Lab Assignment 2: Avocado Prices

Instructions

Submit your .html file to the assignment on PolyLearn.

Introduction

In this lab we're going to be looking at avocado prices! The dataset comes to us from kaggle and represents weekly retail scan data: avocado.csv. A description of the data can be found at the Hass Avocado Board website.

```r
library(tidyverse)
avo <- read.csv("https://www.dropbox.com/s/vsc1dkosz6nwake/avocado.csv?dl=1")
```

Exercises

1) Which region sold the most bags of small organic avocados in 2017?

Hint: TotalUS does not count as a region!

```r
avo %>%
  filter(type == "organic", region != "TotalUS", year == 2017) %>%
  group_by(region) %>%
  summarize(tot_small_bags = sum(Small.Bags)) %>%
  top_n(1, tot_small_bags)
```

```
## # A tibble: 1 x 2
##   region    tot_small_bags
##   <fct>              <dbl>
## 1 Northeast       2166706.
```

2) Use `separate()` to split the `Date` variable into year, month, and day. In which month is the highest volume of avocado sales?

```r
avo %>%
  filter(region == "TotalUS") %>%
  separate(Date, into = c("Year", "Month", "Day"), sep = "-") %>%
  group_by(Month) %>%
  summarize(tot_sales = sum(Total.Volume)) %>%
  top_n(1, tot_sales)
```

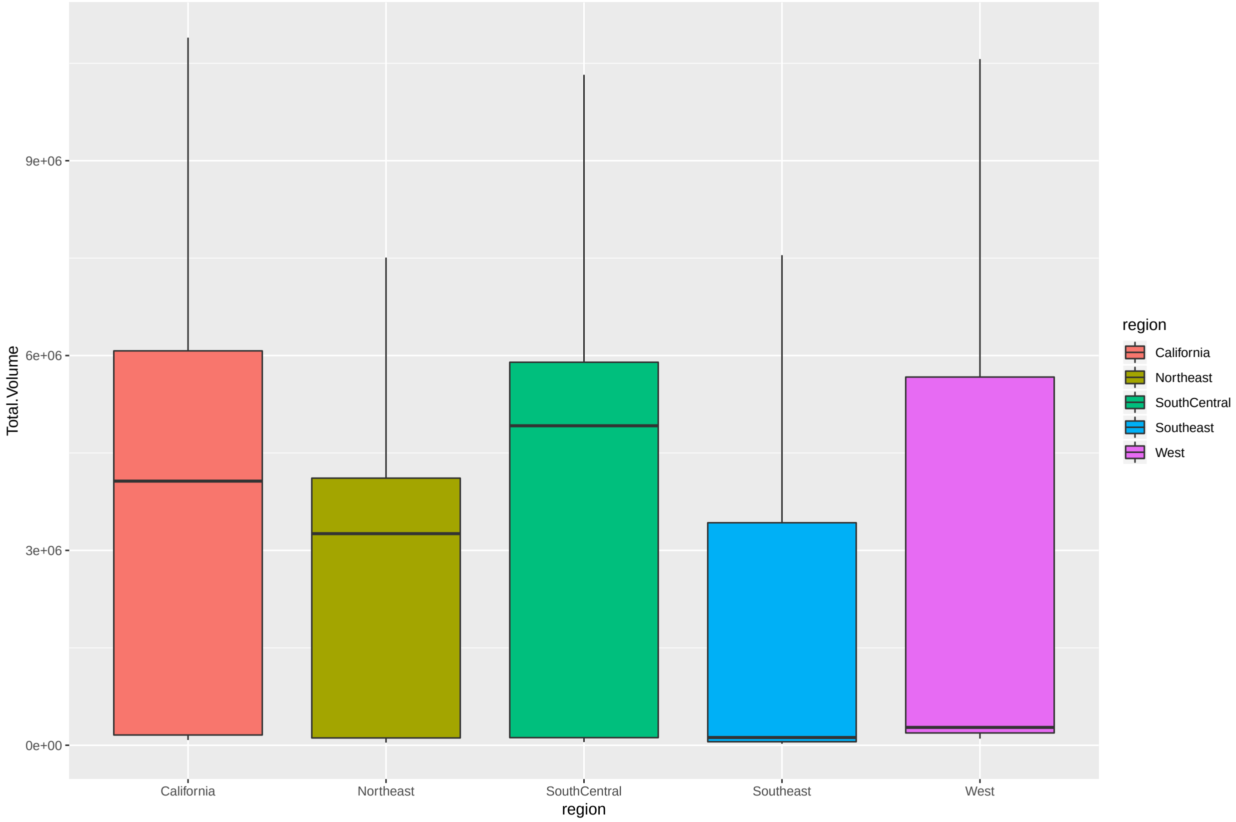
```
## # A tibble: 1 x 2
##   Month  tot_sales
##   <chr>      <dbl>
## 1 01    304528384.
```

3) Which regions sell the most avocados by volume? Plot side-by-side boxplots of Total Volume for only the 5 regions with the highest averages for the Total Volume variable.

Hint: Once you narrow down to the top 5 regions, you can use ***pull()*** *to save the vector of region names for later use.*

```
top_5 <- avo %>%
    filter(region != "TotalUS") %>%
    group_by(region) %>%
    summarize(avg_tot_vol = mean(Total.Volume)) %>%
    top_n(5, avg_tot_vol) %>%
    pull(region)

avo %>%
    filter(region %in% top_5) %>%
    ggplot(aes(x = region, y = Total.Volume, fill = region)) +
    geom_boxplot()
```



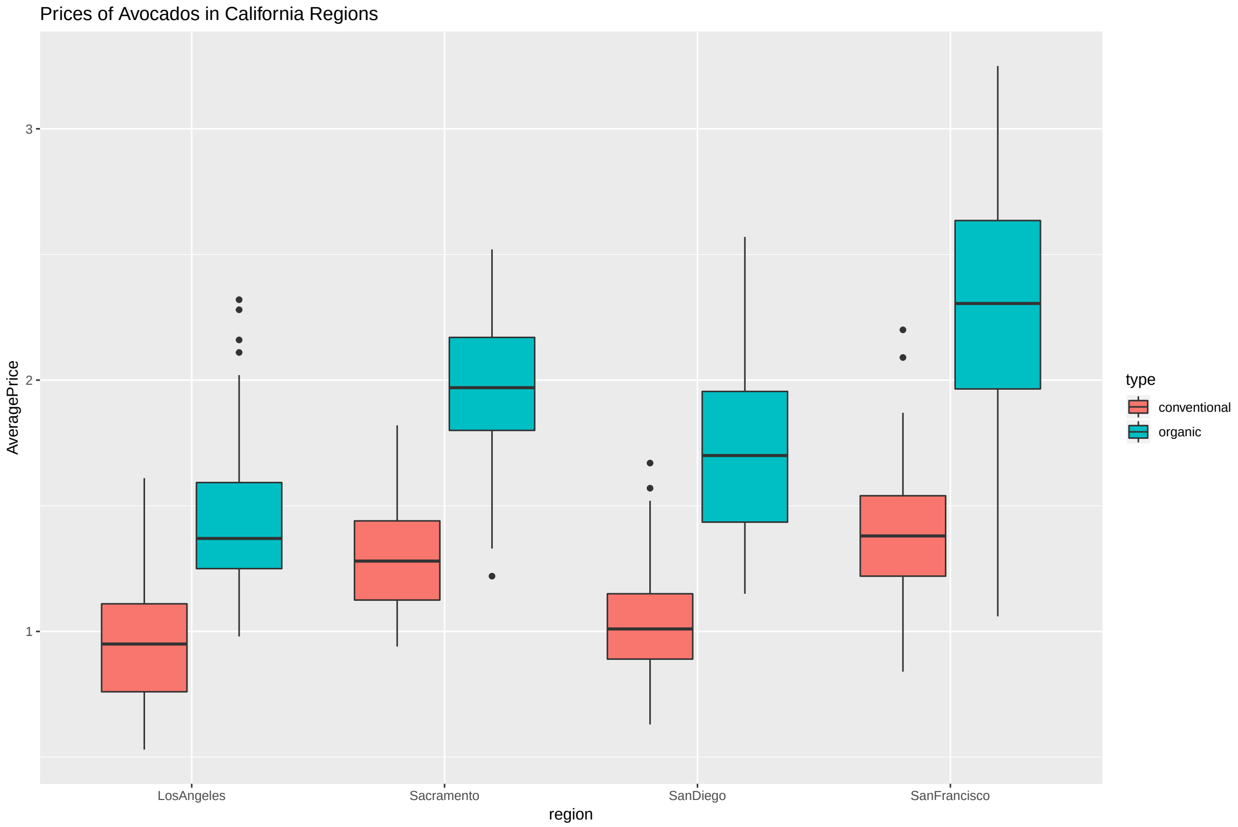
The following four California regions are in this dataset: LosAngeles, SanDiego, Sacramento, SanFrancisco. Answer the following questions about the California regions only.

Hint: These questions will require restructuring of your data!

4) In which regions is the price of organic versus conventional avocados most different? Support your answer with a few summary numbers and a plot.

```
cali <- avo %>% filter(region %in% c("LosAngeles", "SanDiego", "Sacramento", "SanFrancisco"))

cali %>%
  ggplot(aes(x = region, y = AveragePrice, fill = type)) +
  geom_boxplot() +
  ggtitle("Prices of Avocados in California Regions")
```

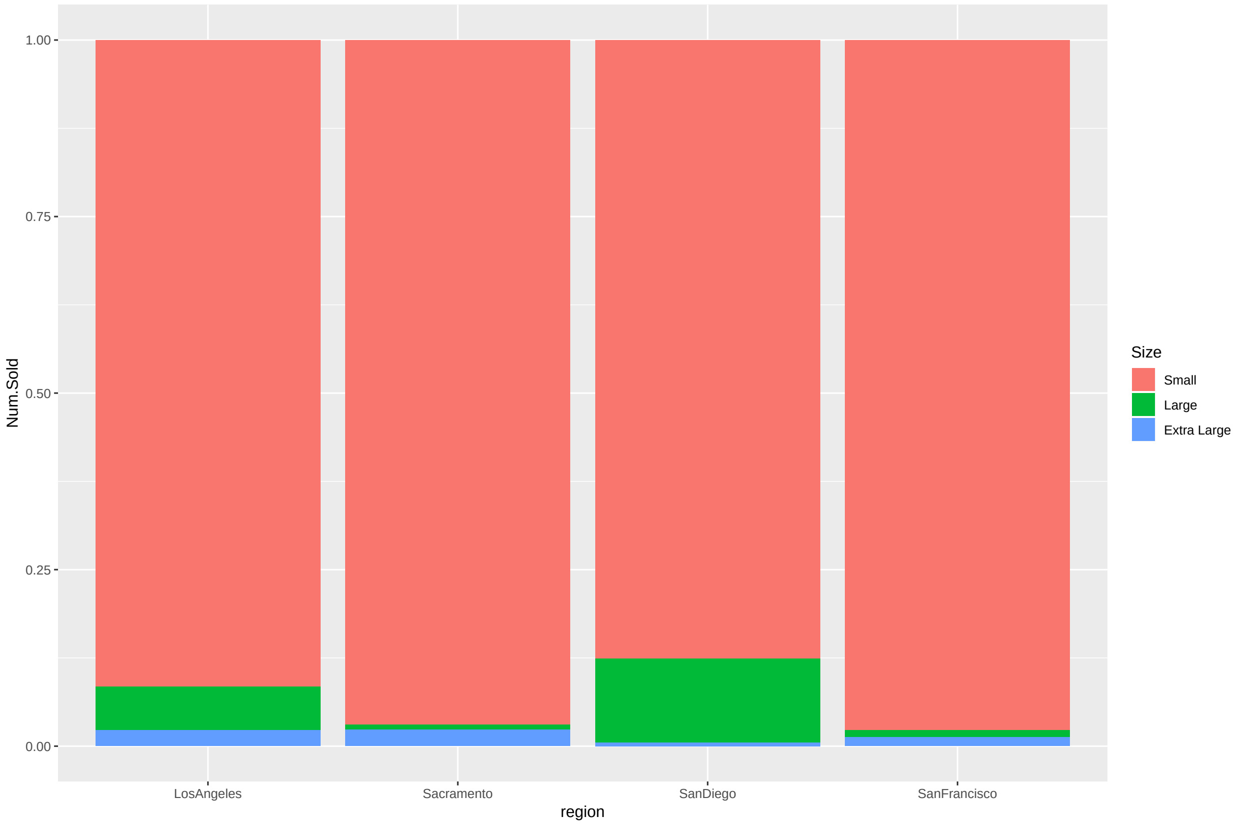


```
cali %>%
  group_by(region, type) %>%
  summarize(avg_price = mean(AveragePrice)) %>%
  spread(key = type, value = avg_price) %>%
  mutate(
    price_diff = organic - conventional
  )
```

```
## # A tibble: 4 x 4
## # Groups:   region [4]
##   region       conventional organic price_diff
##   <fct>               <dbl>   <dbl>      <dbl>
## 1 LosAngeles          0.960    1.46      0.502
## 2 Sacramento          1.28     1.97      0.688
## 3 SanDiego            1.04     1.72      0.685
## 4 SanFrancisco        1.39     2.25      0.850
```

5) How do their avocado habits differ? Make a plot that shows, for all 4 California regions, the percent of avocado sales that are small, large, or extra large. Separate your plot by conventional vs. organic avocados.

```
cali %>%
  group_by(region, type) %>%
  summarize_at(vars(Small.Bags, Large.Bags, XLarge.Bags), funs(mean)) %>%
  gather(key = Size, value = Num.Sold, -region, - type) %>%
  mutate(
    Size = factor(Size,
                  levels = c("Small.Bags", "Large.Bags", "XLarge.Bags"),
                  labels = c("Small", "Large", "Extra Large"))
  ) %>%
  ggplot(aes(x = region, y = Num.Sold, fill = Size)) +
  geom_col(position = "fill")
```


Num.Sold
LosAngeles
Sacramento
SanDiego
SanFrancisco
region
Size
Small
Large
Extra Large