|  |
| --- |
| **제11회「2023 빅콘테스트」결과보고서** |

|  |  |  |  |
| --- | --- | --- | --- |
|  | \* 해당란에 표시 | | |
| **참가분야** | □ 생성형AI 분야 □ 데이터신기술 분야  □ 정형데이터 분석 분야 □ 비정형데이터 분석 분야  □ 빅데이터플랫폼 활용 분야 | | |
| **세부리그**  \*해당시 체크 | □ 어드밴스드 리그 □ 스타터 리그  \*정형데이터 분석분야에 한함(선택) | | |
| □ 지정주제 리그 □ 자유주제 리그  \*빅데이터플랫폼 활용분야에 한함(선택) | | |
| **개인/팀여부** | □ 개인 □ 팀(총 4명) | **개인/팀명** | What’s your EDA |
| **지도교사명** | \*스타터 리그에 한함(선택) | | |
| **대표ID** | alwaysna0813@naver.com | | |

|  |  |
| --- | --- |
| **결과보고서 작성 안내 사항** | |
| **목차**  **(예시)** | **Ⅰ. 개요**  1. 배경  3. 분석 수행 범위    **Ⅱ. 문제 수행 내용**  1. 분석(수행) 절차  2. 분석(수행) 내용 및 결과    **Ⅲ. 주요 결과 및 시사점**  1. 주요 결과 요약  2. 결과 활용 및 시사점 |
| **작성방향** | - 결과보고서는 **30장 내외** 로 목차를 준수하여 작성하여야 하며, 필요시 목차 구성에 항목을 추가하여 자유롭게 작성  - 그림 및 도표 등 활용 가능  - 출처 명시(참고 문헌/논문, 이미지, 저자, 사이트 URL 등) |
| **글꼴 및**  **글자크기** | - 본문 글꼴 : 맑은 고딕  - 대분류[1, 2, 3] 항목 : 13포인트(굸게)  - 중분류[가, 나, 다] 항목 : 12포인트(굵게)  - 소분류[ 1), 2), 3) ] 항목 : 12포인트, 본문내용 : 10포인트 |

**< 목차 >**

**l. 개요**

1. 분석 배경 및 분석 필요성

가. 기존의 공연 가격 책정 방법의 문제점

나. 클래식 공연의 새로운 가격 책정 방법 도입 필요성

1. 분석 목적
2. 분석 수행 범위

**II. 분석(수행) 절차**

1. 분석(수행) 절차
2. 분석(수행) 내용 및 결과

가. 데이터 수집

1. 데이터 수집 전 기본 작업
2. 예술의 전당 홈페이지 크롤링
3. 크롤링한 정가를 바탕으로 좌석 채우기
4. 예술의 전당 콘서트홀 좌석도 좌표화

나. 1차 EDA

다. 데이터 가공

1. 데이터 정제
2. 데이터 이상치 처리
3. 데이터 결측치 처리
4. 그 외의 전처리

라. 2차 EDA

마. 파생변수 생성

1. 공연 관련 파생변수
2. 고객 관련 파생변수
3. 좌석 관련 파생변수

바. 좌석 그룹핑을 위한 클러스터링

**III. 주요 결과 및 시사점**

1. 주요 결과 요약

2. 결과 활용 및 시사점

|  |  |  |
| --- | --- | --- |
| **Ⅰ.** |  | **개요** |

**1. 문제 배경 및 분석 필요성**

**가. 기존의 공연 가격 책정 방법의 문제점**

현재 공연 티켓 가격의 책정은 주로 투입 제작비와 목표 수익에 따라 좌석 등급을 나누고, 다른 작품과 비슷하게 좌석 등급 별 가격을 산정하는 방식으로 이루어진다.(공연 예술의 동적 가격 결정(2020), 박희영)클래식 공연을 진행하는 예술의 전당의 경우도 가격 책정은 비슷하게 이루어진다. 그러나 고객 특성 보다는 기획사의 경험에 의존하여 진행된 가격 책정은 몇 가지 문제점이 있다.

첫 번째는 공연 수요와 고객의 특성을 파악하지 못해 효과적으로 수익을 창출하지 못하는 문제점이다. 실제로 2017년도 기준 공연예술산업 실태조사에 따르면 우리나라 전체 공연 시설의 재정자립도는 48.6%로 나타났다.(공연 예술의 동적 가격 결정(2020), 박희영) 낮은 재정 자립도는 공연의 질을 낮추고, 결국 소비자가격을 올려 공연의 접근성을 낮추며, 합리적이지 않은 가격으로 소비자는 공연에 대한 신뢰도도 잃을 수 있다. 따라서 효과적인 수익 창출을 위해서는 공연 수요와 고객의 특성을 파악해야 한다.

두 번째는 효과적인 티켓 분배가 이루어지지 않는다. 공연의 이익만큼 중요한 것이 공연의 티켓 예매율이다. 수요가 높은 공연과 좌석의 가격을 조정하는 것도 중요하지만 그렇지 않은 좌석의 구매율을 높이는 일도 중요하다. 이익이 크게 늘지 않더라도 저렴한 가격으로 많은 고객이 방문할 수 있도록 유도한다면 클래식 공연에 대한 관심도도 높아지고, 해당 공연은 증가한 예매율을 바탕으로 더욱 더 많은 고객을 유치할 수 있을 것이다. 또한 국민 문화 향유 기회 확대와 순수 공연 활성화라는 예술의 전당의 설립 목표를 달성할 수 있을 것이다.

**나. 클래식 공연의 새로운 가격 책정 방법 도입 필요성**

흔히 제품의 가격 책정 방법은 효용과 원가 사이의 최적 가격을 결정하는 식으로 이루어진다. 공급자는 순이익을 최대로 할 수 있는 방법으로 생산 비용을 줄이고 매출을 최대로 할 수 있는 최적 가격을 찾는다. 그러나 보통의 제품과 달리 클래식 공연의 경우 마니아층이 많아 가격 민감도가 낮고, 가격 탄력성이 낮다는 특징이 있다. 또한 비교적 높은 가격 탓에 진입 장벽이 있고, 이러한 까닭으로 경제적 여유가 있는 고객이 구매하는 경향이 있다. 이렇게 형성된 특유의 분위기 때문에 클래식 공연은 흔히 ‘고급공연’이라 불리며, 고객과 공연 기획 및 주최사 간의 유대감과 신뢰도는 매우 높다고 볼 수 있다. 따라서 높은 할인율이나 가격, 마케팅 전략보다도 공연의 수준이 고객의 수요에 가장 큰 영향을 미칠 것이라고 예상된다.

그렇기에 공급자에 해당하는 공연자 측은 순이익을 늘리기 위해 공급 비용의 절감을 모색하기 보다는 수요 예측과 그에 따른 적절한 가격 책정을 고려하는 것이 나은 방법이라고 할 수 있다. 이러한 이유로 본 조는 공연자 측의 생산비용은 고려하지 않고 오직 매출을 늘리는 방향으로 분석하고자 하며 클래식 공연 고객의 특성을 반영하여 새로운 가격 책정 방법을 도입할 것이다.

**2. 분석 목적**

수요와 소비자의 특성을 반영한 합리적인 티켓 가격을 책정하는 것을 최종 목표로 한다. 수익 창출을 주요 목표로 하되, 동시에 가격의 합리성도 확보될 수 있고자 한다. 이 과정에서 기존에 매겨진 좌석 등급과는 다른 새로운 좌석 grouping 방법을 모색할 것이고, 새로운 인사이트를 제공하고자 한다.

**3. 분석 수행 범위**

예술의 전당 수요를 늘리는 것과 동시에 수익 창출이 목표이기 때문에 돈을 주고 티켓을 구매한 고객을 대상으로 분석한다. 그렇기에 초대권으로 온 사람들과 전석 초대 공연에 온 사람들은 가격 책정과 좌석 그루핑 분석에서 제외한다. 그러나 eda를 통해 얻을 수 있는 인사이트를 통해 파생변수 생성과 방향성 제시에 사용한다. 또한 2018년~2023년의 데이터 중 코로나 시기와 그렇지 않은 시기가 다른 경향성을 보일 것이기에 따로 분석하여 진행한다.

|  |  |  |
| --- | --- | --- |
| **Ⅱ.** |  | **문제 수행 내용** |

**1. 분석(수행) 절차**

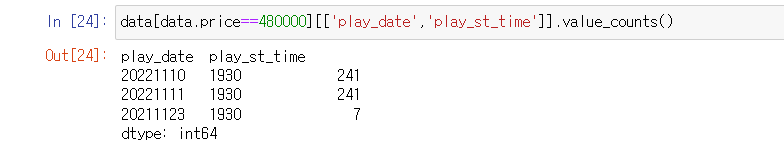
예술의 전당 홈페이지를 기반으로 데이터를 수집하고, 전처리 한 후 좌석 등급 알고리즘으로 고객의 좌석을 파악하였다. 그 후 관련 파생변수를 생성하고 고객 특성과 관련된 변수로 K-means 클러스터링을 통해 공연을 분류한다. 그 후 가설을 설정하고 좌석 그룹핑 및 가격 책정을 한다.

**2. 분석(수행) 내용 및 결과**

**가. 데이터 수집**

1. 데이터 수집 전 기본 작업

원활한 데이터 수집을 위해 기본 작업을 거쳤다. 먼저 모든 행이 결측치인 membership\_type\_6은 제거한다. 또한 본 분석은 콘서트홀에 대해 진행하므로 place열의 값이 콘서트홀인 경우만 추출한다. 다음으로 날짜와 시간에 대한 변수인 tran\_time, play\_st\_time, tran\_date, play\_date, pre\_open\_date, open\_date은 datetime 형식으로 변환해준다. 이 때, play\_date 변수를 홈페이지와 비교해본 결과 실제 공연 날짜보다 7일이 미뤄진 것을 발견하였다. 가령, 데이터에서 가장 높은 가격이었던 48만원짜리 공연을 예술의 전당 홈페이지에서 찾아보면 2022 빈 필하모닉 내한공연 뿐이다. 홈페이지 기준 실제 공연일은 2022년 11월 3일, 2022년 11월 4일, 2022년 11월 16일이었으나 데이터 상에서는 아래의 그림과 같이 2022년 11월 10일, 2022년 11월 11일, 2021년 11월 23일로 모두 7일 미루어졌음을 알 수 있고, 주최 측 문의 결과 데이터 상의 오류라기보다는 데이터 가명화 처리 때문이라 판단하였다. 따라서 play\_date를 7일 당겨준 real\_play\_date 변수를 추가적으로 생성한다.



1. 예술의 전당 홈페이지 크롤링

같은 날 같은 시간에 한 독자적인 공연 751개의 real\_play\_date와 play\_st\_time을 추출한 후 예술의 전당 홈페이지를 참조하여 title을 크롤링해주었다. 그 후 크롤링한 title과 play\_date\_time을 기준으로 입장연령, 관람시간(분), 가격, 주최, 주관, 후원을 크롤링하여 all\_data 파일로 저장한다. 크롤링 한 가격은 R\_price, S\_price, A\_price, B\_price, C\_price, normal(일반), first\_level(1층석), second\_level(2층석), third\_level(3층석), all(전석) 열에 처리하여 숫자형으로 넣는다. 예를 들어 ‘R석 48만원 / S석 39만원 / A석 29만원 / B석 19만원 / C석 9만원’은 R\_price 열에 480000, S\_price 열에 390000, A\_price 열에 290000, B\_price 열에 190,000, C\_price 열에 90,000을 채운다. 크롤링 한 가격에 숫자가 없는 경우가 있는데 이는 모두 전석 초대, 전석 무관중 공연이었으므로 all 열에 0을 채운다.

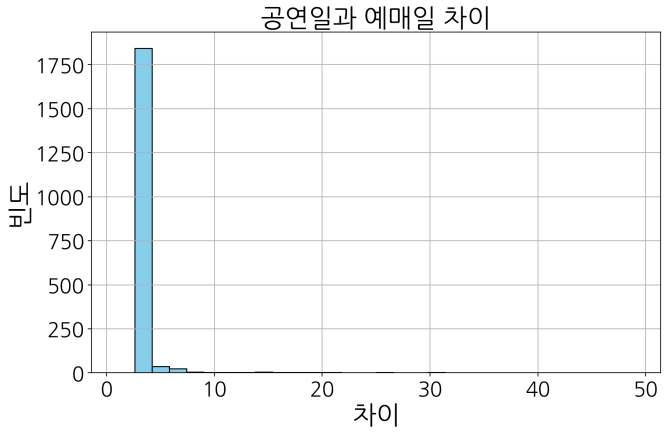
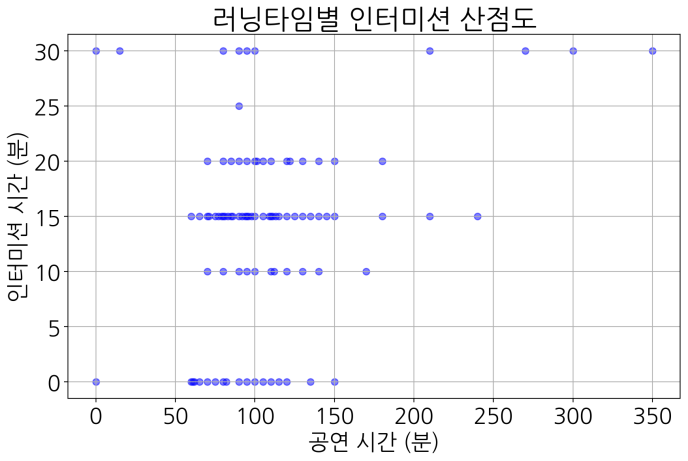
1. 크롤링한 정가를 바탕으로 좌석 채우기

discount\_percent열은 discount\_type을 바탕으로 할인율로 채운다. discount\_type에 ‘%’가 있는 행은 ‘%’ 앞에 있는 숫자를 discount\_percent에 저장하고, ‘일반’이면 할인율이 0%이기 때문에 0을 저장하고, 이에 해당하지 않으면 결측으로 채워준다. 지불한 가격에 할인율을 역적용하여 origin\_price1 열을 형성하고, origin\_price1이 채워져 있는 행에 대해서 실제 좌석을 구하여 seat\_level 열에 채워준다. 단, 원데이터가 마스킹의 문제로 1000원 단위로 버림을 한 데이터이다. 이 경우에 가격을 역적용할 시 소수점으로 나오는 문제가 발생한다. 이를 해결해주기 위해서 역적용한 가격의 허용 범위를 1500원 단위로 두었다. 만약 이 기준을 적용했을 때 그 사람이 어떤 자리에 앉았는지 알 수 있다면 해당하는 자리를 seat\_level 열에 넣어준다. 그러나 이 때 해당하는 가격이 허용 범위 안에 없어 어디에 앉았는지 알 수 없으면 ‘missing’으로, 해당하는 좌석이 2개 이상이면 ‘more’로 채운다. 또한 origin\_price1을 알 수 없는 행에 대해서는 모두 ‘unknown’으로, 모든 행이 0인 공연은 전석초대로 판단하여 ‘free’로 채워주었다. 그 후 가격이 0원이면서 좌석이 unknown인 고객은 모두 좌석을 free로 바꿨다. 또한 좌석은 ‘missing’인데 가격이 모두 0원인 경우에도 ‘free’로 바꿔주었다.

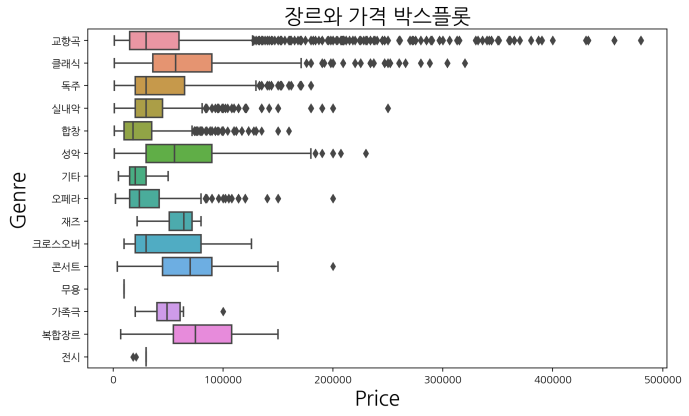
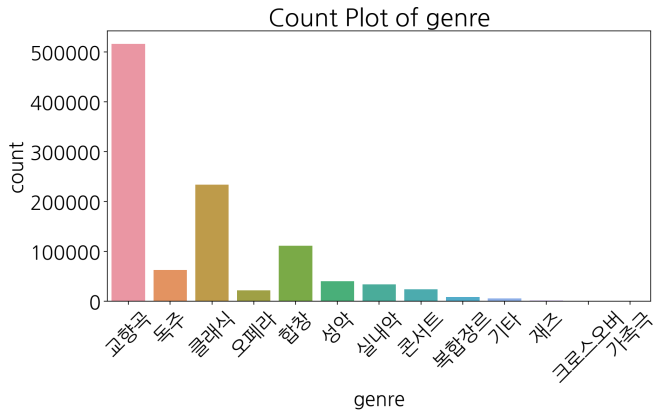
4) 예술의 전당 콘서트홀 좌석도 좌표화

예술의 전당 콘서트홀의 좌석을 좌표화 시키고자 좌석배치도를 이용하여 층, 블록, 열, 번호를 조합하여 좌석배치도와 동일한 모양의 엑셀 파일인 seat.csv를 생성했다. 이후 seat.csv를 데이터프레임으로 변화시킨 후 행과 열을 활용하여 실제 있는 좌석만 좌표로 변환시켰다. 이후 좌석배치도에는 없지만 제공받은 데이터에는 존재하는 좌석을 좌표에 맞게 넣어주었고 seat\_coor.csv로 저장하였다.

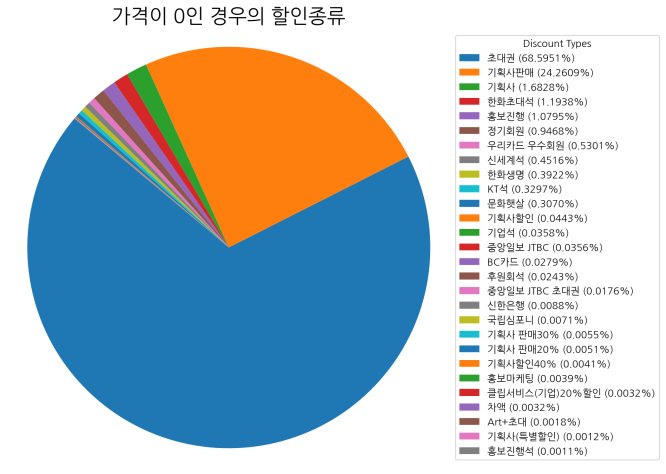
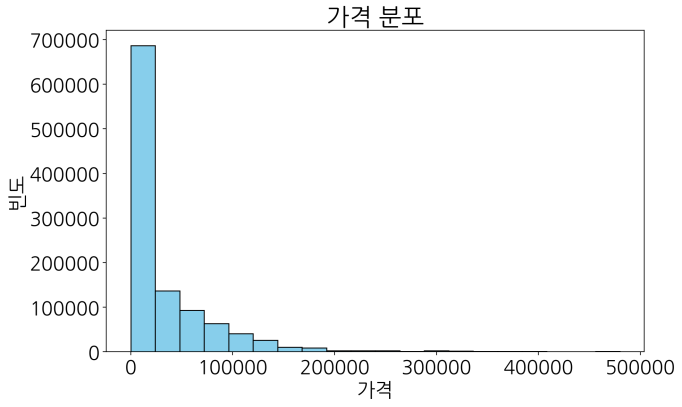
**나. 1차 EDA**

****

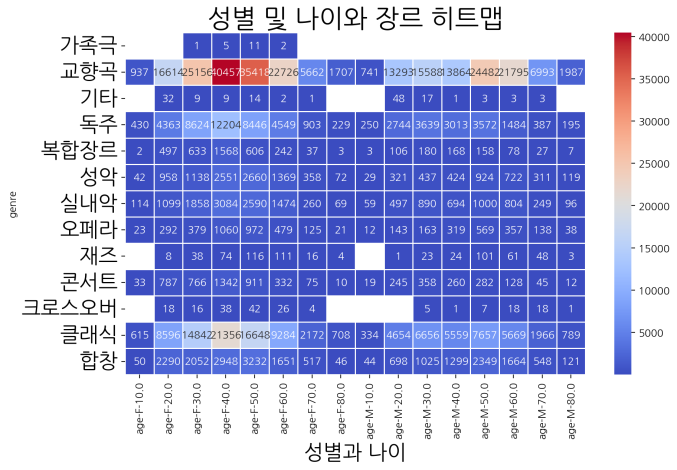
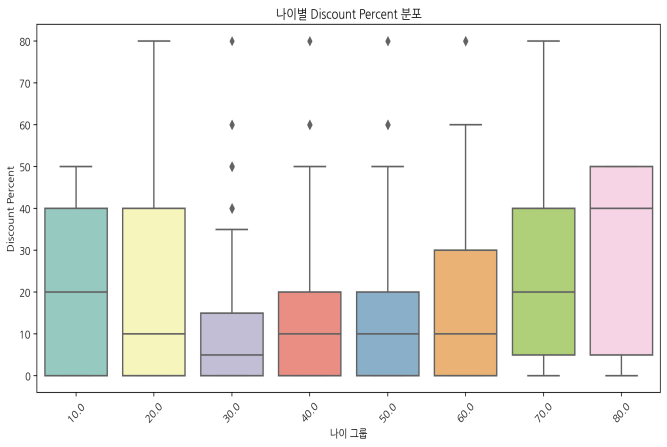
* 공연일과 예매일의 차이에 대한 히스토그램을 시각화해본 결과 4일 차이에 분포가 몰려있었다. 4일 미만으로 차이가 나는 경우는 확인할 수 없었다.
* 러닝타임과 인터미션으로 산점도를 그려본 결과, 러닝타임이 0분, 15분인 공연이 있었다.

****

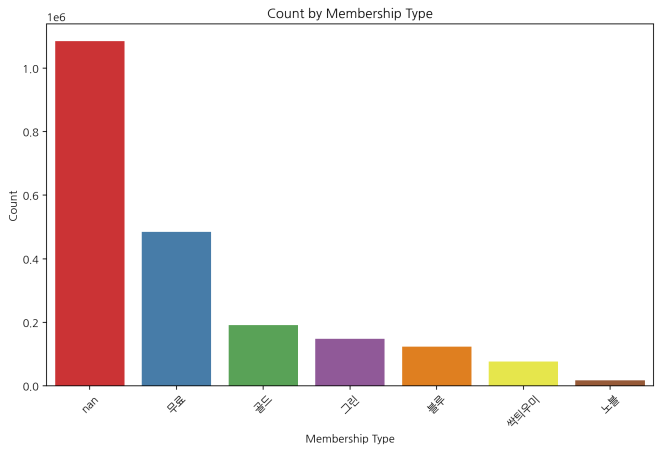
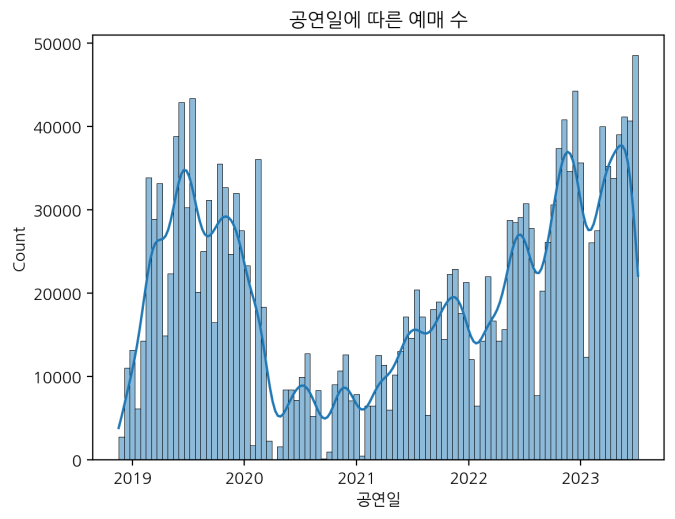
* 장르 히스토그램과 장르별 가격 박스플롯을 그려보았다. 교향곡, 클래식, 합창 순으로 많았고 재즈, 크로스오버, 무용, 가족극, 기타, 복합장르는 거의 없다. 가격은 클래식, 성악, 복합장르가 높은 편에 속했고 기타, 합창이 가장 낮았다.

****

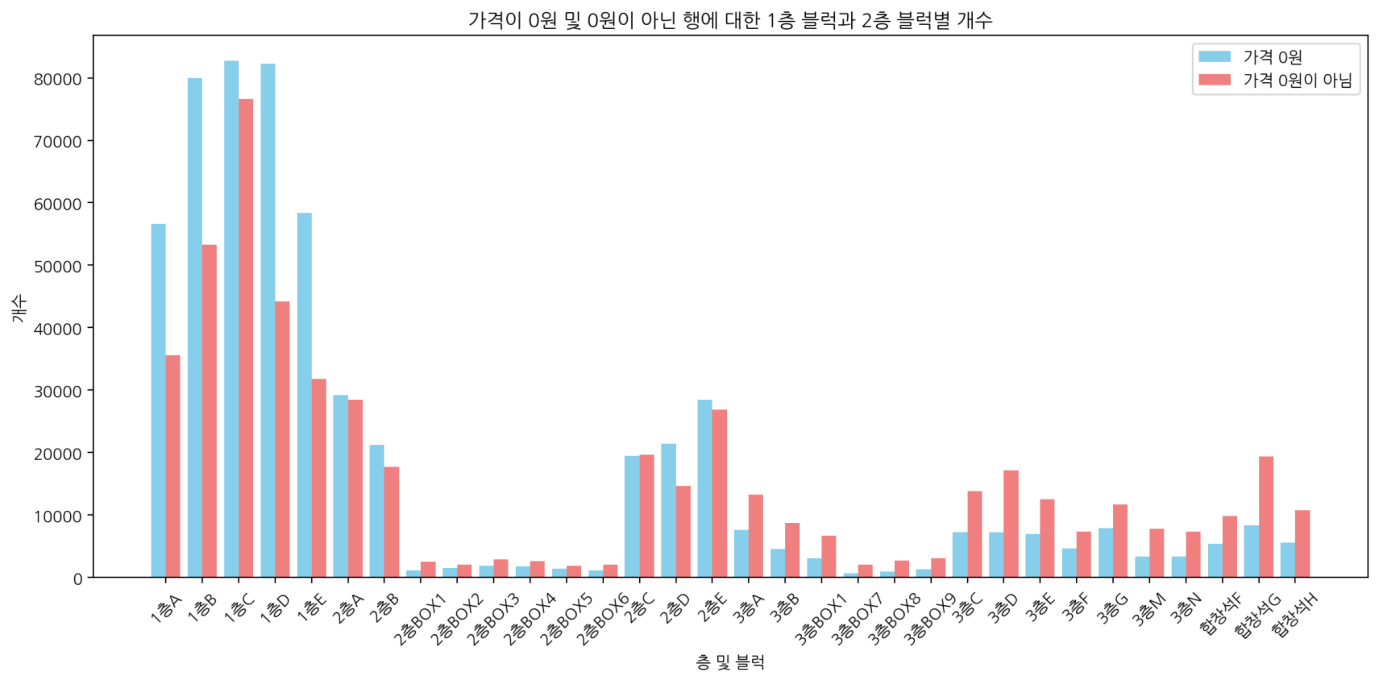
* 가격 분포를 살펴보았는데 가격이 0원인 경우가 가장 많았다. 따라서 가격이 0원일 때 할인받은 내역을 살펴보디 초대권, 기획사판매가 90%를 차지하였다.

****

* 성별과 나이에 따른 장르 선호도를 살펴보았다. 40대와 30대 여성, 50대 남성은 교향곡을 선호하고, 40대 여성은 클래식 장르도 선호하였다. 할인내역의 ‘%’앞 숫자를 추출하여 몇퍼센트 할인을 받았는지 구하였다. 할인 퍼센트와 나이그룹으로 박스 플롯을 그려보았다. 무료멤버쉽 일반등급에 해당하는 30대, 40대, 50대의 할인율이 가장 낮았다.



* 공연일자의 히스토그램을 살펴보았다. 코로나기간동안 예매수가 많이 감소한 것을 볼 수 있었고, 2023년부터는 기존의 예매량과 비슷해질 정도로 코로나의 영향이 없어졌다.
* 멤버쉽 타입에 대한 히스토그램을 살펴보았다. 나이와 관련된 무료 멤버쉽 등급인 노블, 싹틔우미, 일반 중에서 일반이 가장 많았다. 30대~60대 사이의 사람들이 예술의 전당에 많이 가입한다는 사실을 알 수 있었다. 유료 멤버쉽은 연회비가 비싼 순서대로 많았다.



seat에 저장되어있는 좌석을 블록별로 분리하여 히스토그램을 그려보았다. 1층은 제일 중앙인 C블록이 가장 많았고 가장자리로 갈 수록 적어졌다. B블록과 D블록, A블록과 E블록은 대체로 대칭이었다. 2층은 중앙인 C블록보다 가장자리로 갈 수록 값이 많아졌고 역시 대칭이다. 3층은 중앙인 D블록의 값이 가장 많았고, 대칭이었다. 돈을 주고 구매한 경우에 합창석에 앉은 경우가 많아서 합창석이 특정 공연에서는 좋은 자리일 것이라고 예측하였다.

**다. 데이터 가공**

1) 데이터 정제

* 날짜 변수 조정

날짜를 나타내는 변수인 tran\_date, play\_date와 시간을 나타내는 변수인 play\_time과 play\_st\_time을 각각 합쳐준 tran\_date\_time과 play\_date\_time 변수를 생성하고 pre\_open\_date, open\_date와 함께 datetime 형식으로 바꿔준다.

EDA에서 확인한 예매일과 공연일 차이의 최솟값은 4일이다. 당일까지 티켓을 판매하는 공연이 많으므로 차이의 최솟값은 0이 되어야 하므로 예매일을 공연일에 맞춰주었다. 그러나 여기서 play\_date\_time을 7일 앞당겨야 하기 때문에 tran\_date\_time은 원래 날보다 3일 앞당긴다.

* 전석 초대 공연 삭제

전석 초대권으로 이루어진 공연이 존재하였고, 이때의 수익은 0원이라 할 수 있다. 이는 본 분석의 목적과 맞지 않기에 전석 초대인 공연들은 제거해주었다. 크롤링 했을 때 가격에 전석 초대나 전석 무관중으로 진행되는 공연들은 전석 초대로 처리해주었고, 추가로 공연의 고객이 모두 가격이 0원이면 전석 초대로 판단하여 모두 제거하였다.

* 특별 기부 콘서트 삭제

일만원의 나눔, 오만원의 사랑 등의 할인 내역이 있는 공연이 기부 콘서트로 확인 되어 예술의 전당 수익과는 무관하다고 판단하여 제거하였다.

* 행, 열 제거

모든 열이 중복인 14개 행을 삭제하고, 콘서트홀 대관규약 제14조 공연진행석에 의하여 공연진행석을 제거해주었다. 그리고 origin\_price1의 이름을 rev\_dis로 바꿔준다.

* 러닝타임 및 장르 처리

러닝타임이 0분, 15분인 경우가 있었는데 논리적 이상치로 판단하여 주최측에 문의한 결과를 바탕으로 러닝타임이 0일 땐 러닝타임 70, 인터미션 0으로, 러닝타임이 15일 땐 러닝타임 90, 인터미션 15분으로 처리하였다.

장르가 비어있는 13개의 공연의 경우 예술의 전당 홈페이지를 참고하여 장르를 직접 채워주었다. 또한 입장연령이 매우 낮은 경우(48개월 이상, 36개원 이상)는 아동용 공연으로, 콩쿠르가 제목에 들어가는 경우는 콩쿠르로 장르를 바꿔서 채워주었다. 아동용 공연과 콩쿠르는 장르가 아무리 클래식이나 교향곡 등의 장르더라도 오는 고객층의 성격이 매우 달라 좌석 배치와 가격대도 모두 달라질 것이라 생각하여 따로 분류하였다.

* 취소 티켓 처리

같은 좌석이 여러번 예매되었는데 가장 마지막에 예매한 티켓을 제외한 티켓 중에 ticket\_cancel이 2가 아닌 0이 찍혀있는 경우 가장 마지막에 예매한 티켓을 제외하고 전부 제외하였다.

분석의 편의를 위해서 play\_date\_time, tran\_date\_time 순으로 정렬하고 각각 다른 공연을 라벨링하여 label 열로 지정하였다. 공연명이 같더라도 공연일과 공연 시간이 다르면 다른 공연으로 취급하였다.

2) 데이터 이상치 처리

* 초대권인데 price가 50000인 한 행은 초대권은 값을 지불하지 않고 본 고객이라고 생각하여 삭제하였다.
* seat\_level이 ‘missing’이지만 가격이 0원인 것들은 ‘free’로 변경해주었다.
* 공연을 진행한 시간인 running\_time이 0인 공연들을 발견하였고 빅콘테스트 주최측 문의 결과 running\_time과 intermission의 값이 잘못된 오류행인 것을 알 수 있었다. 따라서 running\_time과 intermission을 running\_time이 0일때는 70, 0으로 변경하고 running\_time이 15일때는 90, 15로 각각 변경한다.
* discount\_type이 차액인 경우는 할인 받을 수 있는 증빙 자료를 제출하지 못한 경우에 예매 후 추가로 내야하는 돈이므로 가격대가 대부분 낮다. 또한 주최측에 문의한 결과 seat는 그 고객이 예매한 자리로 들어가는 것이 아니라 공연에서 비어있는 좌석으로 결제된 것처럼 들어가며, 예매 일자가 대부분 공연일이다. 그렇기에 예매 데이터에 속하긴 하나 다른 데이터와 다른 특수한 상황이라고 생각했으며 특히 고객이 실제 앉은 자리를 알 수 없다는 점에서 분석에 왜곡을 일으킬 수 있다고 생각하여 삭제하였다.
* 회원이 아닌데 멤버쉽 타입이 있는 경우의 데이터는 현재를 기준으로 탈퇴한 회원이라 판단하여 예매 당시에는 회원이라 간주하여 member\_yn을 ‘Y’라 바꿔준다.
* 싹틔우미/노블/골드/블루/그린 회원이 아닌데 할인을 받았다면 당시 해당 멤버십이 있었다고 판단하여 age\_mb, gld\_mb, bl\_mb, grn\_mb의 값을 바꿔주었다. 또 유료 멤버십이 있는데 비회원인 경우 age\_mb를 일반으로 채웠다. 왜냐하면 싹틔우미, 일반, 노블 중 가장 비중이 높은 것은 일반이며 유료 회원은 대부분 일반 나이에 해당하는 30~50대가 많기 때문이다.

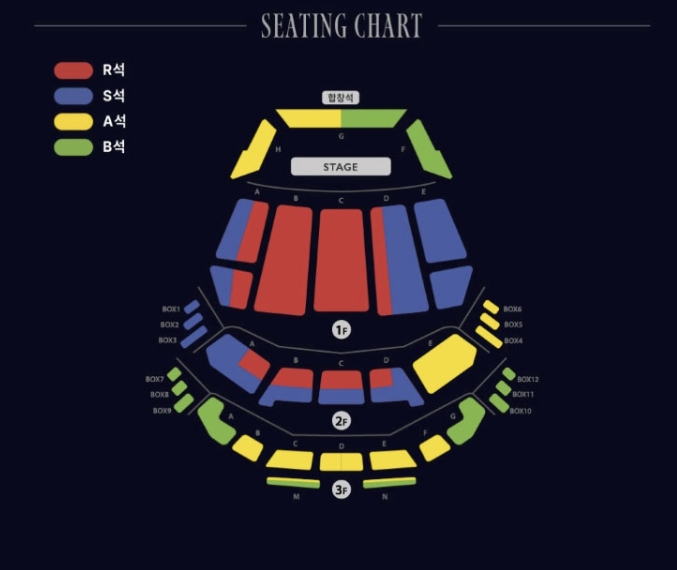
3) 데이터 결측치 처리

* 크롤링이 되지 않은 공연에 대해 결측치를 채워주었다. 같은 공연이지만 시간이 약간 차이나는 14개의 공연들을 발견하였고 확인결과 공연시간이 변경된 것이었다. 이에 실제 공연을 진행한 시간으로 통일하기 위해 변경 전 시간인 7개의 행들을 모두 변경 후 시간으로 바꿔주었다.
* 좌석 등급이 missing인 경우 좌석 등급과 복구한 정가가 맞지 않아 생기는 행으로 오류라고 판단하여 확인 후 수정해주었다. 가령 홈페이지 정보의 날짜 정보가 달라 크롤링이 안되는 경우, 정가를 알 수 있어도 seat\_level을 채우지 못하여 missing으로 처리되는 문제가 생겼다. 이러한 경우에는 직접 해당 날짜를 찾아 수정해주었다. 또한 %가 없어서 복구할 수 없었던 행은 일부 할인율을 홈페이지에서 찾아서 입력해주었다.

1. 좌석 되돌리기

seat\_level이 missing과 unknown으로 결측 처리된 자리는 다음과 같은 알고리즘으로 찾아주었다. 할인 내역에 %가 있는 경우에 정가를 찾을 수 있었다. 하지만 %없는 할인 내역과 초대권 등 price가 0원인 행은 정가를 찾을 수 없었다. 즉, seat\_level이 free, unknown, more의 경우에는 정가를 알 수 없었다. 이에 전체적인 좌석 분포 확인과 정가 복원, 전체 행 좌석 등급을 찾기 위해서 알고리즘을 고안했다. 각 공연별로 거리를 기준으로 좌석을 복구했는데, 가격을 모르는 좌석은 해당 좌석과 가장 가까우며 가격을 알 수 있는 좌석의 등급과 동일하게 복구한다. 조건은 다음과 같다.

첫째, 장르가 독주인 경우 EDA 결과 왼쪽으로 치우쳐져있었다. 그 이유는 피아노가 왼쪽에 치우쳐져있기 때문이라고 판단했다. 따라서 가격 등급도 왼쪽으로 치우쳐져 있을 것이라고 생각했다. 이를 확인하게 위해 장르가 독주인 ‘비킹구르 올라프손 피아노 리사이틀’의 좌석 배치도를 확인했고 실제 왼쪽으로 치우쳐져있는 것을 확인했다.



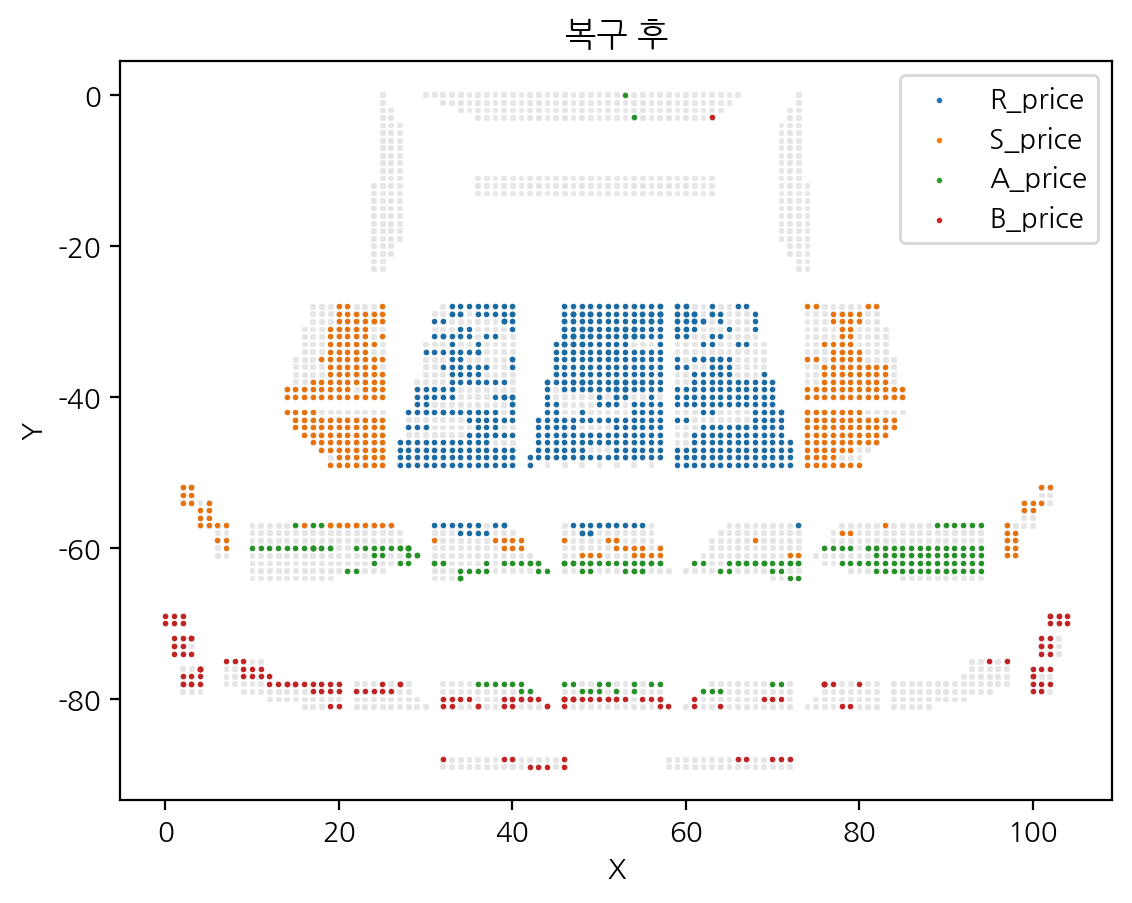
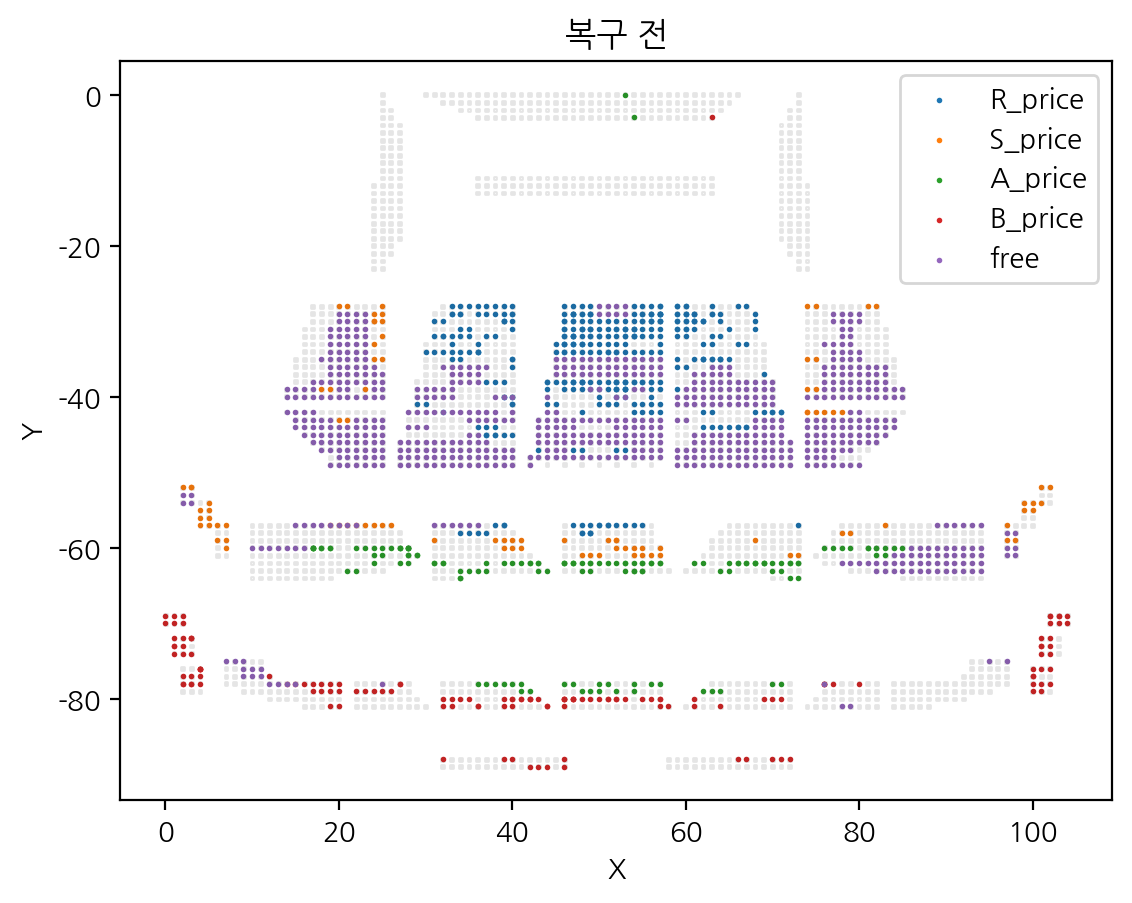
이에 독주의 경우 B블록과 C블록을 기준으로 대칭이 되게끔 가운데 x 좌표를 뺀 후 절댓값을 씌워 거리를 가깝게 하여 가까운 자리로 선택되게 했다.

둘째, 1층의 B,C,D 블록(중앙) / 1층의 A,E 블록(사이드) / 2층의 B,C,D 블록(중앙) / 2층의 A,E 블록(사이드) / 3층의 B,C,D,E,F 블록(중앙) / 3층의 A, G 블록(사이드) / 합창석의 G 블록(중앙) / 합창석의 F,H(사이드) 블록은 각각 해당되는 블록에 알수 있는 좌석이 있을 경우 다른 블록보다 우선적으로 등급이 배정되게 했다. 이때 가운데 블록들은 x값에 가중치를 두어 좌석 배치도의 열끼리 비슷한 등급이 배정되게 했고, 사이드 블록들은 y값에 가중치를 두어 좌석 배치도의 번호끼리 비슷한 등급이 배정되게 하였다. 또한 BOX석의 경우 가중치를 두지않고 같은 층에서 우선적으로 등급이 배정되게 하였다.

셋째, 앞의 경우에서 없을 경우 1층 / 2층 / 3층 / 합창석을 기준으로 등급이 배정되게 했다.

넷째, 앞의 경우에서 없을 경우 전체 거리를 기준으로 배정되게 했는데 합창석의 경우 3층과 등급이 비슷한 경우가 많았기 때문에 합창석에 없더라도 단순 거리순이 아닌 3층부터 고려하도록 했다.

그 결과 좌석의 가격 복구가 잘 됐는지 확인하기 위해 한 공연을 확인해봤다. 복구 전 좌석 등급은 free가 743개, R석이 366개, S석이 94개, A석이 146개, B석이 145개였다. 복구 이후 좌석 등급은 R석이 739개, S석이 379개, A석이 212개, B석이 164개였다. 시각화 결과는 다음과 같다.



(R석 : 파랑, S석 : 주황, A석 : 초록, B석 : 빨강, free : 보라).

1. 그 외의 전처리

* 같은 공연 중 같은 좌석에 대해 ticket\_cancel이 0인 행이 중복되어 나오는 경우를 발견하였다. 하지만 ticket\_cancel이 0인 고객은 실제 공연을 관람한 고객이라 할 수 있다. 따라서 한 공연에서 같은 좌석을 기준으로 ticket\_cancel이 0인 행이 중복되어 나온다는 것은 논리적 오류라 판단하였다. 따라서 중복인 경우 tran\_date를 기준으로 마지막에 예매한 ticket\_cancel이 0인 고객만 남기고 tran\_date\_time이 더 빠른 경우에는 제거해준다.
* performance\_code는 동일하지만 다른 일자에 진행한 공연이 존재하였다. 즉 다른 시간에 공연이 진행되었지만 공연 자체의 이름이 같아 performace\_code로는 unique한 공연을 구분 할 수 없는 문제를 발견하였다. 따라서 play\_date\_time을 기준으로 unique한 공연을 구분할 수 있는 ‘label’ 변수를 생성하였다.
* 본 분석 데이터의 한 행은 고객이 어떤 공연의 좌석을 예매를 했는지, 또한 티켓을 취소했는지를 알 수 있는 데이터이다. 따라서 예매를 하였지만 그 자리를 취소하여 ticket\_cancel이 2가 나오는 행들이 존재한다. 이러한 취소표를 다른 고객이 다시 예매하여 관람을 할 수 있다. 하지만 df에서 같은공연을 기준으로 같은 좌석에 대해 tran\_date\_time으로 정렬한 후 살펴봤을 때, 마지막으로 예매한 사람의 ticket\_cancel이 2인 경우에는 그 전 예매내역에서는 ticket\_cancel이 0이 나올 수 없다. 하지만 df에서 그러한 경우를 발견하였고 오류행이라 판단하여 제거를 해주었다. 이때 제거를 하기 위해 공연의 특정 좌석에서 실제로 관람까지 이루어졌는지를 알 수 있는 cancel\_yn 변수(관람 시 : 0, 비관람 시 : 1)를 생성하였다. 최종적으로 ticket\_cancel이 0이고 cancel\_yn이 1인 행을 제거하였다.
* age는 결측치인데 다른 멤버십을 가지고 있는 행들은 age\_mb를 비회원이 아닌 일반으로 대체했다.

**라. 파생변수 생성**

공연/고객/좌석의 특성을 살리기 위해 파생변수를 생성해준다. 이 때 생성한 파생변수들은 좌석 그룹핑을 위한 클러스터링과 가격 책정을 위한 알고리즘에서 쓰일 것이다. 공연, 고객, 좌석별로의 특징 추출이 가능하며 유의미한 해석이 가능할 것이라 기대한다.

1. 공연 관련 파생변수

* 공연별 취소 건수를 의미하는 cancel\_num 변수를 생성하였다. 공연별로 ticket\_cancel이 2인 행의 수의 합으로 만들어주었다.
* 공연의 play\_date\_time을 사용하여 공연날짜의 요일을 라벨인코딩 해주었다. 이때 월요일을 시작으로 일요일까지 0~6의 값을 각각 바꿔주었다.
* 공연의 관람 건수를 나타내는 tran\_rate 변수를 생성하였다. ticket\_cancel이 0인 행의 수의 합으로 구해주었다. 공연의 인기도를 알 수 있다.
* 사회적 거리두기 단계(2020-08-16 ~ 2021-10-03)를 조사하여 기간별로 단계에 해당하는 숫자를 할당한 covid 변수를 생성하였다.
* 공연별 예매 건수를 의미하는 ticket\_count 변수를 생성하였다. 실제로 관람을 한 고객과 결국 취소한 고객 수를 모두 합한 전체 예매 내역 수를 계산해 구해주었다.
* 각 공연별 존재하는 좌석등급 중 제일 낮은 좌석등급의 가격을 의미하는 min\_price를 생성하였다.
* 기존에 R\_price, S\_price,... , all 로 나눠진 좌석 등급 가격의 이름은 제거하고 가격이 높은 순으로 각각 price1, price2, price3, price4, price5, price6으로 이름을 명명하여 좌석 가격을 정리한다.
* 공연의 가격 등급이 몇개인지를 나타내는 열인 count\_price를 생성한다.

1. 고객관련 파생변수

* membership\_type\_1~6 대신 나이 멤버십(age\_mb)과 유료 멤버십(gld\_mb, bl\_mb, grn\_mb) 열을 만든다. 회원이면 age\_mb에 나이 멤버십인 싹틔우미, 일반, 노블 중 하나를 넣고 비회원이면 비회원으로 채운다. gld\_mb, bl\_mb, grn\_mb엔 1,0으로 유료 멤버십 여부를 채운다.
* 고객의 충성도를 알 수 있는 loyalty 변수를 생성해주었다. 그 후 유료회원은 가중치를 연간 회원비(골드 10, 블루 4, 그린 2)로 하여 더해주고 무료 회원은 1, 비회원은 0으로 하여 회원이 가지고 있는 멤버쉽에 대한 가중치 총합을 만들어준다. 가령 골드, 블루 회원을 가지고 있는 고객은 10\*1+4\*1+1=15가 되는데 그 이유는 기본적으로 무료 회원 멤버십은 기본적으로 회원이라면 가지고 있기 때문에 뒤에 1을 더해주는 것이다. 이 1을 더해주는 작업을 하는 이유는 회원과 비회원의 차이를 두기 위해서이다.
* 공연에서 예매된 좌석 수를 의미하는 ticket\_count를 생성했다. 취소표를 제외한 실제로 그 공연을 가서 본 사람의 수를 의미한다.
* 예매속도를 알 수 있는 like 변수를 만들어준다. 공연을 얼마나 빨리 예매했는지를 따져 고객이 공연에 대해 가지고 있는 관심도를 알 수 있다. like 계산은 다음과 같다. 해당 공연에 선예매 날짜가 존재할 경우엔 (play\_date\_time - tran\_date\_time)/(play\_date\_time - pre\_open\_date), 해당 공연에 선에매 날짜가 존재하지 않을 경우엔 (play\_date\_time - tran\_date\_time)/(play\_date\_time - open\_date)로 계산한다. 하지만 데이터 상으로 1) 예매 오픈일 보다 선예매 오픈일이 더 느린 경우, 2) 선예매가 존재하는 공연 중 선예매 오픈일보다 예매를 더 빨리 한 경우, 3) 선예매가 존재하지 않는 공연 중 예매 오픈일보다 예매를 더 빨리 한 경우가 존재하였다. 이 세가지 경우는 오류 행이라 판단하였고 오류 행 처리는 다음과 같다.

먼저 공연에 오류행이 많으면 가장 빠른 예매일을 pre\_open\_date 혹은 open\_date에 할당한다. 만약 공연에 오류행의 비율이 작다면 그 행 자체를 오류 또는 테스트 데이터로 판단하고 그 행만 삭제해주었다. 이 때 삭제를 해주는 비율의 기준은 전체 티켓 수 5%보다 오류행이 적고, 오류행이 40개 미만일때로 정의한다.

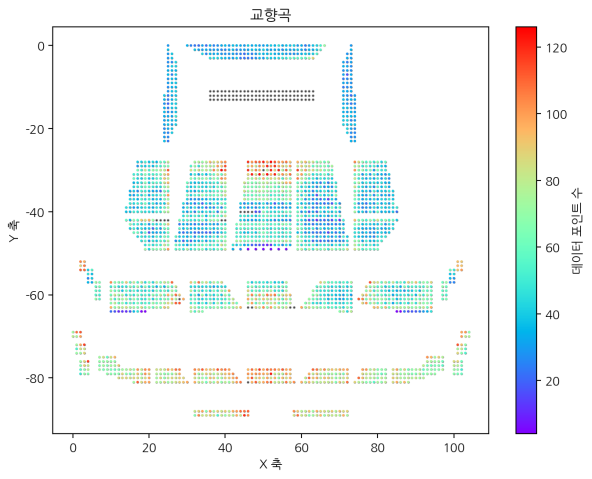
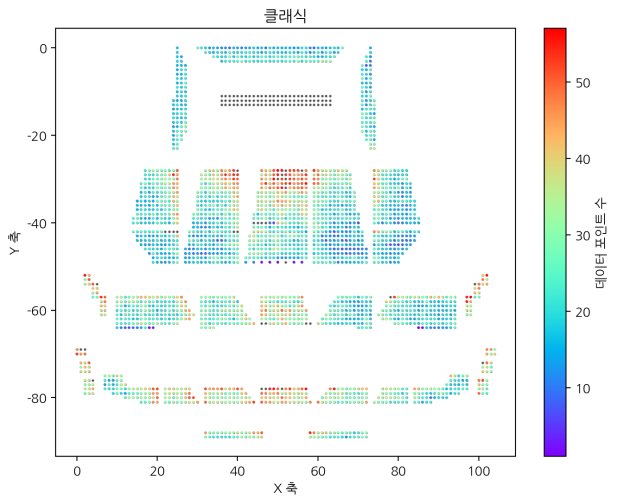
* like 변수를 이용하여 like 공연의 percentile으로 계산 like\_ratio를 생성하였다. 한 공연 안에서 고객들은 각기 다른 like를 갖는데 그 안에서의 percentile으로 계산해주었다. 고객이 예매를 일찍한 경우에 like가 높지만, 공연이 인기있는 공연이라 다른 고객들이 더 일찍 예매했을 때 like\_ratio는 높지 않을 수 있다. 따라서 각 공연 안에서 해당 고객이 다른 고객에 비해 얼마나 관심이 있는지를 나타내는 변수이다.
* 고객이 앉은 좌석의 원래 가격을 수 있는 rev\_seat 변수를 생성하였다. 이전에 가격을 토대로 구할 수 있는 seat\_level을 구한 뒤 구할 수 없는 것은 seat를 기준으로 가까운 자리의 seat\_level로 좌석 등급을 바꿔주었다. 이렇게 알게 된 고객들의 좌석을 기준으로 원래 가격을 채워준다. 이때 좌석이 잘못 매칭된 것이 있어 보정을 해주었다. 실제 좌석의 가격보다 더 높은 값을 내고 볼 리는 없기 때문에 이러한 경우에 대해서는고객이 낸 price보다 비싼 좌석등급 중 가장 낮은 등급의 좌석으로 수정한다.
* 공연의 최소금액으로부터 낸 금액인, 즉 최소 가격으로부터 얼마나 여유 금액이 있는지를 나타내는 afford 열을 생성하였다. 이 때 할인을 %로 받은 행만 처리한다.
* 고객의 할인 타입을 재범주화해준 divide\_discount를 생성한다. 범주는 0, 정가, 회원할인, 외부할인으로 나누어진다.

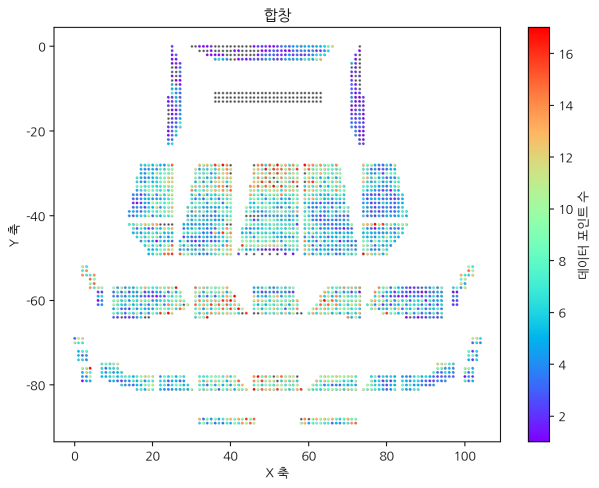
1. 좌석관련 파생변수

* 마지막에 결국 그 좌석이 취소가 되었는지를 나타내는 cancel\_yn 열을 만들어준다. 해당 날짜와 좌석에 앉은 마지막 사람이 그 자리를 취소했다면 1, 취소하지 않았다면 0으로 채운다.
* 공연에 매겨진 좌석 등급의 가격 중 가장 싼 가격인 min\_price 열을 생성한다.
* 한 공연 내에서 특정 좌석에 대한 예매내역수를 알 수 있는 seat\_count 변수를 생성하였다. 특정 공연 내에서의 좌석의 선호도를 알 수 있다.
* 좌석과 무대 사이의 물리적인 거리를 나타내는 distance를 생성하였다. 좌표 데이터 x, y를 사용하여 무대 중심과의 유클라디안 거리를 계산하여 나타내준다. distance가 크면 클수록 실제 공연장에서의 거리가 멀다고 해석할 수 있다.
* 좌석과 무대 사이의 각도를 알 수 있는 angle\_degrees를 생성하였다. 라디안 단위로 각도 계산 후 0~90도 사이로 변환해준다.
* 공연별 좌석별 예매 수인 seat\_count 열을 생성한다. 취소표를 포함하여 그 공연의 좌석별로 얼마나 티켓팅이 있었는지를 나타낸다.
* 해당 좌석의 인기도를 알 수 있는 seat\_hot 변수를 생성하였다. 총 예매 티켓 수 중 그 좌석이 예매된 비중을 min-max 정규화 방법으로 계산하였다. 따라서 0과 1사이의 값을 가진다. 전체적으로 어떤 좌석이 인기가 많은지를 알 수 있다.
* 각 공연의 좌석정가를 큰 순서부터 저장한 price1,price2,...,price6을 만들었다. 공연마다 좌석등급의 분류와 이름이 다르기 때문에(R석,S석 또는 전석 또는 1층석 등) 분석의 편의성을 위하여 등급이 아닌 price1~price6으로 통일해주었다. 만약 기존 좌석 등급이 6종류 미만이면 좌석등급 수만큼은 값이 있고 나머지는 nan이다. 그 후 기존 가격 변수들은 제거해주었다.
* 해당 좌석이 공연 내에서 몇번째로 비싼 좌석인지를 나타내는 변수 rank를 생성해주었다.

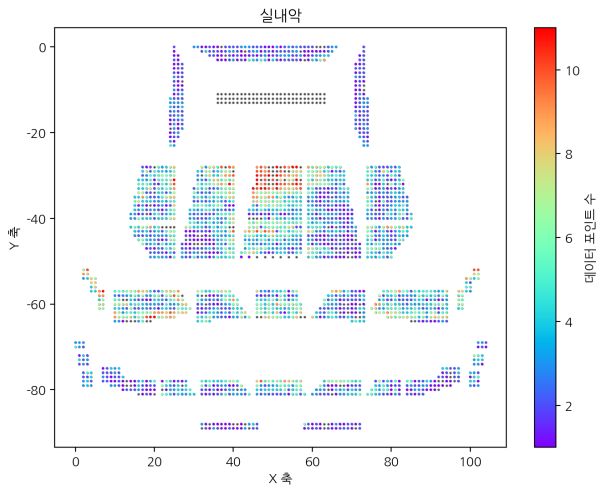
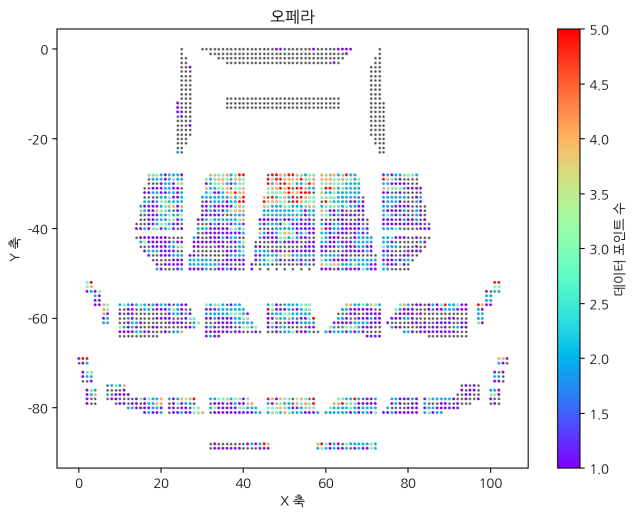
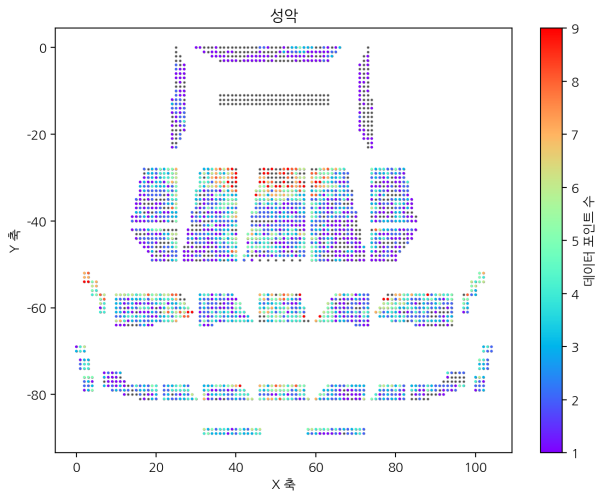
**마. 2차 EDA**

1) 공연별 좌석 시각화

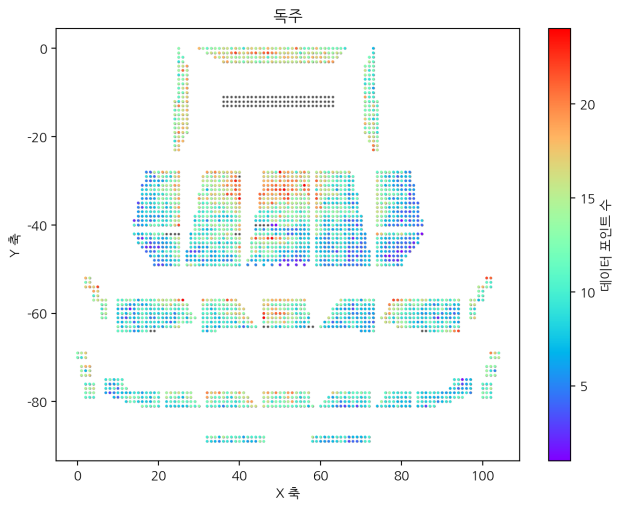




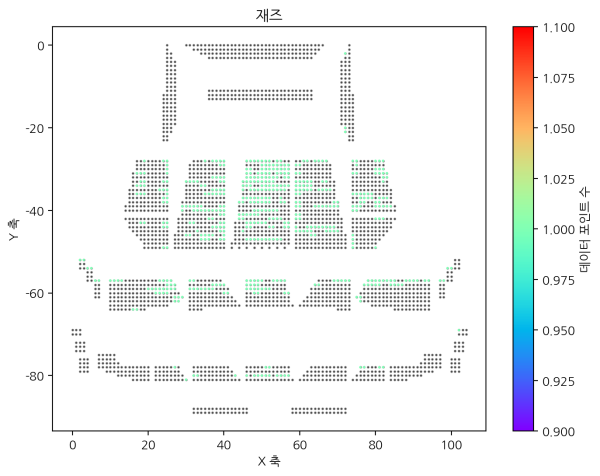
-장르가 클래식, 교향곡, 합창인 경우는 가운데에 몰리는 경향이 있었지만 3층까지도 꽤나 많이 분포된 것으로 보아 규모가 크고 인기가 많은 공연들이라고 판단했다.

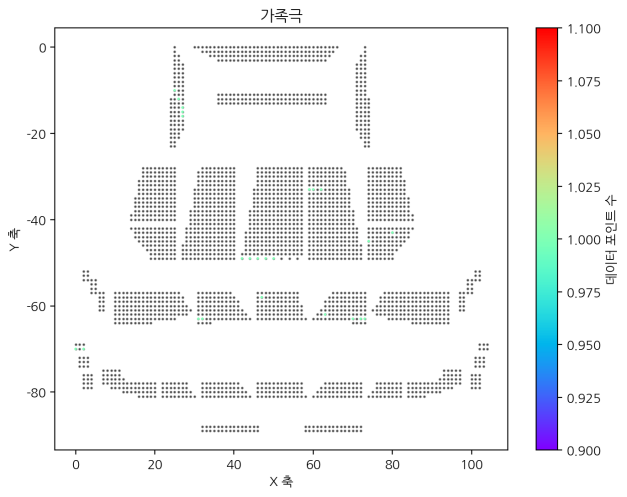
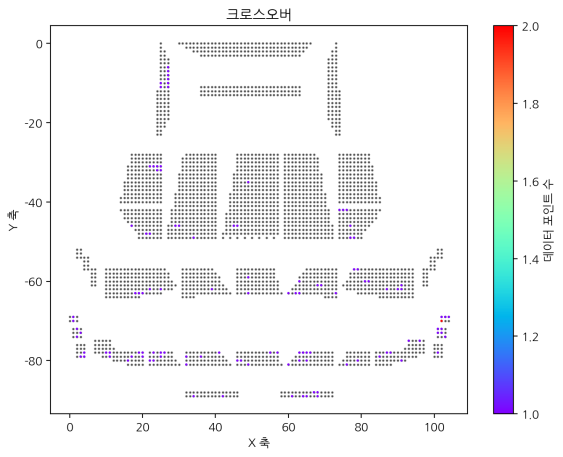


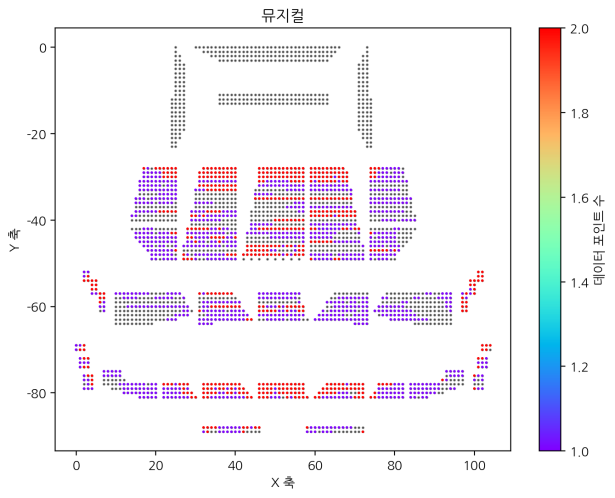
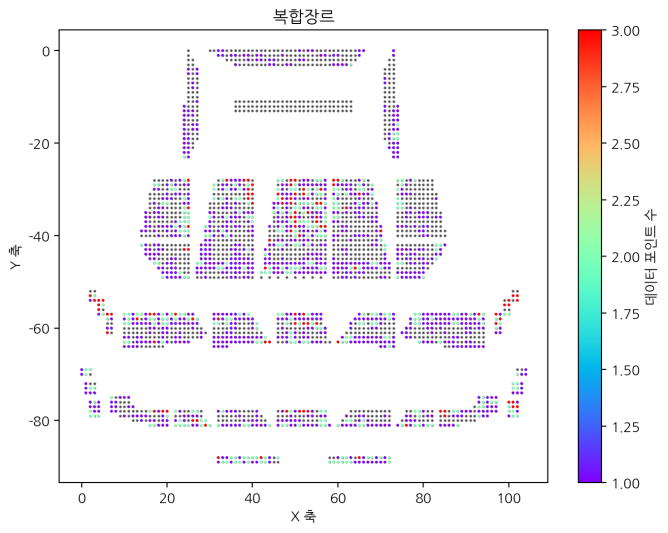
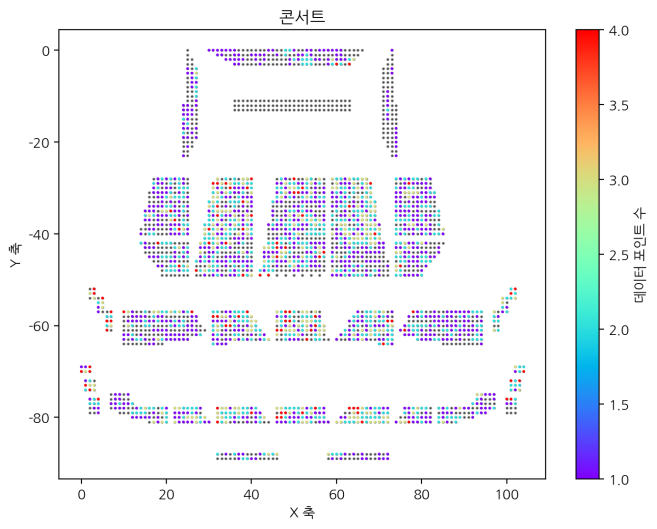
장르가 성악, 오페라, 실내악인 경우 앞쪽에 몰려있는 분포를 띄어서 규모가 작은 공연이라고 판단했다.



좌석 복구에서 언급 했듯이 독주의 경우 왼쪽으로 치우쳐져있는 분포를 보였다. 이는 피아노가 좌측에서 공연하기 때문이라고 확인됐다.







표본이 적은 경우 분포가 뚜렷하게 보이지 않았고 이에 기타 그룹으로 판단했다.

**라. 좌석 그룹핑을 위한 클러스터링**

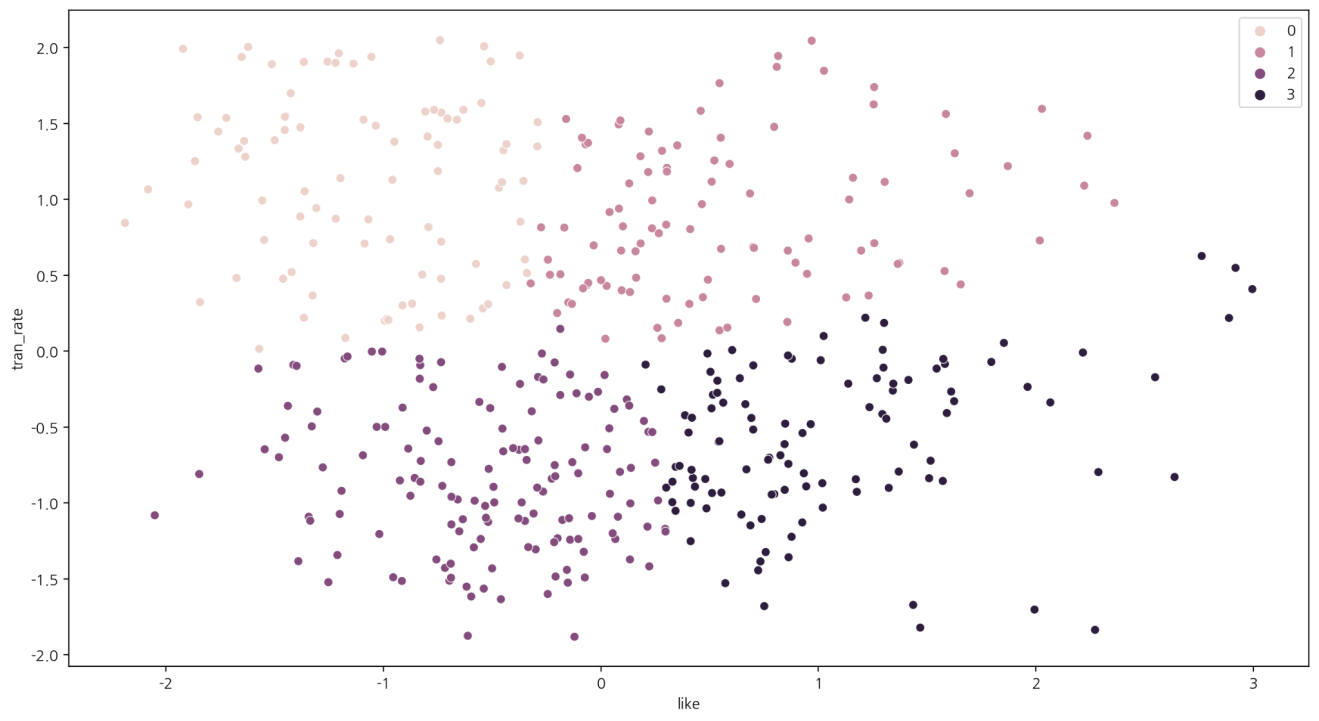
현재까지 분석을 진행한 데이터를 df라 지칭하겠다. 모든 전처리와 eda 이후 원활한 좌석 그룹핑을 위해 클러스터링을 진행한다. 1차적으로 df에서 공연에 대한 변수만을 추출해 각 공연별로 groupby를 시킨 공연 데이터프레임으로 공연에 대한 클러스터링을 진행한다. 이때 공연 클러스터링을 진행하기 위해 사회적 거리두기가 이루어졌던 기간을 기준으로 하여 코로나 데이터인 cov\_data와 코로나가 아닌 not\_cov\_data로 데이터 프레임을 분리한다. 그 후 두 데이터프레임에 대해 각각 공연 클러스터링을 수행해준다. 그 후 다시 df로 돌아와 고객에 대한 클러스터링을 진행해준다.

**1) 공연 클러스터링**

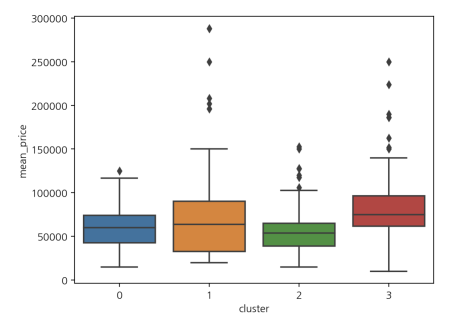
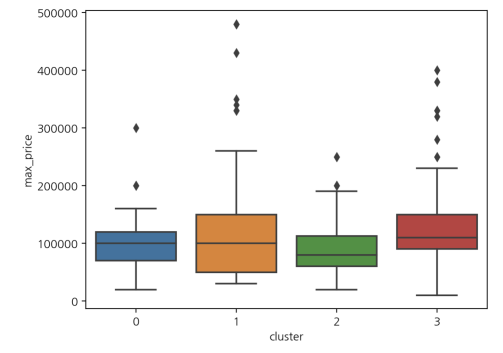
각 공연에 대해 동일한 값을 가지는 변수인 ‘play\_date\_time', 'running\_time', 'intermission', 'ticket\_count', 'cancel\_num', 'tran\_rate', 'days', 'price1', 'price2', 'price3', 'price4', 'price5', 'price6', 'like', ‘covid’, ‘genre’ 변수만 추출하여 play\_date\_time을 기준으로 그룹화를 진행한 공연 데이터프레임을 생성해준다. 이때 like 변수는 공연 변수는 아니지만 공연별 평균값을 구했을 때 공연의 평균적인 예매속도, 즉 평균적인 관심도를 알 수 있다. 따라서 like 변수도 포함하여 공연 데이터프레임을 만들어준다. 다음으로 공연의 특성을 고려할 때, 예술의 전당에서 코로나19 유행 시기 사회거리두기 정책에 따른 좌석 띄워앉기가 이뤄졌던 기간이 존재한다. 좌석 띄워앉기로 인해 애초에 제공되는 좌석의 수에서 차이가 나기에 공연의 특성을 파악할 수 있는 클러스터링에도 영향을 줄 것이라 판단하였다. 따라서 사회적 거리두기가 시행된 기간을 기준으로 분리하여 클러스터링을 따로 진행하여준다. 데이터프레임 분리 후 범주형인 days와 genre 변수를 제외해준 뒤, 나머지 변수들에 대해 pca를 진행한다. 이 때, PCA란 차원 축소 기법으로 고차원 데이터를 저차원의 데이터로 변환함으로써 많은 독립변수들이 존재할 때 생길 수 있는 차원의 저주 문제를 해결해준다. 클러스터링에 필요한 변수를 만들어 내기 위해 공연 데이터프레임에 대해 PCA 기법을 써보았다. 차원축소는 성공적으로 되었으나 주성분끼리의 해석이 명확하지 않아 분석 진행에 어려움을 겪었다. 따라서 PCA 방법을 기각한다.

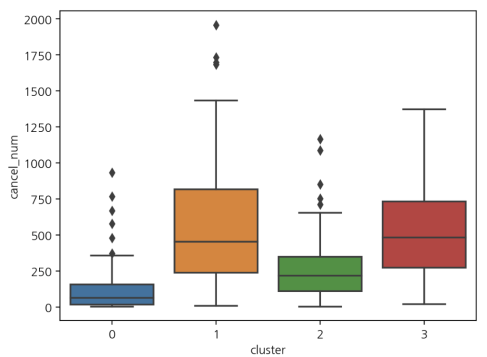
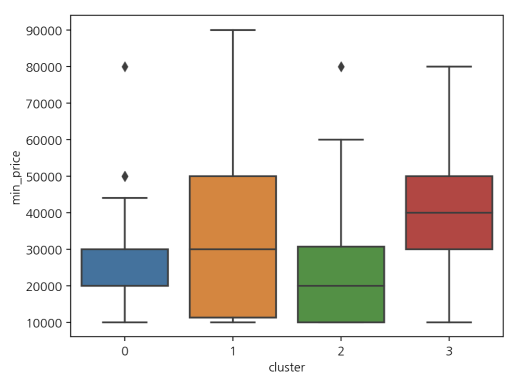
* 코로나 시기가 아닐 때, 공연 클러스터링

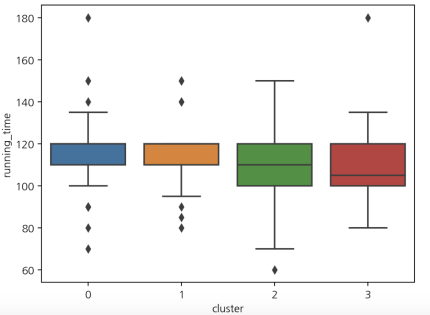
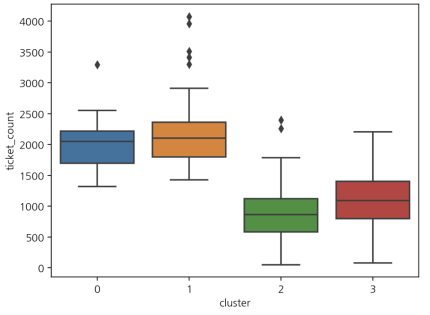
공연별 평균 관심도를 나타내는 like 변수와 예매율을 의미하는 tran\_rate를 사용하여 클러스터링을 진행한다. 클러스터링 방법은 k-means 방법을 사용한다. 이 때 k-means 클러스터링이란 k개의 중심점을 정하여 각 데이터들을 가장 가까운 중심점과 같은 그룹으로 할당한다. 그리고 그룹의 중심점을 다시 업데이트시켜 개별 데이터의 그룹을 할당하는 과정을 반복하는 클러스터링 방법이다. 최적의 군집 수를 정하기 위해 Elbow Method를 사용하였다. Elbow Method란 군집 수에 따른 SSE의 변동을 비교해 최적 군집 수를 정하는 방법이다. k가 4일 때 그래프의 선이 완만하게 꺾여 최적의 군집수라 판단하고 클러스터링을 진행하였다. 클러스터링 결과는 다음과 같다.



시각화를 해보았을 때, 해석의 애매함 없이 군집이 깔끔하게 잘 나온 것을 확인할 수 있다. silhouette\_score 계산 결과 0.52로 클러스터링의 결과가 잘 나왔다고 할 수 있다. 군집에 따른 박스 플랏을 그려보았을 때 다음과 같다.







k-means 군집화 결과와 군집마다의 박스플랏까지 확인 후 찾은 군집들의 특성은 다음과 같다.

1) clust가 0인 경우

관심도(예매속도)가 낮고, 예매율이 높은 공연이다. 박스플랏을 확인해보았을 때, 다른 군집들에 비해 가격대가 낮으며 취소수가 별로 없고 예매 내역 자체의 수는 많아 실제로 본 사람이 많다고 해석된다. 따라서 clust가 0인 공연은 즉 좋은 퀄리티의 공연이 아니지만 가격이 낮다는 이점 때문에 사람들이 결국 많이 보러 오는 공연이라 할 수 있다.

2) clust가 1인 경우

관심도(예매속도)가 높고 예매율도 높은 공연이다. 박스플랏을 확인해보았을 때, 다른 군집들에 비해 가격대가 높으며 취소수와 예매내역 수가 많은, 공연 구매활동이 활발한 공연이라 할 수 있다. 따라서 clust가 1인 공연은 즉, 인기와 수요가 높아 가격대가 높게 형성된 공연이라고 할 수 있다.

3) clust가 2인 경우

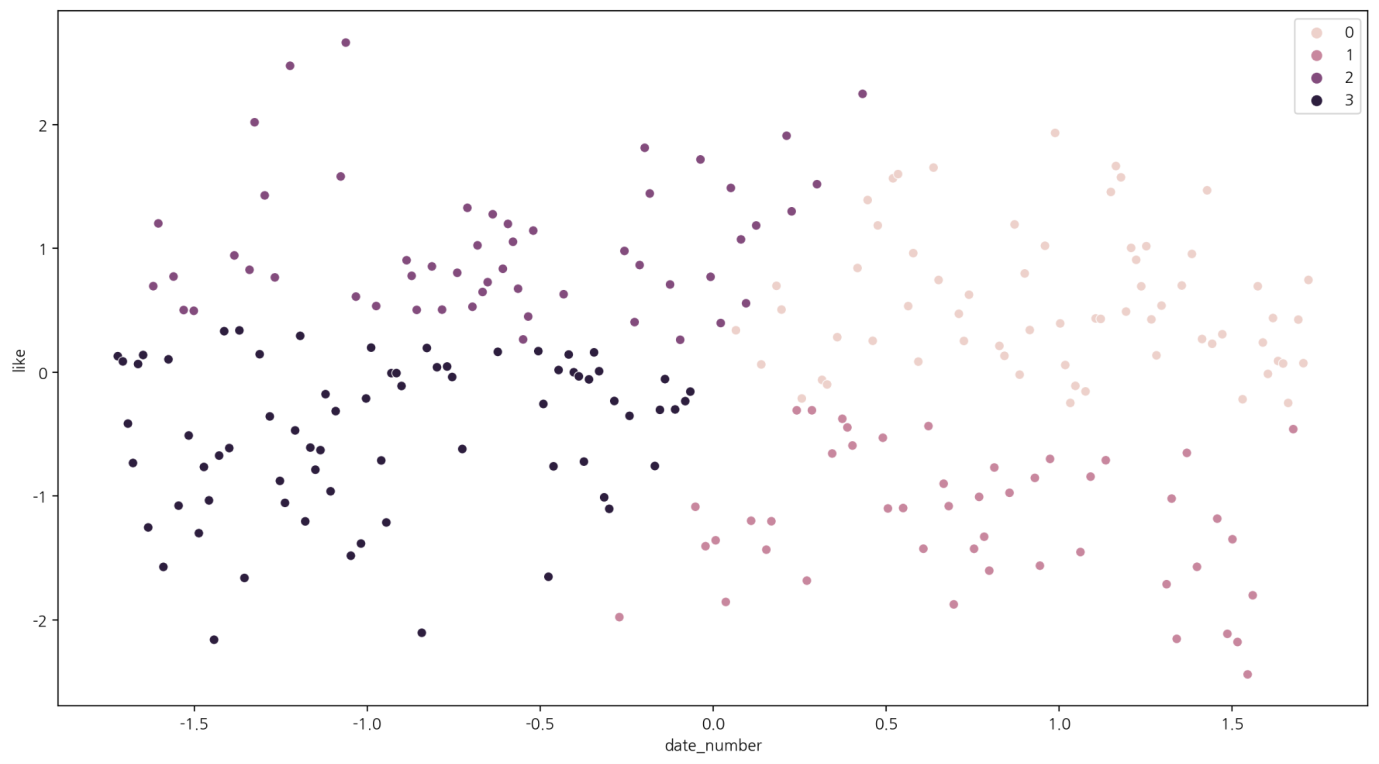
관심도(예매속도)가 낮고, 예매율도 낮은 공연이다. 박스플랏을 확인해보았을 때, 다른 군집들에 비해 가격대가 낮으며, 취소수와 예매내역의 수도 낮은, 공연 구매활동이 저조한 공연이라 할 수 있다. 따라서 clust가 2인 공연은 즉, 인기와 수요가 낮아 다른 공연들에 비해 가격도 저렴하게 책정된 공연이라 할 수 있다.

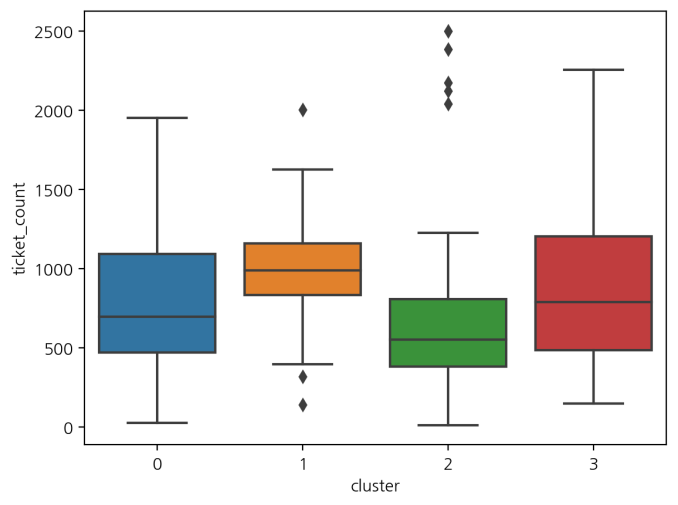
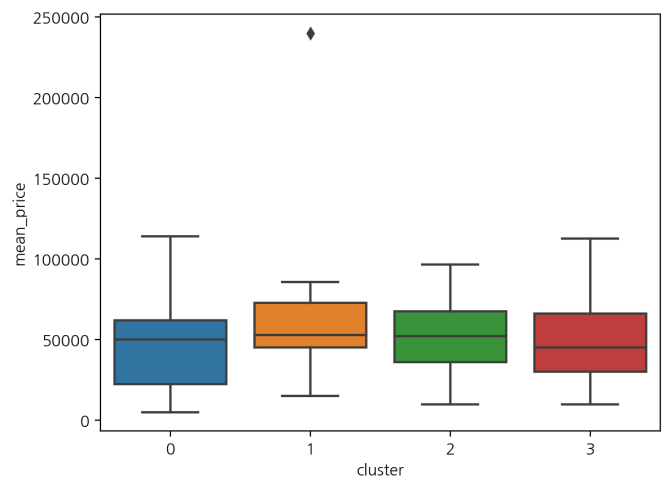
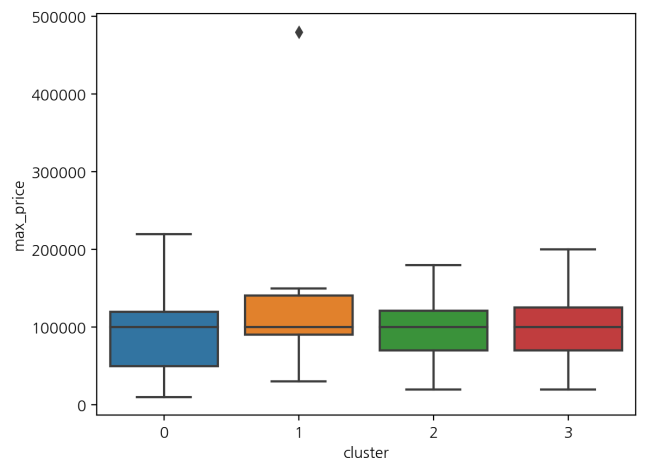
4) clust가 3인 경우

관심도(예매속도)가 높고, 예매율이 낮은 공연이다. 박스플랏을 확인해보았을 때, 다른 군집들에 비해 가격대가 높고 취소수가 높지만 예매내역의 수가 낮은, 실제로 공연을 보러오는 사람이 별로 없는 공연이라 할 수 있다. 따라서 clust가 3인 공연은 즉, 특정 수요층에게만 인기가 있는 (예를 들면 경제적 여유가 남들보다 큰 사람들이 주로 보는 공연)공연이라 할 수 있다.

5) 코로나 시기 공연 클러스터링

코로나 시기의 공연은 특수한 상황이기 때문에 가격 알고리즘에 부적절하다고 판단하였다. 하지만 코로나 기간동안에 공연을 보러 온 고객과, 그 공연의 특성을 살펴보기 위하여 클러스터링을 진행하였다. 코로나 시기의 공연 데이터프레임에 대해 date\_number와 like 변수로 k-means 클러스터링을 수행하였다. date\_number는 코로나 시기에 한 공연들을 날짜 순으로 라벨인코딩 한 변수이고, like는 공연별 평균 관심도를 나타내는 변수이다. 코로나는 확진자 수, 사회적 거리두기 단계 등으로 인해 날짜의 영향을 많이 받기 때문에 날짜와 관련된 변수를 사용하였다. 코로나시기가 아닌 공연의 클러스터링과 동일한 방법을 사용하였고 군집수는 4로 결정하였다 클러스터링 결과는 다음과 같다.





k-means 군집화 결과와 군집마다의 박스플랏까지 확인 후 찾은 군집들의 특성은 다음과 같다.

1) clust가 0인 경우

코로나 말기에 관심도(예매속도)가 높은 공연이다. 예매율이 높은 공연이다. 박스플랏을 확인해보았을 때, 다른 군집들에 비해 가격대가 낮으며 취소수가 높고, 예매 내역 자페의 수는 편차가 크다. 공연의 퀄리티가 좋고 가격이 합리적이라 사람들에게 인기를 얻었으나 코로나때문에 취소를 망설이는 고객들이 많았던 경우라고 정의할 수 있다.

2) clust가 1인 경우

코로나 말기에 관심도(예매속도)가 낮은 공연이다. 박스플랏을 확인해보았을 때, 다른 군집들에 비해 가격이 다소 높고, 취소수가 적고, 예매내역 자체의 수의 편차가 적다. 가격이 다소 높지만 그 돈을 지불하고, 코로나인 것을 감안하고도 공연을 보고싶어하는 사람들이 주로 관람하는 매니아 공연이라고 정의할 수 있다.

3) clust가 2인 경우

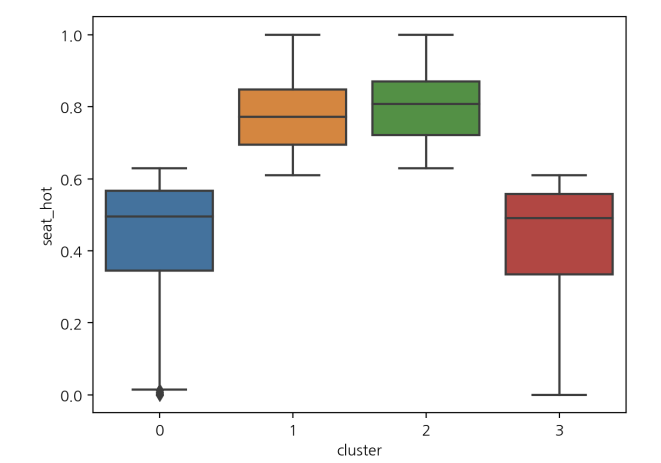
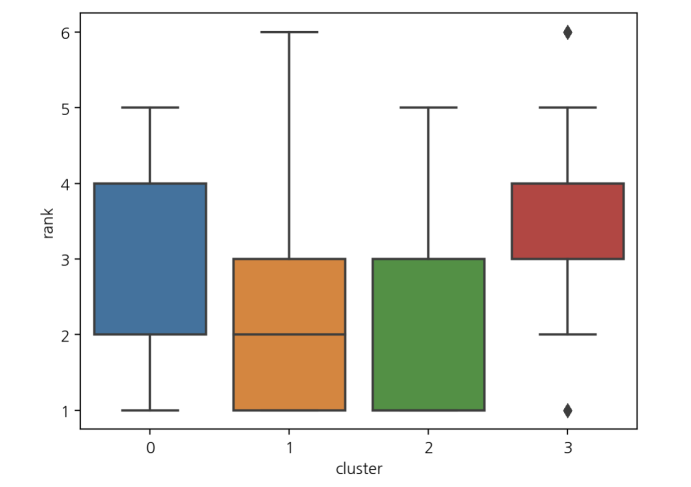
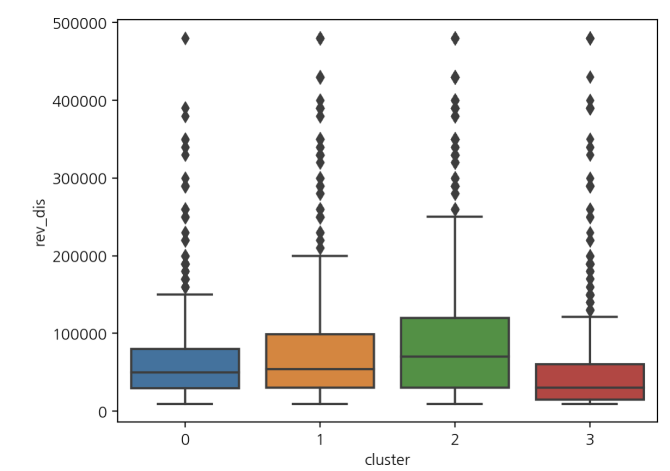
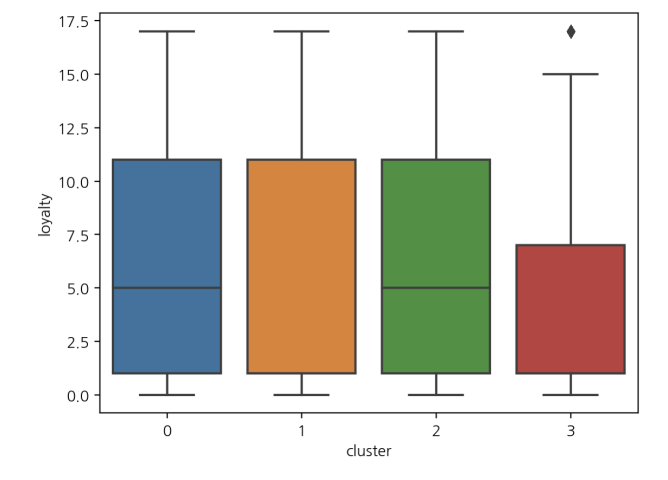
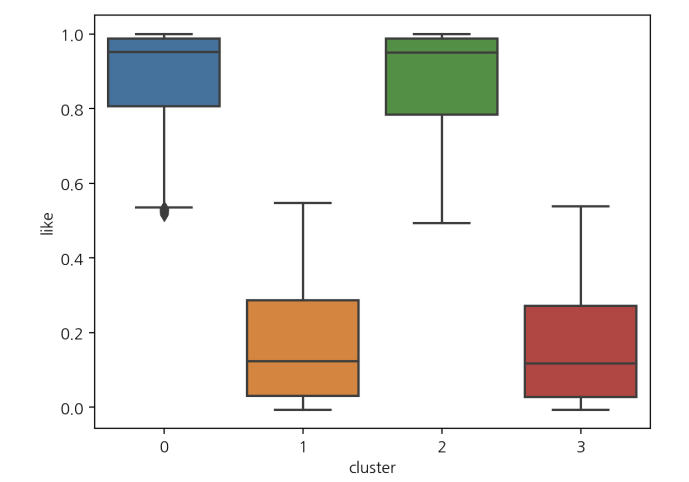
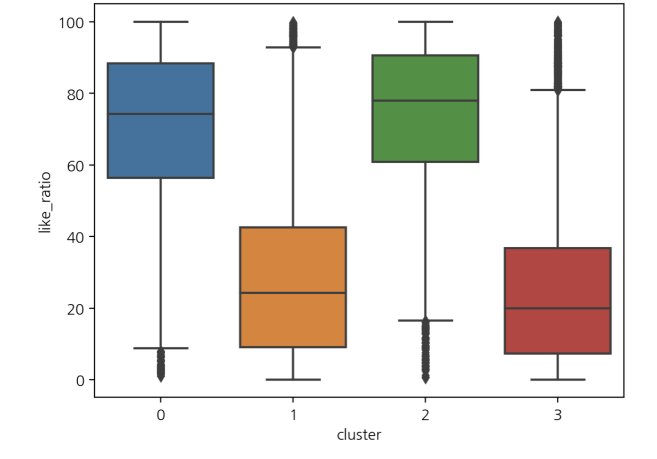
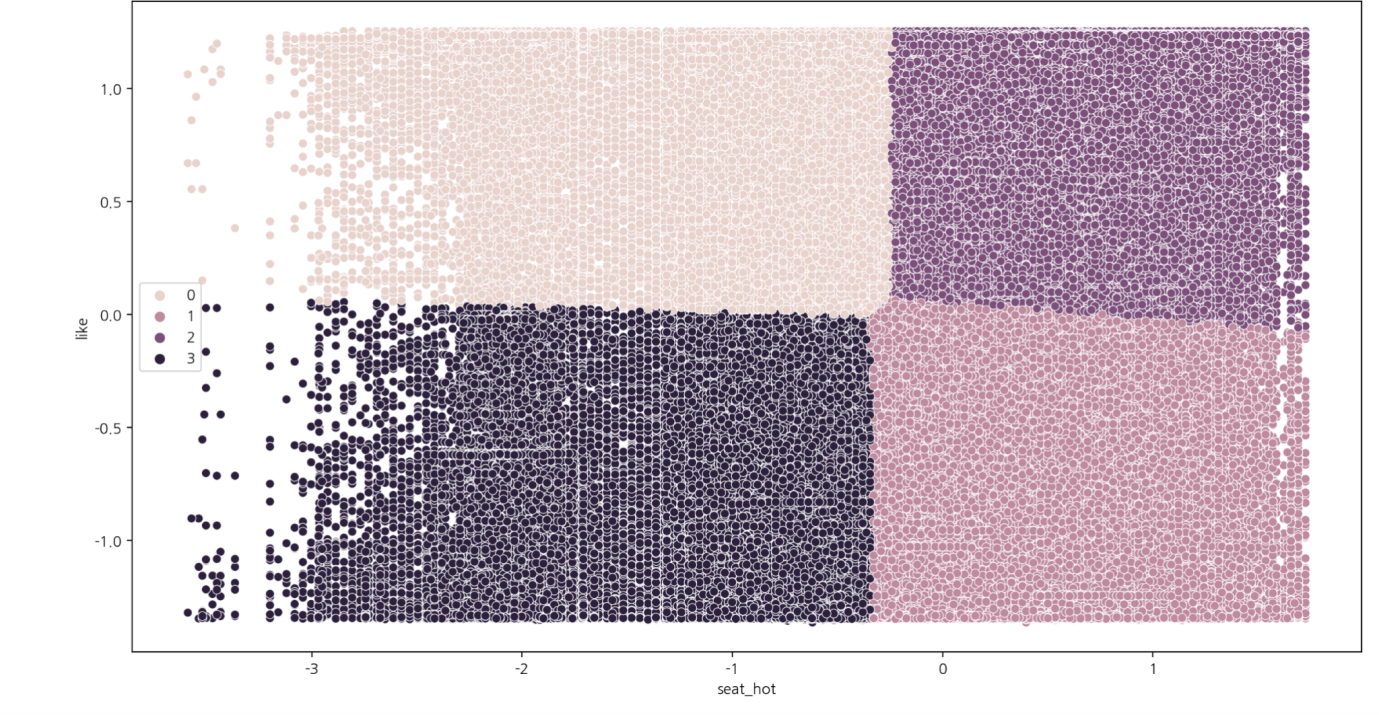
코로나 초기에 관심도(예매속도)가 높은 공연이다. 박스플랏을 확인해보았을 때, 가격과 취소수, 예매내역 자체의 수에서 별다른 특징을 찾을 수 없다. 코로나 초기의 경계단계를 감안하고도 인기가 많았던 공연이라고 정의할 수 있다.

4) clust가 3인 경우

코로나 초기에 관심도(예매속도)가 낮은 공연이다, 박스플랏을 확인해보았을 때, 다른 군집들에 비해 취소수가 낮고 예매 내역 수는 많으면서 편차가 큰 공연이다. 코로나인 것을 감안하고도 공연을 보고싶어하는 사람들이 주로 관람하는 매니아 공연이라고 정의할 수 있다.

**바. 고객 클러스터링**

좌석의 인기도(seat\_hot)와 관심도(like)로 클러스터링을 진행하였고 공연그룹 1,3번(관심도가 높은 그룹)에 고객그룹 2번(관심도가 높아서 좌석의 인기가 높은 곳에 앉은 그룹)이 많은것을보아 공연그룹의 가설을 더욱 신빙성 있게 보충해주었다.

****

**사. 좌석그룹핑 & 가격책정**

1. 알고리즘

공연과 고객의 클러스터링을 진행하여 나온 각 그룹별 특성을 적용하여 알고리즘을 고안했다. 기본틀은 다음과 같다. 각 그룹간 공연의 특성이 유사하게 분류되었으므로 그 공연중 한 사람이 다른 공연에 가게 된다면 그 공연도 볼 것 이다. 이때 afford를 이용해서 가격을 상향하거나 하향할 것이다. 이때 좌석등급이 다른 공연은 유사하지 않다고 판단하여 좌석 등급의 개수에 따라 알고리즘이 달라지게 구성하였다. 좌석 등급은 최소 1개에서 최대 6개까지 있는데 공연개수가 5개 이하인 것은 그룹의 특징이 공연의 특징과 다를 수 있기 때문에 분석하지 않도록 한다. 이때 만약 좌석 등급이 3개라면 그룹 전체의 price1-price3, price2-price3, price3-price3(=0) 평균을 계산한다. 이후 그룹의 평균과 개인의 afford를 비교한다. 이때 인기가 없어 금액을 높여서는 안되는 0번, 2번 그룹은 afford에 n\*표준편차를 더해 금액이 너무 높아지지 않게 조정한다. 최고 좌석을 1.2% 수준으로 맞췄고 1번, 3번 그룹은 금액을 비교적 높여도 되기때문에 앞의 그룹들보다 낮게 가중치를 설정했다. 또한

|  |  |  |
| --- | --- | --- |
| **Ⅲ.** |  | **주요 결과 및 시사점** |

**1. 주요 결과 요약**

1) 데이터 수집 및 파생변수 생성

고객이 어떤 좌석에 앉았는지 정확하게 구하는 것을 목표로 하여 크롤링을 실시했고, EDA를 통해 고객/공연/좌석의 다양한 파생변수를 생성하였다.

2) 고객/공연 클러스터링

주최 측에서 미리 알 수 있는 정보와 관련된 변수는 고객과 공연을 클러스터링하여 각각에 해당하는 경우에 대해 가격 책정이 어떻게 달라져야 하는지, 좌석 비율이 어떻게 분포되어야 하는지에 대한 방법을 고안하여 다양한 인사이트를 제공하였다.

**2. 결과 활용 및 시사점**

**가. 활용 방안**

본 조는 최대한 현실적인 상황에 이 가격 책정 알고리즘을 적용할 수 있도록 하였다. 현재 대관 규약을 살펴보면 대관자와 예술의 전당의 협의 하에 가격과 좌석 그룹 개수, 가격이 책정된다. 그러므로 선택과 판단은 오롯이 대관자와 예술의 전당 측에 있기에 주최 측의 시각에서 제한된 정보를 가지고 가격 책정을 할 수 있는 방법을 고안했다. 본 조가 판단했을 때 예매율과 공연의 고객층처럼 경험적으로 알고 있는 공연의 경향이 있을 것이라 생각하여, 그러한 정보를 바탕으로 가격 책정과 좌석 그루핑을 어떻게 하는 편이 좋을지 다른 비슷한 공연과 비교한 인사이트를 제공한다. 또한 효과적인 고객의 가격 책정을 위해 어떤 변수를 살펴봐야 하는지, 어떤 고객을 타겟으로 해야 이익이 올라가는지 알려줌으로써 더 명확한 가이드라인을 제공하고 고객을 유도할 수 있는 효과적인 전략을 필 수 있다.

**나. 방향성 제시**

공연의 원가와 영업 비용 등 다양한 비용을 생각하여 좌석 등급의 최소 가격을 주최 측에서 결정한다. 그 후 하려는 공연의 예매율과 사람들이 얼마나 빨리 예매했는지를 파악하여 자신이 속한 클러스터의 타겟층의 나이, 평균 구매액 등을 참고해 가격 책정을 하고 좌석 그루핑을 한다면 효과적인 이익 창출을 할 수 있을 것이다.