Midterm Strawberries

AUTHOR
Michael Hyder

Introduction:

In this project, I will clean, reorganize, and begin analyzing data related to strawberries in Florida and California. The data proved to be quite messey. As such, I took several steps to bring it to a workable state.

Data cleaning and organization

Cleaning and organizing data for analysis is an essential skill for data scientists. Serious data analyses must be presented with the data on which the results depend. The credibility of data analysis and modelling depends on the care taken in data preparation and organization.

Public information/citations

This is information provided to us in class. I used it extensively to think about the scope of the project and what I thought was interesting to look at. I additionally used the help of our textbooks and the various ggplot cheat sheets that exist in the R world. Beyond that, I had some trouble rendering the document and used the help of StackOverFlow which is cited below.

https://stackoverflow.com/questions/42340928/knit-error-object-not-found

WHO says strawberries may not be so safe for you–2017March16

Pesticides + poison gases = cheap, year-round strawberries 2019March20

Multistate Outbreak of Hepatitis A Virus Infections Linked to Fresh Organic Strawberries-2022March5

Strawberry makes list of cancer-fighting foods-2023May31

What is the question?

These were some general questions offered to help us think about what might be cool in the dataset.

- Where they are grown? By whom?
- Are they really loaded with carcinogenic poisons?
- Are they really good for your health? Bad for your health?
- Are organic strawberries carriers of deadly diseases?
- When I go to the market should I buy conventional or organic strawberries?

Read the file

This is my initial cleaning process. I read teh data and had to separate the Data Item column first. Then, I wanted to tag where the organic, processing, and fresh data was. Unfortunately, there was not much data for these three categories, but later I will do my best to find insights with it. I then wanted to find what chemicals could be interesting (and feasible) to analyze. I removed any rows that did not have a value for the chemicals and found the 5 chemicals that had the most data entries. That is the output below.

```
[1] "CHEMICAL, FUNGICIDE: (CAPTAN = 81301)"
[2] "CHEMICAL, FUNGICIDE: (CYPRODINIL = 288202)"
[3] "CHEMICAL, FUNGICIDE: (FLUDIOXONIL = 71503)"
[4] "CHEMICAL, FUNGICIDE: (THIRAM = 79801)"
[5] "CHEMICAL, INSECTICIDE: (NOVALURON = 124002)"
```

Chemicals

I decided to use Novaluron, Captan, and Thiram because they had sufficient data to find real insights and were fairly interesting. Novaluron is an insecticide that is commonly used to remove pests like beetles and catepillars from fruits. It is relatively safe for beneficial insects, such as bees, and has low toxicity to humans. Captan is a fungicide that aims to prevent spores from growing on the surface of foods. It is used on strawberries to prevent mold usually. It is moderately toxic to humans and is classified as a probable human carcinogen. Thiram is another fungicide that is used to prevent seed rot in strawberries. It is moderately toxic to humans and will cause skin and eye itchiness. It is banned for food use in the European Union.

Once I found the chemicals I wanted, I created my chem_data table that had all the information regarding specifically these 3 chemicals. Then I removed any unnecessary columns. With that cleaned data, I could now find some relationships!

Total Use of Each Chemical by State

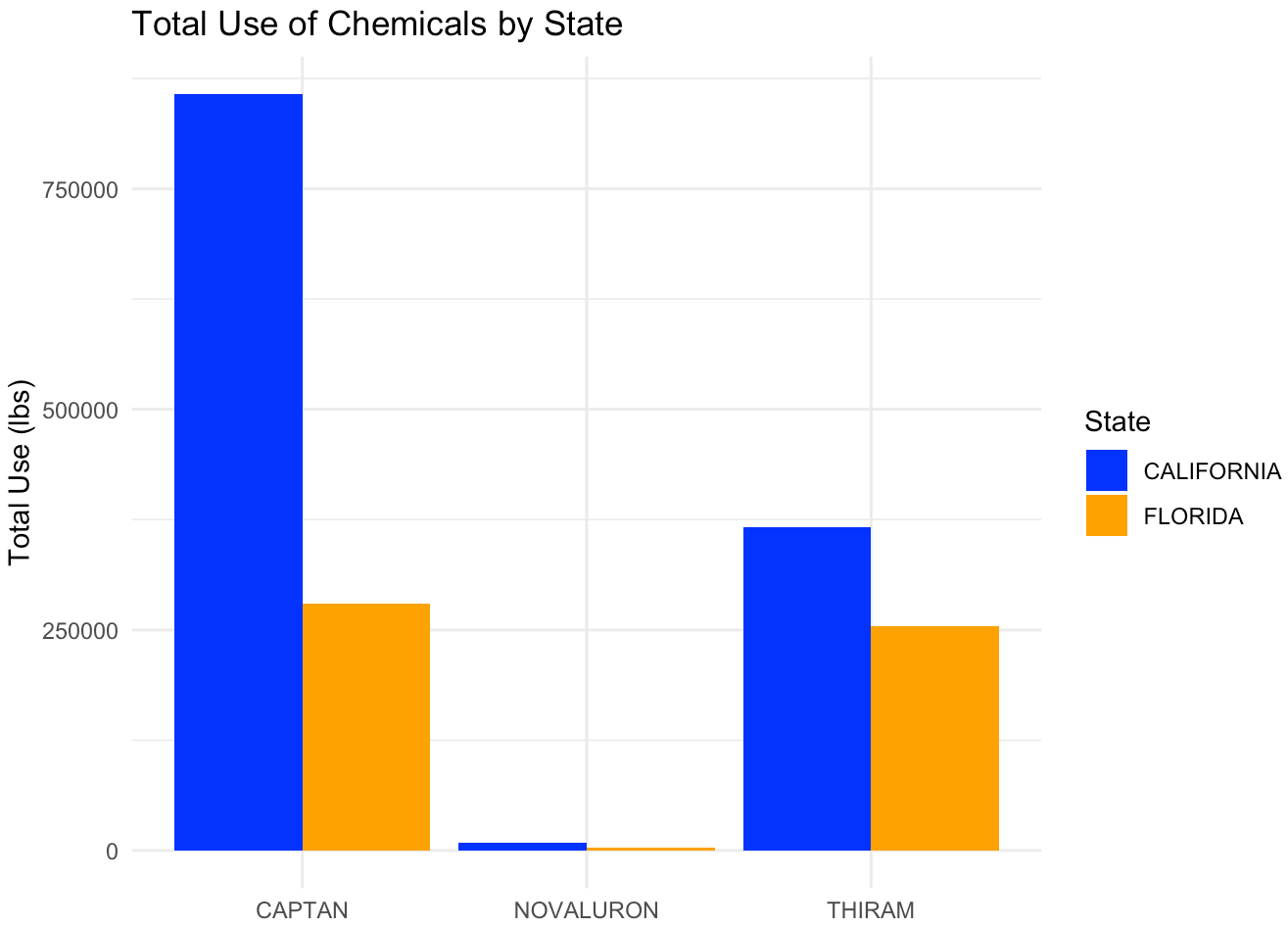
Chemical	State	Total Use (lbs)
CAPTAN	CALIFORNIA	856885
CAPTAN	FLORIDA	279309
NOVALURON	CALIFORNIA	9011
NOVALURON	FLORIDA	2722
THIRAM	CALIFORNIA	365944
THIRAM	FLORIDA	254679

Above is a table of the total amount of each chemical used in each state. Below is a table of the total amount of each chemical used in each year by both states combined.

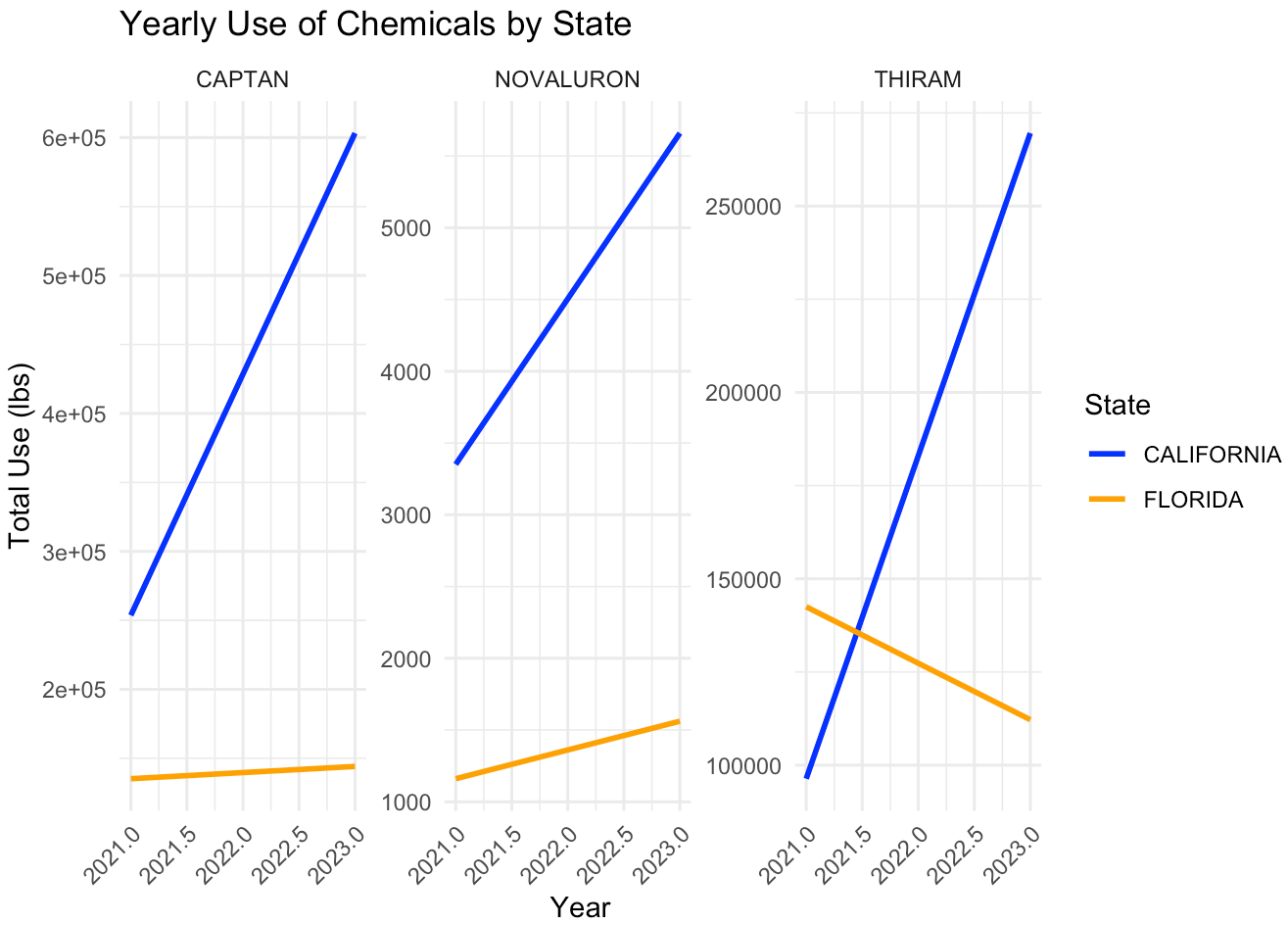
Yearly Use of Each Chemical

Chemical	Year	Total Use (lbs)
CAPTAN	2021	388864
CAPTAN	2023	747329
NOVALURON	2021	4512
NOVALURON	2023	7221
THIRAM	2021	238854
THIRAM	2023	381768

From the tables, we can make bar and line graphs. This visualizations will tell us the story of the data.



Above it is clear that California (CA) uses significantly more chemicals than Florida (FL). This is because they simply have more land to cover. Interestingly, CA uses a much greater proportion of Captan than FL does. At the same time, FL uses a much higher proportion of Thiram than CA does: FL uses nearly as much Thiram as Captan. This is concerning, as Thiram is a much more dangerous and harmful chemical than Captan. In both states, Novaluron has a significantly lower usage.



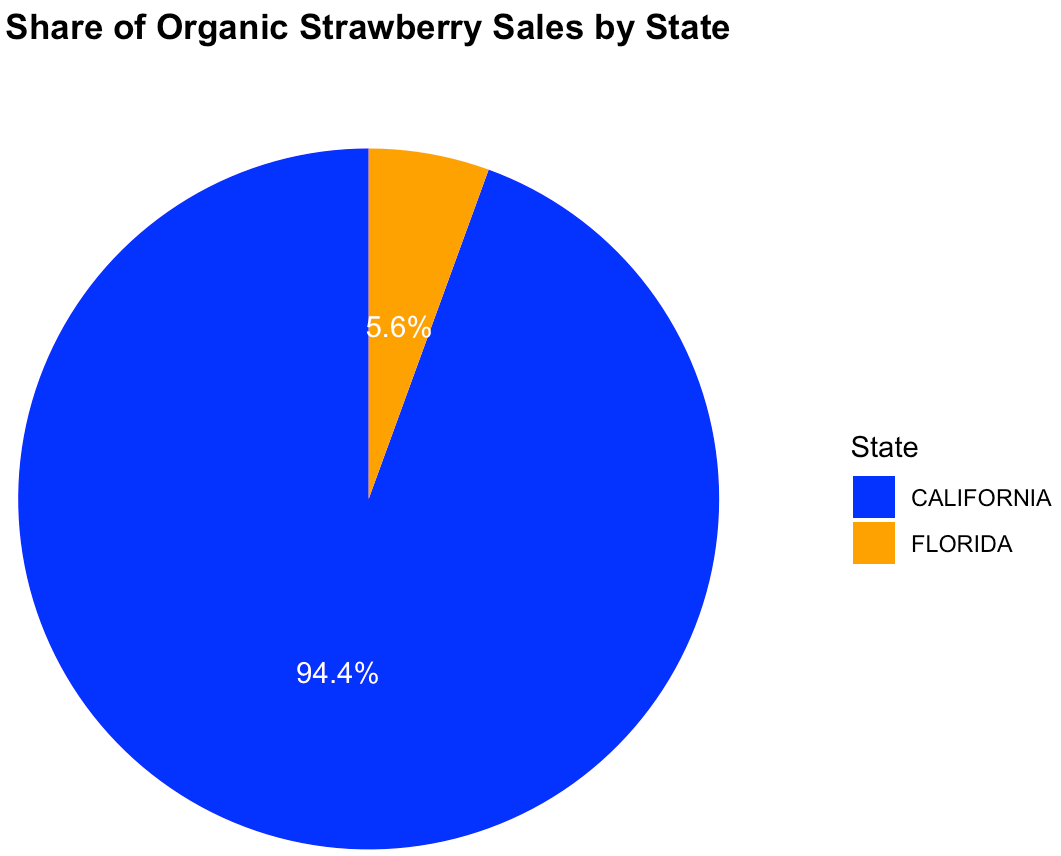
The line chart tells us a very interesting story. For all three chemicals, CA has a significantly larger slope, meaning they are increasing the rate at which they use these chemicals much more drastically than FL. The most surprising finding from this graph however, comes from the Thiram. In 2021, FL actually used more Thiram than CA, despite having much less land to cover. But, by 2023, FL decreased the amount of Thiram it used. This is a great trend as Thiram can be very toxic and minimizing its use makes strawberries all around safer. Unfortunately, CA more than doubled its use of Thiram. This is concerning. The initial bar graph was somewhat deceiving, as it showed FL having a potential Thiram problem, but in reality, CA are the ones who need to restrict the use of Thiram for their strawberries. FL appears to be on the right track with that chemical, however, it may be the case that they are substituting the Thiram with other toxic chemicals that I did not analyze.

Organic Strawberries by State

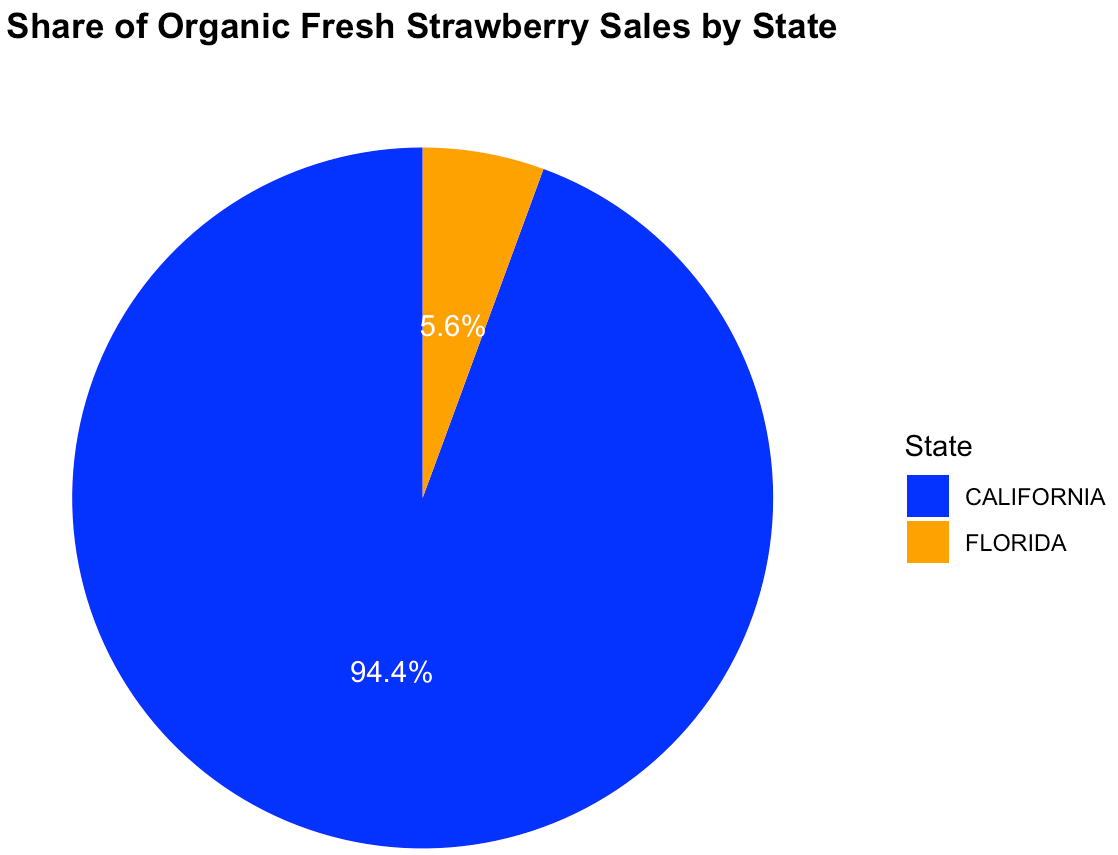
Now, we switch our attention to organic strawberries. We would expect CA to produce a larger proportion of organic strawberries, but how much? And what can we say about the market for organic strawberries at fresh markets? These are the questions I will look to answer here.

To find these answers, I had to do some tricks with the data. Unfortunately, there was not enough data that had the correct units to solve this problem. However, I had data for the amount of strawberries produced in hundreds of pounds. I additionally had partial information about the price of organic strawberries in each state. With this, I was able to estimate the total sales of strawberries in each state

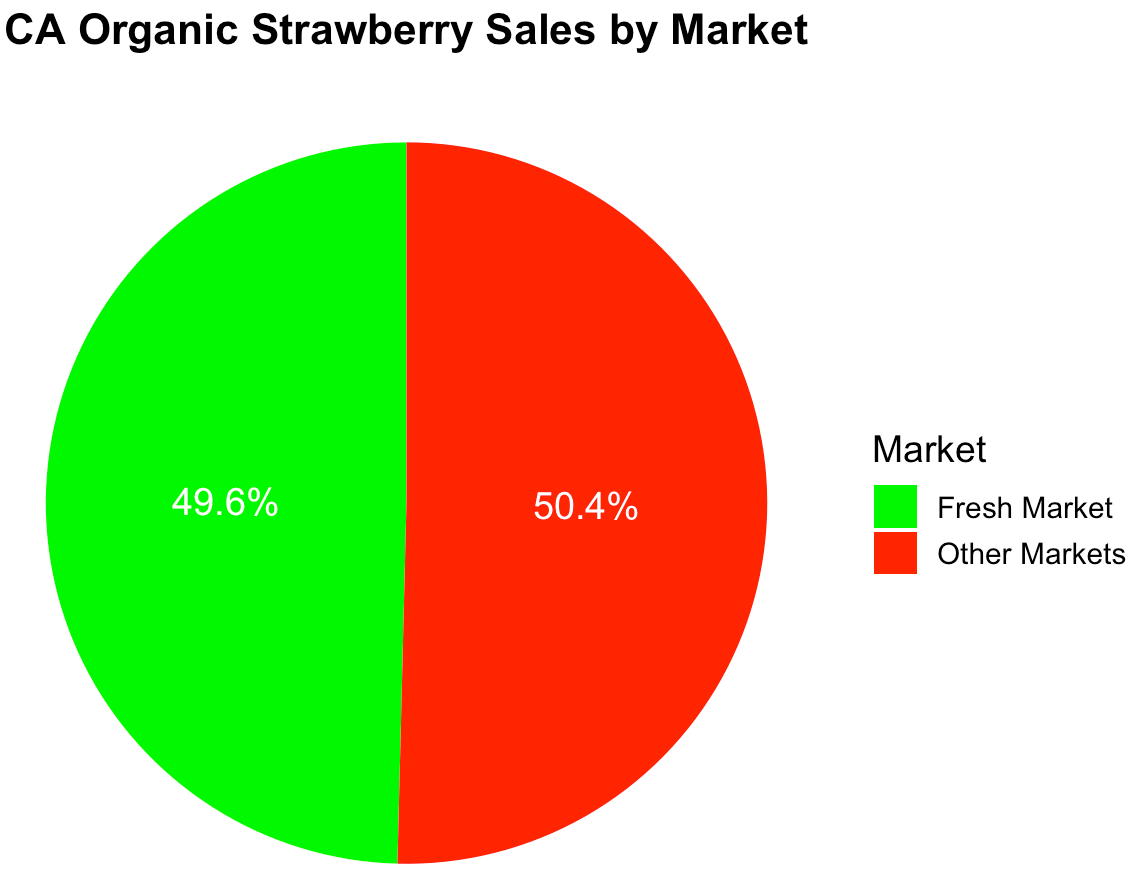
in 2021 (not each data still to get 2023 results). To convert CWT (hundreds of pounds) to dollars, I found where there is CWT in either the item or metric column and added the values of those entries together for each state. Then did the same but with $ instead of CWT. Then, I divided the $ total by the CWT total to find the price per 100 pounds of strawberries. Then, you can multiple this rate by the values in the CWT entries to find the dollar amounts sold. In doing this, I found that in CA the price per 100 pounds was about 110$ where in FL it was about 137$. Perhaps this is because there is more supply in CA.



The estimated values gave me the result above. Approximately 94.4% of the total organic strawberries from the two states came from CA in 2021. Somewhat amazingly, the exact same proportion of organic strawberries that were sold at fresh markets came from CA. This should mean that CA and FL sell the same proportion of their organic strawberries at fresh markets as they do in grocery stores and other venues. This naturally made me wonder what that percentage was. The final (and I think coolest) finding is inspired by this question and visualized in the last pie chart.



Below is the pie chart showing the organic strawberry sales by market in CA in 2021. The data surrounding the processed markets was severely lacking, but for this dataset, we will just classify the sale as organic or not organic. Remarkabley, CA sells almost exactly 50% of its organic strawberries at fresh markets and 50% with other vendors. I find this incredibly interesting, because as we say in the above pie charts, FL must also sell approximately 50% of its organic strawberries at fresh markets. So, we conclude from these findings that about 50% of all organic strawberries are sold at fresh markets from the two largest strawberries producing states in the US.



Conclusions

Regarding chemicals, it appears as thought both states use a significant amount of chemicals on their non-organic strawberries, but we can see trends in FL moving away from very toxic chemicals, such as Thiram, and CA moving towards those chemicals. Now knowing what these chemicals are, and their prevelance in the largest strawberry producing states, I would not buy non-organic strawberries. Now looking at organic strawberries, I can see that fresh markets have an equal share of the market with big grocery stores. However, thinking economically, that largely tells me that organic strawberries are not a great return on investment, as if they were, large grocery stores would want a much larger share of the market. After analyzing this data, I can confidently say that I will think twice before buying non-organic strawberries, and may find myself looking for fresh markets to get my fruits!