



건강기능식품 쇼핑몰

기업재현데이터를 활용한 매출분석 및 마케팅 전략 제안





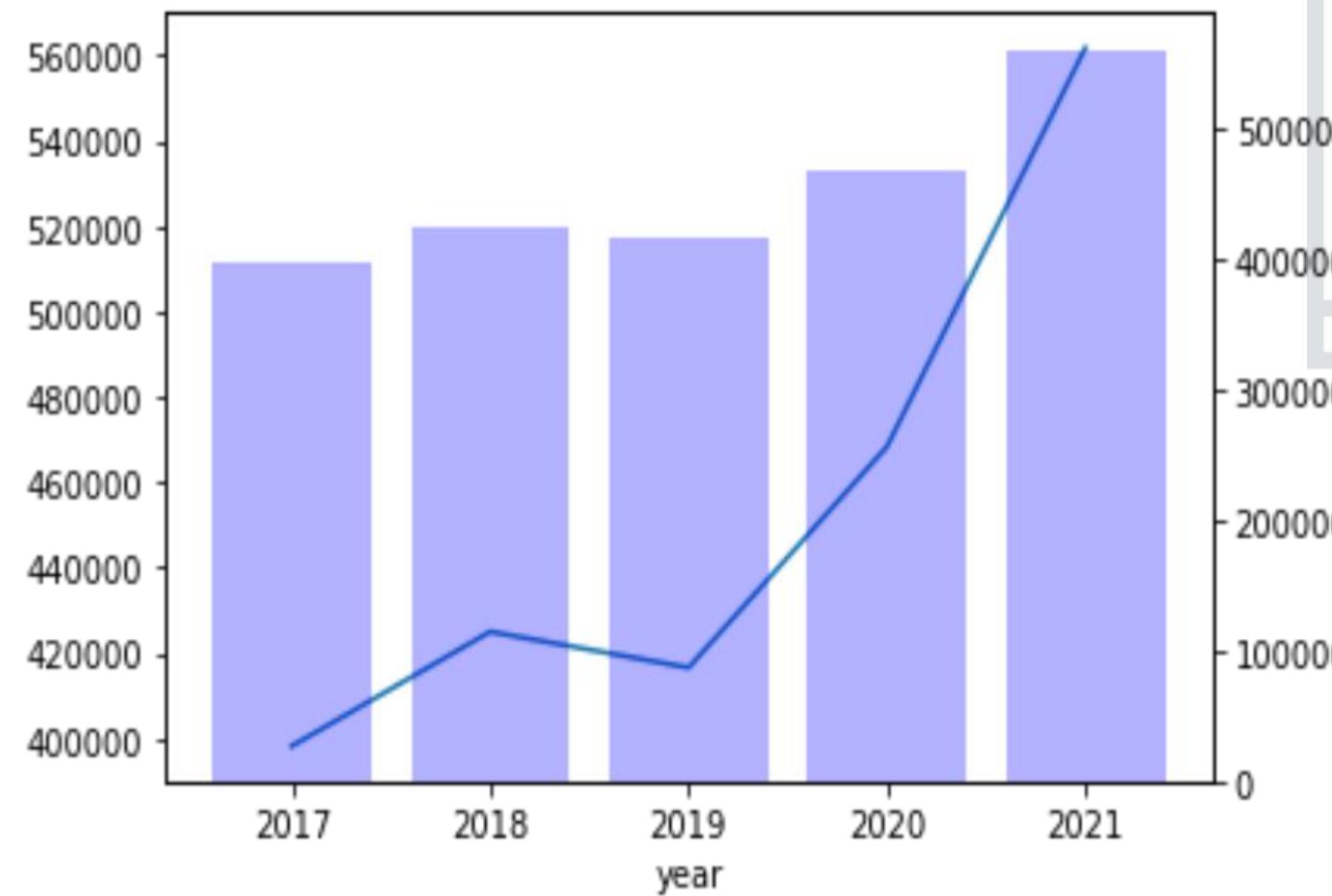
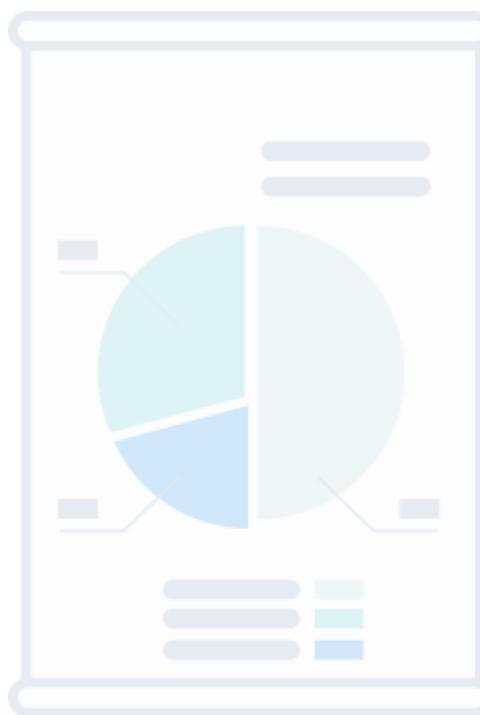
- 01 분석 배경 및 목표
- 02 데이터 EDA
- 03 데이터 처리 및 활용
- 03 결론 및 시사점 / 한계



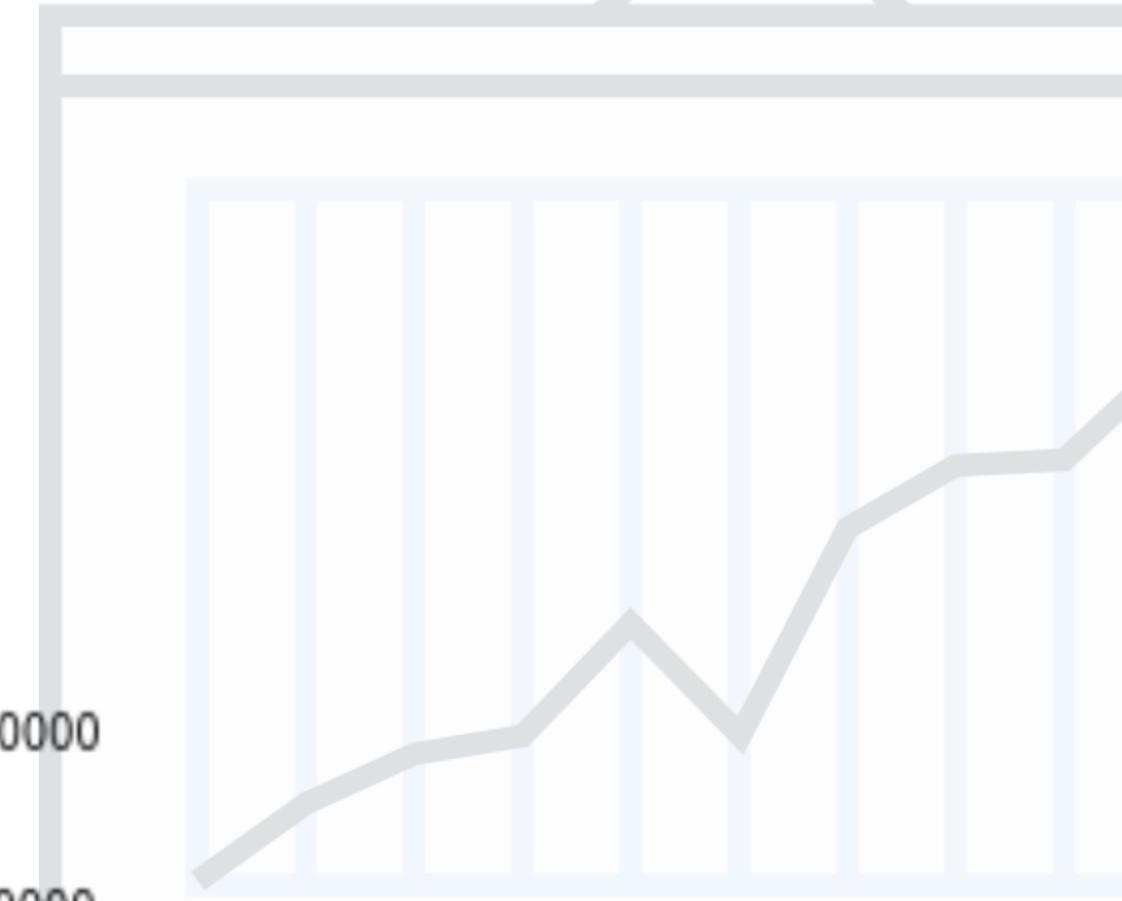


01 분석 배경

쇼핑몰 재현 데이터 판매 추이 시각화



2020년~2021년 기점 매출 급증

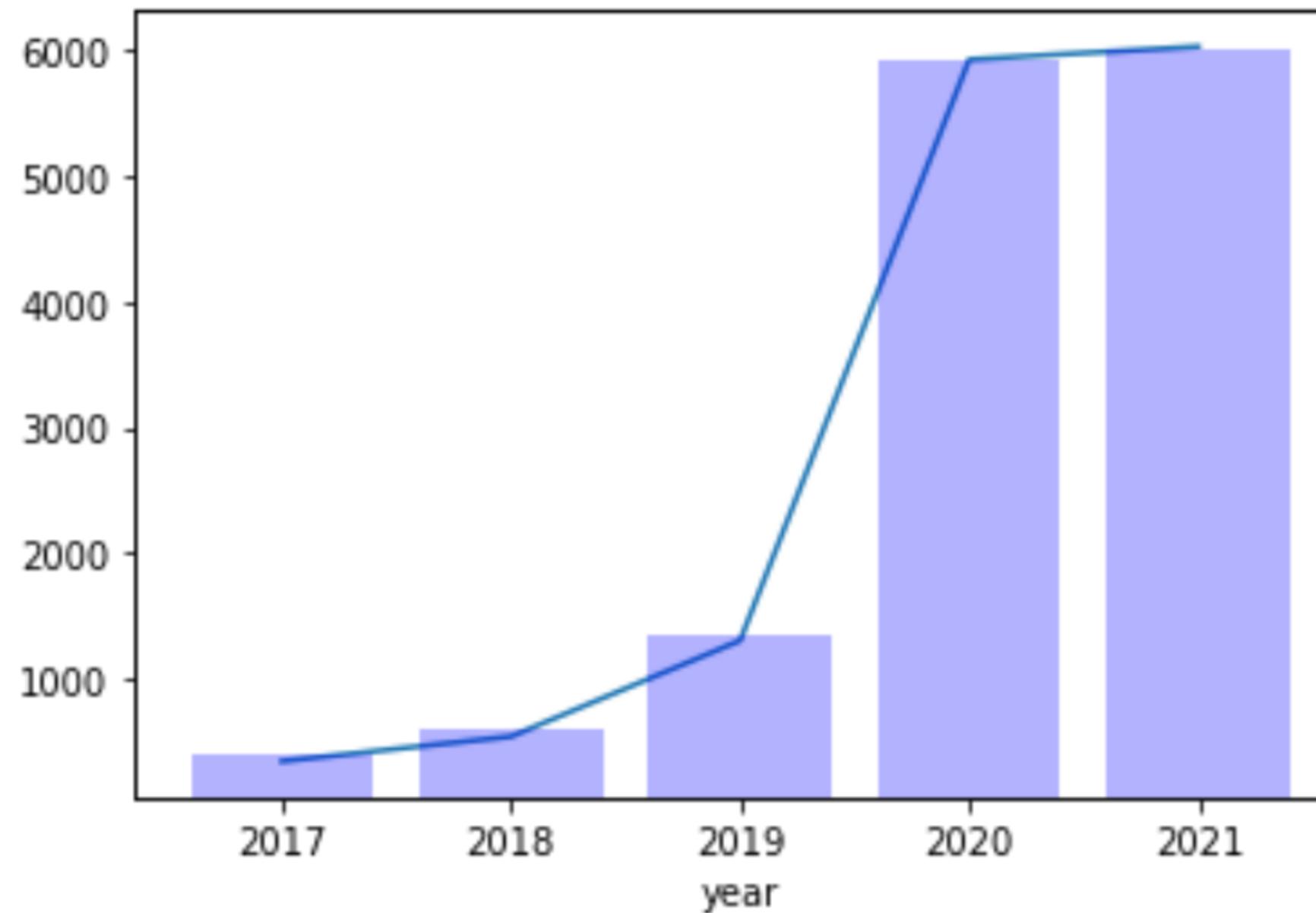




01 분석 배경

'홈트' 키워드 기사 제목 크롤링 (네이버, 약 15,000건)

(1) 연도별 '홈트' 키워드 기사 건수 변화



2020년부터 언급량 급상승

Selenium과 Webdriver를 이용해
제목 혹은 본문에 특정 키워드를 포함된 기사 수집

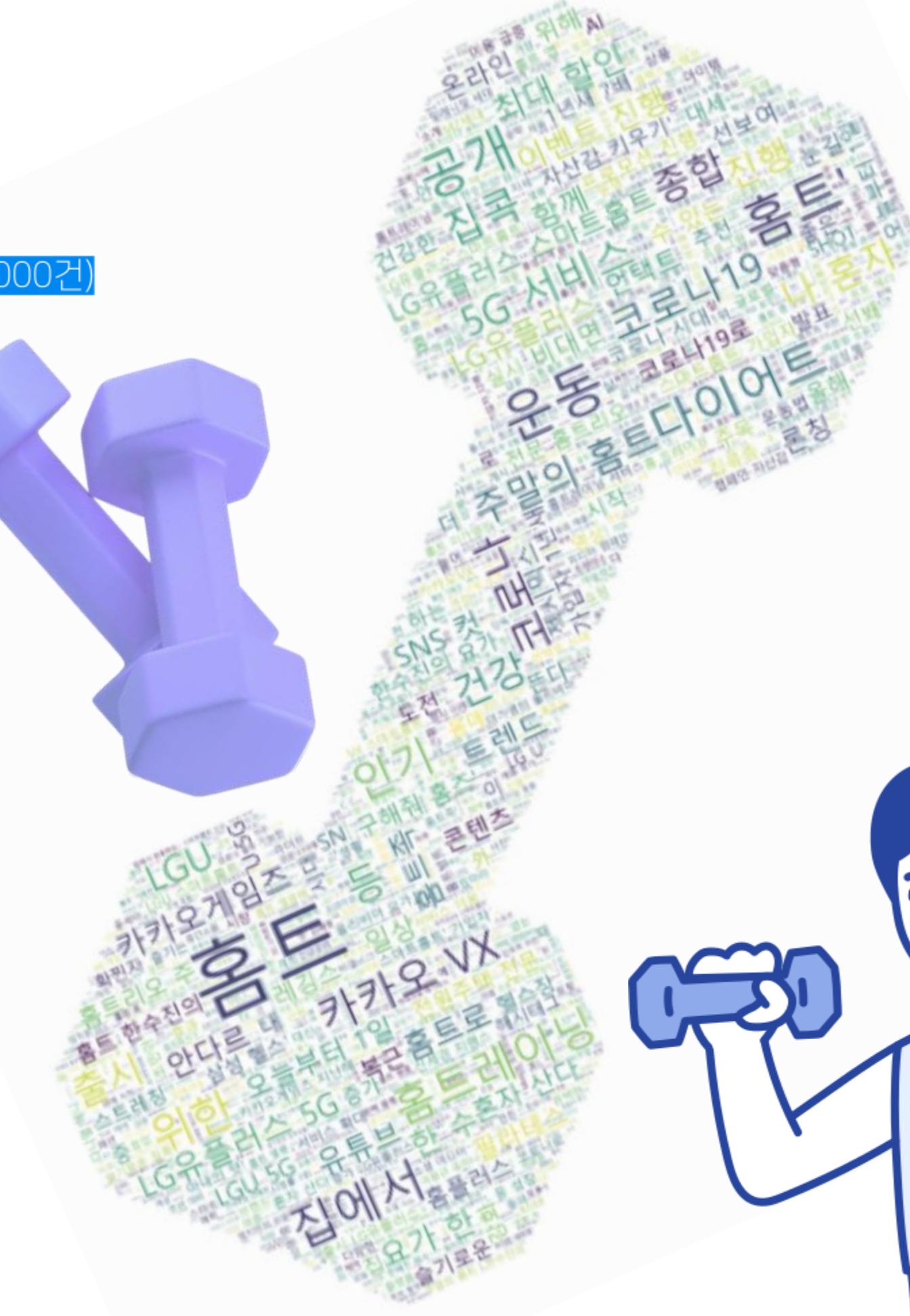




01 분석 배경

'홈트' 키워드 기사 제목 크롤링 (네이버, 약 15,000건)

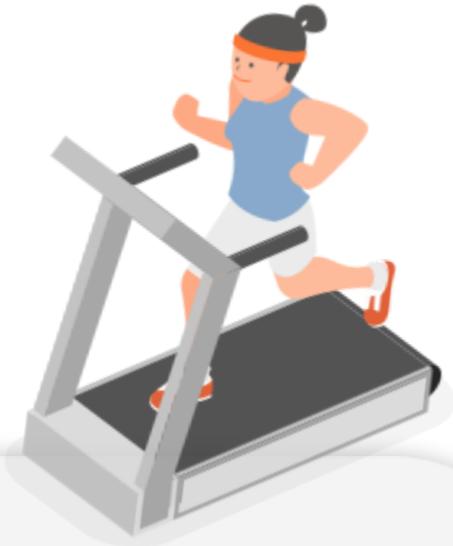
(1) 기사 제목 내 최빈 단어군





01 분석 방향 및 목표

쇼핑몰 매출분석 및 마케팅 전략 제안



주제 1

매출 예측에 유의미한
영향을 미치는 변수 파악



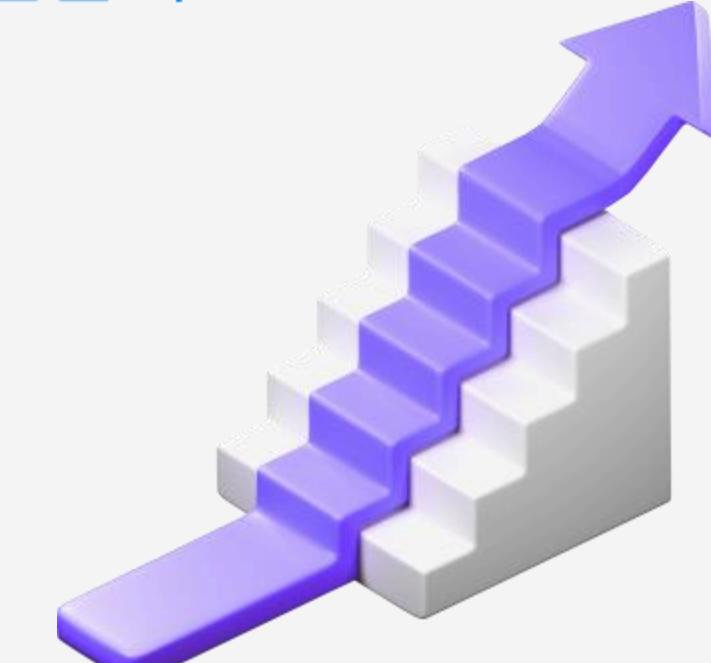
주제 2

시계열 분석을 통한
판매량 예측 모델링



주제 3

연관분석





02 데이터 EDA

(1) 데이터 탐색 및 수집

공공데이터 수집 후 쇼핑몰 재현데이터와 결합

환율 (한국은행, 2017-2021) · 전체소비자 물가(한국은행, 2017-2021) · 건강기능식품 영업 정보(행정안전부, 2017-2021) · 체력단련장(헬스장) 영업 정보(행정안전부, 2017-2021) · 건강기능식품 산업 현황(식품의약품안전청, 2017-2021) · 코로나 거리두기 단계(질병관리청, 직접 수집, 2020-2021) · 네이버뉴스 홈트 관련 기사 (직접 수집, 2017-2021)

데이터셋 구성

구매시간	year	month	카테고리	상품명	상품금액	총상품금액	소비자물가	원/달러(평균)	건기식매출액	건기식영업등록	건강기능식품폐업	온라인쇼핑몰매출규모	상품갯수
2017-01-01	2017	1	부스터	컴뱃 10 0% 웨이 6 7서빙	49	98	97.4	1208	2.24E	0	1	7310479	2

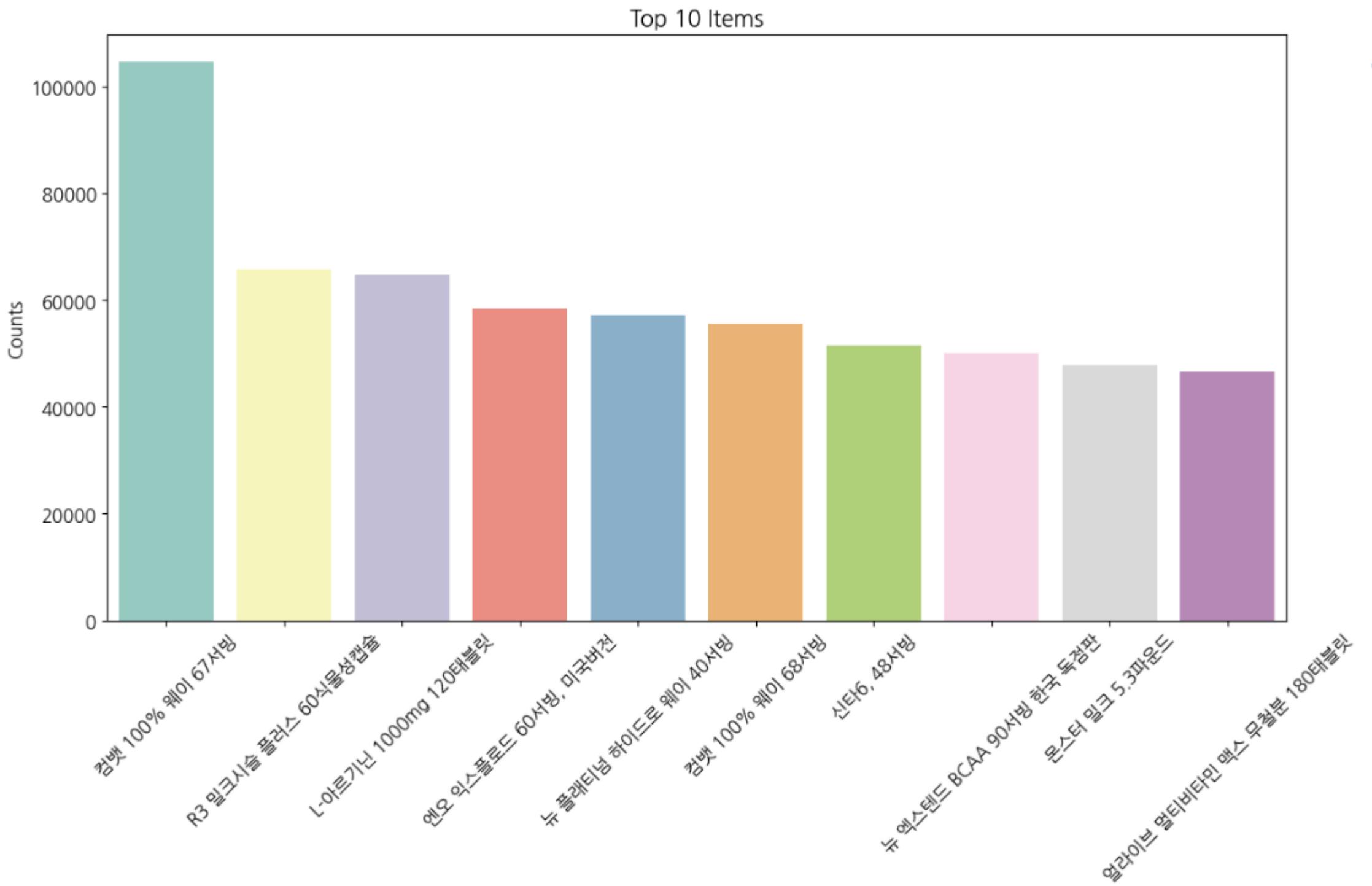
2045388 rows × 14 columns





02 데이터 EDA

(1) 데이터 시각화



판매량 상위 10개

단백질 보충제

간 보호제

피로 회복제

종합비타민

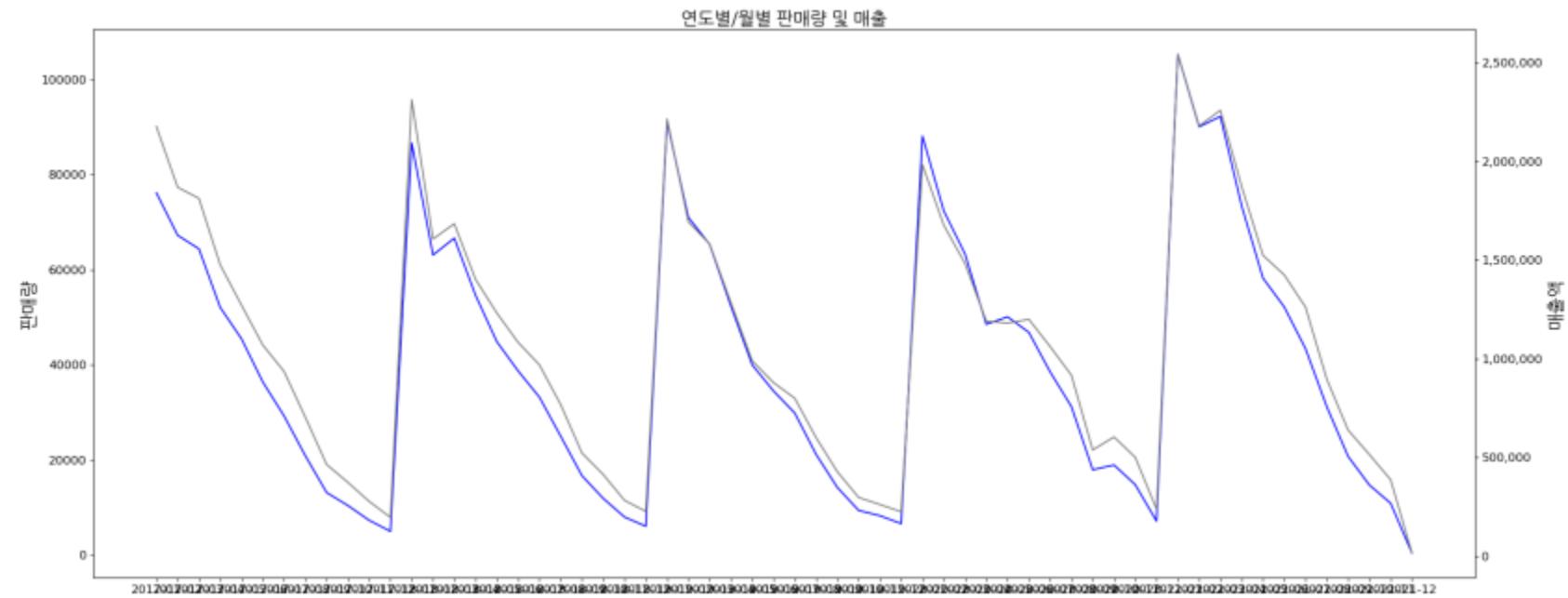
...



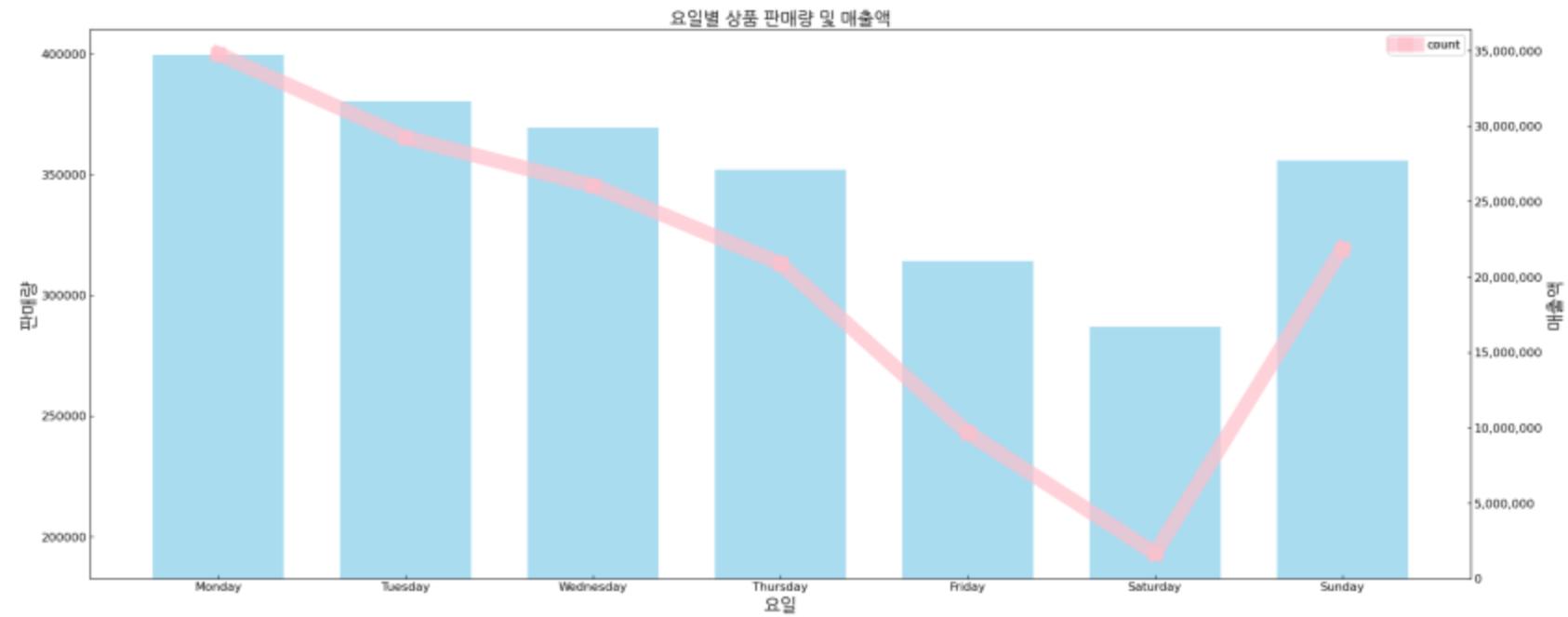


02 데이터 EDA

(1) 데이터 시각화



연도-월별 판매량 및 매출액
상반기 매출 급증하는 계절성 확인



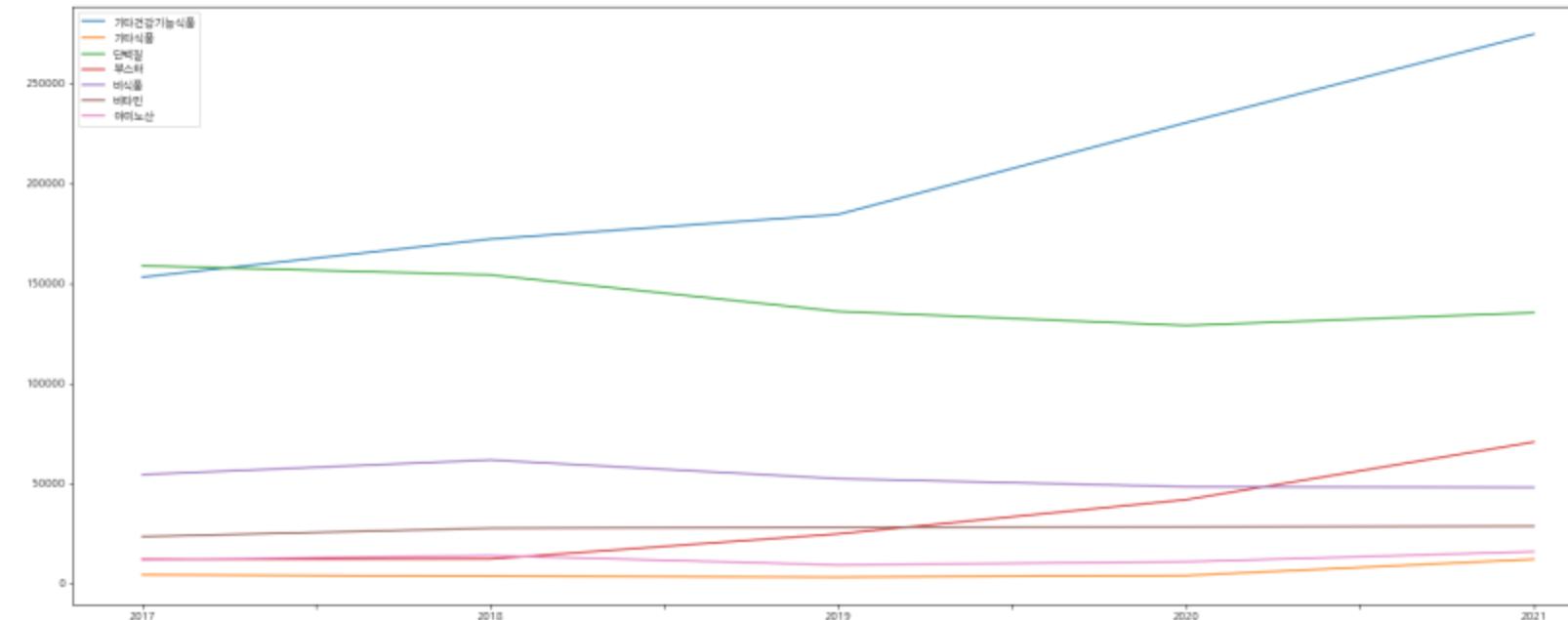
요일별 상품 판매량 및 매출액
주초에 주문량 가장 많음





02 데이터 EDA

(1) 데이터 시각화

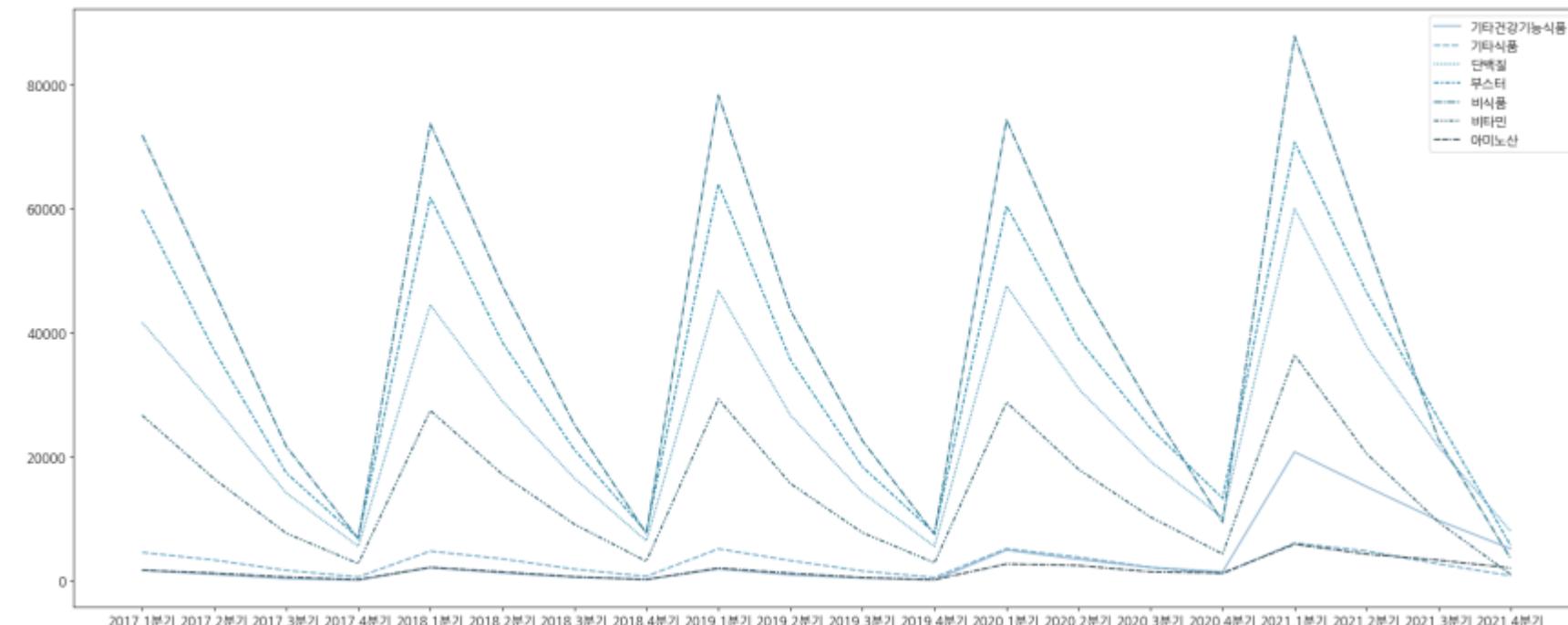


연도별 카테고리별 상품판매량

전체적으로 판매량이 늘어나는 추세를 보였으나 단백질의 경우 수요가 약간 감소하고 부스터가 크게 늘어나는 추세를 보임

분기별 카테고리별 상품판매량

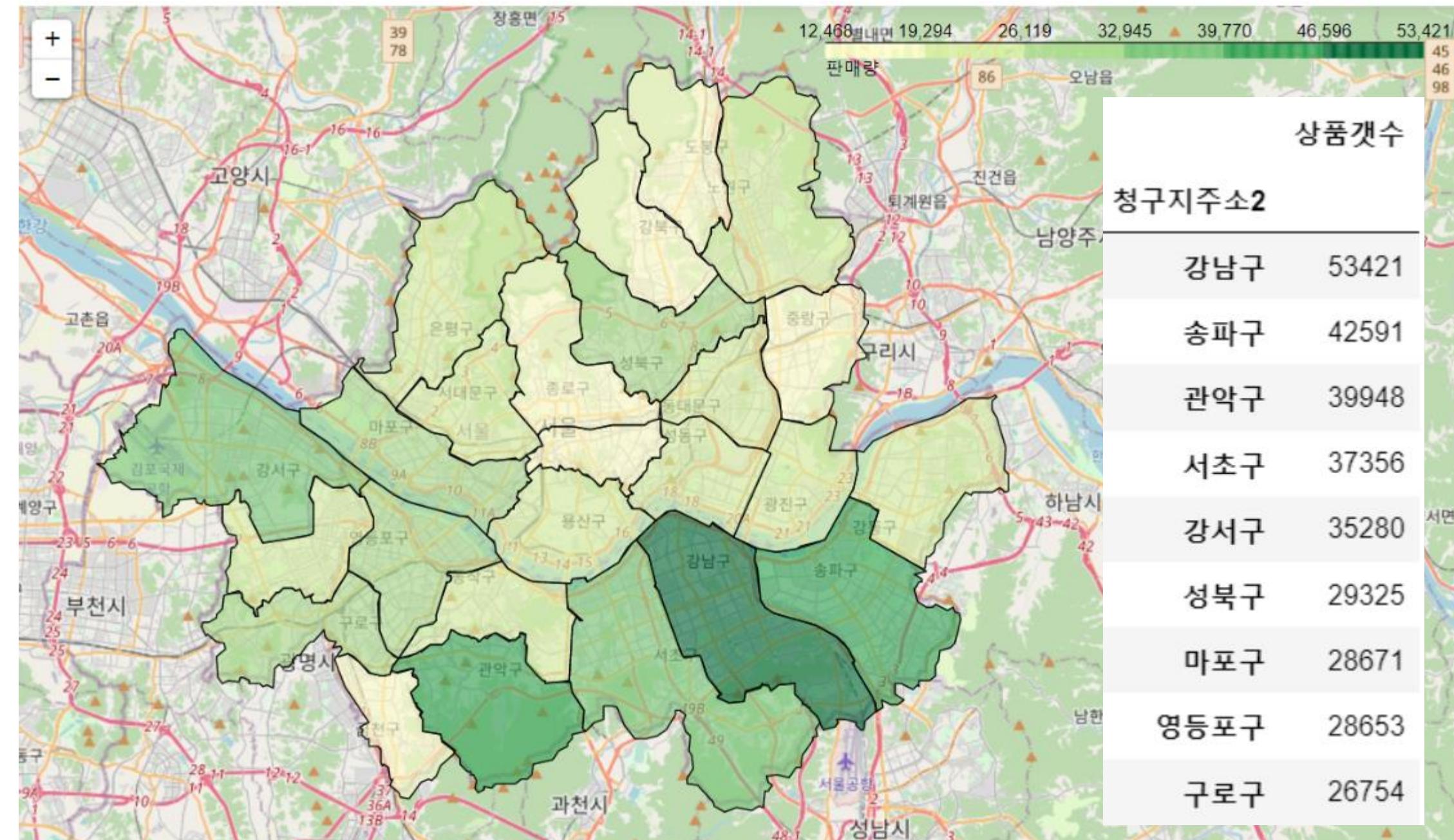
1분기에 판매량이 크게 증가 후 감소하는 추세





02 데이터 EDA

(1) 데이터 시각화



서울 지역 주문량

강남구 주문량 가장 높음
이후 송파구 관악구 강서구 순





02 데이터 EDA

(2) 일자별 매출 예측에 유의미한 변수 파악

Random Forest 활용 Feature Importance 추출

(1) 시간 패턴

- 1년 기준 일자(1일 ~ 365일)
- 1년 기준 주차(1주 ~ 52주)
- 월 기준 일자(1일 ~ 31일)
- 연도 (2017년 ~ 2021년)
- 분기 (1분기 ~ 4분기)
- 월 (1월 ~ 12월)
- 요일 (월요일 ~ 일요일)

```
Train Set 정확도: 0.966122318484883  
Test Set 정확도: 0.912670510022873  
-----설명변수 중요도-----  
          Feature Importance  
0      상품갯수      0.418855  
7      weekofyear    0.270015  
5      dayofyear     0.243858
```

(t+1)일 시점 매출 영향 요인
1년 기준 일자, 1년 기준 주차



소비자 구매 패턴에 시계열적 규칙성이 존재





02 데이터 EDA

(2) 일자별 매출 예측에 유의미한 변수 파악

Random Forest 활용 Feature Importance 추출

(2) 외부 환경 변수

- 건강기능식품 시장 규모
- 건강기능식품 유통업 영업신청/폐업신청
- 온라인쇼핑몰 시장 규모
- 환율
- 소비자물가
- 지하철 수송 인원
- 체력단련업(헬스장) 영업신청/폐업신청
- 코로나19 거리두기 단계

2018~2019년도 매출 영향 요인

건강기능식품 시장 규모

2020년도 매출 영향 요인

건강기능식품 시장 규모, 환율

2021년도 매출 영향 요인

건강기능식품 시장 규모, 환율



연도에 따라 매출에 영향을 미치는 요인 변화

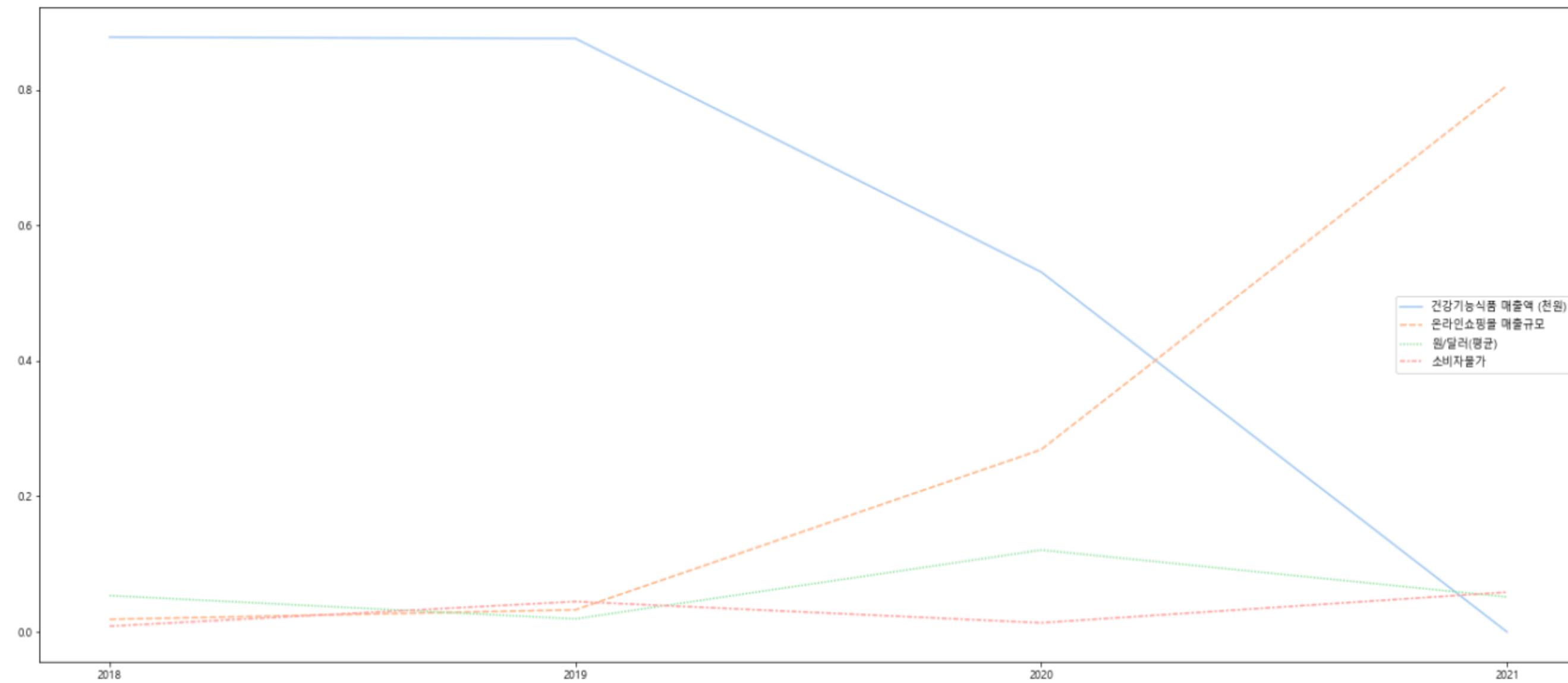




02 데이터 EDA

(2) 매출 예측에 유의미한 영향을 미치는 변수 파악

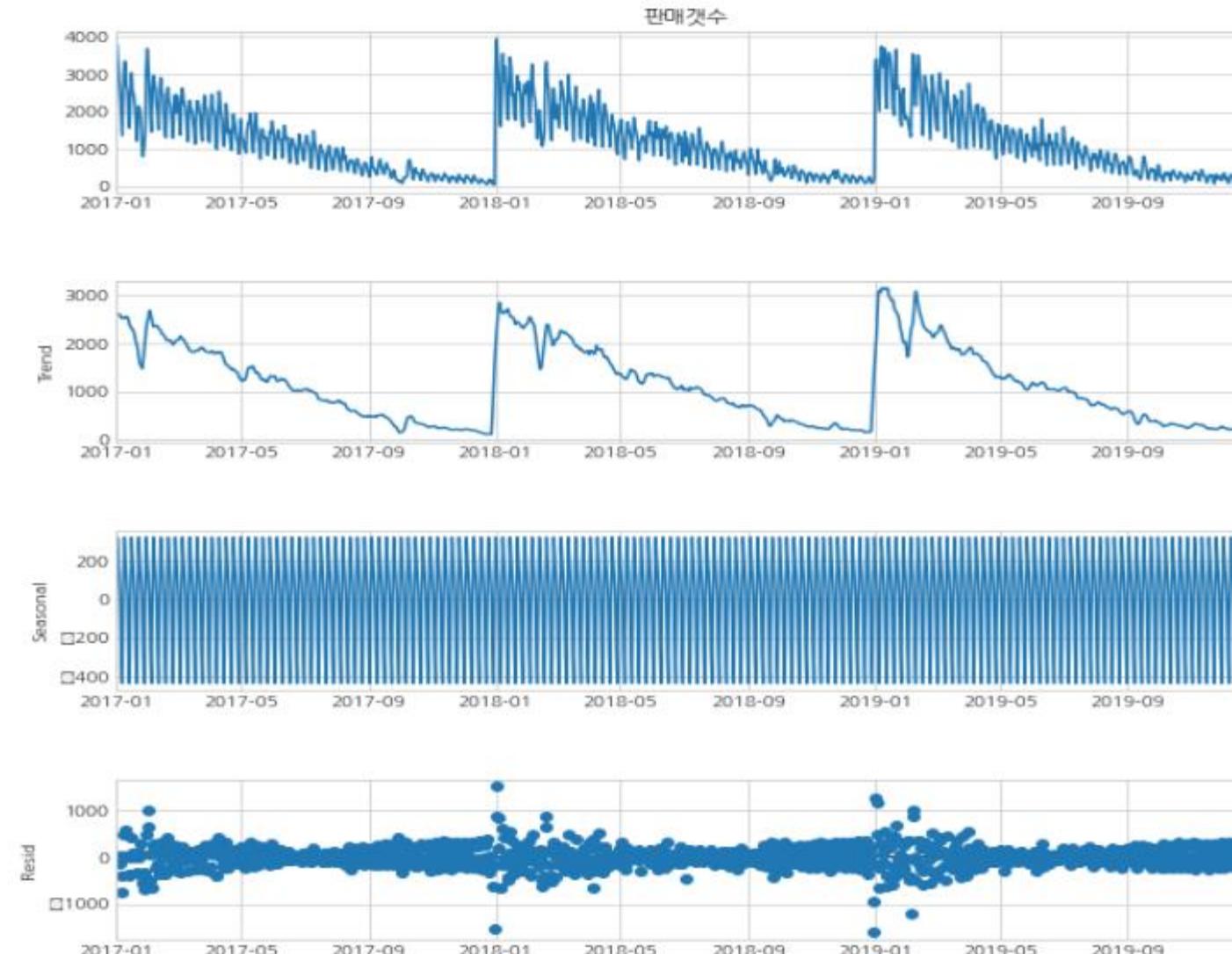
건강기능식품 시장규모 영향력 하락 온라인쇼핑몰 시장규모 영향력 **상승**
영양제 시장 성장에 의존하는 비즈니스 전략 탈피 필요



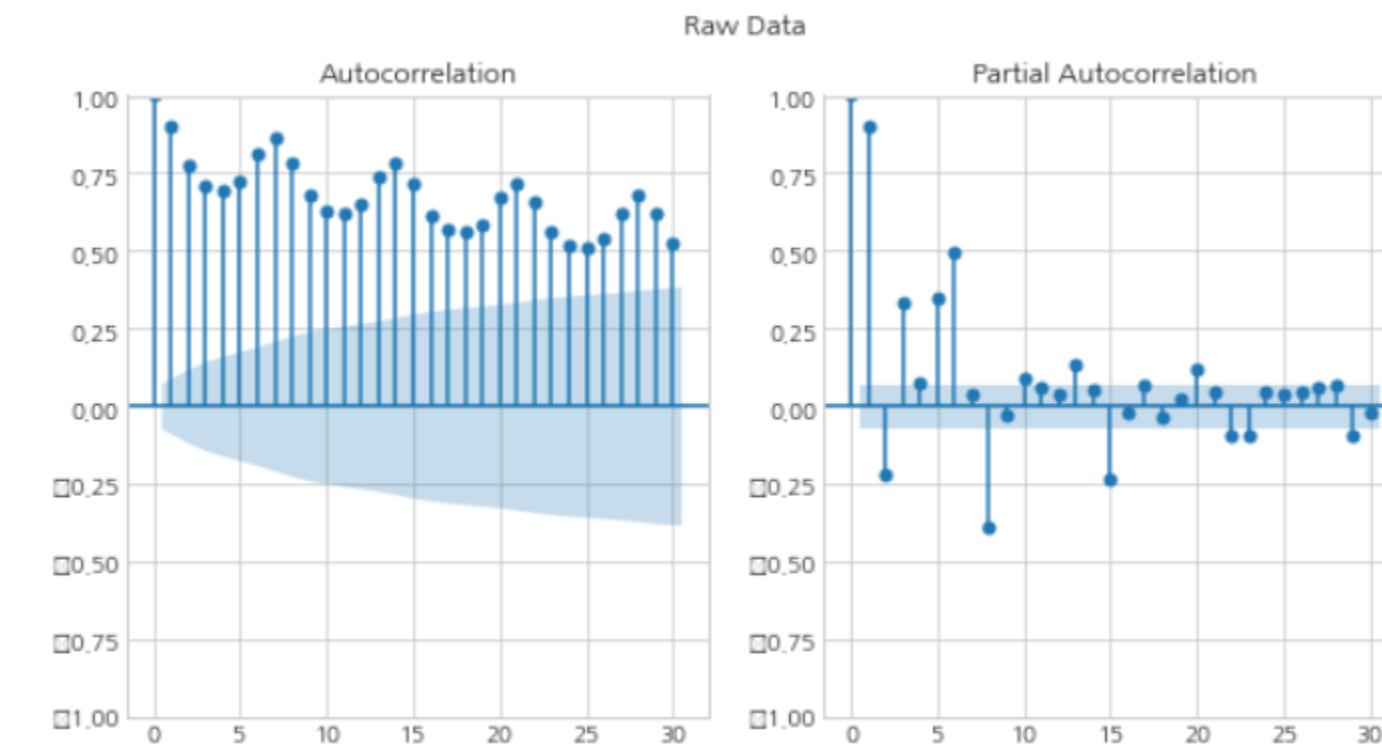


03 데이터 분석

(1) Auto-ARIMA를 활용한 쇼핑몰 매출 시계열 분석



계절성 및 추세 확인



Durbin-Watson Test

p는 1.34로 0보다 크므로 이를 통해 데이터는 자기상관을 가지고 있다고 할 수 있습니다.

ADF테스트 검정결과

p-value는 0.417으로 따라서 위 시계열은 정상성을 띠고 있지 않다고 볼 수 있습니다.



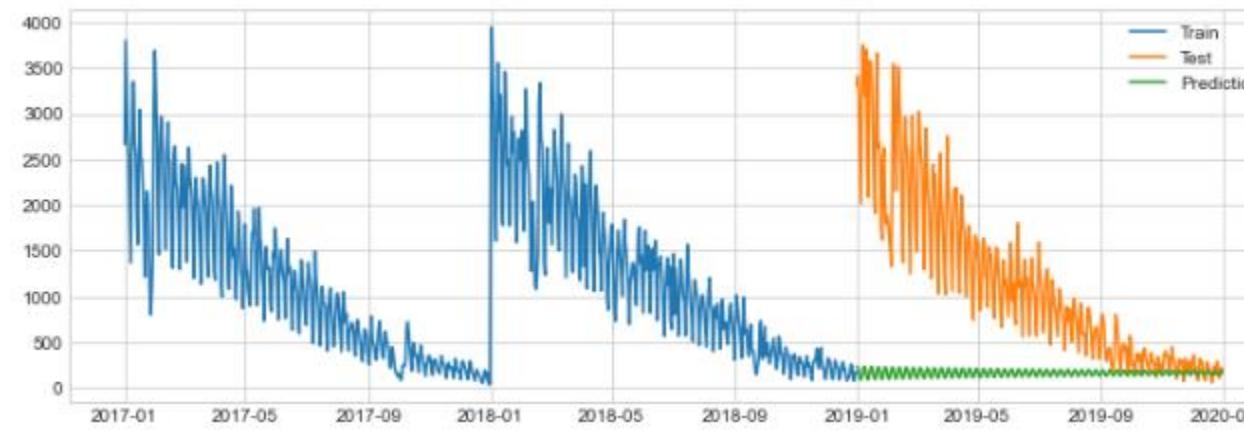


03 데이터 분석

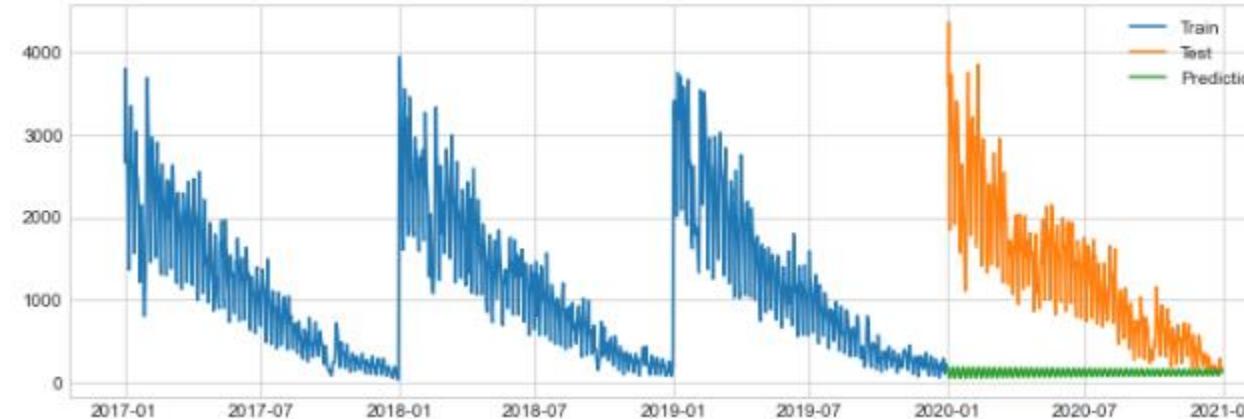
(1) Auto-ARIMA를 활용한 쇼핑몰 매출 시계열 분석

17,18년도 데이터로
19년도 예측

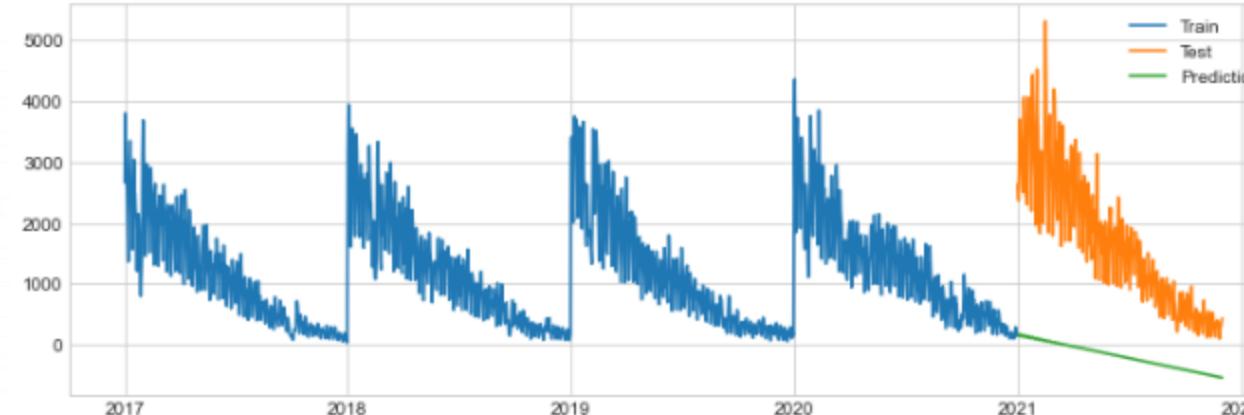
임의보형모델 적용 전



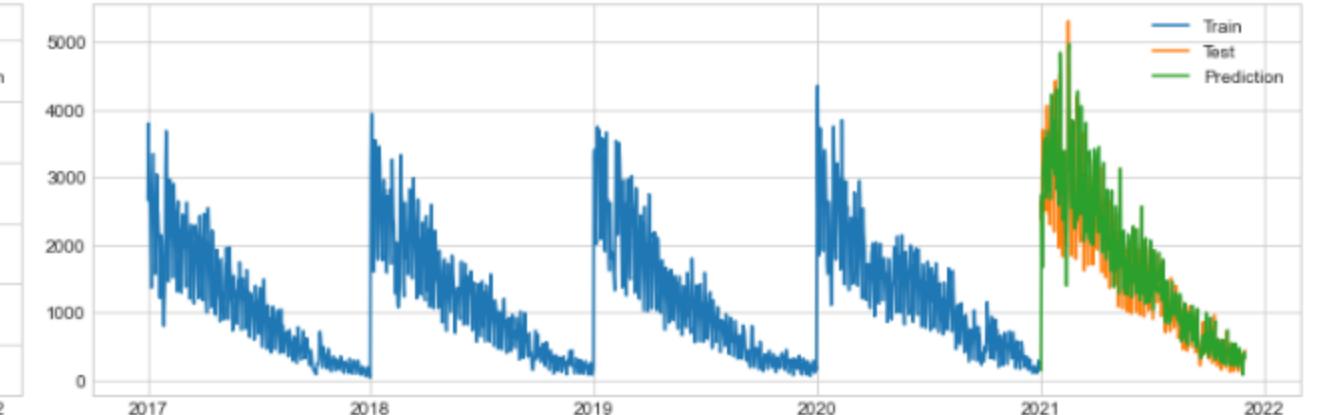
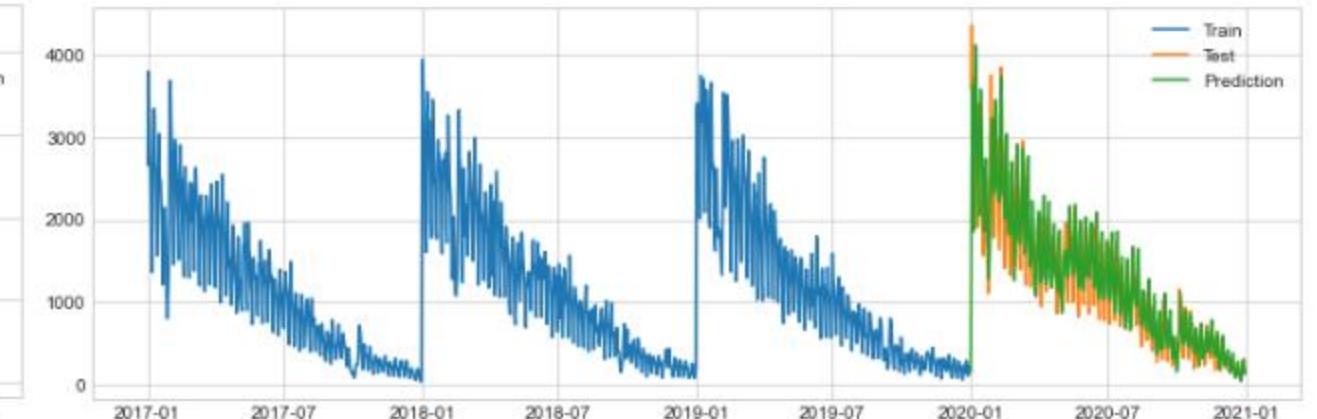
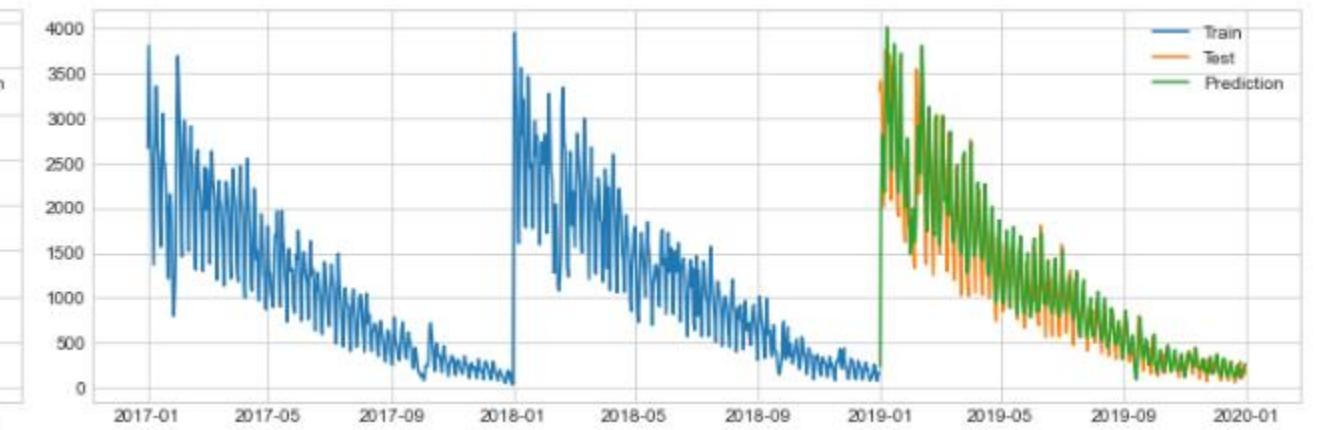
17,18,19년도 데이터로
20년도 예측



17-20년도 데이터로
21년도 예측



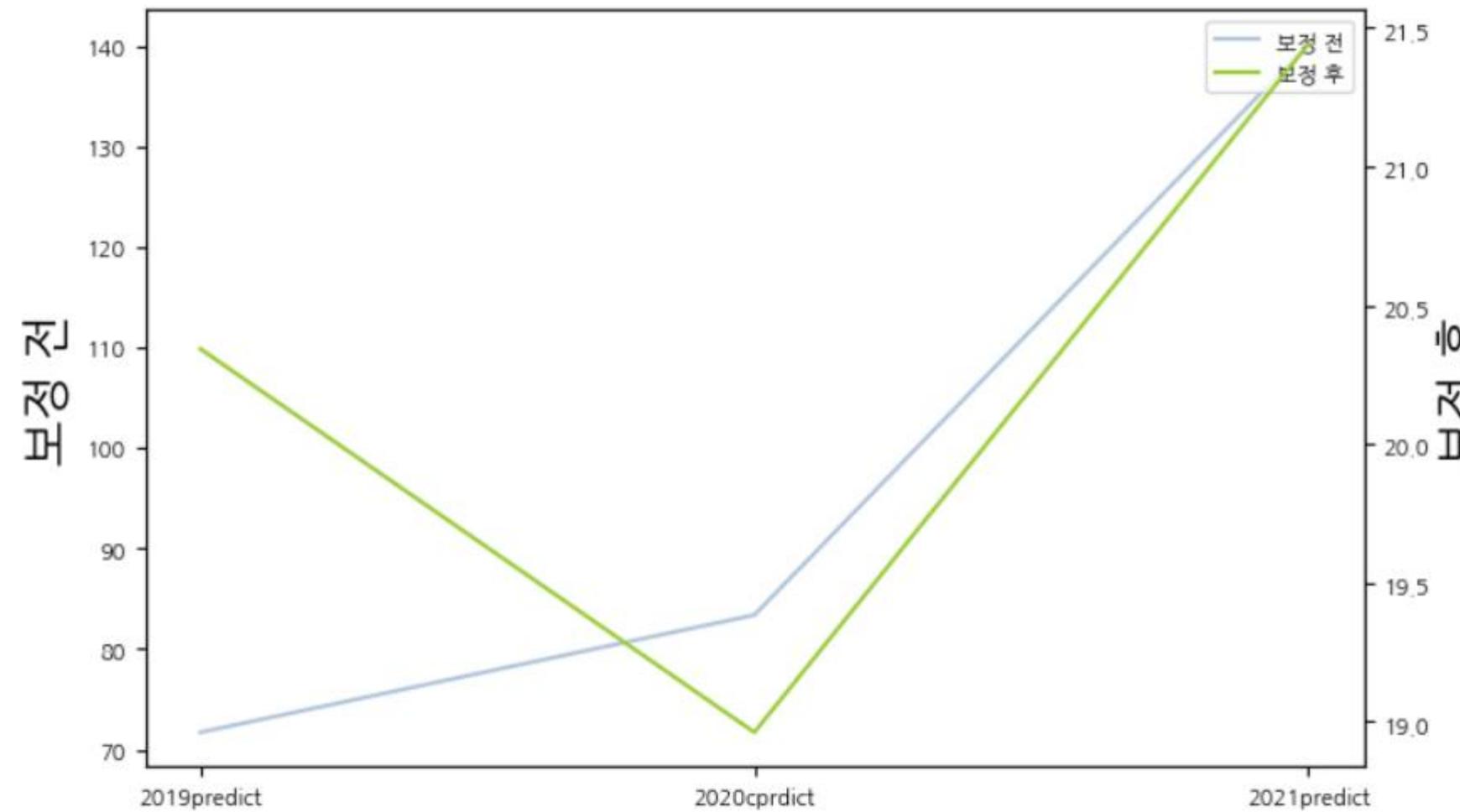
임의보형모델 적용 후





03 데이터 분석

(1) auto-arima를 활용한 쇼핑몰 매출 시계열 분석



2018년, 2019년 매출량을 학습시켜 2020년 매출량을 예측했을 때는 더 좋은 성능을 보였지만,
2021에서는 오히려 성능이 더 떨어졌습니다.

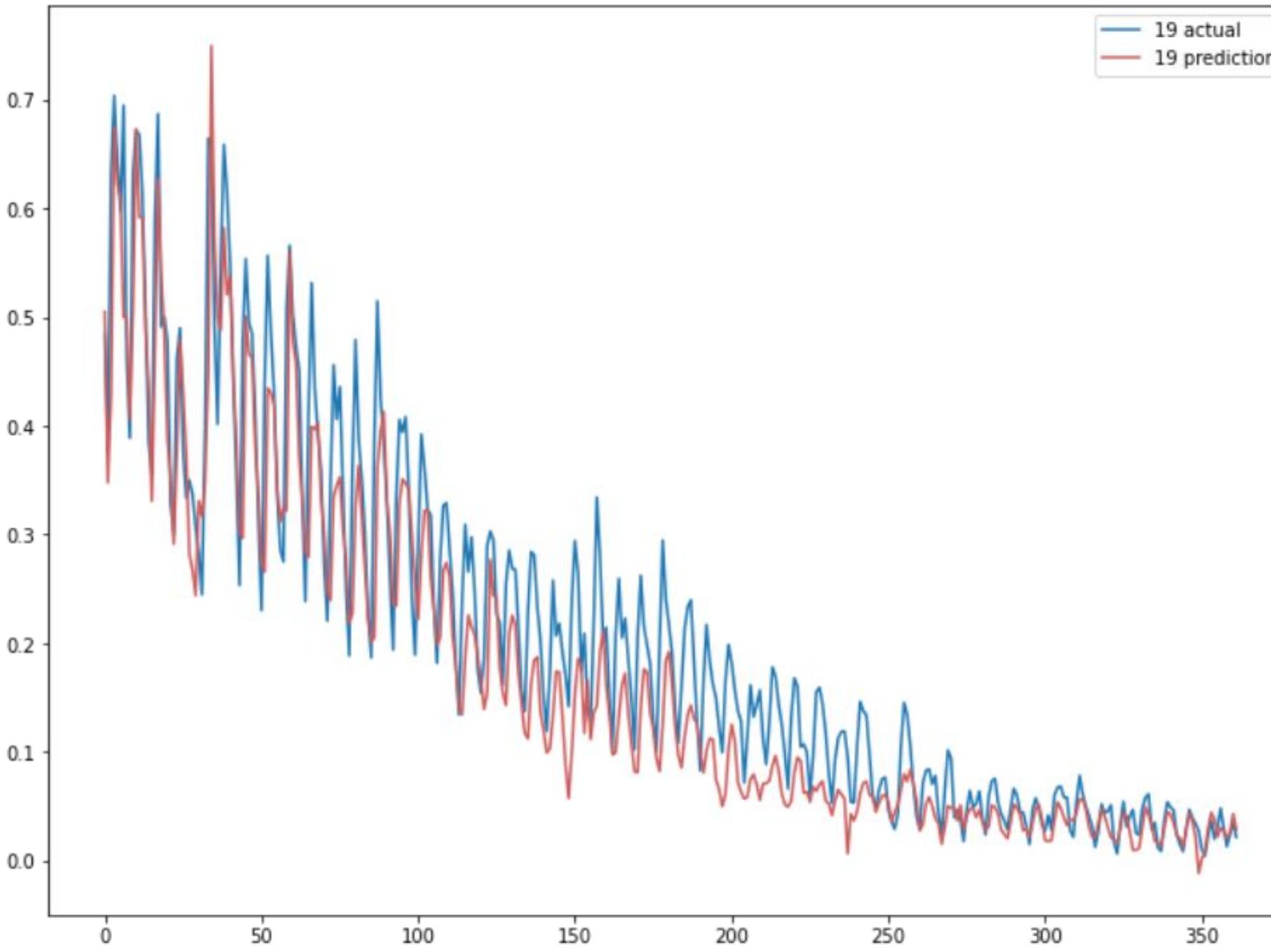
이는 2021년에 매출이 크게 증가했기 때문이라고 해석할 수 있습니다.





03 데이터 분석

(1) 순환신경망 (LSTM)을 활용한 매출량 예측



LSTM Model



Train set : 2017, 2018 Data

Validation set : 2019 Data

평균절댓값백분율오차(MAPE): 27.137

손실 함수(MSE) : 0.00372749

2019년 실제 매출량

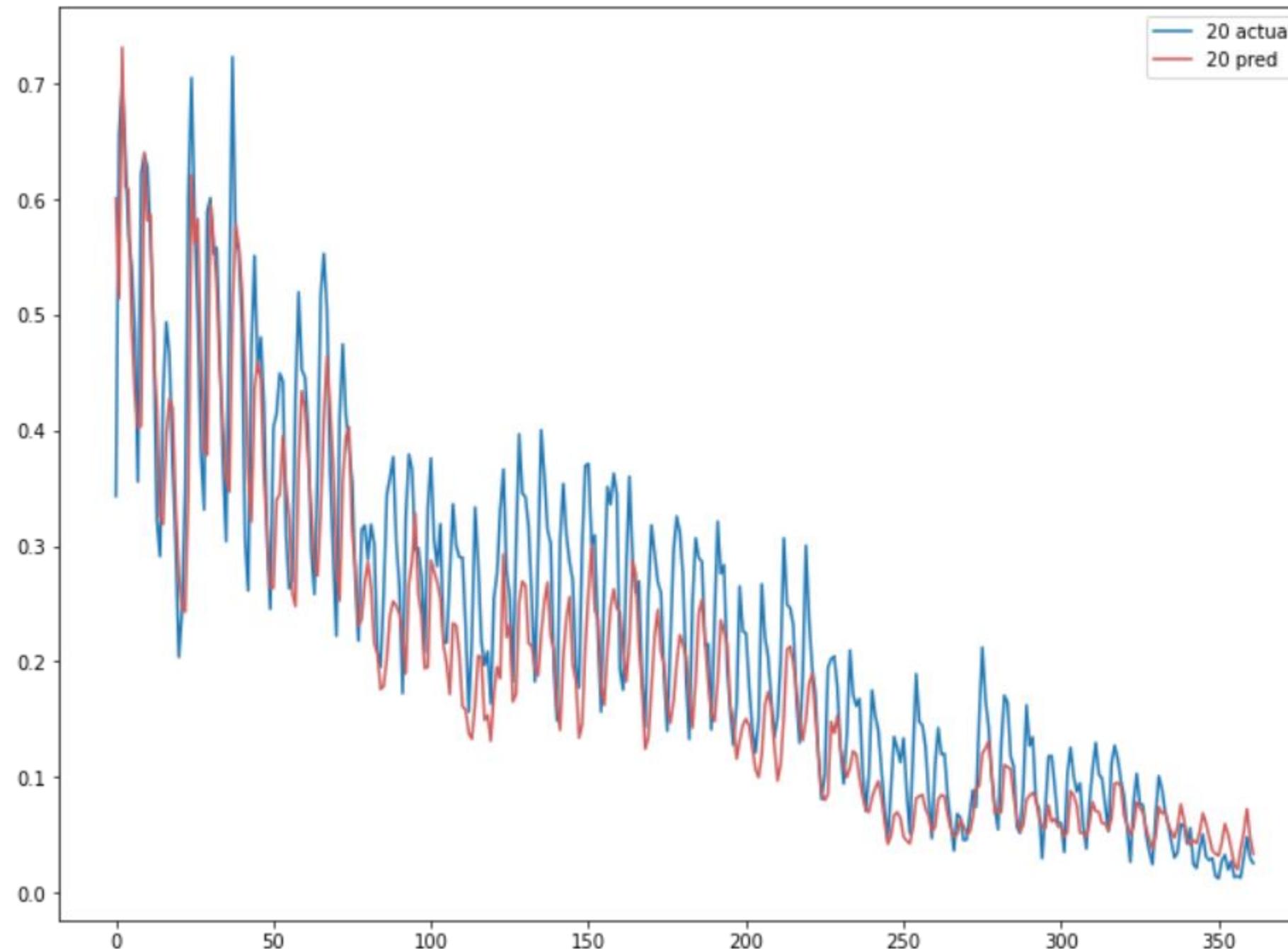
2019년 예측 매출량





03 데이터 분석

(1) 순환신경망 (LSTM)을 활용한 매출량 예측



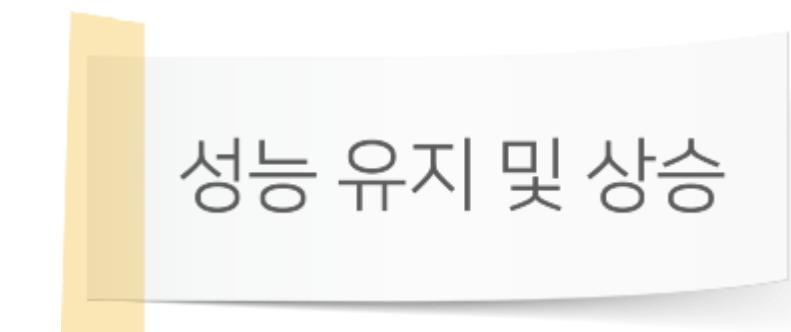
LSTM Model



17, 18 학습 모델을 통해
2020년 매출량 예측 시행

평균절댓값백분율오차(MAPE): 26.633

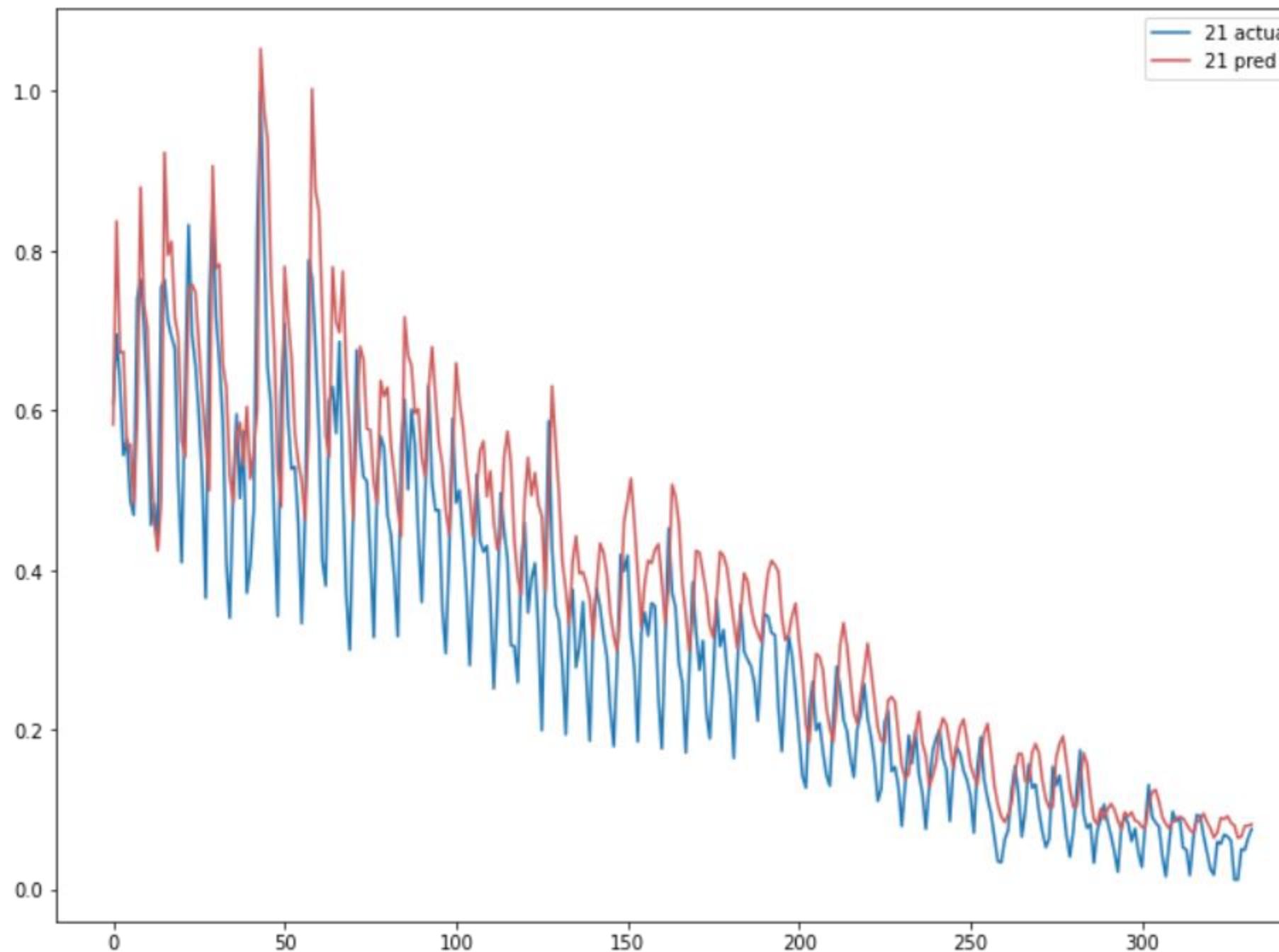
2020년 실제 매출량
2020년 예측 매출량





03 데이터 분석

(1) 순환신경망 (LSTM)을 활용한 매출량 예측



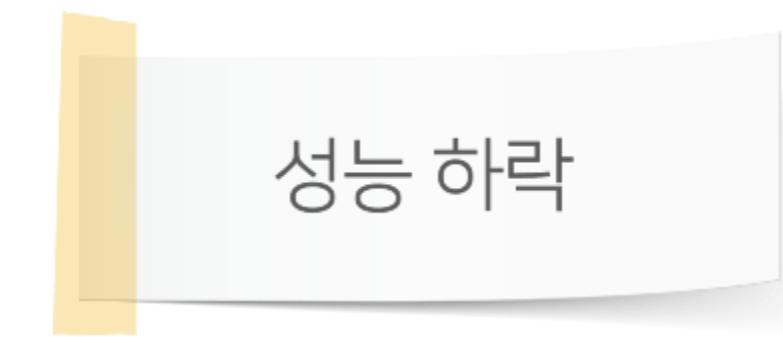
LSTM Model



17, 18 학습 모델을 통해
2021년 매출량 예측 시행

평균절댓값백분율오차(MAPE): 43.775

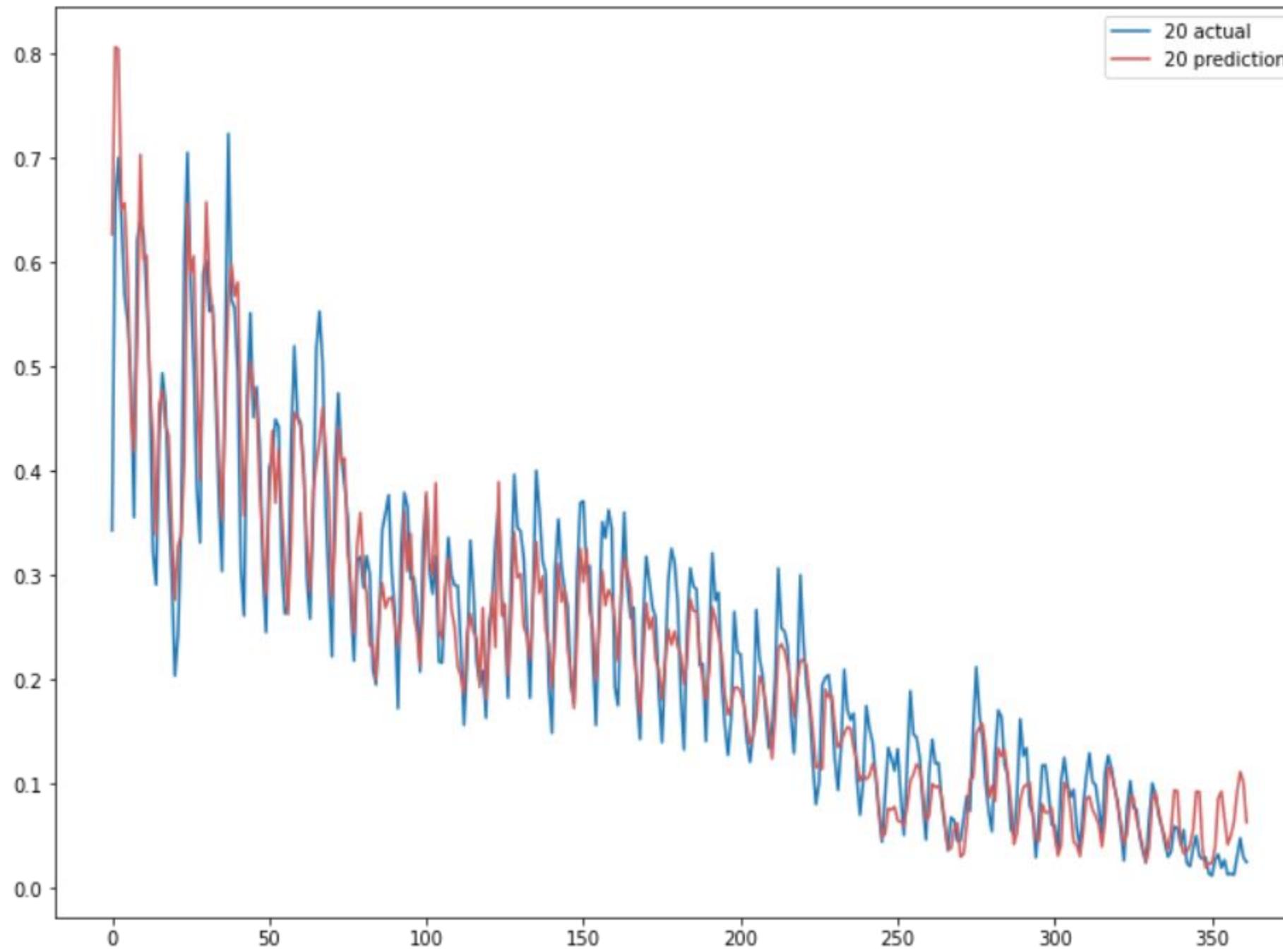
2021년 실제 매출량
2021년 예측 매출량





03 데이터 분석

(1) 순환신경망 (LSTM)을 활용한 매출량 예측



LSTM Model 2



Train set : 2017, 2018, 2019

Validation set : 2020 Data

평균절댓값백분율오차(MAPE): 24.89

손실 함수(MSE) : 0.00224918

2020년 실제 매출량
2020년 예측 매출량

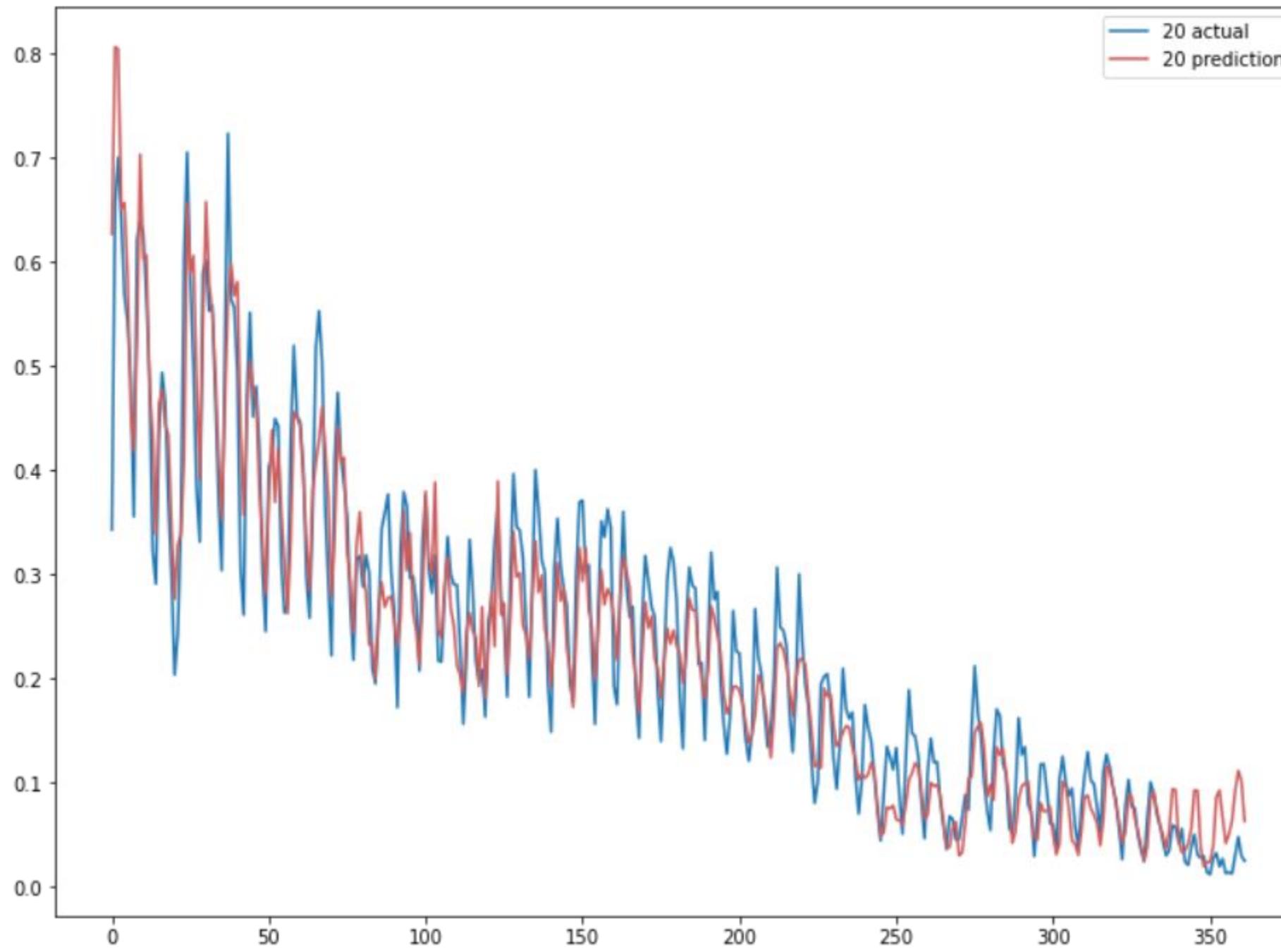
이전 모델에 비해
성능 향상





03 데이터 분석

(1) 순환신경망 (LSTM)을 활용한 매출량 예측



LSTM Model 2



Train set : 2017, 2018, 2019

Validation set : 2020 Data

평균절댓값백분율오차(MAPE): 24.89

손실 함수(MSE) : 0.00224918

2020년 실제 매출량
2020년 예측 매출량

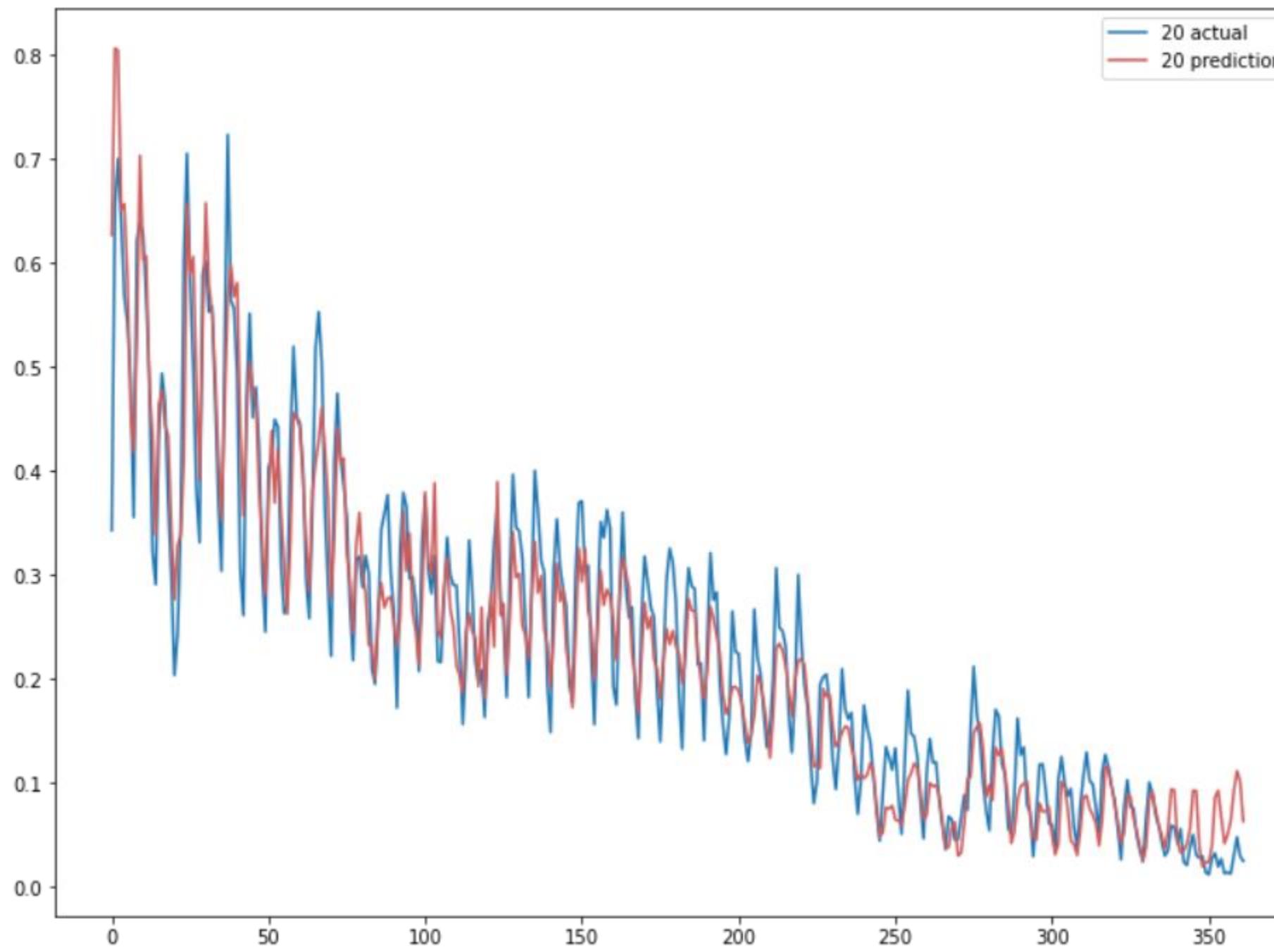
이전 모델에 비해
성능 향상





03 데이터 분석

(1) 순환신경망 (LSTM)을 활용한 매출량 예측



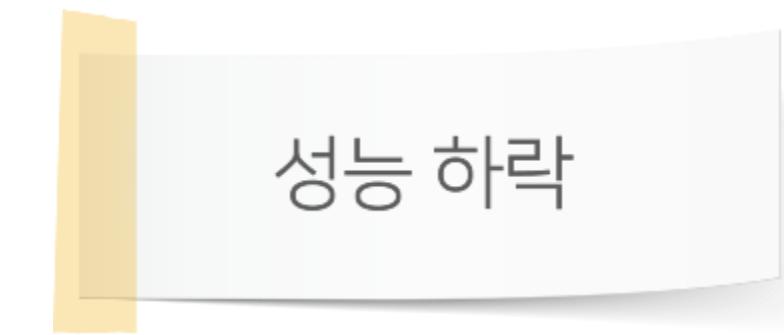
LSTM Model 2



17, 18, 19 학습 모델을 통해
2021년 매출량 예측 시행

평균절댓값백분율오차(MAPE): 50.19

2021년 실제 매출량
2021년 예측 매출량





03 데이터 분석

(2) 상품별 연간 총판매량 예측

트리기반 모델 비교 및 평가

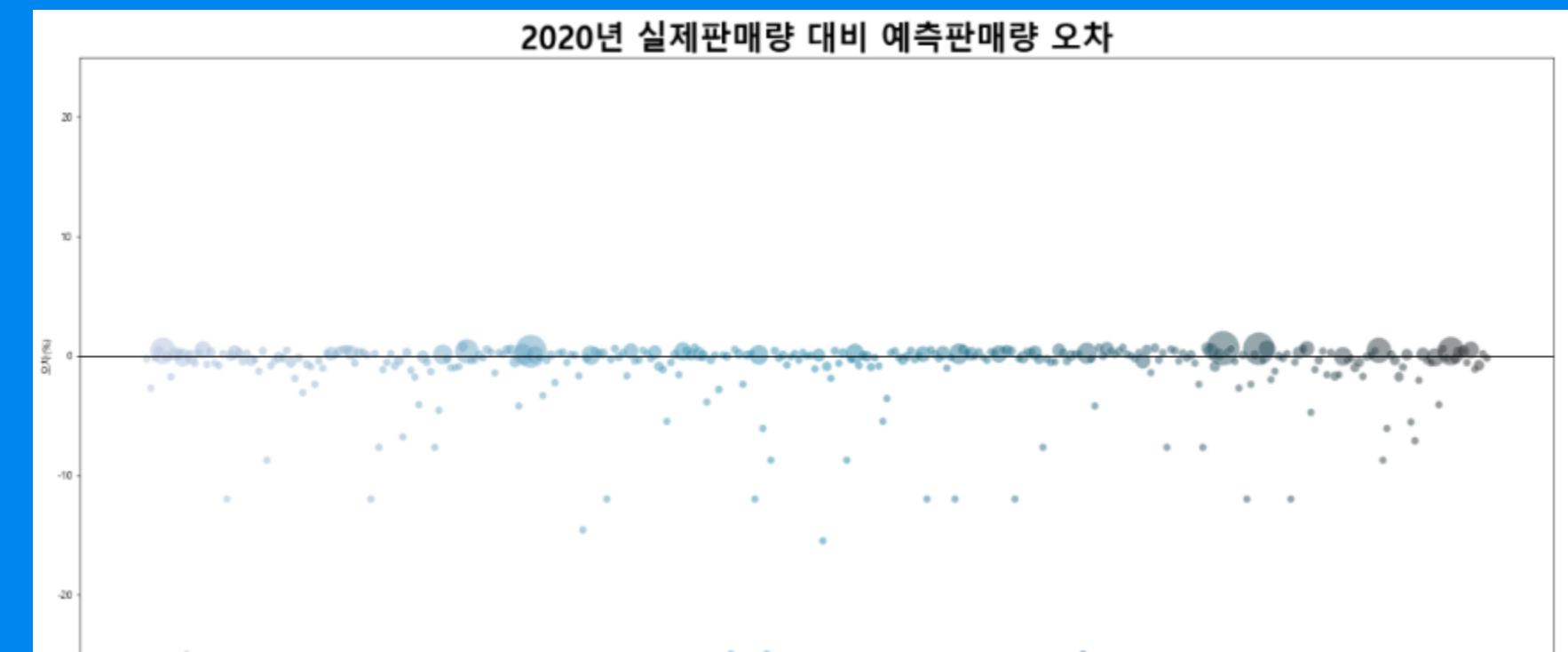
모델명	R2
Random Forest	0.8875
AdaBoost	0.7104
GradientBoosting	0.8342
XGBoost	0.8963

GridSearchCV, RandomizedSearchCV를 이용한 하이퍼 파라미터 최적화

GradientBoosting과 XGBoost 모델의 전반적인 예측 성능 비교 : XGBoost 최종 모델 선정

상품별 연간 판매량 예측 모델

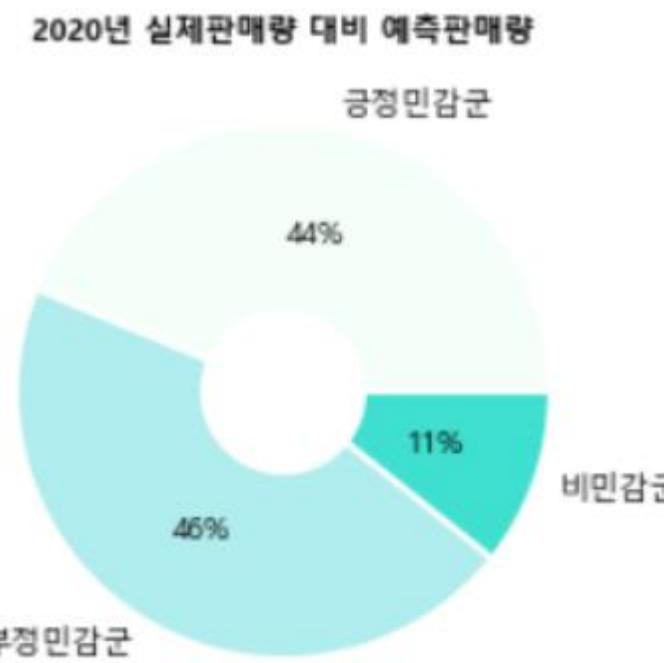
1. 2017년-2018년 데이터로 2019년 예측
2. 2017년-2020년 데이터로 2021년 예측
: 모델 설계 적합성 평가
3. 2017년-2019년 데이터로 2020년 예측
4. 2017년-2020년 데이터로 2021년 예측
: 2020년과 2021년의 급격한 매출 변동 재확인





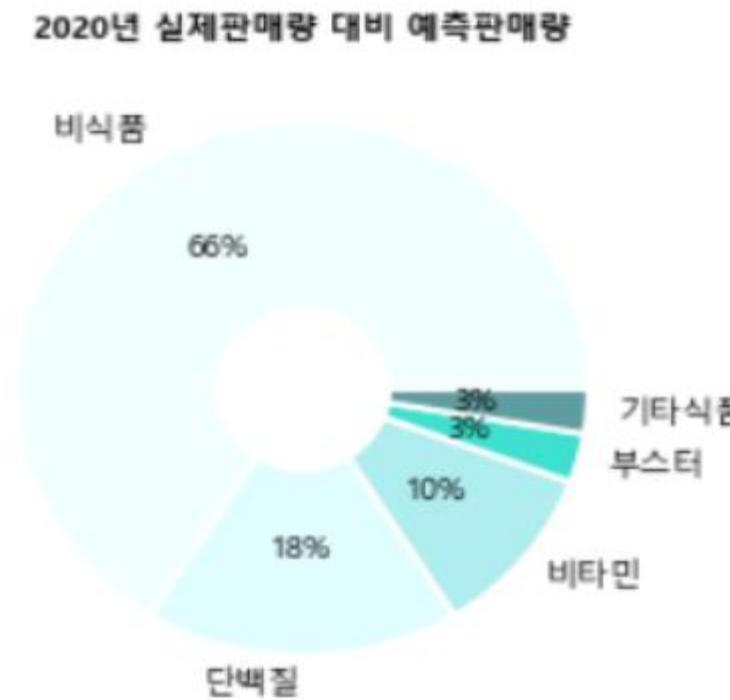
03 데이터 분석

(3) 2020년 실제 판매량 대비 예측 판매량



실제 판매량 대비 예측 판매량으로 변화 민감군 그룹화

부정민감군에 속하는 상품이 더 많음에도 불구하고 2020년 매출 증가 : 통계분석 결과 긍정민감군에 속하는 상품들의 매출과 매출오차가 부정 민감군에 속하는 상품들보다 더 커, 전체적인 매출 증가로 이어짐
긍정민감군과 부정민감군 상품의 수를 조절하기 보다는, 잘 팔리는 상품 을 더 잘 팔리게 하는 마케팅 전략 필요



카테고리별 실제 판매량 대비 예측판매량 비교

긍정민감군 내에서 비식품 카테고리가 차지하는 비중이 66%로 1위
현재 단백질과 부스터 카테고리 중심의 판매전략에서 나아가 상품군의 폭 을 넓히는 사업 영역의 확장 필요



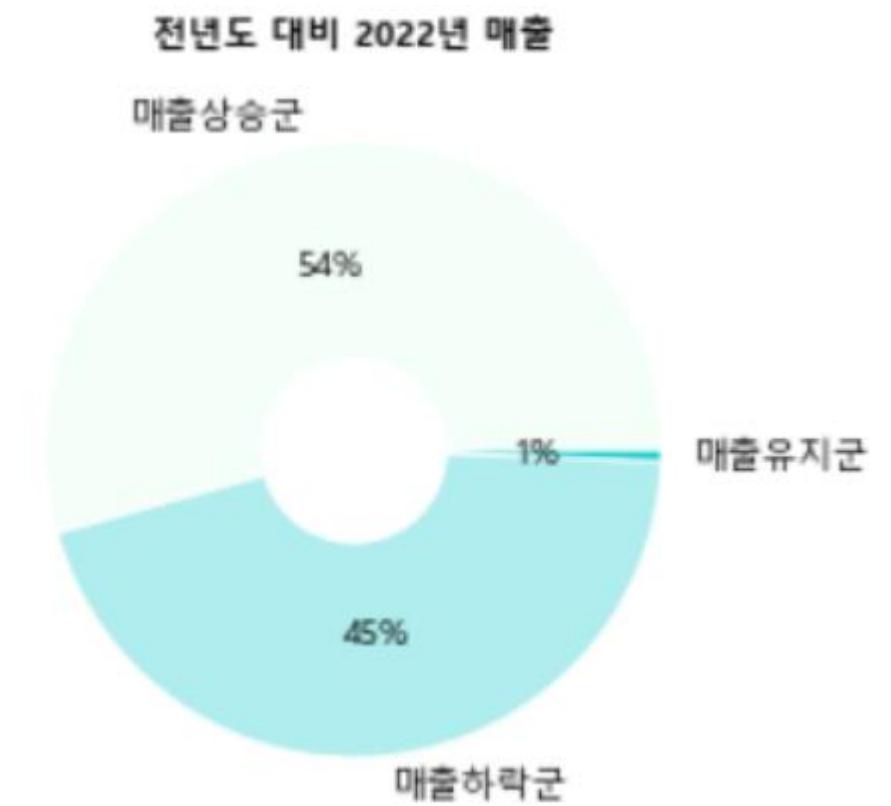
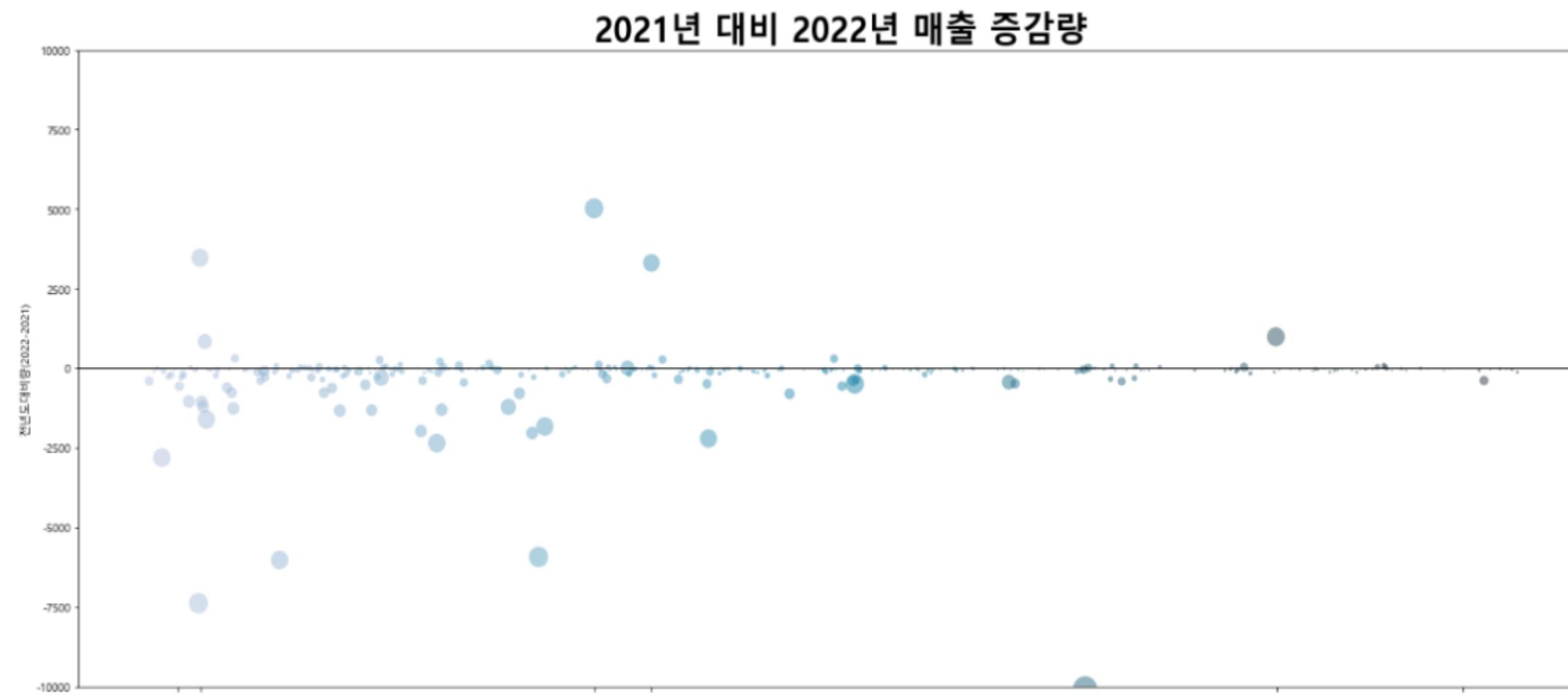


03 데이터 분석

(4) 2022년 매출 예측

2022년 상품별 총매출이 2020년과 2021년에 비해 감소할 것으로 전망

: 매출 상승군에 해당되는 상품 목록이 더 많음에도 불구하고, 쇼핑몰 전체 매출량 감소 예측. 통계분석 결과 매출 상승군의 상승폭 평균이 매출 하락군의 하락폭 평균보다 작아 2022년 전체 매출의 감소로 이어짐. 매출 하락 리스크를 줄이기 위한 새로운 비즈니스 전략 필요





03 데이터 분석

(4) 연관 상품 판매 예측

Apriori algorithm을 이용한 빈발 항목 및 연관 규칙 후보 도출



itemset
0 [100% 웨이 프로틴 56서빙, 100% 골드 스탠다드 웨이 네츄럴 68서빙]
1 [18 인치 폼 률러, 100% 골드 스탠다드 웨이 네츄럴 68서빙]
2 [3XT 맥스 에너지 30서빙, 100% 골드 스탠다드 웨이 네츄럴 68서빙]
3 [100% 골드 스탠다드 웨이 네츄럴 68서빙, 5N 리버 & 오르간 디펜더 270캡슐]
4 [100% 골드 스탠다드 웨이 네츄럴 68서빙, 7일 클렌즈 42식물성캡슐]

빈발 항목 목록 도출 결과





03 데이터 분석

(4) 연관 상품 판매 예측

XGBoostRegressor를 이용한 연관 상품 집합 판매 예측

Train Set 정확도: 0.8404405594875131

Test Set 정확도: 0.8693308912838804

-----설명변수 중요도-----

	Feature	Importance
1	2018	0.942923
0	2017	0.057077

2019년 예측 정확도 및 변수 중요도

Train Set 정확도: 0.9739427322421754

Test Set 정확도: 0.9398845514532838

-----설명변수 중요도-----

	Feature	Importance
3	2020	0.853090
0	2017	0.054490
2	2019	0.053367
1	2018	0.039052

2020년 예측 정확도 및 변수 중요도





04 의의



1. 건강기능식품 쇼핑몰 연매출 영향 요인 파악

건강기능식품 시장 규모 이외의 쇼핑몰 연매출에 영향을 주는 요인을 탐색하여, 쇼핑몰이 외부 환경 요인에 민감하게 반응한다는 특성 발견. 전년도 요인의 변화를 통해 미래 매출을 예측하고 적절한 비즈니스 대응 전략 수립

2. 외부 변화 민감 상품군 파악

리스크 대응 전략 수립을 위해 외부 환경 요인에 민감하게 반응하는 상품군 구별. 변화 민감 상품군을 파악하여 유연한 재고 관리 전략 마련

3. 쇼핑몰 미래 매출 예측

2020년과 2021년 성장에도 불구하고 2022년 미래 매출이 감소할 것이라는 예측을 통해 조기 리스크 대비 전략 수립 가능





05 서비스 활용 방안

기업 매출 향상 및 리스크 회피를 위한 비즈니스 전략 탐색 서비스

서비스 상세 제공 내용

1. 매출에 영향을 미치는 외부 환경 요인 분석하여 리스크 탐색
2. 상품별 리스크 민감도를 분류하여 재고관리 전략 제시
3. 미래 매출 예측을 통해 조기 리스크 발견





03 의의



아리마, lstm 등의 모델 구축하여
다양한 시계열 모델 비교 및 분석

연관분석

- 21년 매출 예측에 대한 구체적 요인까지 도출
- 3일 뒤 매출량을 예측하는 모델이기 때문에
장기적인 매출 분석 불가

- 연관 상품 항목 도출
- 고객별 장바구니 정보의 부재로 정확한 개인
추천 시스템으로의 발전 불가

