

R Notebook

Nelson Brown and Bryce Smith

Introduction

Summary Statistics and Graphics

Quantitative Values

Table 1: First Four Rows for Quantitative Values on Seattle 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

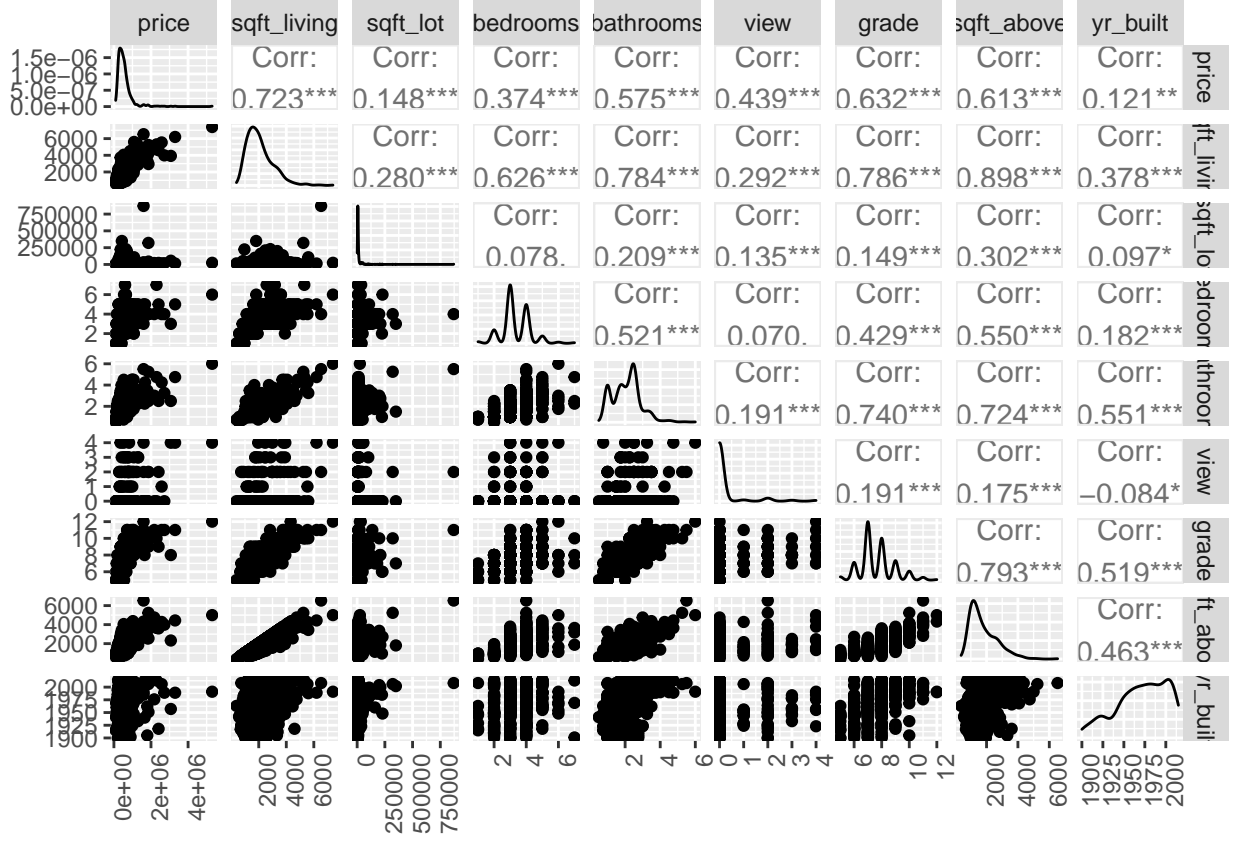
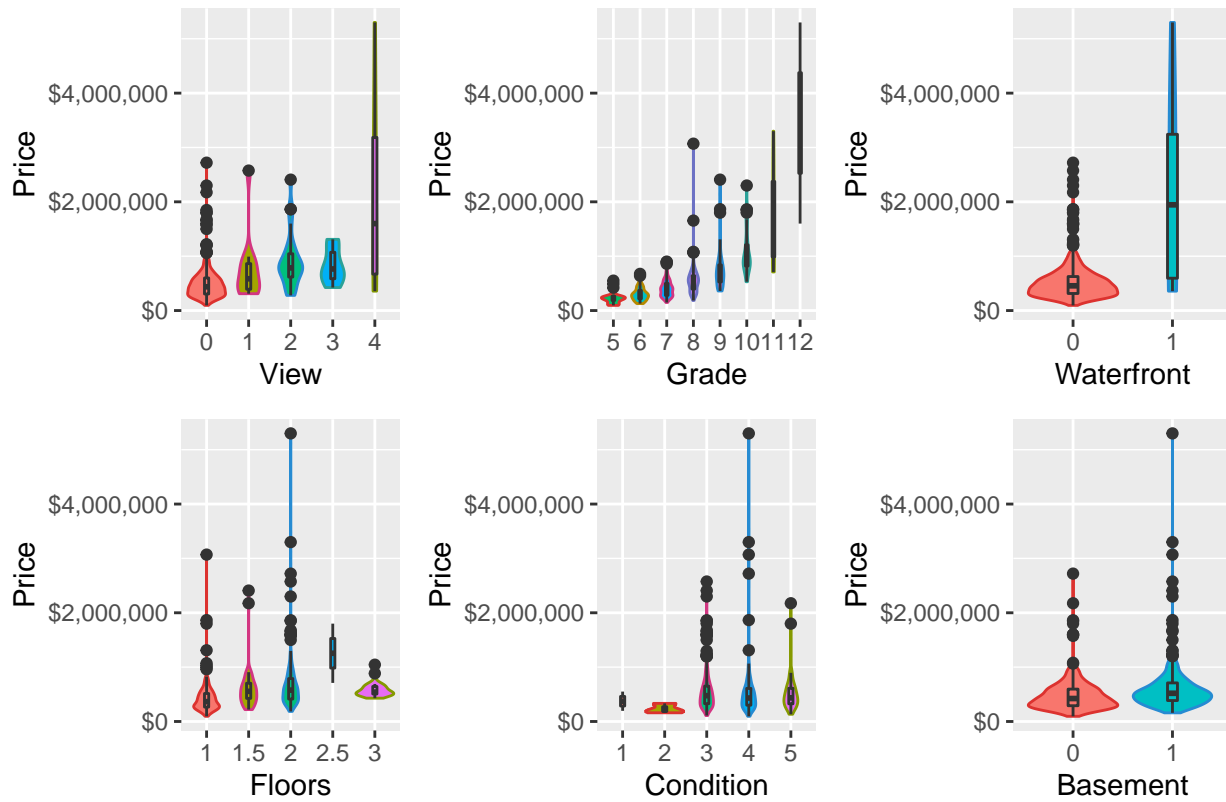


Table 3: Correlation of Quantitative Values

	price	sqft_living	sqft_lot	bedrooms	bathrooms	view	grade	sqft_above	yr_built
price	1	0.723	0.148	0.374	0.575	0.439	0.632	0.613	0.121
sqft_living	0.723	1	0.280	0.626	0.784	0.292	0.786	0.898	0.378
sqft_lot	0.148	0.280	1	0.078	0.209	0.135	0.149	0.302	0.097
bedrooms	0.374	0.626	0.078	1	0.521	0.070	0.429	0.550	0.182
bathrooms	0.575	0.784	0.209	0.521	1	0.191	0.740	0.724	0.551
view	0.439	0.292	0.135	0.070	0.191	1	0.191	0.175	-0.084
grade	0.632	0.786	0.149	0.429	0.740	0.191	1	0.793	0.519
sqft_above	0.613	0.898	0.302	0.550	0.724	0.175	0.793	1	0.463
yr_built	0.121	0.378	0.097	0.182	0.551	-0.084	0.519	0.463	1

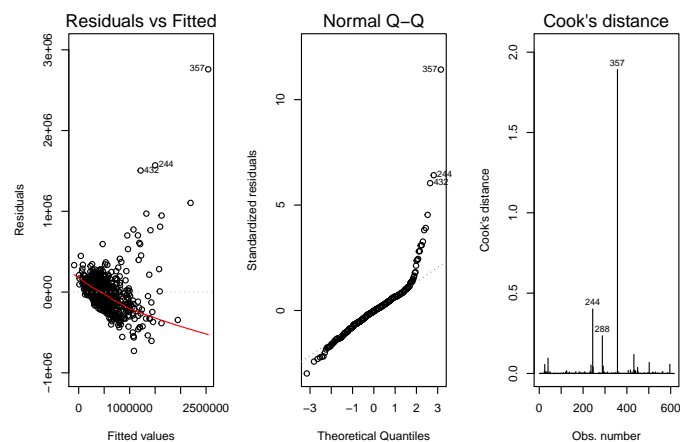
Discrete and Categorical Values

Price by Categorical Variable



Analysis

Initial Model



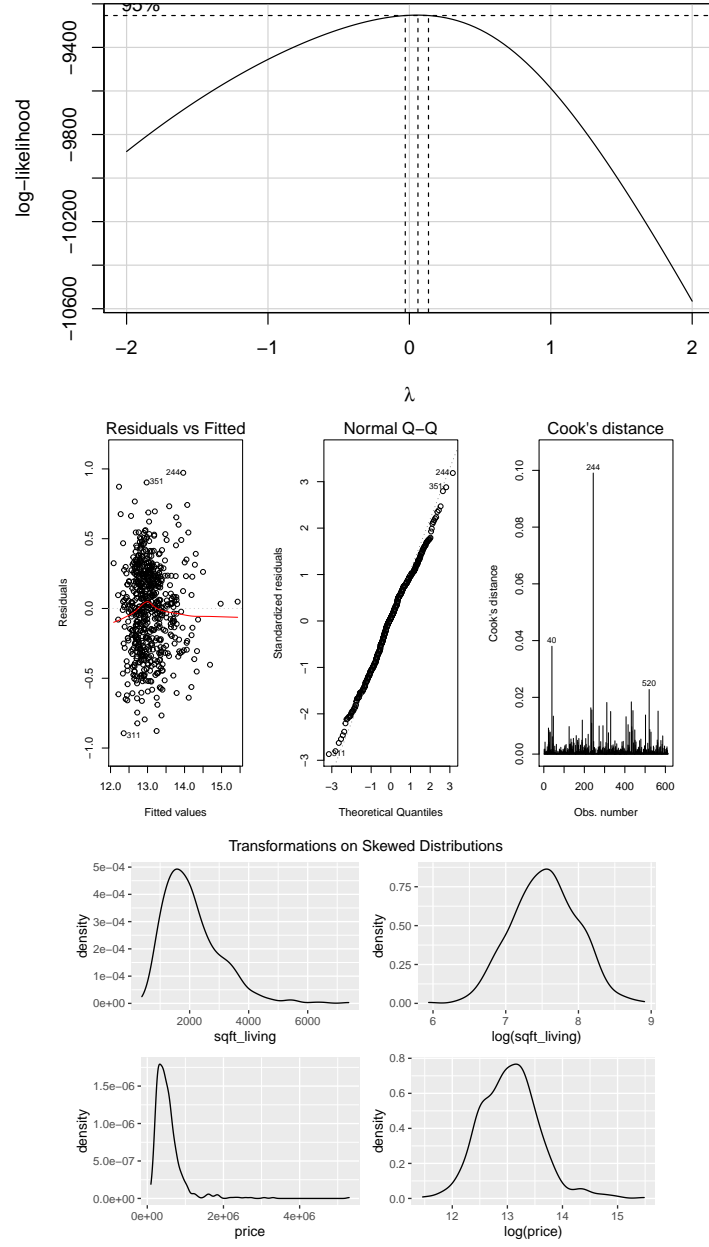


Table 4: VIF Values for Price Model

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
  )
))
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 Seattle Housing Dataframe")

# 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))

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

# 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")

# 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)

```