Thorp Midterm Project

AUTHOR
Aidan Thorp

Preparing Strawberry Data for Analysis

My Assignment is to clean, organize, and explore the strawberry data set. Then turn in a report that describes how my work has set the stage for further analysis and model building.

The Data:

I am already given that the data set contains strawberry farming data with details about conventional and organic cultivation. These data include information about chemicals used in strawberry farming, as well as sales, revenue and expense details.

First I will load all the necessary libraries needed to complete this assignment:

```
# Load necessary libraries
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
library(tidyr)
library(stringr)
```

Next I extracted the strawb_mar6.csv file and moved it to strawb_data. I also wanted to have a quick look at what is inside strawb_data so I did some early exploring of the data set.

```
# Load the dataset
strawb_data <- read.csv("strawb_mar6.csv")

# View the first few rows
head(strawb_data)
```

```
  Program Year Period Week.Ending Geo.Level      State State.ANSI Ag.District
1  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
2  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
3  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
4  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
5  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
6  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
  Ag.District.Code County County.ANSI Zip.Code Region watershed_code Watershed
1               NA     NA          NA       NA     NA              0        NA
2               NA     NA          NA       NA     NA              0        NA
3               NA     NA          NA       NA     NA              0        NA
4               NA     NA          NA       NA     NA              0        NA
5               NA     NA          NA       NA     NA              0        NA
6               NA     NA          NA       NA     NA              0        NA
             Commodity
1 INCOME, NET CASH FARM
2 INCOME, NET CASH FARM
3 INCOME, NET CASH FARM
4 INCOME, NET CASH FARM
5 INCOME, NET CASH FARM
6 INCOME, NET CASH FARM
                                                 Data.Item   Domain
1 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
2 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
3 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
4 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
5 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
6 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
                Domain.Category          Value CV....
1   NET GAIN: (1,000 TO 4,999 $)      6,312,000    9.2
2 NET GAIN: (10,000 TO 24,999 $)     55,328,000    8.0
3 NET GAIN: (25,000 TO 49,999 $)    100,618,000    4.9
4   NET GAIN: (5,000 TO 9,999 $)     13,709,000   13.8
5   NET GAIN: (50,000 OR MORE $) 15,979,024,000    4.7
6  NET GAIN: (LESS THAN 1,000 $)        361,000   15.7
```

```r
# Check column names and structure
str(strawb_data)
```

```
'data.frame':   3584 obs. of  21 variables:
 $ Program         : chr  "CENSUS" "CENSUS" "CENSUS" "CENSUS" ...
 $ Year            : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ...
 $ Period          : chr  "YEAR" "YEAR" "YEAR" "YEAR" ...
 $ Week.Ending     : chr  "" "" "" "" ...
 $ Geo.Level       : chr  "STATE" "STATE" "STATE" "STATE" ...
 $ State           : chr  "CALIFORNIA" "CALIFORNIA" "CALIFORNIA" "CALIFORNIA" ...
 $ State.ANSI      : int  6 6 6 6 6 6 6 6 6 6 ...
 $ Ag.District     : logi  NA NA NA NA NA NA ...
 $ Ag.District.Code: logi  NA NA NA NA NA NA ...
 $ County          : logi  NA NA NA NA NA NA ...
```

```
 $ County.ANSI     : logi  NA NA NA NA NA NA ...
 $ Zip.Code        : logi  NA NA NA NA NA NA ...
 $ Region          : logi  NA NA NA NA NA NA ...
 $ watershed_code  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Watershed       : logi  NA NA NA NA NA NA ...
 $ Commodity       : chr  "INCOME, NET CASH FARM" "INCOME, NET CASH FARM" "INCOME, NET
CASH FARM" "INCOME, NET CASH FARM" ...
 $ Data.Item       : chr  "INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $"
"INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $" "INCOME, NET CASH FARM, OF
OPERATIONS - GAIN, MEASURED IN $" "INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED
IN $" ...
 $ Domain          : chr  "NET GAIN" "NET GAIN" "NET GAIN" "NET GAIN" ...
 $ Domain.Category : chr  "NET GAIN: (1,000 TO 4,999 $)" "NET GAIN: (10,000 TO 24,999 $)"
"NET GAIN: (25,000 TO 49,999 $)" "NET GAIN: (5,000 TO 9,999 $)" ...
 $ Value           : chr  "6,312,000" "55,328,000" "100,618,000" "13,709,000" ...
 $ CV....          : chr  "9.2" "8.0" "4.9" "13.8" ...
```

```
# Summary statistics
summary(strawb_data)
```

```
  Program               Year         Period          Week.Ending
 Length:3584        Min.   :2020   Length:3584        Length:3584
 Class :character   1st Qu.:2021   Class :character   Class :character
 Mode  :character   Median :2022   Mode  :character   Mode  :character
                    Mean   :2022
                    3rd Qu.:2023
                    Max.   :2024
  Geo.Level            State             State.ANSI     Ag.District
 Length:3584        Length:3584        Min.   : 6.00   Mode:logical
 Class :character   Class :character   1st Qu.: 6.00   NA's:3584
 Mode  :character   Mode  :character   Median :12.00
                                       Mean   :18.28
                                       3rd Qu.:25.00
                                       Max.   :50.00
 Ag.District.Code  County        County.ANSI    Zip.Code        Region
 Mode:logical     Mode:logical   Mode:logical   Mode:logical   Mode:logical
 NA's:3584        NA's:3584      NA's:3584      NA's:3584      NA's:3584




 watershed_code Watershed       Commodity          Data.Item
 Min.   :0      Mode:logical   Length:3584        Length:3584
 1st Qu.:0      NA's:3584      Class :character   Class :character
 Median :0                     Mode  :character   Mode  :character
 Mean   :0
 3rd Qu.:0
 Max.   :0
    Domain          Domain.Category       Value              CV....
 Length:3584        Length:3584        Length:3584        Length:3584
```

```
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Data Cleaning

Next I knew from the assignment description I would only need to focus on the states California and Florida, so I filtered strawb_data to only include data where the state is one of those two.

```
strawb_filtered <- strawb_data %>%
  filter(State %in% c("CALIFORNIA", "FLORIDA"))

head(strawb_filtered)
```

```
  Program Year Period Week.Ending Geo.Level      State State.ANSI Ag.District
1  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
2  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
3  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
4  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
5  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
6  CENSUS 2022   YEAR                STATE CALIFORNIA          6          NA
  Ag.District.Code County County.ANSI Zip.Code Region watershed_code Watershed
1               NA     NA          NA       NA     NA              0        NA
2               NA     NA          NA       NA     NA              0        NA
3               NA     NA          NA       NA     NA              0        NA
4               NA     NA          NA       NA     NA              0        NA
5               NA     NA          NA       NA     NA              0        NA
6               NA     NA          NA       NA     NA              0        NA
              Commodity
1 INCOME, NET CASH FARM
2 INCOME, NET CASH FARM
3 INCOME, NET CASH FARM
4 INCOME, NET CASH FARM
5 INCOME, NET CASH FARM
6 INCOME, NET CASH FARM
                                                  Data.Item   Domain
1 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
2 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
3 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
4 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
5 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
6 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
                Domain.Category          Value CV....
1   NET GAIN: (1,000 TO 4,999 $)      6,312,000    9.2
2 NET GAIN: (10,000 TO 24,999 $)     55,328,000    8.0
3 NET GAIN: (25,000 TO 49,999 $)    100,618,000    4.9
4   NET GAIN: (5,000 TO 9,999 $)     13,709,000   13.8
```

```
5   NET GAIN: (50,000 OR MORE $) 15,979,024,000    4.7
6  NET GAIN: (LESS THAN 1,000 $)        361,000   15.7
```

I also wanted to get rid of any unnecessary columns with nothing in them so after looking at the initial heading of the data, I noticed that these columns seemed to have nothing in them. I wanted to check before blindly removing them so I looked to see all unique values in each of these columns. If my suspicions were correct I would remove the column. If not, I would keep the column.

```
# Define the columns to check
cols_to_check <- c("Week.Ending", "Ag.District", "Ag.District.Code", "County",
                   "County.ANSI", "Zip.Code", "Region", "watershed_code", "Waters

# Loop through the columns and print unique values
for (col in cols_to_check) {
  cat("Unique values in", col, ":\n")
  print(unique(strawb_filtered[[col]]))
  cat("\n----------------------\n")
}
```

```
Unique values in Week.Ending :
[1] ""

----------------------
Unique values in Ag.District :
[1] NA

----------------------
Unique values in Ag.District.Code :
[1] NA

----------------------
Unique values in County :
[1] NA

----------------------
Unique values in County.ANSI :
[1] NA

----------------------
Unique values in Zip.Code :
[1] NA

----------------------
Unique values in Region :
[1] NA

----------------------
Unique values in watershed_code :
[1] 0
```

```
----------------------
Unique values in Watershed :
[1] NA

----------------------
```

My suspicions were correct and all of the columns had nothing in them so I will drop all of these columns as they won't help me with my project.

```
strawb_filtered <- strawb_filtered %>%
  select(-c(Week.Ending, Ag.District, Ag.District.Code, County,
            County.ANSI, Zip.Code, Region, watershed_code, Watershed))

head(strawb_filtered)
```

```
  Program Year Period Geo.Level      State State.ANSI             Commodity
1  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
2  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
3  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
4  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
5  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
6  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
                                                 Data.Item   Domain
1 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
2 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
3 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
4 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
5 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
6 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
                 Domain.Category          Value CV....
1   NET GAIN: (1,000 TO 4,999 $)      6,312,000    9.2
2 NET GAIN: (10,000 TO 24,999 $)     55,328,000    8.0
3 NET GAIN: (25,000 TO 49,999 $)    100,618,000    4.9
4   NET GAIN: (5,000 TO 9,999 $)     13,709,000   13.8
5   NET GAIN: (50,000 OR MORE $) 15,979,024,000    4.7
6  NET GAIN: (LESS THAN 1,000 $)        361,000   15.7
```

I also noticed that Value was being kept as a character, which isn't helpful to me. Instead, I converted it to a numeric type, and if the item in Value wasn't a number, it became NA

```
strawb_filtered <- strawb_filtered %>%
  mutate(Value = suppressWarnings(as.numeric(parse_number(Value))))
```

Next I wanted to split the Strawberry data into two parts. One for Census data and Survey data.

```
# Split into two datasets
strawb_census <- strawb_filtered %>% filter(Program == "CENSUS")
strawb_survey <- strawb_filtered %>% filter(Program == "SURVEY")
```

After the original split I still wanted to clean some of the columns a little more. Specifically I wanted to split the Domain.Category up into 3 separate columns. They would be Chemical, Type, and Chemical Name. This way it would be easier to drill deeper into what is happining.

One other quick thing I noticed is that CV had nothing in it for survey data so I decided to drop it in this data set.

```
# Separate 'Domain' into 'Chemical' and 'Type'
strawb_survey <- strawb_survey %>%
  separate(Domain, into = c("Chemical", "Type"), sep = ", ", extra = "merge", fil

strawb_survey <- strawb_survey %>%
  mutate(`Chemical Name` = str_extract(Domain.Category, "\\((.*?)\\)")) %>%  # Ex
  mutate(`Chemical Name` = str_remove_all(`Chemical Name`, "[()]")) %>%  # Remove
  mutate(`Chemical Name` = str_remove(`Chemical Name`, " = \\d+$")) %>%  # Remove
  select(-Domain.Category)  # Drop the Domain.Category column

#Drop CV column from strawb_survey
strawb_survey <- select(strawb_survey, -CV....)
```

Now I wanted to split up my data a little more so it was better organized. I wanted 3 smaller data sets for each item in chemical. One would be for Total, the other two would be for Chemical and Fertilizer respectively.

```
# Step 2: Create filtered datasets
strawb_survey_total <- strawb_survey %>% filter(Chemical == "TOTAL")
strawb_survey_chem  <- strawb_survey %>% filter(Chemical == "CHEMICAL")
strawb_survey_fert  <- strawb_survey %>% filter(Chemical == "FERTILIZER")

head(strawb_survey_chem)
```

```
  Program Year Period Geo.Level      State State.ANSI    Commodity
1  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
2  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
3  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
4  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
5  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
6  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
                                                          Data.Item
1                          STRAWBERRIES - APPLICATIONS, MEASURED IN LB
2                          STRAWBERRIES - APPLICATIONS, MEASURED IN LB
3                          STRAWBERRIES - APPLICATIONS, MEASURED IN LB
4                          STRAWBERRIES - APPLICATIONS, MEASURED IN LB
5 STRAWBERRIES - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG
6 STRAWBERRIES - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG
  Chemical        Type Value                    Chemical Name
1 CHEMICAL   FUNGICIDE    NA                  OXATHIAPIPROLIN
2 CHEMICAL INSECTICIDE    NA                   CYCLANILIPROLE
3 CHEMICAL INSECTICIDE    NA                       PERMETHRIN
```

```
4 CHEMICAL       OTHER    NA ISARIA FUMOSOROSEA STRAIN FE 9901
5 CHEMICAL   FUNGICIDE    NA                  OXATHIAPIPROLIN
6 CHEMICAL INSECTICIDE    NA                   CYCLANILIPROLE
```

```
head(strawb_survey_total)
```

```
  Program Year         Period Geo.Level      State State.ANSI    Commodity
1  SURVEY 2023 MARKETING YEAR     STATE CALIFORNIA          6 STRAWBERRIES
2  SURVEY 2023 MARKETING YEAR     STATE CALIFORNIA          6 STRAWBERRIES
3  SURVEY 2023 MARKETING YEAR     STATE CALIFORNIA          6 STRAWBERRIES
4  SURVEY 2023 MARKETING YEAR     STATE    FLORIDA         12 STRAWBERRIES
5  SURVEY 2023 MARKETING YEAR     STATE    FLORIDA         12 STRAWBERRIES
6  SURVEY 2023 MARKETING YEAR     STATE    FLORIDA         12 STRAWBERRIES
                                                    Data.Item Chemical
1              STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT    TOTAL
2 STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / CWT    TOTAL
3   STRAWBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN $ / CWT    TOTAL
4              STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT    TOTAL
5 STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / CWT    TOTAL
6   STRAWBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN $ / CWT    TOTAL
  Type Value Chemical Name
1 <NA>   121          <NA>
2 <NA>    NA          <NA>
3 <NA>    NA          <NA>
4 <NA>   147          <NA>
5 <NA>    NA          <NA>
6 <NA>    NA          <NA>
```

```
head(strawb_survey_fert)
```

```
  Program Year Period Geo.Level      State State.ANSI    Commodity
1  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
2  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
3  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
4  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
5  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
6  SURVEY 2023   YEAR     STATE CALIFORNIA          6 STRAWBERRIES
                                                                Data.Item
1                         STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB
2                         STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB
3                         STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB
4                         STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB
5 STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG
6 STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG
    Chemical Type  Value Chemical Name
1 FERTILIZER <NA> 393000      NITROGEN
2 FERTILIZER <NA> 216000     PHOSPHATE
3 FERTILIZER <NA> 393000        POTASH
4 FERTILIZER <NA>     NA        SULFUR
```

```
5 FERTILIZER <NA>     13      NITROGEN
6 FERTILIZER <NA>     10     PHOSPHATE
```

```
head(strawb_census)
```

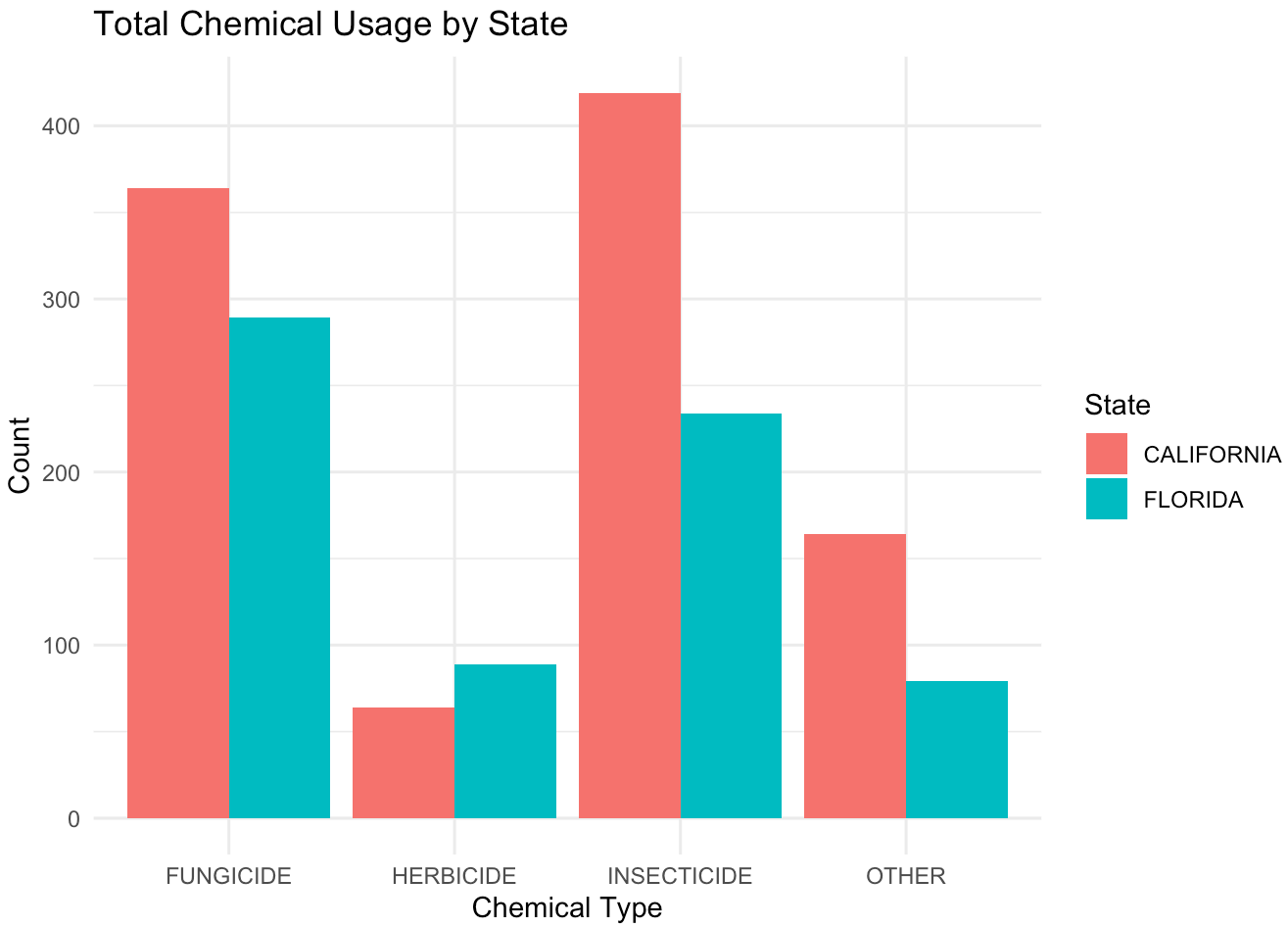
```
  Program Year Period Geo.Level      State State.ANSI             Commodity
1  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
2  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
3  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
4  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
5  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
6  CENSUS 2022   YEAR     STATE CALIFORNIA          6 INCOME, NET CASH FARM
                                                 Data.Item   Domain
1 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
2 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
3 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
4 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
5 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
6 INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $ NET GAIN
                Domain.Category       Value CV....
1   NET GAIN: (1,000 TO 4,999 $)     6312000    9.2
2 NET GAIN: (10,000 TO 24,999 $)    55328000    8.0
3 NET GAIN: (25,000 TO 49,999 $)   100618000    4.9
4   NET GAIN: (5,000 TO 9,999 $)    13709000   13.8
5   NET GAIN: (50,000 OR MORE $) 15979024000    4.7
6  NET GAIN: (LESS THAN 1,000 $)      361000   15.7
```

Visualizations

I was then satisfied that my data was cleaner and better organized I began to think about interesting visualizations that could be insightful. Something that seemed interesting to me was seeing what types of chemicals are used in each state.

I first made a bar graph with the total amount of each type of chemical while also comparing which states use more of each type of chemical

```
ggplot(strawb_survey_chem, aes(x = `Type`, fill = State)) +
  geom_bar(stat = "count", position = "dodge") +
  labs(title = "Total Chemical Usage by State", x = "Chemical Type", y = "Count") +
  theme_minimal()
```



There were several interesting things withing this bar chart. Specifically I found it interesting how many insecticides California uses. Another interesting thing I noticed was herbicides are the only chemical type that are used more in Florida.

Next I wanted to get a full list of all the chemicals in the data set and pick 3 to examine further

```
unique(strawb_survey_chem$`Chemical Name`)
```

```
 [1] "OXATHIAPIPROLIN"
 [2] "CYCLANILIPROLE"
 [3] "PERMETHRIN"
 [4] "ISARIA FUMOSOROSEA STRAIN FE 9901"
 [5] "AZOXYSTROBIN"
 [6] "BACILLUS AMYLOLIQUEFACIENS STRAIN D747"
 [7] "BACILLUS SUBTILIS"
 [8] "BLAD"
 [9] "BORAX DECAHYDRATE"
[10] "BOSCALID"
[11] "BT SUBSP KURSTAKI EVB-113-19"
[12] "CAPTAN"
[13] "CYFLUFENAMID"
[14] "CYPRODINIL"
[15] "DIFENOCONAZOLE"
[16] "FENHEXAMID"
```

```
 [17] "FLUDIOXONIL"
 [18] "FLUOPYRAM"
 [19] "FLUXAPYROXAD"
 [20] "FOSETYL-AL"
 [21] "ISOFETAMID"
 [22] "MEFENOXAM"
 [23] "MONO-POTASSIUM SALT"
 [24] "MYCLOBUTANIL"
 [25] "PENTHIOPYRAD"
 [26] "POLYOXIN D ZINC SALT"
 [27] "PROPICONAZOLE"
 [28] "PYDIFLUMETOFEN"
 [29] "PYRACLOSTROBIN"
 [30] "PYRIMETHANIL"
 [31] "QUINOLINE"
 [32] "SULFUR"
 [33] "TETRACONAZOLE"
 [34] "THIOPHANATE-METHYL"
 [35] "THIRAM"
 [36] "TOTAL"
 [37] "TRIFLOXYSTROBIN"
 [38] "TRIFLUMIZOLE"
 [39] "CARFENTRAZONE-ETHYL"
 [40] "FLUMIOXAZIN"
 [41] "OXYFLUORFEN"
 [42] "PENDIMETHALIN"
 [43] "ABAMECTIN"
 [44] "ACEQUINOCYL"
 [45] "ACETAMIPRID"
 [46] "AZADIRACHTIN"
 [47] "BEAUVERIA BASSIANA"
 [48] "BIFENAZATE"
 [49] "BIFENTHRIN"
 [50] "BT KURSTAK ABTS-1857"
 [51] "BT KURSTAKI ABTS-351"
 [52] "BT KURSTAKI SA-11"
 [53] "CANOLA OIL"
 [54] "CHLORANTRANILIPROLE"
 [55] "CHROMOBAC SUBTSUGAE PRAA4-1 CELLS AND SPENT MEDIA"
 [56] "CYANTRANILIPROLE"
 [57] "CYFLUMETOFEN"
 [58] "ETOXAZOLE"
 [59] "FENBUTATIN-OXIDE"
 [60] "FENPROPATHRIN"
 [61] "FENPYROXIMATE"
 [62] "FLONICAMID"
 [63] "FLUPYRADIFURONE"
 [64] "HEXYTHIAZOX"
 [65] "IMIDACLOPRID"
 [66] "LAMBDA-CYHALOTHRIN"
 [67] "MALATHION"
```

```
 [68] "METHOXYFENOZIDE"
 [69] "NALED"
 [70] "NEEM OIL"
 [71] "NEEM OIL, CLAR. HYD."
 [72] "NOVALURON"
 [73] "PIPERONYL BUTOXIDE"
 [74] "PYRETHRINS"
 [75] "PYRIDABEN"
 [76] "SPINETORAM"
 [77] "SPINOSAD"
 [78] "THIAMETHOXAM"
 [79] "ACIBENZOLAR-S-METHYL"
 [80] "CAPSICUM OLEORESIN EXTRACT"
 [81] "CHLOROPICRIN"
 [82] "DICHLOROPROPENE"
 [83] "FLUTRIAFOL"
 [84] "GARLIC OIL"
 [85] "HYDROGEN PEROXIDE"
 [86] "IRON PHOSPHATE"
 [87] "METALDEHYDE"
 [88] "METAM-POTASSIUM"
 [89] "METAM-SODIUM"
 [90] "PEROXYACETIC ACID"
 [91] "PSEUDOMONAS CHLORORAPHIS STRAIN AFS009"
 [92] "REYNOUTRIA SACHALINE"
 [93] "PYRIOFENONE"
 [94] "ZOXAMIDE"
 [95] "METSULFURON-METHYL"
 [96] "PENOXSULAM"
 [97] "S-METOLACHLOR"
 [98] "BETA-CYFLUTHRIN"
 [99] "ETHYL 2E;4Z"
[100] "OXAMYL"
[101] "CUPRAMMONIUM ACETATE"
[102] "DODECADIEN-1-OL"
[103] "FLUENSULFONE"
[104] "GIBBERELLIC ACID"
[105] "BACILLUS AMYLOLIQUEFAC F727"
[106] "CHLOROTHALONIL"
[107] "COPPER CHLORIDE HYD."
[108] "COPPER HYDROXIDE"
[109] "CYMOXANIL"
[110] "FAMOXADONE"
[111] "IPRODIONE"
[112] "MANCOZEB"
[113] "2,4-D, DIMETH. SALT"
[114] "CLETHODIM"
[115] "GLYPHOSATE ISO. SALT"
[116] "PARAQUAT"
[117] "DIAZINON"
[118] "METHOMYL"
```

```
 [119] "SULFOXAFLOR"
 [120] "CYTOKININS"
 [121] "INDOLEBUTYRIC ACID"
 [122] "BACILLUS AMYLOLIQUEFACIENS MBI 600"
 [123] "BACILLUS PUMILUS"
 [124] "COPPER OCTANOATE"
 [125] "POTASSIUM BICARBON."
 [126] "STREPTOMYCES LYDICUS"
 [127] "GLYPHOSATE POT. SALT"
 [128] "NAPROPAMIDE"
 [129] "BT KURSTAKI EG7841"
 [130] "BT SUB AIZAWAI GC-91"
 [131] "BUPROFEZIN"
 [132] "BURKHOLDERIA A396 CELLS & MEDIA"
 [133] "HELICOVERPA ZEA NPV"
 [134] "PETROLEUM DISTILLATE"
 [135] "POTASSIUM SALTS"
 [136] "PYRIPROXYFEN"
 [137] "SPIROMESIFEN"
 [138] "CAPRIC ACID"
 [139] "CAPRYLIC ACID"
 [140] "MINERAL OIL"
 [141] "PAECILOMYCES FUMOSOR"
 [142] "POTASSIUM SILICATE"
 [143] "COPPER ETHANOLAMINE"
 [144] "DIMETHENAMID"
 [145] "FLUROXYPYR 1-MHE"
 [146] "HALOSULFURON-METHYL"
 [147] "KANTOR"
 [148] "CARBARYL"
 [149] "FENAZAQUIN"
 [150] "ETHEPHON"
```

After some Debugging I realized that Year was easier to use if it was a factor and not a number so I quickly changed it here.

```
# Make sure 'Year' is treated as a factor for better x-axis handling
strawb_census$Year <- as.factor(strawb_census$Year)
```

Bar chart

I decided to pick Sulfur, Thiram, and Potash as my three chemicals to examine further.

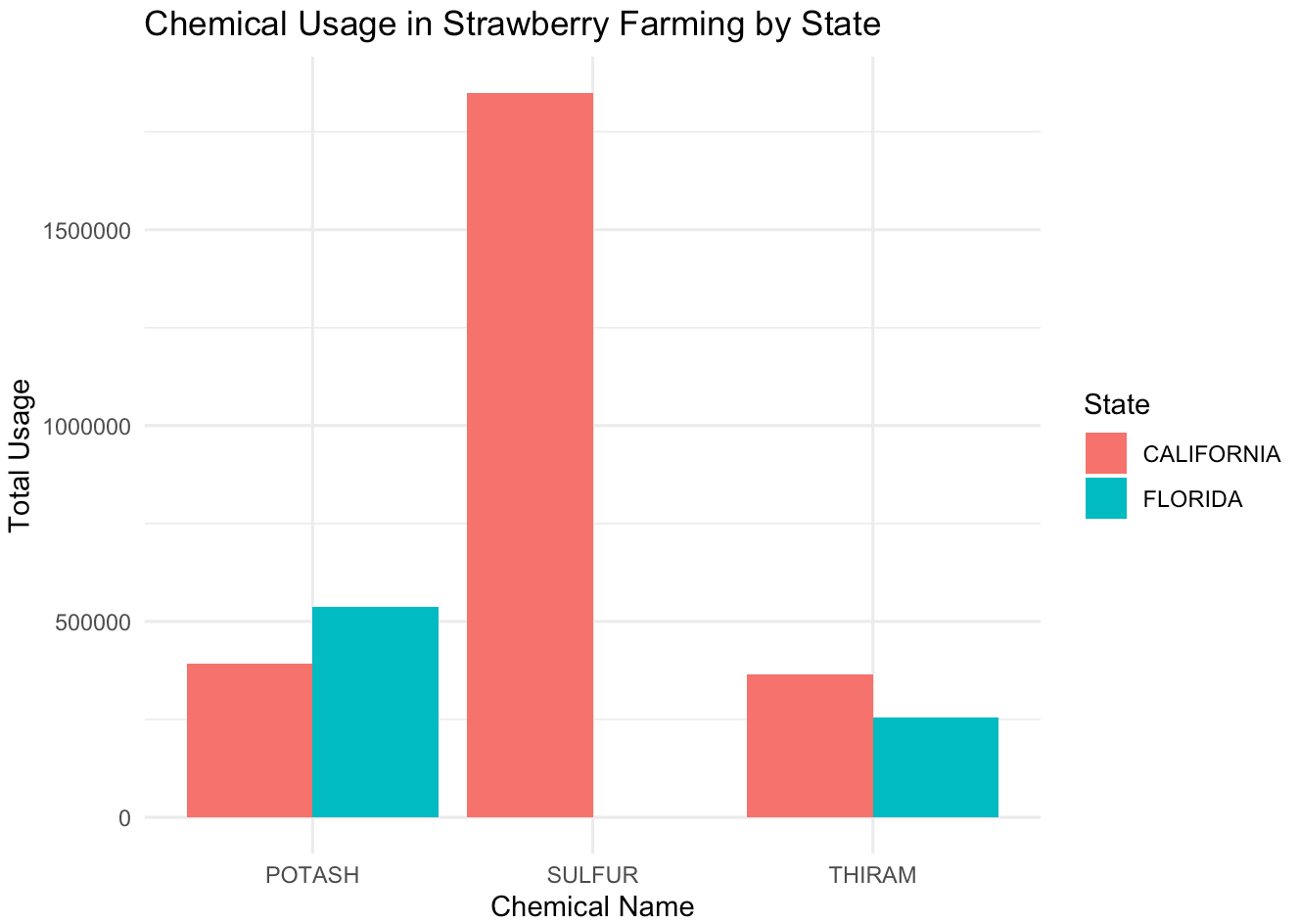
```
# Define the chemicals of interest
chemicals_of_interest <- c("SULFUR", "THIRAM", "POTASH")

# Filter the data for these chemicals
filtered_data <- strawb_survey %>%
  filter(`Chemical Name` %in% chemicals_of_interest)
```

```
# Aggregate data to see usage patterns for each state and chemical
usage_summary <- filtered_data %>%
  group_by(State, `Chemical Name`) %>%
  summarise(Total_Usage = sum(Value, na.rm = TRUE))
```

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

```
# Plot the data
ggplot(usage_summary, aes(x = `Chemical Name`, y = Total_Usage, fill = State)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Chemical Usage in Strawberry Farming by State",
       x = "Chemical Name",
       y = "Total Usage",
       fill = "State") +
  theme_minimal()
```



Interestingly these chemicals are used in large volumes in each of the states. Potash seems like it is used more often in Florida and less in California, but the most interesting part of this graph was how much sulfur is used in California strawberries and how it isn't used in Florida strawberries at all.

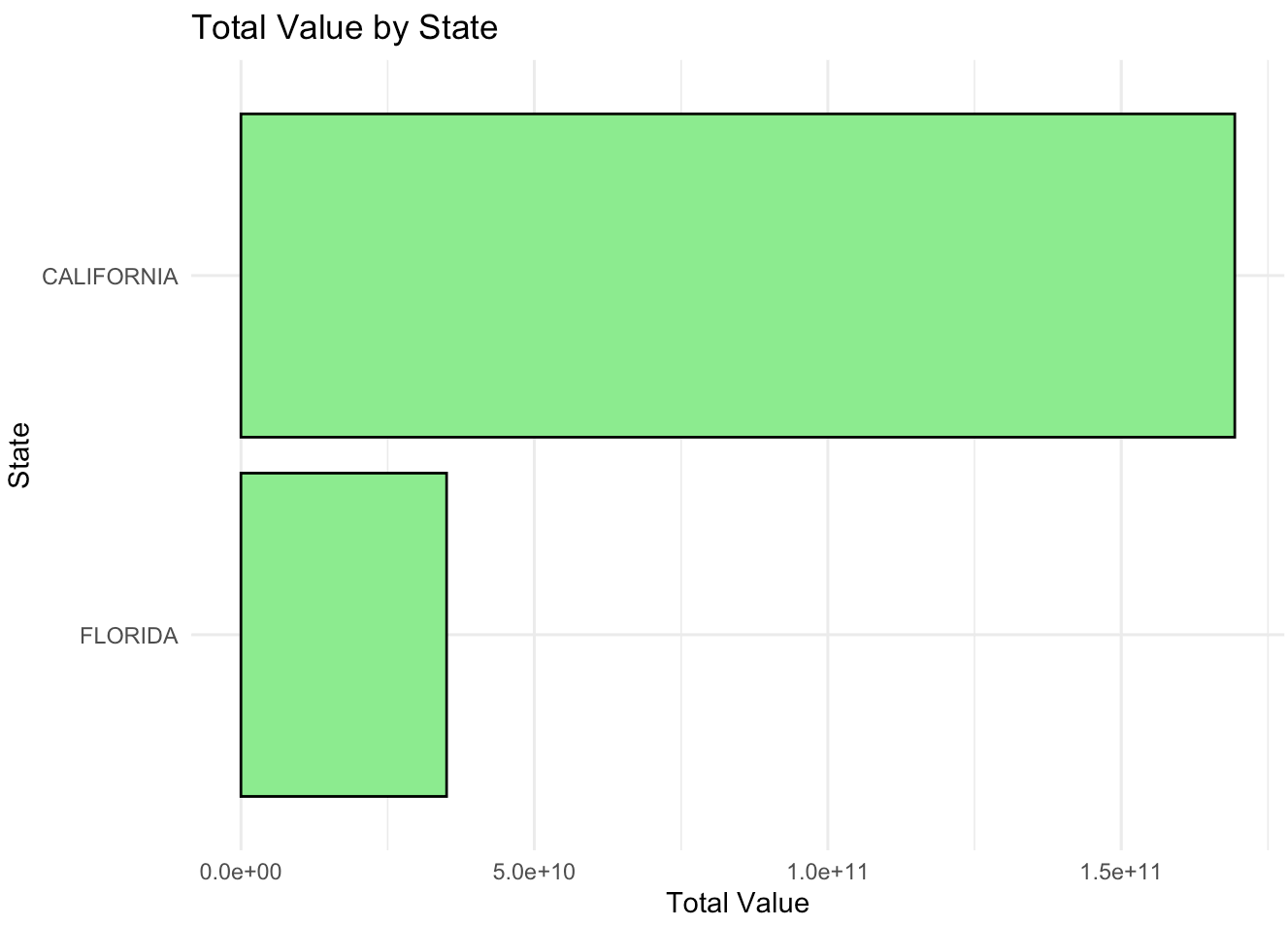
This could be interesting to look at in a future project to see why these products are used so often/little in each state and examine further how these chemicals help/hurt things about the strawberries such as sales and production of them.

Bar chart

Next I wanted to see the total value coming out of each state as it would give me a better perspective of who's producing more strawberries.

```
# Aggregate total value by state
state_value_summary <- strawb_census %>%
  group_by(State) %>%
  summarise(TotalValue = sum(Value, na.rm = TRUE))

# Create a bar plot of total value by state
ggplot(state_value_summary, aes(x = reorder(State, TotalValue), y = TotalValue)) +
  geom_bar(stat = "identity", fill = "lightgreen", color = "black") +
  labs(title = "Total Value by State",
       x = "State",
       y = "Total Value") +
  theme_minimal() +
  coord_flip()  # Flip coordinates to make it easier to read state names
```



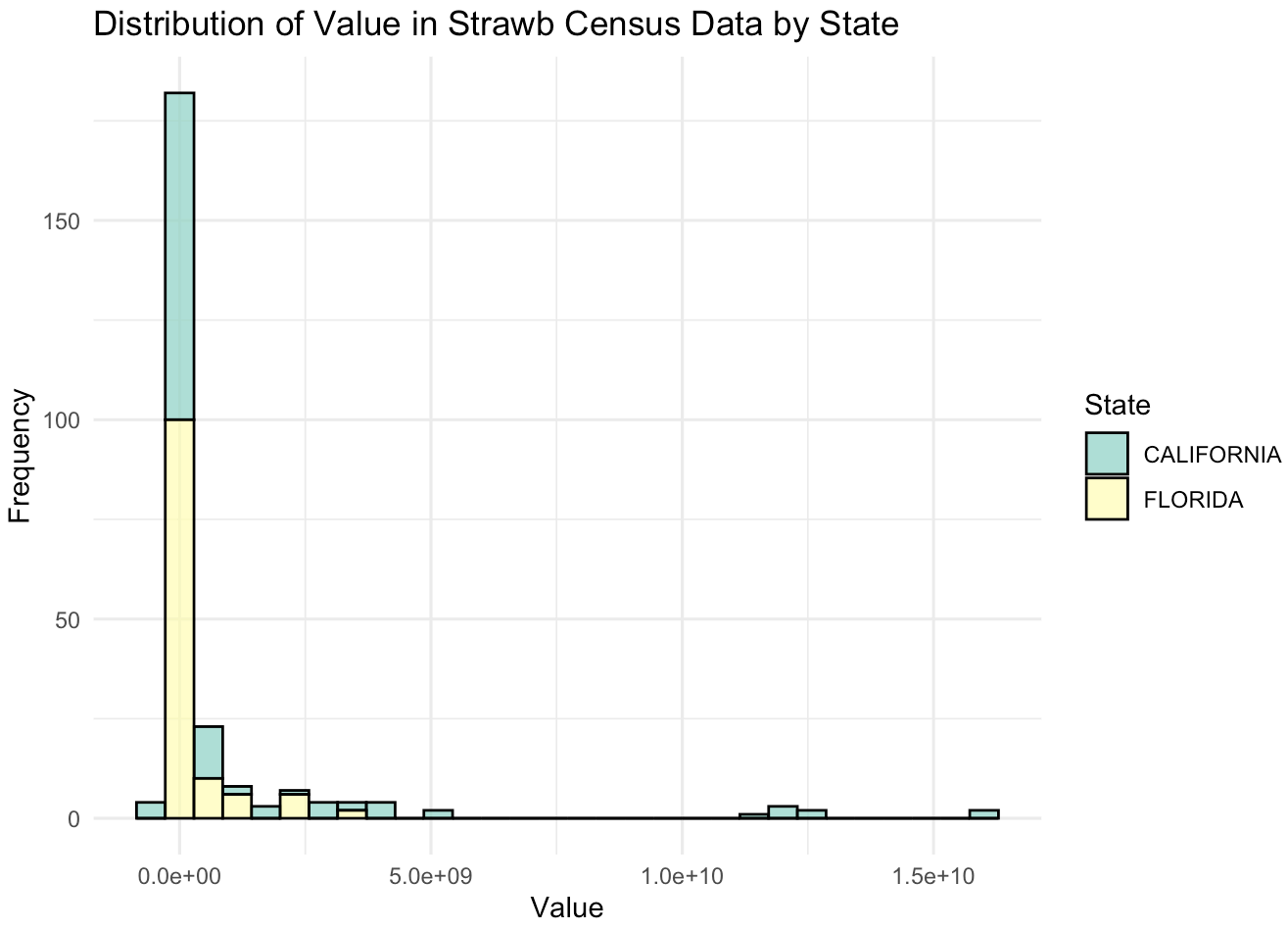
Not surprisingly California produces far more strawberries than Florida. California's total value was over 4 times more than Florida's.

Histogram

Next I wanted to see what the most common Values were as it would give me a better understanding of the frequency of strawberries are being moved for each state.

```
ggplot(strawb_census, aes(x = Value, fill = State)) +
  geom_histogram(bins = 30, color = "black", alpha = 0.7) +
  labs(title = "Distribution of Value in Strawb Census Data by State",
       x = "Value",
       y = "Frequency") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3")
```

```
Warning: Removed 2 rows containing non-finite outside the scale range
(`stat_bin()`).
```



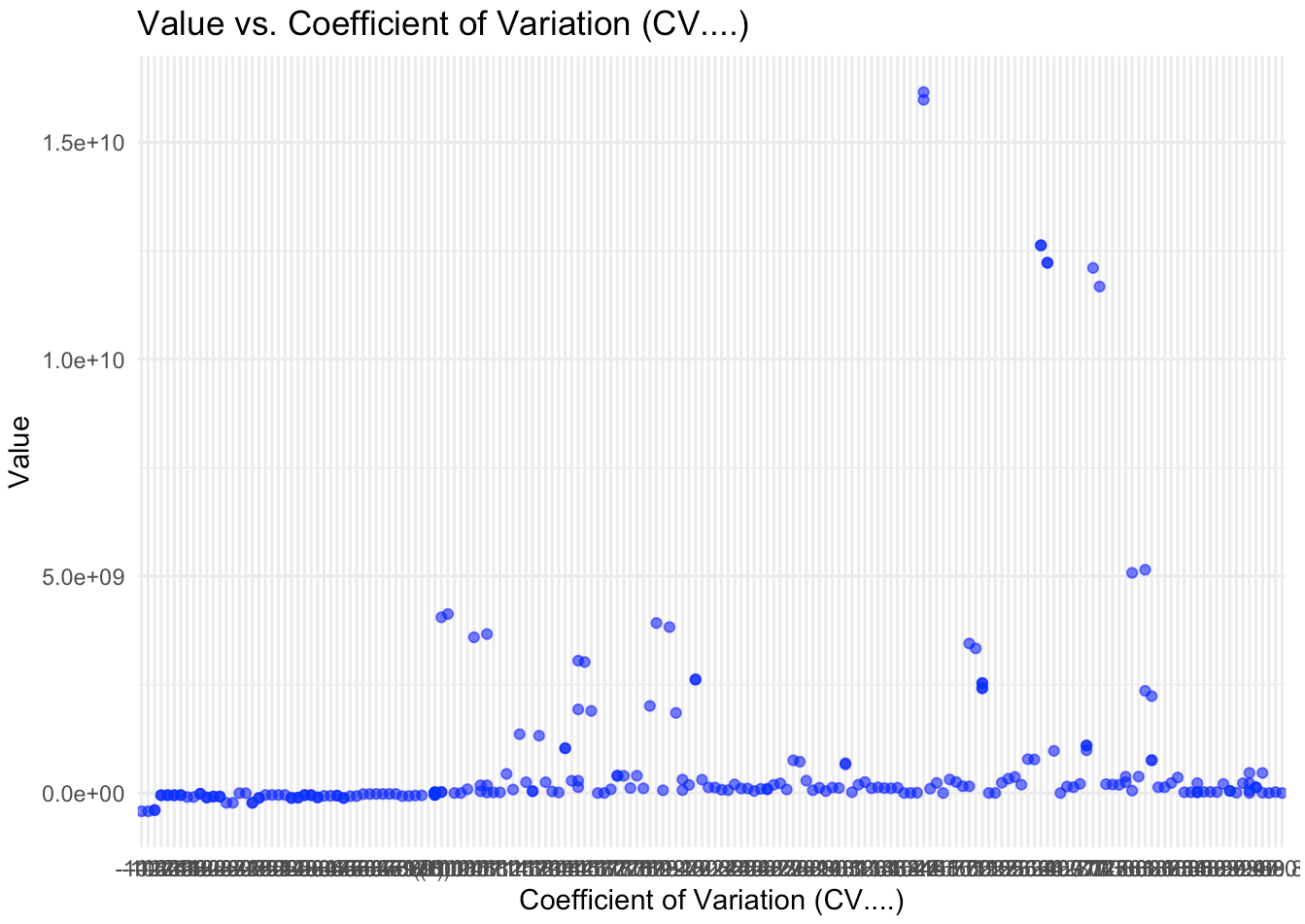
It looks like the most common value was on the smaller end and possibly close to zero for both states, but I also noticed how almost all the large outliers are from California. This could mean that California is moving some of its strawberries in large bulk amounts while Florida isn't.

Scatterplot

Finally, I wanted to see if there was a correlation between Value and the Coefficeint of Variation.

```
# Create a scatter plot of 'Value' vs 'CV....'
ggplot(strawb_census, aes(x = CV...., y = Value)) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(title = "Value vs. Coefficient of Variation (CV....)",
       x = "Coefficient of Variation (CV....)",
       y = "Value") +
  theme_minimal()
```

```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).
```



It looks like the data stays pretty constant with one type of value which makes sense because in the last visual we saw that the highest frequency was right around zero, but when there is an increase in value there does seem to be some trend on increase that could be looked at further in a future project.