음식물 쓰레기 배출량의 영향 요인에 대한 연구.

기초 과학 데이터 실습 과제

조원 :

박지수 20171306

정찬휘 20194970

최광호 20143189

1. 서론

전 세계적으로 환경에 대한 우려가 커지고 있다. 기후 문제는 물론, 환경보호를 위해 일회용 쓰레기를 줄이려는 노력 또한 이루어지고 있는 가운데 많은 프랜차이즈에서 일회용 빨대까지 찾아보기 힘들게 되었다. 이런 상황에 우리 조원들은 한 학기동안 배운 데이터 분석기술을 이용해서 이런 큰 흐름에 어울리는, 유의미한 데이터 분석을 해보고자 이렇게 주제를 선정했다.

우선 음식물 쓰레기 배출량 데이터를 활용하여 다양하게 가공하여 분석해보았고, 회귀분석을 위해 음식물 쓰레기 배출량과 관련 있을 만한 요소들을 추측하여 관련 데이터들을 최대한 모으고, 적합한 데이터 분석 방법을 골라 유의미한 결론들을 도출하고, 그리고 시행됐던 정책들에 대해서 효용성 평가 또한 진행해 보았다. 결과적으로 음식물 쓰레기 배출량을 줄일 수 있는 방법에 대한 영감을 얻고자 했다. (현실적으로 생각해봤을 때, 기초 데이터 분석만으로는 감축을 위한 효과적인 방법 자체를 알아내긴 어렵다고 생각했다.)

다양한 요인을 고려하기 위해 다양한 데이터의 수집이 필요하다 보니, 데이터 전처리에 많은 공을 들였다. 각 정부 기관별로 공유 데이터의 질적 차이(서식, 단위, 공백 등)가 커서 적당히 선별, 처리하였다.

(방대한 데이터를 for 구문과 melt함수를 이용하여 효율적으로 처리해냈다.)

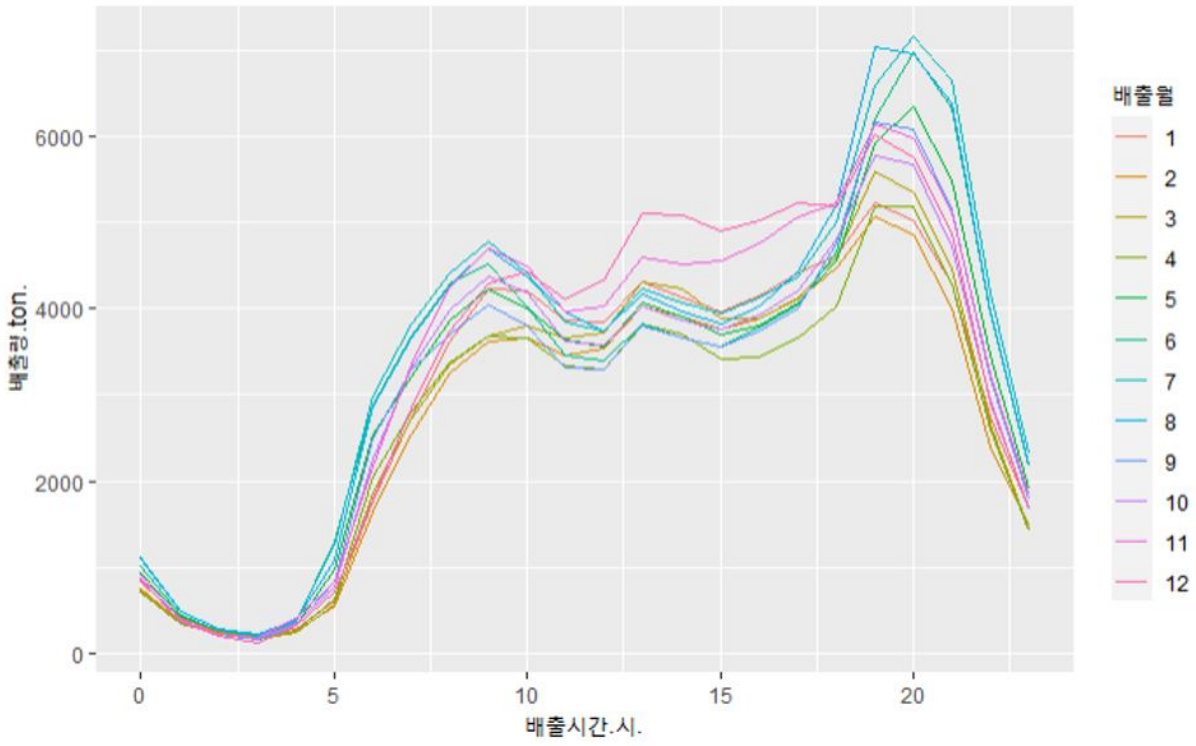
```
whole_05<-whole_m %>% filter(년도<= 2005);whole_05
for (i in 1:17) {
whole_local[i, '평균배출량<=2005'] <- mean((whole_05 %>% filter(시도==levels(factor(whole_05$시도,
levels=c("서울","부산","대구","인천","광주","대전","울산","세종","경기","강원","충북","충남","전북","전남","경북","경남","제주")))[i]))$배출량, na.rm=T);whole_local
}
whole_recent<-whole_m %>% filter(년도 > 2005)
for (i in 1:17) {
whole_local[i, '평균배출량>2005'] <- mean((whole_recent %>% filter(시도==levels(factor(whole_recent$시도,
levels=c("서울","부산","대구","인천","광주","대전","울산","세종","경기","강원","충북","충남","전북","전남","경북","경남","제주")))[i]))$배출량, na.rm=T);whole_local
}
```

```
whole <- read.csv('전국 배출량 데이터 96~19.csv')
whole_s <- whole[-1, ]

whole_s$X2006 <- ((whole_s$X2005 * 2) + (whole_s$X2008)) / 3 # 누락된 2006,07년 데이터는 2005년과 08년도의 3등분한 값으로 각각 대체하겠습니다.
whole_s$X2007 <- ((whole_s$X2005) + (whole_s$X2008 * 2)) / 3

whole_m<- melt(whole_s, id=c('시도별..톤.일','총인구수','세대수','세대당.인구', '지역내.총부가.가치.백만원.','면적.km.2.','생활인구.백만.'))

colnames(whole_m) <- c("시도",'총인구수','세대수','세대당.인구','지역내.총부가.가치.백만원.','면적.km.2.','생활인구.백만.','년도','배출량')

whole_m$년도 = gsub("[X]", "", whole_m$년도)
# dplyr의 rename으로는 변수명을 숫자로 변경이 불가하여, melt함수를 사용하여 년도 데이터를 행으로 보낸 뒤, gsub을 사용하여 값을 대체하였습니다.

whole_m$년도 <- as.numeric(whole_m$년도) # 년도 변수를 numeric 타입으로 변경
# write.csv(whole_m, file='전국 배출량 데이터 96~19(수정).csv')
```

2. 시간 규모 별 음식물 쓰레기 배출량 시각화

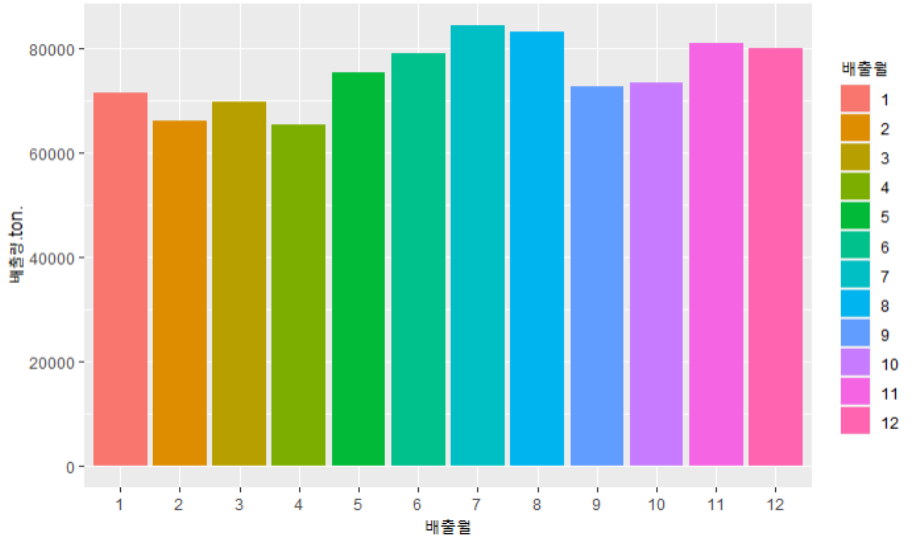
우선 지역 데이터를 다루기에 앞서, 최근 도입된 RFID기술을 활용해 수집된 『한국환경공단 RFID기반 20년도 전국 음식물쓰레기 배출정보』 데이터를 활용해서 일반 가정집의 시간대 별, 요일 별, 계절 별 음식물 쓰레기 배출량을 시각화 해보았다. '퇴근 후'와 '주말', '여름철'에 배출량이 많을 것으로 예상되었다.

2-1. 월 별 배출 시간대에 따른 배출량 추이



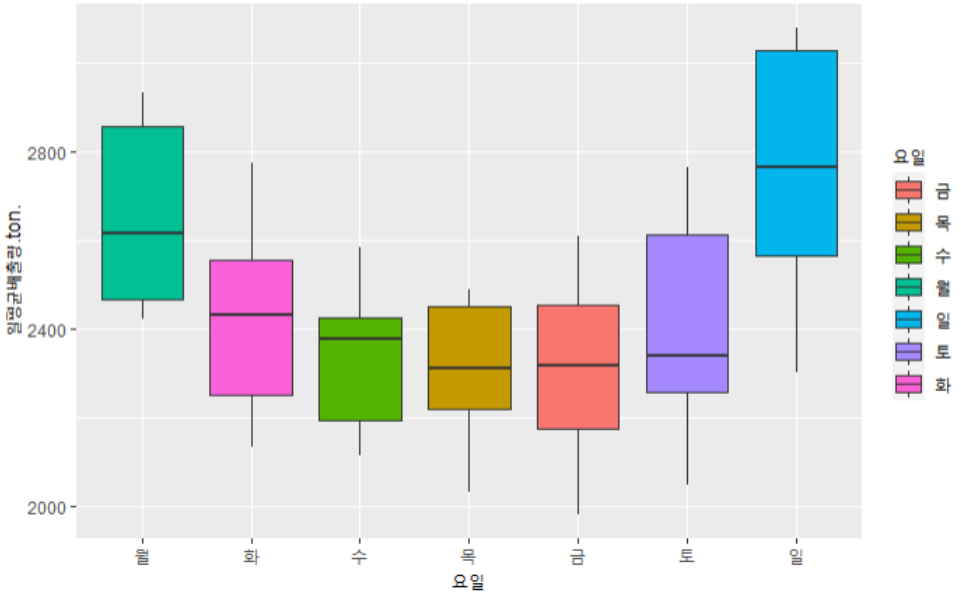
예상대로 퇴근 후 저녁 시간대에 배출량이 많았으며, 월별로 배출시간의 경향 차이는 특별히 없는 것으로 보인다. (배출시간에 계절은 영향이 없다.) 새벽 시간대의 쓰레기 수거는 현명한 판단으로 생각된다.

2-2 : 월별 일일 평균 배출량



월별로 배출량이 큰 차이를 보이지는 않지만 여름철에 배출량이 다소 높은 모습을 보여준다. 일반적으로 생각해봤을 때, 여름철에 음식이 부패되기 쉬워 낭비되는 음식들이 많아 다른 계절들의 경우에 비해 배출량이 크게 많을 것을 것으로 예상되었으나, 큰 차이를 보이지 않는 것으로 보아 계절적 요인에 관계없이 낭비되는 음식물이 많은 것으로 예상할 수 있다. 역시 음식물 쓰레기를 줄이려는 생활습관이 정착되어야 연중 고른 음식물 쓰레기 배출량을 효과적으로 줄일 수 있을 것으로 생각된다.

2-3 : 요일 별 평균 배출량



예상대로 주말(일요일)에 가장 많은 배출량을 기록했고, 다음으로는 월요일에 배출량이 많은 모습이다. Box Plot을 해석해보자면, 배출량이 비교적 많은 토, 일, 월이 데이터의 분포가 넓다. 개인마다 버리러 가는 습관의 차이를 예상해 볼 수 있다. (예를 들어, 주중동안 미뤄뒀던 일을 주말에, 주말 동안 미뤄뒀던 일을 월요일에 해결하는 것이라고 조심스레 예상해 볼 수 있다.)

3. 선형회귀 분석

3-1. 가설 설정

분석에 앞서 음식물쓰레기 배출량에 영향을 줄 만한 요소들을 선별하였다. 그 중에 **당연 '면적(㎢)' '단위면적당 인구 수(백만명/㎢)'** 가 관련이 큰 요인이라고 생각했다. 그 밖에 영향을 줄 만한 요소로 데이터를 구성하여 회귀 분석을 시행하였다.

서울 데이터와 전국 데이터를 따로 가공, 분석하여 변수로 사용할 유의미한 요인들을 얻어 보고 비교하였다.

3-2 서울 구 별 데이터 활용

3-2-1 선형 회귀 분석

먼저 서울의 25개 구, 17년치 음식물 쓰레기 배출량 데이터와 다른 요인과의 선형회귀 분석을 시행하였다. Drop1 함수를 이용하여 보다 정확한 분석을 진행하였다. (유의수준 5%)

* 1회차. (인구 수 변수 제거)

	면적(㎢)	세대 수	인구 수	세대당 인구	GRDP(지역내 총생산)	총 생활 인구수	단위 면적당 인구(백만명/㎢)
P-value	0.1646	0.8161	**0.9561**	0.5107	1.395e-08	0.1314	0.2535
F-value	2.1091	0.0558	**0.0031**	0.4514	101.4291	2.5121	1.3970

* 2회차. (세대 수 변수 제거)

	면적 (㎢)	세대 수	세대당 인구	GRDP(지역내 총생산)	총 생활 인구수	단위 면적당 인구(백만명/㎢)
P-value	0.14560	**0.61310**	0.16560	4.272e-10	0.09635	0.22927
F-value	2.3138	**0.2648**	2.0885	147.0086	3.0784	1.5488

* 3회차. (세대당 인구 수 요소 제거)

	면적(㎢)	**세대당 인구**	GRDP(지역내 총생산)	총 생활 인구수	단위 면적당 인구(백만명/㎢)
P-value	8.499e-09	**0.133550**	1.265e-10	0.002859	2.041e-05
F-value	94.1970	**2.4564**	156.6985	11.7096	31.5557

* 4회차. (완료)

	면적 (㎢)	GRDP (지역내 총생산)	총 생활 인구수 (백만명)	단위 면적당 인구 (백만명/㎢)
P-value	4.757e-10	1.225e-10	0.006993	1.893e-06
F-value	124.8150	145.7029	9.0314	43.8573

* Summary (lm4)

	Intercept	면적(㎢)	GRDP(지역내 총생산)(백만원)	총 생활 인구수(백만명)	단위 면적당 인구 (백만명/㎢)
P-value	1.75e-05	4.76e-10	1.23e-10	0.00699	1.89e-06
t-value	11.172	12.071	-3.005	-3.005	6.622
coefficients	-1.121e+02	4.596	2.701	-5.037	5.187e+03

(Multiple R-squared: 0.9407, Adjusted R-squared: 0.9288)

음식물 쓰레기 배출량(톤) = (-1.121) 10^2 +(4.596)면적(㎢)+(2.701)GRDP(백만원)-(5.037)총 생활 인구수(백만명)+5.187 10^3 단위 면적당 인구수(백만명/㎢)

* 여기서 '총 생활 인구 수' 만이 음의 계수를 가지고 있는데, 이 변수만으로 다시 회귀분석을 하면

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.189e+02  1.182e+01  10.057 6.86e-10 ***
총생활인구수 6.429e-06  5.195e-06   1.237    0.228
```

Cor : 0.2498358 , Multiple R-squared: 0.06242, Adjusted R-squared: 0.02165

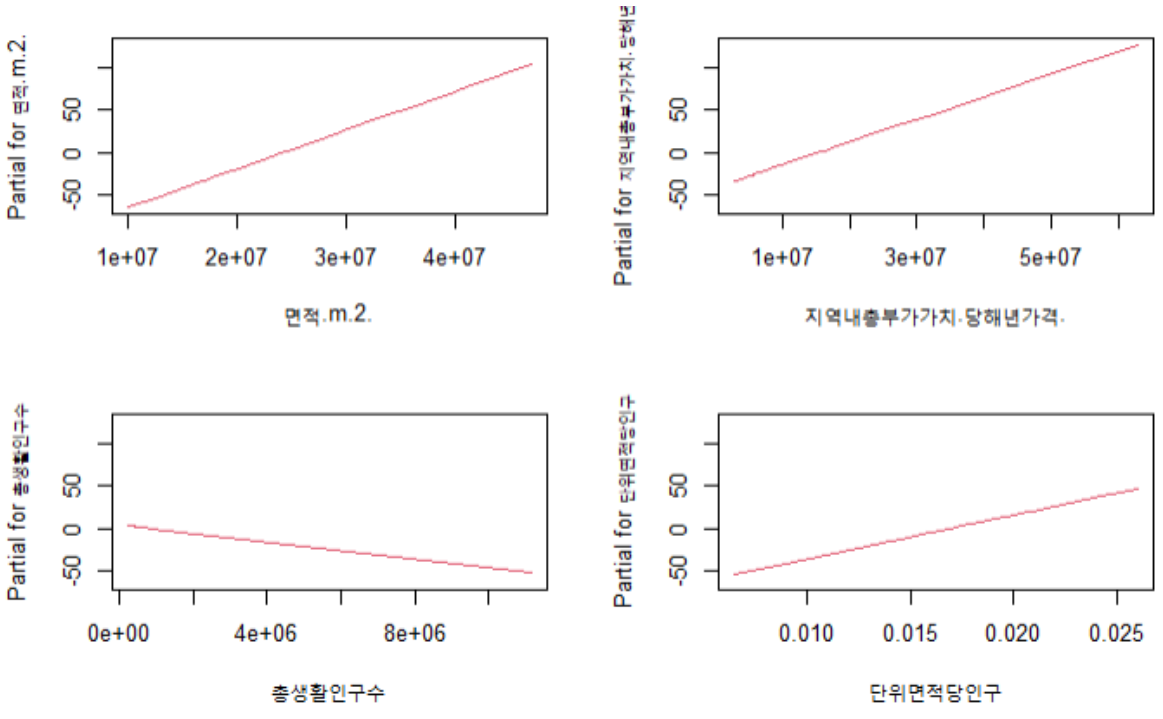
'총 생활 인구 수' 변수가 양수 계수를 가지는 것으로 보이나 두 요소가 큰 상관관계를 가지고 있지 않을 뿐더러, 회귀식이 가지는 설명력도 크지 않은 것으로 보인다.

따라서 처음 소개한 회귀식의 음의 계수는 F-value 가 큰 다른 요소들을 보정하기 위해 나타난 것으로 보인다.

* 처음 예상이 적중함과 동시에, 설명력과 value로 보아 신뢰할 만한 회귀식이 도출되었다.

* 이렇게 4가지 변수들이 음식물 쓰레기 배출량에 영향력을 가짐을 확인할 수 있다. 허나 이 4가지 변수들은 어찌 보면 데이터 분석을 해보지 않아도 알 수 있는 너무 당연한 요인들이다. 하지만 여기서 시사하는 바는 단순 인원수를 의미하는 '총 인구 수' '세대 수' '세대 당 인구 수'는 배제되었다는 것이다. (남은 '생활인구 수'도 큰 비중을 차지하지는 못한다.) 또한 'GRDP'와 '면적 당 인구 수' 가 둘 다 지역적 특수성에 연관되어 있다는 사실 또한 고려한다면, 음식물 쓰레기 배출량에는 단순한 '인구 수'보다는 '지역의 특수성'의 영향이 더 크다는 것을 알 수 있다.

* termplot



각 항목별로 termplot을 만들어서 회귀 모형의 각 변수들의 예측 값에 대응하는 회귀선을 구해보았다.

영향이 큰 요소들일수록 예상되는 회귀선의 기울기 값도 큰 것으로 나타났다.

3-2-2. 정책 효용성 평가

서울시는 2013년 6월부터 쓰레기 종량제 의무화 정책을 시행하였다. 그래서 정책의 효용성을 평가해보고자 정책이 시행된 2013년 이전의 데이터와 그 후의 데이터로, 두 집단으로 나누어 음식물 쓰레기 배출량에 유의미한 변화가 있었는지 t-test를 시행해 보았다. (25개의 구 각각으로 시행, 비교하였다.)

① 2009~2012 vs 2014~2019 :

t = 0.50999, p-value = 0.3063

```
        Welch Two Sample t-test

data:  seoul_gu_sub$`2009<평균배출량<2013` and seoul_gu_sub$`2013<평균배출량<2017`
t = 0.50999, df = 45.811, p-value = 0.3063
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -18.93071       Inf
sample estimates:
mean of x mean of y
  133.884   125.624
```

충분히 작지 않은 p-value 가 산출되어 유의미한 차이가 있었다고 말하기 힘들다. 따라서 정책의 효용성이 있었다고 주장하기 어렵다.

② 2012 vs 2014 :

t = 0.32048, p-value = 0.75

```
        Welch Two Sample t-test

data:  seoul_gu_sub$X2012년 and seoul_gu_sub$X2014년
t = 0.32048, df = 47.643, p-value = 0.75
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -27.55722  38.00522
sample estimates:
mean of x mean of y
  132.468   127.244
```

시행되기 바로 이전해와 바로 다음 해만 비교해도 역시 유의미한 차이가 있었다고 말하기 힘들다.

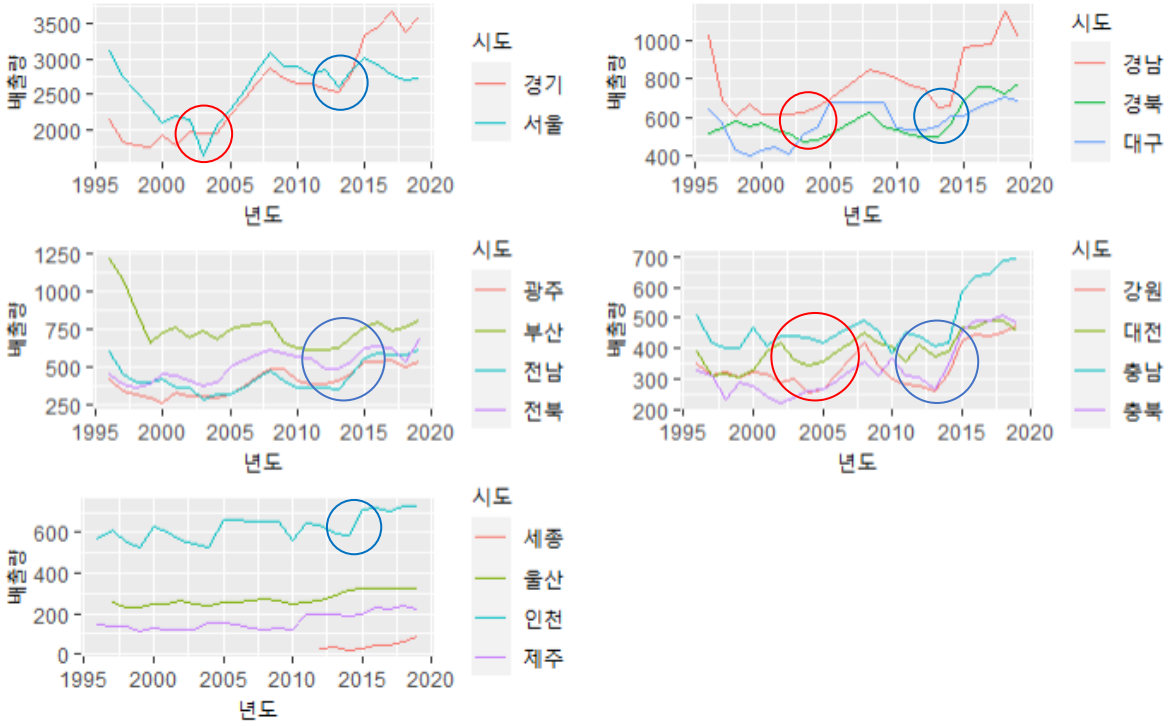
따라서 정책의 효용성을 주장하기는 힘들다고 할 수 있겠다.

또한 후술할 지역별 그래프의 서울의 경우에 이후 오히려 반등하는 모습을 확인할 수 있다.

3-3 전국 행정구역 별 22년치 데이터 활용

전국의 행정구역 17구역의 22년치(96~19년, 중간 2년 누락) 음식물 쓰레기 배출량 데이터를 활용하여 비슷한 회귀 분석을 진행해 보려고 한다.

3-3-1 전체 (17개 행정구역) 추이 그래프



(행정구역 신설로 인해 울산시 97년도, 세종시 12년도부터 데이터가 있다.)

(배출량 단위 톤ton)

96년 데이터 집계가 이루어진 이래 전체적으로 꾸준히 감소하는 경향을 보이다가 2003~4년경, 그리고 2013년경 반등하기 시작하는 모습을 보인다.

3-3-2 전국 데이터 활용 : 회귀 분석

1회차. (세대당 인구 요소 제거)

	인구 수 (백만명)	세대 수	세대당 인구	GRDP(지역내 총생산)	면적 (㎢)	총 생활 인구수(백만명)	단위 면적당 인구(㎢/백만명)
P-value	0.6296	0.4573	**0.9072**	0.7244	0.8725	0.6289	0.1860
F-value	0.2492	0.6032	**0.0144**	0.1323	0.0273	0.2502	2.0497

2회차. (면적(㎢) 요소 제거)

	인구 수 (백만명)	세대 수	GRDP (지역내 생산)	**면적 (㎢)**	총 생활 인구수(백만명)	단위 면적당 인구(㎢/백만명)
P-value	0.4369	0.2649	0.7251	**0.8589**	0.6200	0.1305
F-value	0.6558	1.3950	0.1308	**0.0333**	0.2619	2.7141

3회차. (GRDP 요소 제거)

	인구 수 (백만명)	세대 수	**GRDP(지역내 총생산)**	총 생활 인구수(백만명)	단위 면적당 인구(㎢/백만명)
P-value	0.1605	0.0607	**0.6922**	0.5813	0.0009
F-value	2.2646	4.3624	**0.1651**	0.3228	20.1291

4회차 (총 생활 인구수 요소 제거)

	인구 수 (백만명)	세대 수	**총 생활 인구수(백만명)**	단위면적당 인구(㎢/백만명)
P-value	0.1392	0.0483	**0.4818**	0.0005
F-value	2.5075	4.8315	**0.5268**	21.6056

5회차(완료)

	인구 수 (백만명)	세대 수	단위면적당 인구(㎢/백만명)
P-value	0.0388	0.0014	7.441e^-5
F-value	5.2748	16.2933	32.3528

Summary(lm5)

	Intercept	인구 수 (백만명)	세대 수 (만 세대)	단위 면적당 인구 (㎢/백만명)
P-value	0.86091	0.03890	0.00141	7.44e-05
F-value	-0.179	-2.297	4.036	5.688
coefficients	-3.6249	-242.4	9.918	0.0321

음식물 쓰레기 배출량(톤) = -3.6249 -(242.4)인구수(백만명)+(9.918)세대 수+(0.0321)단위 면적당 인구수(백만명/㎢)

서울시 데이터와 상당히 다른 결과가 나왔다. 전국적으로 보았을 때, 다른 모든 요소들을 제치고 인구의 영향이 더 컸다. 지역적 특수성의 차이는 전국적으로 더 크겠지만 오히려 영향을 크게 미치지 못했다고 판단할 수 있다. (뒤에서 더 자세히 정리하겠다.)

3-3-3 : 정책 시행 검증

2005년 전국적으로 음식물 쓰레기 직매립 금지 정책으로 분리배출이 본격적으로 시행되었다. 따라서 당해 기준 전과 후로 데이터를 나누어 음식물 쓰레기 배출량에 유의미한 변화가 있었는지 t-test를 시행하였다.

1. 평균배출량 비교 :

t-value = -0.4877 , p-value = 0.6853

```
        Welch Two Sample t-test

data:  whole_local$`평균배출량<=2005` and whole_local$`평균배출량>2005`
t = -0.4877, df = 29.402, p-value = 0.6853
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -549.2641       Inf
sample estimates:
mean of x mean of y
 639.7465  762.2847
```

: 높은 p-value를 보인다. 앞선 서울시 정책과 마찬가지로 유의미한 차이가 있었다고 주장하기 힘들다.

2. 인구당 평균 배출량 비교 :

t-value = -2.2868 , p-value = 0.01622 (alternative = less)

```
        Welch Two Sample t-test

data:  whole_local$`단위인구당 배출량<=2005` and whole_local$`단위인구당 배출량>2005`
t = -2.2868, df = 21.493, p-value = 0.01622
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
          -Inf -7.245061e-06
sample estimates:
   mean of x    mean of y
0.0002212998 0.0002504758
```

: 낮은 p-value를 보여 유의미한 차이가 있었다고 주장하기 힘들다.

4. 결론

각 파트마다 간략히 결론을 내려왔고, 이제 종합적인 결론을 내리겠다.

① 우선 시간대별 음식물 쓰레기 배출 데이터를 분석해본 결과, 음식물 처리 시스템의 효율을 증진할 수 있는 인사이트를 얻었다.

② 회귀분석을 해 본 결과, 서울과 전국의 데이터에 영향을 미치는 변수가 서로 달랐다. 서울의 경우 지역의 특수성을 의미하는 변수가, 반대로 전국은 지역적 특수성보다는 보편적 요인이 변수로 채택되었다. 좀 더 풀어서 해석해보면, 서울의 경우 각 구별로 특수성이 보편성보다 강하게 작용하였고, 전국의 경우는 '인구'가 가장 영향이 큰 변수로 채택되었기 때문에, 지자체별 특수성보다는 보편성(이 경우 전국범위의 보편성은 국민성이라 표현할 수 있을 텐데)에 더 영향을 받는 것으로 보인다.

따라서 추측건대, 서울의 경우 전국지자체 중에 구별로 차이(예를 들어 번화한 정도)가 비교적 큰 것으로 보이며, 전국적 보편성을 고려하여 **일반적인 우리의 음식물 낭비 행태 등을 개선해 나가야 한다고 생각해 볼 수 있겠다. (이는 절대 국민성 비하가 아니다 !!)**

③ 서울의 데이터에서 '생활인구' 요인 또한 채택되었는데 앞서 말한 '번화한 정도' 등의 차이에서 배출량의 차이가 발생하여 변수로 채택된 것으로 보인다. 예를 들어 번화가 중심인 강남구가 주로 주거지구인 강동구의 배 이상의 음식물 쓰레기를 배출한다.

5. 추후 연구

결론에서 이어지는 의문이 들었다. 낭비행태가 정말 큰 요인인지를 논리적으로 보일 수 있으며, 그 낭비행태를 개선할 수 있는 방법을 제시할 수 있을까. 관련 연구를 찾아본 결과, 흥미로운 논문을 발견할 수 있었는데, 설문응답을 통해 얻은 데이터로 분석을 하여 '개인적 상대 이익' 이나 '내적 규범적 신념' '자기 성취' 같은 추상적 요인들과 '쓰레기 감량태도, 감량행태' 등과의 연관성을 예측하였다. (연령이나 성별, 주거 형태 등의 요소들도 고려하셨었다. 우리 팀도 해보려고 했으나 윤리적인 문제가 있다고 판단하여 배제하였다.) 때문에 데이터만 확보된다면 추상적 요소 또한 데이터에 담아 연구해보고 싶다.

그리고 연도별 반등구간에서의 원인이나 인접국 일본이 음식물 쓰레기 배출량이 감소 추세임을 고려하여 일본의 음식물 쓰레기 감소 정책의 실효성을 검증해 보면 좋은 연구가 될 것이다.

6. 참고

@ 환경부 환경통계 포털 http://stat.me.go.kr

@ 통계청 공공데이터제공 https://kostat.go.kr

@ 서울 열린 데이터 광장 https://data.seoul.go.k

@ 공공데이터포털 https://www.data.go.kr

@ 위키페디아 각 시도별

@ 환경부 정책 https://me.go.kr/home/web/index.do?menuId=10260

@ 논문 : 쓰레기 감량행태의 영향요인 분석 : 계획행태이론의 적용을 중심으로 (김은희 저)

@ KOSIS 국가 통계 포털 https://kosis.kr