Twin Studies on Smoking

barplot

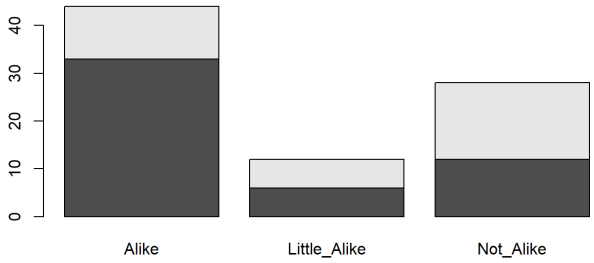
매트릭스를 막대그래프로 그릴 때 매트릭스 모양 그대로 표현하려 함에 유의하여야 한다. 즉, 각 열을 각 막대에 대응시키면서 행으로 주어지는 각각의 값들을 각 막대에 나누어 배분하는 디폴트로 한다(`beside = FALSE`). 따라서 흡연 습관에 대한 쌍둥이 연구의 집계 표를 원하는 막대 그래프로 표현하려면, 즉 쌍둥이의 유형에 따라 닮은 정도를 표현하려면, 현재 나와 있는 표 구조를 전치(transpose)시켜야 한다.

```
library(magrittr)
library(tidyverse)
# par(mfrow = c(1, 2))
#> 제시된 표와 닮은 행열 생성
Nature1 <- matrix(c(33, 11, 6, 6, 12, 16),
                  nrow = 2)
rownames(Nature1) <- c("Identical", "Fraternal")
colnames(Nature1) <- c("Alike", "Little_Alike", "Not_Alike")
Nature1
```

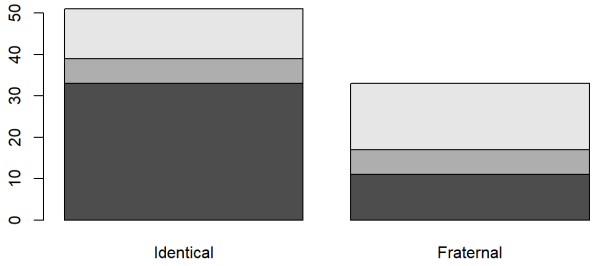
```
##           Alike Little_Alike Not_Alike
## Identical    33            6        12
## Fraternal    11            6        16
```

행렬 구조와 barplot

```
Nature1 %>%
  barplot
```




```
Nature1 %>%
  t %>%
  barplot
```



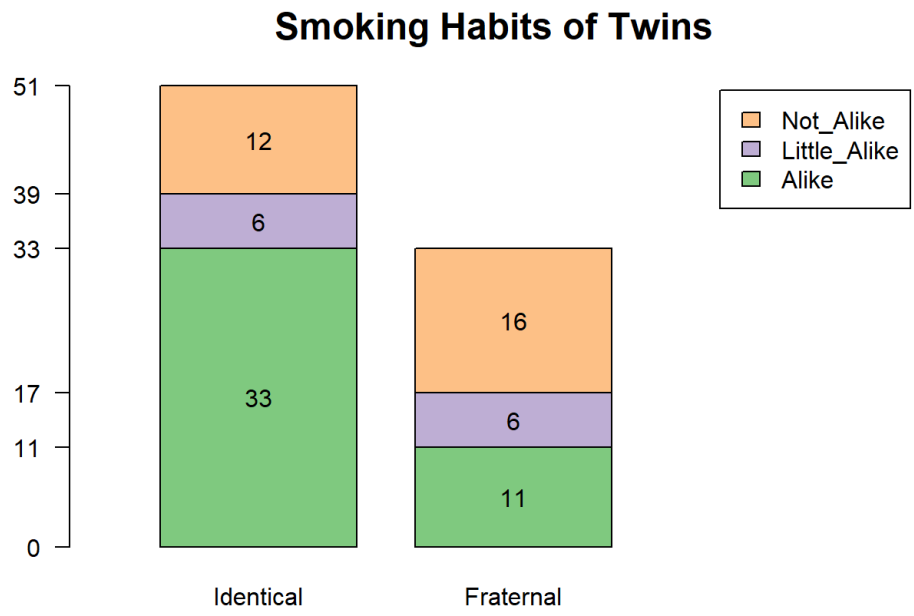

```
# par(mfrow = c(1, 1))
```

`Nature1` 의 구조를 전치(transpose)해 주어야 원하는 모양의 막대그래프가 나올 것임을 알 수 있다. 나머지 이러한 점을 염두에 두고 작성한다.

Stack

```
options(digits = 3)
#> RColorBrewer 패키지를 이용하여 컬러 생성
library(RColorBrewer)
#> "Accent" palette 채택
cols <- brewer.pal(8, "Accent")
#> 막대의 가운데에 추가 정보를 넣기 위한 좌표 설정 함수.
# pos <- function(x){
#   cumsum(x) - x / 2
# }
pos <- . %>% {`-`(cumsum(.), . / 2)}
# pos <- . %>% {cumsum(.) - . / 2}
#> 아래와 같이 작성하면 오류 발생
# pos <- . %>% cumsum(.) - . / 2
#> 텍스트 정보 넣을 좌표를 계산한다.
y1_text <- apply(Nature1,
                 MARGIN = 1,
                 FUN = pos)
```

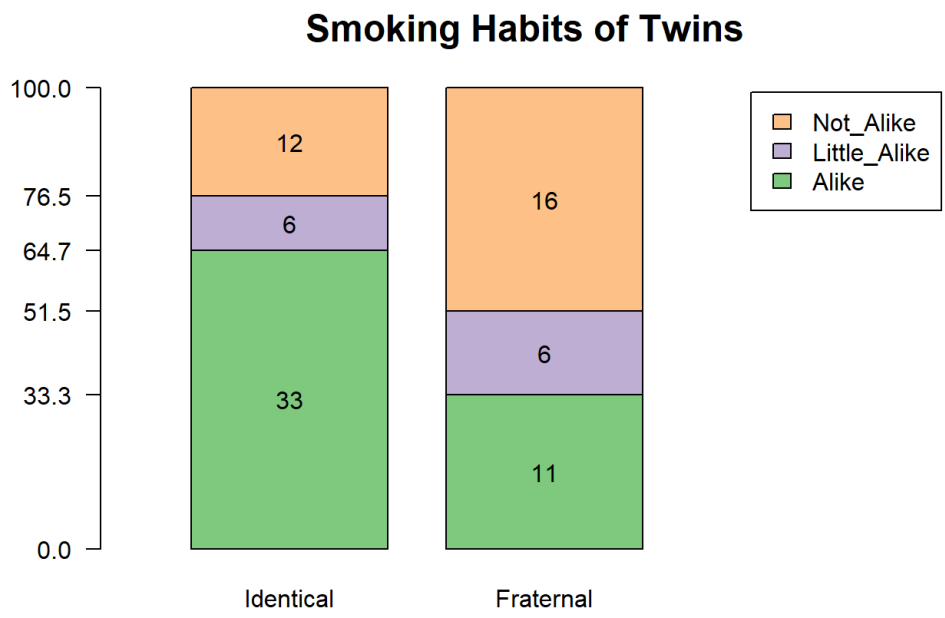
```
b1 <- Nature1 %>%
  t %>%
  barplot(width = 1,
          xlim = c(0, 4),
          space = 0.3,
          col = cols[1:3],
          yaxt = "n")
#> 쌍둥이유형 별로 한 막대에 흡연습관의 닮음 정도를 나타낼 것이므로 `cumsum`함수를 이용하여 막대들이 위치할 좌표를 계산한다. 일란성과 이란성 각각의 수효부터 비교할 수 있도록  막대 높이로 나타내고, 막대 중심에는 해당 속성의 돗수를 표시한다. 원점을 나타내기 위하여 0을 `c`함수 안에 추가하였다. 이를 추가하지 않으면 축이 어떻게 표시되는지 비교한다.
axis(side = 2,
     at = c(0, apply(t(Nature1),
                     MARGIN = 2,
                     FUN = cumsum)),
     labels = format(c(0, apply(t(Nature1),
                                MARGIN = 2,
                                FUN = cumsum)),
                     digits = 3,
                     nsmall = 0),
     las = 2)
#> 막대그래프 작성 과정에서 나온 막대의 좌표와 `pos`함수로 계산한 y좌표를 이용하여 실제 관찰된 쌍둥이 페어의 수효를 표시한다.`y_text`의 구조에 맞추어 `rep()`에서 `each = 3`으로 설정하였다. `bty = ` `"o" 또는 "n"으로 정할 수 있다.
text(x = rep(b1, each = 3),
     y = y1_text,
     labels = t(Nature1))
#> 범례 표시
legend("topright",
       inset = 0.01,
       fill = cols[3:1],
       legend = rev(colnames(Nature1)),
       bty = "o")
#> 메인 타이틀
title(main = "Smoking Habits of Twins",
      cex.main = 1.5)
```



Fill

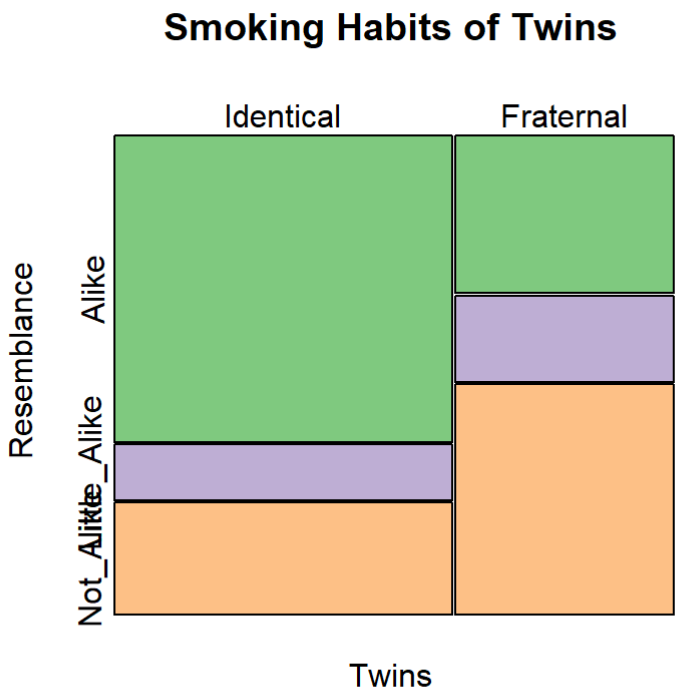
```
#> 쌍둥이유형(일란성/이란성) 별로 흡연습관을 구분하여 각 습관의 백분율을 계산한다. 일란성 쌍둥이와 이란성 쌍둥이의 숫자가 다르기 때문에 공평하게 비교하려면 백분율을 비교하는 것이 타당하다.
Nature1_p <- Nature1 %>%
  prop.table(margin = 1) %>%
  `*`(100)
# Nature1_p <- prop.table(Natrue1, margin = 1) * 100
#> 차곡차곡 쌓아놓은 막대그래프를 그리고(`beside = FALSE`가 디폴트) 추가 정보를 표시할 부분 막대의 가운데 좌표를 저장한다.
#> `width`를 설정하려면 `xlim`도 함께 설정하여야 함에 유의한다.
#> 아래 예시의 경우 막대그래프의 범위를 0에서 4까지로 하면서 첫번째 막대의 중심은 1.5,
#> 두번째 막대의 중심은 3.5에 위치한다.(b1 값으로 파악) 막대의 폭(width)을 1,
#> 두 막대 간의 간격(space) 또한 1로 하여 width, xlim, space 간의 관계를 쉽게 알 수 있도록 하였다.
#> 쌍둥이유형별로 비교하여야 하므로 행렬을 전치시켜서 막대그래프를 그려야 한다.
#>  `yaxt = "n"`을 설정하여 y축에 추가 정보를 넣을 수 있도록 하였다.
#> 텍스트 정보 넣을 좌표를 계산한다.
y1_text_p <- Nature1_p %>%
  apply(MARGIN = 1,
        FUN = pos)
```

```
b1_p <- Nature1_p %>%
  t %>%
  barplot(width = 1,
          xlim = c(0, 4),
          space = 0.3,
          col = cols[1:3],
          yaxt = "n")
#> 쌍둥이유형 별로 한 막대에 흡연습관의 닮음 정도를 나타낼 것이므로 `cumsum`함수를 이용하여 막대들이 위치할 좌표를 계산한다. 일란성과 이란성을 각각 100%로 하고 닮음 정도의 백분율을 막대 높이로 비교할 수 있도록 하되, 막대 중심에는 해당 속성의 돗수를 표시한다. 원점을 나타내기 위하여 0을 `c`함수 안에 추가하였다.
axis(side = 2,
     at = c(0, apply(t(Nature1_p),
                     MARGIN = 2,
                     FUN = cumsum)),
     labels = format(c(0, apply(t(Nature1_p),
                                MARGIN = 2,
                                FUN = cumsum)),
                     digits = 3,
                     nsmall = 1),
     las = 2)
#> 막대그래프 작성 과정에서 나온 막대의 좌표와 `pos`함수로 계산한 y좌표를 이용하여 실제 관찰된 쌍둥이 페어의 수효를 표시한다.`y1_text_p`의 구조에 맞추어 `rep()`에서 `each = 3`으로 설정하였다. `bty = ` `"o" 또는 "n"으로 정할 수 있다.
text(x = rep(b1_p, each = 3),
     y = y1_text_p,
     labels = t(Nature1))
#> 범례 표시
legend("topright",
       inset = 0.01,
       fill = cols[3:1],
       legend = rev(colnames(Nature1)),
       bty = "o")
#> 메인 타이틀
title(main = "Smoking Habits of Twins",
      cex.main = 1.5)
```



Mosaic Plot

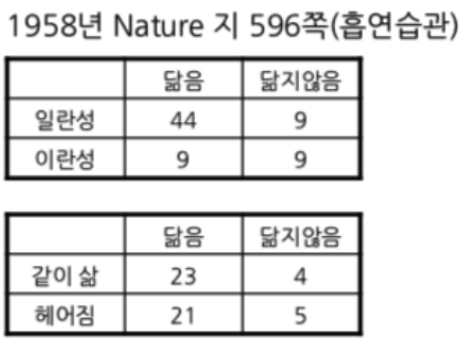
```
mosaicplot(Nature1,
           main = "Smoking Habits of Twins",
           xlab = "Twins",
           ylab = "Resemblance",
           off = c(0.5, 1),
           color = cols[1:3],
           cex.axis = 1,
           las = 0)
```



Nature 1958 version 2

피셔의 논문이 실려 있는 Nature 지의 596쪽에서는 그와 비슷한 연구 결과가 실려 있었다. 피셔가 인용한 보고서가 흡연습관을 세 단계로 구분한 것과는 달리 닮았거나 그렇지 않거나의 이분법으로 나누었다. 이 보고서가 흥미로운 것은 단순히 일란성 쌍둥이와 이란성 쌍둥이의 흡연습관을 비교한 것이 아니라 일란성 쌍둥이들을 다시 어려서 헤어진 경우와 함께 산 경우로 나눠 본 것이다. 일란성 쌍둥이들은 함께 살았든, 헤어져 살았든 흡연습관에 있어서도 놀라울 정도로 닮은 점을 보여 준다.

```
include_graphics("../pics/Nature_1958v2.png")
```



1958년 Nature 지 596쪽(흡연습관)

	닮음	닮지않음
일란성	44	9
이란성	9	9

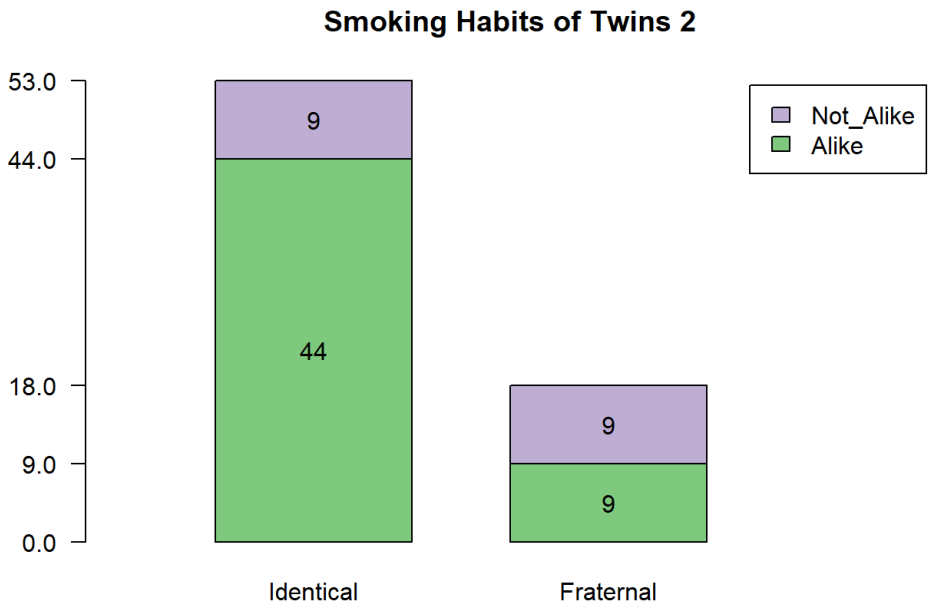
	닮음	닮지않음
같이 삶	23	4
헤어짐	21	5

막대그래프 : Stack

```
Nature2 <- matrix(c(44, 9, 9, 9),
                  nrow = 2)
rownames(Nature2) <- c("Identical", "Fraternal")
colnames(Nature2) <- c("Alike", "Not_Alike")
Nature2
```

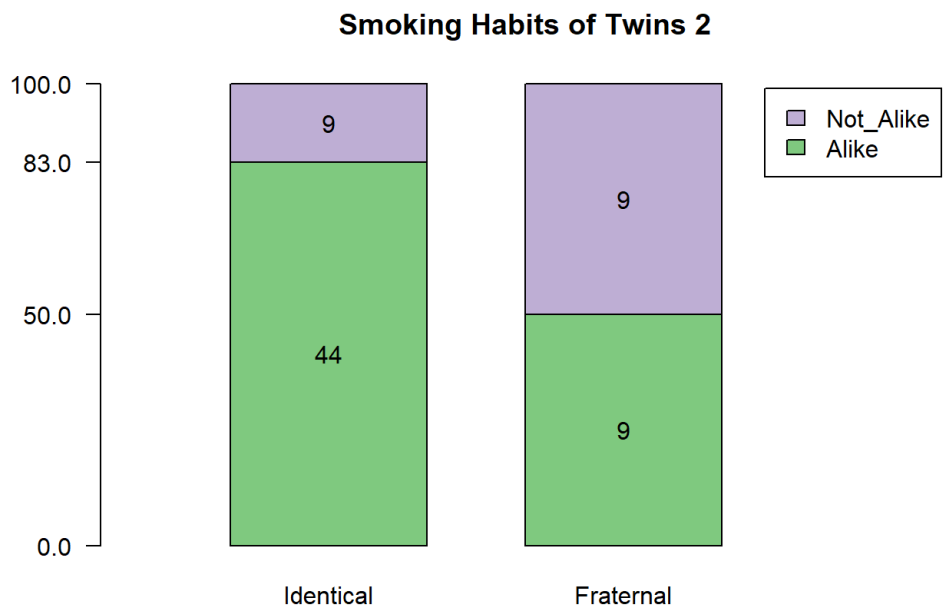
```
##           Alike Not_Alike
## Identical    44         9
## Fraternal     9         9
```

```
y2_text <- apply(Nature2,
                 MARGIN = 1,
                 FUN = pos)
b2 <- barplot(t(Nature2),
              width = 1,
              xlim = c(0, 4),
              space = 0.5,
              col = cols[1:2],
              yaxt = "n")
axis(side = 2,
     at = c(0, apply(t(Nature2),
                     MARGIN = 2,
                     FUN = cumsum)),
     labels = format(c(0, apply(t(Nature2),
                                MARGIN = 2,
                                FUN = cumsum)),
                     digits = 3,
                     nsmall = 1),
     las = 2)
text(x = rep(b2, each = 2),
     y = y2_text,
     labels = t(Nature2))
legend("topright",
       inset = 0.01,
       fill = cols[2:1],
       legend = rev(colnames(Nature2)))
title(main = "Smoking Habits of Twins 2")
```



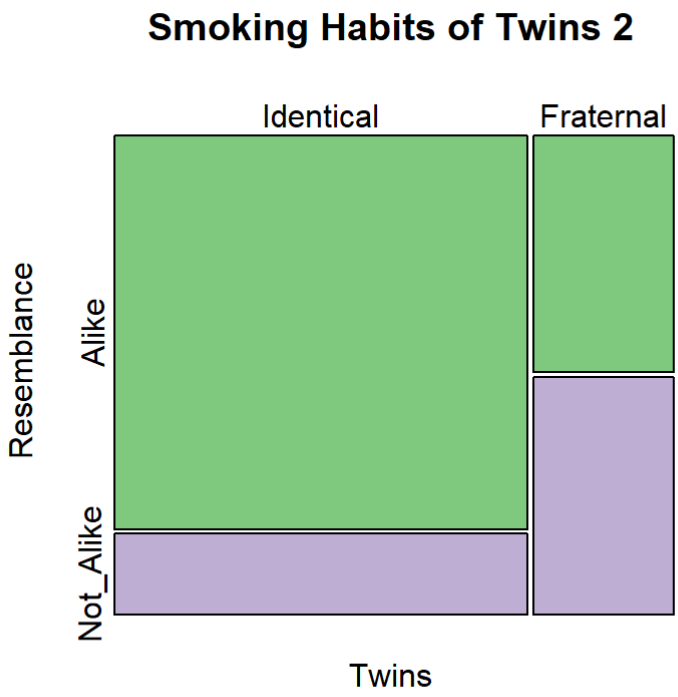
막대그래프 : Fill

```
Nature2_p <- prop.table(Nature2,
                        margin = 1) * 100
y2_text_p <- apply(Nature2_p,
                   MARGIN = 1,
                   FUN = pos)
b2_p <- barplot(t(Nature2_p),
                width = 1,
                xlim = c(0, 4),
                space = 0.5,
                col = cols[1:2],
                yaxt = "n")
axis(side = 2,
     at = c(0, apply(t(Nature2_p),
                     MARGIN = 2,
                     FUN = cumsum)),
     labels = format(c(0, apply(t(Nature2_p),
                                MARGIN = 2,
                                FUN = cumsum)),
                     digits = 3,
                     nsmall = 1),
     las = 2)
text(x = rep(b2_p, each = 2),
     y = y2_text_p,
     labels = t(Nature2))
legend("topright",
       inset = 0.01,
       fill = cols[2:1],
       legend = rev(colnames(Nature2)))
title(main = "Smoking Habits of Twins 2")
```



Mosaic Plot

```
mosaicplot(Nature2,
           main = "Smoking Habits of Twins 2",
           xlab = "Twins",
           ylab = "Resemblance",
           off = 1,
           color = cols[1:2],
           cex.axis = 1)
```



Nature 1958 version 2 : Identical Twins

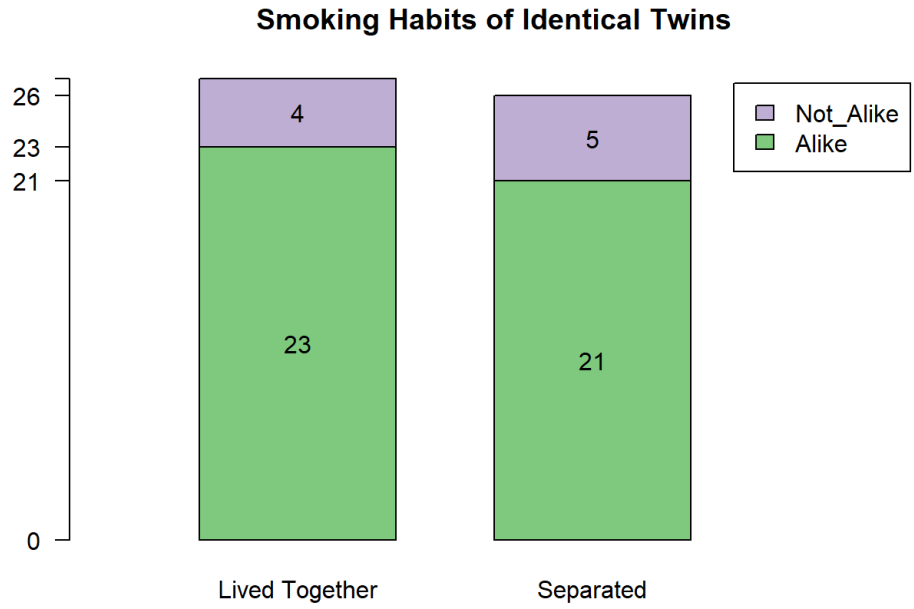
일란성 쌍둥이들만을 대상으로 어렸을 때 헤어졌는지, 함께 살았는지 여부와 흡연습관을 비교한 결과는 놀라울 정도 닮았다는 것을 보여준다.

막대그래프 : stack

```
Nature3 <- matrix(c(23, 21, 4, 5),
                  nrow = 2)
rownames(Nature3) <- c("Lived Together", "Separated")
colnames(Nature3) <- c("Alike", "Not_Alike")
Nature3
```

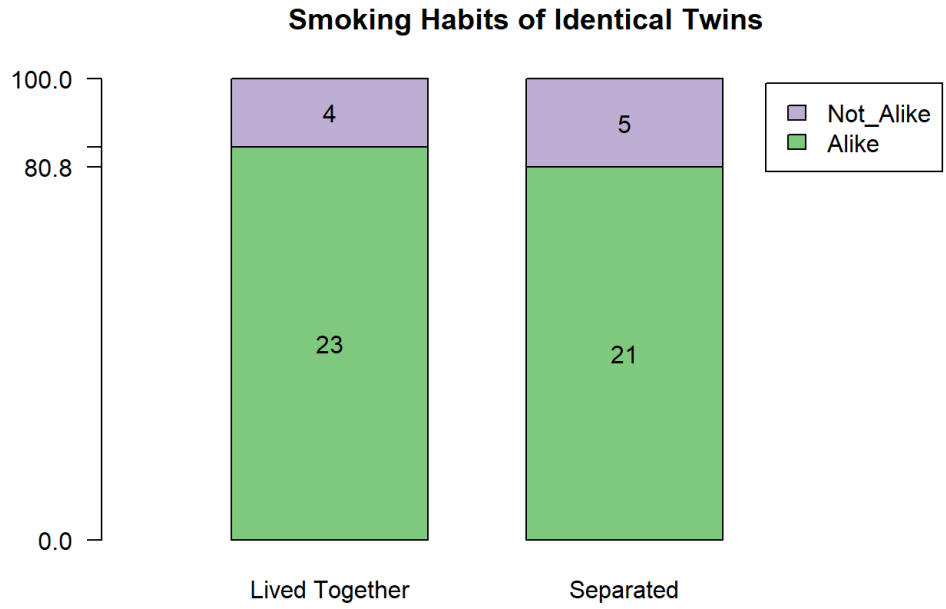
```
##                Alike Not_Alike
## Lived Together    23         4
## Separated         21         5
```

```
c3 <- ncol(Nature3)
b3 <- barplot(t(Nature3),
              width = 1,
              xlim = c(0, 4),
              space = 0.5,
              col = cols[1:2],
              yaxt = "n")
axis(side = 2,
     at = c(0, apply(t(Nature3),
                     MARGIN = 2,
                     FUN = cumsum)),
     labels = format(c(0, apply(t(Nature3),
                                MARGIN = 2,
                                FUN = cumsum)),
                     digits = 3,
                     nsmall = 0),
     las = 2)
y3_text <- apply(Nature3,
                 MARGIN = 1,
                 FUN = pos)
text(x = rep(b3, each = 2),
     y = y3_text,
     labels = t(Nature3))
legend("topright",
       inset = 0.01,
       fill = cols[2:1],
       legend = rev(colnames(Nature2)))
title(main = "Smoking Habits of Identical Twins")
```



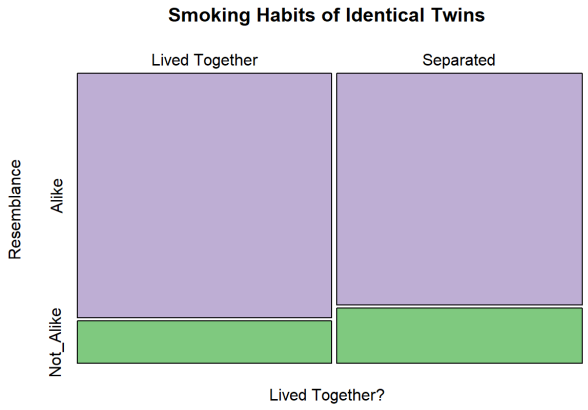
막대그래프 : Fill

```
Nature3_p <- prop.table(Nature3,
                        margin = 1) * 100
b3_p <- barplot(t(Nature3_p),
                width = 1,
                xlim = c(0, 4),
                space = 0.5,
                col = cols[1:2],
                yaxt = "n")
axis(side = 2,
     at = c(0, apply(t(Nature3_p),
                     MARGIN = 2,
                     FUN = cumsum)),
     labels = format(c(0, apply(t(Nature3_p),
                                MARGIN = 2,
                                FUN = cumsum)),
                     digits = 3,
                     nsmall = 1),
     las = 2)
y3_text_p <- apply(Nature3_p,
                   MARGIN = 1,
                   FUN = pos)
text(x = rep(b3, each = 2),
     y = y3_text_p,
     labels = t(Nature3))
legend("topright",
       inset = 0.01,
       fill = cols[2:1],
       legend = rev(colnames(Nature2)))
title(main = "Smoking Habits of Identical Twins")
```



Mosaic Plot

```
mosaicplot(Nature3,
           main = "Smoking Habits of Identical Twins",
           xlab = "Lived Together?",
           ylab = "Resemblance",
           off = 1,
           color = cols[2:1],
           cex.axis = 1)
```



ggplot

tidyverse

깔끔한 데이터 바꾸는 과정에서 유의할 점은 ggplot으로 그릴 때 어떤 변수가 x, y, fill 역할을 할 것인지를 명확히 하여야한다. 어떤 변수를 맨 앞에 위치시키고, 어떤 변수를 그 다음 자리에 위치시키고, 그 다음에 Counts를 위치시킨다. 즉, fill 에 해당하는 변수를 맨 앞에, 그리고 x에 해당하는 변수를 그 다음에, 마지막으로 세번째 변수로 Freq 또는 Counts가 위치하도록 tidy를 적용하면 ggplot으로 그릴 때 상당히 체계적인 접근이 가능해진다. 특히, 막대들의 중간에 추가적인 정보를 삽입하기 위하여 좌표를 계산할 필요가 있을 때 크게 도움이 된다.

```
Nature1_tbl <- Nature1 %>%
  t %>%
  as_tibble %>%
  mutate(Resemblance = row.names(t(Nature1))) %>%
  gather(key = "Twins",
         value = "Counts", -Resemblance) %>%
  mutate(Resemblance = factor(Resemblance,
                              levels = c("Alike",
                                         "Little_Alike",
                                         "Not_Alike")),
         Twins = factor(Twins,
                        levels = c("Identical", "Fraternal")))
Nature1_tbl
```

```
## # A tibble: 6 x 3
##   Resemblance  Twins     Counts
##   <fct>        <fct>      <dbl>
## 1 Alike        Identical     33
## 2 Little_Alike Identical      6
## 3 Not_Alike    Identical     12
## 4 Alike        Fraternal     11
## 5 Little_Alike Fraternal      6
## 6 Not_Alike    Fraternal     16
```

```
Nature2_tbl <- Nature2 %>%
  t %>%
  as_tibble %>%
  mutate(Resemblance = row.names(t(Nature2))) %>%
  gather(key = "Twins",
         value = "Counts", -Resemblance) %>%
  mutate(Twins = factor(Twins,
                        levels = c("Identical", "Fraternal")),
         Resemblance = factor(Resemblance)) %>%
  print
```

```
## # A tibble: 4 x 3
##   Resemblance Twins     Counts
##   <fct>       <fct>      <dbl>
## 1 Alike       Identical     44
## 2 Not_Alike   Identical      9
## 3 Alike       Fraternal      9
## 4 Not_Alike   Fraternal      9
```

```
Nature3_tbl <- Nature3 %>%
  t %>%
  as_tibble %>%
  mutate(Resemblance = row.names(t(Nature3))) %>%
  gather(key = "Separation",
         value = "Counts", -Resemblance) %>%
  mutate(Separation = factor(Separation),
         Resemblance = factor(Resemblance)) %>%
  print
```

```
## # A tibble: 4 x 3
##   Resemblance Separation     Counts
##   <fct>       <fct>           <dbl>
## 1 Alike       Lived Together     23
## 2 Not_Alike   Lived Together      4
## 3 Alike       Separated          21
## 4 Not_Alike   Separated           5
```

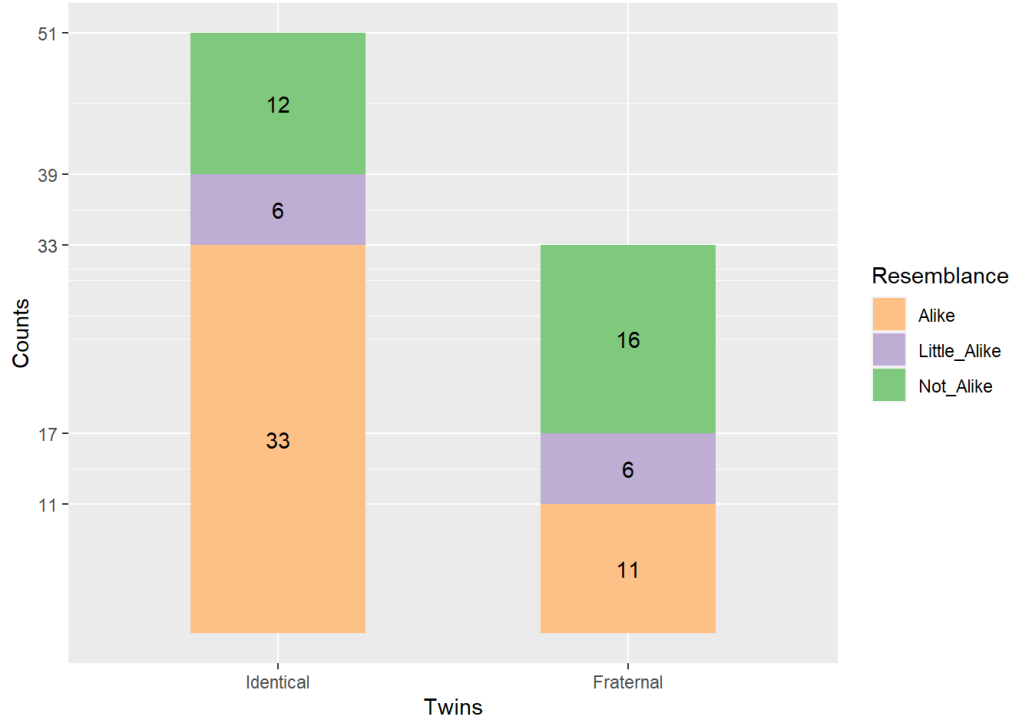
Good old `as.data.frame`

```
Nature1_tbl <- Nature1 %>%
  t %>%
  as.data.frame.table %>%
  `colnames<-`(c("Resemblance", "Twins", "Counts"))
Nature2_tbl <- Nature2 %>%
  t %>%
  as.data.frame.table %>%
  `colnames<-`(c("Resemblance", "Twins", "Counts"))
Nature3_tbl <- Nature3 %>%
  t %>%
  as.data.frame.table %>%
  `colnames<-`(c("Resemblance", "Separation", "Counts"))
```

Nature 1958 Version 1

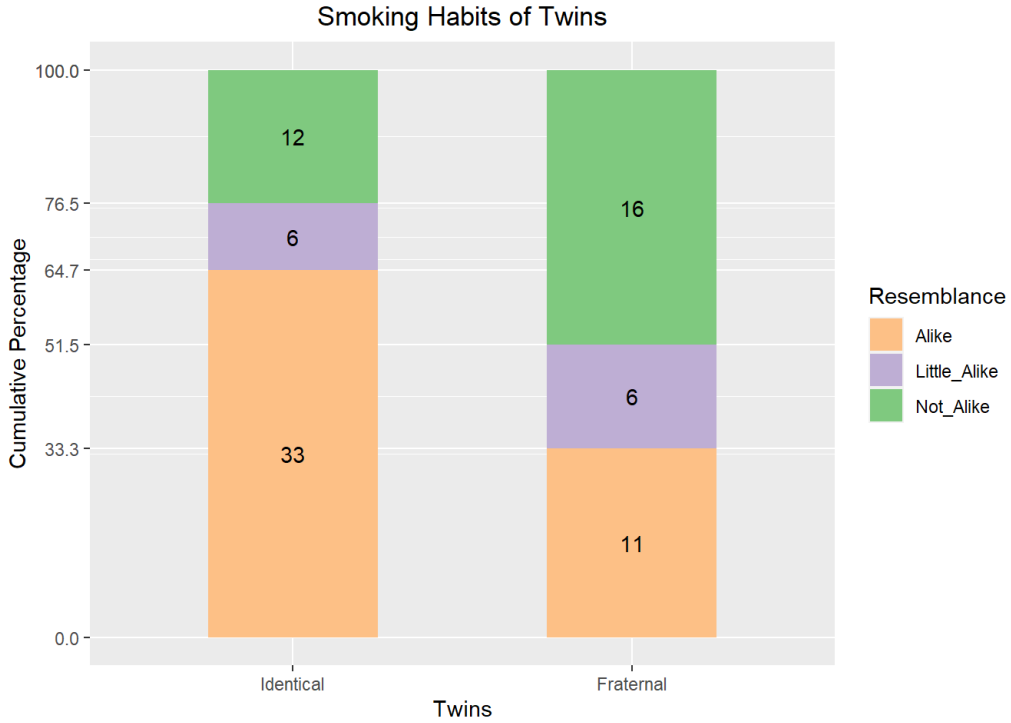
stack

```
y1_breaks <- Nature1_tbl %$%
  tapply(Counts,
         INDEX = Twins,
         FUN = cumsum) %>%
  unlist
y1_text <- Nature1_tbl %$%
  tapply(Counts,
         INDEX = Twins,
         FUN = pos) %>%
  unlist
Nature1_tbl %>%
  ggplot(data = .,
         mapping = aes(x = Twins,
                       y = Counts,
                       fill = Resemblance)) +
  geom_bar(stat = "identity",
           width = 0.5,
           position = position_stack(reverse = TRUE)) +
  geom_text(aes(y = y1_text),
            label = Nature1_tbl$Counts,
            position = "identity") +
  scale_fill_brewer(type = "qual",
                    palette = "Accent",
                    direction = -1) +
  scale_y_continuous(breaks = y1_breaks,
                     labels = y1_breaks)
```



fill

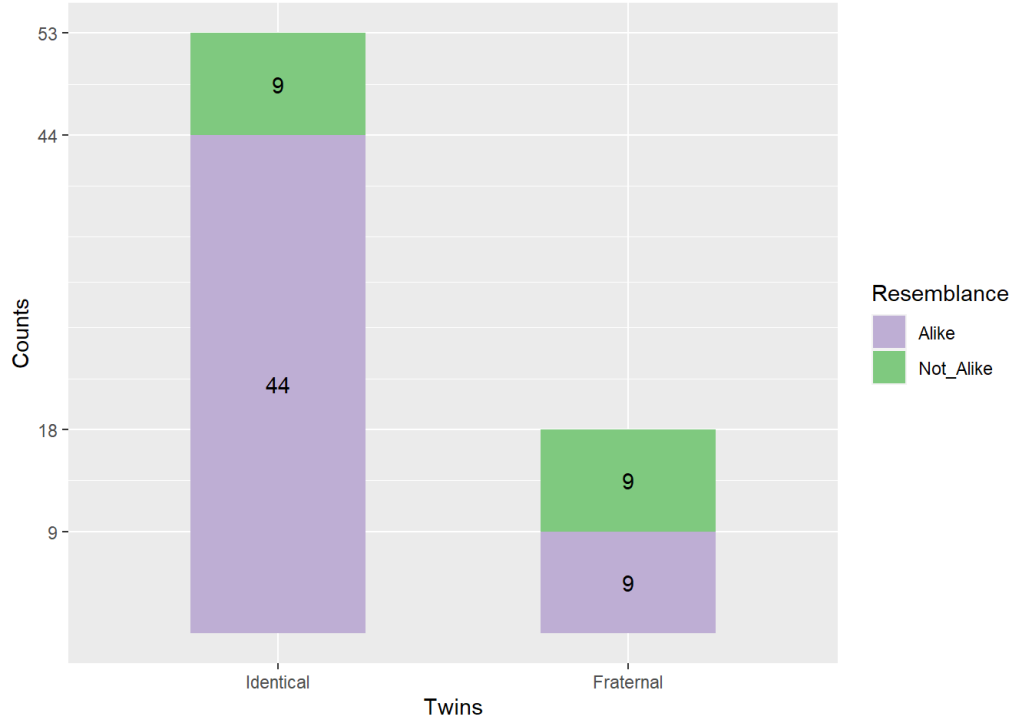
```
y1_fill <- y1_text / (Nature1 %>%
                        apply(MARGIN = 1,
                              FUN = sum) %>%
                        rep(each = 3))
Nature1_tbl %>%
  ggplot(data = .,
         mapping = aes(x = Twins,
                       y = Counts,
                       fill = Resemblance)) +
  geom_bar(stat = "identity",
           width = 0.5,
           position = position_fill(reverse = TRUE)) +
  geom_text(aes(y = y1_fill),
            label = Nature1_tbl$Counts,
            position = "identity") +
  scale_fill_brewer(type = "qual",
                    palette = "Accent",
                    direction = -1) +
  scale_y_continuous(name = "Cumulative Percentage",
                     breaks = c(0,
                                apply(t(Nature1_p),
                                      MARGIN = 2,
                                      FUN = cumsum) / 100),
                     labels = format(c(0,
                                       apply(t(Nature1_p),
                                             MARGIN = 2,
                                             FUN = cumsum)),
                                     digits = 3,
                                     nsmall = 1)) +
  labs(title = "Smoking Habits of Twins") +
  theme(plot.title = element_text(hjust = 0.5))
```



Nature 1958 version 2 : Twin Study

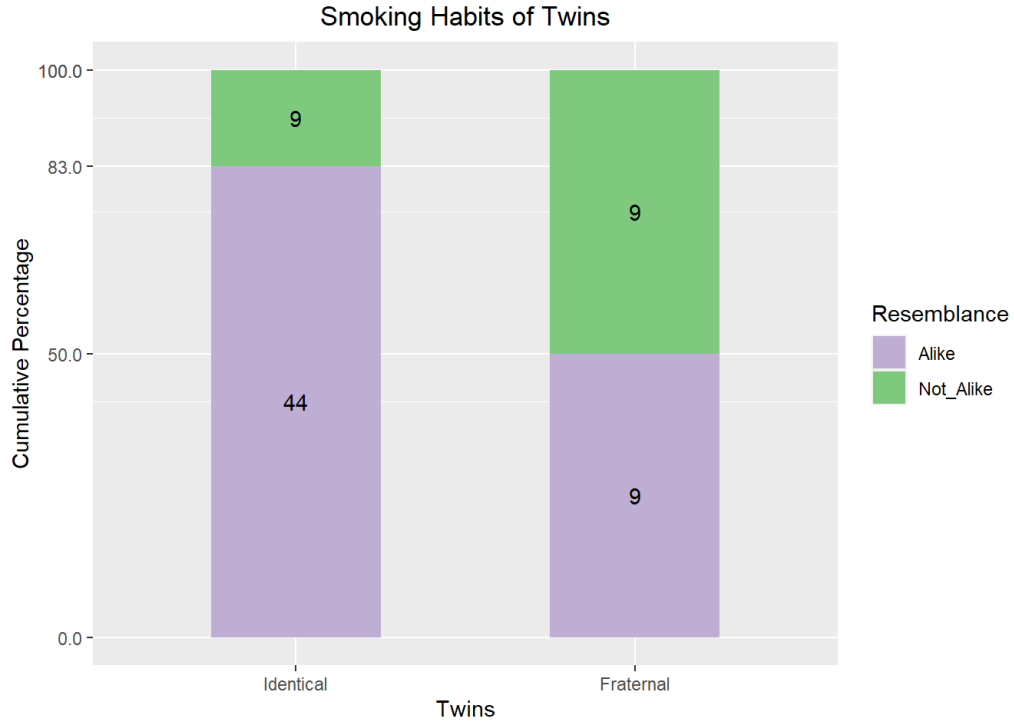
stack

```
y2_breaks <- Nature2_tbl %$%
  tapply(Counts,
         INDEX = Twins,
         FUN = cumsum) %>%
  unlist
y2_text <- Nature2_tbl %$%
  tapply(Counts,
         INDEX = Twins,
         FUN = pos) %>%
  unlist
Nature2_tbl %>%
  ggplot(data = .,
         mapping = aes(x = Twins,
                       y = Counts,
                       fill = Resemblance)) +
  geom_bar(stat = "identity",
           width = 0.5,
           position = position_stack(reverse = TRUE)) +
  geom_text(aes(y = y2_text),
            label = Nature2_tbl$Counts,
            position = "identity") +
  scale_fill_brewer(type = "qual",
                    palette = "Accent",
                    direction = -1) +
  scale_y_continuous(breaks = y2_breaks,
                     labels = y2_breaks)
```



fill

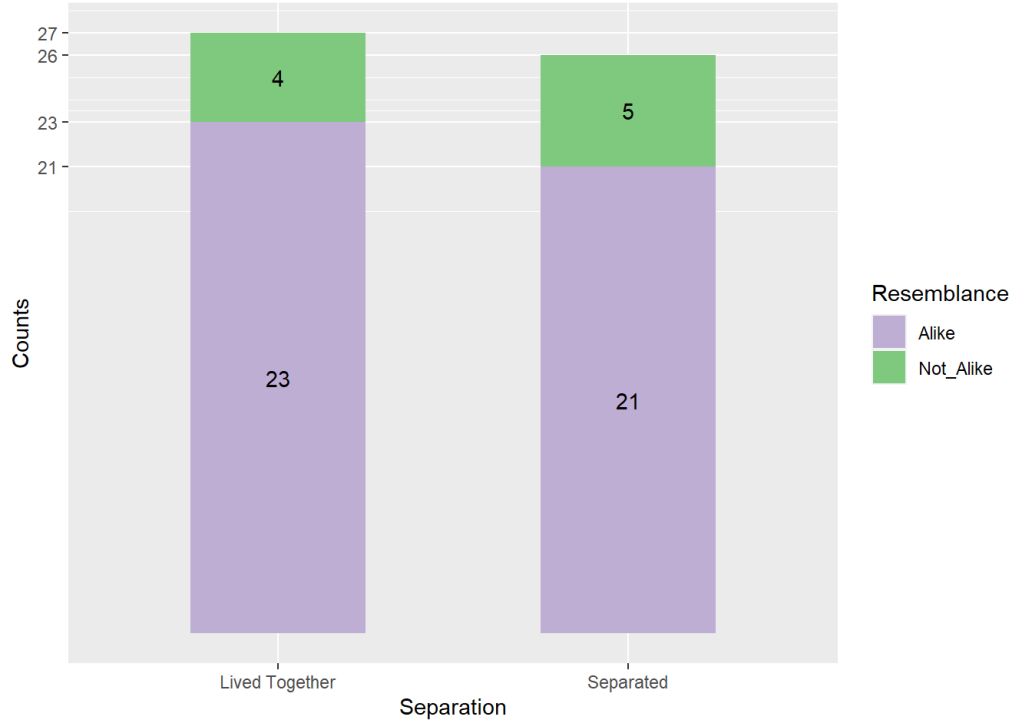
```
y2_fill <- y2_text / (Nature2 %>%
                        apply(MARGIN = 1,
                              FUN = sum) %>%
                        rep(each = 2))
Nature2_tbl %>%
  ggplot(data = .,
         mapping = aes(x = Twins,
                       y = Counts,
                       fill = Resemblance)) +
  geom_bar(stat = "identity",
           width = 0.5,
           position = position_fill(reverse = TRUE)) +
  geom_text(aes(y = y2_fill),
            label = Nature2_tbl$Counts,
            position = "identity") +
  scale_fill_brewer(type = "qual",
                    palette = "Accent",
                    direction = -1) +
  scale_y_continuous(name = "Cumulative Percentage",
                     breaks = c(0,
                                apply(t(Nature2_p),
                                      MARGIN = 2,
                                      FUN = cumsum) / 100),
                     labels = format(c(0,
                                       apply(t(Nature2_p),
                                             MARGIN = 2,
                                             FUN = cumsum)),
                                     digits = 3,
                                     nsmall = 1)) +
  labs(title = "Smoking Habits of Twins") +
  theme(plot.title = element_text(hjust = 0.5))
```



Nature 1958 version 2 : Identical Twins

stack

```
y3_breaks <- Nature3_tbl %$%
  tapply(Counts,
         INDEX = Separation,
         FUN = cumsum) %>%
  unlist
y3_text <- Nature3_tbl %$%
  tapply(Counts,
         INDEX = Separation,
         FUN = pos) %>%
  unlist
Nature3_tbl %>%
  ggplot(data = .,
         mapping = aes(x = Separation,
                       y = Counts,
                       fill = Resemblance)) +
  geom_bar(stat = "identity",
           width = 0.5,
           position = position_stack(reverse = TRUE)) +
  geom_text(aes(y = y3_text),
            label = Nature3_tbl$Counts,
            position = "identity") +
  scale_fill_brewer(type = "qual",
                    palette = "Accent",
                    direction = -1) +
  scale_y_continuous(breaks = y3_breaks,
                     labels = y3_breaks)
```



fill

```
y3_fill <- y3_text / (Nature3 %>%
                        apply(MARGIN = 1,
                              FUN = sum) %>%
                        rep(each = 2))
Nature3_tbl %>%
  ggplot(data = .,
         mapping = aes(x = Separation,
                       y = Counts,
                       fill = Resemblance)) +
  geom_bar(stat = "identity",
           width = 0.5,
           position = position_fill(reverse = TRUE)) +
  geom_text(aes(y = y3_fill),
            label = Nature3_tbl$Counts,
            position = "identity") +
  scale_fill_brewer(type = "qual",
                    palette = "Accent",
                    direction = -1) +
  scale_y_continuous(name = "Cumulative Percentage",
                     breaks = c(0,
                                apply(t(Nature3_p),
                                      MARGIN = 2,
                                      FUN = cumsum) / 100),
                     labels = format(c(0,
                                       apply(t(Nature3_p),
                                             MARGIN = 2,
                                             FUN = cumsum)),
                                     digits = 3,
                                     nsmall = 1)) +
  labs(title = "Smoking Habits of Identical Twins") +
  theme(plot.title = element_text(hjust = 0.5))
```