1. 비교(Compare)의 시각화

비교의 시각화는특정 변수의 변량에 따른 데이터 값들간의 순서를 비교하는 시각화이다. 비교의 시각화는 다른 시각화와 다른 두가지 특징이 있다.

첫번째 특징은 시각화 그래프 내에서 한번 더 통계처리가 필요하다는 특성을 가진다. 보통 비교의 시각화는 데이터에 값에 따라 시각화 한 후 그 값들을 다시 정렬함으로써 시각화를 완성하게 된다. 따라서 데이터 시각화 후에 ggplot객체 내에서 정렬과 같은 통계 처리가 한번 더 일어나게 된다.

두번째 특징은 그 관심의 대상이 데이터 자체의 값 보다는 비교되는 대상 내에서의 상대적 위치에 더 쏠려있다는 점이다. 비교되는 대상 중에 가장 값이 크거나 작은 변량이 무엇인지에 관심이 있는 시각화이다. 비교되는 대상들의 상대적 위치를 명확하게 구분되어야 하기 때문에 비교되는 대상들이 명확하게 구분되어야 하고 이 대상들을 명학하게 구분하기 위해 비교를 위한 변수는 이산형 변수를 사용하는 것이 일반적이다.

비교의 시각화는 막대 그래프, 순위 막대 그래프, 롤리팝 그래프, 도트 그래프 등이 있다.

# 막대 그래프

막대 그래프는 이산형 변수로 구분되는 연속형 수치형 변수를 시각화하기 위해 사용한다. 각각의 이산형 변수에 하나씩의 막대가 표시되고 막대의 길이를 사용하여 데이터 값을 표현한다. 막대 그래프는 데이터를 쉽게 읽을 수 있을 뿐아니라 데이터간의 비교가 상대적으로 쉽다. 따라서 데이터간의 비교에 매우 효과적인 시각화 방법이다.

막대그래프의 가장 일반적인 형태는 세로 막대가 가로축으로 배열되어 있는 형태이다. 과거의 한정된 지면에 표현하던 때나 데이터가 한정적이던 때는 비교할 이산형 변수의 변량이 많지 았았지만 지금같이 데이터가 많이지고 웹사이트와 같이 위아래로 스크롤이 가능한 화면에서는 오히려 가로형 막대를 세로축에 표현하는 경우도 많다.

막대 그래프를 실습해보기 전에 실습에 필요한 데이터를 다음과 같이 데이터를 준비하도록 하겠다. 다음의 데이터는 df\_취업통계 데이터를 과정구분, 대계열, 중계열 별로 그루핑하여 각각의 열의 합계값을 만들고 각각의 중계열별 취업률을 산출한다.

df\_취업통계\_계열별 <- df\_취업통계 |>   
 group\_by(과정구분, 대계열, 중계열) |>   
 summarise(졸업자 = sum(졸업자\_계),   
 취업자 = sum(취업자\_합계\_계),   
 교외취업자 = sum(취업자\_교외취업자\_계),   
 교내취업자 = sum(취업자\_교내취업자\_계),   
 해외취업자 = sum(취업자\_해외취업자\_계),   
 농림어업종사자 = sum(취업자\_농림어업종사자\_계),   
 개인창작활동종사자 = sum(취업자\_개인창작활동종사자\_계),   
 일인창사업자 = sum(`취업자\_1인창(사)업자\_계`),   
 프리랜서 = sum(취업자\_프리랜서\_계),   
 진학자 = sum(진학자\_계),   
 입대자 = sum(입대자),  
 취업불가능자 = sum(취업불가능자\_계),   
 외국인유학생 = sum(외국인유학생\_계),   
 제외인정자 = sum(제외인정자\_계),   
 기타 = sum(기타\_계),   
 미상 = sum(미상\_계),   
 ## 백분률인 취업률은 그 자체로 합계나 평균을 낼 수 없으니 각 그룹별로 재계산  
 취업률 = 취업자 / (졸업자 - (진학자+입대자+취업불가능자+외국인유학생+제외인정자))) |>  
 ## 계열의 표시 순서를 설정하기 위해 레벨을 재조정  
 mutate(대계열 = fct\_relevel(대계열, '인문계열', '사회계열', '교육계열', '자연계열', '공학계열', '의약계열', '예체능계열'))  
  
df\_취업통계\_계열별 |> head()

## # A tibble: 6 x 20  
## # Groups: 과정구분, 대계열 [1]  
## 과정구분 대계열 중계열 졸업자 취업자 교외취업자 교내취업자 해외취업자  
## <chr> <fct> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 대학과정 공학계열 건축 7697 4943 4634 64 6  
## 2 대학과정 공학계열 교통ㆍ운송 4206 2443 2311 37 6  
## 3 대학과정 공학계열 기계ㆍ금속 11422 6206 5759 142 46  
## 4 대학과정 공학계열 기타 3962 2328 2141 65 17  
## 5 대학과정 공학계열 산업 3594 2249 2105 46 8  
## 6 대학과정 공학계열 소재ㆍ재료 6143 2948 2682 85 14  
## # ... with 12 more variables: 농림어업종사자 <dbl>, 개인창작활동종사자 <dbl>,  
## # 일인창사업자 <dbl>, 프리랜서 <dbl>, 진학자 <dbl>, 입대자 <dbl>,  
## # 취업불가능자 <dbl>, 외국인유학생 <dbl>, 제외인정자 <dbl>, 기타 <dbl>,  
## # 미상 <dbl>, 취업률 <dbl>

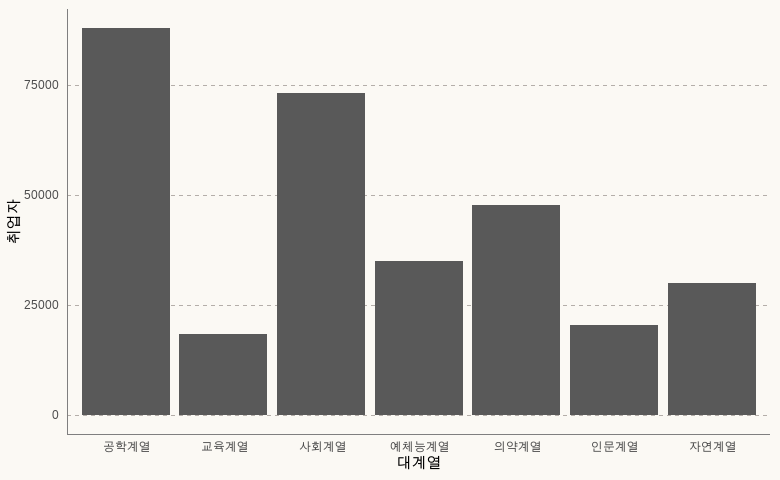
## stack, dodge, fill 타입의 막대 그래프

우리가 가장 흔하게 생각하는 막대 그래프는 각 변량별로 하나의 막대로 데이터를 표현하는 막대그래프이다. 하지만 그 막대를 구성하는 데이터들은 다시 세분화될 수 있다. 이렇게 하나의 막대로 표현된 세부 데이터를 다시 표현해야할 경우를 대비하여 ggplot2는 세가지 형태의 막대 그래프를 지원한다.

* stack : 막대의 세부 그룹을 각각 쌓아 올려 하나의 단일 막대를 생성
* dodge : 막대의 세부 그룹을 각각 옆으로 붙여 세부 그룹 수 만큼의 막대 생성
* fill : 막대의 세부 그룹을 각각 쌓아 올려 하나의 단일 막대를 만들되 전체 막대크 크기를 모두 동일하게 맞춰 세부 그룹들의 비율을 표현

다음의 기본 막대 그래프를에서 파생되는 세가지 형태의 막대 그래프의 형태를 비교해 보도록 한다.

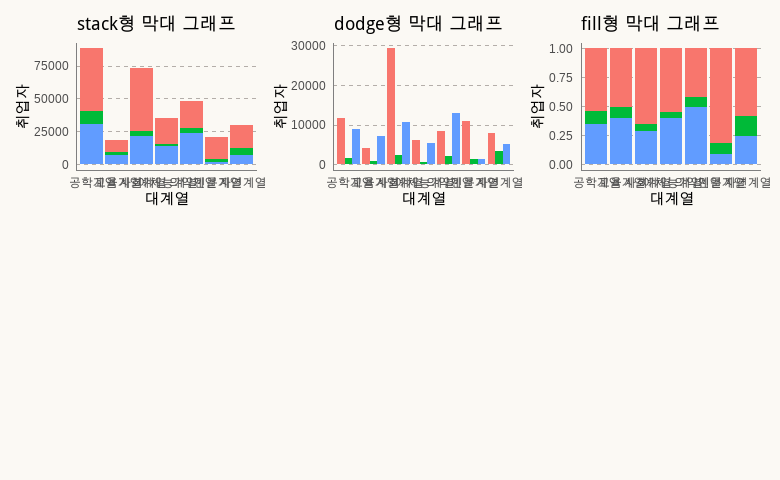
p\_bar\_type <- df\_취업통계\_계열별 |>  
 ## ggplot 객체 생성  
 ggplot()  
  
p\_bar\_type +  
 ## X축을 대계열, Y축을 취업자열로 매핑한 geom\_col 레이어 생성  
 ## geom\_bar를 사용하려면 geom\_bar(aes(x = 대계열, y = 취업자), stat = 'identity')  
 geom\_col(aes(x = 대계열, y = 취업자))



실행결과 8- . 기본형 막대 그래프

geom\_col()과 geom\_bar()의 막대 타입은 막대의 위치를 뜻하는 position 매개변수로 변경할 수 있다. 기본값은 ’stack’으로 전체 막대 길이는 전체 데이터 값이며 각각의 세부 그룹들이 위에 쌓아올려진 형태의 막대 그래프이다.

p\_bar\_type +  
 geom\_col(aes(x = 대계열, y = 취업자, fill = 과정구분), position = 'stack', show.legend = FALSE) +   
 labs(title = 'stack형 막대 그래프')  
  
p\_bar\_type +  
 geom\_col(aes(x = 대계열, y = 취업자, fill = 과정구분), position = 'dodge', show.legend = FALSE) +   
 labs(title = 'dodge형 막대 그래프')  
  
p\_bar\_type +  
 geom\_col(aes(x = 대계열, y = 취업자, fill = 과정구분), position = 'fill', show.legend = FALSE) +   
 labs(title = 'fill형 막대 그래프')



실행결과 8- . 막대그래프의 position 비교

## 데이터 값 넣기

앞서 생성한 막대 그래프들에는 각각의 데이터 값이 표현되지 않아 데이터들을 비교하기가 어렵다. 각각의 막대 그래프의 타입에 따라 데이터 값을 표현하는 방법도 다 다르다.

### stack 막대 그래프의 데이터 값 넣기

stack 막대 그래프에 데이터 값을 넣으려면 가장 효과적인 것이 각각의 세부 막대 중간에 해당 데이터 값을 표현하고 전체 데이터 값을 맨위에 표현해 주는 방법이다. 이를 위해서는 두번의 작업이 필요하다.

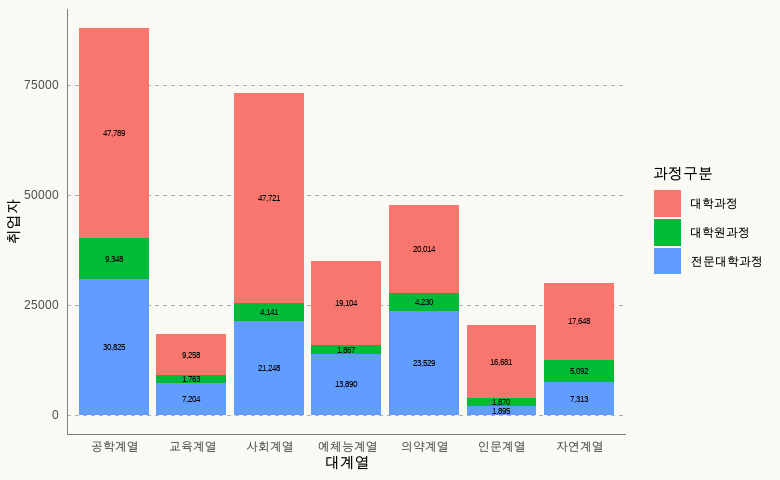
우선 막대를 구성하는 그룹별 합계값을 구해야 한다. 앞의 막대들은 대계열 막대를 과정구분으로 나누었기 때문에 대계열별 과정구분별 데이터 값(취업자수)가 필요하다. 이후 막대 그래프를 그리고 이 안에 geom\_text()로 데이터 값을 표현한다.

각 그룹의 중간에 데이터를 표현하기 위해서는 position\_stack()을 사용한다. 이 함수는 각각의 그룹 위에 기하요소를 올려서 표현해 주는 함수이다.

position\_stack(vjust = 1, reverse = FALSE)  
 - vjust : 수직 정렬 위치 설정  
 - reverse : 순서의 역순을 설정하는 논리값

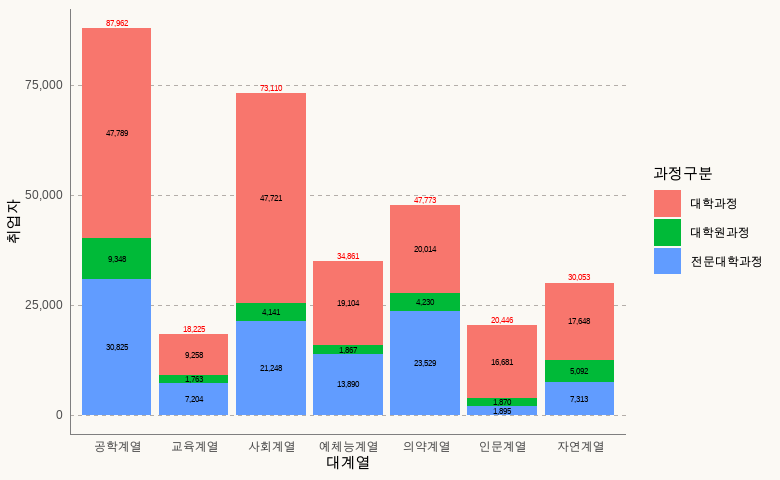
이 함수는 geom\_col(), geom\_bar()이외의 기하요소에서도 사용할 수 있기 때문에 데이터 값을 표현하는데 사용하는 geom\_text()와 geom\_label()에 사용하면 해당 그룹에 데이터 값을 표현할 수 있다. 또 각 막대의 중간에 표현해야 하니 수평정렬 매개변수인 vjust를 0.5로 설정하면 중간에 위치하는 데이터 값을 표현할 수 있다.

## 데이터 값을 표현하기 위해 대계열별 과정구분으로 그루핑하고  
p\_bar\_data\_stack <- df\_취업통계\_계열별 |> group\_by(대계열, 과정구분) |>  
 ## 합계값을 산출  
 summarise(취업자 = sum(취업자)) |>  
 ggplot() +  
 ## X축이 대계열, Y축이 취업자, fill이 과정구분으로 매핑되고 position이 stack으로 설정된 geom\_col 레이어 생성  
 geom\_col(aes(x = 대계열, y = 취업자, fill = 과정구분), position = 'stack') +  
 ## X축이 대계열, Y축이 취업자, lable이 comma형 취업자, fill은 과정구분으로 매핑되고 position은 position\_stack()으로 vjust가 0.5인 stack형으로 설정한 geom\_text 레이어 추가  
 geom\_text(aes(x = 대계열, y = 취업자, label = scales::comma(취업자, accuracy = 1), fill =과정구분), position = position\_stack(vjust = 0.5), size = 2)  
  
p\_bar\_data\_stack



다음으로 전체 데이터 값을 표현해보자. 앞서 세부 그룹의 합계값을 구했기 때문에 이들의 합계값을 산출해주면 된다. 이를 위해 fun을 sum으로 설정한 stat\_summary()를 사용한다.

p\_bar\_data\_stack1 <- p\_bar\_data\_stack +  
 ## X축이 대계열, Y축이 취업자, lable이 comma형 ..y..(fun으로 설정한 함수의 결과값)으로 매핑되고 적용할 함수(fun)은 sum, 표현할 기하요소는 text, vjust는 0.5, size는 2인 stat\_summary 레이어 생성   
 stat\_summary(aes(x = 대계열, y = 취업자, label = scales::comma(..y.., accuracy = 1)), fun = 'sum', geom = 'text', color = 'red', vjust = -0.5, size = 2) +   
 scale\_y\_continuous(labels = scales::comma)  
  
p\_bar\_data\_stack1

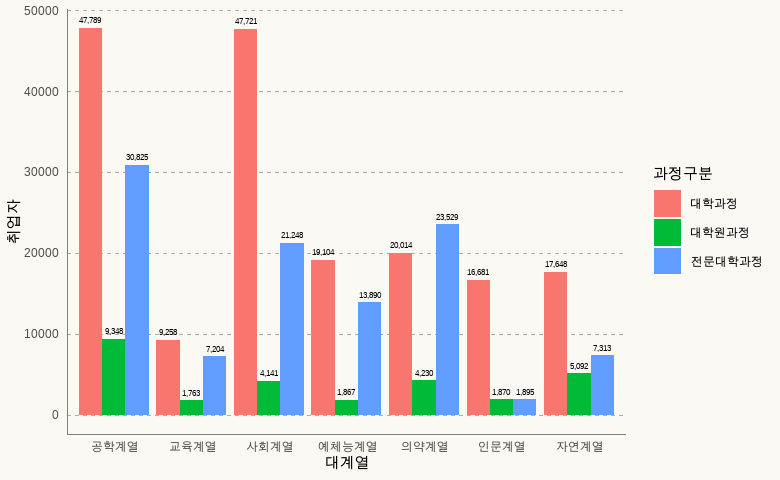


### dodge 막대 그래프의 데이터 값 넣기

dodge 막대 그래프는 stack 막대 그래프보다는 데이터 값을 넣기가 쉽다. 앞서 stack 막대 그래프에 geom\_text()의 위치를 설정하는데에는 postion\_stack()을 사용했다면 dodge 막대 그래프에는 position\_dodge()를 사용한다.

position\_dodge(width = NULL, preserve = c("total", "single"))  
 - width : 옆으로 배열되는 너비 설정, 기본적인 막대의 너비는 0.9임  
 - preserve : 막대 너비를 전체에 맞출지 각각의 개벌 막대에 맞출지 설정

## 데이터 값을 표현하기 위해 대계열별 과정구분으로 그루핑하고  
p\_bar\_data\_dodge <- df\_취업통계\_계열별 |> group\_by(대계열, 과정구분) |>  
 ## 합계값을 산출  
 summarise(취업자 = sum(취업자)) |>  
 ggplot() +  
 ## X축이 대계열, Y축이 취업자, fill이 과정구분으로 매핑되고 position이 stack으로 설정된 geom\_col 레이어 생성  
 geom\_col(aes(x = 대계열, y = 취업자, fill = 과정구분), position = 'dodge') +  
 ## X축이 대계열, Y축이 취업자, lable이 comma형 취업자, fill은 과정구분으로 매핑되고 position은 position\_stack()으로 vjust가 0.5인 stack형으로 설정한 geom\_text 레이어 추가  
 geom\_text(aes(x = 대계열, y = 취업자, label = scales::comma(취업자, accuracy = 1), fill =과정구분), position = position\_dodge(width = 0.9), size = 2, vjust = -1)  
  
p\_bar\_data\_dodge



## 축 라벨에 이미지 넣기

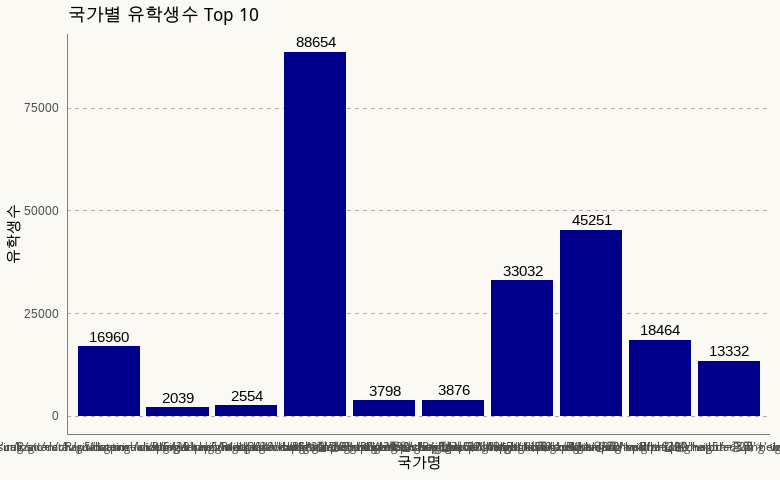
library(readxl)  
df\_nation <- read\_xlsx('파일경로/연도별 유학국가별 유학생수.xlsx', sheet = 'Sheet0', skip = 2, col\_types = c('numeric', 'text', rep('numeric', 25)), col\_names = TRUE)  
  
df\_nation <- df\_nation |>  
 filter(!is.na(학년도), 학제 == '소계') |>  
 select(!contains(c('계', '학제', '기타', '미확인', '그외동남아', '남미'))) |>  
 gather('국가명', '유학생수', -'학년도')  
  
df\_nation\_top10 <- df\_nation |>  
 group\_by(국가명) |>  
 summarise(sum = sum(유학생수)) |>  
 arrange(desc(sum)) |>  
 top\_n(10)  
  
df\_nation\_top10

## # A tibble: 10 x 2  
## 국가명 sum  
## <chr> <dbl>  
## 1 미국 88654  
## 2 캐나다 45251  
## 3 중국 33032  
## 4 필리핀 18464  
## 5 뉴질랜드 16960  
## 6 호주 13332  
## 7 일본 3876  
## 8 영국 3798  
## 9 말레이시아 2554  
## 10 독일 2039

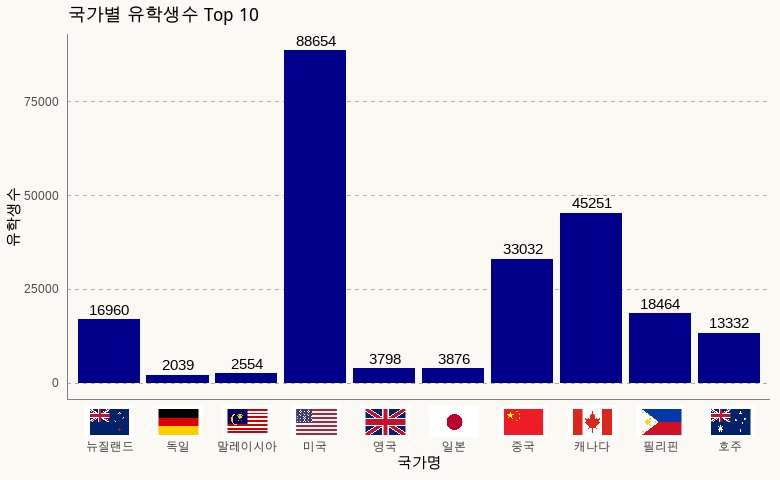
p\_nation\_top10 <- df\_nation\_top10 |>  
 ggplot(aes(x = 국가명, y = sum)) +  
 geom\_col(fill = 'dark blue') +  
 geom\_text(aes(x = 국가명, y = sum, label = sum), vjust = -0.5) +   
 labs(title = '국가별 유학생수 Top 10', y = '유학생수')

flag\_usa <- '아이콘 이미지 저장 폴더 경로/usa.png'  
flag\_canada <- '아이콘 이미지 저장 폴더 경로/can.png'  
flag\_china <- '아이콘 이미지 저장 폴더 경로/chi.png'  
flag\_phi <- '아이콘 이미지 저장 폴더 경로/phi.png'  
flag\_nz <- '아이콘 이미지 저장 폴더 경로/nz.png'  
flag\_aus <- '아이콘 이미지 저장 폴더 경로/aus.png'  
flag\_jap <- '아이콘 이미지 저장 폴더 경로/jap.png'  
flag\_eng <- '아이콘 이미지 저장 폴더 경로/eng.png'  
flag\_mal <- '아이콘 이미지 저장 폴더 경로/mal.png'  
flag\_ger <- '아이콘 이미지 저장 폴더 경로/ger.png'  
  
flags <- data.frame(nations = c('미국', '캐나다', '중국', '뉴질랜드', '필리핀', '호주', '영국', '일본', '말레이시아', '싱가폴'), flag\_path = c(flag\_usa, flag\_canada, flag\_china, flag\_nz, flag\_phi, flag\_aus, flag\_eng, flag\_jap, flag\_mal, flag\_sing))  
  
labels <- setNames(  
 paste0("<img src='", flags$flag\_path, "' width='30' height = '20'> <br> ", flags$nations), flags$nations)

p\_nation\_top10\_1 <- p\_nation\_top10 +  
 scale\_x\_discrete(labels = labels)  
  
p\_nation\_top10\_1

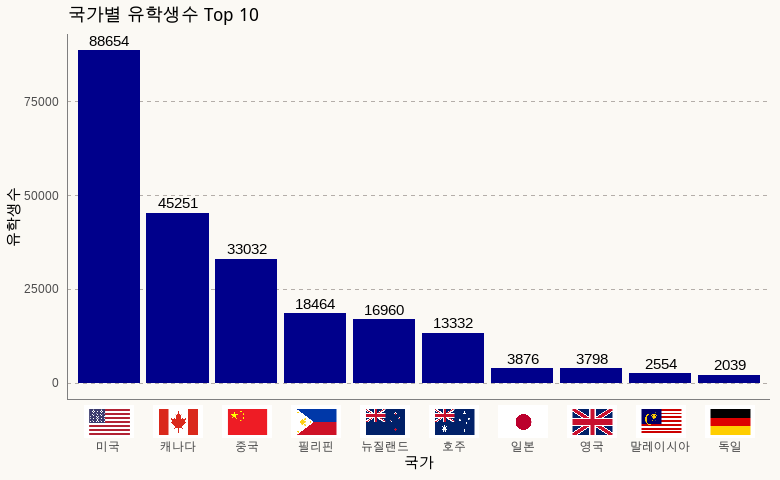


if(!require(ggtext)) {  
 install.packages('ggtext')  
 library(ggtext)  
}  
  
p\_nation\_top10\_1 +  
 theme(axis.text.x = ggtext::element\_markdown())



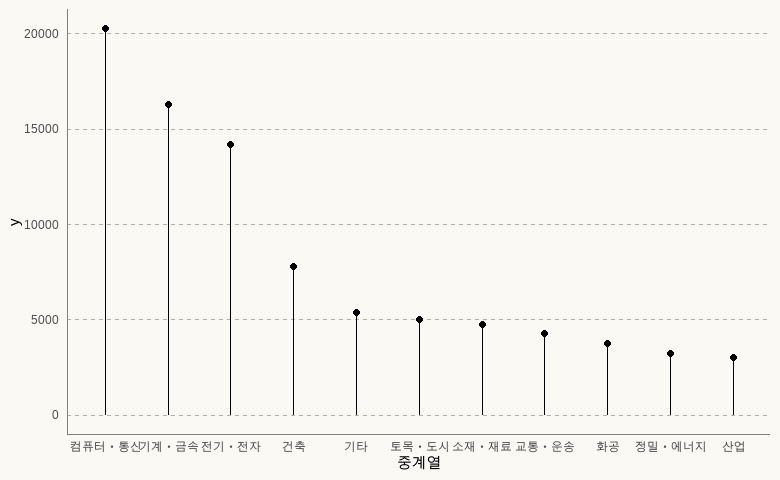
# 순위 막대 그래프

df\_nation\_top10$국가명 <- fct\_reorder(df\_nation\_top10$국가명, desc(df\_nation\_top10$sum))  
  
df\_nation\_top10 |>  
 ggplot(aes(x = 국가명, y = sum)) +  
 geom\_col(fill = 'dark blue') +  
 geom\_text(aes(x = 국가명, y = sum, label = sum), vjust = -0.5) +   
 labs(title = '국가별 유학생수 Top 10', x = '국가', y = '유학생수') +  
 scale\_x\_discrete(labels = labels) +  
 theme(axis.text.x = ggtext::element\_markdown())

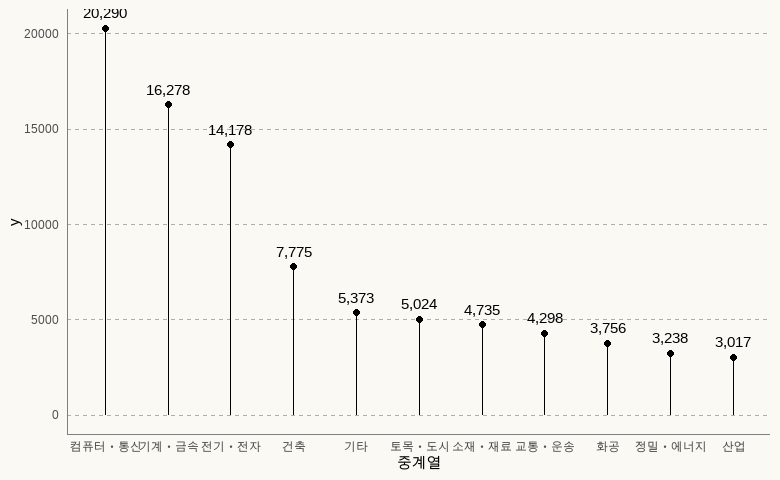


# 롤리팝 그래프

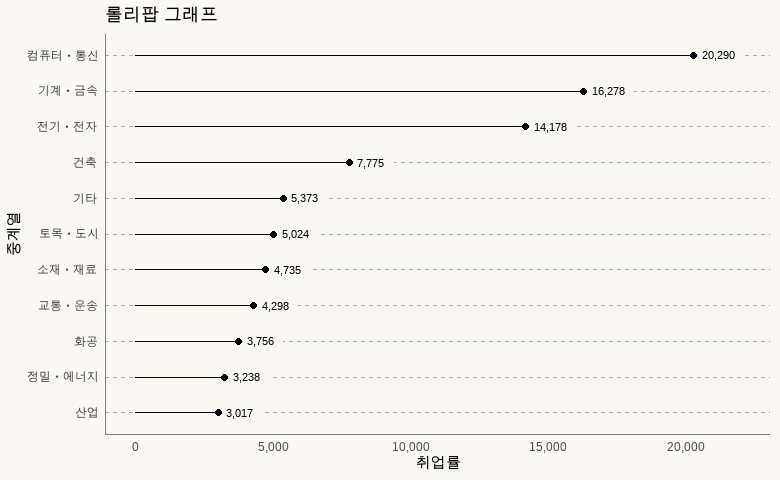
df\_lolipop <- df\_취업통계\_계열별 |> filter(대계열 == '공학계열') |> group\_by(중계열) |>  
 summarise(취업자 = sum(취업자)) |> arrange(desc(취업자))  
  
df\_lolipop$중계열 <- fct\_reorder(df\_lolipop$중계열, desc(df\_lolipop$취업자))  
  
p\_lolipop <- df\_lolipop|>  
 ggplot()  
  
p\_lolipop1 <- p\_lolipop +  
 geom\_segment(aes(x = 중계열, xend = 중계열, y = 0, yend = 취업자)) +  
 geom\_point(aes(x = 중계열, y = 취업자, group = 중계열))  
  
p\_lolipop1



p\_lolipop2 <- p\_lolipop1 +   
 geom\_text(aes(x = 중계열, y = 취업자, label = scales::comma(취업자)), vjust = -1)  
  
p\_lolipop2



p\_lolipop1 +   
 geom\_label(aes(x = 중계열, y = 취업자, label = scales::comma(취업자)), hjust = -0.1, label.size = NA, size = 3, fill = '#fbf9f4') +  
 scale\_x\_discrete(limits = rev) +  
 scale\_y\_continuous(labels = scales::comma, limits = c(0, 22000)) +  
 labs(title = '롤리팝 그래프', x = '중계열', y = '취업률') +   
 coord\_flip()



# 도트 그래프

p\_dot <- df\_lolipop|>  
 ggplot()  
  
p\_dot1 <- p\_dot +  
 geom\_segment(aes(x = 중계열, xend = 중계열, y = min(취업자), yend = max(취업자)), linetype = 'dashed', size = 0.1) +  
 geom\_point(aes(x = 중계열, y = 취업자, group = 중계열), size = 3, color = 'red') +   
 scale\_x\_discrete(limits = rev) +  
 coord\_flip() +   
 labs(title = '도트 그래프', x = '중계열', y = '취업자수')  
  
p\_dot1

