Assignment2

Charlotte MEYER, Yukta BUTALA,Sahaas HARI, Jonathan DAVID

Contents

To complete this assignment, follow these steps:

1. Download the `Assignment2.Rmd` file from Canvas.
2. Open `Assignment2.Rmd` in RStudio.
3. Replace the "Insert All Group Member Names Here" text in the `author:` field with your names.
4. Supply your solutions to the Assignment by editing `Assignment2.Rmd`.
5. Check how your codes work by running in the Console and knitting the codes (clicking `Knit`).
6. When you have completed the homework and have **checked** that your code both runs in the Console and knits correctly when you click `Knit HTML`, rename the R Markdown file to `Assignment2_Team#.Rmd` (e.g., `Assignment2_Team1.Rmd`), and submit on Canvas.

Load Data & libraries

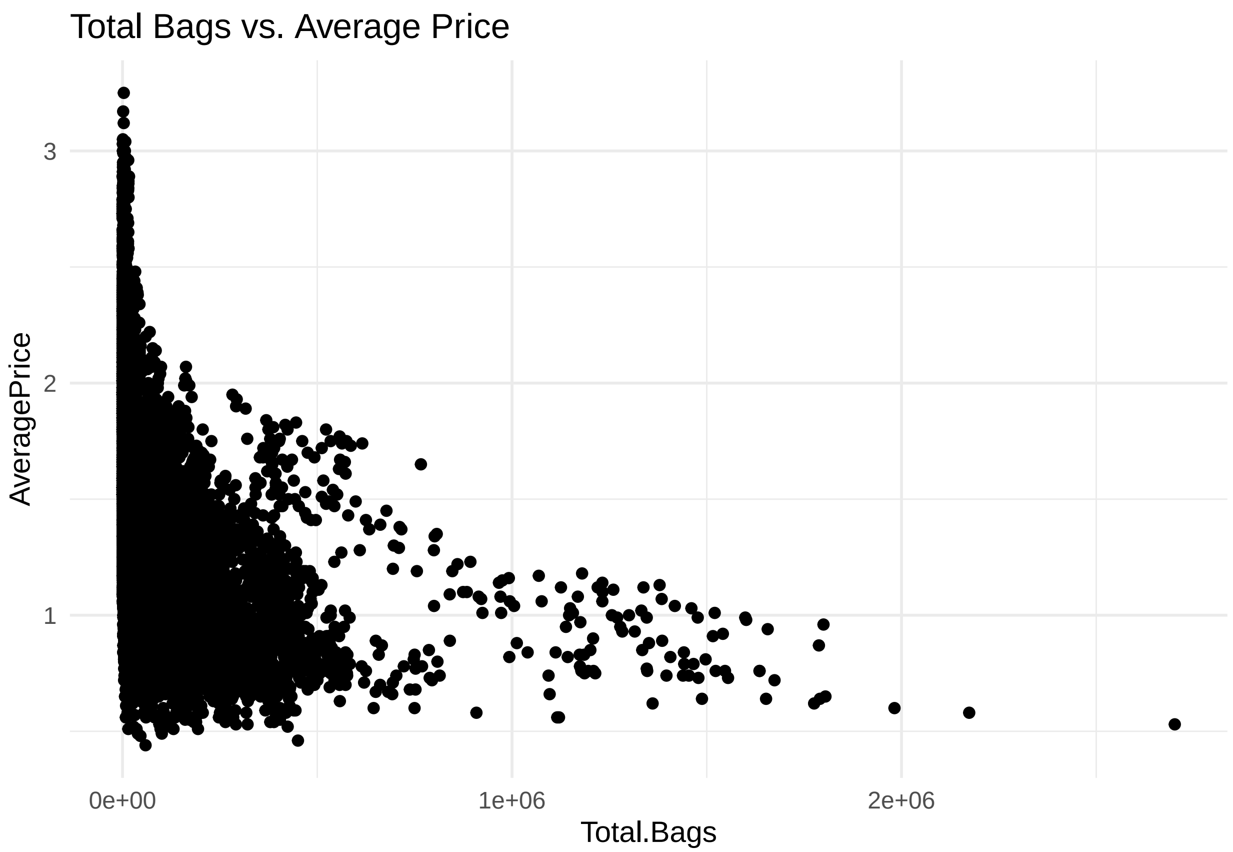
```
# Load the avocado_cleaned.csv file
avocado<-read.csv("/Users/charlottem/Library/Mobile Documents/com~apple~CloudDocs/CMU/Semester 2/Data Vi

# Load the dplyr library
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Load the ggplot2 library
library(ggplot2)

# Load the tidyr library
library(tidyr)
```

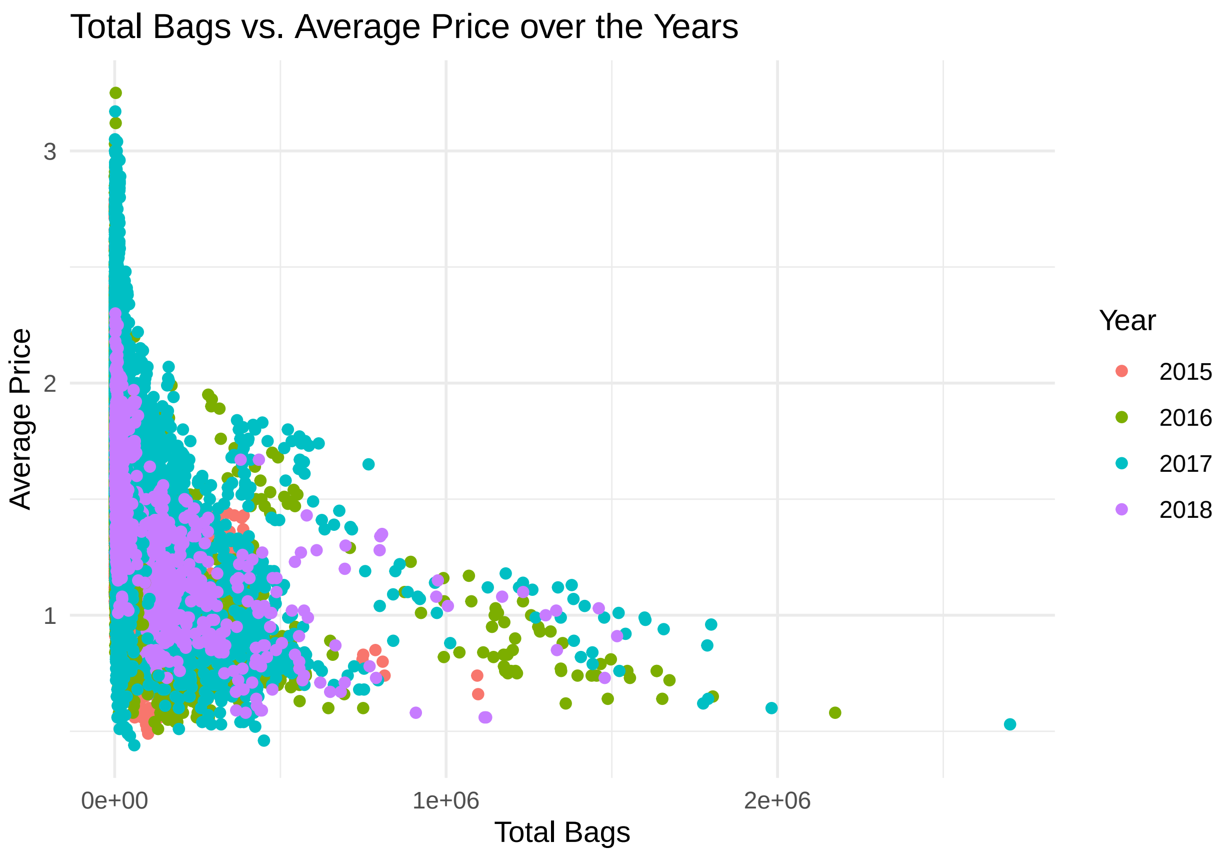
Q1. Scatter Plot

```r
# Plot the Total.Bags vs. AveragePrice on a scatter plot
# In this plot, x axis is Total.Bags and y axis is AveragePrice

ggplot(avocado,aes(x=Total.Bags,y=AveragePrice)) +
  geom_point() +
  labs(title = "Total Bags vs. Average Price",
       x = "Total.Bags",
       y = "AveragePrice") +
  theme_minimal()
```



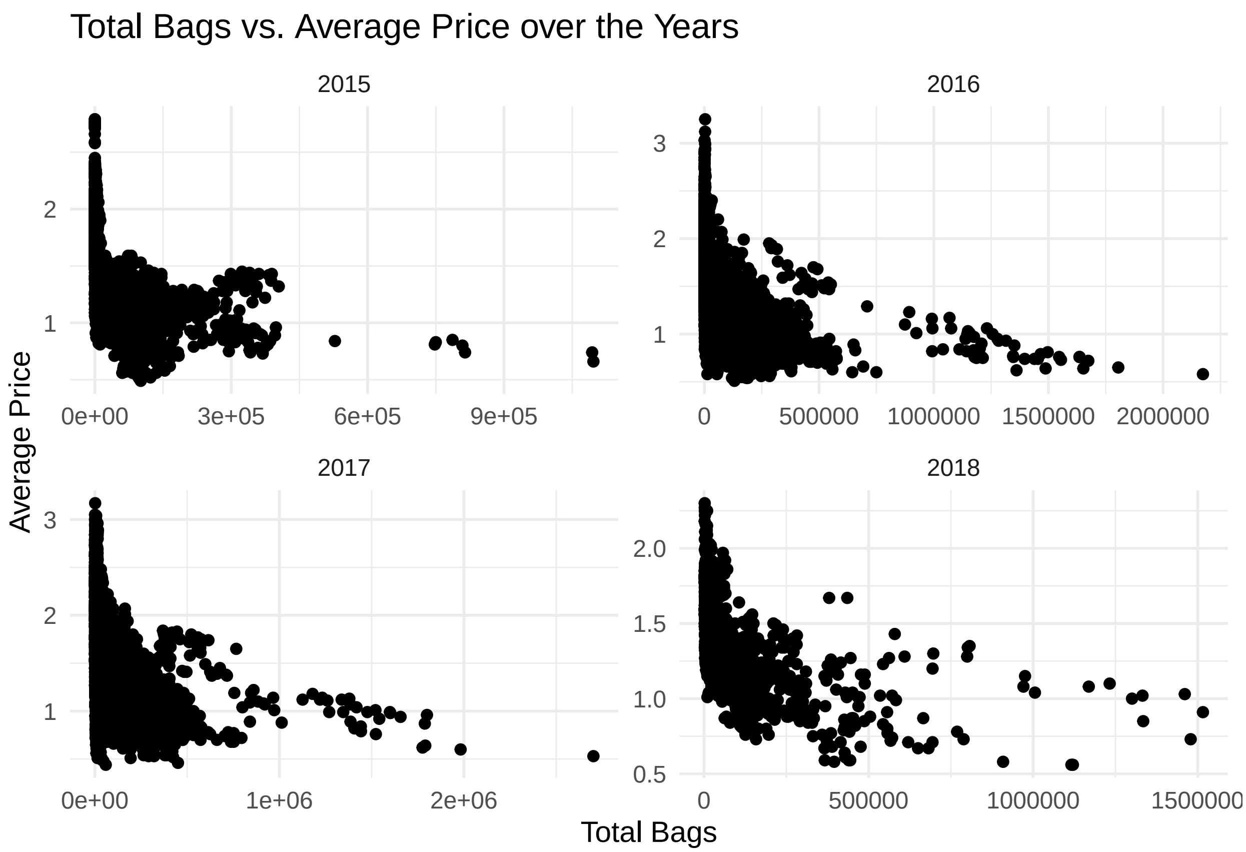
Q2. Scatter Plot with Third Variable

```r
# Plot Total.Bags vs. AveragePrice on a scatter plot and color the points by year
# In this plot, x axis is Total.Bags and y axis is AveragePrice

ggplot(avocado, aes(x = Total.Bags, y = AveragePrice, color = as.factor(year))) +
  geom_point() +
  labs(title = "Total Bags vs. Average Price over the Years",
       x = "Total Bags",
       y = "Average Price",
       color = "Year") +
  theme_minimal()
```



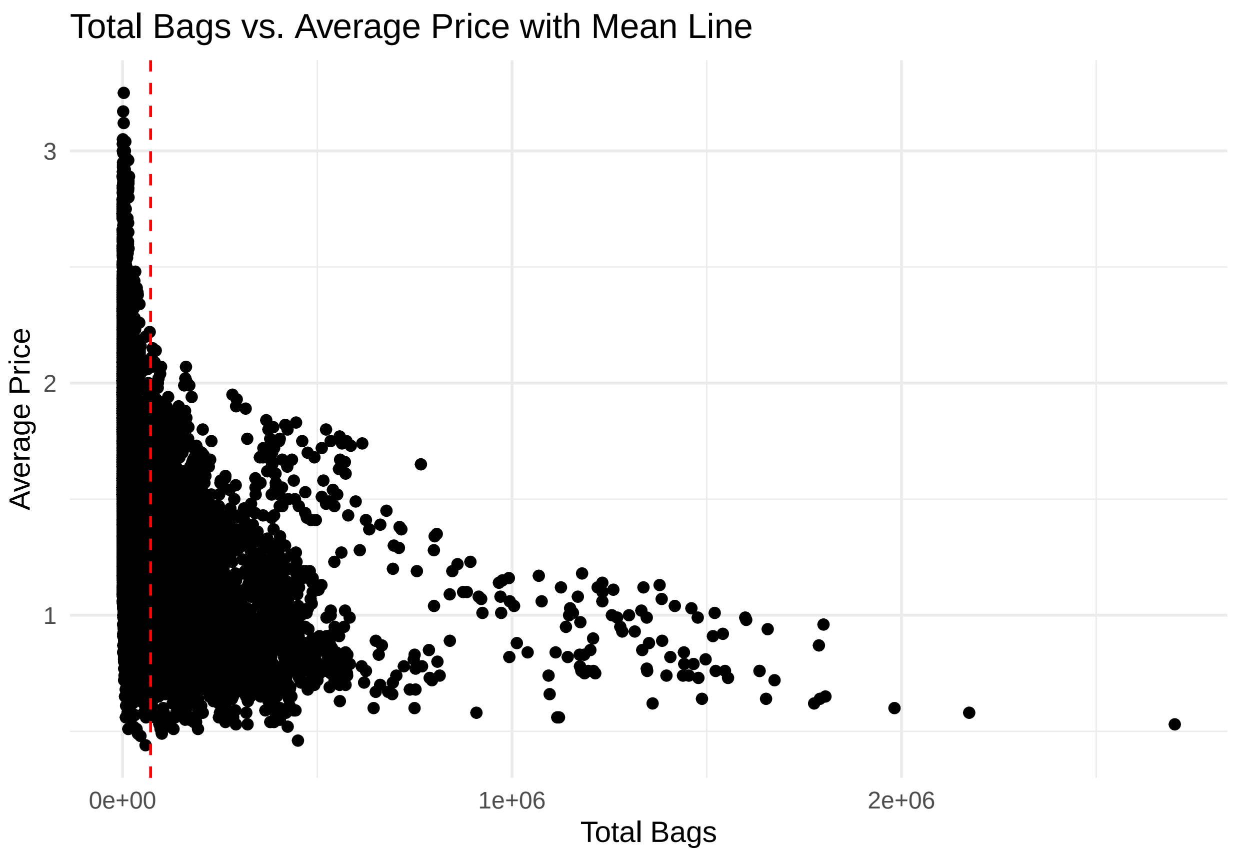
Q3. Multiple Scatter Plots using Facets

```
# Plot Total.Bags vs. AveragePrice on scatter plot for each year separately using facets
# In each sub-plot, x axis is Total.Bags and y axis is AveragePrice

ggplot(avocado, aes(x = Total.Bags, y = AveragePrice)) +
  geom_point() +
  labs(title = "Total Bags vs. Average Price over the Years",
       x = "Total Bags",
       y = "Average Price") +
  facet_wrap(~year, scales = "free") +
  theme_minimal()
```



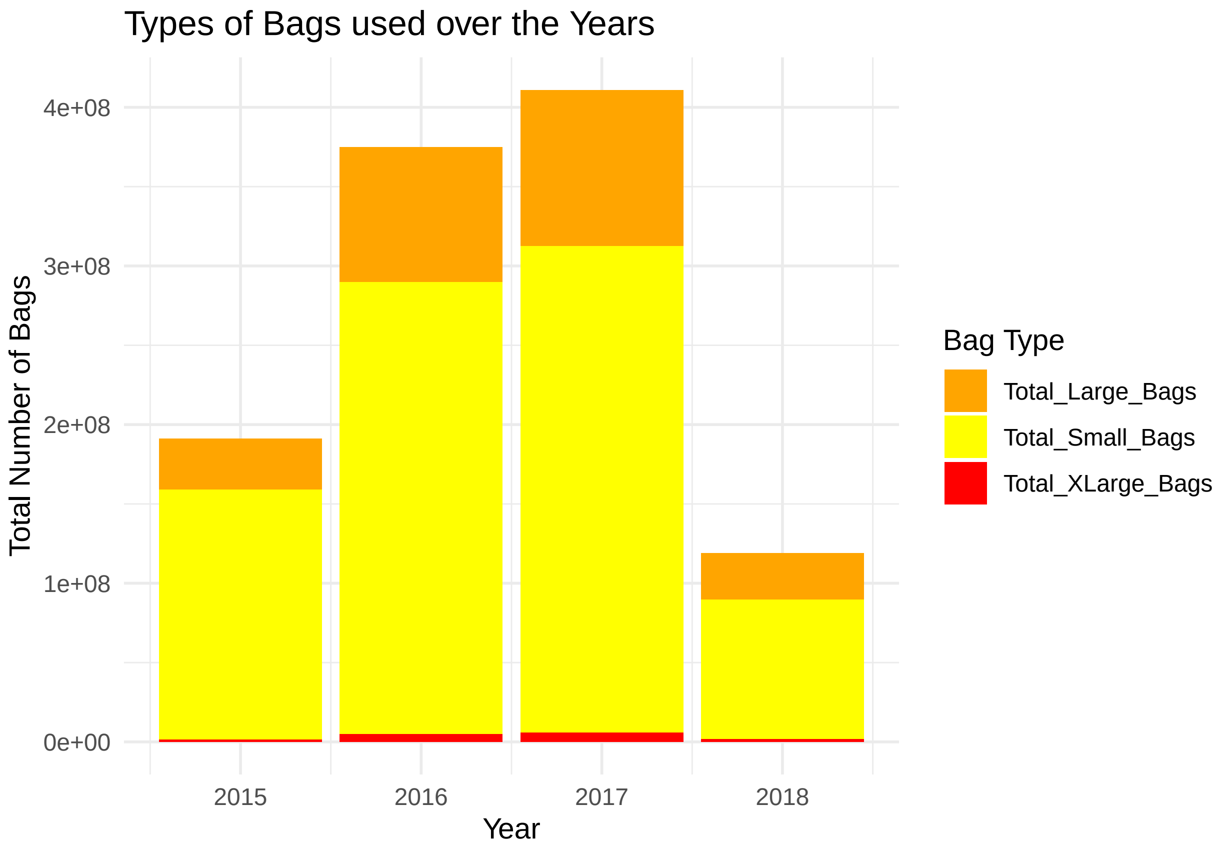
Q4. Scatter Plot with Reference Line Based on Summary Stats

```r
# Plot Total.Bags vs. AveragePrice on scatter plot. (x axis is Total.Bags and y axis is AveragePrice)
# Draw a vertical line corresponding to mean of Total.Bags

ggplot(avocado, aes(x = Total.Bags, y = AveragePrice)) +
  geom_point() +
  geom_vline(xintercept = mean(avocado$Total.Bags), linetype="dashed", color = "red")+
  labs(title = "Total Bags vs. Average Price with Mean Line",
       x = "Total Bags",
       y = "Average Price") +
  theme_minimal()
```



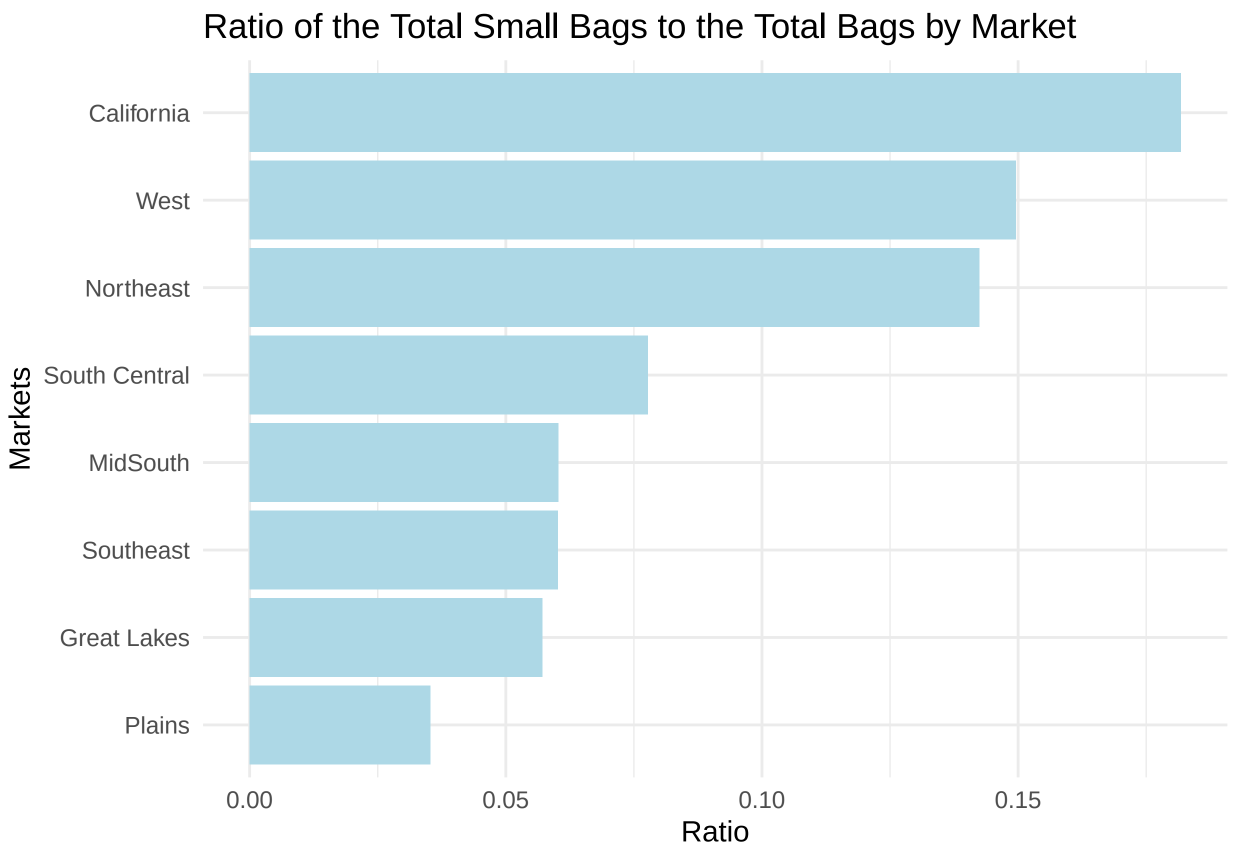
Q5. Bar Chart (* bonus question *)

```
# Plot a bar chart of the total number of Small.Bags, Large.Bags, XLarge.Bags over years.
# Tip: first, aggregate the data by year using group_by();
#      then, convert the aggregated data from wide to long using pivot_longer().
# Note: there are many other approaches to achieve the same goal. feel free to use them if you want.

avocado2 <- avocado %>%
  group_by(year) %>%
  summarise(Total_Small_Bags = sum(Small.Bags),
            Total_Large_Bags = sum(Large.Bags),
            Total_XLarge_Bags = sum(XLarge.Bags)) %>%
  pivot_longer(cols = c(Total_Small_Bags, Total_Large_Bags, Total_XLarge_Bags),
               names_to = "Bag_Type", values_to = "Total_Bags")

ggplot(avocado2,aes(x = year, y = Total_Bags, fill = Bag_Type))+
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Types of Bags used over the Years",
       x = "Year",
       y = "Total Number of Bags",
       fill = "Bag Type") +
  scale_fill_manual(values = c("Total_Small_Bags" = "yellow",
                               "Total_Large_Bags" = "orange",
                               "Total_XLarge_Bags" = "red")) +
  theme_minimal()
```



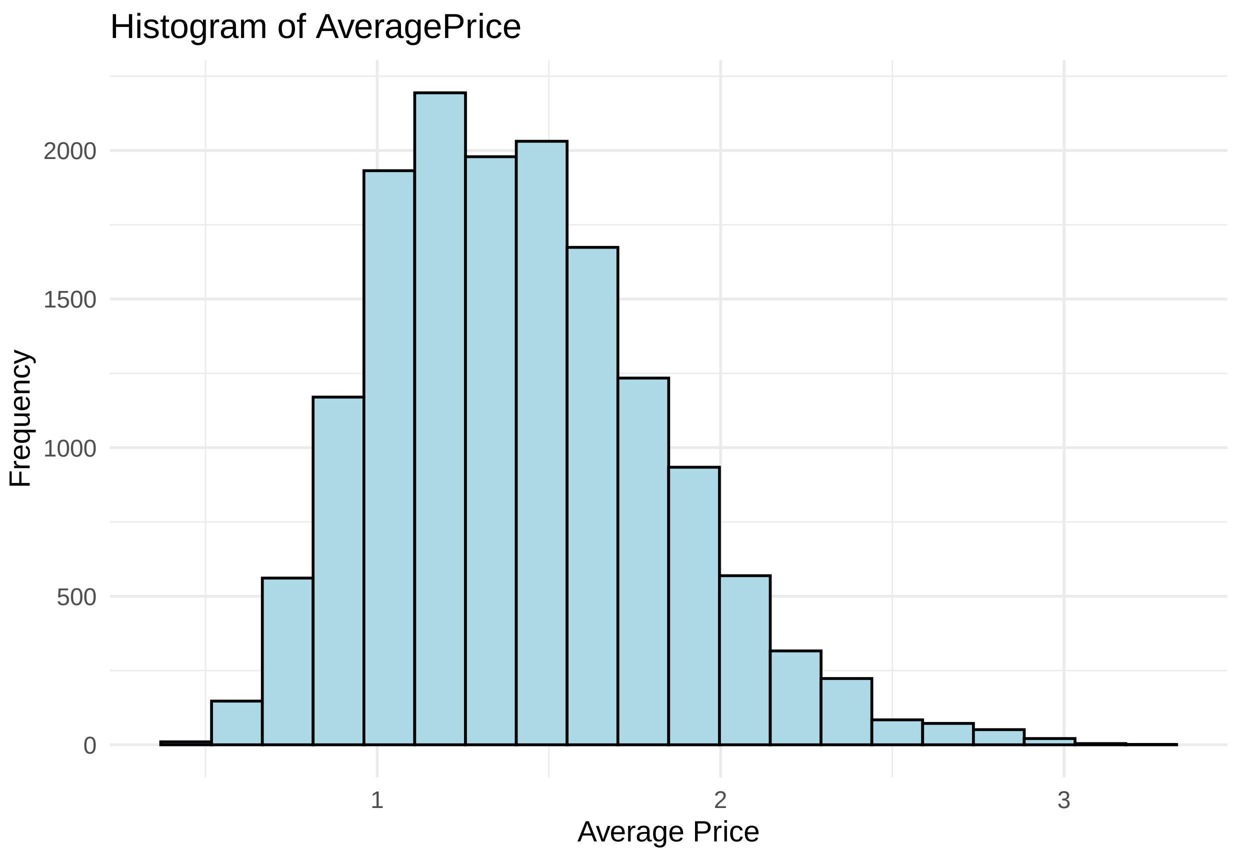
Q6. Horizontal Bar Chart

```
# Plot a horizontal bar chart to show the ratio of total Small.Bags to total Total.Bags for different M
# Sort Markets in descending order based on the Ratio.

avocado %>%
  group_by(Markets) %>%
  summarise(Total_Small_Bags = sum(Small.Bags)) %>%
  mutate(Ratio=Total_Small_Bags/sum(avocado$Total.Bags)) %>%
  arrange(desc(Markets)) %>%
  ggplot(aes(x = Ratio, y = reorder(Markets, Ratio))) +
  geom_bar(stat = "identity", fill = "light blue") +
  labs(title = "Ratio of the Total Small Bags to the Total Bags by Market",
       x = "Ratio",
       y = "Markets") +
  theme_minimal()
```



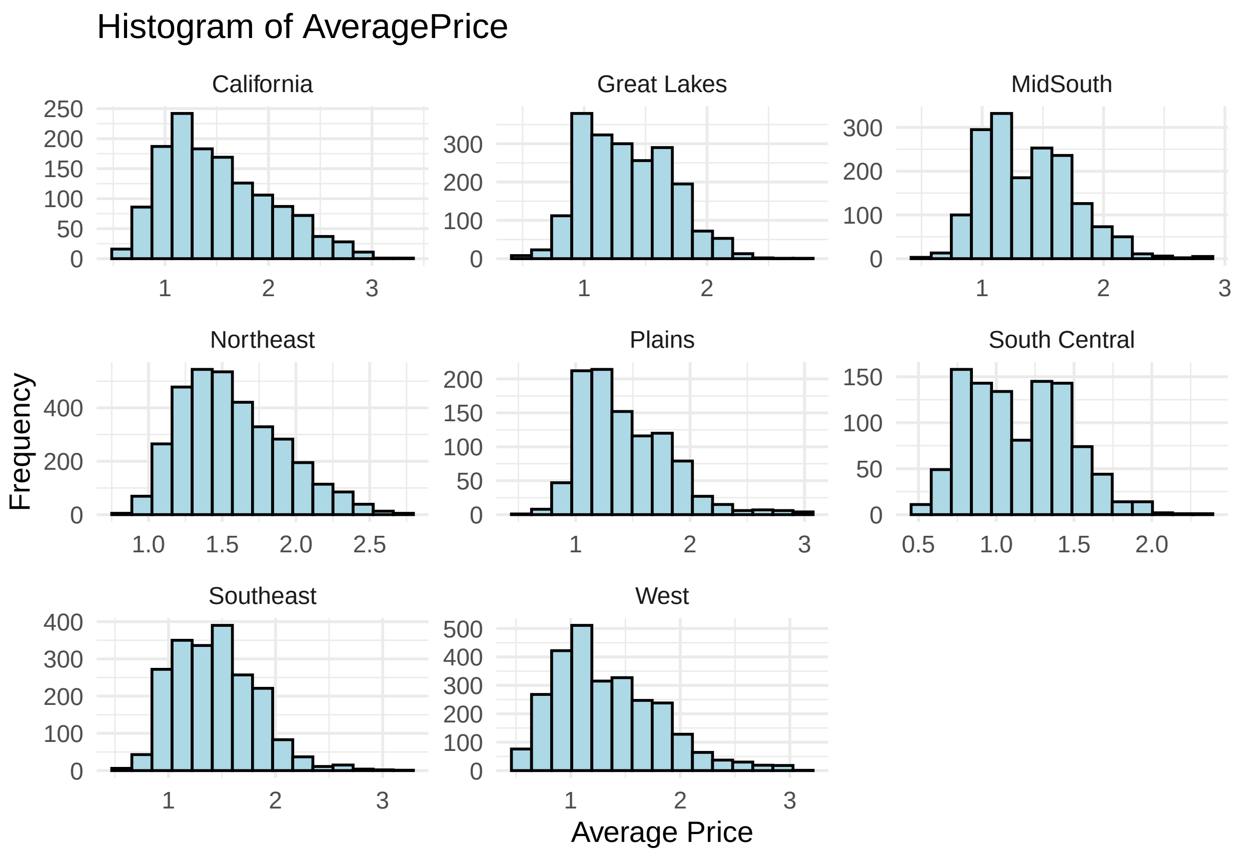
Q7. Basic Histogram

```r
# Draw a histogram of AveragePrice
ggplot(avocado, aes(x = AveragePrice)) +
  geom_histogram(bins=20, fill = "light blue", color = "black") +
  labs(title = "Histogram of AveragePrice",
       x = "Average Price",
       y = "Frequency") +
  theme_minimal()
```



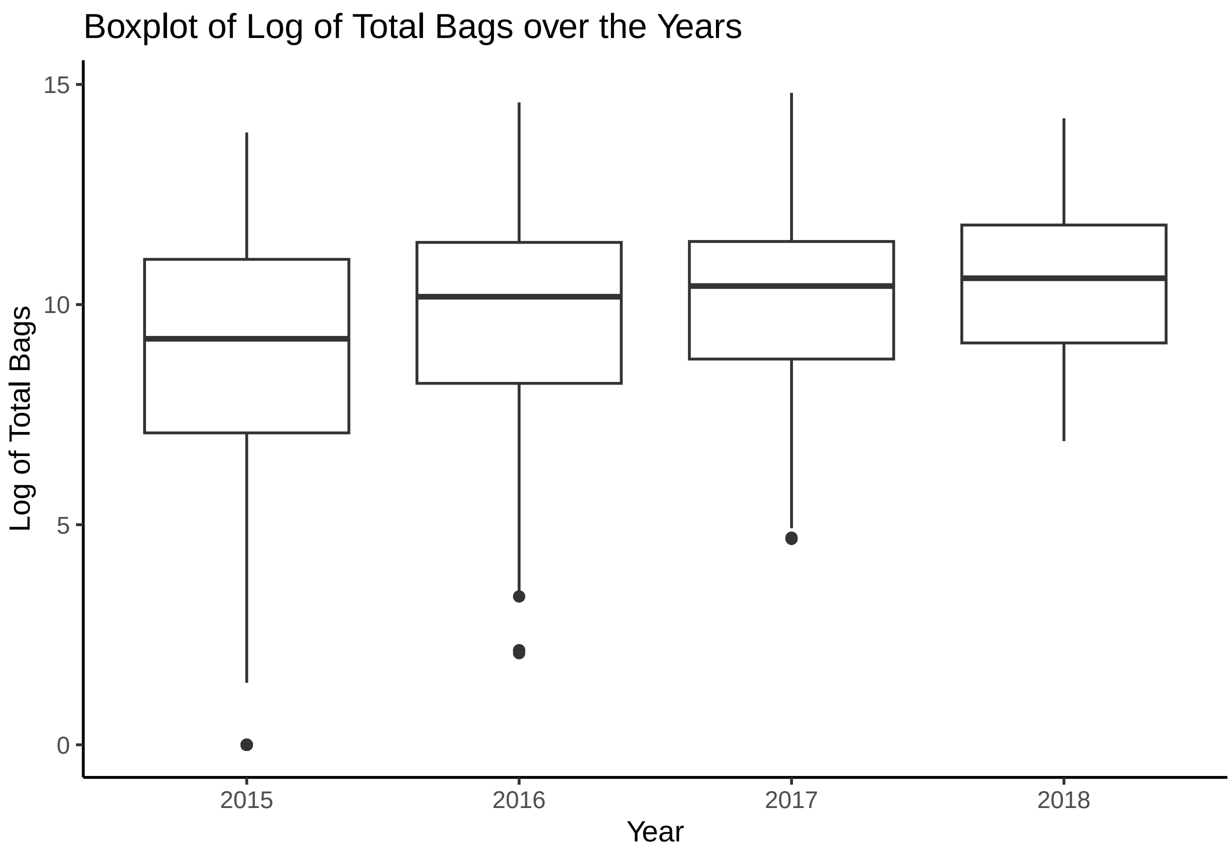
Q8. Multiple Histogram using Facets

```
# Draw histograms of AveragePrice by Markets using facets
ggplot(avocado, aes(x = AveragePrice)) +
  geom_histogram(bins=15, fill = "light blue", color = "black") +
  labs(title = "Histogram of AveragePrice",
       x = "Average Price",
       y = "Frequency") +
  theme_minimal() +
  facet_wrap(~Markets, scales = "free")
```



Q9. Boxplot

```
# Draw boxplots of log(Total.Bags+1) over Years
# note we are apply log() to transform large values, and we apply log(..+1) to avoid log(0)

avocado %>%
  mutate(log_Total_Bags = log(Total.Bags + 1)) %>%
  ggplot(aes(x = as.factor(year), y = log_Total_Bags)) +
  geom_boxplot() +
  labs(title = "Boxplot of Log of Total Bags over the Years",
       x = "Year",
       y = "Log of Total Bags") +
  theme_classic()
```


Boxplot of Log of Total Bags over the Years
Log of Total Bags
15
10
5
0
2015
2016
2017
2018
Year