Berries_EDA

Haoyu Li

10/18/2020

```
library(knitr)
library(tidyverse)
library(magrittr)
library(stats)
```

#Data recap

The data used in this analysis is the cleaned "berries", to be more specific, that is the strawberries I extract out of "berries". More information on this data set can be found on this website : http://quickstats.nass.usda.gov

#Part 1: exploratory data analysis

The exploratory data analysis is to do some basic summaries of the dataset that we are going to look deep into. We will see if there are outliers, large variances, and anything special tha we should pay attention to.

##read cleaned data

```
sberry<-read.csv("strawberry.csv")
sberry%<>%select(-X)
```

#check how many types of chemicals

```
df_1<-sberry%>%select(Year,State,production,Chemical,Value)

df_1%>%group_by(Chemical)%>%summarise(n=n())
```

```
## # A tibble: 9 x 2
##   Chemical          n
##   <fct>         <int>
## 1 ""              358
## 2 "(NITROGEN)"     20
## 3 "(PHOSPHATE)"    20
## 4 "(POTASH)"       20
## 5 "(SULFUR)"       15
## 6 "FUNGICIDE"    1056
## 7 "HERBICIDE"     276
## 8 "INSECTICIDE"  1056
## 9 "OTHER"         399
```

#filter out NA and D value

```
df_1%<>%filter(Chemical !="")
df_1%<>% filter(Value != "(D)")

df_1%<>% filter(Value !=  "(NA)")

df_1%<>%filter(Value != "(Z)")
```

```r
df_1$Value<-as.numeric(gsub(",", "", df_1$Value))
```

#Basic data summary

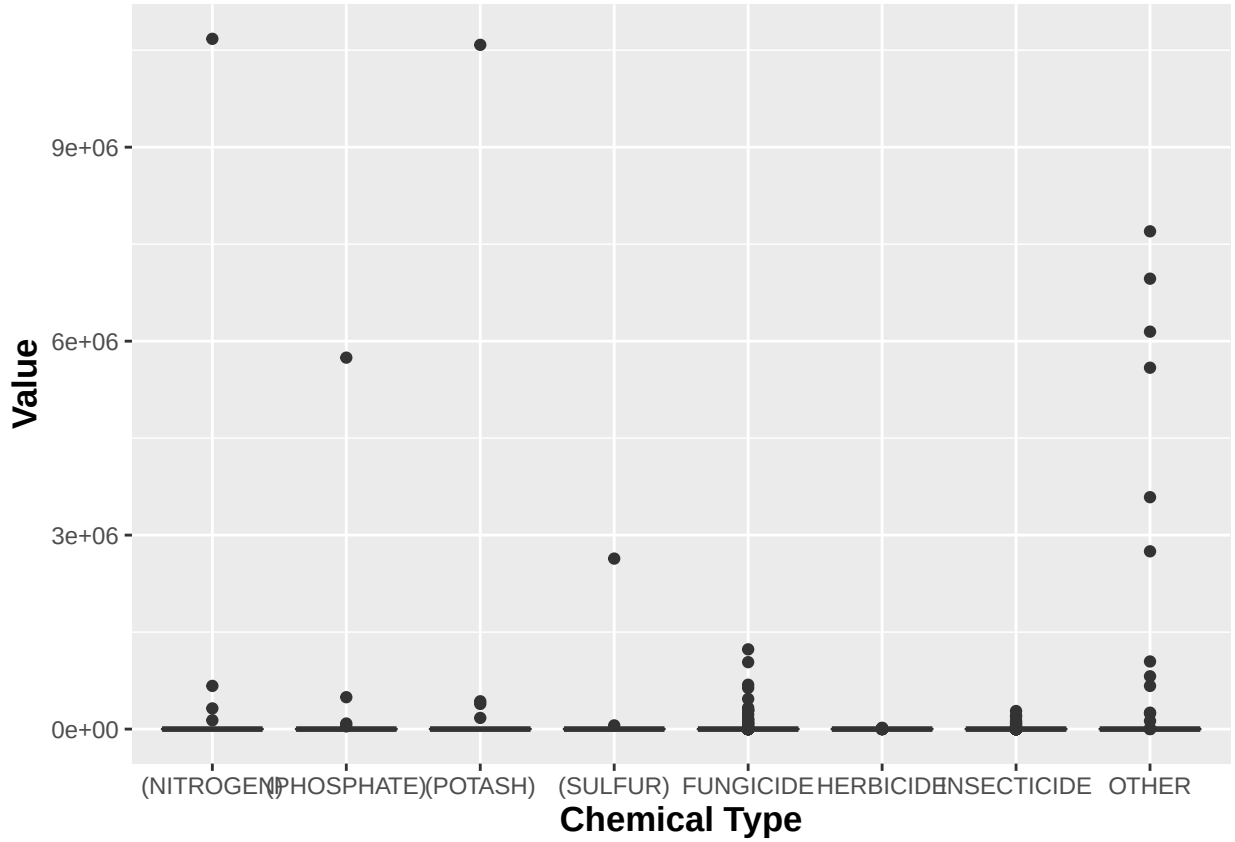
-We do boxplot on the chemical variable to see the data distribution

-No obvious outliers by just looking at the boxplot

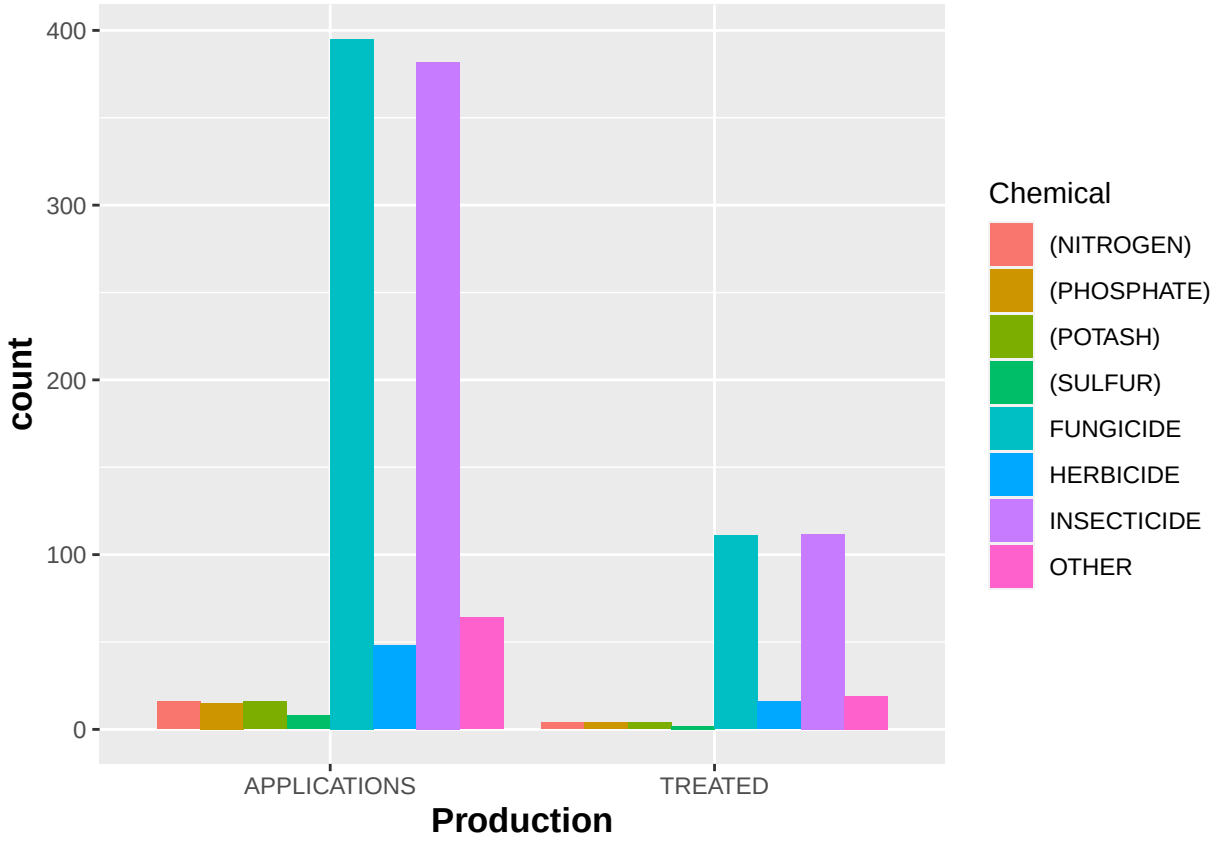
-We also do relatively simple bar chart on the categories of production to get a sense of what kinds of chemicals are used mostly

```r
#summarize the chemical usage
s1<-df_1%>%group_by(State,Chemical)%>%summarise(total = sum(Value))
view(s1)

#boxplot of Chemical
ggplot(data = df_1) +
  geom_boxplot(mapping = aes(x = Chemical, y = Value)) +
  theme(axis.title = element_text(size = 13, face = "bold")) +
  labs(x = "Chemical Type")
```




```r
#count each production by the chemical categories
ggplot(data=df_1)+
  geom_bar(mapping=aes(x=production,fill=Chemical),position = "dodge")+
  theme(axis.title = element_text(size = 13, face = "bold")) +
  labs(x = "Production")
```



#Part 2: PCA

Without going into too much detail, Principal Component Analysis (PCA) can thus be used to reduce the dimensions of the data into fewer components that would retain as much of the variability expressed by the original data as possible. Here, we are interested in the relationship between different chemicals

-prepare data

```
df_2<-df_1%>%select(State,Chemical,Value)
df_2[is.na(df_2)]<-0
df_3<-df_2%>%group_by(State,Chemical)%>%summarise(total = sum(Value))
df_4 <- df_3%>%pivot_wider(names_from = Chemical,values_from = total)
df_4[is.na(df_4)] <-0
df_5<-df_4[,-1]
head(df_5)
```

```
## # A tibble: 4 x 8
##   `(NITROGEN)` `(PHOSPHATE)` `(POTASH)` `(SULFUR)` FUNGICIDE HERBICIDE
##          <dbl>         <dbl>      <dbl>      <dbl>     <dbl>     <dbl>
## 1    11344611.      6238446.  11013590.   2694247.  5446312.    67974.
## 2      456285.       128207.    562297.          0   1285019.      315
## 3           0             0          0          0      2786         0
## 4           0             0          0          0      1503.     1176.
## # ... with 2 more variables: INSECTICIDE <dbl>, OTHER <dbl>
```

#find variable describing the most variance

```
df_5_var <- apply(df_5, 2, var)
print(df_5_var)
```

```
##   (NITROGEN)  (PHOSPHATE)     (POTASH)     (SULFUR)    FUNGICIDE    HERBICIDE
## 3.136437e+13 9.600359e+12 2.937168e+13 1.814742e+12 6.657151e+12 1.138537e+09
##  INSECTICIDE        OTHER
## 3.583071e+11 3.192839e+14
```

```
max(df_5_var)
```

```
## [1] 3.192839e+14
```

```
which(df_5_var == max(df_5_var))
```

```
## OTHER
##     8
```

```
min(df_5_var)
```

```
## [1] 1138537368
```

```
which(df_5_var == min(df_5_var))
```

```
## HERBICIDE
##         6
```

#Correlation matrix

```
cor_m <- cor(df_5)
print(cor_m)
```

```
##             (NITROGEN) (PHOSPHATE)  (POTASH)  (SULFUR) FUNGICIDE HERBICIDE
## (NITROGEN)   1.0000000   0.9998212 0.9999447 0.9992622 0.9804282 0.9990073
## (PHOSPHATE)  0.9998212   1.0000000 0.9995672 0.9998097 0.9765303 0.9996269
## (POTASH)     0.9999447   0.9995672 1.0000000 0.9988032 0.9824437 0.9985084
## (SULFUR)     0.9992622   0.9998097 0.9988032 1.0000000 0.9721433 0.9998915
## FUNGICIDE    0.9804282   0.9765303 0.9824437 0.9721433 1.0000000 0.9711413
## HERBICIDE    0.9990073   0.9996269 0.9985084 0.9998915 0.9711413 1.0000000
## INSECTICIDE  0.9972280   0.9956428 0.9979550 0.9936345 0.9923593 0.9930917
## OTHER        0.9993842   0.9998690 0.9989602 0.9999945 0.9729166 0.9998733
##             INSECTICIDE     OTHER
## (NITROGEN)    0.9972280 0.9993842
## (PHOSPHATE)   0.9956428 0.9998690
## (POTASH)      0.9979550 0.9989602
## (SULFUR)      0.9936345 0.9999945
## FUNGICIDE     0.9923593 0.9729166
## HERBICIDE     0.9930917 0.9998733
## INSECTICIDE   1.0000000 0.9940032
## OTHER         0.9940032 1.0000000
```

#PCA

-By plot the cluster, we can see some of chemical variable clustering together so we can infer these variables may have interactions and relationship that might have effect on the modeling that research team are going to do.

-The purpose the PCA is to help research team get a sense of what variables they should pay attention. That is why the purpose of EDA is to give advice and help team explore the data.

```
pca <- prcomp(df_5, center = TRUE, scale. = TRUE)
print(pca)
```

```
## Standard deviations (1, .., p=4):
```

```
## [1] 2.820135e+00 2.160464e-01 1.269768e-02 1.863993e-16
##
## Rotation (n x k) = (8 x 4):
##                     PC1         PC2         PC3          PC4
## (NITROGEN)  -0.3545342  0.08368169 -0.16286677 -0.793614462
## (PHOSPHATE) -0.3543494  0.17115570 -0.19211789 -0.120526378
## (POTASH)    -0.3545822  0.03503583 -0.14657884  0.549481697
## (SULFUR)    -0.3540259  0.26132883 -0.22222105  0.188568048
## FUNGICIDE   -0.3488599 -0.82892038  0.12920350 -0.001501980
## HERBICIDE   -0.3539197  0.28021423  0.89210491  0.001400009
## INSECTICIDE -0.3540292 -0.26086049 -0.07759069  0.052385761
## OTHER       -0.3540905  0.24597892 -0.21710031  0.124148904
```

```
summary(pca)
```

```
## Importance of components:
##                           PC1     PC2     PC3       PC4
## Standard deviation     2.8201 0.21605 0.01270 1.864e-16
## Proportion of Variance 0.9941 0.00583 0.00002 0.000e+00
## Cumulative Proportion  0.9941 0.99998 1.00000 1.000e+00
```

```
# flip values to positives
pca_1 <- pca
pca_1$rotation <- -pca_1$rotation
pca_1$x <- -pca_1$x
print(pca_1)
```

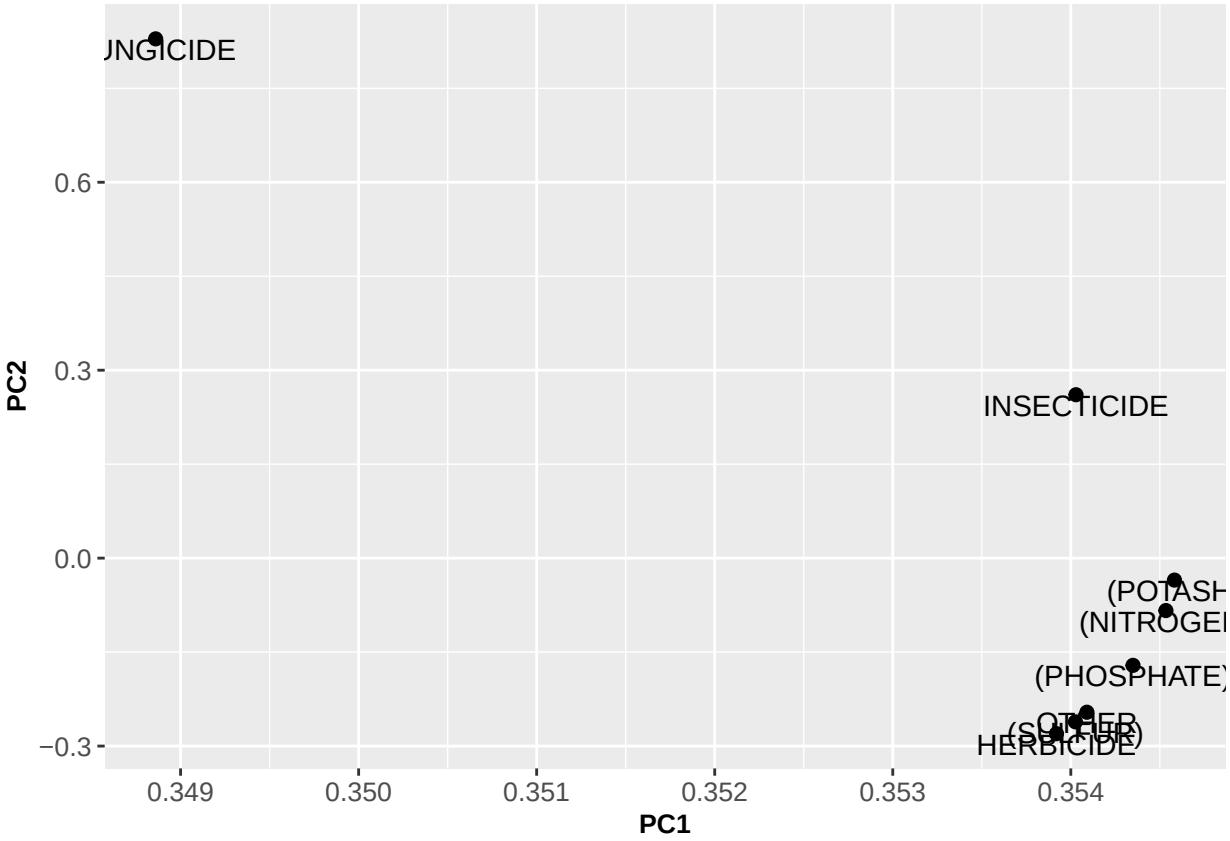
```
## Standard deviations (1, .., p=4):
## [1] 2.820135e+00 2.160464e-01 1.269768e-02 1.863993e-16
##
## Rotation (n x k) = (8 x 4):
##                    PC1         PC2         PC3          PC4
## (NITROGEN)  0.3545342 -0.08368169  0.16286677  0.793614462
## (PHOSPHATE) 0.3543494 -0.17115570  0.19211789  0.120526378
## (POTASH)    0.3545822 -0.03503583  0.14657884 -0.549481697
## (SULFUR)    0.3540259 -0.26132883  0.22222105 -0.188568048
## FUNGICIDE   0.3488599  0.82892038 -0.12920350  0.001501980
## HERBICIDE   0.3539197 -0.28021423 -0.89210491 -0.001400009
## INSECTICIDE 0.3540292  0.26086049  0.07759069 -0.052385761
## OTHER       0.3540905 -0.24597892  0.21710031 -0.124148904
```

```
summary(pca_1)
```

```
## Importance of components:
##                           PC1     PC2     PC3       PC4
## Standard deviation     2.8201 0.21605 0.01270 1.864e-16
## Proportion of Variance 0.9941 0.00583 0.00002 0.000e+00
## Cumulative Proportion  0.9941 0.99998 1.00000 1.000e+00
```

```
# PCA scatter plot
pc_2 <- data.frame(pca_1$rotation[, 1:2])
pc_2$Chemical <- rownames(pc_2)
plot1 <- ggplot(pc_2, aes(x = PC1, y = PC2))
plot1 <- plot1 + geom_point(size = 2) +
  geom_text(aes(label = Chemical), vjust = 1) +
  theme(axis.text = element_text(size = 10),
        axis.title = element_text(size = 10, face = "bold"))
```

```
print(plot1)
```


UNGICIDE
0.6
0.3
0.0
−0.3
PC2
INSECTICIDE
(POTASH
(NITROGE
(PHOSPHATE)
OTHER
(SULFUR)
HERBICIDE
0.349
0.350
0.351
0.352
0.353
0.354
PC1