2020 Fall
예측모델 Project

안암로터리 따릉이 정류장 이용 형태에 관한 분석

2020011135 소규성 2020011132 정의석 2020011136 김지나

Contents

01

Data

Exploratory Data Analysis and

Data Preprocessing

02

Forecasting

Univariate time series,

ARIMA and Multivariate Data Analysis

03

Hidden Markov Model

Analyze and Forecast

Using Hidden Markov Model

04

Result

Conclusion and Insights

1. Data

01 **Data**

Exploratory Data Analysis and Data Preprocessing

02 **Forecasting**

Univariate time series,

ARIMA and Multivariate Data Analysis

03 **Hidden Markov Model**

Analyze and Forecast

Using Hidden Markov Model

04 **Result**

Conclusion and Insights

안암 로터리의 따릉이 하루 대여량은 도대체 얼마일까?

텅 비어 있는 따릉이 정류장

1. Data

사용 데이터 소개

기간 : 2017.06.21 – 2019.12.31

서울 열린데이터 광장

(data.seoul.go.k)

기상기후 빅데이터 분석 플랫폼

날씨마루

(bd.kma.go.kr)

- ✓ 서울시 공공자전거 이용현황
- ✓ 공공자전거 대여소 및 자전거 정보

- ✓ 기상 데이터

  강수량, 풍속, 풍향, 미세먼지

1. Data

Variables

- Data

date	cnt_station	cnt_bike	is_rain	is_holiday	cnt
2017-06-21	743	9855	0	0	13.0
2017-06-22	743	9855	0	0	40.0
2017-06-23	790	10415	0	0	39.0
2017-06-24	790	10415	1	0	28.0
2017-06-25	792	10455	0	0	28.0
...	...	...	...	...	...
2019-12-27	1530	19474	0	0	81.0
2019-12-28	1530	19474	0	0	79.0
2019-12-29	1530	19474	0	0	61.0
2019-12-30	1530	19474	0	0	77.0
2019-12-31	1530	19474	0	0	39.0

924 rows × 10 columns

- Input

Date

Date: 2017.06.21 – 2019.12.31

Is holiday : 공휴일 정보
(0 : 공휴일 X / 1: 공휴일 O)

외부 변수

Station Count : 해당 시점 총 정류장 개수
(numerical value)

Bike Count : 해당 시점 총 자전거 개수
(numerical value)

Is rain : 비(강수량≦3mm) 정보
(0 : 비 X / 1: 비 O)

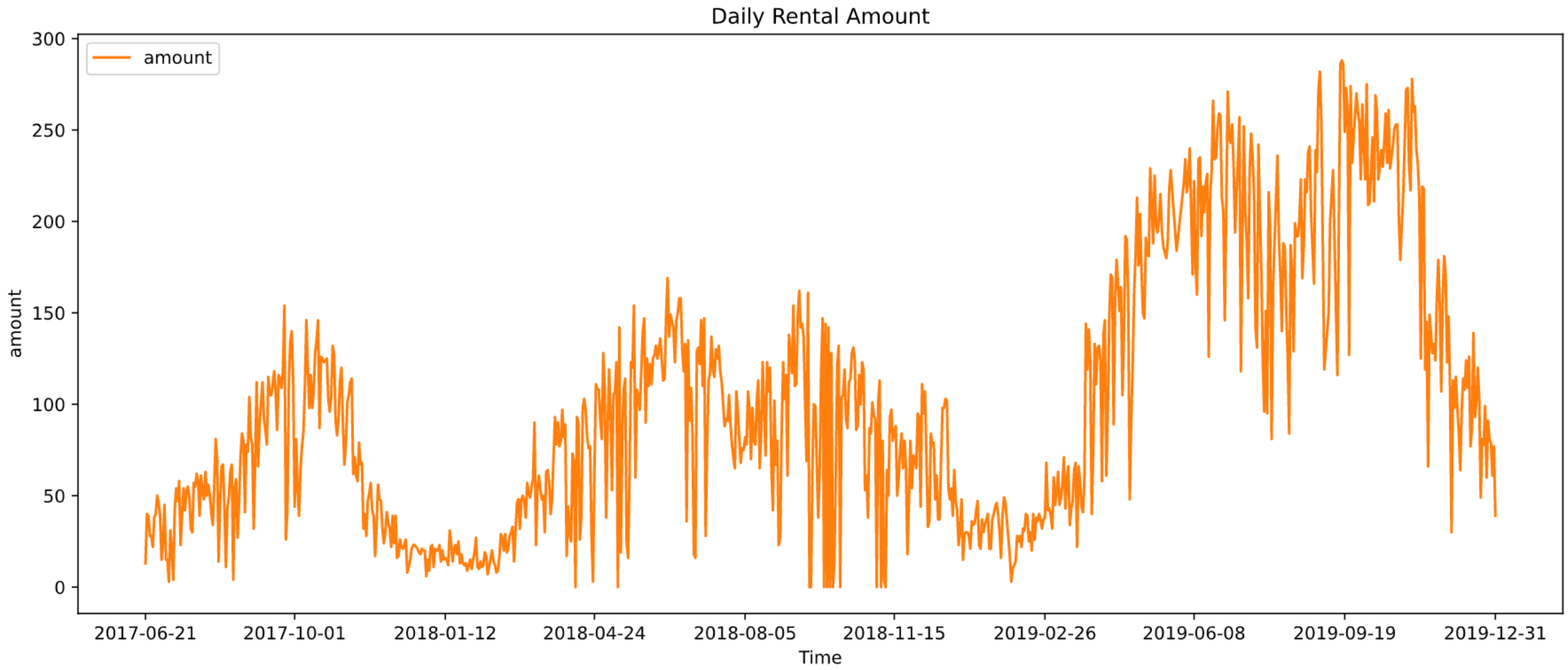
- Target

✓ 안암로터리 정류장 일일 따릉이 대여량

1. Data

안암로터리 정류장 일일 따릉이 대여량

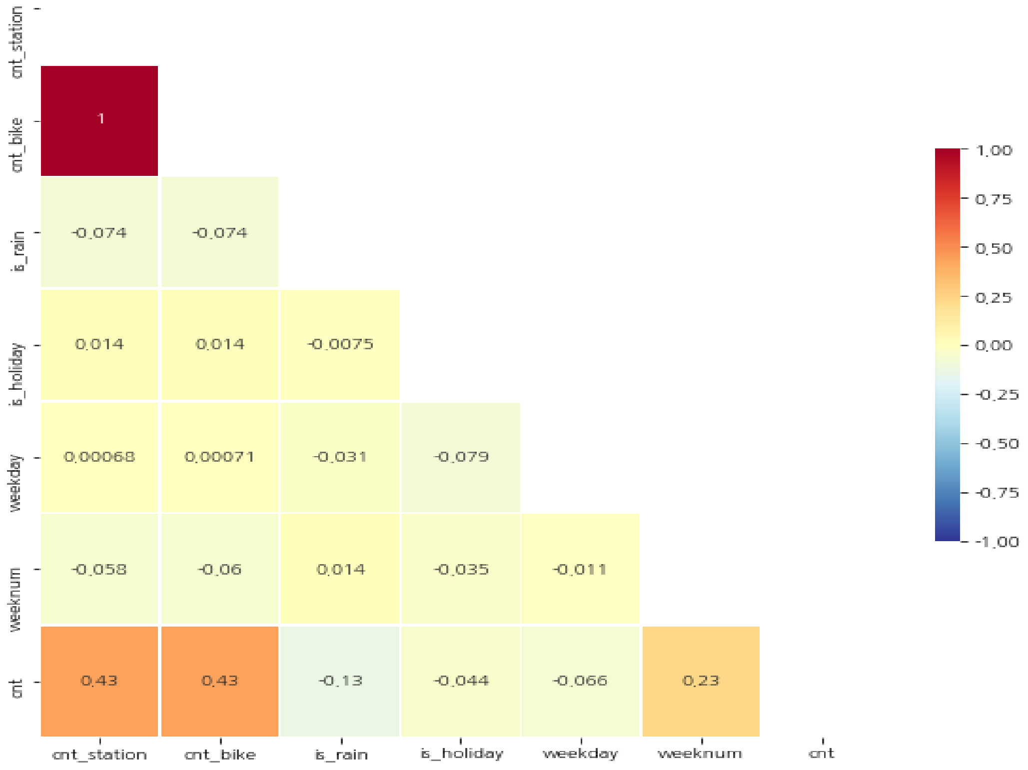
2017-06-21~2019-12-31



1. Data

Exploratory Data Analysis

변수 별 상관관계



Target(일별 대여량)과 높은 상관계수를 보이는 변수

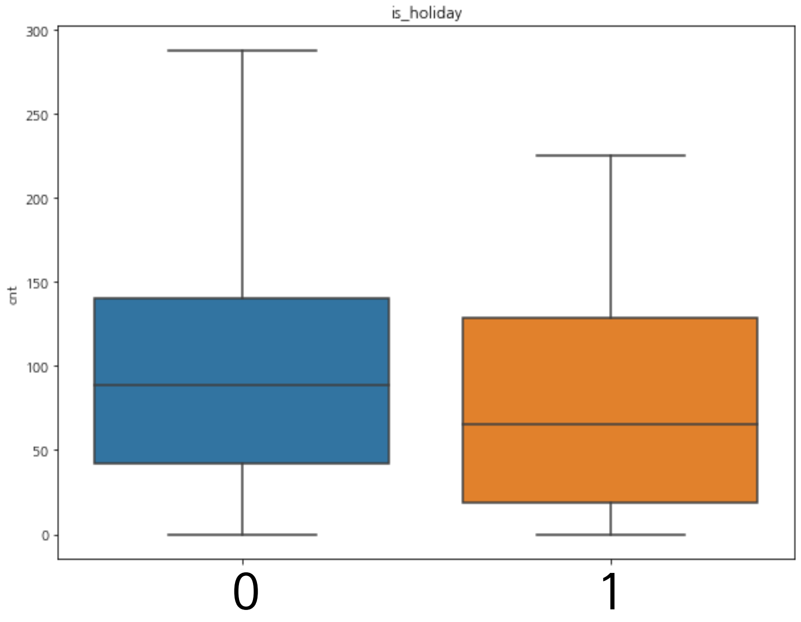
- 강한 양의 상관계수 (r ≧ 0.4)
  - ✓ 전국 배치된 따릉이(자전거)의 수
  - ✓ 전국 설치된 정거장의 수
- 약한 양의 상관계수 (0.1≦ r 〈 0.4)
  - ✓ 연중 주차 수

1. Data

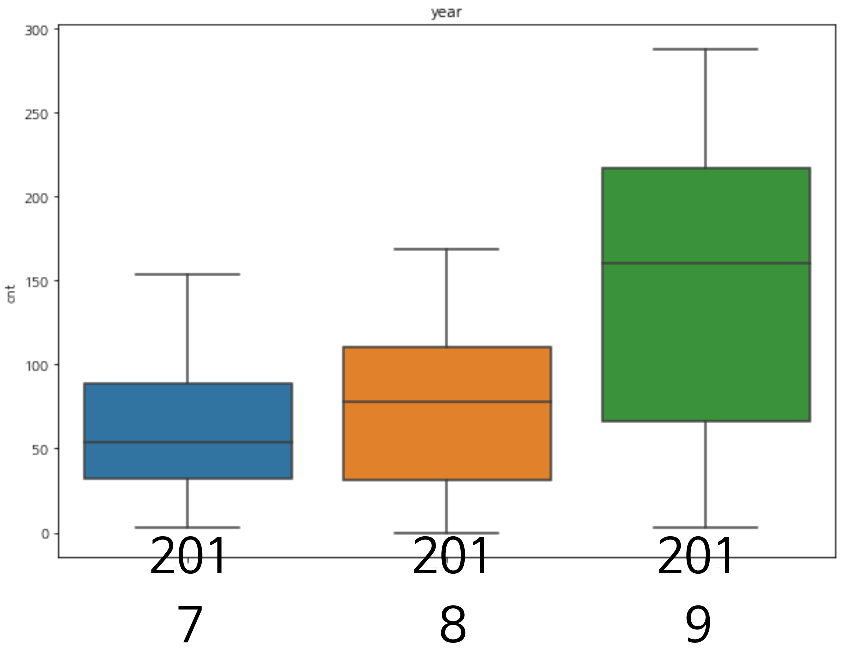
Exploratory Data Analysis

날짜 변수와 일일 대여량

- 공휴일과 일일 대여량 분포



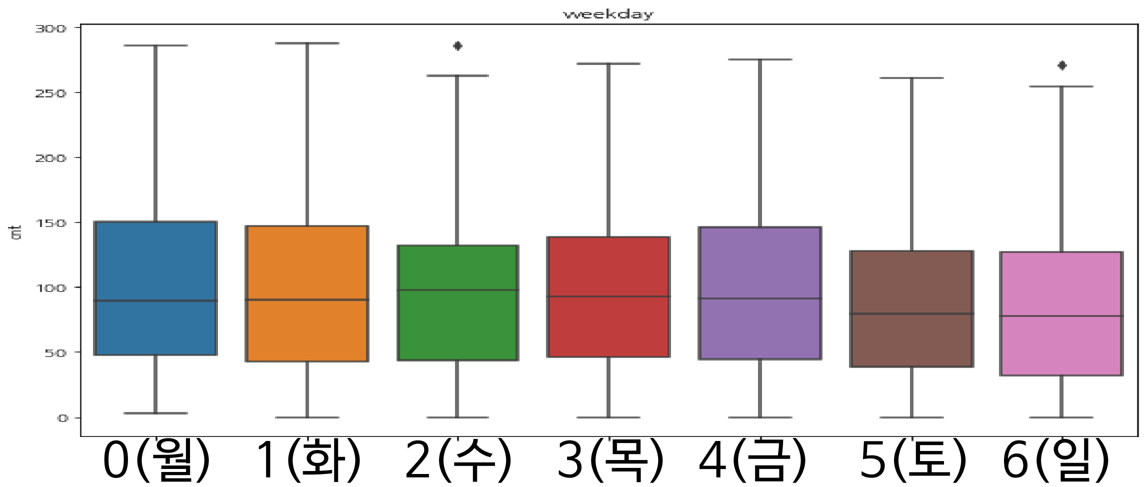
- 연도별 대여량 분포



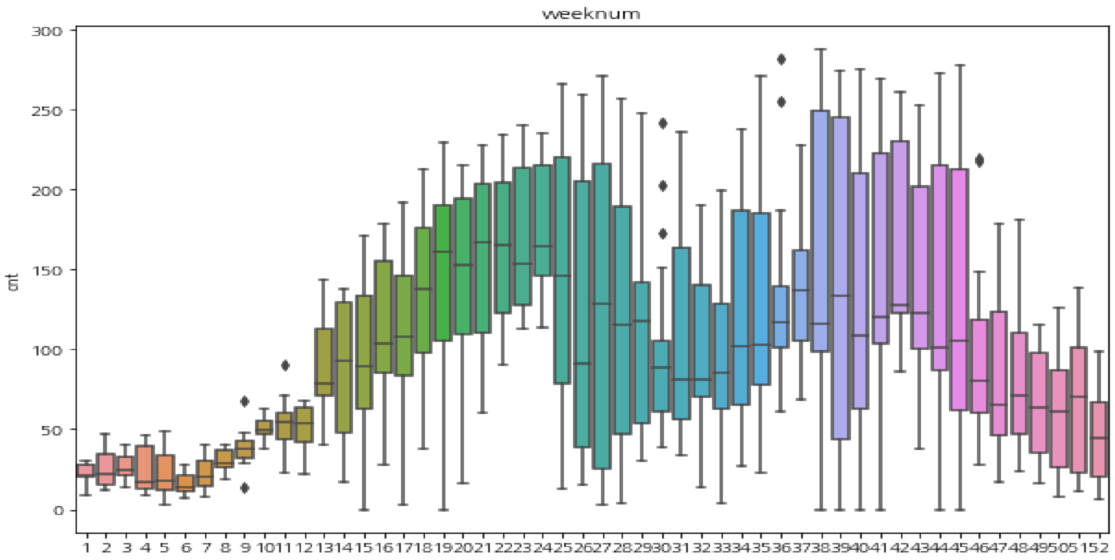
Exploratory Data Analysis

날짜 변수와 일일 대여량

- 요일 별 대여량 분포



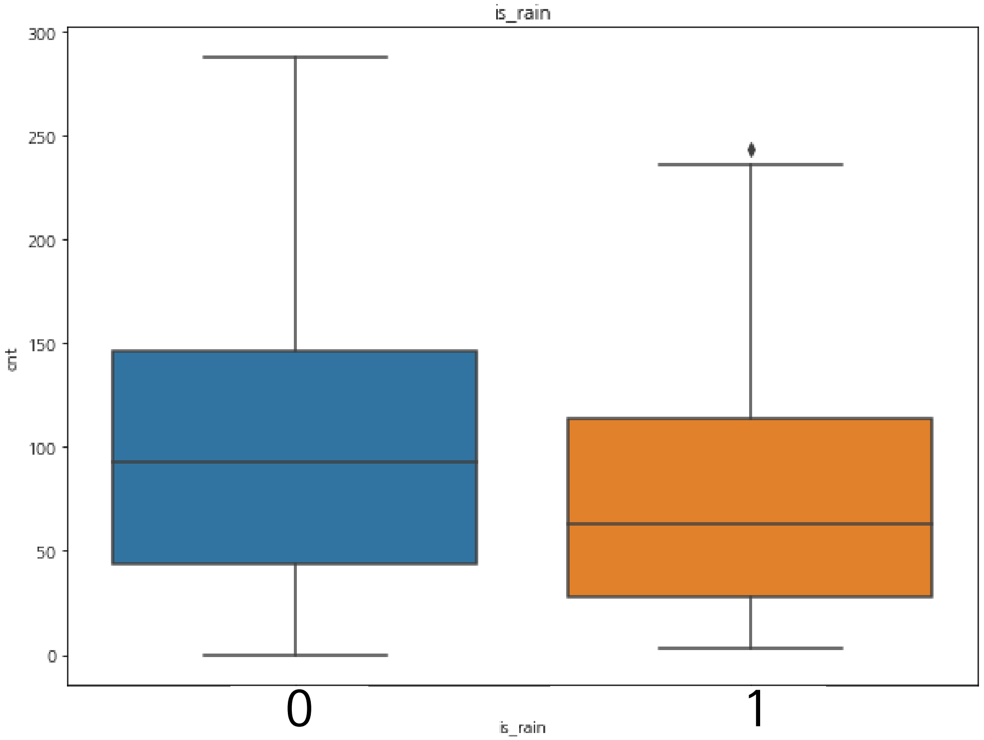
- 주차 별 대여량 분포



1. Data

Exploratory Data Analysis

강수 여부와 일일 대여량



2. Forecasting

01 **Data**

Exploratory Data Analysis and
Data Preprocessing

02 **Forecasting**

Univariate time series,
ARIMA and Multivariate Data Analysis

03 **Hidden Markov Model**

Analyze and Forecast
Using Hidden Markov Model

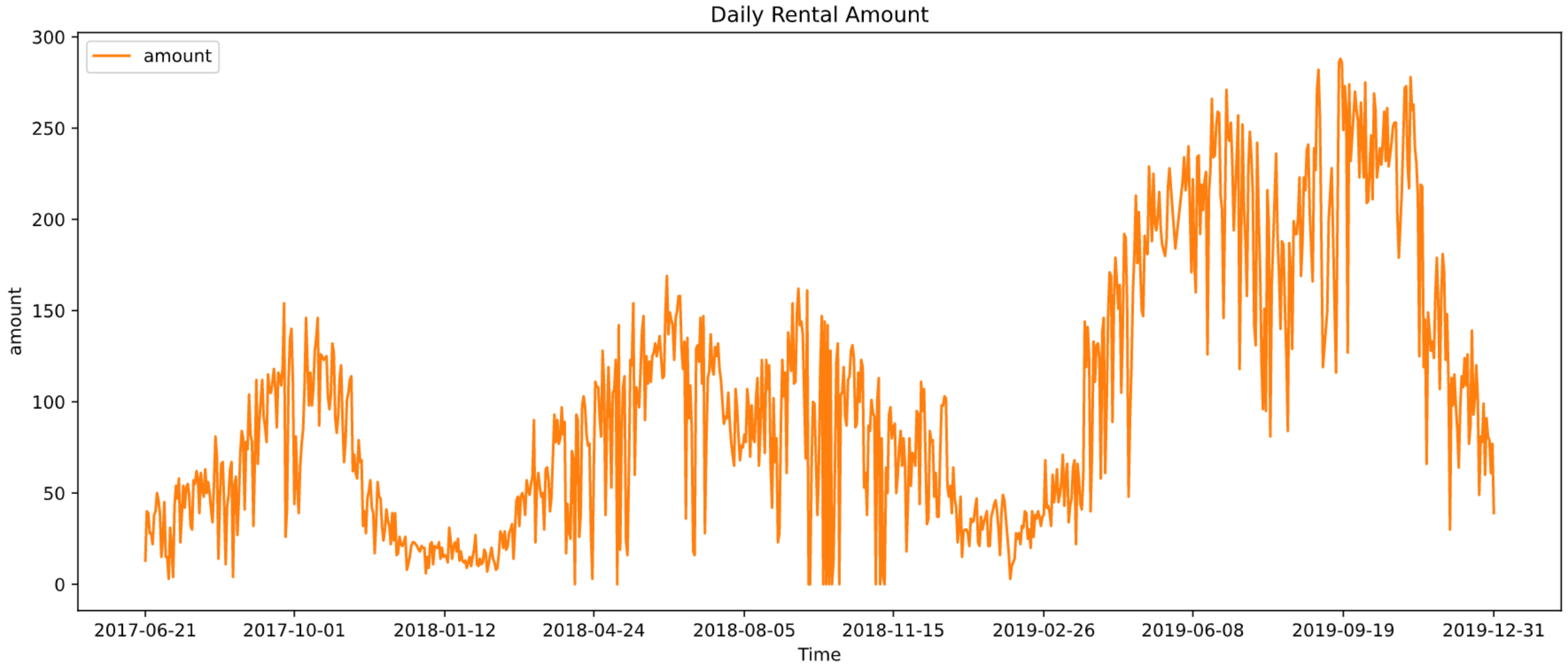
04 **Result**

Conclusion and Insights

2. Forecasting

A. Time Series

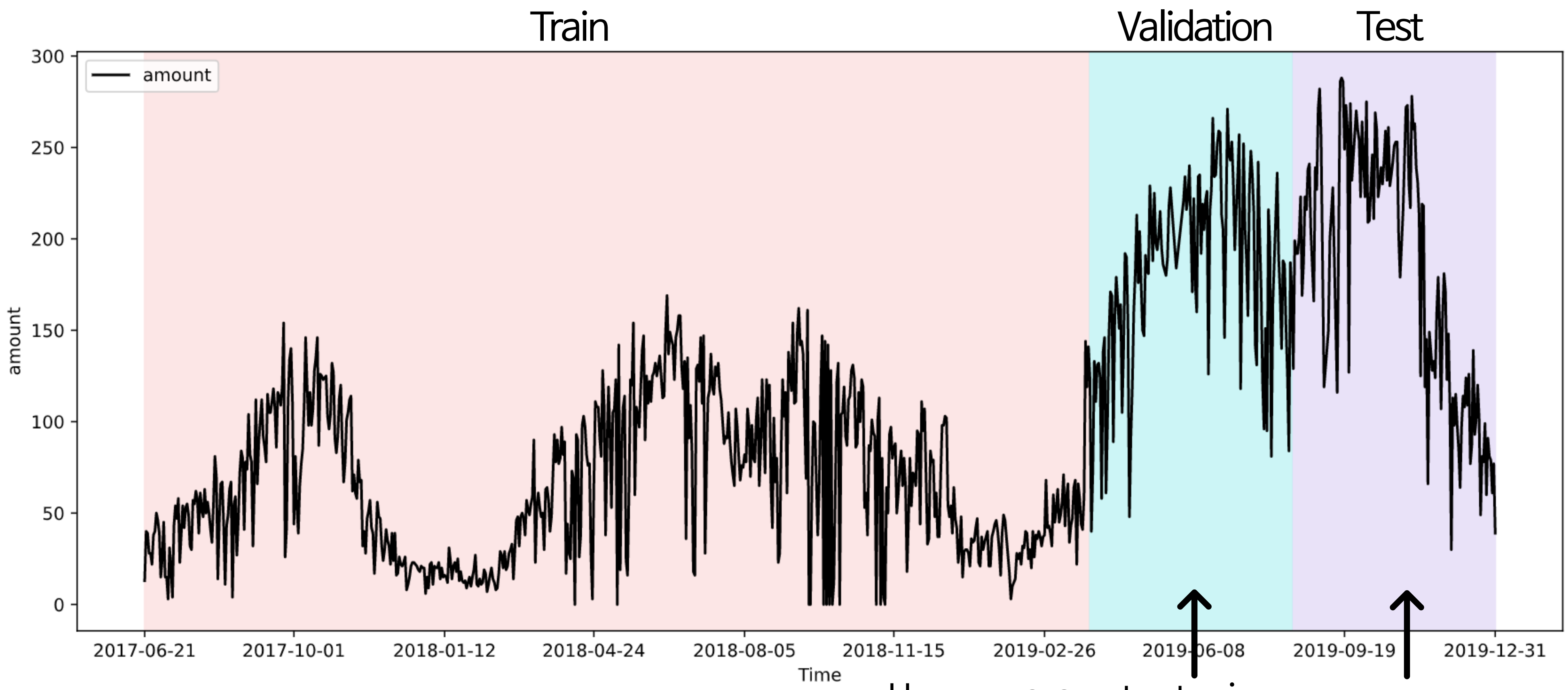
안암로터리 정류장 일일 따릉이 대여량
2017-06-21~2019-12-31



2. Forecasting

A. Time Series

안암로터리 정류장 일일 따릉이 대여량
2017-06-21~2019-12-31



Hyper-parameter tuning
필요한 model에 사용

최종 Best Model 선택

2. Forecasting: Time Series

1) Linear Regression with binary variable

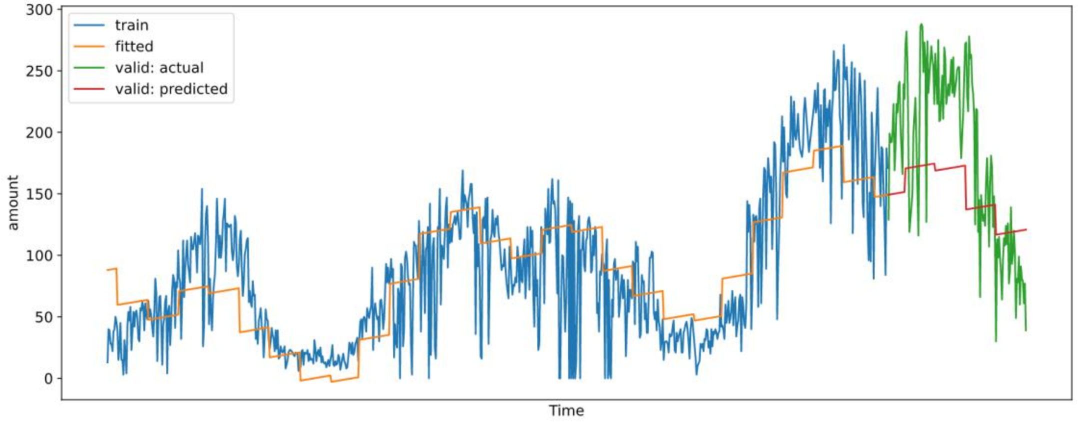
계절성을 반영하는 binary variable 생성 후 Linear Regression 수행

- Input Data

time_index	amount	1	2	3	4	5	6	7	8	9	10	11	12
0	13.0	0	0	0	0	0	1	0	0	0	0	0	0
1	40.0	0	0	0	0	0	1	0	0	0	0	0	0
2	39.0	0	0	0	0	0	1	0	0	0	0	0	0
3	28.0	0	0	0	0	0	1	0	0	0	0	0	0
4	28.0	0	0	0	0	0	1	0	0	0	0	0	0

월 정보를 나타내는 binary variable

- 예측 결과



Forecasting Performance	
MAE	53.31
MAPE	32.34
RMSE	62.91

2) Trigonometric Model

계절성을 반영하는 sine, cosine variable 생성 후 Linear Regression 수행

- Input Data

Model 2

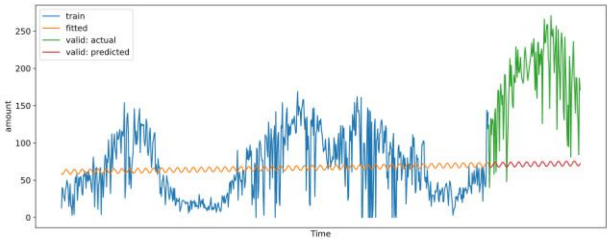
Model 1

time_index	amount	sin_1	cos_1	sin_2	cos_2
0	13.0	0.000000	1.000000e+00	0.000000e+00	1.0
1	40.0	0.500000	8.660254e-01	8.660254e-01	0.5
2	39.0	0.866025	5.000000e-01	8.660254e-01	-0.5
3	28.0	1.000000	6.123234e-17	1.224647e-16	-1.0
4	28.0	0.866025	-5.000000e-01	-8.660254e-01	-0.5

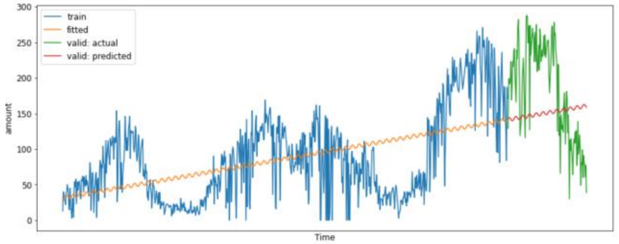
Time을 sine, cosine 값으로 변환한 variable

- 예측 결과

Model 1



Model 2



Forecasting Performance

Model1		Model2	
MAE	111.52	MAE	69.73
MAPE	58.04	MAPE	46.41
RMSE	120.80	RMSE	78.60

2. Forecasting: Time Series

3) Moving Average

일부 과거 데이터의 단순 평균값으로 미래 시점 예측

- 예측 방법

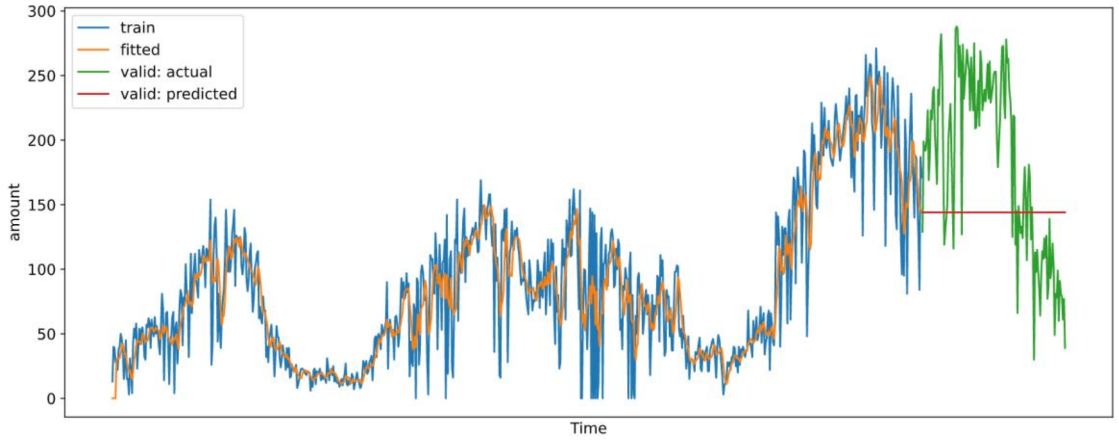
$$F_{t+1} = \frac{1}{n}\sum_{j=t+1-n}^{t} D_i$$

✓ 과거 n개의 data에 대해 동일한 가중치로 미래 예측

✓ 미래의 예측값 모두 동일

✓ Validation set에서,
MSE 기준으로 $n = 5$일 때 성능이 가장 좋음

- 예측 결과 $(n = 5)$



Forecasting Performance	
MAE	68.29
MAPE	42.80
RMSE	78.35

4-1) Simple Exponential Smoothing

모든 과거 데이터를 사용하여 이들의 weighted sum으로 미래 예측

- 예측 방법

$$L_0 = \frac{1}{n}\sum_{i=1}^{n} D_i$$

$$L_{t+1} = \alpha D_{t+1} + (1-\alpha)L_t$$

〈예측〉

$$F_{t+1} = L_t$$

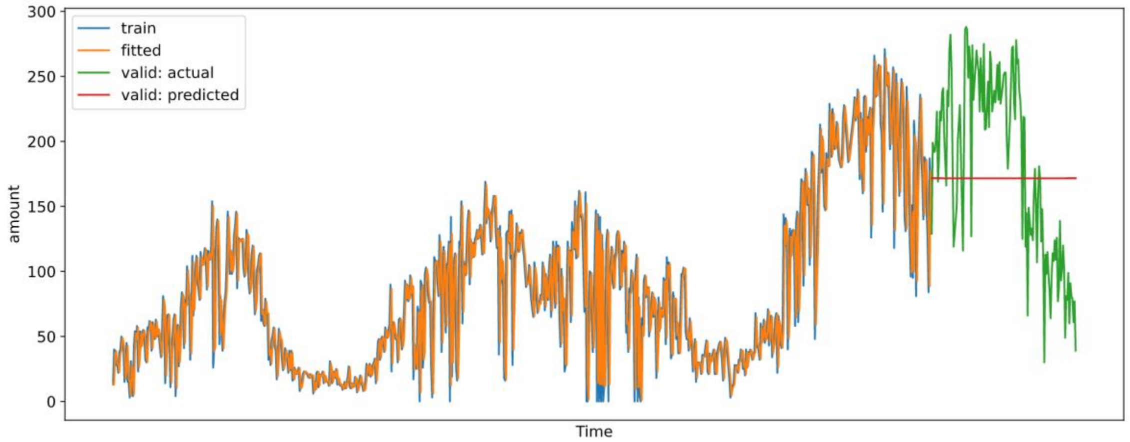
$$F_{t+n} = L_t$$

✓ 과거 모든 data에 대한 weighted sum L_t로 미래 예측

✓ 미래의 예측값 모두 동일

✓ Validation set에서,
MSE 기준으로 $\alpha = 0.9$일 때 성능이 가장 좋음

- 예측 결과 $(\alpha = 0.9)$



Forecasting Performance	
MAE	61.76
MAPE	46.39
RMSE	68.92

4-2) Double Exponential Smoothing

모든 과거 데이터를 사용하여 trend를 반영한 미래 예측

- 예측 방법

$$L_{t+1} = \alpha D_{t+1} + (1-\alpha)(L_t + B_t)$$

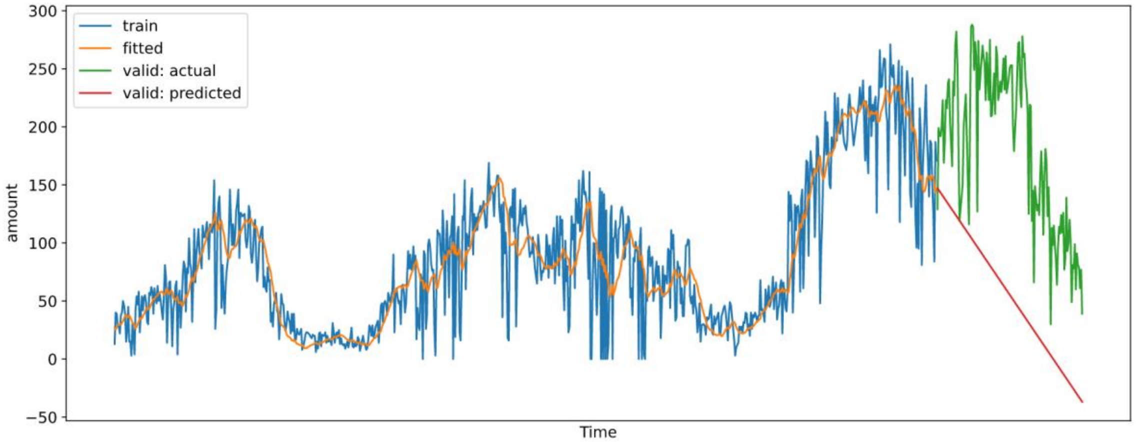
$$B_{t+1} = \beta(L_{t+1} - L_t) + (1-\beta)B_t$$

〈예측〉

$$F_{t+i} = L_t + iB_t \quad i \in \mathbb{Z}$$

- ✓ Trend 예측 가능
- ✓ Validation set에서,
  MSE 기준으로 $\alpha = 0.1,\ \beta = 0.1$일 때 성능이 가장 좋음

- 예측 결과 $(\alpha = 0.1,\ \beta = 0.1)$



Forecasting Performance	
MAE	128.00
MAPE	76.11
RMSE	138.44

2. Forecasting: Time Series

5-1) Additive Holt-Winters Exponential Smoothing

Seasonal variation의 산포가 일정한 time series 예측

- 예측 방법

$$l_t = \alpha(y_T - sn_{T-L}) + (1-\alpha)\left(l_{(T-1)} + b_{(T-1)}\right)$$
$$b_T = \beta\left(l_T - l_{(T-1)}\right) + (1-\beta)b_{(T-1)}$$
$$sn_T = \delta(y_T - l_T) - (1-\delta)sn$$

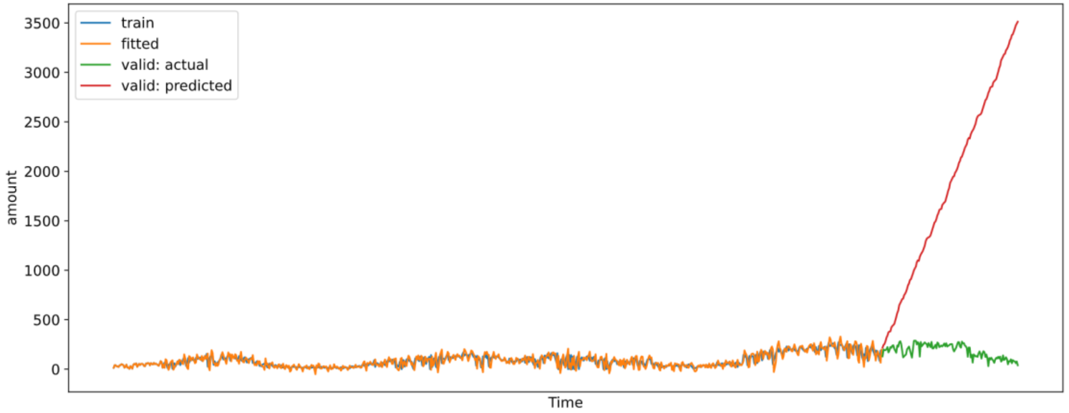
〈예측〉

$$\hat{y}_{T+\tau} = l_T + \tau b_T + sn_{T+\tau-L}$$

✓ Validation set에서,

MSE 기준으로 $\alpha = 0.6,\ \beta = 0.9,\ \delta = 0.1$일 때 성능이 가장 좋음

- 예측 결과 $(\alpha = 0.6,\ \beta = 0.9,\ \delta = 0.1)$



Forecasting Performance	
MAE	1677.11
MAPE	1402.61
RMSE	1952.97

5-2) Multiplicative Holt-Winters Exponential Smoothing

Seasonal variation의 산포가 증가하는 time series 예측

- 예측 방법

$$l_t = \alpha(y_T/sn_{T-L}) + (1-\alpha)\big(l_{(T-1)} + b_{(T-1)}\big)$$
$$b_T = \beta\big(l_T - l_{(T-1)}\big) + (1-\beta)b_{(T-1)}$$
$$sn_T = \delta(y_T/sn_{T-L}) - (1-\delta)sn_T$$

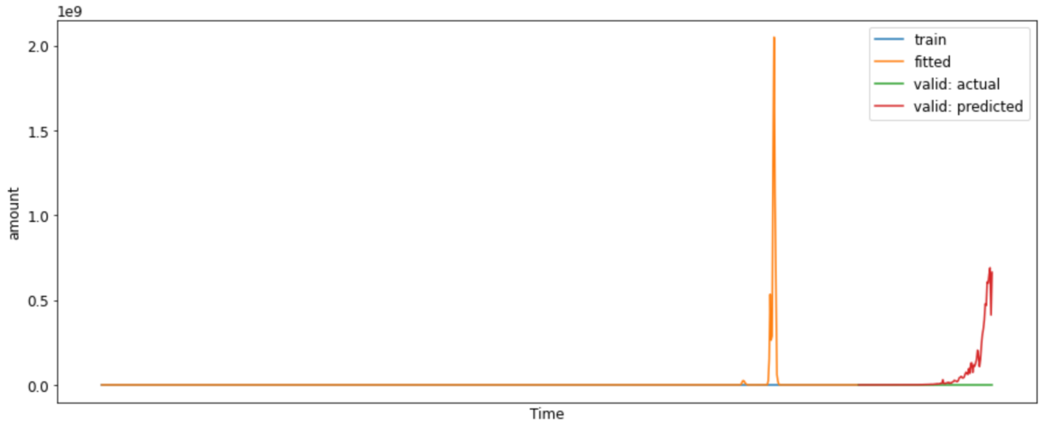
<예측>

$$\hat{y}_{T+\tau} = (l_T + \tau b_T)sn_{T+\tau-L}$$

✓ Validation set에서,

MSE기준으로 $\alpha = 0.6,\ \beta = 0.5, \delta = 0.2$일 때 성능이 가장 좋음

- 예측 결과 $(\alpha = 0.6,\ \beta = 0.5, \delta = 0.2)$



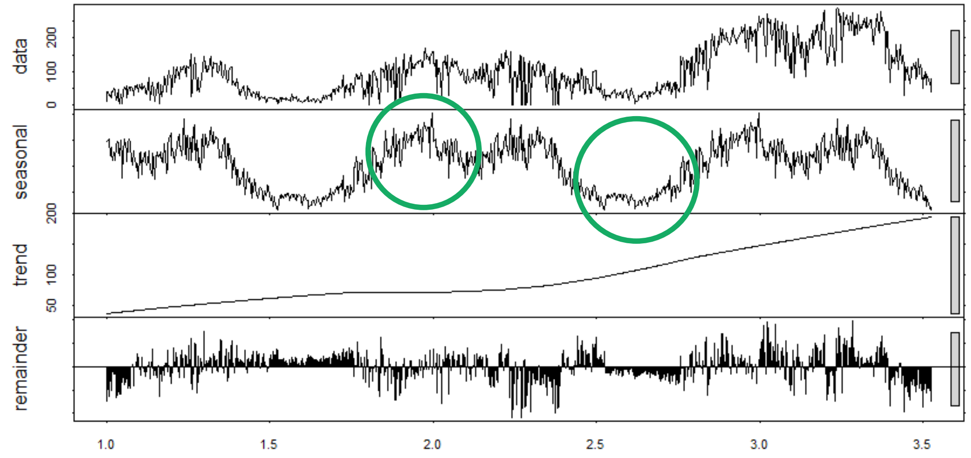
Forecasting Performance	
MAE	60124645.94
MAPE	79370725.26
RMSE	155145463.91

2. Forecasting: Time Series

6) ARIMA

일별 데이터 분석

- Time Series Decomposition



✓연단위(365일) 계절성 확인할 수 있고, 증가 추세를 보이는

데이터로서 특히 2019년에 대여량이 큰 폭 증가

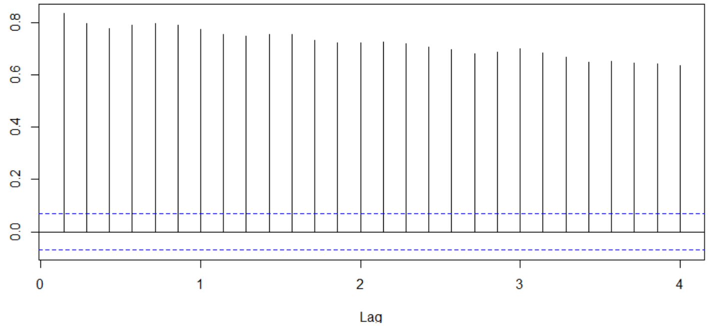
- Durbin-Watson Test

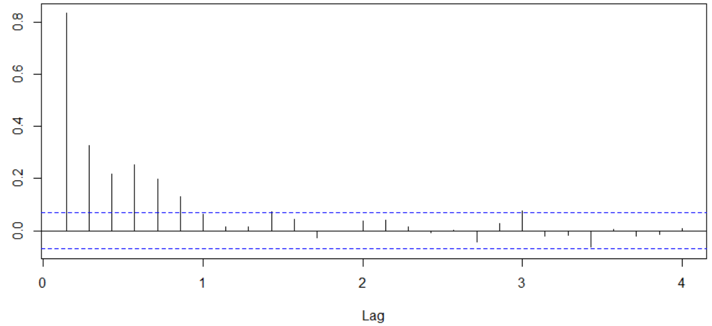
```
        Durbin-Watson test

data:  df ~ seq(1, tot_len, by = 1)
DW = 0.40017, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

✓더빗-왓슨 검정 결과 자기상관성이 있는, 차분 등 전처리가 필요한 데이터

- Augmented Dickey-Fuller Test

```
        Augmented Dickey-Fuller Test

data:  boxcox_df
Dickey-Fuller = -2.2644, Lag order = 9, p-value = 0.4664
alternative hypothesis: stationary
```





✓ACF, PACF 그래프 및 Dickey-Fuller 검정을 통해서도 차분을 통해

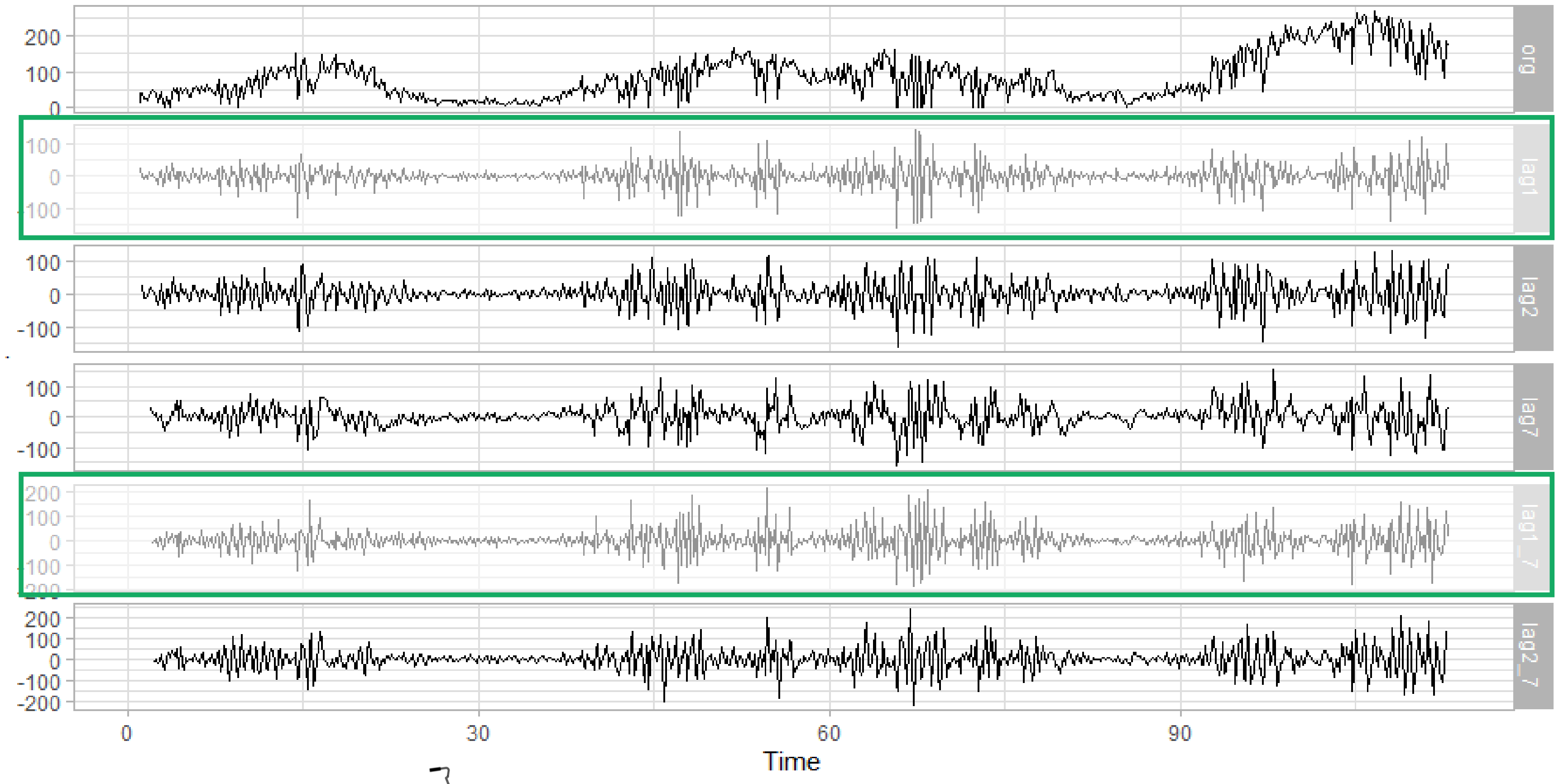
정상 시계열로 만들어야 하는 데이터임을 알 수 있음

2. Forecasting: Time Series

6) ARIMA

일별 데이터 분석

- Differencing



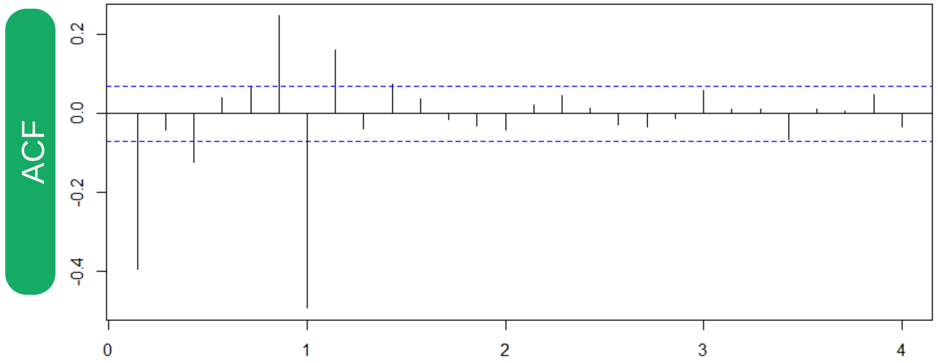
2. Forecasting: Time Series

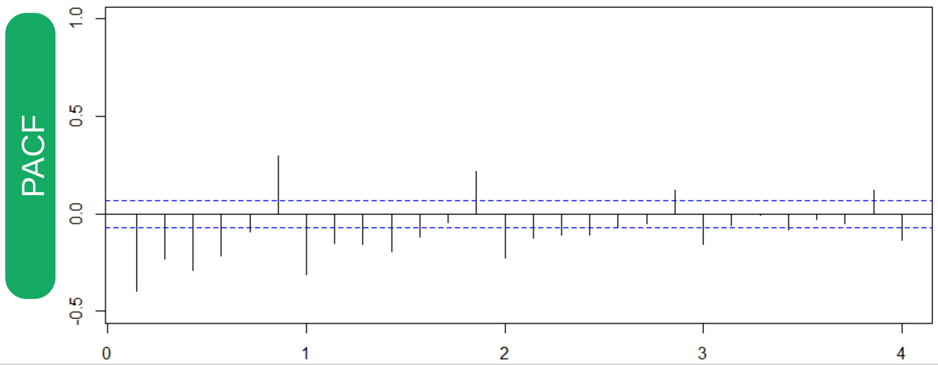
6) ARIMA

일별 데이터 분석

- 1차 차분 결과

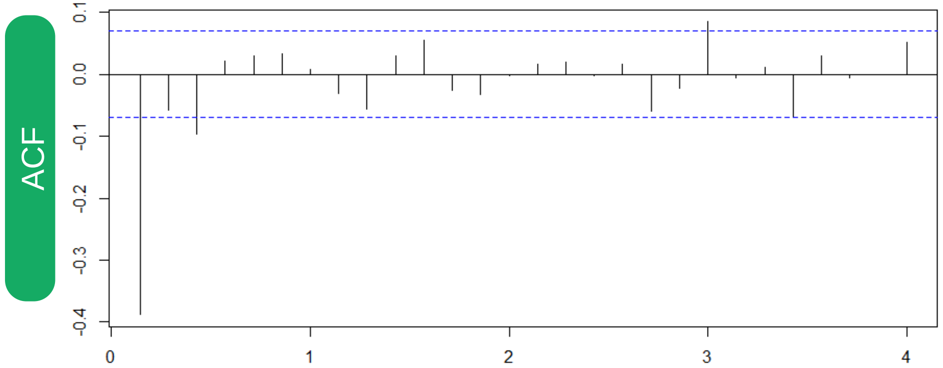
```
        Augmented Dickey-Fuller Test

data:  lagged_df
Dickey-Fuller = -13.047, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

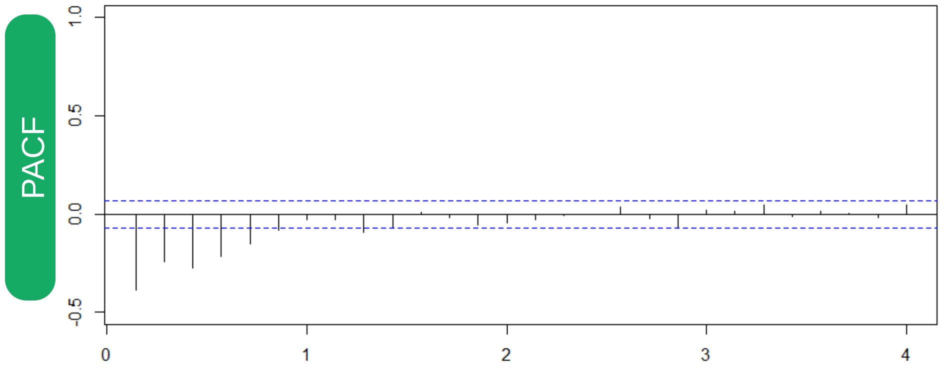




- 1차 + 7차(계절) 차분 결과

```
        Augmented Dickey-Fuller Test

data:  lagged_df
Dickey-Fuller = -16.717, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```





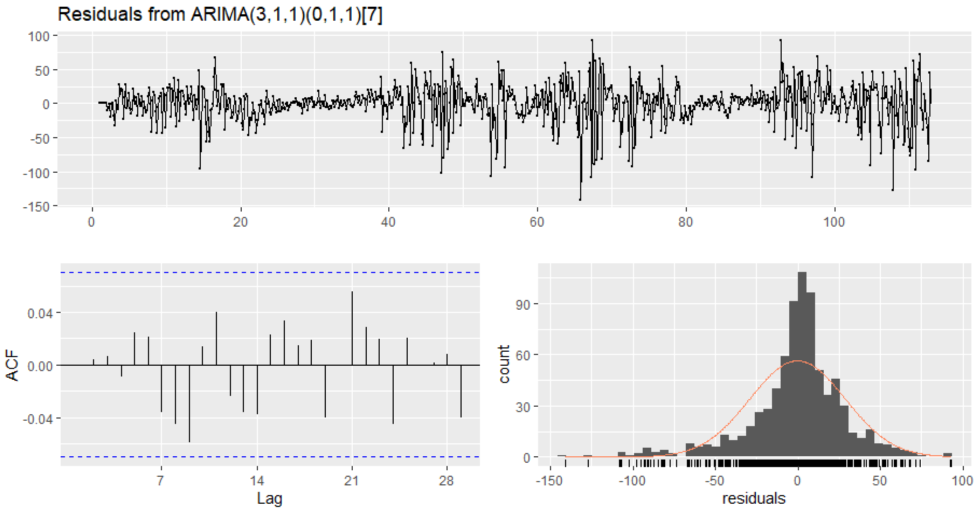
2. Forecasting: Time Series

6) ARIMA

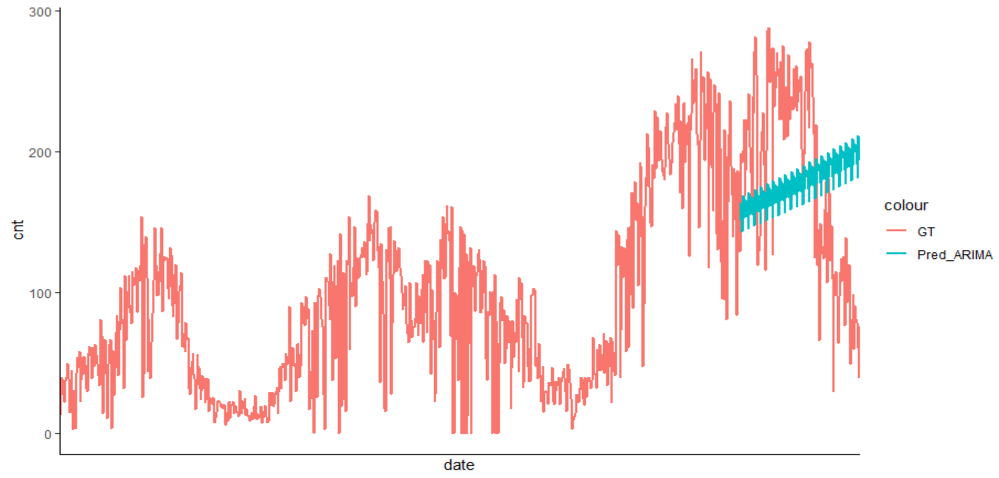
SARIMA (3, 1, 1) (0, 1, 1) [7]

- Result (Ljung-Box test)

```
ARIMA(3,1,1)(0,1,1)[7]

Coefficients:
          ar1     ar2      ar3      ma1     sma1
      -0.0301  0.0547  -0.0285  -0.8210  -1.0000
s.e.   0.0600  0.0555   0.0475   0.0493   0.0355

sigma^2 estimated as 3.572:  log likelihood=-1609.96
AIC=3231.92   AICc=3232.03   BIC=3259.85
```



- Prediction



Forecasting Performance	
MAE	68.14
MAPE	53.30
RMSE	76.08

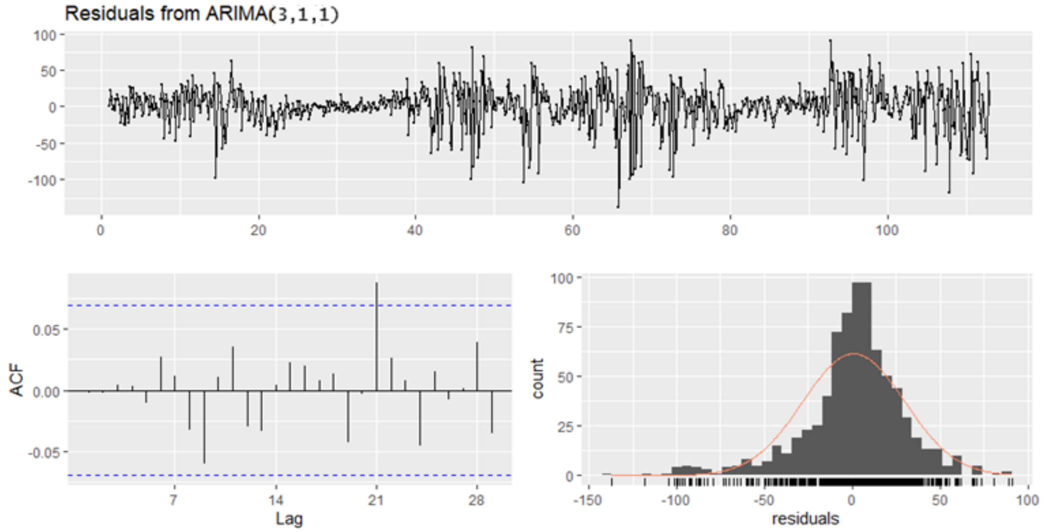
2. Forecasting: Time Series

6) ARIMA

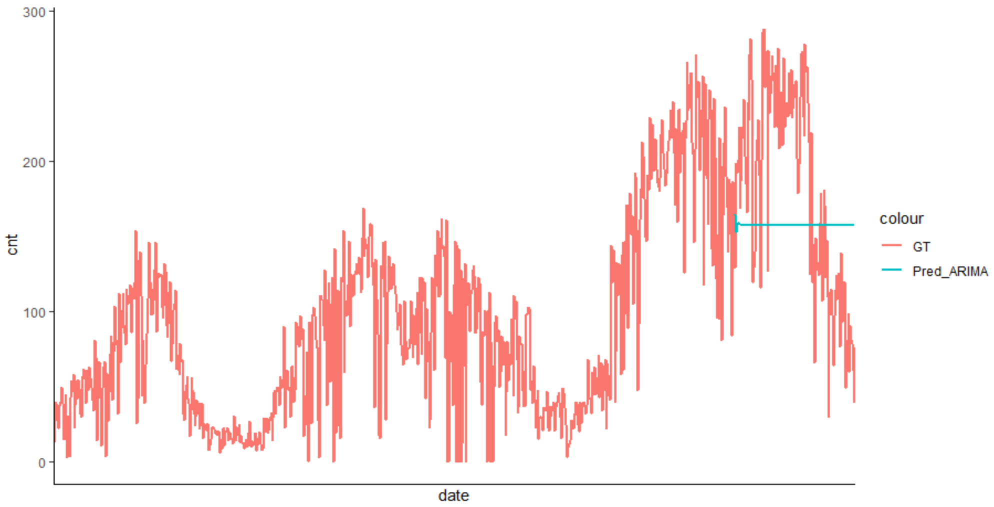
ARIMA(3, 1, 1)

- Result (Ljung-Box test)

```
ARIMA(3,1,1)

Coefficients:
          ar1     ar2      ar3      ma1
      -0.0276  0.0594  -0.0262  -0.8267
s.e.   0.0589  0.0548   0.0472   0.0476

sigma^2 estimated as 3.58:  log likelihood=-1608.87
AIC=3227.74   AICc=3227.82   BIC=3251.06
```



- Prediction



Forecasting Performance

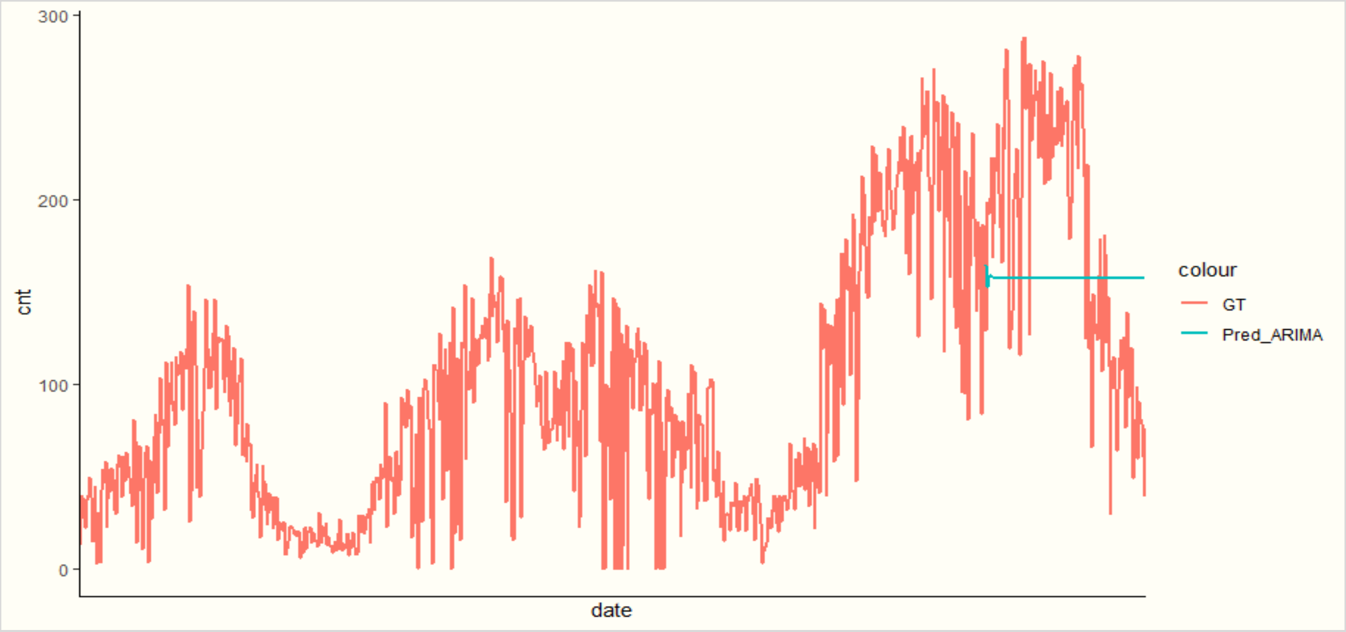
MAE	64.91
MAPE	44.10
RMSE	73.03

최종 모델

2. Forecasting: Time Series

6) ARIMA

일별 데이터 분석 한계점



✓ 예측 결과가 좋지 않고, 특히 test data 기간 내 11~12월 급격한 감소 패턴(계절성)을 설명하지 못함

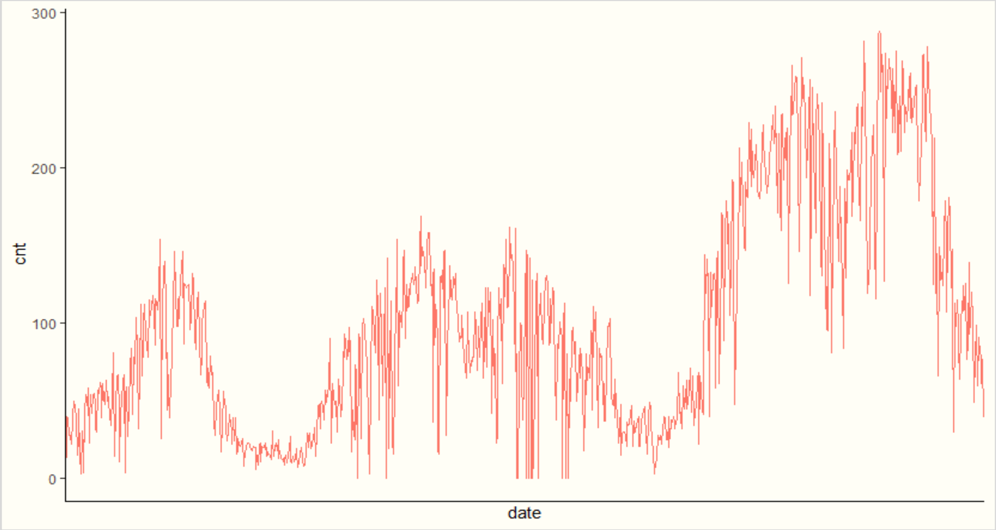
✓ 분석에 사용한 R에서는 lag 최대 길이가 **350**으로 365 단위 계절 차분 불가능할 뿐더러,

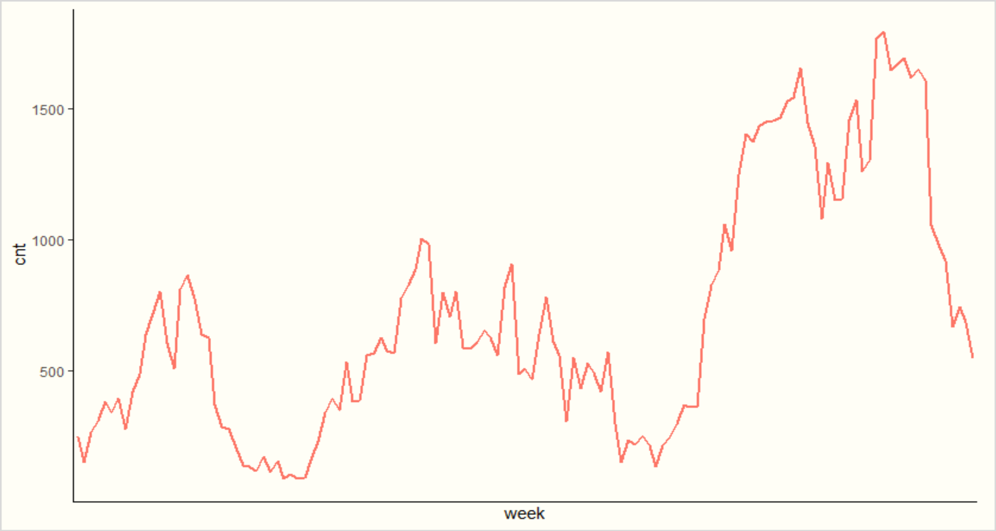
✓365 단위 계절 차분에 대한 의구심

2. Forecasting: Time Series

6) ARIMA

해결방안





✓ 일별 데이터 → 주별 데이터 변환

✓ 365(일) 단위 → 52(주) 단위 계절 차분 사용

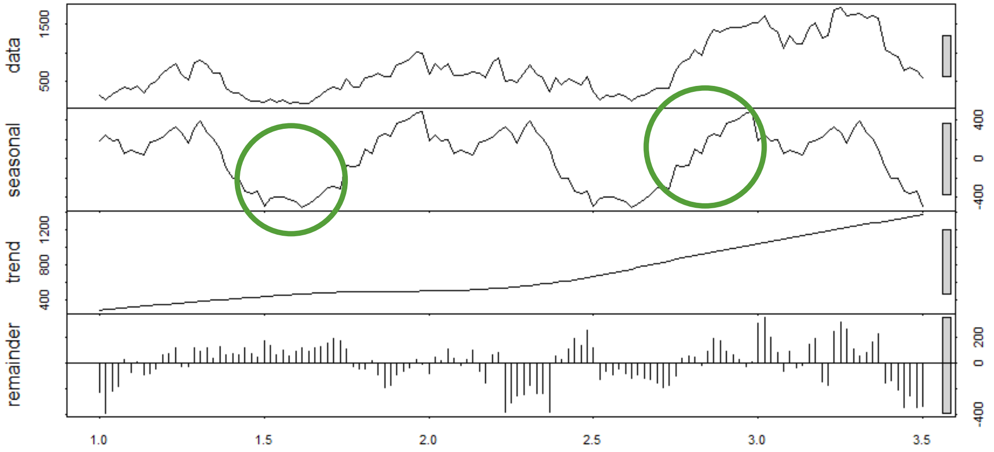
✓ **주별 대여 건수 예측 후 일별 분배** 방식으로 예측 수행

2. Forecasting: Time Series

6) ARIMA

주별 데이터 분석

- Time Series Decomposition



✓연단위(52주) 계절성 확인할 수 있고, 증가 추세를 보이는 데이터로서 특히 2019년에 대여량이 큰 폭 증가

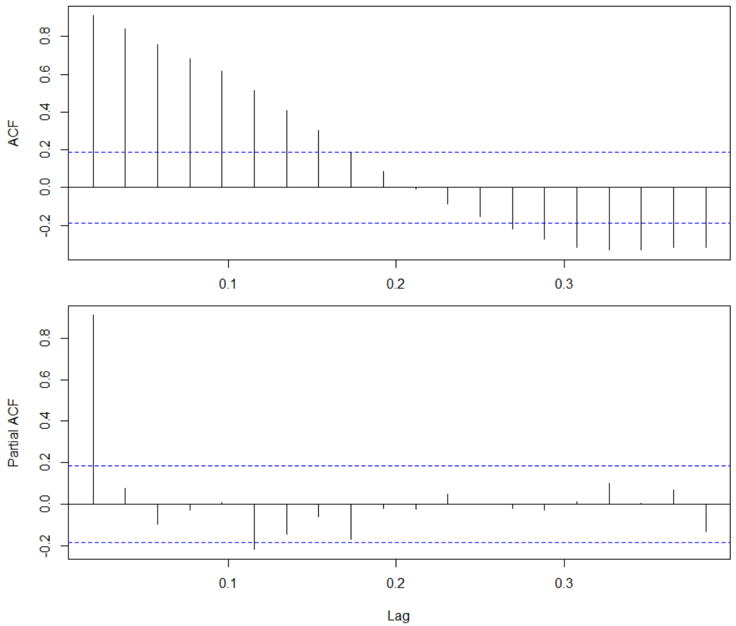
- Durbin-Watson Test

```
        Durbin-Watson test

data:  df_org$cnt ~ seq(1, tot_len, by = 1)
DW = 0.17238, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

✓더빗-왓슨 검정 결과 자기상관성이 있는, 차분 등 전처리가 필요한 데이터

- Augmented Dickey-Fuller Test

```
        Augmented Dickey-Fuller Test

data:  boxcox_df
Dickey-Fuller = -2.1625, Lag order = 4, p-value = 0.5092
alternative hypothesis: stationary
```

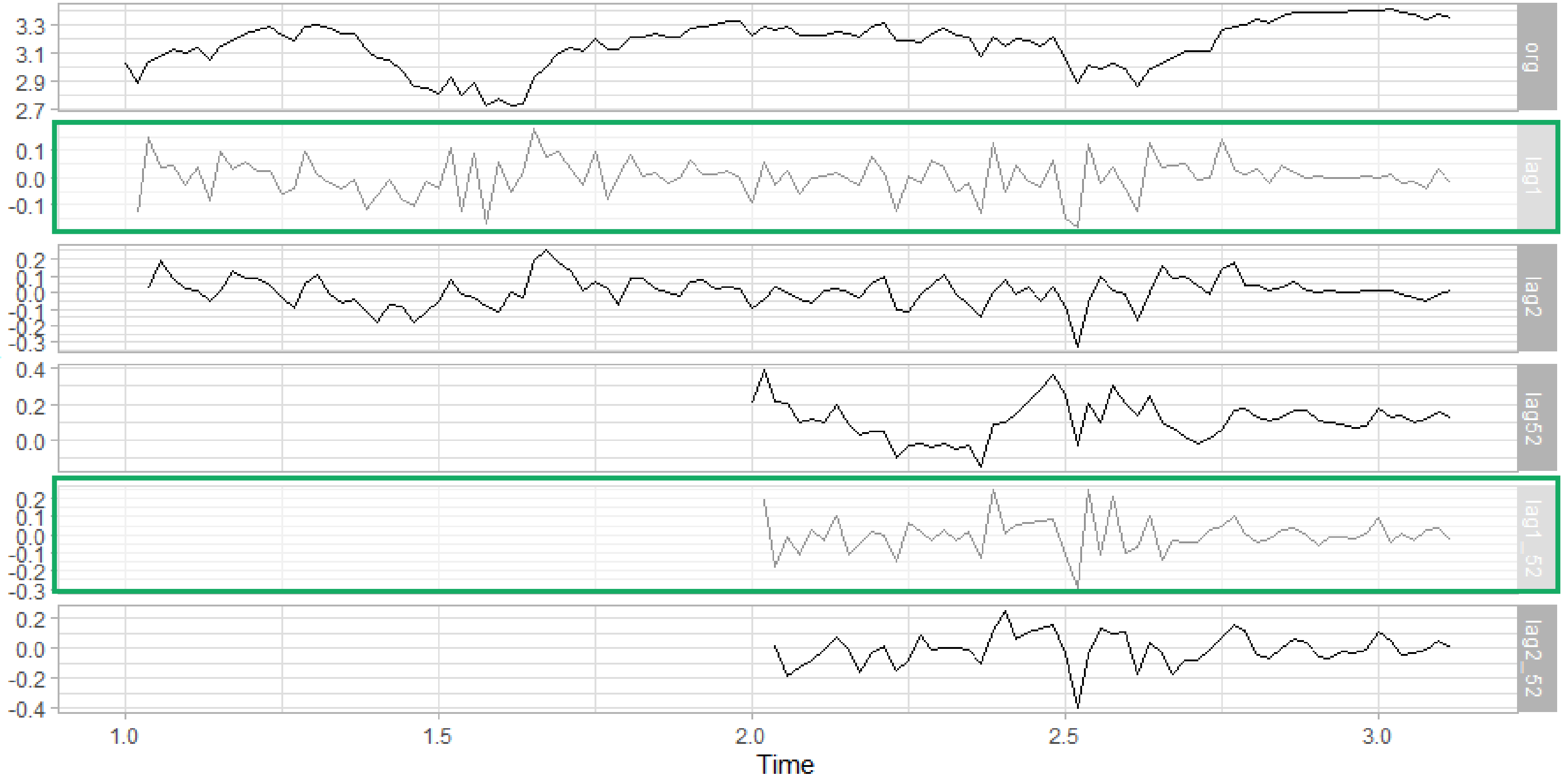


✓ACF, PACF 그래프 및 Dickey-Fuller 검정을 통해서도 차분을 통해 정상 시계열로 만들어야 하는 데이터임을 알 수 있음

2. Forecasting: Time Series

6) ARIMA

주별 데이터 분석



2. Forecasting: Time Series

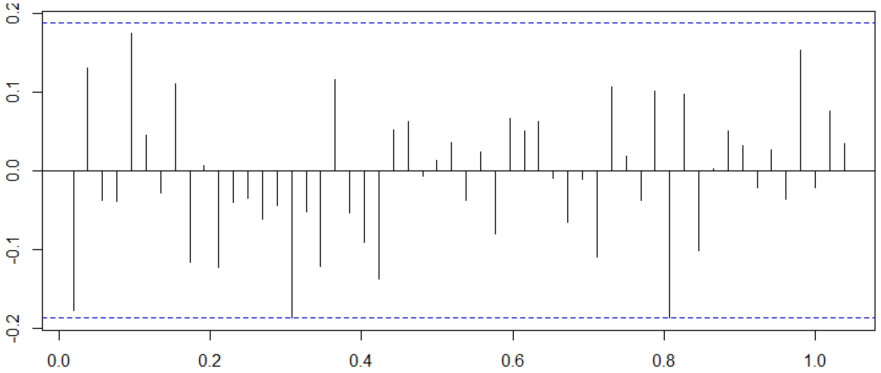
6) ARIMA

주별 데이터 분석

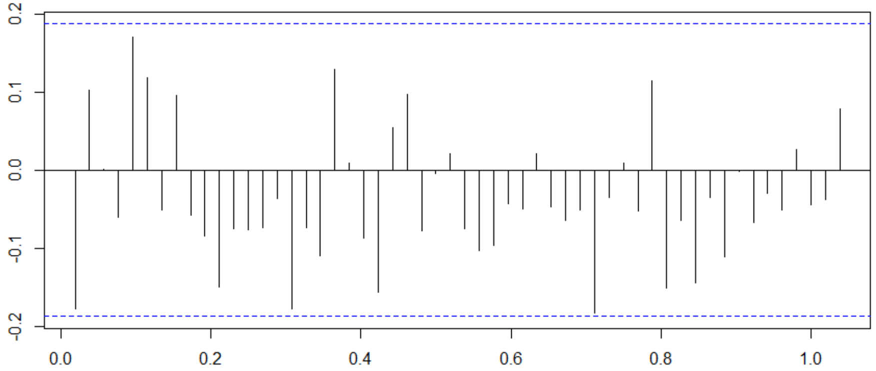
- 1차 차분 결과

```
Augmented Dickey-Fuller Test

data:  lagged_df
Dickey-Fuller = -3.8446, Lag order = 4, p-value = 0.01933
alternative hypothesis: stationary
```

ACF



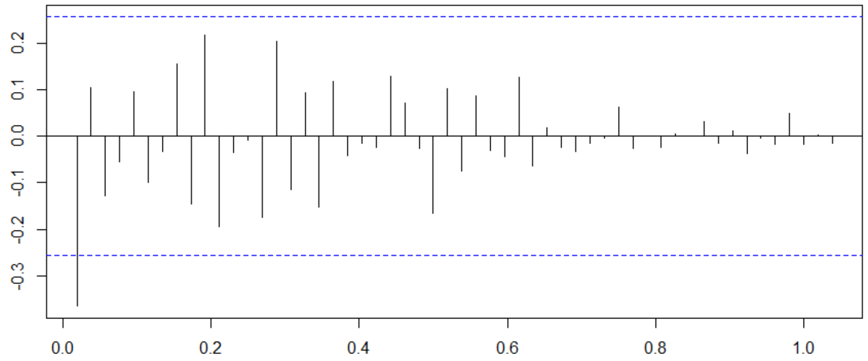
PACF



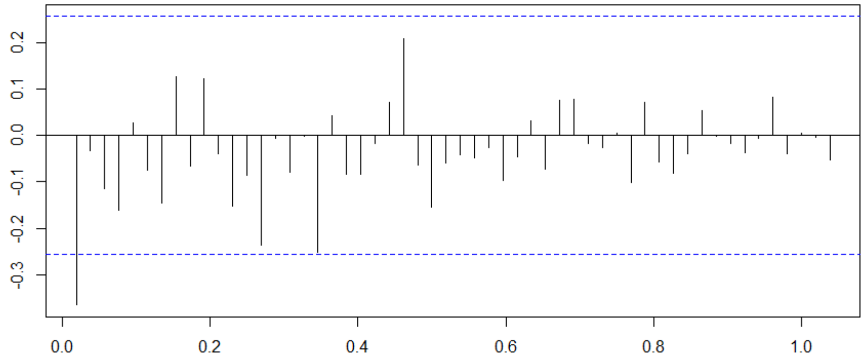
- 1차 + 52차(계절) 차분 결과

```
Augmented Dickey-Fuller Test

data:  lagged_df
Dickey-Fuller = -4.8504, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary
```

ACF



PACF



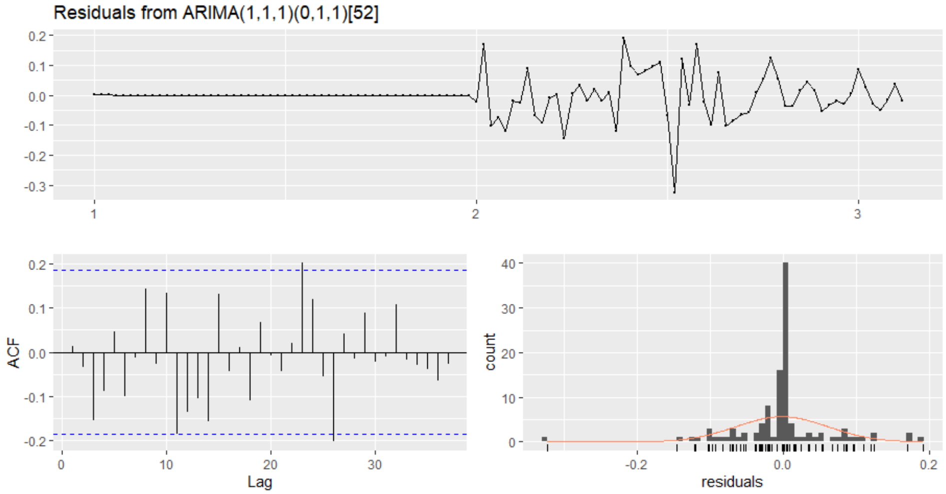
2. Forecasting: Time Series

6) ARIMA

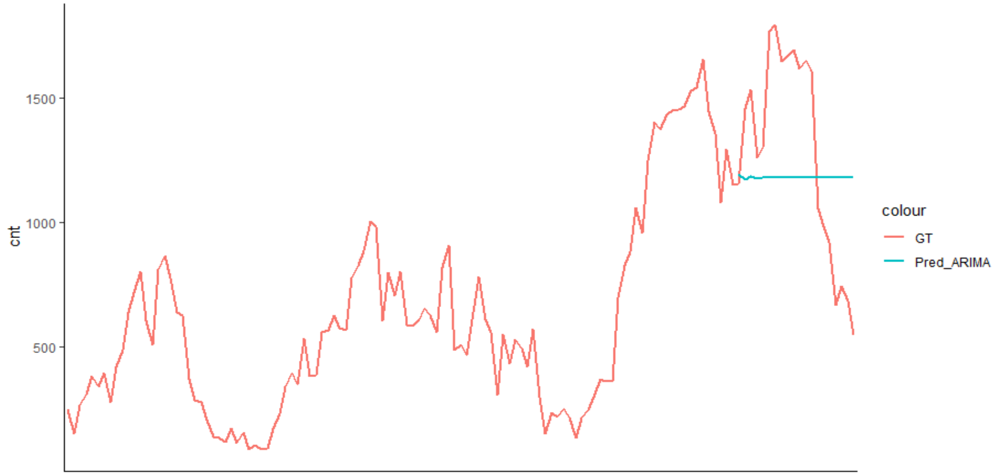
ARIMA(1, 1, 1)

- Result (Ljung-Box test)

```
ARIMA(1,1,1)

Coefficients:
          ar1     ma1
      -0.5094  0.3307
s.e.   0.2824  0.2999

sigma^2 estimated as 0.004468:  log likelihood=142.49
AIC=-278.99   AICc=-278.76   BIC=-270.89
```



- Prediction



Forecasting Performance	
MAE	376.83
MAPE	32.96
RMSE	416.95

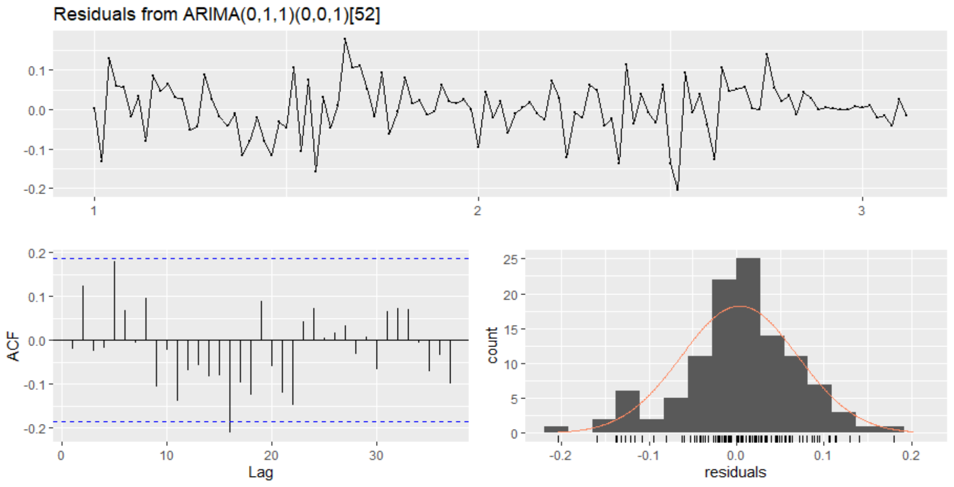
2. Forecasting: Time Series

6) ARIMA

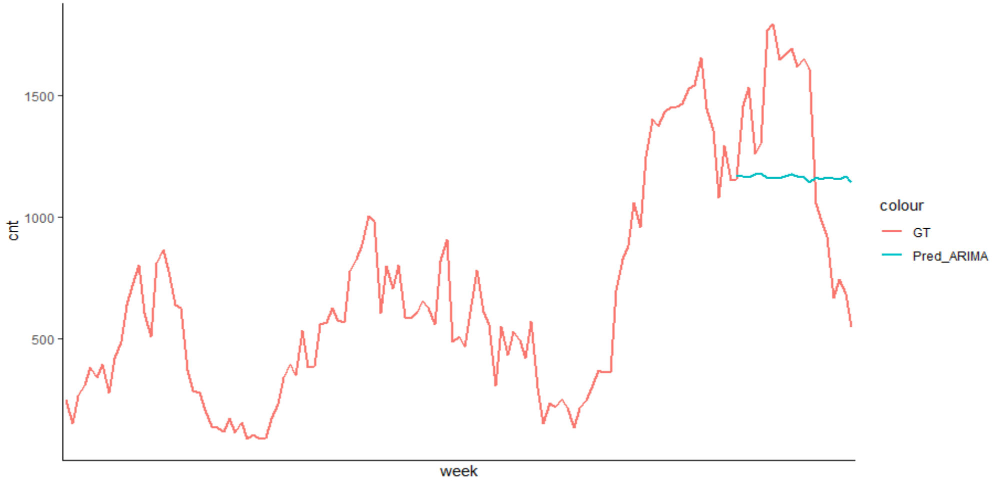
SARIMA (0, 1, 1) (0, 0, 1) [52]

- Result (Ljung-Box test)

```
ARIMA(1,1,1)

Coefficients:
          ar1     ma1
      -0.5094  0.3307
s.e.   0.2824  0.2999

sigma^2 estimated as 0.004468:  log likelihood=142.4971
AIC=-278.99   AICc=-278.76   BIC=-270.89
```



- Prediction



Forecasting Performance	
MAE	376.29
MAPE	33.44
RMSE	419.03

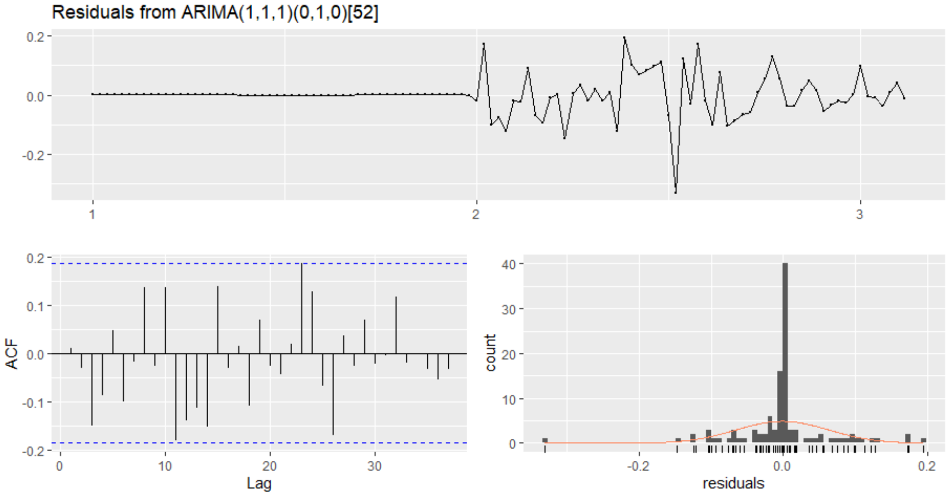
2. Forecasting: Time Series

6) ARIMA

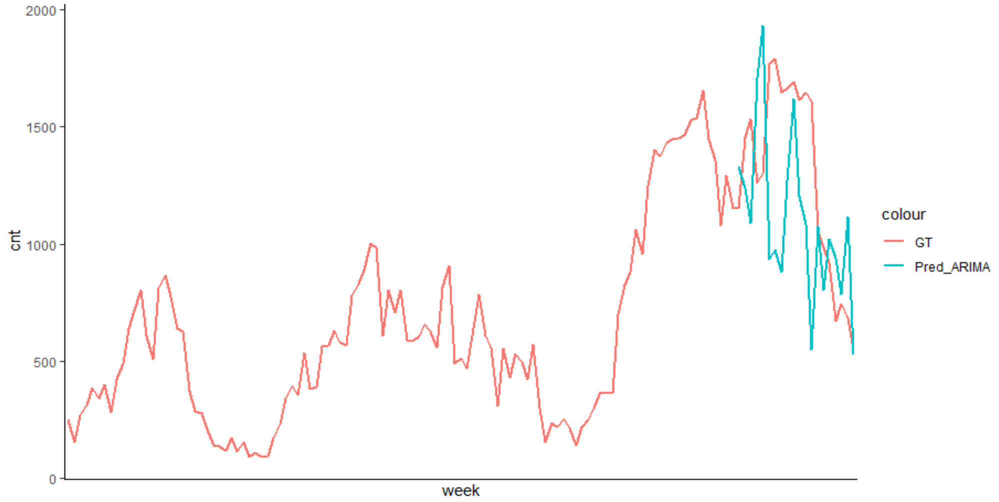
SARIMA(1, 1, 1) (0, 1, 0) [52]

- Result (Ljung-Box test)

```
ARIMA(1,1,1)(0,1,0)[52]

Coefficients:
          ar1      ma1
      -0.2636  -0.1363
s.e.   0.7382   0.7905

sigma^2 estimated as 0.008134:  log likelihood=58.21
AIC=-110.42   AICc=-109.97   BIC=-104.23
```



- Prediction



Forecasting Performance

MAE	395.03
MAPE	28.98
RMSE	494.97

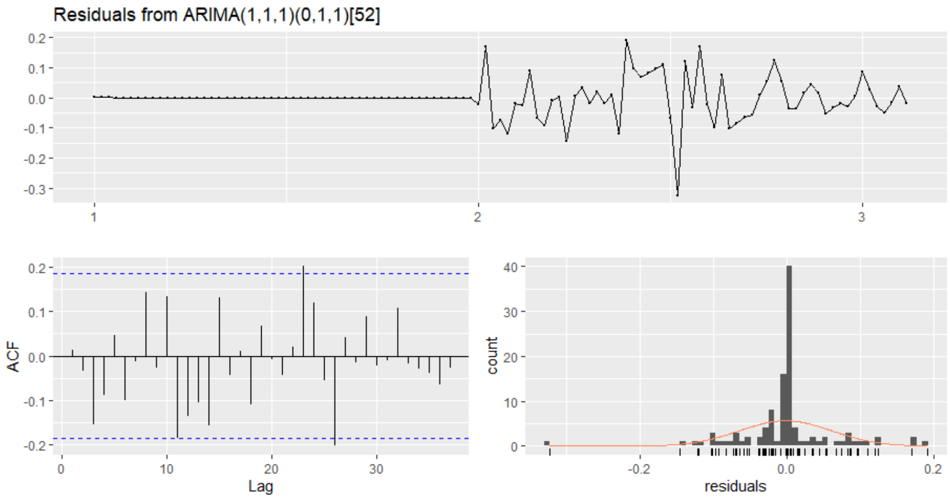
2. Forecasting: Time Series

6) ARIMA

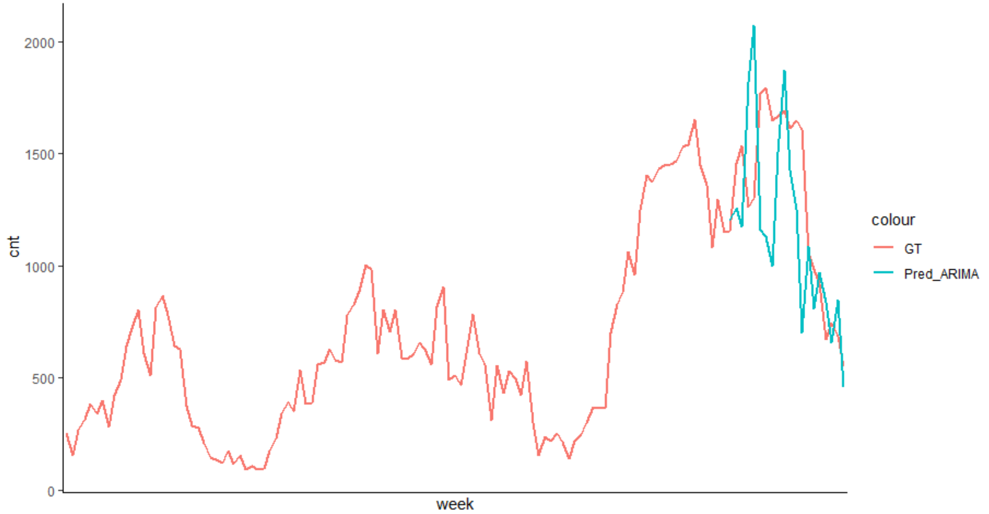
SARIMA (1, 1, 1) (0, 1, 1) [52]

- Result (Ljung-Box test)

```
ARIMA(1,1,1)(0,1,1)[52]

Coefficients:
          ar1      ma1     sma1
      -0.2568  -0.1398  -0.2020
s.e.   0.7589   0.8108   0.6224

sigma^2 estimated as 0.007974:  log likelihood=58.26
AIC=-108.52   AICc=-107.77   BIC=-100.28
```



- Prediction



Forecasting Performance

MAE	322.61
MAPE	23.55
RMSE	417.50

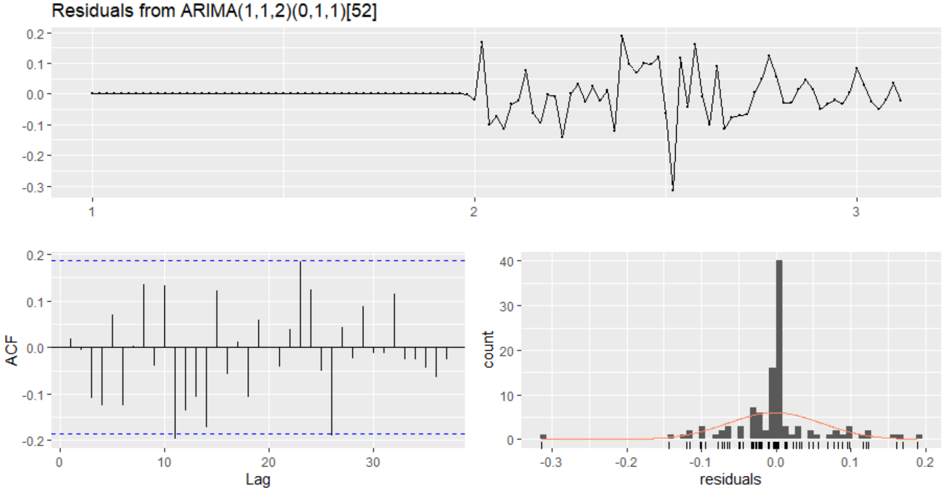
2. Forecasting: Time Series

6) ARIMA

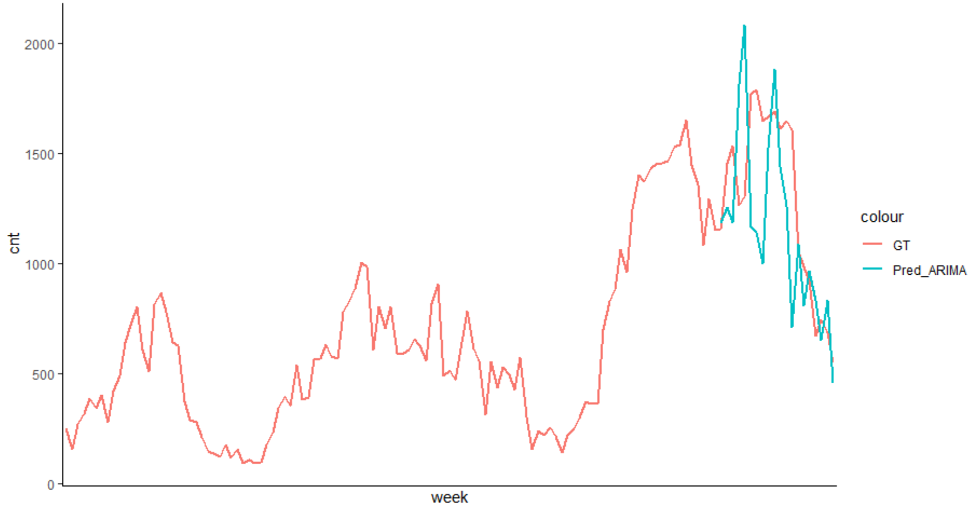
SARIMA(1, 1, 2) (0, 1, 1) [52]

- Result (Ljung-Box test)

```
ARIMA(1,1,2)(0,1,1)[52]

Coefficients:
          ar1     ma1      ma2     sma1
      -0.7983  0.3904  -0.2527  -0.2161
s.e.   0.6218  0.6448   0.3426   0.6365

sigma^2 estimated as 0.008033:  log likelihood=58.42
AIC=-106.84   AICc=-105.69   BIC=-96.54
```



- Prediction



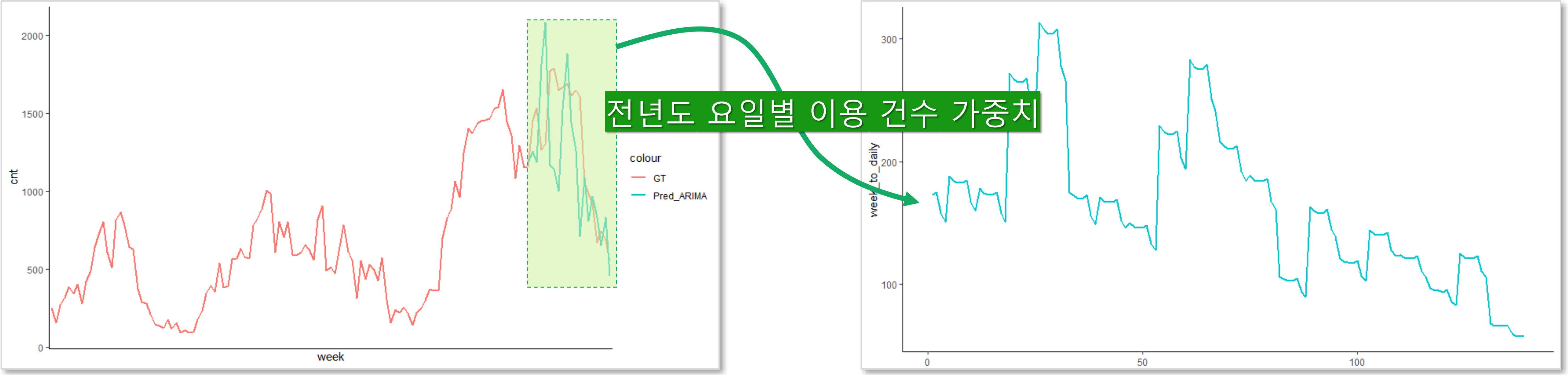
Forecasting Performance	
MAE	318.07
MAPE	23.20
RMSE	413.84

최종 모델

2. Forecasting: Time Series

6) ARIMA

주별 데이터 → 일별 데이터 변환

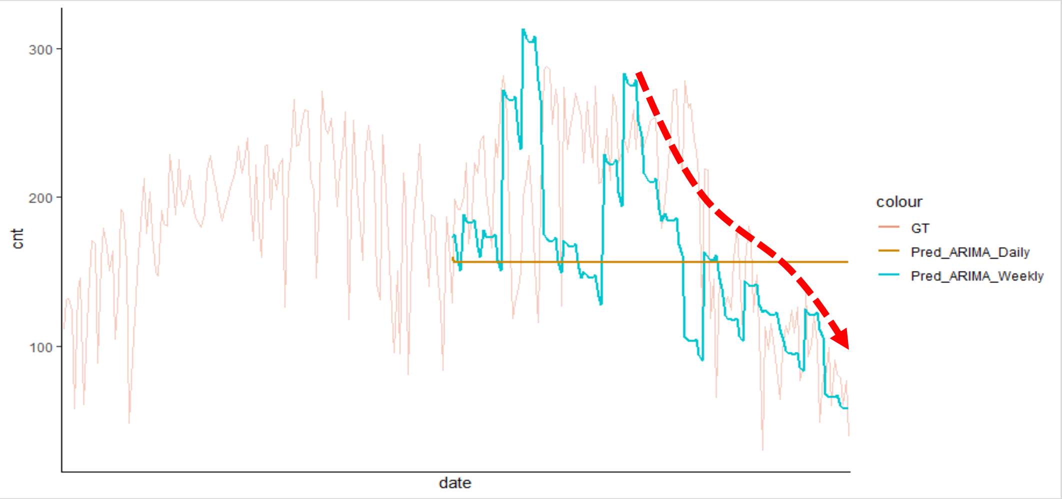


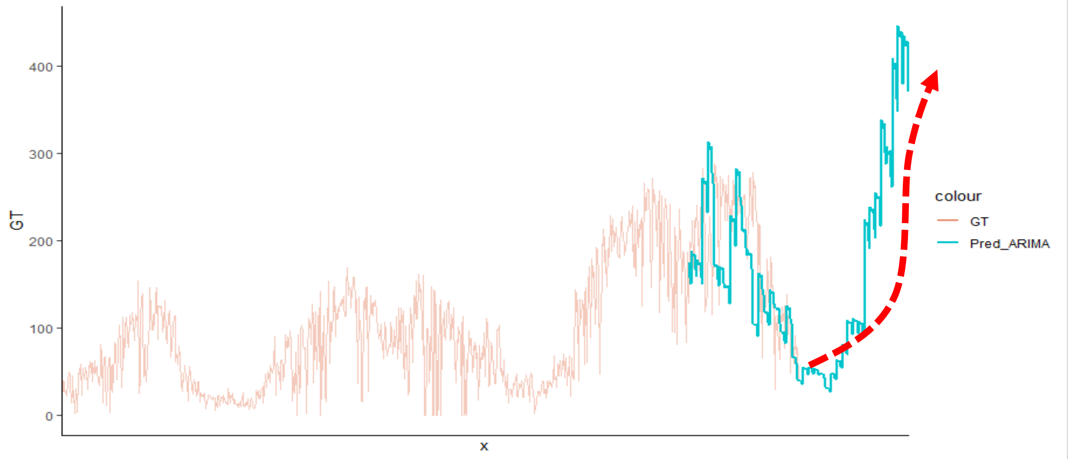
✓ 전년도 요일별 평균 대여 건수를 가중치로 활용하여 주별 예측된 대여 건수를 분배

2. Forecasting: Time Series

6) ARIMA

주별 데이터 vs 일별 데이터





- ✓ MAPE 약 14%p 감소, 기존에 계절성을 반영하지 못해 설명하지 못하던 급격한 감소 추세를 설명할 수 있는 모형 구축
- ✓ 더욱 긴 기간 예측 결과를 보면 3월 말부터 급격히 대여 건수가 반영되었음을 확인할 수 있음

Forecasting Performance

일별 데이터		주별 데이터	
MAE	64.91	MAE	51.51
MAPE	44.10	MAPE	30.57
RMSE	73.03	RMSE	73.03

2. Forecasting
Multivariate Data Analysis

- Data

date	cnt_station	cnt_bike	is_rain	is_holiday	weekday	weeknum	cnt
2017-06-21	743	9855	0	0	2	25	13.0
2017-06-22	743	9855	0	0	3	25	40.0
2017-06-23	790	10415	0	0	4	25	39.0
2017-06-24	790	10415	1	0	5	25	28.0
2017-06-25	792	10455	0	0	6	25	28.0
...	...	...	...	...	...	...	...
2019-12-27	1530	19474	0	0	4	52	81.0
2019-12-28	1530	19474	0	0	5	52	79.0
2019-12-29	1530	19474	0	0	6	52	61.0
2019-12-30	1530	19474	0	0	0	52	77.0
2019-12-31	1530	19474	0	0	1	52	39.0

924 rows × 10 columns

- Input Variables

Date

Weekday : 요일 정보
(월 : 0 , 화 : 1, … , 일 : 6)

Week number : 주차 정보
(1 ~ 52 주차)

Is holiday : 공휴일 정보
(0 : 공휴일 X / 1: 공휴일 O)

외부 변수

Station Count : 해당 시점 총 정류장 개수
(numerical value)

Bike Count : 해당 시점 총 자전거 개수
(numerical value)

Is rain : 비(강수량≦3mm) 정보
(0 : 비 X / 1: 비 O)

2. Forecasting: Multivariate Data Analysis

Support Vector Machine

데이터 집합을 바탕으로 하여 새로운 데이터가 어느 카테고리에 속할지 선형 분류 또는 비선형 분류 판단

- Hyper-parameter

〈 Kernel 〉

- Linear :

$$K(x_1, x_2) = x_1^T\, x_2$$

- Polynomial :

$$K(x_1, x_2) = (x_1^T\, x_2 + c)^d, \qquad c > 0$$

- Sigmoid :

$$K(x_1, x_2) = \tanh\{a(x_1^T\, x_2) + b\}, \qquad a, b > 0$$

- Gaussian(rbf) :

$$K(x_1, x_2) = \exp\{-\frac{||x_1 - x_2||_2^2}{2\sigma^2}\}, \ \sigma \neq 0$$

〈 Hyper-parameter 〉

- C :

  오차 penalty 크기의 Hyper-parameter

- Gamma (γ) :

  결정 경계의 곡률 Hyper-parameter

- Epsilon (ε):

  오차 허용 정도의 Hyper-parameter

Support Vector Machine

Result

- Hyper-parameter tuning

comb	kernel	C	gamma	epsilon	MAE	MAPE	RMSE
1	rbf	100	0.1	10	96.74804	52.19012	104.0959
2	poly	0.1	0.1	100	99.51799	51.74427	109.143
3	poly	0.001	0.001	100	99.51799	51.74427	109.143
4	poly	0.001	0.01	100	99.51799	51.74427	109.143
⋮							
683	sigmoid	100	1	0.01	5908.054	3668.526	6756.238
684	sigmoid	100	1	0.001	5908.055	3668.527	6756.24

- Best Model

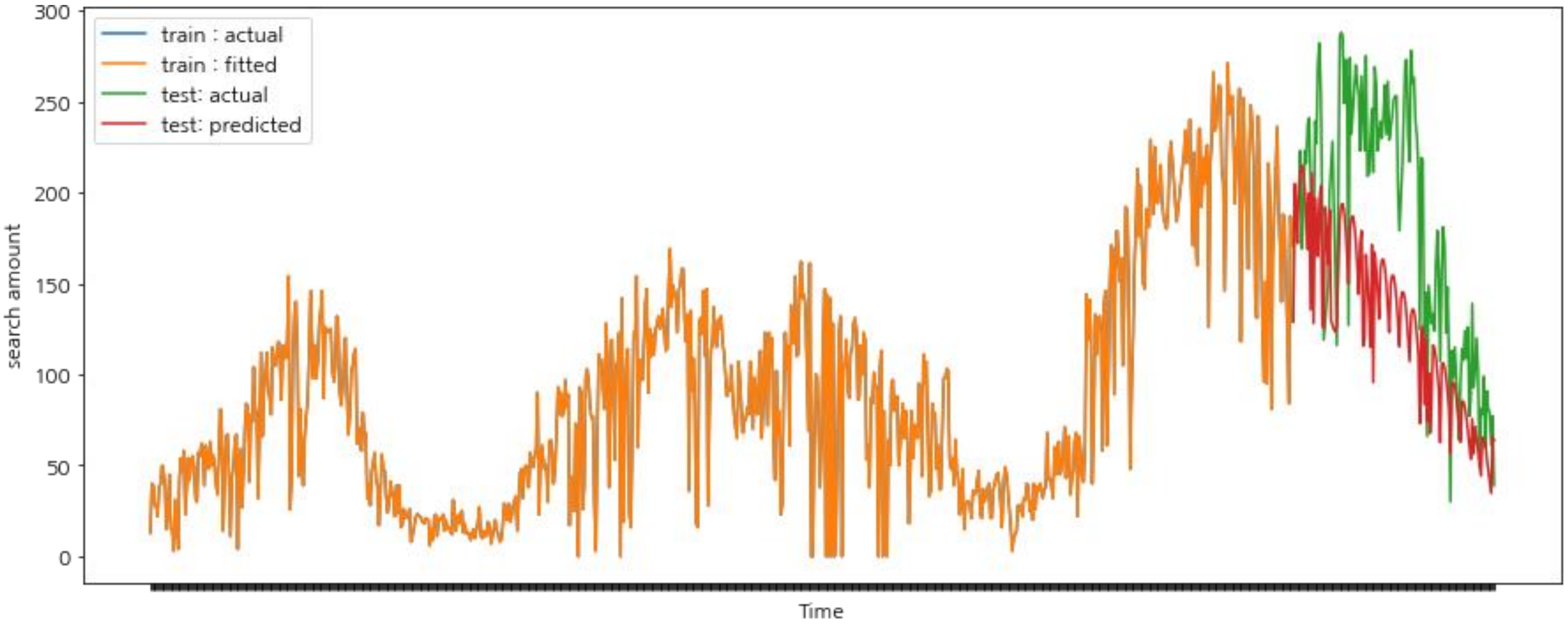
kernel	C	gamma	epsilon
rbf	100	0.1	10

Forecasting Performance	
MAE	59.4873
MAPE	49.0630
RMSE	70.0817

2. Forecasting: Multivariate Data Analysis

Support Vector Regression

Result



2. Forecasting

Best Model

Forecasting Performance

	Forecasting Model	MAE	MAPE▼	RMSE
1	SARIMA(1, 1, 2)(0, 1, 1)[52] (Weekly)	51.51	30.57	73.03
2	Linear Regression	53.31	32.34	62.91
3	Moving Average	68.29	42.80	78.35
4	ARIMA(3, 1, 1) (Daily)	64.91	44.10	73.03
5	Simple Exponential Smoothing	61.76	46.39	68.92
6	Trigonometric Model(model2)	69.73	46.41	78.60
7	SVM	59.49	49.06	70.08
8	Trigonometric Model(model1)	111.52	58.04	120.80
9	Double Exponential Smoothing	128.00	76.11	138.44
10	Additive Holt-Winters Exponential Smoothing	1677.11	1402.61	1952.97
11	Multiplicative Holt-Winters Exponential Smoothing	60124645.94	79370725.26	155145463.91

3. Hidden Markov Model

01 **Data**

Exploratory Data Analysis and
Data Preprocessing

02 **Forecasting**

Univariate time series,
ARIMA and Multivariate Data Analysis

03 **Hidden Markov Model**

Analyze and Forecast
Using Hidden Markov Model

04 **Result**

Conclusion and Insights

3. Hidden Markov Model

비 내리는 날 자전거 타기 불편한 것은 자명한 사실이다.
서울시 강수 여부와 안암로터리 따릉이 대여량의 숨겨진 상관관계를 알아보자.

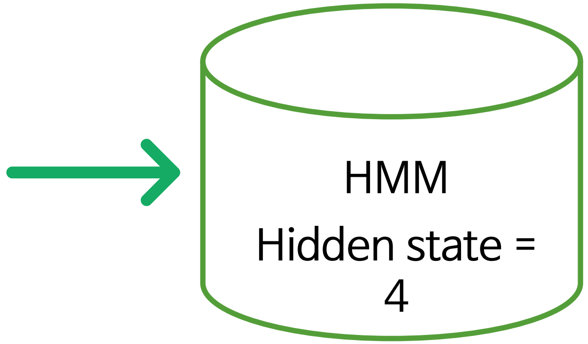
Modeling

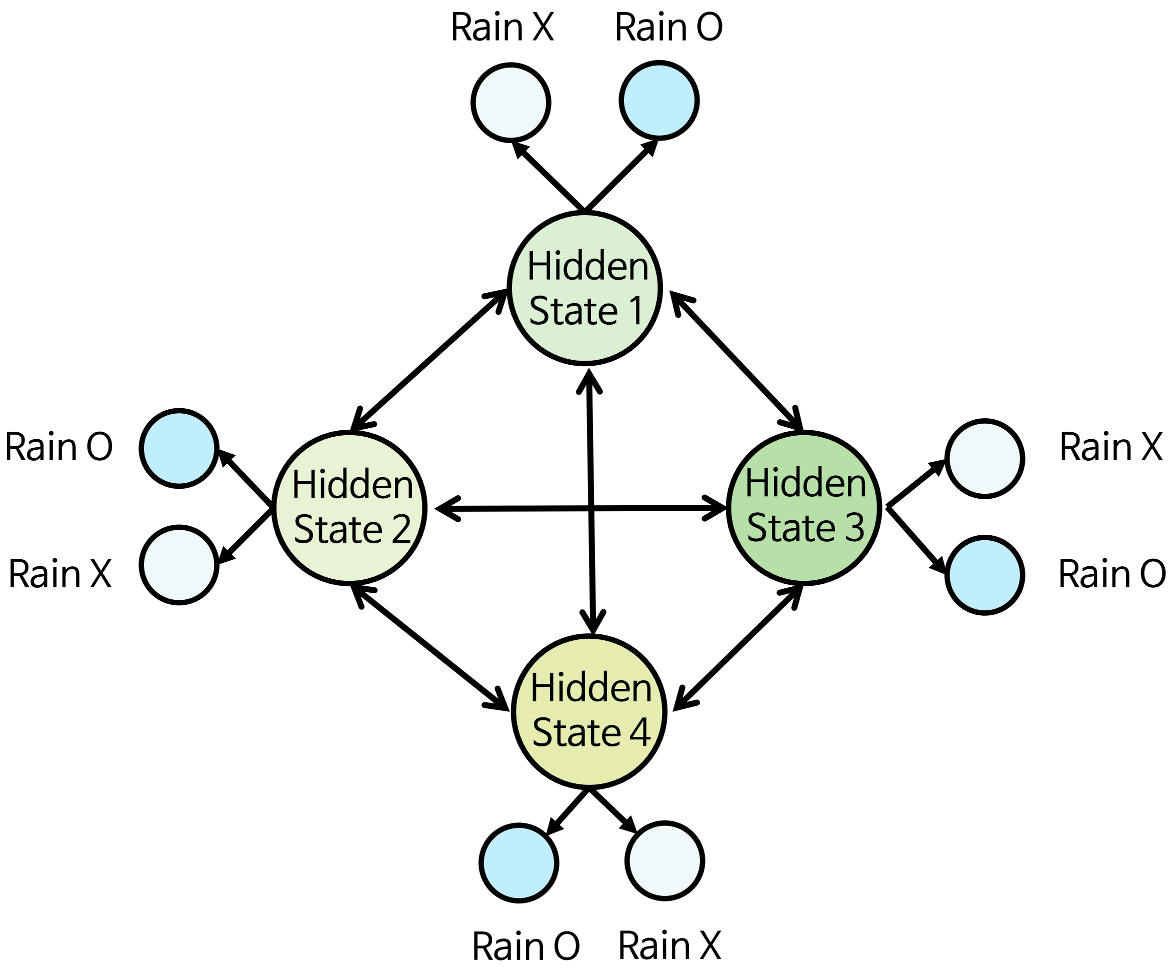
Observable component: 일별 강수 여부, 일일 대여량

- Input
  - ✓ 관찰된 강수 여부와 일일 대여량

date	is_rain	cnt
2017-06-21	0	13.0
2017-06-22	0	40.0
2017-06-23	0	39.0
2017-06-24	1	28.0
2017-06-25	0	28.0
...	...	...
2019-12-27	0	81.0
2019-12-28	0	79.0
2019-12-29	0	61.0
2019-12-30	0	77.0
2019-12-31	0	39.0

924 rows × 2 columns

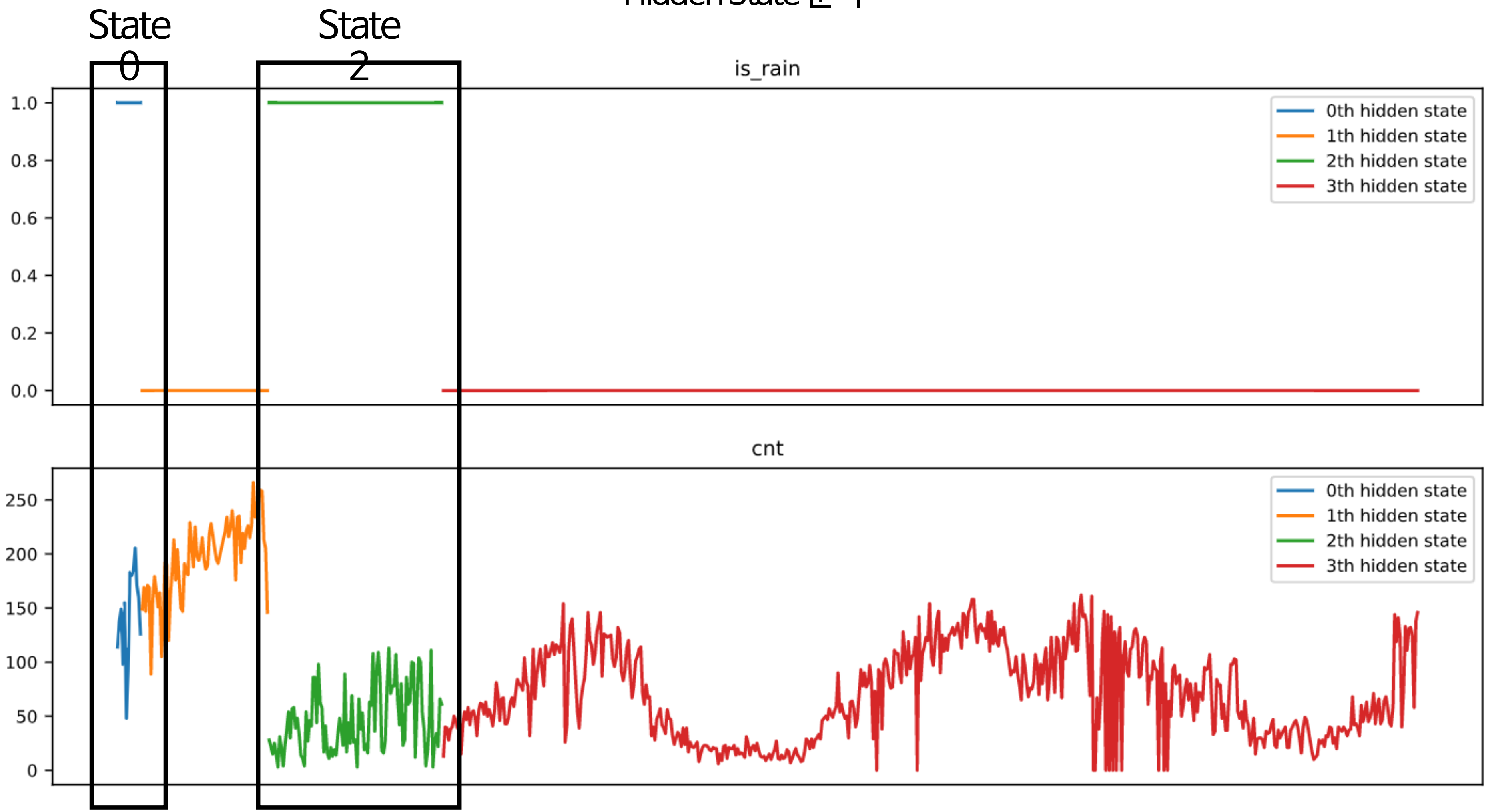




3. Hidden Markov Model

Result

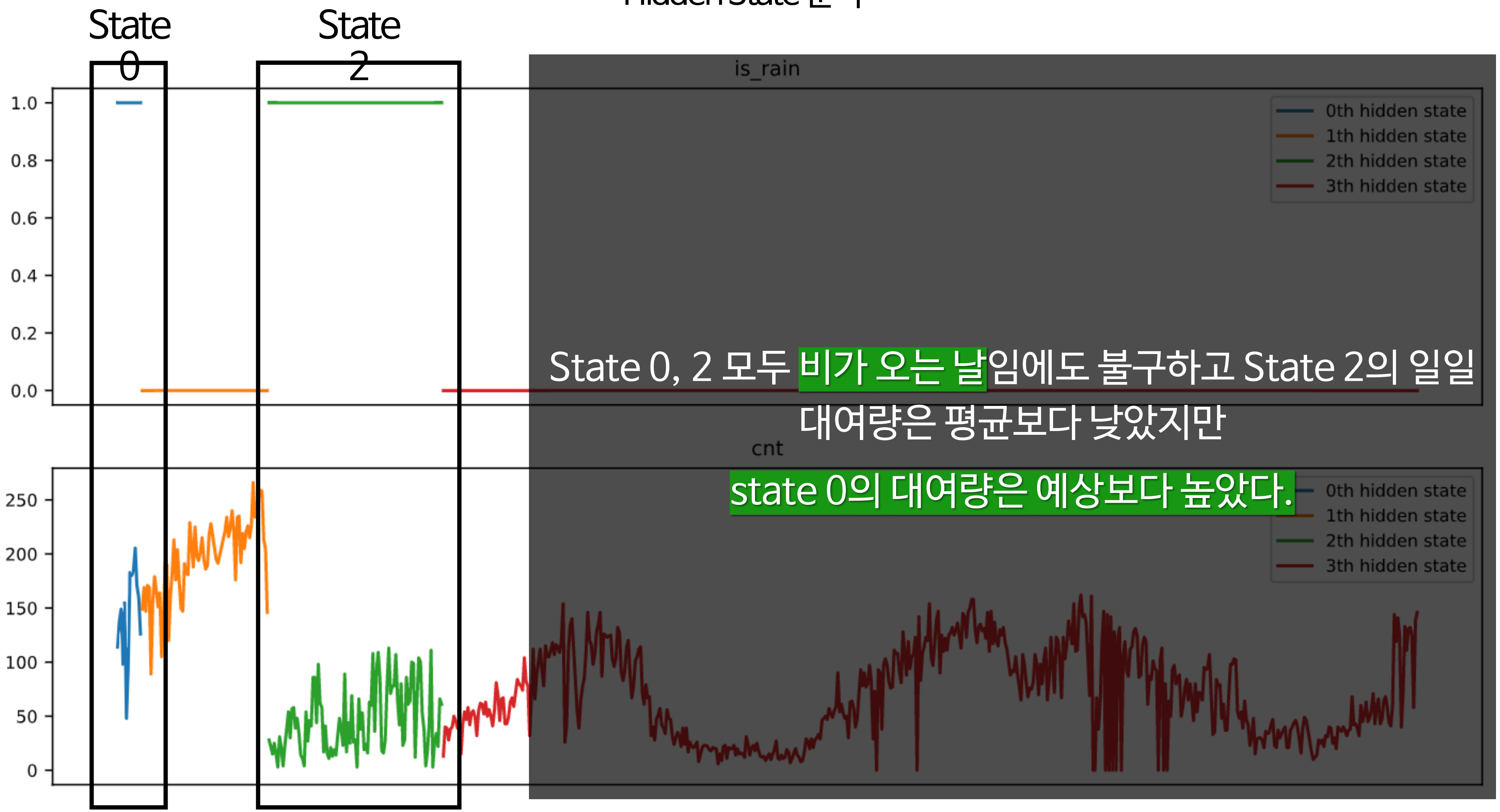
Hidden State 분석



3. Hidden Markov Model

Result

Hidden State 분석



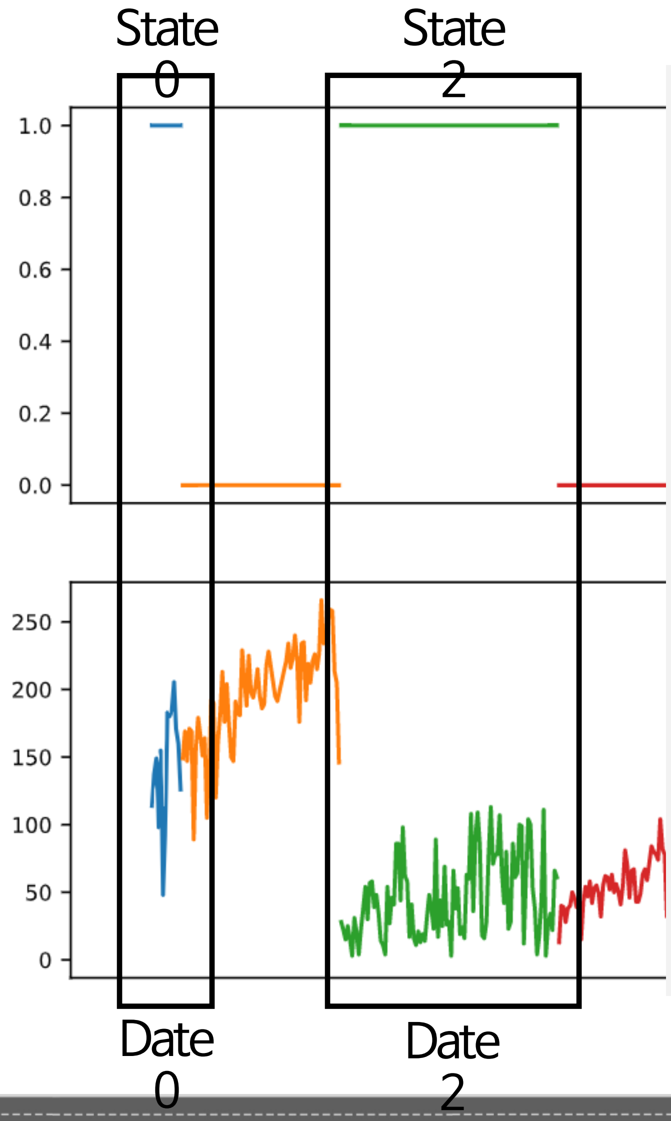
State 0, 2 모두 비가 오는 날임에도 불구하고 State 2의 일일 대여량은 평균보다 낮았지만

state 0의 대여량은 예상보다 높았다.

3. Hidden Markov Model

Result

Hidden State 분석



State 0		
Date	서울 강수량	안암 강수량
2018-06-11	10	0.075
2018-06-14	29	5.6
2018-06-15	16.5	0.183
2019-04-10	15.6	0
⋮	⋮	⋮

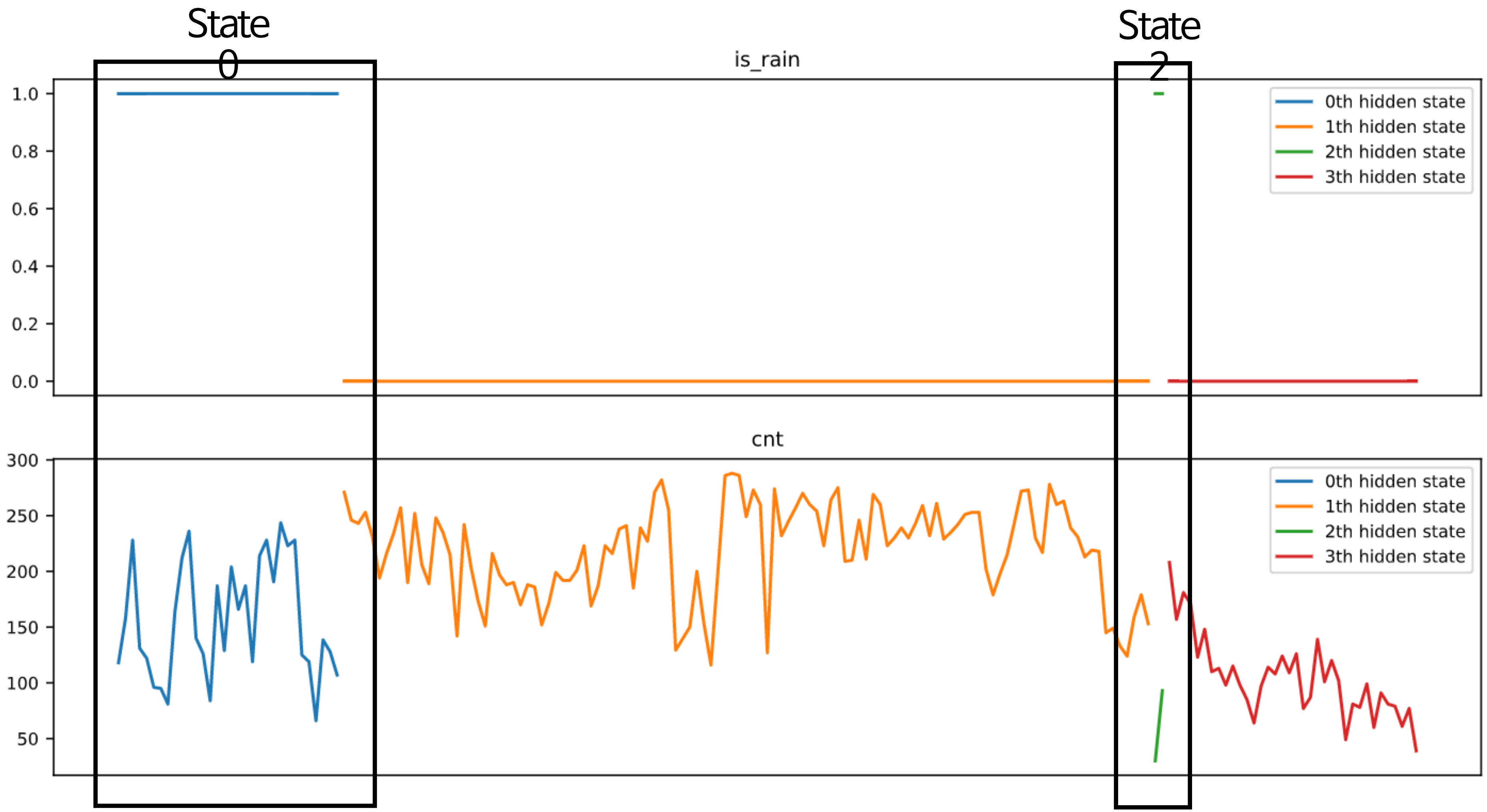
State 2		
Date	서울 강수량	안암 강수량
2017-06-24	8.5	4.458
2017-06-26	7.7	5.895
2017-07-02	6.3	6
2017-07-03	6.2	5.5
⋮	⋮	⋮

- ✓ State 0 : 서울 평균 강수량에 비해 안암의 강수량이 적은 경우
- ✓ State 2 : 서울 평균 강수량과 안암의 강수량이 비슷한 경우

3. Hidden Markov Model

Result

Test



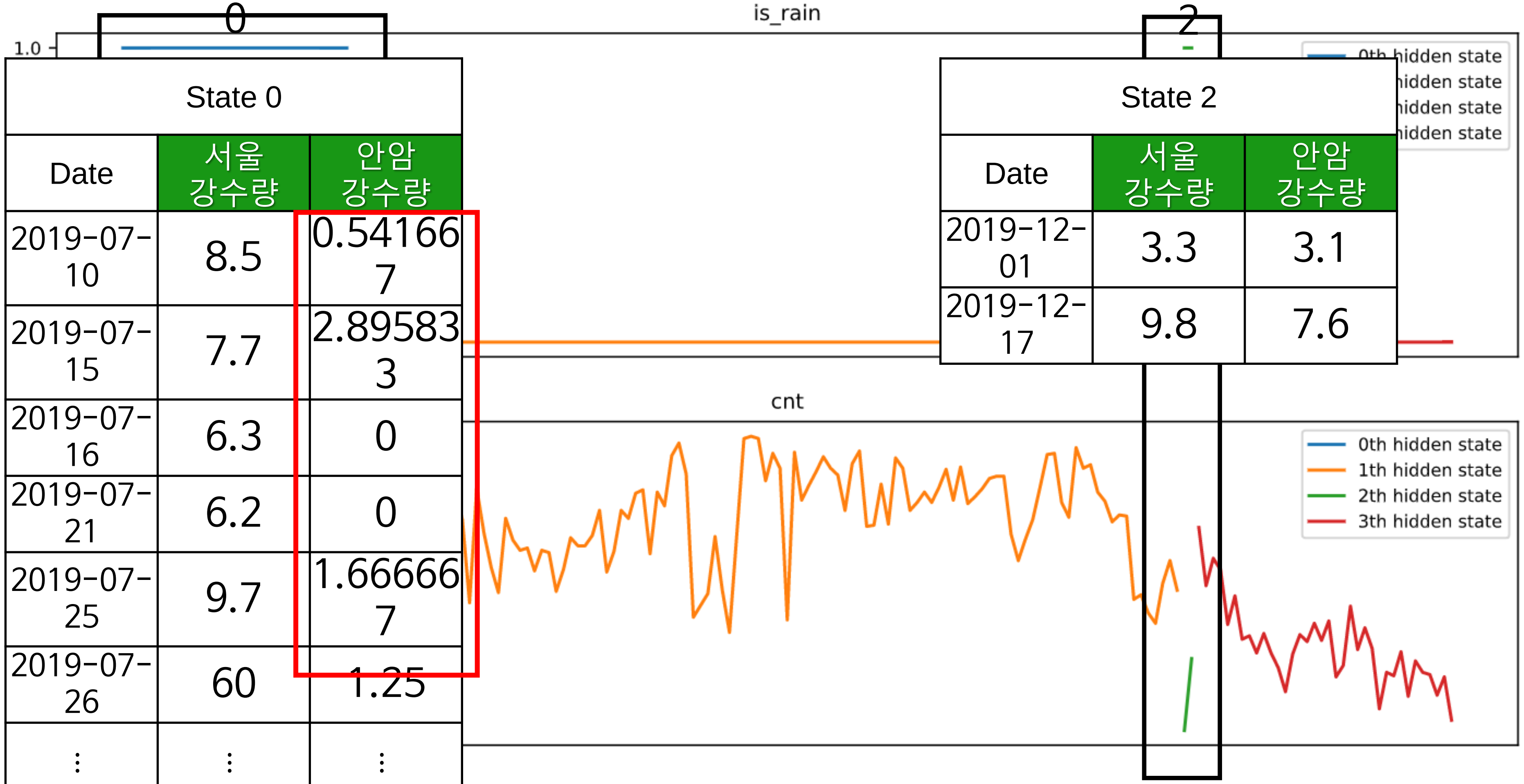
3. Hidden Markov Model

Result

Test

State 0

State 2



State 0		
Date	서울 강수량	안암 강수량
2019-07-10	8.5	0.541667
2019-07-15	7.7	2.895833
2019-07-16	6.3	0
2019-07-21	6.2	0
2019-07-25	9.7	1.666667
2019-07-26	60	1.25
⋮	⋮	⋮

State 2		
Date	서울 강수량	안암 강수량
2019-12-01	3.3	3.1
2019-12-17	9.8	7.6

4. Result

01
Data
Exploratory Data Analysis and
Data Preprocessing

02
Forecasting
Univariate time series,
ARIMA and Multivariate Data Analysis

03
Hidden Markov Model
Analyze and Forecast
Using Hidden Markov Model

04
Result
Conclusion and Insights

4. Result

Conclusion and Insights

TIME SERIES

- 계절성을 반영하는 것이 예측 성능을 결정하는 중요한 요소이다.
- ARIMA는 굉장히 강력한 모형이다.

Multivariate Analysis (SVM)

- 각 시점 데이터를 i.i.d로 가정하여 예측했음에도 좋은 성능을 내는 것을 확인하였다.
- 적절한 외부변수를 사용했을 때 트렌드를 잡아줄 수 있다.

Hidden Markov Model

- 예측에 있어서 미처 고려하지 못한 요소를 발견할 수 있게 도와준다.

감사합니다