Exploratory Data Analysis on Strawberry Chemical Data

Ruihang Han

2024-10-21

Introduction

This analysis explores various aspects of chemical applications, including their types, usage over time, the most heavily used chemicals, and toxicity trends across states. Understanding these trends helps us evaluate agricultural practices and assess the potential environmental and public health impacts.

1. Loading and Cleaning the Dataset

```
# Load necessary libraries
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(PubChemR)
library(stringr)

# Read the dataset (replace with your file path)
data_cleaned <- read.csv("survey_d_chem.csv")

# Convert 'Value' to numeric, replacing (D) and (NA) with NA
data_cleaned$Value <- as.numeric(gsub("[^0-9.]", "", data_cleaned$Value))
```

Explanation:

Purpose: This section loads the necessary libraries and reads the dataset containing information about the chemicals used in agriculture.

Cleaning: It cleans the Value column by replacing non-numeric entries (e.g., "(D)", "(NA)") with NA and converting it to a numeric format.

EDA Objective: The Value column represents the quantity of chemicals used, and cleaning this data is crucial for subsequent analyses to ensure calculations and visualizations are accurate.

2. Retrieving Hazard Data from PubChem (GHS Classification)

Explanation:

Purpose: This section of code uses PubChem's API to retrieve hazard data for each chemical. It looks for hazard classifications according to the Globally Harmonized System (GHS).

GHS_searcher: Searches for the GHS classification in the API response. hazards_retriever: Extracts the relevant hazard statements (e.g., H300: Fatal if swallowed) from the classification. EDA Objective: Hazard data allows us to analyze the potential dangers associated with each chemical. This information will be used later to simulate toxicity scores based on the hazard classifications.

```r
# Define functions for hazard retrieval from PubChemR

GHS_searcher <- function(result_json_object) {
  if (!is.null(result_json_object[["result"]][["Hierarchies"]])) {
    for (i in 1:length(result_json_object[["result"]][["Hierarchies"]][["Hierarchy"]])) {
      if (result_json_object[["result"]][["Hierarchies"]][["Hierarchy"]][[i]][["SourceName"]]
        return(i)
      }
    }
```

```
  }
  return(NA)  # Return NA if GHS Classification not found
}

hazards_retriever <- function(index, result_json_object) {
  if (!is.na(index)) {
    hierarchy <- result_json_object[["result"]][["Hierarchies"]][["Hierarchy"]][[index]]
    output_list <- c()
    for (i in seq_along(hierarchy[["Node"]])) {
      if (str_detect(hierarchy[["Node"]][[i]][["Information"]][["Name"]], "H")) {
        output_list <- c(output_list, hierarchy[["Node"]][[i]][["Information"]][["Name"]])
      }
    }
    return(output_list)
  } else {
    return(NA)  # Return NA if no valid index found
  }
}

# Example: Retrieve hazard data for chemicals
chemical_vec <- unique(data_cleaned$chem_name)

hazard_data <- list()
for (chemical in chemical_vec[1:5]) {  # Limiting to 5 chemicals for demo
  result_f <- tryCatch(
    get_pug_rest(identifier = chemical, namespace = "name", domain = "compound", operation =
    error = function(e) return(NULL)  # Return NULL if there's an error
  )

  if (!is.null(result_f)) {
    ghs_index <- GHS_searcher(result_f)
    hazard_list <- hazards_retriever(ghs_index, result_f)
    hazard_data[[chemical]] <- hazard_list
  } else {
    hazard_data[[chemical]] <- NA
  }
}
```

```
Request failed [404]. Retrying in 2.3 seconds...

Request failed [404]. Retrying in 7.1 seconds...
```

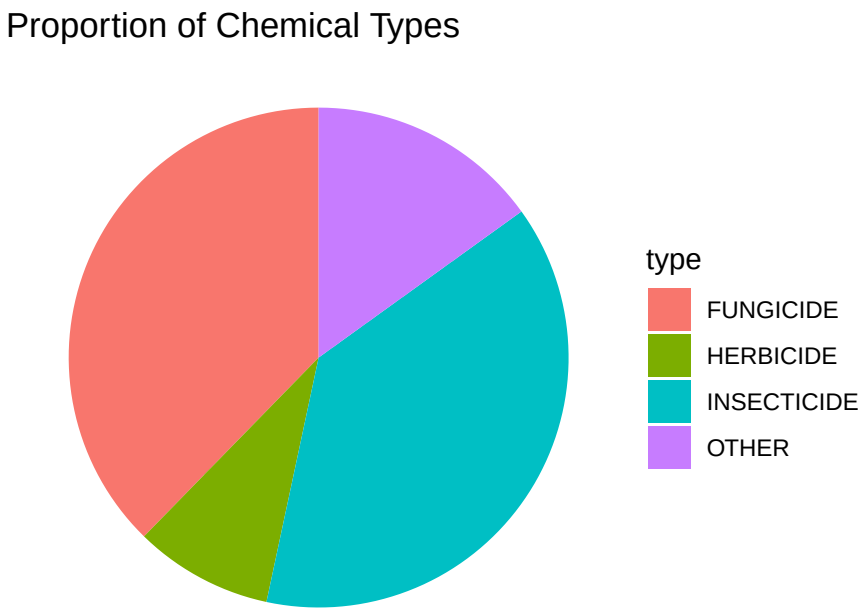
```r
# Print hazards for the chemicals
print(hazard_data)
```

```
$OXATHIAPIPROLIN
 [1] "H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute h
 [2] "H400: Environmental Hazards"
 [3] "Hazard Statement Codes"
 [4] "H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the ac
 [5] "Hazardous to the aquatic environment, acute hazard"
 [6] "Environmental Hazards"
 [7] "Hazard Classes"
 [8] "Hazardous to the aquatic environment, long-term hazard"
 [9] "<img src=\"https://pubchem.ncbi.nlm.nih.gov/images/ghs/GHS09.svg\" style=\"width:40px;h
[10] "Hazard Pictograms"
[11] "ECHA C&P Inventory"
[12] "HCIS, Safe Work Australia (SWA)"

$CYCLANILIPROLE
 [1] "H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute h
 [2] "H400: Environmental Hazards"
 [3] "Hazard Statement Codes"
 [4] "H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the ac
 [5] "Hazardous to the aquatic environment, acute hazard"
 [6] "Environmental Hazards"
 [7] "Hazard Classes"
 [8] "Hazardous to the aquatic environment, long-term hazard"
 [9] "<img src=\"https://pubchem.ncbi.nlm.nih.gov/images/ghs/GHS09.svg\" style=\"width:40px;h
[10] "Hazard Pictograms"
[11] "ECHA C&P Inventory"

$PERMETHRIN
 [1] "H302: Harmful if swallowed [Warning Acute toxicity, oral]"
 [2] "H300: Health Hazards"
 [3] "Hazard Statement Codes"
 [4] "H302+H332: Harmful if swallowed or if inhaled [Warning Acute toxicity, oral; acute toxi
 [5] "H317: May cause an allergic skin reaction [Warning Sensitization, Skin]"
 [6] "H332: Harmful if inhaled [Warning Acute toxicity, inhalation]"
 [7] "H351: Suspected of causing cancer [Warning Carcinogenicity]"
 [8] "H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure]"
 [9] "H371: May cause damage to organs [Warning Specific target organ toxicity, single exposu
[10] "H373: May causes damage to organs through prolonged or repeated exposure [Warning Speci
[11] "H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute h
```

```
[12] "H400: Environmental Hazards"
[13] "H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the ac
[14] "Health Hazards"
[15] "Hazard Classes"
[16] "Hazardous to the aquatic environment, acute hazard"
[17] "Environmental Hazards"
[18] "Hazardous to the aquatic environment, long-term hazard"
[19] "<img src=\"https://pubchem.ncbi.nlm.nih.gov/images/ghs/GHS07.svg\" style=\"width:40px;h
[20] "Hazard Pictograms"
[21] "<img src=\"https://pubchem.ncbi.nlm.nih.gov/images/ghs/GHS08.svg\" style=\"width:40px;h
[22] "<img src=\"https://pubchem.ncbi.nlm.nih.gov/images/ghs/GHS09.svg\" style=\"width:40px;h
[23] "P304+P340: IF INHALED: Remove person to fresh air and keep comfortable for breathing."
[24] "ECHA C&P Inventory"

$`ISARIA FUMOSOROSEA STRAIN FE 9901`
[1] NA

$AZOXYSTROBIN
 [1] "H331: Toxic if inhaled [Danger Acute toxicity, inhalation]"
 [2] "H300: Health Hazards"
 [3] "Hazard Statement Codes"
 [4] "H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure]"
 [5] "H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute h
 [6] "H400: Environmental Hazards"
 [7] "H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the ac
 [8] "Health Hazards"
 [9] "Hazard Classes"
[10] "Hazardous to the aquatic environment, acute hazard"
[11] "Environmental Hazards"
[12] "Hazardous to the aquatic environment, long-term hazard"
[13] "<img src=\"https://pubchem.ncbi.nlm.nih.gov/images/ghs/GHS06.svg\" style=\"width:40px;h
[14] "Hazard Pictograms"
[15] "<img src=\"https://pubchem.ncbi.nlm.nih.gov/images/ghs/GHS08.svg\" style=\"width:40px;h
[16] "<img src=\"https://pubchem.ncbi.nlm.nih.gov/images/ghs/GHS09.svg\" style=\"width:40px;h
[17] "P304+P340: IF INHALED: Remove person to fresh air and keep comfortable for breathing."
[18] "ECHA C&P Inventory"
[19] "HCIS, Safe Work Australia (SWA)"
```

3. Proportion of Chemical Types

```
# Continue with EDA and visualization...
# 1. Proportion of Chemical Types
chemical_type_counts <- data_cleaned %>%
  count(type)

ggplot(chemical_type_counts, aes(x = "", y = n, fill = type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Proportion of Chemical Types") +
  theme_void()
```



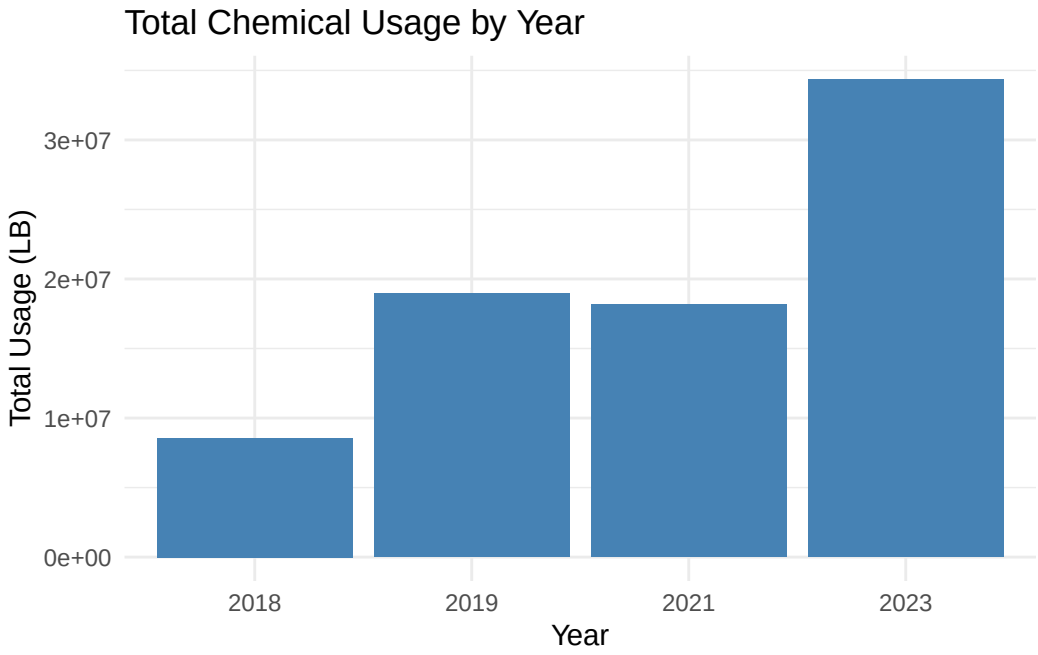
Analysis

- **Purpose**: To determine the relative distribution of different types of chemicals being used in agriculture.
- **Findings**: From the pie chart, we see that **insecticides** dominate chemical applications, followed by **fungicides**. This suggests a high prevalence of pest control measures in agricultural practices. The smaller proportions of herbicides and other chemicals suggest more specific or limited use cases.

- **Implications**: The heavy use of insecticides might indicate a focus on protecting crops from insect-related damage, but it also raises concerns about the long-term ecological impact on pollinators and other non-target organisms.

4. Total Chemical Usage by Year

```
# 2. Total Chemical Usage by Year
usage_by_year <- data_cleaned %>%
  group_by(Year) %>%
  summarise(total_usage = sum(Value, na.rm = TRUE))

ggplot(usage_by_year, aes(x = factor(Year), y = total_usage)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Total Chemical Usage by Year", x = "Year", y = "Total Usage (LB)") +
  theme_minimal()
```



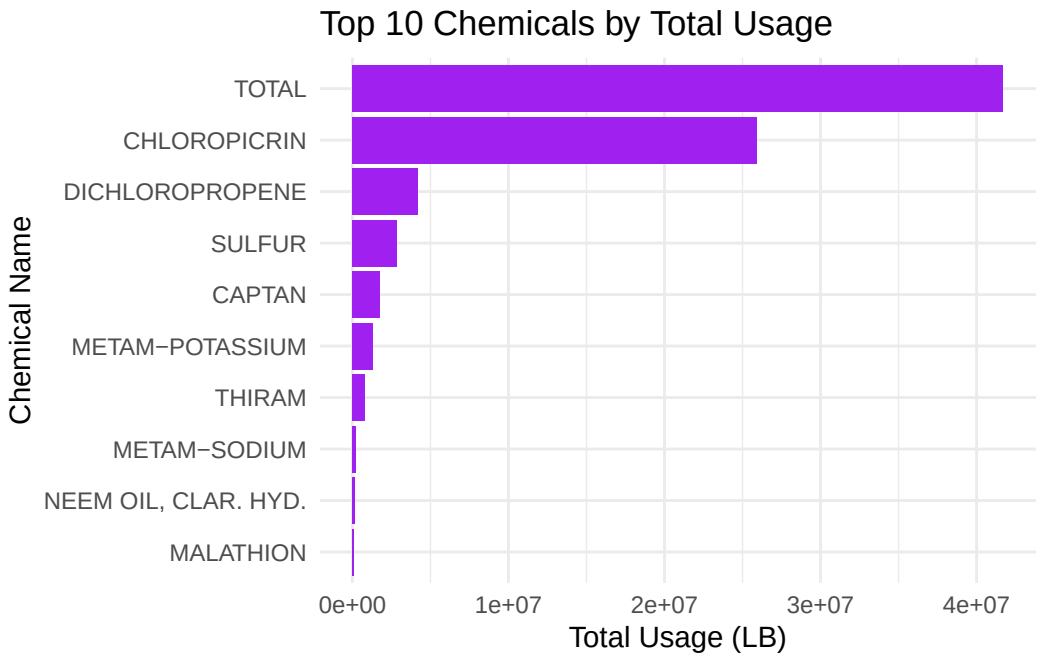
Analysis:

- **Purpose**: To examine the trends in overall chemical usage over several years.

- **Findings**: There is a clear increase in chemical usage from 2018 to 2023, with 2023 showing a significant spike. While 2019 and 2021 had comparable usage levels, the sharp rise in 2023 suggests a dramatic shift, potentially due to increased agricultural activities, changes in regulation, or pest outbreaks.
- **Implications**: The increasing trend in chemical usage raises concerns about sustainability. An important next step would be to investigate whether the increase is driven by larger farm areas, higher intensity of chemical applications, or both. Monitoring the ecological effects of this rise is essential.

5. Top 10 Chemicals by Total Usage

```
# 3. Top 10 Chemicals by Total Usage
top_chemicals <- data_cleaned %>%
  group_by(chem_name) %>%
  summarise(total_usage = sum(Value, na.rm = TRUE)) %>%
  arrange(desc(total_usage)) %>%
  top_n(10, total_usage)

ggplot(top_chemicals, aes(x = reorder(chem_name, total_usage), y = total_usage)) +
  geom_bar(stat = "identity", fill = "purple") +
  coord_flip() +
  labs(title = "Top 10 Chemicals by Total Usage", x = "Chemical Name", y = "Total Usage (LB)"
  theme_minimal()
```



Analysis:

- **Purpose**: To identify the most heavily used chemicals and their relative quantities.
- **Findings**: **Chloropicrin** is by far the most used chemical, followed by **dichloropropene** and **sulfur**. Chloropicrin is known for its use as a soil fumigant, which raises concerns about soil health and long-term contamination. Other notable chemicals, such as **sulfur** and **captan**, are widely used for fungal control.
- **Implications**: The heavy reliance on these top chemicals suggests their central role in modern agriculture. However, such large-scale use demands further investigation into their potential health effects on farmworkers and nearby communities. For example, chloropicrin is a strong irritant and has been associated with respiratory issues.

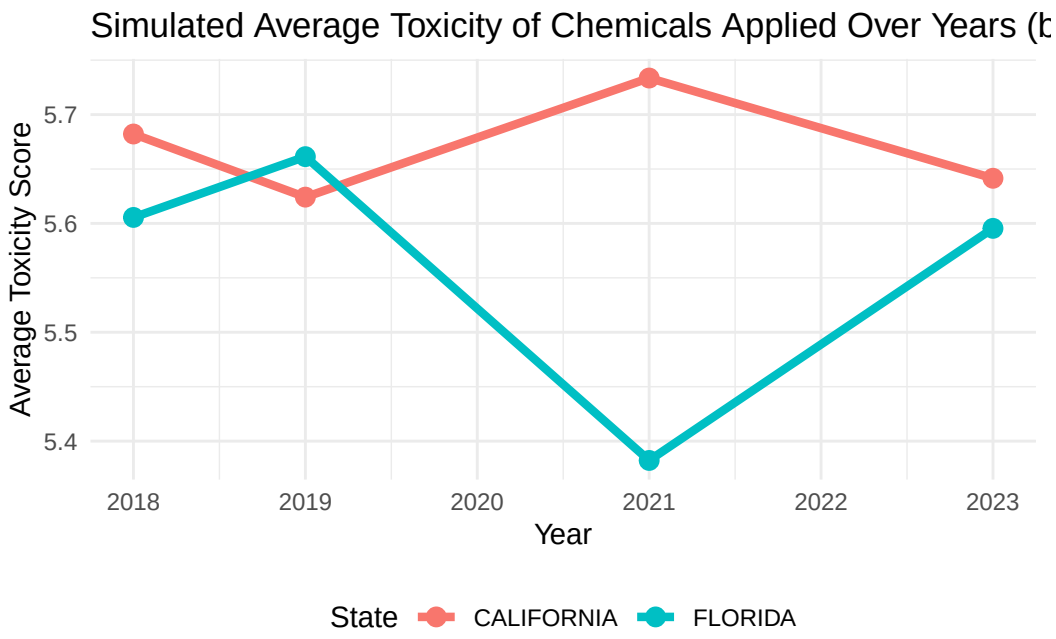
6. Simulating Toxicity Scores Based on GHS Hazard Classes

```
# Load necessary libraries
library(ggplot2)
library(dplyr)

# Simulated dataset
data_cleaned <- read.csv("survey_d_chem.csv")
```

```r
# Convert 'Value' to numeric if needed
data_cleaned$Value <- as.numeric(gsub("[^0-9.]", "", data_cleaned$Value))

# Step 1: Simulate toxicity scores based on GHS hazard classes with additional hazard levels
set.seed(123)  # For reproducibility

toxicity_scores <- list(
  "H300" = 6,  # Fatal if swallowed
  "H301" = 5,  # Toxic if swallowed
  "H302" = 4,  # Harmful if swallowed
  "H304" = 3,  # May be fatal if swallowed
  "H315" = 2,  # Causes skin irritation
  "H319" = 1,  # Causes serious eye irritation
  "H331" = 7,  # Toxic if inhaled
  "H350" = 8,  # May cause cancer
  "H360" = 9,  # May damage fertility or the unborn child (Reproductive toxicity)
  "H361d" = 7, # Suspected of damaging the unborn child (Reproductive toxicity)
  "H400" = 7,  # Very toxic to aquatic life (Acute environmental hazard)
  "H410" = 8   # Very toxic to aquatic life with long-lasting effects (Chronic environmental
)

# Simulate hazard data for chemicals in the dataset
chemicals <- unique(data_cleaned$chem_name)
hazard_data <- list()

# Randomly assign hazard classes to each chemical for the simulation, including the new hazar
for (chemical in chemicals) {
  hazard_data[[chemical]] <- sample(names(toxicity_scores), 1)  # Randomly assign a hazard co
}

# Step 2: Assign a toxicity score to each chemical in the dataset based on the simulated haza
data_cleaned$toxicity_score <- sapply(data_cleaned$chem_name, function(chem_name) {
  hazard <- hazard_data[[chem_name]]
  toxicity_scores[[hazard]]
})

# Step 3: Group data by State and Year, then calculate average toxicity score per state per y
toxicity_by_state_year <- data_cleaned %>%
  group_by(State, Year) %>%
  summarise(average_toxicity = mean(toxicity_score, na.rm = TRUE), .groups = 'drop')

# Step 4: Visualize the toxicity trends across years, split by state
```

```r
ggplot(toxicity_by_state_year, aes(x = Year, y = average_toxicity, color = State, group = Sta
  geom_line(size = 1.5, na.rm = TRUE) +
  geom_point(size = 3, na.rm = TRUE) +
  labs(title = "Simulated Average Toxicity of Chemicals Applied Over Years (by State)",
       x = "Year", y = "Average Toxicity Score") +
  theme_minimal() +
  theme(legend.position = "bottom") +
  scale_color_discrete(name = "State")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Analysis:

- **Purpose**: To evaluate trends in the average toxicity of chemicals used across different states over time.
- **Findings**: California and Florida show distinct toxicity trends. While **California** exhibits a steady increase in average toxicity, **Florida** demonstrates more fluctuation, peaking around 2019 and decreasing by 2021 before rising again in 2023. California's consistent rise may reflect a gradual shift toward more toxic chemicals or an increased focus on intensive pest management.

- **Implications**: The upward trend in toxicity, especially in California, calls for a deeper dive into the types of chemicals being used and the potential environmental and health risks. States that consistently use high-toxicity chemicals should be monitored closely for their environmental impact, including water contamination and harm to biodiversity.

7. Conclusion

Each visualization provides valuable insights into chemical use and toxicity trends in agriculture:

1. **Proportion of Chemical Types** shows that insecticides dominate chemical applications, indicating a strong focus on pest control.
2. **Total Chemical Usage by Year** highlights the increasing reliance on chemicals, particularly in 2023.
3. **Top 10 Chemicals by Total Usage** reveals that chloropicrin is the most heavily used chemical, raising concerns about its long-term environmental effects.
4. **Simulated Toxicity Trends** indicate that California is seeing a steady increase in average chemical toxicity, warranting further investigation into potential risks.

This analysis offers a comprehensive view of chemical usage patterns and potential hazards, laying the groundwork for more in-depth environmental and regulatory studies.