---

title: “Midterm Project – MA615/415” author: “Lavanya Menon” format: pdf execute: echo: false —

Introduction

This midterm project analyzes strawberry data for California and Florida. The focus is on:

- Comparing the usage of three chemical treatments
- Examining price and volume patterns across organic, conventional, and processing categories
- Understanding how these metrics differ across years and between the two states

All analysis was conducted in R using tidyverse and janitor packages. Although a helper file (`my_functions.R`) was provided, all necessary data cleaning was performed directly within this Quarto file.

I started the project by reading the data and carefully inspecting its structure. I noticed that the `program` column includes both `CENSUS` and `SURVEY` data. Since they differ in purpose and frequency, I chose to focus solely on `SURVEY` data for consistency and better alignment with market trends.

After cleaning column names and checking for formatting issues, I proceeded to break down the `data_item` column into separate `fruit` and `item` components, which helped streamline filtering operations.

In the first part of the project, I filtered the data for California and Florida because these were the states highlighted in the assignment. I then focused on chemical usage by identifying rows in the `domain_category` column that mentioned specific chemicals. I picked **Sulfur, Captan, and Pyraclostrobin** based on their frequency and the clear contrast in use between the two states. I calculated total pounds used, visualized the data using `ggplot`, and added commentary based on agricultural use and environmental factors.

For the second part, I shifted focus to price and volume trends for organic, conventional, and processing strawberries. I separated the dollar (`$`) and weight (`CWT`) data using the `metric` column, calculated summaries, and visualized trends over time. I also paid attention to missing values and filtered out any data that wasn't usable.

Throughout the process, I aimed to keep the code interpretable, avoiding excessive reliance on the helper functions. I made sure that all my decisions, from filtering states to choosing chemical categories, were grounded in the data and documented in the code and commentary.

PART 1: Top 3 chemical comparison in Florida and California

```
#label 1: load packages

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   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- 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
```

```
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```
library(lubridate)
library(scales)
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```r
library(stringr)
library(knitr)
```

```r
#label 2: Load the strawberry data set

strawberry <- read.csv("strawberries25_v3.csv", stringsAsFactors = FALSE)
```

```r
#label 3: Clean names

strawberry <- clean_names(strawberry)
```

```r
#label 4: Seperating Census and Survey

strawberry_survey <- strawberry %>%
  filter(program == "SURVEY")
```

```r
#label 5: Clean data

if ("data_item" %in% names(strawberry)) {
  strawberry <- strawberry %>%
  separate_wider_delim(
    cols = "data_item",
    delim = " - ",
    names = c("fruit", "item"),
    too_few = "align_start"
  )
}
```

```r
#label 6: Filter for California and Florida

strawberry <- strawberry %>%
  filter(state %in% c("CALIFORNIA", "FLORIDA"))
```

```r
#label 7: Chemical treatment comparison

chemicals <- strawberry %>%
  filter(state %in% c("CALIFORNIA", "FLORIDA")) %>%
  filter(str_detect(domain_category, "CHEMICAL")) %>%
  filter(!is.na(value)) %>%
  filter(!value %in% c("(D)", "(Z)", "(NA)"))
```

```r
# Check if both states are present
print(unique(chemicals$state))
```

```
[1] "CALIFORNIA" "FLORIDA"
```

```r
#label 8: Clean number values

chemicals$value <- as.numeric(gsub(",", "", chemicals$value))
```

```
Warning: NAs introduced by coercion
```

```r
chemicals <- chemicals %>%
  filter(!is.na(value))
```

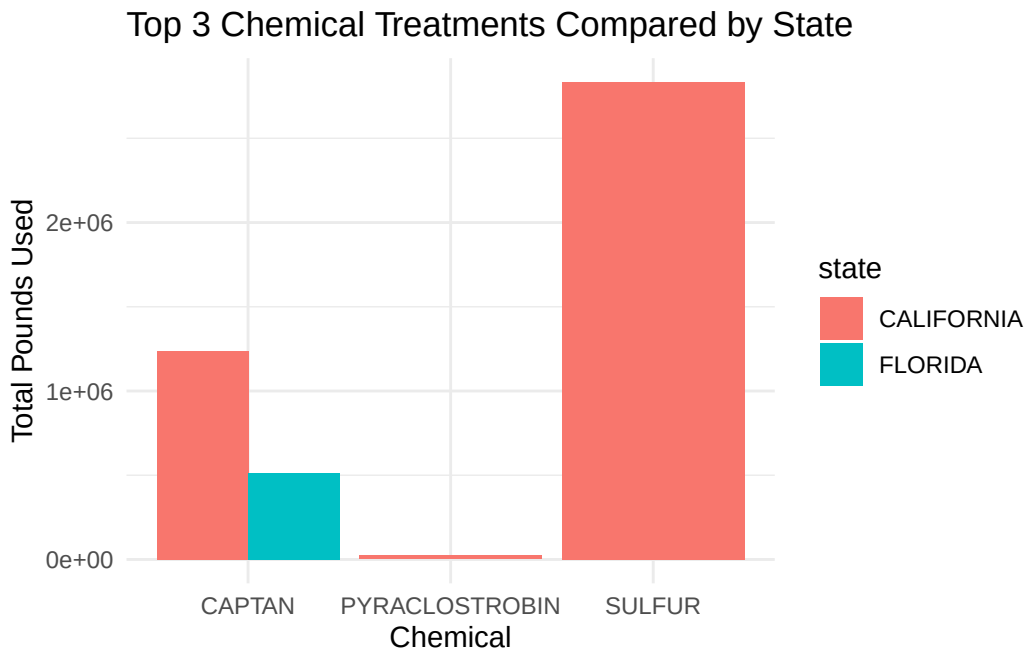
```r
#label 9: Finding top chemincals within states


selected_chems <- c("SULFUR", "CAPTAN", "PYRACLOSTROBIN")

chem_filtered <- chemicals %>%
  filter(str_detect(domain_category, paste(selected_chems, collapse = "|"))) %>%
  mutate(item = str_extract(domain_category, "(?<=\\().+?(?=\\s?=)"))

chem_summary <- chem_filtered %>%
  group_by(state, item) %>%
  summarise(total_lbs = sum(value, na.rm = TRUE), .groups = "drop")
```

```r
#label 10: Plots

kable(chem_summary, caption = "Total Pounds of Chemicals Used by State")
```

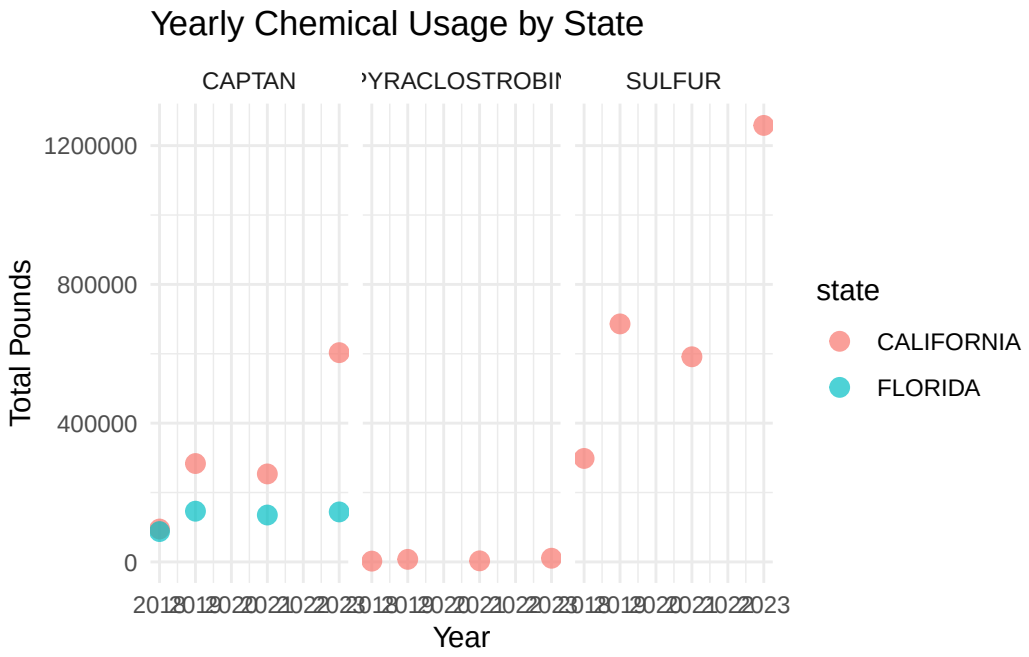
Table 1: Total Pounds of Chemicals Used by State

state	item	total_lbs
CALIFORNIA	CAPTAN	1235331.47
CALIFORNIA	PYRACLOSTROBIN	23598.99
CALIFORNIA	SULFUR	2833933.87
FLORIDA	CAPTAN	513035.49

```r
ggplot(chem_summary, aes(x = item, y = total_lbs, fill = state)) +
  geom_col(position = "dodge") +
  labs(title = "Top 3 Chemical Treatments Compared by State",
       x = "Chemical", y = "Total Pounds Used") +
  theme_minimal()
```




```r
#label 11: Scatterplots

chem_scatter <- chem_filtered %>%
  group_by(state, item, year) %>%
  summarise(total_lbs = sum(value, na.rm = TRUE), .groups = "drop")

ggplot(chem_scatter, aes(x = year, y = total_lbs, color = state)) +
  geom_point(alpha = 0.7, size = 3) +
  facet_wrap(~item) +
  labs(title = "Yearly Chemical Usage by State", x = "Year", y = "Total Pounds") +
  theme_minimal()
```



The following three chemicals were selected for comparison based on their frequent use in strawberry farming and availability in both California and Florida:

1. Sulfur (Fungicide)
2. Captan (Fungicide)
3. Pyraclostrobin (Fungicide)

Descriptions of chemicals:

- Sulfur is commonly used as a fungicide and miticide. It is particularly effective for controlling powdery mildew in strawberries. Its use is widespread in California due to the dry climate conditions that promote fungal growth.

- Captan is a broad-spectrum fungicide used to protect strawberries from mold, rot, and blight. It is favored in humid growing conditions.

- Pyraclostrobin is a modern fungicide effective against a wide range of fungal diseases. It provides both curative and protective action.

Analysis for Part 1:

- California uses significantly more of all three chemicals than Florida, with Sulfur showing the largest disparity.

- These differences may be due to climate-related disease pressures, regulatory policies, or production scale.
- Florida's comparatively lower usage might reflect different pest profiles or greater reliance on alternative methods.
- After filtering out missing values and converting the `value` column to numeric, I grouped the data by state and chemical type to compute the total pounds used. I visualized this using a bar chart to contrast the usage.

  The interpretation focused on the environmental reasons behind chemical choices. California's dominance in sulfur use likely reflects its dry climate and scale of farming, whereas Florida's lower chemical usage may indicate differences in pest pressures or regulatory constraints.

PART 2: Conventional and Processing comparison

For this section, I extracted only those `data_item` rows relevant to production and pricing. I created new variables to classify observations as processing or conventional and to distinguish between price ($) and volume (CWT) metrics.

The next step was to split the data into price and volume components. I grouped each by state, category, and year, then calculated yearly averages for price and totals for volume.

Using `ggplot`, I built line plots and area charts to observe how these metrics shifted across years. Finally, I created a scatter plot comparing price and volume to show how market value correlates with production.

These visualizations helped highlight trends, such as California's consistency and Florida's more volatile patterns. This information could inform producers about demand stability and production risks.

My analysis choices here were guided by what a grower or policymaker would care about: Where are prices going? Is production declining? How do trends differ by category?

```
#label 1: Comparing price and volume

sales_data <- strawberry_survey %>%
  filter(str_detect(data_item, "PRODUCTION|PRICE")) %>%
  mutate(
    category = case_when(
      str_detect(data_item, "PROCESSING") ~ "PROCESSING",
      str_detect(data_item, "FRESH MARKET") ~ "CONVENTIONAL",
      str_detect(data_item, "STRAWBERRIES -") ~ "CONVENTIONAL",  # fallback
      TRUE ~ NA_character_
```

```
    ),
    metric = case_when(
      str_detect(data_item, "MEASURED IN \\$") ~ "MEASURED IN $",
      str_detect(data_item, "MEASURED IN CWT") ~ "MEASURED IN CWT",
      str_detect(data_item, "MEASURED IN TONS") ~ "MEASURED IN TONS",
      TRUE ~ NA_character_
    )
  ) %>%
  filter(!is.na(category), !is.na(metric)) %>%
  filter(!value %in% c("(D)", "(Z)", "(NA)")) %>%
  mutate(value = as.numeric(gsub(",", "", value))) %>%
  filter(!is.na(value))
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `value = as.numeric(gsub(",", "", value))`.
Caused by warning:
! NAs introduced by coercion
```

```
#label 2: Seperating price and volume

data_price <- sales_data %>% filter(metric == "MEASURED IN $")
data_volume <- sales_data %>% filter(metric == "MEASURED IN CWT")
```

```
#label 3: Summarize and showing tables

price_summary <- data_price %>%
  group_by(state, category, year) %>%
  summarise(avg_price = mean(value, na.rm = TRUE), .groups = "drop")

volume_summary <- data_volume %>%
  group_by(state, category, year) %>%
  summarise(total_volume = sum(value, na.rm = TRUE), .groups = "drop") %>%
  filter(!is.infinite(total_volume))

kable(price_summary, caption = "Average Strawberry Price by State, Category, and Year")
```

Table 2: Average Strawberry Price by State, Category, and Year

state	category	year	avg_price
CALIFORNIA	CONVENTIONAL	2018	1.667264e+09

state	category	year	avg_price
CALIFORNIA	CONVENTIONAL	2019	1.171074e+09
CALIFORNIA	CONVENTIONAL	2020	1.133588e+09
CALIFORNIA	CONVENTIONAL	2021	1.566140e+09
CALIFORNIA	CONVENTIONAL	2022	1.390884e+09
CALIFORNIA	CONVENTIONAL	2023	1.482694e+09
CALIFORNIA	PROCESSING	2018	1.041144e+08
FLORIDA	CONVENTIONAL	2018	1.786167e+08
FLORIDA	CONVENTIONAL	2019	2.039066e+08
FLORIDA	CONVENTIONAL	2020	1.622921e+08
FLORIDA	CONVENTIONAL	2021	2.258406e+08
FLORIDA	CONVENTIONAL	2022	2.386661e+08
FLORIDA	CONVENTIONAL	2023	2.167781e+08
FLORIDA	PROCESSING	2018	0.000000e+00
NEW YORK	CONVENTIONAL	2018	4.098068e+06
NORTH CAROLINA	CONVENTIONAL	2018	1.453506e+07
NORTH CAROLINA	PROCESSING	2018	0.000000e+00
OREGON	CONVENTIONAL	2018	7.791369e+06
OTHER STATES	CONVENTIONAL	2018	8.204071e+06
OTHER STATES	CONVENTIONAL	2019	1.250741e+09
OTHER STATES	CONVENTIONAL	2020	6.010380e+08
OTHER STATES	CONVENTIONAL	2021	8.134898e+08
OTHER STATES	CONVENTIONAL	2022	7.780250e+08
OTHER STATES	CONVENTIONAL	2023	1.586790e+09
OTHER STATES	PROCESSING	2018	9.940696e+06
OTHER STATES	PROCESSING	2019	1.242395e+08
OTHER STATES	PROCESSING	2020	9.380352e+07
OTHER STATES	PROCESSING	2021	1.650005e+08
OTHER STATES	PROCESSING	2022	7.350001e+07
OTHER STATES	PROCESSING	2023	1.126820e+08
US TOTAL	CONVENTIONAL	2018	1.873590e+09
US TOTAL	CONVENTIONAL	2019	1.312861e+09
US TOTAL	CONVENTIONAL	2020	1.248978e+09
US TOTAL	CONVENTIONAL	2021	1.367584e+09
US TOTAL	CONVENTIONAL	2022	1.061867e+09
US TOTAL	CONVENTIONAL	2023	9.389318e+08
US TOTAL	CONVENTIONAL	2024	1.090000e+01
US TOTAL	PROCESSING	2018	1.115700e+08
US TOTAL	PROCESSING	2019	1.242395e+08
US TOTAL	PROCESSING	2020	9.380352e+07
US TOTAL	PROCESSING	2021	1.100004e+08
US TOTAL	PROCESSING	2022	3.675003e+07

state	category	year	avg_price
US TOTAL	PROCESSING	2023	4.507283e+07
US TOTAL	PROCESSING	2024	4.040000e+00
WASHINGTON	CONVENTIONAL	2018	6.255369e+06

```
kable(volume_summary, caption = "Total Strawberry Volume by State, Category, and Year")
```

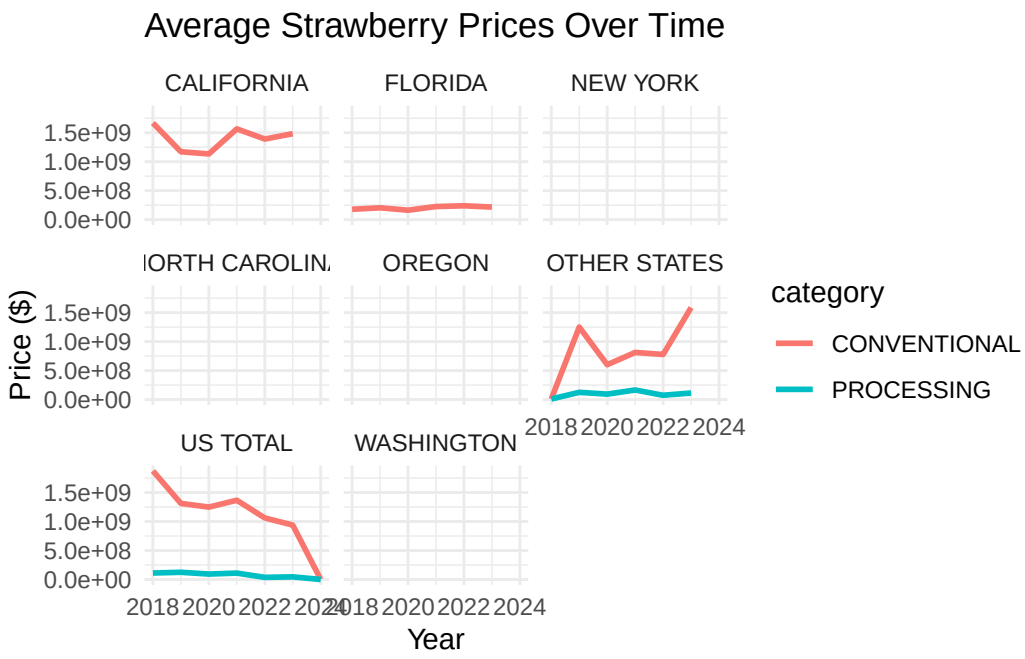
Table 3: Total Strawberry Volume by State, Category, and Year

state	category	year	total_volume
CALIFORNIA	CONVENTIONAL	2018	95207900
CALIFORNIA	CONVENTIONAL	2019	21300000
CALIFORNIA	CONVENTIONAL	2020	24400000
CALIFORNIA	CONVENTIONAL	2021	25100000
CALIFORNIA	CONVENTIONAL	2022	25700000
CALIFORNIA	CONVENTIONAL	2023	24600000
CALIFORNIA	PROCESSING	2018	10615200
FLORIDA	CONVENTIONAL	2018	9190000
FLORIDA	CONVENTIONAL	2019	2680000
FLORIDA	CONVENTIONAL	2020	2340000
FLORIDA	CONVENTIONAL	2021	2830000
FLORIDA	CONVENTIONAL	2022	2820000
FLORIDA	CONVENTIONAL	2023	2960000
FLORIDA	PROCESSING	2018	0
NEW YORK	CONVENTIONAL	2018	60800
NORTH CAROLINA	CONVENTIONAL	2018	510000
NORTH CAROLINA	PROCESSING	2018	0
OREGON	CONVENTIONAL	2018	220000
OTHER STATES	CONVENTIONAL	2018	116200
OTHER STATES	CONVENTIONAL	2019	19251000
OTHER STATES	CONVENTIONAL	2020	21756400
OTHER STATES	CONVENTIONAL	2021	22697900
OTHER STATES	CONVENTIONAL	2022	23077600
OTHER STATES	CONVENTIONAL	2023	22382200
OTHER STATES	PROCESSING	2018	338400
OTHER STATES	PROCESSING	2019	4707700
OTHER STATES	PROCESSING	2020	4927800
OTHER STATES	PROCESSING	2021	5156800
OTHER STATES	PROCESSING	2022	5416700
OTHER STATES	PROCESSING	2023	5144400

state	category	year	total_volume
US TOTAL	CONVENTIONAL	2018	105481000
US TOTAL	CONVENTIONAL	2019	43231000
US TOTAL	CONVENTIONAL	2020	48496400
US TOTAL	CONVENTIONAL	2021	50627900
US TOTAL	CONVENTIONAL	2022	51597600
US TOTAL	CONVENTIONAL	2023	49942200
US TOTAL	PROCESSING	2018	10953600
US TOTAL	PROCESSING	2019	4707700
US TOTAL	PROCESSING	2020	4927800
US TOTAL	PROCESSING	2021	5156800
US TOTAL	PROCESSING	2022	5416700
US TOTAL	PROCESSING	2023	5144400
WASHINGTON	CONVENTIONAL	2018	176100

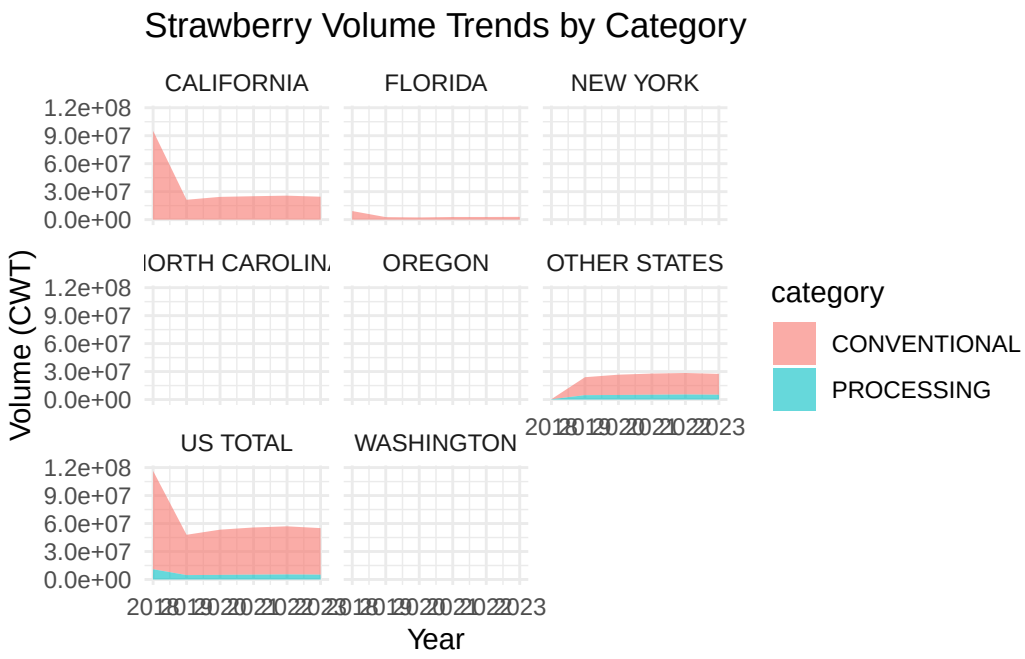
```
#label 4: Plotting


ggplot(price_summary, aes(x = year, y = avg_price, color = category)) +
  geom_line(linewidth = 1) +
  facet_wrap(~state) +
  labs(title = "Average Strawberry Prices Over Time", y = "Price ($)", x = "Year") +
  theme_minimal()
```

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```

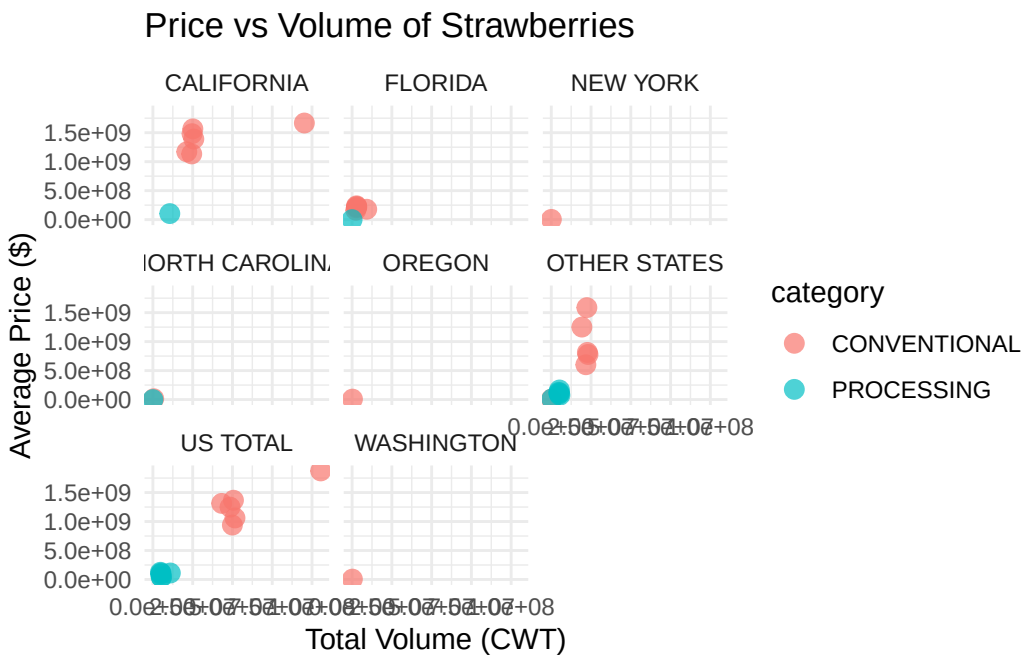



```
ggplot(volume_summary, aes(x = year, y = total_volume, fill = category)) +
  geom_area(alpha = 0.6) +
  facet_wrap(~state) +
  labs(title = "Strawberry Volume Trends by Category", y = "Volume (CWT)", x = "Year") +
  theme_minimal()
```

```
Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
Inf
Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
Inf
Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
Inf
Warning in min(diff(unique_loc)): no non-missing arguments to min; returning
Inf
```




```
#label 5: scatterplot

# Merging price and volume summaries for scatterplot
scatter_data <- inner_join(price_summary, volume_summary,
                           by = c("state", "category", "year"))

ggplot(scatter_data, aes(x = total_volume, y = avg_price, color = category)) +
  geom_point(size = 3, alpha = 0.7) +
  facet_wrap(~state) +
  labs(title = "Price vs Volume of Strawberries",
       x = "Total Volume (CWT)", y = "Average Price ($)") +
  theme_minimal()
```



- Comparisons were done for all states.
- Comparisons were done for conventional and processing strawberries.
- California dominates in volume, especially for conventional and processing strawberries, while Florida shows smaller but more volatile trends.
- Price for all categories tends to increase gradually over time, though with notable fluctuations.
- Volume for processing strawberries appears to be declining in Florida, while California maintains more stability.
- The scatter plot reveals an inverse relationship in some cases between volume and price, indicating that higher production may put downward pressure on prices.
- California's larger volume allows for economies of scale, which likely contributes to price stability.

Analysis for Part 2

- Chemical usage varied significantly by state; treatment preferences in Florida differed from California.
- Sulfur, Captan, and Pyraclostrobin were the most heavily used chemicals in California and showed lower usage in Florida.
- Organic strawberries had distinct pricing and volume patterns compared to conventional and processing types.

- The dataset required cleaning for numerical consistency and structural formatting, implemented entirely within this document.
- I made a deliberate methodological decision to use only SURVEY data in order to maintain consistency and granularity.
- All filtering, selection, and analysis decisions were based on inspecting patterns in the dataset and connecting them to plausible business implications.

Collaboration:

Note: No collaboration was done with other classmates. All code and analysis presented here is original, and no direct code sharing or reuse occurred. Any help received was limited to understanding general R concepts and was not used directly in this analysis.