Assignment F

```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   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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 becom
```

```r
df<- read_csv("strawberry raw.csv")
```

```
Rows: 12969 Columns: 21
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (11): Program, Period, Geo Level, State, State ANSI, watershed_code, Com...
dbl  (1): Year
lgl  (9): Week Ending, Ag District, Ag District Code, County, County ANSI, Z...

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.
```

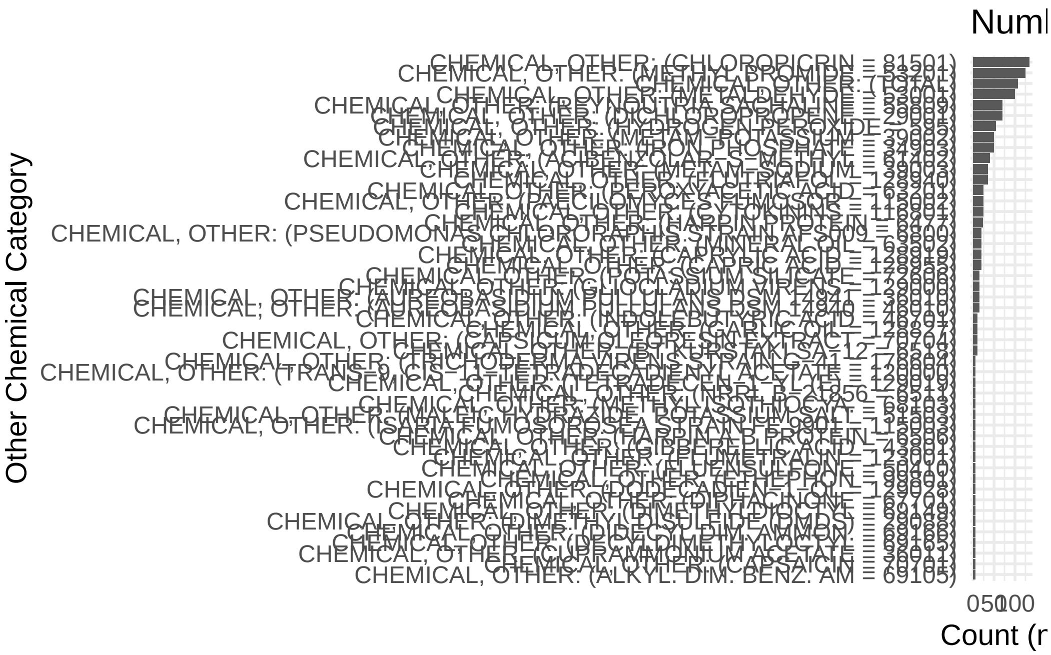
```r
str(df)
```

```
spc_tbl_ [12,969 x 21] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Program          : chr [1:12969] "SURVEY" "SURVEY" "SURVEY" "SURVEY" ...
 $ Year             : num [1:12969] 2023 2023 2023 2023 2023 ...
$ Period          : chr [1:12969] "YEAR" "YEAR" "YEAR" "YEAR" ...
$ Week Ending     : logi [1:12969] NA NA NA NA NA NA ...
$ Geo Level       : chr [1:12969] "STATE" "STATE" "STATE" "STATE" ...
$ State           : chr [1:12969] "CALIFORNIA" "CALIFORNIA" "CALIFORNIA" "CALIFORNIA" ...
$ State ANSI      : chr [1:12969] "06" "06" "06" "06" ...
$ Ag District     : logi [1:12969] NA NA NA NA NA NA ...
$ Ag District Code: logi [1:12969] NA NA NA NA NA NA ...
$ County          : logi [1:12969] NA NA NA NA NA NA ...
$ County ANSI     : logi [1:12969] NA NA NA NA NA NA ...
$ Zip Code        : logi [1:12969] NA NA NA NA NA NA ...
$ Region          : logi [1:12969] NA NA NA NA NA NA ...
$ watershed_code  : chr [1:12969] "00000000" "00000000" "00000000" "00000000" ...
$ Watershed       : logi [1:12969] NA NA NA NA NA NA ...
$ Commodity       : chr [1:12969] "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES
$ Data Item       : chr [1:12969] "STRAWBERRIES - APPLICATIONS, MEASURED IN LB" "STRAWBERRI
$ Domain          : chr [1:12969] "CHEMICAL, FUNGICIDE" "CHEMICAL, INSECTICIDE" "CHEMICAL, 
$ Domain Category : chr [1:12969] "CHEMICAL, FUNGICIDE: (OXATHIAPIPROLIN = 128111)" "CHEMICA
$ Value           : chr [1:12969] "(D)" "(D)" "(D)" "(NA)" ...
$ CV (%)          : logi [1:12969] NA NA NA NA NA NA ...
- 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_logical(),
 ..   `Ag District Code` = col_logical(),
 ..   County = col_logical(),
 ..   `County ANSI` = col_logical(),
 ..   `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_logical()
 .. )
 - attr(*, "problems")=<externalptr>
```

```r
df<- df %>%
  select(where(~!all(is.na(.)))) |> #delete columns with all NA
  select(where(~ n_distinct(.,na.rm = TRUE)>1) ) |>#delete columns with single value, which
  mutate(Year = as.integer(Year),
         Value = str_replace_all(Value, ",", ""),
         Value = as.numeric(Value),
         State = as.factor(State),
         Domain = as.factor(Domain),
         `Domain Category` = as.factor(`Domain Category`))
```

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

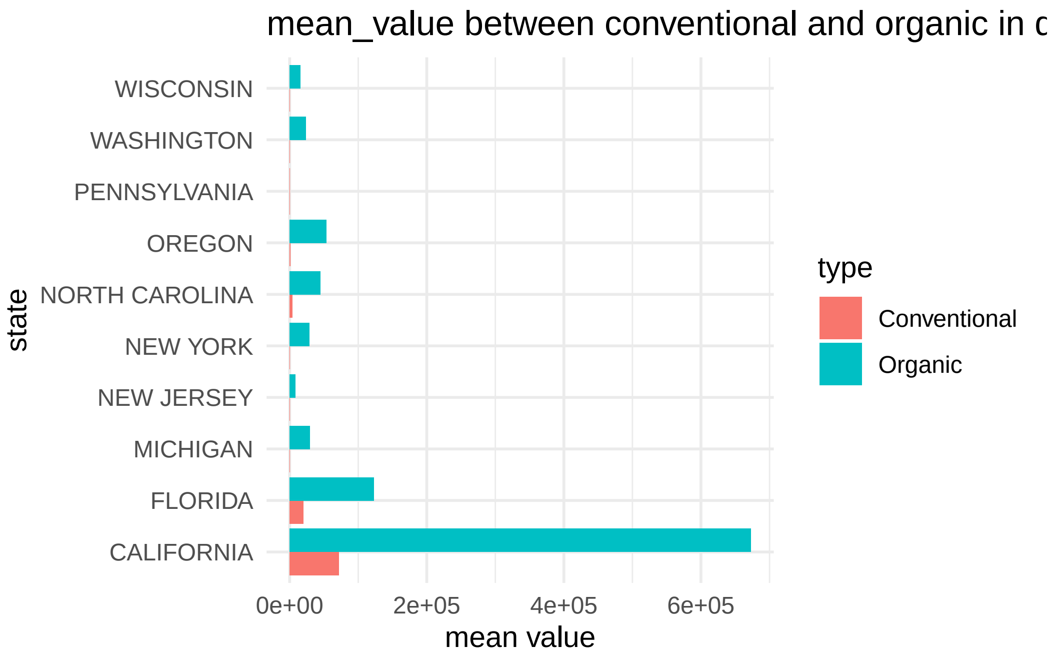
```r
other<- df|> #find those use other chemicals
  filter(Domain == "CHEMICAL, OTHER")
other_che<- other|> #inside the other chemical, group by different other chemicals
  group_by(`Domain Category`)|>
  summarise(mean_value = mean(Value, na.rm = TRUE),
            sd_value = sd(Value, na.rm = TRUE),
            n=n())
view(other_che)
ggplot(other_che,
       aes(x = reorder(`Domain Category`, n),
           y = n)) +
  geom_col() +
  coord_flip() +
  labs(
    title = "Number of Records per Other Chemical Category",
    x = "Other Chemical Category",
    y = "Count (n)") +
  theme_minimal()
```



```r
df<- df|> #detect whether it's conventional or organic
  mutate(type = case_when(
    str_detect(Domain,"CHEMICAL") ~ "Conventional",
    !str_detect(Domain,"CHEMICAL") ~ "Organic"
  )
  )
organic_vs_convention<- df|>
  group_by(State, type)|>
  summarise(
    mean_value = mean(Value, na.rm = TRUE),
    sd_value = sd(Value, na.rm = TRUE),
    n = n()
  ) |>
  arrange(mean_value)
```

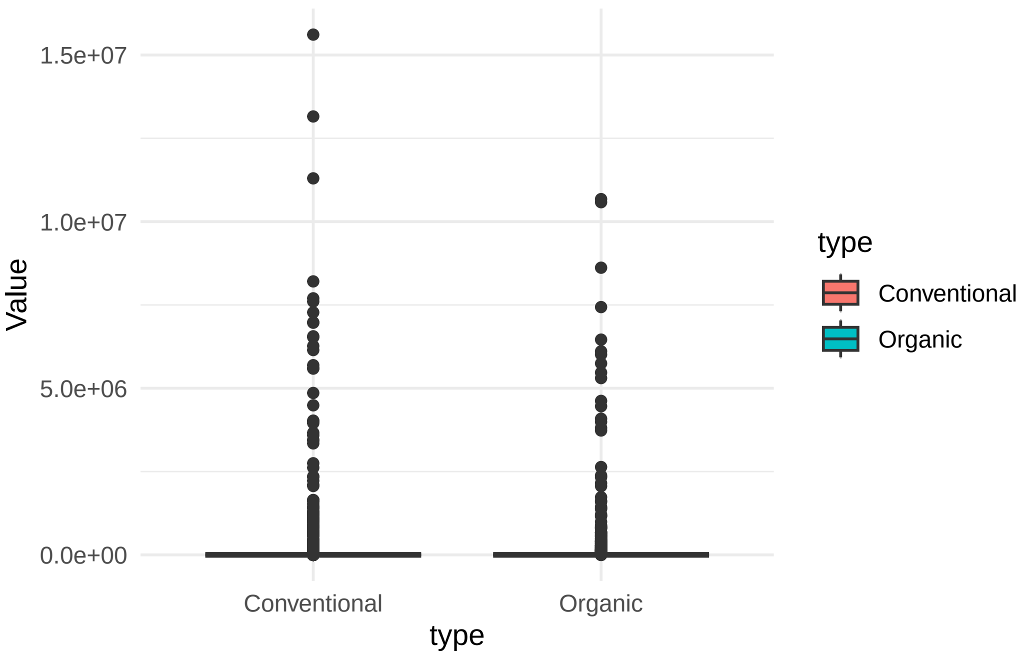
```
`summarise()` has grouped output by 'State'. You can override using the
`.groups` argument.
```

```r
ggplot(organic_vs_convention,aes(State,mean_value,fill = type))+
  geom_col(position = "dodge")+
  coord_flip()+
  labs(title = "mean_value between conventional and organic in differenct states",
       x = "state",
       y= "mean value")+
  theme_minimal()
```



```
ggplot(df, aes(type, Value,fill = type))+
  geom_boxplot() +
  theme_minimal()
```

```
Warning: Removed 4659 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



##Raw data acquisition #The raw strawberry dataset used in this analysis was obtained from the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS), which provides publicly available agricultural data. #Specifically, the data were downloaded from the NASS Quick Stats database (https://quickstats.nass.usda.gov/) Data retrieved: October 2025 Citation:U.S. Department of Agriculture, National Agricultural Statistics Service (NASS). Quick Stats Database. https://quickstats.nass.usda.gov/