Midterm Strawberry - EDA

Yeyang Han, Tianjian Xie, Hao He, Yaquan Yang

2022-11-09

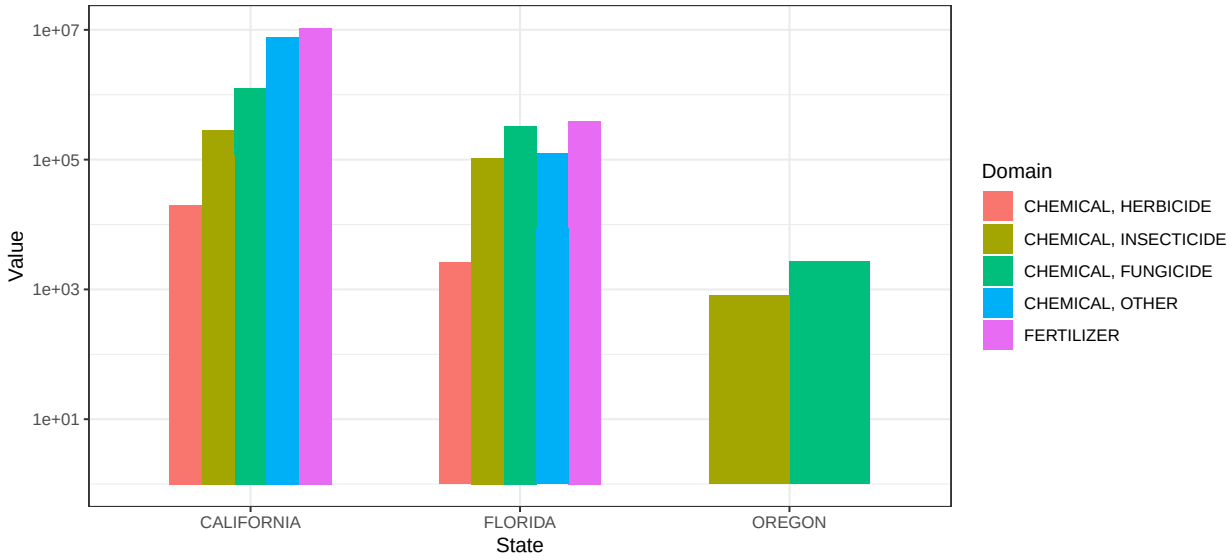
Read data and data wrangling

First, read the dataset into Rstudio. After having an overview of all the columns and their data structure, we start with dropping useless columns. Then we cleaned the Data Item column, but we still need to take care of Domain and Domain Category columns.

```
# strawb
strawb <- read_xlsx("strawberries-2022oct30-a.xlsx")

# Get the column names and index them
cnames <- colnames(strawb)
x <- 1:dim(strawb)[2]

# Delete useless columns
T <- NULL
for(i in x){T <- c(T, dim(unique(strawb[i]))[1])}
drop_cols <- cnames[which(T == 1)]
strawb <-strawb %>% dplyr::select(!all_of(drop_cols))
strawb <-strawb %>% arrange(Year, State)

# Separate many information stored in `Data Item` column into 4 columns
strawb <- strawb %>% separate(col=`Data Item`,
                    into = c("Strawberries", "type", "items", "units"),
                    sep = ",",
                    fill = "right")
```

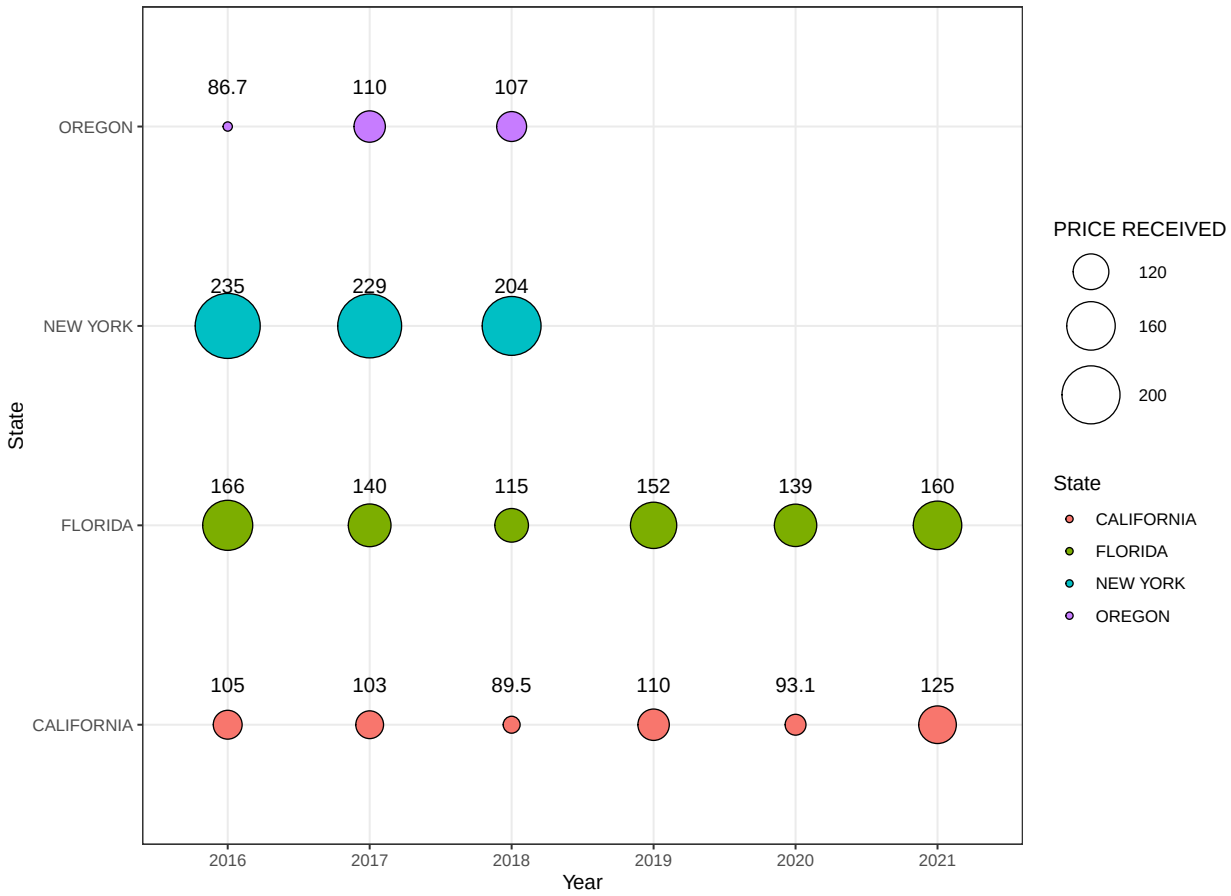
Before we separate these two columns, we create two figures to see if there are any patterns for chemicals used in different states and have a initial view of prices for strawberries.

Figure 1 − CHEMICAL usage in different states



In Figure 1 we can see that there are three states that use chemicals and fertilizers: California, Florida and Oregon. California uses more fertilizers and chemicals overall than the other two states. For California, the most used are fertilizer, followed by chemical-other, fungicide, pesticides and herbicides. Fertilizer and fungicides were the most used in Florida.

Figure 2 −STRAWBERRIES − PRICE RECEIVED



In Figure 2 we can see information on strawberry price received in the five states from 2016-2021. New York and Florida have higher values in this category than other states.

Now, looking at the `domain category` column, it may be better for further analysis if we separate the *strawb* dataset into three subsets: organic, non-organic – and commercial vs chemicals in each. *chemicals used in strawberry cultivation* (pesticides, insecticides, fertilizers, fungicides, herbicides, and others) *sales of organic strawberries sales of non-organic strawberries*

We starts with filtering all the organic entries from the whole dataset to create organic and non-organic subsets. Next, we continue to separate the chemical data from non-organic data because they have more elaborate information of chemicals than that of organic data. Finally, we got all rows of chemical data for creating the chemicals subset.

```
# Find all organic entries

# Type: 62 rows
type_organic <- grep("organic",
                     strawb$type,
                     ignore.case = T)
# Zero rows returned
items_organic <- grep("organic",
                      strawb$items,
                      ignore.case = T)  ## nothing here
# Domain: 62 rows
Domain_organic <- grep("organic",
                       strawb$Domain,
                       ignore.case = T)
# Domain Category: 62 rows
Domain_Category_organic <- grep("organic",
                                strawb$`Domain Category`,
                                ignore.case = T)
org_rows <- intersect(type_organic, Domain_organic)

# Use slice function get organic and non-organic strawberries subsets
strawb_organic <- strawb %>% slice(org_rows, preserve = FALSE)
strawb_non_organic <- strawb %>% filter(!row_number() %in% org_rows)

# Checmicals (used in strawberry cultivation) subset
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)
```

Cleaning three tibbles (subsets)

Chemical data subset

Write a function to drop columns that contain useless information. It's easy to find that after we drop some useless columns, the `units` column still contain NA values in the form of character data. Also, we need to separate Domain Category column to get a detailed and tidy data frame.

```
## drop useless columns
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 <- strawb_chem %>% dplyr::select(!all_of(drop_cols))
after_cols = colnames(strawb_chem)
temp1 <- strawb_chem %>% dplyr::select(units) %>% distinct()
strawb_chem <- strawb_chem %>% separate(col=`Domain Category`,
                          into = c("dc1", "chem_name"),
                          sep = ":",
                          fill = "right")
aa  <- grep("measured in",
            strawb_chem$items,
            ignore.case = T)
length(aa)

temp1 <- strawb_chem %>% dplyr::select(chem_name) %>% unique()
# length(unlist(temp1))

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

strawb_chem <-strawb_chem %>% dplyr::select(Year, State, items, units, dc1, chem_name, Value)
strawb_chem$items <- str_remove_all(strawb_chem$items, "MEASURED IN ")
strawb_chem <- strawb_chem %>% rename(c(category = units, units = items))

## Do all the dc1 entries begen with "Chemical"?

bb  <- grep("CHEMICAL, ",
            strawb_chem$dc1,
            ignore.case = T)
# length(bb)
chem <- 1:2112

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

## Now let's look at these rows in a tibble

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

```
## These non_chem_rows refers to fertilizers

## keep them

fertilizers <- temp1

## Now clean the chem_name column.
## now remove "CHEMICAL, " from the entries in the dc1
## and rename the column chem_types


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

# strawb_chem$dc1 %>% unique()

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

## now fix the chem_name column

## remove the parens

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

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

## separate chem_name and chem_code

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


## now fill in a label fot NA in the category column

## first check that "lb" in the units column corresponds
## to NA in the category column


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

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

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

After getting rid of redundant strings after separating `Domain` and `Domain Category` columns, chemicals now have an tidy individual column for us identify a specific type of chemicals used in strawberry cultivation.Besides, we found 45 entries starts with some other strings: “FERTILIZER” instead of “CHEMICAL”, we decide to keep them for a full analysis of chemicals.

```
##-----------
# Clean up the workspace
```

```
rm(aa,after_cols,b,bb,before_cols,chem,chem_rows,chem_rows_1,chem_rows_2,cnames,
   i,drop_cols,x,T,type_organic,org_rows,ins,ins_2,items_organic,non_chem_rows,
   Domain_Category_organic,Domain_organic,temp1)
```

```
# strawb_chem %>% select(chem_name) %>% distinct()
```

So far, we have a clean chemicals subset. The number of different chemicals listed in our dataset is 172.

Organic strawberry subset

Next subset to clean is organic data. We apply similar approaches to select columns, and expand columns to create a detailed and tidy data frame.

```
##begin to clean organic data
temp1 <- strawb_organic %>% dplyr::select(Year,State) %>% distinct()

# unique(strawb_organic$`Domain`)
# unique(strawb_organic$`Domain Category`)
# unique(strawb_organic$units)
strawb_organic$units <- str_remove_all(strawb_organic$units, "MEASURED IN ")

#clean organic strawberry data
strawb_organic <- strawb_organic[!names(strawb_organic) %in% c("Period","State ANSI","Strawberries", "D
#type, items and units need to be cleaned
strawb_organic
```

```
## # A tibble: 62 x 8
##    Program  Year State      type               items         units Value CV (%~1
##    <chr>   <dbl> <chr>      <chr>              <chr>         <chr> <chr> <chr>
##  1 CENSUS   2016 CALIFORNIA " ORGANIC - SALES" " MEASURED I~  <NA> 2313~ 13.7
##  2 CENSUS   2016 CALIFORNIA " ORGANIC - SALES" " MEASURED I~  <NA> 1446~ 13.5
##  3 CENSUS   2016 CALIFORNIA " ORGANIC"         " FRESH MARK~ " $"  2115~ 13.5
##  4 CENSUS   2016 CALIFORNIA " ORGANIC"         " FRESH MARK~ " CW~ 1221~ 12.9
##  5 CENSUS   2016 CALIFORNIA " ORGANIC"         " PROCESSING~ " $"  1975~ 19.7
##  6 CENSUS   2016 CALIFORNIA " ORGANIC"         " PROCESSING~ " CW~ 2248~ 21
##  7 CENSUS   2016 FLORIDA    " ORGANIC - SALES" " MEASURED I~  <NA> 2455~ 21.9
##  8 CENSUS   2016 FLORIDA    " ORGANIC - SALES" " MEASURED I~  <NA> 14532 23.1
##  9 CENSUS   2016 FLORIDA    " ORGANIC"         " FRESH MARK~ " $"  (D)   (D)
## 10 CENSUS   2016 FLORIDA    " ORGANIC"         " FRESH MARK~ " CW~ (D)   (D)
## # ... with 52 more rows, and abbreviated variable name 1: 'CV (%)'
```

non-organic strawberry data

Similarly, we have a much cleaner non-organic data frame. Now we decided to divide the non-organic data subset into marketing year and year subsets for further analysis.

```
##begin to clean non_organic data
# unique(strawb_non_organic$State) # "CALIFORNIA" "FLORIDA"    "NEW YORK"   "OREGON"

# unique(strawb_non_organic$Strawberries)
# "STRAWBERRIES - PRICE RECEIVED" "STRAWBERRIES"
```

```r
# strawb_non_organic[strawb_non_organic$Strawberries=="STRAWBERRIES - PRICE RECEIVED",]
#all in marketing year

# unique(strawb_non_organic$type)
# " MEASURED IN $ / CWT"; " FRESH MARKET - PRICE RECEIVED";
#" PROCESSING - PRICE RECEIVED"; " BEARING - APPLICATIONS"
# strawb_non_organic %>% dplyr::select(type, items) %>% unique()
# strawb_non_organic[strawb_non_organic$Strawberries=="STRAWBERRIES - PRICE RECEIVED",]
# strawb_non_organic[strawb_non_organic$type==" MEASURED IN $ / CWT",]
# length(which(is.na(strawb_non_organic$items)))

#when Strawberries==STRAWBERRIES - PRICE RECEIVED, tyoe==MEASURED IN $ / CWT, items==NA
#at this time, we will see all Period=="MARKETING YEAR".
#So, is there any relationship between them?
x5 <- strawb_non_organic[strawb_non_organic$Period=="MARKETING YEAR",]

# About all Period=="MARKETING YEAR", Domain=="TOTAL", types==  "NOT SPECIFIED"
# if we do not have "MARKETING YEAR", is there some pattern about value?
x6 <- setdiff(strawb_non_organic,x5)
# only marketing year have " MEASURED IN $ / CWT", " FRESH MARKET - PRICE RECEIVED" and " PROCESSING -
# year only have " BEARING - APPLICATIONS" type
# unique(strawb_non_organic$items)
strawb_non_organic$items <- str_remove_all(strawb_non_organic$items, "MEASURED IN ")


# unique(strawb_non_organic$units) # same operation just like chem
#delete directly

# unique(strawb_non_organic$Domain)
#delete

# unique(strawb_non_organic$`Domain Category`)
strawb_non_organic <- strawb_non_organic %>% separate(col=`Domain Category`,
                          into = c("types", "name"),
                          sep = ":",
                          fill = "right")

# strawb_non_organic %>% dplyr::select(name) %>% unique()
# sum(strawb_non_organic$Domain == strawb_non_organic$types) == dim(strawb_non_organic)[1] #FALSE

# Now, we know the difference between them
x1 <- strawb_non_organic[(strawb_non_organic$Domain == strawb_non_organic$types) == FALSE,]
#All the Domain==TOTAL and types == NOT SPECIFIED
#delete Domain, keep types.
## Do all the types entries begen with "Chemical"?
x2 <- grep("CHEMICAL, ",
           strawb_non_organic$types,
            ignore.case = T)
# length(x2)
##if they are not entries begen with "Chemical", what kind of thing they begin with?
x3 <- strawb_non_organic[grepl("CHEMICAL, ",strawb_non_organic$types),]
# nrow(x3)
x4 <- setdiff(strawb_non_organic, x3)
```

```
# View(x4)
# unique(x4$types)
#"NOT SPECIFIED" "FERTILIZER"
#at this time, we can divide subset into chemical and FERTILIZER
```

By checking the unique values of `types` columns, we decide to divide non-organic dataset into chemical and FERTILIZER subsets.

```
#clean chem names
# strawb_non_organic %>% dplyr::select(name) %>% unique() # 173
#remove the parens
strawb_non_organic$name <- str_remove_all(strawb_non_organic$name, "\\(")
strawb_non_organic$name <- str_remove_all(strawb_non_organic$name, "\\)")
## separate name and code
strawb_non_organic <- strawb_non_organic %>% separate(col = name,
                          into = c("name","code"),
                          sep = "=",
                          fill = "right"
)

#delete useless columns
strawb_non_organic <- strawb_non_organic[!names(strawb_non_organic) %in% c("State ANSI","types")]

#final slice non-organic strawberries dataset into 3 dataset:
#first, slice data according to Period (marketing year and year);
#Next, slice year data according to Domain or chemical
strawb_non_organic_my <- strawb_non_organic[strawb_non_organic$Period=="MARKETING YEAR",]
strawb_non_organic_y <- setdiff(strawb_non_organic, strawb_non_organic_my)
strawb_non_organic_chemical <- strawb_non_organic[grepl("CHEMICAL, ",strawb_non_organic$Domain),]
strawb_non_organic_fertilizers <- setdiff(strawb_non_organic_y, strawb_non_organic_chemical)
```

Now, we have all datasets tidied: * strawb_organic * strawb_non_organic_my strawb_non_organic_chemical strawb_non_organic_fertilizers*

Exploratory Data Analysis

According to the Shopper's Guide to Pesticides in Produce article, people now are more aware of the public health, so organic strawberries will be more and more important in the future since there are less hazardous chemicals used to grow strawberries. This leads us to explore organic strawberry subset first.

organic strawberry

NEW JERSEY have no data for `Value`column in 2019. But we can see its CWT decrease. So, we impute this NA according to proportion.

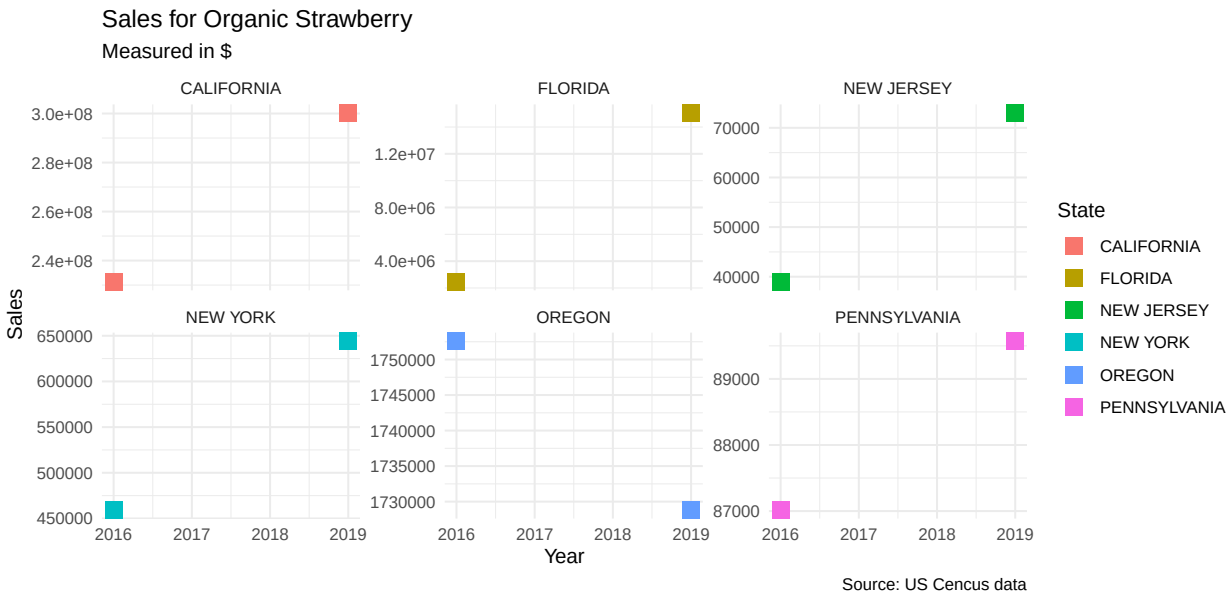
```
e1 <- strawb_organic[strawb_organic$type==" ORGANIC - SALES", ]
# View(e1)


e1$Value <- as.numeric(e1$Value)
```

```
e1[17, "Value"] = 73017.4
e1m <- e1[e1$items == " MEASURED IN $", ]
e1w <- e1[e1$items == " MEASURED IN CWT", ]
datao <- data.frame(Year = e1$Year,State = e1$State, Value = e1$Value)
#first
data1 <- e1m %>% dplyr::select("Year", "State", "Value")
t1 <- e1m %>% pivot_wider(names_from = State, values_from = Value)
```

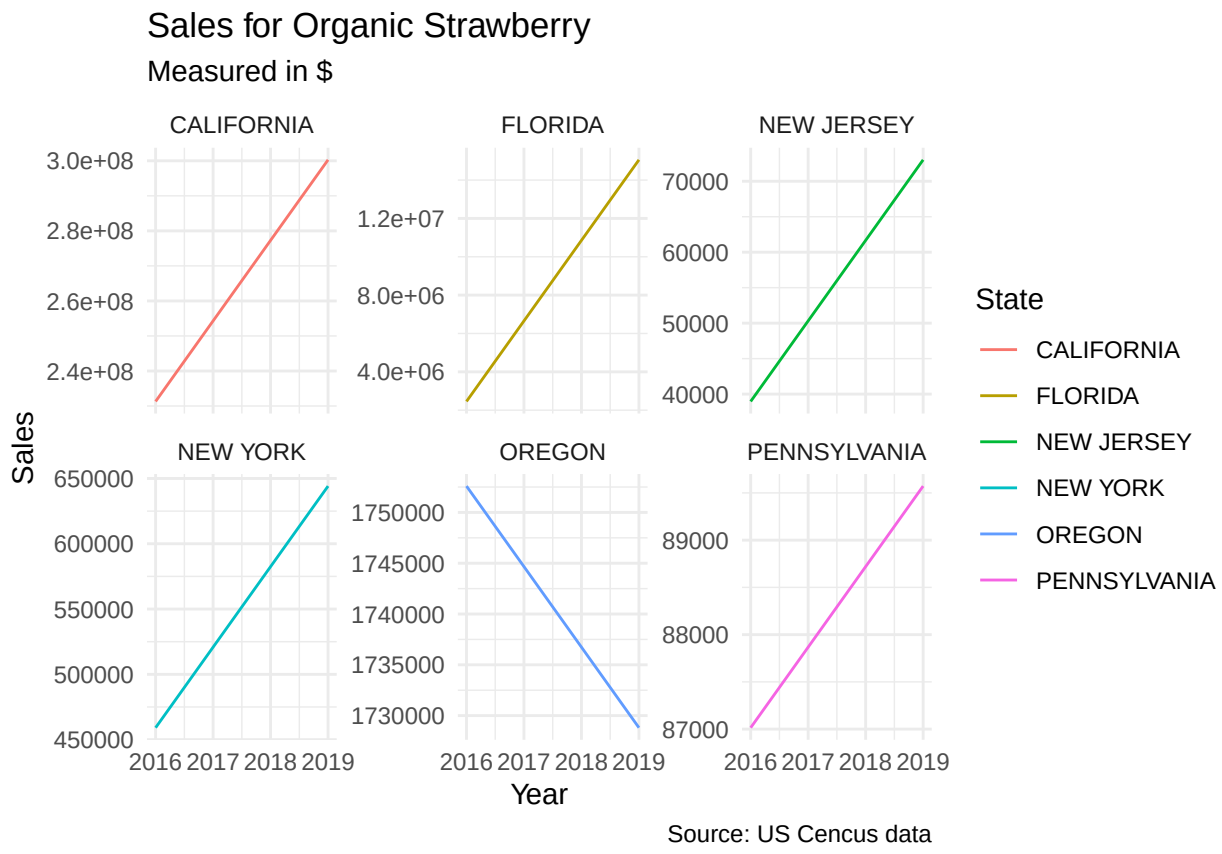
Sales trend for organic strawberries

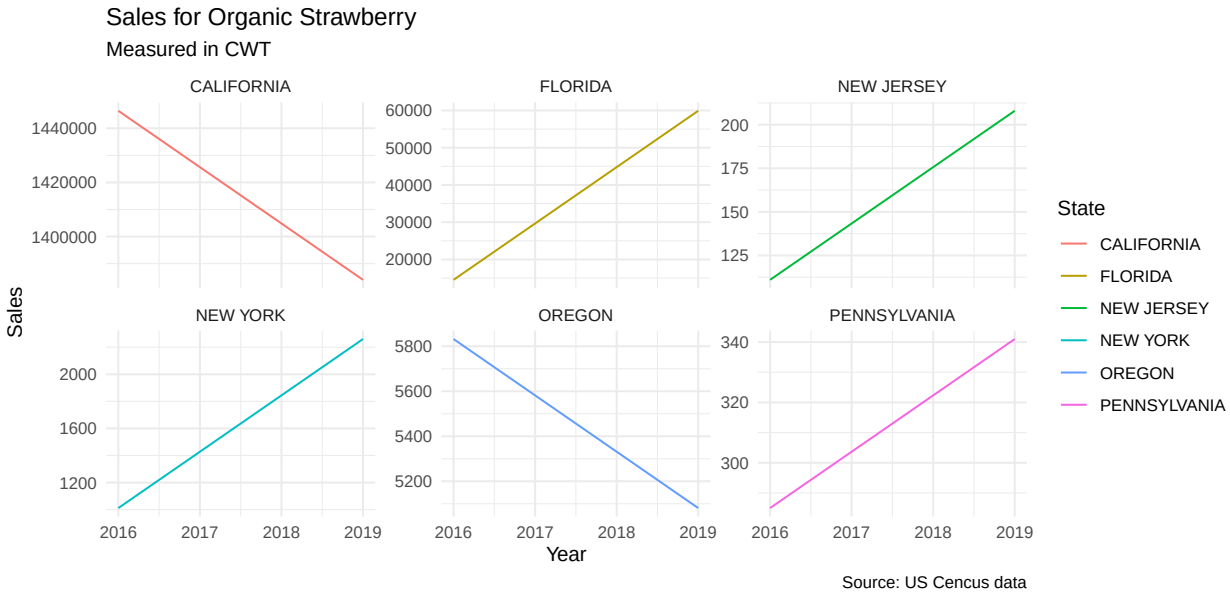
Since the sales data is collected either measured in $ or CWT. We draw two plots for each measurement to avoid any confusion that may exist.

Below is a plot that describes the sales measured in $ for organic strawberries in 2016 and 2019.



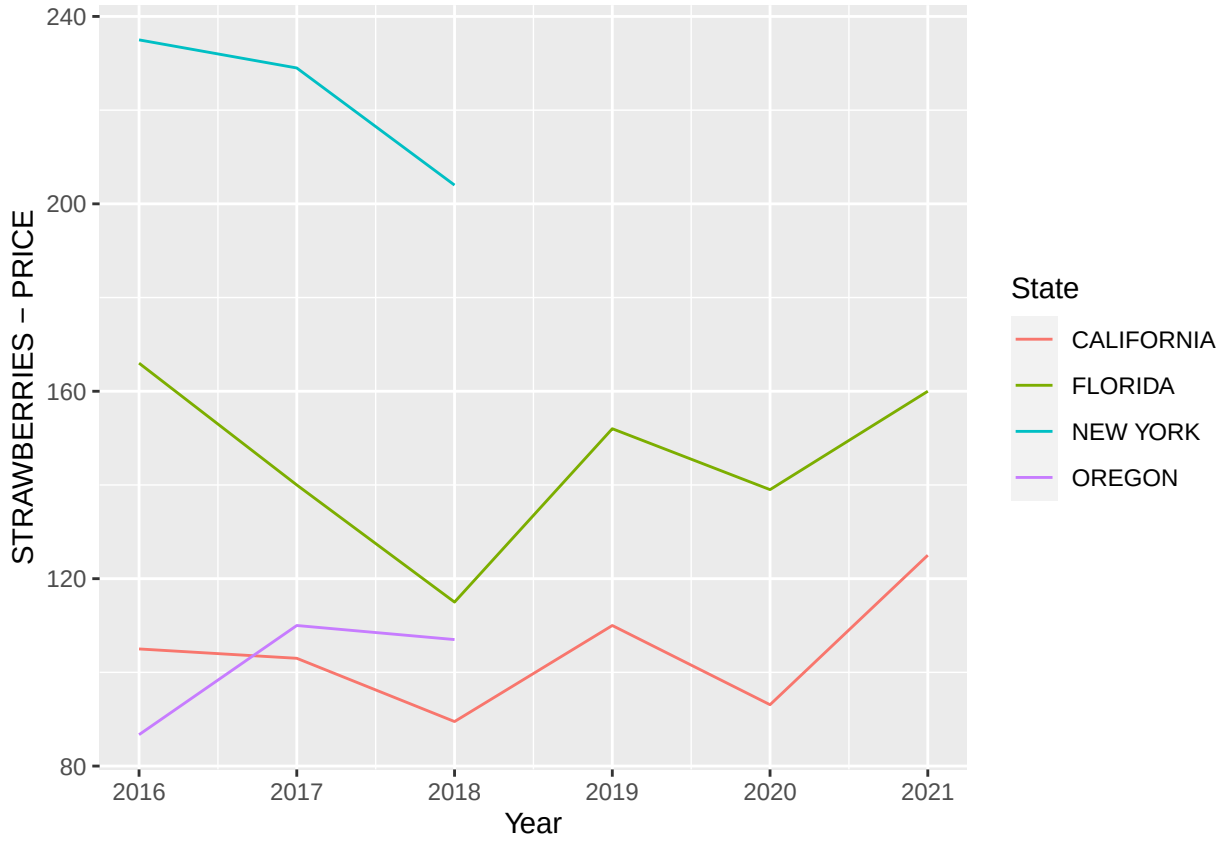
To have a better sense of whethe the sales increase or decrease, we change the plot type into line instead of point.





Above is a plot that describes the sales measured in CWT for organic strawberries. We find there is a discrepancy for sales in California using different measurements. Without full information of the quantities and prices and why some values of coefficient variation are missing, We cannot infer anything about why this discrepancy would appear. We believe that if the conversion between CWT and sales can be added into data preparation, we will be able to discover more convicing results about how sales change between different states during 2016 - 2019.

non_organic:



As for the year of collected data, we find that is we want to analysis sales of per CWT in different states there are so many nan values. However, if we choose the marketing year data, all the strawberries are non organic strawberries. And during 2016-2021, most of states sold of per CWT decreased compared with 2016, except California. In addition, the record of saling of per CWT in New York and Oregon disappeared from 2018. But we do not know why. In the future, we can find more documents to know the reason of that.

fertilizers

```
## [1] "CALIFORNIA" "FLORIDA"
```

```
## [1] " NITROGEN"  " PHOSPHATE" " POTASH"    " SULFUR"
```

These fertilizers are natural and safe for the growth of strawberries.

chemical

Analysis on state

Since California contributes to 80% of the strawberry production and followed by Florida and Oregon, we are interested how many different chemicals are listed by state and if California use more different chemicals than other states.

```
# number of different chemicals listed by different states
# strawb_chem %>% group_by(State) %>% summarise(unqiue_chems = n_distinct(chem_name))
```

Based on our analysis, we can say that California has a higher usage of different chemicals(139) than others (116 for Florida and 21 for Oregon) and this could be one of reasons why California has the highest production.

```
unique(strawb$State)
```

```
## [1] "CALIFORNIA"   "FLORIDA"      "NEW JERSEY"   "NEW YORK"     "OREGON"
## [6] "PENNSYLVANIA"
```

By examine the `State` column, we can see other states including "NEW JERSEY", "NEW YORK" and "PENNSYLVANIA" are involved in strawberry roduction. If we have more information about more states involved , we can perform an regional analysis in U.S where New Jersey, New York, and Pennsylvania can be grouped together as Northeast and Mid-Atlantic United States (Northeast region).

```
## [1] "CHEMICAL"
```

```
## [1] " LB"                      " LB / ACRE / APPLICATION"
## [3] " LB / ACRE / YEAR"
```

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

```
## [1] 171
```

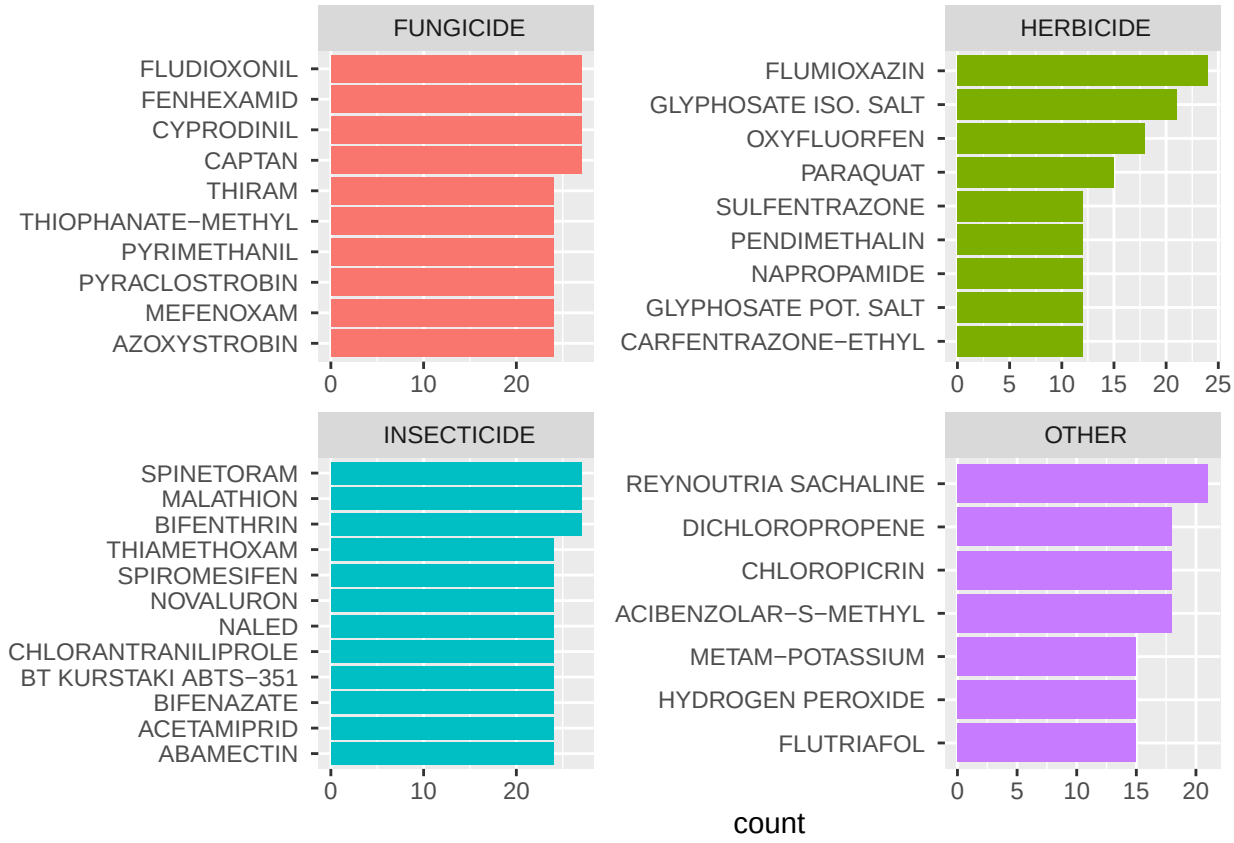
```
## # A tibble: 4 x 2
## # Groups:   detail [4]
##   detail             n
##   <chr>          <int>
## 1 " FUNGICIDE"     771
## 2 " HERBICIDE"     192
## 3 " INSECTICIDE"   795
## 4 " OTHER"         309
```

Hazardous chemicals and safe ones

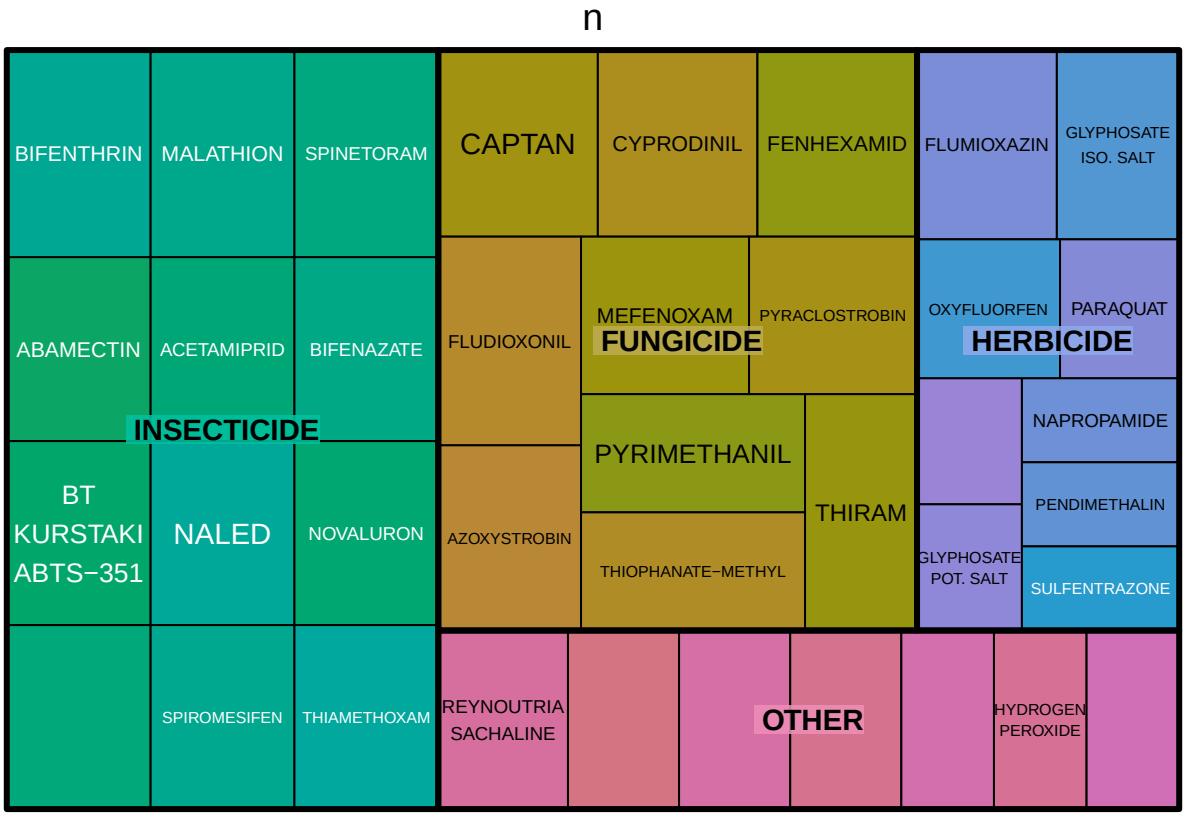
We try to find how many hazardous chemicals mentioned in the "Shoppers Guide to Pesticides in Produce" exists in our dataset. Hazardous chemicals include Carbendazim, Bifenthrin, methyl bromide, 1,3-dichloropropene, chloropicrin, Telone.

```
## Carbendazim, Bifenthrin, methyl bromide, 1,3-dichloropropene, chloropicrin, Telone
df_carbendazim <- grep("carbendazim",
                       e4$`name`, ignore.case = T)
# length(df_carbendazim) #0
df_Bifenthrin <- grep("Bifenthrin",
                       e4$`name`, ignore.case = T)
# length(df_Bifenthrin) #27 #INSECTICIDE
df_methyl_bromide <- grep("methyl bromide",
                       e4$`name`, ignore.case = T)
```

```r
# length(df_methyl_bromide) #3 #other
df_1_3_dichloropropene <- grep("1,3-dichloropropene",
                               e4$`name`,
                               ignore.case = T)
# length(df_1_3_dichloropropene) #0
df_chloropicrin <- grep("chloropicrin",
                        e4$`name`,
                        ignore.case = T)
# length(df_chloropicrin) #18 #other
df_Telone <- grep("Telone",
                  e4$`name`,
                  ignore.case = T)
# length(df_Telone) #0
e5 <- e4 %>%
  group_by(detail) %>%
  count(name) %>%
  arrange(detail,desc(n)) %>%
  slice_max(n, n = 5)
e6 <- e4 %>%
  group_by(detail) %>%
  count(name) %>%
  arrange(detail,desc(n)) %>%
  slice_max(n, n = 5)
e7 <- e6 %>%
  group_by(detail) %>%
  count(detail)
e5 %>%
  ggplot(aes(n,fct_reorder(name, n),fill = detail)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~detail, ncol = 2, scales = "free") +
  labs(x = "count", y = NULL)
```




```
treemap(e5,
        index=c("detail","name"),
        vSize="n",
        type="index"
)
```



A great number of chemicals are used when planting strawberries. We can category them into four types — fungicide, insecticide, herbicide and others. At here, we substract top5 from each categories to see the situation. According to the picture, we find most people choose to use insecticide and the proportion of people who use others as chemicals is lowest.

Based on our research, safe and natural chemicals to use include Neem Oil and Spinosad. We run a quick check if they are in our chemical dataset.

```
# Neem Oil: 27 matches
neem_oil_rows<- grep("NEEM OIL", strawb_chem$chem_name, ignore.case=TRUE)
df_Neem_Oil <- strawb_chem%>% slice(neem_oil_rows)
tableNeem_Oil<- df_Neem_Oil %>% group_by(State) %>% count()

#Spinosad: 12 matches
spinosad_rows<- grep("SPINOSAD", strawb_chem$chem_name, ignore.case=TRUE)
df_Spinosad <- strawb_chem%>% slice(spinosad_rows)
tableSpinosad_state<- df_Spinosad %>% group_by(State) %>% count()
```

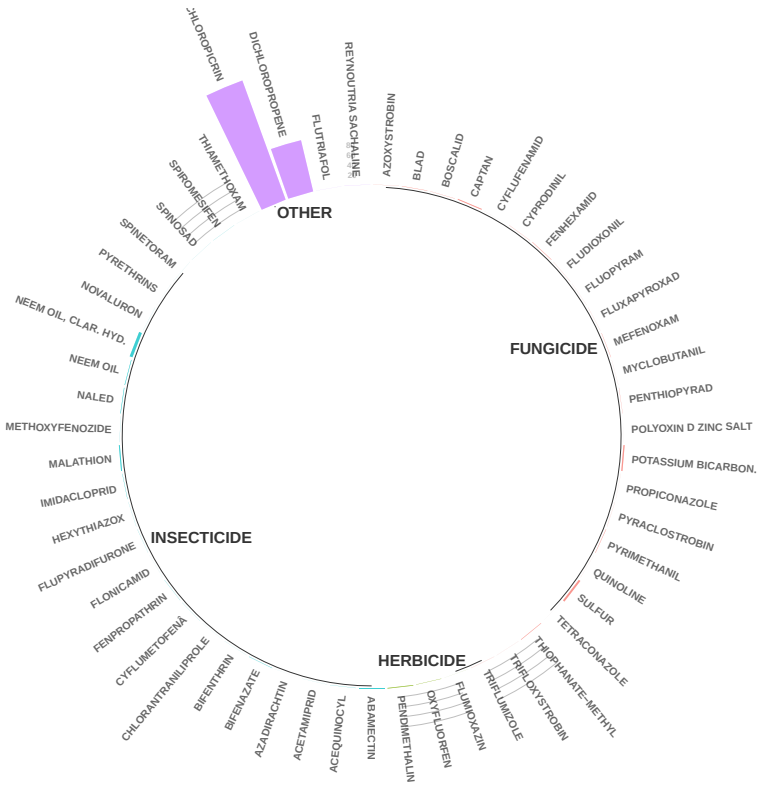
Based on our checking, California has 24 matches of neem oil usage and 12 matches of spinosad (while Florida only has 12 matches of neem oil usage). California seems to be cautious about the chemicals relatively as the vast majority of the fresh strawberries sold in this country are grown there. But because of the limited information we have, we cannot conclude California has the best practice to protect public health.

Abolished pie plot and donut chart graphes

These two graphes show the percentage of usage of Bifenthrin grouped by Year and State. Bifenthrin is a dangerous varity of chemicals widely used in agriculture. It has been banned for agricultural use in European

union countries since July 2019, but it is not a restricted chemical in the United States and is commonly sold in hardware stores. I planned to use these graphs to show the usage of Bifenthrin in California and Florida by years and made a comparison by percentage, but it didn't worked well. So we don't want to rely on these graphs for producing any further analysis.

Circular barplot



This plot is a circular barplot for the LB/Year/App usage of different chemicals in California 2016, grouped by the four different types of chemicals: Fungicide, Herbicide, Insecticide, and the others. The output shows that CHLOROPICRIN is the most common used chemical in measurement of LB/Year/App, and the DICHLOROPROPENE is the second. This two chemicals were used far ahead other types of chemicals. However, CHLOROPICRIN is a very dangerous chemical. It was used as chemical weapons during WWI, and broadly used in agriculture nowadays. From CDC's report, CHLOROPICRIN used to cause toxication of 27 workers in a textile factory, and one death cases is reported. One year before the starting date of this dataset, California starts to limit the usage of this chemical. "In California experience with acute effects of chloropicrin when used as a soil fumigant for strawberries and other crops led to the release of regulations in January 2015 creating buffer zones and other precautions to minimize exposure of farm workers, neighbors, and passersby." ("Control Measures for Chloropicrin". California Department of Pesticide Regulation. January 6, 2015). CHILOROPICRIN is still the most common used chemical after the policy enacted.