# Berry Project

Xijia Luo

October 18, 2020

## Berry Data Cleaning

```
## read the data
ag_data <- read_csv("berries.csv", col_names = TRUE)

## look at number of unique values in each column
ag_data %>% summarize_all(n_distinct) -> aa


## make a list of the columns with only one unique value
bb <- which(aa[1,]==1)

## list the 1-unique valu column names
cn <- colnames(ag_data)[bb]

## remove the 1-unique columns from the dataset
ag_data %<>% select(-all_of(bb))

aa %<>% select(-all_of(bb))


## State name and the State ANSI code are (sort of) redundant
## Just keep the name
ag_data %<>% select(-4)
aa %<>% select(-4)

head(ag_data)
```

```
## # A tibble: 6 x 8
##    Year Period  State  Commodity `Data Item`       Domain `Domain Categor~ Value
##   <dbl> <chr>   <chr>  <chr>     <chr>             <chr>  <chr>            <chr>
## 1  2019 MARKET~ CALIF~ BLUEBERR~ BLUEBERRIES, TAM~ TOTAL  NOT SPECIFIED    2.85
## 2  2019 MARKET~ CALIF~ BLUEBERR~ BLUEBERRIES, TAM~ TOTAL  NOT SPECIFIED    3.56
## 3  2019 MARKET~ CALIF~ BLUEBERR~ BLUEBERRIES, TAM~ TOTAL  NOT SPECIFIED    0.29
## 4  2019 MARKET~ CALIF~ RASPBERR~ RASPBERRIES - PR~ TOTAL  NOT SPECIFIED    2.69
## 5  2019 MARKET~ CALIF~ RASPBERR~ RASPBERRIES, FRE~ TOTAL  NOT SPECIFIED    (D)
## 6  2019 MARKET~ CALIF~ RASPBERR~ RASPBERRIES, PRO~ TOTAL  NOT SPECIFIED    (D)
```

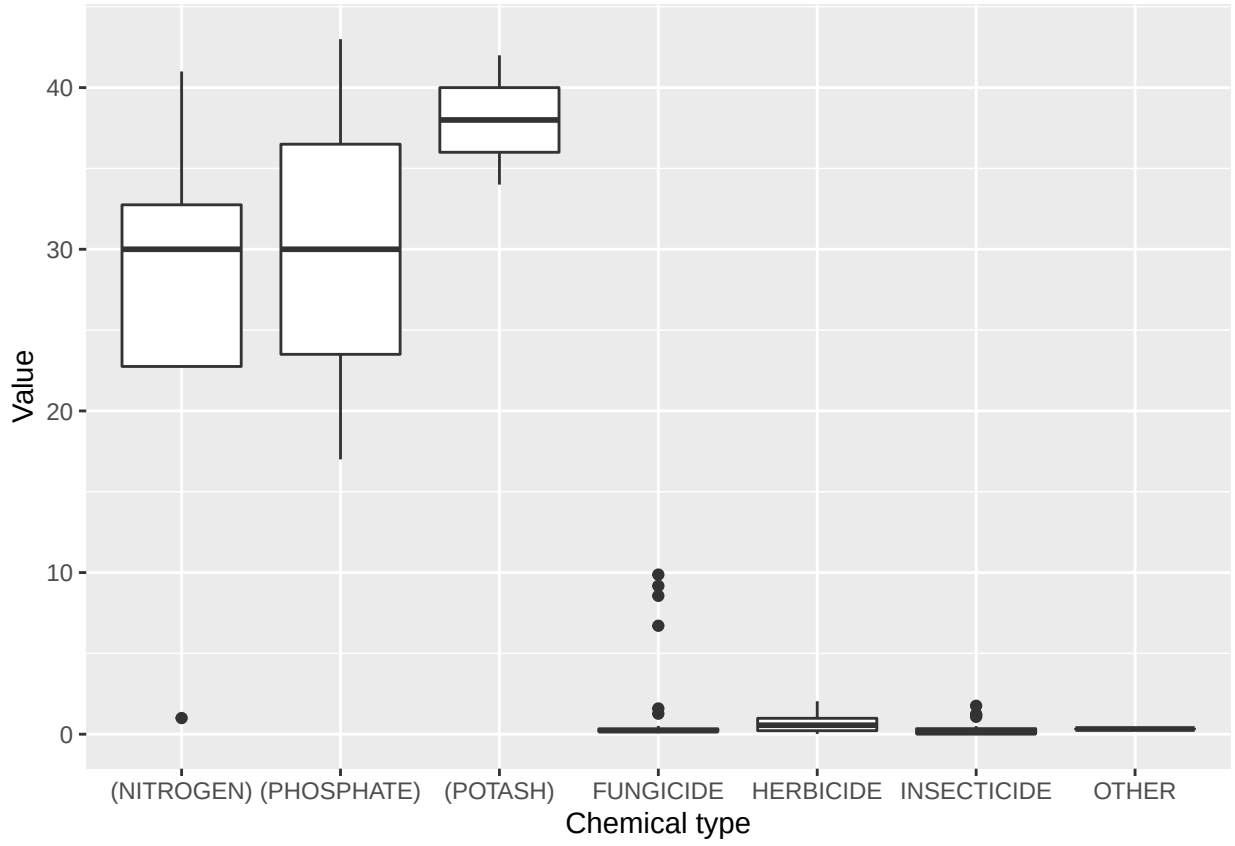
### raspberries

```
rberry <- ag_data %>% filter((Commodity=="RASPBERRIES") & (Period=="YEAR"))
rberry %<>% select(-c(Period, Commodity))

#separate Data Item
rberry %<>% separate(`Data Item`, c("B","type", "meas","what"), sep = ",")
rberry <- rberry %>% filter(B=="RASPBERRIES")

rberry %<>% select(-B)
#separate type
rberry %<>% separate(type,c("b1", "type", "b2", "lab1", "lab2"), " ")
rberry %<>% select(-c(b1,b2))
#separate domain
rberry %<>% separate(Domain, c("D_left", "D_right"), sep = ", ")
rberry %<>% separate(`Domain Category`, c("DC_left", "DC_right"), sep = ", ")
rberry %<>% separate(DC_left, c("DC_left_l", "DC_left_r"), sep = ": ")
rberry %<>% separate(DC_right, c("DC_right_l", "DC_right_r"), sep = ": ")

#remove NA
rberry[is.na(rberry)] <- " "

#remove redundant columns
rberry %<>%  select(-DC_left_l)
rberry %<>% select(-DC_right_l)
rberry %<>% mutate(label = paste(lab1,lab2))
rberry %<>% mutate(D_left = "CHEMICAL", D_left = "")
rberry %<>% mutate(Chemical=paste(D_left, D_right))
rberry %<>% select(-c(D_left, D_right))
rberry %<>% select(Year, State, type, meas,what, label, DC_left_r, DC_right_r, Chemical, Value )


index_meas <- str_detect(rberry$meas, "MEASURED IN")
rberry %<>% mutate(m_in_1 = unlist(map2(index_meas, rberry$meas, f1)))
rberry %<>% mutate(meas = str_replace(rberry$meas, "MEASURED IN.*$", ""))

index_what <- str_detect(rberry$what, "MEASURED IN")
rberry %<>% mutate(m_in_2 = unlist(map2(index_what, rberry$what, f1)))
rberry %<>% mutate(what = str_replace(rberry$what, "MEASURED IN.*$", ""))

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

rberry %<>% rename(Avg = what)
rberry %<>% rename(Marketing = meas, Harvest = label, Chem_family = DC_left_r, Materials = DC_right_r, I
rberry %<>% select(Year, State, type, Marketing,
                   Measures, Avg, Harvest, Chem_family,
                   Materials, Chemical, Value )

rberry %<>% mutate(production = str_trim(paste(Marketing, Harvest)))

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

rberry %<>% mutate(Chemical = str_trim(paste(Chem_family, Chemical)))
```

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


unfoodr <- rberry %<>% filter(production=="APPLICATIONS")


unfoodr %<>% filter(Value != "(D)")

unfoodr %<>% filter(Value !=  "(NA)")

unfoodr %<>% filter(Measures == "MEASURED IN LB / ACRE / APPLICATION")

unfoodr_1 <- unfoodr %>%  select(Year, State, Chemical, Value)
```

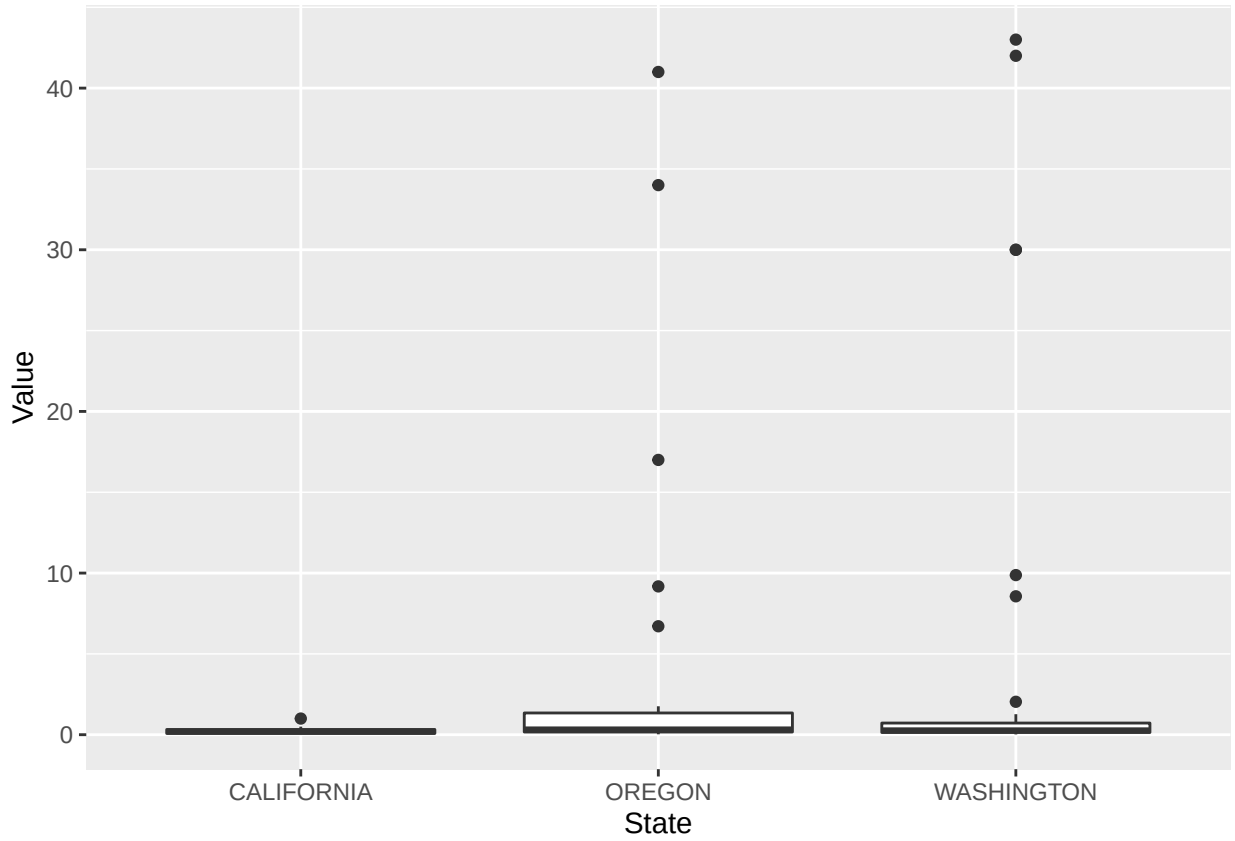
## EDA

```r
unfoodr_1$Value<-as.numeric(unfoodr$Value)
# boxplot of Chemical type
bp1 <- ggplot(unfoodr_1, aes(x = Chemical, y = Value))
bp1 <- bp1 + geom_boxplot() +
  labs(x = "Chemical type")
bp1
```

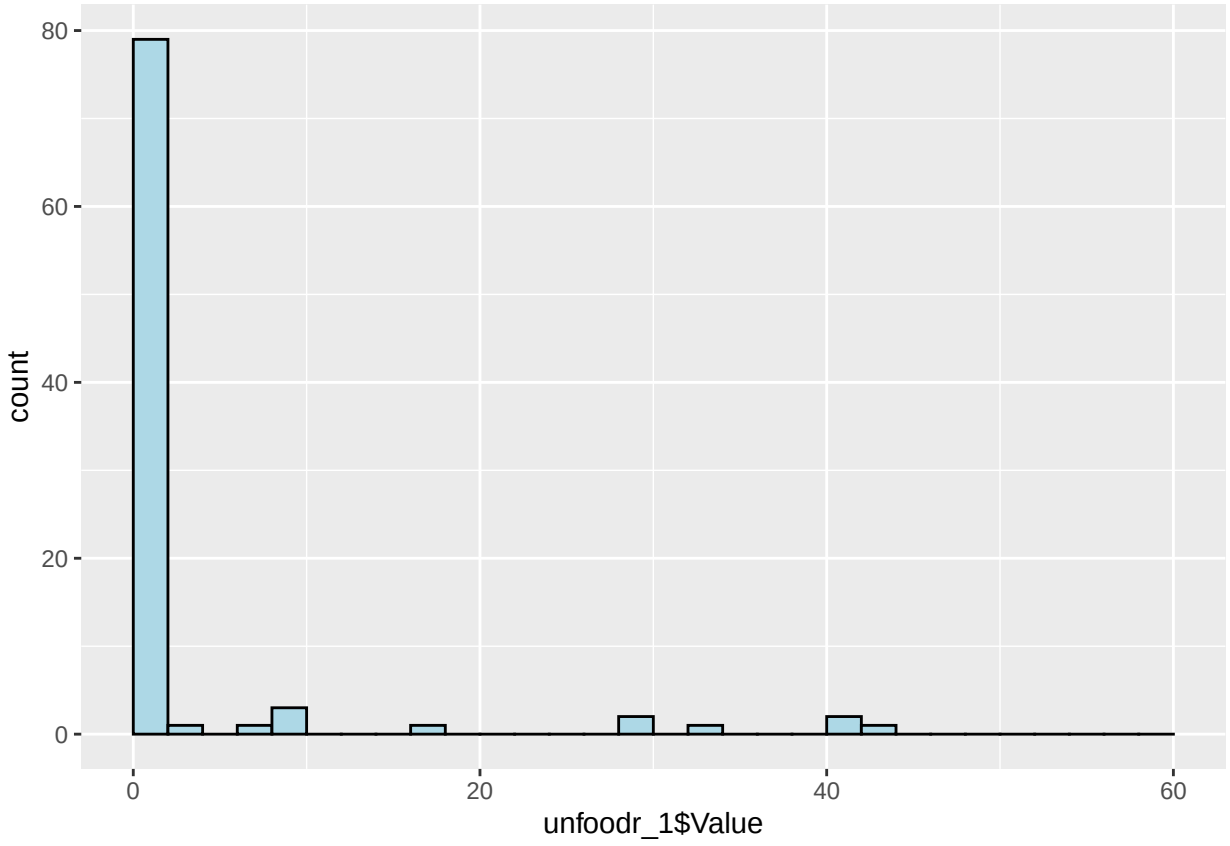



```r
# boxplot of State
bp1 <- ggplot(unfoodr_1, aes(x = State, y = Value))
```

```r
bp1 <- bp1 + geom_boxplot() +
  labs(x = "State")
bp1
```




```r
# histograms of least and most variance variables
h1 <- ggplot(unfoodr_1, aes(unfoodr_1$Value))
h1 <- h1 + geom_histogram(breaks = seq(0, 60, by = 2), col = "black", fill = "light blue")
print(h1)
```




```
summary(unfoodr_1)
```

```
##       Year         State              Chemical            Value
##  Min.   :2015   Length:91          Length:91          Min.   : 0.0170
##  1st Qu.:2015   Class :character   Class :character   1st Qu.: 0.1185
##  Median :2017   Mode  :character   Mode  :character   Median : 0.2810
##  Mean   :2017                                         Mean   : 3.3254
##  3rd Qu.:2017                                         3rd Qu.: 0.9705
##  Max.   :2019                                         Max.   :43.0000
```

## output data

```
write.csv(unfoodr_1, file = "unfoodr_1.csv", row.names = F)
```