Strawberries2_EDA

2024-10-27

Strawberries: Data

This is a project about acquiring strawberry data from the USDA-NASS system and then cleaning, organizing, and exploring the data in preparation for data analysis. To get started, I acquired the data from the USDA NASS system and downloaded them in a csv.

Data cleaning and organization references

"An introduction to data cleaning with R" by Edwin de Jonge and Mark van der Loo

"Problems, Methods, and Challenges in Comprehensive Data Cleansing" by Heiko Müller and Johann-Christoph Freytag

Questions about Strawberries

How are the chemicals classified (e.g., fungicides, insecticides), and which categories are most prevalent? Do certain chemical classes correlate with higher productivity or specific outcomes (e.g., fruit size or yield)?

##Data Cleaning for use

```
# Load libraries
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
```

```
library(tidyr)
library(stringr)
library(readr)

# Load the dataset
strawberries_data <- read_csv("strawberries25_v.csv")
```

```
## 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.
```

```
# View the structure of the dataset
str(strawberries_data)
```

```
## spc_tbl_ [12,669 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Program         : chr [1:12669] "CENSUS" "CENSUS" "CENSUS" "CENSUS" ...
##  $ Year            : num [1:12669] 2022 2022 2022 2022 2022 ...
##  $ Period          : chr [1:12669] "YEAR" "YEAR" "YEAR" "YEAR" ...
##  $ Week Ending     : logi [1:12669] NA NA NA NA NA NA ...
##  $ Geo Level       : chr [1:12669] "COUNTY" "COUNTY" "COUNTY" "COUNTY" ...
##  $ State           : chr [1:12669] "ALABAMA" "ALABAMA" "ALABAMA" "ALABAMA" ...
##  $ State ANSI      : chr [1:12669] "01" "01" "01" "01" ...
##  $ Ag District     : chr [1:12669] "BLACK BELT" "BLACK BELT" "BLACK BELT" "BLACK BELT" ...
##  $ Ag District Code: num [1:12669] 40 40 40 40 40 40 40 40 40 40 ...
##  $ County          : chr [1:12669] "BULLOCK" "BULLOCK" "BULLOCK" "BULLOCK" ...
##  $ County ANSI     : chr [1:12669] "011" "011" "011" "011" ...
##  $ Zip Code        : logi [1:12669] NA NA NA NA NA NA ...
##  $ Region          : logi [1:12669] NA NA NA NA NA NA ...
##  $ watershed_code  : chr [1:12669] "00000000" "00000000" "00000000" "00000000" ...
##  $ Watershed       : logi [1:12669] NA NA NA NA NA NA ...
##  $ Commodity       : chr [1:12669] "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" ...
##  $ Data Item       : chr [1:12669] "STRAWBERRIES - ACRES BEARING" "STRAWBERRIES - ACRES GROWN" "STRA
##  $ Domain          : chr [1:12669] "TOTAL" "TOTAL" "TOTAL" "TOTAL" ...
##  $ Domain Category : chr [1:12669] "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" .
##  $ Value           : chr [1:12669] "(D)" "3" "(D)" "1" ...
##  $ CV (%)          : chr [1:12669] "(D)" "15.7" "(D)" "(L)" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Program = col_character(),
##   ..   Year = col_double(),
##   ..   Period = col_character(),
##   ..   `Week Ending` = col_logical(),
##   ..   `Geo Level` = col_character(),
##   ..   State = col_character(),
##   ..   `State ANSI` = col_character(),
##   ..   `Ag District` = col_character(),
##   ..   `Ag District Code` = col_double(),
##   ..   County = col_character(),
##   ..   `County ANSI` = col_character(),
##   ..   `Zip Code` = col_logical(),
##   ..   Region = col_logical(),
##   ..   watershed_code = col_character(),
##   ..   Watershed = col_logical(),
```

```
##   ..   Commodity = col_character(),
##   ..   `Data Item` = col_character(),
##   ..   Domain = col_character(),
##   ..   `Domain Category` = col_character(),
##   ..   Value = col_character(),
##   ..   `CV (%)` = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# Check the first few rows
head(strawberries_data)
```

```
## # A tibble: 6 x 21
##   Program  Year Period `Week Ending` `Geo Level` State   `State ANSI`
##   <chr>   <dbl> <chr>  <lgl>         <chr>       <chr>   <chr>
## 1 CENSUS   2022 YEAR   NA            COUNTY      ALABAMA 01
## 2 CENSUS   2022 YEAR   NA            COUNTY      ALABAMA 01
## 3 CENSUS   2022 YEAR   NA            COUNTY      ALABAMA 01
## 4 CENSUS   2022 YEAR   NA            COUNTY      ALABAMA 01
## 5 CENSUS   2022 YEAR   NA            COUNTY      ALABAMA 01
## 6 CENSUS   2022 YEAR   NA            COUNTY      ALABAMA 01
## # i 14 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>
```

```
# Get a summary of the dataset
summary(strawberries_data)
```

```
##    Program               Year         Period          Week Ending
##  Length:12669       Min.   :2018   Length:12669       Mode:logical
##  Class :character   1st Qu.:2021   Class :character   NA's:12669
##  Mode  :character   Median :2022   Mode  :character
##                     Mean   :2021
##                     3rd Qu.:2022
##                     Max.   :2024
##
##   Geo Level            State            State ANSI        Ag District
##  Length:12669       Length:12669       Length:12669       Length:12669
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Ag District Code    County           County ANSI        Zip Code
##  Min.   :10.00    Length:12669       Length:12669       Mode:logical
##  1st Qu.:20.00    Class :character   Class :character   NA's:12669
##  Median :50.00    Mode  :character   Mode  :character
##  Mean   :46.18
##  3rd Qu.:62.00
##  Max.   :96.00
##  NA's   :5359
```

```
##   Region        watershed_code     Watershed       Commodity
##  Mode:logical   Length:12669       Mode:logical   Length:12669
##  NA's:12669     Class :character   NA's:12669     Class :character
##                 Mode  :character                  Mode  :character
##
##
##
##
##   Data Item            Domain          Domain Category       Value
##  Length:12669       Length:12669       Length:12669       Length:12669
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     CV (%)
##  Length:12669
##  Class :character
##  Mode  :character
##
##
##
##
```

```
# Rename columns to more readable names if necessary
colnames(strawberries_data) <- str_replace_all(colnames(strawberries_data), "\\s+", "_")

# Check for missing values in each column
colSums(is.na(strawberries_data))
```

```
##          Program             Year           Period      Week_Ending
##                0                0                0            12669
##        Geo_Level            State       State_ANSI      Ag_District
##                0                0              264             5359
## Ag_District_Code           County      County_ANSI         Zip_Code
##             5359             5359             5385            12669
##           Region   watershed_code        Watershed        Commodity
##            12669                0            12669                0
##        Data_Item           Domain  Domain_Category            Value
##                0                0                0                0
##           CV_(%)
##             3965
```

```
# Fill missing values (example: filling with median for numerical columns)
strawberries_data <- strawberries_data %>%
    mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm = TRUE), .)))

# For categorical variables, you may fill missing values with a placeholder like "Unknown"
strawberries_data <- strawberries_data %>%
    mutate(across(where(is.character), ~ ifelse(is.na(.), "Unknown", .)))

# List all column names
colnames(strawberries_data)
```

```
##  [1] "Program"          "Year"             "Period"           "Week_Ending"
##  [5] "Geo_Level"        "State"            "State_ANSI"       "Ag_District"
##  [9] "Ag_District_Code" "County"           "County_ANSI"      "Zip_Code"
## [13] "Region"           "watershed_code"   "Watershed"        "Commodity"
## [17] "Data_Item"        "Domain"           "Domain_Category"  "Value"
## [21] "CV_(%)"
```

```r
# Drop irrelevant columns for chemical analysis using backticks for special characters
strawberries_data <- strawberries_data %>%
    select(-c(`Ag_District`, `Ag_District_Code`, `County`, `County_ANSI`, `Zip_Code`, `watershed_code`,


# Filter out rows based on specific conditions if needed (e.g., removing entries with irrelevant region
strawberries_data <- strawberries_data %>%
    filter(!Region %in% c("Irrelevant_Region1", "Irrelevant_Region2"))

# Step 2: Clean and organize the 'Use', 'Name', and 'Code' columns, and remove 'Domain' and 'Domain_Cat
strawberry_clean <- strawberries_data %>%
  # Extract 'Use' from the 'Domain' column
  mutate(
    Use = case_when(
      str_detect(`Domain`, "FUNGICIDE") ~ "FUNGICIDE",
      str_detect(`Domain`, "INSECTICIDE") ~ "INSECTICIDE",
      str_detect(`Domain`, "HERBICIDE") ~ "HERBICIDE",
      TRUE ~ NA_character_
    ),
    # Extract 'Name' from the 'Domain_Category' column, removing the '= CODE' part
    Name = str_extract(`Domain_Category`, "\\((.*?)\\)"),
    Name = str_replace_all(Name, " = \\d+", ""),  # Remove the '= CODE' part
    Name = str_replace_all(Name, "[()]", ""),    # Remove parentheses around 'Name'
    # Extract 'Code' from the 'Domain_Category' column (after the '=' sign)
    Code = str_extract(`Domain_Category`, "\\d+"),  # Extract only the numeric part of the code
    Code = str_trim(Code)  # Clean up any remaining whitespace
  ) %>%
  # Remove rows where 'Use', 'Name', or 'Code' are NA
  drop_na(Use, Name, Code) %>%
  # Remove the unwanted 'Domain' and 'Domain_Category' columns
  select(-Domain, -`Domain_Category`)

# Detect and remove duplicates
strawberries_data <- strawberries_data %>%
    distinct()

# Save the cleaned dataset
write_csv(strawberries_data, "strawberries_cleaned.csv")
```

Answering Q1. I am going to be using bar charts to show which chemicals are used, as well as what chemicals are in which category, and frequency for each chemicals.

```r
# Load necessary libraries
library(dplyr)
library(ggplot2)
library(readr)
library(stringr)
```

```r
# Load the cleaned dataset
strawberry_clean <- read_csv("strawberries_cleaned.csv")
```

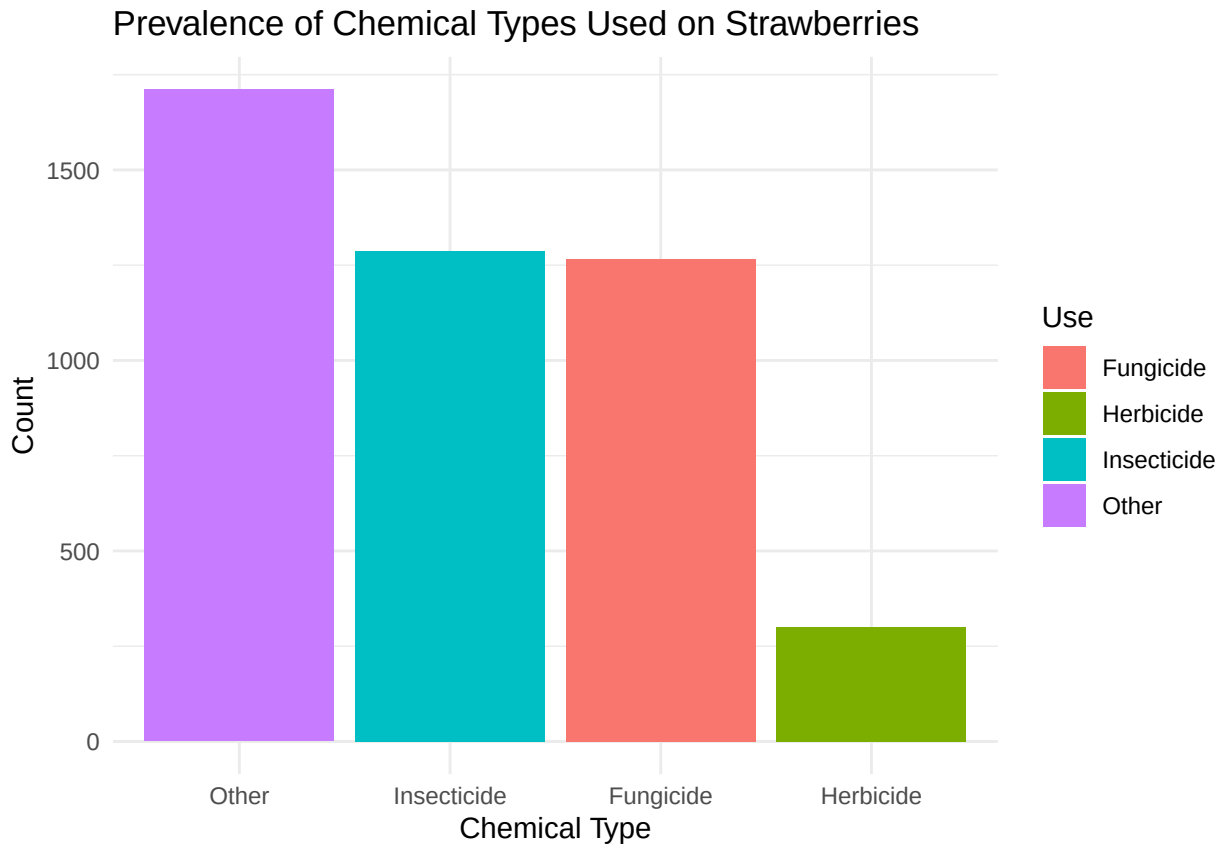
```
## Rows: 7584 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): Program, Period, Geo_Level, State, State_ANSI, Commodity, Data_Ite...
## dbl  (1): Year
## lgl  (2): Week_Ending, Region
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Extract chemical classifications (Use) from 'Domain' column
strawberry_clean <- strawberry_clean %>%
  mutate(
    Use = case_when(
      str_detect(Domain, "FUNGICIDE") ~ "Fungicide",
      str_detect(Domain, "INSECTICIDE") ~ "Insecticide",
      str_detect(Domain, "HERBICIDE") ~ "Herbicide",
      TRUE ~ "Other"
    ),
    # Extract specific chemical names from the 'Domain Category' column
    Chemical_Name = str_extract(`Domain_Category`, "\\((.*?)\\)"),
    Chemical_Name = str_replace_all(Chemical_Name, "[()]", "")  # Remove parentheses
  )

# Filter out rows where 'Use' or 'Chemical_Name' are NA
strawberry_clean <- strawberry_clean %>%
  filter(!is.na(Use) & !is.na(Chemical_Name))

# Count the prevalence of each chemical category
chemical_summary <- strawberry_clean %>%
  group_by(Use) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count))

# Print the summary of chemical types
print(chemical_summary)
```

```
## # A tibble: 4 x 2
##   Use         Count
##   <chr>       <int>
## 1 Other        1711
## 2 Insecticide  1286
## 3 Fungicide    1266
## 4 Herbicide     301
```

```r
# Visualization: Bar chart of chemical types
ggplot(chemical_summary, aes(x = reorder(Use, -Count), y = Count, fill = Use)) +
  geom_bar(stat = "identity") +
  labs(title = "Prevalence of Chemical Types Used on Strawberries",
```

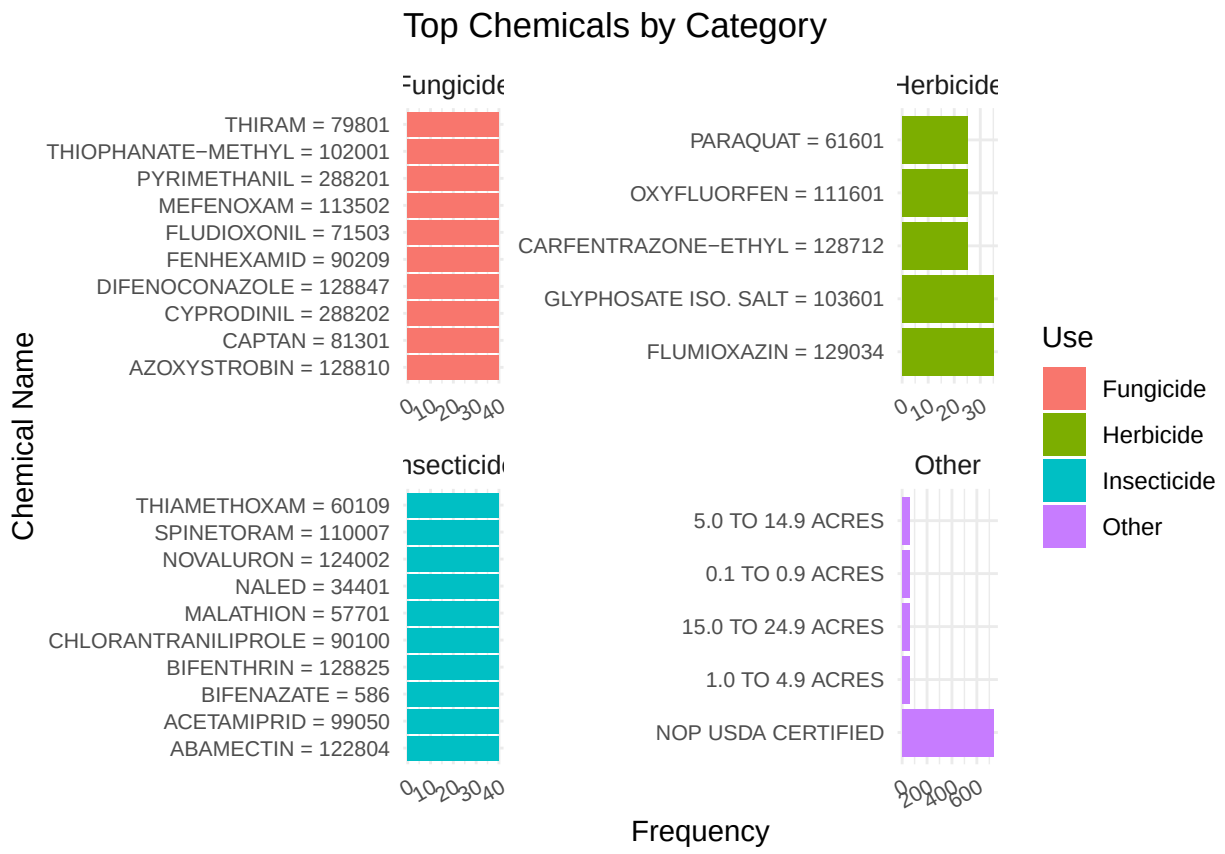
```
       x = "Chemical Type",
       y = "Count") +
  theme_minimal()
```




```
# Visualization: Top chemicals within each category
top_chemicals <- strawberry_clean %>%
  group_by(Use, Chemical_Name) %>%
  summarise(Frequency = n()) %>%
  arrange(desc(Frequency)) %>%
  slice_max(Frequency, n = 5)  # Top 5 chemicals per category
```

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

```
ggplot(top_chemicals, aes(x = reorder(Chemical_Name, -Frequency), y = Frequency, fill = Use)) +
  geom_bar(stat = "identity") +
  facet_wrap(~ Use, scales = "free", nrow = 2) +  # Arrange categories in 2 rows for better spacing
  labs(title = "Top Chemicals by Category",
       x = "Chemical Name",
       y = "Frequency") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 30, hjust = 1, size = 8),  # Adjust text angle and size
        axis.text.y = element_text(size = 8),
        strip.text = element_text(size = 10)) +  # Increase facet label size for clarity
  coord_flip()  # Flip coordinates for horizontal bars
```



As shown, not including other chemicals, Insecticides are the most commonly used chemical type, fungicide being similar bbut a bit smaller and herbicide being used the least.

https://www.cambridge.org/core/journals/weed-technology/article/weed-control-with-and-strawberry-tolerance-to-herbicides-applied-through-drip-irrigation/77FBD1F590F3401C449ACAD43FE1B1DD

This website gives me reasons why herbicides are used the least. Strawberries are sensitive to herbicides, leading to less use of herbicides. For example, oxyfluorfen should be very carefully applied, or else, this could eventually harm the crop.

I would like to look deeper into how other chemicals are preferred for growing strawberries.

##Total Acres Grown by state We will now look at the total Acre of production in Strawberries.

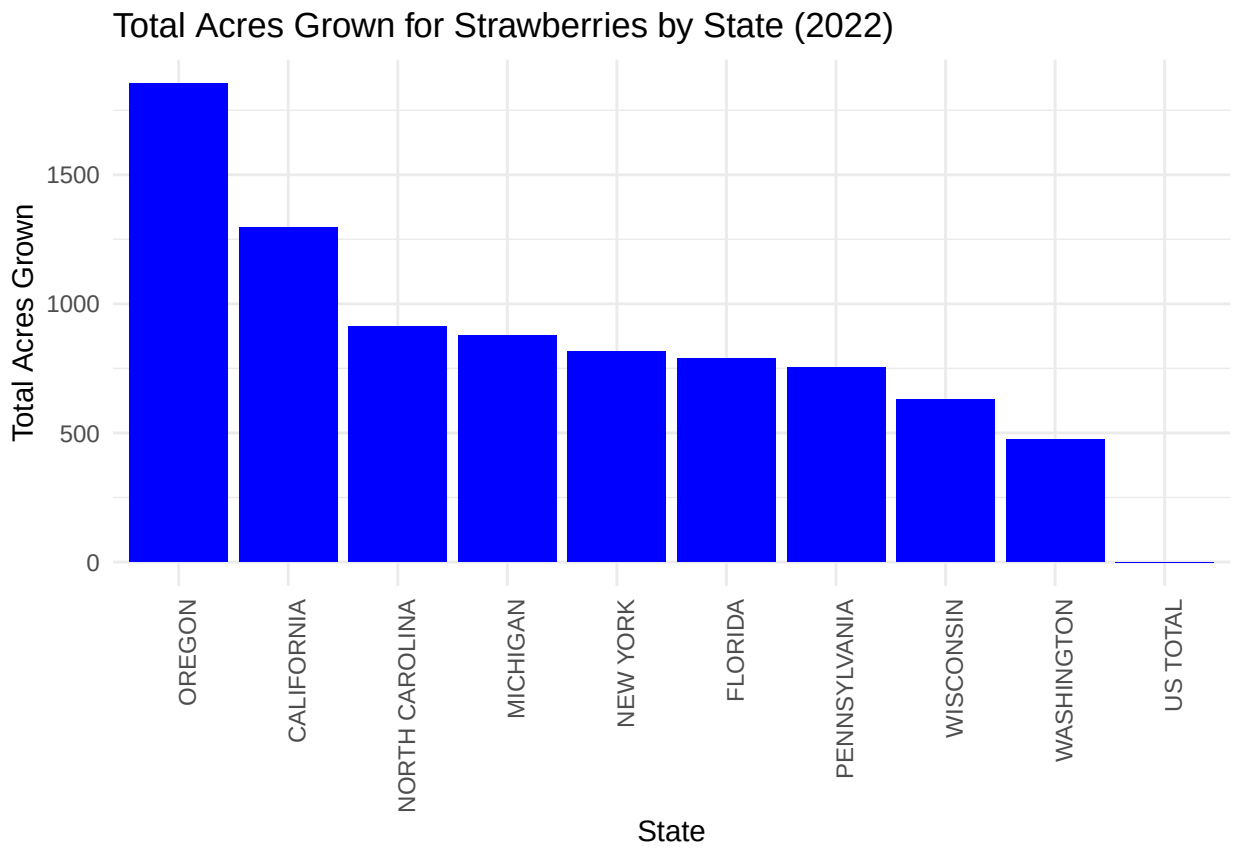
```
# Convert 'Value' column to numeric, replacing '(D)' or other placeholders with NA
strawberry_clean <- strawberry_clean %>%
  mutate(Value = ifelse(Value %in% c("(D)", "(NA)"), NA, as.numeric(Value)))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Value = ifelse(Value %in% c("(D)", "(NA)"), NA,
##   as.numeric(Value))`.
## Caused by warning in `ifelse()`:
## ! NAs introduced by coercion
```

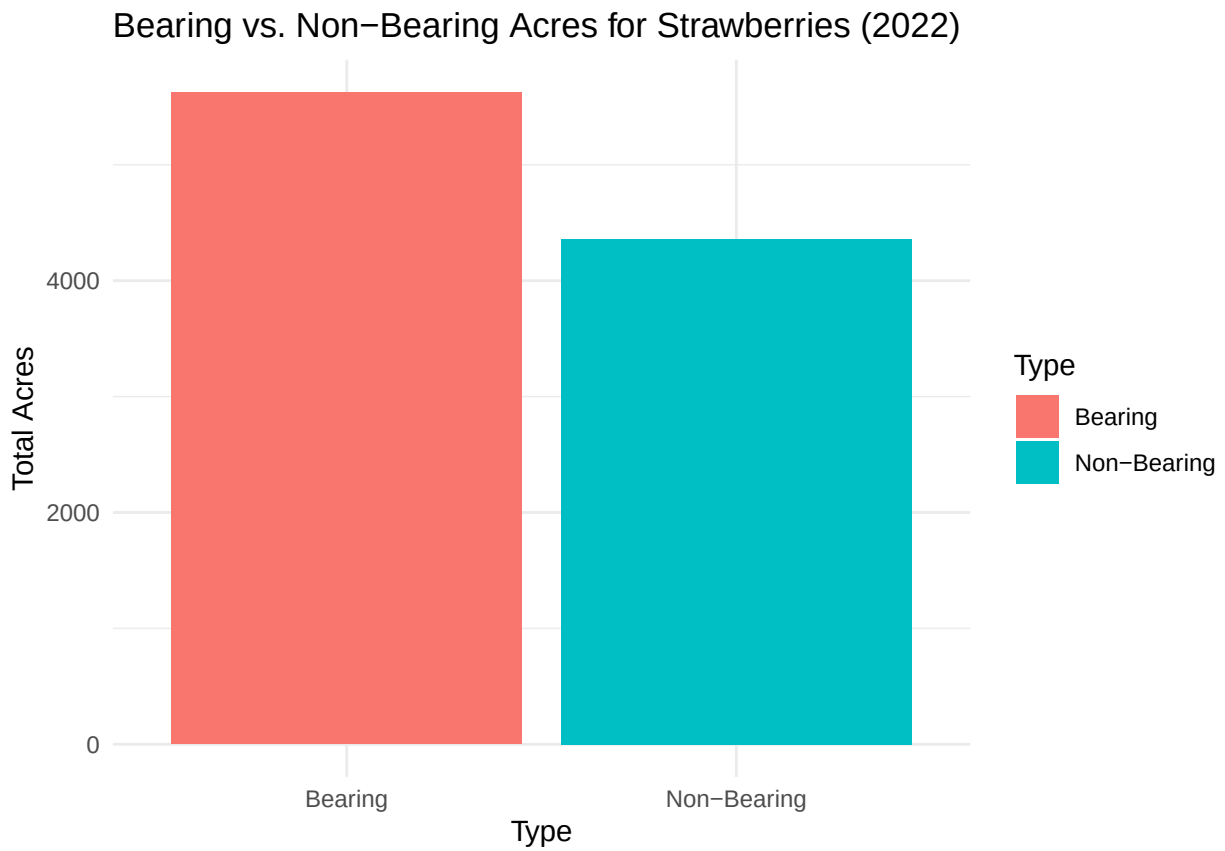
```
acres_data <- strawberry_clean %>%
  filter(str_detect(Data_Item, "ACRES GROWN"))

# Display the aggregated data to verify its content
print(acres_data)
```

```
## # A tibble: 61 x 15
##    Program  Year Period Week_Ending Geo_Level State  State_ANSI Region Commodity
##    <chr>   <dbl> <chr>  <lgl>       <chr>     <chr>  <chr>      <lgl>  <chr>
##  1 CENSUS   2022 YEAR   NA          NATIONAL  US TO~ Unknown    NA     STRAWBER~
##  2 CENSUS   2022 YEAR   NA          NATIONAL  US TO~ Unknown    NA     STRAWBER~
##  3 CENSUS   2022 YEAR   NA          NATIONAL  US TO~ Unknown    NA     STRAWBER~
##  4 CENSUS   2022 YEAR   NA          NATIONAL  US TO~ Unknown    NA     STRAWBER~
##  5 CENSUS   2022 YEAR   NA          NATIONAL  US TO~ Unknown    NA     STRAWBER~
##  6 CENSUS   2022 YEAR   NA          NATIONAL  US TO~ Unknown    NA     STRAWBER~
##  7 CENSUS   2022 YEAR   NA          NATIONAL  US TO~ Unknown    NA     STRAWBER~
##  8 CENSUS   2022 YEAR   NA          STATE     CALIF~ 06         NA     STRAWBER~
##  9 CENSUS   2022 YEAR   NA          STATE     CALIF~ 06         NA     STRAWBER~
## 10 CENSUS   2022 YEAR   NA          STATE     CALIF~ 06         NA     STRAWBER~
## # i 51 more rows
## # i 6 more variables: Data_Item <chr>, Domain <chr>, Domain_Category <chr>,
## #   Value <dbl>, Use <chr>, Chemical_Name <chr>
```

```r
# Filter data for acres grown in 2022 and group by state
acres_by_state <- strawberry_clean %>%
  filter(str_detect(Data_Item, "ACRES GROWN"), Year == 2022) %>%
  group_by(State) %>%
  summarise(Total_Acres = sum(Value, na.rm = TRUE))

# Plot total acres grown by state
ggplot(acres_by_state, aes(x = reorder(State, -Total_Acres), y = Total_Acres)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(title = "Total Acres Grown for Strawberries by State (2022)",
       x = "State",
       y = "Total Acres Grown") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```r
# Filter data for bearing and non-bearing acres in 2022
bearing_acres <- strawberry_clean %>%
  filter(str_detect(Data_Item, "ACRES BEARING"), Year == 2022) %>%
  summarise(Total_Bearing_Acres = sum(Value, na.rm = TRUE))

non_bearing_acres <- strawberry_clean %>%
  filter(str_detect(Data_Item, "ACRES NON-BEARING"), Year == 2022) %>%
  summarise(Total_Non_Bearing_Acres = sum(Value, na.rm = TRUE))

# Combine the two into a single data frame
acres_type <- data.frame(
  Type = c("Bearing", "Non-Bearing"),
  Acres = c(bearing_acres$Total_Bearing_Acres, non_bearing_acres$Total_Non_Bearing_Acres)
)

# Plot bearing vs. non-bearing acres
ggplot(acres_type, aes(x = Type, y = Acres, fill = Type)) +
  geom_bar(stat = "identity") +
  labs(title = "Bearing vs. Non-Bearing Acres for Strawberries (2022)",
       x = "Type",
       y = "Total Acres") +
  theme_minimal()
```



From the Acres Data, we see that Oregon is the state with the biggest Acres of land to grow strawberries. Nationally, there is a bigger proportion of bearing acres than that of non-bearing, showing a good sign of eco-friendly farming, saving the soil.

We will now look at how it differs by state.

##Analysis on Strawberries grown.

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

# Convert 'Value' column to numeric, replacing '(D)' or other placeholders with NA
strawberry_clean <- strawberry_clean %>%
  mutate(Value = ifelse(Value %in% c("(D)", "(NA)"), NA, as.numeric(Value)))

# Check the structure of the cleaned dataset
str(strawberry_clean)
```

```
## tibble [4,564 x 15] (S3: tbl_df/tbl/data.frame)
##  $ Program        : chr [1:4564] "CENSUS" "CENSUS" "CENSUS" "CENSUS" ...
##  $ Year           : num [1:4564] 2022 2022 2022 2022 2022 ...
##  $ Period         : chr [1:4564] "YEAR" "YEAR" "YEAR" "YEAR" ...
##  $ Week_Ending    : logi [1:4564] NA NA NA NA NA NA ...
##  $ Geo_Level      : chr [1:4564] "NATIONAL" "NATIONAL" "NATIONAL" "NATIONAL" ...
##  $ State          : chr [1:4564] "US TOTAL" "US TOTAL" "US TOTAL" "US TOTAL" ...
```

```
##  $ State_ANSI     : chr [1:4564] "Unknown" "Unknown" "Unknown" "Unknown" ...
##  $ Region         : logi [1:4564] NA NA NA NA NA NA ...
##  $ Commodity      : chr [1:4564] "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" ...
##  $ Data_Item      : chr [1:4564] "STRAWBERRIES - ACRES BEARING" "STRAWBERRIES - ACRES BEARING" "STRAW
##  $ Domain         : chr [1:4564] "AREA GROWN" "AREA GROWN" "AREA GROWN" "AREA GROWN" ...
##  $ Domain_Category: chr [1:4564] "AREA GROWN: (0.1 TO 0.9 ACRES)" "AREA GROWN: (1.0 TO 4.9 ACRES)" "A
##  $ Value          : num [1:4564] 963 NA NA NA NA NA NA NA NA NA ...
##  $ Use            : chr [1:4564] "Other" "Other" "Other" "Other" ...
##  $ Chemical_Name  : chr [1:4564] "0.1 TO 0.9 ACRES" "1.0 TO 4.9 ACRES" "100 OR MORE ACRES" "15.0 TO 1
```

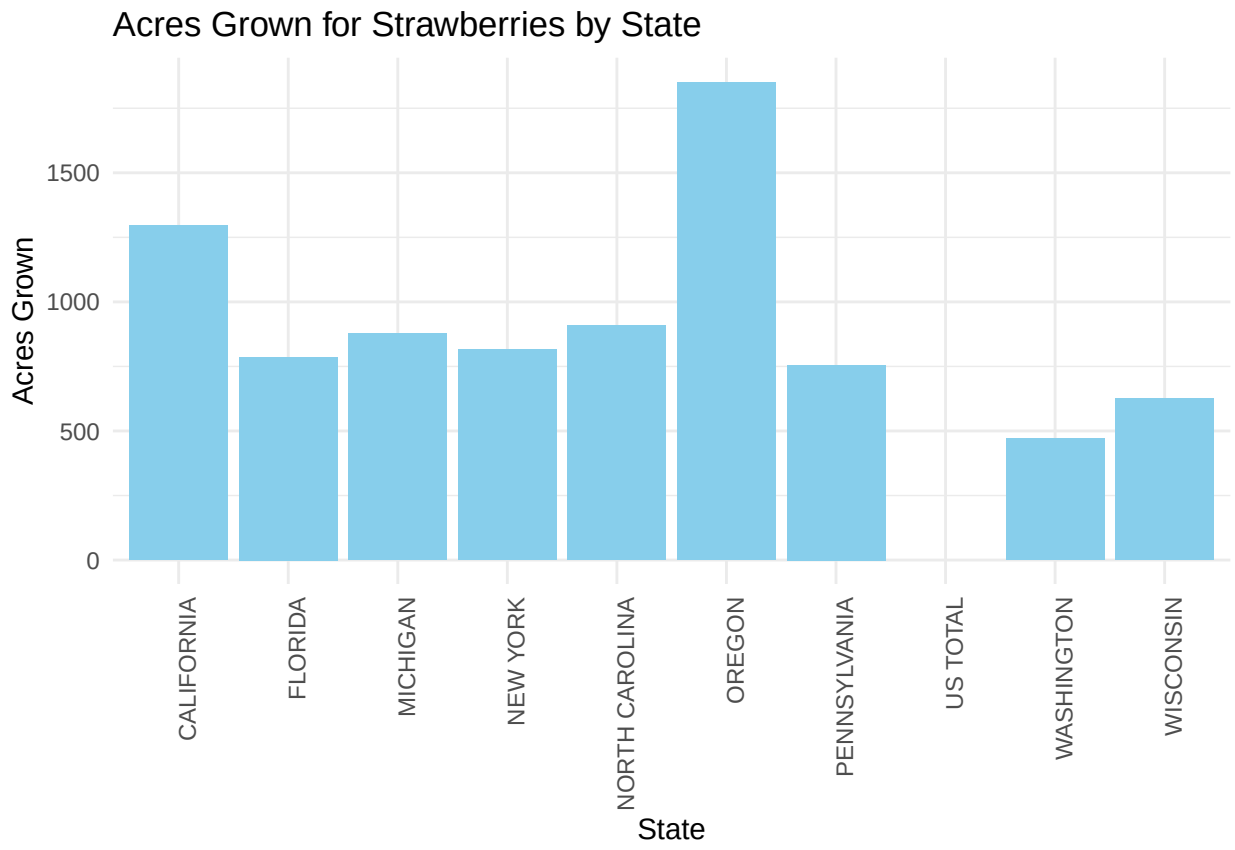
```r
summary(strawberry_clean)
```

```
##    Program               Year         Period          Week_Ending
##  Length:4564        Min.   :2018   Length:4564        Mode:logical
##  Class :character   1st Qu.:2019   Class :character   NA's:4564
##  Mode  :character   Median :2021   Mode  :character
##                     Mean   :2020
##                     3rd Qu.:2022
##                     Max.   :2023
##
##   Geo_Level            State            State_ANSI         Region
##  Length:4564        Length:4564        Length:4564        Mode:logical
##  Class :character   Class :character   Class :character   NA's:4564
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Commodity          Data_Item            Domain          Domain_Category
##  Length:4564        Length:4564        Length:4564        Length:4564
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      Value             Use            Chemical_Name
##  Min.   :  0.017   Length:4564        Length:4564
##  1st Qu.:  1.000   Class :character   Class :character
##  Median :  6.000   Mode  :character   Mode  :character
##  Mean   : 57.194
##  3rd Qu.: 41.000
##  Max.   :963.000
##  NA's   :2601
```

```r
### Visualization 1: Total Acres Grown for Strawberries by State
acres_data <- strawberry_clean %>%
  filter(str_detect(Data_Item, "ACRES GROWN"))

ggplot(acres_data, aes(x = State, y = Value)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Acres Grown for Strawberries by State",
       x = "State",
```

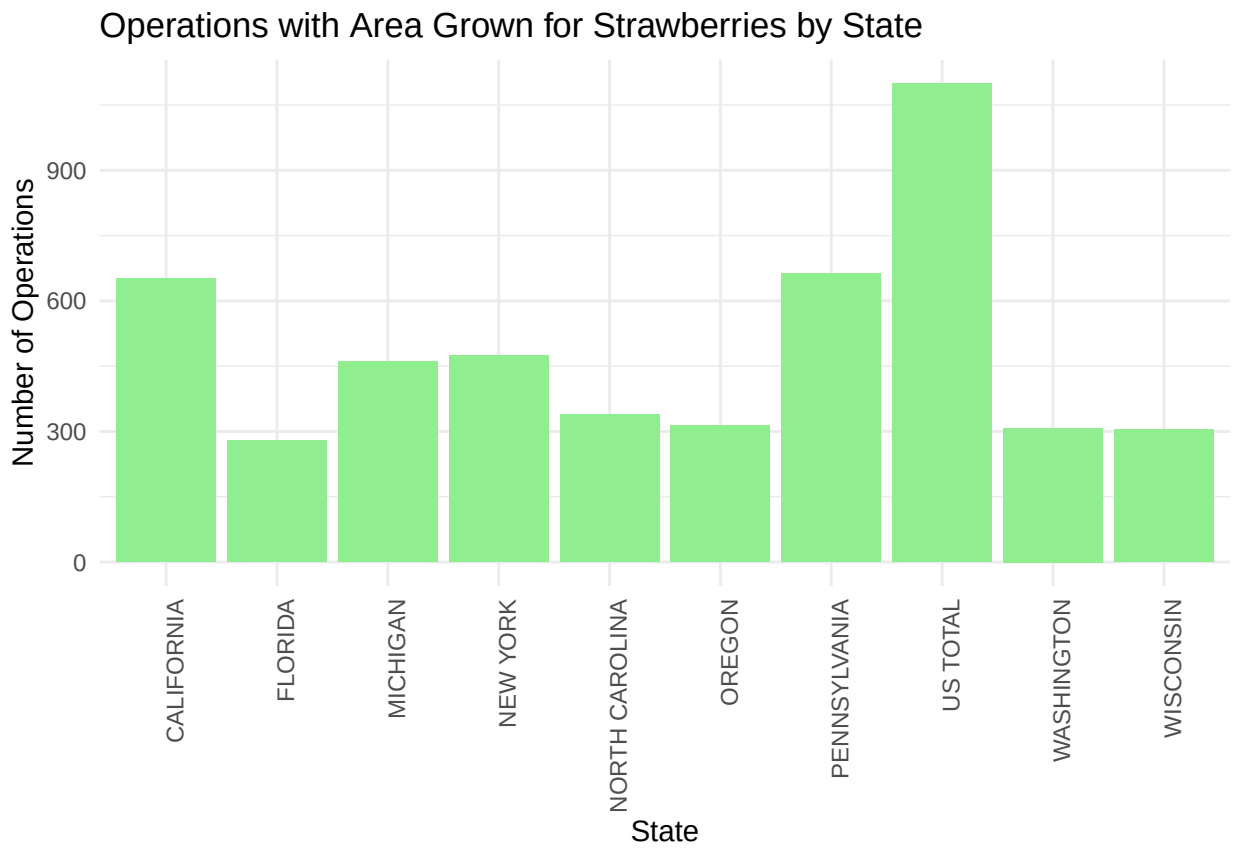
```
       y = "Acres Grown") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Warning: Removed 23 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



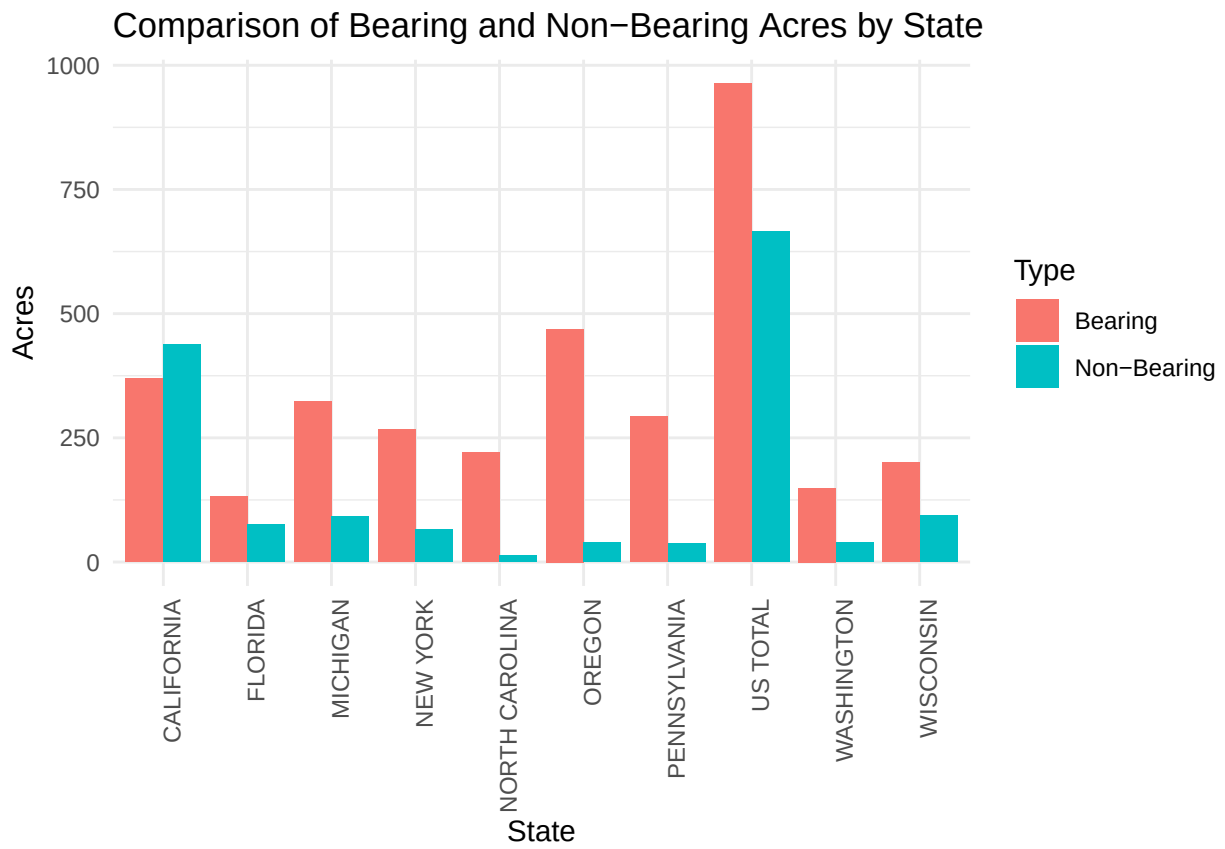

```
### Visualization 2: Operations with Area Grown by State
operations_data <- strawberry_clean %>%
  filter(str_detect(Data_Item, "OPERATIONS WITH AREA GROWN"))

ggplot(operations_data, aes(x = State, y = Value)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Operations with Area Grown for Strawberries by State",
       x = "State",
       y = "Number of Operations") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```




```
### Visualization 3: Comparison of Bearing vs. Non-Bearing Acres by State
bearing_data <- strawberry_clean %>%
  filter(str_detect(Data_Item, "ACRES BEARING"))

non_bearing_data <- strawberry_clean %>%
  filter(str_detect(Data_Item, "ACRES NON-BEARING"))

combined_acres <- rbind(
  bearing_data %>% mutate(Type = "Bearing"),
  non_bearing_data %>% mutate(Type = "Non-Bearing")
)

ggplot(combined_acres, aes(x = State, y = Value, fill = Type)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Comparison of Bearing and Non-Bearing Acres by State",
       x = "State",
       y = "Acres",
       fill = "Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

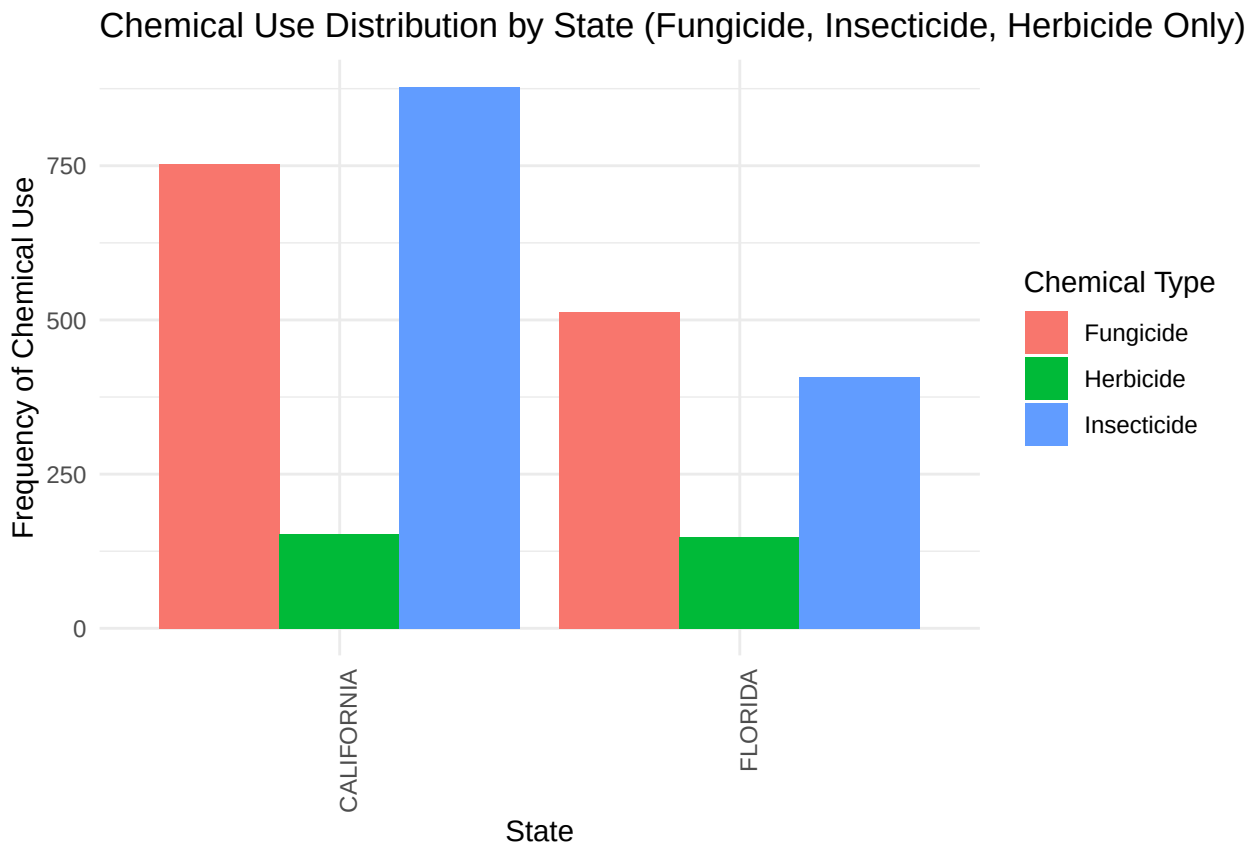
```
## Warning: Removed 51 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```



California: .Across all graphs, California stands out as the leading state in terms of strawberry acreage and operations. This could lead to an understanding of importance in the US strawberry market. Showing a higher number of non-bearing acres suggesting that the state is investing in future production and crop rotation practices Oregon and North Carolina: Also showing significance in portion of non-bearing acres, indicating similar practices to maintain soil health and prepare for future production cycles. US Total: showing a balanced comparison between bearing and non-bearing acres. It represents the nationwide trend of substantial portion of land is kept in non-bearing status to sustain long-term productivity.

We could look deeper into how strawberries farming could actually be a eco-friendly farming in the future.

```
strawberry_clean_filtered <- strawberry_clean %>%
  filter(Use %in% c("Fungicide", "Insecticide", "Herbicide"))

# Group by State and Chemical Use
chemicals_by_state <- strawberry_clean_filtered %>%
  group_by(State, Use) %>%
  summarise(Frequency = n(), .groups = 'drop')

# Plot the distribution of chemical types by state excluding "Other"
ggplot(chemicals_by_state, aes(x = reorder(State, -Frequency), y = Frequency, fill = Use)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Chemical Use Distribution by State (Fungicide, Insecticide, Herbicide Only)",
       x = "State",
       y = "Frequency of Chemical Use",
       fill = "Chemical Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



As you can see, California uses inseciticdes the most, and fungicides as shown, while herbicide is low From here, I am questioning why this is the case, with California being the state with the most operation going on .

The high use of insecticides in California's strawberry fields is due to the state's specific pest challenges. One of the major pests is the lygus bug (Lygus hesperus), which causes significant damage to strawberry crops. The lygus bug is particularly difficult to control due to its mobility and its tendency to migrate into strawberry fields from nearby vegetation. As a result, farmers often resort to using insecticides like malathion, acetamiprid, and novaluron to manage these pests effectively

https://croplifefoundation.wordpress.com/wp-content/uploads/2012/07/combined_document_strawberries.pdf

To further evaluate the EDA, I am planning on examining each chemicals as well as their impacts on the fruits and the yield of product in the future. Also, I will be looking at how the organic raised strawberries differ in chemicals.