

스파르타코딩 데이터 분석 2기
[기초 분석] 3조 0101

2023년 가격 예측을 위한

제주도 특산품의 가격과 연관 변수간 상관관계 분석

팀장: 서영석

팀원: 서민혁, 이민지, 이연수, 이유정

목차

1. 프로젝트 소개

2. 기본 데이터 분석

- 1) 기본 데이터 소개
- 2) 기본 데이터 EDA

3. 추가 데이터 수집 및 분석

3-1. 가락동 경매 가격으로 제주 농산물의 가격 예측하기

- 1) 데이터 소개
- 2) 데이터 전처리 & 분석

3-2. 특산품 유통 가격과 제주도 기상 데이터의 상관관계 분석하기

- 1) 데이터 소개
- 2) 데이터 전처리 & 분석

4. 결론 및 회고

- 1) 결과 정리 및 결론 도출
- 2) 프로젝트 회고

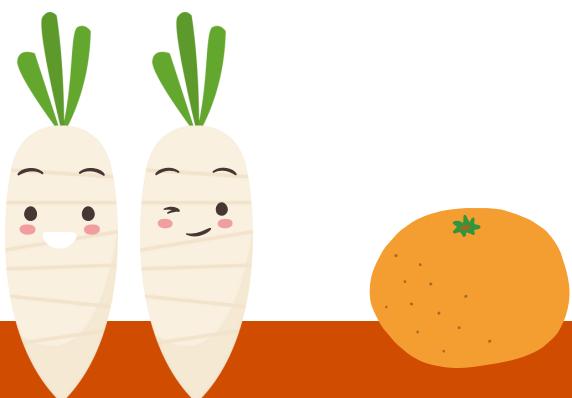
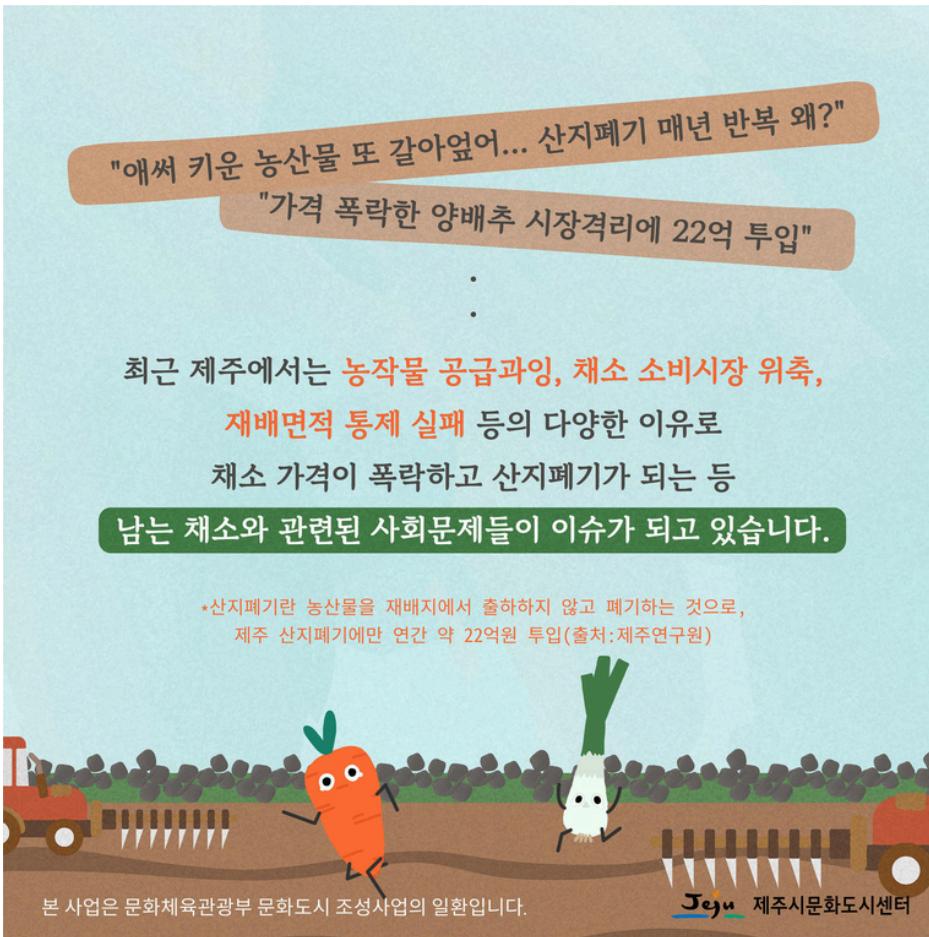


PART 01
프로젝트 소개



PART 01 프로젝트 소개

프로젝트 배경



양배추, 무(월동무), 당근, 브로콜리, 감귤은 제주도의 대표적인 특산품입니다. 최근 제주도에서는 농산품 시장의 공급과 수요의 불균형으로 인하여 산지폐기로 인한 환경오염 및 세금낭비 등 문제가 발생하고 있습니다.

**특산품의 안정적이고 효율적인 수급을 위해
특산품 가격에 대한 관련 요인 분석과 예측이 필요합니다.**



PART 01 프로젝트 소개

프로젝트명

2023년 수익률 예측을 위한 품목별 가격과 변수 간에 상관관계 분석

목표

1. 품목별 가격과 변수 간 상관관계 분석
2. 품목별 매출 추이 및 가격 예측

가설

1. 가격을 기준 하여 특산품 등급 분류 시, 가격은 유의미한 추이를 보일 것이다.
2. 특산품의 재배 기간 기후에 따라 가격에도 영향을 미칠 것이다.

핵심 내용

- 관련 변수 및 데이터 탐색
- 데이터 품질 검증 및 전처리
- 가격과 변수 간 상관관계 분석 및 시각화
- 가격과 추이(시계열) 분석 및 시각화

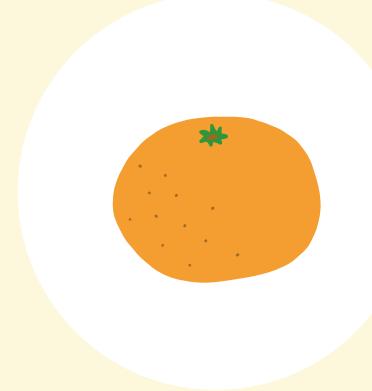
기초 도메인 조사

Q. 각 특산품별 재배 시기 및 적정 재배 환경



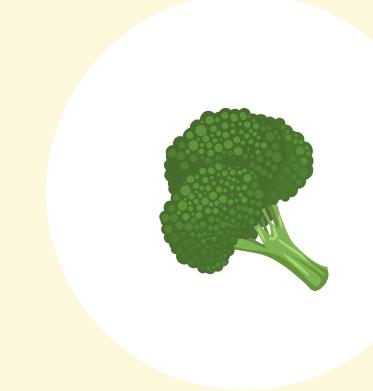
무 (RD)

- 파종 : 9월
- 출하 : 12~1월
- 생육온도 :
 - 발아적온 15~30°C
 - 개화적온 12°C 이하
 - 생육적온 20°C
- 재배적지 :
 - 토양산도 pH 5.5~6.8



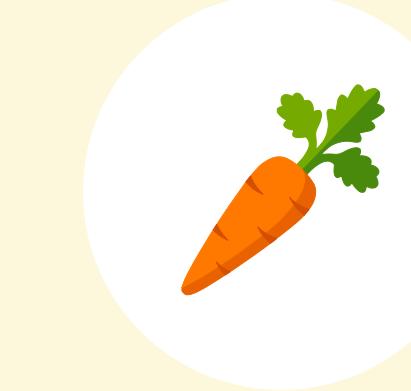
감귤 (TG)

- 발아 :
 - 만감류 : 10월
 - 노지감귤 : 4월
- 출하 :
 - 만감류 : 2~9월
 - 노지감귤 : 10~2월
- 생육온도 :
 - 발아적온 15°C 이상



브로콜리 (BC)

- 파종: 7~8월
- 출하 : 10~11월
- 생육온도 :
 - 발아적온 25°C 이상
 - 생육적온 18~20°C
 - 저장적온 0°C
- 재배적지 :
 - 토양산도 pH 6.0



당근 (CR)

- 파종: 7~8월
- 출하 : 11~3월
- 생육온도 :
 - 발아적온 15~30°C
 - 생육적온 15~20°C
 - 근비대적온 16~21°C
- 재배적지 :
 - 토양산도 pH 6.0~6.5



양배추 (CB)

- 파종: 7~8월
- 출하 : 12~1월
- 생육온도 :
 - 발아적온 25°C 이상
 - 생육적온 15~20°C
 - 저장적온 0~3°C
- 재배적지 :
 - 토양산도 pH 5.5~6.8

PART 02

기본 데이터 분석

기본 데이터 소개

[Main] 제주 특산품 국내 유통 데이터

- 데이터 크기: 59397 (rows) X 13 (columns)
- 데이터 측정 범위: 2019-01-01 ~ 2023-03-03 (1523일)
- 데이터 결측치 : 없음
- 데이터 이상치
 - Supply가 있으나 Price가 0인 데이터 3건 발견

변수명	변수 설명	비고
ID	유통 ID	'item_corporation_location_timestamp' 형태
Timestamp	날짜	'yyyy-mm-dd' 형태 20190101 - 20230303
Item	유통 품목	TG: 감귤/ BC: 브로콜리 RD: 무/ CR: 당근/ CB: 양배추
Corporation	유통 법인	A-F
Location	지역	J: 제주시/ S: 서귀포시
Supply(kg)	유통량	시각화 가독성을 위하여 100kg 단위로 변경
Price(원/kg)	유통당 1kg 당 가격	시각화 가독성을 위하여 원/100kg 단위로 변경

[데이터 명세]

기본 데이터 EDA/전처리

“데이터의 양이 많은 건 좋은데, 2Q 지표까지 0이라니..”

[기본 데이터 정보]

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59397 entries, 0 to 59396
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
 ---  --          -----          --    
 0   ID           59397 non-null   object  
 1   timestamp    59397 non-null   object  
 2   item          59397 non-null   object  
 3   corporation   59397 non-null   object  
 4   location      59397 non-null   object  
 5   supply(kg)   59397 non-null   float64 
 6   price(원/kg) 59397 non-null   float64 
dtypes: float64(2), object(5)
memory usage: 3.2+ MB
```

[기본 데이터 기술 통계량]

	supply(kg)	price(원/kg)
count	59,397.0	59,397.0
mean	11,894.5	1,131.7
std	52,264.0	2,029.9
min	0.0	0.0
25%	0.0	0.0
50%	0.0	0.0
75%	3,800.0	1,519.0
max	1,222,800.0	20,909.0

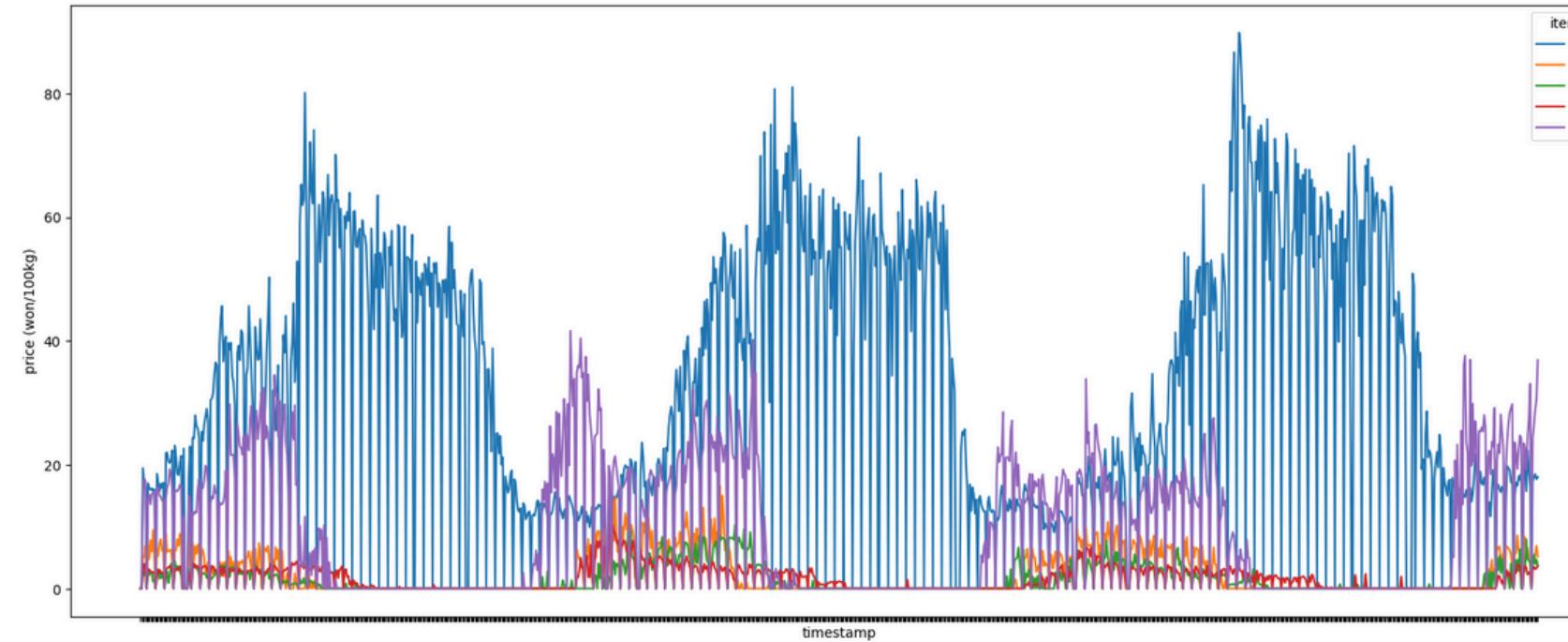


데이터 전처리

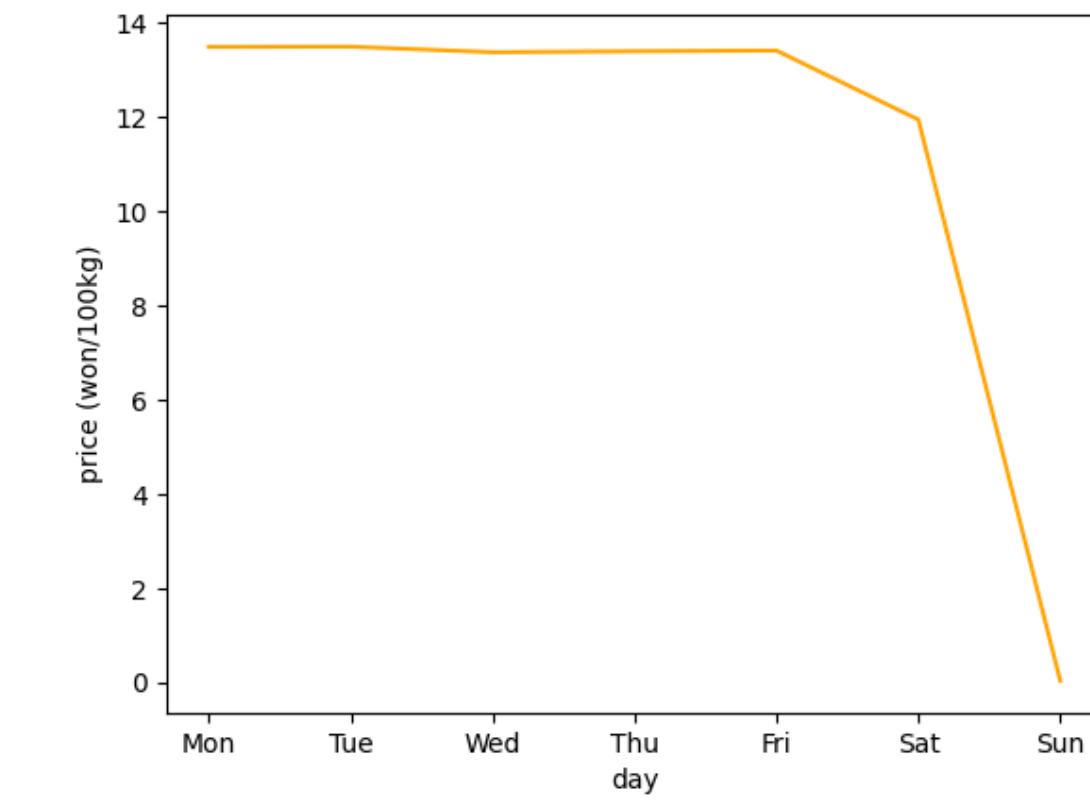
- 유통량이 0일 경우 가격도 0이기 때문에 속성값이 0인 행에 대한 조치 필요
- 기준 단위 스케일링
 - Supply : 1Kg > 100Kg
 - Price : 원/Kg > 원/100Kg

기본 데이터 EDA/전처리

Q. 시간의 흐름에 따라 각 품목의 유통량과 가격 추이는 어떨까?



< 특산품별 가격 추이 - 일 기준 >



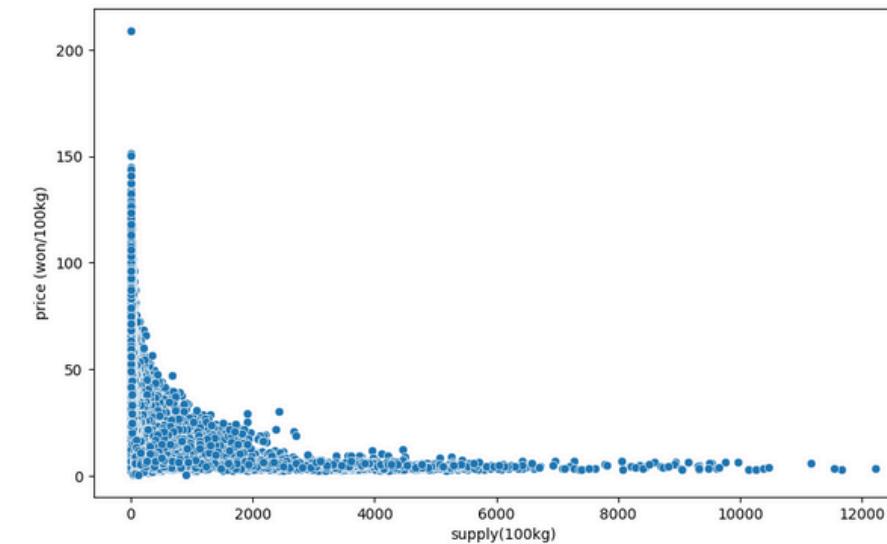
< 요일별 가격 - 요일 기준 >

Insight & Action Memo

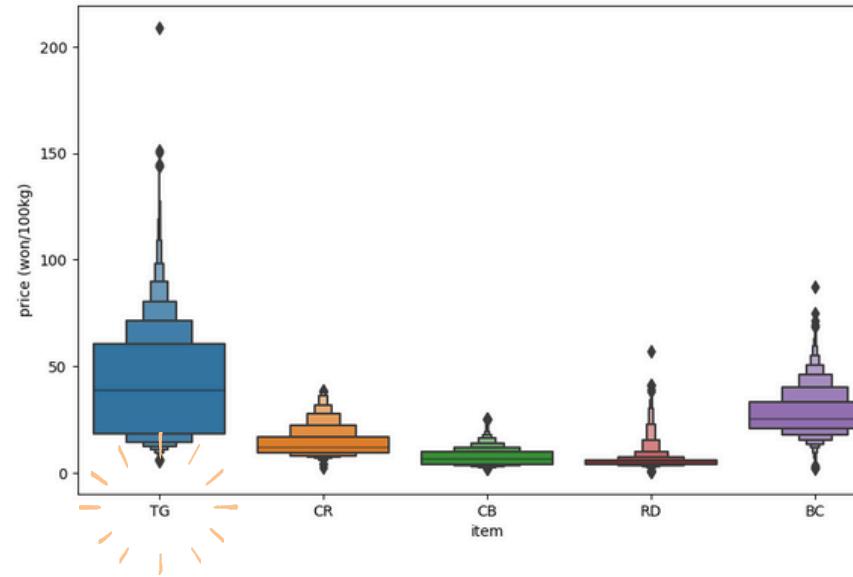
- 일별 가격 추이를 시각화한 결과, 계절성이 확인되어 연도(year), 월(month), 주(week), 요일(day), 연-월(yemon) 컬럼을 추가하였다.
- 특산품마다 각각의 재배/출하 시기가 있기 때문에 가격 추이는 각각 다른 패턴을 갖는다.
- 일요일, 대부분의 공휴일에는 price 컬럼이 0이다.
 - [holidays] 라이브러리 사용하여 공휴일에는 1, 아닌 날엔 0으로 분류하는 컬럼을 추가하였다.
 - 공휴일로 분류된 속성값 중 price 데이터가 0이 아닌 경우, 해당 데이터의 holiday 컬럼 0으로 변경하였다.

기본 데이터 EDA/전처리

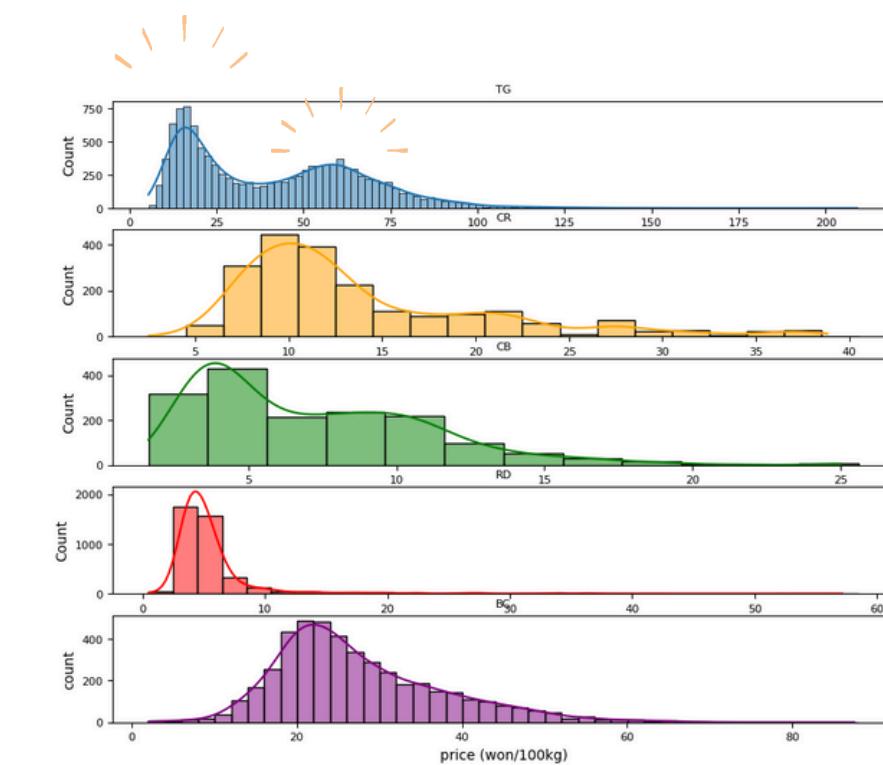
Q. 가격은 어떤 분포를 이루고 있을까?



< 가격/유통량 분포 산점도 >



< 특산품별 가격 분포 박스플롯 >



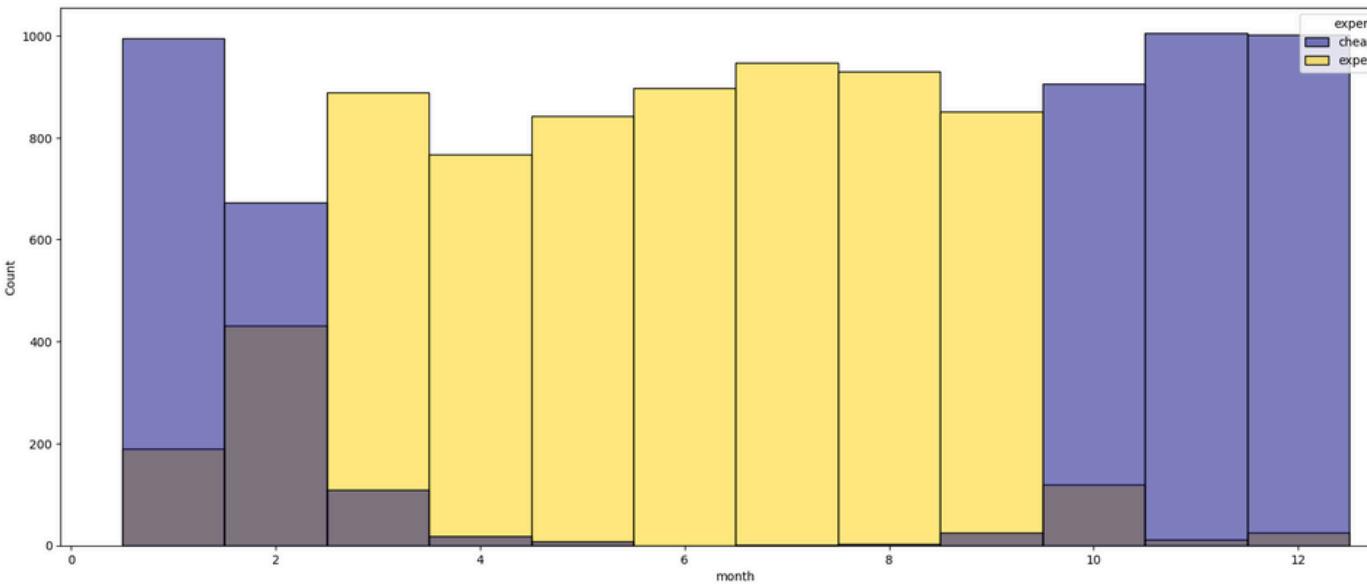
< 특산품별 가격 분포 히스토그램 >

Insight & Action Memo

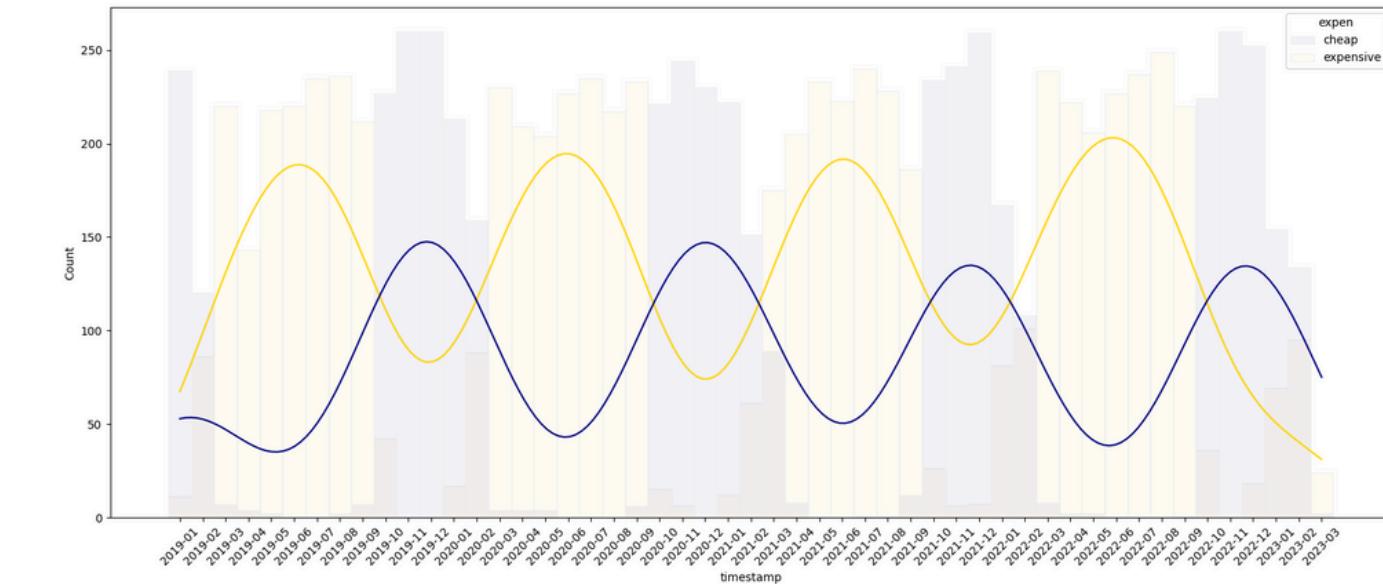
- 가격과 유통량은 약한 음의 상관관계를 보인다.
 - 감귤의 가격은 다른 특산품의 가격에 비해 넓은 폭의 분포를 보였다.
 - 감귤을 제외한 4종의 특산품은 단일 분포의 모양을 볼 수 있다. 감귤의 분포는 두개의 분포 (정규(Gaussian) 분포로 추정) 가 혼합된 모습을 볼 수 있다.
- 위와 같은 현상은 2개의 감귤의 재배종이 섞이면서 발생한 것으로 추론하였다.

기본 데이터 EDA/전처리

Q. 가격을 기준으로 감귤의 종류를 2개로 나눌 수 있을까?



< 감귤 그룹별 월 기준 가격 분포 - 히스토그램 >



< 감귤 그룹별 '연-월' 기준 가격 분포 추이 - 커널 밀도 추정 >

Insight & Action Memo

- 사이킷런의 GMM(Gaussian Mixture Model)을 사용하여 가격을 기준으로 감귤을 2개의 클러스터로 분류하였다.
- 분류된 군집에 대하여 가격 추이를 살펴본 결과 각각의 계절성 패턴을 확인할 수 있었다.
감귤의 특성을 조사한 자료를 참고했을 때 expensive 군집은 만감류(2~9월 출하)로 유추할 수 있고,
cheap 군집은 노지감귤(10~2월 출하)로 유추할 수 있다.

PART 03

데이터 추가 수집 및 분석

데이터 소개

대부분의 농작물은 재배 기간 기온, 강수량 등 기후에 따라 생산량과 가격에 직간접적으로 영향을 많이 받는다.
제주도 특산물의 유통 가격 또한 재배 기간 제주도의 기후와 관계가 있을 것 같으니 상관관계를 분석해볼까?



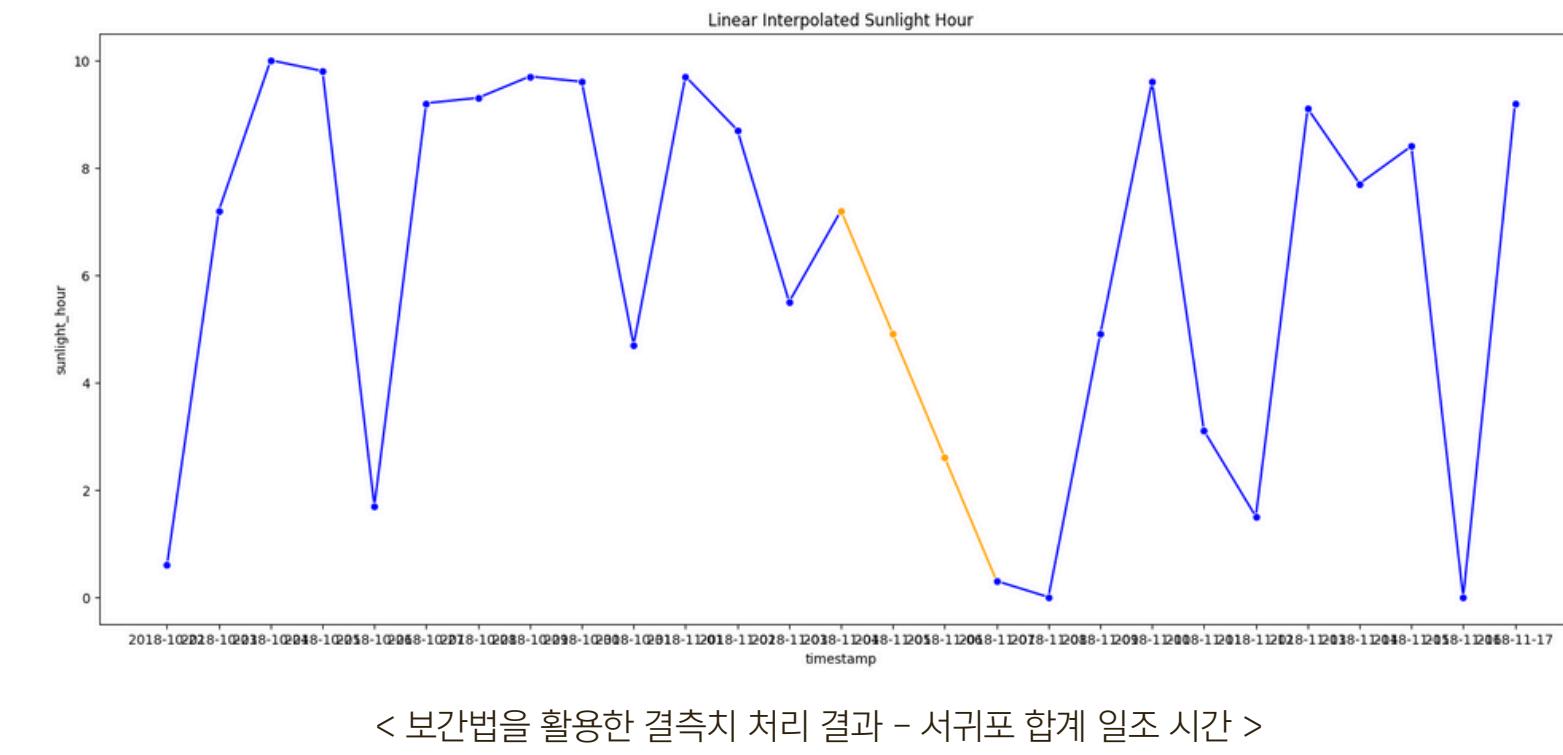
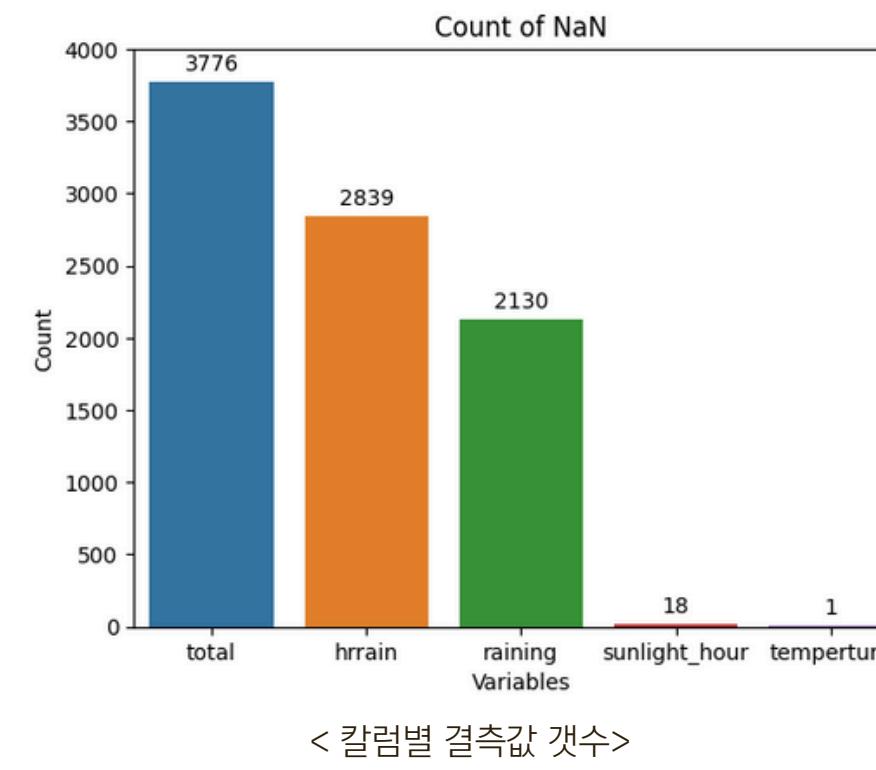
기후 데이터

- 수집 경로 : 공공데이터 포털 내 기상청 지상 일자료 조회서비스
- 데이터 크기: 3776 (rows) X 14 (columns)
- 데이터 수집 범위 : 2018-01-01 ~ 2023-03-03 (1888일)

변수명	변수 설명	비고
station	지점명	J: 제주시, S: 서귀포시
timestamp	날짜	'yyyy-mm-dd' 형태
itemperature	평균 기온	결측치 1개
mintem	최저 기온	-
maxtem	최고 기온	-
raining	일강수량	결측치 2130개
harrain	1시간 최다 강수량	결측치 2839개
wind_speed	풍속	-
sunlight_hour	합계 일조 시간	-

[데이터 명세]

데이터 전처리

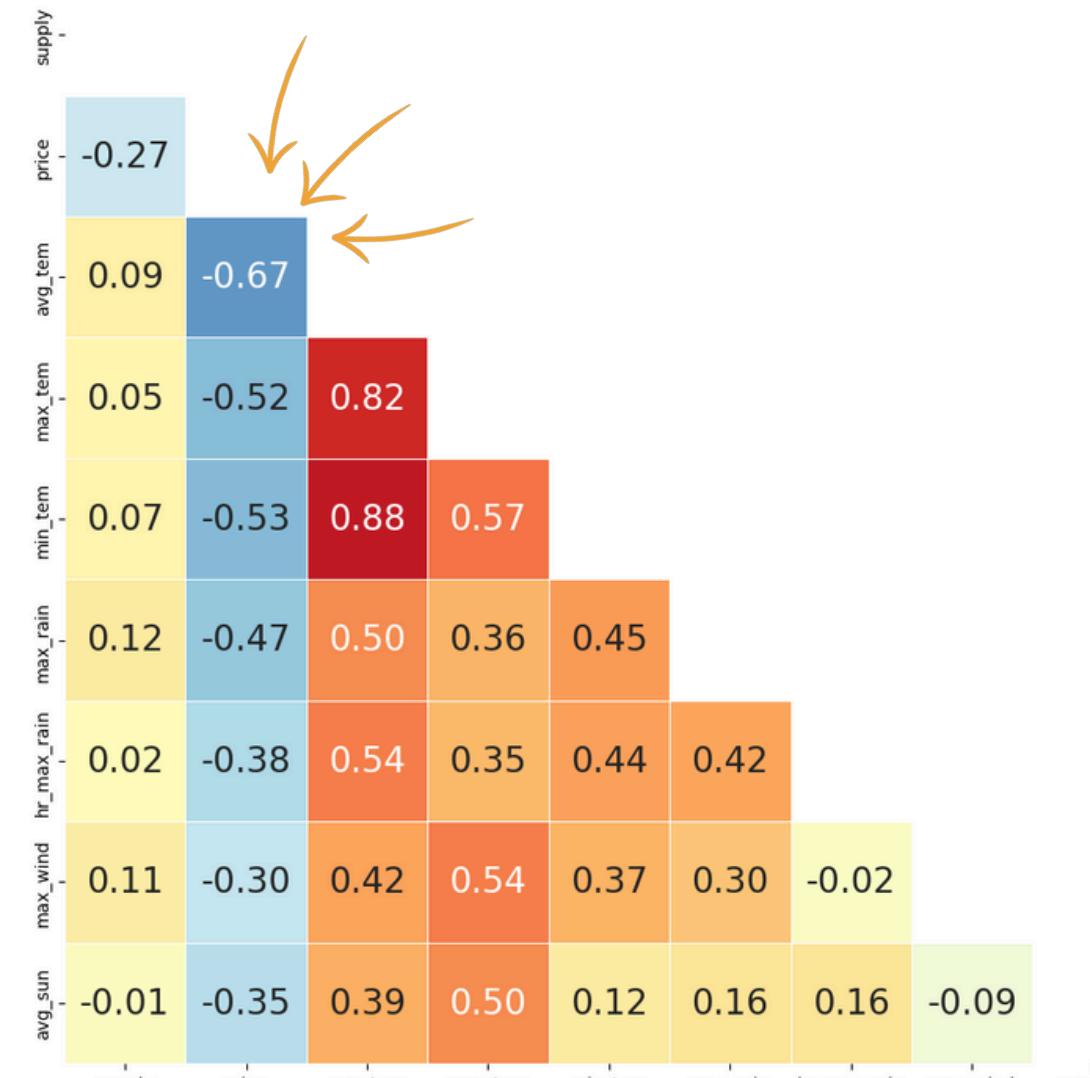


Insight & Action Memo

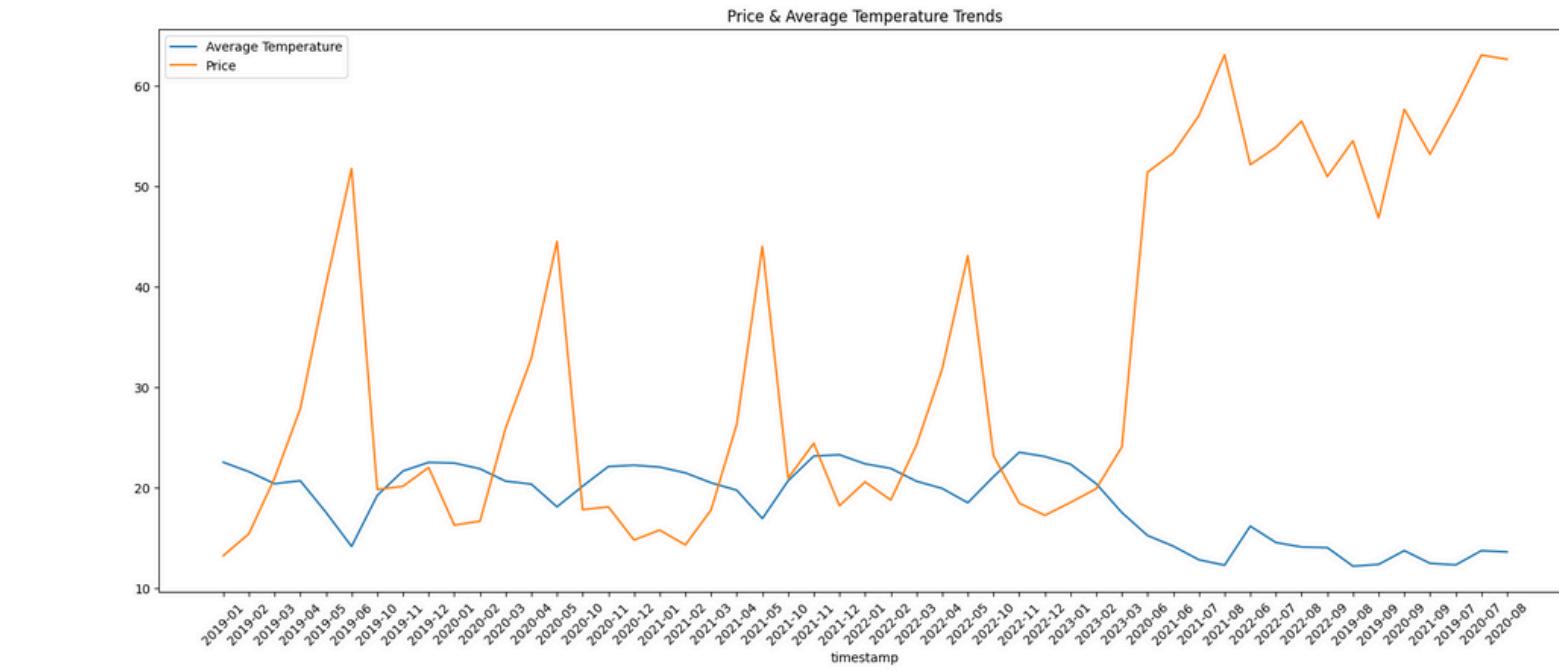
- 최대강수량, 일강수량, 합계, 일조시간, 평균기온 컬럼에서 결측치를 발견했다.
- 선형보간법(Linear Interpolation)을 활용하여 결측치를 채워주었다.
- 결측치 비중이 높은 최대강수량과 일강수량의 경우, 결측치를 희석시킬 수 있도록 제한된 분석을 진행 예정이다
 - 각 특산물의 재배기간을 고려하여 분석 진행 시 약 1분기 단위로 시계열을 그룹화할 예정이다.
 - 분석 시 제한된 집계함수(Max, Mean 등)을 활용할 예정이다.

데이터 분석

Q. 기상 데이터의 기초 통계와 전체 유통량/가격과 유의미한 상관관계가 있을까?



< 기상 데이터의 기초 통계와 유통량/가격 히트맵 >

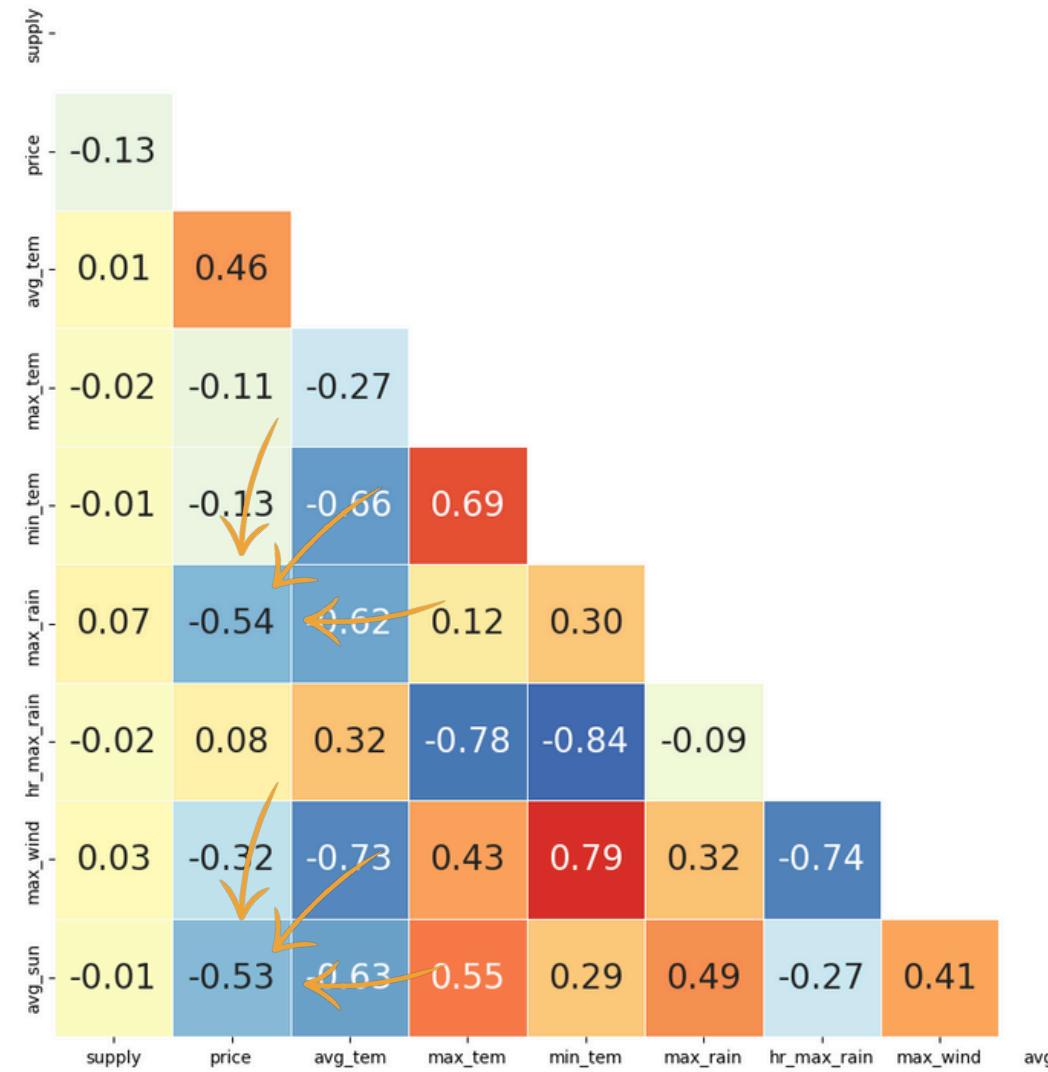


Insight & Action Memo

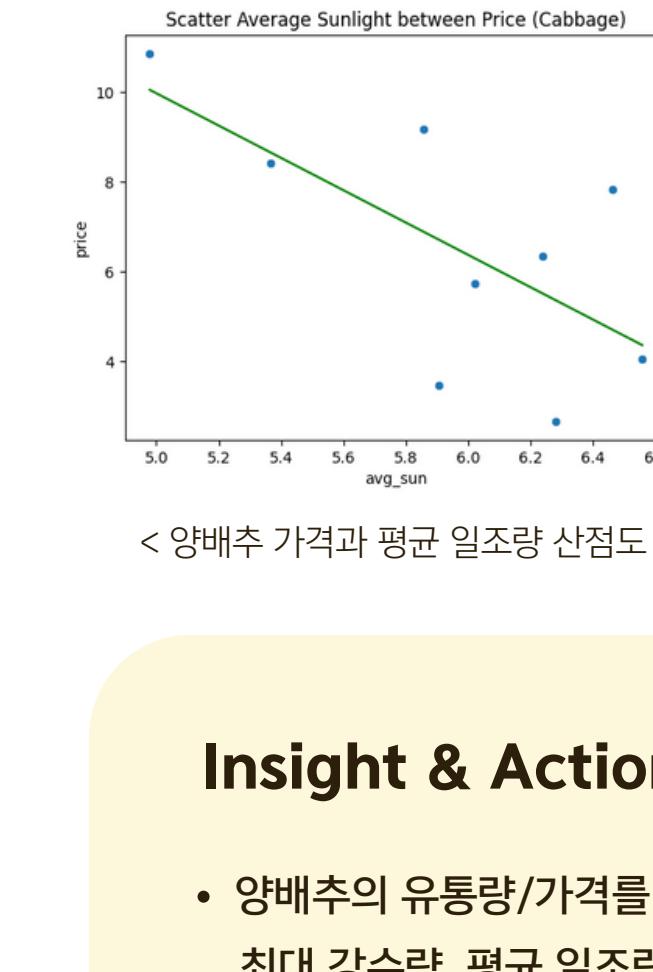
- 각 특산물 별 재배기간 동안의 평균 기온, 최고 기온, 최저 기온, 1일 최다 강수량, 최대 풍속, 일조시간 평균 값을 추출하여 대치하였다.
- 이 중 가격 데이터와 평균 기온 간에 유의미한 음의 상관관계가 발견되었다.
- 각 특산품별 상관관계도 살펴볼 필요가 있다.

데이터 분석

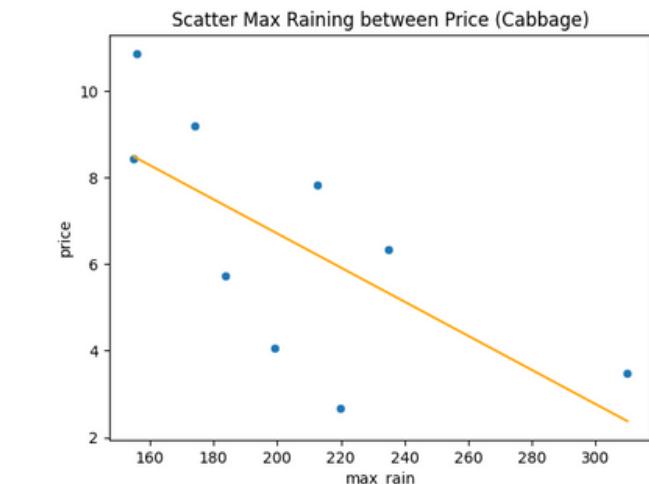
Q. 기상 데이터의 기초 통계와 각 특산품의 유통량/가격과 유의미한 상관관계가 있을까?



< 기상 데이터의 기초 통계와 양배추의 유통량/가격 히트맵 >



< 양배추 가격과 평균 일조량 산점도 >



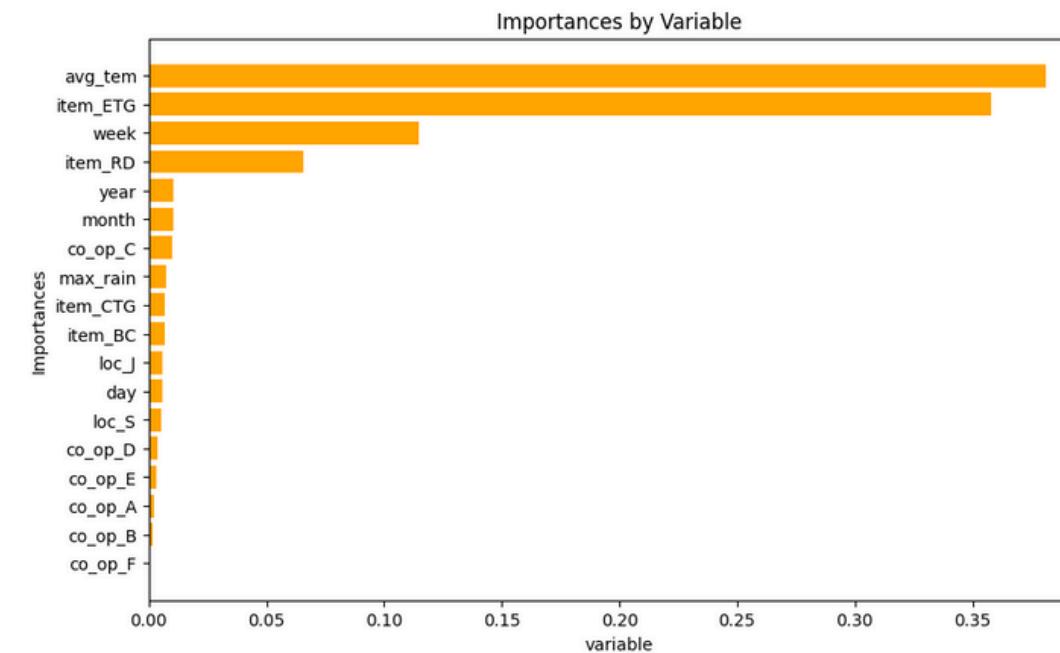
< 양배추 가격과 최대 강수량 산점도 >

Insight & Action Memo

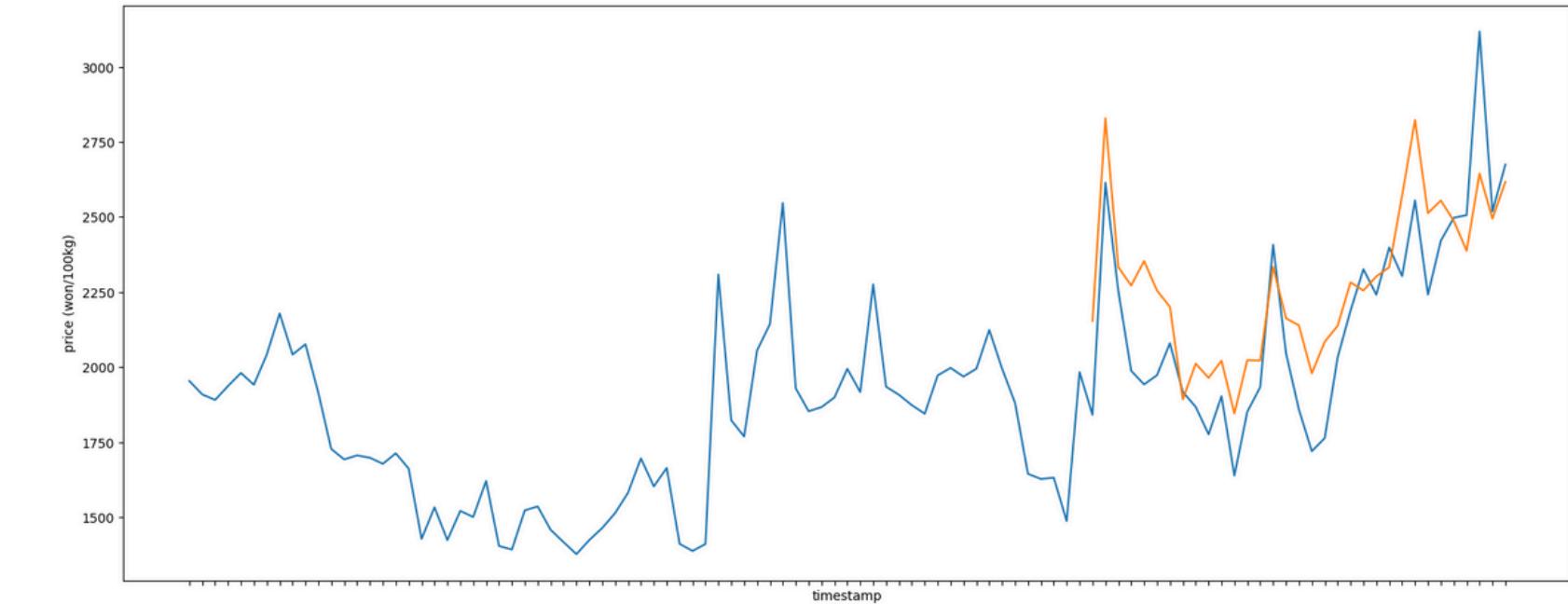
- 양배추의 유통량/가격을 기상 데이터와 대치한 결과
최대 강수량, 평균 일조량과 가격 사이에 유의미한 음의 상관계수를 볼 수 있었다.
이 때 ± 0.5 이상/이하의 값을 유의미하다고 판단하였다.
- 최대 강수량과 평균 일조량을 각각 X 축으로 하는 산점도를 그렸을 때,
우하향하는 추이를 볼 수 있다.

예측 모델링

Q. 상관관계가 있는 기상 요인을 활용하여 가격 예측 모델링을 만들어보자



< 변수 별 가중치 설정 >



< 모델링 결과와 실제 가격 그래프 비교 >

Insight & Action Memo

- 2019–2022년 데이터를 학습 데이터와 검증 데이터, 2023년 데이터를 테스트 데이터로 구분하였다.
- 기상 데이터는 MinMax 정규화 진행 / 범주형 변수는 One-Hot Encoding 진행
- Random Forest Regressor Model을 활용하여 Grid Search를 통해 최적의 하이퍼 파라미터를 도출한 결과로 학습을 진행하며, 그 결과 평균 제곱근 오차값(RMSE)이 약 650 / 결정 계수(R-squared)가 약 0.93 인 모델을 만들 수 있었다.

2023년 가격 예측을 위한

제주도 특산품의 가격과 연관 변수간 상관관계 분석

FOOD