제11회 「2023 빅콘테스트」 결과보고서

* 해당란에 ☑ 표시

참가분야	☐ 생성형AI 분야 ☐ 정형데이터 분석 분야 ☐ 빅데이터플랫폼 활용 분야	☐ 데이터신기술 분야 ☐ 비정형데이터 분석 분야	
세부리그 *해당시 체크	☐ 어드밴스드 리그 *정형데이터 분석분야에 한함(선택)	☐ 스타터 리그	
	☐ 지정주제 리그 *빅데이터플랫폼 활용분야에 한함(선택)	☐ 자유주제 리그	
개인/팀여부	☐ 개인 ☐ 팀(총 4 명)	**개인/팀명**	누가 늙어
지도교사명	*스타터 리그에 한함(선택)		
대표ID	kimje1101@naver.com		

결과보고서 작성 안내 사항	
목차 **(예시)**	**Ⅰ. 개요** 1. 배경 2. 목적 및 필요성 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) - **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. 결과 보고서 요약

본 분석은 클래식 공연 활성화를 위한 가격 모델 수립을 목적으로 한다. 좌석 재분배와 관객 특성을 고려한 가격 정책 제안을 위해 데이터를 분석하고, 좌석 별 가격 모델과 관객 특성 별 가격 모델을 제안한다. 좌석 별 가격 모델에서는 장르와 장소에 따른 군집화를 통해 좌석 등급의 재분배를 시도하고, 이에 기반한 가격 예측 모델을 만든다. 관객 특성 별 가격 모델에서는 요일과 시간대에 따른 할인 혜택 정책을 설계하고, 선예매와 개인 정보의 유의성을 파악하여 선예매 할인 정책을 조정한다. 이를 통해 클래식 공연의 고객 만족도, 수익성, 공공성 측면에서 합리적인 가격을 제시한다.

2. 분석 배경

예술의 전당은 복합적인 문화 예술을 행하고 체험할 수 있는 공공성을 지닌 공간이다. 국민 문화 향유 기회 확대와 순수 공연 활성화의 취지를 가진 공공기관인 만큼, 공연 가격 수립에 대한 고려사항이 복합적으로 존재한다.

예술의 전당 공연 가격 선정에 있어서 보완이 필요한 점으로는 좌석 등급의 보편적인 기준이 모호하다는 점과, 객관적인 장르 별 특징이 반영될 필요성이 존재한다. 또한 일부 공연은 좌석 종류의 다양성이 고려되지 않아, 고객들의 좌석 선택 옵션이 부족하고 이러한 단점을 보완해 고객들의 좌석 선택의 다양성을 통한 주최측인 예술의 전당 수익성 증대와 동시에 고객 만족도를 증대 시킬 필요가 있다.

이에, 좌석 등급과 가격에 관하여 공연장과 기획사가 참고할 수 있는 보편적인 기준안을 제공하며, 고객 만족도와 수익성, 공공성을 두루 충족하는 투명하고 효과적인 좌석 재분배 방법 및 가격 제안 모델을 제안하고자 한다.

예술의 전당은 회원제를 통한 예매 시스템을 제공 중이다. 회원 멤버십 종류로는 크게 무료회원과 유료 회원으로 나뉘며, 유료 회원은 골드/블루/그린 세가지의 선택지가 존재하고 무료 회원은 일반/싹틔우미/노블 의 세가지이며 나이를 기준으로 분류된다. 예술의 전당은 멤버십 타입 별로 할인 정책을 시행 중이며, 각각의 멤버십은 예술의 전당에서 제공하는 다양한 할인 및 이벤트 수혜의 대상이다.

분석 및 task 의 대상인 예술의 전당 콘서트홀은 관람석 총 2505 석으로, 여타의 공연장과 비교할 때 압도적인 규모의 공연장이다. 예술의 전당은 현재 최대 5 개로 분류되는 좌석 등급제로 공연 예매를 시행 중이다. 하지만 해외의 유사 기관인 일본의 산토리 홀과 시드니의 오페라 하우스와 비교했을 때, 예술의 전당 5 개의 좌석 분류는 다양성의 측면에서 다소 부족함을 보인다.

이러한 정보들을 기반해서, 예술의 전당 콘서트홀의 효과적인 가격 수립 모델 수립이라는 과제를 수행하기 위한 스토리 라인은 다음과 같다.

기본적으로 주어진 예술의 전당 예매 데이터를 활용, 추가적인 외부 데이터로 공공데이터 포털의 예술의 전당 공연 및 전시 입장객 현황 데이터와 kopis 공연예술통합 전산망의 Open API 를 통한 공연장 별 공연 목록 및 공연 세부사항 데이터를 활용했다.

앞서 말한 데이터들의 탐색적 분석을 통해 고객, 장르, 공연 데이터에 대한 인사이트 도출 및 시각화를 진행했다. 크게 두가지로 분류되는 회원 비회원 여부로 나눠서 분석을 시행했고, 고객 만족도와 공공성, 그리고 수익성의 측면에서 분석과 인사이트를 도출했다.

다음으로는 결측치 처리 및 이상치 처리 등의 Data Cleansing 을 진행했고 이를 기반해 클러스터링을 진행했다. 고객의 좌석 선호도를 반영한 좌석 재분할을 목적으로 진행했으며 두가지 클러스터링 Method 로 진행 및 비교를 통해 최종적으로 사용한 클러스터링 기법은 K-Means Method 이다. 이를 통해 기존 좌석 분류를 넘어, 장르별 좌석 재분류를 제시한다.

최종적으로는 좌석 별 효과적인 가격 예측 모델을 생성하고, 활용하여 군집별 가격 통계치 및 변동성을 기반한 가격 최적화를 수행하고 좌석 별 가격 등급을 제안한다.

3. 목적 및 필요성

본 분석에서는 클래식 공연 활성화를 위한 효과적 가격 모델 수립을 통해 관람객과 공연 주최 측의 관점에서 양득을 할 수 있는 방안을 제시하고자 한다.

위에서 언급했듯 예술의 전당은 복합적인 문화 예술을 행하고 체험할 수 있는 공공기관이기에, 공연의 가격 수립에 고려해야 할 사항이 복합적으로 존재한다.

이에 고객 만족도 및 좌석 선택의 다양성, 공연자 측의 수익성, 논리 있는 근거를 통한 투명한 가격 등 다양한 요인을 반영해야 할 필요가 있다.

따라서 주어진 데이터를 고객 만족도와 공공성의 여부, 수익성 극대화의 측면에서 분석하여 각 관점에서의 특징을 효과적으로 표현하고자 한다.

먼저, 좌석 등급에 의해 결정되는 가격을 보다 효과적으로 제안하기 위하여 좌석의 재분배를 시도하고자 한다.

장르에 따른 좌석 선호도를 반영한 군집화를 바탕으로 좌석 등급을 재분배하고, 이러한 재분배 결과에 기반하여 고객만족도와 수익성, 공공성을 충족할 수 있는 가격 모델을 제안하고자 하는 바이다.

좌석 재분배 결과를 반영한 가격 모델의 완성을 통해 얻을 수 있는 예상 수익 등으로 정책 시행시의

효과를 입증한다.

기본적으로 가격 모델은 수익 창출 극대화를 목표로 하기 때문에 , 훌륭한 성능의 가격 모델이 완성될 시 예술의 전당은 기관이 가지는 수익성의 향상과 더불어 가격 선정 투명성을 확보할 수 있다.

가격 선정 투명성의 확보를 통해 시장 내 경쟁력을 확보하고 이를 기반으로 선도적 서비스를 제시한다면 이는 국민의 문화 향유 기회 확대의 측면에서 장족의 발전을 이룩할 수 있는 가능성이 될 것이다.

4. 분석 수행 범위

가) 문제 데이터 활용

예술의 전당 측에서 제공한 데이터는 본 분석 연구의 전반적으로 활용되며 공연 티켓 구매 데이터로 해당 구매에서의 좌석 정보, 공연의 정보 등을 담고 있다.
본 팀은 해당 데이터에서의 탐색적 데이터 분석(EDA) 과정을 통해 얻어낸 장소적 특성, 고객 특성, 장르 분석, 공연 수요 분석 등을 통해 분석의 방향성을 설정하였다.

나) 외부 데이터 활용

1. 공연예술통합전산망 오픈 API : Beautiful Soup 를 활용해 얻어낸 JSON 형태의 공연 API 정보를 추출하여 추후 가격 변동성 설명에 활용하였다. 포함된 공연 정보는 공연 이름 , 공연 날짜 ,공연 위치 , 공연자 , 공연 시간 , 좌석 등급표 등이 있다.

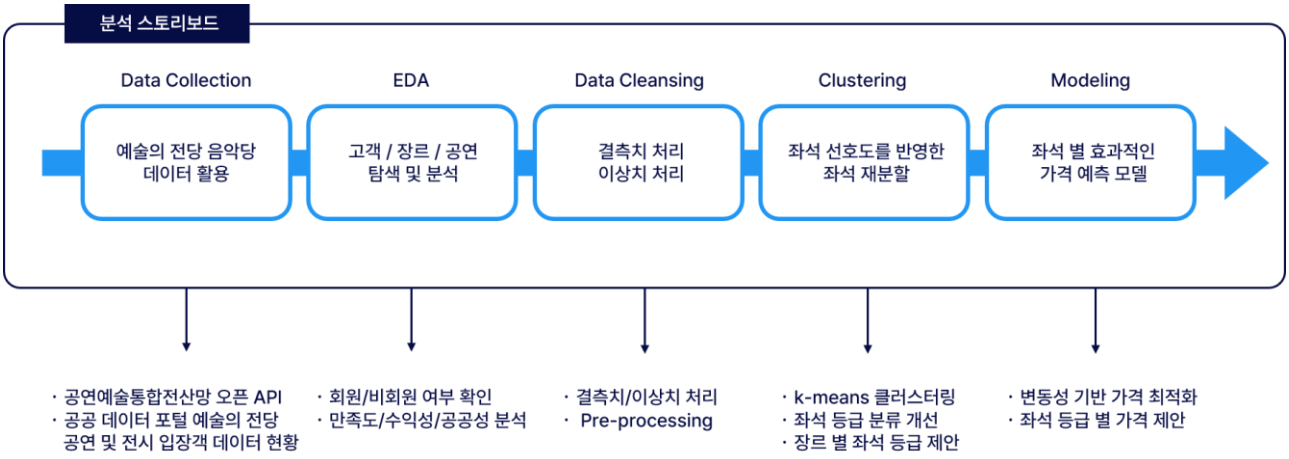
	mt20id	prfnm	prfpdfrom	prfpdto	fcltynm	prfcast	prfcrew	prfruntime	prfage	entrpsnm	pcseguidance	poster
0	PF170622	더멜과 함께하는 크리스마스 지브리 탱고 (2시)	2020.12.25	2020.12.25	예술의전당 [서울] (콘서트홀)	고상지, 최문석, 정솔, 김유성, 윤종수, 박진수, 박용은 등		1시간 40분	만 7세 이상		전석무료	http://www.kopis.or.kr/upload/pfmPoster/PF_PF1...
1	PF170381	KBS교향악단 특별연주회 9 (12.24)	2020.12.24	2020.12.24	예술의전당 [서울] (콘서트홀)	정명훈, 강혜정, 김정미, 박지민, 양준모 등		1시간 10분	만 7세 이상		R석 120,000원, S석 90,000원, A석 50,000원, B석 10,000원	
2	PF170325	KBS교향악단 특별연주회 9 (12.26)	2020.12.26	2020.12.26	예술의전당 [서울] (콘서트홀)	정명훈, 강혜정, 김정미, 박지민, 양준모 등		1시간 10분	만 7세 이상		R석 120,000원, S석 90,000원, A석 50,000원, B석 10,000원	http://www.kopis.or.kr/upload/pfmPoster/PF_PF1...

2. 공공 데이터 포털 예술의 전당 공연 및 전시 입장객 데이터 현황 : 관객수 , 대관 여부 등을 포함하며 추후 '장르' 결측치의 전처리 과정에서 활용하였다.

Ⅱ. 문제 수행 내용

1. 분석(수행) 절차

가. 분석스토리보드



나. 분석 세부 절차

1) 데이터 탐색
   - 탐색적 데이터 분석

2) 데이터 전처리
   - 결측치 및 이상치 처리

3) 좌석 클러스터링
   - 클러스터링 기법 실험
   - 클러스터링 분석 및 시각화
   - 좌석 등급 분류 개선
   - 장르 별 좌석 등급 기준 제안

4) 가격 모델링
   - 타겟 변수 설정
   - 피쳐 엔지니어링
   - 머신러닝/딥러닝 모델링
   - 모델 피쳐 중요도 시각화
   - 변동성 기반 가격 최적화
   - 좌석 등급 별 가격 기준 제안

5) 아이디어 활용
   - 결과 활용 아이디어 제안

2. 분석(수행) 내용 및 결과

Ⅰ. 개요

1. 결과 보고서 요약

2. 분석 배경

2. 목적 및 필요성

3. 분석 수행 범위

가) 문제 데이터 활용

예술의 전당 측에서 제공한 데이터는 본 분석 연구의 전반적으로 활용되며 공연 티켓 구매 데이터로 해당 구매에서의 좌석 정보, 공연의 정보 등을 담고 있다.
본 팀은 해당 데이터에서의 탐색적 데이터 분석(EDA) 과정을 통해 얻어낸 장소적 특성, 고객 특성, 장르 분석, 공연 수요 분석 등을 통해 분석의 방향성을 설정하였다.

나) 외부 데이터 활용

1. 공연예술통합전산망 오픈 API : Beautiful Soup 를 활용해 얻어낸 JSON 형태의 공연 API 정보를 추출하여 추후 가격 변동성 설명에 활용하였다. 포함된 공연 정보는 공연 이름 , 공연 날짜 ,공연 위치 , 공연자 , 공연 시간 , 좌석 등급표 등이 있다.

	mt20id	prfnm	prfpdfrom	prfpdto	fcltynm	prfcast	prfcrew	prfruntime	prfage	entrpsnm	pcseguidance	poster
0	PF170622	더벨과 함께하는 크리스마스 지브리 탱고 (2시)	2020.12.25	2020.12.25	예술의전당 [서울] (콘서트홀)	고상지, 최문석, 정솔, 김유성, 윤종수, 박진수, 박용은 등		1시간 40분	만 7세 이상		전석무료	http://www.kopis.or.kr/upload/pfmPoster/PF_PF1...
1	PF170381	KBS교향악단 특별연주회 9 (12.24)	2020.12.24	2020.12.24	예술의전당 [서울] (콘서트홀)	정명훈, 강혜정, 김정미, 박지민, 양준모 등		1시간 10분	만 7세 이상		R석 120,000원, S석 90,000원, A석 50,000원, B석 10,000원	
2	PF170325	KBS교향악단 특별연주회 9 (12.26)	2020.12.26	2020.12.26	예술의전당 [서울] (콘서트홀)	정명훈, 강혜정, 김정미, 박지민, 양준모 등		1시간 10분	만 7세 이상		R석 120,000원, S석 90,000원, A석 50,000원, B석 10,000원	http://www.kopis.or.kr/upload/pfmPoster/PF_PF1...

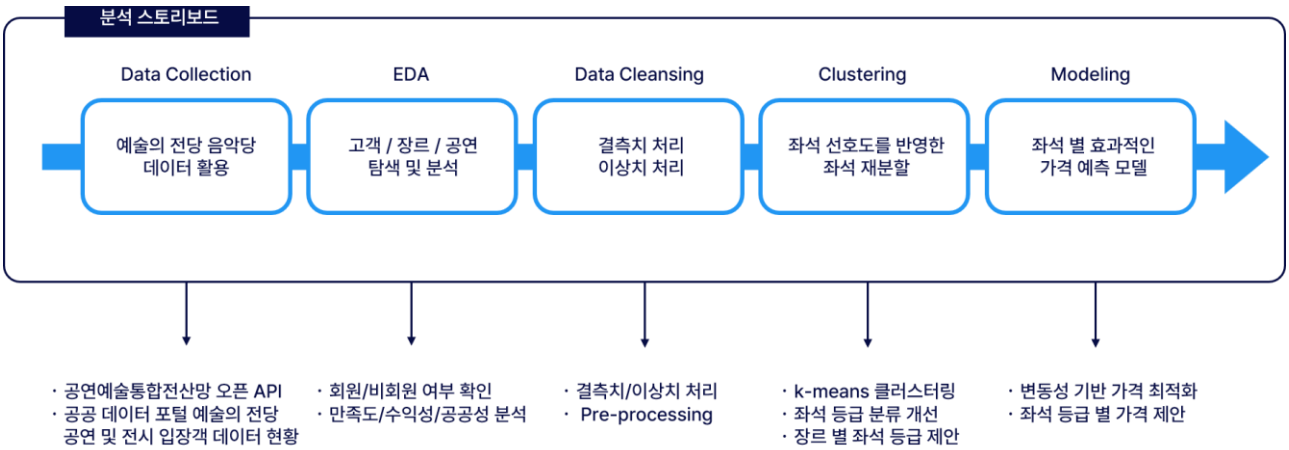
2. 공공 데이터 포털 예술의 전당 공연 및 전시 입장객 데이터 현황 : 관객수 , 대관 여부 등을 포함하며 추후 '장르' 결측치의 전처리 과정에서 활용하였다.

	공간명	작품명	기획대관구분	시작일	종료일	합계
818	예술의전당 콘서트홀	The Voice of DUO 듀오 콘서트 신년음악회	대관	2018-01-02	2018-01-02	848
819	예술의전당 콘서트홀	2018년 새해인사 나눔음악회	대관	2018-01-03	2018-01-03	1391
820	예술의전당 콘서트홀	서울시향 2018 신년음악회	대관	2018-01-07	2018-01-07	2308
821	예술의전당 콘서트홀	2018 신년음악회	기획	2018-01-09	2018-01-09	1916
822	예술의전당 콘서트홀	조성진 피아노 리사이틀 (1.10)	대관	2018-01-10	2018-01-10	2404

Ⅱ. 문제 수행 내용

1. 분석(수행) 절차

가) 분석스토리보드



나) 분석 세부 절차

1) 데이터 탐색
- 탐색적 데이터 분석

2) 데이터 전처리
- 결측치 및 이상치 처리

3) 좌석 클러스터링
- 클러스터링 기법 실험
- 클러스터링 분석 및 시각화
- 좌석 등급 분류 개선
- 장르 별 좌석 등급 기준 제안

4) 가격 모델링
- 타겟 변수 설정
- 피쳐 엔지니어링
- 머신러닝/딥러닝 모델링
- 모델 피쳐 중요도 시각화
- 변동성 기반 가격 최적화
- 좌석 등급 별 가격 기준 제안

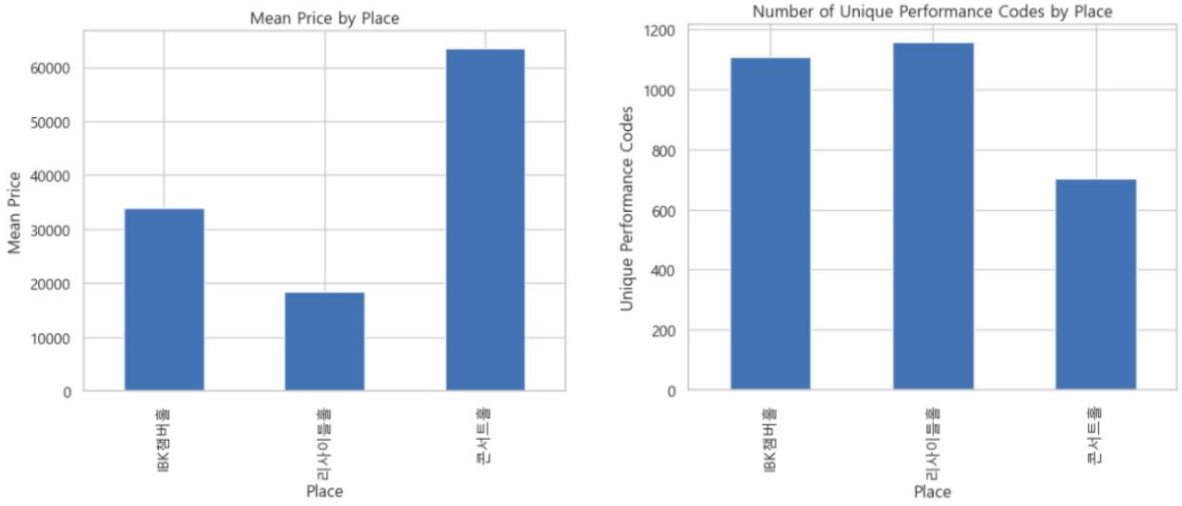
5) 아이디어 활용
- 결과 활용 아이디어 제안

2. 분석(수행) 내용 및 결과

가) EDA (탐색적 데이터 분석)

1. 장소적 특성 예술의 전당 안에서 분석의 주 대상이 되는 콘서트 홀이 가지는 장소적 특성을

파악하였다. 데이터에 포함된 공연장소는 총 3곳 (콘서트 홀, IBK챔버홀 , 리사이틀홀) 이었으며 장소적 특성 분석은 콘서트홀이 다른 두 공연장소에 비해 가지는 차이를 위주로 실시하였다.



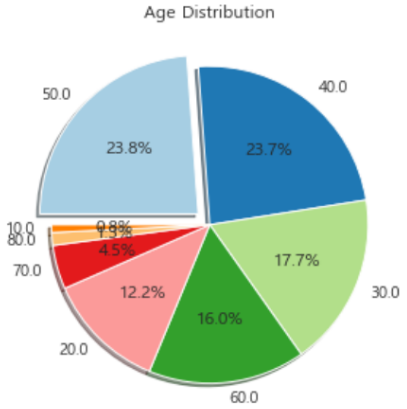
좌측 plot 은 해당 데이터 프레임을 place 별로 groupby 하여 각각의 price 값의 평균으로 집계한 것이다. 이를 통해 콘서트 홀에서 실시하는 공연이 다른 공연장에서 실시하는 공연보다 높은 가격으로 진행된다는 점을 알 수 있다.

우측 Plot 은 해당 데이터 프레임을 place 별로 groupby 하여 각각의 genre 의 unique 한 값의 수를 시각화 한 것이다. 이를 통해 콘서트 홀에서 실시하는 공연이 다른 공연장에서 실시하는 공연들과 장르의 다양성은 비슷하다는 것을 알 수 있다.

위 시각화 결과를 통해, 콘서트홀은 IBK챔버홀의 2배, 리사이틀홀의 3배 정도로 월등히 높게 평균 가격이 형성되어 있지만 콘서트 홀이 타 공연장에 비해서 가지는 공간적 이점 (크기 , 수용인원 등) 에 비해 장르의 다양성은 크게 다르지 않다는 것을 알 수 있다.

2. 고객 특성

문제 데이터를 통해 회원/비회원 고객들에 대한 특성을 파악하였다. 고객들의 인구통계학적 특성을 활용하기 위해서는 age, gender 등의 추가적인 정보들이 필요하다. 그러나 비회원 데이터의 경우에는 age, gender 가 결측치로 채워져있었기에 회원 데이터에 한하여 고객 특성 분석을 실시하였다.



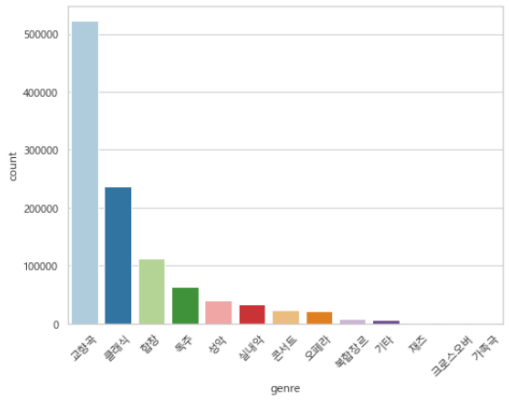
좌측 그림은 전반적인 나이 분포를 보여주는 pieplot 이다. 예술의 전당의 주 수요를 담당하는 고객층은 40대 50대임을 파악할 수 있다. 우측 그림은 성별 분포를 보여주는 barplot 이다. 예술의 전당은 남성에 비해 여성 고객의 수가 더 많음을 확인할 수 있었다.

추가적으로 성별과 나이대를 결합하여 분석을 진행해본 결과 여성의 경우 40대(0.28)-50대-30대- 60대(0.14) 순으로, 남성의 경우 50대(0.24)-60대-30대- 40대(0.15) 순으로 큰 비중을 차지하고 있음을 확인할 수 있었다.

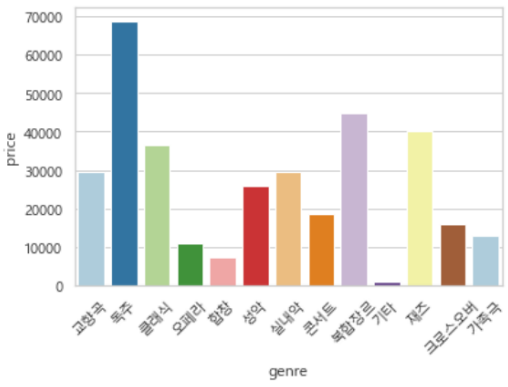
비회원 데이터의 경우에는 예술의 전당측에서 배포하는 여러 형태의 초대권을 통해서 회원가입을 하지 않고 공연을 관람하러 오는 경우가 많음을 파악할 수 있었다.

3. 장르 분석

문제 데이터의 공연 정보 중, 중요하다고 판단한 장르를 중점으로 탐색적 데이터 분석을 실시했다. 장르별로 주로 얻고자 했던 정보는 '어떤 장르의 빈도가 가장 높은가?' 와 '어떤 장르의 공연 가격이 높게 형성 되어있는가?' 였다.



장르 별 공연 횟수



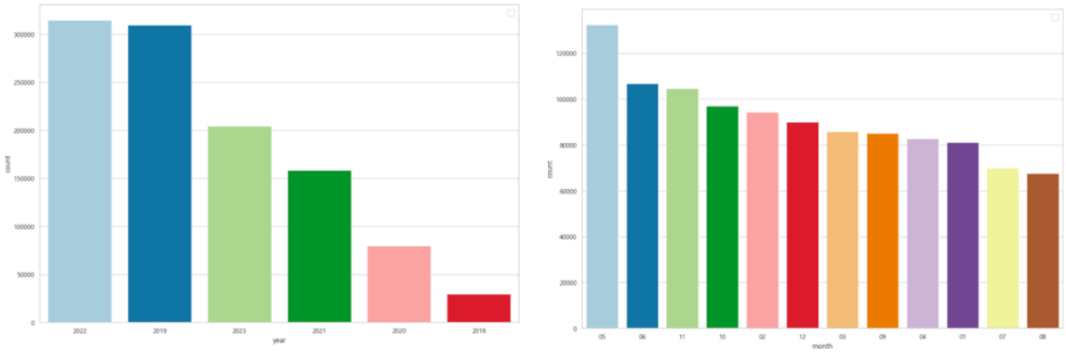
장르 별 평균 가격

좌측 plot 은 장르 별 공연 횟수를 보여주는 시각화 자료로 공연횟수가 교향곡 - 독주 - 실내악 - 클래식 순으로 많다는 것을 보여준다.

우측 plot 은 장르 별 평균 가격을 보여주는 시각화 자료로 독주 - 복합장르 - 재즈 - 클래식 순으로 공연의 평균 가격이 높게 형성되어있다는 것을 보여준다.

4. 공연 수요 분석

문제 데이터는 기본적으로 시간이 기록되는 판매 데이터이기 때문에 시간에 따른 예술의 전당의 수요분석해낼 수 있을 것이라 판단, 관련한 분석을 진행했다.



좌측 Plot 은 연별 수요 , 우측 Plot 은 월별 수요를 시각화한 Plot 이다.

예술의 전당은 코로나가 발생했던 2020년부터 2021년까지 수요가 급격하게 감소하였으며 2022년 하반기에 코로나가 완화되면서 다시금 수요가 증가하는 모습을 보였다.

월 별 수요 분석에 따르면 3월 , 5월 , 6월 , 10월 , 11월의 수요가 가장 높은 것으로 미루어 보아 늦봄~초여름 / 늦가을에 공연 관람을 하는 비중이 높음을 볼 수 있다.

5. 고객 만족도, 수익성, 공공성 분석 결과

위 탐색적데이터 분석 과정을 기반으로 본 팀의 분석 목표였던 고객 만족도, 수익성, 공공성의 측면에서 문제 데이터를 해석하는 과정또한 진행하였다.

해당 성질의 측면에서 데이터를 해석하다는 것은 다시 말해 데이터 속 수치들을 활용하여 해당 성질을 정량화 시키고 여러 측면의 분석을 시행하는 것이라고 할 수 있다.

5-1. 고객만족도

본 팀이 생각하기에 고객 만족도를 대변할 수 있는 정량화된 수치는 '티켓 취소 여부' ,' 재관람율 '등이라고 판단, 문제 데이터의 'ticket_cancel' ,' discount_type' 칼럼에서 정보를 뽑아 수치화 시켰다.

분석 결과 약 79%는 티켓 취소 없이 공연을 관람하였고 재관람 비율은 전체 데이터의 약 0.0005 % 로

판단할 수 있었다.

```
#장르별 공연 평균 수익
df.groupby('genre')['price'].mean().sort_values(ascending=False)
```

```
genre
독주          68748.972489
복합장르       44741.298105
재즈          39985.185185
클래식         36377.315733
실내악         29506.021272
교향곡         29456.259298
성악          25997.994305
콘서트         18621.665691
크로스오버       16018.614271
가족극         12838.709677
오페라         10955.912839
합창           7143.690392
기타           1032.175252
Name: price, dtype: float64
```

5-2. 수익성

탐색적 데이터 분석 과정에서 장르 별 평균 가격대의 편차가 매우 명확한 것을 확인하였다. 이를 통해 본 팀은 장르가 티켓 가격 형성에 영향력을 행사하고 있다고 판단하고 추후 분석 목표가 되는 '가격 형성을 통한 수익성 창출'에 장르가 지대한 영향을 끼칠 것으로 예상하였다.

5-3. 공공성

문제 데이터의 discount_type column 에 따라 취약계층을 정의하였다.
(65세|실버|유족|학생|경로|장애|노블|의사상|임산|청소년|청년|유공|어린이|의상) 이 포함된 discount_type을 취약 계층으로 정의 후, 공연별로 취약 계층 비율을 계산하는 방법으로 정의하였고, 공연 별 취약 계층의 비율을 분석해본 결과 공연 당 0.03% 정도만이 공공성을 띈다는 결과를 얻을 수 있었다.

위 과정을 통해 분석 목표를 기반으로 데이터를 분석하고 보다 구체화된 분석 목표를 설립할 수 있었다. 고객만족도, 수익성, 공공성의 균형점을 모색하며 기존 가격안에 대한 수익성을 분석한 결과 '장르'에 따른 선호도 기반 좌석 분류 및 가격 모델 설립이 수익성 증대에 효과적일 것으로 보였다. 이에 팀의 분석 목표를 장르별 선호도를 반영한 좌석 분류 및 가격 모델 설립으로 구체화하였다.

나) 데이터 전처리(이상치, 결측치)

1. 결측치

본 데이터의 컬럼은 총 24개로 구성되어 있으며, 결측치의 경우 총 11개로 age, gender, membership과 관련된 6가지 컬럼, pre_open_date, open_date, genre의 컬럼에 대한 결측치 처리를 진행했다. age와 gender의 경우, 결측치의 비율이 50%가 넘어 사용하지 못한다고 판단, 따라서 해당 컬럼들을 삭제해주었다.

- pre_open_date(선예매시작일)

pre _open_date에 대한 결측치 비율은 45%이었으며, 해당 pre_open_date(선예매시작일)은 기획자 및 대관자의 선택에 의해 결정된다는 예술의전당 정보를 바탕으로 하였다. 추가적으로 공연의 코드가 같음에도 선예매시작일이 존재하는 경우와 존재하지 않은 경우가 존재함을 파악한 결과, pre_open_date가 NaN 값인 경우, 선예매가 열리지 않은 것으로 판단, 따라서 결측치를 0으로 채워주었다.

- open_date(예매시작일)

Open_date에 대한 결측치 비율은 0.02%이었으며 결측치에 대한 특성을 추출한 결과, performance_code가 479라는 통일된 결과를 얻을 수 있었다. Performance_code가 479인 데이터를 파악한 결과, pre_open_date(선예매시작일)는 NaN 값이었으나, tran_date(예매 거래일자)의 값은 존재함을 파악하였다. 이때 선예매는 열리지 않았다고 가정하였기에, 구매일자가 가장 빠른 날인 '20221129'의 값을 예매 시작일자로 대체해주었다.

- genre(장르)

genre에 대한 결측치 비율은 1.35%이었으며 결측치 처리는 크게 두 가지로 나누어 진행하였다. 1. 해당 performance_code의 genre가 일부만 존재하는 경우, 2. 해당 performance_code의 genre가 전부 비어 있는 경우로 나눌 수 있다.

1. performance_code가 229, 2215인 경우, 같은 공연 코드임에도 불구하고 일부의 데이터에만 gerne가 존재함을 파악하였다. 이에 따라 229의 일부 장르에 대한 결측치는 존재하는 데이터를 활용해 '교향곡'으로 대체했으며, 2215의 경우, '클래식' 장르로 대체해주었다.

2. 해당 공연코드의 장르가 전부 비어 있는 경우, '공연예술통합전산망'의 예술의전당_공연 및 전시 입장객 현황_20220728의 추가 데이터를 활용해 장르를 채워주었다.

- membership_type_2,3,4,5,6(멤버십 타입)

membership_type에 대한 결측치 비율은 각각 75%, 91%, 98%, 99.9%, 100%이었으며, membership_type_6을 제외한 나머지 membership_type의 경우에는 일부 정보에서 유의미한 결과가 존재한다고 판단하여 삭제하지 않고 0으로 채워주었다. 반면 membership_type_6의 경우, 결측치가 100%임에 따라 본 컬럼을 삭제해주었다.

2. 기타 전처리

member_yn == N 중, discount_type이 membership_type에 해당하는 경우, member_yn을 Y로 변경하였다. discount_type에 membership_type에 대한 정보가 존재하므로 일부 membership_type_1에 관련 정보를 채워주었다.

3. 이상치

· running_time(러닝타임) & intermission(휴게시간)

일부 정보만 존재하는 552, 286, 1750, 1491, 2495의 경우, 존재하는 정보로 대체하였다. running_time과 intermission이 0인 경우, 해당 장르의 평균값으로 대체하였다.

· ' [초대권] ' → '초대권' 통일

다) 좌석 클러스터링

1) 클러스터링 사용 목적 - 좌석 선호도 반영

좌석 등급에 의해 결정되는 가격을 보다 효과적으로 제안하기 위하여 좌석의 재분배를 시도하고자 한다. 공연 장르 및 장소의 특징과 고객의 좌석에 대한 선호도를 반영한 군집화를 바탕으로 좌석 등급을 재분배 하고, 이러한 재분배 결과에 기반하여 고객만족도와 수익성, 공공성을 충족할 수 있는 가격 모델을 만들기 위함이다.

클러스터링은 FINCH Method 와 K-Means Method 두가지 방법을 시도했고, 이 중 더 낫다고 판단된 K-Means 클러스터링을 최종적으로 채택해 사용했다. K-Means 가 더 낫다고 판단한 근거는 뒤에서 설명하도록 하겠다.

클러스터링에서 클러스터를 나누는 기준이 되는 특성들은 다음과 같다.

1. 기존 데이터의 seat 정보 4 개 (층, 블록(박스), 열, 좌석 위치)

2. 기존데이터의 파생 변수 2 개 (sell_delta, polularity)

이 중 sell_delta 는 기존 데이터에서 tran_date – open_date 의 연산으로 만든 특성이다. 이를 통해 sell_delta 에 해당 좌석이 얼마나 빨리 팔리는 지 알 수 있고, 이는 time 측면에서 해당 좌석의 인기 및 고객의 선호도를 반영한다.

다음으로 popularity 는 seat_popularity + pre_popularity 의 연산으로 생성한 특성이다. seat_popularity 는 같은 장르 내에서, 해당 좌석의 빈도를 나타낸다. pre_popularity 는 같은 장르 내에서, 선예매로 좌석을 구매한 사람들에 대해, 해당 좌석의 빈도를 나타낸다. 선예매로 좌석을 구매한 것에 대한 가중치를 주기

위함이고, 위 두가지는 해당 좌석에 대한 고객의 선호도를 반영한다. 최종적으로 클러스터링에 사용한 popularity 는 위 두가지의 특성을 더함으로 생성했다. 따라서 최종적으로 좌석 클러스터링에 사용한 변수는 위 6 가지이다.

2) 클러스터링 기법 실험

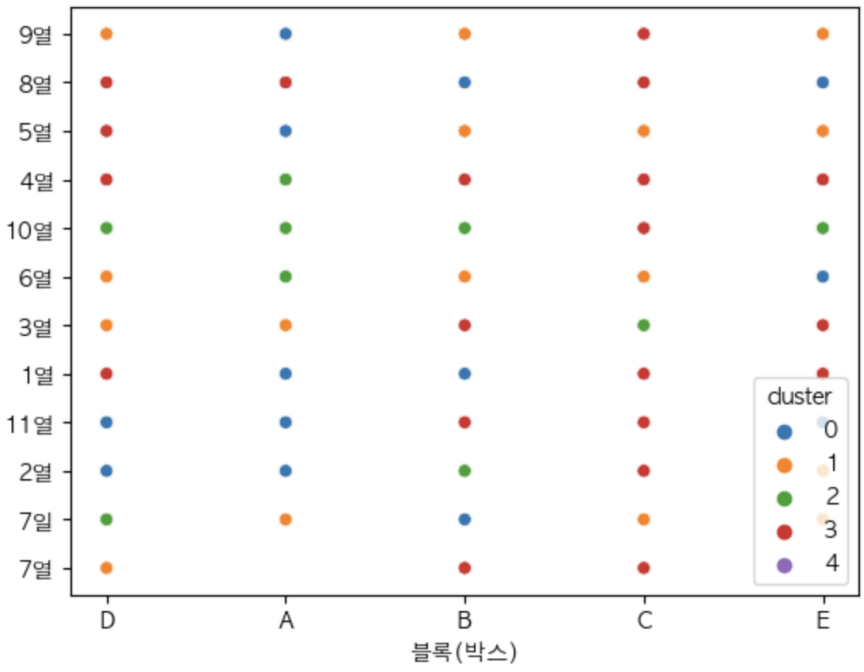
다음으로는 사용한 클러스터링 메소드에 대한 설명이다. 메소드는 FINCH 와 K-Means Method 를 사용했다. 군집화에 사용한 장르는 총 8 가지로, 가장 데이터가 많은 교향곡부터 클래식, 콘서트, 독주, 성악, 오페라, 합창, 실내악이며, 사용하지 않은 장르는 데이터 수가 일반화를 시킬만큼 충분하지 않다고 판단해 클러스터링에서는 제외했다. 군집화는 장르별로 따로 데이터프레임을 만들어 진행했다.

FINCH (First Integer Neighbor Clustering Hierarchy) Method 는 비모수적 클러스터링 방법 중 계층적 클러스터링 기법에 해당한다.

현재 팀에서 다루는 데이터의 크기가 큰 점, 분포를 쉽게 가정할 수 없다는 점, 하이퍼 파라미터의 튜닝 시에 사용자의 주관이 들어갈 수 있다는 점을 고루 고려하여 볼 때, 완벽한 비지도 학습으로 진행되며 계층적 병합 군집 분석으로 대규모 데이터에 적용했을 때에도 성능적인 하락이 없는 것으로 알려져 있는 FINCH Method 를 최초 선정하였다.

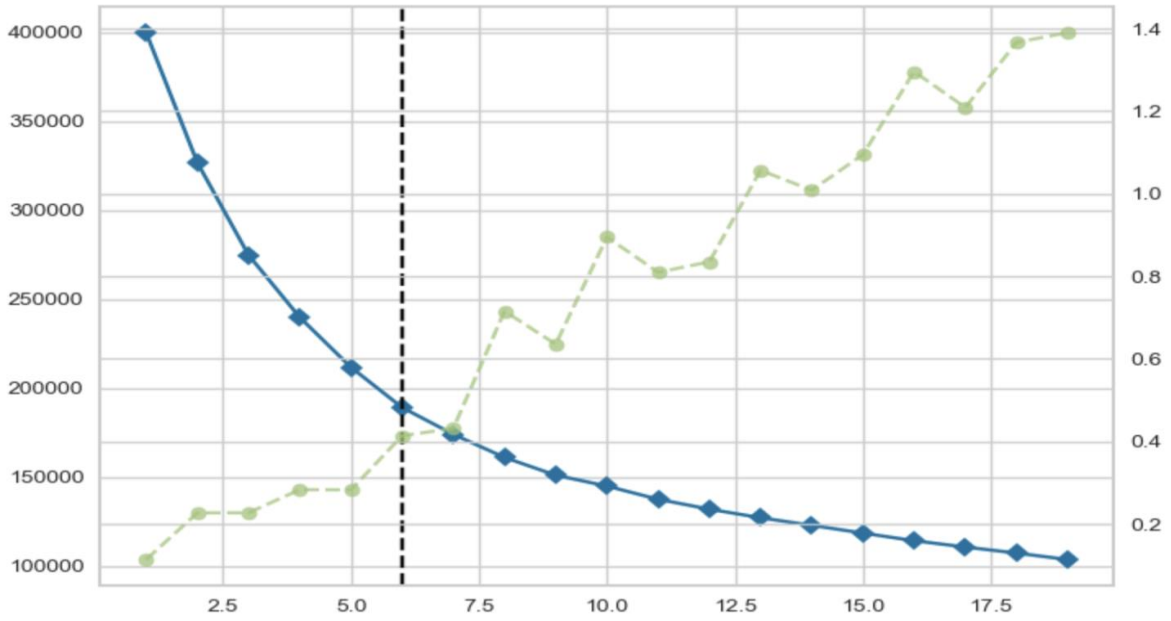
각 장르에 대해 6C2 의 모든 경우의 수로 군집화를 시도했고, 군집 별 가격 특성이 직관적으로 설명이 쉬운 경우와 FINCH method 에서 제공하는 성능지표를 기반으로 장르 별 각 실험의 결과 중 가장 좋은 군집화가 나타난 경우를 택했다.

군집화에 사용된 특성들에 대해 FINCH 알고리즘의 군집이 잘 군집화를 이뤘지만, 좌석 재분배에 사용할 블록(박스), 층 및 열 등의 정보에서는 특정한 패턴을 찾을 수가 없었다.



다음으로 사용한 K-means 는 클러스터의 개수인 n 을 사용자가 지정해야 한다는 한계점을 가진다. 하지만 대용량 데이터에 사용하기 적합하며, 직관적이고 구현이 쉽다는 장점이 있다. 이러한 장점을 기반으로 K-Means Method 를 군집화 Method 의 후보로 선정해 사용했다.

사용자가 군집의 개수를 직접 설정해줘야 한다는 한계점을 보완하기 위해서, 더 객관적인 군집 개수 설정에 도움을 주는 yellowbricks 함수를 이용해 군집의 수를 정했다.

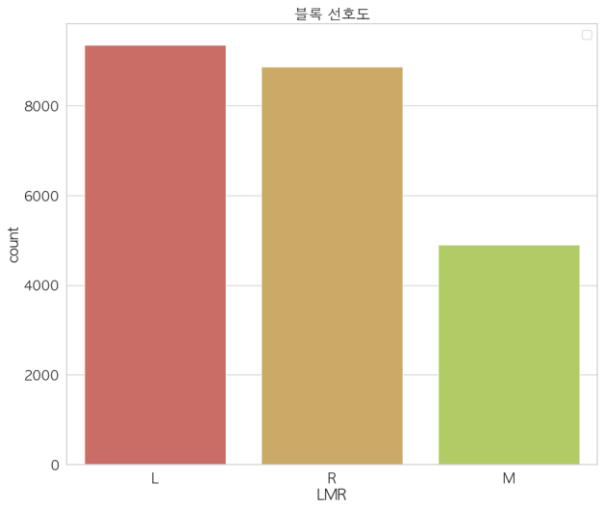


다음은 yellowbricks 를 이용한 군집화의 예시이다. 파란색 그래프는 각 데이터들의 군집 중심과의 평균 거리를 의미하고, 초록 점 그래프는 학습 시간을 나타낸다. 두 그래프를 통해, 최적의 군집 개수를 정해준다. 위의 예시에선 군집 개수를 6 으로 설정했다. 군집화 실행 시에 seed 를 고정해, 최적의 군집 수를 찾고 다시 군집화를 시행했을 때도 동일한 결과가 나오도록 했다. 외에도 사용한 파라미터로는, init: 초기에 군집 중심점의 좌표를 설정할 방식을 말하며 보통은 임의로 중심을 설정하지 않고 일반적으로 k-means++방식으로 최초 설정한다. 하나의 무작위 중심점을 배정하고 두번째 중심점을 첫번째 중심점과 멀리 떨어진 곳에 배치, 그리고 다음 중심점을 1,2 번째 중심점과 멀리 떨어지도록 배치하는 방식으로, 중심점 선정에 있어서 더 신중할 수 있는 파라미터이다.

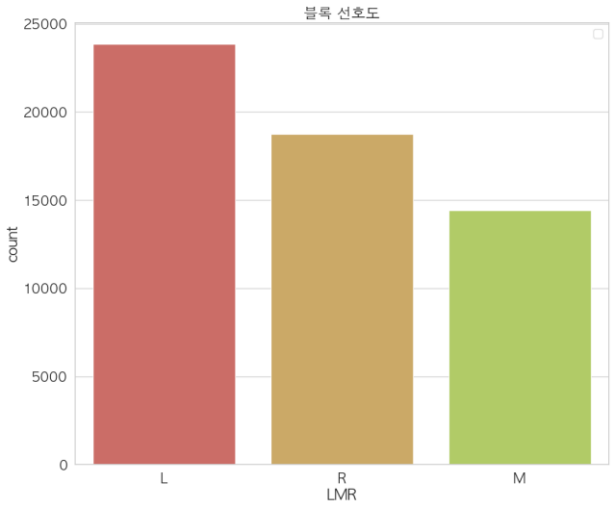
K-Means 에 사용한 특성은 앞서 말한 기존 데이터셋의 seat 정보 4 개에 2 개를 생성해 총 6 개로 진행했다. 군집화 전 StandardScaling 을 진행했고, 군집을 1 부터 20 까지로 설정해 사용했고, 이 중 yellowbricks 함수에 따른 elbow method 로 군집의 개수를 정했다.

장르별 최적의 군집 수는 6~7 개로 비슷한 수로 수렴했다. 클러스터의 결과는 기존 데이터프레임에 열로 추가해, groupby 를 이용해 클러스터 별 가격의 분포를 보고자 했다. 그리고 이를 이용해 좌석 재분배 시, 분배된 각 군집의 적정 가격을 제안하는 모델링의 target 값으로 사용했다. Target 값으로 만든 적정가에 대한 자세한 내용은 뒤에서 설명하겠다.

3) 좌석 클러스터링 시각화



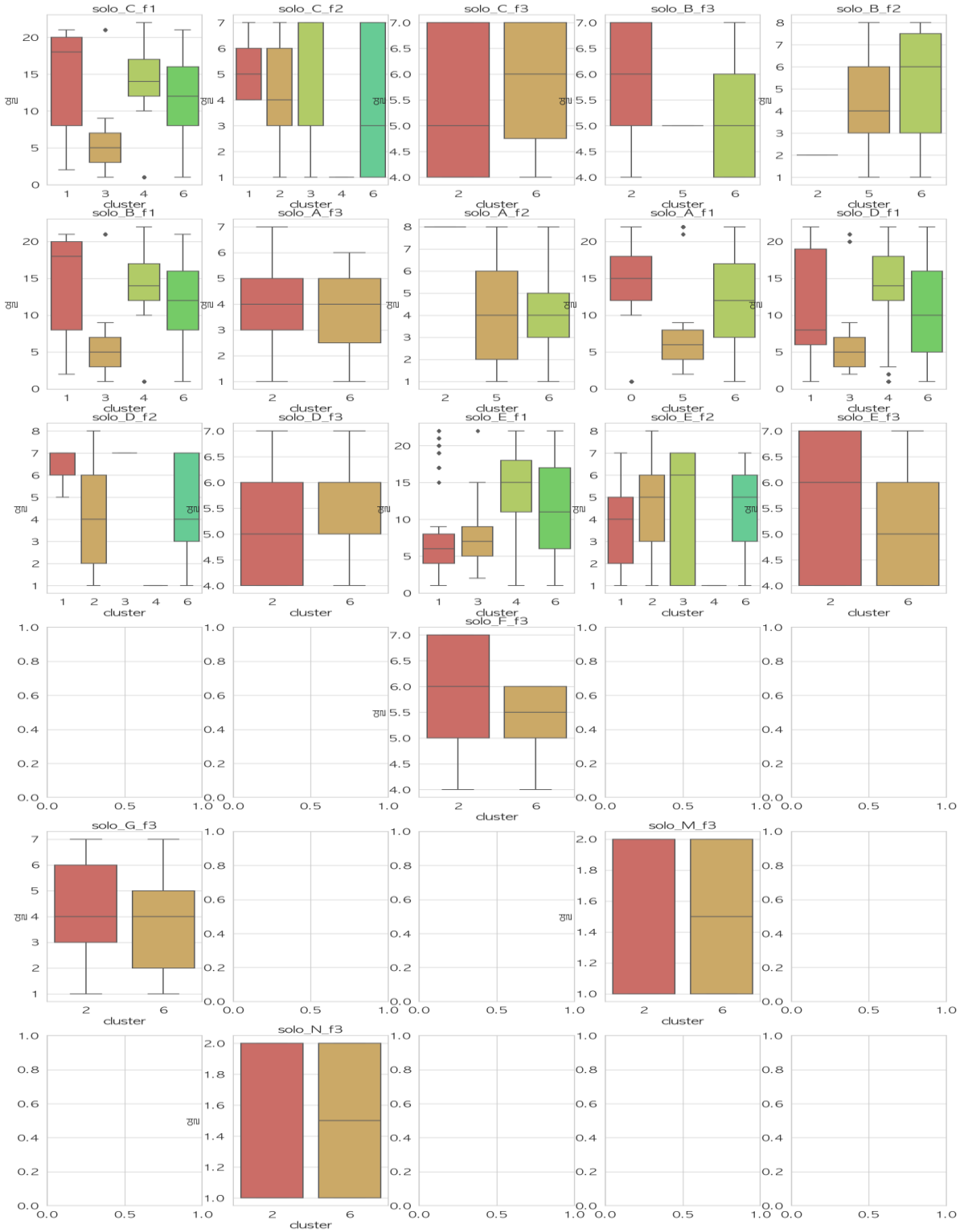
장르별 군집 분석에 앞서, 장르별 블록에 대해 패턴을 파악하고자 했다. M 은 각 층별 중앙 블록인 1,2 층에선 C 블록, 3 층에선 D 블록에 해당한다. L 은 M 에 대하여 왼쪽 모두의 블록들이고, R 은 M 에 대하여 오른쪽 모두의 블록들이다. 대부분의 장르가 왼쪽의 빈도가 더 높았고 오른쪽은 왼쪽보다 항상 조금 낮은 빈도로 나타났다. M 은 가운데에 해당하는 블록 한개에 대한 빈도이므로 L,R 보다 빈도가 다소 높게 나타났다.



위 그래프는 독주 장르에 대한 시각화 그래프이다. 다른 장르들에 비해 L 과 R 의 빈도 차이가 다소 높게 나타났다. 독주는 연주자가 혼자 또는 소수인 경우가 많고, 따라서 시청각적 요소를 고려할 때 관객들은 앞좌석에 앉는 것을 선호할 것이다. 또한 피아노 독주의 경우, 연주자의 손은 많은 관객들의 볼거리일 것이다. 물론 오른쪽에서 피아노의 소리에 집중하는 관객도 다수 존재할 것이나, 위 그래프는 독주 관객들의 경우, 왼쪽 좌석을 선호한다는 가정의 근거가 된다.

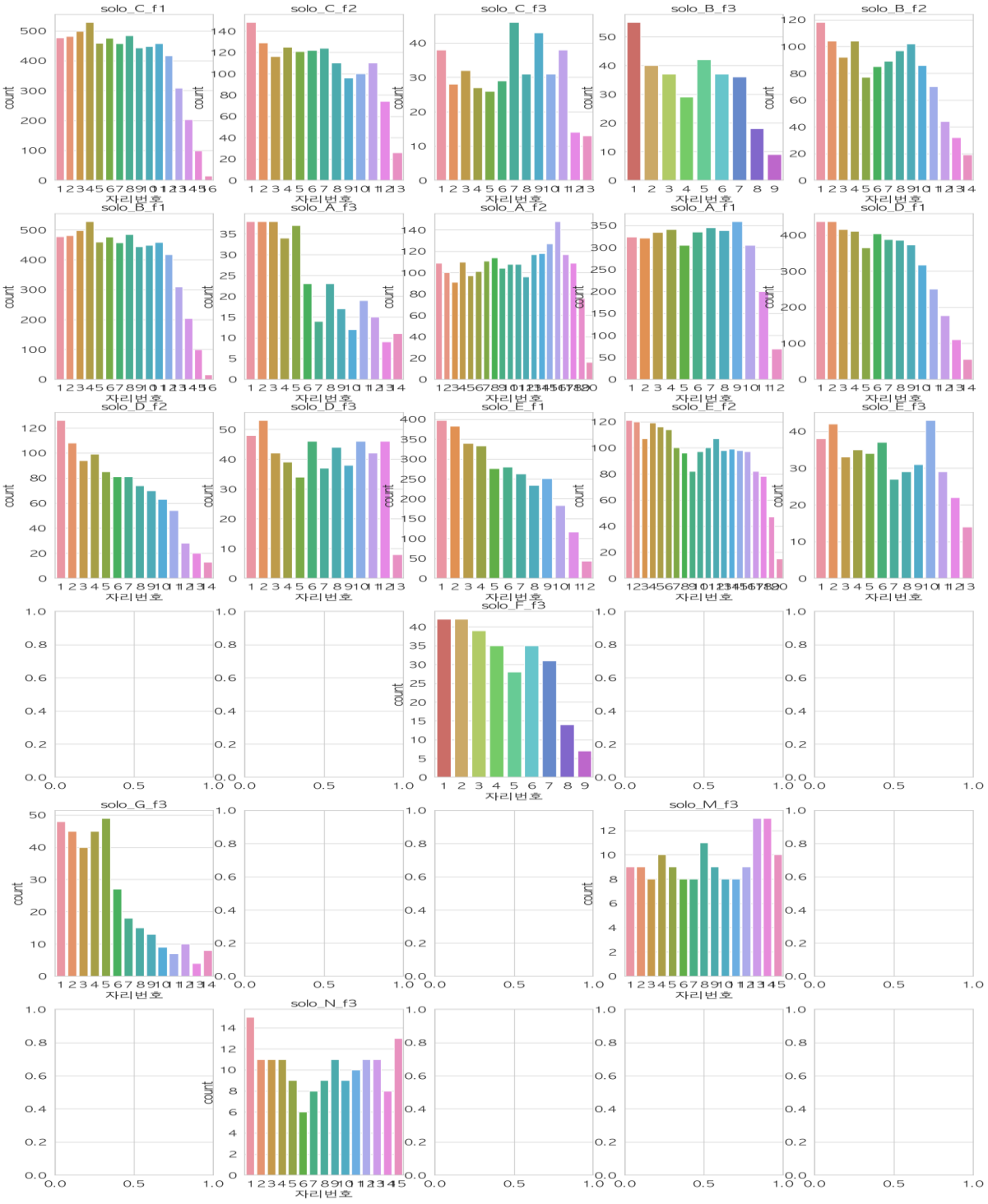
다음으로는 FINCH 클러스터링과 마찬가지로 장르별 군집에 대해, 여러 방향으로 EDA 를 진행했다. 클러스터 열을 추가한 장르별 데이터 프레임을 층 별, 블록(박스) 별로 분할해서 클러스터 별 특징을 분석했다.

다음은 예시인 합창 장르의 층 별 박스 별 각 군집의 좌석 열에 대한 분포를 boxplox 으로 나타낸 것이다. 모든 블록(박스)를 모든 층에 대해 반복문으로 그래프를 그린 것이라 층과 블록의 조합이 존재하지 않을 경우에는 그려지지 않았다.



군집 boxplot 의 1 분위 4 분위 지점과 평균 및 median 값을 고려해 열 별 특이점을 설정할 수 있을 것으로 보인다.

다음은 같은 합창 장르의 층별 장르별, 좌석번호를 왼쪽부터 나열해 시각화로 나타낸 것이다. 기존 콘서트홀의 좌석 정보와 대입해서 볼 때, 블록 별 좌석정보의 변화가 크게 나타나는 곳을 좌석 재분배의 기준점으로 가정했다.



위 두 그래프를 살펴볼 때, K-Means 클러스터링의 결과가 열, 좌석위치 정보에서 좌석 재분배가 더 설명력있다고 판단해서 K-Means 클러스터링의 결과를 최종 task 로 사용했다.

4) 장르 별 좌석 등급 기준 제안

시각화 내용을 바탕으로 기존 콘서트홀 좌석에 대해 Mapping 을 진행했다.

Mapping 에 앞서 공통적 고려해야할 가정으로 다음을 설정했다.

1. 층별 박스별로 데이터를 나누고 군집 별로 가격 순위를 보면서, 가격이 높은 군집의 평균(median)이나 1,3 분위점을 기준으로 좌석 그룹을 나눈다.

2. 한 열만 남기는 것은 지양하고 두열 이상이 묶이는 것을 기준으로 한다.

3. 기존 장르별 좌석 방식과, EDA 를 통해 파악할 수 있는 시청각 요소 등을 고려한다. 각 장르별 현재 진행 중인 공연 중, 좌석 등급이 5 개(최대치)로 나뉜 공연 들의 표본들을 뽑아 엑셀 파일에 대입하고, 이를 토대로 수정을 진행한다.

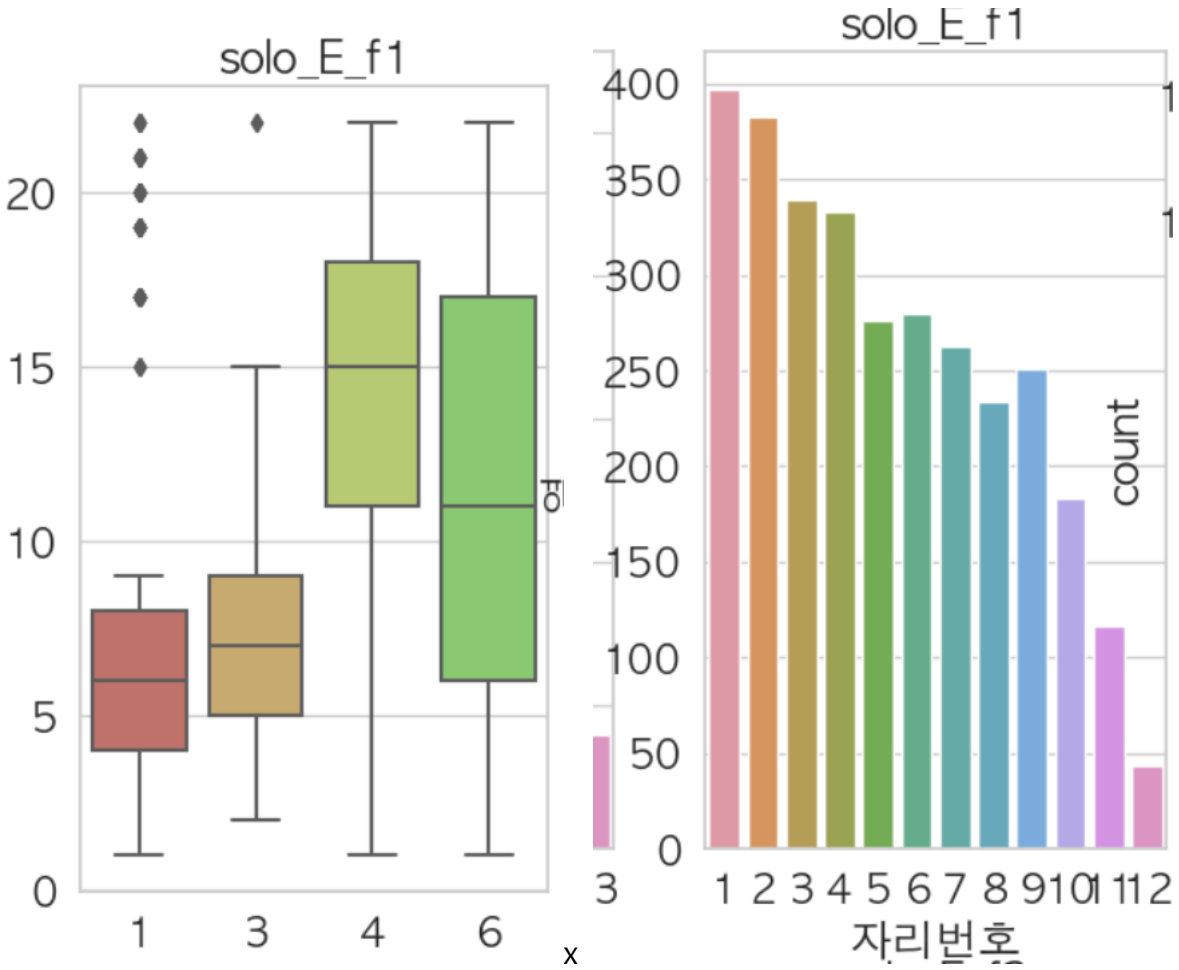
4. 층별 블록별 나눈 데이터프레임에 대해서, 군집 별 평균 price 에 대한 통계량을 고려한다.

첫 번째 가정에 대한 예시로, 4 군집에 대해서 3 군집 이상이 공통적인 부분에는 가격이 비싸던 싸던 수요가 높을 것이다. 3 군집에 대해서 2 군집 이상이 겹친 공통적인 부분에서는, 가격이 비싸던 싸던 수요가 높을 것이다. 이렇게 겹친 부분의 1 칸 안팎을 기준으로 한다면, 그 한 칸을 기준으로 가격에 따라서 관객들이 나뉠 것이다.

이러한 경우가 없을 경우에, 군집 별 특이점으로 보이는 곳을 기준으로 한다. 특이점으로는 군집 여러개의 평균 또는 median 값이 같은 지점, 한 군집과 다른 군집이 만나는 지점을 기준으로 나눈다. 군집이 두개만 나타나는 곳에서는, 군집 두개가 만나는 지점 안팎으로 기준을 정함. 그 부분이 군집 과 가격에 상관없이 좌석을 선택할 지에 대한 기준이 되는 지점일 것이다.

군집이 한개만 나타나는 지점에서는 한 군집의 시작점이나 평균(median), 끝점을 기준으로 나눔

결국 층별, 박스별로 나눈 데이터프레임에 대해 열별 좌석번호 별 시각화를 통해 특이점을 찾아서 좌석 등급을 나눌 지점을 찾고 좌석 재분배 진행

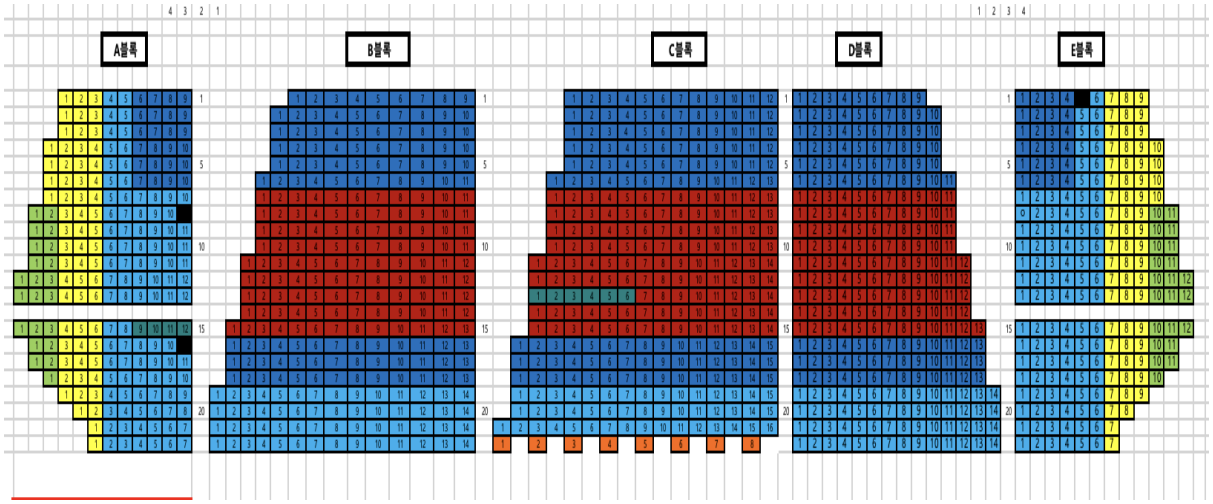


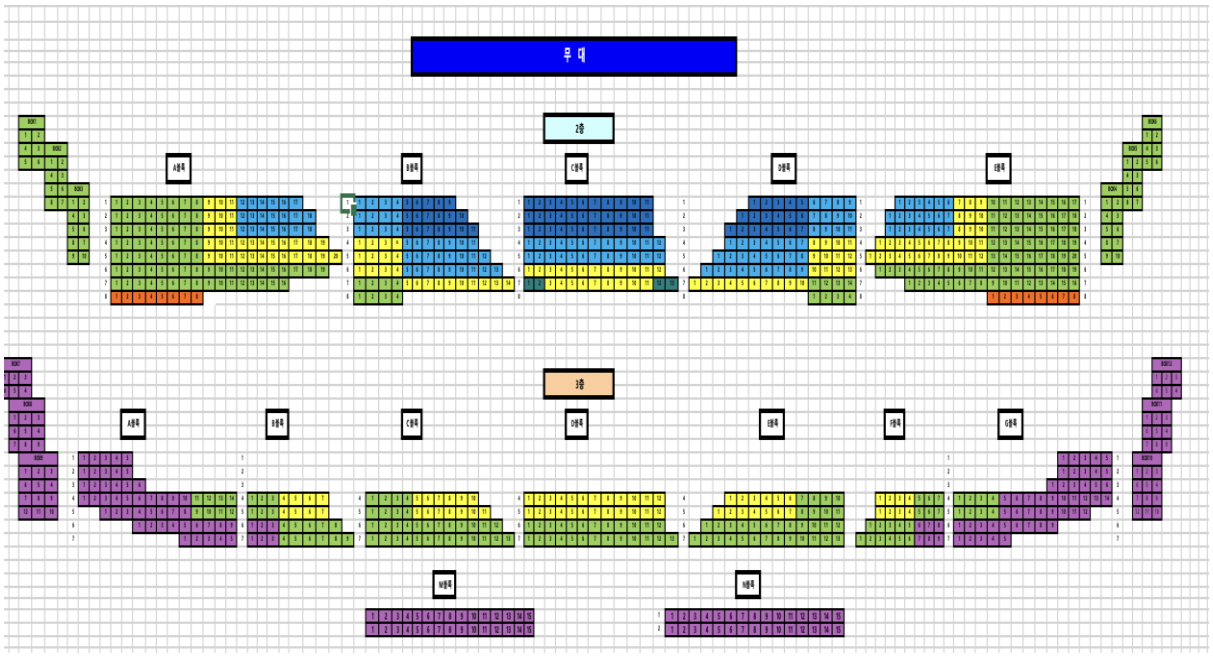
위 그래프는 위의 합창 예시 그래프 중, 1 층의 E 블록에 대한 그래프들이다.

왼쪽은 군집별 열에 대한 분포, 오른쪽은 해당 블록의 자리번호 빈도 Countplot 이다. 위의 가정 및 방법을 토대로 좌석 재분배 예시를 들어보자면,

우선 왼쪽 1,3,6 군집은 6 옆 안팎에서 세 군집이 겹치는 부분이 생긴다. 이 열에 대해서는, 군집별 가격 통계량에 상관없이 각 군집에 해당하는 관객들의 좌석이 겹치는 부분이다. 이 부분을 경계로 삼고, 오른쪽 그래프에서 자리번호 4 번에서 5 번으로 급격한 양상을 띈다. 이 점을 경계로 삼고, 만들어지는 부분에 대해 새로운 좌석 군집을 제시할 수 있을 것이다.

이러한 방법으로 좌석을 재 배치했을 때, 예시는 다음과 같다. 예시는 계속 합창 장르로 이어가겠다.





위에서부터 1층에 해당하는 좌석 재배치 제안 예시이다. 양식은 예술의 전당 홈페이지에서 참고했다. 장르마다 표본으로 뽑은 공연의 좌석 배치는, 현재 예술의 전당 공연 예정인 공연들에 대해서 추출했다.

군집이 6~7개 인 경우에 위와 같은 method로 좌석을 재배치한다면 좌석 등급이 6개로 기존보다 더 다양해 진다. 합창의 경우, 맨 앞 좌석보다 중간 좌석이 더 좋다는 시청각적 요소와 기존의 좌석 배치를 고려했고, 1층 A,E블록과 2층 A,B,D,E 그리고 3층 B,F 등의 블록에서 앞서 말한 메소드를 기반으로 좌석을 데이터에 기반해 재배분했다.

재배분한 6개의 군집에 대하여, 원본 데이터에 가격 순위 열을 추가했다. 이를 통해, K-Means 클러스터링을 기반한 군집 별 가격 통계량과 가격 순위에 기반해 군집 별 적정가를 생성할 수 있다.

가격 통계량에서, 원본 데이터의 price와 discount_type을 통해, 계산된 티켓의 원가를 사용했다. 장르별,

좌석 가격 순위 별, 티켓의 원가의 통계량으로, 적정 가격 예측 모델의 target을 생성했다.

이때 사용된 원가 데이터에서, 통계량은 price가 0인 고객을 제외하고 책정했다.

좌석 재분배는 좌석 선택 다양성으로 인해 고객의 만족도를 충족시킬 수 있으며, 다양한 연령층과 다양한 계층의 사람들에게, 한정된 것이 아닌 여러 좌석 선택지를 줌으로써 공공성에도 기여할 수 있다.

라) 가격 모델링

1. 예측 모델 사용 목적 및 Target 부연설명

Task 6에서는 Clustering의 결과를 바탕으로, 좌석 재분배를 통한 각 군집별 적정가격을 도출해 새로운 target으로 설정하여 예측모델을 수립했다. 이 예측 모델의 목적은 다음과 같다. 클러스터링에 사용한 변수뿐만 아니라 다른 변수를 추가해서 모델링을 진행했을 때, 특성들이 해당 가격을 잘 설명할 수 있는지, 또한 변수 및 모델 자체의 설명력을 검정하는 것이 목적이다.

다음은 모델의 설명력 검정 및 공공성, 수익성, 만족도 측면을 고려해 보다 더 확장된 모델, 그리고 일반화될 수 있는 모델을 설정하였다. 이에 따라 다음과 같은 피처들을 생성하였다.

2. Feature Engineering

1. 공공성 측면

1-1) 공연_취약계층비율

이 피처는 discount_type에 65세|실버|유족|학생|경로|장애|노블|의사상|임산|청소년|청년|유공|어린이|의상이 포함된 데이터를 추출해, 공연_취약계층비율에 대해 정의한 피처이다. 보통, 노인, 학생, 어린이, 장애인 등은 티켓 가격이 일반 관람객보다 저렴할 것이기 때문에, 가격에 영향을 줄 것 같은 피처로 선정하였으며, 공공성 또한 고려할 수 있다는 점에서 생성하였다.

2. 수익성 측면

2-1) 장르별 평균가

이 피처의 경우, 장르별로 평균 가격을 나타낸 피처를 의미한다. 이 피처는 예측 모델의 설명력을 높일 것으로 예상되며, 공연의 특성을 고려하여 가격을 예측하는 데 도움이 될 것이라고 판단해 피처를 생성하였다. 또한, 공연의 수익성 측면에서도 유용한 정보라고 할 수 있다. 장르별 평균 가격이 높을수록 해당 공연들은 더 높은 수익을 창출할 가능성이 있으며, 이러한 정보는 가격 예측 모델에 유용하게 활용될 수

있을 것이다.

2-2) 장르별 공연 비율

이 피처의 경우, 장르별 공연의 비율을 나타낸 피처를 의미한다. 다양성과 수익성, 경쟁력에 대해 좋은 영향을 끼칠 것이라고 판단하여 생성하였다. 장르별 공연 비율은 다양한 공연 유형의 포트폴리오를 반영한다. 다양한 장르의 공연을 제공하는 것은 고객들의 다양한 요구를 충족시키는데 도움이 된다. 이는 수익성을 향상시키는 중요한 요소 중 하나일 수 있다. 또한, 장르별 공연 비율은 가격 정책에도 영향을 미칠 수 있다. 특정 장르의 공연 비율이 높을 경우, 해당 장르의 공연은 고객들에게 더 많이 제공되는 경향이 있을 것이며, 따라서 해당 장르의 평균 가격을 높이거나 낮추는데 영향을 미칠 수 있다.

2-3) 공연 날짜의 상반기 하반기로의 분류

공연이 하반기로 갈수록 공연의 증가, 평균가가 올라간다는 분석 결과를 바탕으로 가격에 영향을 미칠 것이라 판단하여 해당 피처를 생성하였다. 1월부터 6월까지는 0, 7월부터 12월까지는 1로 나타내도록 피처를 생성하였다.

2-4) month == 11 여부

월 별 공연 가격의 평균을 시각화 했을 때, 11월의 공연 평균 가격이 가장 높음을 확인하였다. 이는 날씨가 추워지면서 날씨의 변화로 인해 실내 활동인 공연에 대한 수요 증가로 볼 수 있으며, 이를 통해 공연의 수요 증가에 따른 가격 증가 혹은 공연의 수요 증가에 따라 인기 있는 공연의 개최 등이 이유가 될 수 있다. 따라서, 해당 피처는 가격에 영향을 미칠 것이라 판단하여 해당 피처를 생성해주었다.

3. 만족도 측면

3-1) 재관람 여부

Discount_type을 확인하였을 때, '재관람' 이라는 정보가 있는 할인 혜택이 존재하였다. 이를 통해, 만족도 측면 중에서도 '공연'에 대한 만족도 측면에서 바라보았고, 이를 통해 만족도를 충족시키는 동시에, 가격에도 영향을 줄 것 같다고 판단하여 해당 피처를 생성해주었다.

3-2) 구매 당시 유료회원인 경우

해당 피처는 멤버십 타입 피처와 discount_type을 활용해 제작한 피처이다. 이는 만족도 측면 중, 예술의 전당 시스템 자체에 대한 만족도 부분에서 바라본 관점으로 제작해주었다. 멤버십 타입이 유료회원으로 나타나더라도, 해당 discount_type이 골드/그린/블루/법인 인 경우에만 구매 당시 유료 회원이라고 판단하여 해당 피처로 생성하였다.

3-3) 주말 여부

해당 피처는 해당 공연이 주말에 열리는지 열리지 않는 지의 여부 피처로, 해당 데이터를 분석해본 결과 관람객의 수가 토, 일요일에 가장 높은 분포를 차지했으며, 공연 수 또한 토, 일요일에 가장 많이 열리는 것으로 파악하였다. 주말에는 다양한 날씨, 휴가, 가족 모임 등의 활동을 계획하는 경향이 있으며, 이는 관람 수요 증가로 이어질 수 있다고 판단하였고, 또한 주말에 공연이 가장 많이 열리는 경우에 대해서는 '경쟁력'이라는 요소를 고려해볼 수도 있다. 주말에 공연이 많이 열리는 경우, 다른 공연장과의 경쟁이 치열해질 수 있다고 말할 수 있다. 이는 공연 장소 간의 가격 경쟁을 촉진하고, 주말에 공연을 보다 매력적으로 만들려는 경향이 있다. 따라서 주말 공연의 평균 가격이 상승하는 경향이 있다라고 말할 수 있고, 가격에 영향이 미칠 것이라 판단하여 피처로 생성하였다.

3-4) 자리번호의 범주화

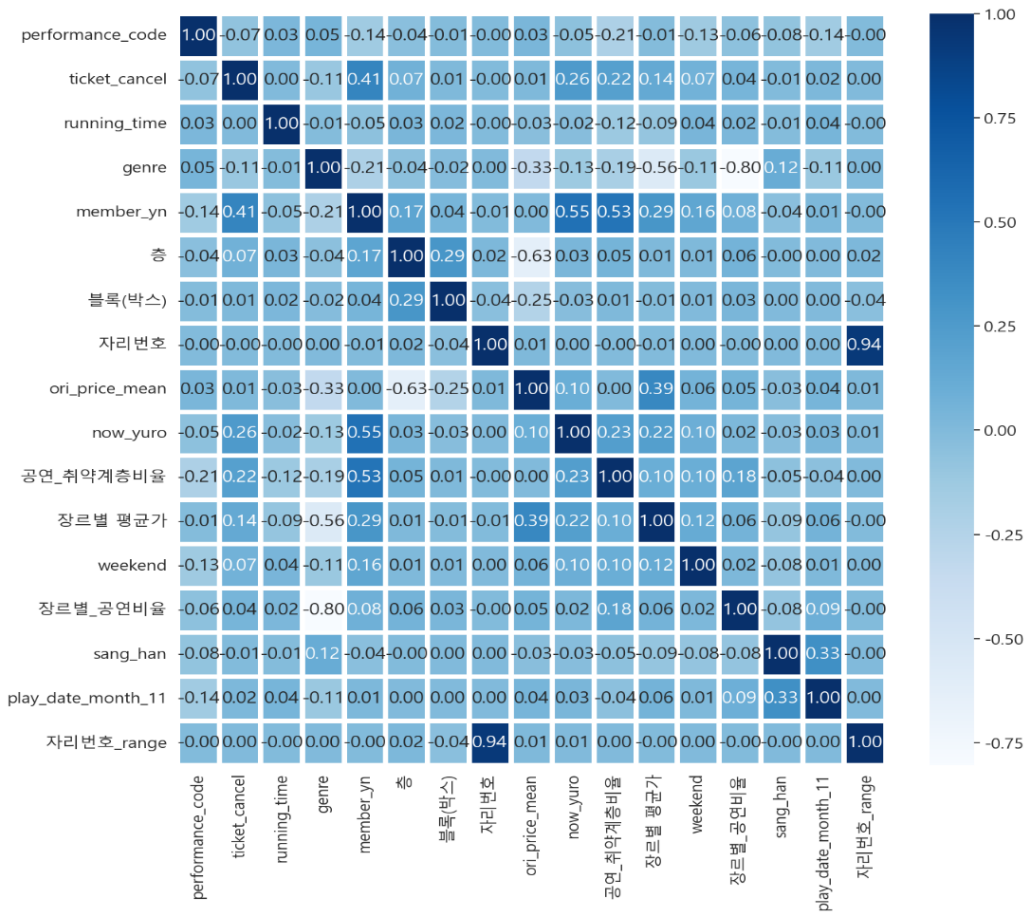
해당 피처는 자리번호의 범주화 피처로, 1이상 5이하는 1로, 6이상 10이하는 2, 11이상 16이하는 3, 그 이상은 4로 반환해주는 피처를 생성해주었다. 자리번호의 범주화 피처는 공연장 내 자리들을 몇 가지 범주로 나눈 정보를 담고 있으며, 이를 통해 가격 설정과 공연장 내 자리 배치를 효율적으로 고려할 수 있도록 하고, 가격 설정에도 효율성을 높이는데 도움이 될 것이라 판단하였다.

위의 생성된 피처에 기존 피처인 performance_code, ticket_cancel, running_time, genre, member_yn, 층, 블록(박스)인 7가지 변수를 추가해주었으며, 이에 따라 최종 모델링에 사용한 피처는 총 16가지의 피처를 사용하였다.

Encoding/Scaler

Encoding 기법의 경우, label Encoding을 사용해주었다. 범주형 변수인 'genre', 'member_yn', '블록(박스)', 'sang_han' 4가지에 대해 인코딩을 진행해주었으며, Scaler의 경우, categorical_feats에 대해 scaling을 진행하였으며, 사용된 변수는 'performance_code', 'ticket_cancel', 'running_time', '층', '자리번호', 'now_yuro', '공연_취약계층비율', '장르별 평균가', 'weekend', '장르별_공연비율', 'play_date_month_11', '자리번호_range', 으로 총 11가지 변수에 대해서 StandardScaler를 사용해주었는데, MinMax Scaler, Robust Scaler와 비교하였을 때, 가장 잘 나와 StandardSclaler를 사용하였다.

피처 간 상관관계



피처 간 상관관계는 대부분 높은 음/양의 상관관계를 지니지 않아, 16가지의 피처를 그대로 모델에 사용하였다.

3. Modeling

Modeling 하기 전, 기존 train 값과, test 값을 train, test, y_train, y_test, 0.3의 비율로 split해주었고, 나누어준 train 값과 y_train 값을 다시 쪼개어 X_train, X_valid, y_train, y_valid, 0.3의 비율로 split 해주어 학습에 활용하였다.

학습과 평가에 쓰는 모델의 후보는 총 6가지였으며, 6가지의 단일 모델 중, 가장 성능이 잘 나온 LGBM Regressor을 최종 모델로 선정하였다. 모델은 Tree 계열인 DecisionTreeRegressor와 ExtraTreeRegressor, RandomForest를 사용하였으며, 두 번째로는 Boosting model로 LGBMRegressor, GradientBoostingRegressor, AdaBoostRegressor을 사용해주었다. 다음은 왜 이 모델을 모델링의 후보로 사용했는지에 대한 다음과 같은 이유로 선택하였다.

우선, Tree 계열인 DecisionTreeRegressor와 ExtraTreeRegressor, RandomForest를 사용하였다. 이러한 트리 기반 모델은 데이터 내의 비선형 관계를 잘 모델링하고, 변수 중요도를 해석하기 쉬워 선택하였다. 또한,

RandomForest는 다수의 트리를 결합하여 과적합을 줄이고 예측 성능을 향상시키는 앙상블 모델이라는 특징을 지닌다.

두 번째는 Boosting 모델로 LGBMRegressor, GradientBoostingRegressor, AdaBoostRegressor을 선택하였다. 이들은 경사 부스팅 알고리즘을 기반으로 하며, 과적합을 줄이고 높은 예측 성능을 제공한다는 특징을 지닌다. LGBMRegressor는 고차원 데이터에서 빠르고 효과적인 모델 학습이 가능하며, GradientBoostingRegressor와 AdaBoostRegressor는 예측 성능이 높고 이상치에 강건한 특성을 갖고 있다.

이러한 모델들은 다양한 데이터 패턴을 다루고 해석 가능한 결과를 제공하기 때문에, 모델링 후보로 선택해주었다. 모델의 다양성을 활용하여 최적의 가격 예측 모델을 개발하고자 하였다.

가) 평가지표

평가지표의 경우, 총 4가지인 RMSE, MAPE, MAE, R^2 Score를 평가지표로 사용해주었다.

선택한 평가지표인 RMSE, MAPE, MAE, R^2 Score를 사용한 이유는 다음과 같다.

1. RMSE (Root Mean Squared Error) : RMSE는 예측값과 실제값 간의 차이를 측정하는 지표로, 예측 오차의 제곱을 평균한 후 제곱근을 취한 값이다. 작은 RMSE 값은 모델의 예측이 실제값과 가깝다는 것을 의미하며, 모델의 예측 정확도를 평가하는 데 사용된다는 특징을 지닌다.

2. MAPE (Mean Absolute Percentage Error) : MAPE는 예측값과 실제값 간의 백분율 오차를 측정하는 지표로, 예측 오차를 실제값으로 나눈 후 평균을 계산한다. 이 지표는 예측 오차가 상대적으로 얼마나 큰지를 나타내며, 예측의 상대적인 정확도를 평가하는 데 사용된다. MAPE는 비율적인 오차를 고려하므로 예측값의 크기에 영향을 받지 않는다는 장점이 존재한다.

3. MAE (Mean Absolute Error) : MAE는 예측값과 실제값 간의 절대값 차이를 평균한 값으로, 예측 오차의 크기를 나타낸다. MAE는 예측 오차의 크기를 고려하는 지표로, 작은 MAE 값은 모델의 예측 정확도가 높다는 것을 의미한다는 특징을 지닌다.

4. R^2 Score (Coefficient of Determination): R^2 Score는 모델이 설명하는 분산의 비율을 나타내는 지표로, 예측값과 실제값 간의 관계를 평가한다. 이 지표는 모델이 데이터의 변동성을 얼마나 잘 설명하는지를 나타내며, 1에 가까울수록 모델이 데이터를 잘 설명한다는 것을 의미한다.

이렇게 네 가지 평가지표를 선택한 이유는 각각 다른 측면에서 모델의 성능을 평가하고자 함이다. RMSE와 MAE는 예측 오차의 크기를 평가하고, MAPE는 상대적인 정확도를 고려하며, R^2 Score는 모델이 데이터를 얼마나 잘 설명하는지를 측정한다. 이를 통해 모델의 다양한 측면에서 예측 성능을 평가하고 최적의 모델을 선택할 수 있을 것이라 판단하여 위의 4가지의 다양한 평가지표를 활용하였다.

나) 후보 Model1

1. Tree model

1-1) DecisionTreeRegressor

```
rmse_score :  9982.71512752902
mape_loss :  0.08481721950399064
MAE Score: 5821.775044023913
R2 Score: 0.8443574166504303
```

1-2) ExtraTreeRegressor

```
rmse_score :  10015.675177043733
mape_loss :  0.0854458410558029
MAE Score: 5866.460636161243
R2 Score: 0.843465285616115
```

1-3) RandomForestRegressor

```
rmse_score :  9131.212630643253
mape_loss :  0.08282827240391658
MAE Score: 5720.507784821421
R2 Score: 0.8660677497057074
```

Tree model 계열 중에서는 RandomForestRegression model, R^2 = 0.87, MAE = 5720, mape = 0.08, rmse = 9882 정도로, 가장 좋은 성능, 좋은 설명력을 지닌다는 것을 확인할 수 있었다.

다) 후보 Model2

2. Boosting Model

2-1) LGBMRegressor

```
rmse_score :  8629.138110698916
mape_loss :  0.09149864785964408
MAE Score: 6193.455378134949
R2 Score: 0.8748460873116153
```

2) GradientBoostingRegressor

```
rmse_score :  9803.890000386706
mape_loss :  0.11488109213564379
MAE Score: 7672.446834367995
R2 Score: 0.8286427971462678
```

2-3) AdaBoostRegressor

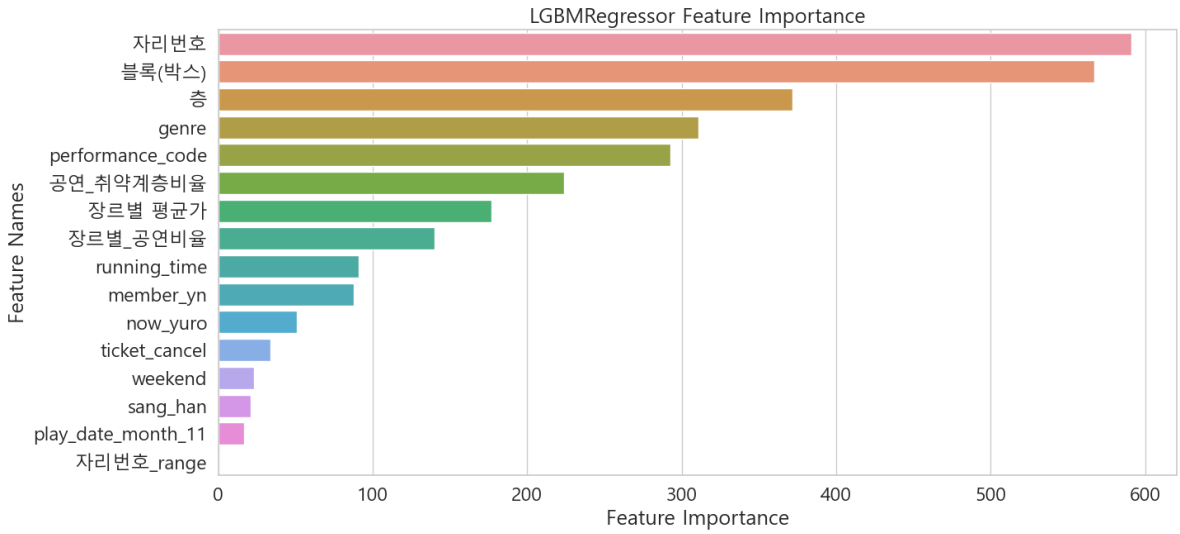
```
rmse_score :  13731.656210187199
mape_loss :  0.18135165695450886
MAE Score: 12154.028402863716
R2 Score: 0.45164935819783614
```

Boosting Model 계열 중에서는 LGBMRegressor model, R^2 = 0.87, MAE = 6193, mape = 0.18 rmse =

13731 정도로, 가장 좋은 성능, 좋은 설명력을 지닌다는 것을 확인할 수 있었다.

이를 통해, 최종 Model은 가장 설명력이 좋고, 성능이 좋았던 LGBMRegressor으로 최종 선정하였다. 이후, optuna를 활용한 tuning도 진행했었으나, 다양한 시도와 실험 결과, 단일 모델을 사용하는 것이 일반화 측면, 설명력 측면에서 가장 뛰어나 LGBMRegressor 단일 모델로 진행하였다.

라) Feature Importance



LGBMRegression에 대한 Feature importance를 시각화한 모습이다.

라) 가격 보정을 통한 최종 가격 제시

1) 가격 보정 필요 이유 - 가격 변동성 반영

군집화를 통해, 군집 별 가격 통계량을 통해 정한 가격은, 적정 가격으로 제안하기엔 담지 못하고 있는 정보가 많다고 판단이 된다. 장르 별 특성은 장르별 군집화를 하며 반영이 되었더라도, 개별적인 공연 별 특성에 대해선 미흡하다.

이에, 기본적인 예매 데이터에서는 얻기 어려운 공연 별 특성을 외부 데이터를 활용해서 인사이트를 도출해 활용하고자 한다.

여기서 얻은 공연별 인사이트를, 변동성이라는 함수를 만들어 가격 제안에 반영한다. 이에 따라, 최종적으로 제시하는 적정 가격은 군집 별 통계치에 변동성 곱해서 책정된다.

가격 변동성을 정의할 때, kopis 공연예술통합 전산망의 Open API를 통한 공연장 별 공연 목록 및 공연 세부사항 데이터를 활용했다.

2) 가격 보정 수식

Open API에서, 2018년부터 2023년까지의 콘서트홀에서 열린 공연 약 1000개의 데이터를 수집했다. 공연들에 대해서, 좌석 등급을 각각 컬럼으로 만들고, 각 좌석 등급에 따른 가격 통계치를 확인했다. 이에 기반해 각 좌석 등급이 고가로 분류되는, 즉 앞선 좌석 등급에 따른 가격 통계치에서, 상위 75%에 해당하는 그룹에 대해 좌석 등급의 개수를 추출했다.

같은 방식으로 하위 25%에 해당하는 그룹에 대해서도 좌석 등급의 개수를 추출해서 각각 칼럼으로 추가했다.

그리고 각 좌석 등급 별, 상위 75%와 하위 25%의 좌석 등급 개수를 value_counts를 통해 확인했다. 그 결과, 높은 구매가의 공연일 수록, 좌석 등급 개수가 많아지는 인사이트를 도출했다.

```
df['S'].describe()
```

```
count      1214.000000
mean      62887.973641
std       48973.509036
min           0.000000
25%       30000.000000
50%       52500.000000
75%       90000.000000
max      390000.000000
Name: S, dtype: float64
```

```
df['class_cnt'][df['S'] < 30000].value_counts(normalize=True)
```

```
2    0.535088
1    0.241228
3    0.206140
4    0.008772
6    0.004386
5    0.004386
Name: class_cnt, dtype: float64
```

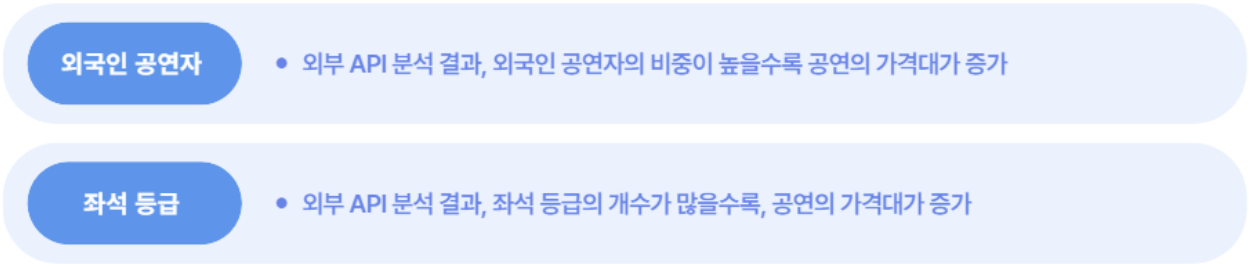
```
df['class_cnt'][df['S'] > 90000].value_counts(normalize=True)
```

```
4    0.570833
5    0.362500
3    0.058333
2    0.008333
Name: class_cnt, dtype: float64
```

위의 df는 OPEN API를 통해 얻은 공연들의 세부사항이 담긴 데이터 프레임이며, 좌석 등급 중 S석에 대한 분석이다. 확인해본 결과, 각 좌석 등급들 모두 비슷한 양상을 보임을 확인했다. 고가의 공연일 수록, 좌석 등급 개수가 많다.

또한 같은 방식으로, 좌석 등급 별 각 25%와 75%에 해당하는 그룹 중에서, 외국인 공연자가 포함된 공연의 빈도를 확인했다.

이를 통해 비슷한 인사이트로, 외국인 공연자가 포함된 공연의 가격은 대체로 고가로 형성되어 있음을 확인했다.



3) 장르 별 좌석 등급에 따른 가격 기준 제안

가격 보정을 위해 사용한 변동성의 함수는, 과제로 주어진 데이터에서는 활용할 수 없다는 한계점이 있다. 주어진 데이터에서는, 익명성을 위해 공연명에 대한 정보가 마스킹 되어 있기 때문에, 공연 별 세부사항에 대한 정보를 얻을 수 없다. 하지만 추후 예술의 전당 측에서 효과적 가격 제안을 위한 모델링을 수행할 때는, 마스킹이 되지 않은 예매 데이터로 활용을 한다는 가정 하에, 제시하는 가격 보정을 사용할 수 있다.

직접적인 모델링을 통해 일반화된 수식을 제공할 순 없지만, 위의 두가지 공연 별 세부적 특성인, 고가의 공연일 수록 외국인 공연자의 빈도가 높고, 좌석 등급의 개수가 많다는 점을 활용해, 외국인 공연자수가 없는 공연에 음의 가중치를, 외국인 공연자가 등장하는 공연에 대해서는 양의 가중치를 부여하는 등의 방식으로 효과적 가격 제안을 할 수 있다.

본 팀이 제시하는 모델은, 최종적으로 고객의 좌석 선호도를 반영한 군집별 특성과 장르 등을 고려한 가격에, 공연별 세부적인 특성을 추가로 반영한 가격을 제안할 수 있다. 이를 통해 본 팀이 대회를 진행하며 고려한 공공성, 고객만족도, 수익성 세가지 측면을 두루 고려한 가격을 책정할 수 있다.

Ⅲ. 주요 결과 및 시사점

1. 주요 결과 요약

가. 좌석 등급별 분할

1) 추가 제출물의 좌석 배치 제안 폴더 확인

2. 결과 활용 및 시사점

1) 추가 제출물의 좌석 배치 제안 폴더 확인

가. 좌석 등급의 다양성 측면 시사점
제시하는 좌석 재배치 예시는 6개의 좌석 군집으로, 기존의 최대 5개의 좌석 배치도를 고려했을 때, 최소 1개 최대 5개의 좌석 종류의 증가를 시사한다. 비교적 예매 가격이 싼, 즉 공공성을 고려하는 공연의 측면에서도 새로이 제시하는 좌석 종류 별 가격에, 변동성을 고려해 최종 가격안을 제시한다면, 예술의 전당과 공연 개최 측의 수익성과 고객의 만족도 및 다양성을 모두 고려하는 가격 모델이 될 수 있을 것이라 생각한다.

나. 결과 활용 아이디어 제시
본 데이터로 군집화를 사용했을 때, 초대권 및 기획사 판매에 관련해 price가 0인 데이터의 비중이 굉장히 많았다. 가격 제시에 있어서 이렇게 많은 비율의 price가 0인 데이터의 활용이 잘 될 수가 없어서 군집화를 제외하고 군집 별 적정 가격 설정에 있어서는 해당 데이터를 사용하지 못했다. 때문에 가격 설정에서 사용되는 데이터의 수가 많이 감소했다는 한계점이 있다. 해당 price가 0인 데이터의 대부분은 회원이 아니며, 성별과 나이 등의 고객관련 특성 또한 결측되어 있다. 추후에 멤버 전환율을 위한 혜택 증대

등의 정책으로 전환율을 높이고, 멤버인 고객들의 데이터를 늘린다면, 성별과 나이 등의 고객 특성도 군집화의 축으로 사용될 수 있으며, 군집 별 적정 가격 제시에도 많은 양의 데이터를 기반해서 책정할 수 있다면, 현재 제시하는 모델보다 더욱 탄탄한 적정 가격 제시 모델이 될 수 있을 것이라 생각한다.
고객 데이터가 군집화에 사용된다면, 기존에는 존재하지 않았던 좌석 종류 별, 고객 특성을 고려한 할인 정책을 제안할 수 있을 것이다. 멤버십 타입 별 할인 정책이 아닌, 군집 분석을 통한 세부적인 고객 타겟을 설정해 할인 정책을 제시한다면, 타겟에 해당되는 고객의 입장에서도 만족도가 증대할 것이고, 위와 같은 방식이라면 취약 계층에 대한 할인 정책 등도 다양하게 고려될 수 있고, 이는 공공성에 기여할 수 있다. 또한 군집 분석을 통해, 가격 탄력성을 고려한다면 특정 군집에 있어서 수익성을 보장받을 수 있다는 장점이 있다.

다. 기대 효과
예술 공연의 가격 산정에 있어서, 시장에서 유사한 경쟁 공연상품의 가격을 비교하여 티켓 가격을 결정하는 시장대응 가격결정 방법은 오늘날 공연기획자들이 주로 많이 사용하는 방법이다. 따라서 경쟁사가 가격을 올리게 되면 자연스레 공연 가격이 올라가게 되고, 이런 인플레 현상을 막기 위해 예술의 전당은 현재의 좌석 등급제를 시행하고 있다.
이에 더해서 고객들의 지난 예매 데이터를 활용한 군집화 및 기계학습을 통한 적정가격 제시 모델은, 고객들에게 객관적이고, 투명한 가격 산정에 대한 근거가 된다. 또한 군집 분석을 통한 타겟 할인 정책이 실행된다면 현재보다 더욱 다양한 계층의 고객 수요가 늘게 될 것이다.
고객 수요가 늘게 됨은, 동시에 예술의 전당 음악당의 경쟁력 증대를 시사한다. 예술의 전당에 대한 경쟁력 증대 및 고객층의 증가는 곧 클래식 공연의 활성화로 이어지게 된다.
공연 활성화와 고객의 증대는 결국 예매 데이터의 증대로 이어지고, 이로 인해 군집 분석 방법을 통한 적정 가격 제시 모델은 더 일반화된 좋은 성능으로 발전할 것으로 기대된다.