Homework2

201982188 통계학과 박현주

2020/9/14

```
#install.packages('alr3')
library(alr3);library(tidyverse)
```

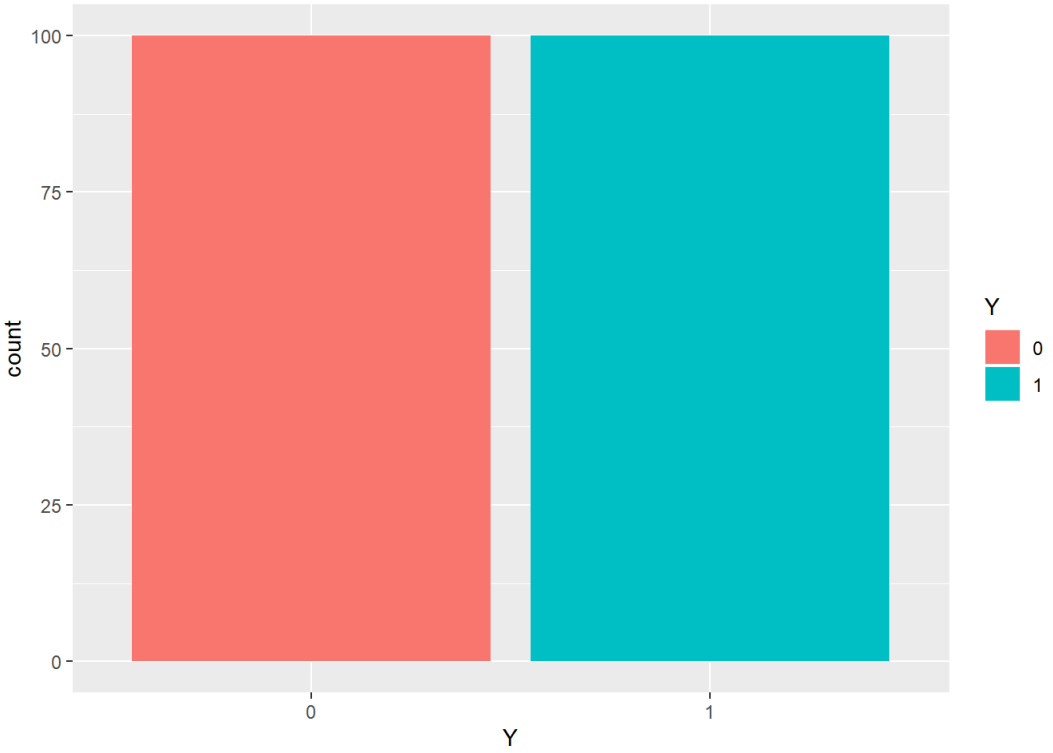
1. Bar chart

```
data(banknote)
banknote$Y <- as.factor(banknote$Y)
banknote %>%glimpse()
```

```
## Observations: 200
## Variables: 7
## $ Length   <dbl> 214.8, 214.6, 214.8, 214.8, 215.0, 215.7, 215.5, 214.5, 21...
## $ Left     <dbl> 131.0, 129.7, 129.7, 129.7, 129.6, 130.8, 129.5, 129.6, 12...
## $ Right    <dbl> 131.1, 129.7, 129.7, 129.6, 129.7, 130.5, 129.7, 129.2, 12...
## $ Bottom   <dbl> 9.0, 8.1, 8.7, 7.5, 10.4, 9.0, 7.9, 7.2, 8.2, 9.2, 7.9, 7....
## $ Top      <dbl> 9.7, 9.5, 9.6, 10.4, 7.7, 10.1, 9.6, 10.7, 11.0, 10.0, 11....
## $ Diagonal <dbl> 141.0, 141.7, 142.2, 142.0, 141.8, 141.4, 141.6, 141.7, 14...
## $ Y        <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

- banknote data는 200개의 관측치와 7개의 변수로 구성된다. 변수에는 지폐에 대한 Length, Left, Right, Bottom, Top, Diagonal과 geunine(진품)인지 counterfeit(가품)인지를 구분해주는 Y가 있다.

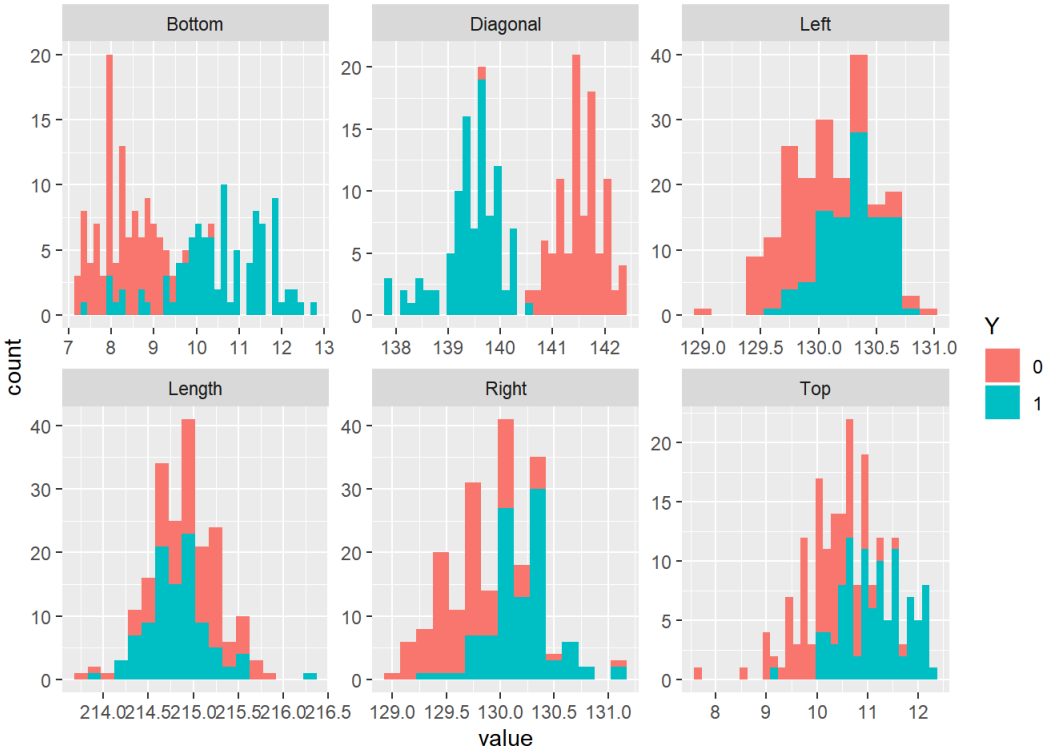
```
banknote %>% ggplot(aes(x=Y,fill=Y))+
  geom_bar()
```



- 위 그림은 bar chart로 Y의 각 구성 요소 빈도를 한 눈에 보기 쉬운 그림이다. x축은 Y의 구성 요소이고 y축은 각 구성 요소의 개수를 의미한다. 빨간색과 파란색으로 Y의 0(genuine)과 1(counterfeit)이 구분되어 있고 각각 100개가 있음을 알 수 있다.

2. Histogram

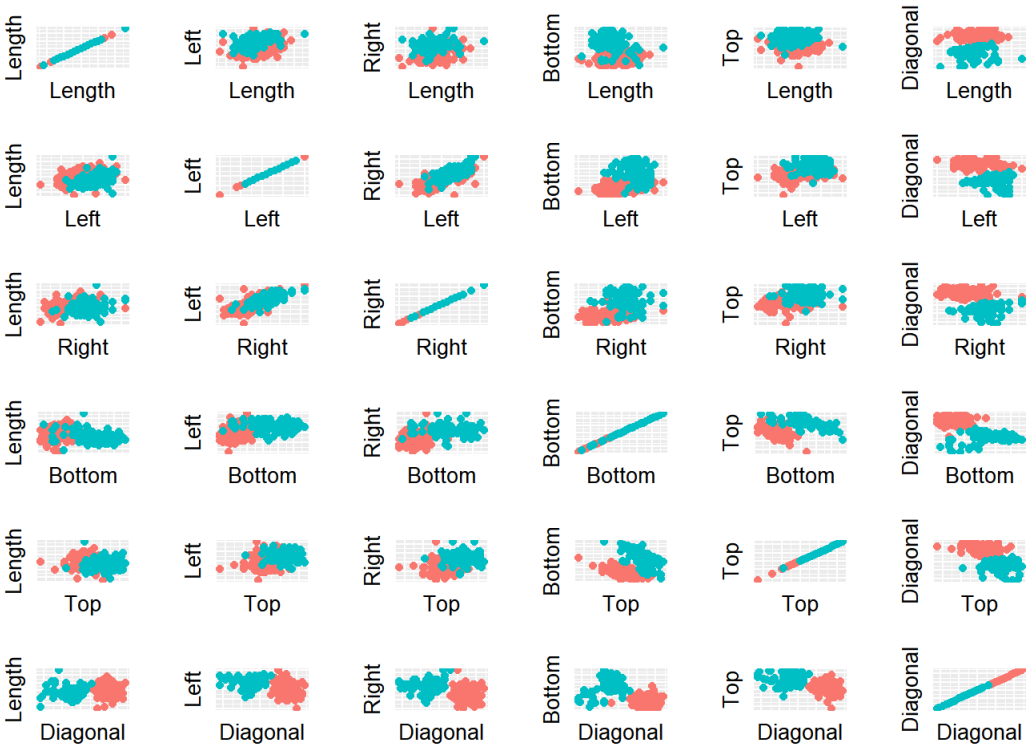
```
banknote %>%
  gather(key='key',value='value',-Y) %>%
  ggplot(aes(x=value, fill=Y)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram(binwidth = .15)
```



- Y를 제외한 6개의 변수를 이용해 위와 같은 히스토그램을 그렸다. 각 변수에 대해 빨간색으로 표현된 0(genuine)과 파란색으로 표현된 1(counterfeit)의 그림을 구분해 그렸다.
- 두 히스토그램의 차이가 가장 뚜렷한 변수는 Diagonal임을 알 수 있다. 위조 지폐일 경우 대각선의 길이가 그렇지 않은 경우보다 더 짧다. Bottom의 경우도 0과 1을 구분해줄 수 있어 보인다. 하단 여백의 너비가 위조 지폐가 더 긴 경우가 많다. 그에 반해 Left, Length, Right은 그 차이가 분명하지 않다.

3. Scatter plot

```
library(magrittr)
#install.packages('multipanelfigure')
library(multipanelfigure)

figure1 <- multi_panel_figure(columns = 6, rows = 6, panel_label_type = "none")


for( i in 1:6){#(ncol(banknote)-1)){
  for( j in 1:6){#(ncol(banknote)-1)){


      #a <- append(a,idx)
      plot_s <- ggplot(banknote,aes(x=banknote[,i],y=banknote[,j], color=Y)) +
              geom_point(show.legend = FALSE)+
              xlab(colnames(banknote)[i])+
              ylab(colnames(banknote)[j])+
        theme(axis.title.x=element_text(size=10),
              axis.text.x=element_blank(),
              axis.ticks.x=element_blank(),
              axis.title.y=element_text(size=10),
              axis.text.y=element_blank(),
              axis.ticks.y=element_blank())

      figrue1 <- figure1 %<>%
                  fill_panel(plot_s, column = j, row = i)

  }
}

figure1
```



- 정품 지폐와 위조 지폐를 완벽하게 구분할 수있는 한 쌍의 측정 값을 찾아보기 위해 위와 같은 산점도를 그렸다. 어떻게 분포돼있는지 알아보는데 목적이 있기 때문에 변수 값들의 범위는 표시하지 않았다.(x축과 y축)
- Y의 값이 확연하게 구분되는 그림은 Diagonal과 다른 변수들간의 산점도이다. 문제 2번에서 histogram을 통해 확인한 결과와 같음을 알 수 있다.
- 정품과 가품의 가장 잘 구분할 수 있는 또 다른 값에는 Bottom과 Top의 측정값이 있다. Bottom과 다른 변수들간의 산점도를 보면 다른 변수 쌍들에 비해 정품과 가품이 주로 구분이 된다는 것을 알 수 있다.