Overall

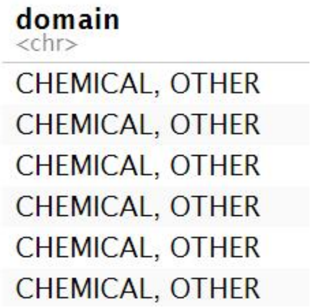
This project analyzes USDA strawberry data focusing on Other Chemicals used across different U.S. states.
To obtain relevant data, our group accessed the USDA National Agricultural Statistics Service (NASS) database under the Environmental section. Within this category, we selected Crops → Fruits → Strawberries, and specifically filtered for data related to chemicals.
We also selected the "State" geographic level to download datasets that include state identifiers.
The purpose of this analysis is to identify regional differences in chemical application, explore the most common "Other Chemicals," and establish a reproducible workflow in R that clearly documents each step.

Data Cleaning and Organizing

The raw dataset contained several columns such as Program, Year, State, Commodity, Data Item, Domain, Domain Category, and Value.
Among these, two columns — Domain and Domain Category — contained combined text information, such as:

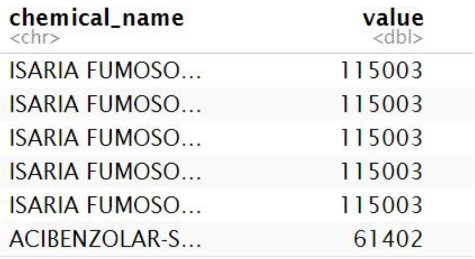
```
Domain                Domain Category
CHEMICAL, FUNGICIDE   CHEMICAL, FUNGICIDE: (OXATHIAPIPROLIN = 128111)
CHEMICAL, INSECTICIDE CHEMICAL, INSECTICIDE: (CYCLANILIPROLE = 26202)
CHEMICAL, INSECTICIDE CHEMICAL, INSECTICIDE: (PERMETHRIN = 109701)
```

The datase talso contained multiple columns with mixed and complex information. In particular, the Domain and Domain Category columns combined the chemical type, chemical name, and numeric value into a single text string, which made it impossible to perform numeric calculations or grouping directly. To make the dataset suitable for analysis, we first filtered the data to include only "Other Chemicals," removing records of fungicides and insecticides that were not relevant to the study.



domain
<chr>
CHEMICAL, OTHER
CHEMICAL, OTHER
CHEMICAL, OTHER
CHEMICAL, OTHER
CHEMICAL, OTHER
CHEMICAL, OTHER

We then extracted the chemical name and chemical value from the text and stored them in separate columns, converting the numeric values into proper number format. Missing values were removed, and all text entries, including chemical names and state names, were standardized to ensure consistency. Redundant columns, such as the original Domain, Domain Category, and Value, were dropped to simplify the dataset.



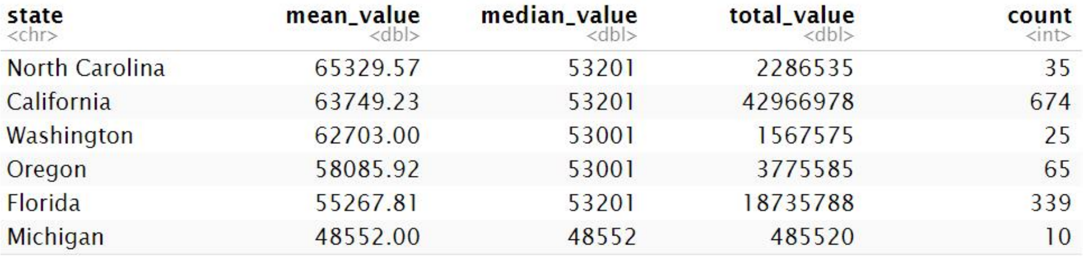
chemical_name <chr>	value <dbl>
ISARIA FUMOSO...	115003
ISARIA FUMOSO...	115003
ISARIA FUMOSO...	115003
ISARIA FUMOSO...	115003
ISARIA FUMOSO...	115003
ACIBENZOLAR-S...	61402

Each of these steps ensured that the resulting data were tidy, structured, and suitable for statistical analysis, allowing us to accurately calculate totals, averages, and make meaningful comparisons across states. This process also made the workflow reproducible and transparent, as every transformation was carefully documented and could be verified or repeated.

data_item	chemical_name	chemical _value
STRAWBERRIES - APPLICATIONS, MEASURED IN LB	ISARIA FUMOSOROSEA STRAIN FE 9901	115003
STRAWBERRIES - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG	ISARIA FUMOSOROSEA STRAIN FE 9901	115003
STRAWBERRIES - APPLICATIONS, MEASURED IN LB / ACRE / YEAR, AVG	ISARIA FUMOSOROSEA STRAIN FE 9901	115003
STRAWBERRIES - APPLICATIONS, MEASURED IN NUMBER, AVG	ISARIA FUMOSOROSEA STRAIN FE 9901	115003

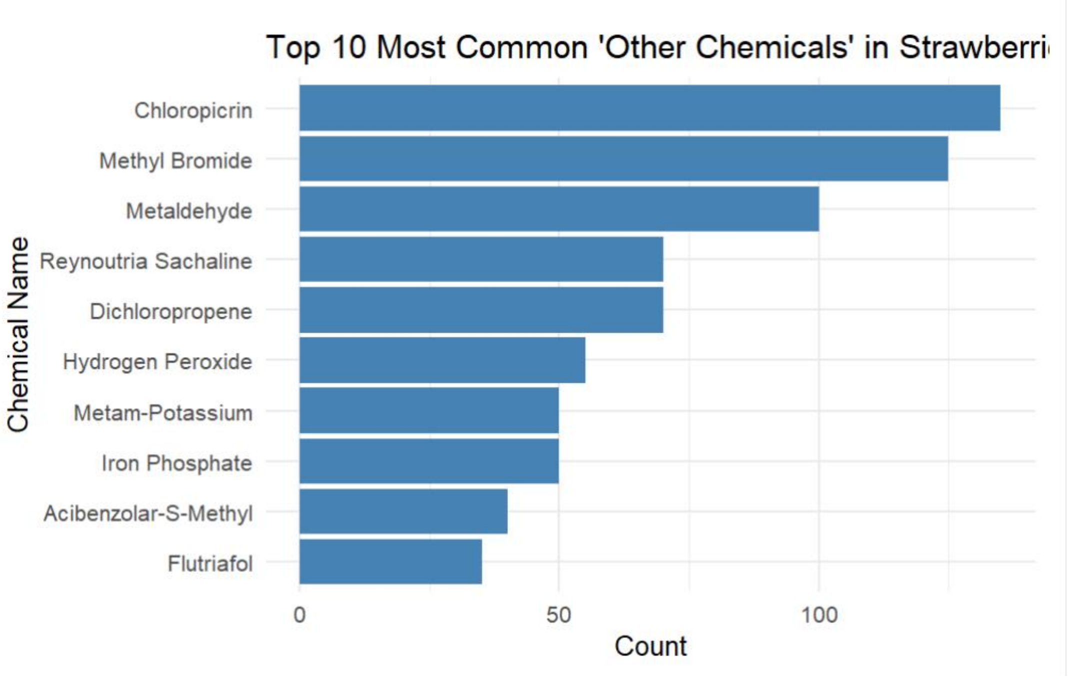
Exploratory Data Analysis

After cleaning and organizing the dataset, we conducted a thorough exploratory data analysis to understand patterns in the usage of “Other Chemicals” across different states. The first step was to examine the distribution of chemical values to identify any extreme values, missing data, or unusual patterns. We observed that some chemicals had very high usage in certain states, while others were rarely applied, which suggested regional preferences or differences in management practices.

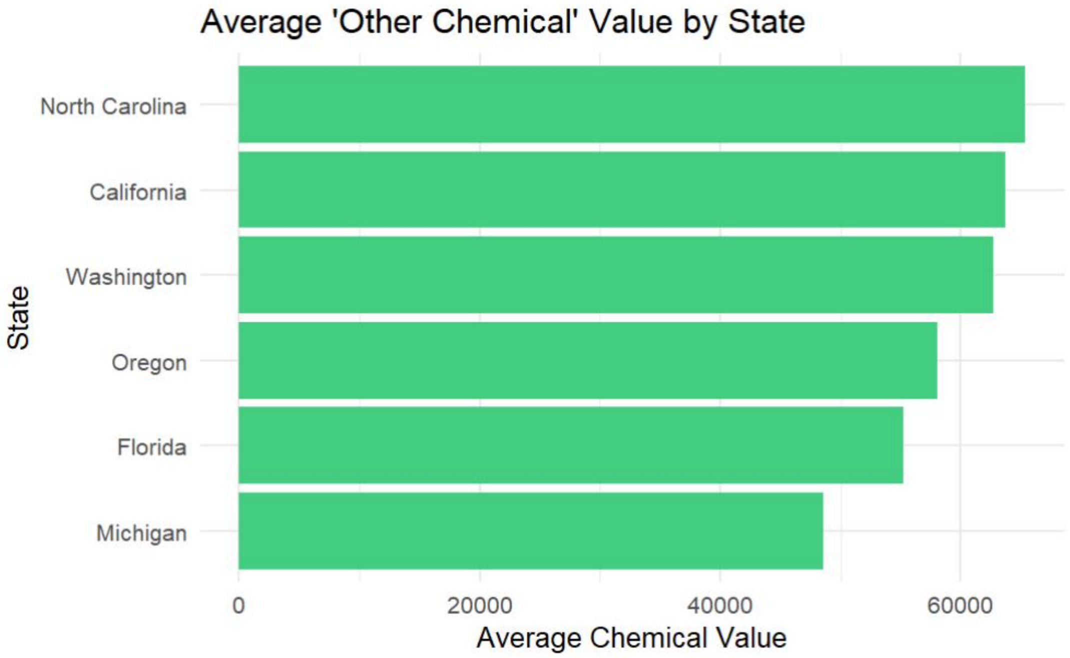


state <chr>	mean_value <dbl>	median_value <dbl>	total_value <dbl>	count <int>
North Carolina	65329.57	53201	2286535	35
California	63749.23	53201	42966978	674
Washington	62703.00	53001	1567575	25
Oregon	58085.92	53001	3775585	65
Florida	55267.81	53201	18735788	339
Michigan	48552.00	48552	485520	10

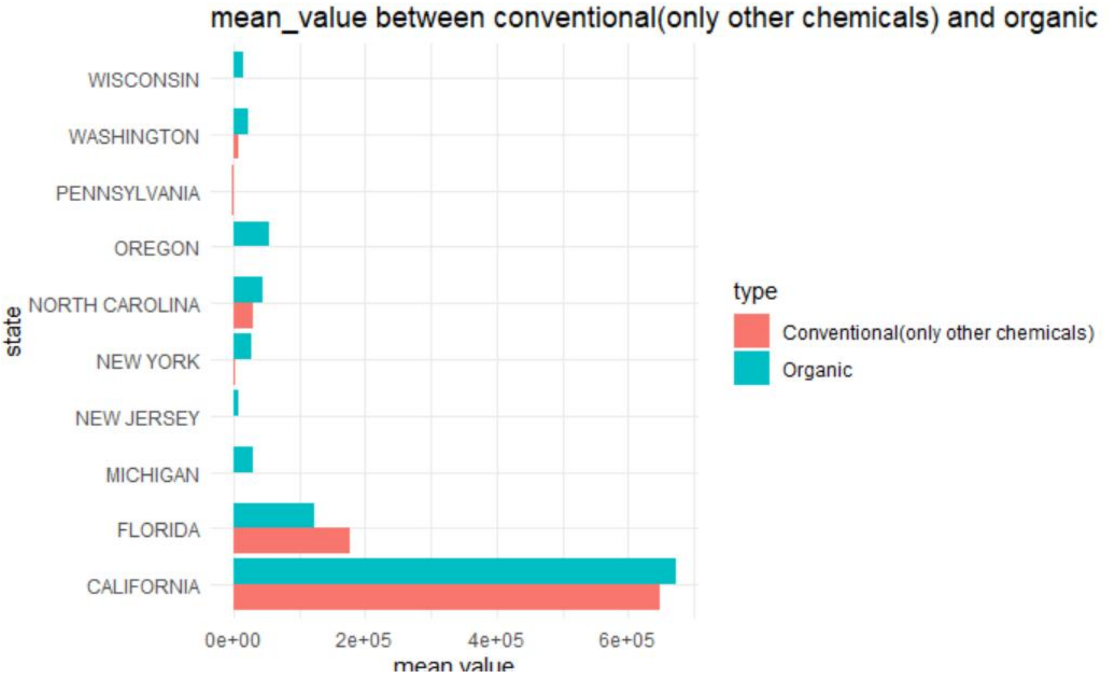
Next, we explored the frequency of each chemical to identify which products were most commonly used in strawberry cultivation. This revealed that a few chemicals dominated the usage, while many others appeared only occasionally. This led us to consider potential relationships between chemical type and state, as well as patterns of concentration versus diversity in chemical applications.

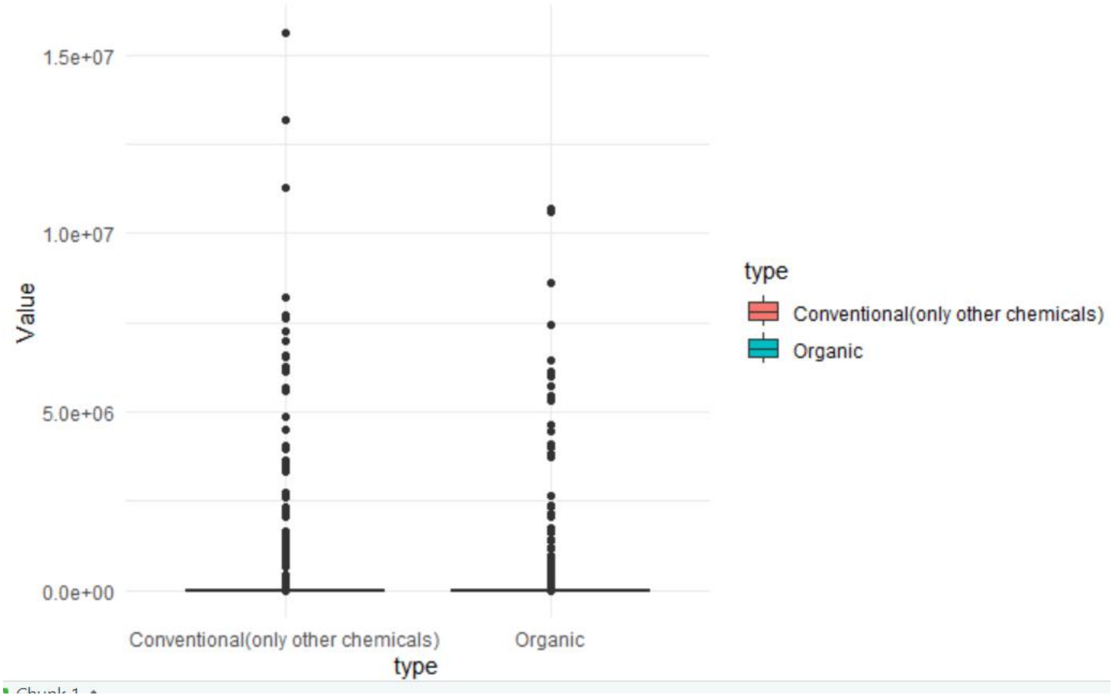


We also compared chemical usage across states by calculating summaries such as the average and total use for each state. This allowed us to see which states applied the most or least of “Other Chemicals,” providing insight into regional differences. During this process, we noticed interesting variations that prompted further questions: for example, why certain chemicals are heavily used in some states and not others, and whether certain states rely on a broader variety of chemicals or focus on just a few.

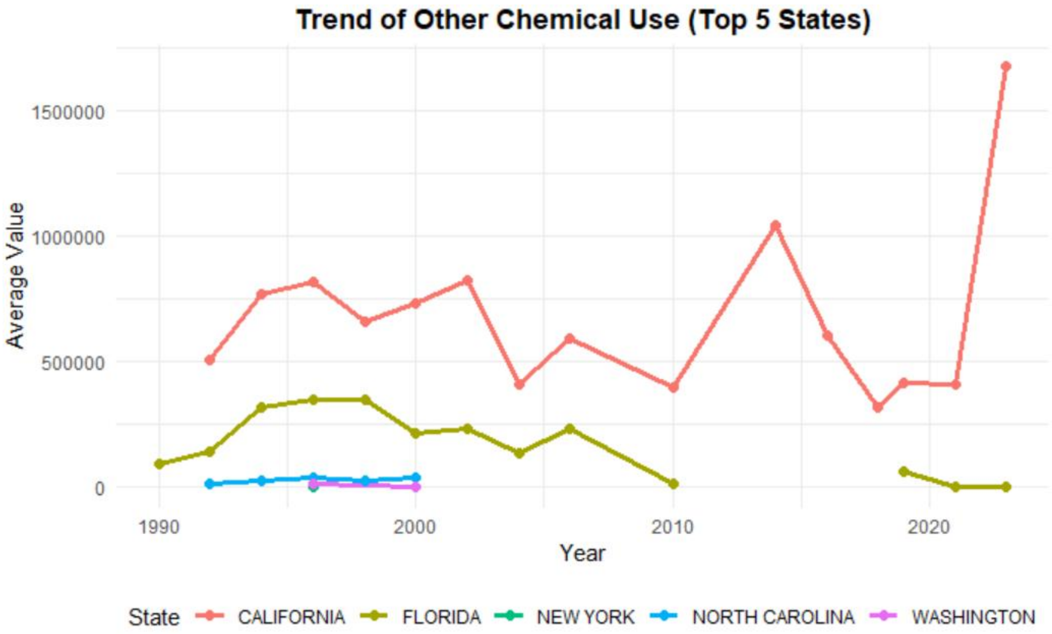


This code filters the dataset to include organic (fertilizer) and conventional (other chemical) records, classifies them into two groups, and then calculates the mean, standard deviation, and sample size by state. It visualizes the differences using a bar chart across states and a boxplot to compare overall distributions between organic and conventional practices.





This table identifies the top five states with the highest average usage of “other chemicals.” It then calculates the yearly average chemical use for these states and visualizes the trend over time using a line plot. The chart helps show which states use the most chemicals and how their usage patterns change across years.



Throughout the EDA process, we kept asking questions about the data, testing assumptions, and seeking patterns that might inform future statistical modeling. This exploratory mindset, driven by curiosity, allowed us to uncover insights that were not immediately obvious from the raw data.

Conclusion

This project showed how important teamwork is when analyzing real data. Our group worked together to clean and organize the USDA strawberry dataset. We discussed how to handle missing values, deciding which ones to remove or keep, and how to make the data consistent. We explored the raw data on the USDA website, carefully selecting the strawberries dataset under environmental data and focusing on the columns that were relevant to “Other Chemicals.”

We spent time understanding each column, what it represented, and how it could affect the analysis. We also figured out how to extract chemical names and values from text fields, and checked each other’s work to make sure it was correct.

Through this assignment, we learned not only technical skills like data cleaning and organizing, but also how to collaborate, discuss decisions as a group, and think carefully about the meaning of each variable. We also gained experience in exploring a dataset and noticing interesting patterns, which helped us understand strawberry chemical use and gave us confidence in handling complex, real-world data.