615strawberry

Yishun Zhang

2024-10-23

Load and Explore the Data Start by loading the dataset and conducting an initial exploration.

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

```
##
## 载入程序包：'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyr)

# Load the dataset
strawberries <- read.csv("C:/Users/17756/Documents/WeChat Files/strawberries25_v3.csv")

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

```
## 'data.frame':    12669 obs. of  21 variables:
##  $ Program         : chr  "CENSUS" "CENSUS" "CENSUS" "CENSUS" ...
##  $ Year            : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ...
##  $ Period          : chr  "YEAR" "YEAR" "YEAR" "YEAR" ...
##  $ Week.Ending     : logi  NA NA NA NA NA NA ...
##  $ Geo.Level       : chr  "COUNTY" "COUNTY" "COUNTY" "COUNTY" ...
##  $ State           : chr  "ALABAMA" "ALABAMA" "ALABAMA" "ALABAMA" ...
##  $ State.ANSI      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Ag.District     : chr  "BLACK BELT" "BLACK BELT" "BLACK BELT" "BLACK BELT" ...
##  $ Ag.District.Code: int  40 40 40 40 40 40 40 40 40 40 ...
##  $ County          : chr  "BULLOCK" "BULLOCK" "BULLOCK" "BULLOCK" ...
##  $ County.ANSI     : int  11 11 11 11 11 11 101 101 101 101 ...
##  $ Zip.Code        : logi  NA NA NA NA NA NA ...
##  $ Region          : logi  NA NA NA NA NA NA ...
##  $ watershed_code  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Watershed       : logi  NA NA NA NA NA NA ...
##  $ Commodity       : chr  "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" ...
##  $ Data.Item       : chr  "STRAWBERRIES - ACRES BEARING" "STRAWBERRIES - ACRES GROWN" "STRAWBERRIES - ACRES NOT BEARING" "STRAWBERRIES - OPERATIONS WITH AREA BEARING" ...
##  $ Domain          : chr  "TOTAL" "TOTAL" "TOTAL" "TOTAL" ...
##  $ Domain.Category : chr  "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" ...
##  $ Value           : chr  " (D)" " 1" " (D)" " 1" ...
##  $ CV....          : chr  " (D)" " 15.7" " (D)" " (D)" ...
```

```
# Check the first few rows to understand the data
head(strawberries)
```

Program <chr>	Year <int>	Period <chr>	Week.Ending <lgl>	Geo.Level <chr>	State <chr>	State.ANSI <int>	Ag.District <chr>	Ag.District.Code <int>
1 CENSUS	2022	YEAR	*NA*	COUNTY	ALABAMA	1	BLACK BELT	40
2 CENSUS	2022	YEAR	*NA*	COUNTY	ALABAMA	1	BLACK BELT	40
3 CENSUS	2022	YEAR	*NA*	COUNTY	ALABAMA	1	BLACK BELT	40
4 CENSUS	2022	YEAR	*NA*	COUNTY	ALABAMA	1	BLACK BELT	40
5 CENSUS	2022	YEAR	*NA*	COUNTY	ALABAMA	1	BLACK BELT	40
6 CENSUS	2022	YEAR	*NA*	COUNTY	ALABAMA	1	BLACK BELT	40

6 rows | 1-10 of 22 columns

```
# Summary statistics of the dataset
summary(strawberries)
```

```
##     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       Min.   : 1.00   Length:12669
##  Class :character   Class :character   1st Qu.: 9.00   Class :character
##  Mode  :character   Mode  :character   Median :21.00   Mode  :character
##                                        Mean   :24.43
##                                        3rd Qu.:39.00
##                                        Max.   :56.00
##                                        NA's   :264
##  Ag.District.Code     County           County.ANSI     Zip.Code
##  Min.   : 10.00   Length:12669       Min.   :  1.00   Mode:logical
##  1st Qu.:20.00   Class :character   1st Qu.: 29.00   NA's:12669
##  Median :50.00   Mode  :character   Median : 69.00
##  Mean   :46.18                      Mean   : 83.82
##  3rd Qu.:62.00                      3rd Qu.:119.00
##  Max.   :96.00                      Max.   :810.00
##  NA's   :5359                       NA's   :5385
##    Region       watershed_code Watershed      Commodity
##  Mode:logical   Min.   :0      Mode:logical   Length:12669
##  NA's:12669     1st Qu.:0      NA's:12669     Class :character
##                 Median :0                     Mode  :character
##                 Mean   :0
##                 3rd Qu.:0
##                 Max.   :0
##
##   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
##
##
##
##
```

1.Data Cleaning Code

```
# Load necessary libraries
library(dplyr)
library(tidyr)

# Remove columns that are entirely empty
strawberries_clean <- strawberries %>%
  select(-c('Week.Ending', 'Zip.Code', 'Region', 'Watershed'))

# Convert non-numeric values in "Value" and "CV (%)" columns to NA
strawberries_clean <- strawberries_clean %>%
  mutate(Value = as.numeric(replace(Value, grepl("[A-Za-z]", Value), NA)),
         CV.... = as.numeric(replace(CV...., grepl("[A-Za-z]", CV....), NA)))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `Value = as.numeric(replace(Value, grepl("[A-Za-z]", Value),
##   NA))`.
## Caused by warning:
## ! 强制改变过程中产生了NA
```

```
# Check the data structure and missing values
summary(strawberries_clean)
```

```
##     Program               Year         Period          Geo.Level
##  Length:12669       Min.   :2018   Length:12669       Length:12669
##  Class :character   1st Qu.:2021   Class :character   Class :character
##  Mode  :character   Median :2022   Mode  :character   Mode  :character
##                     Mean   :2021
##                     3rd Qu.:2022
##                     Max.   :2024
##
##    State            State.ANSI     Ag.District        Ag.District.Code
##  Length:12669       Min.   : 1.00   Length:12669       Min.   : 10.00
##  Class :character   1st Qu.: 9.00   Class :character   1st Qu.:20.00
##  Mode  :character   Median :21.00   Mode  :character   Median :50.00
##                     Mean   :24.43                      Mean   :46.18
##                     3rd Qu.:39.00                      3rd Qu.:62.00
##                     Max.   :56.00                      Max.   :96.00
##                     NA's   :264                        NA's   :5359
##     County          County.ANSI    watershed_code  Commodity
##  Length:12669       Min.   :  1.00   Min.   :0      Length:12669
##  Class :character   1st Qu.: 29.00   1st Qu.:0      Class :character
##  Mode  :character   Median : 69.00   Median :0      Mode  :character
##                     Mean   : 83.82   Mean   :0
##                     3rd Qu.:119.00   3rd Qu.:0
##                     Max.   :810.00   Max.   :0
##                     NA's   :5385
##   Data.Item          Domain          Domain.Category       Value
##  Length:12669       Length:12669       Length:12669       Min.   :    0.00
##  Class :character   Class :character   Class :character   1st Qu.:    1.50
##  Mode  :character   Mode  :character   Mode  :character   Median :    6.00
##                                                           Mean   :   29.91
##                                                           3rd Qu.:   12.00
##                                                           Max.   :1963.00
##                                                           NA's   :5449
##      CV....
##  Min.   :  0.60
##  1st Qu.: 29.50
##  Median : 41.60
##  Mean   : 43.43
##  3rd Qu.: 56.10
##  Max.   : 99.90
##  NA's   :7934
```

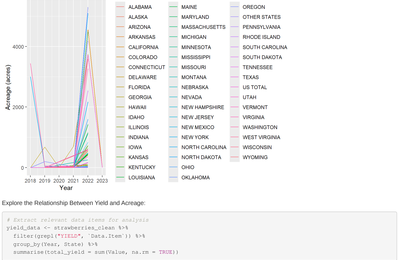
2.Exploratory Data Analysis (EDA) *Calculate Strawberry Acreage: Calculate the* total strawberry acreage by year and state. *Visualize* Changes in *Strawberry Acreage Over Time:*

```
# Calculate total acreage
acreage_data <- strawberries_clean %>%
  filter(grepl("ACRES", `Data.Item`)) %>%
  group_by(Year, State) %>%
  summarise(total_acres = sum(Value, na.rm = TRUE))
```

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

```
# Visualize acreage over the years by state
library(ggplot2)

ggplot(acreage_data , aes(x = Year, y = total_acres, color = State)) +
  geom_line() +
  labs(title = "Change in Strawberry Acreage Over Time",
       x = "Year",
       y = "Acreage (acres)")
```



Explore the Relationship Between Yield and Acreage:

```
# Extract relevant data items for analysis
yield_data <- strawberries_clean %>%
  filter(grepl("YIELD", `Data.Item`)) %>%
  group_by(Year, State) %>%
  summarise(total_yield = sum(Value, na.rm = TRUE))
```

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

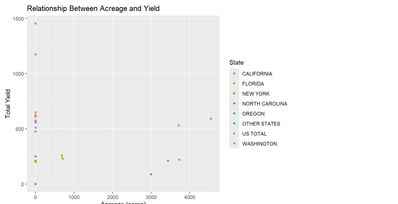
```
merged_data <- merge(acreage_data, yield_data, by = c("Year", "State"))

head(merged_data)
```

	Year <dbl>	State <chr>	total_acres <dbl>	total_yield <dbl>
1	2018	CALIFORNIA	0	1455.0
2	2018	FLORIDA	0	475.0
3	2018	NEW YORK	3000	88.0
4	2018	NORTH CAROLINA	0	250.0
5	2018	OREGON	0	200.0
6	2018	US TOTAL	0	1173.3

6 rows

```
# Visualize the relationship between acreage and yield
ggplot(merged_data, aes(x = total_acres, y = total_yield, color = State)) +
  geom_point() +
  labs(title = "Relationship Between Acreage and Yield",
       x = "Acreage (acres)",
       y = "Total Yield")
```



3.Interpretation of Results Changes in Strawberry Acreage Over Time: The visualization shows that strawberry acreage may vary significantly across states over the years, with some states having larger acreages than others. Relationship Between Acreage and Yield: There appears to be a positive correlation between yield and acreage, but further analysis is needed to understand the impact of other factors (e.g., climate, inputs) on yield.