Berry Project – Strawberry Data Report

Ruxin Liu

10/13/2020

Introduction

As nowadays, more and more people are concerned about health, organic food becomes a more popular choice. In general, there is no universal way to define organic food, but since the main goal of organic agriculture is to optimize the biodiversity as well as the ecological balance, the useage of synthetic chemicals are restricted (food unfolded 2019). In this study, the berry data collected by the United States Department of Agriculture (USDA) database selector and stored online was explored and analyzed to compare the types and quantities of chemicals applied for planting strawberries during different years and in different states.

Methodology

Data Cleaning & Variable Selection

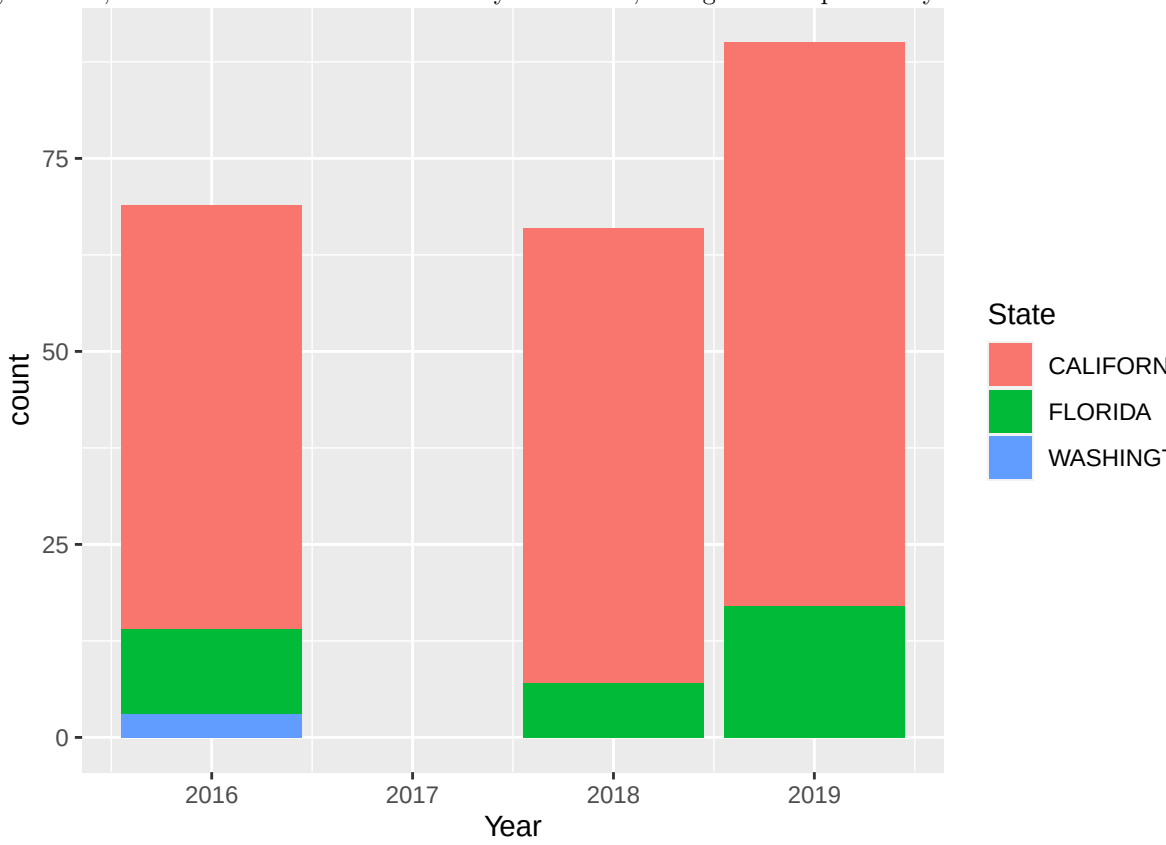
The berry dataset contains information about three kinds of berries, which are blueberries, raspberries and strawberries. In this study, only the data points related to strawberries are considered. Firstly, some of the variables only have one unique value for all the observations, which cannot provide any useful information for the purpose of comparison and can be removed. It turns out that 12 out of the 21 variables only have one unique value, such as "Program", "Week Ending", "Geo Level" and etc. After the first step, the dataset now had 8 variables and 3220 observations in total. Next, since the variable "Data Item" contains lots of different information in one single cell, it needs to be reorganized and renamed. After this step, the variable "Data Item" were separated into 4 new variables, which are "berry" (types of berry), "market" (marketing condition), "harvest" (harvest condition) and "units" (units of values collected). Similarly, the variable "Domain" was separated into "Domain_1" and "Domain_2", and the variable "Domain Category" was separated into "DC_1_left", "DC_1_right", "DC_2_left" and "DC_2_right" first. After carefully checking the unique combinations, it was found that variables "Domain_1" and "DC_1_left" produced exactly the same values. Similarly, it was found that variables "Domain_2" and "DC_2_left" produced exactly the same values. Therefore, only "Domain_1" and "Domain_2" were kept, combined and renamed to be "chemical", which states the type of chemicals being used.

After the above steps, the data still contained much information than needed and required deeper cleaning. The final strawberry data set was selected to include 8 variables, which were Year, State, market, harvest, materials, chemical, Values and units, that all could provide useful information for analysis. All the NA values and D values that are used to withhold to avoid data disclosure were removed from the variable "Value". Then, this tidy data was used to subset a smaller data set of interest, which only looked at observations where chemicals were applied. This was done by only selecting rows with harvest values equal to "APPLICATIONS" and unit values equal to "LB / ACRE / APPLICATION". Now, the smaller data set "unfood" only contained information about the pounds of chemicals used per acre in some states and years, which were all prepared to be analyzed and compared.

All the R code and data used were attached in the Appendix and also in the GitHub repository (https://github.com/Ruxinliu97/Berry).

Exploratory Data Analysis (EDA)

Based on the dataset, it is found that only California, Florida and Washington used chemicals when planting strawberries during the year 2016, 2018 and 2019. In order to see any differences, histograms are plotted by using the ggplot2 package.



Conclusion

Appendix

Code

Importing the data

```
# Load the data
library(tidyverse)
library(magrittr)
berry <- read_csv("berries.csv", col_names = TRUE)
```

```
# Look at number of unique values in each column
berry %>% summarize_all(n_distinct) -> unique
# Make a list of the columns with only one unique value
one <- which(unique[1,] == 1)
# Remove the 1-unique columns from the dataset
berry %<>% select(-all_of(one))
# State name and the State ANSI code are redundant -- keep the name only
berry %<>% select(-4)
```

```r
# Select data on Strawberries only from the berry data
strawberry <- berry %>% filter((Commodity == "STRAWBERRIES") & (Period == "YEAR"))
strawberry %<>% select(-c(Period, Commodity))
```

```r
# The original format of one specific domain type
# (CHEMICAL, INSECTICIDE: (CYFLUMETOFEN= 138831)) is not processed successfully
# with the separate function, therefore they need to be changed manually.
manual <- c(1036, 1148, 1258, 1368, 1480, 1977, 2075, 2171, 2267, 2365,
                      2469, 2542, 2613, 2684, 2757, 2985, 3019, 3052, 3085, 3120)
strawberry$`Domain Category`[manual] <- "CHEMICAL, INSECTICIDE: (CYFLUMETOFEN = 138831)"
```

```r
# Organize the column data item which contains lots of information in a cell
# unique(strawberry$`Data Item`)
strawberry %<>% separate(`Data Item`, c("s1", "s2"), sep = "-")
strawberry %<>% separate(s1, c("berry", "type"), sep = ",")
strawberry %<>% separate(s2, c("data_item", "unit"), sep = ",")
```

```r
unique(strawberry$berry)
# Since all the data selected are about strawberries, there is no need to
# contain name of berries anymore
strawberry %<>% select(-berry)
```

```r
# Organize the column Domain
strawberry %<>% separate(Domain, c("Domain_1", "Domain_2"), sep = ", ")
```

```r
# Organize the column Domain Category
strawberry %<>% separate(`Domain Category`, c("DC_1", "DC_2"), sep = ", ")
strawberry %<>% separate(DC_1, c("DC_1_left", "DC_1_right"), sep = ": ")
strawberry %<>% separate(DC_2, c("DC_2_left", "DC_2_right"), sep = ": ")
```

```r
# Check and delete the redundant variables
paste(strawberry$Domain_1, strawberry$DC_1_left) %>% unique
strawberry %<>%  select(-DC_1_left)
```

```r
# Check and delete the redundant variables
paste(strawberry$Domain_2, strawberry$DC_2_left) %>% unique
strawberry %<>%  select(-DC_2_left)
```

```r
# Change all NA values into a space
strawberry[is.na(strawberry)] <- " "
strawberry %<>% mutate(Domain_1 = "CHEMICAL", Domain_1 = "")
strawberry %<>% mutate(Chemical = str_trim(paste(Domain_1, Domain_2)))
strawberry %<>% mutate(market = str_trim(type))
strawberry %<>% mutate(harvest = str_trim(data_item))
```

```r
strawberry %<>% rename(chem_family = DC_1_right, materials = DC_2_right)
#All the rows with information on chem_family have no information on Chemical,
# so it is reasonable to combine these columns.
strawberry %<>% mutate(chemical = str_trim(paste(chem_family, Chemical)))
```

```r
strawberry %<>% separate(unit, c("u1", "u2", "u3", "u4", "u5", "u6", "u7",
                                 "u8"), sep = " ")
# Only leave the useful information after "measure in"
strawberry[is.na(strawberry)] <- " "
strawberry %<>% mutate(units = str_trim(paste(u4, u5, u6, u7, u8)))
# Select the variables that are kept in the final data
```

```
strawberry %<>% select(Year, State, market, harvest, materials, chemical,
                       Value, units)
```

```
# Check that all the variables are now tidy
# unique(strawberry$Year)
# unique(strawberry$State)
# unique(strawberry$market)
# unique(strawberry$harvest)
# unique(strawberry$chemical)
# unique(strawberry$units)
```

```
# Look at chemicals being applied to strawberry
unfood <- strawberry
unfood <- unfood %<>% filter(harvest == "APPLICATIONS")
# Remove NAs
unfood %<>% filter(Value != "(D)")
unfood %<>% filter(Value !=  "(NA)")
unfood %<>% filter(units == "LB / ACRE / APPLICATION")
unfood_1 <- unfood %>%  select(Year, State, chemical, Value)
unfood_1 %<>% pivot_wider(names_from = chemical, values_from = Value)
```

```
# Convert the characters into numeric values for further calculation and analysis
for(i in 1 : nrow(unfood_1)) {
  unfood_1$FUNGICIDE[[i]] <- as.numeric(unfood_1$FUNGICIDE[[i]])
}
for(i in 1 : nrow(unfood_1)) {
  unfood_1$HERBICIDE[[i]] <- as.numeric(unfood_1$HERBICIDE[[i]])
  unfood_1$INSECTICIDE[[i]] <- as.numeric(unfood_1$INSECTICIDE[[i]])
  unfood_1$OTHER[[i]] <- as.numeric(unfood_1$OTHER[[i]])
}
```

```
# Calculate the total count of each type of chemical usage
unfood_1 %<>% mutate(total_fungi = "NA")
unfood_1 %<>% mutate(total_herb = "NA")
unfood_1 %<>% mutate(total_insect = "NA")
unfood_1 %<>% mutate(total_other = "NA")
for(i in 1 : nrow(unfood_1)) {
  unfood_1$total_fungi[i] <- length(unfood_1$FUNGICIDE[[i]])
  unfood_1$total_herb[i] <- length(unfood_1$HERBICIDE[[i]])
  unfood_1$total_insect[i] <- length(unfood_1$INSECTICIDE[[i]])
  unfood_1$total_other[i] <- length(unfood_1$OTHER[[i]])
}
```

```
library(ggplot2)
unfood %>%
ggplot(aes(x = Year)) +
  geom_bar(aes(fill = State))
```

```
unfood %>%
ggplot(aes(x = Year)) +
  geom_bar(aes(fill = chemical))
```

Reference of R Packages

1. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43),1686, https://doi.org/10.21105/joss.01686
2. 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
3. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Bibliography

1. MA615 notes (class 14 - class 17)
2. food unfolded (2019). *What is organic food | Is it really chemical free?* [online]. Available from:https://www.foodunfolded.com/how-it-works/what-is-organic-food-is-it-really-chemical-free [accessed 18 October 2020].
3. USDA (2019). *Quick Stats* [online]. Available from: https://quickstats.nass.usda.gov/ [accessed 16 October 2020].
4. USDA (2019). *Quick Stats* [online]. Available from: https://quickstats.nass.usda.gov/results/D416E96E-3D5C-324C-9334-1D38DF88FFF1 [accessed 16 October 2020].