# MID_Group4_EDA615

2022-11-08

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyve
rse 1.3.2 ──
## ✓ ggplot2 3.3.6      ✓ purrr   0.3.4
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.0      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_co
nflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ 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)
library(dplyr)
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to
use these themes.
##       Please use hrbrthemes::import_roboto_condensed() to install Ro
boto Condensed and
##       if Arial Narrow is not on your system, please see https://bit.
ly/arialnarrow
```

```
strawb <- read_xlsx("/Users/yirong/Desktop/strawberries-2022oct30-a.xls
x",
                    col_names = T)
```

## Clean Whole data set

```
cnames <- colnames(strawb)
```

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

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))

## Arrange the data frame by year and state.
strawb %<>% arrange(Year, State)

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

## Build 4 subsets

strawb_organic strawb_non_organic strawb_chem

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

items_organic <- grep("organic",
                      strawb$items,
                      ignore.case = T)  ## nothing here

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)


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)
```

##Clean strawb_organic

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

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

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

yy<- grep("MEASURED IN", strawb_organic$items, ignore.case = T)
length(yy)==sum(is.na(strawb_organic$units))

## [1] TRUE

strawb_organic$units<-coalesce(strawb_organic$units, strawb_organic$ite
ms)

strawb_organic$units<- str_remove_all(strawb_organic$units, "MEASURED I
N ")

strawb_organic$items<- str_remove_all(strawb_organic$items, "- SALES")
```

```
strawb_organic$items<- str_remove_all(strawb_organic$items, "MEASURED I
N ")

strawb_organic$items[strawb_organic$items==" $"]<-""

strawb_organic$items[strawb_organic$items==" CWT"]<-""

strawb_organic %<>% rename(Markets = items)

strawb_organic %<>% select(Year, State,Markets, units, Value,`CV (%)`)
```

##Clean Strawb_non_organic

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

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

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

strawb_non_organic %<>% separate(col=`Domain Category`,
                           into = c("dc1", "chem_name"),
                           sep = ":",
                           fill = "right")
strawb_non_organic$Domain[strawb_non_organic$Domain=="TOTAL"]<-"TOTAL/N
OT SPECIFIED"

aa  <- grep("CWT", strawb_non_organic$type,ignore.case = T)
length(aa)

## [1] 18

cc<- grep("/", strawb_non_organic$type,ignore.case = T)
length(cc)

## [1] 18

bb<-sum(is.na(strawb_non_organic$items))
bb

## [1] 18

strawb_non_organic$items<-strawb_non_organic$items %>%
  replace_na("MEASURED IN $ / CWT")
```

```
strawb_non_organic %<>% select(Year, State, items, units, dc1, chem_nam
e, Value)

strawb_non_organic %<>% rename(category = units)

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

strawb_non_organic %<>% rename(units = items)

strawb_non_organic$dc1 <- str_remove_all(strawb_non_organic$dc1, "CHEMI
CAL, ")

strawb_non_organic$dc1 %>% unique()

## [1] "NOT SPECIFIED" "FUNGICIDE"     "HERBICIDE"     "INSECTICIDE"
## [5] "OTHER"         "FERTILIZER"

strawb_non_organic%<>% rename(chem_types = dc1)

strawb_non_organic$chem_name <- str_remove_all(strawb_non_organic$chem_
name, "\\(")

strawb_non_organic$chem_name <- str_remove_all(strawb_non_organic$chem_
name, "\\)")

strawb_non_organic %<>% separate(col = chem_name,
                         into = c("chem_name","chem_code"),
                         sep = "=",
                         fill = "right"
)
qq <- grep("ACRE", strawb_non_organic$units, ignore.case = T)
ww<-grep("AVG",strawb_non_organic$category,ignore.case = T)
length(qq)==length(ww)

## [1] TRUE

strawb_non_organic %<>% select(Year, State,units, chem_types, chem_name,
 chem_code, Value)
dd <- grep("NOT SPECIFIED", strawb_non_organic$chem_types, ignore.case
= T)
sum(is.na(strawb_non_organic$chem_name))==length(dd)

## [1] TRUE

strawb_non_organic$chem_name %<>% replace_na("NONE")
strawb_non_organic$chem_code[strawb_non_organic$chem_name == "NONE"]<-"
NONE"
```

```
strawb_non_organic$chem_code[strawb_non_organic$chem_name== " TOTAL"]<-
"TOTAL"
```

##Clean Strawb_chem

```
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)

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


strawb_chem %<>% select(Year, State, items, units, dc1, chem_name, Valu
e)

strawb_chem %<>% rename(category = units)

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

strawb_chem %<>% rename(units = items)

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

strawb_chem$dc1 %>% unique()

## [1] "FUNGICIDE"   "HERBICIDE"   "INSECTICIDE" "OTHER"       "FERTILI
ZER"

strawb_chem %<>% rename(chem_types = dc1)

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"
)

strawb_chem %<>% select(Year, State,units, chem_types, chem_name, chem_
code, Value)
```

###Build safe_chem and poisons_chem

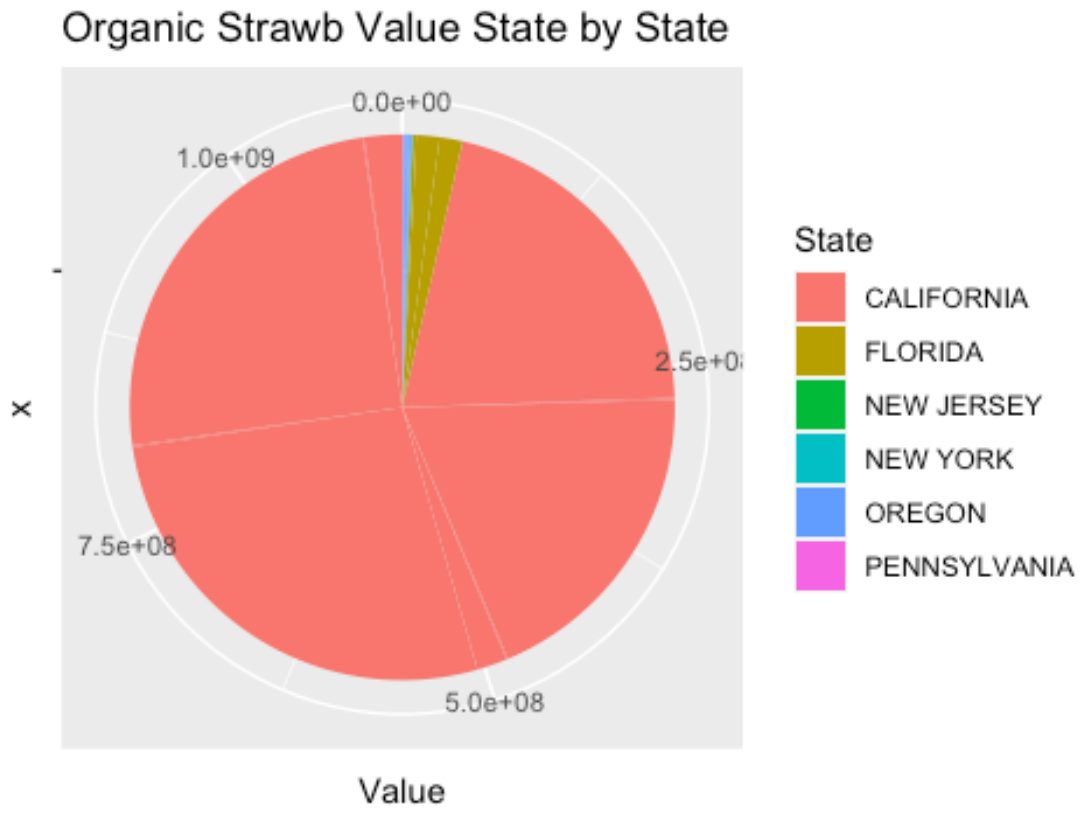
Poisons Chemicals mentioned in the article:BIFENTHRIN128825,METHYL BROMIDE,CHLOROPICRIN

Safe Chemicals: PHOSPHATE,POTASSIUM BICARBON.

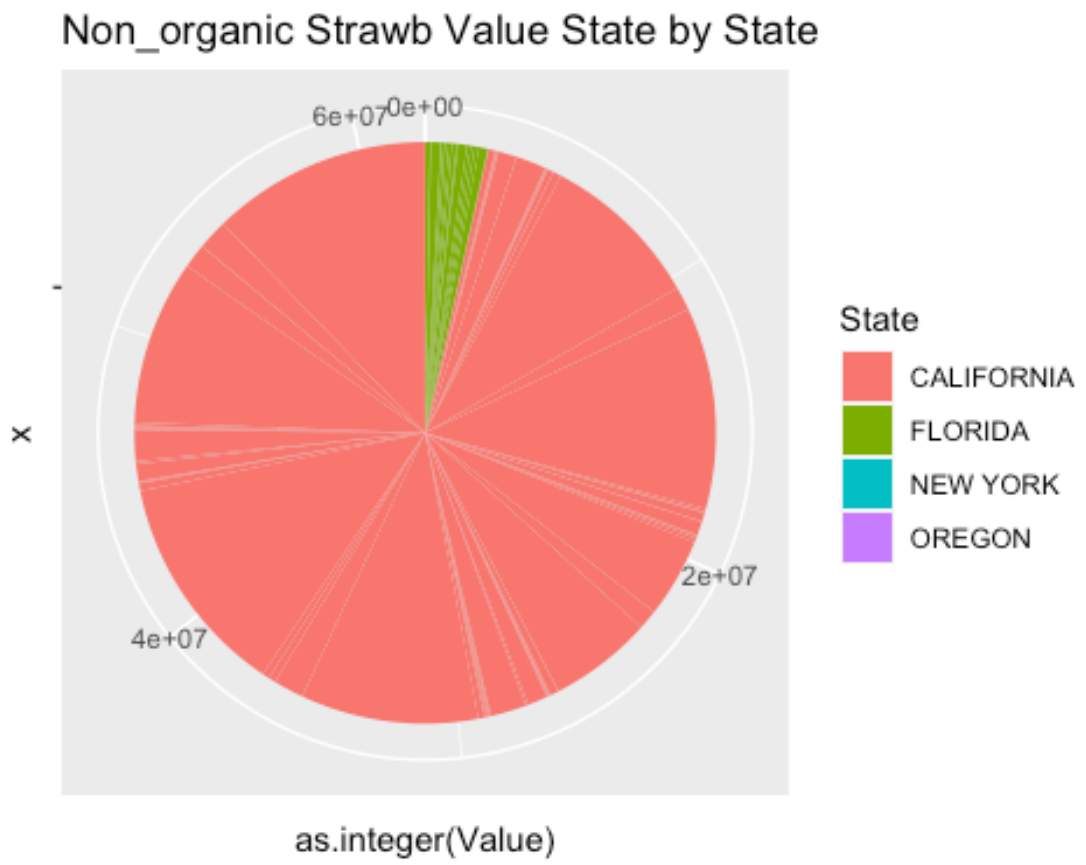
```
poisons_chem<-subset(strawb_chem,chem_name==" BIFENTHRIN "|chem_name==" 
 METHYL BROMIDE "|chem_name==" CHLOROPICRIN ")
safe_chem<-subset(strawb_chem,chem_name==" PHOSPHATE" |chem_name==" POT
ASSIUM BICARBON. " )

strawb_organic_eda<-strawb_organic
strawb_organic_eda$Value<-as.numeric(strawb_organic_eda$Value)

## Warning: NAs introduced by coercion

strawb_organic_eda<-na.omit(strawb_organic_eda)

strawb_non_organic_eda<-strawb_non_organic
strawb_non_organic$Value<-as.numeric(strawb_non_organic$Value)

## Warning: NAs introduced by coercion

strawb_non_organic_eda<-na.omit(strawb_non_organic_eda)

strawb_chem_eda<-strawb_chem
strawb_chem_eda$Value<-as.numeric(strawb_chem_eda$Value)

## Warning: NAs introduced by coercion

strawb_chem_eda<-na.omit(strawb_chem_eda)

poisons_chem_eda<-poisons_chem
poisons_chem_eda$Value<-as.numeric(poisons_chem_eda$Value)

## Warning: NAs introduced by coercion

poisons_chem_eda<-na.omit(poisons_chem_eda)

safe_chem_eda<-safe_chem%<>% select(Year, State,units, chem_types, chem
_name, Value)
safe_chem_eda$Value<-as.numeric(safe_chem_eda$Value)

## Warning: NAs introduced by coercion

safe_chem_eda<-na.omit(safe_chem_eda)
```

##EDA

```
p1<-ggplot(data = strawb_organic_eda, aes(x = "", y = Value, fill = Sta
te)) + geom_bar(stat = "identity") +
  labs(title = "Organic Strawb Value State by State")+
  coord_polar("y")
p1
```

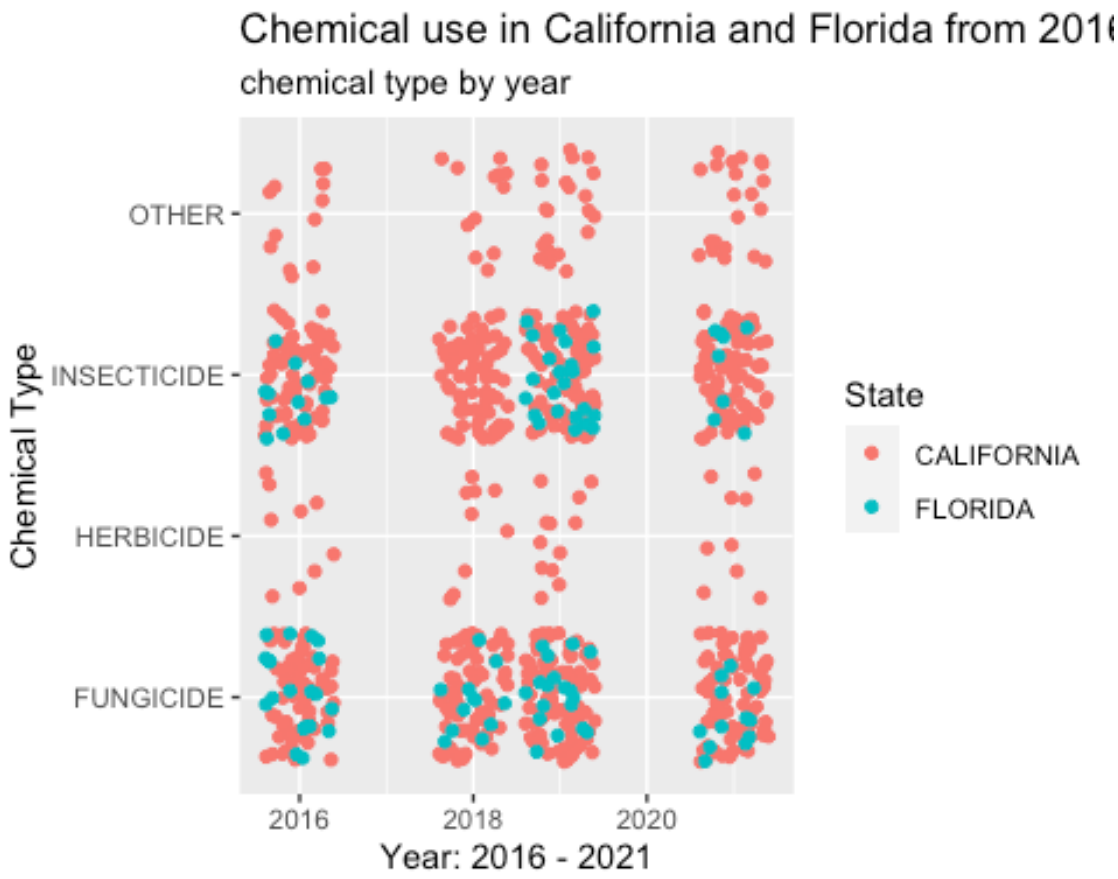



```
p2<-ggplot(data = strawb_non_organic_eda, aes(x = "", y = as.integer(Va
lue), fill = State)) + geom_bar(stat = "identity") +
  labs(title = "Non_organic Strawb Value State by State")+
  coord_polar("y")
p2

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning: Removed 1233 rows containing missing values (position_stac
k).
```



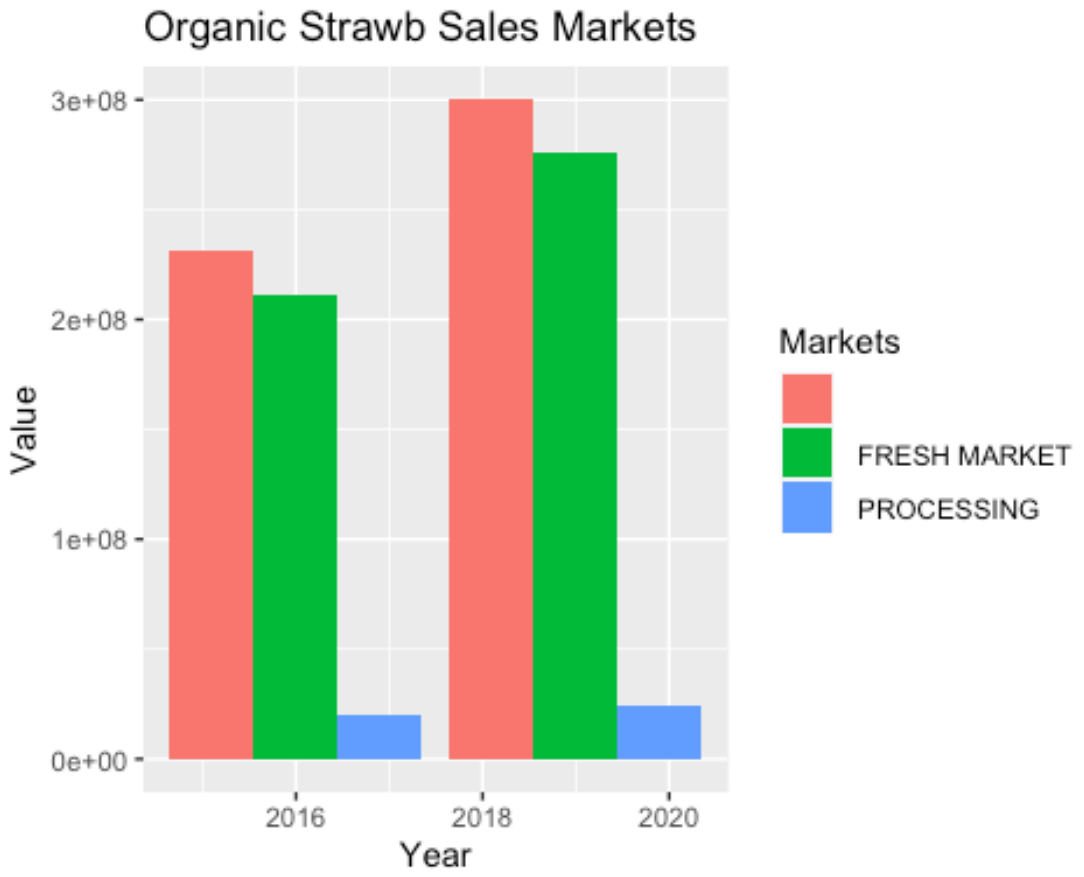

##For Plot 1 & 2 The first bar plot is organic strawberry value by each state, and the second bar plot is about the non-organic strawberry value by each state. From two plots, we can find that the value in California has the biggest proportion.

```
p3<-ggplot(data = strawb_chem_eda, aes(x = Year, y = chem_types, col =
State)) +  geom_jitter()+
  xlab('Year: 2016 - 2021') + ylab('Chemical Type') + labs(title = 'Che
mical use in California and Florida from 2016 to 2021', subtitle = 'che
mical type by year')
p3
```



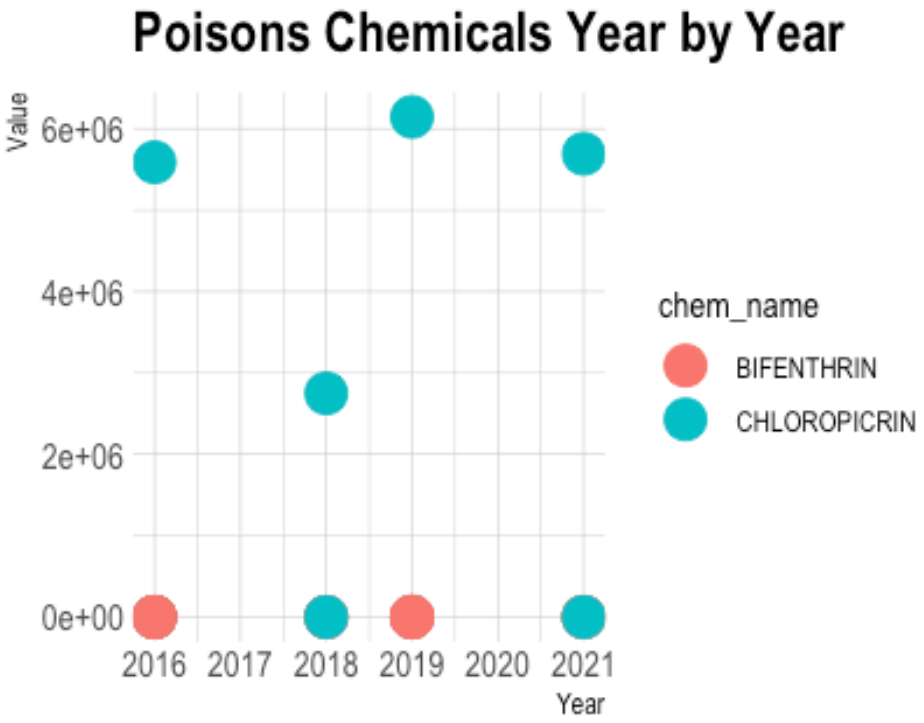

## From the plot, we can find that insecticide and fungicide Widely used in strawberry farming in California and Florida. In 2018, Florida did not use insecticide, and Florida did not use herbicide from 2016 to 2021.

```
p4<-ggplot(strawb_organic_eda, aes(fill=Markets, y=Value, x=Year)) +
    geom_bar(position="dodge", stat="identity")+labs(title = 'Organic S
trawb Sales Markets')
p4
```



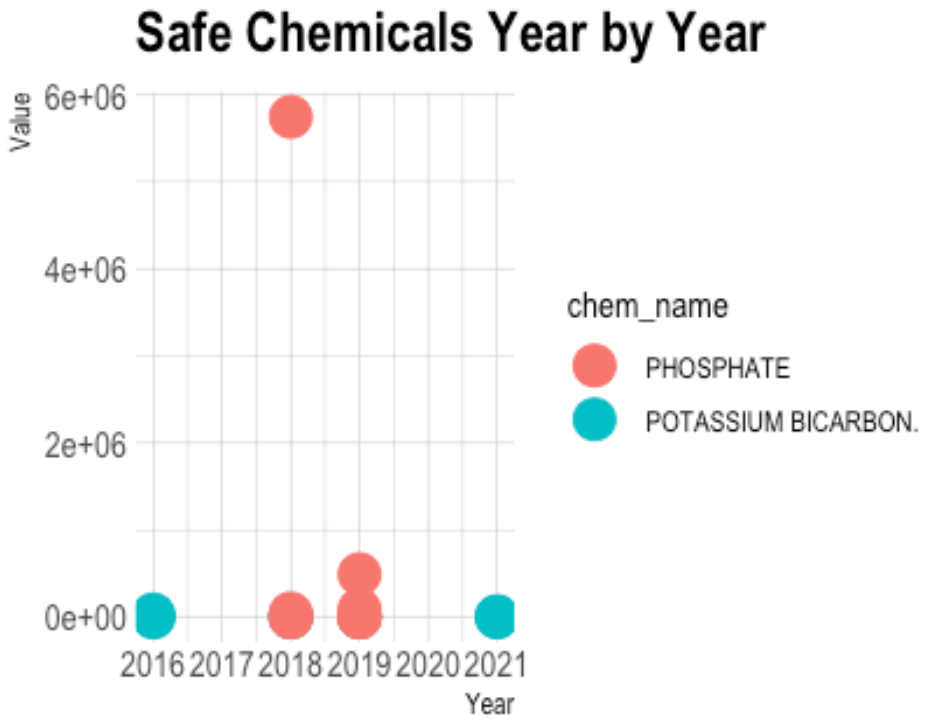

We can find that fresh market has clear increase and also has a larger percentage than procesing.

```
p5 <- ggplot(poisons_chem_eda, aes(x=Year, y=Value, color=chem_name)) +
geom_point(size=6) +theme_ipsum()+labs(title="Poisons Chemicals Year by
 Year")
p5
```



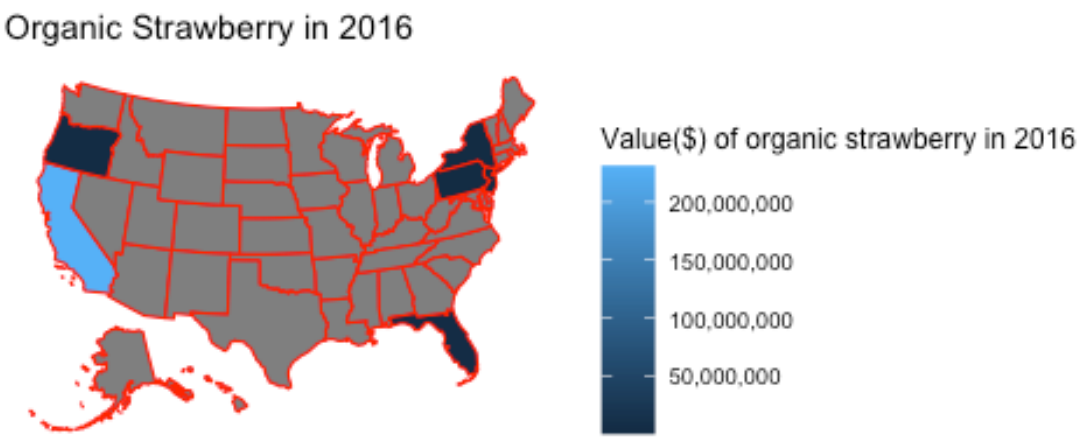

These two chemicals are toxic, and we can find that chloropicrin used in 2016, 2019, and 2021 can get huge value. Due to bifenthrin banned from use, so we can not see this chemical in 2020 and 2021.

```
p6 <- ggplot(safe_chem_eda, aes(x=Year, y=Value, color=chem_name)) +
    geom_point(size=6) +
    theme_ipsum()+labs(title="Safe Chemicals Year by Year")
p6
```

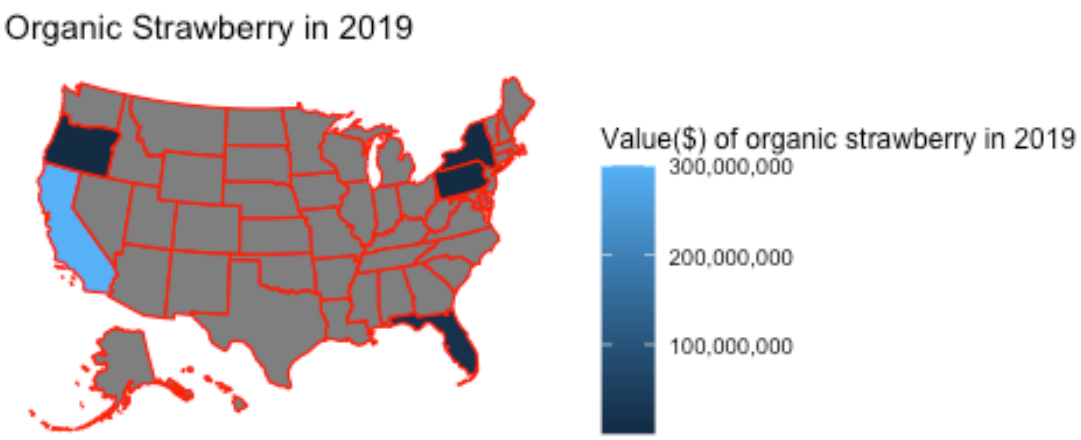



These two chemicals are safe, we can find that phosphate used in 2018 get huge value.

```
s13 <- filter(strawb, Domain == 'ORGANIC STATUS'&
                items==' MEASURED IN $'&
                Value!='(D)')

a2016 <- c(6,12,34,36,41,42)
b2016 <- c(231304956,2455805,38966,459144,1752592,87015)

a2019 <- c(6,12,36,41,42)
b2019 <- c(300277717,15055709,644155,1728809,89572)

library(usmap)
library(ggplot2)
df <- data.frame(
  fips = a2016,
  data = b2016
)
plot_usmap(data = df, values = "data",color="red")+
  scale_fill_continuous(name = "Value($) of organic strawberry in 2016",
 label = scales::comma)+
  theme(legend.position = "right") +labs(title="Organic Strawberry in 2
016")
```

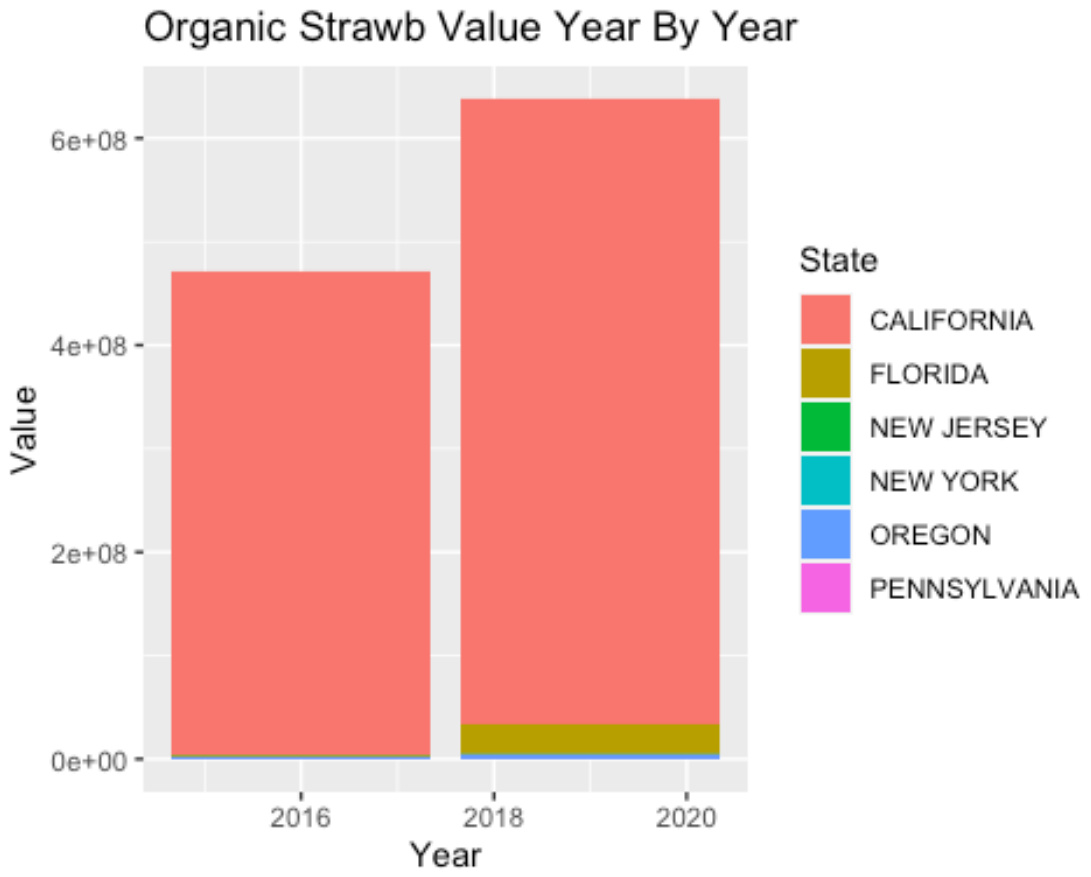



```
df <- data.frame(
  fips = a2019,
  data = b2019
)
plot_usmap(data = df, values = "data",color="red")+
  scale_fill_continuous(name = "Value($) of organic strawberry in 2019",
 label = scales::comma)+
  theme(legend.position = "right")+labs(title="Organic Strawberry in 20
19")
```



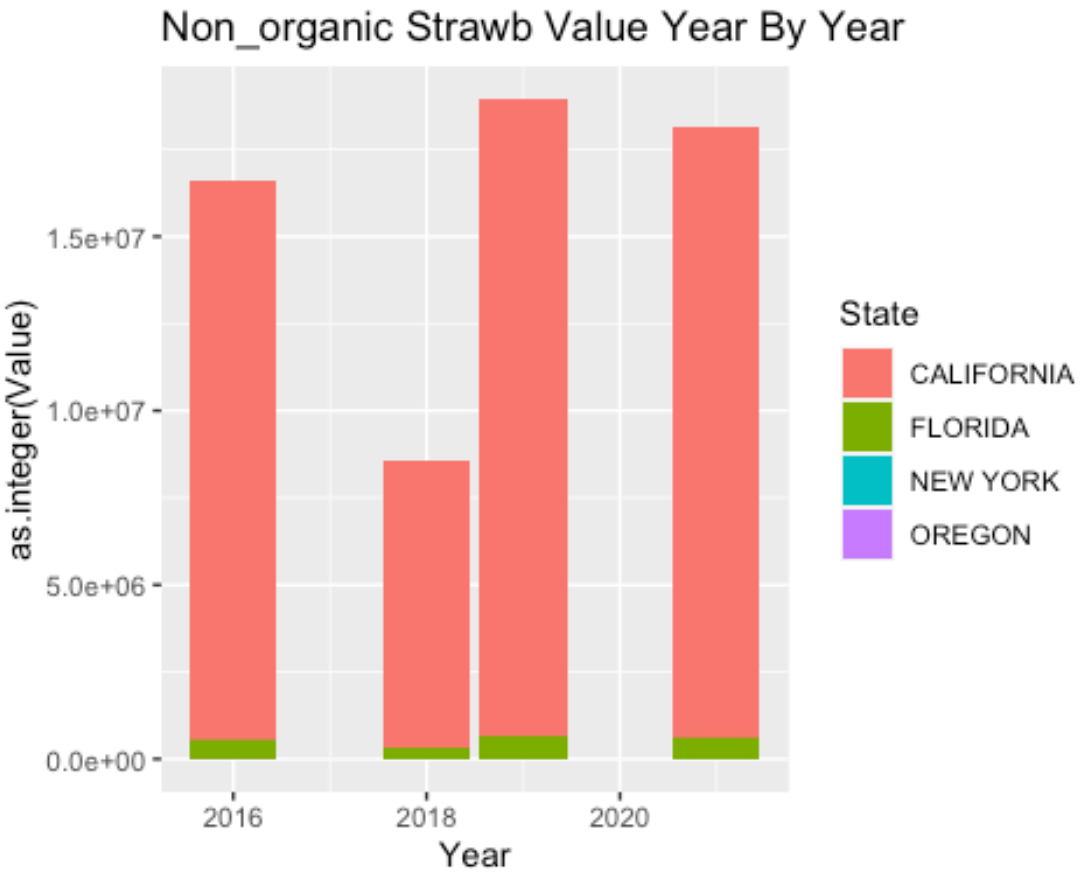

Compare 2016 and 2019, Sales of strawberries in California consistently lead.

```
p7 <- ggplot(data = strawb_organic_eda, mapping = aes(
  x = Year, y = Value, fill = State))
p7 + geom_col()+labs(title = "Organic Strawb Value Year By Year")
```




```
p8 <- ggplot(data = strawb_non_organic_eda, mapping = aes(
  x = Year, y = as.integer(Value), fill = State))+
  labs(title = "Non_organic Strawb Value Year By Year")
p8 + geom_col()
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning: Removed 1233 rows containing missing values (position_stac
k).
```




According to these two plots, the sales value of organic strawberry is much higher than non_organic strawberry in general.

`

```
```