

상점 매출 예측

feat. 신용카드 매출

21년 4월 R과 파이썬을 활용한 빅데이터 분석 전문가 양성 과정 발표



“ **상환 기간의 매출을 예측하여 신용 점수가 낮거나
담보를 가지지 못하는 우수 상점들에 금융 기회를 제공** ”

‘당신의 가게도 대출을 받을 수 있을까?’

00. CONTEXT

01. 문제 분석

- 데이터 선정 이유
- 데이터 분석 목표
- 데이터 분석 과정

02. EDA

- raw data
- 시간 컬럼 처리
- 환불 데이터 제거
- 일별, 월별 데이터
- 처리 스토어별 데이터 분포

03. modeling

- 정상성 분석
- ARIMA
- ETS
- HoltWinters

04. 분석 결과

- 3개월 매출 예측
- 시사점 및 한계점

05. Q & A

- 질의 응답

01. 문제 분석

- 데이터 선정 이유
- 데이터 분석 목표
- 데이터 분석 과정

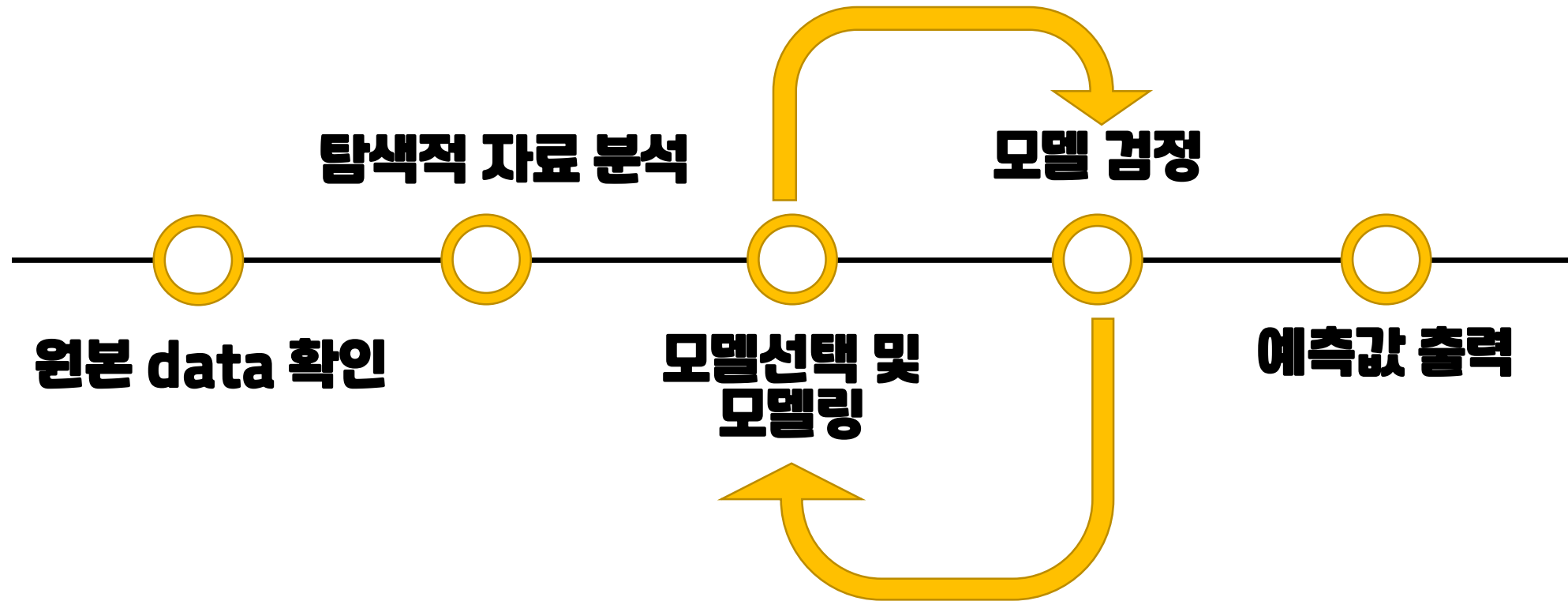


미래 예측, 시계열 분석

시간의 흐름에 따라 관찰된 자료를 분석해 미래의 값을 예측하고 경향, 주기, 계절성 파악하여 활용하는
분석방법

“시계열 분석을 통해 각 스토어별 3개월 매출 예측”





02. 탐색적 자료 분석

- RAW DATA
- 시간 컬럼 처리
- 환불 데이터 처리
- 일별, 월별 데이터 처리
- 스토어별 데이터 분포

1. 데이터 자료 구조

```
> str(data)
'data.frame': 6556613 obs. of 9 variables:
 $ store_id      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ card_id       : int  0 1 2 3 4 5 6 7 8 9 ...
 $ card_company  : chr   "b" "h" "c" "a" ...
 $ transacted_date : chr  "2016-06-01" "2016-06-01" "2016-06-01" "2016-06-01" ...
 $ transacted_time : chr  "13:13" "18:12" "18:52" "20:22" ...
 $ installment_term: int  0 0 0 0 0 0 0 0 0 0 ...
 $ region        : chr   "" "" "" "" ...
 $ type_of_business: chr  "기타 미용업" "기타 미용업" "기타 미용업" "기타 미용업" ...
 $ amount        : num  1857 857 2000 7857 2000 ...
```

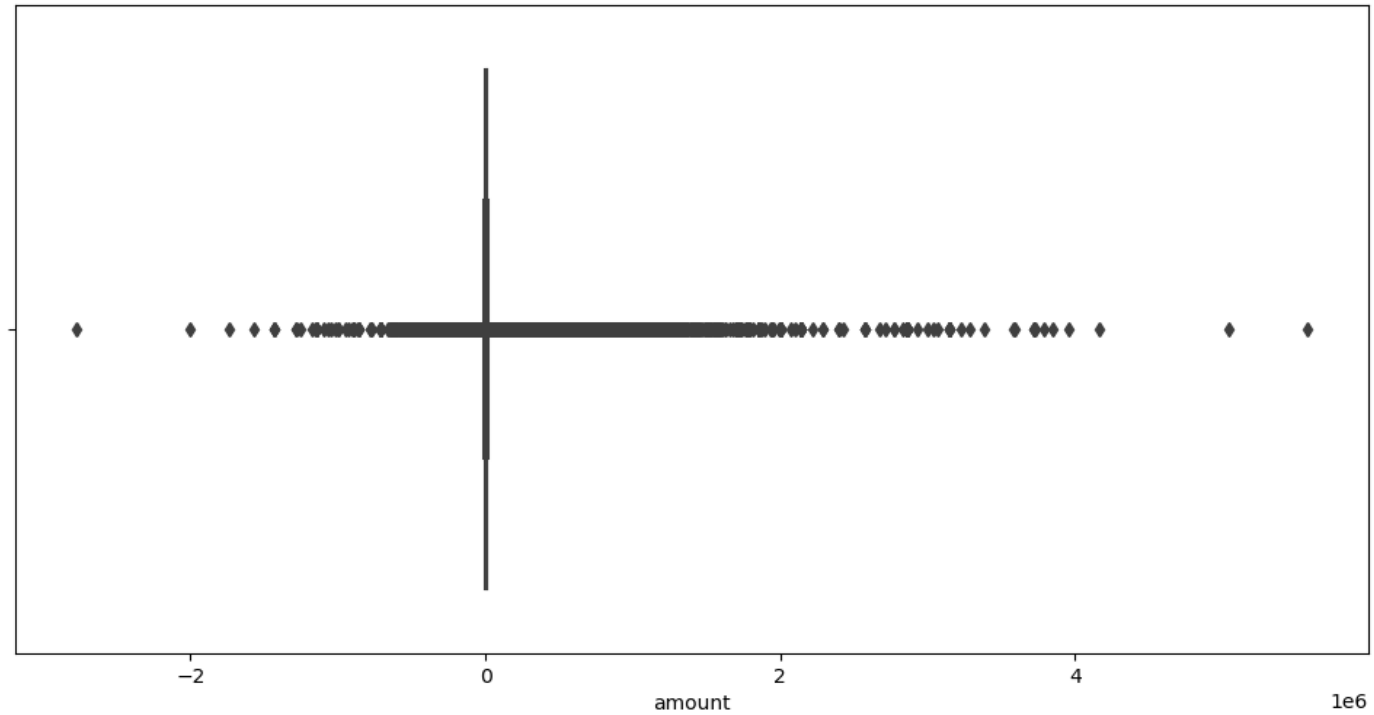
- 종속 변수 : amount
- 독립 변수 : store_id, card_id, card_company, transacted_date, transacted_time, installment_term, region, type_of_business

1. 데이터 자료 구조

```
> str(data)
'data.frame': 6556613 obs. of 9 variables:
 $ store_id      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ card_id       : int  0 1 2 3 4 5 6 7 8 9 ...
 $ card_company  : chr   "b" "h" "c" "a" ...
 $ transacted_date : chr  "2016-06-01" "2016-06-01" "2016-06-01" "2016-06-01" ...
 $ transacted_time : chr  "13:13" "18:12" "18:52" "20:22" ...
 $ installment_term: int  0 0 0 0 0 0 0 0 0 0 ...
 $ region        : chr   "" "" "" "" ...
 $ type_of_business: chr  "기타 미용업" "기타 미용업" "기타 미용업" "기타 미용업" ...
 $ amount        : num  1857 857 2000 7857 2000 ...
```

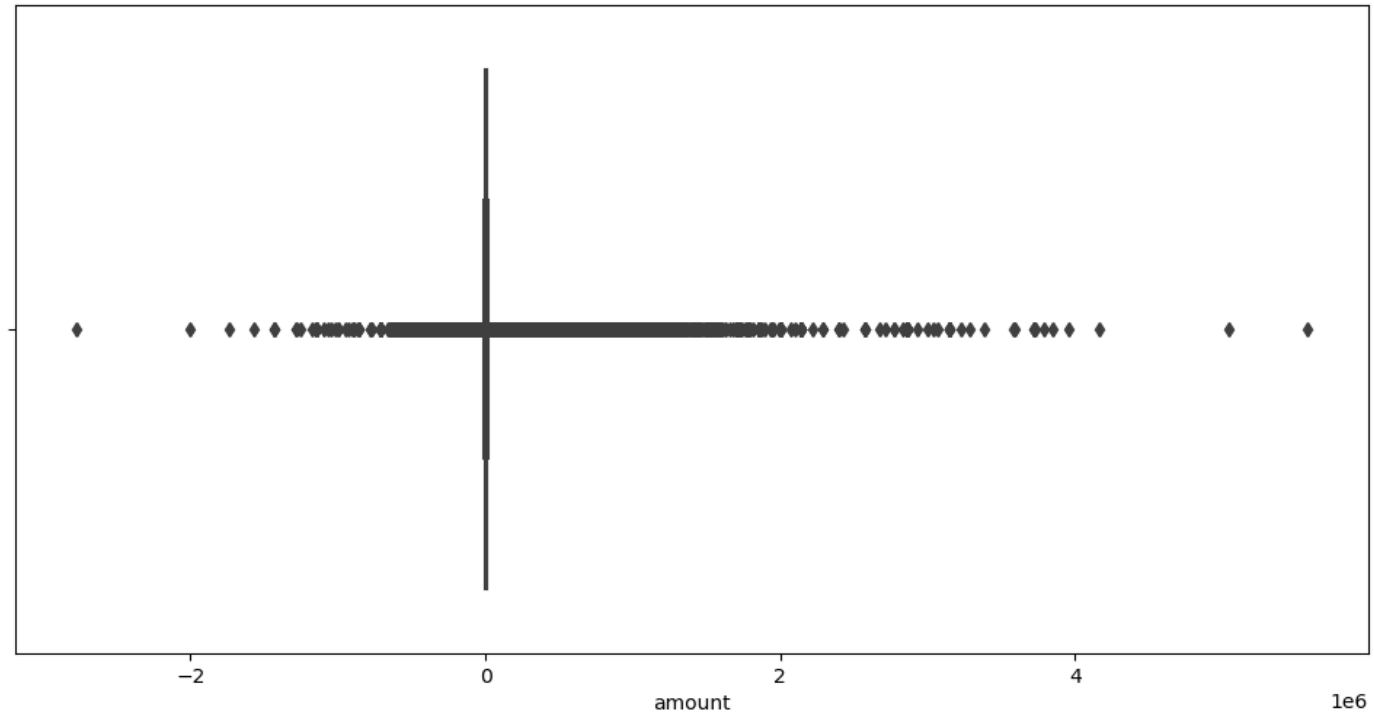
- 종속 변수 : amount
- 독립 변수 : store_id, card_id, card_company, transacted_date, transacted_time, installment_term, region, type_of_business

2. 전체 데이터 분포



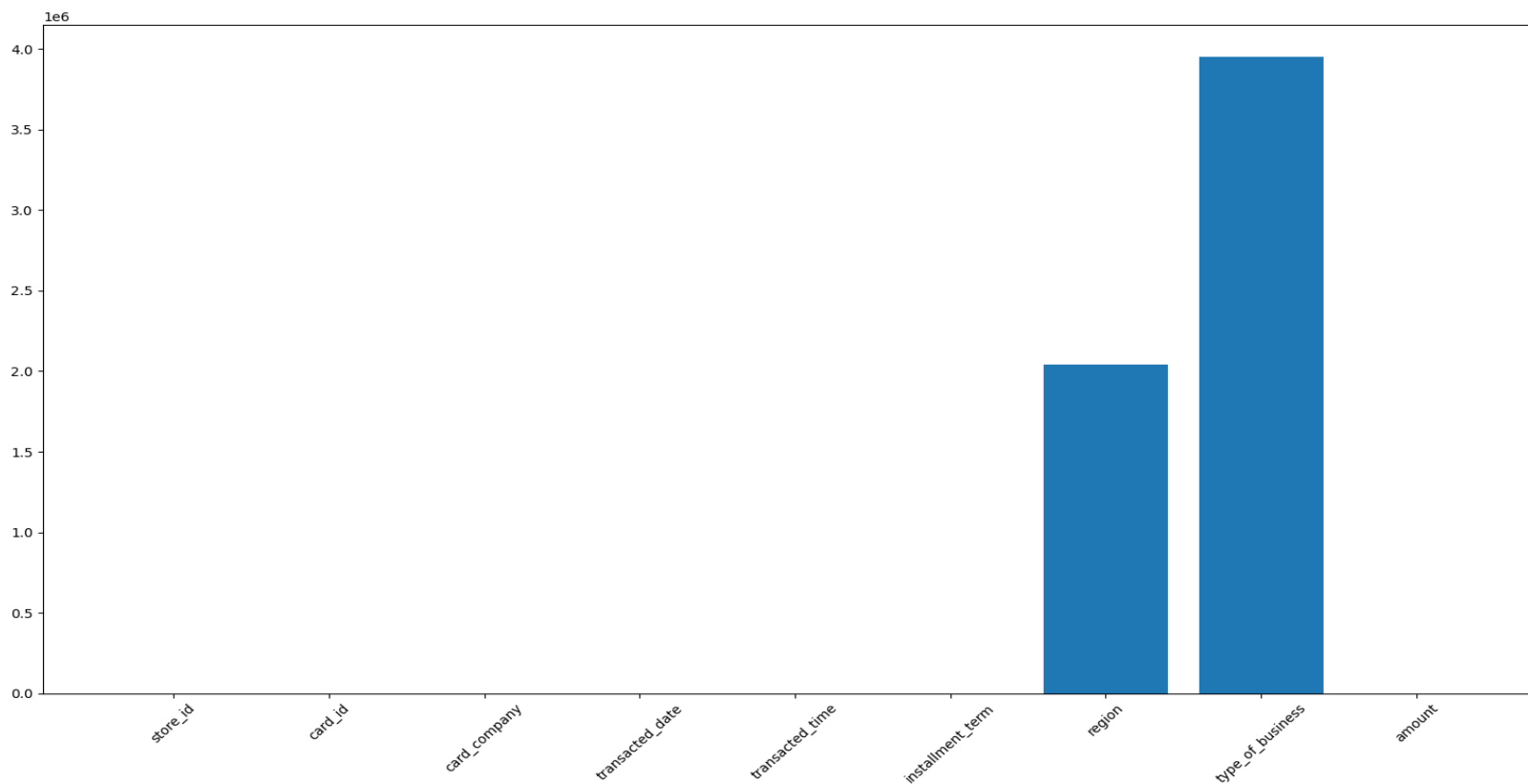
“마이너스 값 존재?”

2. 전체 데이터 분포



“환불 값 존재!”

3. 결측치 확인



“결측치 존재”
region, type of business 결측치

1. 시간 컬럼 합치기

**Transacted_date +
Transacted_time**

Date 컬럼 생성

2. 환불 데이터 제거

ALL AMOUNT

store_id	card_id	amount
0	38	14285.714286
0	39	9000.000000
0	40	8571.428571
0	40	-8571.428571
0	40	8571.428571
0	41	1857.142857
0	42	12857.142857
0	43	1428.571429
0	44	7142.857143
0	45	1857.142857
0	46	7142.857143
0	47	1857.142857



PLUS AMOUNT

store_id	card_id	amount
0	38	14285.714286
0	39	9000.000000
0	40	8571.428571
0	41	1857.142857
0	42	12857.142857
0	43	1428.571429
0	44	7142.857143
0	45	1857.142857
0	46	7142.857143
0	47	1857.142857
0	48	7857.142857
0	49	14285.714286

**환불 내역과 결제내역의
STORE_ID, CARD_ID, AMOUNT 가 같을 때
1 : 1 대응 시켜 데이터 삭제**

2. 환불 데이터 제거

ALL AMOUNT

store_id	card_id	amount
0	38	14285.714286
0	40	8571.428571
0	40	-8571.428571
0	40	8571.428571
0	42	12857.142857
0	43	1428.571429
0	44	7142.857143
0	45	1857.142857
0	46	7142.857143
0	47	1857.142857



PLUS AMOUNT

store_id	card_id	amount
0	38	14285.714286
0	39	9000.000000
0	40	8571.428571
0	41	1857.142857
0	42	12857.142857
0	43	1428.571429
0	44	7142.857143
0	45	1857.142857
0	46	7142.857143
0	47	1857.142857
0	48	7857.142857
0	49	14285.714286

**환불 내역과 결제내역의
STORE_ID, CARD_ID, AMOUNT 가 같을 때
1 : 1 대응 시켜 데이터 삭제**

2. 환불 데이터 제거

ALL AMOUNT

store_id	card_id	amount
0	38	14285.714286
0	40	8571.428571
0	40	-8571.428571
0	40	8571.428571
0	42	12857.142857
0	43	1428.571429
0	44	7142.857143
0	45	1857.142857
0	46	7142.857143
0	47	1857.142857



PLUS AMOUNT

store_id	card_id	amount
0	38	14285.714286
0	39	9000.000000
0	40	8571.428571
0	41	1857.142857
0	42	12857.142857
0	43	1428.571429
0	44	7142.857143
0	45	1857.142857
0	46	7142.857143
0	47	1857.142857
0	48	7857.142857
0	49	14285.714286

**환불 내역과 결제내역의
STORE_ID, CARD_ID, AMOUNT 가 같을 때
1 : 1 대응 시켜 데이터 삭제**

3. 일별, 월별 데이터 변환

일별 데이터

	store_id	transacted_date	amount
1	0	2016-06-01	12571.429
2	0	2016-06-02	40571.429
3	0	2016-06-03	18142.857
4	0	2016-06-04	31714.286
5	0	2016-06-05	10428.571
6	0	2016-06-06	17285.714
7	0	2016-06-07	NA
8	0	2016-06-08	NA
9	0	2016-06-09	35000.000
10	0	2016-06-10	53000.000
11	0	2016-06-11	64428.571
12	0	2016-06-12	51857.143
13	0	2016-06-13	48714.286
14	0	2016-06-14	NA
15	0	2016-06-15	11428.571

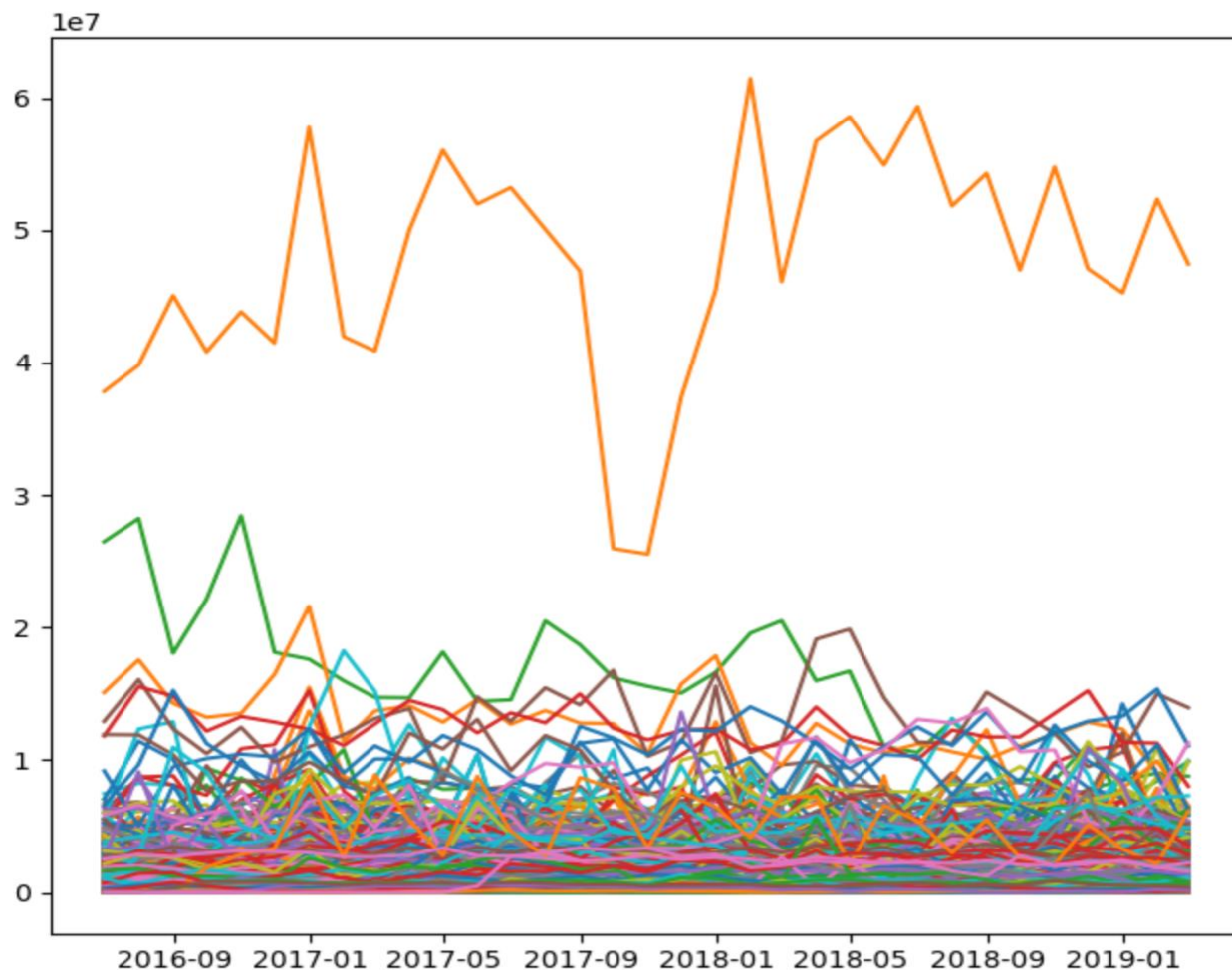
월별 데이터

	store_id	transacted_date	amount
1	0	2016-06	747000.0
2	0	2016-07	1005000.0
3	0	2016-08	871571.4
4	0	2016-09	897857.1
5	0	2016-10	835428.6
6	0	2016-11	697000.0
7	0	2016-12	761857.1
8	0	2017-01	585642.9
9	0	2017-02	794000.0
10	0	2017-03	720257.1
11	0	2017-04	685285.7
12	0	2017-05	744428.6
13	0	2017-06	682000.0
14	0	2017-07	728285.7
15	0	2017-08	749000.0

시계열 데이터 일자로 변환 →

월별로 downsampling

4. 스토어별 데이터 분포 확인



**각각의 스토어별로
모두 다른 추세와 계절성을 보임**

**즉, 하나의 시계열 모델로
설명이 어려우므로
스토어별로 모델링 진행**

5. 사용 모델 설명

✓ MA(이동평균)

- 데이터의 평균값이 시간에 따라 변화하는 경향
- 입력 데이터의 정상성이 가정 되지 않음
- 잔차(white noise)의 의미
- MA 는 미래와 과거의 가중치가 같지만 지수평활은 미래에 더 가중치를 둔다.

✓ ARIMA 모델

- AR(자기상관) + I + MA(이동평균)
- 자기상관
: 자기 자신 이전의 값이 이후의 값에 영향을 미치는 상황
- 가정 : 모델이 정상성을 가지고 있어야 함

✓ AUTO ARIMA

ARIMA 의 모델 중 AIC(아카이케값)이 가장 낮은 모수를 추천해주는 모델

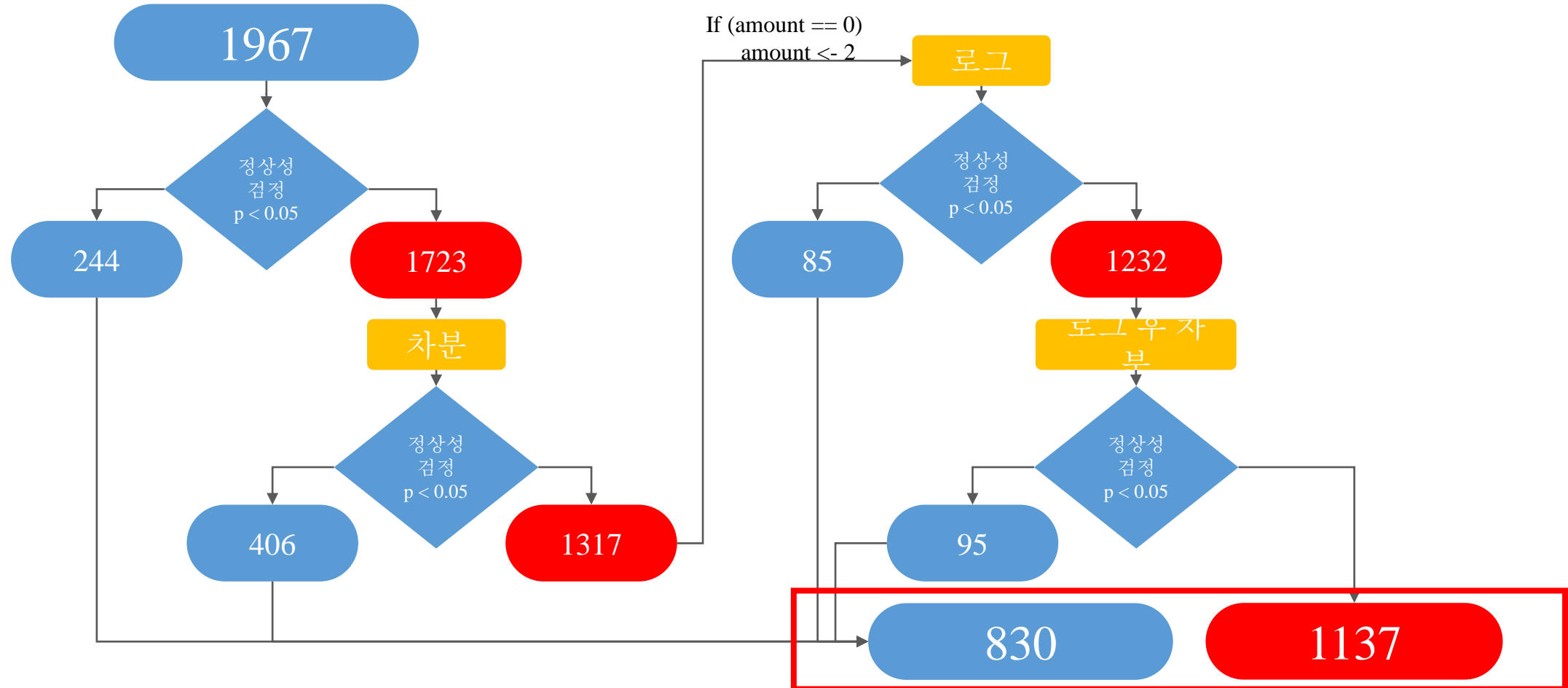
5. 사용 모델 설명

ETS / holt winters 모델

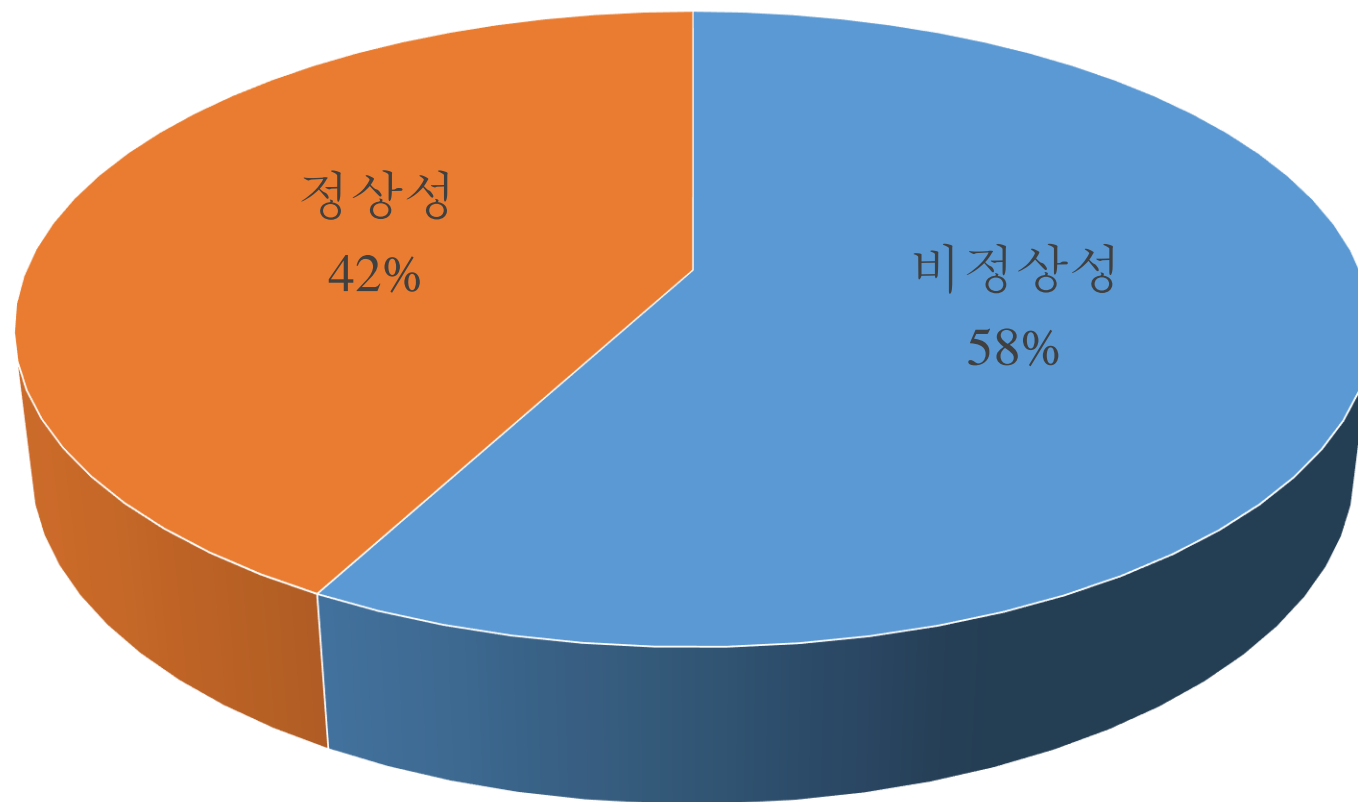
- 지수평활기법 중 추세와 계절성을 고려한 모델
- 과거의 모든 자료에 동일한 가중치 X
- 시스템이 변화한 최근 시점에 큰 가중치를 두어 예측 -> 지수평활 기법 중 ETS, Holt-winters 모델

03. modeling

- 정상성 검정
- Auto Arima
- ETS
- HoltWinters




정상성/비정상성 비율



■ 비정상성 ■ 정상성

1. 파라미터 제어

$$(1 - \phi_1 B - \dots - \phi_p B^p) \quad (1 - B)^d y_t = c + (1 + \theta_1 B + \dots + \theta_q B^q) \varepsilon_t$$



AR(p)

↑

d differences

↑

MA(q)

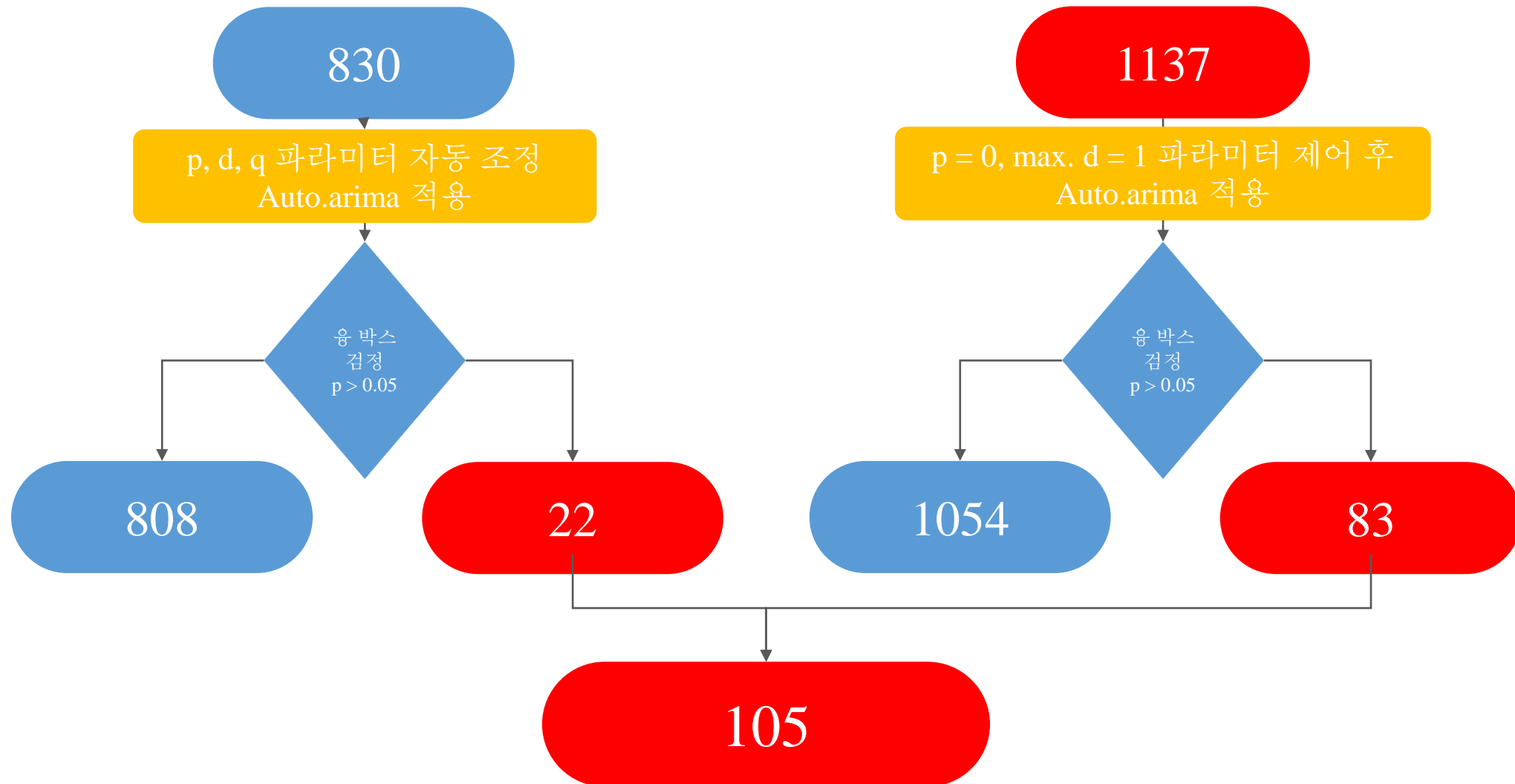
p = 자기회귀 부분의 차수;
d = 1차 차분이 포함된 정도;
q = 이동 평균 부분의 차수.

830

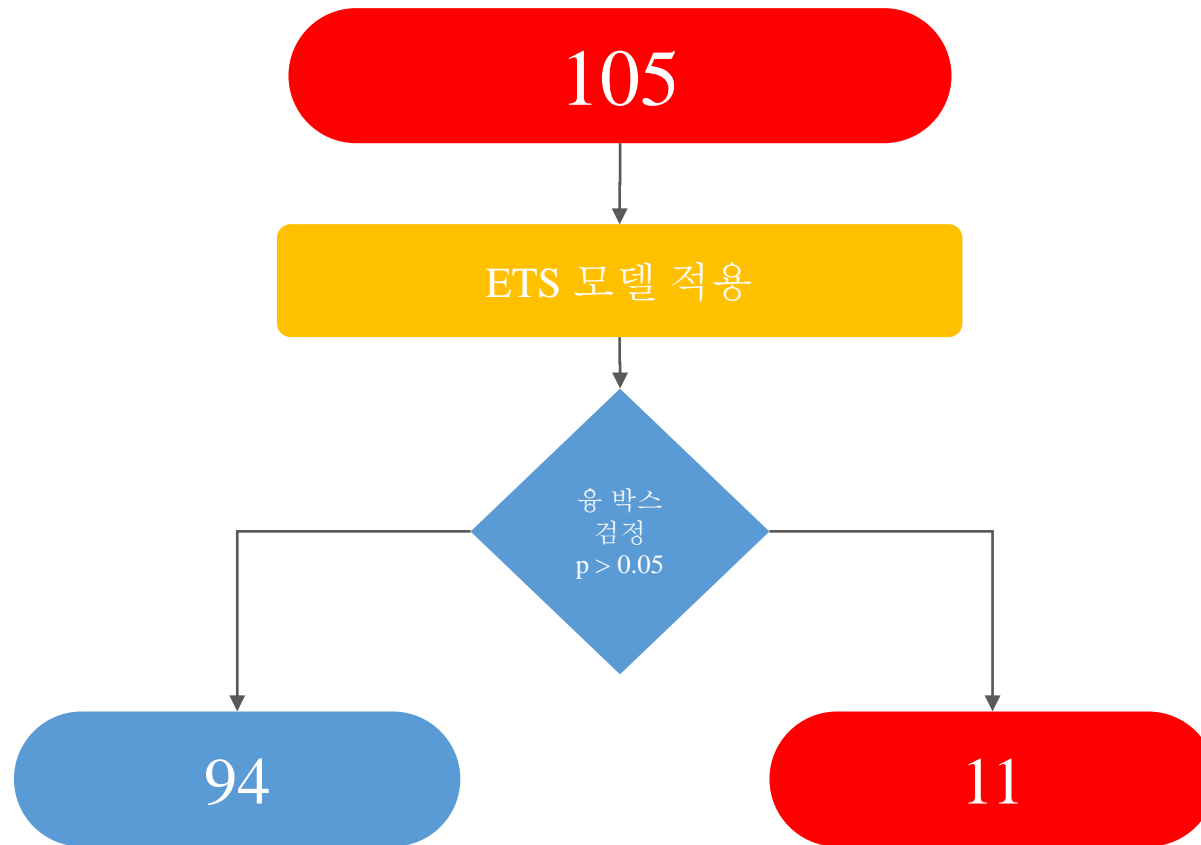
자동으로 최적 파라미터 (p, d, q) 설정

1137

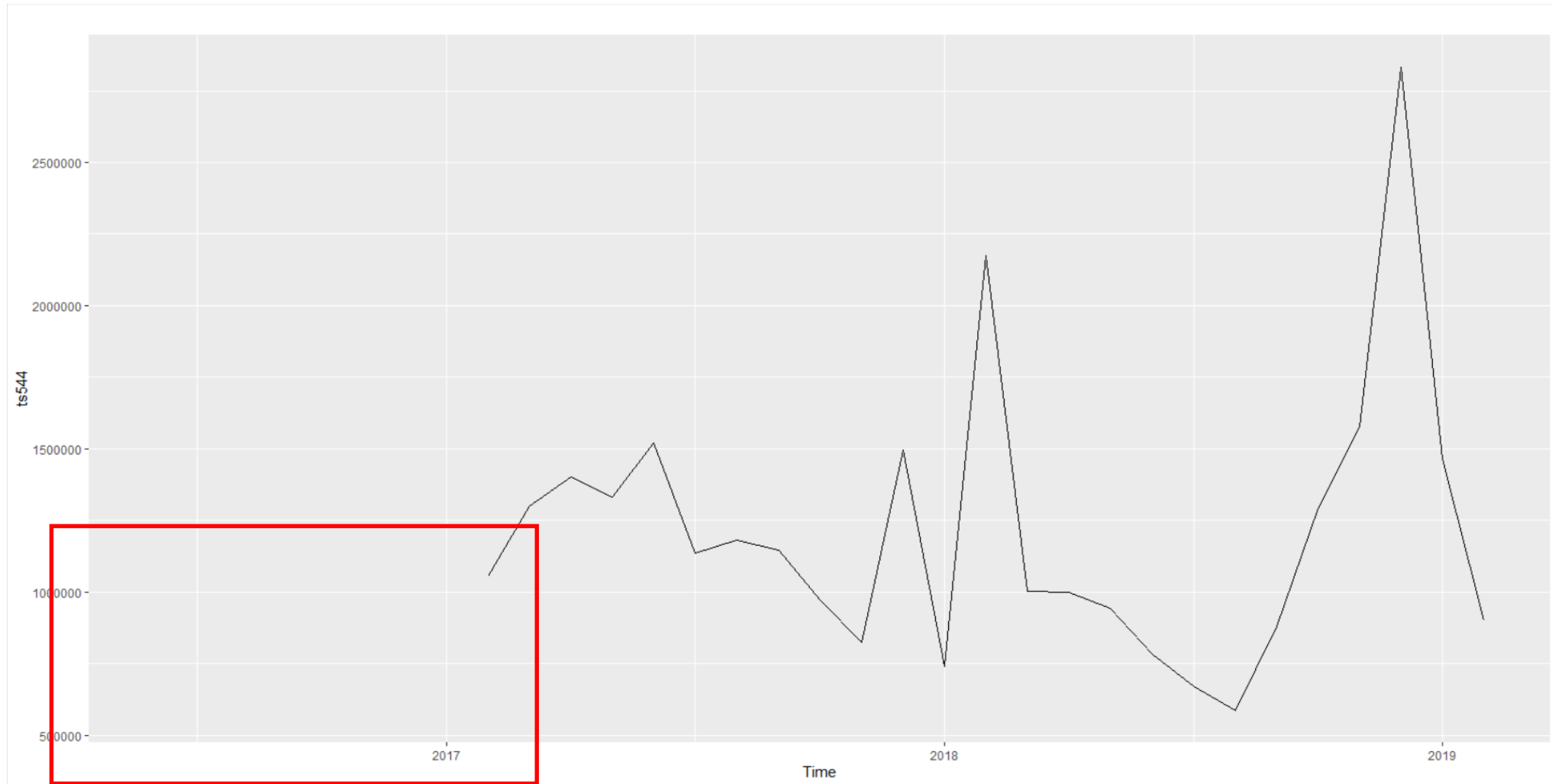
파라미터 (p = 0, max.d = 1, q) 조정



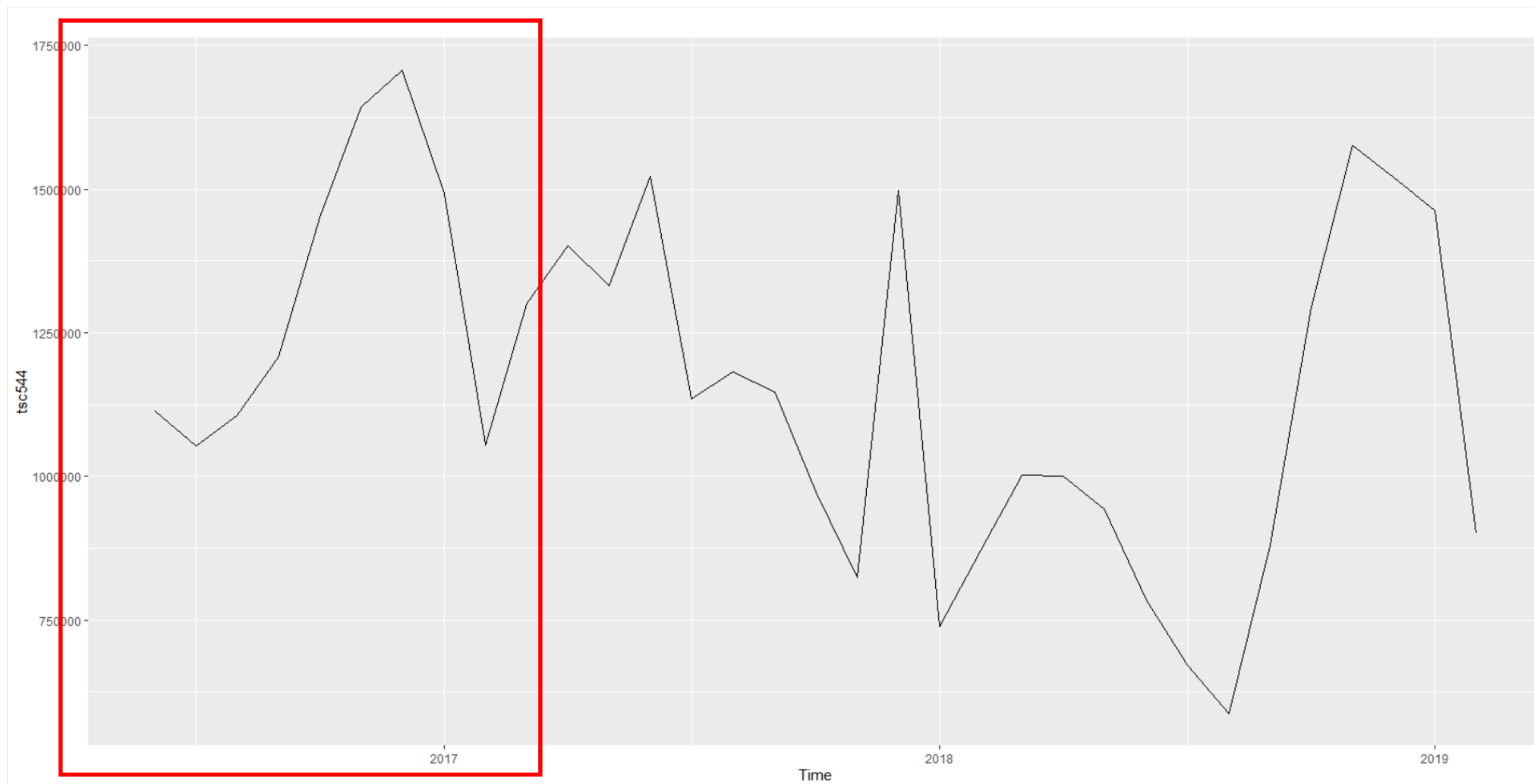
1. 105개 스토어 모델링



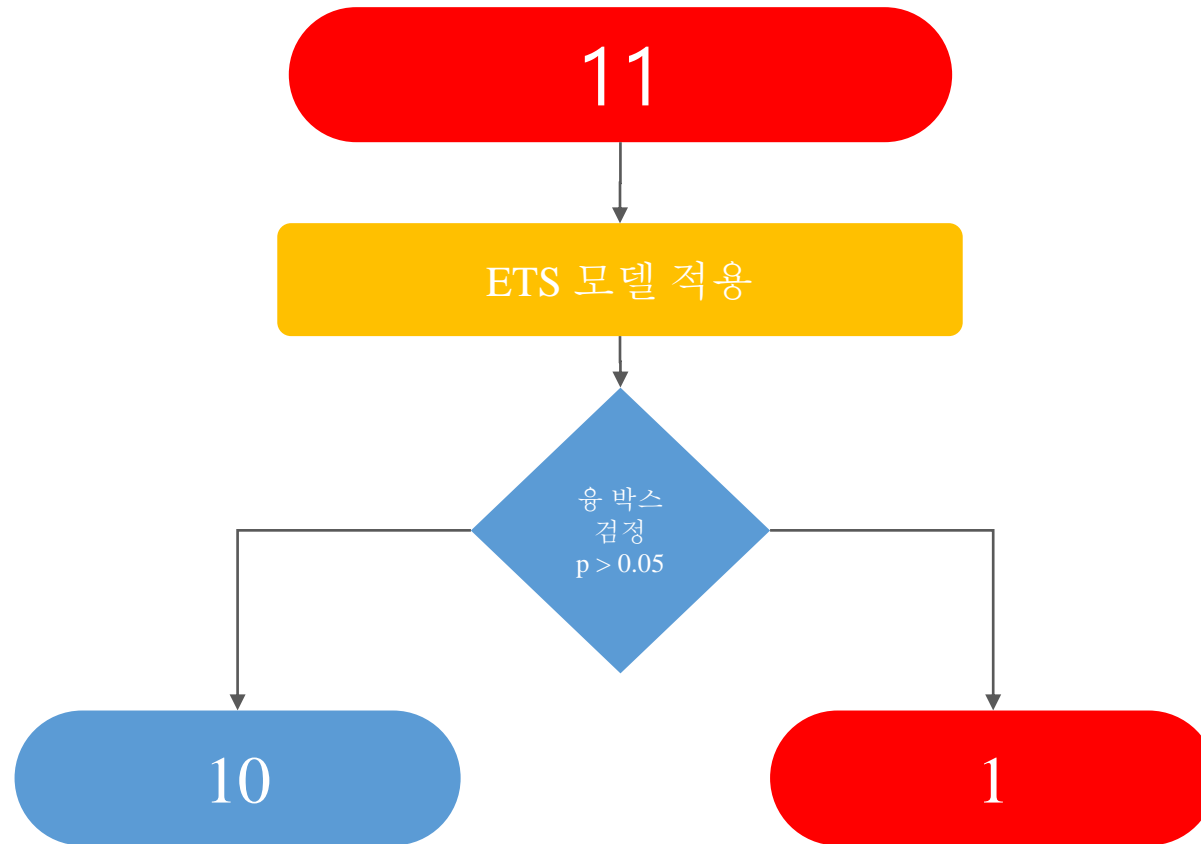
2. 이상치 및 결측치 확인

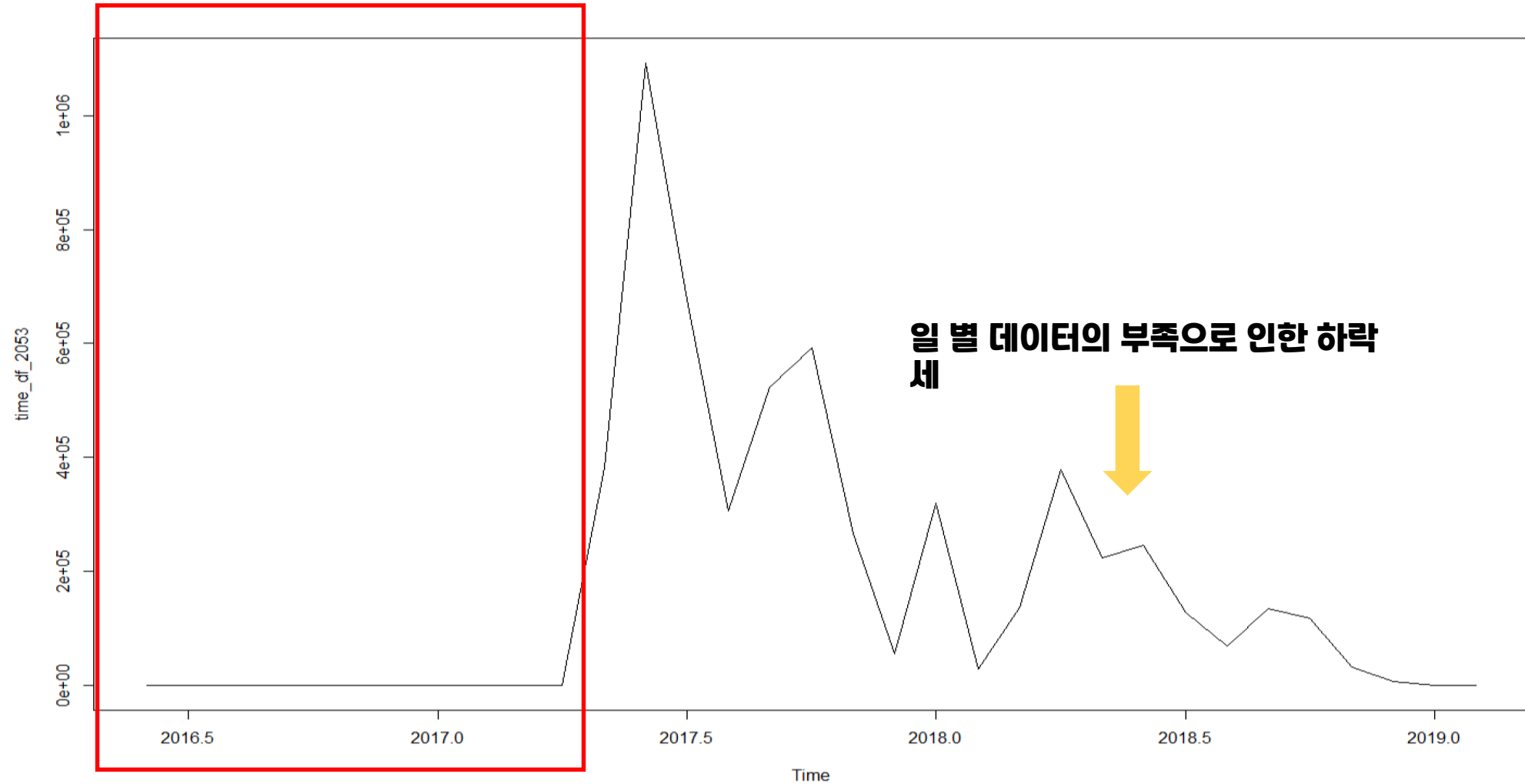


3. tsclean 적용



4. tsclean 후 ETS 모델링





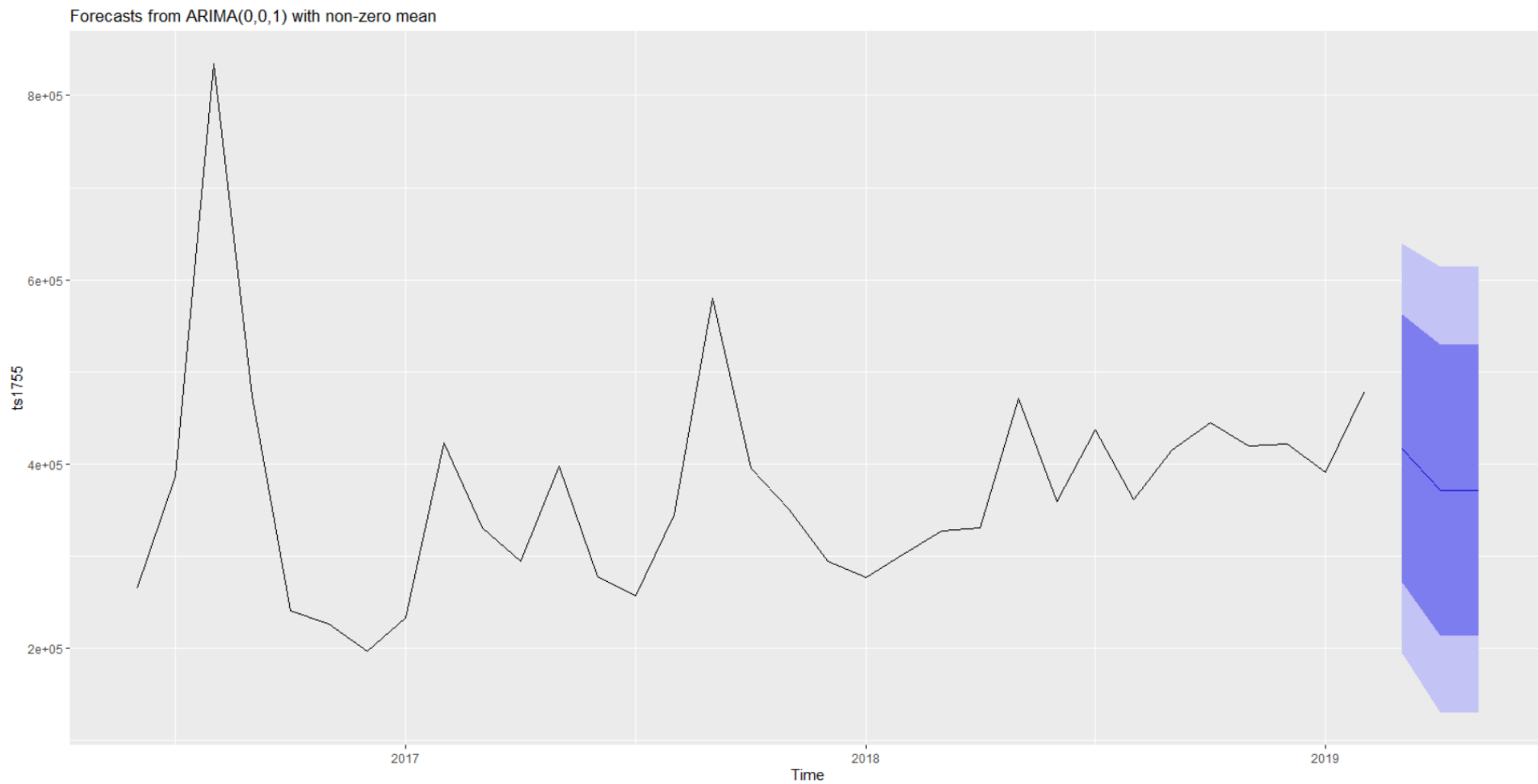
04. 분석 결과

- 3개월 매출 예측
- 시사점 및 한계점

3개월 매출 예측

4. 분석 결과

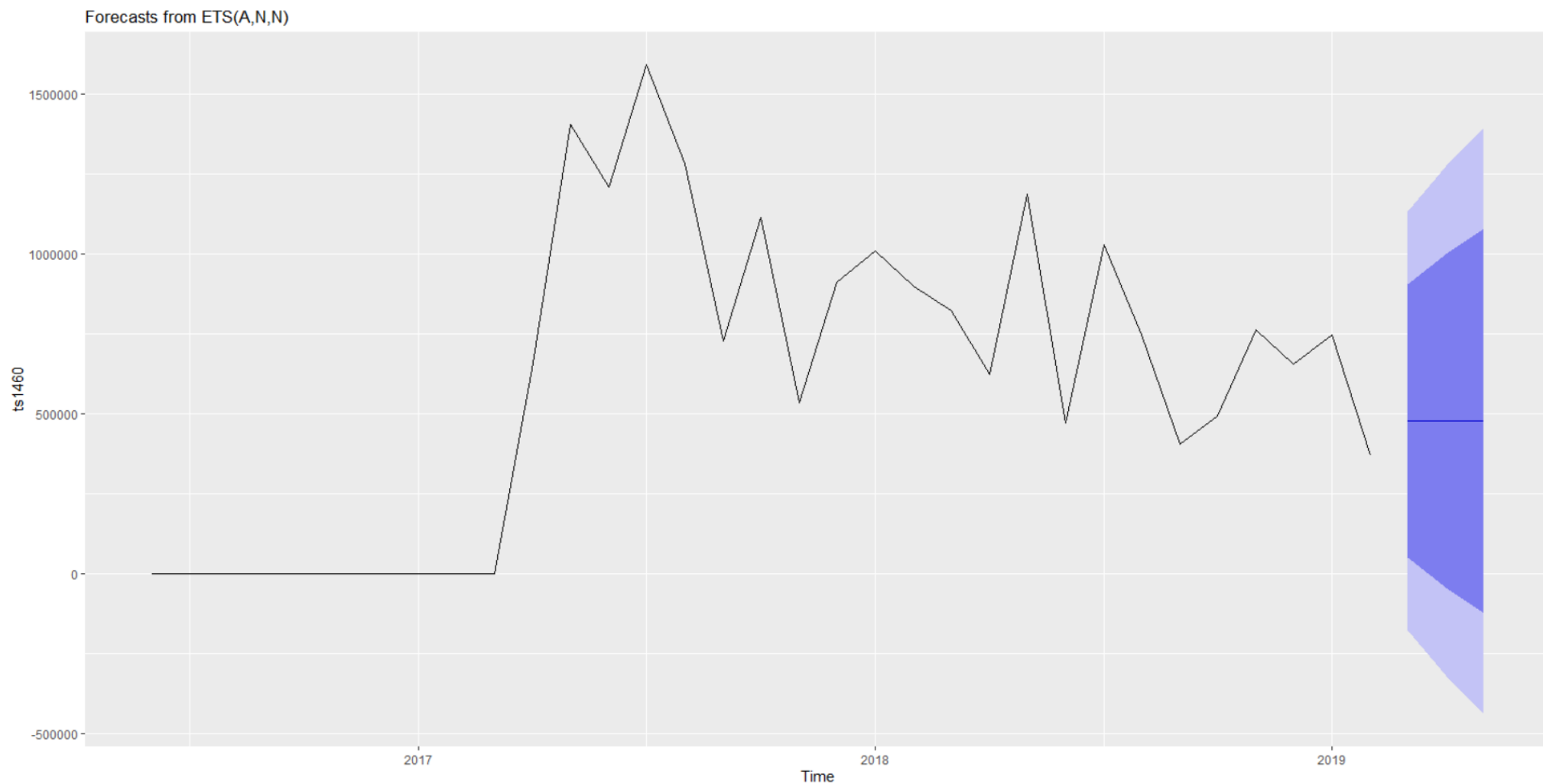
1. ARIMA 모델



3개월 매출 예측

4. 분석 결과

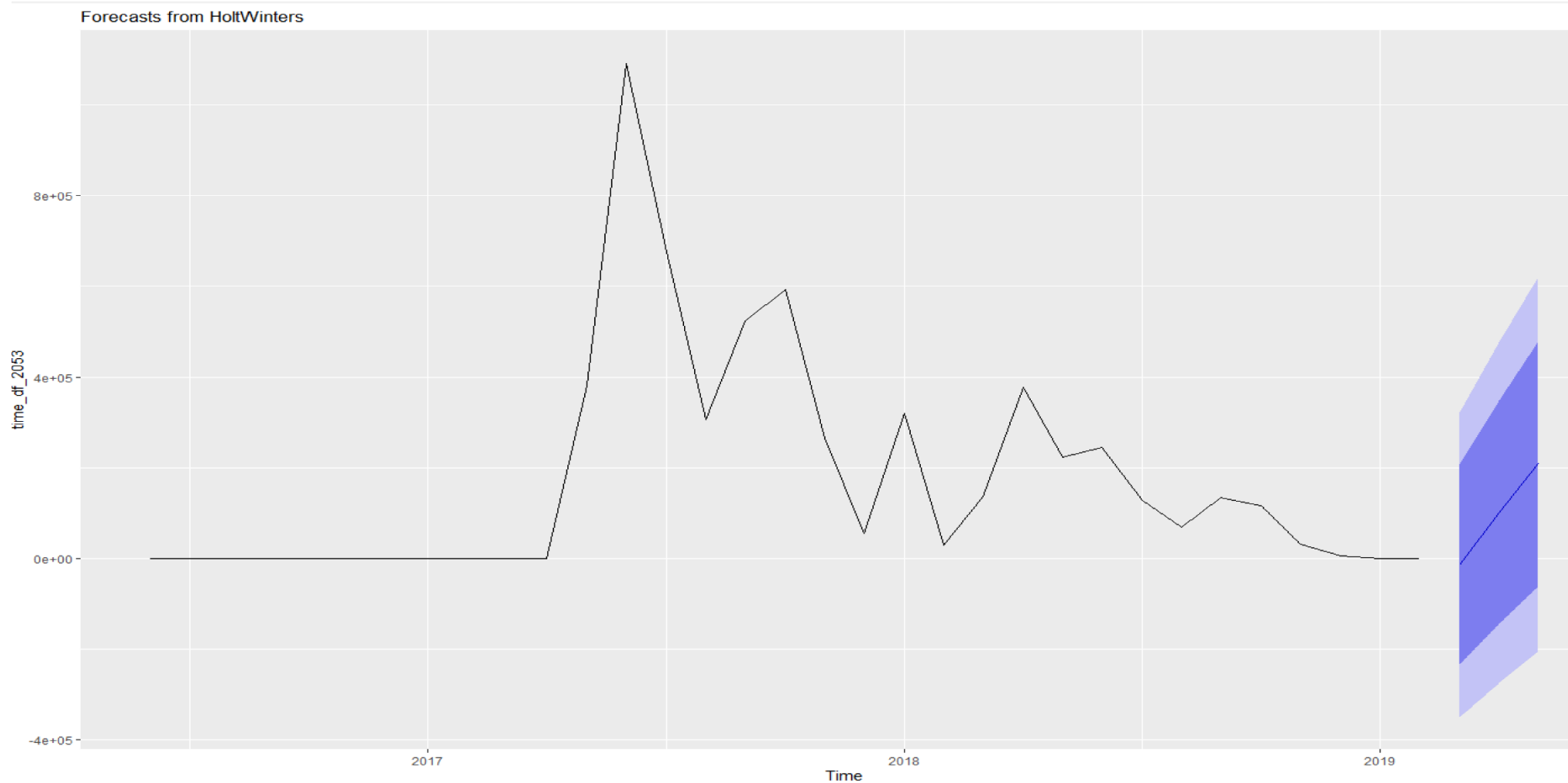
2. ETS 모델



3개월 매출 예측

4. 분석 결과

3. HoltWinters 모델



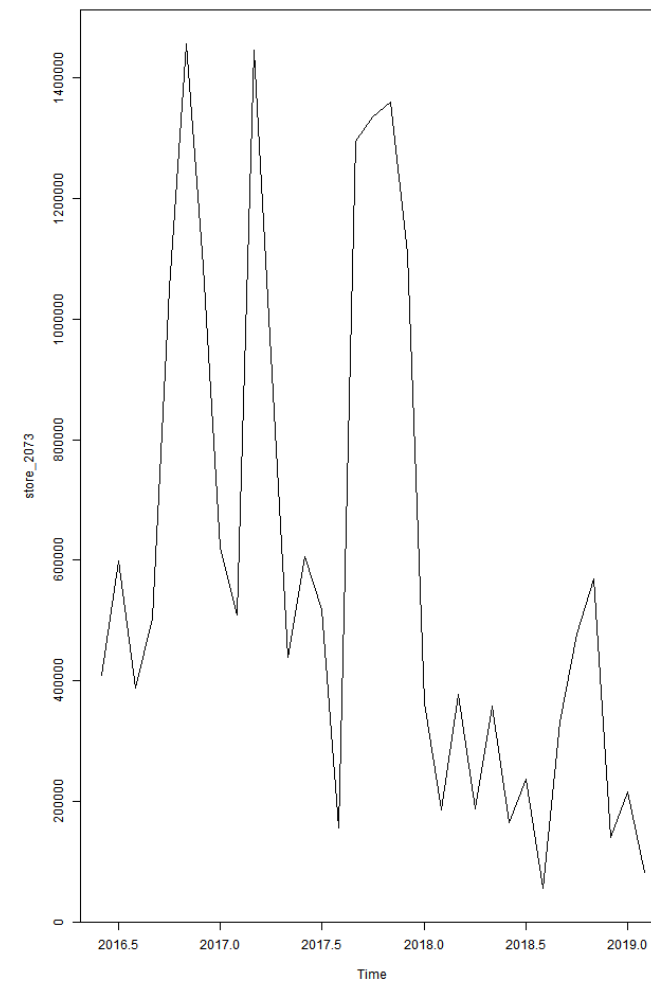
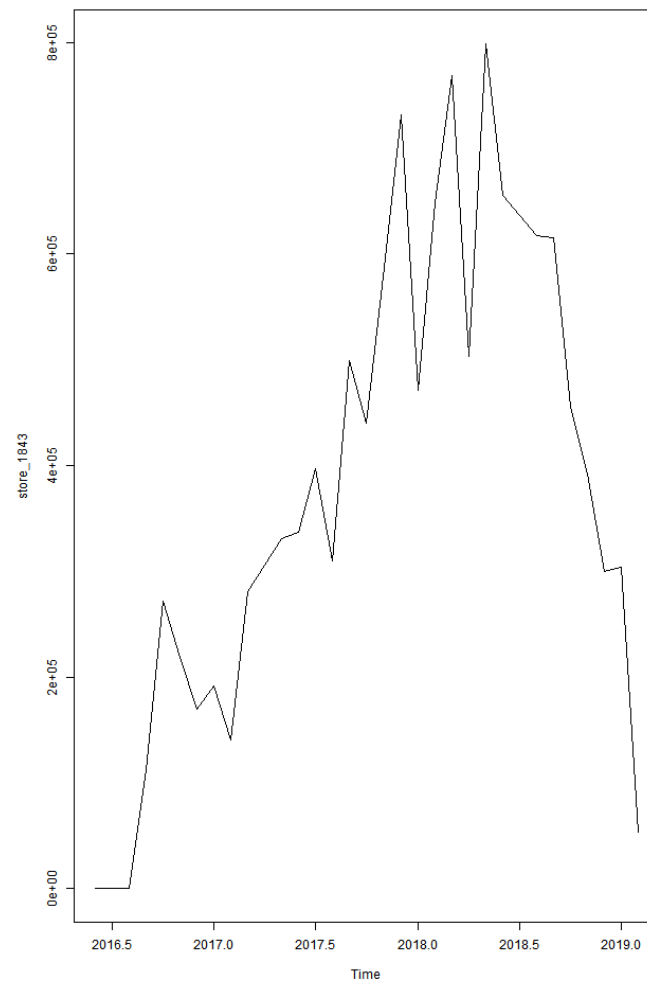
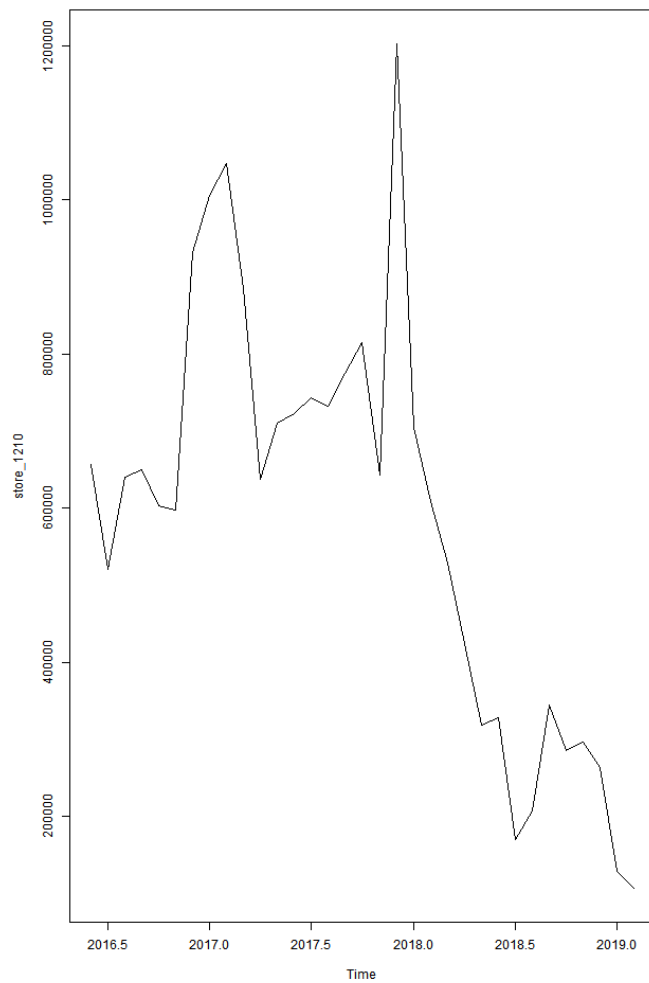
3개월 매출 예측

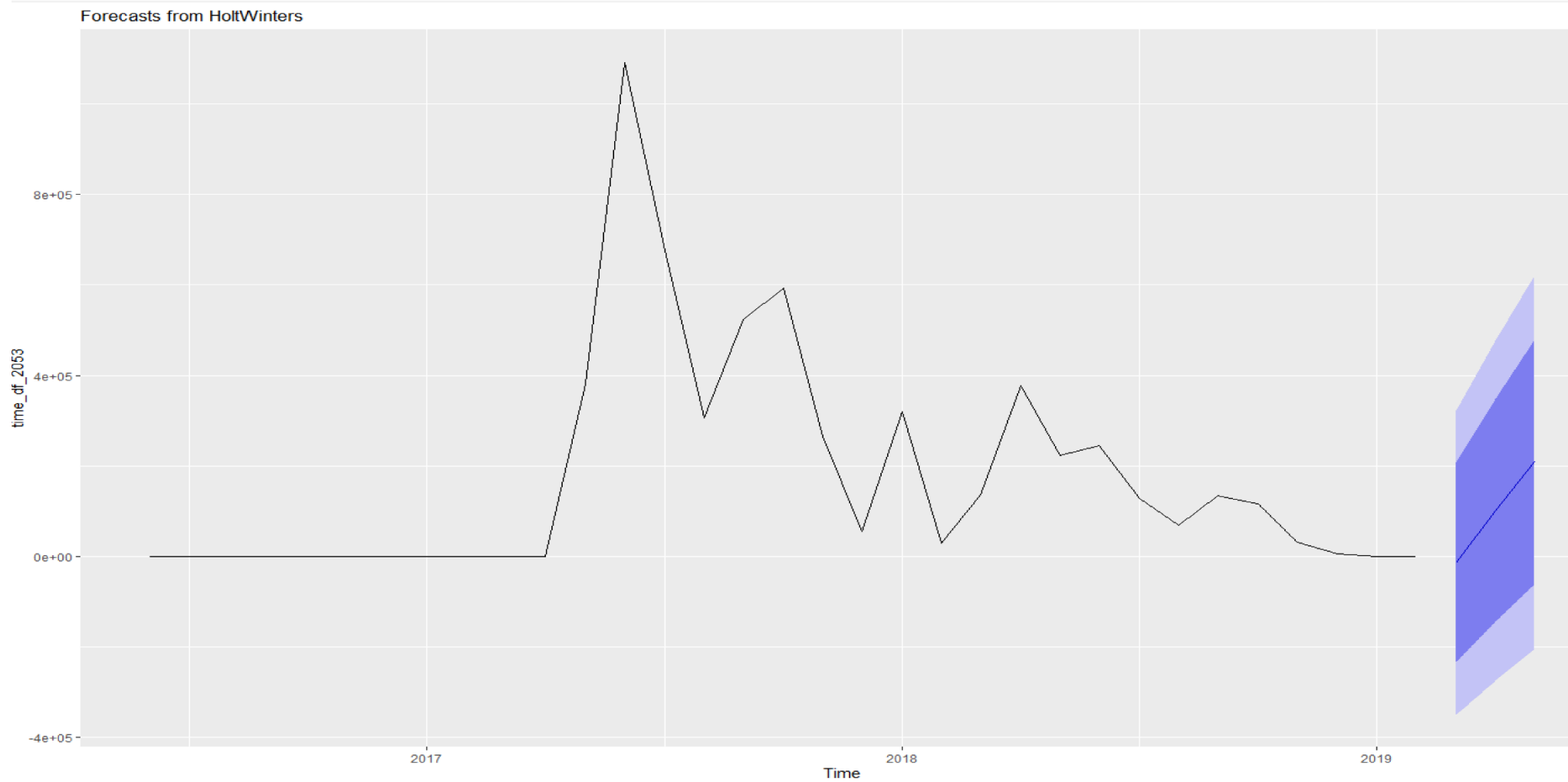
4. 분석 결과

4. 예측 결과

	store_id	predict
1	0	2006844.62
2	1	288621.02
3	2	1276695.13
4	4	2705904.54
5	5	829139.11
6	6	7391571.43
7	7	782086.64
8	8	3894985.71
9	9	1217142.86
10	10	2047847.48
11	11	1735285.71
12	12	1093309.50
13	13	3296871.43
14	14	12315454.00
15	15	2960458.41
16	16	644142.86
17	17	815721.48
18	18	1083085.71
19	19	6359139.43
20	20	5056046.46
21	22	454414.29
22	23	1851718.68
23	24	2864492.86
24	25	427626.66
25	26	1195839.87
26	27	5419393.71
27	28	3620348.25
28	29	2563739.80
29	30	403971.43
30	31	2103318.92

각 스토어별 3개월 매출 예측 총 합





05. Q & A

- **질의응답**



Q & A

The background image shows a modern clothing store. On the left, a rack of various jackets and shirts is visible. In the center, two industrial-style pendant lamps hang from a wooden ceiling. Below them, more clothes are hanging on a rack. To the right, a large mirror reflects the interior, showing two people looking at items. The overall aesthetic is clean and minimalist with a focus on natural materials like wood.

Thank You

21년 4월 R과 파이썬을 활용한 빅데이터 분석 전문가 양성 과정 발표