Strawberry HW

帖达玉

2024-10-02

```
library(knitr)
library(kableExtra)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter()     masks stats::filter()
## ✖ dplyr::group_rows() masks kableExtra::group_rows()
## ✖ dplyr::lag()        masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stringr)

options(echo = FALSE, digits = 3,
        scipen = 999, warn = FALSE, message = FALSE)
```

```
strawberry <- read_csv("strawberries25_v3.csv", col_names = TRUE)
```

```
## Rows: 12669 Columns: 21
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (15): Program, Period, Geo Level, State, State ANSI, Ag District, County...
## dbl  (2): Year, Ag District Code
## lgl  (4): Week Ending, Zip Code, Region, Watershed
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
library(dplyr)
library(stringr)

strawberry <- strawberry %>%
  mutate(Category = case_when(
    Domain == "Total" ~ NA_character_,
    str_detect(Domain, "CHEMICAL") ~ str_trim(str_remove(Domain, "CHEMICAL, ")),
    TRUE ~ Domain
  ))

unique(strawberry$Category)
```

```
## [1] "TOTAL"           "AREA GROWN"      "ORGANIC STATUS" "FUNGICIDE"
## [5] "INSECTICIDE"     "OTHER"           "HERBICIDE"      "FERTILIZER"
```

```
strawberry <- strawberry %>%
  mutate(
    Name = case_when(
      Category == "TOTAL" ~ NA_character_,
      str_detect(`Domain Category`, fixed(Category)) & str_detect(`Domain Category`, "\\(.*=.*\\)") ~
        str_extract(`Domain Category`, "(?<=\\().*?(?=\\s?=)"),
      str_detect(`Domain Category`, fixed(Category)) & str_detect(`Domain Category`, "\\(.*\\)") ~
        str_extract(`Domain Category`, "(?<=\\().*?(?=\\))"),
      TRUE ~ NA_character_
    ),
    Number = case_when(
      Category == "TOTAL" ~ NA_real_,
      str_detect(`Domain Category`, fixed(Category)) & str_detect(`Domain Category`, "\\(.*=.*\\)") ~
        as.numeric(str_extract(`Domain Category`, "(?<=\\=\\s?).*?(?=\\))")),
      str_detect(`Domain Category`, fixed(Category)) & str_detect(`Domain Category`, "\\(.*\\)") ~
        NA_real_,
      TRUE ~ NA_real_
    )
  )

head(strawberry)
```

```
## # A tibble: 6 × 24
##   Program  Year Period `Week Ending` `Geo Level` State   `State ANSI`
##   <chr>   <dbl> <chr>  <lgl>        <chr>       <chr>   <chr>
## 1 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 2 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 3 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 4 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 5 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 6 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## # ℹ 17 more variables: `Ag District` <chr>, `Ag District Code` <dbl>,
## #   County <chr>, `County ANSI` <chr>, `Zip Code` <lgl>, Region <lgl>,
## #   watershed_code <chr>, Watershed <lgl>, Commodity <chr>, `Data Item` <chr>,
## #   Domain <chr>, `Domain Category` <chr>, Value <chr>, `CV (%)` <chr>,
## #   Category <chr>, Name <chr>, Number <dbl>
```

```
strawberry <- strawberry %>%
  mutate(Category = case_when(
    `Domain Category` == "NOT SPECIFIED" ~ NA_character_,
    TRUE ~ Category
  ))

head(strawberry)
```

```
## # A tibble: 6 × 24
##   Program  Year Period `Week Ending` `Geo Level` State   `State ANSI`
##   <chr>   <dbl> <chr>  <lgl>        <chr>       <chr>   <chr>
## 1 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 2 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 3 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 4 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 5 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 6 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## # ℹ 17 more variables: `Ag District` <chr>, `Ag District Code` <dbl>,
## #   County <chr>, `County ANSI` <chr>, `Zip Code` <lgl>, Region <lgl>,
## #   watershed_code <chr>, Watershed <lgl>, Commodity <chr>, `Data Item` <chr>,
## #   Domain <chr>, `Domain Category` <chr>, Value <chr>, `CV (%)` <chr>,
## #   Category <chr>, Name <chr>, Number <dbl>
```

```
strawberry <- strawberry %>%
  mutate(information = gsub("STRAWBERRIES\\s*", "", `Data Item`))

head(strawberry)
```

```
## # A tibble: 6 × 25
##   Program  Year Period `Week Ending` `Geo Level` State   `State ANSI`
##   <chr>   <dbl> <chr>  <lgl>        <chr>       <chr>   <chr>
## 1 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 2 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 3 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 4 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 5 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## 6 CENSUS   2022 YEAR   NA           COUNTY      ALABAMA 01
## # ℹ 18 more variables: `Ag District` <chr>, `Ag District Code` <dbl>,
## #   County <chr>, `County ANSI` <chr>, `Zip Code` <lgl>, Region <lgl>,
## #   watershed_code <chr>, Watershed <lgl>, Commodity <chr>, `Data Item` <chr>,
## #   Domain <chr>, `Domain Category` <chr>, Value <chr>, `CV (%)` <chr>,
## #   Category <chr>, Name <chr>, Number <dbl>, information <chr>
```

```
strawberry_chemical <- strawberry %>%
  filter(grepl("CHEMICAL|FERTILIZER", `Domain Category`))

head(strawberry_chemical)
```

```
## # A tibble: 6 × 25
##   Program  Year Period `Week Ending` `Geo Level` State      `State ANSI`
##   <chr>   <dbl> <chr>  <lgl>        <chr>       <chr>      <chr>
## 1 SURVEY   2023 YEAR   NA           STATE       CALIFORNIA 06
## 2 SURVEY   2023 YEAR   NA           STATE       CALIFORNIA 06
## 3 SURVEY   2023 YEAR   NA           STATE       CALIFORNIA 06
## 4 SURVEY   2023 YEAR   NA           STATE       CALIFORNIA 06
## 5 SURVEY   2023 YEAR   NA           STATE       CALIFORNIA 06
## 6 SURVEY   2023 YEAR   NA           STATE       CALIFORNIA 06
## # ℹ 18 more variables: `Ag District` <chr>, `Ag District Code` <dbl>,
## #   County <chr>, `County ANSI` <chr>, `Zip Code` <lgl>, Region <lgl>,
## #   watershed_code <chr>, Watershed <lgl>, Commodity <chr>, `Data Item` <chr>,
## #   Domain <chr>, `Domain Category` <chr>, Value <chr>, `CV (%)` <chr>,
## #   Category <chr>, Name <chr>, Number <dbl>, information <chr>
```

```
strawberry_AREA <- strawberry %>%
  filter(grepl("AREA GROW|ORGANIC STATUS", `Domain Category`))
head(strawberry_AREA)
```

```
## # A tibble: 6 × 25
##   Program  Year Period `Week Ending` `Geo Level` State    `State ANSI`
##   <chr>   <dbl> <chr>  <lgl>        <chr>       <chr>    <chr>
## 1 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 2 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 3 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 4 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 5 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 6 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## # ℹ 18 more variables: `Ag District` <chr>, `Ag District Code` <dbl>,
## #   County <chr>, `County ANSI` <chr>, `Zip Code` <lgl>, Region <lgl>,
## #   watershed_code <chr>, Watershed <lgl>, Commodity <chr>, `Data Item` <chr>,
## #   Domain <chr>, `Domain Category` <chr>, Value <chr>, `CV (%)` <chr>,
## #   Category <chr>, Name <chr>, Number <dbl>, information <chr>
```

```
strawberry_AREA <- strawberry_AREA %>%
  mutate(
    Min = case_when(
      str_detect(Name, "100 OR MORE ACRES") ~ 100,
      str_detect(Name, "TO") ~ as.numeric(str_extract(Name, "^[0-9.]+")),
      TRUE ~ NA_real_
    ),
    Max = case_when(
      str_detect(Name, "100 OR MORE ACRES") ~ "MORE",
      str_detect(Name, "TO") ~ str_extract(Name, "(?<=TO )[0-9.]+"),
      TRUE ~ NA_character_
    )
  )

head(strawberry_AREA)
```

```
## # A tibble: 6 × 27
##   Program  Year Period `Week Ending` `Geo Level` State    `State ANSI`
##   <chr>   <dbl> <chr>  <lgl>        <chr>       <chr>    <chr>
## 1 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 2 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 3 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 4 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 5 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## 6 CENSUS   2022 YEAR   NA           NATIONAL    US TOTAL <NA>
## # ℹ 20 more variables: `Ag District` <chr>, `Ag District Code` <dbl>,
## #   County <chr>, `County ANSI` <chr>, `Zip Code` <lgl>, Region <lgl>,
## #   watershed_code <chr>, Watershed <lgl>, Commodity <chr>, `Data Item` <chr>,
## #   Domain <chr>, `Domain Category` <chr>, Value <chr>, `CV (%)` <chr>,
## #   Category <chr>, Name <chr>, Number <dbl>, information <chr>, Min <dbl>,
## #   Max <chr>
```

```
strawberry_cleaned=combine(strawberry_AREA,strawberry_chemical)
```

```
## Warning: `combine()` was deprecated in dplyr 1.0.0.
## ℹ Please use `vctrs::vec_c()` instead.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
write.csv(strawberry_cleaned, file = "strawberry_cleaned.csv", row.names = FALSE)
```