STAA 551 Case Study - Home Price Regression Model

Nelson Brown and Bryce Smith

Introduction

This case study focuses on creating a regression model based on a data set of transactions provided for the Seattle area in with properties of homes such as their date built, number of bedrooms, square footage of the living space, and other features that will be described in the next section. We explore the methods of building this model up from those predictors which provide the strongest correlation with our price response variable, removing those predictors which have collinearity with others included in the model, explore interactions, and examine those models which explain the variation in the response. We perform a diagnostic and residual analysis on the best performing model to ensure it does not violate basic assumptions of the OLS model, and utilize transformations to modify the model to fit within those assumptions at the cost of a marginal amount of explained variation. Finally, we examine the results of the final model and describe some of the uses and limitations of the model.

Summary Statistics and Graphics

Our housing data frame had a number of quantitative predictors. One set are square footage of features of the house such as `sqft_living` or the amount of total space that can be lived in (sans attic, basement, or garage space), `sqft_lot` or square footage of the entire property, `sqft_above` which is living space above the basement level, and `yr_built` or the year the home was built. The first few rows of these values are displayed in the table below.

Table 1: First Four Rows for Quantitative Values on Seattlle Housing Dataframe

price	sqft_living	sqft_lot	bedrooms	bathrooms	view	grade	sqft_above	yr_built
359,950	1,570	6,975	3	1.750	0	7	1,040	1,979
909,950	3,050	8,972	5	3.750	0	9	3,050	2,014
318,000	1,570	12,506	3	1.750	0	8	1,570	1,959
272,000	1,390	10,660	4	1.750	0	7	1,030	1,960
475,000	2,320	10,046	4	2.500	0	7	2,320	2,006
907,000	1,340	6,000	3	1.500	1	9	1,340	1,927

Summary statistics for these data points (there exist 613 total rows in the dataset with no missing values) are shown in Table @ref(summary-stats).

Text text text.

Quantitative Values

Text text text.

Text text text.

Table 2: Summary Statistics for Values on Seattle Housing Dataframe

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
price	545,427.700	408,545.900	95,000	315,000	631,500	5,300,000
sqft_living	2,073.669	963.763	380	1,370	2,550	7,390
sqft_lot	15,967.970	46,698.890	740	5,100	10,585	871,200
bedrooms	3.352	0.876	1	3	4	7
bathrooms	2.092	0.805	0.500	1.500	2.500	6.000
view	0.204	0.695	0	0	0	4
grade	7.635	1.217	5	7	8	12
sqft_above	1,793.571	873.153	380	1,130	2,313	6,530
yr_built	1,971.210	29.939	1,900	1,951	1,998	2,015

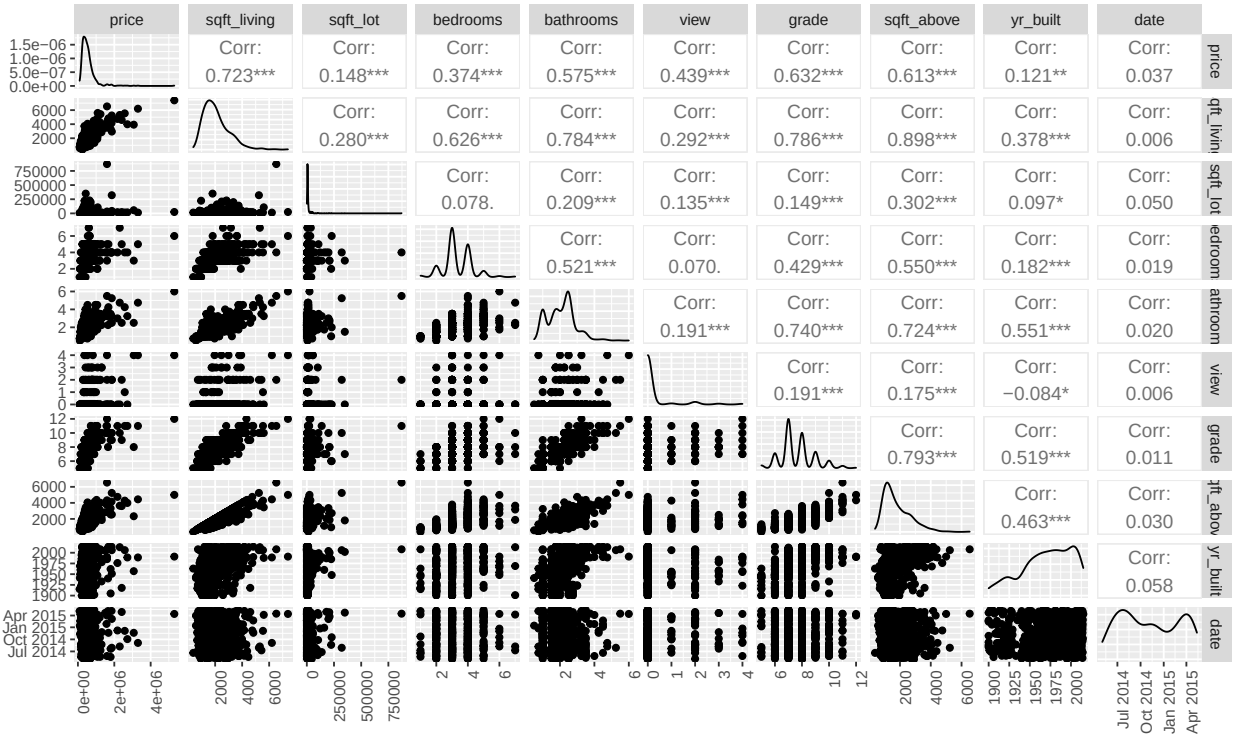

Figure 1: Correlogram of Quantitative or Ordinal Predictors and Response Variable

Discrete and Categorical Values

Text text text

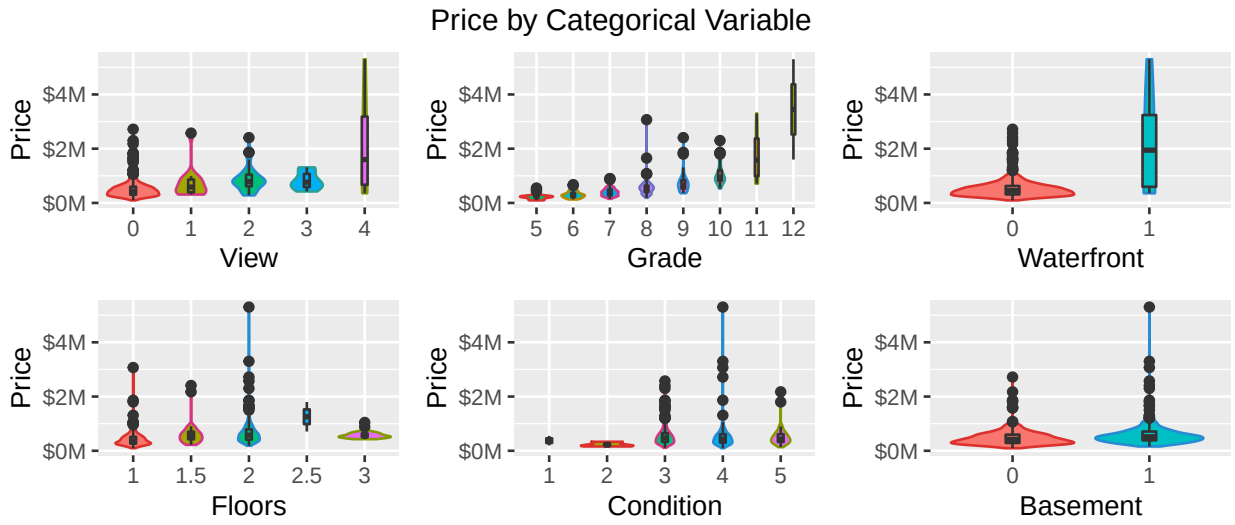

Figure 2: Violin Plots of Categorical Values Interpreted as Factors

Analysis

Bottom-Up Model from Variable Added Last t-test

This analysis began with analyzing a correlogram of the dataset which was presented in figure ___. Since there was a single most correlated predictor `sqft_living` with our response variable `price`, and the second highest predictor `sqft_above` was a directly proportional quantity to `sqft_living` which yielded high colinearity we decided to build a model bottom-up starting with `price` predicted by `sqft_living`. We would sequentially add predictors to the model while assessing model performance.

As seen in model (1) of table 3 on the next page, it was clear that we had strong evidence for a linear association between price and `sqft_living` and that the adjusted R^2 amount of explained variation accounted for roughly 52% of the variation in price. This would remain throughout our analysis the most significant predictor that we could ascertain contributed to the model. We then proceeded to add `grade` to the model. Adding `grade` to the model proved to be helpful, as we noticed a significant p-value for the variables added last test which was 0.000251.

Next we added bathrooms into our model and received a p-value of 2.11e-04 for the variables added last t-test so it was safe to conclude that it was not helpful to our model with `sqft_living` and `grade` already included. However, our next variable we added to our model, `view`, did prove to be helpful with a p-value of 2e-16.

The model with these three variables determine an adjusted R^2 which explained roughly 59% of the variation. This model is shown as model (2) in table 4. We next looked at bathrooms and bedrooms into the model based on the correlation with price. `bedrooms` was chosen first as it had only a 0.428 correlation with `grade` versus the 0.740 correlation of `bathrooms` with `grade`. It was clear with the results from the summary table that bedrooms hardly explained any more variation in price, and was not helpful in our model so we removed that predictor. `bathrooms` showed a similar result and was not included in our analysis.

We did want to try the variable `yr_built` in the model since we believed it was important to attempt to add a variable that represented the age of the house. Adding `yr_built` into our model proved to be significant with `sqft_living`, `grade` and `view` already present in the model. The results of this model are annotated as model (2) in Table 3.

Adding Categorical Indicator Predictors

At this point in our analysis we concluded to stop adding quantitative and quantitative monotonic ordinal variables, and looked at the categorical variables which could be treated as indicator or dummy variables in our model. Starting with waterfront, we found that this indicator was helpful in our model and increased our adjusted R^2 value to approximately 0.65.

Adding basement and renovated as indicators to our model, in the same fashion as we added waterfront to the model, ended up not being helpful to our model with `sqft_living`, `grade`, `view`, and `yr_built` already present. We also added these two indicators, separately, with `waterfront` in the model, and these predictors added were not helpful either. We were satisfied with the five predictors: `sqft_living`, `grade`, `view`, `yr_built`, and `waterfront` being used in our model.

Comparison of Bottom-Up Model with Top-Down Model using ANOVA tables

We wanted to ensure that our model doesn't disagree with an alternative analysis built from a full model in that we did not miss any potential predictors that may add to our model's explained variation while offering strong evidence that they linearly associated with the predictor. To accomplish this, we first compared a model with nearly every predictor with the exception of those predictors that could be eliminated due to their intrinsic nature or prior inspection. For instance, the id field or date of transaction was not considered in the full model. An ANOVA was conducted between these models with an F-test statistic value of 3.969

and associated p-value 6.64e-04 indicated that their may be some predictors which may be linearly associated with the response variable.

We then removed the predictors from the nearly full model by examining those with the weakest p-value evidence, and ended up with a reduced model which discarded `floors`, `sqft_above`, and `sqft_lot`. The reduced model when compared with the full model through an ANOVA table had a F-test statistic of 2.002 and associated p-value of 0.093 which gave fair evidence that the removed predictors were not linearly associated with `price`. Comparing the adjusted R^2 value of this reduced model of 0.661 with the model we arrived out built bottom-up analysis using variable last added t-tests from initial predictors and it's adjusted R^2 of 0.621 indicates that the additional explained variation is minimal.

Exploring Interactions

Proceeding with model 2 in our analysis, we needed to assess the correlations between our predictors. When doing this, it was clear that we had evidence of colinearity between `sqft_living` and `grade`. With this knowledge we sought to assess whether an interaction between these two variables would be helpful to our model. As seen in model (3) in Table 3, adding this interaction term proved to be helpful in our model and increased the variation in `price` explained by the model to 0.706 as expressed in the adjusted R^2. At this point, we moved on to the diagnostic plots and residuals analysis. Models (4) and (5) in Table 3 will be explained in the following section.

Table 3: Iteratively Built Model Comparison

	Dependent variable:				
	price			log(price)	
	(1)	(2)	(3)	(4)	(5)
sqft_living	306.516***	210.259***	−338.813***		
	(11.846)	(17.565)	(52.678)		
log(sqft_living)				0.185	0.426***
				(0.147)	(0.046)
grade		103,280.400***	−22,648.160	−0.025	0.236***
		(14,665.420)	(18,045.770)	(0.153)	(0.018)
view		128,209.200***	63,479.420***	0.109***	0.115***
		(15,686.520)	(16,038.800)	(0.023)	(0.020)
yr_built		−2,834.238***	−2,411.571***	−0.004***	−0.005***
		(406.927)	(361.830)	(0.001)	(0.001)
waterfront			668,962.300***	0.026	
			(110,486.600)	(0.157)	
sqft_living:grade			61.812***		
			(5.796)		
log(sqft_living):grade				0.033*	
				(0.019)	
Constant	−90,183.990***	4,881,654.000***	5,119,662.000***	18.658***	16.976***
	(27,085.020)	(764,312.600)	(673,590.900)	(1.387)	(0.992)
Observations	613	613	613	613	613
R^2	0.523	0.624	0.709	0.648	0.646
Adjusted R^2	0.522	0.621	0.706	0.644	0.643
Residual Std. Error	282,443.100 (df = 611)	251,492.200 (df = 608)	221,542.500 (df = 606)	0.319 (df = 606)	0.320 (df = 608)
F Statistic	669.475*** (df = 1; 611)	251.761*** (df = 4; 608)	245.871*** (df = 6; 606)	185.617*** (df = 6; 606)	277.155*** (df = 4; 608)

Note: *p<0.1; **p<0.05; ***p<0.01

Diagnostic Plots and Residuals Analysis

Satisfied with the model derived up to this point , we proceeded to check the diagnostic plots to check the validity of our assumptions. After review of these plots, it was clear we had problems addressing our assumptions of linearity, mean zero for the errors and constant variance of the errors. This can be shown with the fanning out of the points in the residuals vs fitted plot, and the lack of linearity in the normal QQ plot. We can also see that we have observations with very large Cook's Distance, which may be negatively affecting our model. Checking the residuals versus the predicted values confirmed what we saw in the previous three plots and we concluded that our assumptions were violated.

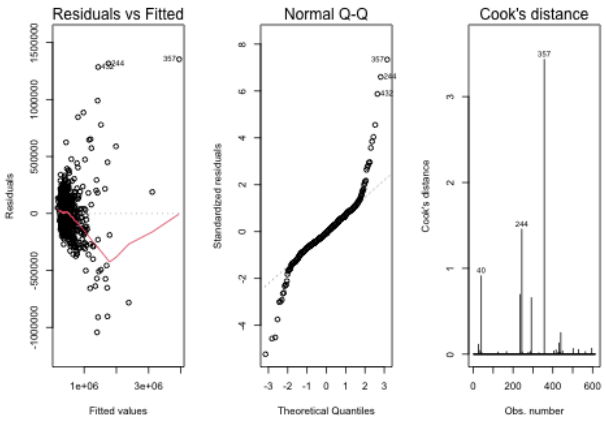

Figure 3: Diagnostic Plots

To address the violations of our assumptions, we entertained the idea of transforming some of the variables used in the model. We could see from the correlogram that both our response, `price`, and predictor, `sqft_living`, had heavily right skewed distributions. We decided to apply a logarithmic transformation after observing a box-cox plot indicating that the power transformation would be justified. We reapplied these into our model.

Doing so we saw our interaction between the now logarithmic transformation of sqft_living and grade become not helpful to our model. We also, saw that the indicator, waterfront, was no longer helpful to the model. This can be seen in model (4) of Table 3. In an effort to satisfy our assumptions of the model, it was decided that removing both of these variables from the model was necessary. This is shown in our final model (5) of Table 3.

As a quick check we looked into the VIF values for the predictors and determined that we weren't worried about collinearity between the predictors.

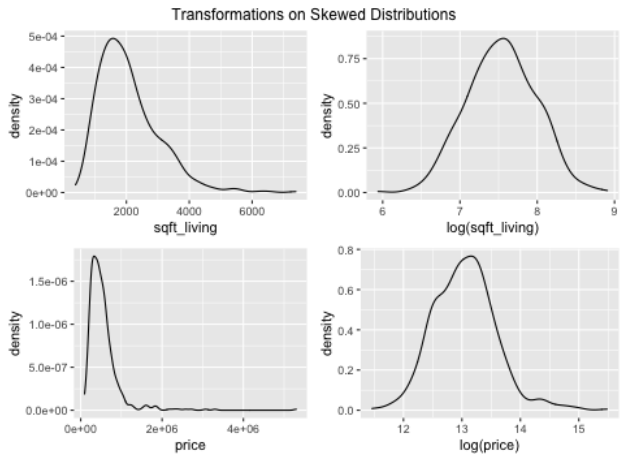

Figure 4: Transformations of Skewed Distributions

Table 4: VIF for Final Price Model

log(sqft_living)	grade	view	yr_built
2.543	2.874	1.112	1.438

Lastly, in an effort to address the values that showed large Cook's Distance in the diagnostic plots, we compared which leverage points in our model also had large residuals. We found observations that satisfied both these criteria, and concluded that these points needed to be addressed with the client.

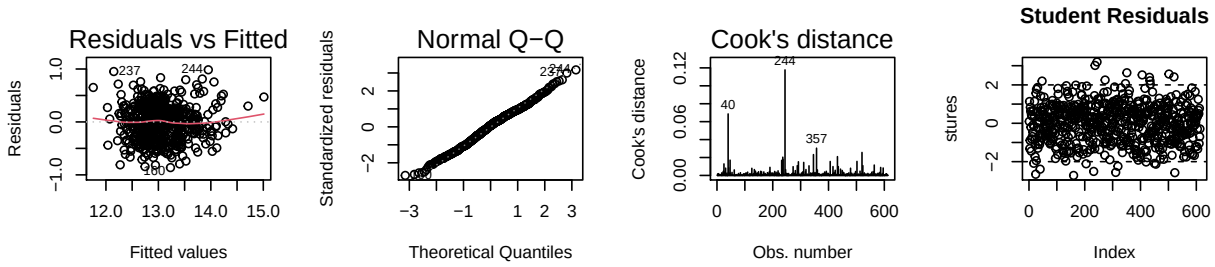

Figure 5: Diagnostic Plots of Final Model

Table 5: High Student Residual and Leverage Values

	price	sqft_living	view	grade	yr_built
40	357,000	2,460	4	7	1,955
47	550,000	1,660	0	5	1,933
244	3,070,000	3,930	4	8	1,957
346	245,000	380	0	5	1,963
432	2,720,000	3,990	0	11	1,989
520	675,000	930	2	6	1,951

Results and Conclusions

Our final fitted model is given by the following

$$\log(\text{price}) = 16.98 + 0.43\log(\text{sqftliving}) + 0.24\text{grade} + 0.12\text{view} - 0.0045\text{yr\_built}$$

log(sqft_living): Holding all else constant, a one percent increase in the living square footage of a home would yield a 0.426% increase in the average price of the home. This can be more easily understood as a 10 percent increase in the living square footage of a home would yield a 4.26% increase in the average price of the home.

grade Holding all else constant, a one level increase in grade will result in an increase in average price of 26.67 percent.

view Holding all else constant, a one level increase in view will result in an increase in average price of 12.23 percent.

yr_built Holding all else constant, a one year increase in the year the house was built will result in a -.45% decrease in average housing price.

Using this model to predict housing price would be beneficial for a stakeholder with the understanding of some key aspects of the model. We will first start with the year the house was built. In our model we have determined that if all else in the model was held constant, the price of a house built closer to present day, then say 5 years beforehand, would produce a lower estimation of price. Lets look at an example of this. Using a house with a living square footage of 2000, housing unit grade of 10, and have been viewed 1 time by a potential buyer, we would see a price of $839,675.10 if it was built in 1990. Holding living square footage, grade, and view constant, and we adjust that house to be built in 2000, we would then estimate an average price of $802,326.9, result in a decrease in housing price of $37,348.24.

Appendix: All Code for This Report

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, dev='pdf')
library(car)
library(ggplot2)
library(grid)
library(gridExtra)
library(ggthemes)
library(lubridate)
library(GGally)
library(scales)
library(dplyr)
library(stargazer) # Used for latex tables to summarize the data and models

# SECTION: HEADER AND INTRODUCTION, INITIAL DATA READ

# Helper function for formatting
sci.notation <- function(value) {
  formatC(value, format = "e", digits = 2)
}

# Read the Data
df <- read.csv('Seattle.csv', strip.white = TRUE, stringsAsFactors = FALSE)
# Clean the Data
df$date <- ymd(substr(df$date,1,nchar(df$date) - 7)) # Convert string to date object
df$was_renovated <- as.factor(with(df,
     ifelse(yr_renovated > 0,
            1,
            0
          )
 ))
df$has_basement <- as.factor(with(df,
     ifelse(sqft_basement > 0,
            1,
            0
          )
 ))

# SECTION: SUMMARY STATISTICS AND GRAPHCIS

# Define our quantitative values of interest
df.quant <- df %>% dplyr::select(price,
                          sqft_living,
                          sqft_lot,
                          bedrooms,
                          bathrooms,
                          view,
                          grade,
                          sqft_above,
                          yr_built,
                          date)

# Print head of initial dataframe
```

```
stargazer(head(df.quant %>% dplyr::select(-date)),
          rownames=FALSE,
          summary=FALSE,
          header=FALSE,
          title="First Four Rows for Quantitative Values on Seattlle Housing Dataframe")

# Summarize initial dataframe
stargazer(df.quant,
          header=FALSE,
          omit.summary.stat=c('N'),
          title="Summary Statistics for Values on Seattle Housing Dataframe",
          label="summary-stats")

# SECTION: QUANTITATIVE VALUES

# Quantitative Data Pairs
ggpairs(df.quant, progress=FALSE) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# SECTION: DISCRETE AND CATEGORICAL VALUES

plot_price_by_cat <- function(df, cat_var, cat_var_name) {
  ggplot(df, aes(x=cat_var, y=price, fill=cat_var)) +
    scale_colour_solarized("red") +
    geom_violin(aes(color=cat_var)) +
    geom_boxplot(width=0.1) +
    xlab(cat_var_name) +
    ylab("Price") +
    scale_y_continuous(labels = scales::dollar_format(scale = .000001, suffix="M")) +
    theme(legend.position="none")
}

q <- plot_price_by_cat(df, as.factor(df$view), "View")
r <- plot_price_by_cat(df, as.factor(df$grade), "Grade")
t <- plot_price_by_cat(df, as.factor(df$waterfront), "Waterfront")
w <- plot_price_by_cat(df, as.factor(df$floors), "Floors")
x <- plot_price_by_cat(df, as.factor(df$condition), "Condition")
y <- plot_price_by_cat(df, as.factor(df$has_basement), "Basement")

grid.arrange(grobs=list(q, r, t,
                        w, x, y),
             ncol=3,
             top="Price by Categorical Variable")

# SECTION BOTTOM-UP MODEL FROM VARIABLE ADDED LAST t-TEST

lm.price.1 <- lm(formula = price ~ sqft_living, data = df)
lm.price.2 <- lm(formula = price ~ sqft_living + grade, data=df)
lm.price.3.not.included <- lm(formula = price ~ sqft_living + grade + bathrooms,
                              data = df)
lm.3.pvalue <- sci.notation(summary(lm.price.3.not.included)$coefficients[3,4])
lm.price.4 <- lm(formula = price ~ sqft_living + grade + view,
            data = df)
```

```r
lm.price.5 <- lm(formula = price ~
                   sqft_living +
                   grade +
                   view +
                   yr_built,
                 data = df)

# SECTION: ADDING CATEGORICAL INDICATOR PREDICTORS

lm.price.5b <- lm(formula = price ~
                   sqft_living +
                   grade +
                   view +
                   yr_built +
                   waterfront,
                  data = df)

# SECTION: COMPARISON OF BOTTOM-UP MODEL WITH TOP-DOWN MODEL USING ANOVA

# Top-down model building using ANOVA analysis between a nearly
# full model and our bottom-up analysis, a nearly full model
# and a reduced model, and finally comparative metrics between
# the reduced model and the model built from a bottom-up analysis.


lm.nearfull <- lm(price ~
                    bedrooms +
                    bathrooms +
                    sqft_living +
                    sqft_lot +
                    floors +
                    waterfront +
                    view +
                    condition +
                    grade +
                    sqft_above +
                    yr_built , data=df)
anova.1 <- anova(lm.nearfull, lm.price.5b)
anova.1.p.value <-formatC(anova.1$`Pr(>F)`[2], format = "e", digits = 2)
lm.reduced <- lm(price ~ bedrooms +
                   bathrooms +
                   sqft_living +
                   waterfront +
                   view +
                   grade +
                   yr_built, data=df)
anova.2 <- anova(lm.nearfull, lm.reduced)
anova.2.p.value <-round(anova.2$`Pr(>F)`[2], 3)
# collect adjusted R^2
nearfull.r.adjusted <- summary(lm.nearfull)
reduced.r.adjusted <-round(summary(lm.nearfull)$adj.r.squared,3)
lm5.r.adjusted <-round(summary(lm.price.5)$adj.r.squared,3)
```

```r
# SECTION: EXPLORING INTERACTIONS

lm.price.6 <- lm(formula = price ~
                   sqft_living +
                   grade +
                   view +
                   yr_built +
                   waterfront +
                   sqft_living*grade,
                   data = df)

# Corrections and transformations applied based on diagnostic plots
# (performed before diagnostic plots as to provide input to table 4
#  used in the report which we didn't want separated too far from
#  the bottom-up model building section of analysis)

lm.price.7 <- lm(formula = log(price) ~
                   log(sqft_living) +
                   grade +
                   view +
                   yr_built +
                   waterfront +
                   log(sqft_living)*grade,
                   data = df)
# Final model
lm.price.8 <- lm(formula =
                   log(price) ~
                   log(sqft_living) +
                   grade +
                   view +
                   yr_built,
                   data = df)
lm.price.final <- lm.price.8
stargazer(lm.price.1,
          lm.price.5,
          lm.price.6,
          lm.price.7,
          lm.price.8,
          header=FALSE,
          align=T,
          label="ModelComparison",
          title="Iteratively Built Model Comparison")

# SECTION: DIAGNOSTIC PLOTS AND RESIDUAL ANALYSIS

# In order to save space, we had to flow some text around these
# diagnostic figures, but I think it worked well to keep the
# figures near their explanation.

png("diagnostics_prior_to_transform.png", height = 350)
# 2. Create the plot
par(mfrow = c(1,3))
plot(lm.price.6,c(1,2,4))
```

```r
# 3. Close the file
dev.off()
# Box-cox examining most likely model predictor exponent
png("boxcox.png", height = 350)
boxCox(lm.price.6, main="Box-Cox Plot of model")
dev.off()

png("transformations.png", height = 350)
par(mfrow = c(1,2))
p <- ggplot(df, aes(x=sqft_living)) + geom_density()
q <- ggplot(df, aes(x=log(sqft_living))) + geom_density()
r <- ggplot(df, aes(x=price)) + geom_density()
s <- ggplot(df, aes(x=log(price))) + geom_density()
grid.arrange(grobs=list(p, q,
                        r, s),
             ncol=2,
             top="Transformations on Skewed Distributions")
dev.off()

# VIF Table
out <- capture.output(stargazer(vif(lm.price.final),
          title="VIF for Final Price Model",
          header=FALSE,
          align=FALSE))

out <- sub(" \\\\centering", "", out)
cat(out)
# Student Residuals and Hat Values Computed
par(mfrow = c(1,4))
plot(lm.price.final,c(1,2,4))
stures <- rstudent(lm.price.final)
plot(stures, main="Student Residuals")
abline(h=2, lty=2)
abline(h=-2, lty=2)
# Calculate intersection of student residuals with high leverage
hv <- as.matrix(hatvalues(lm.price.final))
mn <- mean(hatvalues(lm.price.final))
stures_vals <- which(abs(stures) > 2)
lev_vals <- which(hv > 2*mn)
index_values <- intersect(lev_vals,stures_vals)
df.new <- df %>% dplyr::select(price,
                        sqft_living,
                        view,
                        grade,
                        yr_built)
stargazer(df.new[index_values,],
          header=FALSE,
          title="High Student Residual and Leverage Values",
          summary=FALSE)
```