도서관 빅데이터를 활용한
20대 성별에 따른 대출 트랜드 분석

202044069 배예진

1. 데이터 수집 방법

사용 데이터 : 도서관 정보나루 제공 데이터
https://data4library.kr/

수집 방법 : 도서관 정보나루 사이트의 '인기대출도서'
데이터를 수집한다.

조건 1) 기간(금년), 성별(여성, 남성), 연령(20대)
조건 2) 기간(작년, 분기별 설정), 성별(여성, 남성), 연령(20대)

⇒ 여성과 남성의 데이터를 따로 받아야 하므로
동시에 선택하지 않고 각각 선택하여 csv 파일을 다운 받는다.

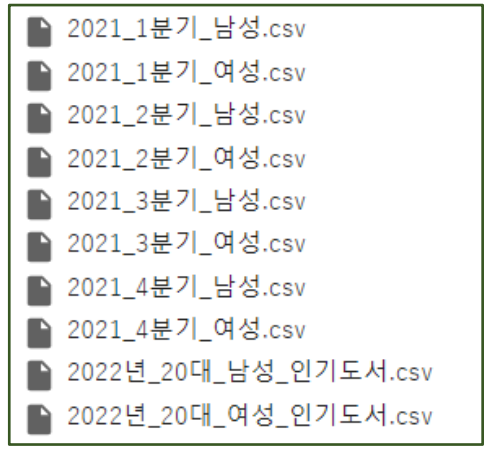
https://www.data4library.kr/loanDataL

1. 데이터 수집 방법

[최종 사용 데이터]

1) 20대 성별에 따른 대출 트렌드 분석용 : 2022년_20대_남성_인기도서
2022년_20대_여성_인기도서

2) 20대 성별, 분기별 대출 트랜드 분석용 : 2021_n분기_남성, 2021_n분기_여성



[csv 파일 구성 내용]

순위, 서명, 저자, 출판사, 출판년도, 권, ISBN, ISBN부가기호, KDC, 대출건수 (200건)

	A	B	C	D	E	F	G	H	I	J
1	순위	서명	저자	출판사	출판년도	권	ISBN	ISBN부가기	KDC	대출건수
2	1	지구 끝의	지은이: 김	Giant Boo	2021		9.79E+12	3810	813.7	5176
3	2	달러구트	지은이: 이	팩토리나인	2020		9.79E+12	3810	813.7	5028
4	3	시선으로부	지은이: 정	문학동네	2020		9.79E+12	3810	813.7	4891
5	4	우리가 빛	지은이: 김	허블	2019		9.79E+12	3810	813.7	4763

2. 데이터 전처리

1) 한글 사용을 위한 폰트 적용

```
[ ]  # 단계 1: 폰트 설치
     import matplotlib.font_manager as fm

     !apt-get -qq -y install fonts-nanum > /dev/null
     fontpath = '/usr/share/fonts/truetype/nanum/NanumBarunGothic.ttf'
     font = fm.FontProperties(fname=fontpath, size=9)
     fm._rebuild()
```

```
[ ]  # 단계 2: 런타임 재시작
     import os
     os.kill(os.getpid(), 9)
```

```
[1]  # 단계 3: 한글 폰트 설정
     import matplotlib.pyplot as plt
     import matplotlib as mpl
     import matplotlib.font_manager as fm

     # 마이너스 표시 문제
     mpl.rcParams['axes.unicode_minus'] = False

     # 한글 폰트 설정
     path = '/usr/share/fonts/truetype/nanum/NanumGothicBold.ttf'
     font_name = fm.FontProperties(fname=path, size=18).get_name()
     plt.rc('font', family=font_name)
     fm._rebuild()
```

2) 라이브러리 불러오기

```
import pandas as pd
import seaborn as sns
```

데이터 가공 : pandas
데이터 시각화 : seaborn

2. 데이터 전처리 – 20대 성별에 따른 대출 트랜드 분석

3) csv -> 데이터프레임 변환

```
# 파일 경로
file_path_man = '/content/2022년_20대_남성_인기도서.csv'
file_path_woman = '/content/2022년_20대_여성_인기도서.csv'


# 데이터프레임 변환(인코딩 에러 -> encoding 파라미터 활용)
df_man = pd.read_csv(file_path_man, encoding='cp949')
df_woman = pd.read_csv(file_path_woman, encoding='cp949')
```

4) 성별 열 추가

```
#성별 열 추가
df_man['성별'] = '남성'
print('2022년_20대_남성_인기도서')
print(df_man.head())

df_woman['성별'] = '여성'
print('2022년_20대_여성_인기도서')
print(df_woman.head())
```

남녀 csv를 각각 따로 다운 받아서 합칠 예정이므로,
구분을 위해 성별 열을 추가한다.

2. 데이터 전처리 – 20대 성별에 따른 대출 트랜드 분석

5) 데이터프레임 병합, 가공

```
# 데이터프레임 합치기(행 방향으로 합치기)
df = pd.concat([df_man, df_woman])
print(df.info()) #400row 확인


#합친 데이터 프레임 가공
#권,ISBN 부가기호 열 삭제
df.drop(['권', 'ISBN부가기호'], axis=1, inplace=True)

#KDC 결측치 처리(396 non-null, 4건의 NaN)
df['KDC'].fillna(method='bfill', inplace=True)
print(df.info())
```

- 남녀 데이터프레임을 행 방향으로 합친다.

각 데이터가 200건이므로 concat 이후에 info()로 400건의 데이터를 확인한다.

- 필요한 데이터만 사용하기 위해 '권', 'ISBN부가기호' 열을 삭제한다.

- KDC(한국십진분류법) 열에 결측값이 있어 결측값을 제거한다.

2. 데이터 전처리 – 20대 성별에 따른 대출 트랜드 분석

6) 데이터 전처리 전후 비교

```
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   순위        400 non-null    int64
 1   서명        400 non-null    object
 2   저자        400 non-null    object
 3   출판사       400 non-null    object
 4   출판년도      400 non-null    object
 5   권         41 non-null     float64
 6   ISBN      400 non-null    float64
 7   ISBN부가기호  394 non-null    float64
 8   KDC       396 non-null    float64
 9   대출건수      400 non-null    int64
 10  성별        400 non-null    object
```

```
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   순위      400 non-null    int64
 1   서명      400 non-null    object
 2   저자      400 non-null    object
 3   출판사     400 non-null    object
 4   출판년도    400 non-null    object
 5   ISBN    400 non-null    float64
 6   KDC     400 non-null    float64
 7   대출건수    400 non-null    int64
 8   성별      400 non-null    object
```

열이 삭제되고, 결측값이 제거된 것을 확인할 수 있다.

2. 데이터 전처리 – 20대 성별에 따른 대출 트랜드 분석

7) 성별 기준 group by 연산

```
#성별별 대출건수 합산을 위한 groupby 연산
grouped_sex = df.groupby(df['성별'])['대출건수'].sum()

# 시리즈 -> 데이터프레임 변환
df_sex = pd.Series(grouped_sex, index=['남성', '여성']).rename_axis('성별')
print(grouped_sex)
print(df_sex)

#성별별 대출건수 시각화
# colors = sns.color_palette('hls') ## 색상 지정
colors=['darkseagreen', 'mediumpurple']
df_sex.plot(kind='bar', color=colors)
```

성별을 기준으로 대출건수를 합산하기 위해 groupby를 통해 시리즈를 만든다. 만들어진 시리즈를 데이터프레임을 변환하여 plot 으로 시각화 한다.

2. 데이터 전처리 – 20대 성별에 따른 대출 트랜드 분석

8) KDC(한국십진분류법) 기반 분류를 위해 함수를 적용한다.

```
def kdcFunc(x) :
 if (x > 0 and x <100):
  return '총류'
 elif (x >=100 and x<200) :
  return '철학'
 elif (x >=200 and x<300) :
  return '종교'
 elif (x >=300 and x<400) :
  return '사회과학'
 elif (x >=400 and x<500) :
  return '자연과학'
 elif (x >=500 and x<600) :
  return '기술과학'
 elif (x >=600 and x<700) :
  return '예술'
 elif (x >=700 and x<800) :
  return '언어'
 elif (x >=800 and x<900) :
  return '문학'
 elif (x >=900 and x<1000) :
  return '역사'
 else :
  return '미상'

df['sorted_KDC'] = df.apply(lambda x:kdcFunc(x['KDC']), axis=1)
```

000 총류	100 철학	200 종교	300 사회과학	400 자연과학
010 도서학, 서지학	110 형이상학	210 비교종교	310 통계학	410 수 학
020 문헌정보학	120 인식론,인과론,인간학	220 불 교	320 경제학	420 물리학
030 백과사전	130 철학의 체계	230 기독교	330 사회학,사회문제	430 화 학
040 강연집,수필집,연설문집	140 경 학	240 도 교	340 정치학	440 천문학
050 일반연속간행물	150 동양철학,사상	250 천도교	350 행정학	450 지 학
060 일반학회,단체,협회,기관	160 서양철학	260 신 도	360 법 학	460 광물학
070 신문,언론,저널리즘	170 논리학	270 힌두교,브라만교	370 교육학	470 생명과학
080 일반전집,총서	180 심리학	280 이슬람교(회교)	380 풍속,예절,민속학	480 식물학
090 향토자료	190 윤리학,도덕철학	290 기타 제종교	390 국방,군사학	490 동물학

500 기술과학	600 예술	700 언어	800 문학	900 역사
510 의 학	610 건축물	710 한국어	810 한국문학	910 아시아
520 농업,농학	620 조각,조형예술	720 중국어	820 중국문학	920 유 럽
530 공학,공업일반,토목공학,환경공학	630 공예,장식미술	730 일본어,기타아시아제어	830 일본문학,기타아시아문학	930 아프리카
540 건축공학	640 서 예	740 영 어	840 영미문학	940 북아메리카
550 기계공학	650 회화,도화	750 독일어	850 독일문학	950 남아메리카
560 전기공학,전자공학	660 사진예술	760 프랑스어	860 프랑스문학	960 오세아니아
570 화학공학	670 음 악	770 스페인어,포르투갈어	870 스페인,포르투갈문학	970 양극지방
580 제조업	680 공연예술,매체예술	780 이탈리아어	880 이탈리아문학	980 지 리
590 생활과학	690 오락,스포츠	790 기타제어	890 기타제문학	990 전 기

csv에는 KDC가 숫자로 입력되어 있어
한글로 변환한 새로운 열을 추가하기 위해 함수를 적용한다.

2. 데이터 전처리 – 20대 성별에 따른 대출 트랜드 분석

9) 함수 적용 후 데이터 확인

```
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   순위         400 non-null    int64
 1   서명         400 non-null    object
 2   저자         400 non-null    object
 3   출판사        400 non-null    object
 4   출판년도       400 non-null    object
 5   ISBN        400 non-null    float64
 6   KDC         400 non-null    float64
 7   대출건수       400 non-null    int64
 8   성별         400 non-null    object
 9   sorted_KDC  400 non-null    object
dtypes: float64(2), int64(2), object(6)
```

```
문학      220
사회과학    103
철학       37
총류       13
자연과학      9
기술과학      9
역사        6
예술        3
Name: sorted_KDC, dtype: int64
```

	출판사	출판년도	ISBN	KDC	대출건수	성별	sorted_KDC
0	김영사	2015	9.788930e+12	909.00	973	남성	역사
1	미래엔	2020	9.791160e+12	332.60	879	남성	사회과학
2	현대문학	2012	9.788970e+12	833.60	842	남성	문학
3	팩토리나인	2020	9.791170e+12	813.70	783	남성	문학
4	Snowfox(스노우폭스북스)	2020	9.791190e+12	327.04	730	남성	사회과학
..	...	...	...	...	...	..	...
195	해냄	2018	9.790000e+12	189.00	761	여성	철학
196	문학동네	2001	9.790000e+12	879.00	760	여성	문학
197	소미미디어	2020	9.790000e+12	833.60	759	여성	문학
198	은행나무	2014	9.790000e+12	813.70	757	여성	문학
199	문학동네	2017	9.790000e+12	813.62	756	여성	문학

[400 rows x 10 columns]

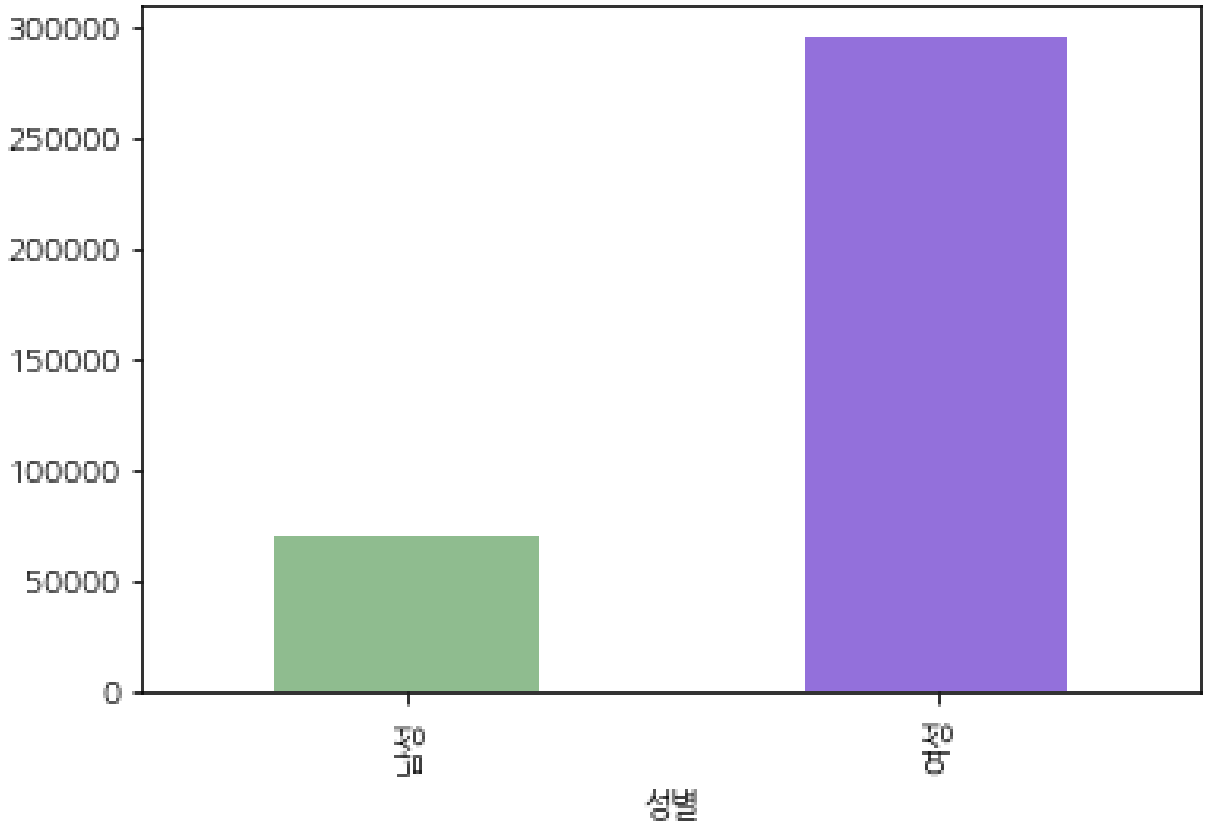
df.info()를 통해 sorted_KDC가 추가된 것을 확인.

df['sorted_KDC'].value_counts()를 통해 고유값 확인

KDC 가 함수를 통해 sorted_KDC 열에 한글로 입력되어 있음을 확인.

3. 데이터 시각화 – 20대 성별에 따른 대출 트랜드 분석

1) 성별에 따른 대출건수 시각화




```
#성별별 대출건수 시각화
colors=['darkseagreen', 'mediumpurple']
df_sex.plot(kind='bar', color=colors)
```

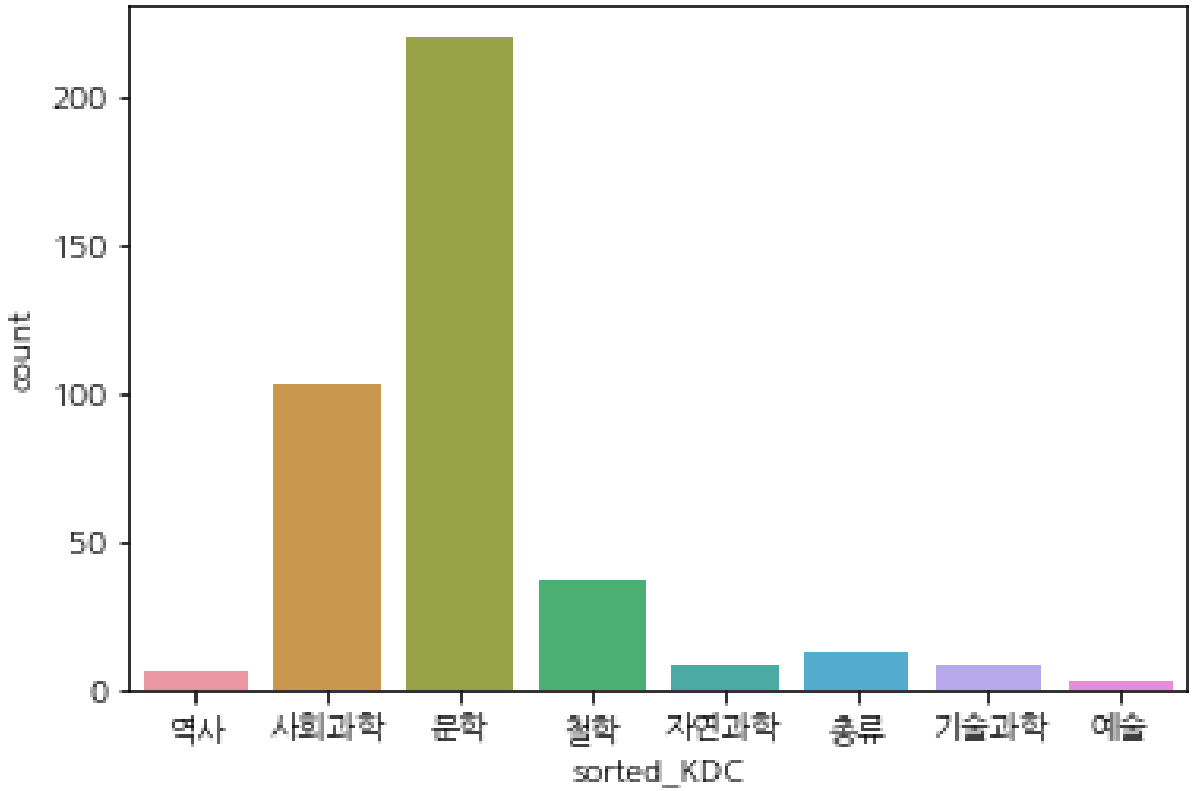
[2022년 12월 10일까지 남녀 대출 건수 비교]

남성 70985
여성 295929

=> 여성이 남성의 4배 이상 대출건수가 많은 것을 확인할 수 있다.

3. 데이터 시각화 – 20대 성별에 따른 대출 트랜드 분석

2) KDC에 따른 대출건수 시각화(성별 통합)



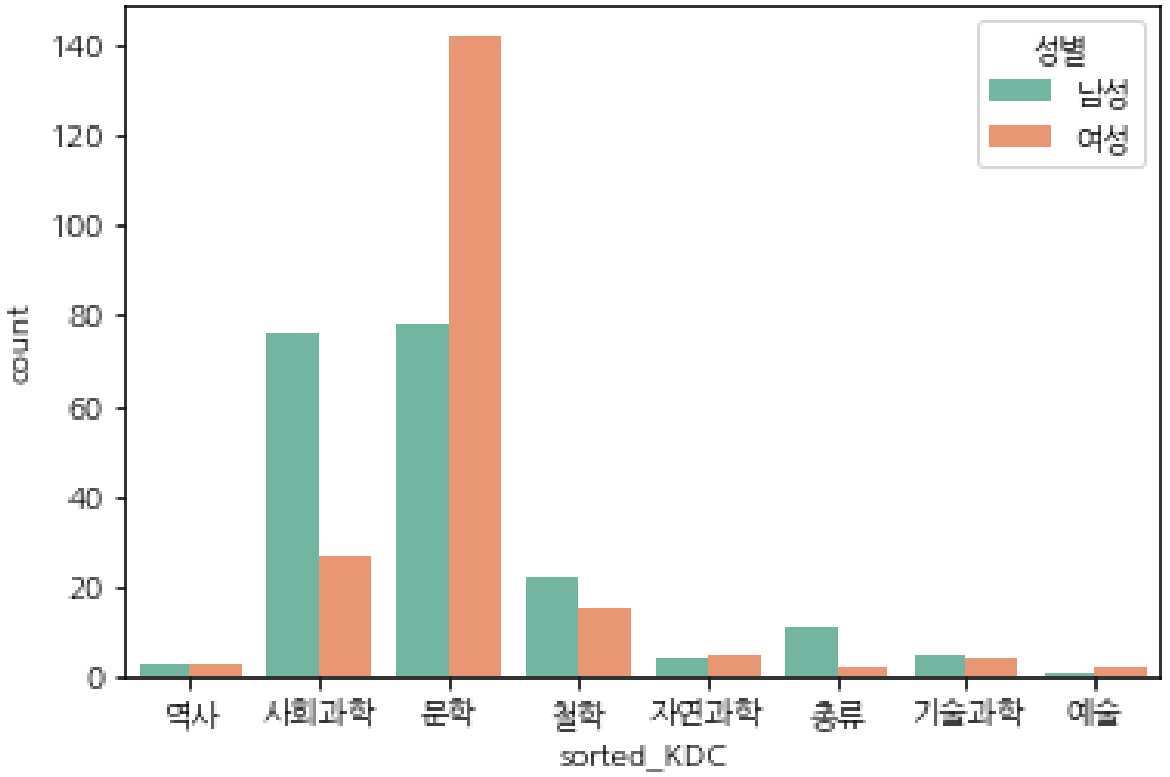

```
#20대 전체 기준 KDC 그래프

sns.set_palette("hls")
sns.countplot(data=df, x='sorted_KDC')
```

문학 -> 사회과학 -> 철학 -> 총류 순으로 대출건수가 많음.

특히 문학과 사회과학이 압도적임을 그래프를 통해 확인할 수 있다.

3. 데이터 시각화 – 20대 성별에 따른 대출 트랜드 분석

3) 성별과 KDC에 따른 대출건수 시각화




```
# 성별별 KDC 분류
sns.set_palette("Set2")
sns.countplot(data=df, x='sorted_KDC', hue='성별')
```

여성의 경우 대출건수 1, 2위인 문학과 사회과학의 차이가 많이 나지만,

남성의 경우 비등함을 그래프로 확인할 수 있다.

2. 데이터 전처리 – 2021년 기준, 20대 성별 · 분기별 대출건수와 KDC 분석

1) csv -> 데이터프레임 변환

```
# csv 불러오기
file_path_man1 = '/content/2021_1분기_남성.csv'
file_path_man2 = '/content/2021_2분기_남성.csv'
file_path_man3 = '/content/2021_3분기_남성.csv'
file_path_man4 = '/content/2021_4분기_남성.csv'

file_path_woman1 = '/content/2021_1분기_여성.csv'
file_path_woman2 = '/content/2021_2분기_여성.csv'
file_path_woman3 = '/content/2021_3분기_여성.csv'
file_path_woman4 = '/content/2021_4분기_여성.csv'
```

8개의 CSV를 앞서 변환한 방법과 같은 방식으로
데이터 프레임으로 변환한다.

2. 데이터 전처리 – 2021년 기준, 20대 성별 · 분기별 대출건수와 KDC 분석

2) 성별, 분기 열 추가

```
# 성별, 분기 열 추가
man_list = [df_man1, df_man2, df_man3, df_man4]
woman_list = [df_woman1, df_woman2, df_woman3, df_woman4]
num_list = [1, 2, 3, 4]

for i, j in zip(man_list, num_list):
  i['성별'] = '남성'
  i['분기'] = str(j)+ '분기'

for i, j in zip(woman_list, num_list) :
  i['성별'] = '여성'
  i['분기'] = str(j)+ '분기'
```

분기별 분석을 위해,
각 csv 별로 분기 열을 추가한다.

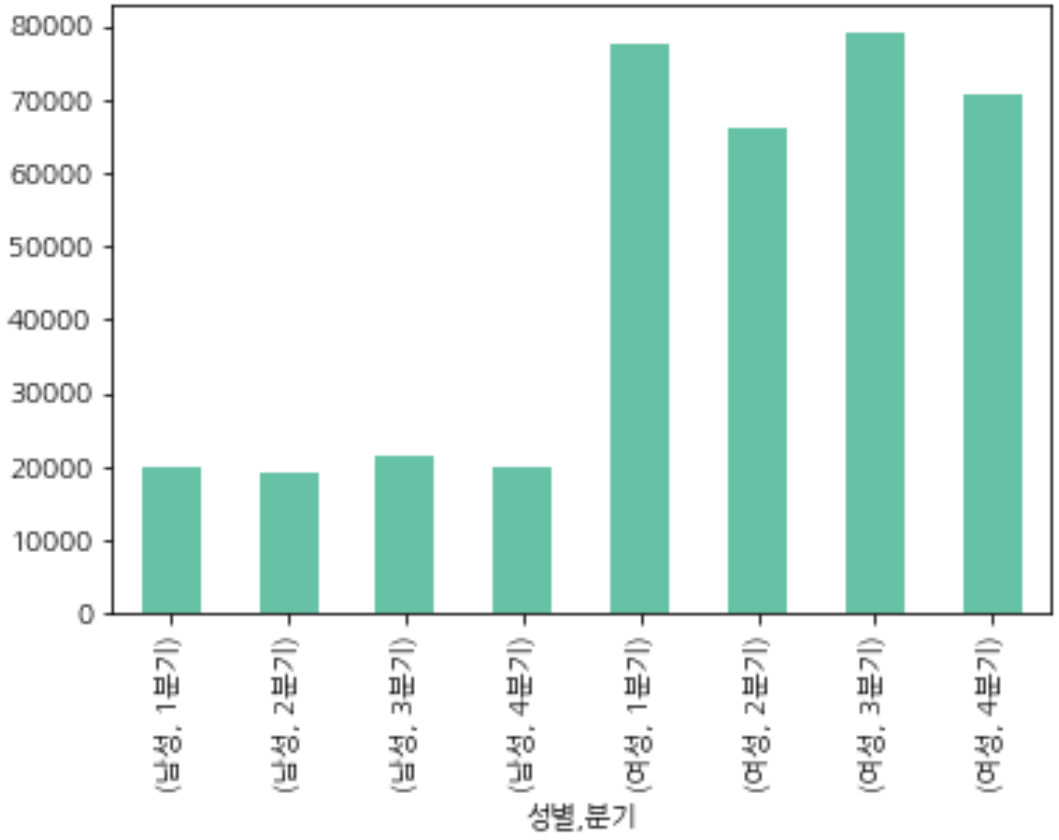
3) 데이터프레임 병합, 가공

8개의 데이터프레임을 병합하고
불필요한 행을 삭제한 뒤,
KDC 결측치를 처리한다.

(기존 2-5번과 동일)

3. 데이터 시각화 – 2021년 기준, 20대 성별 · 분기별 대출건수와 KDC 분석

1) 분기별 대출건수 비교



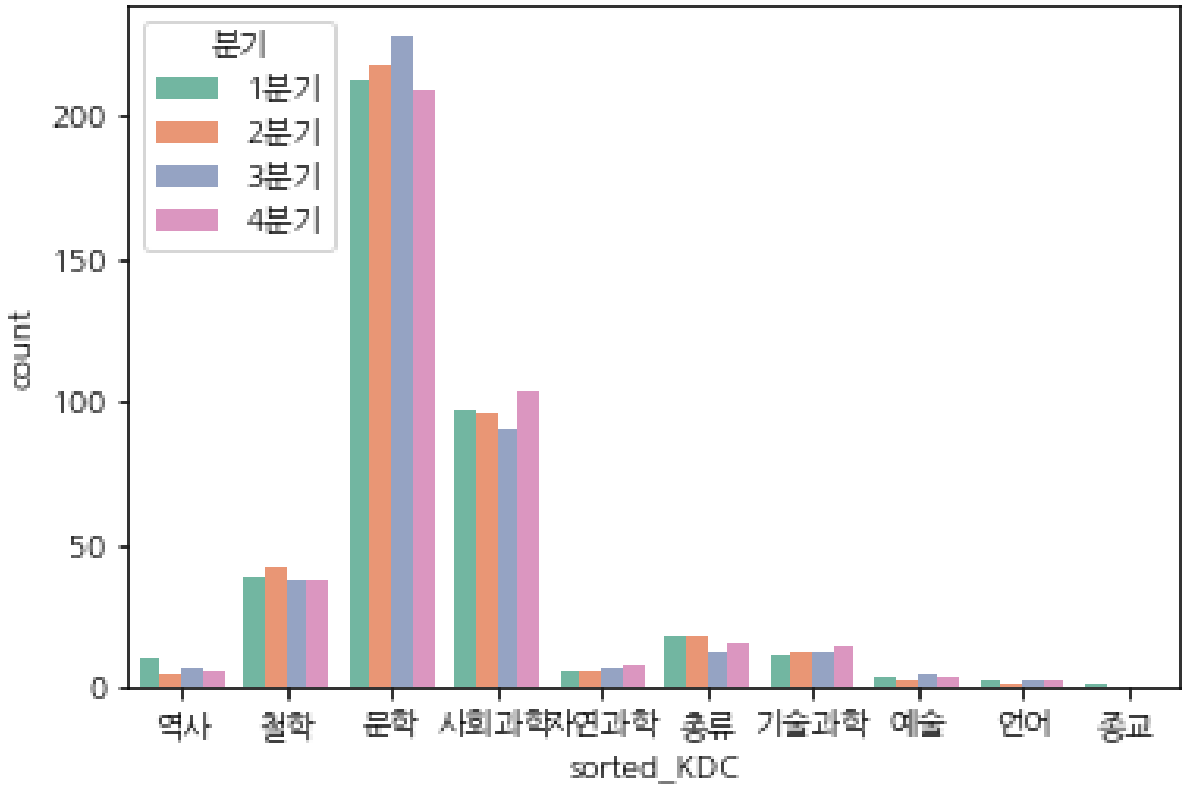

```
# 분기별 대출량
grouped_quater = df_quater.groupby(['성별', '분기'])['대출건수'].sum()
print(grouped_quater)
grouped_quater.plot.bar()
```

남성과 여성 둘 다 1,3분기에 대출건수가 많고,
2, 4분기에 대출건수가 줄어듦을 확인할 수 있다.

성별	분기	
남성	1분기	20124
	2분기	19200
	3분기	21538
	4분기	20121
여성	1분기	77371
	2분기	65949
	3분기	78994
	4분기	70568

3. 데이터 시각화 – 2021년 기준, 20대 성별 · 분기별 대출건수와 KDC 분석

2) 분기별 대출도서의 KDC 비교 (성별 통합)




```
# 분기별 대출도서 유형 분류
sns.set_palette("Set2")
sns.countplot(data=df_quater, x='sorted_KDC', hue='분기')
```

그래프를 통해 분기별 대출 도서의 KDC 건수를 비교한다.

문학은 4분기에 대출건수가 감소한 것에 비해,
사회과학은 4분기에 대출건수가 증가한 것을 확인할 수 있다.