Avocado Project

Evelyn Chen, Christen Parzych

10/12/2019

Research Topic: 'Guac'-Bottom: Exploring Avocados' Fluctuating Prices

Introduction

The avocado is America's 'it' fruit, so much so that it has become a running joke —headlines tell us that millennials choose avocados over, say, homeownership (Cummings, 2019). In spite of the hyperbole, the avocado has become an obsession Americans love to love. The evidence backs this claim: within the week of October 6, 2019, alone, U.S. consumers purchased 48,778,842 pounds of avocado—that's a lot of avocado toast (Hass Avocado Board, 2019)! Mexico leads the world in avocado production, followed by the United States. California produces 90% of American avocados, with Florida and Hawaii rounding out the other 10% (Dekevich, 2018).

This exploratory report utilizes data from the Hass Avocado Board website, published in May 2018 and compiled into a CSV. The Hass Avocado Board is an agricultural advocacy group founded to promote the consumption of avocados within the United States. The dataset includes weekly retail scan data for national retail (grocery, mass, club, drug, dollar, and military) volume and price, including 18,249 observations of 13 variables from 2015 - 2018.

We seek to explore a central question: are the price fluctuations in U.S. avocado-producing regions steadier/smaller than in non-avocado-producing regions? We are also interested in investigating whether the average price of avocados is lower in avocado-producing regions. Finally, to what extent does domestic origin or avocado type predict the average price of avocados?

```
#load dataset
avocado<-read_csv(file="avocado.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_double(),
##   Date = col_date(format = ""),
##   AveragePrice = col_double(),
##   `Total Volume` = col_double(),
##   `4046` = col_double(),
##   `4225` = col_double(),
##   `4770` = col_double(),
##   `Total Bags` = col_double(),
##   `Small Bags` = col_double(),
##   `Large Bags` = col_double(),
##   `XLarge Bags` = col_double(),
##   type = col_character(),
##   year = col_double(),
##   region = col_character()
## )
```

Data

We elected to use the entire dataset for this project, rather than a sample. The data represents weekly retail scan data for national retail volume and price from 2015 - 2018. This retail scan data comes directly from retailers' cash registers, reflecting actual retail sales of Hass avocados. The dataset has 18,249 observations of 13 variables, descriptions of which are located in the table below:

```
variable_table <- matrix(c("Date","Date of observation",
              "AveragePrice","Average price of a single avocado",
              "Total Volume","Total volume of avocados sold in pou
nds",
              "4046","Total number of small/medium PLU 4046 avocad
os sold",
              "4225","Total number of large PLU 4225 avocados sol
d",
              "4770","Total number of extra-large PLU 4770 avocado
s sold",
              "Total bags","Total number of bags sold",
              "Small bags","Total number of small bags sold",
              "Large bags","Total number of large bags sold",
              "Xlarge bags","Total number of extra-large bags sol
d",
              "Type","Conventional or organic",
              "Year","Year sold",
              "Region","City, state, or region sold"), ncol=2, byr
ow=TRUE)

colnames(variable_table) <- c("Variable Name","Description")
rownames(variable_table) <- c(
              1,
              2,
              3,
              4,
              5,
              6,
              7,
              8,
              9,
              10,
              11,
              12,
              13)

variable_table <- as.table(variable_table)
```

```
kable(variable_table)
```

Variable Name	Description
Date	Date of observation
AveragePrice	Average price of a single avocado
Total Volume	Total volume of avocados sold in pounds
4046	Total number of small/medium PLU 4046 avocados sold
4225	Total number of large PLU 4225 avocados sold
4770	Total number of extra-large PLU 4770 avocados sold
Total bags	Total number of bags sold
Small bags	Total number of small bags sold
Large bags	Total number of large bags sold
Xlarge bags	Total number of extra-large bags sold
Type	Conventional or organic
Year	Year sold
Region	City, state, or region sold

The majority of the variables in our dataset are continuous: “AveragePrice,” “Total Volume,” “4046,” “4225,” “4770,” “Total bags,” “Small bags,” “Large bags,” and “Xlarge bags.” The variables “Type,” “Year,” and “Region” are categorical.

Because we are exploring price fluctuations in various regions of the country, our analysis focuses on the following variables: average price, avocado size, avocado type, purchase year, and purchase region. Based on the available data, we first identified a normal range for avocado prices according to region and then determined when avocado prices fell outside of this range from 2015 to 2018. The Hass Avocado Board provides rich and accessible data, so we also integrated more in-depth, region-specific analyses into our report as well.

Tidying Data

Our data required minimal tidying; however, there were a few opportunities to clean up the data. First we renamed most of the columns in a format that would work with the ‘ggplot’ package later on in the project. This involved removing character spaces and replacing with underscores. We then created a new variable, called ‘Volume_Rank,’ to establish a percentile rank of total volume.

```
##renaming column so that it works in ggplot
colnames(avocado)[colnames(avocado)=="AveragePrice"] <- "Average_Price"
colnames(avocado)[colnames(avocado)=="Total Volume"] <- "Total_Volume"
colnames(avocado)[colnames(avocado)=="Total Bags"] <- "Total_Bags"
colnames(avocado)[colnames(avocado)=="Small Bags"] <- "Small_Bags"
colnames(avocado)[colnames(avocado)=="Large Bags"] <- "Large_Bags"
colnames(avocado)[colnames(avocado)=="XLarge Bags"] <- "XLarge_Bags"
colnames(avocado)[colnames(avocado)=="type"] <- "Type"
colnames(avocado)[colnames(avocado)=="year"] <- "Year"
colnames(avocado)[colnames(avocado)=="region"] <- "Region"

#create new variable, 'Volume_Rank': percentile rank of 'Total_Volume'
avocado<-avocado%>%mutate(Volume_Rank=percent_rank(Total_Volume))
```

The data is largely consistent, with the exception of ‘region,’ which contains a mix of cities, states, and regions of the country. To maintain consistency throughout our analysis, we elected to recode this variable into regions. We used the Hass Avocado Board’s existing regional classifications for the continental U.S. as our consistent characterization:

```
region_table <- matrix(c(
              "California","California",
              "West","Washington, Oregon, Idaho, Nevada, New Mexic
o, Montana, Colorado, Utah, Arizona, Wyoming",
              "South Central","Texas, Oklahoma, Arkansas, Louisian
a",
              "Plains","North Dakota, South Dakota, Nebraska, Minn
esota, Iowa, Kansas, Missouri",
              "Great Lakes","Wisconsin, Indiana, Illinois, Ohio, M
ichigan",
              "Southeast","Alabama, Mississippi, Georgia, South Ca
rolina, Florida",
              "Mid-South","Tennessee, Kentucky, North Carolina, Vi
rginia, West Virginia, Maryland",
              "Northeast","Maine, Vermont, New Hampshire, Massachu
setts, Rhode Island, Connecticut, New York, Pennsylvania, New Jers
ey, Delaware"), ncol=2, byrow=TRUE)

colnames(region_table) <- c("Region Name","State(s)")
rownames(region_table) <- c(
              1,
              2,
              3,
              4,
              5,
              6,
              7,
              8)

region_table <- as.table(region_table)

kable(region_table)
```

Region Name	State(s)
California	California

Region Name	State(s)
West	Washington, Oregon, Idaho, Nevada, New Mexico, Montana, Colorado, Utah, Arizona, Wyoming
South Central	Texas, Oklahoma, Arkansas, Louisiana
Plains	North Dakota, South Dakota, Nebraska, Minnesota, Iowa, Kansas, Missouri
Great Lakes	Wisconsin, Indiana, Illinois, Ohio, Michigan
Southeast	Alabama, Mississippi, Georgia, South Carolina, Florida
Mid-South	Tennessee, Kentucky, North Carolina, Virginia, West Virginia, Maryland
Northeast	Maine, Vermont, New Hampshire, Massachusetts, Rhode Island, Connecticut, New York, Pennsylvania, New Jersey, Delaware

We then created a new column ("US_region") and recategorized the data to align with published categorizations. This recoding resulted in eight observational labels for the region variable: California, West, South Central, Plains, Great Lakes, Southeast, Mid-South, and Northeast.

```
avocado$US_region<-NA

avocado$US_region[avocado$Region=="Albany"]<-"Northeast"
avocado$US_region[avocado$Region=="Atlanta"] <- "Southeast"
avocado$US_region[avocado$Region=="BaltimoreWashington"] <- "Mid-South"
avocado$US_region[avocado$Region=="Boise"] <- "West"
avocado$US_region[avocado$Region=="Boston"] <- "Northeast"
avocado$US_region[avocado$Region=="BuffaloRochester"] <- "Northeast"
avocado$US_region[avocado$Region=="California"] <- "California"
avocado$US_region[avocado$Region=="Charlotte"] <- "Mid-South"
avocado$US_region[avocado$Region=="Chicago"] <- "Great Lakes"
avocado$US_region[avocado$Region=="CincinnatiDayton"] <- "Great Lakes"
avocado$US_region[avocado$Region=="Columbus"] <- "Great Lakes"
avocado$US_region[avocado$Region=="DallasFtWorth"] <- "South Central"
avocado$US_region[avocado$Region=="Denver"] <- "West"
avocado$US_region[avocado$Region=="Detroit"] <- "Great Lakes"
avocado$US_region[avocado$Region=="GrandRapids"] <- "Great Lakes"
avocado$US_region[avocado$Region=="GreatLakes"] <- "Great Lakes"
avocado$US_region[avocado$Region=="HarrisburgScranton"] <- "Northeast"
avocado$US_region[avocado$Region=="HartfordSpringfield"] <- "Northeast"
avocado$US_region[avocado$Region=="Houston"] <- "South Central"
avocado$US_region[avocado$Region=="Indianapolis"] <- "Great Lakes"
avocado$US_region[avocado$Region=="Jacksonville"] <- "Southeast"
avocado$US_region[avocado$Region=="LasVegas"] <- "West"
avocado$US_region[avocado$Region=="LosAngeles"] <- "California"
avocado$US_region[avocado$Region=="Louisville"] <- "Mid-South"
avocado$US_region[avocado$Region=="MiamiFtLauderdale"] <- "Southeast"
avocado$US_region[avocado$Region=="Midsouth"] <- "Mid-South"
avocado$US_region[avocado$Region=="Nashville"] <- "Mid-South"
avocado$US_region[avocado$Region=="NewOrleansMobile"] <- "Southeast"
```

```
avocado$US_region[avocado$Region=="NewYork"] <- "Northeast"
avocado$US_region[avocado$Region=="Northeast"] <- "Northeast"
avocado$US_region[avocado$Region=="NorthernNewEngland"] <- "Northe
ast"
avocado$US_region[avocado$Region=="Orlando"] <- "Southeast"
avocado$US_region[avocado$Region=="Philadelphia"] <- "Northeast"
avocado$US_region[avocado$Region=="PhoenixTucson"] <- "West"
avocado$US_region[avocado$Region=="Pittsburgh"] <- "Northeast"
avocado$US_region[avocado$Region=="Plains"] <- "Plains"
avocado$US_region[avocado$Region=="Portland"] <- "West"
avocado$US_region[avocado$Region=="RaleighGreensboro"] <- "Mid-Sou
th"
avocado$US_region[avocado$Region=="RichmondNorfolk"] <- "Mid-Sout
h"
avocado$US_region[avocado$Region=="Roanoke"] <- "Mid-South"
avocado$US_region[avocado$Region=="Sacramento"] <- "California"
avocado$US_region[avocado$Region=="SanDiego"] <- "California"
avocado$US_region[avocado$Region=="SanFrancisco"] <- "California"
avocado$US_region[avocado$Region=="Seattle"] <- "West"
avocado$US_region[avocado$Region=="SouthCarolina"] <- "Mid-South"
avocado$US_region[avocado$Region=="SouthCentral"] <- "South Centra
l"
avocado$US_region[avocado$Region=="Southeast"] <- "Southeast"
avocado$US_region[avocado$Region=="Spokane"] <- "West"
avocado$US_region[avocado$Region=="StLouis"] <- "Plains"
avocado$US_region[avocado$Region=="Syracuse"] <- "Northeast"
avocado$US_region[avocado$Region=="Tampa"] <- "Southeast"
avocado$US_region[avocado$Region=="West"] <- "West"
avocado$US_region[avocado$Region=="WestTexNewMexico"] <- "West"

##footnote: We elected to categorize 'WestTexNewMexico' as 'West,'
rather than as 'South Central.'
```

Observations titled “WestTexNewMexico” did not align with any of the eight published categorizations. Because the description is relatively specific to a geographic area located in the far western United States, we elected to include these observations under “West.” In terms of data limitations, this dataset only includes data related to the continental United States. As a result, we are missing information related to a major U.S. avocado producer: Hawaii. Furthermore, the dataset's sales and volume information does not differentiate between domestic and imported avocados, so we are unable to explore the effect of a region's proximity to an international avocado producer (i.e., South Central and Mexico) on avocado prices. Additionally, 338 rows of observations in the 'avocado' dataset contained sales information from the region categorized as "TotalUS." It is unclear what this region referred to, and we were unable to locate additional information. As a result we decided it was appropriate to exclude this subset of data and not include it in our data analysis.

```
#Deleted rows of data: avocado$region == 'TotalUS'

avocado_tidy<-avocado%>%filter(str_detect(Region, "TotalUS", negat
e = TRUE))
```

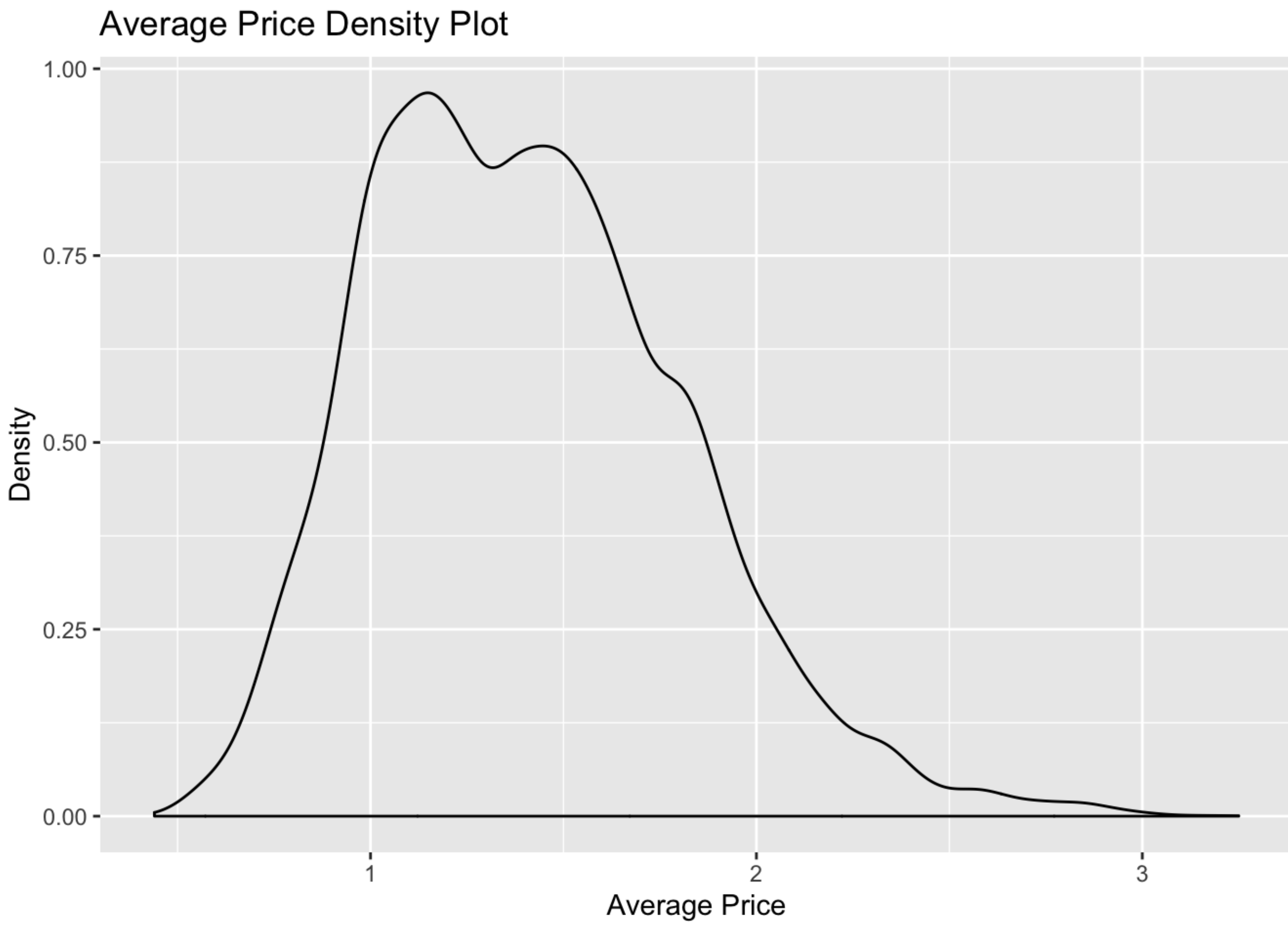
This action deleted 338 total observations from our dataset. We created a new dataset titled 'avocado_tidy' for the remainder of our analysis. The 'avocado_tidy' dataset now contains 17,911 observations of 16 variables.

Exploratory Data Analysis

In this project, we explore price fluctations in avocado sales throughout the United States. The dataset shows weekly retail data of avocado sales from 2015 - 2018.

Our primary outcome of interest is the average price of avocados - "Average_Price." This is a continuous variable. To determine the average price distribution of avocados in this dataset, we used the 'ggplot' command to create a density plot for the "AveragePrice" variable.

```
gg<-ggplot(avocado_tidy,aes(x=Average_Price))
gg<-gg+geom_density()
gg<-gg+xlab("Average Price")
gg<-gg+ylab("Density")
gg<-gg+ggtitle("Average Price Density Plot")
gg
```




```
summary(avocado_tidy$Average_Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.440   1.100   1.370   1.408   1.670   3.250
```

The unconditional mean of “Average_Price” is $1.41, and the average avocado prices range from $0.44 - $3.25. The density plot shows that the distribution of average avocado prices is unimodal and approximately normal. Verifying that the distribution of this outcome variable is approximately normal is important in running subsequent statistical analyses.

Changes in Average Price by Year

We predict that the average price of avocados varies from year to year: “Year” is a likely predictor for average price, as the market trends reflect the effects of supply, demand, and market inflation. In the commands below, we calculated the average prices of avocado for each year. To do this, we generated a dataset called “avocado_year.” Using the ‘group_by’ and ‘summarize’ command, we created a new variable –“avg_price_year”– set equal to the mean average price for each year.

```
avocado_year<-avocado_tidy%>%group_by(Year)%>%summarize(avg_price_
year=mean(Average_Price))
avocado_year
```

Year <dbl>	avg_price_year <dbl>
2015	1.377821
2016	1.340056
2017	1.516610
2018	1.348294

4 rows

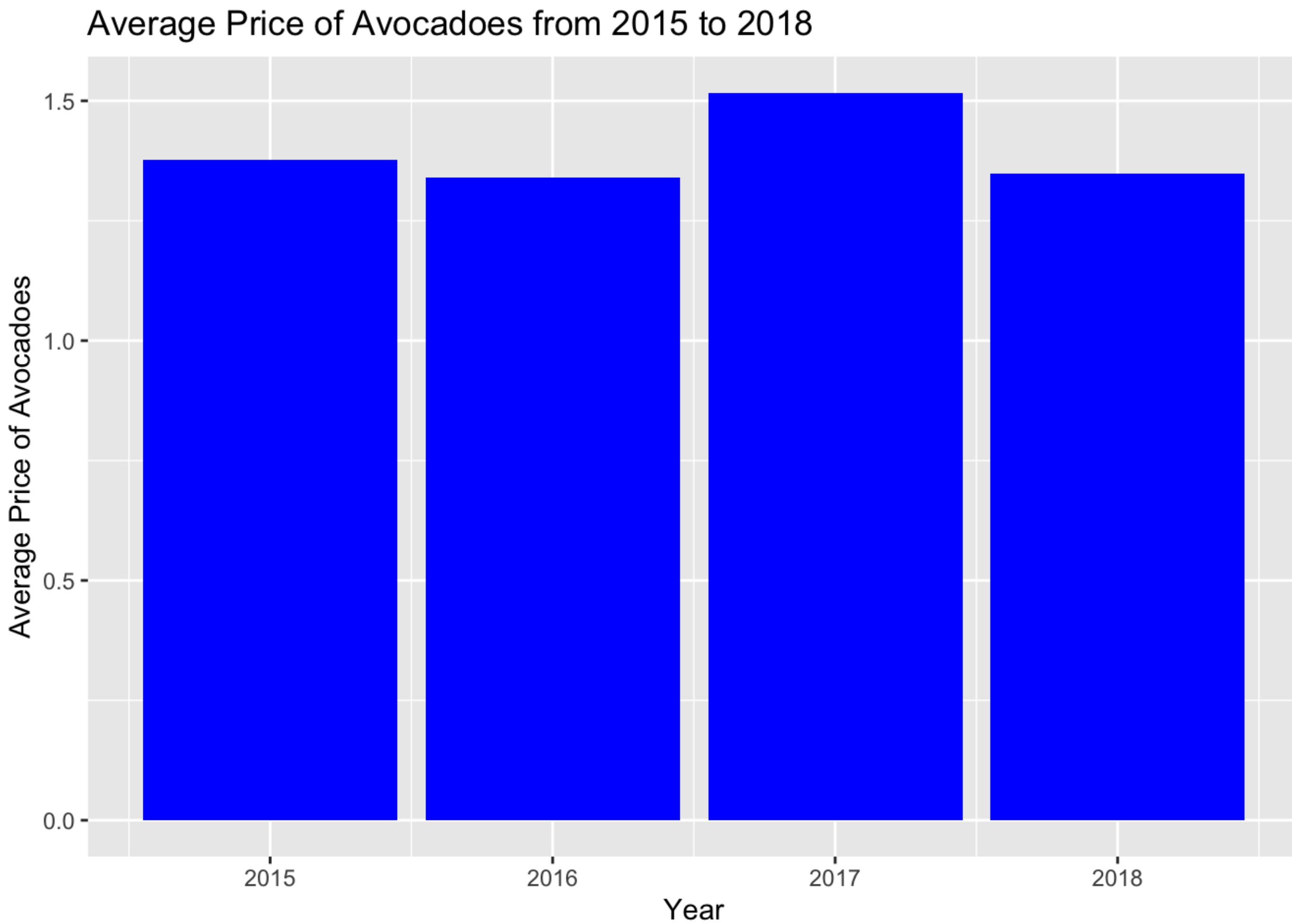
```
avocado_year_table <- avocado_year
colnames(avocado_year_table)<-c("Year","Average Price")
kable(avocado_year_table, align=rep('l', length(avocado_year_table
[,1])))
```

Year	Average Price

Year	Average Price
2015	1.377821
2016	1.340056
2017	1.516610
2018	1.348294

The average price of avocadoes in 2015 was $1.38. The average price in 2017 increased to $1.52 before dropping back down to $1.35 in 2018. We created a bar graph to present this information:

```
## Bar Plot with aesthetics: average price as height, year as cate
gory
gg<-ggplot(avocado_year,aes(x=Year,y=avg_price_year))
gg<-gg+geom_bar(stat="Identity",fill="blue")
gg<-gg+ylab("Average Price of Avocadoes")
gg<-gg+ggtitle("Average Price of Avocadoes from 2015 to 2018")
gg
```



Changes in Average Volume of Avocadoes Sold by Year

We speculate that the average price of avocadoes in 2017 was particularly high because the volume of avocados sold was particularly low. The total volume of avocados sold ("Total_Volume") is a continuous variable. The volume of avocados sold ranges from 85 pounds to 11,274,749 pounds. The mean total volume is 539,259 pounds. To explore this further, we used the 'group_by' and 'summarize' commands to determine the average volume of avocadoes sold in 2015, 2016, 2017, and 2018, respectively. We then created a bar plot from this data to help visualize the observed patterns.

```
summary(avocado_tidy$Total_Volume)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##       85    10571   100154   539259   400177 11274749
```

```
avocado_vol<-avocado_tidy%>%group_by(Year)%>%summarize(avg_vol=mea
n(Total_Volume))
avocado_vol
```

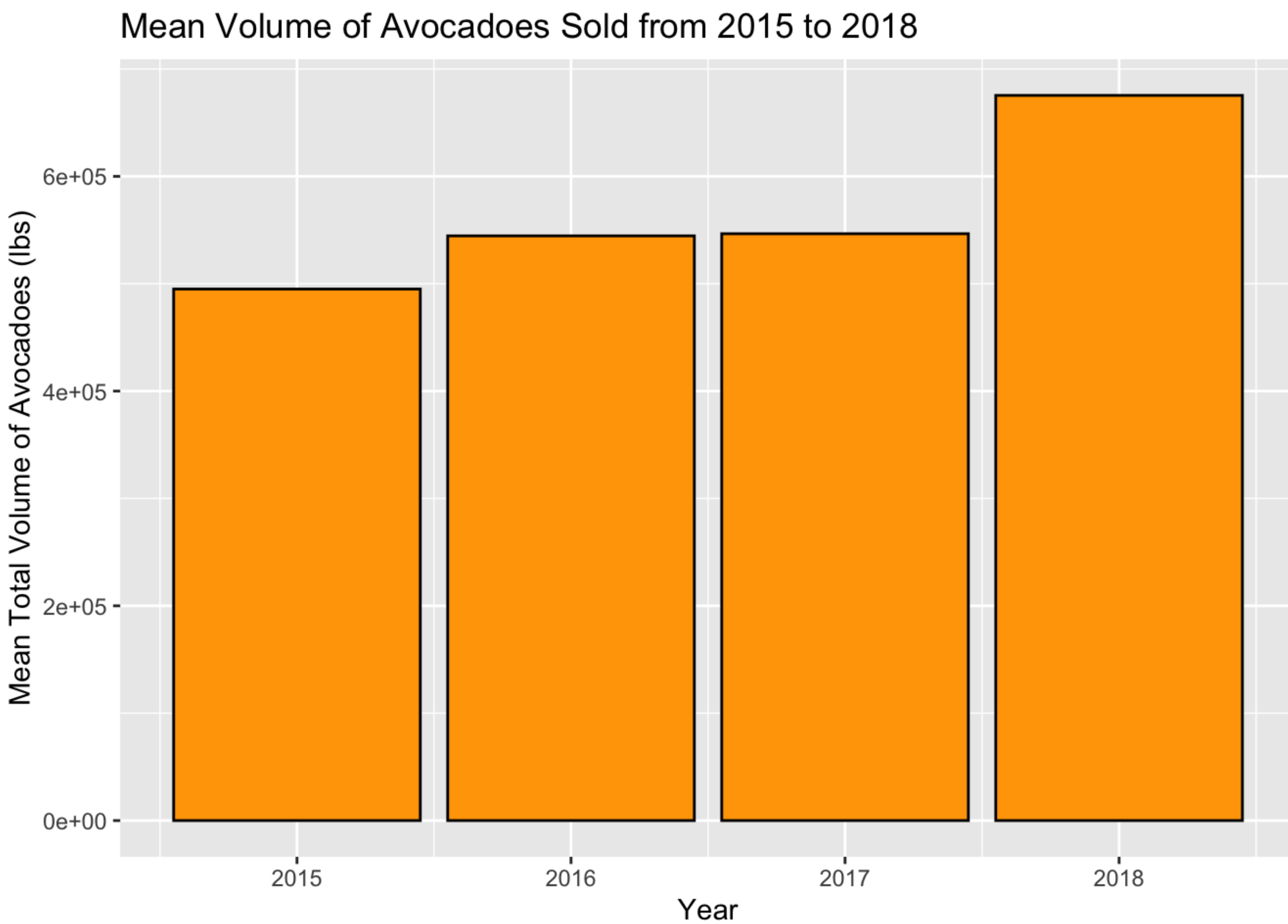
Year <dbl>	avg_vol <dbl>
2015	495048.7
2016	544581.1
2017	546583.4
2018	675397.9

4 rows

```
avocado_vol_table <- avocado_vol
colnames(avocado_vol_table)<-c("Year","Average Total Volume")
kable(avocado_vol_table, align=rep('l', length(avocado_vol_table[,
1])))
```

Year	Average Total Volume
2015	495048.7
2016	544581.1
2017	546583.4
2018	675397.9

```
## Bar Plot with aesthetics: mean total volume of avocadoes sold i
s the outcome variable, sorted by year
gg<-ggplot(avocado_vol,aes(x=Year,y=avg_vol))
gg<-gg+geom_bar(stat="Identity",colour="black",fill="orange")
gg<-gg+ylab("Mean Total Volume of Avocadoes (lbs)")
gg<-gg+ggtitle("Mean Volume of Avocadoes Sold from 2015 to 2018")
gg
```



Avocado sales increased fairly steadily from 2015 to 2018, with 2018 representing the highest volume sold, at 675,397.9 pounds. However, based on the trends observed in the previous two bar plots, there does not appear to be a strong association between the volume of avocadoes sold and the average price.

Two Predictors: Summarizing Average Price by Avocado Type and Year

Perhaps the average price of avocadoes varied based on whether or not the avocadoes sold were grown by conventional or organic methods of farming. Organic farming requires highly regulated farming methods and compliance mechanisms related to pest control and fertilization and increases production cost for farmers (Cernansky, 2018).

```
## Summarize average price by type and year
avocadoes_typeyear<-avocado_tidy%>%
  group_by(Year,Type)%>%
  summarize(avg_typeyear=mean(Average_Price))

avocadoes_typeyear
```

Year <dbl>	Type <chr>	avg_typeyear <dbl>
2015	conventional	1.079198
2015	organic	1.676552
2016	conventional	1.106705
2016	organic	1.573407
2017	conventional	1.296269
2017	organic	1.737107
2018	conventional	1.129167
2018	organic	1.567421

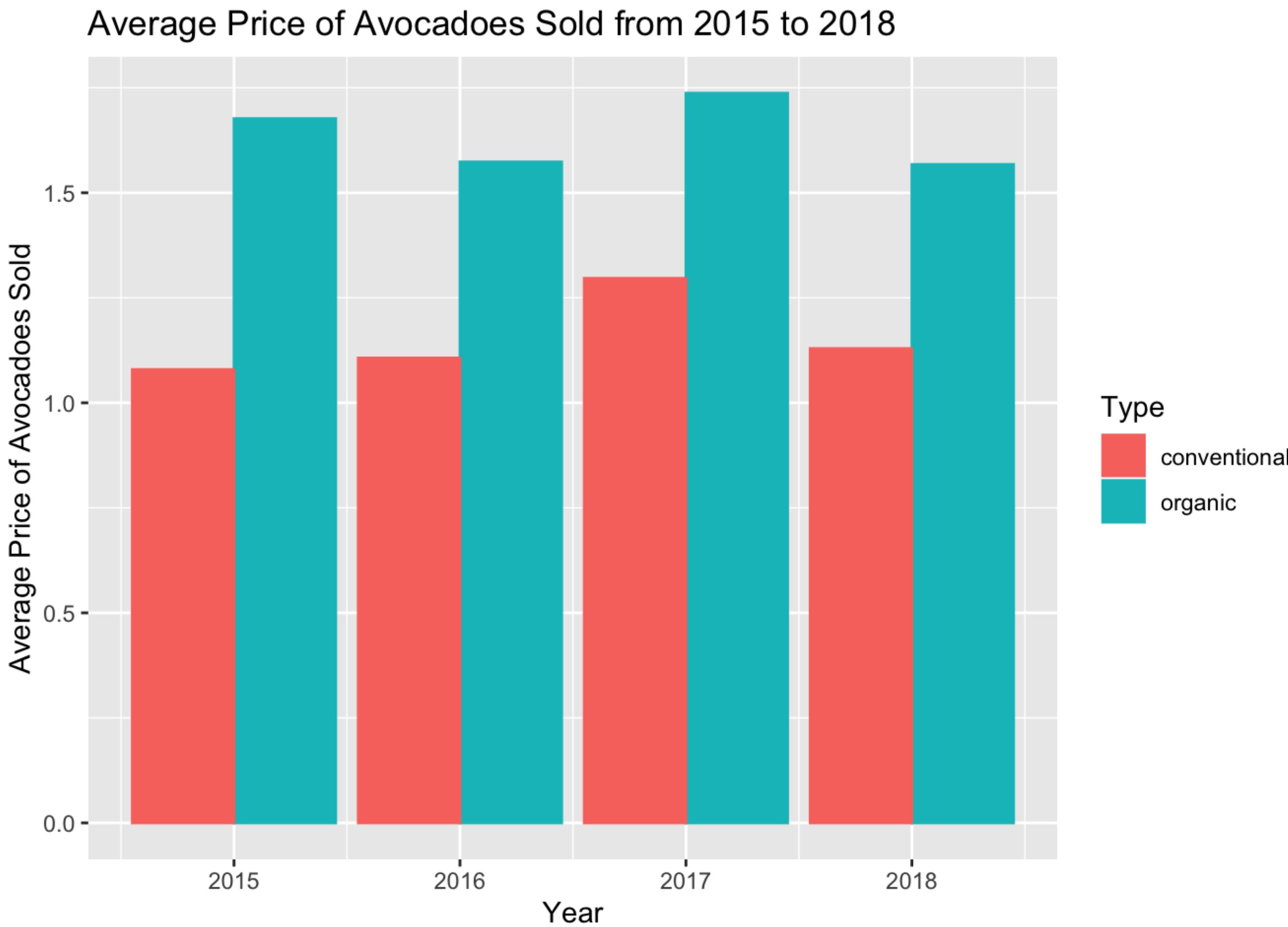
8 rows

```
avocadoes_typeyear_table <- avocadoes_typeyear

colnames(avocadoes_typeyear_table)<-c("Year","Type","Average Pric
e")
kable(avocadoes_typeyear_table, align=rep('l', length(avocadoes_ty
peyear_table[,1])))
```

Year	Type	Average Price
2015	conventional	1.079198
2015	organic	1.676552
2016	conventional	1.106705
2016	organic	1.573407
2017	conventional	1.296269
2017	organic	1.737107
2018	conventional	1.129167
2018	organic	1.567421

```
##bar plot
gg<-ggplot(avocadoes_typeyear,aes(x=Year,y=avg_typeyear,color=Typ
e))
gg<-gg+geom_bar(stat="identity",aes(fill=Type),position="dodge")
gg<-gg+ylab("Average Price of Avocadoes Sold")+xlab("Year")
gg<-gg+ggtitle("Average Price of Avocadoes Sold from 2015 to 2018"
)
gg
```



As predicted, organic avocadoes have a consistently higher average price than conventional avocadoes. In 2017 the average price of avocadoes was $1.52, and we anticipate that this rise was due to the increase in both conventional and organic avocaodes sold that year. The average price of avocadoes sold in 2017 was $1.30 and $1.74 for conventional and organic avocadoes, respectively.

Differences in Average Price Within the Continental U.S.

About 90% of the avocado production in the United States takes place in California (Dekevich, 2018). Florida and Hawaii produce most of the remaining 10%. (Dekevich, 2018). We anticipate that the average price of avocadoes will be lower in these regions, because of decreased shipping and storage costs.

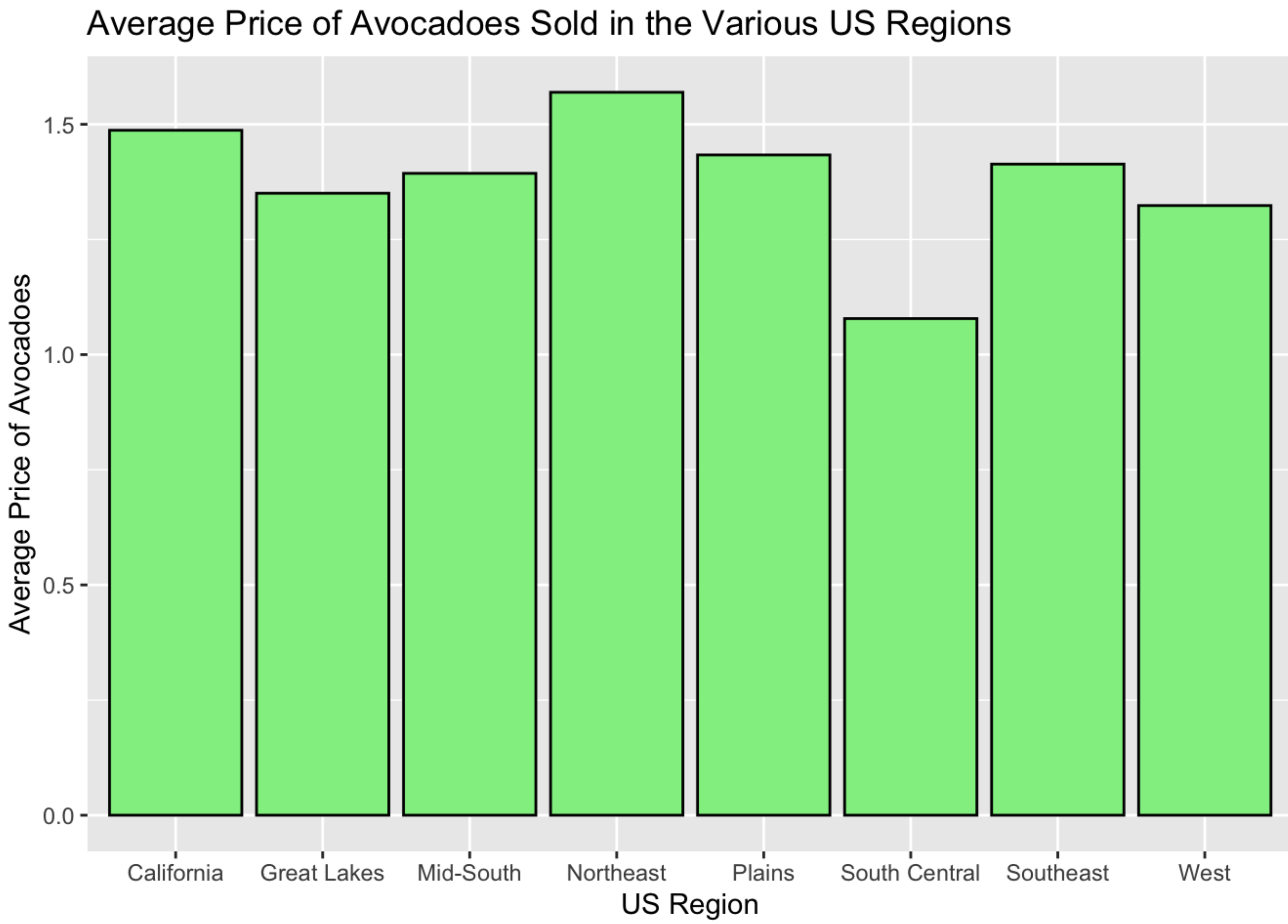
```
## Summarize average price by region
avocado_region<-avocado_tidy%>%
  group_by(US_region)%>%
  summarize(avg_region=mean(Average_Price))

avocado_region
```

US_region <chr>	avg_region <dbl>
California	1.487053
Great Lakes	1.350342
Mid-South	1.393498
Northeast	1.569486
Plains	1.433565
South Central	1.078254
Southeast	1.413609
West	1.323603

8 rows

```
## Bar Plot with aesthetics: average price of avocadoes sold, grou
ped by region
gg<-ggplot(avocado_region,aes(x=US_region,y=avg_region))
gg<-gg+geom_bar(stat="Identity", colour = "black", fill ="light gr
een")
gg<-gg+ylab("Average Price of Avocadoes")+xlab("US Region")
gg<-gg+ggtitle("Average Price of Avocadoes Sold in the Various US
 Regions")
gg
```



It is surprising that the regions that are producing the majority of avocadoes in the United States also have the highest average prices. The average price of avocadoes in California is $1.49, which is $0.08 higher than the unconditional mean of average prices, $1.41. The average price of avocadoes in the Southeast region of the US, which includes Florida, is $1.41, equal to the unconditional mean of average prices. It is apparent that proximity to avocado production is not a likely factor that influences avocado sales price.

In addition, the bar graph shows that South Central has the lowest average price of avocados - $1.08. It is plausible this region has a particularly low average price because of its proximity to Mexico, a country that leads the world in avocado production. However, the dataset's sales and volume information does not differentiate between domestic and imported avocados, so we are unable to verify this hypothesis.

Two Predictors: Summarizing Average Price by Year and U.S. Region

Next we sought to determine whether average prices were influenced by both year and U.S. region of purchase.

```
## Summarize average price by year and US Region
avocadoes_yearRegion<-avocado_tidy%>%
  group_by(Year, US_region)%>%
  summarize(avg_yearRegion=mean(Average_Price))%>%ungroup()%>%arra
nge(US_region)

avocadoes_yearRegion
```

Year <dbl>	US_region <chr>	avg_yearRegion <dbl>
2015	California	1.363538
2016	California	1.455365
2017	California	1.647151
2018	California	1.452500
2015	Great Lakes	1.329148
2016	Great Lakes	1.297473
2017	Great Lakes	1.438868
2018	Great Lakes	1.280298
2015	Mid-South	1.363707
2016	Mid-South	1.319498

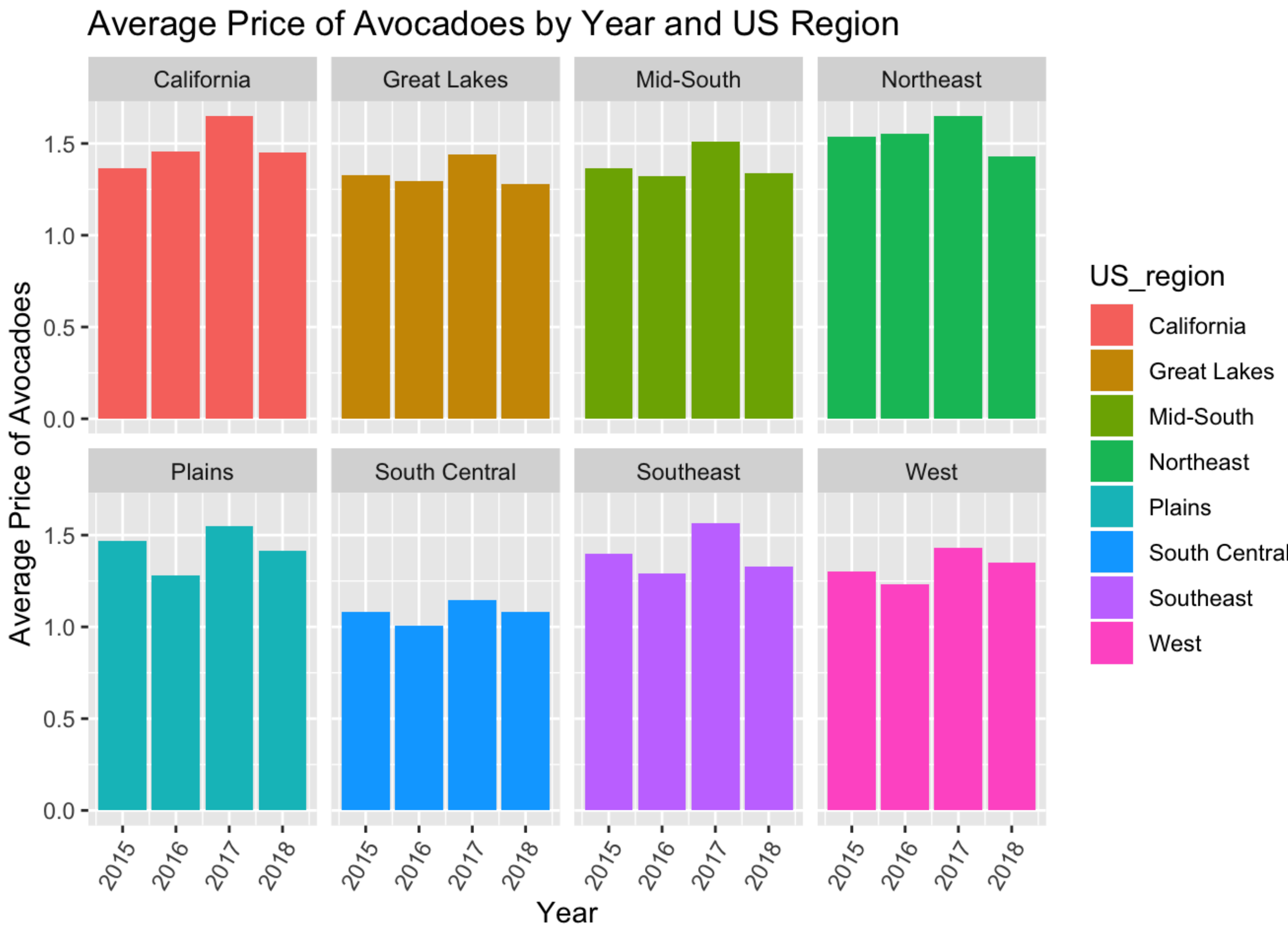
1-10 of 32 rows Previous **1** 2 3 4 Next

```
avocadoes_yearRegion_table <- avocadoes_yearRegion
colnames(avocadoes_yearRegion_table)<-c("Year","US Region","Average Price")
kable(avocadoes_yearRegion_table, align=rep('l', length(avocadoes_yearRegion_table[,1])))
```

Year	US Region	Average Price
2015	California	1.363538
2016	California	1.455365
2017	California	1.647151
2018	California	1.452500
2015	Great Lakes	1.329148
2016	Great Lakes	1.297473
2017	Great Lakes	1.438868
2018	Great Lakes	1.280298
2015	Mid-South	1.363707
2016	Mid-South	1.319498
2017	Mid-South	1.508103
2018	Mid-South	1.337083
2015	Northeast	1.539038
2016	Northeast	1.552605
2017	Northeast	1.648173
2018	Northeast	1.427045
2015	Plains	1.470625

Year	US Region	Average Price
2016	Plains	1.281490
2017	Plains	1.551132
2018	Plains	1.412708
2015	South Central	1.079327
2016	South Central	1.005000
2017	South Central	1.147956
2018	South Central	1.083194
2015	Southeast	1.400797
2016	Southeast	1.290357
2017	Southeast	1.566941
2018	Southeast	1.326012
2015	West	1.301604
2016	West	1.233248
2017	West	1.428141
2018	West	1.349630

```
gg<-ggplot(avocadoes_yearRegion,aes(x=Year,y=avg_yearRegion))
gg<-gg+geom_bar(stat="identity",aes(fill=US_region),position="dodg
e")
gg<-gg+facet_wrap(~US_region,ncol=4)
gg<-gg+ylab("Average Price of Avocadoes")+xlab("Year")
gg<-gg+theme(axis.text.x = element_text(angle = 60, hjust = 1))
gg<-gg+ggtitle("Average Price of Avocadoes by Year and US Region")
gg
```



One of the central questions we aimed to explore was whether or not price fluctuations in U.S. avocado-producing regions were steadier than non-avocado-producing regions. This faceted bar chart show that prices in California, one of U.S.' primary avocado-producing regions, fluctuated more than the other regions in the dataset. Prices in California in 2015 - 2018 ranged from $1.36 to $1.64, a difference of $0.28. The prices in the Southeast, which includes Florida, another avocado-producing region, also ranged from $1.29 to $1.57, a difference of $0.26. The difference in average prices in California and the Southeast is greater than the price fluctuations in the South Central, Great Lakes, West, Northeast, and Mid-South regions.

Two Predictors: Summarizing the Volume of Avocadoes Sold by Year and U.S. Region

Keeping in mind that avocadoes play a large role in the culture and cuisine of California, it is plausible that California's higher average price is due to increased demand for and/or interest in avocadoes, especially in comparison to other regions less familiar with the fruit. Please note that, due to population and size differences, we recognize that it may be difficult to form conclusions based on volume data alone. We would prefer to examine per capita consumption of avocadoes, but the data does not allow for this approach. Nevertheless, we decided to examine whether the volume of avocadoes sold in various regions of the country substantially differed. We are especially interested in how California compares to South Central. Avocadoes are commonly used in Tex-Mex and Mexican cuisine, so we predict that South Central would have a similarly high volume of avocado sales. To further explore the possible trends in our data, we examined the volume of avocadoes sold in 2015 - 2018 in the various regions of the United States.

```
## Summarize total volume by year and US Region
avo_Vol_Yr_Region<-avocado_tidy%>%
  group_by(Year, US_region)%>%
  summarize(Vol_Yr_Region=mean(Total_Volume))%>%ungroup()%>%arrang
e(US_region)

avo_Vol_Yr_Region
```

Year <dbl>	US_region <chr>	Vol_Yr_Region <dbl>
2015	California	1033923.4
2016	California	1121771.9
2017	California	1072422.5
2018	California	1235987.9
2015	Great Lakes	365931.6
2016	Great Lakes	381389.8

Year <dbl>	US_region <chr>	Vol_Yr_Region <dbl>
2017	Great Lakes	397560.1
2018	Great Lakes	492370.2
2015	Mid-South	267947.5
2016	Mid-South	294030.0

1-10 of 32 rows Previous **1** 2 3 4 Next

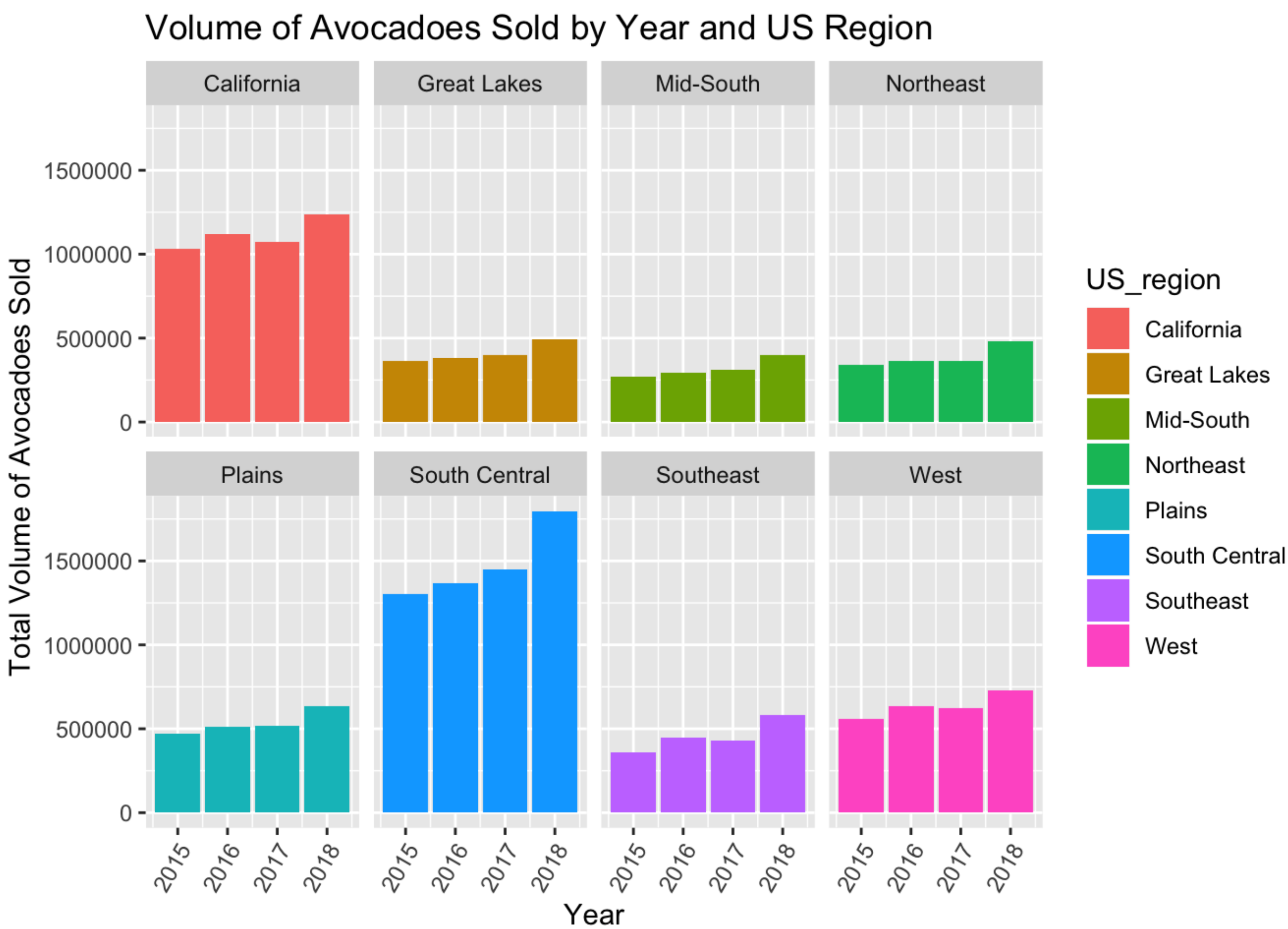
```
avo_Vol_Yr_Region_table <- avo_Vol_Yr_Region
colnames(avo_Vol_Yr_Region_table)<-c("Year","US Region","Mean Tota
l Volume")
kable(avo_Vol_Yr_Region_table, align=rep('l', length(avo_Vol_Yr_Re
gion_table[,1])))
```

Year	US Region	Mean Total Volume
2015	California	1033923.4
2016	California	1121771.9
2017	California	1072422.5
2018	California	1235987.9
2015	Great Lakes	365931.6
2016	Great Lakes	381389.8
2017	Great Lakes	397560.1
2018	Great Lakes	492370.2
2015	Mid-South	267947.5
2016	Mid-South	294030.0

Year	US Region	Mean Total Volume
2017	Mid-South	308588.8
2018	Mid-South	398854.9
2015	Northeast	337785.0
2016	Northeast	363231.3
2017	Northeast	366260.4
2018	Northeast	480670.7
2015	Plains	468948.9
2016	Plains	509067.8
2017	Plains	515415.8
2018	Plains	636786.1
2015	South Central	1301773.6
2016	South Central	1366983.1
2017	South Central	1448920.4
2018	South Central	1798027.4
2015	Southeast	357397.5
2016	Southeast	444479.6
2017	Southeast	430342.2
2018	Southeast	580873.1
2015	West	559433.1
2016	West	634661.1
2017	West	625469.1

Year	US Region	Mean Total Volume
2018	West	728746.5

```
gg<-ggplot(avo_Vol_Yr_Region,aes(x=Year,y=Vol_Yr_Region))
gg<-gg+geom_bar(stat="identity",aes(fill=US_region),position="dodg
e")
gg<-gg+facet_wrap(~US_region,ncol=4)
gg<-gg+ylab("Total Volume of Avocadoes Sold")+xlab("Year")
gg<-gg+theme(axis.text.x = element_text(angle = 60, hjust = 1))
gg<-gg+ggtitle("Volume of Avocadoes Sold by Year and US Region")
gg
```



As predicted, South Central and California have the highest volume of avocadoes sold in 2015 - 2018.

Models and Methods

Since our predominant outcome variable, “Average_Price,” and predictor variables are continuous, we chose to implement a regression model to further investigate our central questions. Based on the density plot of “Average_Price” in our exploratory data analysis, the distribution of this variable is approximately normal, which helps in interpreting the results.

Simple Regression: Model of Average Price as a function of Avocado Type

```
#convert the variable "Type" into a binary variable
avocado_tidy$Avo_Type<-NA

avocado_tidy$Avo_Type[avocado_tidy$Type=="conventional"]<-"0"
avocado_tidy$Avo_Type[avocado_tidy$Type=="organic"]<-"1"

#Model 1: simple regression.
#linear model of average price (dependent variable) as a function
 of type of avocado
mod1 <-lm(avocado_tidy$Average_Price~avocado_tidy$Avo_Type)
summary(mod1)
```

```
##
## Call:
## lm(formula = avocado_tidy$Average_Price ~ avocado_tidy$Avo_Typ
e)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21604 -0.20604 -0.02929  0.19071  1.59396
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.159285   0.003370   344.0   <2e-16 ***
## avocado_tidy$Avo_Type1 0.496751   0.004767   104.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.319 on 17909 degrees of freedom
## Multiple R-squared:  0.3775, Adjusted R-squared:  0.3775
## F-statistic: 1.086e+04 on 1 and 17909 DF,  p-value: < 2.2e-16
```

```
rmse(mod1,avocado_tidy)
```

```
## [1] 0.3189373
```

This simple linear regression model indicates that there is a statistically significant relationship between the type of avocado and average price. We can reject the null hypothesis that the coefficient is zero; the association is not likely due to random chance. The coefficient demonstrates that organic avocadoes are predicted to have an increased average price of $0.50, and the intercept indicates that conventional avocadoes are predicted to have an average price of $1.16. The RMSE value shows that the error in this model is approximately 0.32; that is, the model is on average $0.32 off in predicting the average price.

Multiple Regression: Model of Average Price as a function of Avocado Type and Volume

```
##adding in percentile rank of total_volume as an additional predi
ctor, in addition to avocado type

mod2<-lm(Average_Price~as.factor(Avo_Type)+
          Volume_Rank,
          data=avocado_tidy)

summary(mod2)
```

```
##
## Call:
## lm(formula = Average_Price ~ as.factor(Avo_Type) + Volume_Rank,
##     data = avocado_tidy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14523 -0.20602 -0.02733  0.17799  1.61620
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.447019   0.011173  129.51   <2e-16 ***
## as.factor(Avo_Type)1  0.309935   0.008358   37.08   <2e-16 ***
## Volume_Rank          -0.395014   0.014653  -26.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3127 on 17908 degrees of freedom
## Multiple R-squared:  0.4018, Adjusted R-squared:  0.4017
## F-statistic:  6014 on 2 and 17908 DF,  p-value: < 2.2e-16
```

```
rmse(mod2,avocado_tidy)
```

```
## [1] 0.3126565
```

The two-predictor, multiple regression model indicates that both avocado type and volume are significant predictors of average price. By adding in the second predictor, Volume_Rank, we lowered the RMSE of our model down from 0.319 to 0.313. The RMSE value of our two-predictor model indicates that our model is approximately $0.31 off on average in predicting the average price of avocadoes.

Classification Model

We then utilized the classification model to predict how likely it is that the average price of avocados is above or below the unconditional mean of $1.41. By making the dependent variable, average price of avocados, a binary variable, we gain a better understanding of the extent to which certain predictors influence the average price of avocadoes. The following command first converts our continuous dependent variable to a binary variable:

‘Average_Price_Binary’ = 1 if ‘Average_Price’ is greater than or equal to 1.41.
‘Average_Price_Binary’ = 0 if ‘Average_Price’ is less than 1.41.

```
avocado_tidy$Average_Price_Binary<-NA

avocado_tidy$Average_Price_Binary[avocado_tidy$Average_Price>"1.4
1"]<-"1"
avocado_tidy$Average_Price_Binary[avocado_tidy$Average_Price=="1.4
1"]<-"1"
avocado_tidy$Average_Price_Binary[avocado_tidy$Average_Price<"1.4
1"]<-"0"
```

Next we determined the proportion of average prices that were above or below the unconditional mean of 1.41.

```
table(avocado_tidy$Average_Price_Binary)
```

```
##
##    0    1
## 9477 8434
```

```
prop.table(table(avocado_tidy$Average_Price_Binary))
```

```
##
##         0         1
## 0.5291162 0.4708838
```

Approximately 47% of the data indicates an average avocado price greater than or equal to the unconditional mean of 1.41, and approximately 53% of the dataset is below this unconditional mean. The below cross-tab table shows average prices above or below the unconditional mean by raw count and proportion.

```
avocado_tidy%>%
  count(Average_Price_Binary)%>% # Count numbers of observations a
bove 1.41
  mutate(p=prop.table(n))%>% #mutate for proportions using prop.ta
ble
  kable(format="markdown") # output to table
```

Average_Price_Binary	n	p
0	9477	0.5291162
1	8434	0.4708838

We then cross-tabulated this information by year. Average prices were greater than or equal to the unconditional mean (1.41) 43.4% of the time in 2015, 40.1% of the time in 2016, 58.6% of the time in 2017, and 42.3% of the time in 2018.

```
table1 <- prop.table(table(avocado_tidy$Year,avocado_tidy$Average_
Price_Binary),margin=1)
colnames(table1)<-c("Below Average Price","Equal to or Above Avera
ge Price")
kable(table1, align=rep('l', length(table1[,1])))
```

	Below Average Price	Equal to or Above Average Price
2015	0.5655961	0.4344039
2016	0.5990566	0.4009434
2017	0.4139957	0.5860043
2018	0.5762579	0.4237421

Next we again cross-tabulated this information, this time by U.S. region. The dataset shows that in California and Southeast, regions that produce avocados, 50% and 49% of the weekly retail sales were greater than the unconditional mean of $1.41. In contrast, 84% of the weekly sales in South Central were below the unconditional mean.

```
table2 <- prop.table(table(avocado_tidy$US_region,avocado_tidy$Ave
rage_Price_Binary), margin=1)
colnames(table2)<-c("Below Average Price","Equal to or Above Avera
ge Price")
kable(table2, align=rep('l', length(table2[,1])))
```

	Below Average Price	Equal to or Above Average Price
California	0.5011834	0.4988166
Great Lakes	0.5743872	0.4256128
Mid-South	0.5453649	0.4546351
Northeast	0.3633674	0.6366326

	Below Average Price	Equal to or Above Average Price
Plains	0.5192308	0.4807692
South Central	0.8412229	0.1587771
Southeast	0.5059172	0.4940828
West	0.6120434	0.3879566

```
# Linear model
lm_mod_2<-lm(Average_Price_Binary~
             US_region+
             Avo_Type,
           data=avocado_tidy,y=TRUE,na.exclude=TRUE);summary(lm_mo
d_2)
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singula
r.ok, ...) :
##  extra argument 'na.exclude' will be disregarded
```

```
##
## Call:
## lm(formula = Average_Price_Binary ~ US_region + Avo_Type, data
= avocado_tidy,
##     y = TRUE, na.exclude = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93316 -0.19755 -0.09172  0.24883  0.90828
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.202286   0.009804  20.632  < 2e-16 **
*
## US_regionGreat Lakes   -0.073204   0.012271  -5.966 2.48e-09 **
*
## US_regionMid-South     -0.044181   0.011689  -3.780 0.000158 **
*
## US_regionNortheast      0.137816   0.011303  12.192  < 2e-16 **
*
## US_regionPlains        -0.018047   0.017534  -1.029 0.303360
## US_regionSouth Central -0.340039   0.015305 -22.218  < 2e-16 **
*
## US_regionSoutheast     -0.004734   0.012271  -0.386 0.699679
## US_regionWest          -0.110567   0.011691  -9.457  < 2e-16 **
*
## Avo_Type1               0.593062   0.005758 103.001  < 2e-16 **
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3853 on 17902 degrees of freedom
## Multiple R-squared:  0.4045, Adjusted R-squared:  0.4042
## F-statistic:  1520 on 8 and 17902 DF,  p-value: < 2.2e-16
```

This model indicates that the type of avocado (conventional, organic) and whether the avocadoes were sold in the Great Lakes, Mid-South, Northeast, South Central, and West regions were significant predictors of whether an avocado was sold above average price. We then ran predictions based on this linear model. Everything above 0.5 was predicted to be above average price, everything below 0.5 was predicted to be below average price.

```
#Predictions
avocado_tidy<-avocado_tidy%>%
  add_predictions(lm_mod_2)%>% ## Add in predictions from the mode
l
  rename(pred_lm=pred)%>% ## rename to be predictions from ols (l
m)
  mutate(pred_lm_out=ifelse(pred_lm>=.5,1,0))
```

We then created a table that shows the predictions of the model against what actually happened.

```
pred_table<-table(avocado_tidy$Average_Price_Binary,avocado_tidy$p
red_lm_out)
pred_table
```

```
##
##         0    1
##    0 7741 1736
##    1 1723 6711
```

```
prop.table(pred_table)
```

```
##
##               0          1
##    0 0.43219251 0.09692368
##    1 0.09619787 0.37468595
```

```
rownames(pred_table)<-c("Predicted 0","Predicted 1")
colnames(pred_table)<-c("Actually 0","Actually 1")
pred_table
```

```
##
##               Actually 0 Actually 1
##   Predicted 0       7741       1736
##   Predicted 1       1723       6711
```

The prediction table above indicates that 7,741 avocado sales were accurately predicted to be less than the unconditional average of $1.41, and 6,711 avocado sales were accurately predicted to be above the unconditional average. However, 1,736 avocado sales were predicted to be below the unconditional average, but in fact, they were above $1.41. Furthermore, 1,723 avocado sales were predicted to be above the unconditional average, but in fact they were below $1.41.

Experimental Model

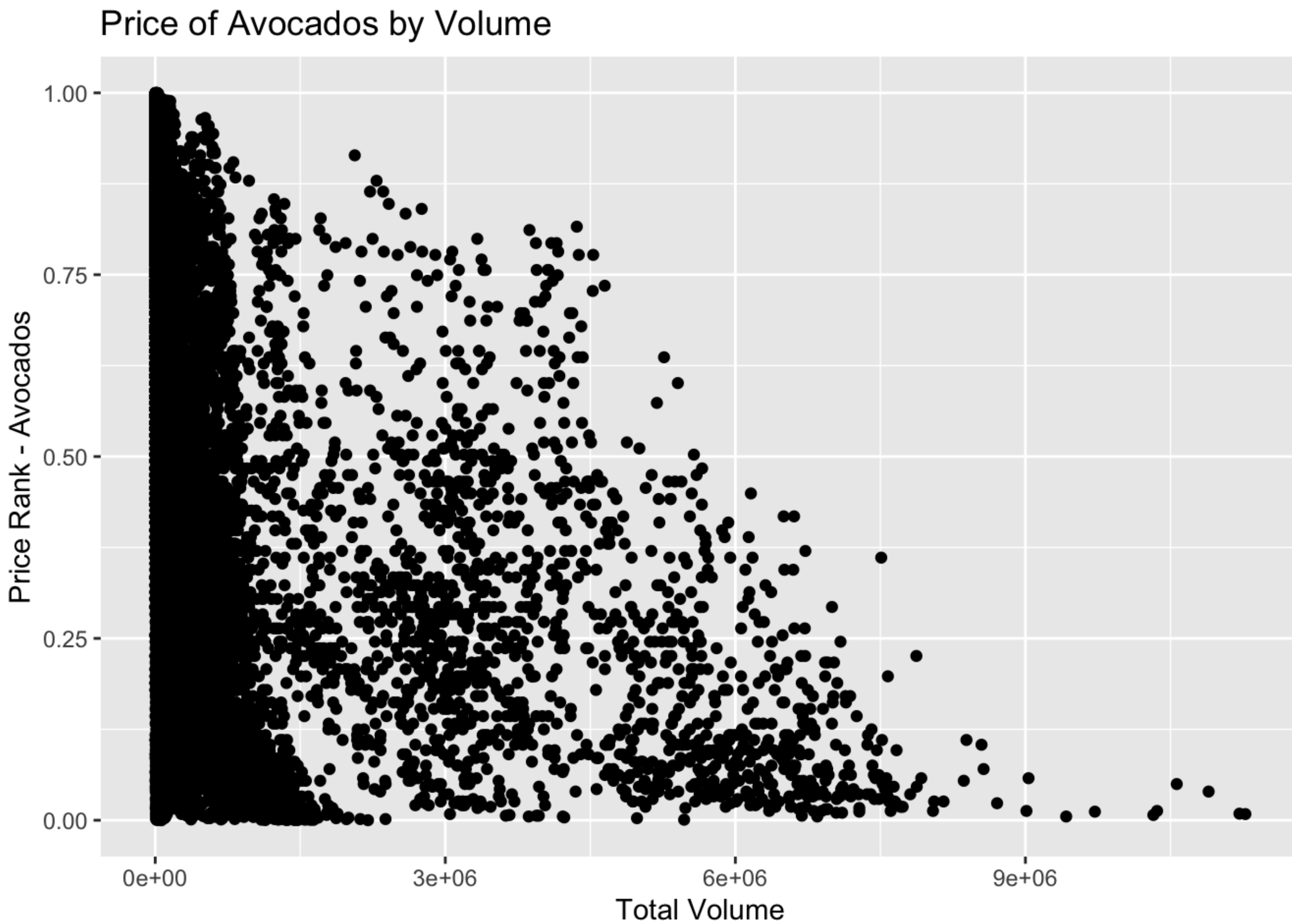
Next we will determine how well our model predicts outcomes outside our sample by creating both a testing and training dataset. The training data will be used to generate our predictions and train our model, while the testing data will be used to validate these predictions and determine how accurate our model is at predicting outcomes.

First we create a simple model that predicts the average price of avocados as a function of avocado type, volume, and year of sale.

```
avocado_tidy_model<-avocado_tidy%>%
  select(Average_Price,Total_Volume,Avo_Type,Year)%>%
  mutate_all(funs(as.numeric))%>%
  mutate(price_rank=percent_rank(Average_Price))%>%
  tbl_df()
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

```
gg<-ggplot(avocado_tidy_model, aes(x=Total_Volume,y=price_rank))
gg<-gg+geom_point()
gg<-gg+ylab("Price Rank - Avocados")+xlab("Total Volume")
gg<-gg+ggtitle("Price of Avocados by Volume")
gg
```



This scatterplot demonstrates that volume of avocado sales and price is negatively correlated - as volume increases, price decreases.

Next we define the model to determine the effect of total volume of avocados sold and avocado type on the average price.

```
## Define the model
mod3_formula<-formula(price_rank~Total_Volume+
                        Avo_Type)
## Run the model against all of the data
basic.mod<-lm(mod3_formula,
              data=avocado_tidy_model); summary(basic.mod)
```

```
##
## Call:
## lm(formula = mod3_formula, data = avocado_tidy_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68219 -0.15812 -0.00196  0.16904  0.64006
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.398e-01  2.786e-03  121.96   <2e-16 ***
## Total_Volume -2.772e-08  1.473e-09  -18.82   <2e-16 ***
## Avo_Type      3.431e-01  3.607e-03   95.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2195 on 17908 degrees of freedom
## Multiple R-squared:  0.4237, Adjusted R-squared:  0.4237
## F-statistic:  6584 on 2 and 17908 DF,  p-value: < 2.2e-16
```

The basic linear model indicates that both avocado type and volume are significant predictors of average price. The RMSE value of our linear model indicates that our model is approximately $0.22 off on average in predicting the average price of avocadoes.

We will now use the `crossv_kfold` command to create a list of datasets from the original dataset. Each has a testing and training dataset. We set the command to 30 folds, so 1/30 of the data will be held out for testing.

```
avocado_tidy_model_cf<-avocado_tidy_model%>%
  crossv_kfold(30)
avocado_tidy_model_cf
```

train	test	.id
<list>	<list>	<chr>

train	test	.id
<list>	<list>	<chr>
<S3: resample>	<S3: resample>	01
<S3: resample>	<S3: resample>	02
<S3: resample>	<S3: resample>	03
<S3: resample>	<S3: resample>	04
<S3: resample>	<S3: resample>	05
<S3: resample>	<S3: resample>	06
<S3: resample>	<S3: resample>	07
<S3: resample>	<S3: resample>	08
<S3: resample>	<S3: resample>	09
<S3: resample>	<S3: resample>	10

1-10 of 30 rows Previous **1** 2 3 Next

We then run the model on each training dataset by first converting them into tibbles. We apply the predictions from the model to each testing dataset, and finally pull the RMSE from each.

```
tic()
rmse_mod3<-avocado_tidy_model_cf %>%
  mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
  mutate(model = map(train, ~ lm(mod3_formula,
                                 data = .))) %>%
  mutate(rmse = map2_dbl(model, test, rmse)) %>% ## apply model, g
et rmse
  select(.id, rmse) ## pull just id and rmse
toc()
```

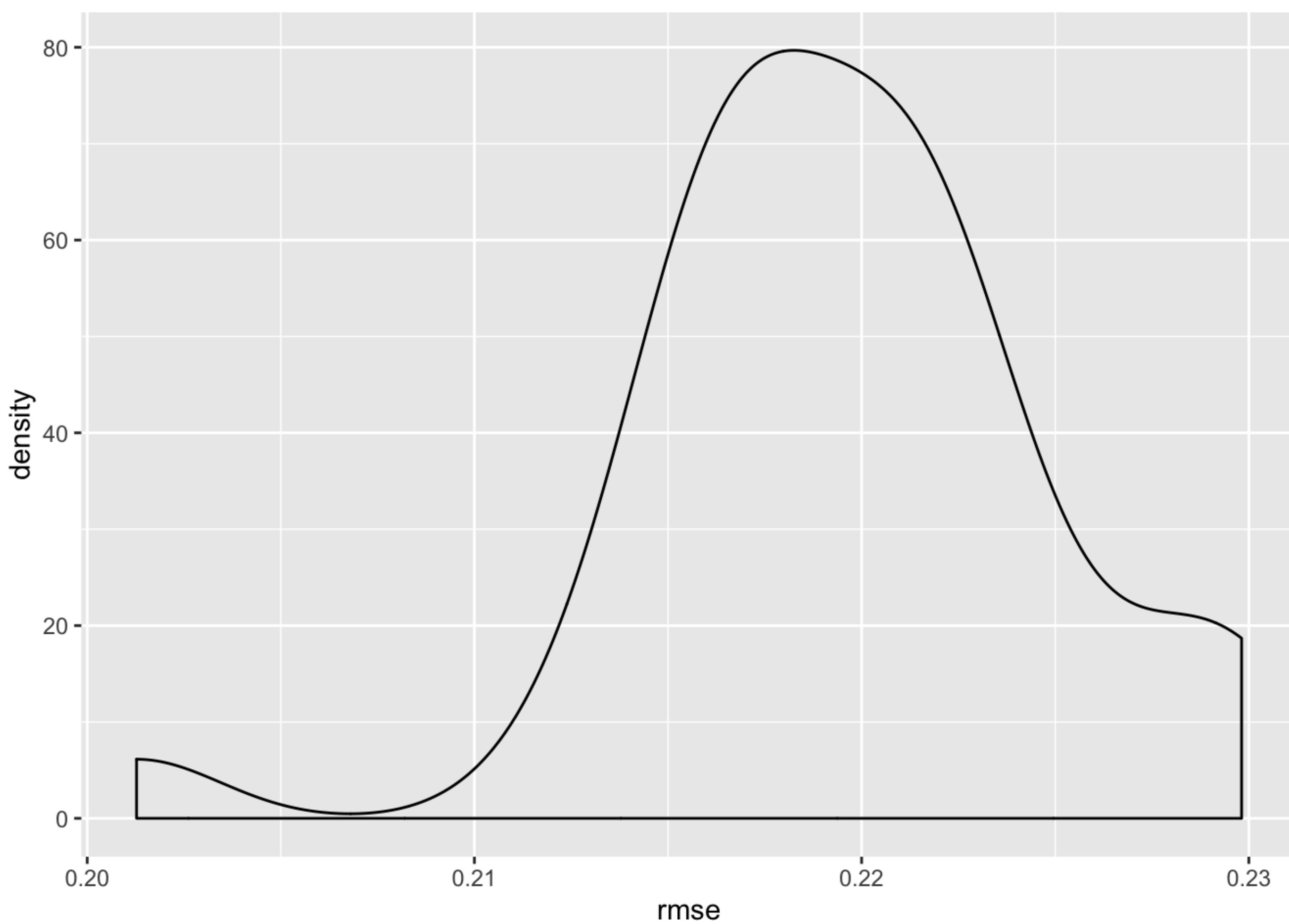
```
## 0.902 sec elapsed
```

We then used ‘ggplot’ to determine the range of our RMSE.

```
summary(rmse_mod3$rmse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2013  0.2161  0.2191  0.2195  0.2225  0.2298
```

```
gg<-ggplot(rmse_mod3,aes(rmse))
gg<-gg+geom_density()
gg
```



The code below demonstrates the minimum, maximum, and RMSE for ‘rmse_mod3,’ respectively:

```
round(summary(rmse_mod3$rmse)[1],4)
```

```
##   Min.
## 0.2013
```

```
round(summary(rmse_mod3$rmse)[6],4)
```

```
##   Max.
## 0.2298
```

```
round(summary(rmse_mod3$rmse)[3],4)
```

```
## Median
## 0.2191
```

As this shows, the rmse for the crossfold validations goes from the a minimum of 0.2013 to a maximum of 0.2298, with a median of 0.2191. The range of RMSE is narrow, 0.0242.

Full Cross Validation: Random Partition

Another way of testing the model's ability to predict data is to utilize the full cross validation, creating random splits of the dataset into training and testing data. The `crossv_mc` command provides for a generalization of the crossfold command. For this command, we can specify the proportion to be randomly held out in each iteration, via `test=p` where `p` is the proportion to be held out.

```
avocado_tidy_model_cv<-avocado_tidy_model%>%
  crossv_mc(n=100,test=.2)
avocado_tidy_model_cv
```

train	test	.id
<list>	<list>	<chr>
<S3: resample>	<S3: resample>	001
<S3: resample>	<S3: resample>	002
<S3: resample>	<S3: resample>	003
<S3: resample>	<S3: resample>	004
<S3: resample>	<S3: resample>	005
<S3: resample>	<S3: resample>	006
<S3: resample>	<S3: resample>	007
<S3: resample>	<S3: resample>	008
<S3: resample>	<S3: resample>	009
<S3: resample>	<S3: resample>	010

1-10 of 100 rows Previous **1** 2 3 4 5 6 ... 10 Next

The `avocado_tidy_model_cv` dataset is a dataset of 100 test-training pairs generated. The testing dataset is .2 of the sample, the proportion of observations that is held out for testing, and it's different every time.

```
tic()
mod3_rmse_cv<-avocado_tidy_model_cv %>%
  mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
  mutate(model = map(train, ~ lm(mod3_formula, data = .)))%>%
  mutate(rmse = map2_dbl(model, test, rmse))%>%
  select(.id, rmse) ## pull just id and rmse

mod3_rmse_cv
```

.id	rmse
<chr>	<dbl>

.id <chr>	rmse <dbl>
001	0.2190828
002	0.2179631
003	0.2204345
004	0.2219141
005	0.2172512
006	0.2204865
007	0.2196124
008	0.2187576
009	0.2180522
010	0.2187900

1-10 of 100 rows Previous **1** 2 3 4 5 6 ... 10 Next

```
toc()
```

```
## 1.939 sec elapsed
```

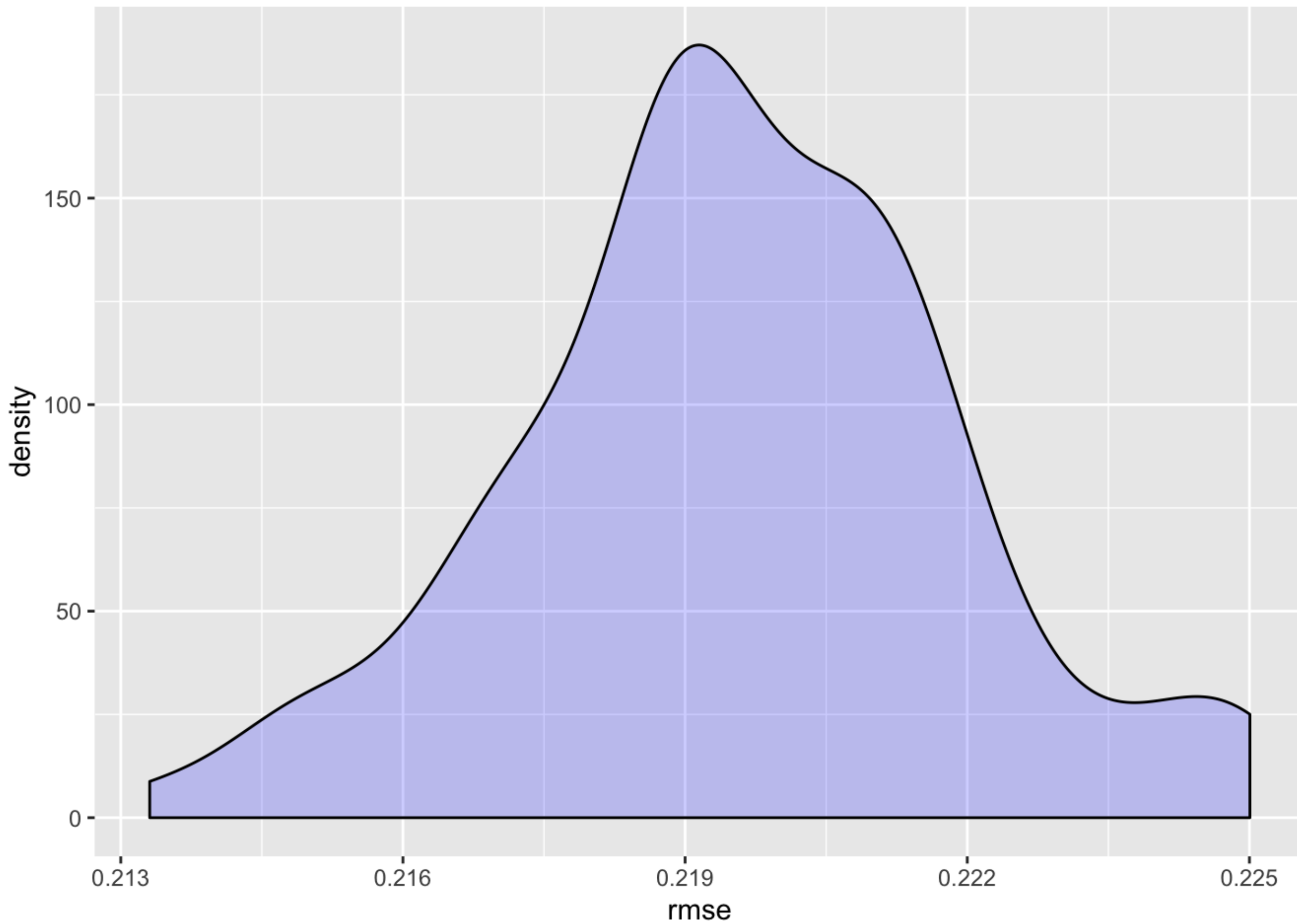
```
summary(mod3_rmse_cv$rmse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2133  0.2184  0.2194  0.2195  0.2211  0.2250
```

```
gg<-ggplot(mod3_rmse_cv,aes(rmse))
gg<-gg+geom_density(bins=50,fill="blue",alpha=.2)
```

```
## Warning: Ignoring unknown parameters: bins
```

```
gg
```



The code below demonstrates the minimum, maximum, and RMSE for ‘mod3_rmse_cv,’ respectively:

```
round(summary(mod3_rmse_cv$rmse)[1],4)
```

```
##   Min.
## 0.2133
```

```
round(summary(mod3_rmse_cv$rmse)[6],4)
```

```
##  Max.
## 0.225
```

```
round(summary(mod3_rmse_cv$rmse)[3],4)
```

```
## Median
## 0.2194
```

As this shows, the rmse for the crossfold validations goes from the a minimum of 0.2133 to a maximum of 0.225, with a median of 0.2194.

Selecting Between Models

We then compare the two cross-validated models to see which performed better:

```
tic()
## Define the model
mod4_formula<-formula("price_rank ~
                      Total_Volume+
                      Avo_Type+
                      Year")



mod4_rmse_cv<-avocado_tidy_model_cv %>%
  mutate(train = map(train, as_tibble)) %>% ## Convert to tibbles
  mutate(model = map(train, ~ lm(mod4_formula, data = .)))%>%
  mutate(rmse = map2_dbl(model, test, rmse))%>%
  select(.id, rmse) ## pull just id and rmse

summary(mod4_rmse_cv$rmse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2109  0.2164  0.2178  0.2178  0.2193  0.2234
```

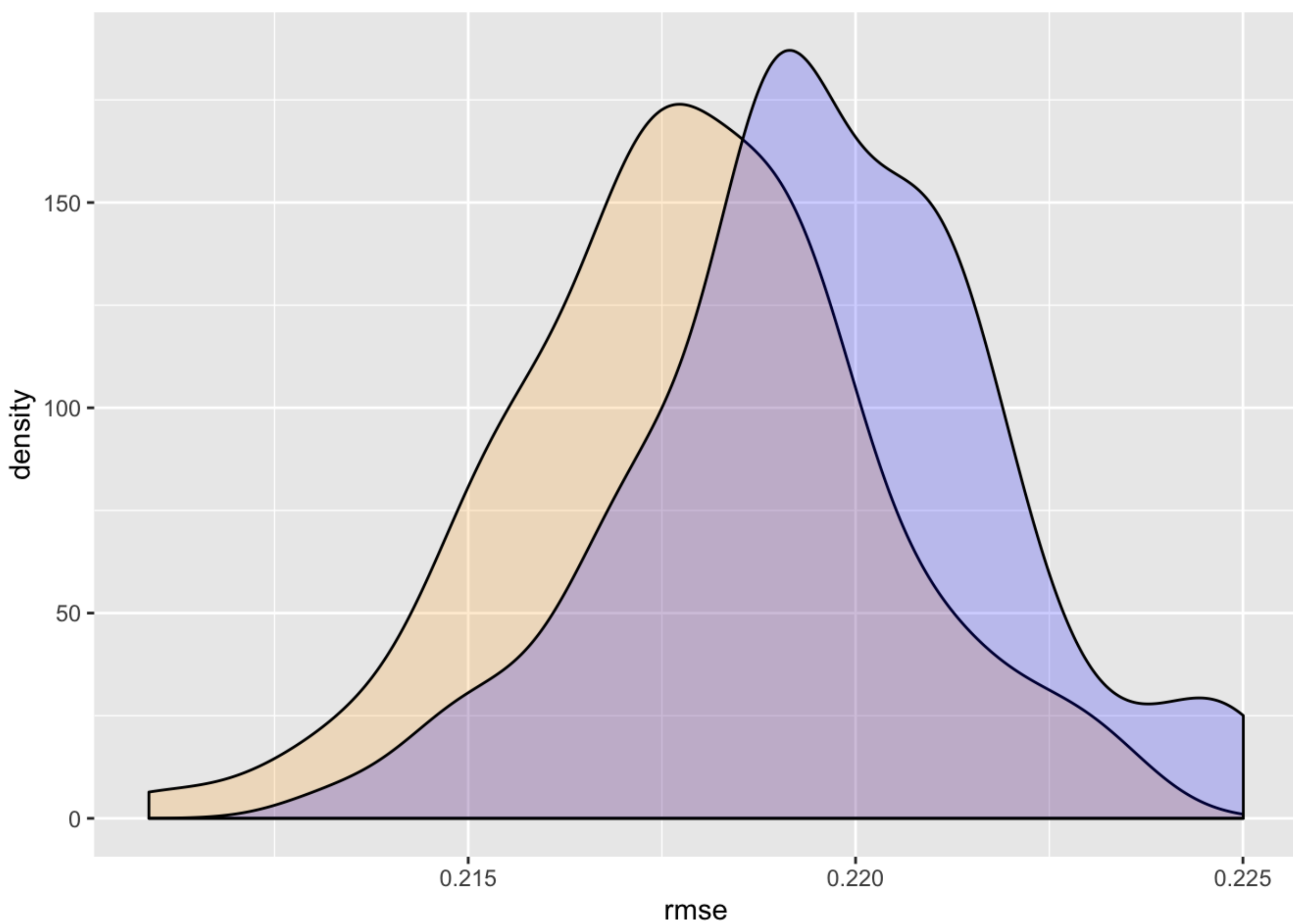
```
summary(mod3_rmse_cv$rmse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2133  0.2184  0.2194  0.2195  0.2211  0.2250
```

```
toc()
```

```
## 1.782 sec elapsed
```

```
gg<-ggplot(mod4_rmse_cv,aes(x=rmse))
gg<-gg+geom_density(fill="orange",alpha=.2)
gg<-gg+geom_density(data=mod3_rmse_cv,aes(x=rmse),fill="blue",alph
a=.2)
gg
```



Although we observe overlap in the performance between the two models, model 4 (orange) depicts a lower RMSE for out-of-sample predictions. The mean RMSE value in model 4 was 0.2178, which is slightly lower than the mean RMSe value in model 3, 0.2196. Furthermore, the maximum RMSE value of model 4, 0.2267, is lower than the maximum RMSE value of model 3, 0.2281. This shows that model 4 is a more accurate model.

Machine Learning

We could let the computer choose a model from a set of candidate variables. Here we use stepwise regression, which involves proposing variables and tasking the computer to evaluate its ability to lower RMSE. The below commands allow the computer to select the covariates that predict the outcome variable.

```
#Tuning model parameters
avocado_tidy_model<-avocado_tidy_model%>%select(-Average_Price)


fitControl<-trainControl(method="boot",
                         p=.2)


fit1<-train(price_rank~Total_Volume+
              Avo_Type,
            method="lm",
            data=avocado_tidy_model,
            trControl=fitControl)

summary(fit1)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68219 -0.15812 -0.00196  0.16904  0.64006
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.398e-01  2.786e-03  121.96   <2e-16 ***
## Total_Volume -2.772e-08  1.473e-09  -18.82   <2e-16 ***
## Avo_Type      3.431e-01  3.607e-03   95.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2195 on 17908 degrees of freedom
## Multiple R-squared:  0.4237, Adjusted R-squared:  0.4237
## F-statistic:  6584 on 2 and 17908 DF,  p-value: < 2.2e-16
```

```
fit1$results
```

	intercept <lgl>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	Rsquar
1	TRUE	0.2193044	0.4252843	0.1790779	0.001672324	0.0091

1 row

```r
## Stepwise Regression
fit2<-train(price_rank~.,
            data=avocado_tidy_model,
            method="glmStepAIC",
            trControl=fitControl)
```

```
## Start:  AIC=-3860.4
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              844.89 -3860.4
## - Year          1   859.04 -3564.8
## - Total_Volume  1   865.32 -3434.4
## - Avo_Type      1  1272.05  3466.5
## Start:  AIC=-3703.55
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              852.32 -3703.5
## - Year          1   868.33 -3372.2
## - Total_Volume  1   869.24 -3353.4
## - Avo_Type      1  1276.88  3534.4
## Start:  AIC=-3709.22
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              852.05 -3709.2
## - Year          1   863.29 -3476.4
## - Total_Volume  1   870.73 -3322.9
## - Avo_Type      1  1283.28  3623.9
## Start:  AIC=-3681.63
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              853.36 -3681.6
## - Year          1   865.14 -3438.1
## - Total_Volume  1   872.72 -3281.9
## - Avo_Type      1  1275.07  3508.9
## Start:  AIC=-3658.15
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              854.48 -3658.2
## - Year          1   867.72 -3384.7
```

```
## - Total_Volume  1   872.31 -3290.2
## - Avo_Type      1  1299.20  3844.7
## Start:  AIC=-3988.66
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              838.86 -3988.7
## - Year          1   855.35 -3642.0
## - Total_Volume  1   855.58 -3637.1
## - Avo_Type      1  1274.39  3499.4
## Start:  AIC=-3908.47
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              842.62 -3908.5
## - Year          1   857.77 -3591.4
## - Total_Volume  1   859.20 -3561.5
## - Avo_Type      1  1293.20  3761.8
## Start:  AIC=-3678.13
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              853.53 -3678.1
## - Year          1   866.02 -3419.8
## - Total_Volume  1   871.44 -3308.1
## - Avo_Type      1  1278.08  3551.2
## Start:  AIC=-3576.12
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              858.40 -3576.1
## - Year          1   870.79 -3321.4
## - Total_Volume  1   876.58 -3202.8
## - Avo_Type      1  1299.11  3843.5
## Start:  AIC=-3364.27
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
```

```
## <none>                868.62 -3364.3
## - Year           1   881.96 -3093.3
## - Total_Volume   1   884.06 -3050.6
## - Avo_Type       1  1295.95  3799.9
## Start:  AIC=-3952.26
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>                840.56 -3952.3
## - Year           1   853.48 -3681.2
## - Total_Volume   1   858.08 -3584.9
## - Avo_Type       1  1277.05  3536.7
## Start:  AIC=-3772.59
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>                849.04 -3772.6
## - Year           1   862.41 -3494.7
## - Total_Volume   1   867.01 -3399.5
## - Avo_Type       1  1288.22  3692.7
## Start:  AIC=-3737.9
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>                850.69 -3737.9
## - Year           1   863.20 -3478.4
## - Total_Volume   1   866.78 -3404.1
## - Avo_Type       1  1280.32  3582.5
## Start:  AIC=-3621.89
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>                856.21 -3621.9
## - Year           1   871.96 -3297.4
## - Total_Volume   1   872.87 -3278.9
## - Avo_Type       1  1287.48  3682.4
## Start:  AIC=-4152.69
## .outcome ~ Total_Volume + Avo_Type + Year
```

```
##
##                  Df Deviance     AIC
## <none>                831.21 -4152.7
## - Total_Volume  1   847.95 -3797.6
## - Year          1   848.03 -3795.9
## - Avo_Type      1  1270.64  3446.6
## Start:  AIC=-3472.15
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                  Df Deviance     AIC
## <none>                863.40 -3472.1
## - Year          1   874.23 -3250.9
## - Total_Volume  1   880.59 -3121.1
## - Avo_Type      1  1304.64  3919.6
## Start:  AIC=-3904.01
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                  Df Deviance     AIC
## <none>                842.83 -3904.0
## - Year          1   856.70 -3613.7
## - Total_Volume  1   861.72 -3509.0
## - Avo_Type      1  1277.03  3536.4
## Start:  AIC=-3611.63
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                  Df Deviance     AIC
## <none>                856.70 -3611.6
## - Year          1   869.94 -3339.0
## - Total_Volume  1   874.53 -3244.8
## - Avo_Type      1  1284.90  3646.5
## Start:  AIC=-3878.89
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                  Df Deviance     AIC
## <none>                844.02 -3878.9
## - Year          1   856.39 -3620.2
## - Total_Volume  1   862.92 -3484.1
## - Avo_Type      1  1288.18  3692.1
```

```
## Start:  AIC=-3740.86
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>               850.54 -3740.9
## - Year          1   865.64 -3427.8
## - Total_Volume  1   868.24 -3374.1
## - Avo_Type      1  1281.66  3601.2
## Start:  AIC=-3558.06
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>               859.27 -3558.1
## - Year          1   872.52 -3285.9
## - Total_Volume  1   876.68 -3200.8
## - Avo_Type      1  1289.31  3707.9
## Start:  AIC=-3914.11
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>               842.36 -3914.1
## - Year          1   856.09 -3626.4
## - Total_Volume  1   862.82 -3486.3
## - Avo_Type      1  1278.26  3553.7
## Start:  AIC=-4007.32
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>               837.98 -4007.3
## - Year          1   854.13 -3667.4
## - Total_Volume  1   857.26 -3602.0
## - Avo_Type      1  1288.04  3690.2
## Start:  AIC=-3856.67
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>               845.06 -3856.7
## - Year          1   857.20 -3603.3
```

```
## - Total_Volume  1   864.68 -3447.6
## - Avo_Type      1  1286.19  3664.5
## Start:  AIC=-3623.41
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              856.14 -3623.4
## - Year          1   869.65 -3345.0
## - Total_Volume  1   873.53 -3265.3
## - Avo_Type      1  1291.09  3732.6
## Start:  AIC=-3769.66
## .outcome ~ Total_Volume + Avo_Type + Year
##
##                Df Deviance     AIC
## <none>              849.18 -3769.7
## - Year          1   863.00 -3482.5
## - Total_Volume  1   867.27 -3394.0
## - Avo_Type      1  1283.14  3622.0
```

```
summary(fit2)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.70741  -0.15503   0.00351   0.16892   0.64049
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.928e+01  3.492e+00  -16.98   <2e-16 ***
## Total_Volume -2.856e-08  1.462e-09  -19.53   <2e-16 ***
## Avo_Type      3.423e-01  3.578e-03   95.66   <2e-16 ***
## Year          2.957e-02  1.732e-03   17.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.0474215
7)
##
##     Null deviance: 1497.54  on 17910  degrees of freedom
## Residual deviance:  849.18  on 17907  degrees of freedom
## AIC: -3769.7
##
## Number of Fisher Scoring iterations: 2
```

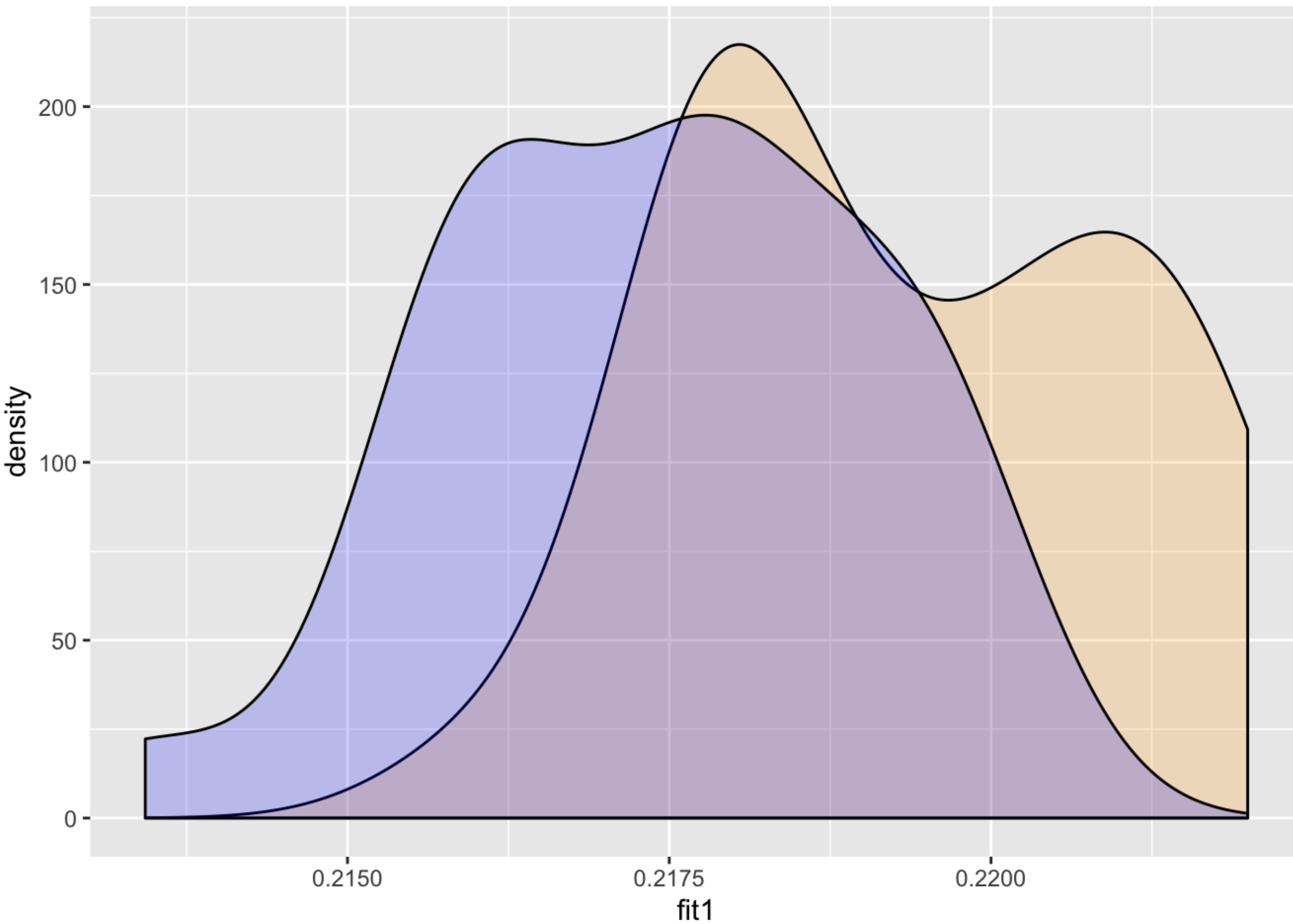
```
fit2$results
```

	parameter <fctr>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>	RMSESD <dbl>	Rsqua
1	none	0.2174165	0.4331066	0.1771545	0.001643104	0.0083

1 row

```
rmse_data<-tbl_df(data.frame(fit1$resample$RMSE,fit2$resample$RMS
E))
names(rmse_data)<-c("fit1","fit2")

gg<-ggplot(rmse_data,aes(x=fit1))
gg<-gg+geom_density(fill="orange",alpha=.2)
gg<-gg+geom_density(aes(x=fit2),fill="blue",alpha=.2)
gg
```



When we implemented the stepwise regression, we tasked the computer to look at every possible model to run among the variables available, and choose the model of best fit. The model that the computer likes the best includes year, total volume of avocadoes sold, and avocado type as the predictors. They are all statistically significant predictors. Volume is negatively associated with average price of avocadoes, and organic avocadoes is correlated with a higher sales price. The results indicated that even with this method of cross-validation, the average RMSE was 0.2175, which is slighly better than the model 4 RMSE value from above.

The ggplot shows the distribution of error. The orange region shows the range of error when we used our simple model with just a couple of covariates, and the blue region shows the distribution of RMSE when we were predicting using a set of covariates as suggested by the computer. The model performance is slightly better for the stepwise regression that was suggested by the computer.

Concluding Remarks

Our analysis set out to determine three central questions: (1) do price fluctuations within avocado-producing regions remain steadier than in non-avocado-producing regions; (2) how do prices fluctuate by region; and (3) does the origin (U.S. region) or type of avocado (conventional vs. organic) influence the price of the avocado?

Our results were surprising: prices in California, one of U.S.' primary avocado-producing regions, fluctuated more than the other regions in the dataset. The difference in average prices in California and the Southeast was greater than the price fluctuations in the South Central, Great Lakes, West, Northeast, and Mid-South regions. In future analyses, it might be useful to explore the various environmental and economic factors that might influence the volatility of prices in these areas.

Contrary to what we initially anticipated, avocado-producing regions, specifically California and South Central, exhibited higher Hass avocado prices than other U.S. regions. Though suprising, this conclusion is not definitive, as we are lacking data from one of the top-three U.S. producers of Hass Avocados: Hawaii. Furthermore, we do not know what percentage of the avocados sold in a region were US-grown. It's likely, for example, that some of the avocados sold in South Central were imported from Mexico. The dataset does not clearly state the origin of the avocados sold, which limits our interpretations of the analysis.

On the other hand, we utilized a simple linear regression model to determine that there was a statistically significant relationship between avocado price and type, with organic avocados having consistently higher prices than conventional avocados.

References

Cernansky, R. (2018, November 20). Organic farming: Why we don't have more organic farms. Retrieved, October 25, 2019, from https://www.nationalgeographic.com/environment/future-of-food/organic-farming-crops-consumers/ (https://www.nationalgeographic.com/environment/future-of-food/organic-farming-crops-consumers/).

Cummings, W. (2017, May 16). Millionaire to Millennials: Your avocado toast addiction is costing you a house. Retrieved October 14, 2019, from https://www.usatoday.com/story/money/2017/05/16/millionaire-tells-millennials-your-avocado-addiction-costing-you-house/101727712/ (https://www.usatoday.com/story/money/2017/05/16/millionaire-tells-millennials-your-avocado-addiction-costing-you-house/101727712/).

Dekevich, D. (2018). Avocados. Retrieved October 14, 2019, from https://fsi.colostate.edu/avocados/ (https://fsi.colostate.edu/avocados/).

Hass Avocado Board. (n.d.). Retrieved October 14, 2019, from https://hassavocadoboard.com/ (https://hassavocadoboard.com/).

Khazan, O. (2015, June 13). The Selling of the Avocado. Retrieved October 14, 2019, from https://www.theatlantic.com/health/archive/2015/01/the-selling-of-the-avocado/385047/ (https://www.theatlantic.com/health/archive/2015/01/the-selling-of-the-avocado/385047/).