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(forecast)
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
```

Check whether the dataset contains the missing value

```r
sum(is.na(df))
```

```
## [1] 0
```

The overall dataset do not contain any missing value

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

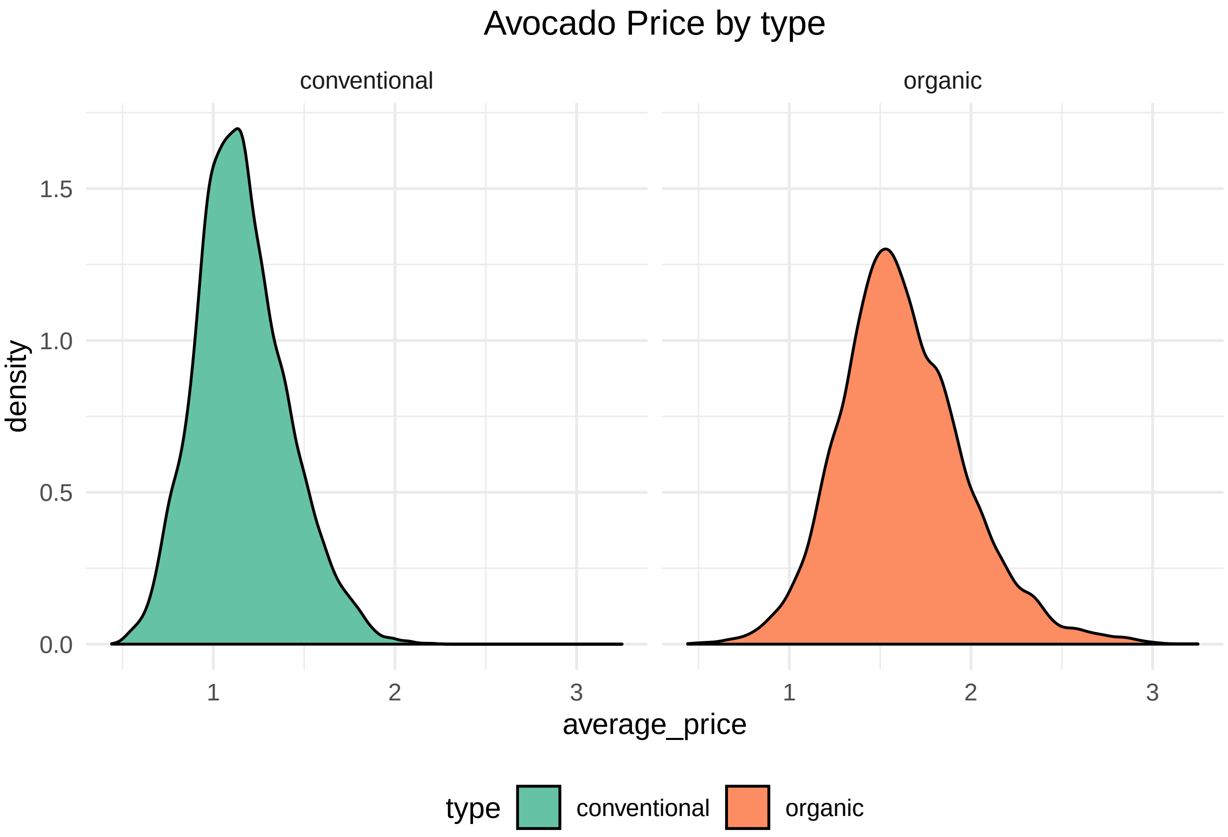
Exploratory Data Analysis

```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 avocados

```
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
```

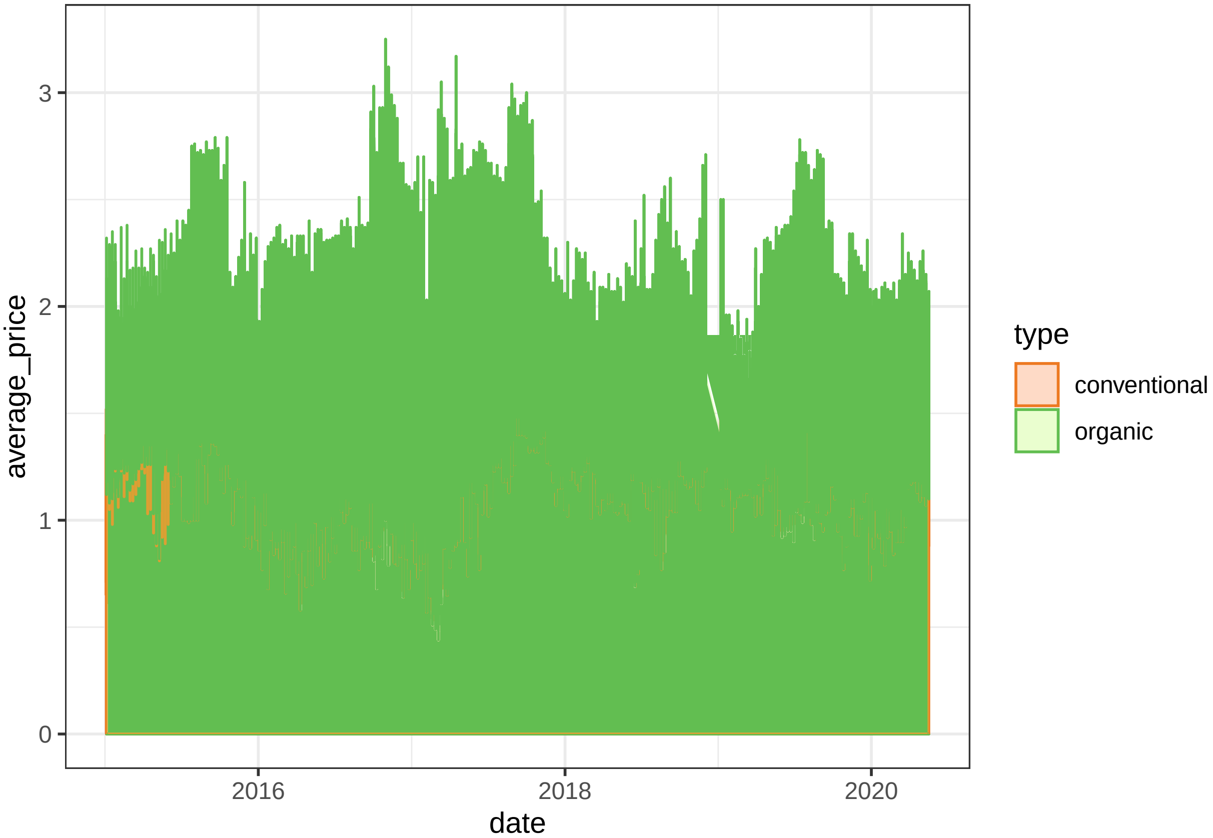
```
vol_type <- df %>% group_by(type) %>% summarise(average_volume = round(mean(total_volume),3),average_pr
vol_type
```

```
## # 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
```

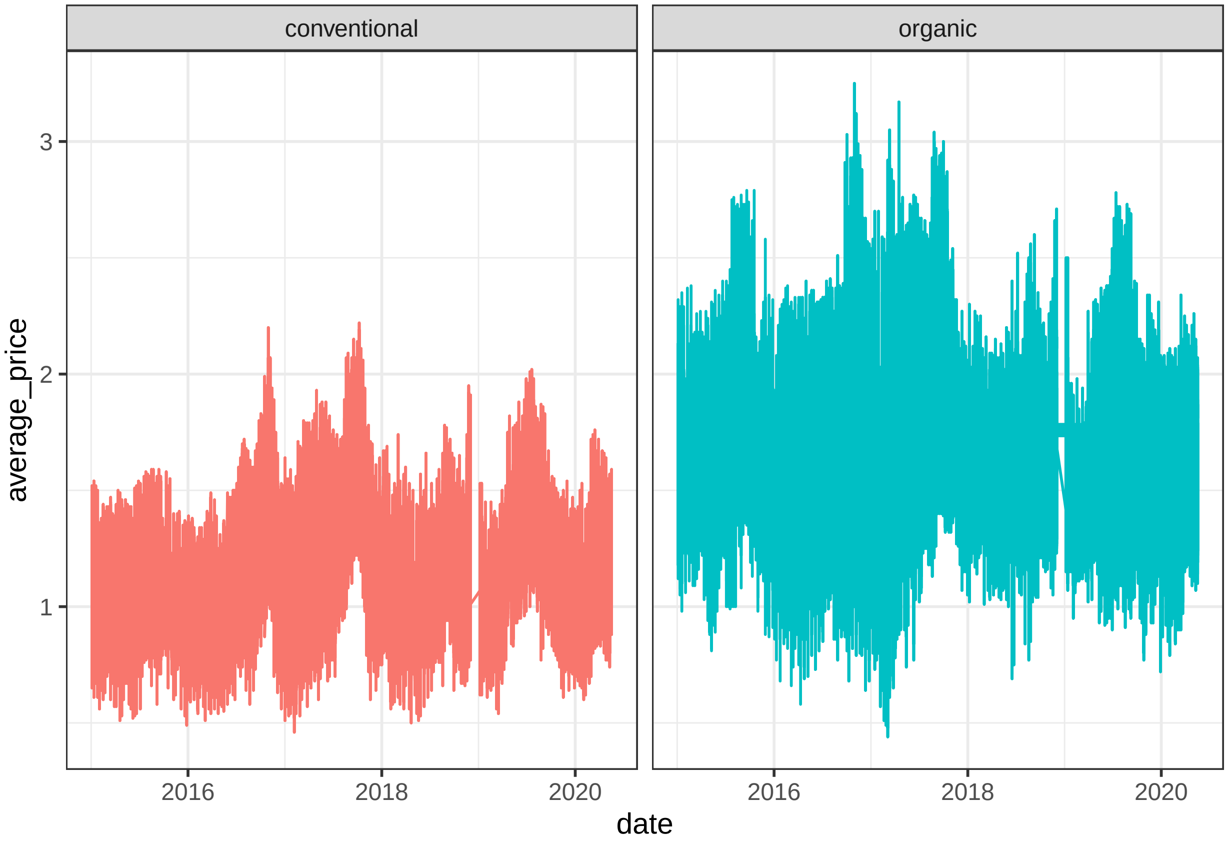
As can be seen from the density plot and the table in avocados. - there are two types of avocado: organic and conventional - organic avocado share a small percent (3.2%) of volume but has a high price (1.62) - conventional avocado share a large percent (96.8) of volume but has a relative low price (1.16)

Avocado price with the Date

```r
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_bw()+
  scale_color_manual(values = c("#ED7921","#62BE51"))+
  scale_fill_manual(values = c("#FD833E","#B8FC5F")
)
```



```r
ggplot(data=df, aes(x=date, y=average_price,col=type))+
  geom_line()+
  facet_wrap(~ type)+
  theme_bw()+
  theme(legend.position = "position")
```



### Relationship between Prices and Total on either conventional or organic avocados

```r
organic <- df %>% select(type,average_price,total_volume,date) %>% filter(type=="organic")
#head(organic)
conventional <- df %>% select(type,average_price,total_volume,date) %>% filter(type=="conventional")
#head(conventional)
```

```r
library(tibbletime)
```

Filter the data into two categories, conventional or organic

```
## Warning: package 'tibbletime' was built under R version 3.6.2
```

```
##
## Attaching package: 'tibbletime'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

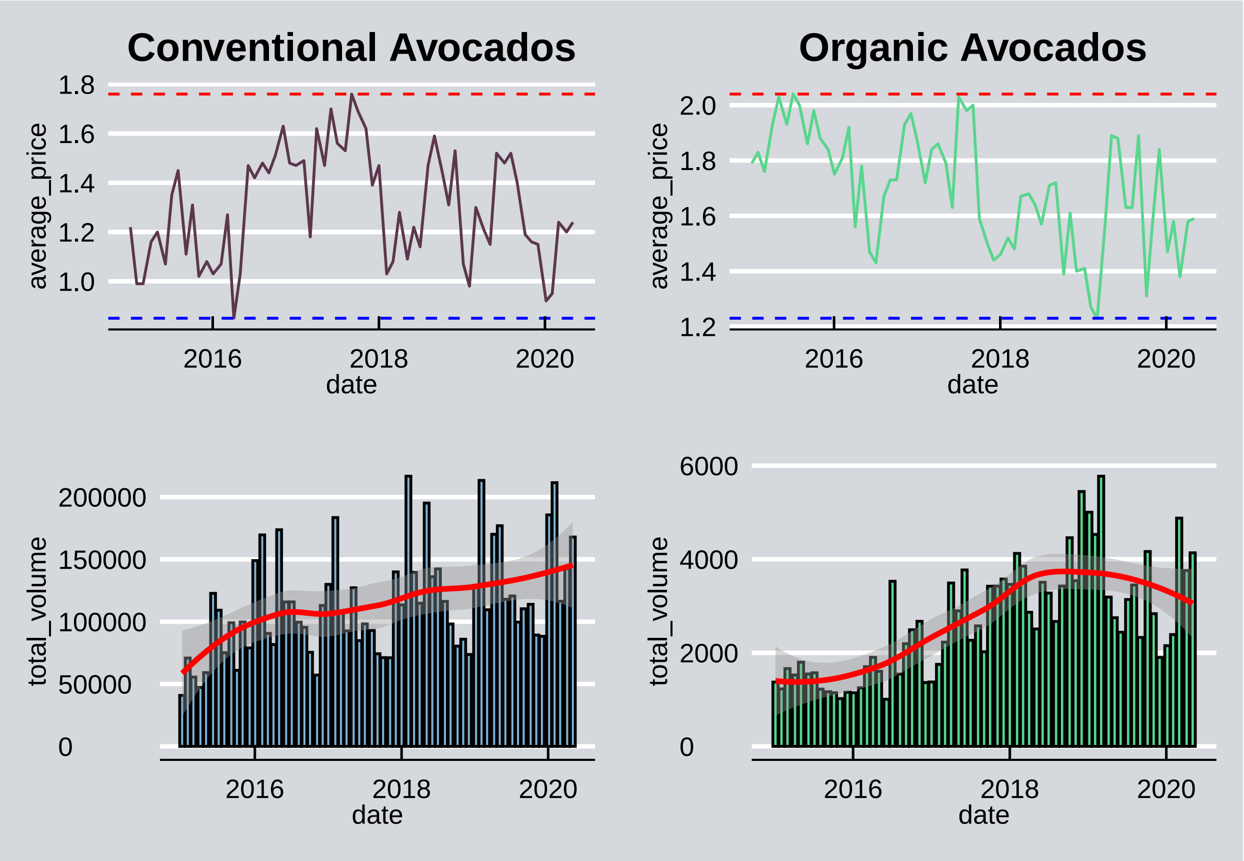
```r
organic <- as_tbl_time(organic,index = date) %>% as_period('1 month')
conventional <- as_tbl_time(conventional,index = date) %>% as_period('monthly')
```

```r
library(ggplot2)
library(ggthemes)
library(cowplot)
```

Monthly avocados price in either conventional or organic avocados

```
##
## ********************************************************
## Note: As of version 1.0.0, cowplot does not change the
##   default ggplot2 theme anymore. To recover the previous
##   behavior, execute:
##   theme_set(theme_cowplot())
## ********************************************************
##
## Attaching package: 'cowplot'
## The following object is masked from 'package:ggthemes':
##
##     theme_map
```

```
options(repr.plot.width=8, repr.plot.height=6)

## average-price with time series
conventional_monthly <- conventional %>%
    ggplot(aes(x=date,y=average_price))+
    geom_line(color="#5C374C")+
    theme_economist()+
    theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill = "#D5D8DC"))+
    labs(title = "Conventional Avocados")+
    geom_hline(yintercept = max(conventional$average_price),linetype="dashed",color = "red")+
    geom_hline(yintercept = min(conventional$average_price),linetype="dashed",color = "blue")

organic_monthly <- organic %>%
    ggplot(aes(x=date,y=average_price))+
    geom_line(color="#58D68D")+
    theme_economist()+
    theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill = "#D5D8DC"))+
    labs(title = "Organic Avocados")+
    geom_hline(yintercept = max(organic$average_price),linetype="dashed",color = "red")+
    geom_hline(yintercept = min(organic$average_price),linetype="dashed",color = "blue")

## create a volume chart
conventional_volume <- conventional %>%
    ggplot(aes(x=date,y=total_volume))+
    geom_bar(stat = 'identity',fill="#7FB3D5",color="black")+
    theme_economist()+
    theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill = "#D5D8DC"))+
    geom_smooth(method = "loess",color="red")

organic_volume <- organic %>%
    ggplot(aes(x=date,y=total_volume))+
    geom_bar(stat = 'identity',fill='#58D68D',color="black")+
    theme_economist()+
    theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill = "#D5D8DC"))+
    geom_smooth(method = "loess",color ="red")

plot_grid(conventional_monthly,organic_monthly,conventional_volume,organic_volume,nrow = 2,ncol = 2)
```



```
#plot_grid(conventional_monthly,conventional_volume,nrow = 2)
```

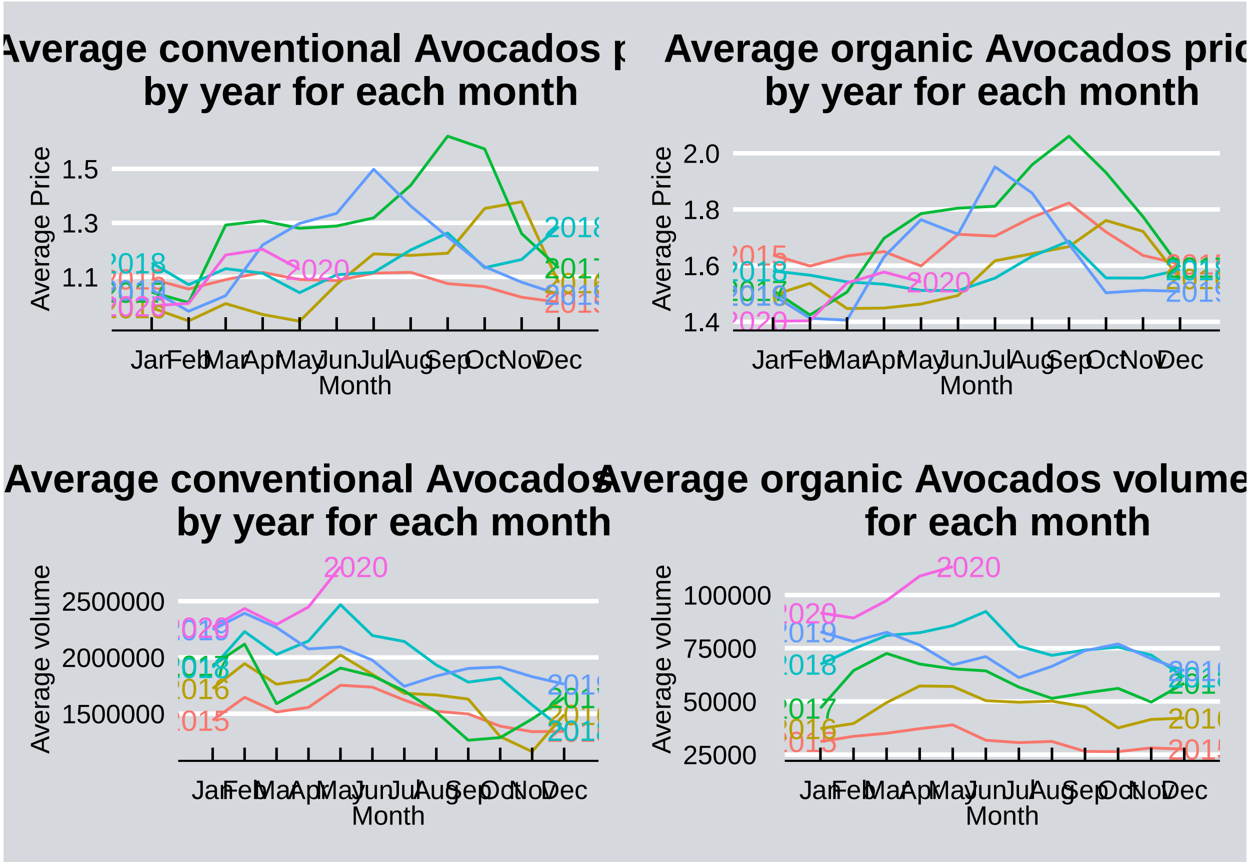
Patterns among the years in each month (Autoplot library)

```
## Process the data into year and month format
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.6.2

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

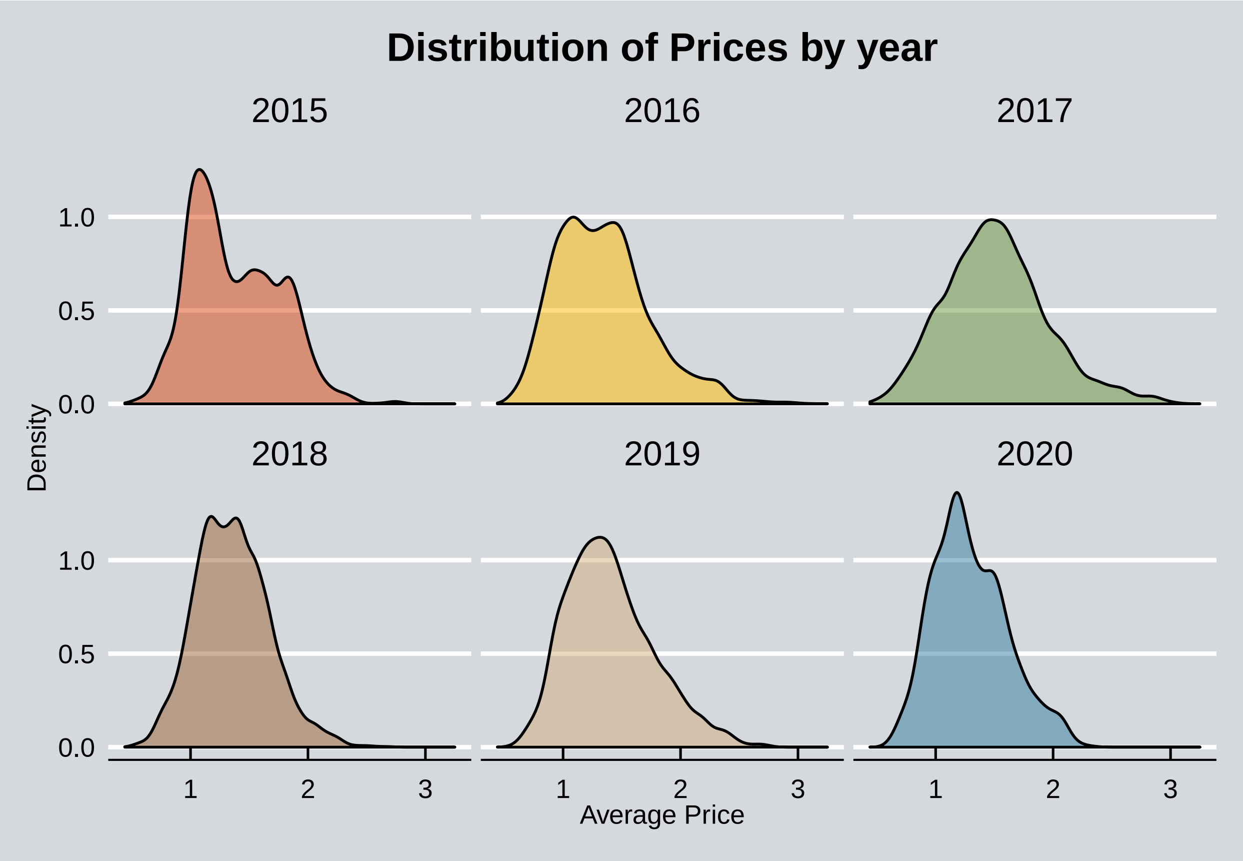
```
seasonal_df <- read.csv("/Users/yuxuan/Desktop/INT301-Avocado-prediction/avocado-updated-2020.csv")
seasonal_df$month_year <- format(as.Date(seasonal_df$date),"%Y-%m")
seasonal_df$month <- format(as.Date(seasonal_df$date),"%m")

## Change the month from a Date format into a numerical foramt, then convert to the three letter format
seasonal_df$monthabb <- sapply(seasonal_df$month, function (x) month.abb[as.numeric(x)])
seasonal_df$monthabb <- factor(seasonal_df$monthabb,levels=month.abb)
seasonal_df$monthabb <- factor(seasonal_df$monthabb)

## Set the figure size
options(repr.plot.width=10,repr.plot.height=8)

## Analyze the price by month
```

```
conv_price <- seasonal_df %>% select(type,year,monthabb,average_price) %>% filter(type=="conventional")

org_price <- seasonal_df %>% select(type,year,monthabb,average_price) %>% filter(type=="organic") %>% g

conv_price <- ts(conv_price$avg,start = 2015,frequency = 12)
org_price <- ts(org_price$avg,start = 2015,frequency = 12)

## Analyze the volume by month
conv_volume <- seasonal_df %>% select(type,year,monthabb,total_volume) %>% filter(type=="conventional")

org_volume <- seasonal_df %>% select(type,year,monthabb,total_volume) %>% filter(type=="organic") %>% g

conv_volume <- ts(conv_volume$avg,start = 2015,frequency = 12)
org_volume <- ts(org_volume$avg,start = 2015,frequency = 12)

byyearplot_price_conv <- ggseasonplot(conv_price,year.labels = TRUE,year.labels.left = TRUE)+
theme_economist()+
theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#D5D8DC"))+
labs(title = "Average conventional Avocados price \n by year for each month", y="Average Price")+
scale_fill_manual(values = c("#922B21", "#EE865D", "#DDCD5E", "#59BEC4", "#048B9F", "#114676"))

byyearplot_price_org <- ggseasonplot(org_price,year.labels = TRUE,year.labels.left = TRUE)+
theme_economist()+
theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#D5D8DC"))+
labs(title = "Average organic Avocados price \n by year for each month", y="Average Price")+
scale_fill_manual(values = c("#922B21", "#EE865D", "#DDCD5E", "#59BEC4", "#048B9F", "#114676"))

byyearplot_volume_conv <- ggseasonplot(conv_volume,year.labels = TRUE,year.labels.left = TRUE)+
theme_economist()+
theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#D5D8DC"))+
labs(title = "Average conventional Avocados volume \n by year for each month", y="Average volume")+
scale_fill_manual(values = c("#922B21", "#EE865D", "#DDCD5E", "#59BEC4", "#048B9F", "#114676"))

byyearplot_volume_org <- ggseasonplot(org_volume,year.labels = TRUE,year.labels.left = TRUE)+
theme_economist()+
theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#D5D8DC"))+
labs(title = "Average organic Avocados volume by year \n for each month", y="Average volume")+
scale_fill_manual(values = c("#922B21", "#EE865D", "#DDCD5E", "#59BEC4", "#048B9F", "#114676"))

plot_grid(byyearplot_price_conv,byyearplot_price_org,byyearplot_volume_conv,byyearplot_volume_org,nrow =
```



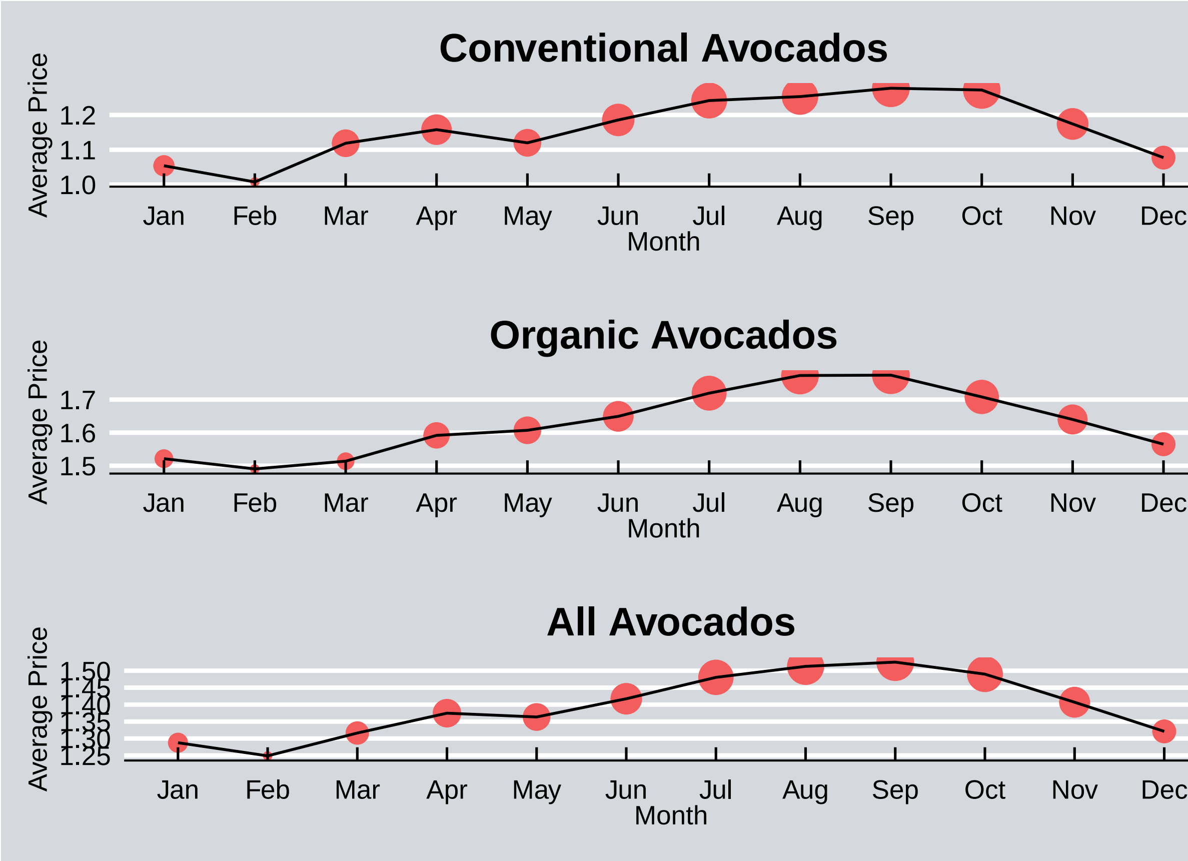
Seasonal patterns analysis

```
ggplot(seasonal_df,aes(x=average_price,fill=as.factor(year)))+
geom_density(alpha=0.5)+
theme_economist()+
facet_wrap(~year)+
theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#D5D8DC"))+
guides(fill=FALSE)+
labs(title = "Distribution of Prices by year",x='Average Price',y='Density')+
scale_fill_manual(values = c("#DA4511", "#FFBD00", "#6A953F", "#9A6233", "#D3AE7C", "#307CA1"))
```



Seasonality patterns

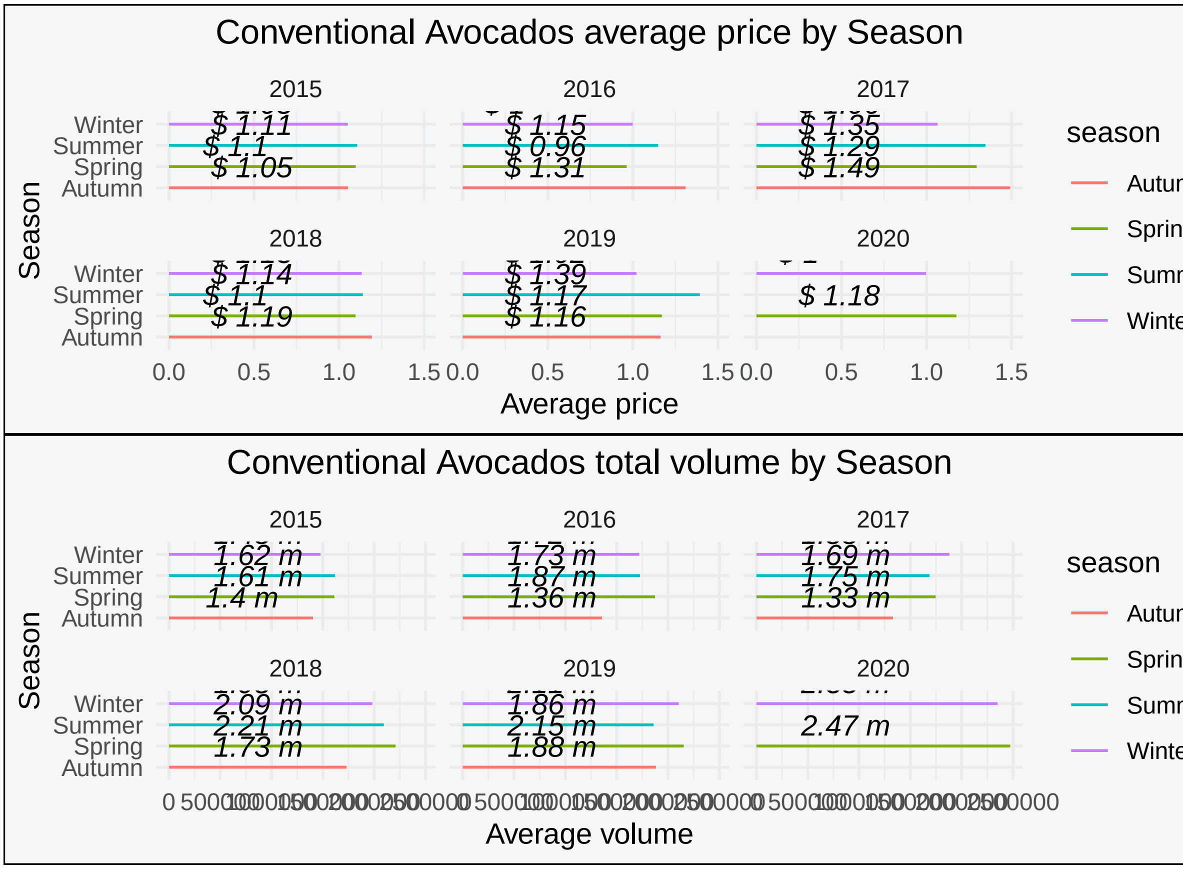
```
options(repr.plot.width=10,repr.plot.height=8)
conv_patterns <- seasonal_df %>% select(monthabb,average_price,type) %>% filter(type=="conventional") %
  ggplot(aes(x=monthabb, y=avg))+
  geom_point(color="#F35D5D",aes(size=avg))+
  geom_line(group=0)+
  theme_economist()+
  theme(legend.position = "none",plot.title = element_text(hjust = 0.5),plot.background = element_rect(
  labs(title = "Conventional Avocados",x="Month",y="Average Price")

organic_patterns <- seasonal_df %>% select(monthabb,average_price,type) %>% filter(type=="organic") %>%
ggplot(aes(x=monthabb,y=avg))+
  geom_point(color="#F35D5D",aes(size=avg))+
  geom_line(group=0)+
  theme_economist()+
  theme(legend.position = "none",plot.title = element_text(hjust = 0.5),plot.background = element_rect(
  labs(title = "Organic Avocados",x="Month",y="Average Price")

whole_patterns <- seasonal_df %>% select(monthabb,average_price,type) %>% group_by(monthabb) %>% summar
  ggplot(aes(x=monthabb,y=avg))+
  geom_point(color="#F35D5D",aes(size=avg))+
  geom_line(group=0)+
  theme_economist()+
  theme(legend.position = "none",plot.title = element_text(hjust = 0.5),plot.background = element_rect(
  labs(title = "All Avocados",x="Month",y="Average Price")


plot_grid(conv_patterns,organic_patterns,whole_patterns,nrow = 3)
```



Monthly analysis

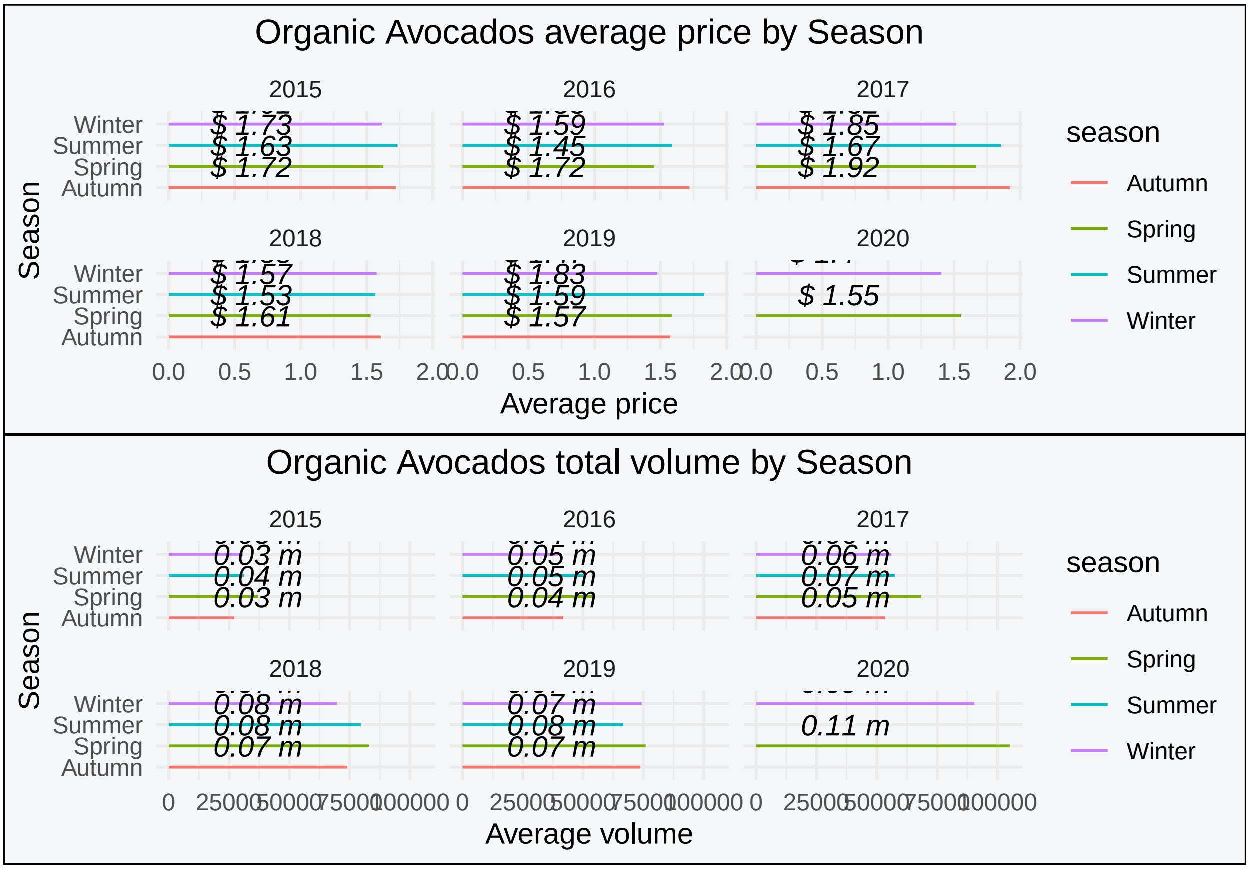
```
#conv_patterns
```

```
options(repr.plot.width=8,repr.plot.height=6)

## seperate the month into four seasons
seasonal_df$season <- ifelse(seasonal_df$month %in% c("03","04","05"),"Spring",
ifelse(seasonal_df$month %in% c("06","07","08"),"Summer",
ifelse(seasonal_df$month %in% c("09","10","11"),"Autumn","Winter")))

## Prepare to analyze the results
seasonality_plot_conventional_price <- seasonal_df %>% select(season,year,average_price,type) %>% filte
  ggplot(aes(x=season,y=avg,color=season))+
  geom_segment(aes(x=season,xend=season,y=0,yend=avg))+
  coord_flip()+
  facet_wrap(~as.factor(year))+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#F4F6F7"))+
  labs(title = "Conventional Avocados average price by Season",x="Season",y="Average price")+
  geom_text(aes(x=season,y=0.01,label=paste0("$ ",round(avg,2))),hjust=-0.5,vjust=-0.5,size=4,color="bla
seasonality_plot_conventional_volume <- seasonal_df %>% select(season,year,total_volume,type) %>% filte
  ggplot(aes(x=season,y=avg,color=season))+
  geom_segment(aes(x=season,xend=season,y=0,yend=avg))+
  coord_flip()+
  facet_wrap(~as.factor(year))+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#F4F6F7"))+
  labs(title = "Conventional Avocados total volume by Season",x="Season",y="Average volume")+
  geom_text(aes(x=season,y=0.01,label=paste0(round(avg/1000000,2)," m")),hjust=-0.5,vjust=-0.5,size=4,co

plot_grid(seasonality_plot_conventional_price,seasonality_plot_conventional_volume,nrow = 2)
```



Seasonal patterns

```
## Prepare to analyze the results
options(repr.plot.width=8,repr.plot.height=6)

seasonality_plot_organic_price <- seasonal_df %>% select(season,year,average_price,type) %>% filter(typ
  ggplot(aes(x=season,y=avg,color=season))+
  geom_segment(aes(x=season,xend=season,y=0,yend=avg))+
  coord_flip()+
  facet_wrap(~as.factor(year))+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#F4F6F7"))+
  labs(title = "Organic Avocados average price by Season",x="Season",y="Average price")+
  geom_text(aes(x=season,y=0.01,label=paste0("$ ",round(avg,2))),hjust=-0.5,vjust=-0.5,size=4,color="bla

seasonality_plot_organic_volume <- seasonal_df %>% select(season,year,total_volume,type) %>% filter(typ
  ggplot(aes(x=season,y=avg,color=season))+
  geom_segment(aes(x=season,xend=season,y=0,yend=avg))+
  coord_flip()+
  facet_wrap(~as.factor(year))+
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5),plot.background = element_rect(fill="#F4F6F7"))+
  labs(title = "Organic Avocados total volume by Season",x="Season",y="Average volume")+
  geom_text(aes(x=season,y=0.01,label=paste0(round(avg/1000000,2)," m")),hjust=-0.5,vjust=-0.5,size=4,c

plot_grid(seasonality_plot_organic_price,seasonality_plot_organic_volume,nrow = 2)
```



levels(df$geography)