

Assignment 1

Jinran Li

2024-10-02

```
library(knitr)
library(kableExtra)
library(tidyverse)
library(stringr)
library(dplyr)
library(readr)
```

```
strawberry <- read_csv("strawberries25_v3.csv", col_names = TRUE)
```

```
## Rows: 12669 Columns: 21
## -- Column specification -----
## Delimiter: ","
## chr (15): Program, Period, Geo Level, State, State ANSI, Ag District, County...
## dbl (2): Year, Ag District Code
## lgl (4): Week Ending, Zip Code, Region, Watershed
##
## 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.
```

```
glimpse(strawberry)
```

```
## Rows: 12,669
## Columns: 21
## $ Program      <chr> "CENSUS", "CENSUS", "CENSUS", "CENSUS", "CENSUS", "~
## $ Year         <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 202~
## $ Period       <chr> "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YE~
## $ 'Week Ending' <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ 'Geo Level'   <chr> "COUNTY", "COUNTY", "COUNTY", "COUNTY", "COUNTY", "~
## $ State        <chr> "ALABAMA", "ALABAMA", "ALABAMA", "ALABAMA", "ALABAM~
## $ 'State ANSI'  <chr> "01", "01", "01", "01", "01", "01", "01", "01", "01~
## $ 'Ag District' <chr> "BLACK BELT", "BLACK BELT", "BLACK BELT", "BLACK BE~
## $ 'Ag District Code' <dbl> 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40, 40,~
## $ County       <chr> "BULLOCK", "BULLOCK", "BULLOCK", "BULLOCK", "BULLOC~
## $ 'County ANSI' <chr> "011", "011", "011", "011", "011", "011", "101", "1~
## $ 'Zip Code'    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Region       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ watershed_code <chr> "00000000", "00000000", "00000000", "00000000", "00~
## $ Watershed     <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ Commodity     <chr> "STRAWBERRIES", "STRAWBERRIES", "STRAWBERRIES", "ST~
## $ 'Data Item'   <chr> "STRAWBERRIES - ACRES BEARING", "STRAWBERRIES - ACR~
```

```
## $ Domain          <chr> "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL~
## $ 'Domain Category' <chr> "NOT SPECIFIED", "NOT SPECIFIED", "NOT SPECIFIED", ~
## $ Value           <chr> "(D)", "3", "(D)", "1", "6", "5", "(D)", "(D)", "2"~
## $ 'CV (%)'        <chr> "(D)", "15.7", "(D)", "(L)", "52.7", "47.6", "(D)", ~
```

```
strawberry <- strawberry |>
  filter(`Geo Level` == "NATIONAL" | `Geo Level` == "STATE")
```

```
state_acres <- strawberry %>%
  filter(State %in% c("CALIFORNIA", "FLORIDA")) %>%
  filter(str_detect(`Data Item`, "ACRES")) %>%
  mutate(Type = ifelse(str_detect(`Data Item`, "BEARING"), "Conventional", "Organic")) %>%
  select(State, Type, Value) %>%
  mutate(Value = as.numeric(Value)) %>%
  group_by(State, Type) %>%
  summarize(Total_Acres = sum(Value, na.rm = TRUE), .groups = "drop") %>%
  ungroup()
```

```
## Warning: There was 1 warning in 'mutate()'.
## i In argument: 'Value = as.numeric(Value)'.
## Caused by warning:
## ! NAs introduced by coercion
```

```
# Plotting the data
ggplot(state_acres, aes(x = State, y = Total_Acres, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Strawberry Acreage in California and Florida",
       x = "State",
       y = "Total Acres") +
  theme_minimal()
```

A bar chart comparing the total acres of conventional and organic farming in California and Florida. The y-axis is labeled 'Total Acres' and ranges from 0 to 2000 with major gridlines every 500 units. The x-axis is labeled 'State' and has two categories: 'CALIFORNIA' and 'FLORIDA'. For each state, there are two bars: a salmon-colored bar for 'Conventional' and a teal-colored bar for 'Organic'. In California, the conventional bar is approximately 1950 acres and the organic bar is approximately 1300 acres. In Florida, the conventional bar is approximately 500 acres and the organic bar is approximately 2150 acres. A legend on the right side of the chart, titled 'Type', shows a salmon square for 'Conventional' and a teal square for 'Organic'.

State	Conventional	Organic
CALIFORNIA	1950	1300
FLORIDA	500	2150

```
## # A tibble: 6 x 24
##   Program Year Period 'Week Ending' 'Geo Level' State 'State ANSI'
```

```
##   <chr>   <dbl> <chr>   <lgl>           <chr>           <chr>           <chr>
## 1 SURVEY   2023 YEAR   NA               STATE           CALIFORNIA 06
## 2 SURVEY   2023 YEAR   NA               STATE           CALIFORNIA 06
## 3 SURVEY   2023 YEAR   NA               STATE           CALIFORNIA 06
## 4 SURVEY   2023 YEAR   NA               STATE           CALIFORNIA 06
## 5 SURVEY   2023 YEAR   NA               STATE           CALIFORNIA 06
## 6 SURVEY   2023 YEAR   NA               STATE           CALIFORNIA 06
## # i 17 more variables: 'Ag District' <chr>, 'Ag District Code' <dbl>,
## #   County <chr>, 'County ANSI' <chr>, 'Zip Code' <lgl>, Region <lgl>,
## #   watershed_code <chr>, Watershed <lgl>, Commodity <chr>, 'Data Item' <chr>,
## #   Domain <chr>, 'Domain Category' <chr>, Value <chr>, 'CV (%)' <chr>,
## #   Use <chr>, Name <chr>, Code <chr>
```

```
use_counts <- strawberry_cleaned %>%
  group_by(Use) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

print(use_counts)
```

```
## # A tibble: 5 x 2
##   Use      Count
##   <chr>    <int>
## 1 INSECTICIDE 1270
## 2 FUNGICIDE 1250
## 3 OTHER      490
## 4 HERBICIDE 285
## 5 <NA>      64
```

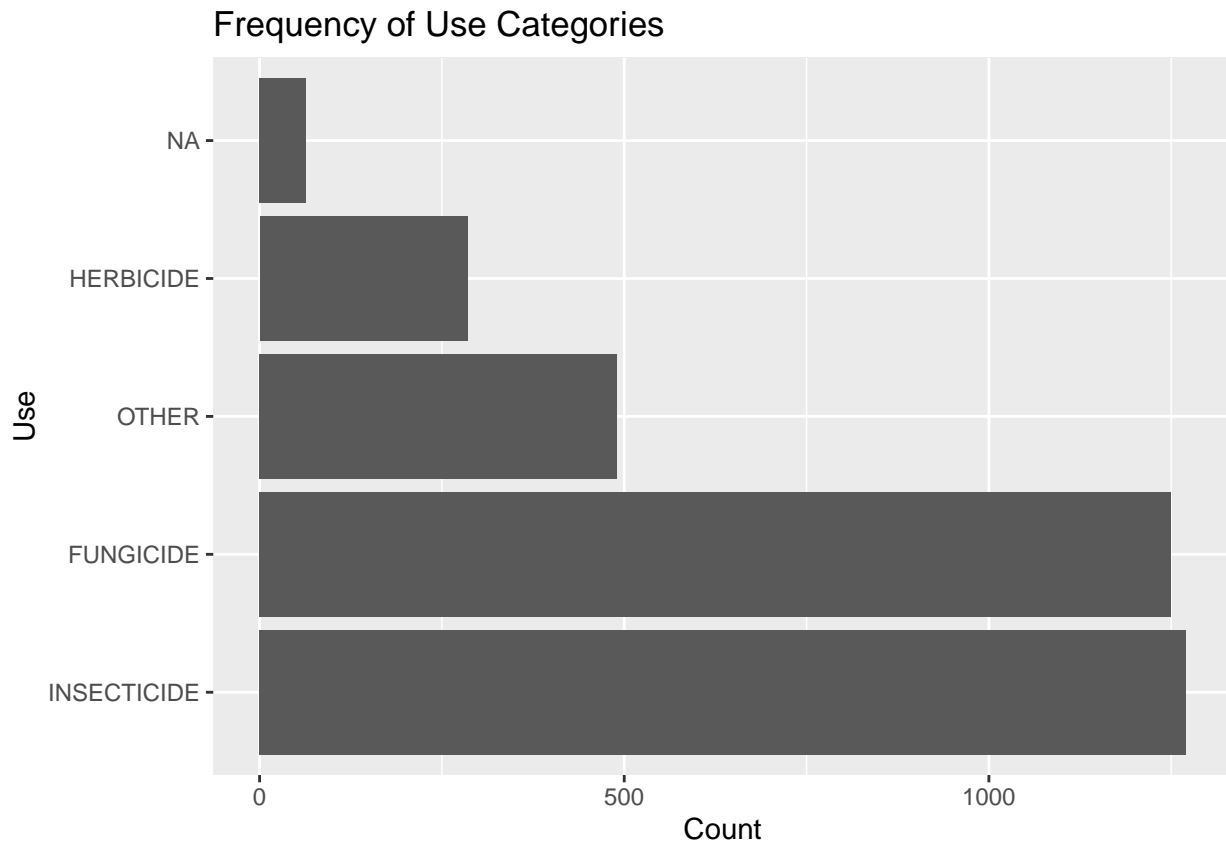
```
strawberry_cleaned <- strawberry_cleaned %>%
  filter(!is.na(Name))

name_counts <- strawberry_cleaned %>%
  group_by(Name) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

print(name_counts)
```

```
## # A tibble: 174 x 2
##   Name      Count
##   <chr>    <int>
## 1 "ABAMECTIN "      40
## 2 "ACETAMIPRID "    40
## 3 "AZOXYSTROBIN "   40
## 4 "BIFENAZATE "     40
## 5 "BIFENTHRIN "     40
## 6 "CAPTAN "         40
## 7 "CHLORANTRANILIPROLE " 40
## 8 "CYPRODINIL "     40
## 9 "DIFENOCONAZOLE "    40
## 10 "FENHEXAMID "     40
## # i 164 more rows
```

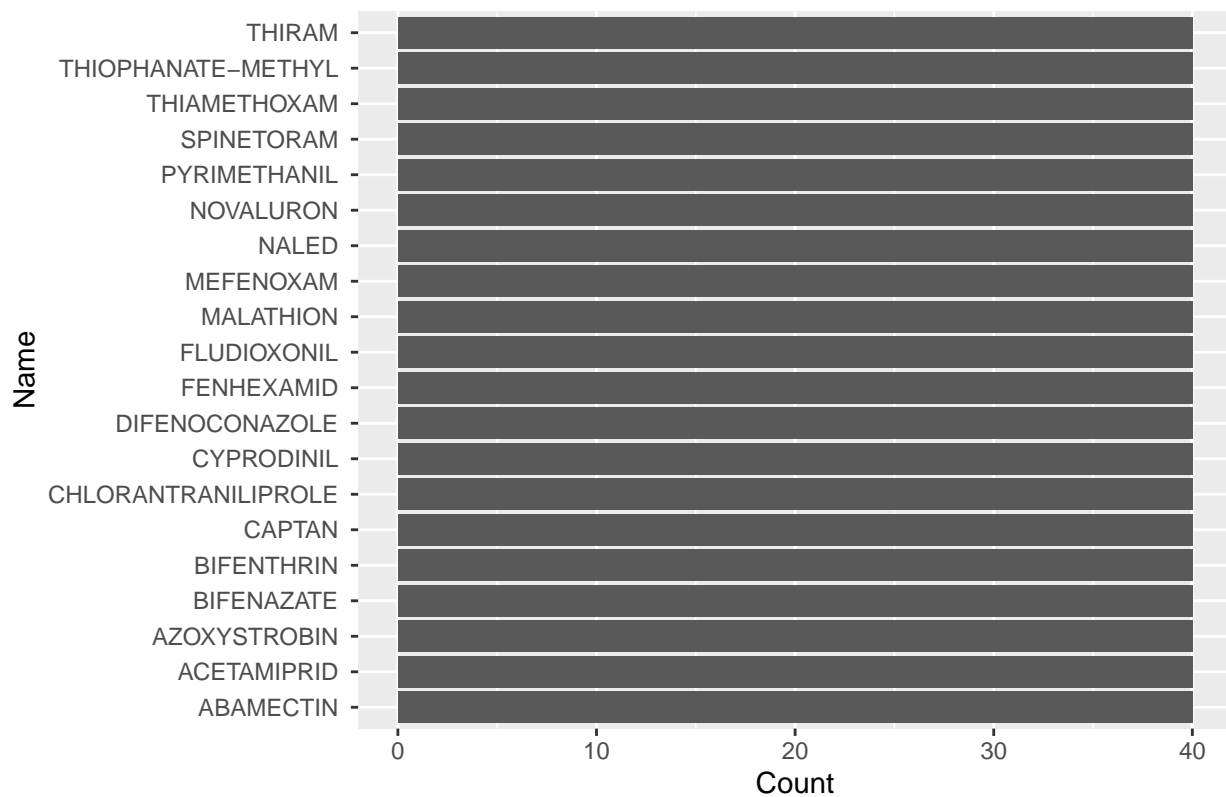
```
ggplot(use_counts, aes(x = reorder(Use, -Count), y = Count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Frequency of Use Categories", x = "Use", y = "Count")
```



```
top_n_names <- name_counts %>%
  top_n(10, wt = Count)

ggplot(top_n_names, aes(x = reorder(Name, -Count), y = Count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Top 10 Most Frequent Name Categories", x = "Name", y = "Count")
```

Top 10 Most Frequent Name Categories



```
insecticide_data <- read.csv("DTXSID9020243 Hazard Point of Departure for human-2024-10-21.csv")
fungicide_data <- read.csv("DTXSID1032359 Hazard Point of Departure for human-2024-10-21.csv")
```

```
insecticide_summary <- insecticide_data %>%
  mutate(Main_Effect = str_extract(Critical.effect, "[^/|]+")) %>%
  group_by(Main_Effect) %>%
  summarize(
    Mean_Dose = mean(as.numeric(Value), na.rm = TRUE),
    Occurrences = n()
  ) %>%
  arrange(desc(Occurrences))
```

```
fungicide_summary <- fungicide_data %>%
  mutate(Main_Effect = str_extract(Critical.effect, "[^/|]+")) %>%
  group_by(Main_Effect) %>%
  summarize(
    Mean_Dose = mean(as.numeric(Value), na.rm = TRUE),
    Occurrences = n()
  ) %>%
  arrange(desc(Occurrences))
```

```
cat("INSECTICIDE Main Health Effects Summary:\n")
```

```
## INSECTICIDE Main Health Effects Summary:
```

```
print(insecticide_summary)
```

```
## # A tibble: 13 x 3
##   Main_Effect          Mean_Dose Occurrences
##   <chr>              <dbl>      <int>
## 1 pathology microscopic-intestine small    798.        18
## 2 in life observation-body weight    285.        12
## 3 -                79.7         11
## 4 in life observation-clinical signs    262.        10
## 5 reproductive performance-postimplantation loss    58.3         6
## 6 development        37.5         4
## 7 developmental malformation-bone      20         4
## 8 organ weight-uterus    82.5         4
## 9 body weight          25         2
## 10 hematology-lymphocyte    857         2
## 11 Other:Decreased mean body weights    12.5         1
## 12 mortality            10         1
## 13 reduced weight       12.5         1
```

```
cat("\nFUNGICIDE Main Health Effects Summary:\n")
```

```
##
## FUNGICIDE Main Health Effects Summary:
```

```
print(fungicide_summary)
```

```
## # A tibble: 24 x 3
##   Main_Effect          Mean_Dose Occurrences
##   <chr>              <dbl>      <int>
## 1 in life observation-body weight    299.        45
## 2 organ weight-liver    390.        20
## 3 pathology microscopic-kidney    283.        13
## 4 developmental malformation-bone    328.        12
## 5 pathology microscopic-liver    242.        12
## 6 organ weight-kidney      37.6        11
## 7 organ weight-adrenal gland    639.         8
## 8 organ weight-heart    243.         8
## 9 hematology-platelet    441.         5
## 10 hematology-leukocyte (wbc) count differential    701         4
## # i 14 more rows
```