# Strawberry Report

Jingwen Xu

2020/10/18

## Project Objective

The data berries is from the survey in states of America about all kinds of agricultural information about blueberry, raspberry and strawberry.Through this program, we hope to examine our understanding and use of tidyverse, data cleaning and organization, EDA, r markdown and shiny.

## Project Method and Progress

### 1.Acquire and read the data

These data were collected from the USDA database selector: https://quickstats.nass.usda.gov

The data were stored online and then downloaded as a CSV file.

### 2.Clean the data

We found that there are many columns with only one unique and these columns should be selected out. After the initial cleaning operation, we can get the following data table:

| Year | Period | State | Commodity | Data Item |
|---|---|---|---|---|
| 2019 | MARKETING YEAR | CALIFORNIA | BLUEBERRIES | BLUEBERRIES, TAME - PRICE R |
| 2019 | MARKETING YEAR | CALIFORNIA | BLUEBERRIES | BLUEBERRIES, TAME, FRESH M |
| 2019 | MARKETING YEAR | CALIFORNIA | BLUEBERRIES | BLUEBERRIES, TAME, PROCESS |
| 2019 | MARKETING YEAR | CALIFORNIA | RASPBERRIES | RASPBERRIES - PRICE RECEIVE |
| 2019 | MARKETING YEAR | CALIFORNIA | RASPBERRIES | RASPBERRIES, FRESH MARKET |
| 2019 | MARKETING YEAR | CALIFORNIA | RASPBERRIES | RASPBERRIES, PROCESSING - P |

#### Strawberries

Professor Haviland had completed the data operations with blueberries. Now, I want to operate on commodity strawberries only with period "Year" from data cleaning to EDA.

**(1)Data cleaning** Firstly, I need to split the columns whose arguments consist of several unique combined by "," or "-" and then select out the redundant columns so that I can get the following data table:

| Year | State | label | market | meas |
|---|---|---|---|---|
| 2019 | CALIFORNIA | ACRES HARVESTED | | |
| 2019 | CALIFORNIA | ACRES PLANTED | | |
| 2019 | CALIFORNIA | PRODUCTION | MEASURED IN $ | |
| 2019 | CALIFORNIA | PRODUCTION | MEASURED IN CWT | |
| 2019 | CALIFORNIA | YIELD | MEASURED IN CWT / ACRE | |
| 2019 | CALIFORNIA | | BEARING - APPLICATIONS | MEASURED IN LB |
| 2019 | CALIFORNIA | | BEARING - APPLICATIONS | MEASURED IN LB |
| 2019 | CALIFORNIA | | BEARING - APPLICATIONS | MEASURED IN LB |
| 2019 | CALIFORNIA | | BEARING - APPLICATIONS | MEASURED IN LB |
| 2019 | CALIFORNIA | | BEARING - APPLICATIONS | MEASURED IN LB |

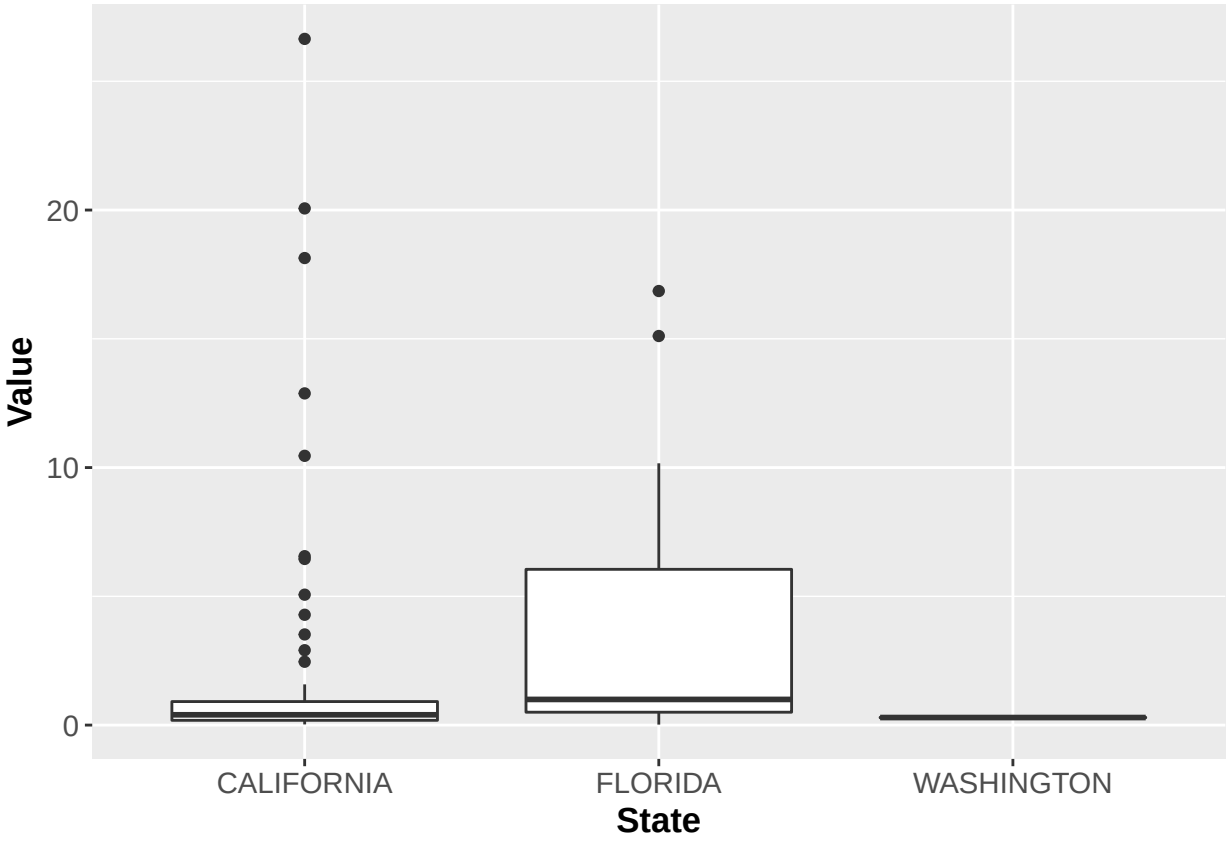
**(2)Data organization** After the data cleaning towards the columns, I found that there are the same entries in different columns which may result in the disarray of data. So I separate out these entries as a new column and select out the redundancy. The process of data organization produced the final tidy dataset.

| Year | State | type | production | Measures | Avg | C |
|---|---|---|---|---|---|---|
| 2019 | CALIFORNIA | | ACRES HARVESTED | | | |
| 2019 | CALIFORNIA | | ACRES PLANTED | | | |
| 2019 | CALIFORNIA | | PRODUCTION | MEASURED IN $ | | |
| 2019 | CALIFORNIA | | PRODUCTION | MEASURED IN CWT | | |
| 2019 | CALIFORNIA | | YIELD | MEASURED IN CWT / ACRE | | |
| 2019 | CALIFORNIA | BEARING | APPLICATIONS | MEASURED IN LB | | I |
| 2019 | CALIFORNIA | BEARING | APPLICATIONS | MEASURED IN LB | | I |
| 2019 | CALIFORNIA | BEARING | APPLICATIONS | MEASURED IN LB | | I |
| 2019 | CALIFORNIA | BEARING | APPLICATIONS | MEASURED IN LB | | I |
| 2019 | CALIFORNIA | BEARING | APPLICATIONS | MEASURED IN LB | | I |

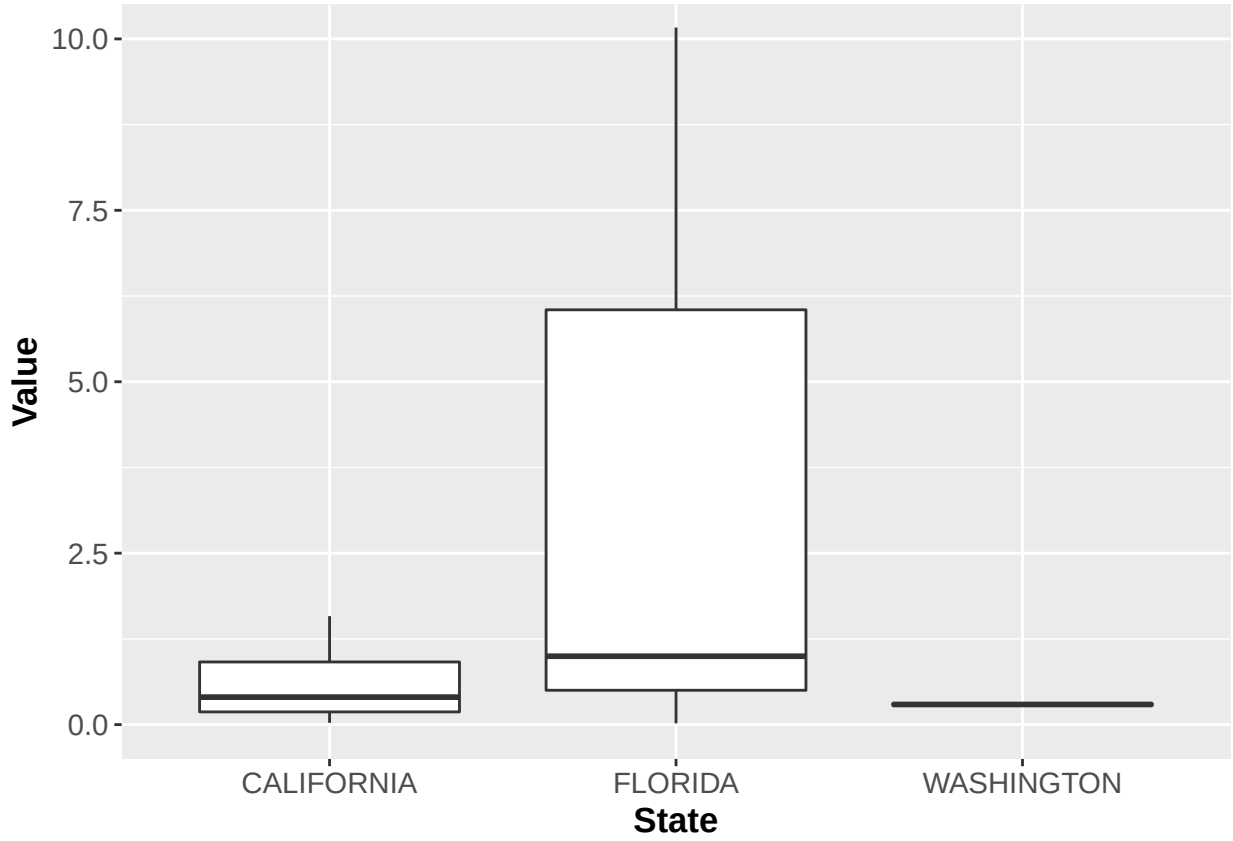
## 3.EDA

Before EDA, I deleted the rows with (NA) or (D) values and turn the character value to numeric value so that we can plot them. There are many variables and all kinds of type of data making the analysis difficult. I choose the data of different states with chemical fungicide measured in lb/acre/year as my object.

**(1)Make boxplot of data** The boxplot can visually display data dispersion, so I make a boxplot to show different data dispersion of different States in America.




We can see there are outliers in the boxplot. Now, I will exclude outliers to make the new boxplot.




In the boxplot, there are many outliers in California and the median value of Florida is larger than that of California. Washington has data of only one year, so its data are mostly zero.

**(2)Explor the data property** (a)Find number of zeros in each state

```
## CALIFORNIA    FLORIDA WASHINGTON
##          1         60         75
```

According to the result, the data of California is most completed with only one zero value. Both Florida and Washington have more than 75 percentage of zero values.

(b)Find the upper inner fence value for each state

```
## CALIFORNIA    FLORIDA WASHINGTON
##     1.9315     0.0000     0.0000
```

It's not surprising to get such a result. Combining the boxplot and maxsb1 values, the upper inner fence value of California is truely about 1.9315. And due to mostly zero values, the upper inner fence values of Florida and Washington are both zero.

(c)Find variable describing the most variance and make histograms of least and most variance variables

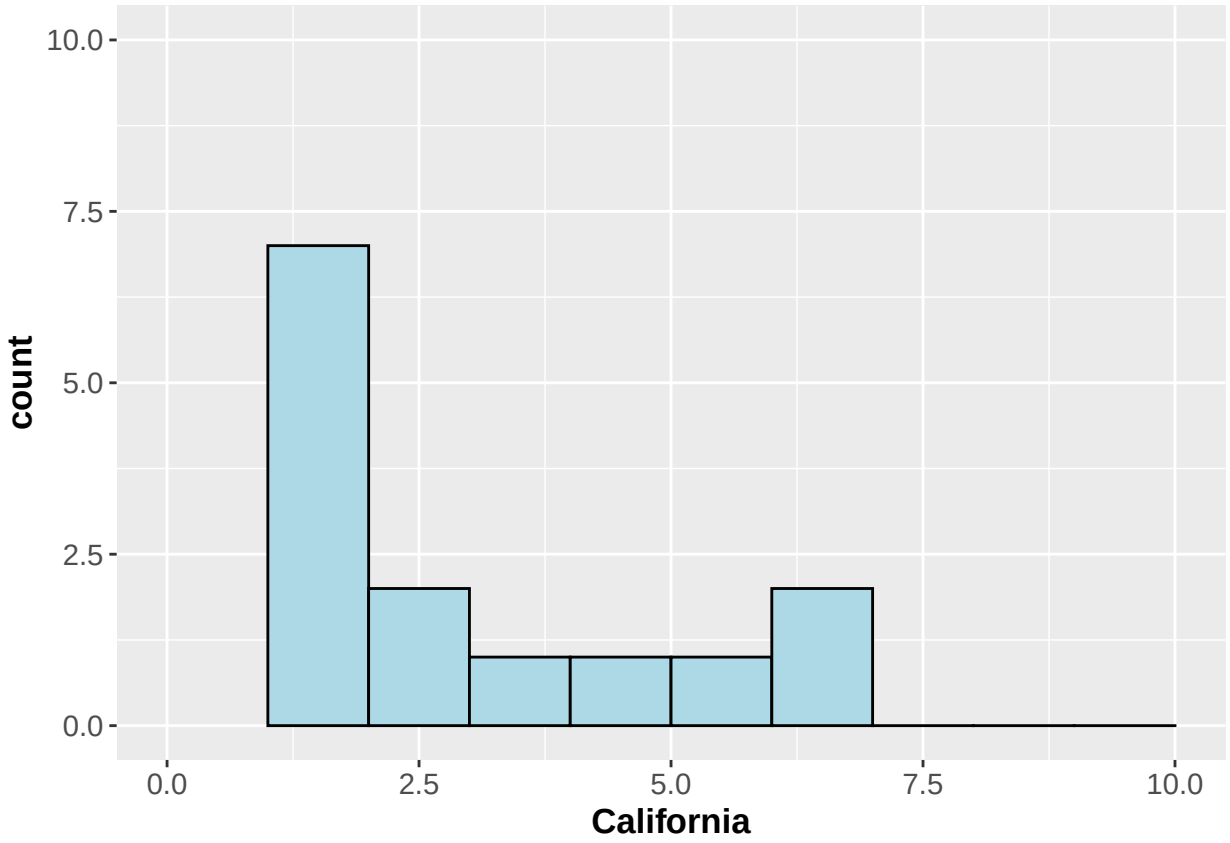
```
##   CALIFORNIA      FLORIDA   WASHINGTON
## 21.216745001  9.403129726  0.002222554
```

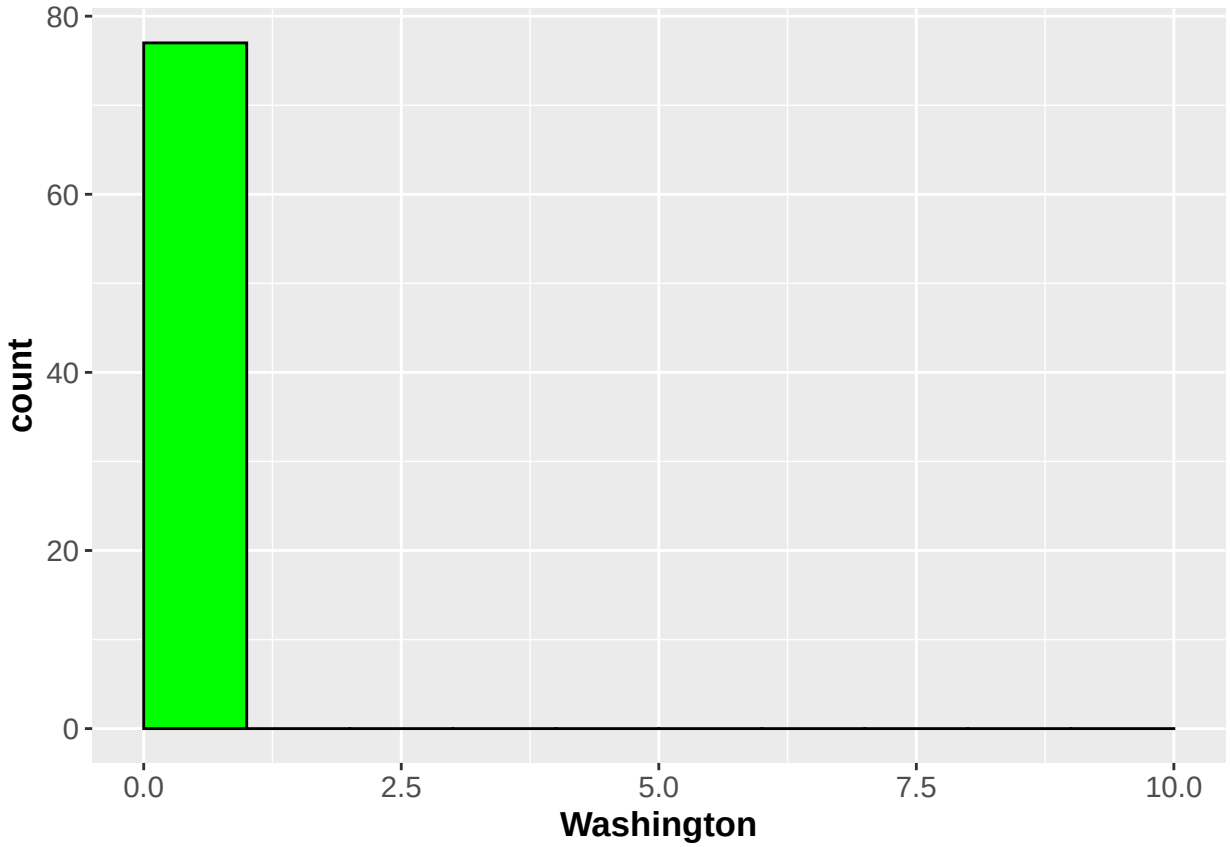
```
## [1] 21.21675
```

```
## CALIFORNIA
##          1
```

```
## [1] 0.002222554
```

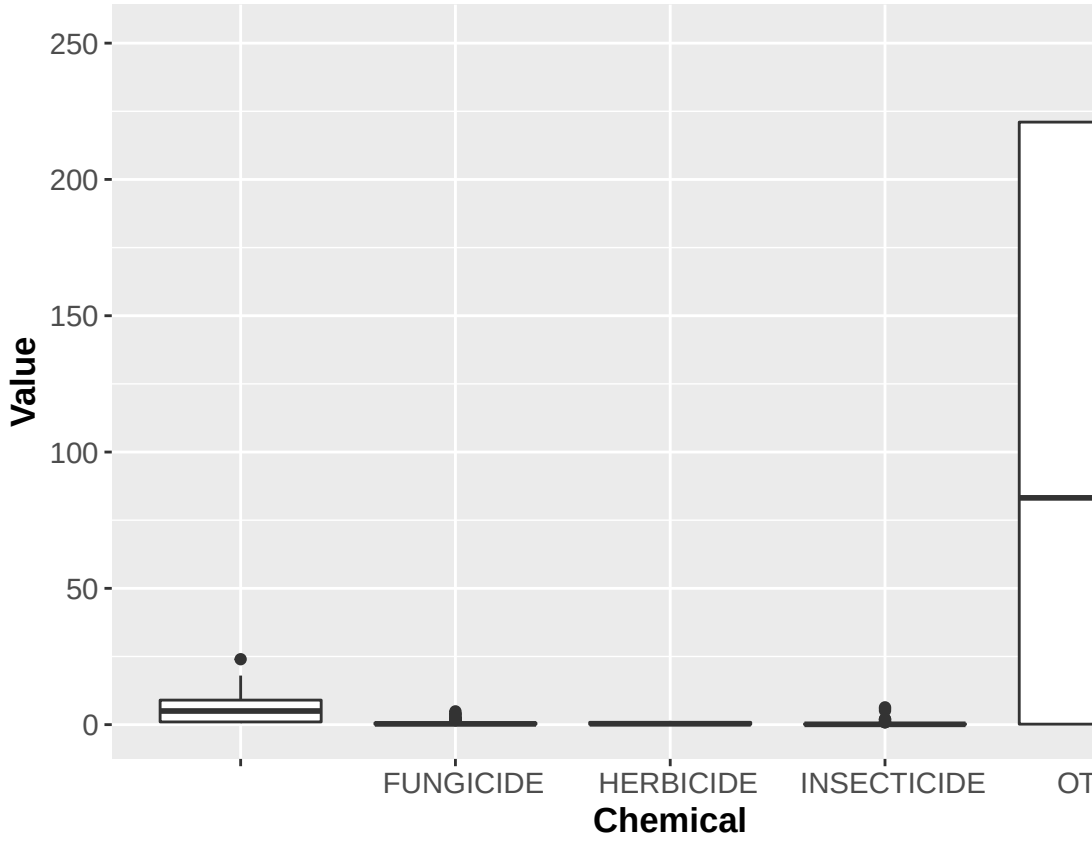
```
## WASHINGTON
##          3
```






Referring to the variance computation and the histograms, California is the most variance variable and Washington is the least one because it has only two positive values.

Another dataset I choose is the data of production "Application" measured in lb/acre/application.




**(1)Make boxplot of data**

There are few outliers in this box plot, so I don't need to make a new boxplot.

**(2)Explor the data proverty** (a)Find number of zeros in each state

```
##     FUNGICIDE     HERBICIDE   INSECTICIDE         OTHER
##           129           213           134           209           211
```

There are about half zero values in Fungicide and Insecticide, more than 75 percentage of zero values in Herbicide and other chemicals.

(b)Find the upper inner fence value for each state

```
##     FUNGICIDE     HERBICIDE   INSECTICIDE         OTHER
##      0.533125      0.000000      0.217500      0.000000      0.000000
```

Due to mostly zero values, the upper inner fence value of herbicide and insecticide is 0. And the upper inner fence values of fungicide and insecticide are reasonable referring to the boxplot.

(c)Find variable describing the most variance and make histograms of least and most variance variables

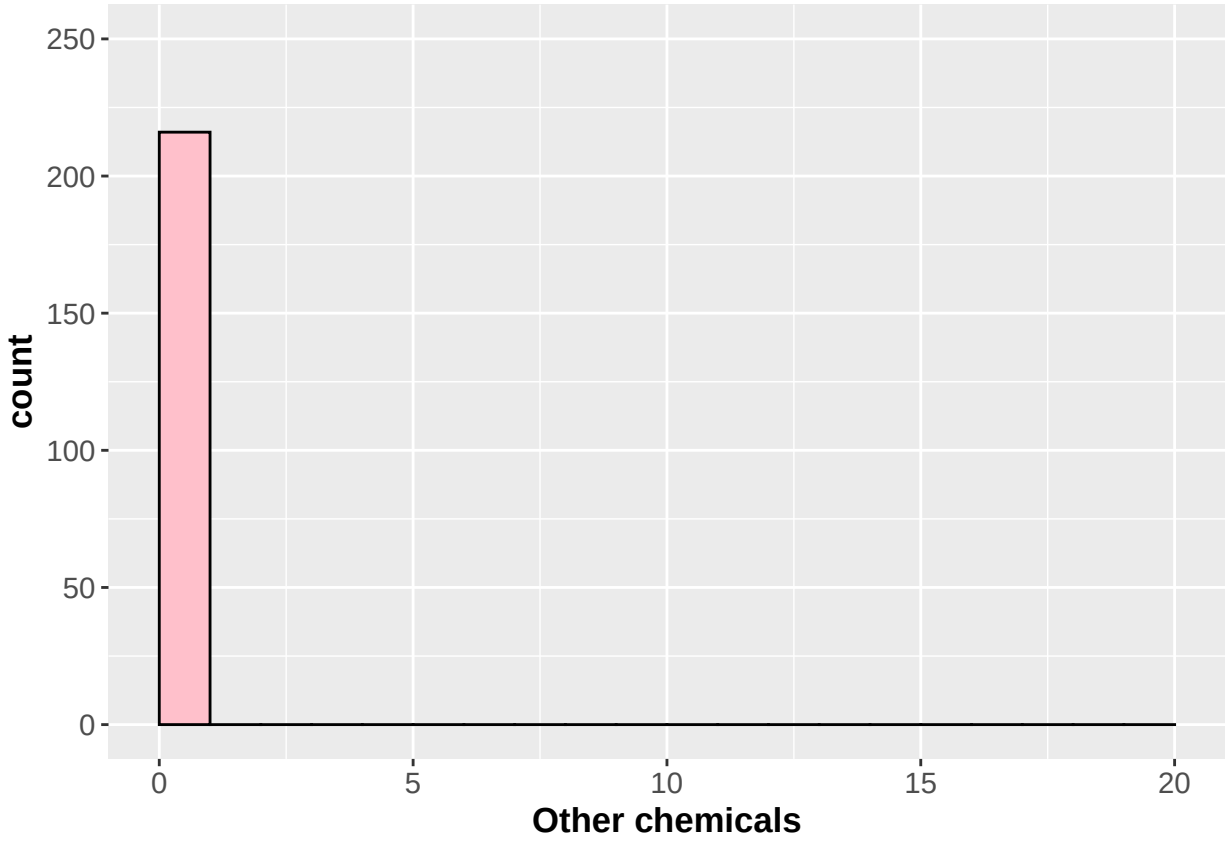
```
##     FUNGICIDE     HERBICIDE   INSECTICIDE         OTHER
##    0.50663561    0.02508259    0.50218668 1310.45024893    5.84927530
```

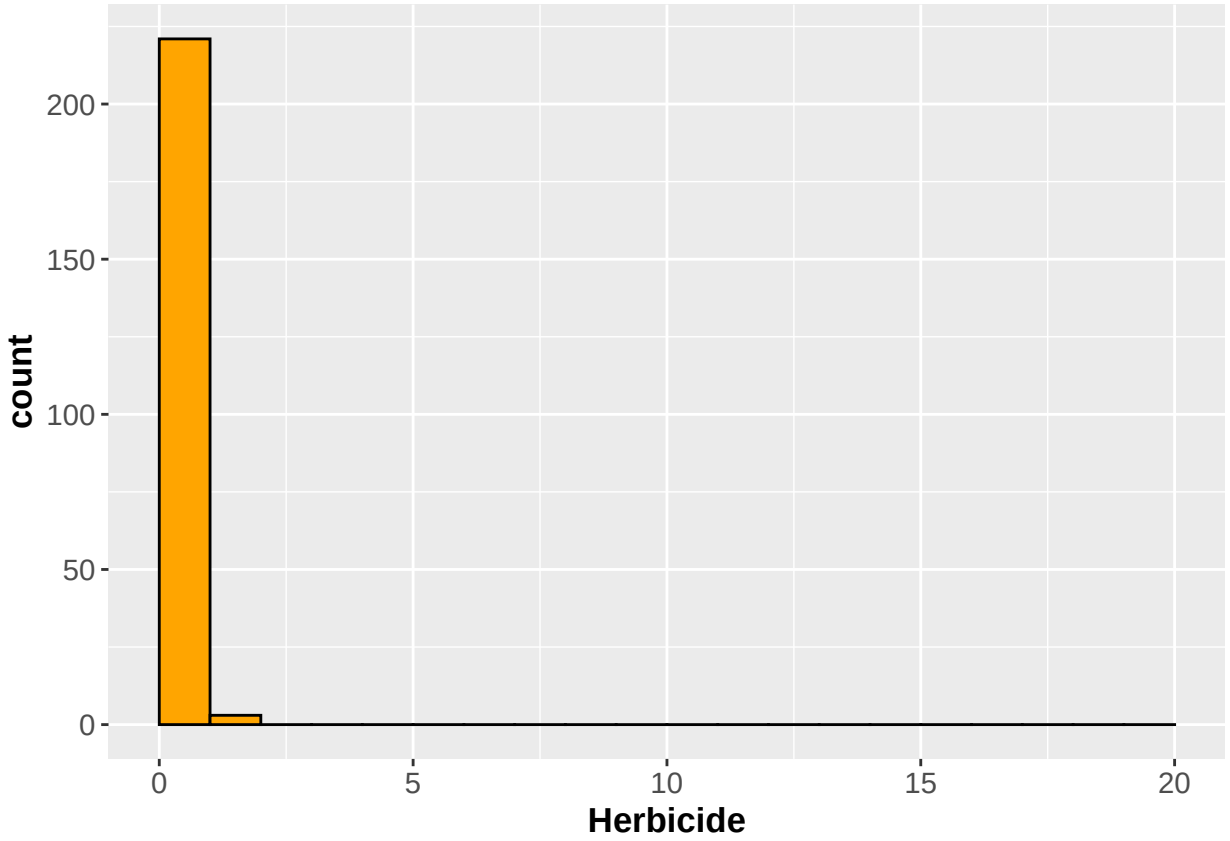
```
## [1] 1310.45
```

```
## named integer(0)
```

```
## [1] 0.02508259
```

```
##   HERBICIDE
##           2
```


250
200
150
100
50
0
count
0
5
10
15
20
Other chemicals




The information and component in Other chemical is obscured and complicated, so its variance is large. Herbiside has only 21 non-zero values with a tiny variance.

## Project Conclusions

According to the two datasets that I had analyzed, if we control the variables, the analysis will be more precise but with a lot of zero value in the filtered dataset. And it's difficult to interpret the results of EDA because there is only one measurable variable. But this is indeed a good training to do data cleaning and organization with nearly all the common cleaning methods. If we want to have a deeper analysis on this dataset, we may need to go on refining the data, especially dealing with the empty values.

## Reference

(1)H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

(2)Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

(3)Stefan Milton Bache and Hadley Wickham (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. https://CRAN.R-project.org/package=magrittr.

(4)Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra.

(5)United States Department of Agriculture.(2020). National Agricultural Statistics Service Quick Stats. Available from:https://quickstats.nass.usda.gov/results/D416E96E-3D5C-324C-9334-1D38DF88FFF1