

Big Data Analytics Programming

Week-11. Time Series Analysis

Jungwon Seo, 2020-Fall

Time Series Analysis

Overview

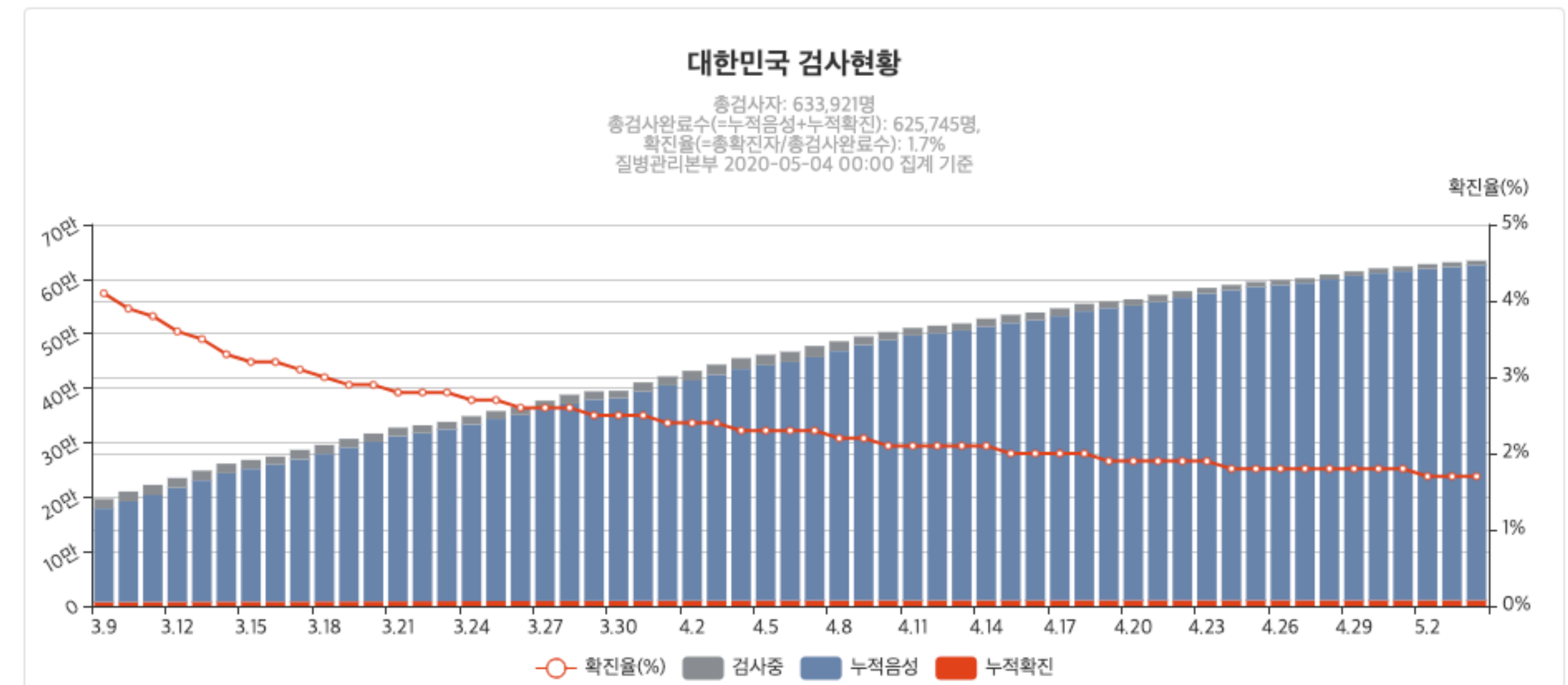
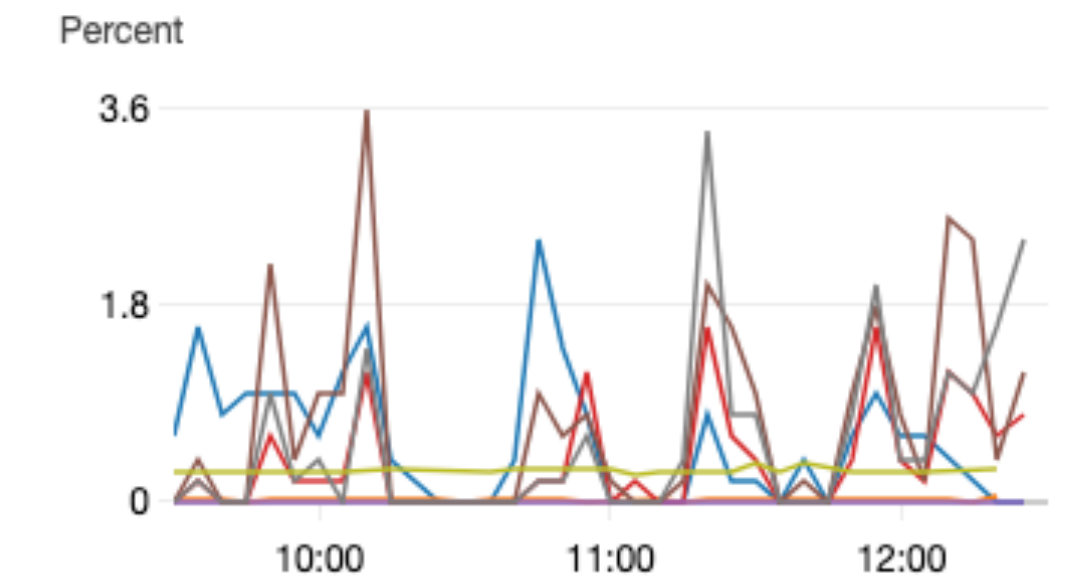
- 시계열 데이터의 정의
- 시계열 데이터의 시각화
- 시계열 데이터 분석의 응용
 - Forecasting
 - Anomaly Detection

Time Series

시계열 데이터의 정의와 목적

- 일정 시간 간격으로 배치된 데이터들의 수열
- 어떠한 규칙에 의해 시계열 데이터가 생성 되는지를 아는 것이 목표
- 지금까지 배워온 데이터는
 - X와 y가 존재하거나 X만 존재해왔다면
 - 시계열 데이터의 경우 y만 존재
- 예제
 - 주식가격, 신종 코로나 확진자, 각종 장비 로그데이터

CPU Utilization Average

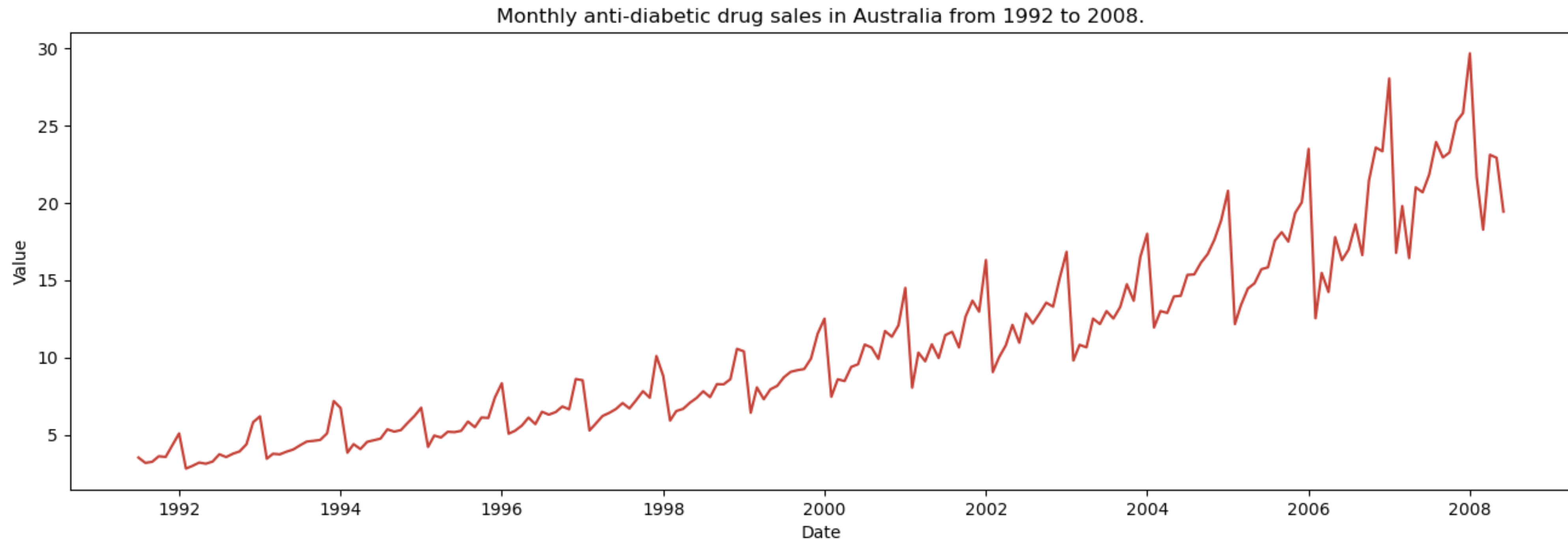


Visualization

시계열 데이터 시각화

Line Chart

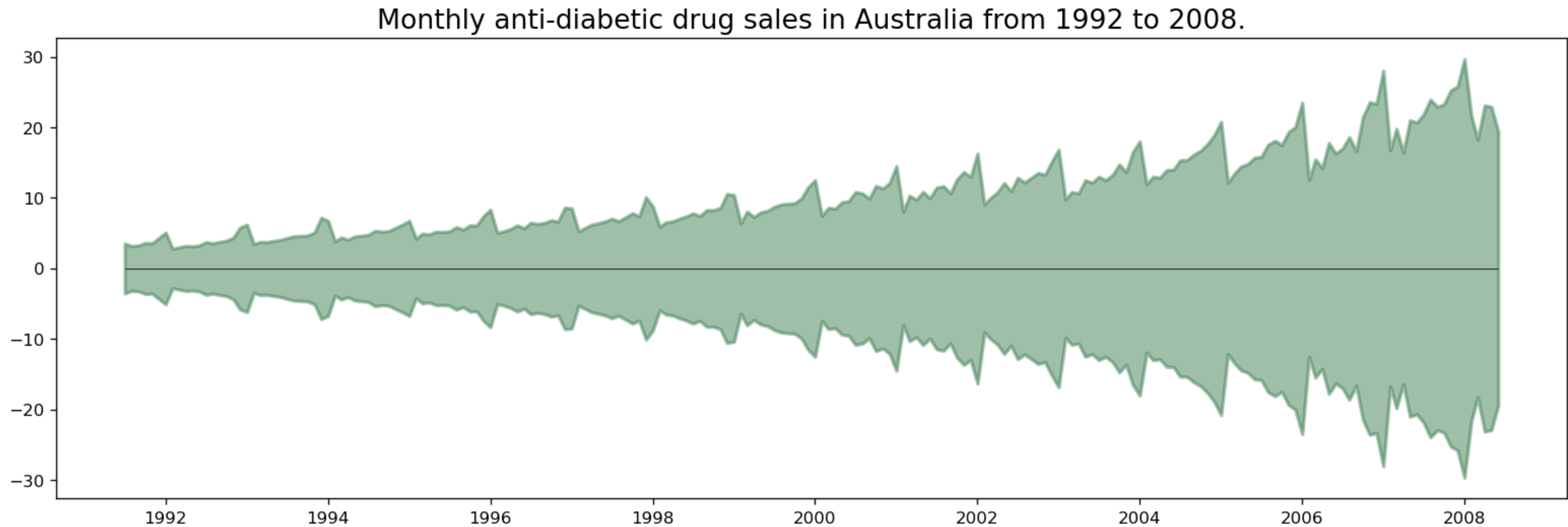
- 기본적으로 line chart로 표현 가능



시계열 데이터 시각화

Two-sided view

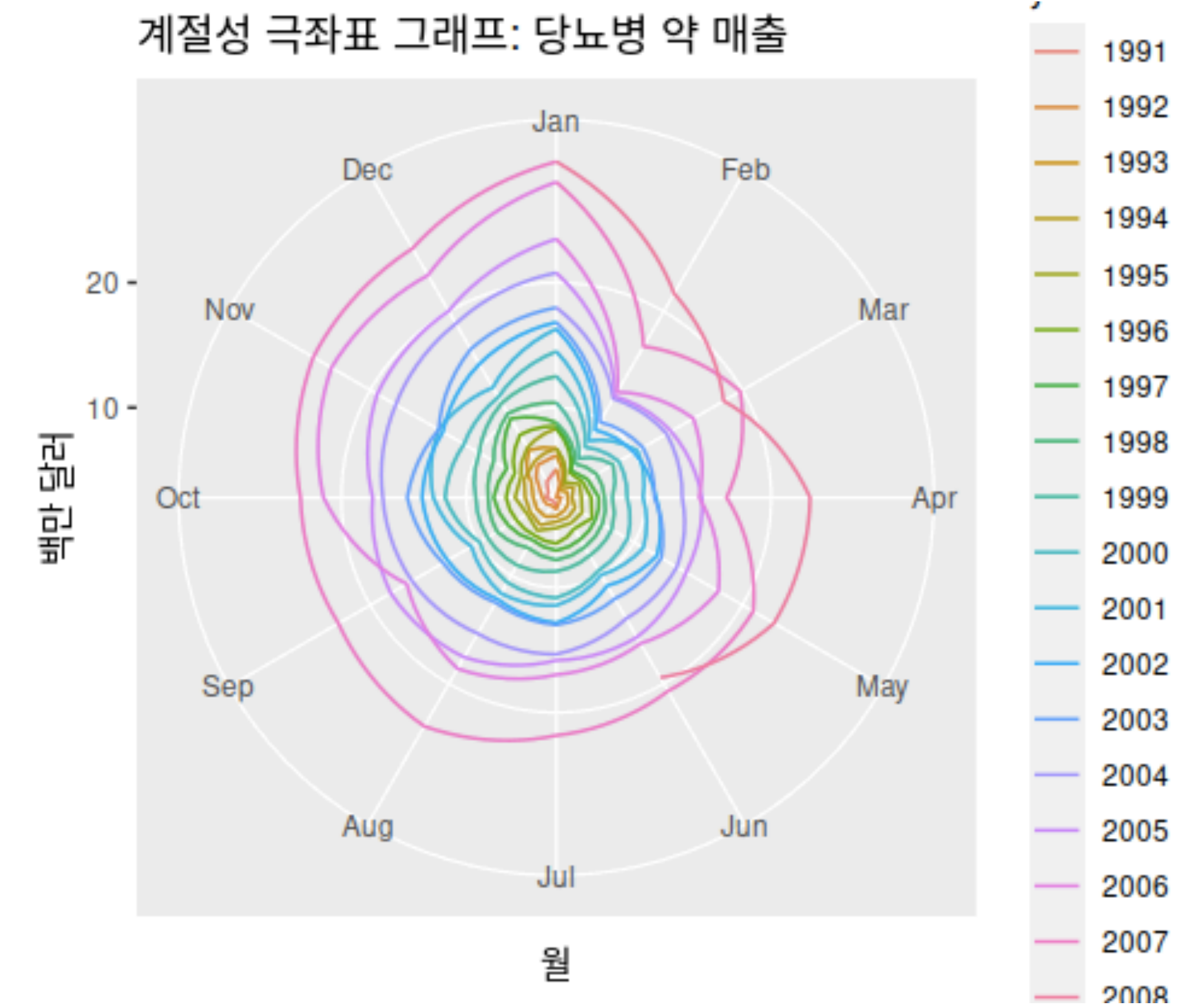
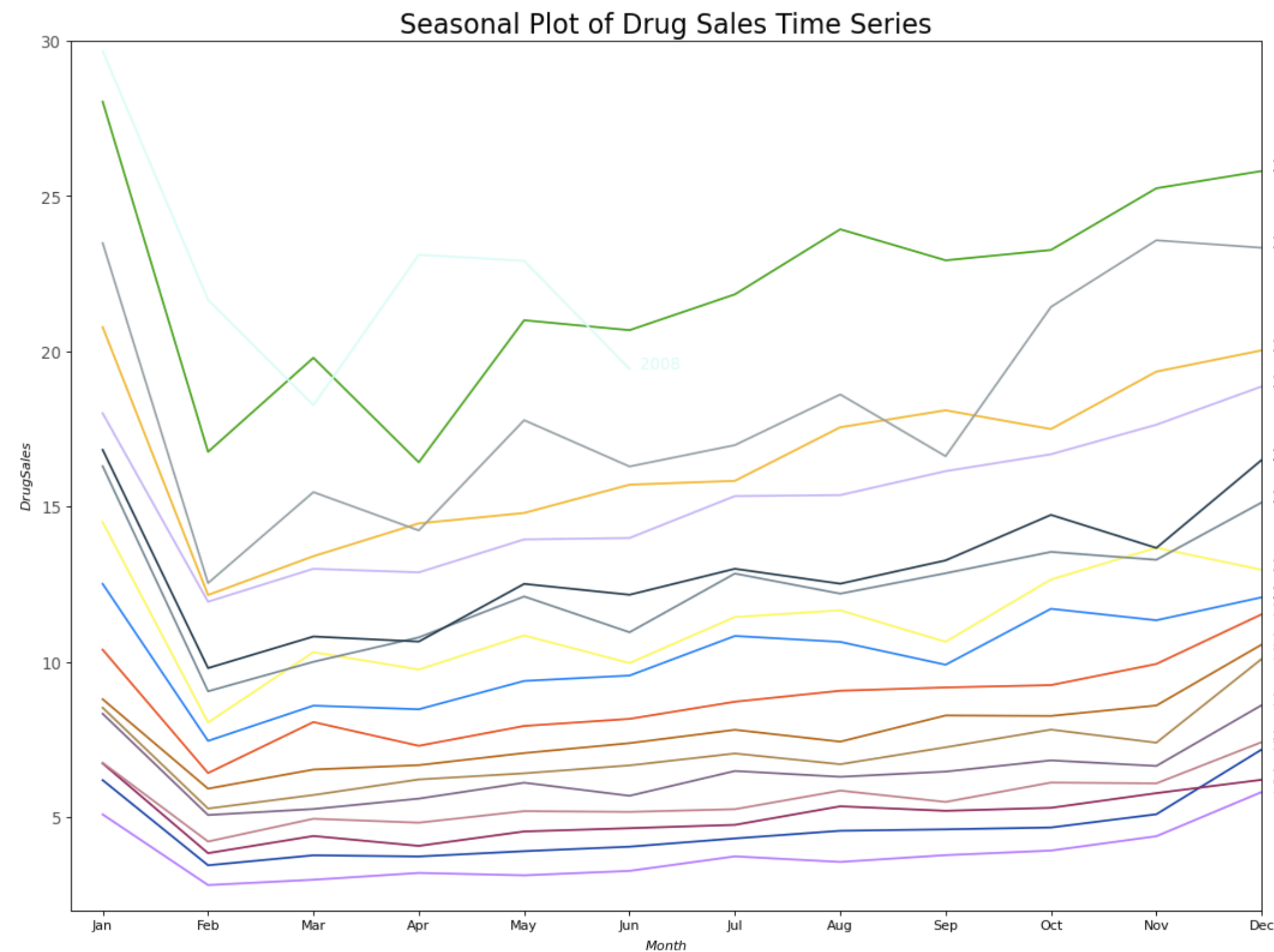
- 모든 데이터가 양수이고, y값의 증가를 강조하고 싶다면?



시계열 데이터 시각화

Seasonal Graph

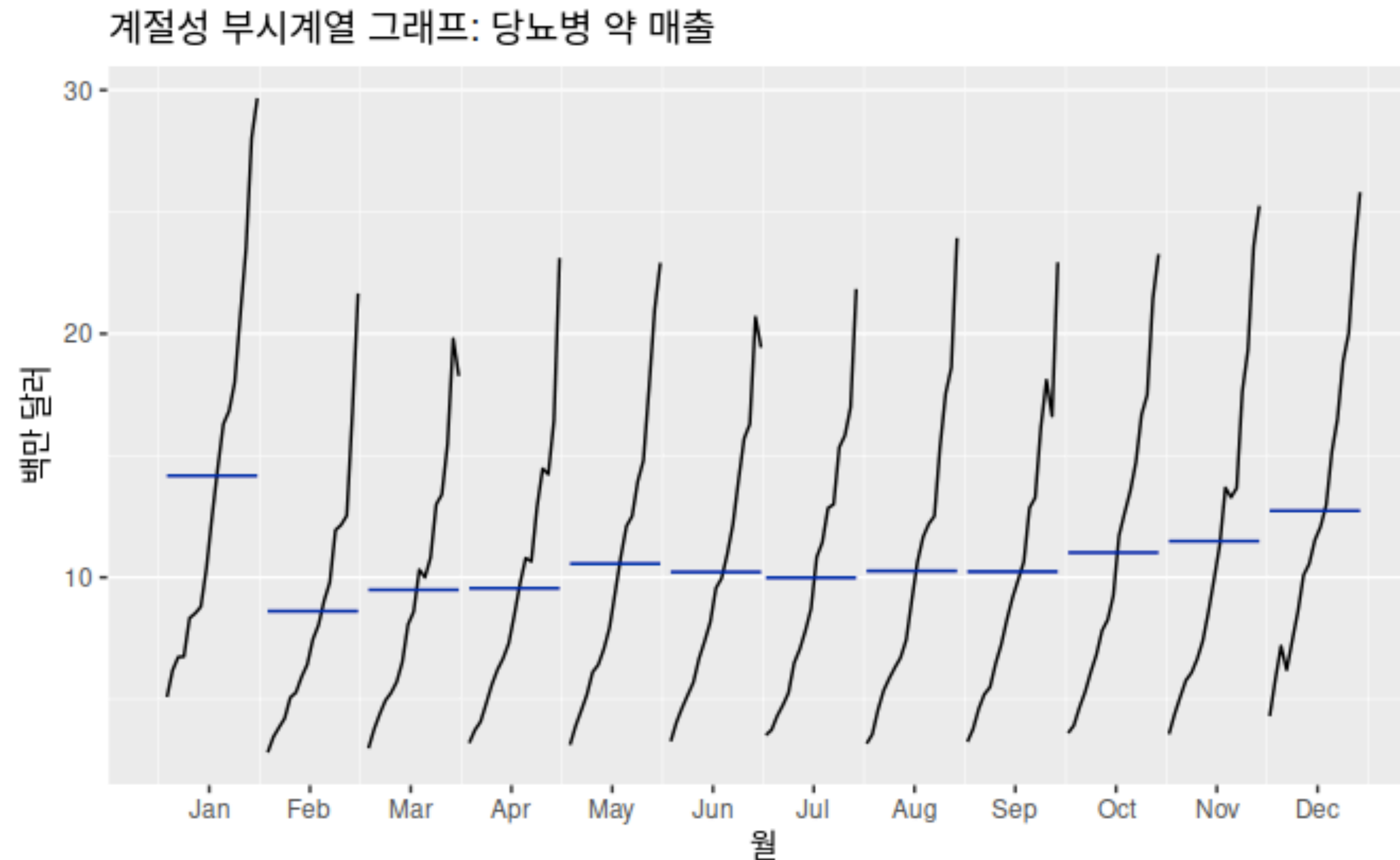
- 매달 또는 매년 존재하는 계절성을 비교하고 표현하고 싶다면?



시계열 데이터 시각화

Seasonal Graph

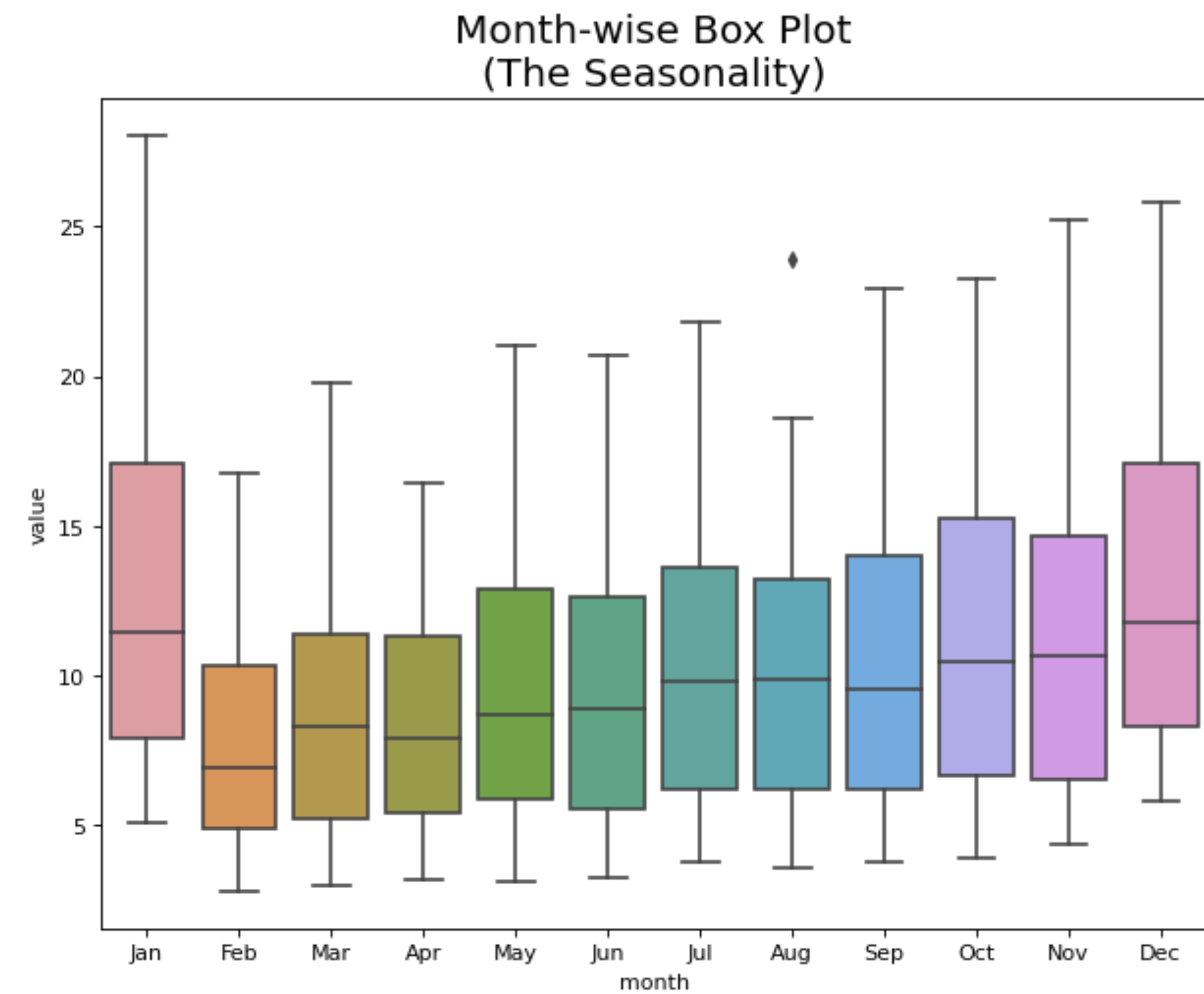
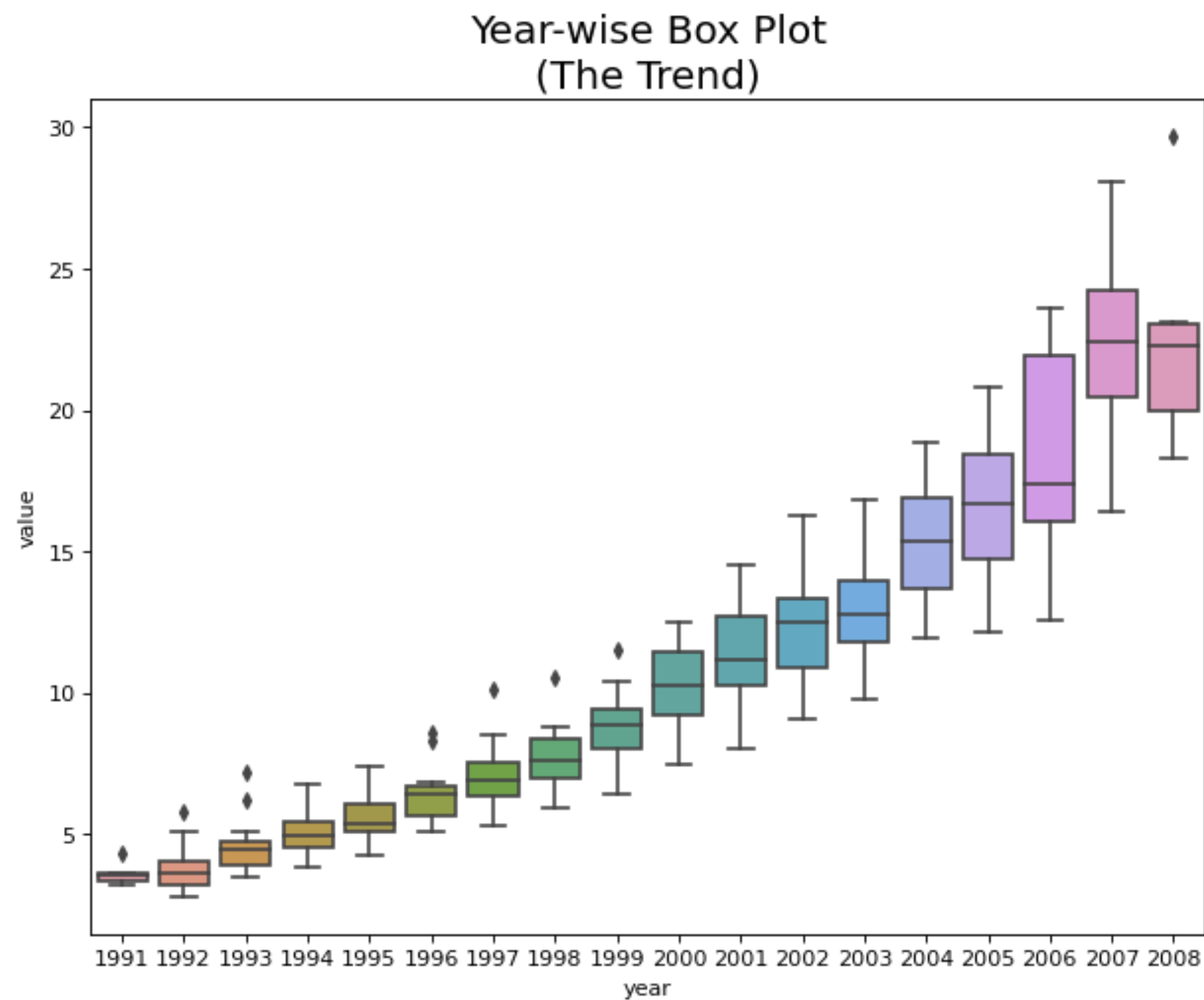
- 계절성 무시계열 그래프



시계열 데이터 시각화

Time Series Box Plot

- 통계적인 데이터 값을 포함하여 표현하고 싶다면,

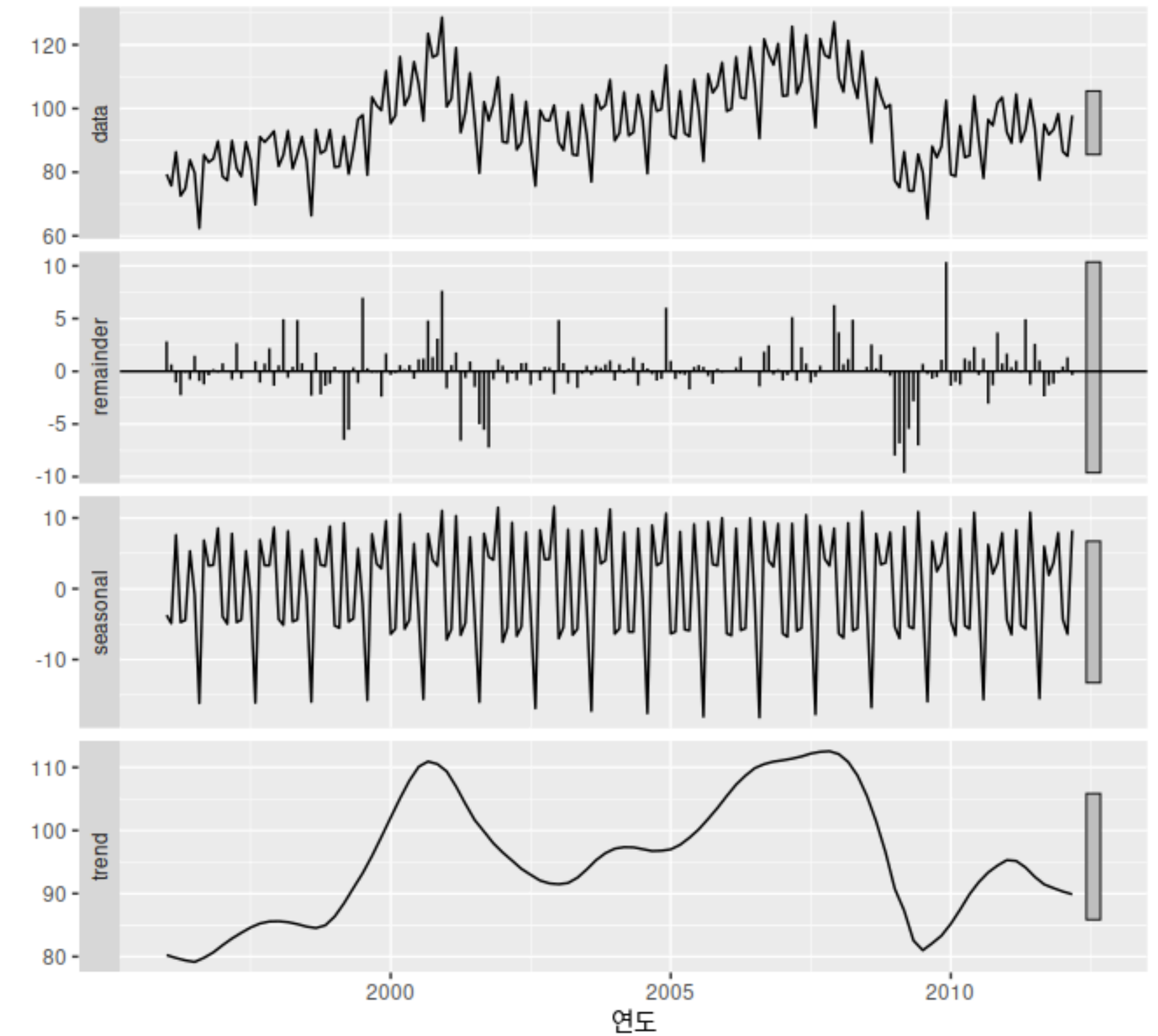


Time Series Decomposition

Time Series Decomposition

시계열 데이터의 구성 성분

- 일반적으로 시계열 데이터는
 - 추세-주기 성분, 계절성 성분, 나머지 성분으로 구성된다고 가정
- 덧셈분해 (Additive Decomposition)
 - $y_t = S_t + T_t + R_t$
 - S: Seasonal, T: Trend, R: Remainder
 - 추세-주거나 계절성의 변동이 시간에 의해 변하지 않는다면, 덧셈 분해가 적합
- 곱셈분해(Multiplicative Decomposition)
 - $y_t = S_t \times T_t \times R_t$
 - 시간에 의해, S나 T가 영향을 받는다면, 곱셈 분해가 적합
 - 영향을 받는다 : 예를들면 시간이 지날 수록 계절성의 변동성이 작아지거나 커진다



Time Series Decomposition

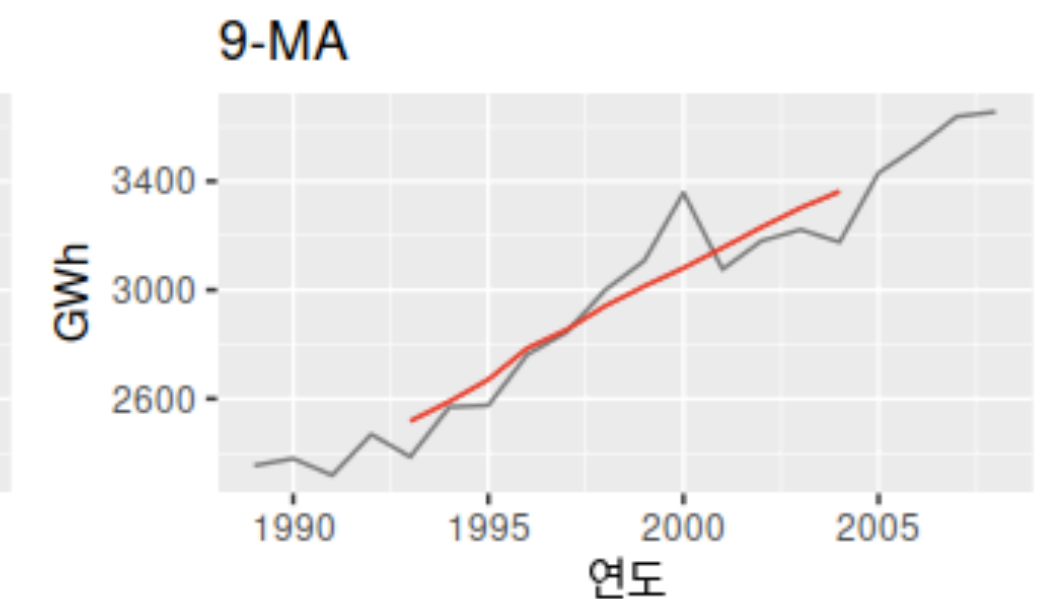
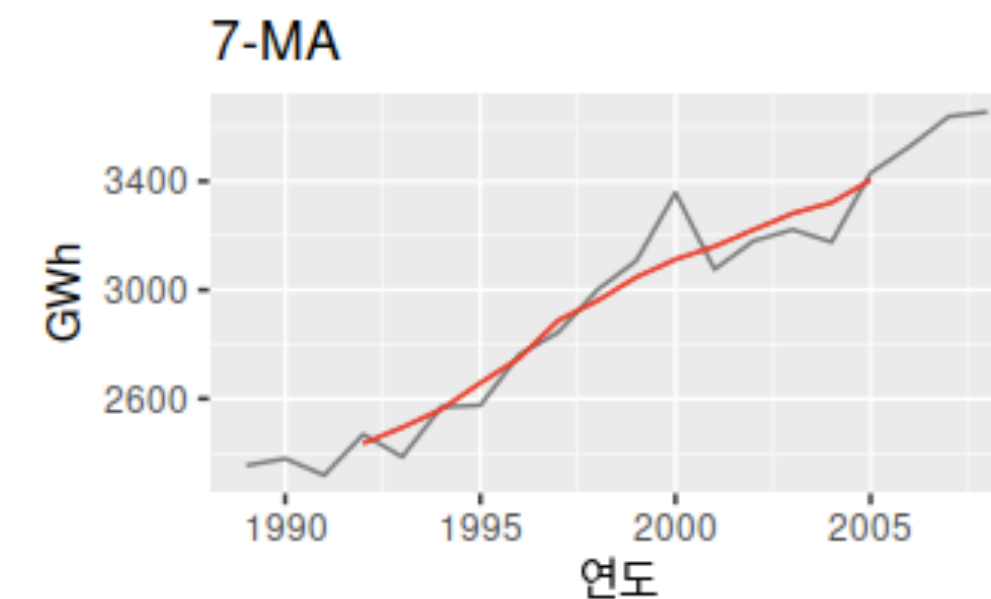
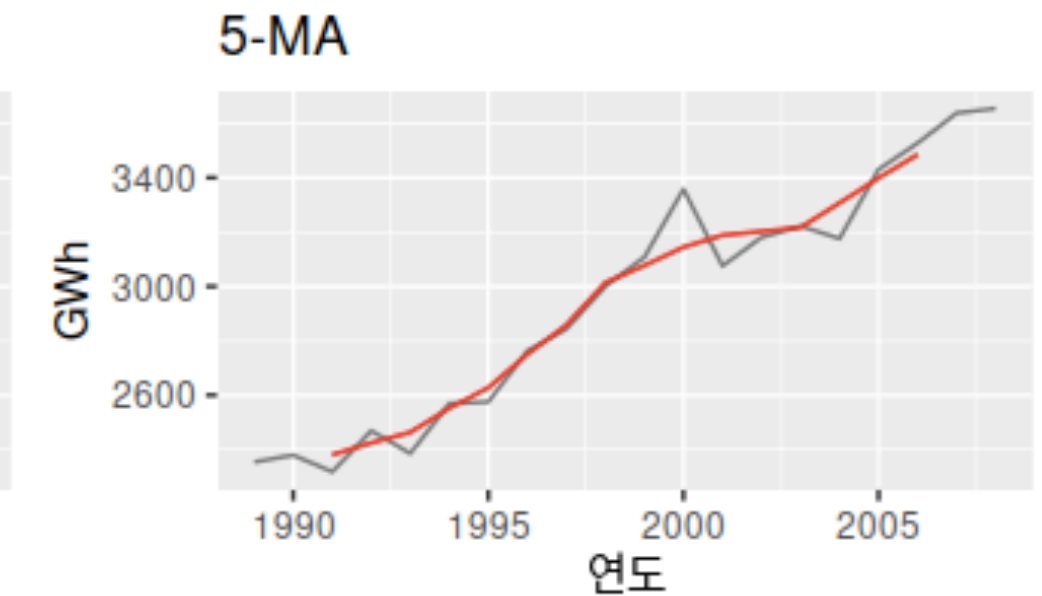
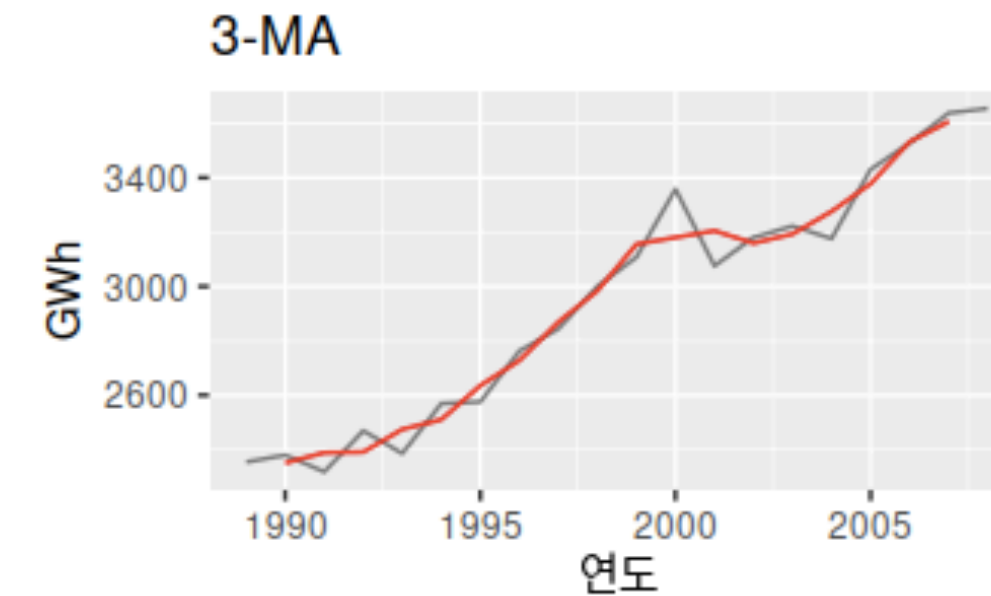
이동평균평활

- Moving Average Smoothing

- 추세-주기를 측정하기 위해
- 차수(order) m의 이동평균 방정식

- $$\hat{T}_t = \frac{1}{m} \sum_{j=-k}^k y_{t+j}$$

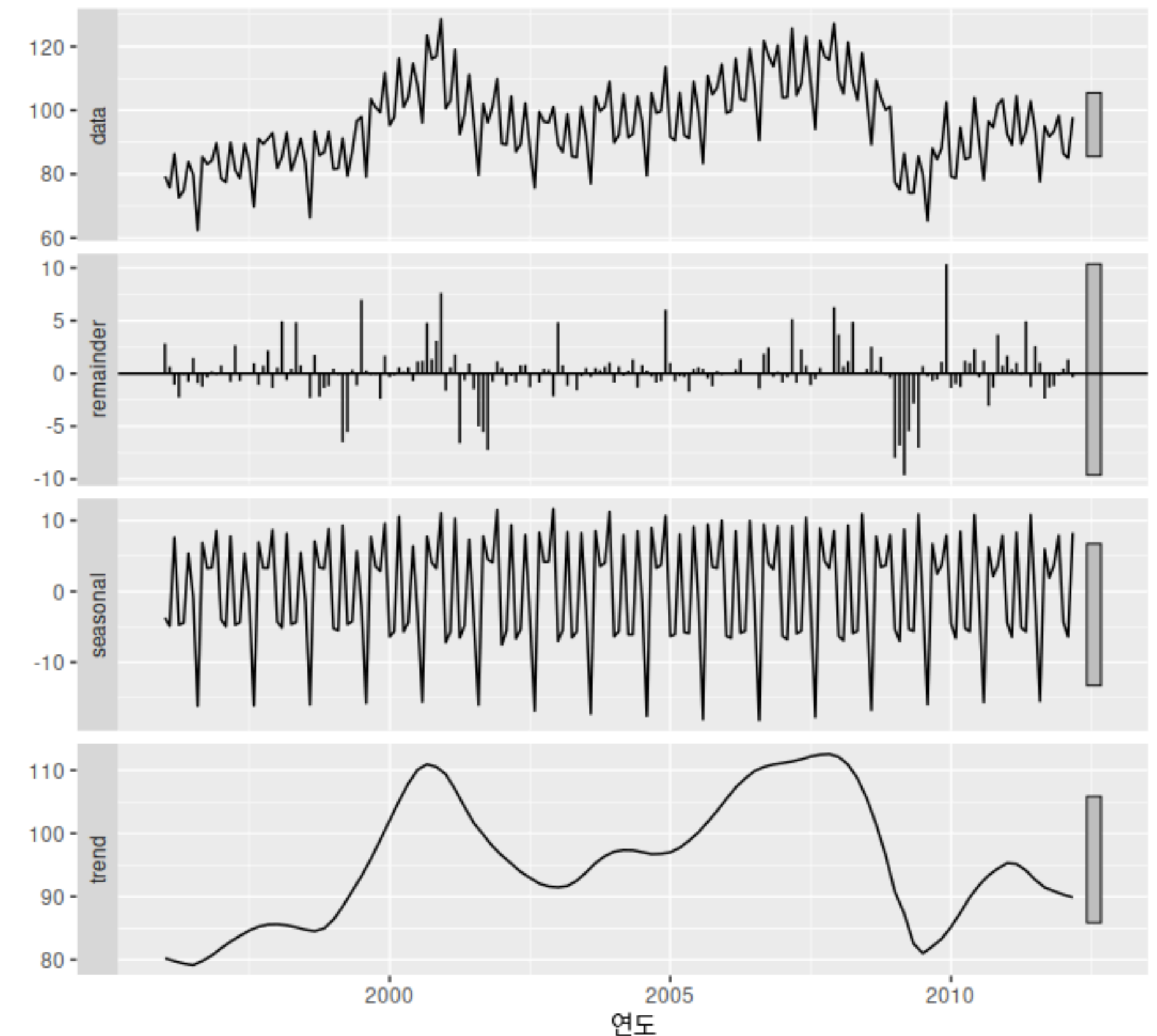
- $m = 2k + 1$
- k기간 안의 시계열 값을 평균하여 시간 t의 추세-주기를 측정
- 주식에서 골든 크로스
 - 단기 이동 평균선이 장기 이동평균선을 뚫는 지점



Time Series Decomposition

곱셈, 덧셈 분해 (고전적인 방식)

- 계절성 지수(seasonal indices) m 결정
 - 분기별 $m=4$, 월별 $m=12$, 주별 패턴이 있는 일별 데이터 $m=7$
 - 1단계
 - m 이 짝수(even)이면 $2 \times m$ -MA, m 이 홀수(odd)이면 m -MA를 사용하여, 추세-주기 성분을 계산 \hat{T}_t
 - 2단계: 추세를 제거한 시계열 계산
 - 덧셈분해: $y_t - \hat{T}_t$, 곱셈분해: y_t / \hat{T}_t
 - 3단계: 계절성분 제거
 - 10년치의 월별 ($m=12$) 데이터라고 가정 했을때, (추세가 제거된 상태에서), 모든 3월의 시계열 값의 평균을 구한다음 각각의 3월에서 평균을 제거, 나머지 월에 대해서도 마찬가지로 적용, 이렇게 평균이 제거된 시계열을 나열하면 \hat{S}_t
 - 4단계: Remainder 계산
 - 덧셈분해: $\hat{R}_t = y_t - \hat{T}_t - \hat{S}_t$, 곱셈분해: $\hat{R}_t = y_t / \hat{T}_t \hat{S}_t$
- 고전적인 방법의 한계
 - 이동평균 평활의 계산 방식상 처음 m 개 또는 마지막 m 개에 대한 데이터에 대한 추세 추정값 X
 - 과도한 smoothing
 - Outlier를 다루는데 적합하지 X



Time Series Decomposition

STL Decomposition

- Seasonal and Trend decomposition using Loess
 - 1990년도에 제안 (by R. B. Cleveland, Cleveland, McRae, & Terpenning)
 - Loess : Local regression, 비선형관계를 추정하기 위한 기법
 - 월별/분기별 등 어떤 종류의 계절 성도 다를 수 있음
 - 계절적인 성분이 시간에 따라 변해도 OK (1950년도 부터 전기 사용량?)
 - 추세-주기의 smoothing 정도를 사용자가 조절 가능
 - 두개의 파라미터를 선택
 - Trend-cycle window, Seasonal window

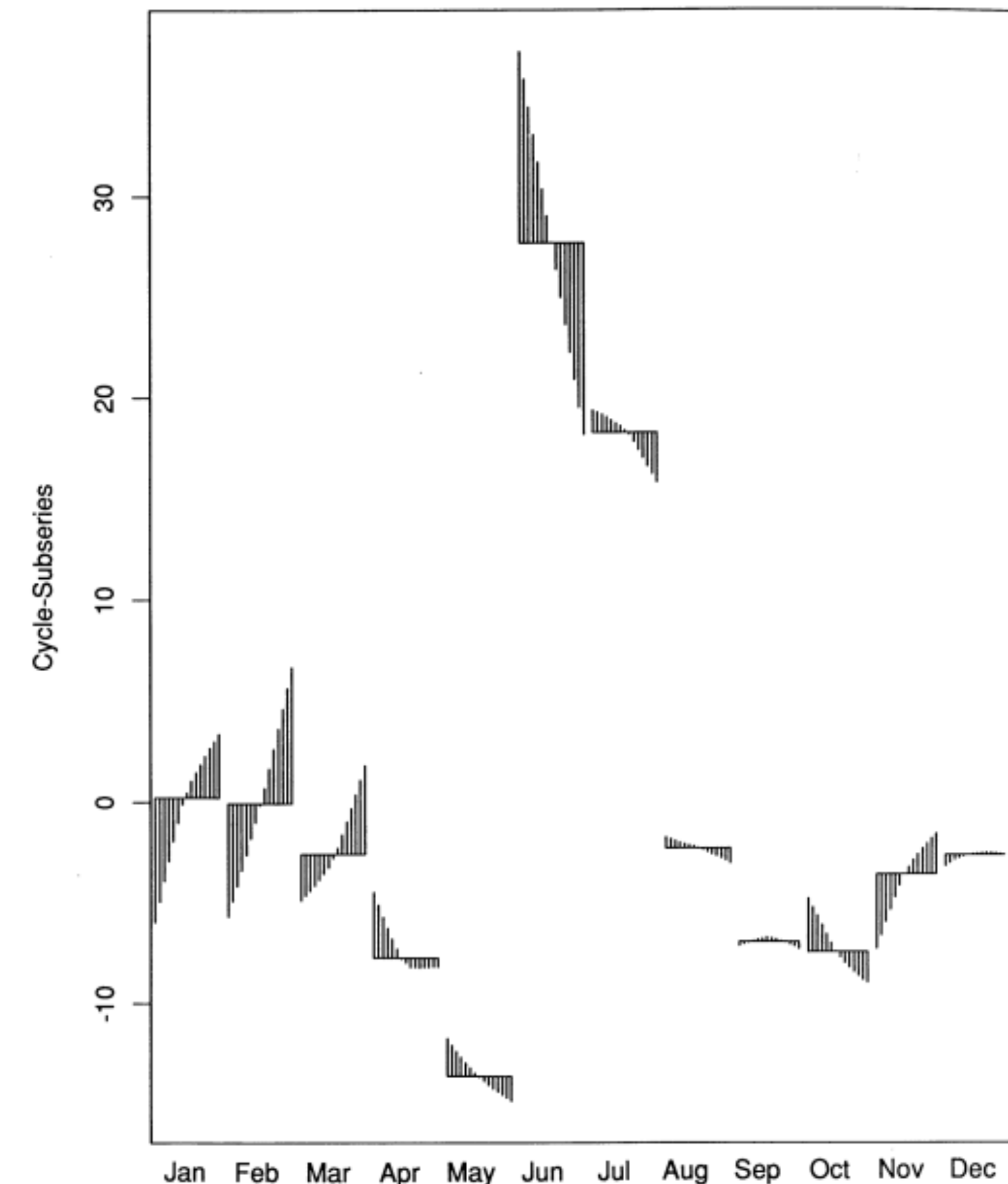


Fig. 6. Cycle-Subseries Plot for U.S. Unemployed Males Ages 16-19. The units on the vertical scale are tens of thousands.

Forecasting

Forecasting

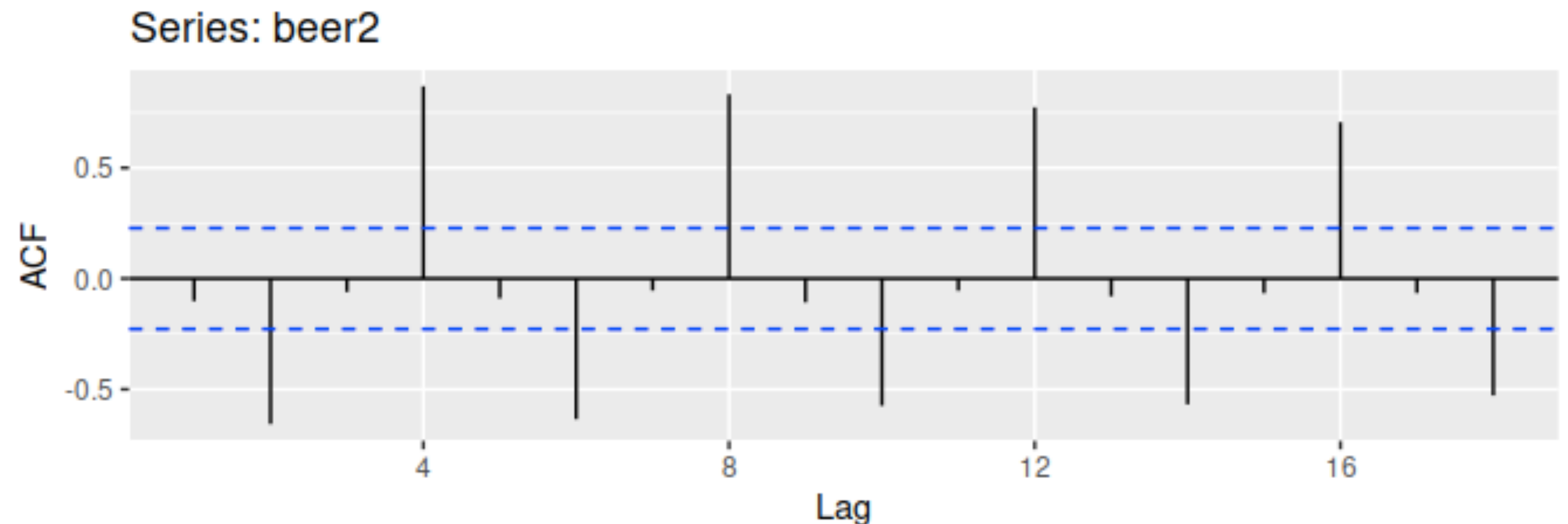
Auto Correlation

- 자기상관

- 시계열의 시차 값(lagged values) 사이의 선형 관계
- r1은 yt와 yt-1 사이의 관계, r2는 yt와 yt-2까지의 관계

- $$r_k = \frac{\sum_{t=k+1}^T (y_t - \hat{y})(y_{t-k} - \hat{y})}{\sum_{t=1}^T (y_t - \hat{y})^2}$$

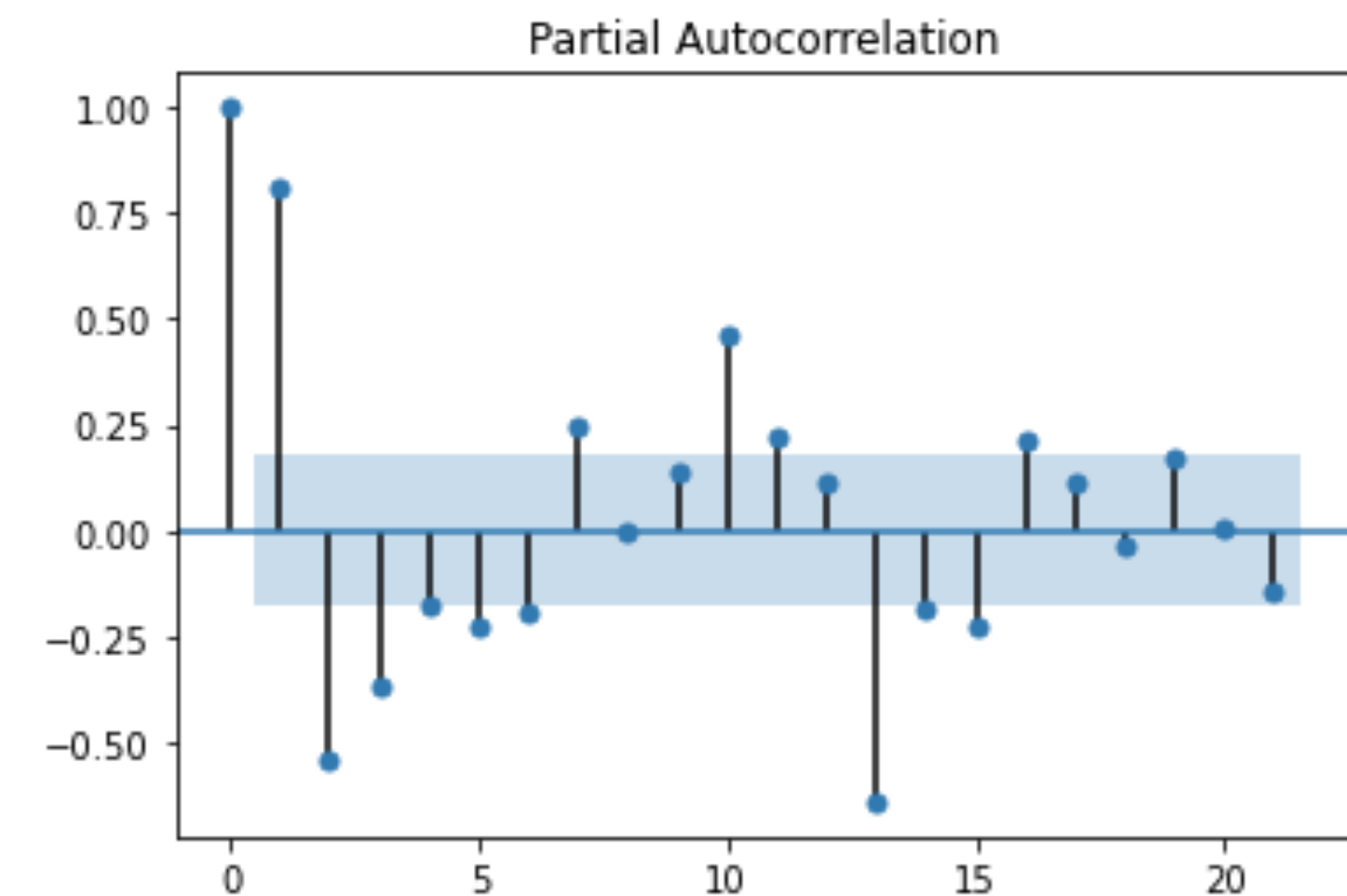
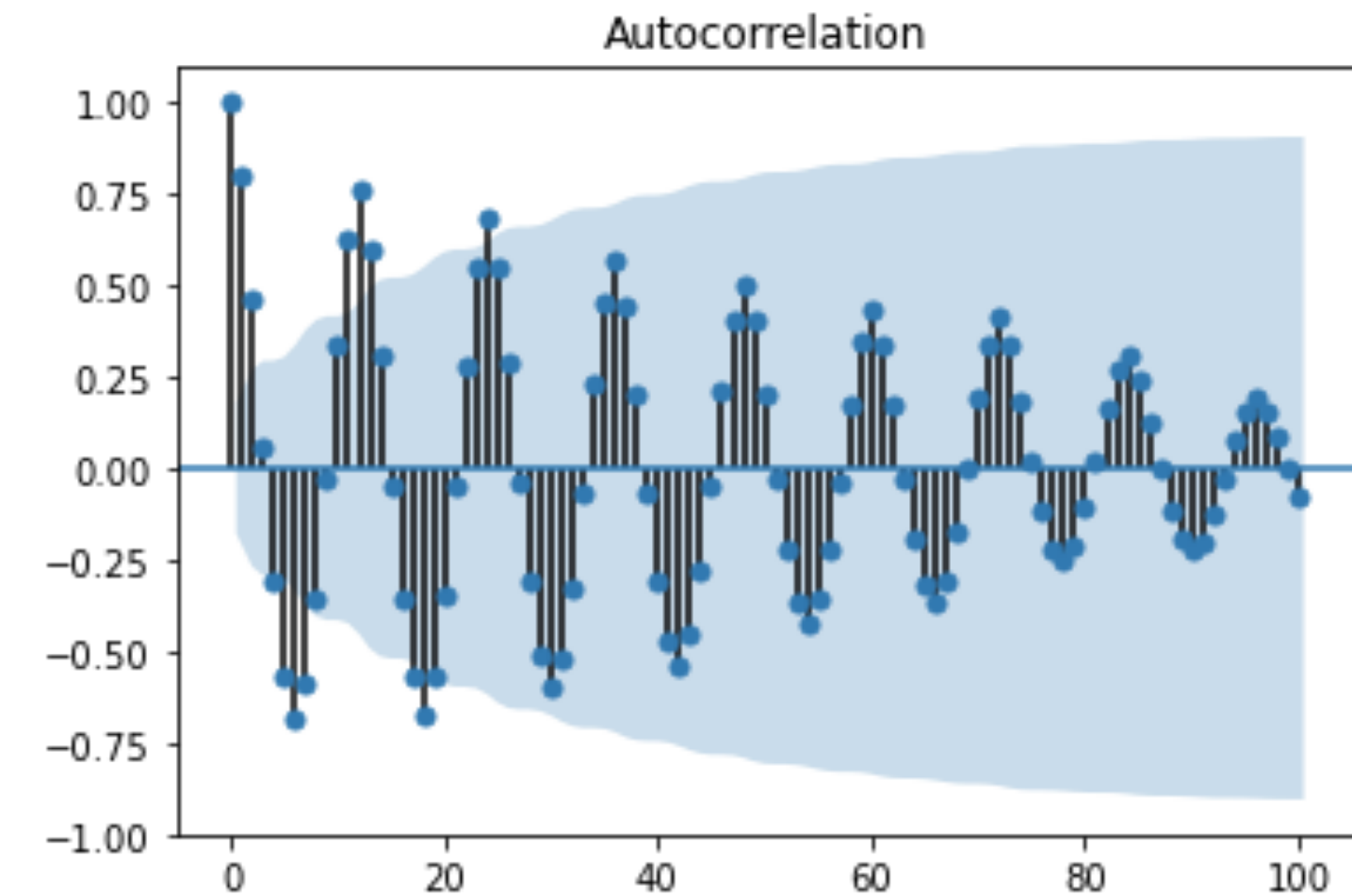
r1	r2	r3	r4	r5	r6	r7	r8	r9
-0.102	-0.657	-0.060	0.869	-0.089	-0.635	-0.054	0.832	-0.108



Forecasting

Autocorrelation

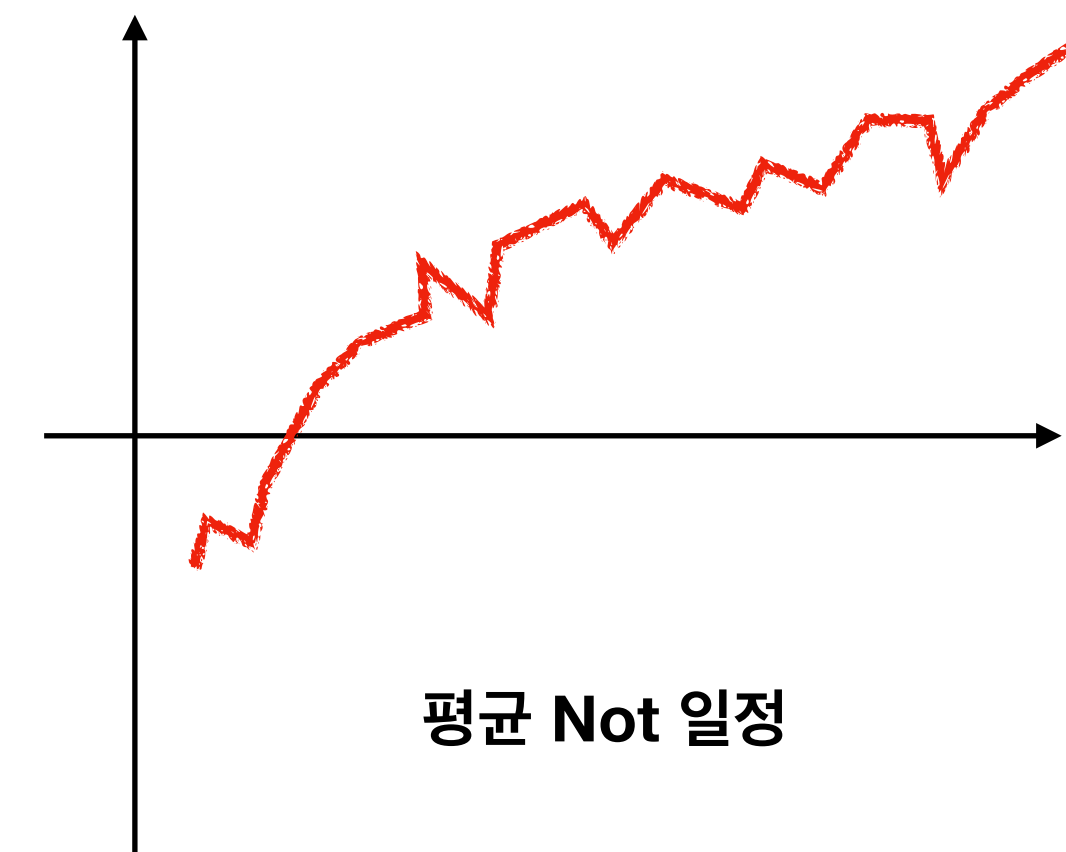
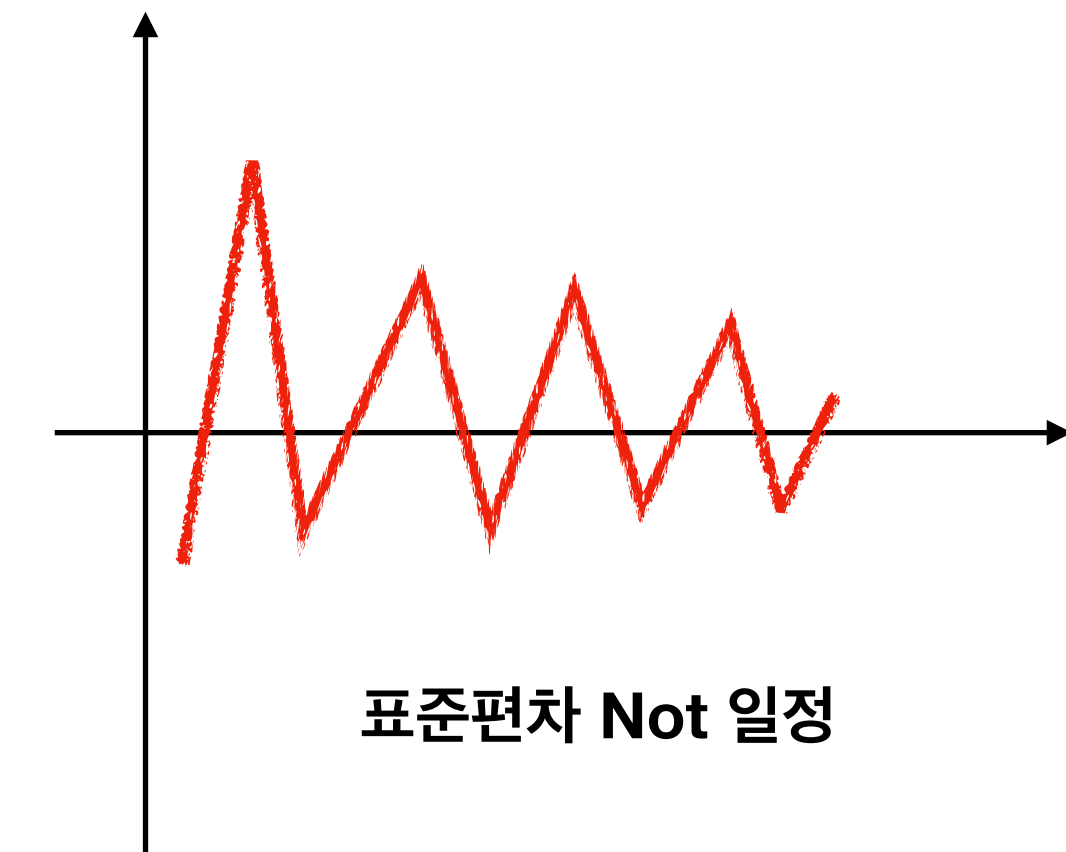
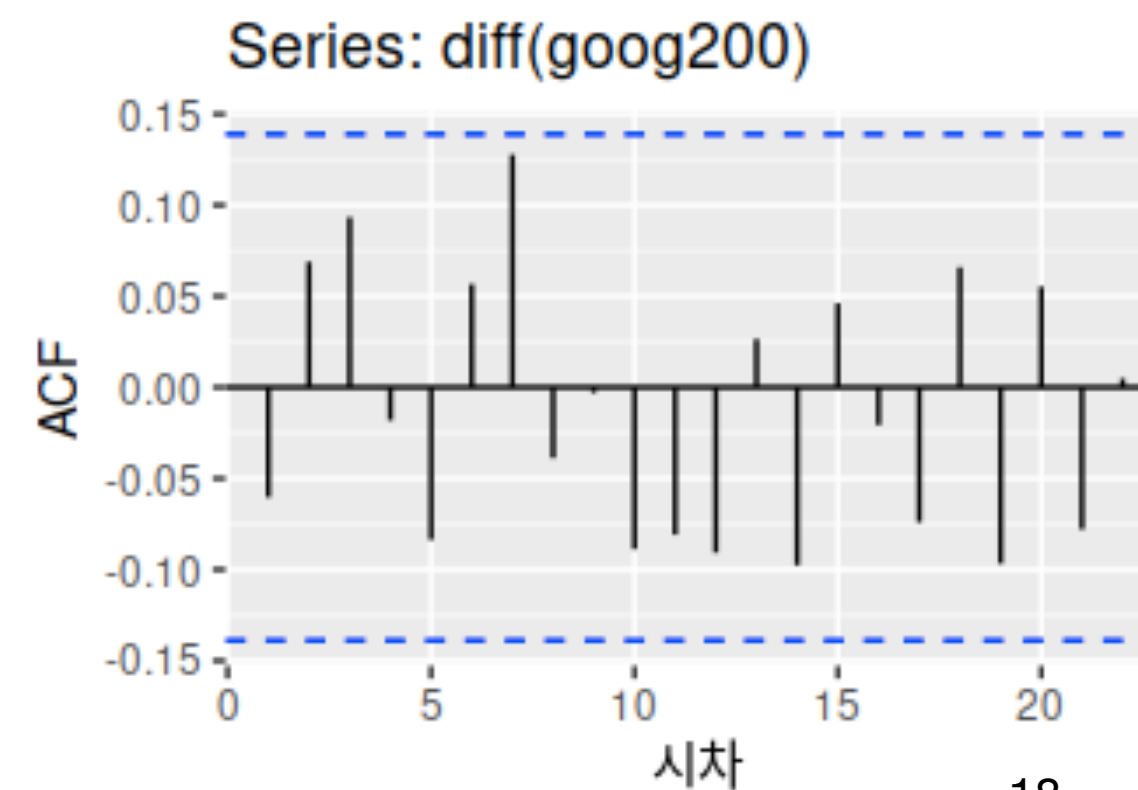
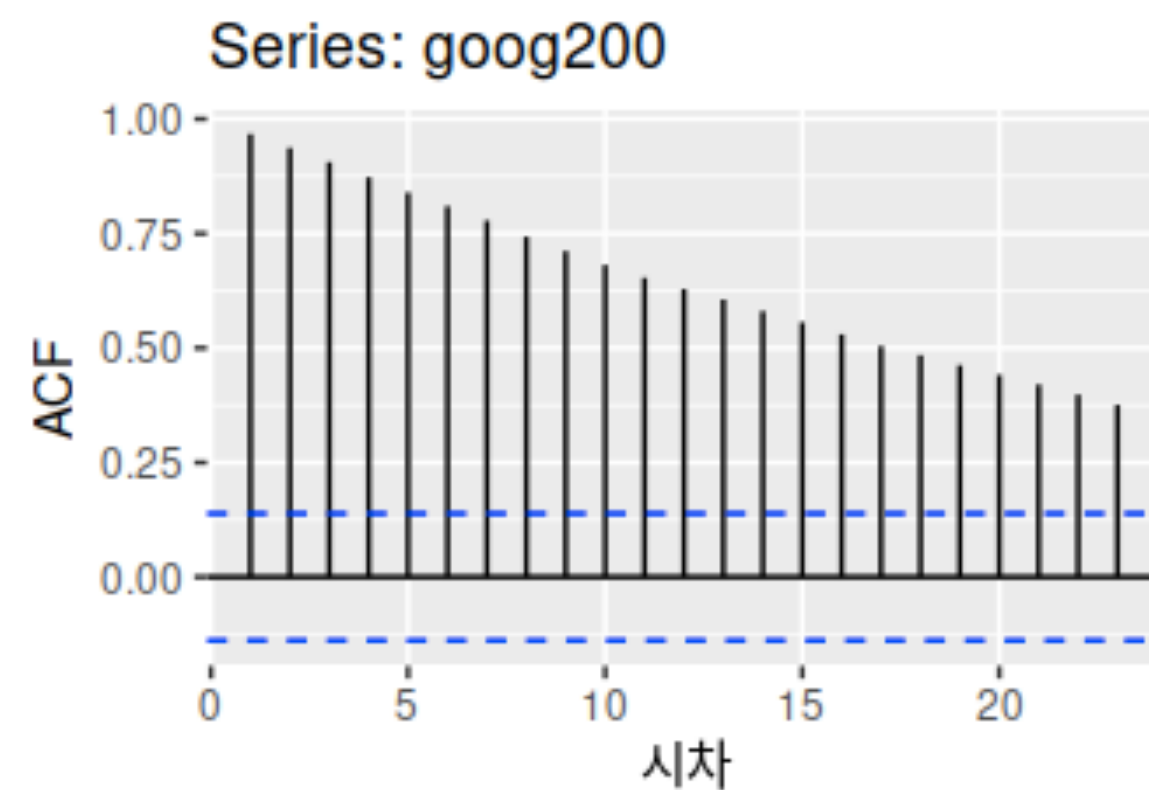
- ACF: Autocorrelation Function
 - 이전 시점을 모두 고려하여 두 시점의 상관관계를 구한다
 - t_3 와 t_4 의 상관관계를 구할때, t_1, t_2 도 같이 고려
 - MA의 차수를 결정할 때 사용
- PACF: Partial Autocorrelation Function
 - 이전 시점은 고려하지 않고 두시점의 상관관계를 구한다
 - t_4 는 t_3 에 의해서만 영향을 받는다고 가정
 - AR의 차수를 결정 할 때 사용



Forecasting

Stationarity

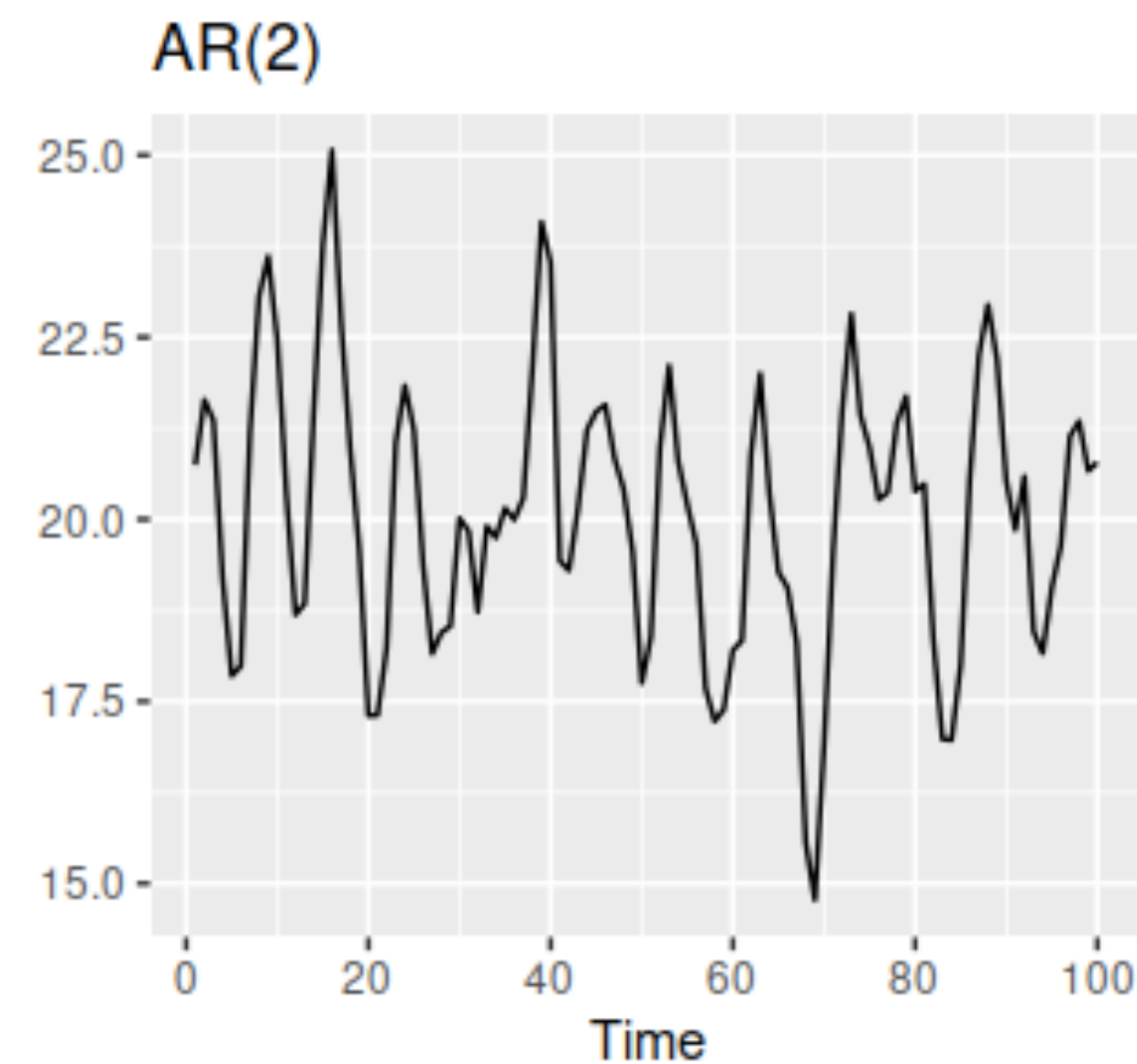
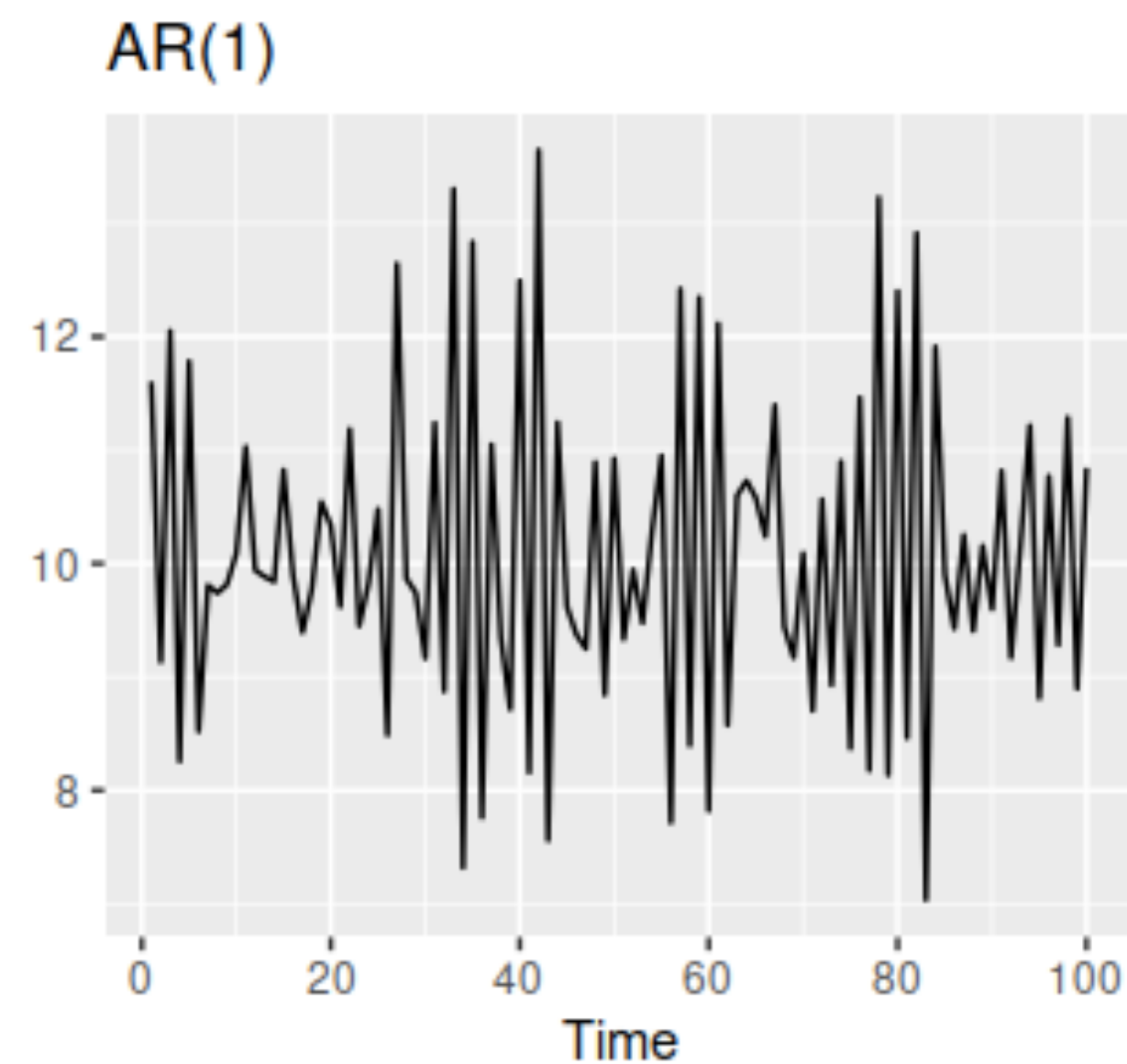
- 정상성의 조건
 - 평균이 일정 (No Trend)
 - 표준편차가 일정
 - No Seasonality
- 어떻게 정상성을 만족하게 만들 수 있을까?
 - 차분(Differencing)을 활용



Forecasting

Autoregressive Model (AR)

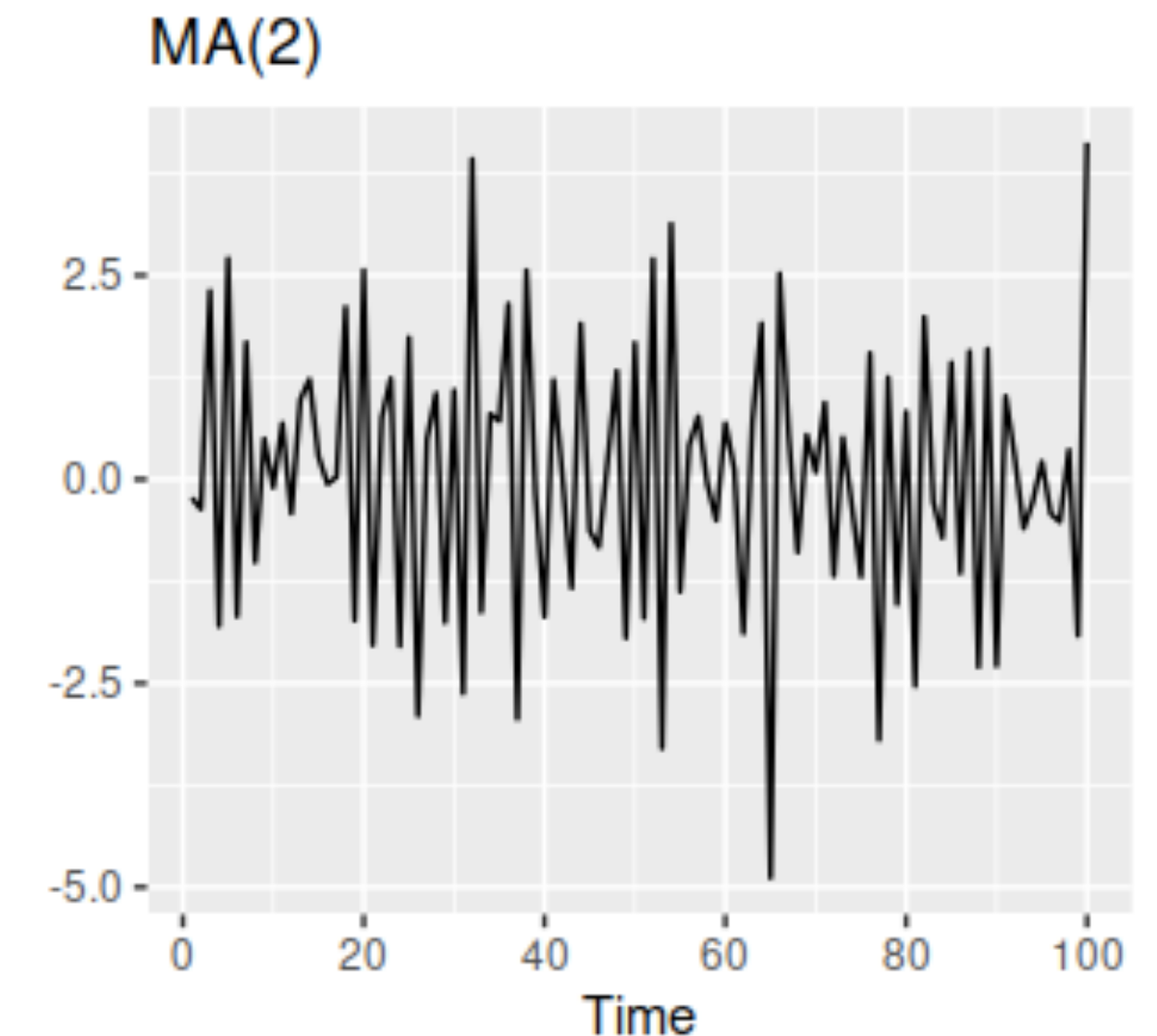
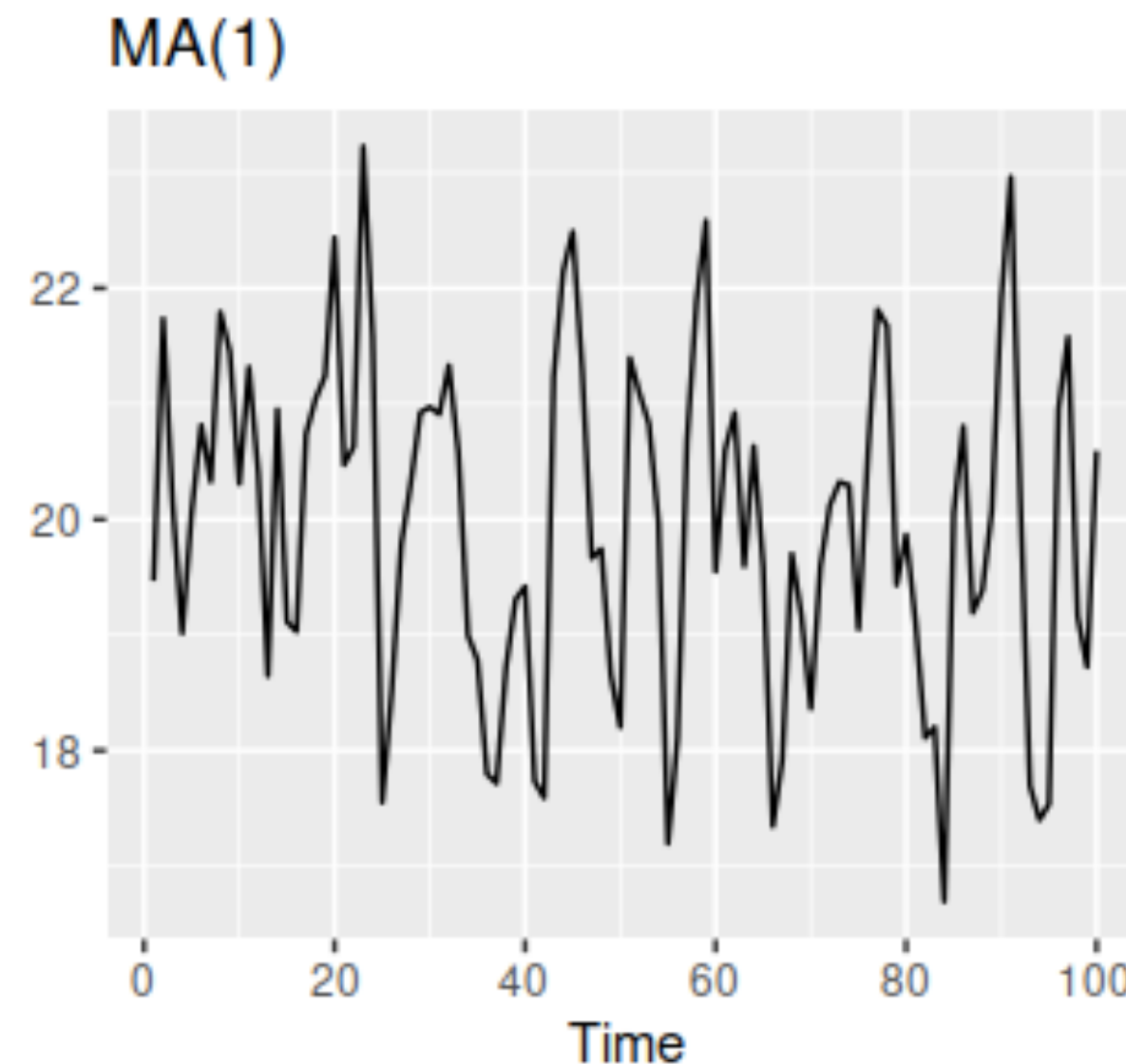
- 자기회귀 모델에서는, 변수의 과거 값의 선형 조합을 이용하여 관심 있는 변수를 예측
 - $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$
 - AR(p)에서 p는 PACF에서 확인



Forecasting

Moving Average Model

- 이동 평균 모델은 회귀처럼 보이는 모델에서 과거 예측 오차(forecast error)을 이용합니다.
 - 이동 평균 평활과 다름!
 - 과거의 주기-추세를 측정할 때 사용
 - MA Model은 예측 할때 사용
 - $y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$
 - MA(q)에서 q는 ACF에서 확인



Forecasting

ARMA

- Auto
Regressive
Moving
Average
- 과거의 값으로 미래를 예측(AR), 과거에 발생하는 오류로 미래의 오류를 예측 (MA)
- ARMA(1,1): $\ell_t = \beta_0 + \beta_1 \ell_{t-1} + \phi_1 \epsilon_{t-1} + \epsilon_t$
- ARMA(p,q)는 ACF와 PACF를 활용해 찾는다

Forecasting

ARIMA

- 이동평균을 누적한 자기회귀
 - Auto
Regressive
Integrated(누적)
Moving
Average
 - $y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t$
- Trend가 존재하는 데이터에 대응하기 위해
- ARIMA(p,d,q)에서 d는 몇차 차분(difference)이냐에 대한 변수
 - 1차 차분: 인접한 값간의 차이
 - 2차 차분: 인접한 값 간의 차이의 차이

Forecasting

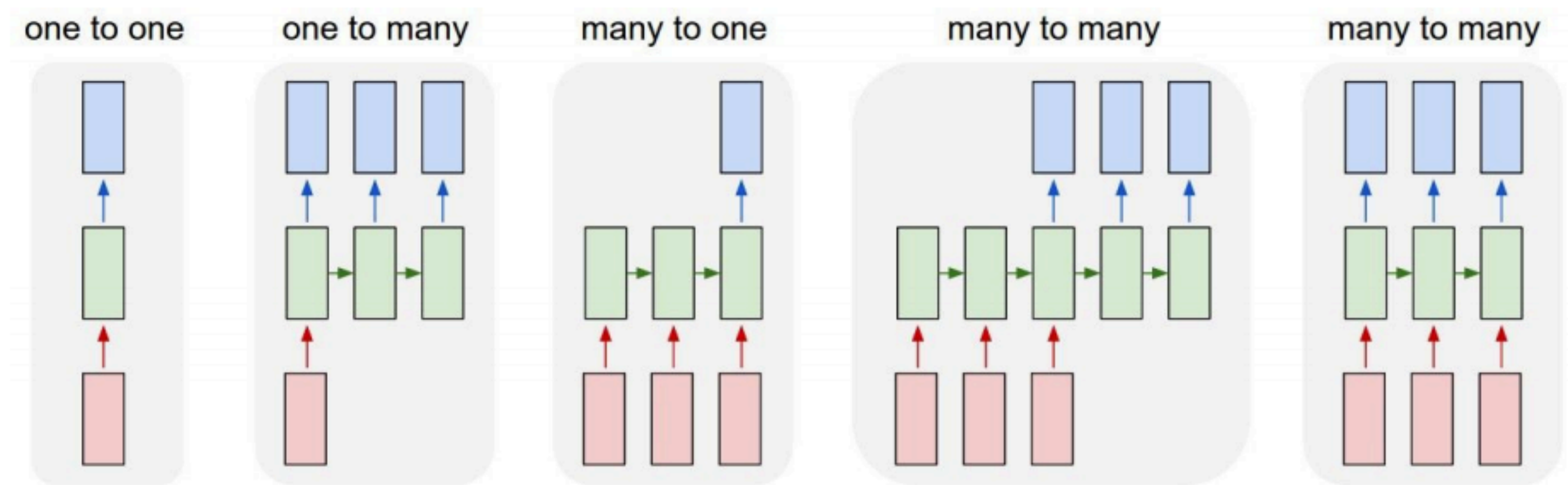
SARIMA

- 계절성 ARIMA 모델 : 계절성이 있는 데이터에 대응하기위해
 - **Seasonality**
Auto
Regressive
Integrated
Moving
Average
- $\text{ARIMA}(\overset{\text{비계절성 부분}}{p,d,q}) (\overset{\text{계절성 부분}}{P,D,Q})m$
 - m = seasonal factor
 - p : AR 차수, d : 차분의 차수, q : MA 차수
 - 대문자 P,D,Q 는 계절성에 대한 p,d,q
- $\text{ARIMA}(1,1,1)(1,1,1)4$
 - $(1 - \phi_1 B) (1 - \Phi_1 B^4)(1 - B)(1 - B^4)y_t = (1 + \theta_1 B) (1 + \Theta_1 B^4)\varepsilon_t$

Forecasting

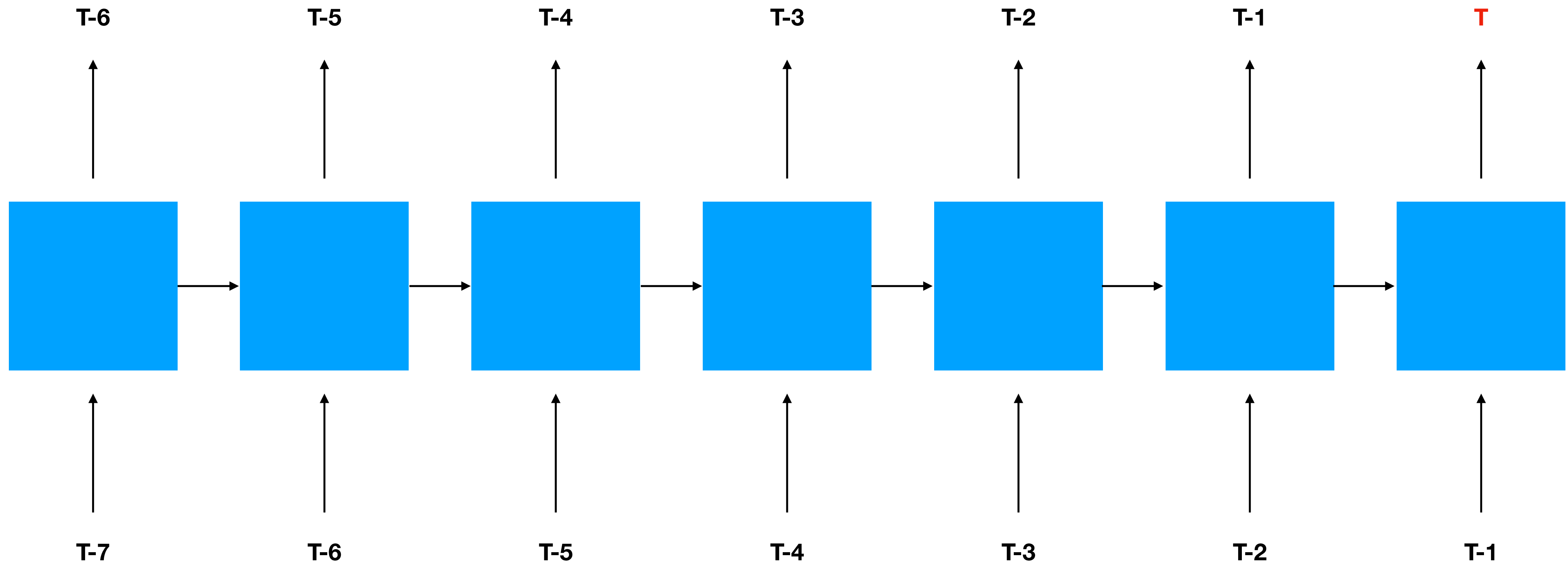
Recurrent Neural Network (RNN)

- 데이터 특성을 파악함에 있어서는, 전통적인 방식이 작용을 하지만, 예측에 있어서는 명확한 한계가 존재
- 많은 테스크들이 딥러닝 기반의 방법론을 쓰는 이유처럼, 만약 패턴을 인공지능망이 알아서 결정하게 한다면?
- Recurrent Neural Networks



Forecasting

Recurrent Neural Network (RNN)



References

Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. [OTexts.com/fpp2](https://otexts.com/fpp2). Accessed on 2020-11-19.

E.O.D