1. 분포의 시각화

# 히스토그램

## bin, binwidth

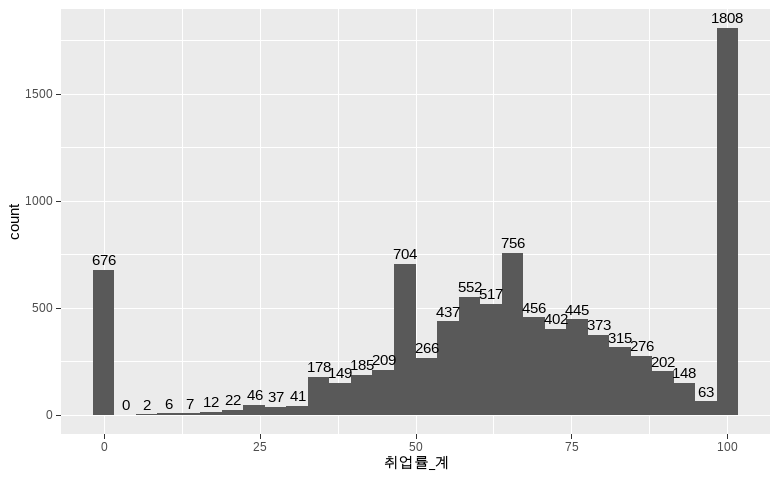
p\_histo <- df\_취업통계 |>  
 ggplot(aes(x = 취업률\_계))  
  
p\_histo +  
 geom\_histogram()



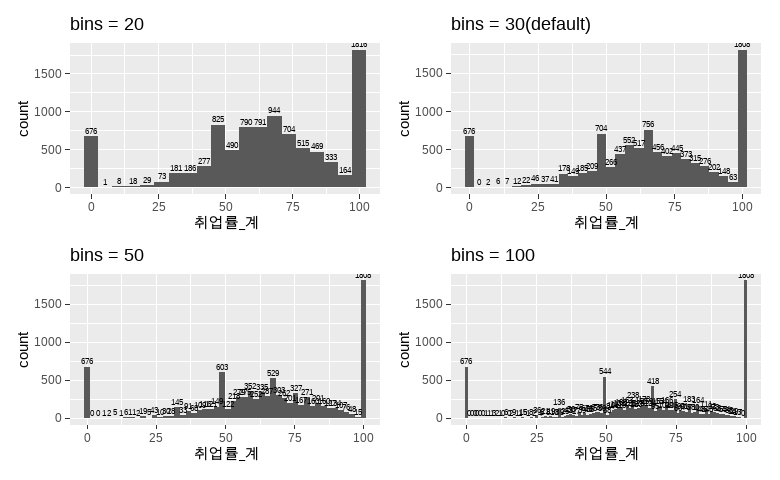
p\_histo +   
 geom\_histogram(aes(y = ..count..))



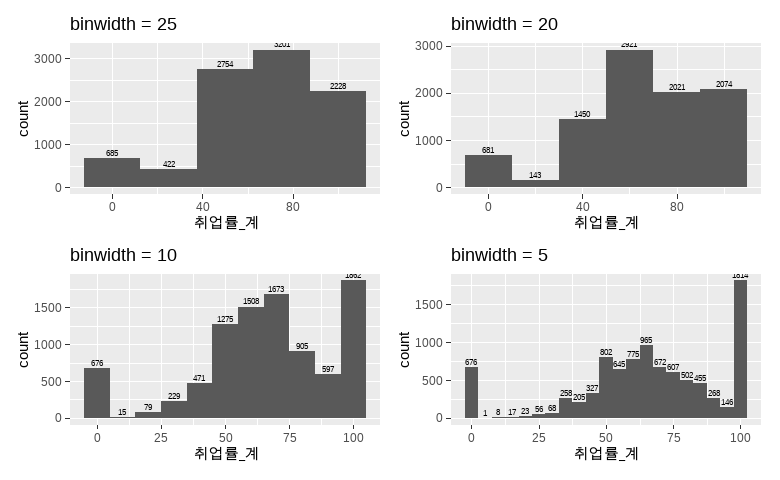
p\_histo +  
 geom\_histogram(aes(y = ..count..)) +  
 stat\_bin(aes(x = 취업률\_계,y=..count.., label=..count..), geom="text", vjust=-.5)



p\_histo +  
 geom\_histogram(aes(y = ..count..), bins = 20) +  
 stat\_bin(aes(y=..count.., label=..count..), bins = 20, geom="text", vjust=-.5) +  
 labs(title = 'bins = 20')  
  
p\_histo +  
 geom\_histogram(aes(y = ..count..), bins = 30) +  
 stat\_bin(aes(y=..count.., label=..count..), bins =30, geom="text", vjust=-.5) +  
 labs(title = 'bins = 30(default)')  
  
p\_histo +  
 geom\_histogram(aes(y = ..count..), bins = 50) +  
 stat\_bin(aes(y=..count.., label=..count..), bins = 50, geom="text", vjust=-.5) +  
 labs(title = 'bins = 50')  
  
p\_histo +  
 geom\_histogram(aes(y = ..count..), bins = 100) +  
 stat\_bin(aes(y=..count.., label=..count..), bins = 100, geom="text", vjust=-.5) +  
 labs(title = 'bins = 100')

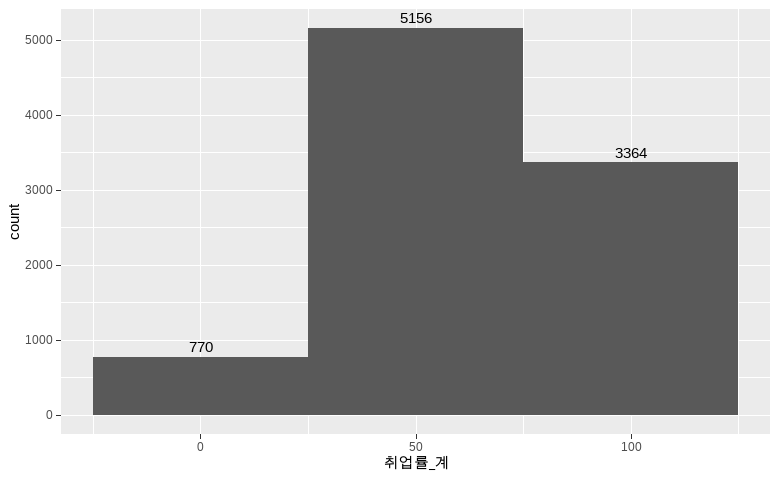


p\_histo +  
 geom\_histogram(aes(y = ..count..), binwidth = 25) +  
 stat\_bin(aes(y=..count.., label=..count..), binwidth = 25, geom="text", vjust=-.5) +  
 labs(title = 'binwidth = 25')  
  
p\_histo +  
 geom\_histogram(aes(y = ..count..), binwidth = 20) +  
 stat\_bin(aes(y=..count.., label=..count..), binwidth = 20, geom="text", vjust=-.5) +  
 labs(title = 'binwidth = 20')  
  
p\_histo +  
 geom\_histogram(aes(y = ..count..), binwidth = 10) +  
 stat\_bin(aes(y=..count.., label=..count..), binwidth = 10, geom="text", vjust=-.5) +  
 labs(title = 'binwidth = 10')  
  
p\_histo +  
 geom\_histogram(aes(y = ..count..), binwidth = 5) +  
 stat\_bin(aes(y=..count.., label=..count..), binwidth = 5, geom="text", vjust=-.5) +  
 labs(title = 'binwidth = 5')

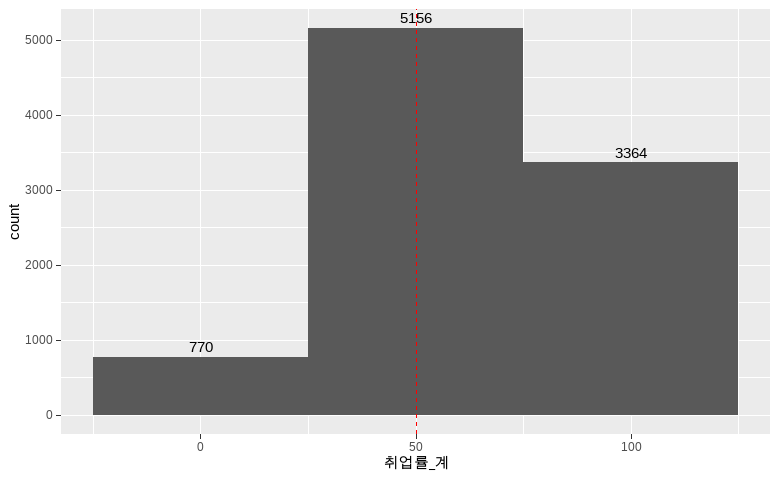


## bin 분할 원리

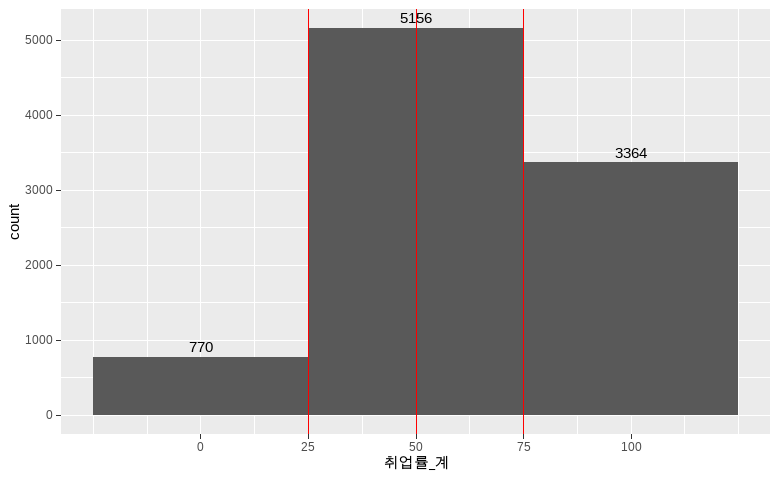
p\_histo2\_1 <- p\_histo +  
 geom\_histogram(aes(y = ..count..), bins = 3) +  
 stat\_bin(aes(y=..count.., label=..count..), bins = 3, geom="text", vjust=-.5)  
  
p\_histo2\_1



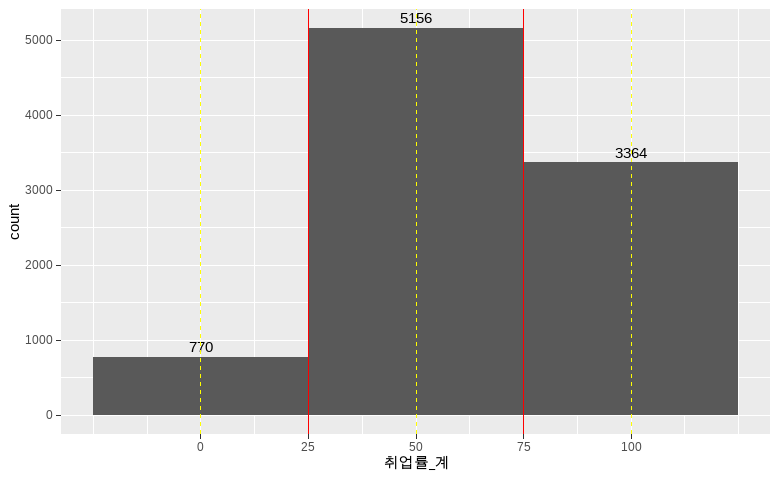
range.x <- max(df\_취업통계$취업률\_계) - min(df\_취업통계$취업률\_계)  
  
p\_histo2\_1 +   
 geom\_vline(aes(xintercept = range.x/2), color = 'red', linetype = 2)



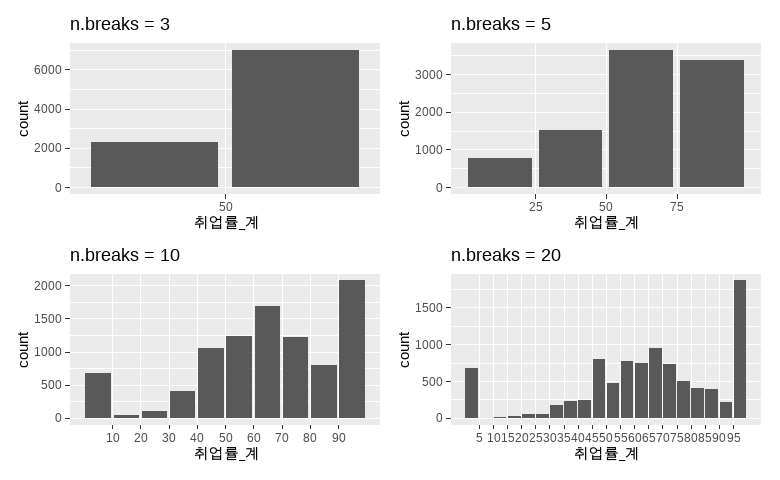
p\_histo2\_1 +   
 geom\_vline(aes(xintercept = range.x/2), color = 'red', linetype = 2) +   
 geom\_vline(xintercept = range.x/2, color = 'red') + ## X축의 1/2 위치  
 geom\_vline(xintercept = range.x/2/2, color = 'red') + ## X축의 1/4 위치   
 geom\_vline(xintercept = range.x/2/2\*3, color = 'red') + ## X축의 3/4 위치  
 scale\_x\_continuous(breaks = c(min(df\_취업통계$취업률\_계),   
 range.x/2/2,   
 range.x/2,  
 range.x/2/2\*3,  
 max(df\_취업통계$취업률\_계)  
 )  
 )



p\_histo2\_1 +   
 geom\_vline(xintercept = 50, color = 'yellow', linetype = 2) +   
 geom\_vline(xintercept = 0, color = 'yellow', linetype = 2) +   
 geom\_vline(xintercept = 100, color = 'yellow', linetype = 2) +   
 geom\_vline(xintercept = 25, color = 'red') + ## X축의 1/4 위치   
 geom\_vline(xintercept = 75, color = 'red') + ## X축의 3/4 위치  
 scale\_x\_continuous(breaks = c(min(df\_취업통계$취업률\_계),   
 range.x/2/2,   
 range.x/2,  
 range.x/2/2\*3,  
 max(df\_취업통계$취업률\_계)  
 )  
 )

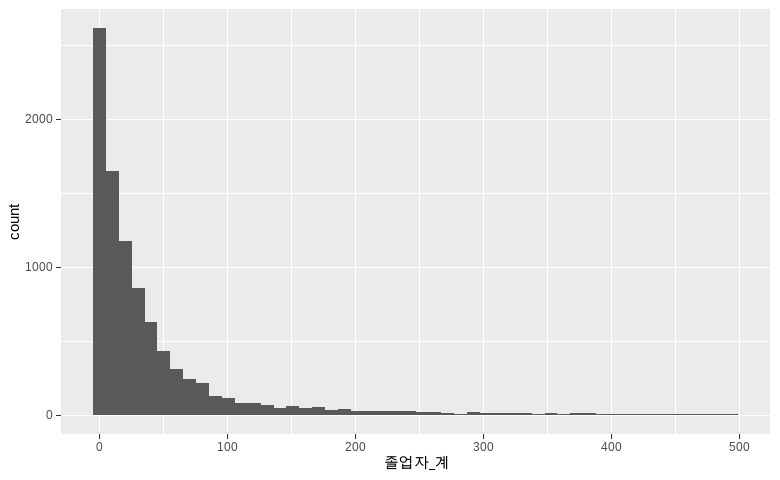


p\_histo +  
 geom\_bar() +  
 scale\_x\_binned(n.breaks = 3, right = T) +  
 labs(title = 'n.breaks = 3')  
  
p\_histo +  
 geom\_bar() +  
 scale\_x\_binned(n.breaks = 5, right = T) +  
 labs(title = 'n.breaks = 5')  
  
p\_histo +  
 geom\_bar() +  
 scale\_x\_binned(n.breaks = 10, right = T) +  
 labs(title = 'n.breaks = 10')  
  
p\_histo +  
 geom\_bar() +  
 scale\_x\_binned(n.breaks = 20, right = T) +  
 labs(title = 'n.breaks = 20')

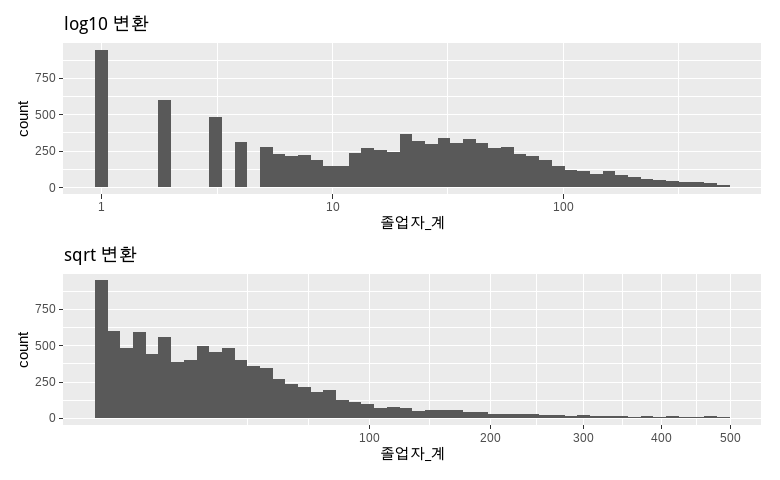


## 히스토그램의 축 변환

p\_histo3 <- df\_취업통계 |> filter(졸업자\_계 < 500) |>  
 ggplot()  
  
p\_histo3\_1 <- p\_histo3 +  
 geom\_histogram(aes(x = 졸업자\_계), bins = 50)  
  
p\_histo3\_1

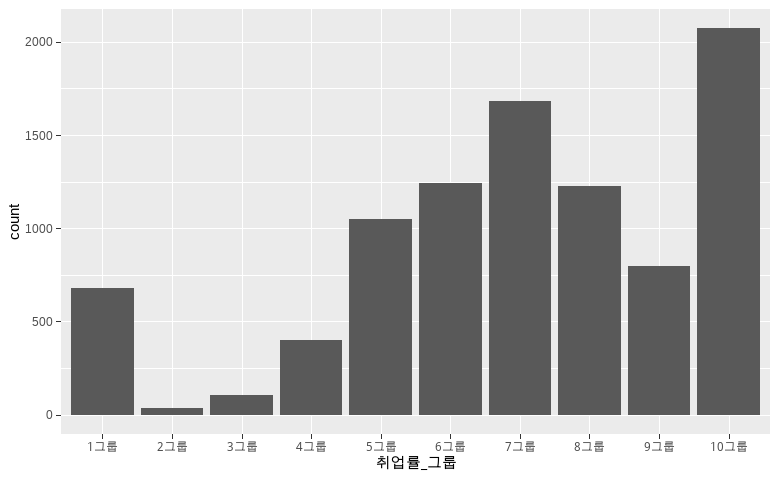


p\_histo3\_1 +  
 scale\_x\_log10() +  
 labs(title = 'log10 변환')  
  
p\_histo3\_1 +  
 scale\_x\_sqrt() +  
 labs(title = 'sqrt 변환')

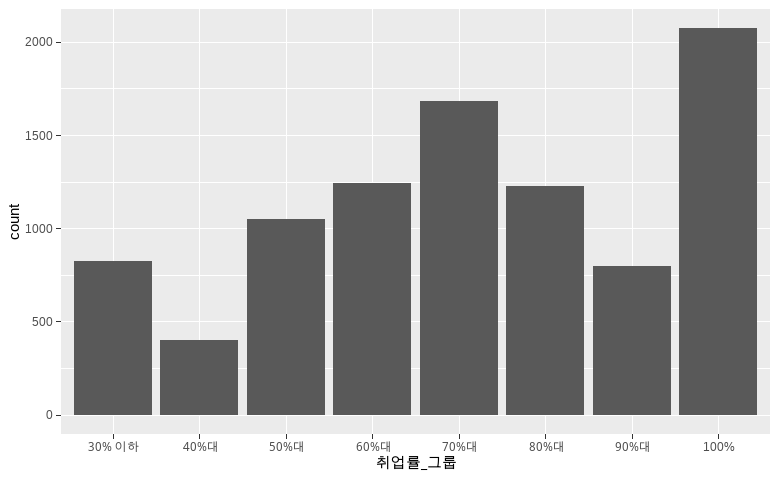


## 사용자 정의 히스토그램

df\_취업통계$취업률\_그룹 <- cut(df\_취업통계$취업률\_계, breaks = 10, label = paste0(1:10, '그룹'))  
  
df\_취업통계 |>  
 ggplot() +  
 geom\_bar(aes(x = 취업률\_그룹, y = ..count..))



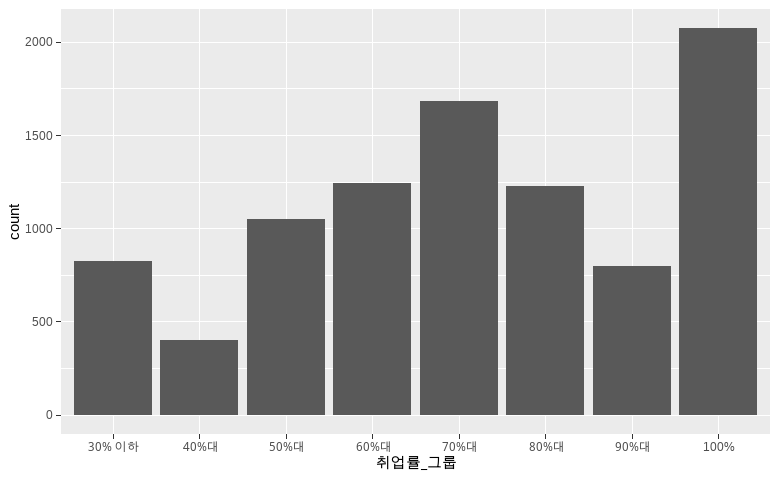
df\_취업통계$취업률\_그룹 <- cut(df\_취업통계$취업률\_계, breaks = c(-Inf, 30, 40, 50, 60, 70, 80, 90, 100), label = c('30% 이하', '40%대', '50%대', '60%대', '70%대', '80%대', '90%대', '100%'))  
  
df\_취업통계 |>  
 ggplot() +  
 geom\_bar(aes(x = 취업률\_그룹, y = ..count..))



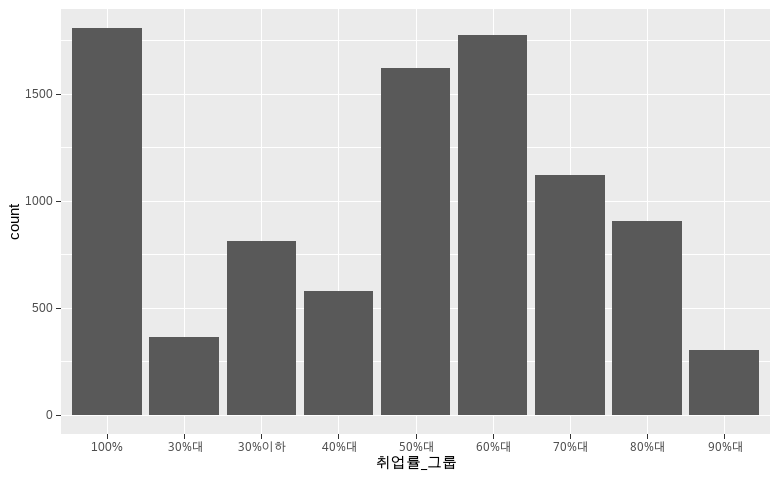
df\_취업통계 |>  
 mutate(취업률\_그룹 = case\_when(  
 취업률\_계 < 30 ~ '30%이하',   
 취업률\_계 >= 30 & 취업률\_계 < 40 ~ '30%대',   
 취업률\_계 >= 40 & 취업률\_계 < 50 ~ '40%대',   
 취업률\_계 >= 50 & 취업률\_계 < 60 ~ '50%대',   
 취업률\_계 >= 60 & 취업률\_계 < 70 ~ '60%대',   
 취업률\_계 >= 70 & 취업률\_계 < 80 ~ '70%대',   
 취업률\_계 >= 80 & 취업률\_계 < 90 ~ '80%대',   
 취업률\_계 >= 90 & 취업률\_계 < 100 ~ '90%대',  
 TRUE ~ '100%'  
 ))

## # A tibble: 9,290 x 37  
## 조사기준일 학제 과정구분 대계열 중계열 소계열 학과코드 학과명 학위구분  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 2020.12.31 전문대~ 전문대학~ 의약계~ 간호 간호 C060101~ 간호학과 <NA>   
## 2 2020.12.31 전문대~ 전문대학~ 의약계~ 간호 간호 C060101~ 간호학~ <NA>   
## 3 2020.12.31 전문대~ 전문대학~ 인문계~ 언어~ 일본어 C010101~ 관광일~ <NA>   
## 4 2020.12.31 전문대~ 전문대학~ 인문계~ 언어~ 일본어 C010101~ 관광일~ <NA>   
## 5 2020.12.31 전문대~ 전문대학~ 인문계~ 언어~ 일본어 C010101~ 관광일~ <NA>   
## 6 2020.12.31 전문대~ 전문대학~ 인문계~ 언어~ 일본어 C010101~ 일본어과 <NA>   
## 7 2020.12.31 전문대~ 전문대학~ 인문계~ 언어~ 일본어 C010101~ 비즈니~ <NA>   
## 8 2020.12.31 전문대~ 전문대학~ 인문계~ 언어~ 일본어 C010101~ 일본어~ <NA>   
## 9 2020.12.31 전문대~ 전문대학~ 인문계~ 언어~ 일본어 C010101~ 일본어~ <NA>   
## 10 2020.12.31 전문대~ 전문대학~ 인문계~ 언어~ 중국어 C010102~ 관광중~ <NA>   
## # ... with 9,280 more rows, and 28 more variables: 졸업자\_계 <dbl>,  
## # 취업률\_계 <dbl>, 취업자\_합계\_계 <dbl>, 취업자\_교외취업자\_계 <dbl>,  
## # 취업자\_교내취업자\_계 <dbl>, 취업자\_해외취업자\_계 <dbl>,  
## # 취업자\_농림어업종사자\_계 <dbl>, 취업자\_개인창작활동종사자\_계 <dbl>,  
## # 취업자\_1인창(사)업자\_계 <dbl>, 취업자\_프리랜서\_계 <dbl>, 진학률\_계 <dbl>,  
## # 진학자\_계 <dbl>, 취업불가능자\_계 <dbl>, 외국인유학생\_계 <dbl>,  
## # 제외인정자\_계 <dbl>, 기타\_계 <dbl>, 미상\_계 <dbl>, ...

df\_취업통계 |>  
 ggplot() +  
 geom\_bar(aes(x = 취업률\_그룹, y = ..count..))



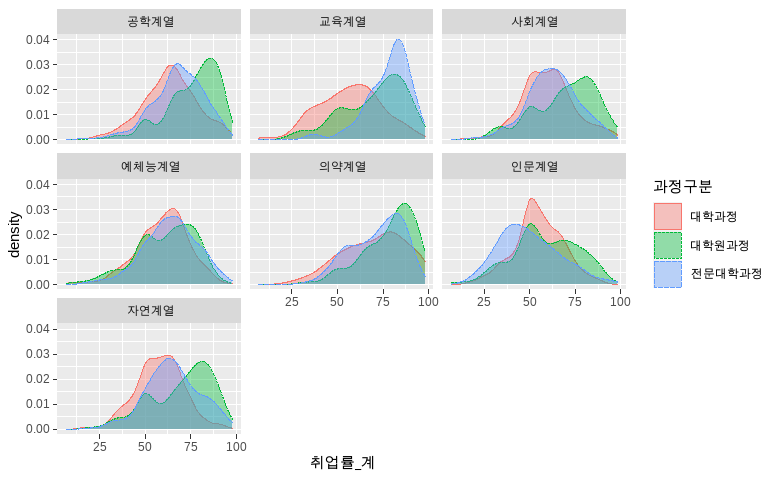
df\_취업통계$취업률\_그룹 <- case\_when(   
 df\_취업통계$취업률\_계 < 30 ~ '30%이하',   
 df\_취업통계$취업률\_계 >= 30 & df\_취업통계$취업률\_계 < 40 ~ '30%대',   
 df\_취업통계$취업률\_계 >= 40 & df\_취업통계$취업률\_계 < 50 ~ '40%대',   
 df\_취업통계$취업률\_계 >= 50 & df\_취업통계$취업률\_계 < 60 ~ '50%대',   
 df\_취업통계$취업률\_계 >= 60 & df\_취업통계$취업률\_계 < 70 ~ '60%대',   
 df\_취업통계$취업률\_계 >= 70 & df\_취업통계$취업률\_계 < 80 ~ '70%대',   
 df\_취업통계$취업률\_계 >= 80 & df\_취업통계$취업률\_계 < 90 ~ '80%대',   
 df\_취업통계$취업률\_계 >= 90 & df\_취업통계$취업률\_계 < 100 ~ '90%대',  
 TRUE ~ '100%'  
 )  
  
df\_취업통계 |>  
 ggplot() +  
 geom\_bar(aes(x = 취업률\_그룹, y = ..count..))



# 밀도 분포 그래프

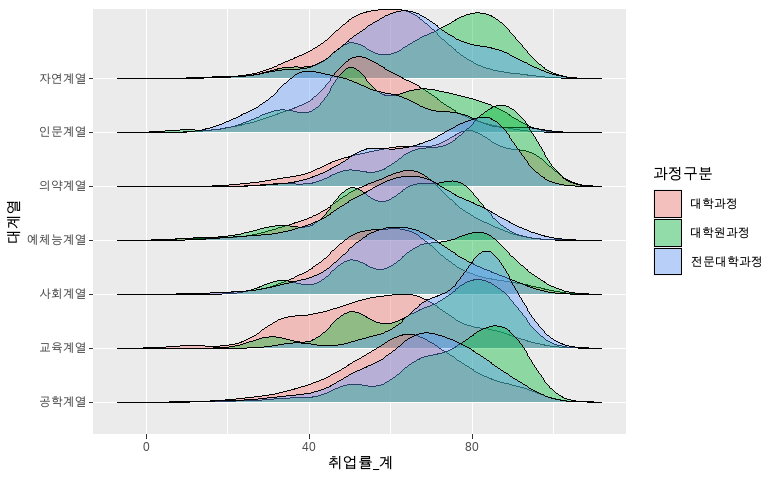
## 다중 밀도 분포 그래프

p\_density <- df\_취업통계 |> filter(취업률\_계 != 100, 취업률\_계 != 0) |>  
 ggplot()  
  
p\_density +  
 geom\_density(aes(x = 취업률\_계, color = 과정구분, fill = 과정구분, linetype = 과정구분), alpha=0.4, position = 'identity') +   
 facet\_wrap(~대계열)



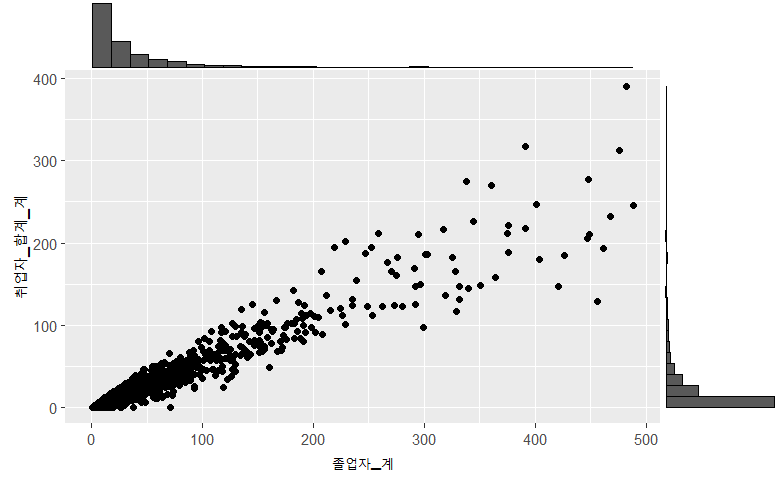
## ridge 그래프

if(!require('ggridges')) {  
 install.packages('ggridges')  
 library(ggridges)  
}  
  
p\_density +  
 geom\_density\_ridges(aes(x = 취업률\_계, y = 대계열, fill = 과정구분), alpha=0.4, position = 'identity')



## marginal

if(!require('ggExtra')) {  
 install.packages('ggExtra')  
 library(ggExtra)  
}  
  
p\_marginal <- df\_취업통계\_sample |>  
 ggplot() +  
 geom\_point(aes(x = 졸업자\_계, y = 취업자\_합계\_계))  
  
  
 ggMarginal(p\_marginal, df\_취업통계\_sample, x = 졸업자\_계, y = 취업자\_합계\_계, type = 'histogram')



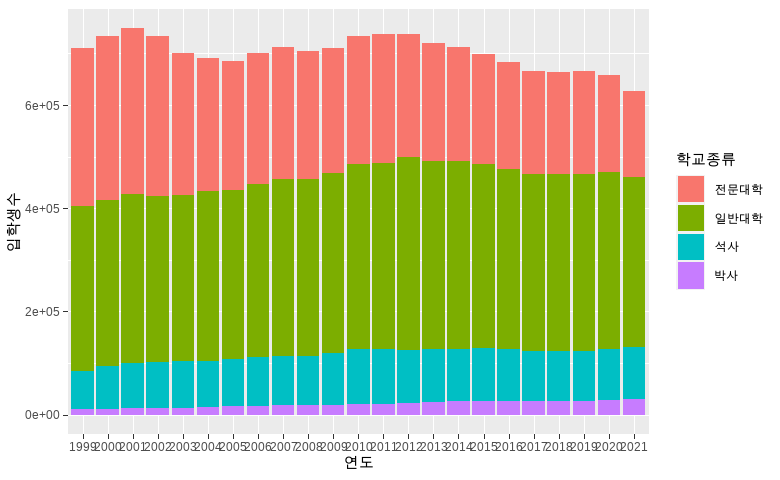
# 막대 그래프

## stack, dodge, fill

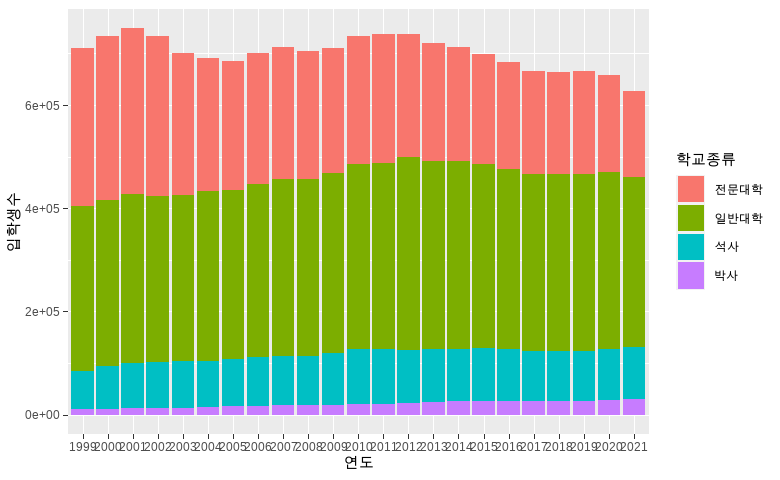
p\_col <- df\_입학자\_long |>  
 filter(학교종류 %in% c('전문대학', '일반대학', '석사', '박사'), 지역 == '전체') |>  
 mutate(학교종류 = fct\_relevel(학교종류, '전문대학', '일반대학', '석사', '박사')) |>  
 ggplot()  
  
p\_col +  
 geom\_col(aes(x = 연도, y = 입학생수))



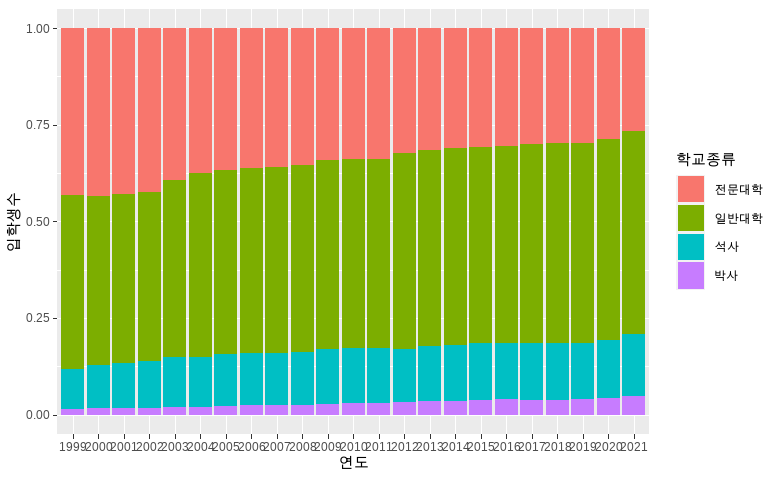
p\_col +  
 geom\_col(aes(x = 연도, y = 입학생수, fill = 학교종류))



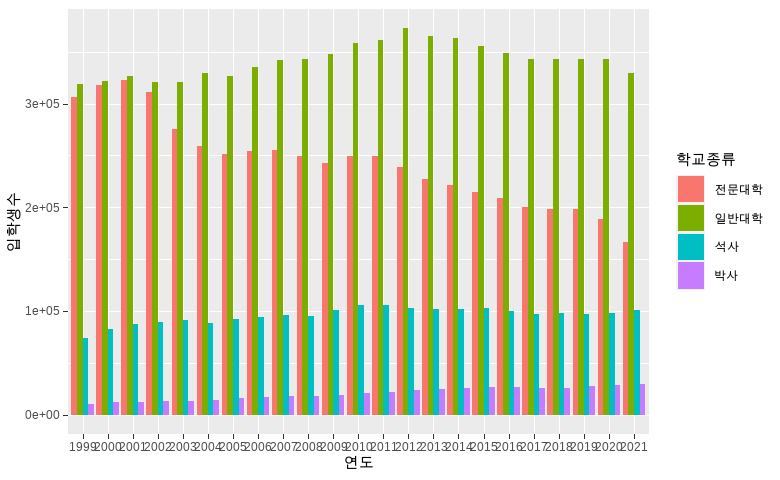
p\_col +  
 geom\_col(aes(x = 연도, y = 입학생수, fill = 학교종류), position = 'stack')



p\_col +  
 geom\_col(aes(x = 연도, y = 입학생수, fill = 학교종류), position = 'fill')

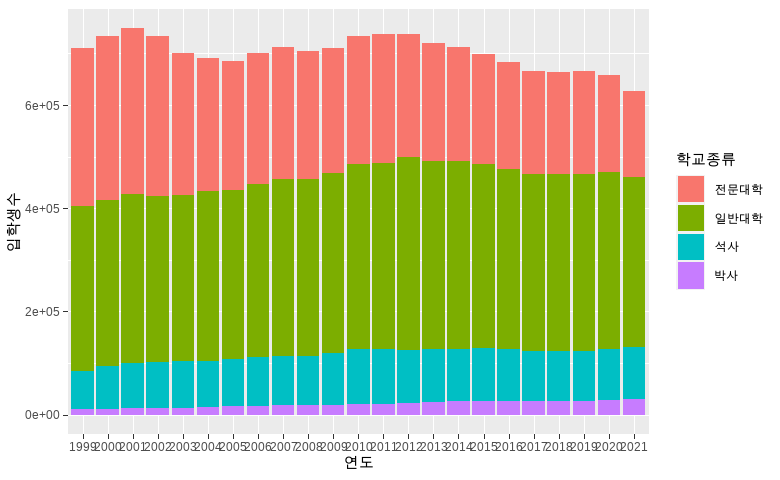


p\_col +  
 geom\_col(aes(x = 연도, y = 입학생수, fill = 학교종류), position = 'dodge')

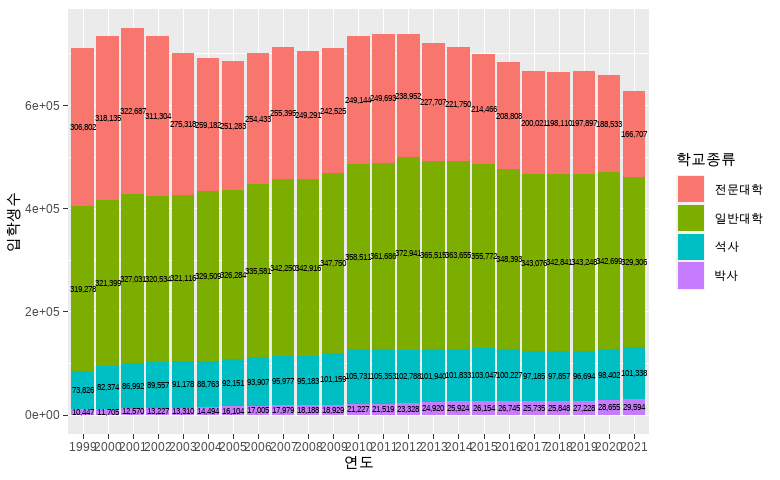


## 데이터 값 넣기

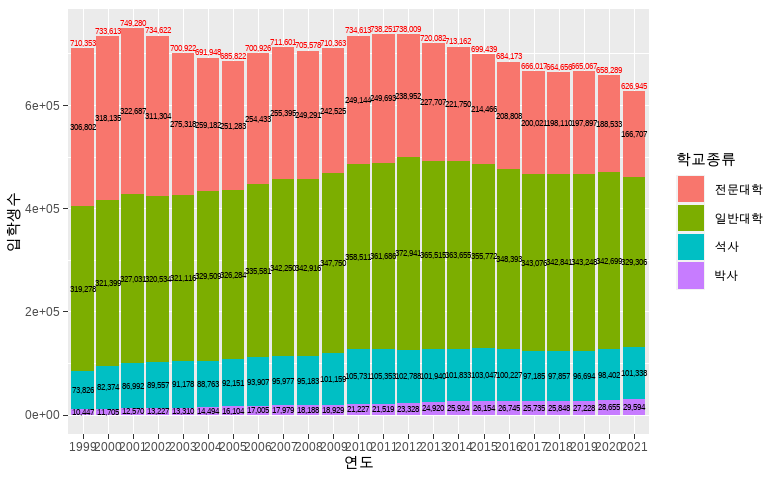
p\_col1\_1 <- p\_col +  
 geom\_col(aes(x = 연도, y = 입학생수, fill = 학교종류))  
  
p\_col1\_1



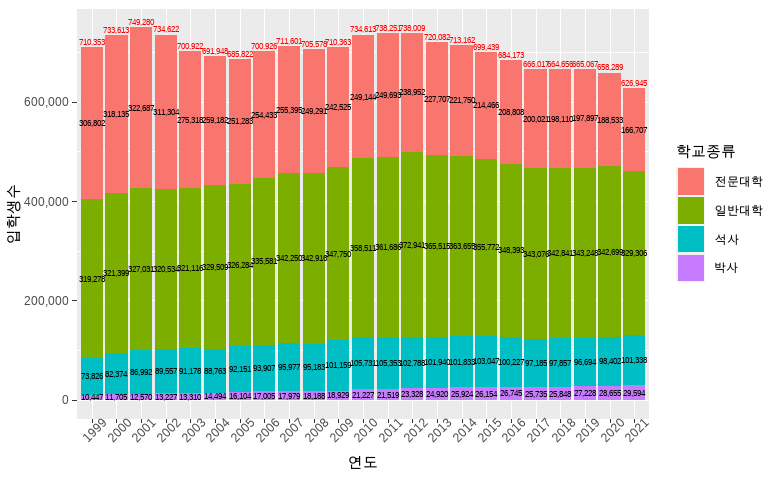
p\_col1\_2 <- p\_col1\_1 +   
 geom\_text(aes(x = 연도, y = 입학생수, label = scales::comma(입학생수), fill =학교종류), position = position\_stack(vjust = 0.5), size = 2)  
  
p\_col1\_2



p\_col1\_3 <- p\_col1\_2 +   
 stat\_summary(aes(x = 연도, y = 입학생수, label = scales::comma(stat(y), accuracy = 1)), fun = 'sum', geom = 'text', color = 'red', vjust = -0.5, size = 2, inherit.aes = FALSE)  
  
p\_col1\_3

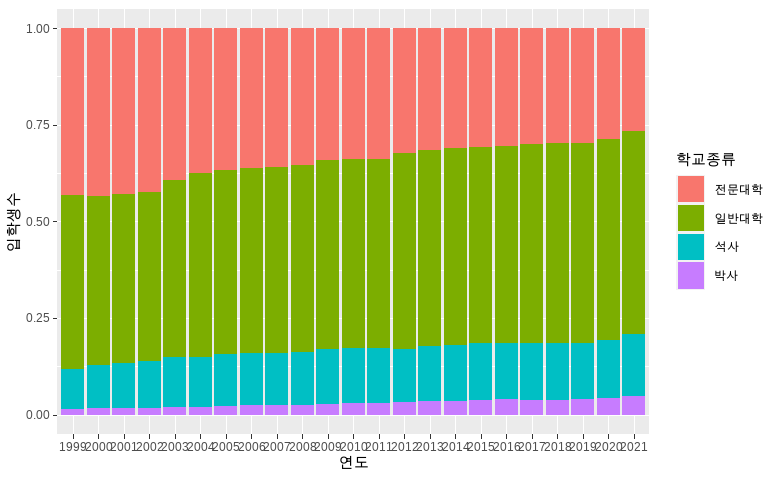


p\_col1\_4 <- p\_col1\_3 +   
 scale\_y\_continuous(labels = scales::comma) +  
 theme(axis.text.x = element\_text(angle = 45))   
  
p\_col1\_4

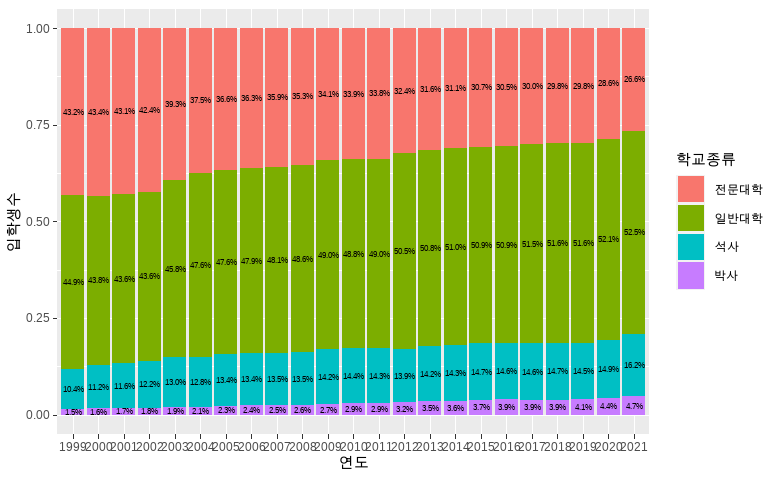


## 가로형 막대 그래프

p\_col\_2\_1 <- p\_col +  
 geom\_col(aes(x = 연도, y = 입학생수, fill = 학교종류), position = 'fill')  
  
p\_col\_2\_1



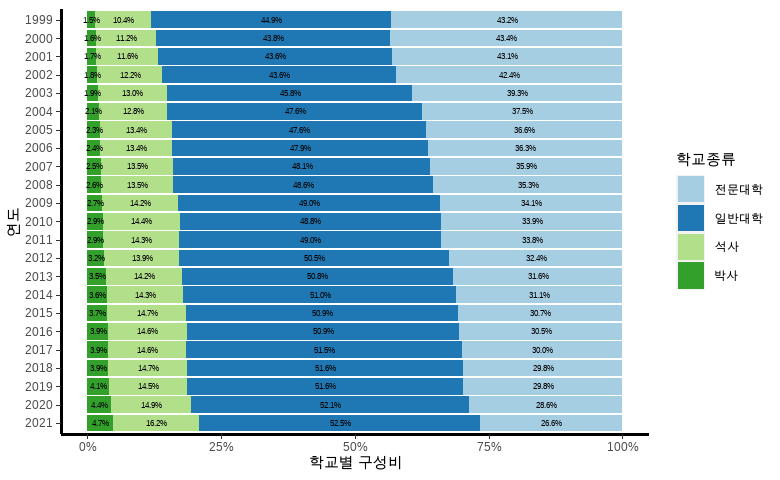
p\_col\_2\_2 <- p\_col\_2\_1 +   
 geom\_text(data = (df\_입학자\_long |>  
 filter(학교종류 %in% c('전문대학', '일반대학', '석사', '박사'), 지역 == '전체') |>  
 mutate(학교종류 = fct\_relevel(학교종류, '전문대학', '일반대학', '석사', '박사')) |>  
 group\_by(연도) |>  
 mutate(비율 = 입학생수 / sum(입학생수)) |>  
 ungroup()),   
 aes(x = 연도, y = 비율, label = scales::percent(비율, accuracy = 0.1), fill = 학교종류), position = position\_stack(vjust = 0.5), size = 2)  
  
p\_col\_2\_2



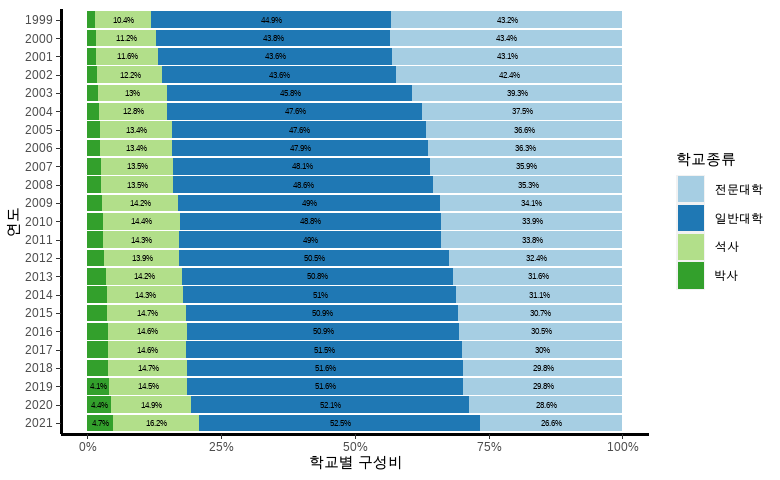
p\_col\_2\_3 <- p\_col\_2\_2 +   
 coord\_flip()  
  
p\_col\_2\_3



p\_col\_2\_3 +   
 scale\_y\_continuous(labels = scales::percent\_format(suffix = "%", prefix = "")) +  
 scale\_x\_discrete(limits=rev) +  
 # scale\_y\_discrete(labels = scales::percent\_format()) +  
 labs(y="학교별 구성비") +  
 theme(axis.line = element\_line(size=1, colour = "black"),  
 panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),  
 panel.border = element\_blank(), panel.background = element\_blank()) + scale\_fill\_brewer(palette="Paired")



p\_col\_2\_1 +   
 geom\_text(data = (df\_입학자\_long |>  
 filter(학교종류 %in% c('전문대학', '일반대학', '석사', '박사'), 지역 == '전체') |>  
 mutate(학교종류 = fct\_relevel(학교종류, '전문대학', '일반대학', '석사', '박사')) |>  
 group\_by(연도) |>  
 mutate(비율 = 입학생수 / sum(입학생수)) |>  
 ungroup()),   
 aes(x = 연도, y = 비율, label = ifelse(비율 > 0.04, paste0(round(비율, 3) \* 100,"%"), ''), fill = 학교종류), position = position\_stack(vjust = 0.5), size = 2) +   
 coord\_flip() +   
 scale\_y\_continuous(labels = scales::percent\_format(suffix = "%", prefix = "")) +  
 scale\_x\_discrete(limits=rev) +  
 # scale\_y\_discrete(labels = scales::percent\_format()) +  
 labs(y="학교별 구성비") +  
 theme(axis.line = element\_line(size=1, colour = "black"),  
 panel.grid.major = element\_blank(), panel.grid.minor = element\_blank(),  
 panel.border = element\_blank(), panel.background = element\_blank()) + scale\_fill\_brewer(palette="Paired")

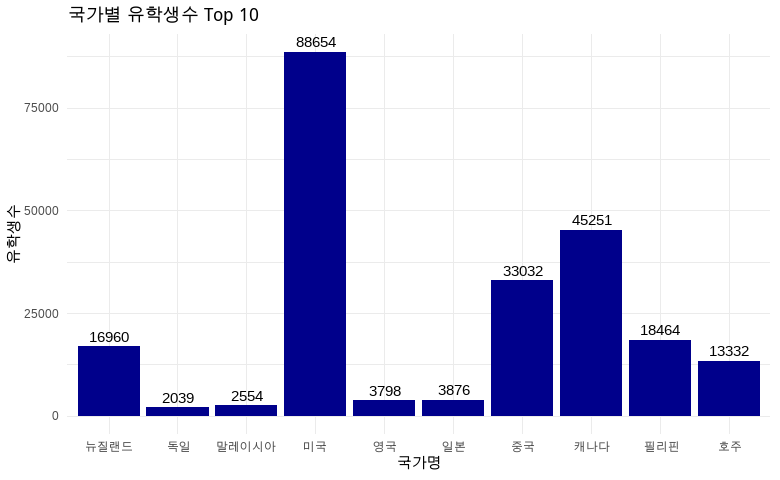


## 축 라벨에 이미지 넣기

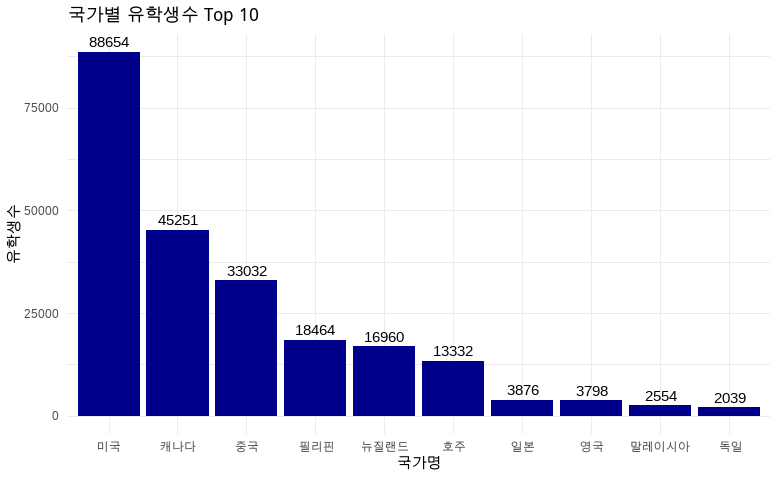
library(readxl)  
df\_nation <- read\_xlsx('파일경로/연도별 유학국가별 유학생수.xlsx', sheet = 'Sheet0', skip = 2, col\_types = c('numeric', 'text', rep('numeric', 25)), col\_names = TRUE)  
  
df\_nation <- df\_nation |>  
 filter(!is.na(학년도), 학제 == '소계') |>  
 select(!contains(c('계', '학제', '기타', '미확인', '그외동남아', '남미')))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

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) +   
 theme\_minimal() +   
 labs(title = '국가별 유학생수 Top 10', y = '유학생수')

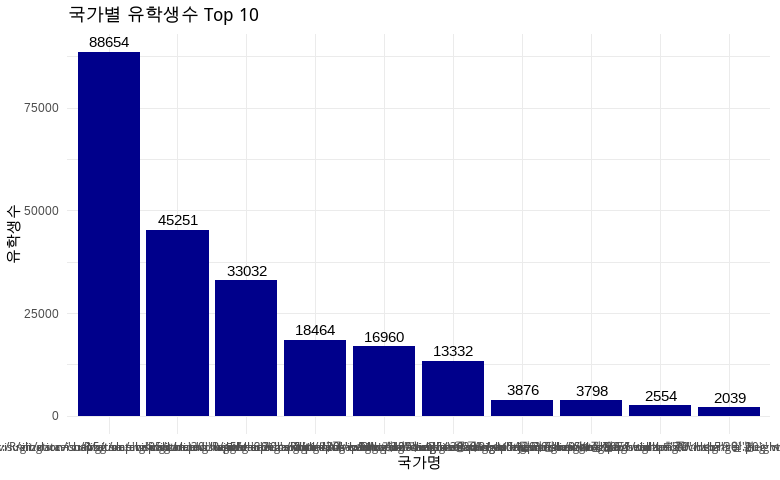


df\_nation\_top10$국가명 <- fct\_reorder(df\_nation\_top10$국가명, desc(df\_nation\_top10$sum))  
  
p\_nation\_top10 <- df\_nation\_top10 |>  
 ggplot(aes(x = 국가명, y = sum)) +  
 geom\_col(fill = 'dark blue') +  
 geom\_text(aes(x = 국가명, y = sum, label = sum), vjust = -0.5) +   
 theme\_minimal() +   
 labs(title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



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())

