Unit1 Final Assignment Berries

Hao Shen

2020/10/17

Summary

In this assignment, we do EAD for data set berries from USDA. After data cleaning and according to data set attributions, we divided the whole berries date set into two parts:

- Bearing: contains information about farm chemical usage.
- Market: with price, yield, area related information.

Besides, we deploy a shiny app Berries-shiny for data display and the document recorded the whole processes is in Berries-rmd.

Data Cleaning

Main Problem: How to split strings?

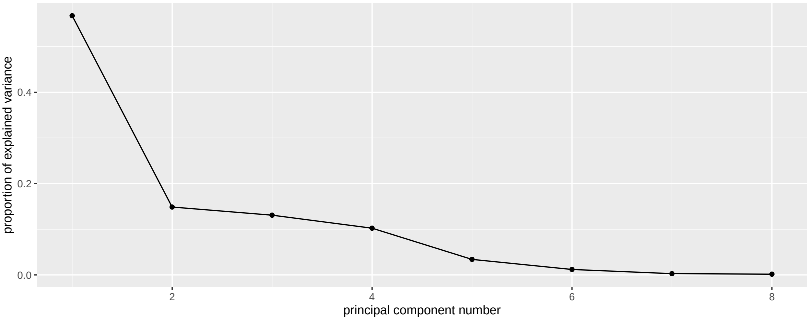
- Multiple variables contained in just one column
- For example: 'CHEMICAL, FUNGICIDE: (CAPTAN=81301)'
- Contains three parts of information: Domain, Category, Name

Solution: Plug '#' as the separator into right places

```r
str_replace_all('CHEMICAL, FUNGICIDE: (CAPTAN=81301)',
                pattern=c(', '='#',': '='#'))%>%
  strsplit(split="#")
```

Domain	Category	Name
CHEMICAL	FUNGICIDE	(CAPTAN=81301)

PCA for bearing



PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
0.5676	0.7163	0.8472	0.9495	0.9835	0.9955	0.9983	1

Just as the plot shows, it requires 4 out of 8 principal components

Equations for market

For market data, we just found their relationships are extremely simple:

- ACRE_HARVEST(ACRE)=PROD(LB)/YIELD(LB/ACRE)
- M_PRICE_RECEIVED($/LB)=M_PROD($)/M_PROD(LB)
- M_PROD($)=M_PRICE_RECEIVED($/LB)*M_PROD(LB)
- M_PROD(LB)=M_PROD($)/M_PRICE_RECEIVED($/LB)
- NS_PROD(LB)=PROD(LB)-UTILIZED_PROD(LB)
- PRICE_RECEIVED($/LB)=U_PROD($)/U_PROD(LB)
- P_PRICE_RECEIVED($/LB)=P_PROD($)/P_PROD(LB)
- P_PROD($)=P_PRICE_RECEIVED($/LB)*P_PROD(LB)
- P_PROD(LB)=P_PROD($)/P_PRICE_RECEIVED($/LB)
- PROD(LB)=NS_PROD(LB)+U_PROD(LB)
- U_PROD($)=PRICE_RECEIVED($/LB)*U_PROD(LB)
- U_PROD(LB)=M_PROD(LB)+P_PROD(LB)
- YIELD(LB/ACRE)=PROD(LB)/ACRE_HARVEST(ACRE)