< 빅콘테스트 데이터분석분야 퓨처스리그 >

제주도 음식물 쓰레기양 예측을 통한 배출량 감소 방안 도출

팀명	에코탐라
팀원	김소은, 김서린, 신정아, 이윤지

목차

1 서론
 1.1 분석 배경
 1.2 제공데이터 활용해 추이 확인
 1.3 분석 내용 요약

2 데이터 전처리 및 사용변수
 2.1 구분변수
 2.2 설명변수
 2.3 반응변수

3 활용 알고리즘
 3.1 상관분석
 3.2 회귀분석
 3.3 XGBoost
 3.4 ARIMA 모형

4 최종예측 결과
 4.1 행정동별 예측 결과
 4.2 행정동 “알 수 없음” 데이터 예측 결과

5 결론
 5.1 분석결과 활용 및 시사점

1. 서론

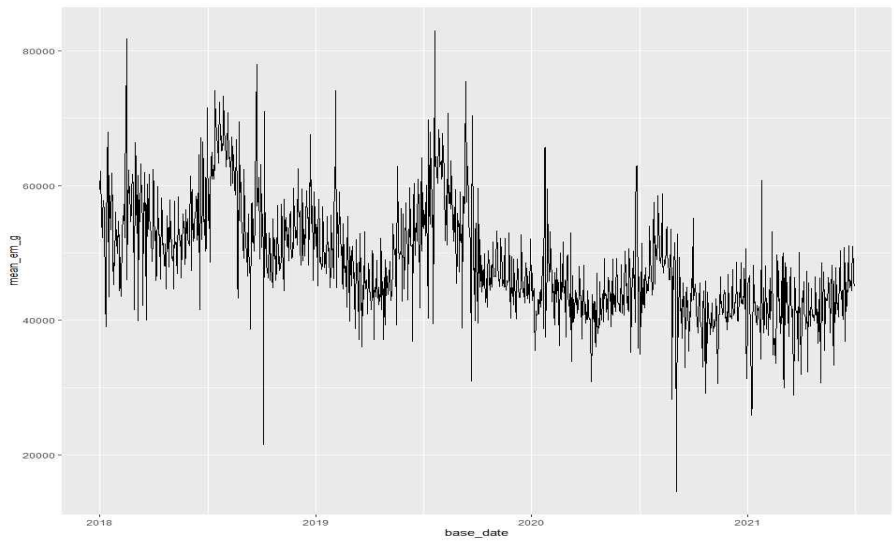
1.1 분석 배경

제주는 청정 자연환경을 자랑하는 대한민국 유일한 유네스코 3관왕 지역이다. 그러나 전국적으로 1인당 생활 폐기물 배출량이 가장 높은 지역으로 쓰레기 배출이 제주의 최대 현안으로 보고 있다. 도민들도 이에 관한 인식이 있으나 음식물 쓰레기의 요인이 복합적으로 문제 해결에 어려움을 겪고 있다. 이런 상황에서 분석을 통해 음식물 쓰레기의 주요 요인을 찾고, 읍면동별 음식물 쓰레기 배출량을 예측하여 배출량을 감소할 수 있는 방안을 제안하고자 한다.

1.2 제공데이터 활용해 추이 확인

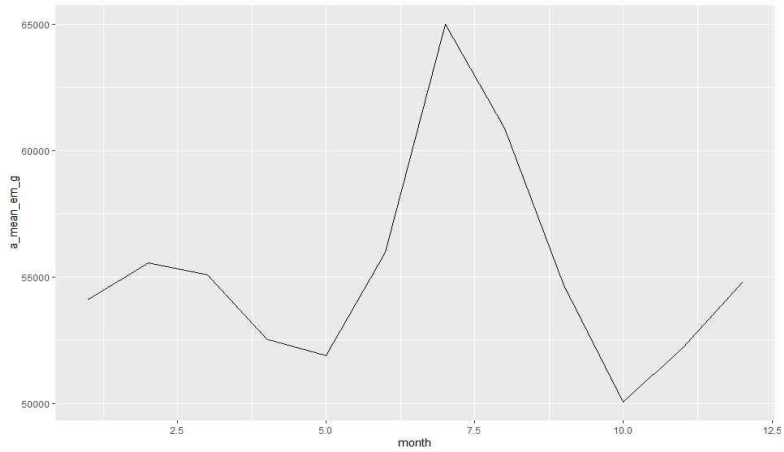
본격적인 분석에 앞서 분석의 목적을 이해하고 분석 방향을 잡기 위해 제공데이터를 이용해 그래프를 그려 추이를 확인했다.

1.2.1 음식물쓰레기 배출량(g)

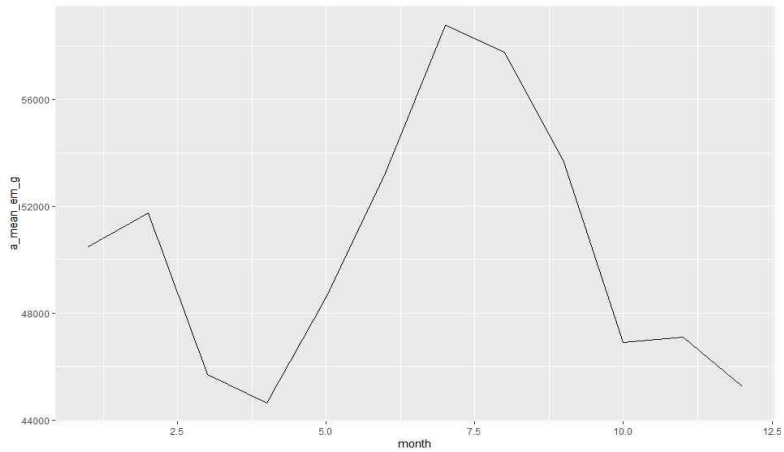


[그림 1] 시간에 따른 설명변수(y)인 음식물 쓰레기 배출량(g)의 일일 그래프

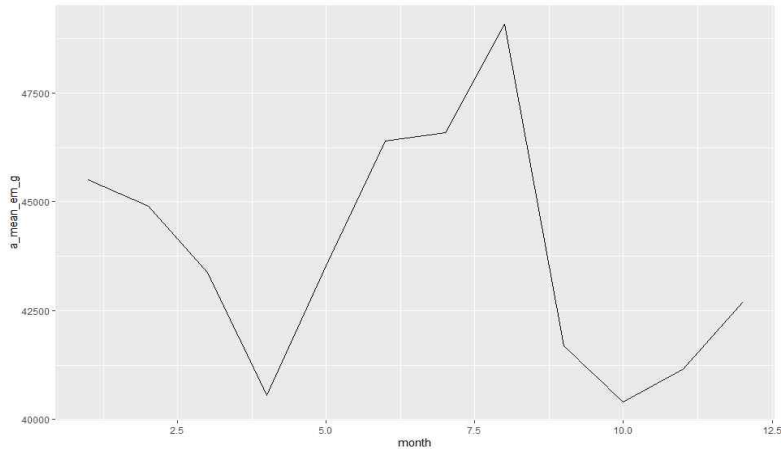
그래프를 통해 2018년부터 지속적으로 음식물 쓰레기 배출량이 점차 감소하는 추세임을 확인할 수 있다. 특히 코로나가 발발한 2020년부터는 음식물 쓰레기 배출량이 많이 감소했다.



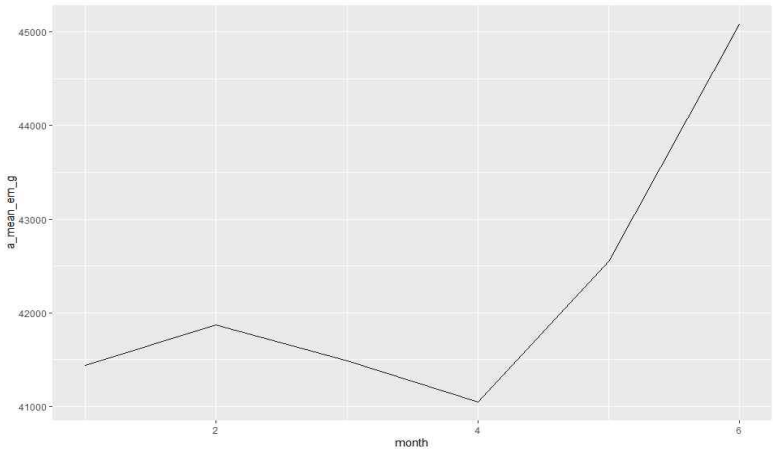
[그림 2] 2018년 월별 음식물 쓰레기 배출량



[그림 3] 2019년 월별 음식물 쓰레기 배출량



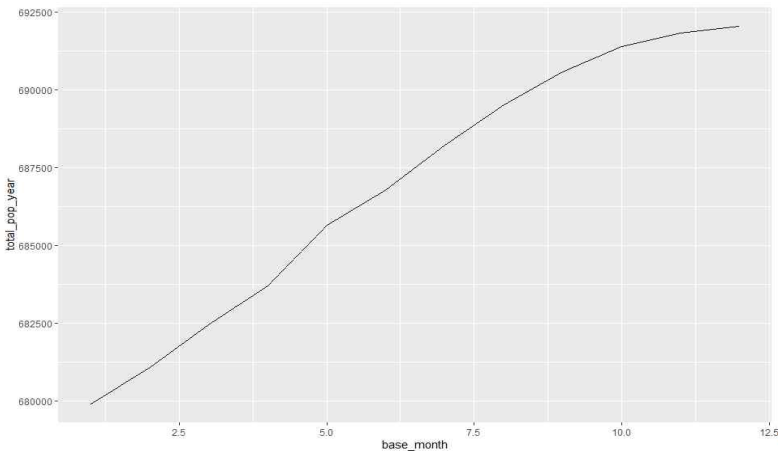
[그림 4] 2020년 월별 음식물 쓰레기 배출량



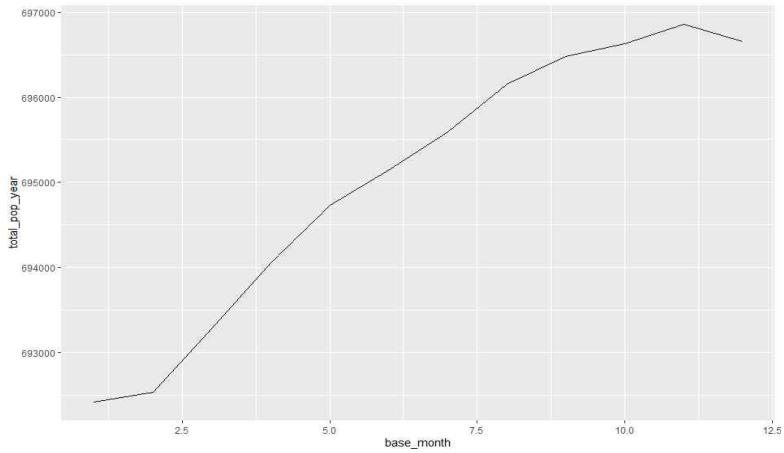
[그림 5] 2021년(1월~6월) 월별 음식물 쓰레기 배출량

월별 그래프를 통해 여름 휴가철인 6월, 7월, 8월에 음식물 쓰레기 배출량이 급증하는 것을 확인할 수 있다. 또한, 겨울 휴가철인 12월, 1월, 2월에도 음식물 쓰레기 배출량이 증가한 모습을 보인다. 이를 통해 대부분 휴가철에 음식물 쓰레기 배출량이 많음을 확인할 수 있다.

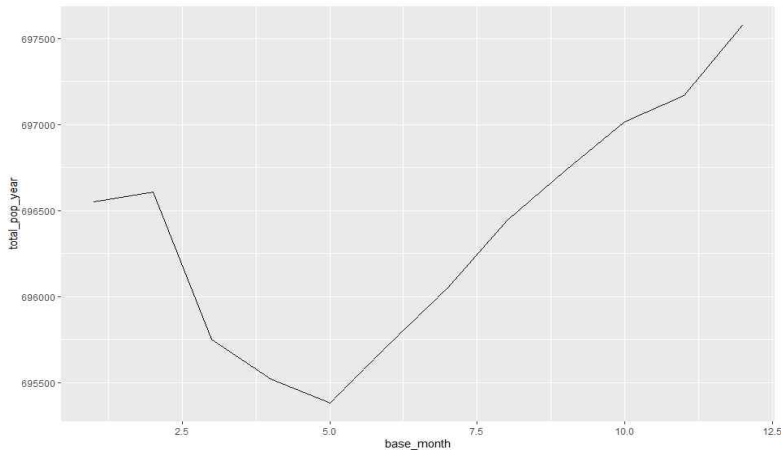
1.2.2 제주 총거주 인구수



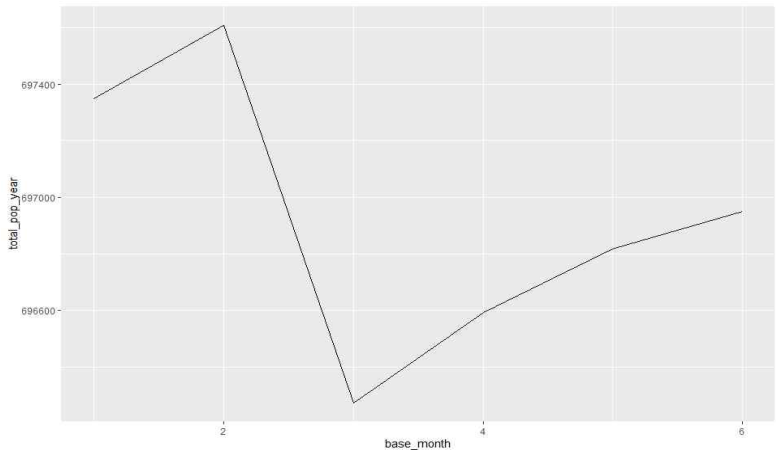
[그림 6] 2018년 월별 제주 총거주 인구수



[그림 7] 2018년 월별 제주 총거주 인구수



[그림 8] 2019년 월별 제주 총거주 인구수



[그림 9] 2021년(1월~6월) 월별 제주 총거주 인구수

위 그래프를 통해 2018년부터 2019년까지 지속적으로 총거주 인구수가 증가하다 2020년 초부터 감소, 증가를 반복함을 알 수 있다.

음식물 쓰레기 배출량이 확연히 감소하기 시작한 2020년에 계속 증가하던 총거주 인구수가 감소, 증가하는 모습을 보여 두 변수 사이에 관련성이 있을 것으로 생각된다. 설명변수(y) 음식물 쓰레기 배출량과 다른 변수들 사이에도 관련성이 있는지 확인하고 최종적으로 음식물 쓰레기 배출량을 감소시킬 방안을 강구하고자 본 분석을 실시한다.

1.3 분석내용 요약

인구, 소비내역, 기후 등 다양한 요인 중 제주도 음식물 쓰레기의 주요 요인을 찾는다. 그 요인들과 음식물 쓰레기 배출량 사이의 관계를 알아보고 2021년 7월, 8월 배출량을 예측할 수 있는 모델을 제시한다. 분석 후 배출량 감소를 위한 합리적인 정책을 모색해보고 예상되는 긍정적인 결과를 도출하여 제주만의 차별점이 있는 정책 마련에 도움이 되는 분석을 하고자 한다.

1.3.1. 제공데이터 추이 확인
- 분석 목적 설정
- 분석 방향 설정

1.3.2. 데이터 전처리
- 외부데이터 불러오기
- 각 변수를 읍면동별 월별 데이터로 변환

1.3.3. 상관분석
- 외부데이터의 적합성 확인

1.3.4. 회귀분석
- 외부데이터와 반응변수 사이의 인과관계 확인

1.3.5. XGBoost
- 파라미터 조절을 통한 최적 모델 선정
- 주요 변수 선정

1.3.6. ARIMA model

- 시계열 데이터의 정상성 확인
- 각 독립변수의 2021년 7월, 8월 값 예측

1.3.7. 최종 예측 및 시각화
- 1.3.5에서 선정한 주요변인을 1.3.6 과정을 통해 2021년 7월, 8월 값 예측
- Qgis를 이용해 음식물 쓰레기 배출량(y)의 예측값 시각화
- 행정동 "알 수 없음" 데이터 예측

1.3.8. 결론
- 예측 결과를 통한 제주도의 음식물 쓰레기 배출량이 감소할 수 있는 방안 제안

2 데이터 전처리와 변수

2.1 구분변수

1) base_date (배출일자)

- 타입 : DATETIME
- YYYY-MM-DD
- 분석에 사용될 설명변수들을 base_date를 이용해 월 단위 데이터로 변환해 사용

2) emd_cd (행정동 코드)

- 타입 : STRING
- 43개 행정동 코드 + 알 수 없음
- 분석에 사용될 각 데이터를 emd_cd별로 정리한 후 분석
- “알 수 없음” 데이터는 따로 모아서 분석

2.2 설명변수

1) korean (내국인 유동인구)

- 거주인구(res_pop_cnt) + 근무인구(work_pop_cnt) + 방문인구(visit_pop_cnt)
- 타입 : FLOAT
- 해당 시각 정각에 측정한 거주, 근무, 방문인구(명)의 합
- 거주/근무/방문인구 : 1~24시 해당 시간 정각에 측정한 인구
  거주지/근무지/방문지와 근무지 외 지역에 머문 시간(분) / 60분
- 시간별 내국인 유동인구 데이터를 월별 내국인 유동인구 데이터로 정리해 사용
- 행정동별 월 내국인 유동인구 수로 변환 후 분석에 사용

2) long_term_frgn (장기체류 외국인 유동인구)

- 거주인구(res_pop_cnt) + 근무인구(work_pop_cnt) + 방문인구(visit_pop_cnt)
- 타입 : FLOAT
- 해당 시각 정각에 측정한 거주, 근무, 방문인구(명)의 합
- 거주/근무/방문인구 : 1~24시 해당 시간 정각에 측정한 인구
  거주지/근무지/방문지와 근무지 외 지역에 머문 시간(분) / 60분
- 시간별 장기체류 외국인 유동인구 데이터를 월별 장기체류 외국인 유동인구 데이터로 정리해 사용
- 행정동별 월 장기체류 외국인 유동인구 수로 변환 후 분석에 사용

3) short_term_frgn (단기체류 외국인 유동인구)

- 타입 : STRING

- 해당 시각 정각에 측정한 방문인구(명)
- 시간별 단기체류 외국인 유동인구 데이터를 월별 단기체류 외국인 유동인구 데이터로 정리해 사용
- 행정동별 월별 단기체류 외국인 유동인구 수로 변환 후 분석에 사용

4) resident (총 거주인구)
- 타입 : INT
- 행정동별 총 거주인구
- 주민등록 거주인구(resid_reg_pop)와 외국인 거주인구(foreign_pop)의 합
- 행정동별 월별 총 거주인구 수 데이터 그대로 사용

5) card_cnt (음식 관련 카드 결제건수)
- 타입 : INT
- 단위 : 건
- 일별 음식 관련 카드 결제건수 데이터를 월별 음식 관련 카드 결제건수 데이터로 정리해 사용
- 행정동별 월 음식 관련 카드 결제건수로 변환 후 분석에 사용

6) card_amt (결제금액)
- 타입 : INT
- 단위 : 원
- 일별 음식 관련 카드 결제금액 데이터를 월별 음식 관련 카드 결제금액 데이터로 정리해 사용
- 행정동별 월 음식 관련 카드 결제금액으로 변환 후 분석에 사용

7) waste_cnt (배출건수)
- 타입 : INT
- 배출거점지역 음식물 쓰레기 배출건수(건)
- 일별 음식물 쓰레기 배출건수 데이터를 월별 음식물 쓰레기 배출건수 데이터로 정리해 사용
- 행정동별 월 음식물 쓰레기 배출건수로 변환 후 분석에 사용

8) detached (단독주택), apt (아파트), town (연립주택), multiplex (다세대주택), commercial_building (비거주용 건물 내 주택)
- 타입 : INT
- 외부데이터 KOSIS의 '주택의 종류별 주택-읍면동(2015,2020), 시군구(2016~2019)' 이용

- 2018년 데이터 : 제주도시, 서귀포시의 년 단위 주택 수
(12개월간의 주택 수를 동일하게 가정)
- 2019년 데이터 : 제주도시, 서귀포시의 년 단위 주택 수
(12개월간의 주택 수를 동일하게 가정)
- 2020년 데이터 : 행정동별 년 단위 주택 수
(12개월간의 주택 수를 동일하게 가정)
- 2021년 데이터 : 데이터가 없어 2020년 데이터 사용

9) distancing (제주 사회적 거리두기 단계)
- 타입 : float
- 외부 데이터 ‘covid19.jeju'와 기사를 통해 자료 수집
- 제주의 사회적 거리두기 단계를 일별로 정리한 후 월평균으로 구하여 이용
- 사회적 거리두기 단계 체계가 등장하기 이전은 0단계라고 가정

10) temp (월 평균기온의 총합)
- 외부 데이터
- 타입 : float
- 외부데이터 기상청과 제주데이터허브를 통해 자료 수집
- 행정동별 일평균기온을 행정동별 월평균기온의 총합으로 정리 후 사용

11) rain (월별 강수량의 총합)
- 외부 데이터
- 타입 : float
- 외부데이터 기상청과 제주데이터허브를 통해 자료 수집
- 행정동별 일강수량을 행정동별 월강수량의 총합으로 정리 후 사용

2.3 반응변수
1) waste_amt (음식물 쓰레기 배출량)
- 타입 : INT
- 배출거점지역의 음식물 쓰레기 총 배출량(g)
- 반응변수(y)
- 일별 음식물 쓰레기 배출량 데이터를 월별 음식물 쓰레기 배출량 데이터로 정리해 사용
- 행정동별 월 음식물 쓰레기 총 배출량으로 변환 후 분석에 사용

3 활용 알고리즘

3.1 상관분석

3.1.1 상관분석의 의미와 선정이유

상관분석(correlation analysis)는 두 변수 간에 어떤 선형적 또는 비선형적 관계를 갖는지 분석하는 방법으로 상관계수를 이용해 측정한다.

상관계수(correlation coefficient)란 두 변수 사이의 선형 관계 정도를 수치화한 계수로 -1~1 사이의 값을 갖는다. 상관계수의 기본 가정은 선형성, 동변량성, 두 변인의 정규분포성, 무선독립표본이다. +1은 완벽한 양의 선형 상관관계, 0은 선형 상관관계 없음, -1은 완벽한 음의 선형 상관관계를 의미한다.

외부데이터 temp(월 평균기온의 총합)와 waste_amt(음식물 쓰레기 배출량), rain(월별 강수량의 총합)과 waste_amt(음식물 쓰레기 배출량) 각각의 관계의 정도를 알아보기 위해 상관분석을 실시했다.

3.1.2 emd_cd, em_cnt, em_g, temp, rain의 상관분석

```
In [22]: corr=food_waste_sort_201819.corr(method='pearson')
         corr
```

Out[22]:

	emd_cd	em_cnt	em_g	temp	rain
emd_cd	1.000000	-0.451129	-0.382820	0.053998	-0.070709
em_cnt	-0.451129	1.000000	0.986177	-0.004420	0.173686
em_g	-0.382820	0.986177	1.000000	-0.022951	0.129840
temp	0.053998	-0.004420	-0.022951	1.000000	0.333577
rain	-0.070709	0.173686	0.129840	0.333577	1.000000

[그림 10] 상관분석 결과

- em_g와 temp의 상관계수 : -0.02
- em_g와 rain의 상관계수 : 0.13
- 두 변수 모두 상관계수가 0에 가까워 음식물 쓰레기 배출량과는 연관이 없다고 판단

3.2 회귀분석

3.2.1 회귀분석의 의미와 선정이유

회귀분석(regression analysis)은 관찰된 연속형 변수들에 대해 변수 간의 인과관계를 밝히고 모형을 적합하여 관심 있는 변수를 예측하거나 추론하기 위한 분석방법이다. 즉, 설명변수로 반응변수가 반응변수에 미치는 영향을 일반화한다. 선형회귀분석은 선형성, 등분산성, 독립성, 비상관성, 정규성을 가정한다.

외부데이터 temp(월 평균기온의 총합), rain(월별 강수량의 총합)과 waste_amt(음식물 쓰레기 배출량) 사이의 인과관계를 살펴보기 위해 회귀분석을 실시했다. 이는 날씨에 따라 음식 배달 빈도가 달라지며, 이에 따라 음식물 쓰레기 배출량 역시 달라질 것이란 생각에서 출발했다.

3.2.2 temp(월 평균기온의 총합)와 waste_amt(음식물 쓰레기 배출량)의 회귀분석

OLS Regression Results

Dep. Variable:	em_g	**R-squared (uncentered):**	0.208
Model:	OLS	**Adj. R-squared (uncentered):**	0.207
Method:	Least Squares	**F-statistic:**	229.1
Date:	Thu, 09 Sep 2021	**Prob (F-statistic):**	4.02e-46
Time:	23:42:53	**Log-Likelihood:**	-17139.
No. Observations:	873	**AIC:**	3.428e+04
Df Residuals:	872	**BIC:**	3.428e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
temp	1.21e+05	7994.329	15.135	0.000	1.05e+05	1.37e+05

Omnibus:	190.241	**Durbin-Watson:**	0.101
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	355.314
Skew:	1.287	**Prob(JB):**	6.99e-78
Kurtosis:	4.773	**Cond. No.**	1.00

[그림 11] 회귀분석 결과

- 결정계수(R^2)는 0.208, 수정결정계수(Adjust R^2)는 0.207로 waste_amt의 변동의 약 20%만이 temp의 변동에 의해 설명됨을 의미
- temp와 waste_amt 사이의 회귀식의 정확도는 매우 낮으므로 설명변수 temp로서의 역할을 하기 어렵다고 판단해 제거

3.2.3 rain(월별 강수량의 총합)와 waste_amt(음식물 쓰레기 배출량)의 회귀분석

[4]:

OLS Regression Results

Dep. Variable:	em_g	**R-squared (uncentered):**	0.327
Model:	OLS	**Adj. R-squared (uncentered):**	0.327
Method:	Least Squares	**F-statistic:**	424.6
Date:	Thu, 09 Sep 2021	**Prob (F-statistic):**	3.62e-77
Time:	23:44:11	**Log-Likelihood:**	-17068.
No. Observations:	873	**AIC:**	3.414e+04
Df Residuals:	872	**BIC:**	3.414e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
rain	1.603e+05	7780.776	20.605	0.000	1.45e+05	1.76e+05

Omnibus:	138.458	**Durbin-Watson:**	0.392
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	260.269
Skew:	0.948	**Prob(JB):**	3.04e-57
Kurtosis:	4.888	**Cond. No.**	1.00

[그림 12] 회귀분석 결과

- 결정계수(R^2)와 수정결정계수(Adjust R^2)가 0.327로 waste_amt의 변동의약 32.70%만이 rain 변동에 의해 설명됨을 의미
- rain과 waste_amt 사이의 회귀식의 정확도는 매우 낮으므로 설명변수 rain으로서의 역할을 하기 어렵다고 판단해 제거

3.2.4 temp(월 평균기온의 총합), rain(월별 강수량의 총합)와 waste_amt(음식물 쓰레기 배출량)의 회귀분석

OLS Regression Results

Dep. Variable:	em_g	**R-squared:**	0.022
Model:	OLS	**Adj. R-squared:**	0.020
Method:	Least Squares	**F-statistic:**	9.694
Date:	Thu, 09 Sep 2021	**Prob (F-statistic):**	6.86e-05
Time:	23:26:57	**Log-Likelihood:**	-16915.
No. Observations:	873	**AIC:**	3.384e+04
Df Residuals:	870	**BIC:**	3.385e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.977e+07	3.25e+06	18.409	0.000	5.34e+07	6.61e+07
temp	-1.817e+04	8667.493	-2.096	0.036	-3.52e+04	-1156.920
rain	4.275e+04	9829.308	4.350	0.000	2.35e+04	6.2e+04

Omnibus:	265.817	**Durbin-Watson:**	0.170
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	608.572
Skew:	1.669	**Prob(JB):**	7.08e-133
Kurtosis:	5.363	**Cond. No.**	654.

[그림 13] 회귀분석 결과

- 결정계수(R^2)는 0.022, 수정결정계수(Adjust R^2)는 0.020로 waste_amt의 변동의 약 2%만이 temp와 rain 변동에 의해 설명됨을 의미
- temp, rain과 waste_amt 사이의 회귀식의 정확도는 매우 낮으므로 설명변수 temp, rain으로서의 역할을 하기 어렵다고 판단해 제거

3.3 XGBoost

3.3.1 XGBoost의 개념과 선정이유

XGBoost는 머신러닝 기법 중 의사결정나무(Decision tree)를 기반으로 한 앙상블 방법으로 Boosting을 기반으로 한다. XGBoost는 효율성과 유연성, 휴대성이 뛰어나고 여러 파라미터를 조절해가며 최적의 모델을 만드는 유연한 러닝 시스템을 가진다. 또, 과적합(over-fitting)을 방지하고 시각화가 쉬우며 빠르게 학습하고 예측할 수 있으면서도 높은 성능을 나타내 현업에서 많이 사용된다. 이런 이유로 XGBoost를 활용해 음식물 쓰레기 배출량의 주요 요인을 선정하고자 한다.

3.3.2 XGBoost 통해 최적 모델 찾기

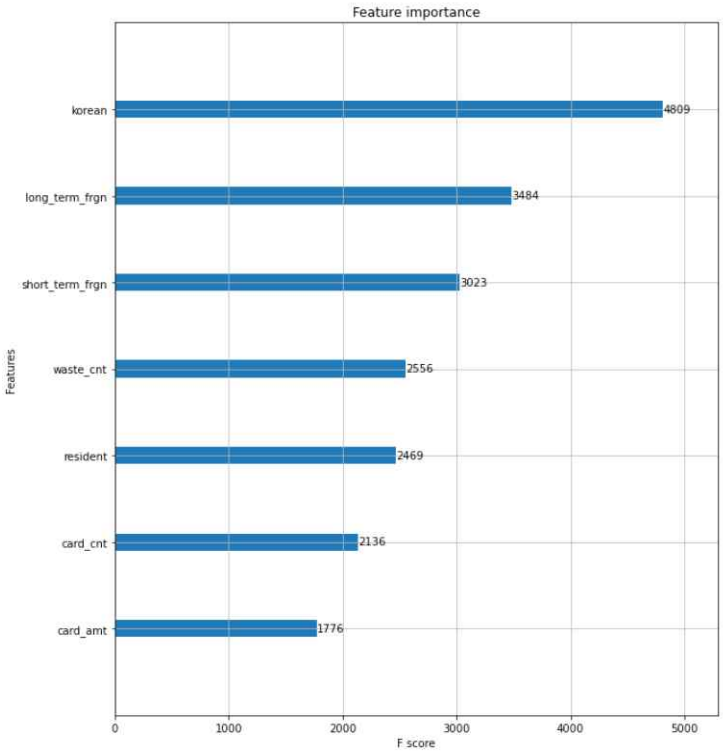
- 설명변수 waste_amt(음식물 쓰레기 배출량, g)의 범위가 매우 크기 때문에 RMSE 대신 RMSLE 사용
- RMSLE란 예측값과 실제값에 로그를 씌운 후에 차이를 비교하는 방법

① ver 1
- 제공데이터의 변수만을 이용해 XGBoost 모델링
- 설명변수(x) :

korean (내국인 유동인구)	card_cnt (음식 관련 카드 결제건수)	resident (총 거주인구)
long_term_frgn (장기체류 외국인 유동인구)	card_amt (결제금액)	
short_term_frgn (단기체류 외국인 유동인구)	waste_cnt (배출건수)	

[표 1] ver 1의 설명변수

- 전체 데이터셋을 학습용 80%, 테스트용 20%로 분할
- max_depth(트리 당 최대 깊이)가 7, eta(학습률)가 0.1, 목적함수(objective)가 reg:linear(회귀)인 하이퍼 파라미터 설정하고 xgboost 실행
- 오류함수의 평가성능지표는 rmsle
- 부스팅 반복횟수는 1000
- 조기중단을 위한 최소 반복횟수는 100
- 결과 : eval-rmsle = 0.09863



[그림 14] XGBoost ver 1에서 요인들의 중요도

2 ver 2-1

- 제공데이터의 변수만을 이용해 XGBoost 모델링
- ver 1 의 변수 중 korean, long_term_frgn을 거주/근무/방문인구로 분류
- ver 1 의 변수 중 resident를 - 주민등록/외국인 거주인구(foreign_pop) & 성별(fe, m)로 분류
- ver 1 의 변수 중 card_cnt 를 home(배달, 간식, 식품, 마트/슈퍼마켓, 농축수산물) / out(한식, 패스트푸드, 주점 및 주류 판매, 양식, 뷔페, 아시아음식)으로 분류

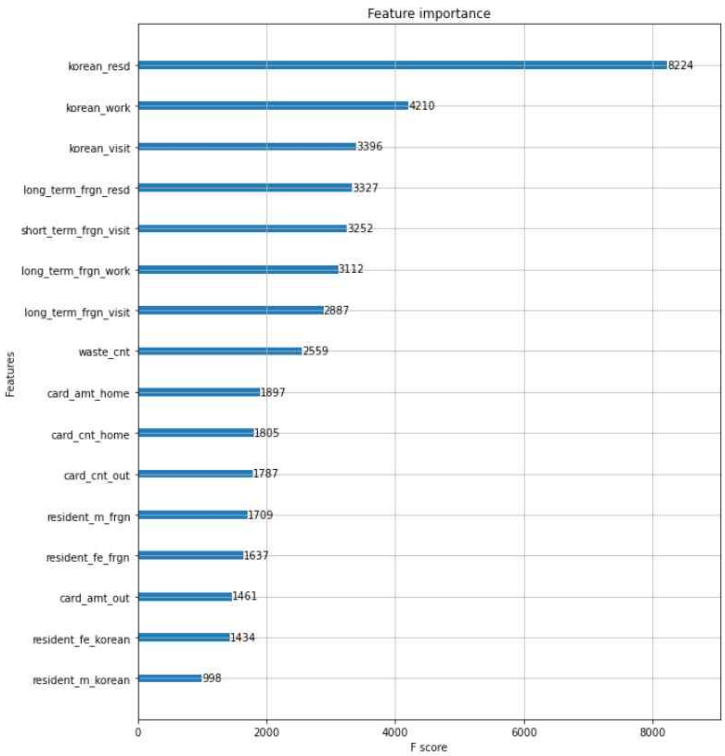
- 설명변수(x) :

korean_resd (내국인 유동인구, 거주인구)	resident_fe_frgn (총 거주인구, 여성, 외국인 거주인구)
korean_work (내국인 유동인구, 근무인구)	resident_m_korean (총 거주인구, 남성, 주민등록 거주인구)
korean_visit (내국인 유동인구, 방문인구)	resident_m_frgn (총 거주인구, 남성, 외국인 거주인구)
long_term_frgn_resd (장기체류 외국인 유동인구, 거주인구)	card_cnt_home (음식 관련 카드 결제건수, 가정)

long_term_frgn_work (장기체류 외국인 유동인구, 근무인구)	card_amt_home (결제금액, 가정)
long_term_frgn_visit (장기체류 외국인 유동인구, 방문인구)	card_cnt_out (음식 관련 카드 결제건수, 외식)
short_term_frgn_visit (단기체류 외국인 유동인구)	card_amt_out (결제금액, 외식)
resident_fe_korean (총 거주인구, 여성, 주민등록 거주인구)	waste_cnt (배출건수)

[표 2] ver 2_1의 설명변수

- 전체 데이터셋을 학습용 80%, 테스트용 20%로 분할
- max_depth(트리 당 최대 깊이)가 7, eta(학습률)가 0.1, 목적함수(objective)가 reg:linear(회귀)인 하이퍼 파라미터 설정하고 xgboost 실행
- 이 때, 오류함수의 평가성능지표는 rmsle
- 부스팅 반복횟수는 1000
- 조기중단을 위한 최소 반복횟수는 100
- 결과 : eval-rmsle = 0.08959
- ver 1(0.09863) > ver 2-1(0.08959)
- ver 2-1 선택



[그림 15] XGBoost ver 2_1에서 요인들의 중요도

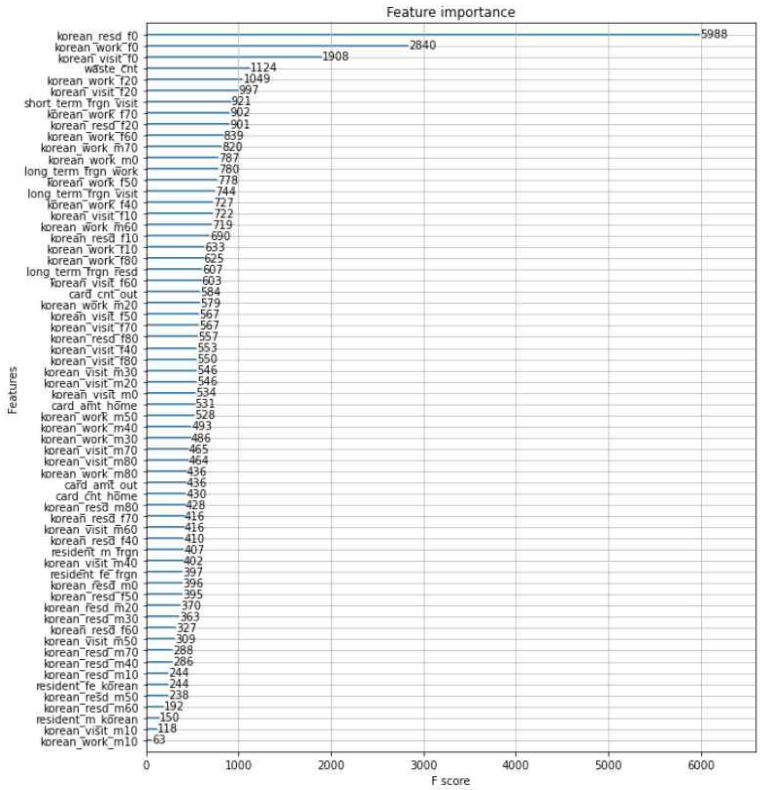
3 ver 2-2

- 제공데이터의 변수만을 이용해 XGBoost 모델링
- ver 2-1 의 변수 중 korean_resd, korean_work, korean_visit을 각각 성별, 연령대로 분류
- 설명변수 (x) :

<table>
<tr><td rowspan="2">korean_resd_
(내국인 유동인구, 거주인구)</td><td>f0 ~ f80 (여성, 0대 ~ 80대)</td><td rowspan="2">resident_m_korean
(총 거주인구, 남성, 주민등록 거주인구)</td></tr>
<tr><td>m0 ~ m80 (남성, 0대 ~ 80대)</td></tr>
<tr><td rowspan="2">korean_work_
(내국인 유동인구, 근무인구)</td><td>f0 ~ f80 (여성, 0대 ~ 80대)</td><td rowspan="2">resident_m_frgn
(총 거주인구, 남성, 외국인 거주인구)</td></tr>
<tr><td>m0 ~ m80 (남성, 0대 ~ 80대)</td></tr>
<tr><td rowspan="2">korean_visit_
(내국인 유동인구, 방문인구)</td><td>f0 ~ f80 (여성, 0대 ~ 80대)</td><td rowspan="2">card_cnt_home
(음식 관련 카드 결제건수, 가정)</td></tr>
<tr><td>m0 ~ m80 (남성, 0대 ~ 80대)</td></tr>
<tr><td colspan="2">long_term_frgn_resd
(장기체류 외국인 유동인구, 거주인구)</td><td>card_amt_home
(결제금액, 가정)</td></tr>
<tr><td colspan="2">long_term_frgn_work
(장기체류 외국인 유동인구, 근무인구)</td><td>card_cnt_out
(음식 관련 카드 결제건수, 외식)</td></tr>
<tr><td colspan="2">long_term_frgn_visit
(장기체류 외국인 유동인구, 방문인구)</td><td>card_amt_out
(결제금액, 외식)</td></tr>
<tr><td colspan="2">short_term_frgn_visit
(단기체류 외국인 유동인구)</td><td>waste_cnt
(배출건수)</td></tr>
<tr><td colspan="2">resident_fe_korean
(총 거주인구, 여성, 주민등록 거주인구)</td><td></td></tr>
<tr><td colspan="2">resident_fe_frgn
(총 거주인구, 여성, 외국인 거주인구)</td><td></td></tr>
</table>

[표 3] ver 2_2의 설명변수

- 전체 데이터셋을 학습용 80%, 테스트용 20%로 분할
- max_depth(트리 당 최대 깊이)가 7, eta(학습률)가 0.1, 목적함수(objective)가 reg:linear(회귀)인 하이퍼 파라미터 설정하고 xgboost 실행
- 이 때, 오류함수의 평가성능지표는 rmsle
- 부스팅 반복횟수는 1000
- 조기중단을 위한 최소 반복횟수는 100
- 결과 : eval-rmsle = 0.08853
- ver 2-1(0.08959) > ver 2-2(0.08853)
- ver 2-2 선택



[그림 16] XGBoost ver 2_2에서 요인들의 중요도

4 ver 3-1

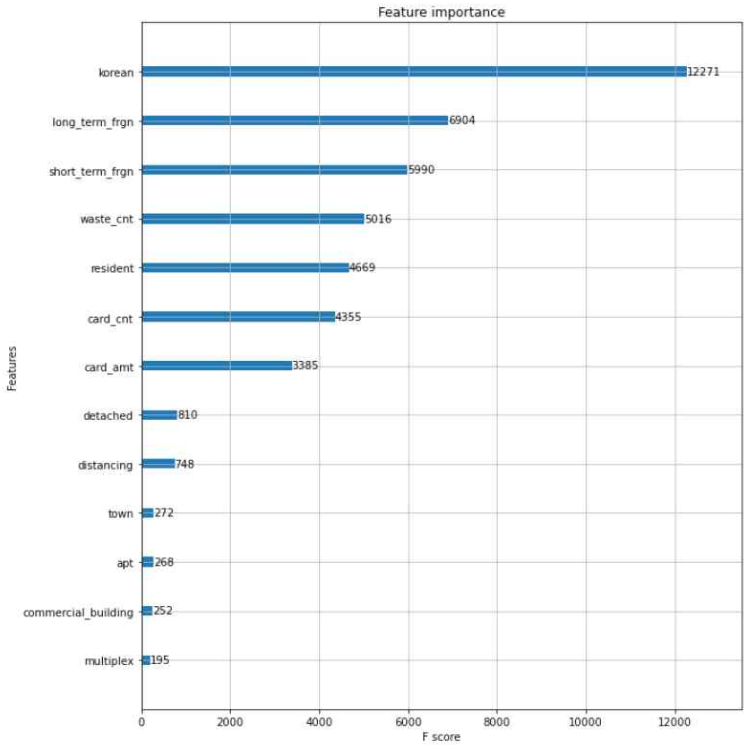
- ver 1의 제공데이터 변수 + 외부데이터 total_house(총 주택 수)의 종류(단독주택, 아파트, 연립주택, 다세대주택, 비거주용 건물내 주택) + distancing(제주 코로나 단계)를 이용해 XGBoost 모델링

- 설명변수 (x) :

korean (내국인 유동인구)	detached (단독주택)	waste_cnt (배출건수)
long_term_frgn (장기체류 외국인 유동인구)	apt (아파트)	
short_term_frgn (단기체류 외국인 유동인구)	town (연립주택)	
resident (총 거주인구)	multiplex (다세대주택)	
card_cnt (음식 관련 카드 결제건수)	commercial_building (비거주용 건물내 주택)	
card_amt (결제금액)	distancing (제주 코로나 단계)	

[표 4] ver 3_1의 설명변수

- 전체 데이터셋을 학습용 80%, 테스트용 20%로 분할
- max_depth(트리 당 최대 깊이)가 7, eta(학습률)가 0.1, 목적함수(objective)가 reg:linear(회귀)인 하이퍼 파라미터 설정하고 xgboost 실행
- 이 때, 오류함수의 평가성능지표는 rmsle
- 부스팅 반복횟수는 1000
- 조기중단을 위한 최소 반복횟수는 100
- 결과 : eval-rmsle = 0.07914
- ver 2-2(0.08853) > ver 3-1(0.07914)
- ver 3-1 선택



[그림 17] XGBoost ver 3_1에서 요인들의 중요도

5 ver 3-2

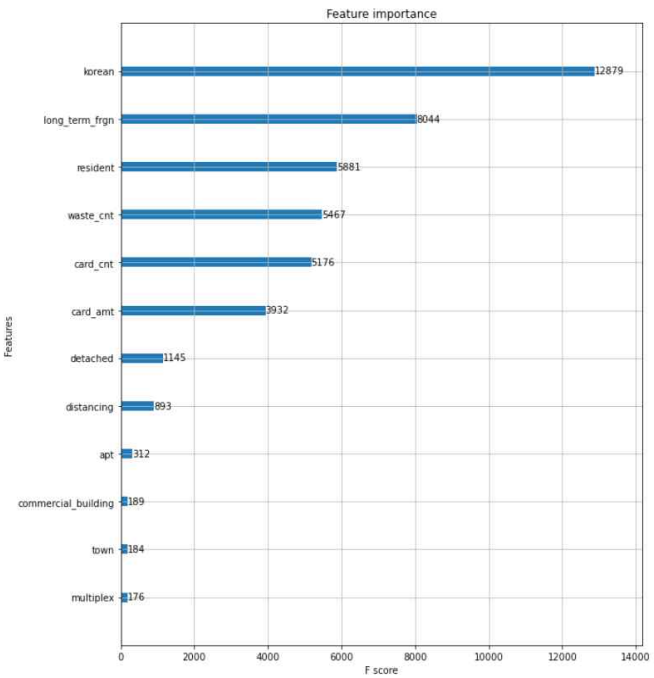
- 시계열 ARIMA 모델링 결과 short_term_frgn(단기체류 외국인 유동인구)의 2021년 7, 8월 값 중 마이너스(-) 다수 존재
- ver 3-1의 변수 중 short_term_frgn(단기체류 외국인 유동인구)를 제거하고 XGBoost 모델링

- 설명변수 (x) :

korean (내국인 유동인구)	detached (단독주택)
long_term_frgn (장기체류 외국인 유동인구)	apt (아파트)
resident (총 거주인구)	town (연립주택)
card_cnt (음식 관련 카드 결제건수)	multiplex (다세대주택)
card_amt (결제금액)	commercial_building (비거주용 건물내 주택)
waste_cnt (배출건수)	distancing (제주 코로나 단계)

[표 5] ver 3_2의 설명변수

- 전체 데이터셋을 학습용 80%, 테스트용 20%로 분할
- max_depth(트리 당 최대 깊이)가 7, eta(학습률)가 0.1, 목적함수(objective)가 reg:linear(회귀)인 하이퍼 파라미터 설정하고 xgboost 실행
- 오류함수의 평가성능지표는 rmsle
- 부스팅 반복횟수는 1000
- 조기중단을 위한 최소 반복횟수는 100
- 결과 : eval-rmsle = 0.07985
- ver 3-1(0.07914) < ver 3-2(0.07985)
- ver 3-1 선택



[그림 18] XGBoost ver 3_2에서 요인들의 중요도

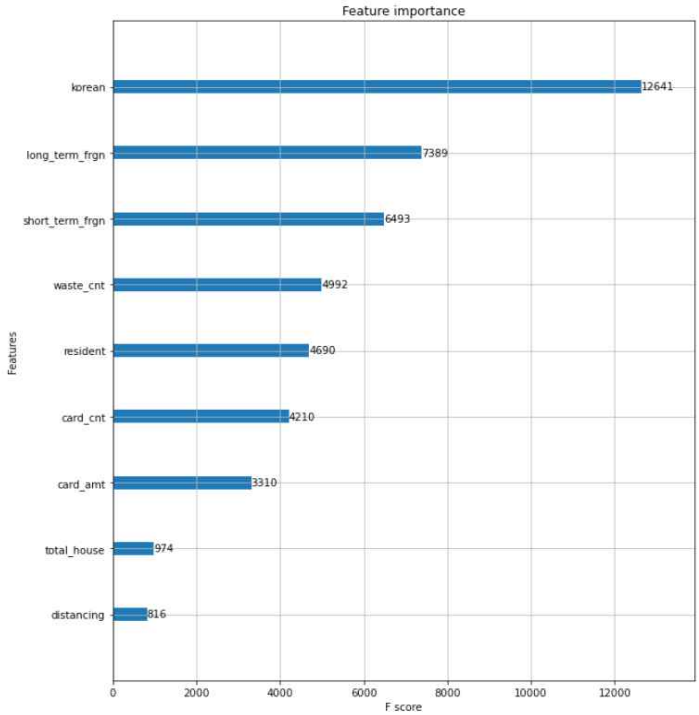
⑥ ver 3_3

- ver 3_1의 total_house(총 주택 수)의 모든 종류를 통합한 total_house만 사용
- ver 1의 제공데이터 변수의 제공데이터 변수 + 외부데이터 total_house(총 주택 수) + distancing(제주 코로나 단계)를 이용해 XGBoost 모델링
- 설명변수(x) :

korean (내국인 유동인구)	card_amt (결제금액)
long_term_frgn (장기체류 외국인 유동인구)	waste_cnt (배출건수)
short_term_frgn (단기체류 외국인 유동인구)	total_house (총 주택 수)
resident (총 거주인구)	distancing (제주 코로나 단계)
card_cnt (음식 관련 카드 결제건수)	

[표 6] ver 3_3의 설명변수

- 전체 데이터셋을 학습용 80%, 테스트용 20%로 분할
- max_depth(트리 당 최대 깊이)가 7, eta(학습률)가 0.1, 목적함수(objective)가 reg:linear(회귀)인 하이퍼 파라미터 설정하고 xgboost 실행
- 이 때, 오류함수의 평가성능지표는 rmsle
- 부스팅 반복횟수는 1000
- 조기중단을 위한 최소 반복횟수는 100
- 결과 : eval-rmsle = 0.08281
- ver 3-1(0.07914) < ver 3-3(0.08281)
- ver 3-1 선택



[그림 19] XGBoost ver 3_3에서 요인들의 중요도

⑦ 최종 모델 선택

- 최종적으로 다음 변수를 설명변수로 이용한 XGBoost를 선택

korean (내국인 유동인구)	detached (단독주택)
long_term_frgn (장기체류 외국인 유동인구)	apt (아파트)
short_term_frgn (단기체류 외국인 유동인구)	town (연립주택)
resident (총 거주인구)	multiplex (다세대주택)
card_cnt (음식 관련 카드 결제건수)	commercial_building (비거주용 건물내 주택)
card_amt (결제금액)	distancing (제주 코로나 단계)
waste_cnt (배출건수)	

[표 7] 최종모델(ver 3_1)의 설명변수

3.4 ARIMA model

3.4.1 ARIMA 모형의 의미와 선정이유

ARIMA(Autoregressive Integrated Moving Average) 모형은 시계열 분석(Time Series Analysis) 모형 중 하나이다. 시계열 모형에는 대표적으로 AR 모형, MA 모형, ARMA 모형, ARIMA 모형이 있다. AR(자기상관) 모형은 random variable에 대해 이전 값이 이후에 값에 영향을 미치고 있는 상황을 이야기한다. MA(이동평균) 모형은 시간이 지날수록 어떤 random variable의 평균값이 지속적으로 증가하거나 감소하는 경향이 생길 수 있다. ARMA 모형은 AR 모형과 MA 모형을 합친 모형이다. ARMA 모형이 과거의 데이터를 사용하는 것에 비해 ARIMA 모형은 이를 넘어 과거의 데이터가 지닌 추세(trend)까지 반영한다. 즉, correlation뿐 아니라 cointegration까지 고려한 모델이다.

설명변수들이 추세와 계절적 요인을 지니고, 과거 데이터를 사용해 미래 값을 예측할 수 있는 시계열 ARIMA 모형을 선택했다.

본 분석과정에서 외부변수인 종류별 주택 수 detached(단독주택), apt(아파트), town(연립주택), multiplex(다세대주택), commercial_building(비거주용 건물내 주택)는 시계열 변수보다 최근 데이터와 유사할 것이기에 ARIMA 모형을 이용하지 않고 KOSIS가 제공하는 가장 최근 데이터 2020년 데이터를 이용한다. 또 다른 외부변수 distancing(제주 코로나 단계)는 시계열 데이터가 아니고 질병관리청에서 지정하는 것이므로 ARIMA 모형에 적용하지 않는다.

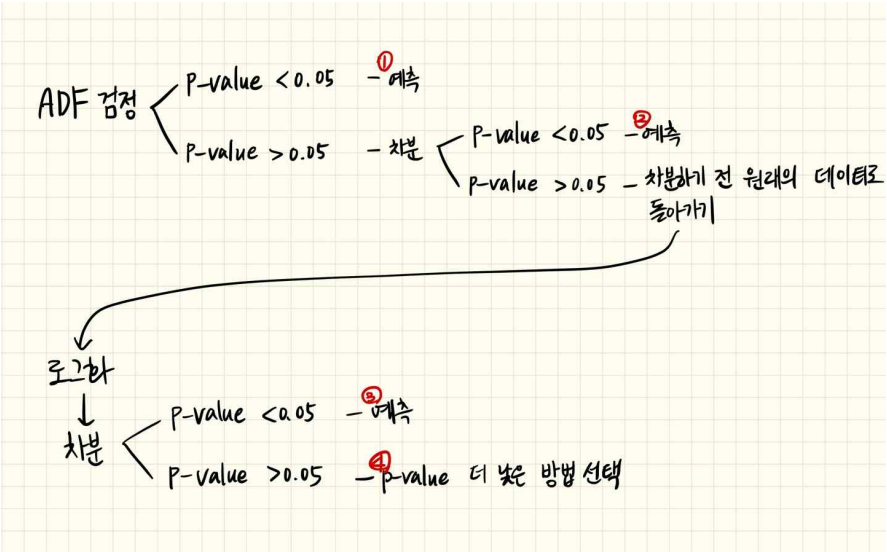
3.4.2 시계열 데이터의 정상성 확인하기

ARIMA 예측모델은 시계열 데이터의 정상성을 가정한다. 여기서 정상성이란 시계열 데이터의 특성이 사건의 흐름에 따라 변하지 않는다는 것을 의미한다. 우리 데이터의 각 변수들은 추세나 계절 요인이 시간이 경과하면서 관측값에 영향을 미치기 때문에 비정상적이다. 따라서 각 데이터들을 정상성을 따르도록 전처리할 필요가 있다. 정상성을 만족하는지 여부는 ADF 검정(귀무가설 : 정상성을 만족하지 않는다)을 통해 확인한다. ①ADF 검정에서 p-value가 0.05보다 작으면 정상적 데이터이므로 그대로 ARIMA 모델에 적용한다. 반면 p-value가 0.05보다 크면 비정상적 데이터이므로 전처리 과정이 필요하다. 전처리 과정은 다음 2가지를 이용했다.

② 차분 ③ 로그화 후 차분

만약 ②와 ③ 과정을 통해서도 정상성을 만족하지 않는 경우, p-value가 더 낮은

방법을 택했다.



[그림 20] 정상성 만족하는 데이터 선택 과정

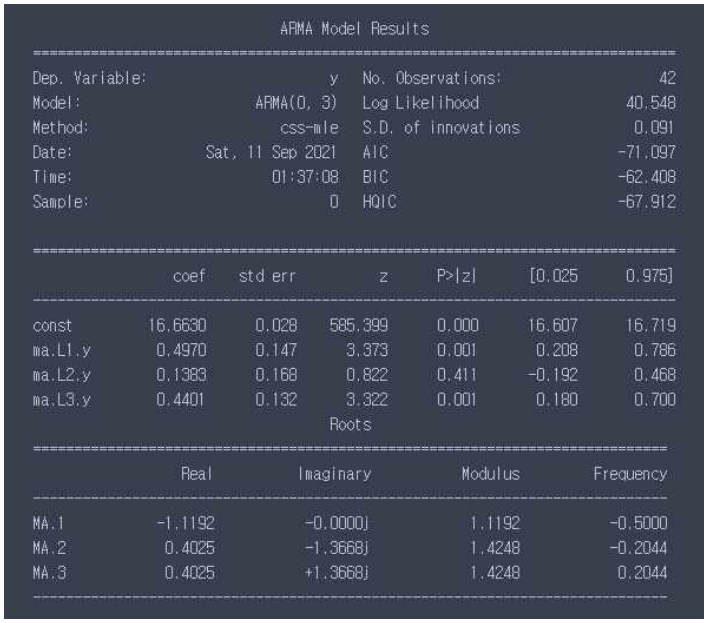
3.4.3 최적화 ARIMA 모형 찾기

ARIMA 모형은 p(AR모형의 lag), d(차분 횟수), q(MA모형의 lag) 3가지 파라미터를 필요로 하는데 각 파라미터 값에 따른 최종결과가 달라질 수 있다. 따라서 우리는 최적의 모델을 선정하기 위해 자동으로 최적화된 모형을 선정해주는 방법 auto_arima 라이브러리를 통해 최종 모델을 택했다.

```
In [8]:
from pmdarima.arima import auto_arima

model_arima= auto_arima(sub_df_log,trace=True, error_action='ignore',suppress_warnings=True,stepwise=False,seasona

model_arima.fit(sub_df_log)
```

```
ARIMA(0,0,0)(0,0,0)[1] intercept   : AIC=-64.907, Time=0.04 sec
ARIMA(0,0,1)(0,0,0)[1] intercept   : AIC=-68.786, Time=0.13 sec
ARIMA(0,0,2)(0,0,0)[1] intercept   : AIC=-68.664, Time=0.08 sec
ARIMA(0,0,3)(0,0,0)[1] intercept   : AIC=-71.097, Time=0.16 sec
ARIMA(0,0,4)(0,0,0)[1] intercept   : AIC=-69.414, Time=0.28 sec
ARIMA(0,0,5)(0,0,0)[1] intercept   : AIC=-68.843, Time=0.45 sec
ARIMA(1,0,0)(0,0,0)[1] intercept   : AIC=-70.618, Time=0.06 sec
ARIMA(1,0,1)(0,0,0)[1] intercept   : AIC=-69.002, Time=0.25 sec
ARIMA(1,0,2)(0,0,0)[1] intercept   : AIC=-67.615, Time=0.47 sec
ARIMA(1,0,3)(0,0,0)[1] intercept   : AIC=-68.574, Time=0.32 sec
ARIMA(1,0,4)(0,0,0)[1] intercept   : AIC=-67.488, Time=0.34 sec
ARIMA(2,0,0)(0,0,0)[1] intercept   : AIC=-69.221, Time=0.13 sec
ARIMA(2,0,1)(0,0,0)[1] intercept   : AIC=-67.210, Time=0.36 sec
ARIMA(2,0,2)(0,0,0)[1] intercept   : AIC=-65.905, Time=0.24 sec
ARIMA(2,0,3)(0,0,0)[1] intercept   : AIC=-66.456, Time=0.52 sec
ARIMA(3,0,0)(0,0,0)[1] intercept   : AIC=-67.231, Time=0.26 sec
ARIMA(3,0,1)(0,0,0)[1] intercept   : AIC=-65.057, Time=0.19 sec
ARIMA(3,0,2)(0,0,0)[1] intercept   : AIC=-63.197, Time=0.45 sec
ARIMA(4,0,0)(0,0,0)[1] intercept   : AIC=-67.376, Time=0.57 sec
ARIMA(4,0,1)(0,0,0)[1] intercept   : AIC=-65.203, Time=0.17 sec
ARIMA(5,0,0)(0,0,0)[1] intercept   : AIC=-65.433, Time=0.29 sec

Best model:  ARIMA(0,0,3)(0,0,0)[1] intercept
Total fit time: 5.764 seconds


ARIMA(order=(0, 0, 3), scoring_args={}, seasonal_order=(0, 0, 0, 1),
      suppress_warnings=True)
```

[그림 21] 최적화 ARIMA 모형 찾는 코드



ARMA Model Results

Dep. Variable:	y	No. Observations:	42
Model:	ARMA(0, 3)	Log Likelihood	40.548
Method:	css-mle	S.D. of innovations	0.091
Date:	Sat, 11 Sep 2021	AIC	-71.097
Time:	01:37:08	BIC	-62.408
Sample:	0	HQIC	-67.912

	coef	std err	z	P>|z|	[0.025	0.975]
const	16.6630	0.028	585.399	0.000	16.607	16.719
ma.L1.y	0.4970	0.147	3.373	0.001	0.208	0.786
ma.L2.y	0.1383	0.168	0.822	0.411	-0.192	0.468
ma.L3.y	0.4401	0.132	3.322	0.001	0.180	0.700

Roots

	Real	Imaginary	Modulus	Frequency
MA.1	-1.1192	-0.0000j	1.1192	-0.5000
MA.2	0.4025	-1.3668j	1.4248	-0.2044
MA.3	0.4025	+1.3668j	1.4248	0.2044

[그림 22] ARIMA 모델 fitting 결과

3.4.4 각 변수의 2021년 7, 8월 값 예측하기

위 방법을 통해 택한 최종 모형을 토대로 2021년 7, 8월의 값을 예측한다.

3 예측하기

```
In [10]:
# 2단위 이후의 예측결과
fore = model_fit.forecast(steps=2)
print(fore)
```

```
(array([16.72822029, 16.66296014]), array([0.09093318, 0.10154369]), array([[16.54999453, 16.90644605],
       [16.46393836, 16.86198191]]))
```

```
In [11]:
print( np.exp(16.72822029) - 1 )
print( np.exp(16.66296014) - 1 )
```

```
18406607.02683503
17243745.94875778
```

[그림 23] 7, 8월 값 예측 결과

3.4.5 최종 예측값

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		base_date	emd_cd	korean	long_term	short_tern	resident	card_cnt	card_amt	waste_cnt	detached	apt	town	multiplex	commerci	distancing
2	0	2021-07-31 0:00	50110250	8684900	469672.4	387.9343	24457.77	292.3133	10681272	23313.07	5542	954	1085	608	277	1.83871
3	1	2021-08-31 0:00	50110250	8084718	487440	0	24460.11	294.1351	10741037	29166.39	5542	954	1085	608	277	3.451613
4	2	2021-07-31 0:00	50110253	18406607	767679.8	24078.52	38348.8	612.1654	19944449	43546.58	9367	1059	1739	1601	335	1.83871
5	3	2021-08-31 0:00	50110253	17243746	738378.2	20402.22	38397.63	616.3762	20061969	44514.03	9367	1059	1739	1601	335	3.451613
6	4	2021-07-31 0:00	50110256	7144597	171266	564.4438	15952.53	201.4658	6808693	14990.38	5532	134	231	380	244	1.83871
7	5	2021-08-31 0:00	50110256	6598143	173496.5	0	15948.57	203.3238	6853216	15767.14	5532	134	231	380	244	3.451613
8	6	2021-07-31 0:00	50110259	10780175	220062.2	2557.875	26001.59	412.395	13038767	30488.83	6281	759	1493	1108	254	1.83871
9	7	2021-08-31 0:00	50110259	9823373	225628.4	1513.767	26048.55	415.072	13118792	37415.76	6281	759	1493	1108	254	3.451613
10	8	2021-07-31 0:00	50110310	3504487	177430.4	408.2155	9474.992	107.4637	3917870	8443.9	3864	24	217	81	104	1.83871
11	9	2021-08-31 0:00	50110310	3530873	159549.1	72.78404	9423.994	108.2395	3948271	8857.604	3864	24	217	81	104	3.451613
12	10	2021-07-31 0:00	50110320	903454.2	106253.9	4035.386	1865.073	16.12822	733150.2	0	783	0	0	20	21	1.83871
13	11	2021-08-31 0:00	50110320	899199.8	140057	3843.123	1859.641	22.74553	733150.2	0	783	0	0	20	21	3.451613
14	12	2021-07-31 0:00	50110330	1036901	84212.47	2654.885	1749.101	24.44483	1314676	0	658	0	6	7	50	1.83871
15	13	2021-08-31 0:00	50110330	851732.2	89352.46	3264.023	1747.041	24.71129	1319304	0	658	0	6	7	50	3.451613
16	14	2021-07-31 0:00	50110510	1955595	65233.29	0	2545.641	71.99481	2637701	4054.042	439	115	15	143	84	1.83871
17	15	2021-08-31 0:00	50110510	1936183	65019.21	0	2547.373	73.17678	2767798	6124.225	439	115	15	143	84	3.451613
18	16	2021-07-31 0:00	50110520	7728271	80042.41	0	32894.81	720.3612	21271597	94619.08	3584	5116	411	1302	203	1.83871
19	17	2021-08-31 0:00	50110520	7755445	79387.23	0	32850.18	722.4284	21316556	96909.37	3584	5116	411	1302	203	3.451613
20	18	2021-07-31 0:00	50110530	6778608	96754.09	212.439	7660.252	215.8736	6036634	14378.7	726	1226	64	239	138	1.83871
21	19	2021-08-31 0:00	50110530	6787681	97579.11	0	7650.038	216.828	6045260	19980.01	726	1226	64	239	138	3.451613
22	20	2021-07-31 0:00	50110540	19556448	224018	591.4194	50149.84	1258.694	35389538	130913.5	3609	6553	2521	3647	486	1.83871
23	21	2021-08-31 0:00	50110540	19470913	211817.2	0	50198.9	1264.613	35491471	126991.9	3609	6553	2521	3647	486	3.451613
24	22	2021-07-31 0:00	50110550	4613551	60672.98	0	13622.07	345.7505	10140450	34512.6	1879	1147	179	712	153	1.83871
25	23	2021-08-31 0:00	50110550	4567034	61316.68	0	13599.81	347.1392	10169719	34958.25	1879	1147	179	712	153	3.451613

[그림 24] ARIMA 모델로 예측한 결과

4 최종예측 결과

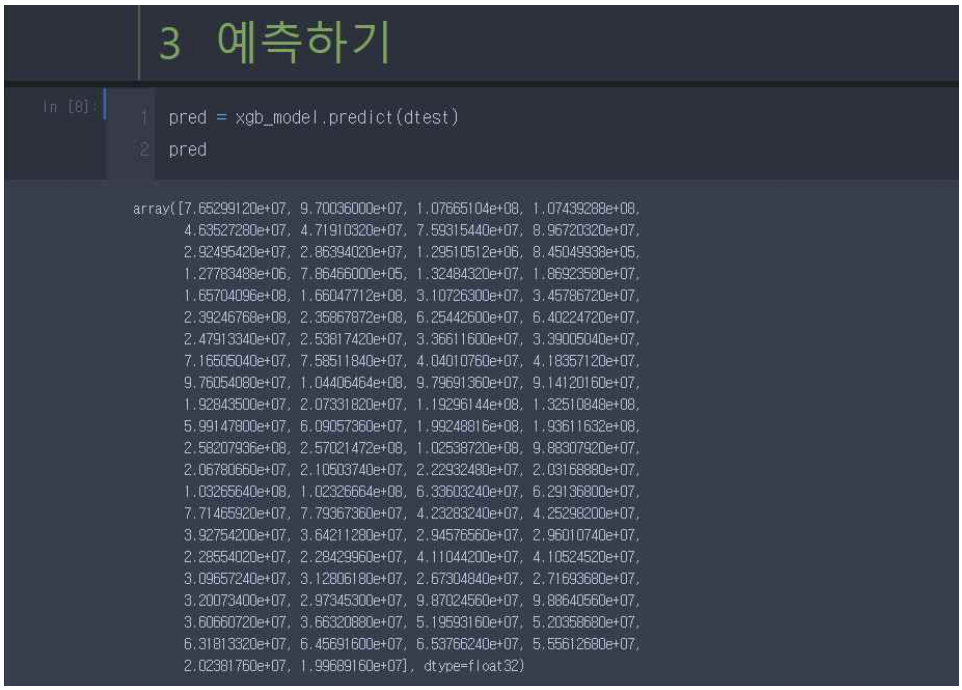
4.1 행정동별 음식물 쓰레기 배출량 예측

- XGBoost ver_3_1의 모델을 사용하여 예측함

korean (내국인 유동인구)	detached (단독주택)
long_term_frgn (장기체류 외국인 유동인구)	apt (아파트)
short_term_frgn (단기체류 외국인 유동인구)	town (연립주택)
resident (총 거주인구)	multiplex (다세대주택)
card_cnt (음식 관련 카드 결제건수)	commercial_building (비거주용 건물내 주택)
card_amt (결제금액)	distancing (제주 코로나 단계)
waste_cnt (배출건수)	

[표 8] 최종모델(ver 3_1)의 설명변수

- 학습용 : (18.01.~21.06. 데이터), 테스트용 : (21.07.~21.08. 데이터)
- 이 때 테스트용 데이터의 X_feature 값들은 ARIMA 모형으로 예측한 값임



In [12]:

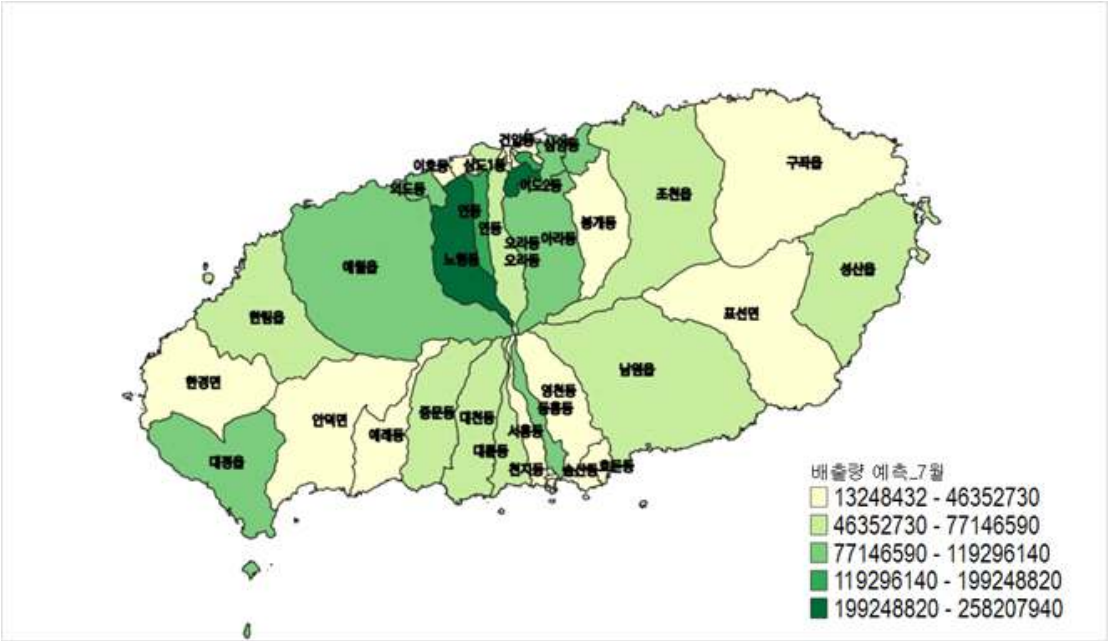
```
result1 = pd.pivot_table(result, index='emd_cd', columns='base_date', values='predict')
result1
```

base_date	2021-07-31	2021-08-31
emd_cd		
50110250	7.652991e+07	9.700360e+07
50110253	1.076651e+08	1.074393e+08
50110256	4.635273e+07	4.719103e+07
50110259	7.593154e+07	8.967203e+07
50110310	2.924954e+07	2.863940e+07
50110320	1.295105e+06	8.450499e+05
50110330	1.277835e+06	7.864660e+05
50110510	1.324843e+07	1.869236e+07
50110520	1.657041e+08	1.660477e+08
50110530	3.107263e+07	3.457867e+07
50110540	2.392468e+08	2.358679e+08
50110550	6.254426e+07	6.402247e+07
50110560	2.479133e+07	2.538174e+07
50110570	3.366116e+07	3.390050e+07
50110580	7.165050e+07	7.585118e+07
50110590	4.040108e+07	4.183571e+07
50110600	9.760541e+07	1.044065e+08
50110610	9.796914e+07	9.141202e+07
50110620	1.928435e+07	2.073318e+07
50110630	1.192961e+08	1.325108e+08
50110640	5.991478e+07	6.090574e+07
50110650	1.992488e+08	1.936116e+08
50110660	2.582079e+08	2.570215e+08
50110670	1.025387e+08	9.883079e+07
50110680	2.067807e+07	2.105037e+07
50110690	2.229325e+07	2.031689e+07
50130250	1.032656e+08	1.023267e+08
50130253	6.336032e+07	6.291368e+07
50130259	7.714659e+07	7.793674e+07
50130310	4.232832e+07	4.252982e+07
50130320	3.927542e+07	3.642113e+07
50130510	2.945766e+07	2.960107e+07
50130520	2.285540e+07	2.284300e+07
50130530	4.110442e+07	4.105245e+07
50130540	3.096572e+07	3.128062e+07
50130550	2.673048e+07	2.716937e+07
50130560	3.200734e+07	2.973453e+07
50130570	9.870246e+07	9.886406e+07
50130580	3.606607e+07	3.663209e+07
50130590	5.195932e+07	5.203587e+07
50130600	6.318133e+07	6.456916e+07
50130610	6.537662e+07	5.556127e+07
50130620	2.023818e+07	1.996892e+07

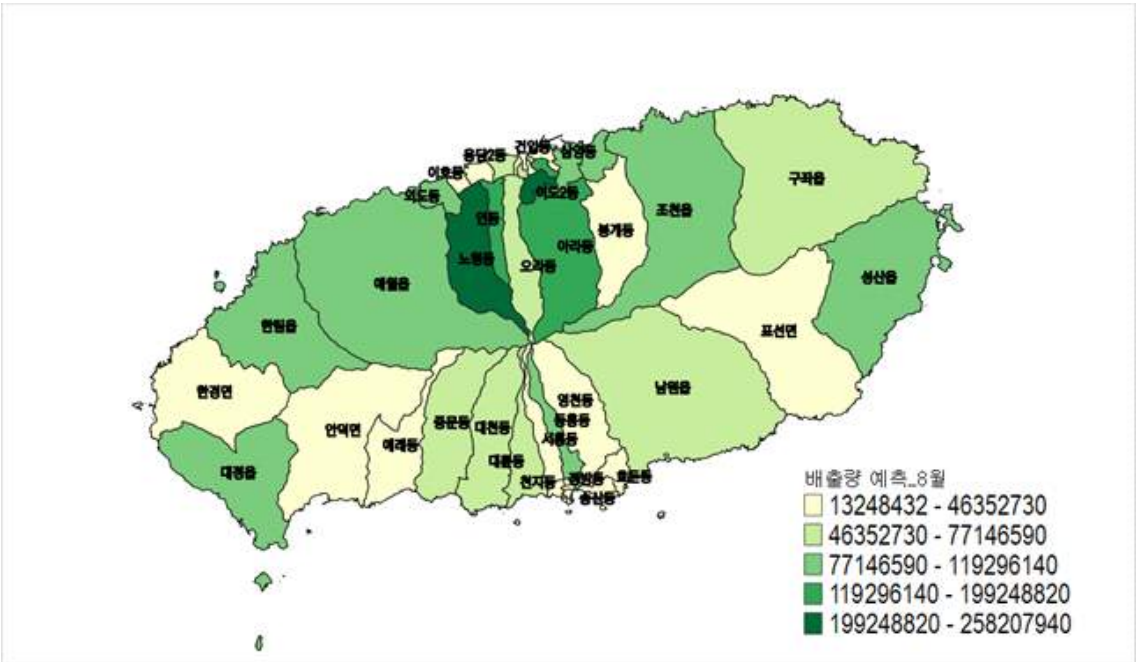
[그림 25] 최종 예측 결과

- Qgis를 이용한 예측값 시각화

Qgis 앱을 사용해 음식물 쓰레기 배출량을 5단계로 나누어 단계 구분 시각화 수행했다. 제주도 시가지 지역을 중심으로 배출량이 높은 것을 시각적으로 재확인했다.



[그림 26] 2021년 7월 배출량 시각화

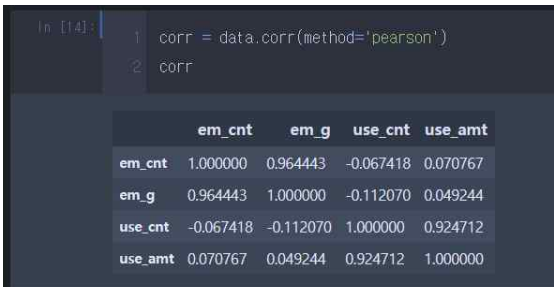


[그림 27] 2021년 8월 배출량 시각화

4.2 “알 수 없음” 데이터 음식물 쓰레기 배출량 예측

제공데이터 중 행정동이 “알 수 없음”인 데이터는 음식물 쓰레기 배출량 데이터와 카드소비량 데이터뿐 이었음

4.2.1 음식물 쓰레기 배출량과 카드결제금액, 카드결제건수 상관관계 파악



```
In [14]:
corr = data.corr(method='pearson')
corr
```

	em_cnt	em_g	use_cnt	use_amt
em_cnt	1.000000	0.964443	-0.067418	0.070767
em_g	0.964443	1.000000	-0.112070	0.049244
use_cnt	-0.067418	-0.112070	1.000000	0.924712
use_amt	0.070767	0.049244	0.924712	1.000000

[그림 28] pearson 상관계수 코드

- em_g와 use_cnt의 상관계수는 -0.11, em_g와 use_amt의 상관계수는 0.05
- 두 변수 모두 상관계수가 0에 가까우므로 연관이 없다고 판단

4.2.2 음식물 쓰레기 배출량과 카드결제금액, 카드결제건수 회귀분석

- em_g와 waste_cnt의 회귀분석 결과
- 결정계수(R^2)는 0.403, 수정결정계수(Adjust R^2)는 0.388로 waste_amt의 변동의 약 40%만이 card_cnt 변동에 의해 설명됨을 의미
- rain과 card_cnt 사이의 회귀식의 정확도는 매우 낮으므로 설명변수 card_cnt로서의 역할을 하기 어렵다고 판단

OLS Regression Results

Dep. Variable:	em_g	R-squared (uncentered):	0.403
Model:	OLS	Adj. R-squared (uncentered):	0.388
Method:	Least Squares	F-statistic:	27.66
Date:	Sat, 11 Sep 2021	Prob (F-statistic):	4.86e-06
Time:	19:55:13	Log-Likelihood:	-700.24
No. Observations:	42	AIC:	1402.
Df Residuals:	41	BIC:	1404.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
use_cnt	561.7619	106.821	5.259	0.000	346.033	777.490

Omnibus:	8.807	Durbin-Watson:	0.090
Prob(Omnibus):	0.012	Jarque-Bera (JB):	6.334
Skew:	0.815	Prob(JB):	0.0421
Kurtosis:	2.019	Cond. No.	1.00

[그림 29] 회귀분석 결과

- 음식물 쓰레기 배출량 - 카드결제금액 회귀분석 결과
- 결정계수(R^2)는 0.430, 수정결정계수(Adjust R^2)는 0.416로 waste_amt의 변동의 약 43.0%만이 card_amt 변동에 의해 설명됨을 의미
- rain과 card_amt 사이의 회귀식의 정확도는 매우 낮으므로 설명변수 card_amt로서의 역할을 하기 어렵다고 판단

OLS Regression Results

Dep. Variable:	em_g	R-squared (uncentered):	0.430
Model:	OLS	Adj. R-squared (uncentered):	0.416
Method:	Least Squares	F-statistic:	30.88
Date:	Sat, 11 Sep 2021	Prob (F-statistic):	1.84e-06
Time:	19:55:45	Log-Likelihood:	-699.28
No. Observations:	42	AIC:	1401.
Df Residuals:	41	BIC:	1402.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
use_amt	0.0170	0.003	5.557	0.000	0.011	0.023

Omnibus:	8.069	Durbin-Watson:	0.101
Prob(Omnibus):	0.018	Jarque-Bera (JB):	6.274
Skew:	0.828	Prob(JB):	0.0434
Kurtosis:	2.081	Cond. No.	1.00

[그림 30] 회귀분석 결과

- 따라서 “알 수 없음”으로 분류된 음식물 쓰레기 배출량 데이터에 대해 카드소비량으로 예측하기엔 부적절하다고 판단함

4.2.3 음식물 쓰레기 배출량 시계열분석(arima)

- 적당히 예측할 수 있는 변수가 부족해 “알 수 없음”으로 분류된 음식물 쓰레기 배출량 데이터만 가지고 ARIMA 모델을 예측함
- 전체 과정은 위와 같음

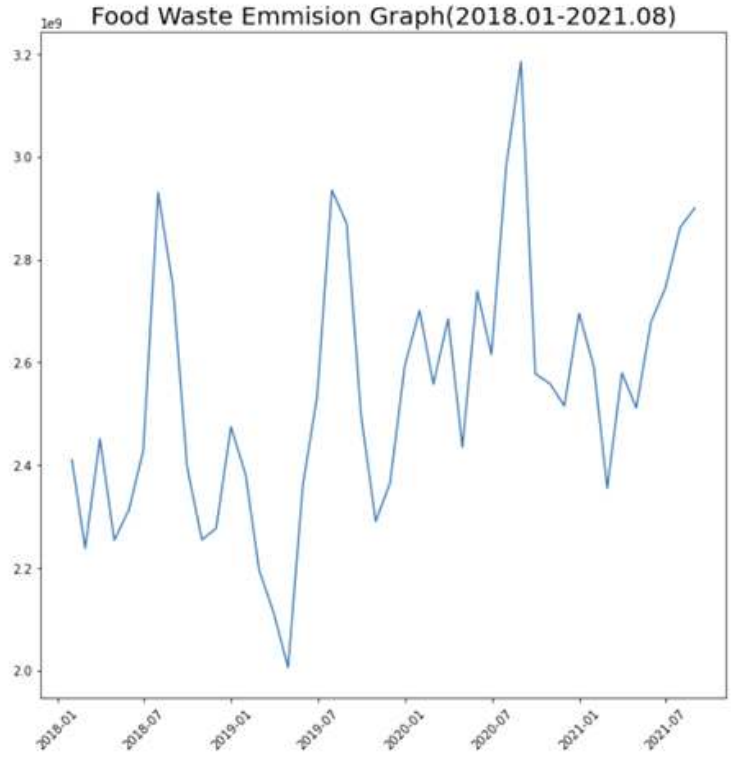
[그림 29] "알 수 없음“ 데이터 음식물 쓰레기 배출량 ARIMA 예측 결과

5 결론

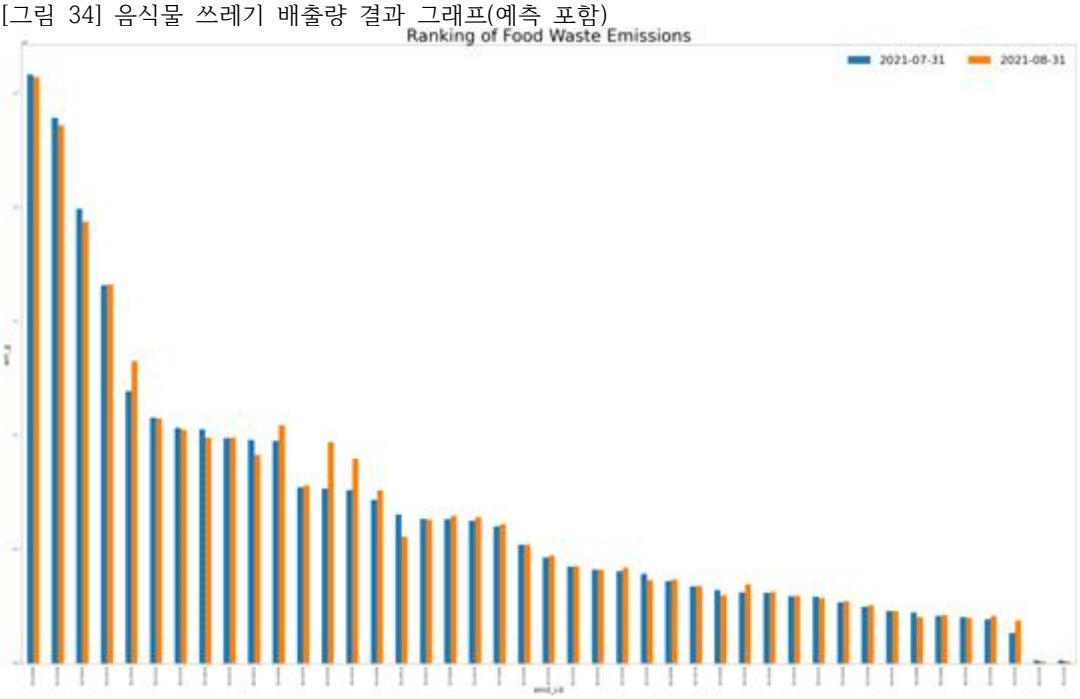
5.1 분석결과 활용 및 시사점

분석 결과 제주도 음식물 쓰레기 배출량의 주요 요인은 내국인 유동인구, 장기체류 외국인 유동인구, 단기체류 외국인 유동인구로 모두 유동인구와 관련이 있었다. 또, 2020년에 음식물 쓰레기 배출량의 감소를 보았을 때 코로나로 이동이 축소되어 유동인구가 감소했음을 알 수 있다.
특히 예측한 7월, 8월 음식물 쓰레기 배출량까지의 선그래프를 보았을 때 여름 관광 성수기에 값이 높아짐을 확인했다. 즉. 관광지의 음식물 쓰레기 배출량이 높다는 의미로 코로나 이후 이동이 자유로워지면 배출량이 증가할 것으로 보인다. 따라서 관광객을 대상으로 한 대책 마련이 필요하다. 예를 들어 관광지 주변 음식점의 음식물 쓰레기 배출량 감소를 위해 식당에서 음식을 남기지 않으면 인증 사진을 통한 이벤트를 진행하거나 음식을 남길 경우 환경부담금을 부과할 수 있다. 또, 관광지 곳곳에 음식물 쓰레기 배출에 경각심을 가질 수 있는 문구를 부착한다. 외국인 관광객을 위해 다양한 언어로 부착하는 방법도 고려해볼 수 있다.

읍면동별 배출량 순위를 막대그래프로 나타냈을 때 노형동, 이도2동, 연동 등 주택이 많은 시가지 지역이 가장 높음을 확인했다. 따라서 관광객뿐만 아니라 도민들의 노력도 필요함을 공익광고를 제작해 경각심을 심어줄 필요가 있다.



[그림 34] 음식물 쓰레기 배출량 결과 그래프(예측 포함)



[그림 35] 읍면동별 2021년 7, 8월 음식물 쓰레기 예측 결과

분석을 통해 앞으로 마주할 포스트코로나 시대에서 제주도의 음식물을 섭취와 배출 패턴의 변화를 파악할 수 있었다. 사회적, 환경적으로 다른 지역과는 차별화된 제주도

만의 음식물 쓰레기 배출량 감소 정책을 위해 관광지, 시가지 등 읍면동별 혹은 성수기와 비수기의 시기별로 나누어 정책을 만들 수 있다. 지속적인 데이터 분석을 통해 다양한 요인과 인과관계를 밝히려는 노력이 계속되어야 한다.