Exploring Avocados

March 2, 2022

The Dataset

The data was pulled from https://www.kaggle.com/neuromusic/avocado-prices Some relevant columns in the dataset are shown below:

- Date - The date of the observation
- AveragePrice - the average price of a single avocado
- type - conventional or organic
- year - the year
- likes - Number of likes of the Talk
- Region - the city or region of the observation
- Total Volume - Total number of avocados sold

Many thanks to the Hass Avocado Board for sharing this data!!

1. What three regions have the highest mean Average Price for conventional avocados?
2. What three regions have the lowest mean total volume of conventional avocados sold?
3. Create a pairwise boxplot for the average price of organic avocados sold across the years. (Remember you will need to 'factor' years in your ggplot). Looking at the plot is there any significant difference in average price across years?
4. Visualize and describe the distribution for average price of organic avocados in New York. Be sure to include a thorough description.
5. Visualize the distribution for total volume of conventional avocados sold in Houston in the year 2015. Be sure to include a thorough description.
6. Visualize the average price of conventional avocados vs. organic avocados in California using a pairwise boxplot. Does there appear to be a significant in average price across the different types?
7. Visualize the total volume of avocados sold in Houston vs. California using a pairwise boxplot. Does there appear to be a significant difference in total volume across the two regions?

Solution:

```
    ## Load Packages
library(ggplot2)
library(tidyverse)

#### Load the dataset
```

```r
avo <- read.csv('/data/datasets/avocado.csv')
glimpse(avo)

#### Question 1 Find top three regions by average price conventional
avo %>% filter(type=='conventional')  %>%
  group_by(region) %>% summarise(average = mean(AveragePrice)) %>% arrange(desc(
    average))
  # Highest prices are in Hartford, NewYork, SanFran

#### Question 2 Find bottom three regions by total volume sold conventional
avo %>% filter(type=='conventional') %>%
  group_by(region) %>%
  summarise(average = mean(Total.Volume)) %>% arrange(average)
  # Lowest volume in Syracuse, Boise, and Spokane

#### Question 3 Pairwise boxplot for organic across years
avo %>% filter(type == 'organic') %>%
  ggplot(aes(y=AveragePrice,fill=factor(year))) +
  geom_boxplot() +
  theme_classic()

#### Question 4  Average price of organic avocados in New York
avo %>% filter(type == 'organic' & region == 'NewYork') %>%
  ggplot(aes(x=AveragePrice)) +
  geom_histogram(fill='green', col='black')+
  theme_classic()+
  labs(x='Average price of avocados', title='Distribution for the average price of
    avocados')

#### Question 5 Total volume for conventional avocados in Houston 2015
avo %>% filter(region == 'Houston' & type == 'conventional' & year==2015) %>%
  ggplot(aes(x=Total.Volume))+
  geom_histogram(fill='darkgreen', col='black')+
  labs(x='Total Volume Sold')

#### Question 6 Average price for conventional vs. organic avocadoes in California
avo %>% filter(region=='California') %>%
  ggplot(aes(y=AveragePrice,fill=type)) +
  geom_boxplot()

#### Question 7  Total volume of organic sold in Houston vs. California
avo %>% filter(type == 'organic') %>%
  filter(region == 'Houston' | region == 'California') %>%
  ggplot(aes(y=Total.Volume, fill=region)) +
  geom_boxplot() +
  labs(y='Total volume sold')+
  theme_classic()


```