
set.seed(1003024416)

setwd("X:/Dropbox/sta302/final project/house-prices-advanced-regression-techniques")
plotwd<-"X:/Dropbox/sta302/final project/writeup/plots/"
table<-read.csv("train.csv", header=TRUE)
data<-cbind("ln_SalePrice"=log(table[, "SalePrice"]),
            table[, c("YearRemodAdd", "MasVnrArea", "TotRmsAbvGrd")])
data[, "BsmtFinSF"]=table$BsmtFinSF1+table$BsmtFinSF2
data<-data[complete.cases(data), ]  # exclude incomplete data points

# Split data into model building set and validation set
sample_size<-floor(.5*nrow(data))
picked<-sample(seq_len(nrow(data)), size=sample_size)
train<-data[picked,]
attach(train)
test<-data[-picked,]
cor(train)
# perform the linear fit
multi.fit<-lm(ln_SalePrice~YearRemodAdd + MasVnrArea + TotRmsAbvGrd + BsmtFinSF)
predictors<-c("YearRemodAdd", "MasVnrArea", "TotRmsAbvGrd", "BsmtFinSF")
multi.res<-resid(multi.fit)
fitted<-fitted(multi.fit)
summary(multi.fit)

# Model selection using AIC
step<-stepAIC(multi.fit, direction="backward")
step$anova

# F critical value
crit<-qf(.95, length(multi.fit$coefficients)-1, multi.fit$df)
fstatistic<-summary(multi.fit)$fstatistic[[1]]  #dendf is denominator dof
print(crit)
print(fstatistic)

# Anova table
anova_table<-anova(multi.fit)
MSRes<-anova_table[nrow(anova_table), 3]
s2<-MSRes  # estimator of variance
```

```r
# Model Validation
beta<-as.numeric(data.matrix(coef(multi.fit)))  # beta coefficients
X<-data.matrix(cbind("1"=vector("numeric", length=dim(test)[1])+1,
                     test[predictors]))  # test data matrix
predictions<-as.numeric(X %*% beta)
test_response<-as.numeric(data.matrix(test[1]))
residuals<-log(test_response) - predictions
MSPE<-sum((residuals)**2)/(dim(test)[1])
MSPE
MSRes

# Outliers
semistudentized<-residuals/var(residuals)
Pmatrix<-X %*% solve(t(X) %*% X) %*% t(X)  # projection matrix
del_res<-residuals/(1-diag(Pmatrix))
n<-dim(train)[1]
p_prime<-length(coef(multi.fit))
mean_leverage<-2*p_prime/n

t<-as.numeric(rstudent(multi.fit))
t_crit<-qt(1-(.05/(2*n)), n-p_prime-1)
outliers_i<-which(abs(t) > t_crit)
pii<-c(diag(Pmatrix)[[outliers_i[1]]], diag(Pmatrix)[[outliers_i[2]]])
t[outliers_i]  # outliers
mean_leverage
which(as.numeric(diag(Pmatrix))>mean_leverage)

# Measure of influence of outlying observations
DFFITS<-abs(t[outliers_i]*sqrt(pii/(1-pii)))
DFFITS

# VIF
vif(multi.fit)
mean(vif(multi.fit))

# Diagnostics plots
w=3*4
h=3*2
png(paste(plotwd, "diagnostics.png", sep=""), width=w, height=h, units="in", res=600)
layout(matrix(c(1,2,3,4),2,2)) # yields 4 graphs/page
plot(multi.fit)
dev.off()

# Pairs plot
png("X:/Dropbox/sta302/final project/writeup/plots/pairs-plot.png",
    width = w, height = h, units = 'in', res = 600)
pairs(ln_SalePrice ~ YearRemodAdd + MasVnrArea + TotRmsAbvGrd + BsmtFinSF, data = train)
dev.off()
cor(train)

# Residuals Plots
png(paste(plotwd, "a1.png", sep=""), width=w, height=h, units="in", res=600)
layout(matrix(c(1,2,3,4),2,2)) # yields 4 graphs/page
```

```
plot(x=YearRemodAdd, y=multi.res,
     ylab="Residuals", xlab="YearRemodAdd",
     main="Residual plot of YearRemodAdd")
abline(0, 0, col='red')
plot(x=MasVnrArea, y=multi.res,
     ylab="Residuals", xlab="MasVnrArea",
     main="Residual plot of MasVnrArea")
abline(0, 0, col='red')
plot(x=TotRmsAbvGrd, y=multi.res,
     ylab="Residuals", xlab="TotRmsAbvGrd",
     main="Residual plot of TotRmsAbvGrd")
abline(0, 0, col='red')
plot(x=BsmtFinSF, y=multi.res,
     ylab="Residuals", xlab="BsmtFinSF",
     main="Residual plot of BsmtFinSF")
abline(0, 0, col='red')
dev.off()
```

Appendix: R Code for Interior Model

Daleep Singh

```
setwd("C:/Users/Daleep/Desktop/sta302/Final Project")
prices=read.csv('train.csv', header=TRUE)
prices3<-prices[-c(1299,524),]
attach(prices3)


#-------------------------------------------------------------------------
---------------
#                Internal Squarefootage vs Salesprice regression
#-------------------------------------------------------------------------
---------------
#--------------------------------
# Polynomial Regression
#--------------------------------
par(mfrow = c(2,2))


frstflr <- X1stFlrSF-mean(X1stFlrSF)
sndflr <- X2ndFlrSF-mean(X2ndFlrSF)
bsmt <- TotalBsmtSF - mean(TotalBsmtSF)
grgarea <- GarageArea - mean(GarageArea)

frstflr_Sqrd <- I(frstflr^2)
sndflr_Sqrd <- I(sndflr^2)
bsmt_Sqrd <- I(bsmt^2)
grgarea_Sqrd <- I(grgarea^2)

frstflrXsndflr <-I(frstflr*sndflr)
frstflrXbsmt<- I(frstflr*bsmt)
frstflrXgrgarea<- I(frstflr*grgarea)
sndflrXbsmt<-I(sndflr*bsmt)
sndflrXgrgarea<-I(sndflr*grgarea)
bsmtXgrgarea<-I(bsmt*grgarea)
logSalePrice <- log1p(SalePrice)

multi <- lm(log1p(SalePrice)~frstflr+ sndflr +bsmt+ grgarea +
frstflr_Sqrd + sndflr_Sqrd + bsmt_Sqrd
            +grgarea_Sqrd + I(frstflr*sndflr) + I(frstflr*bsmt)
+I(frstflr*grgarea)+
              I(sndflr*bsmt) + I(sndflr*grgarea)+
              I(bsmt*grgarea),data = prices)
summary(multi)

updated_multi <- lm(log1p(SalePrice)~frstflr+ sndflr +bsmt+ grgarea +
sndflr_Sqrd + grgarea_Sqrd + I(frstflr*sndflr) + I(frstflr*bsmt)+
                      I(sndflr*grgarea)+ I(bsmt*grgarea),data =
prices)
summary(updated_multi)

hist(resid(updated_multi))
```

```r
#--------------------------------
# MultiCollinearity
#--------------------------------

pairs(~log1p(SalePrice)+frstflr+ sndflr +bsmt+ grgarea + frstflr_Sqrd +
sndflr_Sqrd + bsmt_Sqrd
      +grgarea_Sqrd + I(frstflr*sndflr) + I(frstflr*bsmt)
+I(frstflr*grgarea)+
        I(sndflr*bsmt) + I(sndflr*grgarea)+
        I(bsmt*grgarea), data = prices)
collinearity <- cbind(logSalePrice,frstflr, sndflr ,bsmt, grgarea,
frstflr_Sqrd , sndflr_Sqrd , bsmt_Sqrd
                      ,grgarea_Sqrd ,frstflrXsndflr  ,
frstflrXbsmt,frstflrXgrgarea,
                      sndflrXbsmt , sndflrXgrgarea,
                      bsmtXgrgarea)
cor(collinearity)

VIF <- vif(updated_multi)
VIF
VIFbar <- mean(vif(updated_multi))
VIFbar
#--------------------------------
# Closeness of fit
#--------------------------------
plot(x = fitted(updated_multi),y = resid(updated_multi),xlab = "Fitted
Values",ylab = "Residuals",
     main = "Residual vs Fitted")
abline(0,0)

#--------------------------------
# Normality of errors
#--------------------------------
qqnorm(rstandard(updated_multi),ylab = "Standardized Residuals")
qqline(rstandard(updated_multi))

#--------------------------------
# AIC
#--------------------------------

library(MASS)
step = stepAIC(outliers_removed, direction = "backward")
step$anova # display results

plot(updated_multi)

outlier1299_Removed <-lm(log1p(SalePrice)~frstflr+ sndflr +bsmt+
grgarea + sndflr_Sqrd + grgarea_Sqrd + I(frstflr*sndflr) +
I(frstflr*bsmt)+
                         I(sndflr*grgarea)+ I(bsmt*grgarea),data =
prices2)
summary(outlier1299_Removed)
plot(outlier1299_Removed)
#after deleting two outliers we see that sndflr_sqrd is not significant
```

```
any more. So
#we remove it...

outliers_removed <-lm(log1p(SalePrice)~frstflr+ sndflr +bsmt+ grgarea +
grgarea_Sqrd + I(frstflr*sndflr) + I(frstflr*bsmt)+
           I(sndflr*grgarea)+ I(bsmt*grgarea),data = prices3)

summary(outliers_removed)

plot(outliers_removed)

VIF <- vif(outliers_removed)
VIF
VIFbar <- mean(vif(outliers_removed))
VIFbar

#Validation set analysis
#testing 50x50 split
train <- prices3[1:1020,]
validation <-prices[1021:1458,]
attach(train)

frstflr <- X1stFlrSF-mean(X1stFlrSF)
sndflr <- X2ndFlrSF-mean(X2ndFlrSF)
bsmt <- TotalBsmtSF - mean(TotalBsmtSF)
grgarea <- GarageArea - mean(GarageArea)

frstflr_Sqrd <- I(frstflr^2)
sndflr_Sqrd <- I(sndflr^2)
bsmt_Sqrd <- I(bsmt^2)
grgarea_Sqrd <- I(grgarea^2)

frstflrXsndflr <-I(frstflr*sndflr)
frstflrXbsmt<- I(frstflr*bsmt)
frstflrXgrgarea<- I(frstflr*grgarea)
sndflrXbsmt<-I(sndflr*bsmt)
sndflrXgrgarea<-I(sndflr*grgarea)
bsmtXgrgarea<-I(bsmt*grgarea)
logSalePrice <- log1p(SalePrice)

#MSE
model_MSE <- lm(log1p(SalePrice)~frstflr+ sndflr +bsmt+ grgarea +
grgarea_Sqrd + I(frstflr*sndflr) + I(frstflr*bsmt)+
              I(sndflr*grgarea)+ I(bsmt*grgarea),data = train)
summary(model_MSE)
MSE <- sum((resid(model_MSE))^2)/(nrow(train)-10)

#MSPE
attach(validation)
frstflr <- X1stFlrSF-mean(X1stFlrSF)
sndflr <- X2ndFlrSF-mean(X2ndFlrSF)
bsmt <- TotalBsmtSF - mean(TotalBsmtSF)
grgarea <- GarageArea - mean(GarageArea)
```

```
frstflr_Sqrd <- I(frstflr^2)
sndflr_Sqrd <- I(sndflr^2)
bsmt_Sqrd <- I(bsmt^2)
grgarea_Sqrd <- I(grgarea^2)

frstflrXsndflr <-I(frstflr*sndflr)
frstflrXbsmt<- I(frstflr*bsmt)
frstflrXgrgarea<- I(frstflr*grgarea)
sndflrXbsmt<-I(sndflr*bsmt)
sndflrXgrgarea<-I(sndflr*grgarea)
bsmtXgrgarea<-I(bsmt*grgarea)
logSalePrice <- log1p(SalePrice)

model_MSPE <- lm(log1p(SalePrice)~frstflr+ sndflr +bsmt+ grgarea +
grgarea_Sqrd + I(frstflr*sndflr) + I(frstflr*bsmt)+
                    I(sndflr*grgarea)+ I(bsmt*grgarea),data =
validation)

summary(model_MSPE)
MSPE <- sum((resid(model_MSPE))^2)/(nrow(validation))
```

Appendix: R Code for Exterior Model

Shirley Ching

```
setwd("C:/Users/Miellu/Desktop/STA302FinalProj")
train<-read.csv("train.csv", header=TRUE)
library(MASS)
library(car)
```

```
## Loading required package: carData
```

```
#Applying log transformation to SalePrice to normalize the left-skewed distribution
ln_SalePrice = log(train$SalePrice)
hist(ln_SalePrice)

new_data <- cbind(ln_SalePrice, train$WoodDeckSF, train$OpenPorchSF, train$EnclosedPorch,
                  train$X3SsnPorch, train$ScreenPorch)
colnames(new_data)<-c("ln_SalePrice", "WoodDeckSF", "OpenPorchSF", "EnclosedPorch"
                      ,"X3SsnPorch", "ScreenPorch")


#Explanatory Data Analysis: Correlation matrix and pairs plot
cor(new_data)
pairs(~ ln_SalePrice + WoodDeckSF +  OpenPorchSF + EnclosedPorch + X3SsnPorch +
        ScreenPorch, data = train)


#Model of all data points without XSsnPorch (3 Seasons Porch)... Temporarily
#checking for assumptions
multi.fit = lm(ln_SalePrice~WoodDeckSF +  OpenPorchSF + EnclosedPorch + ScreenPorch,
               data = train)
summary(multi.fit)
par(mfrow=c(2,2))
plot(multi.fit)

#Residual vs. Predictor Variable plots
multi.res = resid(multi.fit)
fitted=fitted(multi.fit)
plot(train$WoodDeckSF, multi.res, ylab="Residuals", xlab="WoodDeckSF",
     main="Residual Plot of WoodDeckSF")
abline(h=0, col="Red")
plot(train$OpenPorchSF, multi.res, ylab="Residuals", xlab="OpenPorchSF",
     main="Residual Plot of OpenPorchSF")
abline(h=0, col="Red")
plot(train$EnclosedPorch, multi.res, ylab="Residuals", xlab="EnclosedPorch",
     main="Residual Plot of EnclosedPorch")
abline(h=0, col="Red")
```

```r
plot(train$ScreenPorch, multi.res, ylab="Residuals", xlab="ScreenPorch",
     main="Residual Plot of ScreenPorch")
abline(h=0, col="Red")



#Validation data set
valid.dat=train[1:730, ]
par(mfrow=c(2,3))
hist(valid.dat$SalePrice, main="Histogram of SalePrice from Validation Model")
multi.fit = lm(SalePrice~WoodDeckSF +  OpenPorchSF + EnclosedPorch + ScreenPorch,
               data=valid.dat)
summary(multi.fit)

multi.res = resid(multi.fit)
fitted=fitted(multi.fit)
plot(fitted, multi.res,
     ylab="Residuals", xlab=" fitted values",
     main="Residual plot")
abline(h=0, col="Red")

multi.stdres = rstandard(multi.fit)
qqnorm(multi.stdres,
       ylab="Standardized Residuals",
       xlab="Normal Scores",
       main="Normal Q-Q plot")
qqline(multi.stdres)


#Applying log transformation on SalePrice for validation data set
valid.ln_SalePrice = log(valid.dat$SalePrice)
hist(valid.ln_SalePrice, main="Histogram of log(SalePrice) from Validation Model")
multi.fit = lm(valid.ln_SalePrice~WoodDeckSF +  OpenPorchSF + EnclosedPorch +
                 ScreenPorch, data=valid.dat)
summary(multi.fit)

multi.res = resid(multi.fit)
fitted=fitted(multi.fit)
plot(fitted, multi.res,
     ylab="Residuals", xlab=" fitted values",
     main="Residual plot")
abline(h=0, col="Red")

multi.stdres = rstandard(multi.fit)
qqnorm(multi.stdres,
       ylab="Standardized Residuals",
       xlab="Normal Scores",
       main="Normal Q-Q plot")
qqline(multi.stdres)

#Finding AIC values
multi.fit = lm(valid.ln_SalePrice~WoodDeckSF +  OpenPorchSF + EnclosedPorch +
                 ScreenPorch, data=valid.dat)
```

```
step = stepAIC(multi.fit, direction = "both")
step$anova

#Model-building set with summary
build.dat=train[731:1460, ]
build.ln_SalePrice = log(build.dat$SalePrice)
multi.fit = lm(build.ln_SalePrice~WoodDeckSF +  OpenPorchSF + EnclosedPorch +
                 ScreenPorch, data=build.dat)
summary(multi.fit)
plot(multi.fit)

#Finding MSPE and MSRes
multi.fit = lm(valid.ln_SalePrice~WoodDeckSF +  OpenPorchSF + EnclosedPorch +
                 ScreenPorch, data=valid.dat)
summary(multi.fit)
SS.multi = sum(resid(multi.fit)^2)
MSPE = SS.multi/730

multi.fit = lm(build.ln_SalePrice~WoodDeckSF +  OpenPorchSF + EnclosedPorch +
                 ScreenPorch, data=build.dat)
summary(multi.fit)
SS.multi = sum(resid(multi.fit)^2)
MSRes = SS.multi/725


#Looking for improper functional form
multi.fit = lm(build.ln_SalePrice~WoodDeckSF +  OpenPorchSF + EnclosedPorch +
                 ScreenPorch, data=build.dat)
multi.res = resid(multi.fit)
fitvalues=fitted(multi.fit)

par(mfrow=c(2,3))
plot(build.dat$WoodDeckSF, multi.res,
     ylab="Residuals", xlab="WoodDeckSF",
     main="Residual plot")
abline(h=0, col="Red")
plot(build.dat$OpenPorchSF, multi.res,
     ylab="Residuals", xlab="OpenPorchSF",
     main="Residual plot")
abline(h=0, col="Red")
plot(build.dat$EnclosedPorch, multi.res,
     ylab="Residuals", xlab="EnclosedPorch",
     main="Residual plot")
abline(h=0, col="Red")
plot(build.dat$ScreenPorch, multi.res,
     ylab="Residuals", xlab="ScreenPorch",
     main="Residual plot")
abline(h=0, col="Red")

plot(fitted, multi.res,
     ylab="Residuals", xlab="Fitted Values",
     main="Residual plot")
abline(h=0, col="Red")
```

```r
#Finding studentized deleted residuals for outliers in y-direction
t <- rstudent(multi.fit)
Pii <- hatvalues(multi.fit)
n <- length(build.ln_SalePrice)
alpha <- 0.05
p_prime = length(coef(multi.fit))
t_crit <- qt(1-alpha/(2*n),n-p_prime-1)
round(t,2)
t_crit
which(abs(t) > t_crit)


#Testing for multicollinearity
VIF <- vif(multi.fit)
VIF
VIFbar <- mean(vif(multi.fit))
VIFbar


#Diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # yields 4 graphs/page
plot(multi.fit)

#Finding projection matrix value for outliers in x-direction
sum(Pii)
temp = sum(Pii) *2 / 730
which (Pii > temp)

#Finding cooks distance for influential points
cooks.distance(multi.fit)
border = qf(.95, df1=5, df2=725)
which(cooks.distance(multi.fit) > border)
```

Appendix: R code for Qualitative Model

Khizer Asad

All R Code

```
setwd("C:/Users/khize/OneDrive/Desktop/STA302/Project/Datasets")
train<-read.csv("train.csv", header=TRUE)
library(plyr)
library(MASS)
library(car)

#Random sampling for the training and test sets, code taken from Dionysius
set.seed(1002605120)
sample_size<-floor(.5*nrow(train))
picked<-sample(seq_len(nrow(train)), size=sample_size)
training<-train[picked,]
testing<-train[-picked,]

#Function to sort the non-binary variables in chronological order by means
# for discrete predictors
test <- lm(formula = SalePrice ~   Foundation  , data = training)
sorter <- function(set) {
  counts <- count(set)
  options <- nrow(counts)
  prices <- integer(options)
  for (i in 1:length(set)) {
    for (x in 1:options) {
      if (set[i] == counts[x,1]) {
        prices[x] <- prices[x] + training$SalePrice[i]
      }
    }
  }
  prices <- prices/counts[2]
  output <- data.frame(counts, prices)
  output<- output[order(output$freq.1),]
  D <- set
  for (i in 1:nrow(output)) {
   D[D == toString(output[i,1])] <- i
  }
  return(as.numeric(D))
}

#Defining the final predictors and response variable
SalePrice <- training$SalePrice
Neighborhood <- sorter(training$Neighborhood)
Exterior <- sorter(training$Exterior1st)
Type <- training$BldgType  ## BUILDING TYPE 3
```

```r
Type[Type == "2fmCon" ] <- "Twnhs"
Type[Type == "Duplex" ] <- "Twnhs"
Type[Type == "TwnhsE" ] <- "Twnhs"
Type[Type == "Twnhs" ] <- "Attached"
Type[Type == "1Fam" ] <- "Detached"
CentralAir <- training$CentralAir

#Regression model
categ <- lm(formula = log(SalePrice) ~  CentralAir  + Type + Neighborhood +
Exterior)

#Explanatory Analysis section figures
Type.I <- as.numeric(factor(Type))
CentralAir.I <- as.numeric(factor(CentralAir))
all.data <- data.frame(log(SalePrice), Neighborhood, Exterior, Type.I,
CentralAir.I)
pairs(all.data)
cor(all.data)

#Model Development
# Significance of Estimates
summary(categ)
# Model Selection
step <- stepAIC(categ, direction = "both")
step$anova
# Model Validation
CentralAir.t <- testing$CentralAir
Type.t <- testing$BldgType
Type.t[Type.t == "2fmCon" ] <- "Twnhs"
Type.t[Type.t == "Duplex" ] <- "Twnhs"
Type.t[Type.t == "TwnhsE" ] <- "Twnhs"
Type.t[Type.t == "Twnhs" ] <- "Attached"
Type.t[Type.t == "1Fam" ] <- "Detached"
Neighborhood.t <- sorter(testing$Neighborhood)
Exterior.t <- sorter(testing$Exterior1st)
SalePrice.t <- testing$SalePrice
test.reg <- lm(formula = log(SalePrice.t)~ CentralAir.t + Type.t +
Neighborhood.t + Exterior.t )
anova(test.reg)

# Diagnostics
# IFF
Residual <- resid(categ)
b.type <- as.numeric(factor(Type))
b.type[b.type == 1] <- 0
b.type[b.type == 2] <- 1
b.centralair <- as.numeric(factor(CentralAir))
b.centralair[b.centralair == 1] <- 0
b.centralair[b.centralair== 2] <- 1
```

```r
par(mfrow = c(2,2))
plot(Neighborhood, Residual, main = "Residual Plot for Neighborhood")
abline(h = 0, col = "red")
plot(Exterior, Residual, main = "Residual Plot for Exterior Material")
abline(h = 0, col = "red")
plot(b.type, Residual, main = "Residual Plot for Building Type")
abline(h = 0, col = "red")
plot(b.centralair, Residual, main = "Residual Plot for Central AC")
abline(h = 0, col = "red")
# Outliers, taken from week 6 code
t <- rstudent(categ)
Pii <- hatvalues(categ)
n <- length(log(SalePrice))
alpha <- 0.05
p_prime = length(coef(categ))
t_crit <- qt(1-alpha/(2*n),n-p_prime-1)
round(t,2)
t_crit
which(abs(t) > t_crit)
outlier.t <- (2*(sum(Pii)))/length(Pii)
which(Pii > outlier.t)
which(Pii > 0.5)
# Cook's Distance for influential points, taken from Shirley
CD <- cooks.distance(categ)
border =qf(.95, df1=4, df2=725)
which(CD>border)
# Multi-collinearity test
VIF <-vif(categ)
VIF
meanVIF <- sum(VIF[,3])/nrow(VIF)
meanVIF
# Diagnostic Plots
layout(matrix(c(1,2,3,4),2,2))
plot(categ)

#Allocated Codes legend code
all.codes <- function(set) {
  counts <- count(set)
  options <- nrow(counts)
  prices <- integer(options)
  for (i in 1:length(set)) {
    for (x in 1:options) {
      if (set[i] == counts[x,1]) {
        prices[x] <- prices[x] + train$SalePrice[i]
      }
    }
  }
  prices <- prices/counts[2]
  allocated_code <- c(1:options)
  output <- data.frame(counts, prices)
```

```
  output<- output[order(output$freq.1),]
  output$allocated_code <- allocated_code
  return(output[,c(1,4)])
}
all.codes(train$Neighborhood)
all.codes(train$Exterior1st)
```

Allocated Codes for Neighborhood and Exterior varibales:

```
all.codes(train$Neighborhood)
       x allocated_code
1 MeadowV              1
  IDOTRR               2
  BrDale               3
  BrkSide              4
  Edwards              5
3 OldTown              6
9  Sawyer              7
  Blueste              8
3   SWISU              9
5 NPkVill             10
3   NAmes             11
2 Mitchel             12
 SawyerW              13
7  NWAmes             14
  Gilbert             15
  Blmngtn             16
  CollgCr             17
  Crawfor             18
  ClearCr             19
1 Somerst             20
5 Veenker             21
4  Timber             22
2 StoneBr             23
5 NridgHt             24
4 NoRidge             25
```

```
all.codes(train$Exterior1st)
       x allocated_code
 BrkComm              1
 AsphShn              2
  CBlock              3
 AsbShng              4
 MetalSd              5
4 Wd Sdng             6
5 WdShing             7
2  Stucco             8
 HdBoard              9
 Plywood             10
 BrkFace             11
3 VinylSd            12
 CemntBd             13
1   Stone             14
 ImStucc             15
```

Group Member Contributions

Khizer Asad: Completed the model development and analysis of Qualitative factors, and worked on proof-reading as well as some formatting.
Shirley Ching: Worked on the Exterior category model of sale price prediction in Section 3 along with proofreading of other sections.
Dionysius Indraatmadja: Developed and created the analysis for the Premium model sections, wrote the introduction to the Section 3 Model Development, and contributed to Section 1 Introduction and the introduction to Section 2 Exploratory Data Analysis.
Daleep Singh: Wrote the interior exploratory analysis, the interior model development, interior conclusion, helped proof read.