midterm_EDA

Hanzhang Song

2022-11-09

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(readxl)

strawb <- read_xlsx("strawberries-2022oct30-a.xlsx", col_names = T)

cnames <- colnames(strawb)
x <- 1:dim(strawb)[2]

unique(strawb[1])
```

```
## # A tibble: 2 x 1
##   Program
##   <chr>
## 1 CENSUS
## 2 SURVEY
```

```
unique(strawb[2])
```

```
## # A tibble: 6 x 1
##    Year
##   <dbl>
## 1  2019
## 2  2016
## 3  2021
## 4  2020
## 5  2018
## 6  2017
```

```
unique(strawb[3])
```

```
## # A tibble: 2 x 1
##   Period
##   <chr>
## 1 YEAR
## 2 MARKETING YEAR
```

```
T <- NULL
for(i in x){T <- c(T, dim(unique(strawb[i]))[1])}

drop_cols <- cnames[which(T == 1)]

strawb %<>% select(!all_of(drop_cols))

strawb %<>% arrange(Year, State)

colnames(strawb)
```

```
##  [1] "Program"         "Year"            "Period"          "State"
##  [5] "State ANSI"      "Data Item"       "Domain"          "Domain Category"
##  [9] "Value"           "CV (%)"
```

```
temp1 <- strawb %>% select(`Data Item`) %>%
  distinct()
strawb2 <- strawb %>% separate(col=`Data Item`,
                               into = c("Strawberries", "items", "units"),
                               sep = ",",
                               fill = "right")
```

```
## Warning: Expected 3 pieces. Additional pieces discarded in 1422 rows [3, 4, 5,
## 6, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124,
## 125, ...].
```

```
strawb3 <- strawb %>% separate(col=`Data Item`,
                               into = c("Strawberries", "type", "items", "units"),
                               sep = ",",
                               fill = "right")
```

```
rm(strawb2, strawb3)

strawb %<>% separate(col=`Data Item`,
                     into = c("Strawberries", "type", "items", "units"),
                     sep = ",",
                     fill = "right")

r_thiram <- grep("THIRAM", strawb$`Domain Category`)
r_thiram_1 <- grep("Thiram",
                   strawb$`Domain Category`,
                   ignore.case = T)
df_carbendazim <- grep("carbendazim",
                       strawb$`Domain Category`, ignore.case = T)
df_Bifenthrin <- grep("Bifenthrin",
                      strawb$`Domain Category`, ignore.case = T)
df_methyl_bromide <- grep("methyl bromide",
                          strawb$`Domain Category`, ignore.case = T)
df_1_3_dichloropropene <- grep("1,3-dichloropropene",
                               strawb$`Domain Category`,
                               ignore.case = T)
df_chloropicrin <- grep("chloropicrin",
                        strawb$`Domain Category`,
                        ignore.case = T)
df_Telone <- grep("Telone",
                  strawb$`Domain Category`,
                  ignore.case = T)
temp1 <- strawb %>% select(Strawberries) %>%
  distinct()
pr_rec <- grep("STRAWBERRIES - PRICE RECEIVED",
               strawb$Strawberries,
               ignore.case = T)


type_organic <- grep("organic",
                     strawb$type,
                     ignore.case = T)
Domain_organic <- grep("organic",
                       strawb$Domain,
                       ignore.case = T)
Domain_Category_organic <- grep("organic",
                                strawb$`Domain Category`,
                                ignore.case = T)

same <- (intersect(type_organic, Domain_organic)==
           intersect(type_organic, Domain_organic))
length(same)==length(type_organic)
```

```
## [1] TRUE
```

```
org_rows <- intersect(type_organic, Domain_organic)

strawb_organic <- strawb %>% slice(org_rows, preserve = FALSE)
```

```
strawb_non_organic <- strawb %>% filter(!row_number() %in% org_rows)


temp1 <- strawb_non_organic %>% select(type) %>%
  distinct()
chem_rows <- grep("BEARING - APPLICATIONS",
                  strawb_non_organic$type,
                  ignore.case = T)
chem_rows_1 <- grep("chemical",
                    strawb_non_organic$Domain,
                    ignore.case = T)
ins <- intersect(chem_rows, chem_rows_1)
chem_rows_2 <- grep("chemical",
                    strawb_non_organic$`Domain Category`,
                    ignore.case = T)
ins_2 <- intersect(chem_rows, chem_rows_2)

strawb_chem <- strawb_non_organic %>% slice(chem_rows, preserve = FALSE)

rm(x, T, drop_cols, temp1, r_thiram, r_thiram_1,
   df_carbendazim, df_Bifenthrin, df_methyl_bromide,
   df_1_3_dichloropropene, df_chloropicrin, df_Telone,
   pr_rec, type_organic, items_organic, Domain_organic,
   Domain_Category_organic, same, org_rows, chem_rows,
   chem_rows_1, chem_rows_2, ins, ins_2, cnames, i)
```

```
## Warning in rm(x, T, drop_cols, temp1, r_thiram, r_thiram_1, df_carbendazim, :
## object 'items_organic' not found
```

```
before_cols = colnames(strawb_chem)
T = NULL
x = length(before_cols)

for(i in 1:x){
  b <- length(unlist(strawb_chem[,i] %>% unique()) )
  T <- c(T,b)
}

drop_cols <- before_cols[which(T == 1)]
strawb_chem %<>% select(!all_of(drop_cols))
after_cols = colnames(strawb_chem)

temp1 <- strawb_chem %>% select(units) %>% distinct()

strawb_chem %<>% separate(col=`Domain Category`,
                          into = c("dc1", "chem_name"),
                          sep = ":",
                          fill = "right")

temp1 <- strawb_chem %>% select(chem_name) %>% unique()
length(unlist(temp1))
```

```
## [1] 172
```

```
aa  <- grep("measured in",
            strawb_chem$items,
            ignore.case = T)
length(aa)
```

```
## [1] 2112
```

```
sum(strawb_chem$Domain == strawb_chem$dc1) == dim(strawb_chem)[1]
```

```
## [1] TRUE
```

```
strawb_chem %<>% select(Year, State, items, units, dc1, chem_name, Value)

strawb_chem %<>% rename(category = units)

strawb_chem$items <- str_remove_all(strawb_chem$items, "MEASURED IN ")

strawb_chem %<>% rename(units = items)

bb  <- grep("CHEMICAL, ",
            strawb_chem$dc1,
            ignore.case = T)
length(bb)
```

```
## [1] 2067
```

```
chem <- 1:2112

non_chem_rows <- setdiff(chem, bb)
length(non_chem_rows)
```

```
## [1] 45
```

```
temp1 <- strawb_chem %>% slice(non_chem_rows)

fertilizers <- temp1

rm(temp1, temps, temp3, aa, bb)
```

```
## Warning in rm(temp1, temps, temp3, aa, bb): object 'temps' not found

## Warning in rm(temp1, temps, temp3, aa, bb): object 'temp3' not found
```

```
strawb_chem$dc1 <- str_remove_all(strawb_chem$dc1, "CHEMICAL, ")

strawb_chem$dc1 %>% unique()
```

```
## [1] "FUNGICIDE"   "HERBICIDE"   "INSECTICIDE" "OTHER"       "FERTILIZER"
```

```
strawb_chem %<>% rename(chem_types = dc1)

bb  <- grep("BIFENTHRIN",
            strawb_chem$chem_name,
            ignore.case = T)

bifen <- strawb_chem %>% slice(bb)

strawb_chem$chem_name <- str_remove_all(strawb_chem$chem_name, "\\(")

strawb_chem$chem_name <- str_remove_all(strawb_chem$chem_name, "\\)")


strawb_chem %<>% separate(col = chem_name,
                          into = c("chem_name","chem_code"),
                          sep = "=",
                          fill = "right"
)

aa <- which(strawb_chem$units == " LB")

bb <- which(is.na(strawb_chem$category))

sum(aa==bb)==length(aa)
```

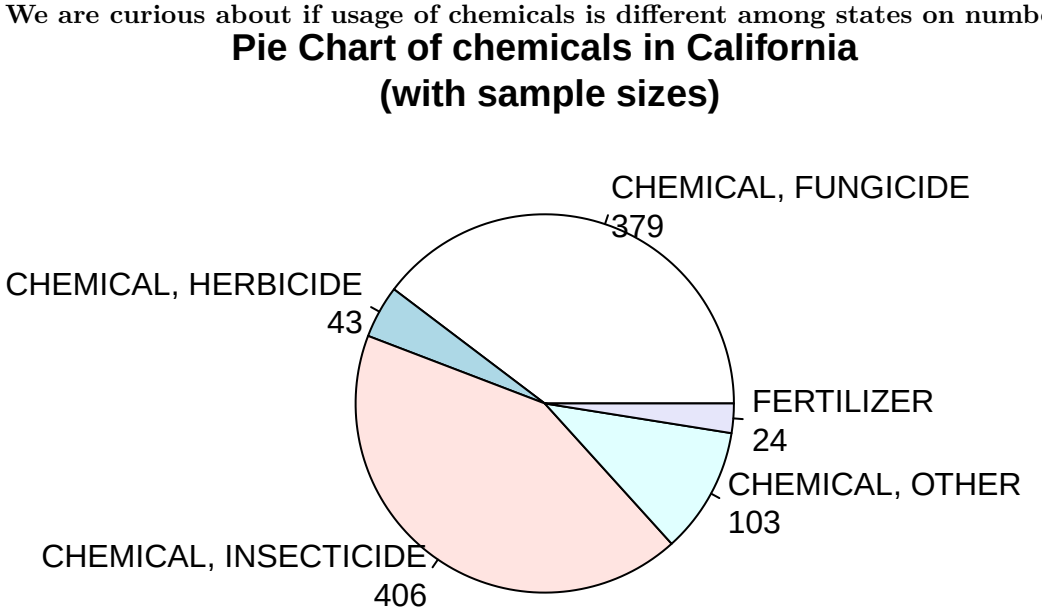
```
## [1] TRUE
```

```
FL <- filter(strawb, State == 'FLORIDA' &
               Domain != 'ORGANIC STATUS' &
               Domain != 'TOTAL' &
               Value != '(D)' &
               Value != '(Z)')
FL_chems <- FL %>%
  group_by(`Domain`) %>%
  summarise(`Count` = n())

CA <- filter(strawb, State == 'CALIFORNIA' &
               Domain != 'ORGANIC STATUS' &
               Domain != 'TOTAL' &
               Value != '(D)' &
               Value != '(Z)')
CA_chems <- CA %>%
  group_by(`Domain`) %>%
  summarise(`Count` = n())

OR <- filter(strawb, State == 'OREGON' &
               Domain != 'ORGANIC STATUS' &
               Domain != 'TOTAL' &
               Value != '(D)' &
               Value != '(Z)')
```

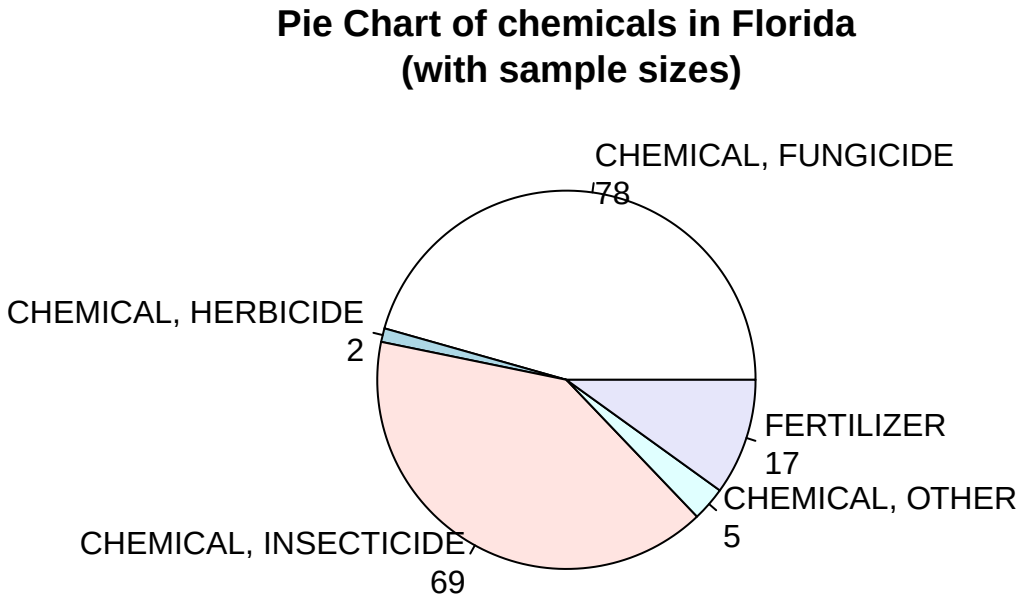
```
OR_chems <- OR %>%
  group_by(`Domain`) %>%
  summarise(`Count` = n())
```

```
CAtable <- table(CA$Domain)
lbls <- paste(names(CAtable), "\n", CAtable, sep="")
pie(CAtable, labels = lbls,
   main="Pie Chart of chemicals in California\n (with sample sizes)")
```

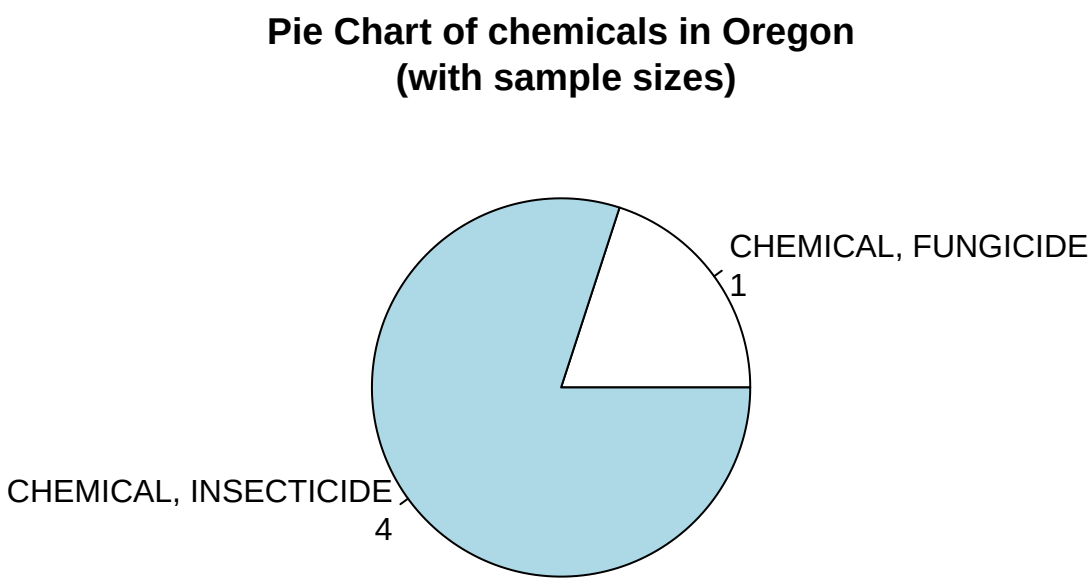
We are curious about if usage of chemicals is different among states on number of categories.




```
FLtable <- table(FL$Domain)
lbls <- paste(names(FLtable), "\n", FLtable, sep="")
pie(FLtable, labels = lbls,
   main="Pie Chart of chemicals in Florida\n (with sample sizes)")
```




```
ORtable <- table(OR$Domain)
lbls <- paste(names(ORtable), "\n", ORtable, sep="")
pie(ORtable, labels = lbls,
   main="Pie Chart of chemicals in Oregon\n (with sample sizes)")
```



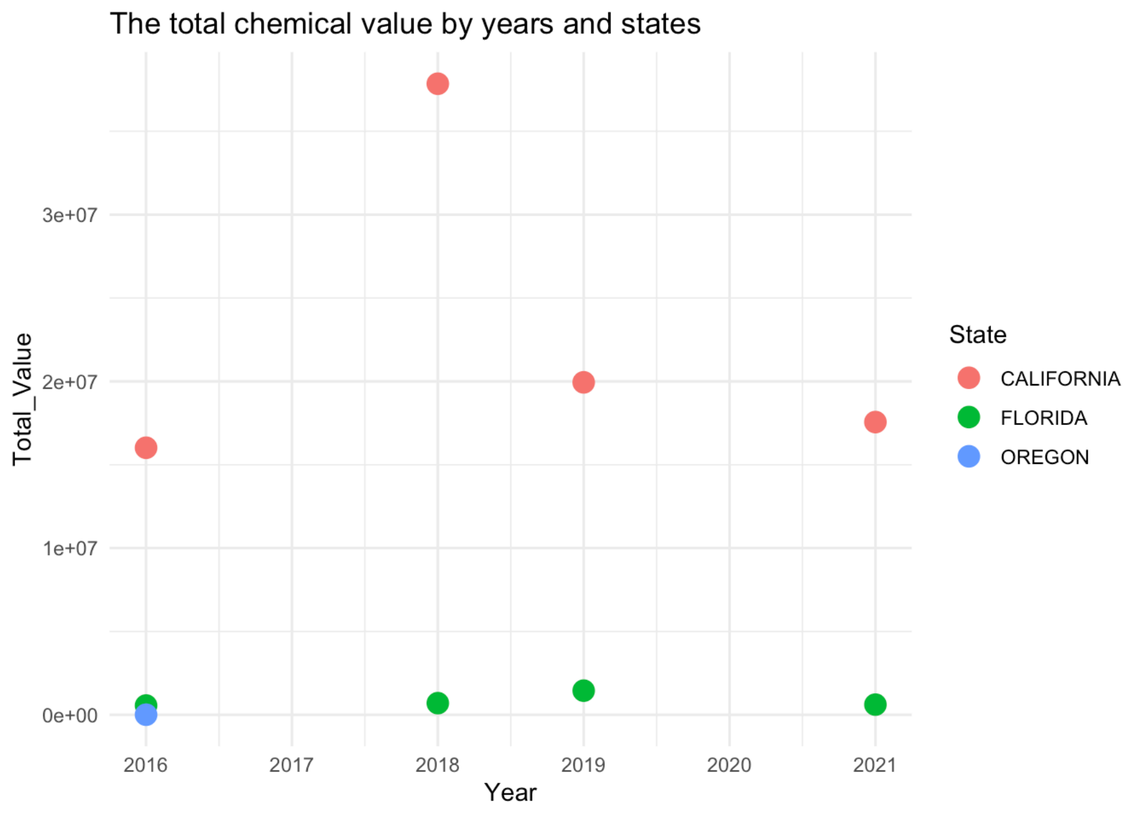
As we can see from those pie charts, California not only has the heaviest usage of chemicals on strawberries, which is 955, but also has most domains of 5. Florida also has 5 domains of chemicals, but the number of chemicals is much less than California, using 171 chemicals in total. The situation is the best in Oregon, only 5 kinds of chemicals, and they are divided into 2 domains. We also observed that all three states rely on the use of fungicide the most compare to other chemicals. Based on our research, they are sprayed to prevent strawberry fungal disease, such as powdery mildew and fruit rot, grow resistant varieties.

```
## [1] "FUNGICIDE"   "HERBICIDE"   "INSECTICIDE" "OTHER"       "FERTILIZER"
```

```
data_year_state <- group_by(strawb_chem,Year,State) %>% summarise(Total_Value = sum(Value))
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

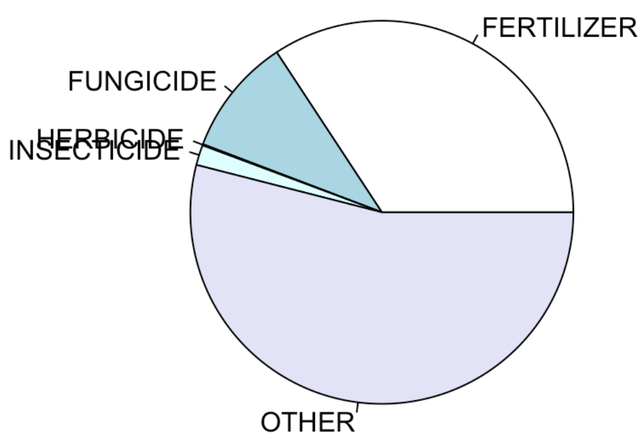
```
ggplot(data_year_state) +
  aes(x = Year, y = Total_Value, colour = State) +
  geom_point(shape = "circle", size = 4L) +
  scale_color_hue(direction = 1) +
  labs(title = "The total chemical value by years and states") +
  theme_minimal()
```




```
rm(data_year_state)
```

For Oregon state, there is only chemical value in 2016. For Florida state, chemical values are used in 2016, 2018, 2019, 2021, but in small amounts. For California state, chemical values are used in 2016, 2018, 2019, 2021 with large amounts.

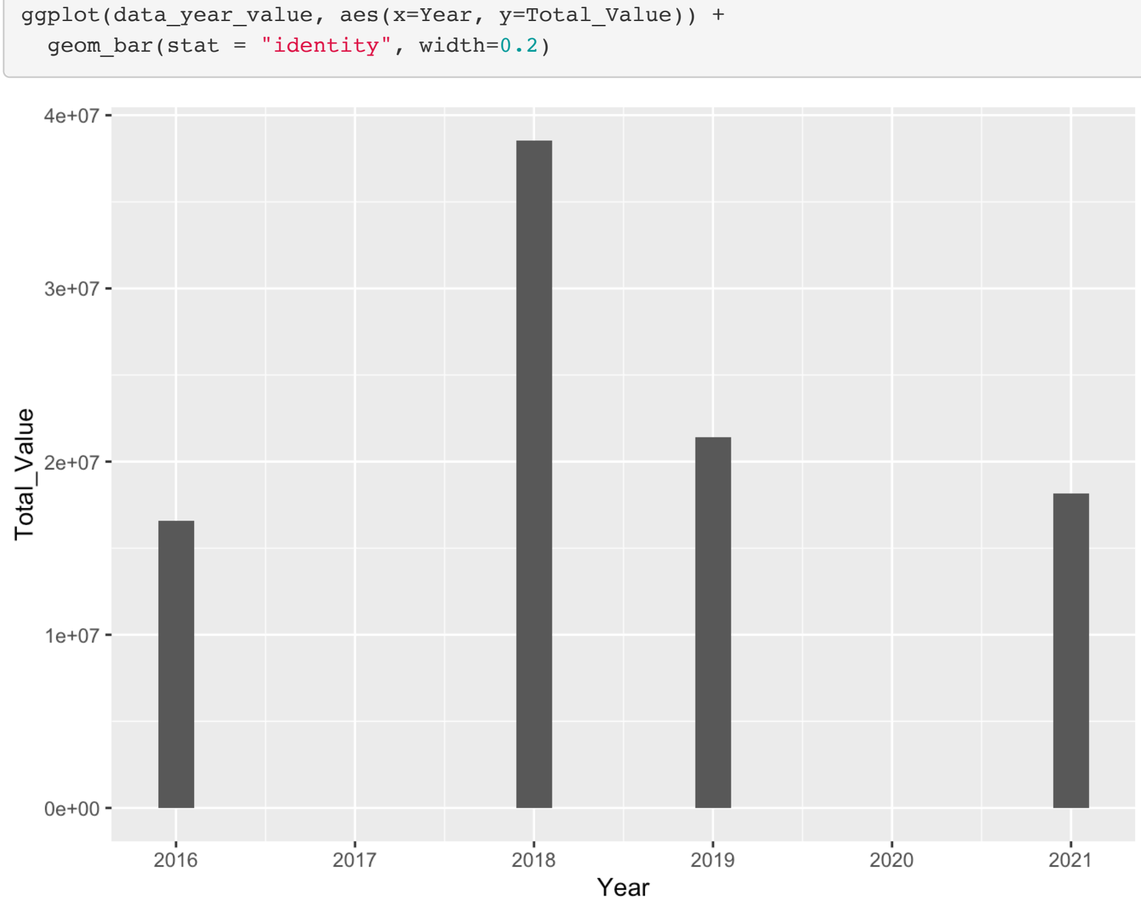
```
data_year_type <- group_by(strawb_chem,chem_types) %>% summarise(Total_Value = sum(Value))
pie(data_year_type$Total_Value, labels = data_year_type$chem_types)
```



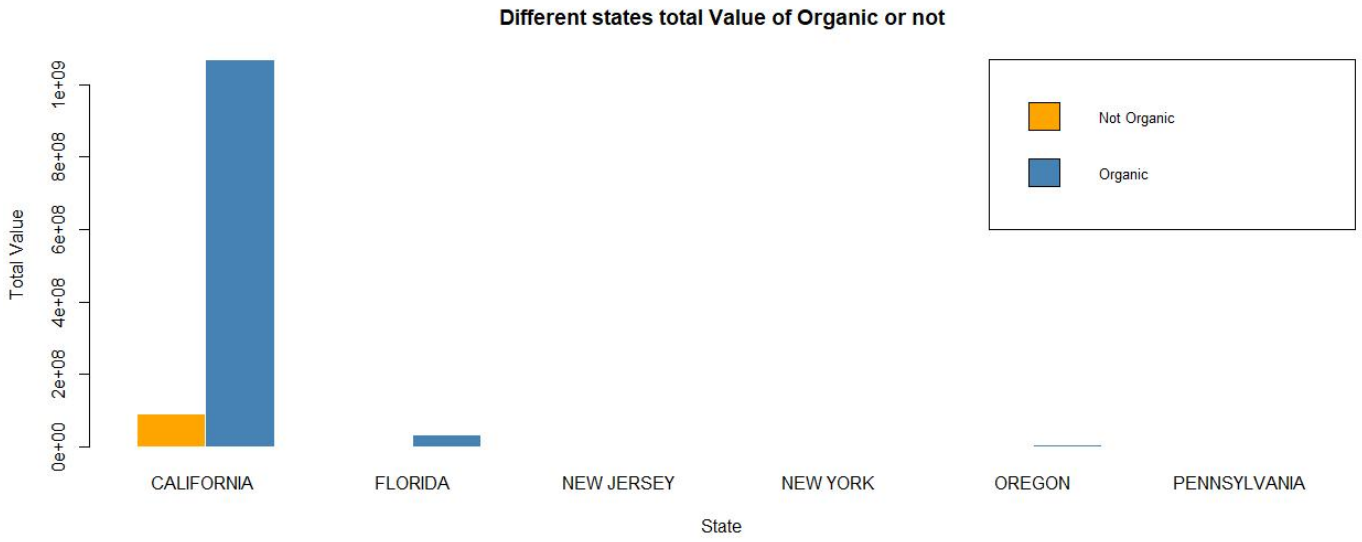

```
rm(data_year_type)
```

As shown in the pie chart, excluding other types, the use of fertilizers is the highest, followed by the use of fungicide.In this case, we should use most fertilizers in chemical categories.

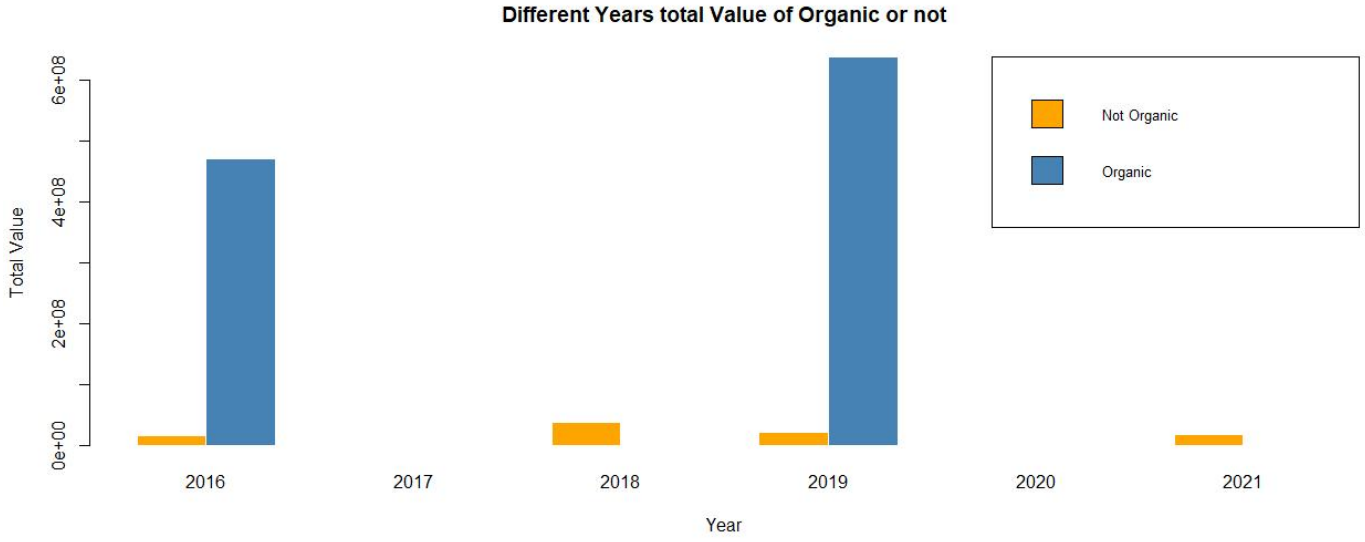
```
data_year_value <- group_by(strawb_chem,Year) %>% summarise(Total_Value = sum(Value))
ggplot(data_year_value, aes(x=Year, y=Total_Value)) +
  geom_bar(stat = "identity", width=0.2)
```



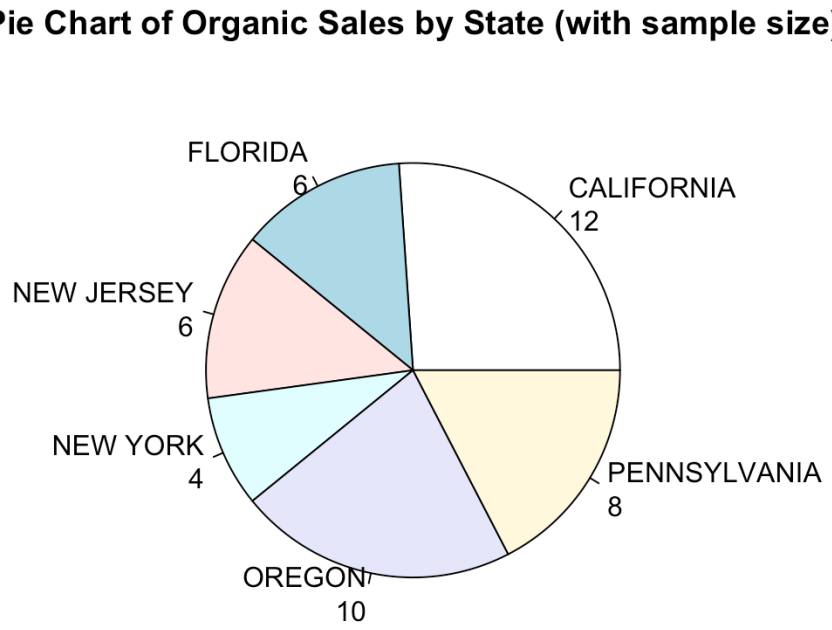
As the bar chart shows, the usage of chemical fertilizers is highest in 2018 and the usage of chemical fertilizers if zero in 2017 and 2020.



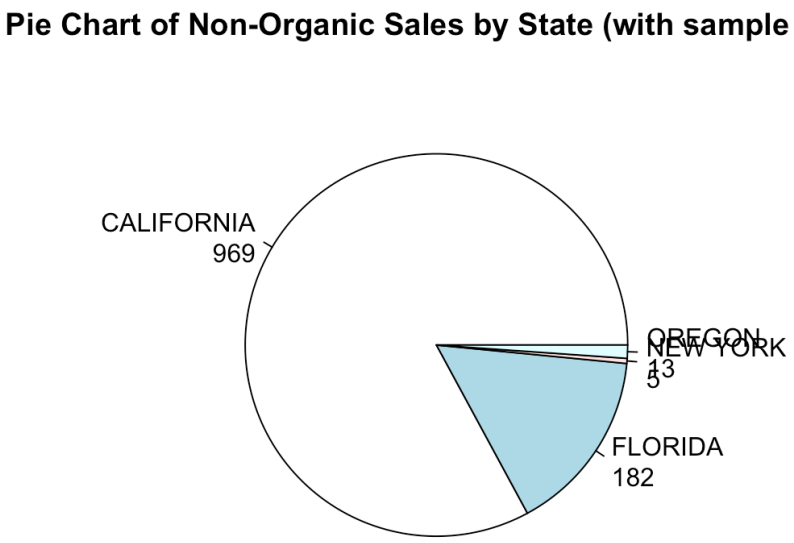
California sales the highest total values of strawberries. The sales value of California is much higher than the sum of 5 other states.
California, Florida, and Oregon sales organic strawberries. The total sales value of organic strawberries of all three states is higher than their sales of not organic strawberries.



In the year 2019 sales the highest total values of strawberries.
Only in 2016 and 2019 sales organic strawberries. In 2016 and 2019, the sales value of organic strawberries is much higher than the sales value of not organic strawberries.



From the above graph, we can find that the California has the most count of sales in organic strawberry in the dataset, and the New York has the least sales in count.



From the above graph, we can find that the California has the most count of sales in non-organic strawberry in the dataset, and the New York has the least sales in count. So we can understand that the California has the most sample size and the New York has the least sample size. Comparing this to the Organic and Non- organic inside each state, we can find more about how it influence the sales.