Stawberries

AUTHOR
Truc Minh Nguyen

PUBLISHED
October 28, 2024

Strawberries: Data

This is a project about acquiring strawberry data from the USDA-NASS system and then cleaning, organizing, and exploring the data in preparation for data analysis. To get started, I acquired the data from the USDA NASS system and downloaded them in a csv.

Fruit Growers News (1)

Questions about Strawberries

- Where they are grown? By whom?
- Are they really loaded with carcinogenic poisons?
- Are they really good for your health? Bad for your health?
- Are organic strawberries carriers of deadly diseases?
- When I go to the market should I buy conventional or organic strawberries?
- Do Strawberry farmers make money?
- How do the strawberries I buy get to my market?

Strawberry data source and parameters

The data set for this assignment has been selected from:

[USDA_NASS_strawb_2024SEP25.

The data have been stored on NASS here: USDA_NASS_strawb_2024SEP25 .

For the assignment, I stored the csv I downloaded on the MA615 Blackboard as strawberries25_v3.csv.

The data was originally collected at the county, state, and national levels, but the degree of missingness at the state level was too high, so I dropped the county-level data.

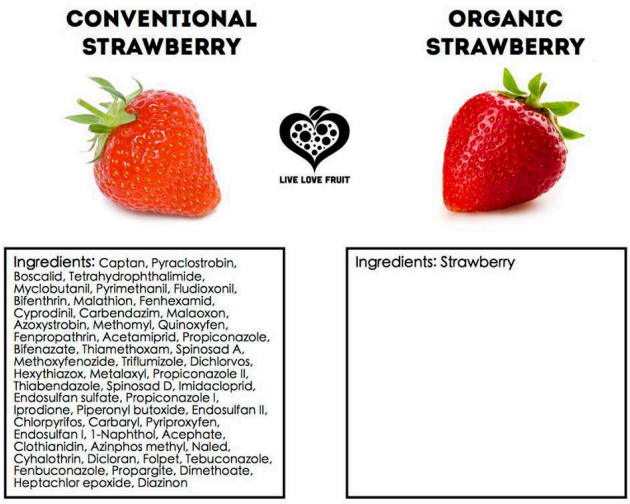
There are 5,359 rows and 21 column in the initial data set. The only complete year is 2022, although there is data for years 2018 through 2024.

To work with the data, define a function to remove columns with only single value in all its rows.

To work with this data, split the Census data from the Survey data.

Census data cleaning and organizing

we're examining census data because it's different from survey data



A Path to Health (2)

Survey data cleaning and organizing

Shift data into alignment function

Examine Domain

now look at totals

there are two markets for Strawberries – Fresh Marketing and Processing

make a table for each

from the Survey Totals

we have reports for

Markets: Fresh and Processing Operations: Growing and Production

California and Florida chemicals

```
library(webchem)
etox_basic(7242)
```

```
$`7242`
$`7242`$cas
[1] "82657-04-3"

$`7242`$ec
# A tibble: 2 × 0

$`7242`$gsbl
<unspecified> [0]

$`7242`$synonyms
# A tibble: 0 × 2
# ℹ 2 variables: name <chr>, language <chr>

$`7242`$source_url
[1] "https://webetox.uba.de/webETOX/public/basics/stoff.do?language=en&id=7242"


attr(,"class")
[1] "etox_basic" "list"
```

California and Florida fertilizers

Chemicals usd in strawberry cultivaion

Six deadly carcinogens from WHO list

captafol

ethylene dibromide also

glyphosate See also 1
2
3
4

malathion 1 2

diazinon 1 2 3

dichlorophenyltrichloroethane (DDT) 1 2 [3](https://www.epa.gov/ingredients-used-pesticide-products/ddt-brief-history-and-status#:~:text=DDT%20(dichloro%2Ddiphenyl%2Dtrichloroethane,both%20military%20and%20civilian%20populations.)

For contrast

Azadirachtin 1 2 3

Sources of agricultural chemical information

pubChem

EPA search

[ETOX]](https://webetox.uba.de/webETOX/index.do) webchem R pkg

Safety Data Sheets

for EPA number lookup epa numbers

Active Pesticide Product Registration Informational Listing

CAS for Methyl Bromide

pesticide chemical search

toxic chemical dashboard

pubChem

The EPA PC (Pesticide Chemical) Code is a unique chemical code number assigned by the EPA to a particular pesticide active ingredient, inert ingredient or mixture of active ingredients.

Investigating toxic pesticides

start here with chem PC code

Pesticide Product and Label System

Search by Chemical

CompTox Chemicals Dashboard

Active Pesticide Product Registration Informational Listing

OSHA chemical database

Pesticide Ingredients

NPIC Product Research Online (NPRO)

Databases for Chemical Information

Pesticide Active Ingredients

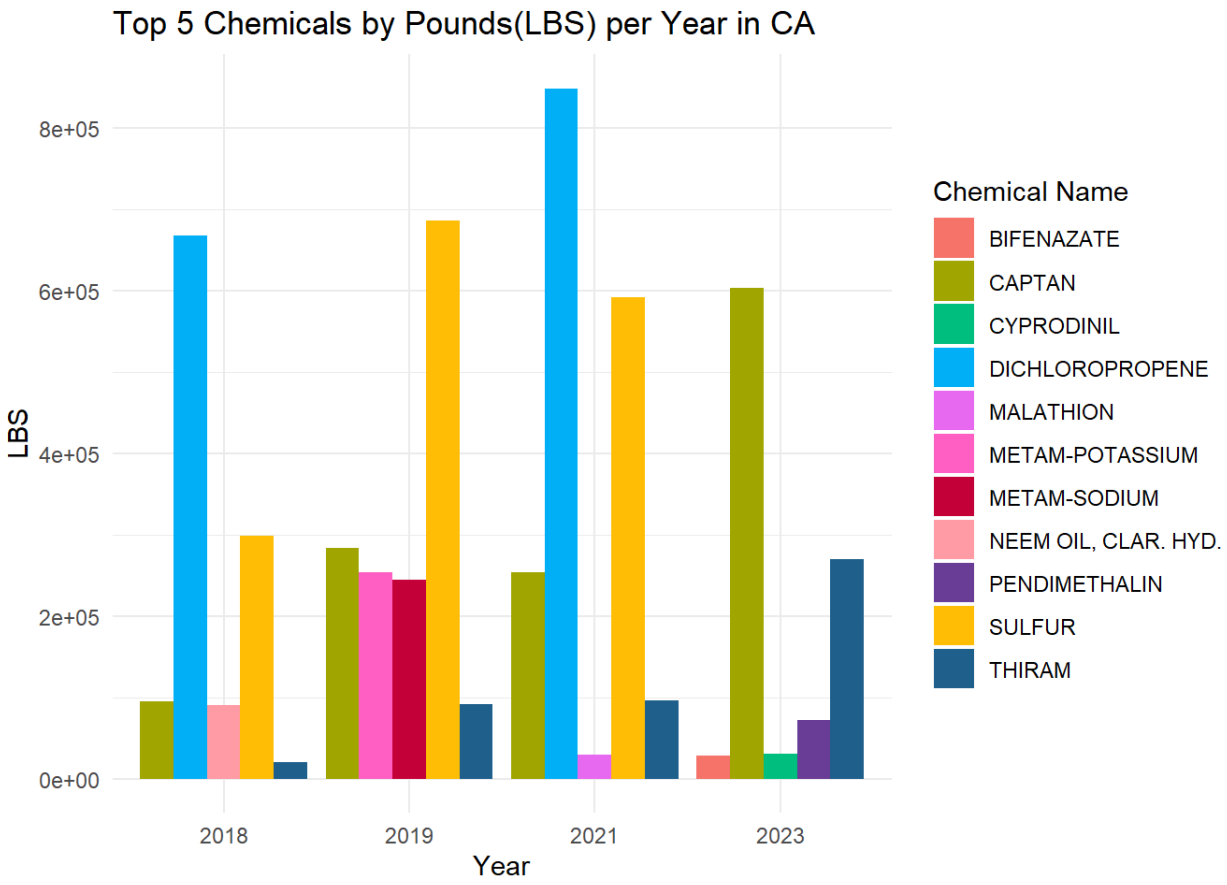
TSCA Chemical Substance Inventory

I want to explore the top 5 chemicals used by lbs in California for the past few years.

```
# Make the value col in chem_ca numeric
chem_ca$Value <- str_replace(chem_ca$Value, ",", "")
chem_ca$Value <- as.numeric(chem_ca$Value)
```

```
Warning: NAs introduced by coercion
```

```r
# Filter by measure unit of LB
chem_ca_lb <- chem_ca |> filter(measure == "LB") |> filter(!is.na(Value), chem_name != "TOTAL") |> arrang


# Get the top 5 highest values per year
top5_CA <- chem_ca_lb |>
  group_by(Year) |>
  top_n(5, Value)

# Plotting
# Define set chemical colors:
chemical_colors <- c(
  "BIFENAZATE" = "#F8766D",
  "CAPTAN" = "#A3A500",
  "CYPRODINIL" = "#00BF7D",
  "DICHLOROPROPENE" = "#00B0F6",
  "MALATHION" = "#E76BF3",
  "METAM-POTASSIUM" = "#FF61C3",
  "METAM-SODIUM" = "#C70039",
  "NEEM OIL, CLAR. HYD." = "#FF9DA7",
  "PENDIMETHALIN" = "#6A3D9A",
  "SULFUR" = "#FFC107",
  "THIRAM" = "#1F618D"
)

top5_CA_no_imp <- ggplot(top5_CA, aes(x = as.factor(Year), y = Value, fill = chem_name)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = chemical_colors)+
  labs(title = "Top 5 Chemicals by Pounds(LBS) per Year in CA",
       x = "Year",
       y = "LBS",
       fill = "Chemical Name") +
  theme_minimal()

top5_CA_no_imp
```




```
# Based on the results, I want to start my investigation with the chemical, sulfur.
```

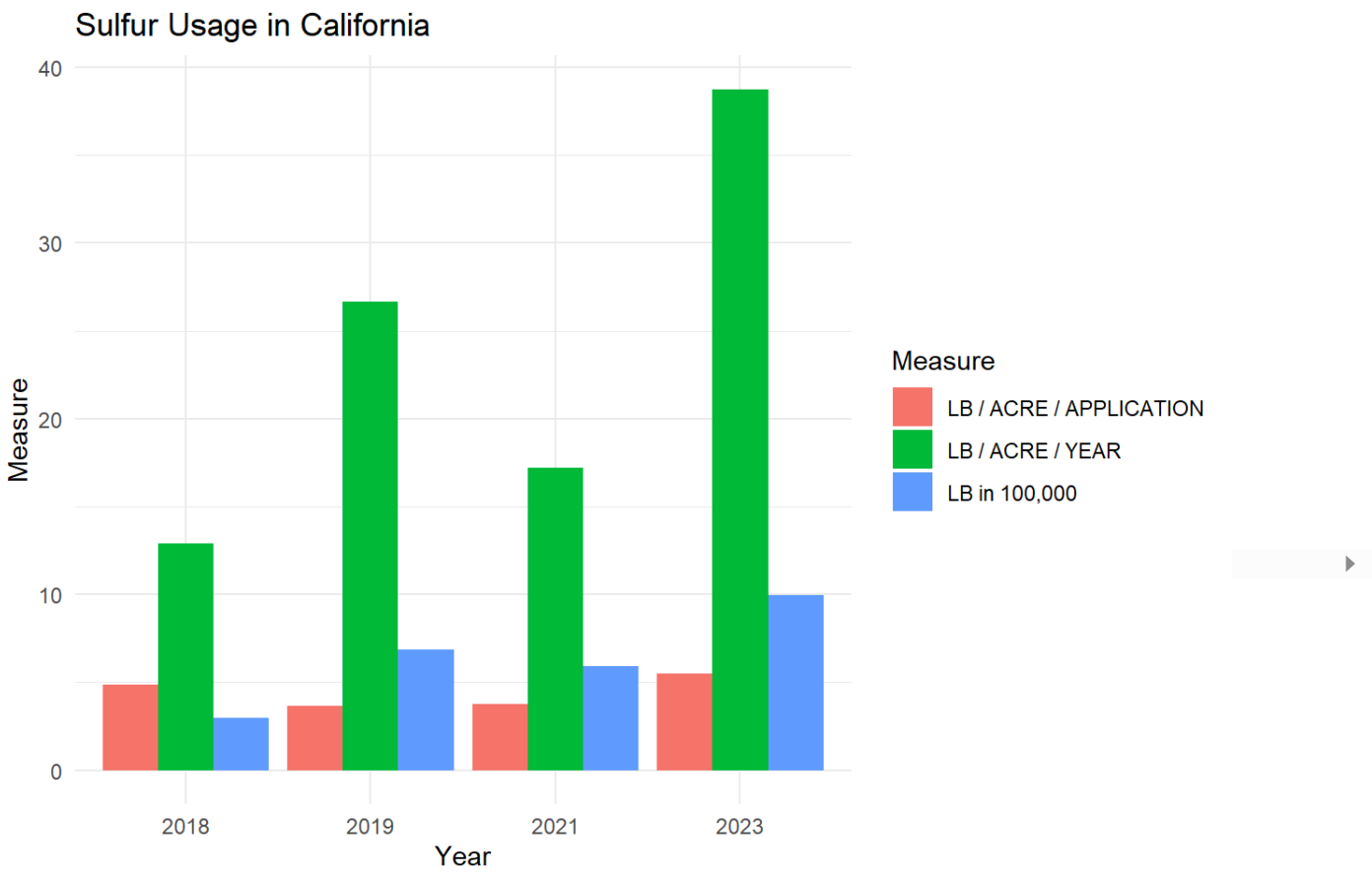
Let's get a dataframe containing all rows with chem_name == "SULFUR".

```
chem_ca_sulfur <- chem_ca |> filter(chem_ca$chem_name == "SULFUR")
```

Looking at the graph, above, I noticed sulfur was missing in 2023. Looking at the dataframe chem_ca_sulfur, I see that in 2023, sulfur contained values for other units such as lb/acre/year and lb/acre/application. It is missing for unit of lb because I took out the NA values. Therefore, I have to impute the lb value for sulfur usage in 2023.

```
#Create proportion df of lb/(lb/acre/app + lb/acre/yr) for sulfur for 2021, 2019 2018
proportion_ca_sulfur <- chem_ca_sulfur |>
  filter(Year %in% c(2021, 2019, 2018)) |>
  group_by(Year) %>%
  summarize(
    LB = sum(Value[measure == "LB"]),
    LB_Acre_Application = sum(Value[measure == "LB / ACRE / APPLICATION"]),
    LB_Acre_Year = sum(Value[measure == "LB / ACRE / YEAR"]),
    Proportion = LB / (LB_Acre_Application + LB_Acre_Year)
  )

#calculate average proportion
average_proportion_factor <- mean(proportion_ca_sulfur$Proportion, na.rm = TRUE)

sum_application_year_2023 <- sum(chem_ca_sulfur$Value[chem_ca_sulfur$Year == 2023 & chem_ca_sulfur$measur
```

```r
#impute value for 2023
chem_ca_sulfur <- chem_ca_sulfur |>
  mutate(Value = ifelse(Year == 2023 & measure == "LB" & is.na(Value),
                        round(sum_application_year_2023 * average_proportion_factor), Value))

#regraph with just sulfur
#scale lb by dividing by 100,000
chem_ca_sulfur <- chem_ca_sulfur |>
  mutate(Value = ifelse(measure == "LB" & !is.na(Value), Value / 100000, Value))

chem_ca_sulfur <- chem_ca_sulfur |>
  mutate(measure = ifelse(measure == "LB", "LB in 100,000", measure))

chem_ca_sulfur <- chem_ca_sulfur |> filter(measure != "NUMBER")

ggplot(chem_ca_sulfur, aes(x = as.factor(Year), y = Value, fill = measure)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Sulfur Usage in California",
       x = "Year",
       y = "Measure",
       fill = "Measure") +
  theme_minimal()
```



EDA of Sulfur Usage in California: The relationship between LB and LB/ACRE/YEAR seems consistent for 2018, 2019, and 2021. In 2021, even though LB was about 100,000 less than 2019, the farms that used sulfur that year the ratio

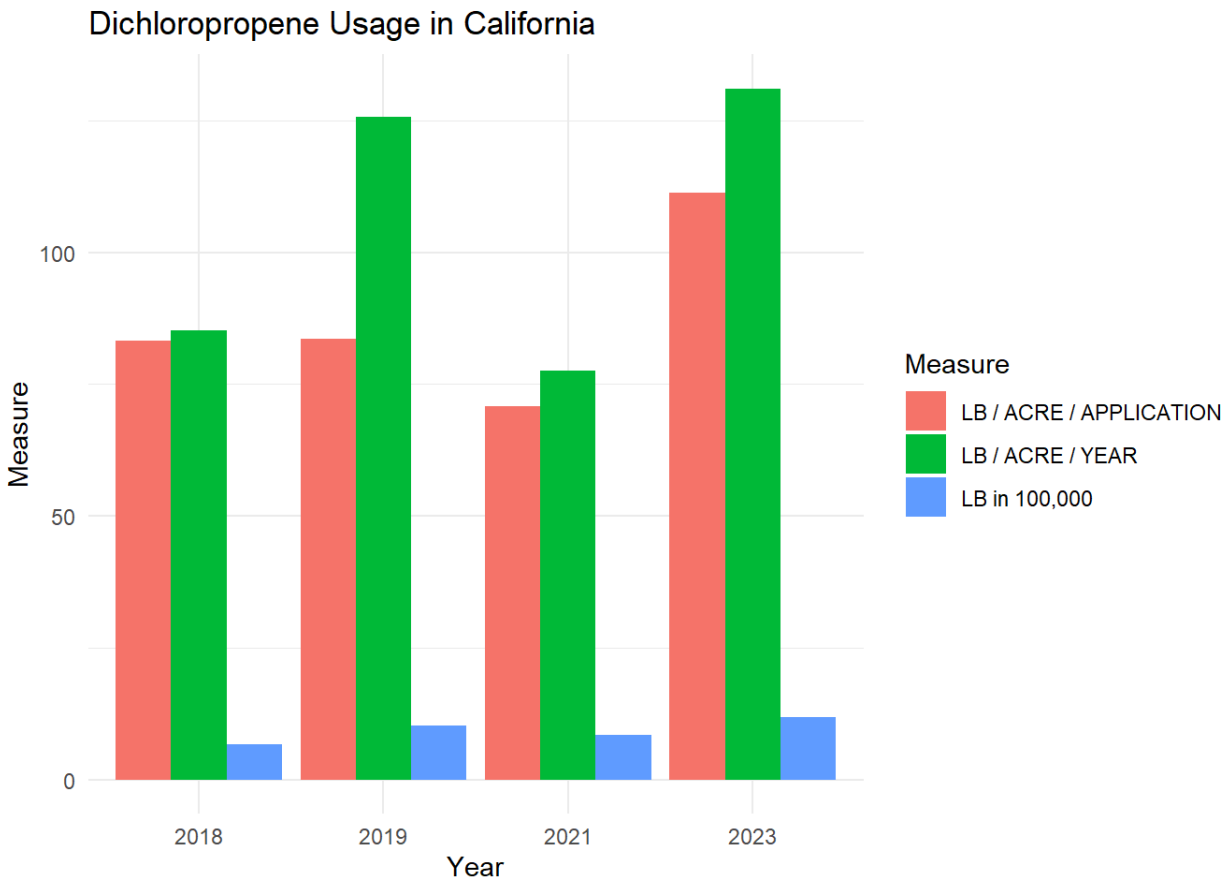
of lb/acre to lb is greater than the same ratio for 2019. It makes me wonder if the concentration might have been increased during the usage.

Conclusion of Sulfur: based on what I have seen with sulfur, I am incline to believe that other chemicals that didn't show up in the top 5 for all the years but showed up in some of the years might be missing and needs imputation. Let's continue with Dichloropropene.

```
chem_ca_dichlo <- chem_ca |> filter(chem_ca$chem_name == "DICHLOROPROPENE")
```

```
#impute the missing value for LB in 2023 and 2019 for dichloropropene using the same method we used
#for sulfur.

#create proportion df of lb/(lb/acre/app + lb/acre/yr) for dichloropropene for 2021, 2018
proportion_ca_dichlo <- chem_ca_dichlo |>
  filter(Year %in% c(2021, 2018)) |>
  group_by(Year) %>%
  summarize(
    LB = sum(Value[measure == "LB"]),
    LB_Acre_Application = sum(Value[measure == "LB / ACRE / APPLICATION"]),
    LB_Acre_Year = sum(Value[measure == "LB / ACRE / YEAR"]),
    Proportion = LB / (LB_Acre_Application + LB_Acre_Year)
  )

#calculate average proportion
average_proportion_factor <- mean(proportion_ca_dichlo$Proportion, na.rm = TRUE)

sum_application_year_2023 <- sum(chem_ca_dichlo$Value[chem_ca_dichlo$Year == 2023 & chem_ca_dichlo$measur
sum_application_year_2019 <- sum(chem_ca_dichlo$Value[chem_ca_dichlo$Year == 2019 & chem_ca_dichlo$measur


#impute value for 2023 & 2019
chem_ca_dichlo <- chem_ca_dichlo |>
  mutate(Value = ifelse(Year == 2023 & measure == "LB" & is.na(Value),
                        round(sum_application_year_2023 * average_proportion_factor), Value))
chem_ca_dichlo <- chem_ca_dichlo |>
  mutate(Value = ifelse(Year == 2019 & measure == "LB" & is.na(Value),
                        round(sum_application_year_2019 * average_proportion_factor), Value))


#regraph with just dichlo
#scale lb by dividing by 100,000
chem_ca_dichlo <- chem_ca_dichlo |>
  mutate(Value = ifelse(measure == "LB" & !is.na(Value), Value / 100000, Value))

chem_ca_dichlo <- chem_ca_dichlo |>
  mutate(measure = ifelse(measure == "LB", "LB in 100,000", measure))

chem_ca_dichlo <- chem_ca_dichlo |> filter(measure != "NUMBER")

ggplot(chem_ca_dichlo, aes(x = as.factor(Year), y = Value, fill = measure)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Dichloropropene Usage in California",
       x = "Year",
```

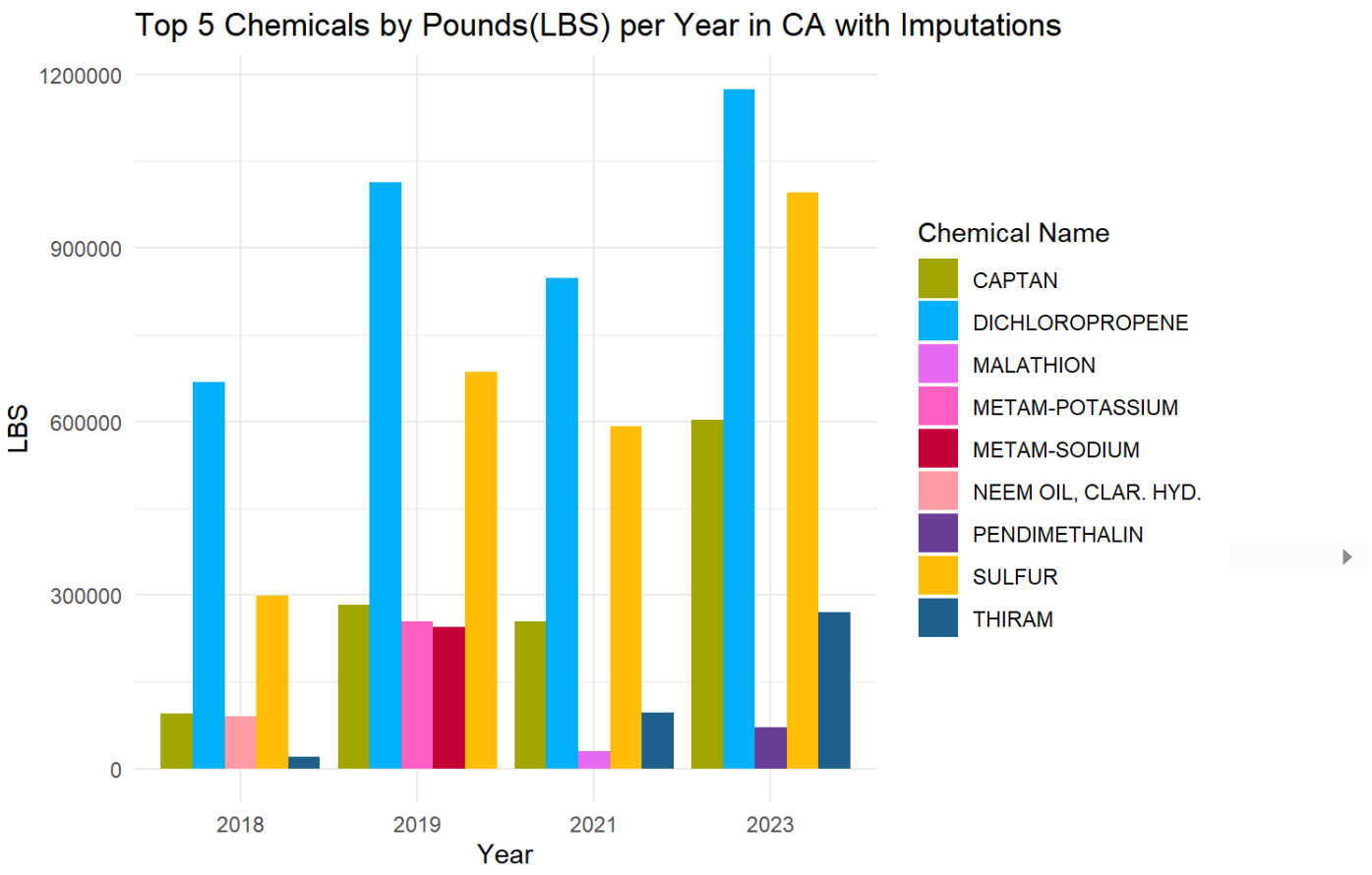
```
      y = "Measure",
      fill = "Measure") +
  theme_minimal()
```



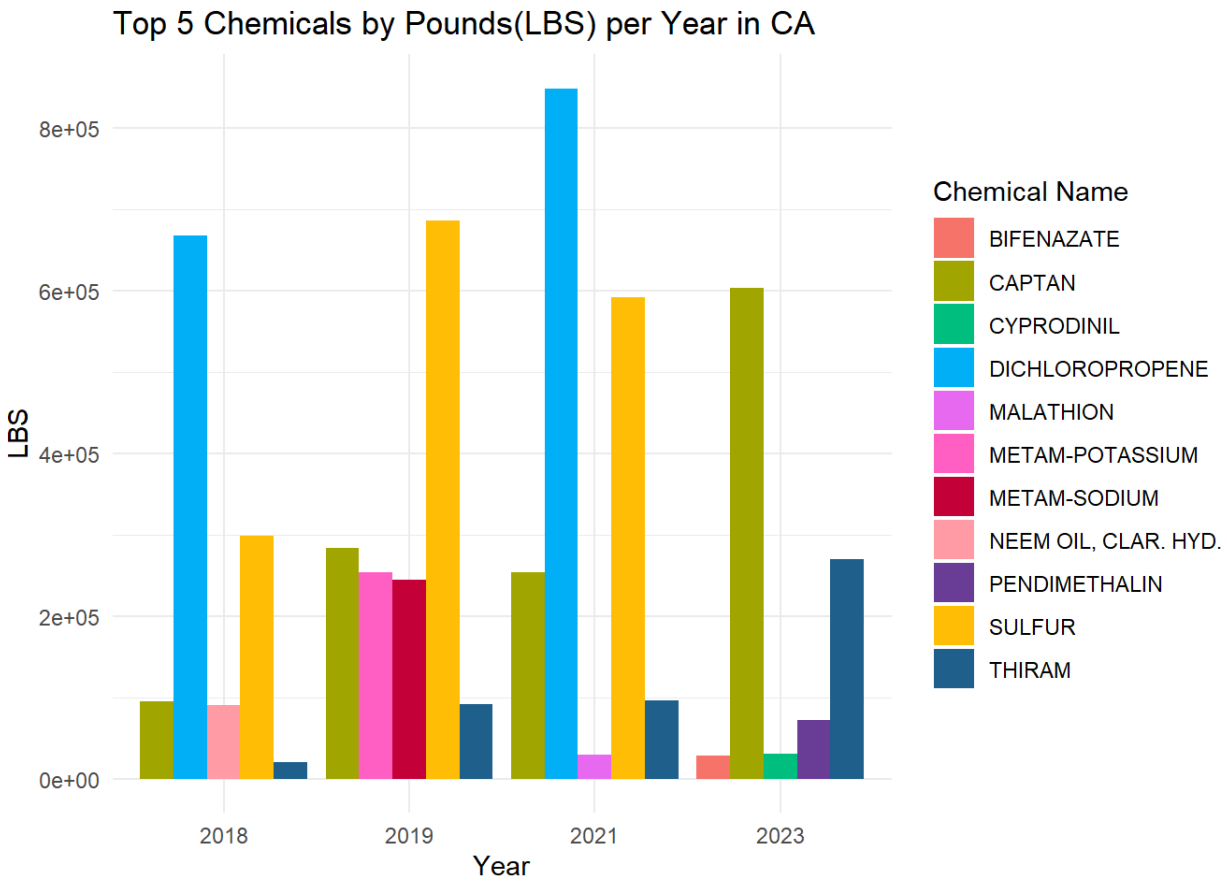
Conclusion: I'm not sure why the lb/acre/year increased so much between 2018 and 2019 while the lb/acre/application remained fairly the same...It is important to note that the imputation is and estimate that comes with a standard error. It provides a predicted value but there are uncertainty with imputing missing data. Let's take a look at how the the top 5 chemicals usage in CA appears with the imputations for missing values of sulfur and dichloropropene.

```
top5_CA_imp <- top5_CA |> filter(chem_name != "SULFUR", chem_name != "DICHLOROPROPENE")
top5_CA_imp <- rbind.data.frame(top5_CA_imp,chem_ca_sulfur,chem_ca_dichlo)

# Let's mutate the units for LB back to LB instead of in 100,000 LB
top5_CA_imp$measure[top5_CA_imp$measure == "LB in 100,000"] <- "LB"
top5_CA_imp$Value[top5_CA_imp$measure == "LB" & top5_CA_imp$chem_name == "SULFUR"] <- top5_CA_imp$Value[t
top5_CA_imp$Value[top5_CA_imp$measure == "LB" & top5_CA_imp$chem_name == "DICHLOROPROPENE"] <- top5_CA_im

# Let's regraph with the imputed values for sulfur and dichloropropene

# Get the top 5 highest values per year (considering imputations)
top5_CA_imp <- top5_CA_imp |>
  group_by(Year) |>
  top_n(5, Value)

# Plotting
```

```
chemical_colors <- c(
  "BIFENAZATE" = "#F8766D",
  "CAPTAN" = "#A3A500",
  "CYPRODINIL" = "#00BF7D",
  "DICHLOROPROPENE" = "#00B0F6",
  "MALATHION" = "#E76BF3",
  "METAM-POTASSIUM" = "#FF61C3",
  "METAM-SODIUM" = "#C70039",
  "NEEM OIL, CLAR. HYD." = "#FF9DA7",
  "PENDIMETHALIN" = "#6A3D9A",
  "SULFUR" = "#FFC107",
  "THIRAM" = "#1F618D"
)

ggplot(top5_CA_imp, aes(x = as.factor(Year), y = Value, fill = chem_name)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = chemical_colors)+
  labs(title = "Top 5 Chemicals by Pounds(LBS) per Year in CA with Imputations",
       x = "Year",
       y = "LBS",
       fill = "Chemical Name") +
  theme_minimal()
```




```
top5_CA_no_imp
```



Conclusion: I could go back and find other top 5 CA chemicals that might need imputations but just looking at the usage graphs, it's clear the the most used chemicals are dichloropropene and sulfur for CA, with some margin of error.

Let's look at how many of the top 5 chemicals in CA contains hazards that are "fatal". I skipped Dichloropropene and Neem Oil because I couldn't readily obtain it from the GHS Search. So I will only do the other 9 chemicals that showed up in the top 5 for CA for year 2018, 2019, 2021, 2023.

Int-Enviroguard (3)

```
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 Classification (UNEC
      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(hierarchy[["Node"]
    output_list[i]<-hierarchy[["Node"]][[i]][["Information"]][["Name"]]
    i<-i+1
  }
  return(output_list[!is.na(output_list)])
}
```

```
chemical_vec<-c("SULFUR", "CAPTAN", "THIRAM", "METAM-POTASSIUM", "METAM-SODIUM", "MALATHION", "PENDIMETHA

hazard_func <- function(chem_name){
  result<-get_pug_rest(identifier = chem_name, namespace = "name", domain = "compound",operation="classif
  return(hazards_retriever(GHS_searcher(result),result))
}
```

```
chem_hazard <- sapply(chemical_vec,hazard_func)
```

I am going to use str_detect to search for the word “Fatal” in any of the chemicals in the chem_hazard vector. This will return a vector of true or false for each chemical in chem_hazard. I will apply a sum for each chemical to see that if it returns 1, that means there was a TRUE detection of “Fatal”.

```
sulfur_fatsum <- sum(str_detect(chem_hazard$SULFUR,"Fatal"))
captan_fatsum <- sum(str_detect(chem_hazard$CAPTAN, "Fatal"))
thiram_fatsum <- sum(str_detect(chem_hazard$THIRAM, "Fatal"))
meta_potas_fatsum <- sum(str_detect(chem_hazard$`METAM-POTASSIUM`, "Fatal"))
meta_sod_fatsum <- sum(str_detect(chem_hazard$`METAM-SODIUM`, "Fatal"))
malathion_fatsum <- sum(str_detect(chem_hazard$MALATHION, "Fatal"))
pend_fatsum <- sum(str_detect(chem_hazard$PENDIMETHALIN, "Fatal"))
cyprodinil_fatsum <- sum(str_detect(chem_hazard$CYPRODINIL, "Fatal"))
bifenazate_fatsum <-sum(str_detect(chem_hazard$BIFENAZATE, "Fatal"))

sulfur_fatsum
```

```
[1] 0
```

```
captan_fatsum
```

```
[1] 1
```

```
thiram_fatsum
```

```
[1] 1
```

```
meta_potas_fatsum
```

```
[1] 0
```

```
meta_sod_fatsum
```

```
[1] 0
```

```
malathion_fatsum
```

```
[1] 0
```

```
pend_fatsum
```

```
[1] 0
```

```
cyprodinil_fatsum
```

```
[1] 0
```

```
bifenazate_fatsum
```

```
[1] 0
```

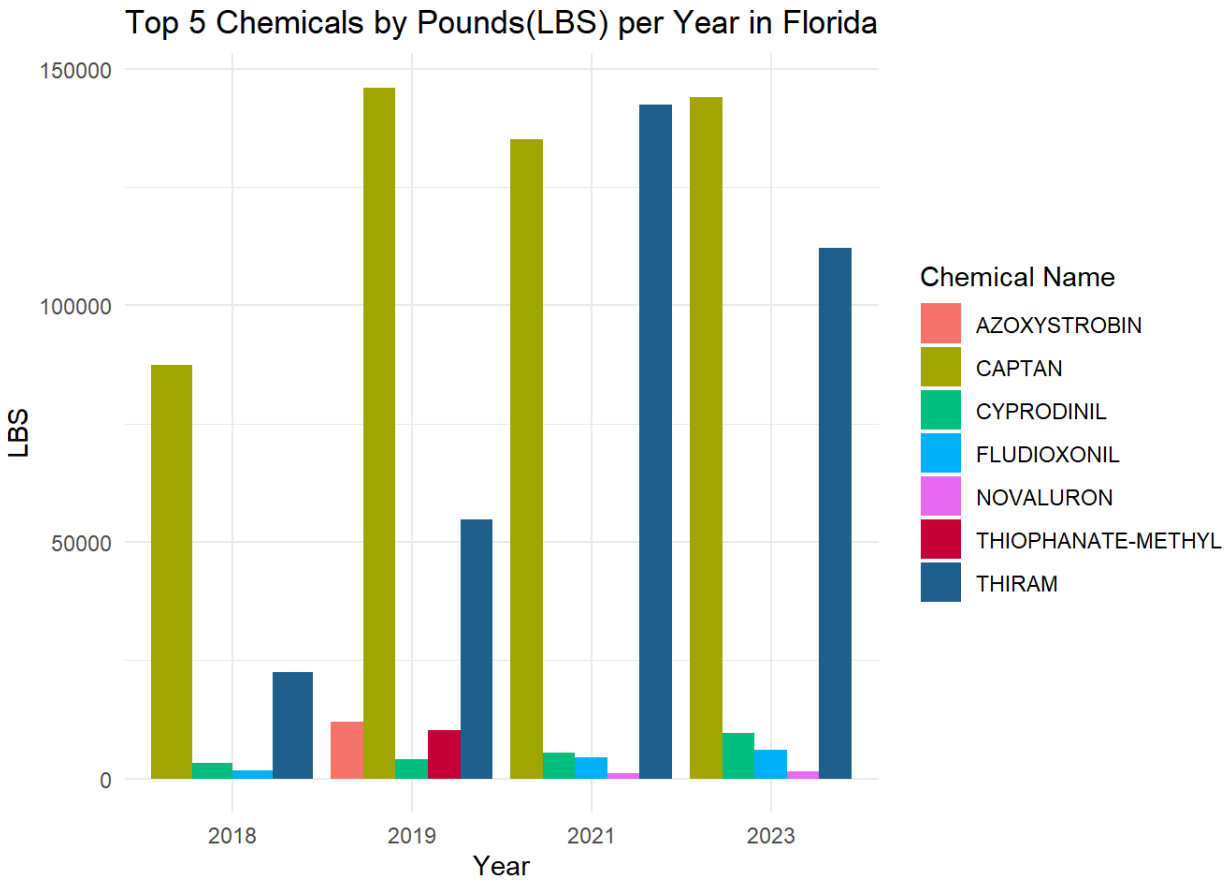
```
#We see that 2 of the 9 chemicals contain fatal hazards: Captan and Thiram
```

Let's take a look at Florida's top 5 chemicals by LB per year.

```
#make the value col in chem_fl numeric
chem_fl$Value <- str_replace(chem_fl$Value, ",", "")
chem_fl$Value <- as.numeric(chem_fl$Value)
```

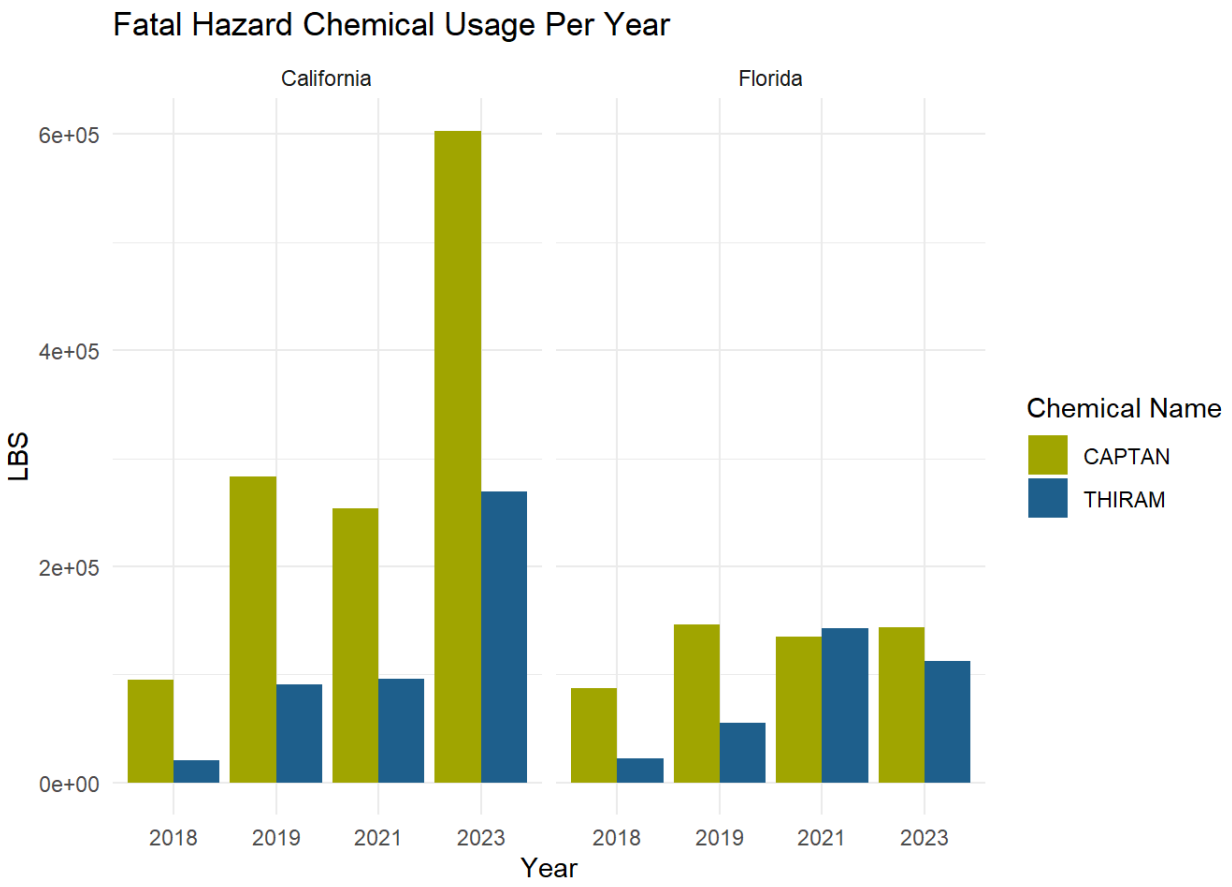
```
Warning: NAs introduced by coercion
```

```
#filter by measure unit of LB
chem_fl_lb <- chem_fl |> filter(measure == "LB") |> filter(!is.na(Value), chem_name != "TOTAL") |> arrang


# Get the top 5 highest values per year
top5_FL <- chem_fl_lb |>
  group_by(Year) |>
  top_n(5, Value)

# Plotting
chemical_colorsfl <- c(
  "AZOXYSTROBIN" = "#F8766D",
  "CAPTAN" = "#A3A500",
  "CYPRODINIL" = "#00BF7D",
  "FLUDIOXONIL" = "#00B0F6",
  "NOVALURON" = "#E76BF3",
  "THIOPHANATE-METHYL" = "#C70039",
  "THIRAM" = "#1F618D"
)
ggplot(top5_FL, aes(x = as.factor(Year), y = Value, fill = chem_name)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = chemical_colorsfl)+
  labs(title = "Top 5 Chemicals by Pounds(LBS) per Year in Florida",
       x = "Year",
       y = "LBS",
       fill = "Chemical Name") +
  theme_minimal()
```



Conclusion: Florida seems to really only have a “top 2” instead of “top 5”. And the top 2 happen to be Captan and Thiram, the chemicals that contain “fatal” hazards in the California top 5!!!! Let’s plot the two chemicals values for florida and california to compare.

```
top5_CA$state <- c("California")
top5_FL$state <- c("Florida")
fatal_chem <- rbind(top5_CA,top5_FL)
fatal_chem <- subset(fatal_chem, chem_name %in% c("CAPTAN", "THIRAM"))

chemical_colors_fatal <- c(
  "CAPTAN" = "#A3A500",
  "THIRAM" = "#1F618D"
)
ggplot(fatal_chem, aes(x = as.factor(Year), y = Value, fill = chem_name)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = chemical_colors_fatal)+
  labs(title = "Fatal Hazard Chemical Usage Per Year",
       x = "Year",
       y = "LBS",
       fill = "Chemical Name") +
  theme_minimal()+
  facet_wrap(~state)
```



Florida seems to use a more close proportion of Captan and Thiram whereas California appears to prefer Captan.. I wonder why that is..

EDA Conclusion: Through this exercise, I've learned that the purpose of EDA is to show you what story the data can tell and what story it doesn't show clearly. However, if the variety of the data is there, you can use the info to make the best guess, but be aware that is is uncertainty. My biggest shock reveal of this project is finding out that the top chemical usage in the top states in the U.S contain fatal hazard chemicals. It makes me wonder about the regulation put in place to be able to use these chemicals and how strict is the application being monitor. I can see that the strawberry business is a big business in the U.S and perhaps that's because it's one of the favorite fruits, that creates a supply and demand situation. I wonder for the less favorable fruits if there are as many chemicals used on them....

Citation for images:

1. Fruit Growers News. (2023, June 29). Survey: California strawberry acreage increases lead to optimism. Fruit Growers News. Retrieved from https://fruitgrowersnews.com/news/survey-california-strawberry-acreage-increases-lead-to-optimism/

2. A Path to Health. (2013, May 29). Strawberries: Conventional vs. organic. A Path to Health. Retrieved from https://apathtohealth.wordpress.com/2013/05/29/strawberries-conventional-vs-organic/

3. Int-Enviroguard. (n.d.). What is the GHS? Understanding the Globally Harmonized System of Classification and Labeling of Chemicals. Int-Enviroguard. Retrieved from https://int-enviroguard.com/blog/what-is-the-ghs/