최종 보고서

4조 (남지원, 박준민, 유건욱)

1. Data Structure

시설 명	소공연장명	시설 정보		공연 정보
		시설 내부 정보	시설 주변 정보	
key	key	KOPIS (www.kopis.or.kr) 제공 공연시설 excel 및 크롤링 시설 총 좌석수, 소공연장 좌석수, 시설 특성, 개관 연도, 주소	kakao map API 크롤링 시설 주변 편의점, 주차장, 주요소, 지하철, 문화 시설, 관광 명소, 숙박 업소, 음식점, 카페 총 개수	KOPIS (www.kopis.or.kr) 제공 공연 excel (2018.01.01 ~ 2019.12.31) 해당 시설에서 공연한 공연 제목, 공연 장르, 공연 시작 날짜, 공연 종료 날짜

< 그림1 : 2차 보고까지의 예상 Data Structure >

시설 명	소공연장명	시설 정보		공연 정보	예매 정보
		시설 내부 정보	시설 주변 정보		
key	key	KOPIS (www.kopis.or.kr) 제공 공연시설 excel 및 크롤링 시설 총 좌석수, 소공연장 좌석수, 시설 특성, 개관 연도, 주소	kakao map API 크롤링 시설 주변 편의점, 주차장, 주요소, 지하철, 문화 시설, 관광 명소, 숙박 업소, 음식점, 카페 총 개수, 위도, 경도	KOPIS (www.kopis.or.kr) 제공 공연 excel (2018.01.01 ~ 2019.12.31) 해당 시설에서 공연한 국악, 무용, 뮤지컬, 복합, 연극, 오페라, 클래식 공연 수 해당 시설에서 공연한 총 공연 수	KOPIS (www.kopis.or.kr) 제공 공연시설 통계 API 활용 (2018.01.01 ~ 2019.12.31) 해당 극장 총 예매 수

< 그림2 : 최종 Integrated Data Structure >

1) 공연 정보의 구조 변화

각각의 공연 시설의 해당 소공연장에서 2018년, 2019년 공연한 모든 공연의 이름과 장르, 공연 기간을 넣으려고 하였습니다. 하지만 각 공연장별로 공연의 수의 차이가 컸습니다. 예를 들어 A공연장에서는 10개의 공연을 했지만, B공연장은 0개인 케이스가 많았습니다. 따라서 2차 보고서에서 계획 했던대로 진행한다면 데이터 손실이 클 것으로 판단했습니다. 팀원들과 상의해본 결과 의미있는 정보는 장르 뿐이라고 판단했습니다. 최종적으로 각각의 공연 시설에서 2018년, 2019년에 진행한 장르별 공연 수와 총 공연 수만 남겼습니다. 자세한 Data Structure는 다음과 같습니다.

시설 명	소공연장	시설 정보	공연 명	장르	시작일	종료일
시설A	공연장 a	시설정보 A, a	공연 1	장르 1	시작 1	종료 1
			공연 2	장르 2	시작 2	종료 2
	공연장 b	시설 정보 A, b	공연 3	장르 3	시작 3	종료 3
...	...	...	...	...	...	...

< 그림 3 : 2차 보고까지의 자세한 예상 Data Structure >

시설 명	소공연장	시설 정보	국악	무용	...	총공연 수	총예매 수
시설A	공연장 a	시설정보 A, a	n_a1 개	n_a2 개	...	N_a1 개	N_a2 개
	공연장 b	시설 정보 A, b	n_b1 개	n_b2 개	...	N_b1 개	N_b2 개
...	...	...	...	...	...	...	...

< 그림 4 : 자세한 최종 Integrated Data Structure >

2) 데이터 추가

모든 시설에 대한 주소는 있었지만 주소의 위도 경도도 추후 데이터 분석 할 때 필요하다고 판단했습니다. 각 주소에 해당하는 위도 경도를 추가로 크롤링 해서 최종 데

이터에 추가했습니다. 그리고 KOPIS에 새롭게 추가된 정보중에서 유의미한 데이터를 찾아보았습니다. 여러 정보들이 있었지만, 1차적으로 통합된 데이터에 추가할 수 있는 데이터는 예매량만 유일하다고 판단되어서 예매량 데이터만 추가하게 되었습니다.

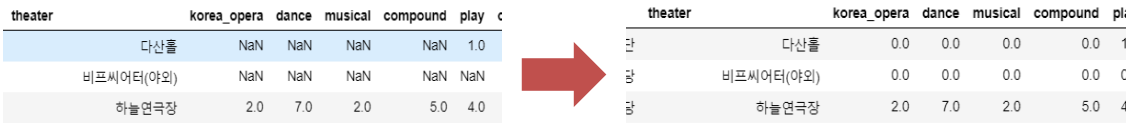
2. NA 처리 및 오류 수정

1) 시설 내부 정보, 시설 주변 정보 합치기

시설 주변 정보의 경우 KOPIS에서 제공된 주소를 기반으로 했기 때문에 공연 정보 data의 주소 정보에 left join을 해서 통합했습니다. 이 과정에서 주변시설이 NA가 되는 경우가 있었고, 확인해보니 주소가 잘못되어 있어서 수정 후 재결합시켰습니다.

2) 시설 정보, 공연 정보 합치기

공연 정보 data를 시설 명, 소공연장 명을 index로 하여 pivot table을 만들어 시설 정보와 시설, 소공연장 별 공연 장르 수를 나타내는 data로 변경했습니다. 이 과정에서 공연하지 않은 장르나 공연이 없었던 시설에 NA가 발생하였고, 이를 모두 0으로 바꾸었습니다. 최종적으로 시설 명, 소공연장 명을 key로 시설 정보에 left join하여 통합했습니다.

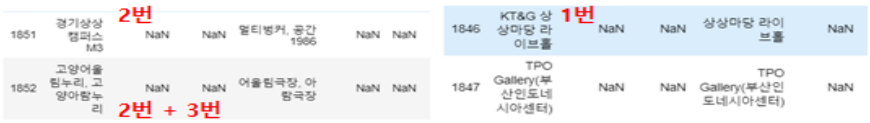


theater	korea_opera	dance	musical	compound	play
다산홀	NaN	NaN	NaN	NaN	1.0
비프씨어터(야외)	NaN	NaN	NaN	NaN	NaN
하늘연극장	2.0	7.0	2.0	5.0	4.0

theater	korea_opera	dance	musical	compound
다산홀	0.0	0.0	0.0	0.0
비프씨어터(야외)	0.0	0.0	0.0	0.0
하늘연극장	2.0	7.0	2.0	5.0

< 그림 5 : 공연 data 구조 변경 후 NA 처리 >

이 과정에서 예상보다 많은 데이터 손실이 일어나 원인을 살펴 본 결과, (1) 시설 정보에 없는 공연장, (2) 소공연장 2 개 관이 합쳐져 있음, (3) 시설 2개가 합쳐져 있음 등 총 3개의 경우를 발견할 수 있었습니다.



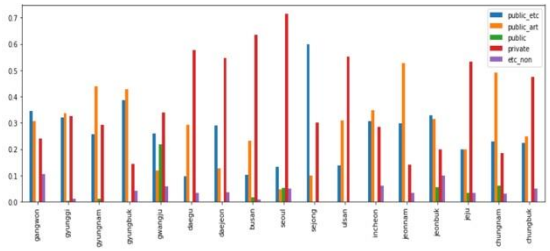
< 그림 6 : 데이터 손실 원인 >

총 87개의 데이터가 세 가지 이유로 NA값을 가졌고 이를 직접 하나하나 확인하여 수정한 후 데이터 손실을 줄일 수 있었습니다.

3) 예매 정보 합치기

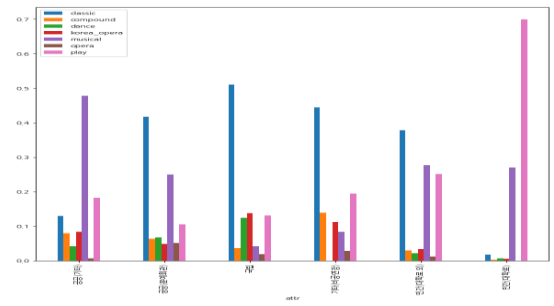
공연시설 데이터와 예매 수 데이터는 KOPIS(www.kopis.or.kr)에서 받은 같은 형식의 자료이기에 합치는 과정에서 형식의 차이로 인해 발생하는 NA는 없었습니다. 다만, 2018년, 2019년에 공연하지 않은 공연 시설의 예매 정보가 없었기 때문에 NA가 발생하였고 이는 0으로 채웠습니다.

3. 탐색적 자료 분석 (EDA)



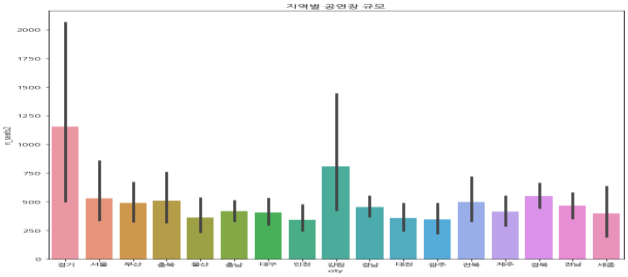
< 그림 7 : 지역별 공연시설 특성 비율 >

특별시, 광역시, 특별자치도의 민간 공연 시설의 비율 높은 것을 확인할 수 있었습니다. 민간 공연 시설의 비율이 0.5가 넘는 지역은 대구, 대전, 부산, 서울, 울산, 제주로 광주와 인천을 제외한 한국의 모든 특별시, 광역시, 특별자치도가 해당됩니다. 국립 공연 시설은 광주에서 특이하게 높은 비율을 가졌습니다.

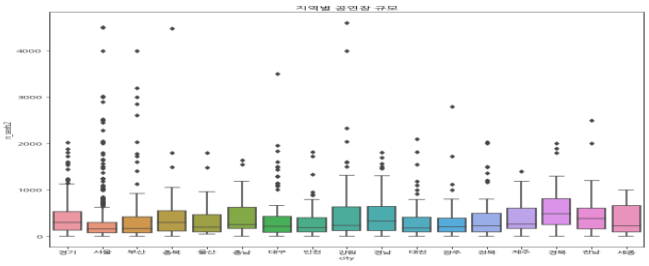


<그림 8: 공연 시설 특징별 장르 비율>

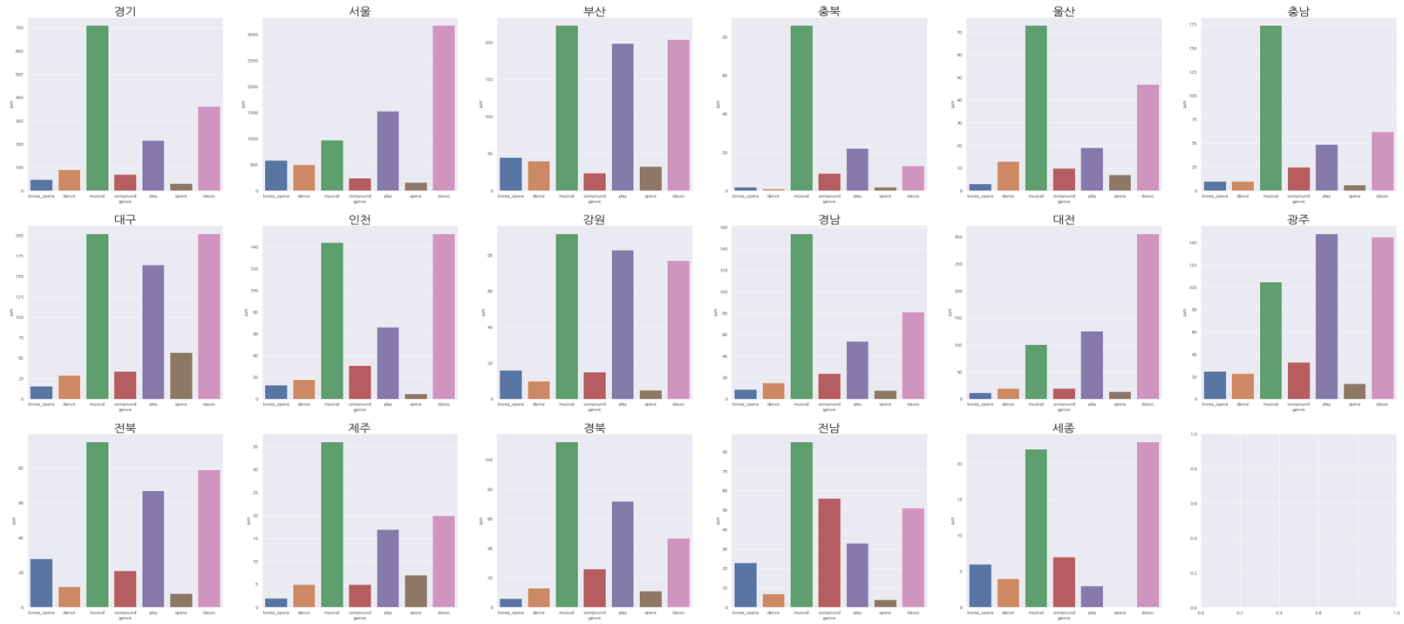
대학로 부근에서는 연극과 뮤지컬의 비율이 매우 높은 것을 확인할 수 있습니다. 혜화를 보더라도 연극만 하는 공연장이 많은 것을 알 수 있습니다. 공공(문예회관), 국립, 기타, 대학로 외 민간 시설의 경우 클래식 공연을 많이 하고 있습니다. 아무래도 클래식과 뮤지컬의 경우 수요가 많고 수익이 많을 것으로 예상됩니다. 공공(기타시설)의 경우 뮤지컬이 매우 높은 것을 확인할 수 있습니다. 국립 시설의 경우 전통 공연이 많이 이루어지는 것을 확인할 수 있습니다.



< 그림 9 : 지역별 공연시설 평균 규모>



< 그림 10 : 지역별 공연시설 규모>

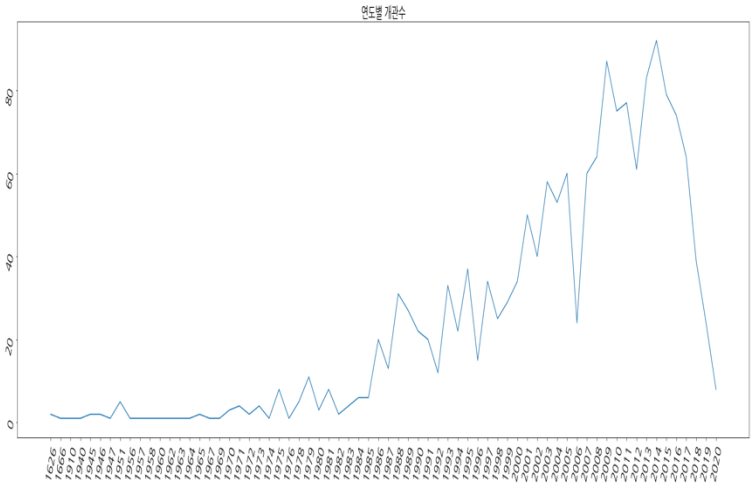


<그림 11: 지역별 공연 장르 비율>

지역별 공연시설 규모의 평균은 경기도에서 가장 높았지만, box-plot을 그려본 결과, 큰 규모의 시설은 경기도보다 서울과 부산에 현저히 많았습니다. 특히 서울의 경우, 공연시설의 수가 경기도의 2.5배, 부산의 7배 등으로 많은 것을 확인할 수 있었습니다. 이를 통해 서울에는 큰 규모의 공연시설도 많지만, 작은 규모의 공연시설 또한 매우 많은 것을 예상할 수 있습니다.

서울, 부산, 충북, 대구, 강원에만 약 3000석 이상의 규모를 갖는 시설을 보유하는 것을 확인했습니다. 3000석 이상의 좌석의 대부분은 벡스코, 엑스코 등의 대규모 전시시설, 종합운동장 등의 대규모 체육시설인 것을 확인했습니다.

지역별 공연장의 전체적인 특징을 보게 되면 클래식과 뮤지컬 공연 그리고 연극의 비율이 높은 것을 볼 수 있습니다. 수익을 생각해보면 이 세가지 공연 장르가 가장 많은 수익을 내기 때문이라고 생각합니다. 대부분의 지역에서는 인기가 가장 많은 뮤지컬의 비율이 높은 것을 볼 수 있습니다. 대전, 서울, 세종, 인천은 다른 지역에 비해 클래식 공연이 많은 것을 확인할 수 있습니다.



<그림 12: 연도별 개관 수>

연도별 개관수를 보게 되면 1980년대부터 조금씩 오르는 양상을 보이다가 2000년대부터 상승폭이 급격하게 오르는 것을 볼 수 있습니다. 이는 80년도부터 공연 관련 산업들이 성장하기 시작했고, 2000년대부터 공연의 규모가 늘면서 공연장의 수가 급격하게 성장했기 때문입니다..