Berries

Simu Huang

10/18/2020

Acquire the data

```
## read the data

ber <- read.csv("D:\\Documents\\R Project\\615\\Berry\\berries.csv")

## look at number of unique values in each column
a_1 <- ber %>% summarize_all(n_distinct)

## delete some column with constant values
ber %<>% select( - c(Program, Week.Ending, Geo.Level, Ag.District, Ag.District.Code, County,
 County.ANSI, Zip.Code, Region, watershed_code, Watershed, CV.... ))
```

choose the strawberry as our study object and then clean the data

```
## straberries

b_3 <- ber %>% filter(Commodity=="STRAWBERRIES")

## reorder the data by state
b_3 <- b_3[order(b_3$State.ANSI),]

## delete some values that cannot be classified
b_3 %<>% filter(Domain.Category != "NOT SPECIFIED")
b_3 %<>% filter(Period == "YEAR")


## revise some columns and remove some duplicate information which can make the data more int
uitive and make it easier for us to use the values in the following steps
b_3 %<>% separate(Data.Item, c("d1","d2"), sep="-")

unique(b_3$d1)
```

```
## [1] "STRAWBERRIES, BEARING "
```

```
unique(b_3$d2)
```

```
## [1] " APPLICATIONS, MEASURED IN LB"
## [2] " APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG"
## [3] " APPLICATIONS, MEASURED IN LB / ACRE / YEAR, AVG"
## [4] " APPLICATIONS, MEASURED IN NUMBER, AVG"
## [5] " TREATED, MEASURED IN PCT OF AREA BEARING, AVG"
```

```
b_3 %<>% select(-d1)

b_3 %<>% separate(d2, c("b1","b2"), sep=",")
unique(b_3$b1)
```

```
## [1] " APPLICATIONS" " TREATED"
```

```
unique(b_3$b2)
```

```
## [1] " MEASURED IN LB"
## [2] " MEASURED IN LB / ACRE / APPLICATION"
## [3] " MEASURED IN LB / ACRE / YEAR"
## [4] " MEASURED IN NUMBER"
## [5] " MEASURED IN PCT OF AREA BEARING"
```

```
b_3[is.na(b_3)] <- " "
for(i in 1:length(b_3$Year)){
      b_3$`Domain.Category`[i]<-str_replace(b_3$`Domain.Category`[i],"NOT SPECIFIED","NOT SPE
CIFIED, ")
}
b_3 %<>%  separate(Domain.Category, c("D1", "Domain2"), sep = ":")
b_3 %<>% select(-D1)

b_3 %<>% separate(Domain, c("Domain1_1","Domain1"), sep=",")
unique(b_3$Domain1_1)
```

```
## [1] "CHEMICAL"   "FERTILIZER"
```

```
unique(b_3$Domain1)
```

```
## [1] " FUNGICIDE"   " HERBICIDE"   " INSECTICIDE" " OTHER"       NA
```

```
b_3 %<>% select(-Domain1_1)



## remove the NA
b_3 %<>% na.omit(b_3)
b_3$Value <-  as.numeric(as.numeric(b_3$Value))

## check
summary(b_3)
```

```
##       Year         Period             State             State.ANSI
##  Min.   :2016   Length:2787        Length:2787        Min.   : 6.00
##  1st Qu.:2016   Class :character   Class :character   1st Qu.: 6.00
##  Median :2018   Mode  :character   Mode  :character   Median : 6.00
##  Mean   :2018                                         Mean   :12.16
##  3rd Qu.:2019                                         3rd Qu.:12.00
##  Max.   :2019                                         Max.   :53.00
##
##   Commodity             b1                 b2              Domain1
##  Length:2787        Length:2787        Length:2787        Length:2787
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    Domain2              Value
##  Length:2787        Min.   :  0.010
##  Class :character   1st Qu.:  0.313
##  Mode  :character   Median :  1.600
##                     Mean   : 38.161

##                     3rd Qu.: 18.000
##                     Max.   :900.000
##                     NA's   :1822
```

```
## take the values and put them in separate columns
aa <- str_extract(b_3$Domain2, "[0-9].*$")
aa <- as.numeric(str_replace_all(aa, "[[:punct:]]", " "))

bb <- str_split(b_3$Domain2, "[0-9].*$")
bb <- unlist(bb)
bb <- str_trim(bb)
bb <- bb[-which(bb=="")]
bb <-str_replace_all(bb, "[[:punct:]]", " ")

## add the new column
straw1 <- b_3 %>% mutate(Domain = bb, Domain_value = aa)
straw1 %<>% select(-Domain2)
straw2  <-  na.omit(straw1)

head(straw2)
```

	Y…	Period	State	State.ANSI	Commodity	b1	b2	Domai
	<int>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>
8	2019	YEAR	CALIFORNIA	6	STRAWBERRIES	APPLICATIONS	MEASURED IN LB	FUNGI
14	2019	YEAR	CALIFORNIA	6	STRAWBERRIES	APPLICATIONS	MEASURED IN LB	FUNGI
16	2019	YEAR	CALIFORNIA	6	STRAWBERRIES	APPLICATIONS	MEASURED IN LB	FUNGI

	Y…	Period	State	State.ANSI	Commodity	b1	b2	Domai
	<int>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>
35	2019	YEAR	CALIFORNIA	6	STRAWBERRIES	APPLICATIONS	MEASURED IN LB	FUNGI
42	2019	YEAR	CALIFORNIA	6	STRAWBERRIES	APPLICATIONS	MEASURED IN LB	HERBI
53	2019	YEAR	CALIFORNIA	6	STRAWBERRIES	APPLICATIONS	MEASURED IN LB	INSEC

6 rows | 1-10 of 12 columns

```
write.csv(straw2, "D:\\Documents\\R Project\\615\\Berry\\straw_cleaned.csv")
```

exclude the outliers

```
## summarize the value by group

s2 <- straw2 %>% group_by(State) %>% summarize(total=sum(Value))
s3 <- straw2 %>% group_by(Year, State) %>% summarize(total=sum(Value))


print(s2)
```

```
## # A tibble: 3 x 2
##   State       total
##   <chr>       <dbl>
## 1 CALIFORNIA 27410.
## 2 FLORIDA     5694.
## 3 WASHINGTON   440.
```

```
print(s3)
```

```
## # A tibble: 7 x 3
## # Groups:   Year [3]
##    Year State       total
##   <int> <chr>       <dbl>
## 1  2016 CALIFORNIA  6613.
## 2  2016 FLORIDA     2539.
## 3  2016 WASHINGTON   440.
## 4  2018 CALIFORNIA 10108.
## 5  2018 FLORIDA      234.
## 6  2019 CALIFORNIA 10689.
## 7  2019 FLORIDA     2921.
```

display the structure of s1 and check the maximum and minimum

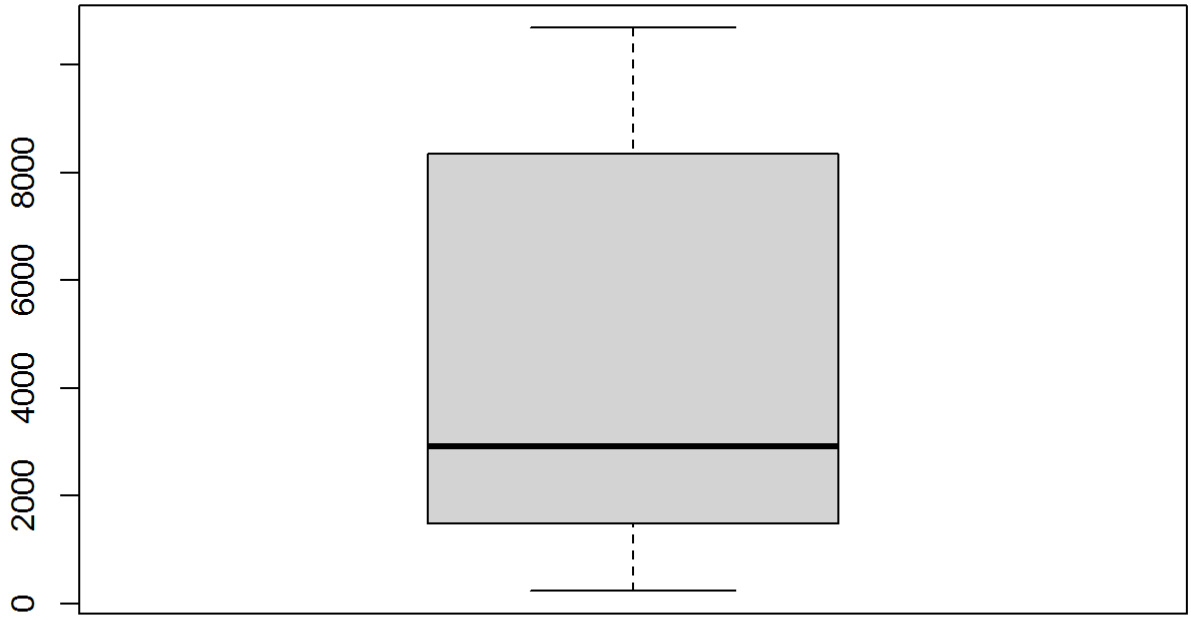
```
## boxplot
str(s3)
```

```
## tibble [7 x 3] (S3: grouped_df/tbl_df/tbl/data.frame)
##  $ Year : int [1:7] 2016 2016 2016 2018 2018 2019 2019
##  $ State: chr [1:7] "CALIFORNIA" "FLORIDA" "WASHINGTON" "CALIFORNIA" ...
##  $ total: num [1:7] 6613 2539 440 10108 234 ...
##  - attr(*, "groups")= tibble [3 x 2] (S3: tbl_df/tbl/data.frame)
##   ..$ Year : int [1:3] 2016 2018 2019
##   ..$ .rows: list<int> [1:3]
##   .. ..$ : int [1:3] 1 2 3
##   .. ..$ : int [1:2] 4 5
##   .. ..$ : int [1:2] 6 7
##   .. ..@ ptype: int(0)
##   ..- attr(*, ".drop")= logi TRUE
```

```
summary(s3)
```

```
##       Year         State               total        
##  Min.   :2016   Length:7           Min.   :  233.7  
##  1st Qu.:2016   Class :character   1st Qu.: 1489.5  
##  Median :2018   Mode  :character   Median : 2921.3  
##  Mean   :2017                      Mean   : 4792.0  
##  3rd Qu.:2018                      3rd Qu.: 8360.4  
##  Max.   :2019                      Max.   :10689.2
```

```
boxplot(s3$total)
```



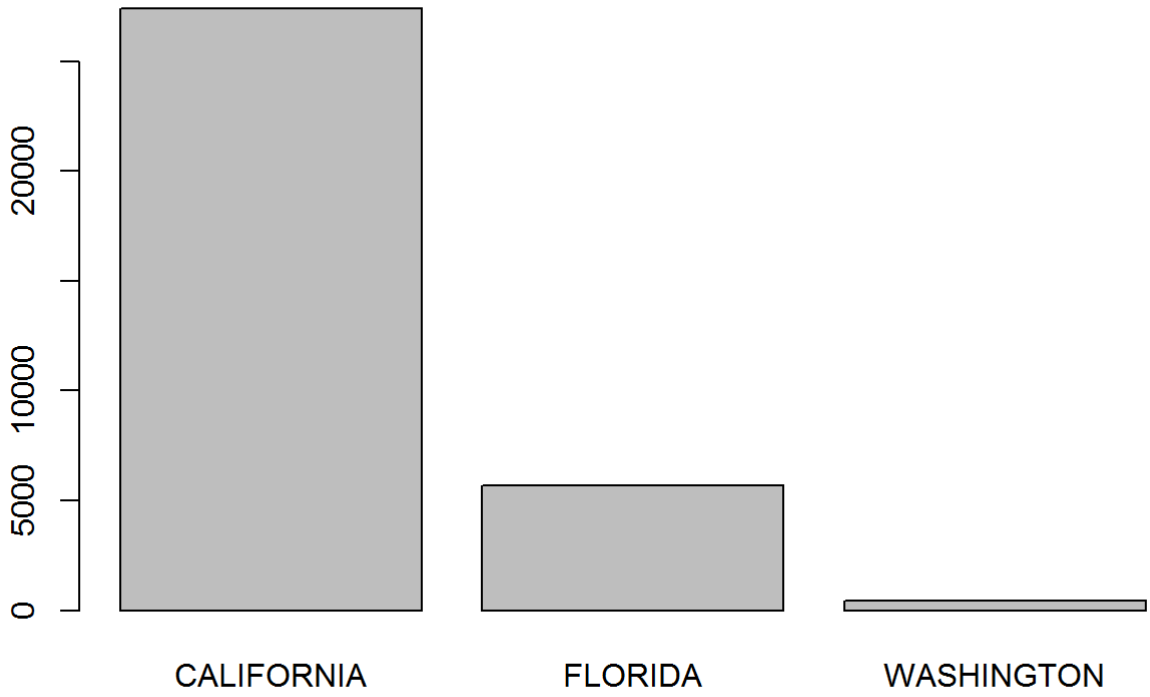

```
## check for outliers
boxplot.stats(s3$total)$out
```

```
## numeric(0)
```

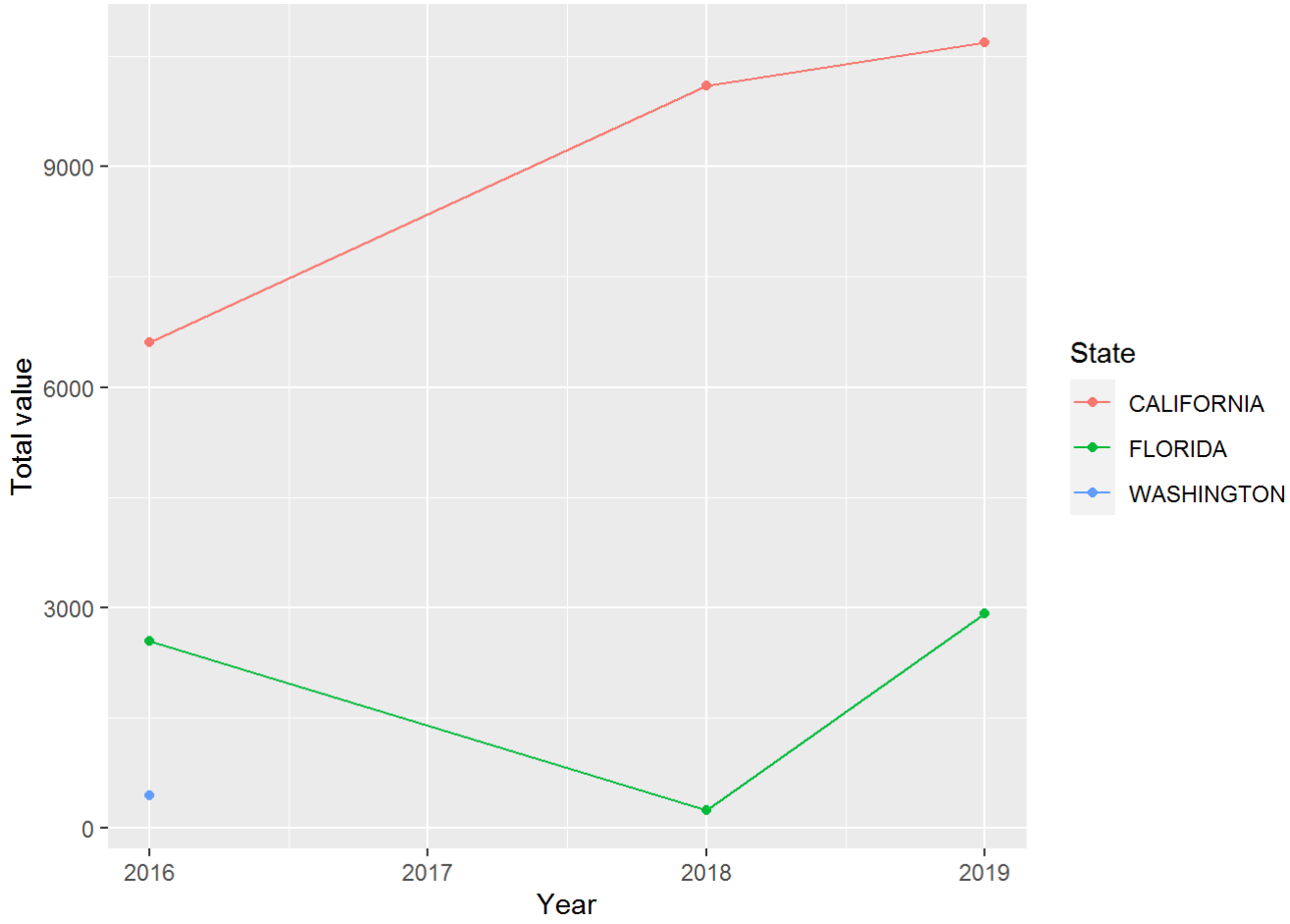
```
## find the location of the outlier
x_out <- which(s3$total %in% boxplot.stats(s3$total)$out)
print(x_out)
```

```
## integer(0)
```

```
## barplot of the total of each state
barplot(s2$total, names.arg = s2$State)
```




```
## line plot of total amount per year in each country
ggplot(data= s3,aes(x=Year,y=total,group=State,color=State))+
  geom_line()+
  geom_point()+
  labs(y="Total value")
```



summary data according to different categories

```
# unique(straw2$State)
# cal <- straw2 %>% filter(State == "CALIFORNIA")
# flo <- straw2 %>% filter(State == "FLORIDA")
# was <- straw2 %>% filter(State == "WASHINGTON")

## value for each country for each year
p_1 <- straw2 %>% group_by(Year,State) %>%
  summarise(count =n(),                          #number
                  max_mon = max(Value),          #maximum
                  min_mon = min(Value),          #minimum
                  avg_sales = mean(Value))       #average
p_1
```

Year <int>	State <chr>	count <int>	max_mon <dbl>	min_mon <dbl>	avg_sales <dbl>
2016	CALIFORNIA	231	700	0.019	28.62609
2016	FLORIDA	46	800	0.010	55.20126
2016	WASHINGTON	14	200	0.155	31.40364
2018	CALIFORNIA	239	900	0.019	42.29393
2018	FLORIDA	16	96	0.221	14.60506
2019	CALIFORNIA	292	900	0.019	36.60694

Year <int>	State <chr>	count <int>	max_mon <dbl>	min_mon <dbl>	avg_sales <dbl>
2019	FLORIDA	63	700	0.045	46.36937

7 rows

```
## every country over the years
p_2 <- straw2 %>% group_by(State) %>%
  summarise(count =n(),                          #number
                  max_mon = max(Value),          #maximum
                  min_mon = min(Value),          #minimum
                  avg_sales = mean(Value),       #average
                  sum_sales = sum(Value))        #sum
p_2
```

State <chr>	count <int>	max_mon <dbl>	min_mon <dbl>	avg_sales <dbl>	sum_sales <dbl>
CALIFORNIA	762	900	0.019	35.97126	27410.103
FLORIDA	125	800	0.010	45.55367	5694.209
WASHINGTON	14	200	0.155	31.40364	439.651

3 rows

```
## the overall situation for each year
p_3 <- straw2 %>% group_by(Year) %>%
  summarise(       count =n(),                   #number
                  max_mon = max(Value),          #maximum
                  min_mon = min(Value),          #minimum
                  avg_sales = mean(Value),       #average
                  sum_sales = sum(Value))        #sum
p_3
```

Year <int>	count <int>	max_mon <dbl>	min_mon <dbl>	avg_sales <dbl>	sum_sales <dbl>
2016	291	800	0.010	32.96060	9591.535
2018	255	900	0.019	40.55659	10341.931
2019	355	900	0.019	38.33943	13610.497

3 rows

```
p_4 <- straw2 %>% group_by(Domain) %>% summarise(count =n())
p_4
```

Domain <chr>	count <int>
ABAMECTIN =	20
ACEQUINOCYL =	12

Domain <chr>	count <int>
ACETAMIPRID =	17
AZADIRACHTIN =	14
AZOXYSTROBIN =	20
BACILLUS PUMILUS =	2
BACILLUS SUBTILIS =	6
BIFENAZATE =	21
BIFENTHRIN =	22
BLAD =	8

1-10 of 74 rows Previous **1** 2 3 4 5 6 … 8 Next

```
## the variance
stra_var <- straw2 %>% group_by(State) %>% summarise(var = var(Value))
print(stra_var)
```

```
## # A tibble: 3 x 2
##   State         var
##   <chr>       <dbl>
## 1 CALIFORNIA 15183.
## 2 FLORIDA    18005.
## 3 WASHINGTON  3361.
```

```
## correlation
straw3 <- data.frame(straw2$Value, straw2$Domain_value)
cor<- cor(straw3)
print(cor)
```

```
##                     straw2.Value straw2.Domain_value
## straw2.Value         1.000000000         0.005045674
## straw2.Domain_value  0.005045674         1.000000000
```

```
cor.test(straw2$Domain_value, straw2$Value, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  straw2$Domain_value and straw2$Value
## t = 0.15129, df = 899, p-value = 0.8798
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.06028593  0.07033424
## sample estimates:
##         cor
## 0.005045674
```

From the result, we can find that people prefer using ABAMECTIN, ACETAMIPRID, and AZOXYSTROBIN. And less people choose to use ACEQUINOCYL. In recent years the value of strawberries in California has been far higher than anywhere else, especially in 2018 and 2019. The value of strawberries fluctuated more in California, and was more stable in Florida.
To sum up, we can conclude from the above results that the different additives used in each place had no significant effect on the value of the strawberry.

Citation

[1] EDA.rmd, Haviland Wright

[2] ag_data.Rmd, Haviland Wright

[3] NASS