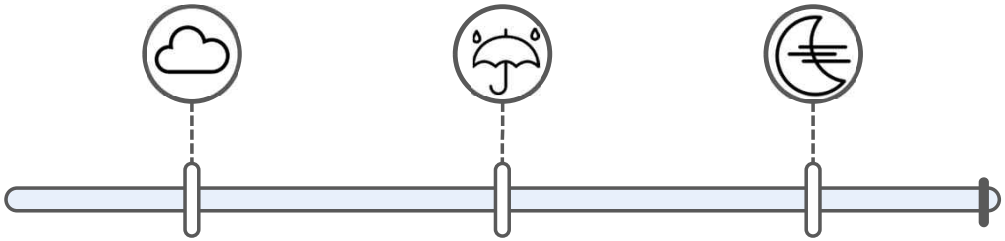
날씨에 따른 온라인 구매 예측

VARX 모델에 기반한 분석 모델 제안

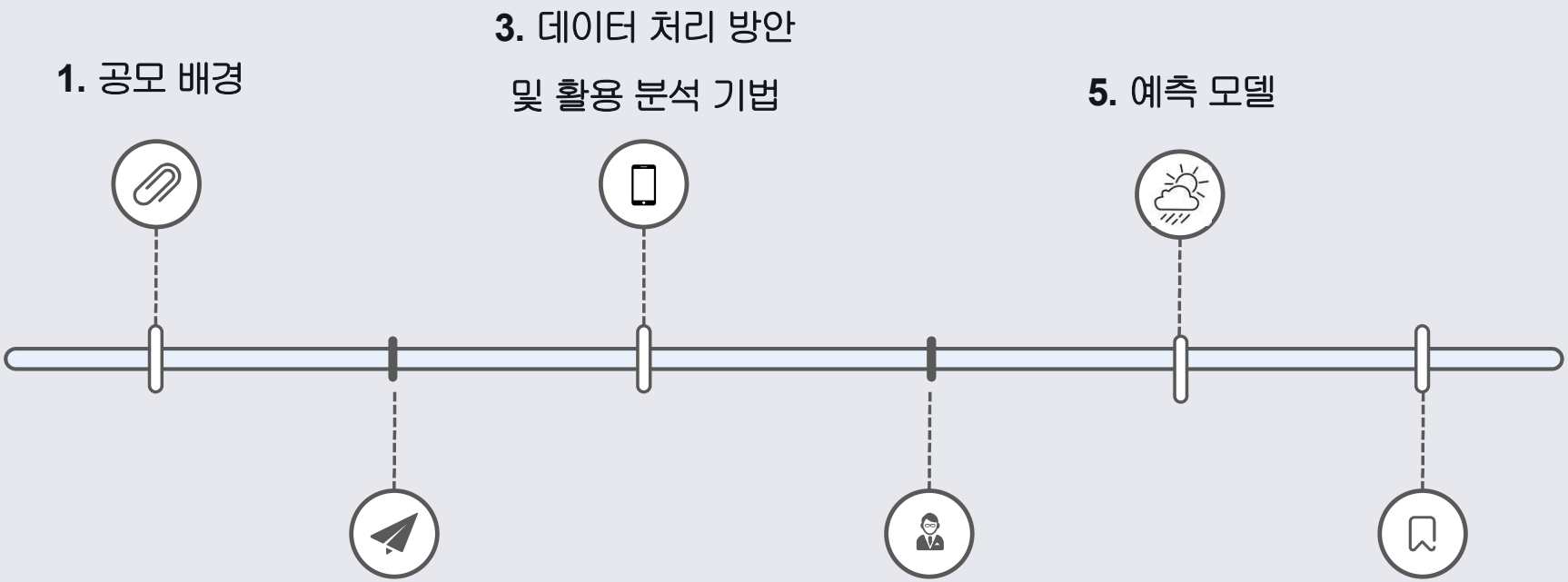
2021 날씨 빅데이터 콘테스트



김예지 김희진 안다영 이미경 이인선 하성연

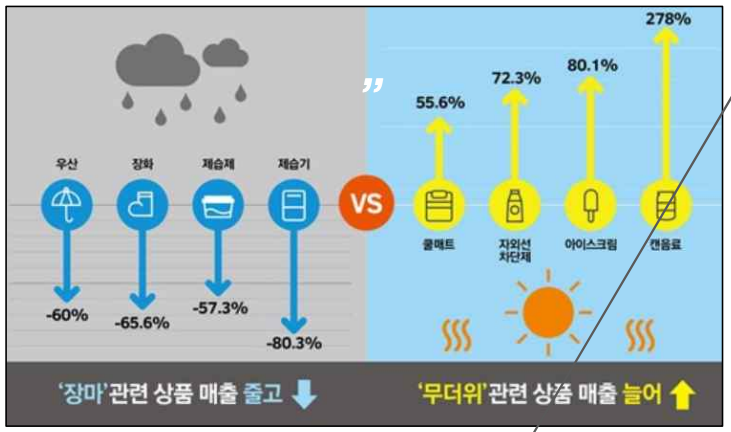
unnormal

contents


1. 공모 배경
2. 활용 데이터 정의
3. 데이터 처리 방안
및 활용 분석 기법
4. 소비 트랜드 분석
5. 예측 모델
6. 결과 활용방안
및 제안

1 공모배경
1.1 날씨와 소비 패턴

"늑장 장마, 소비 패턴 바꿨다"



(출처 :MNB 뉴스)

티몬에 따르면, 지난해 6월 초 히트상품은 제습기와 우산, 장화 등 **장마**관련 상품이 많았다. 기상청이 6월말 **장마가 올 것으로 예상해 미리 대비**하는 고객들이 많았기 때문. 유통업계에서도 장마 관련 행사를 잇따라 열며 **관련 상품을 집중적으로 홍보**했다.

"날씨가 가장 효과적인 판매 전략 : 날씨 마케팅"

기상·기후

↓

소비패턴 파악

↓

날씨 마케팅 →

"날씨가 덥지 않으면
20만원을 돌려드립니다."
<삼성 에어컨>

"비 내리는 날~
반디 네일에는
혜택이 내린다!"
<BANDI>

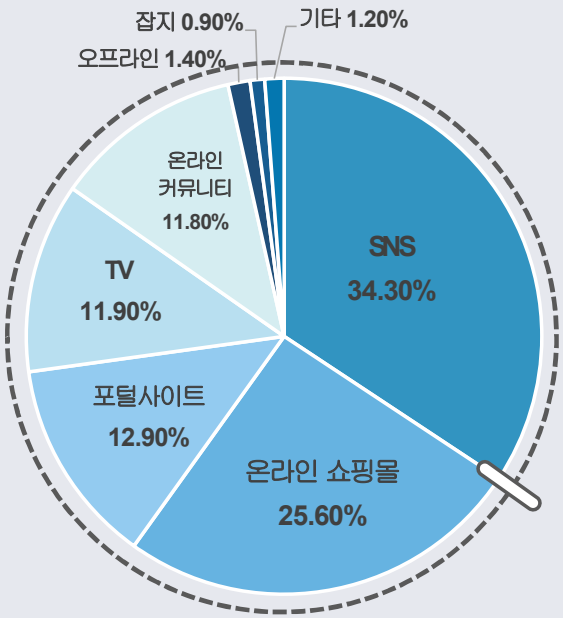
"맑은 날, 빙수구매시
천원 추가하시면
아이스크림 **50g** 더!
<나뚜루팜>

→ 마케팅 효과 증대 → 경쟁력 제고

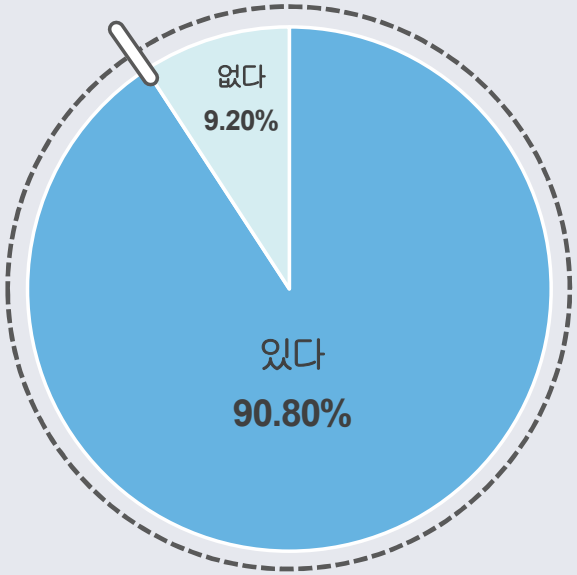
"기상기후정보를 활용한 날씨 마케팅을 통해 경쟁력 확보"

1. 공모배경
1.1 날씨와 소비 패턴

Q1. 평소 상품정보를 접하는 채널은 어디인가요?



Q2. SNS에서 상품을 접하고 구매한 경험이 있나요?



"SNS로 상품 접하고 구매"

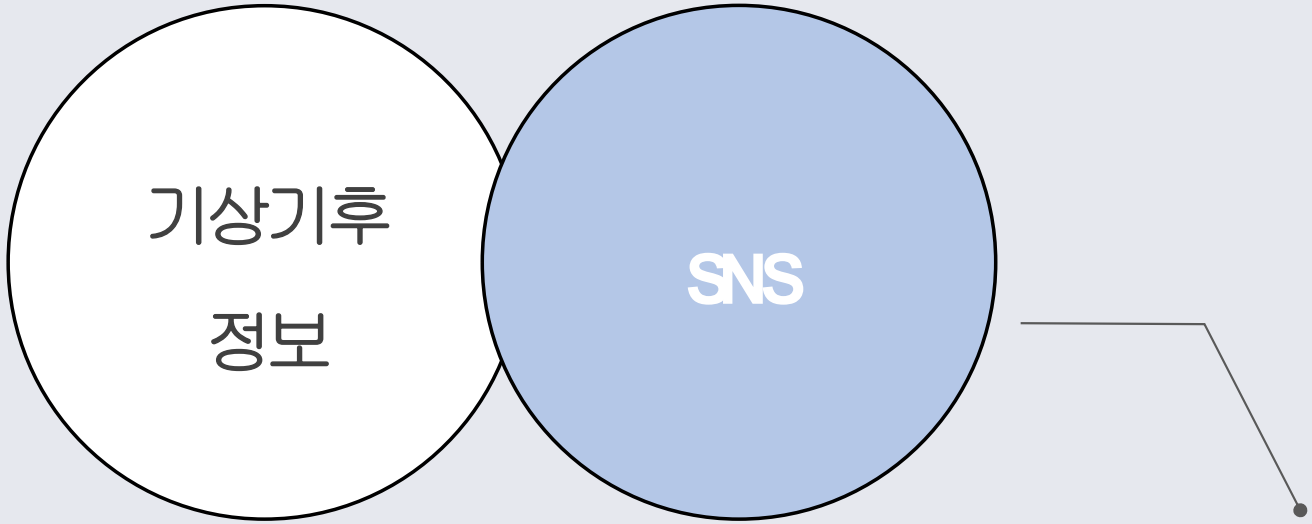
SNS 채널 → 상품정보 파악 → 구매에 영향

"SNS 채널을 활용한 마케팅을 통해 구매 촉진 "

(출처 : 디지털투데이)

1.2 공모배경 분석목표

"기상기후정보를 활용한 소비패턴 파악"



" SNS 채널을 통한 구매 촉진 "

"기상기후정보와 **SNS**채널을 활용한 소비패턴 예측의 정확성 제고 및 마케팅 효과의 극대화"

2 활용데이터 정의

출처	기상청 날씨마루	㈜엠코퍼레이션	㈜바이브컴퍼니
데이터 내역	기온, 강수량, 상대습도, 풍속, 미세먼지	온라인 구매이력	소셜 데이터
기간	2018.01.01 – 2019.12.31	2018.01.01 – 2019.12.31	2018.01.01 – 2019.12.31

변수

STN_ID	지점번호 (공통)
TMA	관측시각
AVG_TA	평균 기온
MAX_TA	최고 기온
MIN_TA	최저 기온
SUM_RN	합계 강수량
AVG_RHM	평균 상대습도
PM10_PM	PM10 미세먼지 농도
PM25_PM	PM25 미세먼지 농도 (초미세먼지)

date	날짜
sex	성별
age	나이
big_cat	상품 대분류명
sm_cat	상품 소분류명
qty	구매 수량

date	날짜
big_cat	상품 대분류명
sm_cat	상품 소분류명
cnt	10만 건 당 건수

3. 데이터 처리 방안 및 활용 분석 기법
3.1 결측값 처리

데이터 정제

날씨변수 결측값

결측치 총 7개
(비율 0.00029%)

변수명	변수설명
Avg_ta	평균 기온
Max_ta	최고 기온
Min_ta	최저 기온
⋮	
Avg_ws	평균 풍속
Max_ws	최대 풍속
Max_ins_ws	최대 순간 풍속
Sum_rn	합계 강수량
Avg_rhm	평균 상대습도
Min_rhm	최소 상대습도
Pm10	미세먼지
Pm25	초미세먼지

결측값 처리 방법

“ MICE ”

(Multivariate Imputation by Chained Equations)

- **MCAR(Missing Completely at random)**값을 다중대체
- 결측치의 비율이 매우 작아 완전 제거법으로 처리할 수 있지만 날짜별 기상데이터를 모두 활용해 분석하기 위해 대체 선택

1. 다른 모든 변수를 사용해 결측치 예측
2. 결측치가 채워진 완성된 데이터셋 여러 개 생성
3. 각각의 완성된 데이터셋에 통계 모형 적용
4. 각각의 분석 결과를 하나로 통합
5. 원 데이터의 결측치 대체

3. 데이터 처리 방안 및 활용 분석 기법
3.1 결측값 처리

데이터 정제

온라인 구매건수 결측값

날짜	성별	연령대	대분류	소분류	구매량
2019-09-19	F	20	뷰티	린스	7
2019-09-19	M	20	뷰티	린스	3
2019-09-19	F	30	뷰티	린스	27
2019-09-19	F	40	뷰티	린스	32
2019-09-20	F	20	뷰티	린스	9

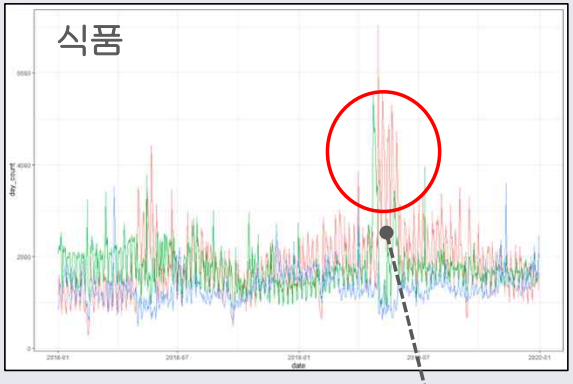
날짜	성별	연령대	대분류	소분류	구매량
2019-09-19	F	20	뷰티	린스	7
2019-09-19	M	20	뷰티	린스	3
2019-09-19	F	30	뷰티	린스	27
2019-09-19	M	30	뷰티	린스	0
2019-09-19	F	40	뷰티	린스	32
2019-09-19	M	40	뷰티	린스	0
2019-09-19	F	50	뷰티	린스	0
2019-09-19	M	50	뷰티	린스	0
2019-09-19	F	60	뷰티	린스	0
2019-09-19	M	60	뷰티	린스	0
2019-09-20	F	20	뷰티	린스	9

1. 구매건수가 없는 날짜 존재.

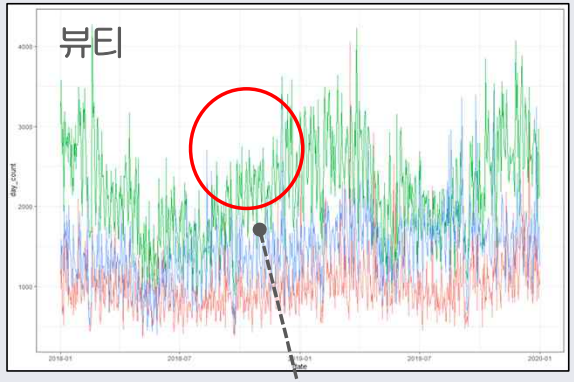
→ 결측값이 존재하는 제품 **[** 식품 : **33**개, 뷰티 : **27**개, 냉난방가전 : **34**개**]**

2. 성별, 연령대별 구매건수에 대한 결측값을 **0**으로 대체

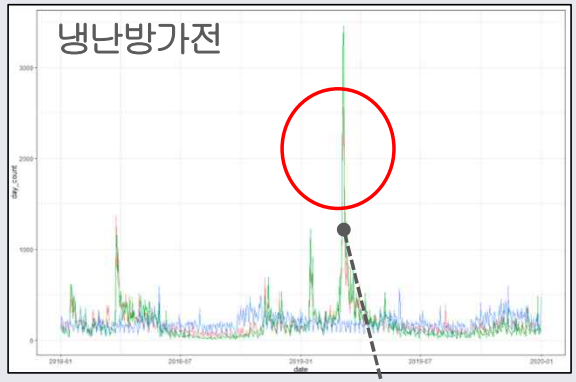
상품 대분류별 구매량 TOP3



생수 → 여름 구매↑



기초화장용 크림 → 가을, 겨울 구매↑



공기정화 용품
공기청정기
→ 봄 구매↑

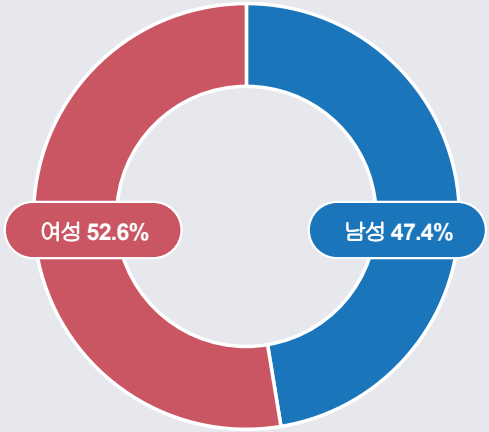
세가지 대분류 그래프 모두 계절성을 보임.

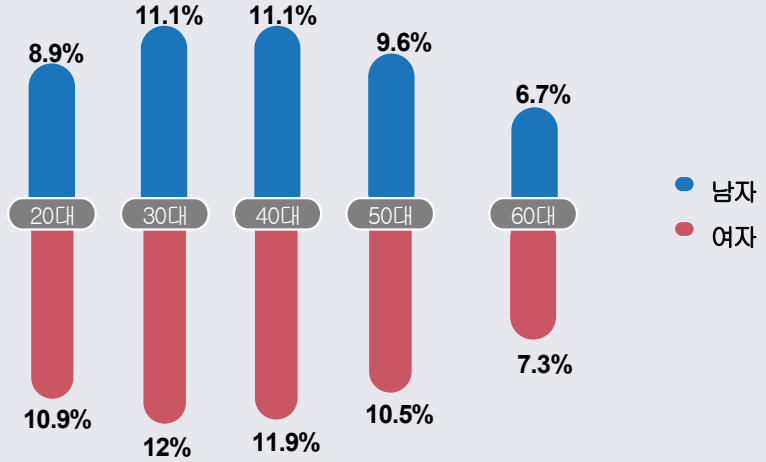
↓

제품 구매가 날씨에 영향을 받을 것이라고 예측할 수 있음.

성별 / 연령대별 구매건수의 차이 검정

T-test & ANOVA





성별, 연령에 따른 대분류별 구매건수 평균 차이 유의

구분		여성 (n = 1,081,977)	남성 (n = 974,922)	p-value
전체		25.363	12.156	**<.0001**
대분류	식품	24.525	13.378	**<.0001**
	뷰티	29.517	10.773	**<.0001**
	가전	9.726	8.152	**<.0001**

구분		20대 (n = 406,097)	30대 (n = 476,260)	40대 (n = 473,038)	50대 (n = 414,271)	60대 (n = 287,233)	p-value
전체		18.009	28.856	24.662	12.574	4.742	**<.0001**
대분류	식품	16.445	29.473	25.768	13.457	5.137	**<.0001**
	뷰티	22.205	30.956	25.277	12.010	4.179	**<.0001**
	가전	5.607	12.688	11.949	7.078	3.283	**<.0001**

연령대 별 구매 카테고리

- **20**대 여성 구매 카테고리

- 60대 남성 구매 카테고리

3. 데이터 처리 방안 **및 활용 분석 기법**
3.3 온라인 구매이력 – 재분류

카테고리 재분류

- 날씨 변수에 따른 수요예측을 위해 카테고리의 명확한 기준을 재설정.

- 온라인 구매 내역 및 소셜 데이터의 대분류(big_ cat) 기준에 따라 **카테고리 기준을 재설정**.
  (large_ cat 이라는 **대분류** 변수 추가)

- **식품 데이터**는 **중분류** (mid_ cat) 변수를 추가하여 보다 세분화.

Big_cat	Large_cat	Mid_cat	Sm_cat
식품	건강기능식품	건강기능식품	홍삼 분말/ 환
	음료류	탄산음료류	탄산음료
	견과 종실류	견과 종실류	헛개/ 가시오가피
합계	**27개**	**61개**	**212개**
뷰티	페이셜케어		기초 화장용 미스트
	바디케어		바디 보습제
	선케어		선파우더
합계	**15개**		**131개**
냉난방 가전	난방기기		온수매트
	에어컨		스탠드형 에어컨
	건조기		의류건조기
합계	**9개**		**40개**

(참고 : 식품 공전, 랄라블라, LG생활전자)

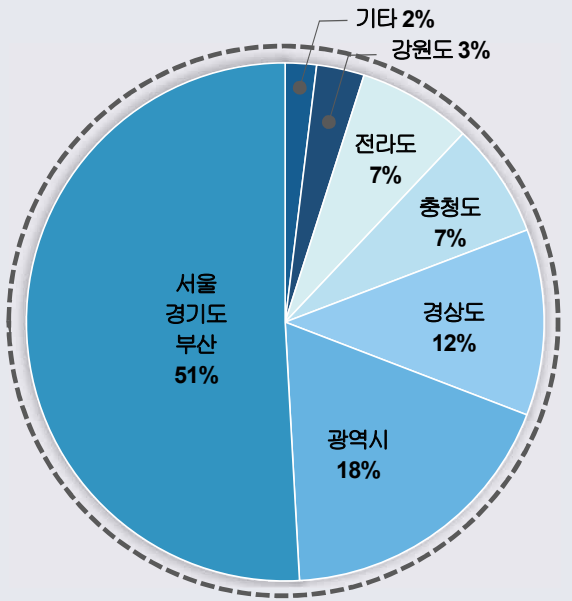
날씨지수 생성 개요

온라인 구매내역 소셜 데이터	날씨 데이터
지역 정보가 없음.	지역 정보가 있음.
두 데이터를 함께 활용하기 위해서는 지역에 국한 되지 않는 **"일별 대표 기상 데이터"** 가 필요.	

PCA ↓ 차원 축소

date	ta_index	rn_index	wind_index	pm_index
날짜	기온 지수	강수량 지수	풍속 지수	미세먼지 지수
2018-01-01	-4.22459	-2.88059	-0.51503	0.051328
2018-01-02	-4.11965	-2.69721	0.060365	-1.16792
2018-01-03	-4.65496	-2.84326	0.029904	0.537802
⋮				

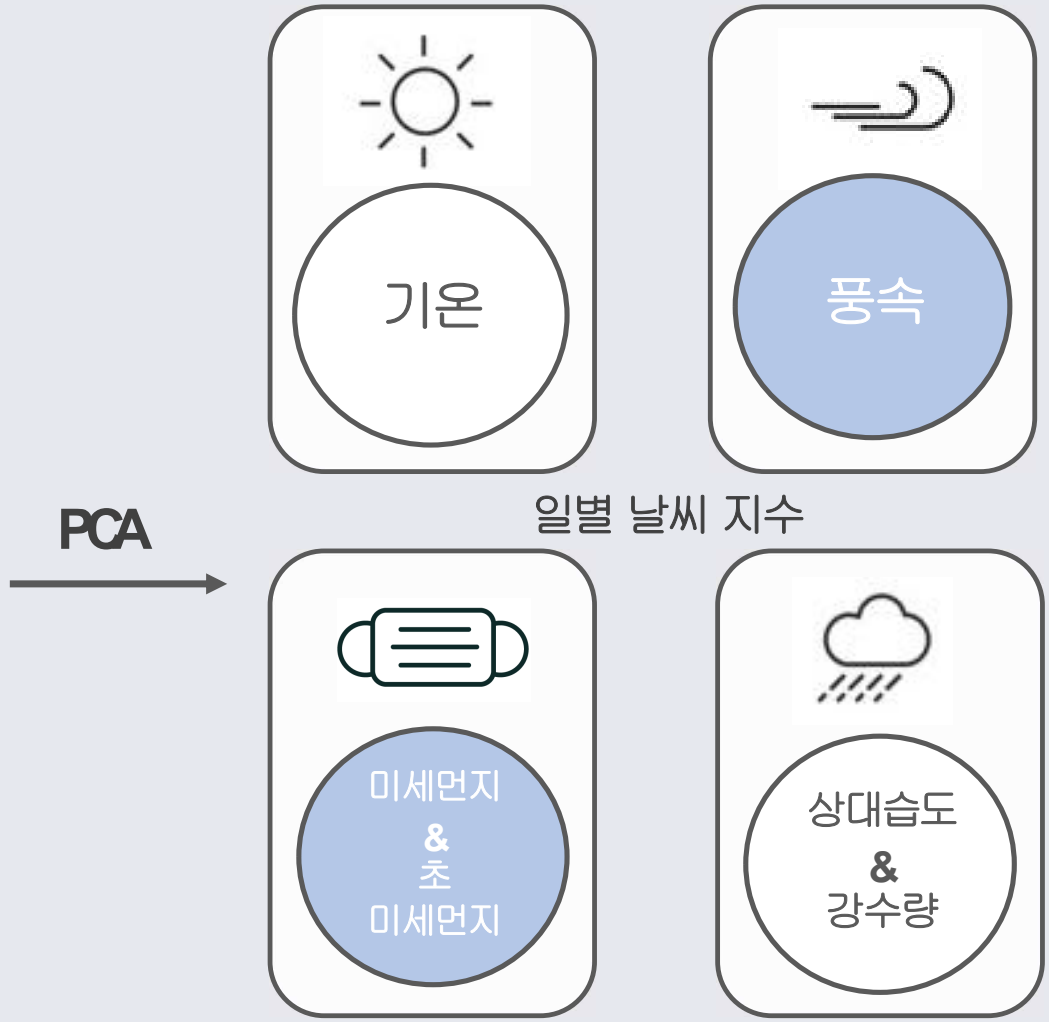
지역별 인구수 현황



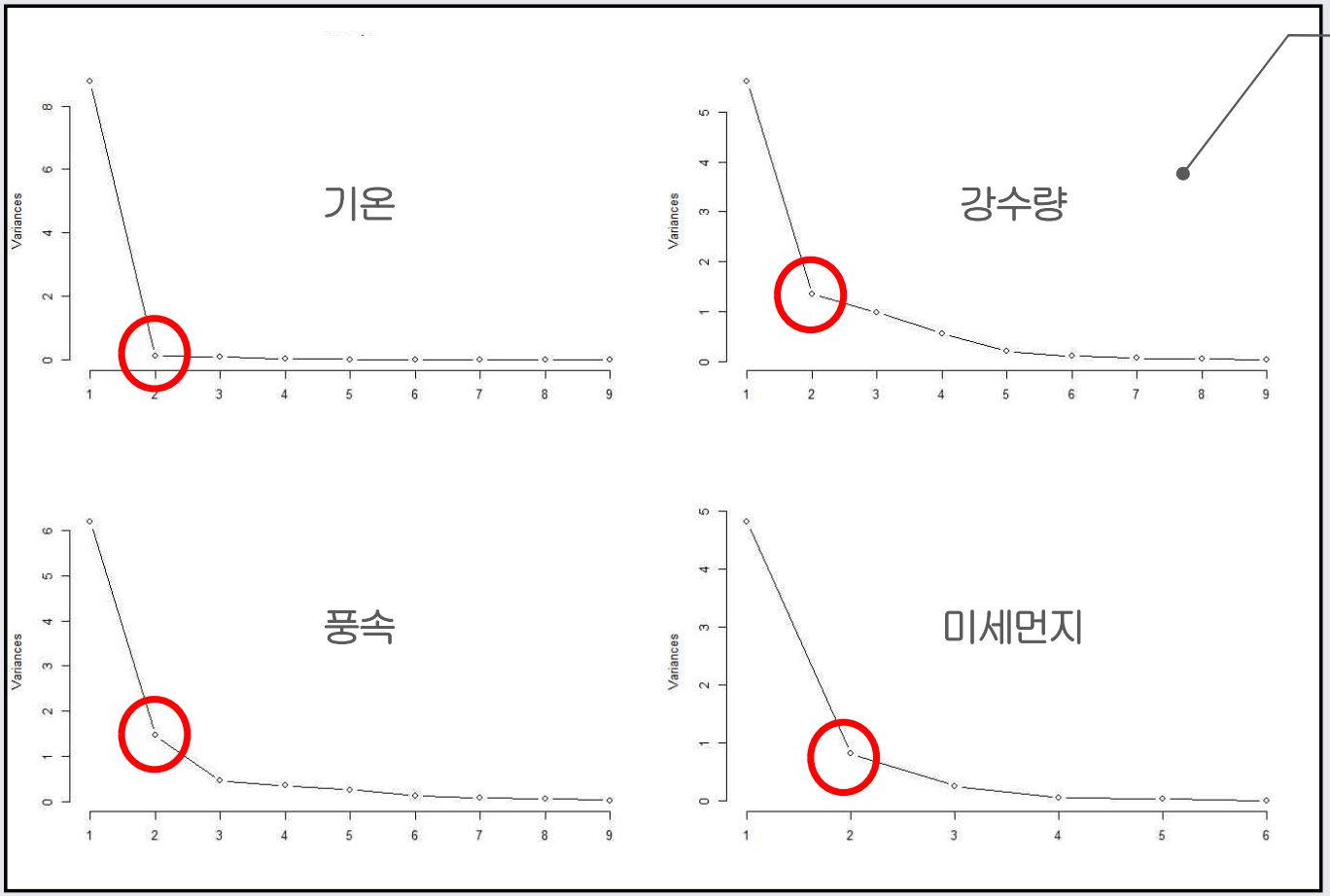
날씨 데이터의 경우, 전체 인구수
(∝**구매자의 비율**)의 절반 이상을 차지하는
서울, 경기, 부산 지역의 데이터를 이용.

날씨지수 생성

변수명	변수설명
Avg_ta	평균 기온
Max_ta	최고 기온
Min_ta	최저 기온
⋮	
Avg_ws	평균 풍속
Max_ws	최대 풍속
Max_ins_ws	최대 순간 풍속
Sum_rn	합계 강수량
Avg_rhm	평균 상대습도
Min_rhm	최소 상대습도
Pm10	미세먼지
Pm25	초미세먼지



날씨지수 생성



Scree plot

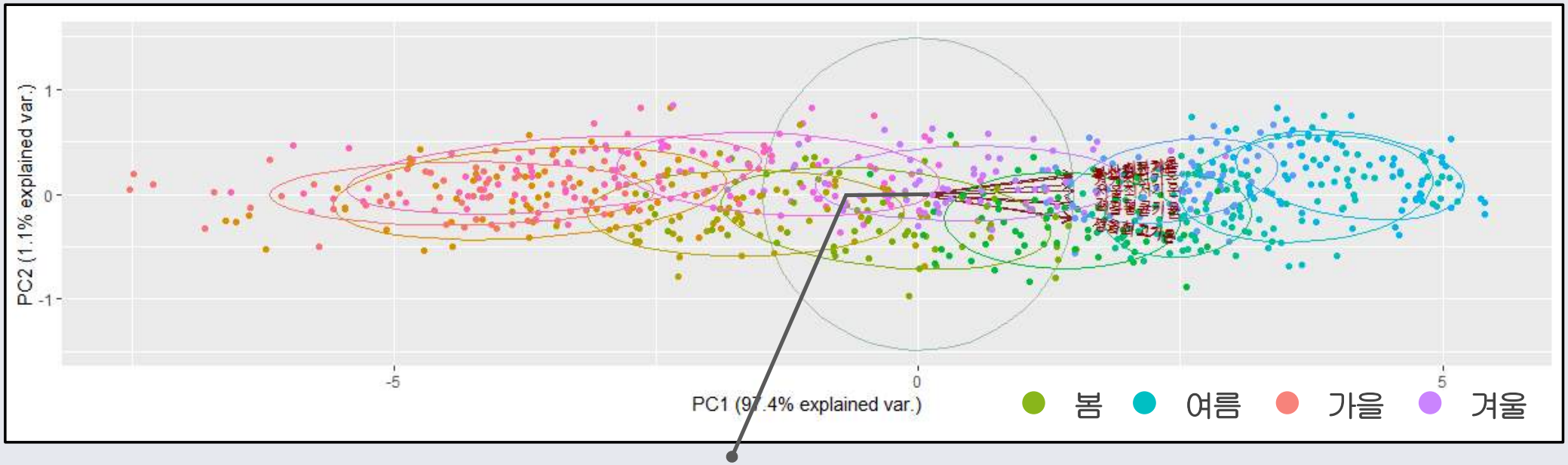
: 기울기가 급격하게 변하는 **elbow**의 개수에 따라 주성분의 개수 선정

기온, 강수량, 풍속, 미세먼지
모두 첫번째 주성분만을 사용하여
날씨 지수 생성

날씨지수 생성

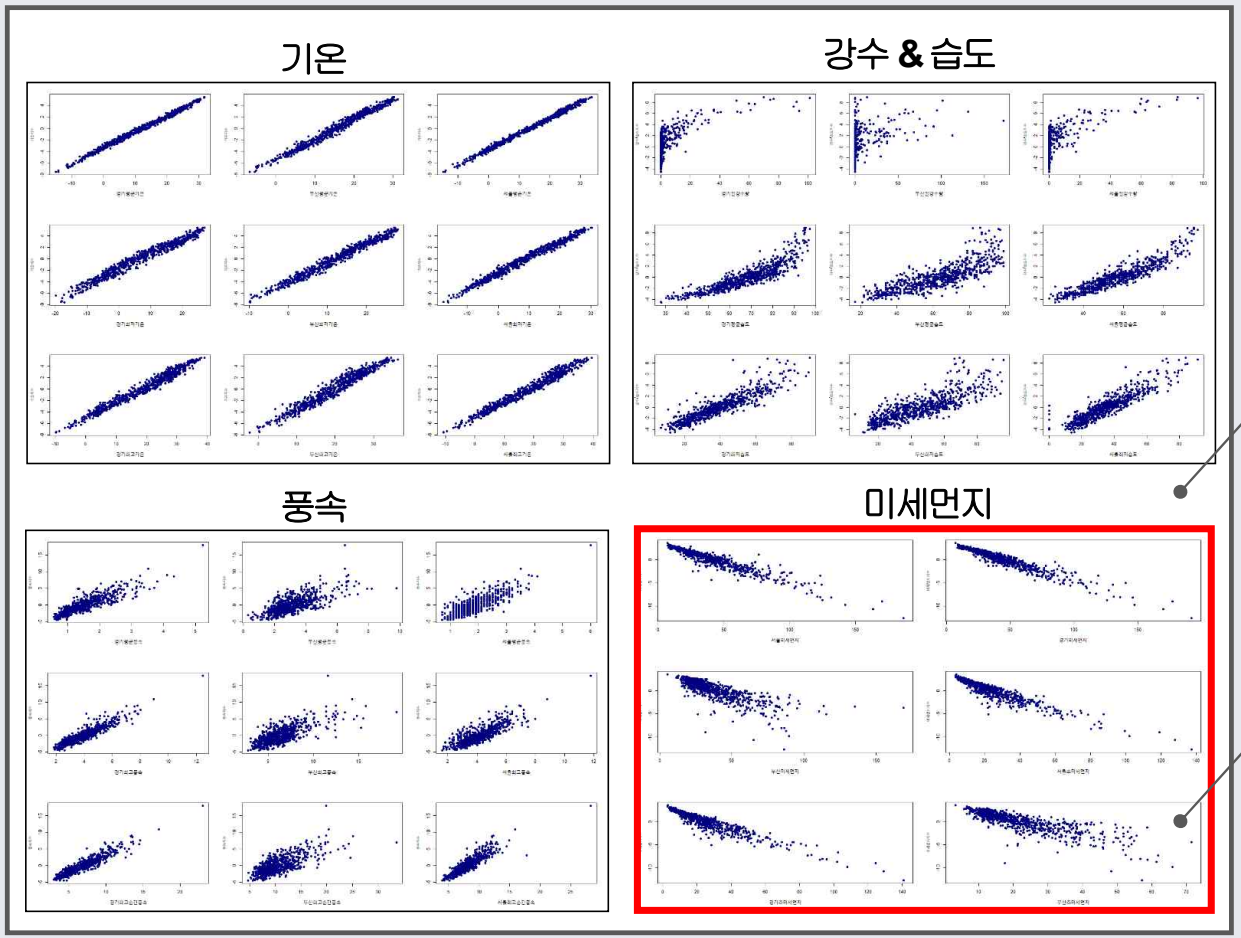
예) 기온지수

= 0.336 * 경기평균기온 **+ 0.334** * 부산평균기온 **+ 0.336** * 서울평균기온 **+ 0.332** * 경기최저기온 **+ 0.333** * 부산최저기온 **+ 0.335** * 서울최저기온 **+ 0.332** * 경기최고기온 **+ 0.329** * 부산최고기온 **+ 0.333** * 서울최고기온



PC2에 대해서는 실제 기온값의 변동성이 작으므로 **PC1**을 주성분으로 사용

날씨지수 생성



기온 ∝ 기온 지수

강수 & 습도 ∝ 강수/ 기온 지수

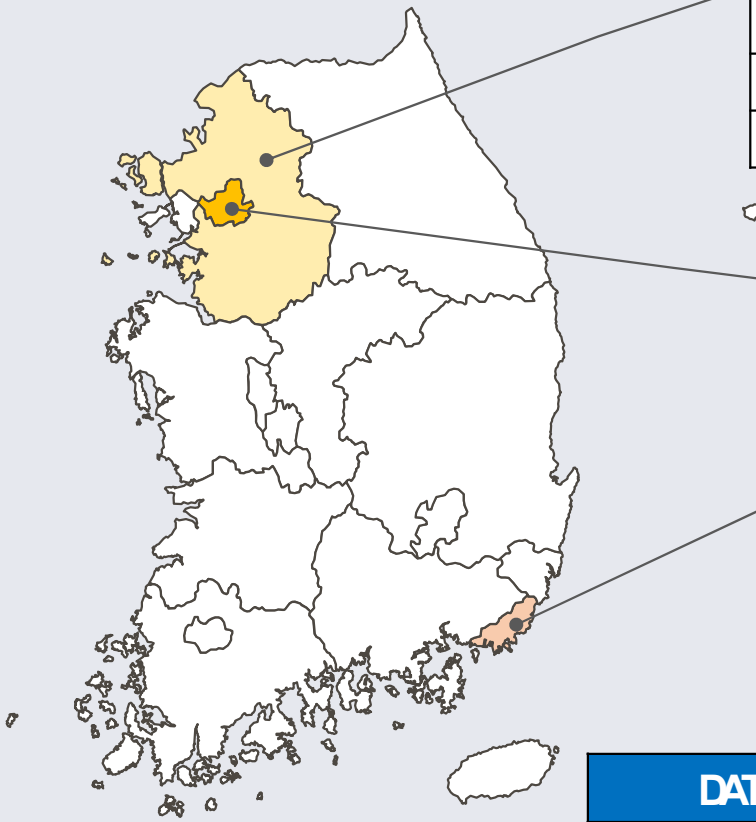
풍속 ∝ 풍속 지수

미세먼지 농도 $\propto \frac{1}{\text{미세먼지 지수}}$

날씨지수와 모두 양의 상관관계를 보이나 미세먼지만 지수와 음의 상관관계를 보임

3. 데이터 처리 방안 및 활용 분석 기법
3.4 날씨지수 생성 – PCA

- 날씨 지수 예시



기온	풍속	미세먼지	습도
AVG 24.3	AVG 1.44	미세먼지 44.65	AVG 77
MIN 20.9	MAX 3.37	초미세먼지 30.1	MIN 59.6
MAX 29.2	순간MAX 5.68		강수 0.04

기온	풍속	미세먼지	습도
AVG 24.7	AVG 1.7	미세먼지 39.9	AVG 67
MIN 22.4	MAX 5.68	초미세먼지 30.7	MIN 48.5
MAX 29	순간MAX 6.8		강수 0

기온	풍속	미세먼지	습도
AVG 22.2	AVG 2	미세먼지 22.1	AVG 88
MIN 20.8	MAX 3.9	초미세먼지 15.2	MIN 70.4
MAX 25.5	순간MAX 7.7		강수 0

DATE	기온지수	풍속지수	미세먼지지수	강수/ 습도지수
2019.06.28	3.0585	2.0314	-1.6537	0.3192

제품군 군집화 (hierarchical clustering)

- 날씨에 큰 영향을 받지 않는 품목, 즉 구매량의 변동폭이 작은 제품을 클러스터링을 통해 제거.
- 대분류별 최적의 k(군집의 수) 를 도출.

소분류별 구매량에 따른
scale을 맞추기 위해
표준화 진행

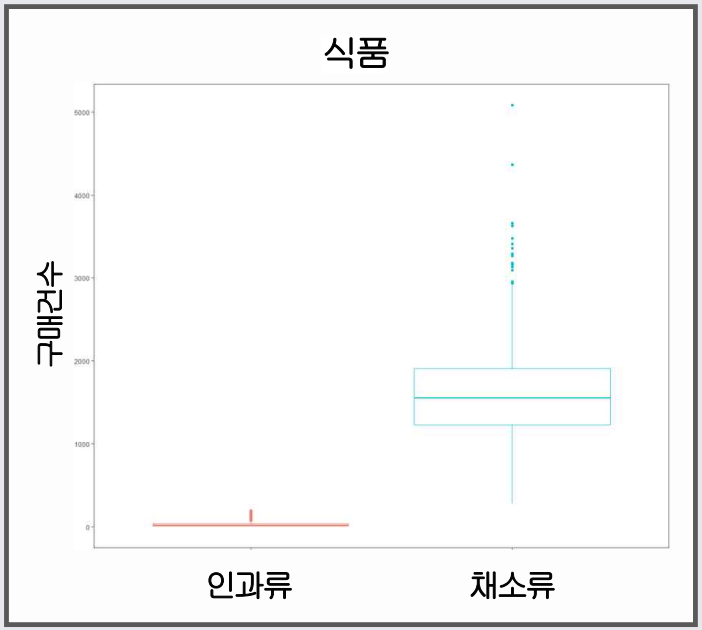
↓

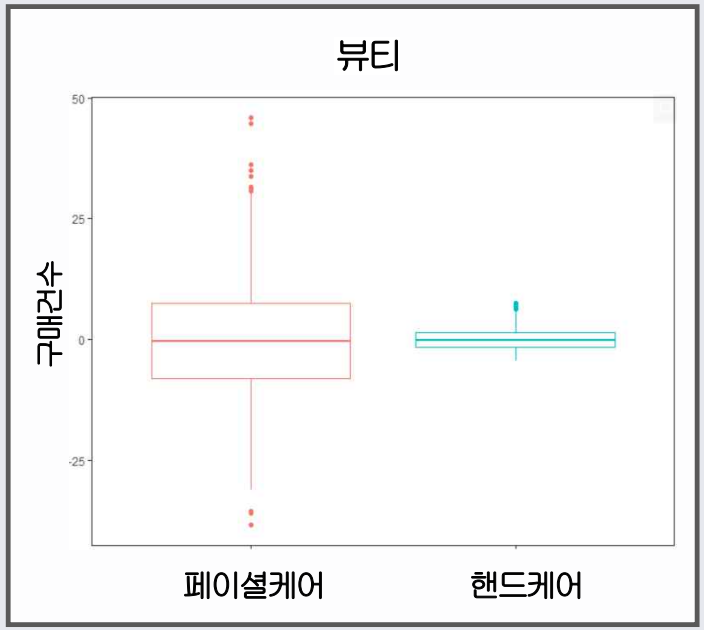
표준화된
최대 구매량 – 최소 구매량
에 대해 클러스터링 시행

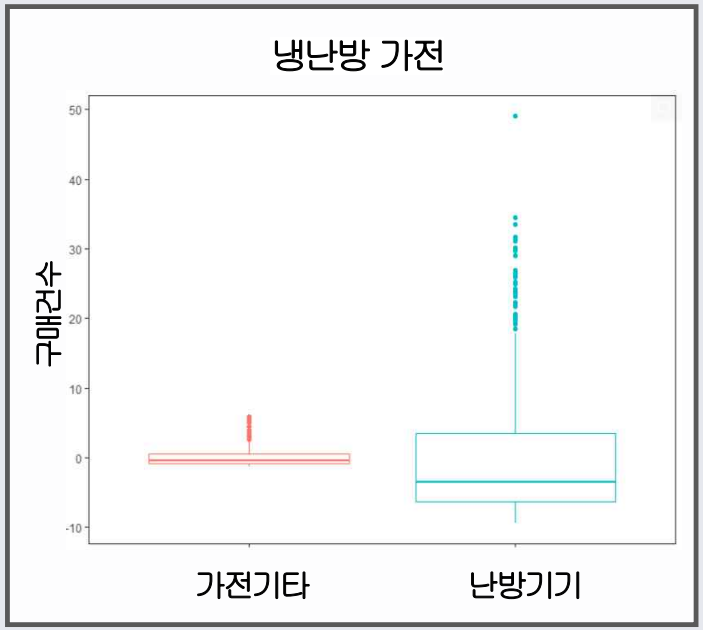
(구매량 – 평균) / 표준편차

대분류	중분류	소분류	일자별 구매량	표준화
식육가공품 및 포장육	식육 가공품	돼지 곱창	524 408 468 ···	- 1.22 - 1.55 - 1.38 ···
		장조림/ 카레용 돈육	11 15 12 ···	- 0.50 0.26 - 0.31 ···
		갈비용 우육	2 2 1 ···	- 1.06 - 1.06 - 1.20 ···

제품군 군집화 (hierarchical clustering)



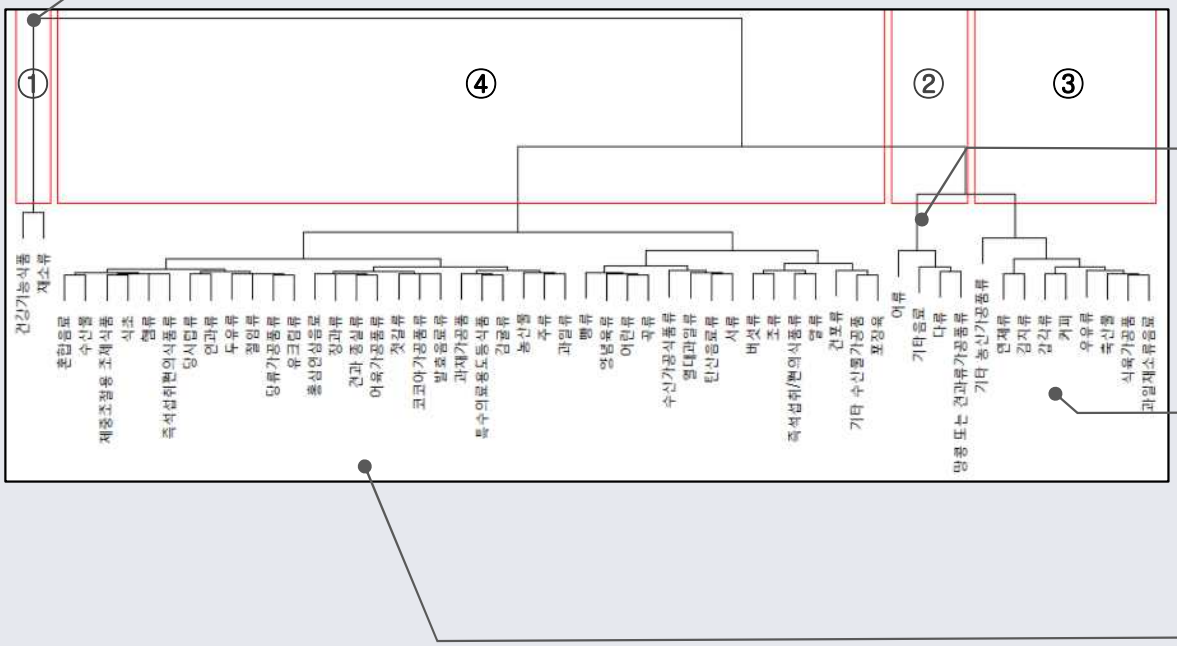




“제품군별 구매량 크기의 차이가 있어 표준화 후 클러스터링 진행”

3. 데이터 처리 방안 및 활용 분석 기법
3.5 온라인 구매이력 – Clustering

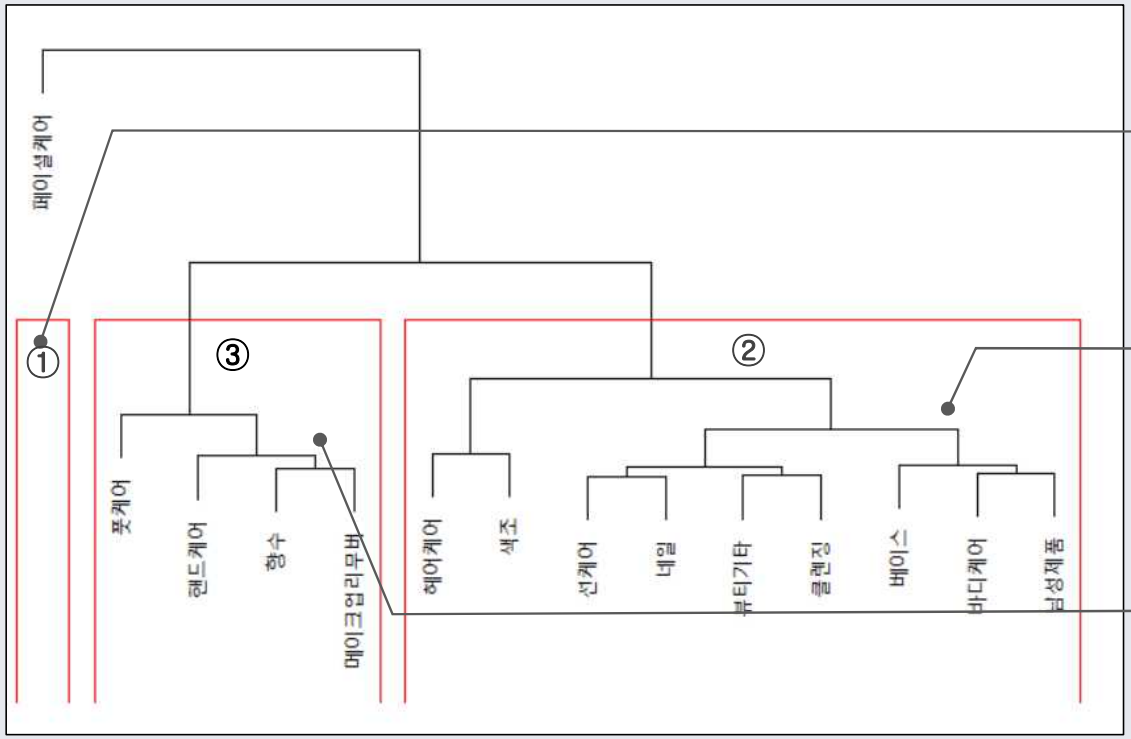
식품 (k = 4)



중분류	최대 – 최저 구매량
건강기능식품	97.4
채소류	75.1
어류	51.6
다류	45.9
땅콩 또는 견과류 가공품류	44.7
기타음료	43.1
기타 농산가공품류	29.9
연체류	28.6
⋮	
과일채소류음료	24.3
버섯류	21.0
조류	20.8
⋮	
인과류	6.09

클러스터링 결과, 군집 ①+ ②+ ③을 모델링에 사용 (④ 제외).

뷰티 (k = 3)

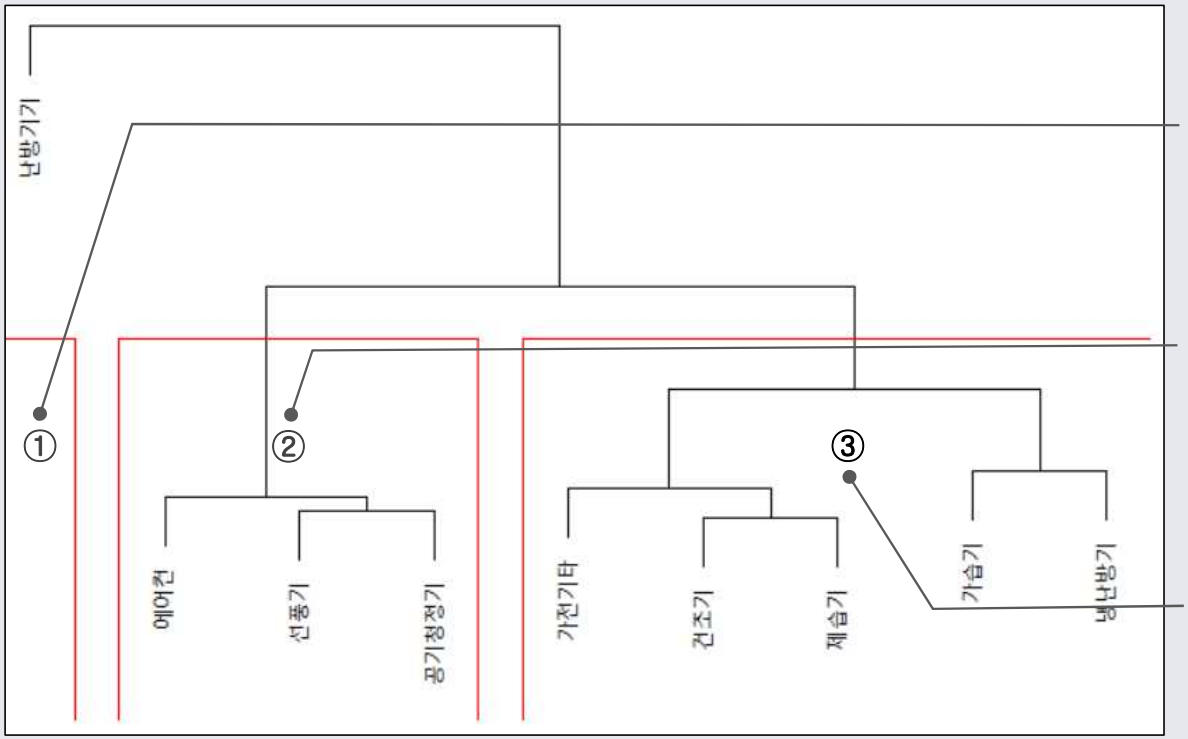


대분류	최대 – 최저 구매량
페이셜케어	84.3
헤어케어	46.3
색조	41.9
…	
클렌징	28.5
풋케어	19.5
핸드케어	12.0
…	
메이크업 리무버	7.98

클러스터링 결과, 군집 ① + ②을 모델링에 사용 (③ 제외).

3. 데이터 처리 방안 및 활용 분석 기법
3.5 온라인 구매이력 – Clustering

냉난방 가전 (k = 3)



대분류	최대 – 최저 구매량
난방기기	58.5
에어컨	31.3
선풍기	29.8
공기청정기	29.0
가습기	20.5
냉난방기	15.6
...	
가전기타	7.11

클러스터링 결과, 군집 ① + ②을 모델링에 사용 (③ 제외).

다중회귀분석 (Multiple Linear Regression Analysis)

- 클러스터링 결과 선정된 제품군에 대해 소셜 데이터의 문서건수 변수(**SNS_cnt**)를 추가하여 다중회귀분석 시행.
- 단계별 변수선택으로 유의하지 않은 변수 제거. **(p-value > 0.05)**
- 제품군별 구매에 영향을 미치는 날씨 변수 도출 및 소셜 데이터의 영향력 검정.

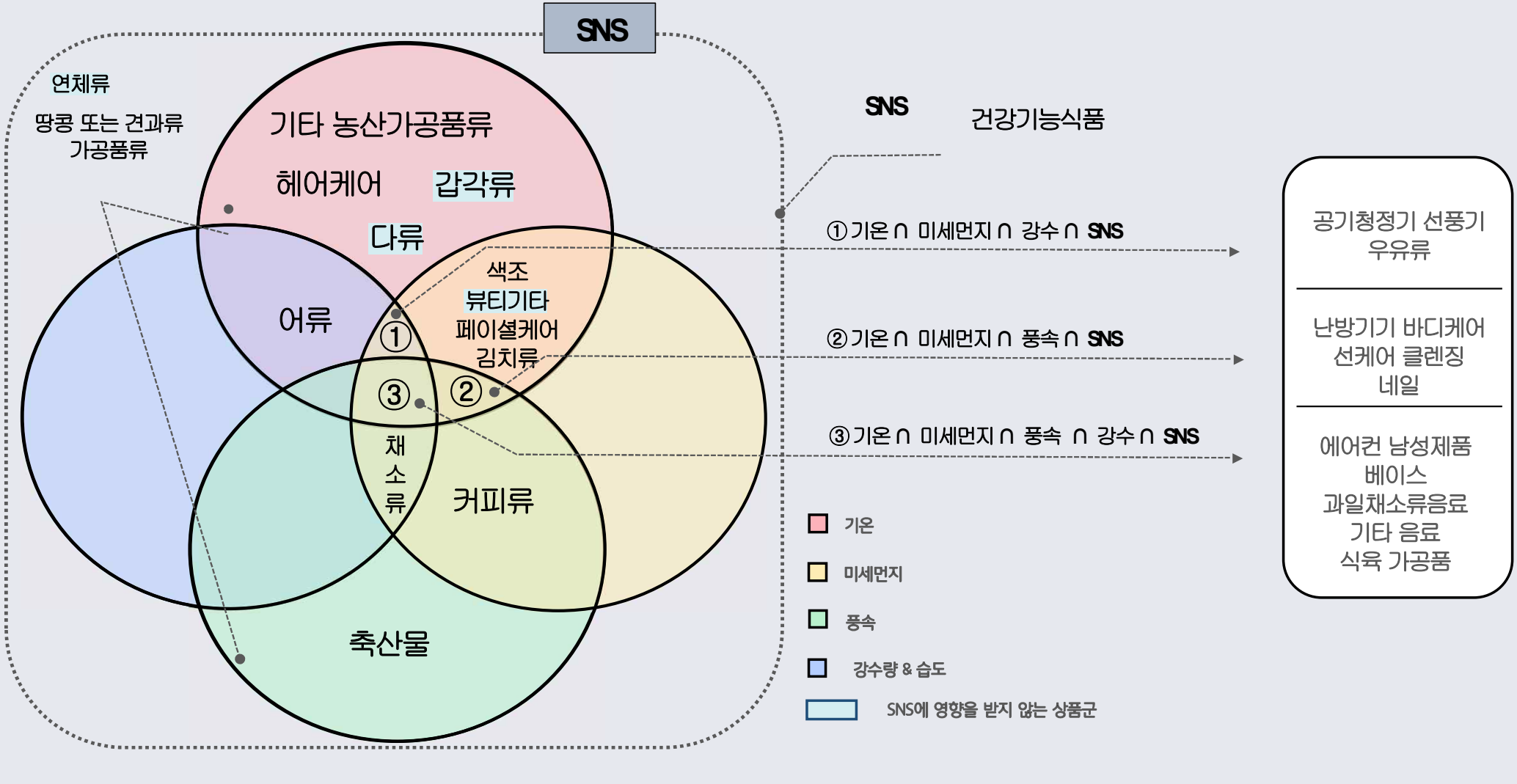
VIF를 통해 다중공선성 없음을 확인. (**VIF < 10** 이면 **FALSE**)

Ex) 냉난방 가전 _ 공기청정기

	SNS 문서건수	기온 지수	강수량 지수	풍속 지수	미세먼지 지수
VIF	FALSE	FALSE	FALSE	FALSE	FALSE
	(Intercept)	SNS 문서건수	기온 지수	강수량 지수	미세먼지 지수
회귀계수	-146.961	5.895	-9.650	-6.966	11.634

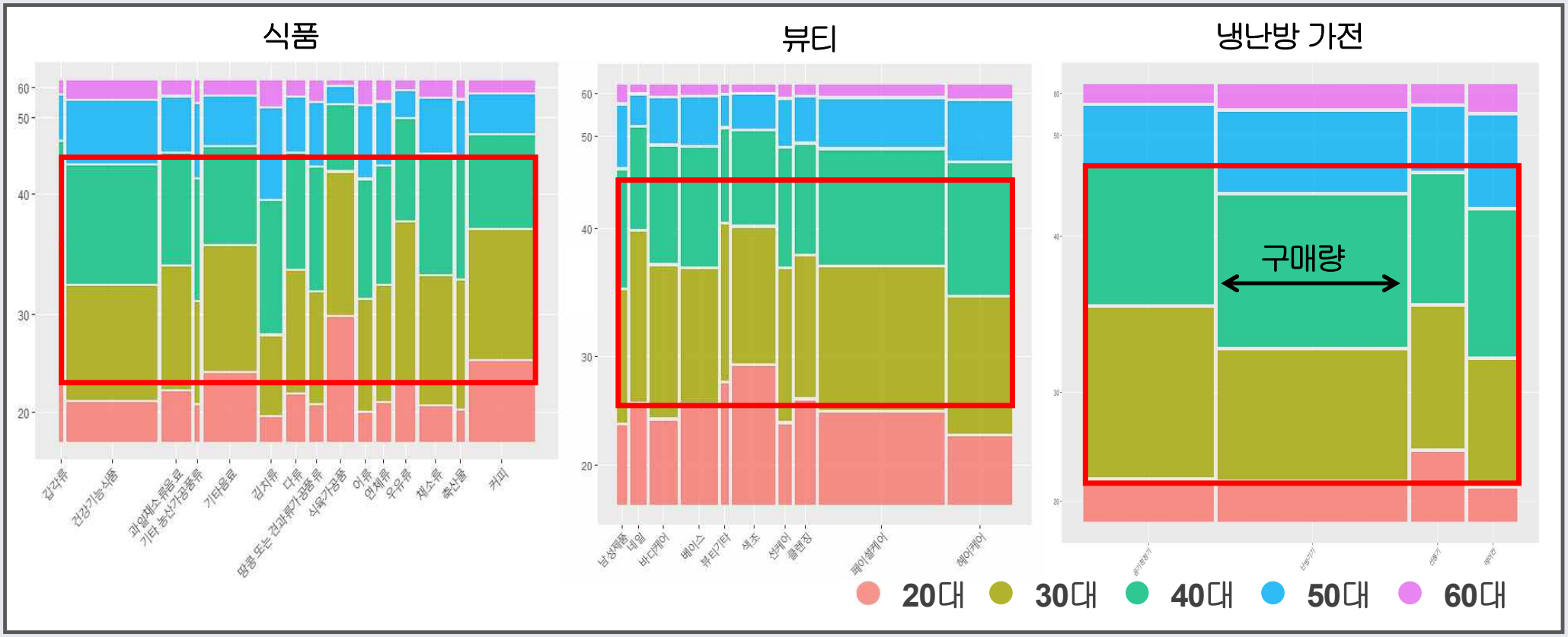
공기청정기의 구매량은 **SNS**, 기온, 강수/ 습도, 미세먼지에 영향을 받음.

3.6 데이터 처리 방안 및 활용 분석 기법
다중회귀분석



4.1 소비 트랜드 분석
상품 구매 연령대

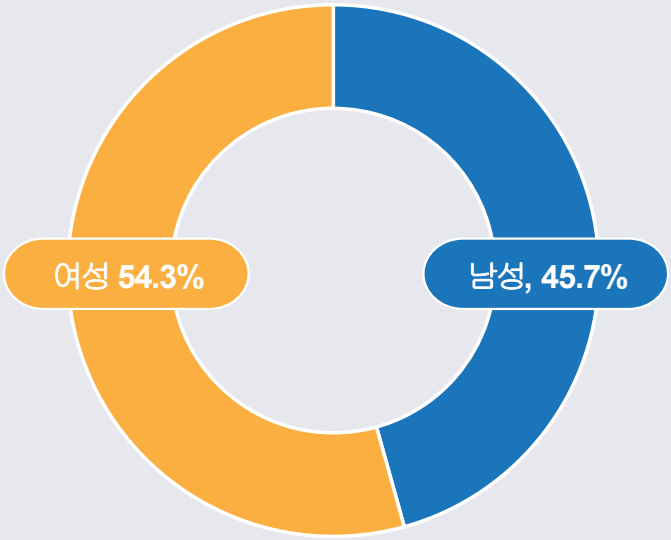
연령대

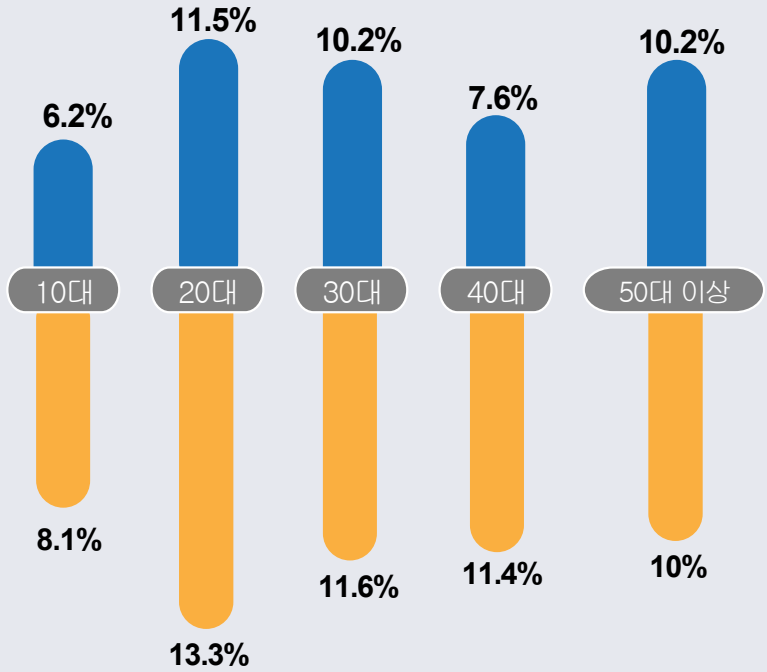


날씨에 민감한 제품군 별 소비층의 연령대를 확인 해 본 결과, 주 소비층이 **30, 40**대 인 것을 알 수 있음.

소셜 미디어

SNS 모바일 앱 사용자 분포



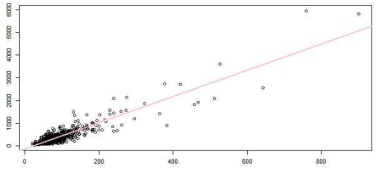
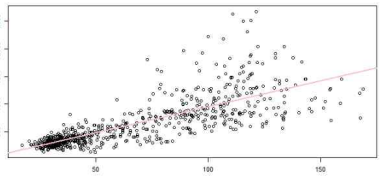
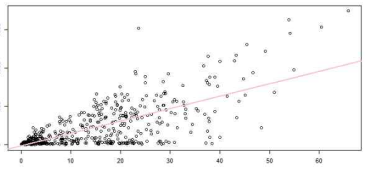
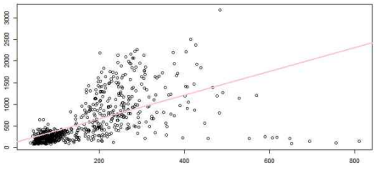
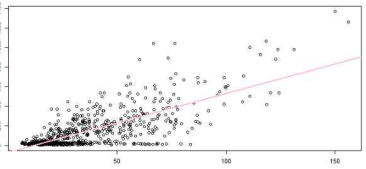


남성보다는 여성이 **SNS**를 더 활용하며,

20, 30, 40대가 전체 **SNS** 이용 비율의 약 **70%**를 차지

소셜 미디어

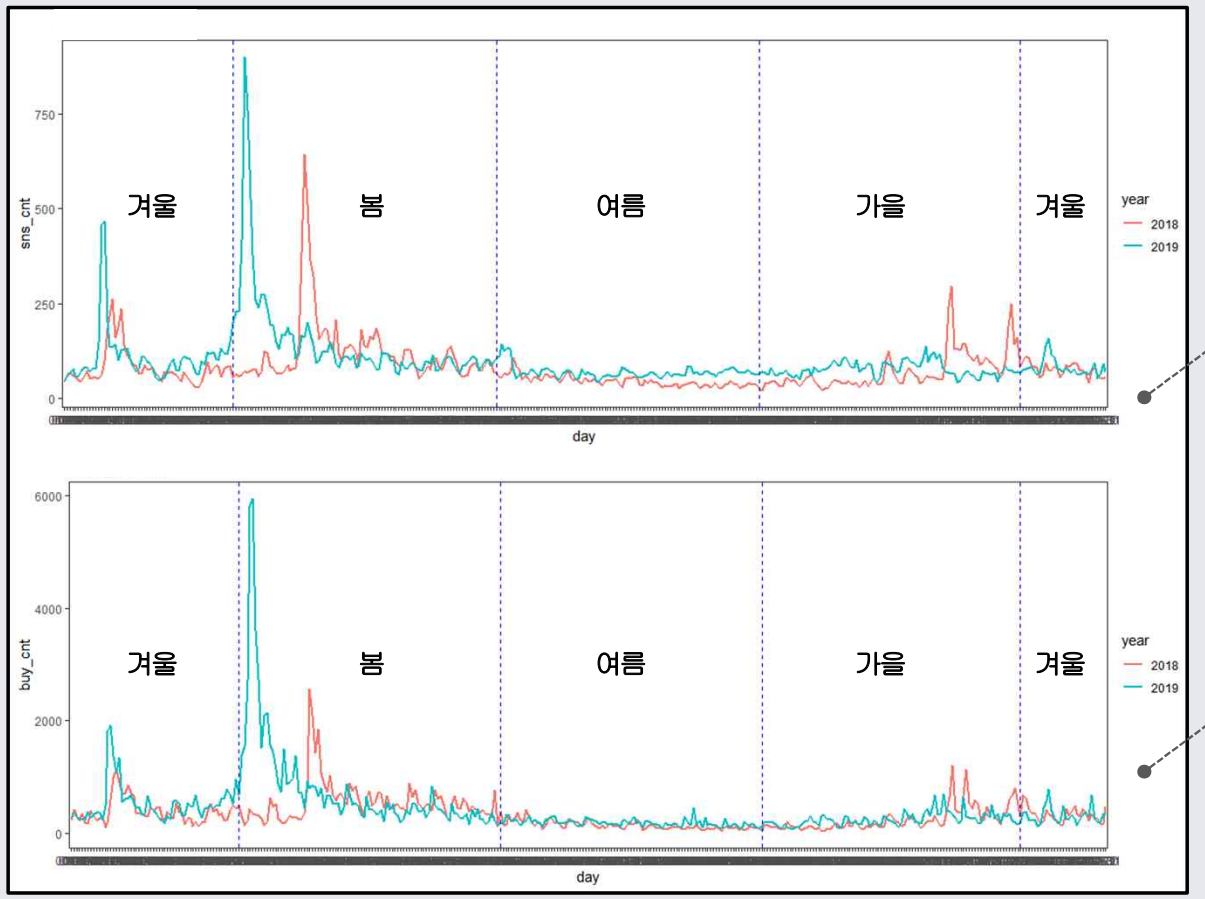
- 날씨에 민감한 제품군 중, **SNS** 문서 건수와 온라인 구매 수량과의 상관계수가 가장 높은 상품 군 **Top 5**

공기청정기	선케어	선풍기	에어컨	난방기기
				
0.915	**0.779**	**0.758**	**0.749**	**0.605**

↓

"SNS 문서 건수와 온라인 구매 건수간의 경향성 확인"

소셜 미디어



SNS 홍보가 적절한 예시 _ 공기청정기

SNS 문서건수

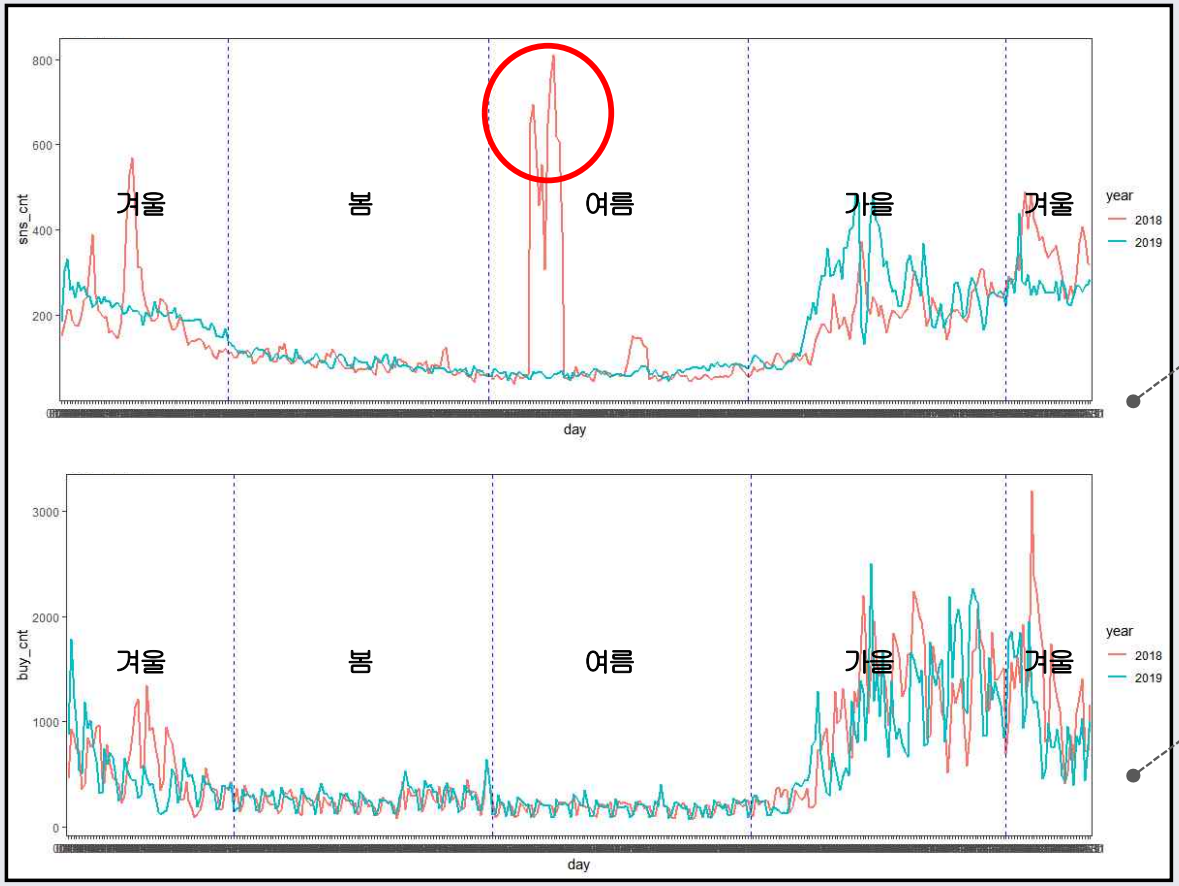
온라인 구매 건수

공기청정기의 SNS의 문서 건수와
온라인 구매 수량의 그래프
패턴이 거의 동일.

각 계절별 소비자의 관심도와
구매량 패턴이 유사함.

**소셜 미디어를 활용한 홍보를
통해 제품에 대한 관심이
실제 구매로 이어지도록 유도.**

소셜 미디어



SNS 홍보가 부적절한 예시 _ 난방기기

SNS 문서건수

온라인 구매 건수

난방기기의 SNS의 문서 건수와
온라인 구매 수량의 그래프
패턴이 대체로 유사함.

하지만 여름에 SNS 상에서의
난방기기의 관심도가
실제 구매 건수로
이어지지는 않음.

**해당시기에 기온에 민감한
제품군인 난방기기를
소셜 미디어를 통해
홍보를 하는 것은 부적절함.**

5. 예측 모델
5.1 모델 선택 배경

시계열분석

- 날씨에 따른 제품군별 수요예측을 위해 시간에 따라 변화하는 날씨의 특성을 반영한 시계열 모형을 고려.
- 사용 변수 : **2018~19**년간 **28**개의 각 제품군별 구매건수, 기온지수, 강수/ 습도지수, 풍속지수, 미세먼지지수

다중회귀결과 날씨에 민감한 제품군이 아니므로 제외

종속변수(Y) : 제품군별 구매건수

설명변수(X) : 날씨 지수

	날짜	갑각류	건강기능식품	공기청정기	...	페이셜케어	헤어케어	기온 지수	강수/ 습도 지수	풍속 지수	미세먼지 지수
Time: 2018 ~ 2019	2018-01-01	100	4051	243	…	8137	2742	-4.224586	-2.880586	-0.51514927	0.05140376
	2018-01-02	108	4999	425	…	9659	3392	-4.119653	-2.629586	0.06208944	-1.16783476
	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	⋮
	2019-12-30	215	5249	347	…	8043	3411	-2.838804	1.438220	4.034096	4.034096
	2019-12-31	283	4678	316	…	6401	2576	-5.430106	-2.629586	4.309431	4.309431

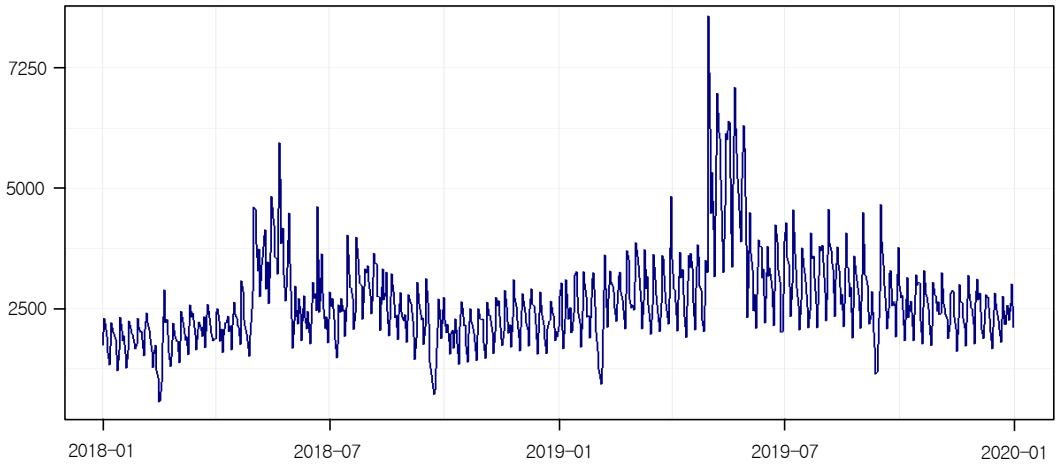
5. 예측 모델
5.1 모델 선택 배경

시계열분석

- 시계열 데이터

: 시간의 경과에 따라서 얻어진 관측값/ 통계량으로
연속한 시간의 관측치가 연관되어 있음

X : 날씨 정보 → 시계열 데이터
Y : 날짜별 구매건수 → 시계열 데이터



- 분석

기온, 강수/ 습도, 풍속, 미세먼지 지수	
일반(비시계열) 모형	시계열 모형
날씨 지수(**X**) ↓ 제품군별 구매량(**Y**)	날씨 지수(**X**), 제품군별 과거 구매량(Y_{t-i}) ↓ 제품군별 현재 구매량(Y_t)
제품군별 구매량 예측	

5. 예측 모델
5.1 모델 선택 배경

시계열분석

- 다변량 반응변수

제품군 1
갑각류

$$y_1 = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,730} \end{pmatrix}$$

$y_{1,1}$ → 2018-01-01

$y_{1,730}$ → 2019-12-31

제품군 2
공기청정기

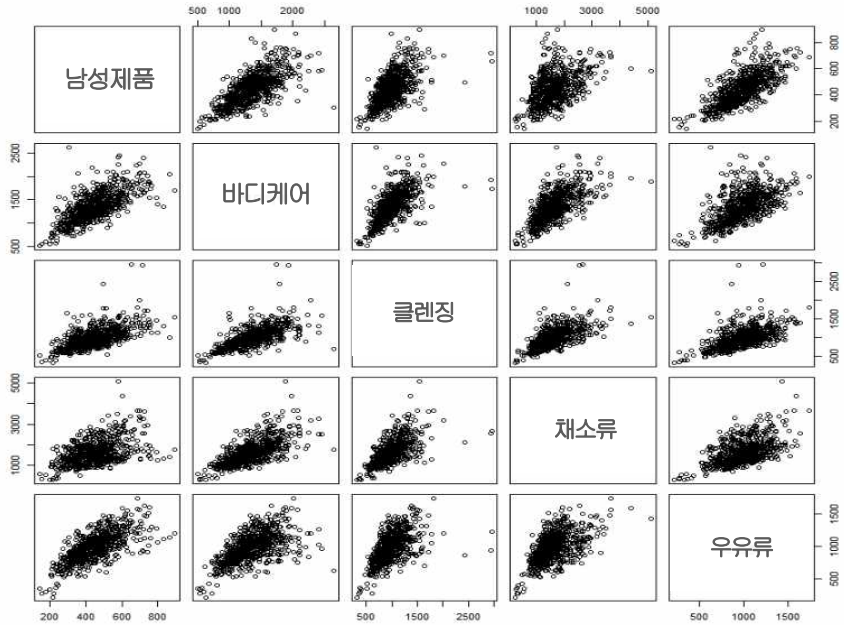
$$y_1 = \begin{pmatrix} y_{2,1} \\ \vdots \\ y_{2,730} \end{pmatrix}$$

⋮ ⋮

제품군 28
헤어케어

$$y_{28} = \begin{pmatrix} y_{28,1} \\ \vdots \\ y_{28,730} \end{pmatrix}$$

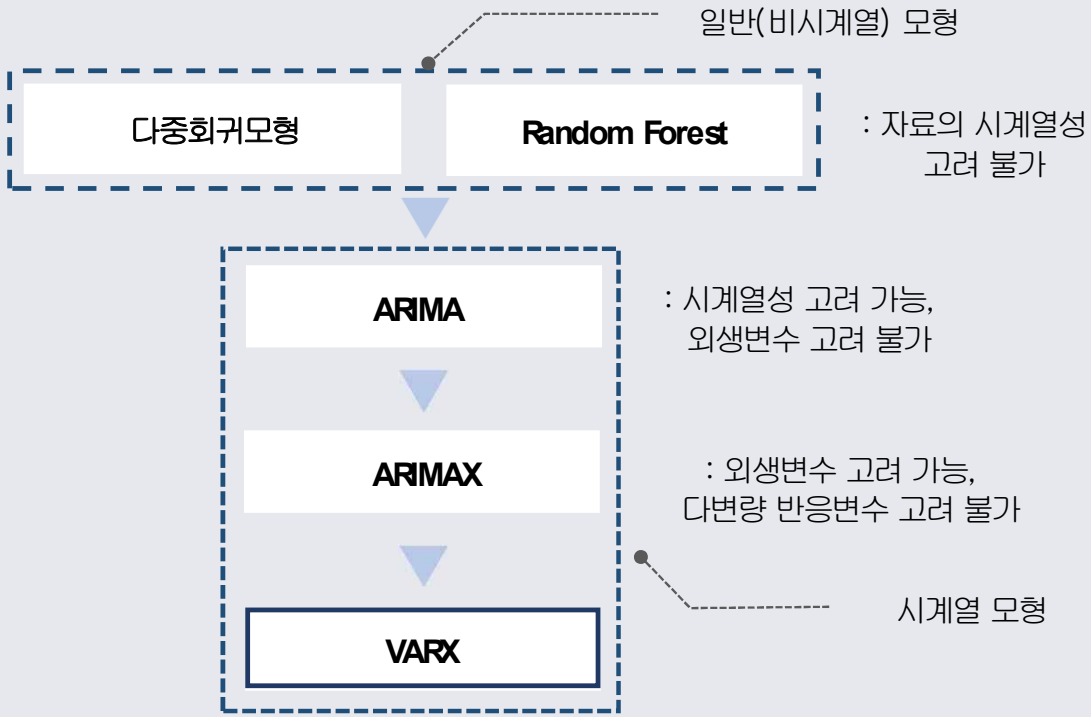
- 제품군 사이의 연관성



서로 강한 상관관계를 가지는 제품군 존재

5.2 예측 모델
모델 비교

모형 \ 데이터 특성	설명변수	시계열성	다변량	Y간 연관성
다중회귀모형	O	X	X	X
Random Forest	O	X	O	X
ARIMA	X (외생변수)	O	X	X
ARIMAX	O (외생변수)	O	X	X
VARX	O (외생변수)	O	O	O



→ 최종적으로 시계열성, 외생변수, 다변량 반응변수를 동시에 고려해 분석할 수 있는 **VARX** 모형을 선택 .

VARX (Vector AutoRegressive with Exogenous Model)
회귀 분석에서 선택된 제품군 29개를 하나의 벡터로 만들어 한 번에 시계열 분석을 시행.

VARX

$$Y_t = \sum_{i=1}^{p} \Theta_i \, Y_{t-i} + \sum_{j=1}^{n} \beta_j \, X_j + \varepsilon_t$$

(p : number of lag , n : number og exogenous)

$\boldsymbol{Y_t}$: 제품군별 t 시점 구매량 벡터, $\boldsymbol{X_j}$: 날씨변수 (기온지수, 강수/ 습도지수, 풍속지수, 미세먼지지수) , $\boldsymbol{\varepsilon_t}$: t 시점 오차벡터

VARX 적용 시계열 데이터 예측 사례

(금융 데이터: 통화정책 효과 예측)

Y : 내생변수 / X : 외생변수

날짜	지역총생산 ($GRDP_i$)	소비자 물가지수 (CPI_i)	…	기준금리 (SR_t)	…
2018-01	3.6	104.45	…	1.5	…
⋮	⋮	⋮	…	⋮	…

구매량 예측에 VARX 모델 적용

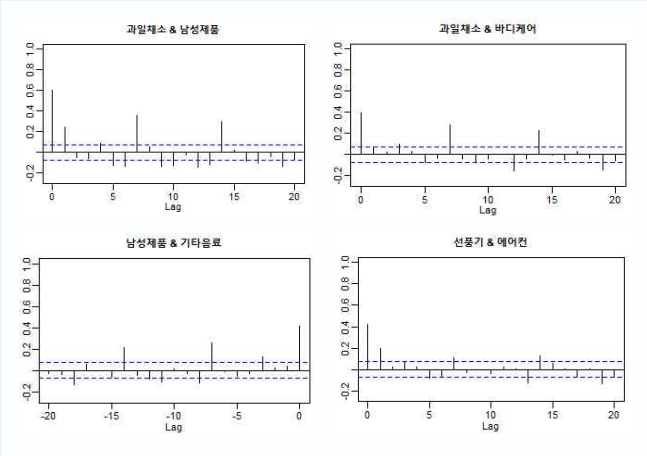
Y : 제품군별 구매건수 / X : 날씨 지수

	날짜	갑각류 (Y_1)	…	헤어케어 (Y_{28})	기온지수 (X_1)	…	미세먼지지수 (X_4)
Time : 2018 ~ 2019	2018-01-01	100	…	2742	-4.2245	…	0.0514
	2018-01-02	108	…	3392	-4.1197	…	-1.1678
	⋮	⋮		⋮	⋮		⋮
	2019-12-31	283	…	2576	-5.4301	…	4.3094

VARX

ACF) 시계열 관측치 y_t 와 y_{t-k} 간 상관의 측도

유의한 큰 값이면 해당 시차에 대한 상관 有



제품군 여러 개를 한 번에 고려했을 때
시차에 대한 상관 있음

제품군을 동시에 모형화해야 함

<u>다변량</u> 반응변수 시계열 모형인 **VARX**를 사용

VAR select — **AR** 차수 결정

AIC(Akaike Information Criterion) 비교

Model	AIC
VARX(1,4)	290.678
VARX(2,4)	290.493
VARX(3,4)	290.624

AIC
가장 작은
모형 선택

<u>외생변수(4개)</u>
기온지수, 강수/ 습도지수, 풍속지수, 미세먼지지수
<u>AR(2)</u>
→ **VARX(2,4)** 모형이 가장 적합

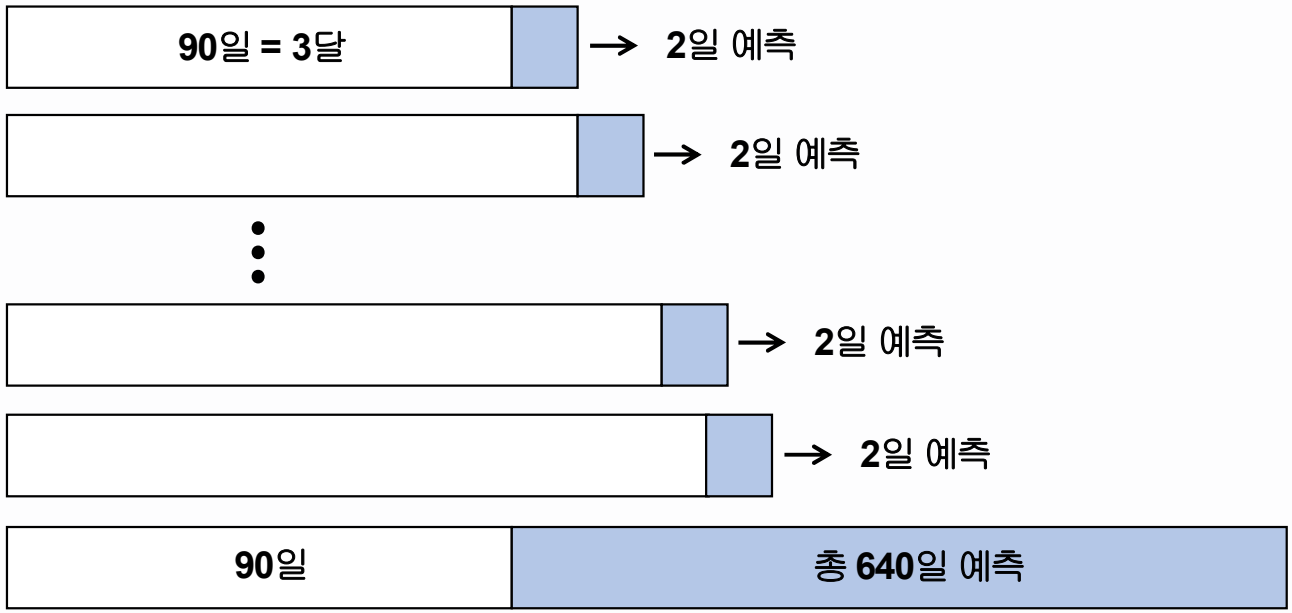
VARX

Rolling Walk Forward

시계열 모형 예측의 대표적 방법

장기간의 추세가 모델의 분산을
과도하게 높여 발생하는
낮은 예측력 문제 완화

Ex) Window = 90 days
h = 2

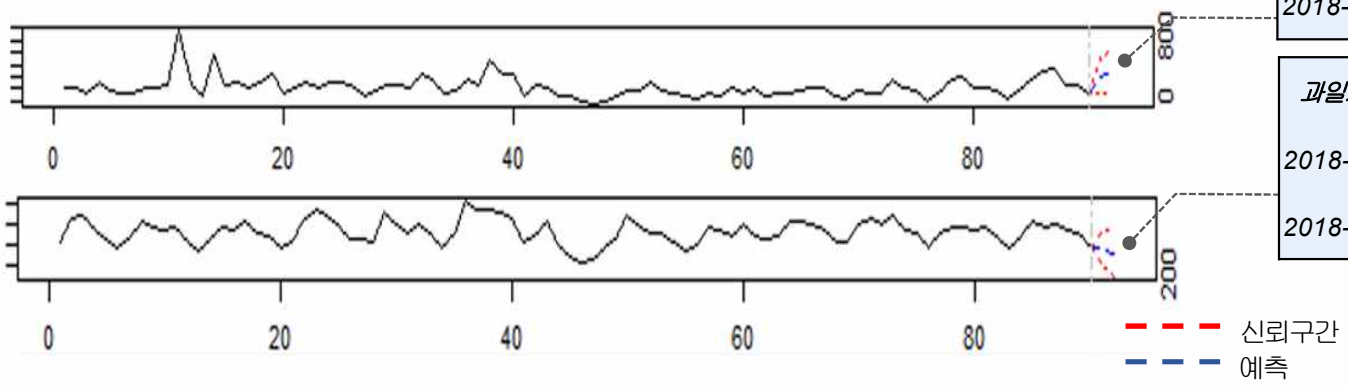


VARX

예측결과

예시) **Rolling Walk Forward**

2018-01-01 ~ 2018-03-31 → **2일 예측**



갑각류	*예측값*	*lower*	*upper*	*α*
2018-04-01	*152.6205*	*77.13087*	*228.1101*	*75.48963*
2018-04-02	*164.455*	*67.78261*	*261.1274*	*96.67237*

과일채소류	*예측값*	*lower*	*upper*	*α*
2018-04-01	***907.9564***	***488.621***	***1327.292***	***419.3354***
2018-04-02	***750.7322***	***151.6996***	***1349.765***	***599.0326***

5 예측 모델
.3 VARX

VARX VARX 모델을 사용하여 2019-06-28 구매량을 예측.

	제품군	예측 구매량		제품군	예측 구매량
1	갑각류	195.981	15	색조	1814.676
2	공기청정기	91.44391	16	선케어	676.3235
3	과일채소류음료	1364.001	17	선풍기	481.2186
4	기타.농산가공품류	193.3409	18	식육가공품	1908.954
5	기타음료	2431.978	19	어류	518.023
6	김치류	1069.625	20	에어컨	213.1371
7	난방기기	28.09676	21	연체류	751.4151
8	남성제품	315.3373	22	우유류	838.223
9	네일	1104.814	23	채소류	2254.805
10	다류	443.3668	24	축산물	246.3193
11	땅콩 또는 견과류 가공품류	269.8598	25	커피	2604.019
12	바디케어	1202.255	26	클렌징	1051.397
13	베이스	1430.925	27	페이셜케어	6360.275
14	뷰티기타	352.2093	28	헤어케어	3097.295

VARX

모델 예측 정확도(accuracy) 비교

- ✓ 시간의 흐름(시계열)을 반영하지 않은 모델 (Linear Regression, Random Forest) 과 시간의 흐름을 반영한 모델 (ARIMAX, VARX) 비교
- ✓ 시계열 모델 예측 정확도가 더 우수함
- ✓ 시계열 모델 중 다변량을 반영한 모델 (VARX) 의 예측 정확도가 더 우수함
- ✓ 정확도 순서 : **VARX** > ARIMAX > Linear Regression > Random Forest

MODELS	RMSE
Linear Regression	**315.2934**
Random Forest	**355.5274**
ARIMAX	**282.2624**
VARX	**272.6845**

RMSE: Root mean squared error

6.1 결과 활용방안 및 제안
VARX 모델 사용 예시

동서식품주식회사

내일 날씨에는 어떤 브랜드의
제품이 많이 팔릴지 VARX
모형 한번으로 구매량 예측 가능.

Maxim　KANU　Maxwell House　T.O.P

더 세분화 한 제품도 VARX 모형
이용하여 KANU의 어떤 종류가 많이
팔릴 지 예측 가능.

SNS & 날씨 마케팅

(1) 날씨와 **SNS**에 영향을 받는 제품을 파악.

(2) 모델을 통하여 어떤 날씨의 제품의 구매량 예측.

(3) 제품의 주 소비층의 연령대를 파악.

• 연령대 별 주 **SNS** 플랫폼

(출처: happist.com)

(4) 주 소비층의 연령대가 많이 쓰는 **SNS**에 집중적으로 홍보를 한다면 효율적인 수익 창출 가능.

↓

날씨에 따라 수요를 예측하여 적절한 시기에 **SNS**홍보를 시행.
주 소비층을 타겟팅 하여 홍보에 적절한 **SNS**플랫폼 선택.

6 결과 활용방안 및 제안
.2 SNS 전략

- 구체적인 제안 방향 및 기대효과

- 날씨 데이터 **+ VARX** 모델 **=** 주간 상품 구매 예측 활용

: 날씨를 고려한 예측을 통해 구매를 더 유도하고 싶은 제품에 대한 레시피 및 상품 판매 사이트를 함께 제안.

예시)

2019.06.28 (금) 서울

흐림/ 비/
안개/
이슬비
22/ 29°C

날씨 데이터

VARX

구매를 유도하고 싶은 제품 파악

SNS

사이트

Kurly
SSG.COM
롯데마트

레시피 및 상품 판매 사이트 제안

6.2 결과 활용방안 및 제안
SNS 전략

예시)

2019.6.28.	서울	흐림/ 비/ 안개/ 이슬비 22/ 29℃			
기온	풍속	미세먼지 초미세먼지		상대습도 강수량	
20℃	2m/ s	19 μg/m³	14 μg/m³	56%	4.5mm

구매 유도 제품 **PICK** ⇣ **VARX** (수요예측)

제품	키위	망고	감귤	…
일별 예측 수요량	5	150	30	…

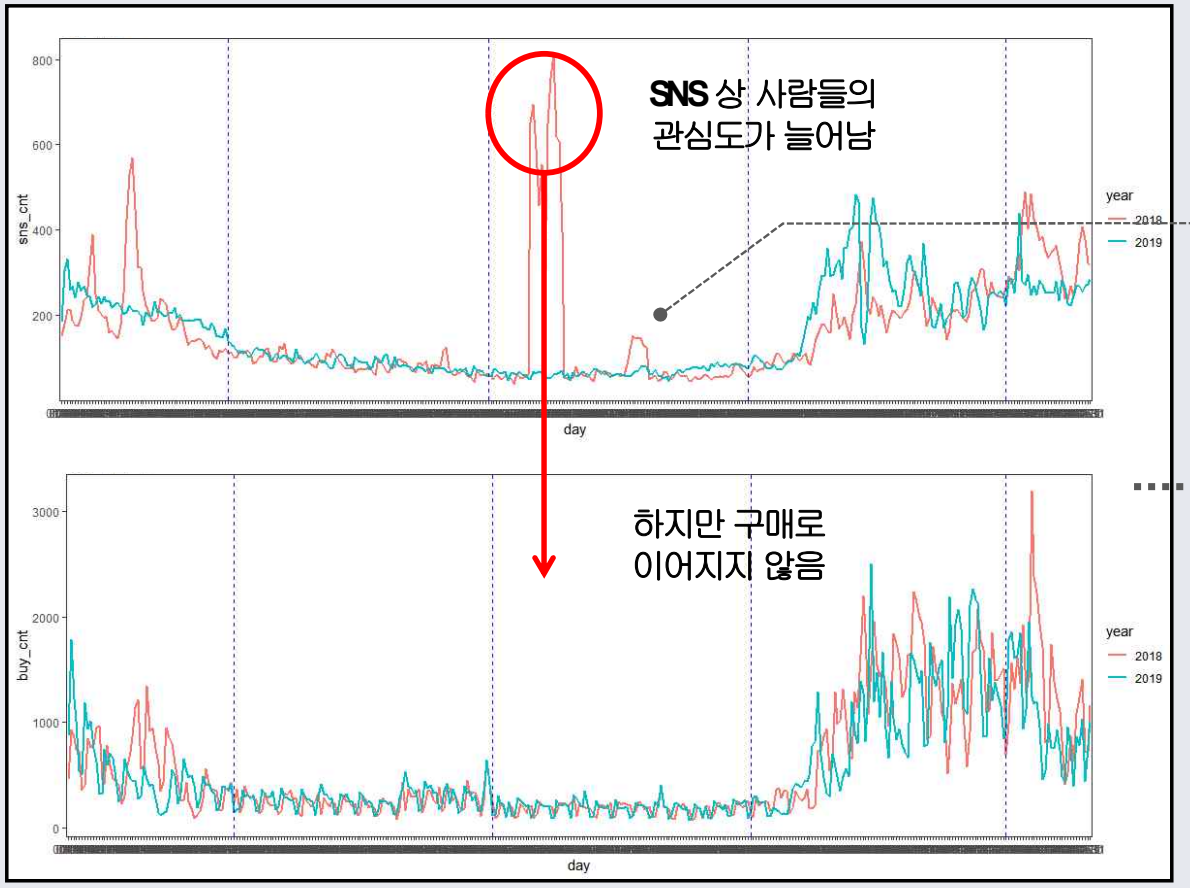
⇣ **SNS** 홍보

망고 관련 레시피 & 재료 구매 사이트(Kurly SSG 롯데마트) 제시

⇣

상품에 대한 관심도가 증가하여
구매 전환까지 이어질 것을 기대해 볼 수 있음.

SNS & 날씨 마케팅



SNS의 노출 건수를 늘려도
구매로 이어지지 않을 것.

해당 시기에 **SNS**의 홍보 효과가
부적절해 보일 때

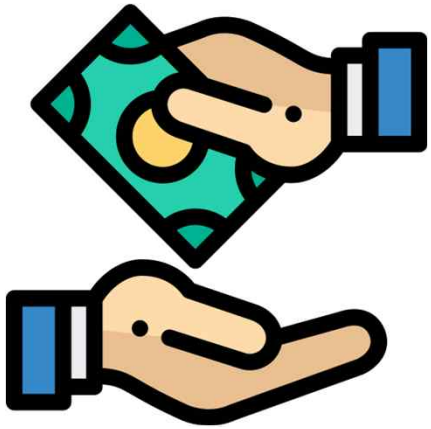
SNS 홍보가 아닌 구매량을 늘릴 수 있는
다른 마케팅 전략이 필요

6. 결과 활용방안 및 제안
.2 장바구니 전략

방안 및 기대효과

장바구니에 상품을 담아둔
잠재고객에게!

날씨의 영향으로
잘 소비되지 않는
제품에 대한 할인 쿠폰을 배포.



고민하고 있던 소비자들의
구매를 이끌어 낼 수 있음.
(효율적인 구매전환)

날씨로 인해 구매량이 저조할 것으로 예상되는 상품의 구매 증진을 위한 효과적인 해결 방안.

참고문헌

1. WILMS, Ines, et al. Interpretable vector autoregressions with exogenous time series. ***arXiv:1711.03623***, 2017.

2. SEPTIANI, Ayu; SUMERTAJAYA, I. Made; AIDI, Muhammad Nur. Vector Autoregressive X (VARX) Modeling for Indonesian Macroeconomic Indicators and Handling Different Time Variations with Cubic Spline Interpolation. ***Repositories-Dept. of Statistics, IPB University***, 2019, 175-180.

3. 김기호, et al. 통화정책 효과의 지역적 차이에 대한 분석. *보험금융연구*, 2015, 26.4: 3-37.

4. 한재윤, et al. 기계학습과 롤링 윈도우 기법을 활용한 주식시장 및 환율 예측 모델 구현. *한국지능정보시스템학회 학술대회논문집*, 2017, 69-70.

END