Strawberry EDA Report

Yuchen Huang

Data acquisition and assessment

Data sources

The data have been sourced from the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA). This data pertains to strawberries cultivated in the United States, detailing their market categories and associated pesticide use. The information was gathered through both census and survey methods.

Another data is "The WHO Recommended Classification of Pesticides by Hazard", so that we can use the form to determine the toxicity of the pesticides used in the strawberry data.

Assumptions and motivations

While strawberries contain compounds that might help lower the risk of cancer, their overall health benefits remain debated. Some sources raise concerns over the potentially toxic pesticides used on strawberries, suggesting that these chemicals, which might not be completely removed after washing, could have negative reproductive effects. Conversely, the Centers for Disease Control and Prevention issued a food safety alert in 2022 linking fresh organic strawberries to a multistate outbreak of Hepatitis A Virus infections. Therefore, we aim to use the NASS data on U.S. strawberries to investigate whether organic strawberries or conventionally grown strawberries carry pesticides that are more harmful to human health, as well as factors related to the production and sale of strawberries.

Data cleaning and organization

Remove NA columns

After loading libraries and reading data strawberry, The first step of data cleaning is removing all the columns containing only NA.

```
[1] "Columns dropped:"
 [1] "Week Ending"      "Geo Level"        "Ag District"      "Ag District Code"
 [5] "County"           "County ANSI"      "Zip Code"         "Region"
 [9] "watershed_code"   "Watershed"        "Commodity"
```

Check if every line of strawberries data is associated with a state

Next, we examine the data and check if every line of strawberries data is associated with a state. Then we figure out the states that containing most data of strawberries.

```
[1] "Every row has value in the State column."
```

```
[1] "The state with the most rows of data is CALIFORNIA with 1886 rows."
```

The state with the most row of data is CALIFORNIA with 1886 rows.

Examine California

Now we have California as the state with most data, so we want to examine California.

Here's the composite columns:

Census: Data Item, Domain Category

Survey: Data Item, Domain, Domain Category

Based of the result we found in California data, we can find a way to separate the strawberry data.

Separate CENSUS and SURVEY into two Data Frames

In the strawberry data frame, The CENSUS rows contains marketing, sales, and production data. The SURVEY rows contain rows which may be redundant with the CENSUS rows and chemical application rows.

Clean and organize Census data frame

Seperate `Data Item` into columns by ","

In this step, we divide the `Item data` column into 4 columns, containing different information like crop type and fruit type.

Create a `Fresh Market` Column

```r
strwb_census <- strwb_census |>
    mutate(`Fresh Market` = temp2, .after = temp2)
strwb_census$`Fresh Market` <- strwb_census$`Fresh Market` |> str_replace("^MEA.*", "")
strwb_census$`Fresh Market` <- strwb_census$`Fresh Market` |> str_replace("^P.*", "")
strwb_census$`Fresh Market`[is.na(strwb_census$`Fresh Market`)] <- ""

strwb_census$temp2 <- strwb_census$temp2 |> str_replace("^FRE.*", "")

strwb_census$`Fresh Market` <- strwb_census$`Fresh Market` |> str_replace("^FRESH MARKET -
```

Here we create a column named `Fresh Market` containing only the information after "Fresh Market".

Create a `Process Market` Column

Here we create a column named `Fresh Market` containing only the information after "Fresh Market", and remove NA's from prop_acct, temp2, and temp3.

Clean up the data into named columns

Clean and organize Survey data frame

After observing the survey data, we can see that columns `Data Item` and `Domain Category` need to be cleaned and organized.

Separate `Data Item` into columns by ","

Create a `Totals` column

Create a `Fresh Market` column

Create a `Process Market` column

Create a `Not Sold` column

Create a `Utilized` column

Create a `Bearing` column

Create a `Metric` column and clean up

Separate `Domain Category` into columns by ","

Create a `Cide Type` column

Create `Chemical Name` and `PC` columns

Create a `CAS` column

Create a `Class` Column

Clean up the survey data

References

Material about strawberries

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

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

Technical references

In their handbook "An introduction to data cleaning with R" by Edwin de Jonge and Mark van der Loo, de Jonge and van der Loo go into detail about specific data cleaning isssues and how to handle them in R.

"Problems, Methods, and Challenges in Comprehensive Data Cleansing" by Heiko Müller and JohannChristoph Freytag is a good companion to the de Jonge and van der Loo handbook, offering additional insights.

Initial Questions

- How are the sales of organic strawberries in different markets?
- How are the sales of non-organic strawberries in different markets?
- Which has better sales, organic or non-organic strawberries?
- What is the toxicity of non-organic strawberries that use different kinds of pesticides?

The data

Here's some explanation of the columns' names

`Value` : Understand it combining with metrics. For example, the metric is lb, then we should understand the value as the production of this type of strawberry is `Value` lb.

`Cide` : The type of pesticide that is applied on the strawberry.

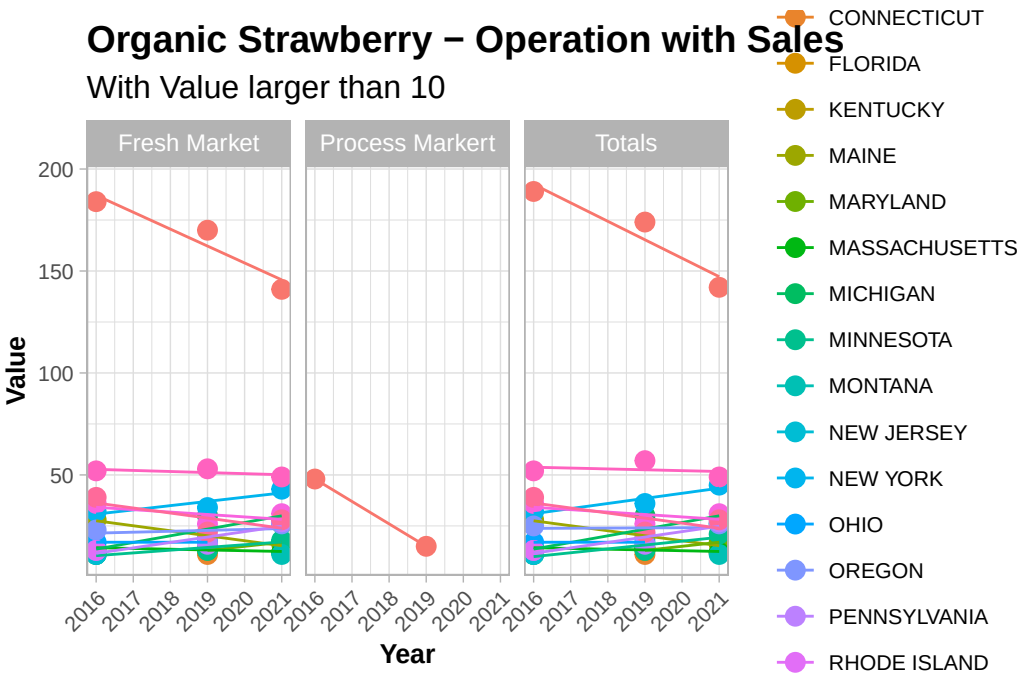
`Bearing` : Application of bearing means that the pesticide was applied on the fruit.

EDA

Census Data

In the following two plots I will focus on Census Data, and analysze the data from different markets in different metrics.

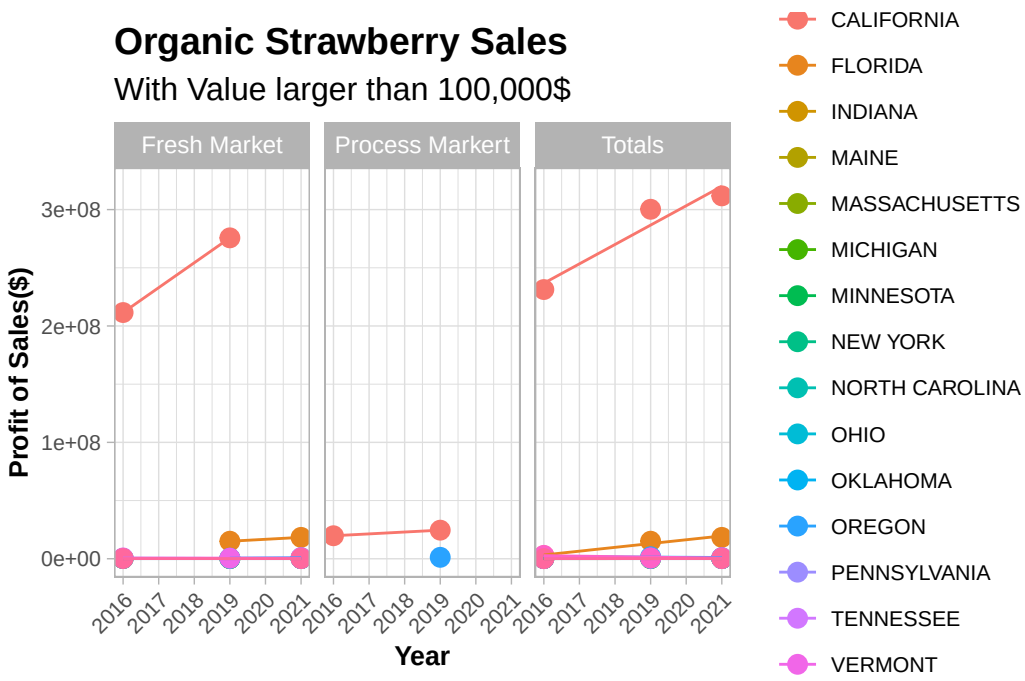
Operations with sales



In this form I consider all the organic stawberry have operations with sale, and pick the States that have value larger than 10, then categorize them into three markets of strawberry. Also, I add a linear regression line for each state, and the line indicates how the value of operations changing as time goes by.

From this Plot we can see that California provide the most operation with sales, and the trend in time scale is slightly decreasing for most states. Also, the Process Market is not working good for all the states.

Sales of Organic Strawberry

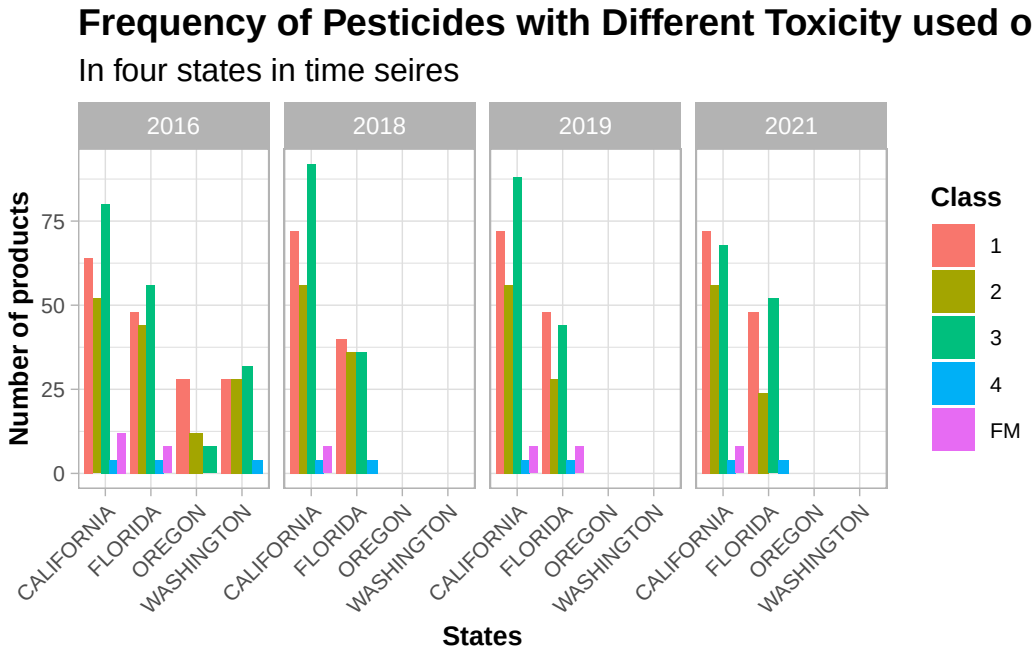


Here I focus on the Profit of Sales of organic strawberries, and I focus only on those states that created over 100,000 $ profit of saling organic strawberry. Just like the previous graph, I observe the data in the scales of 3 markets.

From this Plot we can see that California provide the most sales profit in $, and the trend in time scale is increasing for most states. However, still, the Process market is not doing well.

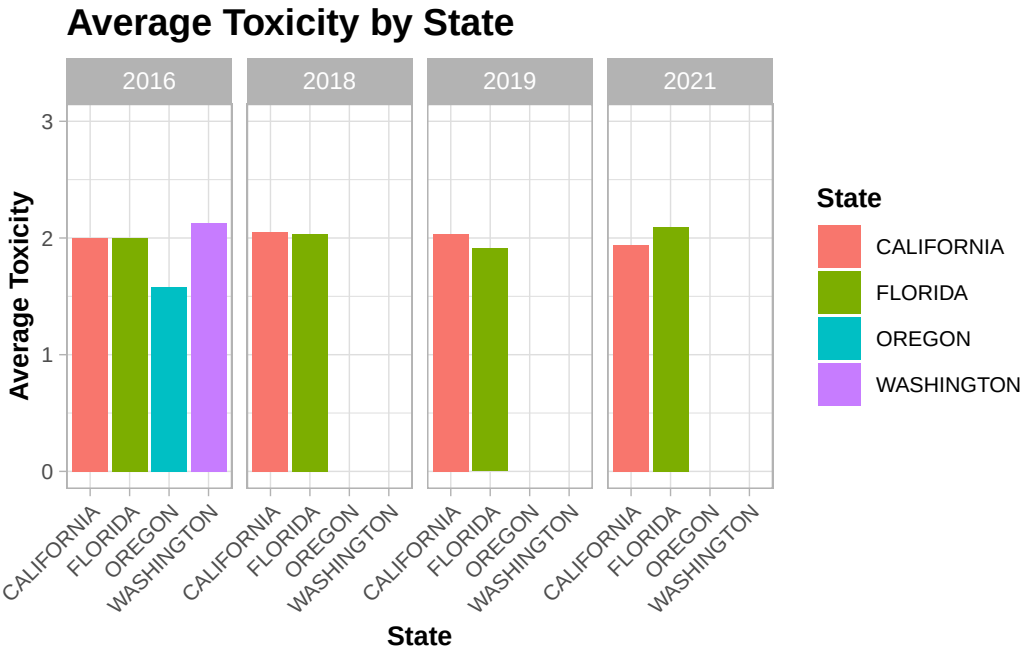
Survey Data

Toxicity



The class represents the toxicity of pesticide, with 1 meaning minimally toxicity and 4 meaning severe toxicity. FM here means Fumigant. We can observe that in the past 2016, 2018, 2019 years, California used "level 3" moderate toxicity on the largest proportion of strawberry product, and on average, the other three states used moderate toxicity pesticides most frequently.

Average Toxicity



Here I calculate the average toxicity of non-organic strawberries in different states in 2016, 2018, 2019, 2021. If we just look at 2016 data, Oregon has the least toxic strawberry. However, since we are missing the data of Oregon in the following three years, we cannot get a conclusion that Oregon has the healthiest strawberry. Comparing the strawberry from California and Florida, we can say the toxicity in these two regions is similar.