415_Midterm_Proj

Before Starting the Project

Data Set

The csv file named "strawb_mar6.csv", which contains strawberry cultivation data in different states and measurements.

Main Focus

1. Pick three chemical treatments used for conventional strawberries in both states and contrast their use. Try to find chemicals with divergent use patterns between the states. Produce, tables, plots, and descriptions of the chemicals, how they are used, and how their use differs between California and Florida.
2. Compare the production and sales of organic and conventional strawberries and strawberries sold for processing. Show differences in price and volume between California and Florida. How do price, cost, and volume relationships change over the years?

Data Cleaning

Install packages that are needed

```
#install required packages
#install.packages("knitr")
#install.packages("kableExtra")
#install.packages("tidyverse")
#install.packages("stringr")
#install.packages("ggplot2")
#install.packages("scales")
```

```
library(knitr)
library(kableExtra)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter()     masks stats::filter()
x dplyr::group_rows() masks kableExtra::group_rows()
x dplyr::lag()        masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(stringr)
library(ggplot2)
```

Run the data cleaning method as same as `USDA-NASS starwberries.qmd`

```
# Label for the code chunk
#| label: read data - glimpse

# Read the CSV file "strawb_mar6.csv" into a data frame called 'strawberry'
# - 'col_names = TRUE' tells read_csv to treat the first row as column names
# - 'show_col_types = FALSE' suppresses printing of column type information
strawberry <- read_csv("strawb_mar6.csv",
                       col_names = TRUE,
                       show_col_types = FALSE)

# Source the file "my_functions.R" to load custom functions defined in that script
source("my_functions.R")
```

```
# remove columns from a data frame that contain only a single unique value
strawb <- strawberry |> drop_one_value_col()
```

Exploring Data

```r
# assume data is a tibble
# n_show is the number of rows to show

show_unique <- function(data, nrows=10 ){
  # make a tibble items to hold the data to show
  # browser()
  a <- nrows * dim(data)[2]  # number of cells in items
  items <- rep(" ", a) # items will coerce everything to char
  dim(items) <- c(nrows ,dim(data)[2]) # shape items
  items <- as_tibble(items)
  colnames(items) <- colnames(data)
  # browser()
  for(i in 1:dim(data)[2]){

    col_items <- unique(data[,i])
    # row_ex is the number of rows needed
    # to make the column length conformable with items
    row_ex <- nrows - dim(col_items)[1]
    if(row_ex >= 0){
      ex_rows <- tibble(rep(" ",row_ex))
      colnames(ex_rows) <- colnames(col_items)
      col_add <- rbind2(col_items, ex_rows)

    } else if(row_ex < 0){
      col_add <- col_items[1:10,]

    }

    items[,i] <- col_add

  }

  return(items)
}

#test <- show_unique(strawb, 10)
```

```r
#|label: split strawb into census and survey pieces

strw_census <- strawb |> filter(Program == "CENSUS")

strw_survey <- strawb |> filter(Program == "SURVEY")
```

```
nrow(strawb) == (nrow(strw_census) + nrow(strw_survey))
```

```
[1] TRUE
```

```
# Remove columns that contain only a single unique value
s_census <- strw_census |> drop_one_value_col(prt_val = TRUE)
```

```
[1] "Looking for single value columns in data frame:  strw_census"
[1] "Columns dropped:"
    Program      Period Week Ending
   "CENSUS"      "YEAR"          NA
```

```
s_survey <- strw_survey |> drop_one_value_col(prt_val = TRUE)
```

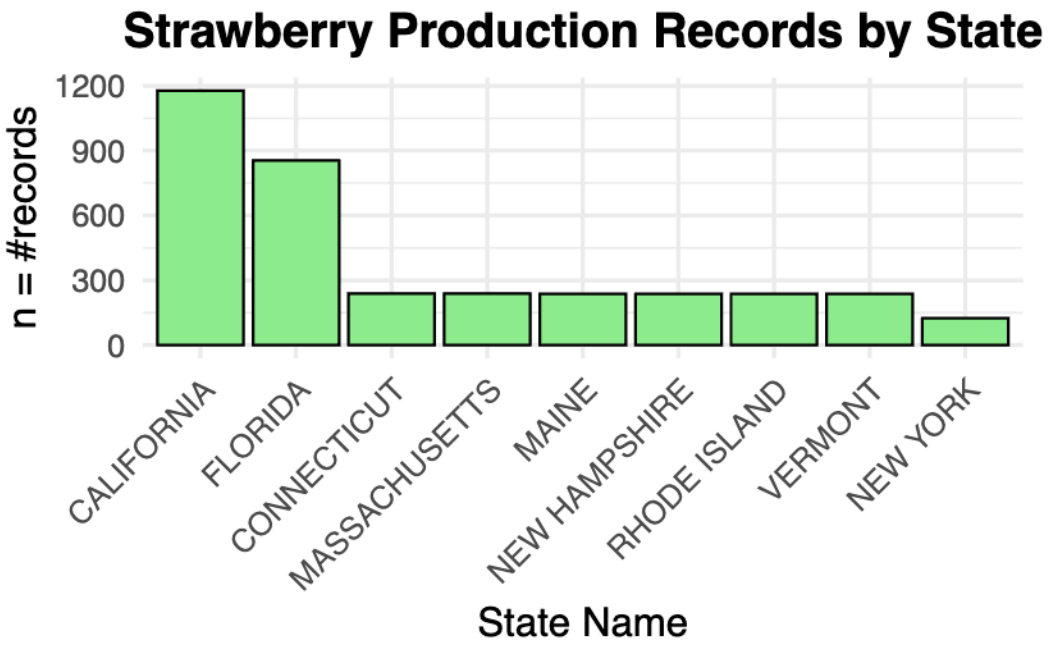
```
[1] "Looking for single value columns in data frame:  strw_survey"
[1] "Columns dropped:"
       Program      Commodity         CV (%)
      "SURVEY" "STRAWBERRIES"             NA
```

```
# Preview up to 10 unique values per column in each data set
unique_cen <- s_census |> show_unique(nrows = 10)
unique_sur <- s_survey |> show_unique(nrows = 10)

# Drop redundant or unnecessary Data
strw_census <- s_census |> select(-`State ANSI`)

# Remove 'State ANSI', 'Week Ending', and 'Period' for Data Cleaning
strw_survey <- s_survey |> select(-`State ANSI`, -`Week Ending`, -Period)

# Removing intermediate and temporary data set
rm(s_census, s_survey, strawberry, items)
```

Strawberry Growth Location

```
# Starberries' grown place
state_all <- strawb |> distinct(State)
state_all1 <- strawb |> group_by(State) |> count()
```

```r
# Improved bar plot (used chatgpt for better visualization of the bar chart)
ggplot(state_all1, aes(x = reorder(State, -n), y = n)) +
  geom_bar(stat = "identity", fill = "lightgreen", color = "black") +
  labs(
    title = "Strawberry Production Records by State",
    x = "State Name",
    y = "n = #records"
  ) +
  theme_minimal(base_size = 15) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(face = "bold", hjust = 0.5)
  )
```




```r
# Show unique values in the 'Domain' column from both datasets
unique(strw_census$Domain)
```

```
[1] "NET GAIN"             "TOTAL"                "NET LOSS"
[4] "AREA OPERATED"        "ECONOMIC CLASS"       "FARM SALES"
[7] "NAICS CLASSIFICATION" "ORGANIC STATUS"
```

```r
unique(strw_survey$Domain)
```

```
[1] "TOTAL"                 "CHEMICAL, FUNGICIDE"   "CHEMICAL, INSECTICIDE"
[4] "CHEMICAL, OTHER"       "CHEMICAL, HERBICIDE"   "FERTILIZER"
```

Part 1: Chemical Analysis

Gathering Chemical Information Data

```r
#Select all the rows that contains chemical use information
chemical <- strw_survey[str_detect(strw_survey$`Domain Category`,
                                   regex("chemical", ignore_case = TRUE)),]

#Taking only CA and FL state data of chemical usage
chemical_CA <- chemical[chemical$State == "CALIFORNIA", ]
chemical_FL <- chemical[chemical$State == "FLORIDA", ]
print(chemical_CA)
```

```
# A tibble: 1,011 x 9
    Year State      Fruit   Category Item  Metric Domain `Domain Category` Value
   <dbl> <chr>      <chr>   <chr>    <chr> <chr>  <chr>  <chr>              <chr>
 1  2023 CALIFORNIA STRAWB~ " MEASU~  <NA> <NA>   CHEMI~ CHEMICAL, FUNGIC~ (D)
 2  2023 CALIFORNIA STRAWB~ " MEASU~  <NA> <NA>   CHEMI~ CHEMICAL, INSECT~ (D)
 3  2023 CALIFORNIA STRAWB~ " MEASU~  <NA> <NA>   CHEMI~ CHEMICAL, INSECT~ (D)
 4  2023 CALIFORNIA STRAWB~ " MEASU~  <NA> <NA>   CHEMI~ CHEMICAL, OTHER:~ (NA)
 5  2023 CALIFORNIA STRAWB~ " MEASU~ " AV~ <NA>   CHEMI~ CHEMICAL, FUNGIC~ (D)
 6  2023 CALIFORNIA STRAWB~ " MEASU~ " AV~ <NA>   CHEMI~ CHEMICAL, INSECT~ (D)
 7  2023 CALIFORNIA STRAWB~ " MEASU~ " AV~ <NA>   CHEMI~ CHEMICAL, INSECT~ (D)
 8  2023 CALIFORNIA STRAWB~ " MEASU~ " AV~ <NA>   CHEMI~ CHEMICAL, OTHER:~ (NA)
 9  2023 CALIFORNIA STRAWB~ " MEASU~ " AV~ <NA>   CHEMI~ CHEMICAL, FUNGIC~ (D)
10  2023 CALIFORNIA STRAWB~ " MEASU~ " AV~ <NA>   CHEMI~ CHEMICAL, INSECT~ (D)
# i 1,001 more rows
```

```r
print(chemical_FL)
```

```
# A tibble: 691 x 9
    Year State   Fruit      Category Item  Metric Domain `Domain Category` Value
   <dbl> <chr>   <chr>      <chr>    <chr> <chr>  <chr>  <chr>              <chr>
```

```
 1  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, FUNGIC~ (D)
 2  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, FUNGIC~ (D)
 3  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, FUNGIC~ (D)
 4  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, HERBIC~ (D)
 5  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, HERBIC~ (D)
 6  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, HERBIC~ (D)
 7  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, INSECT~ (D)
 8  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, INSECT~ (D)
 9  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, INSECT~ (D)
10  2023 FLORIDA STRAWBERR~ " MEASU~ <NA>  <NA>   CHEMI~ CHEMICAL, OTHER:~ (D)
# i 681 more rows
```

```r
#clean the intermediate data set
rm(chemical)
```

In the first question, the three chemical picked from all the possible chemical is captan, thiram, novaluron

Chemical 1: Captan

```r
# First we check the use of Captan
# Use library of stringr functions to select the captan information
# Filter rows from chemical_CA (code from R studio stringr cheat sheet)
captan_CA <- chemical_CA[str_detect(chemical_CA$`Domain Category`,
                                    regex("CAPTAN", ignore_case = TRUE)), ]

# Filter rows from chemical_FL
captan_FL <- chemical_FL[str_detect(chemical_FL$`Domain Category`,
                                    regex("CAPTAN", ignore_case = TRUE)), ]

captan_combined <- rbind(captan_CA, captan_FL)

#clean intermediate steps
rm(captan_CA)
rm(captan_FL)

print(captan_combined)
```

```
# A tibble: 20 x 9
    Year State      Fruit   Category Item  Metric Domain `Domain Category` Value
```

```
   <dbl> <chr>      <chr>   <chr>    <chr> <chr>  <chr>  <chr>             <chr>
 1  2023 CALIFORNIA STRAWB~ " BEARI~ " ME~  <NA>  CHEMI~ CHEMICAL, FUNGIC~ 603,~
 2  2023 CALIFORNIA STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 1.693
 3  2023 CALIFORNIA STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 15.9~
 4  2023 CALIFORNIA STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 9.4
 5  2023 CALIFORNIA STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 88
 6  2021 CALIFORNIA STRAWB~ " BEARI~ " ME~  <NA>  CHEMI~ CHEMICAL, FUNGIC~ 253,~
 7  2021 CALIFORNIA STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 1.662
 8  2021 CALIFORNIA STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 14.2~
 9  2021 CALIFORNIA STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 8.6
10  2021 CALIFORNIA STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 45
11  2023 FLORIDA    STRAWB~ " BEARI~ " ME~  <NA>  CHEMI~ CHEMICAL, FUNGIC~ 144,~
12  2023 FLORIDA    STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 2.012
13  2023 FLORIDA    STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 10.5~
14  2023 FLORIDA    STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 5.2
15  2023 FLORIDA    STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 96
16  2021 FLORIDA    STRAWB~ " BEARI~ " ME~  <NA>  CHEMI~ CHEMICAL, FUNGIC~ 135,~
17  2021 FLORIDA    STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 2.025
18  2021 FLORIDA    STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 21.3~
19  2021 FLORIDA    STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 10.5
20  2021 FLORIDA    STRAWB~ " BEARI~ " ME~ " AVG" CHEMI~ CHEMICAL, FUNGIC~ 61
```

```r
# Now we get the table of captan use data in both California and Florida
```

```r
# now we start to reorganize the data in the order of category at first, Item for the second


captan <- captan_combined |> group_by(Category, Item, State, Year) |> count(Value)

print(captan)
```

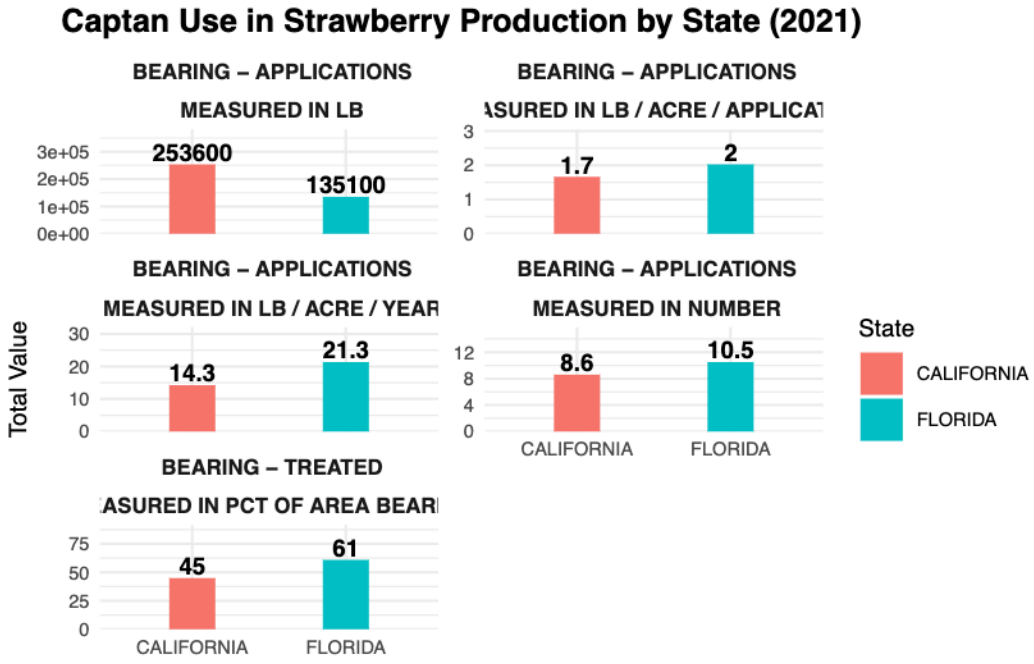
```
# A tibble: 20 x 6
# Groups:   Category, Item, State, Year [20]
   Category                  Item                          State  Year Value     n
   <chr>                     <chr>                         <chr> <dbl> <chr> <int>
 1 " BEARING - APPLICATIONS" " MEASURED IN LB"             CALI~  2021 253,~     1
 2 " BEARING - APPLICATIONS" " MEASURED IN LB"             CALI~  2023 603,~     1
 3 " BEARING - APPLICATIONS" " MEASURED IN LB"             FLOR~  2021 135,~     1
 4 " BEARING - APPLICATIONS" " MEASURED IN LB"             FLOR~  2023 144,~     1
 5 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2021 1.662     1
 6 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2023 1.693     1
 7 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2021 2.025     1
```

```
 8 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2023 2.012     1
 9 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2021 14.2~     1
10 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2023 15.9~     1
11 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2021 21.3~     1
12 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2023 10.5~     1
13 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"       CALI~  2021 8.6       1
14 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"       CALI~  2023 9.4       1
15 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"       FLOR~  2021 10.5      1
16 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"       FLOR~  2023 5.2       1
17 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ CALI~  2021 45        1
18 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ CALI~  2023 88        1
19 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ FLOR~  2021 61        1
20 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ FLOR~  2023 96        1
```

```r
# This table  directly shows the transition of captan use from year 2021 to 2023 in each sta
```

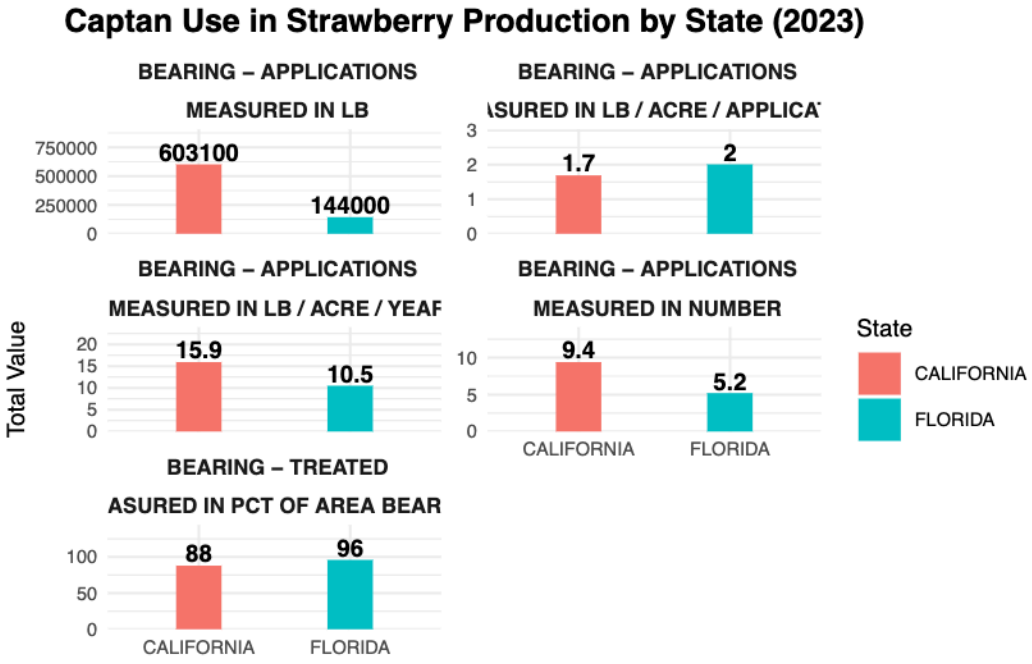
```r
#Then we can make a bar chart to compare the captan use in California and Florida

#Generate Data from 2021 and 2023(using stringr functions)
captan_2021 <- captan[str_detect(captan$Year,
                                 regex("2021", ignore_case = TRUE)), ]
captan_2023 <- captan[str_detect(captan$Year,
                                 regex("2023", ignore_case = TRUE)), ]

#Ensure the value under Value column is numeric
#so in graph it can provide the correct relationship bar
captan_2021$Value <- as.numeric(gsub(",", "", captan_2021$Value))
captan_2023$Value <- as.numeric(gsub(",", "", captan_2023$Value))

#Then we can make the graph (code perfection by chatgpt, for a better visual effect)
ggplot(captan_2021, aes(x = State, y = Value, fill = State)) +
  geom_bar(stat = "identity",
           position = position_dodge(width = 0.9),
           width = 0.3) +
  geom_text(aes(label = round(Value, 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.2, size = 3, fontface = "bold") +
  facet_wrap(~ Category + Item, scales = "free_y", nrow = 3) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.5))) +
  labs(title = "Captan Use in Strawberry Production by State (2021)",
       x = NULL,
       y = "Total Value") +
```

```
  theme_minimal(base_size = 9) +
  theme(
    strip.text = element_text(size = 8, face = "bold"),
    axis.text.x = element_text(),
    axis.text.y = element_text(),
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    panel.spacing = unit(0.4, "lines"),
    legend.position = "right"
  )
```




```
ggplot(captan_2023, aes(x = State, y = Value, fill = State)) +
  geom_bar(stat = "identity",
           position = position_dodge(width = 0.9),
           width = 0.3) +
  geom_text(aes(label = round(Value, 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.2, size = 3, fontface = "bold") +
  facet_wrap(~ Category + Item, scales = "free_y", nrow = 3) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.5))) +
  labs(title = "Captan Use in Strawberry Production by State (2023)",
       x = NULL,
       y = "Total Value") +
```

```
  theme_minimal(base_size = 9) +
  theme(
    strip.text = element_text(size = 8, face = "bold"),
    axis.text.x = element_text(""),
    axis.text.y = element_text(""),
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    panel.spacing = unit(0.4, "lines"),
    legend.position = "right"
  )
```




```
#Remove intermediate Data set
rm(captan_2021,captan_2023)
```

From the graph of 2021, California uses the fungicide that contains captan (a certified carcinogenic chemical) more than Florida in general. However, when we look up into other measurement, the average detection of captan use in California is less than Florida. This is because California Strawberry Production Firm is larger than Florida in size.

In 2023, There are some changes in the data. The measure in lb/acre/year and measure in number of California firm is higher than Florida firm, meaning that California uses captan fungicide more often per year, and the average number that detected captan is higher than Florida.

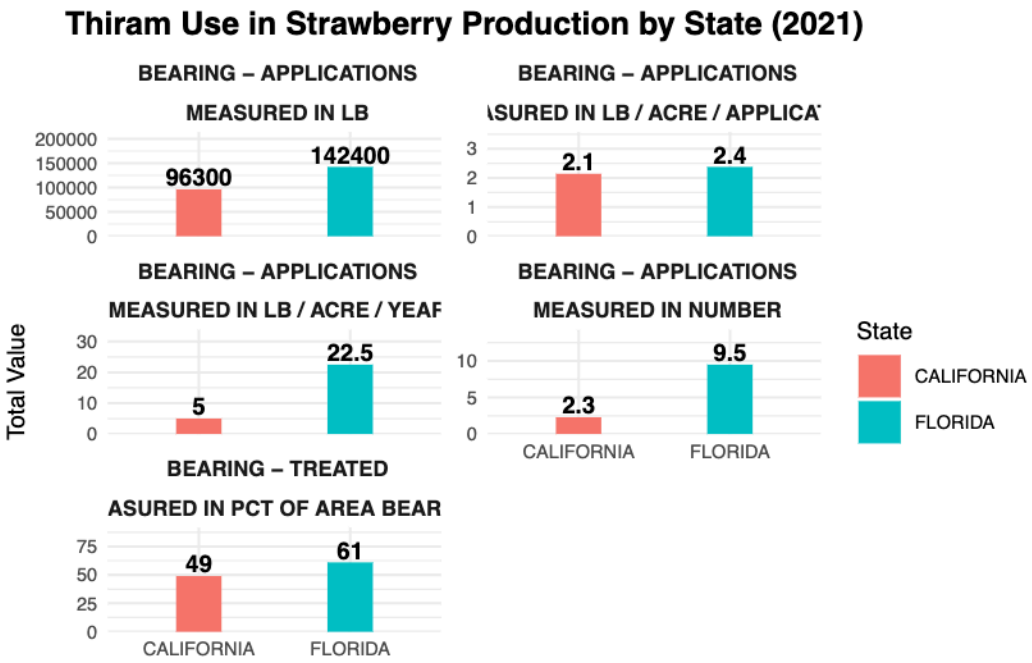
Chemical 2: Thiram

```
#Follows the same step as captan
#Filter rows form chemical_CA
thiram_CA <- chemical_CA[str_detect(chemical_CA$`Domain Category`,
                                    regex("THIRAM", ignore_case = TRUE)), ]

# Filter rows from chemical_FL
thiram_FL <- chemical_FL[str_detect(chemical_FL$`Domain Category`,
                                    regex("THIRAM", ignore_case = TRUE)), ]

#Get the thiram information data and gather them into one data set
thiram_combined <- rbind(thiram_CA, thiram_FL)

#remove intermediate data set
rm(thiram_CA)
rm(thiram_FL)

#Table of Comparison of different year by same state
thiram <- thiram_combined |> group_by(Category,
                                      Item, State, Year) |> count(Value)

print(thiram)
```

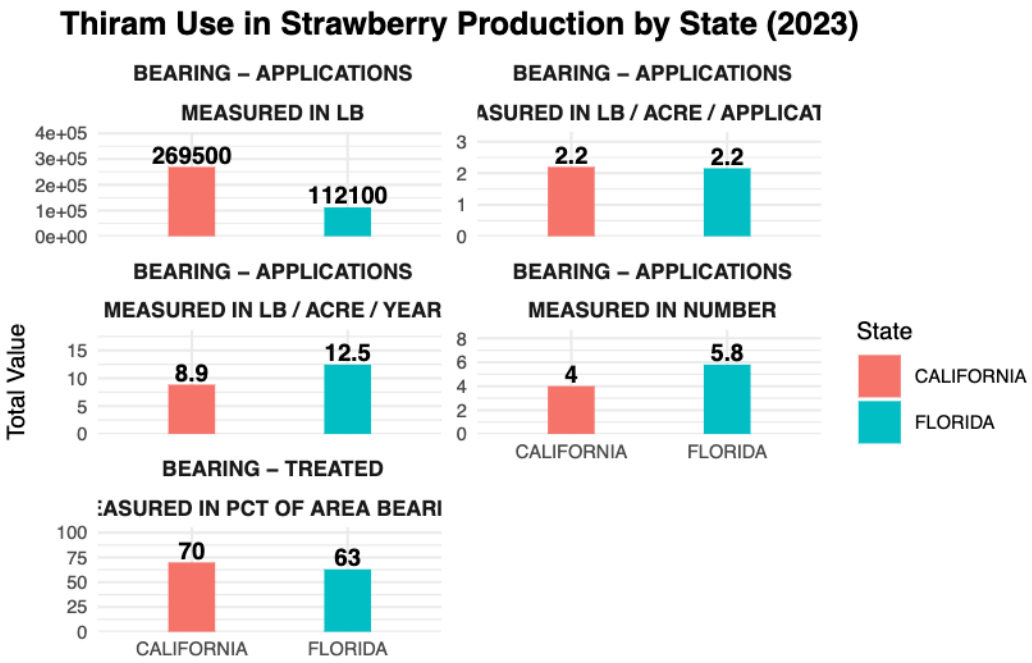
```
# A tibble: 20 x 6
# Groups:   Category, Item, State, Year [20]
   Category                  Item                          State  Year Value     n
   <chr>                     <chr>                         <chr> <dbl> <chr> <int>
 1 " BEARING - APPLICATIONS" " MEASURED IN LB"             CALI~  2021 96,3~     1
 2 " BEARING - APPLICATIONS" " MEASURED IN LB"             CALI~  2023 269,~     1
 3 " BEARING - APPLICATIONS" " MEASURED IN LB"             FLOR~  2021 142,~     1
 4 " BEARING - APPLICATIONS" " MEASURED IN LB"             FLOR~  2023 112,~     1
 5 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2021 2.144     1
 6 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2023 2.201     1
 7 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2021 2.38      1
 8 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2023 2.156     1
 9 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2021 5.029     1
10 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2023 8.873     1
11 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2021 22.5~     1
12 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2023 12.4~     1
13 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"        CALI~  2021 2.3       1
```

```
14 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"       CALI~  2023 4         1
15 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"       FLOR~  2021 9.5       1
16 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"       FLOR~  2023 5.8       1
17 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ CALI~  2021 49        1
18 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ CALI~  2023 70        1
19 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ FLOR~  2021 61        1
20 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ FLOR~  2023 63        1
```

```r
#Rearrange the data for bar chart
thiram_2021 <- thiram[str_detect(thiram$Year,
                                 regex("2021", ignore_case = TRUE)), ]
thiram_2023 <- thiram[str_detect(thiram$Year,
                                 regex("2023", ignore_case = TRUE)), ]

#Ensure the value under Value column is numeric, so in graph it can provide the right relati
thiram_2021$Value <- as.numeric(gsub(",", "", thiram_2021$Value))
thiram_2023$Value <- as.numeric(gsub(",", "", thiram_2023$Value))
#Same year, same domain category and item, but different state record comparison (code perfe

ggplot(thiram_2021, aes(x = State, y = Value, fill = State)) +
  geom_bar(stat = "identity",
           position = position_dodge(width = 0.9),
           width = 0.3) +
  geom_text(aes(label = round(Value, 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.2, size = 3, fontface = "bold") +
  facet_wrap(~ Category + Item, scales = "free_y", nrow = 3) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.5))) +
  labs(title = "Thiram Use in Strawberry Production by State (2021)",
       x = NULL,
       y = "Total Value") +
  theme_minimal(base_size = 9) +
  theme(
    strip.text = element_text(size = 8, face = "bold"),
    axis.text.x = element_text(),
    axis.text.y = element_text(),
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    panel.spacing = unit(0.4, "lines"),
    legend.position = "right"
  )
```




```
ggplot(thiram_2023, aes(x = State, y = Value, fill = State)) +
  geom_bar(stat = "identity",
           position = position_dodge(width = 0.9),
           width = 0.3) +
  geom_text(aes(label = round(Value, 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.2, size = 3, fontface = "bold") +
  facet_wrap(~ Category + Item, scales = "free_y", nrow = 3) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.5))) +
  labs(title = "Thiram Use in Strawberry Production by State (2023)",
       x = NULL,
       y = "Total Value") +
  theme_minimal(base_size = 9) +
  theme(
    strip.text = element_text(size = 8, face = "bold"),
    axis.text.x = element_text(),
    axis.text.y = element_text(),
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    panel.spacing = unit(0.4, "lines"),
    legend.position = "right"
  )
```




```
# Remove intermediate data set
rm(thiram_2021,thiram_2023)
```

From the Graph of 2021, we can see that Thiram use in Florida is higher than California in all the aspect, meaning that Florida firms uses chemical product that contains thiram higher in quantity and frequency.

From the Graph of 2023, California is using thiram included product more than Florida, leading to a higher bar compared to Florida in 4 measurement. However, in number Florida is still a little bit higher than California

Chemical 3: Novaluron

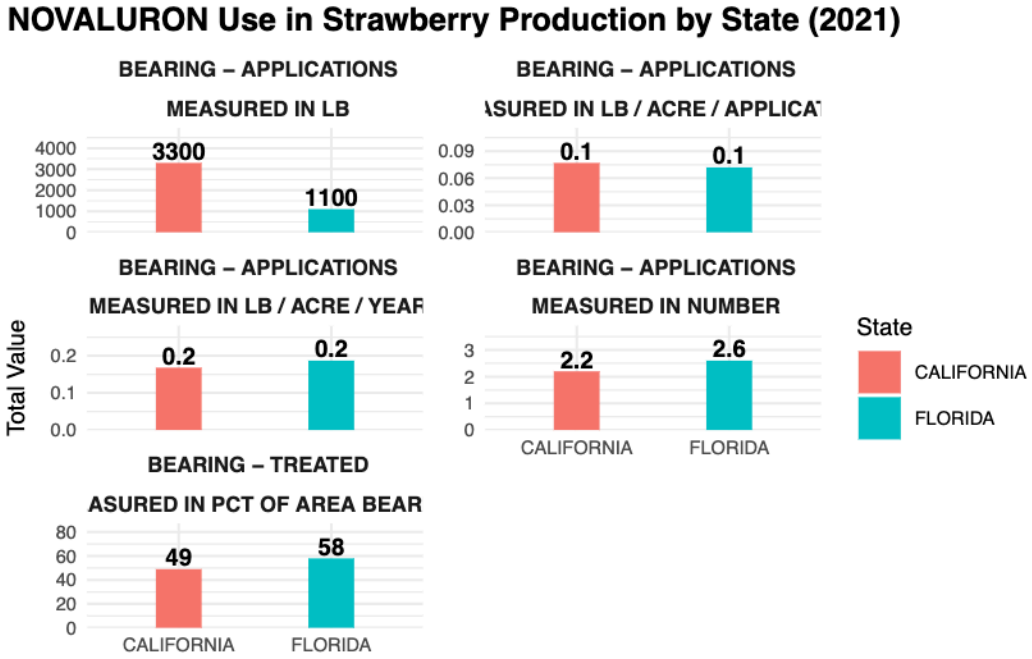
```
#Follows the same step as captan
#Filter rows form chemical_CA
nov_CA <- chemical_CA[str_detect(chemical_CA$`Domain Category`,
                                 regex("NOVALURON", ignore_case = TRUE)), ]

# Filter rows from chemical_FL
nov_FL <- chemical_FL[str_detect(chemical_FL$`Domain Category`,
                                 regex("NOVALURON", ignore_case = TRUE)), ]
```

```r
#Get the thiram information data and gather them into one data set
nov_combined <- rbind(nov_CA, nov_FL)

#remove intermediate data set
rm(nov_CA)
rm(nov_FL)

#Table of Comparison of different year by same state
nov <- nov_combined |> group_by(Category, Item, State, Year) |> count(Value)

print(nov)
```

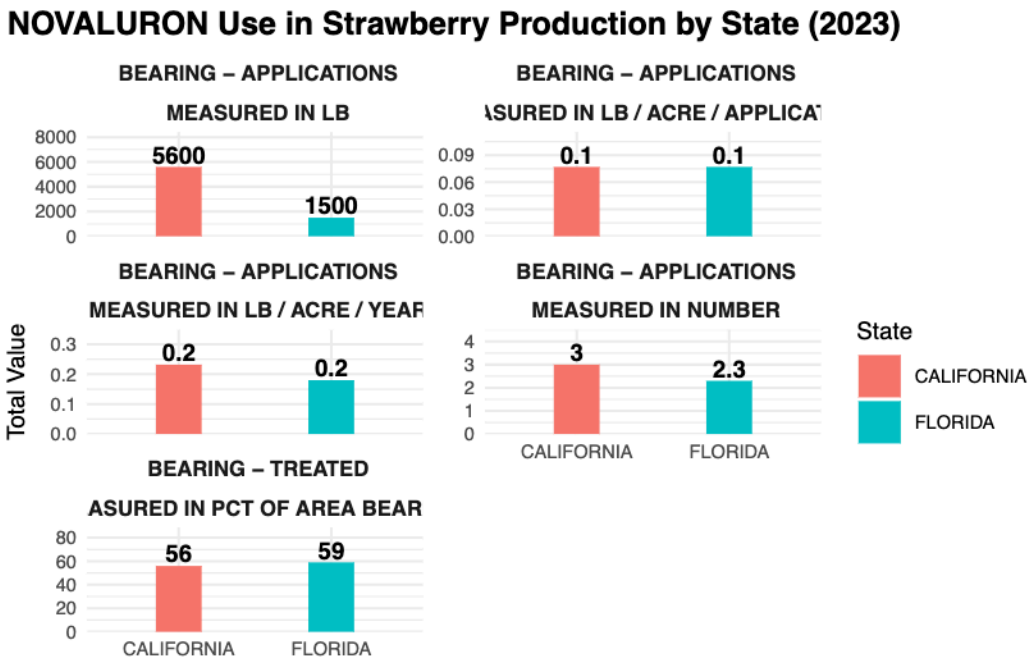
```
# A tibble: 20 x 6
# Groups:   Category, Item, State, Year [20]
   Category                  Item                          State  Year Value     n
   <chr>                     <chr>                         <chr> <dbl> <chr> <int>
 1 " BEARING - APPLICATIONS" " MEASURED IN LB"             CALI~  2021 3,300     1
 2 " BEARING - APPLICATIONS" " MEASURED IN LB"             CALI~  2023 5,600     1
 3 " BEARING - APPLICATIONS" " MEASURED IN LB"             FLOR~  2021 1,100     1
 4 " BEARING - APPLICATIONS" " MEASURED IN LB"             FLOR~  2023 1,500     1
 5 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2021 0.077     1
 6 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2023 0.077     1
 7 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2021 0.072     1
 8 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2023 0.077     1
 9 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2021 0.168     1
10 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ CALI~  2023 0.232     1
11 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2021 0.187     1
12 " BEARING - APPLICATIONS" " MEASURED IN LB / ACRE / ~ FLOR~  2023 0.18      1
13 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"        CALI~  2021 2.2       1
14 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"        CALI~  2023 3         1
15 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"        FLOR~  2021 2.6       1
16 " BEARING - APPLICATIONS" " MEASURED IN NUMBER"        FLOR~  2023 2.3       1
17 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ CALI~  2021 49        1
18 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ CALI~  2023 56        1
19 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ FLOR~  2021 58        1
20 " BEARING - TREATED"      " MEASURED IN PCT OF AREA ~ FLOR~  2023 59        1
```

```r
#Rearrange the data for bar chart
nov_2021 <- nov[str_detect(nov$Year, regex("2021", ignore_case = TRUE)), ]
nov_2023 <- nov[str_detect(nov$Year, regex("2023", ignore_case = TRUE)), ]
```

```r
#Ensure the value under Value column is numeric,
#so in graph it can provide the right relationship bar
#higher value has a higher bar
nov_2021$Value <- as.numeric(gsub(",", "", nov_2021$Value))
nov_2023$Value <- as.numeric(gsub(",", "", nov_2023$Value))

#Same year, same domain category and item, but different state
#record comparison (code perfection by chatgpt, for a better visualization)

ggplot(nov_2021, aes(x = State, y = Value, fill = State)) +
  geom_bar(stat = "identity",
           position = position_dodge(width = 0.9),
           width = 0.3) +
  geom_text(aes(label = round(Value, 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.2, size = 3, fontface = "bold") +
  facet_wrap(~ Category + Item, scales = "free_y", nrow = 3) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.5))) +
  labs(title = "NOVALURON Use in Strawberry Production by State (2021)",
       x = NULL,
       y = "Total Value") +
  theme_minimal(base_size = 9) +
  theme(
    strip.text = element_text(size = 8, face = "bold"),
    axis.text.x = element_text(),
    axis.text.y = element_text(),
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    panel.spacing = unit(0.4, "lines"),
    legend.position = "right"
  )
```




```
ggplot(nov_2023, aes(x = State, y = Value, fill = State)) +
  geom_bar(stat = "identity",
           position = position_dodge(width = 0.9),
           width = 0.3) +
  geom_text(aes(label = round(Value, 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.2, size = 3, fontface = "bold") +
  facet_wrap(~ Category + Item, scales = "free_y", nrow = 3) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.5))) +
  labs(title = "NOVALURON Use in Strawberry Production by State (2023)",
       x = NULL,
       y = "Total Value") +
  theme_minimal(base_size = 9) +
  theme(
    strip.text = element_text(size = 8, face = "bold"),
    axis.text.x = element_text(),
    axis.text.y = element_text(),
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    panel.spacing = unit(0.4, "lines"),
    legend.position = "right"
  )
```




```
# Remove intermediate data set
rm(nov_2021,nov_2023)
```

Novaluron is a kind of chemical that have a negative environmental impact.

From the graph of 2021, we can see California uses Novaluron treatment more than Florida in total. However, Florida uses novaluron more often per year and slightly higher in numbers and area bearing.

From the graph of 2023, we can see California still uses Novaluron more than Florida in total, and since the total quantity of california increases a lot. The other per capita measurement is higher than Florida, except measured in area bearing. It means more percentage of the Florida strawberry is treated with novaluron compare to California.

Part 2: Business Analysis

Data Cleaning and Reorganization

```
# recall that organic information is and sales stuff is all in census data
org_info <- strw_census[str_detect(strw_census$Category, regex("ORGANIC",
                                                               ignore_case = TRUE)), ]
```

```r
conv_info <- strw_census[!str_detect(strw_census$Category, regex("ORGANIC",
                                                                 ignore_case = TRUE)), ]
# filter the data by state
org_CA <- org_info[str_detect(org_info$State, regex("CALIFORNIA", ignore_case = TRUE)), ]
org_FL <- org_info[str_detect(org_info$State, regex("FLORIDA", ignore_case = TRUE)), ]
conv_CA <- conv_info[str_detect(conv_info$State, regex("CALIFORNIA", ignore_case = TRUE)), ]
conv_FL <- conv_info[str_detect(conv_info$State, regex("FLORIDA", ignore_case = TRUE)), ]

#remove the intermediate data set
rm(org_info)
rm(conv_info)

#after this, we need to select the data that we want to compare and put them in the same data

org_CA1 <- org_CA[str_detect(org_CA$Category, regex("SALES", ignore_case = TRUE)), ]
org_FL1 <- org_FL[str_detect(org_FL$Category, regex("SALES", ignore_case = TRUE)), ]
conv_CA1 <- conv_CA[str_detect(conv_CA$Domain, regex("TOTAL", ignore_case = TRUE)), ]
conv_FL1 <- conv_FL[str_detect(conv_FL$Domain, regex("TOTAL", ignore_case = TRUE)), ]

#We now have the organic and convential straberry data for California and Florida
org <- rbind(org_CA1,org_FL1)
conv <- rbind(conv_CA1,conv_FL1)

#remove the unecessary data set
rm(org_CA,org_CA1,org_FL,conv_CA,conv_CA1,conv_FL,conv_FL1)

#we can see the table
print(org)
```

```
# A tibble: 4 x 11
   Year State     Commodity Fruit Category Item  Metric Domain `Domain Category`
  <dbl> <chr>     <chr>     <chr> <chr>    <chr> <chr>  <chr>  <chr>
1  2021 CALIFORN~ STRAWBER~ STRA~ " ORGAN~ " ME~ <NA>   ORGAN~ ORGANIC STATUS: ~
2  2021 CALIFORN~ STRAWBER~ STRA~ " ORGAN~ " ME~ <NA>   ORGAN~ ORGANIC STATUS: ~
3  2021 FLORIDA   STRAWBER~ STRA~ " ORGAN~ " ME~ <NA>   ORGAN~ ORGANIC STATUS: ~
4  2021 FLORIDA   STRAWBER~ STRA~ " ORGAN~ " ME~ <NA>   ORGAN~ ORGANIC STATUS: ~
# i 2 more variables: Value <chr>, `CV (%)` <chr>
```

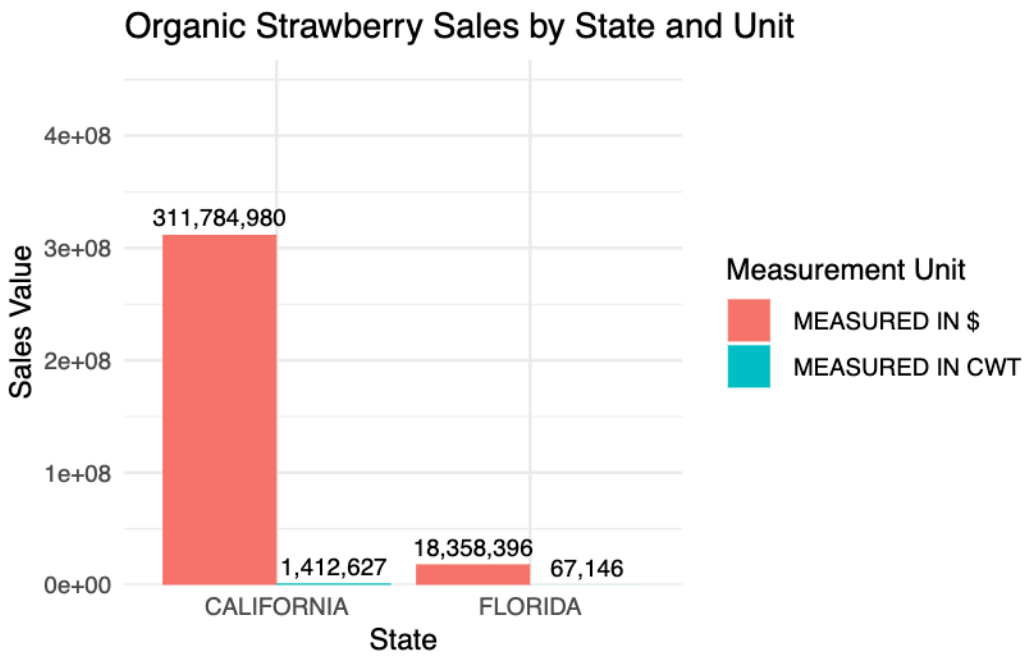
```r
print(conv)
```

```
# A tibble: 10 x 11
```

```
    Year State    Commodity Fruit Category Item  Metric Domain `Domain Category`
   <dbl> <chr>    <chr>     <chr> <chr>    <chr> <chr>  <chr>  <chr>
 1  2022 CALIFOR~ INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
 2  2022 CALIFOR~ INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
 3  2022 CALIFOR~ INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
 4  2022 CALIFOR~ INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
 5  2022 CALIFOR~ INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
 6  2022 FLORIDA  INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
 7  2022 FLORIDA  INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
 8  2022 FLORIDA  INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
 9  2022 FLORIDA  INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
10  2022 FLORIDA  INCOME, ~ INCO~ " NET C~ " OF~ " MEA~ TOTAL  NOT SPECIFIED
# i 2 more variables: Value <chr>, `CV (%)` <chr>
```

Organic Strawberry Data Visualization

```r
#make sure the value is numeric value
org$Value <- as.numeric(gsub(",", "", org$Value))

#then we can visualized the data in scatter plot for comparison

#Bar chart: Value by State, Measurement Unit
ggplot(org, aes(x = State, y = Value, fill = Item)) +
  geom_col(position = position_dodge(width = 0.9)) +
  geom_text(aes(label = scales::comma(Value)),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.5))) +
  labs(title = "Organic Strawberry Sales by State and Unit",
       x = "State",
       y = "Sales Value",
       fill = "Measurement Unit") +
  theme_minimal()
```



This bar chart shows organic strawberry sales in California and Florida, measured in two units: dollars (MEASURED IN $) and hundredweight (CWT).

California shows significantly higher sales in both units, with over $311 million compared to Florida's $18 million. The difference in scale between dollar and CWT values highlights the economic weight of California's organic strawberry market.

Conventional Strawberry Data Visualization

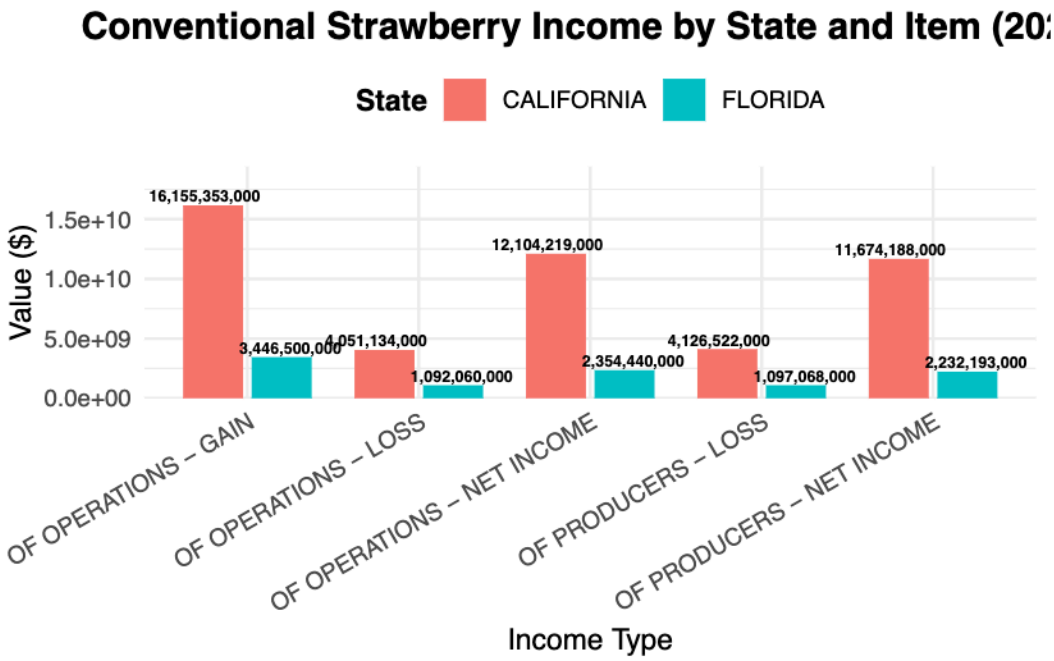
```
#make sure the number in Value column is numeric value for graphing
conv$Value <- as.numeric(gsub(",", "", conv$Value))

#now we can make a bar chart to visualize the date for conventional strawberries


library(ggplot2)
library(scales)
```

```
Attaching package: 'scales'
```

```
The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```r
ggplot(conv, aes(x = Item, y = Value, fill = State)) +
  geom_col(position = position_dodge(width = 0.8), width = 0.7) +
  geom_text(aes(label = comma(Value)),
            position = position_dodge(width = 1),
            vjust = -0.3, size = 2, fontface = "bold") +
  labs(
    title = "Conventional Strawberry Income by State and Item (2022)",
    x = "Income Type",
    y = "Value ($)",
    fill = "State"
  ) +
  scale_y_continuous(labels = label_number(scale_cut = cut_short_scale())) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.2))) +
  theme_minimal(base_size = 11) +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    axis.text.x = element_text(angle = 30, hjust = 1),
    legend.position = "top",
    legend.title = element_text(face = "bold")
  )
```

```
Scale for y is already present.
Adding another scale for y, which will replace the existing scale.
```



This chart compares conventional strawberry farm income in California and Florida across five income categories in 2022.

California shows a clear lead over Florida across all income categories related to conventional strawberry farming. The most notable difference is in "Operations – Gain," where California generated approximately $16.2 billion, compared to Florida's $3.4 billion.

Additionally, California reports significantly higher net income for both operations and producers. These figures highlight California's dominant position in the conventional strawberry industry, both in terms of production scale and overall profitability.

General Conclusion

From the data cleaning and analyze above, many potential problems is reflecting during the whole process.

In the aspect of human health, the production of strawberry still contains risks of extremely toxic chemicals. In the analysis is Captan and Thiram. Using chemical treatment in fruit production industry is inevitable. But the bar chart above, analyzing chemical detection, could be a preliminary guideline of picking strawberries by qualities, for both companies or individuals who wants to consume strawberries.

In the aspect of Environment Impact, the bar chart comparison of Novaluron shows that strawberry production is still inevitably negatively affect the environment by high dose using of this chemical. Thus, lowering the usage of chemical or developing new product that are more environmental friendly is needed.

In the business analysis section, we can see California dominates the strawberry production firm. in both organic and conventional. Calling back to what we saw from chemical data, California strawberry firm now should consider a more environmental and human health friendly production process, and become a leading role of such development.

```
#here is the code to clean all the previos analysis and rerun the program, you can use it by
#rm(list = ls())
```