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.

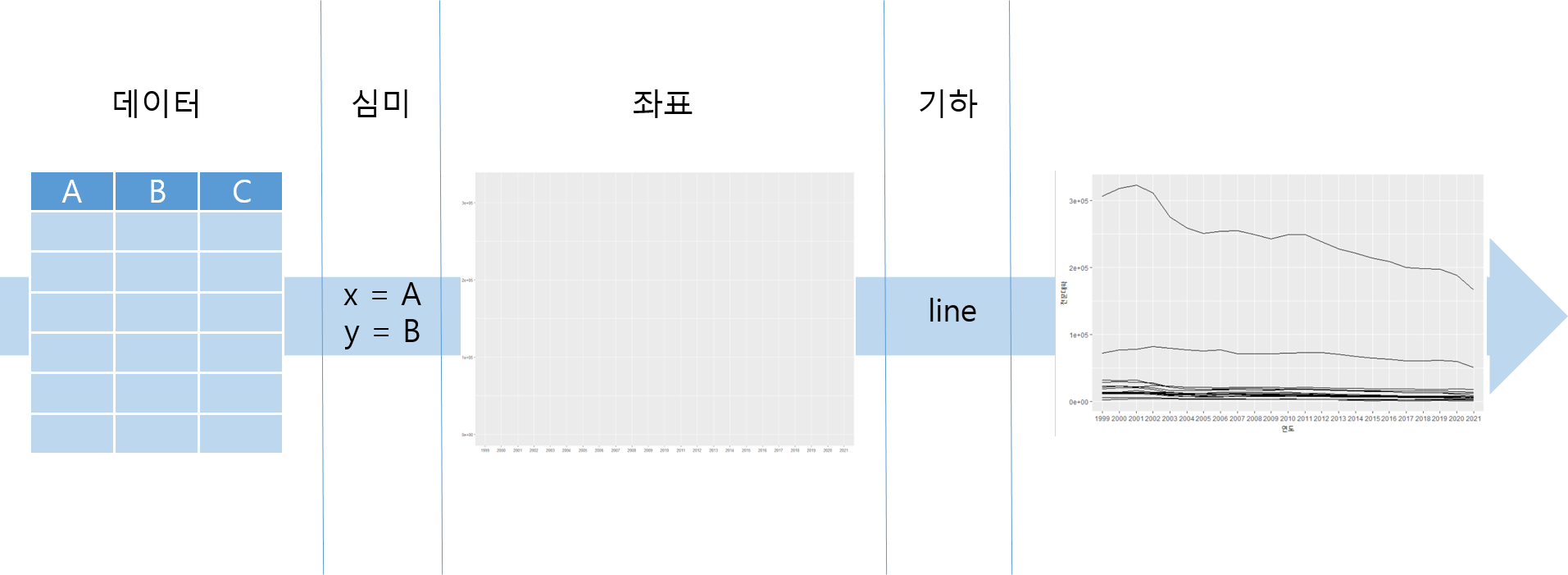
* data(데이터)
* 시각화에서 표현해야할 데이터를 지정한다. 하나의 ggplot2 시각화에는 하나 이상의 데이터를 사용할 수 있는데 최소 하나 이상의 데이터가 필수적으로 포함되어야 한다. ggplot2에서 지원하는 데이터 타입은 데이터프레임이나 tibble이다. 내부적으로 데이터프레임은 tibble로 변환되어 사용된다.
* Aesthetics(심미요소)
* Aesthetics는 데이터 시각와에서 시각적 속성과 해당 시각적 속성을 연결시키는 매핑 정보를 표현한다. Aesthetics로 매핑될 수 있는 시각적 속성은 X, Y축, 색깔, 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.

