Strawberries in the United States: Exploratory Data Analysis

GRS 615: Data Science in R

Nicole Kingdon

October 18, 2023

Introduction

Strawberries & Positive Health

Strawberries are a fruit that holds several minerals, vitamins, and nutrients (CRAIG 1997), which have positive implications on human health (Afrin et al. 2016). Specifically, strawberries have been found to help reduce likelihood of cancer, diabetes, obesity, neurodegeneration, cardiovascular disease, and metabolic syndrome (see Figure 1) (Afrin et al. 2016). Although strawberries as a healthy food is the norm, pesticides appear to be harming the beneficial factors of this fruit.

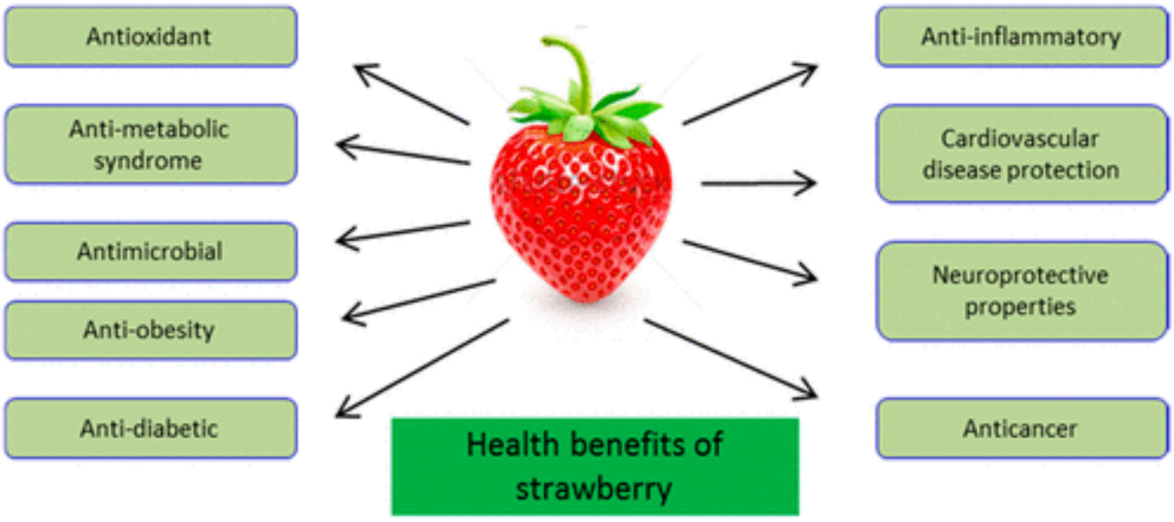

Figure 1: Health benefits of strawberries (Afrin et al., 2016)

Strawberries & Pesticides

Pesticides are used on fruit and vegetable crops, including strawberries, with hopes to increase the quantity (Fenik, Tankiewicz, and Biziuk 2011). Pesticides are made of chemical compounds to reduce or completely eliminate pests from impacting crops (Afrin et al. 2016). These chemical compounds may increase the yield of the crop, but may have a large risk on human health. Additionally, they may contaminate bodies of water and soil with the chemicals, help pests develop resistance to the chemicals, and impact helpful organisms from persisting in areas where pesticides are used. Overall, there are positive and negative impacts of using pesticides (see Figure 2), but it is important to further examine these impacts, specifically on strawberries, to understand the implications of using such.

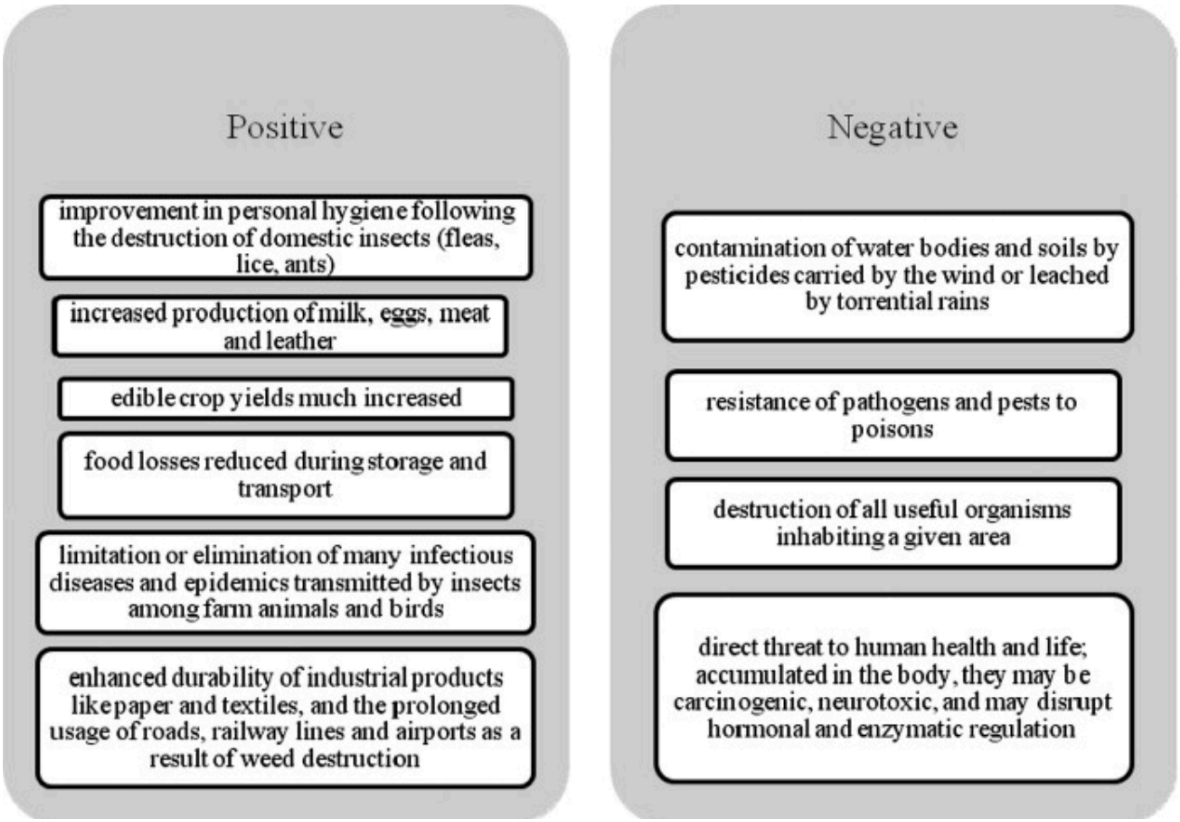

Figure 2: Positive and negative implications of pesticides on fruit and vegetable crops (Fenik et al., 2011)

Organic Strawberries

Strawberries that use alternative and safer methods to production are considered organic (Verteramo Chiu and Gomez 2023). Along with these alternative methods comes a higher cost

of production and care, which increases the price to the consumer.

Analyzing Strawberry Production

The literature varies on if organic or non-organic is environmentally better (Afrin et al. 2016). This exploratory data analysis on production of strawberries, which encompasses both processed (non-organic) and fresh (organic) market data, will help us better understand production of strawberries in the United States.

Data Acquisition & Assessment

USDA-NASS Data

The data was acquired from U.S. Department of Agriculture (USDA) and the National Agricultural Statistics Service (NASS). The data was uploaded for data cleaning and organizing and exploratory data analysis by Professor Haviland Wright, who chose the following data: USDA-NASS.

The data frame uploaded to R is titled `strawberry` (see below).

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

```
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~
```

Census Data

The data offers census data based on state that represents fresh market (organic) and process market (non-organic) sales.

```
Rows: 864
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~
```

Survey Data

Additionally, it holds survey information for each state, specifically indicating pesticides and bacterium used to preserve strawberry crop yield. In addition, it offers fresh and process market data.

```
Rows: 3,450
Columns: 21
$ Program            <chr> "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SURVEY", "~
$ Year               <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 202~
$ Period             <chr> "MARKETING YEAR", "MARKETING YEAR", "MARKETING YEAR~
$ `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> "CALIFORNIA", "CALIFORNIA", "CALIFORNIA", "FLORIDA"~
$ `State ANSI`       <chr> "06", "06", "06", "12", "12", "12", NA, NA, NA, "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 - 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,~
```

States

There were 47 states (c("ALASKA", "CALIFORNIA", "CONNECTICUT", "FLORIDA", "GEORGIA", "IDAHO", "ILLINOIS", "INDIANA", "IOWA", "KENTUCKY", "LOUISIANA", "MAINE", "MARYLAND", "MASSACHUSETTS", "MICHIGAN", "MINNESOTA", "MONTANA", "NEBRASKA", "NEW HAMPSHIRE", "NEW JERSEY", "NEW YORK", "NORTH CAROLINA", "OHIO", "OKLAHOMA", "OREGON", "PENNSYLVANIA", "RHODE ISLAND", "SOUTH CAROLINA", "SOUTH DAKOTA", "TENNESSEE", "VERMONT", "WASHINGTON", "WEST VIRGINIA", "WISCONSIN", "ALABAMA", "ARIZONA", "COLORADO", "KANSAS", "MISSOURI", "NEVADA", "NEW MEXICO", "VIRGINIA", "ARKANSAS", "NORTH DAKOTA", "TEXAS", "UTAH", "OTHER STATES")) with two states considered as "other states".

Years

The data was from the years c(2021, 2019, 2016, 2022, 2020, 2018, 2017).

Assumptions & Motivations

Census Data

The `census` data was a nation-wide collection of data about the fresh and process markets related to strawberries. This data has values that are indicated as `(D)`, which are data that was withheld upon request by the strawberry market in that particular state. This could leave out important information in the data.

Survey Data

The `survey` data was collected via a survey sent out to each state in the United States. There were only 11 out of 47 states who returned the survey (c("CALIFORNIA", "FLORIDA", "OTHER STATES", "NEW YORK", "NORTH CAROLINA", "OREGON", "WASHINGTON", "MICHIGAN", "OHIO", "PENNSYLVANIA", "WISCONSIN")), which includes the "other states". (The "other states" did not have any data relating to pesticides and bacterium.) This is only a 23% response rate, which is not comprehensive of all the states and the entire United States process market. The states that did return the survey will still be able to show a report of pesticide and bacterium usage on their processed strawberry crops.

Data Cleaning & Organizing

R Packages

The following R packages were used to clean and organize the data:

```
library(knitr)
library(kableExtra)
library(tidyverse)
library(stringr)
library(dplyr)
```

Organization

The data was organized into two data frames: `census` and `survey`. The `census` data frame was cleaned and organized to show fresh and process market sales, and the `survey` data frame was prepared to show pesticide and bacterium data.

Cleaning

Initial Cleaning

The following initial data cleaning derived from Professor Wright.

Removed columns with a single value in all columns

Is every line associated with a state?

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

The data is organized by state. The state with the most rows is CALIFORNIA.

Examine the California data

List of the composite columns

Census: Data Item, Domain Category

Survey: Data Item, Domain, Domain Category

Separating Data Frames

The following separation of data frames derived from Professor Wright.

The two new data frames are as follows: `strwb_census`, which holds all the `CENSUS` rows, and `strwb_survey`, which holds all the `SURVEY` rows.

Census

After splitting `CENSUS` and `SURVEY` rows into two data frames, Professor Wright has first organized the CENSUS data.

Separated composite columns and cleaned the Value column.

Composite columns in the `strwb_census`: Data Item, Domain category

Column separators in CENSUS: ",", "-", ":"

Separated `Data Item` into columns by ",".

Created a "Process Market" column.

Removed NA's from prop_acct, temp2, and temp3.

Combined temp2 with temp3 to create Metric column. Removed parts of string that did not matter. Relocated columns.

The Value column was transformed.

I finished cleaning and organizing the `strwb_census` data frame, which is detailed below.

First, I selected particular columns that had necessary data.

Next, I removed the "," from the `Value` column and transformed them into numeric values. This introduced rows with `NA` values.

After that, I cleaned up the `CV (%)` column by changing the values to numbers, instead of strings. This also introduced rows with `NA` values.

Furthermore, I am going to omit all rows in the `Value` and `CV (%)` columns with `NA` values. These are being omitted because they hold no meaning. Only the `Value` column had to be adjusted, as the `CV (%)` column did not have any values if the `Value` column also did not.

Finally, to complete the `strwb_census` cleaning and organizing, I am going to arrange the `State` column to be in ascending order.

Survey

Professor Wright had organized the SURVEY data frame splitting the marketing, and production data from the chemical application data. In the strawberry data frame, The `CENSUS` rows contain marketing, sales, and production data. The `SURVEY` rows contain rows which may be redundant with the CENSUS rows and chemical application rows. These rows contain fresh and process market sales data, which have been removed.

Began cleaning and organizing `strwb_survey`.

Drop one-value columns in `strwb_survey_chem`.

Dropped one-value columns in `strwb_survey_mkt`.

Survey: Market

Now, I further cleaned up both `strwb_survey_mkt` and `strwb_survey_chem`. First, I worked with the `strwb_survey_mkt`.

To begin, I made the `Value` column into numeric values, which introduced `NA` values. Then, I changed the format of the numeric values, so they did not appear in scientific notation.

To reduce the duplicity of `strwb_survey_mkt` with `strwb_census`, I separated the fresh and process market data in `strwb_survey_mkt` from the data on the pesticide application. With this, I created two new data frames: `strwb_survey_mkt` and `strwb_survey_pest`.

Survey: Chemical

Finally, I finished cleaning and organizing `strwb_survey_chem`.

First, I selected the relevant columns.

From here, I split up `temp43`.

Furthermore, I cleaned up both of the new `chemical` and the `PC#` columns by removing the unnecessary parentheses.

I also want to separate the `chemical` column by chemical type and chemical name.

Trimmed off the spaces on both sides of the `chemical` column.

Trimmed the same to `chem_type`.

Finally, trimmed the same with `PC#`.

Now to remove the rows with `NA` values in the `PC#` column.

Addition to Strwb_Survey_Chem: WHO Chemical Toxicity

Now, to better understand the `strwb_survey_chem` data, I added two columns of data. We are first going to begin with "chemical toxicity" (`toxicity`).

To gather this information, I used the World Health Organization (WHO)'s classification of pesticides by hazard [WHO, 2019]. Table 1 shows the WHO's toxicity classification for pesticides.

Additionally, some chemicals are presented as fatal or toxic if inhaled, as they are gaseous or volatile fumigants [WHO, 2019]. Others are classified as "unlikely to present acute hazard" by WHO [WHO, 2019], which means that they will not present any hazard if used properly. Furthermore, they are also classified as "no significant acute toxicity" when they are not in the WHO classification and are found to be non-toxic, mostly discovered through the Environmental Protection Agency (EPA) [EPA]

I searched through each table to find each pesticide. Not all pesticides were on the table. To find the missing chemical toxicities, I used the large language model, Chat GPT [Chat GPT], found information through EPA Pesticide Fact Sheets [EPA-Acibenzolar], [EPA-Ammonium.], [EPA-Aureobasidium.], [EPA-Canola], [EPA-Caprylic], [EPA-Capasaicin], [EPA-Clethodim], [EPA-Cyfluefenamid], [EPA-Cytokinin], [EPA-Indole.], [EPA-Iron-Phos.], [EPA-Mefenoxam], [EPA-Metam-sodium], [EPA-Polyoxin], [EPA-Potassium], [EPA-Potassium-salts], [EPA-Potassium-silicate], [EPA-Spiromedifen], [EPA-STREPTOMYCES-LYDICUS], [EPA-SULFENTRAZONE], [EPA-SULFUR], and in other sources (Kilani-Morakchi, Morakchi-Goudjil, and Sifi 2021), [PubChem-Capric], [Carfentrazone-ethyl], [Pub-Chem-Copper.], [Copper-Octanoate], [ACS-Copper-Oxide], [Cyprodinil], [DECYLDIMETHYLOCTYL], [Didecyl.], [Dodine], [Sodium-Ferric-Ethyl.], [Garlic-Oil], [Glyphosate], [Hydrogen-Peroxide], [Isofetamind], [Methoxyfenozide], [Mineral-oil], [Mono-Potassium-Salt], [Mustard-oil], [Peroxyacetic-acid], [Quinoline], [CDC-PYRACLOSTROBIN], [SOYBEAN-OIL]. The rest of

Class	Description	LD_{50} for the Rat (mg/kg Body Weight)			
		Oral		Dermal	
		Solids	Liquids	Solids	Liquids
Ia	Extremely hazardous	≤5	≤20	≤10	≤40
Ib	Highly hazardous	5–50	20–200	10–100	40–400
II	Moderately hazardous	50–500	200–2,000	100–1,000	400–4,000
III	Slightly hazardous	>500	>2,000	>1,000	>4,000

doi:10.1371/journal.pmed.1000357.t001

Figure 3: Table 1. World Health Organization Pesticide Toxicity Classification

the missing values were found to be bacterium [Chat GPT], which do not have a chemical toxicity, as well as `COPPER ETHANOLAMINE`, `KANTOR`, and `HALOSULFURON-METHYL`, which did not have any clear information on toxicity [Halosulfuron.] [Copper-Ethanolamine].

I classified each chemical toxicity by the string `Highly hazardous`, `Moderately hazardous`, `Unlikely to present acute hazard`, `Slightly hazardous`, `Fatal if inhaled`, `Toxic if inhaled`, `No significant acute toxicity` and `Not specified`, based on the WHO chemical toxicity rating system [WHO, 2019].

Addition to Strwb_Survey_Chem: CAS Registry Number

Now, the second added column represents each pesticide's CAS Registry Number (`CAS#`). A CAS Registry Number allows each chemical compound, including molecular formulas, chemical structures, generic, systematic, common, and trade names, to have a clear identification number [CAS].

To find the `CAS#`s of each pesticide, I used [WHO, 2019], [CHAT GPT], and, mainly, the United States Environmental Protection Agency's Pesticide Chemical Search [EPA-search]. After discovering that some of the rows in `chemical` are actually bacterium, all values that do not have `CAS#`s are indicated as `bacteria`. All the other chemicals were matched with their appropriate `CAS#`.

Quickly, I reorganized `strwb_survey_chem` to have the columns in a different order.

Websites

[CAS](https://www.cas.org/cas-data/cas-registry)

[EPA-search](https://ordspub.epa.gov/ords/pesticides/f?p=chemicalsearch:1)

[EPA Ammonium.](https://www3.epa.gov/pesticides/chem_search/reg_actions/registration/fs_PC-031802_01-Nov-06.pdf)

[WHO, 2019](https://www.who.int/publications/i/item/9789240005662)

[Wikipedia](https://en.wikipedia.org/wiki/Toxicity_class)

References

Afrin, Sadia, Massimiliano Gasparrini, Tamara Y. Forbes-Hernandez, Patricia Reboredo-Rodriguez, Bruno Mezzetti, Alfonso Varela-López, Francesca Giampieri, and Maurizio Battino. 2016. "Promising Health Benefits of the Strawberry: A Focus on Clinical Studies." *Journal of Agricultural and Food Chemistry* 64 (22): 4435–49. https://doi.org/10.1021/acs.jafc.6b00857.

CRAIG, WINSTON J. 1997. "Phytochemicals." *Journal of the American Dietetic Association* 97 (10): S199–204. https://doi.org/10.1016/s0002-8223(97)00765-7.

Fenik, Jolanta, Maciej Tankiewicz, and Marek Biziuk. 2011. "Properties and Determination of Pesticides in Fruits and Vegetables." *TrAC Trends in Analytical Chemistry* 30 (6): 814–26. https://doi.org/10.1016/j.trac.2011.02.008.

Kilani-Morakchi, Samira, Houda Morakchi-Goudjil, and Karima Sifi. 2021. "Azadirachtin-Based Insecticide: Overview, Risk Assessments, and Future Directions." *Frontiers in Agronomy* 3 (July). https://doi.org/10.3389/fagro.2021.676208.

Verteramo Chiu, Leslie J., and Miguel I. Gomez. 2023. "A Tale of Two Strawberries: Conventional and Organic Open-Field Production in California." *Sustainability* 15 (19): 14363. https://doi.org/10.3390/su151914363.