MA615_Berries_Kunyu Liu

Kunyu Liu

10/15/2020

Introduction

The berries data were collected from the USDA database and stored online[^1], which contains blueberries, raspberries, and strawberries. *** # Data Cleaning

Data Import

```r
berries = read.csv('berries.csv', header = TRUE)
im1 = berries %>%
  select(Year,Period,State,Commodity,Data.Item,Domain,Domain.Category,Value)
head(im1)
```

```
##   Year         Period      State   Commodity
## 1 2019 MARKETING YEAR CALIFORNIA BLUEBERRIES
## 2 2019 MARKETING YEAR CALIFORNIA BLUEBERRIES
## 3 2019 MARKETING YEAR CALIFORNIA BLUEBERRIES
## 4 2019 MARKETING YEAR CALIFORNIA RASPBERRIES
## 5 2019 MARKETING YEAR CALIFORNIA RASPBERRIES
## 6 2019 MARKETING YEAR CALIFORNIA RASPBERRIES
##                                                               Data.Item Domain
## 1               BLUEBERRIES, TAME - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 2 BLUEBERRIES, TAME, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 3   BLUEBERRIES, TAME, PROCESSING - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 4                     RASPBERRIES - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 5       RASPBERRIES, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 6         RASPBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
##   Domain.Category Value
## 1   NOT SPECIFIED  2.85
## 2   NOT SPECIFIED  3.56
## 3   NOT SPECIFIED  0.29
## 4   NOT SPECIFIED  2.69
## 5   NOT SPECIFIED   (D)
## 6   NOT SPECIFIED   (D)
```

Initial Screening of the Data

There are many categorical variables, we need to replace many (D),(NA),(X) and (Z) with NA in Value, because this column is defined as categorical.

```
im1$Value <- as.numeric(im1$Value)
```

```
## Warning:      NA
```

```
# Replace (D),(NA),(X) and (Z) with NA
im1[im1 =="(D)"] = NA
im1[im1 =="(NA)"] = NA
im1[im1 =="(X)"] = NA
im1[im1 =="(Z)"] = NA

# summary the new dataset
summary(im1)
```

```
##       Year         Period             State            Commodity
##  Min.   :2015   Length:13238       Length:13238       Length:13238
##  1st Qu.:2016   Class :character   Class :character   Class :character
##  Median :2017   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2017
##  3rd Qu.:2019
##  Max.   :2019
##
##   Data.Item           Domain          Domain.Category       Value
##  Length:13238       Length:13238       Length:13238       Min.   :  0.000
##  Class :character   Class :character   Class :character   1st Qu.:  0.550
##  Mode  :character   Mode  :character   Mode  :character   Median :  1.831
##                                                           Mean   : 49.564
##                                                           3rd Qu.: 26.000
##                                                           Max.   :960.000
##                                                           NA's   :8854
```

Further data cleaning on strawberries

Cleaning - Data Item

Use filter function for extracting data of strawberries

```
im2 = im1 %>% filter(Commodity=="STRAWBERRIES")
summary(im2)
```

```
##       Year         Period             State            Commodity
##  Min.   :2015   Length:3476        Length:3476        Length:3476
##  1st Qu.:2016   Class :character   Class :character   Class :character
##  Median :2018   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2017
##  3rd Qu.:2019
##  Max.   :2019
##
##   Data.Item           Domain          Domain.Category       Value
##  Length:3476        Length:3476        Length:3476        Min.   :  0.000
##  Class :character   Class :character   Class :character   1st Qu.:  0.307
##  Mode  :character   Mode  :character   Mode  :character   Median :  2.000
##                                                           Mean   : 63.618
##                                                           3rd Qu.: 37.000
##                                                           Max.   :960.000
##                                                           NA's   :2247
```

```
strawberry1 = im2 %>% drop_na()

pre = strawberry1$Data.Item
m1 = gsub(" - ",",",pre)

unit1 = str_extract_all(m1, "MEASURED.*[^./AVG]|ACRES")
unit1 = str_replace(unit1, ",","")
unit1 = trimws(1)
```

```
type1 = str_extract_all(m1,"(FRESHMARKET)|(PROCESSING)")
type_data = data.frame(Market.Channel=as.character(type1))
type_data[type_data=="character(0)"] = NA
```

EDA

Data exploration

Summarize Data

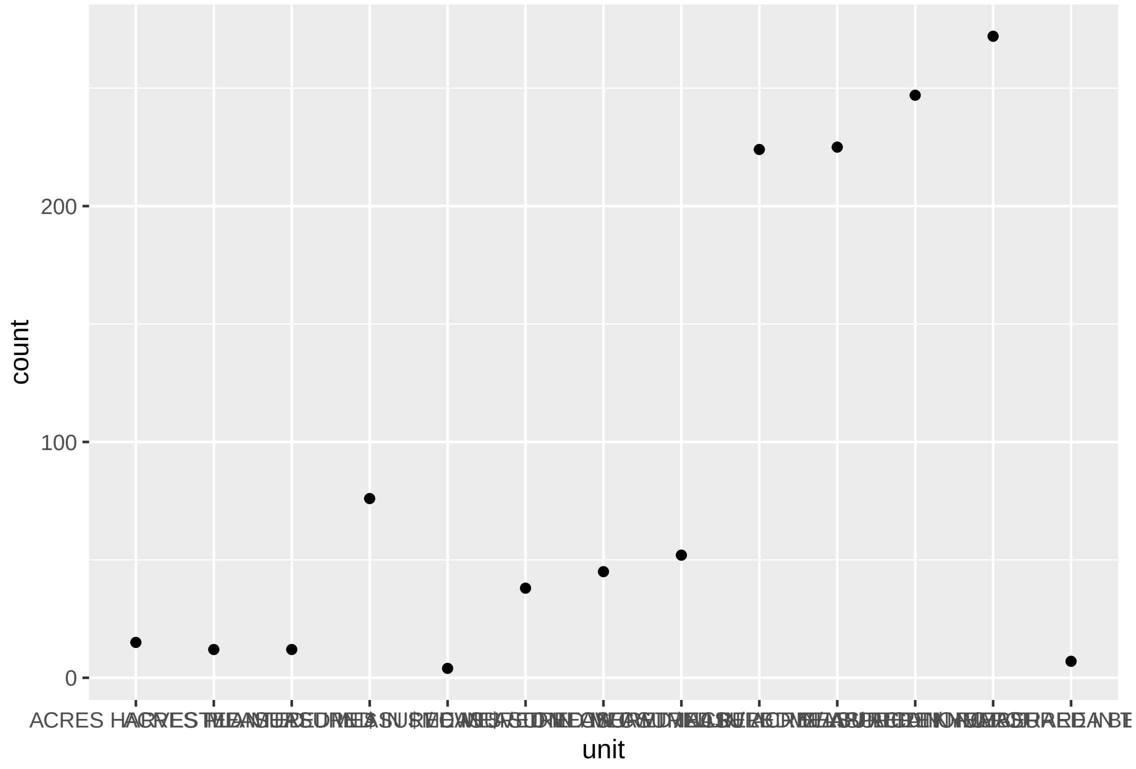
Because the measurement of each data are different, we need to group and summarize them.

```
# Measurement of the strawberry
strawberry1$unit = str_extract_all(m1,"MEASURED IN.*[^, /AVG]|ACRES.*")
strawberry1$unit = as.character(strawberry1$unit)
sum1 = strawberry1 %>%
  group_by(unit)%>%
  summarize(
    count=n(),
    value=sum(Value)
    )
```

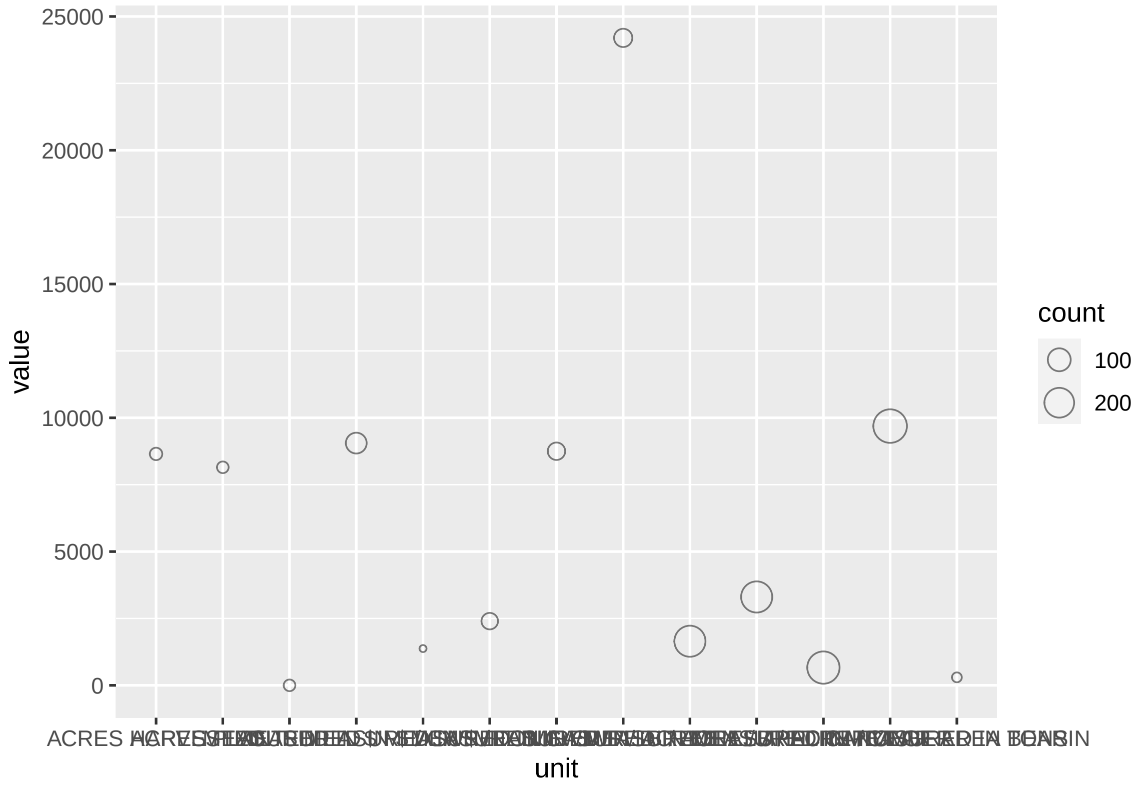
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Plot the whole dataset

```
a1 = ggplot(data = sum1, mapping = aes(x = unit, y = count))+
   geom_point()
print(a1)
```



```
a2 = ggplot(data = sum1, mapping = aes(x = unit, y = value, size = count)) +
  geom_point(shape=21, alpha = 0.5)
print(a2)
```



From the first plot, we can see the number of each measurement. The second plot shows the reason for seperate variable item.

Further EDA

Creat a new frame in order to exact data to do the further EDA

```
unit_new = strawberry1 %>%
  group_by(unit)%>%
  summarize(
    state=State,
    year= Year,
    count=n(),
    value=Value
    )
```

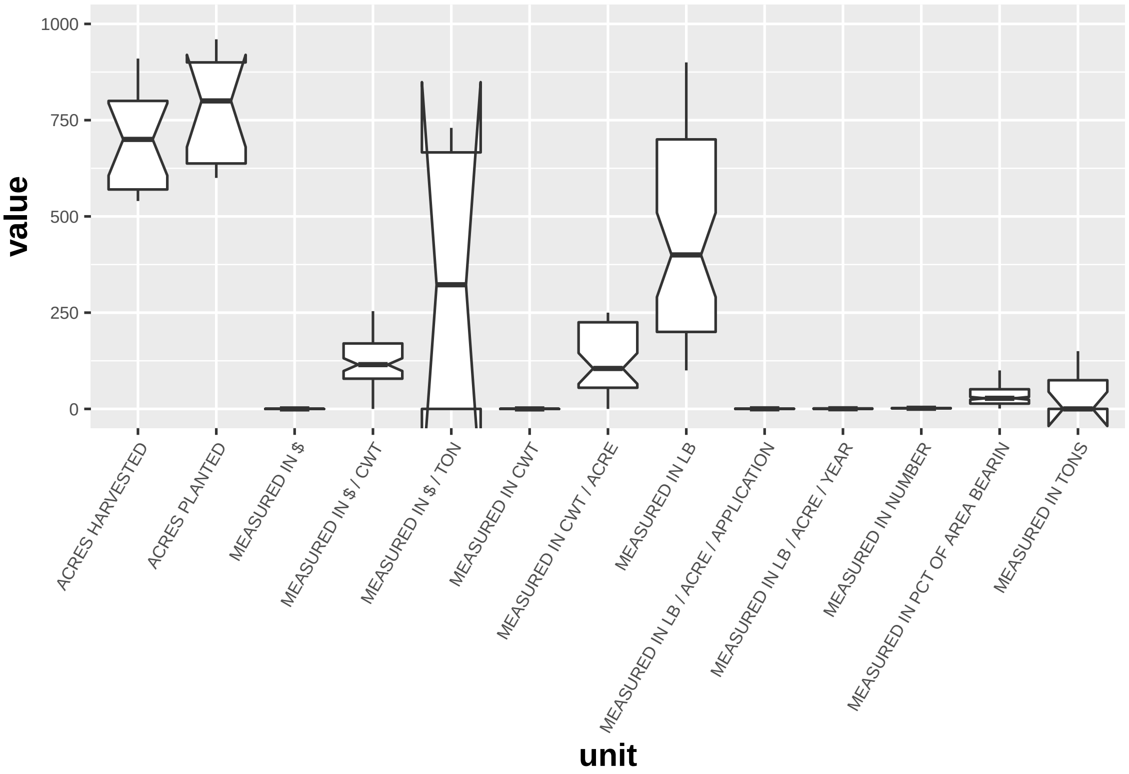
```
## `summarise()` regrouping output by 'unit' (override with `.groups` argument)
```

```
tail(unit_new)
```

```
## # A tibble: 6 x 5
## # Groups:   unit [1]
##   unit             state           year count value
##   <chr>            <chr>          <int> <int> <dbl>
## 1 MEASURED IN TONS NORTH CAROLINA  2018     7     0
## 2 MEASURED IN TONS FLORIDA         2018     7     0
## 3 MEASURED IN TONS NORTH CAROLINA  2018     7     0
## 4 MEASURED IN TONS NORTH CAROLINA  2017     7   149
## 5 MEASURED IN TONS NORTH CAROLINA  2017     7   150
## 6 MEASURED IN TONS FLORIDA         2016     7     0
```

```
# Do a plot, excluding outliers
boxplot = ggplot(unit_new, aes(x = unit, y = value))+
  geom_boxplot(outlier.colour = NA,notch = TRUE) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        axis.text = element_text(size = 7),
        axis.title = element_text(size = 13, face = "bold")) +
  coord_cartesian(ylim = c(0, 1000))
print(boxplot)
```

```
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
## notch went outside hinges. Try setting notch=FALSE.
```



Creat a data frame contain MEASURED IN LB

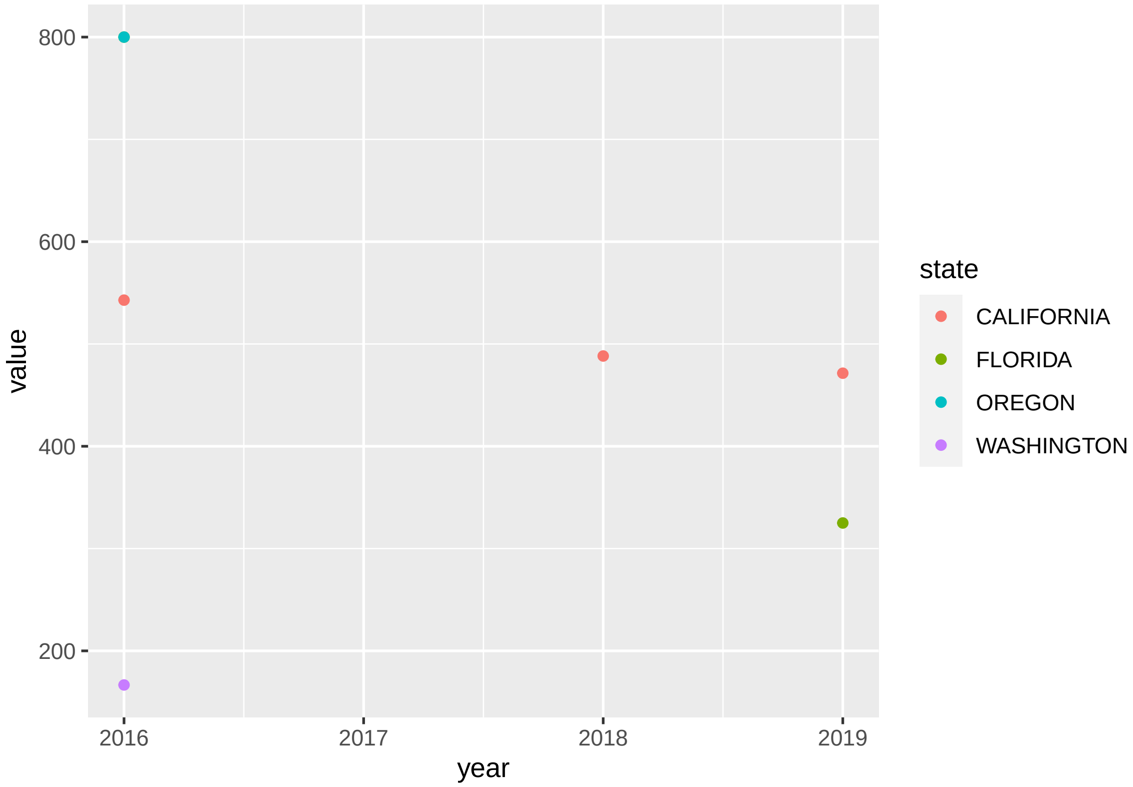
```
LB = filter(unit_new,unit=="MEASURED IN LB" )
LB$value = as.numeric(LB$value)
LB$value[LB$value ==0] = NA
LB_new = group_by(LB,year,state)
LB_final = summarize(LB_new, value = mean(value, na.rm = TRUE))
```

```
## `summarise()` regrouping output by 'year' (override with `.groups` argument)
```

```
summary(LB_final)
```

```
##       year         state               value
##  Min.   :2016   Length:7           Min.   :166.7
##  1st Qu.:2016   Class :character   1st Qu.:398.2
##  Median :2016   Mode  :character   Median :488.2
##  Mean   :2017                      Mean   :513.5
##  3rd Qu.:2018                      3rd Qu.:671.4
##  Max.   :2019                      Max.   :800.0
```

```
# Making plot
a3 = ggplot(LB_final, aes(x = year, y = value))+
  geom_point(aes(color=state))
print(a3)
```

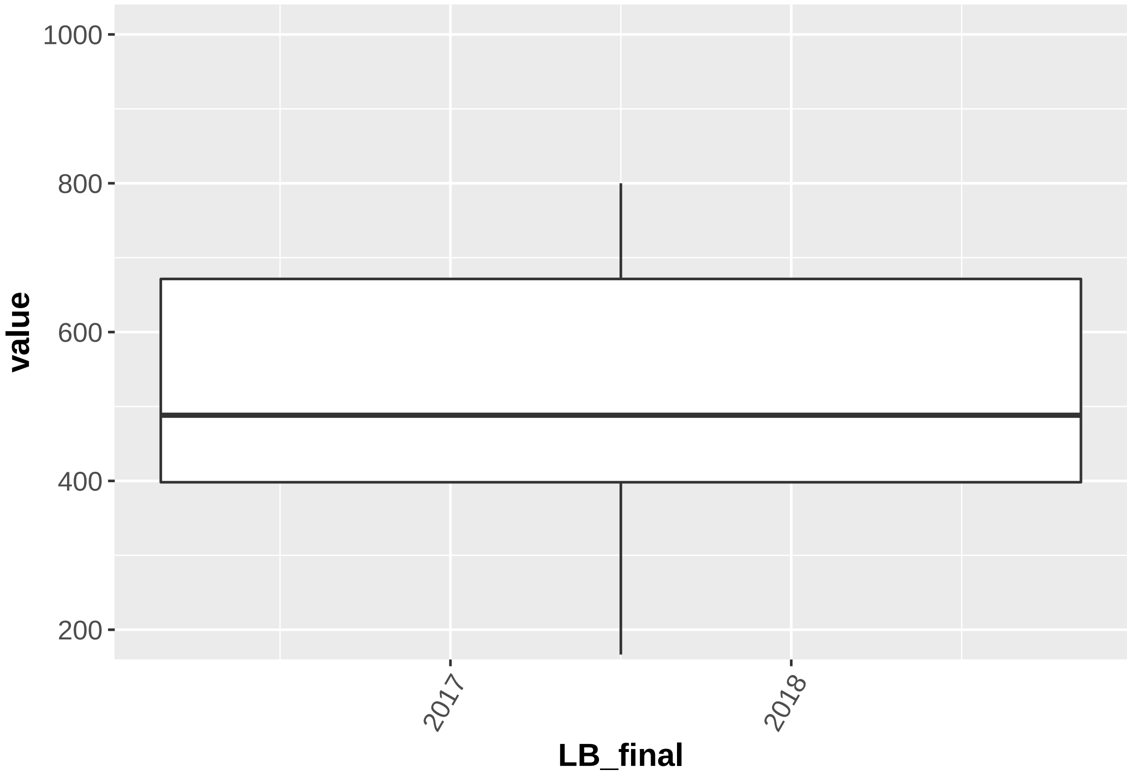


From the plot above, we can see that California always has the highest value in each year when we measured in LB, except in year 2016.

Making boxplot of MEASURED IN LB

```
# excluding outliers
bp1 = ggplot(LB_final, aes(x = year, y = value))
bp1 = bp1 + geom_boxplot(outlier.colour = NA) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        axis.text = element_text(size = 11),
        axis.title = element_text(size = 13, face = "bold")) +
  coord_cartesian(ylim = c(200, 1000)) +
  labs(x = "LB_final")
print(bp1)
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



From the boxplot, we can find the value for LB_final is between 400-700

Creat a data frame contain MEASURED IN $ / CWT

```
CWT = filter(unit_new,unit=="MEASURED IN $ / CWT" )
CWT$value = as.numeric(CWT$value)
CWT$value[CWT$value ==0] = NA
CWT_new = group_by(CWT,year,state)
CWT_final = summarize(CWT_new, value = mean(value, na.rm = TRUE))
```

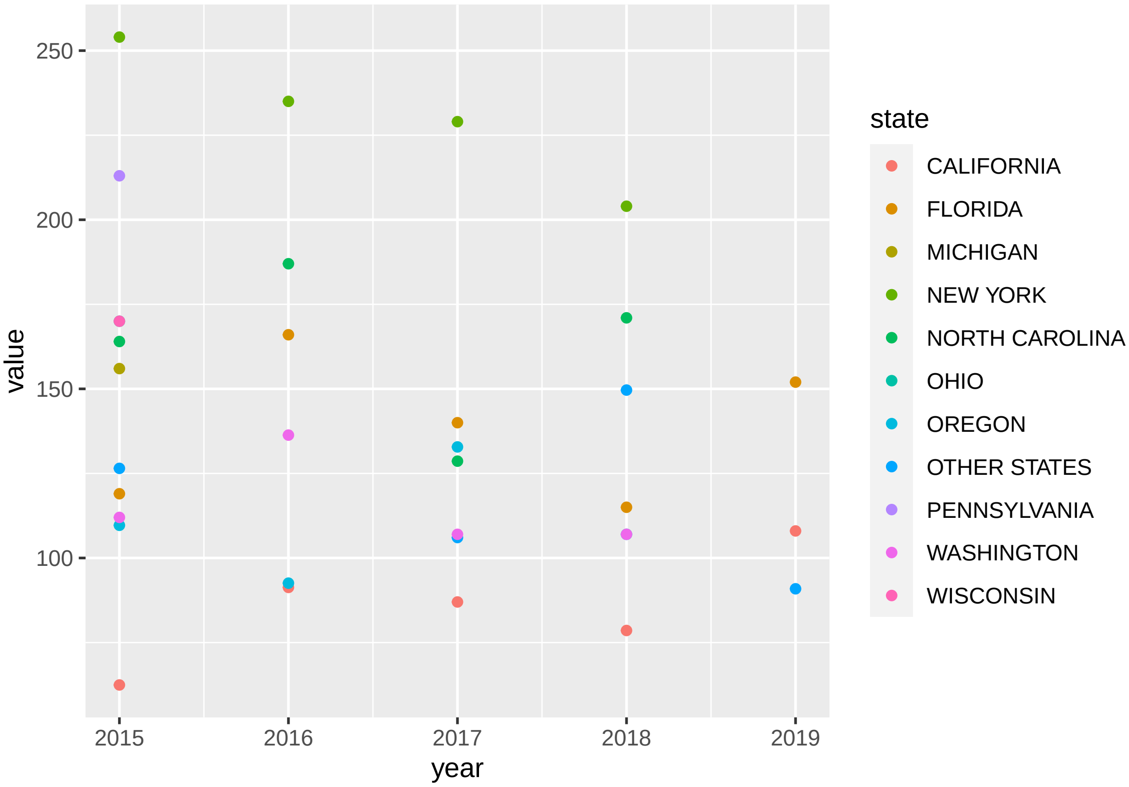
```
## `summarise()` regrouping output by 'year' (override with `.groups` argument)
```

```
summary(CWT_final)
```

```
##       year          state               value       
##  Min.   :2015   Length:35          Min.   : 62.47  
##  1st Qu.:2015   Class :character   1st Qu.:107.00  
##  Median :2016   Mode  :character   Median :130.73  
##  Mean   :2017                      Mean   :140.54  
##  3rd Qu.:2018                      3rd Qu.:169.00  
##  Max.   :2019                      Max.   :254.00  
##                                    NA's   :1       
```

```
# Making plot
a4 = ggplot(CWT_final, aes(x = year, y = value))+
  geom_point(aes(color=state))
print(a4)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



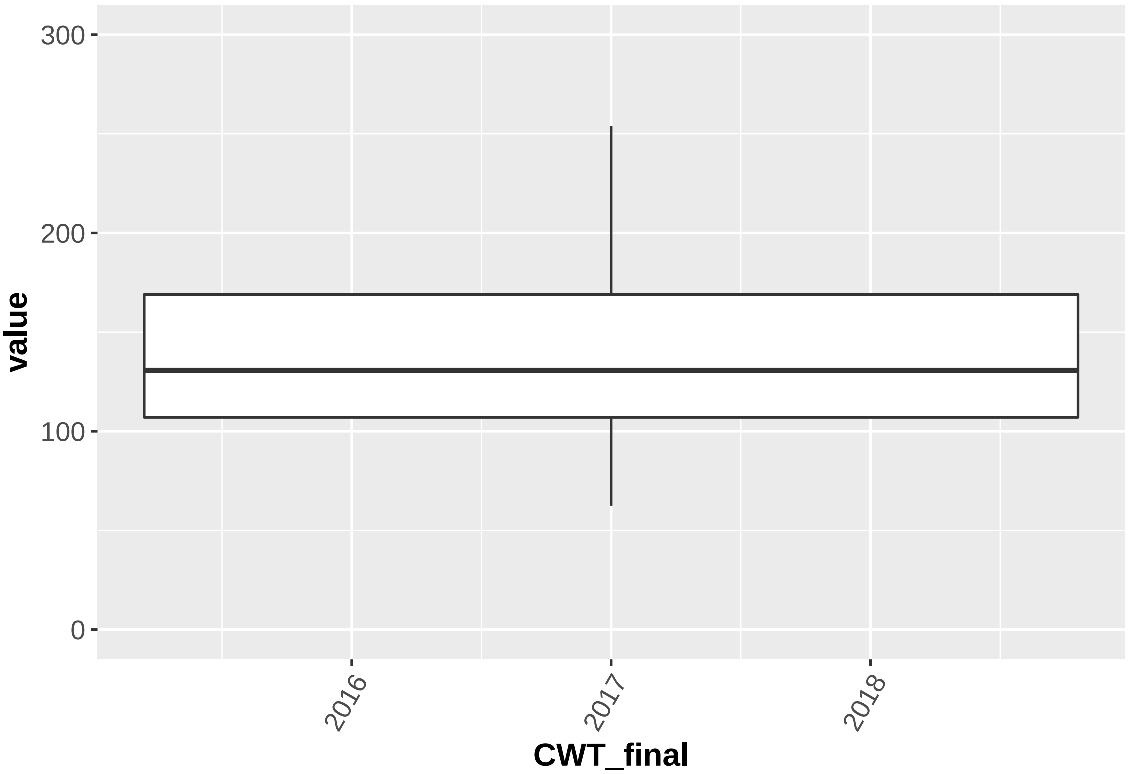
From the plot above, we can see that New York has the highest value in each year when we measured in $ / CWT.

Making boxplot of MEASURED IN $ / CWT

```r
# excluding outliers
bp2 = ggplot(CWT_final, aes(x = year, y = value))
bp2 = bp2 + geom_boxplot(outlier.colour = NA) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1),
        axis.text = element_text(size = 11),
        axis.title = element_text(size = 13, face = "bold")) +
  coord_cartesian(ylim = c(0, 300)) +
  labs(x = "CWT_final")
print(bp2)
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



From the boxplot, we can find the value for CWT_final is between 100-200

Discussion

From the analysis we did above, we can conclude that the California is a good state for buying strawberry, but further analysis is needed in better determine this conclusion, because some states have missing values for some variables. Thus, we need further analysis to find out all the states have the same measurement.

References

(^1): Berry Dataset(https://quickstats.nass.usda.gov/results/D416E96E-3D5C-324C-9334-1D38DF88FFF1)

Guided by Chenghao Meng & Yuxin Wang