

## 2 차 중간 보고서

4 조 (남지원, 박준민, 유건욱)


### 1. Data 크롤링

#### A. 공연시설 data crawling

공연예술 통합전산망 (<http://www.kopis.or.kr/por/main/main.do>)

사이트에서 공연시설 이름, 공연장 이름, 객석 수, 시설 특성, 개관연도, 주소에 대한 db를 엑셀로 제공받은 자료는 하나의 공연시설에 여러 개의 공연장이 있을 때에도 하나의 행으로 나타납니다. 각 공연장에서 공연한 공연을 column으로 추가할 계획이므로 하나의 공연장을 하나의 행으로 나누는 작업을 진행했습니다.

	name	n_theater	n_seats	theater_name	n_seats2
0	(재)경기문화재단	1	154	다산홀	154
1	(재)영화의전당	3	4877	하늘연극장비프시어터(야외)민간 다들러스	84104.000036
2	1M SPACE	1	0	1M SPACE	0



	name	n_theater	n_seats	theater_name	n_seats2
0	(재)영화의전당	3	4877	하늘연극장	841
1	(재)영화의전당	3	4877	비프시어터(야외)	4,000
2	(재)영화의전당	3	4877	민디들러스	36
3	3.15아트센터	2	1667	대극장	1,182

또한, 이 자료에는 해당 시설의 주요시설(장애인 시설, 편의시설, 주차시설) 등에 대한 data가 없어 웹 크롤링을 통해 공연시설 이름, 주요시설을 크롤링하고 각 주요시설이 있으면 1, 없으면 0으로 나타내고 공연시설 이름을 key로 사용하여 통합하였습니다.

통합된 데이터를 확인해보니 주소와 주요시설 변수에 각각 하나씩의 missing value가 있었습니다. 확인해본 결과 방콕에 있는 공연시설이 data에 들어가 있어 삭제처리 했습니다. 다운 받은 엑셀의 공연시설 이름과 스크롤링 한 공연시설 이름이 달라 주요시설 변수가 missing 되어 채워 넣었습니다.

	name	n_theater	n_seats	theater_name	n_seats2	attr	open	city	gu	address	pyuneui	jangaein	jucha
0	(재)경기문화재단	1	154	다산홀	154	공공(기타)	2001	경기	수원시	경기도 수원시 팔달구 인계로 178 (인계동)	1.0	1.0	1.0
1	1M SPACE	1	0	1M SPACE	0	민간(대학교 외)	2019	서울	서대문구	서울특별시 서대문구 연세로4길 27 (창천동)	0.0	0.0	0.0

#### B. 주변시설

##### 1) 네이버 지도 크롤링

네이버 지도에서 주변 시설을 표시해줄 수 있었기 때문에 selenium 으로크

롤링을 시도했습니다. 하지만 네이버 지도에서 직접 크롤링 하는 것에 제한이 있는 것인지 데이터가 추출되지 않는 현상이 있었습니다. 따라서 이 방식은 포기하고 api 를 활용해서 데이터를 추출하기로 했습니다.

## 2) 카카오 지도 api 활용

카카오 지도에서 데이터를 추출한 방식은 다음과 같습니다.

1. 공연장의 주소 데이터를 이용해서 좌표 데이터를 추출한다.
2. 추출된 좌표 데이터를 기준을 반경 500M 내에 있는 주변시설 데이터를 추출한다.

저희가 지정한 주변시설 데이터는 다음과 같습니다.

편의점, 주차장, 주유소, 지하철역, 문화시설, 숙박, 음식점, 카페

카카오 지도 api 를 활용해서 데이터를 추출하던 중 12 개의 데이터에서 좌표가 추출되지 않는 현상을 발견했습니다. 그 중 10 개는 좌표를 재탐색해서 주변 시설 데이터를 추출했고, 나머지 2 개는 해외 공연장과 폐관된 공연장이어서 탐색을 할 수 없었습니다. 해당 데이터는 삭제하기로 했습니다.

name	n_theater	n_seats	theater_name	n_seats2	attr	open	city	gu	address	CS2	PK6	OL7	SW8	CT1	AT4	AD5	FD6	CE7
(재)경기문화재단	1	154	다산홀	154	공공(기타)	2001	경기	수원시	경기도 수원시 팔달구 인계로 178	25.0	27.0	1.0	0.0	12.0	2.0	25.0	464.0	78.0
IM SPACE	1	0	IM SPACE	0	민간(대학로 외)	2019	서울	서대문구	서울특별시 서대문구 연세로 4길 27	58.0	69.0	1.0	2.0	18.0	5.0	93.0	1167.0	267.0

## C. 공연

공연 정보는 KOPIS 에서 제공하는 엑셀 파일을 이용했고, 그 파일을 공연장 지역 데이터 파일과 합친 다음 장르별/지역별 분류를 해 보았습니다.

title	genre	start_date	end_date	place	hall	seat_number
리미트	연극	2018.04.25	오론한	JTN 아트홀(구, 대학로예술극장)	2관	510
스크루지의 크리스마스 파티 (수원)	뮤지컬	2018.11.24	2019.01.01	KBS 수광아트홀	KBS 수광아트홀	200
공룡 애니메이션 (동두천)	뮤지컬	2018.12.01	2018.12.02	동두천시민회관	동두천시민회관	550

공연명과 장르, 공연일자, 장소, 객석수

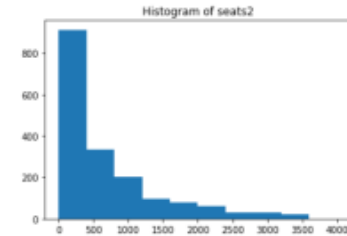
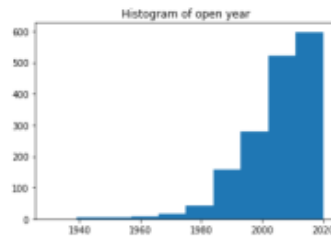
KBS홀	KBS홀 [부산]
세종문화회관	한국소리문화의전당

데이터를 합치는 중 같은 이름의 소속 기관의 경우 분류가 섞여 결과값이 정확하게 나오지 않았습니다. 같은 이름의 세부사항을 삭제하고 한 카테고리 묶으니 지역 분류에 적합하지 않아 복원하였습니다.

## 2. 각 자료의 주요 변수 기초 분석

### A. 공연시설

	n_seats	n_seats2	open
count	1846.000000	1846.000000	1846.000000
mean	2030.112134	581.122969	1777.940412
std	11466.726770	3337.307978	634.006376
min	0.000000	0.000000	0.000000
25%	120.750000	100.000000	1993.000000
50%	400.000000	201.500000	2005.000000
75%	1029.000000	480.000000	2012.000000
max	151495.000000	100000.000000	2020.000000



시설 좌석 수, 공연장 좌석 수, 개관연도

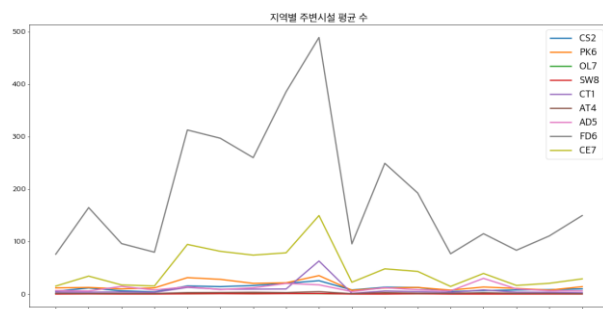
	pyuneui	jangaen	jucha	count
0.0	0.0	0.0	0.0	289
	0.0	1.0	1.0	340
	1.0	0.0	0.0	9
	1.0	1.0	1.0	348
1.0	0.0	0.0	0.0	67
	0.0	1.0	1.0	199
	1.0	0.0	0.0	9
	1.0	1.0	1.0	585

	서울	경기	부산	대구	경남	강원	전북	경북	충남	전남	대전	광주	인천	충북	제주	울산	세종
city	724	234	107	92	89	75	70	70	65	57	55	50	49	40	30	29	10

지역별 공연시설 수

주요시설 유무 별 공연시설 수

### B. 주변시설

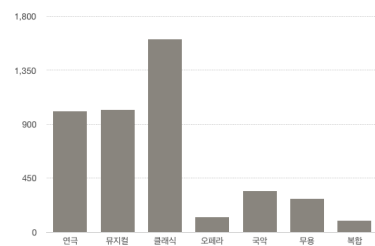


지역별 시설 평균 개수

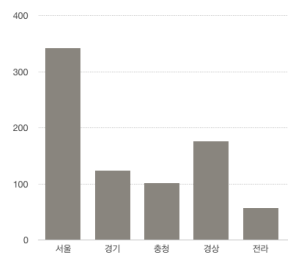
	CS2	PK6	OL7	SW8	CT1	AT4	AD5	FD6	CE7
city									
강원	402.0	882.0	52.0	0.0	270.0	97.0	496.0	5693.0	1137.0
경기	2709.0	2938.0	127.0	66.0	835.0	239.0	1364.0	38541.0	7995.0
경남	548.0	877.0	70.0	4.0	358.0	110.0	1300.0	8553.0	1541.0
경북	294.0	824.0	68.0	1.0	225.0	69.0	542.0	5580.0	1081.0
광주	772.0	1563.0	38.0	24.0	630.0	139.0	693.0	15630.0	4727.0
대구	1299.0	2578.0	110.0	73.0	843.0	270.0	785.0	27311.0	7490.0
대전	892.0	1118.0	64.0	26.0	512.0	183.0	655.0	14293.0	4079.0
부산	2185.0	2290.0	137.0	92.0	1031.0	299.0	2170.0	41233.0	8408.0
서울	18397.0	25402.0	587.0	815.0	45503.0	3280.0	12730.0	353426.0	108140.0
세종	78.0	72.0	4.0	0.0	13.0	7.0	47.0	956.0	225.0
울산	377.0	334.0	18.0	0.0	169.0	81.0	350.0	7225.0	1393.0
인천	613.0	617.0	38.0	31.0	229.0	77.0	402.0	9434.0	2114.0
전남	276.0	425.0	53.0	0.0	175.0	70.0	346.0	4378.0	825.0
전북	465.0	938.0	77.0	0.0	552.0	180.0	2091.0	8050.0	2748.0
제주	232.0	322.0	23.0	0.0	105.0	39.0	296.0	2503.0	492.0
충남	547.0	485.0	54.0	3.0	214.0	70.0	407.0	7198.0	1331.0
충북	394.0	562.0	54.0	0.0	100.0	23.0	238.0	5982.0	1158.0

지역별 시설 수

### C. 공연



공연장르 별 공연 수



지역 별 공연 수

### 3. 최종 Data Snapshot

name		n_theater	n_seats	theater_name	n_seats2	attr	open		
(재)경기문화재단		1	154	다산홀	154	공공(기타)	2001		
2001 아울렛키즈홀 [구로]		1	100	2001 아울렛키즈홀	100	민간(대학로 외)	2017		
2001 아울렛키즈홀 [구로]		1	100	2001 아울렛키즈홀	100	민간(대학로 외)	2017		
open	city	gu	address				pyuneui	jangaein	
2001	경기	수원시	경기도 수원시 팔달구 인계로 178 (인계동)				1	1	
2017	서울	구로구	서울특별시 구로구 중앙로1길 36 (고척동)				1	1	
2017	서울	구로구	서울특별시 구로구 중앙로1길 36 (고척동)				1	1	
jangaein	jucha	CS2	PK6	OL7	SW8	CT1	AT4	AD5	FD6
1	1	25	27	1	0	12	2	25	464
1	1	14	16	2	0	1	0	2	237
1	1	14	16	2	0	1	0	2	237
FD6	CE7	show			genre	start	end		
464	78	경기인형극제, 어린왕자			연극	20180717	20170718		
237	38	크리스마스 가족음악회			클래식	20191223	20191223		
237	38	춤추는 호랑이와 해님달님			뮤지컬	20191231	20190105		

변수명	변수 설명	자료 형태
name	공연장 이름	str
n_theater	공연장 내 공연장 수	int
n_seats	공연시설 총 좌석 수	int
theater_name	공연시설 내 공연장 이름	str
n_seats2	공연장 내 공연장 좌석 수	int
attr	공연시설 특징	str
open	공연시설 개관 연도	int
city	공연시설 위치(시)	str
gu	공연시설 위치(구)	str
address	공연시설 주소	str
CS2	편의점	int
PK6	주차장	int
OL7	주유소	int
SW8	지하철역	int
CT1	문화시설	int
AT4	관광명소	int
AD5	숙박	int
FD6	음식점	int
CE7	카페	int
genre	공연장르 (연극, 뮤지컬 등)	str
show	공연명	str
start	공연 시작 날짜	str
end	공연 종료 날짜	str

공연시설 변수 설명 코드북

### 4. 추후 진행 방향

1 차 계획서를 작성할 때에는 없었던 기간별/지역별/장르별/가격대별/공연통계가 신설되어 유용하게 쓰일 수 있는 데이터가 있다면 추가할 계획입니다. 크롤링 하다보니 공연장별 위도/경도 데이터가 유용할 수 있을 것 같아 추가할 계획입니다.

오류가 있거나 전처리가 필요한 데이터를 수정해 나갈 계획입니다.