Berries Project

Zhitian Liu

Project discription

In this project, we want to firstly clean the dataset “berry” and then do EDA, biuld a shiny app to access the EDA easily ### 1 Data Cleaning

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------
--------- tidyverse 1.3.0 --
```

```
## √ ggplot2 3.3.2     √ purrr   0.3.4
## √ tibble  3.0.3     √ dplyr   1.0.2
## √ tidyr   1.1.2     √ stringr 1.4.0
## √ readr   1.3.1     √ forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------
--- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

Read the data and select meaningful columns

These data were collected from the USDA database selector. It contains information of 3 kinds of berries: blueberry;raspberry and strawberry. We only look at the strawberry part in the rest of project. After look at the data, we find that the berries data had only 8 out of 21 columns containing meaningful data.

```
berry = read.csv("C:/Users/Lenovo/Desktop/berries.csv")
head(berry)
```

```
##   Program Year         Period Week.Ending Geo.Level      State State.ANSI
## 1  SURVEY 2019 MARKETING YEAR          NA     STATE CALIFORNIA          6
## 2  SURVEY 2019 MARKETING YEAR          NA     STATE CALIFORNIA          6
## 3  SURVEY 2019 MARKETING YEAR          NA     STATE CALIFORNIA          6
## 4  SURVEY 2019 MARKETING YEAR          NA     STATE CALIFORNIA          6
## 5  SURVEY 2019 MARKETING YEAR          NA     STATE CALIFORNIA          6
## 6  SURVEY 2019 MARKETING YEAR          NA     STATE CALIFORNIA          6
##   Ag.District Ag.District.Code County County.ANSI Zip.Code Region
## 1          NA               NA     NA          NA       NA     NA
## 2          NA               NA     NA          NA       NA     NA
## 3          NA               NA     NA          NA       NA     NA
## 4          NA               NA     NA          NA       NA     NA
## 5          NA               NA     NA          NA       NA     NA
## 6          NA               NA     NA          NA       NA     NA
##   watershed_code Watershed   Commodity
## 1              0        NA BLUEBERRIES
## 2              0        NA BLUEBERRIES
## 3              0        NA BLUEBERRIES
## 4              0        NA RASPBERRIES
## 5              0        NA RASPBERRIES
## 6              0        NA 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 CV....
## 1   NOT SPECIFIED  2.85     NA
## 2   NOT SPECIFIED  3.56     NA
## 3   NOT SPECIFIED  0.29     NA
## 4   NOT SPECIFIED  2.69     NA
## 5   NOT SPECIFIED   (D)     NA
## 6   NOT SPECIFIED   (D)     NA
```

```
meaningful_berry = berry %>% select(Year,
                                    Period,
                                    State,
                                    Commodity,
                                    Data.Item,
                                    Domain,
                                    Domain.Category,
                                    Value)
```

Look at the column-“value”

The “value” column contains the value of berries (the kind of value is described in the column “data.item”), This should be a numeric list.

```
meaningful_berry$Value=as.numeric(meaningful_berry$Value)
```

```
## Warning: 强制改变过程中产生了NA
```

```
str(meaningful_berry$Value)
```

```
##  num [1:13238] 2.85 3.56 0.29 2.69 NA NA 108 NA NA 2.64 ...
```

Look at the Strawberries

Now we concentrate on the strawberry observations, use filter function to extract the strawberry data.

```
strawberry=meaningful_berry %>% filter(Commodity=="STRAWBERRIES")
#And we only use the data with period "year"
#strawberry=strawberry %>% filter(Period=="YEAR")
# Since there are a lot of NA in the "value" column, we choose to delete these observations.
strawberry=strawberry %>% drop_na()
# Summary of the dataset
summary(strawberry)
```

```
##       Year         Period             State            Commodity
##  Min.   :2015   Length:1229        Length:1229        Length:1229
##  1st Qu.:2016   Class :character   Class :character   Class :character
##  Median :2018   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2018
##  3rd Qu.:2019
##  Max.   :2019
##   Data.Item            Domain          Domain.Category        Value
##  Length:1229        Length:1229        Length:1229        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
```

Cleaning “Data Item” column

The most difficult part to clean is “data column”, it includes lot of information. The most important information, in my opinion, is the measurement of value.

```
straw_item = strawberry$Data.Item
# Replace "-" with ","
straw_item = gsub(" - ", ",", straw_item)

#extract the measurement of value in each observations
strawberry$measurement = str_extract_all(straw_item, "MEASURED IN.*[^, /AVG]|ACRES.*")
strawberry$measurement = str_replace(strawberry$measurement, ",", "")
strawberry$measurement = trimws(strawberry$measurement)
head(strawberry)
```

```
##   Year         Period        State    Commodity
## 1 2019 MARKETING YEAR   CALIFORNIA STRAWBERRIES
## 2 2019 MARKETING YEAR      FLORIDA STRAWBERRIES
## 3 2019 MARKETING YEAR OTHER STATES STRAWBERRIES
## 4 2019 MARKETING YEAR OTHER STATES STRAWBERRIES
## 5 2019           YEAR   CALIFORNIA STRAWBERRIES
## 6 2019           YEAR   CALIFORNIA STRAWBERRIES
##                                                    Data.Item
## 1               STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT
## 2               STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT
## 3 STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / CWT
## 4   STRAWBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN $ / CWT
## 5                     STRAWBERRIES - YIELD, MEASURED IN CWT / ACRE
## 6             STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB
##               Domain                                    Domain.Category Value
## 1              TOTAL                                      NOT SPECIFIED 108.0
## 2              TOTAL                                      NOT SPECIFIED 152.0
## 3              TOTAL                                      NOT SPECIFIED 129.0
## 4              TOTAL                                      NOT SPECIFIED  52.8
## 5              TOTAL                                      NOT SPECIFIED 580.0
## 6 CHEMICAL, FUNGICIDE CHEMICAL, FUNGICIDE: (BORAX DECAHYDRATE = 11102) 300.0
##             measurement
## 1    MEASURED IN $ / CWT
## 2    MEASURED IN $ / CWT
## 3    MEASURED IN $ / CWT
## 4    MEASURED IN $ / CWT
## 5 MEASURED IN CWT / ACRE
## 6         MEASURED IN LB
```

Finalise the data which we would use to do the eda

Now, in the strawberry dataset, we have the state column, the year column, also we have the value and measurement for each observations. I decided to use these 4 variables to do a exploratory data analysis.

```
strawberry = strawberry %>% select(Year, State, Value, measurement)
head(strawberry)
```

```
##   Year        State Value            measurement
## 1 2019   CALIFORNIA 108.0    MEASURED IN $ / CWT
## 2 2019      FLORIDA 152.0    MEASURED IN $ / CWT
## 3 2019 OTHER STATES 129.0    MEASURED IN $ / CWT
## 4 2019 OTHER STATES  52.8    MEASURED IN $ / CWT
## 5 2019   CALIFORNIA 580.0 MEASURED IN CWT / ACRE
## 6 2019   CALIFORNIA 300.0         MEASURED IN LB
```

2 Exploratory Data Analysis

Check how many kinds of measurements are there in the strawberry dataset.

Now we can do some EDA using strawberry data. Fistly, we can check how many kinds of measurements are there in the strawberry dataset.

```
strawberry_measure <- strawberry %>%
  group_by(measurement) %>%
  summarize(count = n(),)
```

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

```
view(strawberry_measure)
```

Hence, We can notice that there are 13 different measurements.

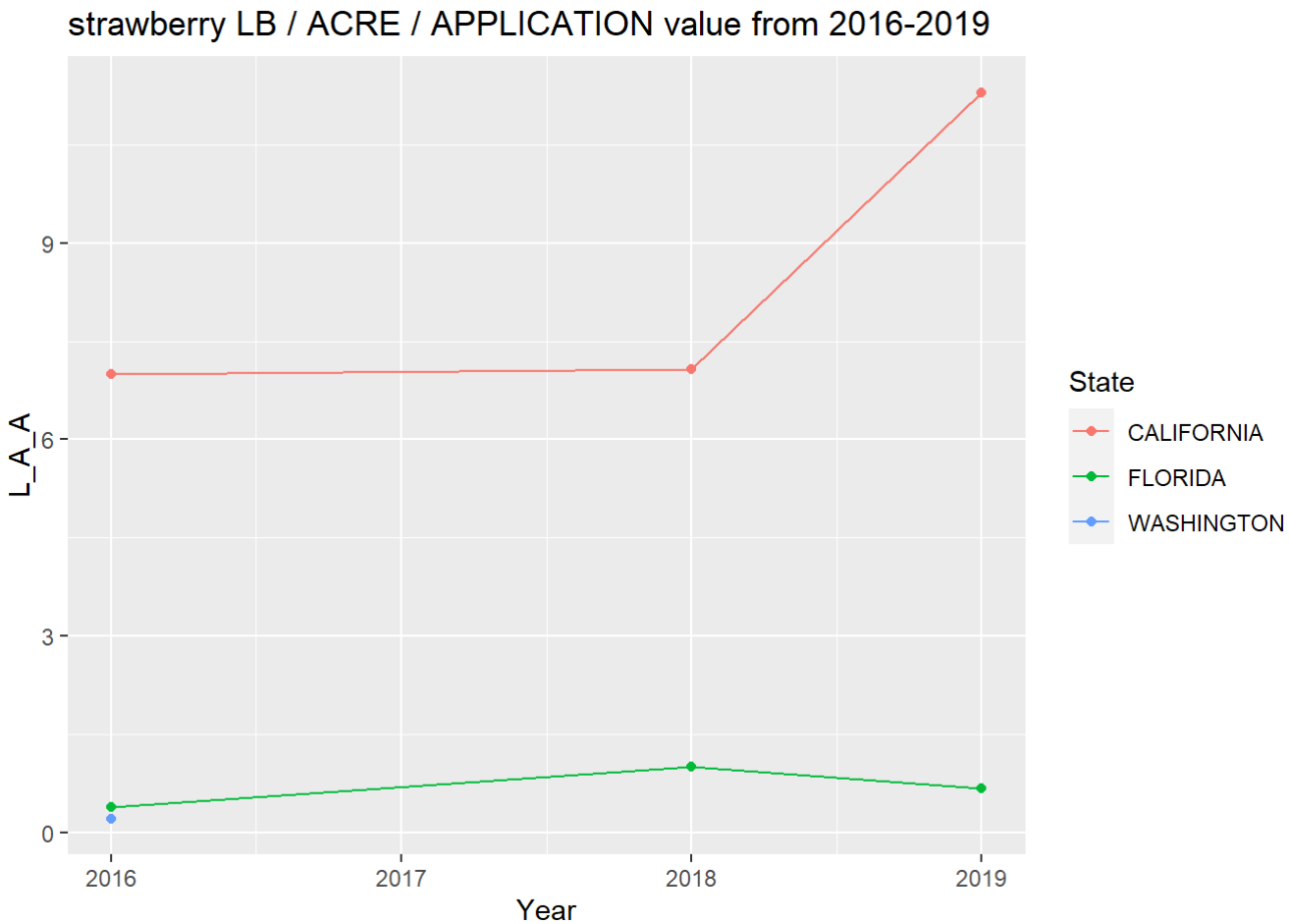
Value measured in LB / ACRE

Firstly we can look at the mean values which are MEASURED IN LB / ACRE / APPLICATION and MEASURED IN LB / ACRE / YEAR in different state through time. We generate figures to show the trend of mean values changing with years in different states.

```
par(mfrow=c(1,2))
data_L_A_A = filter(strawberry, measurement == "MEASURED IN LB / ACRE / APPLICATION")
data_L_A_A = data_L_A_A %>% group_by(State, Year) %>% summarise(L_A_A = mean(Value))
```

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

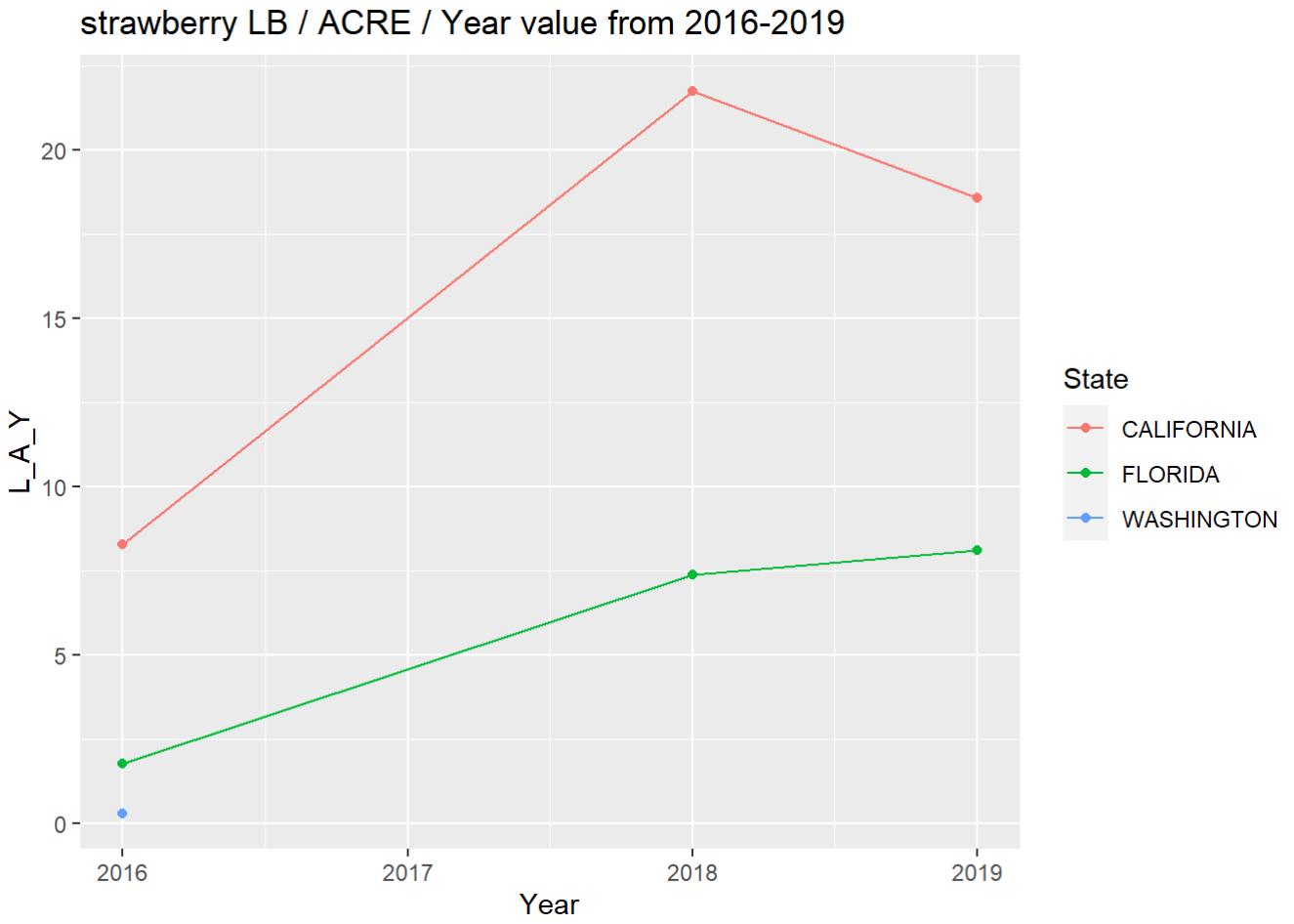
```
ggplot(data = data_L_A_A, mapping = aes(x = Year, y = L_A_A)) + geom_point(aes(color =
                                                                                 State)) + geom
_line(aes(color = State)) +
  ggtitle("strawberry LB / ACRE / APPLICATION value from 2016-2019")
```



```
data_L_A_Y = filter(strawberry, measurement == "MEASURED IN LB / ACRE / YEAR")
data_L_A_Y = data_L_A_Y %>% group_by(State, Year) %>% summarise(L_A_Y = mean(Value))
```

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

```
ggplot(data = data_L_A_Y, mapping = aes(x = Year, y = L_A_Y)) + geom_point(aes(color =
                                                                                 State)) + geom
_line(aes(color = State)) +
  ggtitle("strawberry LB / ACRE / Year value from 2016-2019")
```



From the figures, we know that the production efficiency of strawberry in California is much more higher than production efficiency in Washington. Note that we miss the data in 2017. We only have the data of Washington in 2016, it has the lowest production rate in 2016 among all 3 states.

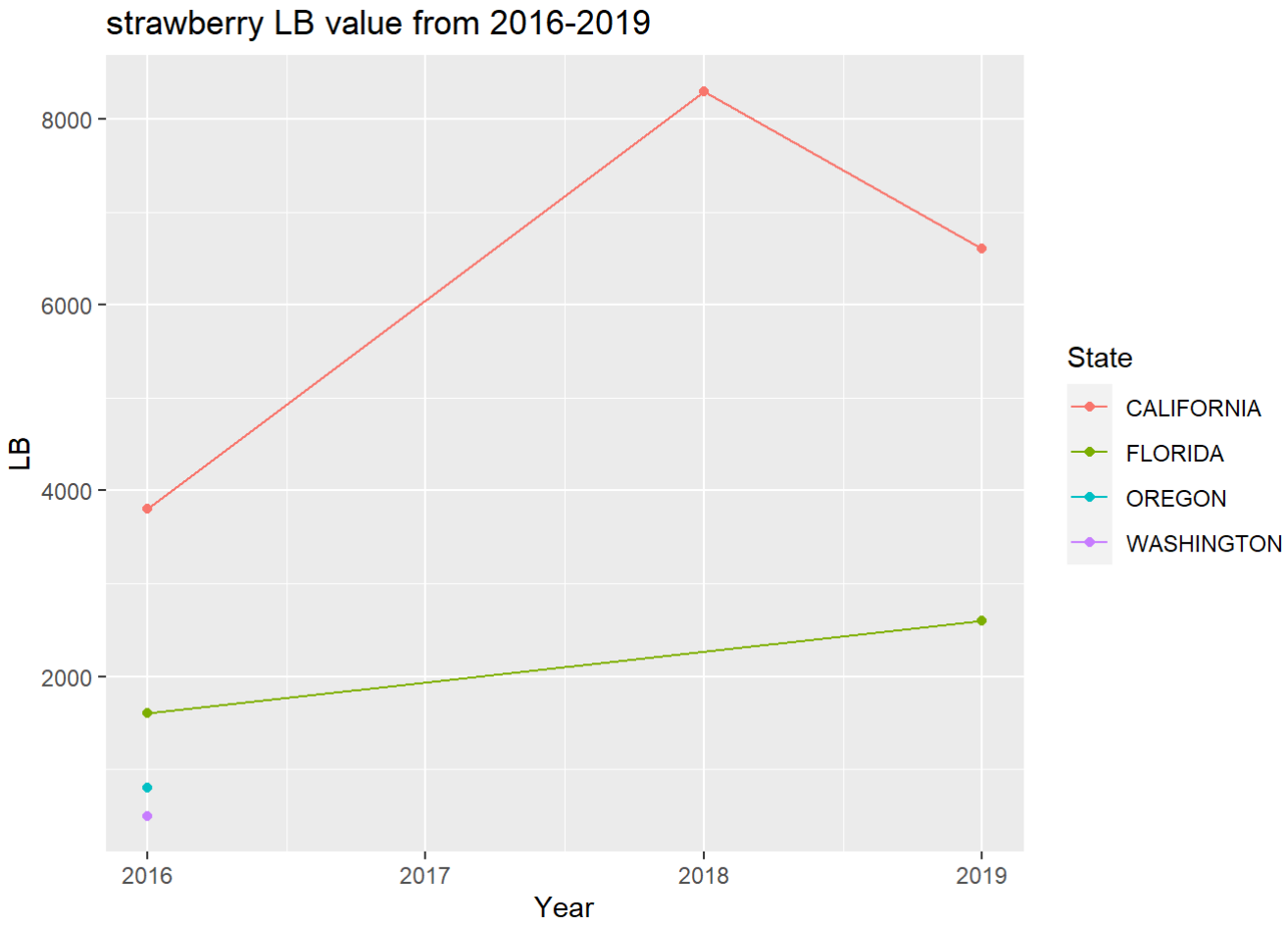
Value measured in LB

After knowing the efficiency of strawberry production, we now look at the production capacity. that is, the observations measured in LB.

```
data_LB = filter(strawberry, measurement == "MEASURED IN LB")
#We sum up the values in different observations to get the total strawberry production in each
 state each year.
data_LB = data_LB %>% group_by(State, Year) %>% summarise(LB = sum(Value))
```

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

```
ggplot(data = data_LB, mapping = aes(x = Year, y = LB)) + geom_point(aes(color =
                                                                           State)) + geom
_line(aes(color = State)) +
  ggtitle("strawberry LB value from 2016-2019")
```



The above figure shows that the production output of strawberry in Washington is still the lowest in 2016. The amount of strawberry produced in California is still the highest among all states.

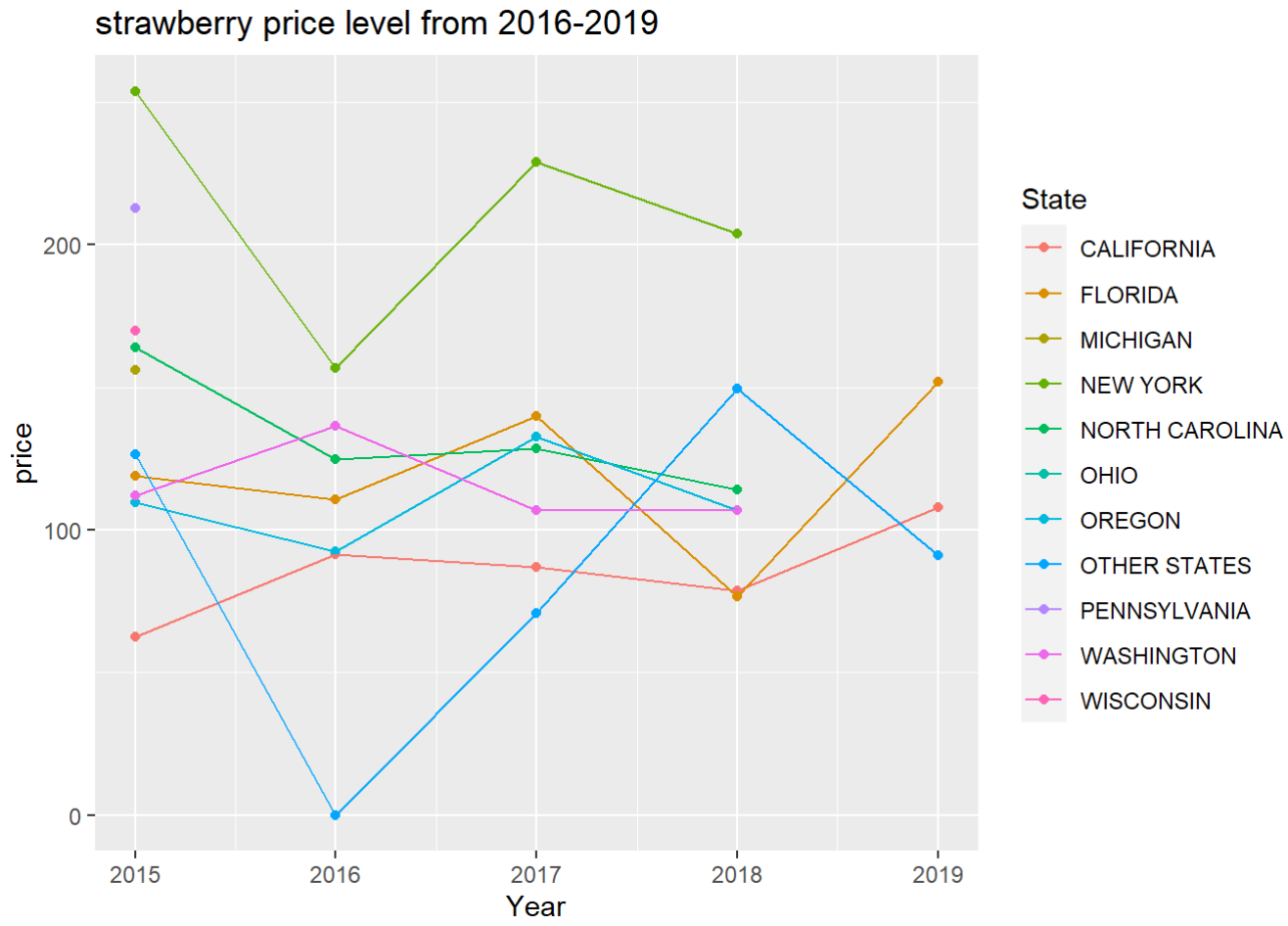
The price value measured in $/CWT

Now let's look at the price level of strawberry in different states.

```
data_price = filter(strawberry, measurement == "MEASURED IN $ / CWT")
data_price = data_price %>% group_by(State, Year) %>% summarise(price = mean(Value))
```

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

```
ggplot(data = data_price, mapping = aes(x = Year, y = price)) + geom_point(aes(color =
                                                                                 State)) + geom
_line(aes(color = State)) +
  ggtitle("strawberry price level from 2016-2019")
```



From the above figure, we can get the information that the strawberry in New York is most expensive. California's strawberry was kept in a very low price level through all these years.

3 Conclusion

After doing some EDA, we know that California is the best place to buy strawberries. Because it has higher production capacity and lower price level.

reference

Some thoughts of data cleaning and EDA come from my classmates, and I agree with it.