Strawberry

Yibing Wang

2024-10-02

```
#devtools::install_github("UrbanInstitute/urbnmapr")
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(ggplot2)
library(sf)
```

```
## Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE
```

```
library(mapview)
library(urbnmapr)
```

```
strawberry <- read.csv("strawberries25_v3.csv", header = TRUE)
head(strawberry)
```

```
##   Program Year Period Week.Ending Geo.Level   State State.ANSI Ag.District
## 1  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 2  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 3  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 4  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 5  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 6  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
##   Ag.District.Code  County County.ANSI Zip.Code Region watershed_code Watershed
## 1               40 BULLOCK          11       NA     NA              0        NA
## 2               40 BULLOCK          11       NA     NA              0        NA
## 3               40 BULLOCK          11       NA     NA              0        NA
## 4               40 BULLOCK          11       NA     NA              0        NA
## 5               40 BULLOCK          11       NA     NA              0        NA
## 6               40 BULLOCK          11       NA     NA              0        NA
##      Commodity                                       Data.Item Domain
## 1 STRAWBERRIES                    STRAWBERRIES - ACRES BEARING  TOTAL
## 2 STRAWBERRIES                      STRAWBERRIES - ACRES GROWN  TOTAL
## 3 STRAWBERRIES                STRAWBERRIES - ACRES NON-BEARING  TOTAL
## 4 STRAWBERRIES     STRAWBERRIES - OPERATIONS WITH AREA BEARING  TOTAL
## 5 STRAWBERRIES       STRAWBERRIES - OPERATIONS WITH AREA GROWN  TOTAL
## 6 STRAWBERRIES STRAWBERRIES - OPERATIONS WITH AREA NON-BEARING  TOTAL
##   Domain.Category Value CV....
## 1   NOT SPECIFIED   (D)    (D)
## 2   NOT SPECIFIED     3   15.7
## 3   NOT SPECIFIED   (D)    (D)
## 4   NOT SPECIFIED     1    (L)
## 5   NOT SPECIFIED     6   52.7
## 6   NOT SPECIFIED     5   47.6
```

Data Cleaning

```
# To clean the data set, I need to identify and remove any column that has the same value for every row.
# For example, the column "watershed_code" has the value "0" in all rows. We can remove those columns.
drop_col <- function(df) {
  df %>% select_if(~ length(unique(.)) > 1)
}
strawberry_clean <- drop_col(strawberry)
head(strawberry_clean)
```

```
##   Program Year Period Geo.Level   State State.ANSI Ag.District Ag.District.Code
## 1  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 2  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 3  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 4  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 5  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 6  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
##    County County.ANSI                                       Data.Item Domain
## 1 BULLOCK          11                    STRAWBERRIES - ACRES BEARING  TOTAL
## 2 BULLOCK          11                      STRAWBERRIES - ACRES GROWN  TOTAL
## 3 BULLOCK          11                STRAWBERRIES - ACRES NON-BEARING  TOTAL
## 4 BULLOCK          11     STRAWBERRIES - OPERATIONS WITH AREA BEARING  TOTAL
## 5 BULLOCK          11       STRAWBERRIES - OPERATIONS WITH AREA GROWN  TOTAL
## 6 BULLOCK          11 STRAWBERRIES - OPERATIONS WITH AREA NON-BEARING  TOTAL
##   Domain.Category Value CV....
## 1   NOT SPECIFIED   (D)    (D)
## 2   NOT SPECIFIED     3   15.7
## 3   NOT SPECIFIED   (D)    (D)
## 4   NOT SPECIFIED     1    (L)
## 5   NOT SPECIFIED     6   52.7
## 6   NOT SPECIFIED     5   47.6
```

```
# Next, I will check the "State" column to make sure there is no missing value.
state_1 <- strawberry_clean %>%
  group_by(State) %>%
  count()
count(state_1)
```

```
## # A tibble: 52 × 2
## # Groups:   State [52]
##    State           n
##    <chr>       <int>
##  1 ALABAMA         1
##  2 ALASKA          1
##  3 ARIZONA         1
##  4 ARKANSAS        1
##  5 CALIFORNIA      1
##  6 COLORADO        1
##  7 CONNECTICUT     1
##  8 DELAWARE        1
##  9 FLORIDA         1
## 10 GEORGIA         1
## # ℹ 42 more rows
```

```
sum(state_1$n) == dim(strawberry_clean)[1]
```

```
## [1] TRUE
```

```
# The sum of the total number of rows(State) is equal to the total row number of the dataset(cleaned). There is n
o missing value in the column "State"
```

```
# After checking the State info, I could use one state as an example to to help me understand the structure and c
ontent of the dataset before analyzing the entire one.
state_summary <- strawberry_clean %>%
  group_by(State) %>%
  summarize(count = n())
print(state_summary)
```

```
## # A tibble: 52 × 2
##    State       count
##    <chr>       <int>
##  1 ALABAMA       154
##  2 ALASKA         41
##  3 ARIZONA        47
##  4 ARKANSAS      120
##  5 CALIFORNIA   2575
##  6 COLORADO      105
##  7 CONNECTICUT    70
##  8 DELAWARE       22
##  9 FLORIDA      1569
## 10 GEORGIA       284
## # ℹ 42 more rows
```

```
# As we can see from the output, California is the largest strawberries producer in the U.S.
```

```
# Analyzing the California strawberry data. I will filter the dataset base on "California" first, and then split
the data by "Census" and "Survey".
cali_census <- strawberry_clean %>%
  filter(State == "CALIFORNIA", Program == "CENSUS") %>%
  select(Year, `Data.Item`, Value)
head(cali_census)
```

```
##   Year                                   Data.Item Value
## 1 2022                STRAWBERRIES - ACRES BEARING   (D)
## 2 2022                  STRAWBERRIES - ACRES GROWN   (D)
## 3 2022 STRAWBERRIES - OPERATIONS WITH AREA BEARING     3
## 4 2022   STRAWBERRIES - OPERATIONS WITH AREA GROWN     3
## 5 2022                STRAWBERRIES - ACRES BEARING   (D)
## 6 2022                  STRAWBERRIES - ACRES GROWN   (D)
```

```
cali_survey <- strawberry_clean %>%
  filter(State == "CALIFORNIA", Program == "SURVEY") %>%
  select(Year, Period, `Data.Item`, Value)
head(cali_survey)
```

```
##   Year         Period
## 1 2023 MARKETING YEAR
## 2 2023 MARKETING YEAR
## 3 2023 MARKETING YEAR
## 4 2023           YEAR
## 5 2023           YEAR
## 6 2023           YEAR
##                                                            Data.Item  Value
## 1               STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT    121
## 2 STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / CWT    (D)
## 3   STRAWBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN $ / CWT    (D)
## 4                                 STRAWBERRIES - ACRES HARVESTED 42,700
## 5                                   STRAWBERRIES - ACRES PLANTED 43,100
## 6                       STRAWBERRIES - APPLICATIONS, MEASURED IN LB    (D)
```

```
# I've noticed that in the Data.Item column, there are multiple info.
# After my first time trying to split the column, some info won't switch to the correct column. I've checked the
raw data, I think it might because there are some "-" instead of ",", so I need to standardize it first.
process_line <- function(line) {
  line <- as.character(line)
  # Replace any kind of dash (-, –, -, etc.) with a common dash (regular hyphen)
  line <- gsub("[---]", "-", line)
  # Split by "-" to get the main components
  parts <- unlist(strsplit(line, " - "))
  fruit <- "Strawberries"

  # Identify Category, Item, and Metric
  if (length(parts) == 2) {
    #separate Item and Metric
    item_metric <- unlist(strsplit(parts[2], ","))
     # Remove "STRAWBERRIES"
    category <- trimws(gsub("^STRAWBERRIES,? ?", "", parts[1]))

    #if the category is empty, NA
    if (category == "") {
      category <- NA
    }

    item <- trimws(ifelse(length(item_metric) > 0, item_metric[1], "N/A"))
    metric <- trimws(ifelse(length(item_metric) > 1, item_metric[2], "N/A"))

  } else if (length(parts) == 3) {
    # If three parts are found, the second part is Category and the third is Item + Metric
    category <- trimws(gsub("^STRAWBERRIES,? ?", "", parts[2]))
    if (category == "") {
      category <- NA
    }

    item_metric <- unlist(strsplit(parts[3], ","))
    item <- trimws(ifelse(length(item_metric) > 0, item_metric[1], "N/A"))
    metric <- trimws(ifelse(length(item_metric) > 1, item_metric[2], "N/A"))

  } else {
    category <- trimws(gsub("^STRAWBERRIES,? ?", "", parts[1]))
    if (category == "") {
      category <- NA
    }

    item <- "N/A"
    metric <- "N/A"
  }
  return(list(Fruit = fruit, Category = category, Item = item, Metric = metric))
}

strawberry_clean <- cbind(strawberry_clean, do.call(rbind, lapply(strawberry_clean$Data.Item, function(x) {
  as.data.frame(process_line(x), stringsAsFactors = FALSE)
})))
head(strawberry_clean)
```

```
##   Program Year Period Geo.Level   State State.ANSI Ag.District Ag.District.Code
## 1  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 2  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 3  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 4  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 5  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 6  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
##    County County.ANSI                                       Data.Item Domain
## 1 BULLOCK          11                    STRAWBERRIES - ACRES BEARING  TOTAL
## 2 BULLOCK          11                      STRAWBERRIES - ACRES GROWN  TOTAL
## 3 BULLOCK          11                STRAWBERRIES - ACRES NON-BEARING  TOTAL
## 4 BULLOCK          11     STRAWBERRIES - OPERATIONS WITH AREA BEARING  TOTAL
## 5 BULLOCK          11       STRAWBERRIES - OPERATIONS WITH AREA GROWN  TOTAL
## 6 BULLOCK          11 STRAWBERRIES - OPERATIONS WITH AREA NON-BEARING  TOTAL
##   Domain.Category Value CV....        Fruit Category
## 1   NOT SPECIFIED   (D)    (D) Strawberries     <NA>
## 2   NOT SPECIFIED     3   15.7 Strawberries     <NA>
## 3   NOT SPECIFIED   (D)    (D) Strawberries     <NA>
## 4   NOT SPECIFIED     1    (L) Strawberries     <NA>
## 5   NOT SPECIFIED     6   52.7 Strawberries     <NA>
## 6   NOT SPECIFIED     5   47.6 Strawberries     <NA>
##                            Item Metric
## 1                 ACRES BEARING    N/A
## 2                   ACRES GROWN    N/A
## 3             ACRES NON-BEARING    N/A
## 4     OPERATIONS WITH AREA BEARING    N/A
## 5       OPERATIONS WITH AREA GROWN    N/A
## 6 OPERATIONS WITH AREA NON-BEARING    N/A
```

```
# The "domain.Category" column also has multiple info.
DC_1 <- strawberry_clean %>%
  group_by(Domain.Category) %>%
  count()
count(DC_1)
```

```
## # A tibble: 191 × 2
## # Groups:   Domain.Category [191]
##    Domain.Category                                                            n
##    <chr>                                                                  <int>
##  1 AREA GROWN: (0.1 TO 0.9 ACRES)                                             1
##  2 AREA GROWN: (1.0 TO 4.9 ACRES)                                             1
##  3 AREA GROWN: (100 OR MORE ACRES)                                            1
##  4 AREA GROWN: (15.0 TO 24.9 ACRES)                                           1
##  5 AREA GROWN: (25.0 TO 49.9 ACRES)                                           1
##  6 AREA GROWN: (5.0 TO 14.9 ACRES)                                            1
##  7 AREA GROWN: (50.0 TO 99.9 ACRES)                                           1
##  8 CHEMICAL, FUNGICIDE: (AZOXYSTROBIN = 128810)                               1
##  9 CHEMICAL, FUNGICIDE: (BACILLUS AMYLOLIQUEFAC F727 = 16489)                 1
## 10 CHEMICAL, FUNGICIDE: (BACILLUS AMYLOLIQUEFACIENS MBI 600 = 129082)         1
## # ℹ 181 more rows
```

```
strawberry_clean <- strawberry_clean %>%
  separate_wider_delim(cols = `Domain.Category`, delim = ": ",
                       names = c("use", "details"),
                       too_many = "error", too_few = "align_start") %>%
  mutate(
    name = str_extract(details, "(?<=\\().*?(?=\\=)"),
    code = str_extract(details, "(?<=\\= ).*?(?=\\))")
  )
strawberry_clean$use <- gsub("^CHEMICAL, ", "", strawberry_clean$use)
head(strawberry_clean)
```

```
## # A tibble: 6 × 22
##   Program  Year Period Geo.Level State   State.ANSI Ag.District Ag.District.Code
##   <chr>   <int> <chr>  <chr>     <chr>        <int> <chr>                  <int>
## 1 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 2 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 3 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 4 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 5 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 6 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## # ℹ 14 more variables: County <chr>, County.ANSI <int>, Data.Item <chr>,
## #   Domain <chr>, use <chr>, details <chr>, Value <chr>, CV.... <chr>,
## #   Fruit <chr>, Category <chr>, Item <chr>, Metric <chr>, name <chr>,
## #   code <chr>
```

```
#for value and cv, there are letters inside, I need to change them to NA.
strawberry_clean$Value <- as.numeric(as.character(strawberry_clean$Value))
```

```
## Warning: NAs introduced by coercion
```

```
strawberry_clean$CV.... <- as.numeric(as.character(strawberry_clean$CV....))
```

```
## Warning: NAs introduced by coercion
```

```
head(strawberry_clean)
```

```
## # A tibble: 6 × 22
##   Program  Year Period Geo.Level State   State.ANSI Ag.District Ag.District.Code
##   <chr>   <int> <chr>  <chr>     <chr>        <int> <chr>                  <int>
## 1 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 2 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 3 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 4 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 5 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 6 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## # ℹ 14 more variables: County <chr>, County.ANSI <int>, Data.Item <chr>,
## #   Domain <chr>, use <chr>, details <chr>, Value <dbl>, CV.... <dbl>,
## #   Fruit <chr>, Category <chr>, Item <chr>, Metric <chr>, name <chr>,
## #   code <chr>
```

```
#delate data.item
strawberry_clean <- strawberry_clean %>%
  select(-Data.Item)
head(strawberry_clean)
```

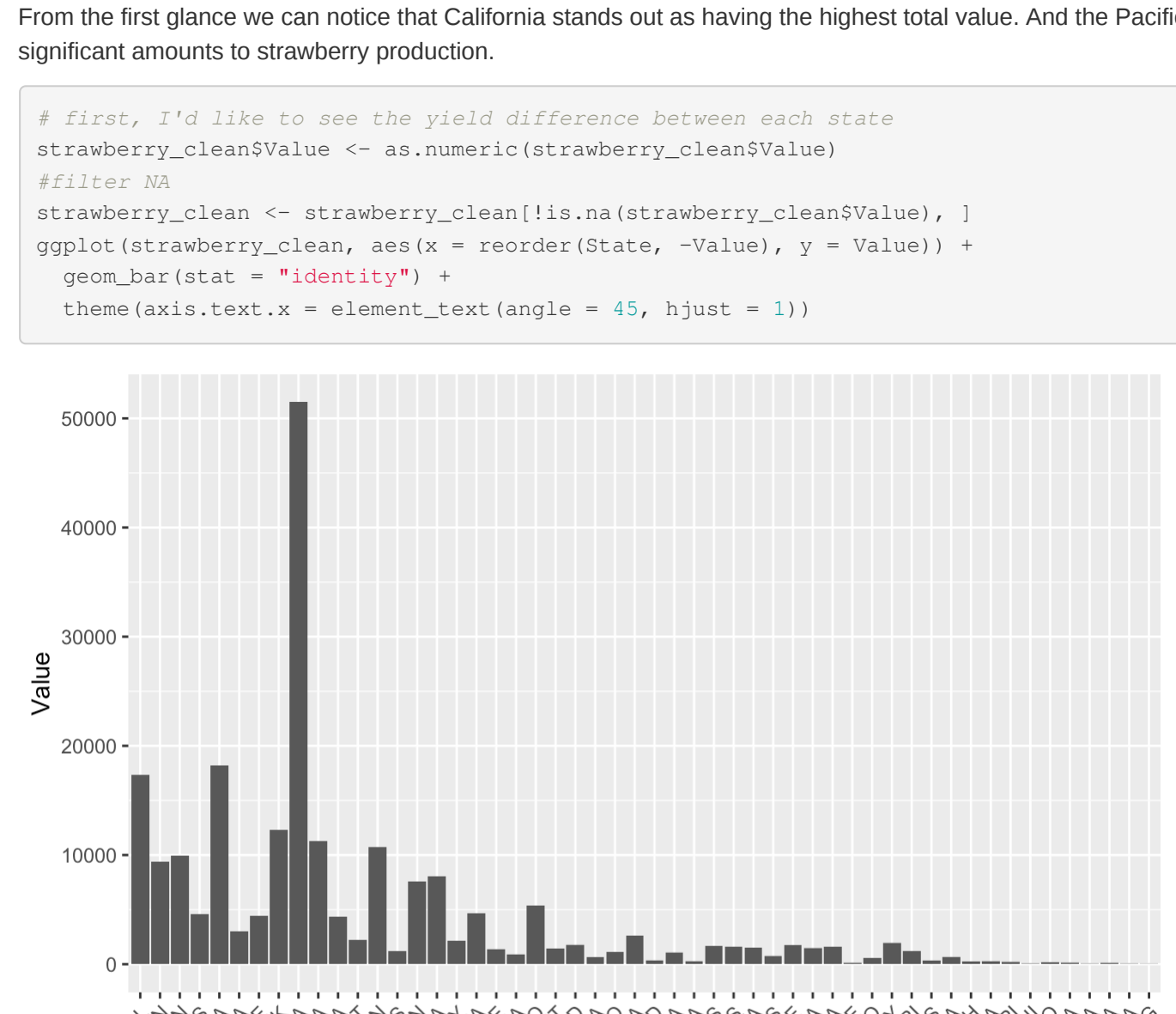
```
## # A tibble: 6 × 21
##   Program  Year Period Geo.Level State   State.ANSI Ag.District Ag.District.Code
##   <chr>   <int> <chr>  <chr>     <chr>        <int> <chr>                  <int>
## 1 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 2 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 3 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 4 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 5 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## 6 CENSUS   2022 YEAR   COUNTY    ALABAMA          1 BLACK BELT                40
## # ℹ 13 more variables: County <chr>, County.ANSI <int>, Domain <chr>,
## #   use <chr>, details <chr>, Value <dbl>, CV.... <dbl>, Fruit <chr>,
## #   Category <chr>, Item <chr>, Metric <chr>, name <chr>, code <chr>
```

Data visualizing

```
counties <- get_urbn_map(map = "states", sf = TRUE)
strawberry_clean$State.ANSI <- as.character(strawberry_clean$State.ANSI)
strawberry_clean$State.ANSI <- str_pad(strawberry_clean$State.ANSI, width = 2, pad = "0")


strawberry_map <- counties %>%
  left_join(strawberry_clean, by = c("state_fips" = "State.ANSI"))
```

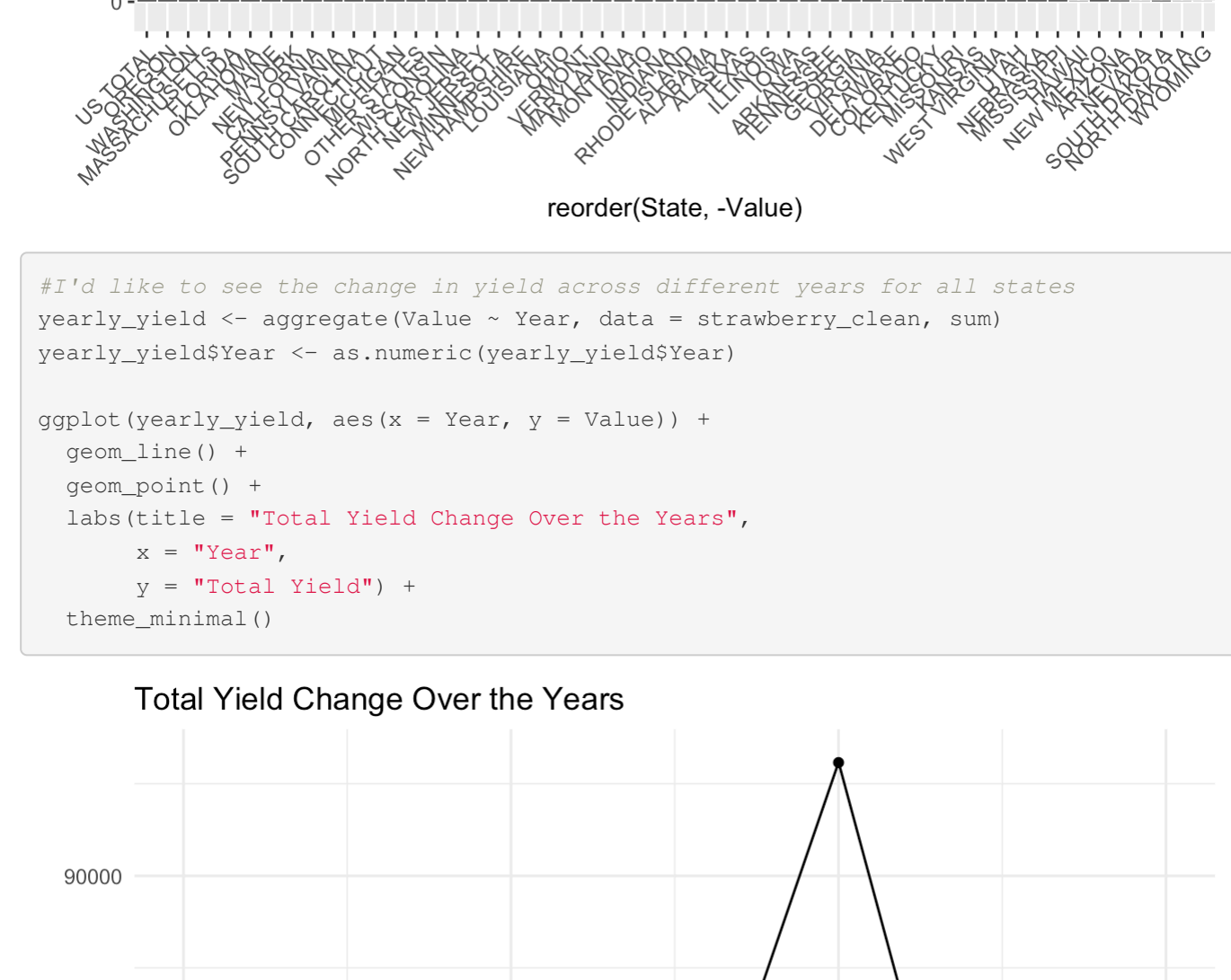
```
## old-style crs object detected; please recreate object with a recent sf::st_crs()
```

```
strawberry_count <- strawberry_map %>%
  group_by(State) %>%
  summarise(total_value = sum(Value, na.rm = TRUE))

mapview(strawberry_count, zcol = 'total_value')
```

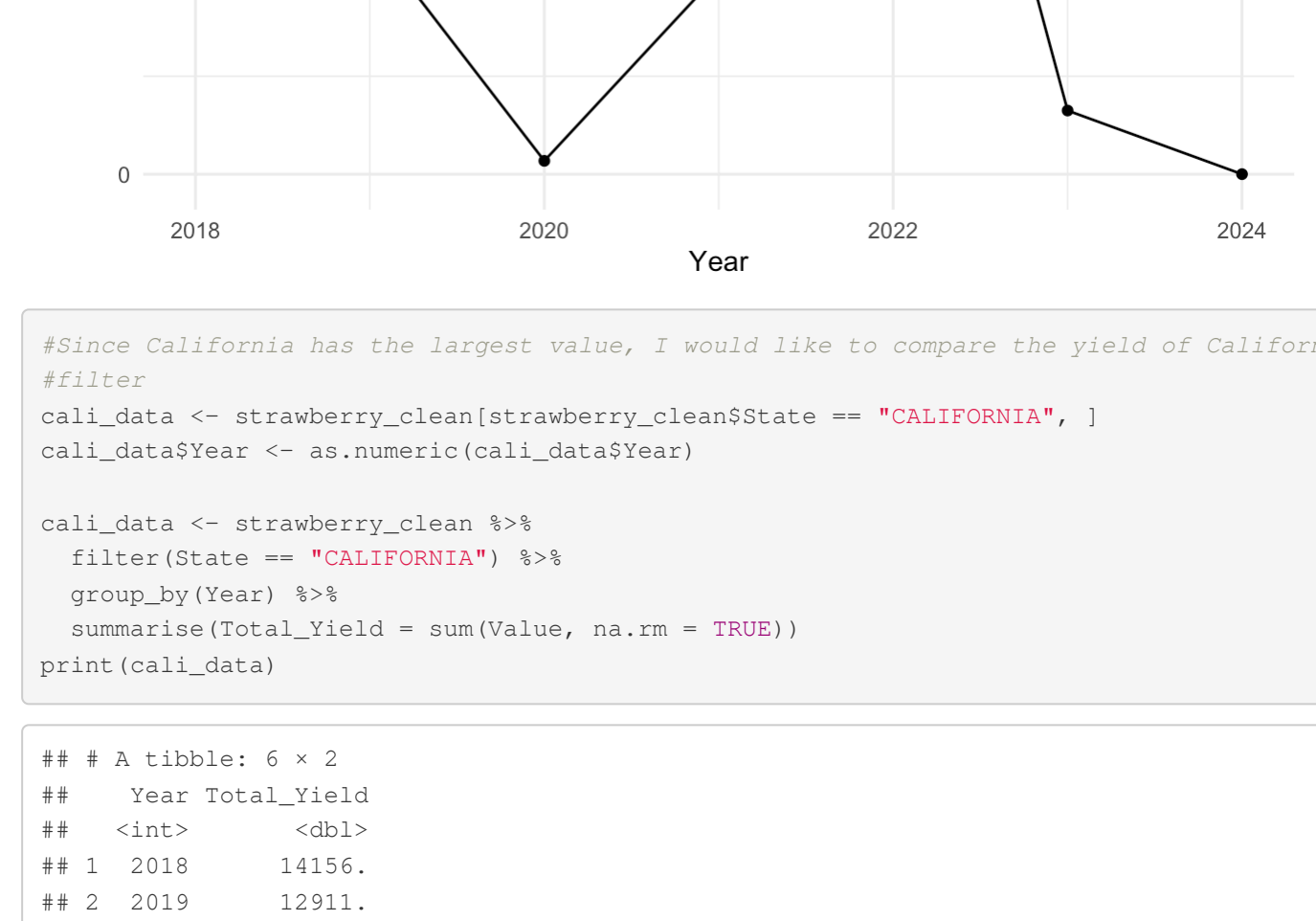


From the first glance we can notice that California stands out as having the highest total value. And the Pacific Northwest as a whole contribute significant amounts to strawberry production.

```
# first, I'd like to see the yield difference between each state
strawberry_clean$Value <- as.numeric(strawberry_clean$Value)
#filter NA
strawberry_clean <- strawberry_clean[!is.na(strawberry_clean$Value), ]
ggplot(strawberry_clean, aes(x = reorder(State, -Value), y = Value)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



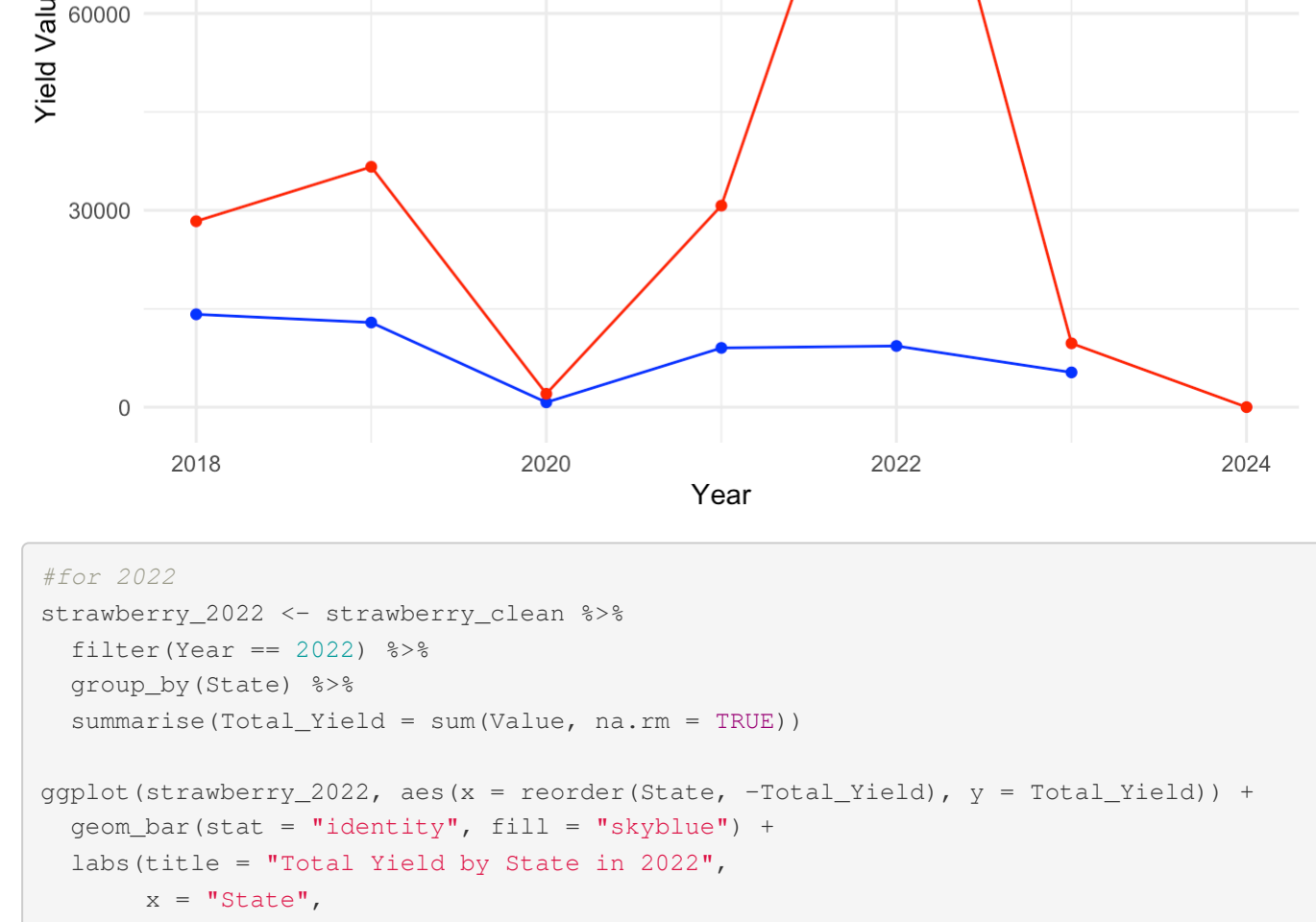

```
#I'd like to see the change in yield across different years for all states
yearly_yield <- aggregate(Value ~ Year, data = strawberry_clean, sum)
yearly_yield$Year <- as.numeric(yearly_yield$Year)

ggplot(yearly_yield, aes(x = Year, y = Value)) +
  geom_line() +
  geom_point() +
  labs(title = "Total Yield Change Over the Years",
       x = "Year",
       y = "Total Yield") +
  theme_minimal()
```



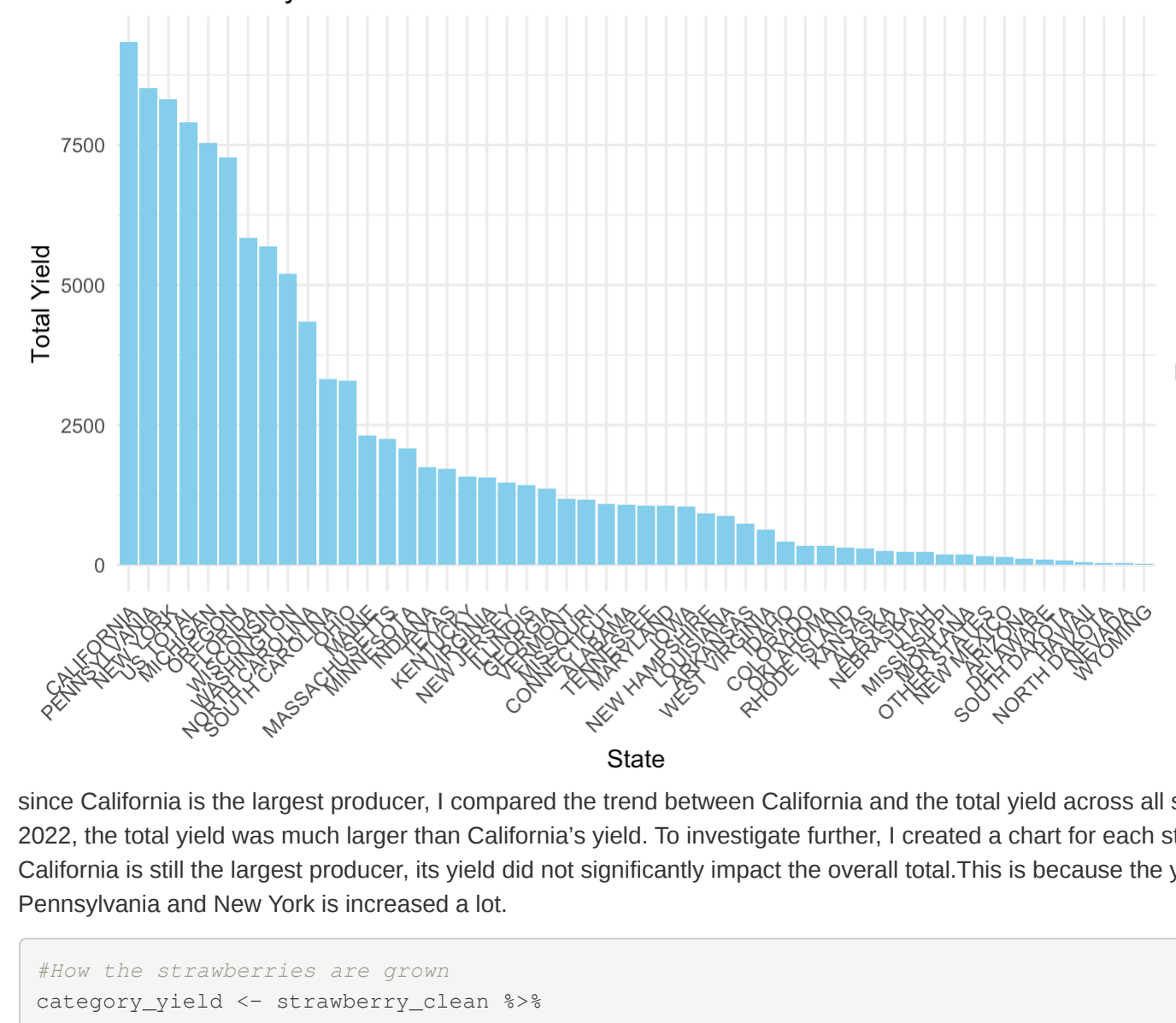

```
#Since California has the largest value, I would like to compare the yield of California and all states
#filter
cali_data <- strawberry_clean[strawberry_clean$State == "CALIFORNIA", ]
cali_data$Year <- as.numeric(cali_data$Year)

cali_data <- strawberry_clean %>%
  filter(State == "CALIFORNIA") %>%
  group_by(Year) %>%
  summarise(Total_Yield = sum(Value, na.rm = TRUE))
print(cali_data)
```

```
## # A tibble: 6 × 2
##    Year Total_Yield
##   <int>       <dbl>
## 1  2018      14156.
## 2  2019      12911.
## 3  2020        743.
## 4  2021       9039.
## 5  2022       9334
## 6  2023       5307.
```

```
ggplot() +
  geom_line(data = cali_data, aes(x = Year, y = Total_Yield), color = "blue") +
  geom_line(data = yearly_yield, aes(x = Year, y = Value), color = "red") +
  geom_point(data = cali_data, aes(x = Year, y = Total_Yield), color = "blue") +
  geom_point(data = yearly_yield, aes(x = Year, y = Value), color = "red") +
  labs(title = "California vs All States Yield Comparison",
       x = "Year",
       y = "Yield Value") +
  theme_minimal()
```

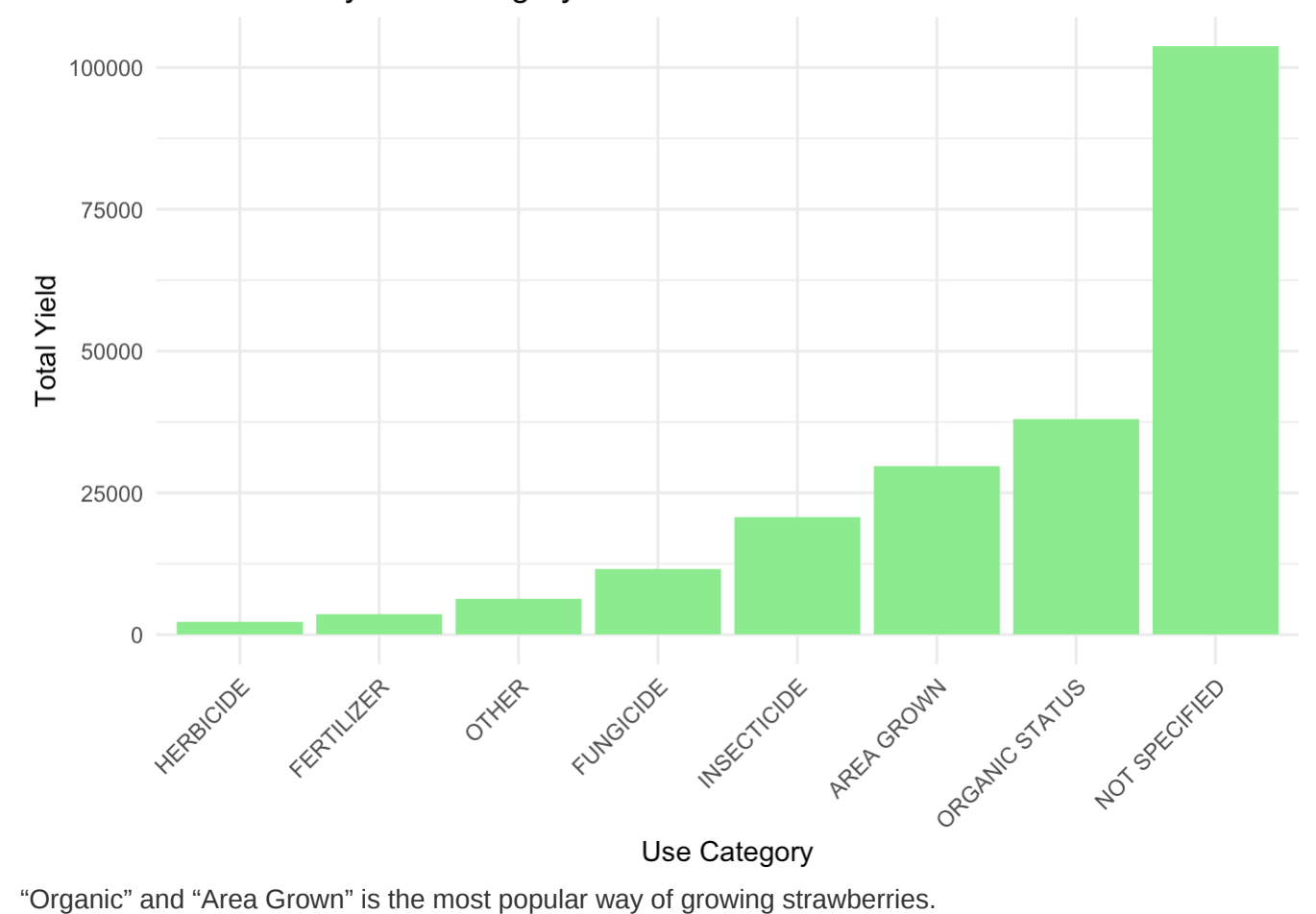



```
#for 2022
strawberry_2022 <- strawberry_clean %>%
  filter(Year == 2022) %>%
  group_by(State) %>%
  summarise(Total_Yield = sum(Value, na.rm = TRUE))

ggplot(strawberry_2022, aes(x = reorder(State, -Total_Yield), y = Total_Yield)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Total Yield by State in 2022",
       x = "State",
       y = "Total Yield") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



I plotted the total yield per year and, since California is the largest producer, I compared the trend between California and the total yield across all states. However, I noticed that in 2022, the total yield was much larger than California's yield. To investigate further, I created a chart for each state's yield and found that although California is still the largest producer, its yield did not significantly impact the overall total.This is because the yield of other states, such as Pennsylvania and New York is increased a lot.

```
#How the strawberries are grown
category_yield <- strawberry_clean %>%
  filter(!is.na(use) & !is.na(Value)) %>%
  group_by(use) %>%
  summarise(Total_Yield = sum(Value, na.rm = TRUE))

ggplot(category_yield, aes(x = reorder(use, Total_Yield), y = Total_Yield)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Total Yield by Use Category", x = "Use Category", y = "Total Yield") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



by comparing the data we have, “Organic” and “Area Grown” is the most popular way of growing strawberries.

```
#write.csv(strawberry_clean, "strawberry_clean.csv", row.names = FALSE)
```