hw

2023-10-14

##Data acquisition and assessment: The data set for this assignment has been selected from: USDA_NASS (https://quickstats.nass.usda.gov)
The data have been stored on NASS here: USDA_NASS_strawb_2023SEP19 (https://quickstats.nass.usda.gov/results/45FBC825-B104-38E2-9802-839F5F3C7036)

Data cleaning and organization

```
library(knitr)
library(kableExtra)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter()     masks stats::filter()
## ✖ dplyr::group_rows() masks kableExtra::group_rows()
## ✖ dplyr::lag()        masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(stringr)
library(dplyr)
library(tidyr)
library(ggplot2)
```

Outline the approach taked to clean and organize the data.

1. take a look at the data set and an overview of this data.

2. drop one-item columns because they are not useful when we are analyzing.

3. understand which column has missing values and what should I do to drop missing values.

4. deal with the data Item part and separate to census and survey data.

5. divide the data to weight and sales so that we can have same measure of values.

6. do some visualizations to see the pattern

    a. the chemical from domain.
    b. the value by state, from weight and sales.
    c. the value by whether organic, from weight and sales.

```
straw = read.csv("strawberry.csv",header = TRUE)
dim(straw)
```

```
## [1] 4314   21
```

```
head(straw)
```

```
##   Program Year Period Week.Ending Geo.Level  State State.ANSI Ag.District
## 1  CENSUS 2021   YEAR          NA     STATE ALASKA          2          NA
## 2  CENSUS 2021   YEAR          NA     STATE ALASKA          2          NA
## 3  CENSUS 2021   YEAR          NA     STATE ALASKA          2          NA
## 4  CENSUS 2021   YEAR          NA     STATE ALASKA          2          NA
## 5  CENSUS 2021   YEAR          NA     STATE ALASKA          2          NA
## 6  CENSUS 2021   YEAR          NA     STATE ALASKA          2          NA
##   Ag.District.Code County County.ANSI Zip.Code Region watershed_code Watershed
## 1               NA     NA          NA       NA     NA              0        NA
## 2               NA     NA          NA       NA     NA              0        NA
## 3               NA     NA          NA       NA     NA              0        NA
## 4               NA     NA          NA       NA     NA              0        NA
## 5               NA     NA          NA       NA     NA              0        NA
## 6               NA     NA          NA       NA     NA              0        NA
##      Commodity                                                   Data.Item
## 1 STRAWBERRIES               STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES
## 2 STRAWBERRIES         STRAWBERRIES, ORGANIC - PRODUCTION, MEASURED IN CWT
## 3 STRAWBERRIES                STRAWBERRIES, ORGANIC - SALES, MEASURED IN $
## 4 STRAWBERRIES              STRAWBERRIES, ORGANIC - SALES, MEASURED IN CWT
## 5 STRAWBERRIES STRAWBERRIES, ORGANIC, FRESH MARKET - OPERATIONS WITH SALES
## 6 STRAWBERRIES  STRAWBERRIES, ORGANIC, FRESH MARKET - SALES, MEASURED IN $
##           Domain                      Domain.Category Value CV....
## 1 ORGANIC STATUS ORGANIC STATUS: (NOP USDA CERTIFIED)     2    (H)
## 2 ORGANIC STATUS ORGANIC STATUS: (NOP USDA CERTIFIED)   (D)    (D)
## 3 ORGANIC STATUS ORGANIC STATUS: (NOP USDA CERTIFIED)   (D)    (D)
## 4 ORGANIC STATUS ORGANIC STATUS: (NOP USDA CERTIFIED)   (D)    (D)
## 5 ORGANIC STATUS ORGANIC STATUS: (NOP USDA CERTIFIED)     2    (H)
## 6 ORGANIC STATUS ORGANIC STATUS: (NOP USDA CERTIFIED)   (D)    (D)
```

It has 4313 obs and 21 columns. and as we can see there are a lot of missing values in this data.

```
## define function
drop_one_value_col <- function(df){
col_name <- NULL
col_val <- NULL
suppressWarnings({
for(i in 1:dim(df)[2]){
if((df |> distinct(df[,i]) |> count()) == 1){
  col_name = c(col_name, colnames(df[i]))
  col_val = c(col_val, df[1,i])
} }
})

if(is.null(col_name)){return("No Columns to drop")}else{
   col_val = unlist(col_val)
   attributes(col_val) = NULL
   drp = data.frame(col_name, col_val)
   return(drp)
   }
}

str <- drop_one_value_col(straw)

# str |> kable(caption = "Dropped Single-Value Columns: names and values")

str <- str$col_name

strawberry <- straw|> select(!all_of(str))
head(strawberry)
```

```
##   Program Year Period  State State.ANSI
## 1  CENSUS 2021   YEAR ALASKA          2
## 2  CENSUS 2021   YEAR ALASKA          2
## 3  CENSUS 2021   YEAR ALASKA          2
## 4  CENSUS 2021   YEAR ALASKA          2
## 5  CENSUS 2021   YEAR ALASKA          2
## 6  CENSUS 2021   YEAR ALASKA          2
##                                                      Data.Item         Domain
## 1               STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES ORGANIC STATUS
## 2         STRAWBERRIES, ORGANIC - PRODUCTION, MEASURED IN CWT ORGANIC STATUS
## 3                STRAWBERRIES, ORGANIC - SALES, MEASURED IN $ ORGANIC STATUS
## 4              STRAWBERRIES, ORGANIC - SALES, MEASURED IN CWT ORGANIC STATUS
## 5 STRAWBERRIES, ORGANIC, FRESH MARKET - OPERATIONS WITH SALES ORGANIC STATUS
## 6  STRAWBERRIES, ORGANIC, FRESH MARKET - SALES, MEASURED IN $ ORGANIC STATUS
##                        Domain.Category Value CV....
## 1 ORGANIC STATUS: (NOP USDA CERTIFIED)     2    (H)
## 2 ORGANIC STATUS: (NOP USDA CERTIFIED)   (D)    (D)
## 3 ORGANIC STATUS: (NOP USDA CERTIFIED)   (D)    (D)
## 4 ORGANIC STATUS: (NOP USDA CERTIFIED)   (D)    (D)
## 5 ORGANIC STATUS: (NOP USDA CERTIFIED)     2    (H)
## 6 ORGANIC STATUS: (NOP USDA CERTIFIED)   (D)    (D)
```

After drop this columns, we have about 10 columns left and looks like some missing values are already gone.

```
#drop missing values, not only for na, but also some values that can not be understand.

is_na<- sapply(strawberry, function(column) sum(is.na(column)))
is_na
```

```
##         Program            Year          Period           State      State.ANSI
##               0               0               0               0              86
##       Data.Item          Domain Domain.Category           Value          CV....
##               0               0               0               0               0
```

```
# only ANSI have some NAs, and we can delete these observation
strawberry = strawberry[!is.na(strawberry$State.ANSI),]
```

```
# Also, we see the value part and CV part has some value that I am not understand, and I
want convert them to NA and, delete the comma in the number.
# by checking the unique of the value, we can see that Value part need to deal with D,N
A,Z
#,and CV need to H,D.

strawberry$CV....[strawberry$CV.... %in% c("(H)", "(D)")] <- NA

strawberry$Value[strawberry$Value %in% c(" (D)"," (NA)"," (Z)")] <- NA

#delete them
strawberry = strawberry[!is.na(strawberry$CV....),]
strawberry = strawberry[!is.na(strawberry$Value),]

# delete comma
strawberry$Value <- gsub(",", "", strawberry$Value)

# Convert the 'Value' column to numeric
strawberry$Value <- as.numeric(strawberry$Value)
strawberry$CV....<- as.numeric(strawberry$CV....)

head(strawberry)
```

```
##    Program Year Period      State State.ANSI
## 8   CENSUS 2021   YEAR CALIFORNIA          6
## 9   CENSUS 2021   YEAR CALIFORNIA          6
## 10  CENSUS 2021   YEAR CALIFORNIA          6
## 11  CENSUS 2021   YEAR CALIFORNIA          6
## 12  CENSUS 2021   YEAR CALIFORNIA          6
## 14  CENSUS 2021   YEAR CALIFORNIA          6
##                                                        Data.Item         Domain
## 8                 STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES ORGANIC STATUS
## 9           STRAWBERRIES, ORGANIC - PRODUCTION, MEASURED IN CWT ORGANIC STATUS
## 10                 STRAWBERRIES, ORGANIC - SALES, MEASURED IN $ ORGANIC STATUS
## 11               STRAWBERRIES, ORGANIC - SALES, MEASURED IN CWT ORGANIC STATUS
## 12  STRAWBERRIES, ORGANIC, FRESH MARKET - OPERATIONS WITH SALES ORGANIC STATUS
## 14 STRAWBERRIES, ORGANIC, FRESH MARKET - SALES, MEASURED IN CWT ORGANIC STATUS
##                          Domain.Category     Value CV....
## 8  ORGANIC STATUS: (NOP USDA CERTIFIED)       142   19.2
## 9  ORGANIC STATUS: (NOP USDA CERTIFIED)   1413251   51.6
## 10 ORGANIC STATUS: (NOP USDA CERTIFIED) 311784980   46.0
## 11 ORGANIC STATUS: (NOP USDA CERTIFIED)   1412627   51.7
## 12 ORGANIC STATUS: (NOP USDA CERTIFIED)       141   20.4
## 14 ORGANIC STATUS: (NOP USDA CERTIFIED)   1401384   50.6
```

deal with part with data item

```r
# Extract Organic Status
strawberry <- strawberry %>%
  mutate(Organic_Status = ifelse(str_detect(`Data.Item`, "ORGANIC"), 1, 0))

# Extract Market Type
strawberry<- strawberry %>%
  mutate(Market_Type = case_when(
    str_detect(`Data.Item`, "FRESH MARKET") ~ "FRESH MARKET",
    str_detect(`Data.Item`, "PROCESSING") ~ "PROCESSING",
    TRUE ~ "GENERAL"
  ))

# Extract Data Type
strawberry<- strawberry %>%
  mutate(Data_Type = case_when(
    str_detect(`Data.Item`, "OPERATIONS WITH SALES") ~ "OPERATIONS",
    str_detect(`Data.Item`, "PRODUCTION, MEASURED IN CWT") ~ "PRODUCTION_CWT",
    str_detect(`Data.Item`, "SALES, MEASURED IN \\$") ~ "SALES_$",
    str_detect(`Data.Item`, "SALES, MEASURED IN CWT") ~ "SALES_CWT",
    TRUE ~ NA_character_
  ))

# View the first few rows
head(strawberry[, c("Data.Item", "Organic_Status", "Market_Type", "Data_Type")])
```

```
##                                                        Data.Item Organic_Status
## 8                STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES              1
## 9          STRAWBERRIES, ORGANIC - PRODUCTION, MEASURED IN CWT              1
## 10                STRAWBERRIES, ORGANIC - SALES, MEASURED IN $              1
## 11              STRAWBERRIES, ORGANIC - SALES, MEASURED IN CWT              1
## 12  STRAWBERRIES, ORGANIC, FRESH MARKET - OPERATIONS WITH SALES              1
## 14 STRAWBERRIES, ORGANIC, FRESH MARKET - SALES, MEASURED IN CWT              1
##     Market_Type      Data_Type
## 8       GENERAL     OPERATIONS
## 9       GENERAL PRODUCTION_CWT
## 10      GENERAL        SALES_$
## 11      GENERAL      SALES_CWT
## 12 FRESH MARKET     OPERATIONS
## 14 FRESH MARKET      SALES_CWT
```

Separate CENSUS and SURVEY into two Data Frames

```
strwb_census <- strawberry |> filter(Program == "CENSUS")

strwb_survey <- strawberry |> filter(Program == "SURVEY")


strawberry_weight = strawberry[strawberry$Data_Type %in% c("PRODUCTION_CWT","SALES_CW
T"),]

strawberry_sale = strawberry[strawberry$Data_Type %in% c("OPERATIONS","SALES_$"),]
```

Visulization part

chemical discussion

```
unique(strawberry$Domain)
```

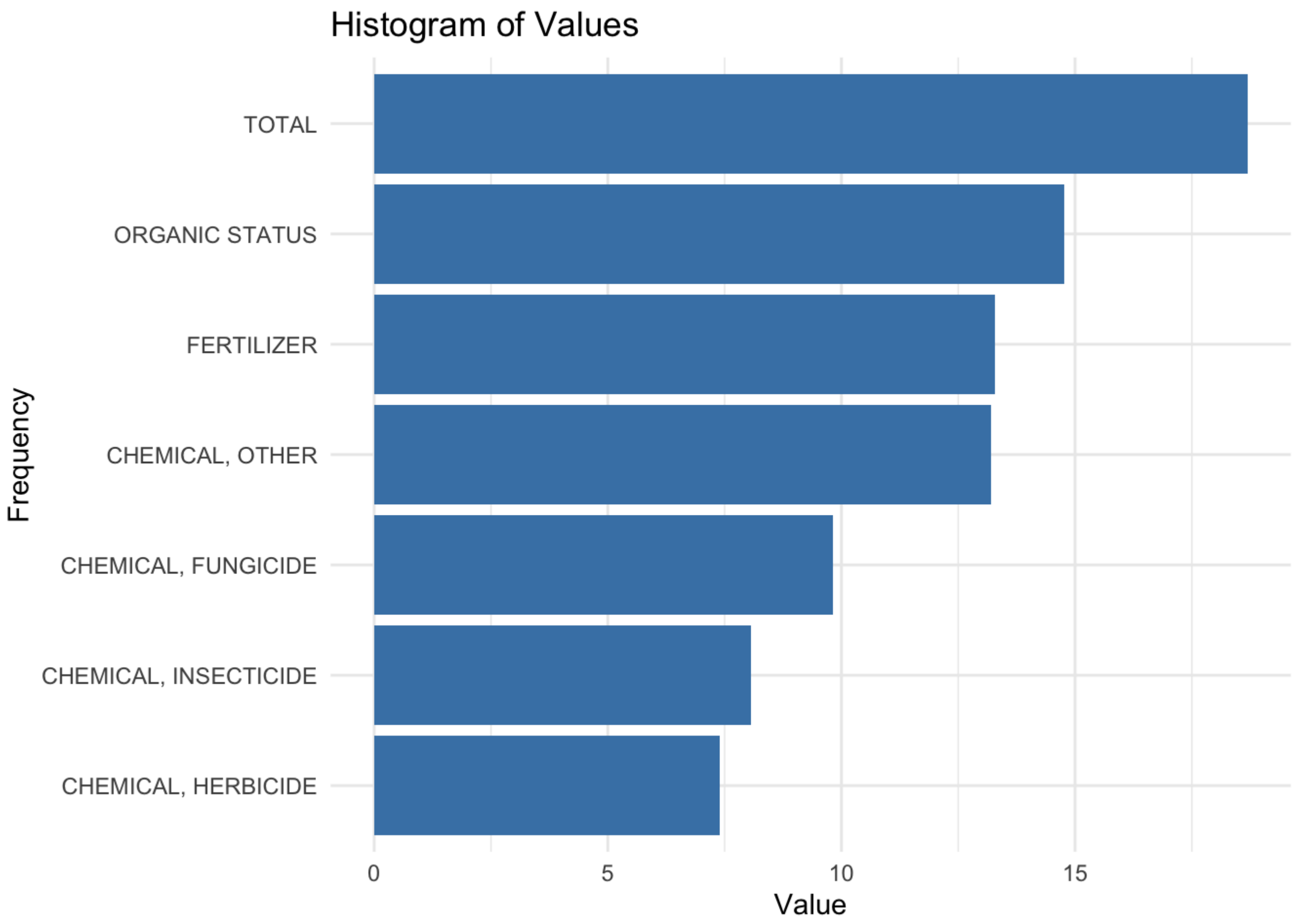
```
## [1] "ORGANIC STATUS"        "TOTAL"                 "CHEMICAL, FUNGICIDE"
## [4] "CHEMICAL, HERBICIDE"   "CHEMICAL, INSECTICIDE" "CHEMICAL, OTHER"
## [7] "FERTILIZER"
```

```
domain_statistics <- strawberry%>%
  group_by(Domain) %>%
  summarise(Count = n(),
            Mean = mean(Value, na.rm = TRUE),
            Median = median(Value, na.rm = TRUE),
            Max = max(Value, na.rm = TRUE))

# Print results
print(domain_statistics)
```

```
## # A tibble: 7 × 5
##   Domain                Count       Mean    Median        Max
##   <chr>                 <int>      <dbl>     <dbl>      <dbl>
## 1 CHEMICAL, FUNGICIDE     515     18302.      1.6     1233500
## 2 CHEMICAL, HERBICIDE      62      1641.      1.36      19600
## 3 CHEMICAL, INSECTICIDE   503      3201.      1.24     279600
## 4 CHEMICAL, OTHER          94    543781.     83.4     7698900
## 5 FERTILIZER               55    589756.     18      10676000
## 6 ORGANIC STATUS          556   2590336.    111     311784980
## 7 TOTAL                   304 130595923. 117500    3030953000
```

```
ggplot(domain_statistics, aes( x = log(Mean),y =reorder(Domain, Mean))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Histogram of Values", x = "Value", y = "Frequency") +
  theme_minimal()
```

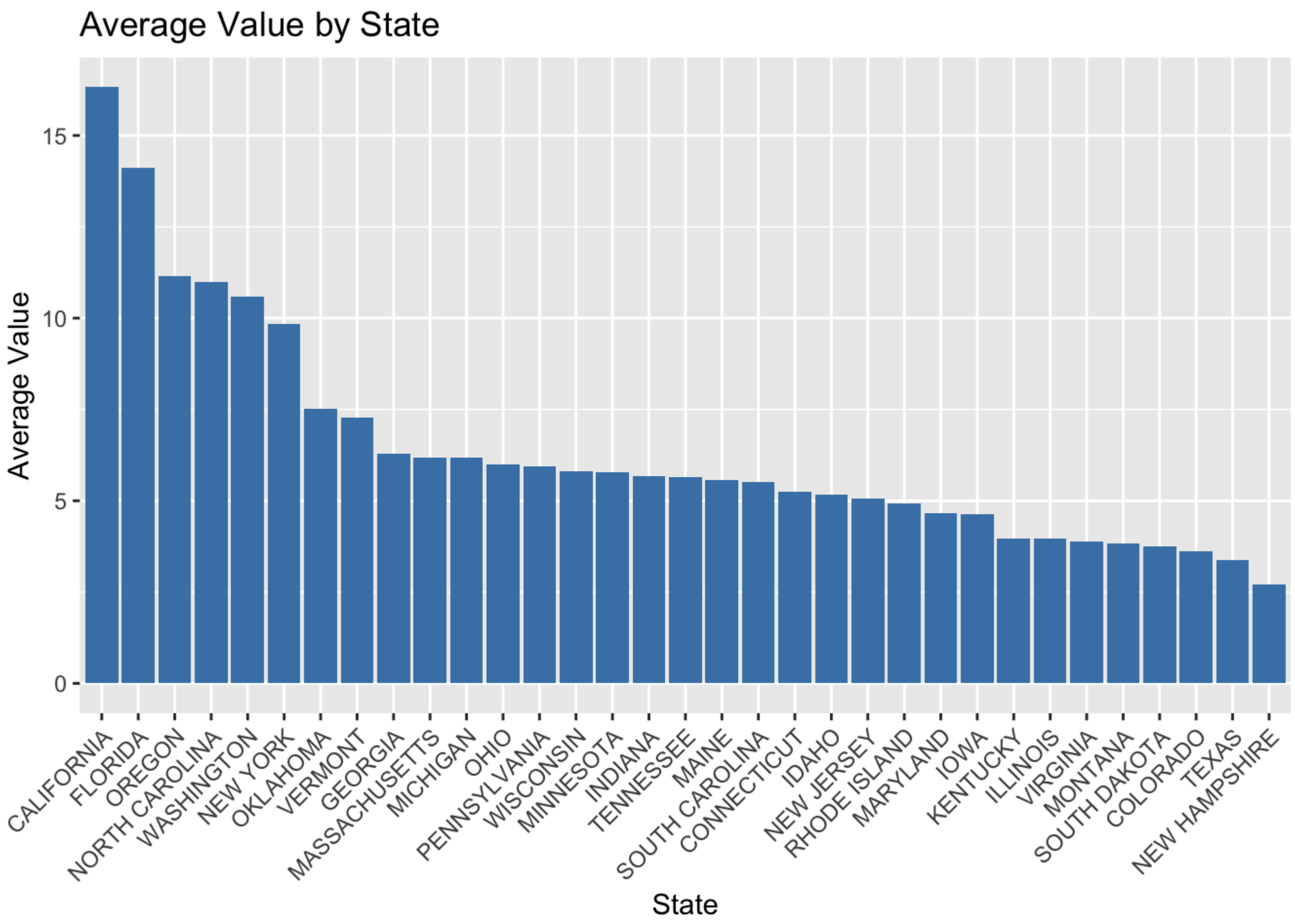


From here, we can see that for different chemical, there may also have different mean values, and ORGANIC tend to have higher values then using chemicals ones. And chmical others also seems will have

know the weight value by state

```
state_statistics <- strawberry_weight %>%
  group_by(State) %>%
  summarise(
    Count = n(),
    Mean = mean(Value, na.rm = TRUE),
    Median = median(Value, na.rm = TRUE),
    Min = min(Value, na.rm = TRUE),
    Max = max(Value, na.rm = TRUE),
    SD = sd(Value, na.rm = TRUE)
  ) %>%
  arrange(-Mean)

ggplot(state_statistics, aes(x = reorder(State, -Mean), y = log(Mean))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Average Value by State", x = "State", y = "Average Value")
```



```
state_statistics
```

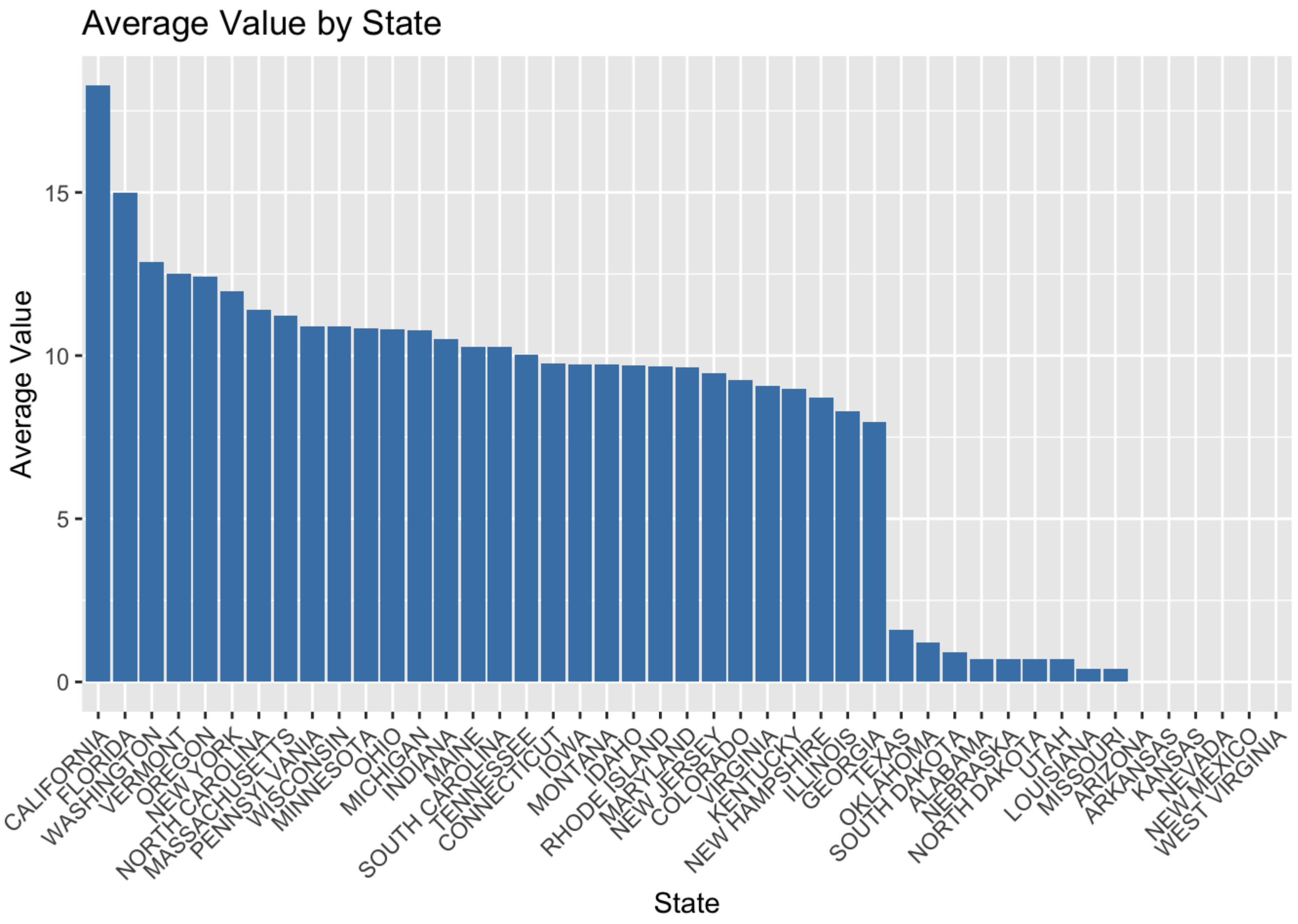
```
## # A tibble: 33 × 7
##    State          Count      Mean   Median   Min      Max        SD
##    <chr>          <int>     <dbl>    <dbl> <dbl>    <dbl>     <dbl>
##  1 CALIFORNIA        49 12260892. 6075000      0 28938000 11512292.
##  2 FLORIDA           38  1340405. 1947500      0  3020000  1235848.
##  3 OREGON            30    69786.   33850      0   232800    79856.
##  4 NORTH CAROLINA    34    58786.    2850      0   149000    66280.
##  5 WASHINGTON        30    39832.   10550.    23   122200    46246.
##  6 NEW YORK          23    18785.    2260      0    50400    20182.
##  7 OKLAHOMA           3     1858     1858   1858     1858        0
##  8 VERMONT            9     1442     1202   1051     2073      478.
##  9 GEORGIA            9      536.     279     19     1309      591.
## 10 MASSACHUSETTS      9      484.     485    251      715      201.
## # ℹ 23 more rows
```

California has super large average value,so I decide to use log to have a clear picture. From here we can see that California is super large and other Florida is second large

know the sales value by state

```
state_statistics <- strawberry_sale %>%
  group_by(State) %>%
  summarise(
    Count = n(),
    Mean = mean(Value, na.rm = TRUE),
    Median = median(Value, na.rm = TRUE),
    Min = min(Value, na.rm = TRUE),
    Max = max(Value, na.rm = TRUE),
    SD = sd(Value, na.rm = TRUE)
  ) %>%
  arrange(-Mean)

ggplot(state_statistics, aes(x = reorder(State, -Mean), y = log(Mean))) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Average Value by State", x = "State", y = "Average Value")
```



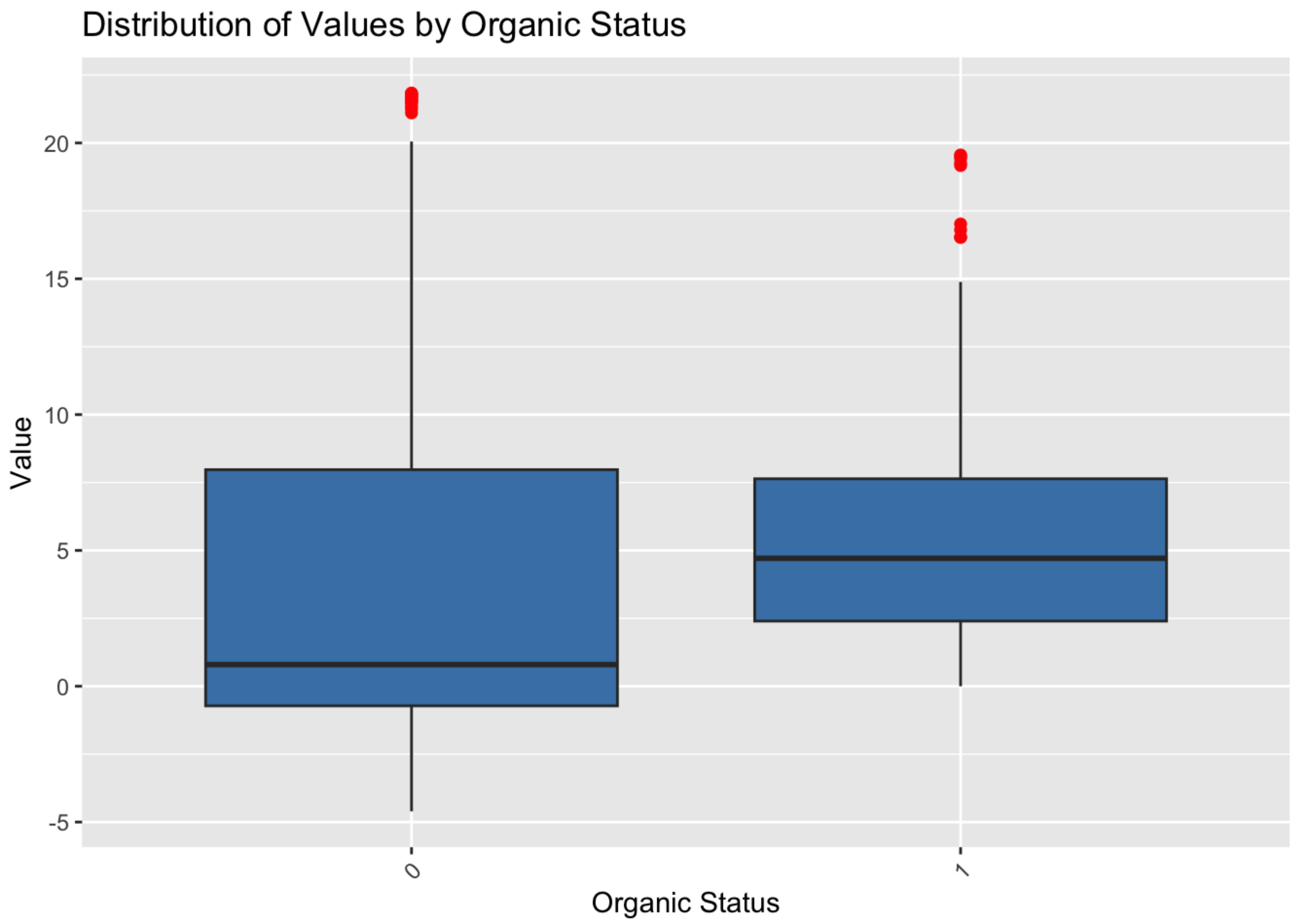
```
state_statistics
```

```
## # A tibble: 45 × 7
##    State          Count      Mean   Median   Min       Max         SD
##    <chr>          <int>     <dbl>    <dbl> <dbl>     <dbl>      <dbl>
##  1 CALIFORNIA        16 85934462.    186.      7 311784980 127669073.
##  2 FLORIDA           10  3256729.     12       2  15055709   6265575.
##  3 WASHINGTON        14   385898.     52.5     4   2917197    807769.
##  4 VERMONT           12   273026. 240170      26    670886    292502.
##  5 OREGON             9   250933.     25       6   1752592    587427.
##  6 NEW YORK          11   157876      36       2    644155    274300.
##  7 NORTH CAROLINA    10    89182      10       4    358487    146352.
##  8 MASSACHUSETTS     12    73982   47420.     12    204896     84108.
##  9 PENNSYLVANIA      12    54257.  43521      13    148898     60457.
## 10 WISCONSIN         12    53362   25075      22    141852     63192.
## # ℹ 35 more rows
```

As we compare sales and weight, California and Florida is similar in their position, but there are a lot of same height in sales have less height in weight. I assume it may be caused by these states sales more non-organic straws than other sates.

```
# Box plot of values distribution by organic status
ggplot(strawberry, aes(x = as.factor(Organic_Status), y = log(Value))) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2, fill = "stee
lblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of Values by Organic Status", x = "Organic Status", y = "Va
lue") +
  scale_y_continuous(labels = scales::comma)
```

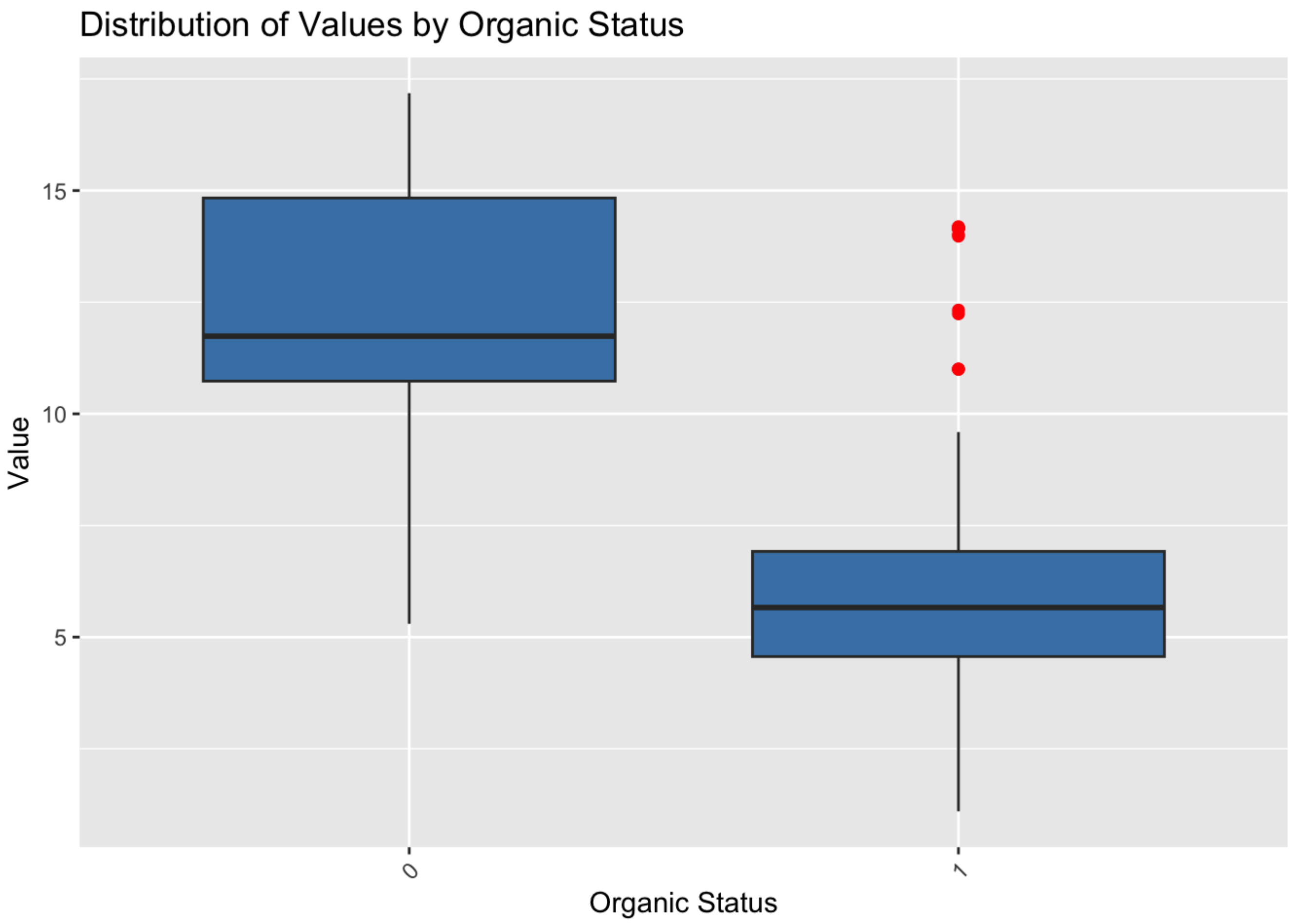
```
## Warning: Removed 42 rows containing non-finite values (`stat_boxplot()`).
```



The organic tends to have higher value than non-Organic.

```
ggplot(strawberry_weight , aes(x = as.factor(Organic_Status), y = log(Value))) +
  geom_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2, fill = "stee
lblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of Values by Organic Status", x = "Organic Status", y = "Va
lue") +
  scale_y_continuous(labels = scales::comma)
```

```
## Warning: Removed 23 rows containing non-finite values (`stat_boxplot()`).
```



The weight of non_organic is far higher than organic, which is accord with common sense that organic is expensice and less.

Reference:

NASS help (https://quickstats.nass.usda.gov/tutorials)

Quick Stats Glossary (https://quickstats.nass.usda.gov/src/glossary.pdf)

Quick Stats Column Definitions (https://quickstats.nass.usda.gov/param_define)

stats by subject (https://www.nass.usda.gov/Statistics_by_Subject/index.php?sector=CROPS)

Databases for Chemical Information (http://npic.orst.edu/ingred/cheminfo.html)

Pesticide Active Ingredients (http://npic.orst.edu/ingred/active.html)

TSCA Chemical Substance Inventory (https://www.epa.gov/tsca-inventory)

glyphosate (https://ordspub.epa.gov/ords/pesticides/f?p=CHEMICALSEARCH:3::::1,3,31,7,12,25:P3_XCHEMICAL_ID:2478)