R마크다운으로 한글 저장

## 스파게티 선 그래프

데이터를 시각화하는 방법으로 많이 사용되는 그래프 중에 하나가 선 그래프이다. 선 그래프는 각각의 데이터 포인트들을 선으로 연결한 그래프이다. 일반적으로 데이터가 축 변수의 흐름에 따라 지속적으로 변화하는 추세를 나타내기 위해 사용하는 그래프로써 일반적으로 좌측에서 우측으로 데이터 포인트를 연결하면서 그려진다.

단변량 일때는 변수에 따른 추세(trend)를 나타내지만 다변량일 경우 추세뿐 아니라 데이터 간의 상관관계를 표현할 때도 사용될 수 있다. 보통의 경우 X축에 설정된 변수의 변화에 따른 Y축 데이터 값을 선으로 연결하여 그려지고, X축의 변수는 연속형(continuous) 변수와 이산형(discrete) 변수 모두 사용될수 있지만 비연속형 변수(discontinuous)는 사용하지 않는 것이 좋다. 예를 들어 이산형이지만 연속형 변수인 연도(2010, 2015, 2020 등)는 가능하지만 수치의 크기가 의미가 없는 모델 번호, 상품명 등의 변수는 좌측의 값과 우측의 값의 위치에 따른 의미가 없어 사용하지 않는 것이 좋다. Y축은 일반적으로 연속형 수치 변수가 사용된다.

앞서 언급한 바와 같이 선 그래프는 단변량과 다변량을 모두 표현한다. 다변량 선 그래프는 같은 X, Y좌표에서 변수의 분류에 따라 여러 개의 선을 동시에 표현하는 선 그래프를 말한다. 하지만 한 그래프에서 너무 많은 선이 표현되면 데이터의 흐름을 알아내기가 어렵다. 이와 같은 선 그래프를 스파게티(spaghetti) 선 그래프라고 한다. 하지만 여러 범주를 가진 변수에 대한 선 그래프를 그리기 위해서는 어쩔수 없이 스파게티 선 그래프를 그려야 할 때가 있다.

이번 포스트에서는 스파게티 선 그래프를 어떻게 효율적으로 그려야하는지를 알아보겠다.

### 데이터 Import

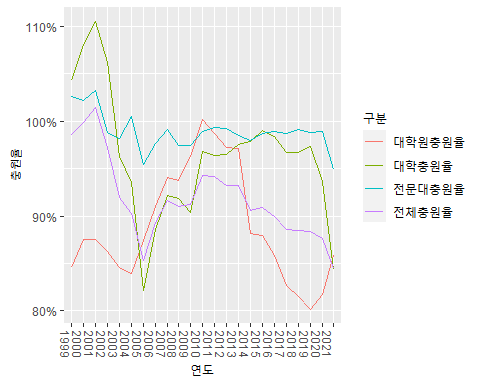
이번 포스트에서 사용하는 데이터는 [한국교육개발원 교육통계서비스 홈페이지](https://kess.kedi.re.kr)의 [고등교육기관의 연도별 신입생 충원율 데이터](https://kess.kedi.re.kr/kessTheme/zipyo?itemCode=03&uppCd1=030202&menuId=m_02_03_01#)를 활용하겠다.

library(readxl)  
library(tidyverse)  
  
df <- read\_excel('./연도별 신입생 충원율 현황.xlsx', skip = 2, na = '-', col\_names = F, col\_types = c(rep('numeric', 13)))  
  
## 파일에서 불러온 데이터 중에서 연도와 충원율에 해당하는 열만을 사용한다.   
df.충원율 <- df |> select(1, 4, 7, 10, 13)  
  
## 데이터프레임의 열 이름을 적절히 바꾸어준다.   
names(df.충원율) <- c('연도', '전체충원율', '전문대충원율', '대학충원율', '대학원충원율')  
  
## 전체 데이터 중 연도가 적절치 않은 데이터가 있다. 적절한 데이터만 남긴다.  
df.충원율 <- df.충원율 |>  
 filter(연도 >= 1999)

### 선 그래프 그리기

전처리된 데이터를 사용하여 선 그래프를 그려본다. 그리기 전에 선 그래프를 그리기에 적당하게 넓은 형태의 데이터프레임을 긴 형태의 데이터프레임으로 바꾸어 준다. 이렇게 변환된 데이터를 사용하여 ‘구분’ 열의 4가지 범주에 따른 다변량 선 그래프는 다음과 같이 생성할 수 있다.

df.충원율 <- pivot\_longer(df.충원율, 2:5, names\_to = '구분')  
  
df.충원율 |>   
 ggplot(aes(x = as.factor(연도), y = value)) +  
 geom\_line(aes(group = 구분, color = 구분)) +  
 labs(x = '연도', y = '충원율') +  
 scale\_y\_continuous(labels = scales::number\_format(suffix = '%')) +  
 theme(axis.text.x = element\_text(angle = -90))



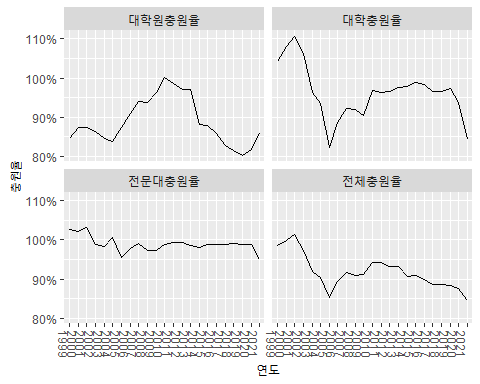
위의 그래프를 보면 총 4개의 선이 그려지는데 이를 구분하기 위해 선의 색깔을 달리하였다. 하지만 구분이 잘 되지 않아 보인다. 이렇게 복잡하게 그려진 다변량 선 그래프가 스파게티 선 그래프이다. 일반적으로 스파게티 선 그래프는 효율적이지 않은 시각화 표현 방법이라고 알려져 있다.

이를 해결하는 방법으로 세가지 방법을 사용해보겠다.

### 1. 분할(facet)을 사용하는 방법

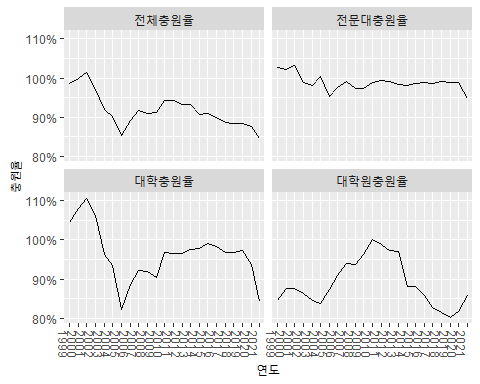
하나의 그래프에 여러개의 범주를 가지는 변수를 표현할 때 많이 사용하는 방법이 각각의 범주별로 분할하여 그래프를 표현하는 방법이다. ggplot2에서는 facet\_wrap()과 facet\_grid()를 사용하여 그래프를 변수의 범주별로 분할한다.

df.충원율 |>   
 ggplot(aes(x = as.factor(연도), y = value)) +  
 geom\_line(aes(group = 구분)) +  
 labs(x = '연도', y = '충원율') +  
 scale\_y\_continuous(labels = scales::number\_format(suffix = '%')) +  
 theme(axis.text.x = element\_text(angle = -90)) +  
 facet\_wrap(~구분)



일단 분할 그래프의 순서가 별로 합리적으로 보이지 않는다. fct\_reorder를 사용하여 순서를 설정해보자.

df.충원율$구분 <- fct\_relevel(df.충원율$구분, c('전체충원율', '전문대충원율', '대학충원율', '대학원충원율') )  
  
df.충원율 |>   
 ggplot(aes(x = as.factor(연도), y = value)) +  
 geom\_line(aes(group = 구분)) +  
 labs(x = '연도', y = '충원율') +  
 scale\_y\_continuous(labels = scales::number\_format(suffix = '%')) +  
 theme(axis.text.x = element\_text(angle = -90)) +  
 facet\_wrap(~구분)

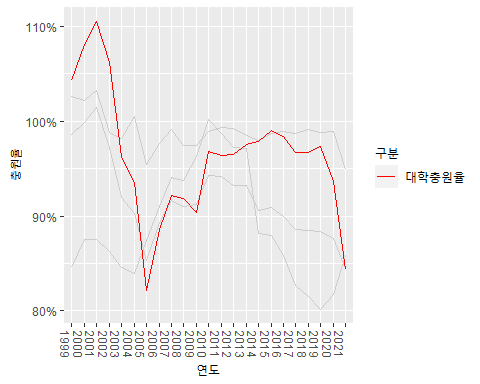


### 2. 특정 범주만 강조하는 방법

보통 스파게티 선 그래프를 그리는 이유는 동일 변수의 여러 범주에 대비하여 특정 범주의 흐름을 파악하는 경우가 많다. 이 경우 강조하고 싶은 범주를 강조하고 나머지 비교 대상이 되는 범주를 흐릿하게 표현하면 효과적으로 스파게티 선 그래프를 활용할 수 있다. 여기에서는 대학 충원율을 강조해보도록 한다.

우선 흐릿한 회색(grey80) 선으로 전체 스파게티 선 그래프를 그린다. 이후 강조하고자 하는 데이터만 필터링 한데 데이터를 사용하여 강조색을 사용한 선 그래프를 그린다.

df.충원율 |>   
 ggplot(aes(x = as.factor(연도), y = value)) +  
 geom\_line(aes(group = 구분), color = 'grey80') +  
 geom\_line(data = df.충원율 |> filter(구분 == '대학충원율'), aes(group = '구분', color = 구분)) +  
 labs(x = '연도', y = '충원율') +  
 scale\_y\_continuous(labels = scales::number\_format(suffix = '%')) +  
 theme(axis.text.x = element\_text(angle = -90)) +  
 scale\_color\_manual(values = c('대학충원율' = 'red'))



### 3. facet과 강조를 모두 사용

강조를 사용한 스파게티 선 그래프는 변수의 전체 범주에서 특정 범주를 확인하는데 적절하다. 하지만 한번에 하나의 강조만을 사용할 수 밖에 없다는 단점이 있는데 모든 범주에 대해 각각의 강조 선 그래프를 그리려면 어떻게 해야할까?

간단하게 생각하면 강조 그래프를 4개 그린다음 patchwork패키지의 기능을 사용하여 네개의 선 그래프를 하나로 붙이면 가능할 것이다.

하지만 이 포스트에서는 facet\_wrap()과 강조 방법을 사용하여 표현해보겠다.

이를 위해서는 먼저 임시로 사용될 데이터프레임이 하나 필요하다. 이 데이터프레임은 원본 데이터 프레임과 동일한 구조를 지니지만 하나 다른점은 facet\_wrap()에서 사용되는 분할 변수 열 이름을 달리 저장한다는 점이다. 다음과 같이 생성한다.

df.충원율.temp <- df.충원율 |> mutate(구분1 = 구분) |> select(-구분)  
  
head(df.충원율.temp)

## # A tibble: 6 x 3  
## 연도 value 구분1   
## <dbl> <dbl> <fct>   
## 1 1999 98.6 전체충원율   
## 2 1999 103. 전문대충원율  
## 3 1999 104. 대학충원율   
## 4 1999 84.6 대학원충원율  
## 5 2000 99.9 전체충원율   
## 6 2000 102. 전문대충원율

원리는 facet\_wrap()의 분할 변수를 통해 그래프를 분할 시키고 동일한 범주를 가졌지만 열 이름이 다른 데이터로 흐릿한 선들을 그려준다. 흐릿한 선들은 facet\_wrap()에 지정되는 분할 열과 동일한 열이름이 아니기 때문에 분할되어 그려지지 않고 전체 분할 그래프에 공통적으로 그려지게 된다.

df.충원율 |>   
 ggplot(aes(x = as.factor(연도), y = value)) +  
 geom\_line(data = df.충원율.temp, aes(group = 구분1), color = 'grey75') +  
 geom\_line(aes(group = 구분)) +  
 labs(x = '연도', y = '충원율') +  
 scale\_y\_continuous(labels = scales::number\_format(suffix = '%')) +  
 theme(axis.text.x = element\_text(angle = -90)) +  
 facet\_wrap(~구분)

