R Notebook

Nelson Brown and Bryce Smith

Introduction

This case study focuses on creating a linear model

Summary Statistics and Graphics

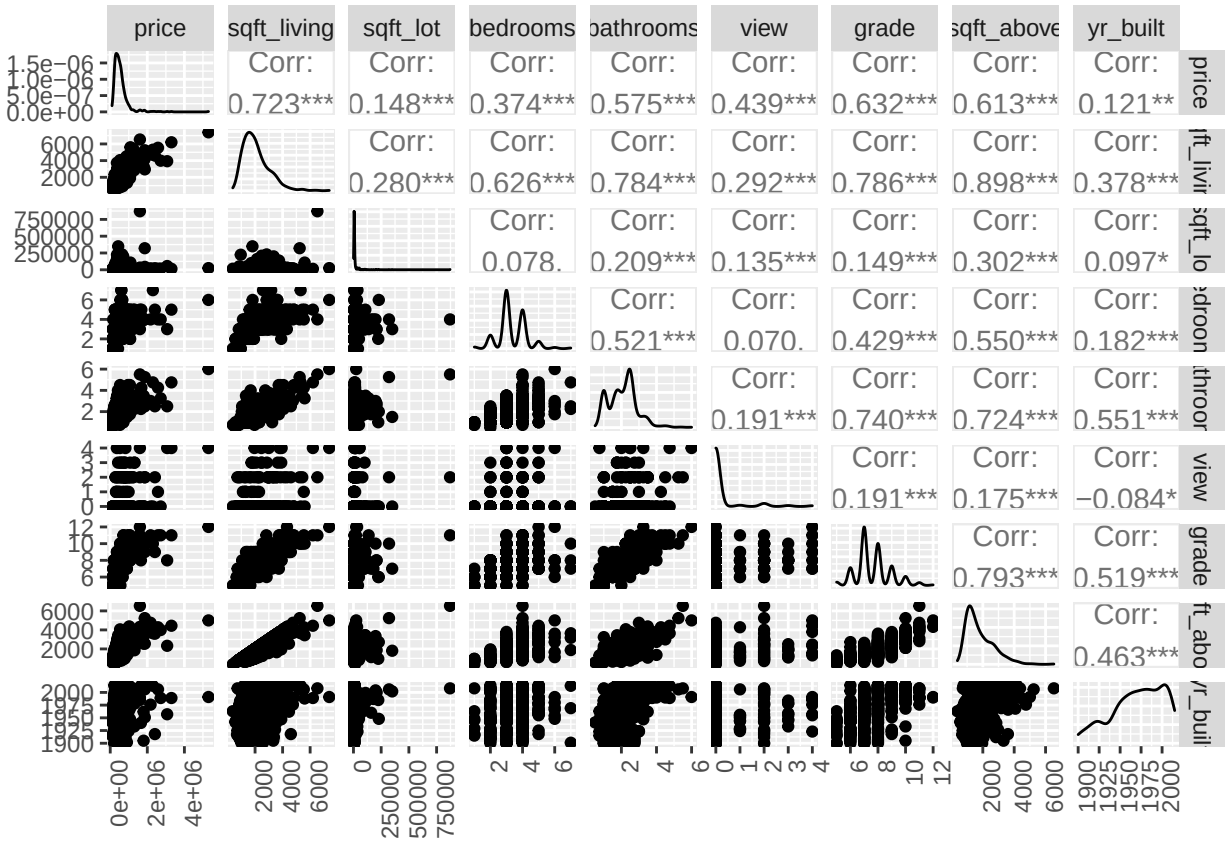
Quantitative Values

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***
0.148***
0.374***
0.575***
0.439***
0.632***
0.613***
0.121**
0.280***
0.626***
0.784***
0.292***
0.786***
0.898***
0.378***
0.078.
0.209***
0.135***
0.149***
0.302***
0.097*
0.521***
0.070.
0.429***
0.550***
0.182***
0.191***
0.740***
0.724***
0.551***
0.191***
0.175***
−0.084*
0.793***
0.519***
0.463***

Discrete and Categorical Values

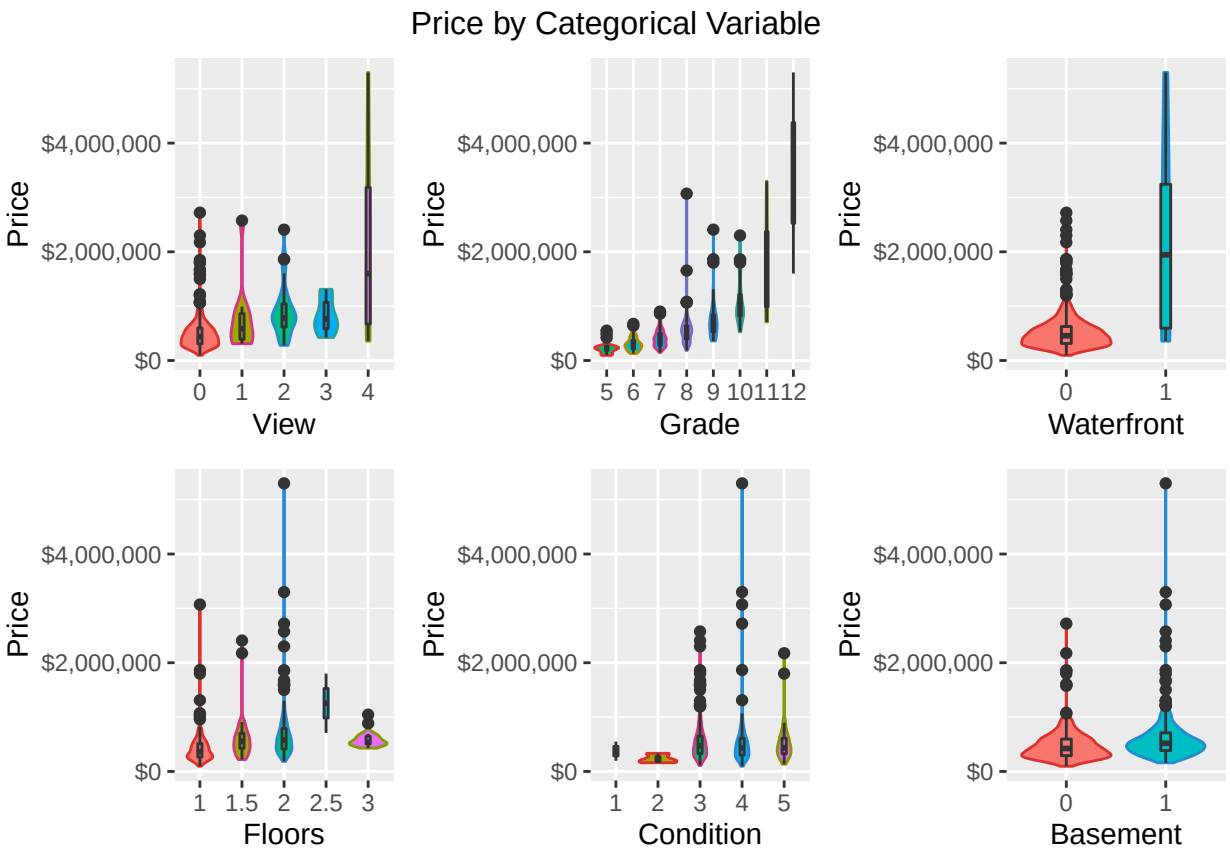
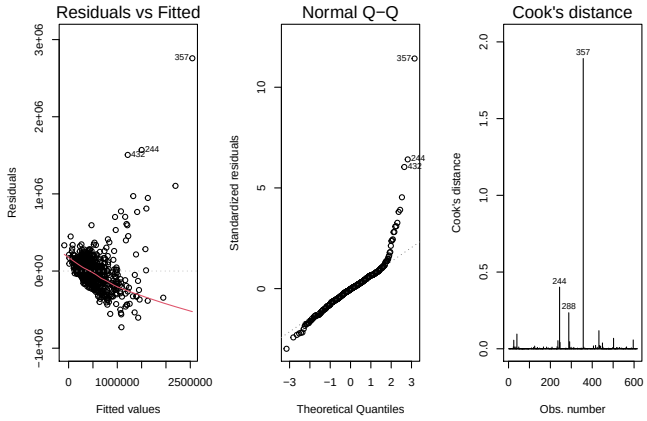


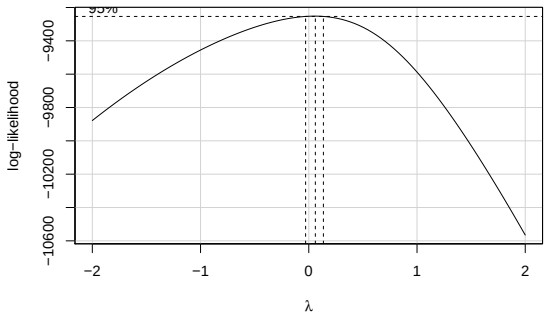
Table 3: Model comparison showing iteratively adding predictors with strong supporting evidence

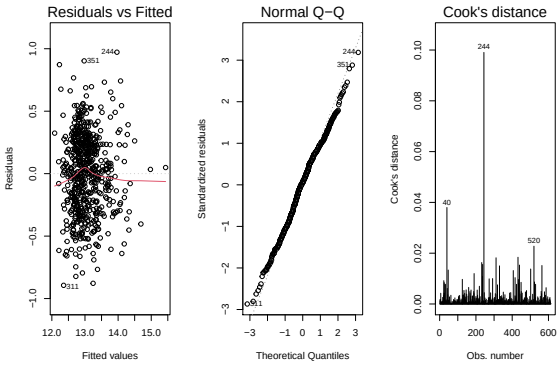
	Dependent variable:			
	price			
	(1)	(2)	(3)	(4)
sqft_living	306.516***	251.551***	210.834***	210.259***
	(11.846)	(18.980)	(18.237)	(17.565)
grade		55,364.790***	64,247.590***	103,280.400***
		(15,032.910)	(14,071.130)	(14,665.420)
view			151,294.300***	128,209.200***
			(15,919.200)	(15,686.520)
yr_built				−2,834.238***
				(406.927)
Constant	−90,183.990***	−398,893.000***	−413,126.400***	4,881,654.000***
	(27,085.020)	(88,005.520)	(82,206.790)	(764,312.600)
Observations	613	613	613	613
R^2	0.523	0.533	0.594	0.624
Adjusted R^2	0.522	0.532	0.591	0.621
Residual Std. Error	282,443.100 (df = 611)	279,583.300 (df = 610)	261,118.100 (df = 609)	251,492.200 (df = 608)
F Statistic	669.475*** (df = 1; 611)	348.403*** (df = 2; 610)	296.388*** (df = 3; 609)	251.761*** (df = 4; 608)
Note:				*p<0.1; **p<0.05; ***p<0.01

Analysis

Initial Model







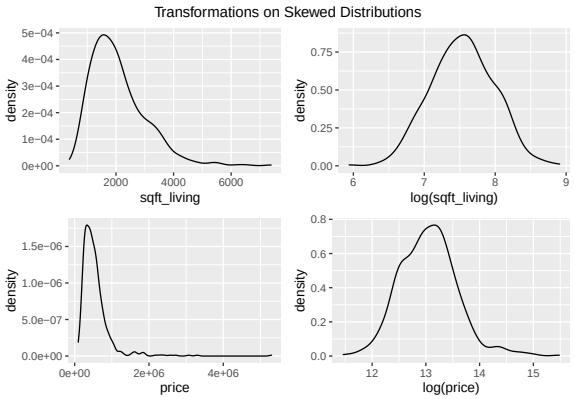


Table 4: VIF Values for Price Model

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.2 <- lm(formula = price ~ sqft_living + grade,
            data = df)
lm.price.3 <- lm(formula = price ~ sqft_living + grade + view,
            data = df)
lm.price.final <- lm(formula = price ~ sqft_living + grade + view + yr_built,
            data = df)
stargazer(lm.price.1,
          lm.price.2,
          lm.price.3,
          lm.price.final,
          header=FALSE,
          align=T,
          label="ModelComparison",
          title="Model comparison showing iteratively adding predictors with strong supporting evidence
```

```r
# 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)
```