CAAP Statistics
Final Project Overview

Jul 21, 2022

Overview

- **Goal: Hands-on application of statistical concepts you learned from lecture to the actual dataset!**
  - Please apply as many the statistical concepts as possible that you learned from the lecture to the actual dataset.
  - Describe how the concepts give you an insight about the dataset
  - Try to visualize your insight and explain to the audience

Step1 - Find a dataset of interest

- Sources
  - Openintro
  - Kaggle
  - UCI Machine Learning Repository
- Specify the reason why you choose the data — why would that dataset be interesting?
- Encourage to find the dataset having more numerical variables; at this moment, it is easier to deal with numerical variables than categorical one.

Step1 - Find a dataset of interest

- Avocado dataset from Kaggle
- Why? I love avocado so much!
- I want to know the factors that might potentially effect the avocado pricing—size, region grown, whether it is organic or not?

Data

https://www.kaggle.com/datasets/neuromusic/avocado-prices

```
   avo = read.csv("avocado.csv")
   avo= avo[,-1]
   head(avo)
```

```
##         Date AveragePrice Total.Volume   X4046     X4225  X4770 Total.Bags
## 1 2015-12-27         1.33     64236.62 1036.74  54454.85  48.16    8696.87
## 2 2015-12-20         1.35     54876.98  674.28  44638.81  58.33    9505.56
## 3 2015-12-13         0.93    118220.22  794.70 109149.67 130.50    8145.35
## 4 2015-12-06         1.08     78992.15 1132.00  71976.41  72.58    5811.16
## 5 2015-11-29         1.28     51039.60  941.48  43838.39  75.78    6183.95
## 6 2015-11-22         1.26     55979.78 1184.27  48067.99  43.61    6683.91
##   Small.Bags Large.Bags XLarge.Bags         type year region
## 1    8603.62      93.25           0 conventional 2015 Albany
## 2    9408.07      97.49           0 conventional 2015 Albany
## 3    8042.21     103.14           0 conventional 2015 Albany
```

Step2 - Work on summary statistics

- Try to understand the data
- Try to use the concepts we have learned from the lecture to do exploratory data anlysis:
  - Numerical summary: mean, median, variance
  - Graphical summary
    - boxplot, histogram, scatterplot
    - barplot, mosaic plot

Data - variables

HASS

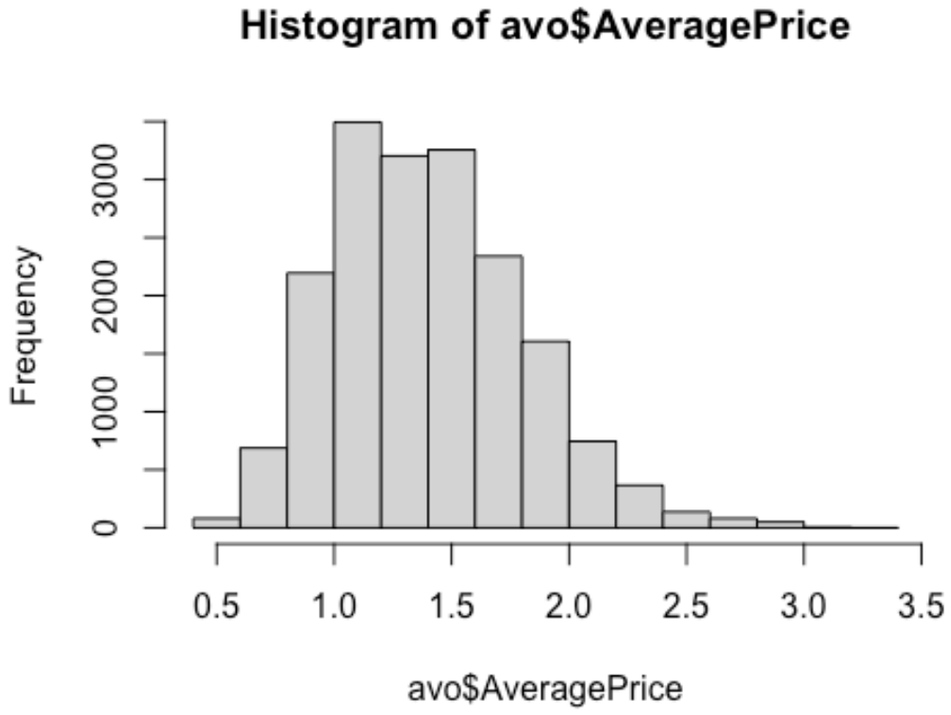
Hass, Small
#60 size and smaller

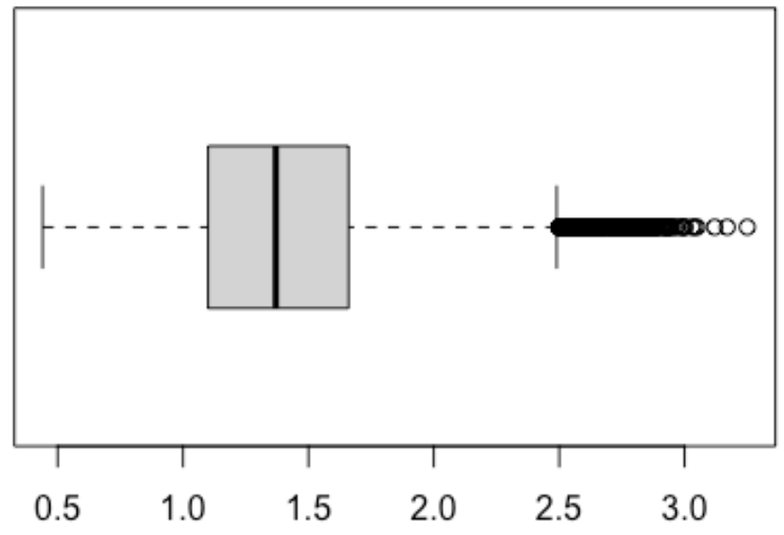
Hass, Large
#40 & #48 sizes

Hass, Extra Large
#36 and larger

- `Date` : The date of the observation
- `AveragePrice` : the average price of a single avocado
- `TotalVolume` : Total number of avocados sold
- `X4046` : Total number of avocados with PLU 4046 sold
- `X4225` : Total number of avocados with PLU 4225 sold
- `X4770` : Total number of avocados with PLU 4770 sold
- `Total.Bags`
- `Small.Bags`
- `Large.Bags`
- `XLarge.Bags`
- `type`: conventional vs organic
- `year`

Step2 - Work on summary statistics





Step2 - Work on summary statistics

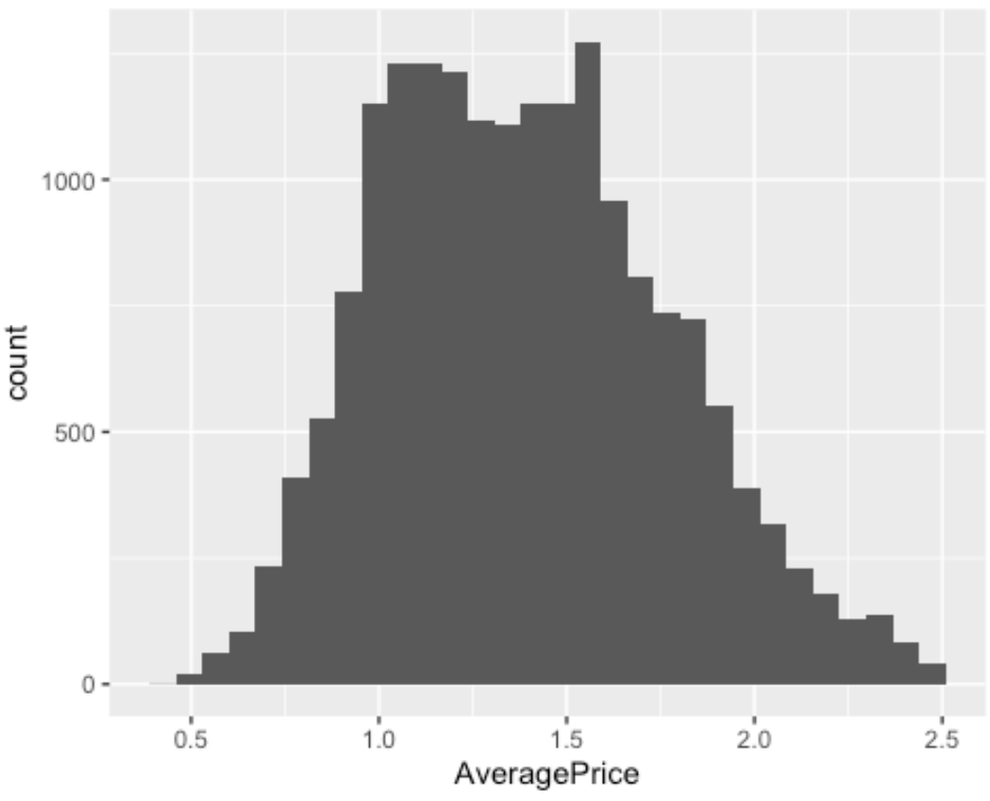
```
summary(avo$AveragePrice)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.440   1.100   1.370   1.406   1.660   3.250
mean(avo$AveragePrice)
## [1] 1.405978
sd(avo$AveragePrice)
## [1] 0.4026766
```

Step3 - Observe any trends, outliers

- At this step, you may find interesting questions that can be answered by the data
- Based on the summary at step2, determine if there is any association between the variables
- If there are outliers, try to find the reason for the outliers — input error or true outliers?

Step3 - Observe any trends, outliers

```
avo %>%
  filter(AveragePrice < 2.5)%>%
  ggplot(aes(x=AveragePrice))+
  geom_histogram()
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

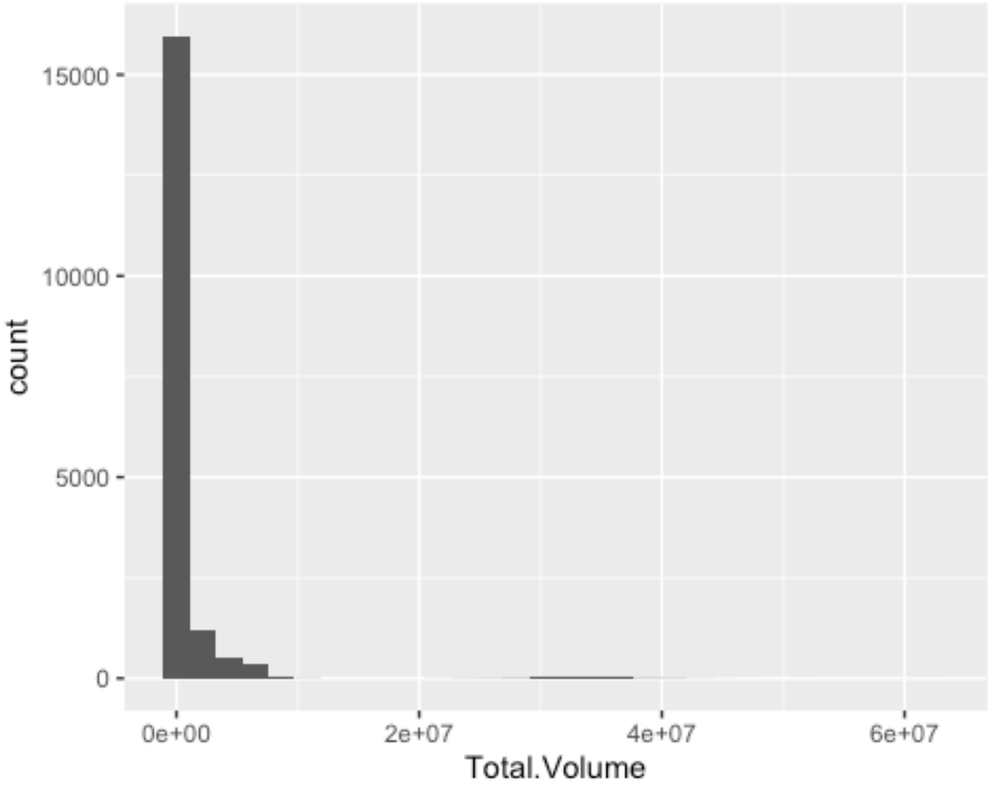


Step3 - Observe any trends, outliers

```
avo %>%
  filter(AveragePrice>=2.5)%>%
  group_by(region)%>%
  summarise(count =n(), av = mean(Total.Volume))%>%
  arrange(desc(count))
## # A tibble: 24 × 3
##    region              count     av
##    <chr>               <int>  <dbl>
##  1 SanFrancisco           54 18210.
##  2 Spokane                17  2667.
##  3 Seattle                16 26129.
##  4 HartfordSpringfield    15 11222.
##  5 RaleighGreensboro      14 11314.
##  6 Charlotte              10  8980.
##  7 Portland               10 14788.
##  8 Boise                   9  1584.
##  9 Sacramento              8  5403.
## 10 Atlanta                 7 15414.
## # … with 14 more rows
```

Step3 - Observe any trends, outliers

```
avo %>%
  ggplot(aes(x=Total.Volume))+
  geom_histogram()
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



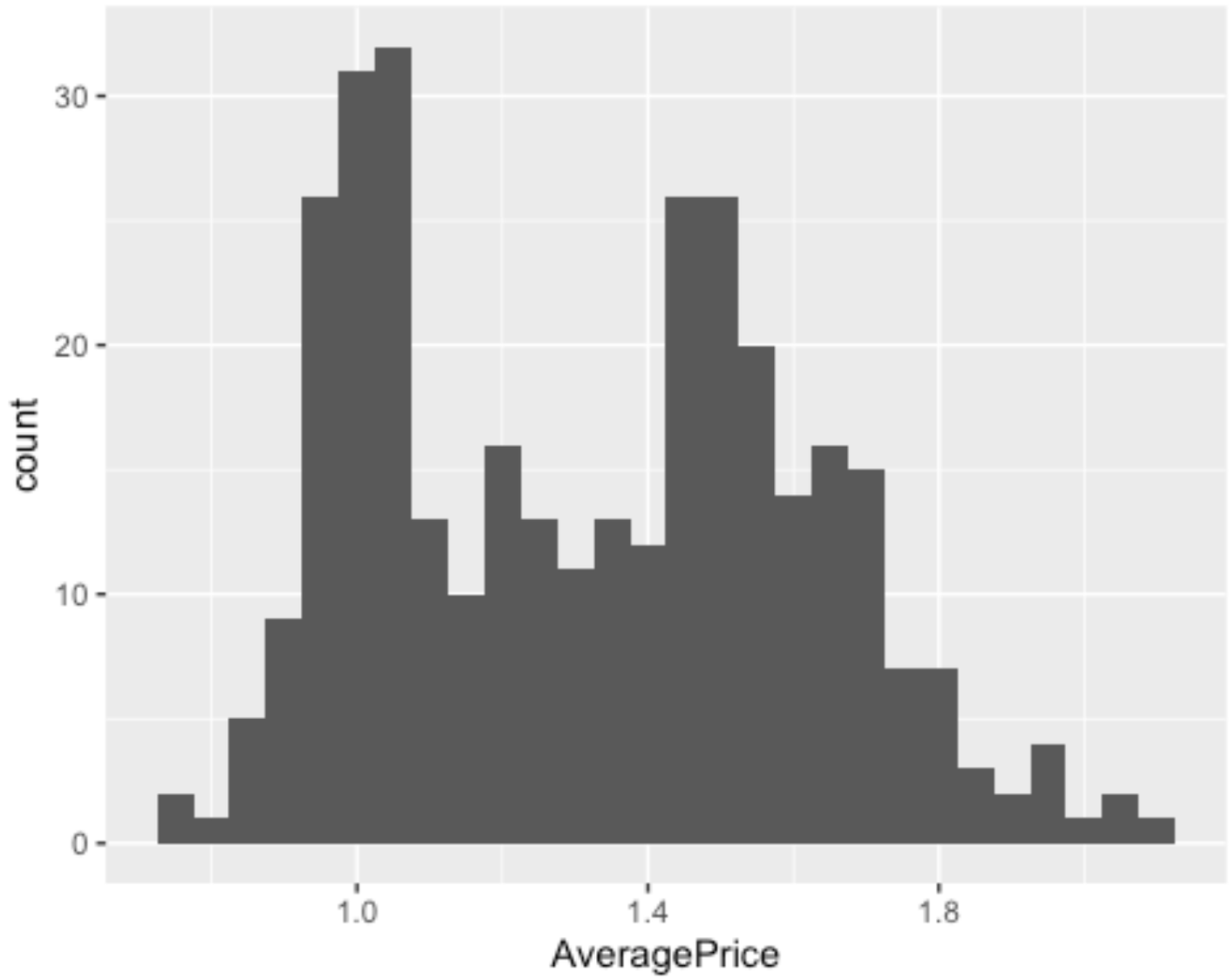
Step3 - Observe any trends, outliers

```
avo%>%
  arrange(desc(Total.Volume))%>%
  head(n = 5)
```

```
##         Date AveragePrice Total.Volume    X4046    X4225     X4770 Total.Bags
## 1 2018-02-04         0.87     62505647 21620181 20445501 1066830.2   19373134
## 2 2017-02-05         0.77     61034457 22743616 20328162 1664383.1   16298296
## 3 2016-02-07         0.76     52288698 16573574 20470573 2546439.1   12698112
## 4 2017-05-07         1.09     47293922 17076651 13549103  863471.9   15804696
## 5 2016-05-08         0.82     46324530 14223305 17896392 1993645.4   12211188
##   Small.Bags Large.Bags XLarge.Bags         type year  region
## 1   13384587    5719097    269451.0 conventional 2018 TotalUS
## 2   12567156    3618271    112870.0 conventional 2017 TotalUS
## 3    9083373    3373078    241661.5 conventional 2016 TotalUS
## 4   11228050    4324231    252415.5 conventional 2017 TotalUS
## 5    8747757    3342781    120650.1 conventional 2016 TotalUS
```

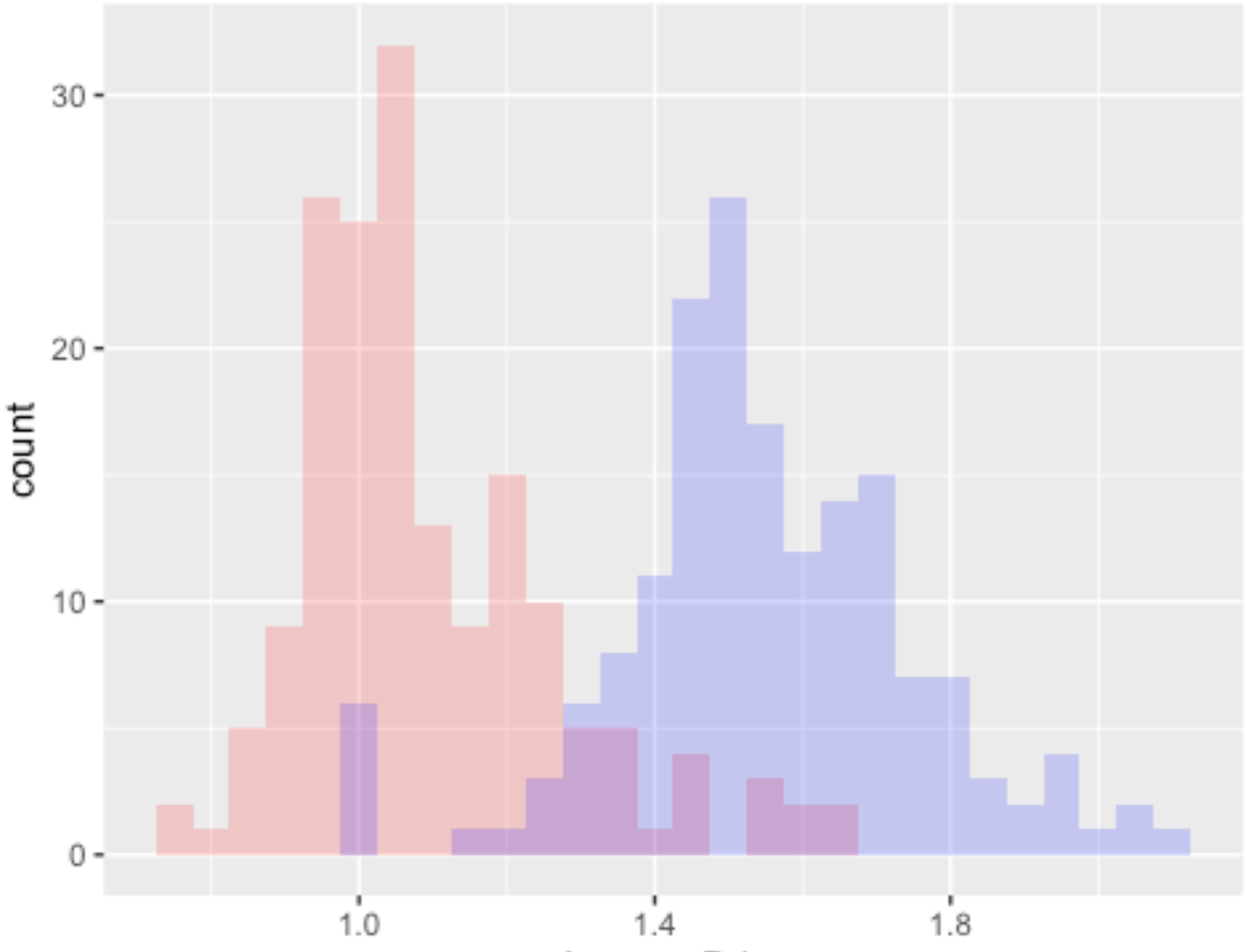
Step3 - Observe any trends, outliers

```
avo %>%
  filter(region == "TotalUS")%>%
  ggplot(aes(x=AveragePrice))+
  geom_histogram(binwidth = 0.05)
```



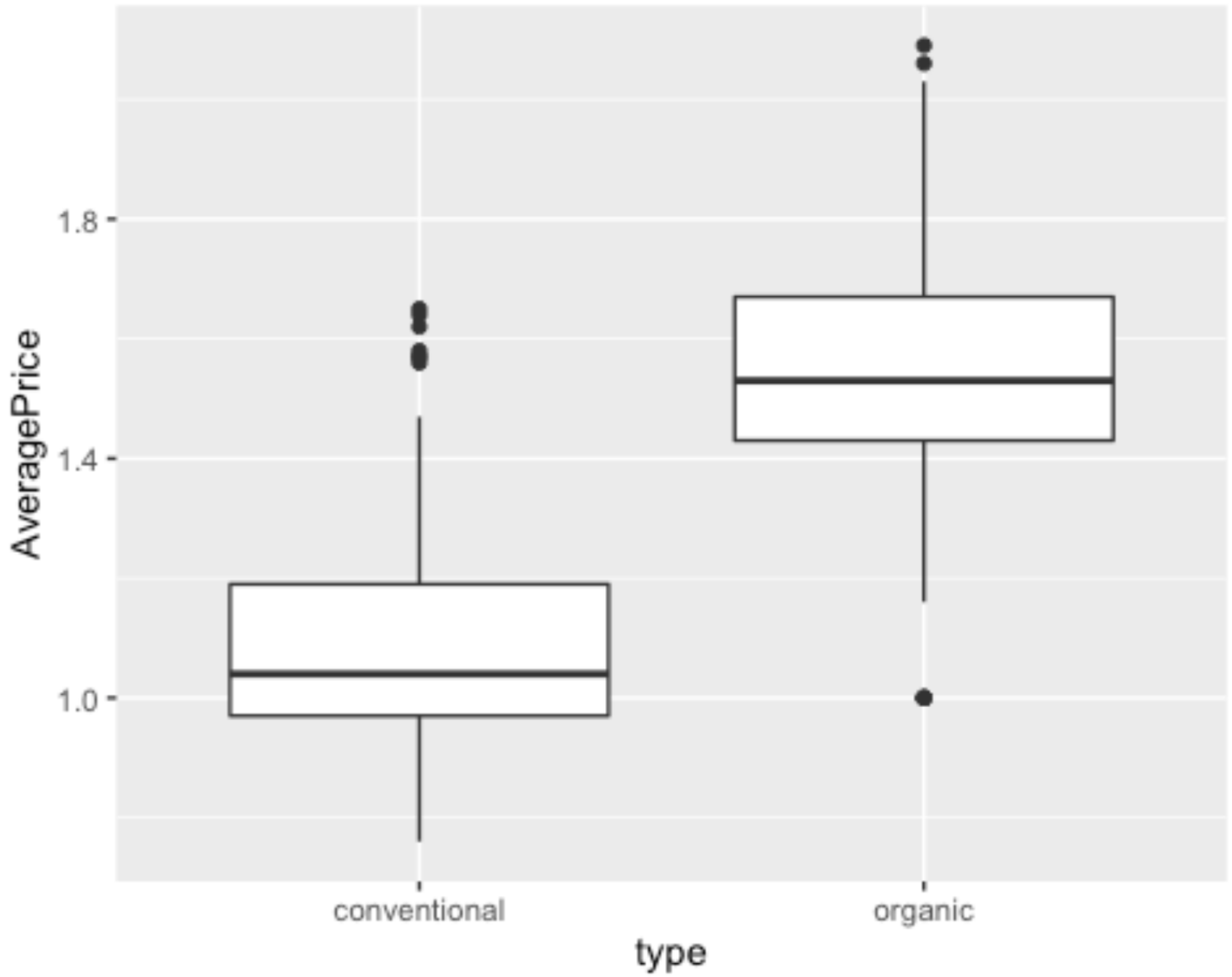
Step3- Observe the trend, outliers

```
avo %>%
  filter(region == "TotalUS")%>%
  ggplot(aes(x=AveragePrice))+
  geom_histogram(data=subset(avo,region=="TotalUS" & type == "conventional"),fill =
"red", alpha = 0.2, binwidth = 0.05) +
  geom_histogram(data=subset(avo,region=="TotalUS" & type == "organic"),fill = "blue",
alpha = 0.2, binwidth = 0.05)
```



Step3- Observe the trend, outliers

```
avo %>%
  filter(region=="TotalUS")%>%
  ggplot(aes(x=type, y =AveragePrice))+
  geom_boxplot()
```



Step4 - Test a hypothesis of interest

- Specify the hypothesis that you want to test
- Try both ways:
  - Randomized simulation
  - Hypothesis testing framework using test statistics
- Interpret the result and persuade the audience why this finding is interesting or appealing

Step5 - Visualize and Present

- When it comes to real world, delivering the result in appealing way is as important as the data analysis itself.
- Try to show your data analysis procedure thoroughly with the code attached
- Please clearly show the reason why you choose the particular statistical concepts/ method to demonstrate your analysis