데이터시각화 텀프로젝트

201811139 강호정

201818162 강진우

201818945 김선재

- 자료소개 -

- 자료 내용

2021년도의 국내 박물관 시설의 시/도, 개관년월일, 관람인원, 관람료 등을 정리한 현황 통계 데이터

- 선정 기준

데이터를 선정하기에 앞서, 조원들이 공통적으로 '분석이 용이한 데이터보단, 데이터 분석에 목적이 있는 데이터를 찾자.' 라고 의견이 모아짐. 그렇게 데이터를 찾는 중 위 데이터를 발견하였고, 문득 대한민국에서 관람객이 제일 많은 박물관은 어디일까, 관람객은 어떤 요인에 영향을 받아서 변하는 걸까 의문을 갖게 되었다.

- 분석 목적 및 의미 –

- 어떠한 변수들이 일평균 관람인원에 영향을 미치는지 알아보고, 관람인 이 많은 곳 또는 적은 곳 원인을 분석.

- 관람객이 많은 박물관 , 지정문화재가 있는 박물관 , 국보가 있는 박물관의 위치를 지도에 표시 및 밀도 표현

- 이 자료 분석을 통해 관람객수에 영향을 미치는 요소들과 유명한 박물관의 분포를 알아내어 추후 박물관 관리자들의 경영지표로 활용가능

- 변수 소개 –

이름/시도/도로명주소/일,연 평균 관람 인원 / 개관년월일 / 국립,사립,공립,대학/ 오디오가이드 제공여부 / 건물면적 / 프로그램총계 / 관람료 등 총 28개의 변수.

-데이터 전처리-

- index 이용하여 필요 없는 열 지움(일반 관람료 / 특별전시 관람료 / 자원봉사자 등 총 6개)

- character형식으로 되어있는 개관년월일 데이터를 format함수를 사용해 날짜 형식으로 바꾼 후, 년도만 추출함 (수치형 변수로 변경)

- 반응변수 일평균 관람인원에 따라 내림차순으로 정리

```
     이름               시도            도로명주소         국립.사립.공립.대학    개관년월일
Length:869         Length:869         Length:869         Length:869         Min.   :1935
Class :character   Class :character   Class :character   Class :character   1st Qu.:2000
Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median :2006
                                                                            Mean   :2003
                                                                            3rd Qu.:2011
                                                                            Max.   :2019
                                                                            NA's   :22
 오디오가이드.제공여부     건물면적          전시실.면적      전시실.유물.교체횟수 문화상품점.면적
Length:869         Min.   :    51.3   Min.   :     0   Length:869         Min.   :   0.00
Class :character   1st Qu.:   721.0   1st Qu.:   330   Class :character   1st Qu.:  16.50
Mode  :character   Median :  1681.0   Median :   677   Mode  :character   Median :  38.85
                   Mean   :  4998.3   Mean   :  1886                      Mean   : 114.67
                   3rd Qu.:  3817.5   3rd Qu.:  1287                      3rd Qu.:  87.00
                   Max.   :583288.0   Max.   :220571                      Max.   :6600.00
                   NA's   :28         NA's   :11                          NA's   :626
   매점면적          주차대수        소장유물종류       소장유물.개수.개설.이후.
Length:869         Length:869         Length:869         Min.   :     1.0
Class :character   Class :character   Class :character   1st Qu.:   600.5
Mode  :character   Mode  :character   Mode  :character   Median :  2155.0
                                                         Mean   : 14696.6
                                                         3rd Qu.:  7918.0
                                                         Max.   :964079.0
                                                         NA's   :62
 지정문화재.등.명칭 지정문화재.등.개수 기획.특별전.연.횟수 프로그램.총계       연개관일수
Length:869         Length:869         Length:869          Min.   :  0.00   Min.   :  0.0
Class :character   Class :character   Class :character    1st Qu.:  3.00   1st Qu.:293.0
Mode  :character   Mode  :character   Mode  :character    Median :  6.00   Median :310.0
                                                          Mean   : 10.04   Mean   :297.3
                                                          3rd Qu.: 11.75   3rd Qu.:315.0
                                                          Max.   :331.00   Max.   :365.0
                                                          NA's   :175      NA's   :26
   연관람인원       일평균.관람인원     일반관람료     특별전시관람료   학예인력.학예직.총인원
Min.   :     0   Min.   :    0.0   Min.   :    0   Min.   :    0   Min.   : 0.000
1st Qu.:  8810   1st Qu.:   30.0   1st Qu.: 2000   1st Qu.: 2000   1st Qu.: 1.000
Median :  30000   Median :   98.0   Median : 3750   Median : 4000   Median : 1.000
Mean   : 114183   Mean   :  352.4   Mean   : 4889   Mean   : 4464   Mean   : 3.084
3rd Qu.: 101860   3rd Qu.:  325.9   3rd Qu.: 6000   3rd Qu.: 5250   3rd Qu.: 3.000
Max.   :4000000   Max.   :10959.0   Max.   :20000   Max.   :13000   Max.   :88.000
NA's   :31        NA's   :31        NA's   :513     NA's   :757     NA's   :573
 일반인력.일반직.공무원  자원.봉사자        관람료          인력
Min.   :  1.000       Min.   :  0.00   Min.   :    0   Min.   :  0.00
1st Qu.:  2.000       1st Qu.:  6.75   1st Qu.:    0   1st Qu.:  0.00
Median :  3.000       Median : 20.00   Median :    0   Median :  0.00
Mean   :  6.239       Mean   : 73.24   Mean   : 2578   Mean   : 17.86
3rd Qu.:  7.000       3rd Qu.: 52.00   3rd Qu.: 4000   3rd Qu.:  5.00
Max.   :128.000       Max.   :886.00   Max.   :26000   Max.   :892.00
NA's   :547           NA's   :697
```

- summary를 이용하여 요약표를 보고 수치형/범주형 변수로 나눴고 전체적으로 수치형 변수의 최대값이 유독 큰 것을 확인할수 있었다 (이상치 존재 할 것이라 추정)

- 전체적으로 모든 변수의 결측값(빈칸)이 많다는 점을 볼 수 있다

- **수치형 변수**

  건물면적/전시실 면적/문화상품점 면적/프로그램총계/연개관일수/

  개관년월일/관람료/인력/소장유물개수/연관람인원/일평균 관람인원/

  일반인력 일반직 공무원/자원봉사자/학예인력 학예직 총인원/

  특별전시관람료/일반관람료 (16개)

- **지역변수형**

  -도로명주소 (1개)

- **범주형(텍스트)**

  국립 사립 공립 대학/오디오가이드 제공여부/

  전시실 유물 교체횟수/이름/시도/주차대수/소장유물 종류/

  소정유물 개수 개설 이후/지정문화재 등 명칭/지정문화재 등 개수/

  기획 특별전 연 횟수/(11개)

- **반응변수(y)**

  일평균 관람인원

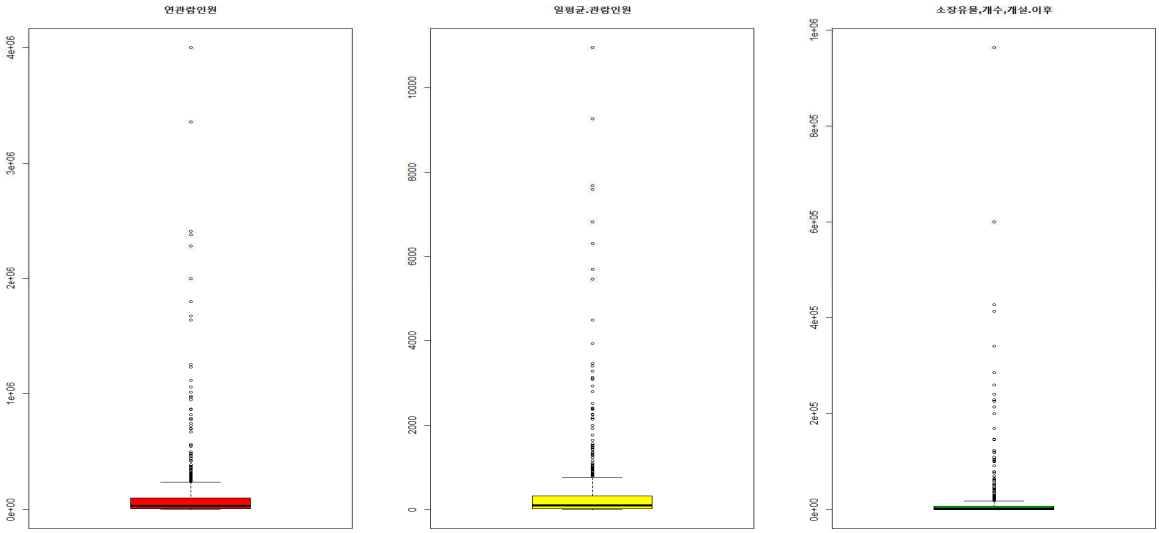
- **데이터 시각화 (수치형) –**

Summary를 보았을 때 결측값이 많았고 이상치가 있을거라 추측했는데,

그래서 이상치를 잘 보여줄수 있는 boxplot을 선택했다. 수치형 변수들의 결측값을 없애고, 새로운 변수에 저장하여 각각 그려보았다.

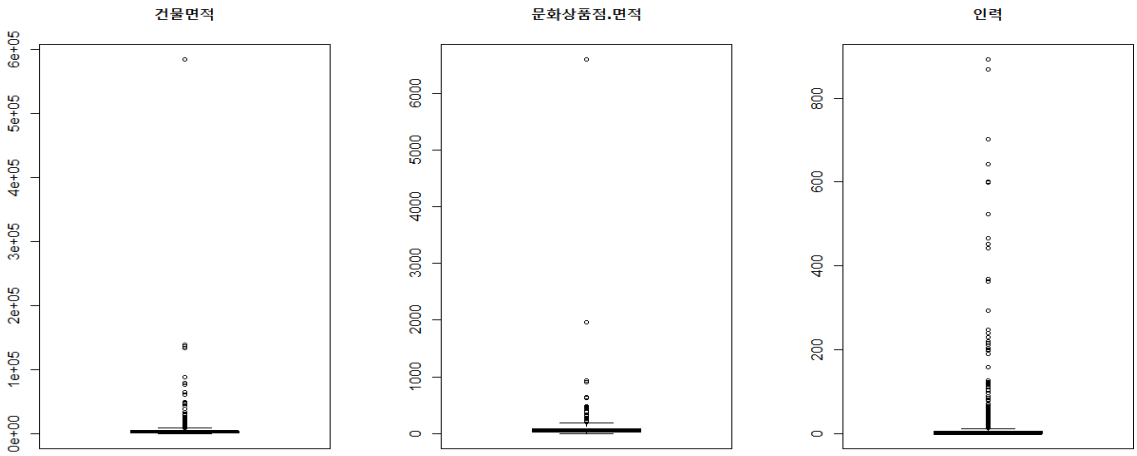
(결측값이 많아서 생기는 왜곡 현상 방지)

연관람인원 / 일평균 관람인원 / 소장유물 개수 개설 이후 (#1.1)



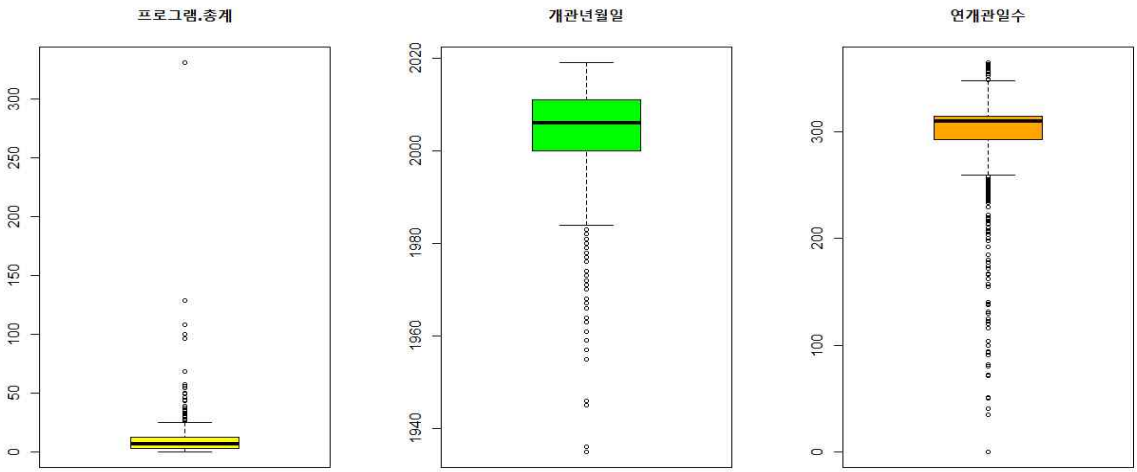
- 3개 그래프 모두 낮은 숫자에 분포가 많은 것을 볼 수 있고 분포가 많은 곳 말고도 점이 많은 것을 볼수있다 (이상치 다수 존재)

건물면적 / 문화상품점 면적 / 인력 (#1.2)



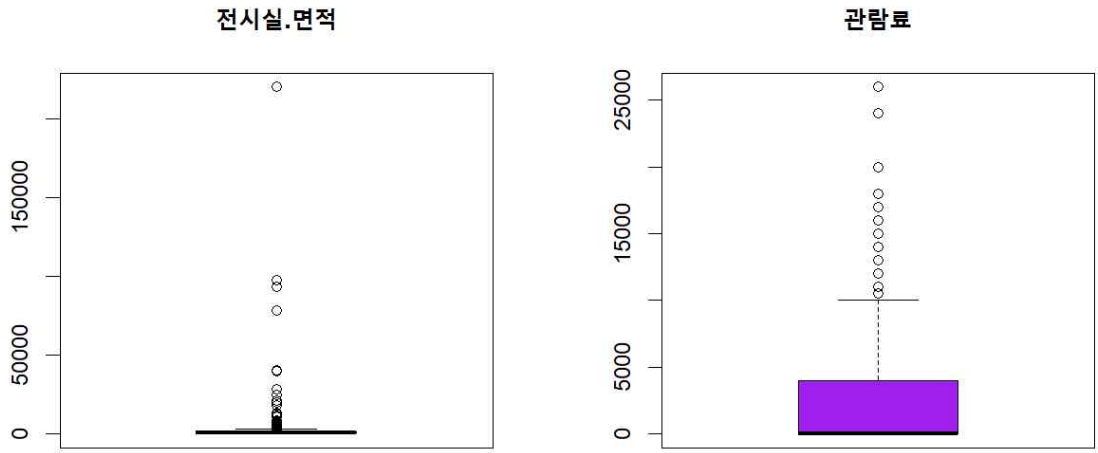
- 3그래프 모두 낮은 숫자에서 가장 많은 분포를 보여줬고 건물면적과 문화상품점 면적은 낮은 쪽에 많이 분포하는 반면에 인력의 이상치 분포는 다양한 수치로 분포했다(면적은 거의 비슷)

프로그램 총계 / 개관년월일 / 연개관일수 (#1.3)



- 프로그램 총계는 낮은 숫자에 많이 분포하고 수치가 크게 차이는 없었다.
- 개관연월일과 연개관일수의 분포는 높은 수치에 가장 많은 데이터가 분포했다 (2000년~2010년대에 개관한 박물관이 많으며 / 대부분 박물관의 연개관일수는 약 300일이었다.)

전시실 면적 / 관람료 (#1.4)



- 전시실 면적은 낮은쪽에 대부분 분포 / 관람료는 거의 0~5000 이다

- 수치형 boxplot 총 해석

이상치가 있는 부분이 많았고, 거의 대부분의 수치형 변수 이상치 부분들이

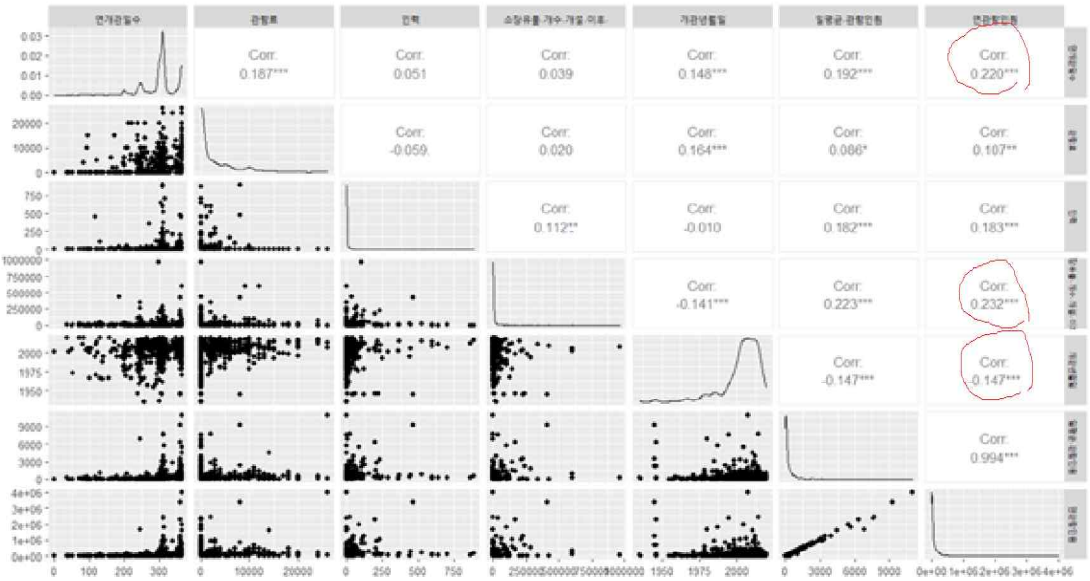
겹쳤는데, 특이하게 개관년월일 , 연개관일수 데이터는 이상치가 반대로 되어있는 것을 볼 수 있었다.

→ 새로운 것이 많을 신생 박물관에 사람이 많이 갈 것이라고 추측했지만

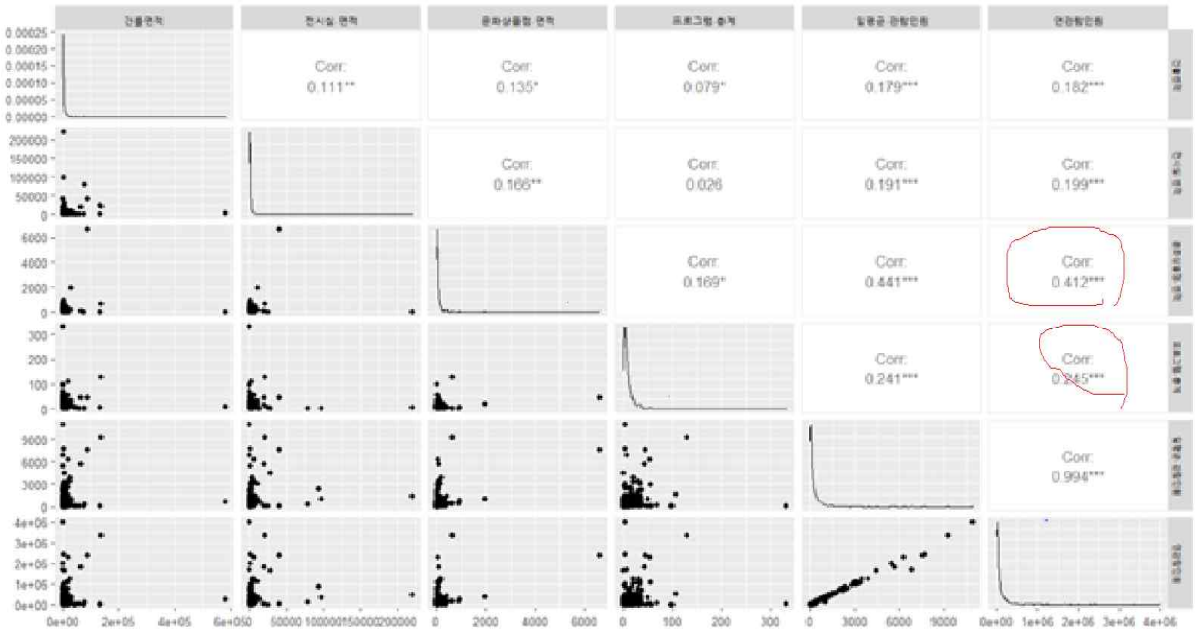
역사가 깊거나 연개관일수가 긴 박물관에 사람들이 많이 갈 것 이라고 그래프를 통해 추론했습니다.

- 수치형 상관관계 시각화 –

수치형 상관관계를 확인 하기위해 ggpairs 함수를 이용해 산점도와 상관계수를 확인 해 보았다.



- 연개관일수 (0.192) , 소장유물 개수 개설 이후(0.233) , 개관년월일(음의상관관계/-0.147) 이 반응변수인 일평균관람인원과의 관계에서 의미 있어 보인다.



- 문화상품점 면적(0.441), 프로그램 총계(0.241)가 반응변수인 일평균관람인원과의 관계에서 의미 있어 보인다.

Box plot과 상관관계 그래프를 통해 의미있어 보이는 5개의 수치형 변수들을 뽑아 보았다.

→ 5개의 수치형 변수를 의미있는 범주형 변수와 묶어 확인할 예정

- 중간 정리-

- 의미 있는 수치형 변수

문화상품점 면적 / 프로그램총계 / 소장유물 개수 개설 이후 /

개관년월일(음의상관관계) / 연개관일수

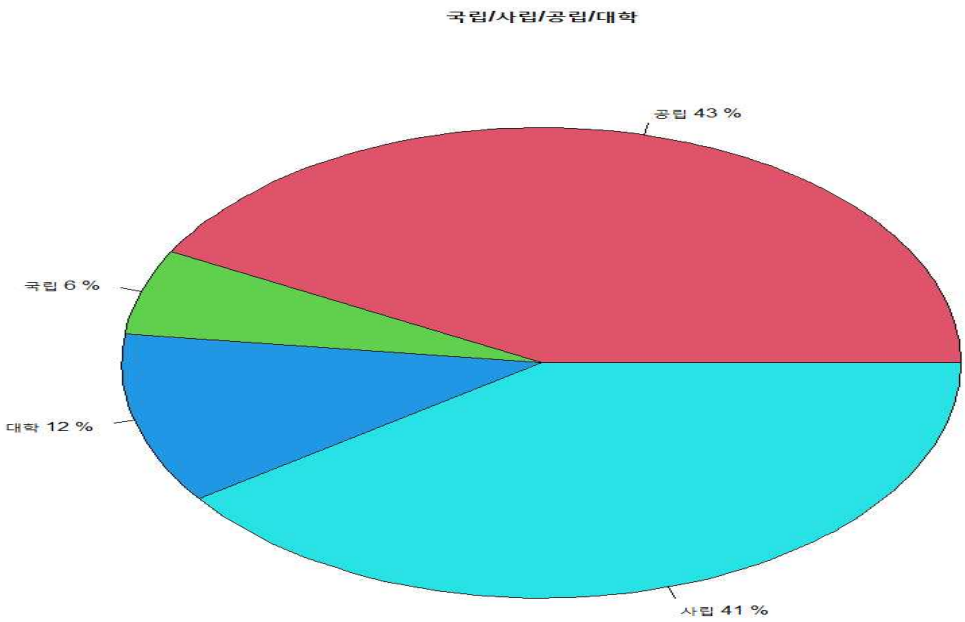
- 의미 있는 범주형 변수

시도 / 사립 공립 대학 국립 / 오디오

→ 수치형은 x 축 / y축은 일평균관람인원 / 범주형은 색으로 표현하여 나타내볼 것이다

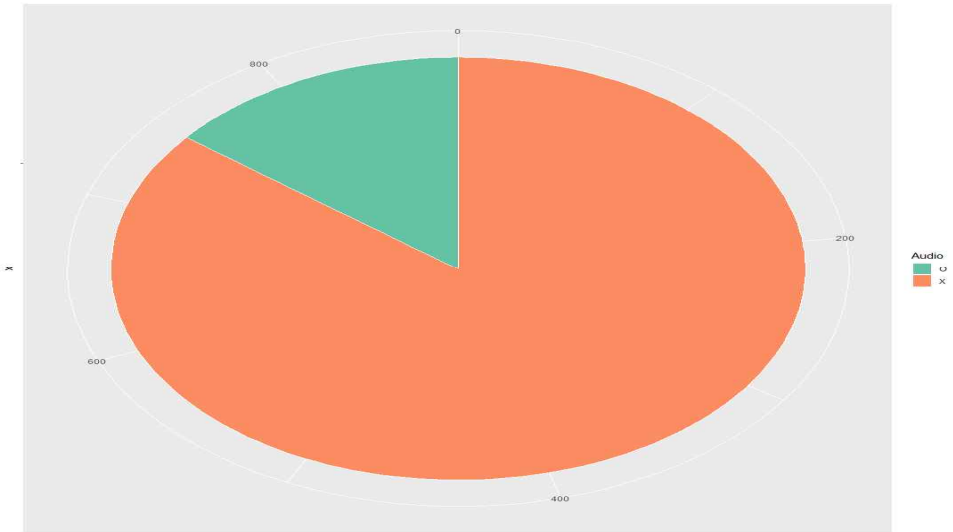
- 범주형 시각화– #2-1

사립 / 공립 / 대학 / 국립, 범주를 파이차트로 시각화 해보았다.



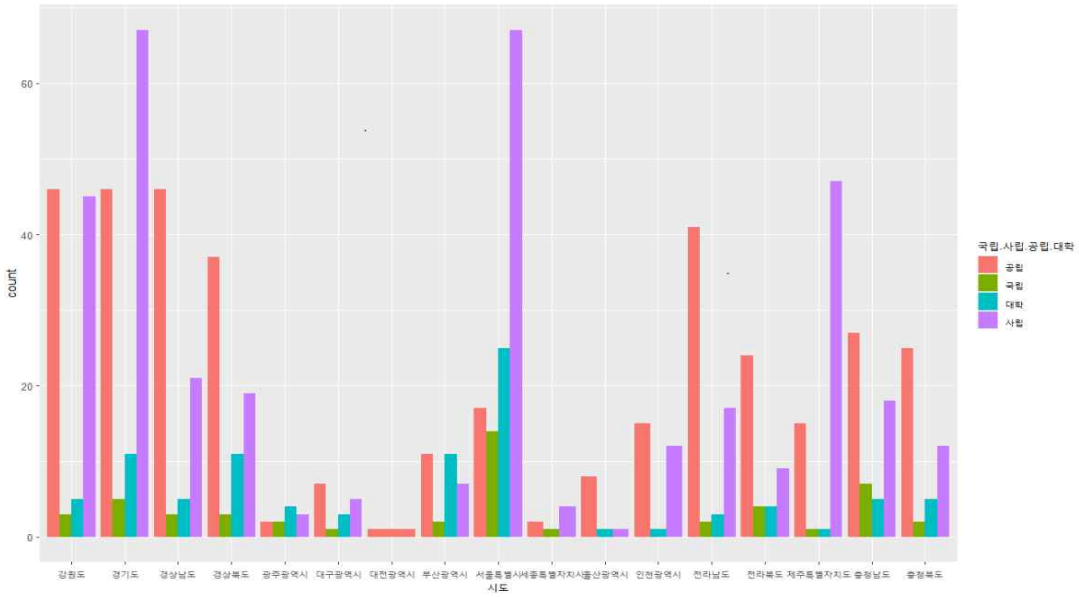
→ 공립>사립>대학>국립 순서로 많음을 알 수 있었다. (사립, 공립이 전체의 대부분을 차지)

오디오 가이드 유무를 파이차트를 통해 나타내 보았다. (#2.2)



→ 오디오 가이드 유무의 비율은 x가 O에 6배정도 된다는 것을 알 수 있었다.

시도는 범주를 여러 개의 범주를 잘 표현해 줄 수 있는 막대그래프로 범례에는 국립.공립.사립.대학 을 나타내어 시각화 해보았다. (#2.3)



12개의 막대차트를 보고 박물관 수가 많은곳 3개까지 정리해보았다.

-공립: 경남 = 경기 = 강원도 (3지역 거의 비슷 / 오히려 서울은 적었다 /)

-사립: 경기도 > 서울 > 제주도 (압도적 1,2,3등 / 수도권 지역)

-대학: 서울 > 경북 > 경기도 (서울이 압도적1등 / 서울에 대학이 많아서)

-국립: 서울 > 충남 > 경기도

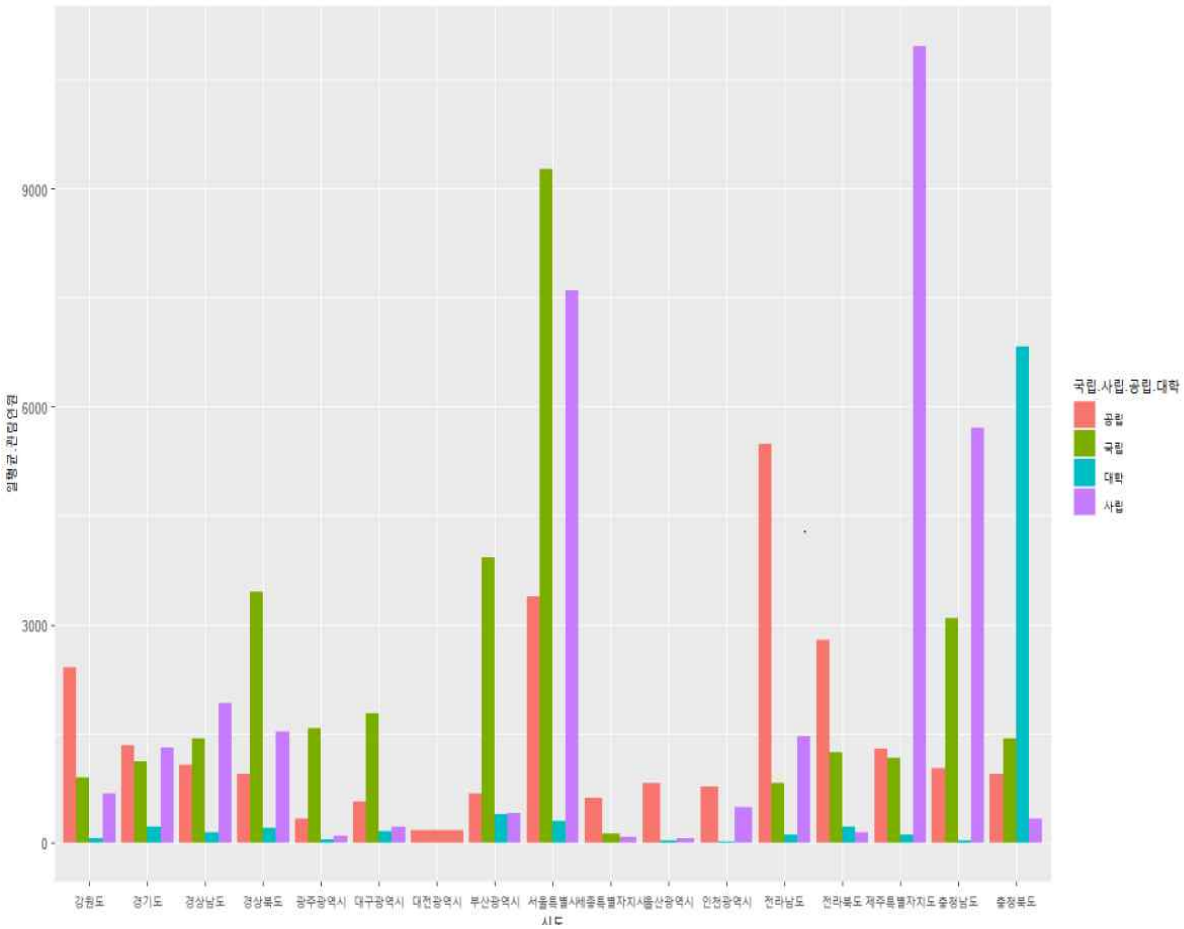
→ 지역 별 국립/사립/공립/대학의 비율과 분포를 알 수 있었는데 전체적으로 수도권 지역에 박물관이 많이 있는 것을 볼 수 있다.

범주형 시각화 중간정리

- **➔ 공립 > 사립 > 대학 > 국립 순서로 박물관이 많다**
- **➔ 오디오가이드 제공 안해주는 곳이 해주는 곳보다 훨씬 많다**
- **➔ 전체적으로 수도권 지역에 박물관이 많은데 수도권이 아닌 지역도 박물관 수가 많은 곳이 있다 (충남,경남,제주 등)**

-반응변수를 고려한 범주형 시각화 – (#3.1)

앞에서 그린 시도별 국립.공립.사립.대학을 나타낸 막대그래프에 일평균 관람인원을 추가해서 그려보았다.



- 사립, 공립, 대학, 국립에 따른 데이터 시도 박물관 수 분포에서 공립은 경남, 경기, 강원도가 박물관 수가 많았는데, 일평균 관람인원은 전남, 서울, 전북 순으로 전혀 다른 곳이 관람인원이 많다는 것을 알 수 있었다.

→ 공립 박물관 수는 일평균 관람인원과 크게 비례하지 않음을 알 수 있다

- 사립은 경기도, 서울, 제주도 순으로 많았고, 다른 지역은 낮은 비율을 띄었는데 일평균 관람객수는 제주, 서울, 충남 순으로 나왔다.

→ 충남이 박물관수는 적지만 의외로 일평균 관람인원은 많다는 것을 알 수 있다

- 대학은 서울, 경북, 경기도처럼 큰 도시에 박물관이 많았는데

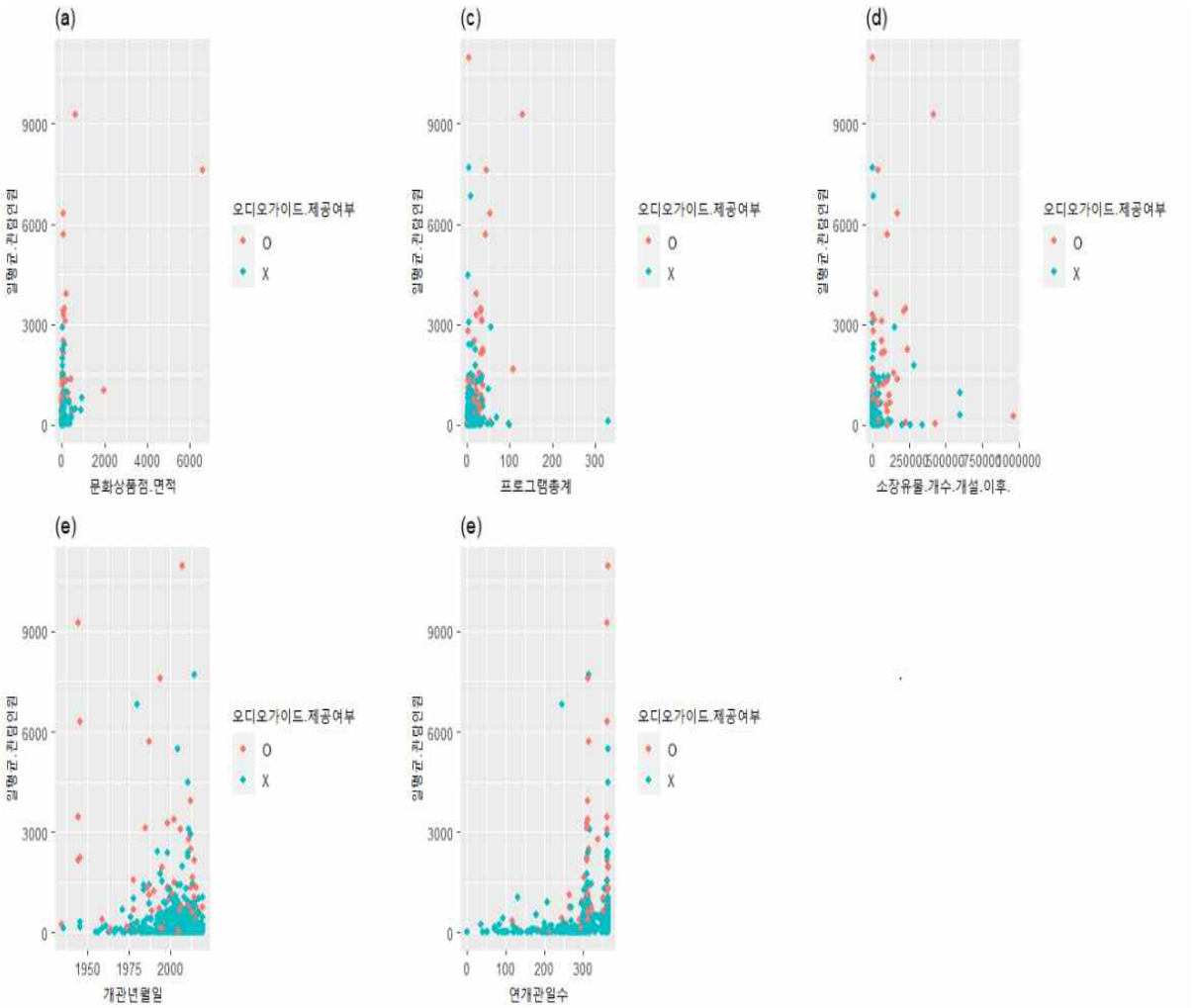
관람인원은 오히려 작은 도시인 충북이 가장 많은 것을 알 수 있다.

- 국립은 서울 > 충남 > 경기도 순으로 박물관이 많았는데 관람인원은 서울에 압도적으로 많은 것을 볼 수 있다.

의미 있는 수치형 변수들(x) 별 일평균 관람인원(y) , 오디오가이드 제공여부

의미 있다고 보여지는 수치형 변수들에 따른 일평균 관람인원을

산점도로 한 화면에 나타내어 분석하였다**. 분석 산점도(#3.2)**



- (a) 문화상품점 면적과 일평균 관람인원이 적은 곳에 오디오 가이드가 제공되지 않는 양상을 띄고, 오디오 가이드가 제공되는 곳은 인원이 많은 것으로 보인다.

- (b) 프로그램 총계는 일평균 관람인원과, 프로그램 총계 두 수치 모두 낮은 쪽에 오디오 가이드가 제공되지 않는 양상을 띄고, 오디오 가이드가 제공되는 곳에 인원이 많은 것으로 보인다.

- (c) 소장 유물 개수 또한 적은곳에는 사람밀도와 오디오가 거의 제공되어지지않는 모습을 띄고 오디오가이드가 제공되어지는곳에 사람이 많은것처럼 보인다

→ 위 3개 그래프의 문화상품점 면적 , 프로그램총계 , 소장유물 개수가 적을수록 일평균 관람인원도 적은 분포를 보이며 오디오도 제공하지 않는 것처럼 보인다.

- (e) 개관년월일은 다른 변수들과 반대로 분포는 최신건물에 많지만,

사람수가 많은 곳은 오디오 가이드를 제공해주는 것처럼 보인다.

→ 일평균 관람인원이 2000이하인 신생박물관은 오히려 오디오 가이드를 제공해주지 않는곳이 많다

- (e2) 연개관일수는 전체적으로 넓고 고른 분포를 띄고, 일평균 개관일수가 높은 곳에만 거의 오디오 가이드를 제공해주는 것처럼 보인다

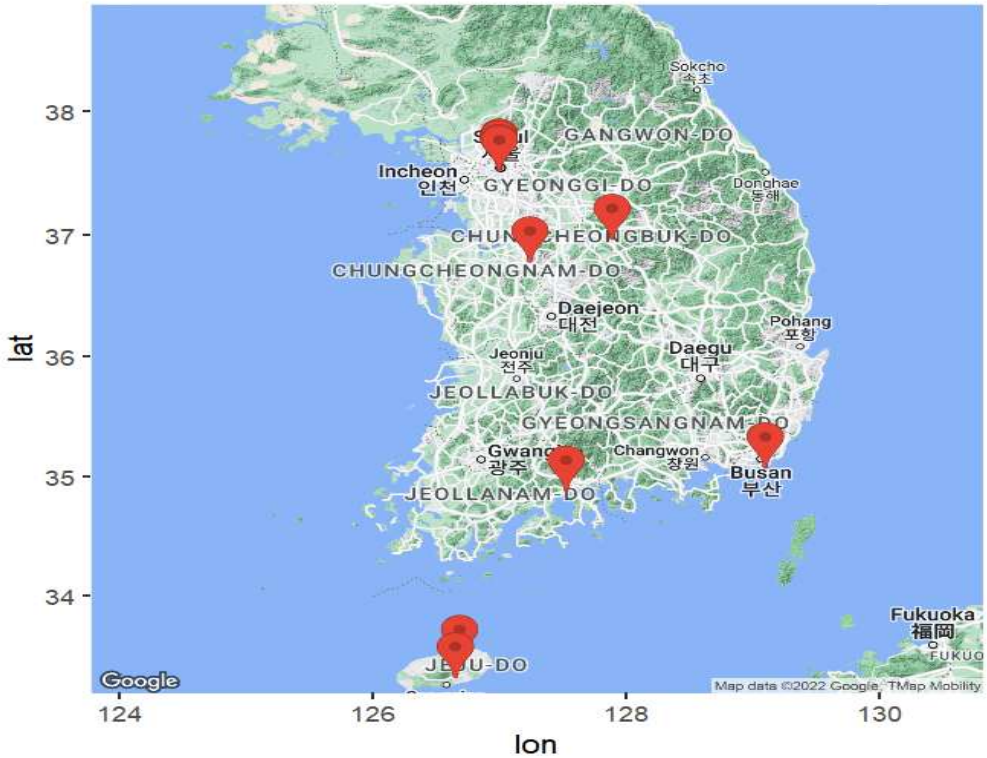
→ 연 개관일수가 많은 곳에 오디오를 제공해 주는곳이 많이 분포해있다.

지도 시각화

지도 (#4.1)

일평균관람인원을 기준으로 한 상위 10개의 박물관의 위치를 지도에 표현했다.

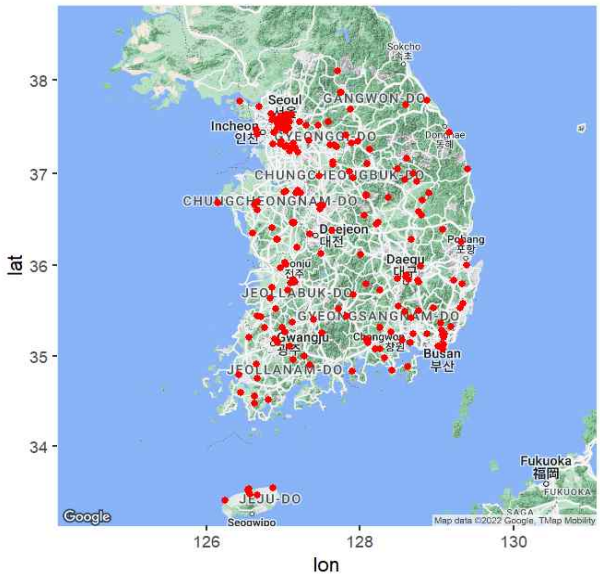
→서울 4(국립) , 제주2(사립), 충북1(대학), 전남1(공립), 충남1(사립) , 부산1(국립)으로 찍혔다.

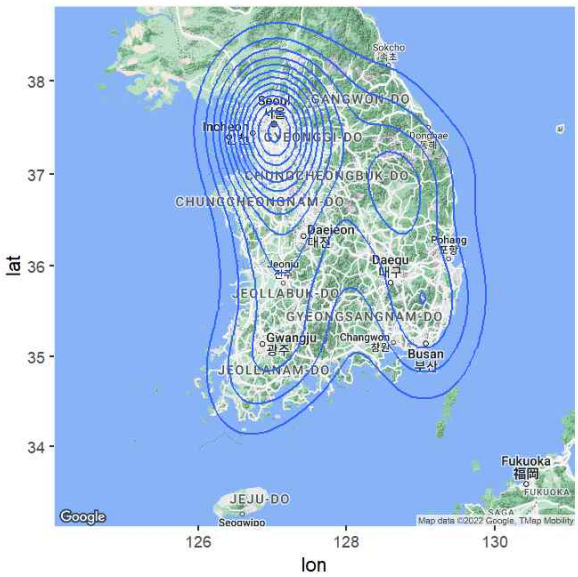


지도 (#4.2)

지정문화재가 있는 박물관의 위치를 지도에 표현했고, 밀도를 표현했다.

→ 서울/경기에 지정문화재가 있는 박물관이 몰려있다.

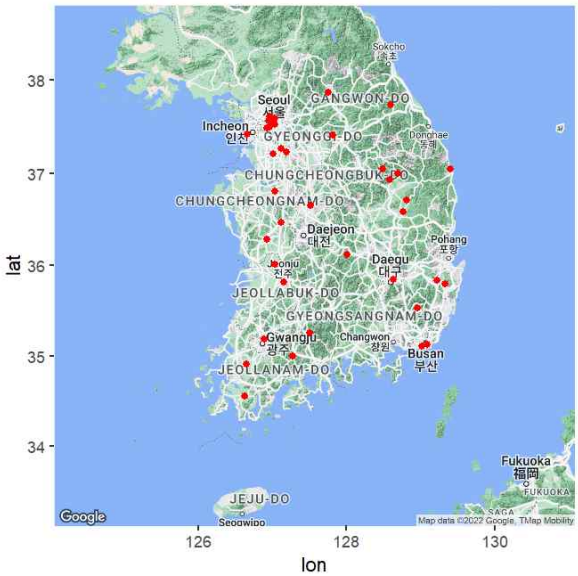


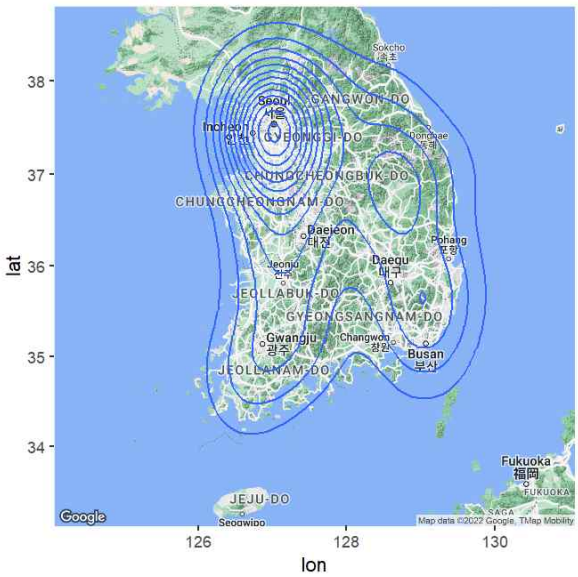


지도 (#4.3)

국보가 있는 박물관의 위치를 지도에 표현했고, 밀도를 표현했다.

→ 서울/경기에 국보가 있는 박물관이 몰려있다.





- 결론 -

- 오디오 가이드는 대부분 제공해주지 않지만, 일평균 관람인원이 많이 분포하는 곳은 대부분 제공해줬다.

- 관람객수는 신생박물관보다 오래된 역사가 깊은 특정 박물관에 많았다.

- 서울이나 경기도 같은 대도시에 사립,공립,대학,국립 박물관수가

전체적으로 많았는데, 지방에도 관람객수가 탑 10안에도 들 정도로 많은 곳이 있었다.

- 문화예술부분에서도 서울 쏠림 현상을 막기 위해 관람객이 많은 지방 박물관들이 위 통계 지표 데이터를 활용하여 박물관을 유지 및 개선 시킬 필요가 있다.