Untitled

# III. 데이터 시각화를 위한 ggplot2

지금까지 데이터 시각화를 위한 R의 기초 사용법에 대해 알아보았다. 이제 본격적으로 ggplot2 패키지를 위주로 데이터 시각화 방법을 알아보도록 하자.

## 1. ggplot2 란

ggplot2 패키지는 R에서 데이터 시각화를 위해 가장 널리 사용되는 패키지이다. 이 패키지는 R-Studio의 수석 데이터 사이언티스트인 Hadley Wickham이 주도적으로 개발한 패키지로 2005년 발간된 Leland Wilkinson의 The Grammar of Graphics을 기본으로 작성되었다.

The Grammar of Graphics는 데이터를 효과적으로 요소를 다음의 7가지로 구분하였다. ggplot2에서는 이 7가지 요소를 사용하여 데이터를 시각화하도록 각종 함수들을 제공하고 있다.

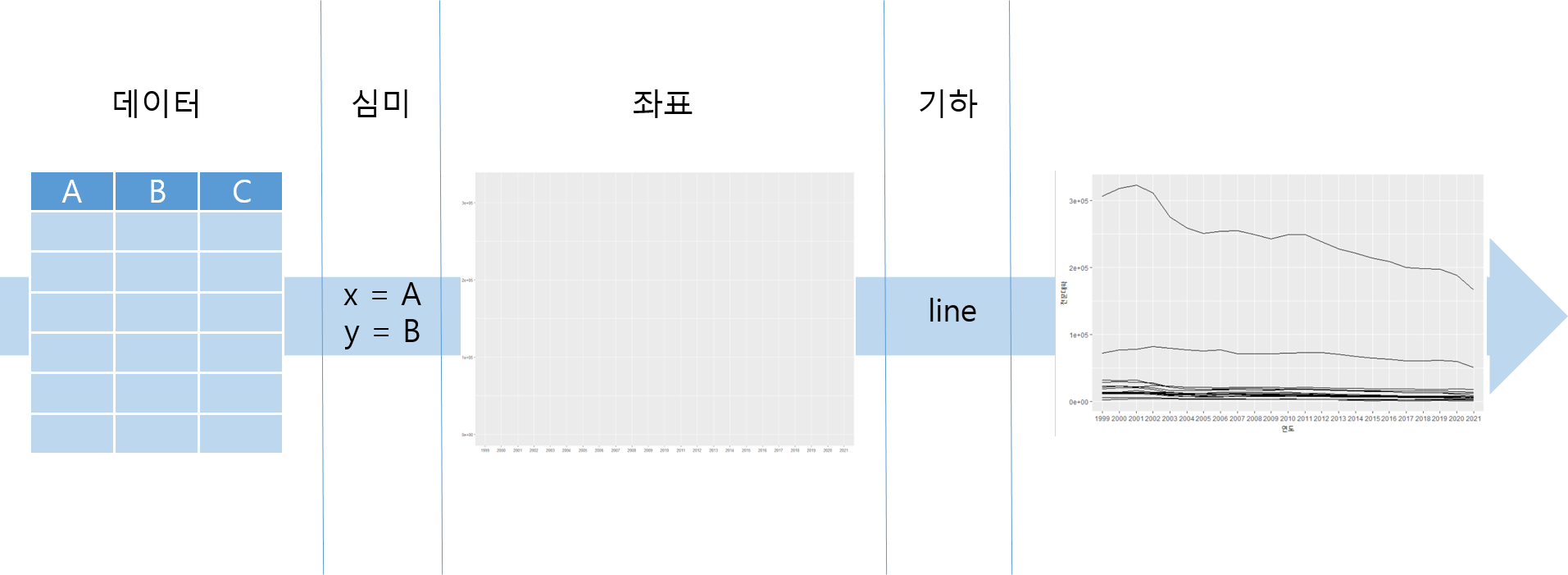
* data(데이터)
* 시각화에서 표현해야할 데이터를 지정한다. 하나의 ggplot2 시각화에는 하나 이상의 데이터를 사용할 수 있는데 최소 하나 이상의 데이터가 필수적으로 포함되어야 한다. ggplot2에서 지원하는 데이터 타입은 데이터프레임이나 tibble이다. 내부적으로 데이터프레임은 tibble로 변환되어 사용된다.
* Aesthetics(심미요소)
* Aesthetics는 데이터 시각와에서 시각적 속성과 해당 시각적 속성을 연결시키는 매핑 정보를 표현한다. Aesthetics로 매핑될 수 있는 시각적 속성은 X, Y축, 색깔, 크기 등이다.
* Geometries(기하요소)
* Geometries는 데이터 시각화에서 실질적으로 데이터를 표현하는데 사용되는 기하학적 도형을 말한다. 기하학적 도형은 점, 선, 막대 등이 있는데 각각의 기하 요소는 각각의 레이어로 시각화되고 이 기하 요소의 레이어를 여러개 사용하여 동시에 여러개의 기하요소를 사용할 수 있다.
* Facet(분할요소)
* 데이터 시각화에 표현되는 데이터가 일변량(univariate)이 아닌 다변량(multivariate)인 경우 하나의 시각화에 각 변량의 기하요소를 여러개 표현하면 여러 기하요소가 혼재되어 데이터 시각화를 통한 분석이 어렵게 된다. 이런 경우 다변량 기하 요소를 일변량화하여 일변량 시각화로 분할하여 표현할 수 있다.
* Statistics(통계요소)
* 데이터 시각화에 표현될 데이터가 원데이터가 아닌 mean, median 등 통계처리되어 표현할 경우 사용되는 요소이다.
* Coordinates(좌표요소)
* 데이터 시각화에 사용되는 좌표계를 설정한다. 사용되는 좌표계는 우리가 흔히 X, Y 축의 2차원 좌표계인 데카르트 좌표계(Cartesian Coordinates)나 극좌표계(Polar Coordinates) 등이 제공된다.
* Theme(테마요소)
* 시각화 제목, 축 제목, 축 단위, 범례 등 데이터 시각화의 전반적인 디자인을 꾸며주는 각종 요소를 말한다. 자신만의 Theme을 만들거나 미리 정의된 Theme을 적용하여 사용할 수 있다.

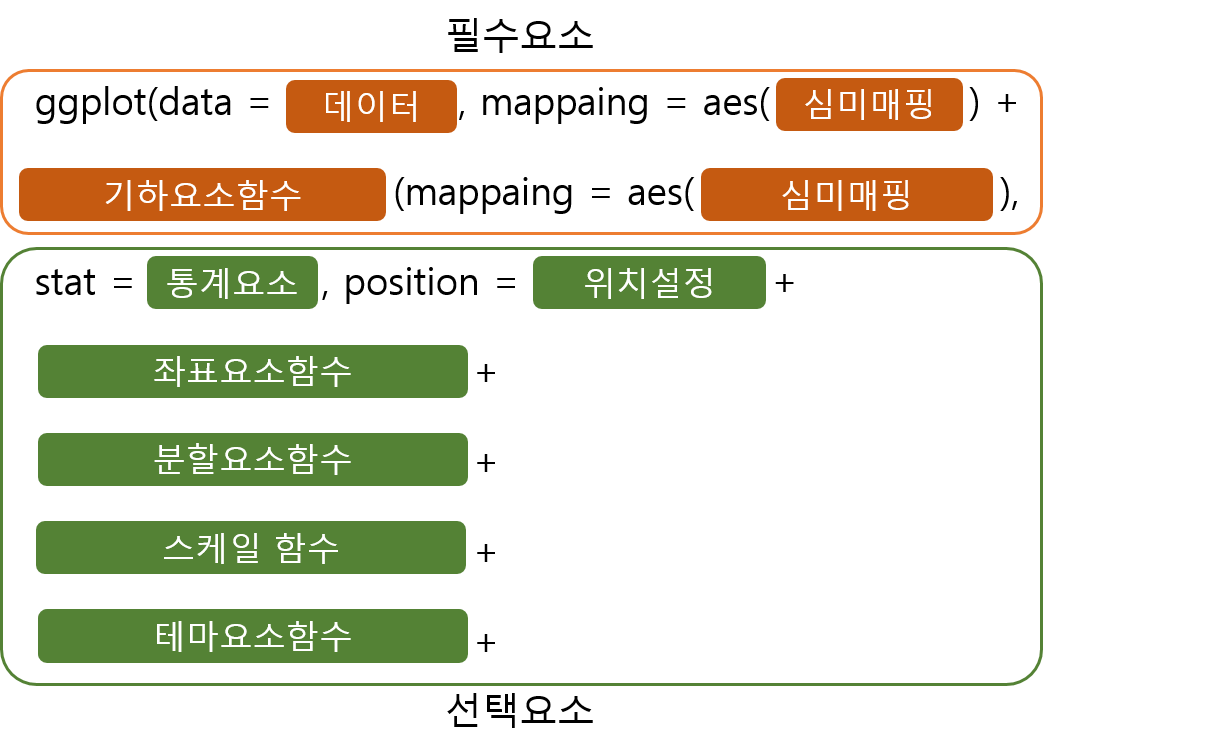
ggplot2 는 일정하고 반복적인 코드를 사용하여 일관된 문법을 제공한다는 점이 가장 큰 장점이다. 함수의 사용법이 일관적이어서 편리하고 읽기가 쉽다는 점과 그래픽의 완성도가 또한 큰 장점이기 때문에 R 사용자들이 가장 애용하는 데이터 시각화 패키지이다.

ggplot2 는 레이어를 추가하는 방식으로 여러 타입의 데이터 시각화 요소들을 동시에 표현할 수 있다. 위에서 설명한 7가지 레이어를 적절히 겹쳐서 시각화를 표현한다.

## 2. ggplot2 문법

ggplot2 에서 사용하는 7가지 요소중에 필수적으로 필요한 요소는 데이터, 좌표요소, 기하요소의 세가지이다. 하지만 데이터와 좌표 요소를 심미요소를 통해 연결하게 되기 때문에 결국 데이터, 좌표요소, 기하요소와 심미요소가 가장 기본적인 ggplot2 의 요소라고 할 수 있겠다. 이 중 좌표 요소는 별다른 함수를 사용하지 않아도 기본적으로 설정되고 나머지 데이터, 기하요소, 심미요소는 ggplot2 에서 제공하는 각종 함수를 사용하여 지정해야 한다.

  
ggplot2 의 문법은 아래의 그림과 같다. 시작은 반드시 ggplot() 함수부터이다. 이후는 순서에 큰 영향은 없으나 일반적으로 geom\_ 으로 시작하는 기하요소 함수를 사용하며 이 기하요소 함수 안에서 aes() 함수를 사용하여 심미 요소를 매핑한다. 각각의 기하요소에 공통적으로 적용될 심미요소의 매핑은 ggplot() 에 넣어 준다. ggplot() 를 사용하여 여러개의 레이어를 중첩하여 사용하기 위해서는 문법에서 제공하는 각각의 함수를 파이프가 아닌 + 기호 연결하여 사용한다.



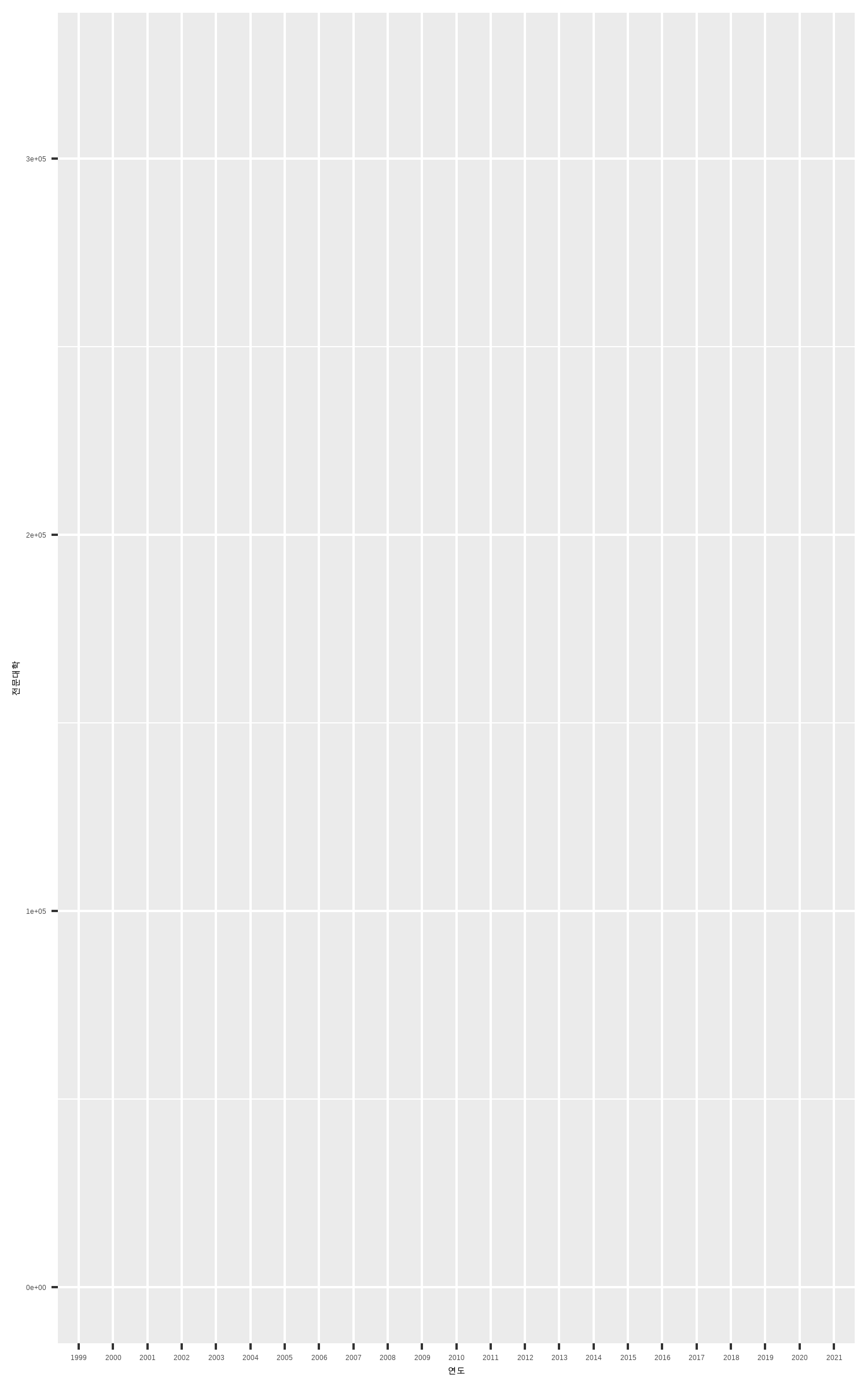
### 2.1 ggplot()

ggplot() 는 ggplot() 객체를 만들기 위해 초기화하는 함수이다. ggplot() 은 시각화를 위한 대상 데이터를 선언하고 전체 레이어에서 공통적으로 사용될 심미 요인들을 지정하는데 사용된다. ggplot() 의 사용법과 주요 매개변수는 다음과 같다.

ggplot(data = NULL, mapping = aes(), ...)  
 - data : ggplot 객체를 위해 사용할 전체 레이어에 공통으로 사용될 기본 데이터프레임을 선언  
`- mapping : aes()를 사용하여 전체 레이어에 공통으로 사용될 심미 요인 선언

ggplot() 에는 데이터에 연결된 기하요소가 선언되지 않기 때문에 데이터 시각화가 완전히 완성되지는 않는다. 하지만 데이터 심미요인의 매핑 결과는 볼 수 있다. 아래의 코드를 실행하면 앞 장에서 읽어들인 입학자 데이터를 X축과 Y축에 매핑한 결과를 볼 수 있다.

## df\_입학자 데이터를 사용하여 x축은 연도열, y축은 전문대학열을 매핑하는 ggplot 객체 생성  
ggplot(df\_입학자, aes(x = 연도, y = 전문대학))



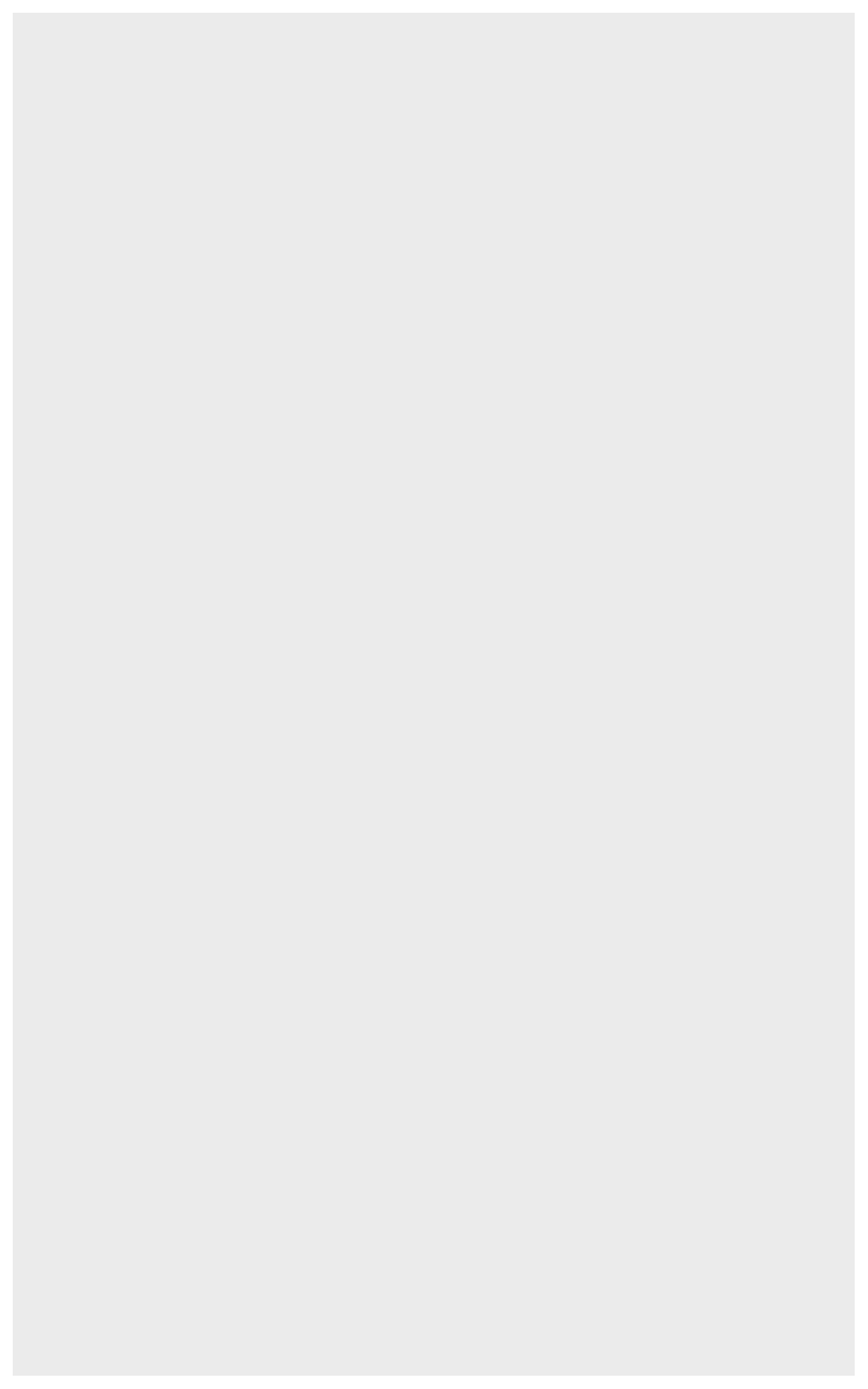
위의 코드에서 사용된 aes() 는 심미요인을 매핑을 생성하는데 사용되는 함수이다. 심미요인 매핑은 어떤 데이터의 열(변수)가 어떤 시각화 요인들로 연결되는지를 선언하는 방법을 말한다. 위의 코드에서 aes() 에 x 시각화 요인중에 X축을 나타내는 매개변수인데 이 x 가 df\_입학자 데이터프레임의 연도열과 연결되었다. 또 y 는 시각화 요인 중 Y축을 나타내는 매개변수로 이 y 가 df\_입학자의 전문대학열과 연결되었다. 이 함수는 ggplot() 뿐아니라 기하요소를 정의하는 geom\_\* 함수에서도 공통적으로 사용된다. aes() 의 사용법과 주요 매개변수는 다음과 같다.

aes(x, y, ...)  
 - x : X축에 매핑될 데이터 열  
 - y : y축에 매핑될 데이터 열

앞서 설명한 바와 같이 ggplot() 에 선언되는 데이터와 심미 요인은 전체 ggplot 객체의 레이어에 공통으로 적용되는 데이터와 심미요인이다. ggplot() 에 선언되는 심미요인은 X축과 Y축의 매핑만 하는 것이 일반적이다. 모든 레이어에 공통적으로 적용되는 심미요인이 많지 않기 때문에 미리 정의해봐야 일부 레이어에서만 사용되기 떄문이다. 물론 ggplot() 에서 선언된 심미요인들이 각각의 레이어에서 다시 선언되는 경우에는 ggplot() 에서 선언된 데이터와 심미 요인보다 각각의 레이어에서 선언되는 데이터와 심미 요인이 해당 레이어에서 우선된다.

만약 모든 레이어에 공통적으로 적용되는 데이터나 심미 요인들이 없다면 매개변수 없이도 사용될 수 있고 데이터만 선언될 수도 있다. 하지만 데이터 선언없이 심미요인의 선언은 불가하다. 데이터가 없거나 심미요인이 없다면 아래와 같이 빈 ggplot 객체가 생성된다.

ggplot(df\_입학자)

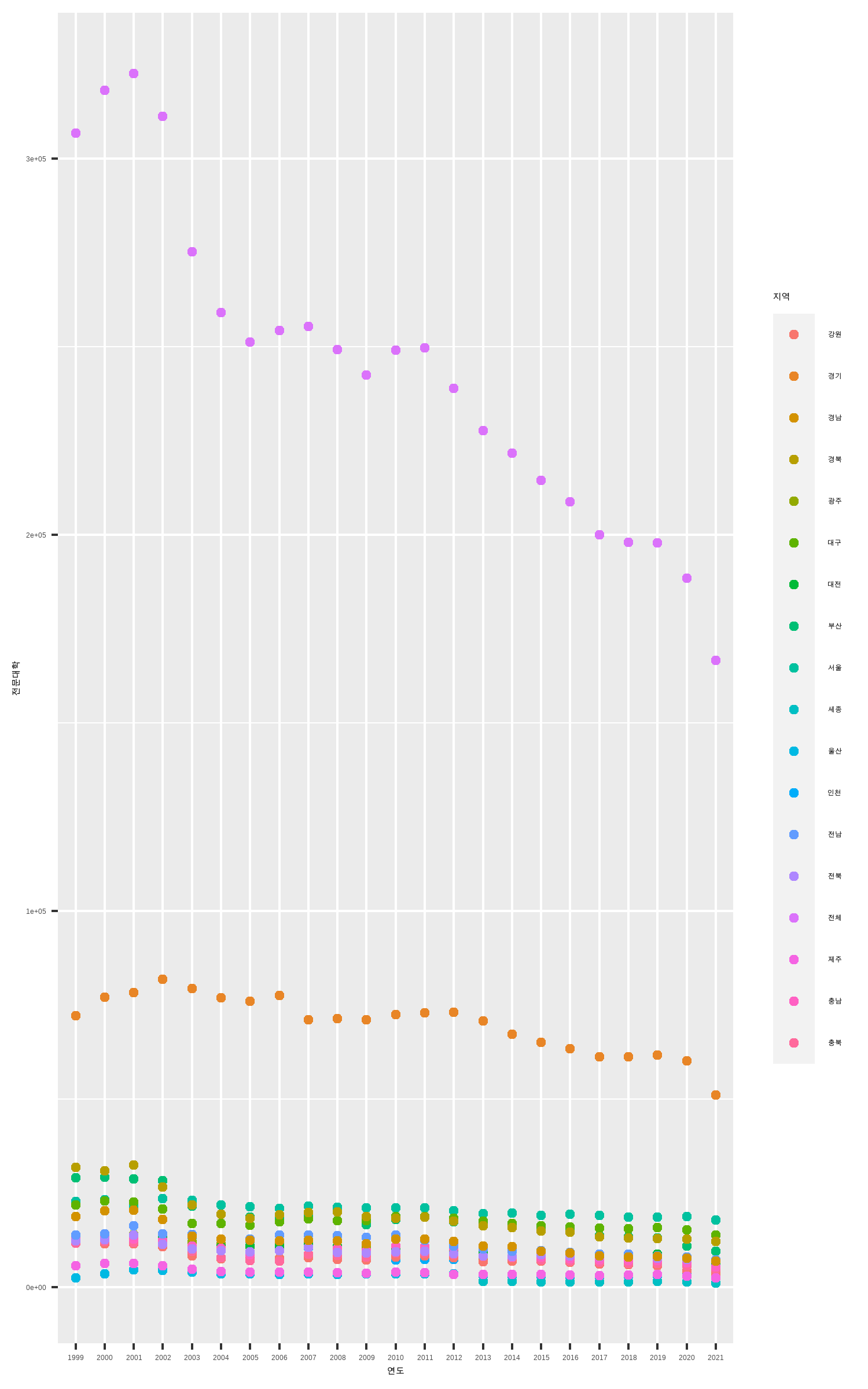


### 2.2 심미요소

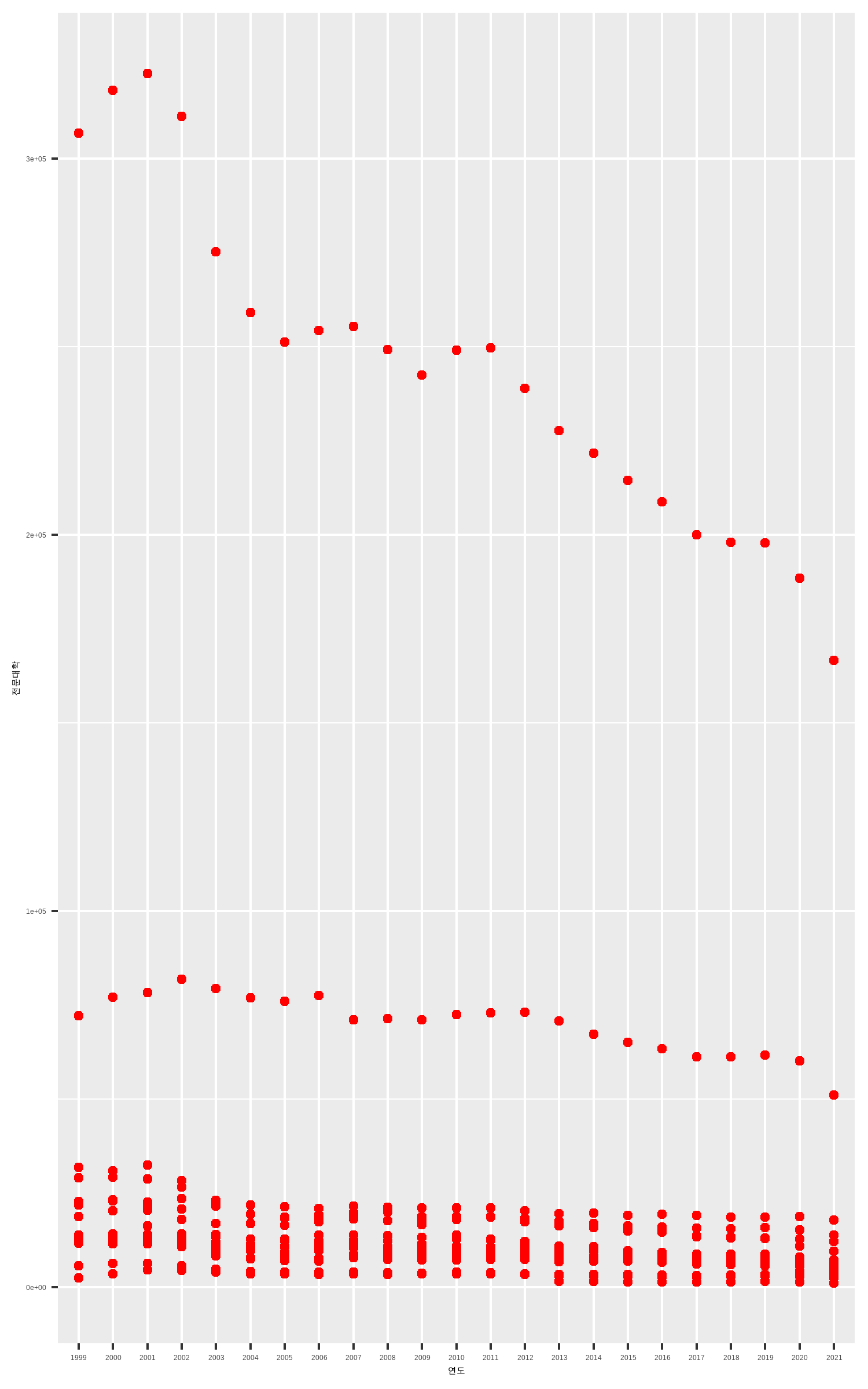
심미요소는 데이터를 표현하는데 사용되는 필요한 요소들을 통칭한다. 즉 심미요소로 지정할 수 있는 모든 요인은 ggplot 객체의 X, Y 좌표내에 표현되는 기하요소의 표현을 위한 요소들로 시각적 요소와 데이터 변수간의 매핑을 통해서 구현된다.

심미요소는 aes() 를 사용하여 매핑할 수 있고 고정값을 설정할 수 있다. 이 부분은 심미요인의 사용에 가장 혼동을 유발하는 부분이다. 심미요소를 매핑한다는 것은 심미요소가 데이터 변수에 의해 변경되어야 하는 경우 사용한다. 즉 심미요소가 변수에 대응됨으로써 변수의 변량에 따라 해당 심미요소들이 바뀌어서 표현된다. 그러나 고정값을 설정한다는 것은 변수에 대응되는 것이 아닌 특정값에 고정되도록 설정하기 때문에 고정값으로 설정된 기하요소들은 변수의 변량이 변경되어도 같은 심미요소로 표현된다. 심미요소는 사용하고자 하는 기하요소 함수(geom\_\*()) 내에서 사용되지만 매핑할때는 반드시 aes() 함수를 사용하여 데이터 열이나 매핑 변수를 설정하여야 하고 고정값을 설정할 때는 반드시 aes() 함수 바깥에서 선언되어야 한다. 다음의 예를 살펴보자.

df\_입학자 |> ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_point(aes(color = 지역))



df\_입학자 |> ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_point(color = 'red')



위의 코드에서 차이는 color 심미요인이 aes() 내에 df\_입학자의 열이 매핑된 것과 aes() 밖에서 고정값(red)로 설정된 것이다. 결과에서 보이듯이 매핑된 color 는 지역 변수의 변량에 따라 color 가 자동적으로 바뀌어 데이터의 구분이 확연히 보인다. 반면 aes() 밖에서 고정값인 red값으로 설정된 코드는 전체 값들이 모두 red 값으로 표현되어 지역적으로 구분되지 않는다.

설정이 가능한 심미요인들은 다음과 같다.

#### 2.2.1 위치(x , y, xend, yend)

x와 y는 기하요소가 표시될 X축의 위치와 Y축의 위치 설정에 필요한 데이터 열의 매핑을 설정한다. 일부 선을 그리거나 사각형을 그리는 geom\_segment() 나 geom\_rect() 와 같은 기하요소 함수에서는 x, y 부터 시작하여 xend, yend까지 기하요소를 그린다. 다른 심미요소와는 달리 x, y, xend, yend는 매핑과 고정값 설정시 모두 반드시 aes() 함수안에서 사용되야 한다.

#### 2.2.2 color

기하요인의 외곽선 색상을 설정한다. R에서 사용되는 색상은 색상 이름으로 설정하거나 RGB코드값으로 설정할 수 있다.

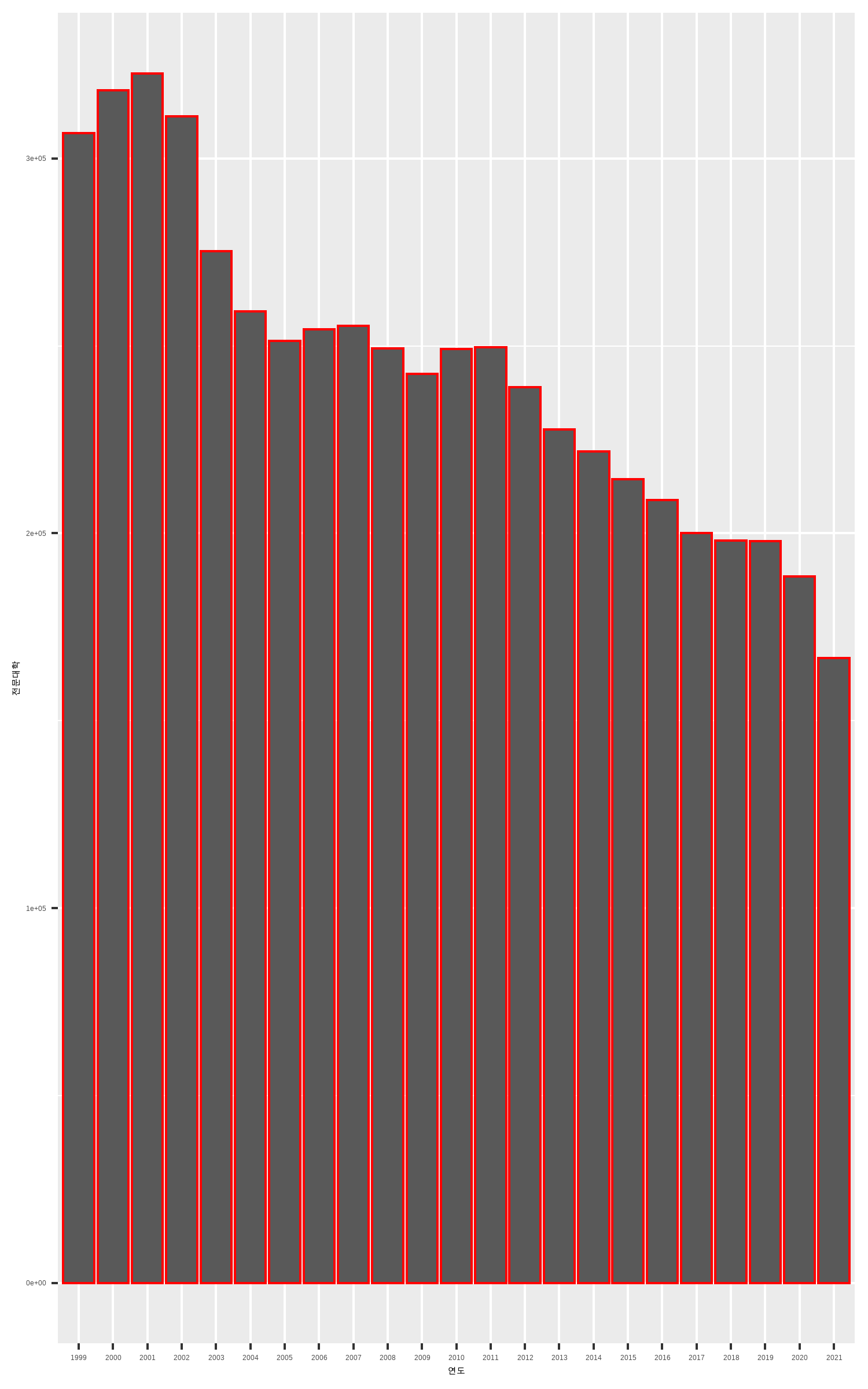
R에서 미리 정의된 색상 이름은 총 657개로 colors() 를 사용하면 확인이 가능하다.

## R에서 미리 정의된 색상 이름 출력, 지면 관계상 10개만 출력  
colors() |> head(10)

## [1] "white" "aliceblue" "antiquewhite" "antiquewhite1"  
## [5] "antiquewhite2" "antiquewhite3" "antiquewhite4" "aquamarine"   
## [9] "aquamarine1" "aquamarine2"

RGB 색상 코드는 HTML/CSS에서와 같이 RGB 코드를 16진수 값(00에서 FF)을 사용하여 2자리씩 정의하는데 “#” 접두사로 붙은 문자열로 설정한다. 예를 들어 red는은 “#FF0000”으로 표현된다.

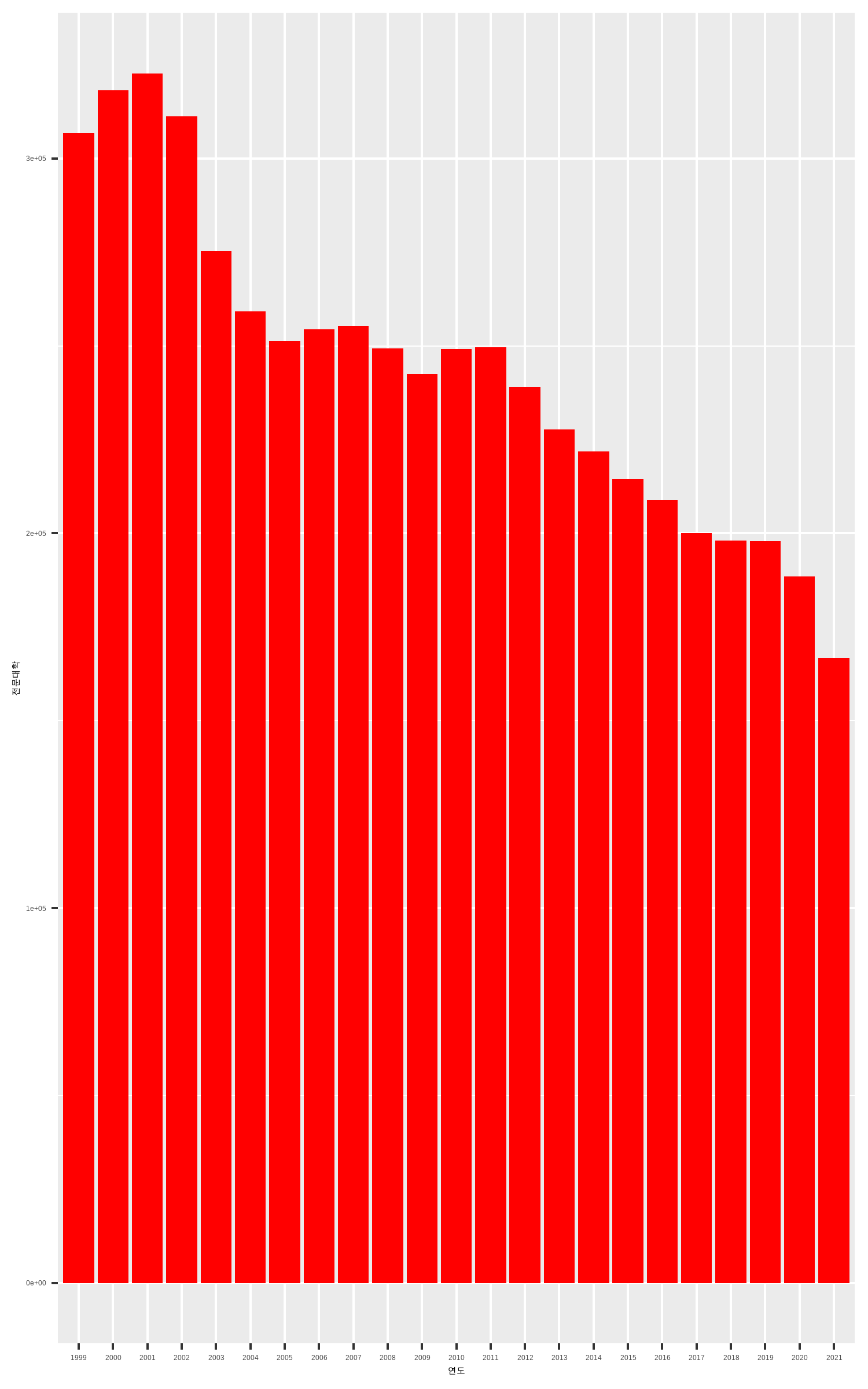
## df\_입학자의 지역이 '전체'인 데이터를 시각화하는데 막대의 외곽선 색상을 빨강으로 설정  
df\_입학자 |> filter(지역 == '전체') |>  
 ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_col(color = '#ff0000')



#### 2.2.3 fill

기하요인의 내부 색상의 설정을 설정한다. 색상의 설정은 color 설정과 동일하다.

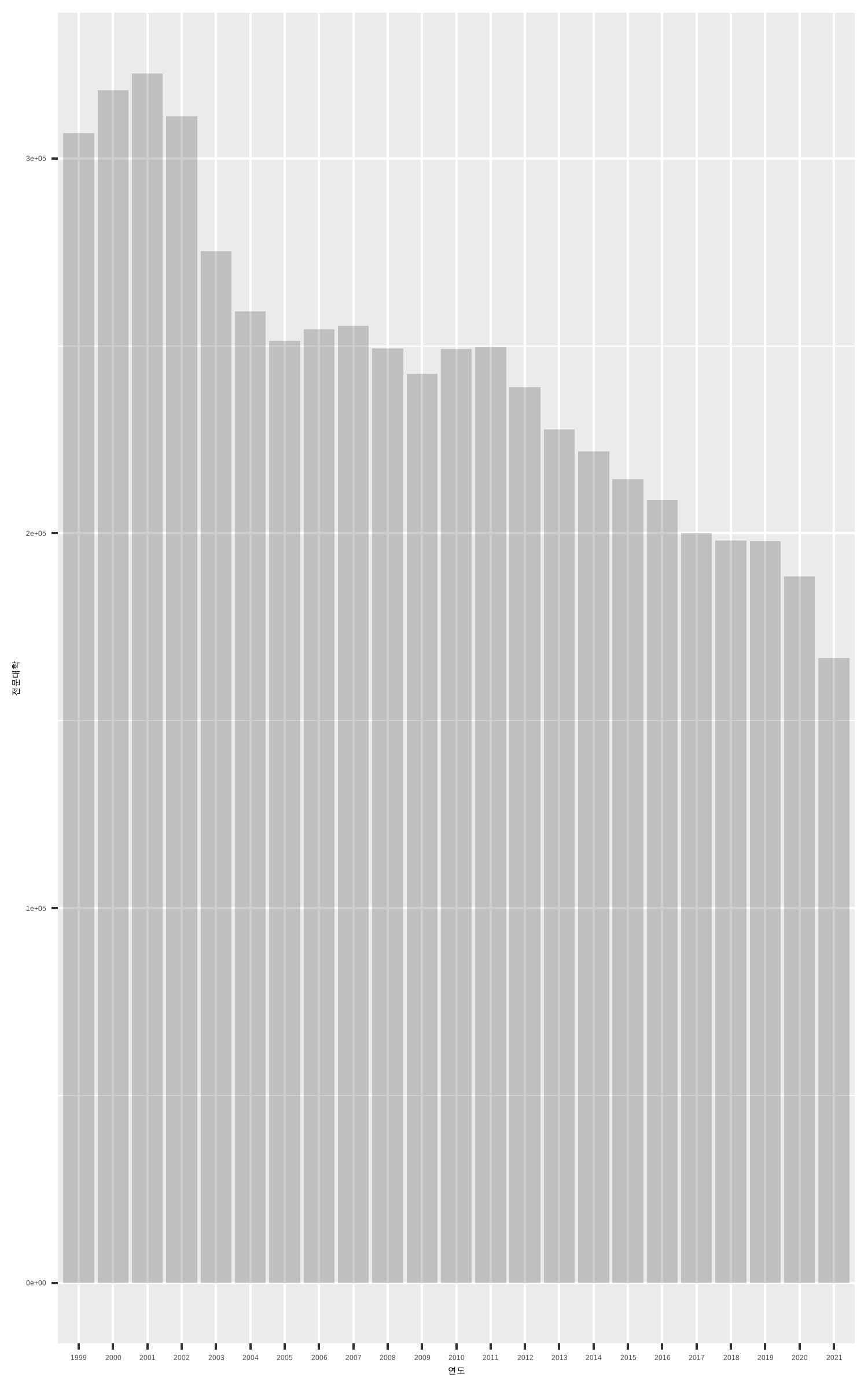
## df\_입학자의 지역이 '전체'인 데이터를 시각화하는데 막대의 내부 색상을 빨강으로 설정  
df\_입학자 |> filter(지역 == '전체') |>  
 ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_col(fill = '#ff0000')



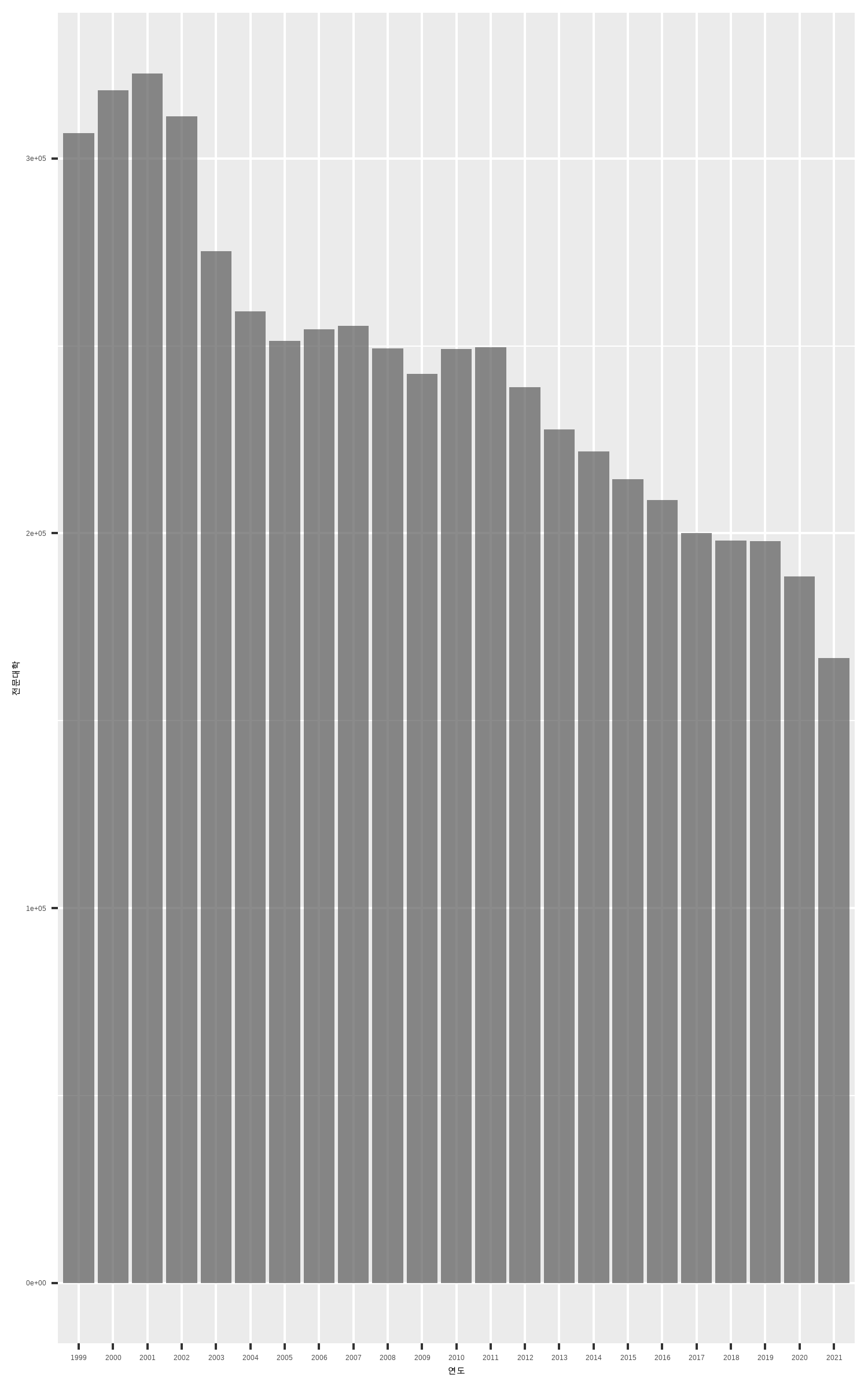
#### 2.2.4 alpha

기하요인의 투명도를 설정한다. alpha 는 정수형 수치로 설정하는데 0부터 1사이의 값을 가진다. 0에 가까울수록 투명해지고 1에 가까울 수록 불투명해진다.

## df\_입학자의 지역이 '전체'인 데이터를 시각화하는데 막대의 투명도를 0.3으로 설정  
df\_입학자 |> filter(지역 == '전체') |>  
 ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_col(alpha = 0.3)

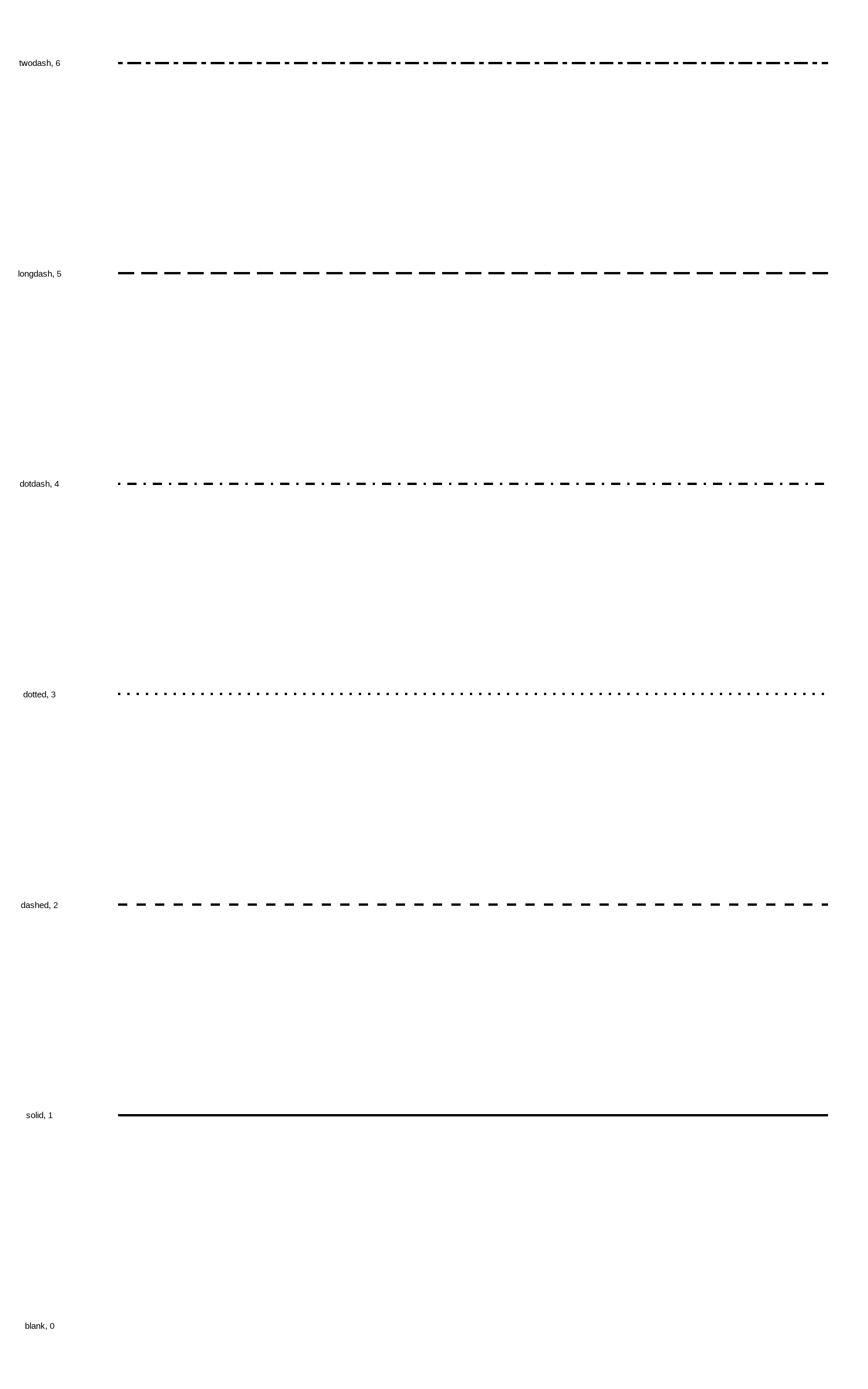


## df\_입학자의 지역이 '전체'인 데이터를 시각화하는데 막대의 투명도를 0.7로 설정  
df\_입학자 |> filter(지역 == '전체') |>  
 ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_col(alpha = 0.7)

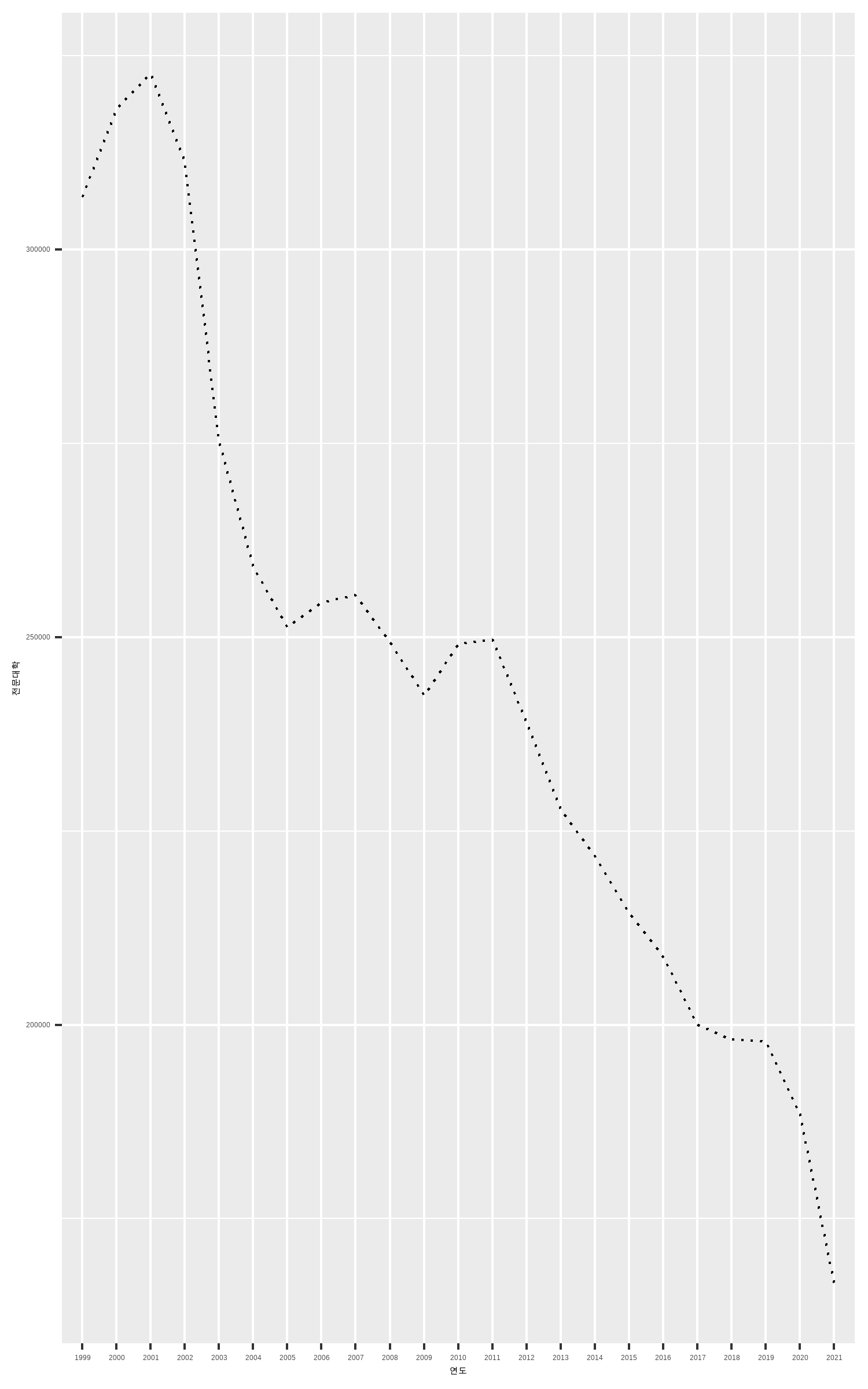


#### 2.2.5 linetype

기하요인 중 선으로 그려지는 기하 요인의 선 타입을 결정한다. R에서는 총 7가지의 선 타입을 제공하는데 0부터 6까지의 숫자나 선 타입의 이름을 사용하여 설정할 수 있다.



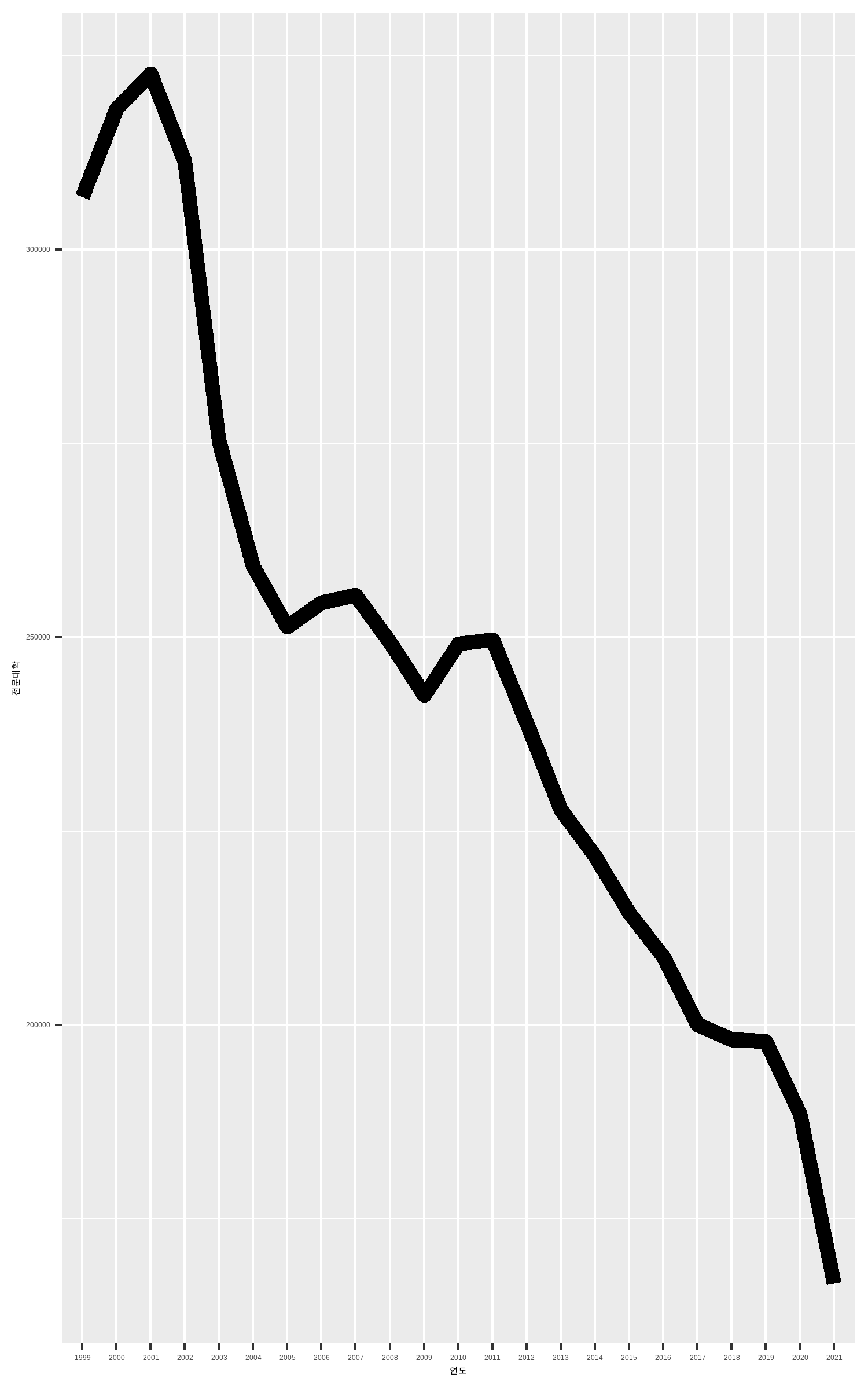
## df\_입학자의 지역이 '전체'인 데이터를 시각화하는데 선의 타입을 'dashed'로 설정  
df\_입학자 |> filter(지역 == '전체') |>  
 ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_line(aes(group = 1), linetype = 'dotted')



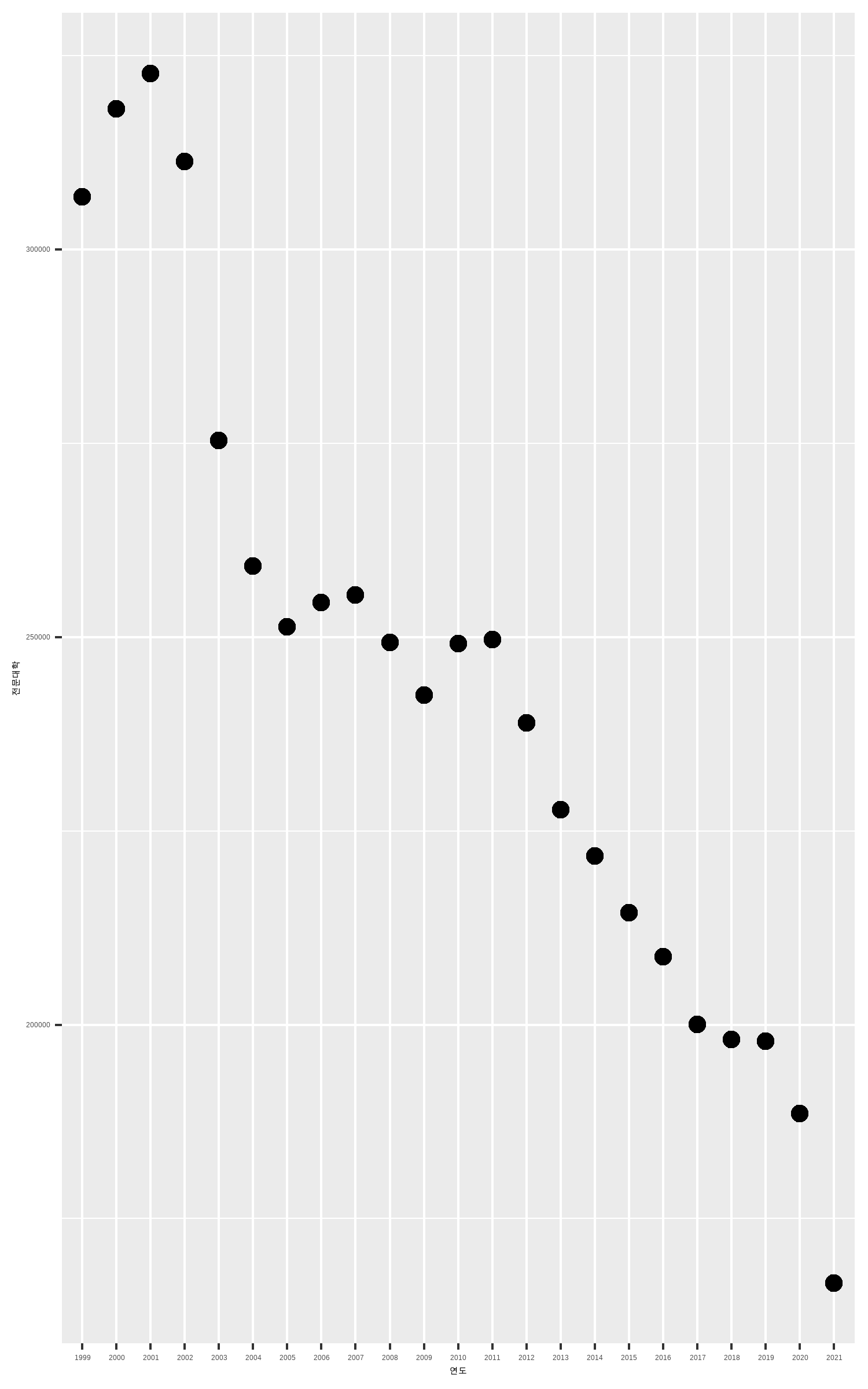
#### 2.2.6 size

size 는 기하요소의 크기를 설정한다. 기하요소가 점이면 점의 크기, 선이면 선의 굵기를 설정한다. 점의 크기를 결정할 때는 반지름의 길이를 밀리미터 단위로 지정한다. 또 선의 굵기도 밀리미터 단위로 지정할 수 있다.

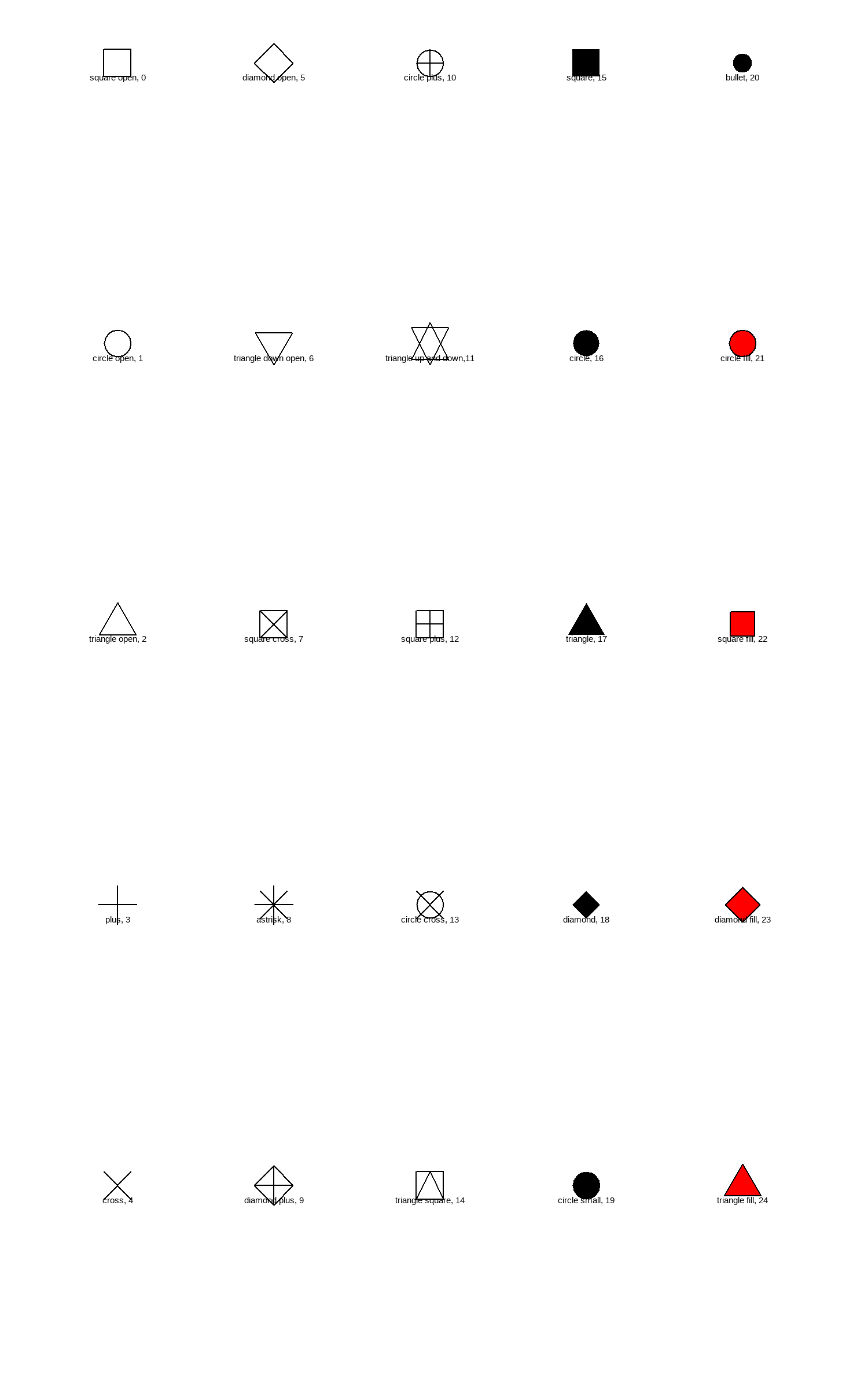
df\_입학자 |> filter(지역 == '전체') |>  
 ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_line(aes(group = 1), size = 3)



df\_입학자 |> filter(지역 == '전체') |>  
 ggplot(aes(x = 연도, y = 전문대학)) +  
 geom\_point(size = 3)



#### 2.2.7 shape



### 2.3 기하요소

기하요소는 ggplot() 로 생성된 초기화 ggplot 객체에 데이터를 표현하는 방법을 지정하는 요소이다. 2.기하요소에는 여러가지가 있지만 우리가 흔히 생각하는 것은 점(Point), 선(Line), 막대(Bar, Col) 등이 대표적이다.

기하요소를 생성하기 위해서는 geom\_\*() 함수를 사용하고 각각의 geom\_\*() 함수를 호출할 때마다 각각의 기하요소 레이어가 생성되고 이 레이어들이 계속 겹쳐서 그려짐으로써 데이터 시각화가 진행된다. 기하요소를 위한 geom\_\*() 의 주요 함수는 각각의 함수에 따라 선언되는 시각화 요소들이 다르지만 데이터, 심미요소 매핑, 기하요소 지정, 통계 변환, 위치 조정 등이 선언된다.

기하요소를 선택할 때는 각각의 기하요소에 따라 표현되는 값의 제한이 있다. 예를 들어 막대 그래프의 경우 Y축은 연속된 정수값이 표현되는 것이 가능하지만 X축에는 연속된 정수값이 아닌 값의 구별이 가능한 이산 값(discrete value)가 와야한다. 그래야 분리된 하나의 이산값에 하나의 막대가 표현될 수 있다. 이렇게 표현하고자 하는 값의 종류에 따라 적합한 기하요소를 선택하여야 한다. 값의 종류는 연속된 값(continuous value), 분리된 이산값(discrete value) 나 팩터(factor)의 여부, 단변량, 다변량의 여부 등으로 나눌 수 있다.

#### 2.3.1 연속된 단변량 기하요소

연속된 단변량 기하요소는 수치형 데이터 하나를 시각화할 때 사용할 수 있는 기하요소이다. ggplot() 는 보통 X축과 Y축의 2차원 표현이 기본이기 때문에 하나의 데이터만이 정의되면 나머지 하나의 데이터는 자동적으로 결정된다. 이렇게 자동적으로 결정되는 데이터는 보통 면적(Area), 밀도분포(Density), 도수 분포(Histogram) 등이다.

##### 2.3.1.1 geom\_histogram

geom\_histogram 은 도수분포표를 그리는 기하요소 함수이다. 도수분포는 초등학교때 배우는 가장 기본적인 막대그래프로 각각의 변수 변량에 따른 데이터의 개수를 표현하는 시각화 방법이다. 변수 변량에 따른 데이터의 개수를 표현하기 때문에 X축 데이터만 설정하면 데이터를 자동적으로 분석하여 X축에 매핑된 변수의 변량별로 데이터 개수를 산출하게 되고 이 개수를 막대 그래프로 표현하게 된다. 따라서 도수분포는 막대그래프에 속하는 종류 중 하나일 뿐이다.

앞에서 막대그래프는 연속된 수치값이 아닌 분리된 이산값이 X축에 매핑되어야 한다고 설명는데 연속된 단변량 기하요소에 geom\_histogram() 이 속하는 것은 왜일까? geom\_histogram() 은 연속된 수치값을 X축에 매핑한다. 하지만 내부적으로 적절한 단위로 전체 X값을 분리하여 이산값으로 만들어 준 후에 막대그래프를 생성해 준다. 이 과정이 통계요인이고 geom\_histogram() 에서 사용되는 유일한 통계요인은 연속된 값을 층화하여 구간하는 방법인 binning이다. 이처럼 자동적으로 계산되는 binning 떄문에 연속된 단변량 기하요소에 속한다.

geom\_histogram 의 사용법과 주요 매개변수는 다음과 같다.

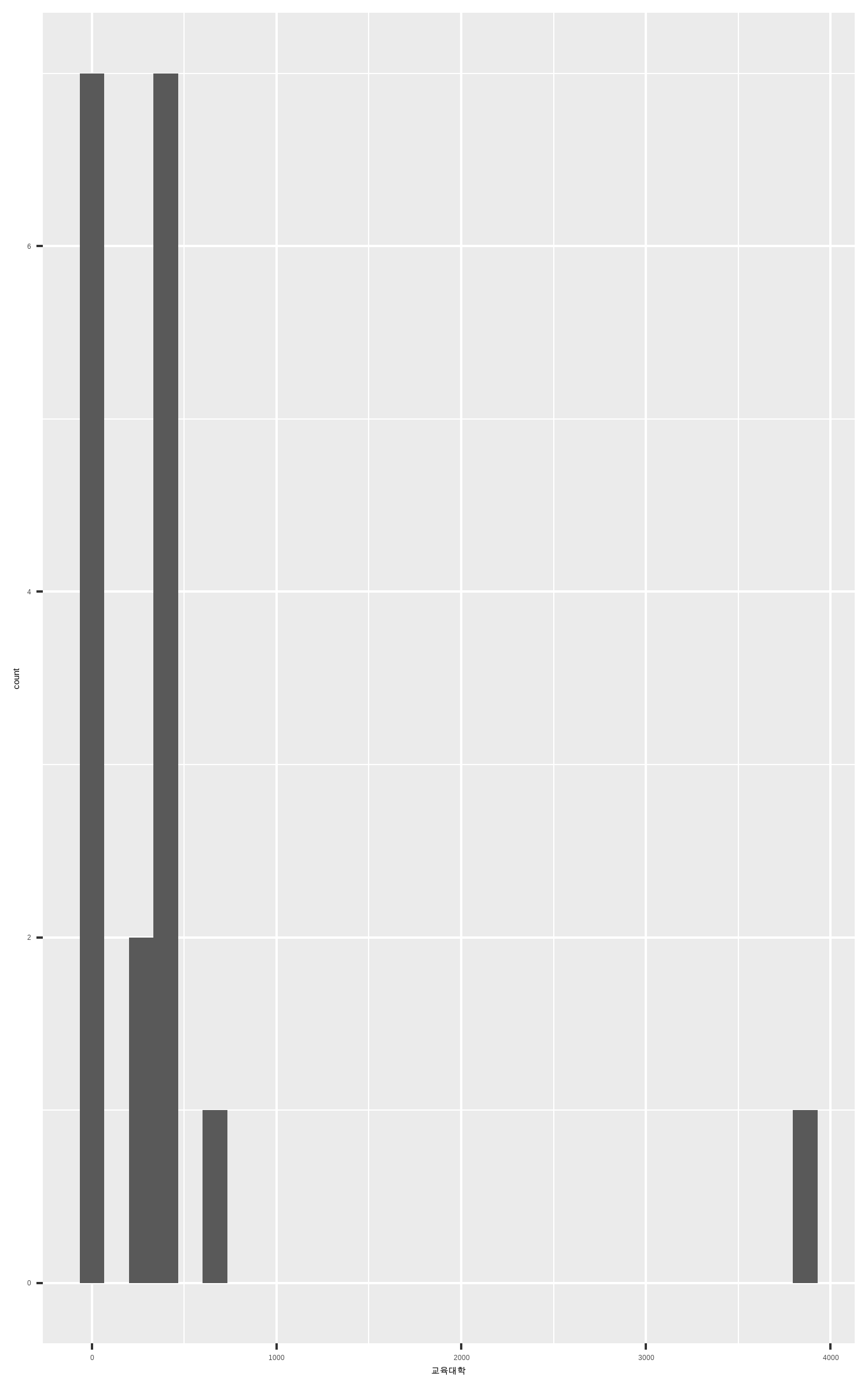
geom\_histogram(mapping = NULL, data = NULL, bins, binwidth, ...)  
 - mapping : aes()를 사용하여 매핑할 심미요소, 생략되면 ggplot()에 정의된 심미매핑 사용  
 - data : 시각화를 위해 사용될 데이터, 생략되면 ggplot()에 정의된 데이터 사용  
 - bins : X축을 나누는 bin의 개수 설정  
 - binwidth : X축을 나누는 bin의 너비 설정, 숫자벡터를 사용할 수 있다. (bin과 binwidth는 동시에 사용될 수 없다)

geom\_histogram() 에 매핑될 수 있는 심미요인은 다음과 같다.

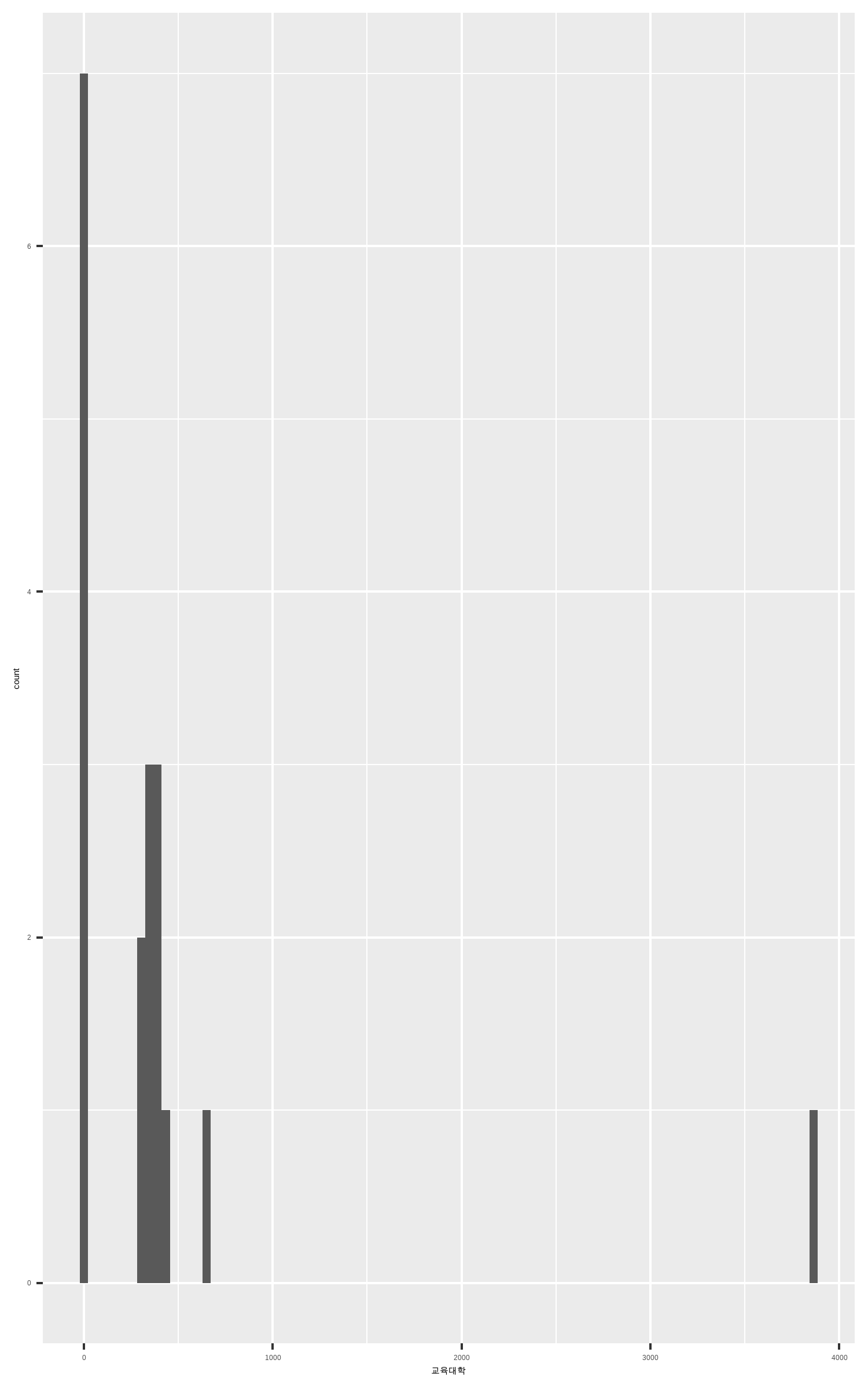
mapping = aes(x, y, alpha, color, fill, linetype, size, weight)

df\_입학자를 사용하여 geom\_historgram() 을 그려보자.

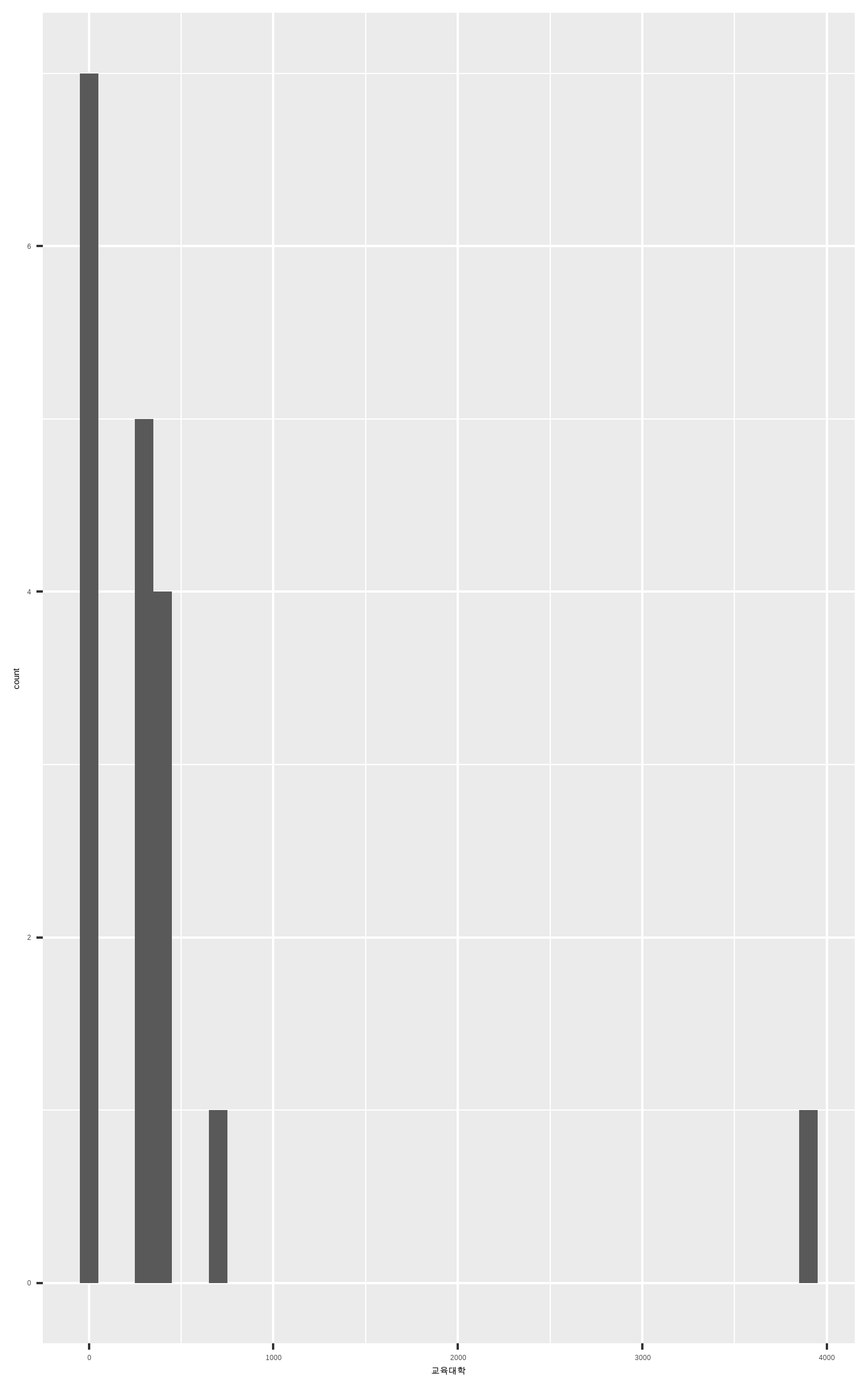
## df\_입학자에서 연도가 2021인 데이터만 추출하여 ggplot 객체를 생성하고 p\_histogram에 저장  
p\_histogram <- df\_입학자 |> filter(연도 == 2021) |>  
 ggplot()  
  
## p\_histogram에 geom\_histogram 레이어를 생성하는데 x축을 교육대학열로 매핑, binning 옵션을 주지 않았으므로 bins = 30이 기본값으로 설정됨   
p\_histogram +  
 geom\_histogram(aes(x = 교육대학))



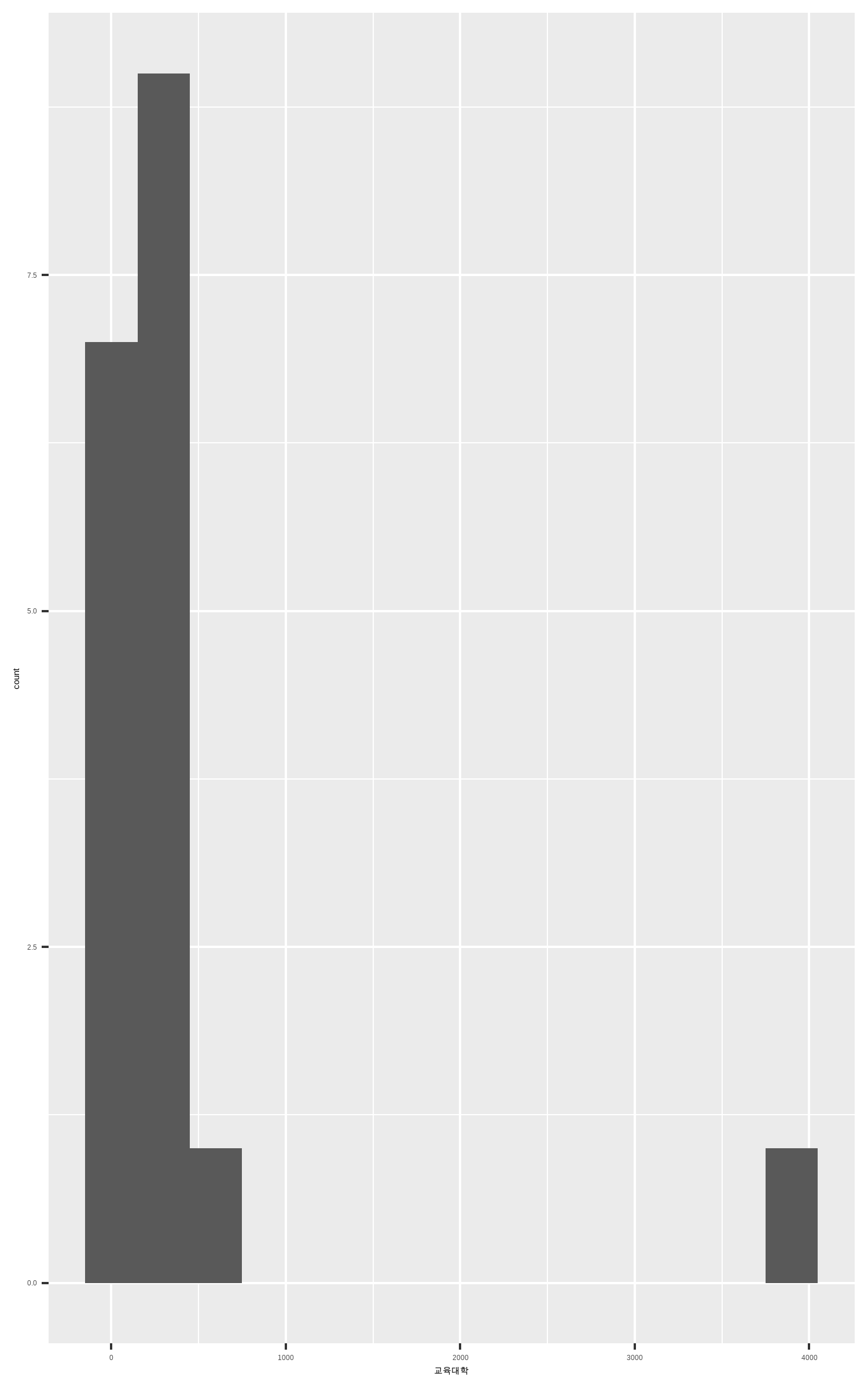
## p\_histogram에 geom\_histogram 레이어를 생성하는데 x축을 교육대학열로 매핑, bins = 90으로 설정   
p\_histogram +  
 geom\_histogram(aes(x = 교육대학), bins = 90)



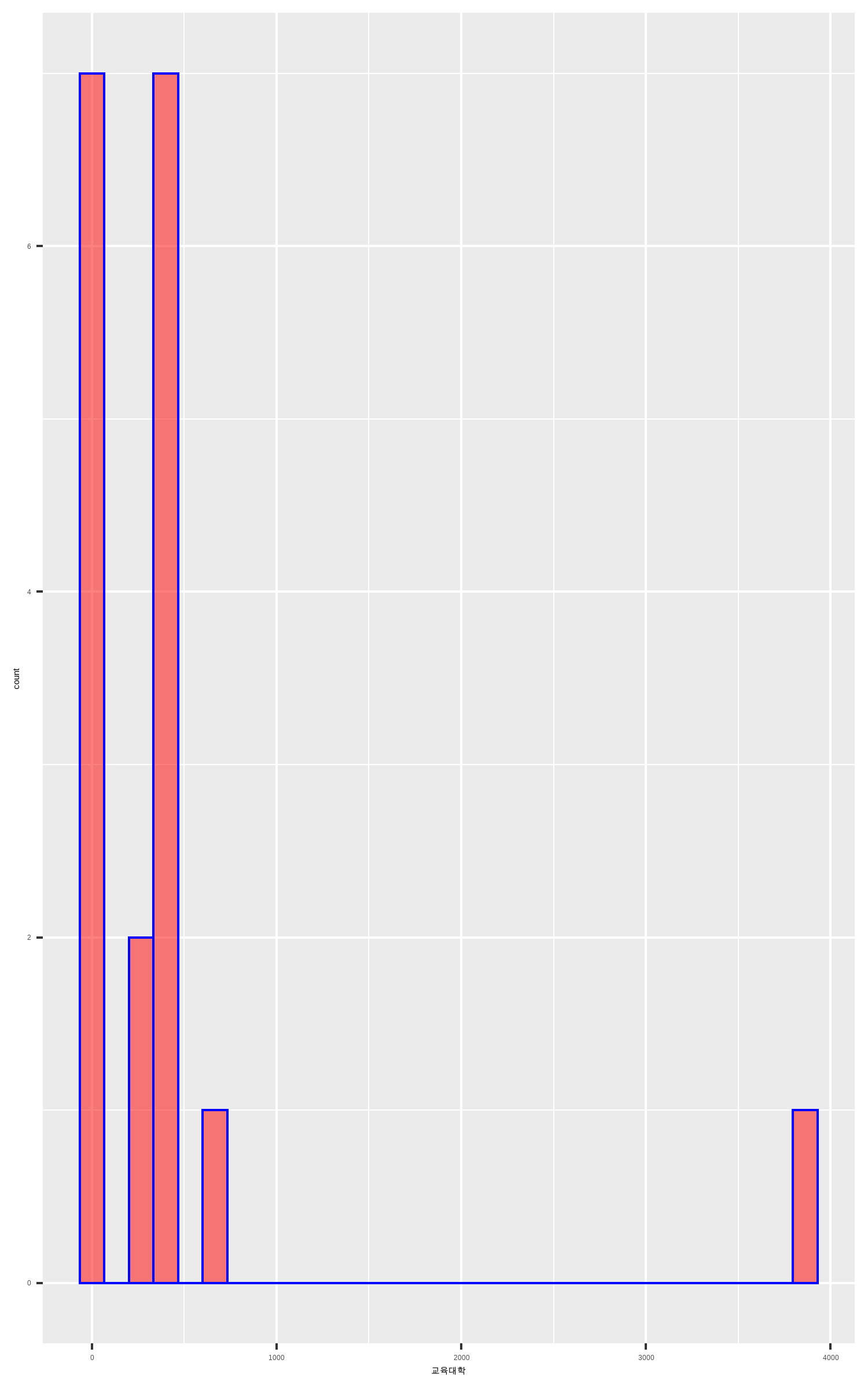
## p\_histogram에 geom\_histogram 레이어를 생성하는데 x축을 교육대학열로 매핑, binwidth = 100으로 설정   
p\_histogram +  
 geom\_histogram(aes(x = 교육대학), binwidth = 100)



## p\_histogram에 geom\_histogram 레이어를 생성하는데 x축을 교육대학열로 매핑, binwidth = 300으로 설정   
p\_histogram +  
 geom\_histogram(aes(x = 교육대학), binwidth = 300)



p\_histogram +   
 geom\_histogram(aes(x = 교육대학), color = 'blue', fill = 'red', alpha = 0.5)



##### 2.2.1.2