R Notebook

Nelson Brown and Bryce Smith

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE)
```

Introduction

Summary Statistics and Graphics

Quantitative Values

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

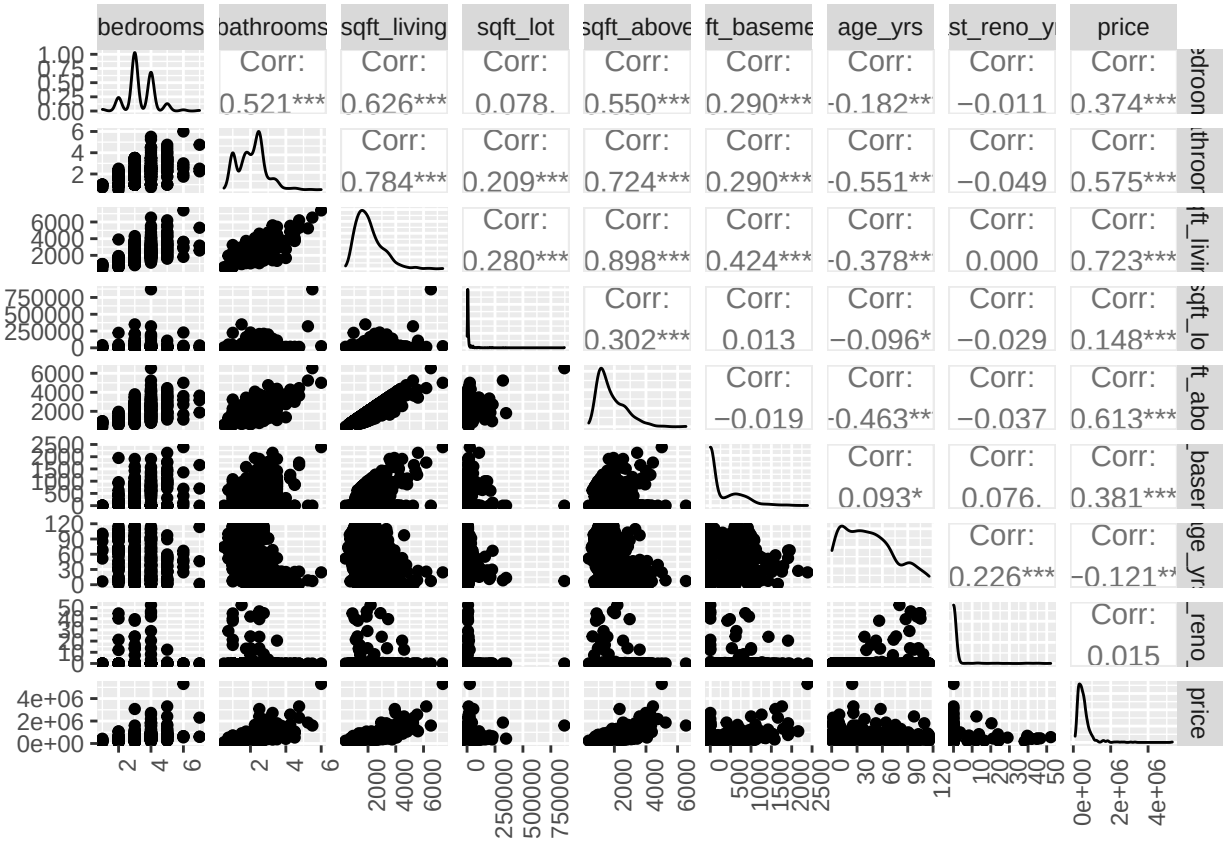
bedrooms	bathrooms	sqft_living	sqft_lot	sqft_above
3	1.750	1,570	6,975	1,040
5	3.750	3,050	8,972	3,050
3	1.750	1,570	12,506	1,570
4	1.750	1,390	10,660	1,030

Table 2: First Four Rows for Quantitative Values on Seattlle Housing Dataframe

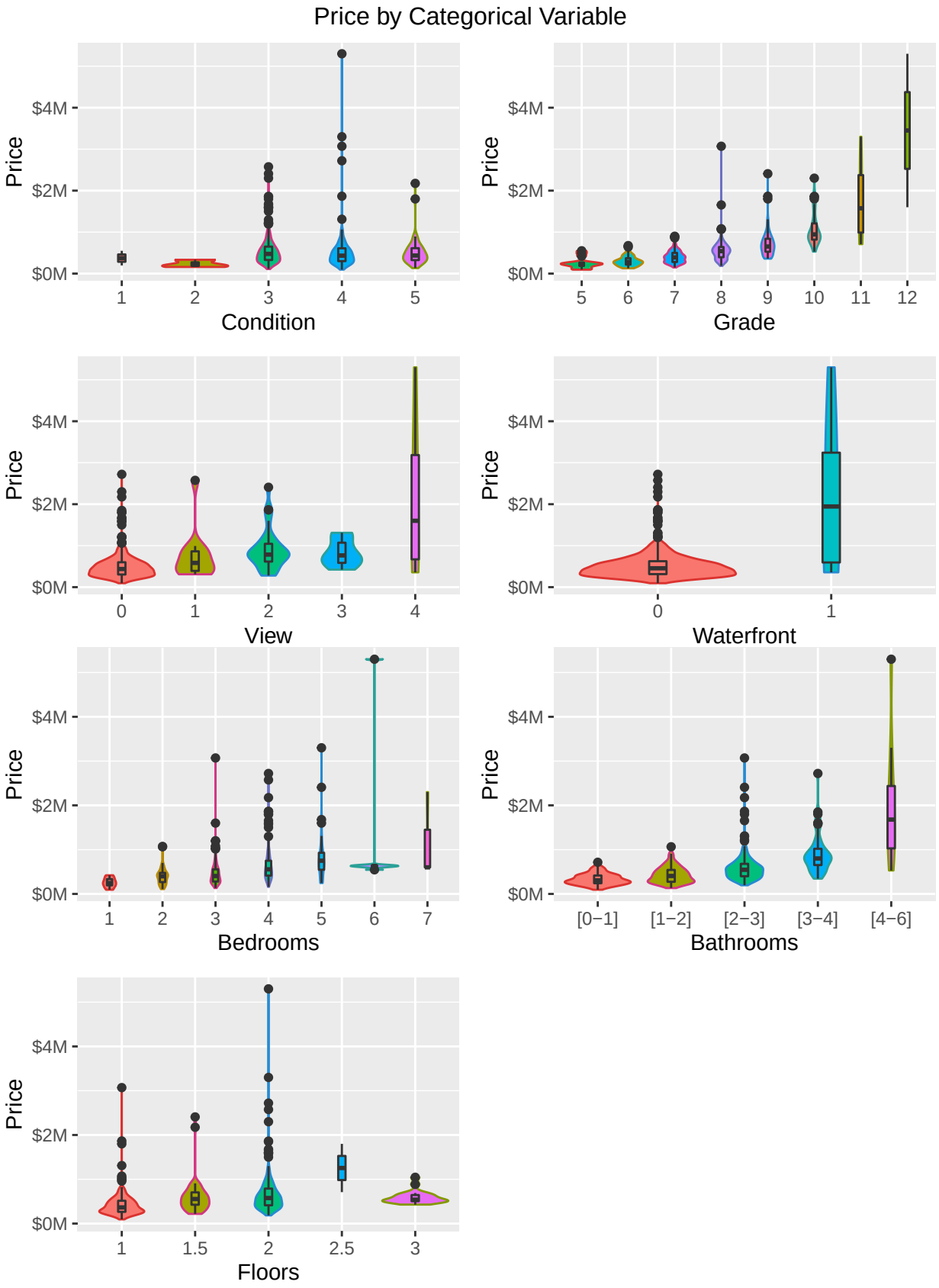
sqft_basement	age_yrs	last_reno_yrs	price
530	35.712	0	359,950
0	1.288	0	909,950
0	56.118	0	318,000
360	54.751	0	272,000

Table 3: 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
bedrooms	3.352	0.876	1	3	4	7
bathrooms	2.092	0.805	0.500	1.500	2.500	6.000
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
floors	1.479	0.535	1	1	2	3
waterfront	0.010	0.099	0	0	0	1
view	0.204	0.695	0	0	0	4
condition	3.388	0.641	1	3	4	5
grade	7.635	1.217	5	7	8	12
sqft_above	1,793.571	873.153	380	1,130	2,313	6,530
sqft_basement	280.098	424.835	0	0	570	2,390
yr_built	1,971.210	29.939	1,900	1,951	1,998	2,015
yr_renovated	84.527	401.973	0	0	0	2,014
age_yrs	43.669	29.943	0.274	16.830	64.219	115.381
last_reno_yrs	0.932	5.554	0	0	0	52



Discrete and Categorical Values



Analysis

Initial Model

Table 4: Initial Data Model

	Dependent variable:
	price
sqft_living	306.516***
	t = 25.874
Constant	−90,183.990***
	t = −3.330
Observations	613
R^2	0.523
Adjusted R^2	0.522
Residual Std. Error	282,443.100 (df = 611)
F Statistic	669.475*** (df = 1; 611)
Note:	*p<0.1; **p<0.05; ***p<0.01

Results and Conclusions

Appendix: All Code for This Report

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE)
library(dplyr)
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
library(gsubfn)

replace_numbers = function(x, cutoff=4, digits=3, scipen=-7) {
  ifelse(nchar(x) < cutoff, x, prettyNum(as.numeric(x), digits=digits, scientific=scipen))
}
sci_notation <- function(star) {
  cat(gsubfn("([0-9\\.]+)", ~replace_numbers(x), star), sep='\n')
}

# 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$date) - 7)) # Convert string to date object
df$date_built <- as.Date(ISOdate(df$yr_built,1,1))
df$age_yrs <- as.double(df$date - df$date_built)/365.
df$last_reno_yrs <- with(df,
     ifelse(yr_renovated == 0,
            0,
            as.double(date-as.Date(ISOdate(df$yr_renovated,1,1)))/365
            )
     )

# Quantitative Values Section
quant.columns <- c(4:7,13:14, 18:19, 3)

# Print head of initial dataframe
stargazer(df[1:4,quant.columns[1:5]],
          rownames=FALSE,
          summary=FALSE,
          header=FALSE,
          title="First Four Rows for Quantitative Values on Seattlle Housing Dataframe")

# Print head of initial dataframe
stargazer(df[1:4,quant.columns[6:length(quant.columns)]],
          rownames=FALSE,
          summary=FALSE,
          header=FALSE,
          title="First Four Rows for Quantitative Values on Seattlle Housing Dataframe")

# Summarize initial dataframe
stargazer(df[,-c(1)],
```

```
          header=FALSE,
          omit.summary.stat=c('N'),
          title="Summary Statistics for Values on Seattle Housing Dataframe")

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

# Discrete and Categorical Values Section
df$condition <- as.factor(df$condition)
df$grade <- as.factor(df$grade)
df$waterfront <- as.factor(df$waterfront)
df$view <- as.factor(df$view)
df$bedrooms <- as.factor(df$bedrooms)
df$floors <- as.factor(df$floors)
cat.discrete.price.columns <- c(3,9,10,11)

tags <- c("[0-1]", "[1-2]","[2-3]", "[3-4]", "[4-6]")
bgroup <- as_tibble(select(df, price, bathrooms)) %>%
  mutate(tag = case_when(
  bathrooms == 0.25|bathrooms == 0.5|
    bathrooms == 0.75|bathrooms == 1.00 ~ tags[1],
  bathrooms == 1.25|bathrooms == 1.5|
    bathrooms == 1.75|bathrooms == 2.00 ~ tags[2],
  bathrooms == 2.25|bathrooms == 2.5|
    bathrooms == 2.75|bathrooms == 3.00 ~ tags[3],
  bathrooms == 3.25|bathrooms == 3.5|
    bathrooms == 3.75|bathrooms == 4.00 ~ tags[4],
  bathrooms > 4.00 & bathrooms <= 6.00 ~ tags[5],
))
df$bath_group <- bgroup$tag

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")
}

p <- plot_price_by_cat(df, df$condition, "Condition")
q <- plot_price_by_cat(df, df$grade, "Grade")
r <- plot_price_by_cat(df, df$view, "View")
s <- plot_price_by_cat(df, df$waterfront, "Waterfront")
t <- plot_price_by_cat(df, df$bedrooms, "Bedrooms")
u <- plot_price_by_cat(df, df$bath_group, "Bathrooms")
v <- plot_price_by_cat(df, df$floors, "Floors")

grid.arrange(grobs=list(p, q,
                        r, s),
```

```
             ncol=2,
             top="Price by Categorical Variable")

grid.arrange(grobs=list(t, u,
                        v),
             ncol=2,
             top=NULL)

# Initial Model
lm.initial <- lm(price ~ sqft_living, data=df)
stargazer(lm.initial,
          header=FALSE,
          title="Initial Data Model",
          report='vc*t')
```