Strawberry Data Cleaning-Chemicals EDA

Ruijian Maggie Lin

2024-10-19

Strawberry Data Cleaning - Chemicals EDA

Read Strawberry Data I cleaned

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

```
## Rows: 12669 Columns: 20
## -- Column specification -------------------------------------------------------
## Delimiter: ","
## chr (17): Program, Period, Geo Level, State, State ANSI, Ag District, County...
## dbl  (3): Year, Ag District Code, Code
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

1. Create new data table when `Domain` == Chemical & `Metric` == LB for California state

```
calif <- strawberry %>%  filter(State=="CALIFORNIA")
calif_chem_lb_o <- calif %>% filter(str_detect(Domain, "CHEMICAL"), Metric == "LB")

# Remove rows where "TOTAL" is detected in Domain Category
calif_chem_lb <- calif_chem_lb_o %>%
  filter(!str_detect(`Domain Category`, "TOTAL"))
```

2. Create new data tables for each different kind of Chemical in `Domain`

```
unique(calif_chem_lb$Domain)
```

```
## [1] "CHEMICAL, FUNGICIDE"   "CHEMICAL, INSECTICIDE" "CHEMICAL, OTHER"
## [4] "CHEMICAL, HERBICIDE"
```

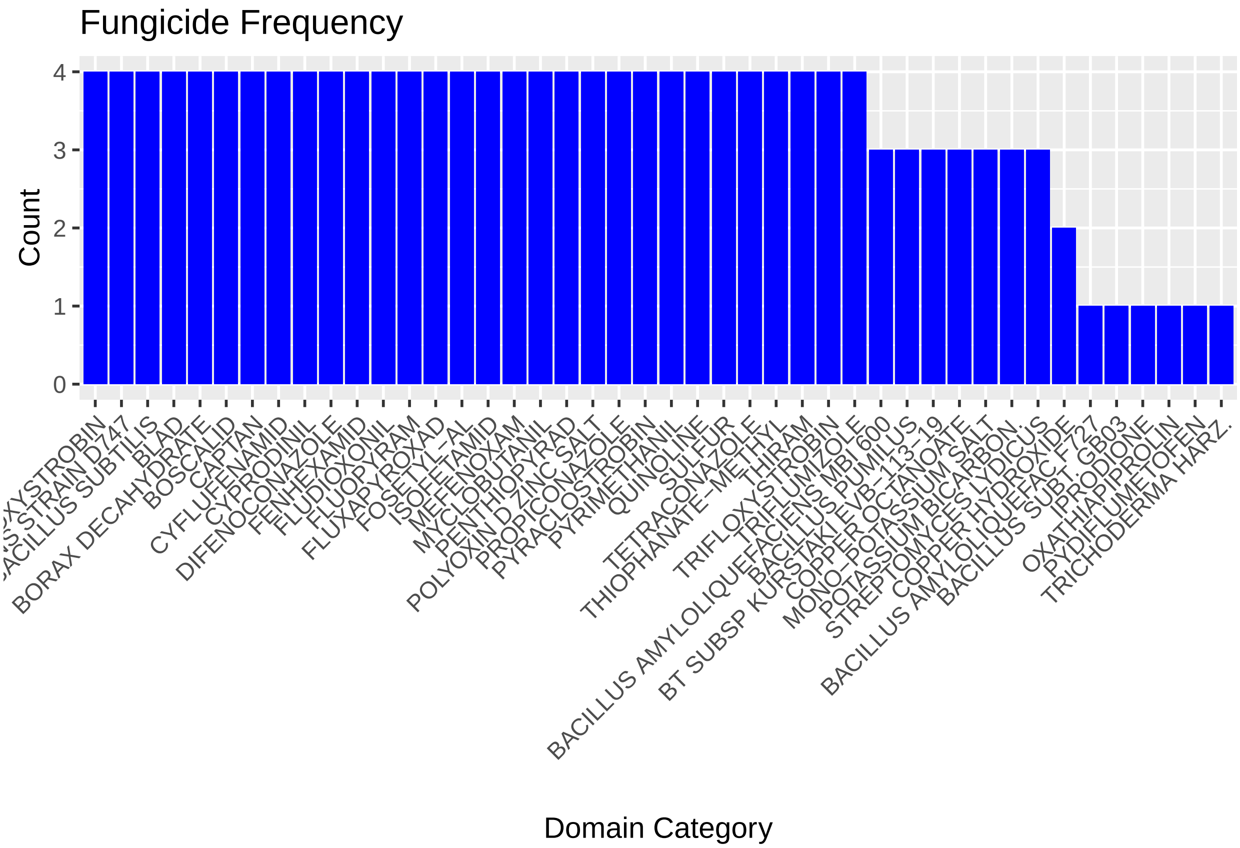
```
fungicide <- calif_chem_lb %>% filter(str_detect(Domain, "FUNGICIDE"))
insecticide <- calif_chem_lb %>% filter(str_detect(Domain, "INSECTICIDE"))
other <- calif_chem_lb %>% filter(str_detect(Domain, "OTHER"))
herbicide <- calif_chem_lb %>% filter(str_detect(Domain, "HERBICIDE"))
```

```r
# Count frequencies for each chemical type
fungicide_counts <- fungicide %>%
  group_by(`Domain Category`) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

insecticide_counts <- insecticide %>%
  group_by(`Domain Category`) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

other_counts <- other %>%
  group_by(`Domain Category`) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

herbicide_counts <- herbicide %>%
  group_by(`Domain Category`) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))
```

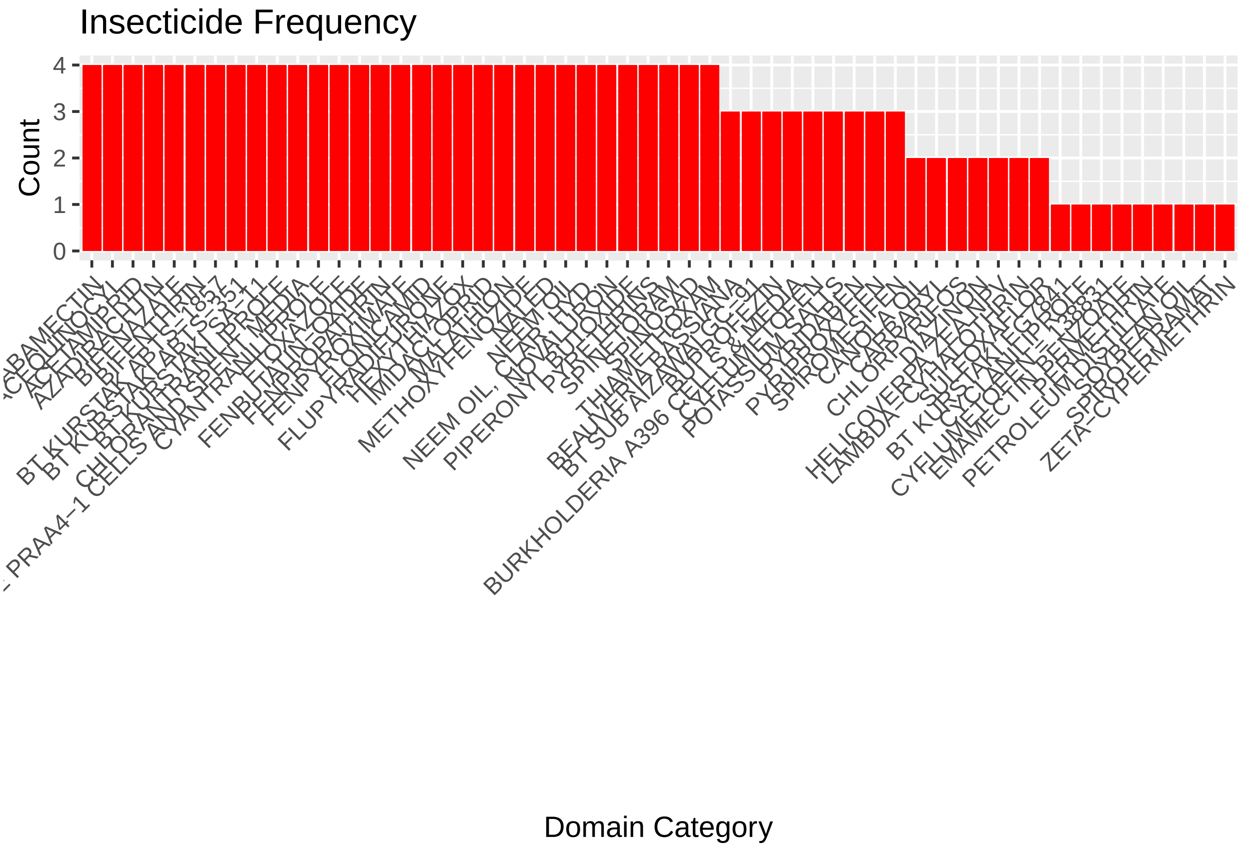
Count the frequency of each chemical name in `Domain Category` for each specific chemical type (from Step 2)

```r
# Fungicides
ggplot(fungicide_counts, aes(x = reorder(`Domain Category`, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Fungicide Frequency", x = "Domain Category", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

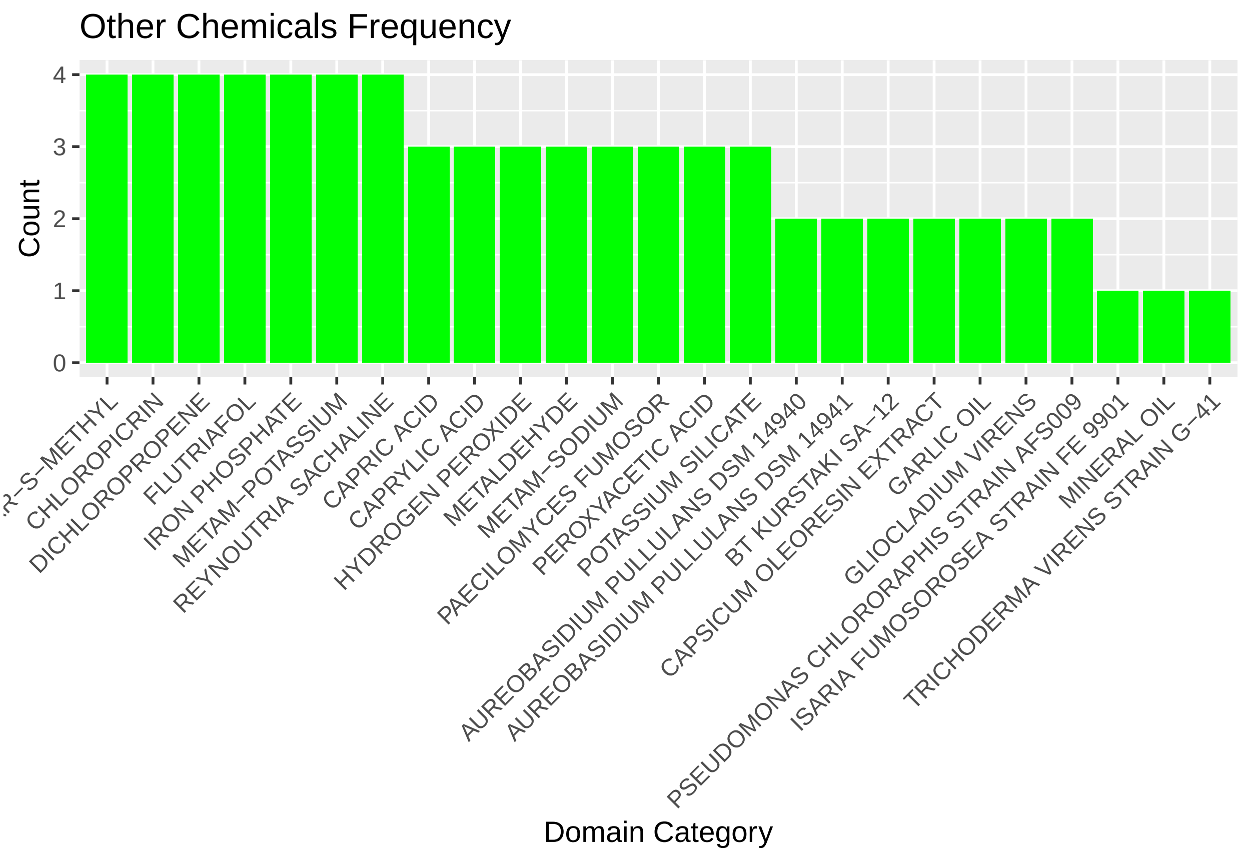
Graph histograms for each specific chemical type to show the frequency of each chemical name visually



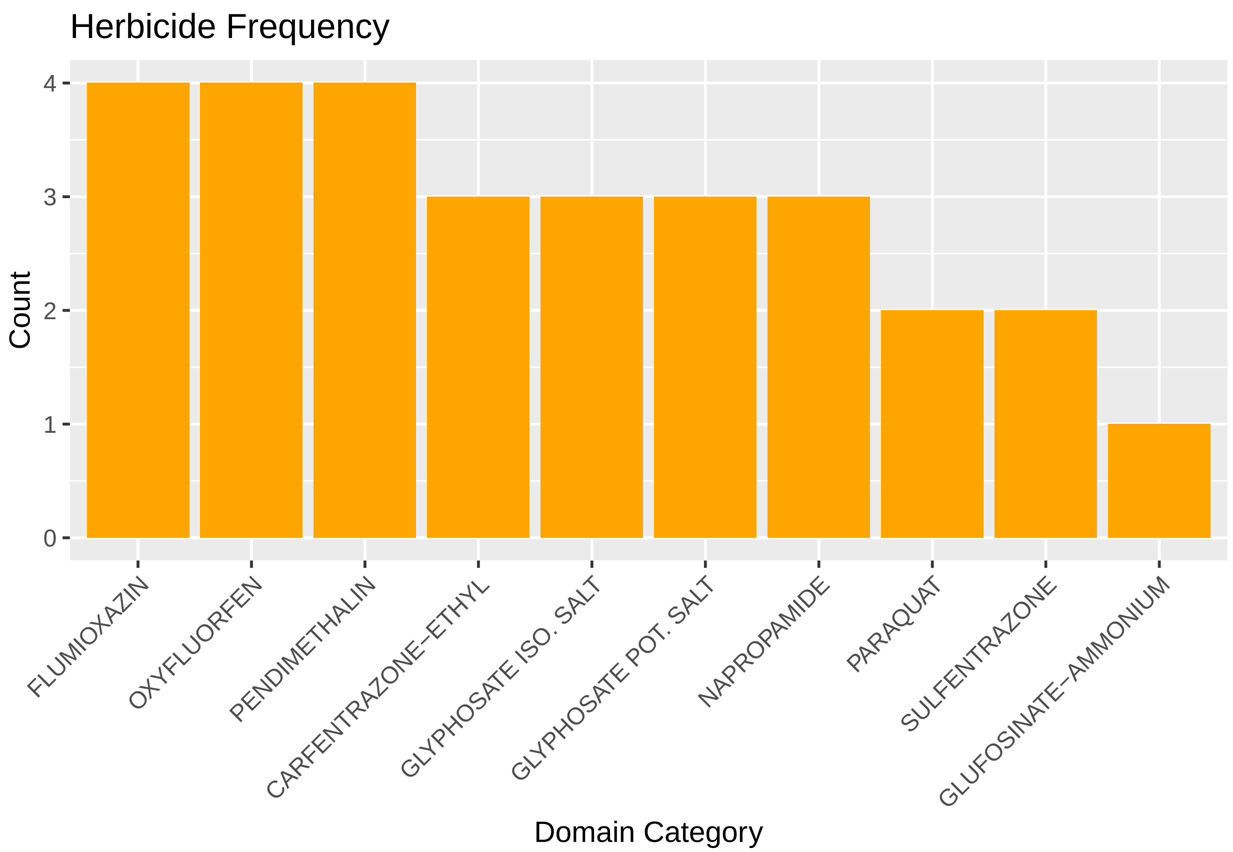
```r
# Insecticides
ggplot(insecticide_counts, aes(x = reorder(`Domain Category`, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(title = "Insecticide Frequency", x = "Domain Category", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
# Others
ggplot(other_counts, aes(x = reorder(`Domain Category`, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "green") +
  labs(title = "Other Chemicals Frequency", x = "Domain Category", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
# Herbicides
ggplot(herbicide_counts, aes(x = reorder(`Domain Category`, -Count), y = Count)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(title = "Herbicide Frequency", x = "Domain Category", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
# Count the number of different chemicals for each chemical type
num_fungicides <- fungicide %>%
  summarise(Unique_Chemicals = n_distinct(`Domain Category`))

num_insecticides <- insecticide %>%
  summarise(Unique_Chemicals = n_distinct(`Domain Category`))

num_others <- other %>%
  summarise(Unique_Chemicals = n_distinct(`Domain Category`))

num_herbicides <- herbicide %>%
  summarise(Unique_Chemicals = n_distinct(`Domain Category`))

list(
  Fungicide = num_fungicides$Unique_Chemicals,
  Insecticide = num_insecticides$Unique_Chemicals,
  Other = num_others$Unique_Chemicals,
  Herbicide = num_herbicides$Unique_Chemicals
)
```

```
## $Fungicide
## [1] 44
##
## $Insecticide
## [1] 56
##
```

```
## $Other
## [1] 25
##
## $Herbicide
## [1] 10
```

Based on the result above, I could observe that in California, there are more different chemicals using for insecticide than other and less chemicals are use for herbicide.

Question: Based on the natural concept, why there are more different chemicals using for insecticide than other, and less chemicals are used for herbicide?

- The analysis I could get based on the result above: The number of chemicals used for insecticides is higher because of the diversity of insect pests, their life cycle complexity, and the development of resistance. In contrast, herbicides are used against a narrower range of targets (weeds) and are constrained by the need to selectively kill weeds without harming crops, resulting in fewer distinct chemicals in use.

3. The chemical names with highest and lowest frequency for Insecticides and Herbicide respectively

```r
# Insecticide: Find the chemical with the highest and lowest frequency
insecticide_freq <- insecticide %>%
  count(`Domain Category`, sort = TRUE)

insecticide_max <- insecticide_freq %>%
  slice_max(n, n = 1)

insecticide_min <- insecticide_freq %>%
  slice_min(n, n = 1)

# Herbicide: Find the chemical with the highest and lowest frequency
herbicide_freq <- herbicide %>%
  count(`Domain Category`, sort = TRUE)

herbicide_max <- herbicide_freq %>%
  slice_max(n, n = 1)

herbicide_min <- herbicide_freq %>%
  slice_min(n, n = 1)
```

Based on the result above, highest frequency = 4; lowest frequency = 1:

Insecticide:

- 10 chemicals with highest frequency: ABAMECTIN, ACEQUINOCYL, ACETAMIPRID, AZADIRACHTIN, BIFENAZATE, BIFENTHRIN, BT KURSTAK ABTS-1857, BT KURSTAKI ABTS-351, BT KURSTAKI SA-11, CHLORANTRANILIPROLE.

- 9 chemicals with lowest frequency: BT KURSTAKI EG7841, CYCLANILIPROLE, CYFLUMETOFEN = 138831, EMAMECTIN BENZOATE, PERMETHRIN, PETROLEUM DISTILLATE, SOYBEAN OIL, SPIROTETRAMAT, ZETA-CYPERMETHRIN.

Herbicide:

- 3 chemicals with highest frequency: FLUMIOXAZIN, OXYFLUORFEN, PENDIMETHALIN.
- 1 chemicals with lowest frequency: GLUFOSINATE-AMMONIUM.

```r
library(PubChemR)

# List all functions in PubChemR
# ls("package:PubChemR")
```

```r
# Define the GHS_searcher function
GHS_searcher <- function(result_json_object) {
  # Check if the necessary parts of the result exist
  if (!is.null(result_json_object[["result"]][["Hierarchies"]])) {
    hierarchies <- result_json_object[["result"]][["Hierarchies"]][["Hierarchy"]]
    if (length(hierarchies) > 0) {
      for (i in 1:length(hierarchies)) {
        if (hierarchies[[i]][["SourceName"]] == "GHS Classification (UNECE)") {
          return(hierarchies[[i]])
        }
      }
    }
  }
  return(NULL)  # Return NULL if GHS Classification is not found
}

# Placeholder for hazards_retriever function
hazards_retriever <- function(ghs_data, full_data) {
  if (is.null(ghs_data)) {
    print("No GHS data found.")
    return(NULL)
  }

  # Assuming you want to extract specific hazard statements or information
  # Here's a simple example of how you might structure this
  hazard_statements <- ghs_data[["GHS Hazard Statements"]]

  if (!is.null(hazard_statements)) {
    print(hazard_statements)
  } else {
    print("No hazard statements available.")
  }
}
```

```r
# Retrieve GHS classification for Acequinocyl
result_f <- get_pug_rest(identifier = "acequinocyl", namespace = "name", domain = "compound",
                         operation = "classification", output = "JSON")
```

```r
# Use the GHS_searcher function to extract GHS data
ghs_data <- GHS_searcher(result_f)

# Retrieve and print the hazards information
hazards_retriever(ghs_data, result_f)
```

(a) For example, choose Acequinocyl from high frequency list and Cyclaniliprole from low frequency list. Assess the chemical's safety and hazards, and analyze why we use those chemicals in high frequency and others in low frequency in Insecticide broadly.

```
## [1] "No hazard statements available."
```

```r
# Retrieve GHS classification for Cyclaniliprole
result_f <- get_pug_rest(identifier = "cyclaniliprole", namespace = "name",
                         domain = "compound", operation = "classification", output = "JSON")

# Use the GHS_searcher function to extract GHS data
ghs_data <- GHS_searcher(result_f)

# Retrieve and print the hazards information
hazards_retriever(ghs_data, result_f)
```

```
## [1] "No hazard statements available."
```

Conclusion: While Acequinocyl and Cyclaniliprole are effective insecticides that play a crucial role in pest management, their high frequency of use must be accompanied by careful consideration of their safety and environmental impacts. Adopting integrated pest management practices that emphasize responsible use and environmental stewardship can help ensure that these chemicals contribute to effective pest control while minimizing adverse effects on human health and ecosystems.

```r


# Retrieve GHS classification for Flumioxazin
result_f <- get_pug_rest(identifier = "flumioxazin", namespace = "name", domain = "compound",
                         operation = "classification", output = "JSON")

# Use the GHS_searcher function to extract GHS data
ghs_data <- GHS_searcher(result_f)

# Retrieve and print the hazards information
hazards_retriever(ghs_data, result_f)
```

(b) For example, choose Flumioxazin from high frequency list and Glufosinate-Ammonium from low frequency list. Assess the chemical's safety and hazards, and analyze why we use those chemicals in high frequency and others in low frequency in Herbicide broadly.

```
## [1] "No hazard statements available."
```

```r
# Retrieve GHS classification for Glufosinate-Ammonium
result_f <- get_pug_rest(identifier = "glufosinate-ammonium ", namespace = "name",
                         domain = "compound", operation = "classification", output = "JSON")
```

```r
# Use the GHS_searcher function to extract GHS data
ghs_data <- GHS_searcher(result_f)

# Retrieve and print the hazards information
hazards_retriever(ghs_data, result_f)
```

```
## [1] "No hazard statements available."
```

Conclusion: The high frequency of Flumioxazin in herbicide applications can be attributed to its broad-spectrum efficacy and effectiveness in controlling resistant weeds, despite its associated health and environmental risks. Conversely, Glufosinate-Ammonium is used less frequently due to its broader acute toxicity concerns and limited target spectrum, along with potential reproductive and organ damage risks. The choice between these herbicides reflects a balance between effectiveness, safety, and environmental stewardship, emphasizing the need for integrated pest management practices that consider the specific requirements of the agricultural context.