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	UTILIZED_PRODUCTION_IN_DOLLAR
2015	ALABAMA	410	540000	1317	960000
2015	ARKANSAS	280	520000	1857	1474000
2015	CALIFORNIA	6200	64100000	10339	120303000
2015	FLORIDA	5500	25300000	4600	82267000
2015	GEORGIA	17200	85000000	4942	104440000

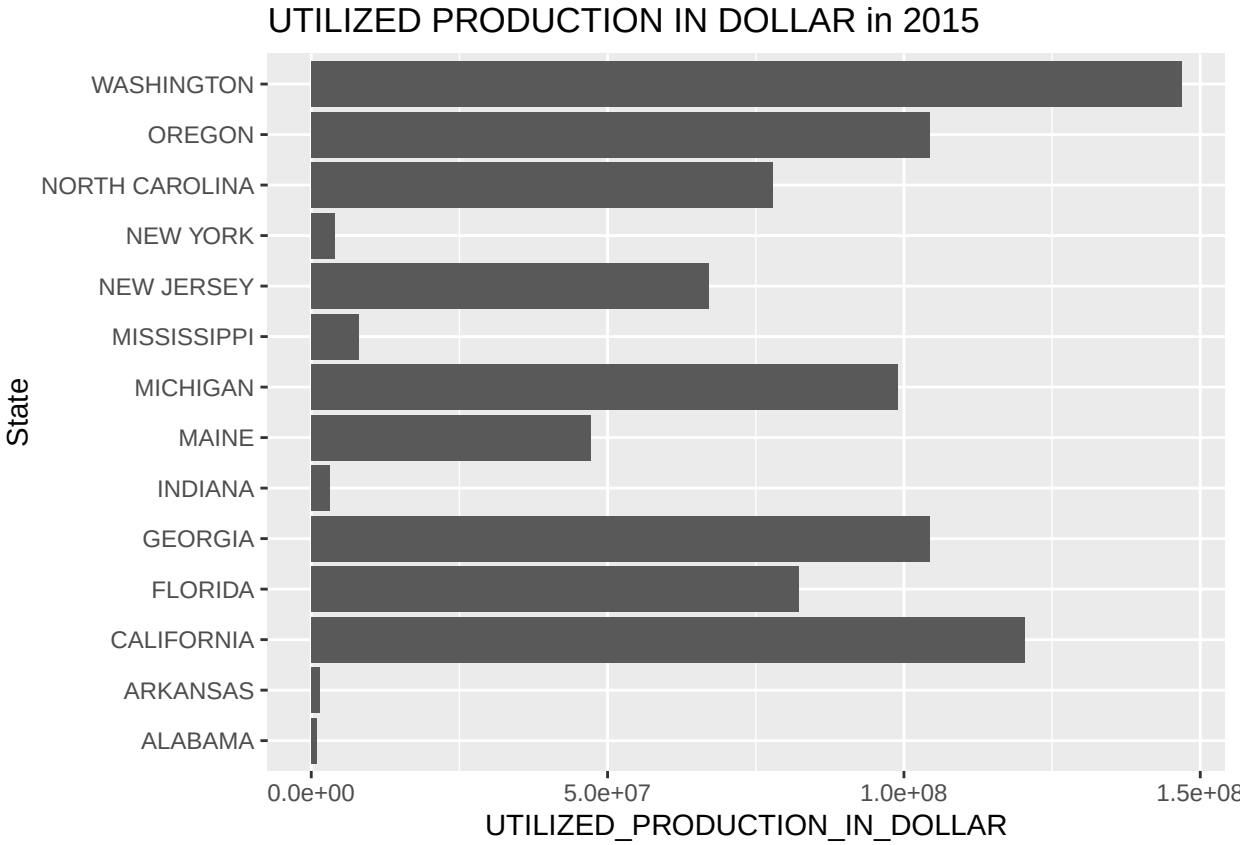
The analysis of "UTILIZED - PRODUCTION MEASURED IN $"

In this part, we focus on the "UTILIZED - PRODUCTION MEASURED IN $".

The "UTILIZED_PRODUCTION_IN_DOLLAR" of states in 2015 are listed below:

Year	State	UTILIZED_PRODUCTION_IN_DOLLAR
2015	WASHINGTON	146847000
2015	CALIFORNIA	120303000
2015	GEORGIA	104440000
2015	OREGON	104307000
2015	MICHIGAN	98925000
2015	FLORIDA	82267000
2015	NORTH CAROLINA	77910000
2015	NEW JERSEY	67064000
2015	MAINE	47180000

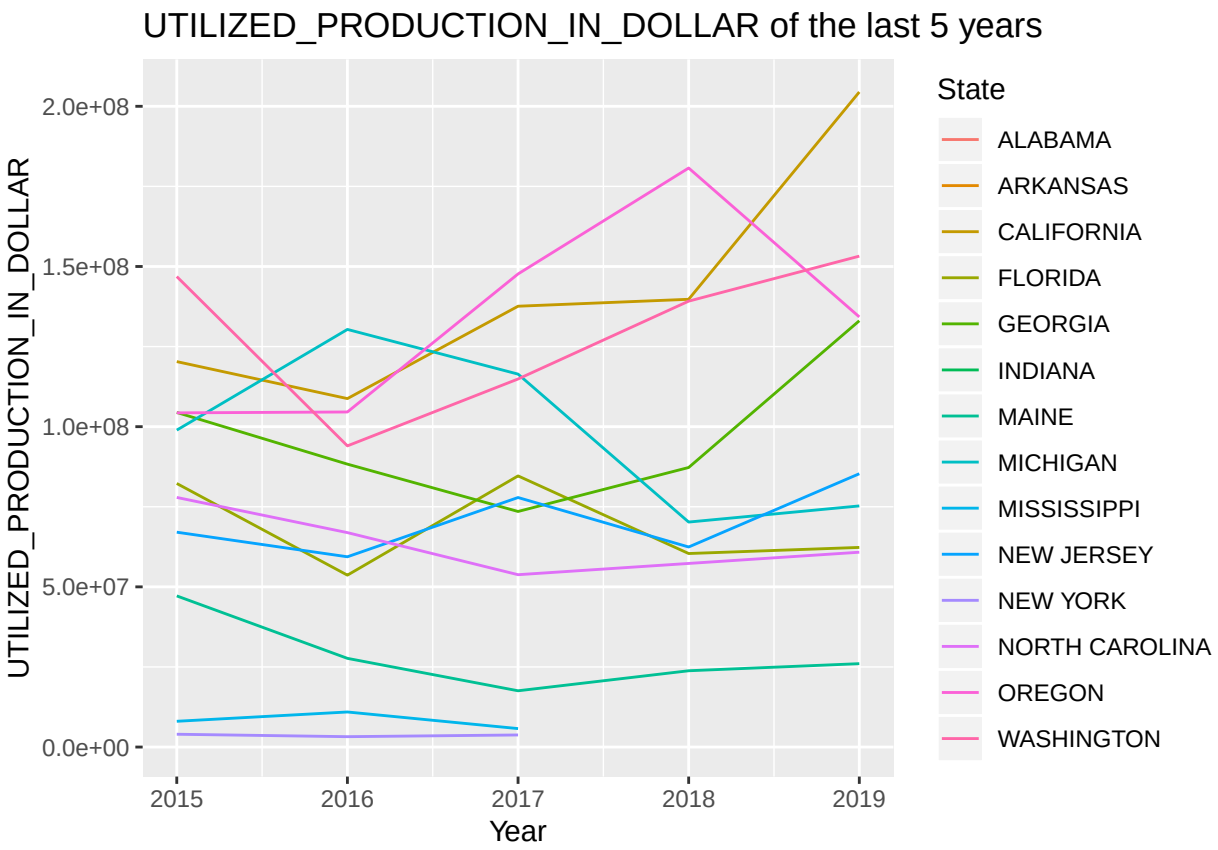
Year	State	UTILIZED_PRODUCTION_IN_DOLLAR
2015	MISSISSIPPI	8040000
2015	NEW YORK	3995000
2015	INDIANA	3048000
2015	ARKANSAS	1474000
2015	ALABAMA	960000



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

Year	State	UTILIZED_PRODUCTION_IN_DOLLAR
2015	CALIFORNIA	120303000
2015	WASHINGTON	146847000
2016	CALIFORNIA	108765000
2016	MICHIGAN	130350000
2017	CALIFORNIA	137596000
2017	OREGON	147665000
2018	CALIFORNIA	139755000
2018	OREGON	180730000
2019	CALIFORNIA	204460000
2019	WASHINGTON	153224000

The UTILIZED PRODUCTION IN DOLLAR of the last 5 years:



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

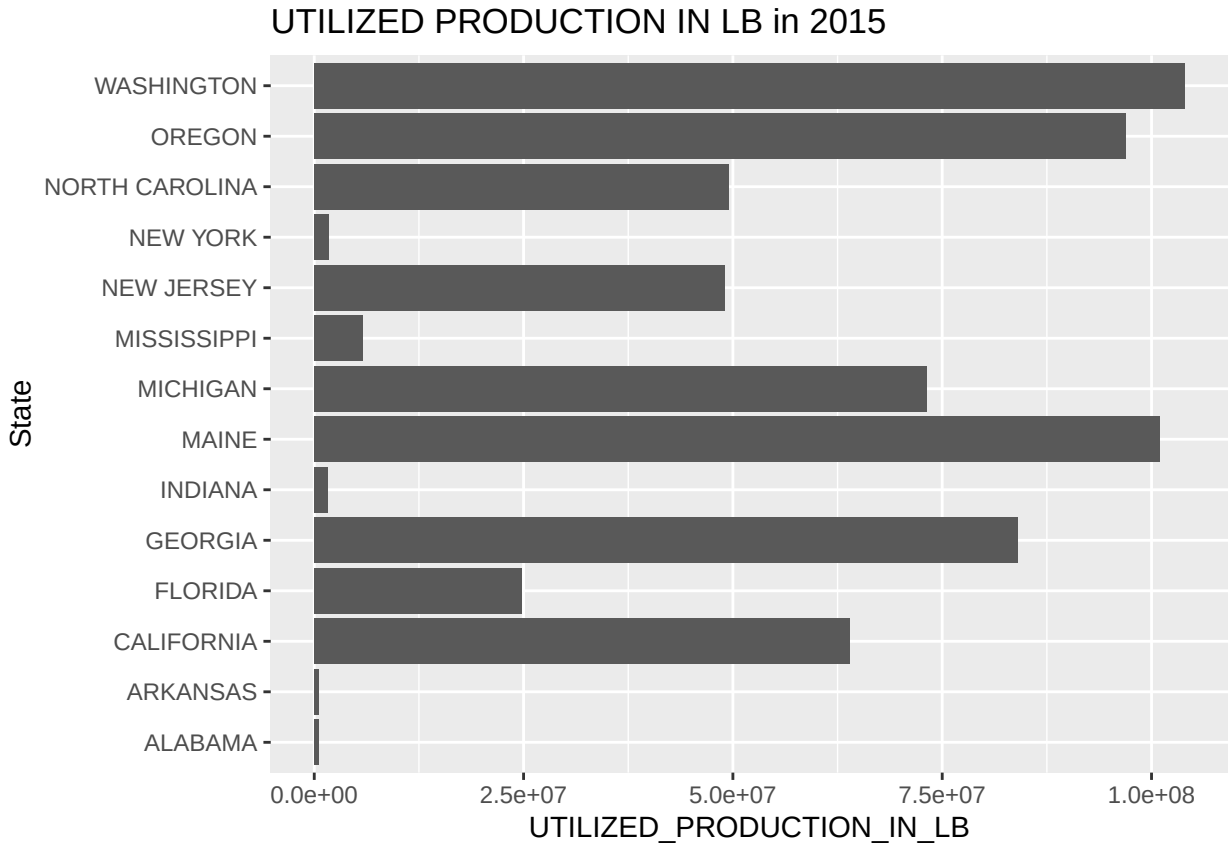
1. The UTILIZED PRODUCTION (in Dollar) of many States were increasing for the last 5 years.
2. Data missing for year 2018 and 2019 for two States.
3. The UTILIZED PRODUCTION (in Dollar) of WASHINGTON was decreasing in 2019.

The analysis of "UTILIZED - PRODUCTION MEASURED IN LB"

In this part, we focus on the "UTILIZED - PRODUCTION MEASURED IN LB".

The "UTILIZED_PRODUCTION_IN_LB" of states in 2015 are listed below:

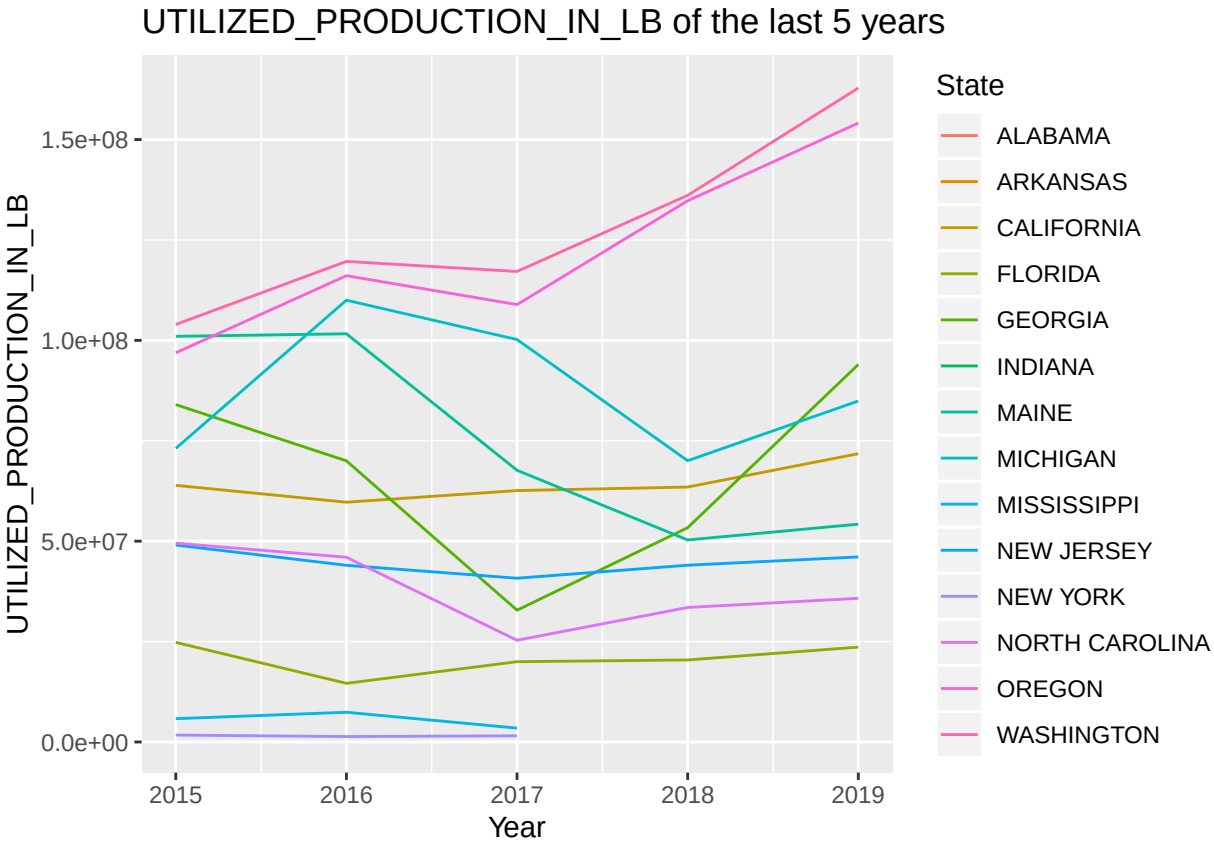
Year	State	UTILIZED_PRODUCTION_IN_LB
2015	WASHINGTON	103950000
2015	MAINE	101000000
2015	OREGON	96900000
2015	GEORGIA	84000000
2015	MICHIGAN	73100000
2015	CALIFORNIA	63900000
2015	NORTH CAROLINA	49500000
2015	NEW JERSEY	49030000
2015	FLORIDA	24800000
2015	MISSISSIPPI	5800000
2015	NEW YORK	1720000
2015	INDIANA	1600000
2015	ARKANSAS	520000
2015	ALABAMA	500000



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

Year	State	UTILIZED_PRODUCTION_IN_LB
2015	MAINE	101000000
2015	WASHINGTON	103950000
2016	OREGON	116100000
2016	WASHINGTON	119650000
2017	OREGON	108900000
2017	WASHINGTON	117150000
2018	OREGON	134750000
2018	WASHINGTON	136100000
2019	OREGON	154100000
2019	WASHINGTON	162830000

The UTILIZED PRODUCTION IN LB of the last 5 years:



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

1. The UTILIZED PRODUCTION (in LB) of many States were increasing for the last 5 years.
2. Data missing for year 2018 and 2019 for two States.
3. The UTILIZED PRODUCTION (in LB) of WASHINGTON and OREGON was decreasing in much fast than other States.

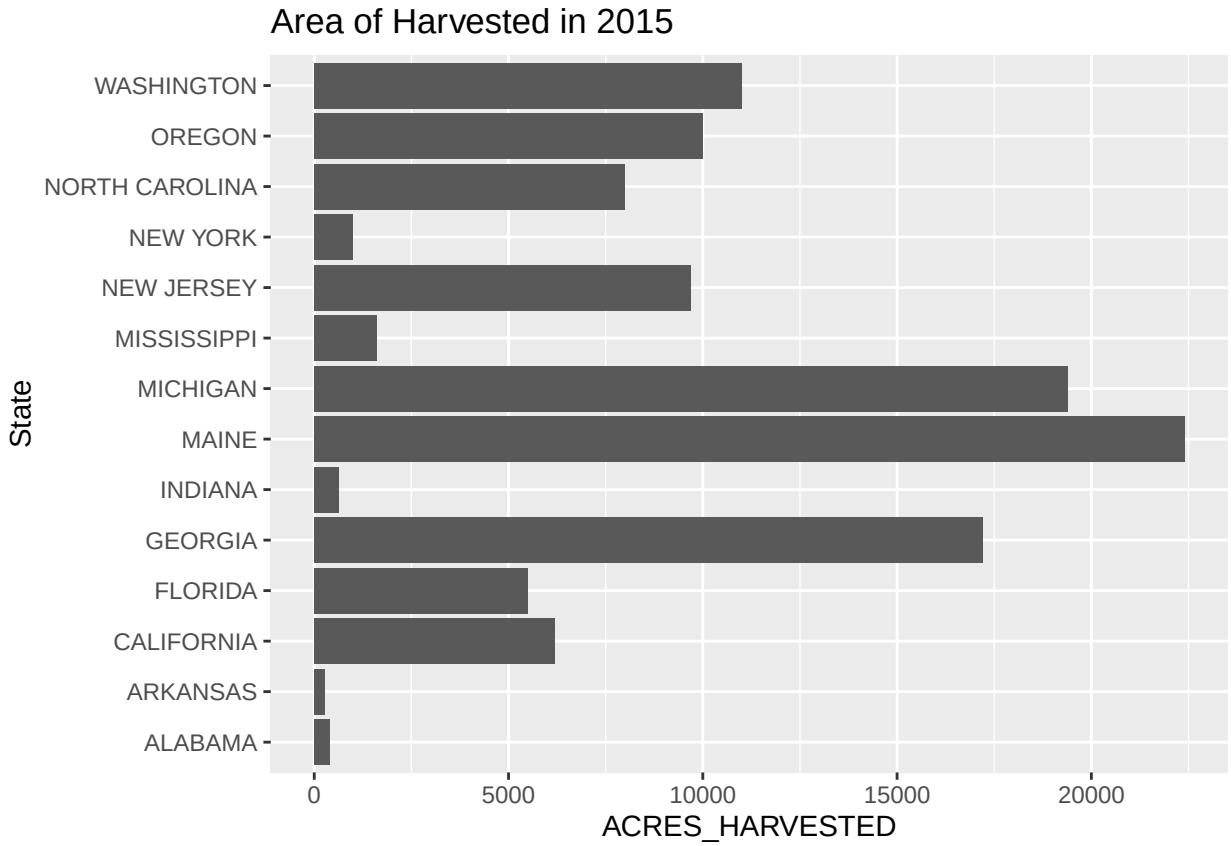
The analysis of Harvested Areas

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	MAINE	22400
2015	MICHIGAN	19400
2015	GEORGIA	17200
2015	WASHINGTON	11000
2015	OREGON	10000
2015	NEW JERSEY	9700
2015	NORTH CAROLINA	8000
2015	CALIFORNIA	6200
2015	FLORIDA	5500
2015	MISSISSIPPI	1600
2015	NEW YORK	1000
2015	INDIANA	630
2015	ALABAMA	410

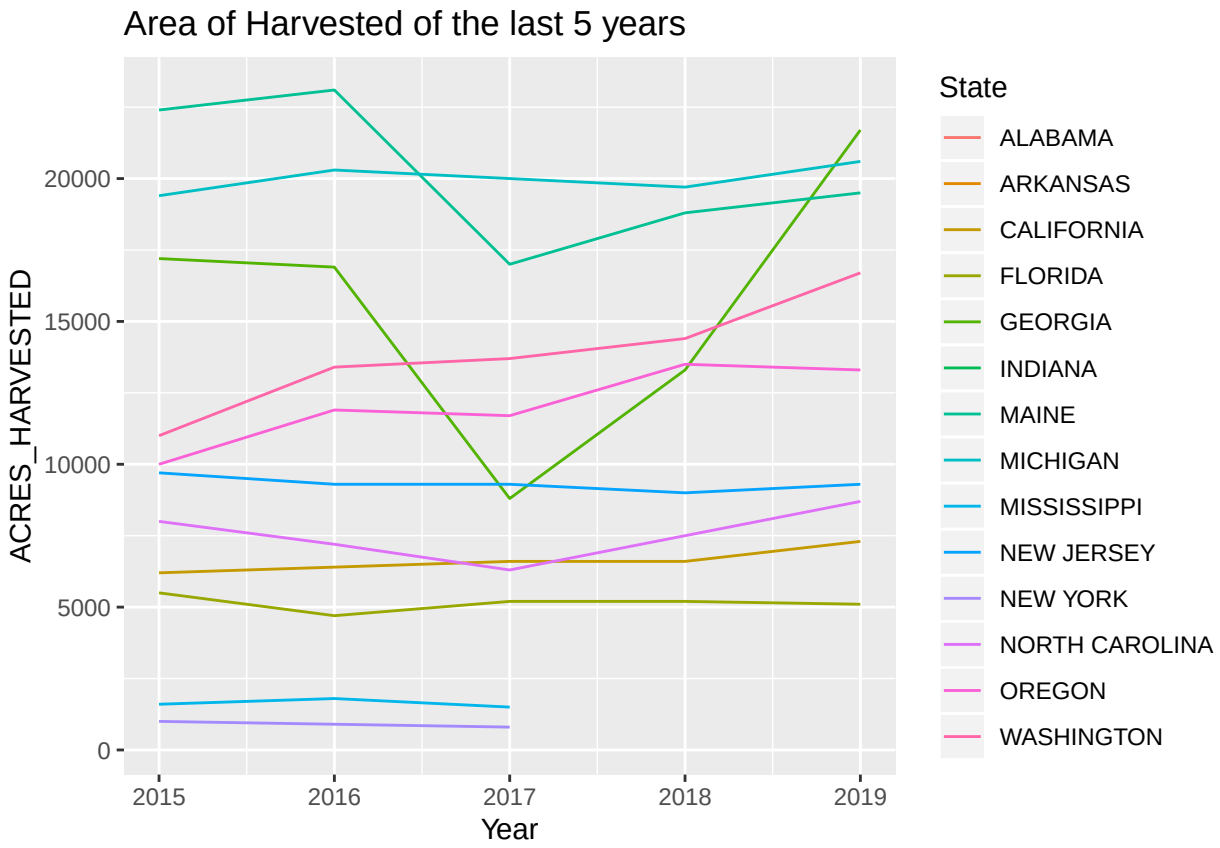
Year	State	ACRES_HARVESTED
2015	ARKANSAS	280



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

Year	State	ACRES_HARVESTED
2015	MAINE	22400
2015	MICHIGAN	19400
2016	MAINE	23100
2016	MICHIGAN	20300
2017	MAINE	17000
2017	MICHIGAN	20000
2018	MAINE	18800
2018	MICHIGAN	19700
2019	GEORGIA	21700
2019	MICHIGAN	20600

The Harvested Areas of the last 5 years:



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

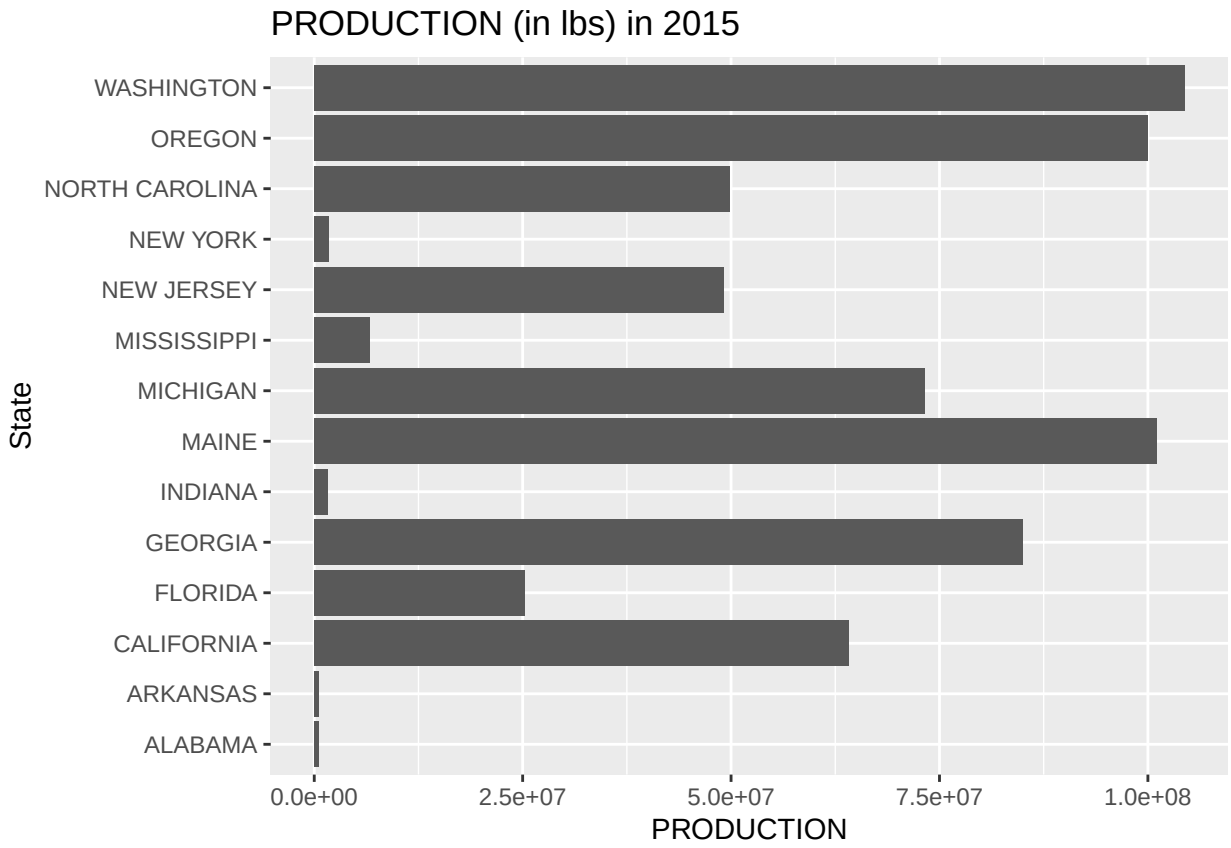
1. The Area of Harvested of 2 States(INDIANA, GEORGIA) were decreasing rapidly and then recoverd.
2. Data missing for year 2018 and 2019 for two States.
3. The Area of Harvested of California was decreasing.

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	WASHINGTON	104400000
2015	MAINE	101110000
2015	OREGON	100000000
2015	GEORGIA	85000000
2015	MICHIGAN	73200000
2015	CALIFORNIA	64100000
2015	NORTH CAROLINA	49900000
2015	NEW JERSEY	49080000
2015	FLORIDA	25300000
2015	MISSISSIPPI	6700000
2015	NEW YORK	1790000
2015	INDIANA	1610000
2015	ALABAMA	540000
2015	ARKANSAS	520000



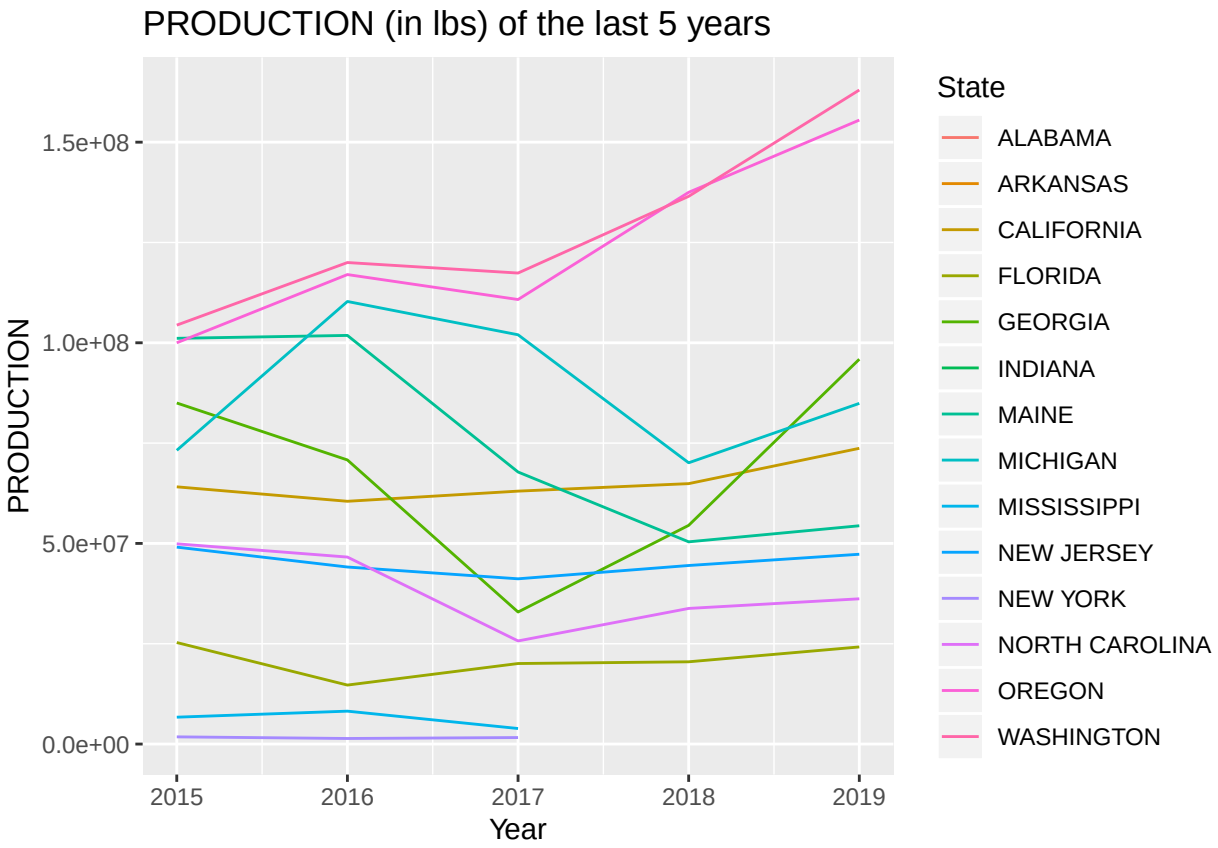
The states which has the top 2 PRODUCTION are listed below:

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

Year	State	PRODUCTION
2015	MAINE	101110000
2015	WASHINGTON	104400000
2016	OREGON	117000000
2016	WASHINGTON	120000000
2017	OREGON	110780000
2017	WASHINGTON	117380000
2018	OREGON	137500000
2018	WASHINGTON	136500000
2019	OREGON	155500000
2019	WASHINGTON	163000000

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 :

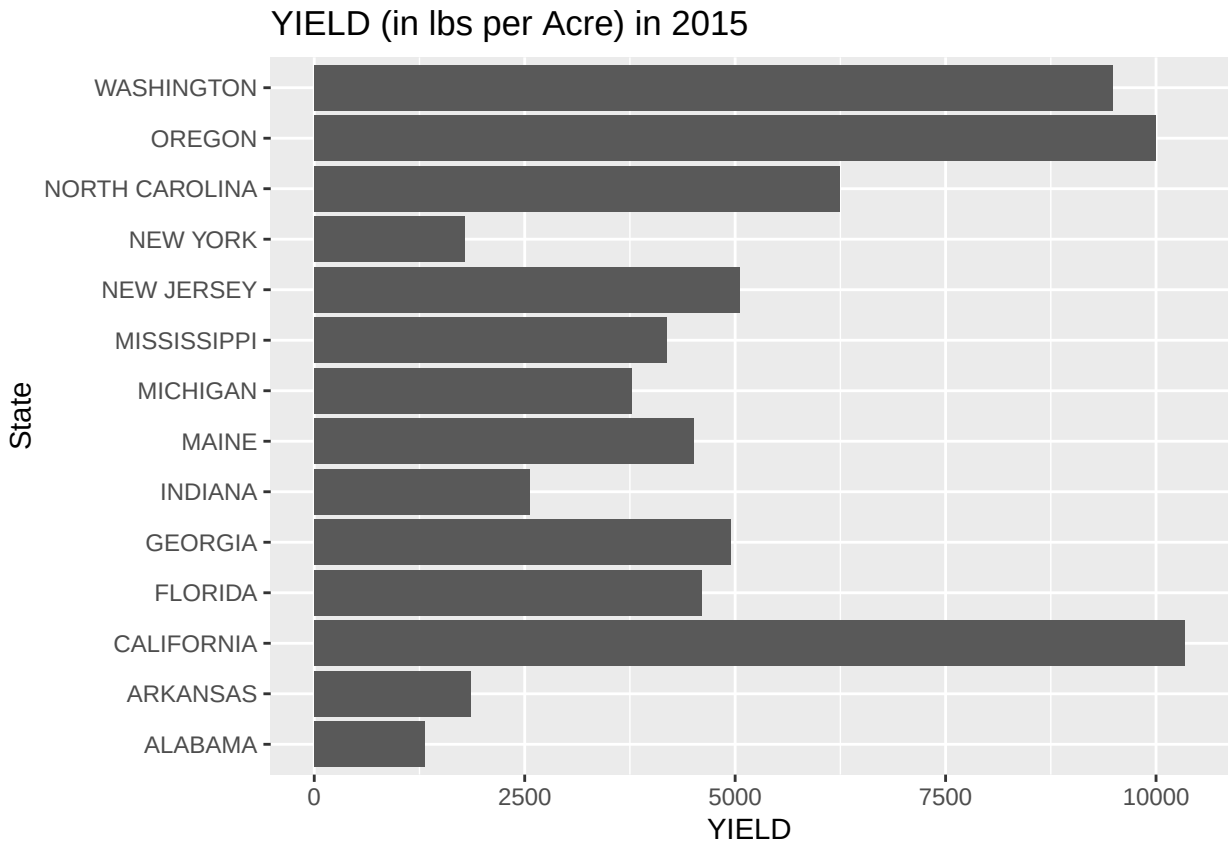
1. The Production of 2 States(CALIFORNIA, WASHINGTON) are much higher than OREGON
2. Data missing for Oregon after year 2017 for two States.

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	10339
2015	OREGON	10000
2015	WASHINGTON	9491
2015	NORTH CAROLINA	6238
2015	NEW JERSEY	5060
2015	GEORGIA	4942
2015	FLORIDA	4600
2015	MAINE	4514
2015	MISSISSIPPI	4188
2015	MICHIGAN	3773
2015	INDIANA	2556
2015	ARKANSAS	1857
2015	NEW YORK	1790
2015	ALABAMA	1317

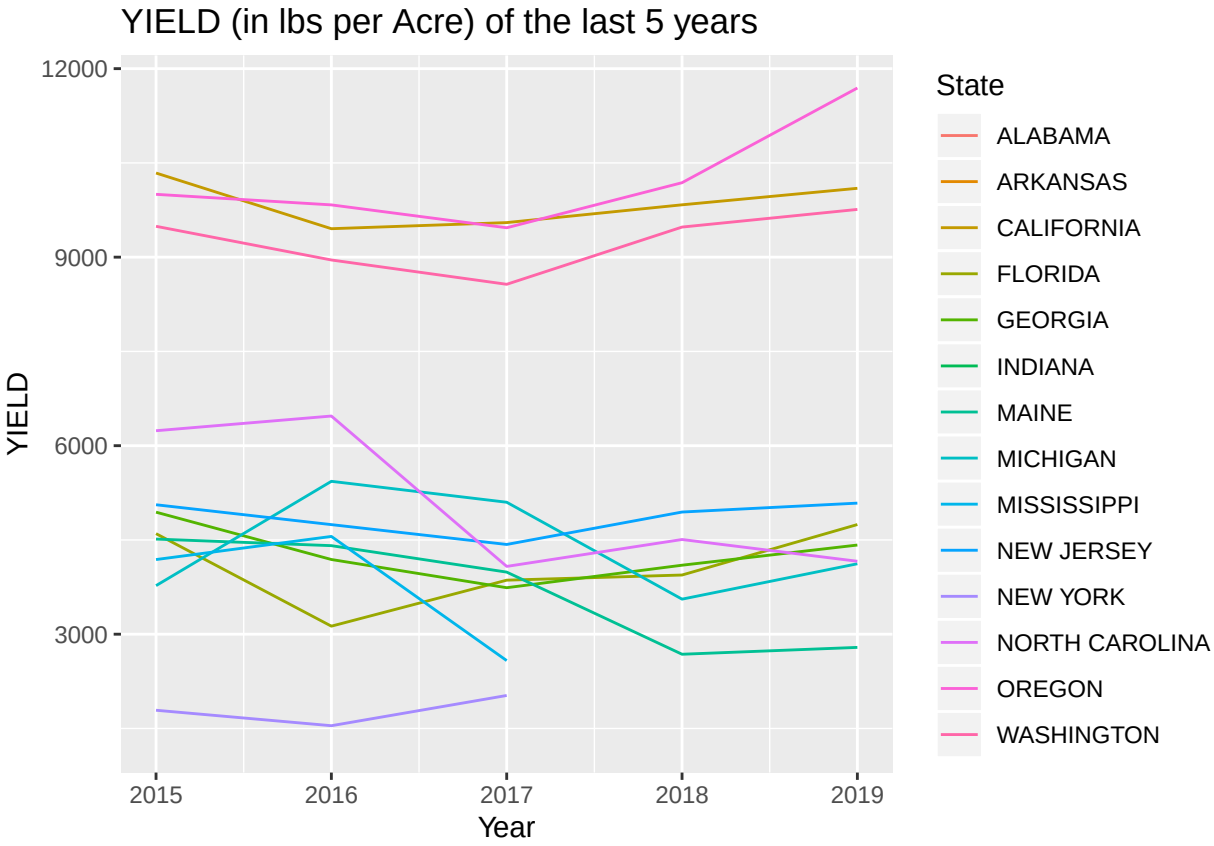


The states which has the top 3 YIELD are listed below:

Year	State	YIELD
2015	CALIFORNIA	10339
2015	OREGON	10000
2016	CALIFORNIA	9453
2016	OREGON	9832
2017	CALIFORNIA	9550
2017	OREGON	9468
2018	CALIFORNIA	9833
2018	OREGON	10185
2019	CALIFORNIA	10096
2019	OREGON	11692

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")
```



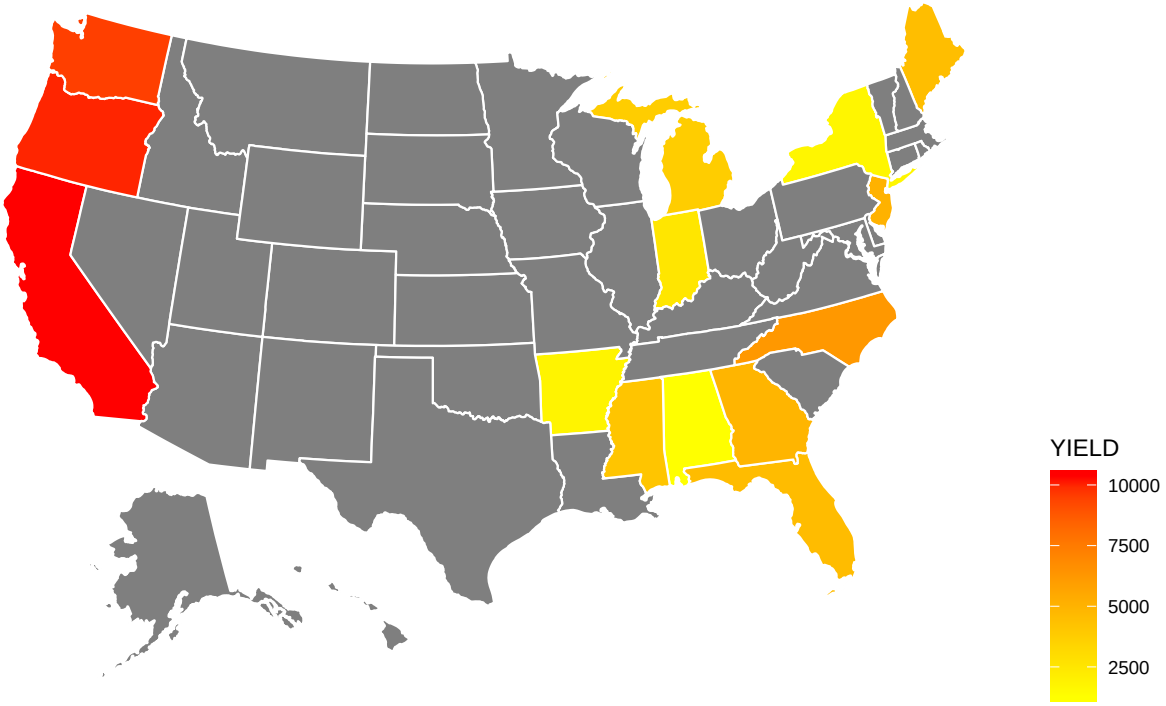
From the above table and graph, it can be seen that :

1. The YIELD of many States was decreasing from 2016 to 2017.
2. Data missing for year 2018 and 2019 for two States.
3. The YIELD of California, ALABAMA and ARKANSAS 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.