MA615 Unit 1 final

Zhaosheng-Xie

read the data

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

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

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

```
Data = berries
```

Data cleaning

```
##Remove single value
Data %<>% select(-c("Program","Geo Level","Ag District","Week Ending",8:15,21))
#Remove NA,D and process (Z)
Data <- filter(Data,Value!='(D)',Value!='(NA)')
Data$Value[which(Data$Value=='(Z)')] <- 0
```

```
##Split Data Item
new <- Data
# new <- separate(new,col = "Data Item",into = c("species", "definition", "unit"),sep = ", ",remov
e = TRUE)
# new <- separate(new,col = "Data Item",into = c("q", "w", "e", "r"),sep = ", ",remove = TRUE)
# new$unit <- tail(strsplit(new$`Data Item`,split=", ")[[1]],1)
```

```
#1.Unit
nr <- nrow(new)
for (i in 1:nr) {
  new$unit[i] <- tail(strsplit(new$`Data Item`,split=",")[[i]],1)

}
```

```
## Warning: Unknown or uninitialised column: `unit`.
```

```
new1 <- new
new2 <- new1
#Replace untidy data
new2$unit[which(new2$unit=="STRAWBERRIES - ACRES HARVESTED"|
                  new2$unit==" WILD - ACRES HARVESTED"|
                  new2$unit==" RED - ACRES HARVESTED"|
                  new2$unit==" TAME - ACRES HARVESTED"|
                  new2$unit=="RASPBERRIES - ACRES HARVESTED"|
                  new2$unit==" BLACK - ACRES HARVESTED")] <- "ACRES HARVESTED"
new2$unit[which(new2$unit=="STRAWBERRIES - ACRES PLANTED")] <- "ACRES PLANTED"
unique(new2$unit)
```

```
##  [1] " MEASURED IN $ / LB"     " MEASURED IN $ / CWT"
##  [3] "ACRES HARVESTED"         " MEASURED IN LB"
##  [5] " MEASURED IN LB / ACRE"  " MEASURED IN $"
##  [7] " AVG"                    "ACRES PLANTED"
##  [9] " MEASURED IN CWT"        " MEASURED IN CWT / ACRE"
## [11] " MEASURED IN $ / TON"    " MEASURED IN TONS"
```

```
#2.Type
new3 <- new2
new3 %<>% separate(`Data Item`, c("B","type", "meas", "what"), sep = ",",remove = FALSE)
```

```
## Warning: Expected 4 pieces. Missing pieces filled with `NA` in 1819 rows [1, 4,
## 5, 6, 7, 8, 11, 14, 17, 20, 21, 26, 27, 28, 29, 30, 33, 34, 35, 36, ...].
```

```
new3 %<>% select(-B)
new3 %<>% separate(type,c("b1", "type", "b2", "lab1", "lab2"), " ")
```

```
## Warning: Expected 5 pieces. Additional pieces discarded in 130 rows [4, 5, 7,
## 26, 28, 33, 46, 129, 533, 1767, 1822, 1823, 1824, 1828, 1829, 1844, 1846, 1847,
## 1853, 1856, ...].
```

```
## Warning: Expected 5 pieces. Missing pieces filled with `NA` in 5594 rows [2, 3,
## 9, 10, 12, 13, 15, 16, 18, 19, 22, 23, 24, 25, 31, 32, 35, 36, 37, 38, ...].
```

```
new3[is.na(new3)] <- " "  ## OK now Data Item has been split into parts
# unique(new3$type)
#I found there is something redundant about new3$type. There will be 3 types in the final data: tame, wild and bearing. So I remove others.
new3$type[which(new3$type=="MEASURED"
                  # new3$type=="FRESH"|
                  # new3$type=="PROCESSING"|
                  # new3$type=="NOT"|
                  # new3$type=="UTILIZED"|
                  # new3$type=="BLACK"|
                  # new3$type=="RED"
                  )] <- " "
new3$type[which(new3$type=="NOT")] <- "NOT SOLD"
new3$type[which(new3$type=="FRESH")] <- "FRESH MARKET"
```

```
#3.Production
#The true values are hided in columns"lab1,lab2,meas,what"
new4 <- new3
new4 %<>% select(-c(`State ANSI`,b1,b2))
#settle these 4 columns and paste into 1 column
# unique(new4$lab1)
new4$lab1[which(new4$lab1=="$"|
                new4$lab1=="-"|
                new4$lab1=="ACRES"|
                new4$lab1=="LB"|
                new4$lab1=="CWT")] <- " "
# unique(new4$lab2)
new4$lab2[which(new4$lab2=="/"|
                new4$lab2=="HARVESTED")] <- " "
# unique(new4$meas)
new4$meas[which(new4$meas==" MEASURED IN $ / LB"|
                new4$meas==" MEASURED IN LB / ACRE"|
                new4$meas==" MEASURED IN LB / ACRE / YEAR"|
                new4$meas==" MEASURED IN $"|
                new4$meas==" MEASURED IN $ / CWT"|
                new4$meas==" MEASURED IN NUMBER"|
                new4$meas==" MEASURED IN CWT"|
                new4$meas==" MEASURED IN LB"|
                new4$meas==" MEASURED IN LB / ACRE / APPLICATION"|
                new4$meas==" MEASURED IN PCT OF AREA BEARING"|
                new4$meas==" MEASURED IN $ / TON")] <- " "
new4$meas[which(new4$meas==" FRESH MARKET - PRICE RECEIVED"|
                new4$meas==" PROCESSING - PRICE RECEIVED"  )] <- "PRICE RECEIVED"
new4$meas[which(new4$meas==" FRESH MARKET - PRODUCTION"|
                new4$meas==" NOT SOLD - PRODUCTION"|
                new4$meas==" PROCESSING - PRODUCTION"|
                new4$meas==" UTILIZED - PRODUCTION")]  <- "PRODUCTION"
new4$meas[which(new4$meas==" UTILIZED - YIELD")] <- "YIELD"
# unique(new4$what)
new4 %<>% select(-what)
```

```
#combine these 3 columns
new4 %<>% mutate(production = str_trim(paste(lab1,lab2,meas)) )
#unique(new4$production)
new4$production[which(new4$production=="PRICE")] <- "PRICE RECEIVED"
new4 %<>% select(-c(lab1,lab2,meas))
#process column production
new4$production[c(4,5,7)] <- "PRICE RECEIVED"
```

```
##onto Domain
new5 <- new4
# new5$Domain %>% unique()

new5 %<>% separate(Domain, c("D_left", "D_right"), sep = ", ")
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 1791 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
# new5$D_left %>% unique()
# new5$D_right %>% unique()

new5[is.na(new5)] <- " "

## And now Domain Category


## new5$`Domain Category` %>% unique()

new5 %<>% separate(`Domain Category`, c("DC_left", "DC_right"), sep = ", ")
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 1922
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 1983
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 2042
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 2107
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 2174
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 3995
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 4052
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 4107
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 4169
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 4233
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 20 rows [184, 259,
## 332, 414, 498, 1336, 1385, 1431, 1478, 1531, 1932, 1993, 2052, 2117, 2184, 4005,
## 4062, 4117, 4179, 4243].
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 1801 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
## looks like DC_left combines labels

head(new5$DC_left %>% unique(),n=20)
```

```
## [1] "NOT SPECIFIED"
## [2] "CHEMICAL"
## [3] "FERTILIZER: (NITROGEN)"
## [4] "FERTILIZER: (PHOSPHATE)"
## [5] "FERTILIZER: (POTASH)"
## [6] "FERTILIZER: (SULFUR)"
## [7] "CHEMICAL, INSECTICIDE: (CYFLUMETOFEN<U+00A0>= 138831)"
```

```
head(new5$DC_right %>% unique(), n=20)
```

```
##  [1] NA
##  [2] "FUNGICIDE: (BOSCALID = 128008)"
##  [3] "FUNGICIDE: (CYPRODINIL = 288202)"
##  [4] "FUNGICIDE: (FLUDIOXONIL = 71503)"
##  [5] "FUNGICIDE: (MYCLOBUTANIL = 128857)"
##  [6] "FUNGICIDE: (PYRACLOSTROBIN = 99100)"
##  [7] "FUNGICIDE: (TOTAL)"
##  [8] "HERBICIDE: (TOTAL)"
##  [9] "INSECTICIDE: (ACEQUINOCYL = 6329)"
## [10] "INSECTICIDE: (BIFENAZATE = 586)"
## [11] "INSECTICIDE: (METHOXYFENOZIDE = 121027)"
## [12] "INSECTICIDE: (PYRETHRINS = 69001)"
## [13] "INSECTICIDE: (SPINETORAM = 110007)"
## [14] "INSECTICIDE: (SPINOSAD = 110003)"
## [15] "INSECTICIDE: (TOTAL)"
## [16] "INSECTICIDE: (ZETA-CYPERMETHRIN = 129064)"
## [17] "OTHER: (TOTAL)"
## [18] "INSECTICIDE: (BT KURSTAKI ABTS-351 = 6522)"
## [19] "FUNGICIDE: (AZOXYSTROBIN = 128810)"
## [20] "FUNGICIDE: (BLAD = 30006)"
```

```
## work on DC_left first

new5 %<>% separate(DC_left, c("DC_left_l", "DC_left_r"), sep = ": ")
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 1922
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 1983
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 2042
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 2107
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 2174
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 3995
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 4052
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 4107
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 4169
```

```
## Warning in gregexpr(pattern, x, perl = TRUE): PCRE error
##  'UTF-8 error: isolated byte with 0x80 bit set'
##  for element 4233
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 5781 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
## new5$DC_left_l %>% unique()
## new5$DC_left_r %>% unique()

## now work on DC_right

head(new5$DC_right %>% unique(), n=20)
```

```
##  [1] NA
##  [2] "FUNGICIDE: (BOSCALID = 128008)"
##  [3] "FUNGICIDE: (CYPRODINIL = 288202)"
##  [4] "FUNGICIDE: (FLUDIOXONIL = 71503)"
##  [5] "FUNGICIDE: (MYCLOBUTANIL = 128857)"
##  [6] "FUNGICIDE: (PYRACLOSTROBIN = 99100)"
##  [7] "FUNGICIDE: (TOTAL)"
##  [8] "HERBICIDE: (TOTAL)"
##  [9] "INSECTICIDE: (ACEQUINOCYL = 6329)"
## [10] "INSECTICIDE: (BIFENAZATE = 586)"
## [11] "INSECTICIDE: (METHOXYFENOZIDE = 121027)"
## [12] "INSECTICIDE: (PYRETHRINS = 69001)"
## [13] "INSECTICIDE: (SPINETORAM = 110007)"
## [14] "INSECTICIDE: (SPINOSAD = 110003)"
## [15] "INSECTICIDE: (TOTAL)"
## [16] "INSECTICIDE: (ZETA-CYPERMETHRIN = 129064)"
## [17] "OTHER: (TOTAL)"
## [18] "INSECTICIDE: (BT KURSTAKI ABTS-351 = 6522)"
## [19] "FUNGICIDE: (AZOXYSTROBIN = 128810)"
## [20] "FUNGICIDE: (BLAD = 30006)"
```

```
new5 %<>% separate(DC_right, c("DC_right_l", "DC_right_r"), sep = ": ")


new5[is.na(new5)] <- " "

##  OK now we need to eliminate the redundancy
## fine and remove redundant columns

## remove column new5$DC_left_l
new5 %<>%  select(-DC_left_l)

## remove column DC_right_l
new5 %<>% select(-DC_right_l)


## remove "Chemical" and joint the columns

new5 %<>% mutate(D_left = "CHEMICAL", D_left = "")

new5 %<>% mutate(Chemical=paste(D_left, D_right))

new5 %<>% select(-c(D_left, D_right))
```

```
## Final tidy data

new5$DC_left_r %>% unique() # rename chemical_family
```

```
## [1] " "           "(NITROGEN)"  "(PHOSPHATE)" "(POTASH)"    "(SULFUR)"
```

```
new5 %<>% rename( Chem_family = DC_left_r, Materials = DC_right_r)
new5 %<>% mutate(Chemical = str_trim(paste(Chem_family, Chemical)))
new5 %<>% select(-c(`Data Item`,Chem_family))
new5 %<>% rename( Type = type, Unit = unit, Production = production)
new5 %<>% select(Year,Period,State,Commodity,Type,Production,Chemical,Materials,Unit,Value)
Tidyberry <- new5
write.table(Tidyberry,"D:/MSSP/Rdata/615/Berry/Tidyberry.csv",col.names = TRUE,sep = ",")
```