포스트 코로나 온라인 소비패턴 예측 및 온라인 시장 트렌드 조사

CONTENTS

01 주세선정

02 데이터수집 및 처리방법

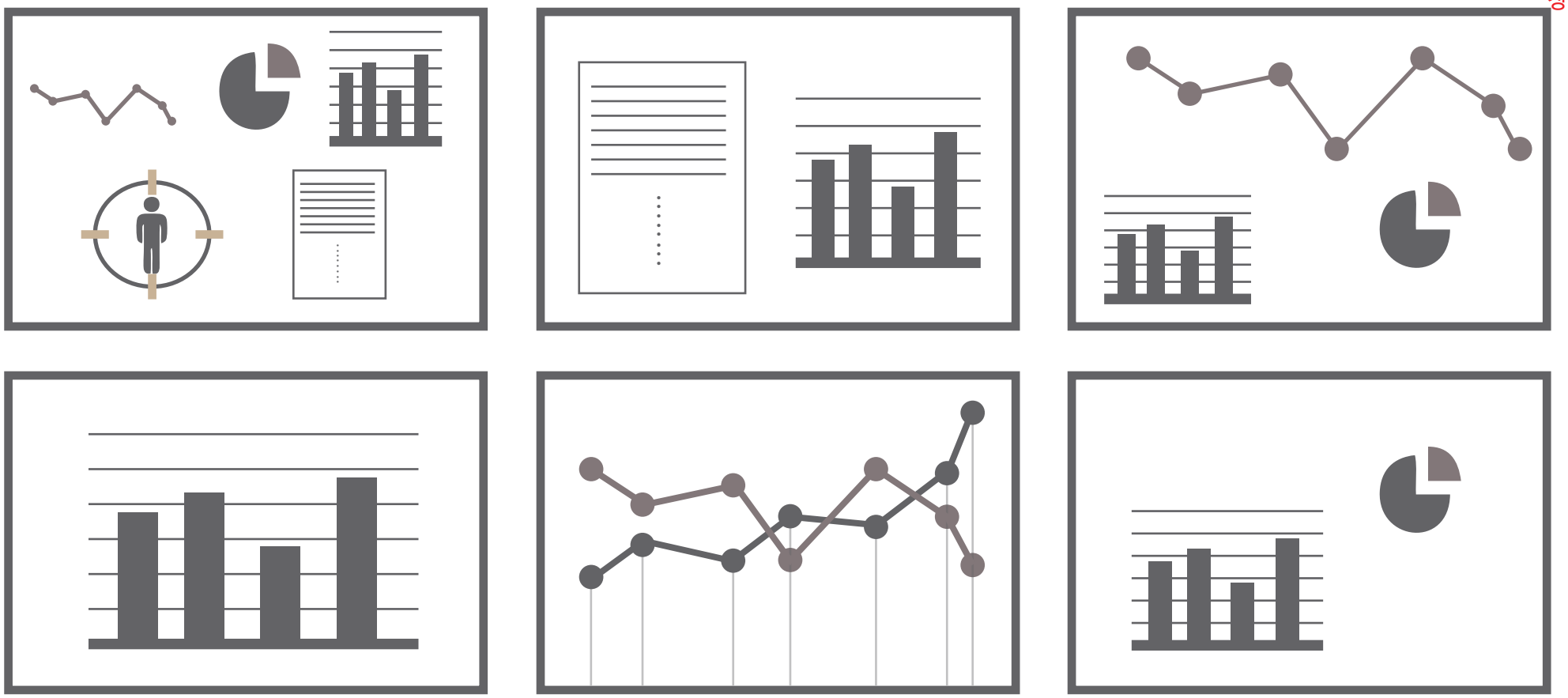
03 EDA 과정 및 결과

04 모형적합 및 성능평가

05 결론 및 한계점

분석목표

온라인 시장 트렌드 조사를 통한 포스트 코로나 시대에 각 산업군별 온라인 소비패턴 변화 예측



> 온라인 시장 사업 진행시 지표로 활용

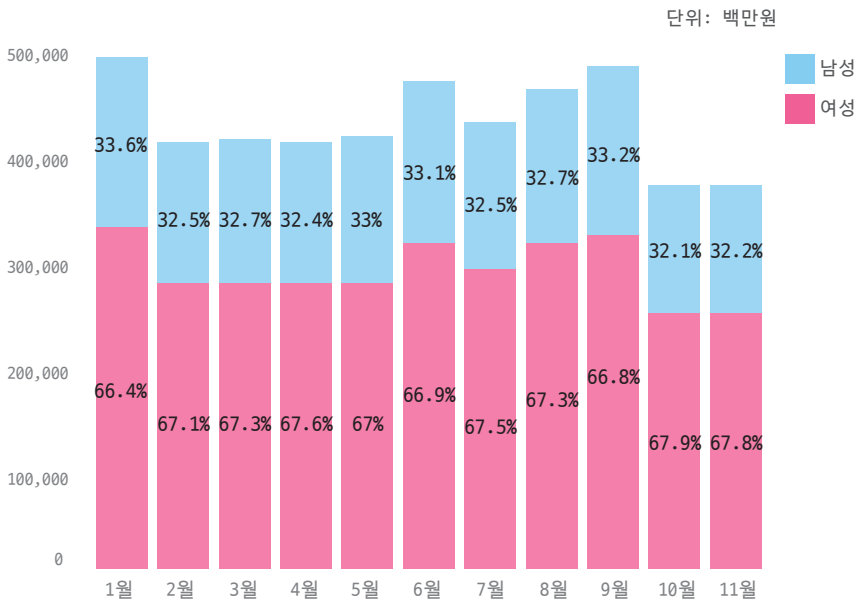
주제선정 배경

출처 : 롯데_오프라인 유통데이터(SDC_통계센터_롯데제공)

온라인 시장, 이베이코리아 관련해서
시킬 수 있는 요소를 넣어보기

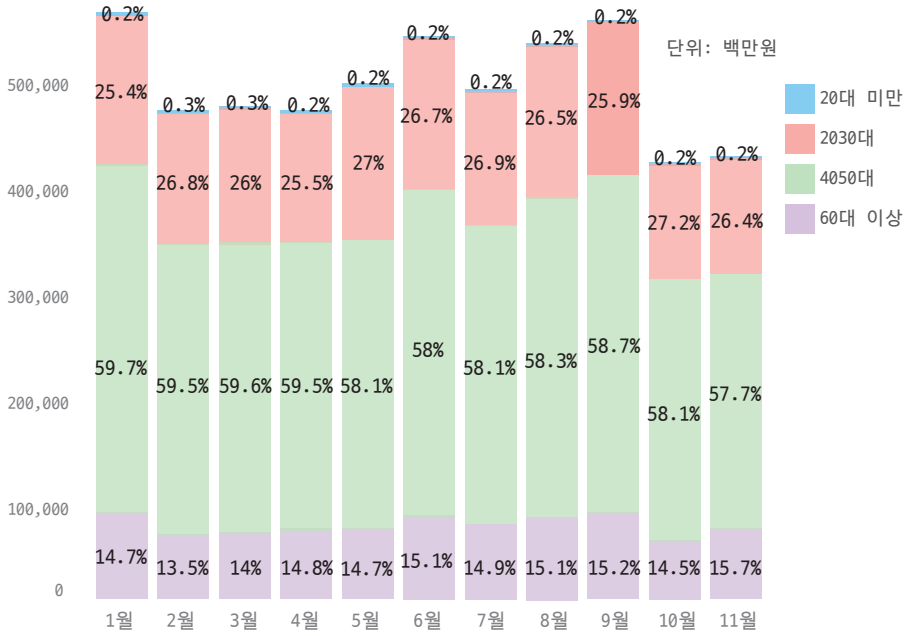
젊은 여성을 공략하는게 마케팅에 중요하다
게 정말 그럴까..? 온라인시장에서도
나? 라는 생각도 들고

성별_매출비율



> 매출비율 남성 1.4% 감소, 여성 1.4% 증가

연령대_매출비율



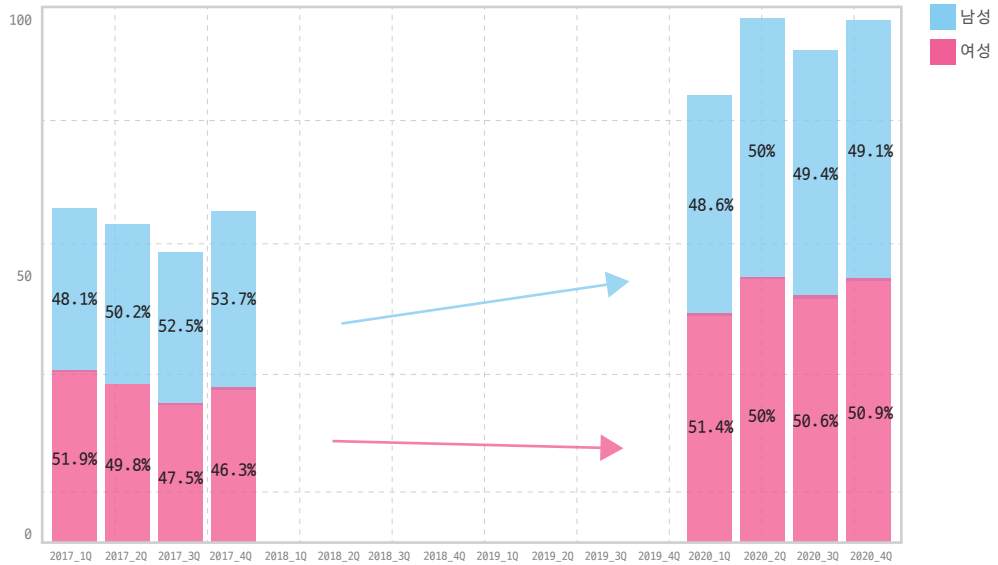
> 매출비율 2030대 1% 증가, 4050대 2% 감소, 60대 이상 1% 증가

매출의 변화 원인이 소비 위축으로 인해 나타나는 것인지 다른 소비 패턴의 변화가 있는 것인지를 파악하고자 하였음

주제선정 배경

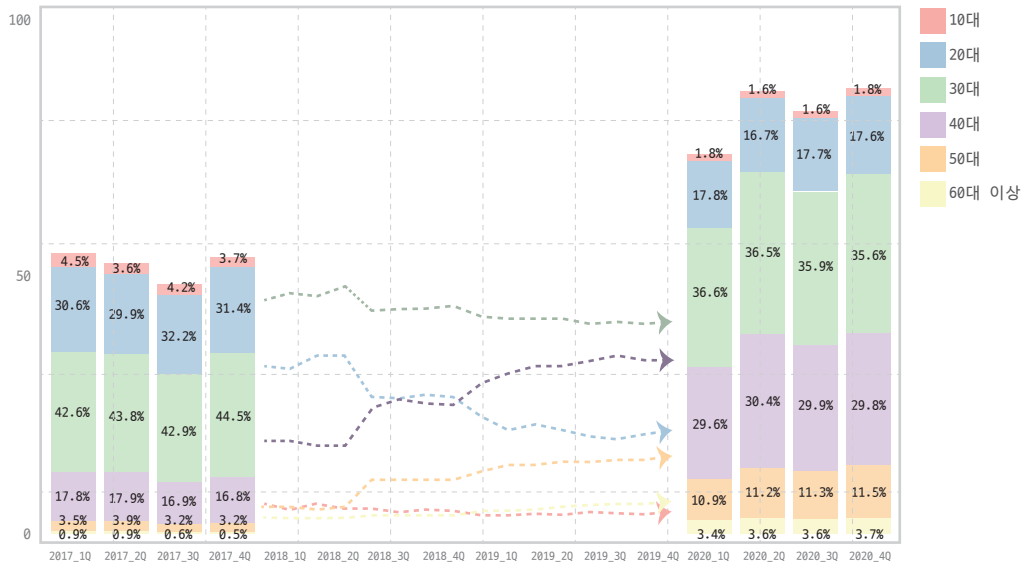
출처 : 네이버 데이터랩 제공(API 데이터)

네이버_데이터랩_성별_온라인 검색추이



> 검색추이비중 남성 9% 증가, 여성 1% 감소

네이버_데이터랩_연령대별_온라인 검색추이



> 검색추이비중 10대 2.7% 감소, 20대 13% 감소, 30대 7% 감소,
40대 12% 증가, 50대 8% 증가, 60대 이상 2% 증가

전체 온라인 검색추이가 점점 상승하는 추세이며, 특히 연령대에서 젊은 층의 비중이 감소하는 동시에 40대 이상의 비중이 모두 증가하는 추세를 보이고 있었다. 이 결과를 통해 온라인 검색 추이가 40대 이상의 신규 유입 등으로 '추측'해볼 수 있었으며, 자세한 리서치 및 통계적 검정을 통해 '트렌드'를 파악하여 향후 각 산업군별 온라인 시장 변화를 예측하고자 하였다.

데이터수집

KOSIS 통계청
https://kosis.kr

온라인쇼핑몰_운영형태별	
년도	월
상품군별	운영형태별

Online몰 / On_Offline몰

온라인쇼핑몰_판매매체별	
년도	월
상품군별	운영형태별

mobile / pc

온라인쇼핑몰_취급상품별	
년도	월
상품군별	운영형태별

종합몰 / 전문몰

경제심리지수
년도
월
경제심리지수

소비자동향조사
년도
월
소비자심리지수

Our World in Data
https://ourworldindata.org

전국_확진자
년도
월
신규확진자_수

네이버데이터랩
https://developers.naver.com

기기별_트렌드조회
기간
비율
group(mo,pc)
카테고리명

성별_트렌드조회
기간
비율
group(F,M)
카테고리명

연령별_트렌드조회
기간
비율
group(age)
카테고리명

SDC통계 데이터센터
https://data.kostat.go.kr

성별_트렌드조회
월
성별
업종구분
이용금액

연령별_트렌드조회
월
연령대
업종구분
이용금액

출처 : 롯데_오프라인 매장 데이터

네이버데이터랩 데이터를 사용함으로써 각 산업군별
타켓 연령층, 성별이 매출액을 예측하는 것에 유의미한지,
어떤 관계가 있는지 파악하고자 하였음 <- 요걸 설명하기 정리해서

기본적으로 KOSIS 통계청에 있는 온라인시장 관련 데이터를 활용하여 전반적인 흐름을 확인하고자 하였고, 코로나데이터를 컬럼으로 추가하여 시장 변화에 대한 예측 정확도와 설명력을 높이고자 하였다. 또한 네이버 데이터랩에 있는 온라인 검색추이를 변수로 채택하여 성별, 연령대에 대한 관심도 및 추세 흐름이 어떻게 변화하고 있는지, 시장에 어떤 영향을 주고 있는지를 확인하고자 했다.

데이터 전처리

1.온라인 시장의 패턴 및 계절성 확인을 위한 '계절','분기' 컬럼 생성

년도	월	판매매체	매출액
2017	01	모바일	345403
2017	01	인터넷	145630
⋮			
2020	12	모바일	569321
2020	12	인터넷	230490

→

년도	월	분기	계절	판매매체	매출액
2017	01	1Q	겨울	모바일	345403
2017	01	1Q	겨울	인터넷	145630
⋮					
2020	12	4Q	겨울	모바일	569321
2020	12	4Q	겨울	인터넷	230490

2. 경기심리지수, 소비자심리지수간의 관계를 파악을 위한 컬럼추가

→

년도	월	분기	계절	판매매체	매출액	경기심리지수	소비자심리지수
2017	01	1Q	겨울	모바일	345403	96.2	93.4
2017	01	1Q	겨울	인터넷	145630	96.2	93.4
⋮							
2020	12	4Q	겨울	모바일	569321	86.1	91.2
2020	12	4Q	겨울	인터넷	230490	86.1	01.2

*경기심리지수 :
*소비자심리지수 :

3. 시장규모의 증가,감소 추세를 예측하기 위한 범주형 컬럼 생성

년도	월	분기	계절	판매매체	매출액	전월대비_증감	경기심리지수	소비자심리지수
2017	01	1Q	겨울	모바일	345403	증가	96.2	93.4
2017	01	1Q	겨울	인터넷	145630	감소	96.2	93.4
⋮								
2020	12	4Q	겨울	모바일	569321	증가	86.1	91.2
2020	12	4Q	겨울	인터넷	230490	감소	86.1	91.2

* 전월대비_증감 = 2017_1월 매출액 - 2017_2월 매출액으로 증가, 감소 컬럼추가

4. 코로나 확진자수 컬럼추가

→

월별	new_case	total_case
1	12	12
2	3139	3139
⋮		
11	8017	34651
12	27117	61761

데이터 전처리

4. KOSIS 통계청, 네이버 데이터랩 데이터 merge를 위한 '중분류' 컬럼 생성

KOSIS
1.컴퓨터 및 주변기기
2.자동차용품
3.가전·전자·통신기기
4.의복
5.신발
6.가방
7.패션용품 및 액세서리
8.스포츠·레져용품
⋮
18. 여행 및 교통서비스
19. 문화 및 레져서비스
20. e쿠폰서비스

중분류
디지털_가전
패션잡화
스포츠_레져
⋮
여가_생활편의

NAVER						
검색기기	카테고리명	기기별_비율	성별	성별_검색비율	연령대	연령대_검색비율
모바일	디지털_가전	30.6%	남성	50.7%	10대	30.6%
인터넷	디지털_가전	10.4%	여성	15.6%	20대	10.4%
모바일	디지털_가전	30.6%	남성	45.7%	30대	38.6%
인터넷	디지털_가전	10.4%	여성	13.6%	40대	17.4%
			⋮			
모바일	여가_생활편의	80.1%	남성	89.7%	30대	17.6%
인터넷	여가_생활편의	25.7%	여성	30.6%	40대	8.9%
모바일	여가_생활편의	80.1%	남성	89.7%	50대	17.6%
인터넷	여가_생활편의	25.7%	여성	30.6%	60대 이상	8.9%

> KOSIS 품목이 21개, NAVER 카테고리가 9개로 상품군별 컬럼이 정확히 일치하지는 않아 추가적으로 중분류컬럼을 생성하여 merge

데이터 전처리

5. 성별, 연령대에 대한 영향력 측정 및 목표변수 중복값 방지를 위한 '선호_성별', '선호_연령대'로 값 추출 후 merge

NAVER

검색기기	카테고리명	기기별_비율	성별	성별_검색비율	연령대	연령대_검색비율
모바일	디지털_가전	30.6%	남성	50.7%	10대	30.6%
인터넷	디지털_가전	10.4%	남성	15.6%	20대	10.4%
모바일	디지털_가전	30.6%	⋮	50.7%	30대	20.6%
인터넷	디지털_가전	10.4%		15.6%	40대	17.4%
					50대	12.6%
					60대 이상	10.4%
모바일	여가_생활편의	30.6%	여성	50.7%	10대	30.6%
인터넷	여가_생활편의	10.4%	여성	15.6%	20대	10.4%
모바일	여가_생활편의	30.6%		50.7%	30대	20.6%
인터넷	여가_생활편의	10.4%		15.6%	40대	11.4%
					50대	30.6%
					60대 이상	10.4%

KOSIS

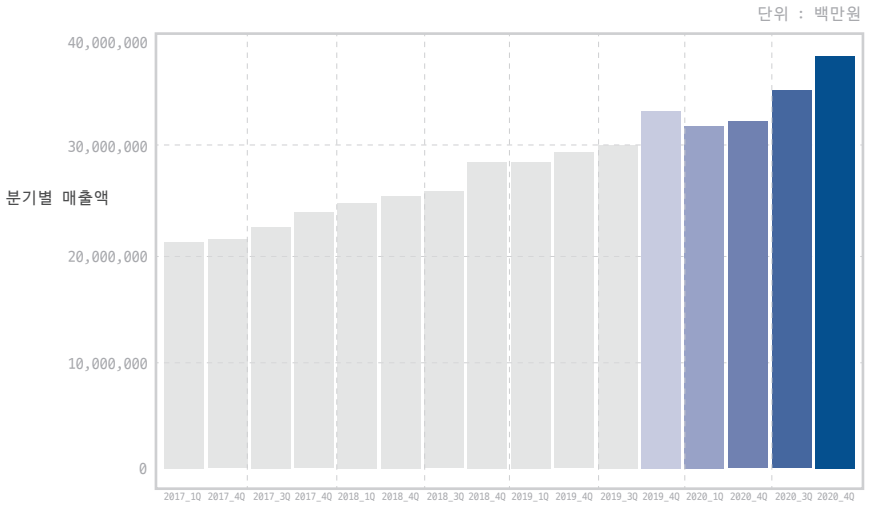
년도	월	분기	계절	판매매체	매출액
2017	01	1Q	겨울	모바일	345403
2017	01	1Q	겨울	인터넷	145630
			⋮		
2020	12	4Q	겨울	모바일	569321
2020	12	4Q	겨울	인터넷	230490

.....

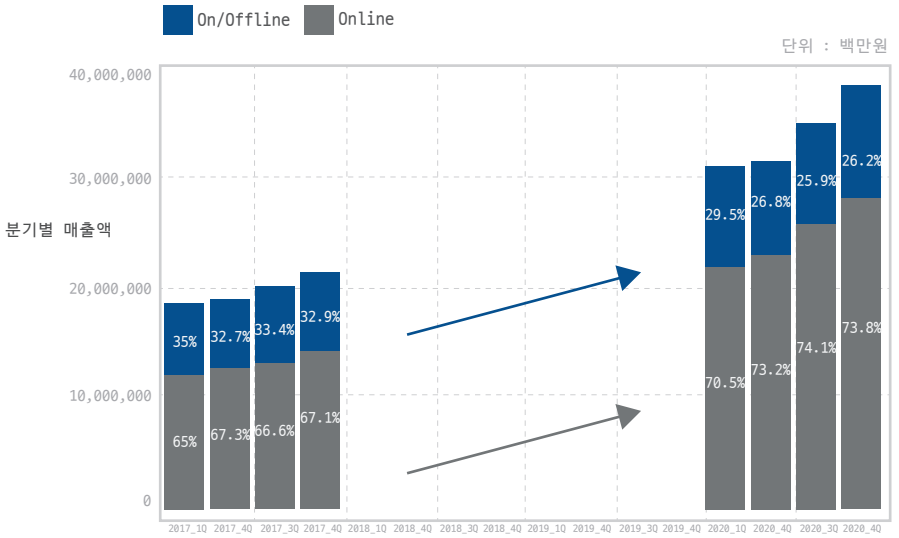
*검색비율(상대적 비율)
2017.1 ~ 2020.12 검색추이 요청시 가장 검색량이 많았던 '월'을 100을 기준하여 나머지 '월'에 검색비율 수치가 생성

> 선호 성별, 선호 연령대 값을 추출함으로써 각 분야별로 높은 수치를 보이고 있는 타켓층이 매출액에 대해 유의미함을 검증하고자 하였음

EDA 과정_1



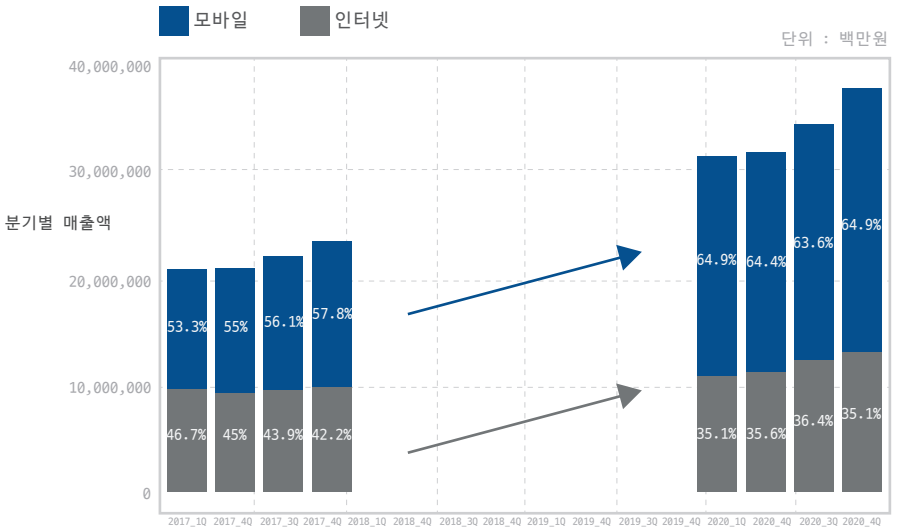
> 2017년도 1분기 기준 2020년 온라인 시장 매출액 64% 증가



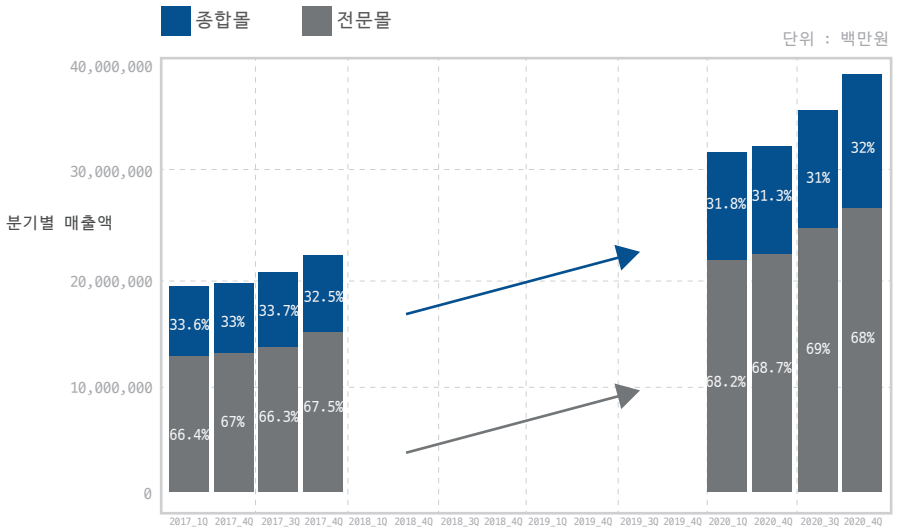
> Online으로만 운영하는 형태가 비중이 8.8% 증가

> 온라인 시장 자체가 크게 확대된 상태에서 운영형태가 어떤 변화를 보여주고 있는지 알 수 있었음

EDA 과정_2



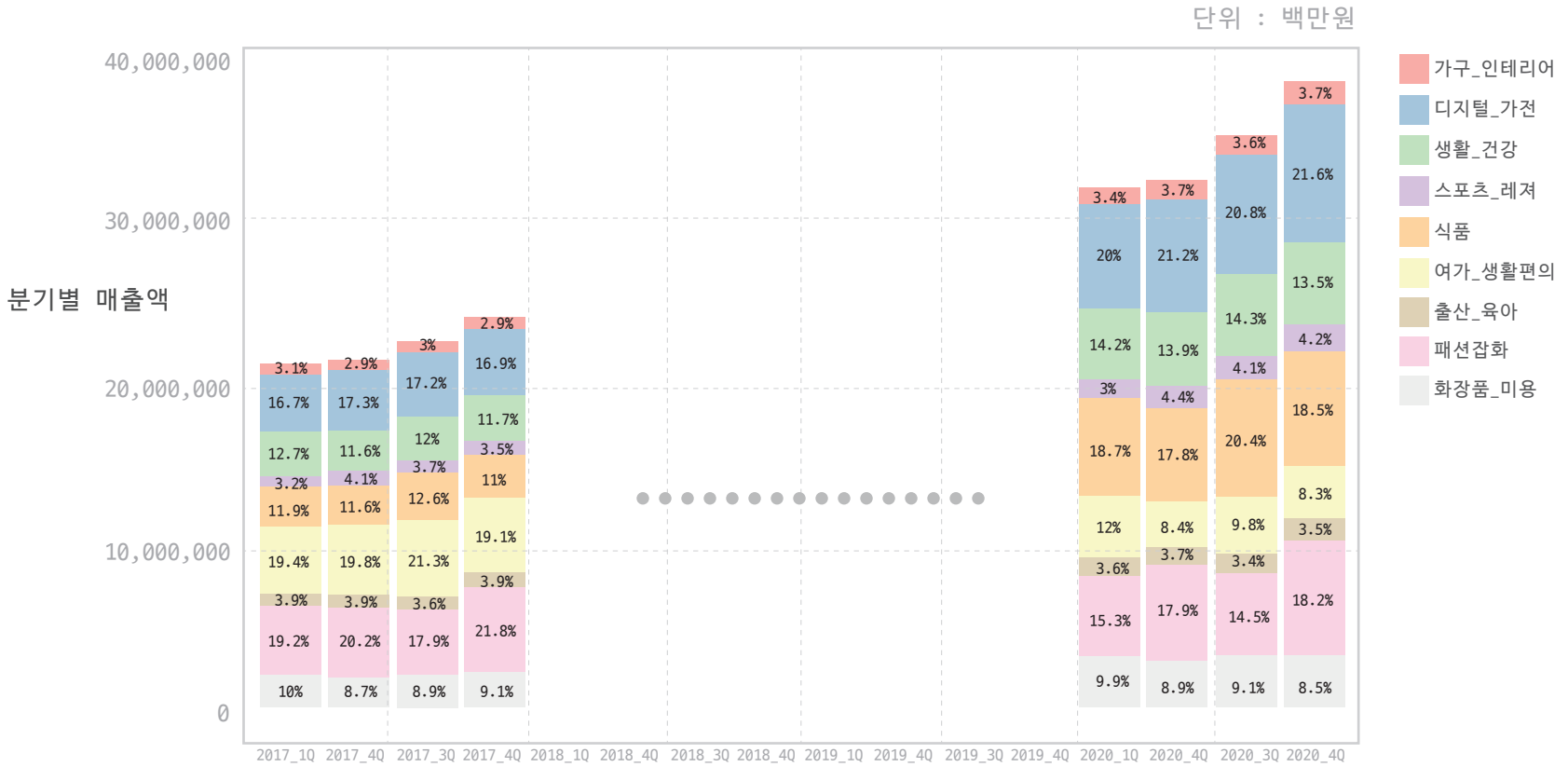
> 모바일 비중이 2017년 기준 11.6% 증가, 인터넷 11.6% 감소



> 취급형태별 매출액은 전문몰이 1.6% 증가

> 시장 자체가 크게 확대된 상태에서 모바일의 비중이 크게 증가함을 알 수 있었고, 취급형태별 매출 비중은 변화폭이 크지 않았음

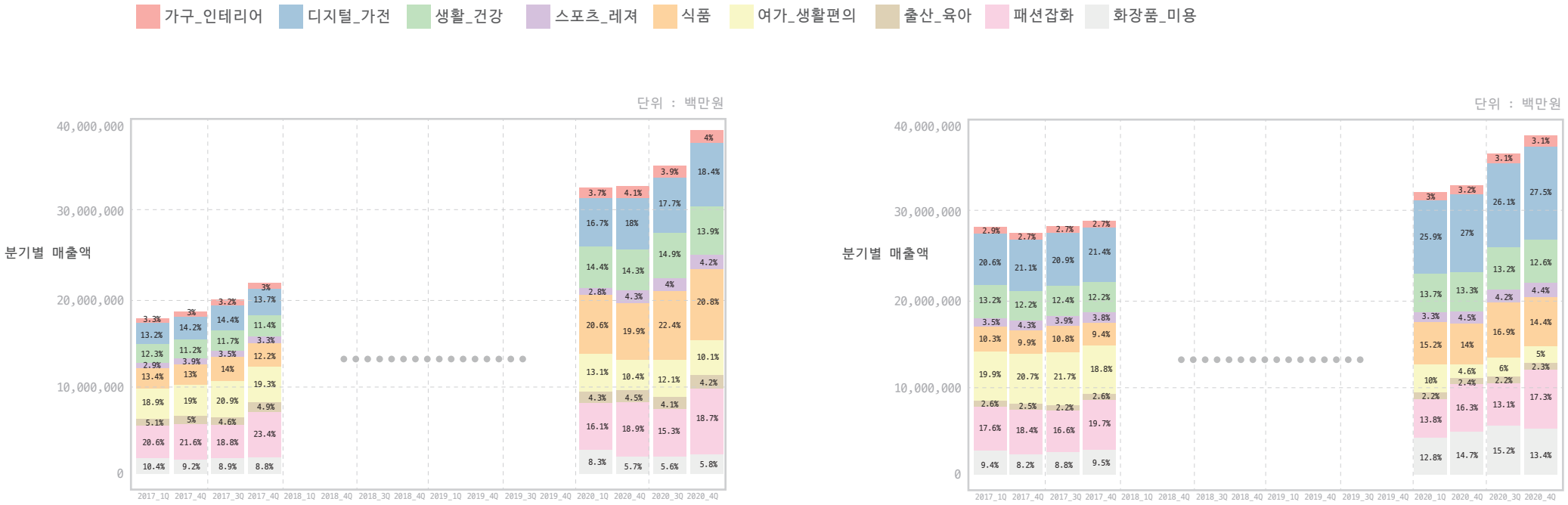
EDA 과정_3



온라인_전체시장

> 산업군별 비중을 살펴보았을때 식품 7.6% 증가, 여가_생활편의가 11.3% 감소로 가장 많은 변화를 보여주고 있었다.

EDA 과정_3



모바일

1위 식품, 2위 패션잡화, 3위 디지털_가전

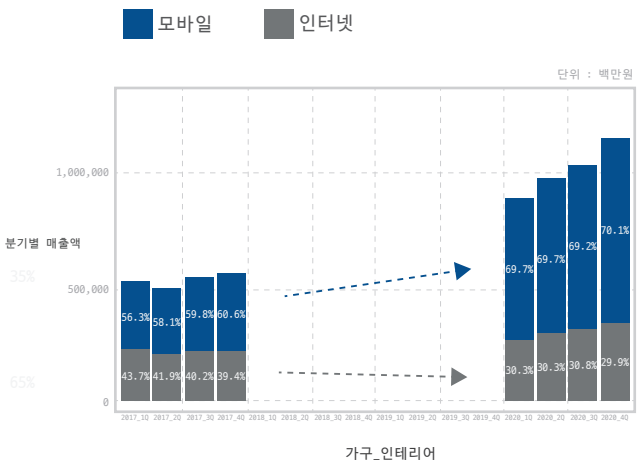
인터넷

1위 디지털_가전, 2위 패션잡화, 3위 식품

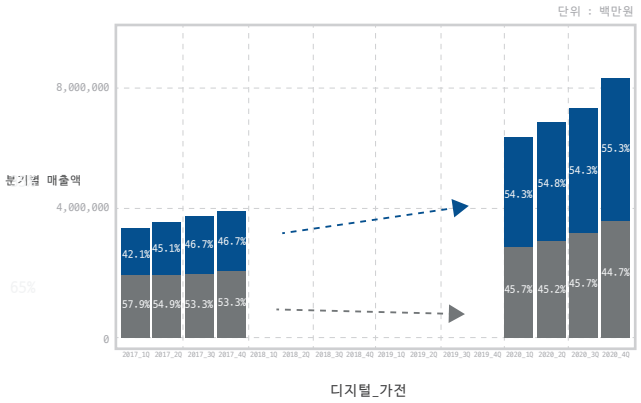
> 모바일과 인터넷을 구분지어 확인해보았을 때 매출비율에 대한 산업군별 순위의 차이가 있었음을 알수있었음

EDA 과정_4

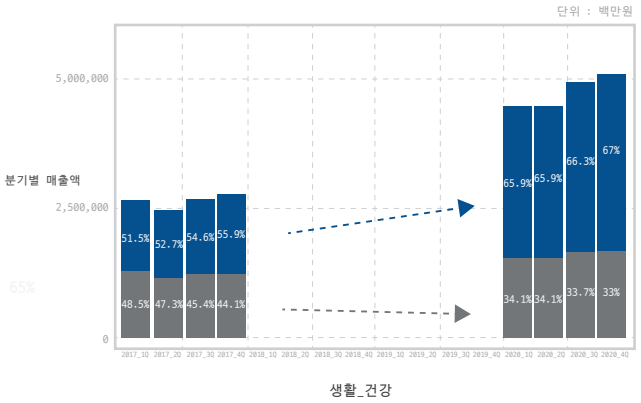
모바일, 인터넷 매출액 비중



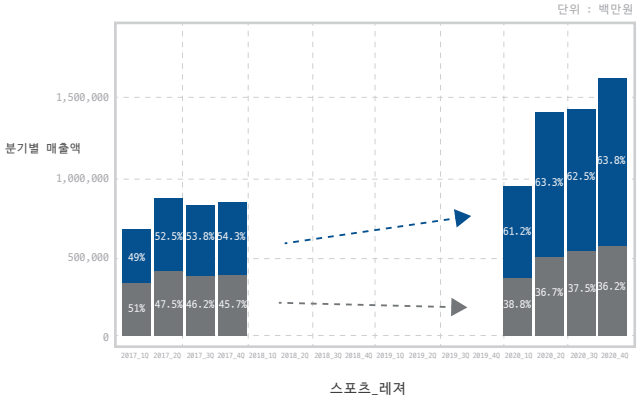
가구_인테리어



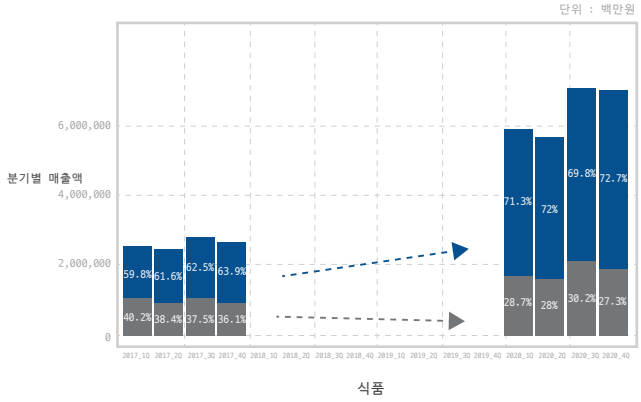
디지털_가전



생활_건강

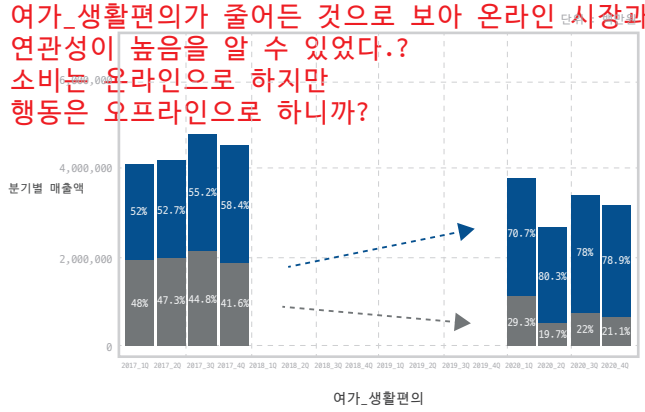


스포츠_레져

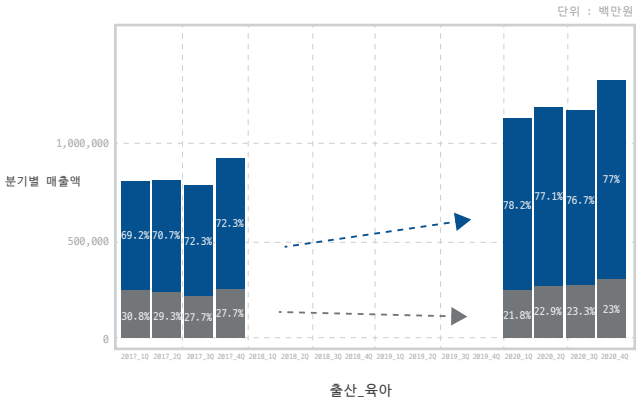


식품

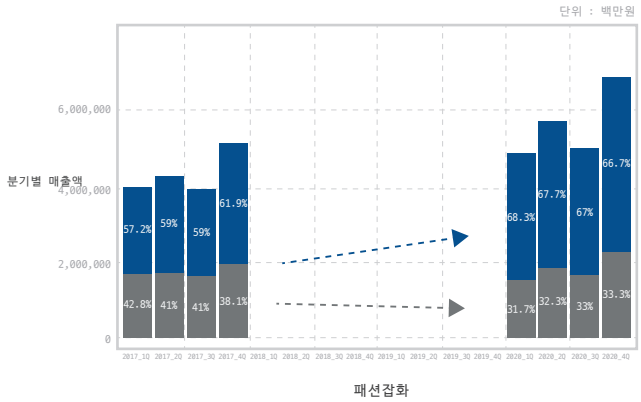
여가_생활편의가 줄어든 것으로 보아 온라인 시장과 오프라인
연관성이 높음을 알 수 있었다.?
소비는 온라인으로 하지만
행동은 오프라인으로 하니까?



여가_생활편의

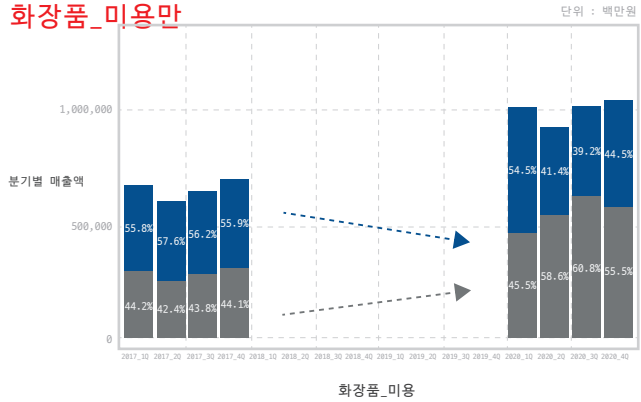


출산_육아



패션잡화

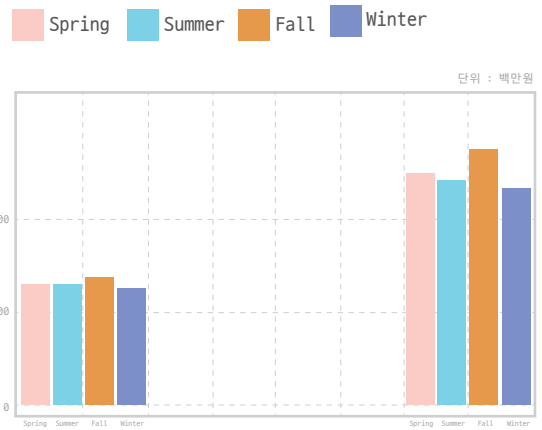
화장품_미용만



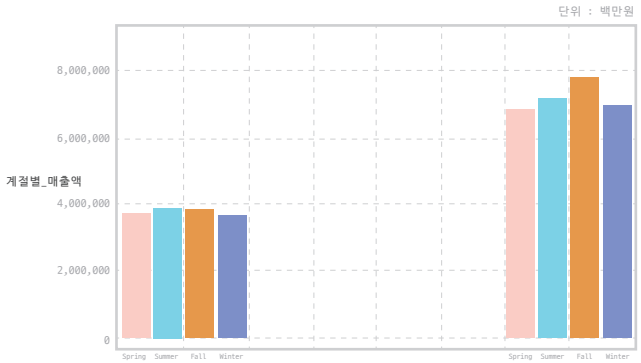
화장품_미용

EDA 과정_5

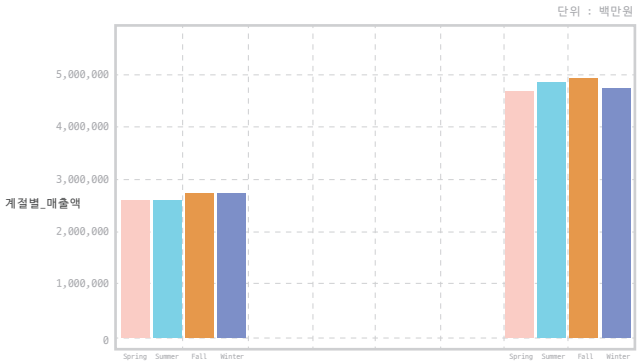
패션잡화, 식품정도를 제외하고는 크게 계절성을 볼수는 없었지만
컬럼추가하여 매출액에 중요도를 확실하게 확인!!#@$@#$



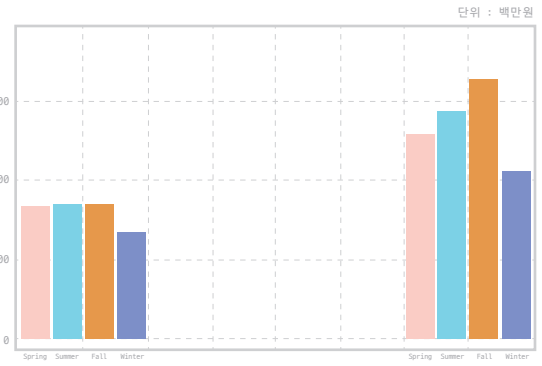
가구_인테리어



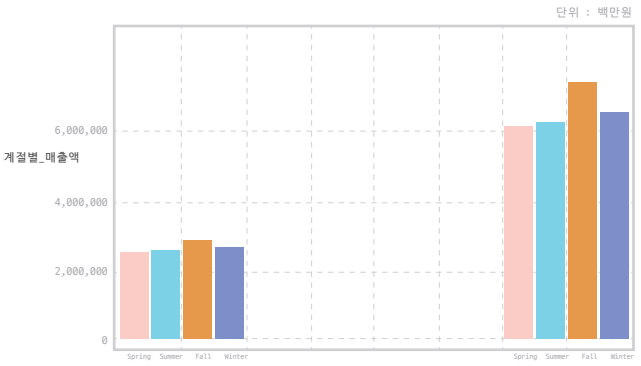
디지털_가전



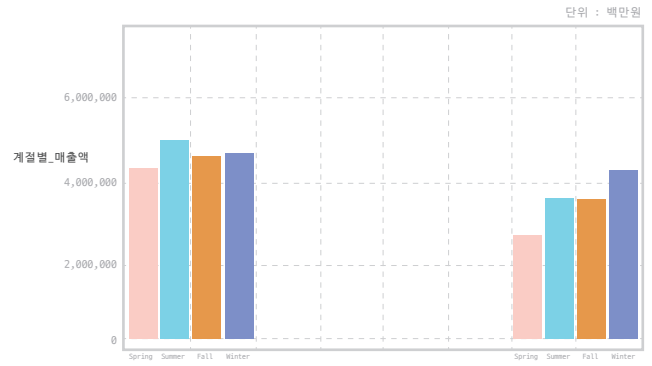
생활_건강



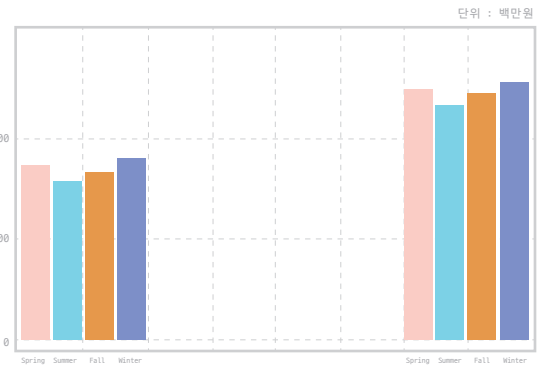
스포츠_레져



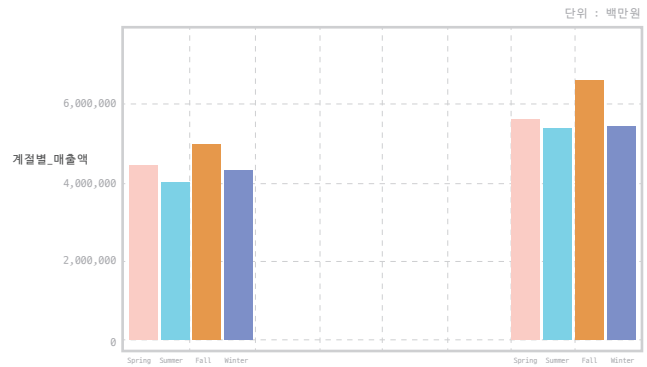
식품



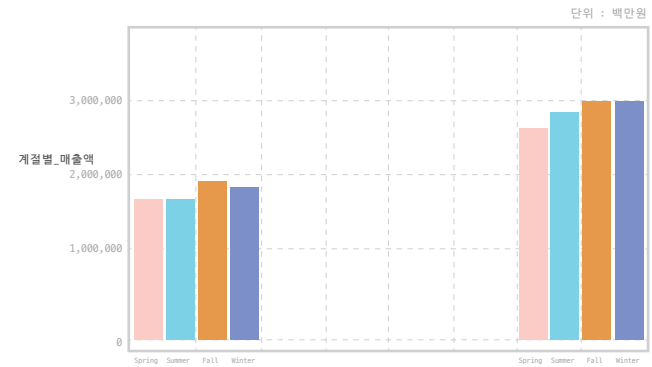
여가_생활편의



출산_육아

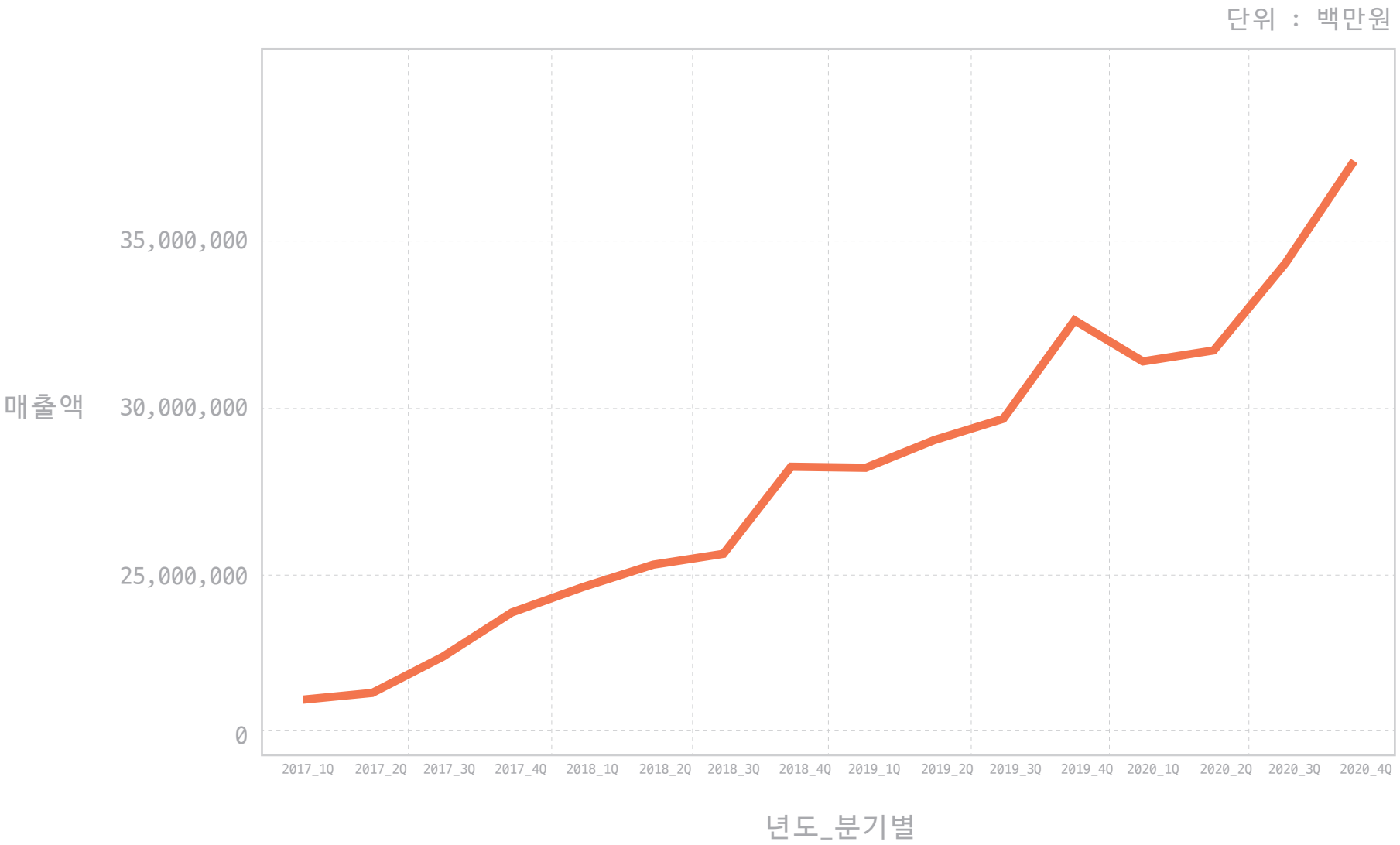


패션잡화



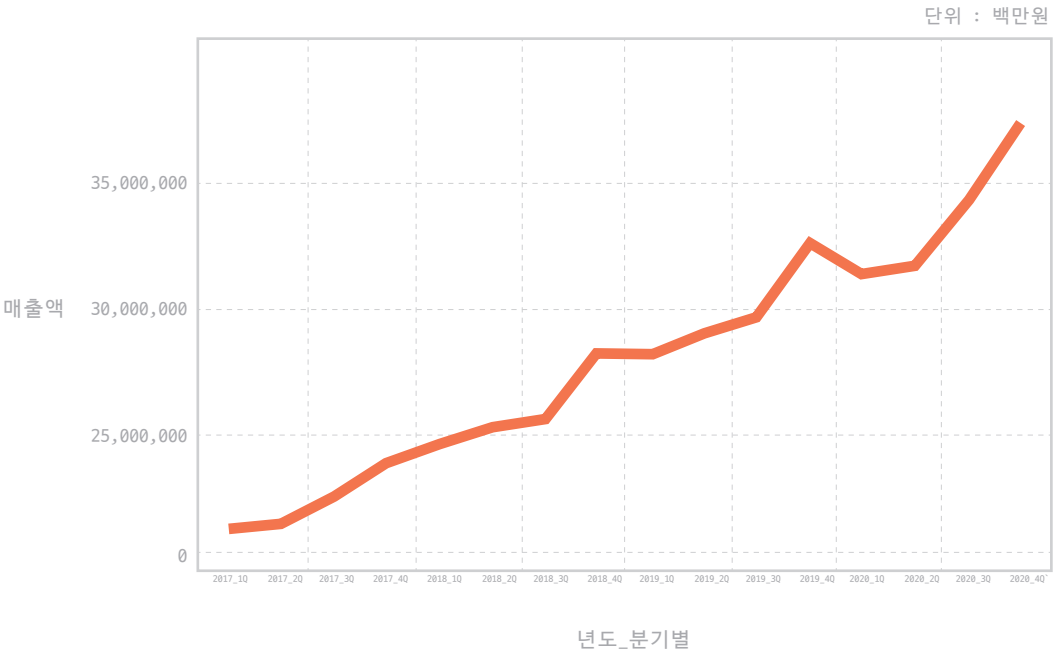
화장품_미용

EDA 과정_6



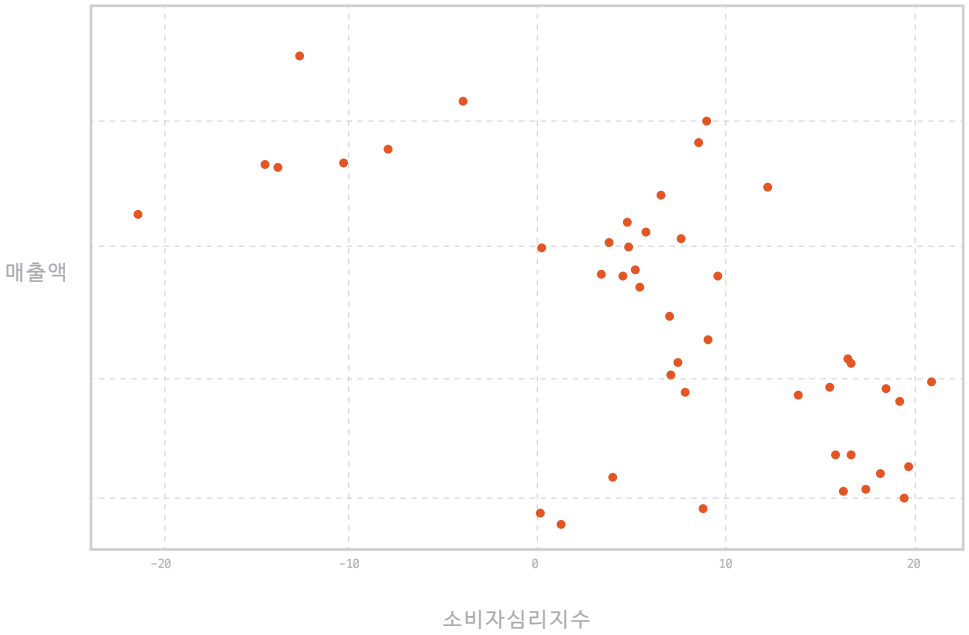
> 시장 자체가 크게 확대된 상태에서 모바일의 비중이 크게 증가함을 알 수 있었고, 취급형태별 매출 비중은 변화폭이 크지 않았음

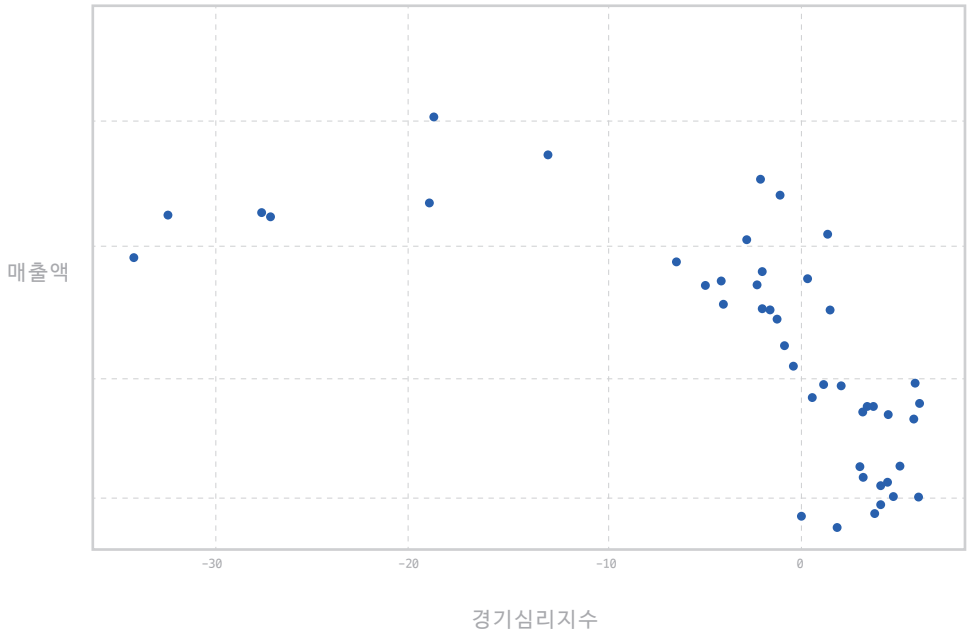
EDA 과정_7



> 매출액에 대한 변화가 경기가 좋아져서 사람들의 소비가 커진건지 확인하기위해 소비자심리지수, 경기심리지수와 함께 확인을 해보았으며, 그 결과 상반되는 패턴을 보여주고 있었음

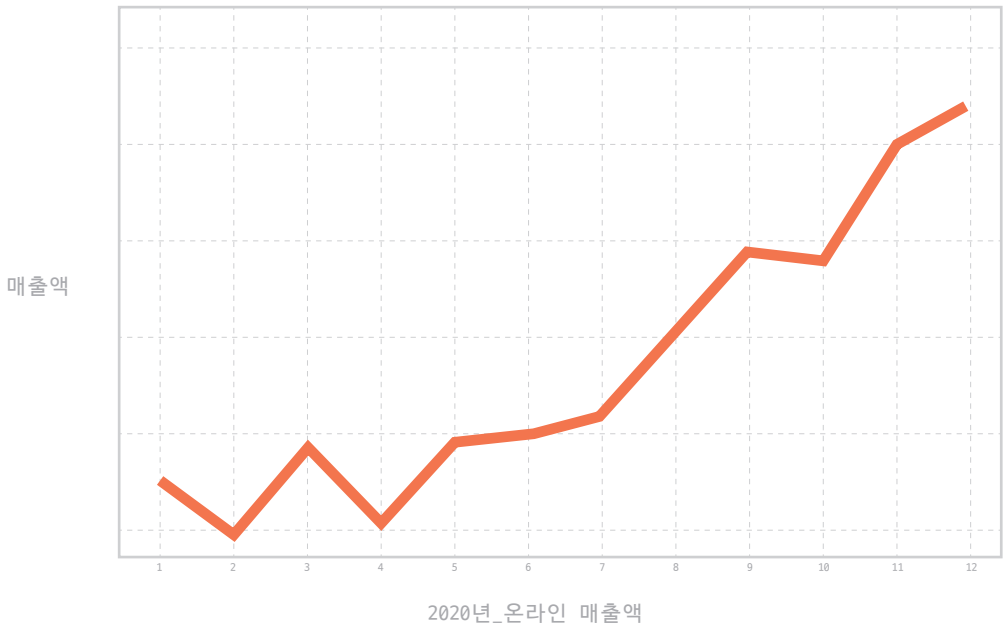
EDA 과정_8

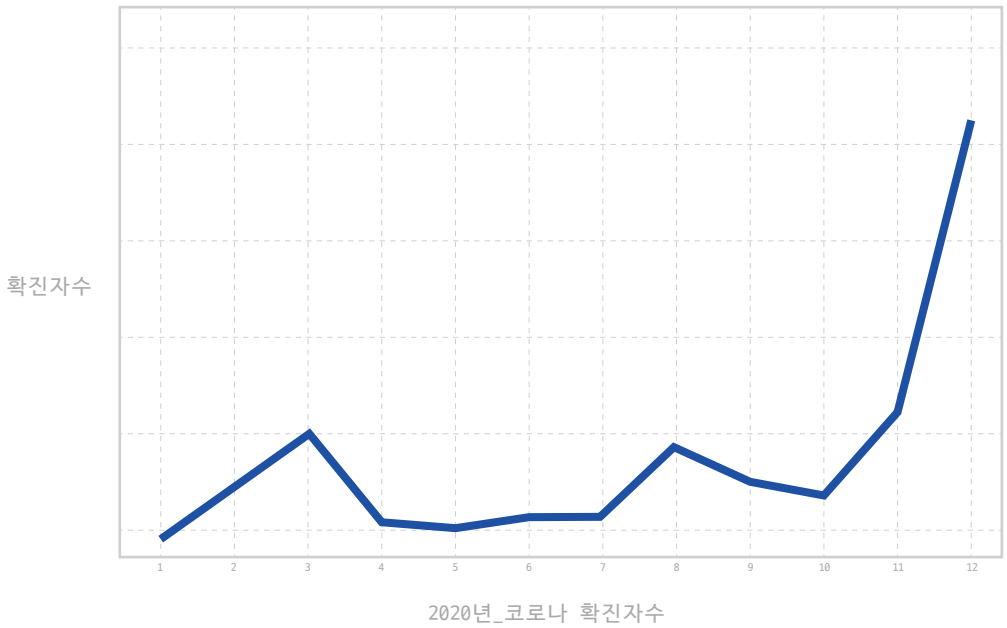




> 조금 더 자세하게 보기위해서 'point plot'으로 그려보았으며, 약간의 음의 선형관계가 나타남을 볼 수 있었음

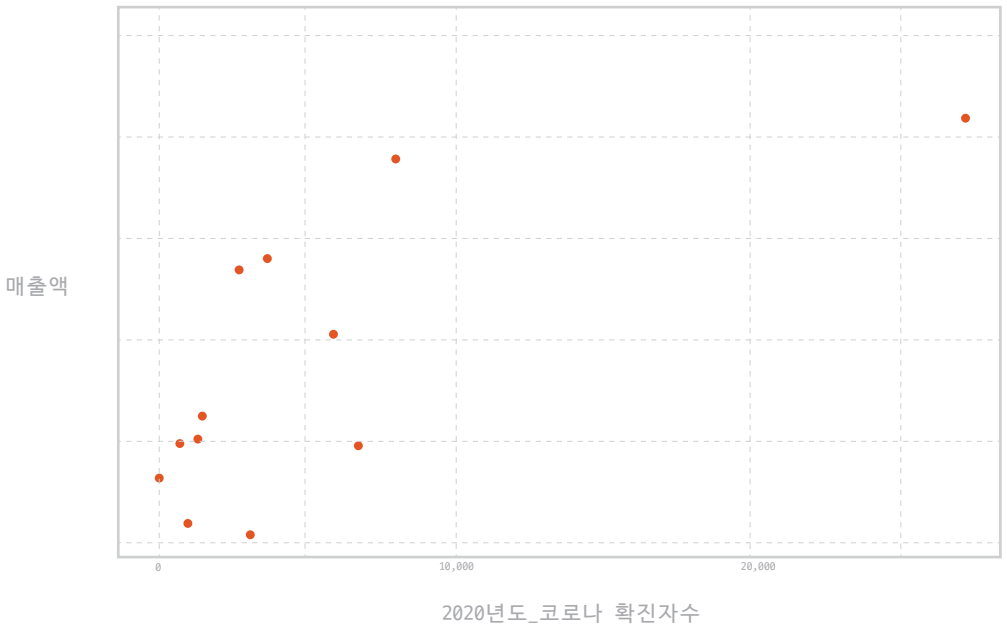
EDA 과정_9

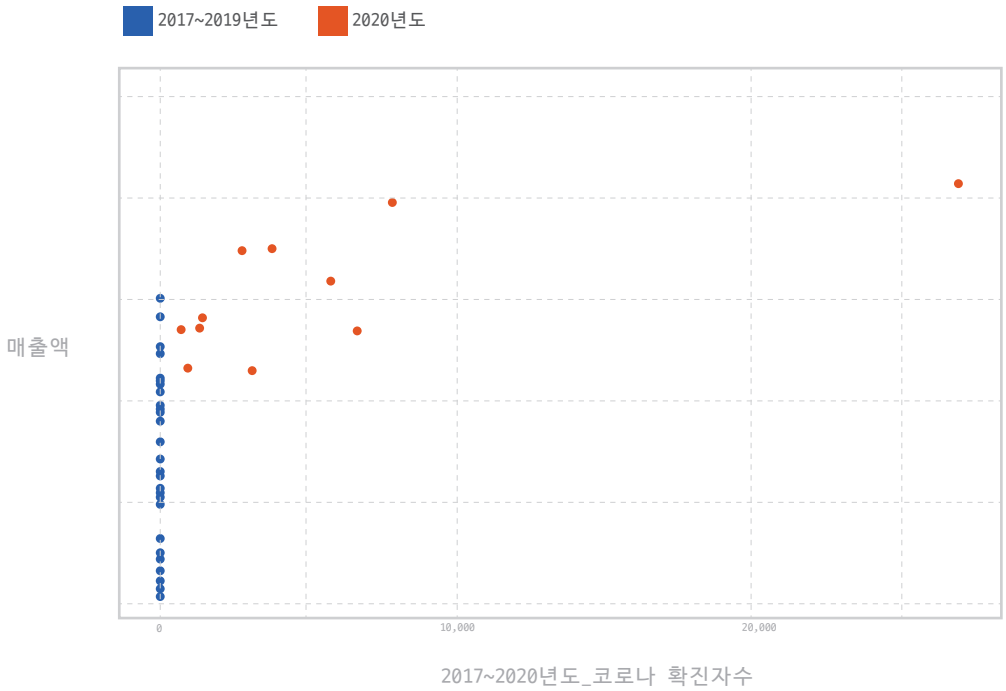




> 코로나와 온라인매출의 관계를 살펴보았고, 자세하게 살펴보기 위해서 우선 2020년도 데이터로만 살펴보았다.
> 둘다 상승을 하고 있었으나 추세가 다르게 보았으며, point plot을 통해 자세하게 보고자 하였다.

EDA 과정_10

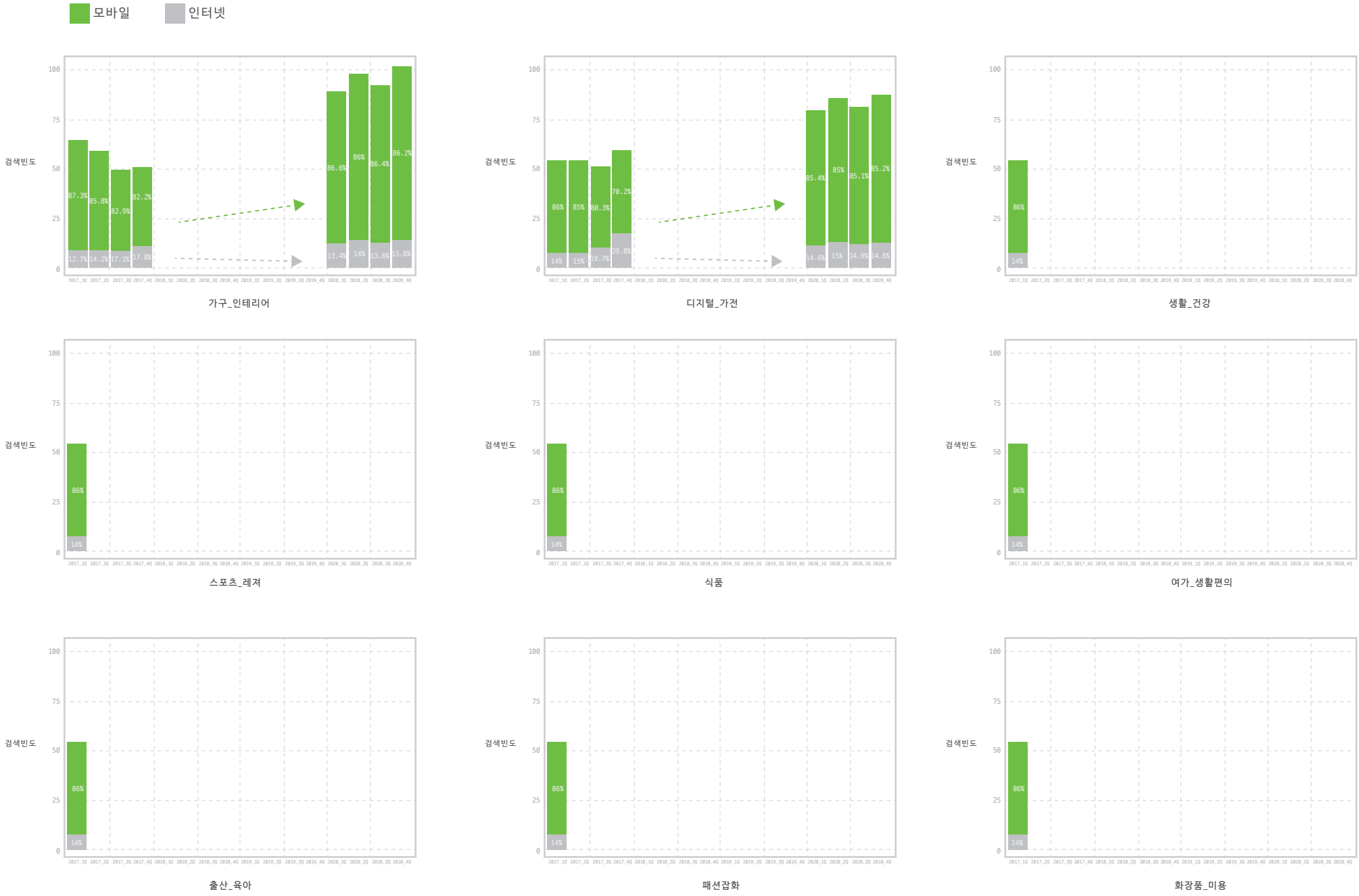




> 2020년도로만 보았을 때 상향하는듯한 모습을 보여주지만 2017~2020년도 범위로 그래프를 나타내보았을 때
이미 시장이 고도화된 상태에서 코로나 변수가 어떤 영향을 주었는지 확인할 필요가 있다고 판단하여 변수로 채택하였음

EDA 과정_10

네이버데이터랩_모바일, 인터넷 검색추이_비중/ 뭐로 더 많이 검색하냐



EDA 결론

앞선 EDA과정을 통해서 온라인과 오프라인의 변화추이와 각 산업군별로 어떤 특징을 보이고 있는지 알 수 있었다.

또한 그 특징들과 매출액간의 선형관계를 보기도 하였는데 EDA에서 발견했던 부분들에 대한 '추측'들이 통계적으로 어떤 의미를 가지고 있는지 모델적합을 통해 확인해보자 하였다.

모형적합

모형적합은 네이버데이터랩에 대한 기기별, 성별, 연령대별 검색추이에 대한 변수의 유의미성을 파악하기 위해

KOSIS DATA 와 KOSIS + NAVER_DATA 두 개의 데이터셋을 만들어서 모형을 적합해보았다.

모형적합

KOSIS_DATA

의사결정나무 : RMSE_182041.3

KOSIS+NAVER_DATA

의사결정나무 : RMSE_134467.1

KOSIS DATA와 NAVER_DATA를 MERGE하여 데이터셋한 결과가 의사결정나무 기준 26% 높은 예측율을 보여주었다.

모형적합

RandomForest

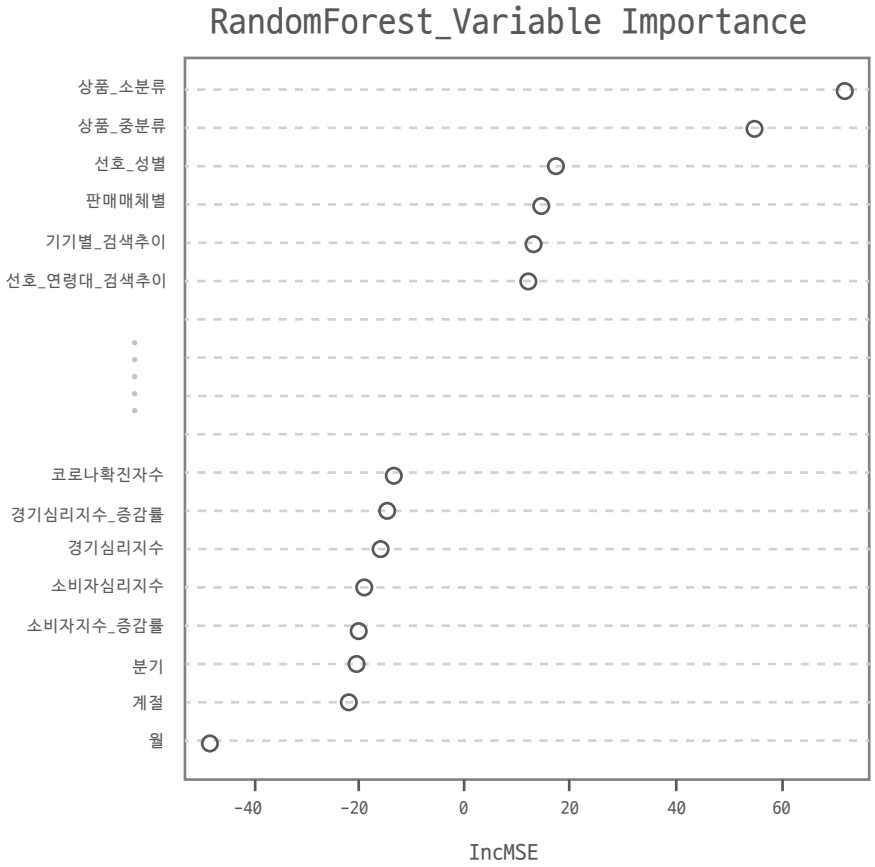
```
library(randomForest)
set.seed(seed = 1234)
fit2 <- randomForest(formula = 매출액 ~.,
                     data = naver_train_data,
                     ntree = 1000,
                     mtry = 3,
                     importance = TRUE,
                     do.trace = 50,
                     keep.forest = TRUE)

real <- naver_test_data$매출액

pred2 <-predict(fit2, newdata = naver_test_data, type = 'response')
RMSE(real, pred2)

importance(fit2, type = 1)
varImpPlot(fit2, main = 'Variable Importance', type = 1)
```

Decision_Tree

```
library(rpart)
Ctrl <- rpart.control(minsplit = 20,
                      cp = 0.01,
                      maxdepth = 10)


set.seed(seed = 1234)
fit3 <- rpart(formula = 매출액~.,
              data = naver_train_data,
              control = Ctrl)


real <- naver_test_data$매출액
pred3 <- predict(object = fit3, newdata = naver_test_data,
                      type = 'vector')

RMSE(real, pred3)
```

최종적으로 KOSIS_DATA + NAVER_DATA를 MERGE한 데이터셋으로 모델을 비교해보았으며, 랜덤포레스트와 의사결정나무로 모형적합을 해본 결과 의사결정나무 모델이 64% 높은 예측율을 보여주었다.

모형적합



Decision_Tree_Variable Importance

Variable Importance				
상품_소분류별 59	상품_중분류별 18	기기별_검색추이 13	판매매체별 4	선호_성별_검색추이 2
선호_연령대_검색추이 1	선호_성별 1	경기심리지수 1	심리지수_증감률 1	코로나확진자수 1

랜덤포레스트와 의사결정나무에서 보여준 변수의 중요를 살펴보았으며, 상품군별 변수가 가장 중요하게 나타났으며 기기별 성별, 연령대별 등 상위권 중요변수에 대한 결과가 유사하게 나타났음을 알 수 있었다.

결론

공통적으로 상품군별 소분류, 중분류에 대해서 중요도가 상위권으로 나타나는것으로보아 온라인시장에서 강하게 나타나는 상품이 있는것으로 생각해볼 수 있었음 하지만 상품군별 매출액 스케일과 온라인에서의 각 상품의 비중을 반영하지는 못해서 특정 상품이 온라인시장에서 '매출이 높다 ' 라기보단 '활성화 ' 가 잘되어있다로 판단할 수 있었음

또한 성별, 연령대에 대한 변수가 높게 나오는것으로보아 적절한 타켓팅에 대한 전략이 유효할 것이라는 것을 추측해볼 수 있었으며, EDA과정에서 나타났던 신규유입으로 판단되는 40대에 대한 적절한 마켓팅 및 전략가이드가 필요할 것으로 생각하였음

결론적으로 온라인 시장이 확대되면서 주타켓층이였던 20대 못지않게 향후 40대에 대한 영향력이 지속적으로 높아질 것으로 보이면서 ( )해야할 것이다.

한계점

1. KOSIS DATA와 NAVER DATA의 표본조사 대상이 다르기 때문에 데이터를 MERGE함에 있어서 100% 신뢰하기는 어려웠다.

2. 추가변수들로 사용한 NAVER DATA가 소비자에 대한 로그기록, 각각의 매출기록처럼 직접적인 데이터가 아닌 상대적비율 데이터이기 때문에 구체적인 인사이트를 도출하지 못한것이 아쉬웠다.

3. 오프라인 데이터가 2017년도부터 있다면 변화추세를 비교할 수 있었을텐데 2020년 데이터로 진행하다보니 한계가 있었다.

감사합니다.