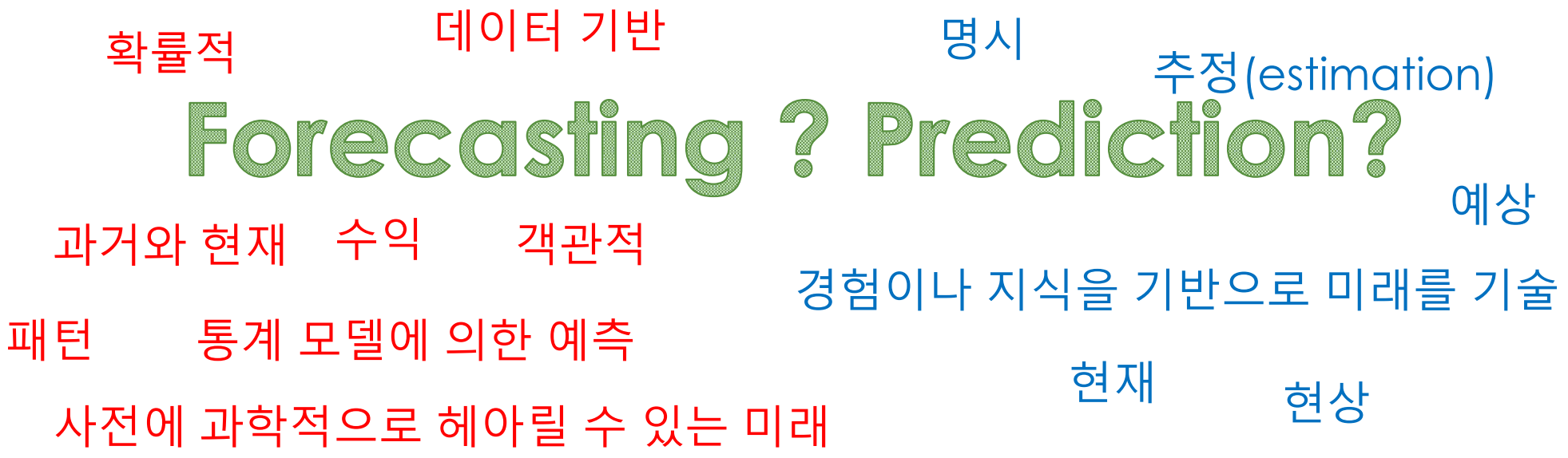
orange 활용 데이터 분석 및 머신 러닝

13차시

시계열분석 (Time Series Analysis)

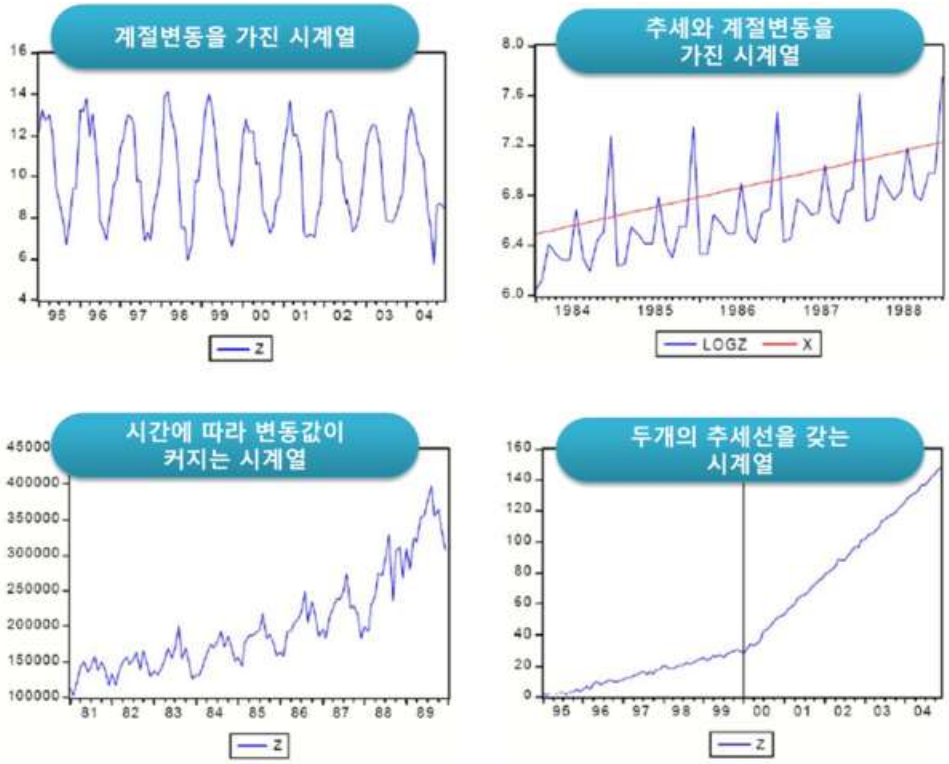
예측이란 무엇인가?



- Forecast : 과거와 현재 데이터에 기반하여 예측을 수행하는 과정이다. [위키피디아]. 미래에 어떤 일이 발생할지를 판단하는 설명이다 [캠브리지 사전]
- Prediction : 라틴어 præ-, "전", dicere, "말하다"의 조합어로 명명되며, 미래의 사건이나 데이터에 대해 서술하는 것이다 [위키피디아]. 미래에 어떤 사건이나 행위가 일어날 것이라고 말하는 것으로써 지식 또는 경험의 결과로써 서술된다 [캠브리지 사전].

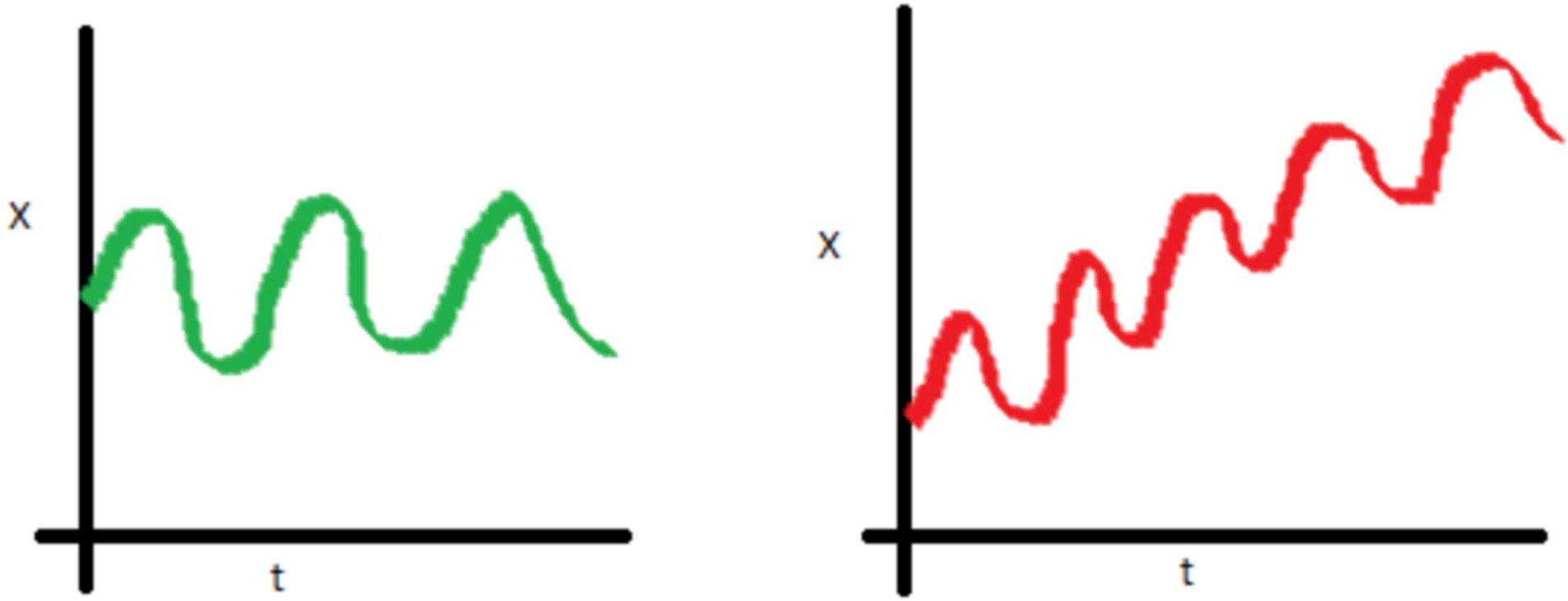
시계열 그림 (Time series Plot)

- 시간의 변화에 따라 시계열 자료의 값이 변하는 것을 나타낸 그림
- 시계열의 특징을 쉽게 파악할 수 있어 해당 자료의 성격에 적합한 분석 방법의 선택에 도움을 줌
- 불규칙변동 / 확률적 변동 : 규칙성이 없이 예측 불가능하게 발생하는 변동 (전쟁, 홍수 화재 파업 등의 원인이 있을 수 있음).
- 체계적 변동 :
  - 추세변동 (trend variation) : 장기간에 걸쳐 어떤 추세로 나타나며 장기간에 걸쳐 지속적으로 증가 또는 감소하거나 일정 상태를 유지하려는 성향을 의미한다.. 예로 국민 총생산, 인구증가율, 기술변화 등..)
  - 순환변동 (syclical variation): 장기적인 일정기간을 주기로 순환적으로 나타나며 경기 변동 곡선과 같이 불황과 회복, 호황과 경기후퇴 등이 수년을 주기로 나타나는 변동이 그 예이다.
  - 계절변동 (seasonal variation) : 계절적 영향과 사회적 관습에 따라 1년 주기로 발생하는 변동요인. 순환주기가 짧은 특징을 지님.



출처 : 경영자를 위한 디지털 전략 가이드, 스마투스 비즈니스 리뷰(http://www.sbr.ai)

Stationary(정상성) vs. Non-Stationary



정상 프로세스 : 시간에 관계없이 평균과 분산이 일정한 시계열 데이터

정상성 (stationary) : 데이터 변동의 안정성

◦ 정상성을 가진 데이터 : 일관된 평균과 분산과 자기상관정도를 보이는 데이터. 모든 시점에 대해서 일정한 평균과 분산을 가진다.

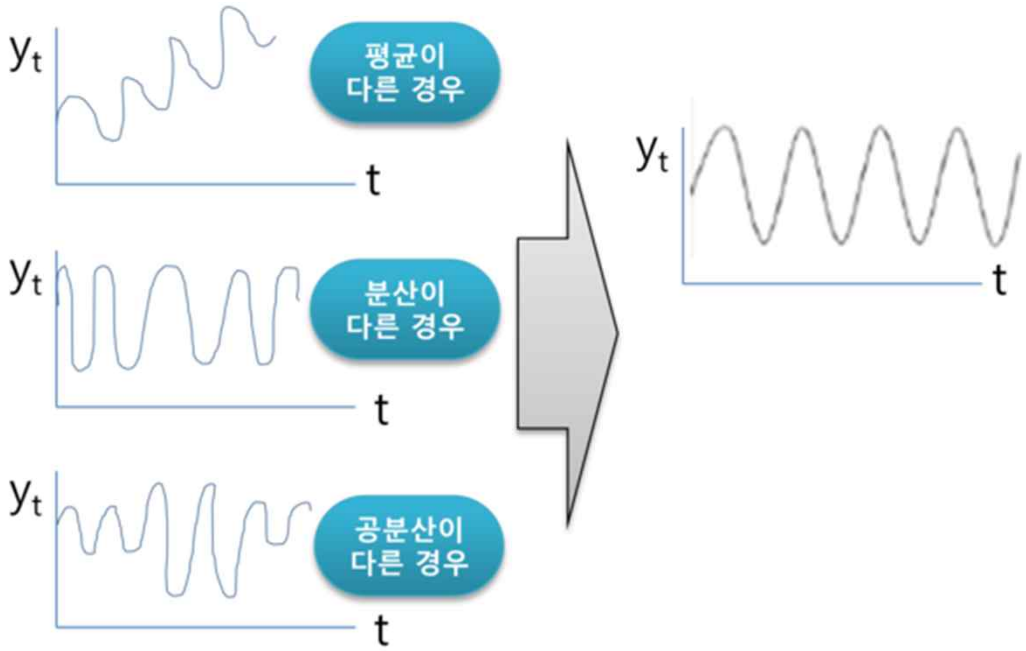
1) 평균이 일정하지 않은 시계열은 차분(difference)을 통해 정상화 할 수 있다.

◦ 차분(difference)은 현 시점 자료에서 전 시점 자료를 뺌.

2) 분산이 일정하지 않은 시계열은 변환(transformation)을 통해 정상화한다.

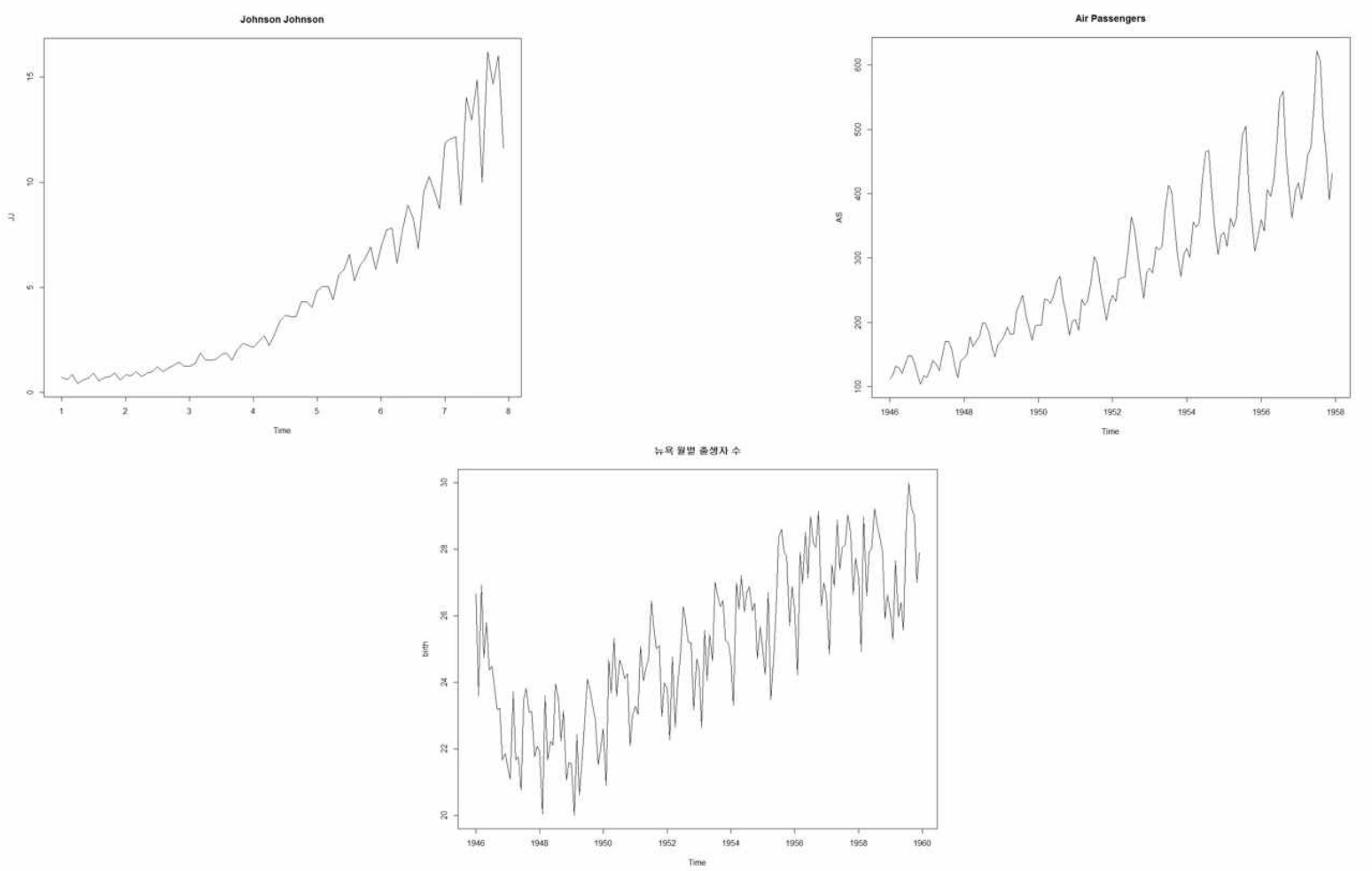
◦ 변환을 통해 정상성을 높이는 방법에는 이동평균법, 지수평활법 등이 있다.

◦ 정상성을 가진 데이터로 만드는 이유 : 시간의 흐름에 따라 증가 혹은 감소 추세가 있는 현상을 연구할 때, 혹은 계절적, 주기적으로 증감 현상을 보이는 이슈를 연구할 때 자연 발생적인, 혹은 문제의 예측 변수와 관계 없는 요인(힘: forces)들의 영향력을 배제하고 순수한 예측변수의 힘을 보기 위한 것



출처 : 경영자를 위한 디지털 전략 가이드, 스마투스 비즈니스 리뷰(http://www.sbr.ai)

대부분의 데이터는 Non-Stationary



[표] 대표적인 시계열 분석 모형들

- 시계열 정보 = 규칙성을 가지는 패턴 + 불규칙한 패턴의 결합
- 규칙성 : 자기상관성 / 이동 평균

종류
AR (자기회귀모형) : Auto regressive model
MA (이동평균모형) : Moving Average model
ARMA (자기회귀이동평균모형) : Autoregressive Moving Average model
ARIMA (자기회귀누적이동평균모형) : Autoregressive Integrated Moving Average model

AR, MA, ARMA

◦ AR – 자기상관(Autocorrelation) 모형

$$y_t = \emptyset_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \ \ldots + \emptyset_p y_{t-p} + \varepsilon_t$$

◦ MA 이동평균(Moving Average) 모형

$$y_t = \theta_0 + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ \ldots + \theta_q \varepsilon_{t-q}$$

◦ ARMA 모형

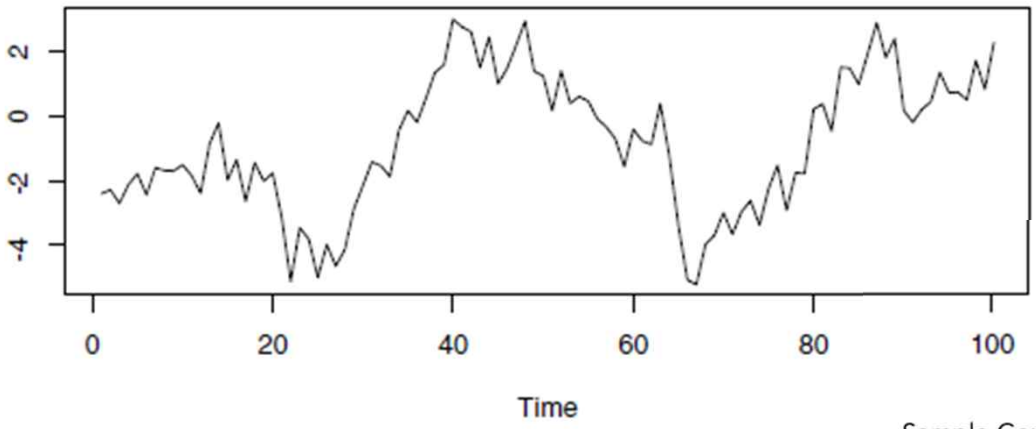
$$\boldsymbol{y_t = \emptyset_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots \emptyset_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ \ldots + \theta_q \varepsilon_{t-q}}$$

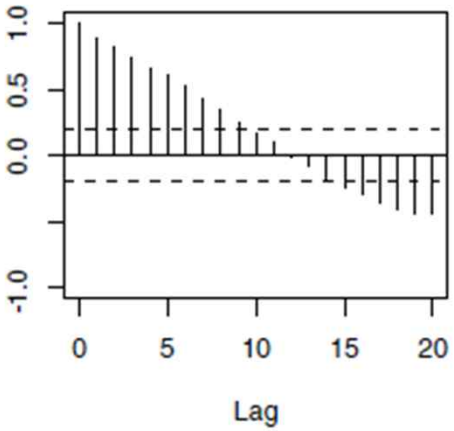
ARIMA (자기회귀누적이동평균모형) : Autoregressive Integrated Moving Average model

- 과거의 데이터변화에 대한 추세관계 (cointegration) 까지 고려한 모델
- Correlation – 서로 간의 선형관계
  - Correlation > 0 => x가 클 때 y값도 큰 값을 가진다.
  - Correlation < 0 => x가 클 때 y값은 작은 값을 가진다.
- Cointegration – 추세관계
  - Cointegration > 0 => x의 값이 이전 값보다 증가하면 y 값은 현재는 작은 값이지만 곧 증가하는 추세로 바뀐다.
  - Cointegration <0 => x의 값이 이전 값보다 증가하면 y 값은 현재는 큰 값이지만 곧 하락하는 추세로 바뀐다.

Autocorrelation

lag 1: 자기 자신과 자기 자신 이전 데이터와의 correlation





Sample Correlation Coefficient

Value of X

Value of Y

Summation: "Take The Sum Of"

Mean of X Variable

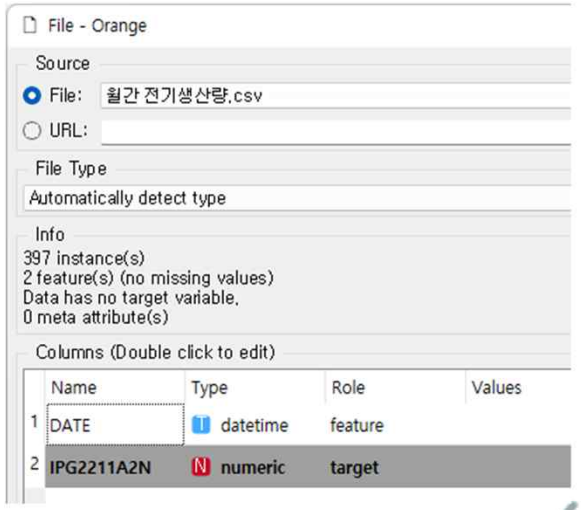
Mean of Y Variable

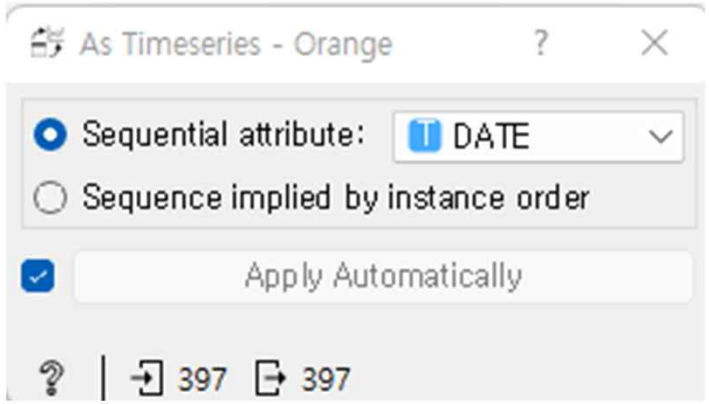
$$r = \frac{\sum \left[\left(x_i - \overline{x} \right) \left(y_i - \overline{y} \right) \right]}{\sqrt{\underbrace{\sum \left(x_i - \overline{x} \right)^2}_{\text{Sum of the squared deviations for X}} * \underbrace{\sum \left(y_i - \overline{y} \right)^2}_{\text{Sum of the squared deviations for Y}}}}$$

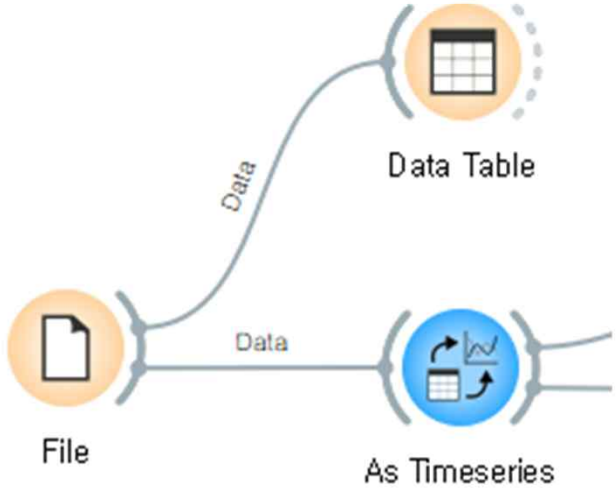
Square Root

월간 에너지 생산량을 예측해보자.

Time Series 위젯

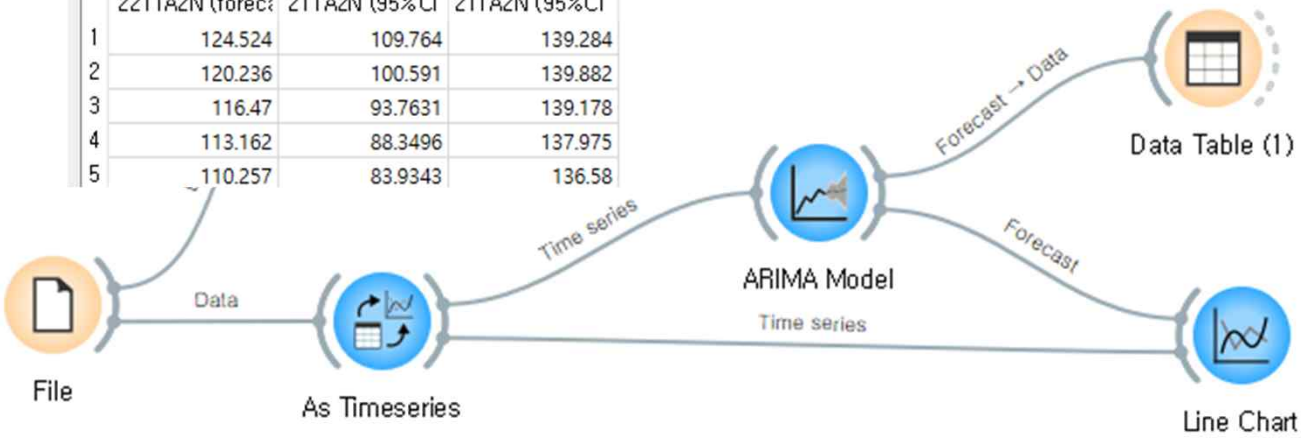


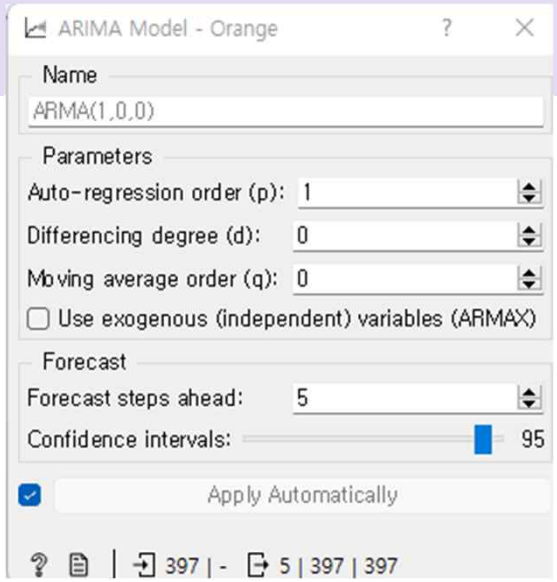


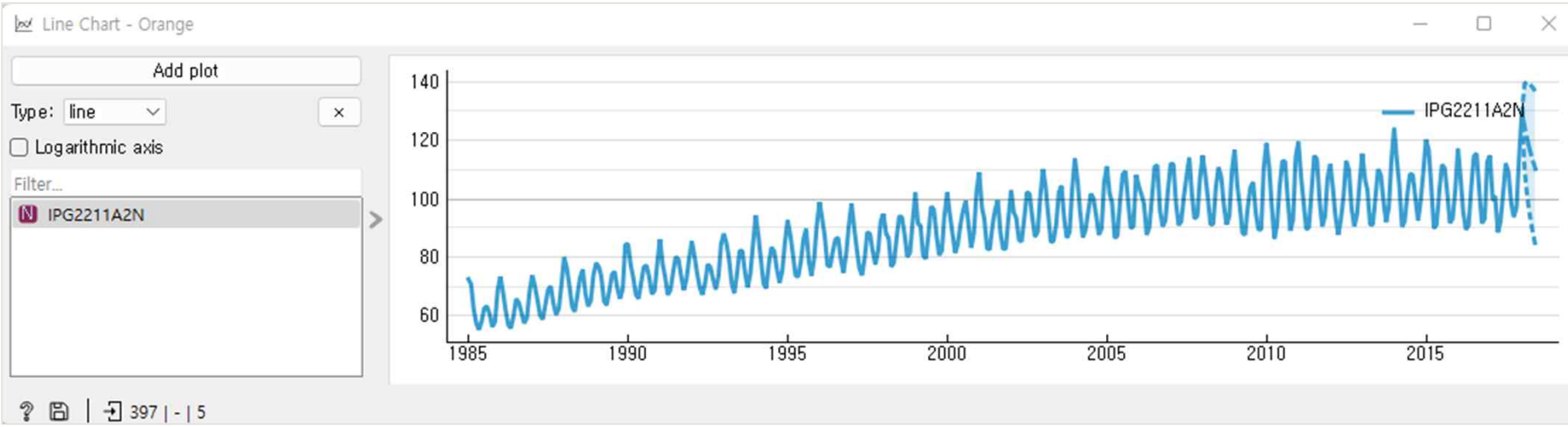


ARIMA MODEL 위젯과 라인차트 시각화

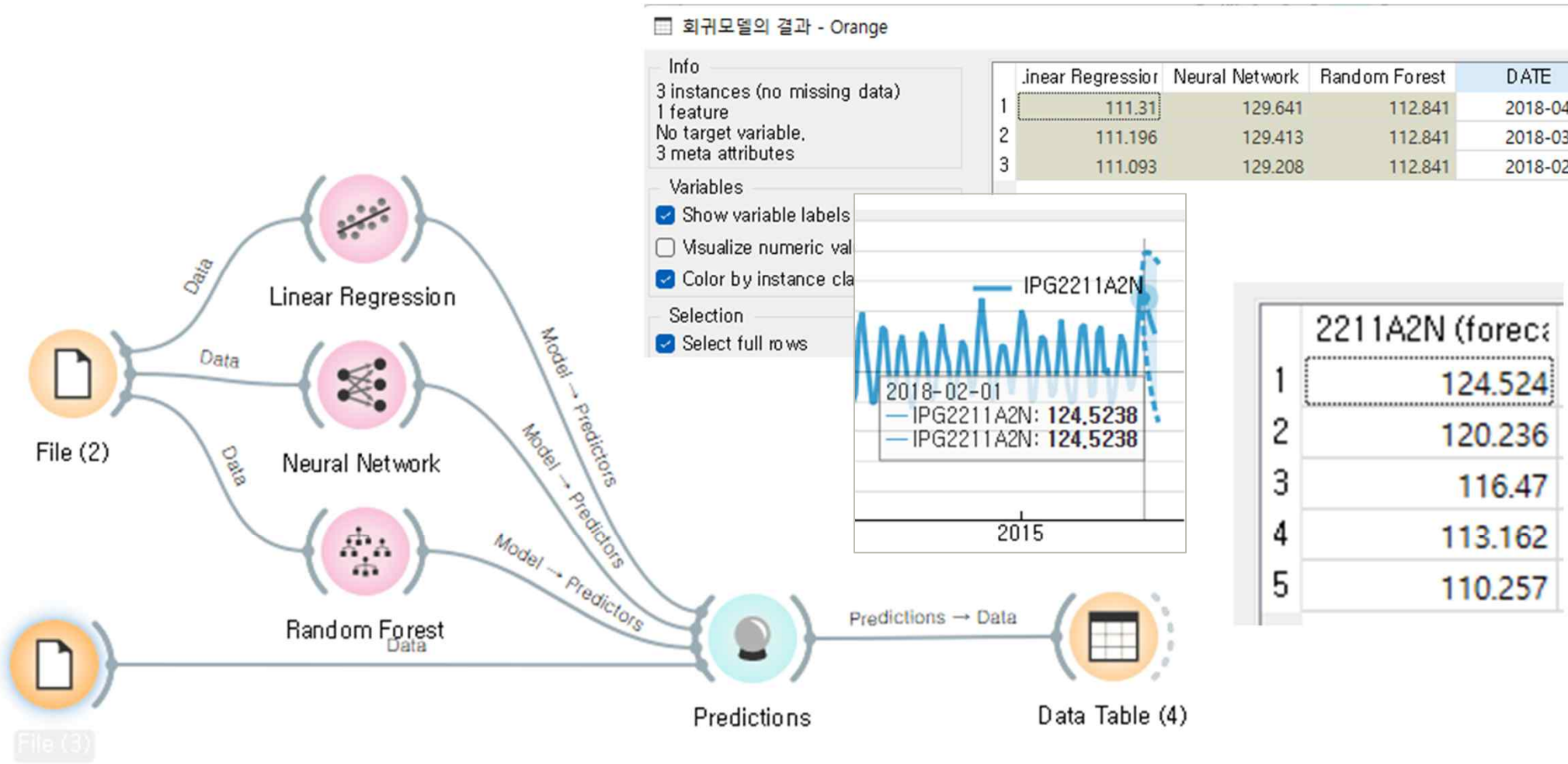
	2211A2N (foreca	211A2N (95%CI	211A2N (95%CI
1	124.524	109.764	139.284
2	120.236	100.591	139.882
3	116.47	93.7631	139.178
4	113.162	88.3496	137.975
5	110.257	83.9343	136.58



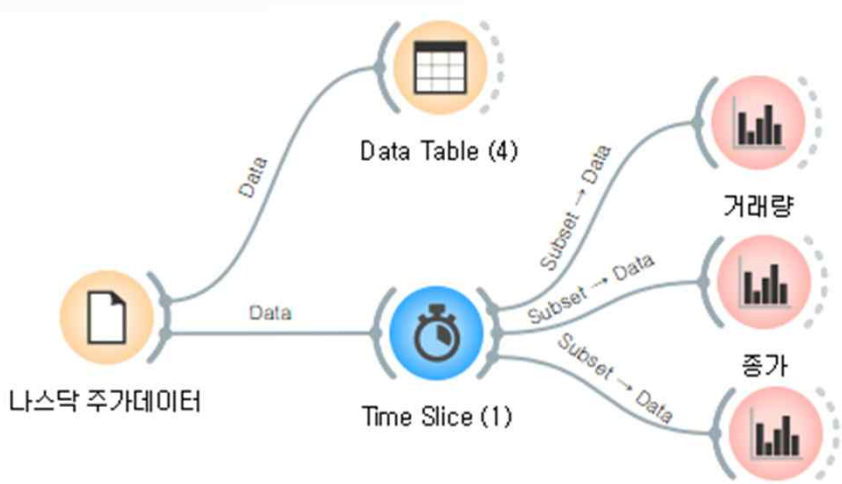


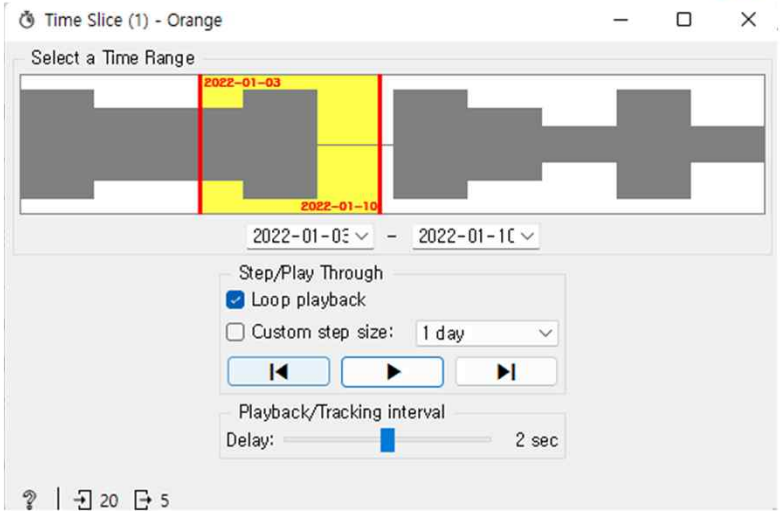


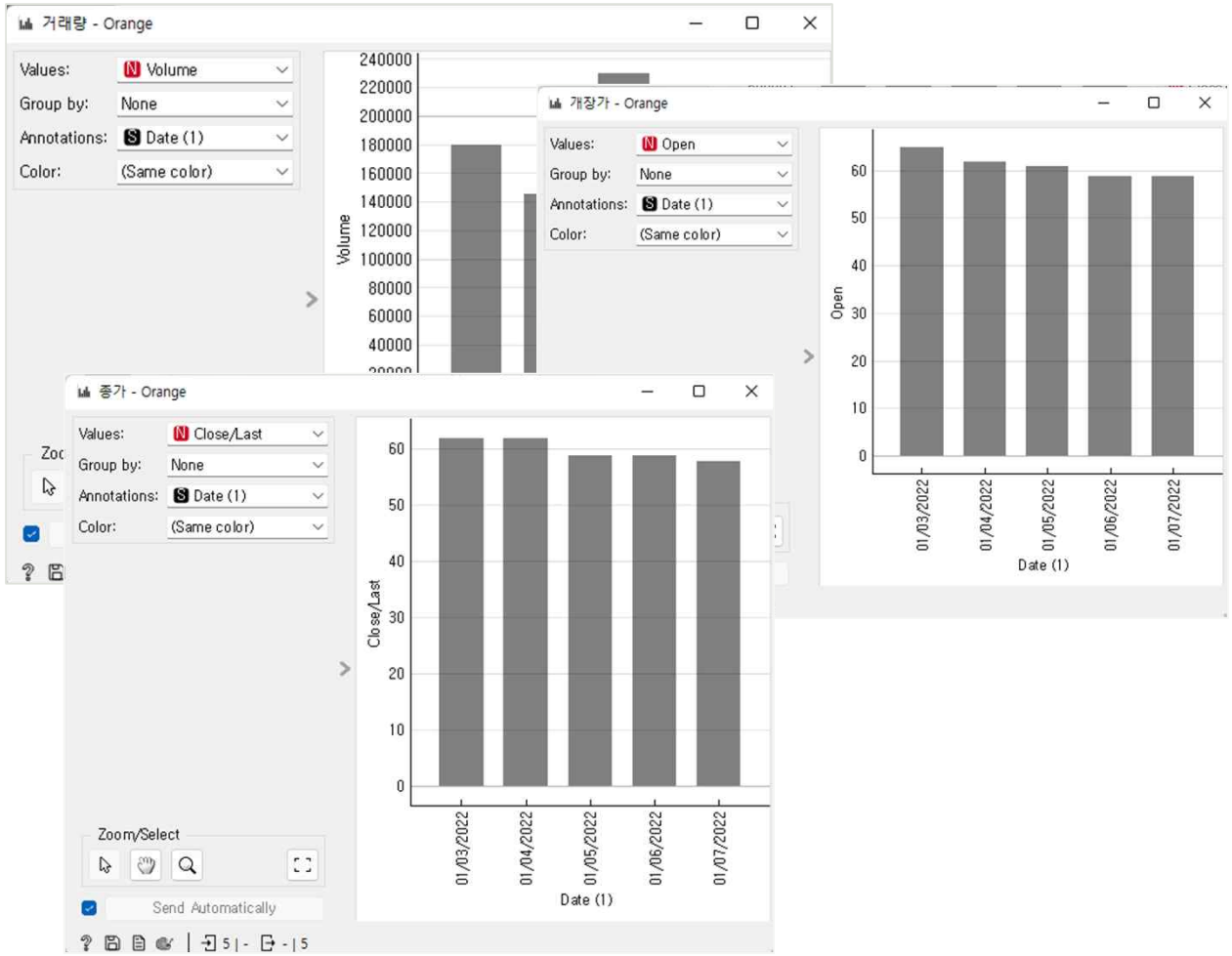
일반 회귀모델로 예측한 결과와의 비교



Time slice 연습 - 나스닥 주가 변동 확인하기







다음 시간에는 데이터 전처리 과정의 필요성 및 오렌지에서 제공되는 전처리 과정을 연습해 보도록 하겠습니다.