MA415 Midterm Project - Isha Mukundan

USDA NASS Strawberry Data Set

This document contains the process and results for data cleaning and exploratory data analysis (EDA) OF A USDA NASS Data Set regarding the production and sales of strawberries between the years of 2020 to 2023 in the states of California and Florida.

Read in Data

The raw USDA NASS Strawberry Data Set that is read in within the code is provided in the associated Github repository under the file name strawb_mar6.csv. An additional file under the name my_functions.R is also provided in the repository which provides function definitions for functions used within the rest of the code- particularly drop_one_value_col().

Clean Data

To provide an initial cleaning of the Strawberry Data Set the function drop_one_value_col() is utilized to remove any redundant columns or columns that are entirely NA- which in both cases are columns that only have one value.

When looking at the cleaned Strawberry Data Set it can be seen that the data set can be further split into two more broad programs of `CENSUS` and `SURVEY` as the information (variables of interest) held within both programs vastly differ. For example, the `CENSUS` data holds information regarding Gains, Losses, and Net Income while the `SURVEY` data provides insight into utilized chemicals. To therefore ease the process of further data cleaning the cleaned Strawberry Data Set will be split into two new data sets: one consisting of `CENSUS` data and one consisting of `SURVEY` data.

```
[1] "Looking for single value columns in data frame:  straw_cen"
[1] "Columns dropped:"
    Program      Period Week Ending 
   "CENSUS"      "YEAR"          NA 

[1] "Looking for single value columns in data frame:  straw_sur"
[1] "Columns dropped:"
       Program      Commodity         CV (%) 
      "SURVEY" "STRAWBERRIES"             NA 
```

Clean Survey Data

As eventually I wanted to gain insight into the usage of chemicals in the production of strawberries, I further cleaned the `SURVEY` Data as the information within this `Program` provided the data regarding the type and usage of chemicals in the growing process. The original `SURVEY` Data consists of columns that have multiple data entries in each as well as columns that hold data that should be placed in a different column.

The `grepl()` function is utilized within the code to recognize patterns within columns of data to shift the location of incorrectly placed data to the correct ones. To tackle the issue regarding multiple data entries in a single cell the `separate_wider_delim()` function is used to split the multiple entries into separate columns based on a chosen delimiter.

First, the non chemical columns in the `SURVEY` data will be cleaned.

Next, the chemical related columns of the `SURVEY` data is cleaned in a similar manner- using separate_wider_delim() to separate columns with multiple values and select() and gsub() to respectively remove unwanted columns and remove unwanted parentheses.

The cleaned Strawberry `SURVEY` data is saved as a .csv for further analysis purposes. This file is found within the Github repository under the name survey_clean.csv

Since further analysis will consider and compare the state of California and Florida I split the cleaned `SURVEY` data into two further data sets: one where the state of interest is California and one where it is Florida.

Clean Census Data

As I wanted to also gain insight into the connection between of sales, production, and net income for strawberries production within the two states, I further cleaned the `CENSUS` Data as the information within this `Program` provided the economic data needed to gain clarity on these topics. Like the `SURVEY` data, the original `CENSUS` Data also consists of columns that have multiple data entries as well as columns that hold data that should be placed in a different column.

```
[1] "INCOME, NET CASH FARM" "STRAWBERRIES"

[1] "Looking for single value columns in data frame:  s_cen_ca_fl_inc"
[1] "Columns dropped:"
                   Year               Commodity                   Fruit 
                 "2022" "INCOME, NET CASH FARM"                "INCOME" 
               Category                  Metric 
       " NET CASH FARM"        " MEASURED IN $" 

[1] "Looking for single value columns in data frame:  s_cen_ca_fl_str"
[1] "Columns dropped:"
                                  Year                              Commodity 
                                "2021"                         "STRAWBERRIES" 
                                 Fruit                                 Domain 
                        "STRAWBERRIES"                       "ORGANIC STATUS" 
                       Domain Category 
"ORGANIC STATUS: (NOP USDA CERTIFIED)" 
```

Now looking just at the data set where the `Commodity` of interest is `INCOME`, the same process undertaken to clean the `SURVEY` data is repeated here for the `CENSUS` data through the use of the functions separate_wider_delim(), select(), and gsub(). Here separate_wider_delim() is used to separate and get rid of the leading “OF OPERATIONS -” and “OF PRODUCERS -” with “F” and then “-” as delimiters.

The values in `FARM SALES` are provided as ranges between two values and to make data visualization easier I split the Range column into its `Lower` and `Upper` bounds. These ranges however had additional information indicating the units of interest ($ or ACRES) and so gsub() was used so that only numerical values were left. There were additionally two special cases where OR MORE and LESS THAN were used, so to ensure only numerical values were left I changed LESS THAN (Value) to go from 0 to the particular value by using gsub() to substitute out LESS THAN for "0-" and used separate_wider_delim to keep 0 in the `Lower` column and (Value) in the Upper bound column. A similar process was utilized for OR MORE but rather than 0 the `Upper` bound for that row was left as NA, as the cap for the Farm Sales value is unknown.

The cleaned Strawberry `CENSUS` data where `Commodity` is `INCOME` is saved as a .csv for further analysis purposes. This file is found within the Github repository under the name census_inc_clean.csv

Since further analysis will consider and compare the state of California and Florida I split the cleaned `CENSUS` data, where the `Commodity` is `INCOME` into two further data sets: one where the state of interest is California and one where it is Florida. Within each data set however, there is additionally a separation between `OPERATIONS` and `PRODUCERS` so each State specific data set will be further split into two new data sets so that there is an individual `OPERATIONS` and `PRODUCERS` for both states.

Now looking at the data set where the `Commodity` of interest is `STRAWBERRIES`, the same process undertaken to clean the earlier `CENSUS` data is repeated here the use of the functions grepl() and select(). Here grepl() is used to move all the entries in the `Item` column that begin with "MEASURED" to the `Metric` column and all entries with "ORGANIC -" to the `Item` column from '`Category`. This second shift leaves everything in Category to have only one value so it can be dropped from the cleaned final data.

The cleaned Strawberry `CENSUS` data where `Commodity` is `STRAWBERRIES` is saved as a .csv for further analysis purposes. This file is found within the Github repository under the name census_str_clean.csv

Since further analysis will consider and compare the state of California and Florida I split the cleaned `CENSUS` data, where the `Commodity` is `STRAWBERRIES` into two further data sets: one where the state of interest is California and one where it is Florida.

Data Analysis

Following the completion of the Data Cleaning portion of the project, I now focus on performing exploratory data analysis to uncover patterns and gain insights into potential trends and explanations that underlie the current production, growth, and sales system of strawberries. By looking into the use of chemicals in the growth of strawberries, insights can be gained regarding the impact of various chemicals on yield, quality, and overall productivity. This can help identify which chemicals are most effective and sustainable in strawberry cultivation, as well as how environmental or regulatory factors might be influencing their usage.

Furthermore, examining how farm sales affect net income provides valuable context for understanding the financial health of strawberry producers. By identifying the farm sales ranges where producers begin to experience positive net income, we can uncover the thresholds at which farm operations become profitable. This information is vital for growers and stakeholders in planning and optimizing production and sales strategies.

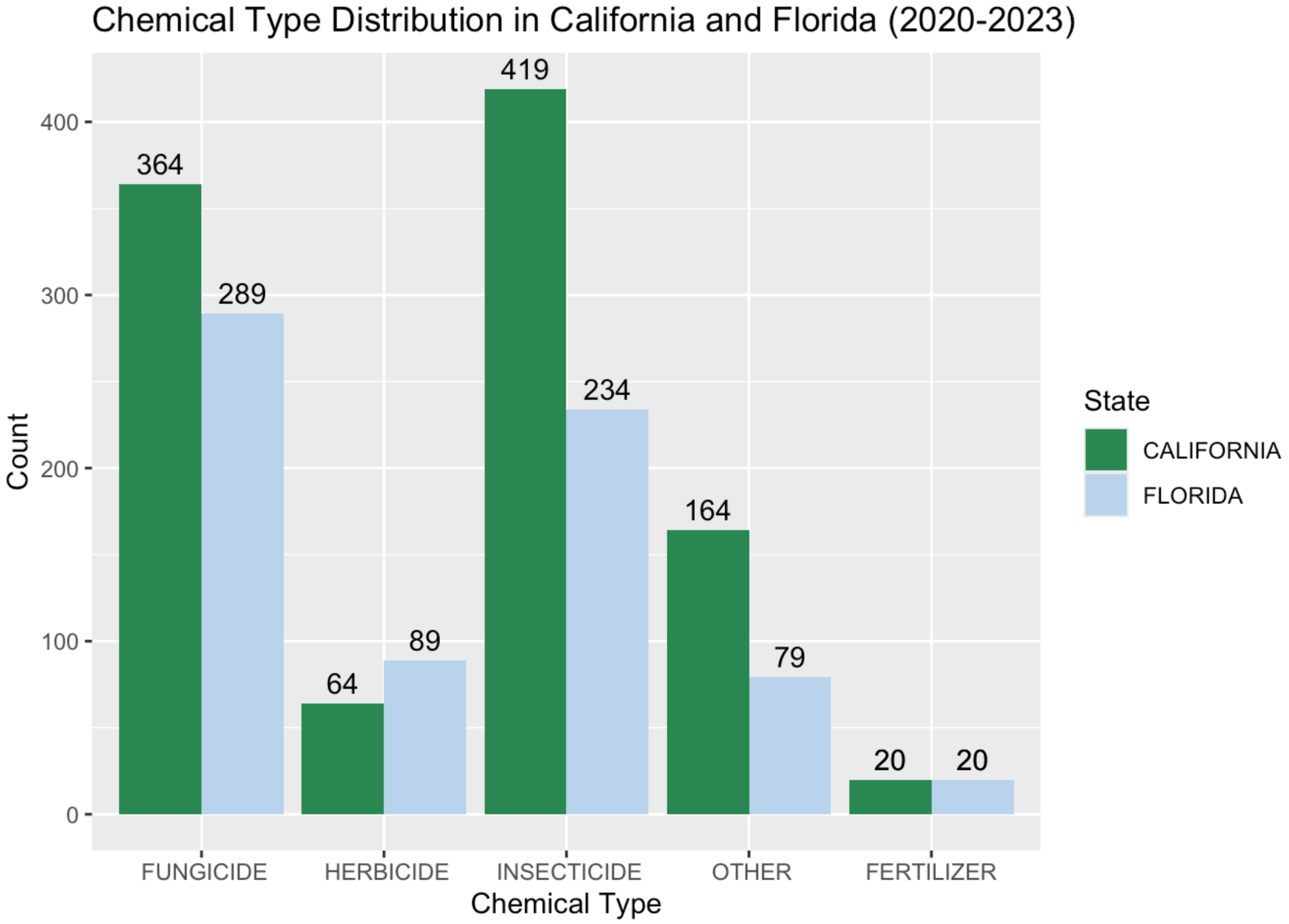
Additionally, understanding how much sales are made based on market type—whether organic, fresh, or processing— can reveal significant insights into market preferences, demand, and pricing dynamics. This information can help to guide producers to tailor their marketing and operational efforts more effectively. Analyzing this data can help to identify which market types are more lucrative and explore the potential for growth in emerging markets such as organic strawberries.

Data Analysis - Chemicals

An important area of focus in this analysis is the use of chemicals in strawberry cultivation. Chemicals such as pesticides, herbicides, and fungicides are commonly used to enhance yields and protect crops from pests and diseases. However, concerns about the environmental impact and potential health risks associated with these chemicals have sparked ongoing debates. By examining the trends in chemical usage, particularly in major strawberry-producing states like California and Florida, this section aims to understand the patterns and implications of chemical applications in strawberry farming. This insight is crucial for identifying sustainable practices that balance productivity with environmental and consumer health considerations.

This data set provides insights into the four main classes of chemicals used (fungicides, herbicide, insecticide, other) as well as fertilizers and so I was interested in the particular breakdown of which class of chemicals was most used in each state and whether the main class of chemical used between the two differed. To visualize this comparison I created a 2 variable bar graph, using ggplot, to show the difference in counts of chemical and fertilizer uses for the four major classes of chemicals between California and Florida.

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



Looking at the produced bar graph, it can be seen how California seems to have a higher reliance on chemicals compared to Florida, as for all chemical types other than herbicides California has a higher count. While fungicides are the most commonly used chemical in Florida, California's most commonly used chemical type is insecticides but it too often uses fungicides- even more than Florida. This discrepancy in the usage amount and type of chemical class used can be due to the fact that California may face greater pest and disease pressures, whereas Florida may focus on different environmental or farming practices that are less reliant on chemicals even when faced with such pressures. These differences in chemical usage could have important implications for sustainability and environmental impact, highlighting the need for tailored strategies in each state to optimize agricultural practices and manage pests and diseases effectively.

Now that more broad insight into the differences between the classes of chemicals between the two states have been found, I am interested in seeing if there are any specific chemicals that are commonly used in both California and Florida and how their usage amounts differ between each other as well as over the years. To help satisfy this curiosity I created a table to visualize the top 10 chemicals used in each state respectively based in the amount used measured in pounds. By filtering each data set based on the values measured in pounds and summing up this measurement for each chemical name, a running total of the total usage of a particular chemical could be found.

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

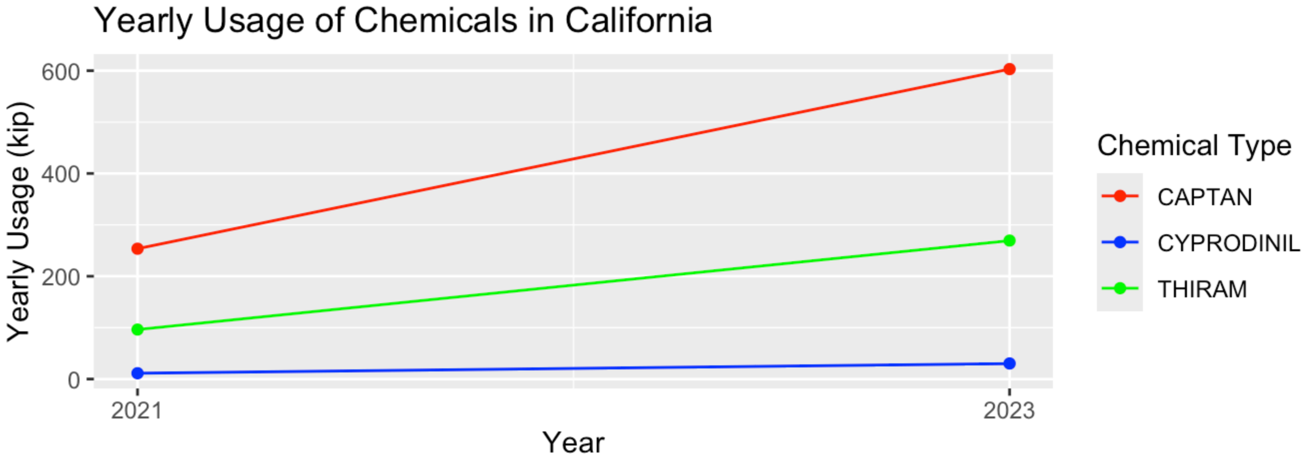
Top 10 Most Used Chemicals in California and Florida

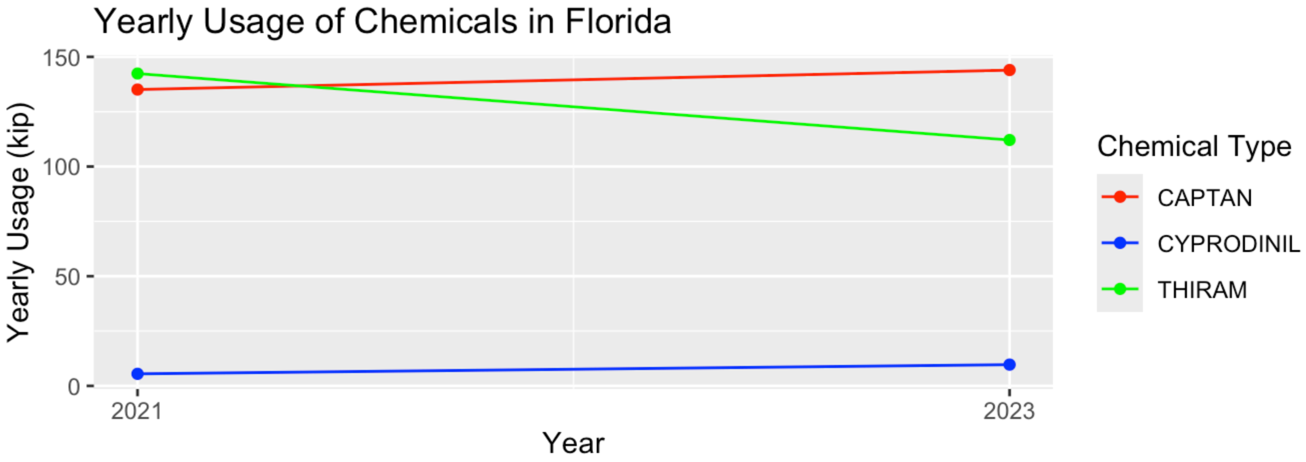
California		Florida	
Name	Total Usage (LB)	Name	Total Usage (LB)
CHLOROPICRIN	16991600	CAPTAN	279100
DICHLOROPROPENE	2491200	THIRAM	254500
SULFUR	1849300	CYPRODINIL	15200
METAM-POTASSIUM	1040400	FLUDIOXONIL	10600
CAPTAN	856700	NOVALURON	2600
THIRAM	365800	SPINETORAM	700
PENDIMETHALIN	82200	ACETAMIPRID	500
MALATHION	48500	CHLORANTRANILIPROLE	400
BIFENAZATE	45400	THIAMETHOXAM	400
CYPRODINIL	41600	ABAMECTIN	100

Based on the table above it can be seen that the three most used chemicals common to both California and Florida (based on their total usage in pounds) are Captan, Thiram, and Cyprodinil- all of which fall under the fungicide class of chemicals. According to PubMed both Captan and Thiram as chemicals exhibit moderate acute toxicity if encountered via oral or inhalation exposure (Cat II and III), while Cyprodinil poses relatively little toxicity risk (Cat III & I4). The fact that these chemicals which are some of the most commonly used in both states do to some extent post toxicity risk mY lead consumers to be more wary about purchasing such products if they are more aware about it in the future. These three chemicals were the top three most used in Florida which follows what was seen in the bar graph above where of the classes of chemicals Florida had the most number of fungicides. The largest use of the fungicides in Florida may imply that in this state Fungi are the largest threat the strawberry growing compared to insects or herbs. The top two chemicals used in California are under the the other class of chemicals, which I found interesting as with the high number of fungicides and insecticides seen in the bar graph to be used in California I was expecting their to be both classes in the top 5 most used chemicals.

After finding the three chemicals that were most commonly used in both states, I was interested in uncovering by how much that total usage of each of these chemicals changed over the years for California and Florida. To visualize this, I create a line plot, using ggplot, that plotted the total usage trajectory of each of the three top chemicals over the years 2021 and 2023 for the two states.

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





From the two line graphs produced above, it can be seen that for both California and Florida the usage of Captan and Cyprodinil have slowly increased over the years while the action of the chemical Thiram changes between the states. In California, the yearly usage of the chemical Thiram also goes up (similar to the other two chemicals), but in Florida the yearly usage of the same chemical drops instead. In California, the upward trend in Thiram usage could be driven by rising pest challenges or changing environmental factors that make this specific chemical more necessary. In contrast, the declining use of Thiram in Florida may suggest a shift in pest management strategies, possibly due to changes in pest pressure or the adoption of alternative chemicals to counter the particular fungi that this chemical in particular addresses. This might be the case rather than a more overarching shift away from fungicides as the other two fungicides (Captan and Cyprodinil) are still being utilized as the years progress.

Data Analysis - Income

An additionally important area of focus in this analysis is the economic trends that follow strawberry production within the two states. Understanding how different levels of sales contribute to profitability is crucial for evaluating the financial sustainability of such farms. Identifying the sale thresholds at which farms begin to experience positive net income, can help uncover key insights that can optimize production and sales strategies. Additionally, examining income in relation to market types—whether organic, fresh, or processing—can provide valuable context for understanding market preferences. This helps identify which market types are more lucrative, allowing producers to better target their efforts and capitalize on high-demand areas.

The data provided within the `CENSUS` portion of the data set provides insight into the correlation between Farm Sales and Net Income and I was interested in seeing the relationship between the two variables, particularly in regards to at which range of farm sales does the farm being to produce a positive net income. To visualize this I created a table that shows the net income of each state for a give range of Farm Sales in dollars.

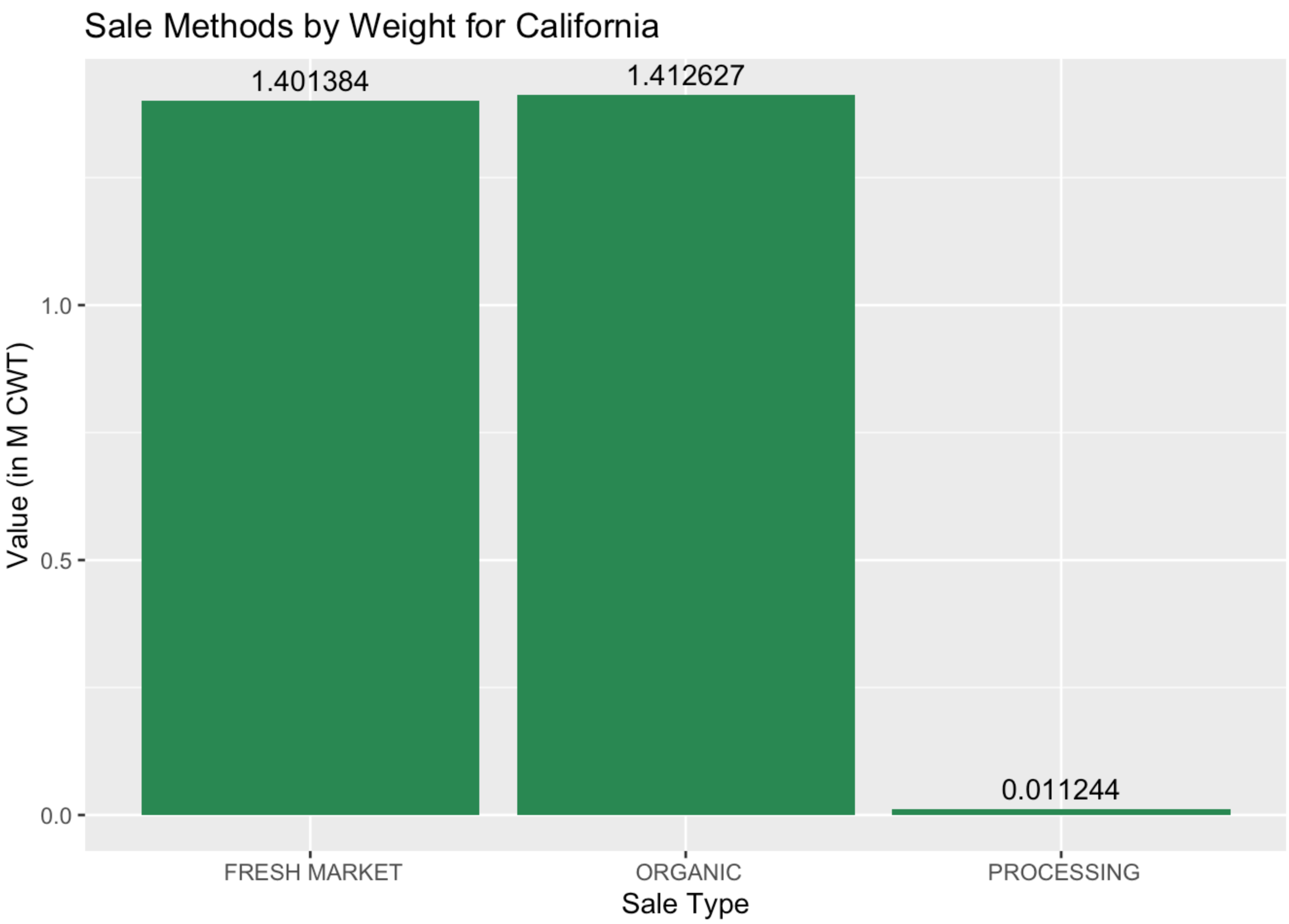
California and Florida Average Net Income in 2022 by Farm Sales

Farm Sales ($)		California		Florida	
Lower	Upper	Value	CV (%)	Value	CV (%)
0	999	-$420.67M	-10.7	-$224.16M	-23.9
1000	2499	-$61.26M	-36.6	-$50.45M	-12.5
2500	4999	-$81.06M	-22.6	-$47.60M	-14.8
5000	9999	-$92.08M	-16.0	-$44.90M	-25.2
10000	24999	-$101.38M	-29.6	-$48.35M	-29.4
25000	49999	-$113.21M	-25.0	-$21.08M	-64.5
50000	99999	-$31.06M	(H)	$16.92M	83.8
100000	249999	-$57.68M	-88.6	$26.36M	84.0
250000	499999	$203.68M	70.5	$104.15M	25.0
500000	999999	$209.18M	88.7	$108.72M	34.0
1000000	NA	$12,219.73M	6.4	$2,412.57M	5.9

Looking at the table above it can be seen how Florida begins to make a profitable (net positive) income at an earlier Farm Sale range when compared to California. For Florida, a positive net income starts as Farm Sales are between $50,000 and $99,999 while California only starts seeing a positive net income when Farm Sales are significantly higher between $250,000 and $499,999. This disparity suggests that Florida farmers can reach profitability at smaller scales, likely due to factors such as lower operational costs or reduced land and labor expenses. Its agricultural industry may also face less competition, allowing producers to become profitable more easily. On the other hand, California's higher costs, driven by factors like labor, land, environmental regulations, and a larger, more competitive agricultural industry, demand higher sales to offset these expenses and reach profitability. Additionally, California's farmers may deal with larger production volumes or more complex farming practices that contribute to higher operational costs. This difference in profitability thresholds also highlights the impact of scale economies in California, where larger-scale operations are needed to cover costs and achieve profit.

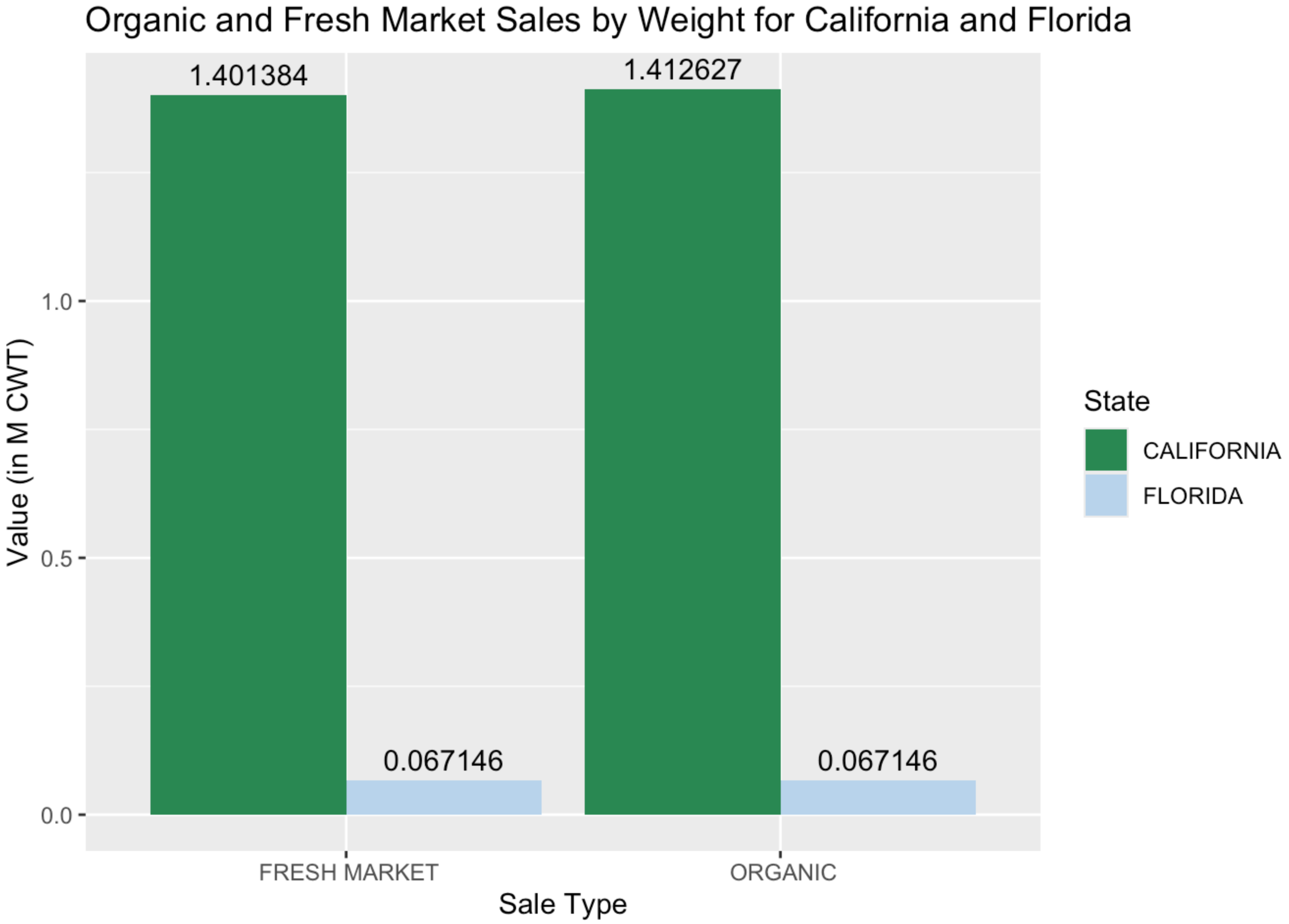
In addition to looking at the affect of Farm Sales on Net Income I was also interested in looking to see how sales in general were affected based on the market type used- whether it is organic, fresh market, or processing. However, the data in the given data set is very limited and missing some values, so the full scope of my question could not be considered but parts of it are able to. California has information for all three market types (organic, fresh market, and processing) when sales are `MEASURED BY CWT`, so I am interested to compare and see which market type produces the most sales for strawberries grown in

California. I visualized this comparison through the generation of a bar graph, using ggplot, the compares the sale methods by weight.



The bar graph highlights that in California, the vast majority of strawberry sales come from the fresh market and organic sectors rather than processing, suggesting a strong consumer preference for fresh and organic produce. The larger pull by the fresh and organic markets to consumers may be to a perception that fresher produce is better quality and more beneficial to those who eat it as well as the environment. This means that it would be import for California farmers to prioritize these markets, as they offer higher profitability (sales) compared to processed strawberries, that could come in the form of jams or jellies, meaning that more land, space, and workforce could be allocated for these sectors.

In the given data set however, Florida has no processing data and California only has hidden values for the Fresh Market and Processing sales in $, so instead I will compare organic and fresh market sales measured by weight/volume (in CWT) between the two states. To make this comparison I will generate a 2 variable bar graph, using ggplot, that compares the Value of product sold, in weight, for each market type between the two states.



From the above bar graph it can be seen that for a given state the weight of produce sold is either exactly the same or very similar regardless of whether the sale type was in Fresh Market or Organic. However, comparing the two states together, it can be seen that by volume in both sale types California sells a significant amount more of strawberries, around 21 times more, than Florida. This stark contrast points to California's larger agricultural scale, more extensive production capacity, and perhaps greater market reach allowing the state to sell much more of its goods. California's dominance in strawberry sales likely reflects its established infrastructure, larger growing areas, and longer growing seasons, while Florida, though a significant producer (the second largest of all the states) seems to operate on a comparatively smaller scale.

Conclusion

In conclusion, the data cleaning and exploratory data analysis (EDA) processes have provided valuable insights into the complexities of strawberry production, sales, and chemical usage across California and Florida. These preliminary analyses have allowed for the identification of key patterns, such as the differing profitability thresholds between the two states and the significant difference in the volume of strawberries sold. However, this is just the beginning as the analysis I provided so far in this document has only scratched the surface of all the information and comparisons that could be generated through the original data set- meaning that are many more areas that need further exploration. Further analysis into this data set as well as creating others through USDA that may focus even more heavily on, for example, the economic situation of strawberry production, could help optimize farming strategies and contribute further to more sustainable practices. Continuing analysis through this cleaned data set can help uncover more valuable insights for strawberry producers and stakeholders in the agricultural industry.