포스트 코로나 (Post COVID19)시대, 20대의 소비 생활

-최종 보고서-

제출 문서	다변량 통계학 학기말 과제
담당 교수	김규성 교수님
제출 기한	2020.12.14
실 체출일	2020.12.12
학과/학년	통계학과 4학년
학번	2014580002
이름	김광률

----------차례----------

1. 서론

1.1 연구 목적

2019년 말부터 현재 2020년까지 COVID19 바이러스(이하 코로나)는 세계인들의 생활양식을 바꾸어 놓았다. 세계는 현재 코로나 이전과 코로나 이후(포스트 코로나시대)로 나누어져서 구분 된다고 해도 과언이 아닐정도로 코로나가 가져온 변화는 극심하다.

코로나의 많은 문제들 중에서도 가장 큰 문제점 중 하나는 바로 경제활동의 변화이다. 사람들이 밖에 나가는 것을 꺼리며, 대면과 만남에 대한 두려움과 함께 각각의 소비패턴이 저하되거나 비대면의 식으로 바뀌어 나가는 사회 현상을 목격 할 수 있었다.

대한민국 또한 코로나에 대해 다양한 정책들과 예방책들을 내놓았으며 그에 따른 사람들의 소비변화가 나타났다. 그 중에서도 우리는 20대의 소비행동에 집중해 보기로 하였다.

20대의 경우 다양한 계층(학생, 직장인, 무직, 사업가, 군인 등)을 가짐으로써 개개인의 소비행동이 특정 나이대보다 개성이 있을 가능성이 있다고 판단하였고, 무엇보다도, 외부로의 사회적 만남(대면)이 그 어떠한 나이대보다 활발한 이들이 소비행동과 활동들이 포스트코로나사회에서 어떠한 특징을 가지는지 찾고자 한다.

나아가 분석의 결과를 통해 여러 인사이트를 찾아 사회현상과 연결시켜서 해석해 보고자 한다.

1.2 문헌 연구

<포스트코로나 시대에서 발견한 소비현상과 떠오르는 물음, 그리고 그것의 답>
https://dacon.io/competitions/official/235618/codeshare/1448?page=2&dtype=vote&ptype=pub
<포스트 코로나 소비와 흥미의 변화는??>
https://dacon.io/competitions/official/235618/codeshare/1419?page=1&dtype=vote&ptype=pub

1.3 데이터 설명

1.3-1(실질적 사용 데이터)

실질적 분석에 사용된 데이터는 2개(Time.csv, index.csv)이다.

그 중에서도 실질적으로 분석을 위해 사용한 변수들에 대한 설명이다.

date	confirmed
날짜(년-월-일) 2020-01-20 ~ 2020-06-30	확진자 수

데이터 파일크기	행 X 열 개수
7KB	164 X 7

period	catm	age	cgi (중요)
기간(연-월) 2019-01 ~ 2020-05	상품 소분류	나이대	카테고리 성장지수

데이터 파일크기	행 X 열 개수
8709KB	127526 X 8

*cgi: 2018년 월평균 대비 매출 성장 비율, 100을 기준으로 이상이면 매출 상승, 이하면 하락을 의미한다.

아래는 데이터의 일부를 발췌한 것이다.

date	time	test	negative	confirmed	released	deceased
2020-01-20	16	1	0	1	0	0
2020-01-21	16	1	0	1	0	0
2020-01-22	16	4	3	1	0	0
2020-01-23	16	22	21	1	0	0
2020-01-24	16	27	25	2	0	0
2020-01-25	16	27	25	2	0	0
2020-01-26	16	51	47	3	0	0
2020-01-27	16	61	56	4	0	0
2020-01-28	16	116	97	4	0	0
2020-01-29	16	187	155	4	0	0
2020-01-30	16	246	199	6	0	0
2020-01-31	16	312	245	11	0	0
2020-02-01	16	371	289	12	0	0
2020-02-02	16	429	327	15	0	0
2020-02-03	16	490	414	15	0	0
2020-02-04	16	607	462	16	0	0

	A	B	C	D	E	F	G	H
1	period	catl	catm	age	gender	sido	sigungu	cgi
2	201901	건강/의료	건강관리용	20	F	서울	관악구	115.3746
3	201901	건강/의료	건강관리용	20	F	서울	광진구	119.5965
4	201901	건강/의료	건강관리용	20	F	서울	도봉구	156.9928
5	201901	건강/의료	건강관리용	20	F	서울	동작구	58.34273
6	201901	건강/의료	건강관리용	20	F	서울	마포구	145.1476
7	201901	건강/의료	건강관리용	20	F	서울	성북구	78.89383
8	201901	건강/의료	건강관리용	20	F	서울	용산구	14.13718
9	201901	건강/의료	건강관리용	20	M	서울	all	102.171
10	201901	건강/의료	건강관리용	20	M	서울	강동구	125.895
11	201901	건강/의료	건강관리용	20	M	서울	강서구	97.52659
12	201901	건강/의료	건강관리용	20	M	서울	광진구	89.5786
13	201901	건강/의료	건강관리용	20	M	서울	금천구	143.7209
14	201901	건강/의료	건강관리용	20	M	서울	동대문구	104.3842

1.3-2(추가 사용 가능 데이터)

아래는 간단하게만 사용되거나 사용되지 않은 데이터들에 대한 설명이다.

이 보고서에서는 중점적으로 다루어지지 않지만 추후 추가적인 분석을 하기에 좋아 보여서 소개하고자 한다.

추가사용 가능 데이터

card_20200717.csv (2020.01.04 ~ 2020.03.05 개인별 카드 사용 내역 데이터)

receipt_dttm	adstrd_code	adstrd_nm	mrhst_induty_cl_code	mrhst_induty_cl_nm	selng_cascnt	salamt
카드사용날짜	행정동코드	행정동이름	업종코드	업종명	매출건수	매출금액

delivery.csv(배달관련 데이터, 2020.01.01 ~ 2020.02.08, 다수의 칼럼을 보유하고 있다.)

TimeAge.csv (코로나 확진자 일자별, 나이대별 데이터. 2020.03.02 ~ 2020.06.30)

date	time	age	confirmed	deceased
날짜	시간	나이대	확진자수	사망자수

PatientInfo_20200717.csv (코로나 확진자 개인별 정보)

patient_id	sex	age	country	province	city				
환자식별번호	성별	나이대	국가	시/도	시/군/구				
infection_case	infected_by	contact_number	symptom_onset_date	confirmed_date	released_date	deceased_date	state		
감염경로	감염자	접촉사람수	증상발현날짜	확진날짜	격리해제날짜	사망날짜	현상황(격리해제, 사망, 격리)		
infection_case									

Policy.csv (정부정책 및 사건)

policy_id	country	type	gov_policy	detail	start_date	end_date
사건번호	국가(Korea로 통일)	사건별대분류	사건별소분류	구체적사건명	사건시작일	사건종료일

Key 데이터

fpopl.csv(지역별로 성별, 연령별 인구수)

base_ymd	tmzon_se_code	sexdstn_se_code	agrde_se_code	adstrd_code	popltn_cascnt
기준날짜	기준시간	성별	연령대별	행정동코드별	인구수

adstrd_master.csv(지역명, 지역코드)

adstrd_code	adstrd_nm	brtc_nm	signgu_nm
행정동코드	행정동명	시/도 명	시/군/구 명

1.4 분석 방법

(1)

데이터를 축소하겠다.(지역의 한정)

인구밀집도가 높고, 분석의 타깃층으로 하는 20대가 많이 몰려있으며, 경제활동도 가장 활발하고, 데이터가 많은 '서울'의 데이터만을 중심으로 분석하고자 한다.

(2)

포스트 코로나시대를 정의하겠다. (코로나 시대 이전 vs 코로나 시대 이후)

일별 확진자수 데이터(Time.csv)를 이용하여 '탐색적데이터분석(EDA)'를 실시하고 기준을 잡아서 포스트코로나시대의 정확한 정의(시기)를 내리도록 하겠다.

(3)

포스트 코로나 이전과 이후의 연령별 소비 패턴의 변화 분석을 하겠다.

먼저 시각화를 통해서 연령별로 소비량(cgi)의 월별 변화 추이를 관찰하고,

'다변량 분산분석(MANOVA)'을 통해 연령별 (코로나 이전 vs 코로나 이후)의 소비량의 변화가 있었는지(통계적으로 유의미하게 변하였다.) 분석해보도록 한다.

(4)

20대들의 cgi변화(코로나 전-중-후)를 카테고리별로 정리한 이후, 카테고리별로 't검정'을 실시한다. 이후 소비패턴이 변화된 카테고리 vs 변화되지 않은 카테고리를 분류한다.

추가적으로 20대와 다른연령(30대~60대, 이하 비 20대)과의 비교를 위해 비 20대의 데이터도 위와 같은 방법으로 분석을 진행하도록 한다.

(5)

(4)의 결과에서 변화된 카테고리와 변화되지 않은 카테고리를 바탕으로 20대vs비20대의 소비패턴의 차이를 통해서 인사이트를 도출한다.

(6)

간단하게 '상관계수를 분석'을 통해 (5)의 결과와 관련하여 카테고리간의 관계를 살펴보도록 한다.

(7)

연령별(20대 vs 비20대)로 시기별(코로나 전-중-후)로 카테고리들의 cgi의 변화된 값들을 통해서 카테고리별 '군집분석'을 하도록 한다.

추가적으로 (4)의 결과와 비교도 해보도록 한다. 이후 인사이트를 도출하도록 한다.

1.5 결과 활용 및 기대 효과

(1)
'분석방법-(2)'에서 EDA를 실시한 결과 특정 시기에 확진자의 증가율이 급격하게 늘어난 날짜(특정사건으로 인하여: 이태원클럽, 신천지 등)가 있었다. 추후 나온 정부 정책들과 일별확진자수의 시계열 그래프를 통하여, 포스트코로나시기를 결정할 수 있다.

(2)
'분석방법–(3)'에서 비20대의 경우에는 소비량(cgi)이 통계적으로 유의미하게 변했다는 결과가 나왔다. 그러나 20대에서는 코로나에도 불구하고 소비량(cgi)의 변화가 유의미하게 변하지 않았다.

(3)
'분석방법-(3),(4)'의 분석을 통해서 코로나 이전과 이후로 비교하였을 때, 연령별(20대, 비 20대) 소비패턴에서 어떠한 품목들이 가장 크게 변동하였는지 알 수 있을 것이라고 생각한다. 또한 연령별(20대 vs 비20대)로 소비량의 변화가 어떤 카테고리별로 차이가 있었는지 알 수 있다.

(4)
'분석방법-(6)'에서 카테고리 품목별 상관계수값과 이를 시각화한 그래프를 통해서 카테고리 품목별로의 관계성에 대해서 알아볼 수 있다.

(5)
'분석방법-(7)'을 통해서 연령별(20대, 비20대) 소비량(cgi)가 변화된(혹은 변화되지 않은) 카테고리들을 군집으로 나누어 분석함으로서 군집(카테고리)간의 특징을 찾아볼 수 있다.

전반적인 기대효과로는 20대들의 지금껏 코로나에 대처한 행동들을 소비량의 변화를 통해 유추해볼 수 있고, 추후 지속되는 코로나시대에 그들이 임해야하는 자세와 태도를 일깨워 줄 수 있을 것이다.

2. 본론

2.1 분석방법 소개

사용 프로그램

기본적인 분석과정은 서론의 분석방법과 같이 진행 될 것이며, 여기서는 분석을 위해서 사용된 패키지, 코드, 함수에 관한것들을 간단하게만 설명하도록 하겠다.

(1) 전처리

Excel: 필터 기능을 사용하였다.

R(패키지): dplyr, readr 등을 사용하였다.

(2) 통계분석

R(함수): manova, t.test, cor 등을 사용하였다.

SAS: PROC GLM, PROC CLUSTER 등을 사용하였다.

(3) 시각화

PowerPoint, Excel: 도표만들기를 사용하였다.

R(패키지): ggplot2, RColorBrewer, corrplot 등을 사용하였다.

SAS: PROC GPLOT, PROC TREE 등을 사용하였다.

2.2 데이터 분석 및 결과 설명

(※분석의 타당성 설명은 결과 설명과 함께 진행 하겠다.)

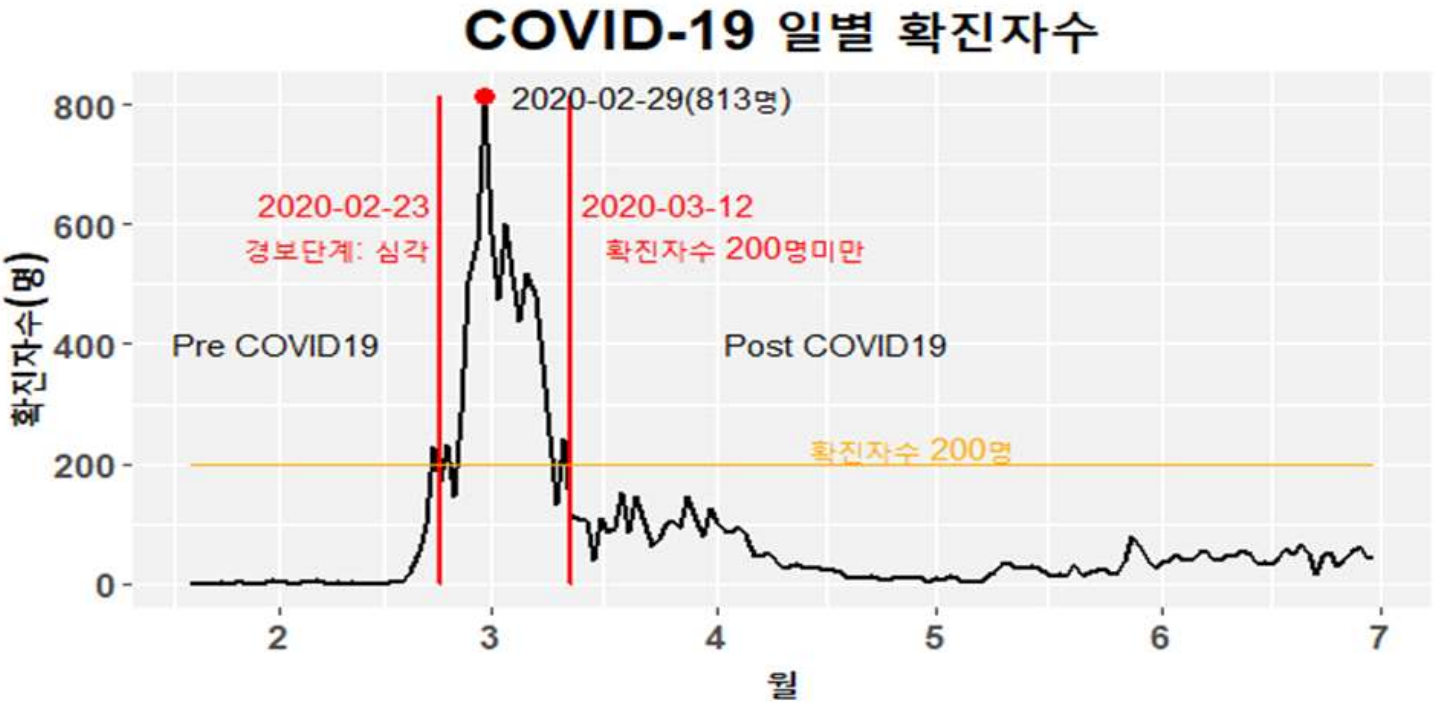
(1) 포스트 코로나기간(Period)의 정의

앞으로의 분석을 위해 시기를 크게 3개

1. 코로나이전(Pre-COVID19)
2. 코로나중(Pre-COVID19)
3. 코로나이후(Post-COVID19)

로 나누도록 하겠다.

아래의 그래프는 한국의 코로나19 일별 확진자 수를 나타낸 것이다.



<그림1>

보는 바와 같이 2월말쯤부터 확진자 수가 급격하게 늘어나기 시작하여, 2020-02-29에 813명으로 최대 확진자 수를 기록하였으며, 이후 확진자수가 조금씩 감소하는 추세를 보인다.

Corona Issue	Policy	Start Date	End Date
	감염병 위기 경보 '관심'	2020-01-03	2020-01-19
국내 첫 확진자 발생	감염병 위기 경보 '주의' 로 격상	2020-01-20	2020-01-27
	감염병 위기 경보 '경계' 로 격상	2020-01-28	2020-02-22
31번째 코로나 확진 (대구 신천지)		2020-02-18	
	감염병 위기 경보 '심각' 로 격상	2020-02-23	
	사회적 거리두기 실시	2020-02-29	2020-03-21
	전국 학교 등교 연기	2020-03-02	2020-04-06
	강화된 사회적 거리두기 실시	2020-03-22	2020-04-19
일일 코로나 발생 10명 이내		2020-04-18	2020-05-09
	약회된 사회적 거리두기 실시	2020-04-20	2020-05-05
용인(66번째) 확진 (이태원 클럽)		2020-05-06	
	생활속 사회적 거리두기 실시	2020-05-06	
	전국 유흥주점 운영 자제	2020-05-08	2020-06-07
	전국 노래방 운영 자제	2020-05-21	2020-06-03
	대중교통 마스크 착용 의무화	2020-05-26	

<그림2>

위의 표는 코로나19사태에 대비한 정부의 주요 지침을 일자별로 가져와 정리한 표이다.(pclicy.csv데이터 이용) 앞에서의 코로나 일별 확진자수 표<그림1>와 함께 참고할 수 있다.

위의 그림들을 바탕으로 우리는 앞으로 코로나시기를 아래와 같이 구분하여 설명하고자 한다. 일별 기간을 아래와 같이 잡은 이유는 <그림1>에서의 확진자 200명선과 <그림2>에서 2020-02-23일에 감염병 위기 경보 '심각'으로 격상한 것을 종합적으로 참조하여 정하였다. 이를 바탕으로 앞으로의 분석을 위하여 월별 기준은 아래와 같이 잡도록 하겠다.

기간구분(일별)

Pre COVID19 : ~ 2020-02-22)

Ing COVID19 : 2020-02-23 ~ 2020-03-12

Post COVID19 : 2020-03-12 ~

기간구분(월별)

Pre COVID19 : ~2020년 1월)

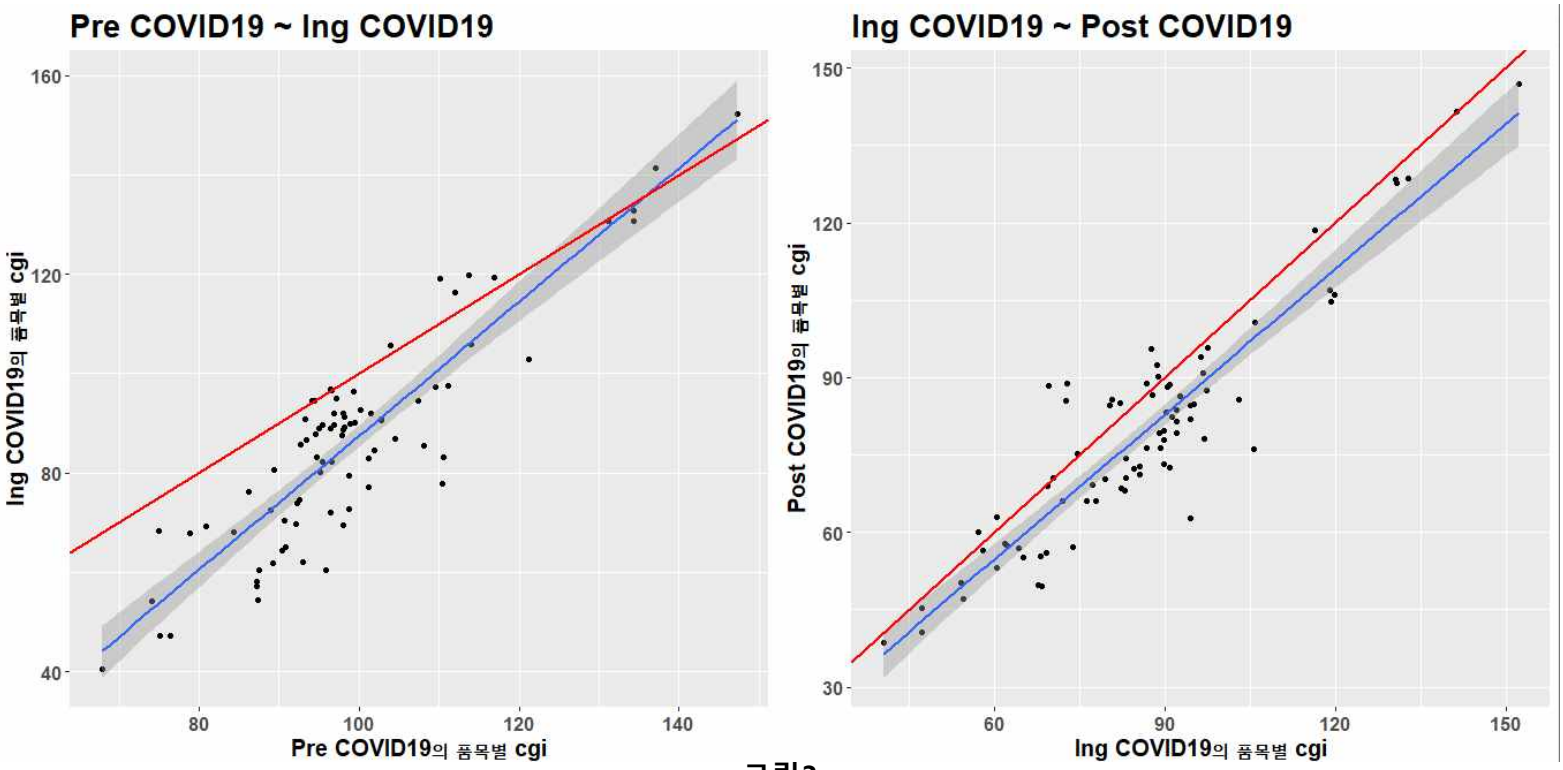
Ing COVID19 : 2020년 2월, 2020년 3월

Post COVID19 : 2020년 3월 ~

(2) 코로나로 인한 cgi의 변화

아래 그림은 코로나 전~중 / 중~후 로 나누어서 품목별 cgi의 산점도를 그려보고 그 추세선(파랑선)을 그린 것이다.

빨간선은 y = x그래프이며 즉 cgi값의 변화가 없다는 가정하에 파란선과 일치해야 하는 선이다.



<그림3>

결과를 보아 알 수 있겠지만, 코로나 전~중(왼쪽 그림)의 경우에는 그 추세선이 일치하지 않아 많은 cgi의 변화가 일어났음을 짐작할 수 있고, 중~후(오른쪽그림)의 경우에는 어느정도 추세선을 따르는 모습을 볼 수 있었으나, 전반적으로 그래프가 y축 아랫방향으로 평행이동한 것을 보아 전체적인 cgi양은 준 것을 어느정도 짐작 가능했다.

(3) 20대 vs 비20대 cgi변화의 MANOVA분석

아래는 (2)에서 분석에 사용하기 위해 전처리한 데이터들의 일부를 발췌한 것이다. 데이터는 연령별, 카테고리별로 코로나 전-중-후 시기의 cgi값의 평균을 계산한 것이다.

	catm	age	pre_cgi	ing_cgi	post_cgi
1	가공식품	20	116.90565	119.23615	104.65071
2	건강관리용품	20	114.00887	105.89657	100.69218
3	기호식품	20	110.36770	77.92078	66.05743
4	담배	20	147.24073	152.22621	147.00256
5	바디/헤어용품	20	108.08911	85.58290	72.72526
6	뷰티소품	20	74.13209	54.15991	50.24942
7	빙과류	20	102.75329	90.51574	88.10114
8	빵류	20	121.23388	102.94132	85.65912
9	생활용품	20	109.56111	97.22850	87.47457
10	신선식품	20	113.77731	119.74410	105.99193
11	애완동물용품	20	98.74781	79.40516	70.30237
12	유제품류	20	110.16759	119.00901	106.88848
13	음료	20	111.14429	97.53343	95.73516
14	제과류	20	103.91464	105.73610	76.11721
15	주류	20	112.04653	116.30847	118.56423
16	화장품	20	107.44231	94.43538	62.74212
17	가공식품	30	101.50393	92.02854	79.20329
18	건강관리용품	30	86.22323	76.19063	66.03792
19	기호식품	30	93.01132	62.16103	57.30047
20	담배	30	134.27280	132.82303	128.62933
21	바디/헤어용품	30	87.39529	54.45211	47.19298
22	뷰티소품	30	67.83461	40.44650	38.64572

<all_cgi>: 전체연령

	catm	age	pre_cgi	ing_cgi	post_cgi
1	가공식품	30	101.50393	92.02854	79.20329
2	건강관리용품	30	86.22323	76.19063	66.03792
3	기호식품	30	93.01132	62.16103	57.30047
4	담배	30	134.27280	132.82303	128.62933
5	바디/헤어용품	30	87.39529	54.45211	47.19298
6	뷰티소품	30	67.83461	40.44650	38.64572
7	빙과류	30	94.52641	87.70131	86.61805
8	빵류	30	110.55631	83.22423	70.45448
9	생활용품	30	96.48575	71.97216	66.02699
10	신선식품	30	98.17880	89.14149	76.28763
11	애완동물용품	30	87.14937	57.07757	60.14193
12	유제품류	30	96.82915	89.76519	77.97219
13	음료	30	98.75643	72.77949	88.85690
14	제과류	30	98.94715	89.89276	73.23939
15	주류	30	94.98095	88.89251	90.15195
16	화장품	30	76.42774	47.16478	45.40699
17	가공식품	40	98.09629	91.25247	82.34817
18	건강관리용품	40	94.74154	83.11299	74.39504
19	기호식품	40	92.18753	73.78171	57.07300
20	담배	40	134.32315	130.69243	128.41279

<all_cgi_3060>: 비20대

(3)-1

H0: 연령대별(20대, 30대, 40대, 50대, 60대) cgi값(매출성장비율)이 기간별(코로나 전-중-후)로 비교해보았을 때, 차이가 유의하지 않다.

H1: not H0.

```
> M = manova(cbind(pre_cgi, ing_cgi, post_cgi)~age, data = all_cgi)
> summary(M, intercept = T, test = "Wilks")
            Df   Wilks approx F num Df den Df  Pr(>F)
(Intercept)  1 0.01174   2132.6      3     76 < 2e-16 ***
age          1 0.87560      3.6      3     76 0.01724 *
Residuals   78
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #연령별 유의미한 차이가 있음
>
> #사후분석(일변량 검정 통계량)
> summary.aov(M)
 Response pre_cgi :
            Df  Sum Sq Mean Sq F value  Pr(>F)
age          1   851.7  851.71  4.5296 0.03647 *
Residuals   78 14666.4  188.03
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response ing_cgi :
            Df Sum Sq Mean Sq F value Pr(>F)
age          1   1009 1008.53  2.2764 0.1354
Residuals   78  34557  443.04

 Response post_cgi :
            Df Sum Sq Mean Sq F value Pr(>F)
age          1    190  189.64  0.4015 0.5282
Residuals   78  36842  472.33
```

위의 분석 결과를 통해서 연령대별로 cgi값이 유의미한 차이(유의수준5%)가 있고, 그 유의미한 차이는 코로나이전의 cgi값에서 발생한 것임을 알 수 있었다.

(3)-2

H0: 비20대(30대, 40대, 50대, 60대) cgi값(매출성장비율)이 기간별(코로나 전-중-후)로 비교해보았을 때, 차이가 유의하지 않다.

H1: not H0.

```
> #비 20대 MANOVA분석 (연령별 ~ (코로나전, 중 , 후), 윌크스 람다 통계량 이용)
> M = manova(cbind(pre_cgi, ing_cgi, post_cgi)~age, data = all_cgi_3060)
> summary(M, intercept = T, test = "Wilks")
            Df   Wilks approx F num Df den Df Pr(>F)
(Intercept)  1 0.01039  1905.49      3     60 <2e-16 ***
age          1 0.95029     1.05      3     60 0.3788
Residuals   62
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

그런데 20대를 제외하고 분석을 해본 결과 이번에는 반대의 결과(H0를 기각하지 못한다.)가 나오는 것을 볼 수 있었다.

즉, 20대를 제외한 나머지 비20대의 cgi값은 비슷한 경향과 값을 보인다는 결과를 알 수 있었고, 따라서 유난히도 특이한 cgi값을 가지는 20대의 소비에 대해서 더 집중을 해보아야겠다는 생각을 할 수 있었다.

(3)-3

추가적으로 연령대별로 코로나 전~중, 중~후 의 cgi 변화(평균값의 차이가 유의미한지, 유의수준5%)를 알아보고자 t-test를 진행해보았다. 아래는 분석의 결과이다.

```
> #20대
> t.test(all_cgi_20$pre_cgi,all_cgi_20$ing_cgi) #변화x

        Welch Two Sample t-test

data:  all_cgi_20$pre_cgi and all_cgi_20$ing_cgi
t = 1.3487, df = 25.587, p-value = 0.1893
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.716393 22.673039
sample estimates:
mean of x mean of y
 110.0958  101.1175

> t.test(all_cgi_20$ing_cgi,all_cgi_20$post_cgi) #변화x

        Welch Two Sample t-test

data:  all_cgi_20$ing_cgi and all_cgi_20$post_cgi
t = 1.357, df = 29.826, p-value = 0.185
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.651354 28.017083
sample estimates:
mean of x mean of y
101.11748  89.93462
> #30대
> t.test(all_cgi_30$pre_cgi,all_cgi_30$ing_cgi) #변화o

        Welch Two Sample t-test

data:  all_cgi_30$pre_cgi and all_cgi_30$ing_cgi
t = 2.6945, df = 25.714, p-value = 0.01225
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.251782 31.668960
sample estimates:
mean of x mean of y
 95.19245  77.23208

> t.test(all_cgi_30$ing_cgi,all_cgi_30$post_cgi) #변화x

        Welch Two Sample t-test

data:  all_cgi_30$ing_cgi and all_cgi_30$post_cgi
t = 0.67144, df = 29.964, p-value = 0.5071
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.66158  21.10496
sample estimates:
mean of x mean of y
 77.23208  72.01039
```

```
> #40대
> t.test(all_cgi_40$pre_cgi,all_cgi_40$ing_cgi) #변화O

        Welch Two Sample t-test

data:  all_cgi_40$pre_cgi and all_cgi_40$ing_cgi
t = 2.3175, df = 26.687, p-value = 0.02839
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.499501 24.771304
sample estimates:
mean of x mean of y
 94.23392  81.09852

> t.test(all_cgi_40$ing_cgi,all_cgi_40$post_cgi) #변화X

        Welch Two Sample t-test

data:  all_cgi_40$ing_cgi and all_cgi_40$post_cgi
t = 1.1401, df = 29.798, p-value = 0.2633
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.211715 21.900202
sample estimates:
mean of x mean of y
 81.09852  73.25428
```

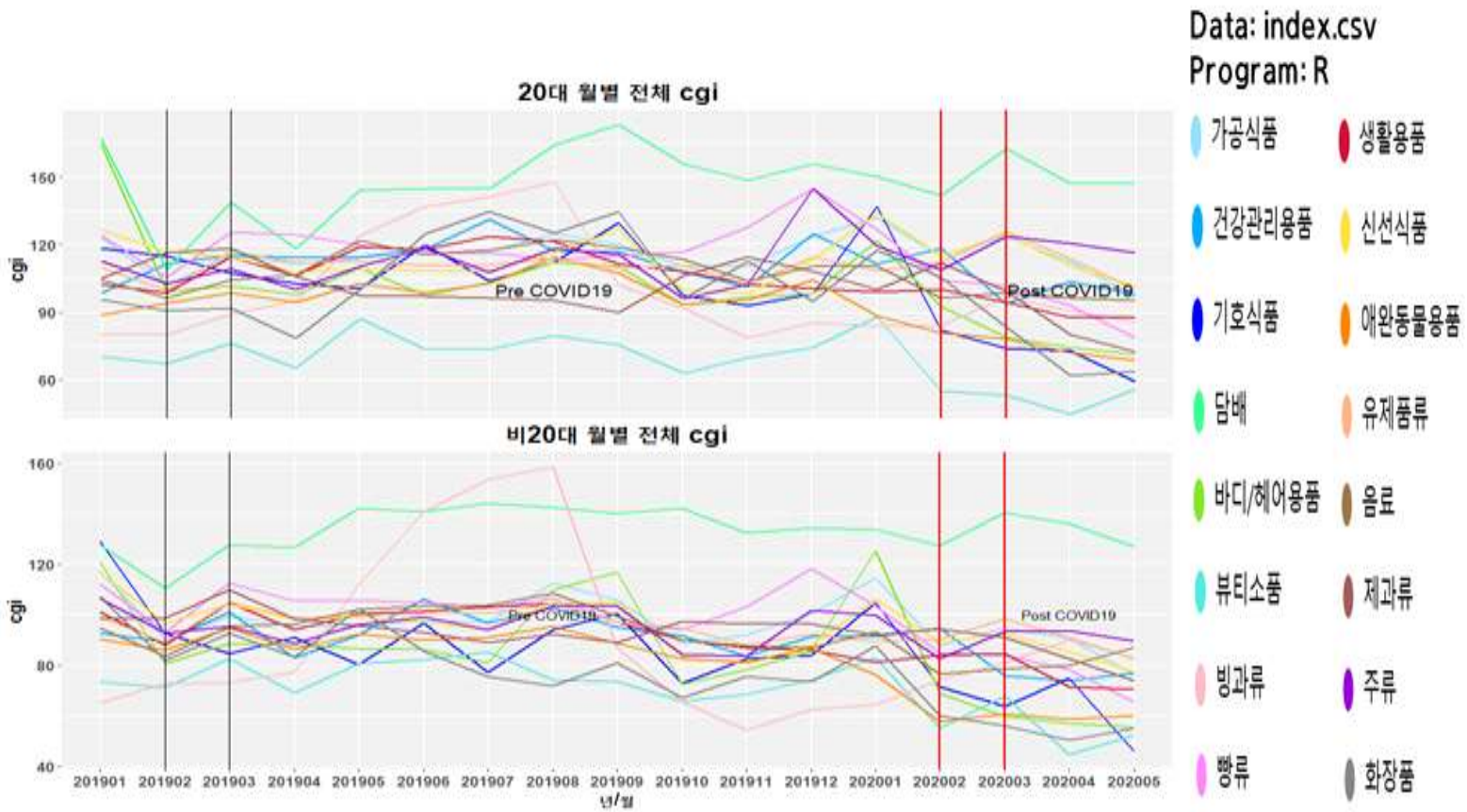
```
> #50대
> t.test(all_cgi_50$pre_cgi,all_cgi_50$ing_cgi) #변화O

        Welch Two Sample t-test

data:  all_cgi_50$pre_cgi and all_cgi_50$ing_cgi
t = 2.0216, df = 25.597, p-value = 0.05379
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.204875 23.542624
sample estimates:
mean of x mean of y
 95.76894  84.10006

> t.test(all_cgi_50$ing_cgi,all_cgi_50$post_cgi) #변화X

        Welch Two Sample t-test

data:  all_cgi_50$ing_cgi and all_cgi_50$post_cgi
t = 0.85972, df = 29.492, p-value = 0.3969
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.721709 21.387406
sample estimates:
mean of x mean of y
 84.10006  77.76722
```

```
> #60대
> t.test(all_cgi_60$pre_cgi,all_cgi_60$ing_cgi) #변화O

        Welch Two Sample t-test

data:  all_cgi_60$pre_cgi and all_cgi_60$ing_cgi
t = 2.6262, df = 25.136, p-value = 0.01449
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.838542 23.443958
sample estimates:
mean of x mean of y
 98.27155  85.13030

> t.test(all_cgi_60$ing_cgi,all_cgi_60$post_cgi) #변화X

        Welch Two Sample t-test

data:  all_cgi_60$ing_cgi and all_cgi_60$post_cgi
t = 0.58399, df = 29.999, p-value = 0.5636
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.783741 15.818797
sample estimates:
mean of x mean of y
 85.13030  81.61277
```

- 분석의 결과를 해석해보자면, 30~60세대는 코로나 전~중으로 즉각적인 cgi의 변화(일반적으로 감소)를 보였지만 20대는 그러하지 않았다.
- 전반적으로 코로나 중~후 에는 cgi의 유의미한 변화가 없었다.
- 그러나 p-value의 값의 크기를 바탕으로 20대의 cgi변화반응이 늦음(중~후 에서의 유의수준 값이 다른 세대에 비해서 가장 작기 때문에)이라고 해석도 해봄직하다.

(4) 20대 vs 비20대의 월X카테고리 cgi변화

그렇다면, 본격적으로 연령을 2개의 그룹으로 나누어 그 cgi값의 변화 추이를 살펴보자.



<그림4>

왼쪽의 검정세로선 2개는 2019년 2월~3월을 나타낸 선이고, 오른쪽의 빨간세로선 2개는 2020년 2월~3월(코로나 중)을 나타낸 선이다.
보이는 바와 같이 20대보다는 비20대의 경우 전반적인 품목들에서 cgi값이 2019년에 비해 2020년이 낮은 값을 보이는 경향을 볼 수 있었다.

(5) 카테고리별 코로나 전-중-후 t-test

아래는 총 16개의 카테고리별 코로나 전-중-후 cgi평균값의 유의미한 차이를 보기 위한 t-test의 결과이다. 카테고리(가공식품~화장품)가 너무 많기에 일부만 발췌하도록 하고 핵심내용은 아래에 표로 정리하여 해석하도록 하겠다.

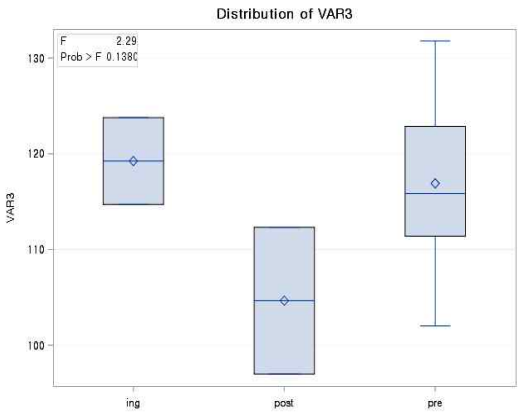
(5)-1

20대

가공식품

Dependent Variable: VAR3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	288.054420	144.027210	2.29	0.1380
Error	14	881.011585	62.929399		
Corrected Total	16	1169.066004			



.

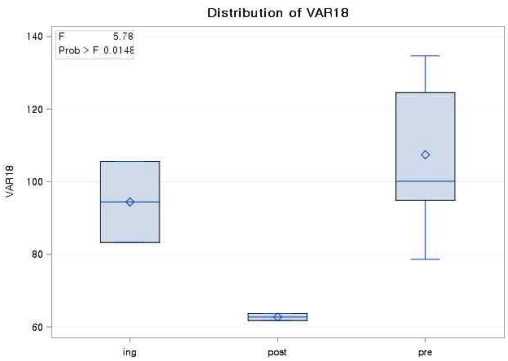
.

.

화장품

Dependent Variable: VAR18

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3551.018468	1775.509234	5.78	0.0148
Error	14	4302.621120	307.330080		
Corrected Total	16	7853.639588			

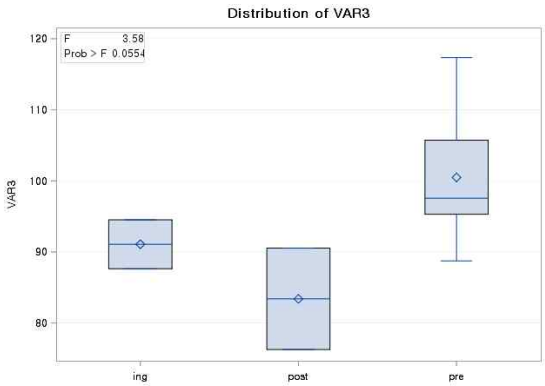


(5)-2

비20대

가공식품

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	595.538799	297.769399	3.58	0.0554
Error	14	1163.323742	83.094553		
Corrected Total	16	1758.862541			



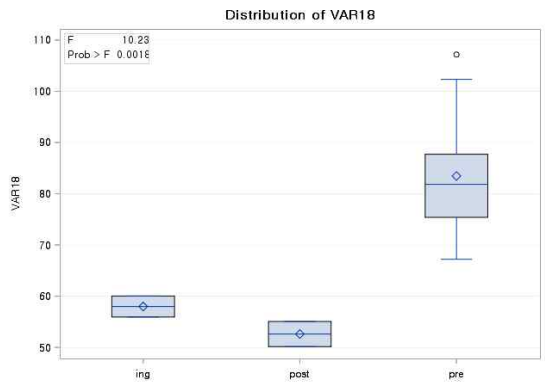
.

.

.

화장품

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2454.552690	1227.276345	10.23	0.0018
Error	14	1678.784652	119.913189		
Corrected Total	16	4133.337343			

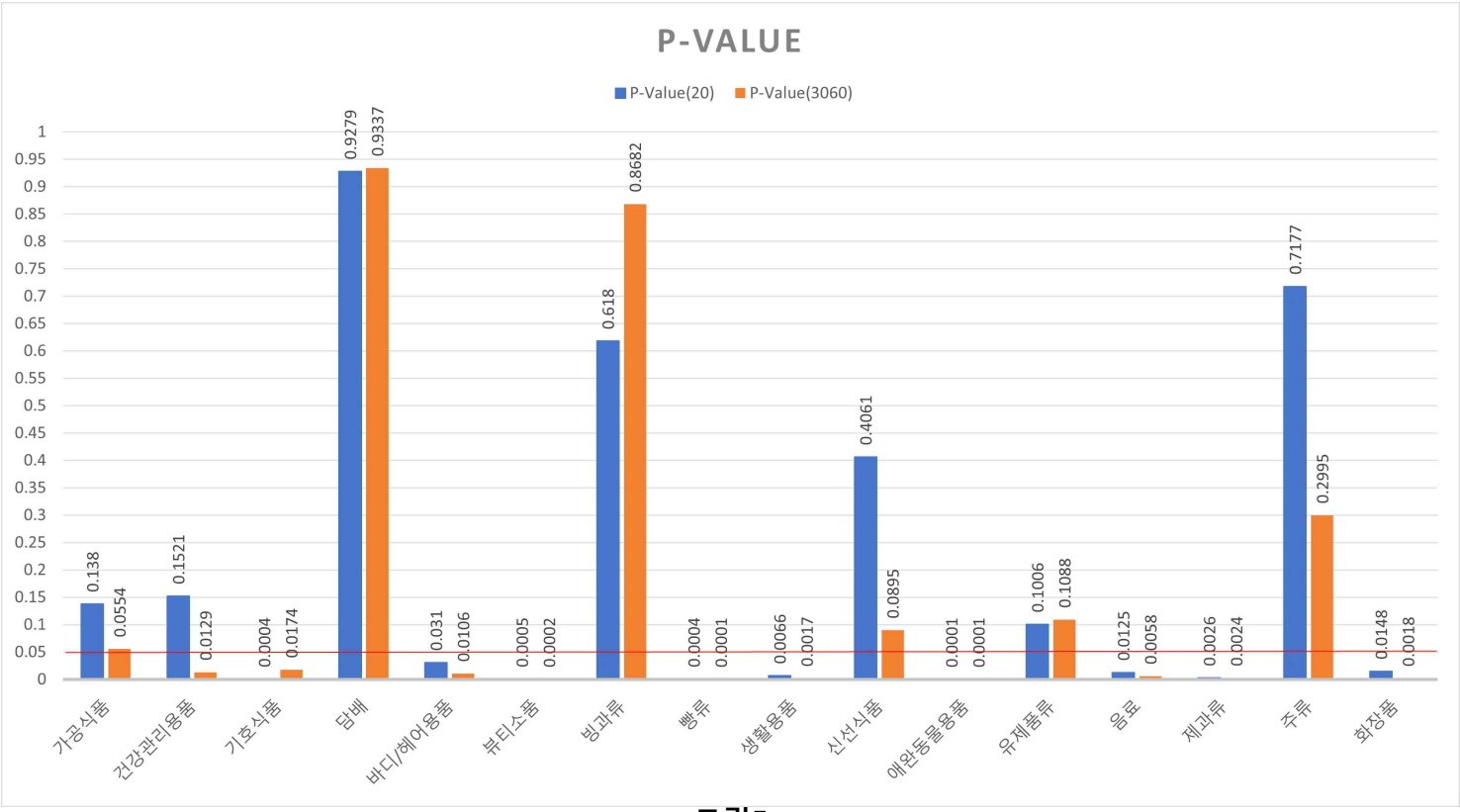


해석을 조금 덧붙이자면 p-value값(형광색으로 표시한부분)이 0.05(유의수준)보다 작은 경우 코로나 전-중-후 의 cgi평균값에 있어서 해당 카테고리항목의 변화가 유의미하게 있었다는 것을 의미한다.

오른쪽의 그림은 전-중-후의 cgi값을 나타내는 box-plot이다.

(5)-3

위의 (5)-1, (5)-2의 결과를 보기 쉽게 정리한 그래프이다.



<그림5>

카테고리별로 t-test를 한 결과의 p-value 값을 정리했으며, 파란색 막대가 20대 주황색 막대가 비20대를 의미한다.

비슷한 결과를 보이는(비슷한 소비변화(cgi)를 보였다는 뜻) 카테고리들도 있었고, 그렇지 못한 카테고리들도 있었다. 위 그래프를 다시 한번 정리해보자.

(5)-4

(5)-3에서 20대vs비20대 의 p-value값을 바탕으로 cgi의 변화유무를 카테고리별로 정리한 표이다.

추가적으로 p-value값을 바탕으로 20대vs비20대의 상대비율을 구하여 정리하여 그 값이 클수록 빨간색, 작을수록 파란색이 되도록 표를 만들었다.

즉, 빨간색일수록 20대와 비20대의 소비경향성과 값의 차이가 큰 카테고리, 파란색일수록 20대와 비20대의 소비경향성과 값의 차이가 작은 카테고리라고 해석 할 수 있다.

품목(전체)	20대	비 20대	p-vaule(20)	p-value(non-20)	p-value(%)
가공식품	X	X	0.138	0.0554	0.71354705
건강관리용품	X	O	0.1521	0.0129	0.92181818
기호식품	O	O	0.0004	0.0174	0.02247191
담배	X	X	0.9279	0.9337	0.4984422
바디/헤어용품	O	O	0.031	0.0106	0.74519231
뷰티소품	O	O	0.0005	0.0002	0.71428571
빙과류	X	X	0.618	0.8682	0.4158256
빵류	O	O	0.0004	0.0001	0.8
생활용품	O	O	0.0066	0.0017	0.79518072
신선식품	X	X	0.4061	0.0895	0.81941082
애완동물용품	O	O	0.0001	0.0001	0.5
유제품	X	X	0.1006	0.1088	0.48042025
음료	O	O	0.0125	0.0058	0.68306011
제과류	O	O	0.0026	0.0024	0.52
주류	X	X	0.7177	0.2995	0.70556429
화장품	O	O	0.0148	0.0018	0.89156627

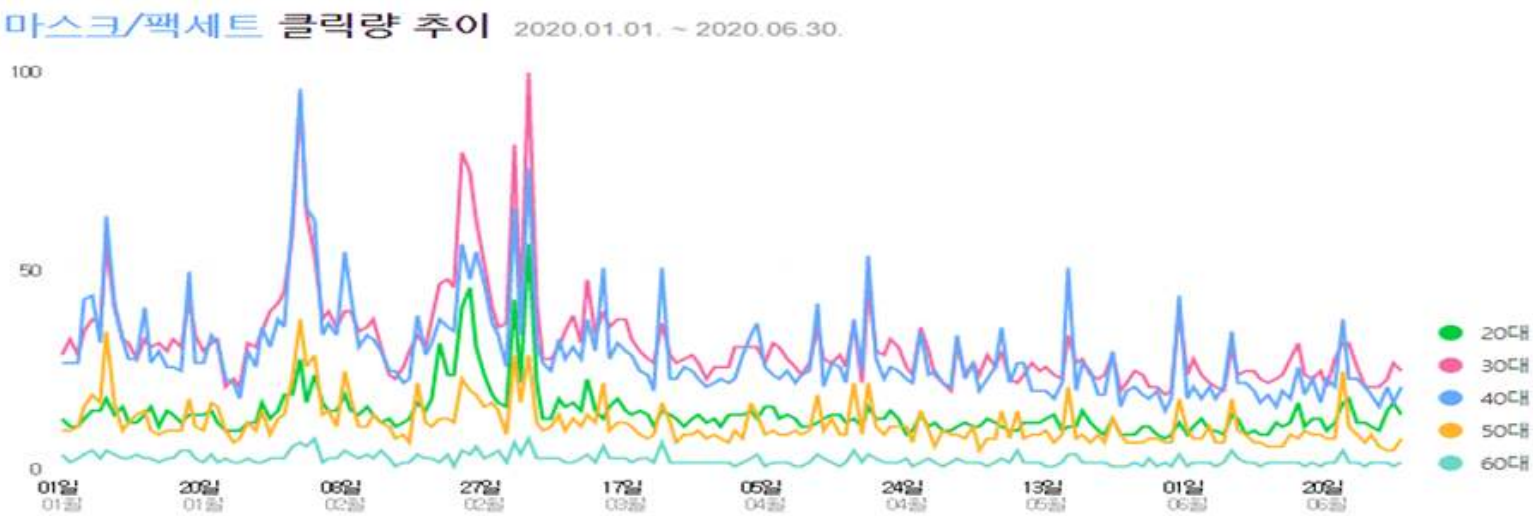
<그림6>

위의 표에서 가장 돋보이는 카테고리는 단연 '건강관리용품'이다. 다른 카테고리들의 경우에는 p-value값이 차이를 보이더라도 유의수준(0.05)을 기준으로 p-value값이 갈리는 경우(즉, cgi변화 차이가 유의미하다/그렇지않다)는 없었는데, 그 정도로 가장 큰 차이를 보인 경우는 '건강관리용품'이 유일했다.

(6) 인사이트 1

위 (5)의 결과를 바탕으로 우리는 20대가 비20대에 비하여 '건강관리용품'의 변화를 보이지 않았다는 사실을 알았다. 그렇다면 코로나로 인하여 가장 급격하게 소비량이 변화된 '건강관리용품'은 무엇이 있을까?

직관적으로 이는 '마스크'라는 상품을 떠올려 볼 수 있다. 실제로 분석을 대상으로 한 시기의 코로나 초창기만 하더라도, 저 당시 젊은층은 코로나 바이러스로 인하여 건강적으로 큰 피해가 없다는 생각들이 있었다. 또한 실제로 저 시기에 마스크를 쓰지 않고 일상생활을 하는 20대들도 꽤 있었다. 그렇다면 이러한 추측이 사실일지 데이터를 바탕으로 확인해보도록 하자.



<그림7>

위 표는 우리가 분석한 시기(2020-01~2020-06) 20대와(초록선) 비20대들의 마스크/팩세트의 검색 클릭량 추이(네이버기준)이다. 50/60대의 경우 다른세대에 비해 상대적으로 인터넷의존도가 높지 않은 세대이기 때문에 낮은 수치를 보인다고 해석가능하지만, 실제로 인터넷을 가장 활발하게 이용하는 20대의 검색량이 다른 세대들에 비해 떨어지는 경향성이 뚜렷한 것은 우리가 초반에 추측했던 것이 사실이라고 해석이 어느정도 가능하다.

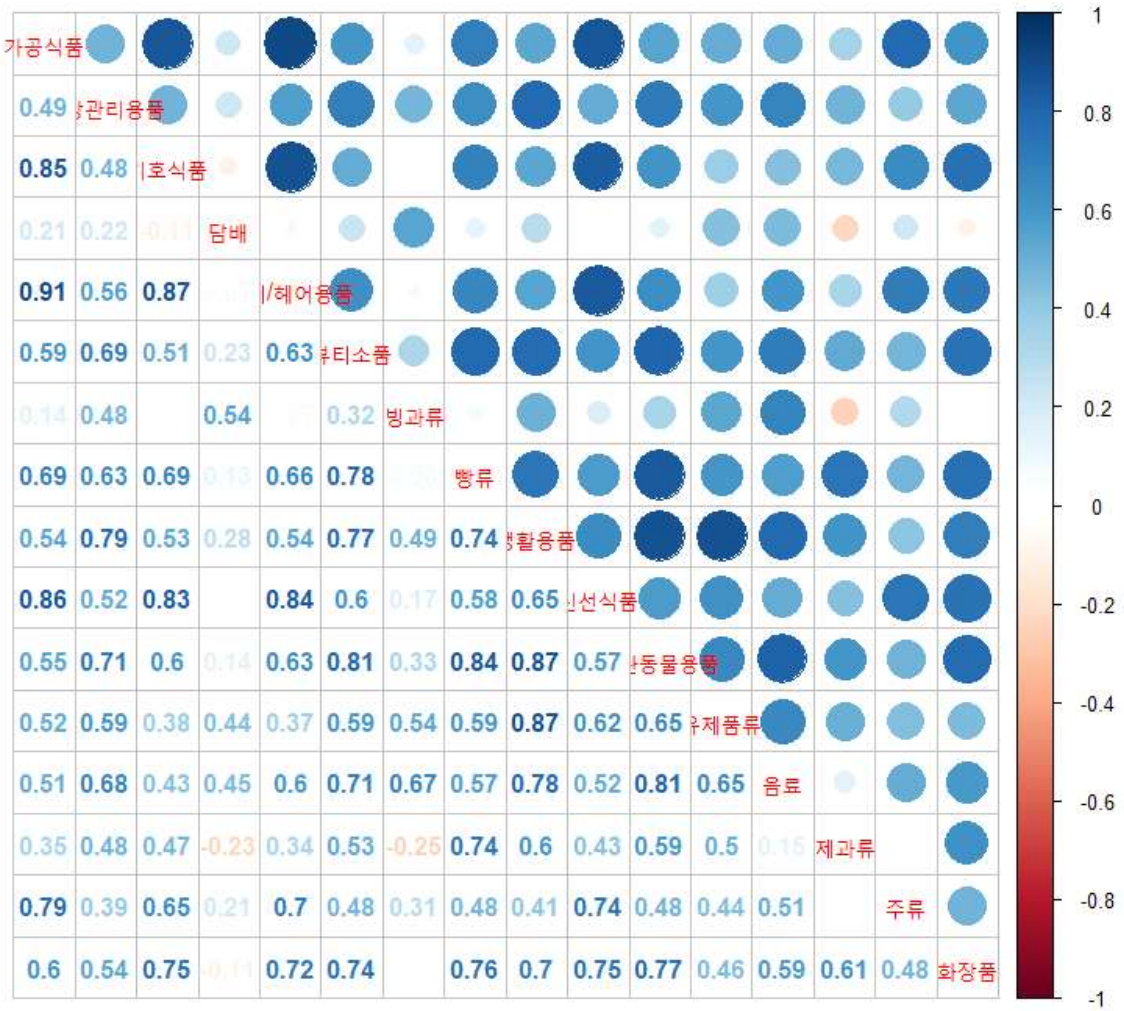
추가적으로 조사해본 결과 전체 검색량 사용자의 경우에는 오히려 20대가 가장 많았고, 네이버의 경우 다른 포털사이트보다 20대의 분포가 훨씬 뛰어난 사이트임을 고려할 때 위와 같은 결과는 더더욱 특별한 결과라고 볼 수 있다.

즉, 코로나 시기에도 불구하고, 20대는 다른 연령에 비하여 마스크에 대한 관심도가 떨어졌으며, 이는 코로나에 대해 20대가 타 세대에 비하여 상대적으로 안일한 생각과 대처를 시사한다는 인사이트를 얻을 수 있었다.

(7) 카테고리별 상관관계 분석(20대 vs 비20대)

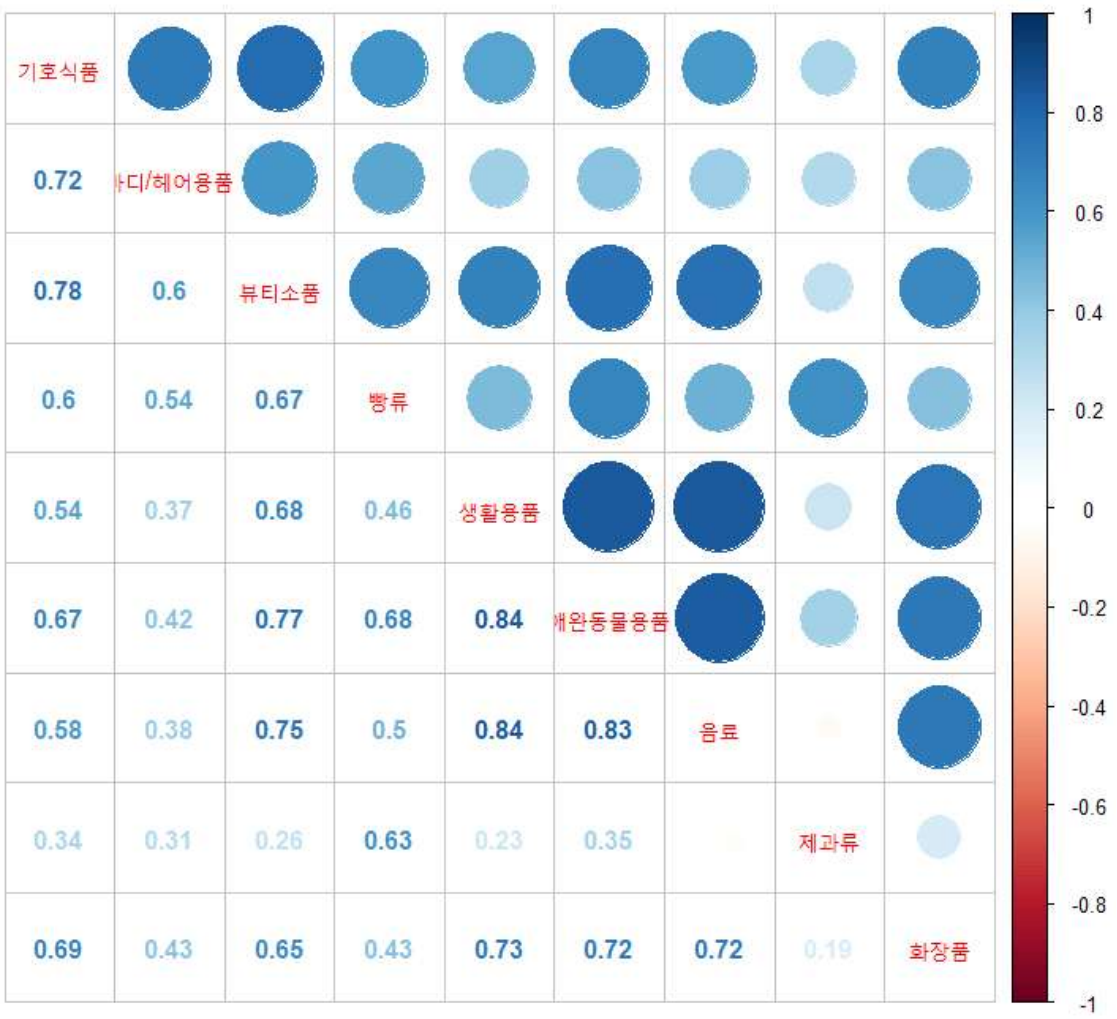
(7)-1 20대

아래는 20대 전체 16개 카테고리들의 월별 cgi값들을 기준으로 구한 상관계수 값과 이를 시각적으로 나타낸 그래프이다.



<그림8>

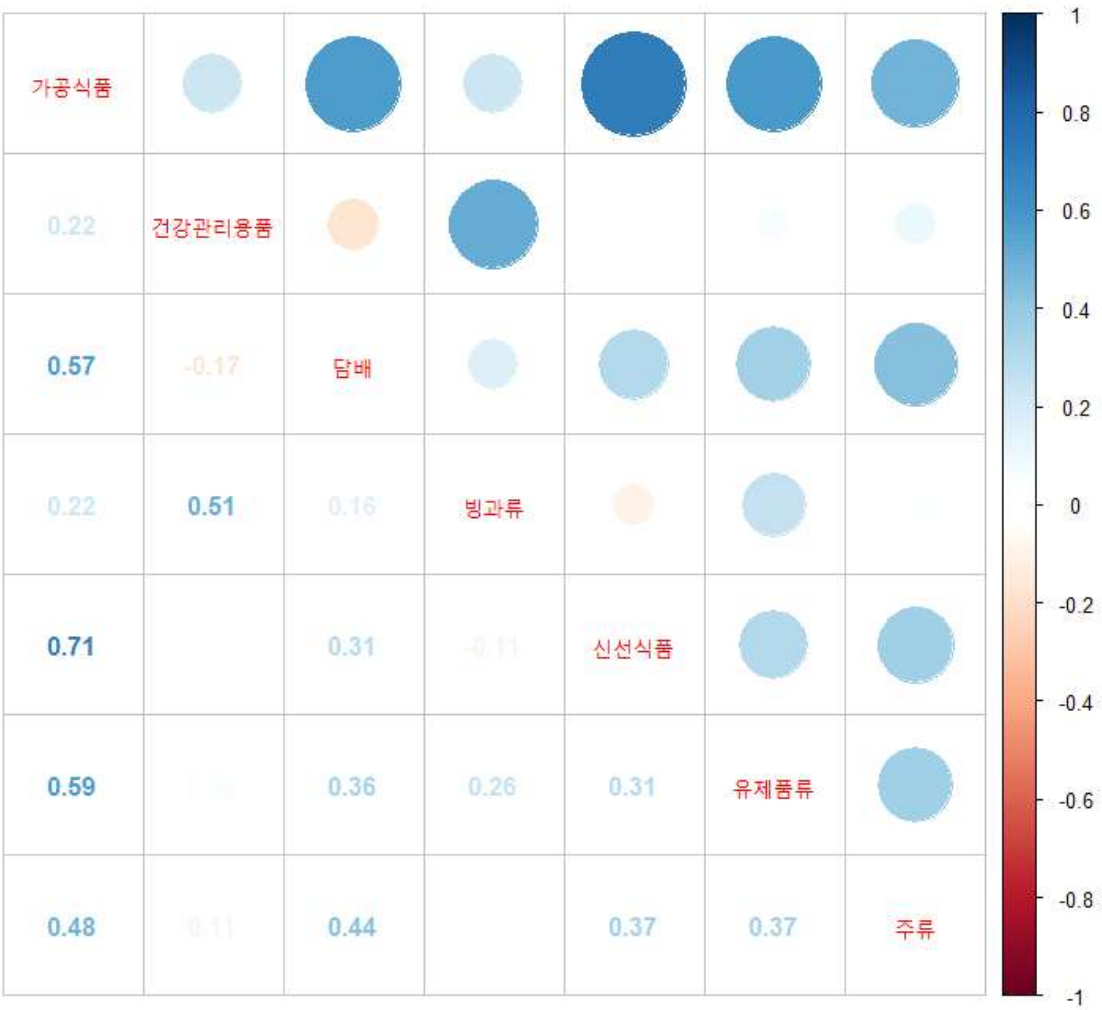
아래는 전체 20대의 카테고리들 중 '5-(4)의 <그림6>'의 결과를 바탕으로 코로나 전-중-후로 cgi가 유의미하게 '변화된' 카테고리들의 상관계수 값과 이를 시각적으로 나타낸 그래프이다.



<그림9>

변한 카테고리들의 공통점을 찾아보자면, 일상생활에 필수적으로 필요하지 않은 품목들인 경우가 많았다. 빵류, 음료, 제과류 의 경우는 카페가 문을 닫거나 축소 운영을 하게되면서 자연스럽게 그 소비가 감소했다고 해석 가능하다. 재미있는 점은 바디/헤어용품, 뷰티소품, 화장품 인데, 이 품목들의 공통점은 주로 밖으로 나가기 위하여 준비하는 과정에서 필요한 물품들이다. 밖으로 나가는 일이 많이 줄었기 때문에 이 품목의 소비가 줄었다고도 해석 가능하고 또 다른 해석으로는 마스크로 인하여 화장이나 꾸밈의 정도가 덜 필요해짐에 따라 소비가 줄었다고도 해석 가능하다.

아래는 전체 20대의 카테고리들 중 '5-(4)의 <그림6>'의 결과를 바탕으로 코로나 전-중-후로 cgi가 '변화되지 않은' 카테고리들의 상관계수 값과 이를 시각적으로 나타낸 그래프이다.

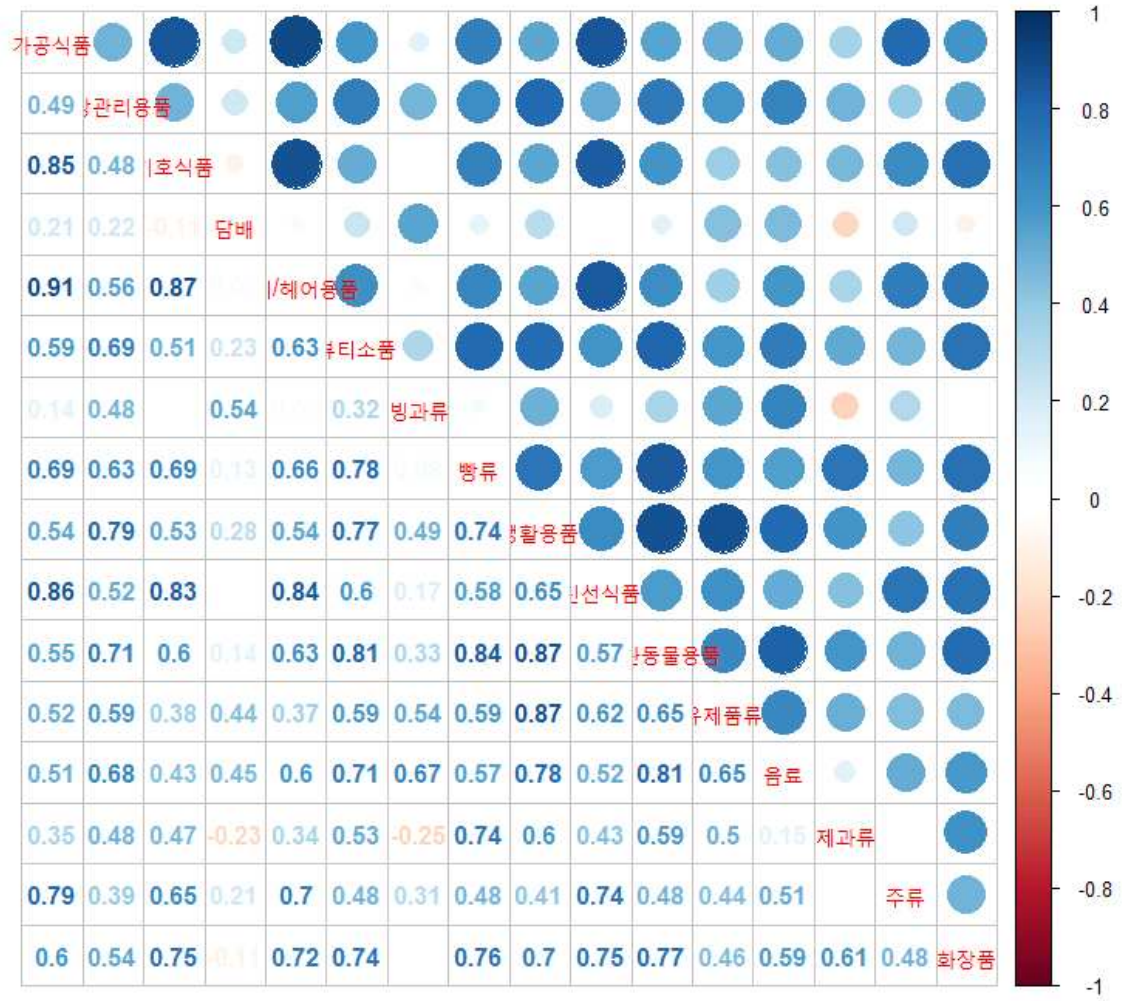


<그림10>

변하지 않은 품목들을 살펴보자면 가공식품, 유제품, 신선식품과 같이 마트나 가게에서 사는 일반적인 식료품은 크게 줄지 않은 것으로 보여진다. 이는 필수적으로 먹어야하는 식생활적인 부분의 관련 소비에 있어서는 크게 영향을 받지 않았다고 해석 가능하다. 또한 이곳에서 재미있는 점은 담배, 주류와 같은 기호품들인데 코로나 사태에도 불구하고 위와 같은 기호품들의 소비는 여전히 크게 영향을 받지 않았다는 점이 흥미로웠다.

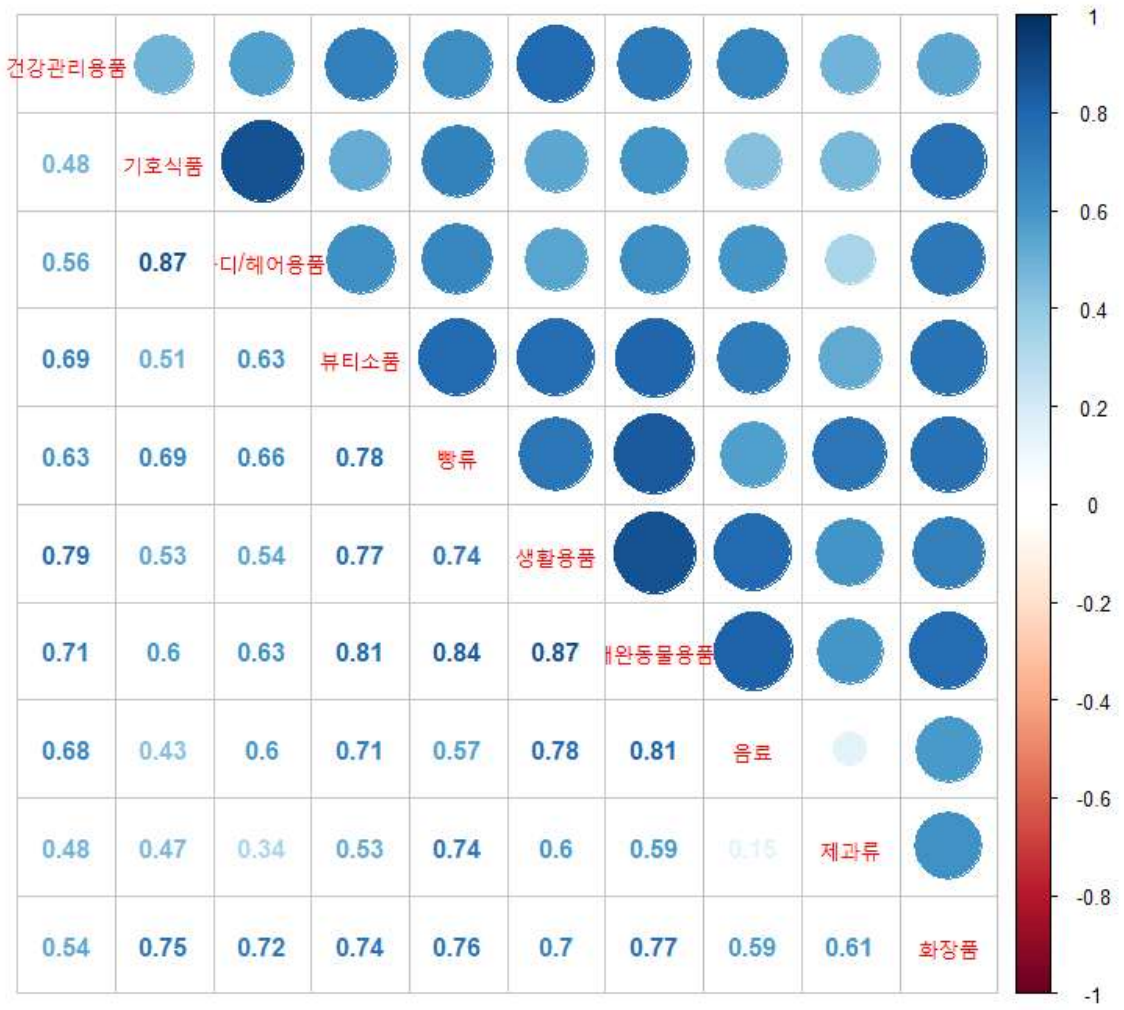
(7)-2 비20대

아래는 비20대 전체 16개 카테고리들의 월별 cgi값들을 기준으로 구한 상관계수 값과 이를 시각적으로 나타낸 그래프이다.



<그림11>

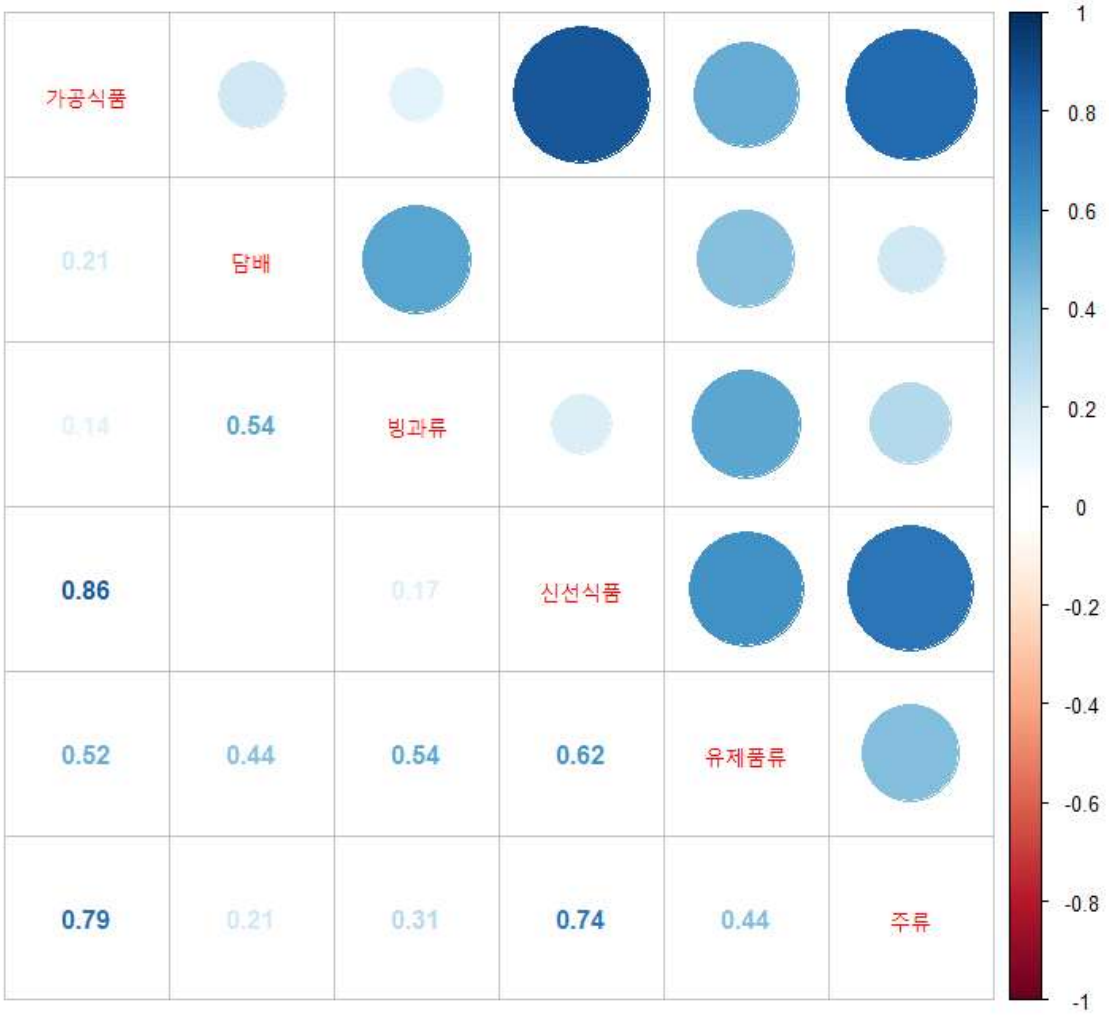
아래는 전체 비20대의 카테고리들 중 '5-(4)의 <그림6>'의 결과를 바탕으로 코로나 전-중-후로 cgi가 유의미하게 '변화된' 카테고리들의 상관계수 값과 이를 시각적으로 나타낸 그래프이다.



<그림12>

변한 카테고리들의 공통점을 찾아보자면, 비20대 또한 위에서 언급했던 '건강관리용품'을 제외하고는 20대와 비슷하게 해석 가능하다. 그러나 그 변한 '정도'의 차이가 있기는 하다.(5-(4)의 <그림6>을 참고)

아래는 전체 비20대의 카테고리들 중 '5-(4)의 <그림6>'의 결과를 바탕으로 코로나 전-중-후로 cgi가 '변화되지 않은' 카테고리들의 상관계수 값과 이를 시각적으로 나타낸 그래프이다.



<그림13>

변하지 않은 품목들을 살펴보자면 이 역시, 비20대 또한 위에서 언급했던 '건강관리 용품'을 제외하고는 20대와 비슷하게 해석 가능하다. 그러나 그 변한 '정도'의 차이가 있기는 하다.(5-(4)의 <그림6>을 참고)

(8) 카테고리별 군집분석(20대 vs 비20대) (feat. Ward의 방법)

아래는 (8)에서 분석에 사용하기 위해 전처리한 데이터들의 일부를 발췌한 것이다. 데이터는 코로나 전-중-후 시기의 cgi값의 평균을 카테고리별로 표준화(scaling) 하여 계산한 것이다.
표준화를 해주지 않으면 애초에 cgi값이 높은 품목 혹은 낮은 품목끼리 군집을 이루게 된다. 따라서 표준화 작업을 통해 우리는 코로나 전-중-후의 cgi값 변화의 비슷한 경향을 띄는 카테고리별의 군집을 묶을 수 있게 된다.

	G20_pre	G20_ing	G20_post
가공식품	0.42221136	0.71964841	-1.1418598
건강관리용품	1.06436314	-0.14443381	-0.9199293
기호식품	1.11542442	-0.29911739	-0.8163070
담배	-0.53650548	1.15375904	-0.6172536
바디/헤어용품	1.07765752	-0.17967601	-0.8979815
뷰티소품	1.14117041	-0.41794981	-0.7232206
빙과류	1.14098281	-0.41680771	-0.7241751
탕류	1.00933125	-0.01893131	-0.9903999
생활용품	1.03656553	-0.07765942	-0.9589061
신선식품	0.08790402	0.95314611	-1.0410501
애완동물용품	1.09656119	-0.23496740	-0.8615938
유제품류	-0.29573933	1.11451525	-0.8187759
음료	1.14810666	-0.46733726	-0.6807694
제과류	0.52161618	0.63134516	-1.1529613
주류	-1.08559287	0.20203833	0.8835545
화장품	0.83663166	0.27091244	-1.1075441

<G20_scale>: 20대

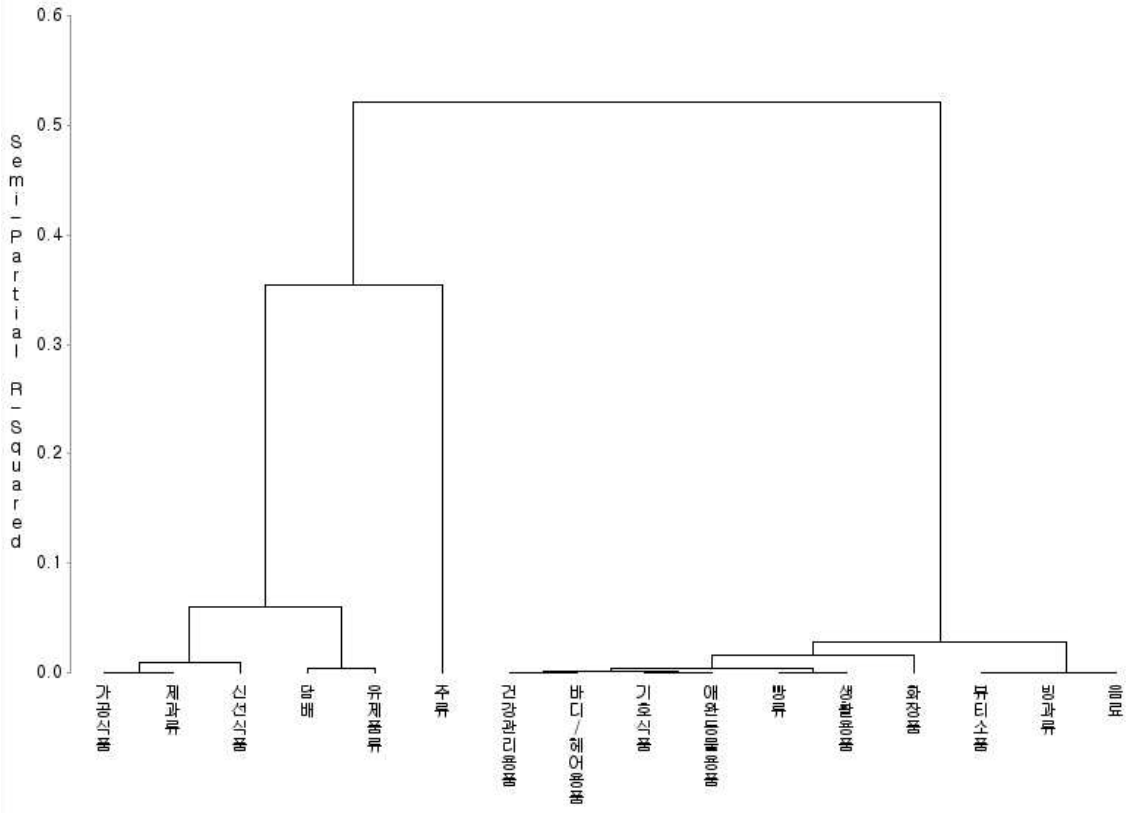
	G3060_pre	G3060_ing	G3060_post
가공식품	-0.7758671	1.12855671	-0.35268957
건강관리용품	-1.0479713	0.94388703	0.10408422
기호식품	1.0221079	-0.04578894	-0.97631899
담배	-1.1080755	0.83533220	0.27274325
바디/헤어용품	0.9195117	0.14511901	-1.06463075
뷰티소품	1.1427977	-0.42818561	-0.71461211
빙과류	0.7835248	0.34278968	-1.12631443
탕류	0.6507450	0.50070148	-1.15144647
생활용품	0.3763997	0.75717964	-1.13357939
신선식품	-0.9019988	1.07533795	-0.17333915
애완동물용품	0.9853094	0.02876071	-1.01407012
유제품류	-0.9682023	1.02902153	-0.06081921
음료	-0.5545210	1.15440350	-0.59988255
제과류	0.2189857	0.87235945	-1.09134516
주류	-1.1499175	0.48403440	0.66588314
화장품	0.2016066	0.88383677	-1.08544338

<G3060_scale>: 비20대

추가적으로 Ward의 방법을 이용한 이유는 군집 내 카테고리들은 동질적이고, 군집 간 카테고리들은 이질적인 분석의 방향을 원했고, 각 군집의 크기가 한쪽에 쏠리지 않는 방향을 원했기 때문이다.

(8)-1 20대

아래는 Ward 방법으로 군집분석을 한 20대의 카테고리별 나무구조 그림이다.



<그림14>

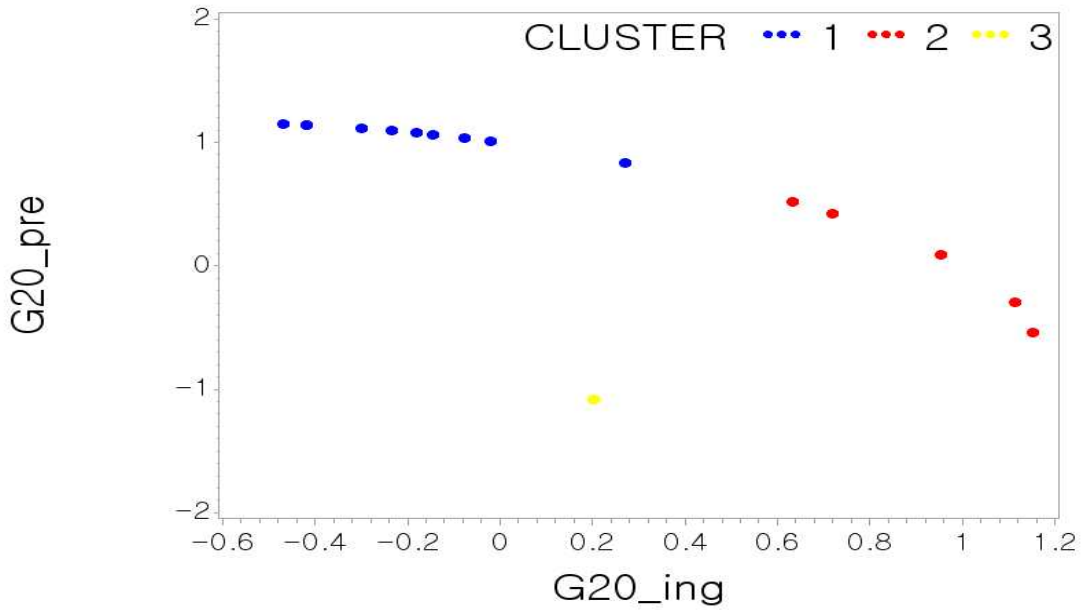
변화 vs 비변화로 나누었었던 (7)번의 분석과 같이 원래는 군집을 2개로 나누어서 (7)의 결과와 비교하려고 하였으나, 나무구조 그림과 아래에 나올 시각화된 산점도를 바탕으로 군집을 3개로 나누는 방향으로 수정하였고, 이에 대한 해석은 마지막에 인사이트에서 함께 하고자 한다.

SAS 시스템

OBS	VAR1	CLUSTER	CLUSNAME	G20_pre	G20_ing	G20_post
1	건강관리용품	1	CL4	1.0643631432	-0.144433813	-0.919929331
2	기호식품	1	CL4	1.1154244164	-0.299117393	-0.816307024
3	바디/헤어용품	1	CL4	1.0776575205	-0.179676005	-0.897981515
4	뷰티소품	1	CL4	1.1411704136	-0.417949807	-0.723220607
5	빙과류	1	CL4	1.1409828052	-0.416807706	-0.724175099
6	빵류	1	CL4	1.0093312497	-0.018931313	-0.990399936
7	생활용품	1	CL4	1.0365655282	-0.077659422	-0.958906106
8	애완동물용품	1	CL4	1.0965611937	-0.234967397	-0.861593797
9	음료	1	CL4	1.148106662	-0.467337261	-0.680769401
10	화장품	1	CL4	0.8366316554	0.27091244	-1.107544095
11	가공식품	2	CL3	0.4222113601	0.7196484097	-1.14185977
12	담배	2	CL3	-0.536505476	1.1537590371	-0.617253561
13	신선식품	2	CL3	0.0879040157	0.9531461132	-1.041050129
14	유제품류	2	CL3	-0.29573933	1.1145152497	-0.81877592
15	제과류	2	CL3	0.5216161838	0.6313451559	-1.15296134
16	주류	3	주류	-1.085592873	0.2020383335	0.8835545399

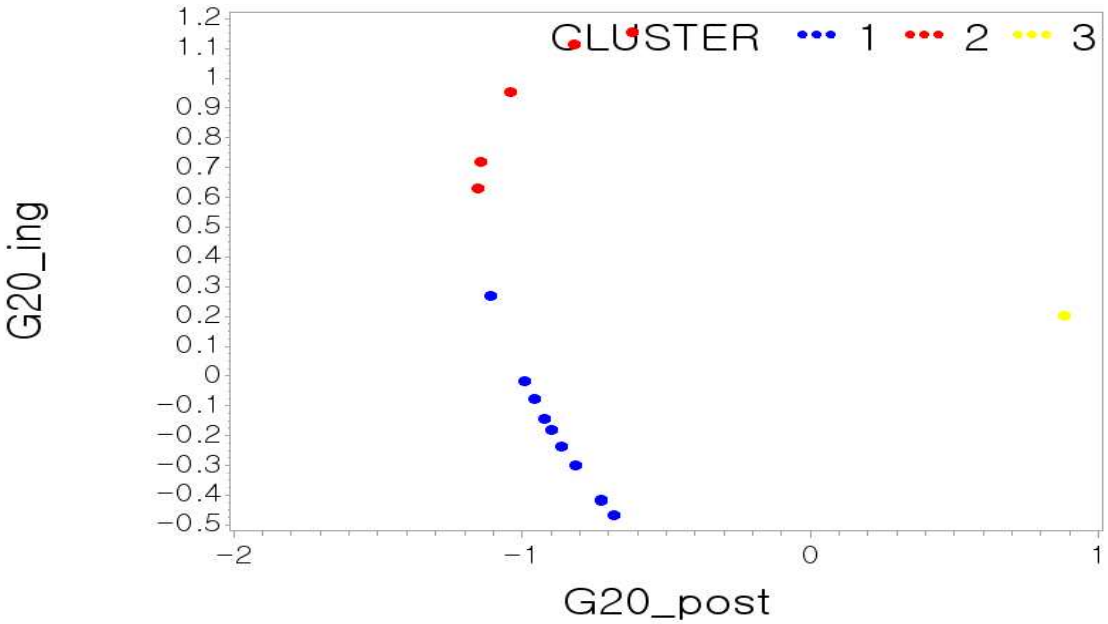
원본 데이터와 각각의 카테고리들이 분류된 군집들을 함께 나타낸 표이다.

아래의 그림은 군집을 3개로 나누었을 때 총 16개의 카테고리들이 어떤식으로 분류되는지 시각화한 산점도이다.



<그림15.1>

코로나 전~중 의 값을 축으로 하여 그린 산점도 이다.



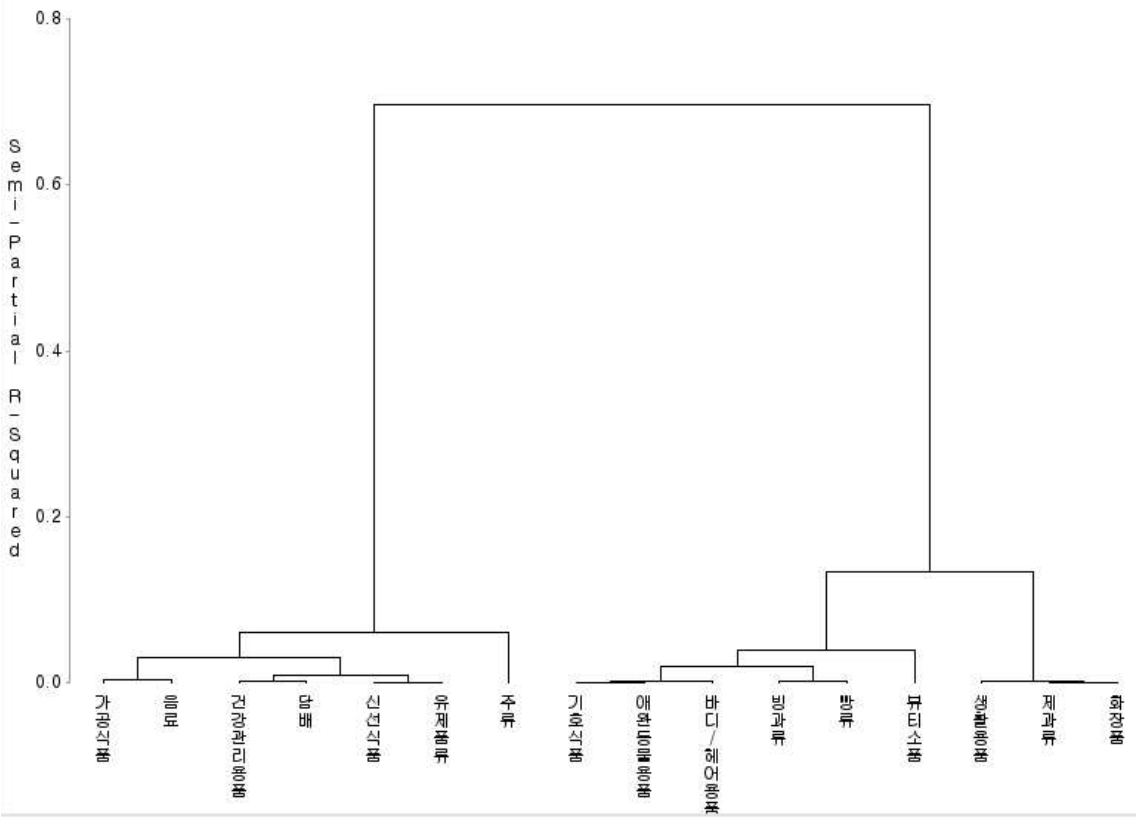
<그림15.2>

코로나 중~후 의 값을 축으로 하여 그린 산점도 이다.

※ 노란색점은 '주류'

(8)-2 비20대

아래는 군집분석을 한 비20대의 카테고리별 나무구조 그림이다.



<그림16>

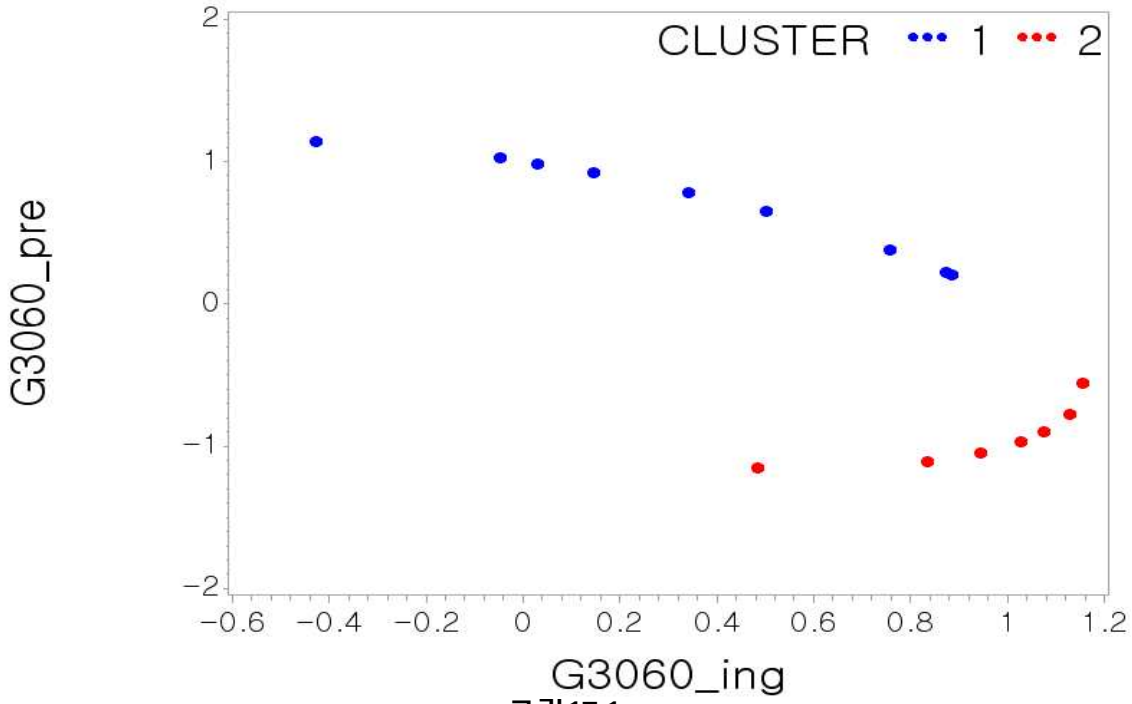
비 20대의 경우에는 20대와 다르게 변화 vs 비변화로 나누었었던 (7)번의 분석과 같이 군집을 2개로 나누어서 (7)의 결과와 비교 해 볼 것이다. 이에 대한 해석 또한 마지막에 인사이트에서 함께 하고자 한다.

SAS 시스템

OBS	VAR1	CLUSTER	CLUSNAME	G3060_pre	G3060_ing	G3060_post
1	기호식품	1	CL2	1.0221079243	-0.045788937	-0.976318987
2	바디/헤어용품	1	CL2	0.9195117399	0.1451190098	-1.06463075
3	뷰티소품	1	CL2	1.1427977158	-0.428185606	-0.71461211
4	빙과류	1	CL2	0.7835247538	0.3427896772	-1.126314431
5	빵류	1	CL2	0.6507449971	0.5007014761	-1.151446473
6	생활용품	1	CL2	0.3763997474	0.7571796388	-1.133579386
7	애완동물용품	1	CL2	0.9853094032	0.0287607134	-1.014070117
8	제과류	1	CL2	0.2189857085	0.8723594486	-1.091345157
9	화장품	1	CL2	0.2016066133	0.8838367698	-1.085443383
10	가공식품	2	CL3	-0.775867144	1.1285567091	-0.352689565
11	건강관리용품	2	CL3	-1.047971251	0.9438870345	0.1040842162
12	담배	2	CL3	-1.10807546	0.835332205	0.2727432547
13	신선식품	2	CL3	-0.9019988	1.075337949	-0.173339149
14	유제품류	2	CL3	-0.968202313	1.0290215272	-0.060819214
15	음료	2	CL3	-0.554520954	1.1544035005	-0.599882547
16	주류	2	CL3	-1.149917539	0.4840343997	0.6658831388

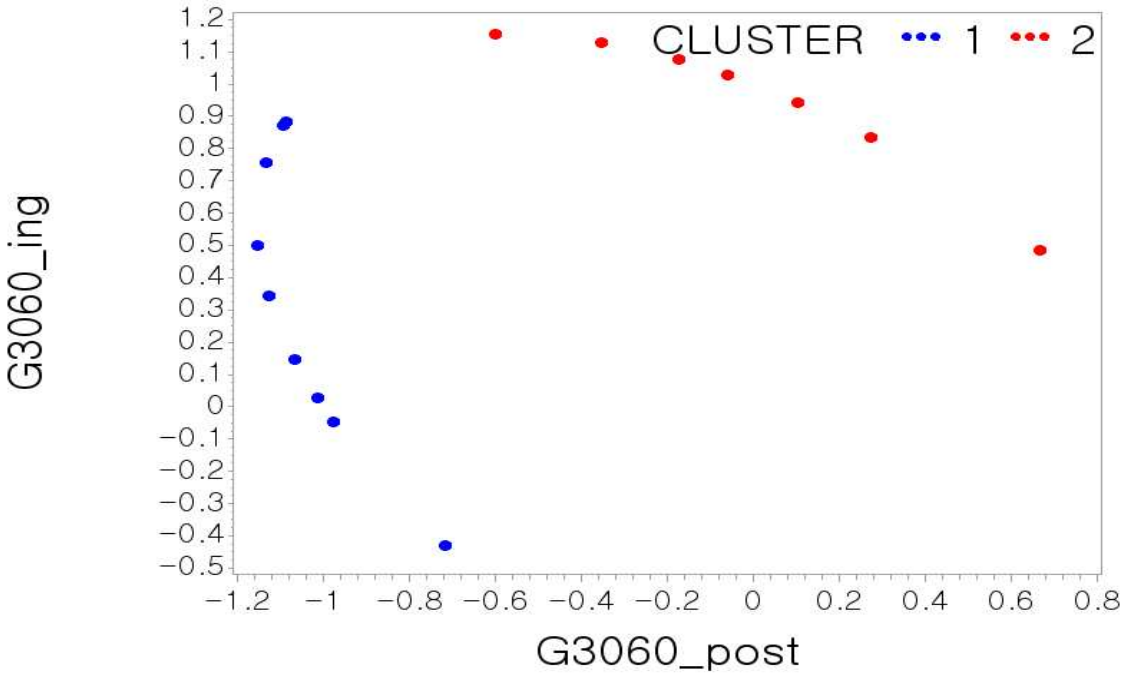
원본 데이터와 각각의 카테고리들이 분류된 군집들을 함께 나타낸 표이다.

아래의 그림은 군집을 2개로 나누었을 때 총 16개의 카테고리들이 어떤식으로 분류되는지 시각화한 산점도이다.



<그림17.1>

코로나 전~중 의 값을 축으로 하여 그린 산점도 이다.



<그림17.2>

코로나 중~후 의 값을 축으로 하여 그린 산점도 이다.

(9) 인사이트 2

(5)와 (8)의 분석을 바탕으로 설명하고자 한다.

품목(전체)	20대	비 20대	p-vaule(20)	p-value(non-20)	p-value(%)
가공식품	X	X	0.138	0.0554	0.71354705
건강관리용품	X	O	0.1521	0.0129	0.92181818
기호식품	O	O	0.0004	0.0174	0.02247191
담배	X	X	0.9279	0.9337	0.4984422
바디/헤어용품	O	O	0.031	0.0106	0.74519231
뷰티소품	O	O	0.0005	0.0002	0.71428571
빙과류	X	X	0.618	0.8682	0.4158256
빵류	O	O	0.0004	0.0001	0.8
생활용품	O	O	0.0066	0.0017	0.79518072
신선식품	X	X	0.4061	0.0895	0.81941082
애완동물용품	O	O	0.0001	0.0001	0.5
유제품	X	X	0.1006	0.1088	0.48042025
음료	O	O	0.0125	0.0058	0.68306011
제과류	O	O	0.0026	0.0024	0.52
주류	X	X	0.7177	0.2995	0.70556429
화장품	O	O	0.0148	0.0018	0.89156627

품목(변하지 않음)	p-value(%)
가공식품	0.713547053
담배	0.4984422
빙과류	0.415825595
신선식품	0.819410815
유제품	0.480420248
주류	0.705564294

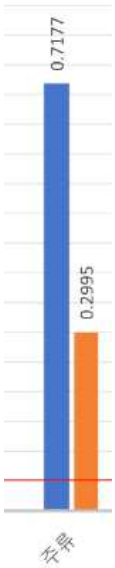
품목(변함)	p-value(%)
기호식품	0.02247191
바디/헤어용품	0.745192308
뷰티소품	0.714285714
빵류	0.8
생활용품	0.795180723
애완동물용품	0.5
음료	0.683060109
제과류	0.52
화장품	0.891566265

<그림18>

(5)의 결과에서는 '건강관리용품'을 제외하고는 20대vs비20대 모두 카테고리들이 코로나 전-중-후를 기준으로 cgi의 변화된품목과 변화되지않은 품목이 위와같은 결과로 나타났다.

또한 (8)의 결과도 아래와 같이 나오는 것을 알 수 있었다. 이 둘의 결과는 묶이게 되는 카테고리 항목들이 상당히 유사한 것을 관찰 할 수 있었고, (7)와 유사한 방향으로 해석이 가능하다.

20대		비20대	
뷰티소품	1	기호식품	1
빙과류	1	바디/헤어용품	1
건강관리용품	1	뷰티소품	1
바디/헤어용품	1	빙과류	1
빵류	1	빵류	1
생활용품	1	생활용품	1
음료	1	애완동물용품	1
기호식품	1	제과류	1
애완동물용품	1	화장품	1
화장품	1	가공식품	2
가공식품	2	건강관리용품	2
제과류	2	담배	2
담배	2	신선식품	2
유제품류	2	유제품류	2
신선식품	2	음료	2
주류	3	주류	2



<그림19>

그러나 주목해 봐야할 것은 바로 주류였다. (5)의 결과로는 알기 힘들었던 주류라는 카테고리가 군집분석의 결과 유난히 20대에서만 매우 다른 성질을 띄는 것을 알 수 있었다. (<그림14>, <그림15.1>, <그림15.2> 참고)

또한 추가적으로 <그림5>의 주류 품목에서도 다른 카테고리에 비하여 20대의 p-value값이 상대적으로 높은 것(주류 카테고리에서 cgi의 변화가 없음)을 확인할 수 있었다.

즉, '주류'라는 카테고리는 20대와 비20대 모두 통계적으로 유의미하게 cgi값이 변한 것은 아니나 20대의 경우가 비20대에 비하여 더더욱 영향을 받지 않았다고 해석이 어느정도 가능하다.

이를 통해서 인사이트를 도출해보자면, '코로나 시대에도 불구하고 20대의 경우 비20대보다 음주를 소비하는데에 있어서 전혀 영향을 받지않고 꾸준히 음주관련 소비를 하였다.' 라는 결과를 도출해볼 수 있다.

3. 결론

3.1 분석 결과 요약

우리는 위의 다양한 분석들을 통해 총 2가지의 인사이트를 도출하였다.
첫 번째로는. '코로나 시기에도 불구하고, 20대는 다른 연령에 비하여 마스크에 대한 관심도가 떨어졌으며, 20대는 타 세대에 비하여 상대적으로 안일한 생각과 대처를 하였다.' 라는 것이고,
두 번째로는, '코로나 시대에도 불구하고 20대의 경우 비20대보다 음주를 소비하는 데에 있어서 전혀 영향을 받지 않고 꾸준히 음주관련 소비를 하였다.' 라는 것이다.

사실 분석을 하고 난 이후 많이 부끄러웠다. 실제 저 당시의 우리 20대들의 인식이 저랬다는 사실은 실제 서울에서 20대 대학생 생활을 하고 있는 나 자신이 누구보다도 잘 알고 있었다. 실제로 객관적인 데이터와 분석의 지표가 통계적 유의성까지 띄면서 20대의 이런 적나라한 사실을 보여주는 것이 놀라울 따름이었다. 코로나를 대처하는 다른 세대와 현저하게 차이를 보이는 우리의 세대의 모습을 보며 많은 반성을 하였다.

3.2 분석의 장점 및 한계점 설명

(1) 분석의 장점
- 분석하고자 하는 개체 수(12만개이상)는 많았기 때문에, 표본의 부족으로 인하여 겪을 수 있는 어려움이 적었다.
- 분석 방향을 잡게 된 타당한 통계적 근거를 들어가며, 분석해 나가는 방향이 좋았던 것 같다.
- 하나의 방법으로 분석한 것 뿐만이 아니라, 큰 맥락으로는 두 가지 방법(1. t-tset를 이용한 p-value비교 2. 군집분석)으로 서로의 분석에서 보기 힘들었던 점을 찾을 수 있었고, 보완할 수 있었다.

(2) 한계점

- 코로나 시기를 일별로 구분하였지만, 실제 분석은 데이터의 한계로 인해 월별 데이터로 분석하였다.
- 코로나 전(13개월) 중(2개월) 후(2개월)로 시기별 데이터의 양 차이로 인한 분산의 차이가 있을 수 있다.
- 대한민국의 특성상 서울과 비서울의 유동인구와 소비량, 주요 세대분포가 다르기 때문에 다른 지역에서는 다른 결과가 나올 수도 있다. 추가적으로 대구의 경우 신천지와 같은 특수한 사건과 지역적정책이 있었기에 또 다른 특색이 있었을 수 있었다고 생각한다.(서울로 지역을 한정한 이유 중 하나)

3.3 추가 연구사항 제안

서론의 1.3-2에서 제안하였던 추가이용가능 데이터들을 통해서 더 구체적인 추가 분석이 가능하다고 생각한다.

(1)

이번분석에서는 카테고리를 cgi변화에 따라서 소분류별로 나누었는데, 개인카드이용내역데이터(card_20200717.csv)를 통해서 카테고리 내역 하나하나의 구체적인 상품들까지 세세하게 분석이 가능할 것이라고도 본다.

(2)

배달관련데이터(delivery.csv)를 통해서 코로나로 인한 언택트 시대에 어떤 카테고리의 배달의 소비량의 얼마나 증가하였는지 연령별, 성별, 지역별로 상품명, 배달량 등을 분석해볼 수 있었다고 생각한다.

(3)

추가적으로 (fpopl.csv, index.csv) 데이터를 함께 이용하여, 코로나 시대 전-중-후로 하여 지역별 사람들의 세대별, 성별 인구수와 유동성을 함께 고려하여 그에 상관되는 소비활동을 지협적으로 분석도 가능해봤을 것이라는 생각을 한다.

※ 참고문헌

데이터 및 차트 출처

NAVER <그림7>

데이콘(데이터: https://dacon.io/competitions/official/235618/data)

※ 사용코드

(1) 포스트코로나의정의(코로나확진자그래프)

```
####################패키지, 라이브러리###################
#install.packages("ggplot2")
library(ggplot2)
############################################################


################데이터 불러오기 및 전처리#################

Time = read.csv("C:\\Users\\GWANGRYUL\\Desktop\\COVID19_DATA\\Time.csv")
Time$date = as.Date(Time$date)
str(Time)

#confirm(확진자수)칼럼을 누적도수 아닌 일별 도수로 바꾸기
confirmed = Time$confirmed
confirmed_new = rep(0,length(confirmed))
confirmed_new[1] = confirmed[1]
for (i in 2:length(confirmed)) {
  confirmed_new[i] = confirmed[i] - confirmed[i-1]
}
confirmed_new
Time$confirmed_new = confirmed_new

#코로나 확진자수 최고치인 날, 확진자수 찾기
max_person = max(confirmed_new)
max_date = Time$date[which.max(Time$confirmed_new)]
max_person
max_date

#코로나중 기간
Time$date[36]
Time$date[52]
confirmed_new[36]
confirmed_new[52]
############################################################

####################그래프그리기#######################

ggplot(Time, aes(x = date, y = confirmed_new)) +
  geom_line(color='black', size = 1) +
  geom_line(mapping = aes(x = date[36]), color = 'red', size = 1) +
  geom_line(mapping = aes(x = date[52]), color = 'red', size = 1) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5, size =20),
        text = element_text(face = "bold", size =15)) +
  ggtitle("COVID-19 일별 확진자수") +  xlab("월") + ylab("확진자수(명)") +
  geom_point(mapping = aes(x = max_date, y = max_person ), color="red", size =3) +
  annotate("text", x= max_date, y=max_person , label="2020-02-29(813명)", color = "black",hjust = -0.1) +
  annotate("text", x= Time$date[1], y=400 , label="Pre COVID19", color = "black",hjust = 0.1) +
  annotate("text", x= Time$date[80], y=400 , label="Post COVID19", color = "black",hjust = 0.2) +
  annotate("text", x= Time$date[37], y=-30 , label="2020-02-24", color = "black",hjust = 0.8) +
  annotate("text", x= Time$date[46], y=-30 , label="2020-03-11", color = "black", hjust = 0)

#기간의 이유 policy데이터 이용
#감염병위기경보단계 "심각"으로 격상 2020-02-23 한 다음날부터#
#코로나 확진자 200명 이상이 유지되는 시기
############################################################
```

(2) 코로나 전-중-후 cgi변화 그래프 & 데이터전처리

```r
####################패키지, 라이브러리####################
#install.packages("dplyr")
#install.packages("readr")
library(dplyr)
library(readr)
############################################################


################데이터 불러오기 및 전처리#################

index <- read_csv("C:/Users/GWANGRYUL/Desktop/COVID19_DATA/index.csv")
head(index)
#View(index)
str(index)


#코로나 전(전연령)
index_pre = index %>%
  select(period, catl, catm, age, cgi) %>%
  filter(  !is.na(index$age) &
             index$gender == 'all' &
             index$sigungu == 'all' &
             index$period >= 201901 & period <= 202001 &
             index$catm != "기타화장품") %>%
  arrange(period, catl, catm)

#코로나 중(전연령)
index_ing = index %>%
  select(period, catl, catm, age, cgi) %>%
  filter(  !is.na(index$age) &
             index$gender == 'all' &
             index$sigungu == 'all' &
             index$period >= 202002 & period <= 202003 &
             index$catm != "기타화장품") %>%
  arrange(period, catl, catm)

```

```r
#코로나 후(전연령)
index_post = index %>%
  select(period, catl, catm, age, cgi) %>%
  filter(    !is.na(index$age) &
               index$gender == 'all' &
               index$sigungu == 'all' &
               index$period >= 202004 &
               index$catm != "기타화장품") %>%
  arrange(period, catl, catm)


#코로나전(상품별,연령별 cgi)
pre_cgi =
  index_pre %>%
  group_by(catm, age) %>%
  summarize(pre_cgi = mean(cgi)) %>%
  arrange(age, catm)

#코로나중(상품별,연령별 cgi)
ing_cgi =
  index_ing %>%
  group_by(catm, age) %>%
  summarize(ing_cgi = mean(cgi)) %>%
  arrange(age, catm)

#코로나후(상품별,연령별 cgi)
post_cgi =
  index_post %>%
  group_by(catm, age) %>%
  summarize(post_cgi = mean(cgi)) %>%
  arrange(age, catm)

#코로나 전중후(상품별,연령별 cgi)
all_cgi = cbind(pre_cgi, ing_cgi[,3], post_cgi[,3])

##################################
```

```r
##################################
library(gridExtra)

#탐색적 데이터분석 EDA

cgi_scatter1 = ggplot(all_cgi, aes(pre_cgi, ing_cgi)) #전~중
cgi_scatter2 = ggplot(all_cgi, aes(ing_cgi, post_cgi)) #중~후

#변화없이 일정하다면 빨간색 선을 따라가야함

#그렇지 못한모습
P1 = cgi_scatter1 +
  geom_point() +
  geom_smooth(formula = y~x, method = 'lm') +
  geom_abline(slope=1, intercept=0, size =1, color = 'red') +
  theme(plot.title = element_text(face = "bold", size =20),
        text = element_text(face = "bold", size =15)) +
  ggtitle("Pre COVID19 ~ Ing COVID19")+
  labs(x = 'Pre COVID19의 품목별 cgi', y = 'Ing COVID19의 품목별 cgi')

#그럭저럭 괜찮은 모습
P2 = cgi_scatter2 +
  geom_point() +
  geom_smooth(formula = y~x, method = 'lm') +
  geom_abline(slope=1, intercept=0, size =1, color = 'red') +
  theme(plot.title = element_text(face = "bold", size =20),
        text = element_text(face = "bold", size =15)) +
  ggtitle("Ing COVID19 ~ Post COVID19")+
  labs(x = 'Ing COVID19의 품목별 cgi', y = 'Post COVID19의 품목별 cgi')

grid.arrange(P1,P2, layout_matrix = rbind(c(1,2)))

##################################
```

(3) 20대vs비20대 의 카테고리별 월별 cgi변화(시각화그래프)

```
#####################시각화#####################

#20대 vs 비 20대 의 품목별 시간순서 cgi변화

#################RGB#############################
#install.packages("RColorBrewer")
library(RColorBrewer)
############################################

#20대 전체 cgi
ggplot(date_catm_cgi_all_20_matrix, aes(x = rownames(date_catm_cgi_all_20_matrix), y = 가공식품))+
  geom_line(aes(group=1),color="#93DAFF", size = 1) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5, size =20),
        text = element_text(face = "bold", size =15)) +
  ggtitle("20대 월별 전체 cgi") +  xlab("년/월") + ylab("cgi")+
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 건강관리용품, group=1),
            color="#00A5FF", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 기호식품, group=1),
            color="#0000FF", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 담배, group=1),
            color="#3DFF92", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = `바디/헤어용품`, group=1),
            color="#80E12A", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 뷰티소품, group=1),
            color="#52E4DC", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 빙과류, group=1),
            color="#FFB6C1", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 빵류, group=1),
            color="#FF82FF", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 생활용품, group=1),
            color="#CD1039", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 신선식품, group=1),
            color="#FFDC3C", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 애완동물용품, group=1),
            color="#FF8200", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 유제품류, group=1),
            color="#FFB182", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 음료, group=1),
            color="#957745", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 제과류, group=1),
            color="#9E5A5A", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 주류, group=1),
            color="#9400D3", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_20_matrix), y = 화장품, group=1),
            color="#828282", size = 1) +
  geom_vline(mapping = aes(xintercept = "202002"), color = 'red', size = 1) +
  geom_vline(mapping = aes(xintercept = "202003"), color = 'red', size = 1) +
  geom_vline(mapping = aes(xintercept = "201902"), color = 'black', size = 0.5) +
  geom_vline(mapping = aes(xintercept = "201903"), color = 'black', size = 0.5) +
  annotate("text", x= 8, y=100 , label="Pre COVID19", color = "black", size = 5) +
  annotate("text", x= 16, y=100 , label="Post COVID19", color = "black", size = 5)


#비20대 전체 cgi
ggplot(date_catm_cgi_all_3060_matrix, aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 가공식품))+
  geom_line(aes(group=1),color="#93DAFF", size = 1) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5, size =20),
        text = element_text(face = "bold", size =15)) +
  ggtitle("비20대 월별 전체 cgi") +  xlab("년/월") + ylab("cgi")+
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 건강관리용품, group=1),
            color="#00A5FF", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 기호식품, group=1),
            color="#0000FF", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 담배, group=1),
            color="#3DFF92", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = `바디/헤어용품`, group=1),
            color="#80E12A", size = 1) +
```

```
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 뷰티소품, group=1),
            color="#52E4DC", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 빙과류, group=1),
            color="#FFB6C1", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 빵류, group=1),
            color="#FF82FF", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 생활용품, group=1),
            color="#CD1039", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 신선식품, group=1),
            color="#FFDC3C", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 애완동물용품, group=1),
            color="#FF8200", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 유제품류, group=1),
            color="#FFB182", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 음료, group=1),
            color="#957745", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 제과류, group=1),
            color="#9E5A5A", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 주류, group=1),
            color="#9400D3", size = 1) +
  geom_line(aes(x = rownames(date_catm_cgi_all_3060_matrix), y = 화장품, group=1),
            color="#828282", size = 1) +
  geom_vline(mapping = aes(xintercept = "202002"), color = 'red', size = 1) +
  geom_vline(mapping = aes(xintercept = "202003"), color = 'red', size = 1) +
  geom_vline(mapping = aes(xintercept = "201902"), color = 'black', size = 0.5) +
  geom_vline(mapping = aes(xintercept = "201903"), color = 'black', size = 0.5) +
  annotate("text", x= 8, y=100 , label="Pre COVID19", color = "black", size = 5) +
  annotate("text", x= 16, y=100 , label="Post COVID19", color = "black", size = 5)


##############################################

```

(4) MANOVA분석(20대vs비20대의 유의미한 차이)

```r
View(all_cgi_3060)
#MANOVA분석 (연령별 ~ (코로나전, 중 , 후), 윌크스 람다 통계량 이용)
M = manova(cbind(pre_cgi, ing_cgi, post_cgi)~age, data = all_cgi)
summary(M, intercept = T, test = "Wilks")
#연령별 유의미한 차이가 있음

#사후분석(일변량 검정 통계량)
summary.aov(M)
#코로나 이전의 cgi양에서 차이가 있고
#코로나 중,후 의 cgi양에서 차이가 없음

#연령별 all_cgi 데이터
all_cgi_20 =
  all_cgi %>%
  filter(age == 20)
all_cgi_30 =
  all_cgi %>%
  filter(age == 30)
all_cgi_40 =
  all_cgi %>%
  filter(age == 40)
all_cgi_50 =
  all_cgi %>%
  filter(age == 50)
all_cgi_60 =
  all_cgi %>%
  filter(age == 60)
#비 20대
all_cgi_3060 =
  all_cgi %>%
  filter(age != 20)

#비 20대 MANOVA분석 (연령별 ~ (코로나전, 중 , 후), 윌크스 람다 통계량 이용)
M = manova(cbind(pre_cgi, ing_cgi, post_cgi)~age, data = all_cgi_3060)
summary(M, intercept = T, test = "Wilks")
#0.3788 유의미한 차이 X -> 즉 20대를 끼고했을 때 유의미한 차이가 있으므로 20대만
#코로나 이전의 소비에서 차이가 있음
#(인사이트)->20대의 소비cgi의 전중후 패턴변화가 다른세대와 차이가 있음

#t.test를 통해서 다시한번 구체적으로 검증해보자
#연령별 전~중, 중~후 의 cgi 변화(평균값의 차이가 유의미한지, 유의수준5%)
#20대
t.test(all_cgi_20$pre_cgi,all_cgi_20$ing_cgi) #변화x
t.test(all_cgi_20$ing_cgi,all_cgi_20$post_cgi) #변화x
#30대
t.test(all_cgi_30$pre_cgi,all_cgi_30$ing_cgi) #변화o
t.test(all_cgi_30$ing_cgi,all_cgi_30$post_cgi) #변화x
#40대
t.test(all_cgi_40$pre_cgi,all_cgi_40$ing_cgi) #변화o
t.test(all_cgi_40$ing_cgi,all_cgi_40$post_cgi) #변화x
#50대
t.test(all_cgi_50$pre_cgi,all_cgi_50$ing_cgi) #변화o
t.test(all_cgi_50$ing_cgi,all_cgi_50$post_cgi) #변화x
#60대
t.test(all_cgi_60$pre_cgi,all_cgi_60$ing_cgi) #변화o
t.test(all_cgi_60$ing_cgi,all_cgi_60$post_cgi) #변화x
```

(5) 카테고리별 t-test결과 cgi변화하였는지(p-value)(20대v비20대나눔)

```
#(전세대)월별, 카테고리별 cgi평균
#코로나전
date_catm_cgi_pre =
  index_pre %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))

#코로나중
date_catm_cgi_ing =
  index_ing %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))

#코로나후
date_catm_cgi_post =
  index_post %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))


#(20대)월별, 그룹별 cgi평균
#코로나전
date_catm_cgi_pre_20 =
  index_pre %>% filter(age == 20) %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))

#코로나중
date_catm_cgi_ing_20 =
  index_ing %>% filter(age == 20) %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))

```

```r
#코로나후
date_catm_cgi_post_20 =
  index_post %>% filter(age == 20) %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))


#(비20대)월별, 그룹별 cgi평균
#코로나전
date_catm_cgi_pre_3060 =
  index_pre %>% filter(age != 20) %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))

#코로나중
date_catm_cgi_ing_3060 =
  index_ing %>% filter(age != 20) %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))

#코로나후
date_catm_cgi_post_3060 =
  index_post %>% filter(age != 20) %>%
  group_by(period,catm) %>%
  summarize(cgi = mean(cgi))

#date_catm_cgi_pre_20
#date_catm_cgi_ing_20
#date_catm_cgi_post_20
#date_catm_cgi_pre_3060
#date_catm_cgi_ing_3060
#date_catm_cgi_post_3060
```

```r
#(20대)월x소품목 별 cgi 행렬
#전
date_catm_cgi_pre_20_matrix = matrix(data = date_catm_cgi_pre_20$cgi, ncol = 16, byrow = T)
date_catm_cgi_pre_20_matrix = cbind(c("pre"),date_catm_cgi_pre_20_matrix)
colnames(date_catm_cgi_pre_20_matrix) = c("시기",sort(unique(index_pre$catm)))
rownames(date_catm_cgi_pre_20_matrix) = sort(unique(index_pre$period))
#중
date_catm_cgi_ing_20_matrix = matrix(data = date_catm_cgi_ing_20$cgi, ncol = 16, byrow = T)
date_catm_cgi_ing_20_matrix = cbind(c("ing"),date_catm_cgi_ing_20_matrix)
colnames(date_catm_cgi_ing_20_matrix) = c("시기",sort(unique(index_ing$catm)))
rownames(date_catm_cgi_ing_20_matrix) = sort(unique(index_ing$period))
#후
date_catm_cgi_post_20_matrix = matrix(data = date_catm_cgi_post_20$cgi, ncol = 16, byrow = T)
date_catm_cgi_post_20_matrix = cbind(c("post"),date_catm_cgi_post_20_matrix)
colnames(date_catm_cgi_post_20_matrix) = c("시기",sort(unique(index_post$catm)))
rownames(date_catm_cgi_post_20_matrix) = sort(unique(index_post$period))


#(비20대)월x소품목 별 cgi 행렬
#전
date_catm_cgi_pre_3060_matrix = matrix(data = date_catm_cgi_pre_3060$cgi, ncol = 16, byrow = T)
date_catm_cgi_pre_3060_matrix = cbind(c("pre"),date_catm_cgi_pre_3060_matrix)
colnames(date_catm_cgi_pre_3060_matrix) = c("시기",sort(unique(index_pre$catm)))
rownames(date_catm_cgi_pre_3060_matrix) = sort(unique(index_pre$period))
#중
date_catm_cgi_ing_3060_matrix = matrix(data = date_catm_cgi_ing_3060$cgi, ncol = 16, byrow = T)
date_catm_cgi_ing_3060_matrix = cbind(c("ing"),date_catm_cgi_ing_3060_matrix)
colnames(date_catm_cgi_ing_3060_matrix) = c("시기",sort(unique(index_ing$catm)))
rownames(date_catm_cgi_ing_3060_matrix) = sort(unique(index_ing$period))
#후
date_catm_cgi_post_3060_matrix = matrix(data = date_catm_cgi_post_3060$cgi, ncol = 16, byrow = T)
date_catm_cgi_post_3060_matrix = cbind(c("post"),date_catm_cgi_post_3060_matrix)
colnames(date_catm_cgi_post_3060_matrix) = c("시기",sort(unique(index_post$catm)))
rownames(date_catm_cgi_post_3060_matrix) = sort(unique(index_post$period))
```

```r
#date_catm_cgi_pre_20_matrix
#date_catm_cgi_ing_20_matrix
#date_catm_cgi_post_20_matrix
#date_catm_cgi_pre_3060_matrix
#date_catm_cgi_ing_3060_matrix
#date_catm_cgi_post_3060_matrix

##(20대)월x소품목 별 cgi 행렬(전중후)
date_catm_cgi_all_20_matrix =
  rbind(date_catm_cgi_pre_20_matrix, date_catm_cgi_ing_20_matrix, date_catm_cgi_post_20_matrix)
#(비20대)월x소품목 별 cgi 행렬(전중후)
date_catm_cgi_all_3060_matrix =
  rbind(date_catm_cgi_pre_3060_matrix, date_catm_cgi_ing_3060_matrix, date_catm_cgi_post_3060_matrix)

#데이터프레임으로 바꿈
date_catm_cgi_all_20_matrix = as.data.frame(date_catm_cgi_all_20_matrix)
date_catm_cgi_all_3060_matrix = as.data.frame(date_catm_cgi_all_3060_matrix)
#숫자형으로 바꿔줌
for (i in 2:17){
  date_catm_cgi_all_20_matrix[,i] = as.numeric(date_catm_cgi_all_20_matrix[,i])
  date_catm_cgi_all_3060_matrix[,i] = as.numeric(date_catm_cgi_all_3060_matrix[,i])
}

write.csv(date_catm_cgi_all_20_matrix,"D:\\M_20.csv")
write.csv(date_catm_cgi_all_3060_matrix,"D:\\M_3060.csv")
##################################################################################3333
#이후 SAS에서 다변량 분산분석을 통해 20vs비20 의 변화된 품목을 보여줌(아래는 SAS코드)
##################################################################################################3
```

```
/*20대*/
 /*데이터를 불러옴*/
proc import datafile = 'D:₩M_20.csv'
out = M_20 dbms = csv replace;
getnames = yes;
run;

/*MANOVA분석*/
 proc glm data = M_20;
class var2; /*그룹변수*/
 model var3 --var18 = var2;
MANOVA H = var2/ PRINTE PRINTH; /*PRINTE가 오차제곱합행렬 , PRINTH가 처리제곱합행렬*/
 run;

/*비20대*/
 /*데이터를 불러옴*/
 proc import datafile = 'D:/M_3060.csv'
out = M_3060 dbms = csv replace;
getnames = yes;
run;

/*MANOVA분석*/
 proc glm data = M_3060;
class var2; /*그룹변수*/
 model var3 --var18 = var2;
MANOVA H = var2/ PRINTE PRINTH; /*PRINTE가 오차제곱합행렬 , PRINTH가 처리제곱합행렬*/
 run;
```

(6) 카테고리별 상관관계분석(시각화그래프)

```
################################################
##카테고리별 상관관계분석

#install.packages("corrplot")
library(corrplot)

##전체##
new_20 <- date_catm_cgi_all_20_matrix[-1]
cor_20 <- cor(new_20)
corrplot.mixed(cor_20)
new_3060 <- date_catm_cgi_all_3060_matrix[-1]
cor_3060 <- cor(new_3060)
corrplot.mixed(cor_3060)

##변화 비변화 나눠서##
a <- c(4,6,7,9,10,12,14,15,17)
b <- c(2,3,5,8,11,13,16)
c <- c(3,4,6,7,9,10,12,14,15,17)
d <- c(2,5,8,11,13,16)

yes_20 <- date_catm_cgi_all_20_matrix[a]
no_20 <- date_catm_cgi_all_20_matrix[b]
yes_3060 <- date_catm_cgi_all_3060_matrix[c]
no_3060 <- date_catm_cgi_all_3060_matrix[d]
cor_yes_20 <- cor(yes_20)
cor_no_20 <- cor(no_20)
cor_yes_3060 <- cor(yes_3060)
cor_no_3060 <- cor(no_3060)
corrplot.mixed(cor_yes_20)
corrplot.mixed(cor_no_20)
corrplot.mixed(cor_yes_3060)
corrplot.mixed(cor_no_3060)
```

(7) 카테고리 군집분석(20대vs비20대)

```r
#군집분석 위한 데이터 만들기
# 코로나 전-중-후 변화별로 카테고리별 cgi변화량 만들기
#20대
G20_pre = date_catm_cgi_all_20_matrix %>% filter(시기 == 'pre')
G20_pre = apply(G20_pre[,-1],2,mean)
G20_ing = date_catm_cgi_all_20_matrix %>% filter(시기 == 'ing')
G20_ing = apply(G20_ing[,-1],2,mean)
G20_post = date_catm_cgi_all_20_matrix %>% filter(시기 == 'post')
G20_post = apply(G20_post[,-1],2,mean)

G20 = as.data.frame(cbind(G20_pre, G20_ing, G20_post))
str(G20)
#비20대
G3060_pre = date_catm_cgi_all_3060_matrix %>% filter(시기 == 'pre')
G3060_pre = apply(G3060_pre[,-1],2,mean)
G3060_ing = date_catm_cgi_all_20_matrix %>% filter(시기 == 'ing')
G3060_ing = apply(G3060_ing[,-1],2,mean)
G3060_post = date_catm_cgi_all_20_matrix %>% filter(시기 == 'post')
G3060_post = apply(G3060_post[,-1],2,mean)

G3060 = as.data.frame(cbind(G3060_pre, G3060_ing, G3060_post))
str(G3060)


#단순 cgi수치를 카테고리별로 표준화 시켜주기

#20대
G20t = as.data.frame(t(G20))
for (i in 1:ncol(G20t)){
  G20t[,i] = scale(G20t[,i])
}
G20_scale = as.data.frame(t(G20t))
G20_scale

#비20대
G3060t = as.data.frame(t(G3060))
for (i in 1:ncol(G3060t)){
  G3060t[,i] = scale(G3060t[,i])
}
G3060_scale = as.data.frame(t(G3060t))
G3060_scale

#이후 SAS로 넘어가서 군집분석 실행
write.csv(G20_scale,"D:\\G20_scale.csv")
write.csv(G3060_scale,"D:\\G3060_scale.csv")



####################아래는 군집분석 SAS코드############################
```

```
/*20대*/
/* 카테고리 나누는 군집분석*/
proc import datafile = 'D:₩G20_scale.csv' dbms = csv out = G20_scale replace;
run;

/*계층적 군집분석*/
/*method =  single(최단연결법), complete(최장연결법), ward(왈드연결법), average(평균연결법), centroid(중심연결법)에 따라서 달라짐*
/*적정군집수는 군집화 인덱스인 pseudo-F값이 전후보다 높거나 pseudo-T 값이 낮아진 단계에서 결정*/


proc cluster data = G20_scale method = ward outtree = out1 standard pseudo rsq;
var G20_pre G20_ing G20_post;
id var1;
copy G20_pre G20_ing G20_post;
/*나무구조그림 그리기, ncl= 군집의개수지정*/
proc tree data = out1 ncl=3 out=cluster_1;
id var1;
proc print data = cluster_1;
run;
/*데이터 정렬 및 병합*/
proc sort data = cluster_1;
by var1;
run;
proc sort data = G20_scale;
by var1;
run;
data cluster_1d;
merge cluster_1 G20_scale;
by var1;
run;
proc sort data = cluster_1d;
by cluster;
run;
proc print data = cluster_1d;
run;
```

```
/*군집별 기초통계량*/
proc means data = cluster_1d;
by cluster;
run;
/*군집분석 시각화*/
proc gplot data=cluster_1d;
   plot G20_pre*G20_ing = cluster / vaxis=axis1 haxis=axis2 legend=legend;
   axis1 label=(h=3 a=90 r=0) value=(h=2);
   axis2 label=(h=3)  value=(h=2);
   symbol1 i=none h=1.5 v=dot  c=blue;
   symbol2 i=none h=1.5 v=dot  c=red;
   symbol3 i=none h=1.5 v=dot  c=yellow;
   legend position=(top right  inside) mode=share label =(h=3 ) value =(h=3);
run;
proc gplot data=cluster_1d;
   plot G20_ing*G20_post = cluster / vaxis=axis1 haxis=axis2 legend=legend;
   axis1 label=(h=3 a=90 r=0) value=(h=2);
   axis2 label=(h=3)  value=(h=2);
   symbol1 i=none h=1.5 v=dot  c=blue;
   symbol2 i=none h=1.5 v=dot  c=red;
   symbol3 i=none h=1.5 v=dot  c=yellow;
   legend position=(top right  inside) mode=share label =(h=3 ) value =(h=3);
run;
```

```
/*비 20대*/
/* 카테고리 나누는 군집분석*/
proc import datafile = 'D:₩G3060_scale.csv' dbms = csv out = G3060_scale replace;
run;

/*계층적 군집분석*/
/*method =  single(최단연결법), complete(최장연결법), ward(왈드연결법), average(평균연결법), centroid(중심연결법)에 따라서 달라짐*/
/*적정군집수는 군집화 인덱스인 pseudo-F값이 전후보다 높거나 pseudo-T 값이 낮아진 단계에서 결정*/


proc cluster data = G3060_scale method = ward outtree = out2 standard pseudo rsq;
var G3060_pre G3060_ing G3060_post;
id var1;
copy G3060_pre G3060_ing G3060_post;
/*나무구조그림 그리기, ncl= 군집의개수지정*/
proc tree data = out2 ncl=2 out=cluster_2;
id var1;
proc print data = cluster_2;
run;
/*데이터 정렬 및 병합*/
proc sort data = cluster_2;
by var1;
run;
proc sort data = G3060_scale;
by var1;
run;
data cluster_2d;
merge cluster_2 G3060_scale;
by var1;
run;
proc sort data = cluster_2d;
by cluster;
run;
proc print data = cluster_2d;
run;
/*군집별 기초통계량*/
proc means data = cluster_2d;
by cluster;
run;
/*군집분석 시각화*/
proc gplot data=cluster_2d;
    plot G3060_pre*G3060_ing = cluster / vaxis=axis1 haxis=axis2 legend=legend;
    axis1 label=(h=3 a=90 r=0) value=(h=2);
    axis2 label=(h=3)  value=(h=2);
    symbol1 i=none h=1.5 v=dot  c=blue;
    symbol2 i=none h=1.5 v=dot  c=red;
    symbol3 i=none h=1.5 v=dot  c=yellow;
    legend position=(top right  inside) mode=share label =(h=3 ) value =(h=3);
run;
proc gplot data=cluster_2d;
    plot G3060_ing*G3060_post = cluster / vaxis=axis1 haxis=axis2 legend=legend;
    axis1 label=(h=3 a=90 r=0) value=(h=2);
    axis2 label=(h=3)  value=(h=2);
    symbol1 i=none h=1.5 v=dot  c=blue;
    symbol2 i=none h=1.5 v=dot  c=red;
    symbol3 i=none h=1.5 v=dot  c=yellow;
    legend position=(top right  inside) mode=share label =(h=3 ) value =(h=3);
run;
```