Report

Jingyi Niu

October 19, 2020

Data Manipulation

The raw data contains three different products: STRAWBERRY, RASPBERRY, STRAWBERRY. Our interest is focus on the RASPBERRY data. Therefore, the `filter` function in `dplyr` package is been used to extract all data about RASPBERRY.

The main difficult in data manipulation is that some key information are combined togather in a single variable, for example: **Data Item** contains the type information, and the type information contains three key informations (ACRES HARVESTED, PRODUCTION, YIELD).

To extrac these informations, the `separate` function in package `tidyr` is been used for many times.

Furthermore, the variable `Value` in this dataset is not stored in numeric for two reasons: 1. the value is in the format of `123,456,789`, the comma needs to be remove. 2. there exists `(D)` values.

To do our data analysis, the variable `Value` need to be converted into numeric. The `str_replace_all` function in the `stringr` package is been used to do this conversion.

Now the data is been manipulated into a tidy format, like this:

Year	State	ACRES_HARVESTED	PRODUCTION	YIELD
2015	CALIFORNIA	11400	672080000	19400
2015	OREGON	2200	28165000	4270
2015	WASHINGTON	9350	221775000	7870
2016	CALIFORNIA	8450	508840000	20100
2016	OREGON	1750	25620000	4870

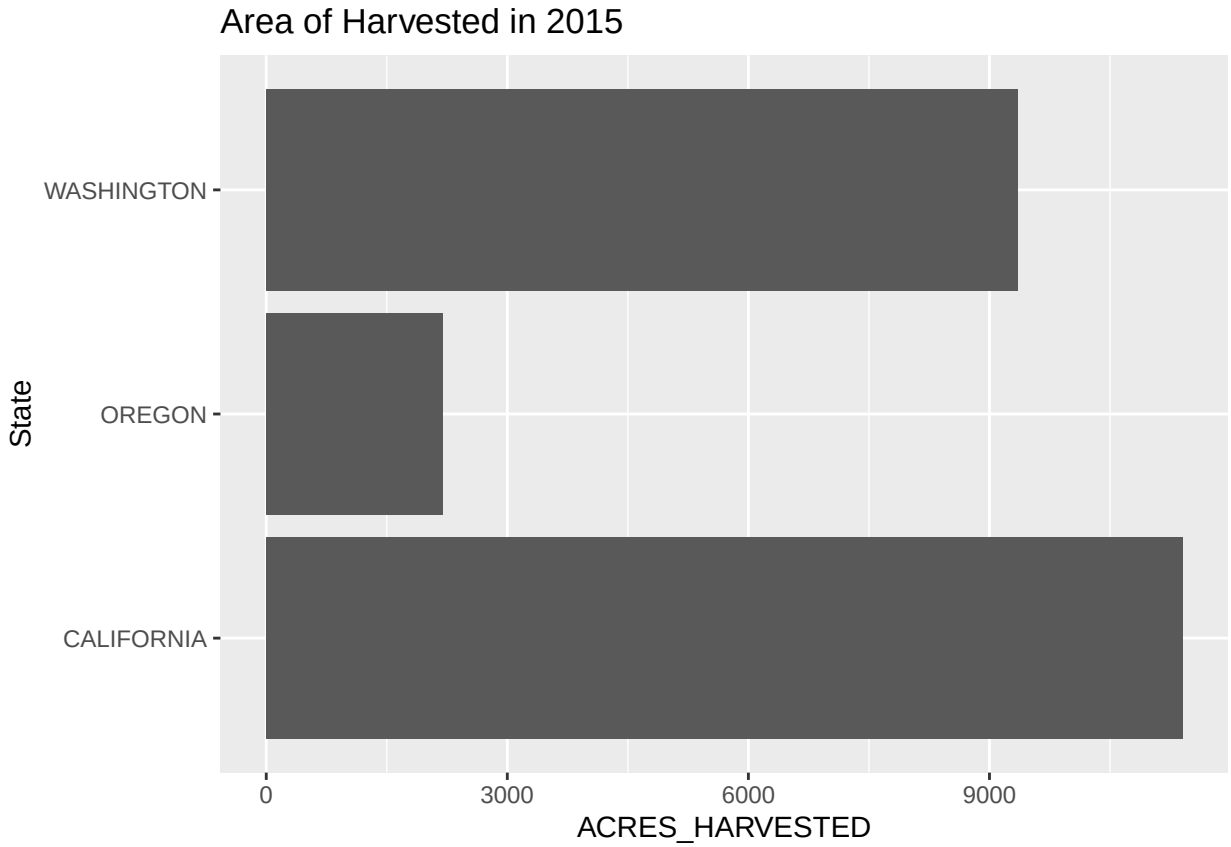
It can be noticed that there exists many NA in the variable `YIELD` and `ACRES HARVESTED`.

The analysis of Harvested Areas

In this part, we focus on the Harvested Areas(in Acres).

The Harvested Areas of states in 2015 are listed below:

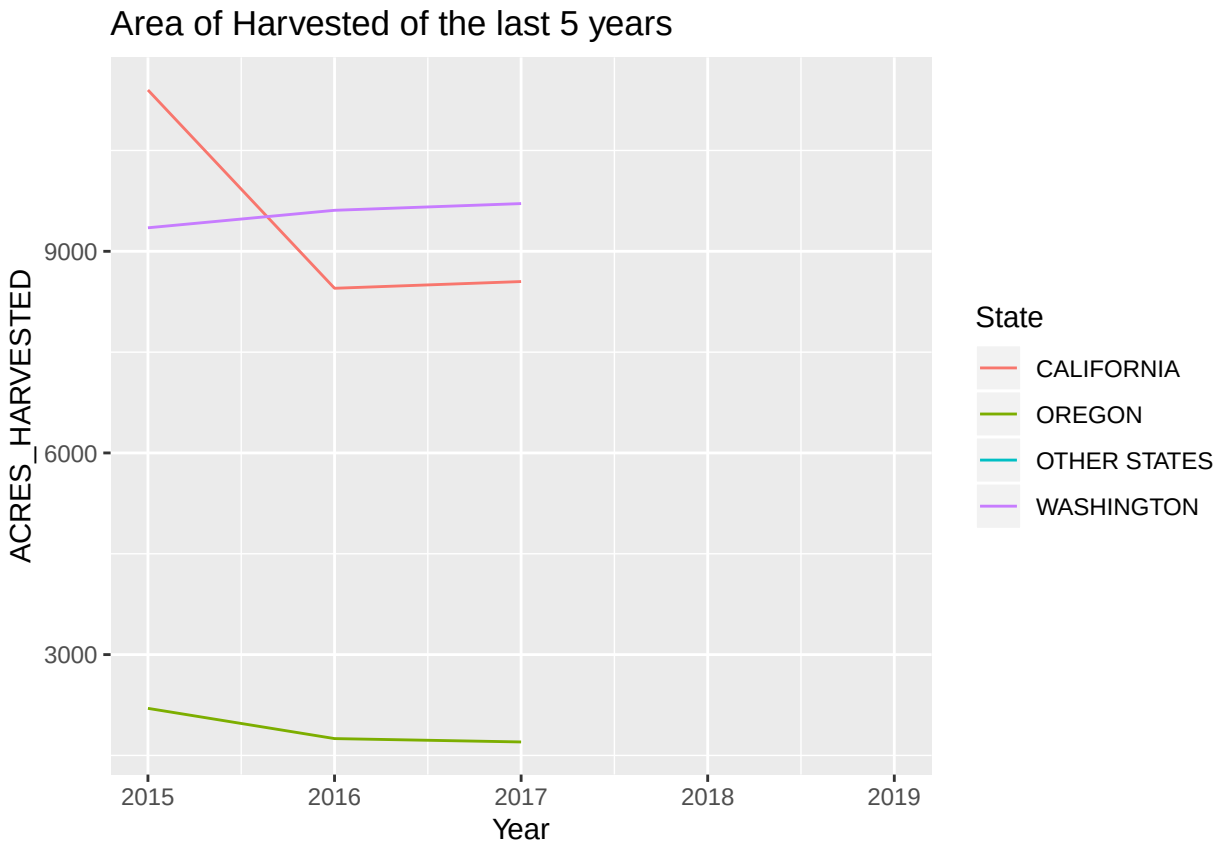
Year	State	ACRES_HARVESTED
2015	CALIFORNIA	11400
2015	WASHINGTON	9350
2015	OREGON	2200



The states which has the top 2 Harvested Areas are listed below:

Year	State	ACRES_HARVESTED
2015	CALIFORNIA	11400
2015	WASHINGTON	9350
2016	CALIFORNIA	8450
2016	WASHINGTON	9610
2017	CALIFORNIA	8550
2017	WASHINGTON	9710

The Harvested Areas of the last 5 years:



From the above table and graph, it can be seen that :

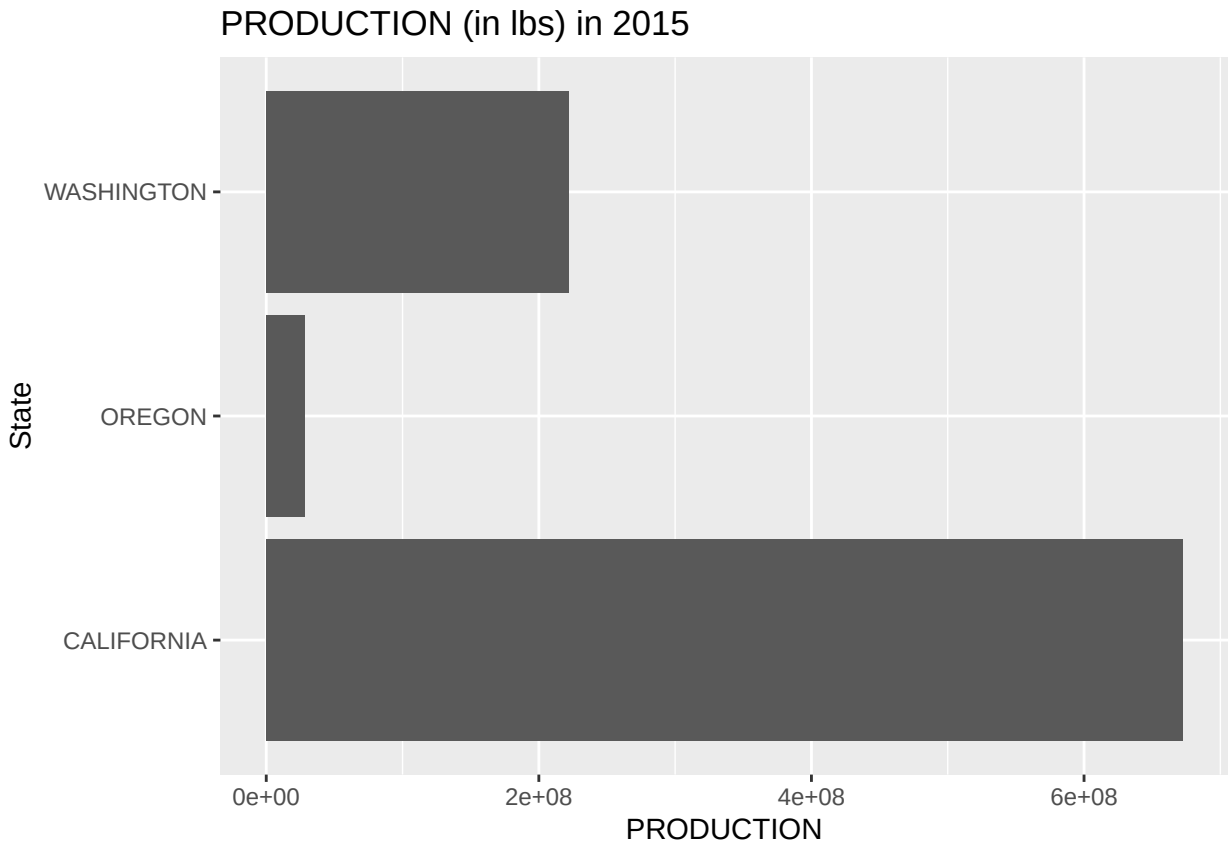
1. The Area of Harvested of 2 States(CALIFORNIA, WASHINGTON) are much higher than OREGON
2. Data missing for year 2018 and 2019.
3. The Area of Harvested of California was decreasing.

The analysis of PRODUCTION

In this part, we focus on the PRODUCTION(in lbs).

The PRODUCTION of states in 2015 are listed below:

Year	State	PRODUCTION
2015	CALIFORNIA	672080000
2015	WASHINGTON	221775000
2015	OREGON	28165000



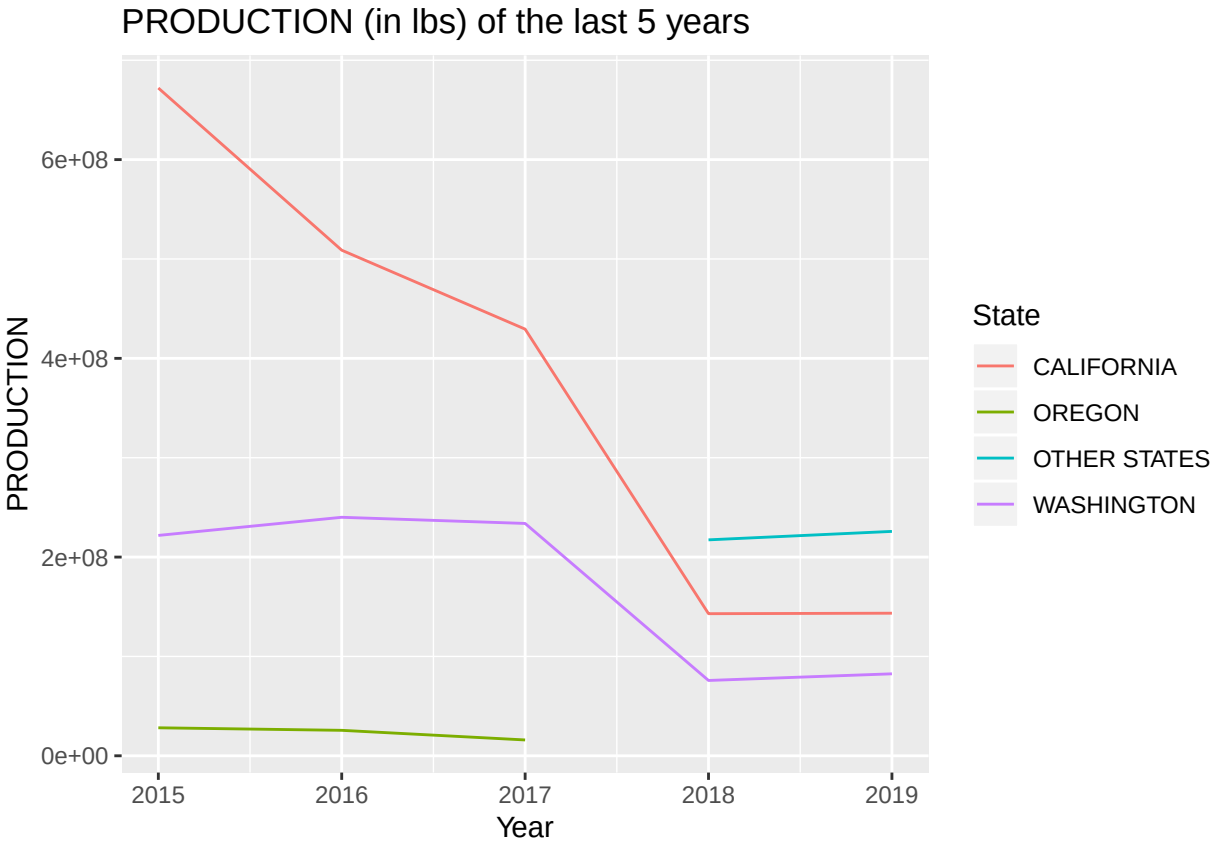
The states which has the top 2 PRODUCTION are listed below:

```
# PRODUCTION (in lbs)  top 3
dat2 %>%
  group_by(Year) %>%
  top_n(2, PRODUCTION) %>%
  select(-YIELD, -ACRES_HARVESTED) %>% knitr::kable()
```

Year	State	PRODUCTION
2015	CALIFORNIA	672080000
2015	WASHINGTON	221775000
2016	CALIFORNIA	508840000
2016	WASHINGTON	240030000
2017	CALIFORNIA	429390000
2017	WASHINGTON	233840000
2018	CALIFORNIA	143000000
2018	OTHER STATES	217320000
2019	CALIFORNIA	143500000
2019	OTHER STATES	225840000

The PRODUCTION of the last 5 years:

```
dat2 %>%
  ggplot(aes(x = Year, y = PRODUCTION, color = State)) +
  geom_line() +
  labs(title = "PRODUCTION (in lbs) of the last 5 years")
```



From the above table and graph, it can be seen that :

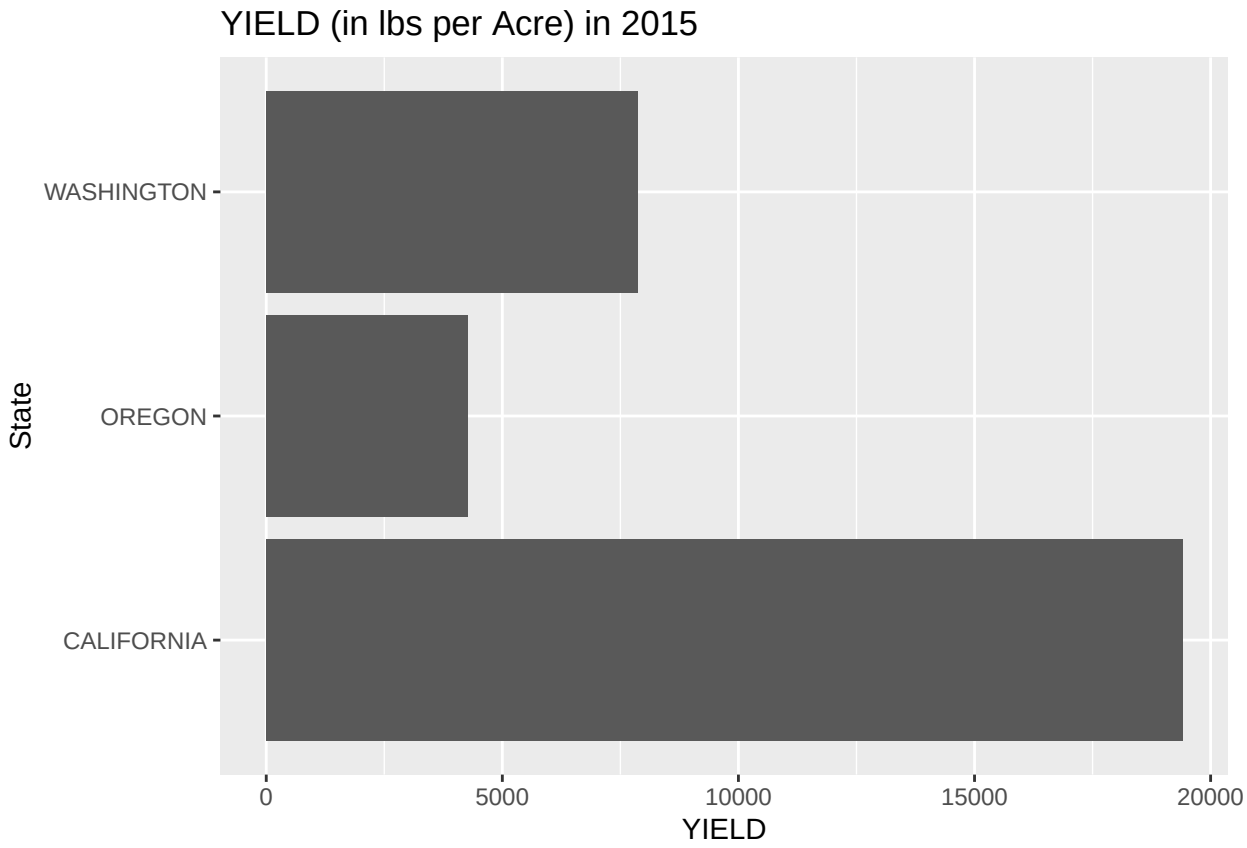
1. The Production of 2 States(CALIFORNIA, WASHINGTON) are much higher than OREGON
2. Data missing for Oregon after year 2017.
3. The Production of 2 States are keep decreasing for the last 5 years.

The analysis of YIELD

In this part, we focus on the YIELD(in lbs per Acre).

The YIELD of states in 2015 are listed below:

Year	State	YIELD
2015	CALIFORNIA	19400
2015	WASHINGTON	7870
2015	OREGON	4270

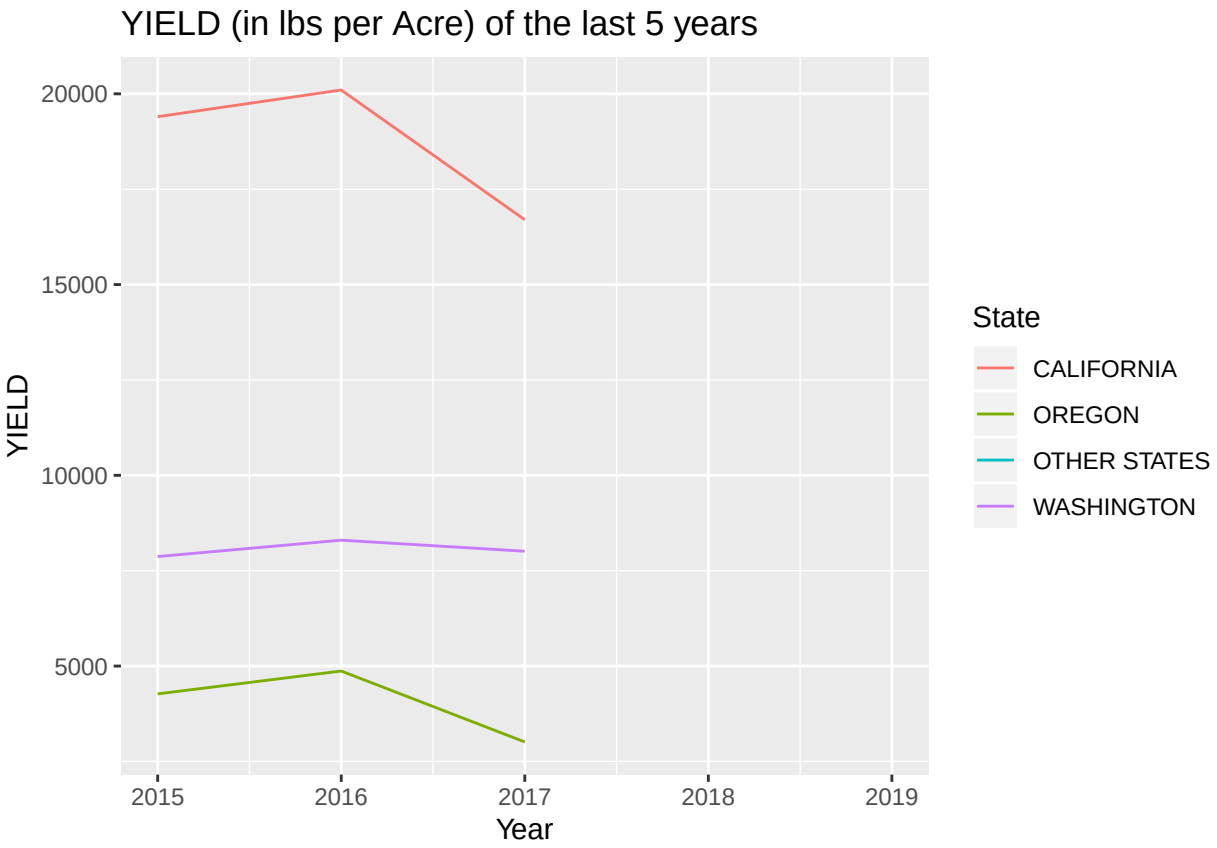


The states which has the top 3 YIELD are listed below:

Year	State	YIELD
2015	CALIFORNIA	19400
2015	WASHINGTON	7870
2016	CALIFORNIA	20100
2016	WASHINGTON	8300
2017	CALIFORNIA	16700
2017	WASHINGTON	8010

The YIELD of the last 5 years:

```
# plot
dat2 %>%
  ggplot(aes(x = Year, y = YIELD, color = State)) +
  geom_line() +
  labs(title = "YIELD (in lbs per Acre) of the last 5 years")
```



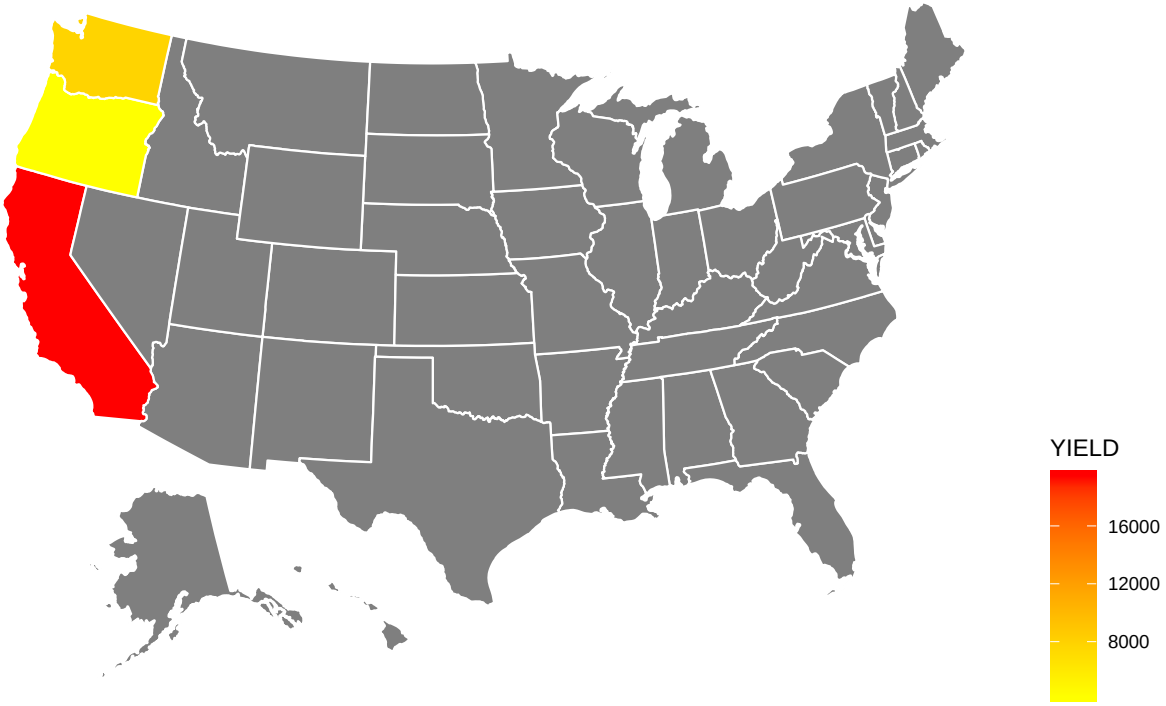
From the above table and graph, it can be seen that :

1. The YIELD of California was decreasing from 2016 to 2017.
2. Data missing for year 2018 and 2019.
3. The YIELD of California is much higher than the rest.

It is very interesting that 3 States (CALIFORNIA, OREGON, WASHINGTON) has much higher YIELD(in lbs per Acre) than the other states. Further study may be needed.

YIELD on the US Map

In this part, we plot the YIELD of 3 States on the map of United States.



It can be seen that the State that have much higher YIELD is located in the middle of **West Coast**. The weather may be a great reason of the high YIELD.