Strawberry Project Report

Yike Hu

2025-10-26

title: “Strawberry Project Report” author: “Yike Hu” date: “2025-10-26” format: pdf —

```
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   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- 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 becor
```

```
file_path <- "other chemicals with state.csv"
df <- read_csv(file_path, show_col_types = FALSE)

cat("Data loaded\n")
```

```
Data loaded
```

```
cat("Rows:", nrow(df), " Cols:", ncol(df), "\n\n")
```

```
Rows: 12969  Cols: 21
```

```r
miss_count <- sapply(df, function(x) sum(is.na(x)))
print(miss_count)
```

```
         Program             Year           Period      Week Ending
               0                0                0            12969
       Geo Level            State       State ANSI      Ag District
               0                0                0            12969
Ag District Code           County      County ANSI         Zip Code
           12969            12969            12969            12969
          Region   watershed_code        Watershed        Commodity
           12969                0            12969                0
       Data Item           Domain  Domain Category            Value
               0                0                0                0
          CV (%)
           12969
```

```r
miss_pct <- round(colMeans(is.na(df)) * 100, 2)
cat("\nMissing percent:\n")
```

```
Missing percent:
```

```r
print(miss_pct)
```

```
         Program             Year           Period      Week Ending
               0                0                0              100
       Geo Level            State       State ANSI      Ag District
               0                0                0              100
Ag District Code           County      County ANSI         Zip Code
             100              100              100              100
          Region   watershed_code        Watershed        Commodity
             100                0              100                0
       Data Item           Domain  Domain Category            Value
               0                0                0                0
          CV (%)
             100
```

```r
cat("\nTotal missing percent:", round(mean(is.na(df)) * 100, 2), "\n")
```

```
Total missing percent: 42.86
```

```
cat("\nTop missing columns:\n")
```

```
Top missing columns:
```

```
print(sort(miss_count, decreasing = TRUE)[1:5])
```

```
     Week Ending      Ag District Ag District Code           County
           12969             12969            12969            12969
     County ANSI
           12969
```

```
dup_rows <- duplicated(df)
n_dup <- sum(dup_rows)
cat("\nDuplicate rows:", n_dup, "\n")
```

```
Duplicate rows: 0
```

```
if (n_dup > 0) {
  cat("\nPreview duplicates:\n")
  print(head(df[dup_rows, ]))
  df <- df[!dup_rows, ]
  cat("\nAfter removing duplicates:", nrow(df), "\n")
}

write_csv(df, "other_chemicals_clean_basic.csv")
cat("\nSaved cleaned data\n")
```

```
Saved cleaned data
```

The *Other Chemicals* dataset contains 12,969 observations and 21 variables. An initial quality assessment was performed to identify missing values and duplicate records. Approximately 42.9% of all entries were missing, mainly from location-related columns such as *County*, *Zip*

Code, *Ag District*, and *Week Ending*, which were entirely empty. These variables were likely excluded in the original USDA data due to aggregation at a higher geographic level.

In contrast, key analytical variables — including *Program*, *Year*, *Commodity*, *Data Item*, *Domain*, and *Value* — were complete with no missing observations. This indicates that the core measurement data are reliable and can be safely used for further analysis.

No duplicate rows were detected, suggesting that the dataset is internally consistent and free from redundant entries. A cleaned version of the dataset was created by removing unnecessary columns and saving the result as **"other_chemicals_clean_basic.csv"** for subsequent analysis.