2021 빅콘테스트 공모전 퓨처스리그 - ECO제주

제주도 음식물 쓰레기양 예측을 통한 배출량 감소 방안 도출

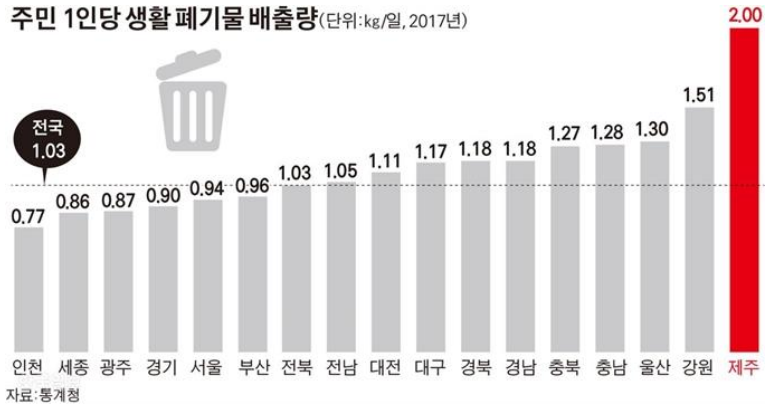
TEAM 넥스트 레벨

조유민 rollingill@yonsei.ac.kr
남승지 seungji07@yonsei.ac.kr
유은영 uare0@yonsei.ac.kr
조유림 rim6@yonsei.ac.kr

Contents

01 분석 배경 및 문제 정의
청정 도시 제주의 오명

지금 제주도는?



"아름다운 자연환경과 각종 볼거리로 여행객들에게 인기 있는 제주도

하지만 꾸준한 생활 폐기물 배출 증가 추세와 더불어

현재 **2019년 1인당 쓰레기배출량 전국 1**위라는 오명을 안고 있습니다."

01 분석 배경 및 문제 정의
제주도 쓰레기 배출 정책 현주소

감량기

01 분석 배경 및 문제 정의
코로나 이후의 변화

코로나의 영향은?

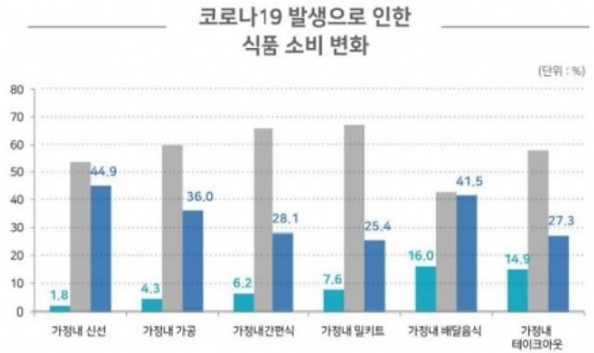
2020년 제주도 관광객 전년대비 33% 감소,
그에 비해 쓰레기 발생량 5.4% 감소

→ **관광객 감소 ≠ 배출량 감소**, 그 원인은?

KREI "2020 식품소비행태조사"

가정 내 식사 증가로 가구당 배출 음식물 쓰레기 양 10% 증가

음식물 쓰레기가 늘어났다고 응답한 비중 10.7% ➔ 16.8%



02 요인 선정 및 분석 목표
연구 가설

음식물 배출량에 영향을 줄 것으로 예상되는 3가지 요인을 설정하여 이를 바탕으로 분석 진행

1 *냉장고 안에서 시들어가는 식재료들, 버려지는 배달 잔반*

특정 업종 **카드 소비** 내역이 늘어나면 배출량이 늘어날까?

2 *천둥번개에 비 오는 날, 집 밖은 위험해*

날씨가 안 좋으면 배출량이 줄어들까?

3 *실시간 검색어는 곧 현대인들의 행동이고 생각*

포털 사이트 **검색량**과 배출량의 연관성이 높지 않을까?

02 요인 선정 및 분석 목표
분석 목표

읍면동별 음식물 쓰레기 배출량 예측을 통해서 배출 증가에 영향을 주는 핵심 요인 파악

이를 바탕으로 배출량 감축을 목표로 하는 정책 또는 제주 시민의 행동 변화를 도출할 수 있는 유인 제시

03 데이터 탐색 결과 및 변수 생성
분석 데이터 소개

03 데이터 탐색 결과 및 변수 생성
음식물 쓰레기 데이터

RFID 기반 종량제

공동주택 내 설치된 장비나 차량에 태그를 인식 후 음식물쓰레기를 배출하는 방식

배출거점별 지불금액(pay_amt)/ 쓰레기 배출량(em_g)은 0.03으로 일정

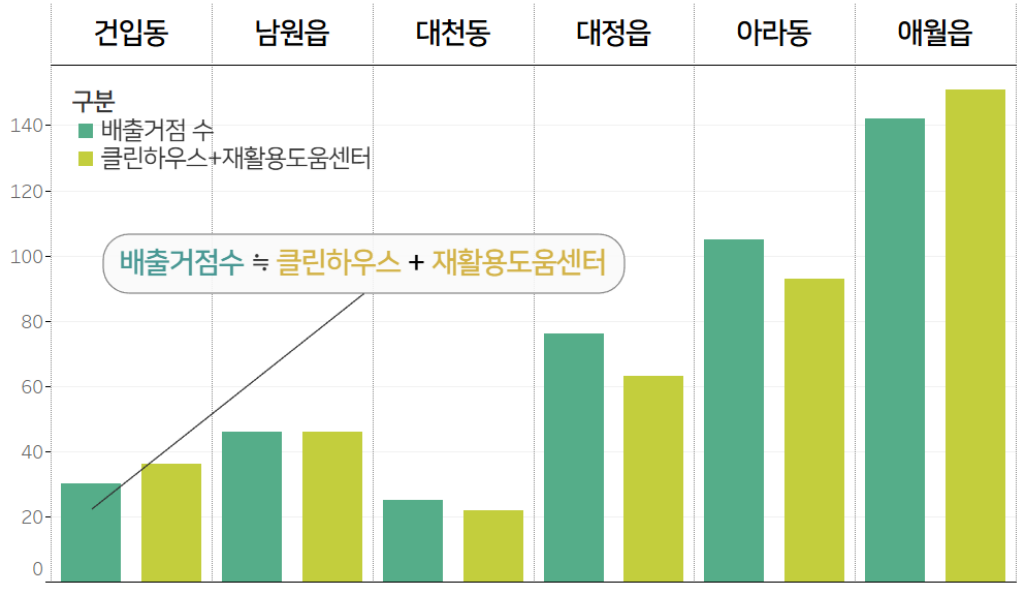
→ RFID 방식으로 **1kg당 30(가정)원**을 지불

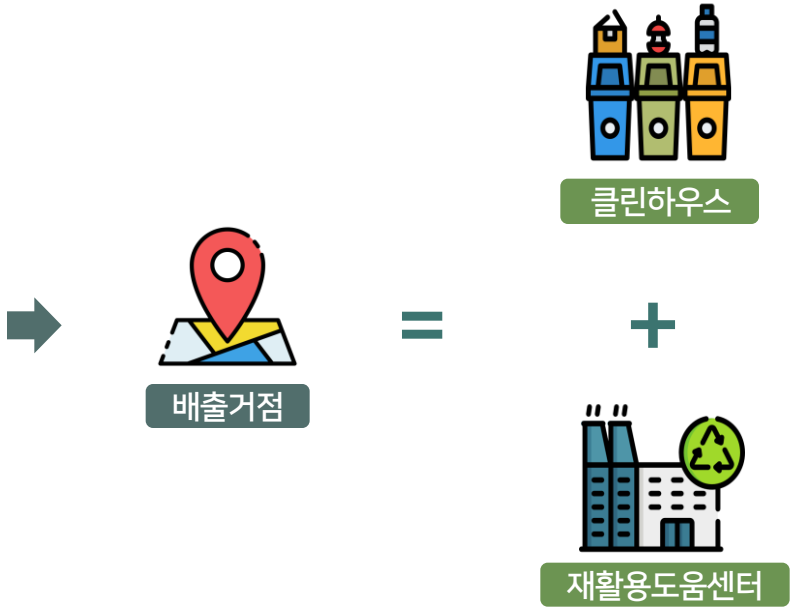
03 데이터 탐색 결과 및 변수 생성
음식물 쓰레기 데이터

배출거점코드

비식별화된 **배출거점**코드 개수를 비교한 결과,

거점개수가 해당 읍면동의 **클린하우스**와 **재활용도움센터**의 합계에 준하는 것을 확인





03 데이터 탐색 결과 및 변수 생성
음식물 쓰레기 데이터

지역 클러스터링

쓰레기 배출량 양상을 고려하여 6개 지역으로 그룹핑

제주_1

제주_2

제주_3

서귀포_1

서귀포_2

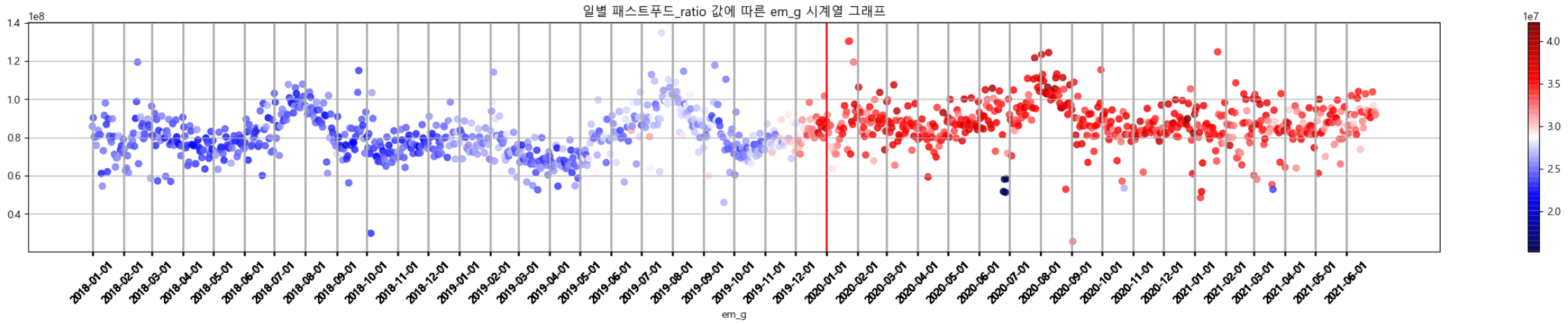
서귀포_3

03 데이터 탐색 결과 및 변수 생성
카드 소비 데이터

건당 결제금액

[건당 결제금액 = 결제금액 / 결제건수]을 의미하는 ratio 파생변수 생성

➔ 업종별 ratio에 따른 배출량 시계열 그래프를 그려 상이한 패턴이 있는지 확인 (붉은색일수록 건당 결제금액이 높은 날)



"코로나를 기점으로 배달 및 패스트푸드 소비량 증가"

03 데이터 탐색 결과 및 변수 생성
카드 소비 데이터

업종 클러스터링

한식 · 패스트푸드 · 아시아음식

주점및주류판매 양식

마트 · 식품 간식

배달

농축수산물

부페

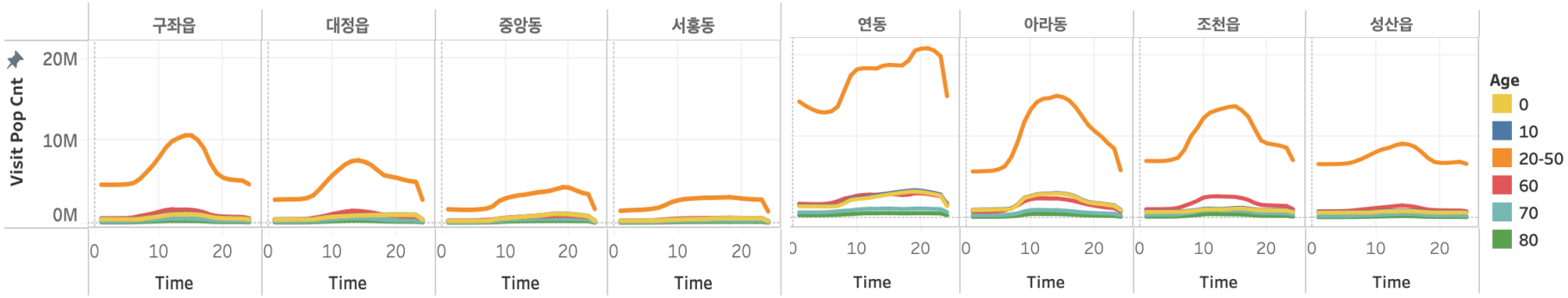
1) 코로나 전후 소비내역 변동폭

2) 쓰레기 배출량과의 상관관계를 고려하여 6개 업종으로 그룹핑

03 데이터 탐색 결과 및 변수 생성
유동인구 데이터

내국인 거주 · 근무 · 방문인구

연령별 방문유동인구를 비교한 결과 **20대-50대**의 활동 양상이 두드러짐



➔ [**출근 이후**(8-19시) 최대 근무유동인구 / **퇴근 이후**(18-8시) 평균 근무유동인구] 를 의미하는 **vpc_daytm_rt** 파생변수 생성

➔ [제주 거주민의 **주 근무 연령대(20-50대)**의 최대 근무유동인구] 를 의미하는 **work_pop_jeju** 파생변수 생성

➔ [제주 거주민의 최대 방문유동인구] 를 의미하는 **visit_pop_jeju** 파생변수 생성

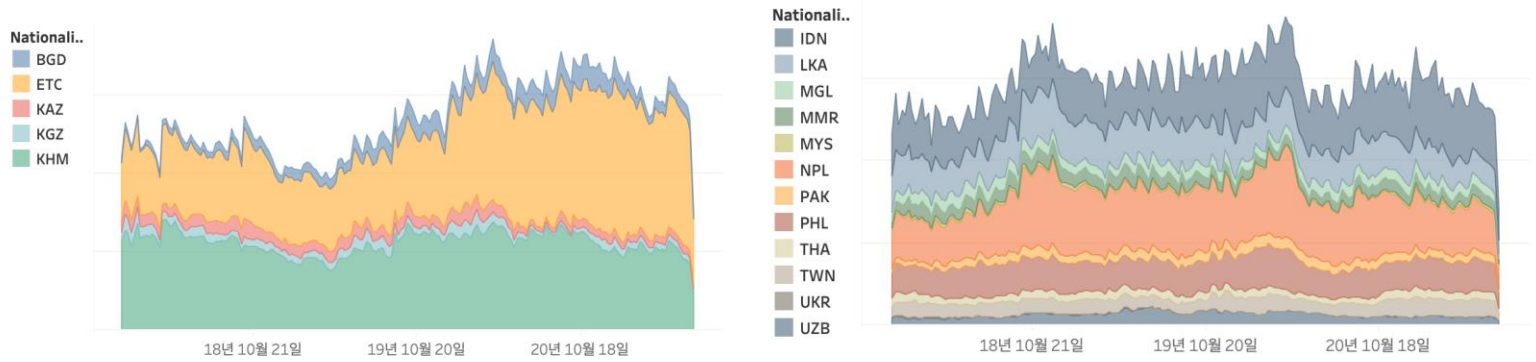
➔ [타지 거주민의 평균 근무유동인구, 최대방문유동인구] 를 의미하는 **work_pop_etc, visit_pop_etc** 파생변수 생성

➔ [제주 거주민의 평균 거주유동인구[를 의미하는 **resd_pop_cnt** 파생변수 생성

03 데이터 탐색 결과 및 변수 생성
유동인구 데이터

외국인 장기체류

지리적 특성과 **인구 규모**를 고려하여 8개 국가로 그룹핑

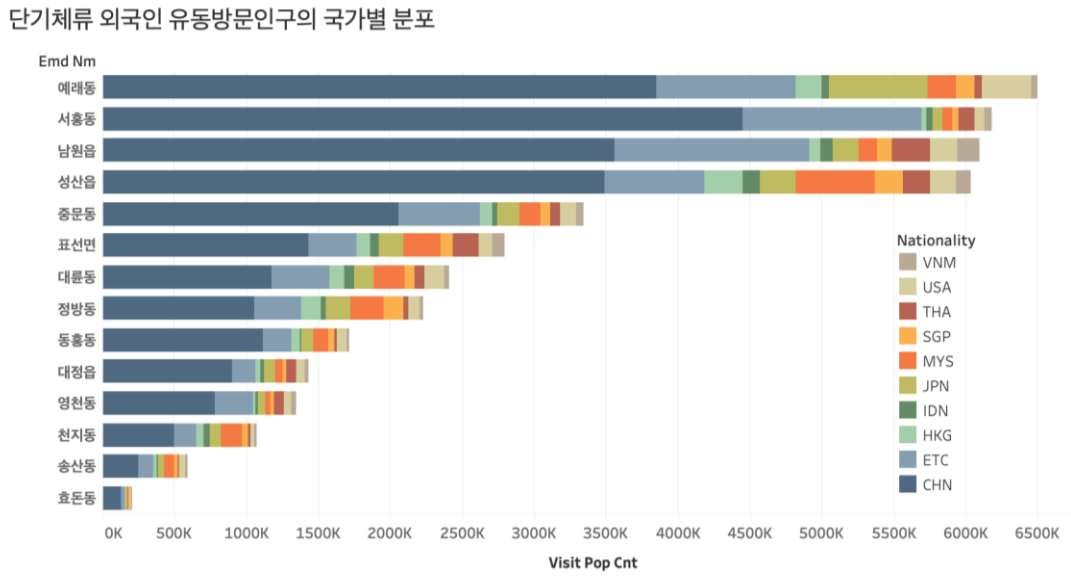


1) 외국인 장기체류 유동인구 파생변수들 간의 상관관계
2) 쓰레기 배출량과의 상관관계를 고려하여 8개의 변수로 축소

03 데이터 탐색 결과 및 변수 생성
유동인구 데이터

외국인 단기체류

코로나 이후 대부분의 국가에서 단기체류 외국인의 유동인구 현저히 감소



1) 읍면동별 최다 방문 국적은 **중국**,

➔ [중국 거주 방문 인구/ 전체 방문 인구] 를 의미하는 파생변수 생성

2) 쓰레기 배출량과의 상관관계를 고려하여 5개의 변수로 축소

03 데이터 탐색 결과 및 변수 생성
기상 데이터

기상

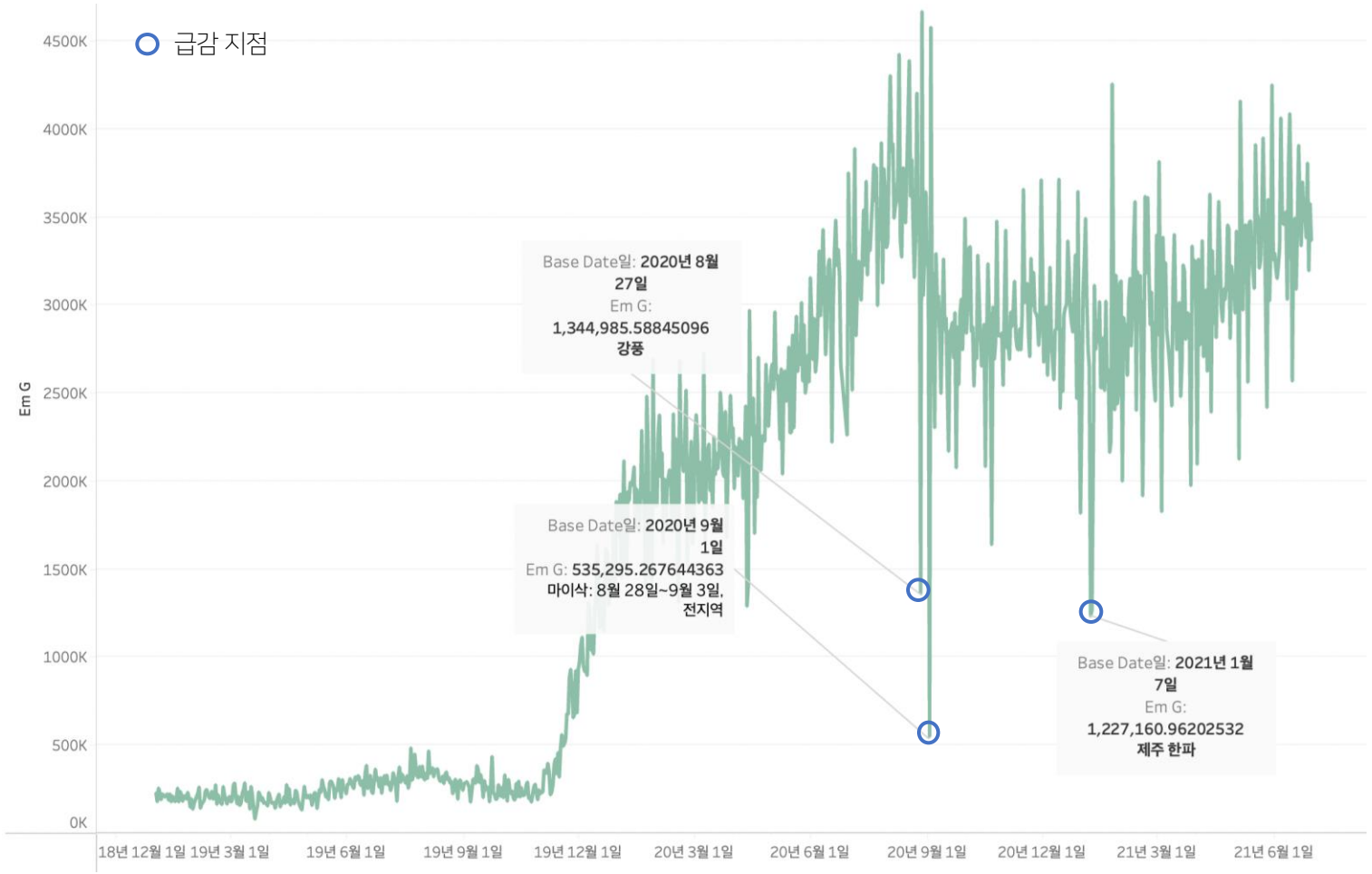
배출량이 급감하는 시기

≈ **자연재해**(강풍, 태풍, 한파)

추가 수집이 필요한 데이터

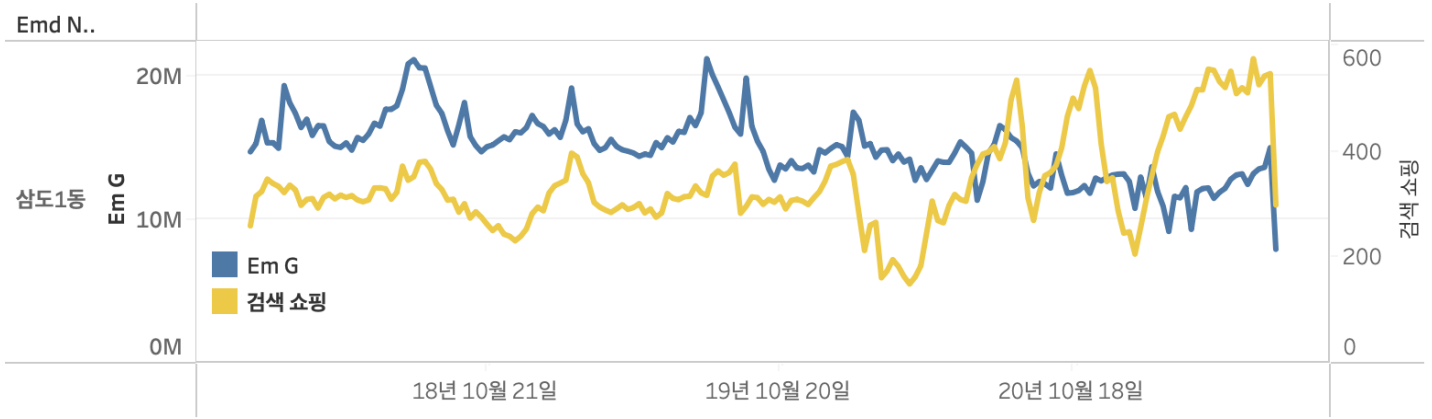
➔ 기상 관측 데이터

➔ 날씨 관련 검색어 데이터



03 데이터 탐색 결과 및 변수 생성
검색어 데이터

실시간 검색어



제주도 관련 전체 검색량 : 데이터: 날씨, 자연재해, 관광, 쇼핑 등의 카테고리로 나누어 검색어의 검색량을 수집

Ex) 전체_관광: [제주 여행, 제주 맛집, 제주 펜션, 제주도 에어비앤비, 제주도 가볼만한 곳 등]

Ex) 검색_태풍 : [제주도태풍,제주도폭풍,제주도해일,제주태풍,제주폭풍,제주해일,솔릭,콩레이,타파,폭우,마이삭,루핏,오마이스 등]

읍면동 별 검색량 : 읍면동별 검색어 데이터 추가를 통해 읍면동별 모델의 성능을 높이고자 함

Ex) 읍면동별검색_관광: [애월 펜션, 애월 맛집, 애월 카페, 애월읍 가볼 만한 곳, 애월읍 바다 등]

03 데이터 탐색 결과 및 분석 개요
사용변수

사용변수

Aa 카테고리	변수
기본 변수	날짜, 읍면동
시간 변수	주말 및 공휴일
거주 인구	내국인 거주인구, 외국인 거주인구, 전체 거주인구, 거주인구 중 외국인 비율, 총 세대수, 인구밀집도, 세대당인구 평균
입도객	전체 입도객 , 내국인 입도객, 외국인 입도객
내국인 유동인구	제주 거주 거주유동인구, 제주 거주 근무유동인구, 타지 거주 근무유동인구, 타지 거주 방문유동인구
검색어	검색_관광, 검색_쇼핑, 검색_쓰레기, 검색_한파, 검색_강풍, 검색_태풍, 검색_장마, 검색_폭염, 검색_우박, 검색_날씨, 검색_미세먼지, 검색_황사, 읍면동별검색_생활, 읍면동별검색_관광, 읍면동별검색_쇼핑
식품 업종별 카드 결제 변수	유형1 카드 결제 건수, 유형2 카드 결제 건수, 유형3 카드 결제 건수, 농축수산물 카드 결제 건수, 배달 카드 결제건수, 배달 카드 결제액
외국인 장기/단기체류 유동인구	중국 국적 거주유동인구, 장기체류 거주 유동인구 전체
날씨	강수, 기온, 습도, 풍속, 태풍, 미세먼지

04 예측 모델 개발 및 결과 분석
분석 개요

STEP 1. 7, 8월 X 예측

STEP 2. 배출량 Y 예측

거주 · 세대

유동 · 카드

기상 · 검색

거주
내국인 / 외국인 / 연령별

세대
인구밀도 / 세대수 / 세대당인구수

입도객
내국인 / 외국인

※ 7, 8월 데이터 누락

카드
유형1-3 / 배달 / 농축수산물

유동인구
내국인 / 외국인 단기·장기체류

쓰레기
배출량

검색량
관광 / 생활 / 쇼핑 / 쓰레기 / 기상

기상
강수 / 기온 / 습도 / 풍속 / 미세먼지 / 태풍

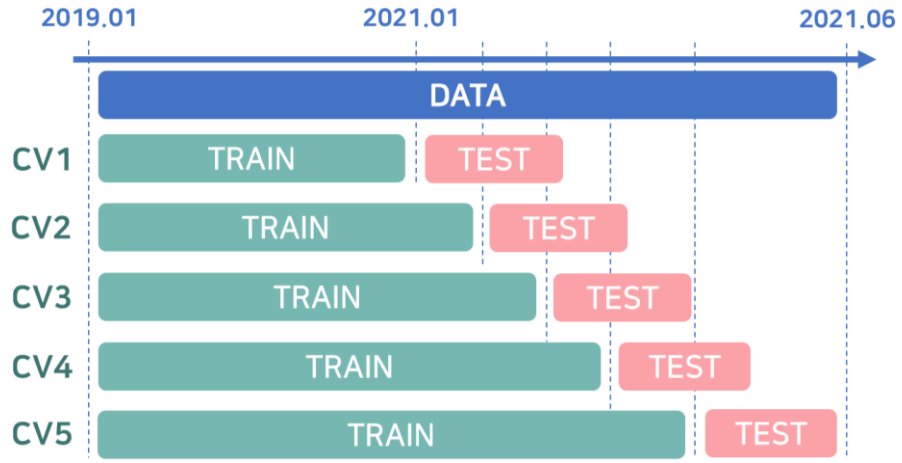
시간
주말 / 공휴일

04 예측 모델 개발 및 결과 분석
시계열 모델 교차검증(Time Series cross-validation)

Time Series CV

일반적인 train-test split 방법을 시계열 데이터에 적용할 경우 미래 시점의 데이터가 모델에 들어갈 우려가 있음

따라서 test set은 train set보다 **미래 데이터**를 사용해야 함.



➔ X 예측 모델의 경우,

20년도 7, 8월과 21년도 4, 5 ,6월의 데이터로 교차검증

➔ Y 예측 모델의 경우 ,

20년도 1월 이후의 데이터로 교차검증

∴ 과적합을 피하고 일반적 모델 생성 & 신뢰성 있는 모델 평가 가능

04 예측 모델 개발 및 결과 분석
X 예측

Problem

7, 8월 배출량 예측에 필요한

카드 · 유동인구 데이터 누락

Solution

7, 8월 데이터가 존재하는 **기상, 검색량 등**의 변수와

Prophet 파생변수 활용해 누락된 데이터 예측

예측에 사용하는 모델

Prophet | **Ridge Regression** | **Random Forest**

PROCESS

STEP 1 Prophet 모형을 활용해 7, 8월 데이터가 있는 변수를 시계열 분해

STEP 2 7, 8월 데이터가 있는 변수와 Prophet 파생변수를 포함해 X 예측

04 예측 모델 개발 및 결과 분석
X 예측

Prophet

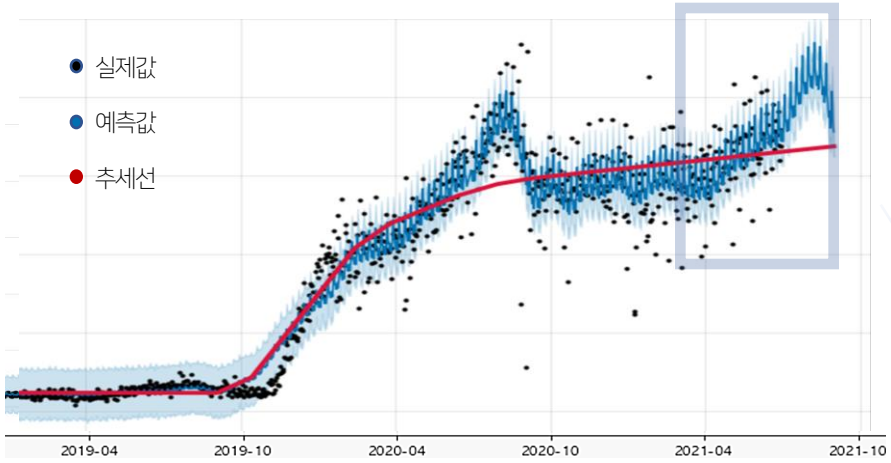
Generalized additive model(GAM)과 유사한 **단변량 시계열** 예측모델

시계열 데이터의 자기상관성을 고려해 **추세와 계절성**으로 분해한 파생변수를 X 예측에 활용

$$y(t) = g(t) + h(t) + s(t) + \epsilon_i$$

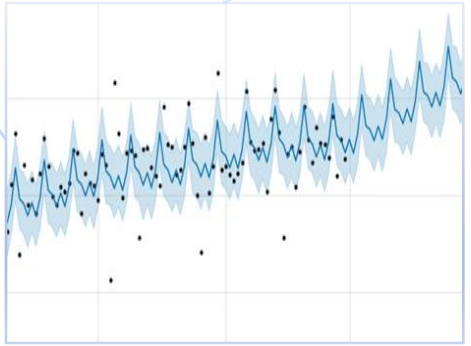
- $y(t)$: 예측값
- $g(t)$: 추세(trend)
- $s(t)$: 계절성(additive)
- 주간 계절성(weekly)

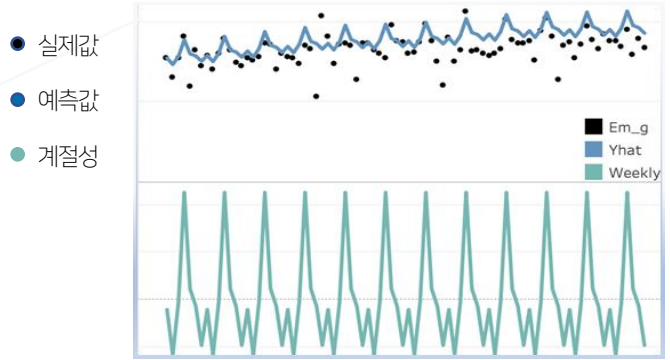
Ex) 애월읍의 쓰레기 배출량 Prophet 예측 결과



파생변수

1. 예측값
2. 추세 + 주간 계절성





04 예측 모델 개발 및 결과 분석
X 예측

Ridge / RF

<기간>

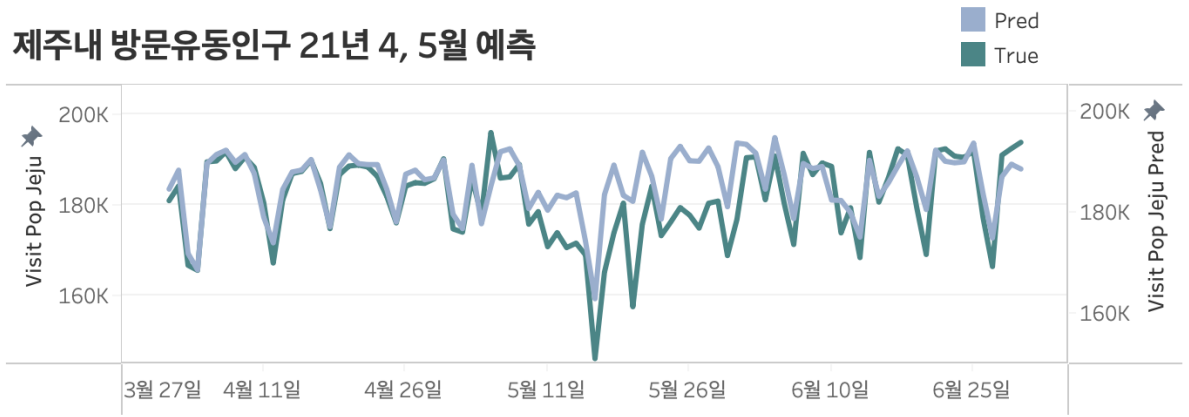
- 전체 기간(2018-01-01-)
- **코로나 발생 이후**의 기간(2019-07-01-)

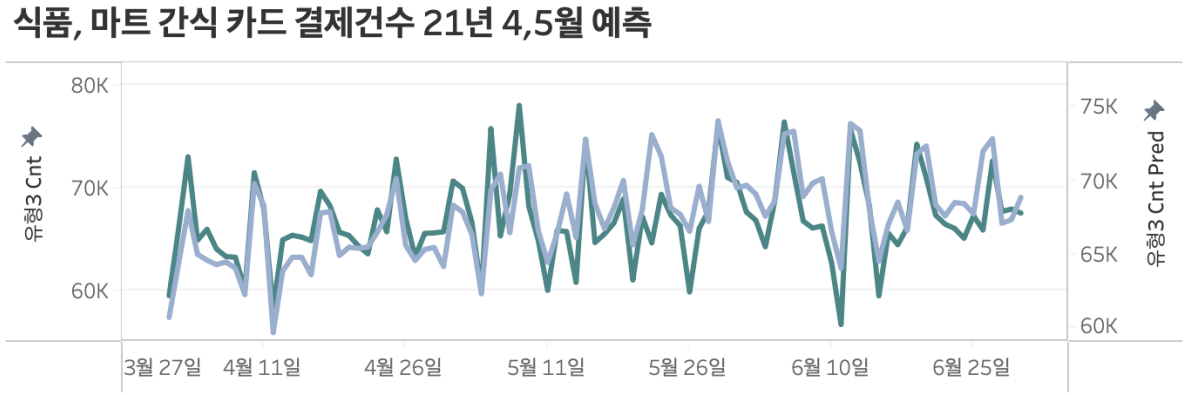
<모형>

- Ridge 회귀
- Random Forest

4개 조합의 모델 중 가장 성능이 좋은 경우를 선택해 최종 X 예측

※ 20년도 7, 8월과 21년도 4, 5 ,6월의 CV 예측 MAPE 기준





04 예측 모델 개발 및 결과 분석
Y 예측

비선형 구조의 시계열 데이터에 적합하도록 수정한 부스팅 모형인 STLB를 사용

PROCESS

STEP 1 X와 Y를 각각 시계열 분해하여 요소별로 예측

STEP 2 3가지 요소를 취합해 원본 데이터 형태로 복원

STEP 3 예측한 Y를 다음 시점의 예측을 위한 lag 변수로 사용

😢 기존의 Boosting / Bagging 모형

시간 관련 변수를 넣어주지 않는 한 시간적 속성을 고려하지 않기 때문에 시계열 예측으로 보기 어려움

😊 STLB

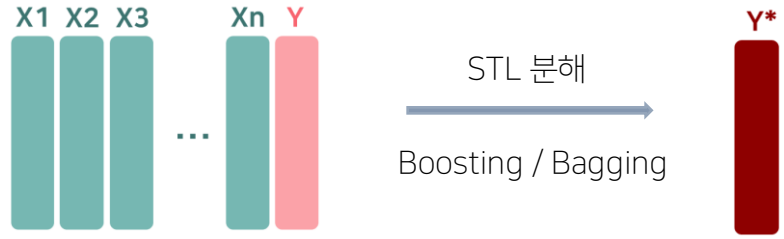
시계열 데이터의 특성을 훼손하지 않고 Tree 모형에 적용 가능

Decision Tree 기반 학습만으로는 잡아내지 못하는 복잡한 패턴을 분해하여 예측하기 때문에 성능이 높아짐

04 예측 모델 개발 및 결과 분석
Y 예측

시계열 분해

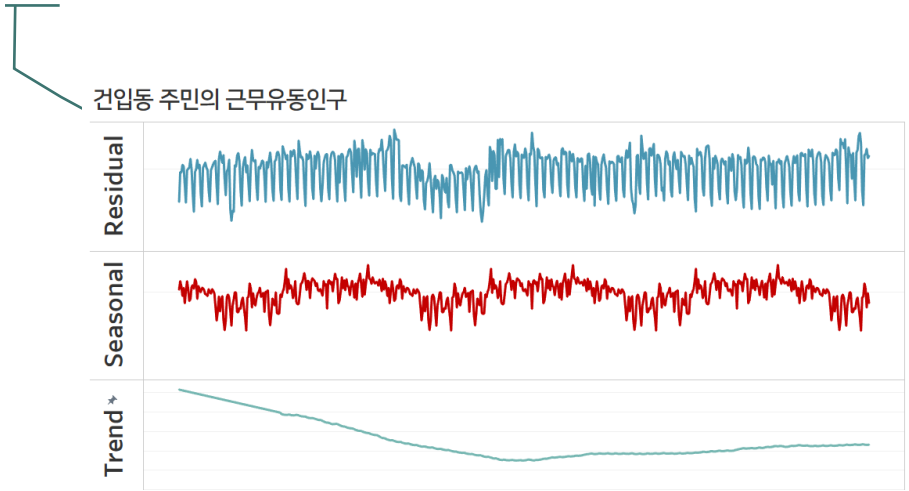
추세(trend), 계절성(seasonal), 잔차(residual)로 시계열 분해한 뒤에 각각의 요소에 대해 예측을 수행 + 원래 형태로 복원

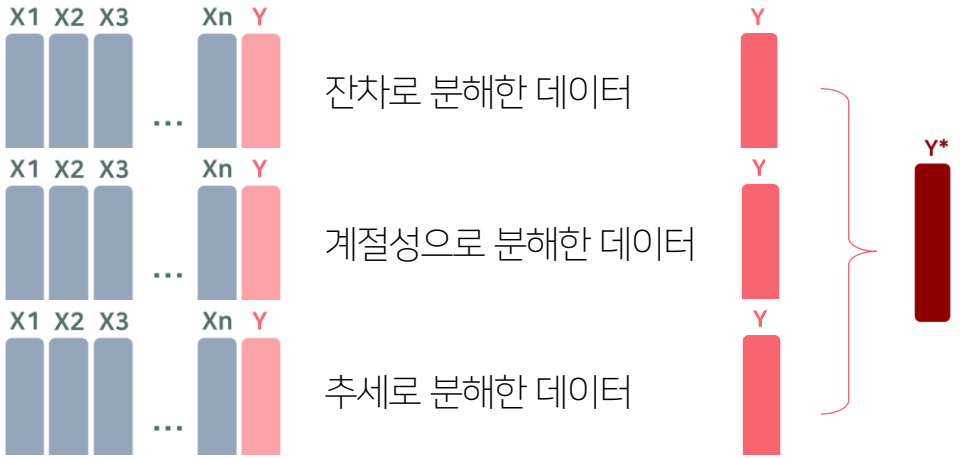


STL 분해

Boosting / Bagging

Y* = (Y의) 잔차 × (Y의) 계절성 × (Y의) 추세



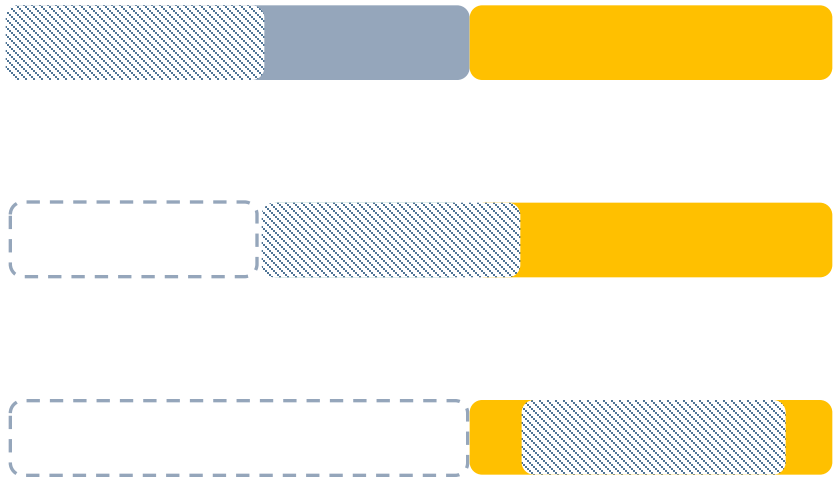


04 예측 모델 개발 및 결과 분석
Y 예측

Lag 변수

과거 배출량을 반영하며 예측하기 위해 n일 후의 배출량을 의미하는 **em_g_shift** 파생변수 생성

Ex) y lag = 7일



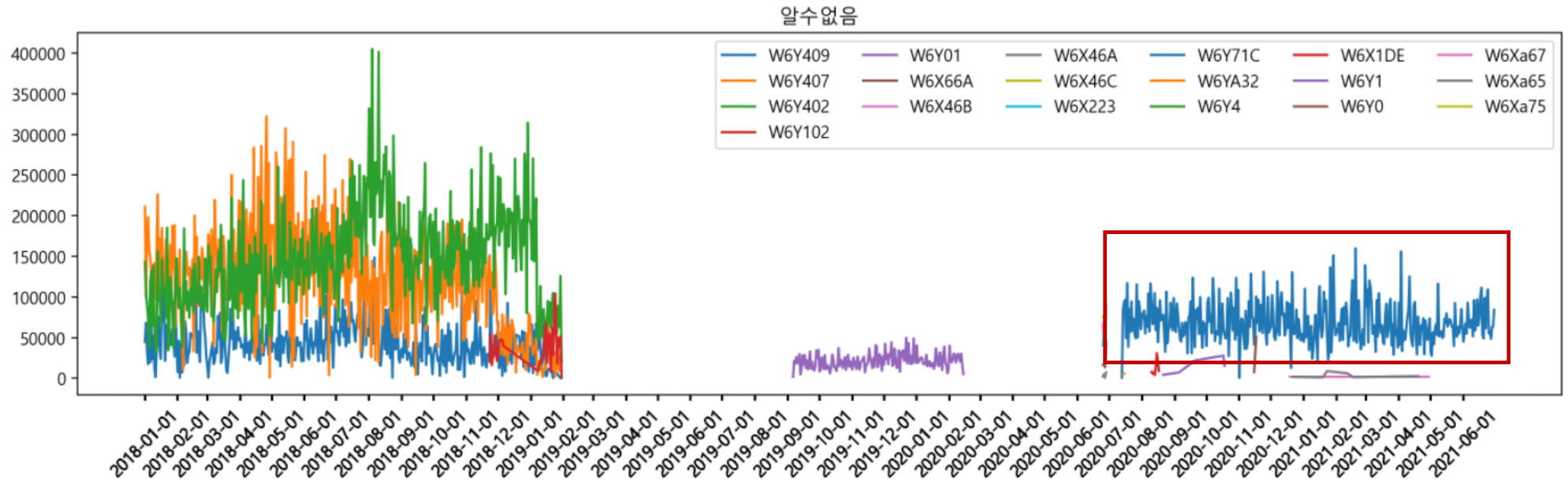
04 예측 모델 개발 및 결과 분석
알수없음 예측

데이터에 쓰레기가 배출된 읍면동, 도시를 알 수 없는 "알수없음" 존재

많은 배출 거점이 있지만, 2020년도 7월부터 최근까지 **W67IC 거점**에서 대부분의 쓰레기가 배출된 것을 확인

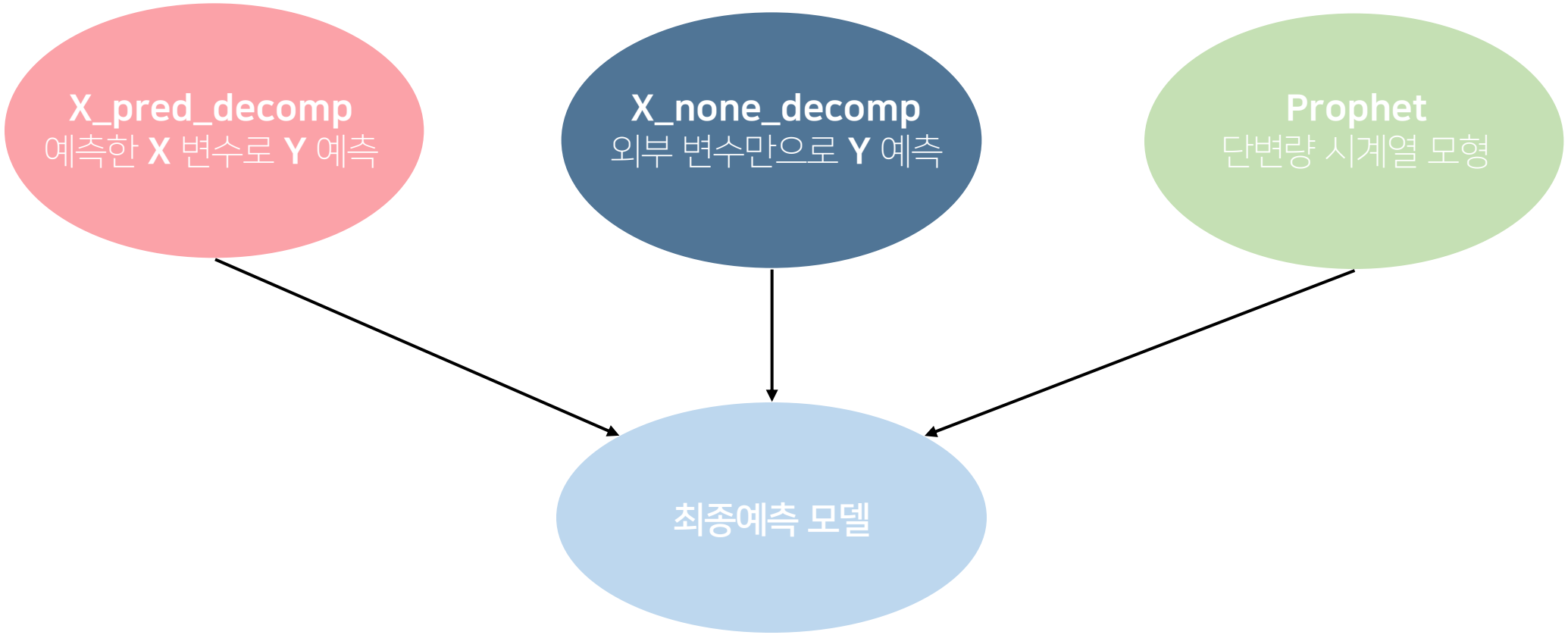
읍면동을 특정할 수 없고 데이터가 상대적으로 적기 때문에

X 변수를 사용하지 않고 **단변량 시계열 모형**인 **Prophet** 모형으로 7, 8월 쓰레기 배출량 예측



04 예측 모델 개발 및 결과 분석
최종 예측 모델: ensemble

Ensemble: 여러 개의 예측 모형을 생성 후 결과를 종합, 하나의 예측결과 도출 → 모델의 안정성으로 성능 향상



05 감소 방안 도출 및 예측 결과 활용 예시
모델링 결과 해석

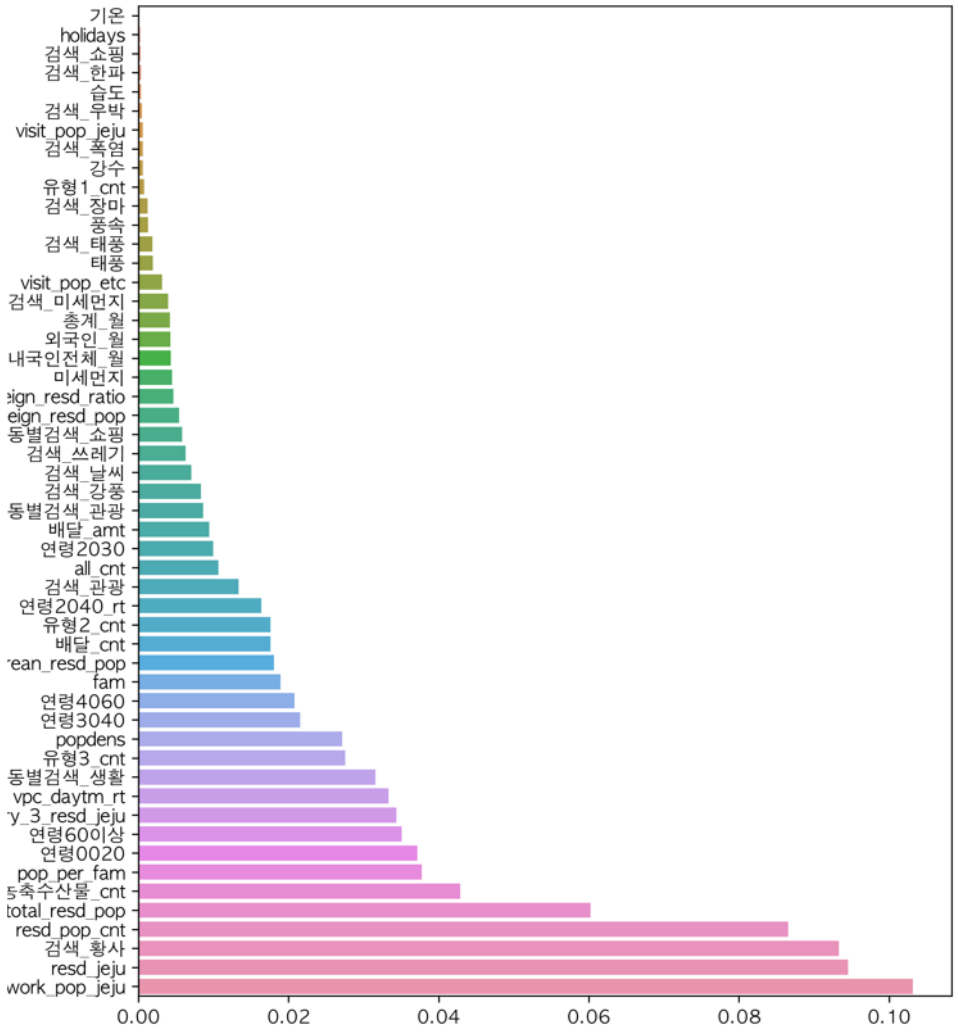
1 제주도 내 쓰레기 배출량 증가의 공통 주요요인 해석

2 읍면동 별 모델링을 통해 읍면동별로 다른 주요요인 해석

3 동네 별 맞춤 쓰레기 배출 감소방안+제주도 공통 쓰레기 배출 감소방안 함께 제시

05 감소 방안 도출 및 예측 결과 활용 예시
X 예측 모델 평가

아라동의 feature importance



X 예측 모델 성능 평가

Aa 이름	X 예측 변수 반영 전	X 예측 변수 반영 후
예래동	MAPE 9.59 →	MAPE 8.46
도두동	MAPE 8.31 →	MAPE 7.79
아라동	MAPE 5.91 →	MAPE 5.5

X 예측 모델링 결과를 Y 예측 모델링에 반영한 결과,

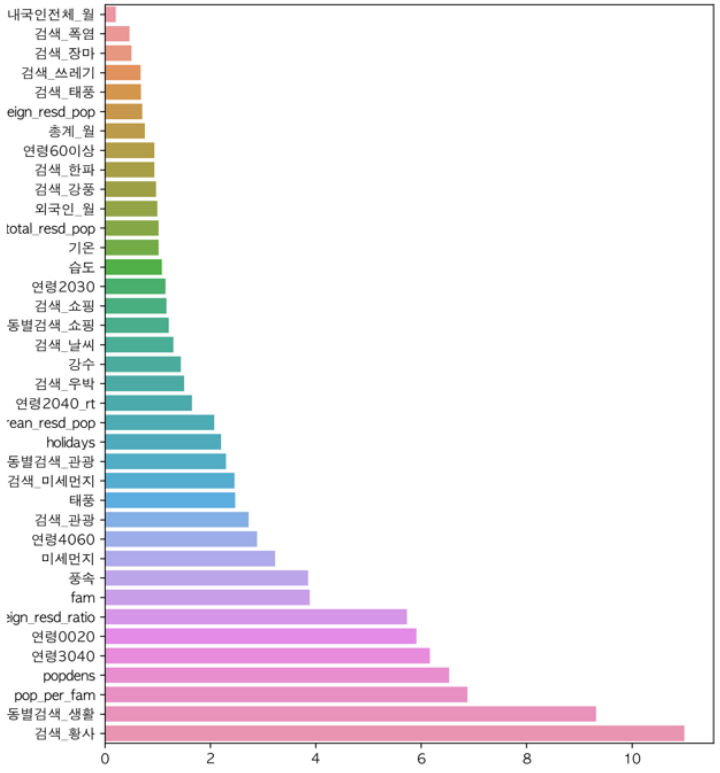
외부변수만 넣었을 때보다 MAPE 감소

→ X 예측 성능 향상에 따라 Y 예측 성능이 향상됨을 확인

특히 **내국인 유동인구 변수, 카드** 변수가 중요

05 감소 방안 도출 및 예측 결과 활용 예시

제주도 공통 쓰레기 배출 주요 요인: 날씨



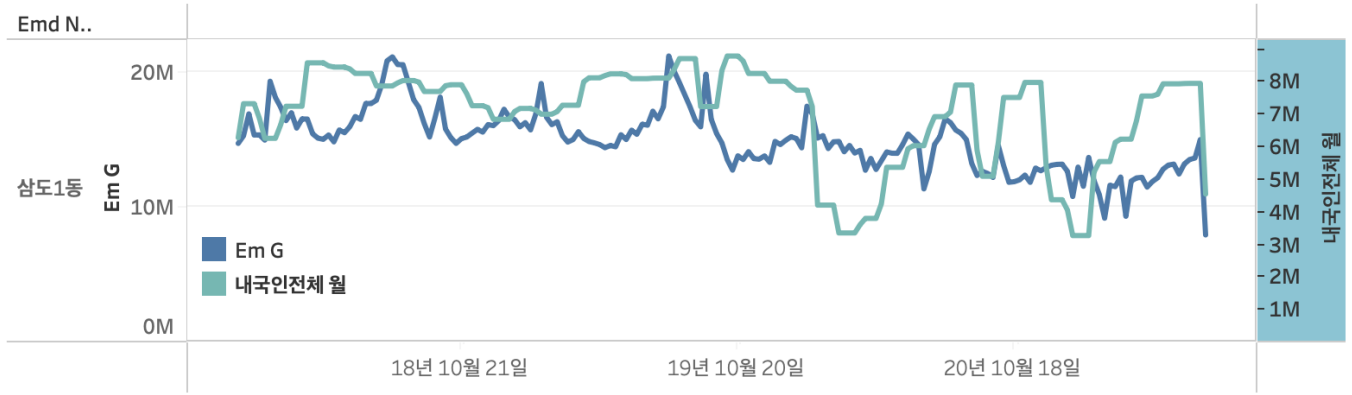
대천동 모델 (월 평균 MAPE 4.04)

날씨가 좋지 않을 때 **날씨를 검색**하는 것과 쓰레기 배출량에 연관성 존재

→ 쓰레기 배출량과 날씨의 상관관계

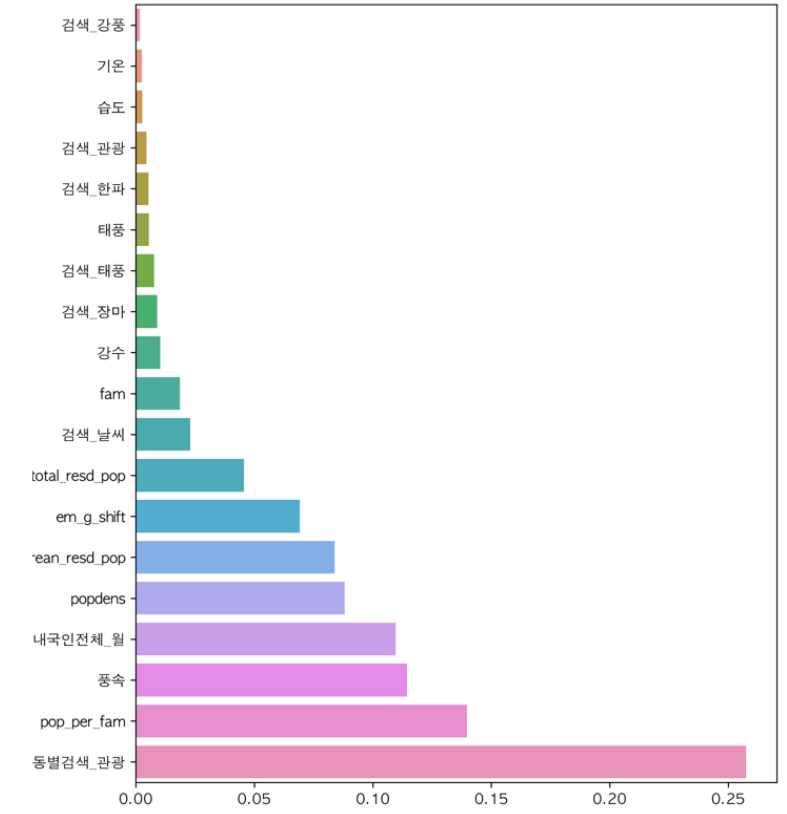
→ 사람들의 행동이 검색량 변수에 반영

05 감소 방안 도출 및 예측 결과 활용 예시
제주도 공통 쓰레기 배출 주요 요인: 관광 관련 변수



관광지의 특성으로 인해 쓰레기 배출량과 **입도객 변수가** 쓰레기 배출량과 연관이 높음

삼도1동 모델 (월별 평균 MAPE: 8.66)



관광 관련 검색어의 feature importance가 1위

05 감소 방안 도출 및 예측 결과 활용 예시

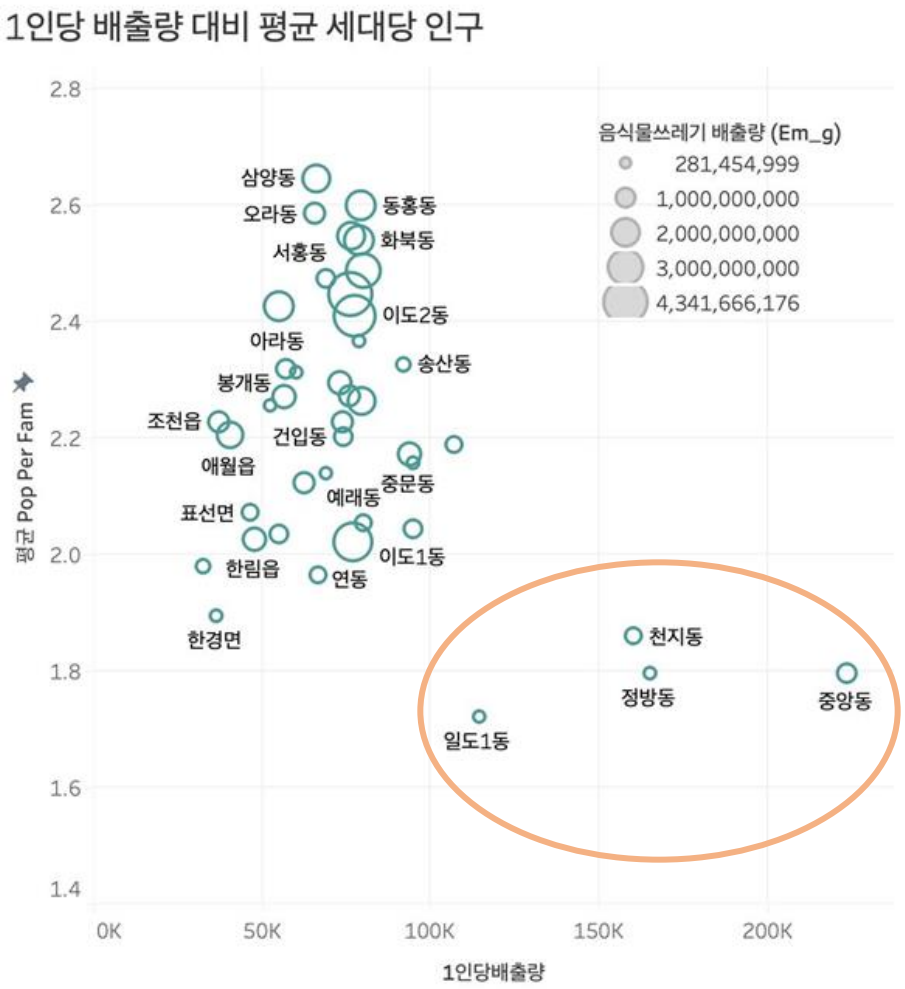
제주도 공통 쓰레기 배출 주요 요인: 카드 결제 건 변수

동네 별 음식 소비의 지표가 되는 카드변수는 em_g와 강한 상관관계를 가짐

음식물 쓰레기 배출량 예측에의 주요 요인

특히 배달 업종 카드 결제 건수의 경우 코로나-19로 인한 배달량 증가로 인해 주요 요인으로 부상

05 감소 방안 도출 및 예측 결과 활용 예시
인당 쓰레기배출량이 높은 읍면동 결과해석

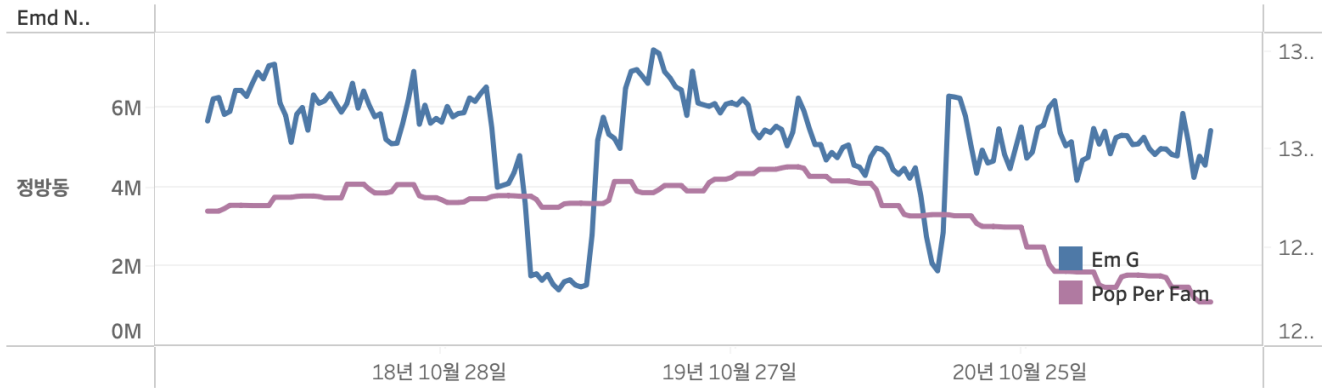


단순 쓰레기 배출량이 아닌 **인당 쓰레기 배출량을 비교**

인당 쓰레기 배출량이 높은 요주의 읍면동 특성은?

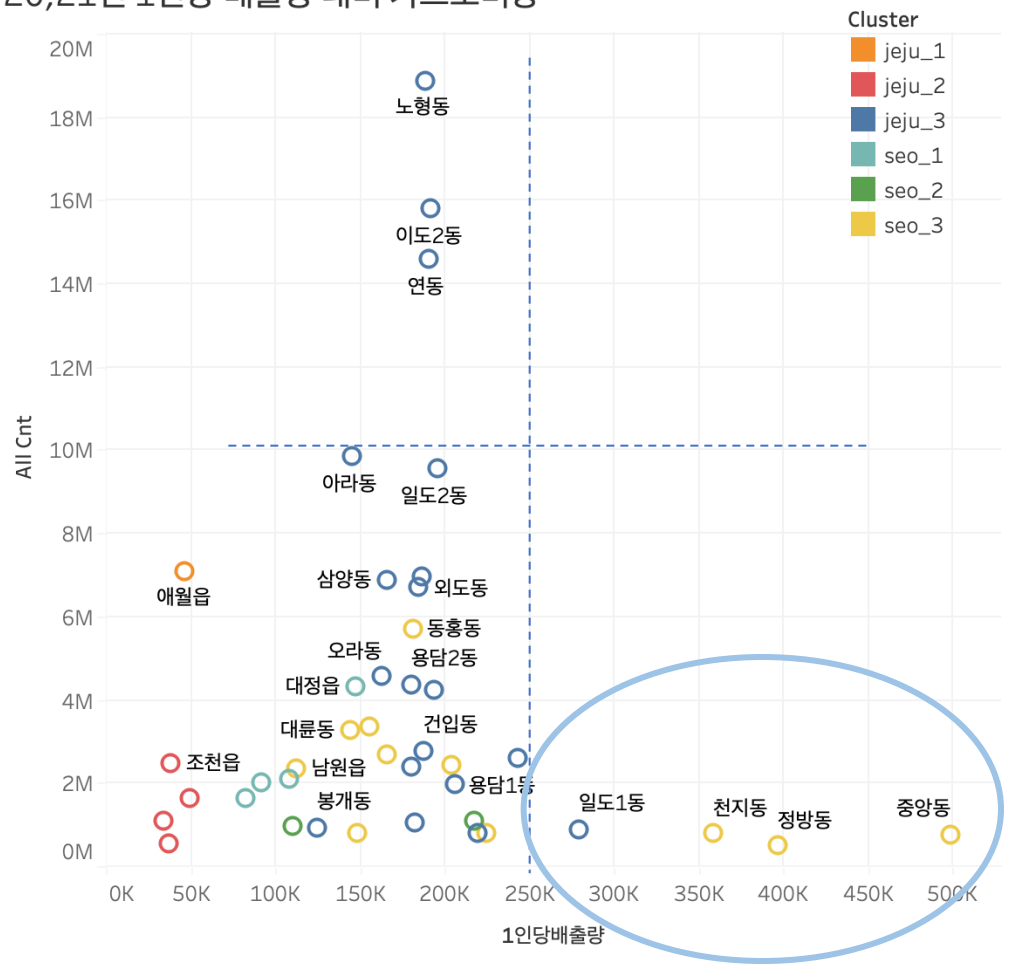
[일도1동, 정방동, 중앙동, 천지동] = 세대당 인구 평균이 가장 낮은 읍면동

⇒ 음식을 남기는 비율이 높은 **1인가구**에서 배출량이 높았을 것으로 유추



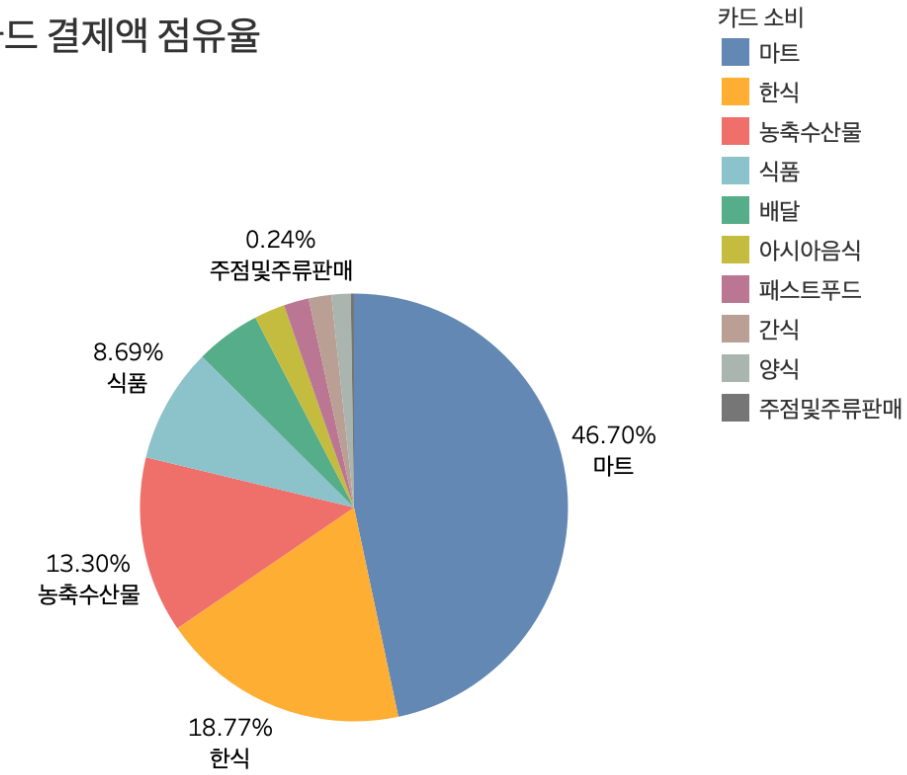
05 감소 방안 도출 및 예측 결과 활용 예시
인당 쓰레기배출량이 높은 읍면동 결과해석

20,21년 1인당 배출량 대비 카드소비량



카드를 많이 쓰는 지역이 아님에도 **인당 배출량 多**

중앙동 8월 카드 결제액 점유율

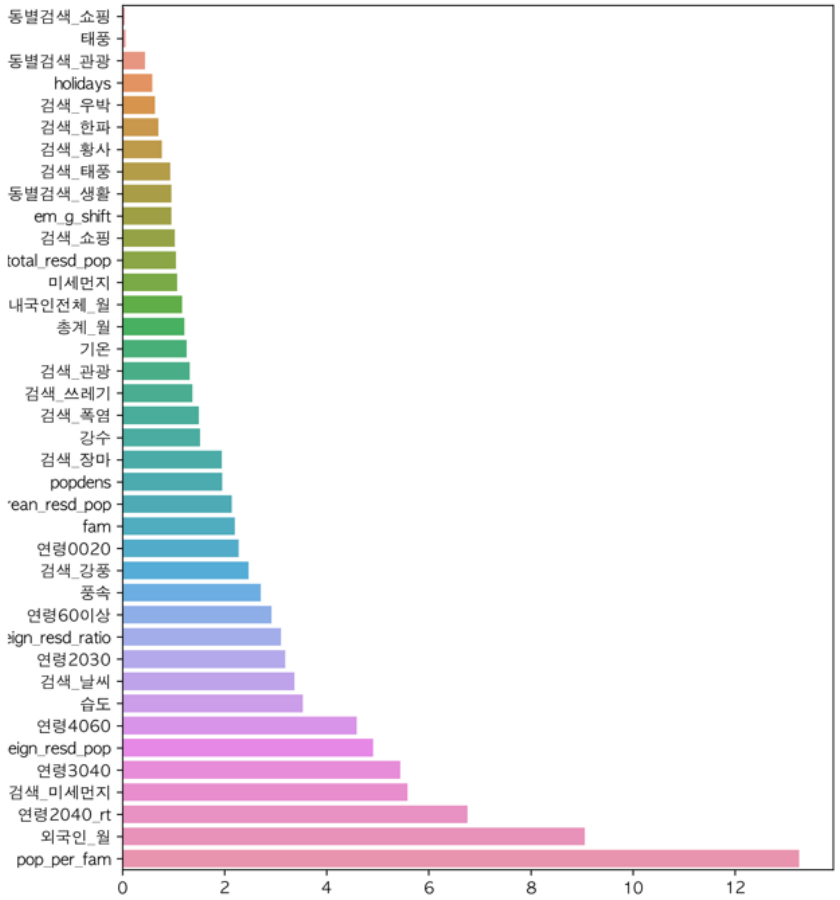


1인당 배출량이 가장 높은 중앙동의 21년도 8월 **카드 결제액 업종별 점유율을 예측**한 결과, 마트가 46.7%로 다른 읍면동에 비해 높음

→ 1인가구가 식재료 등 음식을 남기면서 쓰레기 배출량 증가

05 감소 방안 도출 및 예측 결과 활용 예시
인당 쓰레기배출량이 높은 읍면동 결과해석

정방동 모델 (월별 평균 MAPE: 4.6)



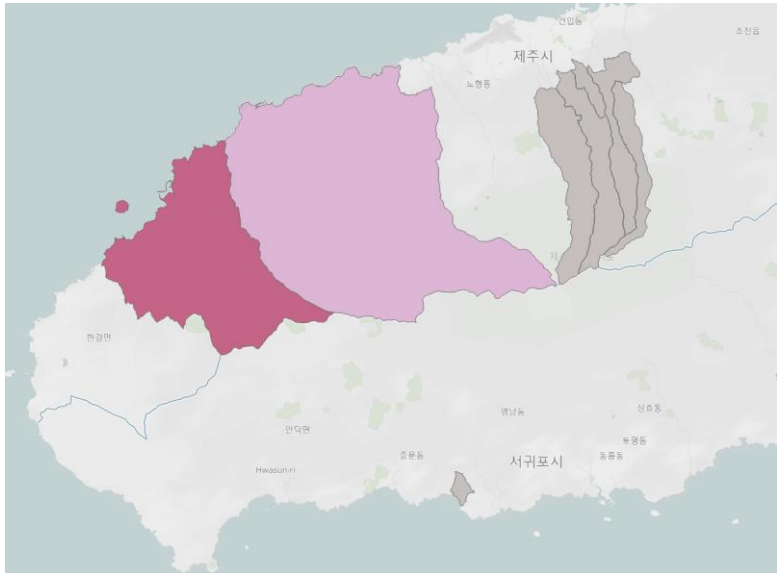
상위 feature importance에 세대당 인구, 인구밀집도, 총세대수 등의 인구 변수 늘 포함

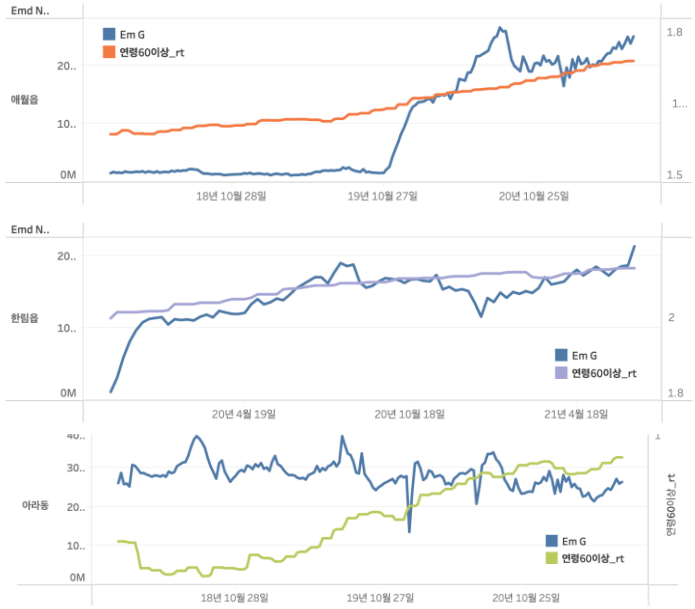
[일도1동, 정방동, 중앙동, 천지동]

→ 해당 읍면동은 세대당 인구가 쓰레기 배출량 예측의 주요 요인으로 나타남

→ 읍면동별 특성 모델에 효과적으로 반영

05 감소 방안 도출 및 예측 결과 활용 예시
인당 쓰레기 배출량이 낮은 읍면동 결과해석



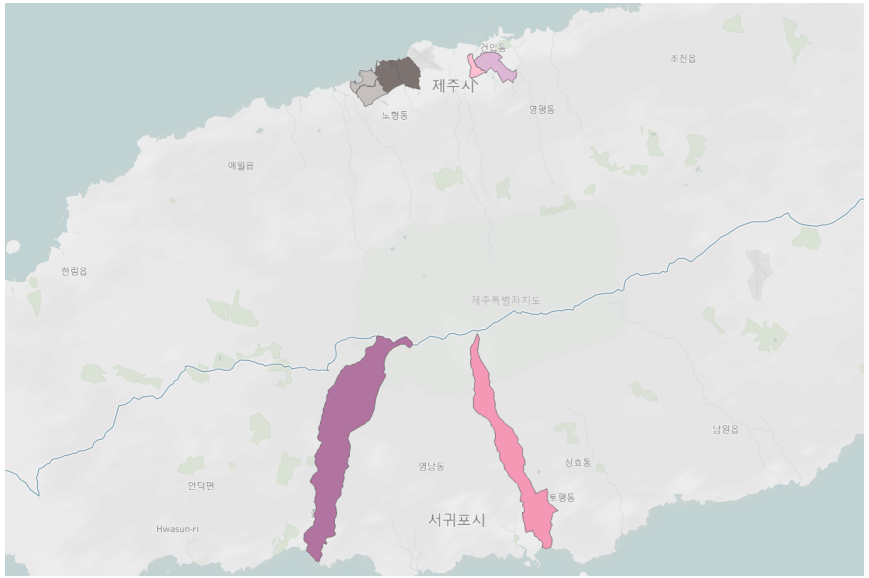


총 쓰레기 배출량은 높은데 인당 쓰레기 배출량이 낮은 읍면동 = [아라동, 애월읍, 한림읍]

연령대가 낮으면 음식물 쓰레기를 더 많이 버리고, 연령대가 높으면 음식물 쓰레기를 더 적게 버린다는 연구결과 (Richter.B)

총 인구에서 20세 이하의 비율이 줄고 60세 이상의 비율이 늘어남에 따라서 쓰레기 배출량 증감에 주요 요인으로 작용

05 감소 방안 도출 및 예측 결과 활용 예시
읍면동 특성 맞춤 감소방안

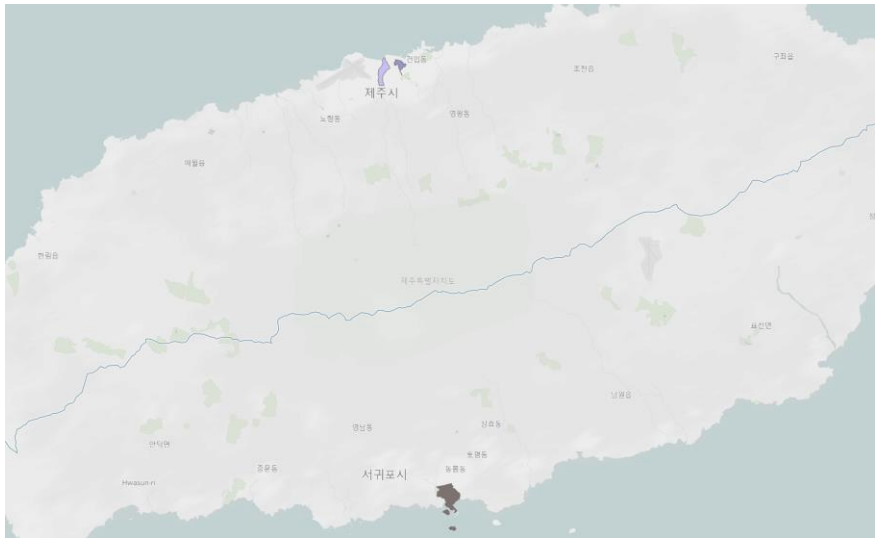


연령대가 낮고 배출량이 높은 지역 = [도두동, 동홍동, 이호동, 이도1동, 일도2동, 중문동]

RFID 방식의 장점을 이용, 쓰레기 배출 시에 **해당 달의 누적 쓰레기 배출량 메시지로 전송**

'인증샷 캡쳐' 문화 및 메신저 사용에 익숙한 낮은 연령대의 **자연스러운 인식 변화** 도모

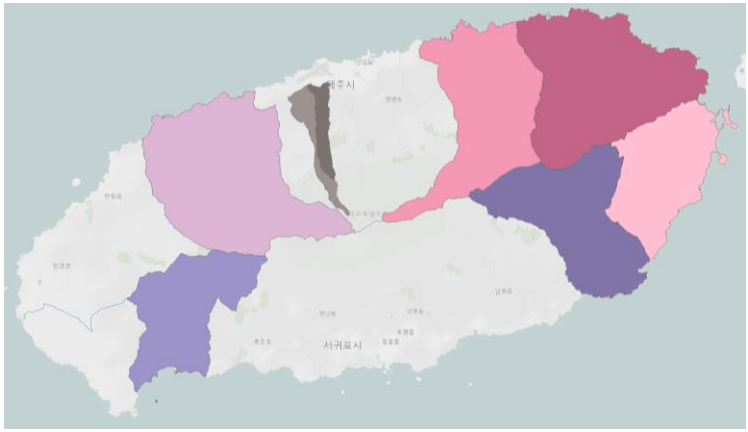
05 감소 방안 도출 및 예측 결과 활용 예시
읍면동 특성 맞춤 감소방안



세대당 인구가 낮고 배출량이 높은 지역 = [용담1동, 일도1동, 정방동, 중앙동, 천지동]

1인 가구가 많은 특성을 고려, 마트 식재료 1인분 씩 판매하도록 독려하는 정책 시행

05 감소 방안 도출 및 예측 결과 활용 예시
읍면동 특성 맞춤 감소방안



관광지 = [구좌읍, 노형동, 성산읍, 안덕면, 애월읍, 연동, 조천읍, 표선면]

관광객 대상으로 쓰레기 배출 관련 관광사업 시행, 제주 **관광객 인식개선**을 통해 쓰레기 배출량 감소 도모

ex) 실제로 플라스틱 쓰레기의 경우 “제주 줍깅”과 같은 사업 존재

음식물 쓰레기로 만든 퇴비 체험 등의 관광사업 시행, SNS 인증 캠페인을 통해 **참여 유도**

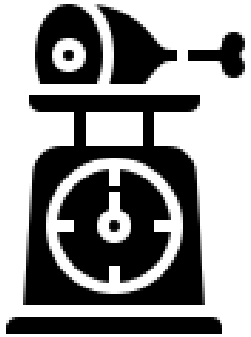
05 클린하우스 최적화 방안

제주도 내 클린하우스 : 쓰레기 배출 시 음식물 쓰레기 계량기(RFID) 사용

가정: 1kg=30원
소형음식점: 1kg=51원
대형음식점: 1kg=106원

제주도 하루 평균 음식물 쓰레기 210~220t. 음식물 쓰레기 처리에 **매년 630억원에 달하는 손실액이 발생**

제주도민의 생활·음식폐기물 부담률 전국 최하위권 → 음식물 쓰레기를 적게 배출할 유인 필요



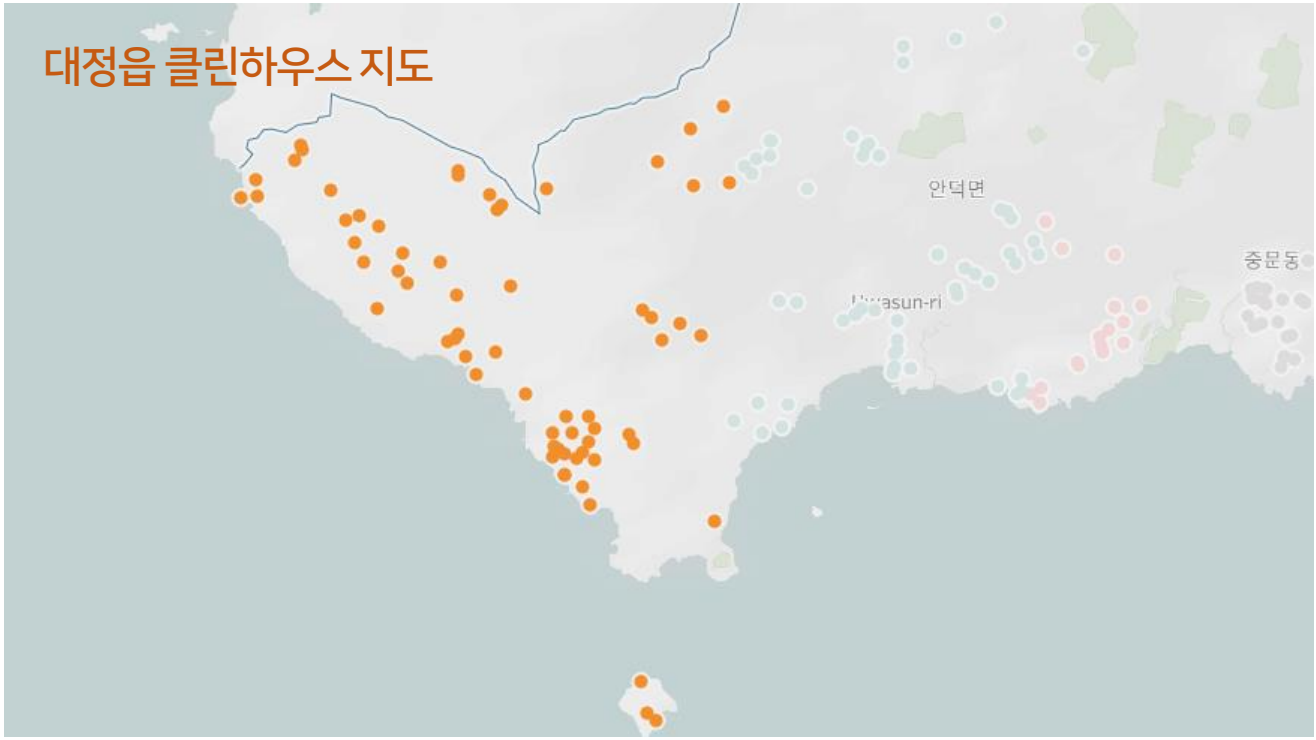
05 클린하우스 최적화 방안

제주도 음식점들은 **음식물 쓰레기 감량기 사용 의무화 제도** 시행 중

음식물쓰레기를 전기로 분쇄·압축·탈수해 배출된 음식물의 70~80%를 감량하여 퇴비 생산

2019년 음식물 쓰레기 배출량 10% 감량 확인, 전 음식점에 적용될 경우 약 50% 감소 예상

05 클린하우스 최적화 방안



유동인구 및 쓰레기 배출량이 높은 **클린하우스에 음식물 쓰레기 감량기** 설치

→ 감량기를 사용하여 쓰레기 배출 시 감소된 무게 만큼 적게 내도록 **인센티브**

05 코로나 이후의 변화

2020년 관광객이 대폭 감소했음에도 쓰레기 배출량은 소폭 감소

코로나 이후 관광객이 원래처럼 돌아올 경우, 쓰레기 배출량의 증가를 피하기 어려움

➔ 쓰레기 감소방안의 적극적인 선전 및 활용 필요

06 참고문헌 및 사용 데이터
분석 데이터 소개

참고문헌

Taylor SJ, Letham B, Forecasting at Scale (2018),
https://peerj.com/preprints/3190.pdf

분석 언어 및 도구

R, Python, Tableau, Excel

사용 데이터

데이터	형식	출처	기준년도	데이터	형식	출처	기준년도
음식물쓰레기	CSV	대회 제공 데이터	2018.01-2021.06	검색어	CSV	네이버 데이터랩	2018.01-2021.08
유동인구	CSV	대회 제공 데이터	2018.01-2021.06	기상관측	CSV	기상자료개방포털	2018.01-2021.08
거주인구	CSV	대회 제공 데이터	2018.01-2021.06	입도객	CSV	열린 데이터 광장	2018.01-2021.08
카드소비	CSV	대회 제공 데이터	2018.01-2021.06	세대정보	CSV	통계청	2018.01-2021.08

Thank you