Strawberry

Ruihang Han

2024-10-02

```
library(readr)

library(dplyr)
```

```
##
##     '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)

straw<-read.csv('strawberries25_v3.csv')

# Define a function to remove columns with only one unique value
remove_single_value_columns <- function(dataframe) {
  dataframe %>% select_if(~ n_distinct(.) > 1)
}

# Apply the function to clean the dataframe
straw_clean <- remove_single_value_columns(straw)

# Group by 'State' and count the number of records for each group
state_counts <- straw_clean %>%
  group_by(State) %>%
  tally()

# Check if the sum of group counts equals the number of rows in the cleaned dataframe
nrow(state_counts)
```

```
## [1] 52
```

```r
sum(state_counts$n) == nrow(straw_clean)
```

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

```r
# Summarize the data by state
state_summary <- straw_clean %>%
  group_by(State) %>%
  summarize(total_records = n())

# Print the summary
print(state_summary)
```

```
## # A tibble: 52 x 2
##    State       total_records
##    <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
```

```r
# Filter for California CENSUS data and select specific columns
california_census <- straw_clean %>%
  filter(State == "CALIFORNIA", Program == "CENSUS") %>%
  select(Year, `Data.Item`, Value)

head(california_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)
```

```r
# Filter for California SURVEY data and select specific columns
california_survey <- straw_clean %>%
  filter(State == "CALIFORNIA", Program == "SURVEY") %>%
  select(Year, Period, `Data.Item`, Value)

# Define a function to process 'Data.Item' strings and extract relevant information
parse_data_item <- function(text) {
  text <- as.character(text)
  text <- gsub("[---]", "-", text)  # Replace all types of dashes with a standard dash
```

```r
  segments <- strsplit(text, " - ")[[1]]  # Split the string by " - "

  fruit <- "Strawberries"  # Set the default fruit name

  # Case 1: If the string splits into 2 parts
  if (length(segments) == 2) {
    category <- str_remove(segments[1], "^STRAWBERRIES,?\\s*") %>% trimws()  # Clean the category
    details <- strsplit(segments[2], ",")[[1]]  # Split the details by ","
    item <- trimws(details[1])  # Extract item
    metric <- ifelse(length(details) > 1, trimws(details[2]), "N/A")  # Extract metric if available
  # Case 2: If the string splits into 3 parts
  } else if (length(segments) == 3) {
    category <- str_remove(segments[2], "^STRAWBERRIES,?\\s*") %>% trimws()
    details <- strsplit(segments[3], ",")[[1]]
    item <- trimws(details[1])
    metric <- ifelse(length(details) > 1, trimws(details[2]), "N/A")
  # Case 3: Default case when only 1 part exists
  } else {
    category <- str_remove(segments[1], "^STRAWBERRIES,?\\s*") %>% trimws()
    item <- "N/A"
    metric <- "N/A"
  }

  # Return the parsed information as a list
  list(Fruit = fruit, Category = category, Item = item, Metric = metric)
}

# Apply 'parse_data_item' function to each row in 'Data.Item' and combine the results
straw_clean <- bind_cols(straw_clean, do.call(rbind, lapply(straw_clean$`Data.Item`, parse_data_item)) %

# Group by 'Domain.Category' and count occurrences
domain_category_counts <- straw_clean %>%
  group_by(Domain.Category) %>%
  tally()

nrow(domain_category_counts)
```

```
## [1] 191
```

```r
# Split 'Domain.Category' column into 'use' and 'details' columns
straw_clean <- straw_clean %>%
  separate(col = `Domain.Category`, into = c("use", "details"), sep = ": ", extra = "drop", fill = "rig
  mutate(
    name = str_extract(details, "(?<=\\().*?(?=\\=)"),  # Extract the name part from the details
    code = str_extract(details, "(?<=\\= ).*?(?=\\))")  # Extract the code part from the details
  )

# Clean up the 'use' column by removing "CHEMICAL, " prefix
straw_clean$use <- str_remove(straw_clean$use, "^CHEMICAL, ")

# Convert 'Value' and 'CV....' columns to numeric
straw_clean$Value <- as.numeric(straw_clean$Value)
```

```
## Warning:      NA
```

```
straw_clean$CV.... <- as.numeric(straw_clean$CV....)
```

```
## Warning:      NA
```

```
# Remove the 'Data.Item' column as it's no longer needed
straw_clean <- straw_clean %>%
  select(-`Data.Item`)

# Display the cleaned dataframe
head(straw_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 Domain           use details Value CV....        Fruit
## 1 BULLOCK          11  TOTAL NOT SPECIFIED    <NA>    NA     NA Strawberries
## 2 BULLOCK          11  TOTAL NOT SPECIFIED    <NA>     3   15.7 Strawberries
## 3 BULLOCK          11  TOTAL NOT SPECIFIED    <NA>    NA     NA Strawberries
## 4 BULLOCK          11  TOTAL NOT SPECIFIED    <NA>     1     NA Strawberries
## 5 BULLOCK          11  TOTAL NOT SPECIFIED    <NA>     6   52.7 Strawberries
## 6 BULLOCK          11  TOTAL NOT SPECIFIED    <NA>     5   47.6 Strawberries
##   Category                            Item Metric name code
## 1                            ACRES BEARING    N/A <NA> <NA>
## 2                              ACRES GROWN    N/A <NA> <NA>
## 3                        ACRES NON-BEARING    N/A <NA> <NA>
## 4              OPERATIONS WITH AREA BEARING    N/A <NA> <NA>
## 5                OPERATIONS WITH AREA GROWN    N/A <NA> <NA>
## 6          OPERATIONS WITH AREA NON-BEARING    N/A <NA> <NA>
```