R Notebook

Nelson Brown and Bryce Smith

Introduction

This case study focuses on creating a linear model

Summary Statistics and Graphics

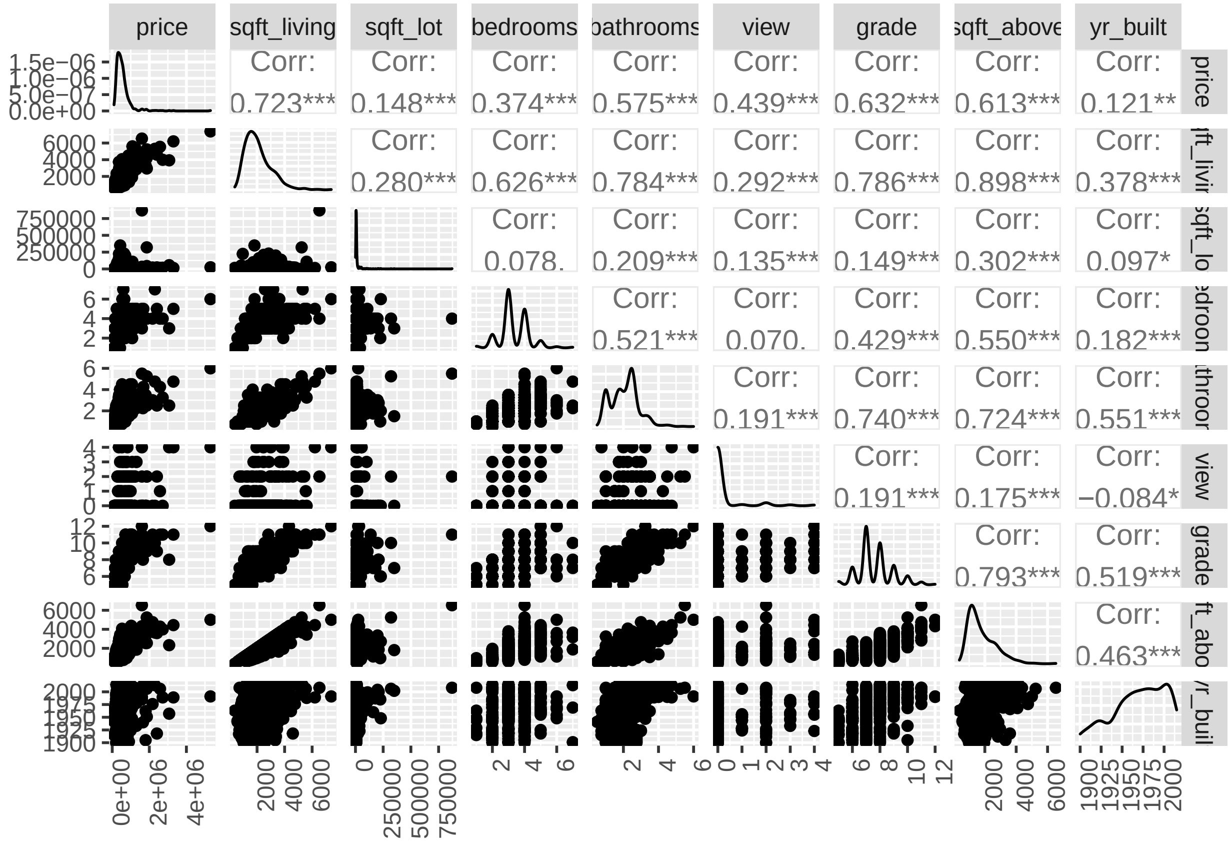
Quantitative Values

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

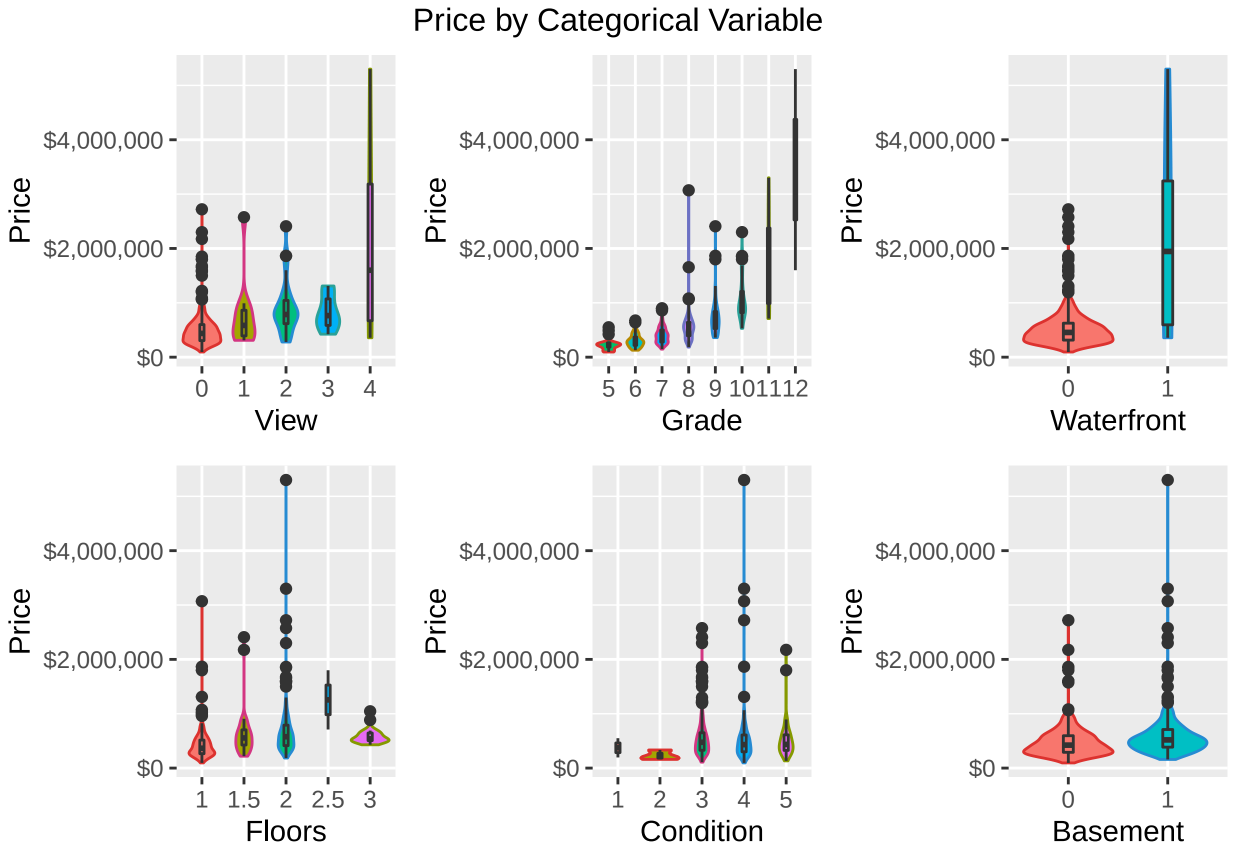
price	sqft_living	sqft_lot	bedrooms	bathrooms	view	grade	sqft_above	yr_built
359, 950	1, 570	6, 975	3	1.750	0	7	1, 040	1, 979
909, 950	3, 050	8, 972	5	3.750	0	9	3, 050	2, 014
318, 000	1, 570	12, 506	3	1.750	0	8	1, 570	1, 959
272, 000	1, 390	10, 660	4	1.750	0	7	1, 030	1, 960
475, 000	2, 320	10, 046	4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

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


price
sqft_living
sqft_lot
bedrooms
bathrooms
view
grade
sqft_above
yr_built
Corr: 0.723***
Corr: 0.148***
Corr: 0.374***
Corr: 0.575***
Corr: 0.439***
Corr: 0.632***
Corr: 0.613***
Corr: 0.121**
Corr: 0.280***
Corr: 0.626***
Corr: 0.784***
Corr: 0.292***
Corr: 0.786***
Corr: 0.898***
Corr: 0.378***
Corr: 0.078.
Corr: 0.209***
Corr: 0.135***
Corr: 0.149***
Corr: 0.302***
Corr: 0.097*
Corr: 0.521***
Corr: 0.070.
Corr: 0.429***
Corr: 0.550***
Corr: 0.182***
Corr: 0.191***
Corr: 0.740***
Corr: 0.724***
Corr: 0.551***
Corr: 0.191***
Corr: 0.175***
Corr: −0.084*
Corr: 0.793***
Corr: 0.519***
Corr: 0.463***
1.5e−06
1.0e−06
5.0e−07
0.0e+00
6000
4000
2000
750000
500000
250000
0
6
4
2
4
3
2
1
0
12
10
8
6
2000
1975
1950
1925
1900
0e+00
2e+06
4e+06
250000
500000
750000

Discrete and Categorical Values



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 9.27e-02 which gave strong 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.

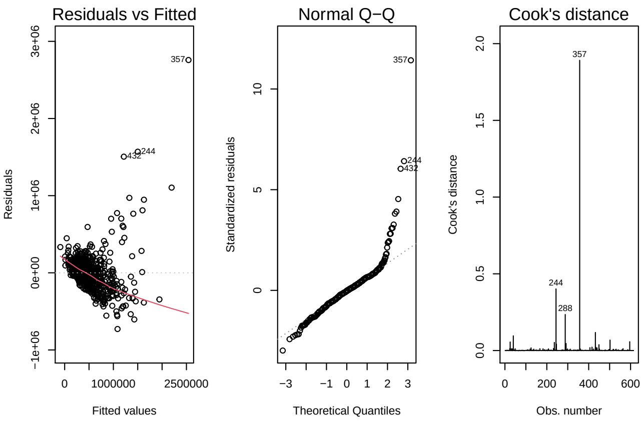
Exploring Interactions

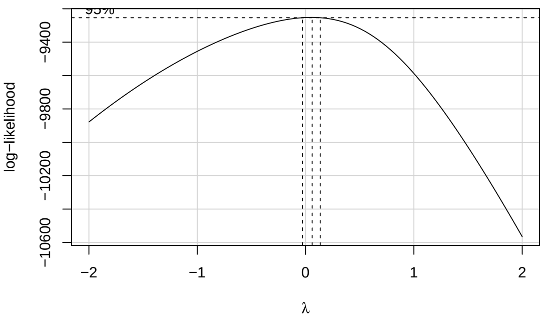
Proceeding with model 2 in our analysis, we needed to assess the correlations between our predictors. When doing this, it was clear that we had evidence of colinearity between sqft_living and grade. With this knowledge we sought to assess whether an interaction between these two variables would be helpful 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.

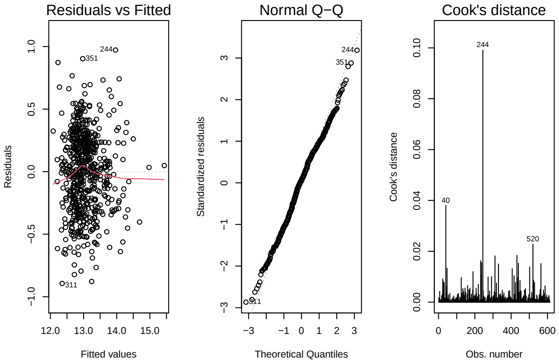
Table 3: Model comparison showing iteratively adding predictors with strong supporting evidence

	Dependent variable:				
	price			log(price)	
	(1)	(2)	(3)	(4)	(5)
sqft_living	306.516*** (11.846)	210.259*** (17.565)	−338.813*** (52.678)		
log(sqft_living)				0.185 (0.147)	0.426*** (0.046)
grade		103,280.400*** (14,665.420)	−22,648.160 (18,045.770)	−0.025 (0.153)	0.236*** (0.018)
view		128,209.200*** (15,686.520)	63,479.420*** (16,038.800)	0.109*** (0.023)	0.115*** (0.020)
yr_built		−2,834.238*** (406.927)	−2,411.571*** (361.830)	−0.004*** (0.001)	−0.005*** (0.001)
waterfront			668,962.300*** (110,486.600)	0.026 (0.157)	
sqft_living:grade			61.812*** (5.796)		
log(sqft_living):grade				0.033* (0.019)	
Constant	−90,183.990*** (27,085.020)	4,881,654.000*** (764,312.600)	5,119,662.000*** (673,590.900)	18.658*** (1.387)	16.976*** (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







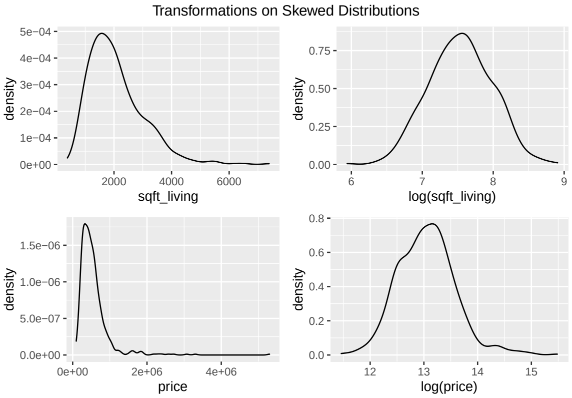


Table 4: VIF Values for Price Model

sqft_living	grade	view	yr_built	has_basement
2.982	3.146	1.155	1.473	1.131

Results and Conclusions

Appendix: All Code for This Report

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

# 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
          )
 ))

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

# Print head of initial dataframe
stargazer(head(df.quant),
          rownames=FALSE,
          summary=FALSE,
          header=FALSE,
          title="First Four Rows for Quantitative Values on Seattlle Housing Dataframe")
```

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

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

# Discrete and Categorical Values Section
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 = 1)) +
    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")
# Bottom up model formulation.

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)

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

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

# 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$`Pr(>F)`[2], format = "e", digits = 2)
lm.reduced <- lm(price ~ bedrooms +
                   bathrooms +
                   sqft_living +
                   waterfront +
                   view +
                   grade +
                   yr_built, data=df)
anova.2 <- anova(lm.nearfull, lm.reduced)
anova.2.p.value <-formatC(anova.2$`Pr(>F)`[2], format = "e", digits = 2)
# 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)


# 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
```

```r
# (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="Model comparison showing iteratively adding predictors with strong supporting evidence

# Initial Model
lm.price <- lm(formula = price ~ sqft_living + grade + view + yr_built + has_basement,
            data = df)
par(mfrow = c(1,3))
plot(lm.price,c(1,2,4))
boxCox(lm.price, main="Box-Cox Plot of model")
par(mfrow = c(1,3))
lm.price <- lm(formula = log(price) ~ sqft_living + grade + view + yr_built + has_basement,
            data = df)
plot(lm.price,c(1,2,4))
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")
# VIF Table
stargazer(vif(lm.price), title="VIF Values for Price Model", header=FALSE)
```