Stawberries

AUTHOR
MA615

PUBLISHED
October 14, 2024

Strawberries: Data

This is a project about acquiring strawberry data from the USDA-NASS system and then cleaning, organizing, and exploring the data in preparation for data analysis. To get started, I acquired the data from the USDA NASS system and downloaded them in a csv.

Data cleaning and organization references

"An introduction to data cleaning with R" by Edwin de Jonge and Mark van der Loo

"Problems, Methods, and Challenges in Comprehensive Data Cleansing" by Heiko Müller and Johann-Christoph Freytag

Questions about Strawberries

- Where they are grown? By whom?
- Are they really loaded with carcinogenic poisons?
- Are they really good for your health? Bad for your health?
- Are organic strawberries carriers of deadly diseases?
- When I go to the market should I buy conventional or organic strawberries?
- Do Strawberry farmers make money?
- How do the strawberries I buy get to my market?

Strawberry data source and parameters

The data set for this assignment has been selected from:

[USDA_NASS_strawb_2024SEP25.

The data have been stored on NASS here: USDA_NASS_strawb_2024SEP25 .

For the assignment, I stored the csv I downloaded on the MA615 Blackboard as strawberries25_v3.csv.

The data was originally collected at the county, state, and national levels, but the degree of missingness at the state level was too high, so I dropped the county-level data.

There are 5,359 rows and 21 column in the initial data set. The only complete year is 2022, although there is data for years 2018 through 2024.

To work with the data, define a function to remove columns with only single value in all its rows.

To work with this data, split the Census data from the Survey data.

Census data cleaning and organizing

we're examining census data because it's different from survey data

we're isolating organic data

Note that straw_cen has only one year: 2022

Current stats Census date has been isolated and split between Organic and Conventional strawberries

imputed values

We'll start with missing value interpolation for census data:

We will first interpolate the Bearing "Grown" values. Then we can interpolate Bearing and Non-bearing from those values as we discussed in lecture.

Now if GROWN is NA, there are three possibilities:

1 Bearing and Non-bearing are both NA
2 Neither Bearing nor Non-bearing are NA
3 One of Bearing nor Non-bearing are NA

The solution for each of these cases is different:

1 Interpolate from the total for area grown(a check of the dataframe shows you that the total is never NA). We'll try and calculate the average ratio of the missing parts across the non-missing data and split the difference from the total across those
2 Sum Bearing and non-bearing. Easy.

3 Calculate the average proportion of grown that bearing or non bearing(as the case may be) is for the state in question. Then we use that to figure out the value.

Applying our functions now to strawberry census:

Now that we have imputed the values, we still have an adjustment to make. Because we are using proportional averages, the values might no longer add up to our totals. So we need to write a function that will scale our imputed values so that the totals add up.

The happy news is because we are doing this by index, we still know which are the values we have imputed and therefore which values we need to scale.

Imputing acres bearing and non-bearing

We can go one better with imputation where we impute case 1 for the "GROWN" categories together with other case 1s for each. A slight modification to the function above plus some other changes achieves this.

CENSUS TABLES

Survey data cleaning and organizing

Shift data into alignment function

Examine Domain

now look at totals

there are two markets for Strawberries – Fresh Marketing and Processing

make a table for each

from the Survey Totals

we have reports for

Markets: Fresh and Processing Operations: Growing and Production

Florida - California - 2018 -2023

California and Florida chemicals

California and Florida fertilizers

Filter FUNGICIDE (fungicide) data and remove missing values

```
survey_d_fung <- survey_d_chem %>%
  filter(type == "FUNGICIDE") %>%
  filter(Value != "(D)" & Value != "(NA)")
```

Filter data for AZOXYSTROBIN

```
survey_d_AZOXYSTROBIN <- survey_d_fung %>%
  filter(chem_name == "AZOXYSTROBIN")

# Display AZOXYSTROBIN usage in different states
print(survey_d_AZOXYSTROBIN)
```

```
# A tibble: 30 × 10
    Year State      mk1     mk2   measure other type  chem_name chem_index Value
   <dbl> <chr>      <chr>   <chr> <chr>   <chr> <chr> <chr>     <chr>      <chr>
 1  2023 CALIFORNIA BEARING APPL… LB      <NA>  FUNG… AZOXYSTR… 128810     3,300
 2  2023 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… AZOXYSTR… 128810     0.234
 3  2023 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… AZOXYSTR… 128810     0.326
 4  2023 CALIFORNIA BEARING APPL… NUMBER  AVG   FUNG… AZOXYSTR… 128810     1.4
 5  2023 CALIFORNIA BEARING TREA… PCT OF… AVG   FUNG… AZOXYSTR… 128810     24
 6  2023 FLORIDA    BEARING APPL… LB      <NA>  FUNG… AZOXYSTR… 128810     100
 7  2023 FLORIDA    BEARING APPL… LB / A… AVG   FUNG… AZOXYSTR… 128810     0.151
 8  2023 FLORIDA    BEARING APPL… LB / A… AVG   FUNG… AZOXYSTR… 128810     0.214
 9  2023 FLORIDA    BEARING APPL… NUMBER  AVG   FUNG… AZOXYSTR… 128810     1.4
10  2023 FLORIDA    BEARING TREA… PCT OF… AVG   FUNG… AZOXYSTR… 128810     3
# ℹ 20 more rows
```

```
# Discussion on the environmental impact of AZOXYSTROBIN
# Load usage amount of AZOXYSTROBIN

azoxystrobin_use <- survey_d_AZOXYSTROBIN %>%
  group_by(State) %>%
  summarize(avg_use = mean(as.numeric(Value), na.rm = TRUE))
```

```
Warning: There were 2 warnings in `summarize()`.
The first warning was:
ℹ In argument: `avg_use = mean(as.numeric(Value), na.rm = TRUE)`.
ℹ In group 1: `State = "CALIFORNIA"`.
Caused by warning in `mean()`:
! NAs introduced by coercion
ℹ Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```
# Print average usage in each state
print(azoxystrobin_use)
```

```
# A tibble: 2 × 2
  State      avg_use
  <chr>        <dbl>
1 CALIFORNIA    5.90
2 FLORIDA      15.9
```

```
# Top three fungicides used in California strawberries
survey_ca_fung_order <- survey_d_fung %>%
  filter(State == "CALIFORNIA", measure == "LB / ACRE / YEAR") %>%
  arrange(desc(as.numeric(Value)))

# Print the top three fungicides used in California

head(survey_ca_fung_order, 3)
```

```
# A tibble: 3 × 10
   Year State      mk1     mk2    measure other type  chem_name chem_index Value
  <dbl> <chr>      <chr>   <chr>  <chr>   <chr> <chr> <chr>     <chr>      <chr>
1  2023 CALIFORNIA BEARING APPLI… LB / A… AVG   FUNG… SULFUR    77501      38.7…
2  2019 CALIFORNIA BEARING APPLI… LB / A… AVG   FUNG… SULFUR    77501      26.6…
3  2019 CALIFORNIA BEARING APPLI… LB / A… AVG   FUNG… FOSETYL-… 123301     18.1…
```

```
# Filter usage of Fosetyl-AL and Potassium Bicarbonate

fosetyl_al_use <- survey_d_fung %>%
  filter(chem_name == "FOSETYL-AL")

potassium_bicarbonate_use <- survey_d_fung %>%
  filter(chem_name == "POTASSIUM BICARBON")


# Print usage of Fosetyl-AL and Potassium Bicarbonate

print(fosetyl_al_use)
```

```
# A tibble: 5 × 10
   Year State      mk1     mk2    measure other type  chem_name chem_index Value
  <dbl> <chr>      <chr>   <chr>  <chr>   <chr> <chr> <chr>     <chr>      <chr>
1  2019 CALIFORNIA BEARING APPLI… LB      <NA>  FUNG… FOSETYL-… 123301     34,2…
2  2019 CALIFORNIA BEARING APPLI… LB / A… AVG   FUNG… FOSETYL-… 123301     3.997
3  2019 CALIFORNIA BEARING APPLI… LB / A… AVG   FUNG… FOSETYL-… 123301     18.1…
4  2019 CALIFORNIA BEARING APPLI… NUMBER  AVG   FUNG… FOSETYL-… 123301     4.5
5  2019 CALIFORNIA BEARING TREAT… PCT OF… AVG   FUNG… FOSETYL-… 123301     5
```

```
print(potassium_bicarbonate_use)
```

```
# A tibble: 0 × 10
# ℹ 10 variables: Year <dbl>, State <chr>, mk1 <chr>, mk2 <chr>, measure <chr>,
#   other <chr>, type <chr>, chem_name <chr>, chem_index <chr>, Value <chr>
```

```
# Filter usage of Captan and Thiram

captan_use <- survey_d_fung %>%
  filter(chem_name == "CAPTAN")

thiram_use <- survey_d_fung %>%
  filter(chem_name == "THIRAM")


# Print usage of Captan and Thiram

print(captan_use)
```

```
# A tibble: 40 × 10
    Year State      mk1     mk2   measure other type  chem_name chem_index Value
   <dbl> <chr>      <chr>   <chr> <chr>   <chr> <chr> <chr>     <chr>      <chr>
 1  2023 CALIFORNIA BEARING APPL… LB      <NA>  FUNG… CAPTAN    81301      603,…
 2  2023 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… CAPTAN    81301      1.693
 3  2023 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… CAPTAN    81301      15.9…
 4  2023 CALIFORNIA BEARING APPL… NUMBER  AVG   FUNG… CAPTAN    81301      9.4
 5  2023 CALIFORNIA BEARING TREA… PCT OF… AVG   FUNG… CAPTAN    81301      88
 6  2023 FLORIDA    BEARING APPL… LB      <NA>  FUNG… CAPTAN    81301      144,…
 7  2023 FLORIDA    BEARING APPL… LB / A… AVG   FUNG… CAPTAN    81301      2.012
 8  2023 FLORIDA    BEARING APPL… LB / A… AVG   FUNG… CAPTAN    81301      10.5…
 9  2023 FLORIDA    BEARING APPL… NUMBER  AVG   FUNG… CAPTAN    81301      5.2
10  2023 FLORIDA    BEARING TREA… PCT OF… AVG   FUNG… CAPTAN    81301      96
# ℹ 30 more rows
```

```
print(thiram_use)
```

```
# A tibble: 40 × 10
    Year State      mk1     mk2   measure other type  chem_name chem_index Value
   <dbl> <chr>      <chr>   <chr> <chr>   <chr> <chr> <chr>     <chr>      <chr>
 1  2023 CALIFORNIA BEARING APPL… LB      <NA>  FUNG… THIRAM    79801      269,…
 2  2023 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… THIRAM    79801      2.201
 3  2023 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… THIRAM    79801      8.873
 4  2023 CALIFORNIA BEARING APPL… NUMBER  AVG   FUNG… THIRAM    79801      4
 5  2023 CALIFORNIA BEARING TREA… PCT OF… AVG   FUNG… THIRAM    79801      70
 6  2023 FLORIDA    BEARING APPL… LB      <NA>  FUNG… THIRAM    79801      112,…
 7  2023 FLORIDA    BEARING APPL… LB / A… AVG   FUNG… THIRAM    79801      2.156
 8  2023 FLORIDA    BEARING APPL… LB / A… AVG   FUNG… THIRAM    79801      12.4…
 9  2023 FLORIDA    BEARING APPL… NUMBER  AVG   FUNG… THIRAM    79801      5.8
10  2023 FLORIDA    BEARING TREA… PCT OF… AVG   FUNG… THIRAM    79801      63
# ℹ 30 more rows
```

```
# Discussion on the dual role of sulfur in fungicides and fertilizers
# Filter sulfur usage in fertilizers

survey_d_sulfur <- survey_d_fert %>%
  filter(chem_name == "SULFUR")


# Filter sulfur usage in fungicides

survey_f_sulfur <- survey_d_fung %>%
  filter(chem_name == "SULFUR")


# Print sulfur usage data

print(survey_d_sulfur)
```

```
# A tibble: 25 × 7
    Year State      mk2          measure                  other chem_name Value
   <dbl> <chr>      <chr>        <chr>                    <chr> <chr>     <chr>
 1  2023 CALIFORNIA APPLICATIONS LB                       <NA>  SULFUR    (D)
 2  2023 CALIFORNIA APPLICATIONS LB / ACRE / APPLICATION AVG   SULFUR    (D)
 3  2023 CALIFORNIA APPLICATIONS LB / ACRE / YEAR        AVG   SULFUR    (D)
 4  2023 CALIFORNIA APPLICATIONS NUMBER                   AVG   SULFUR    (D)
 5  2023 CALIFORNIA TREATED      PCT OF AREA BEARING      AVG   SULFUR    (D)
 6  2023 FLORIDA    APPLICATIONS LB                       <NA>  SULFUR    (D)
 7  2023 FLORIDA    APPLICATIONS LB / ACRE / APPLICATION AVG   SULFUR    (D)
 8  2023 FLORIDA    APPLICATIONS LB / ACRE / YEAR        AVG   SULFUR    (D)
 9  2023 FLORIDA    APPLICATIONS NUMBER                   AVG   SULFUR    (D)
10  2023 FLORIDA    TREATED      PCT OF AREA BEARING      AVG   SULFUR    (D)
# ℹ 15 more rows
```

```
print(survey_f_sulfur)
```

```
# A tibble: 20 × 10
    Year State      mk1     mk2   measure other type  chem_name chem_index Value
   <dbl> <chr>      <chr>   <chr> <chr>   <chr> <chr> <chr>     <chr>      <chr>
 1  2023 CALIFORNIA BEARING APPL… LB      <NA>  FUNG… SULFUR    77501      1,25…
 2  2023 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… SULFUR    77501      5.471
 3  2023 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… SULFUR    77501      38.7…
 4  2023 CALIFORNIA BEARING APPL… NUMBER  AVG   FUNG… SULFUR    77501      7.1
 5  2023 CALIFORNIA BEARING TREA… PCT OF… AVG   FUNG… SULFUR    77501      75
 6  2021 CALIFORNIA BEARING APPL… LB      <NA>  FUNG… SULFUR    77501      591,…
 7  2021 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… SULFUR    77501      3.779
 8  2021 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… SULFUR    77501      17.2…
 9  2021 CALIFORNIA BEARING APPL… NUMBER  AVG   FUNG… SULFUR    77501      4.6
10  2021 CALIFORNIA BEARING TREA… PCT OF… AVG   FUNG… SULFUR    77501      88
11  2019 CALIFORNIA BEARING APPL… LB      <NA>  FUNG… SULFUR    77501      686,…
12  2019 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… SULFUR    77501      3.656
13  2019 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… SULFUR    77501      26.6…
14  2019 CALIFORNIA BEARING APPL… NUMBER  AVG   FUNG… SULFUR    77501      7.3
15  2019 CALIFORNIA BEARING TREA… PCT OF… AVG   FUNG… SULFUR    77501      72
16  2018 CALIFORNIA BEARING APPL… LB      <NA>  FUNG… SULFUR    77501      298,…
17  2018 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… SULFUR    77501      4.865
18  2018 CALIFORNIA BEARING APPL… LB / A… AVG   FUNG… SULFUR    77501      12.88
19  2018 CALIFORNIA BEARING APPL… NUMBER  AVG   FUNG… SULFUR    77501      2.6
20  2018 CALIFORNIA BEARING TREA… PCT OF… AVG   FUNG… SULFUR    77501      64
```

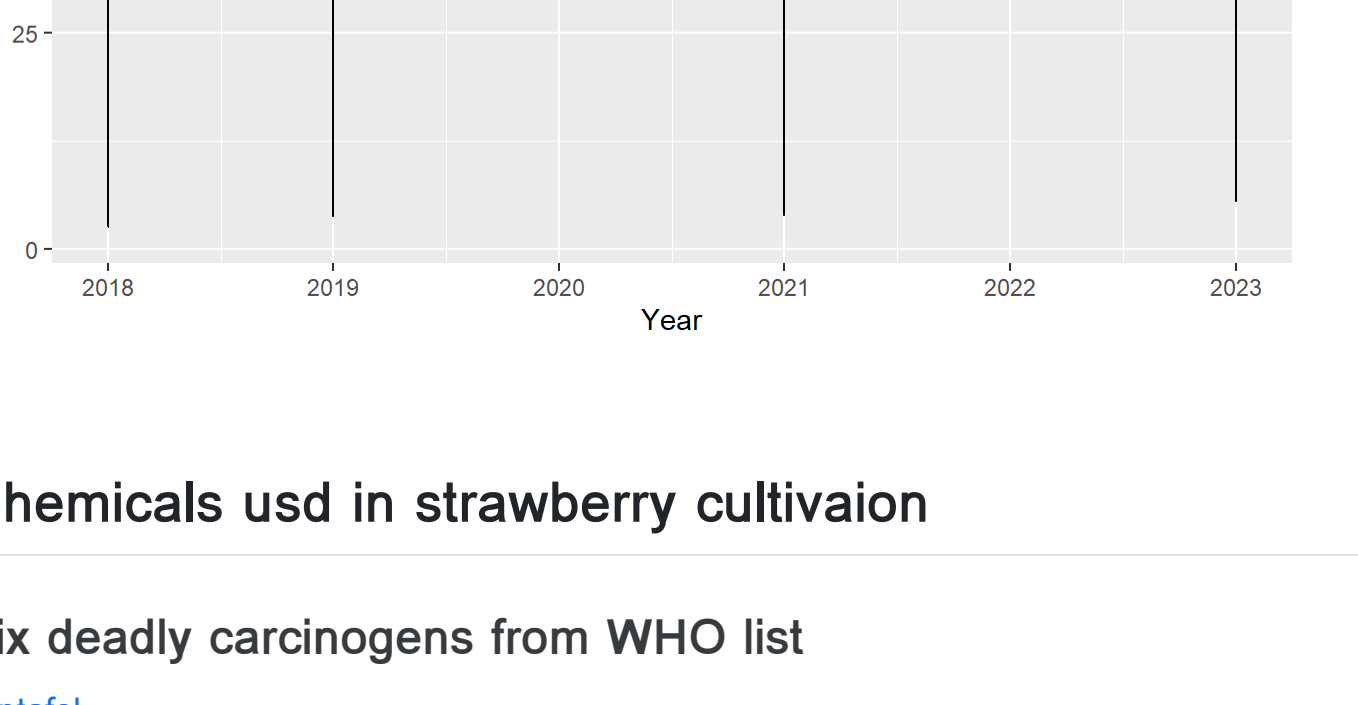
```
# Visualize sulfur usage in California

library(ggplot2)

ggplot(survey_f_sulfur, aes(x = Year, y = as.numeric(Value))) +
  geom_line() +
  labs(title = "Sulfur Usage in California", x = "Year", y = "Usage (LB/acre/year)")
```

```
Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_line()`).
```



Chemicals usd in strawberry cultivaion

Six deadly carcinogens from WHO list

captafol

ethylene dibromide also

glyphosate See also 1
2
3
4

malathion 1 2

diazinon 1 2 3

dichlorophenyltrichloroethane (DDT) 1 2 [3](https://www.epa.gov/ingredients-used-pesticide-products/ddt-brief-history-and-status#:~:text=DDT%20(dichloro%2Ddiphenyl%2Dtrichloroethane,both%20military%20and%20civilian%20populations.)

For contrast

Azadirachtin 1 2 3

Sources of agricultural chemical information

for EPA number lookup epa numbers

Active Pesticide Product Registration Informational Listing

CAS for Methyl Bromide

pesticide chemical search

toxic chemical dashboard

pubChem

The EPA PC (Pesticide Chemical) Code is a unique chemical code number assigned by the EPA to a particular pesticide active ingredient, inert ingredient or mixture of active ingredients.

Investigating toxic pesticides

start here with chem PC code

step 2 to get label (with warnings) for products using the chemical

Pesticide Product and Label System

Search by Chemical

CompTox Chemicals Dashboard

Active Pesticide Product Registration Informational Listing

OSHA chemical database

Pesticide Ingredients

NPIC Product Research Online (NPRO)

Databases for Chemical Information

Pesticide Active Ingredients

TSCA Chemical Substance Inventory

glyphosate