Final.Version-StrawberryMidterm

Aidan Patt

EDA of Strawberries

Libraries

```
library(knitr)
library(kableExtra)
library(tidyr)
library(tidyverse)
library(stringr)
library(esquisse)
library(dplyr)
```

Citations

For Code

stack overflow

sparkbyexamples.com

Data Wrangling with R

R Markdown Cookbook

For Research

Inside Climate News

CA Pesticide Use

Crop Nutrition

EPA R.E.D FACTS: Thiram

The Data

```
# Data set  used for chemical usage information
strawberry = read.csv("strawb_mar6.csv")

# Data set used for sales and production information
sp_straw = read.csv("strawberry_sales&production.csv")
```

For this project two data sets were cleaned and used. The original data set, given in class was used to obtain information about the amount of chemicals used for strawberry farming for California and Florida. The second data set was used to obtain more information about the sales and production of strawberries for each state. The data cleaning process and analysis will focus on California and Florida because they are the largest producers of strawberries in the United States. The document will first focus on data cleaning and visualization for chemical usage and then focus on sales and production respectively.

Chemical Usage

Data Cleaning for Chemical Information

The original data set has a lot of unnecessary values and columns, with myriad different measurements, which needs to be cleaned for any useful analysis to occur. There is no chemical information given in census, so the first thing done was to separate the data into census and survey data sets. After that the survey data set was taken and further separated into two data sets for each state, California and Florida. For each state, since the focus was on individual chemical use, rows totaling all chemical usage were taken out. Additionally, there were many chemicals that had undisclosed data, and these were removed from the data set, in the hopes that there would be similar chemicals for both states where data was disclosed, and so that comparison could be done across these chemicals.

```
#| echo: false
#| label: Cleaning up survey data for California and Florida


source("myfunctions.R")

# getting rid of uneccessary columns
strawberry = drop_one_val_col(strawberry)

# only looking a periods that are a year, and getting rid of
# unecessary/redundant columns
strawberry = strawberry |> filter(Period == "YEAR") |>
  select(!Week.Ending & !Period & !State.ANSI)

## breaking up strawberry into census and survey

## Census
strawb_c = strawberry |> filter(Program == "CENSUS")

## Survey
strawb_s = strawberry |> filter(Program == "SURVEY")


## census for florida and california
straw_c_f = strawb_c |> filter(State == "FLORIDA")

straw_c_c = strawb_c |> filter(State == "CALIFORNIA")


## survey for florida and california
straw_s_c = strawb_s |> filter(State == "CALIFORNIA")
straw_s_f = strawb_s |> filter(State == "FLORIDA")


### Cleaning survey data for florida
# straw_s_f

# Eliminating all columns with just one value
straw_s_f = drop_one_val_col(straw_s_f)


# making new table with just total values
```

```r
tot_straw_s_f = straw_s_f |> filter(Domain == "TOTAL")

# Getting rid of total values from the original table
straw_s_f = straw_s_f |> filter(Domain != "TOTAL")

# Making a column strictly for chemical name
straw_s_f = straw_s_f |> separate_wider_delim(
  cols = Domain.Category,
  delim = " (",
  names = c("domainR", "Chemical.Name"),
  too_many = "merge"
)

# getting rid of redundant column
straw_s_f = straw_s_f |> select(!domainR)

# separating chemical name into the chemical name and number
straw_s_f = straw_s_f |> separate_wider_delim(
  cols = Chemical.Name,
  delim = " = ",
  names = c("Chemical.Name", "Chemical.Number"),
  too_few = "align_start"
)

# cleaning up parenthesis in the columns
straw_s_f$Chemical.Number = gsub(")", "", as.character(straw_s_f$Chemical.Number))
straw_s_f$Chemical.Name = gsub(")", "", as.character(straw_s_f$Chemical.Name))
straw_s_f$Chemical.Name = sub("ETHYL (2E;4Z-DECADIENOATE",
                          paste0("ETHYL (2E;4Z-DECADIENOATE",")"),
                          as.character(straw_s_f$Chemical.Name),
                          fixed = TRUE)

# making a list of undisclosed chemicals and removing total rows
undisclosed_chem_f = straw_s_f |> filter(Chemical.Name != "TOTAL") |>
  filter(Value == " (D)" | Value == " (NA)")

# Removing undisclosed data from the table
straw_s_f = straw_s_f |> filter(Chemical.Name != "TOTAL") |>
  filter(Value != " (D)" & Value != " (NA)")

# splitting Data.item column into multiple columns
straw_s_f = straw_s_f |> separate_wider_delim(
```

```
  cols = Data.Item,
  delim = ",",
  names = c("Fruit", "Bearing Type", "Metric"),
  too_many = "merge"
)

# Getting rid of columns with one value
straw_s_f = drop_one_val_col(straw_s_f)


### Cleaning survey data for California

# straw_s_c
straw_s_c = drop_one_val_col(straw_s_c)

# unique(straw_s_c$Data.Item)
# unique(straw_s_c$Domain)

# making new table with just total values
tot_straw_s_c = straw_s_c |> filter(Domain == "TOTAL")

# Getting rid of total from the table
straw_s_c = straw_s_c |> filter(Domain != "TOTAL")

# Making a column strictly for chemical name
#unique(straw_s_c$Domain.Category)
straw_s_c = straw_s_c |> separate_wider_delim(
  cols = Domain.Category,
  delim = " (",
  names = c("domainR", "Chemical.Name"),
  too_many = "merge"
)

straw_s_c = straw_s_c |> select(!domainR)

straw_s_c = straw_s_c |> separate_wider_delim(
  cols = Chemical.Name,
  delim = " = ",
  names = c("Chemical.Name", "Chemical.Number"),
  too_few = "align_start"
)
```

```
# cleaning up parenthesis in the columns
straw_s_c$Chemical.Number = gsub(")", "", as.character(straw_s_c$Chemical.Number))
straw_s_c$Chemical.Name = gsub(")", "", as.character(straw_s_c$Chemical.Name))

# Making a table of undisclosed chemicals for California
undisclosed_chem_c = straw_s_c |> filter(Chemical.Name != "TOTAL") |>
  filter(Value == " (D)" | Value == " (NA)")

# Removing undisclosed data from the table
straw_s_c = straw_s_c |> filter(Chemical.Name != "TOTAL") |>
  filter(Value != " (D)" & Value != " (NA)")

# splitting Data.item column into multiple columns
straw_s_c = straw_s_c |> separate_wider_delim(
  cols = Data.Item,
  delim = ",",
  names = c("Fruit", "Bearing Type", "Metric"),
  too_many = "merge"
)

# Getting rid of columns with one value
straw_s_c = drop_one_val_col(straw_s_c)
```

Deciding on the Metric for Chemical Use

There were many different measurements given for chemical usage. It seemed important to control for size, since one state might have significantly more farmland than the other, if a metric for land is not taken into account, and data was given based off only the amount chemicals used by state it could lead to misleading differences. However it is also important to get some type of aggregate annual value, rather than amount used at each application. Due to these considerations, the metric used was the average chemical use in lb/acre/year.

Deciding on Chemicals to Compare

There were serveral reasons for the chemicals that were chosen. While dichloropronene is one of highest used chemicals in California, there is no information of it being used in Florida, it is not even a part of the chemicals with usage that is undisclosed. It seemed valuable to look into this differing use in dichloropronene, so it was one of the chemicals that made the list because Califoria used it a lot, and Florida did not at all. Additionally while both use potash as a fertilizer, it was also notable that California's use of it is almost double that of Floridas.

And as for the top ten chemicals, Thiram seemed to be the only one on the list for both where Florida used more on average.

Tables of Chemical Usage for Each State

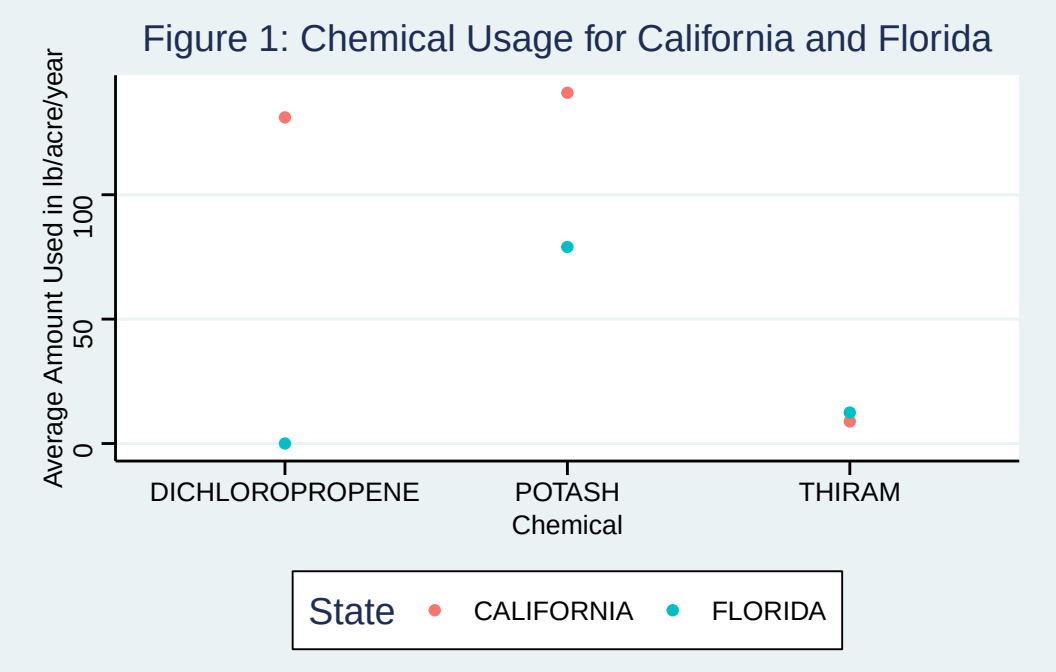
Table 1: Top 10 Chemicals used in California

Type	Chemical Name	Average use in lb/acre/yr
CHEMICAL, OTHER	CHLOROPICRIN	442.413
FERTILIZER	NITROGEN	165.000
FERTILIZER	POTASH	141.000
CHEMICAL, OTHER	DICHLOROPROPENE	131.091
FERTILIZER	PHOSPHATE	88.000
CHEMICAL, FUNGICIDE	SULFUR	38.761
CHEMICAL, FUNGICIDE	CAPTAN	15.932
CHEMICAL, FUNGICIDE	THIRAM	8.873
CHEMICAL, HERBICIDE	PENDIMETHALIN	5.892
CHEMICAL, INSECTICIDE	MALATHION	2.398

Table 2: Top 10 Chemicals used in Florida

Type	Chemical Name	Average use in lb/acre/yr
FERTILIZER	POTASH	79.000
FERTILIZER	NITROGEN	43.000
FERTILIZER	PHOSPHATE	29.000
CHEMICAL, FUNGICIDE	THIRAM	12.456
CHEMICAL, FUNGICIDE	CAPTAN	10.509
CHEMICAL, FUNGICIDE	CYPRODINIL	0.756
CHEMICAL, FUNGICIDE	FLUDIOXONIL	0.470
CHEMICAL, FUNGICIDE	AZOXYSTROBIN	0.214
CHEMICAL, INSECTICIDE	NOVALURON	0.180
CHEMICAL, INSECTICIDE	ACETAMIPRID	0.179

Data Visualization for Usage of Chosen Three Chemicals



Discussion

The reason dichloropropene is used so much in California, seems to be the fact that the state is highest user of insecticides and pesticides in the nation. However dichloropropene is a dangerous fumigant and pesticide that is listed as a carcinogen, so it is concerning that its use is still so high, especially becasue it has caused workers to develop cancer. Potash is a fertilizer for the soil, and a possible reason that California has to use more of it is becasue the soil in california has less natural compounds to promote growth. Additionally, the scale of California's strawberry industry is much larger than Florida's which could lead to less ability for sustainable farming practices, thus leading to a more heavy reliance on potash and other fertilizers. Thiram is a fungicide, that due to its nature to be mobile enough to reach water sources and influence aquatic life, has a very regulated used. It is possible that both California and Florida, are using Thiram as much as they can, hence the close value given for both.

Sales and Production

When examining the sales and production of strawberries for both states, it seemed necessary to find more data. As stated previously, USDA was used to obtain another data set giving

sales and production information for strawberries across California and Florida, over multiple years.

Sales of Strawberries

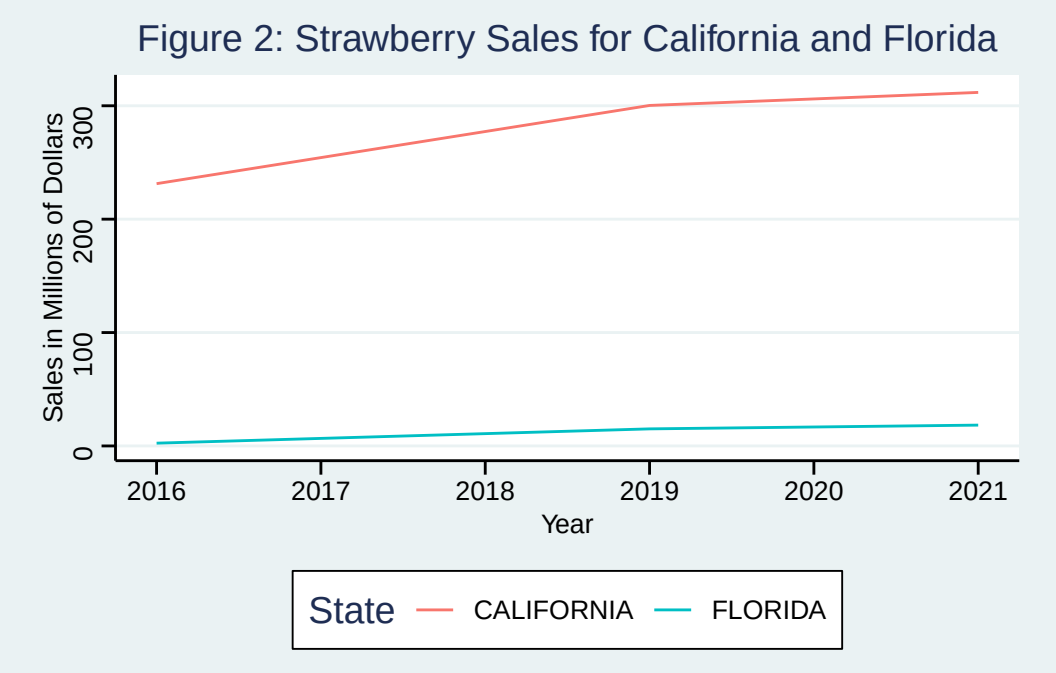
While the new data set provided some more information USDA only gave sales statistics for organic strawberries. Although the new data set showed the different types of markets the organic strawberries were sold in, encompassing processing, conventional, organic, and fresh markets, there was not enough reporting for them, and it still did not seem to be enough data points for meaningful visualization of the different markets. However when looking at sales, without breaking it down into the different markets, there was a decent amount of data for California and Florida.

Table of Sales for California and Florida

Table 3: STRAWBERRY SALES IN DOLLARS

Year	State	Value	CV....
2021	CALIFORNIA	311784980	46.0
2021	FLORIDA	18358396	99.95
2019	CALIFORNIA	300277717	33.1
2019	FLORIDA	15055709	83.4
2016	CALIFORNIA	231304956	13.7
2016	FLORIDA	2455805	21.9

Data Visualization of Sales



Discussion

While it makes sense that California would have more sales than Florida. What is interesting and noticeable is that the rate at which California's sales have increased is higher. The fact that the industry is selling more could explain the states reliance on strawberry farming revenue and why it continues to use harmful carcinogenic chemicals in much larger quanitities than any other state.

Production of Strawberries

For examining the production of strawberries the only options are seemingly total production, production for fresh market, and production for processing. Initially focused on just the plain production of strawberries without a focus on the market.

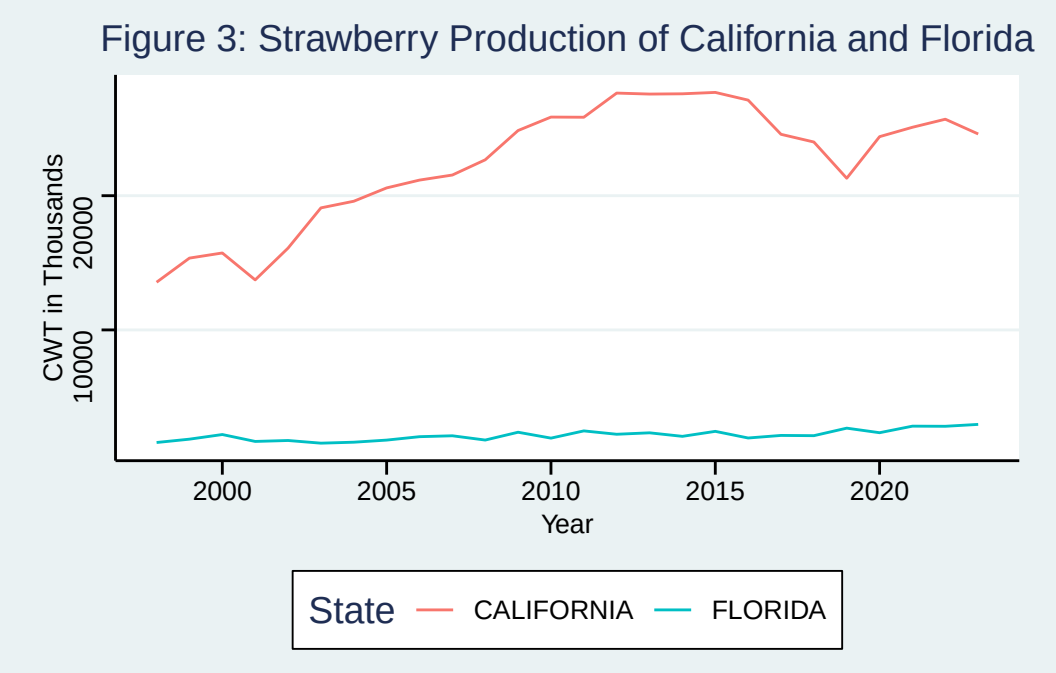
Table of Strawberry Production for California and Florida

```
#| echo: false
#| label creating table of Production of Strawberries


### Table for Production of Strawberries for California and Florida in CWT
P_straw = p_straw |> filter(Econ == "PRODUCTION" & Measure == "MEASURED IN CWT") |>
                    drop_one_val_col()

P_straw = P_straw |> mutate("CWT in thousands" = Value / 1000)

# Generating the table of first 20 values
prod_table = head(P_straw, 20)
prod_table |> select(!Value) |>
  kable( booktabs = TRUE,
          caption = "Amount of Strawberries Produced") %>%
  kable_styling(latex_options = "striped", font_size = 10)
```

Table 4: Amount of Strawberries Produced

Year	State	CWT in thousands
2023	CALIFORNIA	24600.0
2023	FLORIDA	2960.0
2022	CALIFORNIA	25700.0
2022	FLORIDA	2820.0
2021	CALIFORNIA	25100.0
2021	FLORIDA	2830.0
2020	CALIFORNIA	24400.0
2020	FLORIDA	2340.0
2019	CALIFORNIA	21300.0
2019	FLORIDA	2680.0
2018	CALIFORNIA	24000.0
2018	FLORIDA	2120.0
2017	CALIFORNIA	24574.5
2017	FLORIDA	2137.5
2016	CALIFORNIA	27122.0
2016	FLORIDA	1947.5
2015	CALIFORNIA	27697.0
2015	FLORIDA	2442.0
2014	CALIFORNIA	27592.0
2014	FLORIDA	2071.0

Data Visualization of Production of Strawberries



Discussion

The graph of the production seems to be in line with the sales of each state. However there was a dip in production for California between 2015 and 2020, that does not seem to be represented in sales. This could be due to a number of reasons, but probably has most to do with scarcity.