Berry Data Analysis Report

Fall 2020 MA615 Assignment 2

Wendy Liang

Data Description

This report is a limited data exploratory analysis of Berry data from the USDA database selector: https://quickstats.nass.usda.gov. The data were https://quickstats.nass.usda.gov/results/D416E96E-3D5C-324C-9334-1D38DF88FFF1 stored online and then downloaded as a CSV file.

The Berry data displays the survey statistics of three kinds berries based on the simple filtering rules of What, Where, and When.https://quickstats.nass.usda.gov/param_define and https://quickstats.nass.usda.gov/src/glossary.pdf describe different variables of Berry. And My task is to explore these variables and my goal is to support the agricultural research team for analysis and modeling.

Method Overview

There are three steps of my work:

- step1: Data Cleaning and Organizing
- step2: Exploratory Data Analysis
- step3: Shiny App

Analysis was conducted in R Studio (Version 1.3.1073). Data cleaning was performed using package `stringr` and package `magrittr`. EDA was conducted using package `ggplot2` and `dplyr`, creating shiny app were conducted using package `shiny`. In addition, this report was compiled by package `Knitr`.

step1: Data Cleaning

In this step, the most important thing is to separate certain chr variables.

Firstly, I use `separate` function on `Data Item` and delete the `Berries Name`.

```
#### seperate Measure and Berry
ag_data %<>% separate(`Data Item`,c("Berry","Other"),sep = "BERRIES")

#### delete Berry Category
ag_data %<>% select(-Berry)
head(ag_data)
```

Secondly, I use `separate` function with parameter `sep=""` to separate measurement variable, such as "$ / LB".

```
#### Create Measure Method
ag_data %<>% separate(`Other`,c("lab1","Measure"),sep="MEASURED IN")
#check
head(ag_data)
```

```r
ag_data%>%summarize_all(n_distinct)
unique(ag_data$Measure)
```

Then, I use functions in `stringr` package to deal with the data. Instead of splitting characters into different column in Prof's work, I create new column. I use `str_extract` function to extract my goal patterns from each string and save them as new variables (new column). In this way, I create `Type`,`Production` and `Marketing` variables.

```r
#### Create Type, Production, Marketing, Domain,Chemi_family, Materials
ag_data$Type=str_extract(ag_data$lab1,pattern = "(BEARING)|(TAME)|(WILD)")
ag_data$Production=str_extract(ag_data$lab1,pattern = "PRODUCTION")
ag_data$Marketing=str_extract(ag_data$lab1,pattern ="(ACRES HARVESTED)|
(YIELD)|(FRESH MARKET)|
(NOT SOLD)|(PROCESSING)|(UTILIZED)|
(APPLICATIONS)|(TREATED)|(NOT HARVESTED)")
```

Last but not least, I saperate`Domain` and `Domain Category` as the same as dealing with `Measurement`.

```r
ag_data %<>% separate(`Domain`,c("Domain","Chemi_family"),sep=",")
ag_data %<>% separate(`Domain Category`,c("lab2","Materials"),sep="[(]")
ag_data %<>% separate(`Materials`,c("Materials","lab3"),sep="[)]")
#check
head(ag_data)
unique(ag_data$Materials)
```

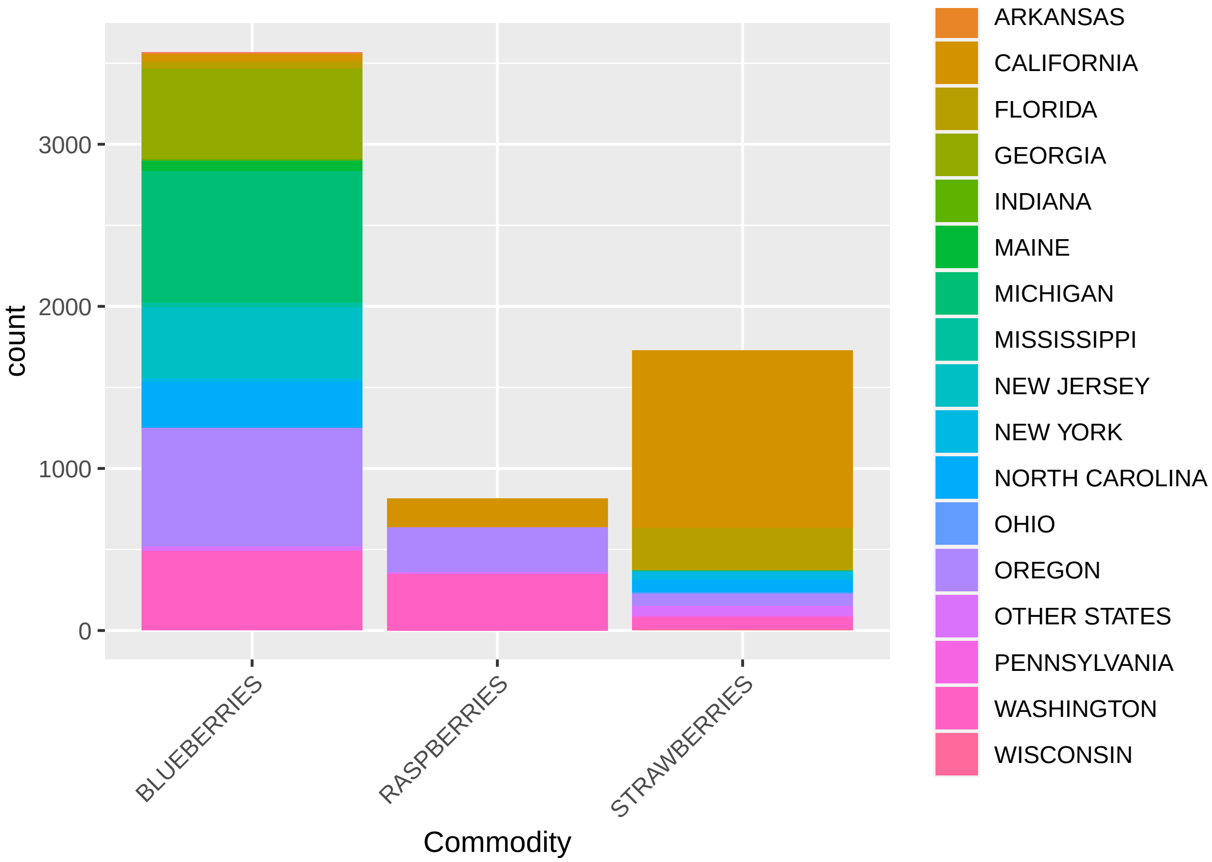
Now, I finish my data cleaning. The following is the result data frame which I export as a csv file.

```
##   X Year      State   Commodity Type Production    Marketing Measure Domain
## 1 1 2019 CALIFORNIA BLUEBERRIES TAME       <NA>         <NA>  $ / LB  TOTAL
## 2 2 2019 CALIFORNIA BLUEBERRIES TAME       <NA> FRESH MARKET  $ / LB  TOTAL
## 3 3 2019 CALIFORNIA BLUEBERRIES TAME       <NA>   PROCESSING  $ / LB  TOTAL
## 4 4 2019 CALIFORNIA RASPBERRIES <NA>       <NA>         <NA>  $ / LB  TOTAL
## 5 5 2019 CALIFORNIA RASPBERRIES <NA>       <NA> FRESH MARKET  $ / LB  TOTAL
## 6 6 2019 CALIFORNIA RASPBERRIES <NA>       <NA>   PROCESSING  $ / LB  TOTAL
##   Chemi_family Materials Value
## 1         <NA>      <NA>  2.85
## 2         <NA>      <NA>  3.56
## 3         <NA>      <NA>  0.29
## 4         <NA>      <NA>  2.69
## 5         <NA>      <NA>   (D)
## 6         <NA>      <NA>   (D)
```

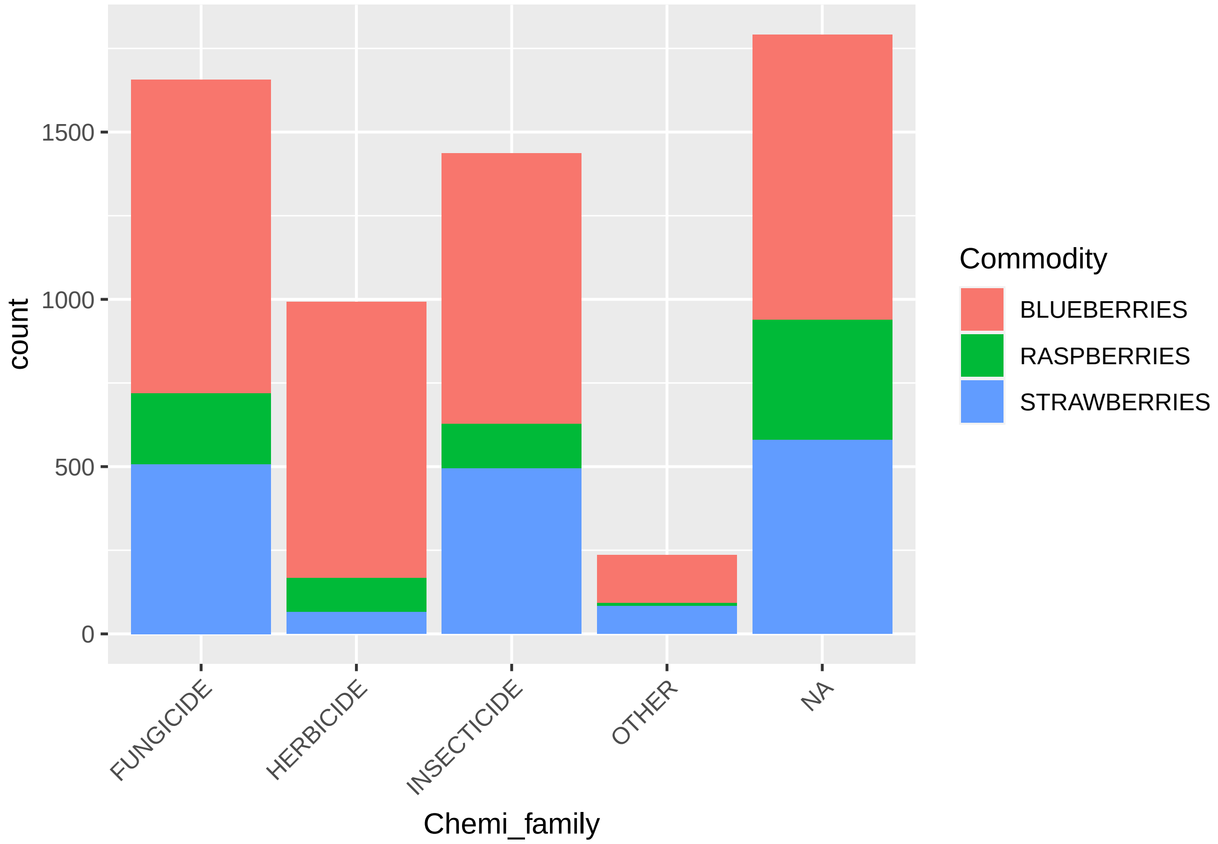
Step2: Explorary Data Analysis

Explore Category Variable

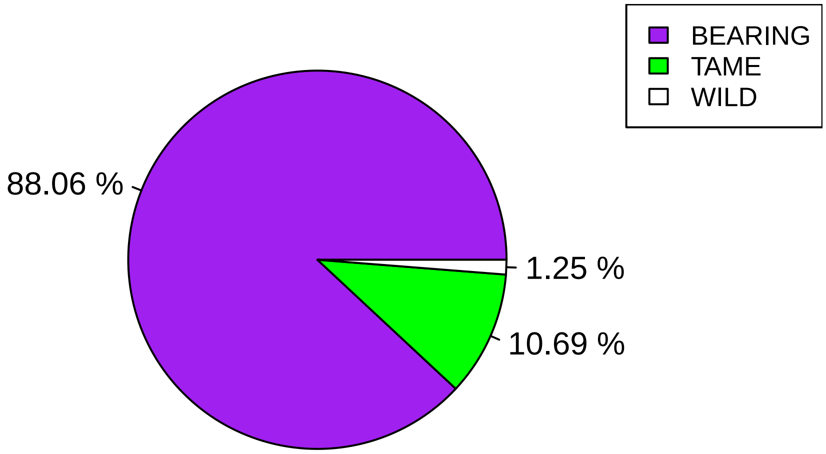
This is a histogram plot for `Commodity`, divided by `State`.



This is a histogram plot for `Chemi_family`, divided by `Commodity`.



This is a pie plot for `Type`, so we can observe 88.06% Berries are BEARING type.



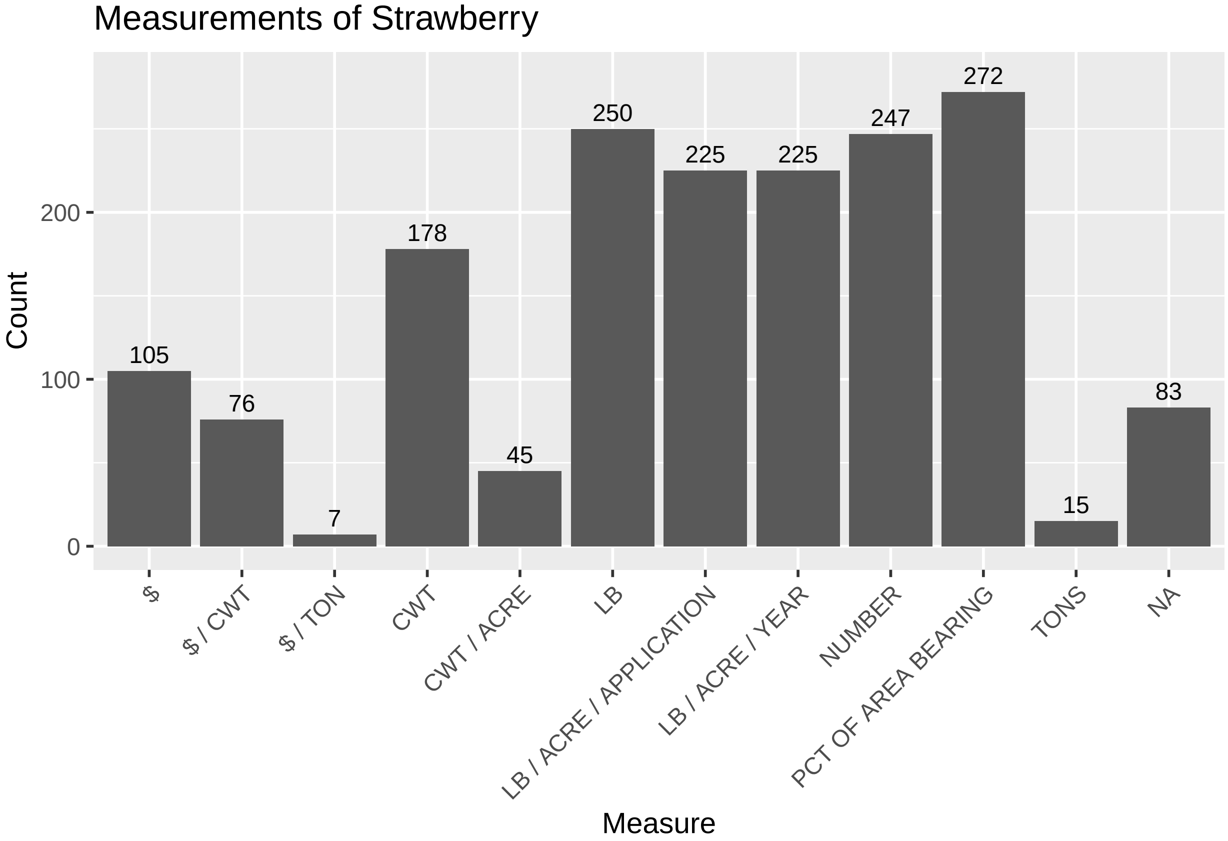
Explore Contious Variable `Value`

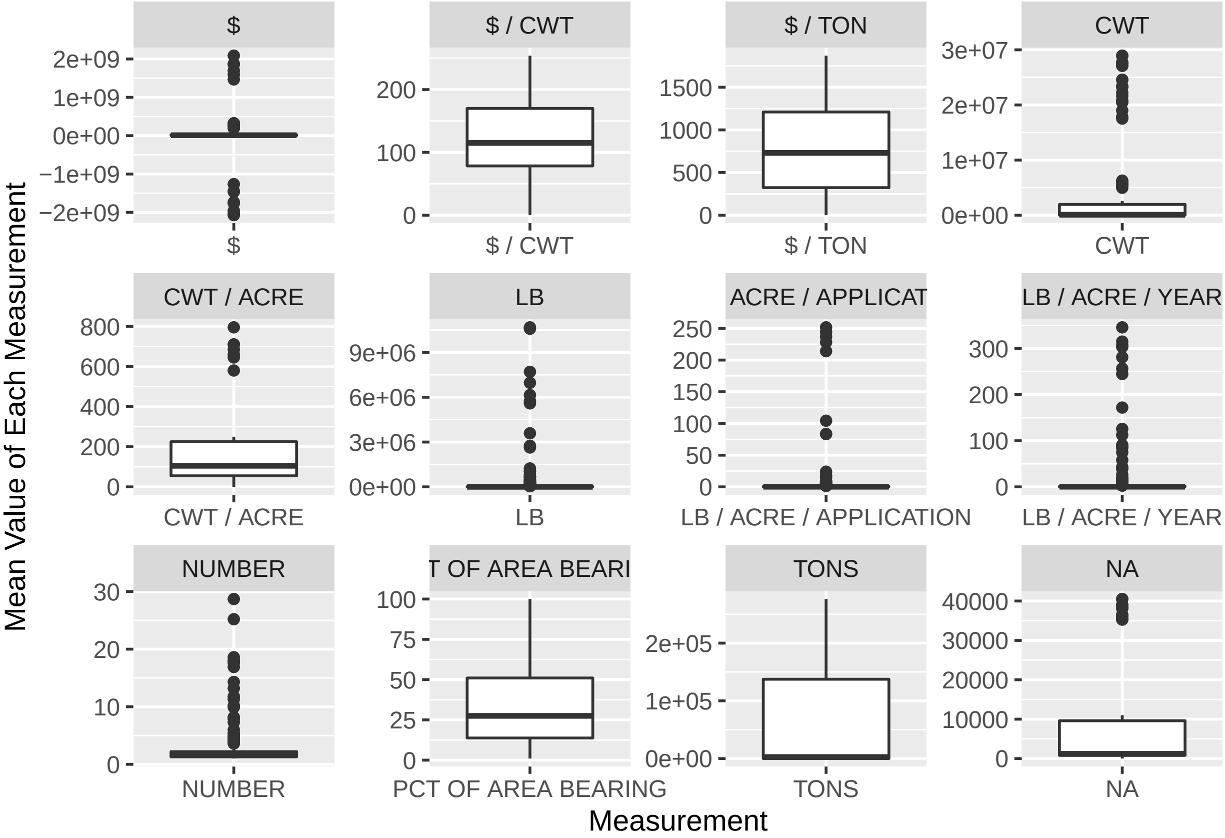
I'm gonna explore the independent variable `Value` of Strawberry dataset.

1. **Overview of `Measurement` and `Value`**

   - There are 12 types of measurements for strawberry in the dataset.
   - This is a plot for count of each measurement.There are 272 *PCT of area bearing* and 7 *$ / Ton*, which are the max and min along all measurements.

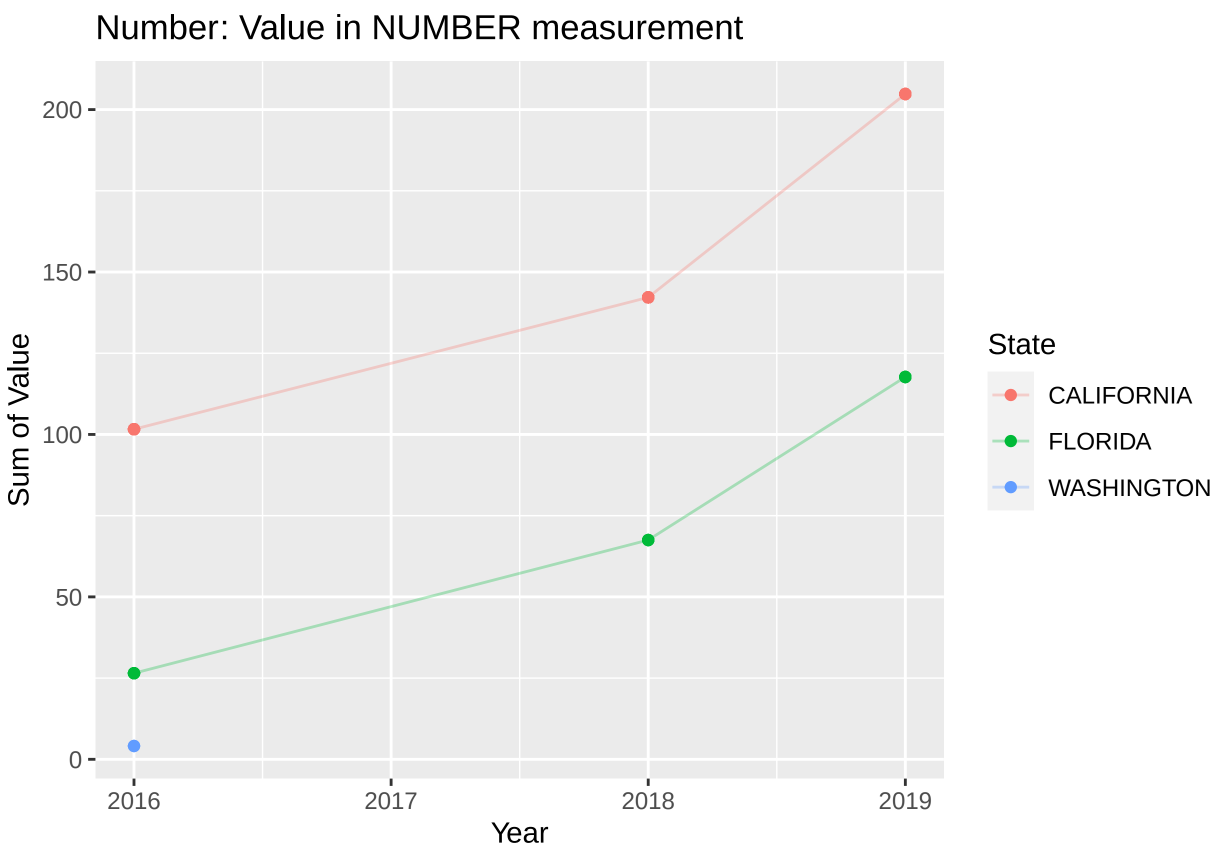
- This are several boxplots for `Value`, divided by `Measurement`.According to the plot, I find the values of different measurements are vary widely.

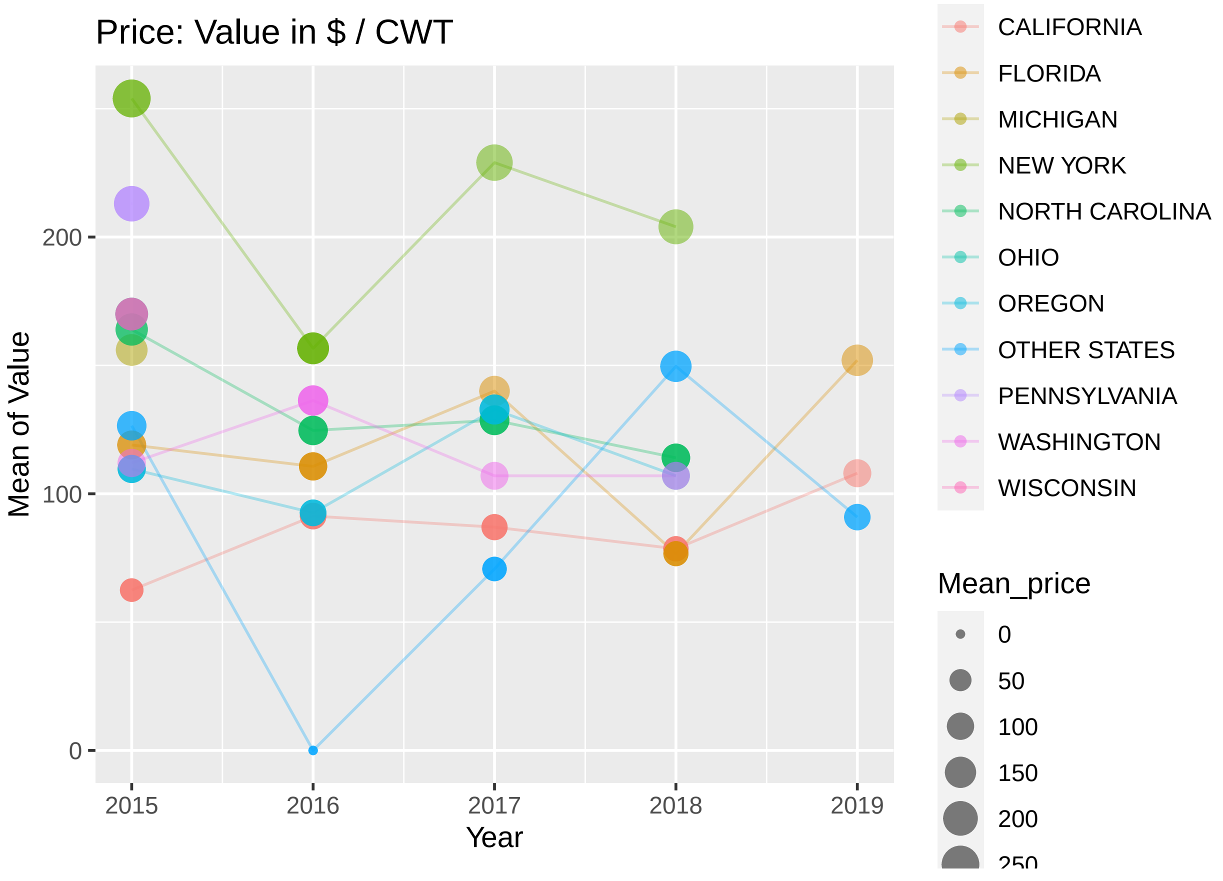


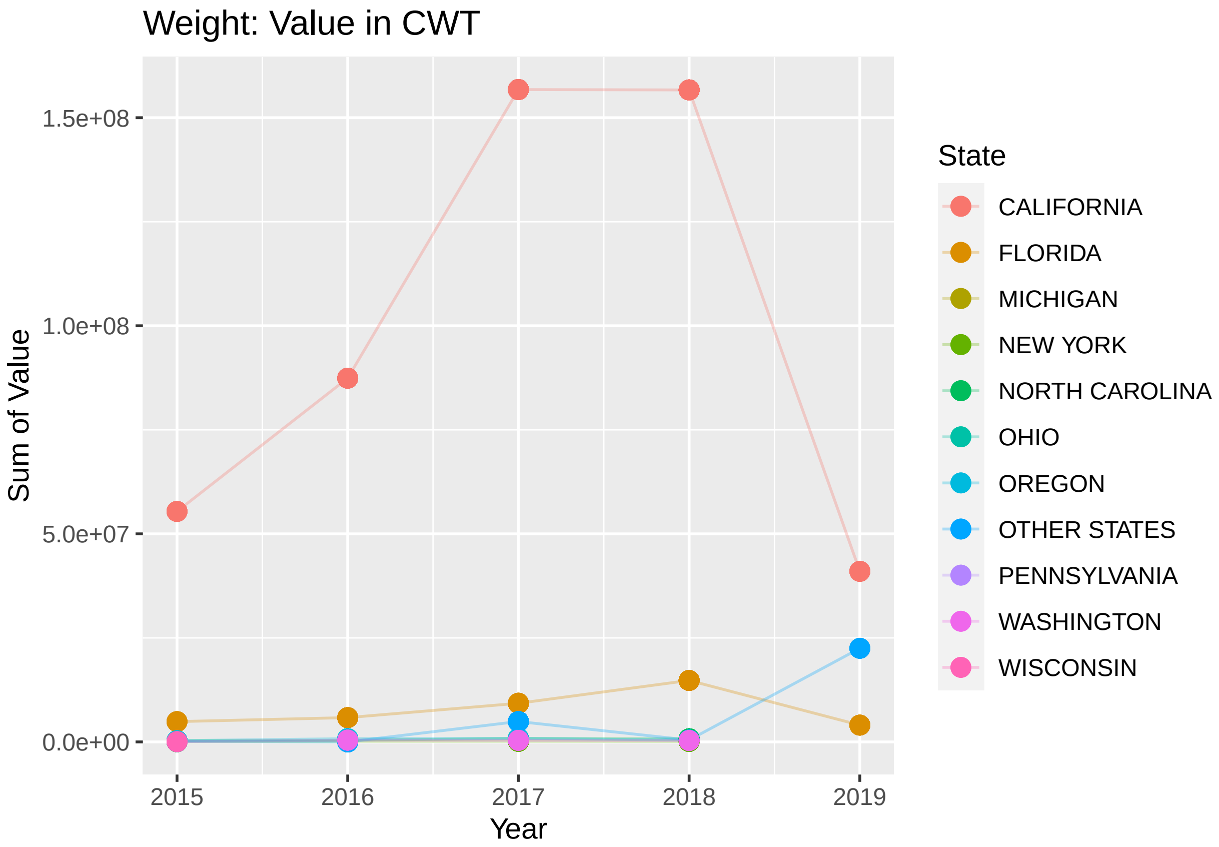


2. `Value` of three different `Measurement`

   - I choose three different measurements to describe three characters of strawberry. `NUMBER` refers to number, `$ /CWT` refers to price, and `TONS` refers to weight. I calculate the sum of `Value` in each year when exploring number and weight. And I calculate the mean of `Value` in each year when exploring price.
   - Using `ggplot2` package,I make three point and line plots. According to these plots, we can find the growing trend and value difference.

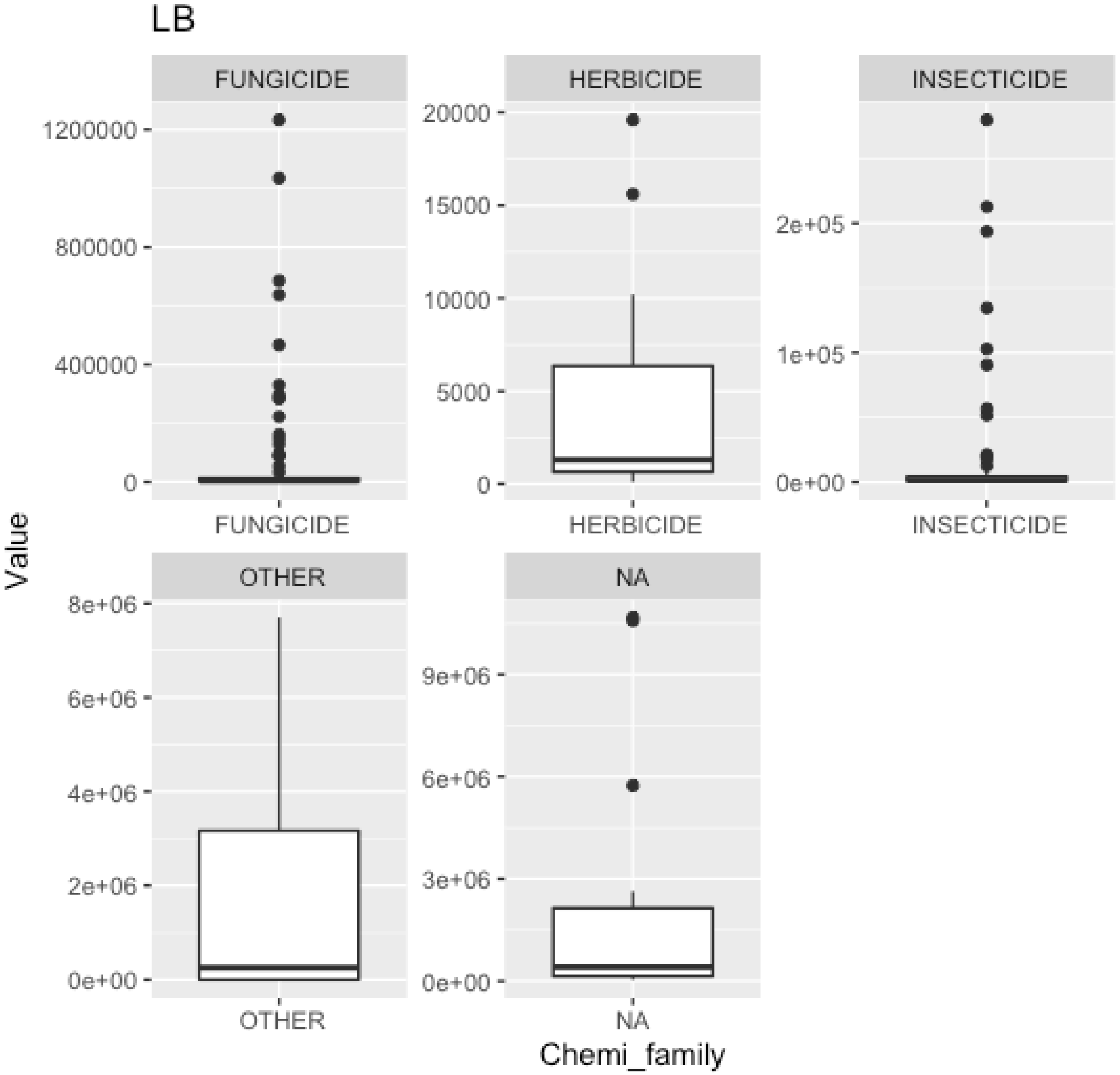

Number: Value in NUMBER measurement
Sum of Value
200
150
100
50
0
2016
2017
2018
2019
Year
State
CALIFORNIA
FLORIDA
WASHINGTON


Price: Value in $ / CWT
Mean of Value
200
100
0
2015
2016
2017
2018
2019
Year
CALIFORNIA
FLORIDA
MICHIGAN
NEW YORK
NORTH CAROLINA
OHIO
OREGON
OTHER STATES
PENNSYLVANIA
WASHINGTON
WISCONSIN
Mean_price
0
50
100
150
200
250



- There are some observations:
  - number of strawberry increases from 2016 to 2019 year
  - California has the most number of strawberry
  - Price varies widely from states and years
  - The strawberry in New York is the most expensive while California is cheapest.
  - The weight of strawberry in California and Michigan decreased a lot in 2019

Value of different Marketing I just display one situation in the report —- Marketing== "APPLICATION" meanwhile Measurement == "LB". This kind of plot indicate the Value distribution of different chemical families.



Step3: Shiny App

Here's the APP link: https://wendy-liang.shinyapps.io/Berry__EDA/?_ga=2.148296823.1134128627.1603185041-870852303.1603185041

Conclusion

Based on the Exploratory Data Analysis, I come up with some personal conclusions. Firstly, California is a suitable place for strawberry to grow and New York is a profitable place for growers. Secondly, there are correlations between chemical family and value in each measurement. Thirdly, strawberries are grown more with time while prices fluctuate a lot each year.

This analysis certainly contains its limitations. There are a large number of NA in the dataset. My analysis only accounts one kind of berries. Further exploration and modeling could include all these berries –

blueberry, raspberry and strawberry.

Reference

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

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

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