Final USDA NASS Data Cleaning

AUTHOR
Yangyu Chen

PUBLISHED
October 28, 2024

Preparing data for analysis

Acquire, explore, clean & structure, EDA

Data cleaning and organization

“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

Strawberries

Questions

Where they are grown? By whom? Strawberries are grown in multiple states and counties across the U.S. For example, the data includes information about strawberry cultivation in Alabama, specifically in Bullock County, and agricultural districts like Black Belt.

Are they really loaded with carcinogenic poisons? The dataset does not include information about pesticide or chemical use, so it is not possible to determine whether the strawberries contain carcinogenic substances.

Are they really good for your health? Bad for your health? There is no information in the dataset about the nutritional value or health impact of strawberries, so this question cannot be answered directly from the data.

Are organic strawberries carriers of deadly diseases? The dataset does not specify whether the strawberries are organic or contain any information regarding potential deadly diseases.

When I go to the market should I buy conventional or organic strawberries? There is no information in the dataset about whether the strawberries are organic, so no recommendation can be made.

Do Strawberry farmers make money? The data includes information about cultivated areas and operations, such as “ACRES BEARING” and “OPERATIONS WITH AREA GROWN,” but it does not provide any economic data to analyze the profitability of strawberry farmers.

How do the strawberries I buy get to my market? Strawberries are harvested, packed, and transported via cold chain logistics to keep them fresh until they reach the market.

The data

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

and has been stored on the blackboard as strawberries25_v3.csv.

- The data set is selected from: USDA_NASS
- There are many missing values in the dataset, so we will clean up the data later.

```
library(knitr)
library(kableExtra)
library(tidyverse)
library(stringr)
library(ggplot2)
library(esquisse)
library(RColorBrewer)
library(viridis)
library(treemap)
```

First let's read the data file and have a glimpse of the data.

```
Rows: 12,669
Columns: 21
$ Program            <chr> "CENSUS", "CENSUS", "CENSUS", "CENSUS", "CENSUS", "…
$ Year               <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 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> "COUNTY", "COUNTY", "COUNTY", "COUNTY", "COUNTY", "…
$ State              <chr> "ALABAMA", "ALABAMA", "ALABAMA", "ALABAMA", "ALABAM…
$ `State ANSI`       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ `Ag District`      <chr> "BLACK BELT", "BLACK BELT", "BLACK BELT", "BLACK BE…
$ `Ag District Code` <dbl> 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40,…
$ County             <chr> "BULLOCK", "BULLOCK", "BULLOCK", "BULLOCK", "BULLOC…
$ `County ANSI`      <dbl> 11, 11, 11, 11, 11, 11, 101, 101, 101, 101, 119, 11…
$ `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     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ 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 - ACRES BEARING", "STRAWBERRIES - ACR…
$ Domain             <chr> "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL…
$ `Domain Category`  <chr> "NOT SPECIFIED", "NOT SPECIFIED", "NOT SPECIFIED", …
$ Value              <chr> "(D)", "3", "(D)", "1", "6", "5", "(D)", "(D)", "2"…
$ `CV (%)`           <chr> "(D)", "15.7", "(D)", "(L)", "52.7", "47.6", "(D)",…
```

Then, let's look for the top 5 states that have the highest strawberry sales.

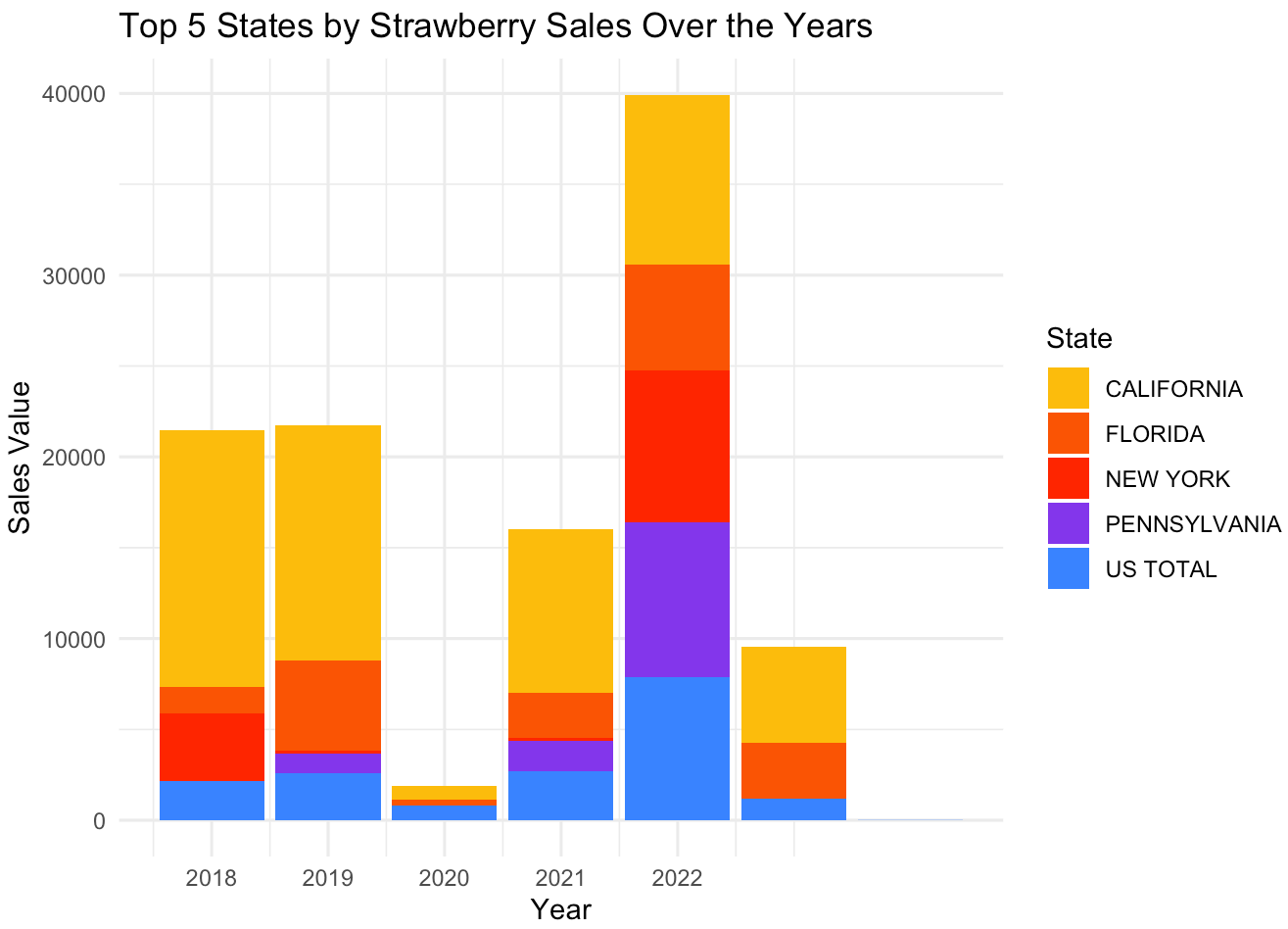
```
# Converts the 'Value' column to numeric type, handling non-numeric inputs
strawberry$Value <- as.numeric(as.character(strawberry$Value), na.rm = F)

# Group by 'State' and 'Year', then sum 'Value'
```

```
grouped <- strawberry |>
  group_by(State, Year) |>
  summarise(Value = sum(Value, na.rm=TRUE), .groups='drop')

# Find the top 5 states for total sales
top_5_states <- names(sort(tapply(grouped$Value, grouped$State, sum), decreasing
paste("Top 5 States by Strawberry Sales:", paste(top_5_states[1:5], collapse = ",
```

```
[1] "Top 5 States by Strawberry Sales: CALIFORNIA, FLORIDA, US TOTAL, NEW YORK,
PENNSYLVANIA"
```

Make a plot.

```
# Data for only the first five states are included
top_5_states_df <- grouped[grouped$State %in% top_5_states, ]

# Use ggplot2 to create a bar chart
ggplot(top_5_states_df, aes(x = Year, y = Value, fill = State)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    title = "Top 5 States by Strawberry Sales Over the Years",
    x = "Year",
    y = "Sales Value"
  ) +
  scale_x_continuous(breaks = seq(2016,2022,1)) +
  scale_fill_manual(values = c("#FFBE0B", "#FB5607", "#FF0000", "#8338EC", "#3A86
  theme(
    plot.title = element_text(hjust = 0.5, size = 20),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12)
  ) +
  theme_minimal()
```



#It can be observed that California consistently records very high sales values each year.

#Data cleaning and organization #During data cleaning, we start by removing columns with identical values in every row. Next, we divide the dataset into two separate DataFrames: one for CENSUS data and another for SURVEY data. Additionally, we clean the 'Value' column. To make the data more suitable for analysis, we also split the string values within the Data Item.

EDA

First, we would like to know about the use of chemicals. For example, we want to know which chemicals are commonly used. So we extract chemical name and code from strwb_survey_chem.

```
# Extract chemical name and code
strwb_survey_chem <- strwb_survey_chem %>%
  mutate(Chemical_Name = str_extract(temp43, "(?<=\\()(.*?)(?= =)"),
         Chemical_Code = str_extract(temp43, "(?<= = )(\\d+)"))
```

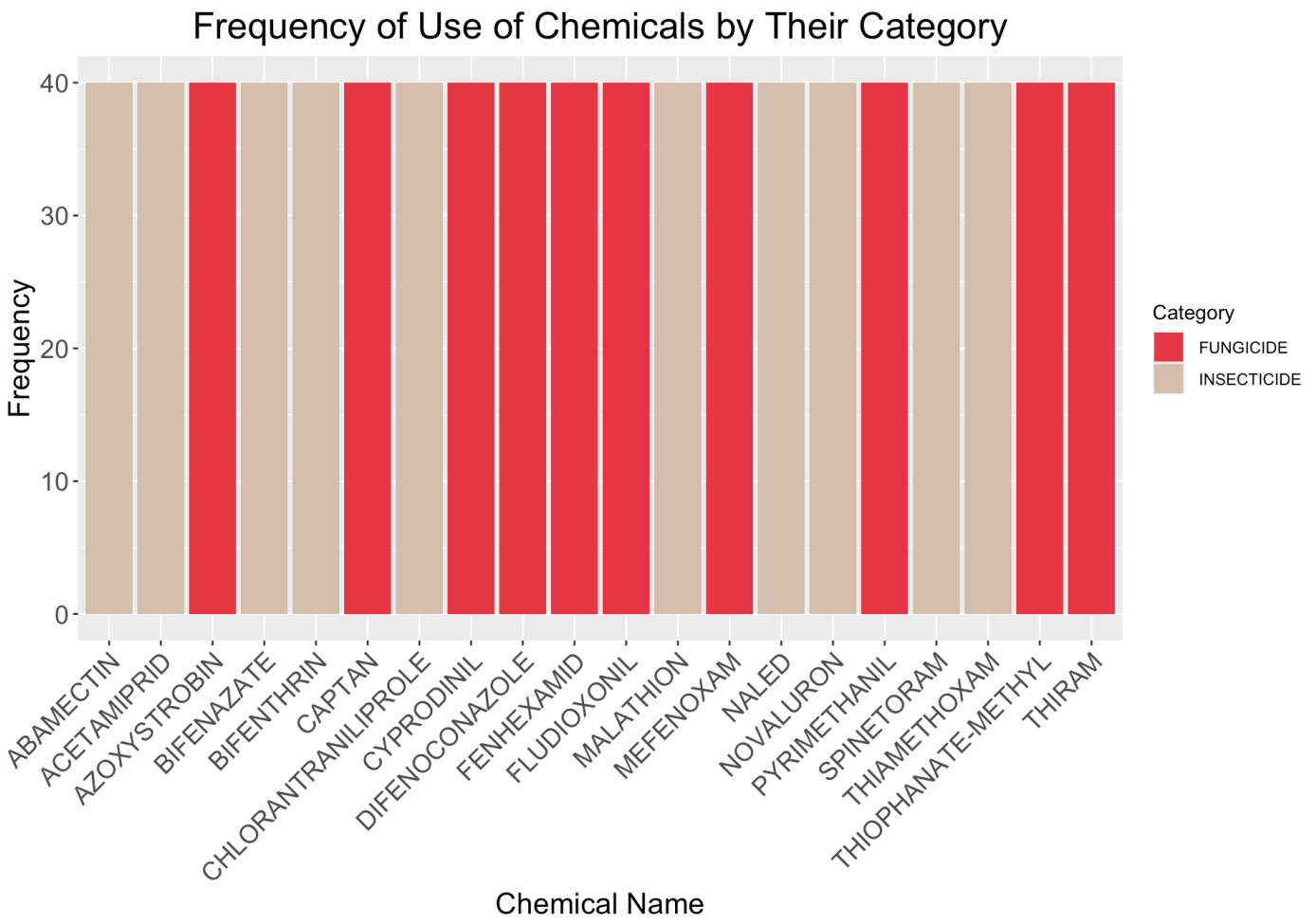
Because the units of measurement of chemicals are inconsistent, it is difficult to calculate the exact amount of use, but we can calculate the frequency of use of chemicals.

```
# Count the frequency of each unique chemical substance in the 'Chemical_Name' co
chemical_frequency <- table(strwb_survey_chem$Chemical_Name)
```

```r
# Sort the chemicals by their frequency from high to low and take the top 20
top_20_chemicals <- head(sort(chemical_frequency, decreasing = TRUE), 20)
filtered_strwb_survey_chem <- strwb_survey_chem[strwb_survey_chem$Chemical_Name %

# Rename
filtered_strwb_survey_chem <- filtered_strwb_survey_chem |>
  rename(Category = temp23)
```

Make a plot of Frequency of Use of Chemicals by Their Category.

```r
ggplot(filtered_strwb_survey_chem) +
  aes(x = reorder(Chemical_Name, -table(Chemical_Name)[Chemical_Name]), fill = Ca
  geom_bar() +
  labs(
    title = "Frequency of Use of Chemicals by Their Category",
    x = "Chemical Name",
    y = "Frequency"
  ) +
  scale_fill_manual(values = c("#E63946", "#D5BDAF", "#1D3557")) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 20),
    axis.title.x = element_text(size = 16),
    axis.title.y = element_text(size = 16),
    axis.text.x = element_text(size = 14, angle = 45,vjust = 1, hjust=1),
    axis.text.y = element_text(size = 14),
  )
```

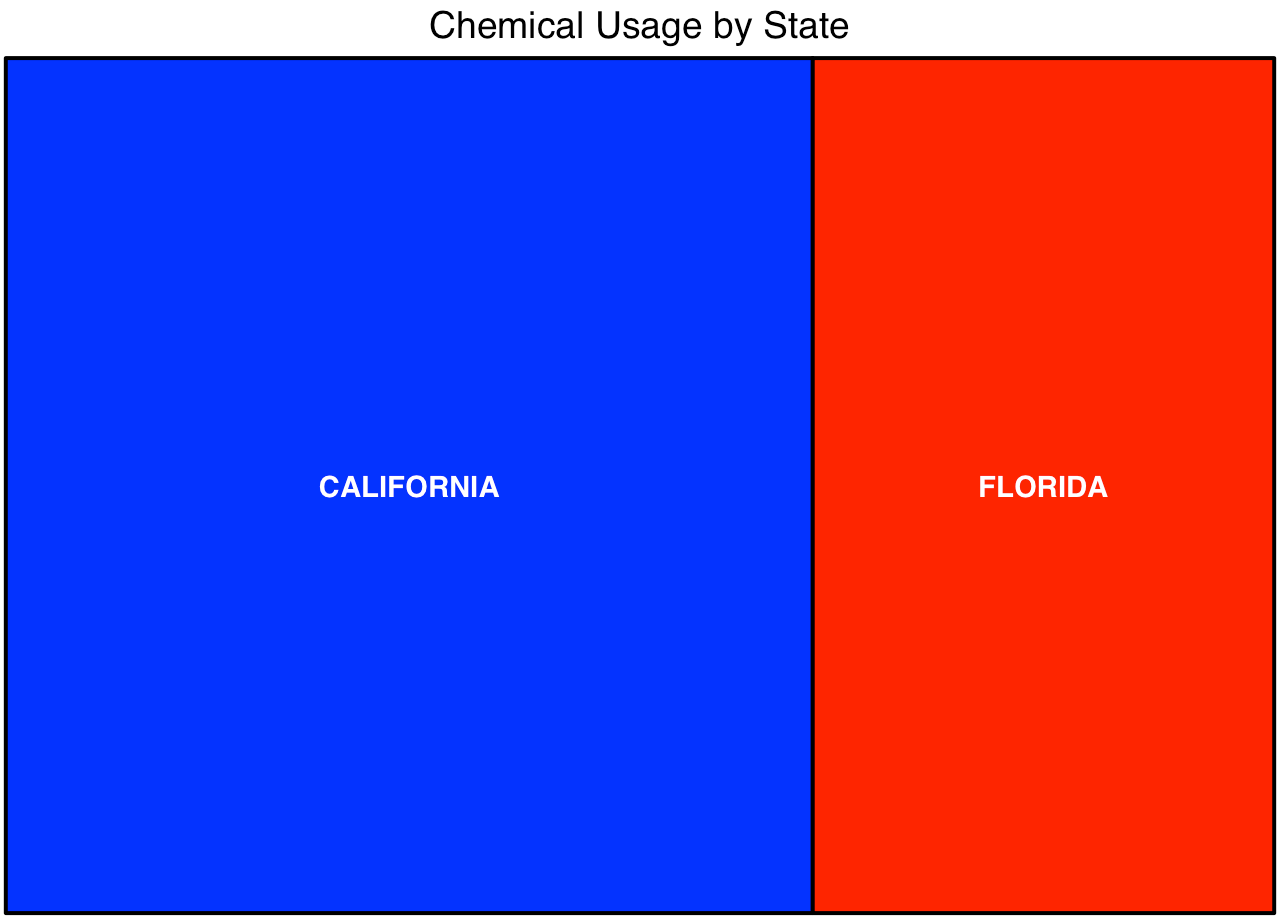


It's very clear that average strawberry price in California is relatively low.

Then, we take a look at the Chemical usage by California and other states.

```
# Install and load required packages
if (!require("treemap")) install.packages("treemap")
library(treemap)
if (!require("dplyr")) install.packages("dplyr")
library(dplyr)

# Group the data by 'State' and count the frequency of each chemical in each stat
state_chemical_count <- strwb_survey_chem %>%
  group_by(State) %>%
  summarize(Chemical_Name_count = n()) %>%
  arrange(desc(Chemical_Name_count))

# Define custom hex colors in a separate column: California as blue (#0000FF), Fl
state_chemical_count <- state_chemical_count %>%
  mutate(Color = case_when(
    State == "CALIFORNIA" ~ "#0000FF",  # Blue for California
    State == "FLORIDA" ~ "#FF0000",     # Red for Florida
    TRUE ~ "#B0B0B0"
  ))
```

```
        # Draw the treemap, using vColor as 'Color' and setting the color palette explici
        treemap(state_chemical_count,
                index = "State",
                vSize = "Chemical_Name_count",
                vColor = "Color",
                type = "color",  # Specify color type to interpret hex colors
                title = "Chemical Usage by State")
```



Needless to say, California has a large strawberry crop, so it obviously uses a lot of chemicals to kill insects.

Therefore, let's take a look at which chemicals are most commonly used in California.

```
        write_csv(strawberry, "final_cleaned_strawberry_data.csv")
```