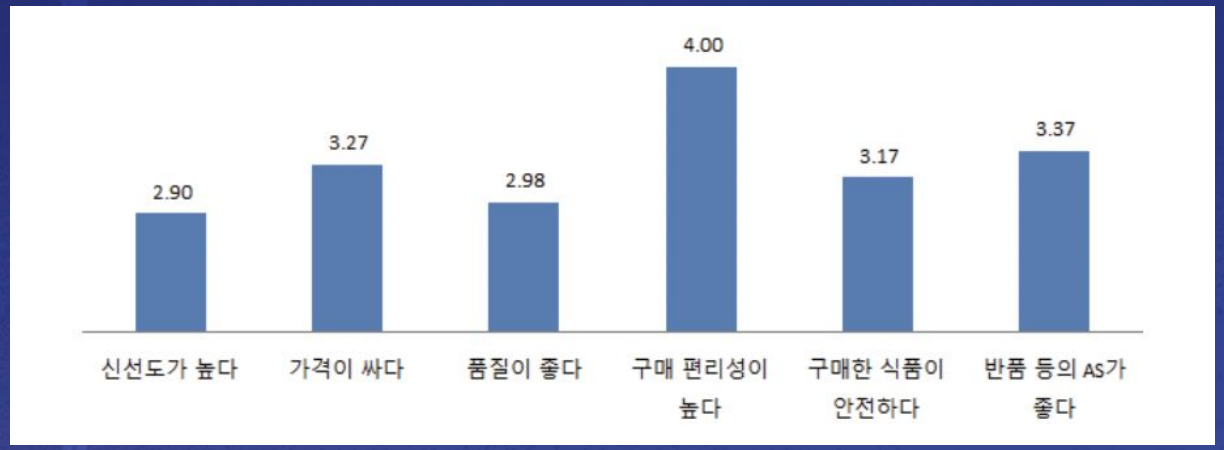
전체 스토리라인

1. 프로젝트 목표 : 무엇을 위한 모델링인가? 기업에겐 어떤 이득이?
2. 데이터 요약 : 데이터 출처, 기간, 품종 등 간략히
3. 데이터 특성 및 이상치, 이상치로 판별한 이유
4. 이상치 및 NaN 처리
5. 평가지표 선정 및 이유
6. 모델 선정 및 이유
7. 베이스라인 모델링
8. 모델 튜닝 : 튜닝을 위해 한 시도들(파생변수, 스케일러 등)
9. 튜닝 후 성능 비교

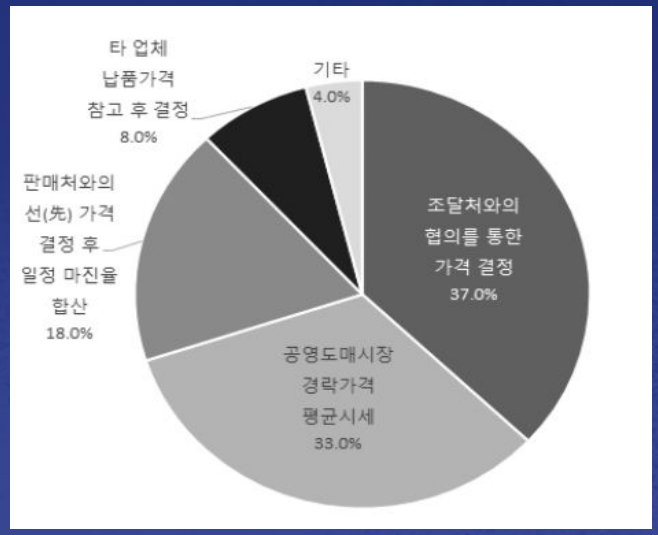
농산물 가격 예측 AI
최종 발표

9조 샛별조

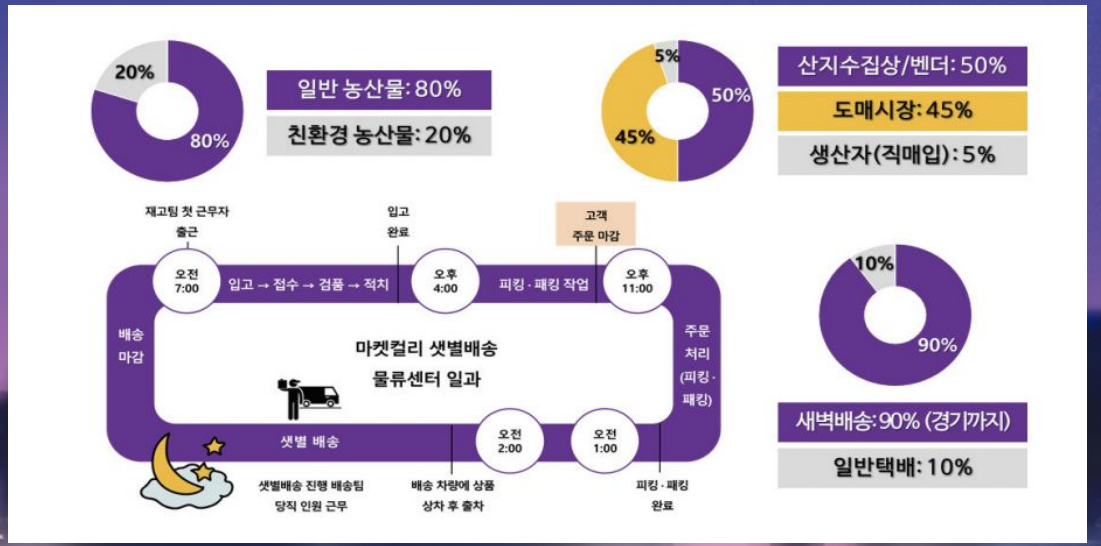
장한결(팀장), 김태성, 최고은



온·오프라인 농식품 소매 유통채널 인식 비교(5점 척도)



농산물 구매가격 결정 방식



마켓컬리 B2C 유형 비즈니스 모델

출처: 한국농촌경제연구원에서 “플랫폼 기반의 농산물 유통서비스 활성화”

INDEX

① 프로젝트 목표

농산물 가격 예측 모델 구현

컬리

최근 소비 트렌드 등 여러 요소를 고려해 주문량 예측

+

농산물 가격 예측 모델 (데이콘)[1)]

↓

적정 가격으로 구매 가능

↓

구매 비용 감소

1) https://dacon.io/competitions/official/235801/overview/description

② 타겟 설정

Dacon 2021
예측대회 품목

2022년 컬리 최다 판매량
농산물(채소, 과일)

배추 무 양파
건고추 마늘
얼갈이배추
양배추
시금치 미나리
당근 파프리카
새송이
샤인머스켓 청상추
백다다기 애호박

토마토

포도
(캠벨얼리)

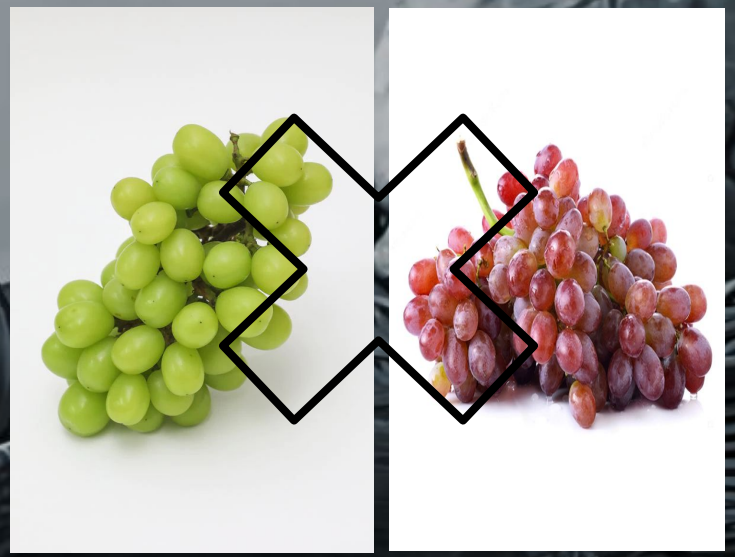
팽이버섯

대파

깻잎

감귤 사과
배 자두
콩나물 호박

③ EDA(Exploratory Data Analysis)



③ EDA(Exploratory Data Analysis)

EDA data :

기간 : 2016년 1월 1일 ~

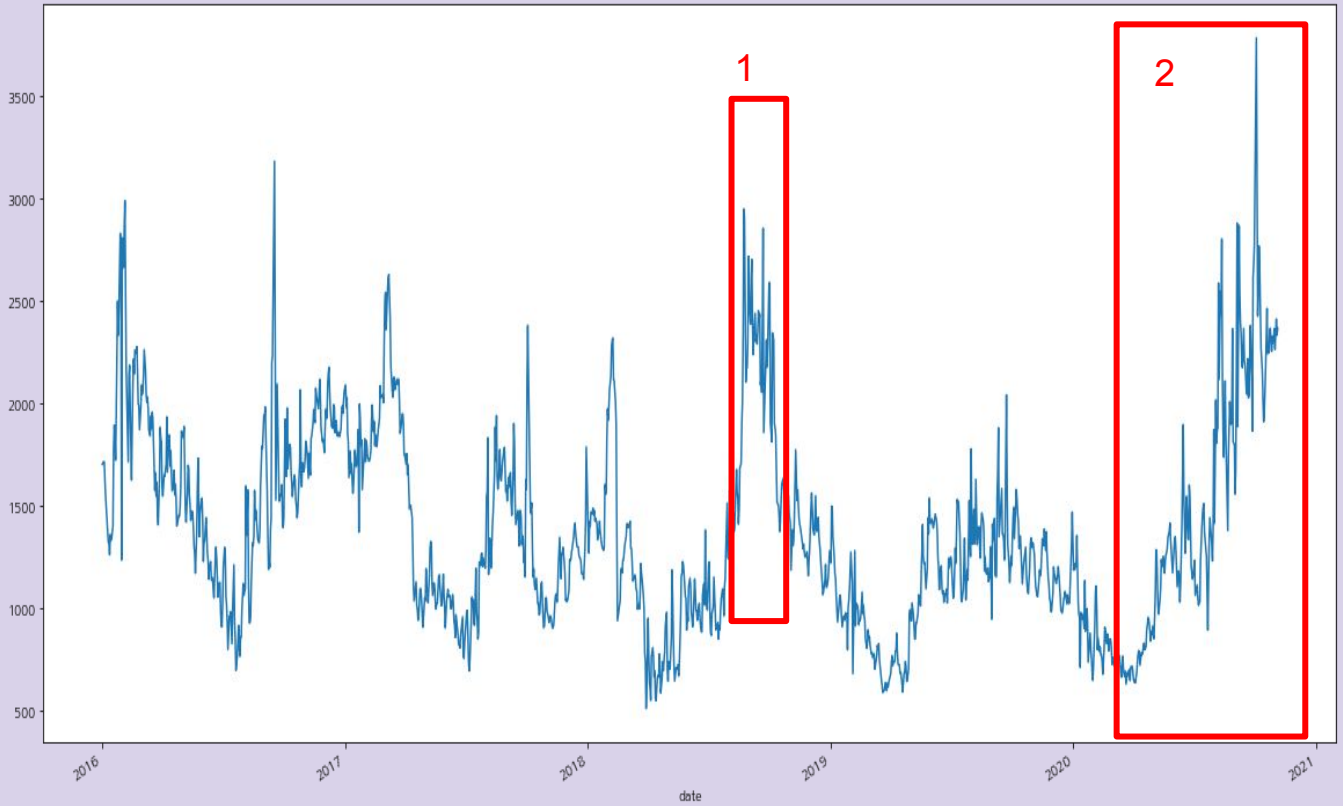
2020년 11월 04일

Data From:

Dacon 2021 농산물 가격 예측대회

https://dacon.io/competitions/official/235801/overview/description

③ EDA(Exploratory Data Analysis)



기간 : 2016년 1월 1일 ~ 2020년 11월 04일

대파

순위	전국			
	폭염일수		열대야일수	
1위	2018년	31.4	2018년	17.7

1 .2018년 폭염으로 역대급 폭염으로 인한 피해

2. 2020년 대파가격 상승의 이유

2020년 여름 장마로 인한 재배면적의 축소

2019년에 폭락한 대파 가격 때문에 재배면적을 정책으로 축소함

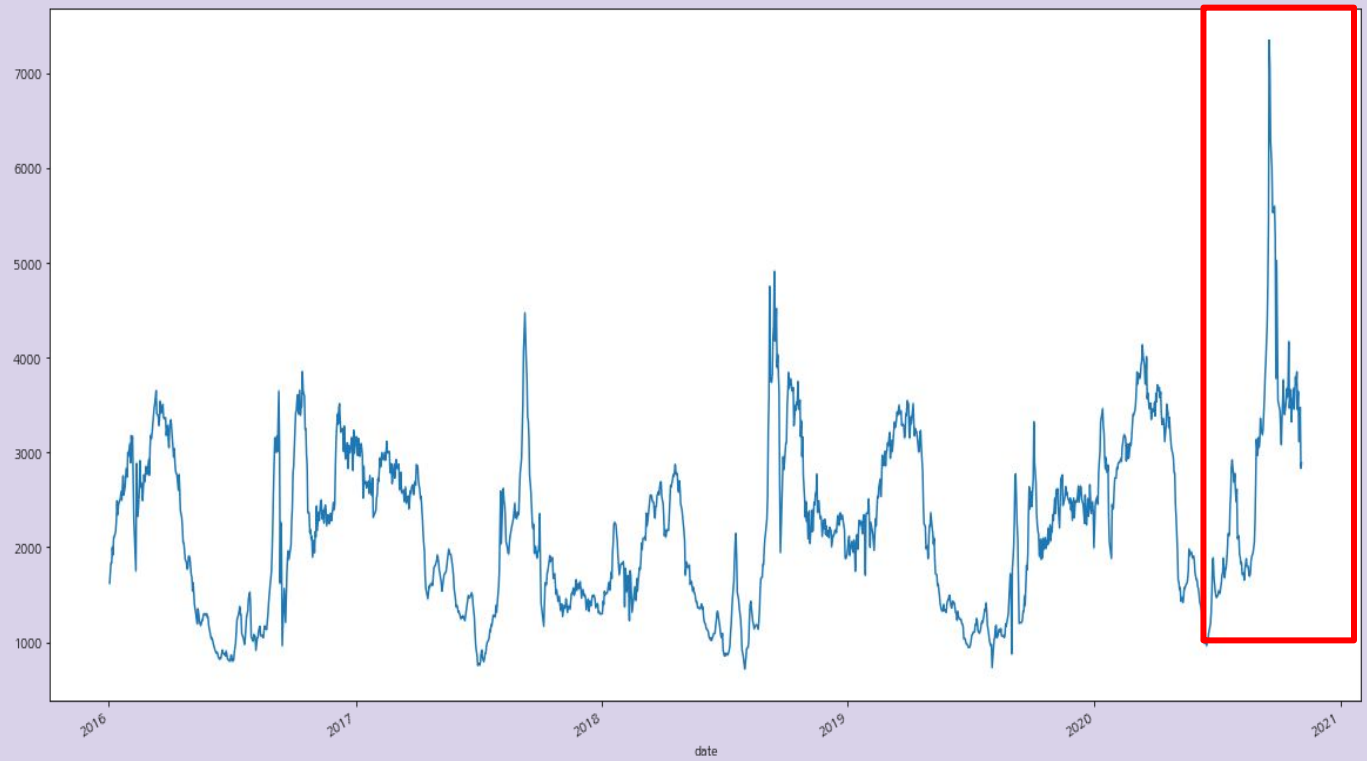
출처 >

https://news.mt.co.kr/mtview.php?no=2021041010050848016

chosun.com/site/data/html_dir/2021/02/09/2021020901666.html

https://www.ifs.or.kr/bbs/board.php?bo_table=News&wr_id=4341

③ EDA(Exploratory Data Analysis)

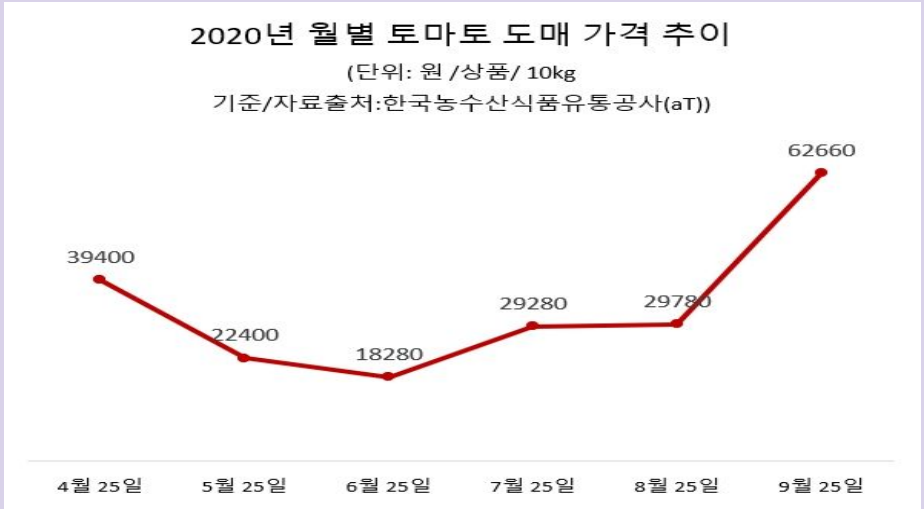


기간 : 2016년 1월 1일 ~ 2020년 11월 04일

#토마토

2020년 이상기후 한파와 폭설로 인한 냉해

[재배방법 및 시기] **** 수확시기 7월 ~ 10월까지 ****

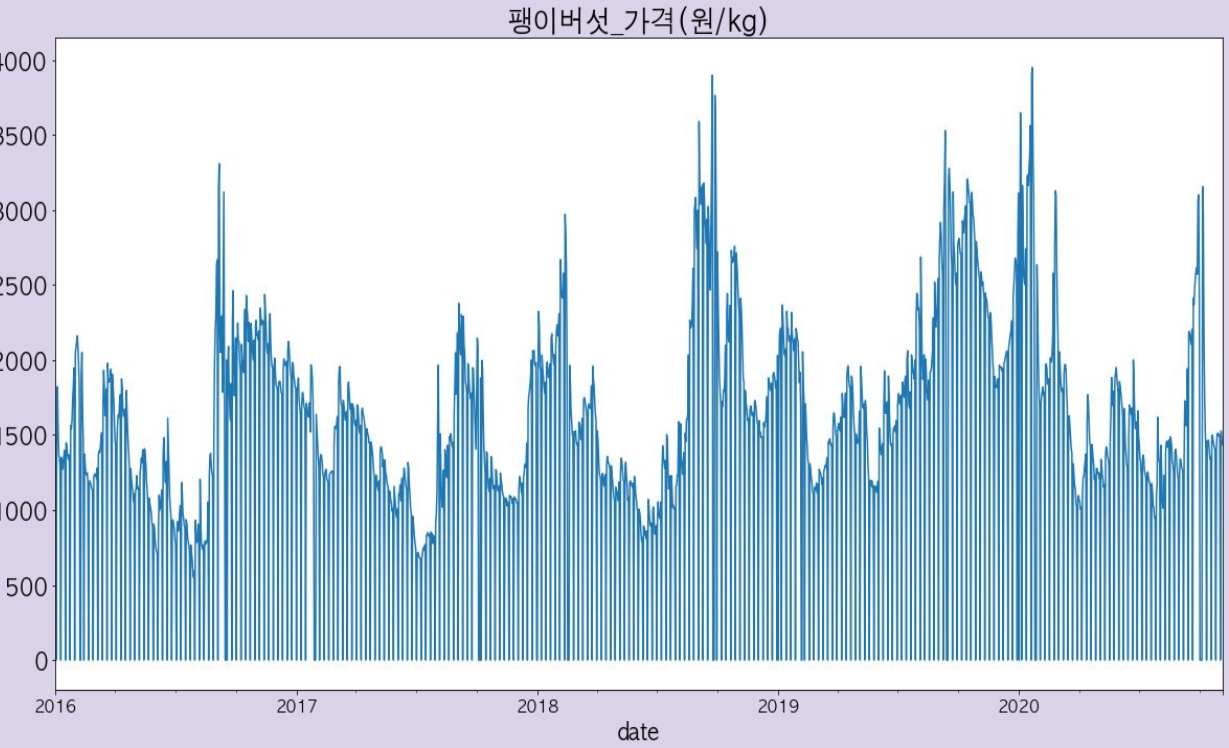


출처 가격 109% 오른 토마토에 전전긍긍 자영업자…"토마토 주스 안해요" (fntimes.c

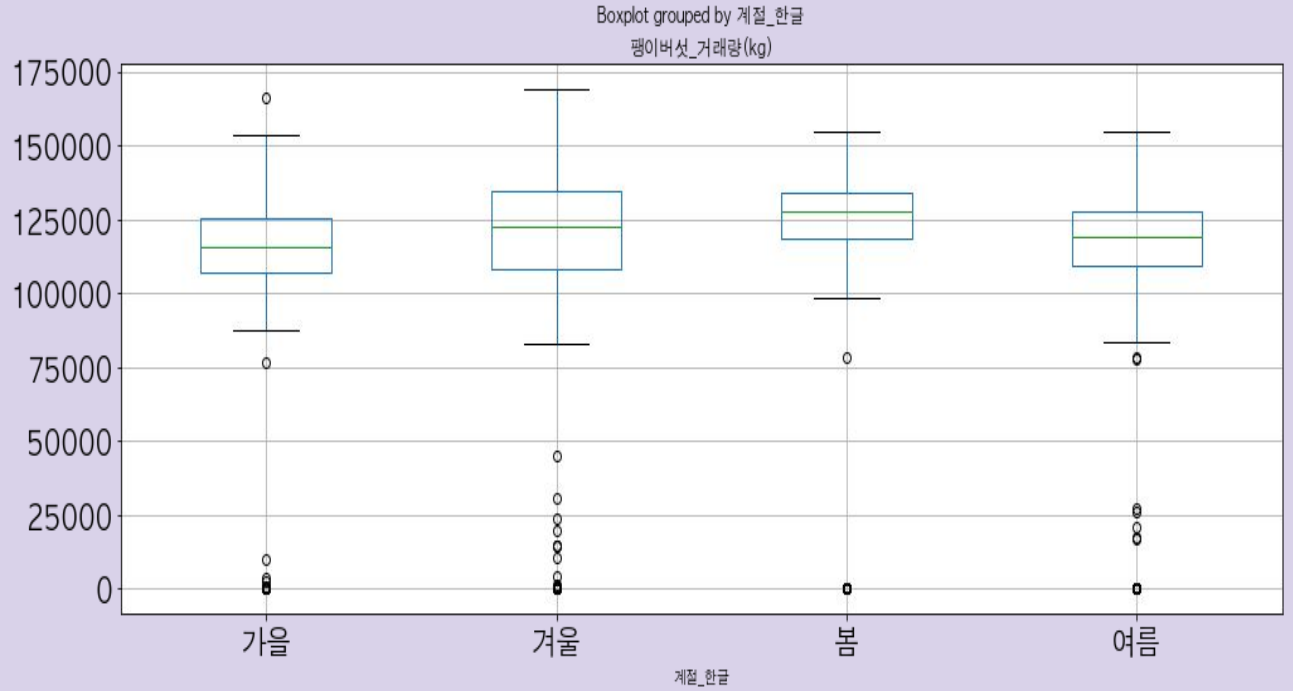
③ EDA(Exploratory Data Analysis)

팽이버섯 기간 : 2016년 1월 1일 ~ 2020년 11월 4일

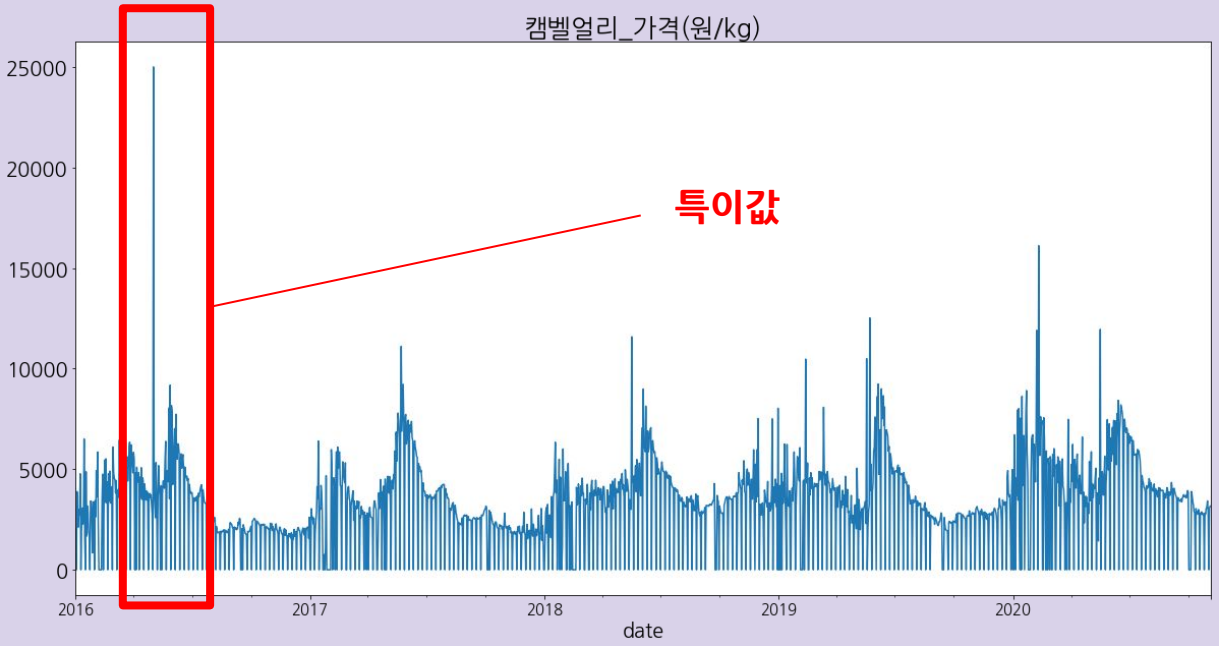
가격



거래량



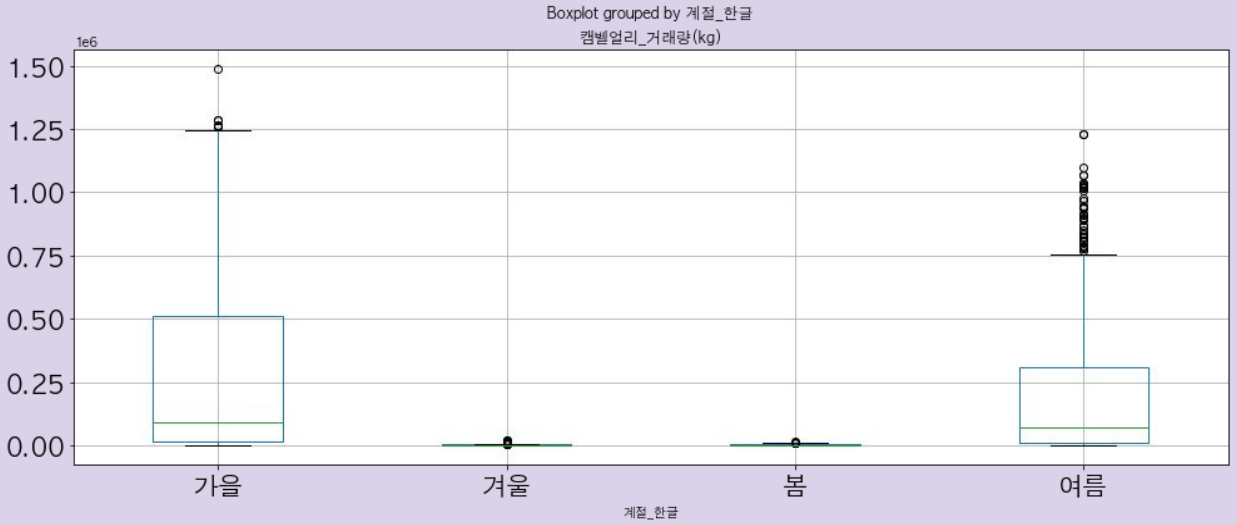
③ EDA(Exploratory Data Analysis)



캠벨얼리(포도)

2016년도 중반 가격 상승이유 >

FTA로 국내 포도 농가 급감 -> 재배면적 줄어들면서 가격 상승

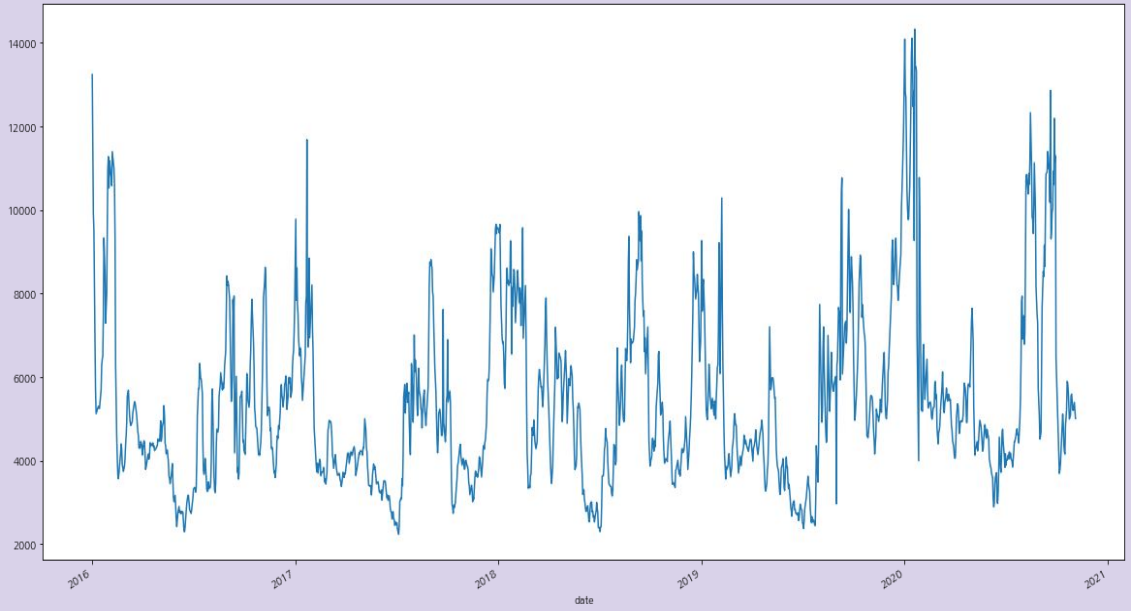


기간 : 2016년 1월 1일 ~ 2020년 11월 04일

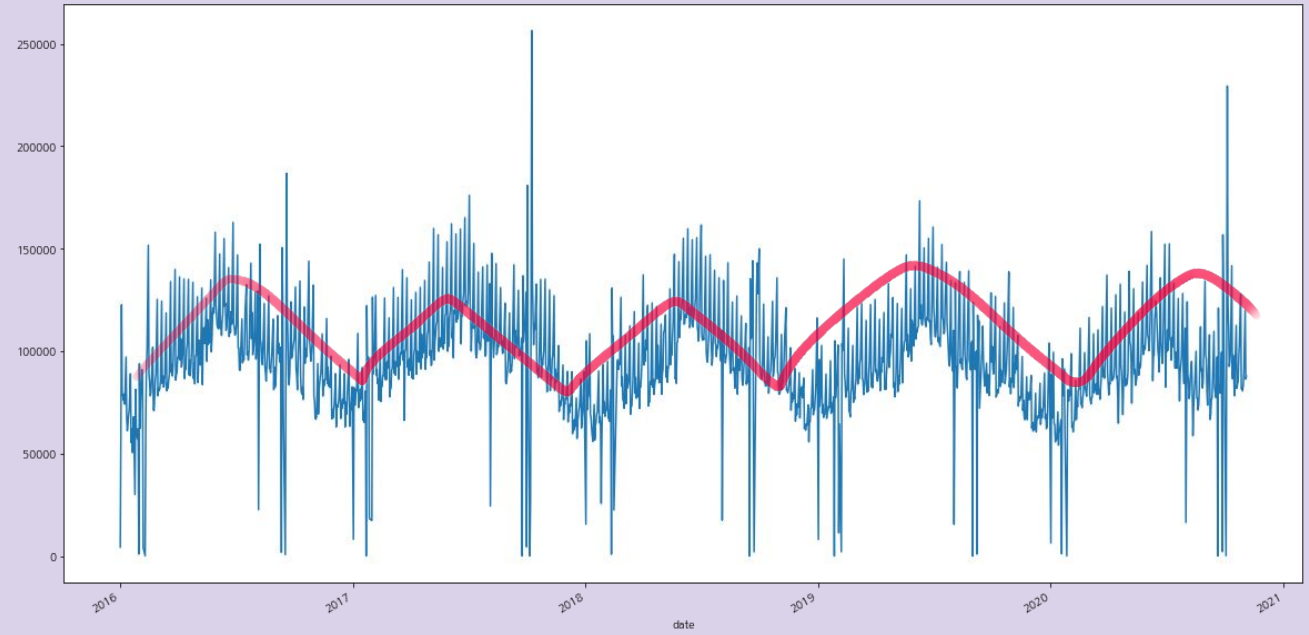
③ EDA(Exploratory Data Analysis)

깻잎

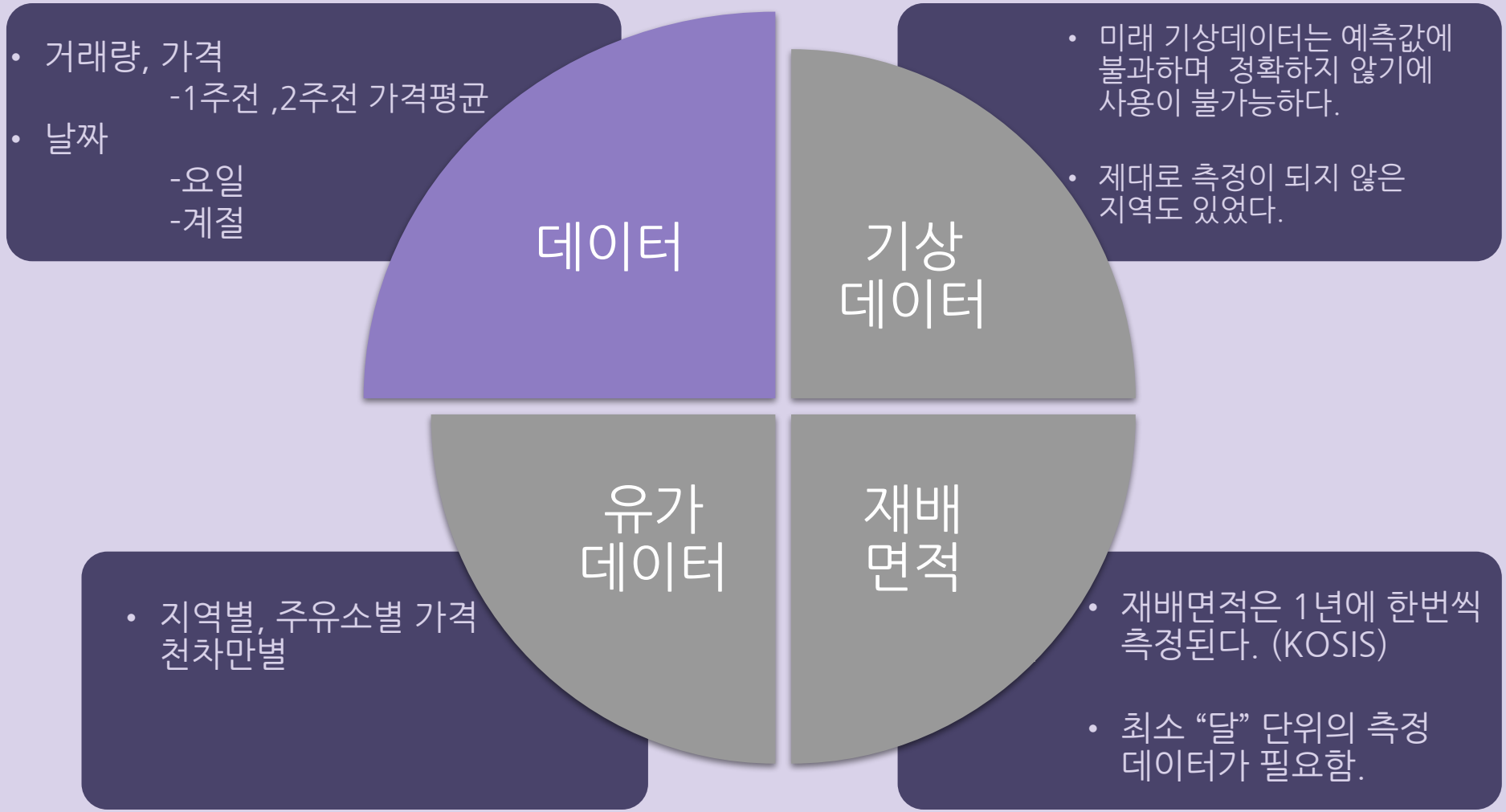
기간 : 2016년 1월 1일 ~ 2020년 11월 04일



주기성을 파악하기 어려움.



③ EDA(Exploratory Data Analysis)



- 거래량, 가격
  -1주전 ,2주전 가격평균
- 날짜
  -요일
  -계절

- 미래 기상데이터는 예측값에 불과하며 정확하지 않기에 사용이 불가능하다.
- 제대로 측정이 되지 않은 지역도 있었다.

- 지역별, 주유소별 가격 천차만별

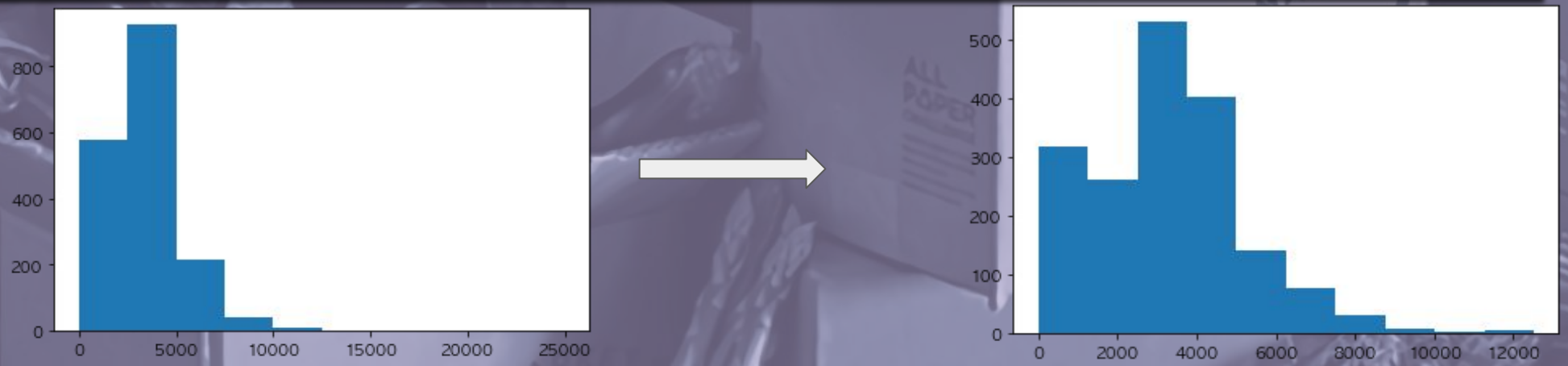
- 재배면적은 1년에 한번씩 측정된다. (KOSIS)
- 최소 “달” 단위의 측정 데이터가 필요함.

④ Data Pre-Processing

결측치 처리

거래량, 가격 데이터에 NAN값은 없음.

이상치 처리



캠벨얼리 가격 히스토그램 (예시)

⑤ Modeling

Regression Metric

RMSE(Root Mean Squared

$$RMSE = \sqrt{\frac{\sum (y - \hat{y})^2}{n}}$$

MSE에 Root를 씌워 에러를 제곱해서 생기는
값이 왜곡될 가능성 적음

⑤ Modeling

다음 3가지 모델링을 시도해서 가장 성능이 좋은 모델을 찾는다.

~~ARIMA~~ ~~XGBOOST~~ ~~Random Forest~~

[Prophet]

트렌드와 **주기적 특성**, 예외적이고 **이벤트**와 같은 휴가철 상황까지도 모델링 가능

-사용방법이 간편하고 적은 정보만으로도 예측가능

[LightGBM]

XGBoost 장점 + 성능(시간 최적화)
메모리 사용량 적음

적은 데이터셋 오버피팅 발생 가능

[LSTM]

단기 메모리와 장기 메모리를 나눠 학습, 두 메모리를 병합해 이벤트 확률 예측

→ 과거의 정보를 훨씬 잘 반영함

⑤ Modeling

다음 3가지 모델링을 시도해서 가장 성능이 좋은 모델을 찾는다.

[Prophet]

다른 모델링과 다른 이질적인 특성

Y값과 DS(시간) 데이터만 가지고 추측하기에 다른 변수를 넣을수가 없어 배제함

[LightGBM]

XGBoost 장점 + 성능(시간 최적화)
메모리 사용량 적음

적은 데이터셋 오버피팅 발생 가능

[LSTM]

단기 메모리와 장기 메모리를 나눠 학습, 두 메모리를 병합해 이벤트 확률 예측

→ 과거의 정보를 훨씬 잘 반영함

⑥ Optimization

Train / Test 단계에서의 각 모델 **베이스라인** 점수 (RMSE)

	토마토	캠벨얼리	팽이버섯	깻잎	대파
LGBM	1285.57	237.75	282.07	342.41	98.26
LSTM	1024.06	120.56	29.81	242.51	130.69

⑥ Optimization

Scaler

StandardScaler MinMaxScaler **RobustScaler** MaxAbsScaler

$$\frac{X_i - X_{med}}{Q_3(X_i) - Q_1(X_i)}$$

이상치(outlier) 영향 최소화

IQR = Q3(75%) – Q1(25%)

⑥ Optimization

날짜 (Date)

요일
(월화수목금토일)

계절
(봄,여름,가을,겨울)

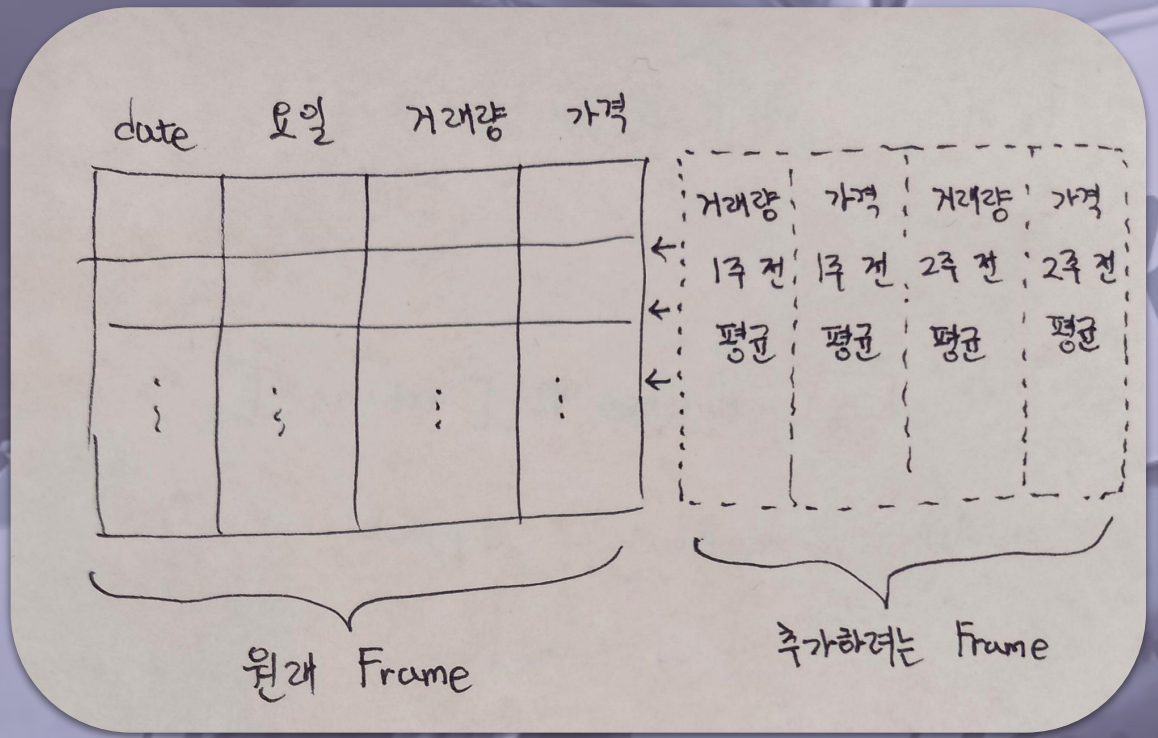
계절	요일_금요일	요일_목요일	요일_수요일	요일_월요일	요일_일요일	요일_토요일	요일_화요일
-2	1	0	0	0	0	0	0
-2	0	0	0	0	0	1	0
-2	0	0	0	0	1	0	0

⑥ Optimization

미래 가격을 예측하기 위해서는 과거의 흐름이 중요

단순 과거 거래량/가격으로는 모델 성능 향상을 기대하기 어려움

1주 전과 2주 전 가격/거래량 평균 추가



⑥ Optimization

	date	요일	대파_거래량(kg)	대파_가격(원/kg)
0	2016-01-01	금요일	0.0	0.0
1	2016-01-02	토요일	92334.0	1704.0
2	2016-01-03	일요일	0.0	0.0
3	2016-01-04	월요일	994328.1	1716.0
4	2016-01-05	화요일	787716.0	1715.0
...	...	...	...	...

[목적]
월요일부터 일요일까지 7일 또는 14일 단위로 분류해서 평균 내는 것

=> 다루고자 하는 데이터프레임은 금요일부터 시작함

데이터프레임의 앞부분(1주 : 10row, 2주 : 17row)을 그 날 거래량과 가격으로 복사

결과

Train / Test 단계에서의 각 모델 **베이스라인** 점수 (RMSE)

	토마토	캠벨얼리	팽이버섯	깻잎	대파
LGBM	1285.57	237.75	282.07	342.41	98.26
LSTM	1024.06	120.56	29.81	242.51	130.69

Train / Test 단계에서의 각 모델 **최종** 점수 (RMSE)

	토마토	캠벨얼리	팽이버섯	깻잎	대파
LGBM	964.77	215.5	402.12	165.77	219.66
LSTM	889.6	282.42	12.66	99.24	222.11

결과

		기준일자 가격 (20201106)	1주 후 (20201113)	2주 후 (20201120)	4주 후 (20201204)
대파	LightGBM	2402.0	2182.34	1979.54	2204.46
	LSTM		2229.93	1956.23	1774.11
깻잎	LightGBM	4831.0	4665.23	4855.3	5916.4
	LSTM		4730.36	6352.39	5696.17
팽이버섯	LightGBM	1408.0	1810.12	1342.96	2212.14
	LSTM		1381.22	1449.34	1883.42
토마토	LightGBM	2806.0	3770.77	2650.97	3173.16
	LSTM		3677.17	3089.88	2235.22
포도 (캠벨얼리)	LightGBM	2797.0	3012.5	2712.13	2842.19
	LSTM		2530.52	2734.32	2842.74

⑦ 회고 및 아쉬운 점

비정형 데이터에 대한 부분

비정형 데이터는 비디오와 오디오 등의 정보를 포함하므로 이를 텍스트로 변환한 후, 자연어 처리 기법으로 언급하는 빈도를 계산하고, 감정 단어를 구분해서 긍부정 데이터로 활용...

현재 수준을 뛰어넘는 난이도이므로 시도하지 못했다.

우리 조 모델에 비정형 데이터 처리 기술을 사용해서 예측을 한다면 더 좋은 성능이 나올 것으로 예상된다.

종속 변수 선택에 대한 부분

Dacon 경진 대회에 있는 농산물이 아닌 다른 농산물들은 과거 가격 데이터를 불러오는 데 있어 문제가 생겨 선택하지 못했다.
(농넷 API 막힘)

⑦ 회고 및 아쉬운 점

모델링 과정에 대한 부분

17page에서 소개했던 모델은 3가지였는데 베이스라인 점수를 산출하는 과정부터 Prophet은 빠져있다.

LightGBM과 LSTM과 달리 Prophet은 생소한 모델이었고 코드 구현 또한 어려웠기에 진행하지 못했다.

만약 Prophet 모델을 돌릴 수 있는 실력을 가지고 앞서 소개한 Pre-processing까지 한다면 더 좋은 성능을 기대해 볼 수 있겠다.

프로젝트 진행 전체

컬리 담당자와의 미팅날은 10월 14일이었는데, 이 날 피드백에 의해 이전까지 진행했던 것의 8할을 갈아엎어야 했다.

교육생 신분으로서 생각하는 프로젝트 목표와, 기업인이 생각하는 프로젝트 목표에 있어 차이가 크다는 것을 알게 되었다.

담당자 미팅을 좀 더 일찍 진행했다면 어땠을까 하는 아쉬움이 들었다.

⑧ 데이터 출처

데이터	출처	주기	데이터 타입
농산물 가격	농넷	일	csv
날씨(기온, 강수량, 습도) 농업기상데이터	기상청 종관기상관측(ASOS)	일	csv
유가	오피넷	일	csv
소비자 물가 지수	KOSIS	월	csv
농산물 수출입	KATI(한국무역통계진흥원)	월	csv
재배면적	KOSIS(국가통계포털)	년	csv
블루베리	농넷	년	csv
실업률	KOSIS	년	csv

참고문헌:

1. “공공 데이터를 이용한 농산물 가격 예측 시스템 개발” https://github.com/seongmoonKang/Data-Analysis-Capstone

2.https://github.com/Kimsejin97/MJ_Capstone

3.https://dacon.io/competitions/official/235801/codeshare/4063?page=1&dtype=recent

https://lightgbm.readthedocs.io/en/v3.3.2/

- LightGBM

https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html

https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM

- LSTM

https://facebook.github.io/prophet/

- PROPHET


Q n A