Strawberries EDA

Jin Wen Lin

Objective

The main goal is to explore the Strawberries data from the USDA-NASS system through data cleaning, organization, as well as finding patterns for the variables in order to get a deeper understanding of the data. We are mainly focusing on the chemical part of the Strawberries data.

Overview of Dataset

First, let us start with looking an overview of the chemical data.

```r
# load data (survey data filtered from part 1)
straw_sur <- read.csv("strawberry_survey.csv")
colnames(straw_sur)[colnames(straw_sur) == "Chemical.Name"] <- "Type"
colnames(straw_sur)[colnames(straw_sur) == "Chemical.Type"] <- "Name"
glimpse(straw_sur)
```

```
Rows: 3,965
Columns: 17
$ Program           <chr> "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SU~
$ Year              <int> 2024, 2024, 2023, 2023, 2023, 2023, 2023, 2023, 2023,~
$ Period            <chr> "YEAR", "YEAR", "MARKETING YEAR", "MARKETING YEAR", "~
$ Geo.Level         <chr> "NATIONAL", "NATIONAL", "NATIONAL", "NATIONAL", "NATI~
$ State             <chr> "US TOTAL", "US TOTAL", "US TOTAL", "US TOTAL", "US T~
$ State.ANSI        <int> NA, NA, NA, NA, NA, 6, 6, 6, 12, 12, 12, NA, NA, NA, ~
$ Ag.District       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ Ag.District.Code  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ County            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ County.ANSI       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ Data.Item        <chr> "STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, ADJUSTE~
$ Domain           <chr> "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL",~
$ Type             <chr> "NOT SPECIFIED", "NOT SPECIFIED", "NOT SPECIFIED", "N~
$ Name             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ Chemical.Code    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
$ Value            <dbl> 1.090000e+01, 4.040000e+00, 1.230000e+02, 1.420000e+0~
$ CV....           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

To make things in Data Item column more clear, I am going to split it. Here are the computations.

```
# Use code from lecture
straw_sur1 <- straw_sur %>% separate_wider_delim(cols = `Data.Item`,
                                                delim = ", ",
                                                names = c("straw",
                                                         "mkt",
                                                         "measure",
                                                         "other"
                                                         ),
                                                too_many = "merge",
                                                too_few = "align_start")


straw_sur2 <- straw_sur1 %>%  separate_wider_delim(cols = "straw",
                                                 delim = " - ",
                                                 names = c("straw",
                                                           "more"),
                                                 too_many = "merge",
                                                 too_few = "align_start"
                                                 )

shift_loc <- function(df, col_name, dat_name, num_col, num_shift){
 # browser()
  col_num = which(colnames(df) == col_name)
  row_num = which(df[,col_num] == dat_name)  ## calcs a vector of rows

  for(k in 1:length(row_num)){
  d = rep(0,num_col) ## storage for items to be moved
  for(i in 1:num_col){
    d[i] = df[row_num[k], col_num + i - 1]
  }
  for(i in 1:num_col){
```

```r
    ra = row_num[k]
    cb = col_num + i - 1
    df[ra, cb] <-  NA
  }
  for(j in 1:num_col){
    rc = row_num[k]
    cd = col_num + j - 1 + num_shift
    df[rc, cd] = d[j]
  }
  }
 # sprintf("Rows adjusted:")
  # print("%d",row_num)
  return(df)
}

straw_sur2 %<>% shift_loc("more", "PRICE RECEIVED", 2, 1 )

straw_sur2 %<>% shift_loc("more", "ACRES HARVESTED", 1, 1 )

straw_sur2 %<>% shift_loc("more", "ACRES PLANTED", 1, 1 )

straw_sur2 %<>% shift_loc("more", "PRODUCTION", 2, 1 )

straw_sur2 %<>% shift_loc("more", "YIELD", 2, 1 )

straw_sur2 %<>% shift_loc("more", "APPLICATIONS", 3, 1 )

straw_sur2 %<>% shift_loc("more", "TREATED", 3, 1 )

rm(straw_sur, straw_sur1)

straw_sur2 <- straw_sur2 %>% select(-more, -State.ANSI, -Ag.District,
                                    -Ag.District.Code, -County, -County.ANSI)

straw_sur <- straw_sur2
rm(straw_sur2)
glimpse(straw_sur)
```

```
Rows: 3,965
Columns: 15
$ Program       <chr> "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SURVE~
$ Year          <int> 2024, 2024, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 20~
$ Period        <chr> "YEAR", "YEAR", "MARKETING YEAR", "MARKETING YEAR", "MAR~
$ Geo.Level     <chr> "NATIONAL", "NATIONAL", "NATIONAL", "NATIONAL", "NATIONA~
$ State         <chr> "US TOTAL", "US TOTAL", "US TOTAL", "US TOTAL", "US TOTA~
$ straw         <chr> "STRAWBERRIES", "STRAWBERRIES", "STRAWBERRIES", "STRAWBE~
$ mkt           <chr> "FRESH MARKET - PRICE RECEIVED", "PROCESSING - PRICE REC~
$ measure       <chr> "ADJUSTED BASE", "ADJUSTED BASE", "MEASURED IN $ / CWT",~
$ other         <chr> "MEASURED IN $ / CWT", "MEASURED IN $ / TON", NA, NA, NA~
$ Domain        <chr> "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "T~
$ Type          <chr> "NOT SPECIFIED", "NOT SPECIFIED", "NOT SPECIFIED", "NOT ~
$ Name          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ Chemical.Code <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
$ Value         <dbl> 1.090000e+01, 4.040000e+00, 1.230000e+02, 1.420000e+02, ~
$ CV....        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

For this data, we are mainly interested in data with states California and Florida. The variables that are interested in are market, measure, chemical type, chemical name, and value.

First, let's explore the variable market.

```
unique(straw_sur$mkt)
```

```
 [1] "FRESH MARKET - PRICE RECEIVED" "PROCESSING - PRICE RECEIVED"
 [3] "PRICE RECEIVED"                "ACRES HARVESTED"
 [5] "ACRES PLANTED"                 "PRODUCTION"
 [7] "YIELD"                         "FRESH MARKET - PRODUCTION"
 [9] "FRESH MARKET"                  "NOT SOLD - PRODUCTION"
[11] "PROCESSING - PRODUCTION"       "PROCESSING"
[13] "UTILIZED - PRODUCTION"         "APPLICATIONS"
[15] "TREATED"                       "BEARING - APPLICATIONS"
[17] "BEARING - TREATED"
```

The above shows the professional terms that are related to the production and pricing of the strawberry market. To further understand the meaning of the above terms, explanation of some terms are provided below.

FRESH MARKET - PRICE RECEIVED: The farmers directly sold the strawberries to the fresh market and received the price for strawberry, measured in $ / CWT.

PROCESSING - PRICE RECEIVED: The price received by producers from the processing market such as using strawberry to make jams etc., measured in $ / TON.

PRODUCTION: The value of strawberries produced, measured in $.

FRESH MARKET - PRODUCTION: The value of strawberries produced for the fresh market, measured in $.

NOT SOLD - PRODUCTION: The weight of unsold strawberries production, measured in cwt.

BEARING - APPLICATIONS: The number of chemicals used for bearing strawberry plants.

Another element that we are interested in is the chemicals.

```
unique(straw_sur$Type)
```

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

```
nrow(straw_sur[straw_sur$Type == "NOT SPECIFIED", ])
```

```
[1] 491
```

```
nrow(straw_sur[straw_sur$Type == "FUNGICIDE", ])
```

```
[1] 1266
```

```
nrow(straw_sur[straw_sur$Type == "INSECTICIDE", ])
```

```
[1] 1286
```

```
nrow(straw_sur[straw_sur$Type == "HERBICIDE", ])
```

```
[1] 301
```

```
nrow(straw_sur[straw_sur$Type == "FERTILIZER", ])
```

```
[1] 115
```

There are four main chemicals treated to strawberries or strawberry plants appeared in this data, they are fungicide, insecticide, herbicide, and fertilizer. we will dig deeper in the following analysis.

Lastly, let's look the information about state.

```
unique(straw_sur$State)
```

```
[1] "US TOTAL"       "CALIFORNIA"     "FLORIDA"        "OTHER STATES"
[5] "NEW YORK"       "NORTH CAROLINA" "OREGON"         "WASHINGTON"
```

```
nrow(straw_sur[straw_sur$State == "CALIFORNIA", ])
```

```
[1] 2295
```

```
nrow(straw_sur[straw_sur$State == "FLORIDA", ])
```

```
[1] 1375
```

```
nrow(straw_sur[straw_sur$State == "NEW YORK", ])
```

```
[1] 25
```

```
nrow(straw_sur[straw_sur$State == "NORTH CAROLINA", ])
```

```
[1] 28
```

```
nrow(straw_sur[straw_sur$State == "OREGON", ])
```

```
[1] 25
```

```
nrow(straw_sur[straw_sur$State == "WASHINGTON", ])
```

```
[1] 25
```

From the above results, we can see that the two states California and Florida have the most observations. Therefore, we are going to focus on strawberry market on these two states.

California Data

```r
# California data
california <- straw_sur %>% filter(str_detect(straw_sur$State, "CALIFORNIA"))
head(california)
```

```
# A tibble: 6 x 15
  Program  Year Period    Geo.Level State straw mkt   measure other Domain Type
  <chr>   <int> <chr>     <chr>     <chr> <chr> <chr> <chr>   <chr> <chr>  <chr>
1 SURVEY   2023 MARKETIN~ STATE     CALI~ STRA~ PRIC~ MEASUR~ <NA>  TOTAL  NOT ~
2 SURVEY   2023 MARKETIN~ STATE     CALI~ STRA~ FRES~ MEASUR~ <NA>  TOTAL  NOT ~
3 SURVEY   2023 MARKETIN~ STATE     CALI~ STRA~ PROC~ MEASUR~ <NA>  TOTAL  NOT ~
4 SURVEY   2023 YEAR      STATE     CALI~ STRA~ ACRE~ <NA>    <NA>  TOTAL  NOT ~
5 SURVEY   2023 YEAR      STATE     CALI~ STRA~ ACRE~ <NA>    <NA>  TOTAL  NOT ~
6 SURVEY   2023 YEAR      STATE     CALI~ STRA~ APPL~ MEASUR~ <NA>  CHEMI~ FUNG~
# i 4 more variables: Name <chr>, Chemical.Code <int>, Value <dbl>,
#   CV.... <lgl>
```

```r
glimpse(california)
```

```
Rows: 2,295
Columns: 15
$ Program       <chr> "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SURVE~
$ Year          <int> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 2023, 20~
$ Period        <chr> "MARKETING YEAR", "MARKETING YEAR", "MARKETING YEAR", "Y~
$ Geo.Level     <chr> "STATE", "STATE", "STATE", "STATE", "STATE", "STATE", "S~
$ State         <chr> "CALIFORNIA", "CALIFORNIA", "CALIFORNIA", "CALIFORNIA", ~
$ straw         <chr> "STRAWBERRIES", "STRAWBERRIES", "STRAWBERRIES", "STRAWBE~
$ mkt           <chr> "PRICE RECEIVED", "FRESH MARKET - PRICE RECEIVED", "PROC~
$ measure       <chr> "MEASURED IN $ / CWT", "MEASURED IN $ / CWT", "MEASURED ~
$ other         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "AVG", "AVG", "AVG",~
$ Domain        <chr> "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "CHEMICAL, ~
$ Type          <chr> "NOT SPECIFIED", "NOT SPECIFIED", "NOT SPECIFIED", "NOT ~
$ Name          <chr> NA, NA, NA, NA, NA, "OXATHIAPIPROLIN", "CYCLANILIPROLE",~
$ Chemical.Code <int> NA, NA, NA, NA, NA, 128111, 26202, 109701, 115003, 12811~
$ Value         <dbl> 121, NA, NA, 42700, 43100, NA, NA, NA, NA, NA, NA, NA, N~
$ CV....        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

The above is the first few rows about the California data. Notice that there are some NAs in the Value column. It is hard to fill or replace the NAs since it is difficult for us to investigate the value of production or sales since some producers or farmers are both involved with processing and fresh market at the same time. Therefore, it is hard to make adjustment about the NAs.

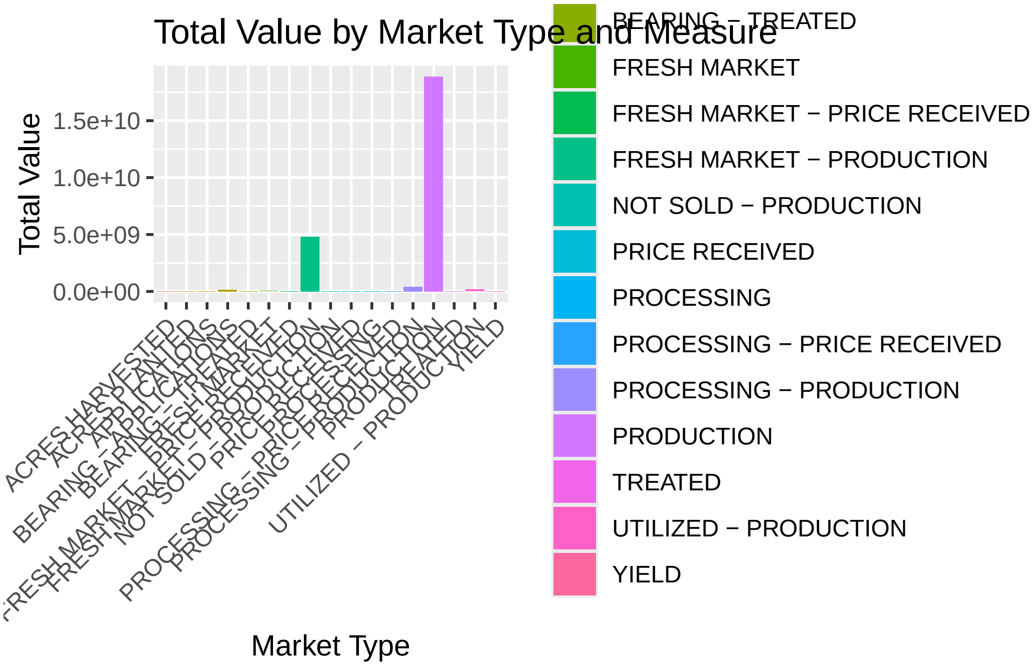
Strawberry Market in California

Now, let's investigate how is the value distributed in strawberry market.

```
# calculate the sum of values for each market
# with their specific measure or units
sum_values <- california %>% group_by(mkt, measure) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
```

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

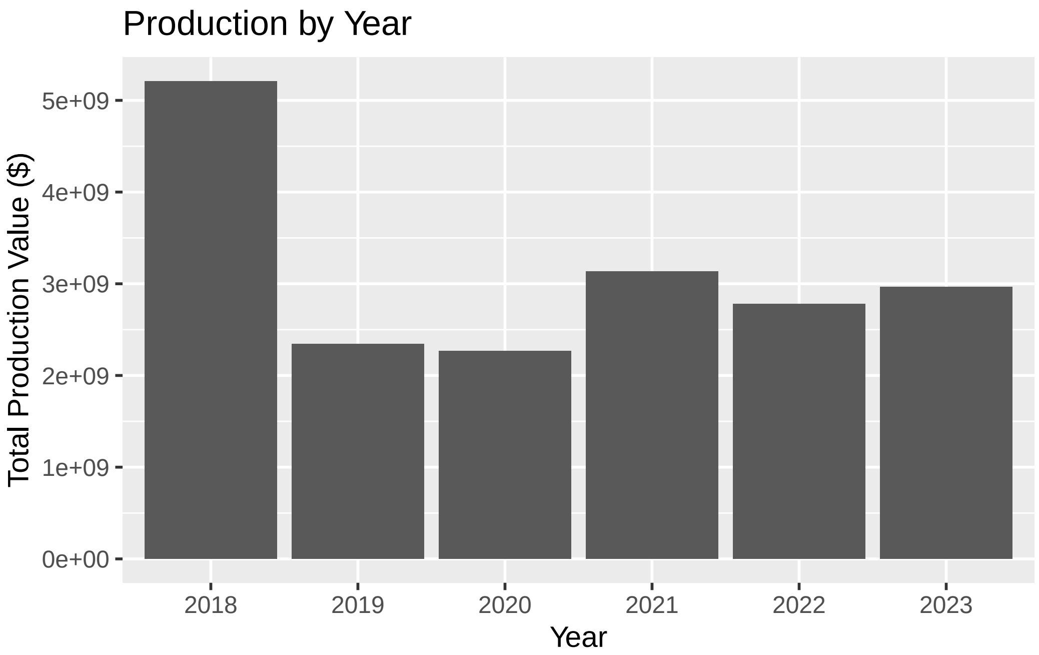
```
# bar graph of the distribution of the strawberry market
ggplot(sum_values, aes(x = mkt, y = sum, fill = mkt)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Value by Market Type and Measure", x = "Market Type", y = "Total Value
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



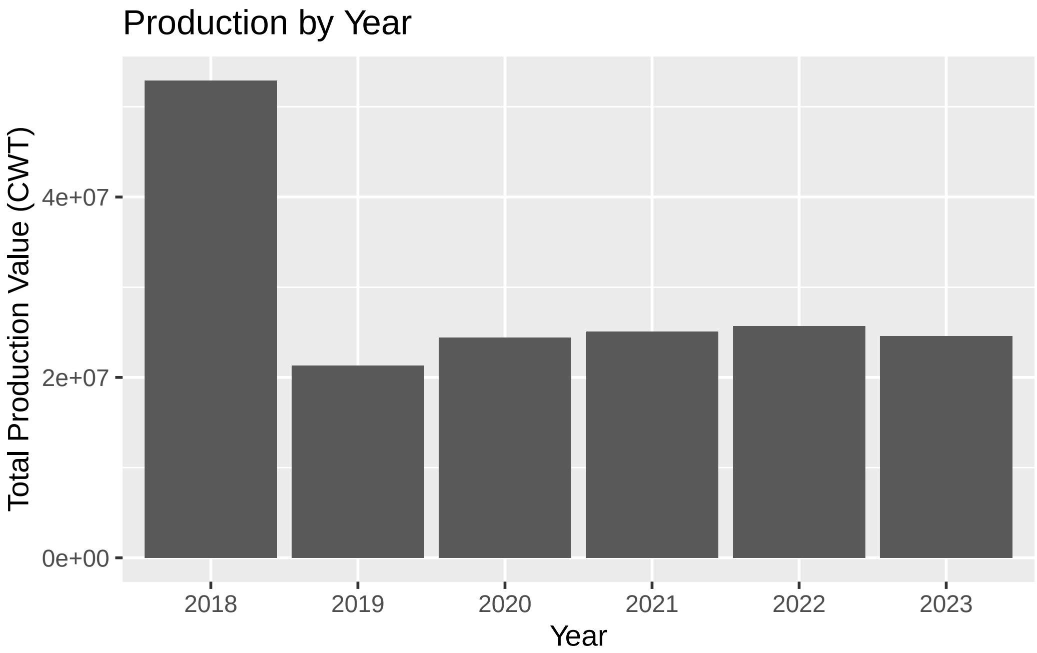
The above bar graph shows how the values are distributed in strawberry market with each specific unit. From here, we can see that the state California is mainly focused on production and fresh market production for strawberry market.

Next, we are going to look if there is an increase in production over the years.

```
# filter out the production under market
production <- california %>%
  filter(mkt == "PRODUCTION", measure == "MEASURED IN $")
production_cwt <- california %>%
  filter(mkt == "PRODUCTION", measure == "MEASURED IN CWT")
# total production for each year ($)
total_prod <- production %>%
  group_by(Year) %>%
  summarise(total_value = sum(Value, na.rm = TRUE))
# total production for each year (CWT)
total_prod_cwt <- production_cwt %>%
  group_by(Year) %>%
  summarise(total_value = sum(Value, na.rm = TRUE))

# bar graph for total production in $
ggplot(total_prod, aes(factor(Year), total_value)) +
  geom_bar(stat = "identity") +
  labs(title = "Production by Year",
       x = "Year",
       y = "Total Production Value ($)")
```



```
# bar graph for tatoal production in CWT
ggplot(total_prod_cwt, aes(factor(Year), total_value)) +
  geom_bar(stat = "identity") +
  labs(title = "Production by Year",
       x = "Year",
       y = "Total Production Value (CWT)")
```



From above, we can see that the year 2018 has the most production in strawberries whether or not it is measured in $ or CWT. There is a sharp decrease in the year of 2019 and the production went back up a little starting from 2020 to 2023.

Chemicals

Now, let's focusing on the factor chemicals.

```
# check the types of chemical in California Data
unique(california$Type)
```

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

Then, let's filter out the data based on each chemical type.

```r
# filter out the corresponding chemical type
## fungicide
fung <- california %>%
  filter(Type == "FUNGICIDE")
## insecticide
insect <- california %>%
  filter(Type == "INSECTICIDE")

## other
other <- california %>%
  filter(Type == "OTHER")

## herbicide
herb <- california %>%
  filter(Type == "HERBICIDE")

## fertilizer
fert <- california %>%
  filter(Type == "FERTILIZER" )
```

Fungicide

Here are the summaries of sum of values measured in different units for each year using the chemical type Fungicide.

```r
# measure in LB
lb_fung <- fung %>%
  filter(str_detect(measure, "MEASURED IN LB"))
lb_fung <- lb_fung %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
lb_fung$Type <- rep("Fungicide", nrow(lb_fung))
lb_fung
```

```
# A tibble: 4 x 3
   Year      sum Type
  <int>    <dbl> <chr>
1  2018  930951. Fungicide
2  2019 2466193. Fungicide
3  2021 2103570. Fungicide
4  2023 4641888. Fungicide
```

```
# measure in percentage of area bearing
pct_fung <- fung %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
pct_fung <- pct_fung %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
pct_fung$Type <- rep("Fungicide", nrow(pct_fung))
pct_fung
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018   556 Fungicide
2  2019  1146 Fungicide
3  2021   998 Fungicide
4  2023  1246 Fungicide
```

```
# measure in number
num_fung <- fung %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
num_fung <- num_fung %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
num_fung$Type <- rep("Fungicide", nrow(num_fung))
num_fung
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018  42.1 Fungicide
2  2019  73.1 Fungicide
3  2021  55.4 Fungicide
4  2023  63.3 Fungicide
```

Insecticide

Here are the summaries of sum of values measured in different units for each year using the chemical type Insecticide.

```
# measure in LB
lb_ins <- insect %>%
  filter(str_detect(measure, "MEASURED IN LB"))
lb_ins <- lb_ins %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
lb_ins$Type <- rep("Insecticide", nrow(lb_ins))
lb_ins
```

```
# A tibble: 4 x 3
   Year     sum Type
  <int>   <dbl> <chr>
1  2018 264032. Insecticide
2  2019 415436. Insecticide
3  2021 222922. Insecticide
4  2023 361713. Insecticide
```

```
# measure in percentage of area bearing
pct_ins <- insect %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
pct_ins <- pct_ins %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
pct_ins$Type <- rep("Insecticide", nrow(pct_ins))
pct_ins
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018   528 Insecticide
2  2019  1070 Insecticide
3  2021   918 Insecticide
4  2023  1018 Insecticide
```

```
# measure in number
num_ins <- insect %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
num_ins <- num_ins %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
num_ins$Type <- rep("Insecticide", nrow(num_ins))
num_ins
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018  39.9 Insecticide
```

```
2  2019  70   Insecticide
3  2021  44   Insecticide
4  2023  42.4 Insecticide
```

Other

Here are the summaries of sum of values measured in different units in each year using the other type of chemical.

```
# measure in LB
lb_o <- other %>%
  filter(str_detect(measure, "MEASURED IN LB"))
lb_o <- lb_o %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
lb_o$Type <- rep("Other", nrow(lb_o))
lb_o
```

```
# A tibble: 4 x 3
   Year       sum Type
  <int>     <dbl> <chr>
1  2018  7007063. Other
2  2019 15393391. Other
3  2021 15202598. Other
4  2023 28558274. Other
```

```
# measure in percentage of area bearing
pct_o <- other %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
pct_o <- pct_o %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
pct_o$Type <- rep("Other", nrow(pct_o))
pct_o
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018   121 Other
2  2019   174 Other
3  2021   264 Other
4  2023   200 Other
```

```
# measure in number
num_o <- other %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
num_o <- num_o %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
num_o$Type <- rep("Other", nrow(num_o))
num_o
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018   4.6 Other
2  2019   9.7 Other
3  2021  10.3 Other
4  2023   5.4 Other
```

Herbicide

Here are the summaries of sum of values measured in different units for each year using the chemical type Herbicide.

```
# measure in LB
lb_h <- herb %>%
  filter(str_detect(measure, "MEASURED IN LB"))
lb_h <- lb_h %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
lb_h$Type <- rep("Herbicide", nrow(lb_h))
lb_h
```

```
# A tibble: 4 x 3
   Year     sum Type
  <int>   <dbl> <chr>
1  2018  10804. Herbicide
2  2019  18304. Herbicide
3  2021  29904. Herbicide
4  2023 165007. Herbicide
```

```
# measure in percentage of area bearing
pct_h <- herb %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
pct_h <- pct_h %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
pct_h$Type <- rep("Herbicide", nrow(pct_h))
pct_h
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018    21 Herbicide
2  2019    57 Herbicide
3  2021    70 Herbicide
4  2023    86 Herbicide
```

```
# measure in number
num_h <- herb %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
num_h <- num_h %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
num_h$Type <- rep("Herbicide", nrow(num_h))
num_h
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018   3.5 Herbicide
2  2019   8.3 Herbicide
3  2021   3.5 Herbicide
4  2023  10.5 Herbicide
```

Fertilizer

Here are the summaries of sum of values measured in different units for each year using the chemical type Fertilizer.

```
# measure in LB
lb_f <- fert %>%
  filter(str_detect(measure, "MEASURED IN LB"))
lb_f <- lb_f %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
lb_f$Type <- rep("Fertilizer", nrow(lb_f))
lb_f
```

```
# A tibble: 3 x 3
   Year      sum Type
  <int>    <dbl> <chr>
1  2018 29641951 Fertilizer
2  2019  1648256 Fertilizer
3  2023  1002430 Fertilizer
```

```
# measure in percentage of area bearing
pct_f <- fert %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
pct_f <- pct_f %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
pct_f$Type <- rep("Fertilizer", nrow(pct_f))
pct_f
```

```
# A tibble: 3 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018   365 Fertilizer
2  2019   226 Fertilizer
3  2023   196 Fertilizer
```

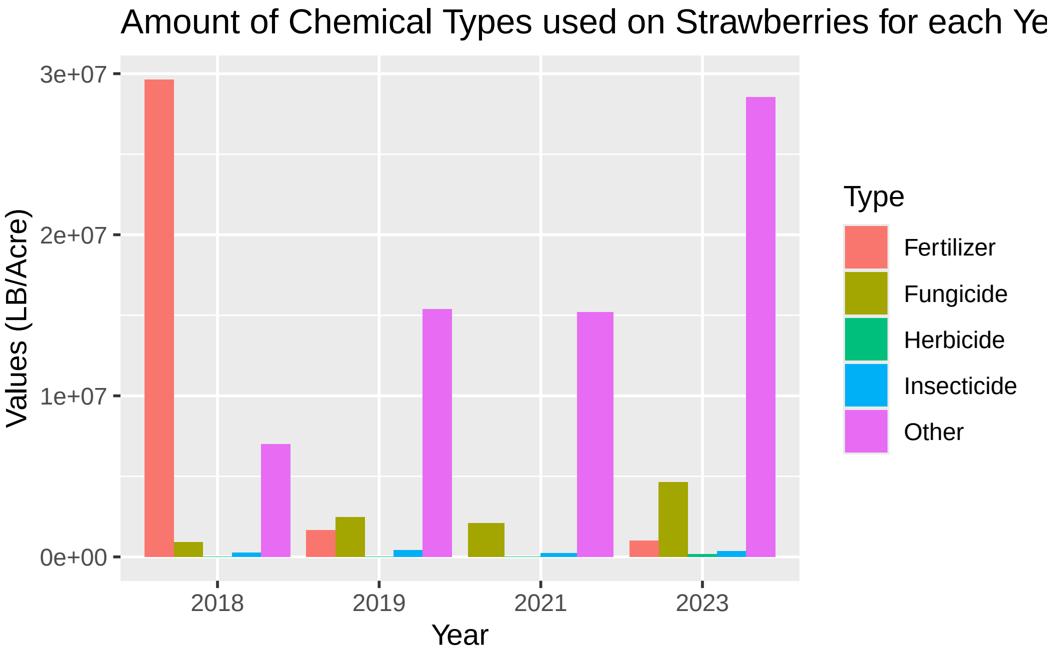
```
# measure in number
num_f <- fert %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
num_f <- num_f %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
num_f$Type <- rep("Fertilizer", nrow(num_f))
num_f
```

```
# A tibble: 3 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018  52.1 Fertilizer
2  2019  43.7 Fertilizer
3  2023  31.7 Fertilizer
```

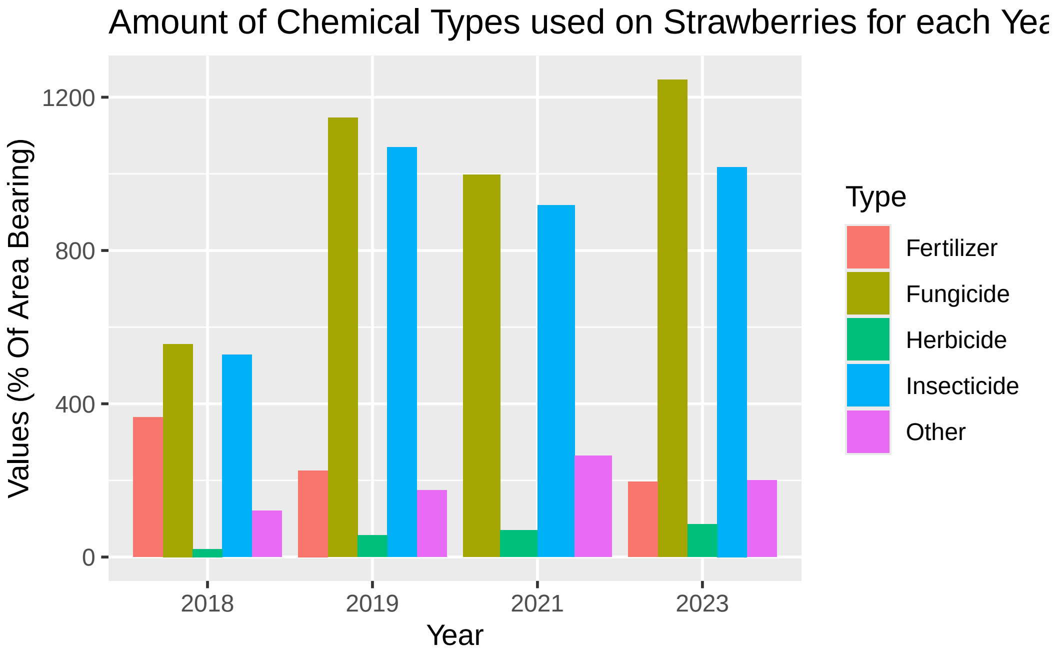
Visualization of The Above Summaries

We will use the above summaries to create a bar graph about each chemical type.

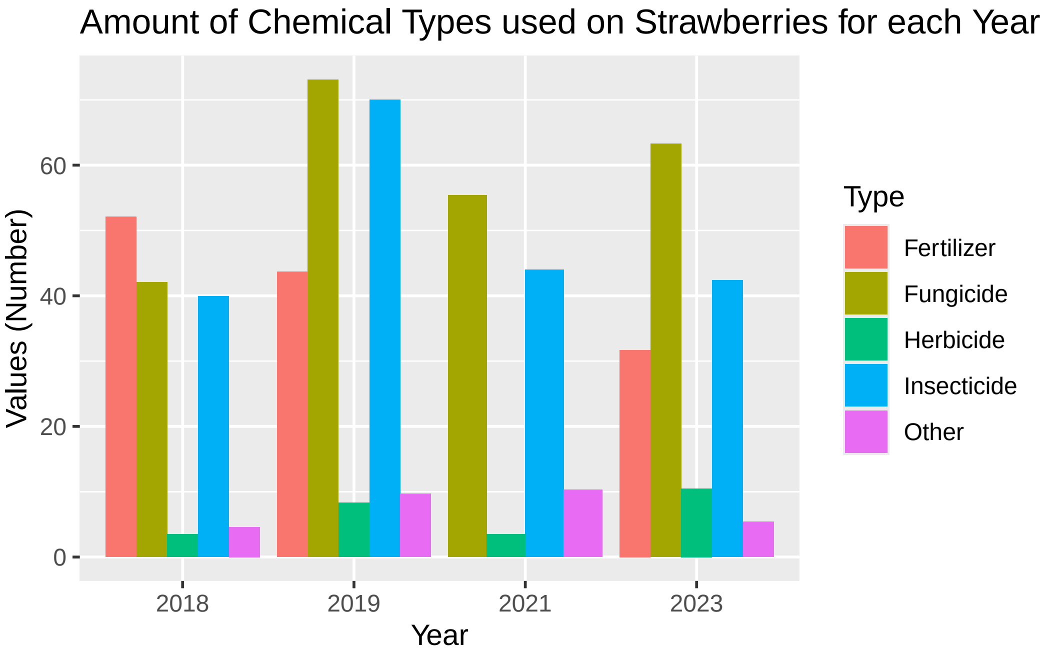
```
# recreate a dataset that contains the summaries above
lb_california <- rbind(lb_fung, lb_ins, lb_o, lb_h, lb_f)
pct_california <- rbind(pct_fung, pct_ins, pct_o, pct_h, pct_f)
num_california <- rbind(num_fung, num_ins, num_o, num_h, num_f)

# Visualizations
## LB/Acre
ggplot(lb_california, aes(x = factor(Year), y = sum, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Amount of Chemical Types used on Strawberries for each Year",
       x = "Year",
       y = "Values (LB/Acre)",
       fill = "Type")
```



```
## PCT
ggplot(pct_california, aes(x = factor(Year), y = sum, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Amount of Chemical Types used on Strawberries for each Year",
       x = "Year",
       y = "Values (% Of Area Bearing)",
       fill = "Type")
```



```r
## number
ggplot(num_california, aes(x = factor(Year), y = sum, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Amount of Chemical Types used on Strawberries for each Year (California)",
       x = "Year",
       y = "Values (Number)",
       fill = "Type")
```



From the above graphs, we can see that if we are focusing on the measurement of LB/Acre, other chemicals are mainly used other chemical types instead of the four types listed, except in 2018, fertilizer used the most. By looking at the measurement of percentage of Area Bearing, we can see that for each year, the most used chemical type on strawberries is fungicide, followed by insecticide. Lastly, by looking the measure in numbers, we can see that fungicide is used the most in the years of 2019, 2021, and 2023. However, in the year of 2018, we can see that fertilizer is used the most.

Florida Data

Now, moving onto the state of Florida.

```
# Florida data
florida <- straw_sur %>% filter(str_detect(straw_sur$State, "FLORIDA"))
```

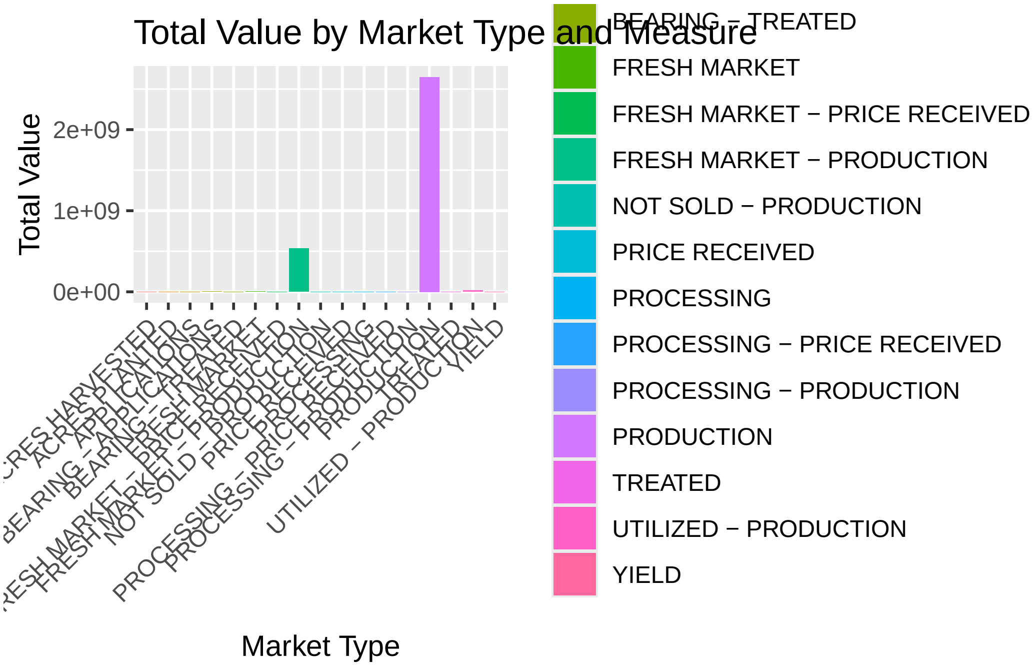
Strawberry Market in Florida

Let’s investigate how is the value distributed in strawberry market in Florida.

```
# calculate the sum of values for each market with their specific measure or units
sum_values_f <- florida %>% group_by(mkt, measure) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
```

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

```
# bar graph of the distribution of the strawberry market
ggplot(sum_values_f, aes(x = mkt, y = sum, fill = mkt)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Value by Market Type and Measure", x = "Market Type", y = "Total Value
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

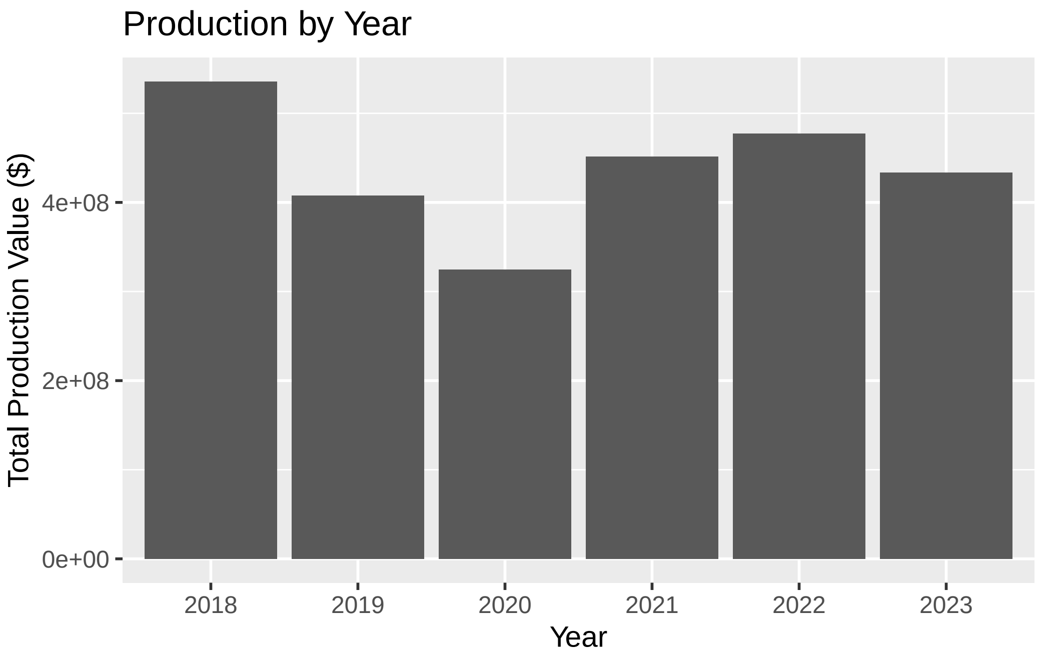


The above bar graph shows how the values are distributed in strawberry market in Florida with each specific unit. From here, we can see that Florida is mainly focused on production and fresh market production for strawberry market, same as California.

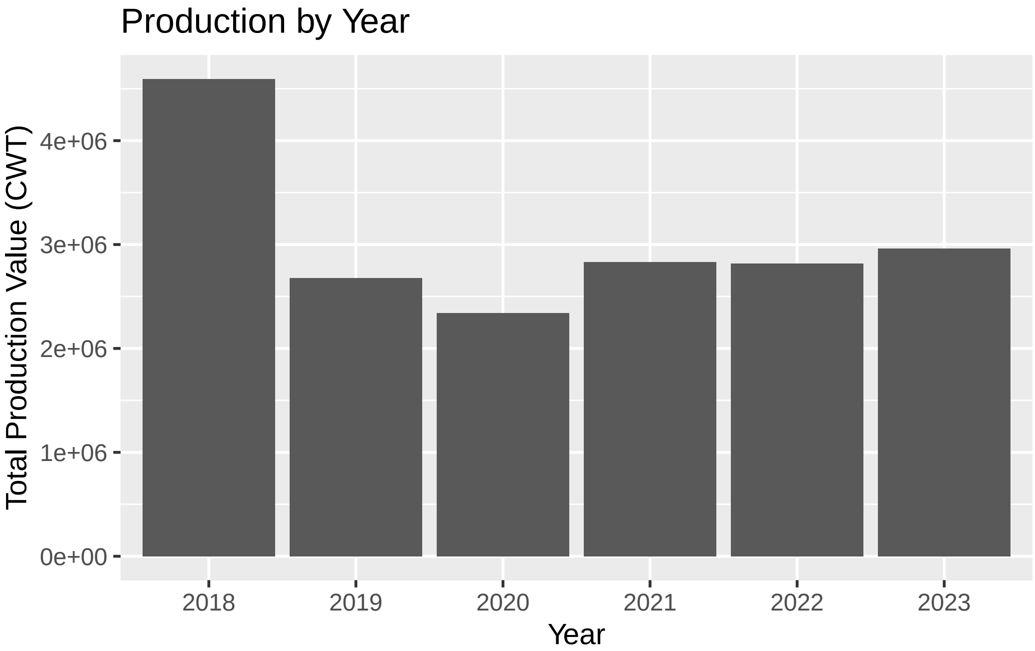
Next, we are going to look if there is an increase in production over the years.

```
# filter out the production under market
production_f <- florida %>%
  filter(mkt == "PRODUCTION", measure == "MEASURED IN $")
```

```
production_cwt_f <- florida %>%
  filter(mkt == "PRODUCTION", measure == "MEASURED IN CWT")
# total production for each year ($)
total_prod_f <- production_f %>%
  group_by(Year) %>%
  summarise(total_value = sum(Value, na.rm = TRUE))
# total production for each year (CWT)
total_prod_cwt_f <- production_cwt_f %>%
  group_by(Year) %>%
  summarise(total_value = sum(Value, na.rm = TRUE))

# bar graph for total production in $
ggplot(total_prod_f, aes(factor(Year), total_value)) +
  geom_bar(stat = "identity") +
  labs(title = "Production by Year",
       x = "Year",
       y = "Total Production Value ($)")
```



```
# bar graph for tatoal production in CWT
ggplot(total_prod_cwt_f, aes(factor(Year), total_value)) +
  geom_bar(stat = "identity") +
  labs(title = "Production by Year",
       x = "Year",
       y = "Total Production Value (CWT)")
```



From the above two bar plots with units in $ and CWT, we can see that there is a sharp decrease in production in CWT from 2018 to 2019, and went back a little start from 2021 to 2023. Then, by looking at the production value in dollars, we can say that there is an overall decrease from 2018 to 2020, and it went back up adn become steady started from 2021 to 2023.

Comparison between California and Florida

Notice that California and Florida faces the similar situation in strawberry production, where 2018 is the peak in production and 2020 the lowest. There must exist factors that impact the overall production in the year of 2019 to 2020. Possible affects could be climate change and COVID-19. I believe the COVID could be the main impact of strawberry production in the year of 2019 to 2020.

Chemicals

Now, let's focusing on the factor chemicals.

```
# check the types of chemical in California Data
unique(florida$Type)
```

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

Same chemical types as California. Now, we are going to investigate the situations of these chemical types in Florida.

Similar to what we did for California, let's filter out the chemical types.

```
# filter out the corresponding chemical type
## fungicide
fung_f <- florida %>%
  filter(Type == "FUNGICIDE")
## insecticide
insect_f <- florida %>%
  filter(Type == "INSECTICIDE")

## other
other_f <- florida %>%
  filter(Type == "OTHER")

## herbicide
herb_f <- florida %>%
  filter(Type == "HERBICIDE")

## fertilizer
fert_f <- florida %>%
  filter(Type == "FERTILIZER" )
```

Fungicide

Summary of values measured in different units for each year using the chemical type Fungicide.

```
# measure in LB
fl_lb_fung <- fung_f %>%
  filter(str_detect(measure, "MEASURED IN LB"))
fl_lb_fung <- fl_lb_fung %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
```

```r
fl_lb_fung$Type <- rep("Fungicide", nrow(fl_lb_fung))
fl_lb_fung
```

```
# A tibble: 4 x 3
   Year     sum Type
  <int>   <dbl> <chr>
1  2018 243321. Fungicide
2  2019 516844. Fungicide
3  2021 590750. Fungicide
4  2023 574729. Fungicide
```

```r
# measure in percentage of area bearing
fl_pct_fung <- fung_f %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
fl_pct_fung <- fl_pct_fung %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_pct_fung$Type <- rep("Fungicide", nrow(fl_pct_fung))
fl_pct_fung
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018   299 Fungicide
2  2019   415 Fungicide
3  2021   319 Fungicide
4  2023   444 Fungicide
```

```r
# measure in number
fl_num_fung <- fung_f %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
fl_num_fung <- fl_num_fung %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_num_fung$Type <- rep("Fungicide", nrow(fl_num_fung))
fl_num_fung
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018  13.9 Fungicide
2  2019  25.6 Fungicide
3  2021  26.2 Fungicide
4  2023  18.6 Fungicide
```

Insecticide

The summary of sum of the values measured in different units for each year using the chemical type Insecticide.

```
# measure in LB
fl_lb_ins <- insect_f %>%
  filter(str_detect(measure, "MEASURED IN LB"))
fl_lb_ins <- fl_lb_ins %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_lb_ins$Type <- rep("Insecticide", nrow(fl_lb_ins))
fl_lb_ins
```

```
# A tibble: 4 x 3
   Year     sum Type
  <int>   <dbl> <chr>
1  2018 102700  Insecticide
2  2019   8402. Insecticide
3  2021   8001. Insecticide
4  2023  15301. Insecticide
```

```
# measure in percentage of area bearing
fl_pct_ins <- insect_f %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
fl_pct_ins <- fl_pct_ins %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_pct_ins$Type <- rep("Insecticide", nrow(fl_pct_ins))
fl_pct_ins
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018    37 Insecticide
2  2019   307 Insecticide
3  2021   184 Insecticide
4  2023   248 Insecticide
```

```
# measure in number
fl_num_ins <- insect_f %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
fl_num_ins <- fl_num_ins %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_num_ins$Type <- rep("Insecticide", nrow(fl_num_ins))
fl_num_ins
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018    0  Insecticide
2  2019  19.6 Insecticide
3  2021   6.5 Insecticide
4  2023   8   Insecticide
```

Other

The summary of sum of the values measured in different units in each year using the other type of chemical.

```
# measure in LB
fl_lb_o <- other_f %>%
  filter(str_detect(measure, "MEASURED IN LB"))
fl_lb_o <- fl_lb_o %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_lb_o$Type <- rep("Other", nrow(fl_lb_o))
fl_lb_o
```

```
# A tibble: 4 x 3
   Year    sum Type
  <int>  <dbl> <chr>
1  2018      0 Other
2  2019 125900 Other
3  2021   8600 Other
4  2023   5100 Other
```

```
# measure in percentage of area bearing
fl_pct_o <- other_f %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
fl_pct_o <- fl_pct_o %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_pct_o$Type <- rep("Other", nrow(fl_pct_o))
fl_pct_o
```

```
# A tibble: 4 x 3
   Year   sum Type 
  <int> <dbl> <chr>
1  2018     0 Other
2  2019    17 Other
3  2021    49 Other
4  2023    22 Other
```

```
# measure in number
fl_num_o <- other_f %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
fl_num_o <- fl_num_o %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_num_o$Type <- rep("Other", nrow(fl_num_o))
fl_num_o
```

```
# A tibble: 4 x 3
   Year   sum Type 
  <int> <dbl> <chr>
1  2018     0 Other
2  2019     0 Other
3  2021     0 Other
4  2023     0 Other
```

Herbicide

The summary of sum of values measured in different units for each year using the chemical type Herbicide.

```
# measure in LB
fl_lb_h <- herb_f %>%
  filter(str_detect(measure, "MEASURED IN LB"))
fl_lb_h <- fl_lb_h %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_lb_h$Type <- rep("Herbicide", nrow(fl_lb_h))
fl_lb_h
```

```
# A tibble: 4 x 3
   Year   sum Type 
  <int> <dbl> <chr>
1  2018     0 Herbicide
```

```
2  2019   300 Herbicide
3  2021  2600 Herbicide
4  2023  9900 Herbicide
```

```r
# measure in percentage of area bearing
fl_pct_h <- herb_f %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
fl_pct_h <- fl_pct_h %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_pct_h$Type <- rep("Herbicide", nrow(fl_pct_h))
fl_pct_h
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018     0 Herbicide
2  2019    15 Herbicide
3  2021     9 Herbicide
4  2023    26 Herbicide
```

```r
# measure in number
fl_num_h <- herb_f %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
fl_num_h <- fl_num_h %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_num_h$Type <- rep("Herbicide", nrow(fl_num_h))
fl_num_h
```

```
# A tibble: 4 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018     0 Herbicide
2  2019     0 Herbicide
3  2021     0 Herbicide
4  2023     0 Herbicide
```

Fertilizer

The summary of sum of values measured in different units for each year using the chemical type Fertilizer.

```
# measure in LB
fl_lb_f <- fert_f %>%
  filter(str_detect(measure, "MEASURED IN LB"))
fl_lb_f <- fl_lb_f %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_lb_f$Type <- rep("Fertilizer", nrow(fl_lb_f))
fl_lb_f
```

```
# A tibble: 3 x 3
   Year    sum Type
  <int>  <dbl> <chr>
1  2018 351037 Fertilizer
2  2019 795103 Fertilizer
3  2023 873189 Fertilizer
```

```
# measure in percentage of area bearing
fl_pct_f <- fert_f %>%
  filter(str_detect(measure, "MEASURED IN PCT OF AREA BEARING"))
fl_pct_f <- fl_pct_f %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_pct_f$Type <- rep("Fertilizer", nrow(fl_pct_f))
fl_pct_f
```

```
# A tibble: 3 x 3
   Year   sum Type
  <int> <dbl> <chr>
1  2018   297 Fertilizer
2  2019   226 Fertilizer
3  2023   128 Fertilizer
```

```
# measure in number
fl_num_f <- fert_f %>%
  filter(str_detect(measure, "MEASURED IN NUMBER"))
fl_num_f <- fl_num_f %>% group_by(Year) %>%
  summarise(sum = sum(Value, na.rm = TRUE))
fl_num_f$Type <- rep("Fertilizer", nrow(fl_num_f))
fl_num_f
```

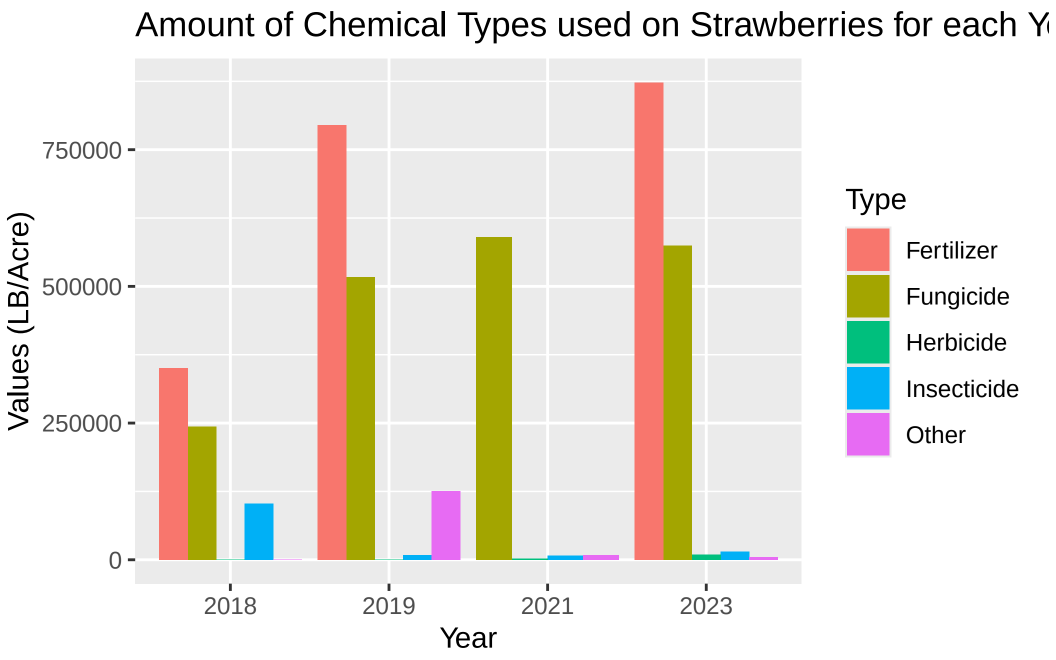
```
# A tibble: 3 x 3
   Year   sum Type
```

```
  <int> <dbl> <chr>
1  2018  53.6 Fertilizer
2  2019  72.5 Fertilizer
3  2023  14.8 Fertilizer
```

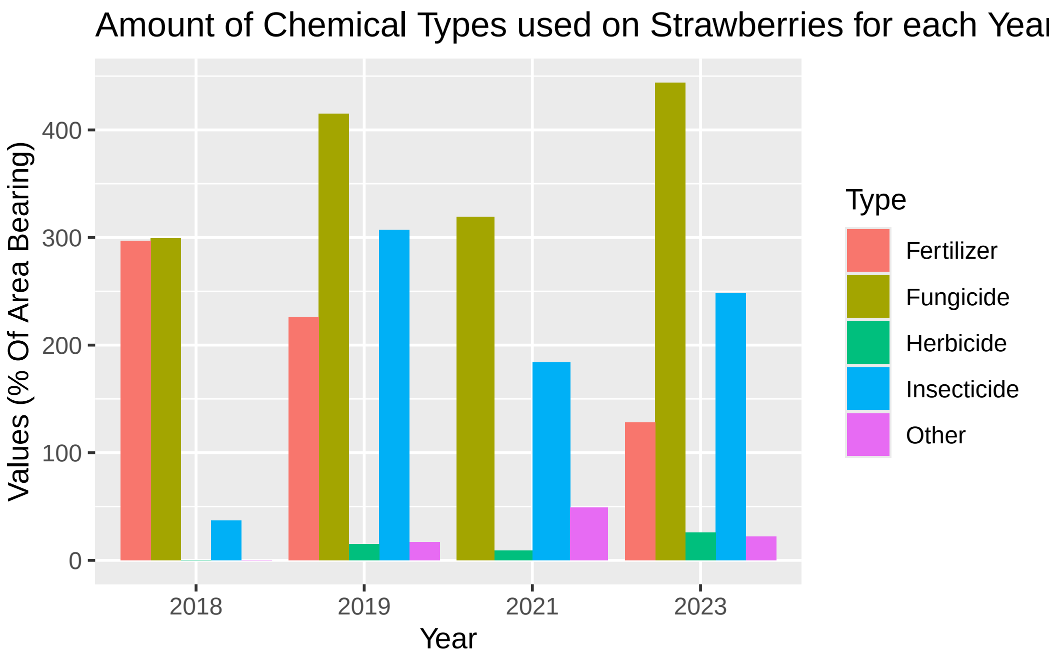
Visualization of The Above Summaries

Here is the visulation based on the information from the above summaries.

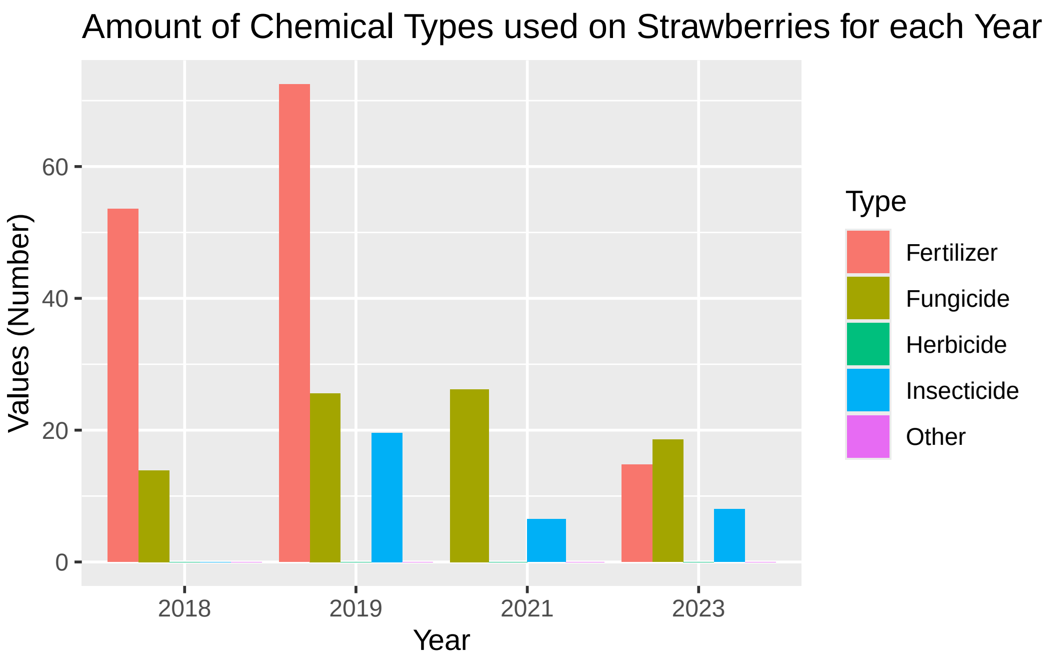
```r
# recreate a dataset that contains the summaries above
lb_fl <- rbind(fl_lb_fung, fl_lb_ins, fl_lb_o, fl_lb_h, fl_lb_f)
pct_fl <- rbind(fl_pct_fung, fl_pct_ins, fl_pct_o, fl_pct_h, fl_pct_f)
num_fl <- rbind(fl_num_fung, fl_num_ins, fl_num_o, fl_num_h, fl_num_f)

# Visualizations
## LB/Acre
ggplot(lb_fl, aes(x = factor(Year), y = sum, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Amount of Chemical Types used on Strawberries for each Year (Florida)",
       x = "Year",
       y = "Values (LB/Acre)",
       fill = "Type")
```



```r
## PCT
ggplot(pct_fl, aes(x = factor(Year), y = sum, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Amount of Chemical Types used on Strawberries for each Year (Florida)",
       x = "Year",
       y = "Values (% Of Area Bearing)",
       fill = "Type")
```



```r
## number
ggplot(num_fl, aes(x = factor(Year), y = sum, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Amount of Chemical Types used on Strawberries for each Year (Florida)",
       x = "Year",
       y = "Values (Number)",
       fill = "Type")
```



By looking at the first graph (measured in LB/Acre), we can see that for all the years except 2019, the most used type of chemical is fertilizer. Overall, the second most used chemical type is fungicide. By looking at the second graph (% area bearing), we can see that again fungicide is used the most. lastly, the third graph (measured in number), in the years of 2018 to 2019, fertilizers are mainly used and for other years, fungicide taken over.

Comparion of Chemicals between California and Florida

Both states used the chemical type fungicide for treatments on bearing plants or for growth of strawberries. However, Florida also considered fertilizer while California considered other chemicals. This comes up to a question of why is California using other chemicals instead of fertilizers. Further exploration is discussed below.

Let's dig deep into the chemical type fertilizer by loading the package from the website PubChem through exploration of hazard information.

```
# Load the PubChem package
library(PubChemR)

# get the different names of chemicals that fall under the fertilizer type
 unique(fert$Name)
```

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

```
# Use the function written from the lecture
GHS_searcher<-function(result_json_object){
  result<-result_json_object
  for (i in 1:length(result[["result"]][["Hierarchies"]][["Hierarchy"]])){
    if(result[["result"]][["Hierarchies"]][["Hierarchy"]][[i]][["SourceName"]]=="GHS Classifi
      return(i)
    }

  }
}

hazards_retriever<-function(index,result_json_object){
  result<-result_json_object
  hierarchy<-result[["result"]][["Hierarchies"]][["Hierarchy"]][[index]]
  i<-1
  output_list<-rep(NA,length(hierarchy[["Node"]]))
  while(str_detect(hierarchy[["Node"]][[i]][["Information"]][["Name"]],"H") & i<length(hierar
    output_list[i]<-hierarchy[["Node"]][[i]][["Information"]][["Name"]]
    i<-i+1
  }
  return(output_list[!is.na(output_list)])
}
```

Use the functions from the lecture, we can scrape the hazard information.

```
# Fertilizer: Nitrogen
result_1<-get_pug_rest(identifier = "nitrogen", namespace = "name", domain = "compound",opera

hazards_retriever(GHS_searcher(result_1),result_1)
```

```
 [1] "H280: Contains gas under pressure; may explode if heated [Warning Gases under pressure]
 [2] "H200: Physical Hazards"
 [3] "Hazard Statement Codes"
 [4] "H281: Contains refrigerated gas; may cause cryogenic burns or injury [Warning Gases und
 [5] "H317: May cause an allergic skin reaction [Warning Sensitization, Skin]"
 [6] "H300: Health Hazards"
 [7] "H319: Causes serious eye irritation [Warning Serious eye damage/eye irritation]"
 [8] "H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute h
 [9] "H400: Environmental Hazards"
[10] "H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aq
```

```
# Fertilizer: Phosphate
result_2<-get_pug_rest(identifier = "phosphate", namespace = "name", domain = "compound",ope

hazards_retriever(GHS_searcher(result_2),result_2)
```

```
logical(0)
```

```
# Fertilizer: Sulfur
result_3<-get_pug_rest(identifier = "sulfur", namespace = "name", domain = "compound",operat

hazards_retriever(GHS_searcher(result_3),result_3)
```

```
 [1] "H228: Flammable solid [Danger Flammable solids]"
 [2] "H200: Physical Hazards"
 [3] "Hazard Statement Codes"
 [4] "H315: Causes skin irritation [Warning Skin corrosion/irritation]"
 [5] "H300: Health Hazards"
 [6] "H317: May cause an allergic skin reaction [Warning Sensitization, Skin]"
 [7] "H319: Causes serious eye irritation [Warning Serious eye damage/eye irritation]"
 [8] "H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure]'
 [9] "H373: May causes damage to organs through prolonged or repeated exposure [Warning Spec
[10] "H413: May cause long lasting harmful effects to aquatic life [Hazardous to the aquatic
[11] "H400: Environmental Hazards"
```

From the above results, we can see that both nitrogen and sulfur are harmful to aquatic life. Notice that both states California and Florida are near oocean. Since these type of fertilizers are harmful to aquatic ecosystem, then it is better to reduce the use. As a result, Florida are using less fertilizers in recent years compared to the years of 2018 and 2019, as well as California.

After going to the PubChem website, it is found that phosphate is not classified as GHS hazard. In addition, PubChem do not include the information for Potash. So an additional reasearch pn potash is done and the reference link are provided at the end. The website ("The Canadian Encyclopedia") stated that the main component of potash is potassium, and mining this chemical would have harmful impact on vegetation, wildlife, and water pollution. Therefore, we can say that fertilizers are harmful to the environment, hence both states are using less fertilizers in the recent years.

Sources of Chemicals

pubChem

[Potash](https://www.thecanadianencyclopedia.ca/en/article/potash#:~:text=Global%20and%20Environment
Because%20potassium%20(the&text=The%20environmental%20impact%20of%20potash,volume%20water%20c