415Midterm- Strawberry Cleaning

Cleaning the Strawberry Data Set

We will be cleaning and analyzing the USDA's NASS data on strawberry production, which gives insight into the agricultural industry. Prior to cleaning the Strawberry data set, it is important to understand the structure of the data and the data itself. Because of this, I used the str() and head() functions to get more information on the data. After this, removing any columns with missing values (NA) was the next step, so I individually deleted any columns with missing values before checking if the columns were deleted from the data set. Additionally I removed the State.ANSI column since it is just the USDA's NASS code assigned to each state which are already listed.

```
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


'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 F
 $ Data.Item       : chr  "INCOME, NET CASH FARM, OF OPERATIONS - GAIN, MEASURED IN $" "INCOM
 $ 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 $)" "NE
 $ Value           : chr  "6,312,000" "55,328,000" "100,618,000" "13,709,000" ...
 $ CV....          : chr  "9.2" "8.0" "4.9" "13.8" ...
```

```
  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
```

```
 [1] "Program"           "Year"              "Period"            "Week.Ending"
 [5] "Geo.Level"         "State"             "State.ANSI"        "Ag.District"
 [9] "Ag.District.Code"  "County"            "County.ANSI"       "Zip.Code"
[13] "Region"            "watershed_code"    "Watershed"         "Commodity"
[17] "Data.Item"         "Domain"            "Domain.Category"   "Value"
[21] "CV...."
```

```
 [1] "Program"           "Year"              "Period"            "Week.Ending"
 [5] "Geo.Level"         "State"             "State.ANSI"        "Ag.District"
 [9] "Ag.District.Code"  "County"            "County.ANSI"       "Zip.Code"
[13] "Region"            "watershed_code"    "Watershed"         "Commodity"
[17] "Data.Item"         "Domain"            "Domain.Category"   "Value"
[21] "CV...."
```

```
 [1] "Program"          "Year"             "Period"           "Geo.Level"
 [5] "State"            "Commodity"        "Data.Item"        "Domain"
 [9] "Domain.Category"  "Value"            "CV...."
```

The next step in my data cleaning is checking if any remaining columns have missing values. To do this, I ran a summary of the missing values per column. Next I used a for loop with the unique() function to check if the remaining columns had the same values in every row, or if there are diverse values. This code also deleted any columns with 1 value or NA values (as a way to double check my previous cleaning work), which disposed of the Geo.Level column since it only had 1 value. To continue cleaning the USDA NASS Strawberry data, I removed any data entries (rows) where the State was not Florida or California and used rbind() to stack the data sets on top of each other, further organizing the data. Eliminating the rows of data on strawberries from Connecticut, Maine, Massachusetts, New Hampshire, New York, Rhode Island and Vermont, allows us to strictly analyze the data from the two states producing the most strawberries.

```
        Program            Year          Period       Geo.Level           State
              0               0               0               0               0
      Commodity       Data.Item          Domain Domain.Category           Value
```

```
              0               0               0               0               0
         CV....
              0
```

```
[1] "CALIFORNIA"    "CONNECTICUT"   "FLORIDA"       "MAINE"
[5] "MASSACHUSETTS" "NEW HAMPSHIRE" "NEW YORK"      "RHODE ISLAND"
[9] "VERMONT"
```

```
[1] "CALIFORNIA" "FLORIDA"
```

Next, I began restructuring the columns to get rid of words that are included in other parts of the data set. Getting rid of the repetitive phrases will help make the data less cluttered and easier to analyze later. I then created a new column called Measure and strategically placed it between Data.Item and Domain to further organize the values that were originally in the Data.Item column better. I then created a for loop that had a similar structure to what I was taught in Python that deleted the word MEASURED and moved any words written after it (if any) to the new Measures column.

```
 [1] "OPERATIONS - GAIN MEASURED IN $"
 [2] "OPERATIONS - LOSS MEASURED IN $"
 [3] "OPERATIONS - NET INCOME MEASURED IN $"
 [4] "PRODUCERS - LOSS MEASURED IN $"
 [5] "PRODUCERS - NET INCOME MEASURED IN $"
 [6] " ORGANIC - ACRES HARVESTED"
 [7] " ORGANIC - SALES MEASURED IN $"
 [8] " ORGANIC - SALES MEASURED IN CWT"
 [9] " ORGANIC FRESH MARKET - SALES MEASURED IN $"
[10] " ORGANIC FRESH MARKET - SALES MEASURED IN CWT"
[11] " ORGANIC PROCESSING - OPERATIONS WITH SALES"
[12] " ORGANIC PROCESSING - SALES MEASURED IN $"
[13] " ORGANIC PROCESSING - SALES MEASURED IN CWT"
[14] "- PRICE RECEIVED MEASURED IN $ / CWT"
[15] " FRESH MARKET - PRICE RECEIVED MEASURED IN $ / CWT"
[16] " PROCESSING - PRICE RECEIVED MEASURED IN $ / CWT"
[17] "- ACRES HARVESTED"
[18] "- APPLICATIONS MEASURED IN LB"
[19] "- APPLICATIONS MEASURED IN LB / ACRE / APPLICATION AVG"
[20] "- APPLICATIONS MEASURED IN LB / ACRE / YEAR AVG"
[21] "- APPLICATIONS MEASURED IN NUMBER AVG"
[22] "- TREATED MEASURED IN PCT OF AREA BEARING AVG"
[23] "- YIELD MEASURED IN CWT / ACRE"
```

```
[24] " BEARING - APPLICATIONS MEASURED IN LB"
[25] " BEARING - APPLICATIONS MEASURED IN LB / ACRE / APPLICATION AVG"
[26] " BEARING - APPLICATIONS MEASURED IN LB / ACRE / YEAR AVG"
[27] " BEARING - APPLICATIONS MEASURED IN NUMBER AVG"
[28] " BEARING - TREATED MEASURED IN PCT OF AREA BEARING AVG"
```

After splitting the columns, there were some NA values in the Measure categories which I then replaced as NOT SPECIFIED. In addition, there were some extraneous symbols and spaces that needed to get removed, so gsub() and trimws() were used to remove such punctuation and spaces in Domain.Category and Data.Item.

In order to analyze the differences between strawberries that are organic, conventional and sold for processing, I created a new column that differentiates the type of strawberries.

```
  Production.Type    n
1    Conventional 1774
2         Organic   13
3      Processing    8
```

I then cleaned the Value data, changing the values to numeric and removing any punctuation before creating the columns Value.clean and Value.numeric and removing the original Value column.

```
Warning in unique(clean_data$Value[is.na(as.numeric(clean_data$Value)) & : NAs
introduced by coercion
```

```
  [1] "6,312,000"      "55,328,000"     "100,618,000"    "13,709,000"    
  [5] "15,979,024,000" "361,000"        "16,155,353,000" "14,782,000"    
  [9] "178,143,000"    "226,112,000"    "42,326,000"     "3,589,297,000" 
 [13] "474,000"        "4,051,134,000"  "1,927,889,000"  "-33,586,000"   
 [17] "376,405,000"    "223,317,000"    "285,520,000"    "282,080,000"   
 [21] "5,147,340,000"  "129,012,000"    "1,355,260,000"  "150,061,000"   
 [25] "2,007,260,000"  "253,660,000"    "-65,060,000"    "12,628,069,000"
 [29] "-106,536,000"   "-68,448,000"    "-81,917,000"    "-114,019,000"  
 [33] "195,437,000"    "-104,287,000"   "-33,686,000"    "248,883,000"   
 [37] "-394,217,000"   "-61,357,000"    "12,620,172,000" "-101,630,000"  
 [41] "-54,427,000"    "-81,200,000"    "-113,783,000"   "210,990,000"   
 [45] "-92,174,000"    "-31,354,000"    "230,302,000"    "-421,320,000"  
 [49] "370,330,000"    "3,050,538,000"  "3,918,351,000"  "773,639,000"   
 [53] "753,277,000"    "65,455,000"     "687,822,000"    "127,298,000"   
 [57] "185,907,000"    "2,622,937,000"  "539,000"        "440,665,000"   
 [61] "-24,247,000"    "-115,016,000"   "12,104,219,000" "14,792,000"
 [65] "178,428,000"    "225,741,000"    "42,488,000"     "3,664,599,000"
 [69] "4,126,522,000"  "1,892,878,000"  "-46,056,000"    "357,332,000"
 [73] "191,307,000"    "253,627,000"    "282,964,000"    "5,075,486,000"
 [77] "126,638,000"    "1,321,748,000"  "134,158,000"    "1,847,524,000"
 [81] "236,584,000"    "-64,878,000"    "12,226,709,000" "-106,305,000"
 [85] "-71,525,000"    "-81,775,000"    "-113,434,000"   "188,319,000"
 [89] "-104,173,000"   "-33,492,000"    "228,455,000"    "-393,713,000"
 [93] "-61,265,000"    "12,219,729,000" "-101,379,000"   "-57,684,000"
 [97] "-81,058,000"    "-113,211,000"   "203,678,000"    "-92,079,000"
[101] "-31,056,000"    "209,182,000"    "-420,669,000"   "328,698,000"
[105] "3,018,756,000"  "3,824,571,000"  "779,227,000"    "721,365,000"
[109] "63,297,000"     "658,069,000"    "110,554,000"    "39,371,000"
[113] "2,612,726,000"  "553,000"        "377,554,000"    "-114,941,000"
[117] "11,674,188,000" "4,228"          "311,784,980"    "1,412,627"
[121] " (D)"           "1,401,384"      "11,244"         "42,700"
[125] " (NA)"          "3,300"          "2,800"          "6,600"
[129] "603,100"        "30,300"         "8,600"          "22,400"
[133] "14,600"         "7,100"          "7,200"          "10,800"
[137] "4,000"          "2,300"          "10,600"         "8,300"
[141] "3,600"          "1,258,100"      "1,300"          "269,500"
[145] "2,338,800"      "3,900"          "71,400"         "89,700"
[149] "12,800"         "28,700"         "4,600"          "2,000"
[153] "1,700"          "7,600"          "6,200"          "5,000"
[157] "5,400"          "3,100"          "19,400"         "7,900"
[161] "5,600"          "3,800"          "231,600"        "11,299,000"
[165] "1,642,600"      "15,611,900"     "393,000"        "216,000"
[169] "43,500"         "40,200"         "1,900"          "2,100"
[173] "253,600"        "11,300"         "16,700"         "7,500"
[177] "1,800"          "2,400"          "4,100"          "3,400"
[181] "1,200"          "591,200"        "9,400"          "96,300"
[185] "1,056,400"      "6,300"          "16,900"         "7,400"
[189] "1,100"          "5,500"          "2,500"          "2,600"
[193] "29,100"         "2,900"          "5,800"          "2,200"
[197] "116,700"        "5,692,600"      "848,600"        "1,040,400"
[201] "3,000"          "7,602,900"      "37,600"         "6,129,000"
[205] "36,478,000"     "57,932,000"     "11,597,000"     "3,333,914,000"
[209] "452,000"        "3,446,500,000"  "22,953,000"     "132,964,000"
[213] "132,339,000"    "49,391,000"     "753,698,000"    "716,000"
[217] "1,092,060,000"  "309,366,000"    "3,887,000"      "232,432,000"
[221] "85,037,000"     "47,097,000"     "84,339,000"     "984,558,000"
[225] "87,509,000"     "131,097,000"    "115,499,000"    "198,473,000"
[229] "75,146,000"     "-50,931,000"    "2,535,231,000"  "-48,226,000"
```

```
[233] "22,511,000"     "-49,401,000"    "-21,504,000"    "96,770,000"
[237] "-44,275,000"    "18,118,000"     "121,833,000"    "-225,686,000"
[241] "-50,501,000"    "2,530,625,000"  "-48,476,000"    "25,977,000"
[245] "-47,676,000"    "-21,358,000"    "104,931,000"    "-45,059,000"
[249] "17,114,000"     "113,832,000"    "-224,969,000"   "8,958,000"
[253] "463,292,000"    "249,540,000"    "1,032,858,000"  "402,459,000"
[257] "1,220,000"      "6,605,000"      "394,633,000"    "-71,410,000"
[261] "154,553,000"    "-3,775,000"     "133,742,000"    "-17,145,000"
[265] "1,367,000"      "2,354,440,000"  "23,056,000"     "133,191,000"
[269] "132,360,000"    "49,116,000"     "758,630,000"    "715,000"
[273] "1,097,068,000"  "306,493,000"    "4,045,000"      "192,224,000"
[277] "66,343,000"     "44,154,000"     "82,161,000"     "971,452,000"
[281] "79,781,000"     "123,693,000"    "108,651,000"    "189,513,000"
[285] "63,683,000"     "-50,885,000"    "2,417,319,000"  "-48,091,000"
[289] "22,755,000"     "-49,325,000"    "-21,228,000"    "96,240,000"
[293] "-44,111,000"    "17,892,000"     "116,503,000"    "-224,874,000"
[297] "-50,454,000"    "2,412,574,000"  "-48,346,000"    "26,361,000"
[301] "-47,600,000"    "-21,079,000"    "104,151,000"    "-44,896,000"
[305] "16,917,000"     "108,718,000"    "-224,155,000"   "7,743,000"
[309] "462,096,000"    "248,349,000"    "1,031,835,000"  "399,019,000"
[313] "1,230,000"      "1,425,000"      "396,363,000"    "-73,456,000"
[317] "156,104,000"    "-3,760,000"     "19,516,000"     "-17,140,000"
[321] "1,887,000"      "2,232,193,000"  "18,358,396"     "67,146"
[325] "14,100"         "144,000"        "9,700"          "6,100"
[329] "112,100"        "302,700"        "9,900"          "1,500"
[333] "12,300"         "5,100"          "283,000"        "52,000"
[337] "538,000"        "135,100"        "4,500"          "142,400"
[341] "303,200"        " (Z)"           "11,700"
```

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

After cleaning the data set, the data was broken up into four data sets: California Survey data (CA_survey), California Census data (CA_census), Florida Survey data (FL_survey) and Florida Census data (FL_census). For additional cleaning, State and Program columns were removed from the four split data sets. In addition, census data and survey data were separated into two different data sets and all data where the Commodity column said STRAWBERRIES were also transferred to a new data set for easier analysis later.

```
[1] "Number of CA survey rows: 1051"
```

```
[1] "Number of CA census rows: 127"
```

```
[1] "Number of FL survey rows: 731"
```

```
[1] "Number of FL census rows: 124"
```

Analyzing the Strawberry Data

Now that the USDA NASS Strawberry Data has been cleaned, there are 2,033 entries remaining with 11 columns. This data set will be much more straightforward to look at when the analysis of the data begins. To begin analyzing the data, I wanted to look at some key differences between California and Florida's strawberry production. To do this I created summary count and metrics tables before forming a bar plot to compare the acreage between the two states. I really enjoyed using knitr:: to create report worthy tables with easily readable information.

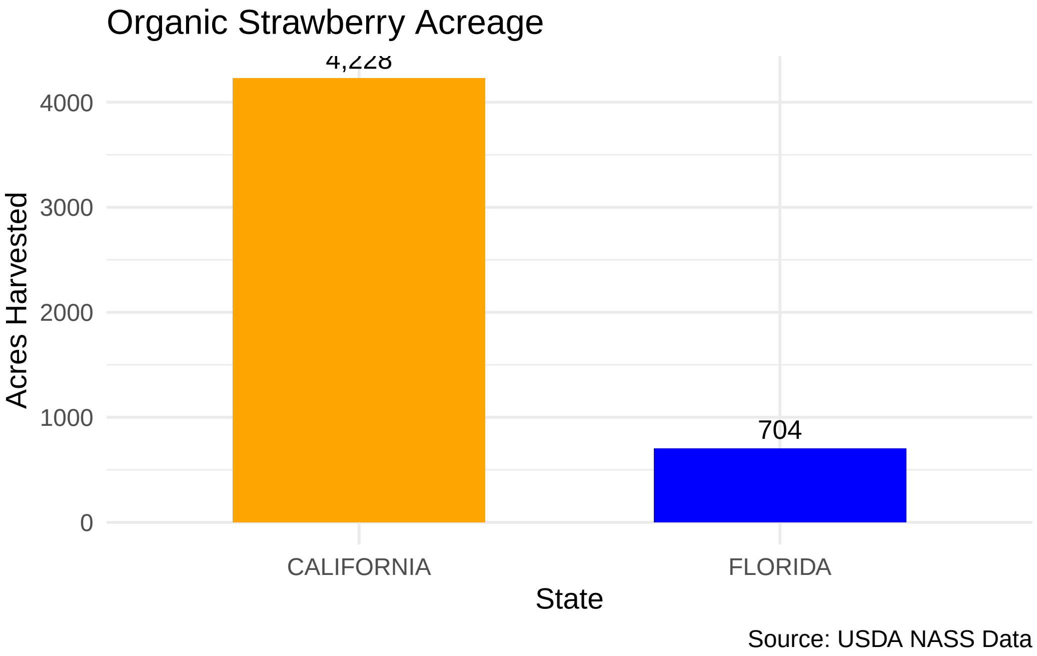
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.4     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
v readr     2.1.5
-- Conflicts ------------------------------------------ tidyverse_conflicts()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

Attaching package: 'scales'


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

    discard


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

    col_factor
```

Table 1: Strawberry Production Summary Statistics

State	Avg Value	Min Value	Max Value	Total Records
CALIFORNIA	606735.4	0.017	311784980	1059
FLORIDA	284626.5	0.017	18358396	736

Table 2: Key Metrics Comparison

State	ORGANIC - ACRES HARVESTED	ORGANIC - SALES
CALIFORNIA	4,228	156,598,804
FLORIDA	704	9,212,771



Source: USDA NASS Data

In order to determine which 3 conventionally used chemicals I should take a closer look at, I filtered out the strawberries so only the ones listed as conventional and used in both Florida and California would show up.

```
[1] "Conventional Domain.Category values in both states:"
```

```
 [1] "ABAMECTIN = 122804"
 [2] "ACETAMIPRID = 99050"
 [3] "ACIBENZOLAR-S-METHYL = 61402"
 [4] "AZOXYSTROBIN = 128810"
 [5] "BACILLUS AMYLOLIQUEFAC F727 = 16489"
 [6] "BACILLUS SUBTILIS = 6479"
 [7] "BIFENAZATE = 586"
 [8] "BIFENTHRIN = 128825"
 [9] "BORAX DECAHYDRATE = 11102"
[10] "BOSCALID = 128008"
```

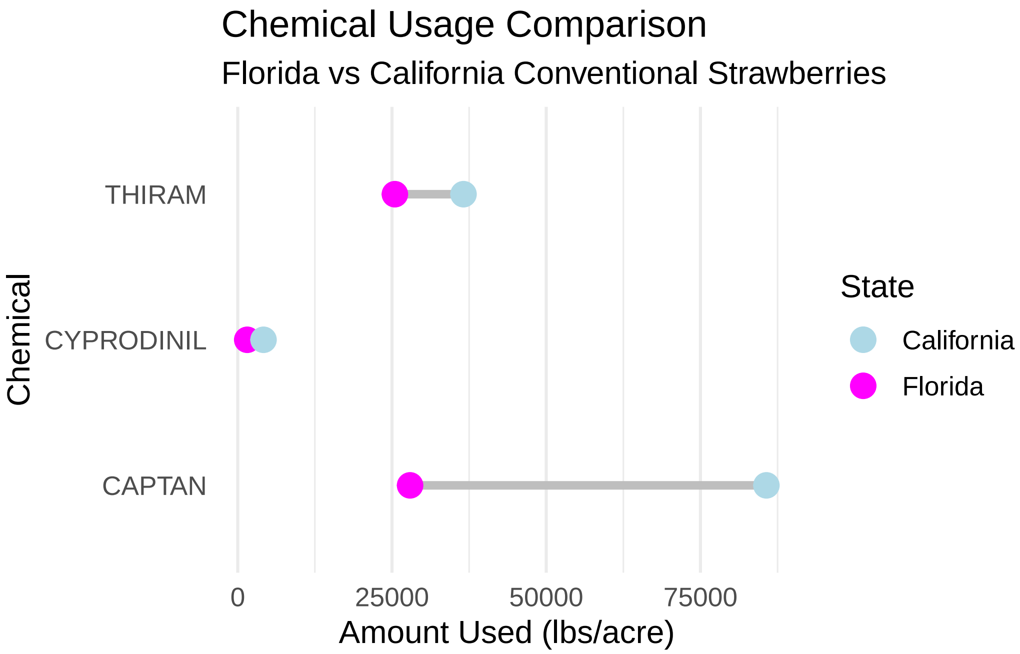
```
[11] "BT KURSTAKI ABTS-351 = 6522"
[12] "CAPTAN = 81301"
[13] "CARFENTRAZONE-ETHYL = 128712"
[14] "CHLORANTRANILIPROLE = 90100"
[15] "CYANTRANILIPROLE = 90098"
[16] "CYFLUFENAMID = 555550"
[17] "CYFLUMETOFEN = 138831"
[18] "CYPRODINIL = 288202"
[19] "DIAZINON = 57801"
[20] "DIFENOCONAZOLE = 128847"
[21] "FENHEXAMID = 90209"
[22] "FENPROPATHRIN = 127901"
[23] "FENPYROXIMATE = 129131"
[24] "FLUDIOXONIL = 71503"
[25] "FLUMIOXAZIN = 129034"
[26] "FLUOPYRAM = 80302"
[27] "FLUPYRADIFURONE = 122304"
[28] "FLUTRIAFOL = 128940"
[29] "FLUXAPYROXAD = 138009"
[30] "FOSETYL-AL = 123301"
[31] "GLYPHOSATE ISO. SALT = 103601"
[32] "HEXYTHIAZOX = 128849"
[33] "IMIDACLOPRID = 129099"
[34] "ISOFETAMID = 270000"
[35] "LAMBDA-CYHALOTHRIN = 128897"
[36] "MALATHION = 57701"
[37] "MEFENOXAM = 113502"
[38] "METAM-POTASSIUM = 39002"
[39] "METHOXYFENOZIDE = 121027"
[40] "MONO-POTASSIUM SALT = 76416"
[41] "NALED = 34401"
[42] "NITROGEN"
[43] "NOT SPECIFIED"
[44] "NOVALURON = 124002"
[45] "OXATHIAPIPROLIN = 128111"
[46] "PARAQUAT = 61601"
[47] "PENTHIOPYRAD = 90112"
[48] "PHOSPHATE"
[49] "PIPERONYL BUTOXIDE = 67501"
[50] "POTASH"
[51] "PROPICONAZOLE = 122101"
[52] "PSEUDOMONAS CHLORORAPHIS STRAIN AFS009 = 6800"
[53] "PYDIFLUMETOFEN = 90110"
```

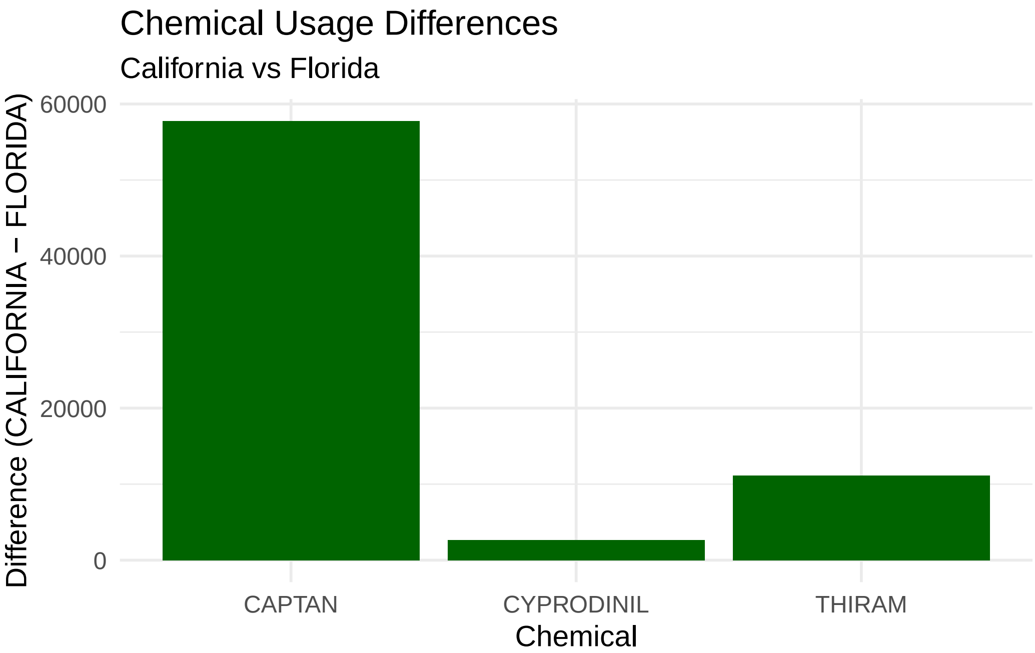
```
[54] "PYRACLOSTROBIN = 99100"
[55] "PYRETHRINS = 69001"
[56] "PYRIMETHANIL = 288201"
[57] "SPINETORAM = 110007"
[58] "SPINOSAD = 110003"
[59] "SPIROMESIFEN = 24875"
[60] "SULFUR"
[61] "SULFUR = 77501"
[62] "TETRACONAZOLE = 120603"
[63] "THIAMETHOXAM = 60109"
[64] "THIOPHANATE-METHYL = 102001"
[65] "THIRAM = 79801"
[66] "TOTAL"
[67] "TRIFLUMIZOLE = 128879"
```

The three chemicals that I chose to isolate were Thiram, Cyprodinil, and Captan. I defined the uses of the three fungicides with their toxicity levels before getting the usage data in California and Florida. I them inputted that usage data into a comparison table. I struggled greatly with this code as cleaning the chemicals and finding interesting data was difficult to do when using coding techniques that were new to me. I then created visualizations (dot plot and bar plot) that are visually appealing while still exhibiting the data clearly. This aspect of the data analysis was fun for me as I got to try different ways to plot and got to customize them. The dot plot compares the usage of the three chemicals between states while the bar plot exhibits the difference in usage between states.

```
[1] "Chemical Comparison for Conventional Strawberries:"
```

```
# A tibble: 3 x 6
  Chemical   Use                           Toxicity FLORIDA CALIFORNIA Difference
  <chr>      <chr>                         <chr>      <dbl>      <dbl>      <dbl>
1 CAPTAN     Fungicide for fruit rot       Moderate  27931.     85688.     57758.
2 CYPRODINIL Fungicide for gray mold       Low        1536.      4174.      2637.
3 THIRAM     Fungicide for seed treatment  High      25468.     36594.     11126.
```





Here is my analysis of the different production types (organic and conventional). First I began by creating a table for production comparison, then sales comparison and then processing comparison.

```
   Production.Type                                        Data.Item    n
1     Conventional                                  ACRES HARVESTED    8
2     Conventional                                     APPLICATIONS   68
3     Conventional                           BEARING - APPLICATIONS 1316
4     Conventional                                BEARING - TREATED  341
5     Conventional                     FRESH MARKET - PRICE RECEIVED    8
6     Conventional                                   PRICE RECEIVED    8
7     Conventional                                          TREATED   17
8     Conventional                                            YIELD    8
9          Organic                        ORGANIC - ACRES HARVESTED    2
10         Organic                                  ORGANIC - SALES    4
11         Organic                     ORGANIC FRESH MARKET - SALES    4
12         Organic ORGANIC PROCESSING - OPERATIONS WITH SALES    1
13         Organic                       ORGANIC PROCESSING - SALES    2
14      Processing                      PROCESSING - PRICE RECEIVED    8
```

```
# A tibble: 2 x 2
  Production.Type Production
  <chr>                <dbl>
1 Conventional        214900
2 Organic               4932
```

```
# A tibble: 1 x 2
  Production.Type     Sales
  <chr>               <dbl>
1 Organic         351461326
```

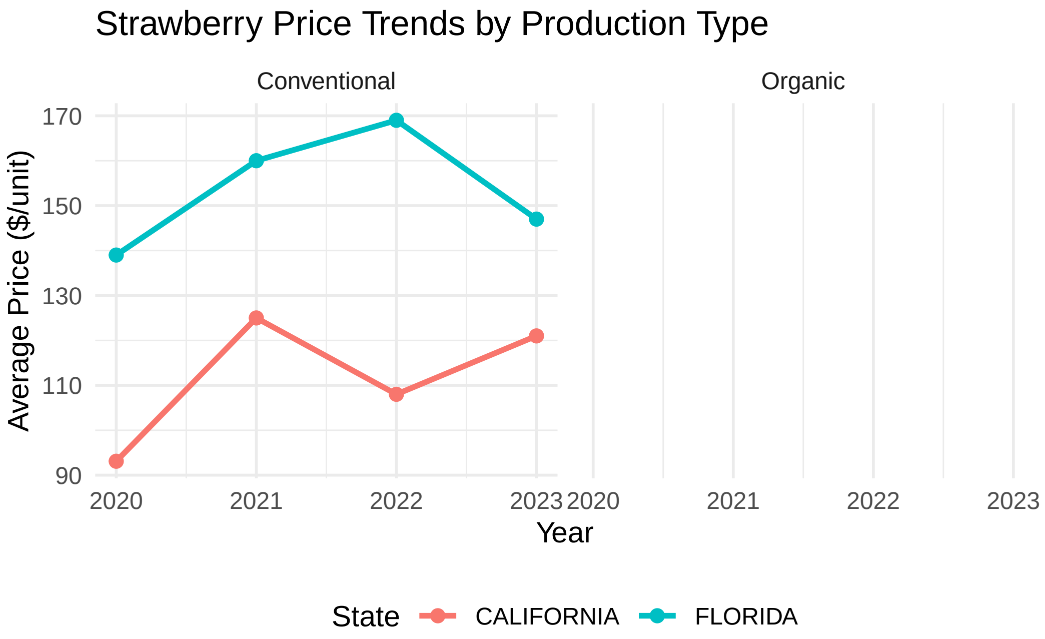
```
# A tibble: 2 x 2
  Production.Type Processing
  <chr>                <dbl>
1 Organic              11251
2 Processing               0
```

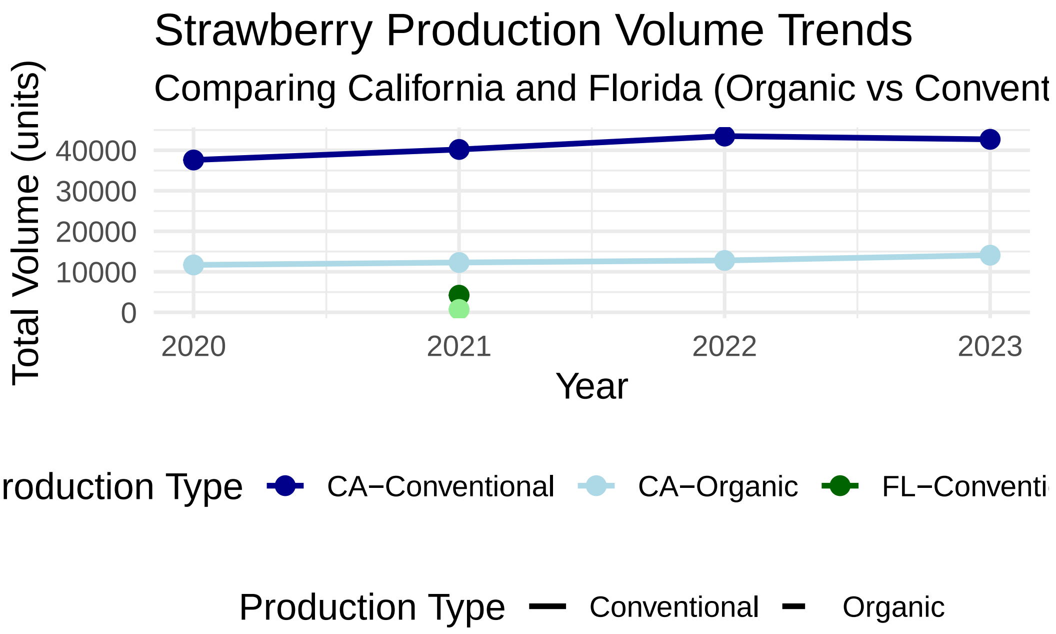
For the second question of the assignment, I created visualizations that show price trends by production type and production volume trends by state and production type. In addition, I also created tables to show the direct comparisons of the state and production type data, something I was introduced to briefly in my past internship. The most difficult part of this project was modifying the code for the visualizations to make them look put together as well as figuring out which functions to choose to make the cleaned data set be used to its full potential.

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

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

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





```
[1] "Annual CA/FL Ratios by Production Type:"

# A tibble: 5 x 4
   Year Production.Type Price_Ratio Volume_Ratio
  <dbl> <chr>                 <dbl>        <dbl>
1  2020 Conventional          0.670         3.21
2  2021 Conventional          0.781         3.27
3  2021 Organic             NaN             6.01
4  2022 Conventional          0.639         3.40
5  2023 Conventional          0.823         3.03

[1] "Annual Differences (CA - FL):"

# A tibble: 5 x 9
  Production.Type  Year Avg_Cost Avg_Price_CALIFORNIA Avg_Price_FLORIDA
  <chr>           <dbl>    <dbl>                <dbl>             <dbl>
1 Conventional     2020      NaN                 93.1               139
2 Conventional     2021      NaN                125                 160
3 Conventional     2022      NaN                108                 169
4 Conventional     2023      NaN                121                 147
5 Organic          2021      NaN                NaN                 NaN
# i 4 more variables: Total_Volume_CALIFORNIA <dbl>,
#   Total_Volume_FLORIDA <dbl>, Price_Diff <dbl>, Volume_Diff <dbl>
```