2024학년도 프로젝트 계획서

도서 대출량에 영향을 주는 환경 요인 분석

2024년 12월 01일

컴퓨터정보공학과

팀원	고태경(202244042)

인하공업전문대학
INHA TECHNICAL COLLEGE
仁荷工業專門大學

본 과제(결과물)는 인하공업전문대학 컴퓨터정보과 빅데이터 처리 교과목의 프로젝트 보고서 입니다.

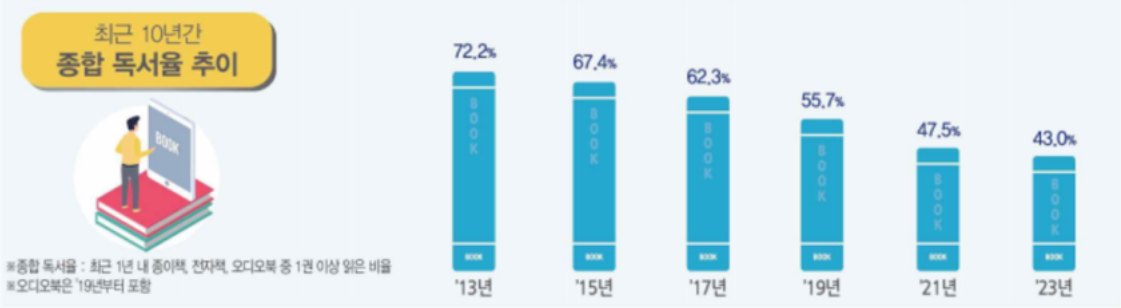
목 차

프로젝트 깃허브 주소

https://github.com/taegyeong0225/bigdata-processing

1. 프로젝트 개요 및 목표

2023년 국민 독서 실태조사에 따르면 성인의 최근 10년간 종합 독서율 추이는 계속 감소 하고 있는 추세이다.



<성인 국민 독서 실태조사 결과(2013~2023)> (출처 : 2023년 국민 독서 실태조사 2p)

이와 함께, 최근에는 MZ 세대의 문해력 부족이 사회적 논란이 되고 있다. 예를 들어, 2020년 8월 광복절 당시 월요일인 17일을 임시공휴일로 지정하면서 '사흘 연휴'라는 표현을 사용하였는데, 이를 '4일 연휴'로 오해하거나, 왜 '3일 연휴'를 '사흘'로 표현했는지 묻는 사례가 있었다. 또한, 2022년 8월에는 웹툰 작가 변덕이 사인회 예약 안내문에서 사용한 '심심한 사과'라는 표현을 본래 뜻인 '매우 깊고 간절하다'로 이해하지 못하고, '지루하고 재미없는 사과'로 해석한 사례가 있었다.

2024년 10월 7일 발표된 '학생 문해력 실태 교원 인식 설문조사'에 따르면, 교사들은 학생들의 문해력 저하의 주요 원인으로 29.2%가 '**독서 부족**'을 꼽았다. 또한, 문해력 개선 방안으로는 32.4%가 '**독서활동 강화**'를 지목한 것으로 나타났다.

출처 : 학생 문해력 실태 교원 인식 설문조사 결과 발표(2024.10.07.), 한국 교원 단체 총연합회

이처럼 문해력 부족 문제가 논란이 되고 있는 가운데, 독서를 장려해야 한다는 사회적 공감대가 형성되고 있다. 이에 따라 문화체육관광부는 2024년 4월 22일에 '제4차 독서문화진흥 기본계획(2024~2028)'을 발표하며, 독서 활성화를 위한 정책적 노력을 강화하고 있다.

책을 접하기 가장 쉬운 장소 중 하나는 도서관이다. 본 프로젝트에서는 지역별 전국 도서관 도서 대출 수 데이터와 다른 공공 데이터를 추가로 결합하여 도서 대출량에 영향을 미치는 환경 요인을 분석하고자 한다. 또한, 이 결과로 대출량 및 독서율을 높일 수 있는 구체적인 방안을 제안하는 것을 목표로 하고 있다.

해당 프로젝트에 앞서 세 가지 가설을 설정하였다.

가설 1. 도서관의 수와 도서 대출 수는 비례할 것이다.

도서관이 많은 지역은 도서관을 접할 기회가 많기 때문에 도서 대출 수가 정비례할 수 밖에 없을 것이다.

가설 2. 도서 대출 수는 학원 같은 교육 시설 수에 영향을 가장 많이 받을 것이다.

교육 시설이 많은 지역에서 독서 교육을 더 많이 받았을 것이며, 도서 대출을 많이 할 것이다.

가설 3. 문화시설 수가 많은 곳일 수록 도서 대출 수가 높을 것이다.

문화 생활, 독서 모두 취미 생활 중 하나이므로, 문화 시설 수가 많을수록 도서 대출량이 높을 것이다.

가설 4. 접근성이 높을 수록 도서 대출 수가 높을 것이다.

도서관이 주변에 없거나 가기 힘들다면 도서 대출을 하기가 꺼려질 수 있기 때문에, 접근성이 높은 수록 도서 대출 수가 높을 것이라고 생각했다.

2. 데이터 수집

전반적으로 공공데이터를 활용하여 데이터 분석을 진행하였다.

- 도담(2022 제1호), 지역별 대출건수 (2022.01~2022.05)

- 전국도서관표준데이터.csv
  2024-08-16까지의 데이터 기준 일자를 가진 데이터가 있음

- 행정경계(시도)
  국토지리정보원 연속수치지형도 행정경계 데이터

- 국토교통부_전국 버스정류장 위치정보

- 디지털 문화역세권 (2022)

- 소상공인시장진흥공단_상가(상권)정보_20220630

3. 프로젝트 범위

데이터 수집 -> 데이터 가공/정제 -> 데이터 분석 -> 데이터 시각화 -> 분석 결과

프로젝트를 위해 필요한 데이터들은 공공데이터를 수집하여 분석을 진행한다. 외부 환경과 도서 대출에 대한 데이터 분석을 진행한 후 결과에 따른 독서율 상승 방안을 제시

4. 프로젝트 과정

1. 지역별 도서 대출 수 알아내기

데이터 전처리

'도담(2022 제1호)'를 통해 '지역별 도서 대출 수 데이터.csv'를 얻는다.

```
import pandas as pd

# 지역별 도서 대출 수 (2022.01~05)
file_path = '/content/drive/MyDrive/bigdata_processing/지역별_도서_대출_수.csv'
df = pd.read_csv(file_path, encoding='utf-8')
df
```

	월	서울	부산	대구	인천	광주	대전	울산	세종	경기	강원	충북	충남	전북	전남	경북	경남	제주	합계
0	1월	2416122	657623	578205	387130	231657	297452	210335	216464	3551797	206332	214415	387931	191199	230191	404976	633240	163884	10978953
1	2월	2101437	555337	489198	339060	201234	238150	183034	195021	3306579	180208	158257	328690	180417	193280	352435	526686	141915	9670938
2	3월	2324436	603978	530220	358794	211297	268950	198683	193148	3411516	182693	183714	353305	196776	216331	370998	547141	145756	10297736
3	4월	2206682	597172	513308	342656	207941	260514	161703	174087	3263590	182810	157633	342153	195313	219834	367276	542052	130859	9865583
4	5월	2022531	562069	489610	317549	196921	248177	152280	169920	3033461	159736	164709	318730	198123	181516	344051	478580	138185	9176148

데이터 전처리

- '월', '합계' 열은 필요없는 열이므로 삭제한다.
- '지역별 도서 대출 수'를 구하기 위해, 열별 합계를 구해 5번째 행에 추가한다.
- 5번째 행(지역별 합계가 구해진 행)만 남긴 데이터를 저장한다.

```
# 첫 열('월'), 마지막 열('합계') 삭제
df = df.drop(columns=['월','합계'])  # 없는 열은 무시)

# 열별 합계 계산
total_row = df.sum(axis=0)

# 월별 합계를 나타내는 5번째 행만 남기기
df = df.iloc[[4]].reset_index(drop=True)

# 결과 확인
print(df)
```

```
        서울      부산      대구      인천      광주      대전      울산      세종      경기  \
0  2022531  562069  489610  317549  196921  248177  152280  169920  3033461

       강원      충북      충남      전북      전남      경북      경남      제주
0  159736  164709  318730  198123  181516  344051  478580  138185
```

- 다른 데이터와 함께 분석할 때, 통일성을 주기 위해 지역과 관련된 column명은 공식적인 현재 행정명으로 사용하기로 결정했다.

```
# 새로운 열 이름 리스트
new_columns = ['서울특별시', '부산광역시', '대구광역시', '인천광역시', '광주광역시',
               '대전광역시', '울산광역시', '세종특별자치시', '경기도', '강원특별자치도',
               '충청북도', '충청남도', '전북특별자치도', '전라남도', '경상북도',
               '경상남도', '제주특별자치도']


# 열 이름 변경
df.columns = new_columns

# 결과 확인
print(df)
```

```
     서울특별시  부산광역시  대구광역시  인천광역시  광주광역시  대전광역시  울산광역시  세종특별자치시  \
0  2022531  562069  489610  317549  196921  248177  152280  169920  3033461

      경기도  강원특별자치도  충청북도  충청남도  전북특별자치도  전라남도  경상북도  경상남도  제주특별자치도
0  159736  164709  318730  198123  181516  344051  478580  138185
```

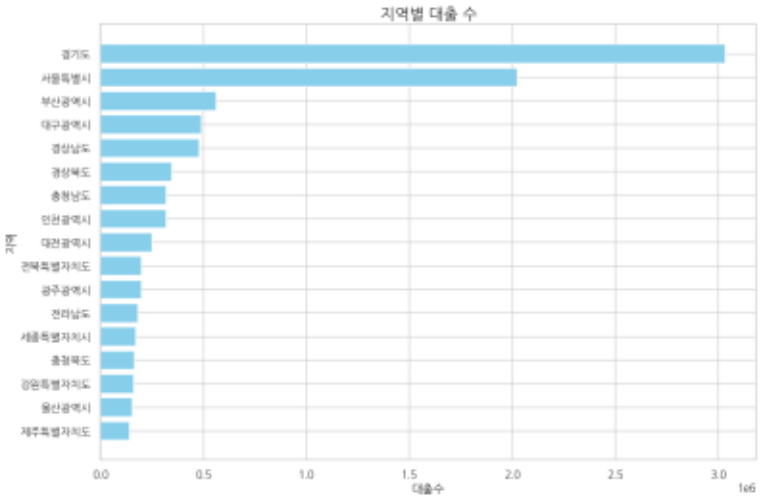
- 지도에 시각화 하기 위해 melt 함수를 이용해 wide format을 long format으로 변환한다.

```
# Wide Format -> Long Format 변환
df_long = df.melt(var_name='지역', value_name='대출수')

# 대출수 기준으로 내림차순 정렬
df_long = df_long.sort_values(by='대출수', ascending=False).reset_index(drop=True)

# 결과 확인
print(df_long)
```

```
          지역     대출수
0         경기도  3033461
1      서울특별시  2022531
2      부산광역시   562069
3      대구광역시   489610
4        경상남도   478580
5        경상북도   344051
6        충청남도   318730
7      인천광역시   317549
8      대전광역시   248177
9    전북특별자치도   198123
10     광주광역시   196921
11       전라남도   181516
12   세종특별자치시   169920
13       충청북도   164709
14   강원특별자치도   159736
15     울산광역시   152280
16   제주특별자치도   138185
```

데이터 시각화

맷플롯립을 이용하여 막대그래프로 시각화한 결과이다.




```
# matplotlib 라이브러리를 불러옵니다.
import matplotlib.pyplot as plt
import matplotlib.font_manager as fm
import matplotlib as mpl

plt.rcParams['font.family'] = 'NanumGothic'
print(plt.rcParams['font.family'], plt.rcParams['font.size'])   # 폰트확인

# 데이터 정렬
df_long_sorted = df_long.sort_values(by='대출수', ascending=False)

# 그래프 크기 설정
plt.figure(figsize=(12, 8))

# 막대그래프 생성
plt.barh(df_long_sorted['지역'], df_long_sorted['대출수'], color='skyblue')

# 그래프 제목 및 축 레이블 설정
plt.title('지역별 대출 수', fontsize=16)
plt.xlabel('대출수', fontsize=12)
plt.ylabel('지역', fontsize=12)

# y축 순서를 뒤집어서 가장 높은 값이 위로 오게 설정
plt.gca().invert_yaxis()

# 그래프 표시
plt.show()
```

경기도, 서울특별시, 부산특별시 순으로 대출량이 많으며, 수도권역 절반 이상을 차지함을 알 수 있었다.

시도 행정 구분 경계 데이터를 통해 지도로 시각화 해보았다.

데이터 수집

디지털 트윈 국토 사이트에서 얻은 N3A_G001000.shp 데이터를 불러온다.

```
import geopandas as gpd

# 압축 해제 후 Shapefile 경로
shp_file_path = "/content/drive/MyDrive/bigdata_processing/N3A_G0010000/N3A_G0010000.shp"

# Shapefile 읽기
gdf = gpd.read_file(shp_file_path)

# 데이터 확인
print(gdf.head())
```

```
                              UFID        BJCD   NAME    DIVI      SCLS  \
0  1000037806845G001101000000000000  4200000000    강원도  HJD004  G0018112
1  1000037709820G001101000000000000  4100000000    경기도  HJD004  G0018112
2  1000035810071G001101000000000002  4800000000   경상남도  HJD004  G0018112
3  1000036811070G001101000000000003  4700000000   경상북도  HJD004  G0018112
4  1000035616034G001101000000000004  2900000000  광주광역시  HJD003  G0018112

      FMTA                                           geometry
0  S2112366  MULTIPOLYGON (((410031.382 508019.255, 410030....
1  S2115251  MULTIPOLYGON (((183707.154 485149.691, 183662....
2  S2112848  MULTIPOLYGON (((299424.537 211971.935, 299423....
3  S2117716  MULTIPOLYGON (((421936.784 342113.335, 421935....
4  S2116931  POLYGON ((178186.692 295823.537, 178268.005 29...
```

데이터 전처리

'NAME'으로 저장되어 있는 shp 파일의 지역명 부분을 현재 행정구역명으로 변경한다.

```
# NAME 열의 고유 값 출력
unique_names = gdf['NAME'].unique()
print(unique_names)
```

```
['강원도' '경기도' '경상남도' '경상북도' '광주광역시' '대구광역시' '대전광역시' '부산광역시' '서울특별시'
 '세종특별자치시' '울산광역시' '인천광역시' '전라남도' '전라북도' '제주특별자치도' '충청남도' '충청북도']
```

- 전북특별자치도 = 전라북도

2024년 1월 18일, '전북특별자치도'가 출범하여 행정구역 명칭이 아래와 같이 변경됨

https://overseas.mofa.go.kr/cn-wuhan-ko/brd/m_22785/view.do?seq=1347269

- 강원특별자치도 = 강원도

2023년 6월 11일, 628년 만에 기존의 강원도에서 강원특별자치도로 변경됨

```
# NAME 열 값 수정
gdf['NAME'] = gdf['NAME'].replace({
    '강원도': '강원특별자치도',
    '전라북도': '전북특별자치도',
})

# 수정된 결과 확인
print(gdf['NAME'].unique())
```

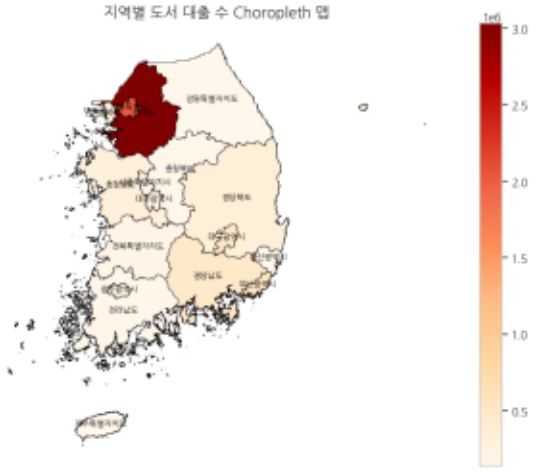
```
['강원특별자치도' '경기도' '경상남도' '경상북도' '광주광역시' '대구광역시' '대전광역시' '부산광역시' '서울특별시'
 '세종특별자치시' '울산광역시' '인천광역시' '전라남도' '전북특별자치도' '제주특별자치도' '충청남도' '충청북도']
```

데이터 시각화

NAME 열을 지역으로 변경하여 지역별 도서 대출량 데이터와 결합 후 choropleth 맵으로 시각화를 진행한다.

```
# 'NAME' 열을 '지역'으로 변경하여 병합 준비
gdf = gdf.rename(columns={'NAME': '지역'})

# 병합
merged_gdf = gdf.merge(df_long, on='지역', how='left')

# 병합 결과 확인
print(merged_gdf.head())
```

```
                                    UFID        BJCD      지역    GIVI     SCLS  \
0  1000378086045G0011101000000000000  4200000000  강원특별자치도  HJD004  G0018112
1  1000377090280G0011101000000000001  4100000000     경기도  HJD004  G0018112
2  1000358180710G0011101000000000002  4800000000    경상남도  HJD004  G0018112
3  1000368111070G0011101000000000003  4700000000    경상북도  HJD004  G0018112
4  1000356160340G0011101000000000004  2900000000   광주광역시  HJD003  G0018112

      FMTA                                           geometry     대출수
0  52112366  MULTIPOLYGON (((410831.382 508819.255, 410830....   159736
1  52115251  MULTIPOLYGON (((383707.154 485149.691, 383662....  3033461
2  52112848  MULTIPOLYGON (((299424.537 211971.935, 299423....   478560
3  52117716  MULTIPOLYGON (((421936.784 342113.335, 421935....   344051
4  52118931  POLYGON ((178186.692 295823.537, 178268.805 29...   196921
```



2. 전국 도서관 분포 현황

데이터 수집

공공데이터 포털에서 수집한 '전국도서관표준데이터.csv'를 활용한다.

```
import pandas as pd

file_path = '/content/drive/MyDrive/bigdata_processing/전국도서관표준데이터.csv'

national_libraries = pd.read_csv(file_path, encoding='CP949')

national_libraries.head()
```

	도서관명	시도명	시군구명	도서관유형	휴관일	평일운영시작시각	평일운영종료시각	토요일운영시작시각	토요일운영종료시각	공휴일운영시작시각	...	운영기관명	도서관전화번호	부지면적	건물면적	홈페이지주소	위도	경도	데이터기준일자	제공기관코드	제공기관명
0	중산도서	경기도	남양주시	작은도서관	공휴일	09:00	18:00	09:00	14:00	00:00	...	중산노인종합복지관	031-520-3457	NaN	234.0	http://www.snlib.go.kr/smal	37.598227	127.108422	2023-08-31	3760000	경기도 남양주
1	와부양떡이	경기도	남양주시	작은도서관	공휴일	09:00	18:00	00:00	00:00	00:00	...	양지재복지관	070-8861-0924	NaN	148.0	http://www.snlib.go.kr/smal	37.458586	127.162658	2023-08-31	3760000	경기도 남양주
2	와천어울림	경기도	남양주시	작은도서관	공휴일	09:00	18:00	00:00	00:00	00:00	...	와천어울복지관	031-736-2779	NaN	83.0	http://www.snlib.go.kr/smal	37.644538	127.12309	2023-08-31	3760000	경기도 남양주
3	진건꿈동지	경기도	남양주시	작은도서관	공휴일	09:00	18:00	00:00	00:00	00:00	...	진건효복지관	031-725-9939	NaN	427.0	http://www.snlib.go.kr/smal	37.640828	127.144453	2023-08-31	3760000	경기도 남양주
4	한울	경기도	남양주시	작은도서관	공휴일	09:00	18:00	00:00	00:00	00:00	...	한울종합사회복지관	031-8012-1177	NaN	190.0	http://www.snlib.go.kr/smal	37.567589	127.175589	2023-08-31	3760000	경기도 남양주

5 rows × 28 columns

데이터 전처리

데이터 정보를 확인해 본 결과, null 값이 있는 행은 존재하지 않아 결측치 처리는 필요하지 않았다. 우리나라 행정구역은 17개인데 19개로 세어지는 문제가 있었다.

```
national_libraries.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3557 entries, 0 to 3556
Data columns (total 28 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   도서관명         3557 non-null   object
 1   시도명          3557 non-null   object
 2   시군구명         3557 non-null   object
 3   도서관유형        3557 non-null   object
 4   휴관일          3557 non-null   object
 5   평일운영시작시각      3557 non-null   object
 6   평일운영종료시각      3557 non-null   object
 7   토요일운영시작시각     3557 non-null   object
 8   토요일운영종료시각     3557 non-null   object
 9   공휴일운영시작시각     3557 non-null   object
 10  공휴일운영종료시각     3557 non-null   object
 11  열람좌석수         3557 non-null   int64
 12  자료수(도서)       3557 non-null   object
 13  자료수(연속간행물)    3557 non-null   int64
 14  자료수(비도서)      3557 non-null   object
 15  대출가능권수        3557 non-null   int64
 16  대출가능일수        3557 non-null   int64
 17  소재지도로명주소      3557 non-null   object
 18  운영기관명         2827 non-null   object
 19  도서관전화번호       3329 non-null   object
 20  부지면적          1873 non-null   object
 21  건물면적          3185 non-null   float64
 22  홈페이지주소        2253 non-null   object
 23  위도            3313 non-null   float64
 24  경도            3313 non-null   float64
 25  데이터기준일자       3557 non-null   object
 26  제공기관코드        3557 non-null   object
 27  제공기관명         3557 non-null   object
dtypes: float64(3), int64(4), object(21)
memory usage: 778.2+ KB
```

```
unique_count = national_libraries['시도명'].nunique()
print("행정구역 갯수 : ", unique_count) # 대한민국의 행정 구역은 총 17개이나, 19개로 세어짐
```

```
행정구역 갯수 :  19
```

```
national_libraries['시도명'].unique()
```

```
array(['경기도', '서울특별시', '전북특별자치도', '경상북도', '제주특별자치도', '인천광역시', '충청남도',
       '울산광역시', '광주광역시', '전라남도', '충청북도', '부산광역시', '경상남도', '강원특별자치도',
       '대전광역시', '대구광역시', '세종특별자치시', '전라북도', '강원도'], dtype=object)
```

이것은 전북특별자치도와 강원특별자치도가 그 전 이름인 전라북도와 강원도가 남아있었기 때문이었다.

새로운 특별자치도 이름으로 통일하였다.

```
national_libraries['시도명'] = national_libraries['시도명'].replace({
    '전라북도': '전북특별자치도',
    '강원도': '강원특별자치도',
})

# 전처리된 결과 확인
print(national_libraries['시도명'].unique())
```

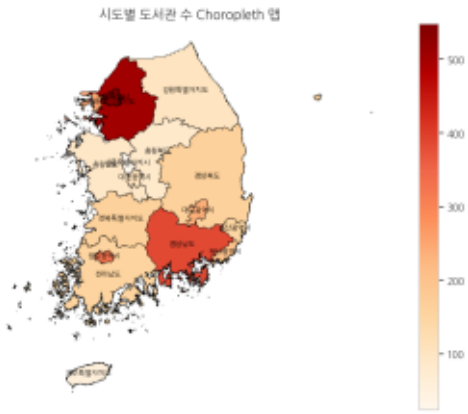
```
['경기도' '서울특별시' '전북특별자치도' '경상북도' '제주특별자치도' '인천광역시' '충청남도' '울산광역시' '광주광역시'
 '전라남도' '충청북도' '부산광역시' '경상남도' '강원특별자치도' '대전광역시' '대구광역시' '세종특별자치시']
```

데이터 시각화

시도명으로 묶어 시도별 전국 도서관 개수를 확인하였다, 대출 도서 수와 마찬가지로 막대그래프와 choropleth 맵으로 시각화를 진행했다.

```
# 시도별 전국 도서관 개수 확인
library_count_by_region = national_libraries['시도명'].value_counts().reset_index()

# 열 이름을 '시도명'과 '도서관 수'로 변경
library_count_by_region.columns = ['시도명', '도서관 수']

# 도서관 수를 기준으로 내림차순 정렬하고 인덱스를 재설정
library_count_by_region = library_count_by_region.sort_values(by='도서관 수', ascending=False).reset_index(drop=True)

# 결과 출력: 시도별 도서관 개수를 출력
print(library_count_by_region)
```

```
      시도명  도서관 수
0     서울특별시     548
1       경기도     506
2     경상남도     309
3     광주광역시     329
4     인천광역시     255
5     대구광역시     237
6     부산광역시     198
7   전북특별자치도     162
8     경상북도     162
9     전라남도     154
10    울산광역시     150
11    대전광역시     149
12  강원특별자치도     103
13    충청남도      93
14    충청북도      85
15  제주특별자치도      57
16  세종특별자치시      24
```



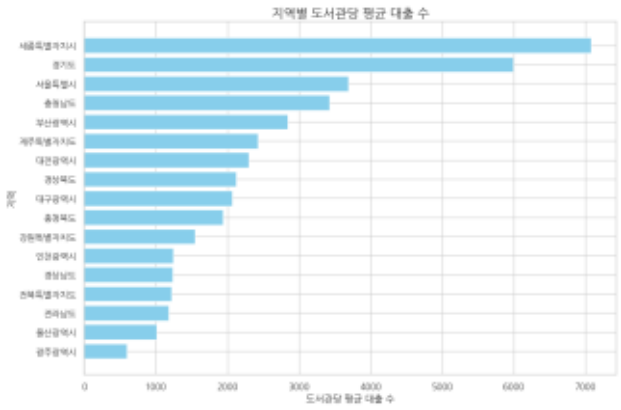
경상남도는 행정구역 중 도서관이 많은 축에 속하지만 도서대출 수는 낮은 편이다.

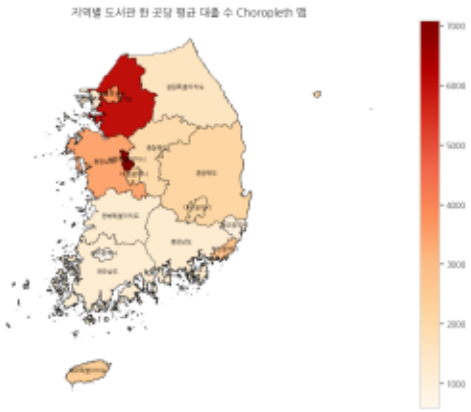
3. 지역별 도서관 한 곳당 평균 대출 수

지역별 도서관 수와 도서 대출 수가 정비례하지 않음에 따라 (도서 대출 수 / 도서관 수)를 통해 데이터 분석을 진행하기로 하였다.

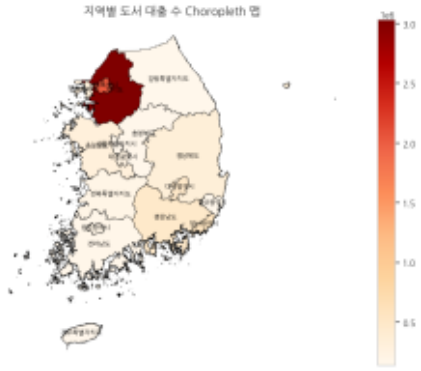
```
import pandas as pd

# library_count_by_region과 df_long_sorted 병합 (지역명을 기준으로)
merged_df = pd.merge(df_long_sorted, library_count_by_region, left_on='지역', right_on='시도명', how='inner')

# '도서관 당 평균 대출 수' 계산
merged_df['도서관당 평균 대출 수'] = merged_df['대출수'] / merged_df['도서관 수']

# 결과 데이터를 정리
average_loans_per_library = merged_df[['지역', '대출수', '도서관 수', '도서관당 평균 대출 수']]

# 내림차순 정렬
average_loans_per_library = average_loans_per_library.sort_values(by='도서관당 평균 대출 수', ascending=False).reset_index(drop=True)

# 결과 출력
print(average_loans_per_library)
```

```
          지역      대출수  도서관 수  도서관당 평균 대출 수
0     세종특별자치시   169920     24   7080.000000
1       경기도  3033461    506   5994.982213
2     서울특별시  2822531    548   3030.750000
3     충청남도   318730     93   3427.204301
4     부산광역시   562069    198   2838.732323
5   제주특별자치도   138185     57   2424.298246
6     대전광역시   248177    108   2297.935185
7     경상북도   344051    162   2123.771605
8     대구광역시   489618    237   2065.864979
9     충청북도   164709     85   1937.752941
10  강원특별자치도   159736    103   1550.834951
11    인천광역시   317549    255   1245.290196
12    경상남도   478598    309   1239.844660
13  전북특별자치도   198123    162   1222.981481
14    전라남도   181516    154   1178.675325
15    울산광역시   152280    150   1015.200000
16    광주광역시   196821    329    598.544073
```

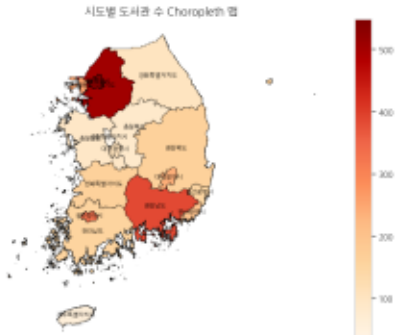




도서 대출 수를 도서관 수로 나누니, 세종특별시가 한 도서관 당 평균 대출 수가 많은 것으로 나타났다.



< 지역별 도서 대출 수 >



< 시도별 도서관 수 >

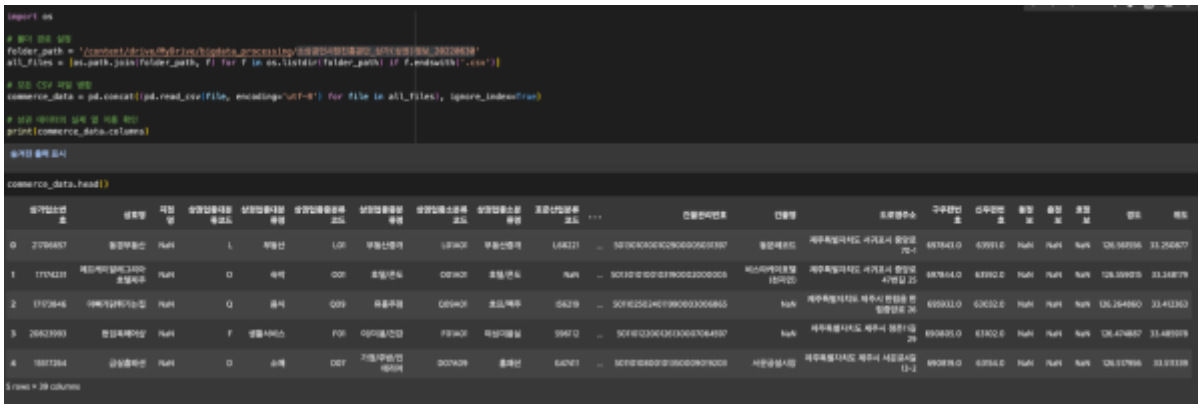
도서관 수와 도서 대출 수 사이에 어느 정도 상관관계가 있지만, 완전한 비례 관계는 아니다.

- 경기도와 서울특별시는 도서관 수와 대출 수 모두 상위권에 속한다.
- 그러나 경상남도의 경우, 도서관 수는 많지만 대출 수는 상대적으로 낮다.
- 세종특별자치시는 가장 적은 도서관 수(24개)를 가지고 있지만, 도서관당 평균 대출 수는 가장 높다(7,080권).

4. 지역별 도서관 한 곳당 평균 대출 수와 주변 상권과 비교

데이터 수집

공공데이터포털에서 전국도서관표준데이터.csv와 소상공인시장진흥공단_상가(상권) 정보_20220630을 다운로드하여 진행한다.



데이터 전처리

'시도명'을 현재 행정구역명으로 변경하고, 필요한 열만 추출한다.

```
import pandas as pd

# 도서관 위치 데이터 읽기
file_path = '/content/drive/MyDrive/bigdata_processing/전국도서관표준데이터.csv'
library_data = pd.read_csv(file_path, encoding='euc-kr')

library_data['시도명'] = library_data['시도명'].replace({
    '전라북도': '전북특별자치도',
    '강원도': '강원특별자치도',
})

# 전처리된 결과 확인
print(national_libraries['시도명'].unique())

# 필요한 열만 추출
library_data = library_data[['도서관명', '시도명', '위도', '경도']]
print(library_data.head())
```

```
['경기도' '서울특별시' '전북특별자치도' '경상북도' '제주특별자치도' '인천광역시' '충청남도' '울산광역시' '광주광역시'
 '전라남도' '충청북도' '부산광역시' '경상남도' '강원특별자치도' '대전광역시' '대구광역시' '세종특별자치시']
    도서관명  시도명          위도          경도
0   판교인생  경기도  37.390227  127.108622
1  하늘달햄이  경기도  37.458106  127.162618
2  하연대흙톡  경기도  37.344538  127.112319
3  한마음점자  경기도  37.408228  127.144153
4      한솔  경기도  37.367389  127.115589
```

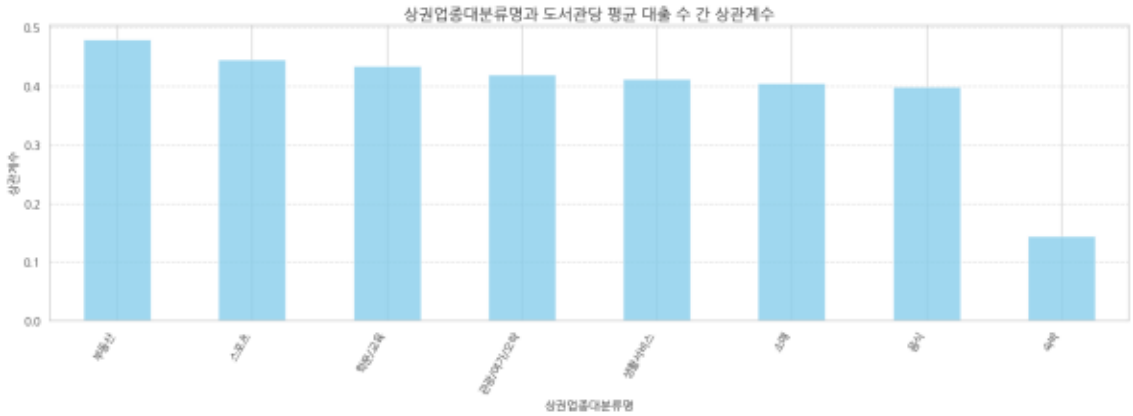
데이터 분석 및 시각화

상권업종대분류명과 도서관당 평균 대출 수 간 상관계수를 계산해보았다.

'시도명','상권업종대분류명' 열만 필터링 한 후 , '시도명'과 '상권업종대분류명'으로 그룹화하고 발생 횟수(개수)를 계산한다. 행이 시도명, 열이 상권업종대분류명, 값이 상권 개수인 피벗 테이블을 생성한다. 두 데이터로 만들어진 테이블을 병합하여 상관계수를 구한 후 시각화 하면 다음과 같다.

```python
import pandas as pd

# 1단계: 필요한 열만 필터링
commerce_big_data_filtered = commerce_data[['시도명', '상권업종대분류명']]

# 2단계: '시도명'과 '상권업종대분류명'으로 그룹화하고 발생 횟수(개수)를 계산
commerce_big_data_grouped = commerce_big_data_filtered.groupby(['시도명', '상권업종대분류명']).size().reset_index(name='상권 개수')

# 3단계: 피벗 테이블 생성 (행: 시도명, 열: 상권업종대분류명, 값: 상권 개수)
commerce_big_data_pivot = commerce_big_data_grouped.pivot(index='시도명', columns='상권업종대분류명', values='상권 개수').fillna(0)

# 4단계: 인덱스 초기화하여 데이터를 더 읽기 쉽게 만듭니다.
commerce_big_data_pivot.reset_index(inplace=True)

# 5단계: 결과 데이터프레임 출력
print(commerce_big_data_pivot)
```

```
상권업종대분류명      시도명  관광/여가/오락   부동산  생활서비스    소매   숙박  스포츠   음식  학문/교육
0              강원도    2047   2407   16388   33200  5627  1136   39692   5295
1              경기도   13148  29132  101963  170840  5844  7690  174428  43507
2             경상남도    3312   5163   26989   48291  4889  1828   60145  10838
3             경상북도    2685   3554   24085   44284  3676  1451   49912   7907
4             광주광역시    1765   3187   14506   24789   493   851   22546   5684
5             대구광역시    2285   4203   19826   33764   654  1293   34788   7252
6             대전광역시    1770   2707   14029   24597   585   854   23311   5183
7             부산광역시    3293   6646   27944   50859  1695  1988   50539   9945
8             서울특별시    8404  17356   69233  101805  2320  5175  117033  24903
9            세종특별자치시     281    960    1884    3093    67   164    4477   1006
10            울산광역시    1302   1710    8703   13188   645   555   17869   3935
11            인천광역시    3146   5610   23220   35608  1660  1410   36979   8138
12            전라남도    2135   2187   18335   42004  3729  1020   39428   6433
13            전라북도    2179   3111   19027   42121  2093  1097   33749   8191
14           제주특별자치도     958   1349    7017   13715  3556   595   20649   2791
15            충청남도    2430   3809   19510   40953  3121  1134   41040   6465
16            충청북도    2173   2779   16239   31719  1539  1048   30597   5762
```



부동산이 가장 높은 상관관계를, 숙박이 가장 적은 상관 관계를 가지고 있음을 알 수 있다.

4-1. 도서관 반경 1km 상권과 도서관당 평균 대출 수와의 상관계수

1. 도서관 데이터와 상권 데이터를 GeoDataFrame으로 변환하고, 도서관 반경 1km 버퍼를 준 뒤, 공간 조인으로 도서관 반경 1km 내 상권 데이터를 추출한다.

지역별 상권업종별 상권 개수를 집계한 뒤, 피벗 테이블을 생성하고, 도서관당 평균 대출 수 데이터와 병합한다. 상관 계수를 계산한 뒤 막대 그래프로 시각화하였다.

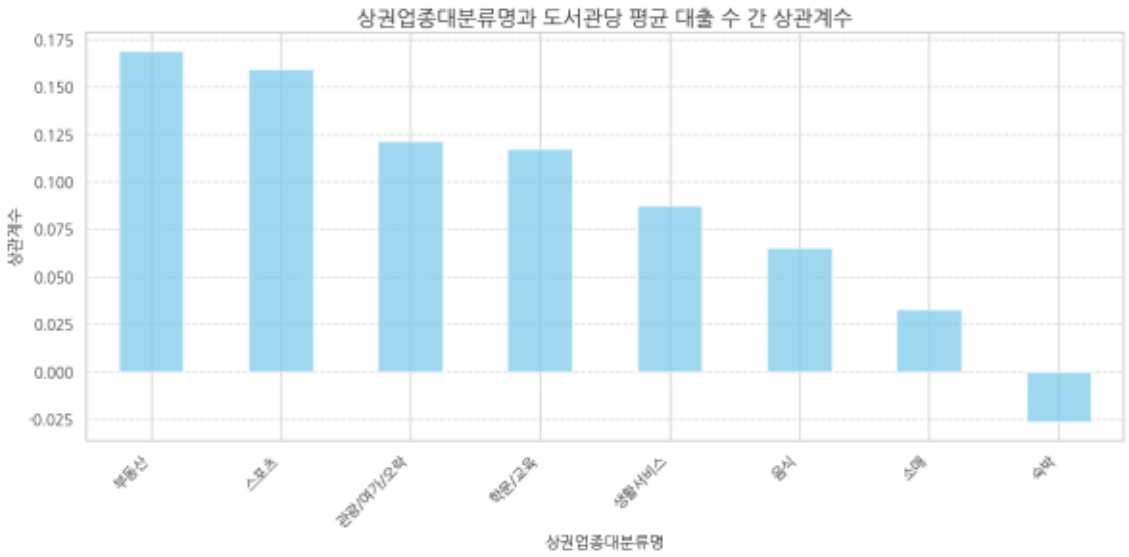
```python
import geopandas as gpd
from shapely.geometry import Point

# Step 1: 도서관 데이터를 GeoDataFrame으로 변환
library_gdf = gpd.GeoDataFrame(
    library_data,
    geometry=gpd.points_from_xy(library_data['경도'], library_data['위도']),
    crs='EPSG:4326'
)

# Step 2: 상권 데이터를 GeoDataFrame으로 변환
commerce_gdf = gpd.GeoDataFrame(
    commerce_data,
    geometry=gpd.points_from_xy(commerce_data['경도'], commerce_data['위도']),
    crs='EPSG:4326'
)

# Step 3: 좌표계를 적합한 단위 (EPSG:3857)로 변경
library_gdf = library_gdf.to_crs(epsg=3857)
commerce_gdf = commerce_gdf.to_crs(epsg=3857)

# Step 4: 도서관 반경 1km 버퍼 생성
library_gdf['buffer'] = library_gdf.geometry.buffer(1000)  # 1km = 1000m

# Step 5: 도서관 버퍼 GeoDataFrame 생성
buffer_gdf = gpd.GeoDataFrame(library_gdf[['buffer']], geometry='buffer', crs=library_gdf.crs)

# Step 6: 공간 조인으로 도서관 반경 1km 내 상권 데이터 추출
filtered_commerce = gpd.sjoin(commerce_gdf, buffer_gdf, how='inner', predicate='intersects')

# Step 7: 필터링된 데이터로 필요한 열만 유지
filtered_commerce = filtered_commerce[['상호명', '상권업종대분류명', '시도명', '위도', '경도']]
```

<1번 과정>

```python
# Step 1: 지역별 상권업종별 상권 개수 집계
filtered_commerce_grouped = filtered_commerce.groupby(['시도명', '상권업종대분류명']).size().reset_index(name='상권 개수')

# Step 2: 피벗 테이블 생성 (행: 지역, 열: 상권업종대분류명, 값: 상권 개수)
filtered_commerce_pivot = filtered_commerce_grouped.pivot(index='시도명', columns='상권업종대분류명', values='상권 개수').fillna(0)

# Step 3: 도서관당 평균 대출 수 데이터와 병합
# average_loans_per_library는 '지역' 열이 포함된 도서관당 평균 대출 수 데이터
average_loans_per_library.rename(columns={'지역': '시도명'}, inplace=True)
merged_data = pd.merge(average_loans_per_library[['시도명', '도서관당 평균 대출 수']], filtered_commerce_pivot, on='시도명', how='inner')

numeric_data = merged_data.select_dtypes(include=['number'])

# Step 4: 상관계수 계산
correlation_result = numeric_data.corr()['도서관당 평균 대출 수'].drop('도서관당 평균 대출 수')

# Step 5: 상관관계 출력
print(correlation_result)

# Step 6: 상관관계 시각화 (막대그래프)
plt.figure(figsize=(12, 6))
correlation_result.sort_values(ascending=False).plot(kind='bar', color='skyblue', alpha=0.8)
plt.title('상권업종대분류명과 도서관당 평균 대출 수 간 상관계수', fontsize=16)
plt.ylabel('상관계수', fontsize=12)
plt.xlabel('상권업종대분류명', fontsize=12)
plt.xticks(rotation=45, ha='right')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```

```
관광/여가/오락     0.121545
부동산           0.168869
생활서비스         0.087395
소매            0.032483
숙박           -0.026707
스포츠           0.159584
음식            0.064956
학문/교육         0.117140
Name: 도서관당 평균 대출 수, dtype: float64
```

<2번 과정>



상권업종과 도서관당 평균 대출 수 사이에는 약한 양의 상관관계가 있음을 알 수 있다.

• 부동산(0.480), 스포츠(0.445), 학문/교육(0.434) 업종과 도서관당 평균 대출 수 사이에 가장 높은 상관관계가 나타났다.
• 숙박업(0.145)을 제외한 모든 업종이 0.4 이상의 상관계수를 보였다.

이는 교육 시설뿐만 아니라 다양한 상권 업종이 도서 대출 수와 관련이 있음을 시사한다.

5. 지역별 도서관 한 곳당 평균 대출 수와 접근성과의 상관 관계

데이터 수집 및 전처리

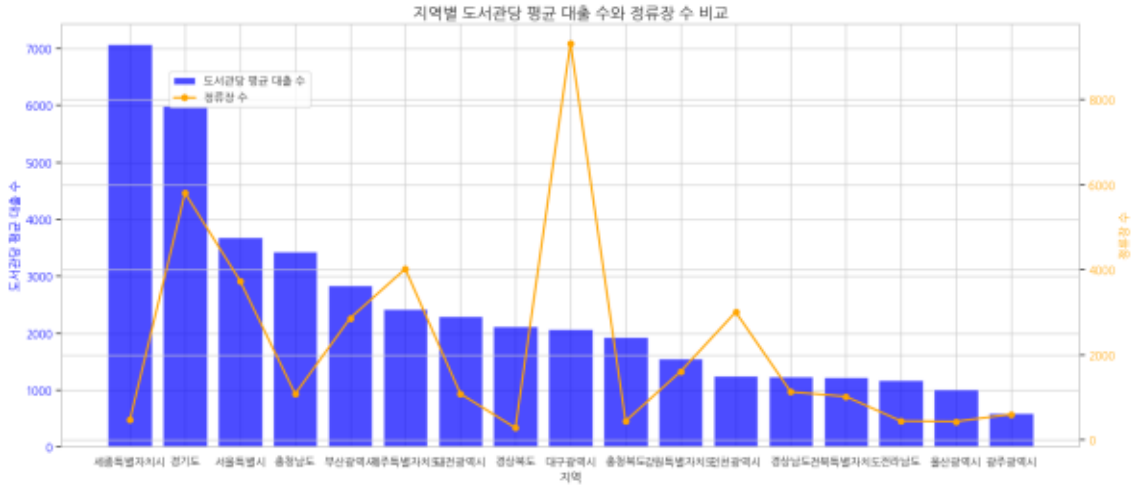
공공데이터 포털에서 수집한 '2022년_전국버스정류장 위치정보_데이터.csv'를 활용한다. 분석에 필요한 위도와 경도 열만 추출한다.

```
import pandas as pd
import geopandas as gpd
from shapely.geometry import Point
import matplotlib.pyplot as plt
import matplotlib.font_manager as fm
import seaborn as sns

# 한글 폰트 설정 (나눔고딕 폰트 사용)
font_path = '/usr/share/fonts/truetype/nanum/NanumGothic.ttf'  # 시스템에 설치된 폰트 경로
font_prop = fm.FontProperties(fname=font_path)
plt.rc('font', family=font_prop.get_name())
plt.rcParams['axes.unicode_minus'] = False  # 마이너스 부호 깨짐 방지

# 데이터 로드
csv_path = '/content/drive/MyDrive/bigdata_processing/국토교통부_전국 버스정류장 위치정보_20221012/2022년_전국버스정류장 위치정보_데이터.csv'

# CSV 파일 로드
station_data = pd.read_csv(csv_path, encoding='utf-8')  # 인코딩 확인 필요

library_location = national_libraries[["위도", "경도"]]
```

데이터 시각화

도서관 데이터와 정류장 데이터를 GeoDataFrame으로 변환한 뒤, 도서관 반경 500m 버퍼를 생성한다. (정류장을 걸어서 갈 수 있는 거리의 마지노선을 500m로 선정하였다.). 공간 조인으로 도서관 반경 500m 정류장을 추출한다.

```
# 도서관 데이터 GeoDataFrame으로 변환
library_gdf = gpd.GeoDataFrame(
    national_libraries,  # 원본 데이터프레임
    geometry=gpd.points_from_xy(
        national_libraries['경도'], national_libraries['위도']
    ),  # 경도, 위도로 포인트 생성
    crs='EPSG:4326'  # WGS84 좌표계
)

# 정류장 데이터 GeoDataFrame으로 변환
station_gdf = gpd.GeoDataFrame(
    station_data,
    geometry=gpd.points_from_xy(station_data['경도'], station_data['위도']),  # 정류장 경도, 위도
    crs='EPSG:4326'  # WGS84 좌표계
)

library_gdf = library_gdf.to_crs(epsg=3857)
station_gdf = station_gdf.to_crs(epsg=3857)

# Step 3: 도서관 반경 500m 버퍼 생성
library_gdf['buffer'] = library_gdf.geometry.buffer(500)  # 500m 반경 생성

# Step 4: 도서관 버퍼 데이터 생성
buffer_gdf = gpd.GeoDataFrame(library_gdf[['buffer', '시도명']], geometry='buffer', crs=library_gdf.crs)

# Step 5: 공간 조인으로 도서관 반경 500m 내 정류장 데이터 추출
filtered_station = gpd.sjoin(
    station_gdf,          # 정류장 데이터
    buffer_gdf,           # 도서관 버퍼 데이터
    how='inner',          # 도서관 버퍼 내에 포함된 정류장만 선택
    predicate='intersects'  # 정류장이 도서관 반경과 교차하는지 확인
)

# Step 6: 필터링된 데이터 확인
print(filtered_station.head())  # 도서관 반경 500m 내 정류장 데이터 출력
```



정류장 수가 많은 지역(예: 서울특별시, 경기도)은 도서관당 대출 수가 평균 이하로 나타났다. 이는 대중교통 접근성이 높더라도 도서관 이용률과 직접적인 상관관계가 낮을 수 있음을 의미한다. 교통 접근성보다는 도서관 프로그램, 시설 수준 등의 요인이 대출률에 더 큰 영향을 미쳤을 가능성이 있다.

6. 디지털 역세권인지에 따른 도서 대출량

데이터 수집

한국문화정보원이 공개한 디지털 문화역세권(2022).csv 데이터를 사용한다.

```
# 데이터 로드
csv_path = '/content/drive/MyDrive/bigdata_processing/디지털 문화역세권 (2022).csv'

# CSV 파일 로드
digital_subway = pd.read_csv(csv_path, encoding='utf-8')  # 인코딩 확인 필요
digital_subway.head()
```

	CTPRVN_NM	SIGNGU_NM	SIGNGU_CD	SCCNT_YM	FCLTY_CL_NM	FCLTY_CO	SEARCH_CO	FILE_NM	BASE_DE
0	서울특별시	종로구	1111000000	20230119	공공도서관	6	2689	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231
1	서울특별시	종로구	1111000000	20230119	박물관	23	41022	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231
2	서울특별시	종로구	1111000000	20230119	미술관	15	135923	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231
3	서울특별시	중구	1114000000	20230119	문예회관	1	79526	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231
4	서울특별시	중구	1114000000	20230119	공공도서관	3	6	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231

데이터 전처리

열 이름이 영어로 되어있는 것을 한국어로 변환해준다.

```
digital_subway

# 새로운 열 이름 리스트
new_columns = ['시도명', '시군구명', '시군구코드', '검색량년월', '시설분류명',
               '시설수', '검색수', '파일명', '기준일자']

# 열 이름 변경
digital_subway.columns = new_columns
digital_subway.head()
```

	시도명	시군구명	시군구코드	검색량년월	시설분류명	시설수	검색수	파일명	기준일자
0	서울특별시	종로구	1111000000	20230119	공공도서관	6	2689	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231
1	서울특별시	종로구	1111000000	20230119	박물관	23	41022	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231
2	서울특별시	종로구	1111000000	20230119	미술관	15	135923	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231
3	서울특별시	중구	1114000000	20230119	문예회관	1	79526	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231
4	서울특별시	중구	1114000000	20230119	공공도서관	3	6	KC_597_DGT_CLT_STATN_BIZAEA_2022	20221231

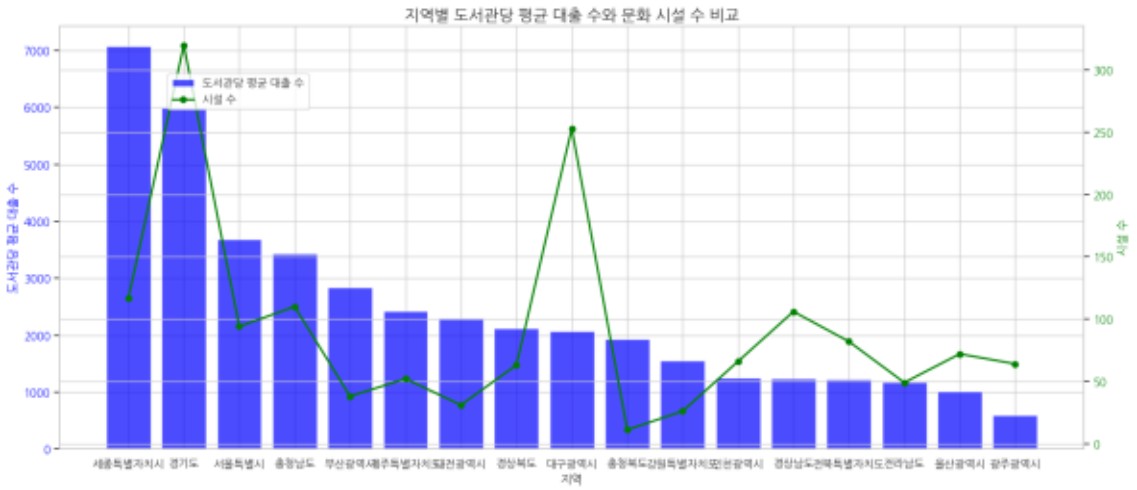
groupby를 이용해 시도명별 문화 시설 수의 합계를 계산한다.

```
# 시도명별 시설수 합계 계산
cultural_facilities_count_per_city = digital_subway.groupby('시도명')['시설수'].sum().reset_index()

# 결과 확인
print(cultural_facilities_count_per_city)
```

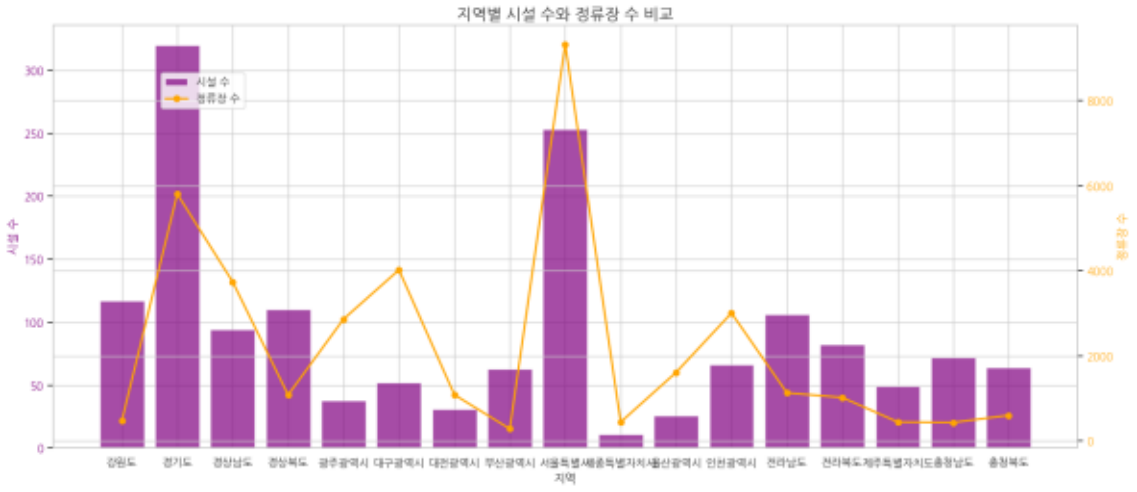
```
        시도명  시설수
0       강원도  117
1       경기도  328
2      경상남도   94
3      경상북도  110
4     광주광역시   38
5     대구광역시   52
6     대전광역시   31
7     부산광역시   63
8     서울특별시  253
9   세종특별자치시   11
10    울산광역시   26
11    인천광역시   66
12     전라남도  106
13     전라북도   82
14  제주특별자치도   49
15     충청남도   72
16     충청북도   64
```

데이터 시각화

시도명별 시설수의 합계와 지역별 도서관당 평균 대출수를 막대그래프와 꺾은 선 그래프로 나타낸다.



디지털 역세권인지와 접근성이 좋은지에 따른 시각화 결과가 비슷하게 나옴을 알 수 있다. 두 요인은 도서 대출량보다 서로에게 영향을 많이 주는 듯하다. (지역 문화시설이 많을수록 정류장 수가 많다.)



결론

가설 1. 도서관의 수와 도서 대출 수는 비례할 것이다.

도서관 수와 도서 대출 수 사이에 어느 정도 상관관계가 있지만, 완전한 비례 관계는 아니다. 세종특별자치시는 가장 적은 도서관 수(24개)를 가지고 있지만, 도서관당 평균 대출 수는 가장 높다(7,080권).

가설 2. 도서 대출 수는 학원 같은 교육 시설 수에 영향을 가장 많이 받을 것이다.

상권업종과 도서관당 평균 대출 수 사이에는 약한 양의 상관관계가 있다. 상업 데이터 특성상 주거공간은 들어가지 않았는데 부동산이 높은 이유에는 주거공간이 근처 있음을 나타내는 것으로 보인다. 스포츠가 많은 이유는 도서관과 함께 스포츠 공간이 또 하나의 취미 생활의 공간이기 때문이라고 생각한다. 그 다음으로 학업이 뒤이은다. 그와 반대로 숙박은 음의 상관관계를 가졌다.

가설 3. 문화시설 수가 많은 곳일 수록 도서 대출 수가 높을 것이다,
가설 4. 접근성이 높을 수록 도서 대출 수가 높을 것이다.

생각보다 영향을 많이 주지 않는 듯 했다. 오히려 문회시설의 수와 접근성이 더 영향을 많이 주었다.

이번 프로젝트를 통해 앞으로 독서율을 높이기 위해 생각해야 할 것들을 정리해보았다.

도서관 접근성 관련

1. 도서관 수 확대
   세종특별자치시의 사례에서 볼 수 있듯이, 도서관 수가 적더라도 도서관당 평균 대출 수가 높을 수 있다. 따라서 단순히 도서관 수를 늘리는 것보다는 전략적으로 도서관을 배치하는 것이 중요하다.
2. 도서관 위치 최적화: 상권 데이터와 도서관 위치를 분석하여 접근성이 높은 곳에 도서관을 설치해야 한다. 특히 부동산(0.480), 스포츠(0.445), 학문/교육(0.434) 업종과 도서관당 평균 대출 수 사이에 높은 상관관계가 나타났으므로, 이러한 업종이 밀집한 지역에 도서관을 설치하는 것이 효과적일 수 있다.

도서관 서비스 개선

1. 도서관 운영 시간 확대
   분석해보지 않았지만 평일, 토요일, 공휴일 운영 시간을 분석하여 이용자들의 니즈에 맞게 운영 시간을 조정하는 것이 도움이 될 수 있다.
2. 열람 좌석 수 증대
   도서관의 열람 좌석 수를 늘려 더 많은 이용자가 동시에 이용할 수 있도록 한다.

상권과 연계한 독서 문화 조성

1. 복합 문화 공간 조성

   상권업종 중 '관광/여가/오락'이 도서관당 평균 대출 수와 양의 상관관계(0.420)를 보였고, 문화시설이 많은 지역이 도서 대출 수가 많은 모습을 보였다.

   도서관과 문화시설을 연계한 복합 공간을 조성한다.

2. 교육 시설과의 연계

   학문/교육 업종과의 상관관계(0.434)를 고려하여, 학교나 학원과 연계한 독서 프로그램을 개발한다. 청소년들은 학교를 통해서 독서 교육을 받을 수 있지만, 성인들은 쉽지 않다. 성인들을 위해 지역 문화센터들이 관련 프로그램을 만드는 등 노력해주어야 한다고 생각한다.

출처

2023년 국민독서실태조사 보고서
http://www.mcst.go.kr/kor/s_notice/notice/noticeView.jsp?pSeq=18001

- 학생 문해력 실태 교원 인식 설문조사 결과 발표(2024.10.07.), 한국 교원 단체 총연합회
https://www.kfta.or.kr/usr/wap/detail.do?app=16527&seq=270000370496

- 중앙일보. (2022년 12월 5일). '심심한 사과' '사흘' 문해력 논란…청년만의 문제 아니다. 중앙일보. https://www.joongang.co.kr/article/25123031

- 빅데이터 분석보고서 『도담: 도서관 빅데이터를 담다』 제1호 발간, 2022-06-29, 빅데이터 정보나루
https://www.data4library.kr/noticeV

- 문화 체육 관광부 (제4차 독서문화진흥기본계획(2024-2028))
https://www.mcst.go.kr/kor/s_policy/dept/deptView.jsp?pSeq=1921&pDataCD=0406000000&pType=04-