Chemical EDA

Suheng Yao

2024-10-19

```
# Read in the cleaned dataset and select records of California and Florida
df <- read.csv("strawberry_cleaned.csv")

# Find states that used chemicals in growing strawberry
states_with_chemical_info <- df %>%
  filter(!is.na(Chemical_Info)) %>%
  distinct(State) %>%
  pull(State)

print(states_with_chemical_info)
```

```
## [1] "CALIFORNIA" "FLORIDA"
```

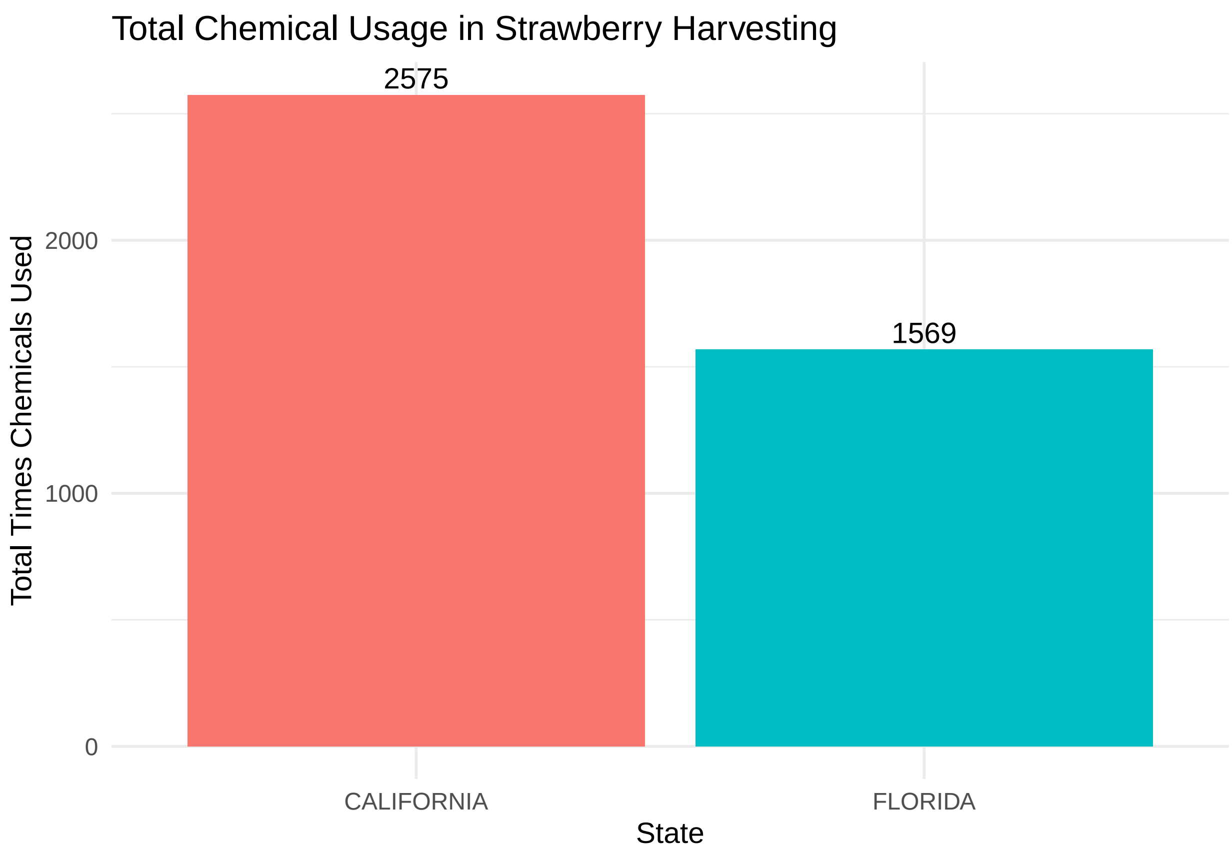
Only California and Florida used chemicals when growing strawberries, we can filter out only those two states and compare the frequency of chemical used:

```
df <- df %>%
  select(-1) %>%
  filter(State == "CALIFORNIA" | State == "FLORIDA")

# Draw a bar plot comparing the frequency
chem_state <- df %>%
  group_by(State) %>%
  summarise(Total_usage=n()) %>%
  arrange(desc(Total_usage))
print(chem_state)
```

```
## # A tibble: 2 x 2
##   State      Total_usage
##   <chr>            <int>
## 1 CALIFORNIA        2575
## 2 FLORIDA           1569
```

```
ggplot(chem_state, aes(x = State, y = Total_usage, fill = State)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Total_usage), vjust = -0.3, size = 4) +
  theme_minimal() +
  labs(title = "Total Chemical Usage in Strawberry Harvesting",
       x = "State",
       y = "Total Times Chemicals Used") +
  theme(legend.position = "none")  # Remove legend
```



The frequency that California used chemicals is almost twice as the frequency of Florida.

```r
# Split the California data and Florida data into two dataframes
df_cal <- df %>%
  filter(State == "CALIFORNIA")

df_flo <- df %>%
  filter(State == "FLORIDA")

# Find the distinct chemicals used in California and Florida
length(unique(df_cal$Chemical_Info))
```

```
## [1] 136
```

```r
length(unique(df_flo$Chemical_Info))
```

```
## [1] 117
```

```r
# Find out chemicals used by California but not used by Florida
diff_chem1 <- setdiff(unique(df_cal$Chemical_Info), unique(df_flo$Chemical_Info))
print(diff_chem1)
```

```
##  [1] "CYCLANILIPROLE"
##  [2] "PERMETHRIN"
##  [3] "ISARIA FUMOSOROSEA STRAIN FE 9901"
```

```
##  [4] "BACILLUS AMYLOLIQUEFACIENS STRAIN D747"
##  [5] "BLAD"
##  [6] "BT SUBSP KURSTAKI EVB-113-19"
##  [7] "POLYOXIN D ZINC SALT"
##  [8] "QUINOLINE"
##  [9] "TRIFLOXYSTROBIN"
## [10] "PENDIMETHALIN"
## [11] "ACEQUINOCYL"
## [12] "AZADIRACHTIN"
## [13] "BEAUVERIA BASSIANA"
## [14] "BT KURSTAKI SA-11"
## [15] "CANOLA OIL"
## [16] "CHROMOBAC SUBTSUGAE PRAA4-1 CELLS AND SPENT MEDIA"
## [17] "ETOXAZOLE"
## [18] "FENBUTATIN-OXIDE"
## [19] "NEEM OIL"
## [20] "NEEM OIL, CLAR. HYD."
## [21] "PYRIDABEN"
## [22] "CAPSICUM OLEORESIN EXTRACT"
## [23] "GARLIC OIL"
## [24] "IRON PHOSPHATE"
## [25] "METALDEHYDE"
## [26] "METAM-SODIUM"
## [27] "BACILLUS AMYLOLIQUEFACIENS MBI 600"
## [28] "BACILLUS PUMILUS"
## [29] "COPPER OCTANOATE"
## [30] "POTASSIUM BICARBON."
## [31] "STREPTOMYCES LYDICUS"
## [32] "BT KURSTAKI EG7841"
## [33] "BT SUB AIZAWAI GC-91"
## [34] "BUPROFEZIN"
## [35] "BURKHOLDERIA A396 CELLS & MEDIA"
## [36] "HELICOVERPA ZEA NPV"
## [37] "PETROLEUM DISTILLATE"
## [38] "POTASSIUM SALTS"
## [39] "PYRIPROXYFEN"
## [40] "CAPRIC ACID"
## [41] "CAPRYLIC ACID"
## [42] "MINERAL OIL"
## [43] "PAECILOMYCES FUMOSOR"
## [44] "POTASSIUM SILICATE"
## [45] "BACILLUS SUBT. GB03"
## [46] "TRICHODERMA HARZ."
## [47] "GLUFOSINATE-AMMONIUM"
## [48] "SULFENTRAZONE"
## [49] "CHLORPYRIFOS"
## [50] "SOYBEAN OIL"
## [51] "ZETA-CYPERMETHRIN"
## [52] "AUREOBASIDIUM PULLULANS DSM 14940"
## [53] "AUREOBASIDIUM PULLULANS DSM 14941"
## [54] "BT KURSTAKI SA-12"
## [55] "GLIOCLADIUM VIRENS"
## [56] "TRICHODERMA VIRENS STRAIN G-41"
## [57] "EMAMECTIN BENZOATE"
```

```
## [58] "SPIROTETRAMAT"
```

```r
diff_chem2 <- setdiff(unique(df_flo$Chemical_Info), unique(df_cal$Chemical_Info))
print(diff_chem2)
```

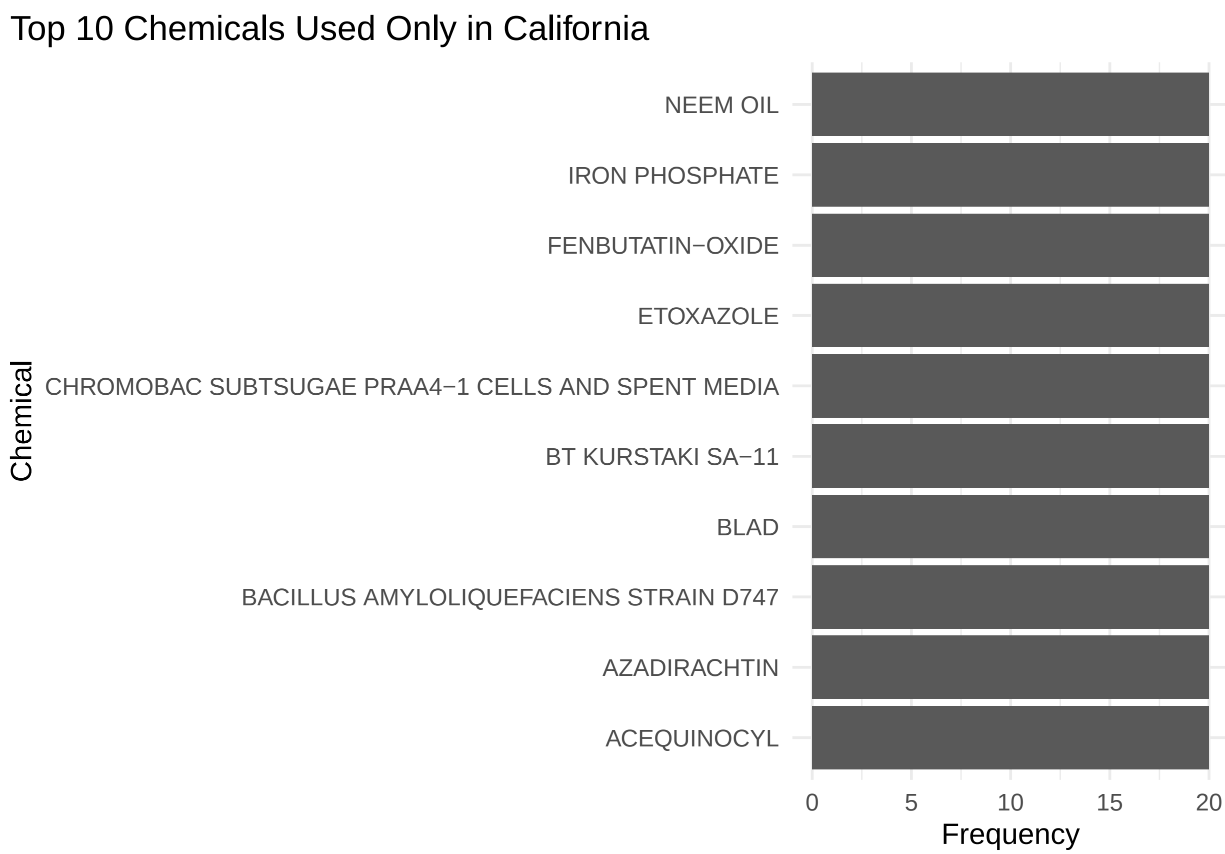
```
##  [1] "PYRIOFENONE"                "ZOXAMIDE"
##  [3] "METSULFURON-METHYL"         "PENOXSULAM"
##  [5] "S-METOLACHLOR"              "BETA-CYFLUTHRIN"
##  [7] "ETHYL 2E;4Z-DECADIENOATE" "OXAMYL"
##  [9] "CUPRAMMONIUM ACETATE"       "DODECADIEN-1-OL"
## [11] "FLUENSULFONE"               "GIBBERELLIC ACID"
## [13] "CHLOROTHALONIL"             "COPPER CHLORIDE HYD."
## [15] "CYMOXANIL"                  "FAMOXADONE"
## [17] "MANCOZEB"                   "2,4-D, DIMETH. SALT"
## [19] "CLETHODIM"                  "METHOMYL"
## [21] "CYTOKININS"                 "INDOLEBUTYRIC ACID"
## [23] "COPPER ETHANOLAMINE"        "DIMETHENAMID"
## [25] "FLUROXYPYR 1-MHE"           "HALOSULFURON-METHYL"
## [27] "KANTOR"                     "FENAZAQUIN"
## [29] "ETHEPHON"                   "DODINE"
## [31] "FLUTOLANIL"                 "2,4-D, TRIISO. SALT"
## [33] "CYPERMETHRIN"               "ALKYL. DIM. BENZ. AM"
## [35] "DECYLDIMETHYLOCTYL"         "DIDECYL DIM. AMMON."
## [37] "DIMETHYLDIOCTYL"            "MUSTARD OIL"
## [39] "DIMETHYL DISULFIDE DMDS"
```

After finding the distinct chemicals in each state, California used 19 more types of chemicals than Florida, and there are 58 types of chemicals used in California but not in Florida, on the other hand, there are 39 types of chemicals used in Florida but not in California. We can first explore those 58 chemicals used only in California:

```r
# Select records with those 58 chemicals
df_chem_cal <- df_cal %>%
  filter(Chemical_Info %in% diff_chem1)

# Create a new column to record the frequency of each chemical used
chem_freq <- df_chem_cal %>%
  count(Chemical_Info) %>%
  arrange(desc(n))

# Plot the top 10 chemicals used in California
ggplot(chem_freq[1: 10,], aes(x=reorder(Chemical_Info, n), y=n))+
  geom_bar(stat = "identity")+
  theme_minimal()+
  coord_flip()+
  labs(x = "Chemical", y = "Frequency",
       title = "Top 10 Chemicals Used Only in California")+
  theme(plot.title.position = "plot")
```



Plot above shows the top 10 most used chemicals in strawberry harvesting, we can further explore the properties of those 10 chemicals:

```
# Find the index of GHS classification
GHS_searcher<-function(result_json_object){
  # check if the chemicals in the database first
  if (is.null(result_json_object) ||
      is.null(result_json_object[["result"]]) ||
      is.null(result_json_object[["result"]][["Hierarchies"]]) ||
      is.null(result_json_object[["result"]][["Hierarchies"]][["Hierarchy"]])){
    return("did not find the chemical in the database")
  }

  result<-result_json_object
  for (i in 1:length(result[["result"]][["Hierarchies"]][["Hierarchy"]])){
    if(result[["result"]][["Hierarchies"]][["Hierarchy"]][[i]][["SourceName"]]=="GHS Classification (UN
      return(i)
    }

  }
}

# Use the output from GHS_searcher to access the hazard information
hazards_retriever<-function(index,result_json_object){

  # Check if GHS_searcher did not find the index
  if (is.character(index) && index == "did not find the chemical in the database") {
```

```
    return(index)
  }

  result<-result_json_object
  hierarchy<-result[["result"]][["Hierarchies"]][["Hierarchy"]][[index]]
  i<-1
  output_list<-rep(NA,length(hierarchy[["Node"]]))
  while(str_detect(hierarchy[["Node"]][[i]][["Information"]][["Name"]],"H") & i<length(hierarchy[["Node
    output_list[i]<-hierarchy[["Node"]][[i]][["Information"]][["Name"]]
    i<-i+1
  }
  return(output_list[!is.na(output_list)])
}

# Find the chemical information for the top 10 chemicals
chem_vec <- chem_freq[1: 10, "Chemical_Info"]
access_hazard <- function(chemical){
  results <- list()
  for (chem in chemical){
      result <- get_pug_rest(identifier = chem,
                             namespace = "name",
                             domain = "compound",
                             operation="classification",
                             output = "JSON")
      ghs_index <- GHS_searcher(result)
      hazards <- hazards_retriever(ghs_index, result)
      results[[chem]] <- list(
        chemical_name = chem,
        chemical_hazards = ifelse(hazards == "did not find the chemical in the database", character(0),
      )

  }

  return(results)
}


hazard_info <- access_hazard(chem_vec)
hazards_df <- do.call(rbind, lapply(hazard_info, function(x) {
  data.frame(
    chemical_name = x$chemical_name,
    hazards = paste(x$chemical_hazards, collapse = ";"),
    stringsAsFactors = FALSE
  )
}))

hazards_df %>%
  knitr::kable(escape = FALSE) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = FALSE,
                position = "left",
                font_size = 12) %>%
  column_spec(1, bold = TRUE) %>%
```

```
  row_spec(0, bold = TRUE, color = "white", background = "#4E79A7")
```

	chemical_name
ACEQUINOCYL	ACEQUINOCYL
AZADIRACHTIN	AZADIRACHTIN
BACILLUS AMYLOLIQUEFACIENS STRAIN D747	BACILLUS AMYL
BLAD	BLAD
BT KURSTAKI SA-11	BT KURSTAKI SA
CHROMOBAC SUBTSUGAE PRAA4-1 CELLS AND SPENT MEDIA	CHROMOBAC SU
ETOXAZOLE	ETOXAZOLE
FENBUTATIN-OXIDE	FENBUTATIN-OX
IRON PHOSPHATE	IRON PHOSPHAT
NEEM OIL	NEEM OIL

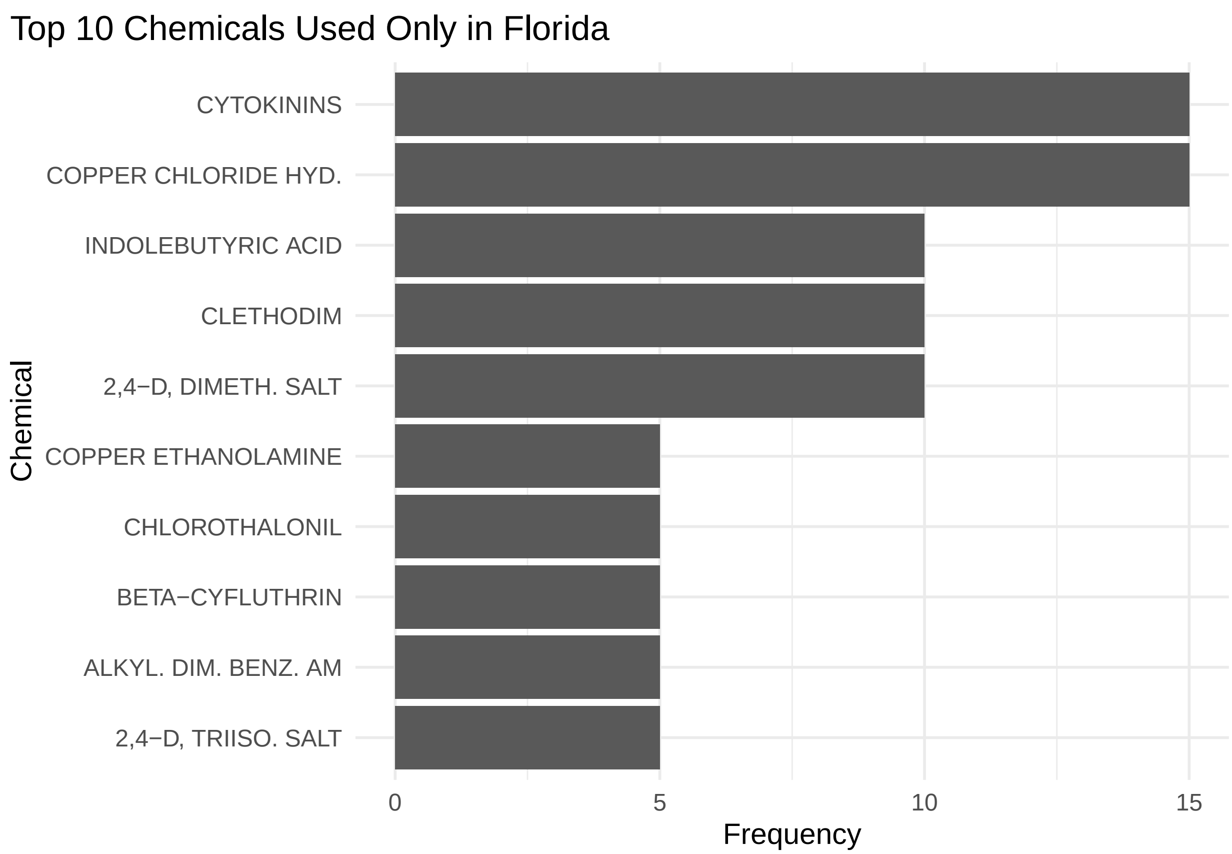
Through the chemical information presented in the above table, four out of the top ten chemicals have harmful effect on the environment. Especially for Fenbutatin oxide, which can be fatal if inhaled and cause serious organ damage. Also, most of those chemicals are very toxic to aquatic life, which will harm the sustainable development of the environment. Thus, the suggestion to related authority is to decrease the usage of those four chemicals in growing strawberry, especially for fenbutatin-oxide and acequinocyl. Although they can be used as pesticide, they do more harm to the environment and humans.

Now, we can further analyze the top 10 chemicals used in strawberry growing in Florida:

```
# Select records with Florida only chemicals
df_chem_flo <- df_flo %>%
  filter(Chemical_Info %in% diff_chem2)

# Create a new column to record the frequency of each chemical used
chem_freq_flo <- df_chem_flo %>%
  count(Chemical_Info) %>%
  arrange(desc(n))

# Plot the top 10 chemicals used in California
ggplot(chem_freq_flo[1: 10,], aes(x=reorder(Chemical_Info, n), y=n))+
  geom_bar(stat = "identity")+
  theme_minimal()+
  coord_flip()+
  labs(x = "Chemical", y = "Frequency",
       title = "Top 10 Chemicals Used Only in Florida")+
  theme(plot.title.position = "plot")
```




```r
chem_vec_flo <- c("INDOLEBUTYRIC ACID",
                  "BETA-CYFLUTHRIN", "CHLOROTHALONIL")
hazard_info_flo <- access_hazard(chem_vec_flo)

hazards_df_flo <- do.call(rbind, lapply(hazard_info_flo, function(x) {
  data.frame(
    chemical_name = x$chemical_name,
    hazards = paste(x$chemical_hazards, collapse = ";"),
    stringsAsFactors = FALSE
  )
}))

hazards_df_flo %>%
  knitr::kable(escape = FALSE) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width = FALSE,
                position = "left",
                font_size = 12) %>%
  column_spec(1, bold = TRUE) %>%
  row_spec(0, bold = TRUE, color = "white", background = "#4E79A7")
```

	chemical_name	hazards
INDOLEBUTYRIC ACID	INDOLEBUTYRIC ACID	H301: Toxic if swallowed [Danger Acute to
BETA-CYFLUTHRIN	BETA-CYFLUTHRIN	H300: Fatal if swallowed [Danger Acute to

CHLOROTHALONIL	CHLOROTHALONIL	H317: May cause an allergic skin reaction

Through the table shown above, three out of top ten chemicals used only in Florida have hazard information. The most toxic chemical is the beta-cyfluthrin, which can be fatal if swallowed and inhaled, and it also has toxic effects to aquatic lives, which is bad for the environment.

In summary, by comparing the chemicals used in California and Florida, California used more types of chemicals than Florida, also, the maximum used times of the most used chemicals in each state were also higher for California(20 vs 15). However, California and Florida both used chemicals that are toxic to the aquatic lives and environment when growing strawberry, and it is important to suggest to the goverment to restrict the amount of harmful chemicals used in insecticides and pesticides, and they should be changed to more environmental friendly substitutes.