서식. 결과보고서

① 참가자 정보	
개인・팀명	디어마이프렌즈
주 제(제목)	40·50대 소비패턴 분석을 통한 공연예술 관객층 확대 방안 탐구

② 결과물 작성

- **개요**

1) 분석 목적

코로나 19는 사회에 큰 파문을 불러왔다. 개인을 넘어 범세계적인 영역으로 나아가 영향을 미치고 있다. 공연예술계 또한 이를 피할 수 없었으며, 이에 따라 공연예술 시장은 큰 타격을 입었다. 이러한 상황에서 공연예술 시장의 위기를 극복하기 위한 새로운 방안이 절대적으로 필요한 상황이다. 팀 디어마이프렌즈는 극복방안으로 40·50대를 새로운 관객층으로 확대할 필요성을 인지하였으며, 그에 따라 이들의 소비패턴을 분석하여 관객층으로 확대할 방안을 탐구하고자 한다.

2) 배경 및 필요성

<표 1> 인구피라미드와 연령대별 소득·소비 추이 및 연평균 증감률 현황

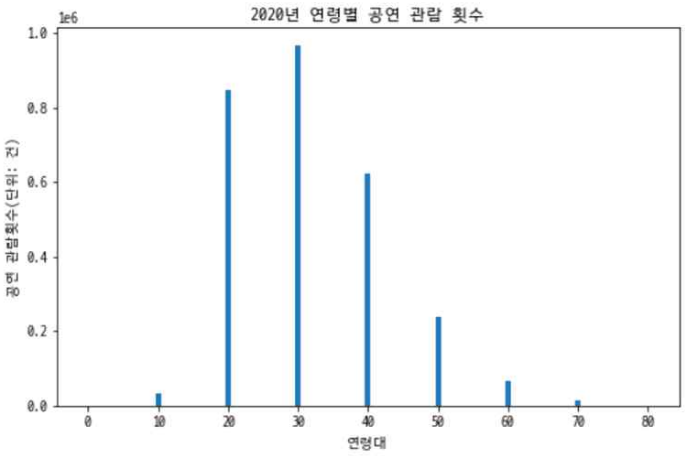
<표 1>의 인구피라미드를 보면 저출생·고령화로 인하여 20·30대보다 40·50대가 많은 비율을 차지한다는 것을 알 수 있다. 40·50대 가구주의 월평균 가구소득 규모가 최근 10년 동안 가장 빠른 속도로 증가하였고 이에 따라 소비 지출액 추이 또한 많이 증가했음을 알 수 있다. 이를 통해 40·50대가 코로나 19로 인해 위축된 공연 예술 산업에 새로운 돌파구가 될 수 있다.

<표 2> 연령별 라이선스뮤지컬 선택 시 고려요인의 중요도

라이선스뮤지컬 선택기준의 중요도	연령별 퍼센트(순위)				
	20대 미만	20대	30대	40대	50대 이상
작품의 유명도/인지도	10.5(1)	9.76(2)	9.84(2)	10.36(1)	9.71(1)
작품의 내용/질/예술성	10.1(2)	10.07(1)	10.31(1)	10.28(2)	9.68(2)
출연배우 유명도/인지도	8(4)	7.8(5)	8.03(4)	8(3)	8.79(3)
가수, 연예인 등 스타의 출연	5.5(7)	5.86(7)	4.81(8)	7.19(4)	7.36(4)
방송/잡지/인터넷 등 광고	5(10)	4.66(10)	4.1(10)	4.62(10)	5.36(9)
입소문/공연 평/관람후기	7.3(5)	8.41(3)	8.49(3)	6.94(5)	5.79(8)
입장권가격/할인여부	8.3(3)	8.05(4)	6.8(5)	6.79(6)	6.86(5)
공연장 유명도/위치/관의시설	5.3(8)	4.75(9)	4.73(9)	4.64(9)	4.54(10)
극단/기획사	2.8(12)	3.35(12)	3.78(11)	3.49(11)	3.32(12)
관객서비스	5.1(9)	3.99(11)	3.51(12)	3.32(12)	4.11(11)
음악	6.5(6)	6.19(6)	6.76(6)	6.55(7)	5.89(7)
연출/작가	3.6(11)	4.97(8)	6.68(7)	5.83(8)	6.61(6)

<표 2>은 연령별 라이선스뮤지컬 선택 시 고려요인의 중요도를 보여준다. 라이선스뮤지컬 선택 시에 입장권가격/할인여부가 20대는 3위를 기록했지만 40대는 6위, 50대는 5위를 기록했다. 이는 20대가 출연배우의 유명도/인지도보다 입장권가격/할인여부에 더 민감하다는 것을 알 수 있다. 하지만 소비여력이 충분한 40·50대는 뮤지컬 구매결정에 있어 입장권가격/할인여부가 미치는 영향이 적다는 것을 알 수 있다. 40·50대의 공연 소비에 유명도/인지도가 영향을 준다는 것을 입증한다면, 기존 공연예술 산업계의 전형적인 마케팅 방법인 스타마케팅의 단점으로 지적받아왔던 티켓 가격 상승을 관객들이 부담해야한다는 것을 보완해 줄 수 있을 것이다.

<표 3> 2020년 연령별 공연 관람 횟수



<표 3>을 통해 40·50대도 20·30대 다음으로 공연을 많이 관람하는 계층임을 알 수 있다. 그렇기 때문에 마케팅 타깃을 소비 여력이 충분한 40·50대로 넓혀본다면 공연예술산업에 다양한 비즈니스 기회를 제공할 수 있다고 기대한다.

3) 차별성 및 독창성

코로나 팬데믹 상황으로 인해 공연예술 시장은 위기를 맞았다. 이에 따라 공연예술 시장 활성화를 위한 새로운 대안이 필요한 상황이라는 것을 부정할 수 없다. 기존 공연예술계는 20·30대 관객층에 치중되어 왔다. 다양한 관객층을 흡수하지 못하고 특정 나이 대에 편협하게 기울어진 경향이 드러난다. 공연시장 활성화를 위해선 관객층의 확대가 절대적으로 필요하다.

(재)국립극단 어린이청소년극연구소는 어린이 청소년 극에 대한 본격적인 연구와 작품개발을 수행하고 있다. 또한 청소년 관객층에 대한 연구와 리서치, 워크숍 등을 수행하고 이 과정에서 얻어지는 내용과 의미를 제작 공연에 반영하여 청소년 연극의 새로운 방향성을 제시할 수 있는 시스템과 제작 여건을 마련하는 데 노력하고 있다고 공식적으로 설명하고 있다. 이처럼 기존 공연예술계의 주 관객층이었던 20·30대에만 집중하기보다 새로운 관객층을 탐색하고 이들을 공연시장으로 유입시키는 일은 필수 불가결한 일이라는 것을 알 수 있다. 새로운 관객층을 끊임없이 탐색하고 이들을 유입시키는 일은 공연예술 시장의 중요한 과제로 작용한다.

이에 근거하여 팀 디어마이프렌즈는 공연예술 시장 활성화를 위해 40·50대 중장년층이 새로운 관객층으로 유입시킬 방안을 도출하였다. 40·50대 중장년층은 최근 새로운 소비계층으로 떠오르고 있는 세대이다. 베이비붐 시대가 중장년층으로 들어서면서 이들은 더 이상 수동적인 소비가 아닌, ‘유튜브’나 온라인 쇼핑을 즐길 수 있다. 또한 삶의 질에 대한 관심이 늘면서 새로운 소비계층으로서의 가능성을 보여주고 있다. 이러한 점을 고려하였을 때, 40·50대를 새로운 관객층을 유입시키는 방안은 설득력을 높인다.

팀 디어마이프렌즈는 40·50대의 소비 유입 가능성을 위한 구체적인 방안을 도출하기 위해 40·50대의 소비패턴을 다중선형회귀분석을 하였다. 다양한 소비 경향성을 분석해 냄으로써 그들의 삶 전반적인 소비 패턴을 파악하고 그에 맞는 방안 혹은 마케팅 방식을 도출해내었다. 기존 공연예술 예매 자료를 통합시킨 KOPIS 빅데이터를 활용하면서도, 이에 국한되지 않고 외부데이터와의 연관성을 찾아내어 인사이트를 도출시켰다. 이제 공연예술 또한 하나의 소비로 작용하기 때문에 도출된 결과를 공연예술 시장과 연관 지어 마케팅 방식을 고안해 내는 것은 의의가 있다고 볼 수 있다.

4) 주요 내용 및 특징

저출생과 고령화로 인해 40·50대의 비율은 증가하고 있다. 40·50대는 20·30세대와 달리 훨씬 월등한 경제력을 갖추고 있으며, 40·50대가 공연예술 시장의 소비자층으로 새롭게 유입될 경우, 시장 활성화를 이루어 낼 수 있다고 예상할 수 있다.

분석은 크게 2가지 방식으로 나누어 진행하였다. 첫 번째는 공연 리뷰 데이터를 크롤링하여 키

워드 빈도수를 도출했다. 리뷰를 크롤링하여 40·50대가 공연을 예매/관람할 때 중요하게 여기는 요소들을 키워드로 밝히고 그 빈도수를 순위화함으로써 그들의 선호도와 가치관을 판단한다. 분석 결과를 이용하여 40·50대가 선호하는 공연을 제작할 수 있도록 방안을 도출시킨다. 두 번째는 소비건수와 공연예매건수 간의 다중선형회귀분석을 진행하였다. 14가지 항목의 소비건수를 독립변수로 공연예매건수를 종속변수로 설정하여 다중선형회귀분석을 진행한 결과 40·50대의 공연예매에 영향을 미치는 소비항목을 확인할 수 있었다. 그리고 분석 결과를 토대로 마케팅 방안을 도출시켰다.

도출된 방안을 공연예술시장에 적용할 경우 기존에 충분히 가능성이 있음에도 불구하고 놓치고 있었던 40·50대라는 새로운 관객층을 유입시킬 수 있을 것이다. 빠르게 변화하는 시대가 도래 하면서 마케팅의 방식은 끊임없이 다양한 방식과 새로운 소비층을 찾아야 하는 추세이다. 공연예술시장 또한 빠르게 변화하는 시대에 맞추어 40·50대라는 새로운 관객층을 유입시키고 그들에게 맞는 마케팅 전략을 수립해야 한다는 사실을 지나치게 간과하고 있다. 40·50대를 집중적으로 분석한 결과를 바탕으로 도출해낸 방안들은 코로나 팬데믹으로 위기를 맞은 공연예술 시장을 다시 활성화할 수 있는 가치가 있을 것이라고 예상한다.

• 세부 내용

1) 분석 방법

가) 공연 리뷰 데이터를 크롤링하여 키워드 빈도수 도출

· 방법론

40·50대가 많이 관람한 공연의 네이버 예약 리뷰를 확인한 후, 리뷰에서 언급한 단어들을 추출하여 빈도수를 확인한다. 이를 통해, 40·50대가 공연 관람 시에 중점적으로 본 부분이 무엇인지 순위화 하여 확인한다.

< 공연명 찾기 >

① 파일 불러오기

pandas 라이브러리를 활용하여, KOPIS 엑셀 파일을 불러온다.

<그림 1> 파일 불러오기

```
import pandas as pd
import seaborn as sns; sns.set()
```

```
df = pd.read_csv('빅데이터 분석 공모전 raw data 추출_2020.12.16-2020.12.31.csv')
```

② KOPIS 데이터에서 연령이 40·50대(1961~1981년생)이고, 네이버 예약으로 예매한 데이터로 범위를 좁히기

네이버 예약을 고른 이유는 네이버 예약 리뷰에 공연 방문 날짜가 적혀있기 때문이다. 공연 방문 날짜를 통해 2020년에 공연을 본 것인지를 확인하고자 하였다. 네이버 예약상의 데이터를 확인하기 위하여 KOPIS 데이터도 네이버 예약의 경우로 범위를 좁혔다.

<그림 2> 조건에 맞는 데이터로 범위 좁히기

```
df = df[(df['연령']>1961) & (df['연령']<= 1981)]
```

```
df = df[df['전송사업자명']=='네이버예약']
```

③ '기획제작사명' 컬럼을 기준으로 예매 건수가 높은 순으로 나열하기

KOPIS 데이터에서 '출연진내용'과 '제작진내용' 컬럼은 Null 데이터가 많았다. 또한 '공연장코드'로 공연을 분리하여 확인할 수 있다고 판단하였으나, '공연장코드'를 인터넷에서 검색하였을 때 공연이 검색되지 않았다. 그래서 '기획제작사명'이 Null 데이터가 거의 없을 뿐만 아니라 공연별로 '기획제작사명'이 다르게 나타나는 것을 확인하여 '기획제작사명'을 기준으로 예매 건수를 확인하였다. 만일 '기획제작사명'이 동일한 공연이 두 개가 있으면, '공연일시'를 참고하여 최종 공연명이 무엇인지 확인하였다.

<그림 3> 내림차순으로 예매 건수 높은 '기획제작사' 확인

```
sorted(count.items(), key=lambda x : x[1],reverse=True)
```

```
[('(주)인사이트엔터테인먼트(제작사)', (재)중구문화재단(충무아트센터)(주최)', 484),
 ('컬처마인(제작사)', (주)컬쳐홀릭(제작사)', (재)국립박물관문화재단(주관)', 컬처마인(주관)', (재)국립박물관문화재단(주최)', 컬처
  310),
 ('광야아트미니스트리(제작사)', 광야아트미니스트리(기획사)', 280),
 ('(주)대구백화점 대백프라자(주최)', (주)교문(주관)', 236),
 ('극단 파릇(제작사)', 232),
 ('광주시립발레단(주관)', 광주문화예술회관(주최)', 219),
 ('극단 뮤다드(기획사)', 극단 뮤다드(주최)', 187),
 ('브이매직엔터테인먼트(제작사)', 164),
 (nan, 146),
 ('(주)신시컴퍼니(제작사)', (주)신시컴퍼니(기획사)', (주)신시컴퍼니(주최)', SBS(주최)', 120),
```

상위 10개의 '기획제작사명' 데이터를 확인하여, 40·50대가 많이 본 공연이 무엇인지 확인하였다.

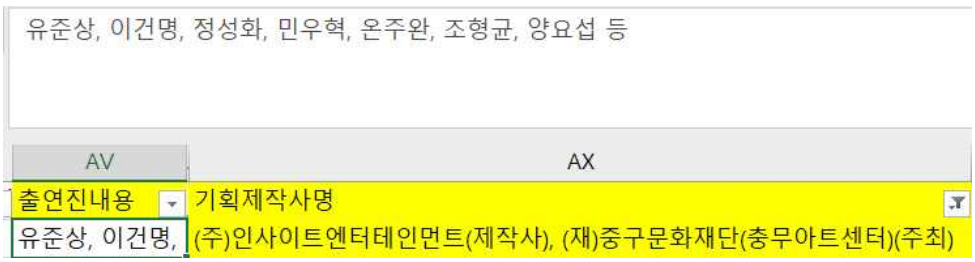
<그림 4> KOPIS 데이터에서 '기획제작사명' 확인 후 '출연진내용' 확인

출연진내용	제작진내용	기획제작사명
고태선, 이원혁, 이희민, 박연수, 한준희	김용민, 조희	(주)미스터이엔티(주최), (주)미스터이엔티(주관)
박연수, 한준희, 고태선, 이원혁, 박현우, 이희민	조희, 김용민, 박인서	(주)미스터이엔티(주최), (주)미스터이엔티(주관), (주)미스터이엔티(기획사)

동일한 '기획제작사명'이 있다 할지라도 공연마다 기획제작사가 여러 개가 있는 경우가 있다.

<그림 4>를 통해 기획제작사가 같은 경우가 있다고 할지라도 '기획제작사명' 컬럼 내에 저장된 '기획제작사명'의 수가 다르기 때문에 다른 공연임을 인지할 수 있었다.

<그림 5> '출연진내용' 확인



KOPIS 데이터에서 '기획제작사명'을 검색하여, 해당 공연의 '출연진내용'을 확인하였다. '출연진내용'을 KOPIS 사이트에서 검색하여 확인하였다.

<그림 6> 공연명 확인

번호	공연명	일시	장소	기획/제작/주최/주관
1	뮤지컬 그날들	2020.11.13 ~ 2021.03.07	충무아트센터 대공연장	(재)중구문화재단(충무아트센터)(주최), (주)인사이트엔터테인먼트(제작)

검색을 통해 해당 공연명이 '그날들'임을 확인하였다.

KOPIS에서 제공한 총 14개 파일 모두 같은 과정을 통해 파일별 상위 10개의 공연명을 확인하였다. 1·2월달의 공연은 상위 공연 예매 건수가 1,000건이 넘는 경우가 많았다. 하지만 3월부터는 1,000건 이상의 공연이 거의 없을 뿐만 아니라 100건 이상 예매한 공연도 드물었다. 이는 코로나19의 여파로 인해 나타난 결과임을 추측할 수 있었다. 그래서 공연 예매 건수를 기준으로 공연을 확인하기보다 각 월의 상위 10개의 공연을 확인하기로 결정하였다. 그리고 공연 기간이 한 달 이내 이외에도 여러 달 동안 하는 공연이 많아 최종적으로 26개의 공연을 확인할 수 있었다. 26개 공연의 리뷰를 파악하기 위해 네이버 예약을 통해 예매한 관람객의 리뷰를 크롤링 하고자 한다.

< 리뷰 크롤링 >

① 크롤링에 필요한 라이브러리 및 크롬 드라이버 불러오기

<그림 7> 라이브러리 및 크롬 드라이버 불러오기

```
from selenium import webdriver as wd
from selenium.webdriver.common.keys import Keys
from bs4 import BeautifulSoup
import pandas as pd
from tqdm import tqdm
import time
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

```
driver = wd.Chrome('C:/Users/user/chromedriver.exe')
```

② 공연의 네이버 예약 리뷰 사이트로 이동하기

관람평 ▶ 리뷰 클릭 ▶ 사이트 주소 복사하여 url 변수에 저장

<그림 8> 공연 관람평 확인

네이버 사이트에서 '뮤지컬 그날들'을 검색한 후, 관람평을 누른다. ▶ 네이버 예약의 리뷰를 확인하여 리뷰를 클릭한다. ▶ 새 리뷰창이 뜨면 목록을 눌러 크롤링 할 사이트의 주소를 복사한다.

<그림 9> 크롤링 할 사이트 첫 화면

<그림 10> 웹 페이지 파싱

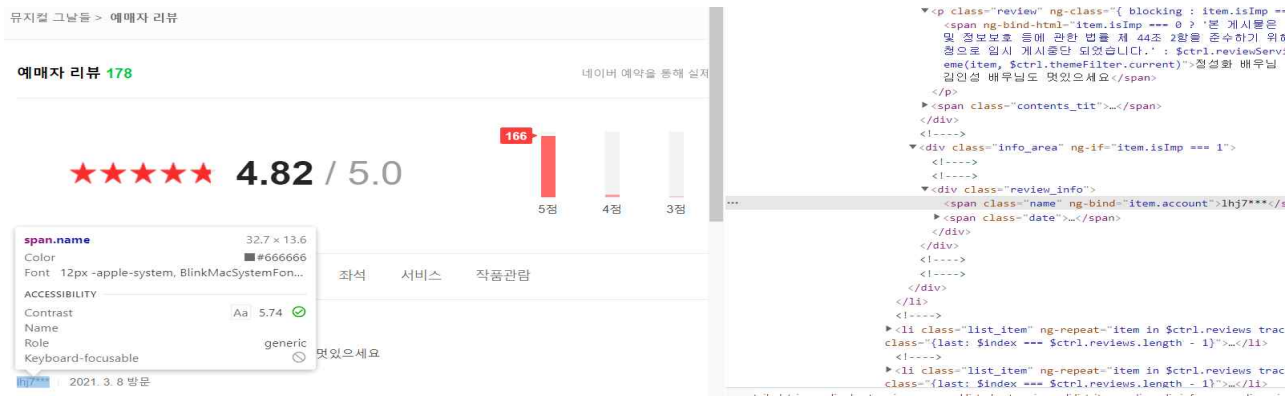
```
############################ 예매 사이트로 이동 ############################
url = 'https://booking.naver.com/review/bizes/421150'
driver.get(url)
time.sleep(2)

html = driver.page_source
soup = BeautifulSoup(html, 'html.parser')
```

사이트로 이동 후 BeautifulSoup 으로 현재 웹 페이지를 파싱한다.

③ 리뷰 크롤링

리뷰 내용, 작성자 ID, 방문 날짜를 크롤링 한 후, 데이터 프레임으로 합쳐 엑셀 파일로 저장하였다.

<그림 11> 크롤링 할 속성 확인



크롤링 하고자 하는 부분을 우 클릭 ▶ 검사를 통해 해당 class 명을 확인한다. BeautifulSoup 의 find_all 함수를 이용하여 해당 class 명 위치의 데이터를 받아온다.

<그림 12> 크롤링

```
################################## 크롤링 ##################################
# 리뷰 내용, 작성자 ID, 방문 날짜
review = []; review_id = []; review_date = [];

for y in tqdm(range(0,int(page))):

    html = driver.page_source
    soup = BeautifulSoup(html, 'html.parser')

    # 리뷰 내용
    w = soup.find_all('p',{'class':'review'})
    for x in w:
        review.append(x.get_text())

    # 작성자 ID
    w = soup.find_all('span',{'class':'name'})
    for x in w:
        # 주최측에서 답변을 단 경우, 답변에 해당하는 내용은 저장X
        if '그날들' not in x.get_text():
            review_id.append(x.get_text())

    # 방문 날짜
    w = soup.find_all('span',{'ng-bind': '$ctrl.getVisitDate(item.completedDateTime || item.useDate)'});
    for x in w:
        review_date.append(x.get_text())

    # 페이지 넘기기 위해 페이지 xpath 주소 저장
    next_page = driver.find_element_by_xpath('//*[@id="container"]/review-list/div/div/div/div[2]/a[2]')
    # review 변수에 저장된 곳 클릭
    next_page.send_keys(Keys.ENTER)
```

받아온 데이터를 get_text 함수를 통해 텍스트화 하여 리스트에 저장한다. 웹 페이지를 넘기기 위하여 페이지 넘기는 버튼의 xpath 주소를 받아 next_page 변수에 저장한다. 그리고 next_page를 클릭하여 다음 페이지로 넘기도록 한다. 또한 작성자 ID를 크롤링 할 때, 작성자 ID의 span class 명과 주최 측 답변자의 span class 명이 같은 것을 확인하였다. 이 결과 작성자 ID를 담은 리스트와 리뷰 내용, 방문 날짜를 담은 리스트의 길이가 다르게 나타나는 경우가 발생하였다. 이를 대비하여, 주최 측 ID가 text에 존재하지 않을 때만 리스트에 담도록 하였다.

<그림 13> 부족한 정보 추가

```
# 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')
```

부족한 정보를 추가하여 리스트에 저장하고, 공연 별로 담았던 리스트들을 하나로 합쳤다.

<그림 14> 제대로 뽑혔는지 확인

```
# 제대로 뽑혔는지 확인
len(a); len(b); len(c); len(d); len(e); len(f); len(g); len(h); len(i); len(j)
type(a); type(b); type(c); type(d); type(e); type(f); type(g); type(h); type(i); type(j)
```

```
690
690
690
690
```

모든 길이와 type이 동일하게 나왔는지 확인하였다.

<그림 15> 데이터프레임 및 엑셀 저장

```
df = pd.DataFrame({ '공연명': d,
                    '공연 장소': e,
                    '공연 기간':f,
                    '공연 시간': g,
                    '관람 연령':h,
                    '가격':i,
                    '장르': j,
                    '리뷰 내용':a,
                    '예약자ID':b,
                    '방문 날짜':c})
```

```
df.to_excel('C:/Users/user/정세영/review_crawling.xlsx')
```

데이터 프레임에 저장하여, 엑셀 파일에 저장을 완료하였다.

<그림 16> 엑셀 파일 'review_crawling.xlsx'

	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	0	장화 신은	국립중앙부	2019.12.14	70분	36개월이상	66,000원	뮤지컬	11세되는	rlqm****	2020. 2. 3
1	1	장화 신은	국립중앙부	2019.12.14	70분	36개월이상	66,000원	뮤지컬	아이들이	huni****	2020. 2. 3
2	2	장화 신은	국립중앙부	2019.12.14	70분	36개월이상	66,000원	뮤지컬	아이가 재	best****	2020. 2. 3
3	3	장화 신은	국립중앙부	2019.12.14	70분	36개월이상	66,000원	뮤지컬	퀄리티가	kw****	2020. 2. 3
4	4	장화 신은	국립중앙부	2019.12.14	70분	36개월이상	66,000원	뮤지컬	취소할까(	bveo*****	2020. 2. 3

엑셀 파일 내에서 방문 날짜가 2020년인 데이터들만 분류 및 저장하여 크롤링을 완료하였다.

< 키워드 분석 >

리뷰 데이터를 형태소 분석기를 통해 토큰화 하여 명사에 해당하는 것만 뽑아 명사별 빈도수를 확인한다. 형태소 분석기는 Okt를 사용하였다.

① 라이브러리 불러오기

형태소 분석기로 Okt라이브러리를 불러왔다. 그리고 명사 빈도수를 count 하기 위해 collections 라이브러리를 불러오고, 맞춤법 체크를 위한 spell_checker라이브러리도 불러왔다.

<그림 17> 라이브러리 불러오기

```
from konlpy.tag import Okt
from collections import Counter
import collections

# 맞춤법 체크
from hanspell import spell_checker
```

```
# Okt 객체 선언
okt = Okt()
```

② 전처리

데이터 정제 과정으로 특수문자와 공백을 제거하였다. 그리고 맞춤법을 검사하여 맞춤법을 체크하였다. 그 이후에 형태소를 분석하여 명사만 분리하였다.

<그림 18> 명사 분리 함수

```
def segmentation_sub(tagged_review):
    noun = []
    noun_count = 0
    for sentence in tagged_review:
        if sentence[1] == 'Noun':
            noun.append(sentence)
            noun_count += 1
        else: continue
    return noun, noun_count
```

토큰화 결과 ‘Noun’에 해당하는 경우만 리스트에 담는다.

<그림 19> 크롤링 함수

```
def noun_segmentation(review):

    ### (1) 정제 ###

    # 텍스트 양옆 공백 제거
    review = review.strip()

    # ParseError 일으키는 특수문자 제거
    review = review.replace("&", " and ")
    review = review.replace("~", "")

    # 텍스트 중간 공백은 하나만 남기기
    review = " ".join(review.split())

    ### (2) 맞춤법 체크 ###
    try:
        spell_check = spell_checker.check(review)
        spell_dict = spell_check.as_dict()

        if spell_dict['result'] == True:
            # 맞춤법 체크 성공 >> True가 나오는 경우
            spelled = spell_dict['checked']
            spell_result = "TRUE"

        else:
            # 맞춤법 체크 실패 >> False가 나오는 경우
            spelled = review
            spell_result = "FALSE"
```

```
    except:
        spelled = review
        spell_result = "FALSE"

    ### (3) 토큰화 및 품사태깅 ###
    tagged_review = okt.pos(spelled)

    ### (4) 명사 분리 ###
    noun, count = segmentation_sub(tagged_review)

    return noun, count, tagged_review
```

데이터 정제 과정, 토큰화를 거친다. 이후 명사를 분리하여 명사, 명사 개수 count, 태그 된 리뷰를 return한다.

<그림 20> 크롤링 한 결과를 최종 리스트에 저장

```
tagged_noun = []
noun_count = []
tagged_review = []

for review in tqdm(raw_review):
    noun, count, tagged = noun_segmentation(review)

    tagged_noun.append(noun)
    noun_count.append(count)
    tagged_review.append(tagged)

    time.sleep(0.05)
```

return한 명사, 명사 개수 count, 태그 된 리뷰 데이터를 최종 리스트인 tagged_noun, noun_count, tagged_review에 저장한다.

③ 추출한 tagged_noun에서 명사만을 뽑아 빈도를 확인하기

tagged_noun은 (‘아이’ , ‘Noun’)의 형태로 저장되어있다. 여기서 ‘아이’ 만을 분리하도록 한다.

<그림 21> 명사 추출 함수 및 명사 추출

```
def sub_func(review):
    noun = []
    for sentence in review:
        noun.append(sentence[0])
    return noun
```

```
noun = []
for review in tqdm(tagged_noun):
    w = sub_func(review)

    noun.append(w)
```

<그림 22> 명사 빈도 저장(예: (’배우’, 1926))

```
# 단일 리스트로 만들기
noun_list = []
for w in noun:
    noun_list += w
```

```
counts = collections.Counter(noun_list)
```

<그림 23> 내림차순으로 정리

```
count_sort = sorted(counts.items(), key=lambda k:k[1], reverse=True)
```

크롤링 한 리뷰에서 명사는 총 62749개가 추출되었다. 그리고 명사 빈도수 측정을 통해 언급된 서로 다른 명사는 총 4385개가 추출되었다. 추출된 명사 빈도수를 내림차순으로 정리하여 어떤 명사가 많이 언급되었는지를 확인할 수 있었다.

④ 최종 결과를 데이터 프레임에 저장하여, 엑셀 파일로 저장하기

전처리 파일과 명사 빈도 파일을 생성한다.

<그림 24> 전처리 파일에 새로운 컬럼 추가

```
# 전처리 파일
raw['Noun count'] = noun_count
raw['Tagged Noun'] = tagged_noun
raw['Noun'] = noun
raw['Tagged review'] = tagged_review
```

전처리 결과 후 뽑힌 명사 count, 태그 된 명사, 명사, 태그 된 리뷰 컬럼을 기존 raw 데이터 프레임에 추가한다.

<그림 25> 전처리 엑셀 파일 저장

```
writer = pd.ExcelWriter(
    'tagged_sentence_review.xlsx',
    engine = 'xlsxwriter'
)

raw.to_excel(writer, sheet_name = "preprocessed")
writer.save()
```

<그림 26> 전처리 엑셀 파일 'tagged_sentence_review.xlsx'

	공연명	공연 장소	공연 기간	공연 시간	관람 연령	가격	장르	리뷰 내용	예약자ID	방문 날짜	Noun count	Tagged Noun	Noun	Tagged review
0	장화 신은	국립중앙박	2019.12.14	70분	36개월이상	66,000원	뮤지컬	11세되는	rlqm****	2020. 2. 3	13	[('세', 'Noun'), ('	['세', '아이'	[('11', 'Number'), ('세'
1	장화 신은	국립중앙박	2019.12.14	70분	36개월이상	66,000원	뮤지컬	아이들이	huni****	2020. 2. 3	1	[('아이', 'Noun')]	['아이']	[('아이', 'Noun'), ('들',
2	장화 신은	국립중앙박	2019.12.14	70분	36개월이상	66,000원	뮤지컬	아이가 재	best****	2020. 2. 3	1	[('아이', 'Noun')]	['아이']	[('아이', 'Noun'), ('가',

전처리를 완료한 엑셀 파일을 생성하였다.

<그림 27> 명사 빈도 엑셀 파일 저장

```
writer = pd.ExcelWriter(
    'noun_frequency.xlsx',
    engine = 'xlsxwriter'
)

counts_df.to_excel(writer, sheet_name = "frequency")
writer.save()
```

<그림 28> 명사 빈도 엑셀 파일 ‘noun_frequency.xlsx’

	Noun	Frequenc
0	배우	1926
1	공연	1457
2	연극	1276
3	연기	1255
4	아이	1121
5	정말	1061
6	시간	1032
7	보고	937
8	뮤지컬	752
9	것	697
10	수	615

명사 빈도 결과를 확인하고, 엑셀 파일을 생성하였다.

나) 소비건수와 공연예매건수 간의 다중선형회귀분석

• 방법론

14가지 항목의 소비건수를 독립변수로 하고 공연예매건수를 종속변수로 하는 다중선형회귀분석을 통하여 40대, 50대의 공연예매에 영향을 미치는 소비항목을 분석한다.

< 데이터 전처리 >

① 파일 불러오기

pandas 라이브러리를 이용하여 KOPIS 데이터와 카드소비변화 데이터를 읽는다. 카드소비 변화 데이터는 한국데이터거래소에서 무료로 제공하는 삼성카드의 ‘코로나로 인한 카드소비 변화’ 데이터를 이용한다.

<그림 29> 카드소비변화 데이터 읽기

```
consumption = pd.read_csv('./data/DATA_SSC_CORONA_MERS.csv')
consumption
```

	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
...	...	...	...	...	...

② 카드소비 변화 데이터에서 연령대가 40·50대이고 '소비일자'가 2020년인 소비건수의 데이터를 추출하기

<그림 30> 카드소비변화 데이터 중 조건에 맞는 데이터 추출

```
consumption = consumption[consumption['소비일자_년'] == 2020]

consumption = consumption[(consumption['연령대'] == '50대') | (consumption['연령대'] == '40대')]
consumption_sum = consumption.groupby(['소비일자', '소비업종'])['소비건수합계'].sum()
consumption_sum = pd.DataFrame(consumption_sum)
consumption_sum.reset_index(inplace = True)

corr_4050 = consumption_sum[consumption_sum['소비업종'] == '가전/가구']['소비일자']
corr_4050 = pd.DataFrame(corr_4050)
corr_4050.set_index('소비일자', inplace = True)

col_list = consumption_sum['소비업종'].unique()
for i in col_list:
    corr_4050 = addCol_4050(i, corr_4050)

def addCol(sector, w_40):
    temp = consumption_40_w[consumption_40_w['소비업종'] == sector][['소비일자', '소비건수합계']]
    temp.set_index('소비일자', inplace = True)
    w_40 = pd.concat([w_40, temp], axis = 1)
    w_40.rename(columns = {'소비건수합계' : '소비건수_' + sector}, inplace = True)
    return w_40
```

카드소비 변화 데이터에서 연령대가 40·50대이고 '소비일자'의 연도가 2020년인 데이터를 추출한다. 해당 데이터는 2015, 2019, 2020년도의 5, 6, 7월 데이터만 포함하고 있다. 날짜를 인덱스로 설정하고 소비항목을 컬럼으로 설정하여 회귀분석을 위한 데이터 세트를 만든다.

③ KOPIS 데이터 중 연령대가 40·50대이고 예매일시 연도가 2020년인 데이터 추출하기

<그림 31> KOPIS 데이터 중 조건에 맞는 데이터 추출

```
data = data[data['예매일시_년'] == 2020]
data = data[(data['예매일시_월'] == 5) | (data['예매일시_월'] == 6) | (data['예매일시_월'] == 7)]

data_4050 = data[((data['연령대'] == 40) | (data['연령대'] == 50)) & (data['예매/취소구분'] == 1)]
data_4050_group = data_4050.groupby('예매일시_날짜')['입장권고유번호'].count()

corr_4050 = addCol_p_4050(data_4050_group, '공연_예매건수', corr_4050)

def addCol_p(temp, colName, w_40):
    w_40 = pd.concat([w_40, temp], axis = 1)
    w_40.rename(columns = {'입장권고유번호' : colName}, inplace = True)
    return w_40
```

출생연도에 따라 실제 나이를 계산한 후 '연령대' 컬럼을 생성하여 연령대를 저장한다. 예매일자 데이터를 연도, 월, 일로 분리하여 2020년에 해당하는 데이터를 추출한다. 카드소비 변화 데이터가 5~7월만 포함하고 있으므로 KOPIS 데이터 또한 예매일시가 5~7월인 데이터로 정제한다. 또한 '예매/취소구분'이 1인 데이터를 추출하여 예매취소를 제외한 예매건수만 사용한다. 날짜 단위로 groupby()를 사용하여 하루의 공연예매건수를 count 하여 저장한다. 앞서 만들었던 소비건수 데이터에 공연예매건수 컬럼을 추가하여 데이터 전처리를 완료한다.

<그림 32> 전처리 후 데이터

	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

< 다중선형회귀분석 >

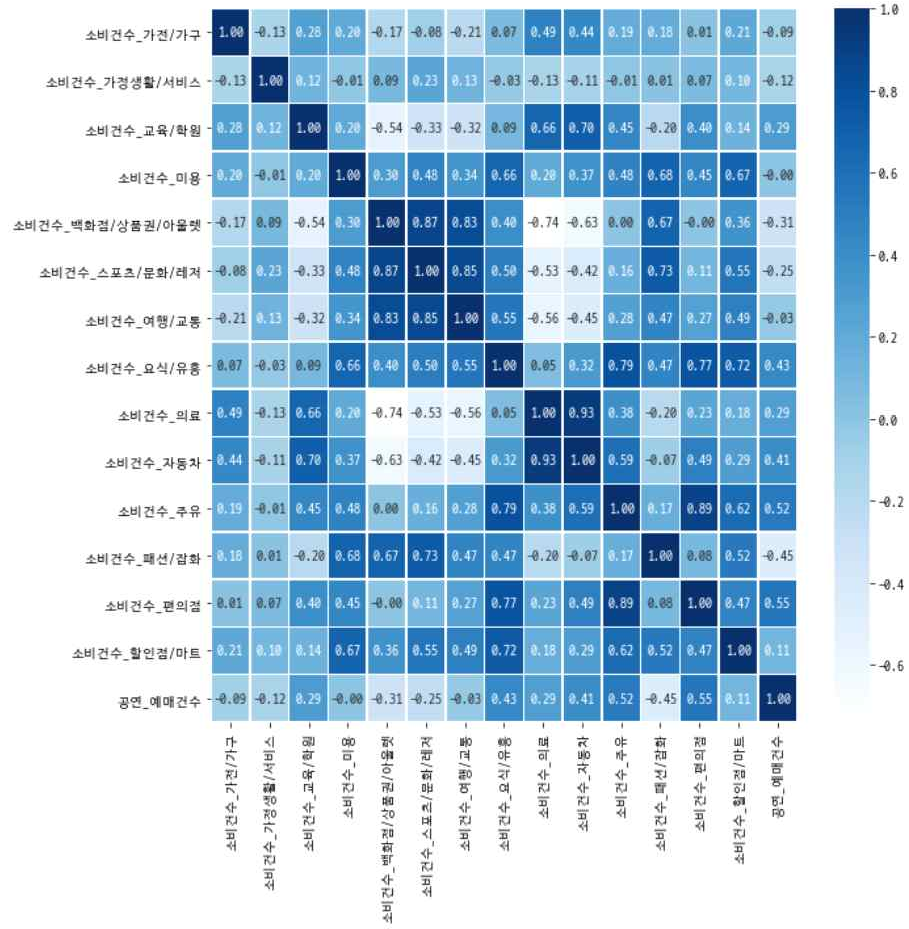
① 히트맵과 산점도 그래프를 통해 공연예매건수와 항목별 소비건수 간의 상관관계를 확인하기

다중회귀분석을 하기 전에 종속변수(공연_예매건수)와 독립변수(소비건수)의 상관관계를 확인하여 선형회귀분석의 가능성을 파악한다. 일부 변수에서 선형성을 보이므로 다중선형회귀분석을 진행한다. 또한 설명변수들 사이에 상관계수가 0.8 이상인 경우 다중공선성이 의심되므로 이를 해

결하기 위한 절차를 진행한다.

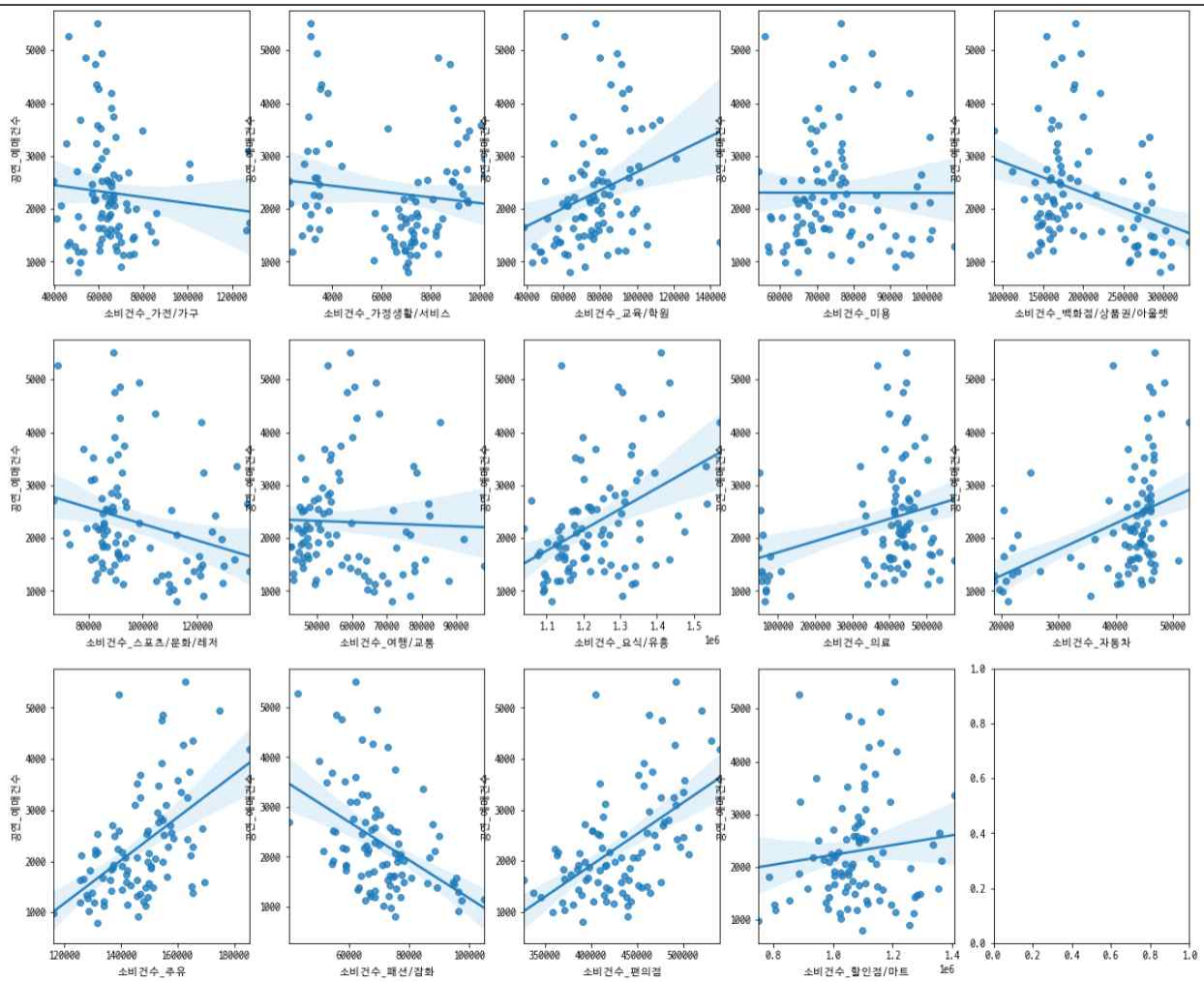
<그림 33> 독립변수와 종속변수간의 상관관계 확인

```
import seaborn as sns
import matplotlib.pyplot as plt
plt.rcParams['font.family'] = 'D2Coding'
plt.figure(figsize = (10, 10))
sns.heatmap(data = corr_w_4050.corr(), annot = True, fmt = '.2f', linewidths = .5, cmap = 'Blues')
plt.show()
```



<그림 34> 종속변수와 각각의 독립변수에 대한 산점도 그래프

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.rcParams['font.family'] = 'D2Coding'

fig, axs = plt.subplots(figsize = (19, 16), ncols = 5, nrows = 3)
x_features = corr_w_4050.drop('공연_예매건수', axis = 1).columns

for i, feature in enumerate(x_features):
    row = int(i / 5)
    col = i % 5
    sns.regplot(x = feature, y = '공연_예매건수', data = corr_w_4050, ax = axs[row][col])
```



② VIF(Variance Inflation Factor)를 통하여 다중공선성 확인

<그림 35> VIF를 통한 다중공선성 확인

```
# 데이터 불러오기
corr_m_4050 = pd.read_csv('./data/corr_m_4050.csv', index_col = 0)

# 다중선형회귀분석
x_data = corr_m_4050.drop('공연_예매건수', axis = 1) # 변수 n개
target = corr_m_4050['공연_예매건수']

# 상수항 추가
x_data1 = sm.add_constant(x_data, has_constant = "add")

from statsmodels.stats.outliers_influence import variance_inflation_factor

vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor(x_data1.values, i) for i in range(x_data1.shape[1])]
vif["features"] = x_data1.columns
```

```
x_data = corr_m_4050.drop(['공연_예매건수', '소비건수_스포츠/문화/레저', '소비건수_백화점/상품권/아울렛',
                           '소비건수_요식/유흥', '소비건수_자동차', '소비건수_미용',
                           '소비건수_교육/학원', '소비건수_주유', '소비건수_의료',
                           '소비건수_가정생활/서비스', '소비건수_할인점/마트', '소비건수_가전/가구'], axis = 1)
x_data1 = sm.add_constant(x_data, has_constant = "add")
```

	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

다중공선성을 확인할 때 쓰는 지표인 VIF를 이용하여, 다중공선성을 가장 크게 유발하는 설명변수를 제거한다. VIF Factor가 10 이상일 경우 다중공선성이 있다고 판단하며, 컬럼을 하나씩 제거하면서 상수를 제외한 모든 컬럼이 10 이하의 값이 될 때까지 반복한다.

③ statsmodels 라이브러리의 OLS를 이용하여 다중선형회귀를 수행하기

다중선형회귀분석을 수행하기 위한 라이브러리 scikit-learn 과 statsmodels 중에서, 가장 적합한 회귀모델을 찾기보다는 관련 변수와 효과 크기를 확인하는 것에 중점을 두기 위해 statsmodels 라이브러리를 사용하여 분석을 진행한다.

<그림 36> OLS 검정

```
# OLS 검정
multi_model = sm.OLS(target, x_data1)
fitted_multi_model = multi_model.fit()
fitted_multi_model.summary()
```

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

최소자승법으로 회귀 모델을 구하는 메서드인 OLS(Ordinary Least Square)를 이용하여 다중선형 회귀분석을 진행한다. VIF를 통해 변수를 제거한 데이터를 적합 시킨 후 결과를 확인한다. 그리고 후진 제거법(Backward Elimination)을 이용하여 중요성이 떨어진다고 판단되는 독립변수를 하나씩 제거한다. 중요한 독립변수를 제거하는 것을 방지하기 위해 통상적인 유의수준인 0.05보다 덜 엄격한 0.1을 우선적인 기준으로 사용하였다. p-value 값이 0.1을 초과할 경우 변수를 제거하되, Adj. R-squared 값에 영향을 미치지 않는 선에서 p-value 값이 0.05를 초과하는 경우도 제거하였다. 그리고 변수를 제거할 때 Adj. R-squared 값과 F-statistics 값을 고려하여 진행한다.

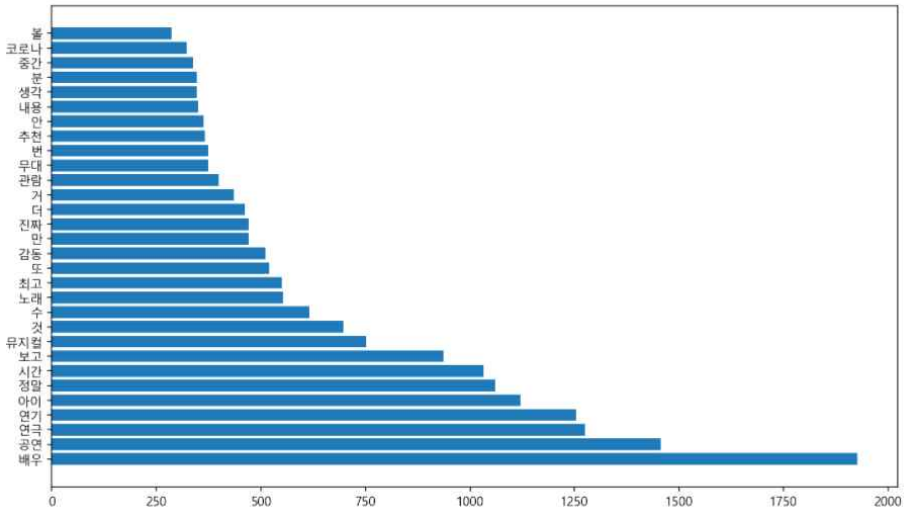
④ 데이터의 범위를 세분화하여 다중회귀분석 시행 후 모델 해석하기

연령대와 성별을 40·50대 전체, 40·50대 남성, 40·50대 여성으로 설정한 3가지의 데이터를 만든다. 먼저 연령대와 성별의 범위가 큰 데이터를 다중회귀분석을 통해 분석하고, 해당 결과에 따라 유의미한 결과를 얻을 것으로 예상되는 데이터의 세분화된 연령과 성별 범위를 재설정한다. 변경된 데이터에 따른 다중회귀분석을 다시 진행한다.

2) 분석 결과

가) 공연 리뷰 데이터를 크롤링 하여 키워드 빈도수 도출 결과

<그림 37> 명사 빈도수 시각화



빈도는 배우 ▶ 공연 ▶ 연극 ▶ 연기 ▶ 아이 ▶ 정말 ▶ 시간 등 순으로 나타난 것을 확인할 수 있다. 여기서 '아이'와 '시간' 명사가 상위에 랭크 되어있다는 것에 주목하였다.

크롤링 파일에서 '아이'와 '시간'이 함께 언급된 리뷰를 찾았다. "아이와 함께 좋은 시간 보냈어요.", "1시간 10분간의 공연시간 동안 아이들이 몰입해서 즐겼습니다.", "아이들 눈높이에 딱 적당한 시간대와 연극이여서 잘 보고 왔어요~^^" 등 아이와 함께한 시간에 초점을 둔 리뷰를 확인할 수 있었다. '아이'와 '시간'이 함께 언급된 리뷰 중에 이와 같은 경향을 띄지 않는 리뷰도 있었다. 하지만 전반적으로 아이와 함께한 시간에 대한 리뷰가 많이 나타난다는 것을 확인했다. 이를 통해, 40·50대가 아이와 함께 시간을 보내기 위해 공연을 관람한다는 경향을 보인다는 결론을 도출해내었다.

또한 크롤링 파일을 확인한 결과 출연배우의 언급이 853건이 있음을 확인하였다. <그림 26>. 이는 9번째로 언급된 '뮤지컬'(752건)보다 많고, 8번째로 언급된 '보고'(937)보다 적은 것을 알 수 있으며, 총 언급 단어 4385개 중 9번째로 많이 언급된 경우이다. 이는 공연을 볼 때 출연배우를 중점으로 관람하는 경우가 적지 않다고 결론을 내릴 수 있다. 출연배우를 언급한 리뷰를 확인하였을 때, "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." 등 공연 관람에 출연배우가 영향을 준 것을 확인할 수 있다. 출연배우 이름을 리뷰에 언급한다는 것은 공연 소비에 있어서 배우의 유명도/인지도가 영향을 준다고 할 수 있다. 따라서 스타마케팅이 40·50대에게 충분한 소비를 불러올 수 있다는 결론을 도출해내었다.

나) 소비건수와 공연예매건수 간의 다중선형회귀분석

· 전체 결과

5개의 모델에서 모든 독립변수의 유의확률(p-value)이 0.05보다 작은 값을 보이며, 잔차의 표준오차(std err)의 값이 0.01 미만이다. 그리고 Prob(F-statistics) 값이 0.05 이하이므로 최소 1개의 독립변수의 기울기가 0이 아니라는 결과를 통해 전체 회귀가 유의미함을 확인하였다.

Adj. R-squared 값은 전체 데이터 중 해당 회귀모델이 설명할 수 있는 데이터의 비율을 모델에 영향을 주는 데이터를 반영하여 조정한 결정계수로, 40·50대 여성의 데이터가 73.2%로 5개의 결과 중 가장 설명력이 높다는 것을 알 수 있다. 또한 F-statistic 값은 F 통계량으로, 도출된 회귀식이 적절한지 볼 수 있고 0과 가까울수록 적절하다는 것을 의미하며 5개의 결과 중 40·50대 남성에서 가장 작은 수치로 회귀식을 설명하고 있다는 것을 알 수 있다.

Adj. R-squared 값 비교: 40·50대 남성 < 40·50대 전체 < 40대 여성 < 50대 여성 < 40·50대 여성

F-statistic 값 비교: 40·50대 남성 < 40·50대 전체 < 50대 여성 < 40·50대 여성 < 40대 여성

40·50대 전체와 40·50대 남성 결과에서 독립변수 '소비건수_편의점' 이 종속변수와 상관도가 가장 큰 것으로 나타나며, 40·50대 여성, 40대 여성, 50대 여성 결과에서 독립변수 '소비건수_요식/유흥' 이 종속변수와 상관도가 가장 큰 것으로 나타난다. 모든 모델에서 독립변수 '소비건수_패션/잡화' 가 1 증가할 때 종속변수는 감소하는 양상을 보인다.

① 40·50대 전체

<그림 38> 40·50대 전체 OLS 결과

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

회귀식: $y = -1196.7868 + 0.0082x_1 - 0.0303x_2 + 0.005x_3$

(x_1: 소비건수_여행/교통, x_2: 소비건수_패션/잡화, x_3: 소비건수_편의점)

Adj. R-squared 값이 0.546으로 추정된 회귀모델로 데이터의 54.6%를 설명할 수 있고,

F-statistics 값은 5개의 모델 중 2번째로 작은 값을 보인다. t(t-test)의 값을 보았을 때 독립변수 '소비건수_편의점' 이 종속변수와의 상관도가 가장 크게 나타나고 있다.

② 40·50대 남성

<그림 39> 40·50대 남성 OLS 결과

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

회귀식: $y = -323.6474 + 0.0036x_1 - 0.0081x_2 + 0.0011x_3$

(x_1: 소비건수_여행/교통, x_2: 소비건수_패션/잡화, x_3: 소비건수_편의점)

Adj. R-squared 값이 0.433으로 추정된 회귀모델로 데이터의 43.3%를 설명할 수 있고, F-statistics 값이 5개의 모델 중 가장 작은 값이기 때문에 F 통계량으로 도출된 회귀식이 비교적 적절하다. t(t-test)의 값을 보았을 때 독립변수 '소비건수_편의점' 이 종속변수와의 상관도가 가장 크게 나타나고 있다.

③ 40·50대 여성

<그림 40> 40·50대 여성 OLS 결과

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

회귀식: $y = -2222.6588 - 0.0112x_1 + 0.008x_2 - 0.0675x_3$

(x_1: 소비건수_여행/교통, x_2: 소비건수_요식/유흥, x_3: 소비건수_패션/잡화)

Adj. R-squared 값이 0.732으로 추정된 회귀모델로 데이터의 73.2%를 설명할 수 있고, F-statistics 값은 5개의 모델 중 4번째로 낮은 값을 보인다. t(t-test)의 값을 보았을 때 독립변수 '소비건수_요식/유흥'이 종속변수와의 상관도가 가장 크게 나타나고 있다.

④ 40대 여성

<그림 41> 40대 여성 OLS 결과

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

회귀식: $y = -1696.7315 + 0.0096x_1 - 0.0978x_2$

(x_1: 소비건수_요식/유흥, x_2: 소비건수_패션/잡화)

Adj. R-squared 값이 0.673으로 추정된 회귀모델로 데이터의 67.3%를 설명할 수 있다. F-statistics 값은 5개의 모델 중 가장 높은 값이기 때문에 비교적 회귀식을 설명하지 못한다고 볼 수 있다. t(t-test)의 값을 보았을 때 독립변수 '소비건수_요식/유흥' 이 종속변수와의 상관도가 가장 크게 나타나고 있다.

⑤ 50대 여성

<그림 42> 50대 여성 OLS 결과

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

회귀식: $y = -281.2682 - 0.0032x_1 - 0.0094x_2 + 0.0043x_3 - 0.0188x_4$

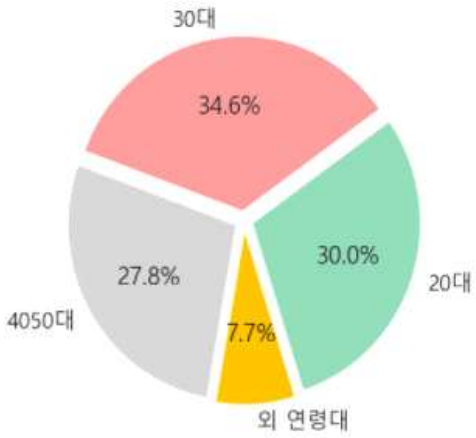
(x_1: 소비건수_백화점/상품권/아울렛, x_2: 소비건수_스포츠/문화/레저,

x_3: 소비건수_요식/유흥, x_4: 소비건수_패션/잡화)

Adj. R-squared 값이 0.730으로 추정된 회귀모델로 데이터의 73%를 설명할 수 있고, F-statistics 값은 5개의 모델 중 3번째로 작은 값을 보인다. t(t-test)의 값을 보았을 때 독립변수 '소비건수_요식/유흥' 이 종속변수와의 상관도가 가장 크게 나타나고 있다.

3) 활용 가능성 및 방향

<그림 43> 연령별 공연 예매 비율



공연을 관람한 전체 연령 데이터 수에서 40·50대 비율은 27.8%로 30대 34.6%, 20대 30% 다음으로 공연 관람을 많이 한 것을 알 수 있다. 40·50대의 비율이 30대, 20대의 비율과 크게 차이 나지 않다는 것을 확인하였다. 이를 통해 40·50대를 타깃으로 한 공연을 제작할 때, 아이와 함께 시간을 보낼 수 있는 공연을 제작한다면 40·50대 공연 소비를 증대시킬 수 있고, 전체 공연 매출에도 기여를 할 수 있다고 본다. 또한 출연배우의 언급이 전체 명사 4385개 중 9번째로 언급된 것을 통해 스타마케팅이 40·50대 공연 소비에 충분한 영향을 줄 것이라고 판단한다.

다중회귀분석 모델로 각 항목의 소비건수 데이터를 이용하여 40·50대 공연예술 예매건수를 예측할 수 있다. 데이터 설명력이 가장 높게 나온 50대 여성 결과의 회귀식에서 여행/교통 소비건수, 요식/유흥 소비건수, 패션/잡화 소비건수가 각각 1 증가할 때 공연예매건수가 0.0112 감소, 0.008 증가, 0.0675 감소한다는 결과를 얻을 수 있었다. 각각의 소비건수를 알 수 있다면 공연예매건수를 예측하여 공연예술 활성화를 위한 전략을 제시하는 것에 도움이 될 수 있다.

또한 공연예술 예매에 영향을 미치는 소비항목을 마케팅에 접목하여 효율적인 효과를 기대할 수 있다. 50대 여성 결과에서 소비건수_요식/유흥이 공연예매건수와 가장 큰 상관도를 가지고 있다는 결과를 얻었다. 이를 통해 요식/유흥 소비가 늘어날 때 공연예술 참여를 유도하는 마케팅을 하여 40·50대의 공연예술 활성화에 기여할 수 있다.

4) 기대효과

향후 키워드 분석을 통해 나온 명사를 기반으로 토픽 모델링을 진행해 볼 수 있다. 토픽 모델링 기법을 활용하여 리뷰 속 주제를 발견하는 연구를 통해 공연 흥행 요소를 확인하여 공연 매출 증대 방안을 제시해 볼 수 있다. 또한 다중선형회귀분석의 결과를 근거로 장바구니 분석을 후속 연구로 진행할 수 있다. 공연을 예매한 관람객의 결제내역을 확인하여 공연 전후로 소비하는 상

품이 무엇인지 분석하고 공연 관람객의 소비 트렌드를 예측하여 공연 매출 증대 및 주변 상권 활성화를 기대할 수 있을 것이다. 공연예술 제작사는 리뷰 크롤링 결과를 통해 인지하게 된 선호 주제를 토대로 공연을 제작하고 다중선형회귀분석의 결과를 통해 40·50대의 예매를 예측하여 관련 기획, 회계 시스템 등을 조절할 수 있을 것이다. 제작사에 있어 수요를 예측하고 이에 맞는 효율적인 공급을 제공하는 것은 무엇보다 우선시되어야 하는 사항이다.

다중선형회귀분석을 통해 파악한 40·50대의 소비 트렌드에 맞추어 공연예술 마케팅을 진행할 수 있다. 예를 들어 공연장 주변 음식점의 영수증을 증빙하였을 때 공연예술 할인을 추가적으로 해준다면 공연예술이 40·50대들에게 더 이상 특별하게 여겨지는 장르가 아니게 된다. 그렇게 접근성을 높임으로써 여가생활의 일부분으로 자리 잡을 수 있도록 유도한다. 또한 패션/잡화를 구매할 경우, 공연예술의 예매율이 낮아지는 것으로 보아 동시에 두 가지를 하기엔 부담이 된다는 추측을 할 수 있다. 그렇기에 특정 패션/잡화 브랜드와 예술을 협업시킴으로써 특정 전시나 공연을 관람을 했을 경우에만 구매할 수 있는 할인된 패션/잡화 굿즈를 판매한다면 이에 대한 문제를 해결할 수 있을 것이다. 더불어 특정 브랜드를 선호하는 경제력을 갖춘 마니아층을 공연예술계로 유입시킬 수 있을 것이다. 40·50대가 경제력을 갖춘 세대이기 때문에 공연예술에 대한 벽을 낮추고 접근성을 높이기만 한다면 공연예술에 소비를 할 것이라고 기대할 수 있다. 이러한 예시와 같이 구체적인 마케팅 방식을 도출할 수 있으며, 이는 40·50대가 새로운 관객층으로 유입되어 공연예술 시장의 활성화를 불러올 것이다.

5) 한계점

크롤링 한 리뷰에서 명사를 분리하기 위해 Okt 형태소 분석기를 사용하였다. ‘Noun’ 에 해당하는 단어를 분류하였으나, 실제로 ‘Noun’ 에 해당하는 단어를 보았을 때 명사가 아닌 경우 (예: ‘보고’, ‘것’) 가 존재하였다. 현재 연구에서는 명사 분류의 정확도가 중점이 되지 않기 때문에 이 문제는 중점적으로 다루지 않았지만, 이후에 Okt 형태소 분석기를 이용하여 텍스트 데이터에서 명사를 분류할 때 명사를 분류하는 기준이 추가적으로 필요할 것이다.

또한 리뷰 데이터에서 나타난 결과가 40·50대 전체의 경향성을 파악하기에는 무리가 있다. 리뷰 사이트에서는 개인정보를 공개하지 않아 리뷰 작성자의 연령을 파악할 수 없기 때문이다. 따라서 리뷰 데이터 분석 당시 KOPIS 데이터에서 40·50대가 많이 본 공연을 기준으로 리뷰를 크롤링하였으나, 리뷰 중에 40·50대가 아닌 작성자들도 분명히 존재할 것이라고 예상된다. 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.

소비건수와 공연예매건수의 다중선형회귀분석에서 삼성카드의 소비건수 데이터가 5월~7월에 대해서만 무료로 제공하고 있어 2020년 전체의 소비패턴을 반영하기에 어려움이 있었다. 2020년 전체 소비건수 데이터를 얻을 수 있었다면, 2020년 한 해의 경향을 확인하고 더 나아가 다중회귀

분석 모델의 설명력 또한 높일 수 있을 것이다.

다중선형회귀분석에서 후진 제거법을 이용하여 다중공선성을 최소화했지만, 그럼에도 모든 모델에서 다중공선성이 다소 높은 수치를 보인다. 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.

• 기타

1) 활용 데이터 목록 및 출처

- KOPIS 데이터
- 네이버 예약 사이트 공연 리뷰 크롤링 (출처: https://booking.naver.com/)

※ 엑셀파일
corr_4050.csv, corr_m_4050.csv, corr_w_4050.csv, corr_w_40.csv, corr_w_50.csv
tagged_sentence_review.xlsx, noun_frequency.xlsx

- KDX 한국데이터거래소, 삼성카드 코로나로 인한 카드소비 변화 (메르스와 비교)
(https://kdx.kr/data/view/27200)

※ 코드 파일
play_frequency.ipynb, review_tokenization.ipynb, 인터파크_review_crawling.ipynb
소비건수_공연예매건수 _다중선형회귀분석.ipynb

2) 참고문헌

김광석, 김수경, 차윤지(2017),「고령사회 진입과 시니어 비즈니스의 기회」,Samjong insight 49호,p.1-7.
김은정(2009),「뮤지컬관객의 티켓구매 결정요인에 관한 연구」,p.89.
김하연, 서대호(2019), 토픽 모델링 기반 국내 공연 흥행 요인 분석,p.108-109.
이지영(2020),『데이터 과학 기반의 파이썬 빅데이터 분석』,p.305-327.
시모야마 테루마사, 마쯔다 유마, 미키 타카유키(2020),『파이썬 데이터 분석 실무 테크닉 100』,p.4-129.

웹페이지
- 국립극단, “국립극단 청소년극”, 2021년 9월 12일 접속, https://www.ntck.or.kr/ko/ntck/child_teen.
- “[회귀분석] 회귀분석 실습(1) - OLS 회귀분석 결과 해석 및 범주형 변수 처리(Statsmodel)“, YSY의 데이터분석 블로그, 2021년 08월 28일 수정, 2021년 9월 11일 접속, https://ysyblog.tistory.com/119.
- “Multiple Linear Regression Analysis 다중 선형 회귀 분석”, 多言數窮, 2011년 1월 19일 수정, 2021년 9월 11일 접속, https://m.blog.naver.com/libido1014/120122338861.
- “Scikit-learn vs. StatsModels: Which, why, and how?”, The Data Incubator, 2017년 11월 8일 수정, 2021년 9월 12일 접속,

https://www.thedataincubator.com/blog/2017/11/08/scikit-learn-vs-statsmodels/.
- sgis통계지리정보서비스, “통계지리정보서비스 “, https://sgis.kostat.go.kr/view/index, 2021년 09월 12일 접속.