📌 40 · 50대 소비패턴 분석을 통한 공연예술 관객층 확대 방안 탐구

팀 디어마이프렌즈 정세영 연주원 이승원 정유지

Table of Contents

목차

001 공모배경

①문제점 분석

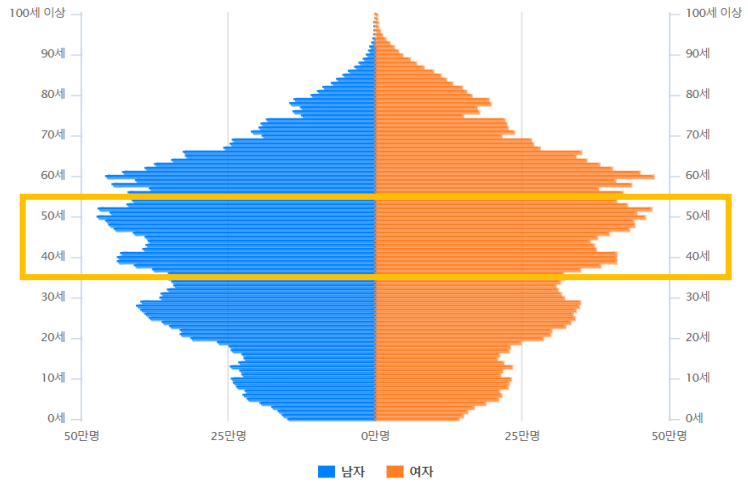
저렴한 가격과 다양한 콘텐츠를 앞세운 유튜브·넷플릭스 등
'가성비' 스트리밍 서비스 공세에 공연장을 찾는 20대 관객이 줄고,
구매력이 높은 30·40대가 이들의 빈자리를 채우고 있다.

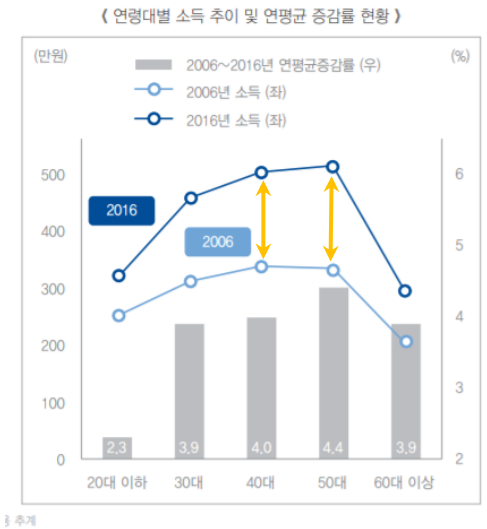
문화예술공연 관람률

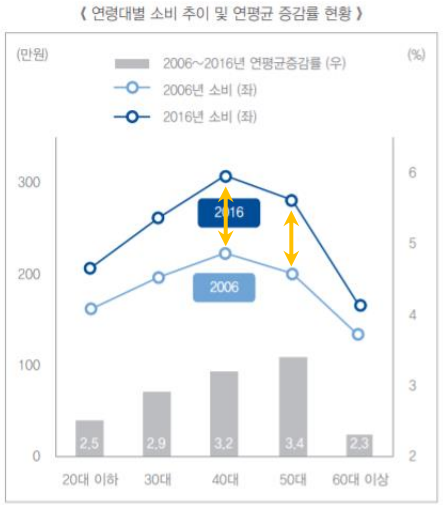
	2009		2011		2013		2015		2017		2019	
전체	52.4	⇔(0%)	54.5	▲(4%)	60.8	▲(12%)	64.5	▲(6%)	64.0	▲(-1%)	63.6	▲(-1%)
남자	50.5	⇔(0%)	51.5	▲(2%)	58.5	▲(14%)	62.0	▲(6%)	61.6	▲(-1%)	61.1	▲(-1%)
여자	54.2	⇔(0%)	57.4	▲(6%)	62.9	▲(10%)	66.9	▲(6%)	66.3	▲(-1%)	66.1	▲(0%)
20세 미만	77.2	⇔(0%)	77.9	▲(1%)	82.6	▲(6%)	84.5	▲(2%)	86.0	▲(2%)	83.8	▼(-3%)
20-29세	79.6	⇔(0%)	78.2	▼(-2%)	83.4	▲(7%)	83.8	▲(0%)	83.8	⇔(0%)	82.8	▼(-1%)
30-39세	68.2	⇔(0%)	70.6	▲(4%)	77.2	▲(9%)	79.2	▲(3%)	78.6	▲(-1%)	79.8	▲(2%)
40-49세	53.4	⇔(0%)	58.7	▲(10%)	67.4	▲(15%)	73.2	▲(9%)	73.7	▲(1%)	74.4	▲(1%)
50-59세	35.0	⇔(0%)	41.2	▲(18%)	48.1	▲(17%)	56.2	▲(17%)	58.0	▲(3%)	58.9	▲(2%)
60세 이상	13.4	⇔(0%)	16.6	▲(24%)	21.7	▲(31%)	28.9	▲(33%)	29.1	▲(1%)	31.2	▲(7%)

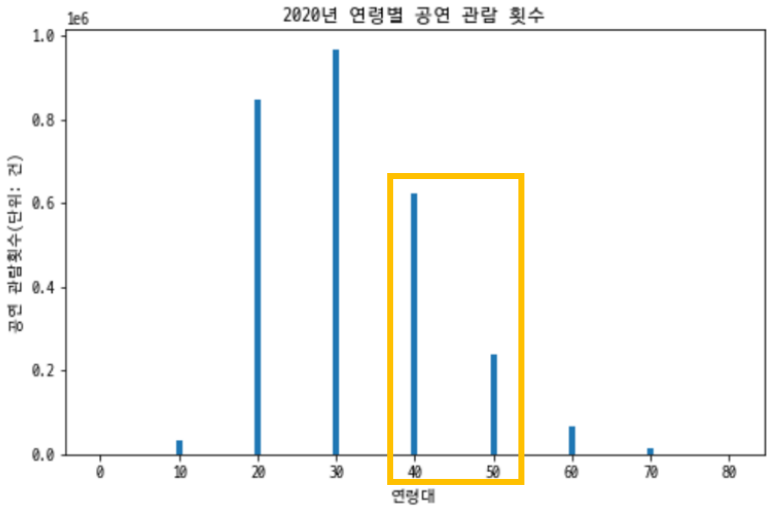
기존 공연예술 시장의
주 타겟이었던 20대 관객이 줄어들고 있다.

②해결방안



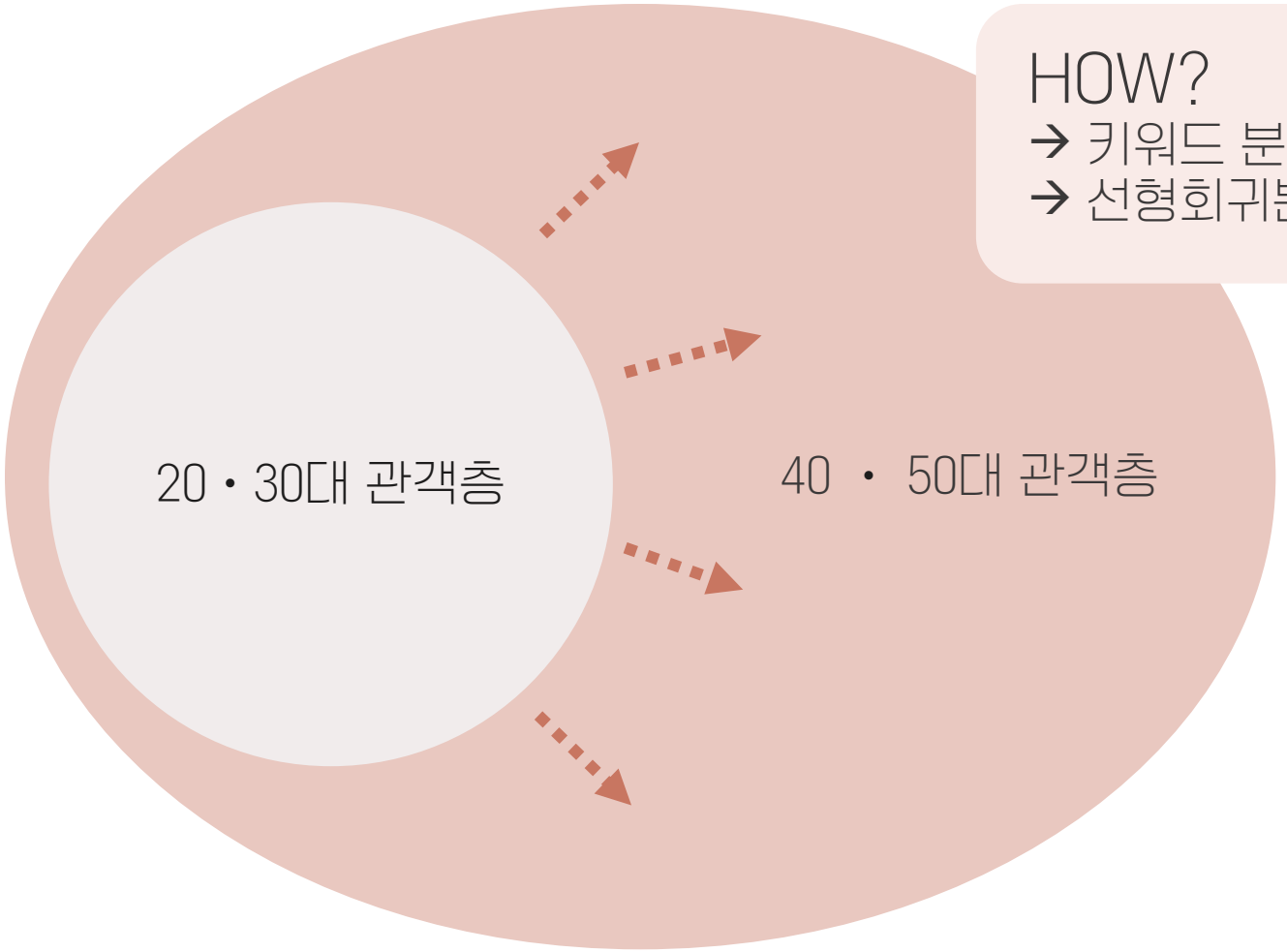






➔ 40・50대 관객층 유입을 통한 공연예술계의 새로운 비즈니스 기회 창출

③차별성 및 독창성



HOW?
→ 키워드 분석
→ 선형회귀분석

002 분석 방법 및 결과

①사용한 데이터

키워드 분석

Ⅰ. KOPIS 데이터

연령이 40·50대인 데이터와 네이버 예약으로 예매한 데이터 추출

Ⅱ. 네이버 예매자 리뷰 데이터

다중회귀분석

Ⅰ. KOPIS 데이터

출생연도를 이용하여 연령대가 40・50대인 데이터 추출한 후, 하루단위로 예매건수가 몇 개인지 집계

Ⅱ. 삼성카드 카드소비 변화 데이터

	소비일자	소비업종	성별	연령대	소비건수합계
0	20150629	편의점	여성	50대	77585
1	20200501	편의점	남성	40대	570937
2	20190531	주유	여성	50대	93502
3	20150508	요식/유흥	남성	30대	950842
4	20200613	교육/학원	여성	20대	14199
...	...	...	...	...	...
38635	20190527	가전/가구	남성	20대	15962
38636	20200720	자동차	남성	30대	36500
38637	20190731	자동차	여성	20대	7328
38638	20150721	여행/교통	여성	30대	35275
38639	20190522	가정생활/서비스	남성	60대이상	1719

38640 rows × 5 columns

연령대가 40대, 50대인 데이터 추출

②키워드 분석

- 분석목적

키워드 분석을 통해 40·50대가 공연을 소비하는 이유를 파악

- 분석방법

Ⅰ.
40·50대가 많이
관람한 공연명 찾기

Ⅱ.
공연 리뷰 크롤링

Ⅲ.
리뷰 속 키워드
분리 및 시각화

002 분석 방법 및 결과

②키워드 분석

Ⅰ.40·50대가 많이 관람한 공연명 찾기 | Ⅱ. 공연 리뷰 크롤링 | Ⅲ. 리뷰 속 키워드 분리 및 시각화

(1) 데이터 범위 좁히기

- KOPIS 데이터 내에서 연령이 40·50대(1961~1981년생) & 네이버 예약으로 예매

(2) KOPIS 데이터의 '기획제작사명' 컬럼을 기준으로 예매 건수가 높은 순으로 나열하기

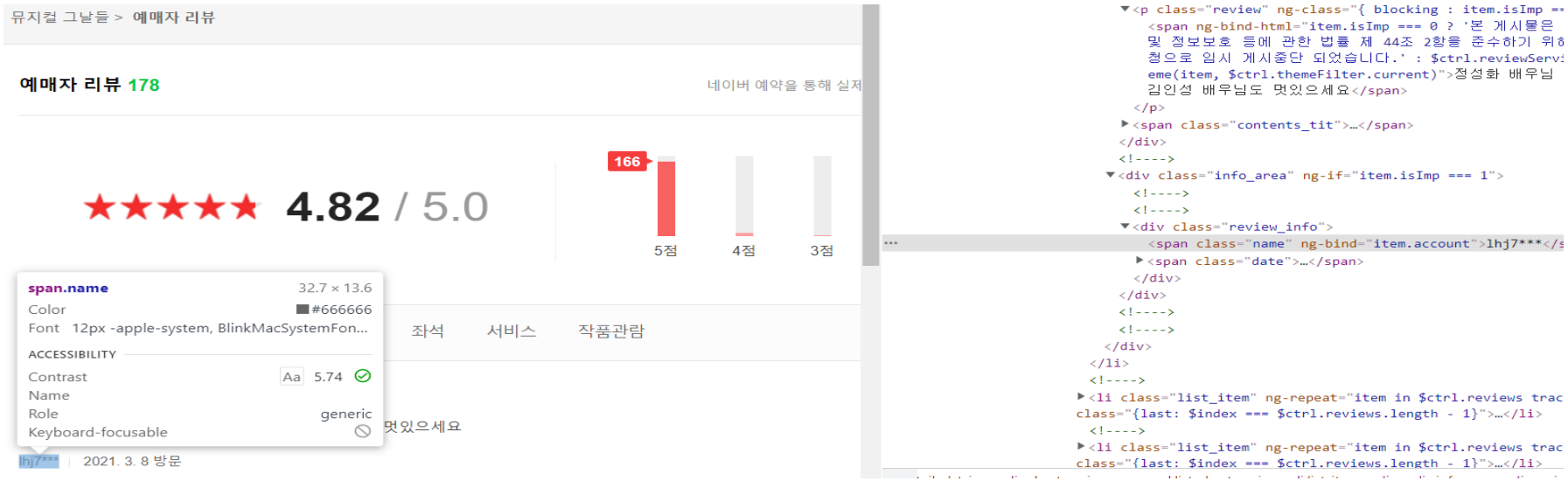
- 상위 10개의 '기획제작사명' 데이터를 확인하여, KOPIS 데이터에서 해당 공연의 '출연진내용' 확인
- 확인한 '출연진내용' 을 KOPIS 사이트에서 검색하여 공연명 확인

KOPIS에서 제공한 총 14개 파일 모두 같은 과정을 통해 파일별 각 월의 상위 10개의 공연명을 확인하였다.
➔ 최종적으로 26개의 공연을 확인하였고, 이 26개의 공연의 리뷰를 네이버 예약 사이트에서 크롤링하고자 한다.

002 분석 방법 및 결과

②키워드 분석

Ⅰ.40·50대가 많이 관람한 공연명 찾기 | Ⅱ. 공연 리뷰 크롤링 | Ⅲ. 리뷰 속 키워드 분리 및 시각화



- 크롤링 하고 싶은 부분의 class 명을 받아 BeautifulSoup의 find_all 함수를 이용하여 텍스트화 후 리스트에 저장

002 분석 방법 및 결과

②키워드 분석

Ⅰ.40·50대가 많이 관람한 공연명 찾기 | Ⅱ. 공연 리뷰 크롤링 | Ⅲ. 리뷰 속 키워드 분리 및 시각화

★★★★½ 4.5

모두 열심히 해주셔서 넘 좋았어요^^
아이가 역시 공주님라푼젤 언니가 젤 좋았다고하네요ㅎㅎ
좋은 공연 감사합니다 엄마 아빠도 즐겁게 관람했어요^^

ycom*** | 2021. 9. 11 방문

- 작성자 ID를 크롤링할 때, 작성자 ID의 span class 명과 주최측 답변자의 span class명이 동일하여 모든 ID가 뽑히는 경우를 확인

➔ 작성자 ID만 뽑기 위해 if 조건을 걸어 주최측의 ID가 아닐 경우에만 ID를 뽑도록 함

└ 안녕하세요, 달밤엔컴퍼니 입니다!
반짝반짝 라푼젤을 찾아주시고 아이와 함께 즐겁게 관람해주셔서 정말 감사드립니다.
소중한 후기에 힘입어 항상 좋은 공연을 위해 노력하겠습니다. 감사합니다.

클래식 가족뮤지컬 반짝 반짝 라푼젤 | 2021. 9. 11 오후 9:41

```
# 공연명, 공연 장소, 공연 기간, 공연 시간, 관람 연령, 가격, 장르
name = []; place = []; period = []; runTime = []; age = []; price = []; genre = [];

for x in range(len(review)):
    name.append('그날들')
    place.append('충무아트센터 대극장')
    period.append('2020.11.13 ~ 2021.03.07')
    runTime.append('165분')
    age.append('8세이상')
    price.append('R석 120,000원 S석 80,000원 A석 50,000원')
    genre.append('뮤지컬')
```

- 공연명, 공연 장소, 공연 기간, 공연 시간 등 부족한 정보를 추가하여 리스트에 저장

002 분석 방법 및 결과

②키워드 분석

Ⅰ.40·50대가 많이 관람한 공연명 찾기 | Ⅱ. 공연 리뷰 크롤링 | Ⅲ. 리뷰 속 키워드 분리 및 시각화

엑셀 파일 저장

A	B	C	D	E	F	G	H	I	J	K
	공연명	공연 장소	공연 기간	공연 시간	관람 연령	가격	장르	리뷰 내용	예약자ID	방문 날짜
0	장화 신은	국립중앙박	2019.12.14	70분	36개월이상	66,000원	뮤지컬	11세되는	rlqm****	2020. 2. 3
1	장화 신은	국립중앙박	2019.12.14	70분	36개월이상	66,000원	뮤지컬	아이들이	huni****	2020. 2. 3
2	장화 신은	국립중앙박	2019.12.14	70분	36개월이상	66,000원	뮤지컬	아이가 재	best****	2020. 2. 3
3	장화 신은	국립중앙박	2019.12.14	70분	36개월이상	66,000원	뮤지컬	퀄리티가	kw****	2020. 2. 3
4	장화 신은	국립중앙박	2019.12.14	70분	36개월이상	66,000원	뮤지컬	취소할까	bveo*****	2020. 2. 3

- 크롤링: 공연명, 공연 장소, 공연 기간, 공연 시간, 관람 연령, 가격, 장르, 리뷰 내용, 예약자 ID, 방문날짜
- 전처리: 방문날짜가 2020년이 아닌 경우의 행 모두 제거

②키워드 분석

(1) 정제
- 텍스트 양 옆 공백 제거
- 특수문자 제거(예: '&', '~')
- 텍스트 중간 공백은 하나만 남기기(공백이 두개 이상인 경우가 존재)

(2) 맞춤법 체크
- 라이브러리 hanspell 이용

(3) 토큰화 및 품사 태깅
- Okt를 이용하여 리뷰 토큰화 및 품사 태깅

(4) 명사 분리
- 품사 태깅된 리뷰에서 'Noun' 에 해당하는 단어 분리

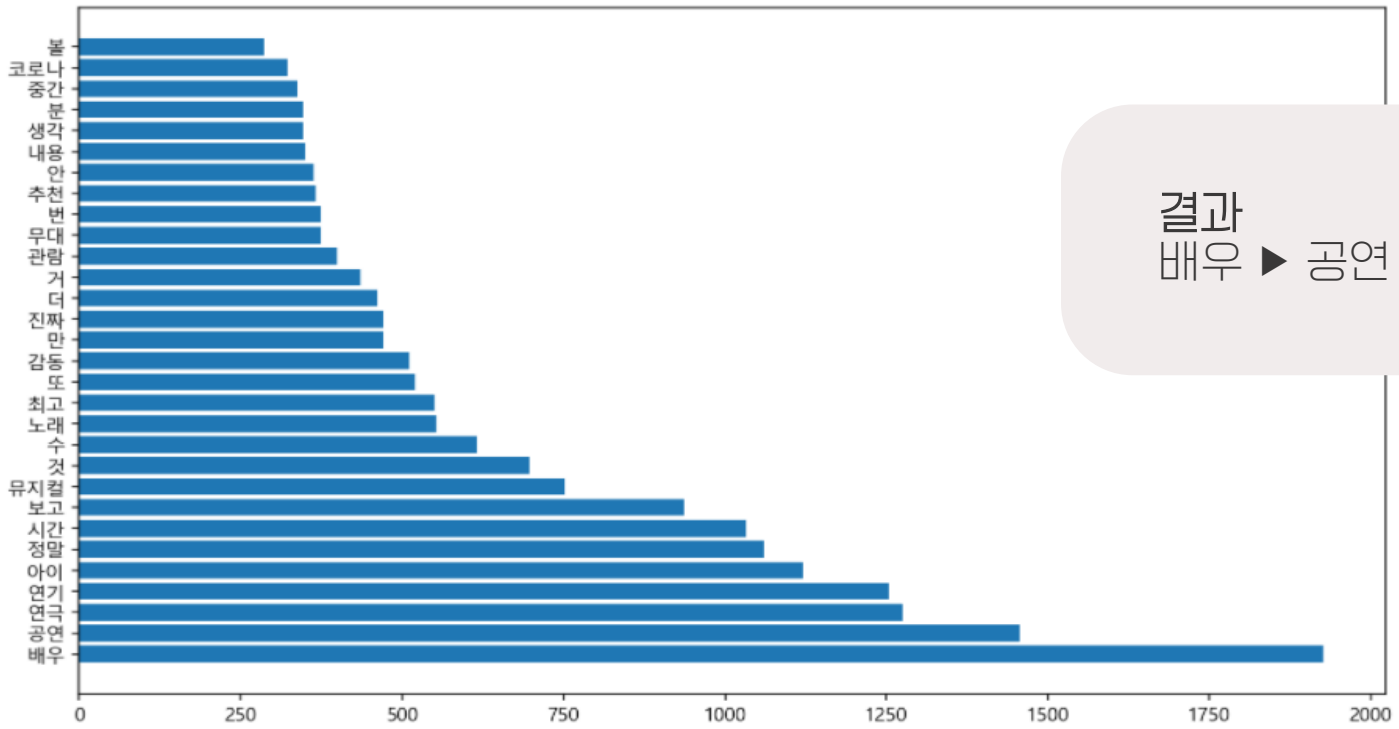
(5) 이중리스트 ➔ 단일리스트
- ('배우', 'Noun') ➔ '배우' 만 추출하기
- 한 문장내에 있는 단어가 한 리스트에 저장되어 있는 상태이므로 단일리스트로 변경

(6) 명사 빈도수 count

(7) 내림차순 정리

(8) 시각화

②키워드 분석



결과
배우 ▶ 공연 ▶ 연극 ▶ 연기 ▶ 아이 ▶ 정말 ▶ 시간

결과적으로 크롤링 파일에서 명사는 총 62749개가 추출되었다. 그리고 명사 빈도수 측정을 통해 언급된 서로 다른 명사는 총 4385개임을 확인하였다. 또한, 가장 많이 언급된 단어는 총 1926번 언급이 되었음을 확인하였다.

002 분석 방법 및 결과

③다중선형회귀분석

	소비건수_가전/가구	소비건수_가정생활/서비스	소비건수_교육/학원	소비건수_미용	소비건수_백화점/상품권/아울렛	소비건수_스포츠/문화/레저	소비건수_여행/교통	소비건수_요식/유흥	소비건수_의료	소비건수_자동차	소비건수_주유	소비건수_패션/잡화	소비건수_편의점	소비건수_할인점/마트	공연_예매건수
2020-05-01	168852	9185	156849	166008	483613	355219	244830	3170690	823668	119264	541531	148908	1285970	2616996	10670
2020-05-02	177098	12298	136121	163239	505411	355140	258005	3176138	847650	111819	527666	152185	1257137	2695141	7993
2020-05-03	113262	5411	94734	116748	520064	324418	226798	2709895	163307	57012	434148	123831	1110042	2309438	7315
2020-05-04	192377	15980	195744	148987	369591	305316	175230	2859839	1213639	169472	532958	133431	1264069	2321018	9062
2020-05-05	155607	10561	122573	122061	614335	368677	165568	2976004	241073	84678	460442	157599	1042202	2411621	7947
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

- 분석목적
  다중선형회귀분석을 통하여 40·50대의 공연예매에 영향을 미치는 소비항목을 분석
- 가설 : 14가지의 소비 경향이 공연예매에 영향을 준다.
  → 예) 40·50대에서 편의점, 패션/잡화 항목에서 연관성이 있을 것이다.
  → 예) 40·50대 남성에서 패션/잡화 항목이 연관이 있을 것이다.
  → 예) 40·50대 여성에서 편의점, 패션/잡화, 자동차, 요식/유흥 항목에서 연관이 있을 것이다.

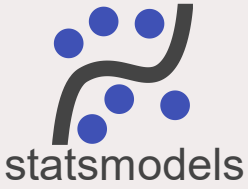
002 분석 방법 및 결과

③다중선형회귀분석

- 분석방법

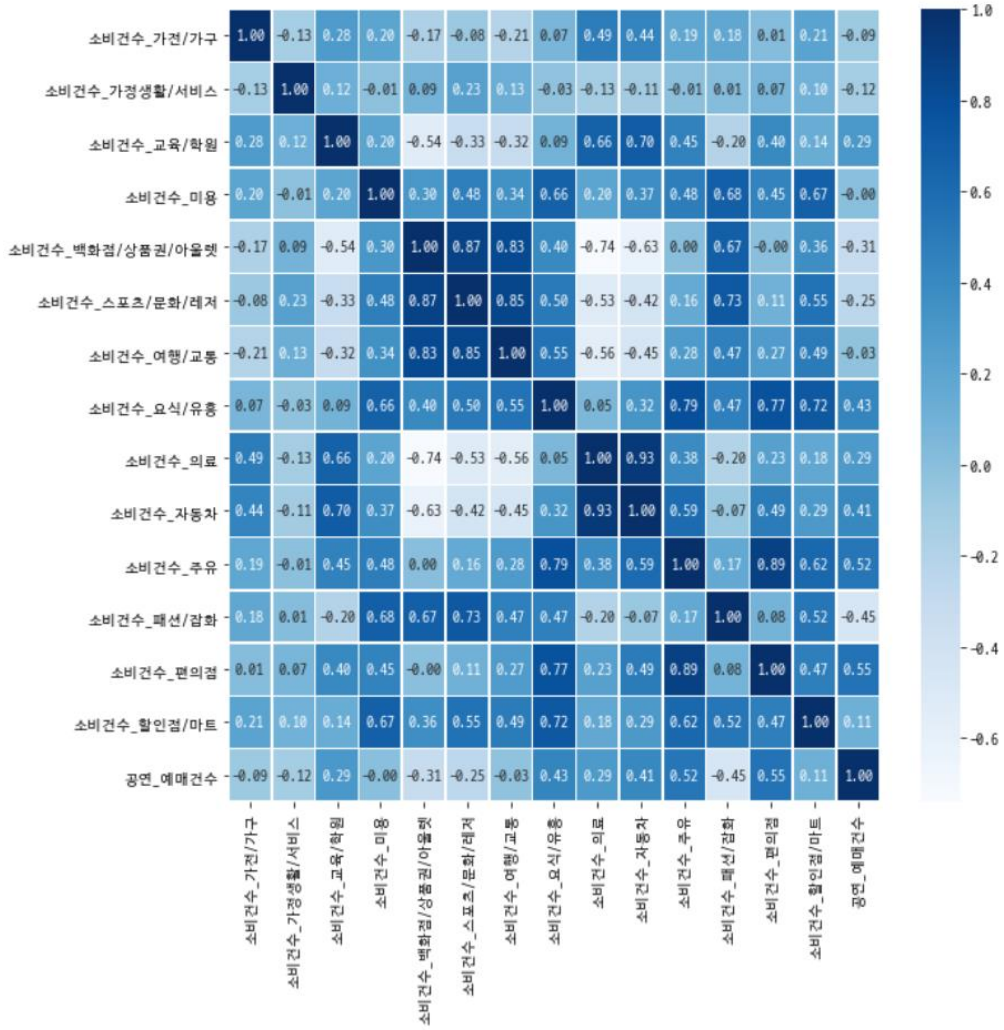
Ⅰ. 선형성 확인 및 다중공선성 확인

Ⅱ. Statsmodels 라이브러리의 OLS를 이용하여 다중선형회귀분석 수행



002 분석 방법 및 결과

③다중선형회귀분석



Ⅰ. 선형성 확인 및 다중공선성 확인

(1). 히트맵과 산점도 그래프를 통해 공연예매건수와 항목별 소비건수 간의 상관관계 확인

- 선형회귀분석이 유의미한 결과를 얻기 위해 지켜야하는 기본 가정 4가지 중에서 선형성을 확인하기 위해 상관관계 확인

- 독립변수들 사이에 상관계수가 0.8이상인 경우가 있음

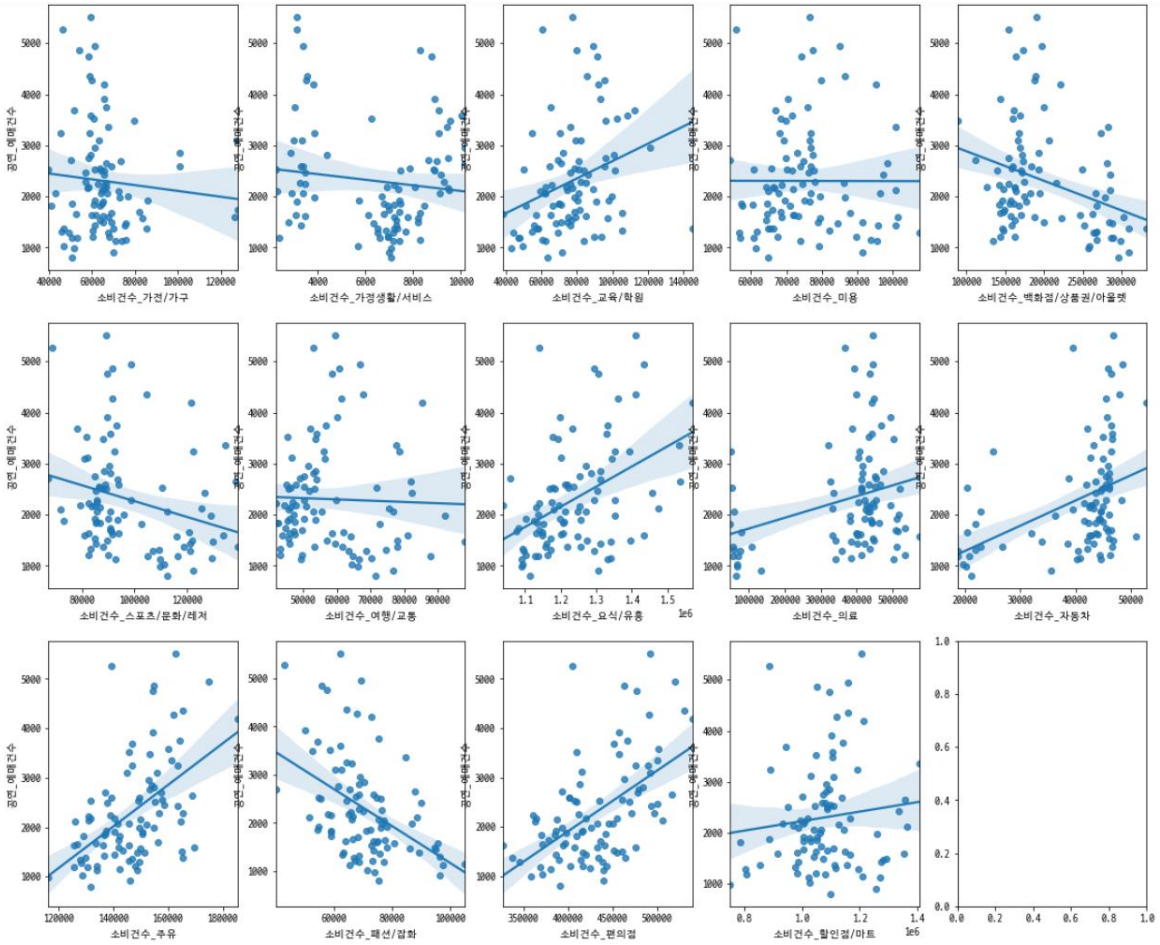
➔ 다중공선성 의심

- 선형회귀분석 기본가정 4가지 중 독립성을 위배

➔ 이를 해결하기 위한 절차 진행

002 분석 방법 및 결과

③다중선형회귀분석



Ⅰ. 선형성 확인 및 다중공선성 확인

(2) 일부 변수에서 선형회귀 4가지 기본가정 중 선형성을 보이기 때문에 다중선형회귀분석 진행

002 분석 방법 및 결과

③다중선형회귀분석

	VIF Factor	features
0	181.030952	const
1	3.311783	소비건수_가전/가구
2	1.375171	소비건수_가정생활/서비스
3	2.167119	소비건수_교육/학원
4	3.292800	소비건수_여행/교통
5	4.913309	소비건수_의료
6	6.717976	소비건수_주유
7	6.245491	소비건수_패션/잡화
8	3.774895	소비건수_편의점
9	7.015010	소비건수_할인점/마트

Ⅰ. 선형성 확인 및 다중공선성 확인

(3) VIF(Variance Inflation Factor)로 다중공선성 확인

- 다중공선성을 가장 크게 유발하는 독립변수를 제거
- VIF Factor가 10 이상일 경우 다중공선성이 있다고 판단
- 컬럼을 하나씩 제거하면서 상수를 제외한 모든 컬럼이 10 이하의 값이 될 때 까지 반복

002 분석 방법 및 결과

③다중선형회귀분석

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.604
Model:	OLS	Adj. R-squared:	0.556
Method:	Least Squares	F-statistic:	12.38
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	9.80e-13
Time:	19:30:53	Log-Likelihood:	-758.08
No. Observations:	92	AIC:	1538.
Df Residuals:	81	BIC:	1566.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1457.0865	1628.327	0.895	0.374	-1782.774	4696.947
소비건수_가전/가구	-0.0076	0.005	-1.684	0.096	-0.017	0.001
소비건수_가정생활/서비스	-0.0560	0.036	-1.559	0.123	-0.128	0.015
소비건수_교육/학원	-0.0038	0.004	-0.985	0.327	-0.012	0.004
소비건수_미용	0.0004	0.009	0.041	0.967	-0.017	0.018
소비건수_여행/교통	0.0134	0.006	2.113	0.038	0.001	0.026
소비건수_자동차	0.0209	0.010	2.096	0.039	0.001	0.041
소비건수_주유	-0.0087	0.008	-1.161	0.249	-0.024	0.006
소비건수_패션/잡화	-0.0251	0.008	-3.170	0.002	-0.041	-0.009
소비건수_편의점	0.0054	0.002	3.056	0.003	0.002	0.009
소비건수_할인점/마트	-0.0001	0.001	-0.173	0.863	-0.002	0.001

Omnibus:	5.432	Durbin-Watson:	1.066
Prob(Omnibus):	0.066	Jarque-Bera (JB):	5.513
Skew:	0.590	Prob(JB):	0.0635
Kurtosis:	2.785	Cond. No.	4.47e+07

Ⅱ. Statsmodels 라이브러리의 OLS를 이용하여 다중선형회귀 수행

(1) 후진 제거법(Backward Elimination)을 이용하여 중요성이 떨어진다고 판단되는 독립변수를 하나씩 제거

- p-value 값이 0.1을 초과할 경우 변수를 제거하되, 결정계수 값에 큰 영향을 미치지 않는 선에서 p-value 값이 0.05를 초과하는 경우도 제거

(2) 데이터의 범위를 세분화하여 다중회귀분석 시행 후 모델 해석하기

- 앞의 OLS 결과를 기반으로 유의미한 결과를 얻을 것으로 예상되는 연령과 성별 범위를 재설정
- 변경된 데이터에 따른 다중회귀분석을 반복하여 진행

002 분석 방법 및 결과

③다중선형회귀분석

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.561
Model:	OLS	Adj. R-squared:	0.546
Method:	Least Squares	F-statistic:	37.41
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	1.11e-15
Time:	19:37:10	Log-Likelihood:	-762.92
No. Observations:	92	AIC:	1534.
Df Residuals:	88	BIC:	1544.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1196.7868	1196.204	-1.000	0.320	-3573.991	1180.418
소비건수_여행/교통	0.0082	0.004	2.076	0.041	0.000	0.016
소비건수_패션/잡화	-0.0303	0.004	-8.027	0.000	-0.038	-0.023
소비건수_편의점	0.0050	0.001	6.002	0.000	0.003	0.007

Omnibus:	11.096	Durbin-Watson:	0.923
Prob(Omnibus):	0.004	Jarque-Bera (JB):	11.334
Skew:	0.817	Prob(JB):	0.00346
Kurtosis:	3.535	Cond. No.	1.61e+07

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.433
Method:	Least Squares	F-statistic:	24.19
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	1.65e-11
Time:	19:57:52	Log-Likelihood:	-618.59
No. Observations:	92	AIC:	1245.
Df Residuals:	88	BIC:	1255.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-323.6474	252.219	-1.283	0.203	-824.879	177.584
소비건수_여행/교통	0.0036	0.001	2.603	0.011	0.001	0.006
소비건수_패션/잡화	-0.0081	0.001	-7.129	0.000	-0.010	-0.006
소비건수_편의점	0.0011	0.000	4.001	0.000	0.001	0.002

Omnibus:	22.193	Durbin-Watson:	1.027
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.324
Skew:	1.174	Prob(JB):	4.29e-07
Kurtosis:	4.460	Cond. No.	1.11e+07

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.732
Method:	Least Squares	F-statistic:	83.72
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	1.08e-25
Time:	20:14:44	Log-Likelihood:	-708.40
No. Observations:	92	AIC:	1425.
Df Residuals:	88	BIC:	1435.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2222.6588	629.543	-3.531	0.001	-3473.744	-971.574
소비건수_여행/교통	-0.0112	0.005	-2.038	0.045	-0.022	-0.000
소비건수_요식/유흥	0.0080	0.001	12.968	0.000	0.007	0.009
소비건수_패션/잡화	-0.0675	0.005	-12.392	0.000	-0.078	-0.057

Omnibus:	9.373	Durbin-Watson:	1.384
Prob(Omnibus):	0.009	Jarque-Bera (JB):	9.334
Skew:	0.659	Prob(JB):	0.00940
Kurtosis:	3.836	Cond. No.	1.38e+07

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.673
Method:	Least Squares	F-statistic:	94.81
Date:	Sat, 11 Sep 2021	Prob (F-statistic):	8.81e-23
Time:	19:12:08	Log-Likelihood:	-693.39
No. Observations:	92	AIC:	1393.
Df Residuals:	89	BIC:	1400.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1696.7315	536.446	-3.163	0.002	-2762.638	-630.825
소비건수_요식/유흥	0.0096	0.001	10.775	0.000	0.008	0.011
소비건수_패션/잡화	-0.0978	0.007	-13.307	0.000	-0.112	-0.083

Omnibus:	12.378	Durbin-Watson:	1.360
Prob(Omnibus):	0.002	Jarque-Bera (JB):	13.582
Skew:	0.764	Prob(JB):	0.00112
Kurtosis:	4.100	Cond. No.	8.20e+06

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	62.54
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	8.53e-25
Time:	20:37:17	Log-Likelihood:	-577.13
No. Observations:	92	AIC:	1164.
Df Residuals:	87	BIC:	1177.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-281.2682	147.372	-1.909	0.060	-574.185	11.649
소비건수_백화점/상품권/아울렛	-0.0032	0.001	-2.995	0.004	-0.005	-0.001
소비건수_스포츠/문화/레저	-0.0094	0.004	-2.271	0.026	-0.018	-0.001
소비건수_요식/유흥	0.0043	0.000	14.070	0.000	0.004	0.005
소비건수_패션/잡화	-0.0188	0.003	-5.564	0.000	-0.026	-0.012

Omnibus:	6.129	Durbin-Watson:	1.285
Prob(Omnibus):	0.047	Jarque-Bera (JB):	5.905
Skew:	0.440	Prob(JB):	0.0522
Kurtosis:	3.875	Cond. No.	5.58e+06

Ⅱ. Statsmodels 라이브러리의 OLS를 이용하여 다중선형회귀 수행

➔ 결과적으로 5개의 모델이 만들어짐

- 40·50대 전체
- 40·50대 남성
- 40·50대 여성
- 40대 여성
- 50대 여성

003 결과 분석

①키워드 분석

1) '아이'와 '시간'이 많이 언급된 것에 주목

➔ "아이와 함께 좋은 시간 보냈어요.", "1시간 10분간의 공연시간 동안 아이들이 몰입해서 즐겼습니다.", "아이들 눈높이에 딱 적당한 시간대와 연극이여서 잘 보고 왔어요 ~^^"

➔ 결론: 40·50대가 아이와 함께 시간을 보내기 위해 공연을 관람한다는 경향을 띈다.

2) 출연배우의 언급 853건(9번째로 많이 언급)

➔ "해나 배우가 나오는 날로 봤어요. 연기도 정말 잘하시는데 가창력도 대단했습니다.", "엄기준, 박건형, 조재윤 배우님 캐스팅으로 정말 재미있게 보았습니다!", "엄건복 페어 친구케미가 완전 최고 특히 박건형배우 역시 능청스러운 연기에 아주 그냥 찰떡입니다."

➔ 결론: 출연배우 이름을 리뷰에 언급한다는 것은 공연 소비에 있어서 배우의 유명도/인지도가 영향을 준다고 할 수 있다. 스타마케팅이 40·50대에게 충분한 소비를 불러올 수 있다.

②다중선형회귀분석

Ⅰ. 전체 결과(모든 모델)

- 결정계수: 40·50대 여성에서 73.2%로 가장 높게 나타남
- F-statistic: 40·50대 남성에서 가장 낮게 나타남
- t(t-test)

- 40·50 전체, 40·50 남성: 소비건수_편의점
- 40·50 여성, 40대 여성, 50대 여성: 소비건수_요식/유흥

➔ 독립변수 '소비건수_패션/잡화' 가 1 증가할 때 종속변수 '공연예매건수' 는 감소하는 양상을 보임

003 결과 분석

②다중선형회귀분석

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.561
Model:	OLS	Adj. R-squared:	0.546
Method:	Least Squares	F-statistic:	37.41
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	1.11e-15
Time:	19:37:10	Log-Likelihood:	-762.92
No. Observations:	92	AIC:	1534.
Df Residuals:	88	BIC:	1544.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1196.7868	1196.204	-1.000	0.320	-3573.991	1180.418
소비건수_여행/교통	0.0082	0.004	2.076	0.041	0.000	0.016
소비건수_패션/잡화	-0.0303	0.004	-8.027	0.000	-0.038	-0.023
소비건수_편의점	0.0050	0.001	6.002	0.000	0.003	0.007

Omnibus:	11.096	Durbin-Watson:	0.923
Prob(Omnibus):	0.004	Jarque-Bera (JB):	11.334
Skew:	0.817	Prob(JB):	0.00346
Kurtosis:	3.535	Cond. No.	1.61e+07

II. 40·50대 전체

- $y = -1196.7868 + 0.0082x_1 - 0.0303x_2 + 0.005x_3$ (x_1: 소비건수_여행/교통, x_2: 소비건수_패션/잡화, x_3: 소비건수_편의점)
- 결정계수 값이 0.546으로 추정된 회귀모델로 데이터의 54.6%를 설명할 수 있음
- F-statistics 값은 5개의 모델 중 2번째로 작은 값을 보임

➔ t(t-test): 독립변수 '소비건수_편의점' 이 종속변수와의 상관도가 가장 크게 나타나고 있음

003 결과 분석

②다중선형회귀분석

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.452
Model:	OLS	Adj. R-squared:	0.433
Method:	Least Squares	F-statistic:	24.19
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	1.65e-11
Time:	19:57:52	Log-Likelihood:	-618.59
No. Observations:	92	AIC:	1245.
Df Residuals:	88	BIC:	1255.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-323.6474	252.219	-1.283	0.203	-824.879	177.584
소비건수_여행/교통	0.0036	0.001	2.603	0.011	0.001	0.006
소비건수_패션/잡화	-0.0081	0.001	-7.129	0.000	-0.010	-0.006
소비건수_편의점	0.0011	0.000	4.001	0.000	0.001	0.002

Omnibus:	22.193	Durbin-Watson:	1.027
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.324
Skew:	1.174	Prob(JB):	4.29e-07
Kurtosis:	4.460	Cond. No.	1.11e+07

Ⅲ. 40·50대 남성

- $y = -323.6474 + 0.0036x_1 - 0.0081x_2 + 0.0011x_3$ (x_1: 소비건수_여행/교통, x_2: 소비건수_패션/잡화, x_3: 소비건수_편의점)
- 결정계수 값이 0.433으로 추정된 회귀모델로 데이터의 43.3%를 설명할 수 있음
- F-statistics 값이 5개의 모델 중 가장 작은 값이기 때문에 F 통계량으로 도출된 회귀식이 비교적 적절함

➔ t(t-test): 독립변수 '소비건수_편의점' 이 종속변수와의 상관도가 가장 크게 나타나고 있음

003 결과 분석

②다중선형회귀분석

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.732
Method:	Least Squares	F-statistic:	83.72
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	1.08e-25
Time:	20:14:44	Log-Likelihood:	-708.40
No. Observations:	92	AIC:	1425.
Df Residuals:	88	BIC:	1435.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2222.6588	629.543	-3.531	0.001	-3473.744	-971.574
소비건수_여행/교통	-0.0112	0.005	-2.038	0.045	-0.022	-0.000
소비건수_요식/유흥	0.0080	0.001	12.968	0.000	0.007	0.009
소비건수_패션/잡화	-0.0675	0.005	-12.392	0.000	-0.078	-0.057

Omnibus:	9.373	Durbin-Watson:	1.384
Prob(Omnibus):	0.009	Jarque-Bera (JB):	9.334
Skew:	0.659	Prob(JB):	0.00940
Kurtosis:	3.836	Cond. No.	1.38e+07

Ⅳ. 40·50대 여성

- $y = -2222.6588 - 0.0112x_1 + 0.008x_2 - 0.0675x_3$ (x_1: 소비건수_여행/교통, x_2: 소비건수_요식/유흥, x_3: 소비건수_패션/잡화)
- 결정계수 값이 0.732으로 추정된 회귀모델로 데이터의 73.2%를 설명할 수 있음
- F-statistics 값은 5개의 모델 중 4번째로 낮은 값을 보임

➔ t(t-test): 독립변수 '소비건수_요식/유흥' 이 종속변수와의 상관도가 가장 크게 나타나고 있음

003 결과 분석

②다중선형회귀분석

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.681
Model:	OLS	Adj. R-squared:	0.673
Method:	Least Squares	F-statistic:	94.81
Date:	Sat, 11 Sep 2021	Prob (F-statistic):	8.81e-23
Time:	19:12:08	Log-Likelihood:	-693.39
No. Observations:	92	AIC:	1393.
Df Residuals:	89	BIC:	1400.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1696.7315	536.446	-3.163	0.002	-2762.638	-630.825
소비건수_요식/유흥	0.0096	0.001	10.775	0.000	0.008	0.011
소비건수_패션/잡화	-0.0978	0.007	-13.307	0.000	-0.112	-0.083

Omnibus:	12.378	Durbin-Watson:	1.360
Prob(Omnibus):	0.002	Jarque-Bera (JB):	13.582
Skew:	0.764	Prob(JB):	0.00112
Kurtosis:	4.100	Cond. No.	8.20e+06

V. 40대 여성

- $y = -1696.7315 + 0.0096x_1 - 0.0978x_2$
  (x_1: 소비건수_ 요식/유흥, x_2: 소비건수_ 패션/잡화)
- 결정계수 값이 0.673으로 추정된 회귀모델로 데이터의 67.3%를 설명할 수 있음
- F-statistics 값은 5개의 모델 중 가장 높은 값이기 때문에 비교적 회귀식을 설명하지 못한다고 볼 수 있음

➔ t(t-test): 독립변수 '소비건수_요식/유흥' 이 종속변수와의 상관도가 가장 크게 나타나고 있음

003 결과 분석

②다중선형회귀분석

OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	62.54
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	8.53e-25
Time:	20:37:17	Log-Likelihood:	-577.13
No. Observations:	92	AIC:	1164.
Df Residuals:	87	BIC:	1177.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-281.2682	147.372	-1.909	0.060	-574.185	11.649
소비건수_백화점/상품권/아울렛	-0.0032	0.001	-2.995	0.004	-0.005	-0.001
소비건수_스포츠/문화/레저	-0.0094	0.004	-2.271	0.026	-0.018	-0.001
소비건수_요식/유흥	0.0043	0.000	14.070	0.000	0.004	0.005
소비건수_패션/잡화	-0.0188	0.003	-5.564	0.000	-0.026	-0.012

Omnibus:	6.129	Durbin-Watson:	1.285
Prob(Omnibus):	0.047	Jarque-Bera (JB):	5.905
Skew:	0.440	Prob(JB):	0.0522
Kurtosis:	3.875	Cond. No.	5.58e+06

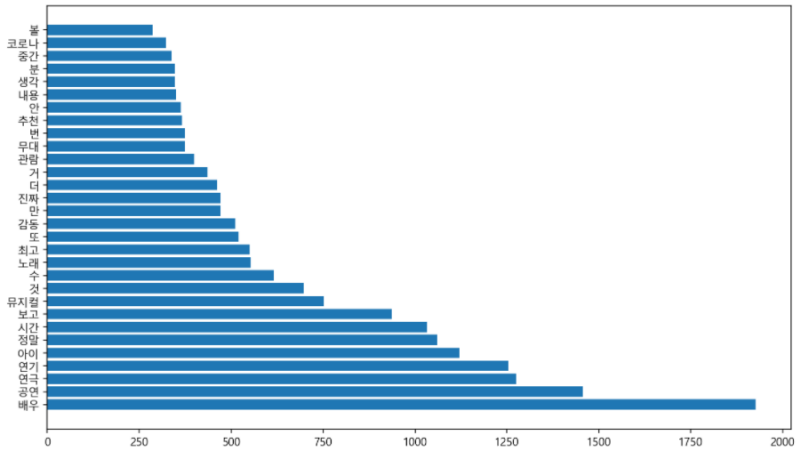
VI. 50대 여성

- $y = -281.2682 - 0.0032x_1 - 0.0094x_2 + 0.0043x_3 - 0.0188x_4$
  (x_1: 소비건수_백화점/상품권/아울렛, x_2: 소비건수_스포츠/문화/레저, x_3: 소비건수_요식/유흥, x_4: 소비건수_패션/잡화)
- 결정계수 값이 0.730으로 추정된 회귀모델로 데이터의 73%를 설명할 수 있음
- F-statistics 값은 5개의 모델 중 3번째로 작은 값을 보임

➔ t(t-test): 독립변수 '소비건수_요식/유흥' 이 종속변수와의 상관도가 가장 크게 나타나고 있음

004 활용방안 및 기대효과

①활용 가능성 및 방안



OLS Regression Results

Dep. Variable:	공연_예매건수	R-squared:	0.742
Model:	OLS	Adj. R-squared:	0.730
Method:	Least Squares	F-statistic:	62.54
Date:	Wed, 08 Sep 2021	Prob (F-statistic):	8.53e-25
Time:	20:37:17	Log-Likelihood:	-577.13
No. Observations:	92	AIC:	1164.
Df Residuals:	87	BIC:	1177.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-281.2682	147.372	-1.909	0.060	-574.185	11.649
소비건수_백화점/상품권/아울렛	-0.0032	0.001	-2.995	0.004	-0.005	-0.001
소비건수_스포츠/문화/레저	-0.0094	0.004	-2.271	0.026	-0.018	-0.001
소비건수_요식/유흥	0.0043	0.000	14.070	0.000	0.004	0.005
소비건수_패션/잡화	-0.0188	0.003	-5.564	0.000	-0.026	-0.012

Omnibus:	6.129	Durbin-Watson:	1.285
Prob(Omnibus):	0.047	Jarque-Bera (JB):	5.905
Skew:	0.440	Prob(JB):	0.0522
Kurtosis:	3.875	Cond. No.	5.58e+06

결과 : 배우 ▶ 공연 ▶ 연극 ▶ 연기 ▶ 아이 ▶ 정말 ▶ 시간

이를 통해 40 • 50대를 타깃으로 한 공연을 제작할 때,
아이와 함께 시간을 보낼 수 있는 공연을 제작한다면
40 • 50대 공연 소비를 증대할 수 있을 것이다.

결과

여행/교통 소비건수 1△ → 0.0112 ▽

요식/유흥 소비건수 1△ → 0.008 △

패션/잡화 소비건수 1△ → 0.0675 ▽

이를 통해 각각의 소비건수를 알 수 있다면 공연예매건수를
예측하여 공연예술 활성화를 위한 전략을 제시할 수 있다.

②기대효과

Ⅰ.
토픽모델링

Ⅱ.
장바구니 분석

Ⅲ.
예매 예측 시스템

Ⅳ.
맞춤 마케팅

004 활용방안 및 기대효과

③한계점

크롤링한 리뷰에서 명사를 분리하기 위해 Okt 형태소 분석기를 사용하였다. → 이후에 Okt 형태소 분석기를 이용하여 텍스트 데이터에서 명사를 분류할 때 명사를 분류하는 기준이 추가적으로 필요할 것이다.

리뷰 데이터에서 나타난 결과가 40・50대 전체의 경향성을 파악하기에는 무리가 있다. → 하지만 실제 관객의 리뷰를 분석함으로써 공연을 본 관객의 선호도를 분석해 공통적으로 나타나는 주제를 발견하였다는 것에 큰 의의를 둔다. 그리고 분석 결과를 토대로 토픽 모델링을 통해 공연의 흥행요인을 분석해 볼 수 있을 것이다.

삼성카드의 소비건수 데이터가 5월~7월에 한해서만 무료로 제공되었다. → 2020년 전체 소비건수 데이터를 얻을 수 있었다면, 2020년 한 해의 경향을 확인하고 더 나아가 다중회귀분석 모델의 설명력 또한 높일 수 있을 것이다.

다소 높은 다중공선성 수치 → 이것은 향후 최적화된 모델을 찾는 것이 용이한 scikit-learn 라이브러리를 이용하여 분석을 진행함으로써 다중공선성 요인을 해결하고 더 높은 설명력을 갖는 결과를 제시할 수 있을 것이다.