

딥러닝 기반의 수산물 수입가격 예측을 통한 최적의 가격 예측 모형 도출

Seafood import price prediction model based on deep learning



Work Team Name & Members

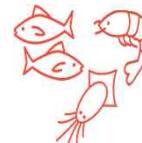
W_W(원터위너)
곽원일(팀장), 송민아, 김유환, 김지훈

Work Schedule

2021.08.24 주제선정, 데이터 수집
2021.08.30 데이터 전처리, 데이터 분석
2021.09.03 모델 설계 및 개발
2021.09.13 모델 평가 및 검증
2021.09.24 장고 웹프레임워크 구현

Work Rule

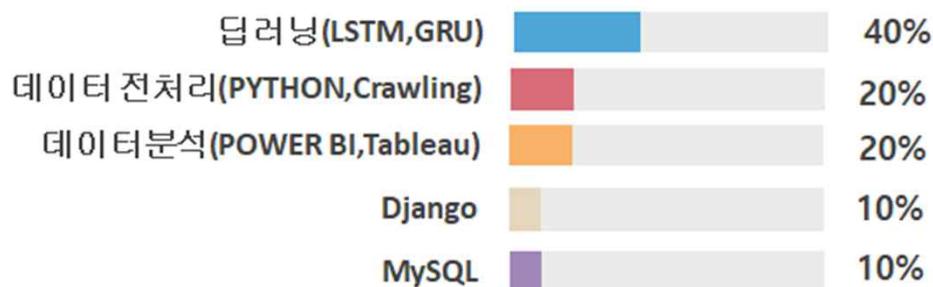
Train Data : 2016년 ~ 2020년 수산물 수입데이터
Test Data : 2021년 1월 ~ 6월 예측
데이터 전처리 방법, 활용 알고리즘 설명



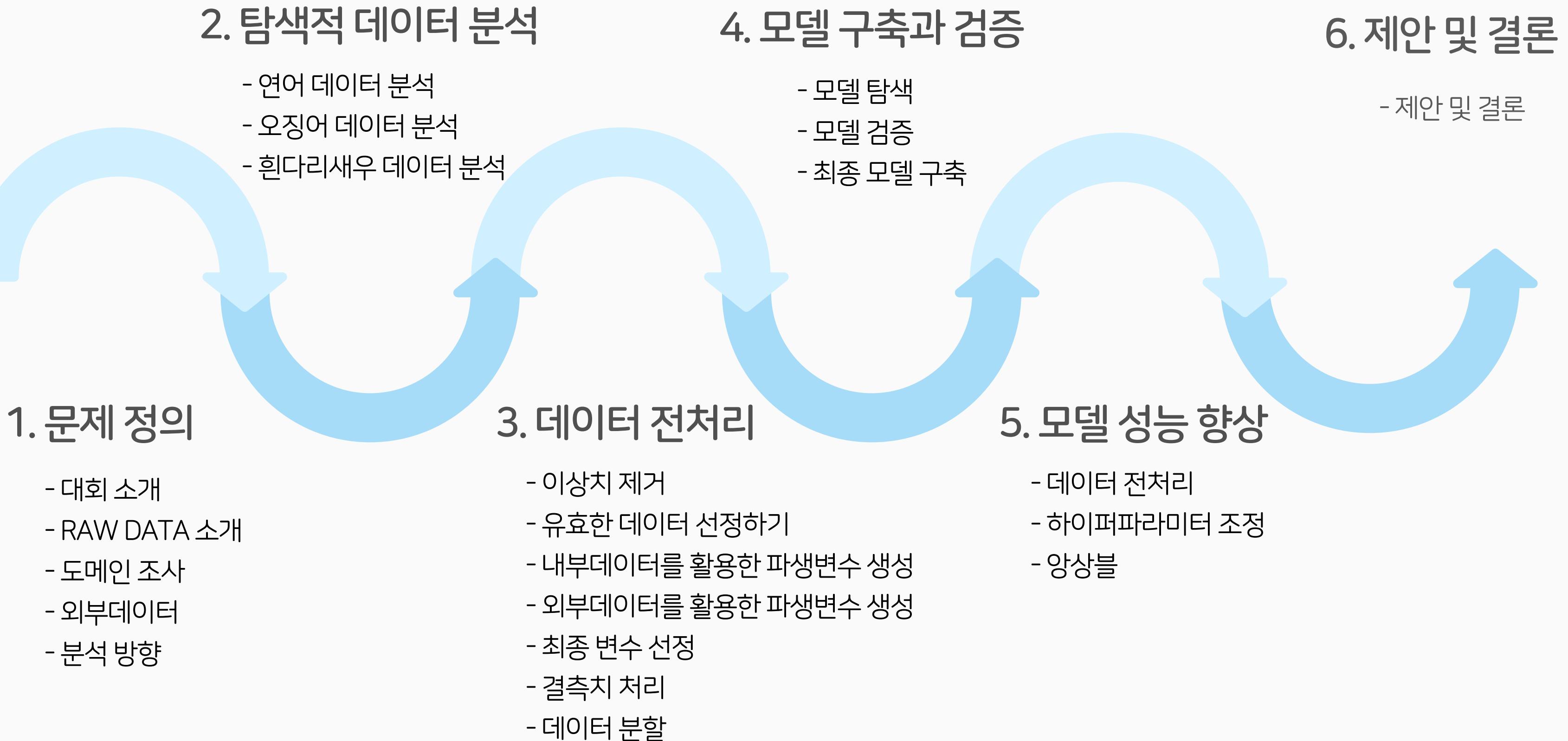
수산물(연어,오징어,흰다리새우) 주별 평균단가 예측 서비스

빅데이터 생태계 조성을 통한 해양수산 뉴딜 정책 실현 및
수산업 이해 관계자의 경영계획 수립 기반 구축

Skills



목 차





대회 설명

주관	한국해양수산개발원
주제	수산물 수입가격 예측을 통한 최적의 가격 예측 모형 도출
목적	빅데이터 생태계 조성을 통한 해양수산 뉴딜 정책 실현 및 수산업 이해관계자의 경영계획 수립 기반 구축
데이터	2016년 ~ 2020년의 5년간 수산물 수입평균단가 데이터
예측시점	2021년 1월 4일 ~ 2021년 6월 28일
예측품목	오징어, 연어, 흰다리새우
평가척도	RMSE



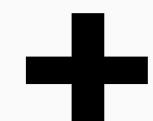
RAW DATA 소개

- 수산물수입평균단가 데이터

REG_DATE	P_TYPE	CTRY_1	CTRY_2	P_PURPOSE	CATEGORY_1	CATEGORY_2	P_NAME	P_IMPORT_TYPE	P_PRICE
2015-12-28	수산물	아르헨티나	아르헨티나	판매용	갑각류	새우	아르헨티나 붉은새우	냉동	7.48
2015-12-28	수산물	바레인	바레인	판매용	갑각류	게	꽃게	냉동	2.92
2015-12-28	수산물	바레인	바레인	판매용	갑각류	게	꽃게	냉동, 절단	3.36
2015-12-28	수산물	칠레	칠레	판매용	패류 명게류	해삼	해삼	건조, 자숙	18.26
2015-12-28	수산물	중국	중국	판매용	어류	서대 박대 페루다	서대	냉동	4.79



2015년 12월 28일 ~
2019년 12월 30일 데이터



2020년 1월 6일 ~
2020년 12월 28일
추가 데이터



51,552개 데이터
&
10개의 변수



도메인 조사

- 수산물 수입가격에 대한 전반적인 지식과 동향 습득

> 한국해양수산개발원 수산물 수입 동향

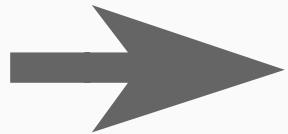
> 수산물수출정보포털

> 유엔식량농업기구 수산양식 통계데이터

> 관세청 수산물 수출입 무역통계



도메인 조사 결과 수입가격은 이상치가 생기는 경우가 빈번



데이터 전처리과정에서 이상치를 어떻게 처리할 것인지가 핵심

 외부데이터
> 원달러 환율 (출처: 서울외국환중개소)

날짜	통화명	환율
2015.12.28	미 달러화(USD)	1,171.00
2015.12.29	미 달러화(USD)	1,165.00
2015.12.30	미 달러화(USD)	1,167.70
2015.12.31	미 달러화(USD)	1,172.00
2016.01.04	미 달러화(USD)	1,172.00

> 두바이유 시세 (출처: Investing)

날짜	종가	오픈	고가	저가	거래량	변동 %
2021년 06월 28일	71.54	71.54	71.54	71.54	-	-0.16%
2021년 06월 25일	71.65	71.65	71.65	71.65	-	0.08%
2021년 06월 24일	71.60	71.60	71.60	71.60	-	0.11%
2021년 06월 23일	71.52	71.52	71.52	71.52	-	0.15%
2021년 06월 22일	71.41	71.41	71.41	71.41	-	-0.09%

> 활용 목적:

1. 환율이 수입가격에 영향을 미칠 것이라 판단
2. 원달러 환율이 가장 대표적인 지표

> 활용 목적:

1. 수출입 국가간 거리를 고려하여 유가가 수입가격에 영향을 미칠 것이라 판단
2. 두바이유가는 유가 중에서도 가장 대표적인 지표



분석방향

1

유효한
데이터로
정제

2

3트랙 분석 및
최적의 모델 생성

3

안정적으로
예측하는
모델 구축

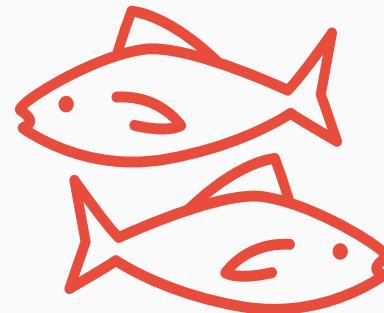
4

시사점
도출

목 차

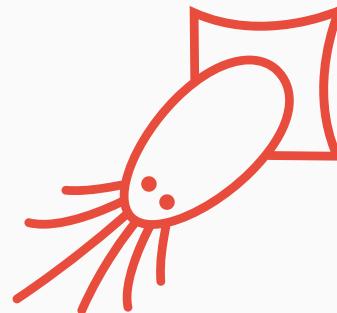


3 트랙 데이터 분석



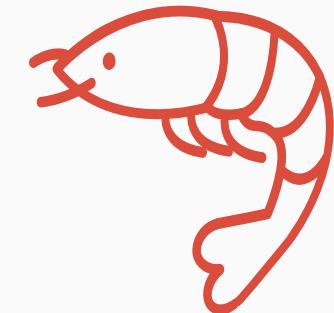
연어

- 1895개 데이터
- 제조국: 11개국
- 수출국: 12개국
- 수입용도: 4개
- 수입형태: 9개



오징어

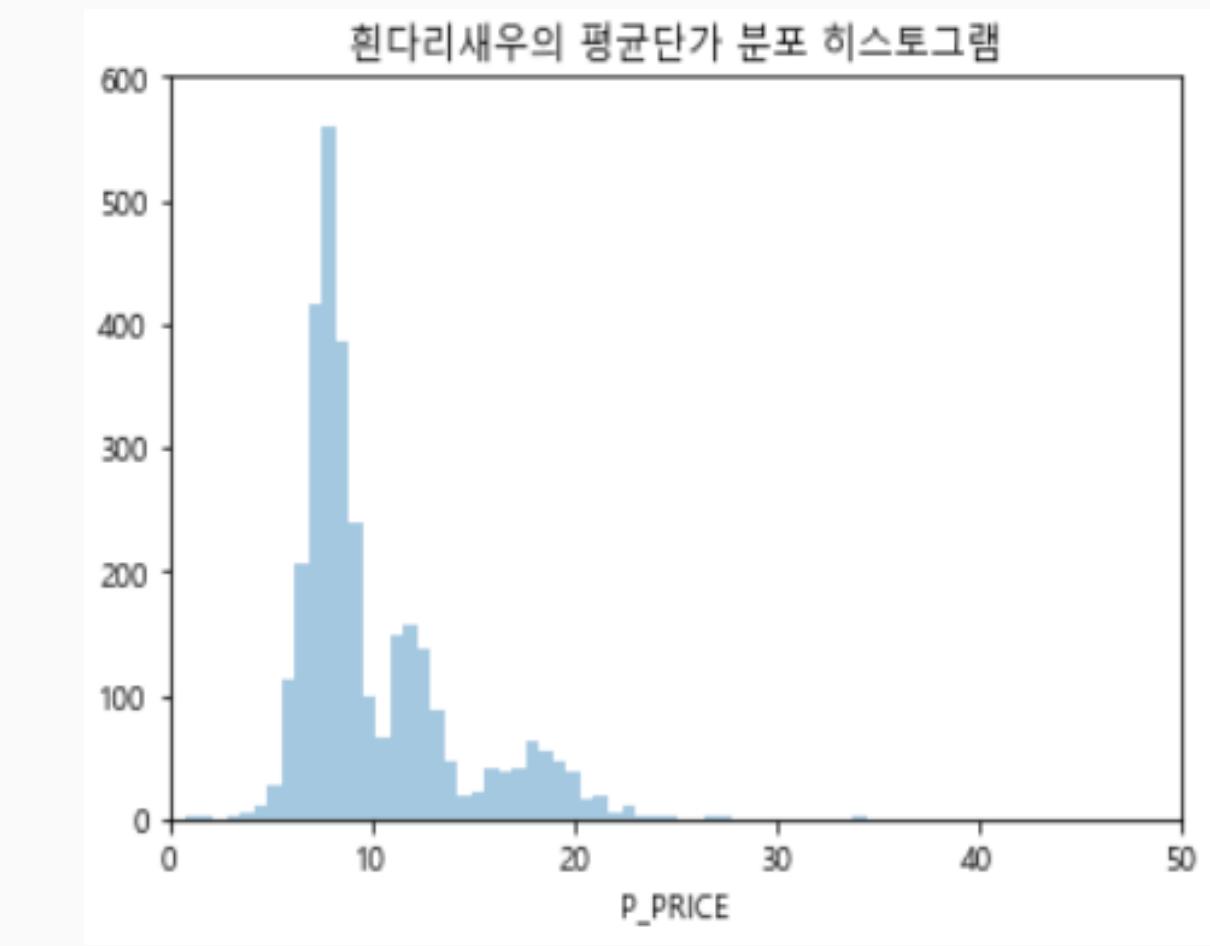
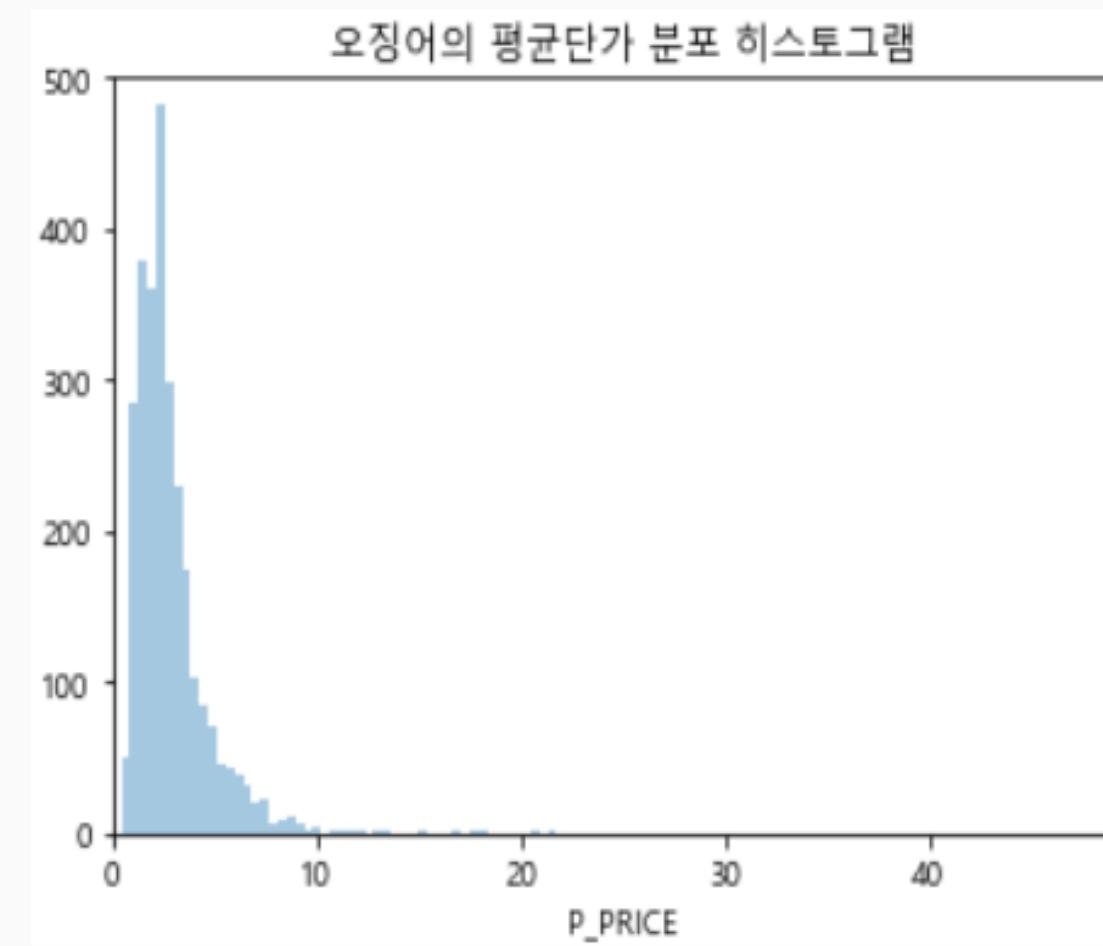
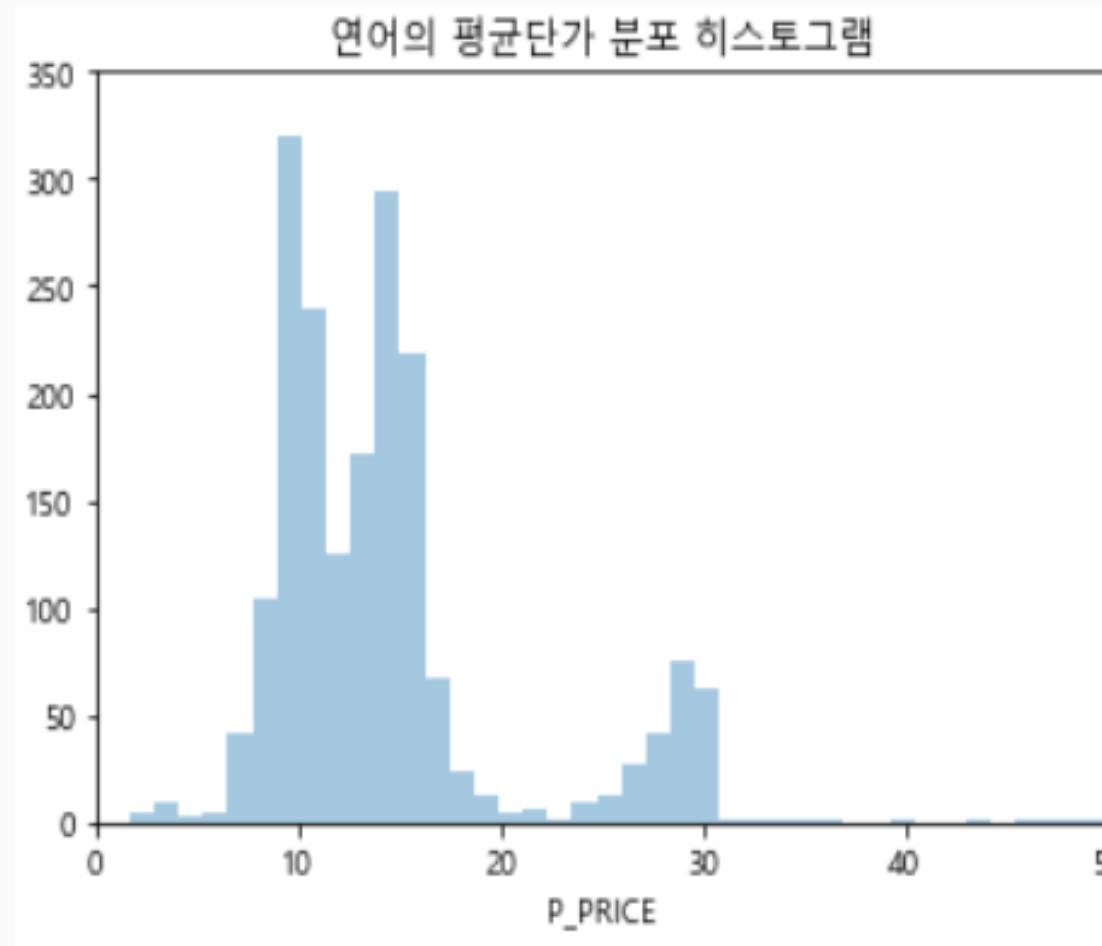
- 2771개 데이터
- 제조국: 18개국
- 수출국: 19개국
- 수입용도: 5개
- 수입형태: 13개



흰다리새우

- 3133개 데이터
- 제조국: 12개국
- 수출국: 13개국
- 수입용도: 2개
- 수입형태: 7개

3 트랙 데이터 분석



68

각 어종별 평균단가의 히스토그램 분석결과 분산과 분포가 제각각이므로 서로 다른 형태의 데이터 전처리과정이 필요

69



연어 데이터 분석



주별 평균가격 : \$ 14.61

주별 평균가격의 최댓값 : \$ 21.40

주별 평균가격의 최솟값 : \$ 11.18

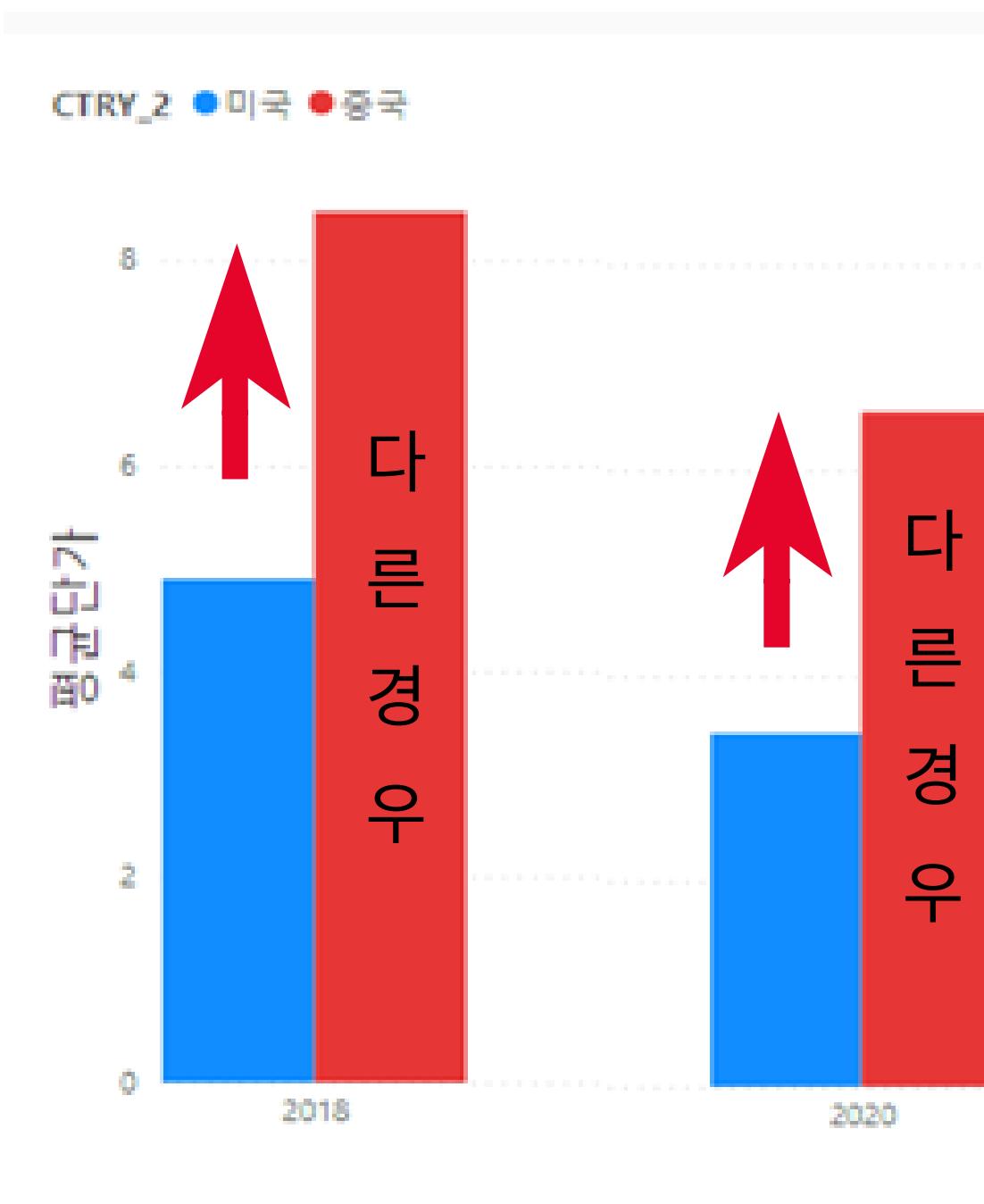
분석 결과

1. 안정적인 흐름을 보이다 이상치가 발생하는 이벤트들 발생
2. 2017년에 발생한 이벤트: 장기적인 이슈
(자사제품제조용 냉장, 필렛(F) 5월부터 7월까지 가격이 폭등)
3. 이상치를 제거하고 이벤트 변수를 찾아내는 것
→ 연어 주별 평균가격 예측 모델 구축의 핵심
4. 2020년 팬데믹으로 인한 급격한 가격변동 이슈 발생
5. 주별 평균가격의 분산이 크기 때문에 안정적인 모델 구축 필요

연어 데이터 분석

- 제조국과 수출국이 다른경우

<제조국과 수출국이 다를 때 가격차이>



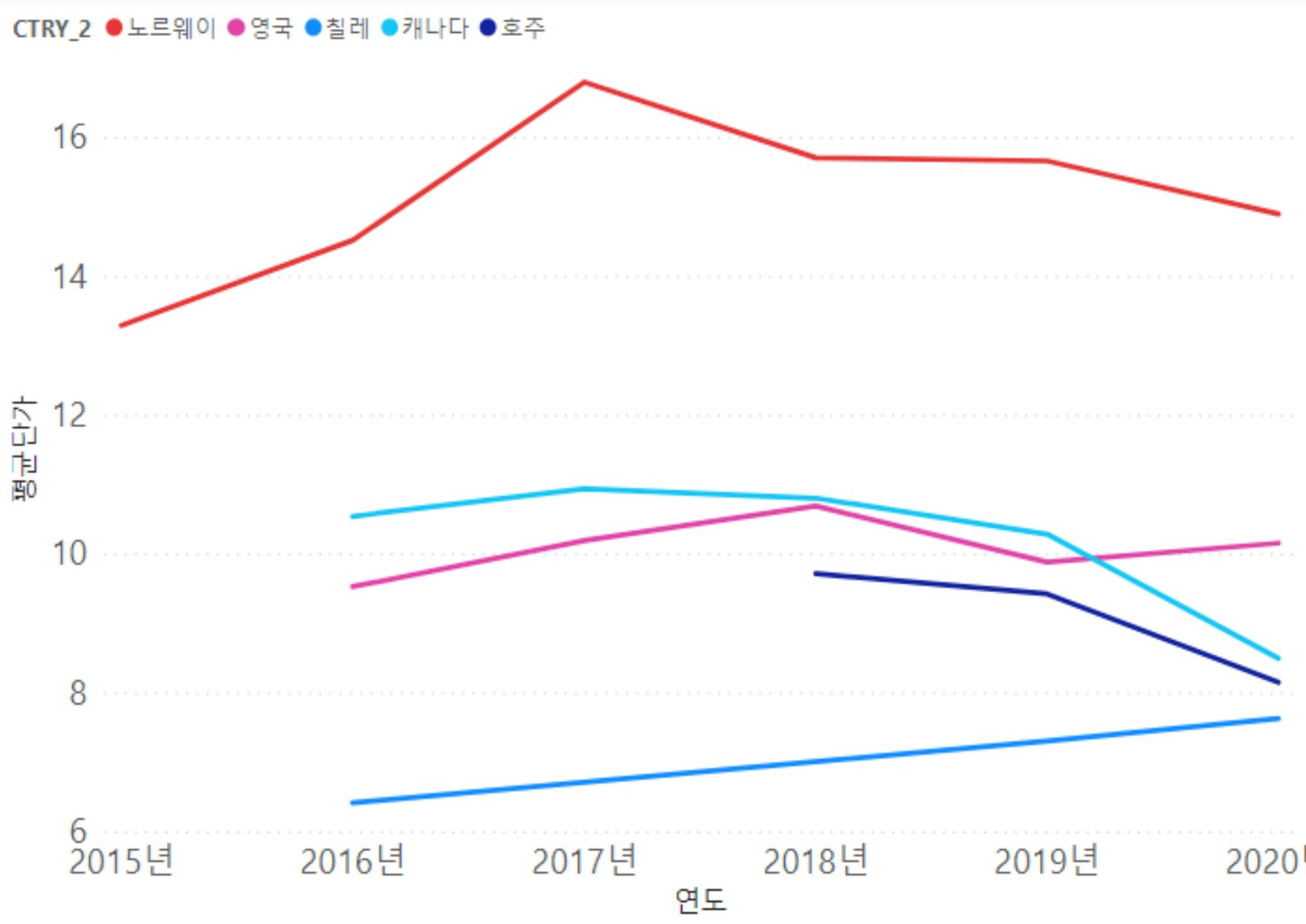
REG_DATE	P_TYPE	CTRY_1	CTRY_2	P_PURPOSE	CATEGORY_1	CATEGORY_2	P_NAME	P_IMPORT_TYPE	P_PRICE
2018-11-05	수산물	노르웨이	베트남	판매용	어류	연어	연어	냉동,슬라이스(S) 프자회감	16.64
2018-12-10	수산물	미국	중국	판매용	어류	연어	연어	냉동,풀랫(F)	8.47
2020-10-19	수산물	미국	중국	판매용	어류	연어	연어	냉동,풀렛(F)	6.57

제조국과 수출국이 다를때 같은 조건임에도 가격이 높음
→ 그러나 총 데이터중 3건에 불과함

연어 데이터 분석

- 수출국이 다른 경우

<연어 상위 5국가 평균단가>



- 노르웨이: 1595건

- 영국: 147건

- 캐나다: 66건

- 호주: 31건

- 칠레: 23건

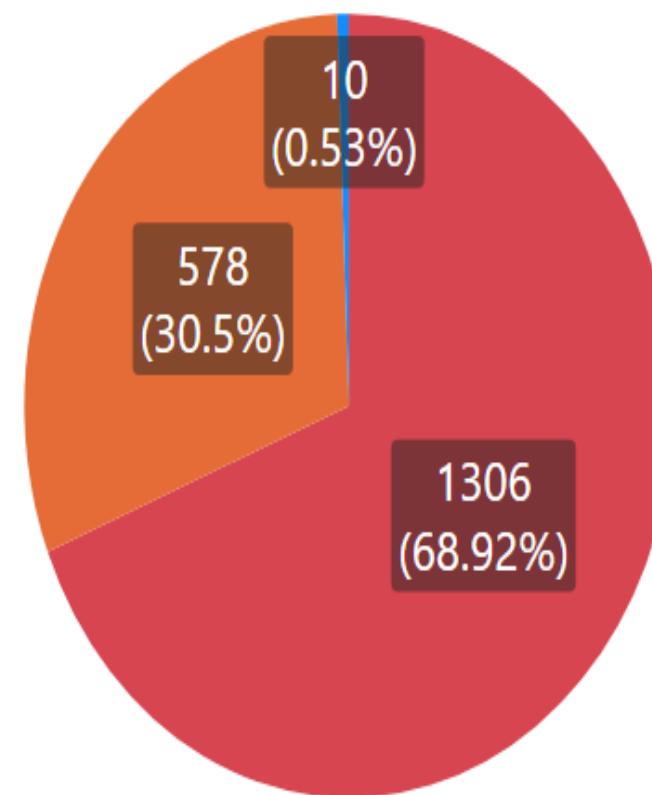
> 노르웨이: 가장 품목이 많으며 단가가 비쌈 → 비싼 판매용, 냉장, 포장횟감, 필렛(F)이 수입되기 때문

연어 데이터 분석

- 수입용도별

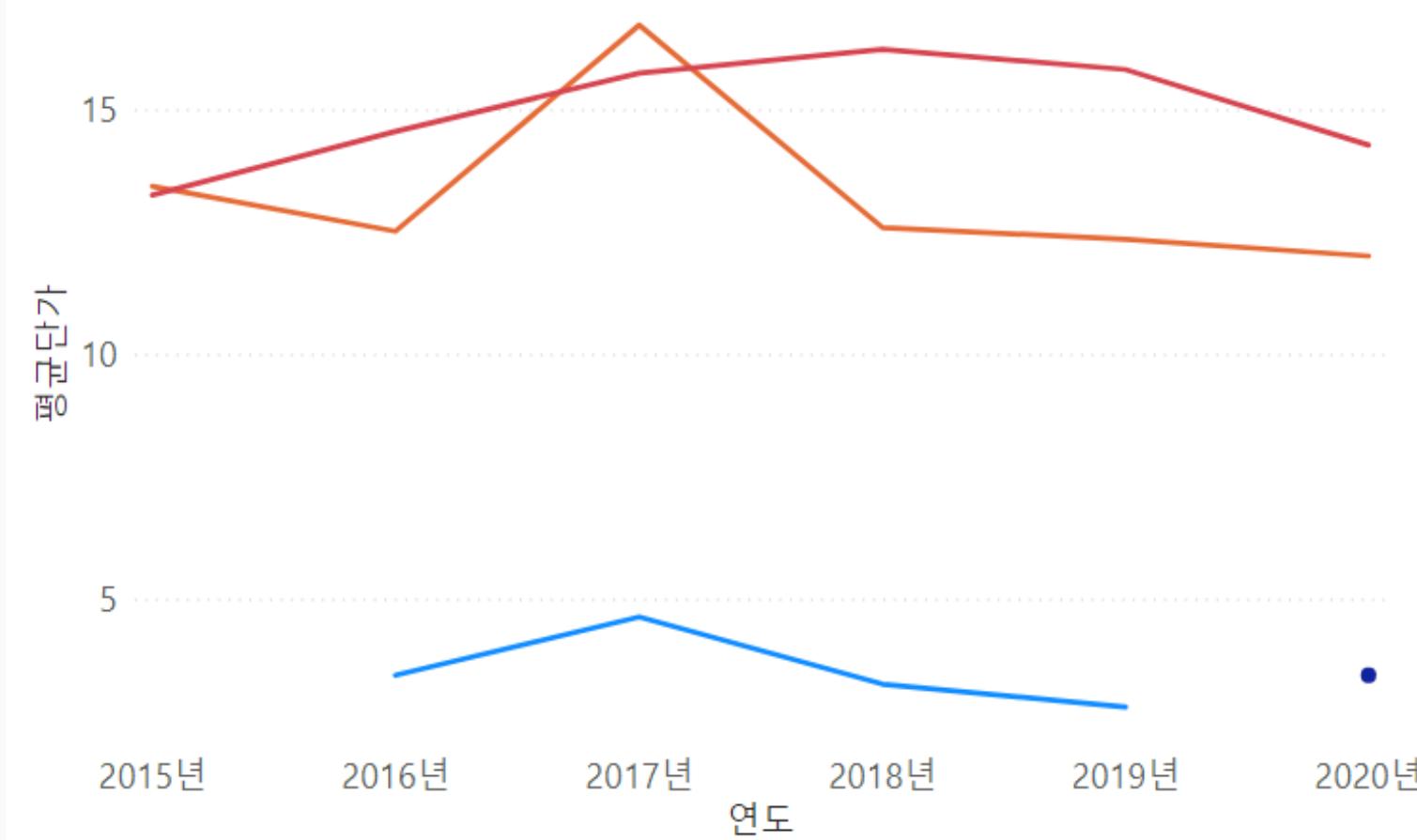
<연어 수입용도별 비율>

P_PURPOSE ● 판매용 ● 자사제품제조용 ● 외화회득용 원료 ● 외화회득용 제품



<연어 수입용도별 연평균단가>

P_PURPOSE ● 외화회득용 원료 ● 외화회득용 제품 ● 자사제품제조용 ● 판매용



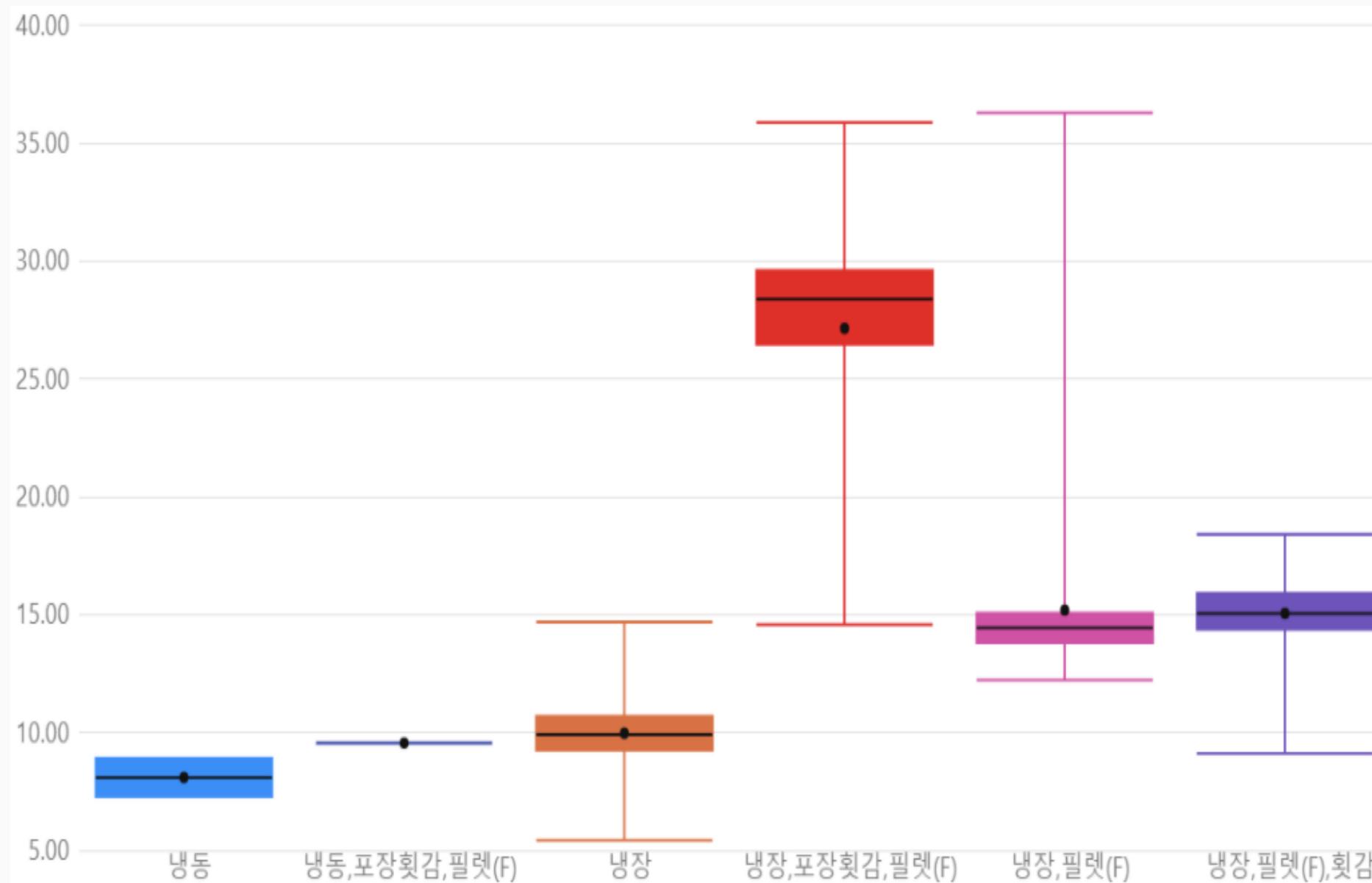
> 자사제품제조용: 전체 데이터의 32%(578건)를 차지, 2017년에 가격 급등



연어 데이터 분석

- 수입형태별

<연어 수입형태별 평균단가 박스플롯>



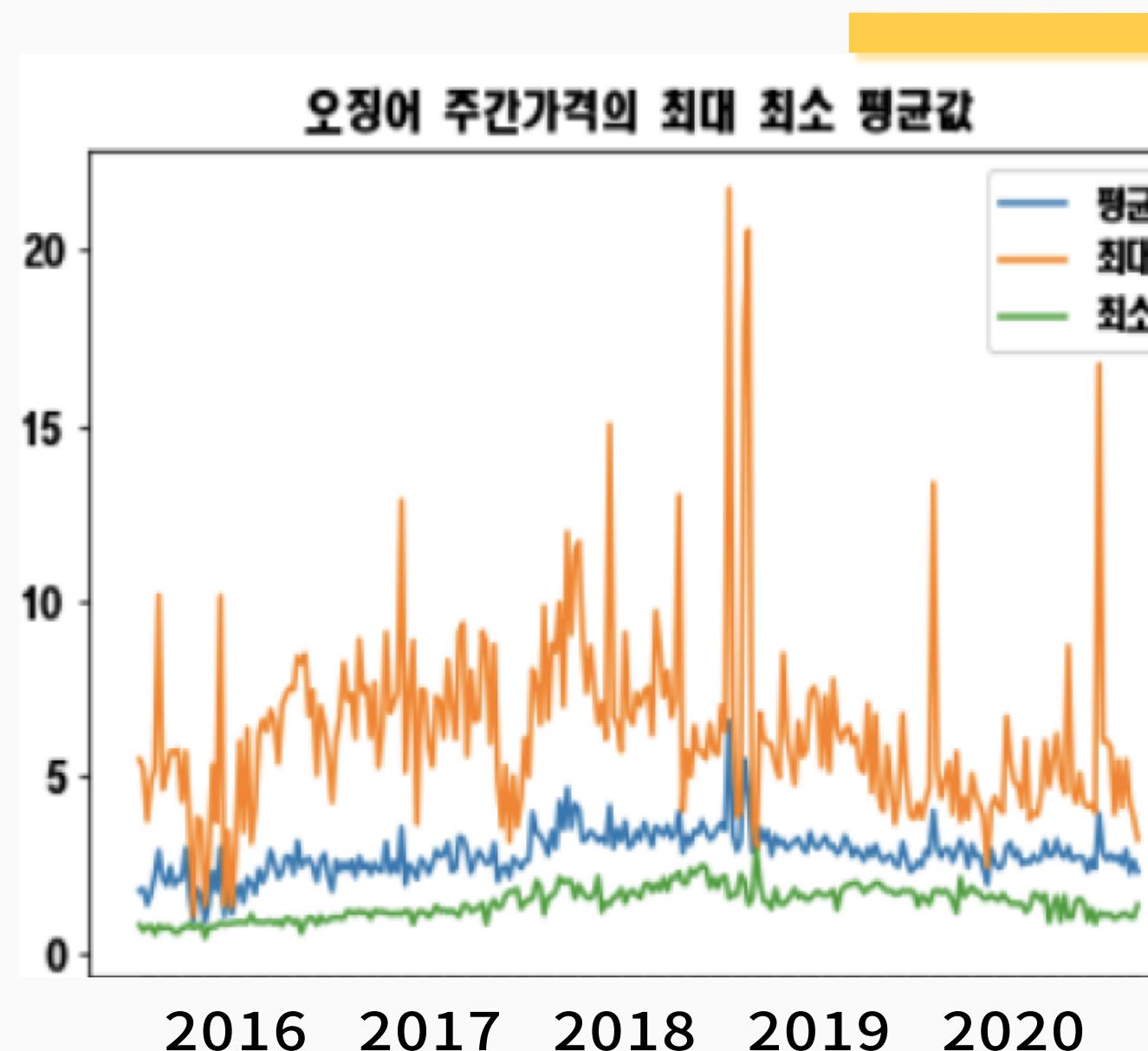
1. 냉장,포장횟감,필렛(F)

- 가장 비쌈, 총 274건

2. 냉장,필렛(F)

- 가격 분산이 크기 때문에 이상치 확인 필요

오징어 데이터 분석



주별 평균가격 : \$ 2.77

주별 평균가격의 최댓값 : \$ 6.56

주별 평균가격의 최솟값 : \$ 0.84

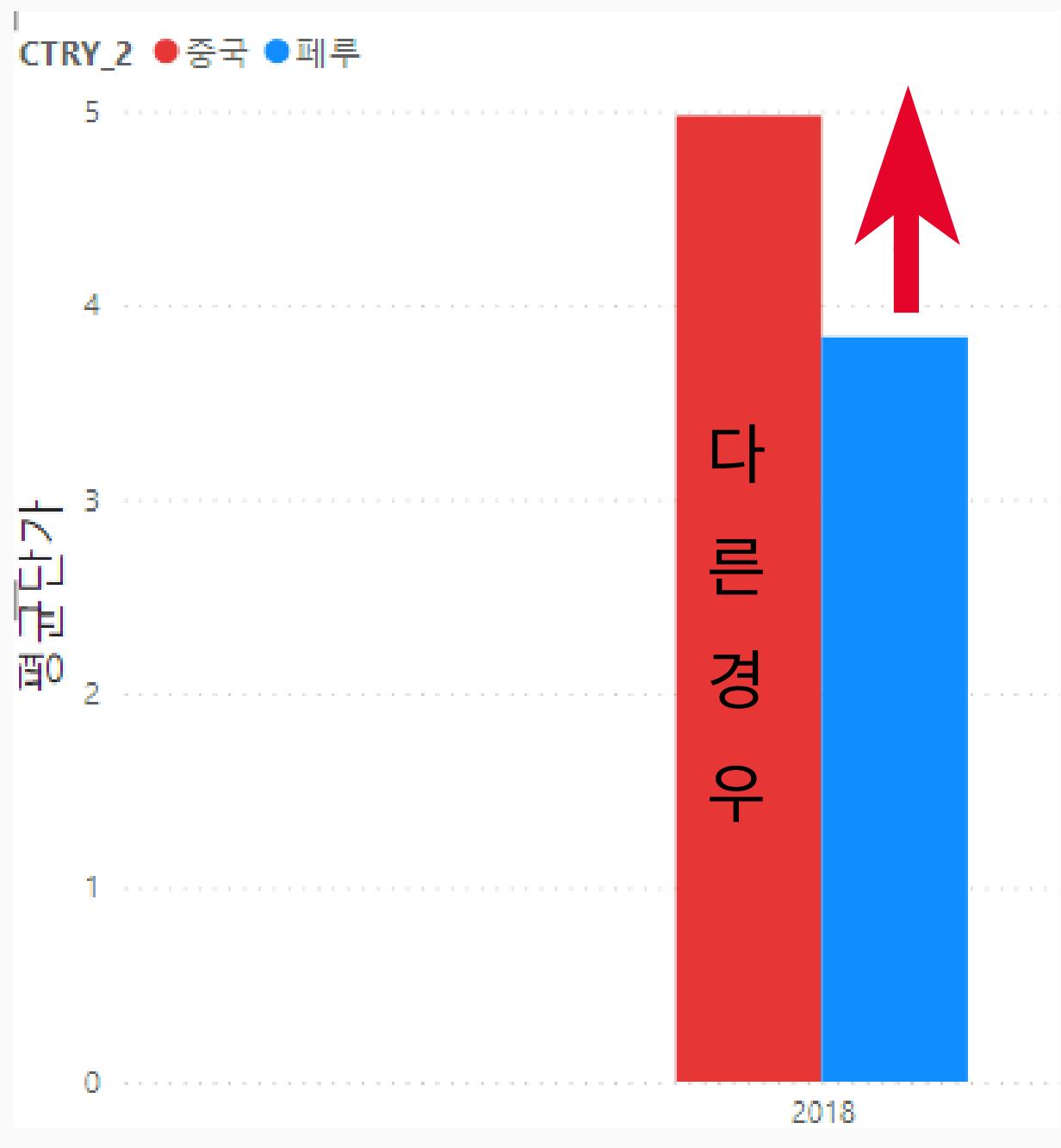
분석 결과

- 가격의 편차가 크기 때문에 딥러닝 모델 학습 시 오버 피팅이 일어나지 않도록 유의
- 수입형태 중 (건조), (냉동, 슬라이스, 포장횟감)의 가격이 높아 최댓값의 분포가 매우 큼
- 최댓값의 큰 분산을 어떻게 조절하느냐가 오징어 가격 예측 모델의 핵심
- 2016년 상반기 가격폭락이 있었음
- 팬데믹 이슈에도 큰 영향을 받지 않은 것으로 판단

오징어 데이터 분석

- 제조국과 수출국이 다른 경우

<제조국과 수출국이 다를 때 가격차이>



REG_DATE	P_TYPE	CTRY_1	CTRY_2	P_PURPOSE	CATEGORY_1	CATEGORY_2	P_NAME	P_IMPORT_TYPE	P_PRICE
2016-01-25	수산물	칠레	일본	판매용	연체류 해물모듬	오징어	오징어	냉동,다리	0.500000
2016-02-01	수산물	대한민국	중국	판매용	연체류 해물모듬	오징어	오징어	냉동	1.700000
2016-02-22	수산물	칠레	일본	판매용	연체류 해물모듬	오징어	오징어	냉동,다리	1.520000
2016-04-25	수산물	대한민국	중국	판매용	연체류 해물모듬	오징어	오징어	냉동	2.150000
2016-05-30	수산물	대한민국	중국	판매용	연체류 해물모듬	오징어	오징어	냉동,지느러미	1.320000
...
2020-11-09	수산물	대한민국	중국	판매용	연체류 해물모듬	오징어	오징어	냉동	3.212229
2020-11-16	수산물	대한민국	중국	판매용	연체류 해물모듬	오징어	오징어	냉동	3.219264
2020-11-23	수산물	뉴질랜드	중국	판매용	연체류 해물모듬	오징어	오징어	냉동	5.300000
2020-12-21	수산물	대한민국	중국	판매용	연체류 해물모듬	오징어	오징어	냉동	3.237740
2020-12-28	수산물	대한민국	중국	판매용	연체류 해물모듬	오징어	오징어	냉동	3.210000

2771건 중 137건

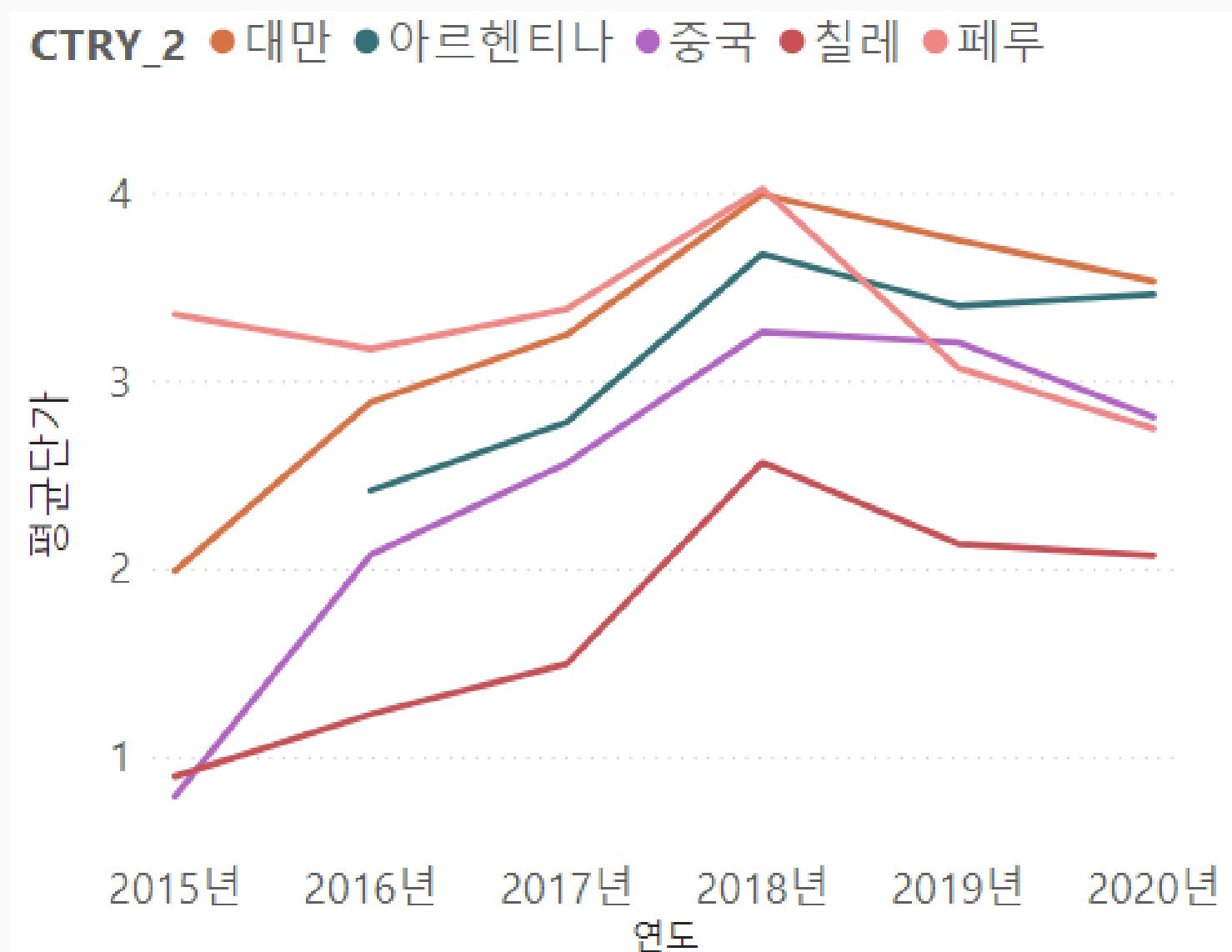
1. 137건으로 다른 품목에 비해 많은 비중을 차지
2. 대부분 가격대가 다른 조건보다 비싸지는 경향을 가지고 있음



오징어 데이터 분석

- 수출국별

<오징어 수출국별 평균단가 >



- 페루: 1013건

- 중국: 811건

- 칠레: 682건

- 아르헨티나: 78건

- 대만: 70건

> 평균단가 가격이 \$1~4 사이로 가격차이가 크지 않음

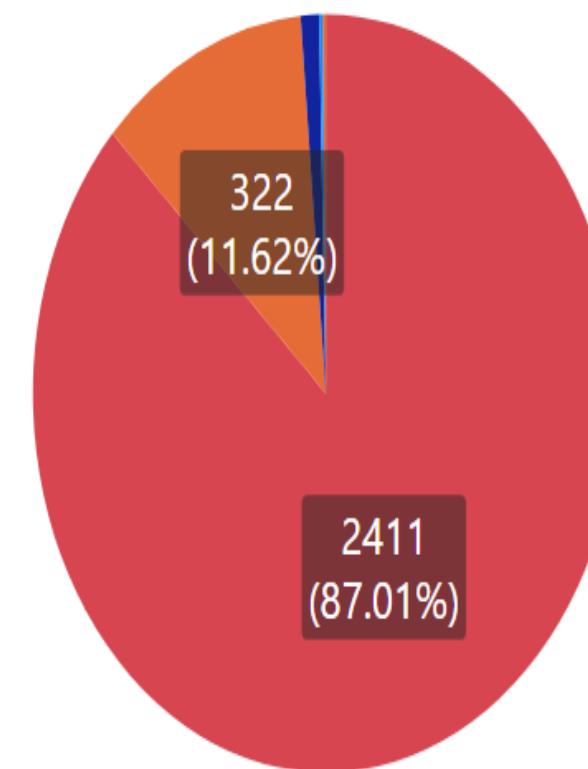
> 칠레: 주 수입형태(냉동,동체/냉동,다리/냉동,지느러미)이 냉동보다 가격이 낮아 평균단가가 낮음.

오징어 데이터 분석

- 수입용도별

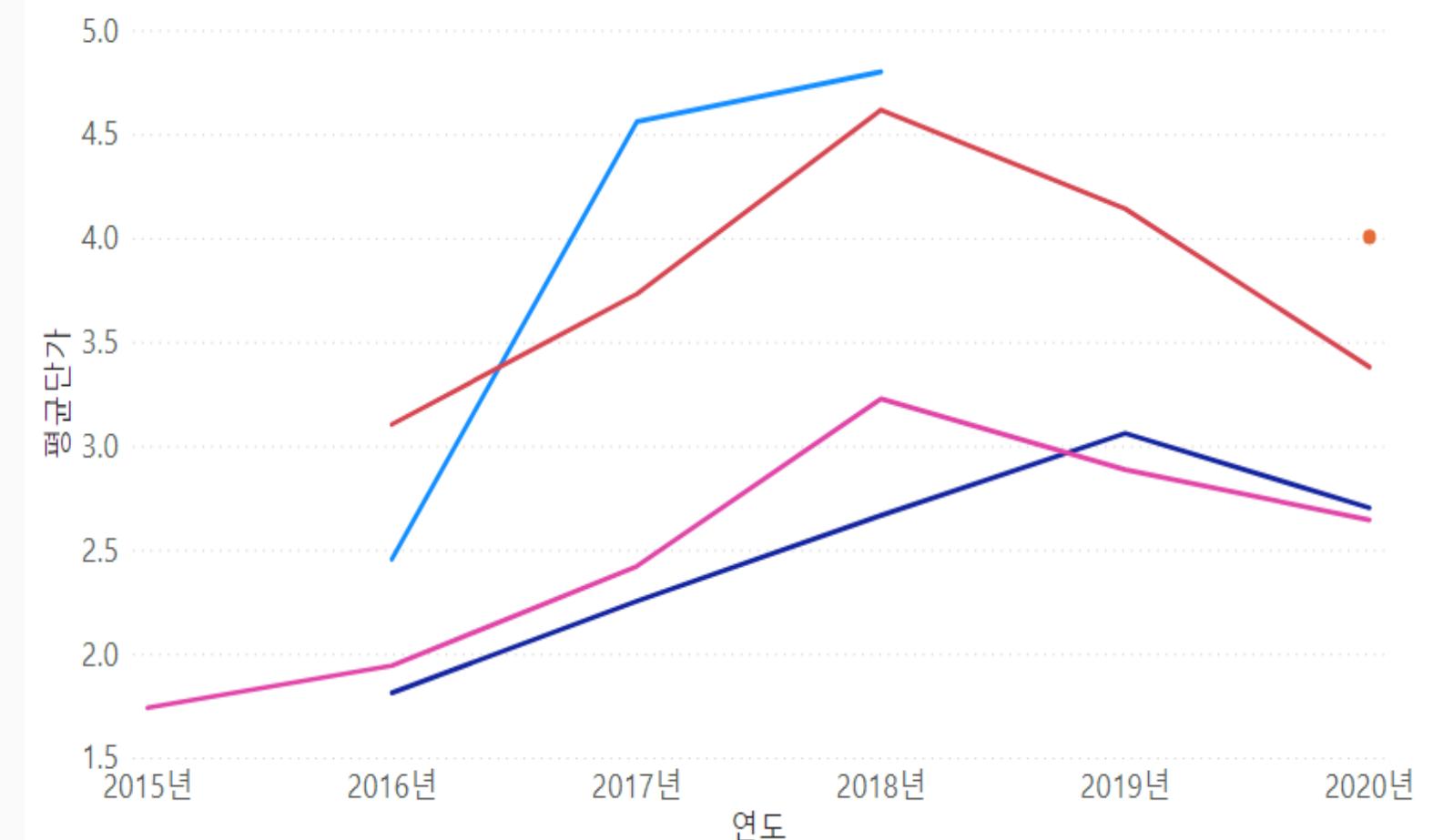
<오징어 수입용도별 비율>

P_PURPOSE ● 판매용 ● 자사제품제조용 ● 외화획득용 원료 ● 반송품(기타) ● 외화획득용 제품



<오징어 수출용도별 평균단가 >

P_PURPOSE ● 반송품(기타) ● 외화획득용 원료 ● 외화획득용 제품 ● 자사제품제조용 ● 판매용

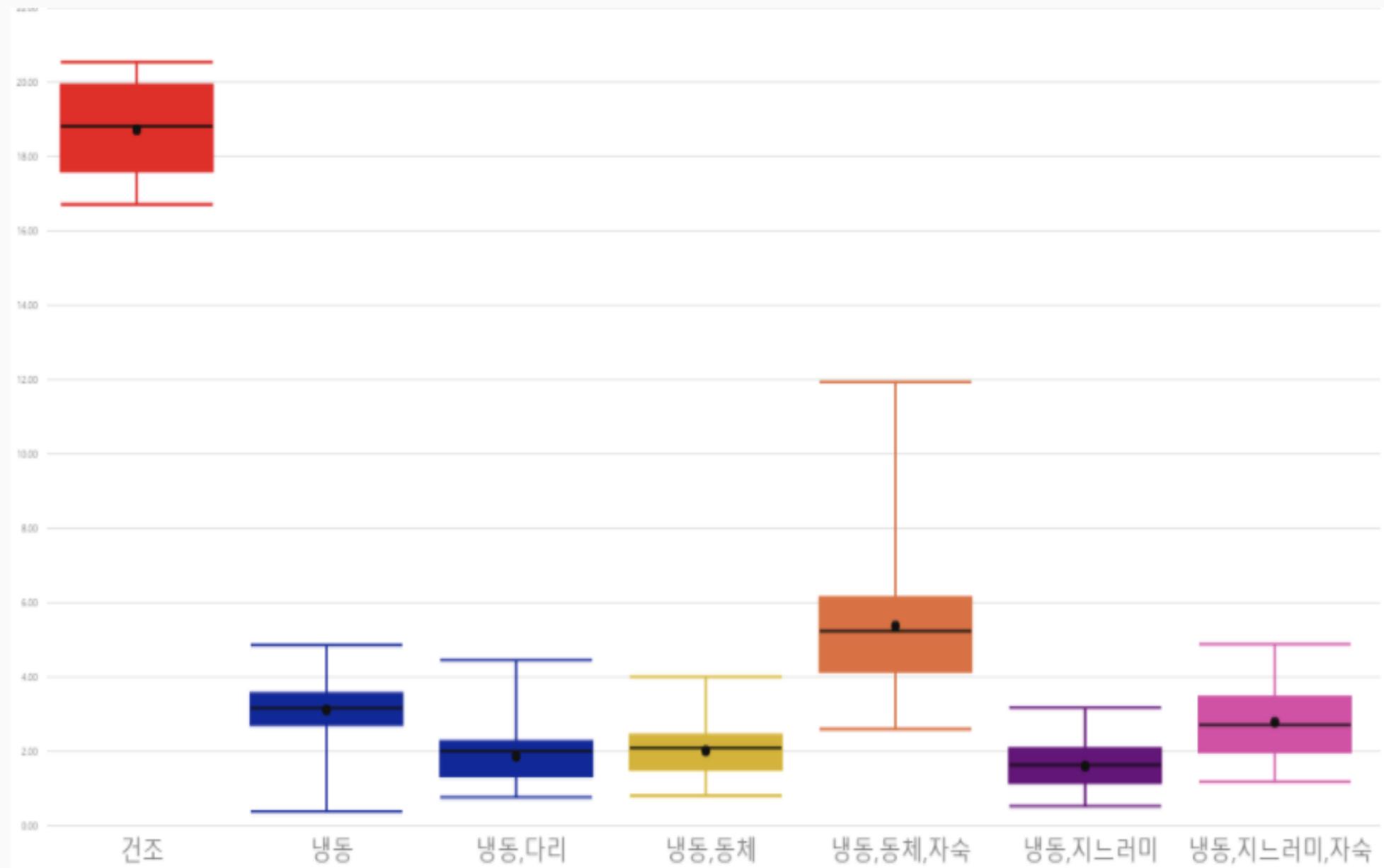


- > 외화획득용 원료와 외화획득용 제품: 수입형태가 다르므로 가격차이가 심함
- > 판매용이 비중이 커 평균단가를 움직이는 변수로 판단

오징어 데이터 분석

- 수입형태별

<오징어 수입형태별 평균단가 박스플롯>

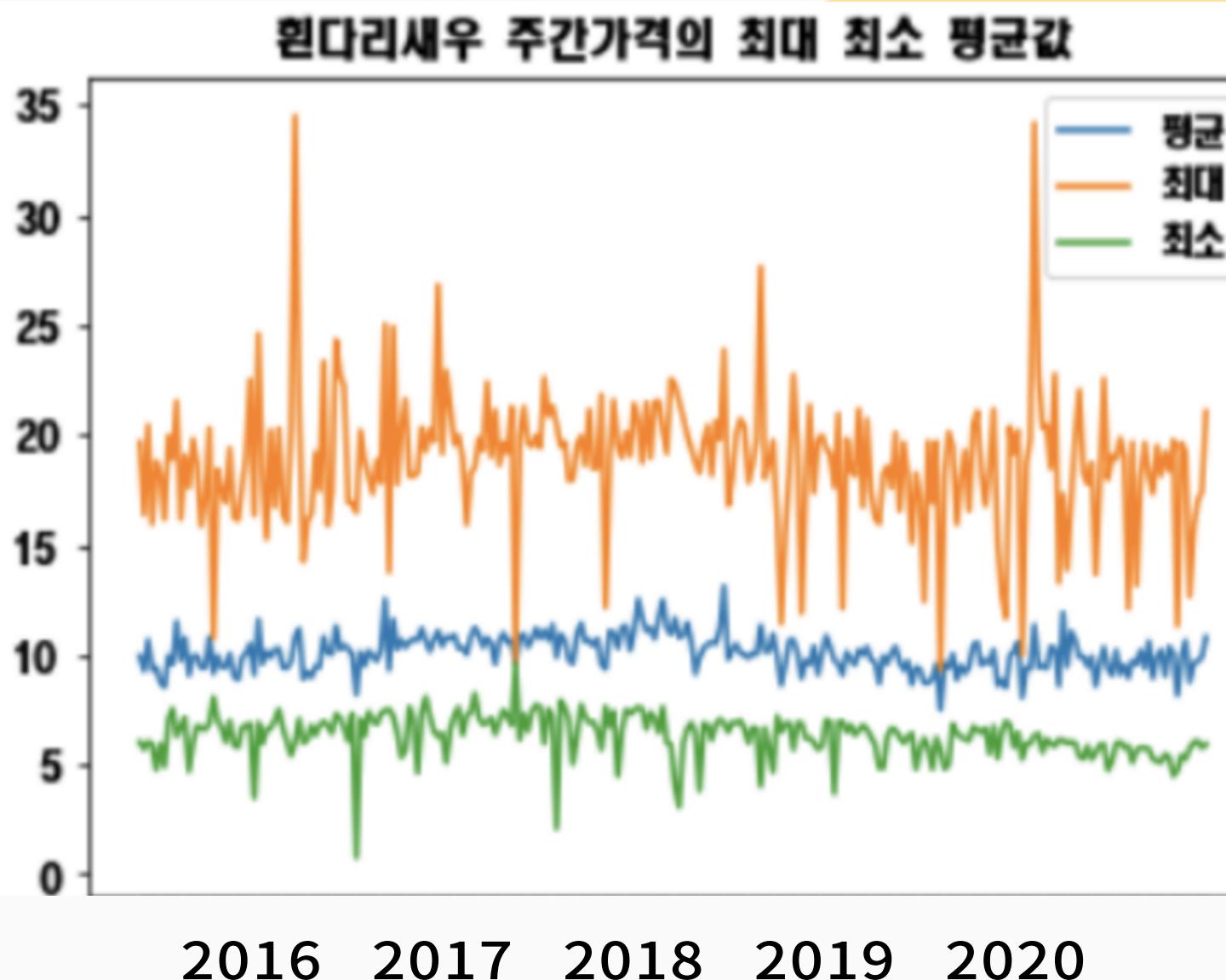


1. 건조

- 다른 품목에 비해 가격이 월등히 높아 주별 평균단가를 폭등시키는 원인
- 그러나 거래 횟수는 총 5건에 불과



undai새우 데이터 분석



주별 평균가격 : \$ 10.08

주별 평균가격의 최댓값 : \$ 13.1

주별 평균가격의 최솟값 : \$ 7.49

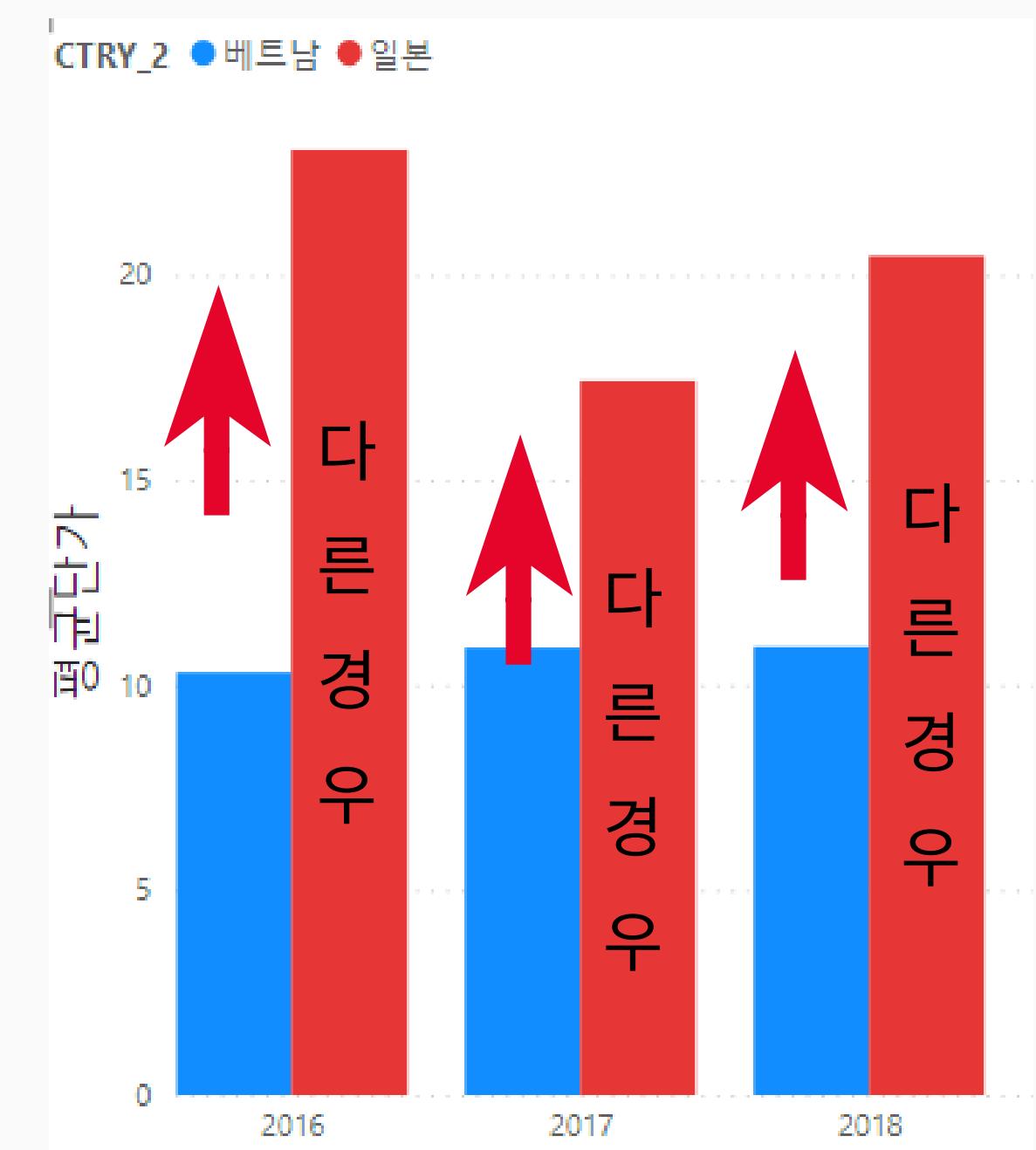
분석 결과

1. 안정적인 시계열 흐름을 보임
2. 이벤트가 발생하여 이상치들이 발견
3. 일정한 값이 지속되지 못하고 분산이 반복되는 현상 발생
4. 평균값을 안정적으로 예측할 수 있는 모델을 구축해야함

흰다리새우 데이터 분석

- 제조국과 수출국이 다른경우

<제조국과 수출국이 다를 때 가격차이>



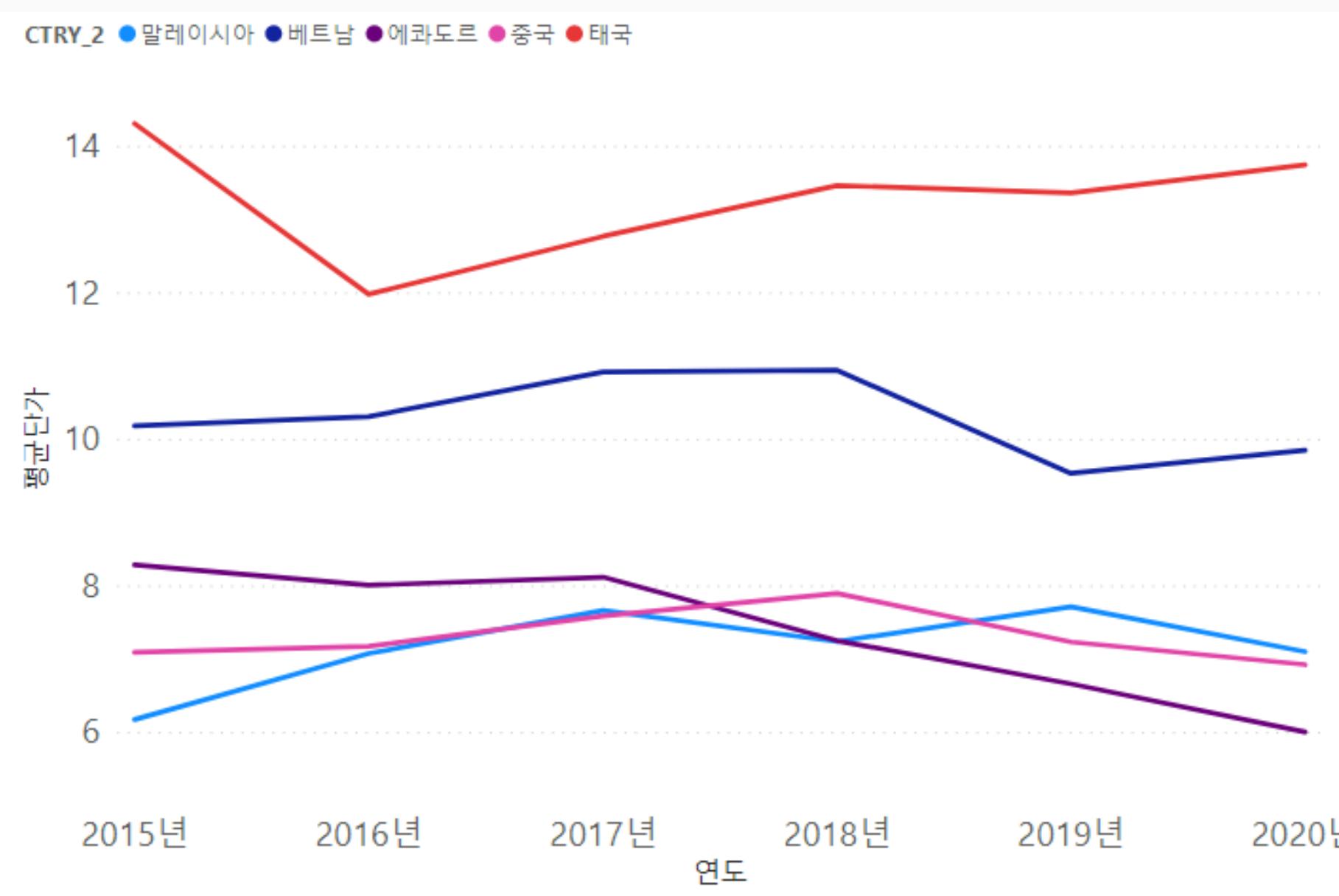
Database table showing 5 rows of data for white shrimp. The last row is circled in orange with the text "31333건 중 5건" overlaid.

REG_DATE	P_TYPE	CTRY_1	CTRY_2	P_PURPOSE	CATEGORY_1	CATEGORY_2	P_NAME	P_IMPORT_TYPE	P_PRICE
2016-02-29	수산물	베트남	일본	판매용	갑각류		새우 흰다리새우	냉동,살,자숙,포장횟감	21.52
2016-07-18	수산물	베트남	일본	판매용	갑각류		새우 흰다리새우	냉동,살,자숙,포장횟감	24.58
2017-08-21	수산물	베트남	일본	판매용	갑각류		새우 흰다리새우	냉동,살,자숙,포장횟감	18.10
2017-11-20	수산물	베트남	일본	판매용	갑각류		새우 흰다리새우	냉동,살,자숙,포장횟감	16.75
2018-07-02	수산물	베트남	일본	판매용	갑각류		새우 흰다리새우	냉동,살,자숙,포장횟감	20.47

흰다리새우 데이터 분석

- 수출국별

<흰다리새우 수출국별 평균단가>



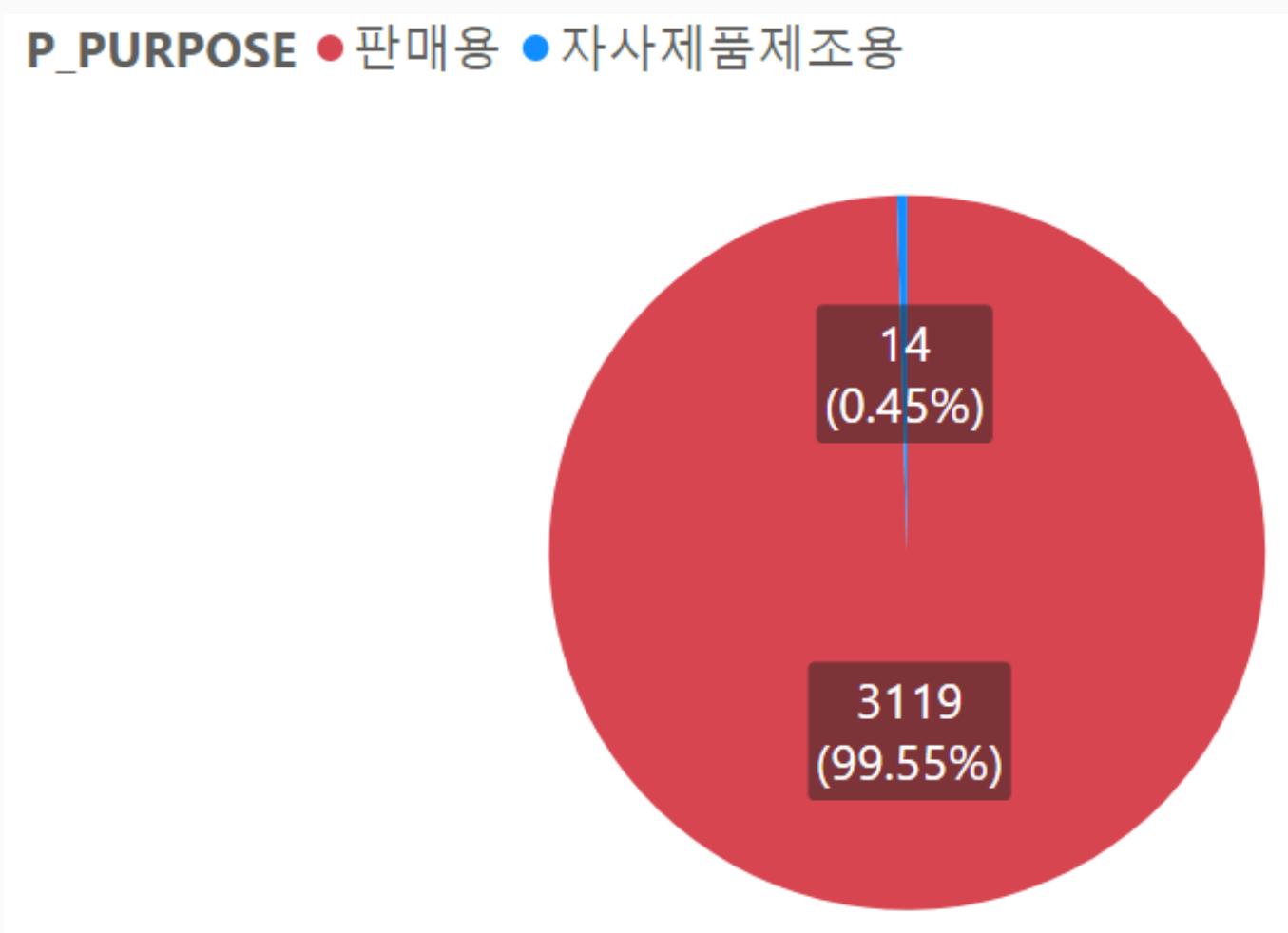
- 베트남: 1214건
- 태국: 883건
- 에콰도르: 250건
- 말레이시아: 233건
- 중국: 183건

> 품목이 많은 태국이 모든 수입형태에서 타국에 비해 가격이 높음

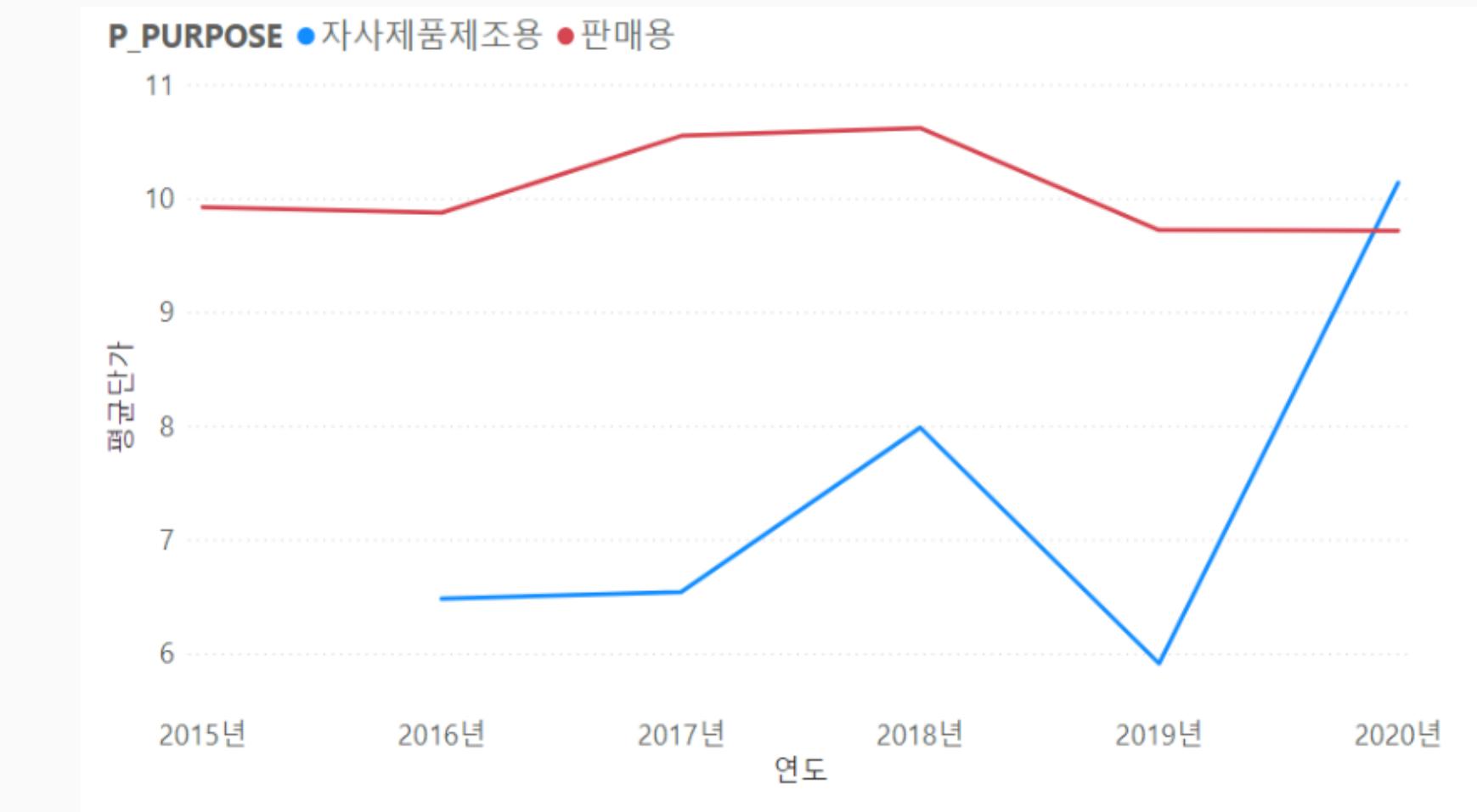


- 수입용도별

<흰다리새우 수입용도별 비율>



<흰다리새우 수출용도별 평균단가>



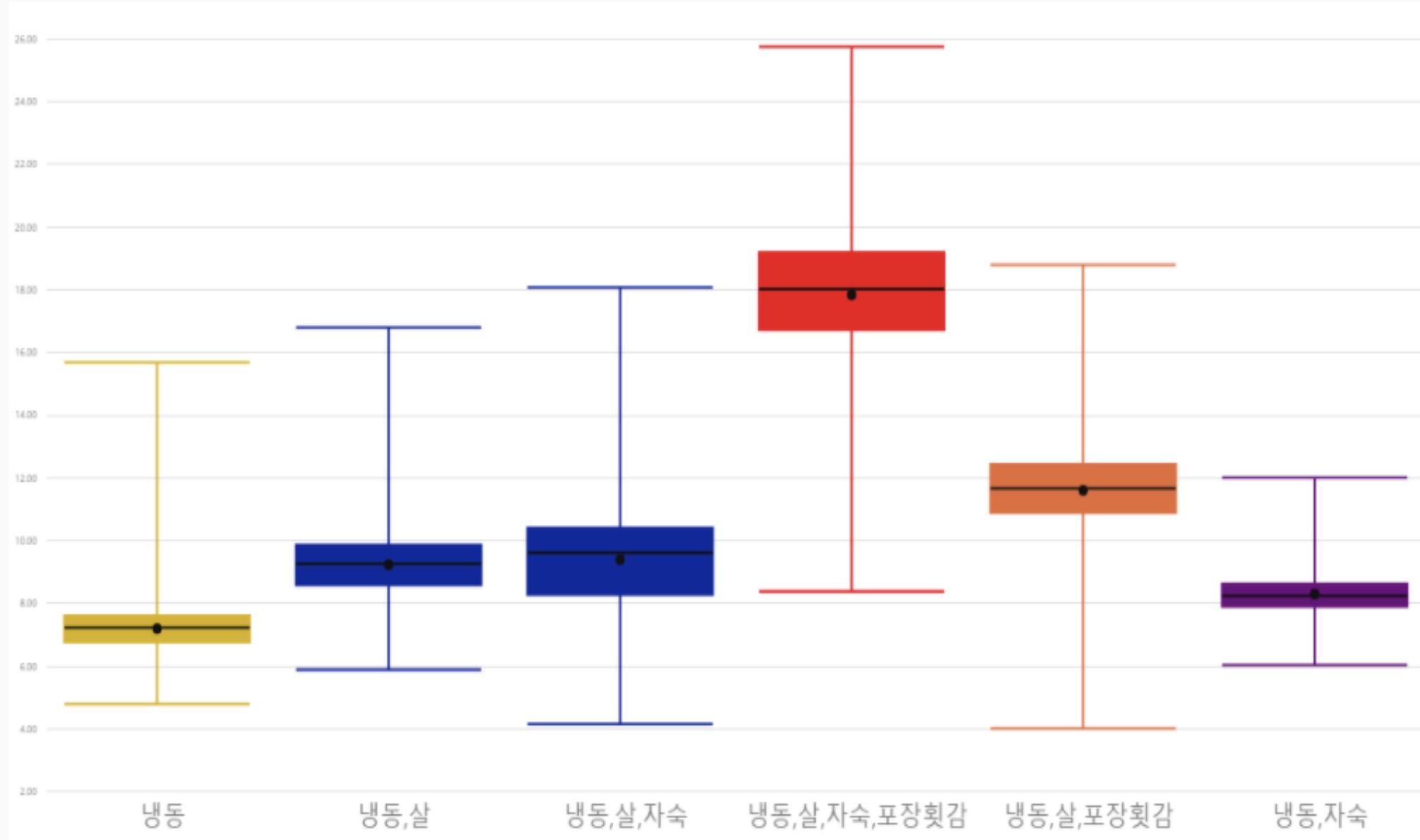
> 자사제품제조용의 비중이 작아 판매용에 집중하여 분석하기로 판단



힌다리새우 데이터 분석

- 수입형태별

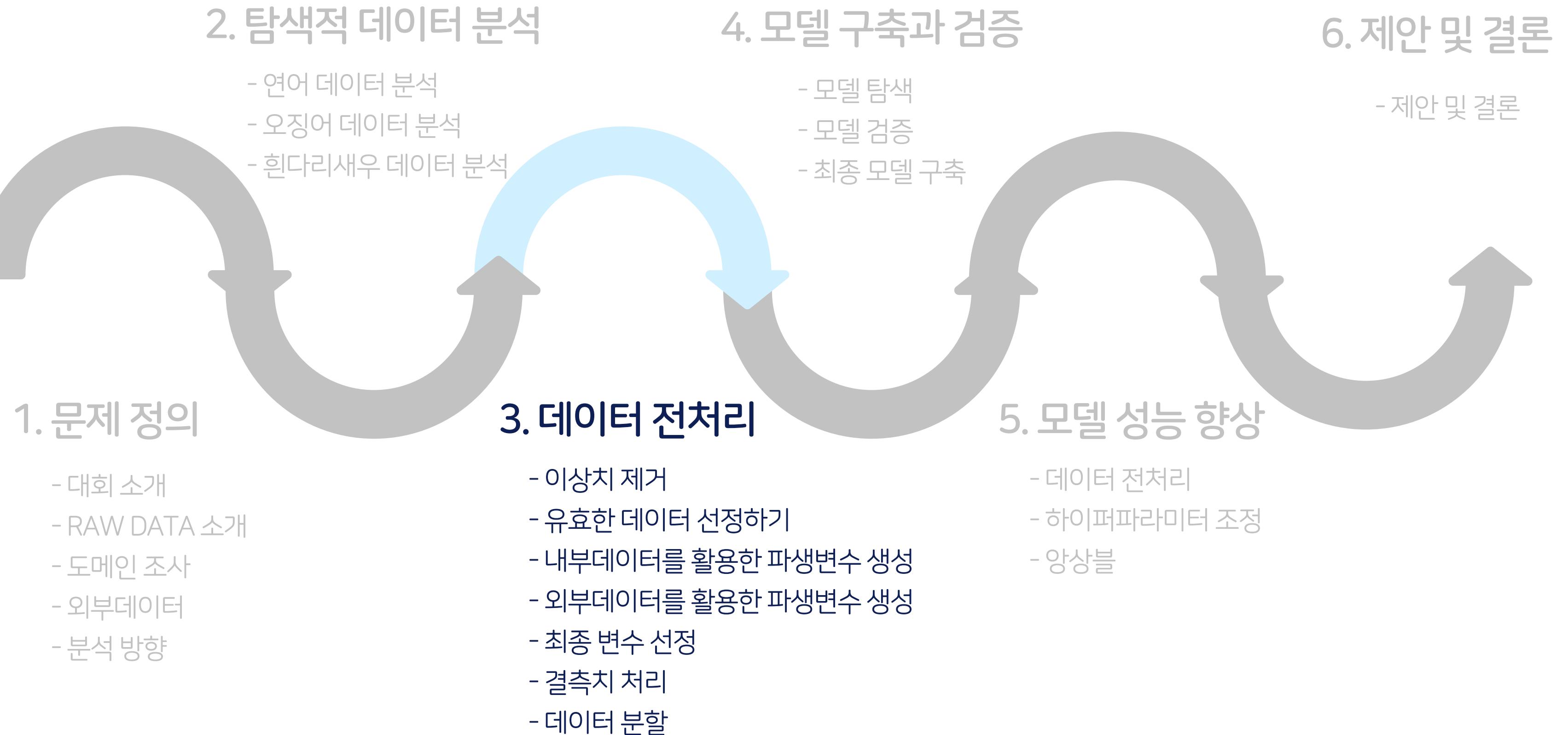
<힌다리새우 수입형태별 평균단가 박스플롯>



1. 냉동,살,자숙,포장횟감

- 다른 품목에 비해 가격이 월등히 높아
주별 평균단가를 폭등시키는 원인

목 차



이상치 예시 - 연어

<연어 냉장,필렛(F) 주별 평균단가>



2017년 가격 폭등: 이상치로 판단 → 안정적인 예측 모델을 위해 평균가격으로 데이터 수정

 이상치 제거 - 연어

연어 29건

1. 전 주 혹은 다음주와 비교하여 30%이상의 가격 변동이 발생한 경우
2. 전 주 혹은 다음 주의 가격 정보가 존재하지 않을 경우,
이전 회차 혹은 다음 회차의 가격 비교



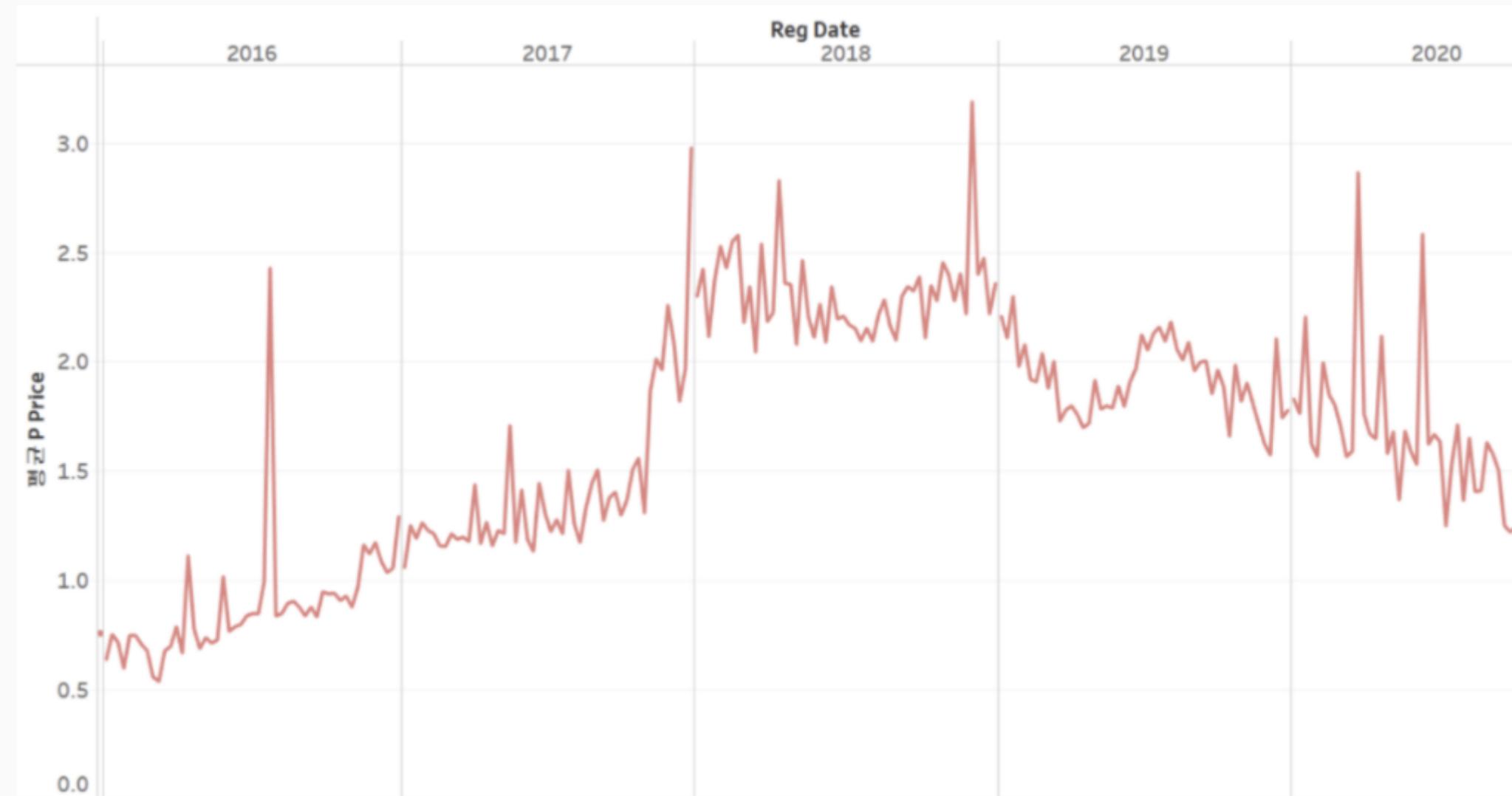
1. 전 주(혹은 이전회차)와 다음주(혹은 다음회차)의 평균가격으로 조정
2. 연어는 많은 비율을 차지하는 노르웨이의 냉장 가격에 대부분 비례



해당 데이터의 가격과 노르웨이 냉장 가격과의 가격 비율에 따라 조정

이상치 예시 - 오징어

<오징어 냉동,지느러미 주별 평균단가>



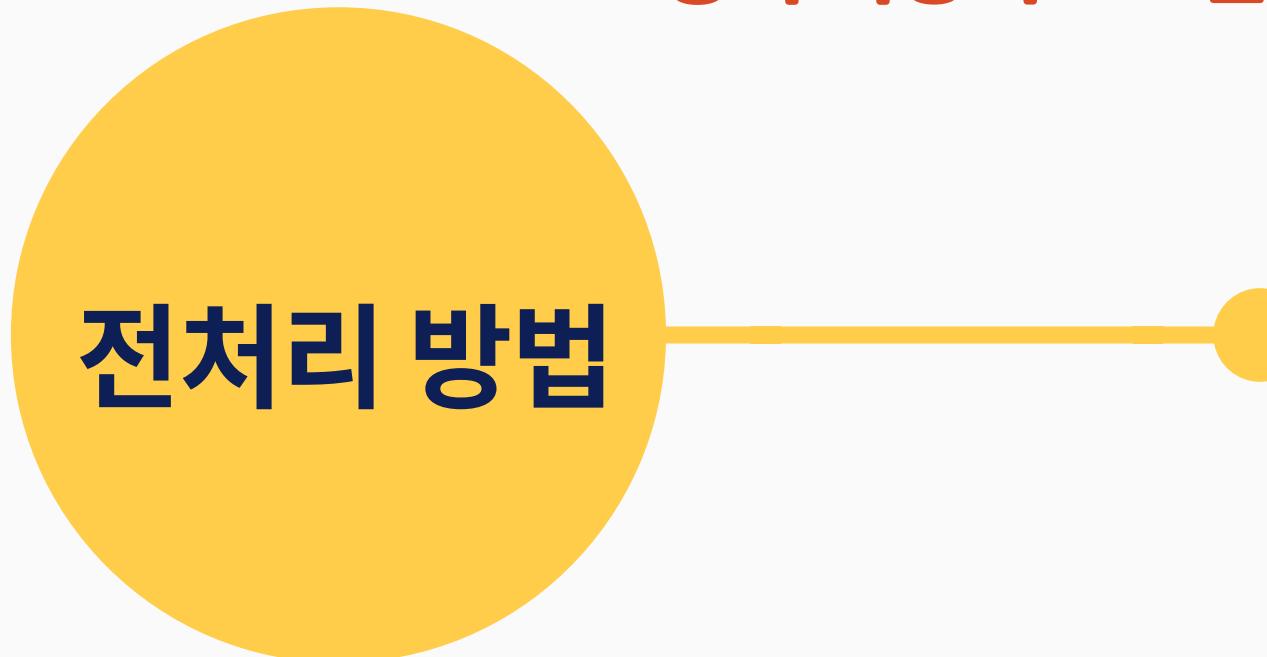
이상치가 많이 발견되는 오징어 가격의 특성상,
이상치 기준을 다른 어종에 비해 높게 잡을 필요성이 있다고 판단

 이상치 제거 - 오징어

1. 전 주와 다음주와 비교하여 두 경우 모두 100%이상의 가격 변동이 발생한 경우

2. 같은 해 같은 조건의 평균가격과 100%이상 차이 날 경우

오징어 이상치 103건

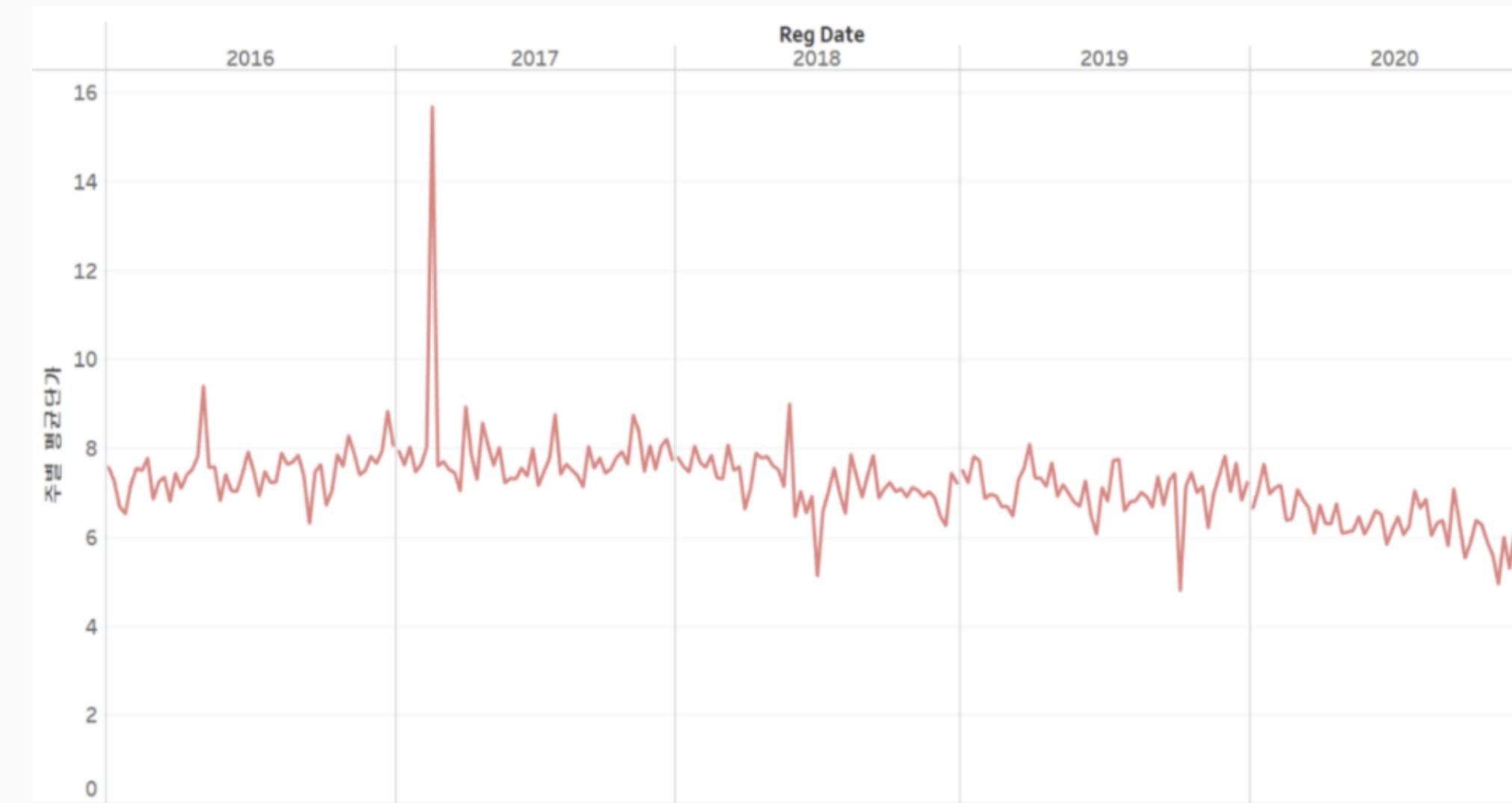


1. 전 주와 다음주의 평균가격으로 조정

2. 변수가 많은 오징어의 특성상 그 주의 다른 변수 가격에 영향을 미친 경우
→ 이상치로 보지 않음



<흰다리새우 냉동 주별 평균단가>



안정적인 시계열 모습 중에 이벤트성으로 이상치를 보여주는 흰다리새우의 특성상,
그 패턴을 파악해 볼 필요성이 있다고 판단

이상치 제거 - 흰다리새우



1. 전 주 혹은 다음주와 비교하여 50%이상의 가격 변동이 발생한 경우
2. 전 주 혹은 다음 주의 가격 정보가 존재하지 않을 경우,
이전 회차 혹은 다음 회차의 가격 비교

흰다리새우 이상치 27건

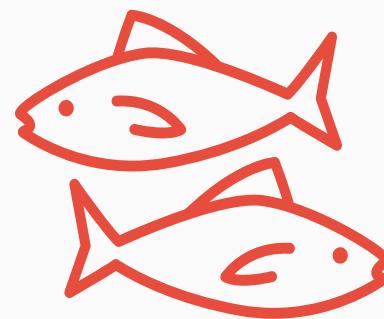


1. 전 주(혹은 이전회차)와 다음주(혹은 다음회차)의 평균가격으로 조정

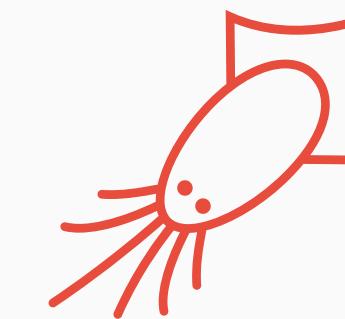


유효한 데이터 선정하기

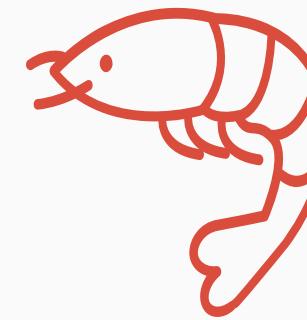
2021년의 **수입 트렌드를 반영하기 위해**
예측 모델에 필요하지 않은 데이터들을 제거하는 작업을 거침.



연어 데이터 제거: 55건



오징어 데이터 제거: 40건



흰다리새우 데이터 제거: 291건



데이터셋 생성



RAW DATA (p_name = 연어)

1895개 데이터, 10개 변수



RAW DATA (p_name = 오징어)

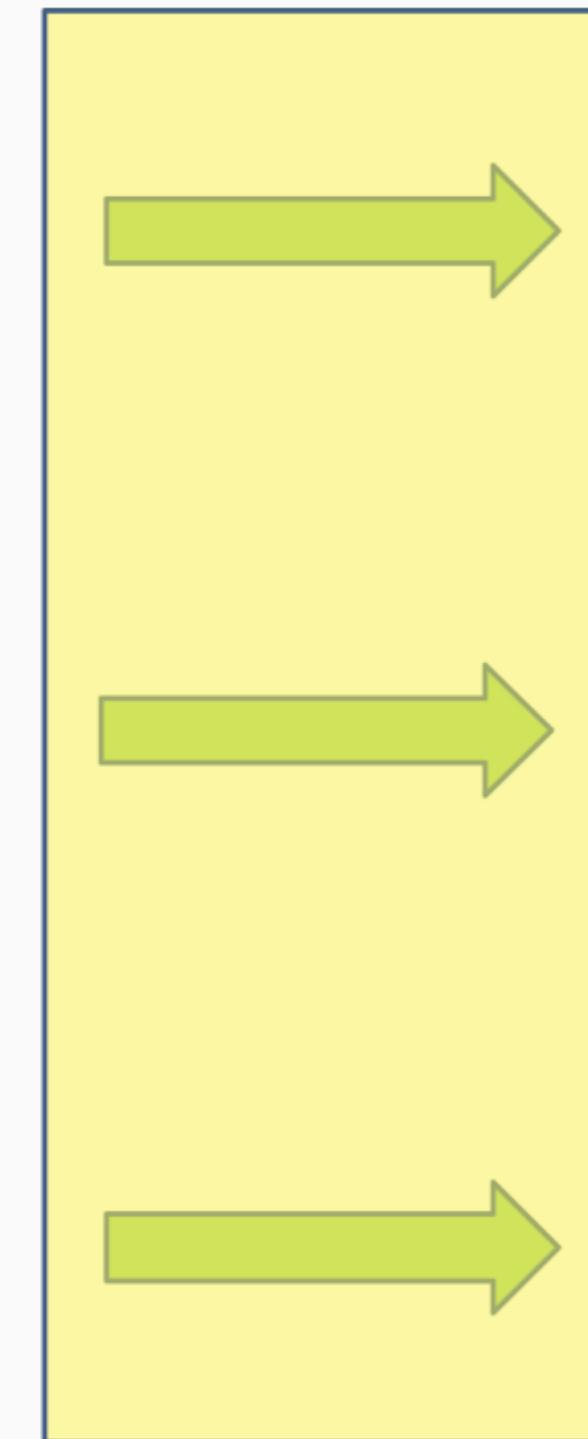
2771개 데이터, 10개 변수



RAW DATA (p_name = 흰다리새우)

3133개 데이터, 10개 변수

이상치 제거, 유효한 데이터 선정



preprocessingdata_salmon_

1840개 데이터, 10개 변수



preprocessingdata_squid

2731개 데이터, 10개 변수



preprocessingdata_shrimp

2842개 데이터, 10개 변수



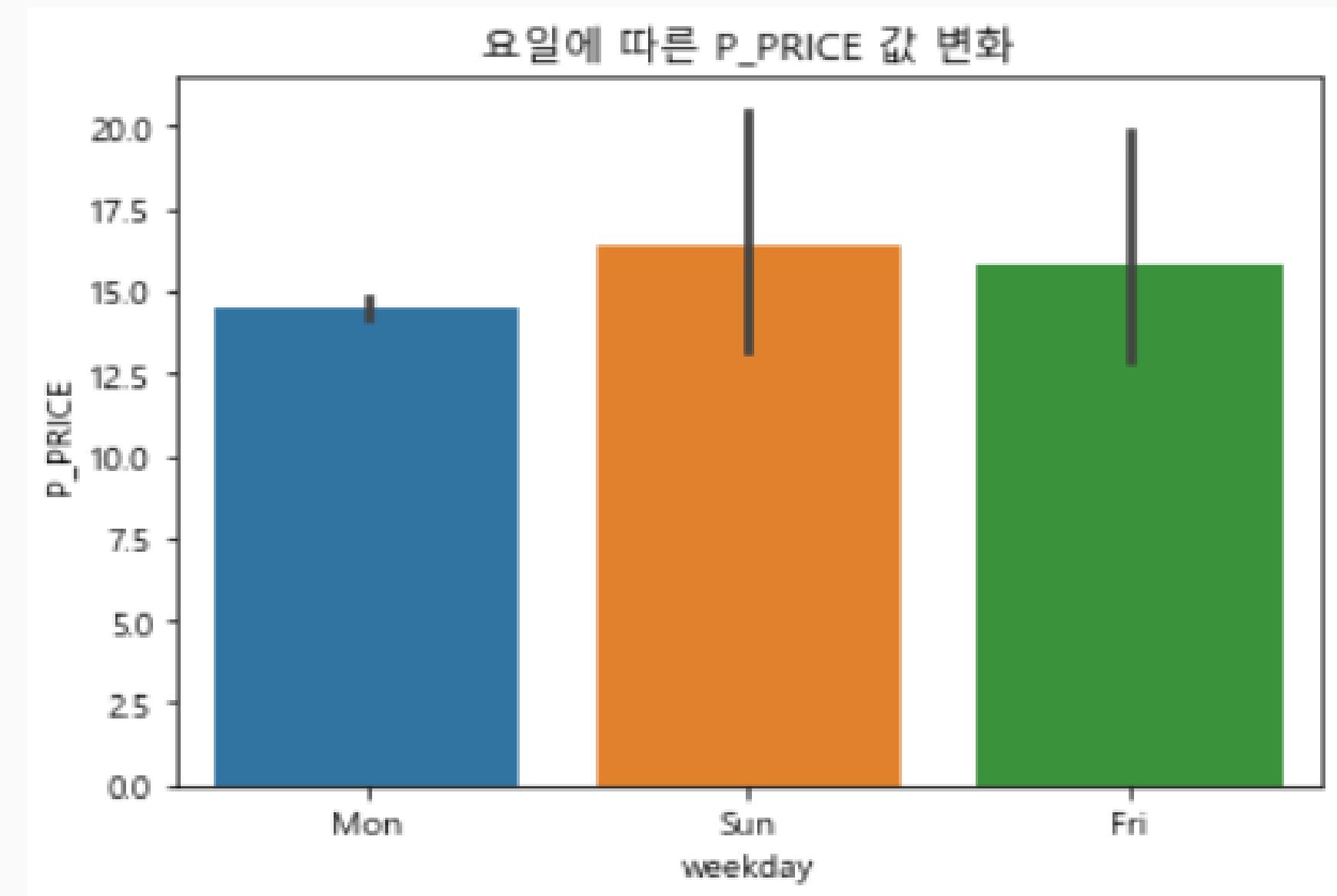
내부데이터를 활용한 파생변수 생성

- RAG_DATE 활용: 요일 변수

요일 변수

RAG_DATE

월 변수



분석결과 : 2017년 1월 1일(일) & 2017년 1월 6일 (금) 제외 모두 월요일
→ 요일 변수는 유의미 하지 않다고 판단



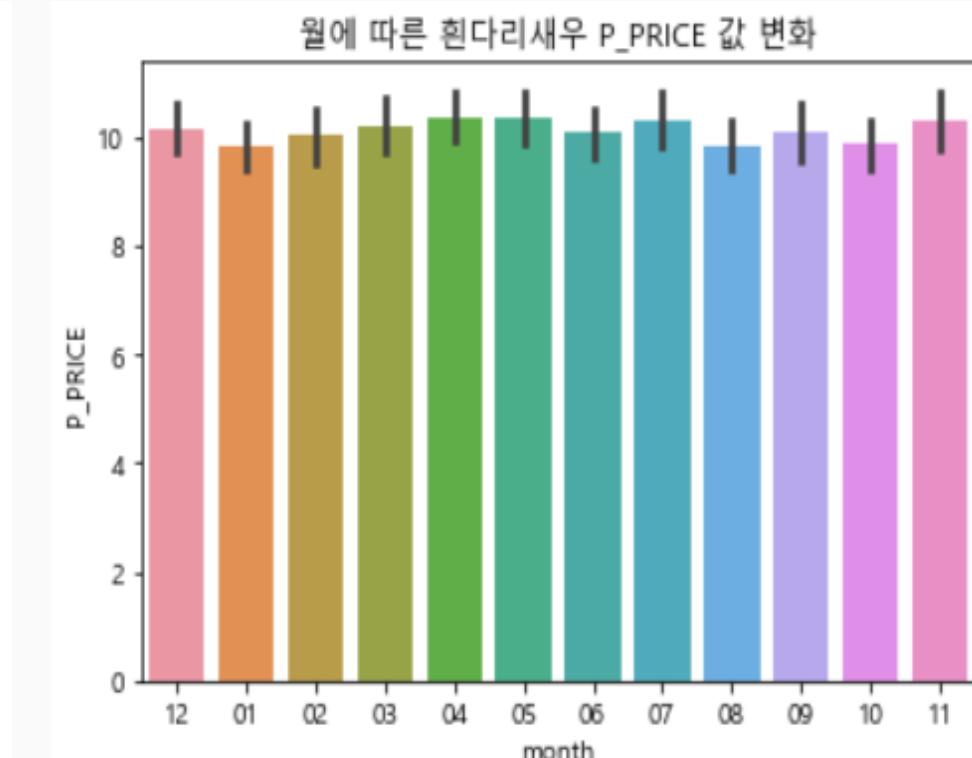
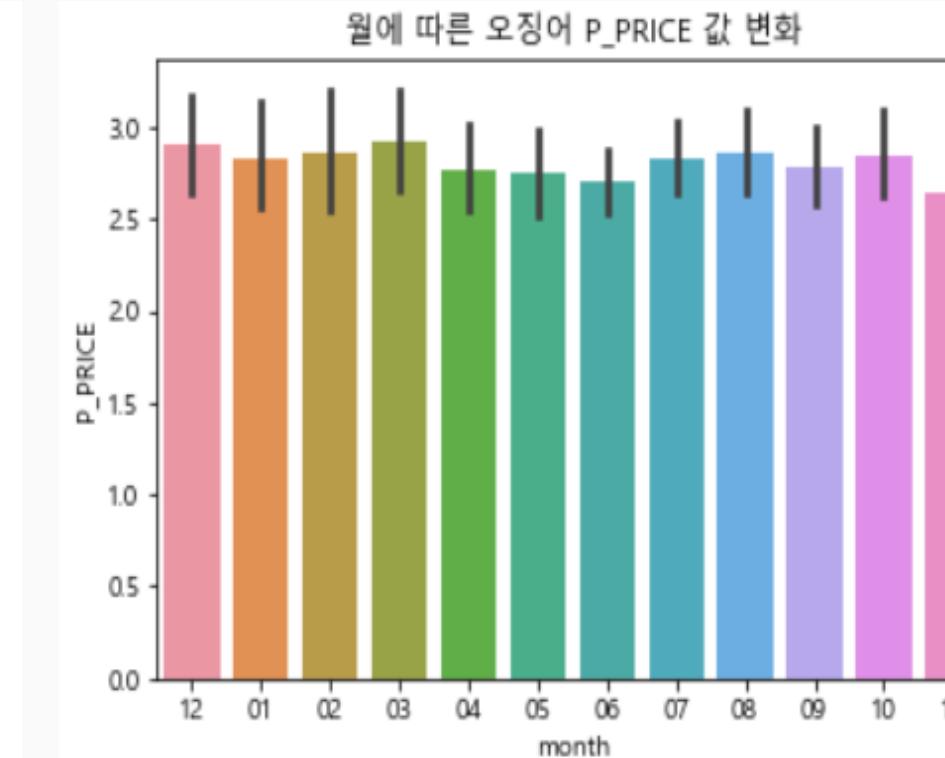
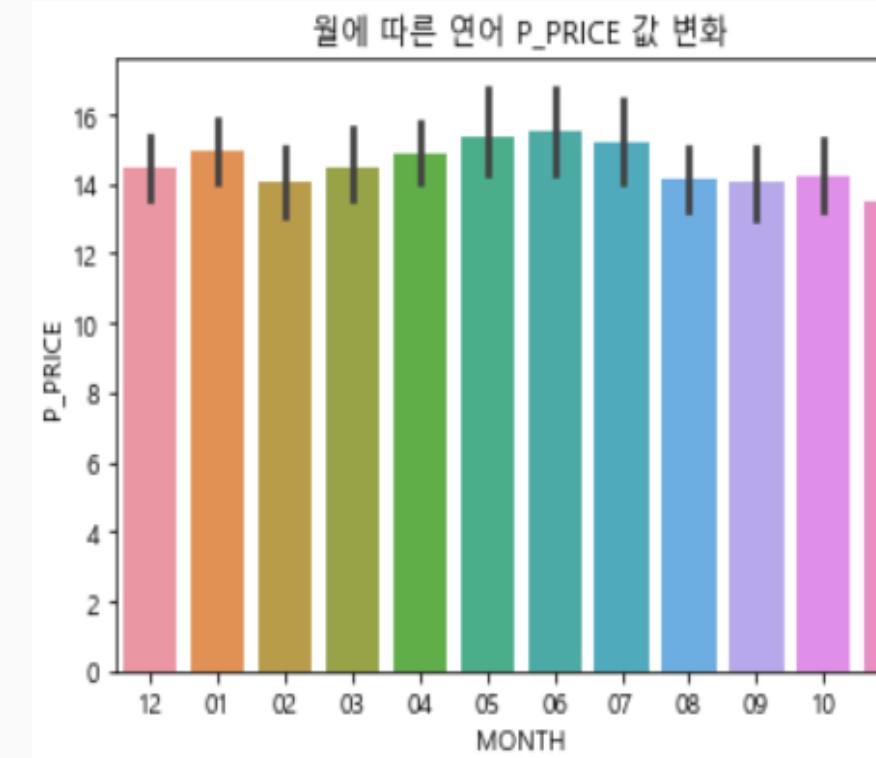
내부데이터를 활용한 파생변수 생성

- RAG_DATE 활용: 월 변수

요일 변수

RAG_DATE

월 변수

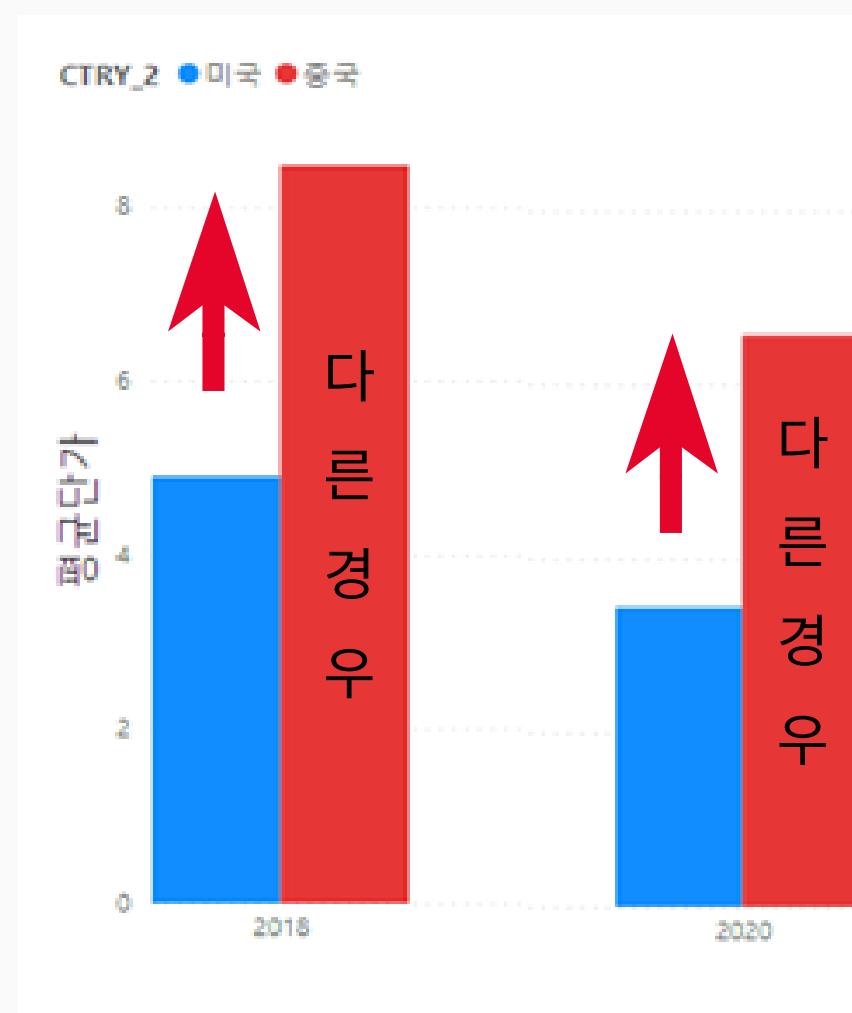


분석결과 : 월 변수는 유의미한 변수로 판단
 → MONTH, MONTH_MEAN 변수생성



내부데이터를 활용한 파생변수 생성

- CTRY_1(제조국)



<연어 제조국과 수출국이 다른 경우 가격차이> <오징어 제조국과 수출국이 다른 경우 가격차이> <흰다리새우 제조국과 수출국이 다른 경우 가격차이>

> 제조국과 수출국이 다른 경우, P_PRICE가 증가하므로 유의미한 변수로 판단 → CTRY_DIF 변수 생성



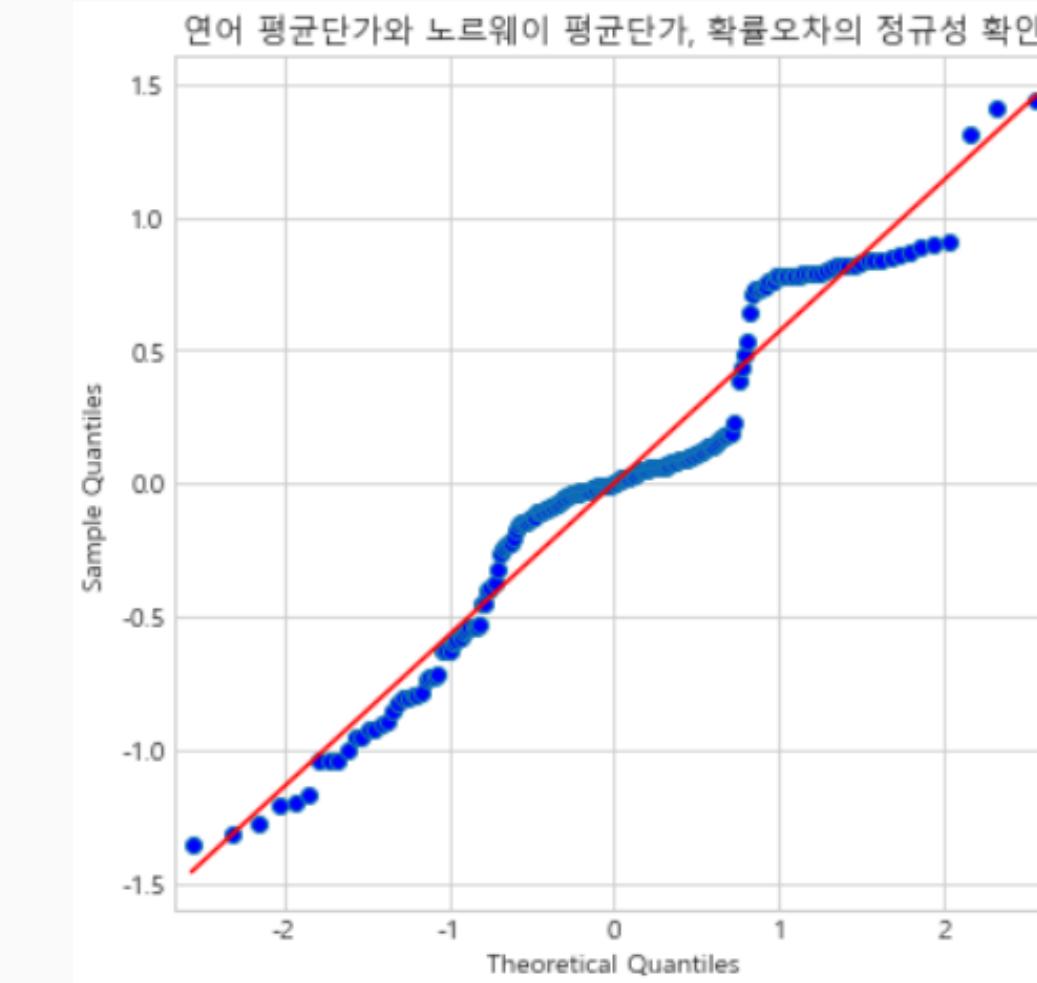
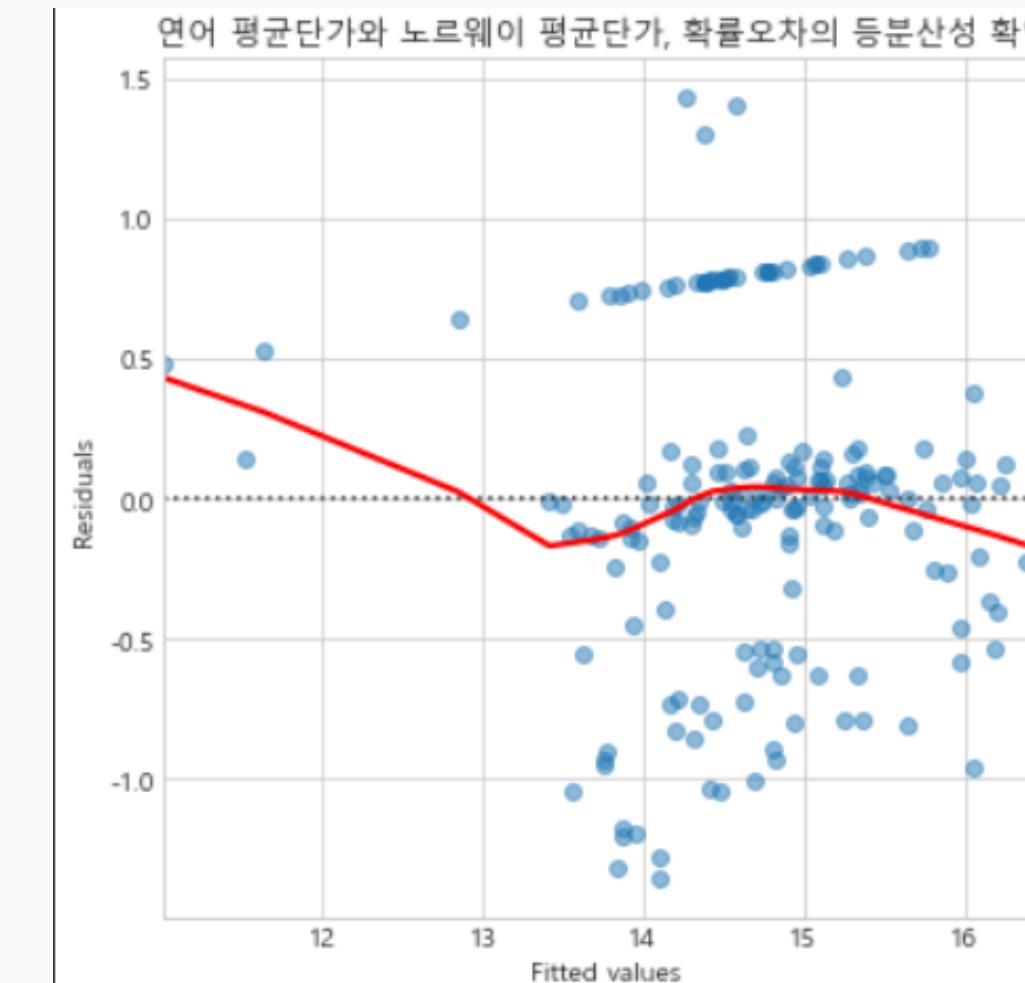
내부데이터를 활용한 파생변수 생성

- CTRY_2(수출국) 활용: 수출국별 평균단가 및 카운트

평균단가

CTRY_2

카운트



- 주별 수입된 수출국의 평균단가와 카운트를 변수들로 생성
 - 연어는 노르웨이의 수입의존성이 높음
- 노르웨이 평균단가는 연어의 주별 평균단가와 상관계수 0.812



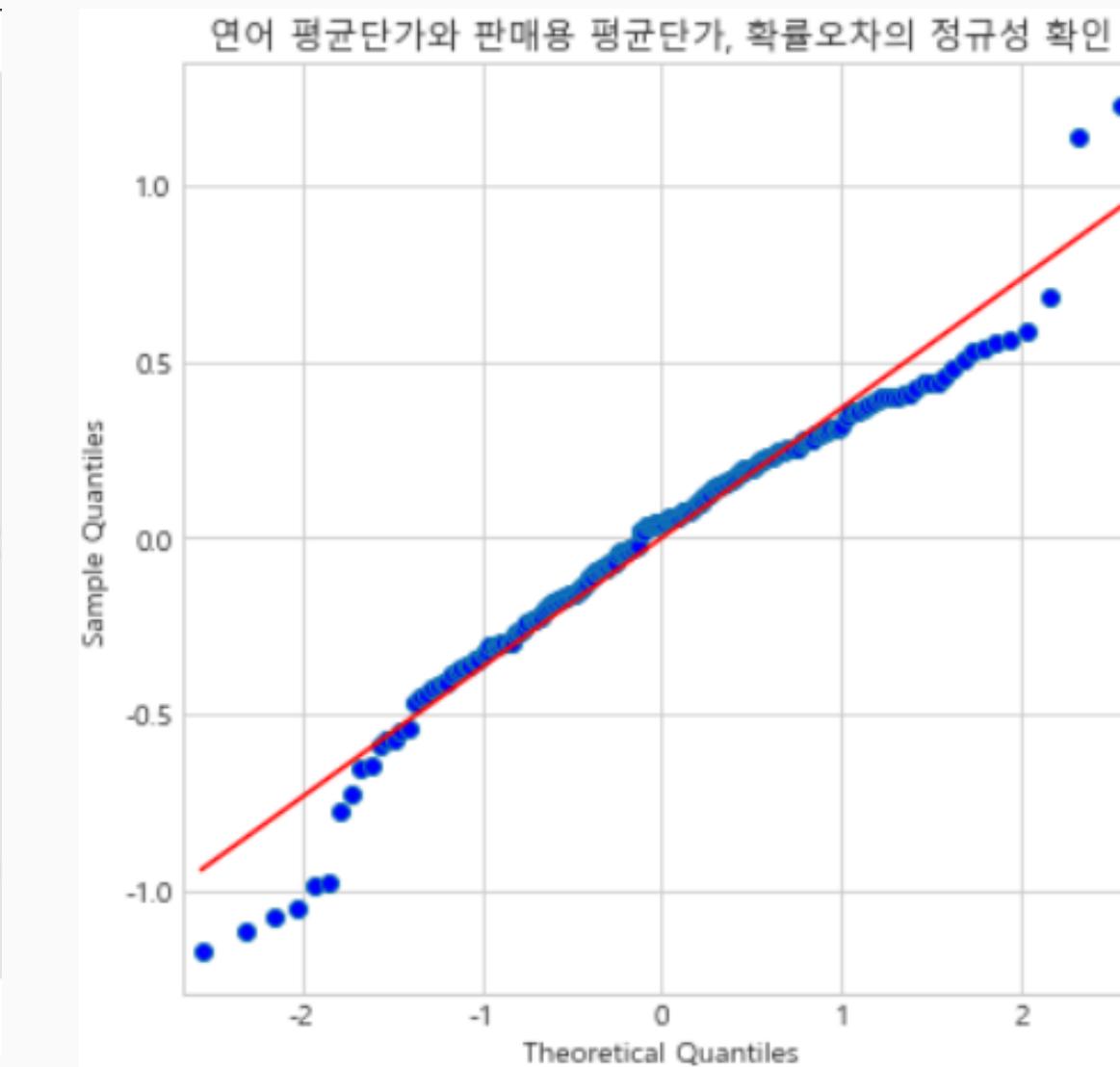
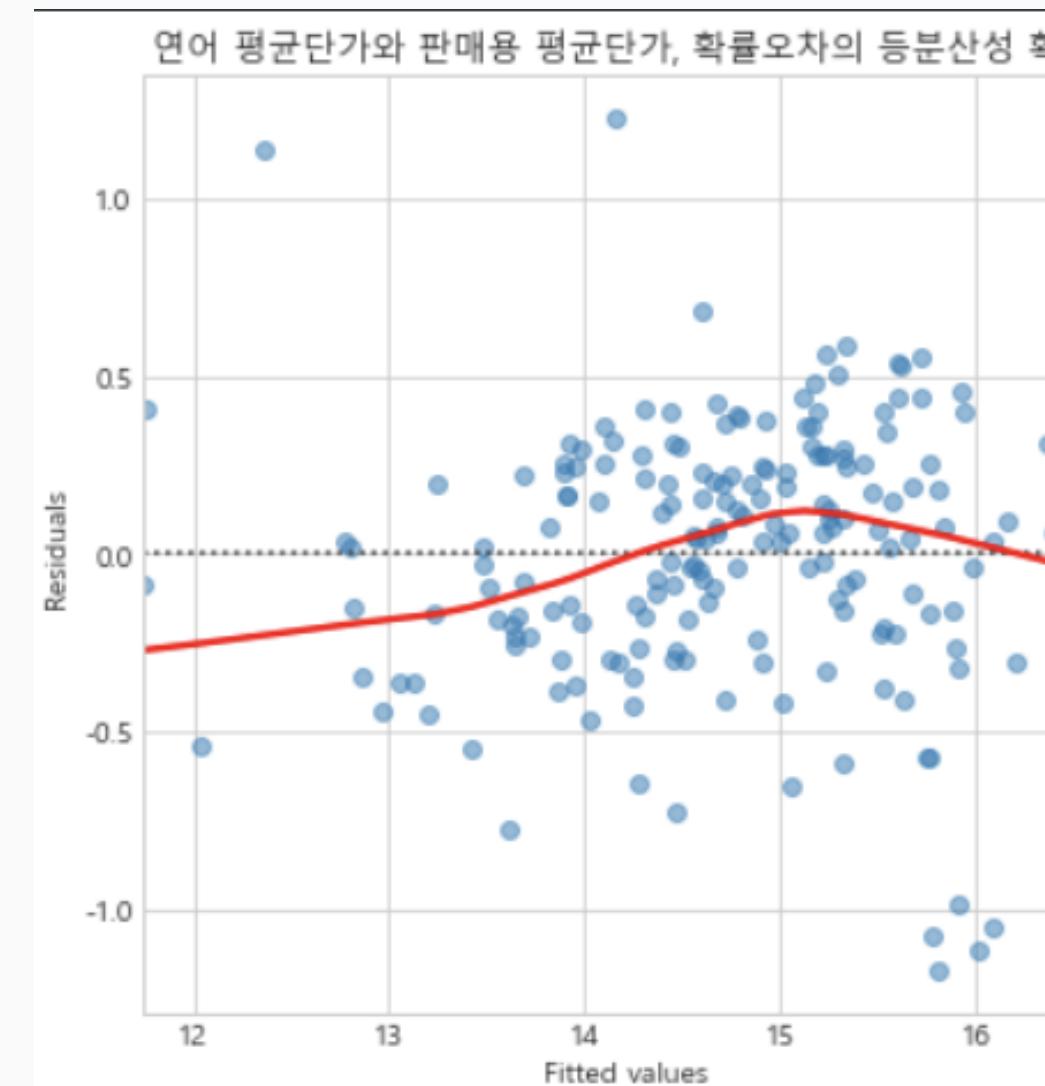
내부데이터를 활용한 파생변수 생성

- 수입용도 활용: 수입용도별 평균단가 및 카운트

평균단가

수입용도

카운트



- 주별 수입된 수입용도의 평균단가와 카운트를 변수들로 생성
- 주별 연어의 평균단가는 판매용 평균단가와 상관계수 0.92



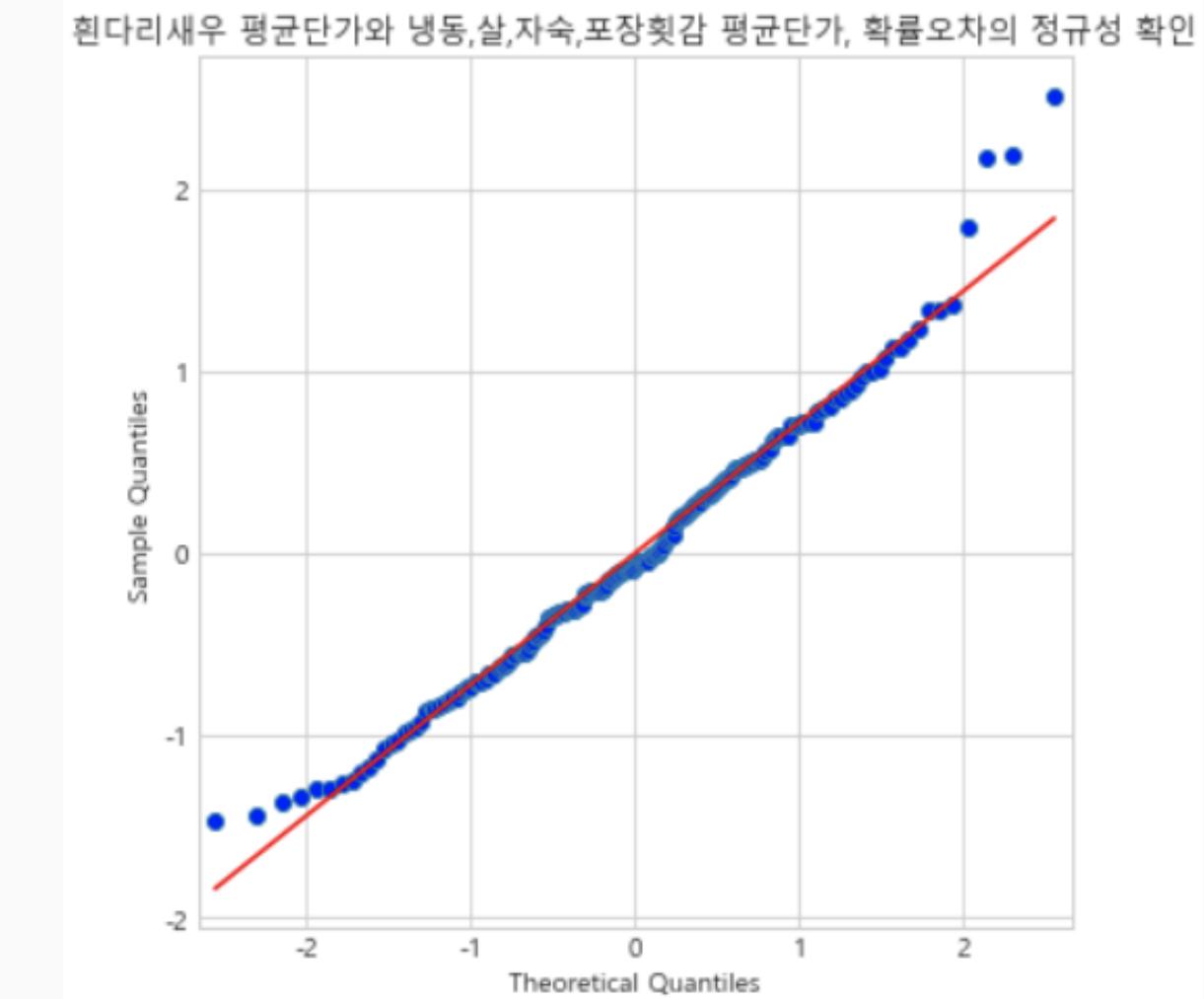
내부데이터를 활용한 파생변수 생성

- 수입형태 활용: 수입형태별 평균단가 및 카운트

평균단가

수입형태

카운트



- 주별 수입된 수입형태의 평균단가와 카운트를 변수들로 생성
- 주별 흰다리새우의 평균단가는 냉동,살,자숙,포장횟감 평균단가와 상관계수 0.52



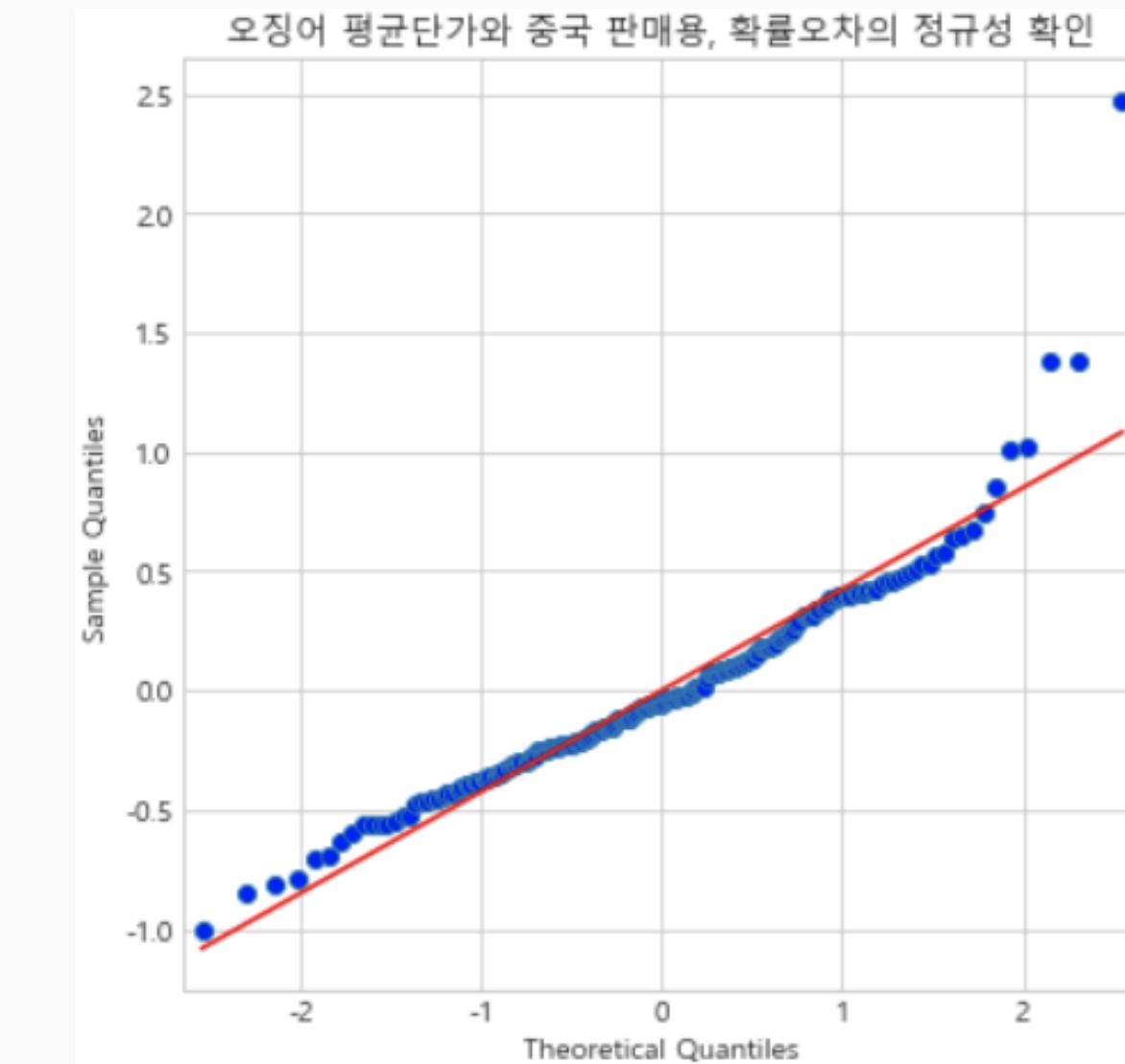
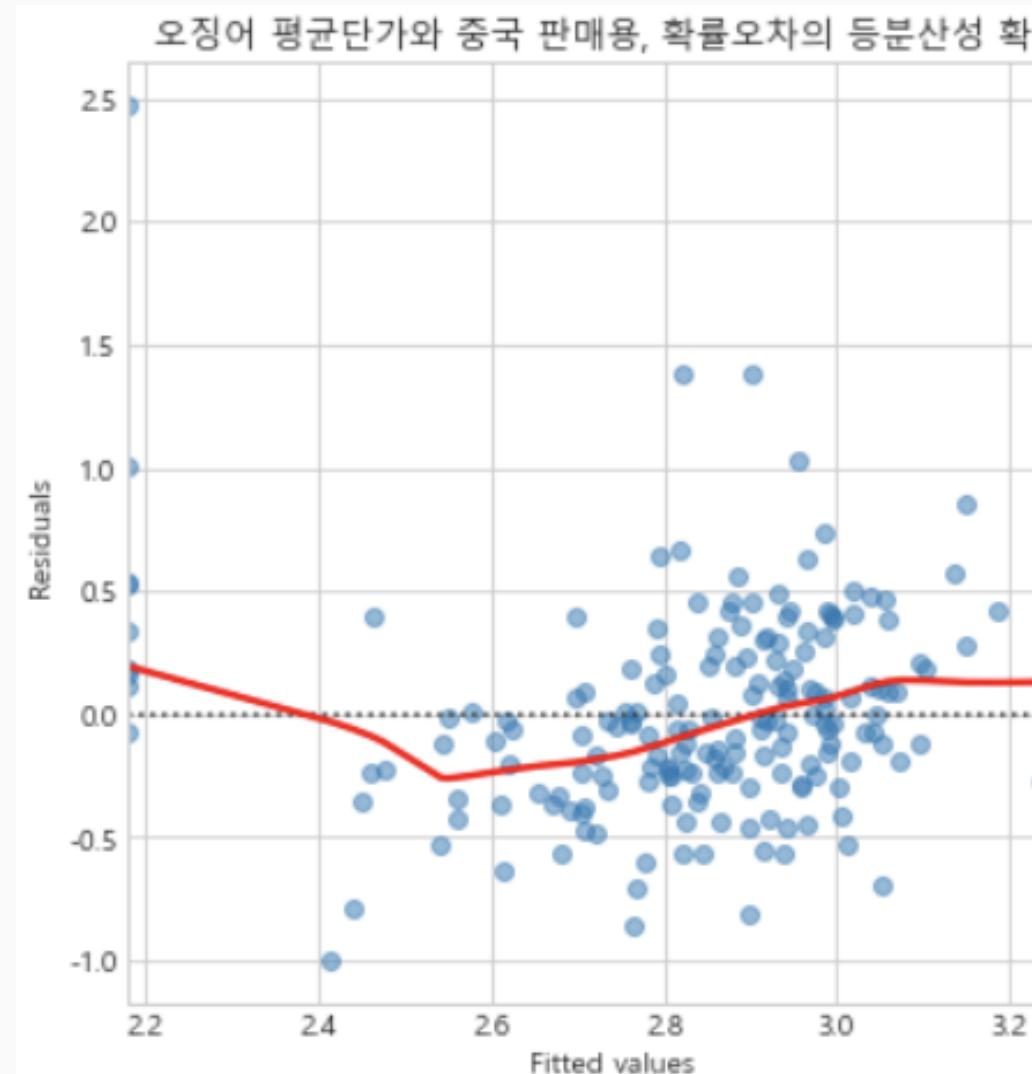
내부데이터를 활용한 파생변수 생성

- 기본변수 조합(ex. 국가 + 수입형태 평균단가 및 카운트)

평균단가

수입형태

카운트

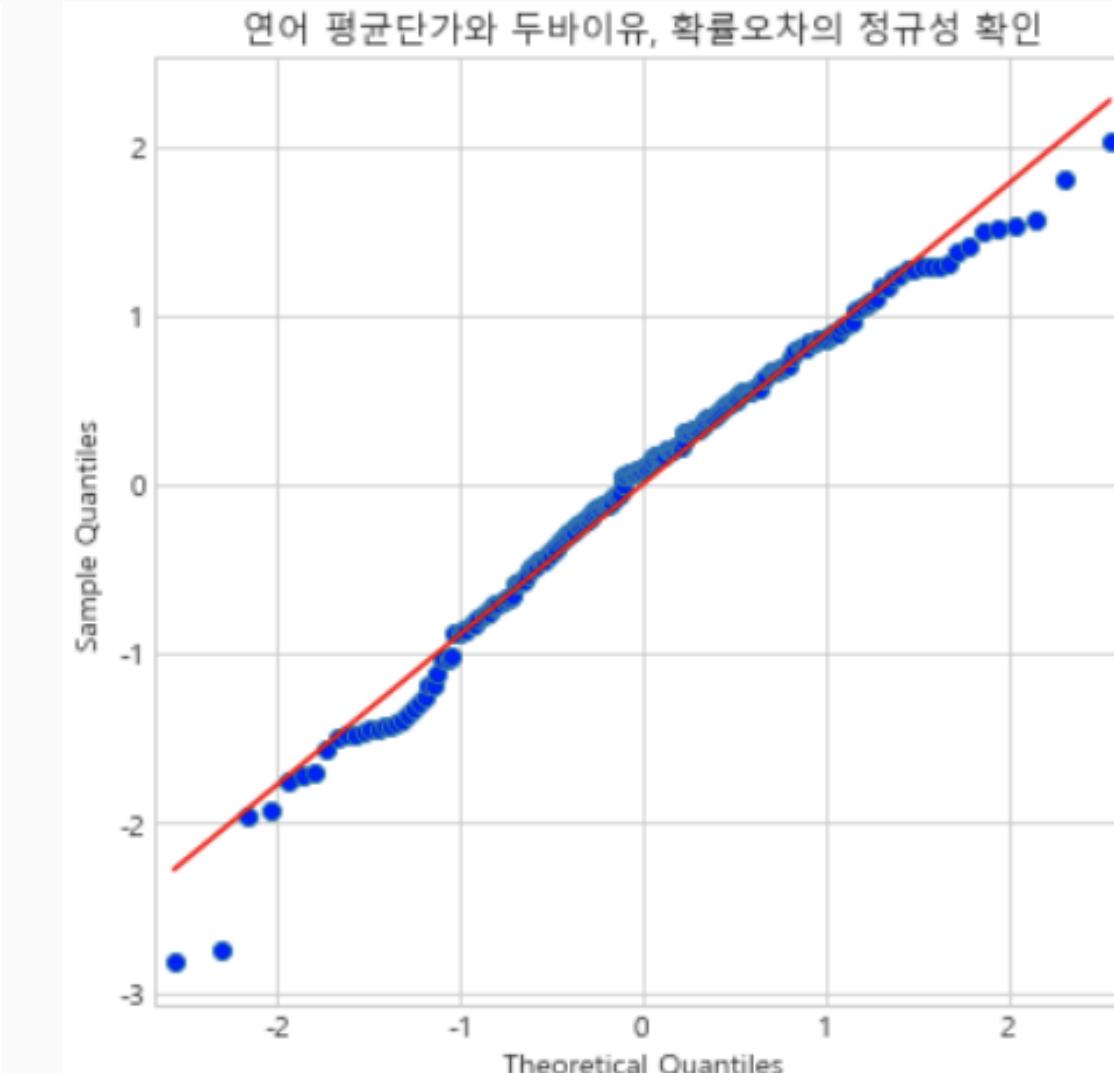
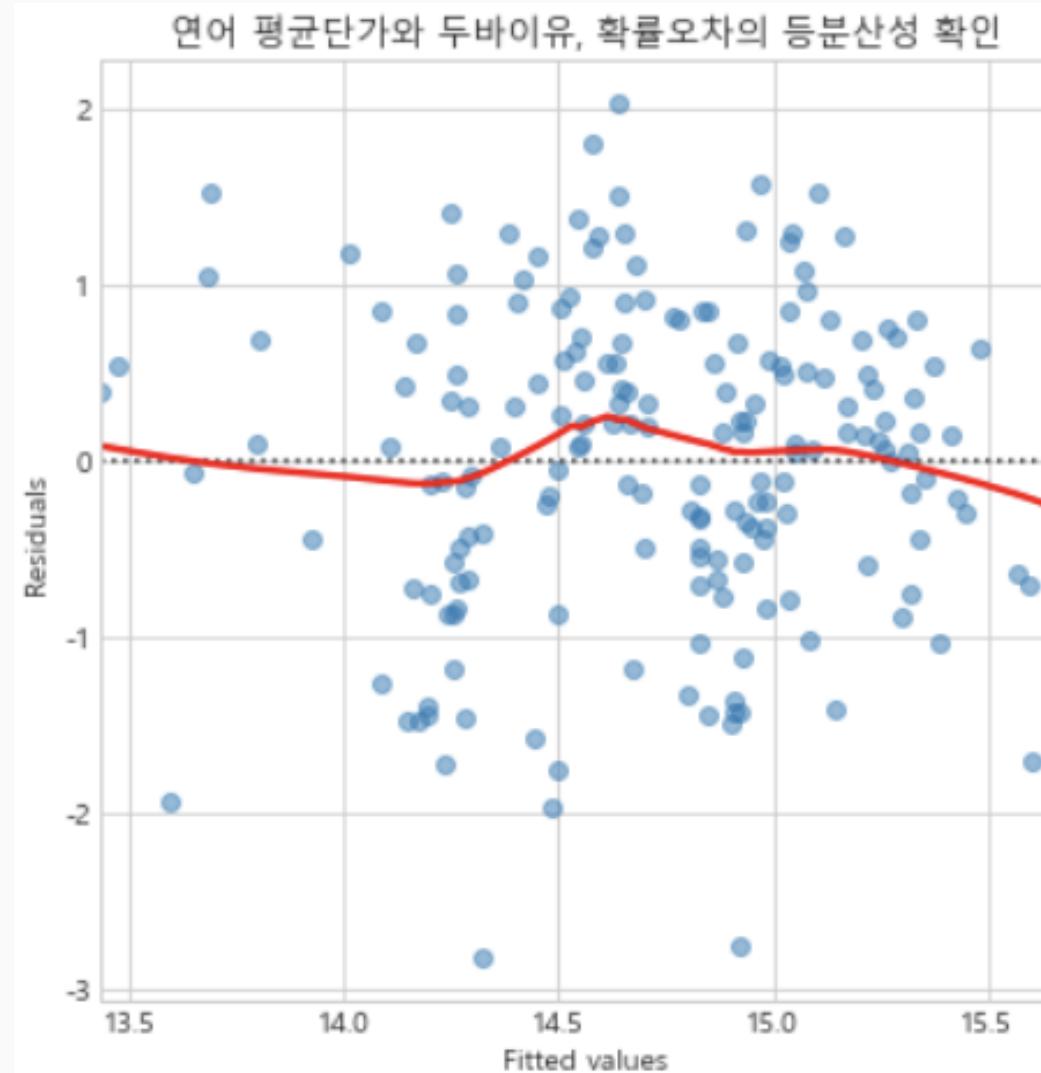
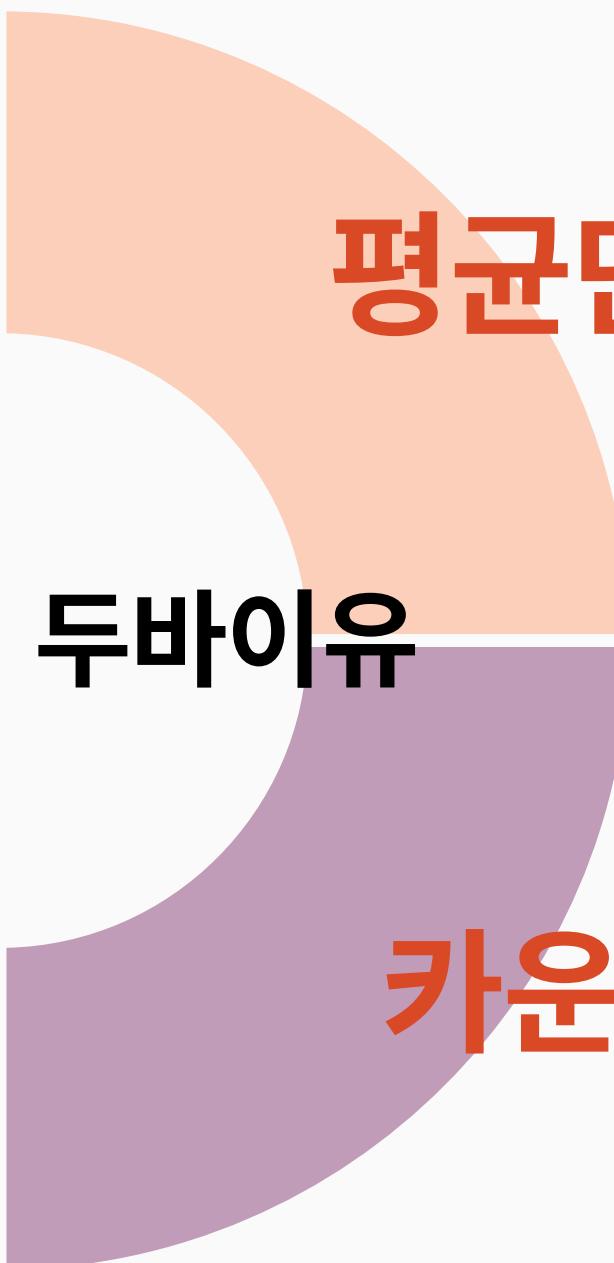


- 주별 수입된 국가별 수입용도의 평균단가와 카운트를 변수들로 생성
- 주별 오징어의 평균단가는 중국 판매용 평균단가와 상관계수 0.48



외부데이터를 활용한 파생변수 생성

- 원 달러 환율, 두바이유

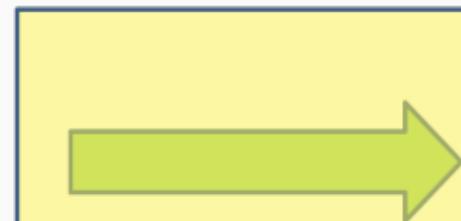


- 외부데이터를 활용해 변수들 생성
- 주별 연어의 평균단가는 두바이유와 상관계수 0.43

 데이터셋 생성

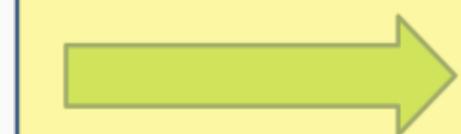
이상치 제거, 유효한 데이터 선정

RAW DATA (p_name = 연어)
1895개 데이터, 10개 변수



RAW DATA (p_name = 오징어)
2771개 데이터, 10개 변수

preprocessingdata_salmon_
1840개 데이터, 10개 변수



RAW DATA (p_name = 흰다리새우)
3133개 데이터, 10개 변수

preprocessingdata_squid
2731개 데이터, 10개 변수



변수 생성, 결측치 처리 (결측치는 0으로 처리)

salmon_initial
235개 데이터, 120개 변수



squid_initial
234개 데이터, 154개 변수



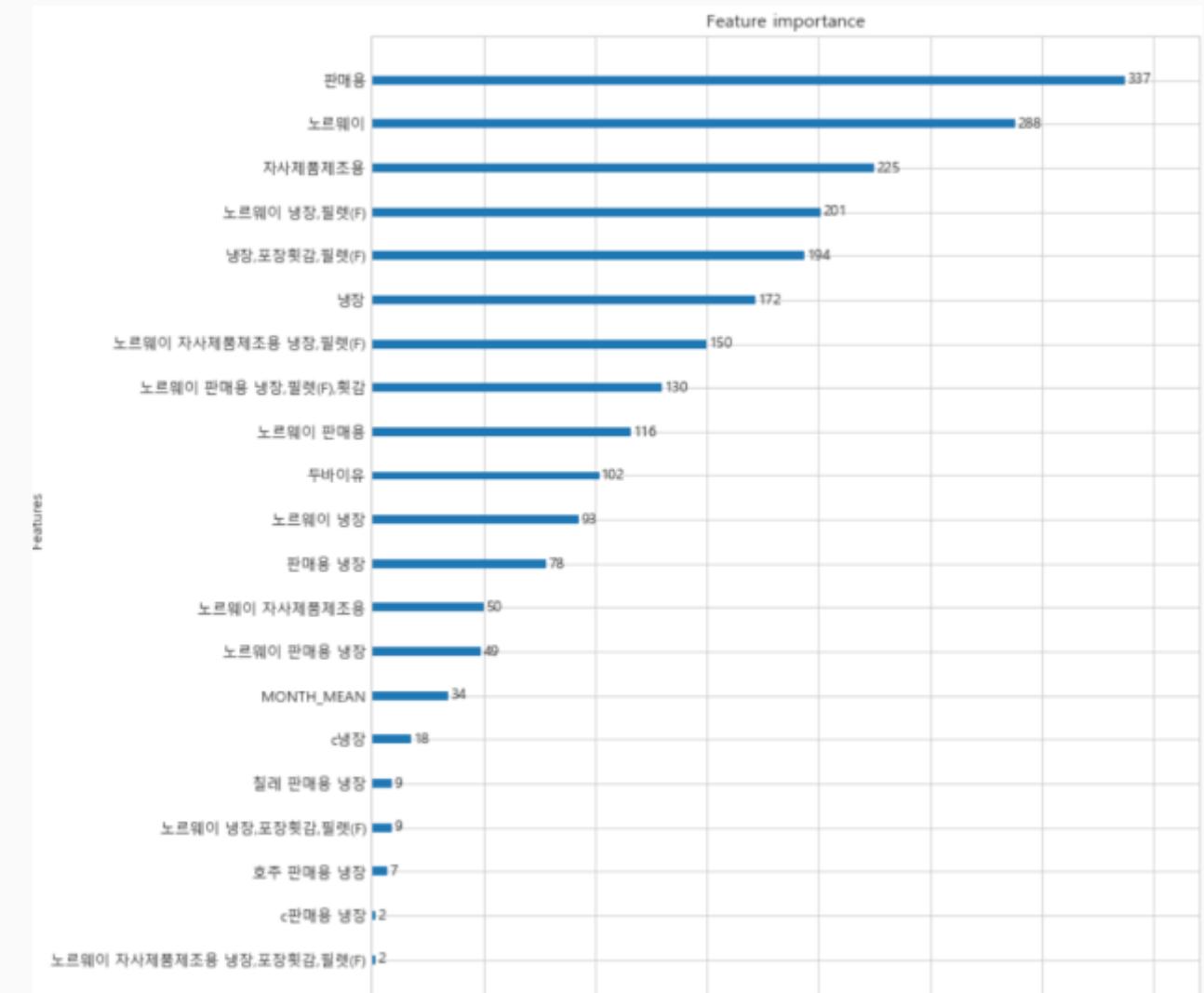
shrimp_initial
234개 데이터, 157개 변수



✓ 최종 변수 선정

노르웨이과 P_PRICE의 상관계수 : 0.8128772225011722
 뉴질랜드과 P_PRICE의 상관계수 : 0.12620909402708946
 아이슬란드과 P_PRICE의 상관계수 : 0.13830680739867215
 아일랜드과 P_PRICE의 상관계수 : 0.005064134970081402
 영국과 P_PRICE의 상관계수 : 0.005617526834984044
 칠레과 P_PRICE의 상관계수 : -0.47544214548658015
 캐나다과 P_PRICE의 상관계수 : -0.0926569241646168
 호주과 P_PRICE의 상관계수 : -0.3344838066981276
 노르웨이과 P_PRICE의 상관계수 : -0.04372819715914675
 뉴질랜드과 P_PRICE의 상관계수 : 0.1262090940270894
 아이슬란드과 P_PRICE의 상관계수 : 0.12597346870610332
 아일랜드과 P_PRICE의 상관계수 : 0.0023285397374101743

OLS Regression Results			
Dep. Variable:	P_PRICE	R-squared:	0.201
Model:	OLS	Adj. R-squared:	0.197
Method:	Least Squares	F-statistic:	46.83
Date:	Wed, 15 Sep 2021	Prob (F-statistic):	1.09e-10
Time:	13:55:15	Log-Likelihood:	-244.89
No. Observations:	188	AIC:	493.8
Df Residuals:	186	BIC:	500.3
Df Model:	1		
Covariance Type:	nonrobust		



상관계수 절댓값 0.4이상

p-value 값 0.05이상

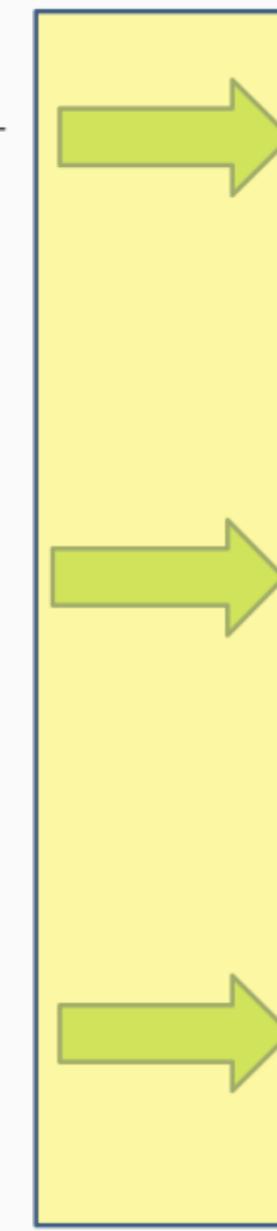
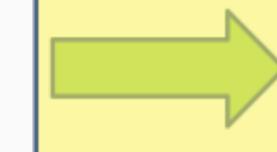
XGBoost 변수 중요도 참고

 데이터셋 생성

이상치 제거, 유효한 데이터 선정

RAW DATA (p_name = 연어)
1895개 데이터, 10개 변수preprocessingdata_salmon_
1840개 데이터, 10개 변수RAW DATA (p_name = 오징어)
2771개 데이터, 10개 변수preprocessingdata_squid
2731개 데이터, 10개 변수RAW DATA (p_name = 흰다리새우)
3133개 데이터, 10개 변수preprocessingdata_shrimp
2842개 데이터, 10개 변수

변수 생성, 결측치 처리

salmon_initial
235개 데이터, 120개 변수squid_initial
234개 데이터, 154개 변수shrimp_initial
261개 데이터, 157개 변수salmon_result
235개 데이터, 26개 변수squid_result
234개 데이터, 10개 변수shrimp_result
234개 데이터, 10개 변수



최종 변수 선정 - 연어

연어 분류 방식	변수명
날짜 관련 변수	month_mean
수출국 관련 변수	norway_p
수입용도 관련 변수	make_p, sale_p
수입형태 관련 변수	cold_p, cold, pack, f_p, cold_c
기본변수를 조합한 파생변수	n_make_p, n_sale_p, n_cold_p, n_cold,pack,f_p, n_cold,f_p, n_make_cold,pack, f_p, n_make_cold,f_p, n_sale_cold_p, n_sale_cold,f,h_p, chile_sale_cold_p, aus_sale_cold_p, n_make_cold,pack,f_p, chile_sale_cold_p, australia_sale_cold_c, sale_cold_p,sale_cold_c
외부데이터를 활용한 파생변수	oil_p
타겟변수	p_price

 최종 변수 선정 - 오징어

오징어 분류 방식	변수명
수출국 관련 변수	china_p, peru_p
수입용도 관련 변수	sale_p
수입형태 관련 변수	freeze, fuselage_p, freeze, fuselage, simmer_p, freeze, fin_p
기본변수를 조합한 파생변수	china_sale_p, peru_sale_p
타겟변수	p_price

 최종 변수 선정 - 흰다리새우

흰다리새우 분류 방식	변수명
수출국 관련 변수	vietnam_p
수입용도 관련 변수	sale_p
수입형태 관련 변수	freeze, boneless_p, freeze, boneless, boil, pack_c
기본변수를 조합한 파생변수	vietnam_freeze, boneless, boil, pack_p, thailand_freeze, boneless, boil, pack_p, vietnam_freeze, boneless, boil, pack_c, thailand_freeze, boneless, boil, pack_c
타겟변수	p_price

목 차

1. 문제 정의

- 대회 소개
- RAW DATA 소개
- 도메인 조사
- 외부데이터
- 분석 방향

2. 탐색적 데이터 분석

- 연어 데이터 분석
- 오징어 데이터 분석
- 흰다리새우 데이터 분석

3. 데이터 전처리

- 이상치 제거
- 유효한 데이터 선정하기
- 내부데이터를 활용한 파생변수 생성
- 외부데이터를 활용한 파생변수 생성
- 최종 변수 선정
- 결측치 처리
- 데이터 분할

4. 모델 구축과 검증

- 모델 탐색
- 모델 검증
- 최종 모델 구축

5. 모델 성능 향상

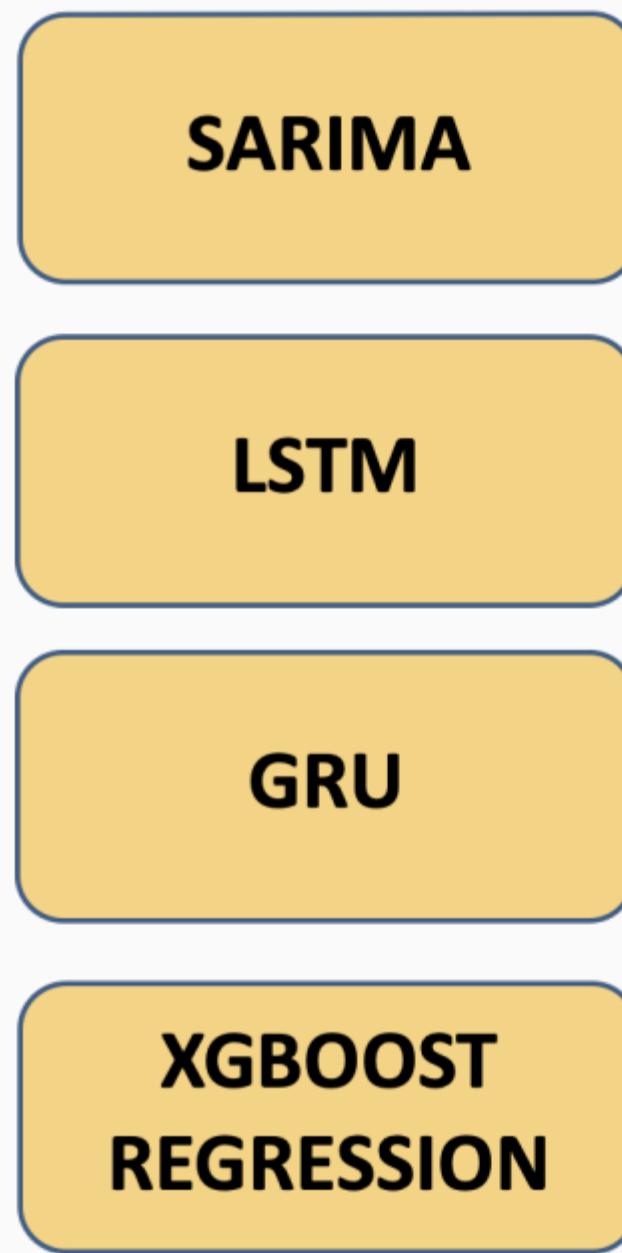
- 데이터 전처리
- 하이퍼파라미터 조정
- 앙상블

6. 제안 및 결론

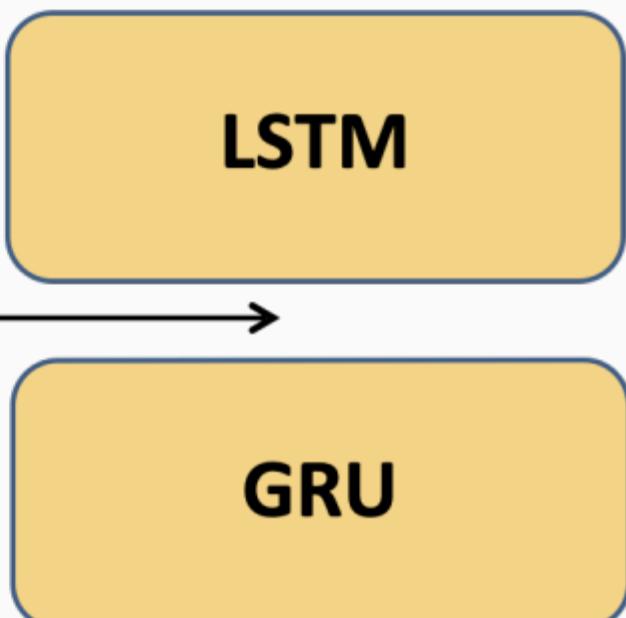
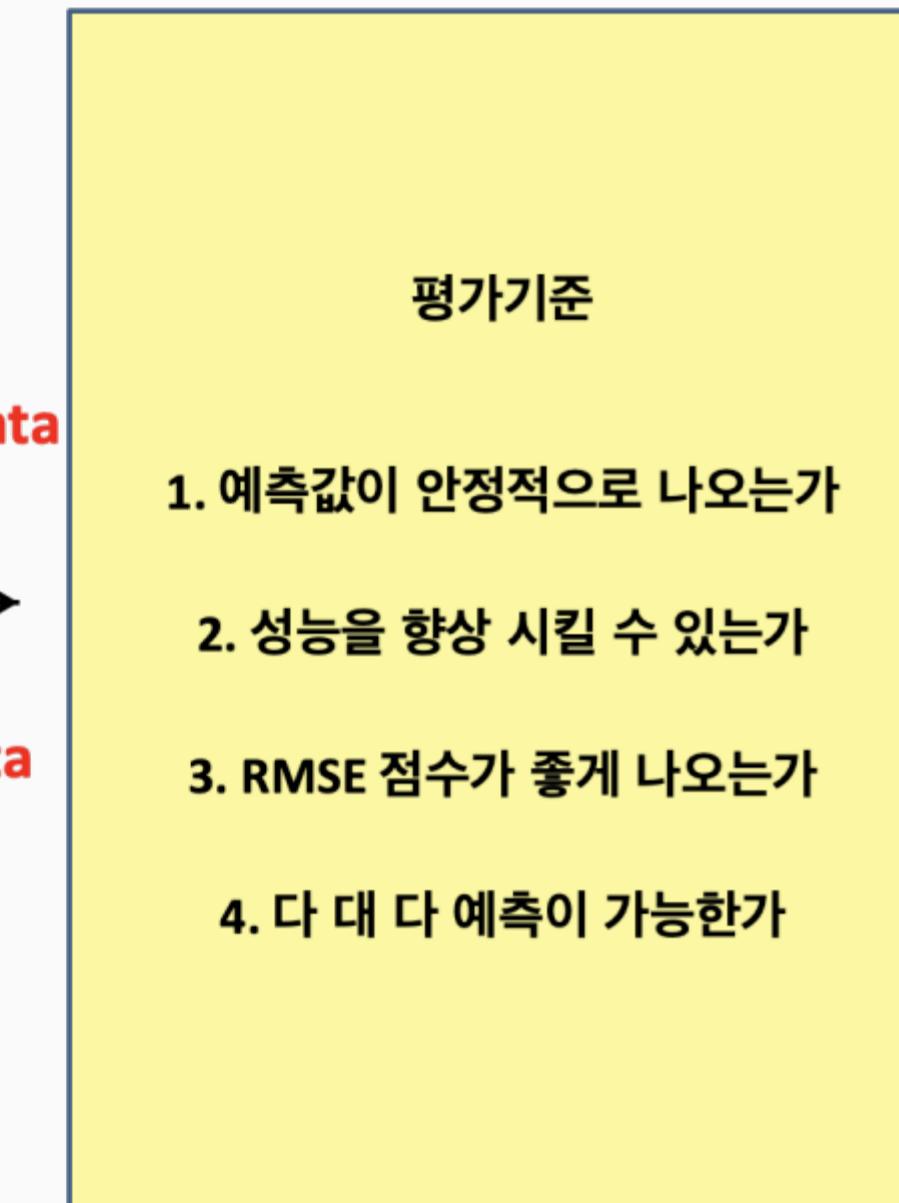
- 제안 및 결론

모델 탐색

대표적인 시계열 예측 모델



평가기준을 통해 2개의 모델 선택





모델 탐색

왜 다 대다 모델인가?

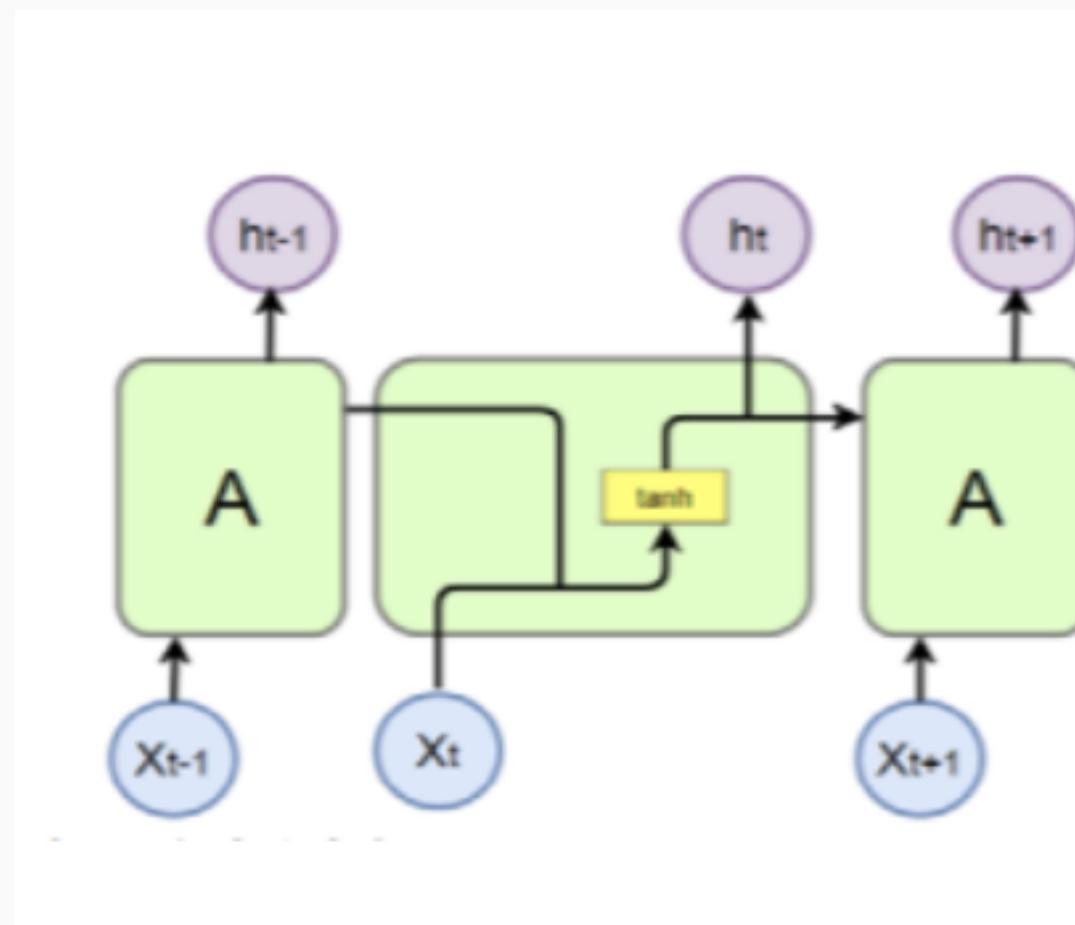
15.73743	12.8103	15.53716
15.76867	12.83593	14.57243
15.69546	12.84282	17.12177
15.32475	12.41241	16.78093
11.49348	10.89697	12.08999
14.33569	11.17672	13.70109
14.65823	11.66244	14.62323
15.4204	12.61998	15.51172
16.88718	16.72896	14.93354
15.68214	13.06554	14.62027

15.73743	12.8103	15.53716
15.76867	12.83593	14.57243
15.69546	12.84282	17.12177
15.32475	12.41241	16.78093
11.49348	10.89697	12.08999
14.33569	11.17672	13.70109
14.65823	11.66244	14.62323
15.4204	12.61998	15.51172
16.88718	16.72896	14.93354
15.68214	13.06554	14.62027

대회의 목적은 수산물 가격 예측을 통해 해양수산업 이해관계자들의 경영계획을 돋기 위함
주별 평균가격 뿐만 아니라 변수로 들어갈 그 주에 들어올 수입품목과 그 가격에 대한 예측도 함으로서
실질적으로 최대한 많은 수산업자들에게 도움을 주기 위하여 다:다 구성을 선정.

모델 탐색

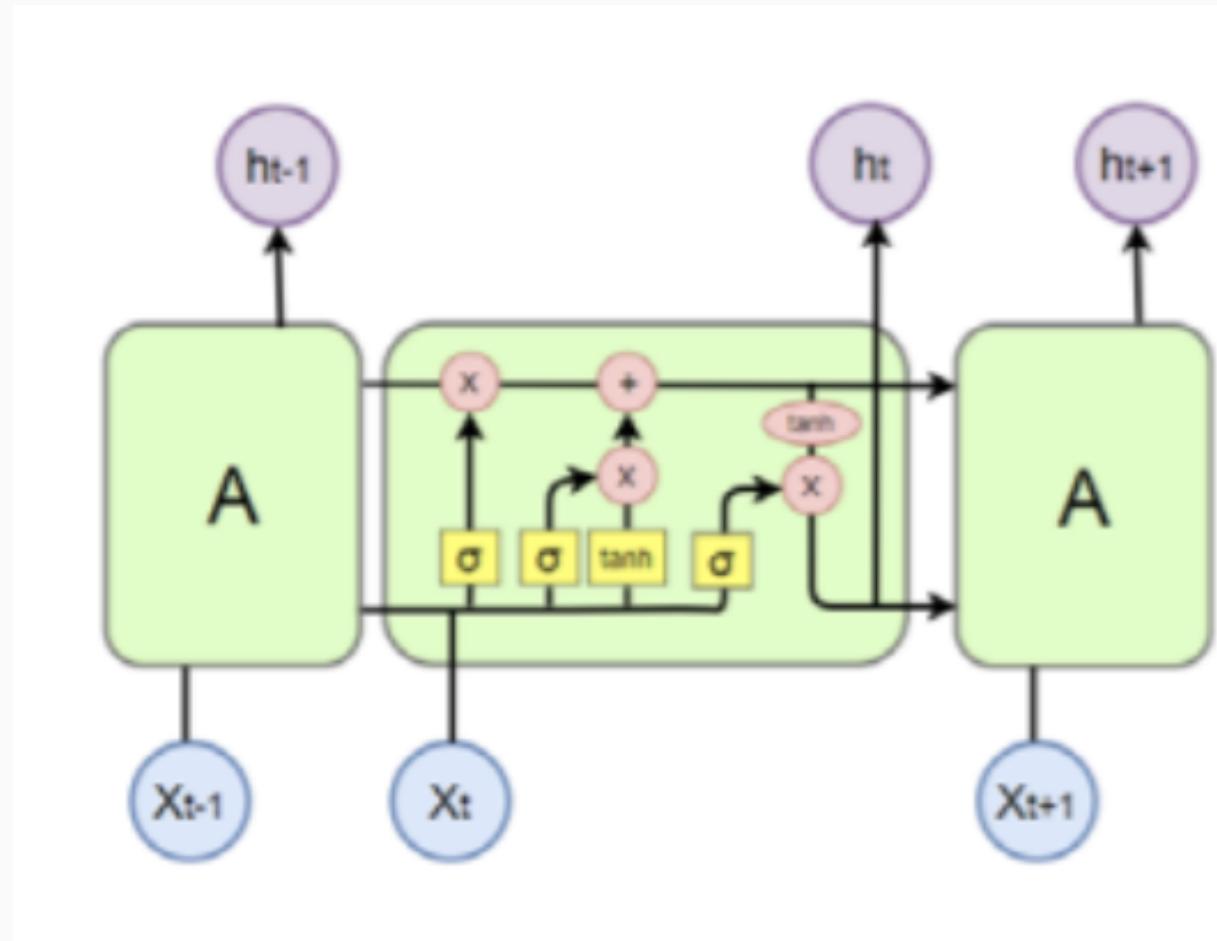
- 순환 신경망(Recurrent Neural Network, RNN)



- 일반신경망에서 시계열 개념이 추가된 것
- 장점: 은닉계층에 이전 정보를 기억시킬 수 있음 → 시계열 데이터 예측이 가능
- 단점: 학습이 반복되어 길어질 경우, 과거의 학습 상태가 사라지는 장기 의존성 문제가 있음

 모델 탐색

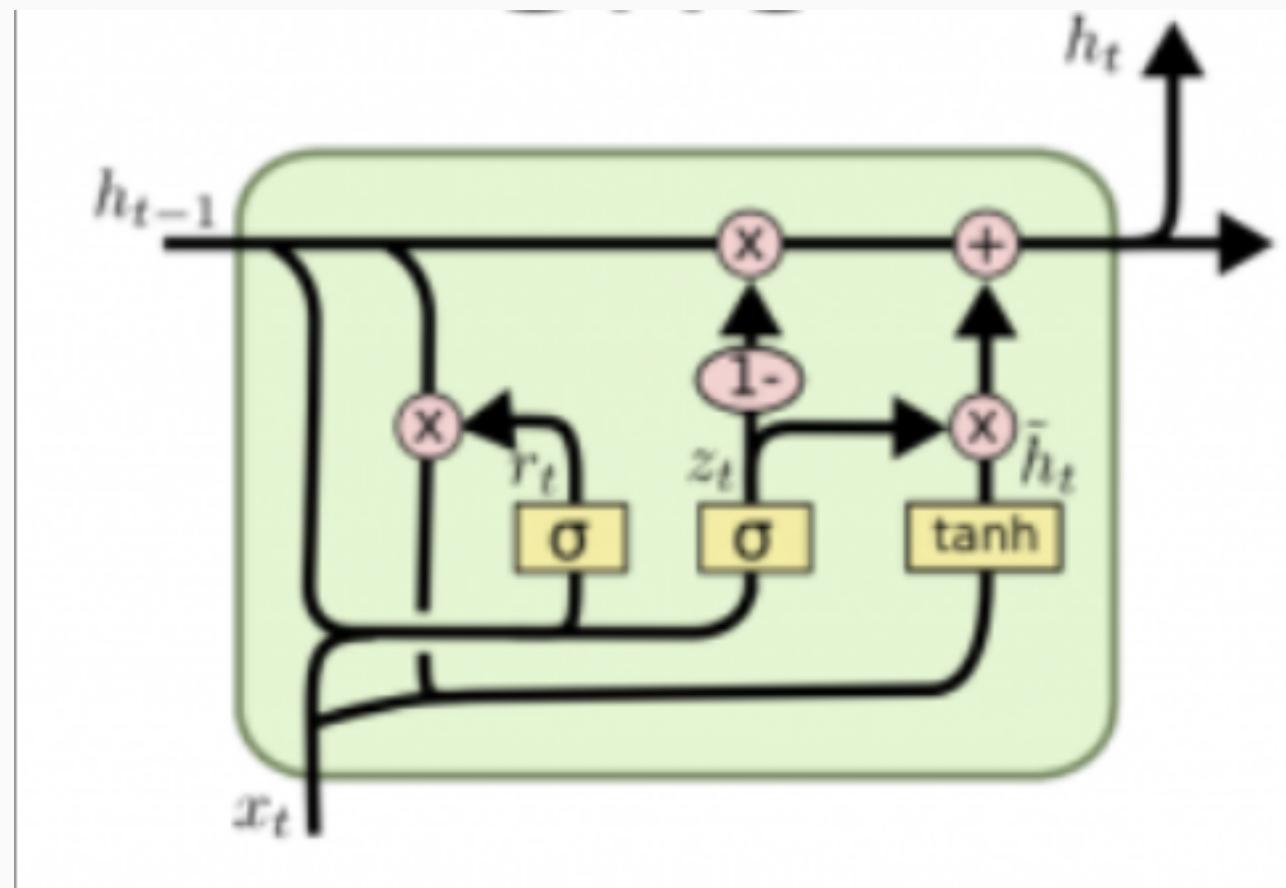
- LSTM(Long Short-Team Memory)



- 순환신경망의 일반적인 신경들 대신 LSTM 셀(Cell)로 교체된 구조를 가짐
- 장점: LSTM 셀들은 입력, 망각, 출력, 게이트를 가져 RNN에 비해 상태를 유지하기에 용이
→ RNN이 가진 장기 의존성 문제를 해결

 모델 탐색

• GRU

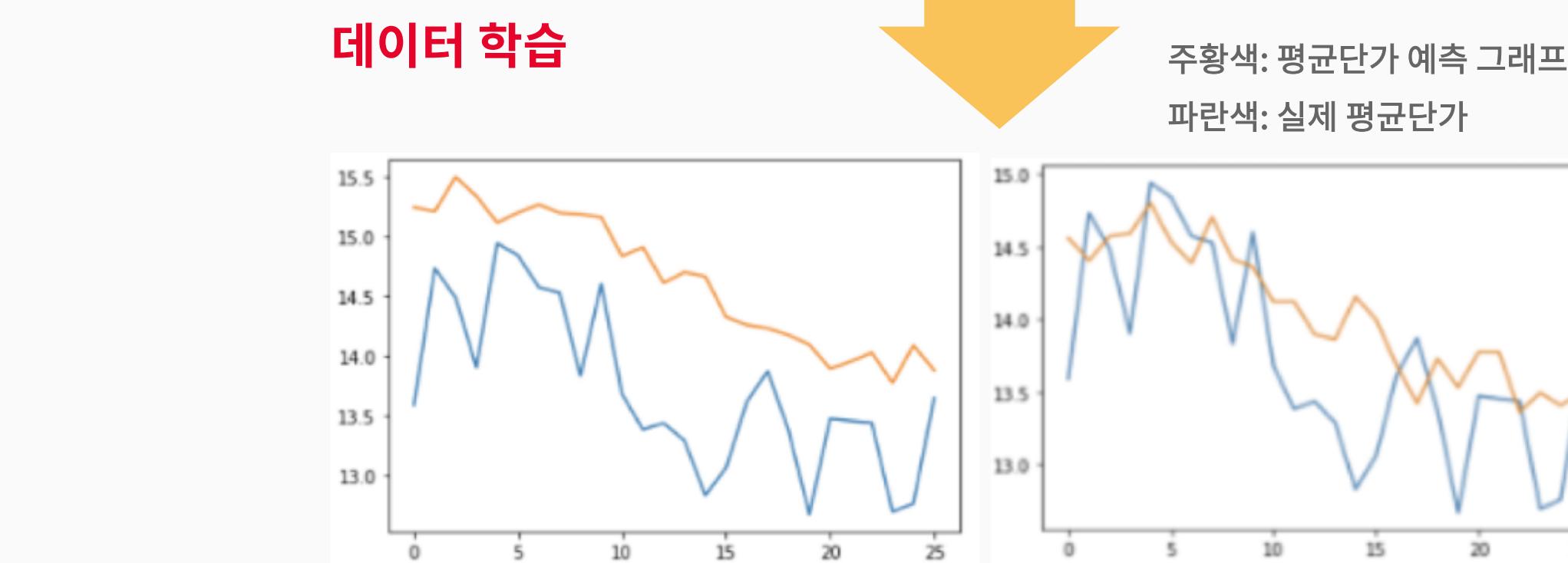
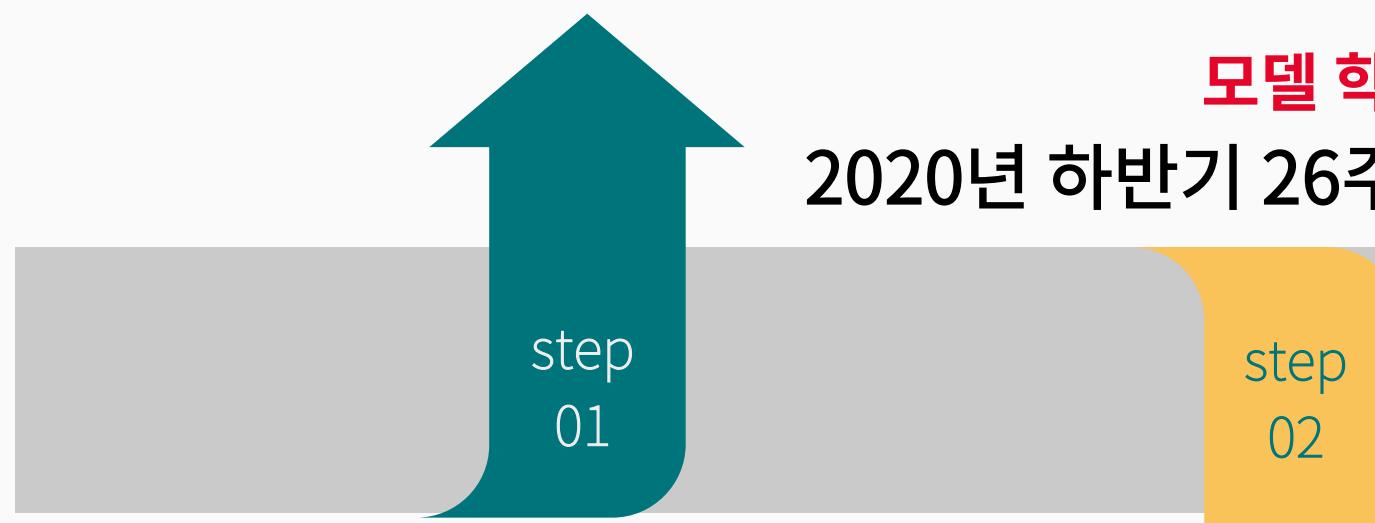


- LSTM셀은 입력, 망각, 출력 게이트를 가지지만 GRU셀은 리셋게이트와 업데이트 게이트 2개의 gate만 사용
- 장점: 기존 LSTM에 비해 더 간단한 구조를 가짐 → 학습할 파라미터가 더 적음



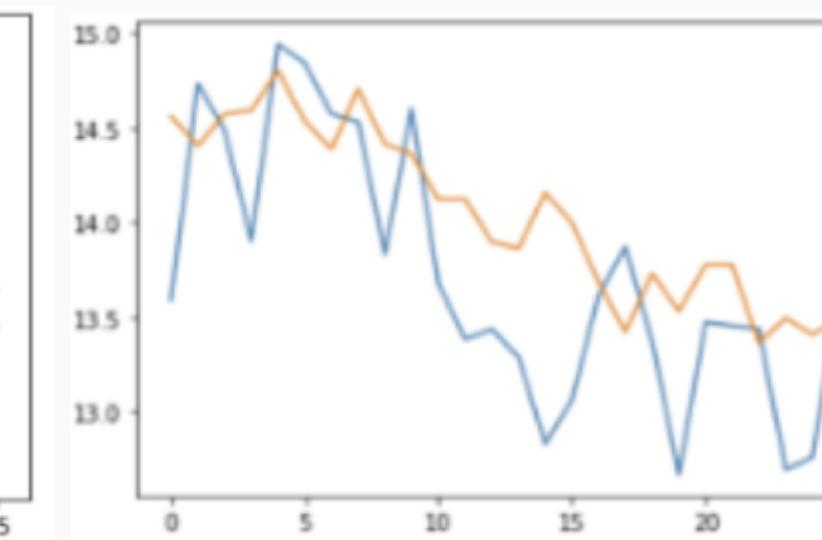
모델 검증

2020년 상반기까지의 연어 데이터를 학습



LSTM

GRU가 좋은 결과를 얻었다고 분석



GRU

LSTM도 가격 동향을 예측하는데 사용 할 수 있다고 판단
→ GRU와 LSTM 둘다 사용하기로 결정

최종 모델 구축

salmon_result

235개 데이터, 26개 변수



squid_result

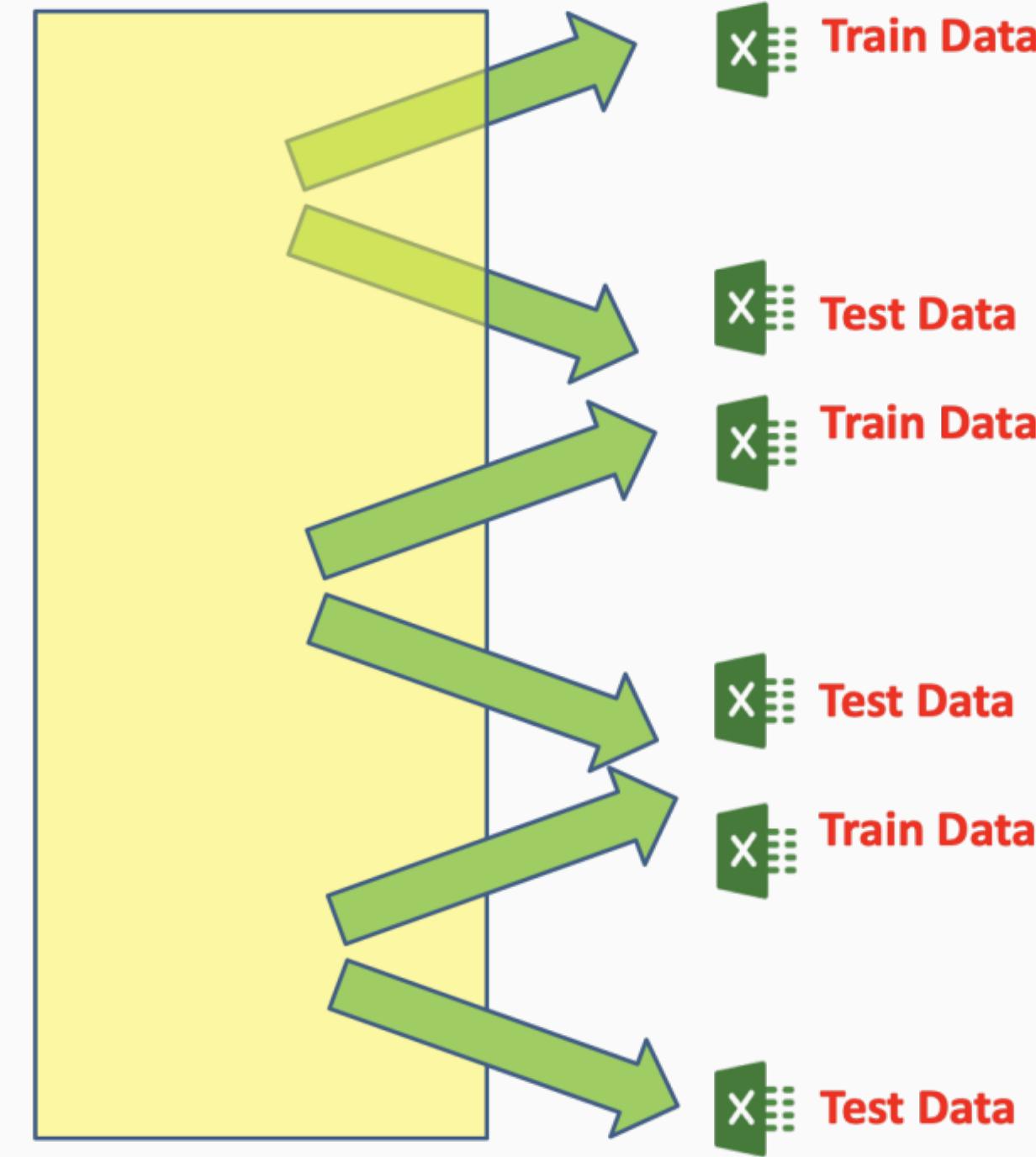
234개 데이터, 10개 변수



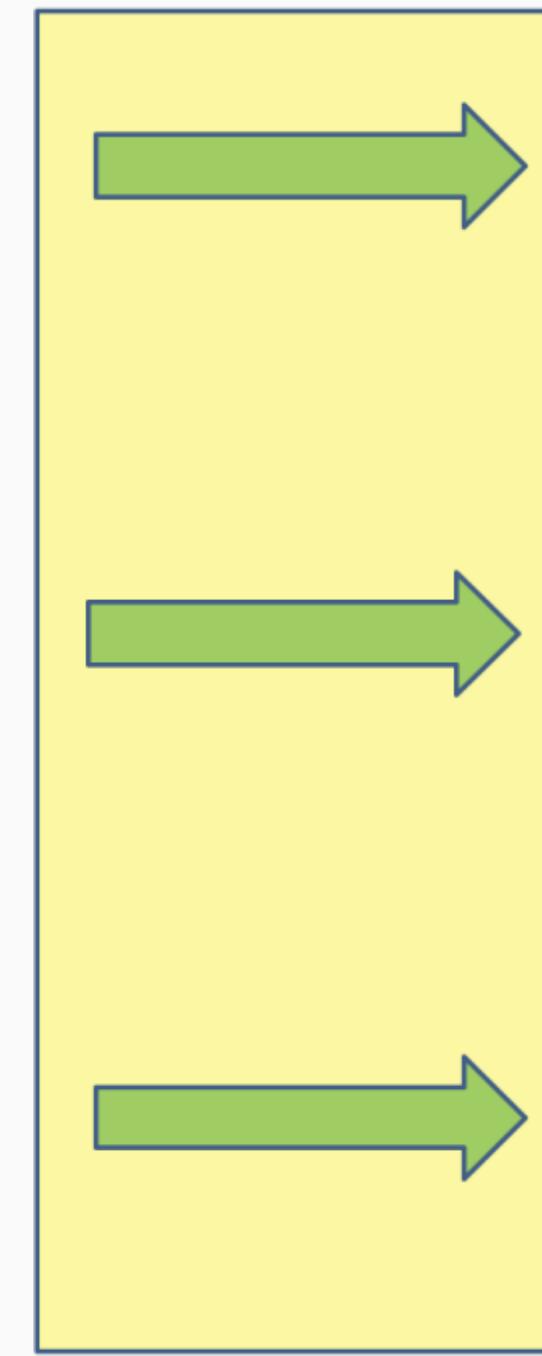
shrimp_result

234개 데이터, 10개 변수

데이터 분할



모델 검증 및 향상

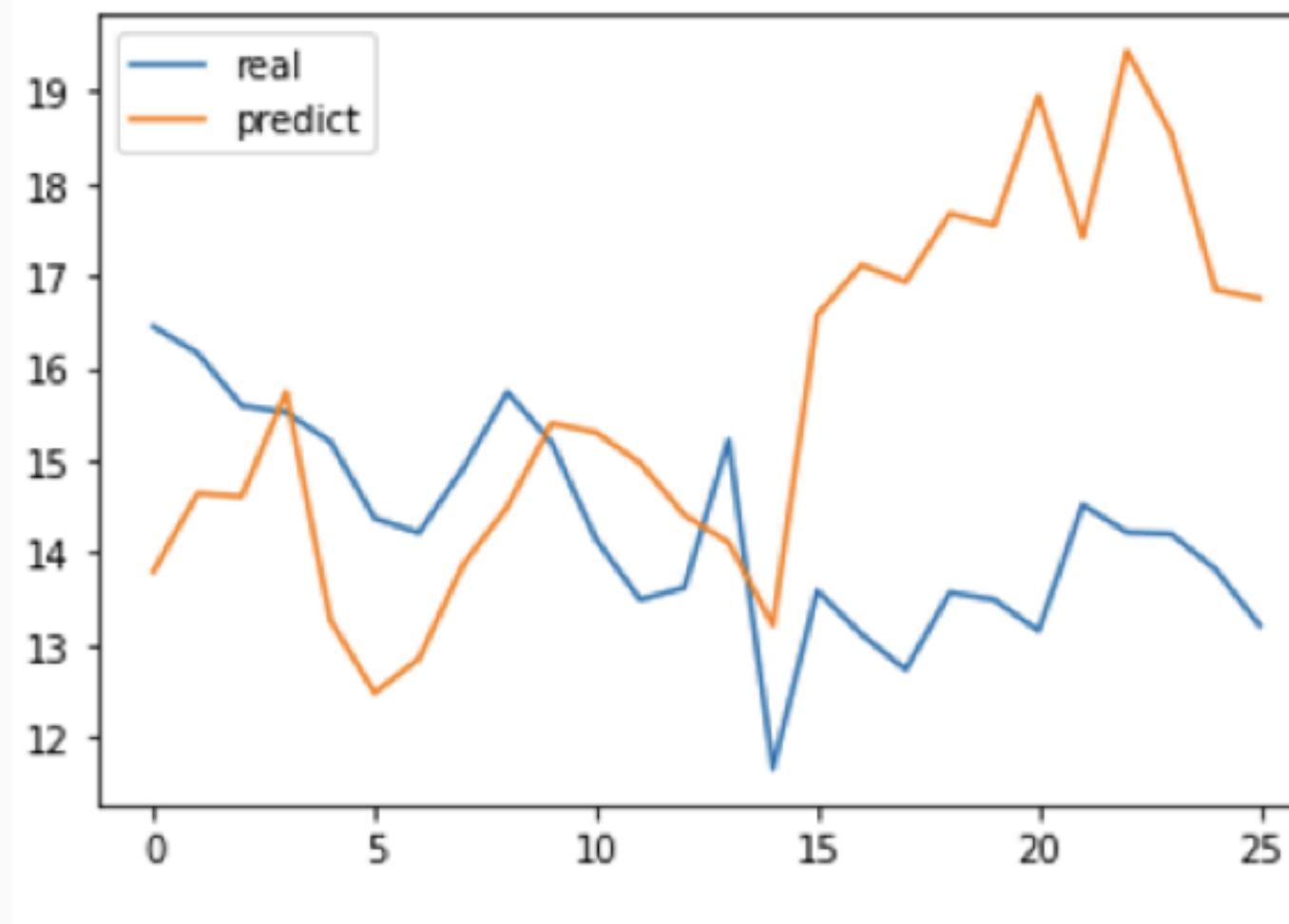


목 차

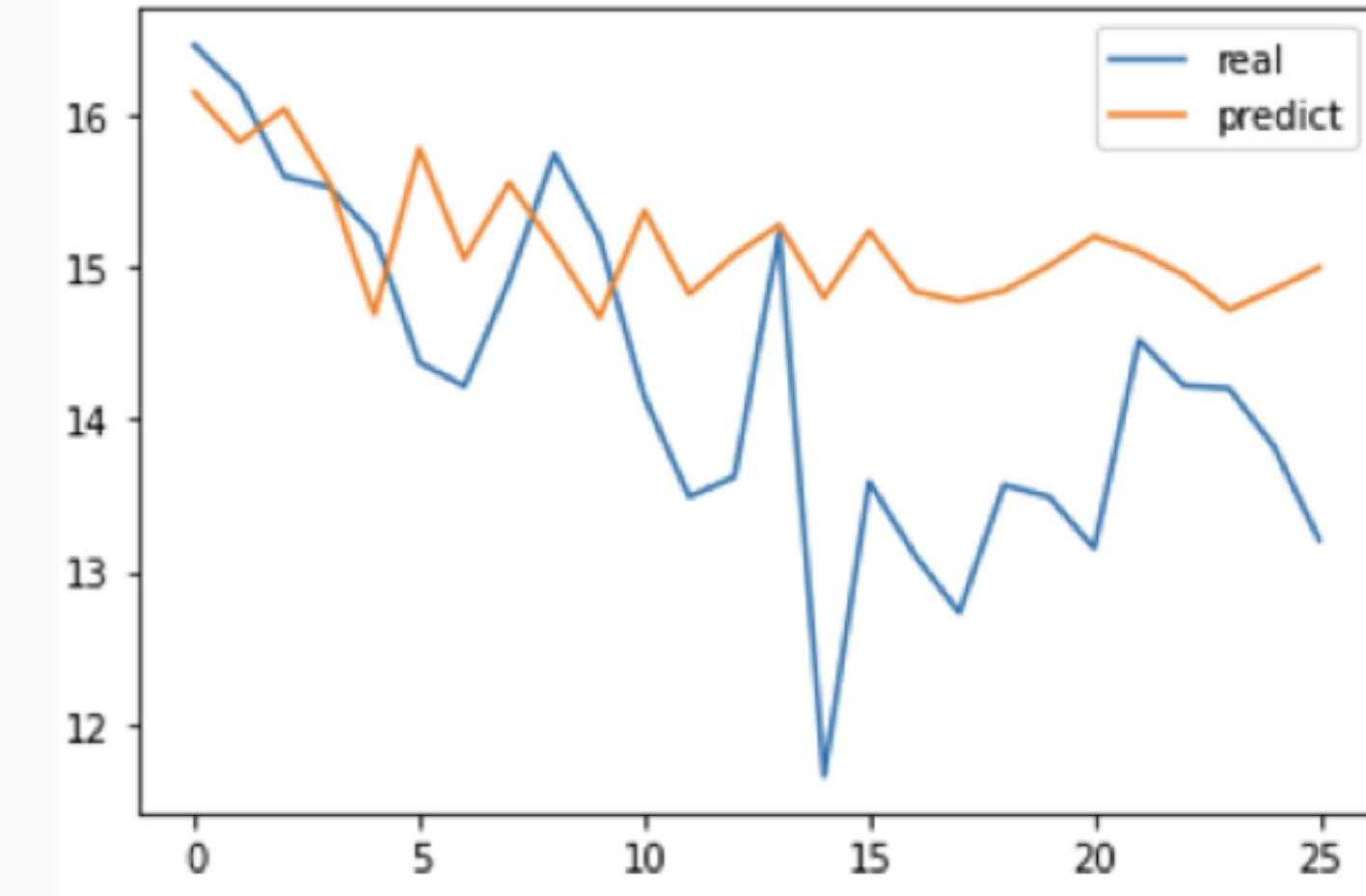


성능 향상을 위한 방법

- 데이터 전처리



제공된 2015년 12월부터 2019년 12월까지의
rawdata를 학습시킨 뒤,
2020년 상반기 26주를 예측한 그래프

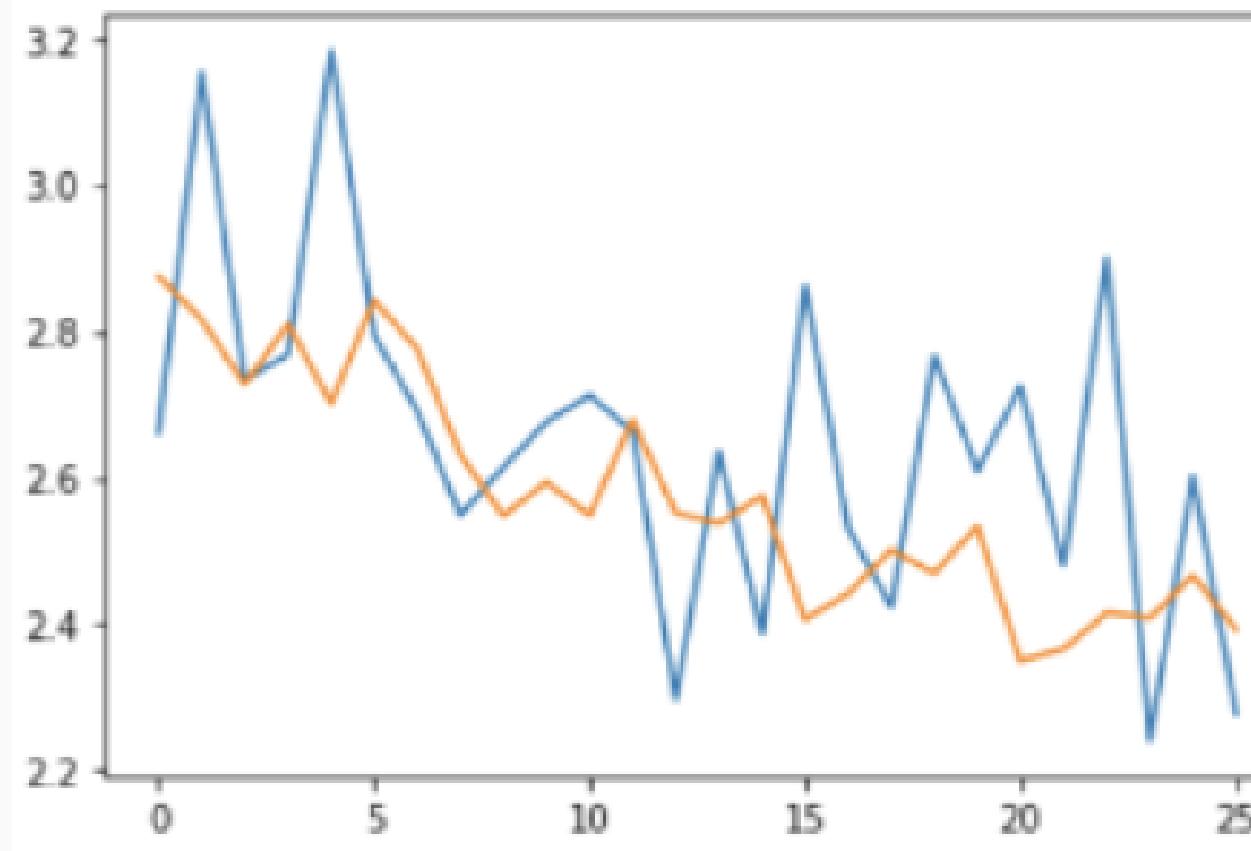


이상치 제거, 외부데이터 활용, 파생변수 생성 등의 방법으로
전처리한 데이터를 동기간 학습, 예측한 그래프

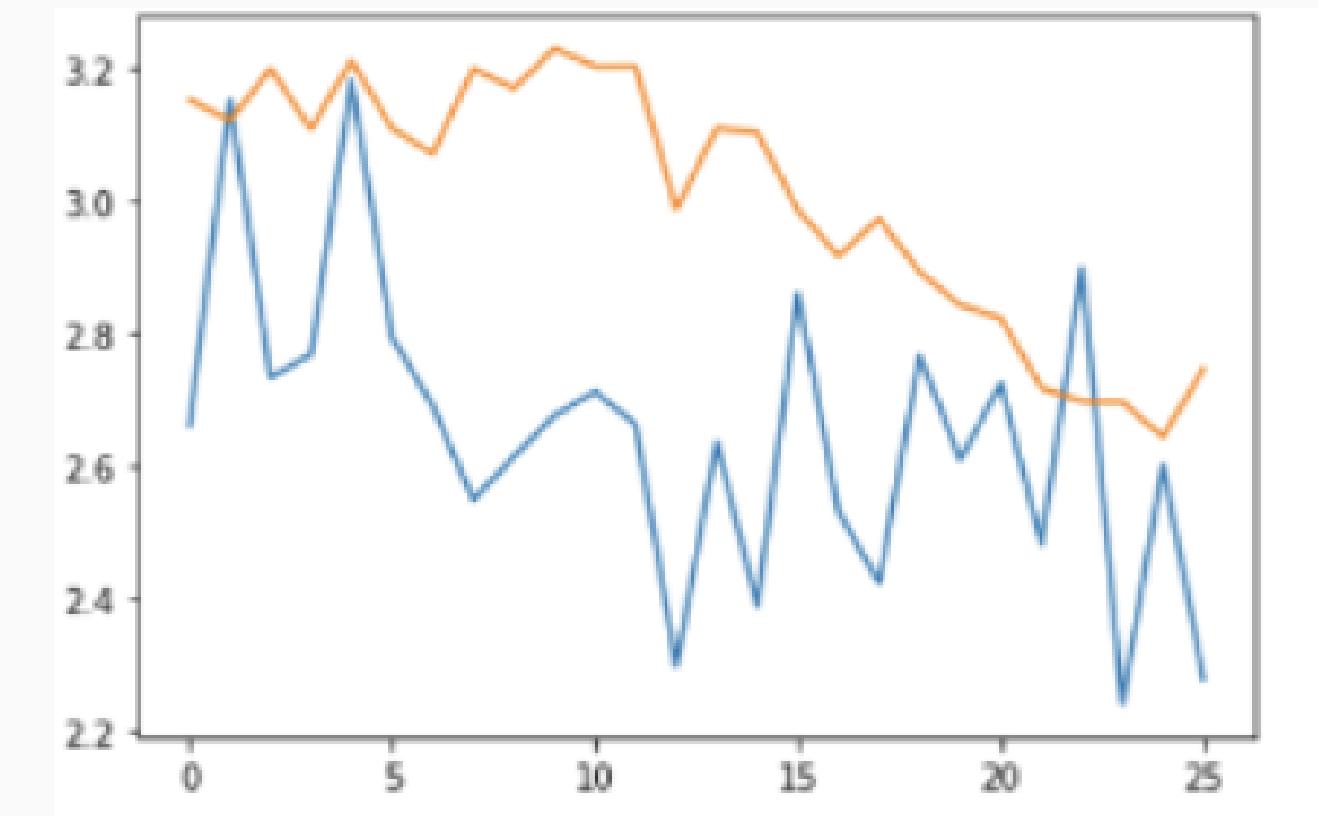


성능 향상을 위한 방법

- 모델의 하이퍼 파라미터 튜닝



> GRU의 첫번째 레이어의 유닛 수가 32,
학습의 배치 사이즈 수가 16 인 경우

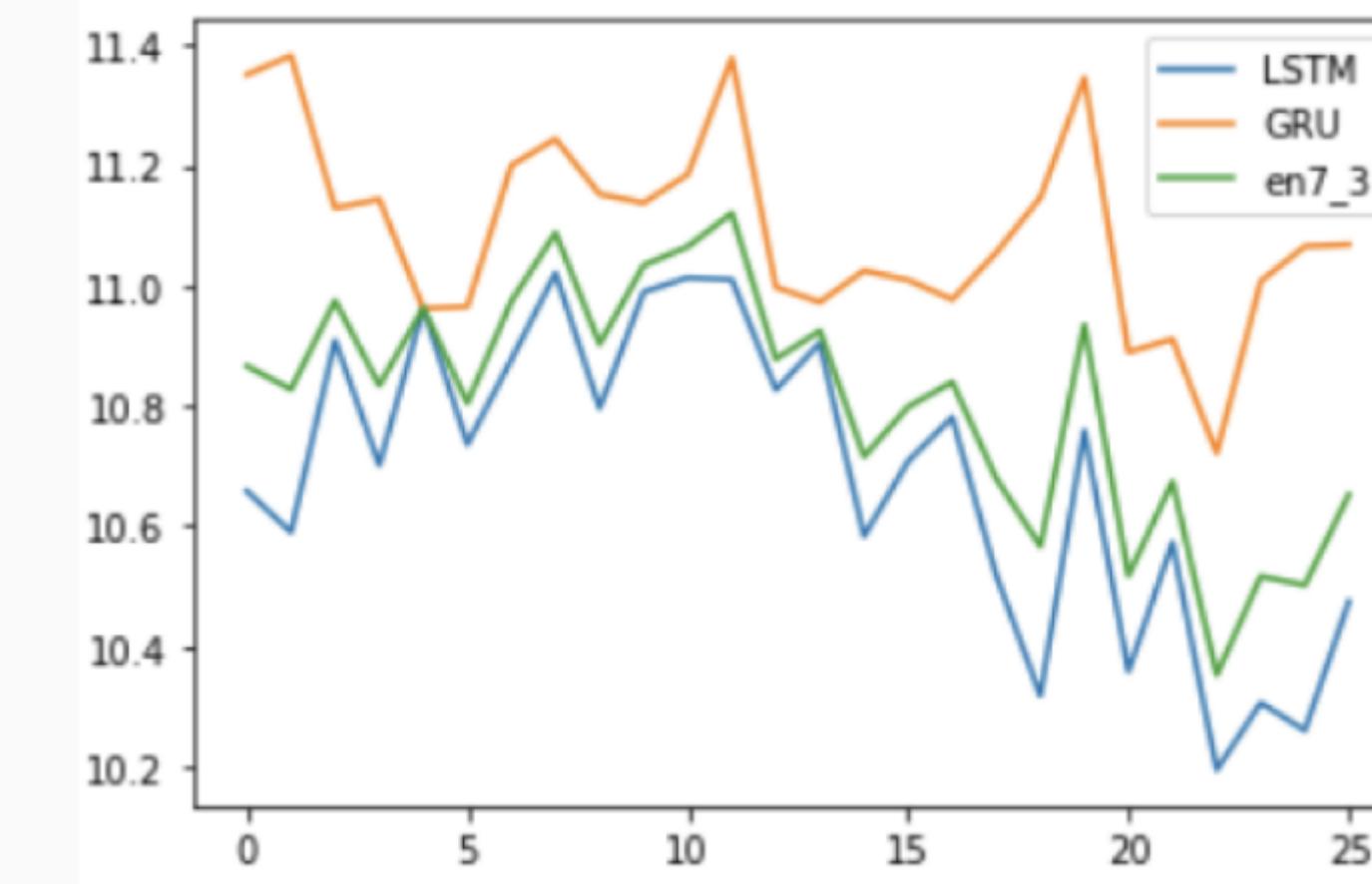
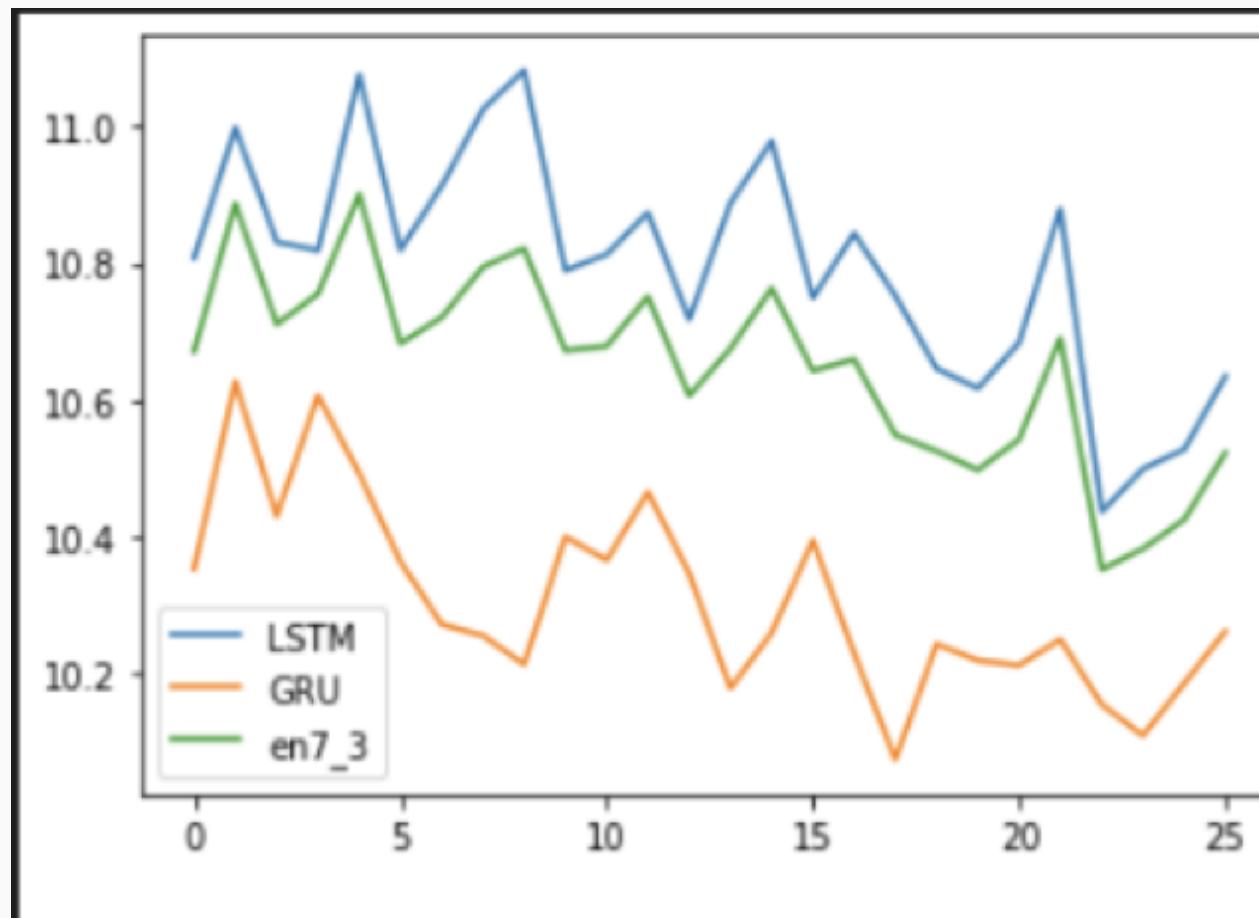


> GRU의 첫번째 레이어의 유닛 수가 32,
학습의 배치 사이즈 수가 8 인 경우



성능 향상을 위한 방법

- 소프트 앙상블

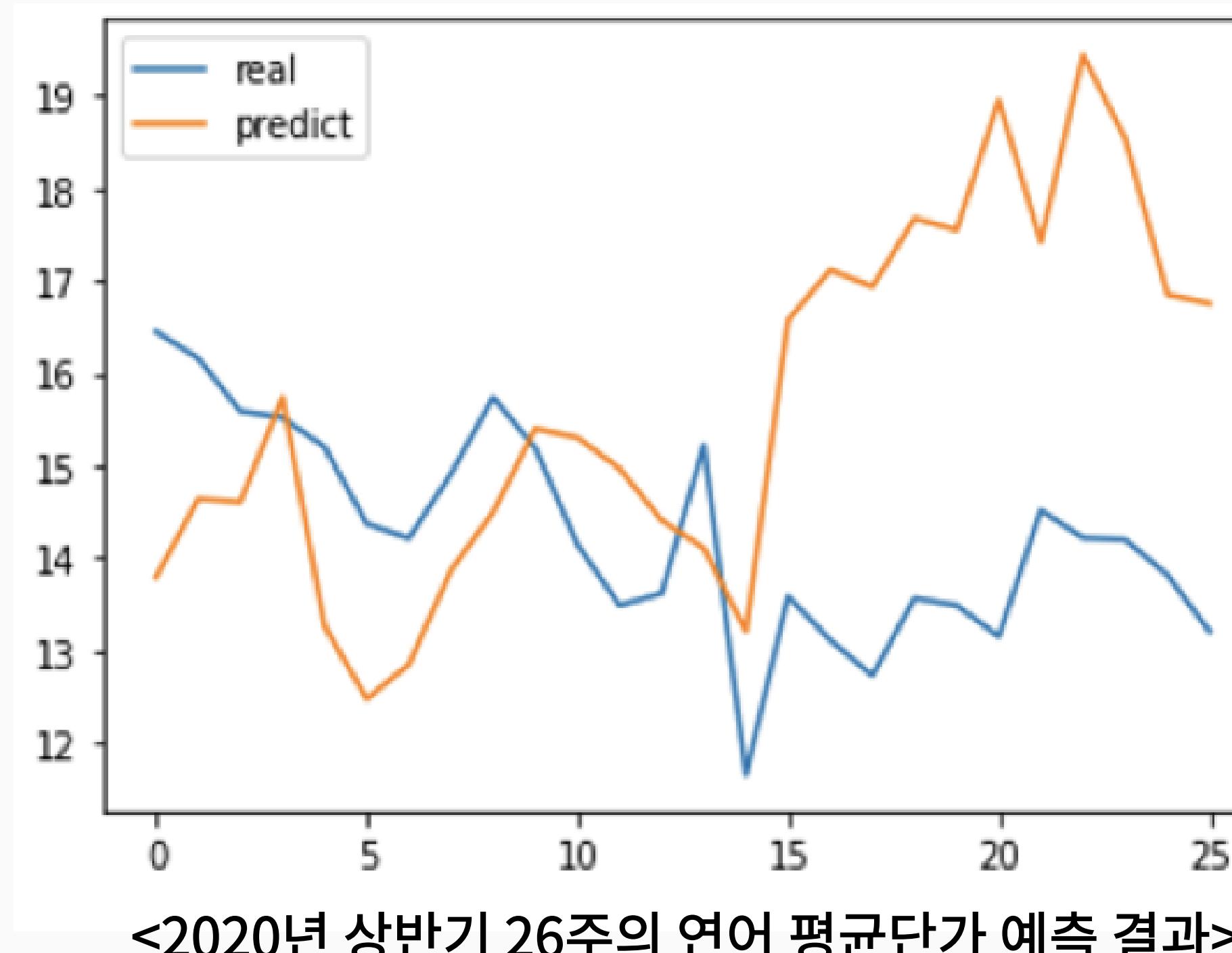


< 흰다리 새우의 2020년까지의 데이터를 학습하여 2021년 26주를 LSTM과 GRU로 예측해본 그래프 >

=> 둘 중 하나의 모델만 사용하게 된다면, 매 예측시마다 예측 값이 쉽게 변동되는 문제가 발생 할 수 있음.
두 모델을 동시에 사용하여 예측 값을 일정 비율로 앙상블 시켜준다면, 보다 안정적인 예측이 가능해짐.

성능 향상을 위한 방법의 예시

- RAW DATA



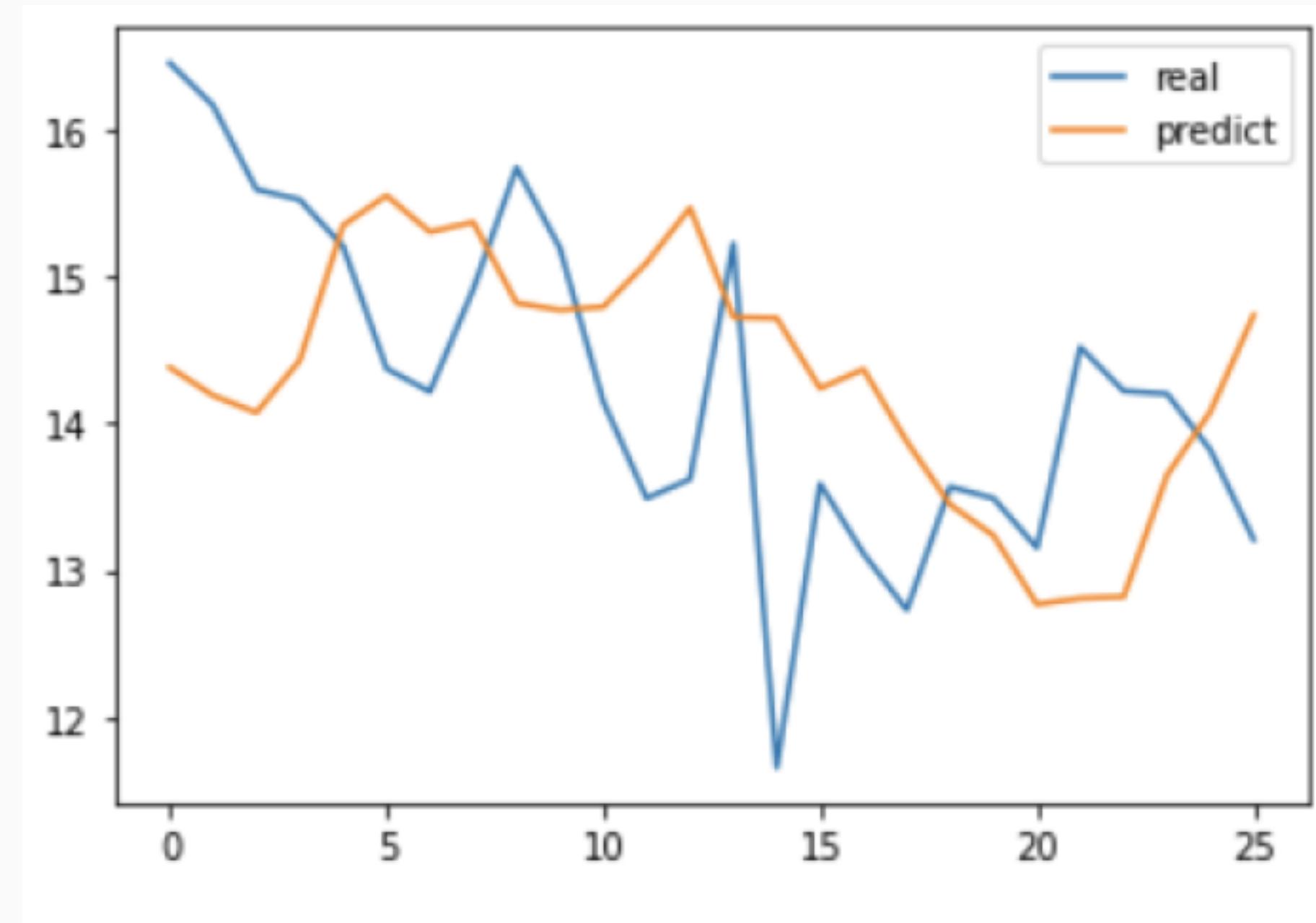
2015년12월부터 2019년12월까지의
rawdata로만 학습한 LSTM 모델



RMSE = 2.878720459665

성능 향상을 위한 방법의 예시

- 이상치 제거



2016년 7월부터 2019년 12월까지의
이상치가 제거된 데이터로 학습한 LSTM 모델



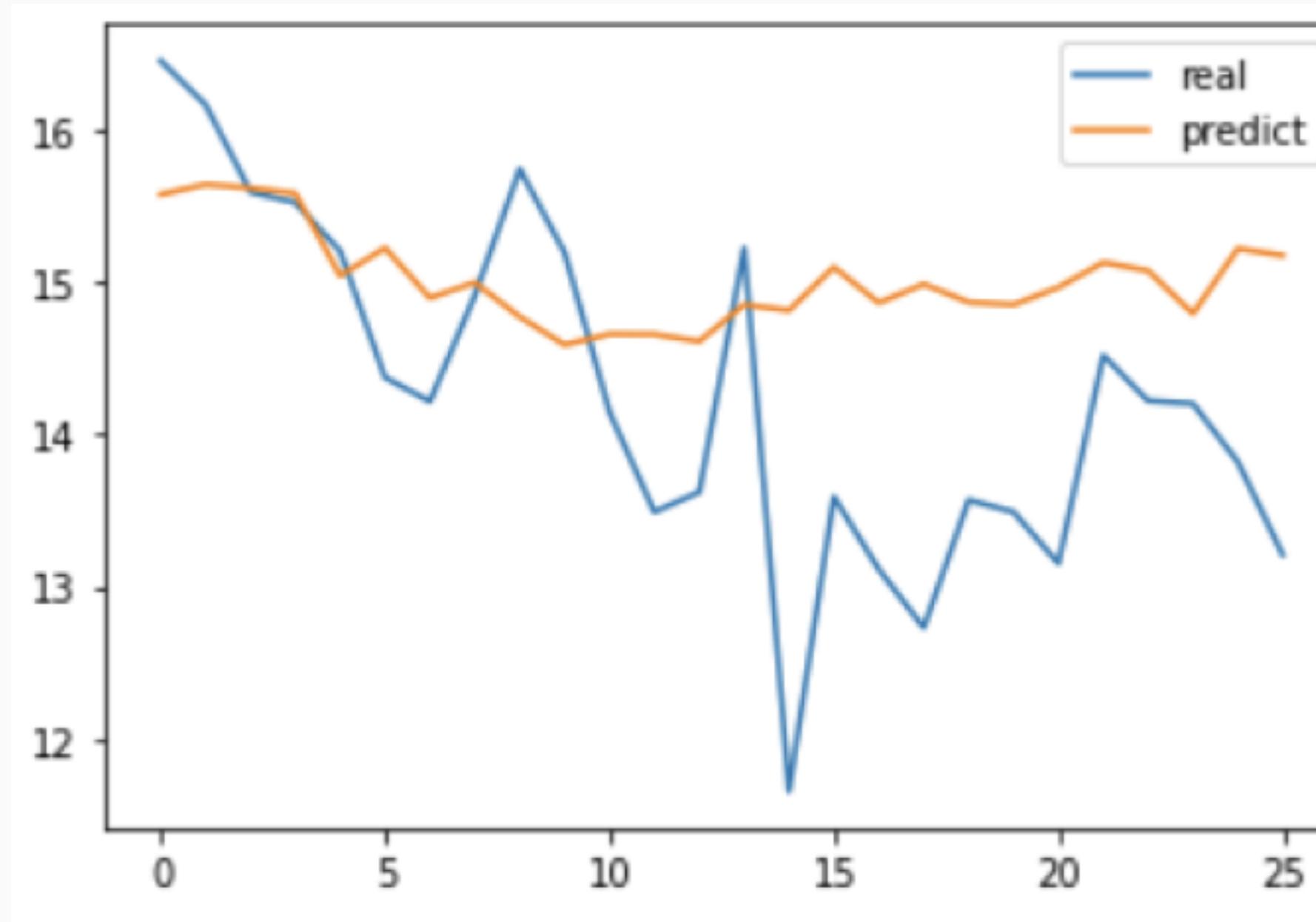
RMSE = 1.37863885904733

RMSE가 52.2% 개선

<2020년 상반기 26주의 연어 평균단가 예측 결과>

성능 향상을 위한 방법의 예시

- 파생변수와 외부변수 활용



<2020년 상반기 26주의 연어 평균단가 예측 결과>

2016년 7월부터 2019년 12월까지의
이상치가 제거된 데이터

+ 파생변수
+ 외부변수



상관계수가 0.4 이상인 변수들만을 뽑아낸
데이터로 학습한 LSTM 모델

RMSE = 1.2110665334308834



성능 향상을 위한 방법의 예시

- 파생변수와 외부변수 활용

1단계

2015년 12월부터 2019년 12월까지의
rawdata로만 학습한 LSTM 모델

RMSE = 2.878720459665

58% 개선

3단계

2016년 7월부터 2019년 12월까지의
이상치가 제거된 데이터

- + 파생변수
- + 외부변수

2단계

2016년 7월부터 2019년 12월까지의
이상치가 제거된 데이터로 학습한 LSTM 모델

RMSE = 1.37863885904733

12.2% 개선

상관계수가 0.4 이상인 변수들만을
뽑아낸 데이터로 학습한 LSTM 모델

RMSE = 1.2110665334308834

목 차

1. 문제 정의

- 대회 소개
- RAW DATA 소개
- 도메인 조사
- 외부데이터
- 분석 방향

2. 탐색적 데이터 분석

- 연어 데이터 분석
- 오징어 데이터 분석
- 흰다리새우 데이터 분석

3. 데이터 전처리

- 이상치 제거
- 유효한 데이터 선정하기
- 내부데이터를 활용한 파생변수 생성
- 외부데이터를 활용한 파생변수 생성
- 최종 변수 선정
- 결측치 처리
- 데이터 분할

4. 모델 구축과 검증

- 모델 탐색
- 모델 검증
- 최종 모델 구축

5. 모델 성능 향상

- 데이터 전처리
- 하이퍼파라미터 조정
- 앙상블

6. 제안 및 결론

- 제안 및 결론



제안 및 결론

최적의 가격예측 모형을 구성하기 위해 다음과 같은 과정이 필요했다.

1. 어종별 특성 이해
2. 어종의 특성에 따른 예측 모델 구성 전략
3. 어종의 특성에 따른 전처리
(노이즈 제거, 파생변수 및 외부변수 파악, 변수들의 상관계수 고려 등)

즉, 최적의 가격 예측 모형을 구성하기 위해서는,
데이터의 분석 및 전처리가 가장 중요함을 알 수 있었다.



제안 및 결론

1. 향상된 예측모델을 제공함으로써 해양수산업 이해관계자들의 효율적인 조업관리, 경영, 계획 수립에 도움을 줄 수 있음
2. 수산물 가격의 장기예측을 통해 국가적인 수입계획 수립에 도움 최적의 수입날짜를 예측하여, 다양한 상세어종 예측단가를 제공해 수입업자들의 수입 계획에 도움을 줄 수 있음
3. 시뮬레이터 구축을 통해 실질적으로 사용 가능한 예측 모델을 구축

THANK YOU!