Strawberry Data EDA

Beiming Yu

10/28/2022

1. Reading the census data and survery dara last week, and we will espeically focus on the data in California state

```r
datacen <- read.csv("straw_cen_cleaned2.csv")
datasur <- read.csv("straw_sur_cleaned2.csv")
str(datasur)
```

```
## 'data.frame':    1432 obs. of  15 variables:
##  $ Program          : chr  "SURVEY" "SURVEY" "SURVEY" "SURVEY" ...
##  $ Year             : int  2024 2024 2023 2023 2023 2023 2023 2023 2023 2023 ...
##  $ Period           : chr  "YEAR" "YEAR" "MARKETING YEAR" "MARKETING YEAR" ...
##  $ Geo.Level        : chr  "NATIONAL" "NATIONAL" "NATIONAL" "NATIONAL" ...
##  $ State            : chr  "US TOTAL" "US TOTAL" "US TOTAL" "US TOTAL" ...
##  $ State.ANSI       : int  -1 -1 -1 -1 -1 6 12 -1 -1 -1 ...
##  $ Market_Type      : chr  "FRESH MARKET" "PROCESSING" "OTHER" "FRESH MARKET" ...
##  $ Measure_Operation: chr  "PRICE RECEIVED, ADJUSTED BASE" "PRICE RECEIVED, ADJUSTED BASE" "PRICE REC
##  $ Unit_of_Measure  : chr  "$ / CWT" "$ / TON" "$ / CWT" "$ / CWT" ...
##  $ Domain           : chr  "TOTAL" "TOTAL" "TOTAL" "TOTAL" ...
##  $ Chemical_Use     : chr  "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" ...
##  $ Chemical_Name    : chr  "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" ...
##  $ Chemical_Code    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Value            : num  10.9 4.04 123 142 43.8 121 147 142 43.8 485 ...
##  $ CV....           : logi  NA NA NA NA NA NA ...
```

And we will find that based on the separation only, the survey data will contain the chemicals.

2. We want to focus ont the chemical used in California, and we will remove the variables with Not specified and Total. And we want to lahtest data which means we will not use data older than 5 years.

```r
unique1<- unique(datasur$Chemical_Name)
unique2<- unique(datasur$Chemical_Code)
ca_chemical <- subset(datasur, State != "California")
ca_chemical1 <- subset(ca_chemical, !(Chemical_Name %in% c("NOT SPECIFIED", "TOTAL")))
head(ca_chemical1)
```

```
##    Program Year Period Geo.Level      State State.ANSI Market_Type
## 19  SURVEY 2023   YEAR     STATE CALIFORNIA          6     BEARING
## 20  SURVEY 2023   YEAR     STATE CALIFORNIA          6     BEARING
## 21  SURVEY 2023   YEAR     STATE CALIFORNIA          6     BEARING
## 22  SURVEY 2023   YEAR     STATE CALIFORNIA          6     BEARING
## 23  SURVEY 2023   YEAR     STATE CALIFORNIA          6     BEARING
## 24  SURVEY 2023   YEAR     STATE CALIFORNIA          6     BEARING
##    Measure_Operation              Unit_of_Measure                Domain
## 19      APPLICATIONS                           LB CHEMICAL, INSECTICIDE
## 20      APPLICATIONS LB / ACRE / APPLICATION, AVG   CHEMICAL, FUNGICIDE
## 21      APPLICATIONS LB / ACRE / APPLICATION, AVG   CHEMICAL, FUNGICIDE
## 22      APPLICATIONS LB / ACRE / APPLICATION, AVG   CHEMICAL, FUNGICIDE
## 23      APPLICATIONS LB / ACRE / APPLICATION, AVG   CHEMICAL, FUNGICIDE
## 24      APPLICATIONS LB / ACRE / APPLICATION, AVG   CHEMICAL, FUNGICIDE
##    Chemical_Use     Chemical_Name Chemical_Code   Value CV....
## 19  INSECTICIDE         ABAMECTIN        122804 300.000     NA
## 20    FUNGICIDE      AZOXYSTROBIN        128810   0.234     NA
## 21    FUNGICIDE BORAX DECAHYDRATE         11102   0.042     NA
## 22    FUNGICIDE          BOSCALID        128008   0.354     NA
## 23    FUNGICIDE            CAPTAN         81301   1.693     NA
## 24    FUNGICIDE        CYPRODINIL        288202   0.316     NA
```

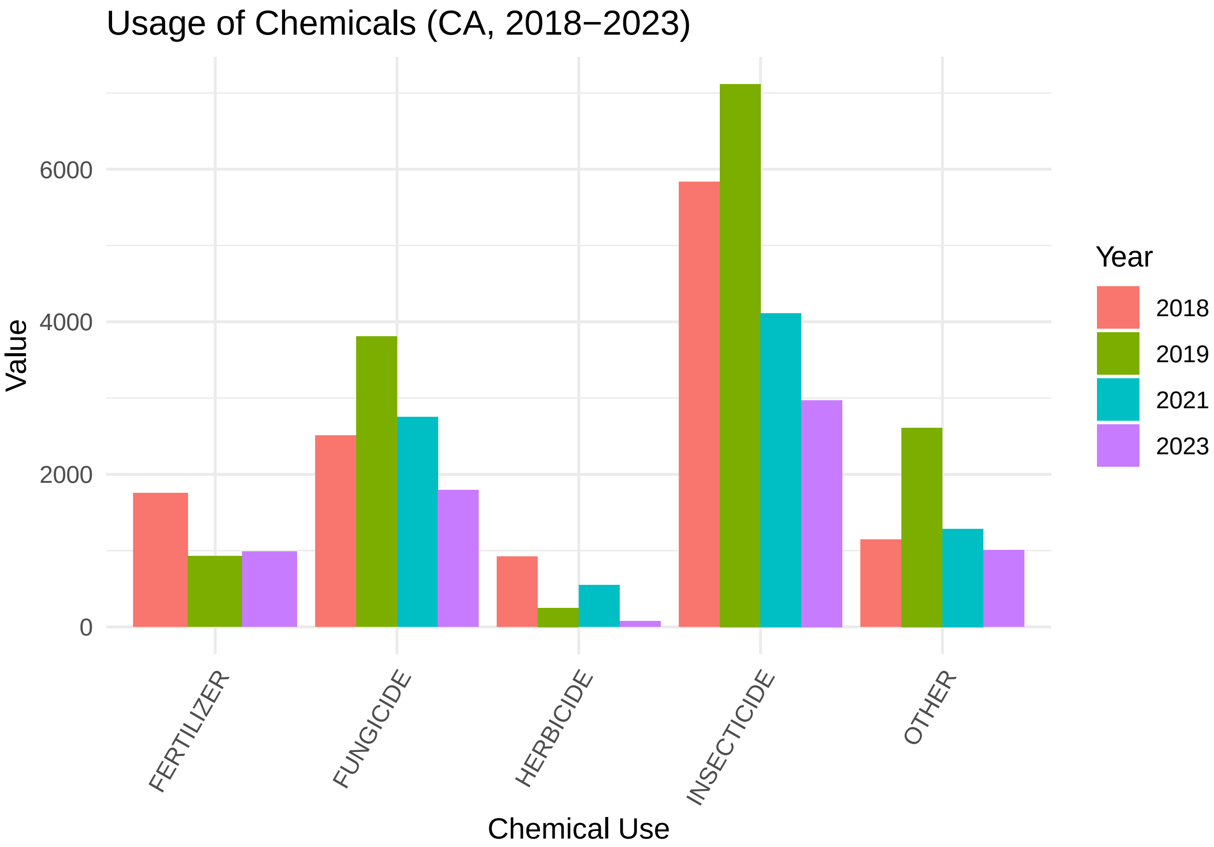
```
ca_chemical2 <- ca_chemical1[ca_chemical1$Year %in% 2018:2023, ]
```

3. Then, based on the reading, we know there are four uses for the chemical including insecticide, fungicide, herbicide and others. We can make a graph to show the total amount of usage in each category for every year. The change of the amount of usage might reflect the change of climate or species in that year in california. To get that data, we will add the value of each chemical based on the Chemical use, and separate them in 5 years.

```
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
```

```
ca_chemical3 <- aggregate(Value ~ Chemical_Use + Year, data = ca_chemical2, FUN = sum, na.rm = TRUE)
ggplot(ca_chemical3, aes(x = Chemical_Use, y = Value, fill = as.factor(Year))) +
    geom_bar(stat = "identity", position = "dodge") +
    labs(title = "Usage of Chemicals (CA, 2018-2023)",
         x = "Chemical Use",
         y = "Value",
         fill = "Year") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



```
#Based on the bar plot we have, we can see some unusual years like 2019, which has a significant increa
```

4. For each category, we cant to identify which is the most popular chemical to use and the trend of usage among each year.

```
ca_chemical_agg <- aggregate(Value ~ Chemical_Name + Year, data = ca_chemical2, FUN = sum)

tenchemicals <- function(year) {
  subset(ca_chemical_agg, Year == year) %>%
    arrange(desc(Value)) %>%
    head(10)
}

top_10_2023 <- tenchemicals(2023)
top_10_2021 <- tenchemicals(2021)

print(top_10_2023)
```

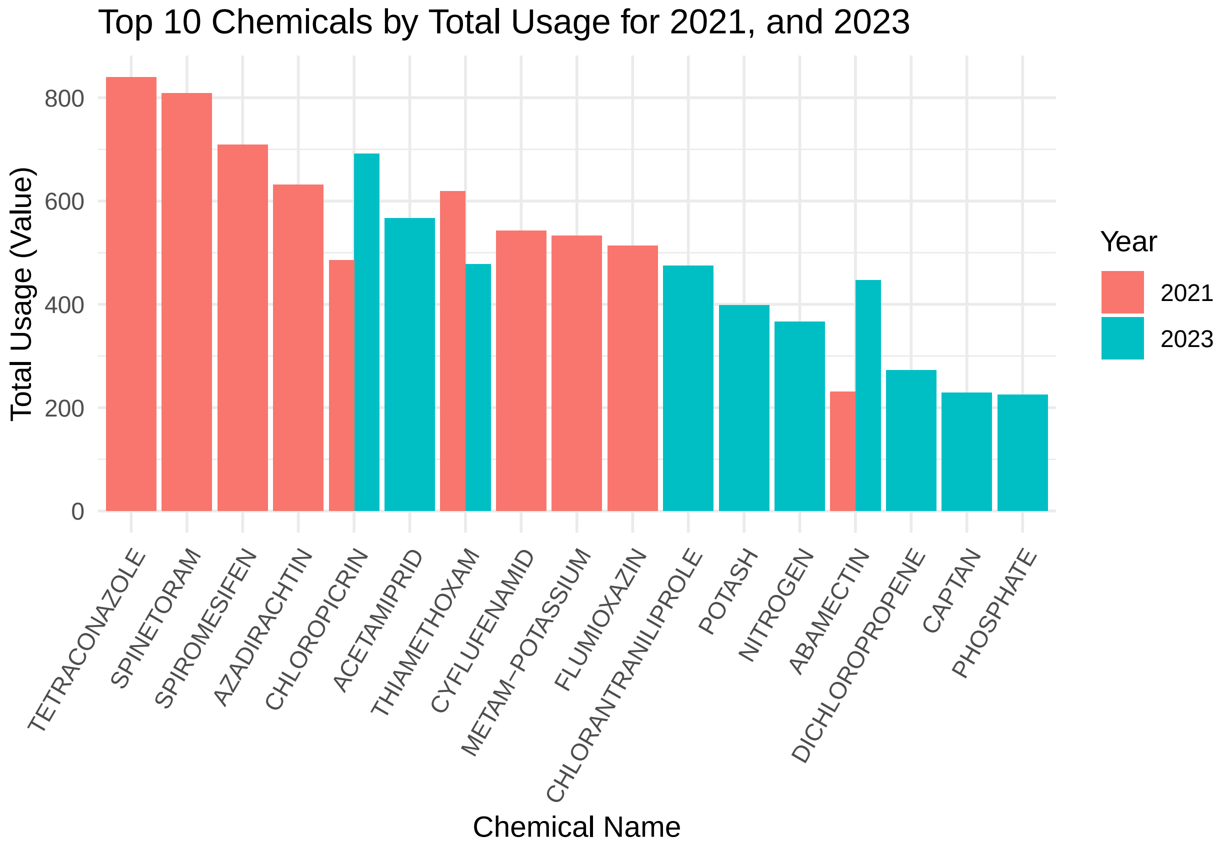
```
##          Chemical_Name Year   Value
## 1          CHLOROPICRIN 2023 691.997
## 2           ACETAMIPRID 2023 566.233
## 3          THIAMETHOXAM 2023 477.876
## 4   CHLORANTRANILIPROLE 2023 474.935
## 5             ABAMECTIN 2023 446.508
## 6                POTASH 2023 398.100
## 7              NITROGEN 2023 366.200
## 8       DICHLOROPROPENE 2023 272.682
## 9                CAPTAN 2023 228.746
## 10          PHOSPHATE 2023 225.200
```

```
print(top_10_2021)
```

```
##      Chemical_Name Year   Value
## 1    TETRACONAZOLE 2021 839.594
## 2       SPINETORAM 2021 808.686
## 3      SPIROMESIFEN 2021 708.626
## 4      AZADIRACHTIN 2021 631.677
## 5      THIAMETHOXAM 2021 618.553
## 6      CYFLUFENAMID 2021 542.354
## 7  METAM-POTASSIUM 2021 533.018
## 8       FLUMIOXAZIN 2021 513.383
## 9      CHLOROPICRIN 2021 485.114
## 10        ABAMECTIN 2021 230.597
```

```
top_10_all <- rbind(top_10_2023, top_10_2021)

ggplot(top_10_all, aes(x = reorder(Chemical_Name, -Value), y = Value, fill = as.factor(Year))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Top 10 Chemicals by Total Usage for 2021, and 2023",
       x = "Chemical Name",
       y = "Total Usage (Value)",
       fill = "Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



```
#Based on the plot, we can find a interesting fact that the popular chemicals are keeping changing in d
```

5. Then I came up with a further question, is that causing by the different usage of the chemicals. As we find the in step 3, farmers used more pesticides in 2019, which means the chemicals used in pesticides will be used more. Therefore, we will separate them in 4 categroies.

```
ca_chemical2_filtered <- subset(ca_chemical2, Chemical_Use == "INSECTICIDE")
ca_chemical_agg1 <- aggregate(Value ~ Chemical_Name + Year, data = ca_chemical2_filtered, FUN = sum)

tenchemicals1 <- function(year) {
  subset(ca_chemical_agg1, Year == year) %>%
    arrange(desc(Value)) %>%
    head(10)
}

top_10_2023_new <- tenchemicals1(2023)
top_10_2021_new <- tenchemicals1(2021)

print(top_10_2023_new)
```

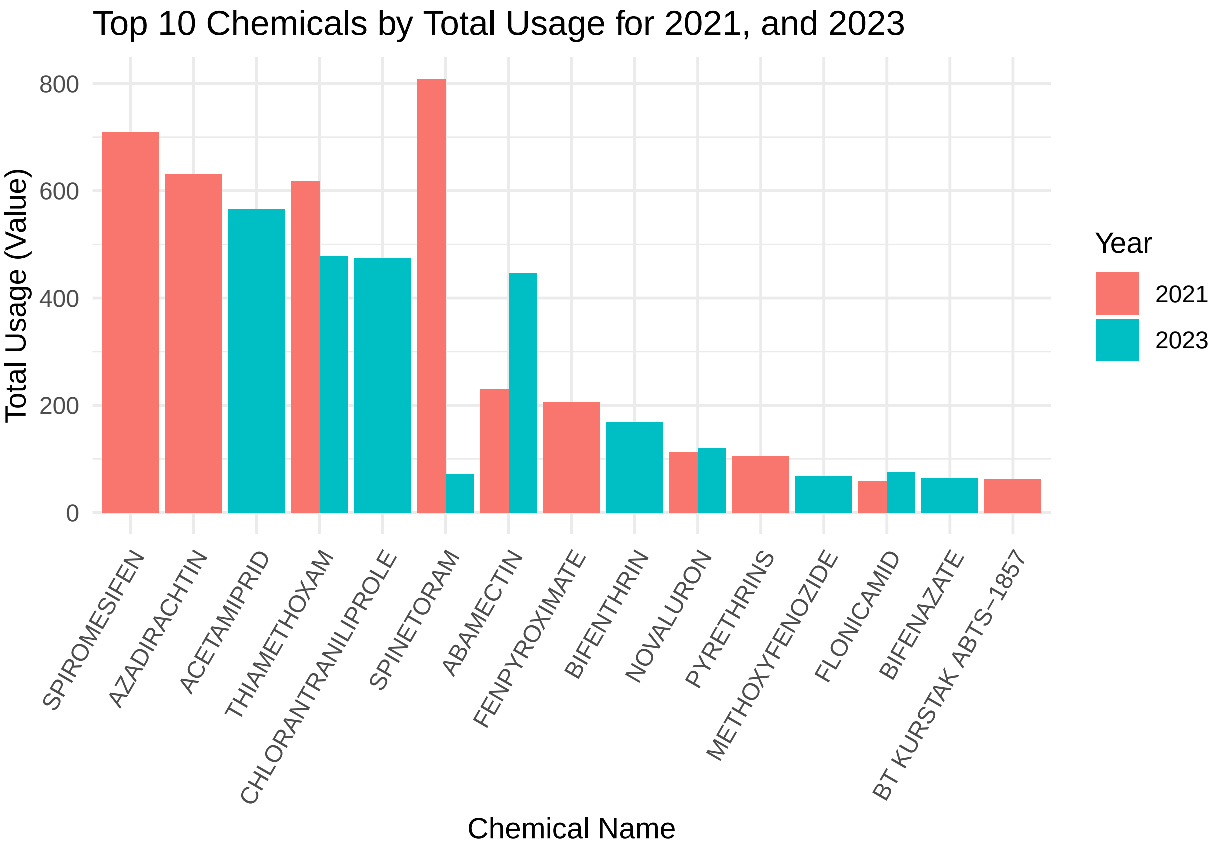
```
##           Chemical_Name Year   Value
## 1           ACETAMIPRID 2023 566.233
## 2          THIAMETHOXAM 2023 477.876
## 3  CHLORANTRANILIPROLE 2023 474.935
## 4             ABAMECTIN 2023 446.508
## 5            BIFENTHRIN 2023 169.247
## 6             NOVALURON 2023 120.866
## 7            FLONICAMID 2023  75.584
## 8            SPINETORAM 2023  71.996
## 9       METHOXYFENOZIDE 2023  68.045
## 10           BIFENAZATE 2023  64.794
```

```
print(top_10_2021_new)
```

```
##           Chemical_Name Year   Value
## 1            SPINETORAM 2021 808.686
## 2           SPIROMESIFEN 2021 708.626
## 3           AZADIRACHTIN 2021 631.677
## 4           THIAMETHOXAM 2021 618.553
## 5              ABAMECTIN 2021 230.597
## 6          FENPYROXIMATE 2021 205.327
## 7              NOVALURON 2021 112.304
## 8             PYRETHRINS 2021 104.708
## 9  BT KURSTAK ABTS-1857 2021  62.500
## 10            FLONICAMID 2021  59.048
```

```
top_10_all_new <- rbind(top_10_2023_new, top_10_2021_new)

ggplot(top_10_all_new, aes(x = reorder(Chemical_Name, -Value), y = Value, fill = as.factor(Year))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Top 10 Chemicals by Total Usage for 2021, and 2023",
       x = "Chemical Name",
       y = "Total Usage (Value)",
       fill = "Year") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



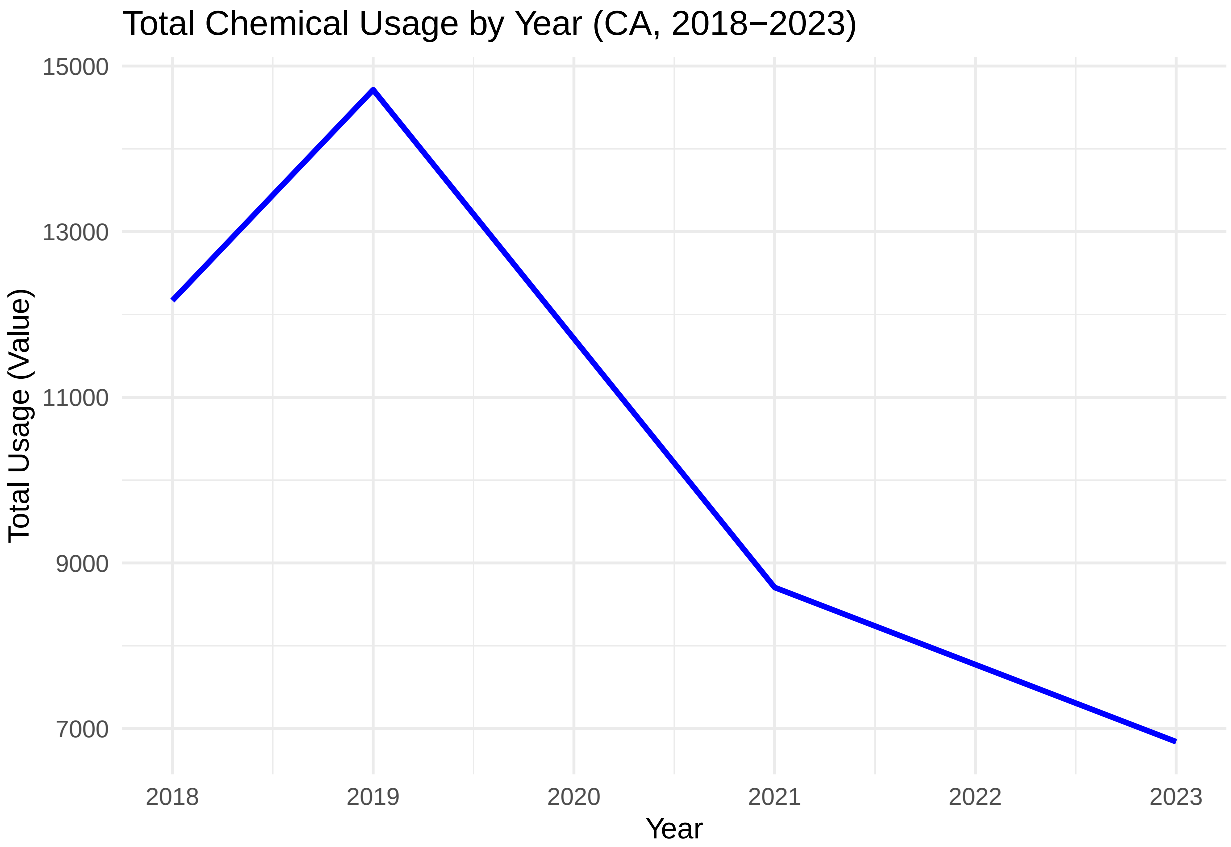
```
#We can see more same chemical are used, but there are are some new chemicals as well.
```

6.Another question we want to explore is that is there any shifts of usage in different chemicals.

```
ca_chemical_total <- aggregate(Value ~ Year, data = ca_chemical2, FUN = sum, na.rm = TRUE)

# Plot the total Value for each year using ggplot
ggplot(ca_chemical_total, aes(x = Year, y = Value)) +
    geom_line(color = "blue", size = 1) +
    labs(title = "Total Chemical Usage by Year (CA, 2018-2023)",
         x = "Year",
         y = "Total Usage (Value)") +
    theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
#We can find the the usage of different chemicals are actually become less and less.
```