Analysis of Berries Dataset

Ruiyi Lian

2020/10/19

1. Data Clean

Load data.

```
dtRaw <- fread("C:/Users/user/Desktop/1900_20201018/berries(3)(1).csv")
dt <- dtRaw
```

Convert value from character to numeric.

```
dt$Value <- str_remove_all(dt$Value,",")
dt$Value <- as.numeric(dt$Value)
```

Distinguish different method of measure.

```
dt$measure <- sub('.*MEASURED IN ', '', dt$`Data Item`)
unique(dt$measure)
```

```
##  [1] "$ / LB"
##  [2] "$ / CWT"
##  [3] "BLUEBERRIES, TAME - ACRES HARVESTED"
##  [4] "LB"
##  [5] "LB / ACRE"
##  [6] "$"
##  [7] "RASPBERRIES - ACRES HARVESTED"
##  [8] "LB / ACRE / APPLICATION, AVG"
##  [9] "LB / ACRE / YEAR, AVG"
## [10] "NUMBER, AVG"
## [11] "PCT OF AREA BEARING, AVG"
## [12] "STRAWBERRIES - ACRES HARVESTED"
## [13] "STRAWBERRIES - ACRES PLANTED"
## [14] "CWT"
## [15] "CWT / ACRE"
## [16] "BLUEBERRIES, WILD - ACRES HARVESTED"
## [17] "$ / TON"
## [18] "TONS"
## [19] "RASPBERRIES, BLACK - ACRES HARVESTED"
## [20] "RASPBERRIES, RED - ACRES HARVESTED"
```

2. Price of Production

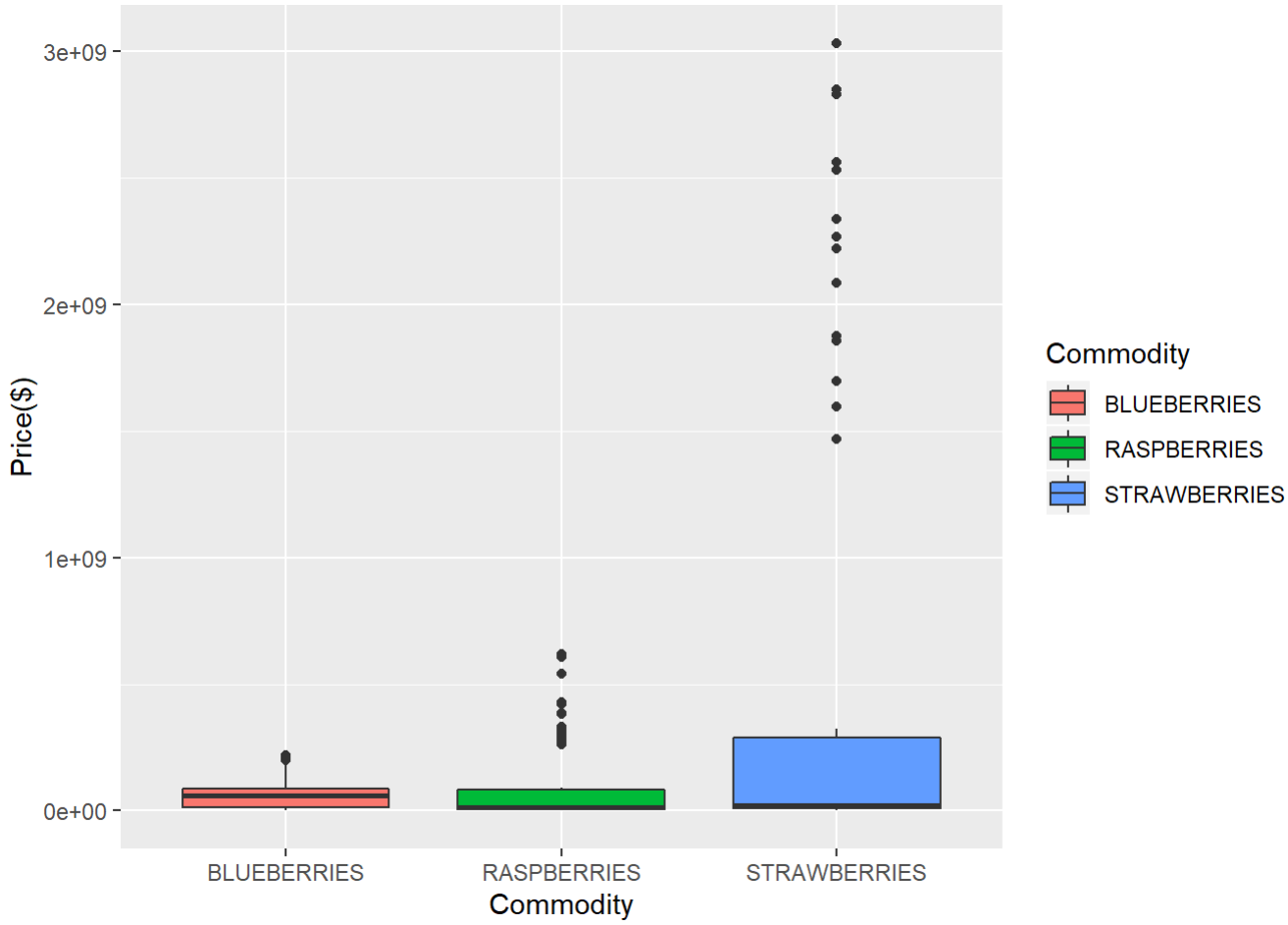
2.1 Select data

```
production <- dt[str_detect(dt$`Data Item`,"PRODUCTION")&
                   dt$measure=="$",]
production <- production[!is.na(production$Value)&
                           production$Value!=0,]
```

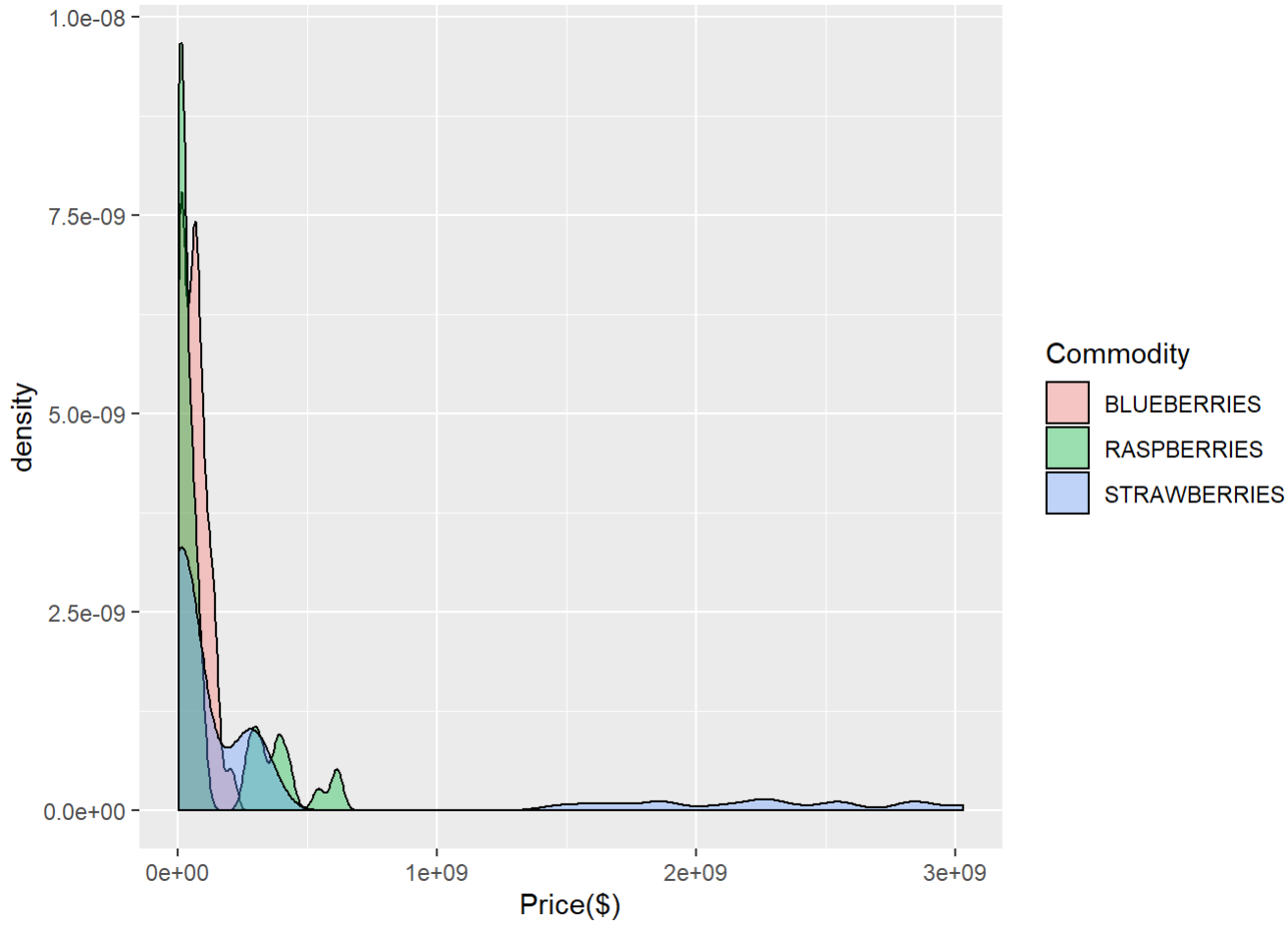
2.2 Affect of Commodity

Confirm whether price of production were affect by commodity or not.

```
ggplot(production,aes(x=Commodity,y=Value,fill=Commodity))+
  geom_boxplot()+
  labs(y="Price($)")
```




```
ggplot(production,aes(x=Value,group=Commodity,fill=Commodity))+
  geom_density(alpha=0.36)+
  labs(x="Price($)")
```



The figures above indicates that strawberries are more expensive than the other two kinds of berries.

```
summary(aov(Value~Commodity,production))
```

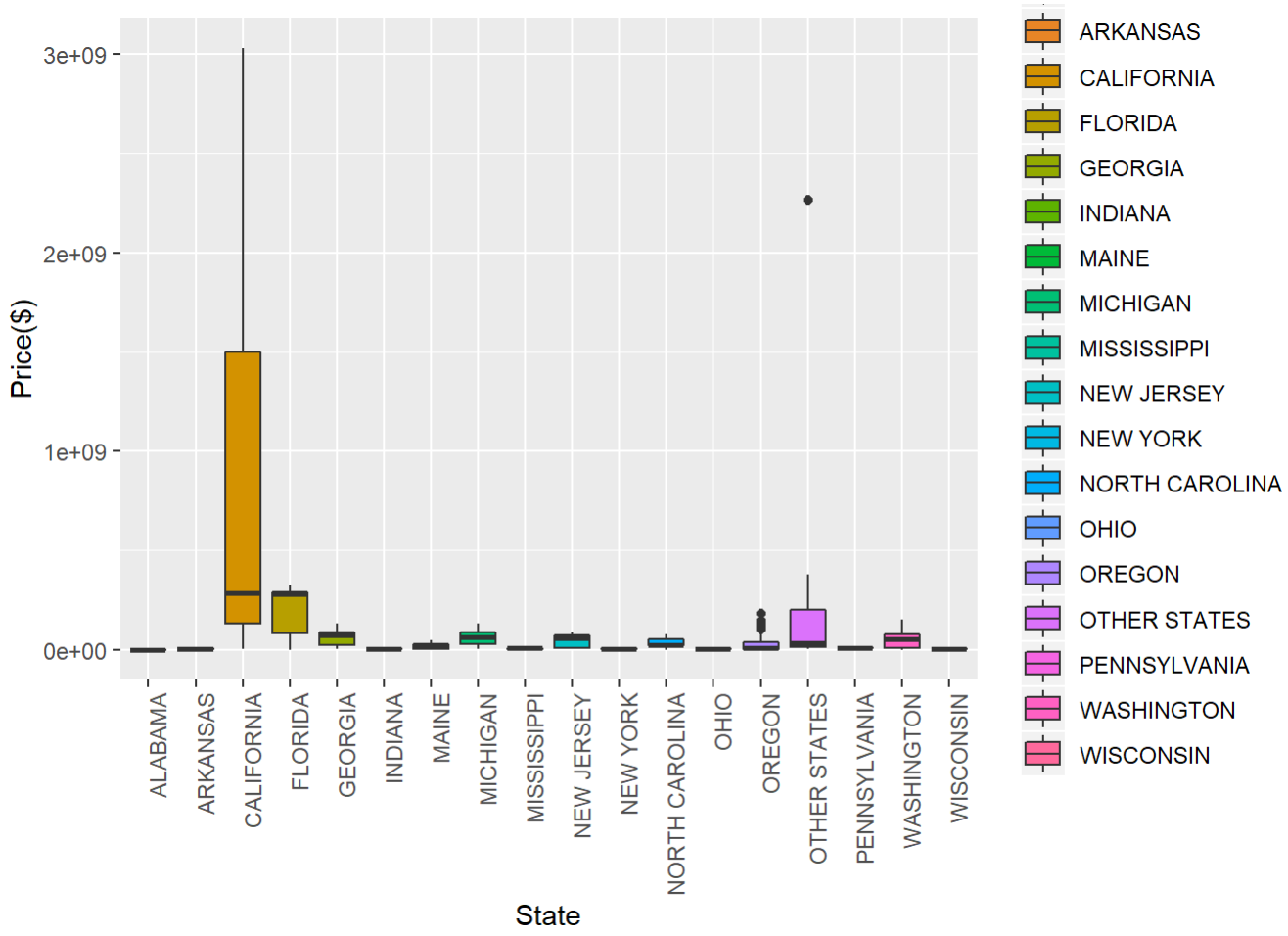
```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## Commodity     2 6.922e+18 3.461e+18   16.49 1.64e-07 ***
## Residuals   292 6.127e+19 2.098e+17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA model show that price of production were affect by commodity.

2.3 Affect of State

Confirm whether price of production were affect by State or not.

```
ggplot(production,aes(x=State,y=Value,group=State,fill=State))+
  geom_boxplot()+
  labs(y="Price($)")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



The figures above indicates that berries in California are more expensive than berries in other states.

```
summary(aov(Value~State,production))
```

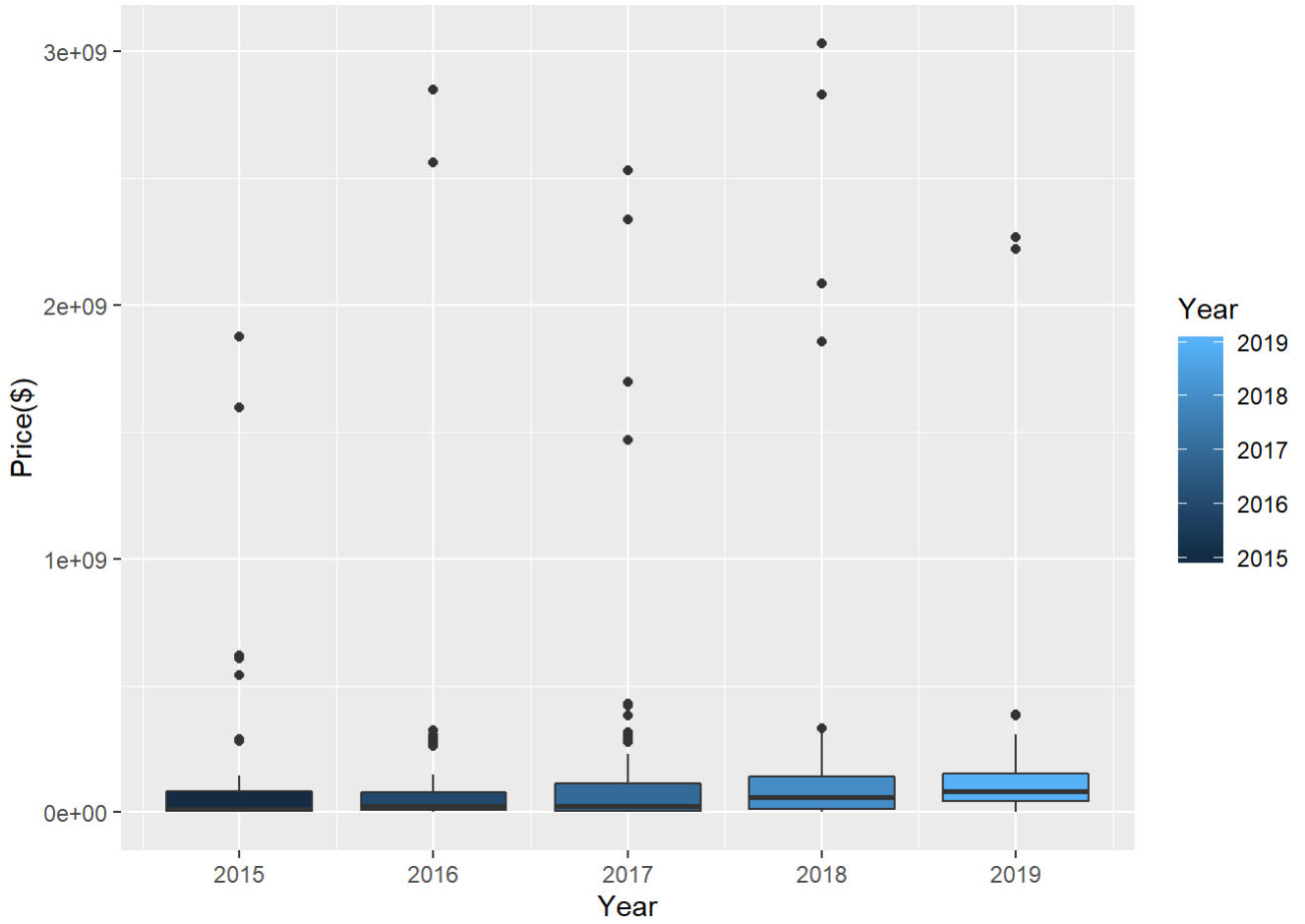
```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## State        17 2.036e+19 1.197e+18   6.934 6.39e-14 ***
## Residuals   277 4.784e+19 1.727e+17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA model show that price of production were affect by State.

2.4 Affect of Year

Confirm whether price of production were affect by year or not.

```
ggplot(production,aes(x=Year,y=Value,group=Year,fill=Year))+
  geom_boxplot()+
  labs(y="Price($)")
```



The figures above indicates that there are no different between price of berries in different years.

```
summary(aov(Value~Year,production))
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Year          1 7.697e+17 7.697e+17   3.345 0.0684 .
## Residuals   293 6.743e+19 2.301e+17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p-value is large than 0.05, we accept H0 and consider that price of production were not affect by Year.

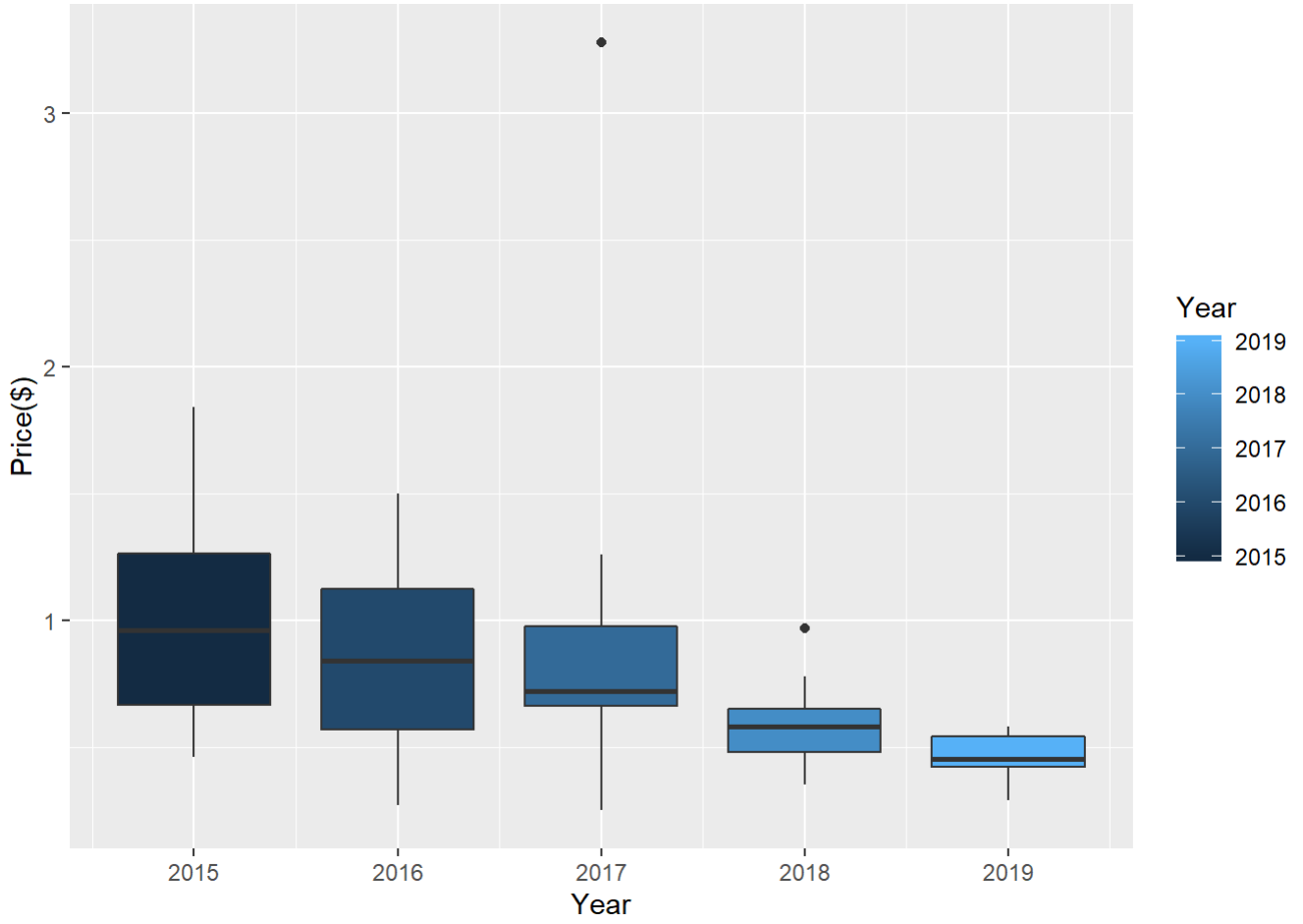
3. Price of procession

3.1 Select Data

```
procession <- dt[str_detect(dt$`Data Item`,"PROCESSING")&
                   dt$measure == "$ / LB",]
procession <- procession[!is.na(procession$Value),]
```

3.2 Affect of Year

```
ggplot(procession,aes(x=Year,y=Value,group=Year,fill=Year))+
  geom_boxplot()+
  labs(y="Price($)")
```



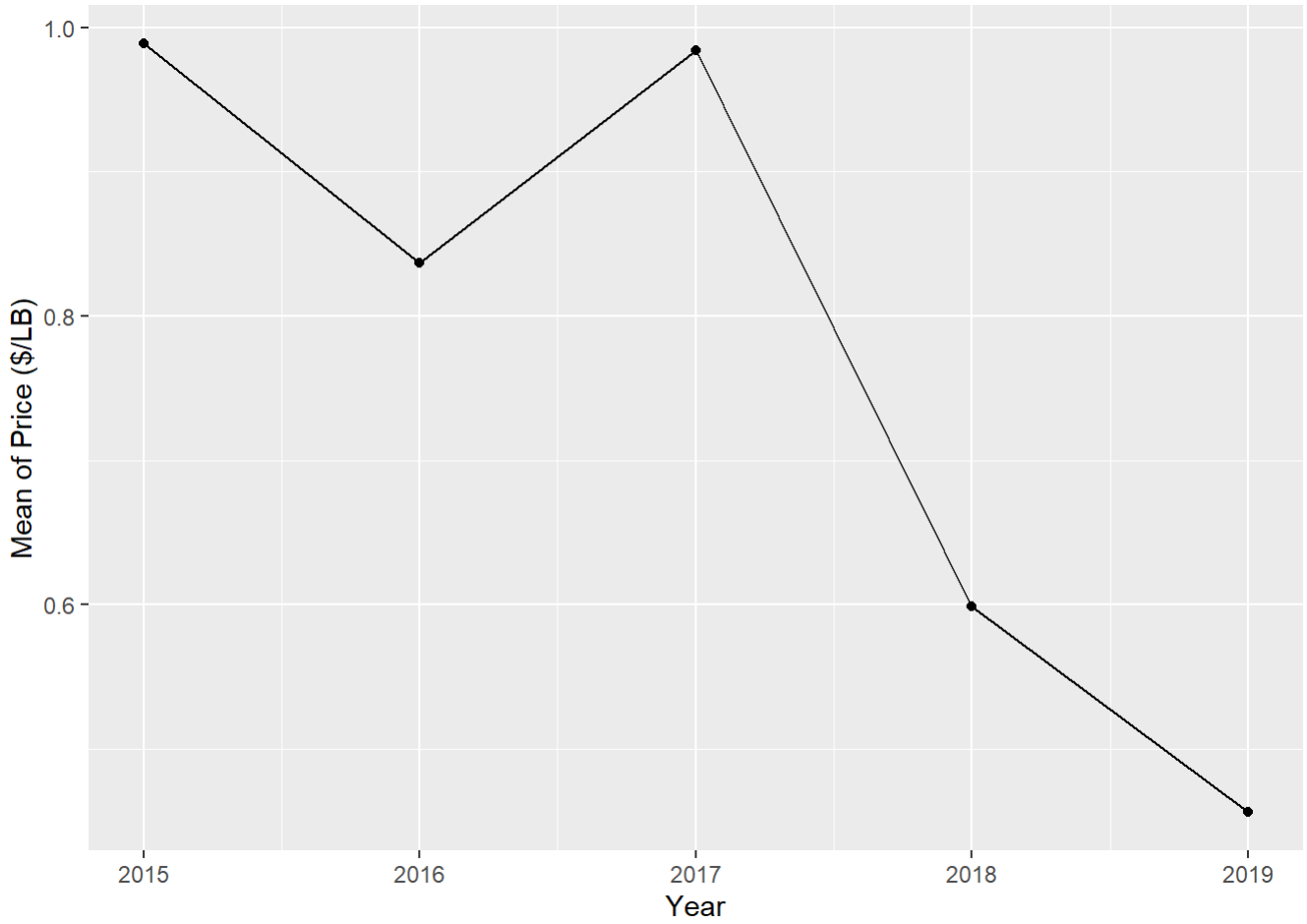

```
summary(aov(Value~Year,procession))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Year          1  1.695  1.6945   8.533 0.00502 **
## Residuals    56 11.121  0.1986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The figure and ANOVA model show that price of procession were affect by year.

3.3 Trend

```
processionYM <- procession[,(meanPrice=mean(Value)),by=Year]
ggplot(processionYM,aes(x=Year,y=V1))+
  geom_line()+
  geom_point()+
  labs(y="Mean of Price ($/LB)")
```



The price of berries procession are decreasing.

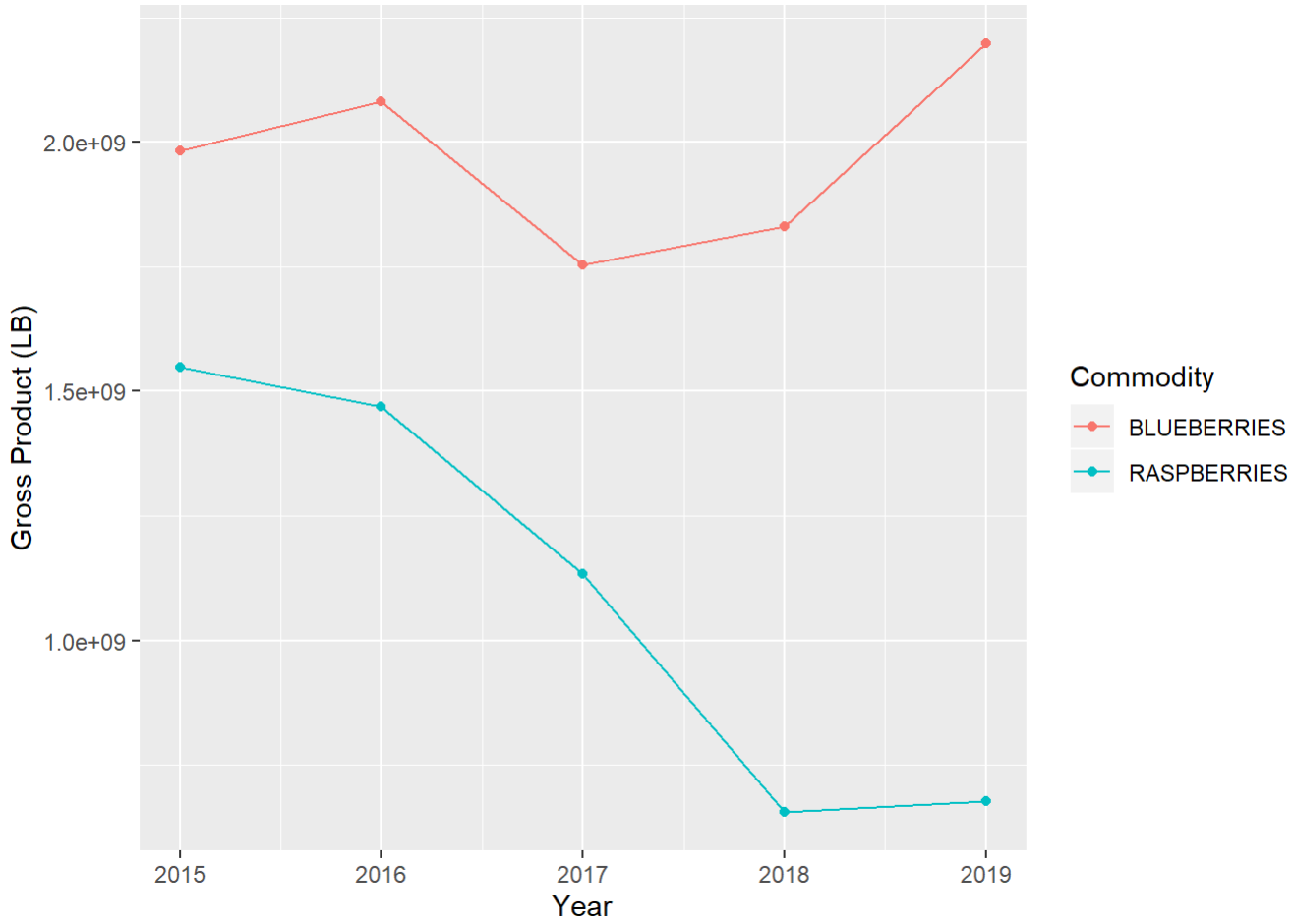
4. Gross Product pf Production

4.1 Select data

```
gpprd <- dt[str_detect(dt$`Data Item`,"PRODUCTION")&
              dt$measure=="LB",]
gpprdYS <- gpprd[,(sum(Value,na.rm = T)),
                 by=c("Year","State","Commodity")]
gpprdC <- gpprd[,(sum(Value,na.rm = T)),
                by=c("Year","Commodity")]
```

4.2 Change of Gross Product of Procession

```
ggplot(gpprdC,aes(x=Year,y=V1,group=Commodity,color=Commodity))+
  geom_line()+
  geom_point()+
  labs(y="Gross Product (LB)")
```



The decrease trend of gross product of raspberries procession is interesting.

4.3 Change of Raspberries in Different States.

```
rasp <- gpprd[gpprd$Commodity=="RASPBERRIES"][,(sum(Value,na.rm = T)),
                                              by=c("Year","State")]
rasp <- dcast(rasp,Year~State,value.var = "V1")
rasp
```

```
##    Year CALIFORNIA   OREGON OTHER STATES WASHINGTON
## 1: 2015 1122680000 56330000           NA  369975000
## 2: 2016 1017680000 51240000           NA  400250000
## 3: 2017  715950000 26720000           NA  389940000
## 4: 2018  286000000       NA    217320000  151600000
## 5: 2019  287000000       NA    225840000  165000000
```

Both of gross product of raspberries procession in California and Washington are decrease.

Citation

1.*Cookbook for R*
2.*Market Analysis of Fresh Berries in the United States*