AI+X선도인재양성기초프로젝트

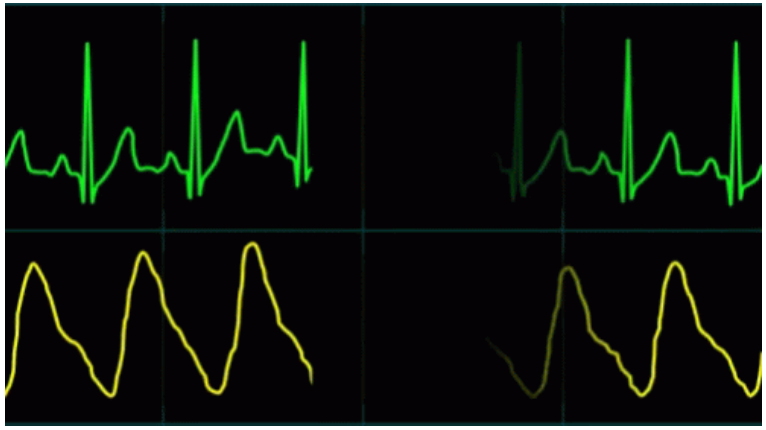
13. 시계열 데이터 분석

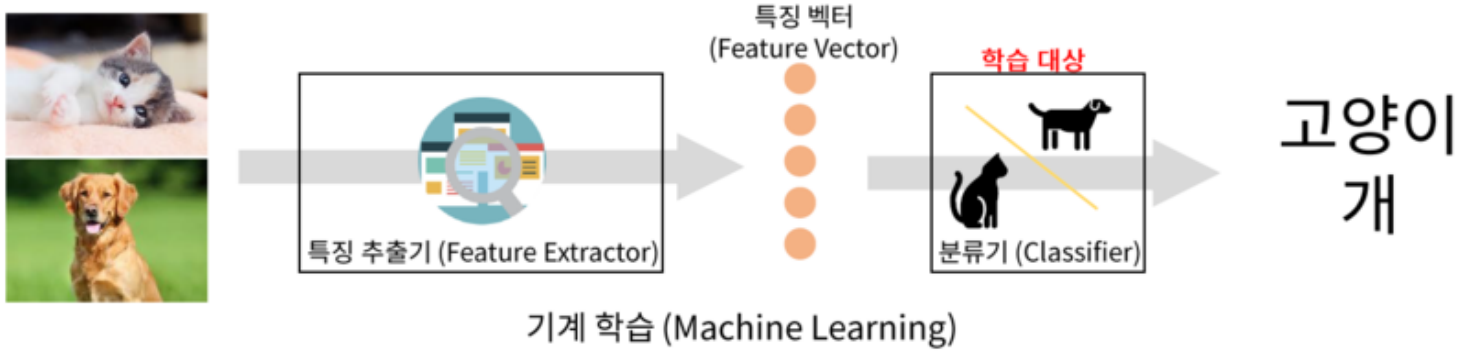
Heenam Yoon

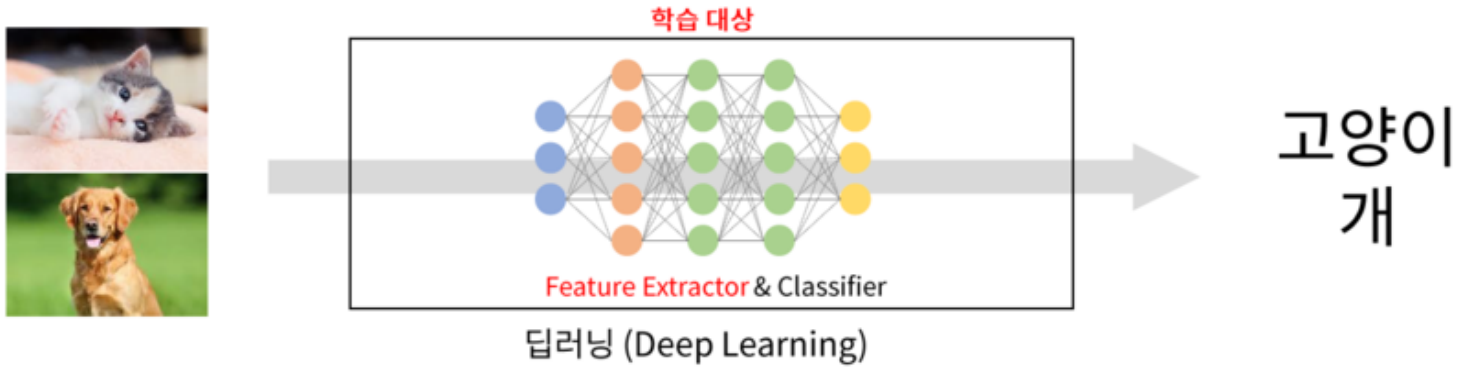
Department of
Human-Centered Artificial Intelligence

E-mail) h-yoon@smu.ac.kr
Room) 0112

시계열 데이터 예

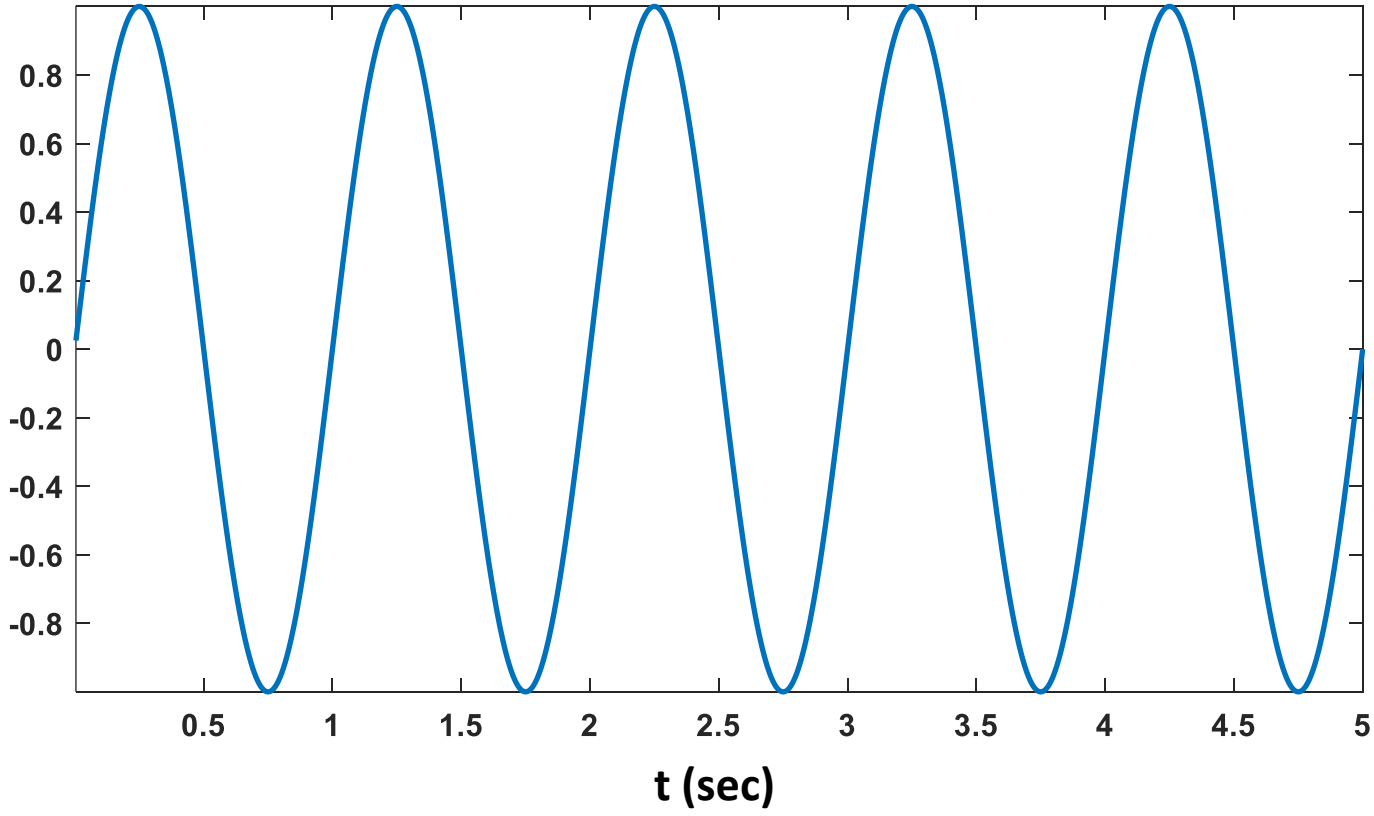




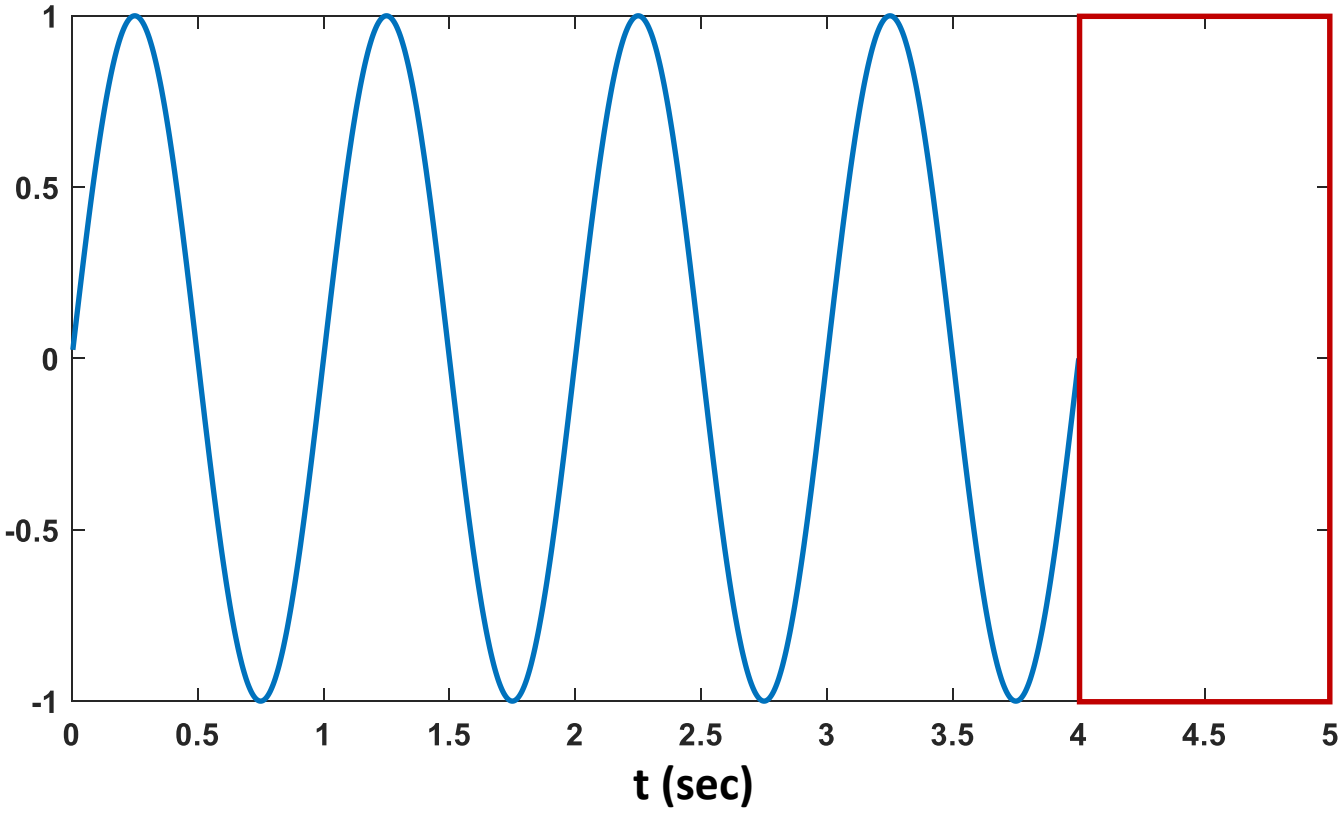


$$y(t) = sin(2\pi t)$$

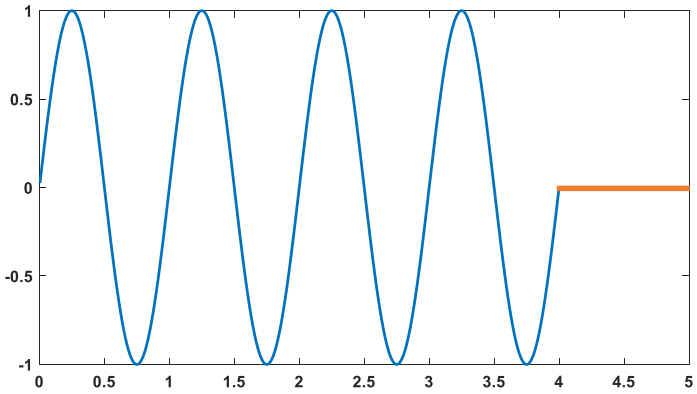
주기: 1초
주파수: 1/T= 1Hz



- 주어진 데이터 (파란색)을 이용하여 unknown값을 예상해보자
- 어떻게 하면 좋을까



마지막 값으로 하겠다



마지막 5개 데이터로
회귀식을 만들어서 해보겠다

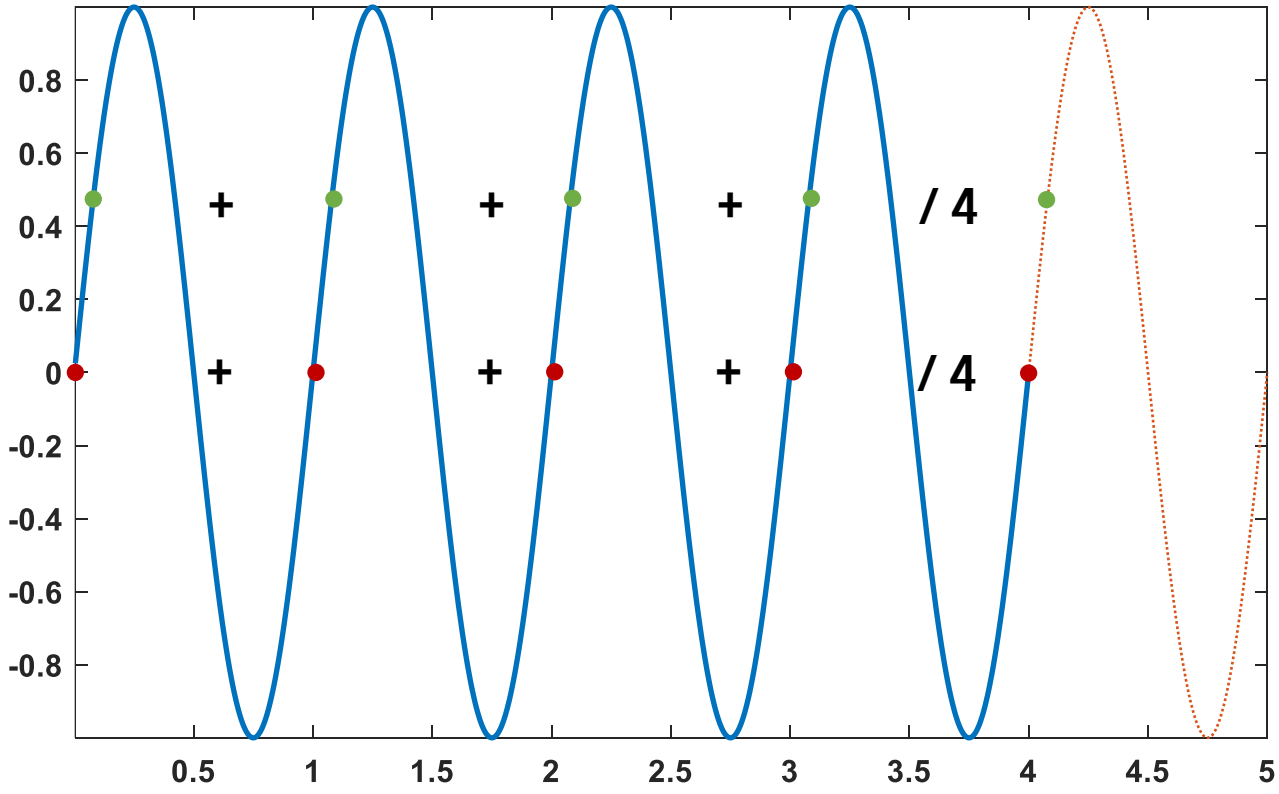
일어나지 않은 일이니 무엇이 정답인지 알 수 없음

$$y(t) = sin(2\pi t)$$

1초 **주기**로 데이터가 반복되고 있음

0, 1, 2, 3초 때의 값들의 평균을 4초 때의 값으로 하겠다

0.1, 1.1, 2.1, 3.1초 때의 값들의 평균을 4.1초 때의 값으로 하겠다



주기를 아는 방법

- 신호처리 관점에서는 주파수변환을 해보면 알 수 있음
- Autocorrelation (자기 상관)을 통해 알 수 있음
- Correlation

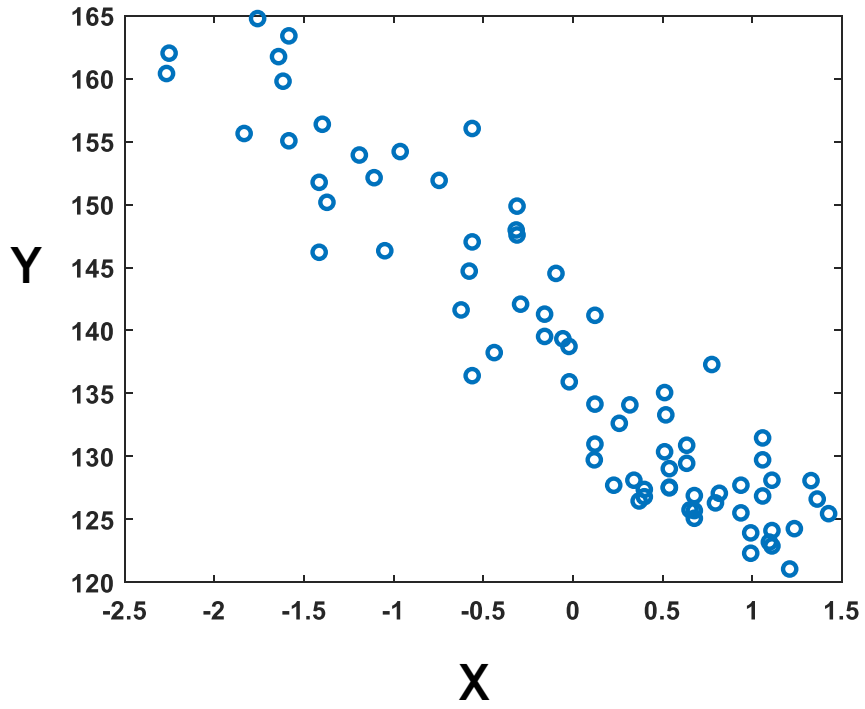
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- Cross correlation
- Auto correlation

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

주기를 아는 방법

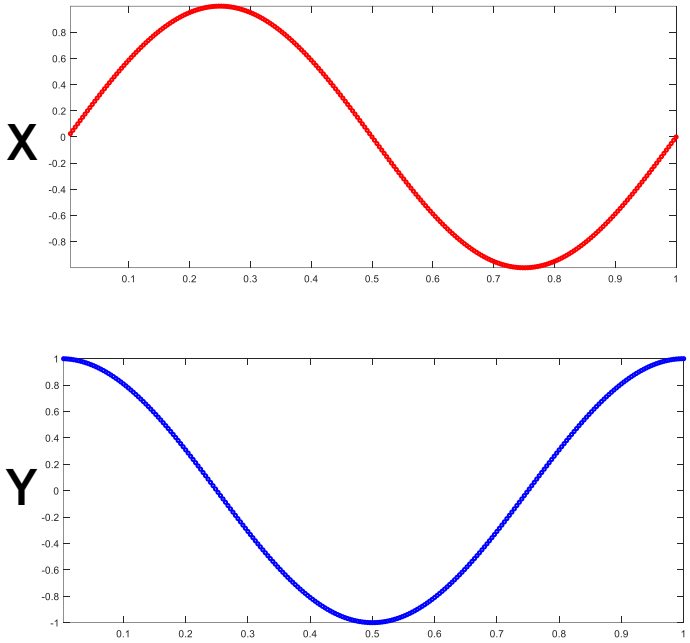
- Correlation

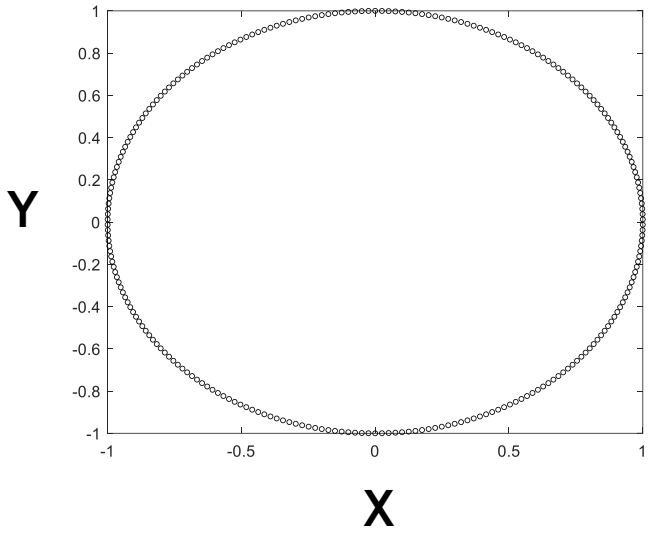


X	Y

주기를 아는 방법

- Cross correlation



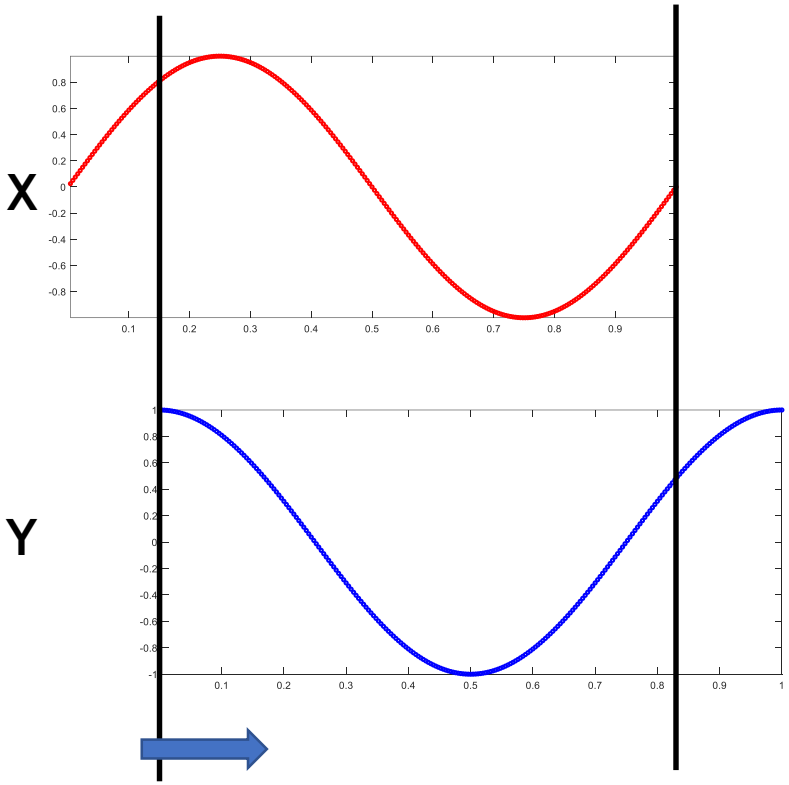


Correlation 계산

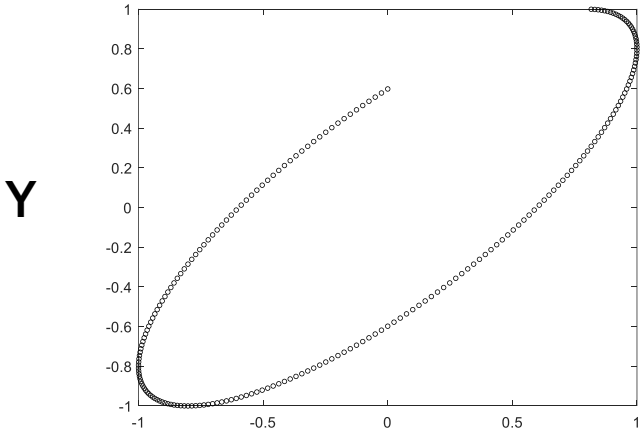
lag=0초

주기를 아는 방법

- Cross correlation



Shifting(sliding)

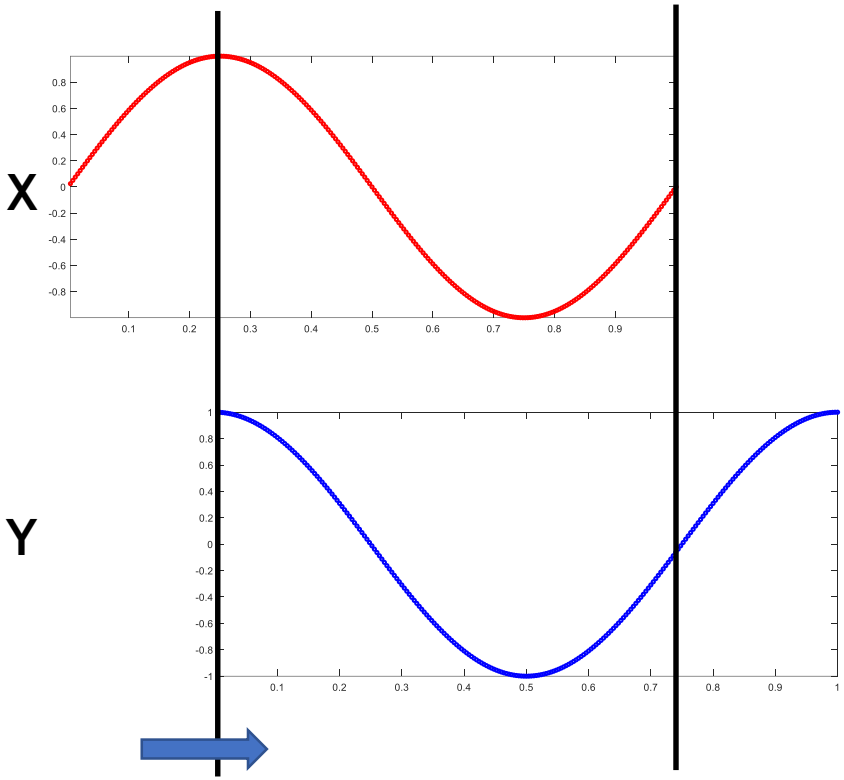


Correlation 계산

lag=0.15초

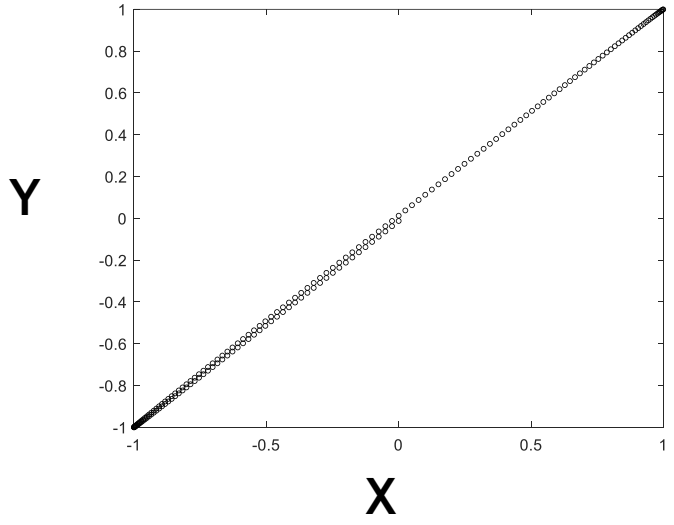
주기를 아는 방법

- Cross correlation



Shifting(sliding)

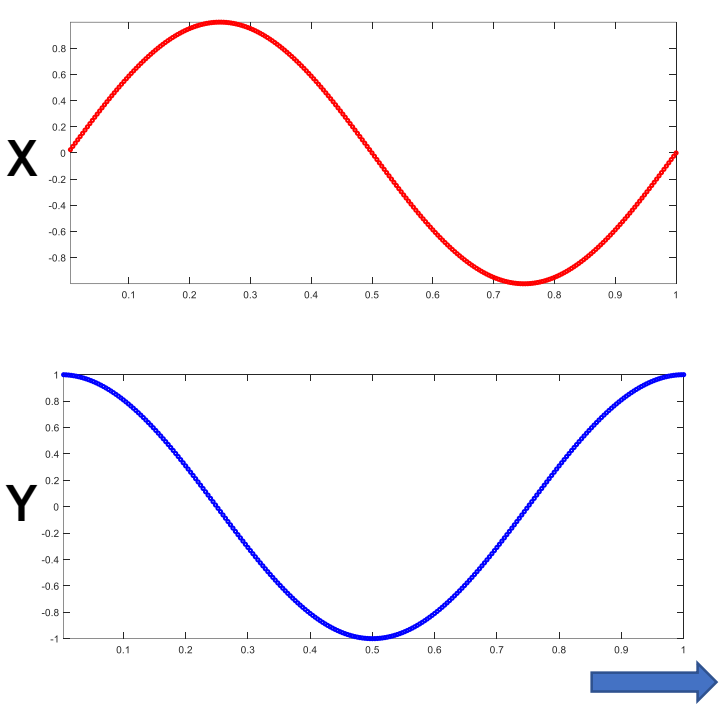
lag=0.25초

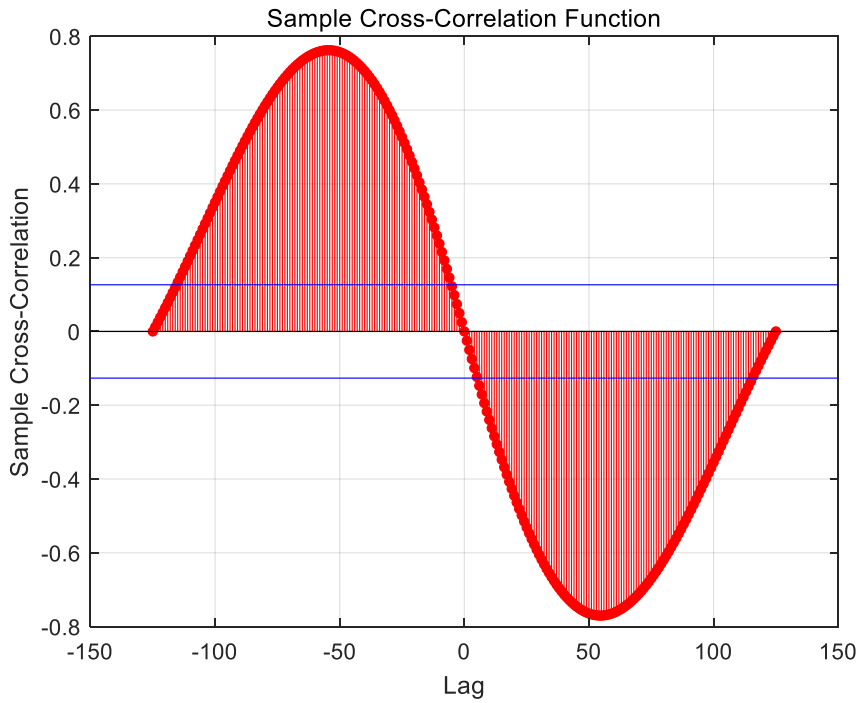


Correlation 계산

주기를 아는 방법

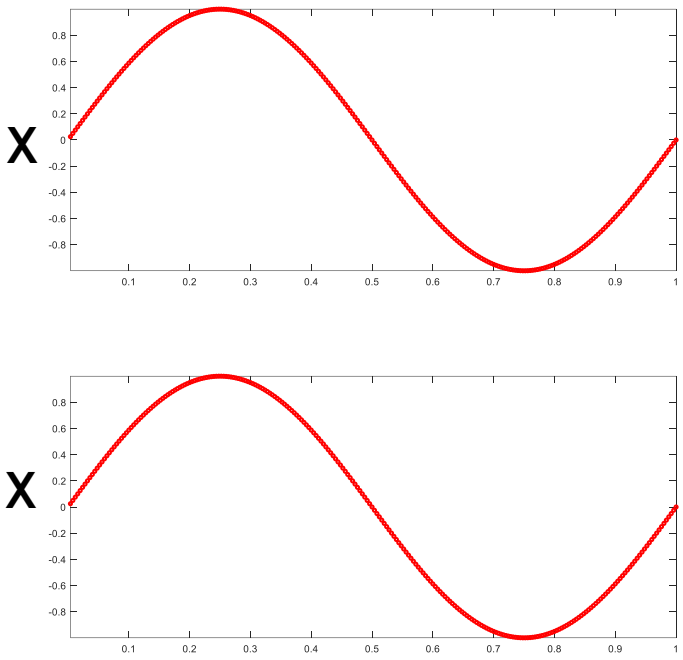
- Cross correlation

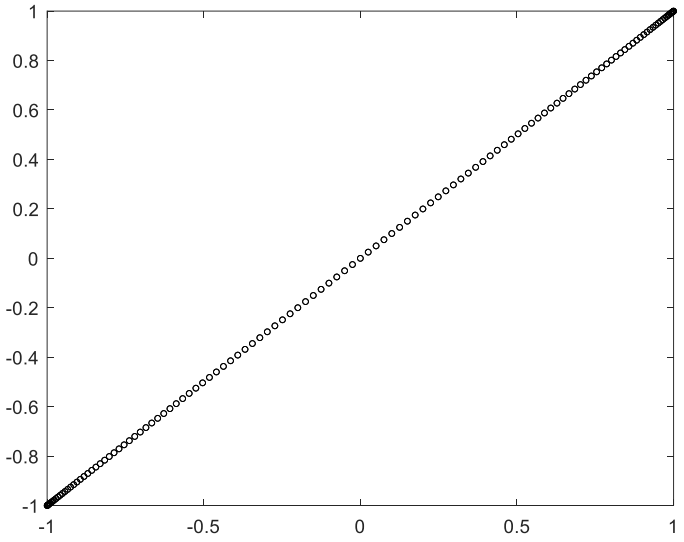




주기를 아는 방법

- Autocorrelation

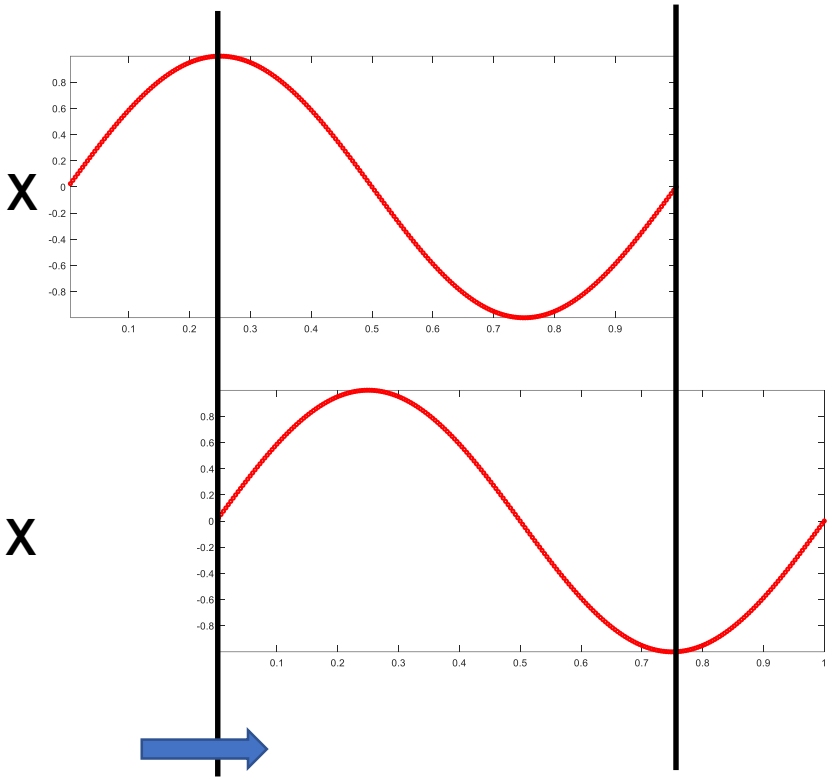




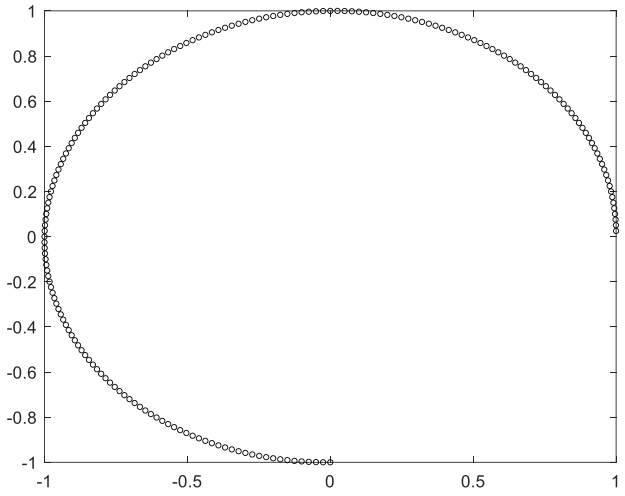
Correlation 계산

주기를 아는 방법

- Cross correlation



Shifting(sliding)

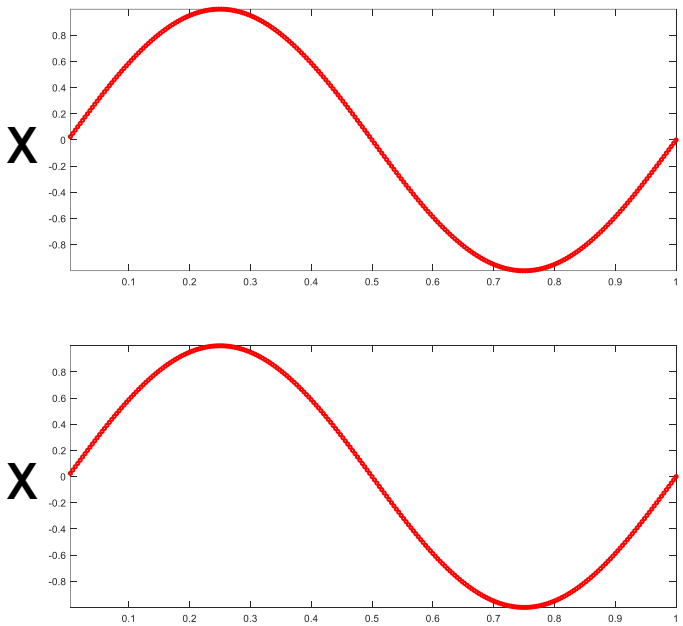


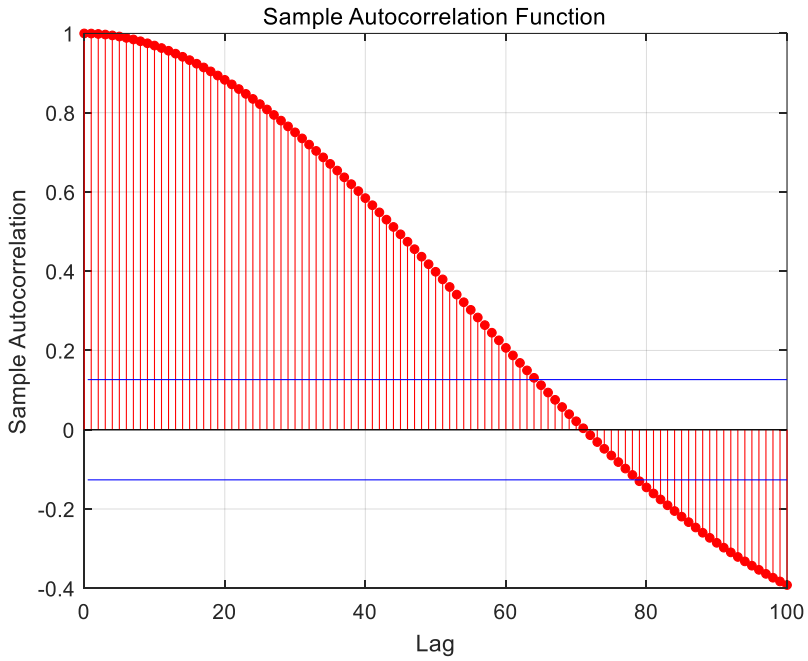
Correlation 계산

lag=0.25초

주기를 아는 방법

- Autocorrelation



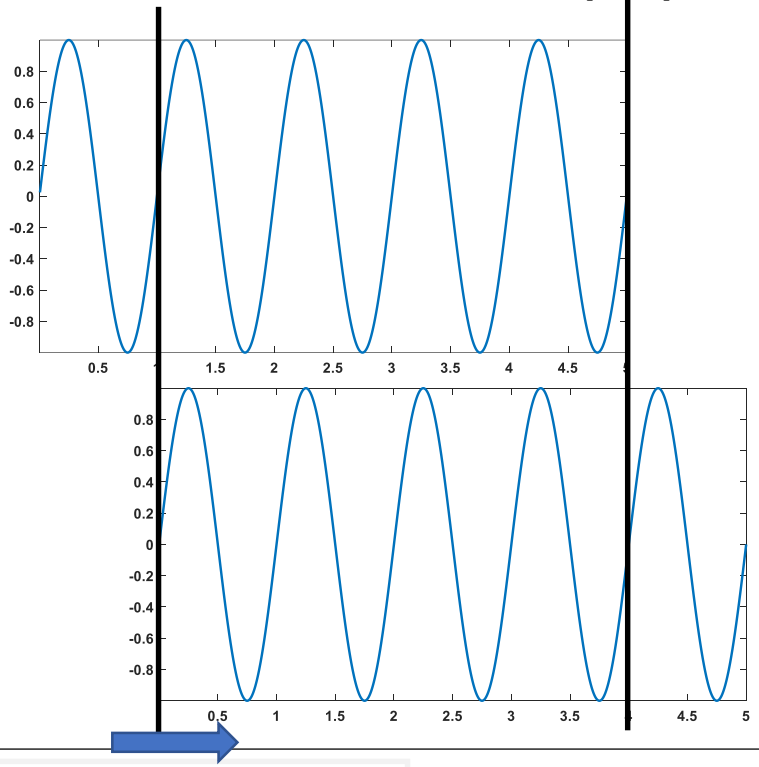


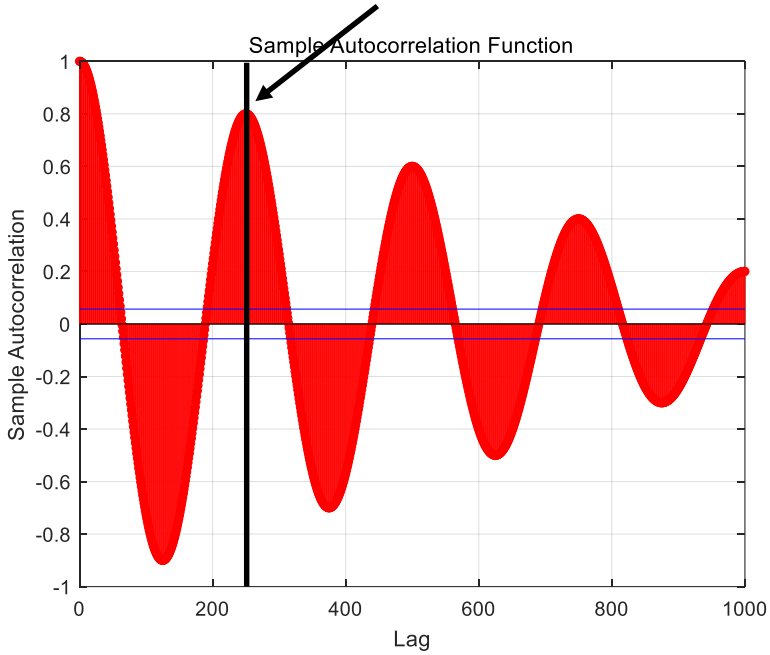
주기를 아는 방법

- Autocorrelation

$y(t) = sin(2\pi t)$

주기: 1초
주파수: 1/T= 1Hz

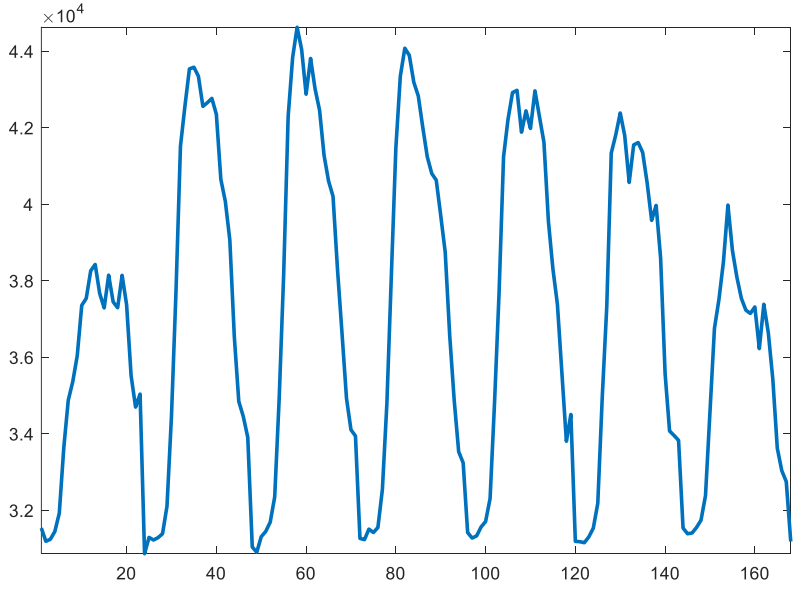




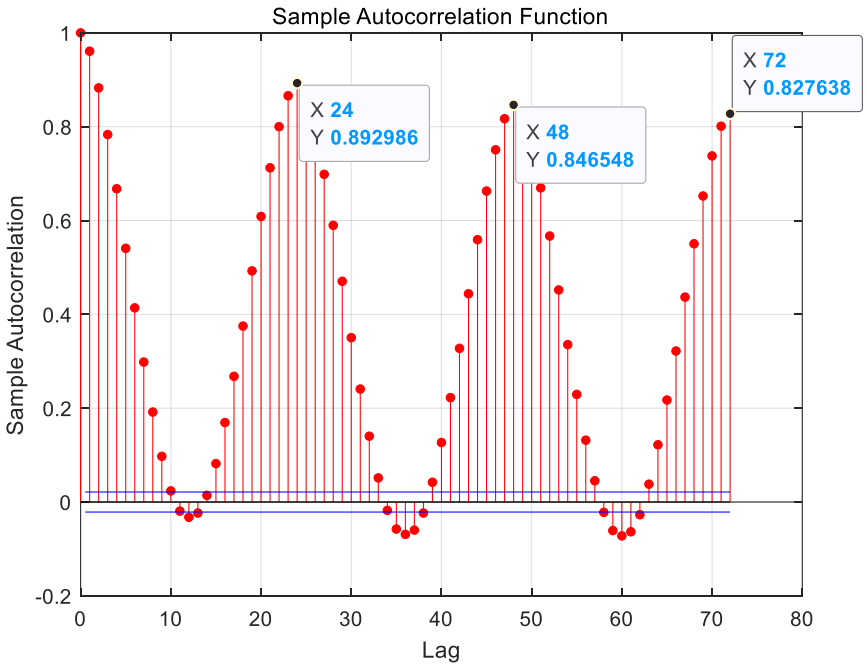
주기를 아는 방법

- Autocorrelation

2017년 생활인구 데이터 7일치

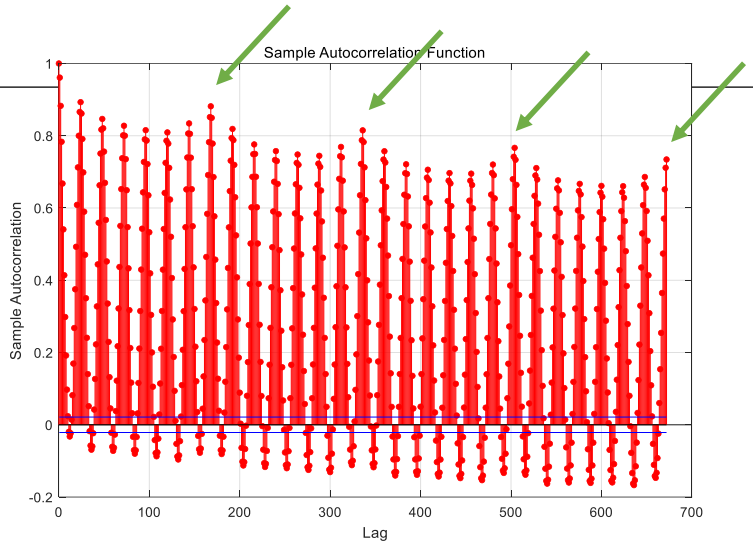


24시간 간격으로 주기가 반복

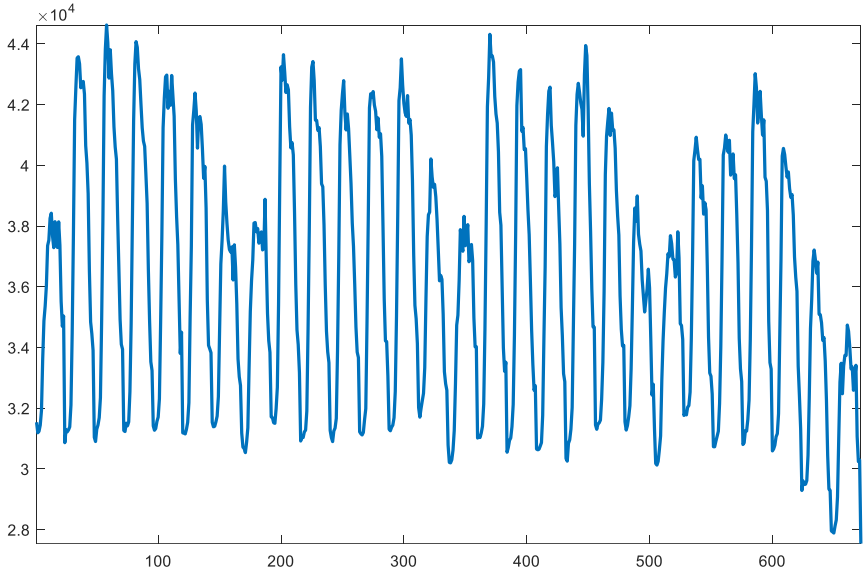


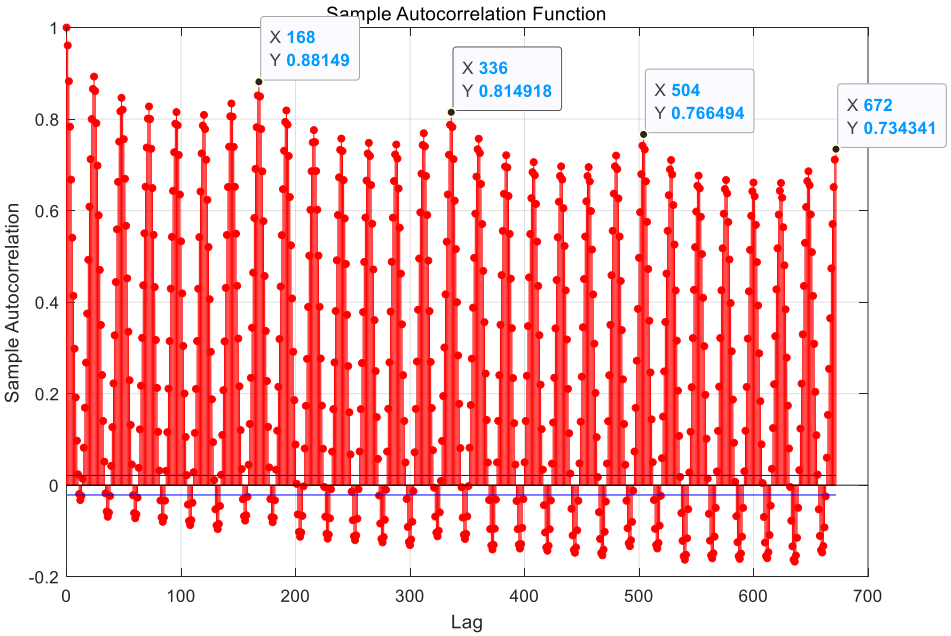
주기를 아는 방법

- Autocorrelation



2017년 생활인구 데이터 28일치





24*7 (1주일)시간 간격으로의 주기도 관찰

왜 24시간, 24*7시간을 shift했는지에 대한 답이었음

In [20]:
```
# 아래에 실습코드를 작성하고 결과를 확인합니다.

df_total['1d'] = df_total['총생활인구수'].shift(24)
df_total['7d'] = df_total['총생활인구수'].shift(24*7)
```

In [7]:
```
df_total.head(30)
```

	year	month	day	hour	종생활인구수	1d
0	2017	1	1	0	31535.2200	NaN
1	2017	1	1	1	31188.9174	NaN
2	2017	1	1	2	31240.4974	NaN
3	2017	1	1	3	31442.4314	NaN
4	2017	1	1	4	31922.7751	NaN
5	2017	1	1	5	33633.7304	NaN
6	2017	1	1	6	34876.8006	NaN
7	2017	1	1	7	35358.9775	NaN
8	2017	1	1	8	36038.7688	NaN
9	2017	1	1	9	37353.1794	NaN
10	2017	1	1	10	37534.7596	NaN
11	2017	1	1	11	38257.1671	NaN
12	2017	1	1	12	38423.5288	NaN
13	2017	1	1	13	37666.9073	NaN
14	2017	1	1	14	37287.4833	NaN
15	2017	1	1	15	38144.0804	NaN
16	2017	1	1	16	37444.9623	NaN
17	2017	1	1	17	37292.5709	NaN
18	2017	1	1	18	38139.0160	NaN
19	2017	1	1	19	37368.8302	NaN
20	2017	1	1	20	35517.1900	NaN
21	2017	1	1	21	34695.3430	NaN
22	2017	1	1	22	35035.7382	NaN
23	2017	1	1	23	30863.1777	NaN
24	2017	1	2	0	31290.0276	31535.2200
25	2017	1	2	1	31221.5248	31188.9174
26	2017	1	2	2	31283.4217	31240.4974
27	2017	1	2	3	31384.6021	31442.4314
28	2017	1	2	4	32104.6669	31922.7751
29	2017	1	2	5	34465.2673	33633.7304

	year	month	day	hour
0	2017	1	1	0
1	2017	1	1	1
2	2017	1	1	2
3	2017	1	1	3
4	2017	1	1	4
5	2017	1	1	5

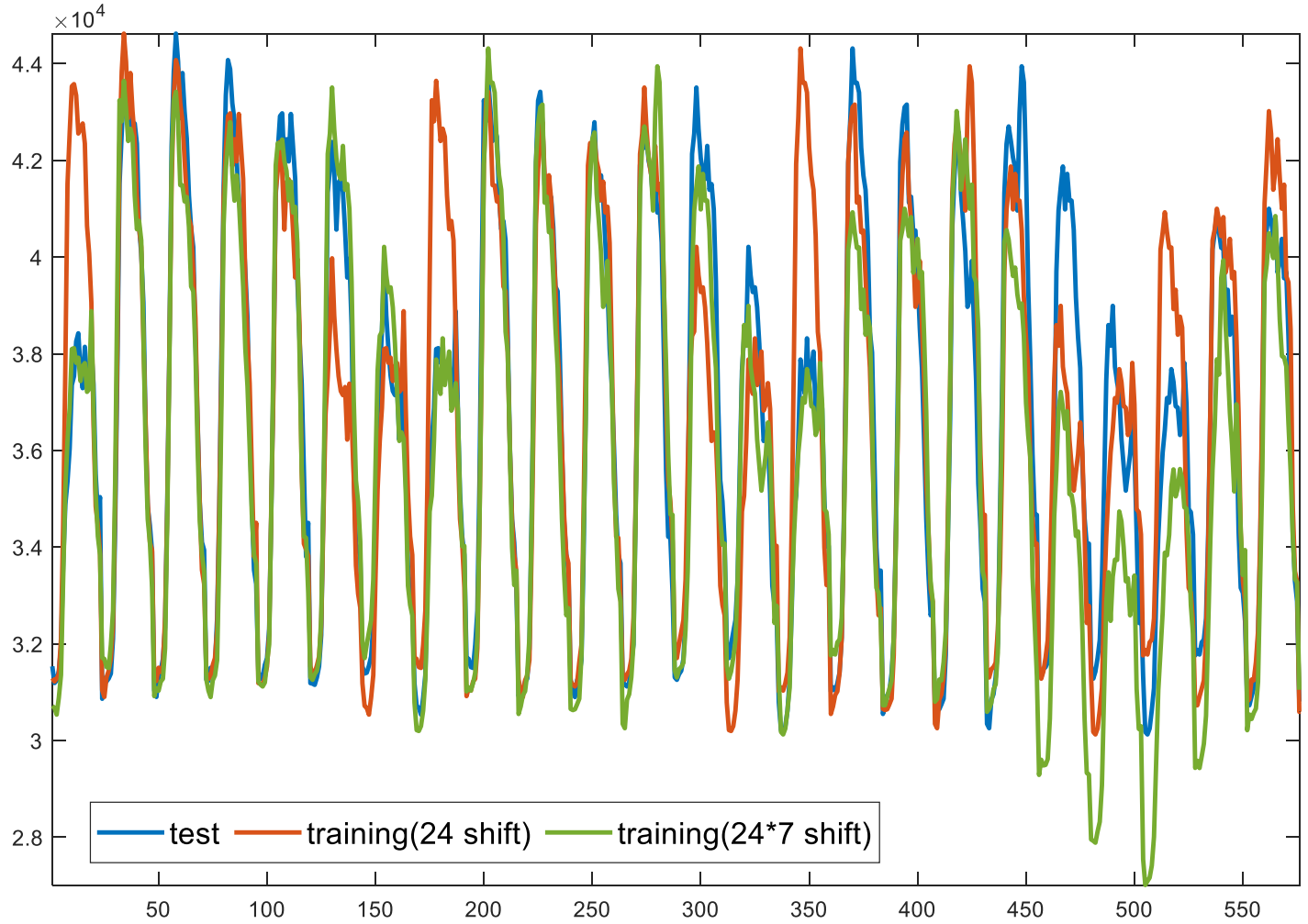
	year	month	day	hour	
24	2017	1	2	0	31290.0276
25	2017	1	2	1	31221.5248
26	2017	1	2	2	31283.4217
27	2017	1	2	3	31384.6021
28	2017	1	2	4	32104.6669
29	2017	1	2	5	34465.2673

Test data

1d		year	month	day	hour
31535.2200	0	2017	1	1	0
31188.9174	1	2017	1	1	1
31240.4974	2	2017	1	1	2
31442.4314	3	2017	1	1	3
31922.7751	4	2017	1	1	4
33633.7304	5	2017	1	1	5

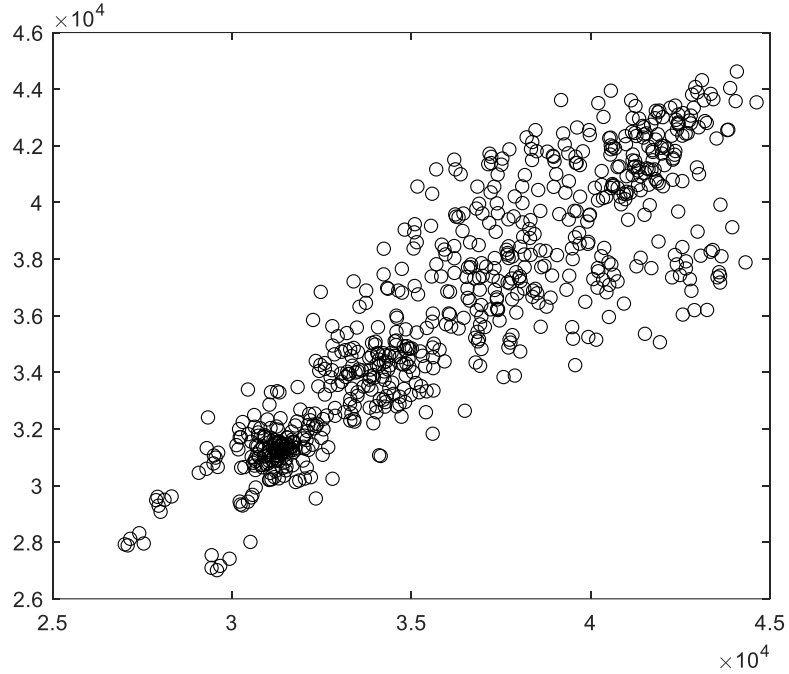
Training data 1

2017년 1월

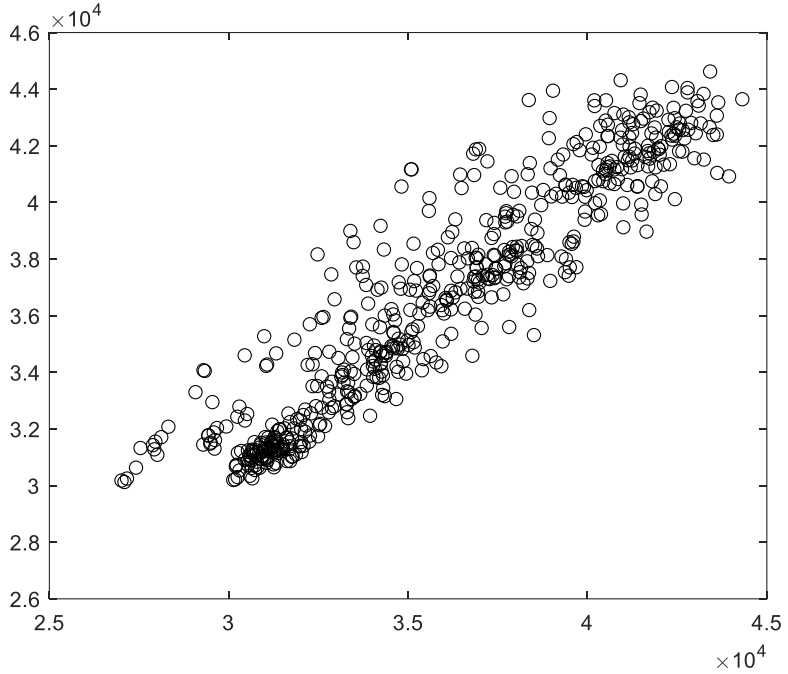


2017년 1월

생활인구

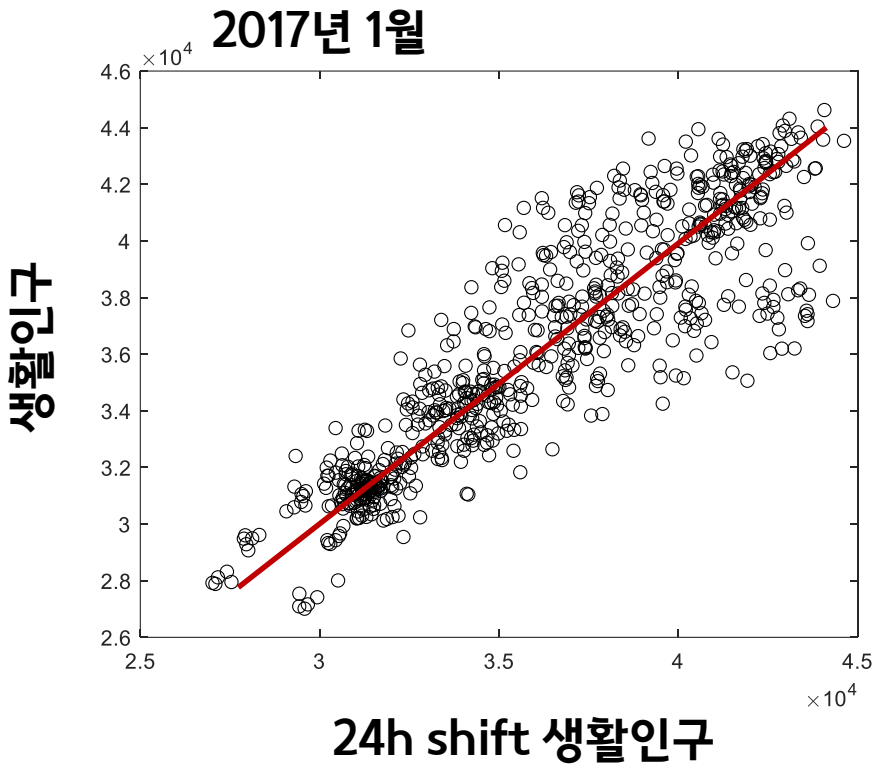


24h shift 생활인구



24*7h shift 생활인구

생활인구를 회귀분석으로 분석한다면?

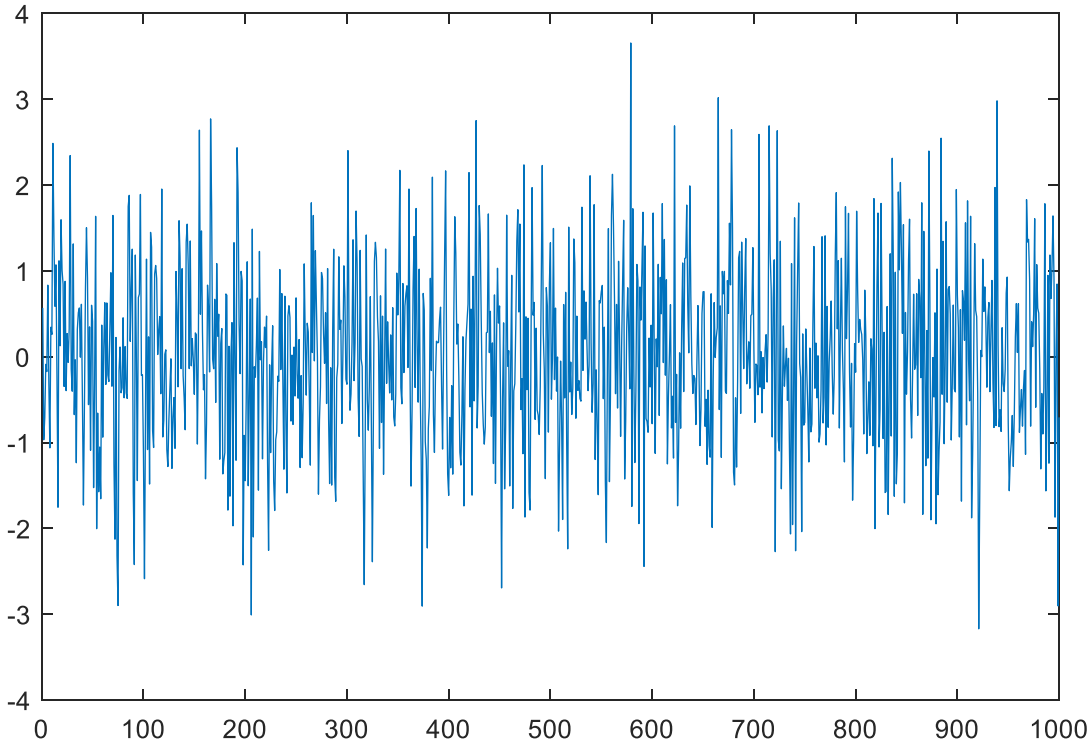


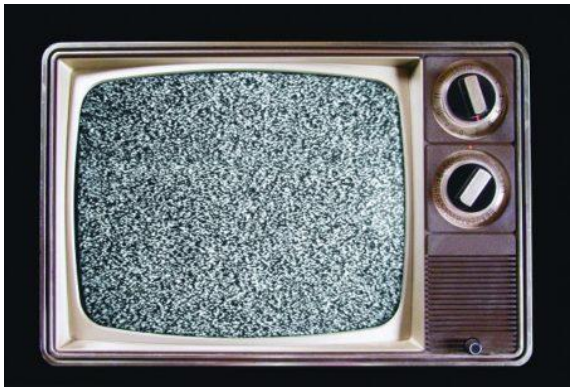
Y=ax+b의 a, b를 결정하면 됨

오차들 = 잔차

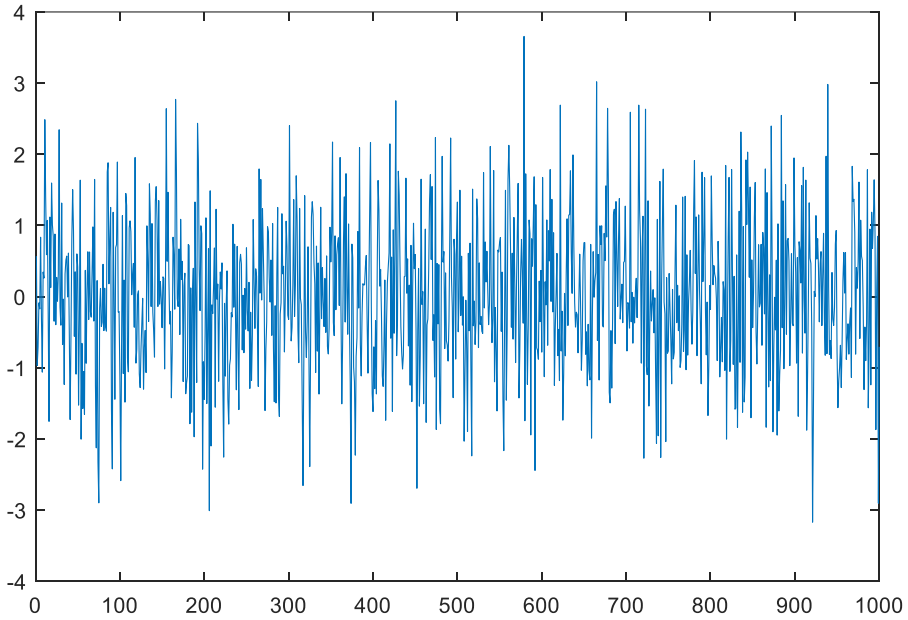
잔차는 white noise의 성질이 있다고 했었음

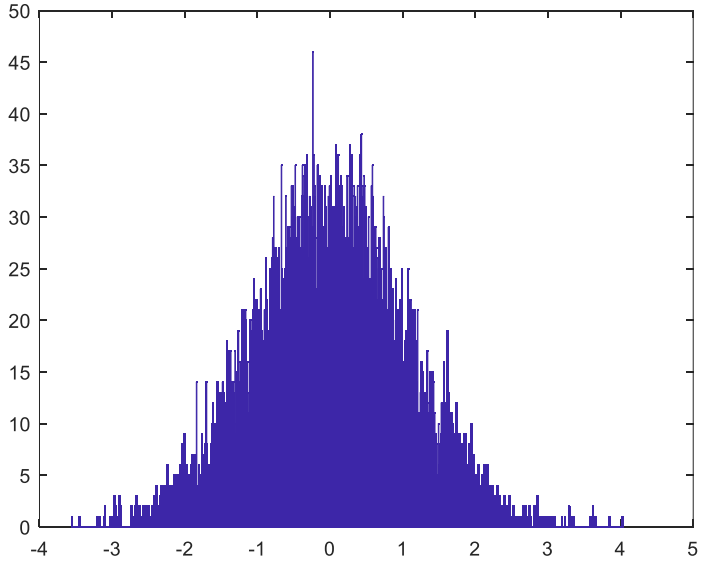
White noise





White noise





선형회귀에서의 가정

- **종속변수와 독립변수간에 선형적 연관성이 있어야 함**
  - 산점도를 통해 종속변수와 설명변수간 선형관계 확인

- **독립변수의 각각의 값에 대해서 종속변수의 값들이 정규분포를 따라야 함 (즉, 잔차의 정규성)**

- **종속 변수 값들의 분산이 모든 독립변수 값에 대해서 동일해야 함 (등분산)**

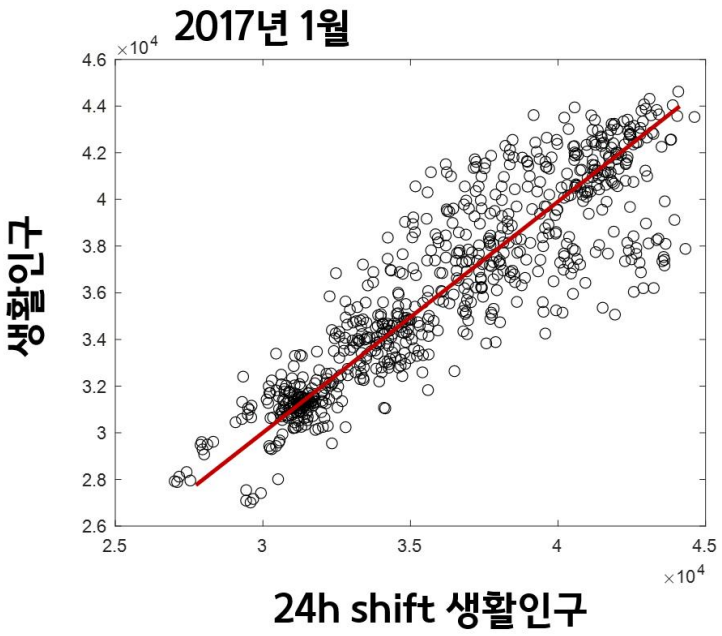
가정에 대한 확인

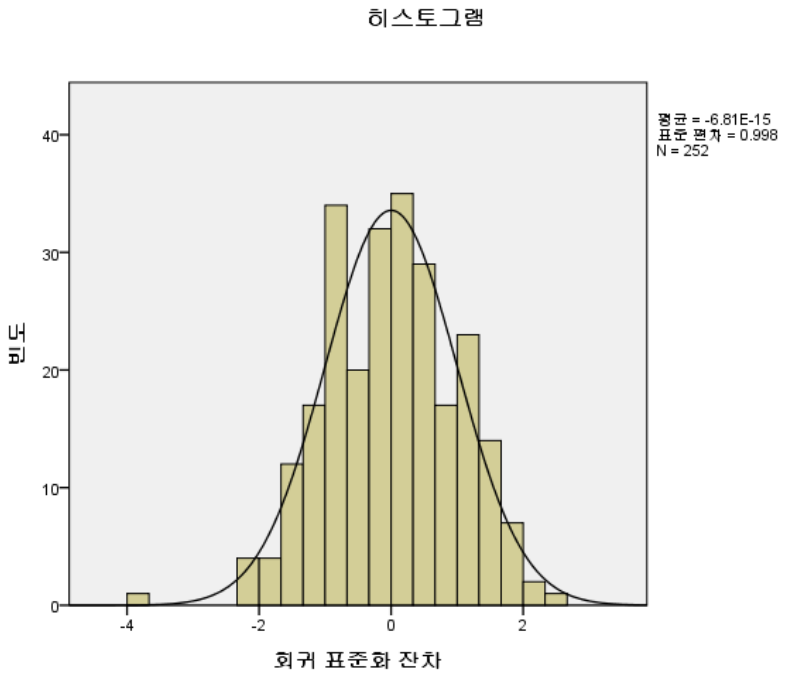
- 회귀 모형이 자료에 적합된 후에 선형 회귀의 가정들이 위배되지 않았는지 반드시 체크
- 만일 가정이 위배되었다면 선형 회귀 모형으로의 적합이 적절하지 않음

회귀 분석

정규성 확인: 잔차의 분포 확인

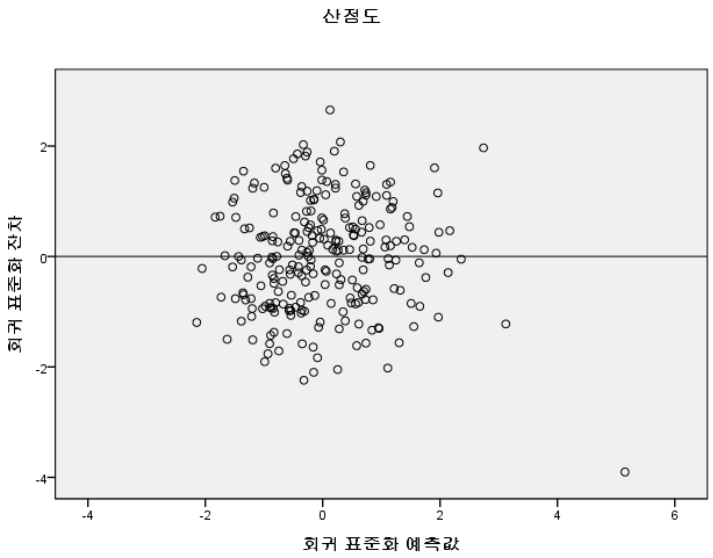
- 잔차의 히스토그램과 정규확률도를 그려봄으로써 정규성 확인



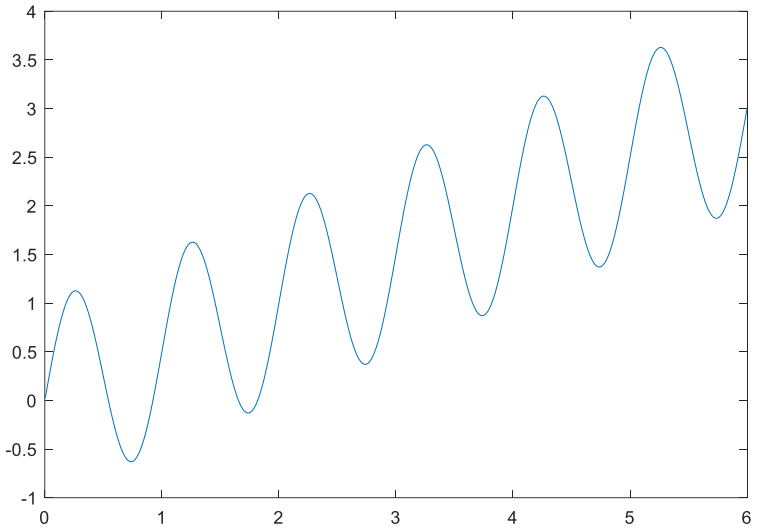


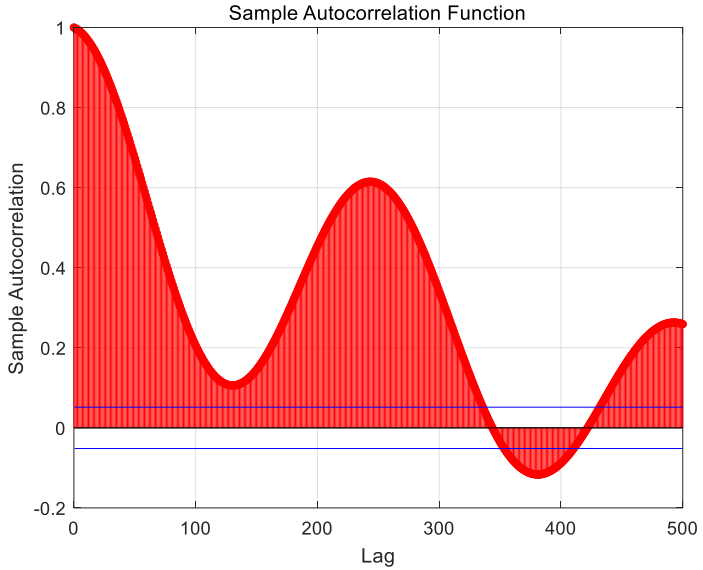
등분산성: 잔차도의 확인

- **잔차와 예측값의 산점도로 등분산성 및 선형성 평가**
  - 0 근처에 골고루 퍼짐
  - 곡선과 같은 특정패턴이 나타나지 않음
  - 예측 값이 증가함에 따라 잔차가 증가 혹은 감소 패턴 없음



$$y(t) = sin(2\pi t) + \frac{1}{2}t$$



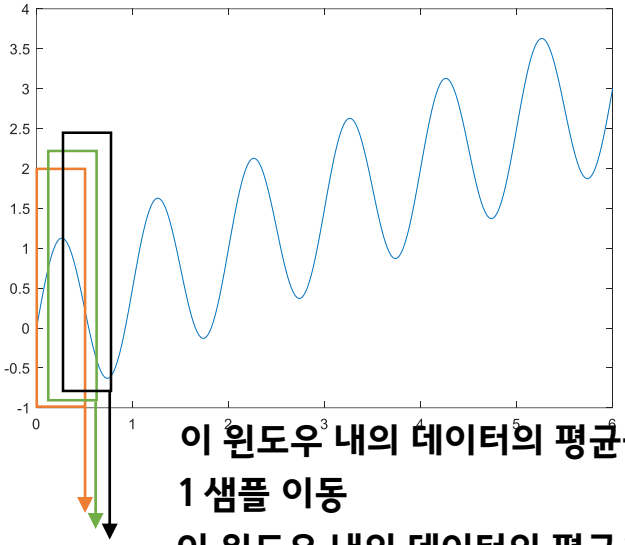


주기 (1초)도 있고, 시간에 따라 증가하는 패턴도 보임
이 패턴은 어떻게 알 수 있을까

이동평균 (moving window average)

- 윈도우: 분석하고자 하는 샘플 수 (시간)
- 이동평균: 해당 윈도우 내의 데이터 샘플의 평균하고, 윈도우를 1샘플씩 이동시켜가며 반복하는 방법

윈도우: 12시간 (12샘플)



이 윈도우 내의 데이터의 평균을 계산

1 샘플 이동

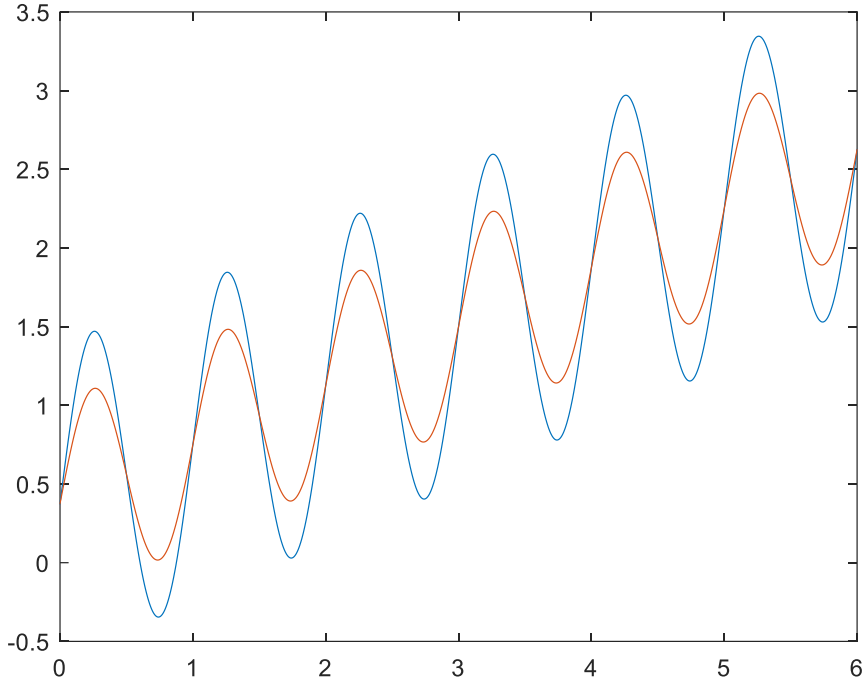
이 윈도우 내의 데이터의 평균을 계산

1 샘플 이동

⋮

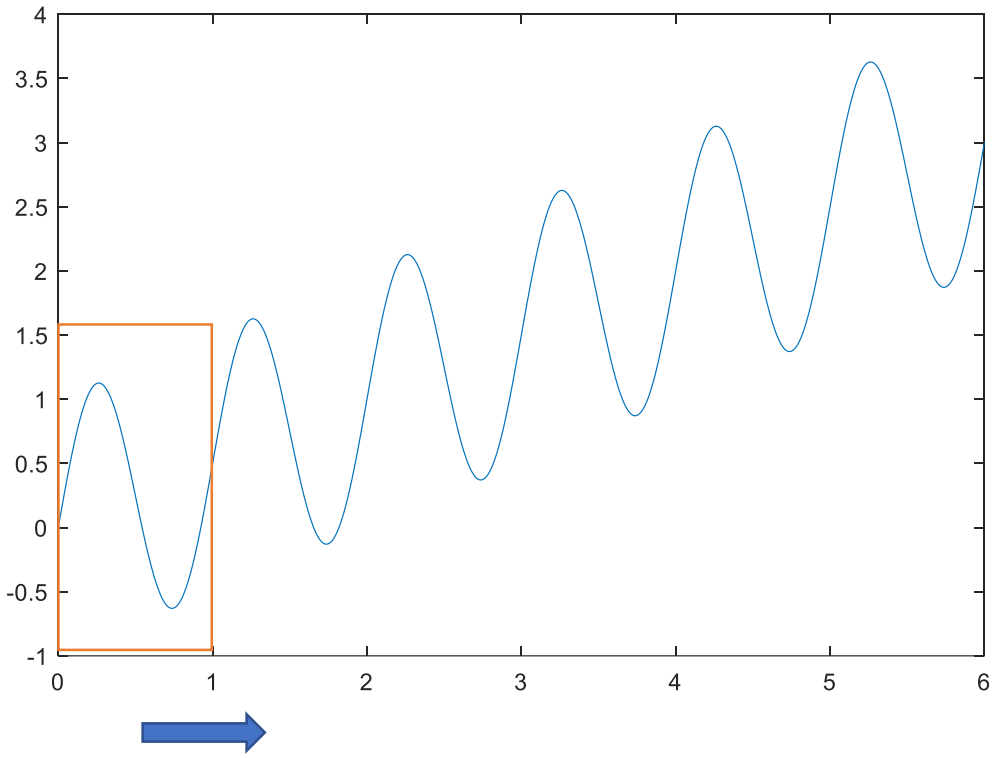
이동평균 (moving window average)

- 윈도우: 12시간 (12샘플), 1시간씩 이동시키며 평균을 구한 결과



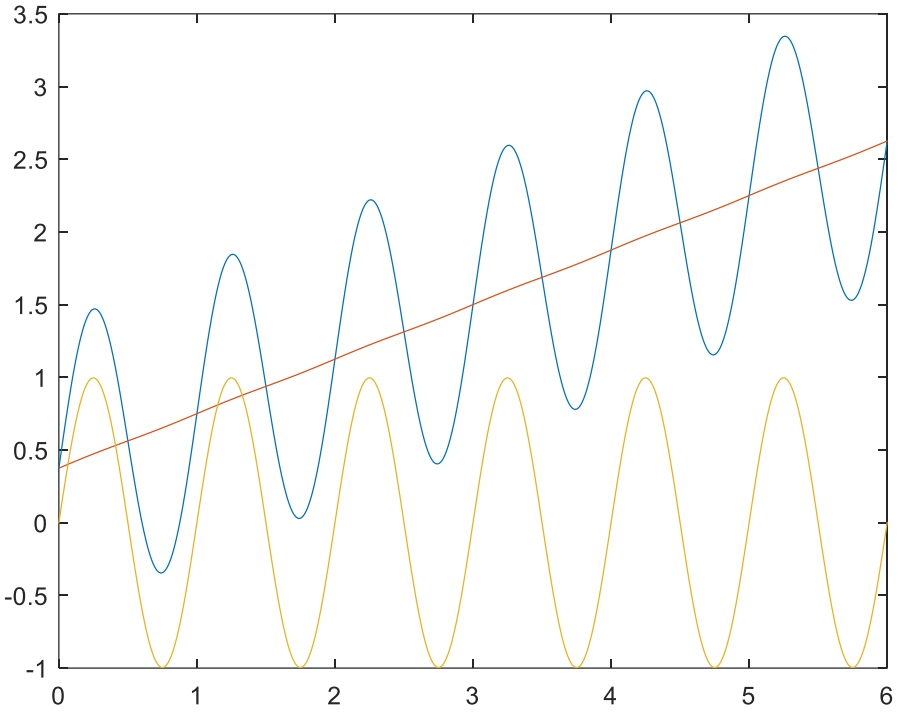
이동평균 (moving window average)

- 윈도우: 24시간 (24샘플), 1시간씩 이동시키며 평균을 구한다면?



이동평균 (moving window average)

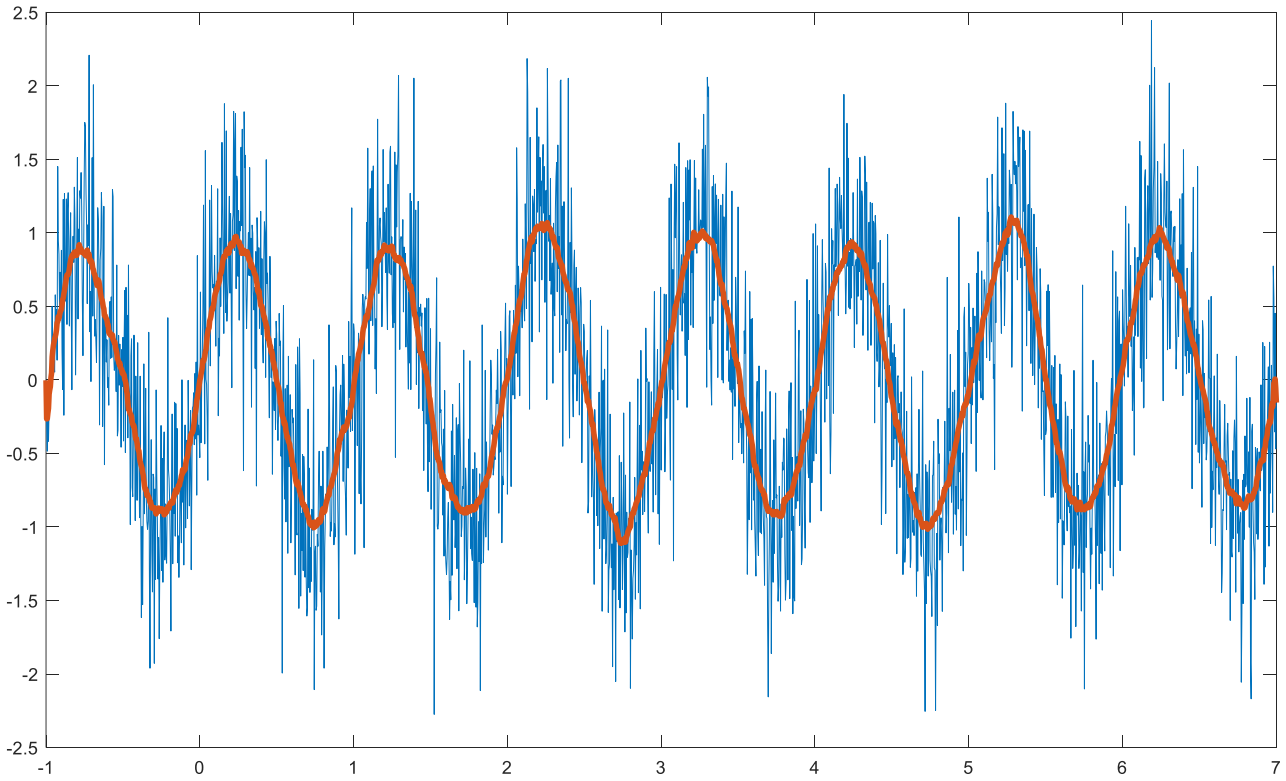
- 윈도우: 24시간 (24샘플), 1시간씩 이동시키며 평균을 구한다면?



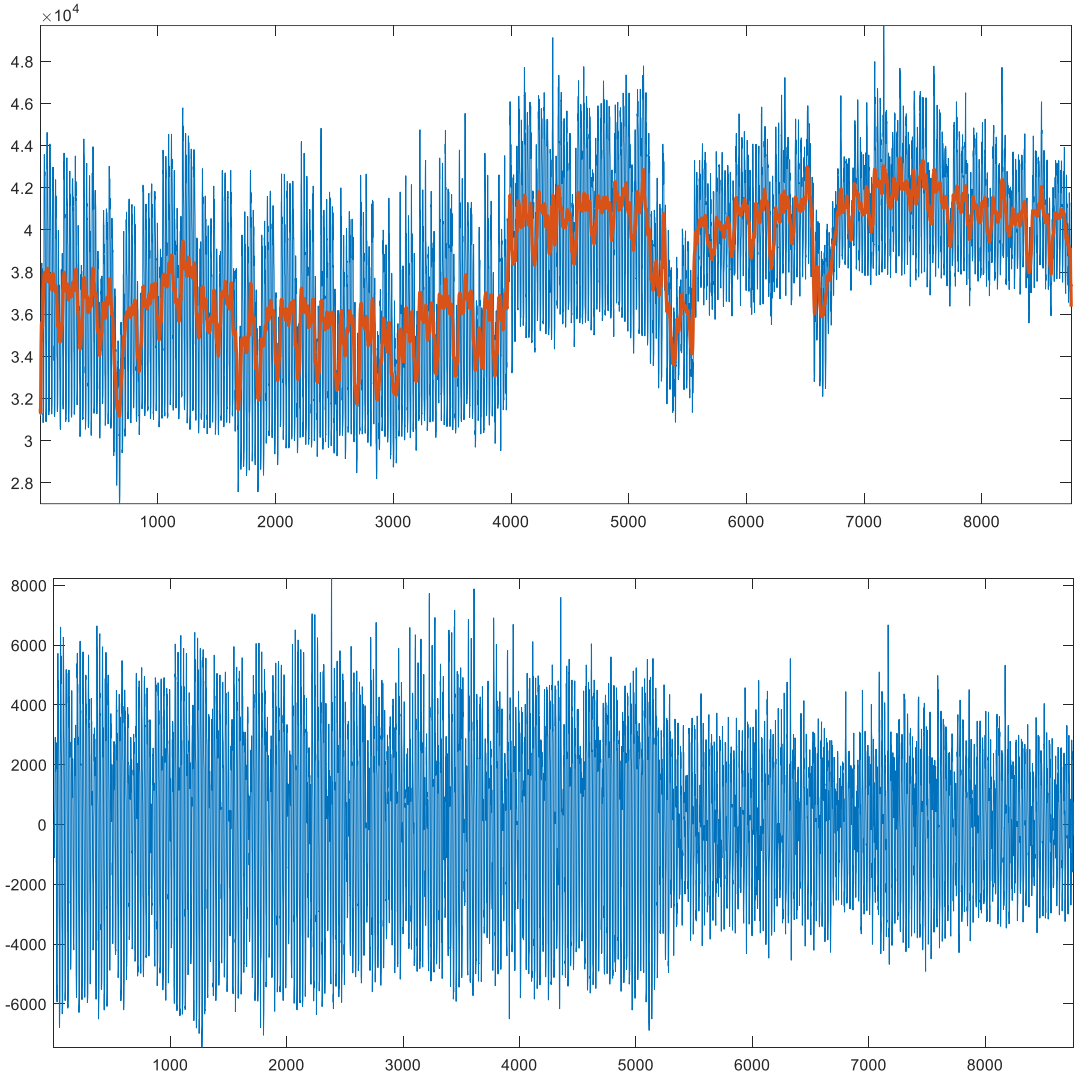
이동평균 (moving window average)

- 노이즈 제거에도 활용

Window size 0.2초

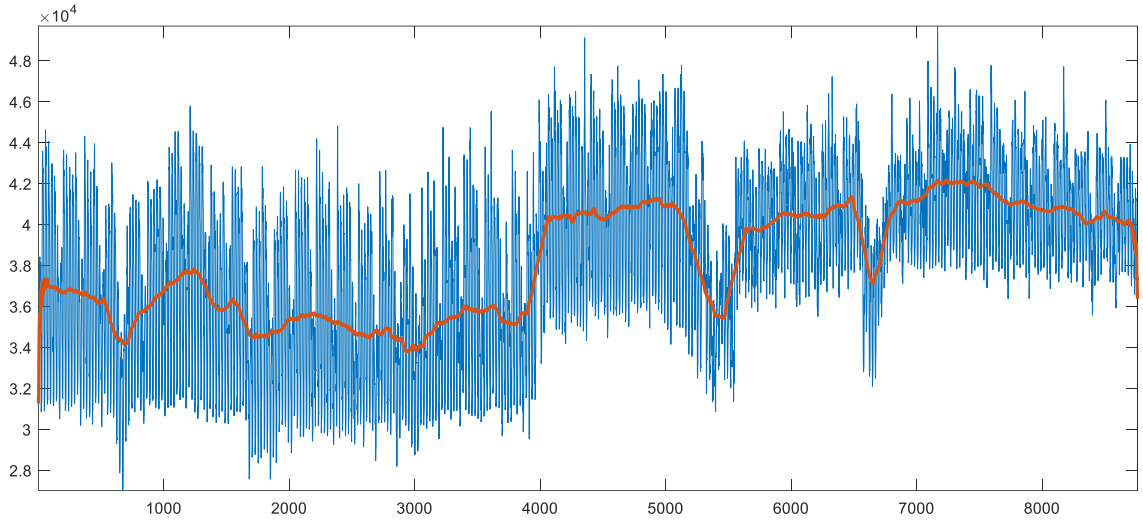


이동평균 (moving window average)

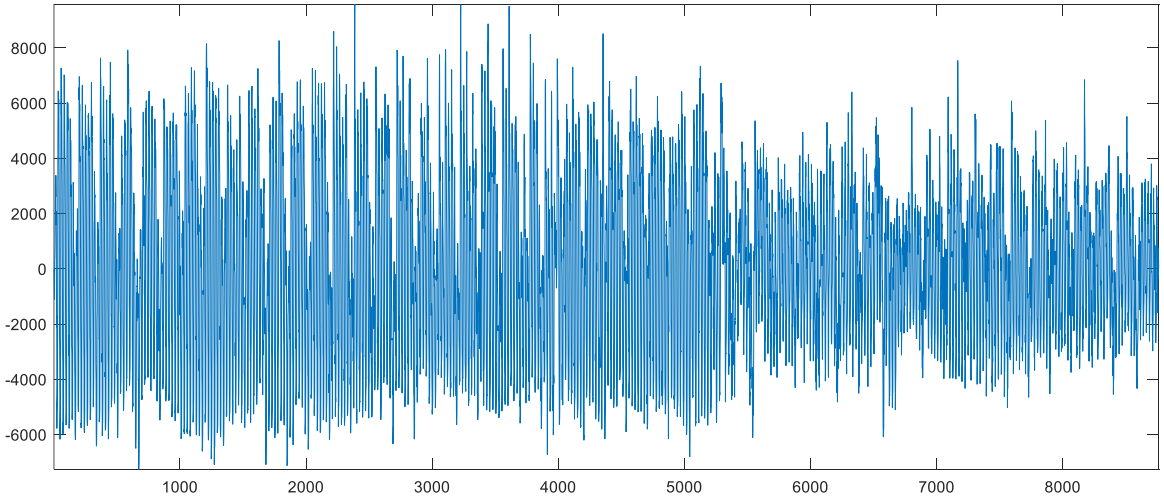


Window size: 24시간

이동평균 (moving window average)



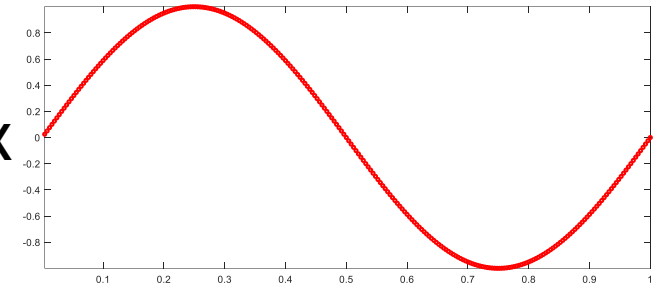
Window size: 24*7시간



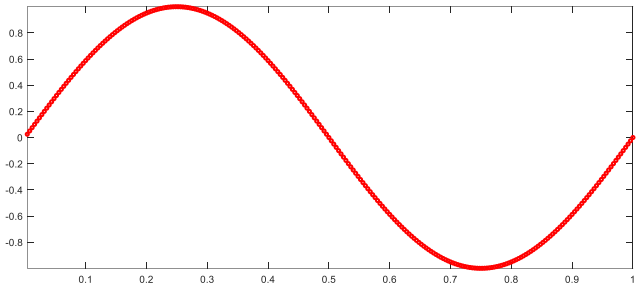
Regression vs. auto regressive model

Correlation vs. autocorrelation

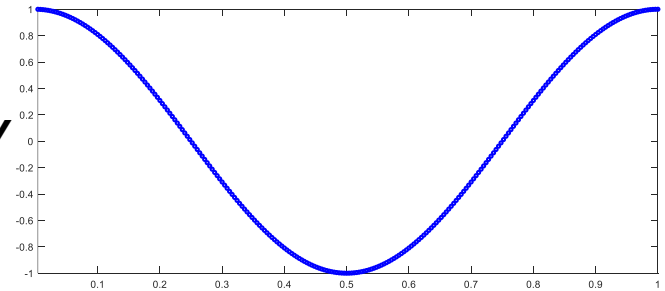
X



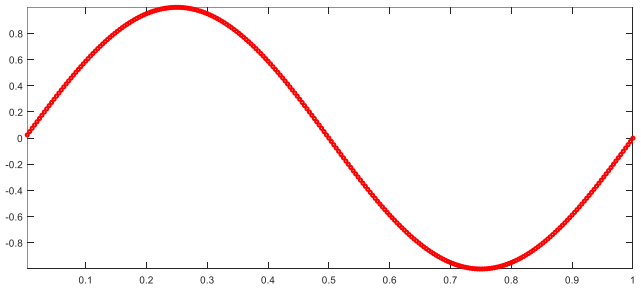
X



Y



X



Regression vs. auto regressive (AR) model

X로 Y를 추정하는 것
Y = aX + b

Lag (time delay) 없음

거주인원(X)으로 행정구의 면적 (Y)을 추정하겠다

X로 X를 추정하는 것

Lag (time delay) 있음

오늘 생활인구로 내일 생활인구를 추정하겠다
오늘 (월) 생활인구로 다음주 월요일 생활인구를 추정하겠다

auto regressive (AR) model

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

	year	month	day	hour	총생활인구수	
0	2017	1	1	0	31535.2200	
1	2017	1	1	1	31188.9174	31535.2200
2	2017	1	1	2	31240.4974	31188.9174
3	2017	1	1	3	31442.4314	31240.4974
4	2017	1	1	4	31922.7751	31442.4314
5	2017	1	1	5	33633.7304	31922.7751
6	2017	1	1	6	34876.8006	33633.7304
7	2017	1	1	7	35358.9775	34876.8006
8	2017	1	1	8	36038.7688	35358.9775
9	2017	1	1	9	37353.1794	36038.7688
10	2017	1	1	10	37534.7596	37353.1794
11	2017	1	1	11	38257.1671	37534.7596
12	2017	1	1	12	38423.5288	38257.1671
13	2017	1	1	13	37666.9073	38423.5288

	year	month	day	hour	총생활인구수		
0	2017	1	1	0	31535.2200		
1	2017	1	1	1	31188.9174	31535.2200	
2	2017	1	1	2	31240.4974	31188.9174	31535.2200
3	2017	1	1	3	31442.4314	31240.4974	31188.9174
4	2017	1	1	4	31922.7751	31442.4314	31240.4974
5	2017	1	1	5	33633.7304	31922.7751	31442.4314
6	2017	1	1	6	34876.8006	33633.7304	31922.7751
7	2017	1	1	7	35358.9775	34876.8006	33633.7304
8	2017	1	1	8	36038.7688	35358.9775	34876.8006
9	2017	1	1	9	37353.1794	36038.7688	35358.9775
10	2017	1	1	10	37534.7596	37353.1794	36038.7688
11	2017	1	1	11	38257.1671	37534.7596	37353.1794
12	2017	1	1	12	38423.5288	38257.1671	37534.7596
13	2017	1	1	13	37666.9073	38423.5288	38257.1671

Moving Average (MA) model

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t,$$

ARMA

Auto Regressive (AR) model

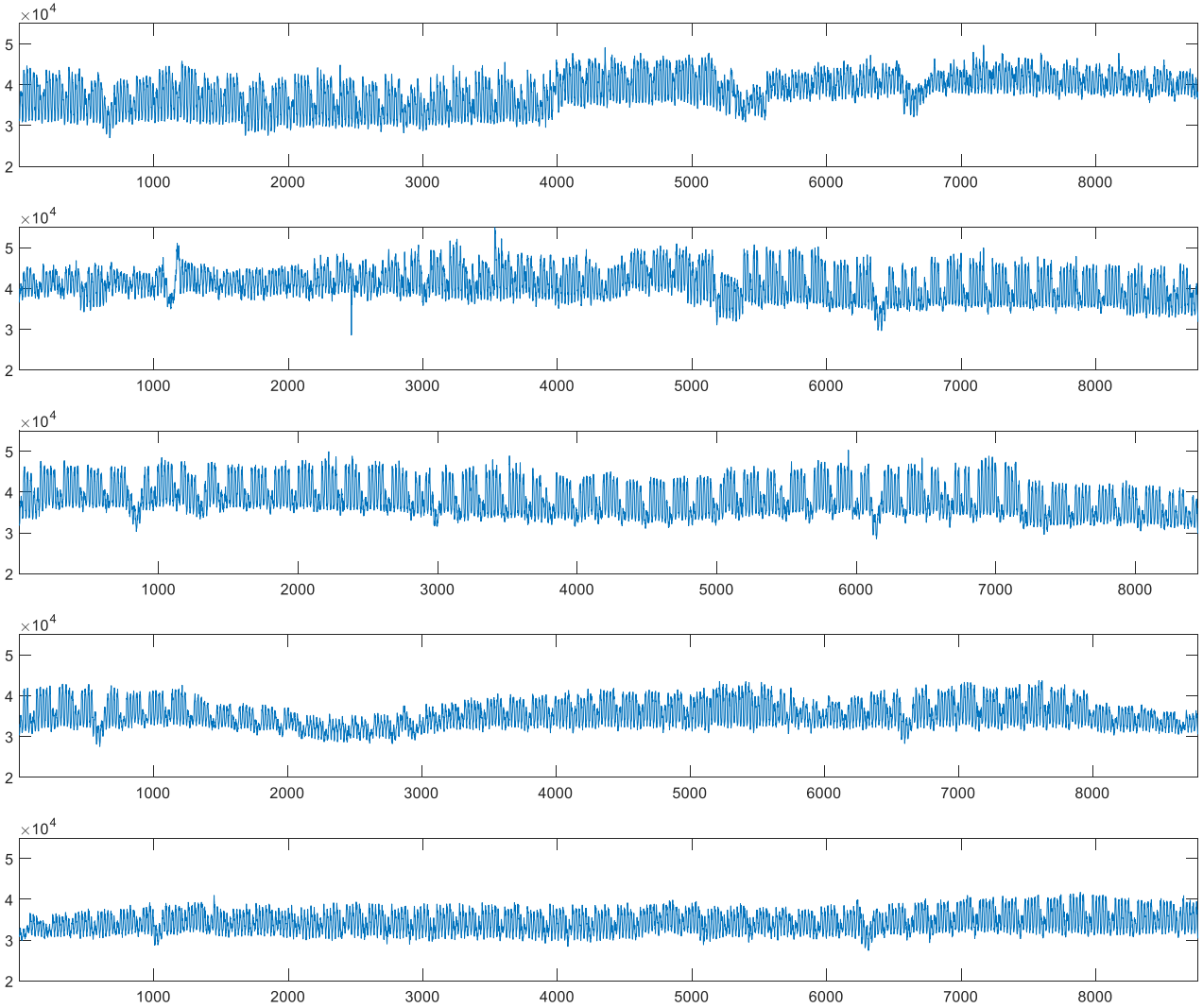
+

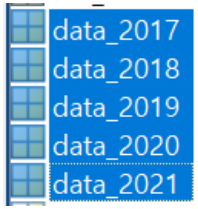
Moving Average (MA) model

ARIMA

autoregressive integrated moving average

그림을 많이 그려보자



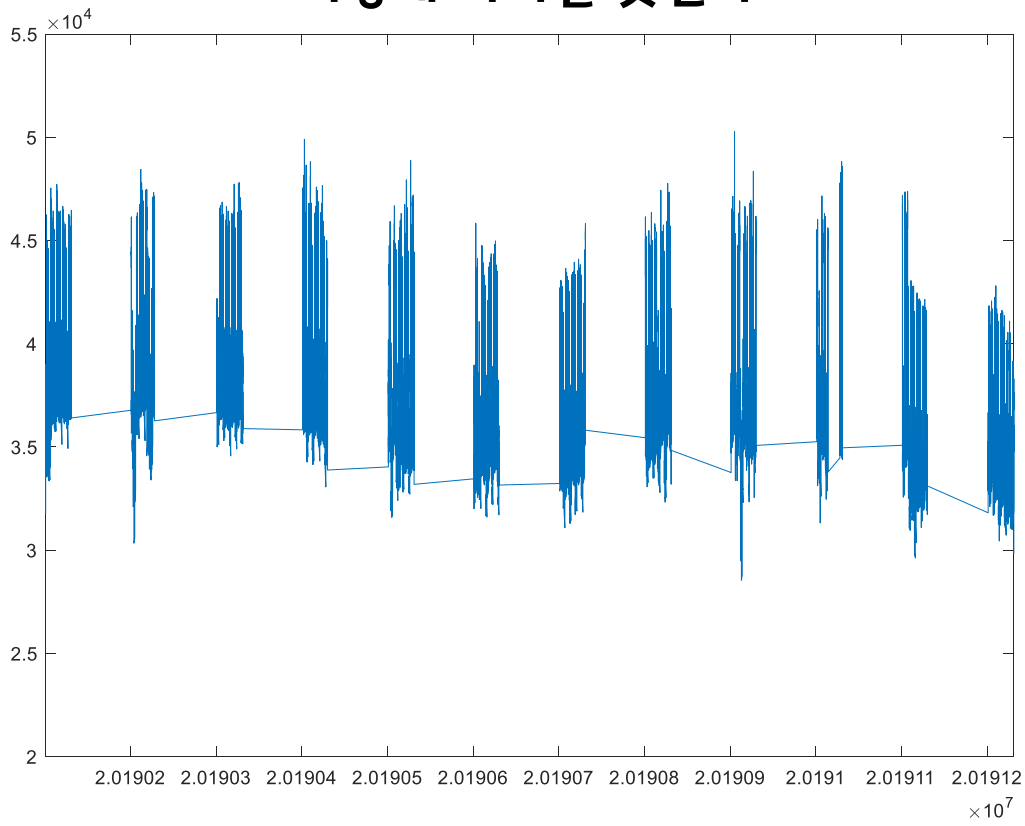


8760x3 double
8760x3 double
8448x3 double
8784x3 double
8760x3 double

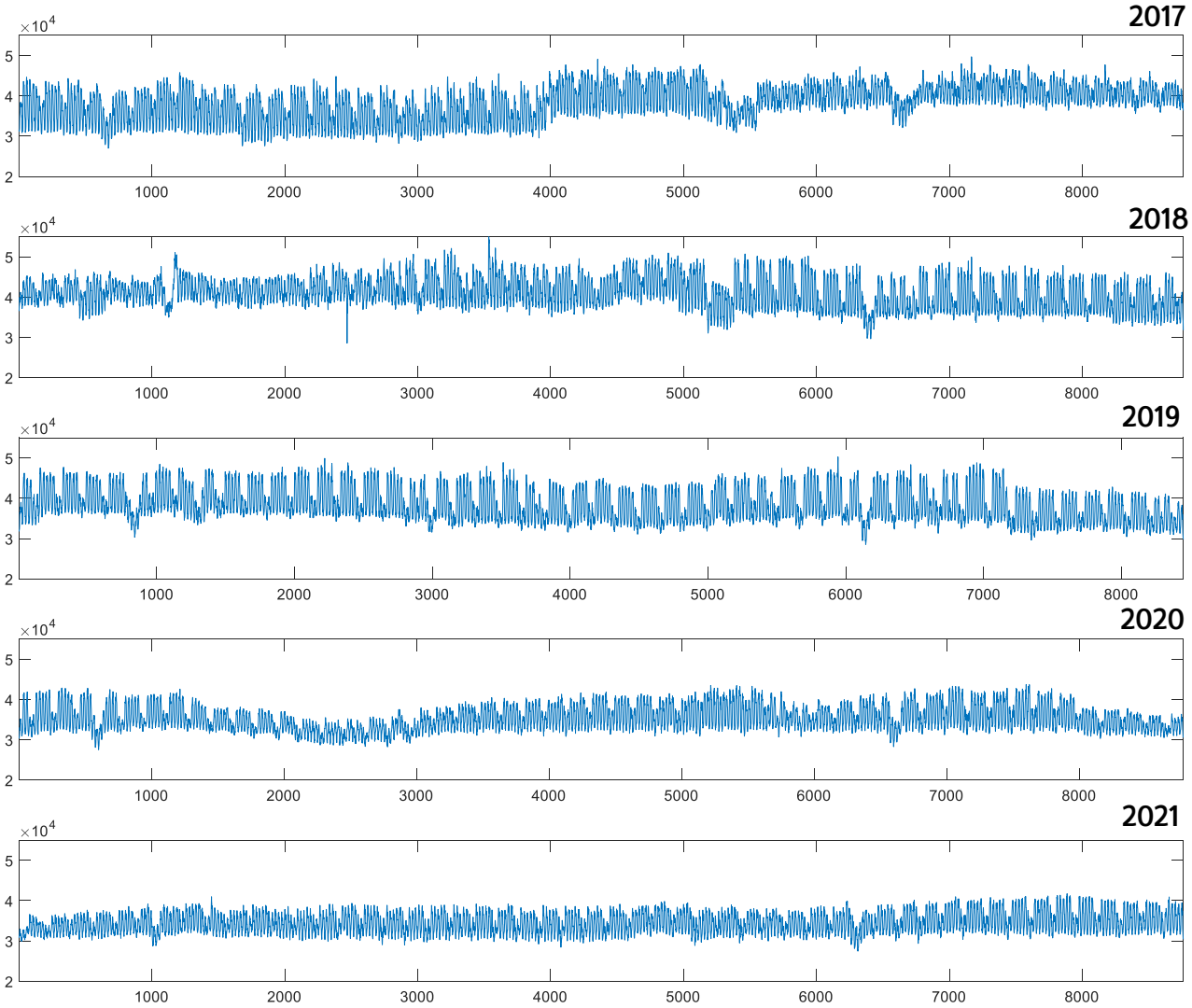
?

결측 데이터가 있다!

어떻게 처리할 것인가?



9	20191014	14	4.5446e+04
0	20191014	15	4.4438e+04
1	20191014	16	4.1662e+04
2	20191014	17	4.1124e+04
3	20191014	18	3.9080e+04
4	20191014	19	3.7066e+04
5	20191014	20	3.5238e+04
6	20191014	21	3.4519e+04
7	20191014	22	3.4339e+04
8	20191014	23	3.3770e+04
9	20191028	0	3.4466e+04
0	20191028	1	3.4671e+04
1	20191028	2	3.5038e+04
2	20191028	3	3.5020e+04
3	20191028	4	3.5522e+04
4	20191028	5	3.7969e+04
5	20191028	6	4.0003e+04
6	20191028	7	4.2298e+04



코로나는 언제 발생?

여행은 언제부터 가능?

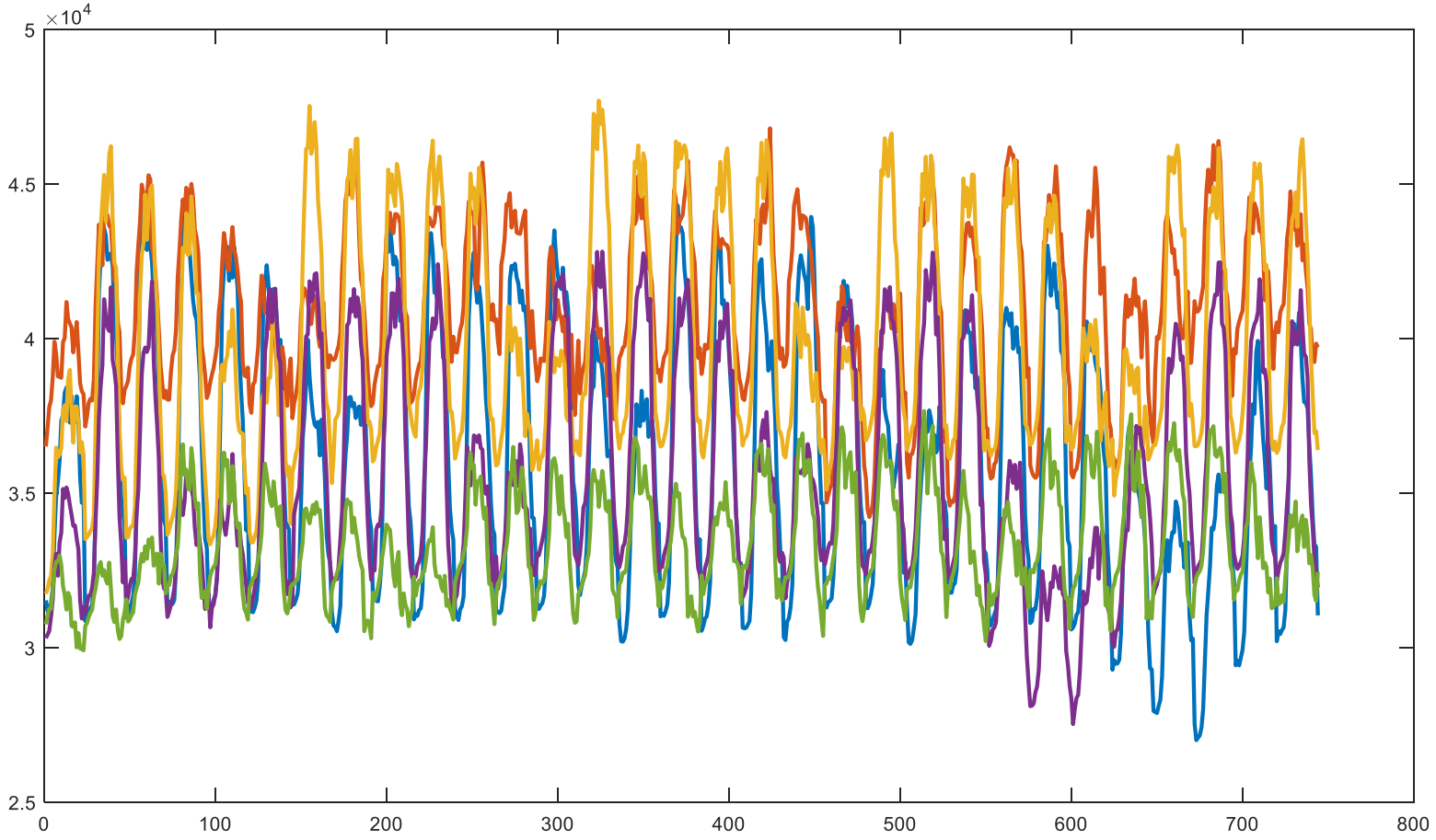
휴가철은 언제?

명절, 공휴일?

1월 1일은 무슨 요일?

윤달?

5년치 1월데이터 한번에 그려보기



- 어떤 모델을 사용할 것인가?
- 데이터 전처리를 어떻게 할 것인가?

- **모델링의 꽃은 데이터 전처리**

Thank you.