Strawberries

Bruce Mallory

10/20/2020

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.

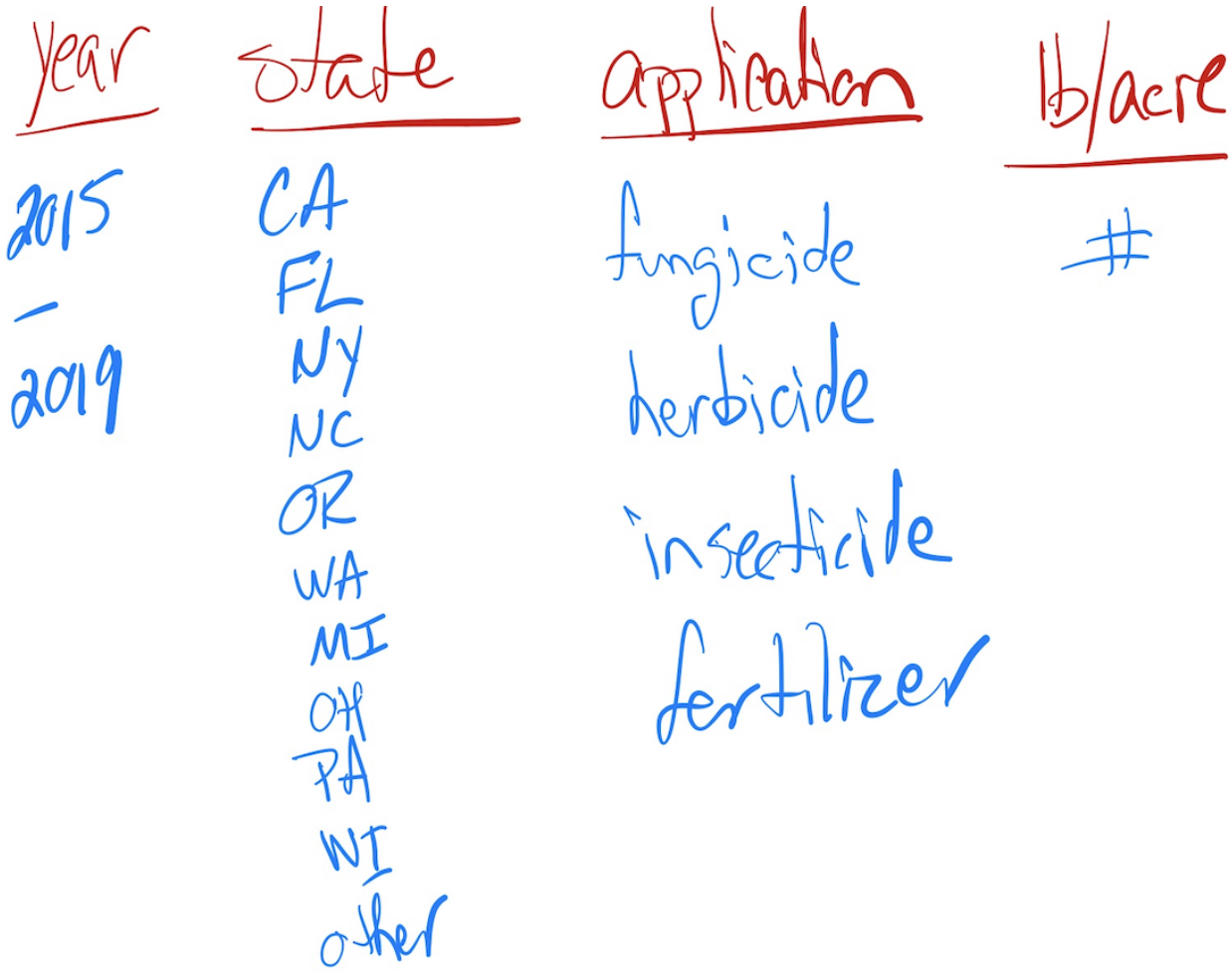

Figure 1: my target data.frame organization

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 to 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 NA columns

After removing all the columns that only had NAs in them, I had these 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" adds no new information to "State"

This leaves 8 columns.

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

(4) Filter for 'Strawberries' and 'Year'

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. After filtering, I removed the "Period" and "Commodity" columns.

This leaves 6 columns.

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

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

In the "Domain" column these are the unique entries:

```
## [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"
```

(6) From the "Data Items" column filter the useful records

In this column there were 5 unique entries

```
##  [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, BEARING - APPLICATIONS, MEASURED IN LB"
##  [7] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG"
##  [8] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB / ACRE / YEAR, AVG"
##  [9] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN NUMBER, AVG"
## [10] "STRAWBERRIES, BEARING - TREATED, MEASURED IN PCT OF AREA BEARING, AVG"
## [11] "STRAWBERRIES, FRESH MARKET - PRODUCTION, MEASURED IN $"
## [12] "STRAWBERRIES, FRESH MARKET, UTILIZED - PRODUCTION, MEASURED IN CWT"
## [13] "STRAWBERRIES, NOT SOLD - PRODUCTION, MEASURED IN CWT"
## [14] "STRAWBERRIES, PROCESSING - PRODUCTION, MEASURED IN $"
## [15] "STRAWBERRIES, PROCESSING, UTILIZED - PRODUCTION, MEASURED IN CWT"
## [16] "STRAWBERRIES, UTILIZED - PRODUCTION, MEASURED IN CWT"
## [17] "STRAWBERRIES, PROCESSING, UTILIZED - PRODUCTION, MEASURED IN TONS"
## [18] "STRAWBERRIES, PROCESSING - PRODUCTION, MEASURED IN CWT"
## [19] "STRAWBERRIES, FRESH MARKET - PRODUCTION, MEASURED IN CWT"
```

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

Var1	Freq
STRAWBERRIES - ACRES HARVESTED	36
STRAWBERRIES - ACRES PLANTED	32
STRAWBERRIES - PRODUCTION, MEASURED IN $	32
STRAWBERRIES - PRODUCTION, MEASURED IN CWT	36
STRAWBERRIES - YIELD, MEASURED IN CWT / ACRE	36

Since the frequencies were almost identical, I decided to use the one that made the most sense to me: "LB/ACRE/YEAR on average." I then filtered in these rows and deleted the "Data Item" column, and renamed the "Value" column to be "Avg lb/acre/yr." This is the head of the data.frame so far:

Year	State	Domain	Domain Category	Avg lb/acre/yr
2019	CALIFORNIA	CHEMICAL, FUNGICIDE	CHEMICAL, FUNGICIDE: (AZOXYSTROBIN = 128810)	0.447
2019	CALIFORNIA	CHEMICAL, FUNGICIDE	CHEMICAL, FUNGICIDE: (BACILLUS AMYLOLIQUEFACIENS MBI 600 = 129082)	(NA)
2019	CALIFORNIA	CHEMICAL, FUNGICIDE	CHEMICAL, FUNGICIDE: (BACILLUS AMYLOLIQUEFACIENS STRAIN D747 = 16482)	(NA)
2019	CALIFORNIA	CHEMICAL, FUNGICIDE	CHEMICAL, FUNGICIDE: (BACILLUS PUMILUS = 6485)	(NA)
2019	CALIFORNIA	CHEMICAL, FUNGICIDE	CHEMICAL, FUNGICIDE: (BACILLUS SUBT. GB03 = 129068)	(NA)

(7) Create an "Application" column.

To do this I first separated the "Domain" column and kept just the back end of each of the entries. I relabeled this column "Application." This gave me the following unique entries in the "Application" column:

```
## [1] "FUNGICIDE"   "HERBICIDE"   "INSECTICIDE" "OTHER"       NA
```

Then I needed to see if all of the NAs in the "Application" column were when a 'FERTILIZER' was used. To do this I wrote a loop that checked each row to see if, when the entry in "Application" was = NA the "Domain Category" contained 'FERTILIZER.' There were 15 instances, so I concluded that I could replace the NAs in the "Application" column with "FERTILIZER" and delete the "Domain Category" column.

```
n<-length(straw$Application)
x<-0
for (i in (1:n)) {
  if((straw$Application[i]=="NA")|(is.na(straw$Application[i])) & (str_detect(straw$`Domain Category`[i]
    x<-x+1
  }
}

print(x)
```

```
## [1] 15
```

Finally, I filtered out the records where "Application" contained 'OTHER,' and where "Avg lb/acre/yr" contained (NA) or (D). This got me down to 225 records.

MY FINAL PRODUCT (n=225):

Year	State	Application	Avg lb/acre/yr
2019	CALIFORNIA	FUNGICIDE	0.447
2019	CALIFORNIA	FUNGICIDE	0.54
2019	CALIFORNIA	FUNGICIDE	0.051
2019	CALIFORNIA	FUNGICIDE	0.508
2019	CALIFORNIA	FUNGICIDE	10.456