R Notebook

Nelson Brown and Bryce Smith

Introduction

This case study focuses on creating a linear model

Summary Statistics and Graphics

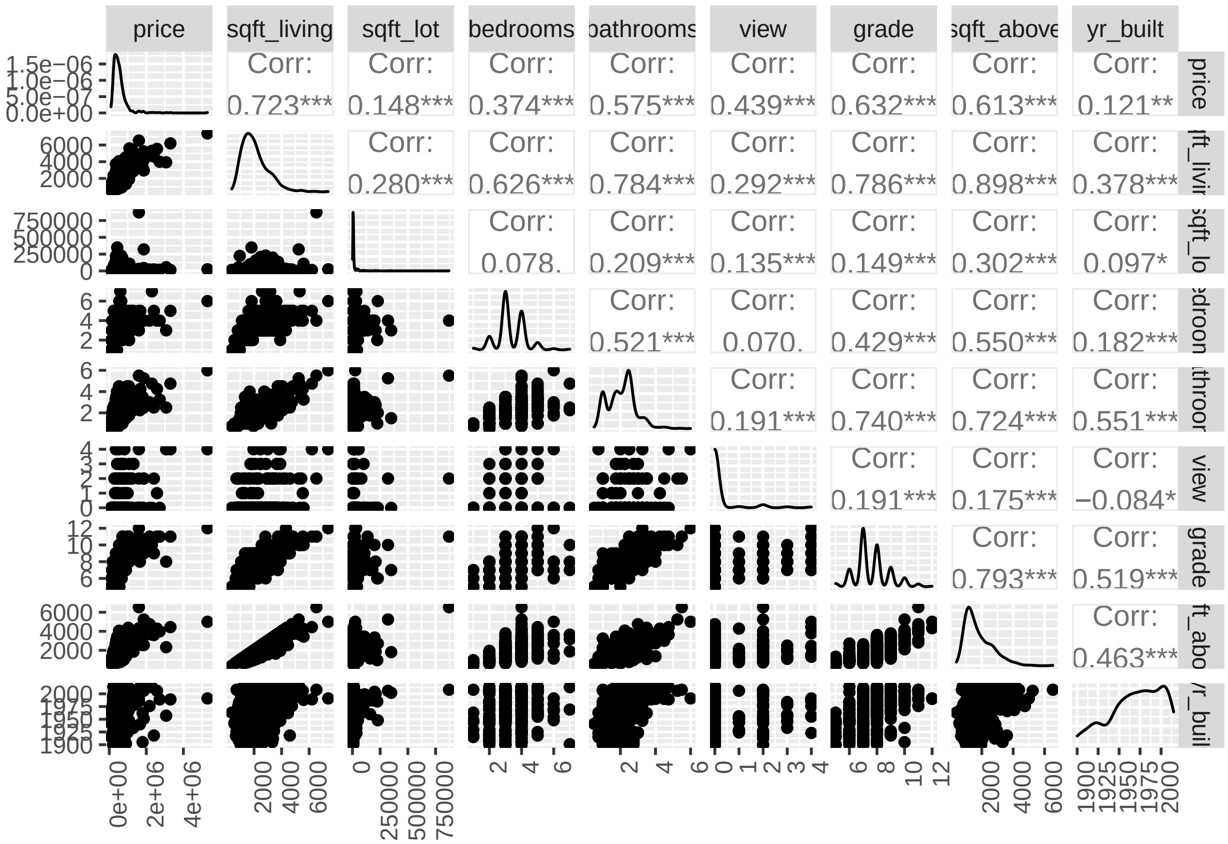
Quantitative Values

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
2000
4000
6000
0
250000
500000
750000
2
4
6
0
1
2
3
4
6
8
10
12
1900
1925
1950
1975
2000

Discrete and Categorical Values

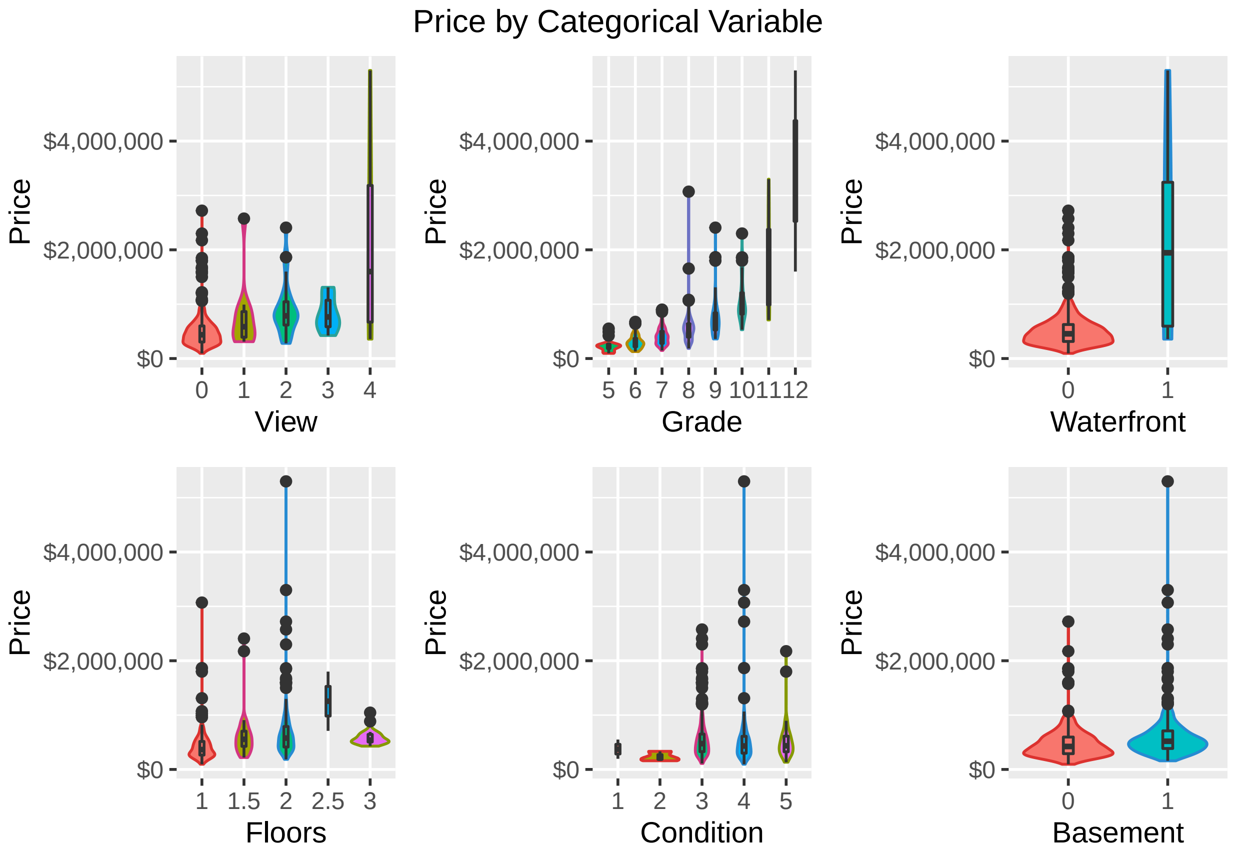


Table 3: Model comparison showing iteratively adding predictors with strong supporting evidence

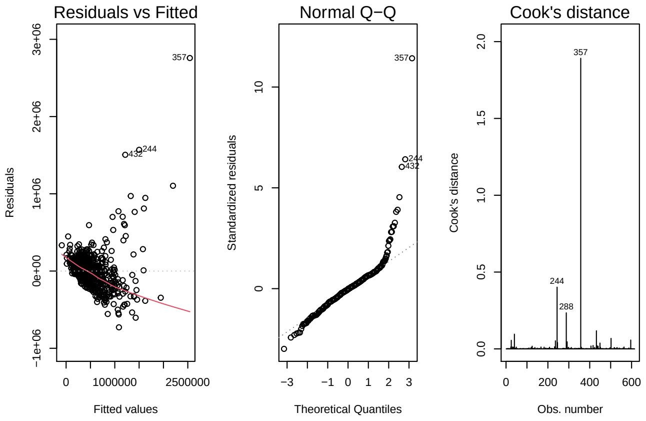
	Dependent variable:				
	price			log(price)	
	(1)	(2)	(3)	(4)	(5)
sqft_living	306.516***	197.951***	−338.813***		
	(11.846)	(16.932)	(52.678)		
log(sqft_living)				0.185	0.426***
				(0.147)	(0.046)
grade		111,296.700***	−22,648.160	−0.025	0.236***
		(14,110.000)	(18,045.770)	(0.153)	(0.018)
view		63,193.660***	63,479.420***	0.109***	0.115***
		(17,464.810)	(16,038.800)	(0.023)	(0.020)
waterfront		869,393.200***	668,962.300***	0.026	
		(118,556.800)	(110,486.600)	(0.157)	
yr_built		−2,922.920***	−2,411.571***	−0.004***	−0.005***
		(390.526)	(361.830)	(0.001)	(0.001)
sqft_living:grade			61.812***		
			(5.796)		
log(sqft_living):grade				0.033*	
				(0.019)	
Constant	−90,183.990***	5,025,533.000***	5,119,662.000***	18.658***	16.976***
	(27,085.020)	(733,417.900)	(673,590.900)	(1.387)	(0.992)
Observations	613	613	613	613	613
R^2	0.523	0.654	0.709	0.648	0.646
Adjusted R^2	0.522	0.651	0.706	0.644	0.643
Residual Std. Error	282,443.100 (df = 611)	241,240.200 (df = 607)	221,542.500 (df = 606)	0.319 (df = 606)	0.320 (df = 608)
F Statistic	669.475*** (df = 1; 611)	229.646*** (df = 5; 607)	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

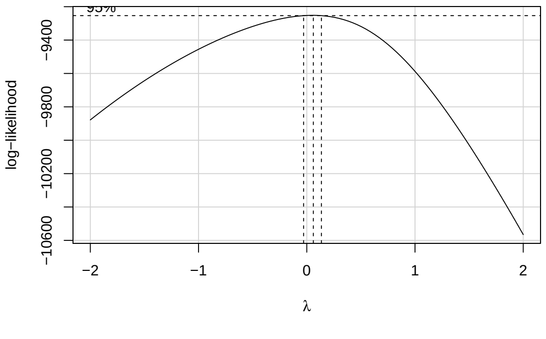
We wanted to ensure that our model doesn't disagree with an alternative analysis built from a full model in that we did not miss any potential predictors that may add to our model's explained variation while offering strong evidence that they linearly associated with the predictor. To accomplish this, 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.

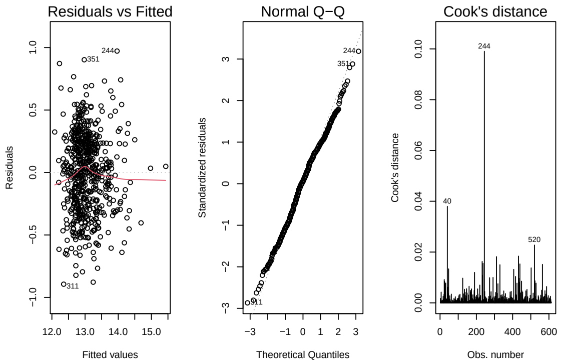
We then removed the predictors from the nearly full model by examining those with the weakest p-value evidence, and ended up with a reduced model which discarded `floors`, `sqft_above`, and `sqft_lot`. Comparing the adjusted R^2 value of this reduced model of 0.661 with the model we arrived out built bottom-up from initial predictors and it's adjusted R^2 of 0.651 indicates that the additional explained variation is minimal. Seeking a simpler model, we feel confident that our model will perform accurately after residual error and examination of the diagnostic plots.

Analysis

Initial Model







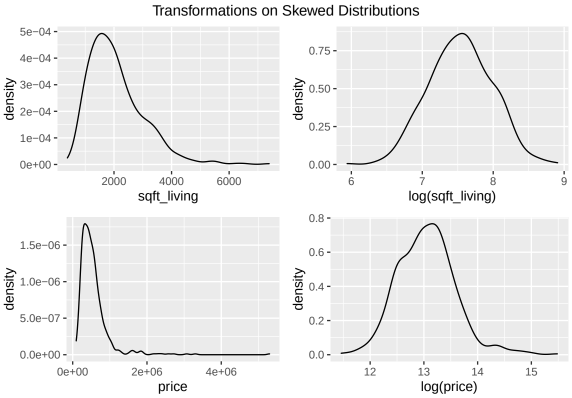


Table 4: VIF Values for Price Model

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 = TRUE, 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

# 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")

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

cor.df.quant <- cor(df.quant)

#Correlation Matrix
#invisible(stargazer(cor.df.quant,
#                     header=FALSE,
#                     title="Correlation of Quantiative Values"))

# 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")
lm.price.1 <- lm(formula = price ~ sqft_living,
            data = df)
lm.price.4 <- lm(formula = price ~ sqft_living + grade + view + waterfront,
            data = df)
lm.price.5 <- lm(formula = price ~ sqft_living + grade + view + waterfront + yr_built,
            data = df)
lm.price.6 <- lm(formula = price ~ sqft_living + grade + view + waterfront + yr_built + sqft_living*grad
            data = df)
lm.price.7 <- lm(formula = log(price) ~ log(sqft_living) + grade + view + waterfront + yr_built + log(s
            data = df)
lm.price.8 <- lm(formula = log(price) ~ log(sqft_living) + grade + view + yr_built,
            data = df)
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

# Top-down model building using ANOVA analysis between full and reduced model.

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

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.5)
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)


# 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)
```