Raspberries Document

Jenna Moscaritolo

10/18/2020

Project Goal

The goal of this project is to further analyze raspberries grown in the years 2015 to 2019 in states categorized by the state they are grown in. There will also be an analysis of fungicides, herbicides, insecticides, and fertilizers (the chemical applications the farmers utilize). The main data that will be looked at for this assignment is the weight of the raspberries in pounds (lbs).

Procedure

1. Data Cleaning

For the data cleaning section, the outputs are the resulting columns from each step.

Step 1.1 Reading the data

The data that was collected is from the United States Department of Agriculture (USDA) database selector.

The specific dataset used in this project is about berry collections throughout the US.

```
##  [1] "Program"          "Year"             "Period"           "Week Ending"
##  [5] "Geo Level"        "State"            "State ANSI"       "Ag District"
##  [9] "Ag District Code" "County"           "County ANSI"      "Zip Code"
## [13] "Region"           "watershed_code"   "Watershed"        "Commodity"
## [17] "Data Item"        "Domain"           "Domain Category"  "Value"
## [21] "CV (%)"
```

Step 1.2. Removing the columns with NAs

There are spots in the data that are filled with "NA" meaning that there was no data recorded for that specific point in time. This could be due to a number of things including a lack of funding or technology malfunctions. Now, the dataset is ready to be utilized without being stopped by the NAs.

```
##  [1] "Program"         "Year"            "Period"          "Geo Level"
##  [5] "State"           "State ANSI"      "watershed_code"  "Commodity"
##  [9] "Data Item"       "Domain"          "Domain Category" "Value"
```

Step 1.3. Removing the columns with one outcome

For instance, the column called "Program" only has outputs of "SURVEY" which does not give us new information necessary for the analysis. Also, the column "State ANSI" is a copy of the column "State" so this is removed as well.

```
## [1] "Year"            "Period"          "State"           "Commodity"
## [5] "Data Item"       "Domain"          "Domain Category" "Value"
```

Step 1.4. Pulling the raspberry data

For this project, the only data at interest is the raspberry data with the column "Period" only consisting of the term "YEAR". So, we take out the data with strawberries, blueberries, or any other berry that is not raspberries along with removing the "MARKETING YEAR" data.

```
## [1] "Year"            "State"           "Data Item"       "Domain"
## [5] "Domain Category" "Value"
```

Step 1.5. Filtering out not unique rows in column "Domain"

There are many rows with information that will not help further analyze the topic and will therefore not be used. So, it is okay to remove that information for this project.

```
## [1] "TOTAL"                 "CHEMICAL, FUNGICIDE"   "CHEMICAL, HERBICIDE"
## [4] "CHEMICAL, INSECTICIDE" "CHEMICAL, OTHER"       "FERTILIZER"
```

```
##  [1] "RASPBERRIES - ACRES HARVESTED"
##  [2] "RASPBERRIES - PRODUCTION, MEASURED IN LB"
##  [3] "RASPBERRIES - YIELD, MEASURED IN LB / ACRE"
##  [4] "RASPBERRIES, FRESH MARKET - PRODUCTION, MEASURED IN $"
##  [5] "RASPBERRIES, FRESH MARKET - PRODUCTION, MEASURED IN LB"
##  [6] "RASPBERRIES, NOT SOLD - PRODUCTION, MEASURED IN LB"
##  [7] "RASPBERRIES, PROCESSING - PRODUCTION, MEASURED IN $"
##  [8] "RASPBERRIES, PROCESSING - PRODUCTION, MEASURED IN LB"
##  [9] "RASPBERRIES, UTILIZED - PRODUCTION, MEASURED IN $"
## [10] "RASPBERRIES, UTILIZED - PRODUCTION, MEASURED IN LB"
## [11] "RASPBERRIES, BLACK - ACRES HARVESTED"
## [12] "RASPBERRIES, BLACK - PRODUCTION, MEASURED IN LB"
## [13] "RASPBERRIES, BLACK, FRESH MARKET - PRODUCTION, MEASURED IN $"
## [14] "RASPBERRIES, BLACK, FRESH MARKET - PRODUCTION, MEASURED IN LB"
## [15] "RASPBERRIES, BLACK, NOT SOLD - PRODUCTION, MEASURED IN LB"
## [16] "RASPBERRIES, BLACK, PROCESSING - PRODUCTION, MEASURED IN $"
## [17] "RASPBERRIES, BLACK, PROCESSING - PRODUCTION, MEASURED IN LB"
## [18] "RASPBERRIES, BLACK, UTILIZED - PRODUCTION, MEASURED IN $"
## [19] "RASPBERRIES, BLACK, UTILIZED - PRODUCTION, MEASURED IN LB"
## [20] "RASPBERRIES, BLACK, UTILIZED - YIELD, MEASURED IN LB / ACRE"
## [21] "RASPBERRIES, RED - ACRES HARVESTED"
## [22] "RASPBERRIES, RED - PRODUCTION, MEASURED IN LB"
## [23] "RASPBERRIES, RED, FRESH MARKET - PRODUCTION, MEASURED IN $"
## [24] "RASPBERRIES, RED, FRESH MARKET - PRODUCTION, MEASURED IN LB"
## [25] "RASPBERRIES, RED, NOT SOLD - PRODUCTION, MEASURED IN LB"
## [26] "RASPBERRIES, RED, PROCESSING - PRODUCTION, MEASURED IN $"
## [27] "RASPBERRIES, RED, PROCESSING - PRODUCTION, MEASURED IN LB"
## [28] "RASPBERRIES, RED, UTILIZED - PRODUCTION, MEASURED IN $"
## [29] "RASPBERRIES, RED, UTILIZED - PRODUCTION, MEASURED IN LB"
## [30] "RASPBERRIES, RED, UTILIZED - YIELD, MEASURED IN LB / ACRE"
## [31] "RASPBERRIES, UTILIZED - YIELD, MEASURED IN LB / ACRE"
## [32] "RASPBERRIES, BLACK, NOT HARVESTED - PRODUCTION, MEASURED IN LB"
## [33] "RASPBERRIES, NOT HARVESTED - PRODUCTION, MEASURED IN LB"
## [34] "RASPBERRIES, RED, NOT HARVESTED - PRODUCTION, MEASURED IN LB"
```

```
## [1] "NOT SPECIFIED"
```

Step 1.6. Cleaning the column data

Here, there are many columns that have a lot of information in them that would be better separated. An example is how "Data Items" begins ever line with "RASPBERRIES, BEARING - . . . " so, the parts before and after the "-" will be separated.

2. New Data Frame

After all of the data cleaning, this table below is the remaining data that is needed to further the project.

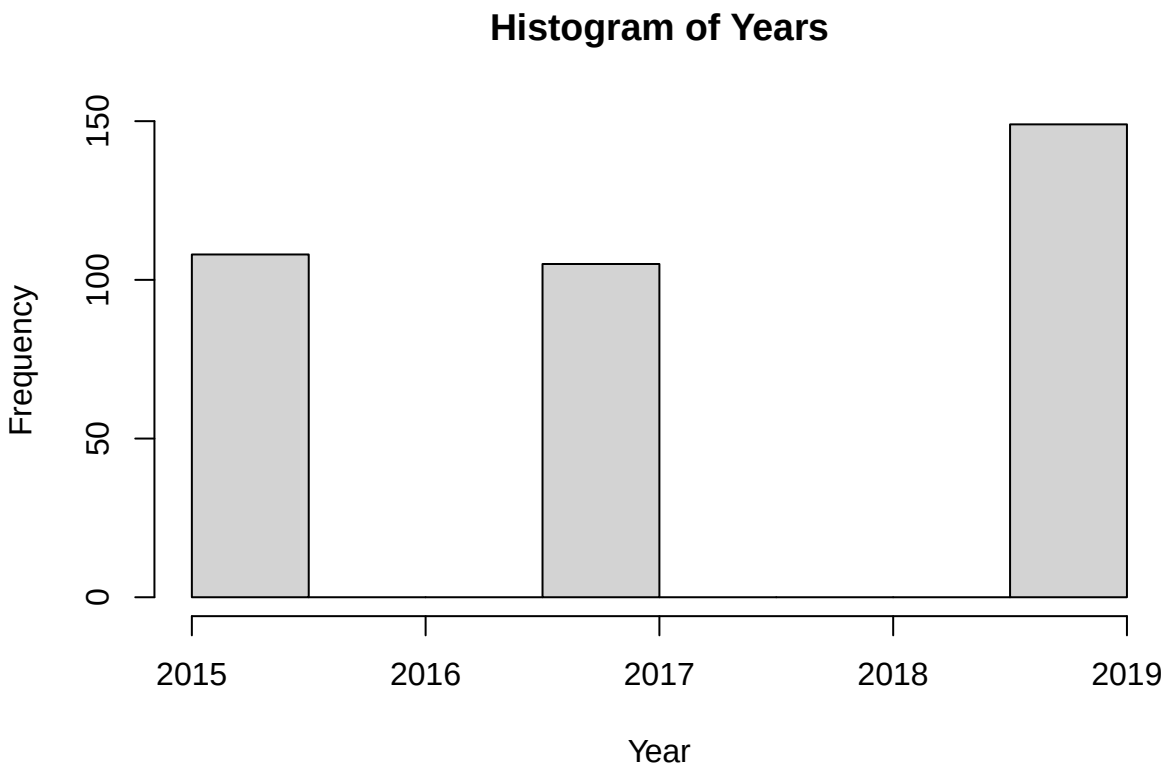
Year	State	Production	Measurement	Chemicals
2019	CALIFORNIA	APPLICATIONS	MEASURED IN LB	(AZOXYSTROBIN = 128810)
2019	CALIFORNIA	APPLICATIONS	MEASURED IN LB	(BACILLUS AMYLOLIQUEFACIEN
2019	CALIFORNIA	APPLICATIONS	MEASURED IN LB	(BACILLUS SUBTILIS = 6479)
2019	CALIFORNIA	APPLICATIONS	MEASURED IN LB	(BORAX DECAHYDRATE = 11102
2019	CALIFORNIA	APPLICATIONS	MEASURED IN LB	(BOSCALID = 128008)
2019	CALIFORNIA	APPLICATIONS	MEASURED IN LB	(CALCIUM POLYSULFIDE = 76702

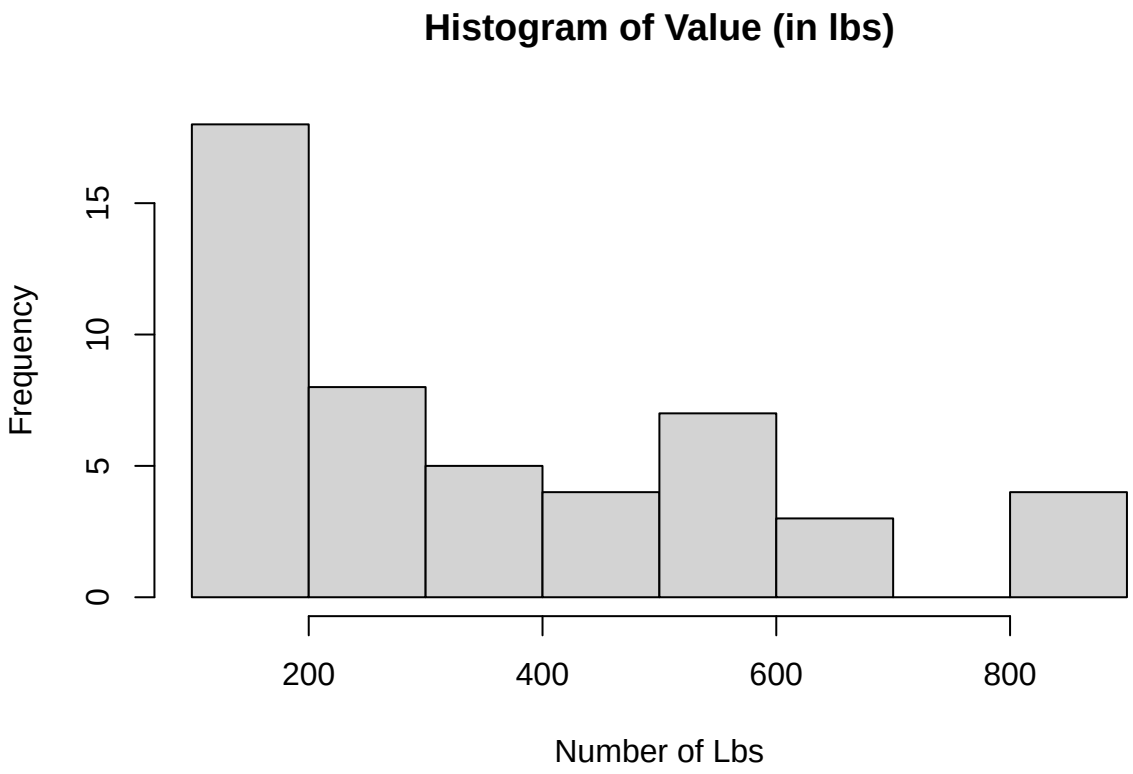
Exploratory Data Analysis

The goal of this project is to explore raspberries grown in the years 2015 to 2019 in states categorized by the state they are grown in. With this, there is no question to answer but rather to practice techniques in data cleaning and presentations. So, this project will be plotting visually and interpreting that information.

3. Histograms

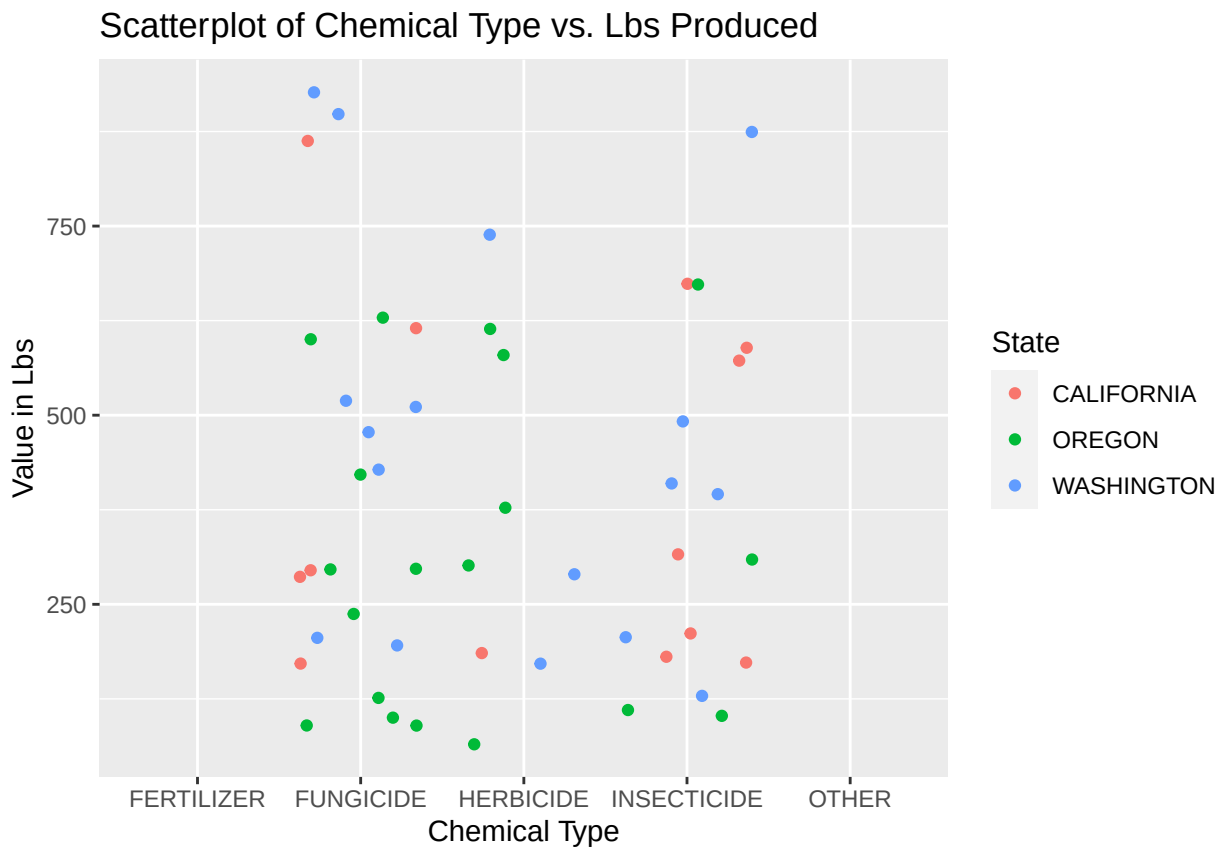
Here are two histograms: one looking at the years and one presenting the number of pounds.


Histogram of Years
Frequency
0
50
100
150
2015
2016
2017
2018
2019
Year



4. Scatterplot

Here is a scatterplot showing the relationship between the chemical type and the pounds produced and color coding the data points by state. The geom_jitter() function ensures that points will not overlap and thus not being visible. Also, this project will use points from California, Oregon, and Washington because they have the most usable data.



5. Calculations

In these three tables below, the averages of the number of pounds of raspberries for each state, year, and chemical type.

State	mean(Value, na.rm = TRUE)
CALIFORNIA	407.6923
OREGON	315.7895
WASHINGTON	458.8235

Year	mean(Value, na.rm = TRUE)
2015	406.6667
2017	388.8889
2019	375.0000

Chemical_Type	mean(Value, na.rm = TRUE)
FERTILIZER	NaN
FUNGICIDE	400.0000
HERBICIDE	377.7778
INSECTICIDE	382.3529
OTHER	NaN

Conclusion

This project requires a lot less information than the dataset initially had. So, many columns were removed and all of the NAs were filtered out as they were not able to help further the analyses. Information like Watershed Code and Week Ending were not necessary for this analysis so they were removed from the subset.

The data from 2016 and 2018 does not have as many points as the other years so it is difficult to make any predictions without assuming multiple unknown variables such as climate, how often the farmer tends to the berries, etc. The Histogram of Years easily proves this statement. Looking at the Histogram of Value (in lbs), the data is right-skewed meaning that the raspberries were, most of the time, at a lower weight.

In the plot called Scatterplot of Chemical Type vs. Lbs Produced, it shows that there were not many data points for the fertilizer and other categories of Chemical Type. So, let's look at the other three types: fungicide, herbicide, and insecticide. It seems that more points, but not that many points are for herbicides either. Herbicides are almost used only by Oregon and Washington farmers whereas those two states and California use fungicide and insecticide often. In terms of weight, Oregon generally has lighter raspberries and Washington has more middle to heavy weights. California is fairly average but also does not have as many data points as the other two states have.

For each state, the last three tables show the weight in lbs, the year, and the chemical type of the raspberries collected. The two tables with Year and Chemical Type did not have that much of a change in the averages of pounds. However, for the table with State, Washington had the heaviest raspberries while Oregon has the lighted raspberries by far.

References

1. Bruce Cowles, Anna Cook. Personalized advice in meetings.
2. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
3. Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra
4. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
5. Stefan Milton Bache and Hadley Wickham (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. https://CRAN.R-project.org/package=magrittr
6. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
7. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.