Strawberries

Bruce Mallory

10/19/2020

My Goal

My overall goal is to create a data.frame that will allow me to look at the type of chemical application (fungicide, herbicide, insecticide, or fertilizer) and examine per acre applications in selected states during selected years.

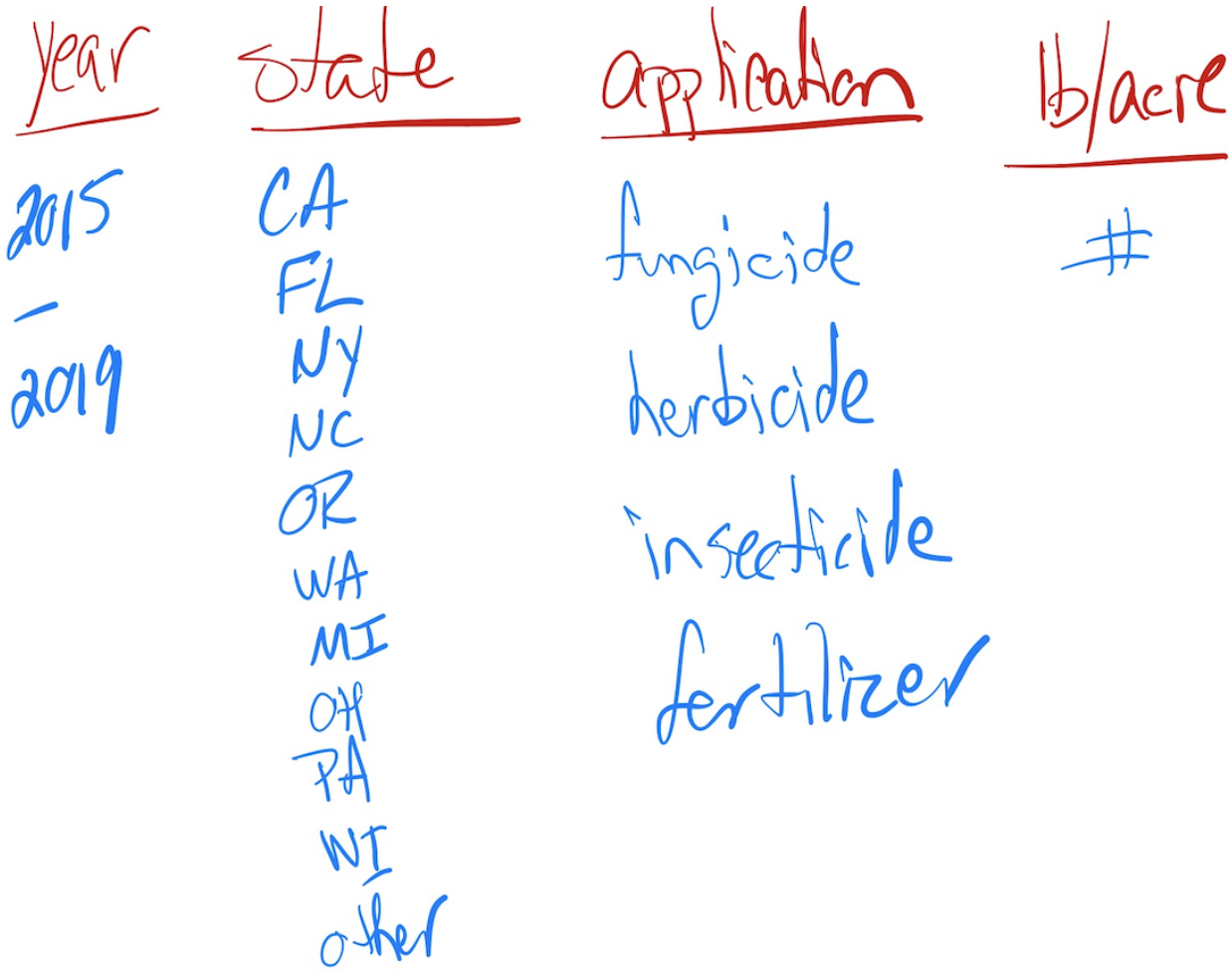

Figure 1: my target data.frame organization

MY STEPS

(1) Read and examine

These data were collected from the USDA database selector: https://quickstats.nass.usda.gov

The data were stored online and then downloaded as a CSV file.

The data has 21 columns.

```
##  [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 (%)"
```

(2) Remove the columns that only had NAs

This leaves 12 columns.

```
##  [1] "Program"         "Year"            "Period"          "Geo Level"
##  [5] "State"           "State ANSI"      "watershed_code"  "Commodity"
##  [9] "Data Item"       "Domain"          "Domain Category" "Value"
```

(3) Remove the columns that provide no new information

"Program" and "Geo Level" have only 1 value. And "State ANSI" is a duplicate of "State"

This leaves 8 columns.

```
## [1] "Year"            "Period"          "State"           "Commodity"
## [5] "Data Item"       "Domain"          "Domain Category" "Value"
```

(4) Select the rows that contain "Strawberries" & only the rows where the 'Period' = "YEAR". Then eliminate the Period and Commodity column.

NOTE: The Period column has three values: "MARKETING YEAR", "YEAR", and "YEAR - AUG FORECAST." I am only keeping the records where Period = "YEAR" so that we have a consistent comparison.

This leaves 6 columns.

```
## [1] "Year"            "State"           "Data Item"       "Domain"
## [5] "Domain Category" "Value"
```

(5) From the "Domain" column filter out unecessary records

In "Domain" column these are the unique enteries:

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

Before filtering out all the records where Domain==“TOTAL” I checked to see what information was in those records in the “Data Item” column and in the “Domain Category” column.

```
##  [1] "STRAWBERRIES - ACRES HARVESTED"
##  [2] "STRAWBERRIES - ACRES PLANTED"
##  [3] "STRAWBERRIES - PRODUCTION, MEASURED IN $"
##  [4] "STRAWBERRIES - PRODUCTION, MEASURED IN CWT"
##  [5] "STRAWBERRIES - YIELD, MEASURED IN CWT / ACRE"
##  [6] "STRAWBERRIES, FRESH MARKET - PRODUCTION, MEASURED IN $"
##  [7] "STRAWBERRIES, FRESH MARKET, UTILIZED - PRODUCTION, MEASURED IN CWT"
##  [8] "STRAWBERRIES, NOT SOLD - PRODUCTION, MEASURED IN CWT"
##  [9] "STRAWBERRIES, PROCESSING - PRODUCTION, MEASURED IN $"
## [10] "STRAWBERRIES, PROCESSING, UTILIZED - PRODUCTION, MEASURED IN CWT"
## [11] "STRAWBERRIES, UTILIZED - PRODUCTION, MEASURED IN CWT"
## [12] "STRAWBERRIES, PROCESSING, UTILIZED - PRODUCTION, MEASURED IN TONS"
## [13] "STRAWBERRIES, PROCESSING - PRODUCTION, MEASURED IN CWT"
## [14] "STRAWBERRIES, FRESH MARKET - PRODUCTION, MEASURED IN CWT"
```

```
## [1] "NOT SPECIFIED"
```

None of these enteries refered to chemical applications. So I was comfortable removing those records. This brought the data.frame down from 3220 rows to 2862 rows

(6) Filter useful records from the “Data Items” column

In this column there were 5 unique enteries

```
## [1] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB"
## [2] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG"
## [3] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB / ACRE / YEAR, AVG"
## [4] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN NUMBER, AVG"
## [5] "STRAWBERRIES, BEARING - TREATED, MEASURED IN PCT OF AREA BEARING, AVG"
```

I made a table to see which of these categories had the most information for me to use. This is the table: