|  |
| --- |
| **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 4 명) | **개인/팀명** | 쿡 |
| **지도교사명** | \*스타터 리그에 한함(선택) | | |
| **대표ID** | soonoolimal@dankook.ac.kr | | |

|  |  |
| --- | --- |
| **결과보고서 작성 안내 사항** | |
| **목차** | **Ⅰ. 개요**  1. 배경  2. 목적 및 필요성  3. 분석 수행 범위  **Ⅱ. 문제 수행 내용**  1. 분석(수행) 절차   1. Phase : 0 2. Phase : 1 3. Phase : 2 4. Phase : 3     **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)  - **생성형AI분야의 경우**, 활용하는 생성형AI 툴 종류 명시 및 생성형AI를 활용한 히스토리 필수 제출(생성형AI에 input/output한 내용이 드러나는 소스) |
| **글꼴 및**  **글자크기** | - 본문 글꼴 : 맑은 고딕  - 대분류[1, 2, 3] 항목 : 13포인트(굵게)  - 중분류[가, 나, 다] 항목 : 12포인트(굵게)  - 소분류[ 1), 2), 3) ] 항목 : 12포인트, 본문내용 : 10포인트 |

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

**1. 배경**

예술의 전당은 순수 공연 활성화를 위해 설립된 다양한 공연을 제공하는 문화예술의 중심지로, 이곳에서 열리는 공연들은 이들만의 독특한 특성을 지니고 있다. 이러한 공연들은 공연의 장르, 악단의 규모, 예술가의 스타일, 음악과 무용의 조합 등, 다양한 요소에 따라 다양한 경험을 제공한다. 따라서 관객들의 공연에 대한 선호도는 크게 다를 것으로 예상되며, 이러한 선호도는 좌석 선택에도 영향을 미칠 것으로 기대된다. 예를 들어, 클래식 음악 콘서트를 선호하는 관객은 음향에 최적화된 좌석을 선호할 것이며, 연극을 좋아하는 관객은 배우와 무대와 가까운 좌석을 선호할 것이다.

한편, 이러한 다양성을 고려하지 않고 획일화된 구역에 책정된 가격은 관객들에게 납득하기 어려움을 초래할 것이고 이는 클래식을 비롯한 예술의 전당에서 진행되는 공연 활성화에 부정적인 영향을 미칠 것이다. 본 보고서는 공연의 특성과 관객의 예매 데이터를 기반으로 한 다양한 특성들을 고려하여 좌석 구역과 가격을 조정하는 모델을 제안함으로써, 관객들에게 더 나은 가치를 제공하고 예술의 전당과 관객 간의 상호 혜택을 극대화하는 방안을 모색할 뿐만 아니라 공연 활성화에 기여할 것을 기대한다.

**2. 목적 및 필요성**

1. **공연 특성을 기반으로 한 좌석 그룹핑 및 가격모델 수립**
2. 공연 주체자의 특성을 기반으로 한 공연 분류

다양한 공연을 진행됨에 따라 관객들이 선호하는 좌석들이 상이할 것으로 예상됨으로 공연 특성에 따라 데이터를 분류하고 분류된 데이터를 대상으로 좌석 그룹핑을 적용한다. 이를 통해 공연 특성을 반영한 좌석 그룹핑이 가능하고, 관객들의 선호 요소들을 적용한 좌석 그룹핑이 가능하다.

1. 공연 분류별 좌석 그룹핑

공연 특성을 기반으로 분류된 데이터를 대상으로 관객의 특성 요소가 반영된 데이터를 추가로 활용하여 좌석 그룹핑을 실시한다. 지금까지 쌓여온 데이터들을 통해 비슷한 특성을 가지는 공연들을 그룹화 하고 그안에서 나타나는 데이터특성을 확인한다. 특성을 반영할 수 있는 변수들을 사용해 그룹화된 공연안에서 좌석 그룹핑을 진행한다. 그렇다면 고객들에게 좌석 가격수립의 타당성을 보여줌으로써 서로의 신뢰를 높일 수 있고 또한 좌석 그룹핑을 결정짓는 변수들을 정량적으로 판단해 좌석 그룹핑과 가격의 직접적인 연관성도 나타낼 수 있다.

**3. 분석 수행 범위**

좌석 그룹핑이 진행됨에 따라 코로나에 따른 좌석 띄어앉기라는 특수성이 적용되는 2020년 3월 20일부터 2022년 4월 18일까지의 공연의 ID는 제외한다. 예술의 전당 음악당 안에서도 클래식홀의 가격 선정을 위해 공연장이 리사이틀홀 및 IBK챔버홀인 공연의 ID는 제외한다. 또한, 가격모델 수립을 위한 객단가 추정에 있어서 객단가 추정이 불가한 좌석이 90%가 넘는 공연의 ID를 제외한다.

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

**1. 분석(수행) 절차**

**가. Phase 0 : 데이터 전처리**

이 단계에서는 데이터 분석을 용이하게 만들기 위한 기본적인 전처리를 수행한다.

**1) 기본적인 전처리**

<변수명 변경과 형변환, 좌석 분해>

변수명을 변경하고, 날짜나 시간에 관련된 변수들을 datetime 객체로 변환하고, format을 통일한다. 또한 seat 변수의 텍스트 데이터를 층과 블록, 박스, 열, 좌석 번호로 분해한다.

<멤버십 분해>

멤버십 유형을 나타내는 변수들을 그들이 가진 정보에 기반하여 각 예매자가 가지고 있는 멤버십 유형을 나타내는 변수를 생성한다. 이때, 데이터에 의하면 한 예매자가 여러 멤버십을 가질 수 있다. 따라서 특정 멤버십 각각에 대하여, 가입했다면 1, 그렇지 않으면 0을 할당한다.

-변환 후

<취소표 전환>

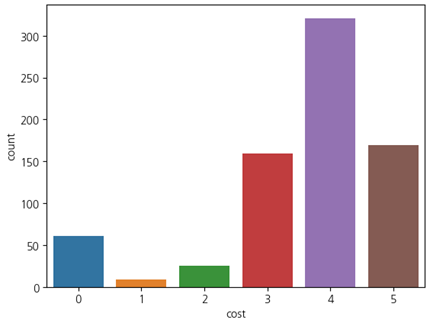
각 공연에 대해 하나의 좌석은 최종적으로 단 한번 예매되어 기록된다. id-seat별로 cancel이 0인 행 중 중복행이 존재하는 경우, 가장 나중에 예매된 정보를 나타내는 행을 제외한 티켓의 cancel에는 2를 할당해 취소표로 전환한다.

<객단가 추가>

각 좌석에 대해 할인이 적용되지 않은 객단가를 추정한다. 이때, 다음의 사항들을 고려한다.

1. price와 dc로부터 얻을 수 있는 할인률(%) 정보 2. 예술의 전당 홈페이지에 명시된 좌석 등급의 총 개수의 상하한 (R, S, A, B, C)

%가 포함된 dc로부터만 할인률을 추정할 수 있으며 dc가 ‘일반’이라면 할인률이 적용되지 않은 경우임을 이용해 객단가를 1차 추정한다. 그 후, 좌석 등급의 최대(소) 개수와 데이터 탐색으로부터 얻은 정보 및 도메인-할인률과 price에 대해 각 공연이 가질 수 있는 모든 경우의 cost 그리드 : \_All\_Price.csv-를 기반으로 객단가를 2차 추정한다. 물론 이 과정은 각 공연들에 대해 독립적으로 수행된다.

****

그 결과, 총 3, 4, 5개의 객단가 종류를 갖는 공연의 수가 가장 많았다. ‘0’은 모든 cost가 1, 2차 추정으로도 추정될 수 없어 모두 결측값인 경우, 다시 말해 애초에 price가 0원인 초대권, 기획사할인으로만 구성된 공연이며, 1과 2의 경우에도 cost의 결측값을 상당한 비율로 포함하고 있음을 확인하였다. 본 경진대회에서 다루는 목적인 ‘여러 관점에서의 복합적인 정보를 고려한 좌석 그룹핑과 그에 대한 가격 모델 수립’에 입각했을 때, 추후 분석은 3, 4, 5개의 객단가 종류를 갖는 공연들에 초점을 맞춘다. 더불어 cost의 일부 결측값들은 이후 분석 과정에서 처리될 것이다.

<좌석 등급화>

기존의 좌석과 가격의 정보 및 위에서 추정한 객단가를 기반으로 각 공연별로 좌석에 등급을 매핑한다.  
 - seat\_rnk : 추정한 cost(객단가)를 통해, 각 공연별로 해당 좌석이 어떤 등급의 좌석인지를 나타낸다.

- total\_rnk : 공연별로 공연의 좌석이 총 몇 등급인지를 나타낸다.

<할인 타입의 범주 조정>

원본 데이터셋에는 매우 다양한 할인 타입이 있다. 이들의 범주를 단순화한다. 이 작업은 크게 2가지 단계로 나눌 수 있다.

1. 사실상 같은 할인 타입으로 볼 수 있는 경우를 통일한다. (e.g. 학생할인10%와 학생 할인 10%, 띄어쓰기 문제)

2. 몇몇 테마를 기반으로 그룹화할 수 있는 할인 타입을 범주화한다.

1. 회원 관련 : 정기회원, 유료회원, 골드, 블루, 그린, 무료, 싹틔우미, 노블

2. 예매자 관련 : 학생, 청년, 경로, 군인, 예비군, 직장인, 임산부, 장애인, 국가유공자, 가족, 예술인, 교직원, 보건의료인, 마니아

3. 예매 관련 : 얼리버드, 당일, 타임세일, 단체, 웰컴, 재관람, 차액, 예매권, 카드

4. 후원 기관 관련 : 호원회원, 단원, 국립합창단회원, KBS교향악단회원, KOSYMI, 심포니회원, 관계자

5. 공공기관 관련 : 강남구, 경기도, 문화릴레이, 문화누리, 문화햇살

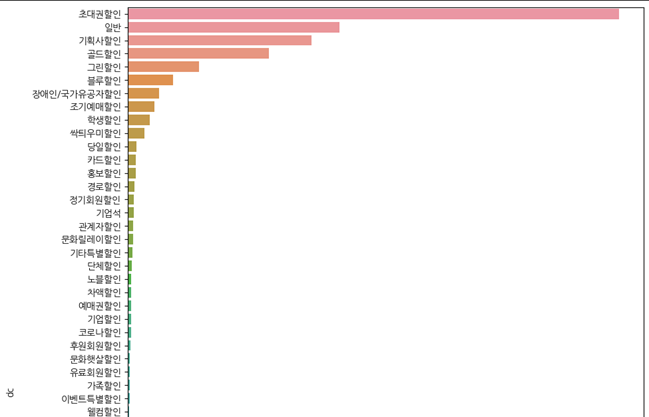
6. 시기 관련 : 연도별 분기 패키지, 기념일, 코로나

7. 공연 관련 : 음악가

8. 특별 할인 관련 : 쿠폰, 이벤트 및 제휴 기관 회원, 유료티켓, 홍보, 블라인드

9. 기업 관련 : 한화, 중앙일보, 신한은행, 신세계, KT

이러한 범주화는 이후 분석을 진행하면서 보다 더 간소화된다.

****

특히, 주목할 점은 기획사할인과 초대권할인의 할인 타입으로 공연을 관람한 사람이 매우 많다는 것이다. 상술했듯 이들은 모두 price의 값이 0이며, 본 경진대회에서 다루는 목적인 ‘여러 관점에서의 복합적인 정보를 고려한 좌석 그룹핑과 그에 대한 가격 모델 수립’에 입각했을 때, 공연을 무료로 관람한 사람들로부터 얻을 수 있는 정보는 매우 적다. 따라서 이후의 분석은 멤버십 관련 할인이나 예매자 관련 정보에 관한 할인 등 예매자에 대한 정보를 추출할 수 있는 할인 타입에 초점을 맞추고 진행될 것이다.

<요일과 시간 보완>

공연 정보로부터 얻을 수 있는 시간적 정보를 나타내는 다음과 같은 변수들을 추가한다.

1. play\_weekday : 공연 요일

2. play\_hour : 공연 시간

3. hour\_cat : 공연 시간의 범주 (launch, afternoon, evening)

4. holiday : 공휴일

5. covid : 코로나 바이러스로 인한 사회적 거리두기 기간

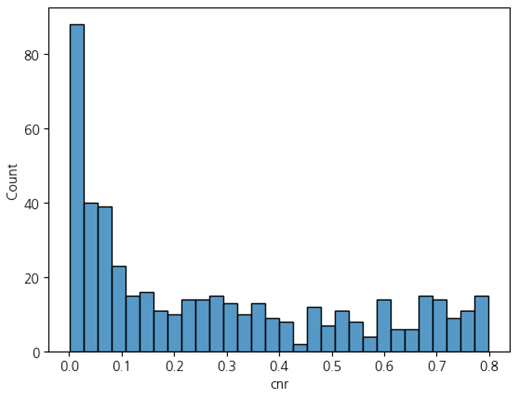
2) 추가적인 전처리

좌석 그룹핑을 위해 좌석의 위치 정보를 정량적으로 반영할 수 있는 변수가 필요하다. 따라서, 예술의 전당 홈페이지로부터 다운 받은 콘서트홀 좌석 배치표를 기반으로, 이를 2차원 그리드에 옮겨 각 좌석의 x, y 좌표를 생성한다. 생성한 좌석의 좌표는 이후 좌석들의 공간적 거리를 이용한 통계량 계산 및 여러 변수들의 값의 추정에서 유용하게 사용될 것이다.

좌석의 블록과 박스가 통합된 sector 변수로 block과 box를 대체한다.

3) KNN을 이용한 객단가 3차 추정

객단가의 NA 비율이 지나치게 높은 공연은 절대 다수의 할인 타입이 초대권, 기획사로 구성됨을 확인했다. 따라서, 객단가의 NA 비율이 0.8을 넘는 공연들은 제외한 후, KNN을 사용하여 객단가의 NA를 채운다. 이때, 객단가의 NA 비율(0 제외)의 분포는 다음과 같다.

****

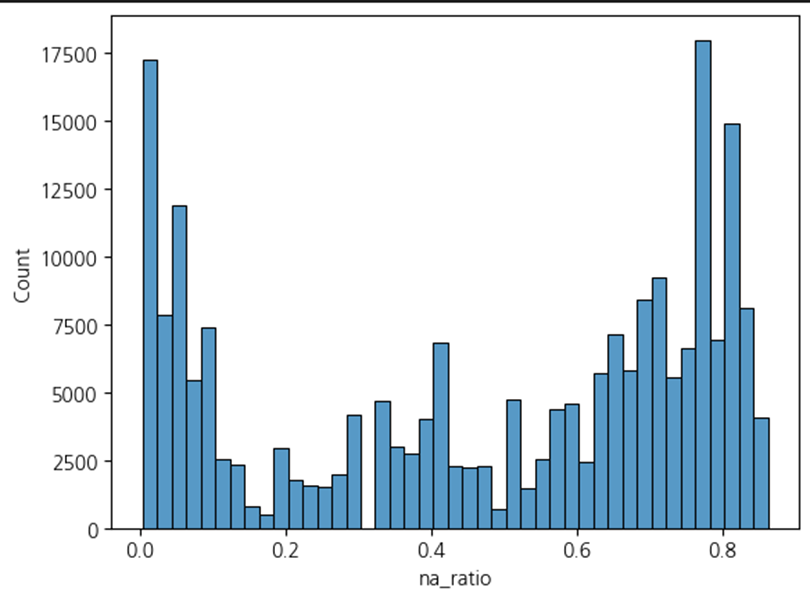
KNN을 사용한 객단가의 NA 전처리는 다음의 3단계로 이루어지며, 단계가 진행될수록 더 공간적 범위를 넓혀 나가며 K개의 주변값(neighbor)를 탐색한다.

1. 식별자 : id-floor-sector

KNN 모델의 식별자를 id-floor-sector로 하여, 생성한 x, y 좌표를 사용해 가장 가까운 주변 3개 좌석의 cost의 최빈값으로 cost의 NA를 채운다. 이때, 식별자를 기준으로 했을 때 데이터 자체가 3개 미만으로 존재하는 경우와, 식별자를 기준으로 했을 때에도 cost가 전부 NA인 경우를 제외하여 다음 단계로 넘긴다.

2. 식별자 : id-floor

KNN 모델의 식별자를 id-floor로 조금 넓혀, 첫 단계에서 처리되지 못한 공연들에 대해 생성한 x, y 좌표를 사용해 가장 가까운 주변 3개 좌석의 cost의 최빈값으로 cost의 NA를 채운다. 이때, 현재 단계에서의 공연별 객단가의 NA 비율의 분포를 보면 NA의 비율이 높은 공연들이 꽤 많다.

****

따라서, 기준값을 정해 식별자를 기준으로 했을 때에도 cost의 NA 비율이 threshold인 80%보다 높은 경우를 제외하여 다음 단계로 넘긴다. 물론 식별자를 기준으로 했을 때 데이터 자체가 3개 미만으로 존재하는 경우도 포함하여 다음 단계로 넘긴다.

3. 식별자 : id

마지막까지 채워지지 않은 cost의 NA는 가장 넓은 단위인 id를 기준으로 한다. 같은 과정을 반복하되, 이제는 cost의 NA가 없도록 모든 경우를 처리한다.

**나. Phase 1 & 2 : 공연 특성을 활용한 대분류와 좌석 그룹핑**

이 단계에서는 공연 특성을 활용한 공연의 대분류와 해당 대분류 내에서의 좌석 그룹핑을 진행한다. 해당 두 과정에서 주요하게 고려한 사항은 두 가지이다.

1. 공연 특성을 활용한 대분류

공연 특성을 활용한 대분류에서는 예술의 전당에서 공연이 런칭되기로 결정된 상태를 가정한다. 즉, 공연과 관련된 정보인 장르, 공연 시간, 러닝타임, 인터미션 등만이 활용할 수 있는 정보이며 본 대회에서 주어진 대부분의 주요 정보를 담고 있는 개인에 관련된 정보-나이, 성별, 멤버십, 예매 거래 시간, 할인 타입- 등은 활용할 수 없다. 이들을 공연의 분류에서 활용하는 것은 data leakage에 해당하기 때문이다.

2. 예매자의 개인 정보를 반영한 좌석 그룹핑

공연의 대분류가 완료되었다면, 이제 본 단계에서 예매자의 개인 정보를 활용해 해당 대분류 안에 좌석 그룹핑을 진행한다. 이 과정에서는 무작위적이고 산발적으로 보이는 다방면의 개인 정보-나이, 멤버십, 할인 타입, 성별 등-에 대해 대표성을 가지면서, 공연의 대분류 사이에서는 상이하지만 공연의 대분류 내에서는 유의미한 설명력을 가지되 좌석이 군집화될 수 있도록 치우치지 않은 분포의 통계량들을 적절하게 선택해 좌석 그룹핑에 해당 통계량들이 고른 영향을 줄 수 있도록 유도하는 것에 초점을 맞추었다.

가능한 모든 공연 특성의 조합을 시도해 공연들을 조합하고, 그에 대해 좌석에 대한 예매자 개인 정보들이 적절하게 상호작용할 수 있는 통계량들을 생성해 좌석의 클러스터링을 시도했다. 그 중 가장 적절한 대분류와 통계량의 조합은 아래와 같다고 판단된다.

<공연 특성을 활용한 공연의 대분류>

원본 데이터셋에 제시된 장르를 범주화하여 공연 대분류의 기준으로 사용했다. 장르 중 교향곡은 공연 악단의 규모가 크다. 반면 클래식, 독주, 실내악, 재즈는 교향곡보다는 그 규모가 작다. 또한 합창, 성악, 콘서트, 오페라, 기타, 복합장르, 크로스오버는 공연의 핵심이 악기 소리가 아닌 사람의 목소리이며 사람이 주체가 된다. 이러한 판단 아래에서 장르를 3개의 범주로 묶어 공연을 대분류하였다. 이 대분류는 ‘split’ 변수에 반영되어 있다.

<예매자의 개인 정보를 활용한 좌석 그룹핑>

위의 대분류 내에서 좌석을 예매한 개개인의 특성들을 적절히 반영할 수 있도록 seat를 index로 여러 통계량들을 생성하였고, 해당 통계량들로 데이터셋을 새롭게 재구축해 좌석 그룹핑에 사용하였다. 생성한 통계량들은 모두 대분류 내에서 계산되었고, 변수가 내포하는 의미에 따라 전체 데이터셋과 예매 데이터셋 중 하나만을 사용하였다. 변수들은 다음과 같다.

A. 예매 수 (book\_cnt)

특정 좌석의 예매 횟수(cancel=0)를 나타낸다. 이 통계량은 좌석에 녹아 있는 예매자의 개인 정보는 주요하게 반영하지 못하지만, 각 좌석들의 인기도 자체를 나타낼 수 있는 가장 강력한 통계량이다.

B. 좌석의 위치 정보

-dist

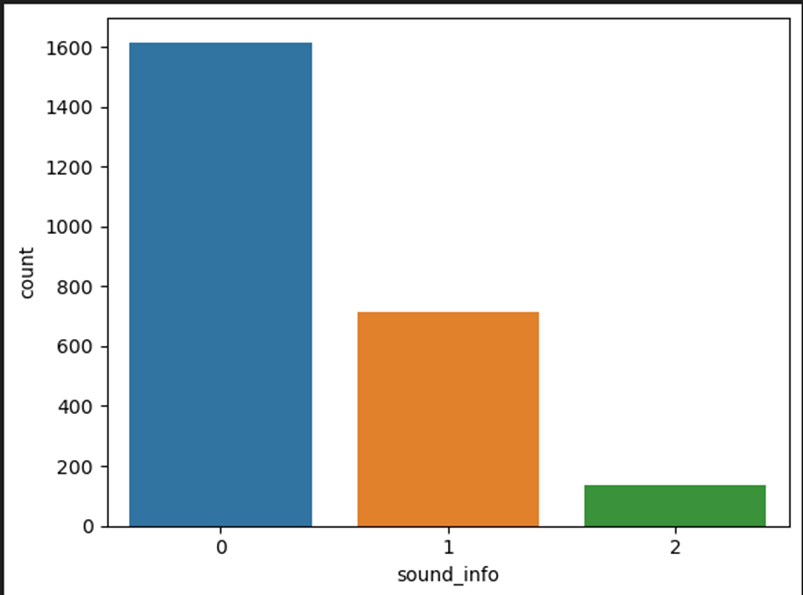
이 변수는 Phase 0의 전처리 단계에서 생성한 (예술의 전당 콘서트홀 좌석 배치도를 기반으로 좌석 각각에 매핑된) (x, y) 좌표와, 무대의 중심으로부터의 거리(dist)이다. 이 통계량은 개개인이 좌석을 구매할 때 가장 우선적으로 고려하는 ‘무대와의 거리’를 직접적으로 반영한다.

-dist\_f\_rnk

이 변수는 층별로 좌석들을 그룹화한 후, 좌석과 무대 사이의 거리(dist)를 정렬해 좌석에 랭킹을 부여한 변수이다. 앞서 생성한 좌석의 그리드와 dist는 무대와의 거리를 반영하지만, ‘층’에 대한 위치 정보를 빈약하게 반영한다. 예술의 전당 홈페이지에서 얻은 좌석 배치도는 층을 구분하지 않고 제시되어 있기 때문이다. 따라서 높이 정보를 반영할 필요가 있으며, 좌석 배치도와 실제 콘서트홀의 좌석 배치의 차이에서 오는 오차를 최소화하기 랭킹을 부여했다.

C. 좌석의 음향 정보 (sound\_info)

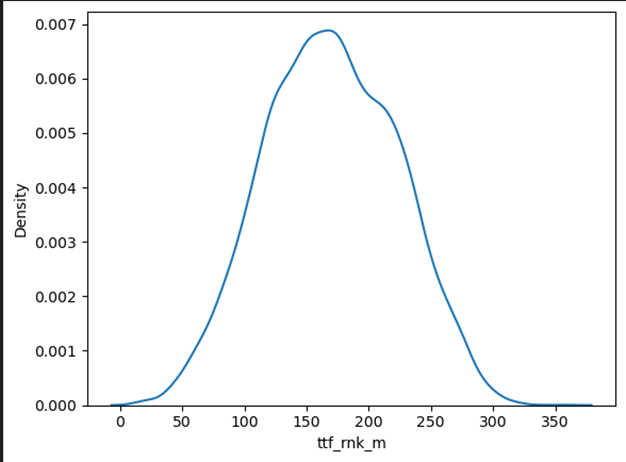
이 변수는 다양한 매체의 보도와 후기를 기반으로 얻은 음향 효과가 극대화되는 좌석을 일반 좌석과 구분하는 변수이다. 예술의 전당에서 열리는 공연을 자주 관람하고, 관심이 많은 고객일수록 이러한 정보의 영향력이 높아질 것이지만, 처음 예술의 전당에서 공연을 관람하거나 평소에 이러한 류의 공연에 관심이 적은 사람들에 대해서는 이 변수의 영향력이 낮아질 것이다. 따라서 본 변수는 좌석의 위치적 효과뿐 아니라 좌석을 예매한 개인의 정보와도 유의미한 연관이 있을 것이다.

****

D. 좌석의 인기도 : 예매 거래 시간의 순위와 가중 순위 (예매 데이터셋 활용)

-ttf\_rnk\_m

이 변수는 공연별로 좌석에 대해 해당 좌석의 예매 시간 1시간 단위로 정렬한 후 랭킹을 부여한 후, 해당 랭킹의 평균을 좌석에 대해 계산한 변수이다. 예매가 오픈된 이후 빠르게 예매가 완료된 좌석일수록 해당 좌석은 인기가 높은 좌석임을 의미하며, 그러한 인기도는 공연별로 다를 것이다. 즉, 이 통계량은 공연 정보와 좌석의 인기도 사이의 연관관게를 나타내어 줄 수 있으며, Phase 1의 공연 대분류와 좌석 사이의 관계성을 클러스터링에 반영하는 데 유의미할 것이다.

****

E. 좌석과 예매자 (전체 데이터셋 활용)

-people 데이터셋

****

이 데이터셋은 전체 데이터셋에 대해 특정 좌석에서의 개인 정보만을 나타낸 데이터셋이다. 이를 활용해 대분류에 속하는 공연들에 대해 아래의 변수를 만들어 각 좌석에 예매자의 개인 정보가 유의미하게 구분되는 것을 유도하였다.

- free\_0\_wr\_m, free\_1\_wr\_m

이 변수는 people 데이터셋을 활용하여 각 좌석의 예약 순서별 순위를 멤버십 정보에 따라 유료 회원과 무료 회원으로 나눈 후, 그들의 평균을 계산한 변수이다. 이때, 아예 예약되지 않은 좌석에는 가장 마지막 순위에 1을 더한 순위를 부여하였다. 이 변수는 앞선 좌석에 대한 통계량들과는 달리 좌석의 위치 정보는 전혀 반영하지 않았으므로, 멤버십을 가진 회원들과 그렇지 않은 회원들 사이의 선호 좌석의 차이를 보여줄 수 있을 것이라 기대한다. 그리고 이러한 효과는 위에서 도입한 다양한 좌석에 관한 통계량들과 함께 상호작용하여, 특정 위치의 좌석 대한 개개인의 정보와 선호도가 좌석 그룹핑에 반영될 것이다.

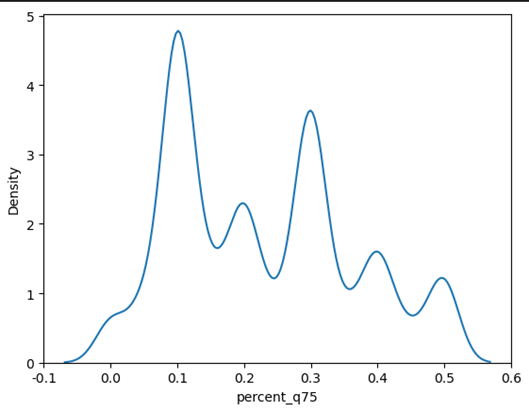
- age\_10\_20, age\_30\_40\_50, age\_60\_70

이 변수는 각 좌석별 특정 연령대에 속하는 사람 수를 나타낸다.

F. 좌석과 할인률 (전체 데이터셋 활용)

-percent\_975

이 변수는 좌석에 대해 할인률의 3사분위수를 구한 통계량이다. 위의 변수들은 좌석에 대해 멤버십이나 위치 정보만을 고려했지, 할인률을 고려하지 않았으므로 그에 관한 정보의 필요성에 의해 도입하였다. 평균이나 중앙값이 아닌 3사분위수를 선택한 이유는, 분명 할인 타입으로서 프로모션이 있었지만 할인률을 알 수 없는 행들이 다수 존재하기 때문에 이들의 값이 0으로 처리되어 전체적인 통계량의 값이 작아지는 효과를 상쇄하고 할인률 정보를 보다 고르게 반영하기 위해서이다.

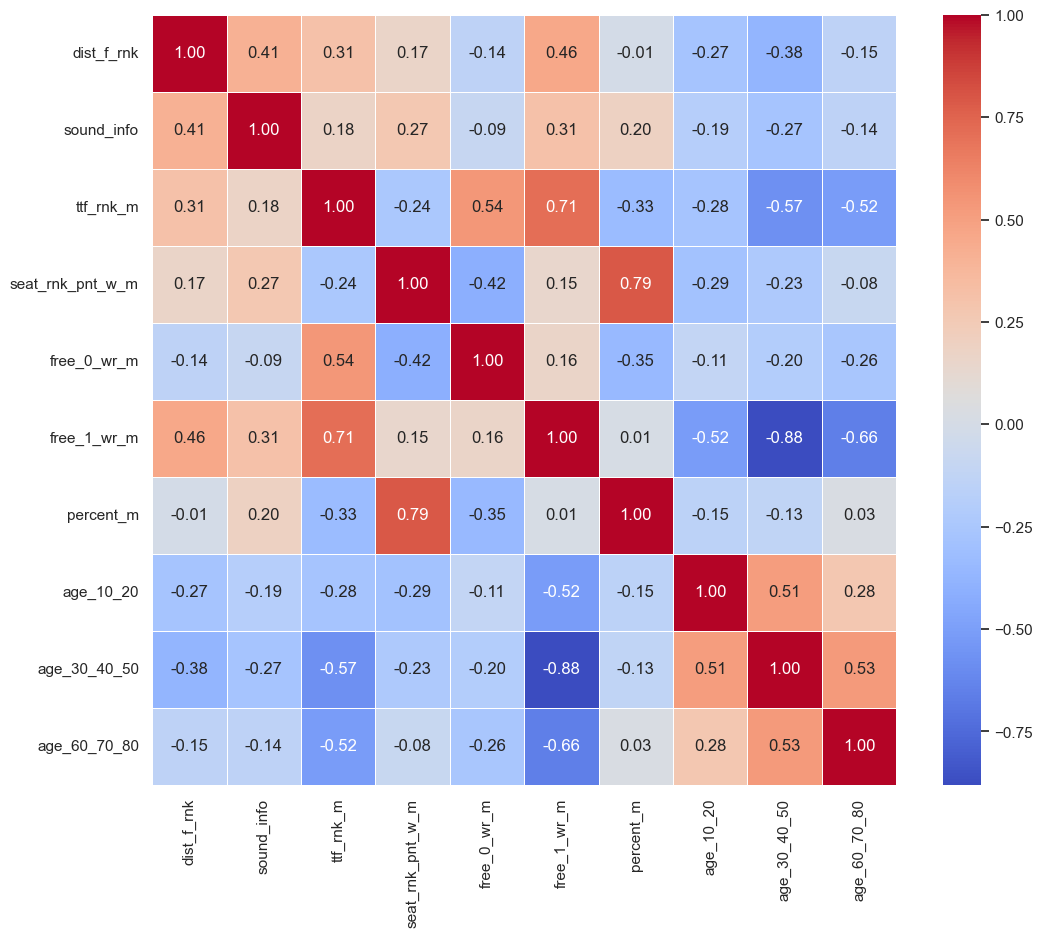
****

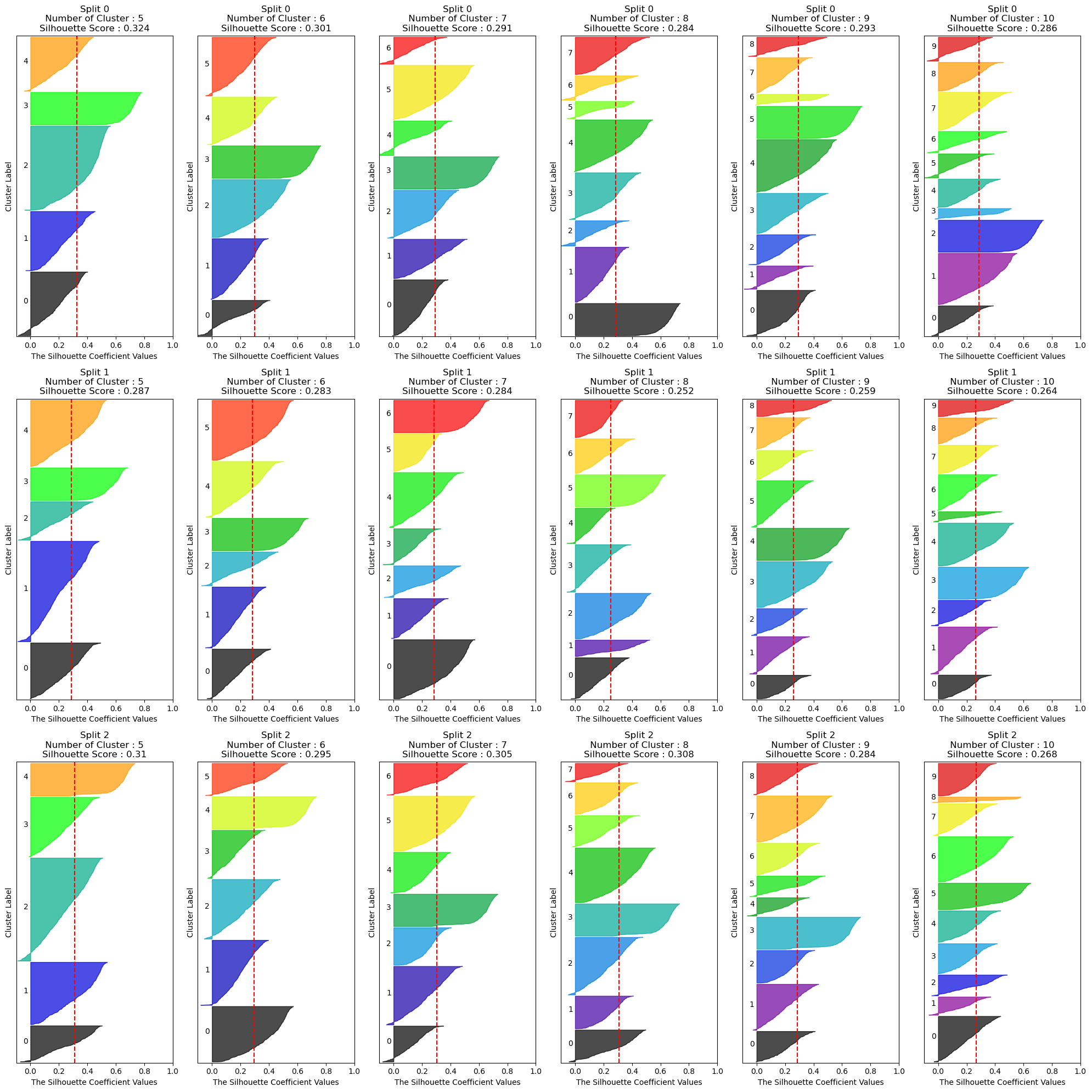
-cost\_m

이 변수는 좌석별 평균 객단가를 계산하였다. 할인률과 객단가의 상호 연관성을 반영하기 위해 추가하였다.

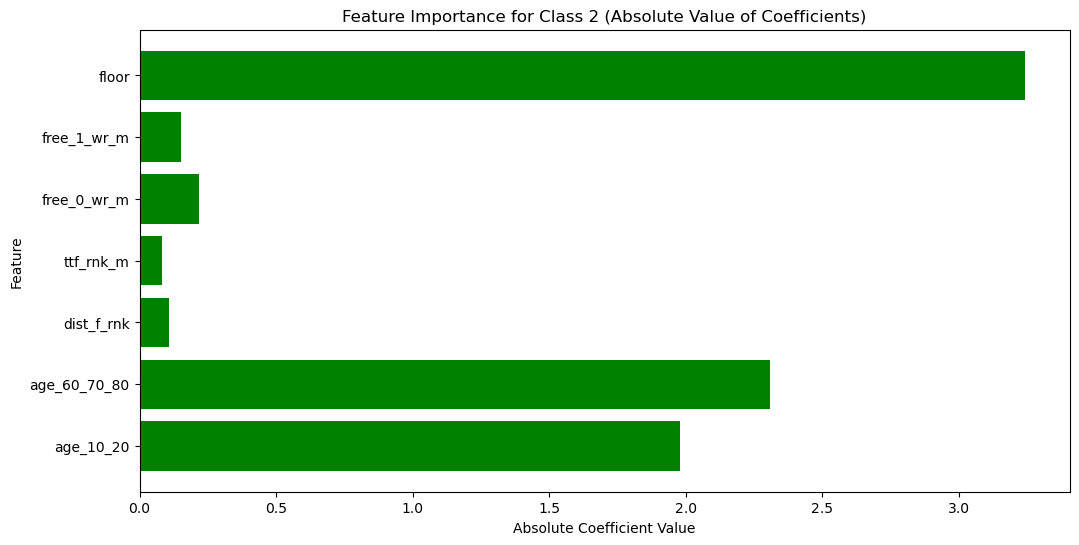
위의 통계량들의 조합을 달리 시도하면서 가장 유의미한 좌석 클러스터를 형성하는 변수를 최종 선택하였다.

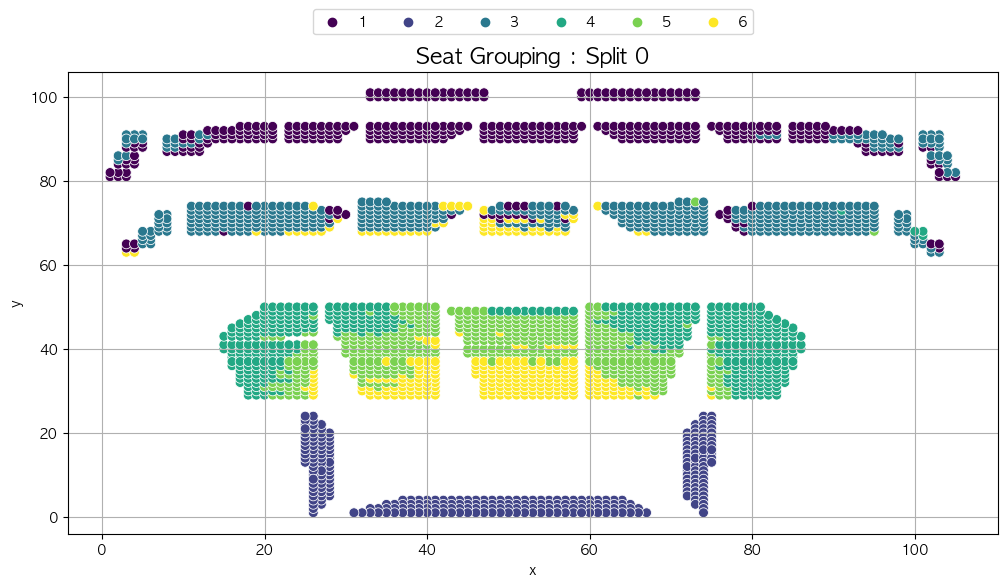
위에서 정의한 변수들을 통해 특성에 맞게 나눈 공연들 안에서 고객들의 특성과 거리, 층과 같은 위치적인 정보를 조합해 최적의 좌석 그룹핑을 진행한다. K-Means는 비지도 학습 알고리즘 중 하나로, 주어진 데이터를 여러 개의 클러스터로 그룹화하는 데 사용된다. 클러스터링은 데이터 포인트를 그룹으로 나누는 과정을 의미하며, 각 클러스터는 비슷한 특성을 가진 데이터 포인트들의 모음이다. 위에서 만든 변수들은 모두 연속형 변수이고 각 관측치들 간의 거리를 측정하기 위해 minmax 스케일링을 모든 변수에 대해 (0,1)에 범위에 대해 진행한다. 변수들의 상관관계가 높게 되면 문제가 생길 수 있기 때문에 상관관계에 대한 히트맵을 그려본 후 상관계수가 0.8이상인 변수의 조합은 kmeans 클러스터링에 포함되지 않게 실행했다.

****

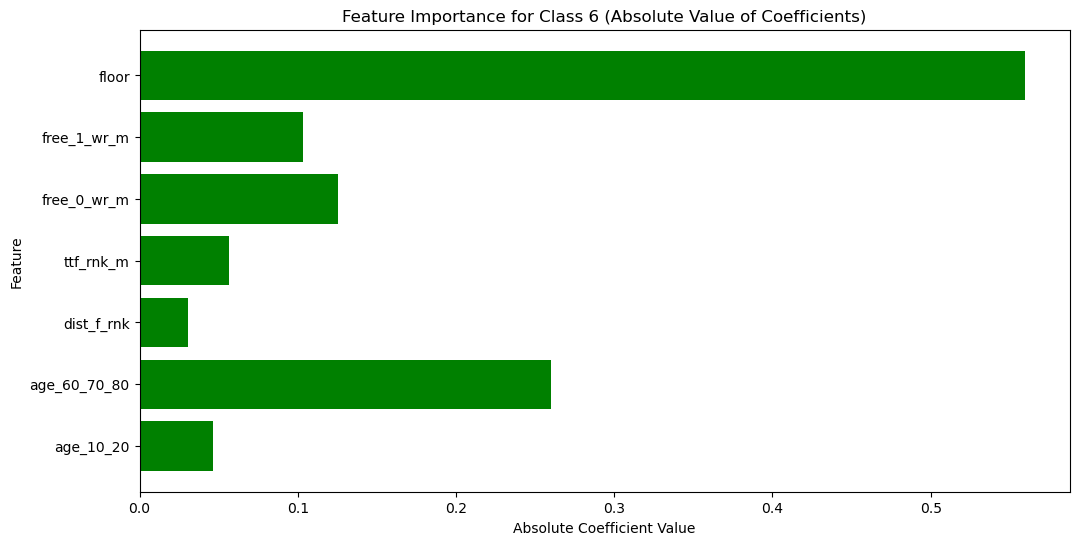
최종 선택된 변수는 'age\_10\_20', 'age\_60\_70\_80', 'dist\_f\_rnk' ,'ttf\_rnk\_m','free\_0\_wr\_m','free\_1\_wr\_m','floor' 로 결정됐고 최적의 k는 군집간의 거리와 군집내의 거리를 계산해 군집이 잘형성됐는가를 판단해주는 실루엣계수의 각 클러스터들의 값과 전체 평균값이 상위인 클러스터를 선택했다. ****

그 후 3개의 장르로 나눠진 공연 데이터에 대해 위에서 결정된 클러스터를 y로 삼아 로지스틱 회귀분석을 진행하고 회귀계수의 절댓값을 통해 각 클러스터내에서 어떤 변수가 의미있는지를 파악했다. 각 공연 데이터 별로 특징이 되는 클러스터 하나씩의 변수 중요도를 보았다.

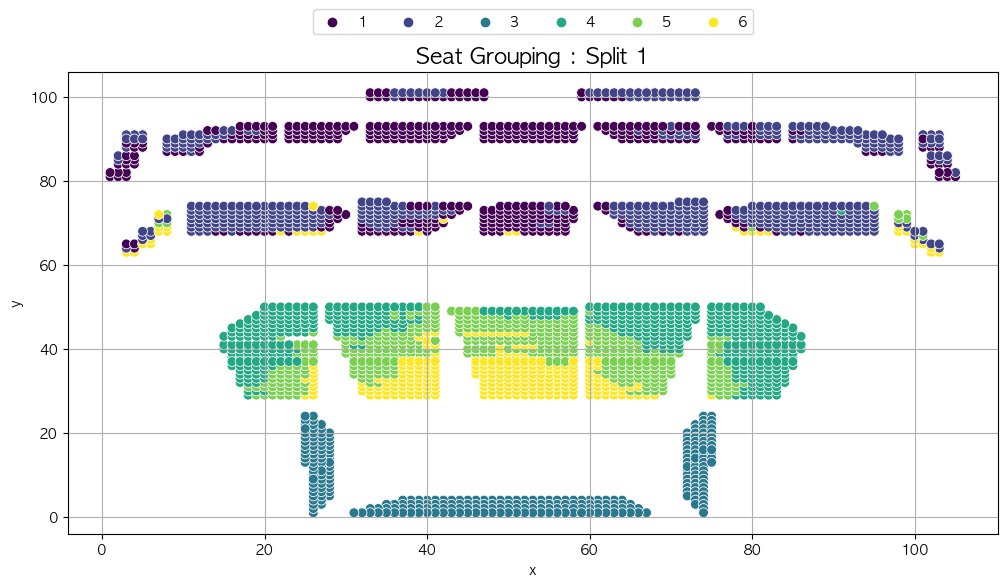
****

장르 0의 데이터에서 2번 클러스터의 변수 중요도를 보면 거리를 판단하는 floor가 가장 큰 중요도를 가지고 있지만 그 뒤로 60,70,80회원들과 10,20회원들의 중요도도 높게 나오는 것을 확인할 수 있다.****

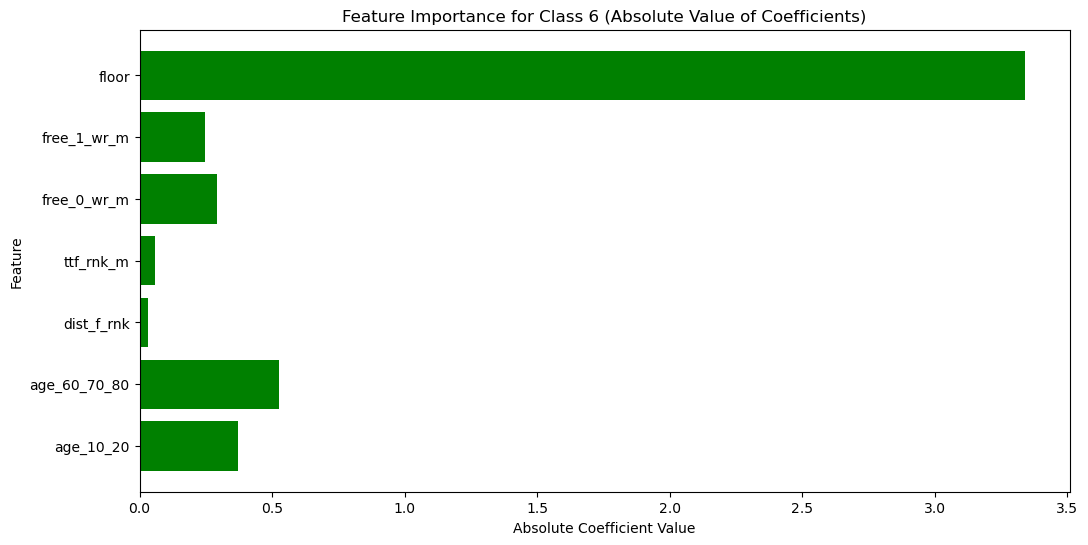
합창석 쪽의 좌석들이 2번 클러스터의 대부분을 차지하고 있는데 합창석이 0층인 것과 10,20,60,70,80대가 교향곡 공연에서 합창석 좌석을 선호하는 내용을 통해 클러스터가 결정됐다고 볼 수 있다.

****

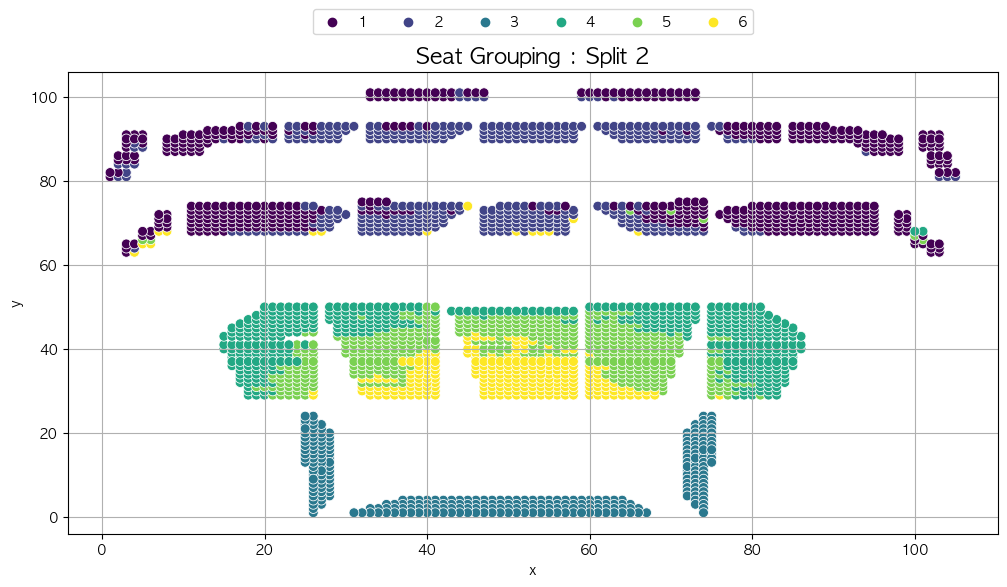
장르 1에서도 floor가 가장 큰 요인을 차지하지만 그 뒤로 60,70,80대의 선호와 전체 회원들의 선호도 포함되어서 클러스터링 된 것을 알 수 있다.



1층 맨앞 가운데의 클러스트에서는 위에서 말한 특성을 가진 고객들이 높은 선호를 보인다고 해석할 수 있다.

****

장르 2인 공연에서는 floor 변수 외에 나머지 변수의 값이 두드러지게 보이는 변수가 없다. 나머지 클러스터에서도 비슷한데

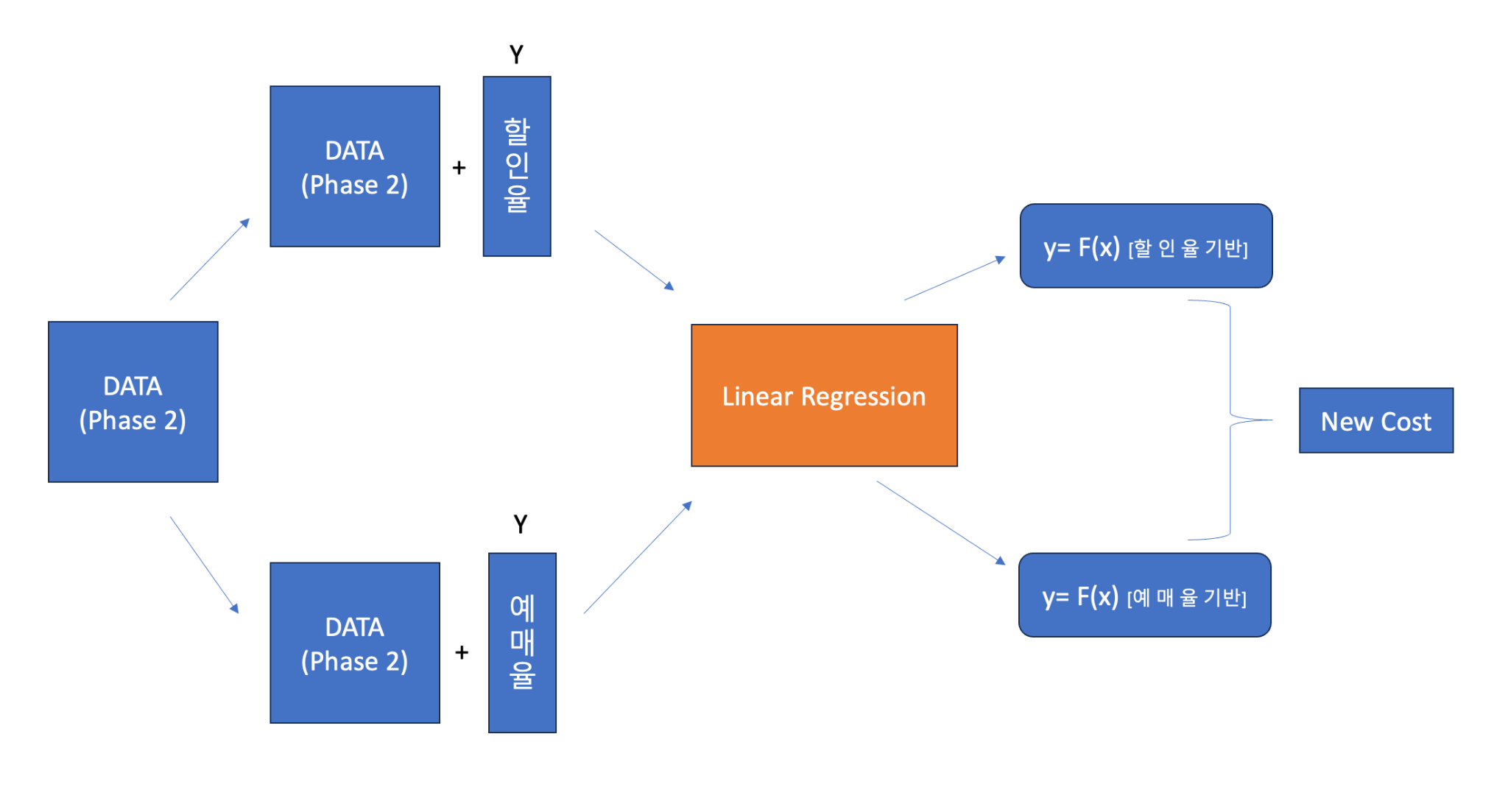
****

이러한 공연에서는 특정 고객들의 선호가 있다기 보다 층과 거리등 위치적인 요소가 좌석을 그룹핑하는데 더 많은 영향을 끼쳤다고 해석할 수 있다**.**

**다. Phase 3**

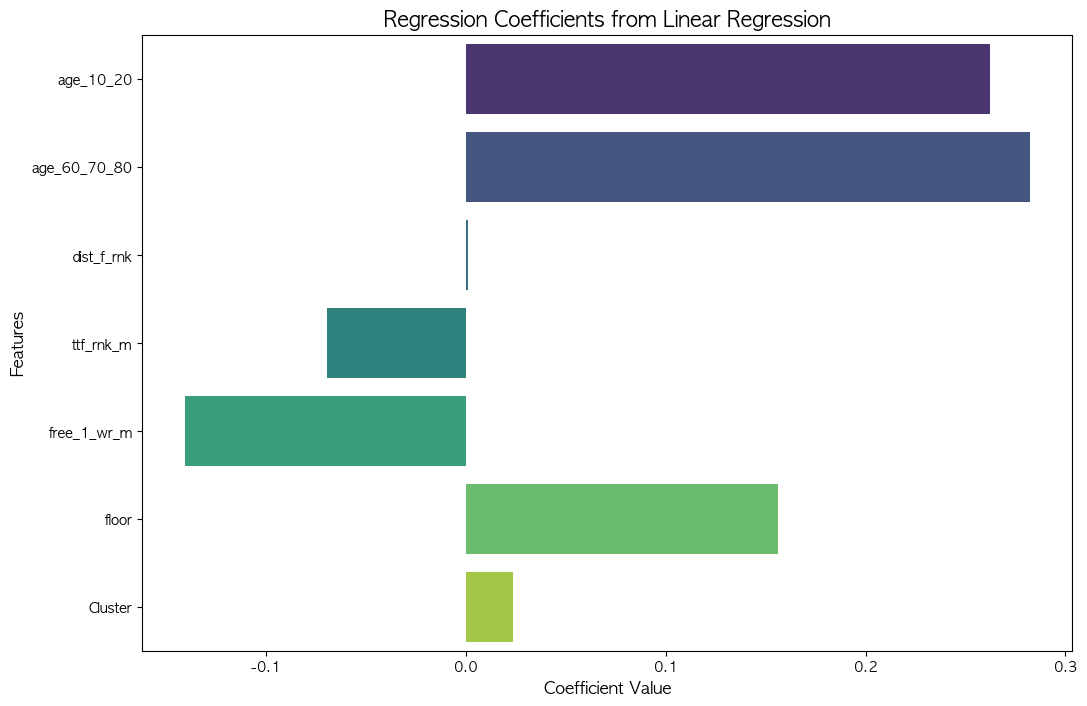
* **가격 모델 수립**

공연 특성을 기준으로 분류된 세 데이터에 대해 Phase 2 에서는 각 분류 별 좌석 클러스터링을 실시하였다. 군집화된 좌석에 대해 그들만이 보유한 특성과 이 좌석들을 이용하는 관객들의 특성을 반영하여 가격을 산정하고자 한다. 먼저, 가격에 직접적인 영향을 주는 두가지 요소에 대해 선형 회귀 모델을 적용한다. 이 때 가격에 직접적인 영향을 주는 두가지 요소로는 각 좌석에 대한 평균 예매율과 평균 할인율이다. 평균 예매율에 대한 데이터 수집은 앞서 나눈 세 데이터에 대해 각 좌석이 예매된 횟수를 해당 데이터에 속해 있는 id의 총 개수로 나눔으로써 확보할 수 있었다. 평균 할인율은 예매 데이터의 할인 타입에 대해 확인 할 수 있는 할인율에 평균을 구함으로써 확보할 수 있었다. 이렇게 확보된 평균 예매율, 평균 할인율을 클러스터링에 활용된 변수와 좌석이 가지고 있는 지리적 데이터로 선형 회귀 모델을 적용했을 때 나오는 회귀 계수를 최종 가격 선정 모델에 활용하였다. 가격 모델의 전체적인 프로세스는 아래와 같다.



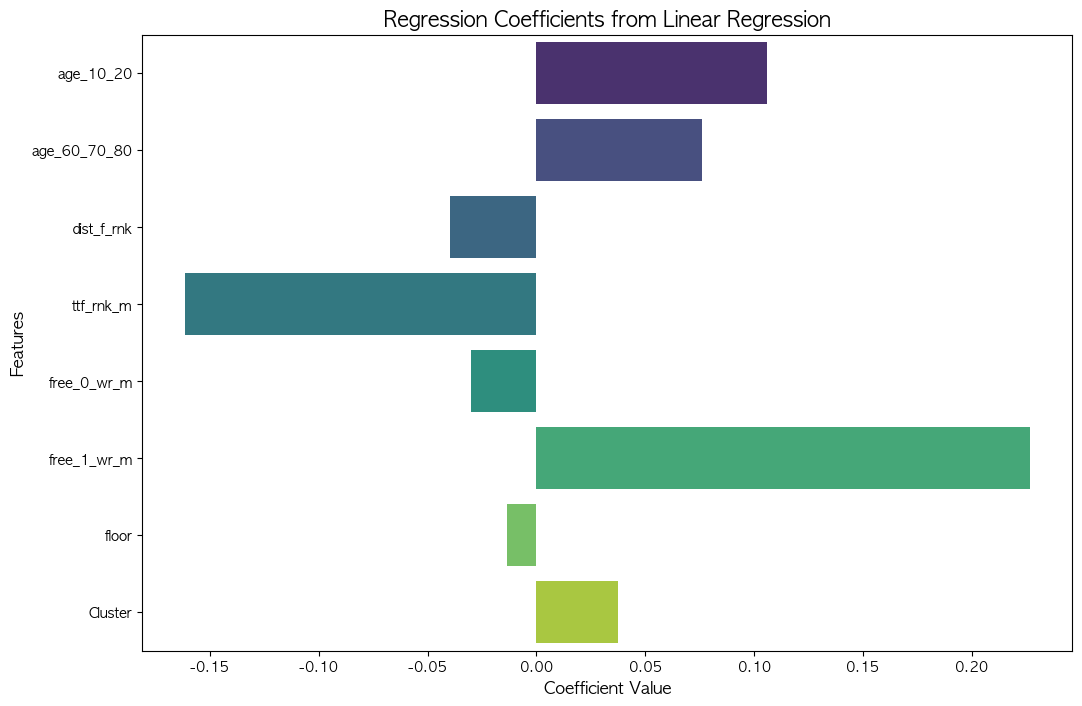
1. Linear Regression - 예매율

예매율의 경우 가격에 큰 영향을 주는 중요한 변수로 예매율이 증가하면 좌석의 가격을 높히고 예매율이 하락하면 좌석의 가격을 낮추는 효과를 기대할 수 있다. 따라서 예매율에 관한 회귀 모델의 경우 클러스터에 사용한 변수들을 활용하여 회귀 모델을 학습 시키고 이로써 확인 할 수 있는 회귀 계수를 토대로 가격 산정에 대한 설명력을 높일 수 있다. 아래는 공연 특성을 기반으로 한 분류 0번 데이터에 좌석 그룹핑을 적용한 후의 데이터를 기반으로한 선형 회귀 모델 학습 결과이다. 0번 분류의 경우 대규모 악단이 공연 주체자로써 좌석 그룹핑은 좌석 선호도가 앞쪽에 국한 되는 것이 아닌 고루 분포되어있음을 확인 할 수 있고, 이 특성을 바탕으로 예매율에 대한 선형 회귀 모델을 적용한 결과, 연령대, ttf\_rnk\_rn, floor, Cluster의 경우 예매율을 상승시키는 요인으로, 거리, free 데이터의 경우 예매율을 하락시키는 요인으로 확인 되었다. 이러한 특성을 가진 예매율은 가격과 비례하는 요인으로 가정하며, 예매율이 높은 자리 일수록 가격이 비싸지고, 예매율이 낮은 자리 일수록 가격이 낮아진다는 조건하에 아래 그래프에서 확인할 수 있는 모든 회귀 계수를 고려하여 선정 가격에 대한 “증가” 요인으로 활용하고자 한다.



1. Linear Regression - 할인율

할인율 또한 가격에 크게 영향을 주는 변수로서 할인율이 상대적으로 높은 자리 일수록 높은 가격으로 책정된 자리이고, 낮은 자리 일수록 낮은 가격으로 책정된 자리라는 가정을 기반으로 할인율을 가격 선정 요소에 적용하고자 한다. 따라서 앞서 언급한 예매율 기반 선형 회귀 모델과 비슷하게 클러스터에 사용한 변수들을 활용하여 회귀 모델을 학습 시키고 이로써 확인 할 수 있는 회귀 계수를 토대로 가격 산정에 대한 설명력을 높일 수 있다. 아래의 그래프는 공연 특성을 기반으로 한 분류 0번 데이터에 좌석 그룹핑을 적용한 후의 데이터를 기반으로한 선형 회귀 모델 학습 결과이다. 클러스터는 커짐에 따라 평균 가격이 증가하는 추이를 보이는데 할인율을 대상으로 하는 회귀 모델의 결과를 보면 클러스터가 커짐에 따라 할인율이 증가하는 것을 확인 할 수 있고, 연령대는 모두 비율이 커짐에 따라 할인율이 증가하는 양상을 보이며, dist\_rnk 는 커짐에 따라 할인율이 감소하는 것으로 보아 가까운 자리 일수록 할인율이 증가한다는 것을 확인할 수 있다. 이러한 특징을 기반으로 이 모델에서 확인할 수 있는 모든 회귀 계수를 고려하여 선정 가격에 대한 “감가” 요인으로 활용하고자 한다.



1. Linear 예매 + Linear 할인

가격 선정에 있어 예매율을 바탕으로하는 회귀 모델을 증가 요인으로, 할인율을 기반으로 한 회귀 모델을 감가 요인으로 적용할 때 기대할 수 있는 이점들은 다양하다. 첫번째로, 지속적으로 가격 선정 모델에 대한 업데이트가 가능하다. 가격 선정에 사용되는 회귀 모델은 모두 관객의 예매 데이터와 좌석 고유 특성의 데이터를 기반으로 학습된다. 이는 데이터가 변함에 따라 모델의 학습 양상이 변화한다는 점과 직결되며, 새로운 공연에 대해 가격을 산정할때 이 공연과 같은 분류에 있는 모든 공연 정보들을 활용해 가격을 추가적으로 보정하거나 새로 선정할 수 있다. 뿐만 아니라 새로 선정된 공연도 추후 관객 정보 데이터를 수집할 수 있으므로 지속적으로 발전하고 업데이트 된다는 장점이 있다. 두번째로, 가격 산정에 대한 타당성 확보가 가능하다. 설명 가능한 모델을 기반으로 다양한 산업, 특히 보수적으로 접근해야하는 금융권과 같은 산업 분야에서 아직까지 활발하게 사용되고 있는 회귀 모델을 기반으로 함에 따라 어떤 요인들이 가격 선정에 영향을 미쳤는지 충분히 설명이 가능하며 신뢰구간을 통해 상한 하한을 설정하여 공연 기획자 측에 선택의 폭을 늘려줄 수 있다. 이를 테면, 공연 특성의 군집 별로 상이하게 드러나는 특정 변수를 제시하고 설명할 수 있다. 마지막으로, 클러스터를 대상으로 하는 것이 아닌 각 좌석 별로 가격을 다르게 할 수 있다는 용이함이 있다. 공연 좌석 그룹핑에 있어 다양한 좌석 가격을 제시하는 것은 관객들에게 혼란을 줄 수 있으나, 공연 기획자 측에서는 이를 활용하여 공연 가격 선정에 다양성을 확보할 수 있고, 신뢰구간과 더불어 이들이 선택할 수 있는 폭을 늘려준다. 본 보고서에서 기술되는 가격 산정 방식은 회귀 모델을 통해 제시하는 각 좌석에 상이한 가격들을 앞서 구분지은 클러스터를 대상으로 그룹화 하여 평균을 내어 7가지의 좌석 그룹핑별 가격을 제시한다. 아래는 활용 예시이다.

.

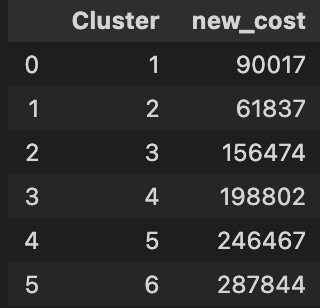
위 의 데이터 프레임은 0번 분류 데이터에 대해 도출된 각 좌석별 가격 모델을 나타낸다. ‘seat’ 열은 예술의 전당 콘서트 홀에서 앉을 수 있는 모든 좌석이며, ‘Cluster’는 0번 분류 데이터에 대해 좌석 그룹핑을 하고 이에 맞게 좌석 그룹핑 번호를 부여한 열이다. ‘appreciation’은 앞서 언급한 ‘예매율' 기반으로 선정된 회귀식에서 기존 보유한 데이터를 적용하여 최종적으로 출력된 ‘증가'요인 이며, ‘depreciation’은 ‘할인율' 기반으로 선정된 회귀식에서 기존 보유한 데이터를 적용하여 최종적으로 출력된 ‘감가'요인이다. 새로운 데이터에 대한 적용은 아래와 같다.



위의 데이터프레임은 앞서 제시한 데이터 프레임에 0번 분류에 해당되는 2022년 12월 27일에 진행된 2631번 공연에 대한 좌석별 cost를 좌석을 기준으로 병합한 것이다. 여기서 확인 할 수 있는 NaN 값들의 이유는 데이터가 전처리 과정에서 예매 데이터를 기반으로 cost 값을 채운 것이므로 예매가 되지 않는 좌석에 대해서는 cost 값을 추론할 수 없기에 발생한다. 이는 새로운 데이터 적용 시 공연 기획사 측이 초기에 제시한 디폴트 cost 값으로 대체가 가능하기에 NaN값에 대해서는 처리하지 않고 진행한다. 이 데이터를 기반으로 New cost 선정식인 아래의 식을 적용하여 New cost 열을 만든다.

**New Cost = Default Cost \* [ 1 + ( Appreciation - Depreciation) ]**

이로써 구할 수 있는 좌석 그룹핑 별 최종 New Cost 는 아래의 데이터 프레임과 같다.



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

**1. 주요 결과 요약**

**가. 프로세스 절차 별 요약**

1. Phase 0

공연 특성 정보를 기반으로 제공된 데이터를 3개의 데이터로 분류함. 분류된 데이터를 대상으로 결측치, 이상치, 활용할 수 있는 새로운 열 생성을 진행하였다.

1. phase 1

Phase 1에서 분류된 데이터를 대상으로 히트맵과 각종 지표를 통해 데이터 분류에 대한 가설을 검증한다.

1. phase 2

세가지로 분류된 데이터를 대상으로 좌석 클러스터를 진행하고 특성에 맞게 군집 번호를 지정함

1. phase 3

군집에 특성에 맞게 예매율과 할인율을 기반으로 계산된 회귀모델을 활용해 가격 모델을 수립함.

**2. 결과 활용 및 시사점**

1. 지속적으로 업데이트 가능한 가격 선정 모델

phase 3 에 언급했듯, 본 보고서에 제안한 가격 모델은 공연 특성 정보와 관객 데이터를 누적하여 사용하고 이를 모델에 반영함으로써 추후에 축적될 데이터에도 충분히 사용가능한 모델을 제시한다. 뿐만 아니라 모델을 설계함에 있어 Data Leakage를 고려하여 적용할 데이터에 관객 예매 데이터가 없어도 가격 선정 요소에 대한 반영이 가능하기에 타당성을 제시할 수 있다.

2. 공연 특성을 기반으로 좌석의 지리적 요인, 관객의 선호도를 반영한 가격 모델

활용 데이터는 가설에서 비롯된 공연 특성별로 집계된 데이터이므로 공연 특성이 반영되었으며 클러스터링에서 사용된 예매 데이터 또한 공연 특성을 기반으로 나누어진 데이터에서 집계된 통계량들, 데이터들이기에 공연 특성, 좌석의 지리적 요인, 관객의 선호 통계를 모두 활용한 가격 모델이다.

3. 고정된 가격 제시가 아닌 공연 기획사의 선택지를 반영한 가격 모델

최종적으로 제시한 가격 모델에 들어가는 Default Cost 는 과거 축적된 데이터를 기반으로 한 좌석 그룹핑 별 평균 Cost도 가능하지만, 공연 기획사 측에서 최초에 제시한 좌석별 Cost를 가지고 누적된 공연 특성, 좌석의 지리적 요인, 관객의 선호도 데이터를 기반으로 기획사에서 제시하는 Cost의 범주에서 크게 벗어나지 않지만 위의 요소들이 반영된 합리적인 가격을 제시함으로써 관객 뿐만 아니라 기획사 측에도 설명력 있는 가격을 제시할 수 있다.

**3. 추가 제언**

데이터에 많은 노이즈가 있어 결정적 요인의 충분한 반영에 있어 부족함이 있지만, 공연 특성 정보에 충분한 활용 및 추가적인 데이터 활용이 가능하면 현재 제시한 가격 모델보다 다양한 요소들이 추가적으로 반영된 합리적이고 설명력 있는 모델을 제시할 수 있을 것이다.