Strawberry 2

Yibing Wang

2024-10-22

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(ggplot2)
```

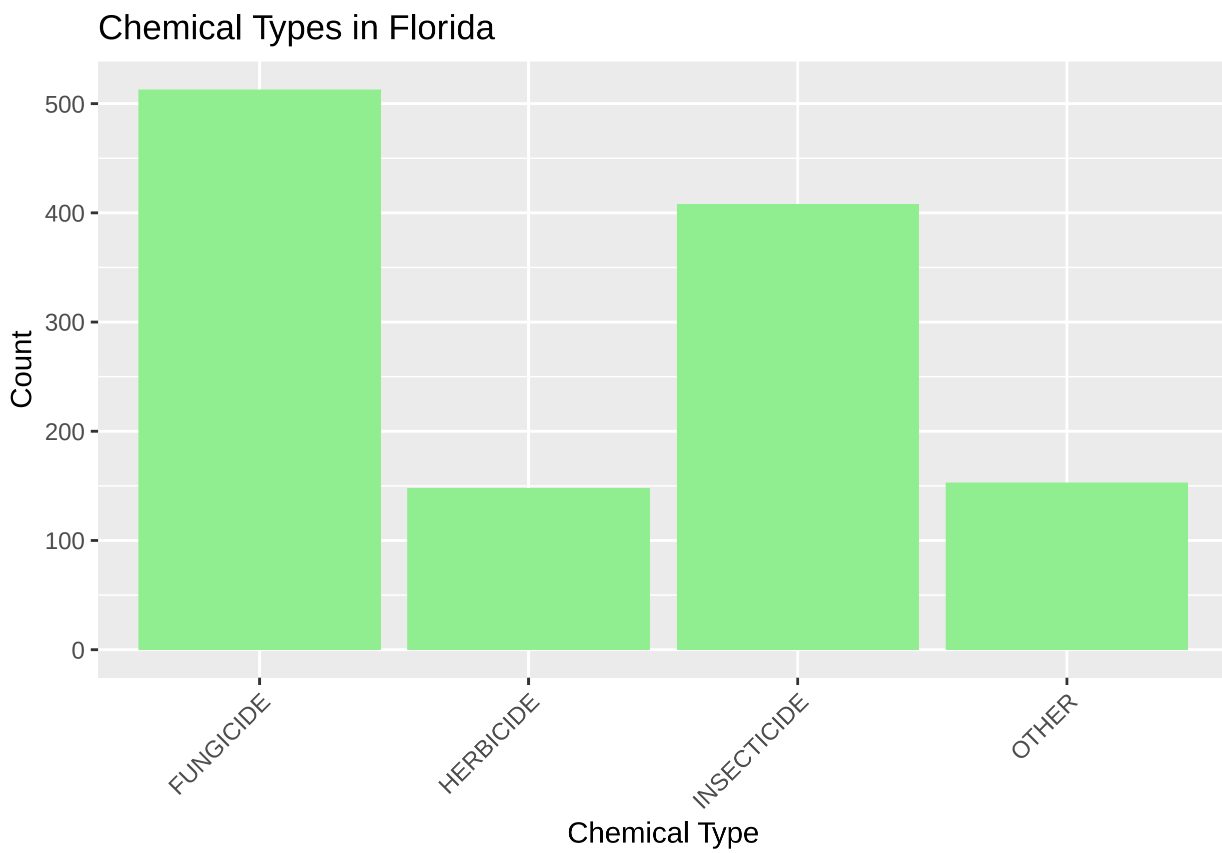
my questions: what's the most popular chemical used in Florida and California? Why we choose A(the most popular one) instead of B (the least popular one)?

```r
#import csv from ver 7
chem_data <- read.csv("chem.csv")
head(chem_data)
```

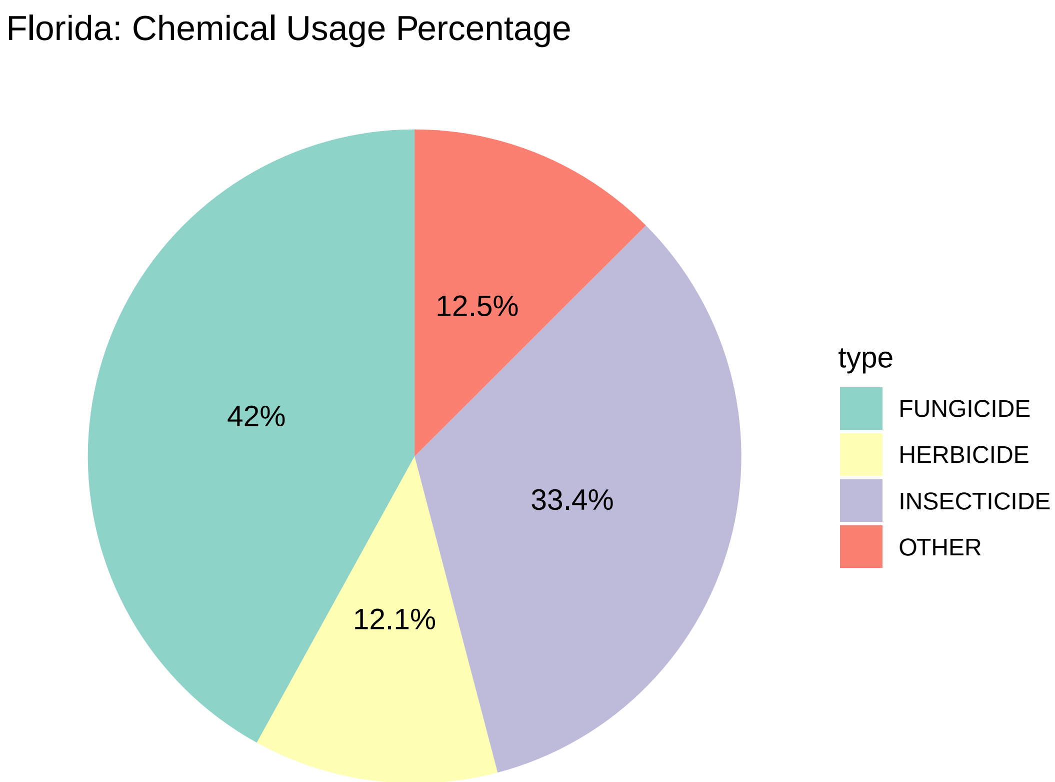
```
##   X Year      State          mk1  mk2                 measure other        type
## 1 1 2023 CALIFORNIA APPLICATIONS <NA>                      LB  <NA>   FUNGICIDE
## 2 2 2023 CALIFORNIA APPLICATIONS <NA>                      LB  <NA> INSECTICIDE
## 3 3 2023 CALIFORNIA APPLICATIONS <NA>                      LB  <NA> INSECTICIDE
## 4 4 2023 CALIFORNIA APPLICATIONS <NA>                      LB  <NA>       OTHER
## 5 5 2023 CALIFORNIA APPLICATIONS <NA> LB / ACRE / APPLICATION   AVG   FUNGICIDE
## 6 6 2023 CALIFORNIA APPLICATIONS <NA> LB / ACRE / APPLICATION   AVG INSECTICIDE
##                         chem_name chem_index Value
## 1                  OXATHIAPIPROLIN     128111   (D)
## 2                   CYCLANILIPROLE      26202   (D)
## 3                       PERMETHRIN     109701   (D)
## 4 ISARIA FUMOSOROSEA STRAIN FE 9901     115003  (NA)
## 5                  OXATHIAPIPROLIN     128111   (D)
## 6                   CYCLANILIPROLE      26202   (D)
```

```r
# I would like to know which kind of chemicals are mostly used in florida and california
#florida
florida_data <- chem_data %>%
```

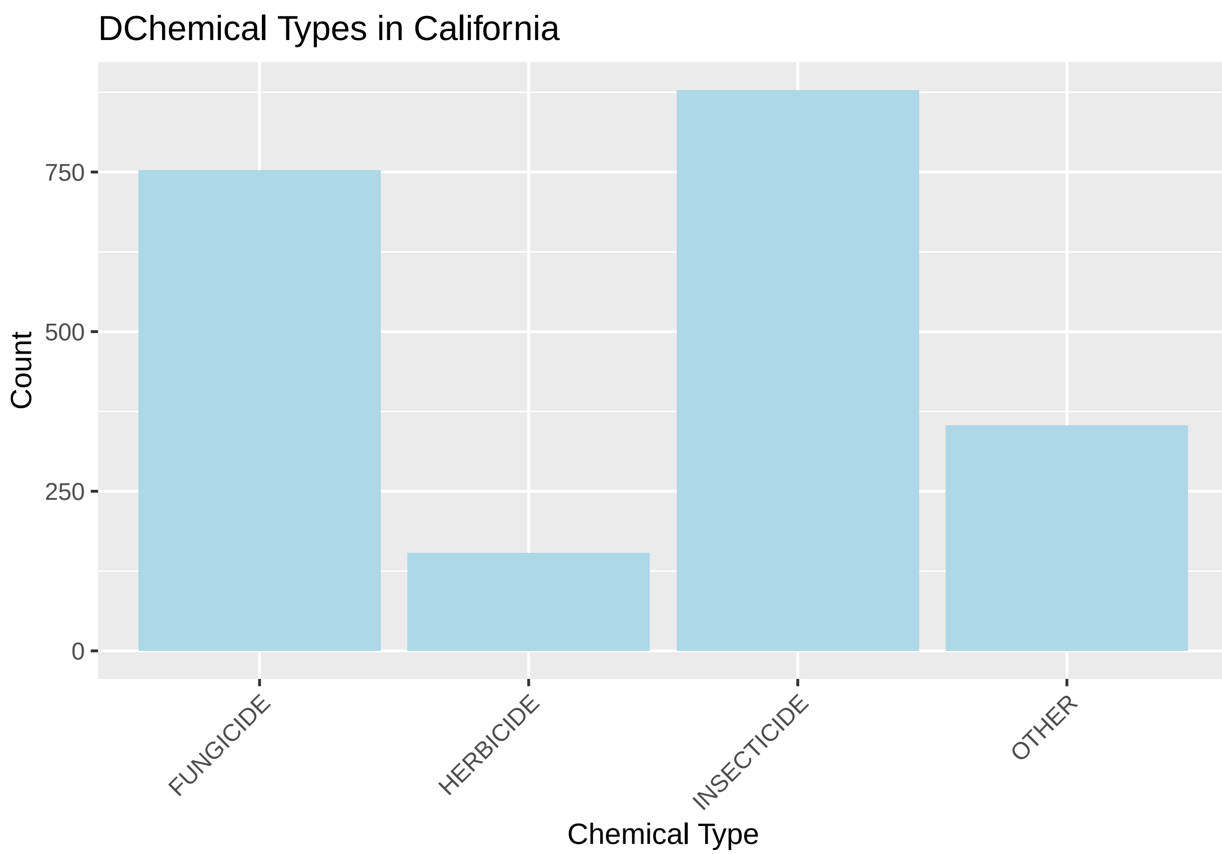
```r
  filter(State == "FLORIDA")

florida_missing <- colSums(is.na(florida_data))

ggplot(florida_data, aes(x = type)) +
  geom_bar(fill = "lightgreen") +
  ggtitle("Chemical Types in Florida") +
  xlab("Chemical Type") +
  ylab("Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



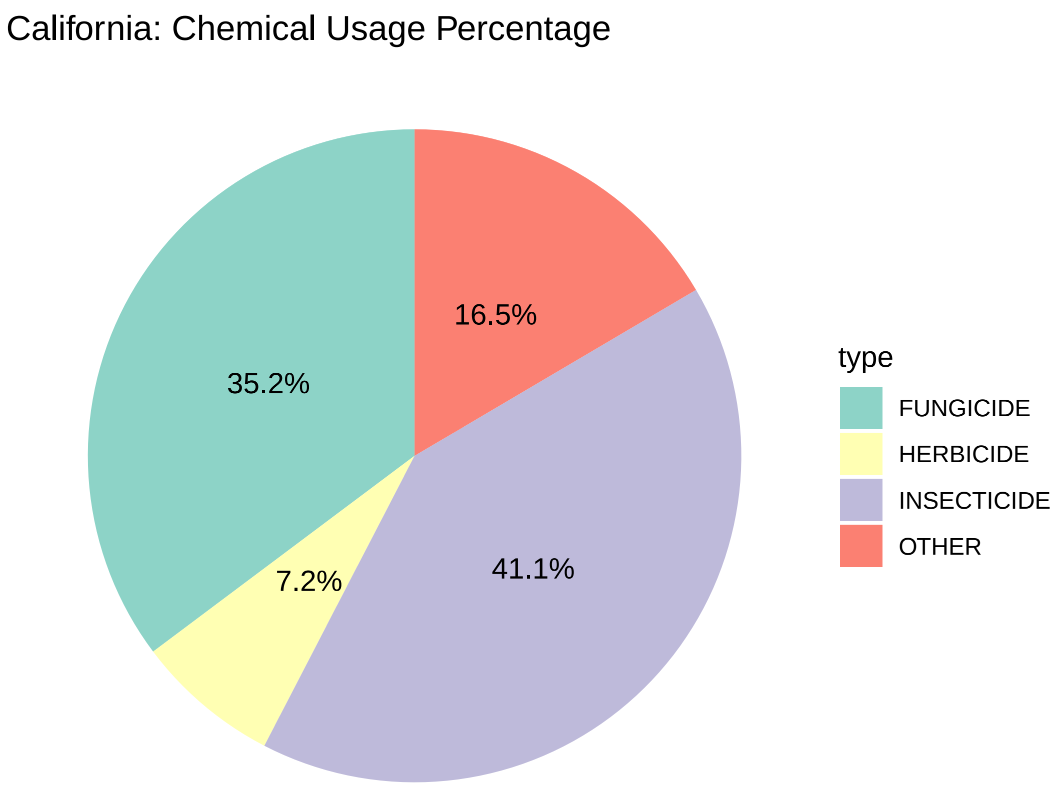

```r
#pie chart
florida_percentage <- florida_data %>%
  count(type) %>%
  mutate(percentage = n / sum(n) * 100)

ggplot(florida_percentage, aes(x = "", y = percentage, fill = type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  ggtitle("Florida: Chemical Usage Percentage") +
  theme_void() +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5)) +
  scale_fill_brewer(palette = "Set3")
```




```r
#Cali
california_data <- chem_data %>%
  filter(State == "CALIFORNIA")

california_missing <- colSums(is.na(california_data))

#pie chart
california_percentage <- california_data %>%
  count(type) %>%
  mutate(percentage = n / sum(n) * 100)

ggplot(california_data, aes(x = type)) +
  geom_bar(fill = "lightblue") +
  ggtitle("DChemical Types in California") +
  xlab("Chemical Type") +
  ylab("Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



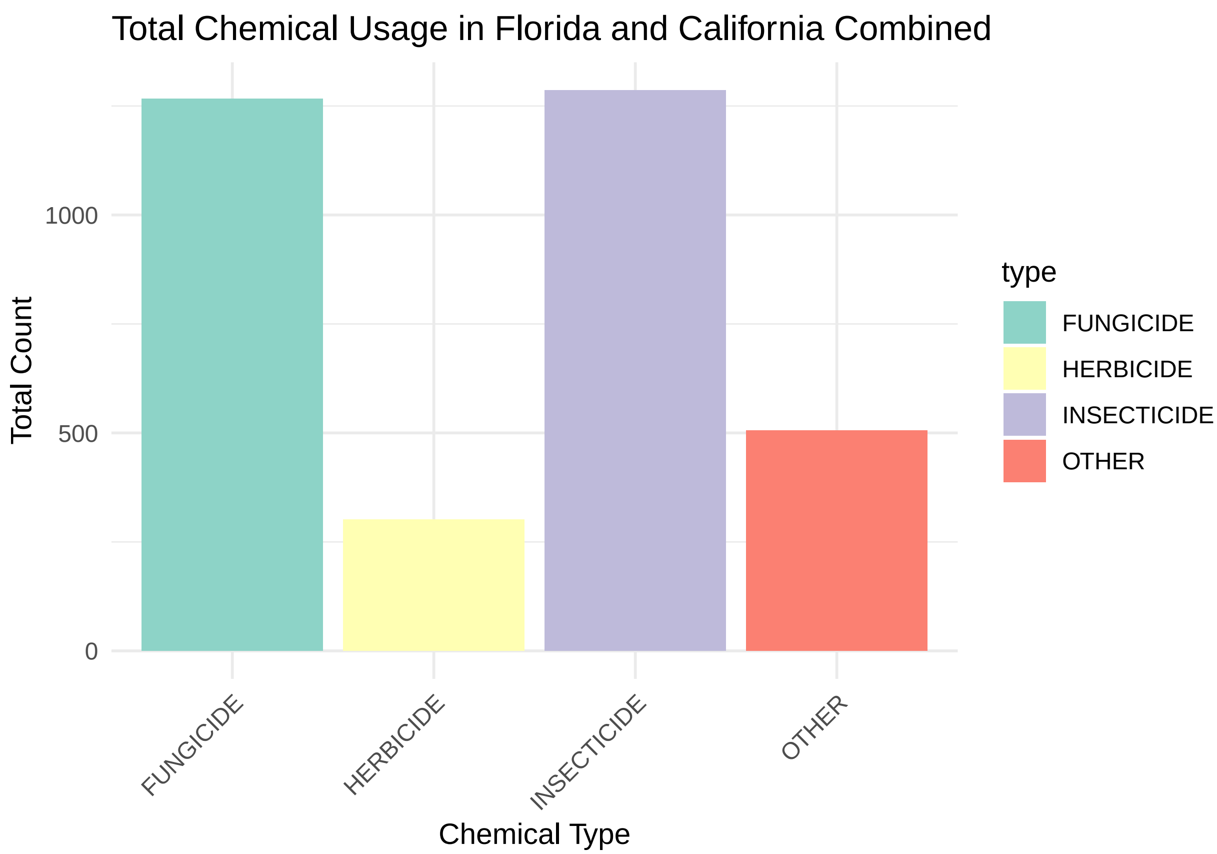

```
ggplot(california_percentage, aes(x = "", y = percentage, fill = type)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  ggtitle("California: Chemical Usage Percentage") +
  theme_void() +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5)) +
  scale_fill_brewer(palette = "Set3")
```



For florida, Fungicide are the most commonly used chemical type followed by the insecticide, but in California, insecticides are more prevalent than fungicides. Herbicide are the least used chemical in both states. As we can see from the pie chart, the difference between the fungicide and insecticide isn't big (less than 10%).

For the next step, I would like to know why would people choose fungicide and insecticide instead of using Herbicide.

```
#graph of chemical use with florida plus california
combined_data <- chem_data %>%
  filter(State %in% c("FLORIDA", "CALIFORNIA")) %>%
  group_by(type) %>%
  summarise(total_count = n())

ggplot(combined_data, aes(x = type, y = total_count, fill = type)) +
  geom_bar(stat = "identity") +
  ggtitle("Total Chemical Usage in Florida and California Combined") +
  xlab("Chemical Type") +
  ylab("Total Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")
```



For fungicide, insecticide and herbicide, find the most popular chemical for each type.

```
filtered_data <- chem_data %>%
  filter(State %in% c("FLORIDA", "CALIFORNIA"))

popular_chemicals <- filtered_data %>%
  group_by(type, chem_name) %>%
  summarise(count = n())
```

```
## `summarise()` has grouped output by 'type'. You can override using the
## `.groups` argument.
```

```
popular_chemicals <- popular_chemicals %>%
  group_by(type) %>%
  filter(count == max(count)) %>%
  ungroup()

print(popular_chemicals)
```

```
## # A tibble: 26 x 3
##    type      chem_name          count
##    <chr>     <chr>              <int>
##  1 FUNGICIDE AZOXYSTROBIN          40
##  2 FUNGICIDE CAPTAN                40
##  3 FUNGICIDE CYPRODINIL            40
```

```
## 4 FUNGICIDE DIFENOCONAZOLE        40
## 5 FUNGICIDE FENHEXAMID            40
## 6 FUNGICIDE FLUDIOXONIL           40
## 7 FUNGICIDE MEFENOXAM             40
## 8 FUNGICIDE PYRIMETHANIL          40
## 9 FUNGICIDE THIOPHANATE-METHYL    40
## 10 FUNGICIDE THIRAM               40
## # i 16 more rows
```

using azoxystrobin for fungicide, flumioxazin for herbicide and abamectin for insecticide as examples for each type, find ad compare the hazards using PubChemR from QMD

```r
#from qmd
library(tidyverse)
library(PubChemR)

GHS_searcher <- function(result_json_object) {
  hierarchies <- result_json_object[["result"]][["Hierarchies"]][["Hierarchy"]]

  for (i in seq_along(hierarchies)) {
    if (hierarchies[[i]][["SourceName"]] == "GHS Classification (UNECE)") {
      return(i)
    }
  }
  # Return NULL if GHS Classification is not found
  return(NULL)
}

hazards_retriever <- function(index, result_json_object) {
  if (is.null(index)) {
    return(NA)  # Return NA if GHS data is not available
  }

  hierarchy <- result_json_object[["result"]][["Hierarchies"]][["Hierarchy"]][[index]]
  nodes <- hierarchy[["Node"]]
  hazard_statements <- c()
  i <- 1

  while (i <= length(nodes) && str_detect(nodes[[i]][["Information"]][["Name"]], "^H")) {
    hazard_statements <- c(hazard_statements, nodes[[i]][["Information"]][["Name"]])
    i <- i + 1
  }
  if (length(hazard_statements) == 0) {
    return(NA)
  }
  return(hazard_statements)
}

# List of chemicals to process
chemical_vec <- c("azoxystrobin", "flumioxazin", "abamectin")

# Initialize an empty list to store results
results_list <- list()
```

```r
for (chemical in chemical_vec) {
  result <- get_pug_rest(
    identifier = chemical,
    namespace = "name",
    domain = "compound",
    operation = "classification",
    output = "JSON"
  )

  ghs_index <- GHS_searcher(result)
  hazards <- hazards_retriever(ghs_index, result)

  # Store the results in a list
  results_list[[chemical]] <- hazards
}

# Convert the results list into a data frame
results_df <- results_list %>%
  enframe(name = "Chemical", value = "Hazard_Statements") %>%
  unnest(cols = c(Hazard_Statements))

# Display the data frame
print(results_df)
```

```
## # A tibble: 23 x 2
##    Chemical     Hazard_Statements
##    <chr>        <chr>
##  1 azoxystrobin H331: Toxic if inhaled [Danger Acute toxicity, inhalation]
##  2 azoxystrobin H300: Health Hazards
##  3 azoxystrobin Hazard Statement Codes
##  4 azoxystrobin H370: Causes damage to organs [Danger Specific target organ tox~
##  5 azoxystrobin H400: Very toxic to aquatic life [Warning Hazardous to the aqua~
##  6 azoxystrobin H400: Environmental Hazards
##  7 azoxystrobin H410: Very toxic to aquatic life with long lasting effects [War~
##  8 flumioxazin  H360: May damage fertility or the unborn child [Danger Reproduc~
##  9 flumioxazin  H300: Health Hazards
## 10 flumioxazin  Hazard Statement Codes
## # i 13 more rows
```

The hazard_statements basically explain my questions. for fungicide (azoxystrobin) and insecticide (abamectin), herbicide (flumioxazin)

All three kinds of chemicals are harmful to human, but herbicide has both acute and chronic health risks.

Same for the environment, Fungicides and insecticides often present high environmental risks, especially in terms of aquatic toxicity, herbicide has long-term environmental risk. Also, the chronic environmental impact of herbicides can contribute to lower usage rates. Fungal and pests can spread rapidly and cause economic losses, but weeds' controlling is usually not that urgent. There are also some non-chemical methods to use instead of herbicide.

In conclusion, Farmers are more likely to choose fungicides and insecticides more frequently than herbicides because of the immediate crop protection they provide against significant threats like fungal diseases and pests. Even though all three chemicals are harmful to environment and human, herbicide tend to have more chronic long-term risks.

For future tasks, I think it would be helpful to compare the potential crop losses due to pests, fungal and weeds diseases. Analyzing how sensitive the major crops grown in Florida and California are to pests, fungal and weeds diseases.