

Report Of Berry Assignments

Zhe Yu

2020/10/18

Introduction

In this berry assignment, I chose the data of raspberry to do cleaning, organizing and some EDA. I tried to find how years, states and chemicals used effected the value of raspberry. Also I find the correlation between different type of chemicals.

Clean data

In this procedure, I selected all the data of raspberry from the raw data. And due to the messy information in column 'Data Item', 'Domain' and 'Domain Category', I tried to separate them into several independent columns.

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

## use '#' to separate 'Data Item' into three parts
rberry$`Data Item` <- str_replace(rberry$`Data Item`, "^RASPBERRIES", "RASPBERRIES#")
rberry$`Data Item` <- str_replace(rberry$`Data Item`, "MEASURED IN", "#MEASURED IN")
rberry %<>% separate(`Data Item`, c("B", "type", "meas"), sep = "#")

## tidy up each new column and continue to separate them
rberry %<>% separate(type, c("Type", "Production"), sep = " - ")
rberry %<>% separate(Type, c("d1", "Type"), sep = ",")
rberry %<>% separate(Production, c("Production", "d2"), sep = ",")
rberry %<>% separate(meas, c("Measures", "Avg"), sep = ", ")

## delete those columns with useless information
rberry %<>% select(-c(B, d1, d2))
rberry[is.na(rberry)] <- " " ## OK now Data Item has been split into parts

#kable(head(rberry, n=10)) %>% kable_styling(fixed_thead = T, font_size = 10)
head(rberry, n=10)
```

```
## # A tibble: 10 x 9
##   Year State Type Production Measures Avg Domain `Domain Category` Value
##   <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 2019 CALIF~ " " ACRES HAR~ " " " " TOTAL NOT SPECIFIED 7,500
## 2 2019 CALIF~ " " PRODUCTION "MEASUR~ " " TOTAL NOT SPECIFIED 143,~
## 3 2019 CALIF~ " " YIELD "MEASUR~ " " TOTAL NOT SPECIFIED 19,1~
## 4 2019 CALIF~ " BEA~ APPLICATI~ "MEASUR~ " " CHEMI~ CHEMICAL, FUNGICI~ (D)
## 5 2019 CALIF~ " BEA~ APPLICATI~ "MEASUR~ " " CHEMI~ CHEMICAL, FUNGICI~ (NA)
## 6 2019 CALIF~ " BEA~ APPLICATI~ "MEASUR~ " " CHEMI~ CHEMICAL, FUNGICI~ (NA)
```

```
## 7 2019 CALIF~ " BEA~ APPLICATI~ "MEASUR~ " " CHEMI~ CHEMICAL, FUNGICI~ (D)
## 8 2019 CALIF~ " BEA~ APPLICATI~ "MEASUR~ " " CHEMI~ CHEMICAL, FUNGICI~ 300
## 9 2019 CALIF~ " BEA~ APPLICATI~ "MEASUR~ " " CHEMI~ CHEMICAL, FUNGICI~ (D)
## 10 2019 CALIF~ " BEA~ APPLICATI~ "MEASUR~ " " CHEMI~ CHEMICAL, FUNGICI~ (D)
```

```
## onto Domain
```

```
##separate 'Domain' and 'Domain category'
```

```
rberry %<>% separate(Domain, c("D_left", "D_right"), sep = ", ")
rberry[is.na(rberry)] <- " "
rberry %<>% separate(`Domain Category`, c("DC_left", "DC_right"), sep = ", ")
```

```
## work on DC_left first
```

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

```
## now work on DC_right
```

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

```
#kable(head(rberry,n=10)) %>% kable_styling(fixed_thead = T, font_size = 10)
head(rberry,n=10)
```

```
## # A tibble: 10 x 13
##   Year State Type Production Measures Avg D_left D_right DC_left_l
##   <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 2019 CALI~ " " ACRES HAR~ " " " " TOTAL " " NOT SPEC~
## 2 2019 CALI~ " " PRODUCTION "MEASUR~ " " TOTAL " " NOT SPEC~
## 3 2019 CALI~ " " YIELD "MEASUR~ " " TOTAL " " NOT SPEC~
## 4 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " CHEMI~ "FUNGI~ CHEMICAL
## 5 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " CHEMI~ "FUNGI~ CHEMICAL
## 6 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " CHEMI~ "FUNGI~ CHEMICAL
## 7 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " CHEMI~ "FUNGI~ CHEMICAL
## 8 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " CHEMI~ "FUNGI~ CHEMICAL
## 9 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " CHEMI~ "FUNGI~ CHEMICAL
## 10 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " CHEMI~ "FUNGI~ CHEMICAL
## # ... with 4 more variables: DC_left_r <chr>, DC_right_l <chr>,
## # DC_right_r <chr>, Value <chr>
```

```
## OK now we need to eliminate the redundancy
```

```
## fine and remove redundant columns
```

```
## remove column rberry$DC_left_l and DC_right_l
```

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

```
## remove "Chemical" and joint the columns
```

```
rberry %<>% mutate(D_left = "CHEMICAL", D_left = "")
rberry %<>% mutate(Chemical=paste(D_left, D_right))
rberry %<>% select(-c(D_left, D_right))
```

```
## select columns that want to reserve
```

```
rberry %<>% select(Year, State, Type, Production, Measures,Avg, DC_left_r, DC_right_r, Chemical, Value )
```

```
#kable(head(rberry,n=10)) %>% kable_styling(fixed_thead = T, font_size = 10)
head(rberry,n=10)
```

```
## # A tibble: 10 x 10
##   Year State Type Production Measures Avg DC_left_r DC_right_r Chemical
##   <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 2019 CALI~ " " ACRES HAR~ " " " " " " " "
## 2 2019 CALI~ " " PRODUCTION "MEASUR~ " " " " " "
## 3 2019 CALI~ " " YIELD "MEASUR~ " " " " " "
## 4 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " " " "(AZOXYST~ " FUNGI~
## 5 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " " " "(BACILLU~ " FUNGI~
## 6 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " " " "(BACILLU~ " FUNGI~
## 7 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " " " "(BORAX D~ " FUNGI~
## 8 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " " " "(BOSCALI~ " FUNGI~
## 9 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " " " "(CALCIUM~ " FUNGI~
## 10 2019 CALI~ " BE~ APPLICATI~ "MEASUR~ " " " " "(CAPTAN ~ " FUNGI~
## # ... with 1 more variable: Value <chr>
```

```
## now let's clean it up
rberry %<>% rename(Chem_family = DC_left_r, Materials = DC_right_r)
rberry %<>% mutate(Chemical = str_trim(paste(Chem_family, Chemical)))
rberry %<>% select(Year, State, Type, Production, Avg, Measures, Materials, Chemical, Value)
#kable(head(rberry,n=10)) %>% kable_styling(fixed_thead = T, font_size = 10)
head(rberry,n=10)
```

```
## # A tibble: 10 x 9
##   Year State Type Production Avg Measures Materials Chemical Value
##   <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 2019 CALIF~ " " ACRES HARV~ " " " " " " 7,500
## 2 2019 CALIF~ " " PRODUCTION " " "MEASURED~ " " "" 143,~
## 3 2019 CALIF~ " " YIELD " " "MEASURED~ " " "" 19,1~
## 4 2019 CALIF~ " BEA~ APPLICATIO~ " " "MEASURED~ "(AZOXYSTROB~ "FUNGIC~ (D)
## 5 2019 CALIF~ " BEA~ APPLICATIO~ " " "MEASURED~ "(BACILLUS A~ "FUNGIC~ (NA)
## 6 2019 CALIF~ " BEA~ APPLICATIO~ " " "MEASURED~ "(BACILLUS S~ "FUNGIC~ (NA)
## 7 2019 CALIF~ " BEA~ APPLICATIO~ " " "MEASURED~ "(BORAX DECA~ "FUNGIC~ (D)
## 8 2019 CALIF~ " BEA~ APPLICATIO~ " " "MEASURED~ "(BOSCALID =~ "FUNGIC~ 300
## 9 2019 CALIF~ " BEA~ APPLICATIO~ " " "MEASURED~ "(CALCIUM PO~ "FUNGIC~ (D)
## 10 2019 CALIF~ " BEA~ APPLICATIO~ " " "MEASURED~ "(CAPTAN = 8~ "FUNGIC~ (D)
```

```
##write.csv(rberry,"rberry.csv",row.names = F)
```

Organize data

The data was cleaned up, because the majority of raspberry is for application, so I filter rows of data in the same situation to be prepared for the EDA part.

```
## look at chemicals being applied to food, and drop "(D)" and "(NA)"
unfood <- rberry %<>% filter(Production=="APPLICATIONS")
unfood %<>% filter(Value != "(D)")
unfood %<>% filter(Value != "(NA)")
unfood %<>% filter(Measures == "MEASURED IN LB / ACRE / APPLICATION")
unfood$Value <- as.numeric(unfood$Value)
```

```
## just I mentioned before I select year, state, chemical and value,
## and make the dataframe more wider
```

```

unfood_1 <- unfood %>% select(Year, State, Chemical, Value)
unfood_1$Value <- as.numeric(unfood_1$Value)
unfood_1 %<>% pivot_wider(names_from = Chemical, values_from = Value)

## because the using the pivot_wider so some data was a list,
## I change them into sum
for (i in 1:6) {
  f <- as.numeric(unlist(unfood_1$INSECTICIDE[i]))
  unfood_1$INSECTICIDE[i] <- sum(f)
}
for (i in 1:6) {
  f <- as.numeric(unlist(unfood_1$FUNGICIDE[i]))
  unfood_1$FUNGICIDE[i] <- sum(f)
}
for (i in 1:6) {
  f <- as.numeric(unlist(unfood_1$HERBICIDE[i]))
  unfood_1$HERBICIDE[i] <- sum(f)
}
for (i in 1:6) {
  f <- as.numeric(unlist(unfood_1$OTHER[i]))
  unfood_1$OTHER[i] <- sum(f)
}
for (i in 1:6) {
  f <- as.numeric(unlist(unfood_1$(NITROGEN)[i]))
  unfood_1$(NITROGEN)[i] <- sum(f)
}
for (i in 1:6) {
  f <- as.numeric(unlist(unfood_1$(PHOSPHATE)[i]))
  unfood_1$(PHOSPHATE)[i] <- sum(f)
}
for (i in 1:6) {
  f <- as.numeric(unlist(unfood_1$(POTASH)[i]))
  unfood_1$(POTASH)[i] <- sum(f)
}

unfood_1$FUNGICIDE <- as.numeric(unfood_1$FUNGICIDE)
unfood_1$INSECTICIDE <- as.numeric(unfood_1$INSECTICIDE)
unfood_1$HERBICIDE <- as.numeric(unfood_1$HERBICIDE)
unfood_1$OTHER <- as.numeric(unfood_1$OTHER)
unfood_1$(NITROGEN) <- as.numeric(unfood_1$(NITROGEN))
unfood_1$(PHOSPHATE) <- as.numeric(unfood_1$(PHOSPHATE))
unfood_1$(POTASH) <- as.numeric(unfood_1$(POTASH))

#kable(head(rberry,n=10)) %>% kable_styling(fixed_thead = T, font_size = 10)
head(rberry,n=10)

## # A tibble: 10 x 9
##   Year State Type Production Avg Measures Materials Chemical Value
##   <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
## 1 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (AZOXYSTROBIN ~ FUNGICI~ (D)
## 2 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (BACILLUS AMYL~ FUNGICI~ (NA)
## 3 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (BACILLUS SUBT~ FUNGICI~ (NA)
## 4 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (BORAX DECAHYD~ FUNGICI~ (D)

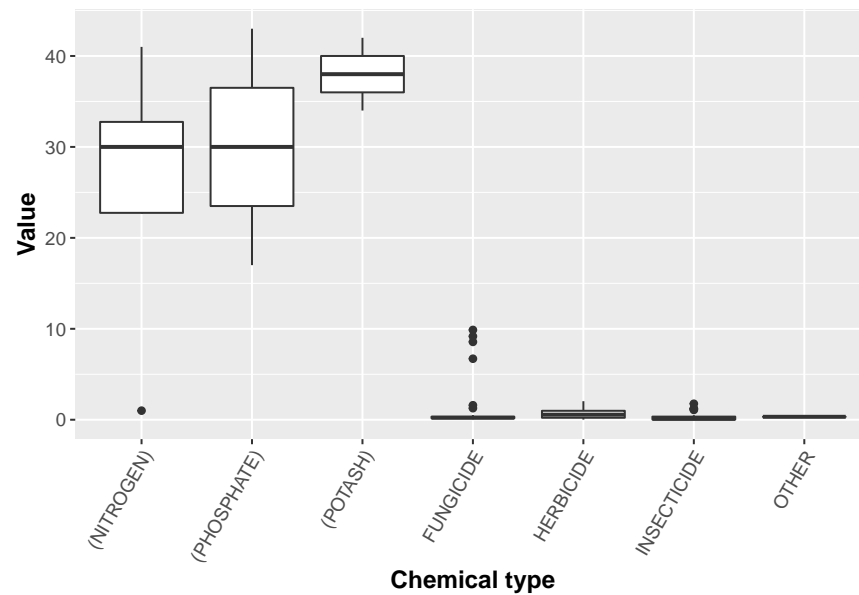
```

```
## 5 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (BOSCALID = 12~ FUNGICI~ 300
## 6 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (CALCIUM POLYS~ FUNGICI~ (D)
## 7 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (CAPTAN = 8130~ FUNGICI~ (D)
## 8 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (COPPER OCTANO~ FUNGICI~ (D)
## 9 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (COPPER OXIDE ~ FUNGICI~ (D)
## 10 2019 CALIF~ " BEA~ APPLICATIO~ " " MEASURE~ (CYPRODINIL = ~ FUNGICI~ 900
```

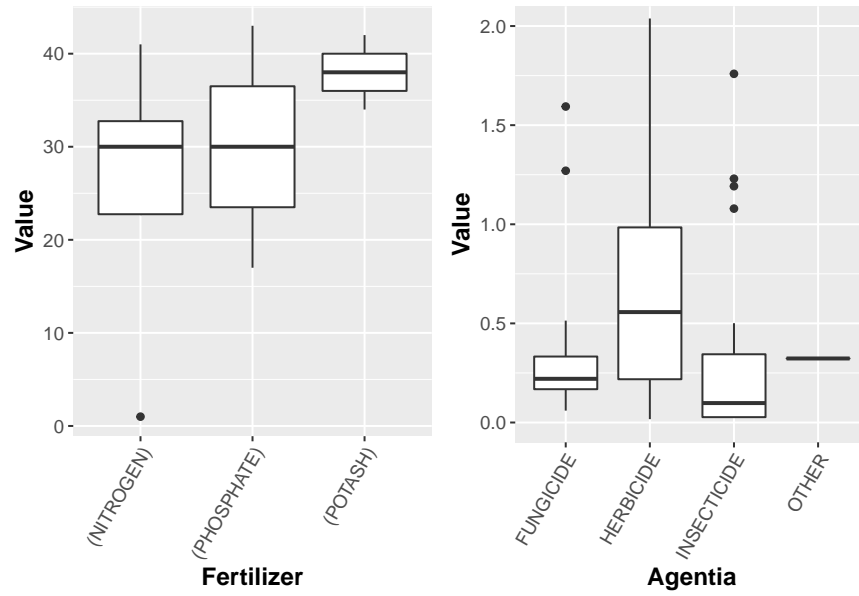
EDA

Chemical type and Value

At first I want to draw the boxplot to see the distribution of value by using different type of chemical, but the range of different chemical is not the same. And also there are outliers that make the plot not useful, so I make some adjustment and create another plot.



so I separate the chemical into two parts: agentia and fertilizer. I create an indicator to distinguish them. Two parts are drawn separately, but I use the `grid.arrange` function to combine them into the same plot. It is clear that the value of raspberry using fertilizer ('nitrogen', 'phosphate' and 'potash') is much higher than those using agentia ('fungicide', 'herbicide', 'insecticide' and 'other') (because 'other' also has a small range so I put it in the agentia).

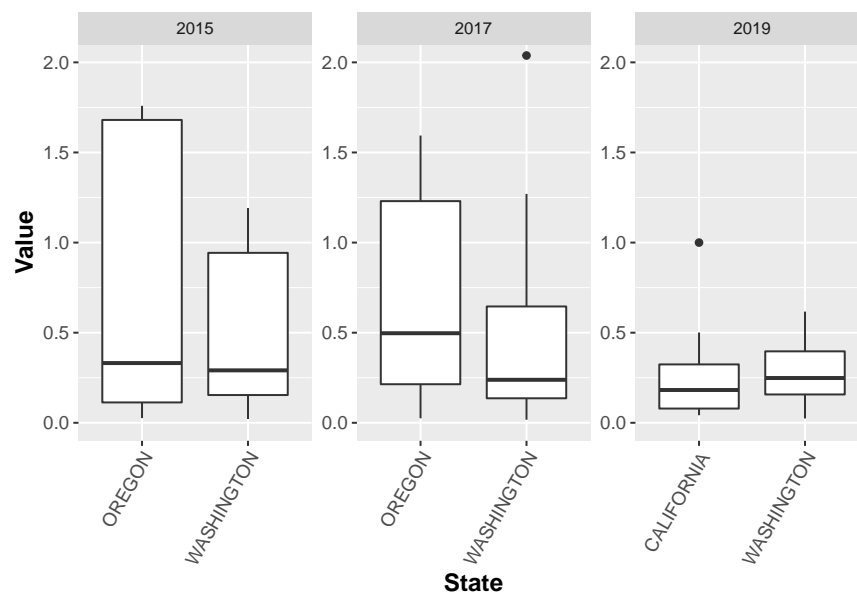


And also can see those outliers here.

```
## # A tibble: 4 x 11
##   Year State Type Production Avg Measures Materials Chemical Value
##   <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <dbl>
## 1  2017 OREG~ " BE~ APPLICATI~ AVG MEASURE~ (CALCIUM~ FUNGICI~ 6.71
## 2  2017 WASH~ " BE~ APPLICATI~ AVG MEASURE~ (CALCIUM~ FUNGICI~ 8.56
## 3  2015 OREG~ " BE~ APPLICATI~ AVG MEASURE~ (CALCIUM~ FUNGICI~ 9.18
## 4  2015 WASH~ " BE~ APPLICATI~ AVG MEASURE~ (CALCIUM~ FUNGICI~ 9.88
## # ... with 2 more variables: Chemicaltype <chr>, Variables <chr>
```

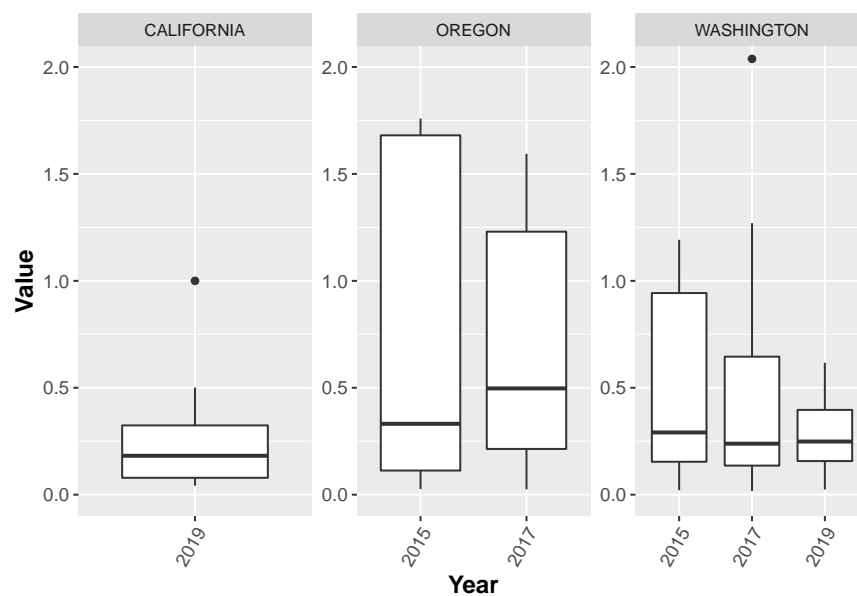
state and value

Look at 2015, the median of value of Oregon and Washington was much the same, around 0.26, but the value of Washington was closer. In 2017, the median of value in Oregon is almost two times of that in Washington. In 2019, there was no obvious difference between the situation of California and Washington.



Year and Value

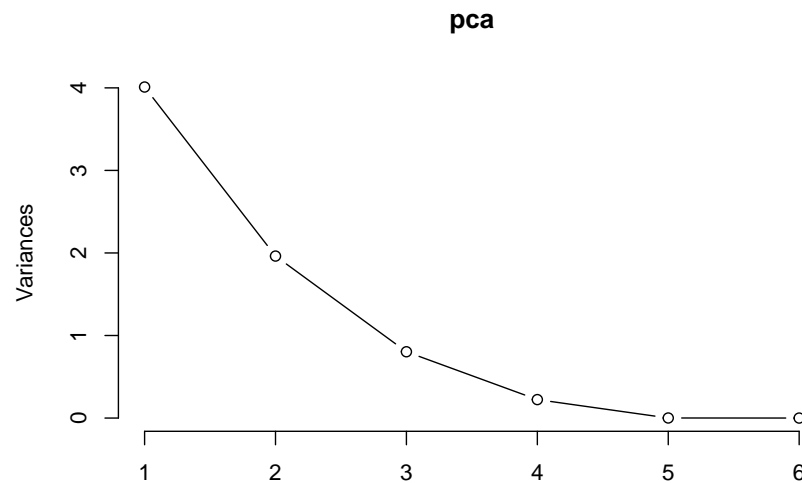
The dataset only have one year of value in California so there is no much to discuss. It seems that Oregon improves a lot from 2015 to 2017, for the median of value in 2017 is almost two times of that in 2015. And the value is closer as well. The same with Oregon, Washington also decreases the range of value, but there was not increasing in median, and even the median decreased in each year.



Correlations between chemical type.

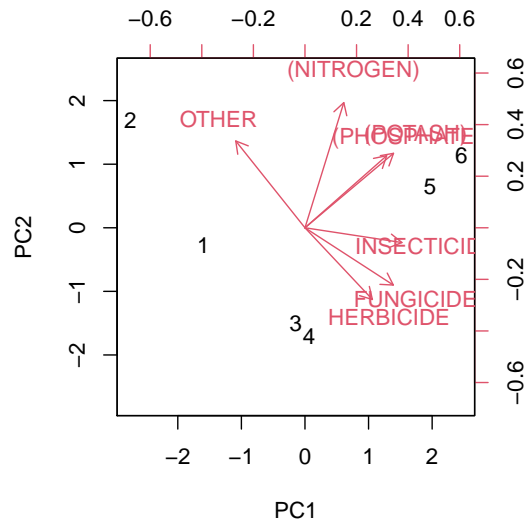
Without going into too much detail, Principal Component Analysis (PCA) can thus be used to reduce the dimensions of the data into fewer components that would retain as much of the variability expressed by the original data as possible. The fewer components would assist in describing the relationship between the original variables by projecting them onto a two-dimensional grid allowing for easier visualization from which similar types of chemical may be grouped together.

```
## Importance of components:
##          PC1    PC2    PC3    PC4    PC5    PC6
## Standard deviation  2.0025 1.4008 0.8962 0.47286 0.03009 1.135e-16
## Proportion of Variance 0.5729 0.2803 0.1147 0.03194 0.00013 0.000e+00
## Cumulative Proportion 0.5729 0.8532 0.9679 0.99987 1.00000 1.000e+00
```



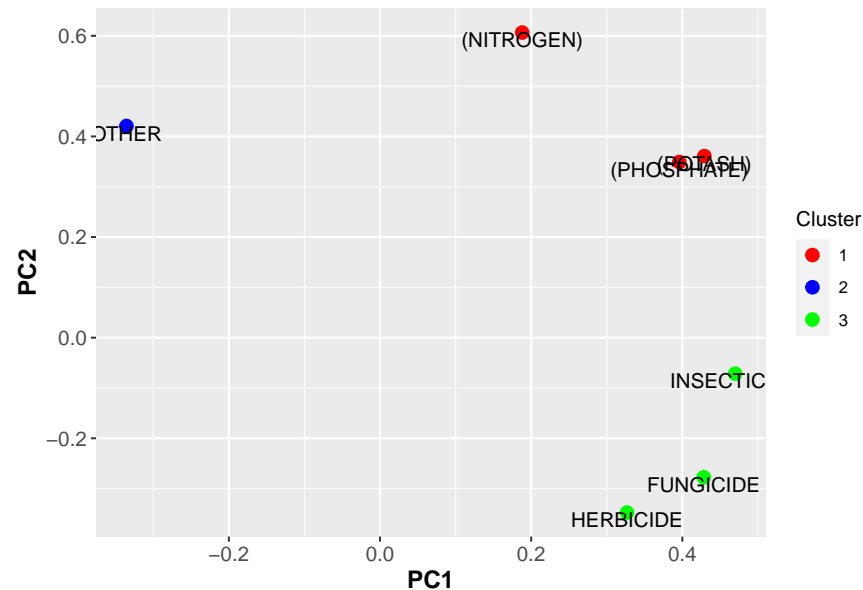
The PCA provides 6 components and 90% of the total variance is attributed to the first 4 components. See in plot above.

```
biplot(pca2.1, scale = 0)
```



And in the biplot I can see the relationship between each variables. The size of the angle between vectors determines the correlation of the variables, which is the desired indicator to achieve the objective for this analysis. A small angle indicates a strong positive correlation, 90 degrees represents no correlation and 180 degrees represents a negative correlation.

For example, Phosphate and potash is almost coincide; others and herbicide have negative correlation.



In this plot, it is clear that chemical can be separate into to three part, just I mentioned before, apart from ‘other’, fertilizer(‘nitrogen’, ‘phosphate’ and ‘potash’) gathered as cluster1, and agentia(‘fungicide’, ‘herbicide’, ‘insecticide’) gathered as cluster3.

Conclusion(What I learned)

In this assignment, cleaning data is a big and important job for the data is a little bit messy in some columns. At first, I cannot understand meaning of each column, So I go to the website USDA to find more information. When cleaning the data, I have to be patient and careful enough, and know what I actually want. In the EDA part, I met lots of difficulties, such as how to put two plots with different scales together, how to use kable to show the dataframe without overflow and the plot is not what I expected, so on. Sometimes, I give up, but most of time I found the solution from textbook or the Internet. And I realized that the type of column is really important for plots(such as character, numeric etc). And I believe that next time I will be better!

Citation

- [1] Exploratory data analysis into the relationship between different types of crime in London
- [2] R for Data Science
- [3] tutorial for grid.arrange
- [4] dmorison/eda-relationships-between-crime-london
- [5] USDA