Exploratory Data Analysis on Strawberry Chemical Data

Ruihang Han

2024-10-20

1. Introduction

This report presents an exploratory data analysis (EDA) of chemical usage in strawberry farming across different U.S. states. The aim of this analysis is to explore how chemical use varies across regions and domains, identify trends and outliers, and provide insights into the reasons behind these differences. Understanding the factors influencing chemical use can help inform better farming practices and policies for a more sustainable future.

2. Data Overview

The dataset includes various variables related to chemical usage, including:

- **State:** The state where the chemical was applied.
- **Value:** The amount of chemical used.
- **Domain:** The category or type of chemical (e.g., Fertilizer, Insecticide, etc.).

Before starting the analysis, we clean the data by removing irrelevant columns and rows with missing values.

3. Data Cleaning

To ensure we work with a clean dataset, we:

Remove irrelevant columns (e.g., Fruit, Category, Item, and Metric). Filter out rows with missing Value, as these entries don't provide any information on chemical usage.

```
# Data cleaning
data_cleaned <- data %>%
  select(-c(Fruit, Category, Item, Metric)) %>%  # Drop irrelevant columns
  filter(!is.na(Value))  # Remove rows with missing 'Value'

# Checking the cleaned data
summary(data_cleaned)
```

```
##    Program               Year         Period           Geo.Level
##  Length:7220        Min.   :2018   Length:7220        Length:7220
##  Class :character   1st Qu.:2022   Class :character   Class :character
##  Mode  :character   Median :2022   Mode  :character   Mode  :character
##                     Mean   :2022
```

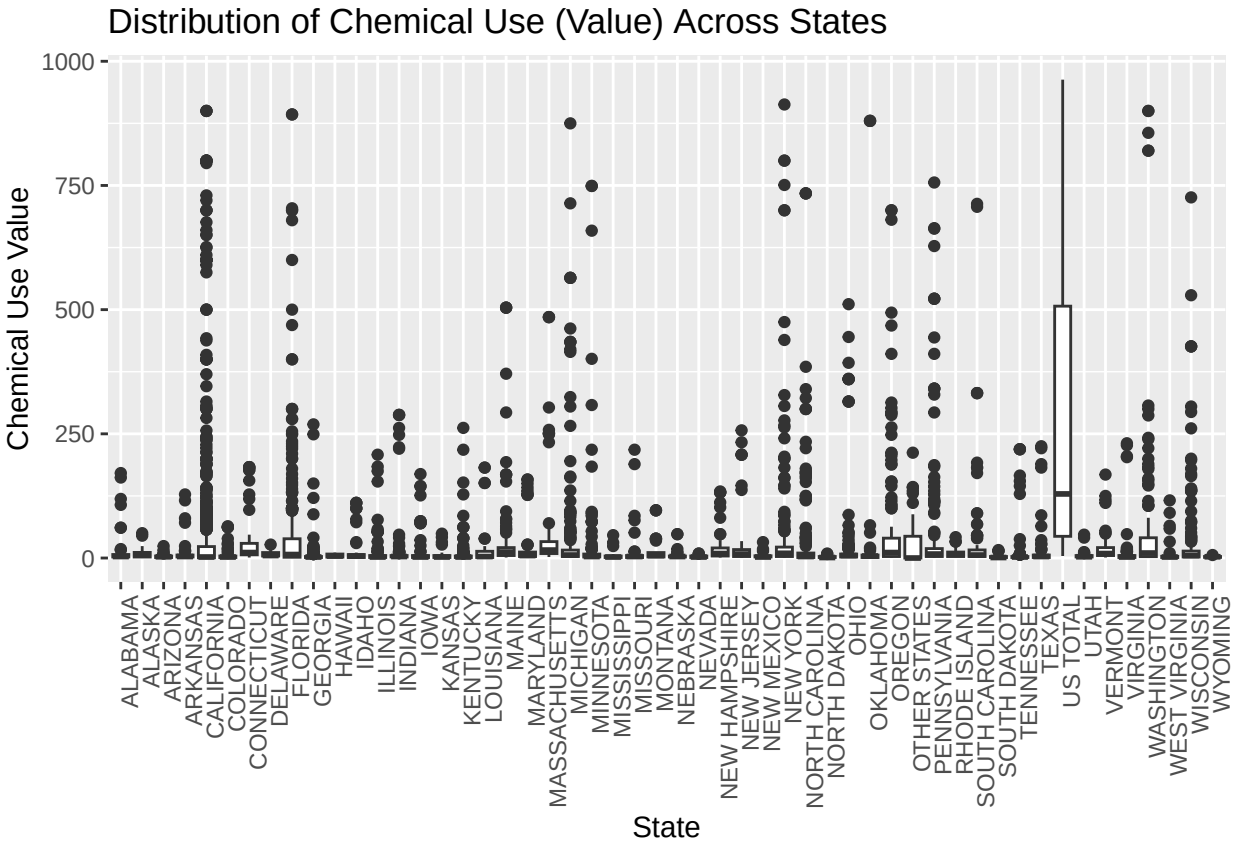
```
##                    3rd Qu.:2022
##                    Max.   :2024
##
##     State              State.ANSI    Ag.District        Ag.District.Code
##  Length:7220        Min.   : 1.00   Length:7220        Min.   :10.00
##  Class :character   1st Qu.:12.00   Class :character   1st Qu.:20.00
##  Mode  :character   Median :26.00   Mode  :character   Median :50.00
##                     Mean   :26.83                      Mean   :45.81
##                     3rd Qu.:41.00                      3rd Qu.:60.00
##                     Max.   :56.00                      Max.   :96.00
##                     NA's   :114                        NA's   :2399
##     County            County.ANSI        Domain             use
##  Length:7220        Min.   :  1.00   Length:7220        Length:7220
##  Class :character   1st Qu.: 27.00   Class :character   Class :character
##  Mode  :character   Median : 67.00   Mode  :character   Mode  :character
##                     Mean   : 82.46
##                     3rd Qu.:117.00
##                     Max.   :810.00
##                     NA's   :2414
##  details            Value             CV....         name            code
##  Mode:logical   Min.   :  0.00   Min.   : 0.60   Mode:logical   Mode:logical
##  NA's:7220      1st Qu.:  1.50   1st Qu.:29.60   NA's:7220      NA's:7220
##                 Median :  4.00   Median :41.60
##                 Mean   : 29.91   Mean   :43.32
##                 3rd Qu.: 12.00   3rd Qu.:55.90
##                 Max.   :963.00   Max.   :99.90
##                                  NA's   :2744
```

4. Analysis and Visualizations

4.1 Distribution of Chemical Use Across States

The following boxplot illustrates the distribution of chemical use across different states. The width of the distribution, along with the presence of outliers, suggests that certain states have farms or regions that use a significantly higher amount of chemicals.

```r
# Distribution of chemical use across states
ggplot(data_cleaned, aes(x = State, y = Value)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Distribution of Chemical Use (Value) Across States",
       x = "State", y = "Chemical Use Value")
```



Insights:

States like California, Florida, and Texas display wide variability in chemical use. The range of values in these states suggests that different regions within the same state might be employing distinct agricultural practices. For instance, larger commercial farms may rely more on chemical inputs to manage pests, boost soil fertility, or improve crop yield consistency.

The presence of multiple outliers, especially in California, may indicate that certain farms are disproportionately dependent on chemical use. These could be regions with higher crop density or more frequent pest issues. California's wide spread and high outliers could also be attributed to the diversity of crops grown, with some crops (e.g., strawberries) requiring more intensive chemical treatments due to their vulnerability to pests and diseases.

Florida and Texas similarly show wide distributions, potentially due to climatic conditions that promote the growth of pests and weeds, necessitating greater use of insecticides and herbicides. For example, the warm and humid environment in Florida creates ideal conditions for pest proliferation, which would explain higher reliance on pesticides.

On the other hand, some states like North Dakota and Wyoming have narrower ranges and lower median chemical use. These states may focus on crops that are less dependent on chemical inputs or may have more stringent regulations limiting chemical use.

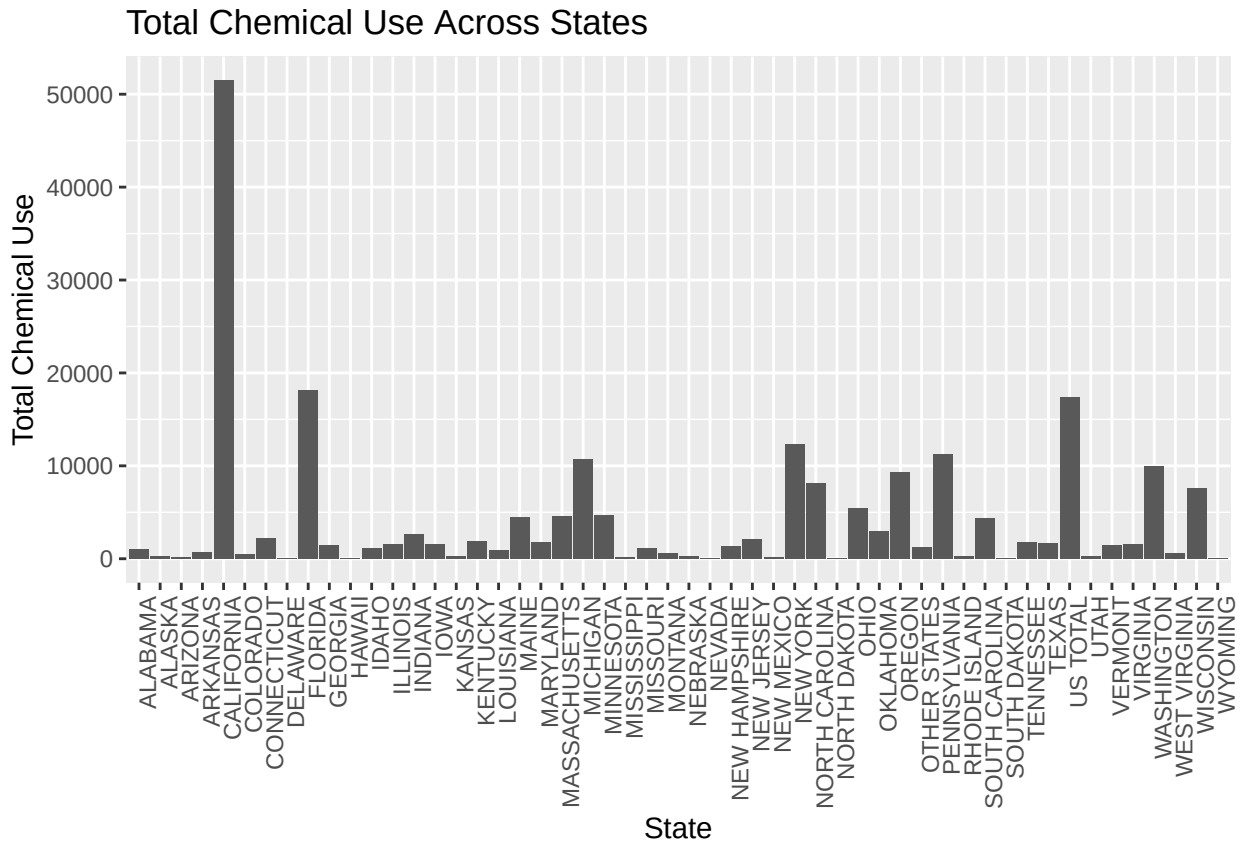
There may also be differences in the type of crops grown in various states, with states producing high-maintenance crops (like strawberries) likely using more chemicals to sustain crop health and yield.

4.2 Total Chemical Use Across States

The bar plot below provides a clearer picture of total chemical use per state. This allows us to see which states contribute the most to overall chemical usage.

```r
# Aggregate data by state for total chemical use per state
state_chemical_use <- data_cleaned %>%
  group_by(State) %>%
  summarise(Total_Chemical_Use = sum(Value, na.rm = TRUE))

# Bar plot of total chemical use by state
ggplot(state_chemical_use, aes(x = State, y = Total_Chemical_Use)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Total Chemical Use Across States",
       x = "State", y = "Total Chemical Use")
```



Insights:

California is the leading state in terms of total chemical use, which is not surprising given its status as the largest agricultural state in the U.S. The combination of large-scale farming operations, diverse crop production, and a climate conducive to year-round farming means that California requires substantial chemical inputs. High-value crops like strawberries, which are grown in abundance in California, are particularly reliant on pesticides and fertilizers, contributing to the state's high usage.

Florida and Texas also show significant total chemical usage. These states are known for producing large quantities of fruits and vegetables, which may require chemical treatments to protect against pests and diseases, especially in their warm climates. Additionally, Florida's high humidity and precipitation levels can exacerbate fungal and pest problems, leading to more fungicide and insecticide applications.

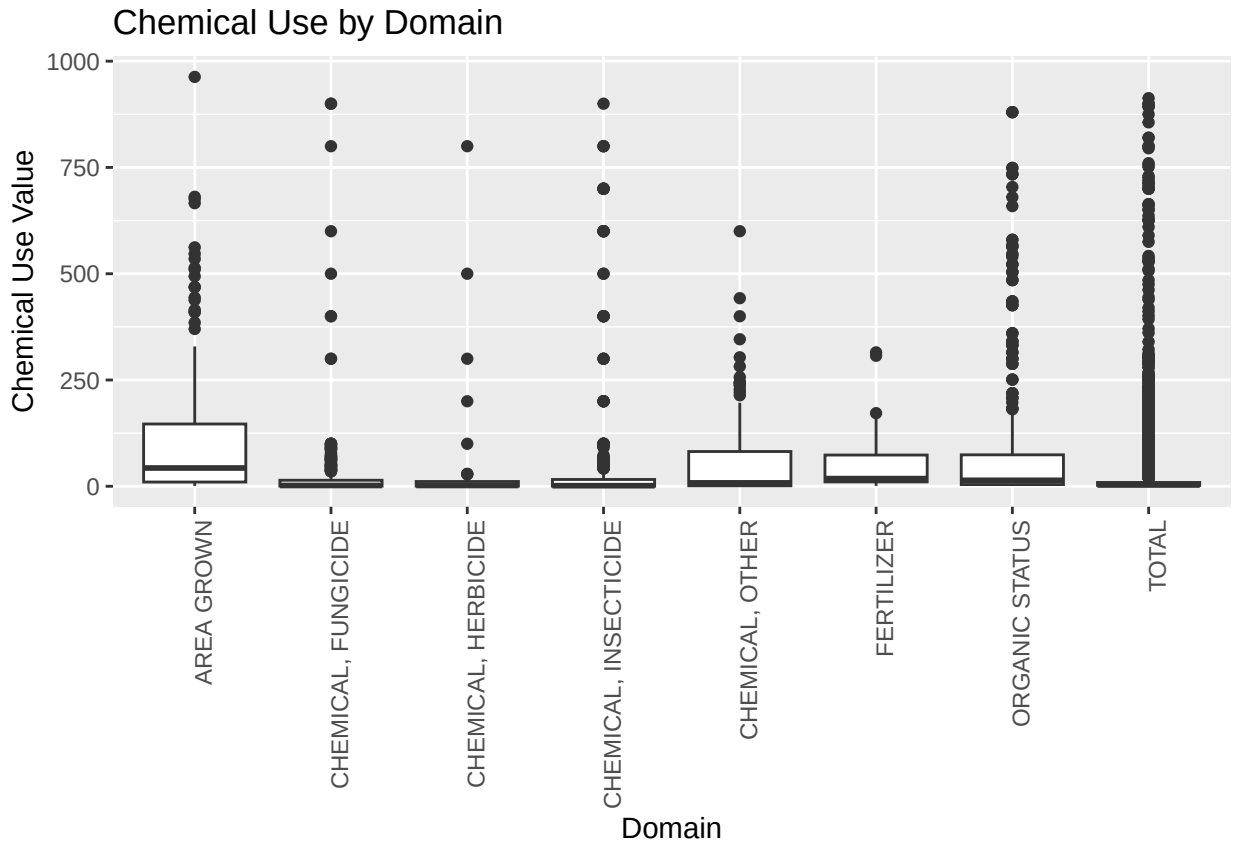
Other states, while contributing less in absolute terms, may have more targeted or specialized agricultural sectors that do not require as much chemical input. States like Maine or Vermont are known for smaller, often organic or semi-organic farming operations, which likely use fewer chemicals.

The large discrepancy between states like California and smaller agricultural states highlights the scale of industrial farming and raises questions about the sustainability and environmental impact of such high levels of chemical usage.

4.3 Chemical Use by Domain

The plot below breaks down chemical use by domain, providing insights into which types of chemicals are most frequently applied across agricultural operations.

```
# Distribution of chemical use by domain
ggplot(data_cleaned, aes(x = Domain, y = Value)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Chemical Use by Domain",
       x = "Domain", y = "Chemical Use Value")
```



Insights:

Fertilizers dominate the chemical usage in most states. This is expected, as fertilizers are crucial for maintaining soil fertility, especially in intensive farming operations where the same land is used for consecutive crop cycles. Fertilizers help replenish nutrients in the soil, ensuring that crops like strawberries continue to produce high yields.

Insecticides and fungicides show lower median usage compared to fertilizers. This could indicate that farmers are employing more selective pest management practices, such as integrated pest management (IPM), which aims to minimize pesticide use through biological control and crop rotation. The lower reliance on pesticides

could also suggest that some farms are adopting organic or sustainable practices, reducing chemical inputs where possible.

The "Other" chemical category includes chemicals that don't fall under the standard pesticide or fertilizer classifications. This category could encompass a range of agricultural inputs, from growth regulators to soil conditioners, which are used to improve crop quality or resilience.

Organic status shows variability in chemical use, which might seem contradictory at first. However, it's possible that certain chemicals allowed under organic certifications (such as specific natural pesticides) are being included here. This highlights the complexity of organic farming, where some chemical inputs are still permissible under certain regulations.

5. Conclusions

This analysis reveals several important insights:

California's Dominance: California's leading role in agriculture is reflected in its high total and per-farm chemical usage. This suggests that California's agricultural practices, especially for high-value crops like strawberries, are heavily dependent on chemicals. Outliers in Chemical Use: The presence of extreme outliers in states like California points to potential inefficiencies or overuse of chemicals in certain farms, raising concerns about sustainability and environmental impact.

Variability Across States: The wide range of chemical use across states could be attributed to differences in climate, crop types, and farm sizes. Warmer, more humid states like Florida may need more chemical treatments to manage pests and diseases, while states with more temperate climates might require fewer chemical interventions.

6. Recommendations

Based on the findings, the following actions are recommended:

Review Chemical Use Practices in High-Usage States: States like California and Florida should review their agricultural practices to ensure that chemical usage is optimized and sustainable, particularly for outlier farms that show signs of overuse.

Encourage Organic Practices: There is an opportunity to expand organic farming practices, especially in states with lower pesticide and fungicide usage. Promoting integrated pest management (IPM) could help reduce reliance on chemicals without sacrificing yield.

State-Specific Policies: Given the variation in chemical use across states, policymakers should tailor agricultural policies to state-specific conditions, ensuring that they reflect local needs while promoting sustainable farming.