INT303 Final project

Yuxuan Wu 1716309

December 1, 2020

Import the libraries

```
library(tidyr)
library(skimr)
library(GGally)
library(viridis)
library(caret)
library(e1071)
library(rpart)
library(xgboost)
library(corrplot)
library(corrgram)
library(ggplot2)
library(ggthemes)
library(psych)
library(scales)
library(treemap)
library(repr)
library(cowplot)
library(magrittr)
library(ggpubr)
library(RColorBrewer)
library(plotrix)
library(ggrepel)
library(tidyverse)
library(gridExtra)
library(lubridate)
library(tibbletime)
library(reshape2)
```

Load the data and return the head of data

```
df <- read.csv("/Users/yuxuan/Desktop/INT301-Avocado-prediction/avocado-updated-2020.csv")
head(df)
```

```
##         date average_price total_volume     X4046     X4225    X4770 total_bags
## 1 2015-01-04          1.22     40873.28   2819.50  28287.42    49.90    9716.46
## 2 2015-01-04          1.79      1373.95     57.42    153.88     0.00    1162.65
## 3 2015-01-04          1.00    435021.49 364302.39  23821.16    82.15   46815.79
## 4 2015-01-04          1.76      3846.69   1500.15    938.35     0.00    1408.19
## 5 2015-01-04          1.08    788025.06  53987.31 552906.04 39995.03  141136.68
## 6 2015-01-04          1.29     19137.28   8040.64   6557.47   657.48    3881.69
##   small_bags large_bags xlarge_bags         type year           geography
## 1    9186.93     529.53           0 conventional 2015               Albany
## 2    1162.65       0.00           0      organic 2015               Albany
## 3   16707.15   30108.64           0 conventional 2015              Atlanta
## 4    1071.35     336.84           0      organic 2015              Atlanta
## 5  137146.07    3990.61           0 conventional 2015 Baltimore/Washington
## 6    3881.69       0.00           0      organic 2015 Baltimore/Washington
```

Explore the data and some clarification

Explain the features

- date - The date of the observation
- average_price - The average price of a single
- total_volume - Total number of avocados sold
- year - The year
- type - conventional or organic
- geography - The city or region of the observation

X4046, X4225, X4770 stands for the PLU code

- Small/Medium Hass Avocado (~3-5oz avocado) | #4046
- Large Hass Avocado (~8-10oz avocado) | #4225
- Extra Large Hass Avocado (~10-15oz avocado) | #4770

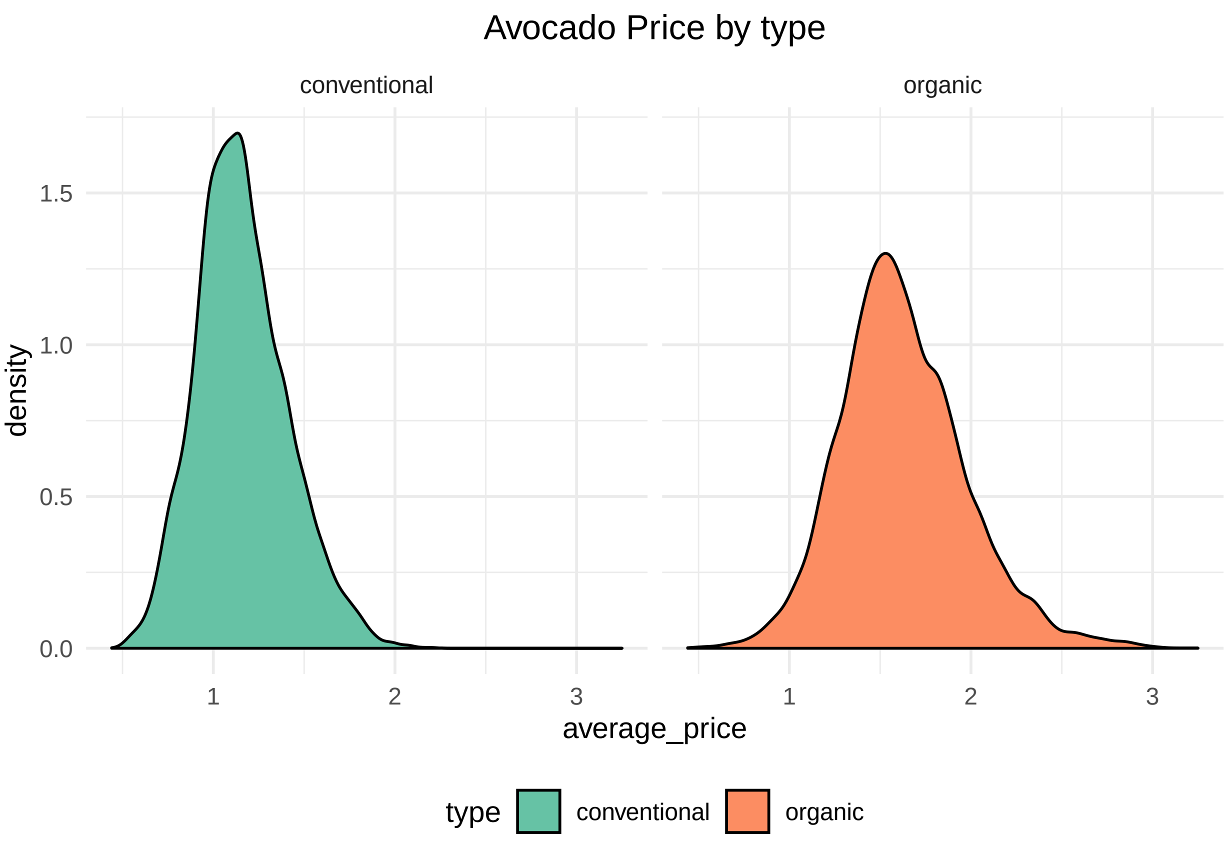
Exploratory Data Analysis

```r
levels(df$type)
```

Density plot of the difference between two avocadoes.

```
## [1] "conventional" "organic"
```

```r
library(ggplot2)
options(repr.plot.width = 8, repr.plot.height = 4)
ggplot(df, aes(x=average_price, fill=type))+
  geom_density()+
  facet_wrap(~type)+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5),legend.position = "bottom")+
  labs(title = "Avocado Price by type")+
  scale_fill_brewer(palette = "Set2")
```



Create a matrix to demonstrate the volume of conventional and organic avocadoes

```
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
```

```
vol_type <- df %>% group_by(type) %>% summarise(average_volume = round(mean(total_volume),3),average_pr
vol_type
```

```
## # A tibble: 2 x 4
##   type         average_volume average_price volume_percent
##   <fct>                 <dbl>         <dbl>          <dbl>
## 1 conventional       1818206.          1.16          96.8
## 2 organic              60127.          1.62           3.20
```

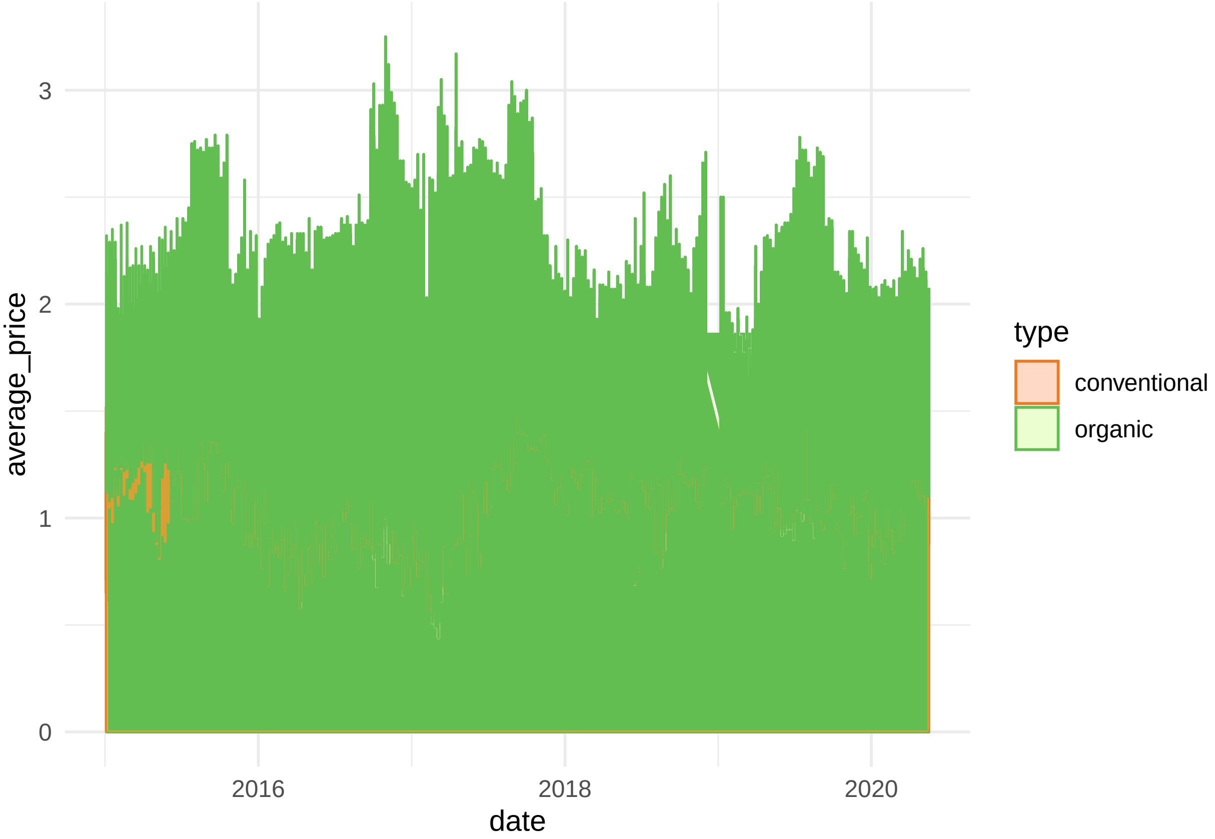
As can be seen from the density plot and the table in avocadoes. - there are two types of avocado: organic and conventional - organic avocado share a small percent (3.2%) of volume but has a high price (1.62) - conventional avocado share a large percent (96.8) of volume but has a relative low price (1.16)

Avocado price with the Date

```
library(ggplot2)
## Change the Date column from factor to the date format
df$date <- as.Date(df$date, "%Y-%m-%d")

## Sort the dates and order the datesets in date
df <- df[order(df$date),]

## Make the plot
df %>% select(date, average_price, type) %>%
  ggplot(aes(x=date,y=average_price))+
  geom_area(aes(color=type,fill=type),alpha=0.3,position=position_dodge(0.8))+
  theme_minimal()+
  scale_color_manual(values = c("#ED7921","#62BE51"))+
  scale_fill_manual(values = c("#FD833E","#B8FC5F")
)
```


average_price
date
type
conventional
organic