# Stawberries: exploratory data analysis

Zhan Gu

2023-10-11

## Assignment

Using our class discussions and this document as a starting point, produce an EDA report. The report should describe the data itself so that readers understand the data sources used in the report and how you cleaned and organized the data for analysis.

The sections below suggest how the report might be organized. The report should be succinct, communicating the information that you believe will be helpful to someone doing a fuller analysis of the data or using the data for model building. Implementation details should be included in commentary that is included in code.

Sections of the document as it was originally presented in class have been commented so that you can see them in the code.

### Data acquisition and assessment

- Data sources
- Assumptions and motivations

### Data cleaning and organization

Drop Single-Value Columns: Identify and drop columns with a single unique value.

Data Organization: Data is organized by state.

Separate Data by Program: Split the dataset into two data frames, strwb_census and strwb_survey, based on the "CENSUS" and "SURVEY" values in the "Program" column.

Split Composite Columns: Decompose columns such as 'Data Item' into separate meaningful columns.

Handle Missing Values: Replaced NA values and footnotes such as '(D)'.

## References

### Material about strawberries

WHO says strawberries may not be so safe for you–2017March16

Pesticides + poison gases = cheap, year-round strawberries 2019March20

Multistate Outbreak of Hepatitis A Virus Infections Linked to Fresh Organic Strawberries-2022March5

Strawberry makes list of cancer-fighting foods-2023May31

### Technical references

In their handbook "An introduction to data cleaning with R" by Edwin de Jonge and Mark van der Loo, de Jonge and van der Loo go into detail about specific data cleaning isssues and how to handle them in R.

"Problems, Methods, and Challenges in Comprehensive Data Cleansing" by Heiko Müller and Johann-Christoph Freytag is a good companion to the de Jonge and van der Loo handbook, offering additional insights.

## The data

The data set for this assignment has been selected from: USDA_NASS The data have been stored on NASS here: USDA_NASS_strawb_2023SEP19

Make relevant observations in the document and in your code about data. Add commentary to the code so that anthoer analysts could use or extend your code.

Discusss missing data, inclding how you handled it. Be careful to point out where NA's are being produced during processing and are not data missing in the original data.

Where it is relevant, include information of how you have organized the data for analysis. It might, for example, be helpful to know that there is both agricultural census data and survey data. It might be helpful to discuss data that appears to be redundant between these two sources.

Make sure you include details in your discussion and in your code about other data and information you used in your work. Cite sources and provide detail that would allow another analyst to reproduce your work.

Table 1: Dropped Single-Value Columns: names and values

| col_name | col_val |
|---|---|
| Week Ending | NA |
| Geo Level | STATE |
| Ag District | NA |
| Ag District Code | NA |
| County | NA |
| County ANSI | NA |
| Zip Code | NA |
| Region | NA |
| watershed_code | 00000000 |
| Watershed | NA |
| Commodity | STRAWBERRIES |

```
Rows: 4,314
Columns: 21
$ Program            <chr> "CENSUS", "CENSUS", "CENSUS", "CENSUS", "CENSUS", "~
$ Year               <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 202~
$ Period             <chr> "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YE~
$ `Week Ending`      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `Geo Level`        <chr> "STATE", "STATE", "STATE", "STATE", "STATE", "STATE~
$ State              <chr> "ALASKA", "ALASKA", "ALASKA", "ALASKA", "ALASKA", "~
$ `State ANSI`       <chr> "02", "02", "02", "02", "02", "02", "02", "06", "06~
$ `Ag District`      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `Ag District Code` <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ County             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `County ANSI`      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `Zip Code`         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Region             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ watershed_code     <chr> "00000000", "00000000", "00000000", "00000000", "00~
$ Watershed          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Commodity          <chr> "STRAWBERRIES", "STRAWBERRIES", "STRAWBERRIES", "ST~
$ `Data Item`        <chr> "STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES", "S~
$ Domain             <chr> "ORGANIC STATUS", "ORGANIC STATUS", "ORGANIC STATUS~
$ `Domain Category`  <chr> "ORGANIC STATUS: (NOP USDA CERTIFIED)", "ORGANIC ST~
$ Value              <chr> "2", "(D)", "(D)", "(D)", "2", "(D)", "(D)", "142",~
$ `CV (%)`           <chr> "(H)", "(D)", "(D)", "(D)", "(H)", "(D)", "(D)", "1~
```

```
Rows: 4,314
Columns: 10
```

```
$ Program           <chr> "CENSUS", "CENSUS", "CENSUS", "CENSUS", "CENSUS", "C~
$ Year              <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021~
$ Period            <chr> "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEA~
$ State             <chr> "ALASKA", "ALASKA", "ALASKA", "ALASKA", "ALASKA", "A~
$ `State ANSI`      <chr> "02", "02", "02", "02", "02", "02", "02", "06", "06"~
$ `Data Item`       <chr> "STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES", "ST~
$ Domain            <chr> "ORGANIC STATUS", "ORGANIC STATUS", "ORGANIC STATUS"~
$ `Domain Category` <chr> "ORGANIC STATUS: (NOP USDA CERTIFIED)", "ORGANIC STA~
$ Value             <chr> "2", "(D)", "(D)", "(D)", "2", "(D)", "(D)", "142", ~
$ `CV (%)`          <chr> "(H)", "(D)", "(D)", "(D)", "(H)", "(D)", "(D)", "19~
```

```
[1] "Every row has value in the State column."
```

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

## EDA

Once the data has been cleaned and organized, you must conduct your own EDA. Be sure to include a discussion of your analysis of the chemical information, including citations for data and other information you have used. Visualizations should play a key role in your analysis. Plots should be labeled and captioned.

```
Rows: 3,450
Columns: 10
$ Program           <chr> "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SURVEY", "S~
$ Year              <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022~
$ Period            <chr> "MARKETING YEAR", "MARKETING YEAR", "MARKETING YEAR"~
$ State             <chr> "CALIFORNIA", "CALIFORNIA", "CALIFORNIA", "FLORIDA",~
$ `State ANSI`      <chr> "06", "06", "06", "12", "12", "12", NA, NA, NA, "06"~
$ `Data Item`       <chr> "STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT"~
$ Domain            <chr> "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL"~
$ `Domain Category` <chr> "NOT SPECIFIED", "NOT SPECIFIED", "NOT SPECIFIED", "~
$ Value             <chr> "108", "(D)", "(D)", "169", "(D)", "(D)", "0", "135"~
$ `CV (%)`          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

Chemical Usage in Strawberry Production Strawberries, like many crops, may undergo various chemical treatments to ensure their growth, quality, and protection against pests. However, some of these chemicals may be harmful or even carcinogenic. It's essential to analyze which chemicals are used.

```
# A tibble: 5 x 10
   Year State      `State ANSI` temp3    temp4 temp23 temp43 temp44 temp45 Value
  <dbl> <chr>      <chr>        <chr>    <chr> <chr>  <chr>  <chr>  <chr>  <chr>
1  2021 CALIFORNIA 06           " MEASU~  <NA> " HER~ " HER~ <NA>   <NA>   (D)
2  2021 CALIFORNIA 06           " MEASU~  <NA> " HER~ " HER~ <NA>   <NA>   (D)
3  2021 CALIFORNIA 06           " MEASU~ " AV~ " HER~ " HER~ <NA>   <NA>   (D)
4  2021 CALIFORNIA 06           " MEASU~ " AV~ " HER~ " HER~ <NA>   <NA>   (D)
5  2021 CALIFORNIA 06           " MEASU~ " AV~ " HER~ " HER~ <NA>   <NA>   (D)
```

**These references have been left in the document to help while you are writing. Cite those you use and drop the rest from the final document.**

NASS help

Quick Stats Glossary

Quick Stats Column Definitions

stats by subject

for EPA number lookup epa numbers

Active Pesticide Product Registration Informational Listing

pc number input pesticide chemical search

toxic chemical dashboard

ACToR – Aggregated Computational Toxicology Resource

comptox 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

International Chemical safety cards

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