berry EDA

Xiaozhou Lu

2020.10.19

1 Import the berry data

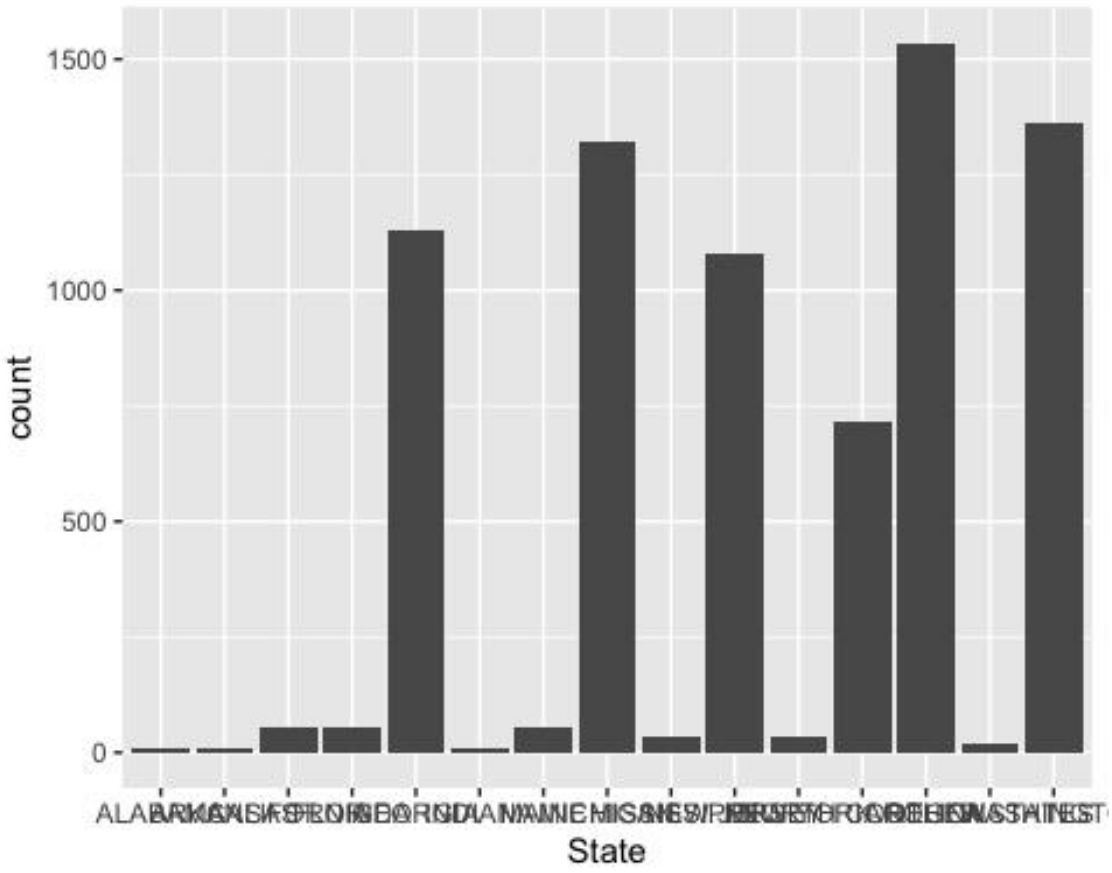
The data has been installed from USDA database selector https://quickstats.nass.usda.gov. We have get the data cleaned and decide to use data with year period and blueberry commodity to conduct our analysis. For the numbers that are not avaliable we represent those with 0. This might not be rigorous but is necessary for analysising.

```
berry<- read.csv("/Users/angryboats/Desktop/MA 615/berries/blueberry.csv", header=T)
berry$Value<- as.numeric((gsub(",","",berry$Value)))
```

2 Yields in Each State

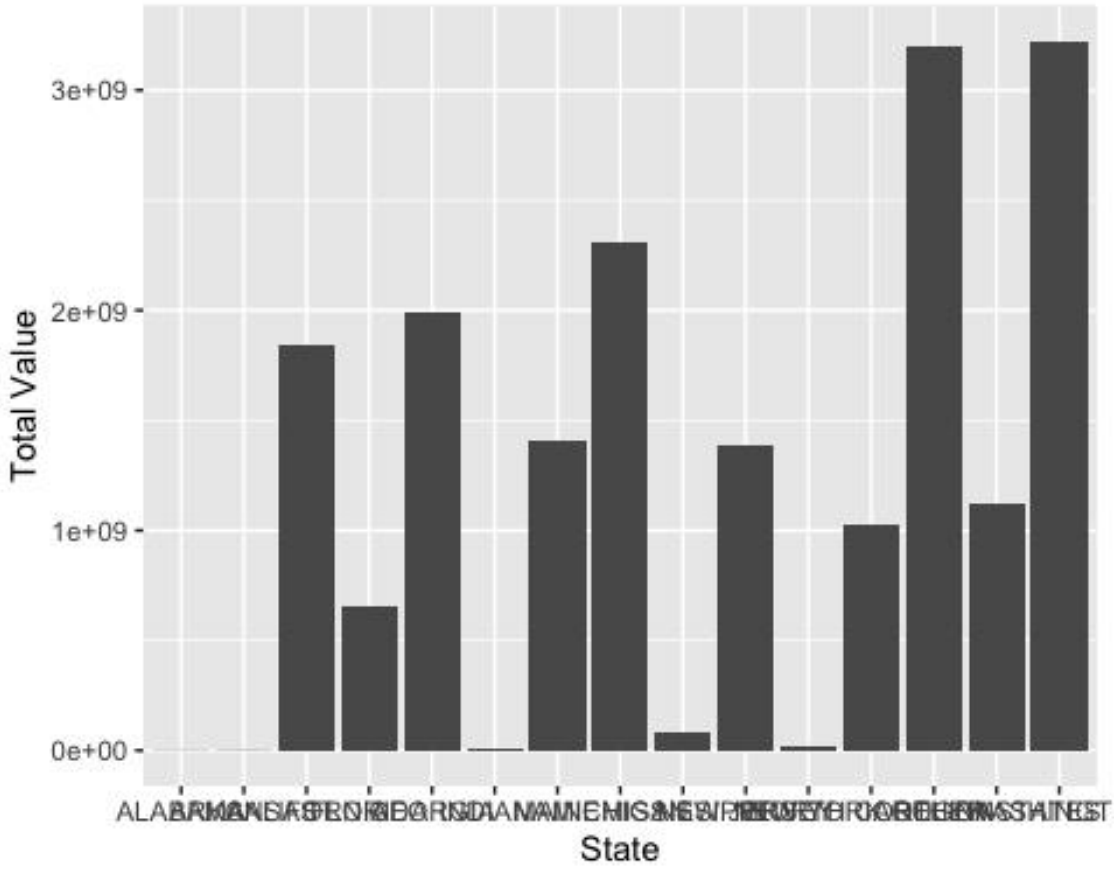
We want to have a first impression on what information the dataset contains and how the data from each state differs from others. We will focus on two aspects. First, the numbers of records in each state. Second, the values from each states, which shows the yields of blueberries in each state.

```
# Number of records in each state
ggplot(data=berry)+
  geom_bar(mapping=aes(x=State))
```



Now we can briefly see the differences of yields between states, but the result is not convincing enough. There could be occasions that one state has less records while has much higher value in each records. So next step is to plot the sum of values in each state.

```
cstate<- c("ALABAMA", "ARKANSAS", "CALIFORNIA", "FLORIDA", "GEORGIA", "
INDIANA", "MAINE", "MICHIGAN", "MISSISSIPPI", "NEW JERSEY", "NEW YORK",
"NORTH CAROLINA", "OREGON", "OTHER STATES", "WASHINGTON" )
value_sum<- rep(NA, 15)
j=1
for (i in cstate){
  berry1<- berry %>% filter(State==i)
  value_sum[j]<- sum(berry1$Value)
  j=j+1
}
svalue<- data.frame(cstate, value_sum)
ggplot(data=svalue, mapping=aes(x=cstate, y=value_sum), fill=obj,group=
factor(1))+
  geom_bar(stat="identity")+
  labs(x="State", y="Total Value")
```



By comparing the two graphs we have plotted above, it confirms our assumption, that there are states without too many records while their total value is very high. So the production of blueberries mainly comes from California, Florida, Georgia, Maine, Michigan, Jersey, North Carolina, Oregon and Washington. It's not difficult to see that these states are mostly located by seaside. So we can roughly say that a seaside environment is good for blueberries to grow.

The Variation of Yields by Year

The data we have ranges from 2015 to 2019. We would like to know how the yields varies year by year. From the graphs above, Washington has the highest value for blueberries production. So we will take the data for Washington as an example.

```
berry_w<- berry %>% filter(State=="WASHINGTON")
w2015<- berry_w %>% filter(Year==2015)
w2016<- berry_w %>% filter(Year==2016)
w2017<- berry_w %>% filter(Year==2017)
w2018<- berry_w %>% filter(Year==2018)
w2019<- berry_w %>% filter(Year==2019)
wvalue<- c(sum(w2015$Value),
        sum(w2016$Value),
        sum(w2017$Value),
        sum(w2018$Value),
```

```
        sum(w2019$Value))
wyear<- c(2015:2019)
washi<- data.frame(wyear, wvalue)
ggplot(data=washi, mapping=aes(x=wyear, y=wvalue))+
  geom_point(shape=13, color="red", size=7)+
  geom_smooth()+
  geom_smooth(method=lm, formula='y~x', color="green")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freed
om.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 2015
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.02
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 4.0804
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : span too sma
ll. fewer
## data values than degrees of freedom.
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinvers
e used at
## 2015
```

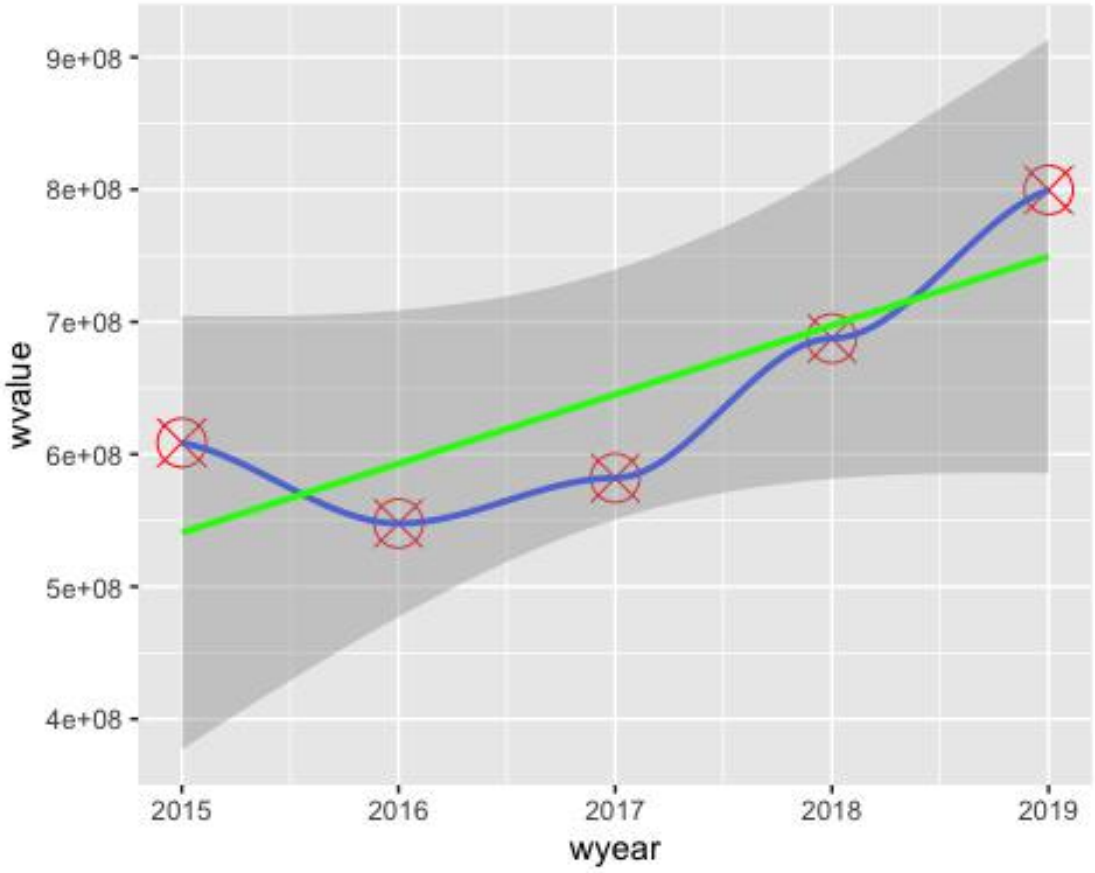
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood
 radius 2.02
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal c
ondition
## number 0
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are ot
```

```
her near
## singularities as well. 4.0804

## Warning in max(ids, na.rm = TRUE): max 里所有的参数都不存在；回覆-Inf
```




```
lm(wvalue~wyear, data=washi)

##
## Call:
## lm(formula = wvalue ~ wyear, data = washi)
##
## Coefficients:
## (Intercept)        wyear
##  -1.046e+11    5.218e+07
```

From the graph above, the yields of blueberry production in Washington state is generally become higher from 2015 to 2019. And we use linear regression to fit an equation, and the coefficient is positive, which helps confirm that the trend is going upwards. However, since the data are only from 2015 to 2019, it is not likely to make a more accurate prediction about the yields years later.