Berries Project-Data cleaning

Haoyu Li

10/18/2020

```
library(knitr)
library(tidyverse)
library(magrittr)
library(kableExtra)
```

#Overview

Every year, the United States Department of Agriculture collect enormous amount of data for different kinds of agriculture products, including animals, crops, and fruits etc. These data have been collected for agriculture research purpose. Thus, it is important for research team not just only to collect data, but to have the ability to clean, organize, and explore the data for future analysis and modeling, since the data that is collected annualy sometimes are just a bunch of data cluster, which means the data is a mess. For this project, we are going to look at the data including three major berries, and we are going to clean, organize, and explore the berries data so that these data can be used for future research purpose.

##1.Data Cleaning

This part mainly includes the process of how the data will be cleaning and what the problems we encounter with the raw data. Eventually, we provide the code to fix these problem and clean the raw data

##(1)Read data and roughly cleaning

The berries data can be found from the USDA database selector : http://quickstats.nass.usda.gov

```
ag_data <- read_csv("berries.csv", col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   Year = col_double(),
##   `Week Ending` = col_logical(),
##   `State ANSI` = col_double(),
##   `Ag District` = col_logical(),
##   `Ag District Code` = col_logical(),
##   County = col_logical(),
##   `County ANSI` = col_logical(),
##   `Zip Code` = col_logical(),
##   Region = col_logical(),
##   watershed_code = col_double(),
##   Watershed = col_logical(),
##   `CV (%)` = col_logical()
## )
```

```
## See spec(...) for full column specifications.
```

```r
head(ag_data,n=20)
```

```
## # A tibble: 20 x 21
##    Program  Year Period `Week Ending` `Geo Level` State `State ANSI`
##    <chr>   <dbl> <chr>  <lgl>         <chr>       <chr>        <dbl>
##  1 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
##  2 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
##  3 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
##  4 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
##  5 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
##  6 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
##  7 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
##  8 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
##  9 SURVEY   2019 MARKE~ NA            STATE       CALI~            6
## 10 SURVEY   2019 MARKE~ NA            STATE       FLOR~           12
## 11 SURVEY   2019 MARKE~ NA            STATE       FLOR~           12
## 12 SURVEY   2019 MARKE~ NA            STATE       FLOR~           12
## 13 SURVEY   2019 MARKE~ NA            STATE       FLOR~           12
## 14 SURVEY   2019 MARKE~ NA            STATE       FLOR~           12
## 15 SURVEY   2019 MARKE~ NA            STATE       FLOR~           12
## 16 SURVEY   2019 MARKE~ NA            STATE       GEOR~           13
## 17 SURVEY   2019 MARKE~ NA            STATE       GEOR~           13
## 18 SURVEY   2019 MARKE~ NA            STATE       GEOR~           13
## 19 SURVEY   2019 MARKE~ NA            STATE       MAINE           23
## 20 SURVEY   2019 MARKE~ NA            STATE       MAINE           23
## # ... with 14 more variables: `Ag District` <lgl>, `Ag District Code` <lgl>,
## #   County <lgl>, `County ANSI` <lgl>, `Zip Code` <lgl>, Region <lgl>,
## #   watershed_code <dbl>, Watershed <lgl>, Commodity <chr>, `Data Item` <chr>,
## #   Domain <chr>, `Domain Category` <chr>, Value <chr>, `CV (%)` <lgl>
```

-By looking at the first 20 rows of data, we can have a feeling that some columns may just have one values, so for these columns, we can disgard them since they may do not have effect or relation with other variables.

```r
#look at how many unique values in each column
aa<-ag_data %>% summarise_all(n_distinct)
print(aa)
```

```
## # A tibble: 1 x 21
##   Program  Year Period `Week Ending` `Geo Level` State `State ANSI`
##     <int> <int>  <int>         <int>       <int> <int>        <int>
## 1       1     5      3             1           1    18           18
## # ... with 14 more variables: `Ag District` <int>, `Ag District Code` <int>,
## #   County <int>, `County ANSI` <int>, `Zip Code` <int>, Region <int>,
## #   watershed_code <int>, Watershed <int>, Commodity <int>, `Data Item` <int>,
## #   Domain <int>, `Domain Category` <int>, Value <int>, `CV (%)` <int>
```

-Now we select the columns with only

-We can also view the column names which only have one unique value

```r
#make a list of the columns which has only one value
bb<-which(aa[1,]==1)
#columns that only have one unique value
cn<-colnames(ag_data)[bb]
print(cn)
```

```
##  [1] "Program"          "Week Ending"      "Geo Level"        "Ag District"
```

```
##  [5] "Ag District Code" "County"           "County ANSI"      "Zip Code"
##  [9] "Region"           "watershed_code"   "Watershed"        "CV (%)"
```

-Now we delete these columns that we list above

```
ag_data %<>% select(-all_of(bb))
head(ag_data,n=20)
```

```
## # A tibble: 20 x 9
##     Year Period State `State ANSI` Commodity `Data Item` Domain `Domain Categor~
##    <dbl> <chr>  <chr>        <dbl> <chr>     <chr>       <chr>  <chr>
##  1  2019 MARKE~ CALI~            6 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
##  2  2019 MARKE~ CALI~            6 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
##  3  2019 MARKE~ CALI~            6 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
##  4  2019 MARKE~ CALI~            6 RASPBERR~ RASPBERRIE~ TOTAL  NOT SPECIFIED
##  5  2019 MARKE~ CALI~            6 RASPBERR~ RASPBERRIE~ TOTAL  NOT SPECIFIED
##  6  2019 MARKE~ CALI~            6 RASPBERR~ RASPBERRIE~ TOTAL  NOT SPECIFIED
##  7  2019 MARKE~ CALI~            6 STRAWBER~ STRAWBERRI~ TOTAL  NOT SPECIFIED
##  8  2019 MARKE~ CALI~            6 STRAWBER~ STRAWBERRI~ TOTAL  NOT SPECIFIED
##  9  2019 MARKE~ CALI~            6 STRAWBER~ STRAWBERRI~ TOTAL  NOT SPECIFIED
## 10  2019 MARKE~ FLOR~           12 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
## 11  2019 MARKE~ FLOR~           12 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
## 12  2019 MARKE~ FLOR~           12 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
## 13  2019 MARKE~ FLOR~           12 STRAWBER~ STRAWBERRI~ TOTAL  NOT SPECIFIED
## 14  2019 MARKE~ FLOR~           12 STRAWBER~ STRAWBERRI~ TOTAL  NOT SPECIFIED
## 15  2019 MARKE~ FLOR~           12 STRAWBER~ STRAWBERRI~ TOTAL  NOT SPECIFIED
## 16  2019 MARKE~ GEOR~           13 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
## 17  2019 MARKE~ GEOR~           13 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
## 18  2019 MARKE~ GEOR~           13 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
## 19  2019 MARKE~ MAINE           23 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
## 20  2019 MARKE~ MAINE           23 BLUEBERR~ BLUEBERRIE~ TOTAL  NOT SPECIFIED
## # ... with 1 more variable: Value <chr>
```

-We view the selected ag_data find out that "State" and "State ANSI" are the same thing, so we get rid one of them

```
ag_data%<>%select(-`State ANSI`)
```

-Now we have initially cleaned our data

```
kable(head(ag_data)) %>%
  kable_styling(font_size=8)
```

Year	Period	State	Commodity	Data Item
2019	MARKETING YEAR	CALIFORNIA	BLUEBERRIES	BLUEBERRIES, TAME - PRICE RECEIVED, MEASURED IN $ / LB
2019	MARKETING YEAR	CALIFORNIA	BLUEBERRIES	BLUEBERRIES, TAME, FRESH MARKET - PRICE RECEIVED, ME
2019	MARKETING YEAR	CALIFORNIA	BLUEBERRIES	BLUEBERRIES, TAME, PROCESSING - PRICE RECEIVED, MEASU
2019	MARKETING YEAR	CALIFORNIA	RASPBERRIES	RASPBERRIES - PRICE RECEIVED, MEASURED IN $ / LB
2019	MARKETING YEAR	CALIFORNIA	RASPBERRIES	RASPBERRIES, FRESH MARKET - PRICE RECEIVED, MEASUREI
2019	MARKETING YEAR	CALIFORNIA	RASPBERRIES	RASPBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN

##(2) Target specific columns

-We have roughly clean the ag_data and shrink it to 8 columns. But when we look closely, there are some columns containing more than one character information, which is not good enough for the data analysis. So now we need to focus on these columns and try to separate the information in these columns

-One of the important purposes of this berries project is to summarize the major kinds of berries data.

-For the simplicity of this project, we choose one of the berries that we want to summarize. Since all of the berries data structure will be the same, if we clean and organize one of the berries data, we should be able to deal with other berries data.

-First we look at how many berries we have

```
berry <- unique(ag_data$Commodity)
print(berry)
```

```
## [1] "BLUEBERRIES"  "RASPBERRIES"  "STRAWBERRIES"
```

##Strawberries

-we choose one of the berries: Strawberries, and the period only equals to "Year" will be included.

```
strawberry <- ag_data %>% filter((Commodity=="STRAWBERRIES") & (Period=="YEAR"))
```

-Now we can delete "Commodity" and"Period" to make the strawberry data more tidy.

```
strawberry %<>% select(-c(Period, Commodity))
head(strawberry,n=20)
```

```
## # A tibble: 20 x 6
##     Year State  `Data Item`          Domain    `Domain Category`          Value
##    <dbl> <chr>  <chr>                <chr>     <chr>                      <chr>
##  1  2019 CALIF~ STRAWBERRIES - ACRE~ TOTAL     NOT SPECIFIED              35,400
##  2  2019 CALIF~ STRAWBERRIES - ACRE~ TOTAL     NOT SPECIFIED              36,000
##  3  2019 CALIF~ STRAWBERRIES - PROD~ TOTAL     NOT SPECIFIED              2,221~
##  4  2019 CALIF~ STRAWBERRIES - PROD~ TOTAL     NOT SPECIFIED              20,50~
##  5  2019 CALIF~ STRAWBERRIES - YIEL~ TOTAL     NOT SPECIFIED              580
##  6  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (AZO~ 5,500
##  7  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BAC~ (NA)
##  8  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BAC~ (NA)
##  9  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BAC~ (NA)
## 10  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BAC~ (NA)
## 11  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BAC~ (NA)
## 12  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BLA~ 1,200
## 13  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BOR~ 300
## 14  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BOS~ 6,200
## 15  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (BT ~ (NA)
## 16  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (CAP~ 283,5~
## 17  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (COP~ (D)
## 18  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (COP~ (D)
## 19  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (CYF~ 400
## 20  2019 CALIF~ STRAWBERRIES, BEARI~ CHEMICAL~ CHEMICAL, FUNGICIDE: (CYP~ 16,600
```

##Dealing with Data item

-Next we clean the "Data item" column since it contains multiple values -We separate 'Data item' into several different columns

```
#Revise the first 5 rows in 'Data item' so their formats are consistent with other rows.
strawberry$`Data Item`[1]<-str_replace(strawberry$`Data Item`[1],"STRAWBERRIES","STRAWBERRIES, TAME")
strawberry$`Data Item`[2]<-str_replace(strawberry$`Data Item`[2],"STRAWBERRIES","STRAWBERRIES, TAME")
strawberry$`Data Item`[3]<-str_replace(strawberry$`Data Item`[3],"STRAWBERRIES","STRAWBERRIES, TAME")
strawberry$`Data Item`[4]<-str_replace(strawberry$`Data Item`[4],"STRAWBERRIES","STRAWBERRIES, TAME")
strawberry$`Data Item`[5]<-str_replace(strawberry$`Data Item`[5],"STRAWBERRIES","STRAWBERRIES, TAME")

strawberry %<>% separate(`Data Item`, c("S","type", "meas", "what"), sep = ",")
```

```
## Warning: Expected 4 pieces. Missing pieces filled with `NA` in 890 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
strawberry %<>% select(-S)
head(strawberry,n=20)
```

```
## # A tibble: 20 x 8
##     Year State  type      meas     what  Domain  `Domain Category`        Value
##    <dbl> <chr>  <chr>     <chr>    <chr> <chr>   <chr>                    <chr>
##  1  2019 CALIF~ " TAME --  <NA>    <NA>  TOTAL   NOT SPECIFIED            35,400
##  2  2019 CALIF~ " TAME --  <NA>    <NA>  TOTAL   NOT SPECIFIED            36,000
##  3  2019 CALIF~ " TAME -- " MEASU~ <NA>  TOTAL   NOT SPECIFIED            2,221~
##  4  2019 CALIF~ " TAME -- " MEASU~ <NA>  TOTAL   NOT SPECIFIED            20,50~
##  5  2019 CALIF~ " TAME -- " MEASU~ <NA>  TOTAL   NOT SPECIFIED            580
##  6  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (A~ 5,500
##  7  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ (NA)
##  8  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ (NA)
##  9  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ (NA)
## 10  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ (NA)
## 11  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ (NA)
## 12  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ 1,200
## 13  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ 300
## 14  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ 6,200
## 15  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (B~ (NA)
## 16  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (C~ 283,5~
## 17  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (C~ (D)
## 18  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (C~ (D)
## 19  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (C~ 400
## 20  2019 CALIF~ " BEARIN~ " MEASU~ <NA>  CHEMIC~ CHEMICAL, FUNGICIDE: (C~ 16,600
```

-We have a new problem, that is– 'type' column has multiple value, and we need to separate them.

```
#make a dataframe to separate the multiple characters in 'type'
ty_1<- str_split(strawberry$type, " ", simplify=TRUE)
head(ty_1, n=20)
```

```
##       [,1] [,2]      [,3] [,4]           [,5]        [,6]
##  [1,] ""   "TAME"    "-"  "ACRES"        "HARVESTED" ""
##  [2,] ""   "TAME"    "-"  "ACRES"        "PLANTED"   ""
##  [3,] ""   "TAME"    "-"  "PRODUCTION"   ""          ""
##  [4,] ""   "TAME"    "-"  "PRODUCTION"   ""          ""
##  [5,] ""   "TAME"    "-"  "YIELD"        ""          ""
##  [6,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
##  [7,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
##  [8,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
##  [9,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [10,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [11,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [12,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [13,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [14,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [15,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [16,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [17,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [18,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
## [19,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
```

```
## [20,] ""   "BEARING" "-"  "APPLICATIONS" ""          ""
```

```
strawberry %<>% separate(type,c("s1", "type", "b1","lab1", "lab2","lab3"), " ")
```

```
## Warning: Expected 6 pieces. Missing pieces filled with `NA` in 3119 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
head(strawberry,n=20)
```

```
## # A tibble: 20 x 13
##     Year State s1    type  b1    lab1  lab2  lab3  meas  what  Domain
##    <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1  2019 CALI~ ""    TAME  -     ACRES HARV~ <NA>   <NA> <NA>  TOTAL
##  2  2019 CALI~ ""    TAME  -     ACRES PLAN~ <NA>   <NA> <NA>  TOTAL
##  3  2019 CALI~ ""    TAME  -     PROD~ <NA>  <NA>  " ME~ <NA>  TOTAL
##  4  2019 CALI~ ""    TAME  -     PROD~ <NA>  <NA>  " ME~ <NA>  TOTAL
##  5  2019 CALI~ ""    TAME  -     YIELD <NA>  <NA>  " ME~ <NA>  TOTAL
##  6  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
##  7  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
##  8  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
##  9  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 10  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 11  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 12  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 13  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 14  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 15  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 16  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 17  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 18  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 19  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## 20  2019 CALI~ ""    BEAR~ -     APPL~ <NA>  <NA>  " ME~ <NA>  CHEMI~
## # ... with 2 more variables: `Domain Category` <chr>, Value <chr>
```

-We get rid of 's1' and 'b1' since two columns do not contain informative value -Replace 'NA' with blank

```
strawberry %<>% select(-c(s1,b1))
strawberry[is.na(strawberry)] <- " "
head(strawberry,n=20)
```

```
## # A tibble: 20 x 11
##     Year State type  lab1  lab2  lab3  meas  what  Domain `Domain Categor~ Value
##    <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>  <chr>            <chr>
##  1  2019 CALI~ TAME  ACRES "HAR~ " "   " "   " "   TOTAL  NOT SPECIFIED   35,4~
##  2  2019 CALI~ TAME  ACRES "PLA~ " "   " "   " "   TOTAL  NOT SPECIFIED   36,0~
##  3  2019 CALI~ TAME  PROD~ " "   " "   " ME~ " "   TOTAL  NOT SPECIFIED   2,22~
##  4  2019 CALI~ TAME  PROD~ " "   " "   " ME~ " "   TOTAL  NOT SPECIFIED   20,5~
##  5  2019 CALI~ TAME  YIELD " "   " "   " ME~ " "   TOTAL  NOT SPECIFIED   580
##  6  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ 5,500
##  7  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ (NA)
##  8  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ (NA)
##  9  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ (NA)
## 10  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ (NA)
## 11  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ (NA)
## 12  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ 1,200
## 13  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ 300
## 14  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ 6,200
```

```
## 15  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ (NA)
## 16  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ 283,~
## 17  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ (D)
## 18  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ (D)
## 19  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ 400
## 20  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " "   CHEMI~ CHEMICAL, FUNGI~ 16,6~
```

##Dealing with Domain

-Another column we need to deal with is 'Domain' -After we separate Domain into different columns, we replace 'NA' with blank entries

```
strawberry %<>% separate(Domain, c("Domain_left", "Domain_right"), sep = ", ")
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 433 rows [1, 2,
## 3, 4, 5, 126, 127, 128, 129, 246, 247, 248, 249, 366, 367, 368, 369, 486, 487,
## 488, ...].
```

```
strawberry[is.na(strawberry)] <- " "
```

##Dealing with Domain category

-We need to separate 'Domain category' as well

```
strawberry %<>% separate(`Domain Category`, c("DC_left", "DC_right"), sep = ", ")
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 30 rows [90, 212,
## 332, 452, 574, 652, 653, 720, 721, 785, 786, 850, 851, 916, 917, 1052, 1164,
## 1274, 1384, 1496, ...].
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 433 rows [1, 2,
## 3, 4, 5, 126, 127, 128, 129, 246, 247, 248, 249, 366, 367, 368, 369, 486, 487,
## 488, ...].
```

```
strawberry$DC_left%>%unique()
```

```
## [1] "NOT SPECIFIED"           "CHEMICAL"
## [3] "FERTILIZER: (NITROGEN)"  "FERTILIZER: (PHOSPHATE)"
## [5] "FERTILIZER: (POTASH)"    "FERTILIZER: (SULFUR)"
```

-After we check the unique value of 'DC_left' and DC_right', we found out that 'DC_left' also contain multiple values. Thus, we continue separate 'DC_left' and 'DC_right'.

-After we separate the two columns, replace all 'NA' value

```
strawberry %<>% separate(DC_left, c("DC_left_l", "DC_left_r"), sep = ": ")
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 3145 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
strawberry %<>% separate(DC_right, c("DC_right_l", "DC_right_r"), sep = ": ")
strawberry[is.na(strawberry)] <- " "
head(strawberry,n=20)
```

```
## # A tibble: 20 x 15
##     Year State type  lab1  lab2  lab3  meas  what  Domain_left Domain_right
##    <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>       <chr>
##  1  2019 CALI~ TAME  ACRES "HAR~ " "   " "   " "   TOTAL       " "
##  2  2019 CALI~ TAME  ACRES "PLA~ " "   " "   " "   TOTAL       " "
##  3  2019 CALI~ TAME  PROD~ " "   " "   " ME~ " "   TOTAL       " "
##  4  2019 CALI~ TAME  PROD~ " "   " "   " ME~ " "   TOTAL       " "
```

```
##  5  2019 CALI~ TAME  YIELD " "   " "   " ME~ " " TOTAL       " "
##  6  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
##  7  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
##  8  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
##  9  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 10  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 11  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 12  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 13  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 14  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 15  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 16  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 17  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 18  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 19  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## 20  2019 CALI~ BEAR~ APPL~ " "   " "   " ME~ " " CHEMICAL    "FUNGICIDE"
## # ... with 5 more variables: DC_left_l <chr>, DC_left_r <chr>,
## #   DC_right_l <chr>, DC_right_r <chr>, Value <chr>
```

-We use code: 'paste(strawberry$Domain_left, strawberry$DC_left_l) %>% unique' to decide remove column DC_left_l

-Same code apply to 'Domain_right' and 'DC_right_l, we decide to remove DC_right_l

```
paste(strawberry$Domain_left,strawberry$DC_left_l) %>% unique
```

```
## [1] "TOTAL NOT SPECIFIED"   "CHEMICAL CHEMICAL"     "FERTILIZER FERTILIZER"
```

```
strawberry %<>%  select(-DC_left_l)
strawberry %<>%  select(-DC_right_l)
```

-Next we combine 'lab1' and 'lab2', before we combine, we test if there is overlaps of two columns and then we combine them

```
#check for lab1, lab2, lab3

##paste(strawberry$lab1, strawberry$lab2,strawberry$lab3) %>% unique()
strawberry %<>% mutate(label = paste(lab1,lab2,lab3))
```

-Also, we test for Domain_left and Domain_right, and then we combine them

```
#test for "chemical" in column Domain_left

##paste(strawberry$Domain_left, strawberry$Domain_right) %>% unique()
## remove "Chemical" and joint the columns

strawberry %<>% mutate(Domain_left = "CHEMICAL", Domain_left = "")
strawberry %<>% mutate(Chemical=paste(Domain_left, Domain_right))

strawberry %<>% select(-c(Domain_left, Domain_right))
strawberry %<>% select(Year, State, type, what, meas, label, DC_left_r, DC_right_r, Chemical, Value )
head(strawberry,n=20)
```

```
## # A tibble: 20 x 10
##     Year State  type   what  meas    label DC_left_r DC_right_r   Chemical Value
##    <dbl> <chr>  <chr>  <chr> <chr>   <chr> <chr>     <chr>        <chr>    <chr>
##  1  2019 CALIF~ TAME   " "   " "     "ACR~ " "       " "          "  "     35,4~
##  2  2019 CALIF~ TAME   " "   " "     "ACR~ " "       " "          "  "     36,0~
```

```
##  3  2019 CALIF~ TAME   " "   " MEAS~ "PRO~ " "       " "          "  "     2,22~
##  4  2019 CALIF~ TAME   " "   " MEAS~ "PRO~ " "       " "          "  "     20,5~
##  5  2019 CALIF~ TAME   " "   " MEAS~ "YIE~ " "       " "          "  "     580
##  6  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(AZOXYSTRO~ " FUNGI~ 5,500
##  7  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BACILLUS ~ " FUNGI~ (NA)
##  8  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BACILLUS ~ " FUNGI~ (NA)
##  9  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BACILLUS ~ " FUNGI~ (NA)
## 10  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BACILLUS ~ " FUNGI~ (NA)
## 11  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BACILLUS ~ " FUNGI~ (NA)
## 12  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BLAD = 30~ " FUNGI~ 1,200
## 13  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BORAX DEC~ " FUNGI~ 300
## 14  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BOSCALID ~ " FUNGI~ 6,200
## 15  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(BT SUBSP ~ " FUNGI~ (NA)
## 16  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(CAPTAN = ~ " FUNGI~ 283,~
## 17  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(COPPER HY~ " FUNGI~ (D)
## 18  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(COPPER OC~ " FUNGI~ (D)
## 19  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(CYFLUFENA~ " FUNGI~ 400
## 20  2019 CALIF~ BEARI~ " "   " MEAS~ "APP~ " "       "(CYPRODINI~ " FUNGI~ 16,6~
```

-Now the problem is that we have entries in both the "what" and "meas" columns that begin with "MEASURED IN" we need to know how many overlaps there are and combine them together

```
## in the column "what"
count_1 <- str_detect(strawberry$what, "MEASURED IN")
sum(count_1)
```

```
## [1] 59
```

```
## in the column "meas"

count_2 <- str_detect(strawberry$meas, "MEASURED IN")
sum(count_2)
```

```
## [1] 2992
```

- We need to combine them into one column.We will separate them from their current column and put them into two columns, then we will test to make sure there aren't any overlaps

-The method we use is to separate the "MEASURED IN" entries in the meas column and form an index of the entries to be separated out

```
f1 <- function(a,b){
  if(a){
    return(b)
  }else{
      return("")
  }
}

strawberry %<>% mutate(m_in_1 = unlist(map2(count_2, strawberry$meas, f1)))
```

-Now we replace "MEASURED IN.*$" value in every entry of meas so we don't need them

```
strawberry %<>% mutate(meas = str_replace(strawberry$meas, "MEASURED IN.*$", ""))
```

-We will do the same process to "what" column as well and then we combine 'm_in_1' and 'm_in_2' columns

```
strawberry %<>% mutate(m_in_2 = unlist(map2(count_1, strawberry$what, f1)))
strawberry %<>% mutate(what = str_replace(strawberry$what, "MEASURED IN.*$", ""))
```

```
strawberry %<>% mutate(units = str_trim(paste(m_in_1, m_in_2)))

#check for unique value in combined column
strawberry$units %>% unique()
```

```
##  [1] ""                                    "MEASURED IN $"
##  [3] "MEASURED IN CWT"                     "MEASURED IN CWT / ACRE"
##  [5] "MEASURED IN LB"                      "MEASURED IN LB / ACRE / APPLICATION"
##  [7] "MEASURED IN LB / ACRE / YEAR"        "MEASURED IN NUMBER"
##  [9] "MEASURED IN PCT OF AREA BEARING"     "MEASURED IN TONS"
```

##(3)Final cleaning up and organizing

The final data cleaning and organizing mainly focus on rename the variables that makes no sense and variables that can be revised in a better form that can be easily interpret

```
#test unique value in every column
#strawberry$what %>% unique()


#strawberry$meas %>% unique()


#strawberry$label %>% unique()


#strawberry$DC_left_r %>% unique()


#strawberry$DC_right_r %>% unique()


#strawberry$Value %>% unique()


#strawberry$units %>% unique()

#rename column names

strawberry %<>% rename(Avg = what)
strawberry %<>% rename(Marketing = meas, Harvest = label, Chem_family = DC_left_r, Materials = DC_right_

#colnames(strawberry)

strawberry %<>% select(Year, State, type, Marketing,
                  Measures, Avg, Harvest, Chem_family,
                  Materials, Chemical, Value )


#combine Marketing and Harvest
strawberry %<>% mutate(production = str_trim(paste(Marketing, Harvest)))


strawberry %<>% select(Year, State, type, production, Measures,
                  Avg, Chem_family, Materials, Chemical, Value)
```

```r
#combine Chem_family and Chemical
strawberry %<>% mutate(Chemical = str_trim(paste(Chem_family, Chemical)))

strawberry %<>% select(Year, State, type, production, Avg, Measures, Materials, Chemical, Value)

head(strawberry,n=20)
```

```
## # A tibble: 20 x 9
##     Year State  type   production  Avg   Measures   Materials     Chemical Value
##    <dbl> <chr>  <chr>  <chr>       <chr> <chr>      <chr>         <chr>    <chr>
##  1  2019 CALIF~ TAME   ACRES HARV~ " "   ""         " "           ""       35,4~
##  2  2019 CALIF~ TAME   ACRES PLAN~ " "   ""         " "           ""       36,0~
##  3  2019 CALIF~ TAME   PRODUCTION  " "   "MEASURED~ " "           ""       2,22~
##  4  2019 CALIF~ TAME   PRODUCTION  " "   "MEASURED~ " "           ""       20,5~
##  5  2019 CALIF~ TAME   YIELD       " "   "MEASURED~ " "           ""       580
##  6  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(AZOXYSTROB~ "FUNGIC~ 5,500
##  7  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BACILLUS A~ "FUNGIC~ (NA)
##  8  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BACILLUS A~ "FUNGIC~ (NA)
##  9  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BACILLUS P~ "FUNGIC~ (NA)
## 10  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BACILLUS S~ "FUNGIC~ (NA)
## 11  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BACILLUS S~ "FUNGIC~ (NA)
## 12  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BLAD = 300~ "FUNGIC~ 1,200
## 13  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BORAX DECA~ "FUNGIC~ 300
## 14  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BOSCALID =~ "FUNGIC~ 6,200
## 15  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(BT SUBSP K~ "FUNGIC~ (NA)
## 16  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(CAPTAN = 8~ "FUNGIC~ 283,~
## 17  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(COPPER HYD~ "FUNGIC~ (D)
## 18  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(COPPER OCT~ "FUNGIC~ (D)
## 19  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(CYFLUFENAM~ "FUNGIC~ 400
## 20  2019 CALIF~ BEARI~ APPLICATIO~ " "   "MEASURED~ "(CYPRODINIL~ "FUNGIC~ 16,6~
```