PART.Ⅳ
연구과제

다이나믹
반응형 탐색적 자료분석

연구배경
연구내용
연구결과
향후계획

다이나믹 반응형 탐색적 자료분석

서울대학교 자연과학대학
박건웅

연구배경

빅데이터의 이해

데이터의 크기, 성격, 정보 등을 고려하여 최적의 데이터 분석 방식을 선택하는 것은 사용자의 편의를 극대화하고 새로운 서비스를 창출할 수 있는 중요한 요소로 작용함. 도서관 대출 데이터에 대한 분석 결과는 기존 도서관 사용자의 이용 패턴을 이해하고, 정보 이용자에게 새로운 행동 양식을 유도하며 나아가 유의미한 정보 및 서비스를 제공할 수 있는 주요한 툴로 활용될 수 있을 것으로 기대됨.

빅데이터 분석은 분석 대상에 대한 최적의 데이터 분석 방법을 선정하고, 데이터 분석을 통해 가치 있는 요소 및 문제를 발굴하는데 의미가 있음. 최적의 데이터 분석 방법을 선정하기 위해서는 해당 데이터에 대한 심도 있는 이해가 선제적으로 필요하며, 데이터의 결합 방식, 크기 및 제공 정보 등을 기본적으로 분석하였음.

연구 방향

통계적 지식이 전혀 없는 사람도 도서관 이용 패턴을 손쉽게 이해할 수 있도록, 직관적으로 이해 가능한 정보 및 서비스 도출하는 것을 방향으로 설정. 특히, 서비스 제공자와 이용자 모두에게 흥미를 유발할 수 있는 탐색적 자료 분석인 (1) 다이내믹 시각화, (2) 반응형 시각화 개발에 집중

연구 방법

- ○ 데이터 융합을 통한 새로운 정보 도출
  - 사람-학과, 사람-과목의 데이터 연계를 통한 새로운 정보를 도출
- ○ 다이내믹/반응형 시각화 구현

- 데이터 시각화는 데이터를 이해하는 가장 효과적인 방법임. 특히 실적을 보여주고, 추세에 대해 커뮤니케이션하며, 새로운 전략의 영향력을 해석하는 등, 주요 행동-결과의 관계를 보여주고자 하는 많은 곳에서 사용되고 있음

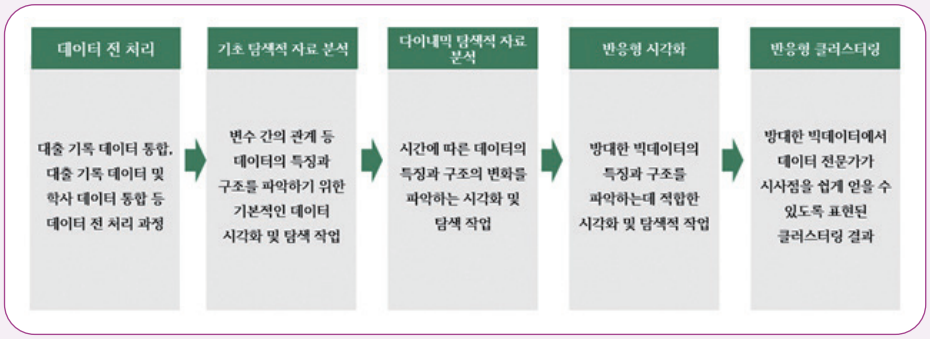
○ 도서관 데이터 시각화 구현 방향

- 다양한 기준으로 이용자들의 도서대출 순위를 보여주는 racing bar chart 구현
- 학과별 시간에 따른 문학과 비문학 도서대출 다이내믹/반응형 시각화 구현

○ 통계적 기계학습 기법을 이용한 데이터 분석(향후)

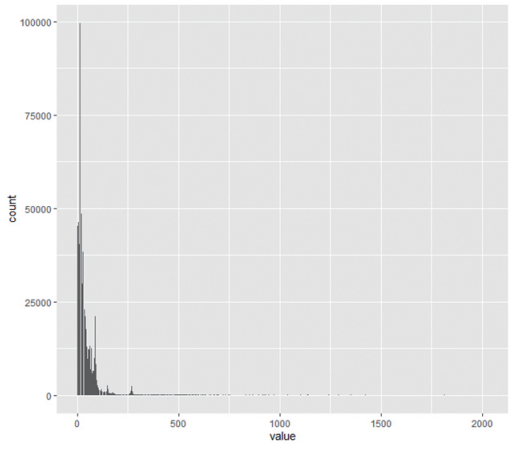
- 향후 연구에서 통계적 기계학습 기법 (예: 반응형 클러스터링)을 활용하여 보다 심도 있는 데이터 분석을 진행할 예정임

연구내용 및 결과



기초 탐색적 자료 분석 (Introductory EDA)

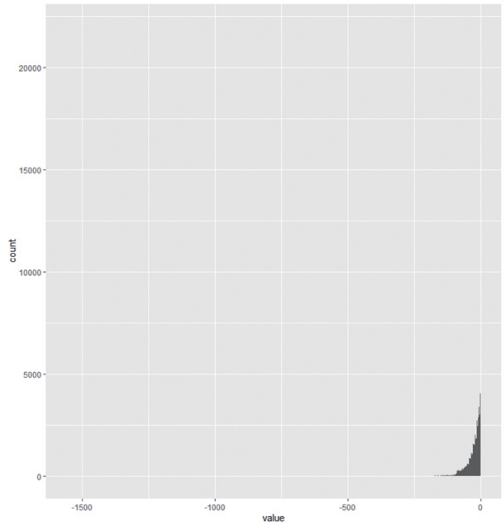
○ 반납날짜 - 대여날짜 그래프



days	count
0	535102
14	99514
7	71165
1	58913
28	58330
13	55535
42	48924
21	48525
15	47635
12	47009

- 0 이하 값들 제거 (음수 존재 288 cases)
- 대여 기간이 긴 책도 있어서 표로 나타내면 다음과 같음
- 14 days 기본 대여 기간, 이후 일주일 단위로 변동 보임

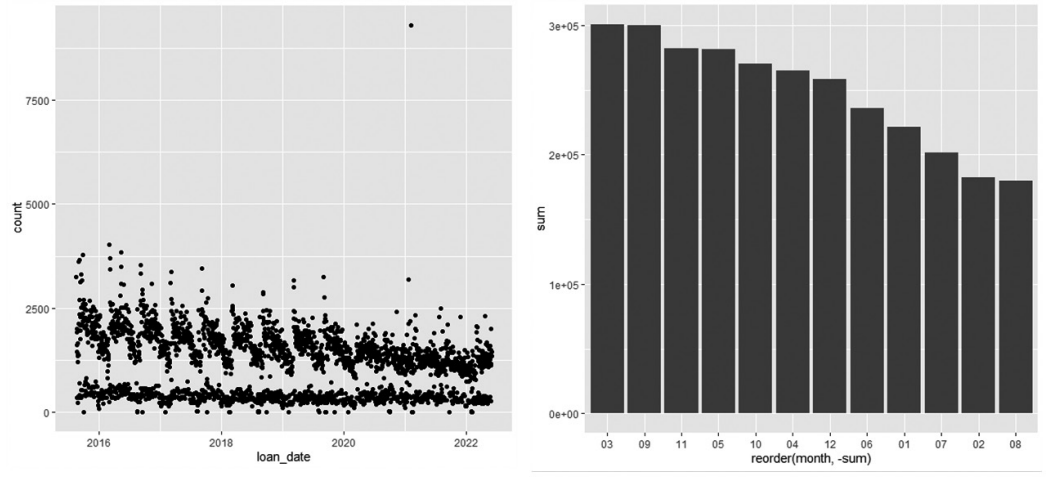
○ 예약 기간 그래프



days	count
0	22009
-1	4063
-7	3547
-6	3382
-2	3330
-4	3044
-3	3020
-5	2924
-8	2723
-14	2605

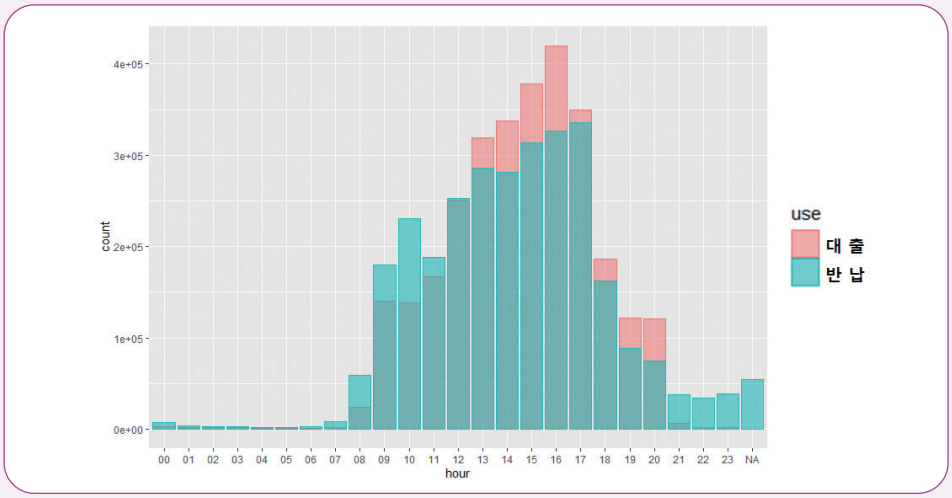
- 당일 예약 제일 많음
- 대체로 예약한 지 일주일 만에 대여

○ 날짜에 따른 도서 대여 건수



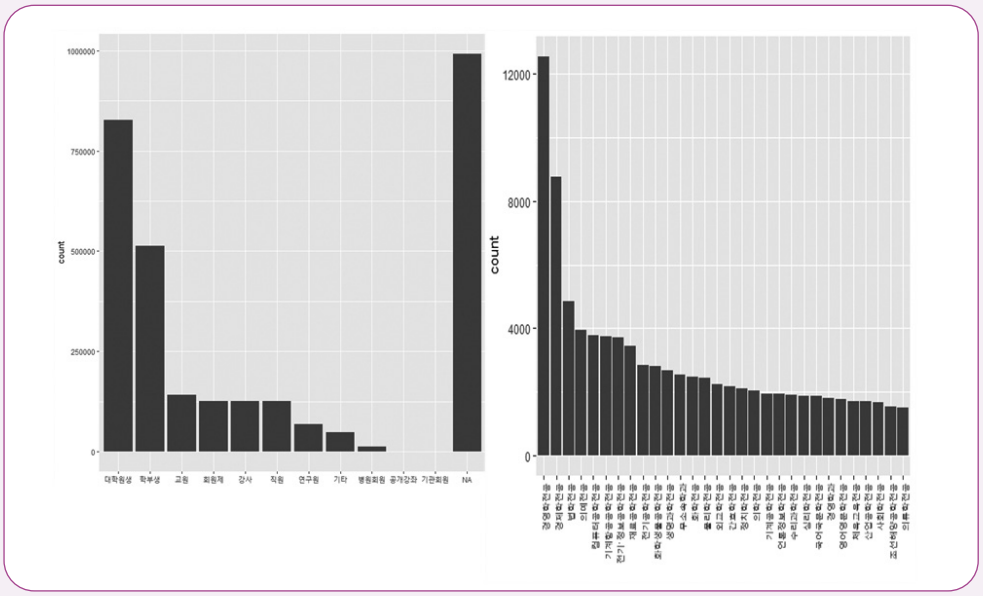
- 날짜에 따라 규칙적인 변동 보임
- 월별 나눠서 살펴보면 개강 달(3, 9), 중간고사 이후(11, 5), 남은 학기(10, 4) 순으로 감소

○ 시간에 따른 도서대출 및 반납



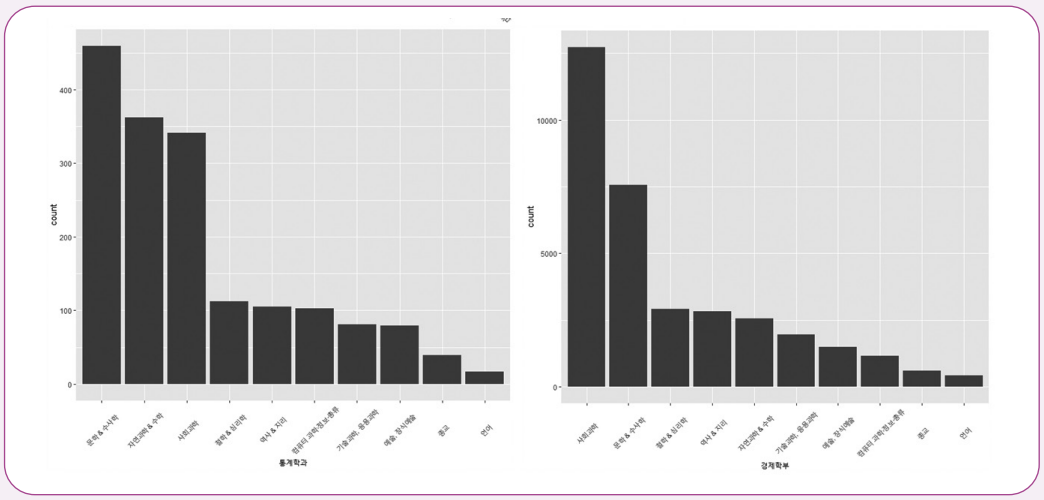
- 대출 시간은 오후에 몰려있고, 반납 시간은 비교적 고르게 퍼져 있음
- 대출과 반납 모두 8AM ~ 8PM에 이루어져 사람들이 학교에 많은 시간대에 주로 이루어지는 것을 알수 있음 하지만 늦은 시간(8PM ~ 11PM)에도 반납이 이루어 지는 것으로 보아 반납은 대출기한 도서대출에 영향 있음을 시사함

○ 사용자 소속에 따른 도서대출



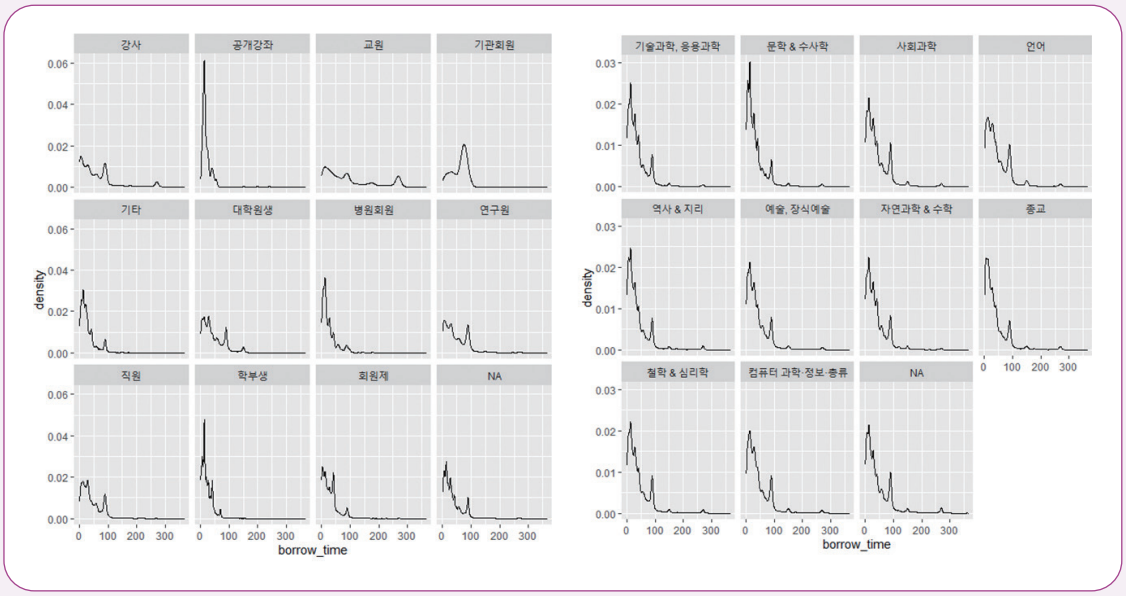
- 대학원생 > 학부생 > 교원 > 회원제 > 강사 > 직원 순으로 도서 이용
- 경영학전공 > 경제학전공 > 법학전공 > 의예 전공 > 컴퓨터공학 순으로 도서 이용
- 결측치 NA와 시간을 고려하지 않아 의미를 찾기 어려움

○ 사용자 소속별 대출 분야



- 통계학과: 문학&수사학 > 자연과학&수학 > 사회과학
- 경제학부: 사회과학 > 문학&수사학 > 철학&심리학
- 사용자 소속 그룹에 따른 책 추천을 달리할 필요성 확인

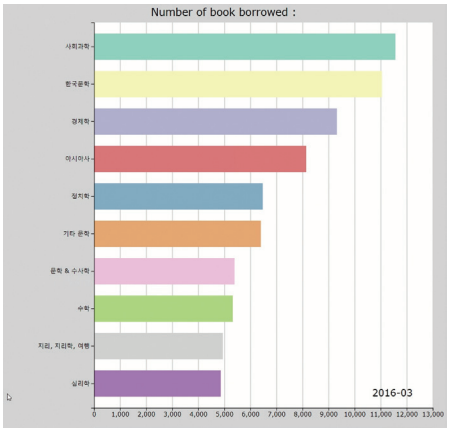
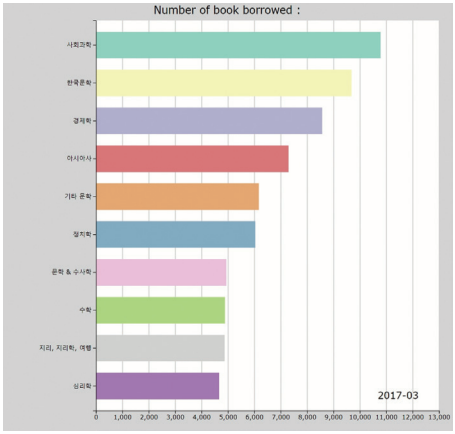
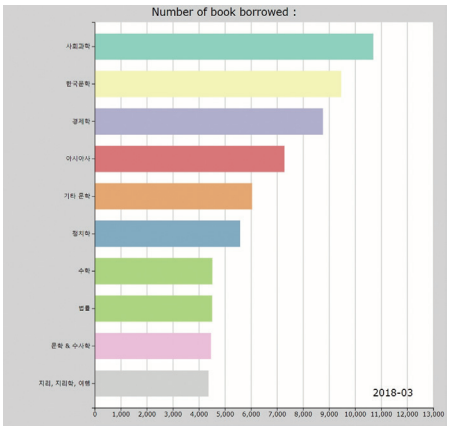
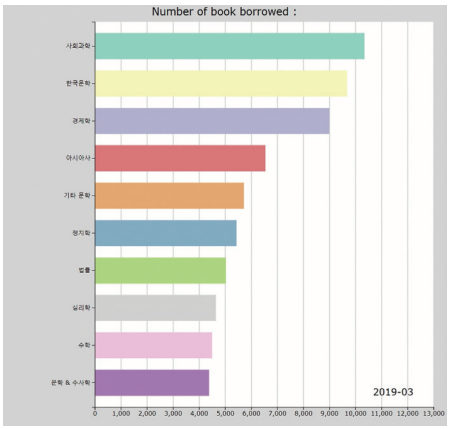
○ 이용자 및 도서 분야에 따른 도서 대여 기간

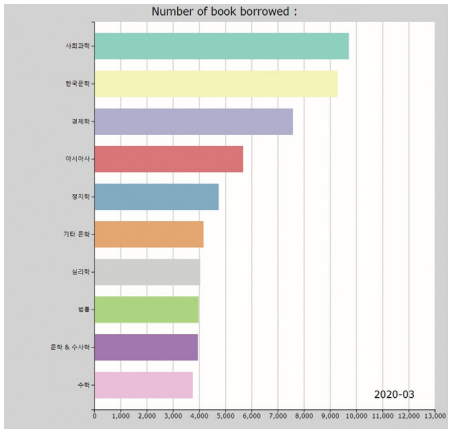
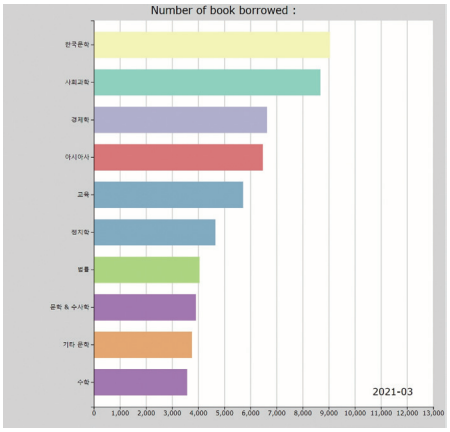


- 대여자 대부분은 책의 대여 기간이 기본적으로 주어지는 14일 대여기간 이내로 짧지만, 일부 강사, 교원과 연구원의 경우 대여연장을 통해 2달 이상씩 기간이 긴 경우가 있음
- 문학&수사학의 경우 책의 대여 기간이 역시 14일 이내로 대부분 짧지만, 언어의 경우 대여 기간이 2달 이상씩 긴 경우가 있음
- 하지만 대부분은 대여기간 패턴이 비슷하며 약 14일 기본 대여 기간 끝나는 날에 반납하는 경우가 많음
- 앞선 8PM - 11PM 반납시간의 패턴과 함께 고려할 때, 책을 실제로 읽는 시간이 짧더라도 여러 가지 이유로 대출 기한에 맞춰서 반납한다고 판단할 수 있음. 이를 기반으로 책 이용이 끝났을 때 바로 반납을 독려하는 제도가 필요하다고 판단됨

다이내믹 탐색적 자료 분석: Racing Bar Chart

○ 분야별 도서대출 순위

〈2016. 3. ~ 2016. 6.〉	〈2017. 3. ~ 2017. 6.〉
	
〈2018. 3. ~ 2018. 6.〉	〈2019. 3. ~ 2019. 6.〉
	

〈2020. 3. ~ 2020. 6.〉	〈2021. 3. ~ 2021. 6.〉
	

○ 2016년~2020년: 사회과학 〉 한국문학 〉 경제학 〉 아시아사 순으로 대여량이 많음

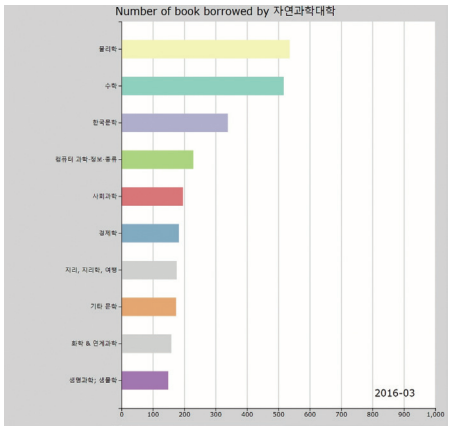
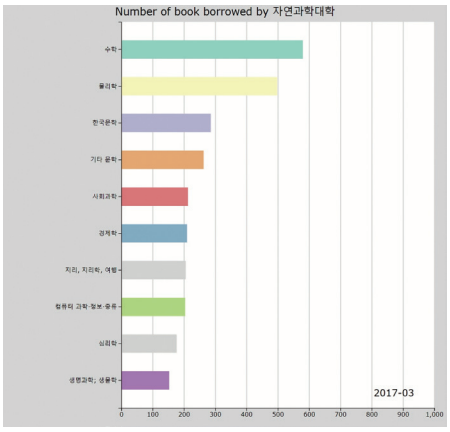
○ 2021년: 한국문학〉 사회과학 〉 경제학〉 아시아사 순으로 대여량이 많아짐
즉, 1등과 2등이 역전됨

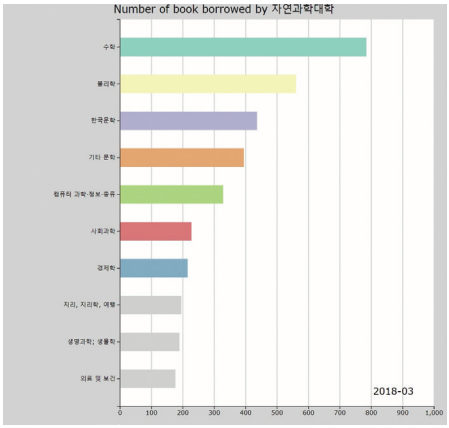
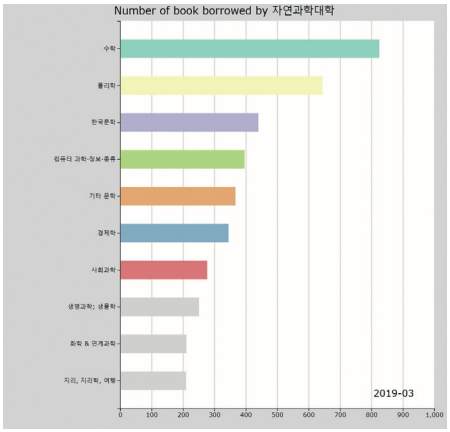
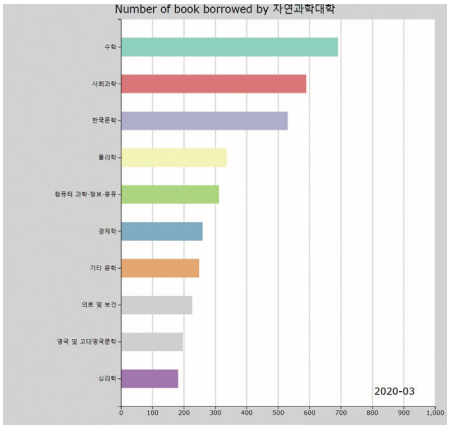
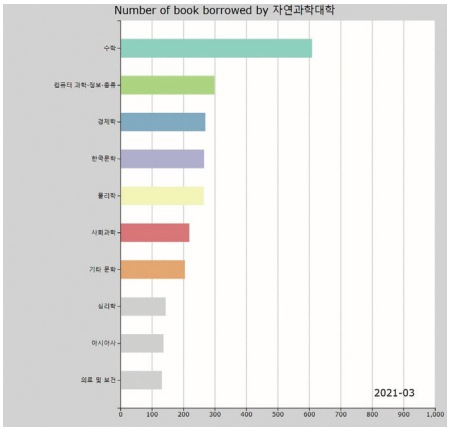
○ 기타문학과 수학은 시간이 갈수록 순위가 많이 떨어지는 경향이 있음

○ 교육과 법률은 시간이 갈수록 순위가 상승하는 경향이 있음

○ 코로나19로 인한 일시적인 영향인지 시대의 흐름인지 원인 파악이 필요

○ 자연과학대학 소속 이용자의 도서 분야별 순위

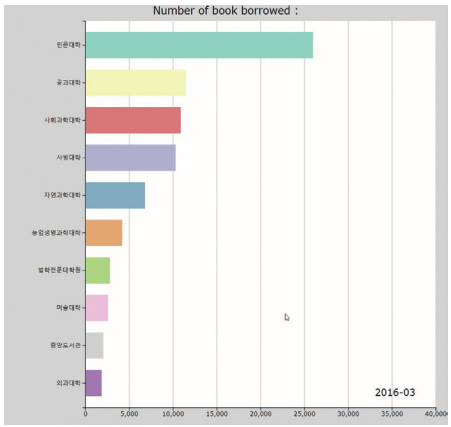
〈2016. 3. ~ 2016. 6.〉	〈2017. 3. ~ 2017. 6.〉
	

〈2018. 3. ~ 2018. 6.〉	〈2019. 3. ~ 2019. 6.〉
	
〈2020. 3. ~ 2020. 6.〉	〈2021. 3. ~ 2021. 6.〉
	

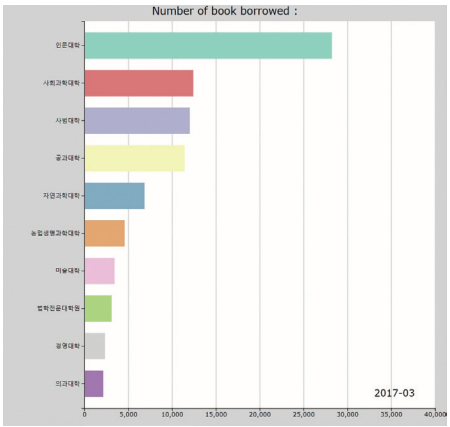
- 2016년: 물리학 > 수학 > 한국문학 분야 도서대출 순위를 보임
- 2017년~2019년: 수학 > 물리학 > 한국문학 분야 순으로 수학 관련 도서는 꾸준히 대출량이 늘어나지만, 물리학은 감소하는 경향을 보임
- 2020년: 수학 > 사회과학 > 한국문학 순으로 앞서 설명한 경향이 더욱 심해짐
- 사회과학 분야 도서대출량을 2020년 급격한 순위 상승을 보임
- 컴퓨터 과학 분야 도서대출량은 2021년에 급격한 순위 상승이 있음
- 현재는 3월~6월 경향을 분석했지만, 다른 시기 순위 및 대출량을 볼 필요성 있음

○ 단과대학별 도서대출량 순위

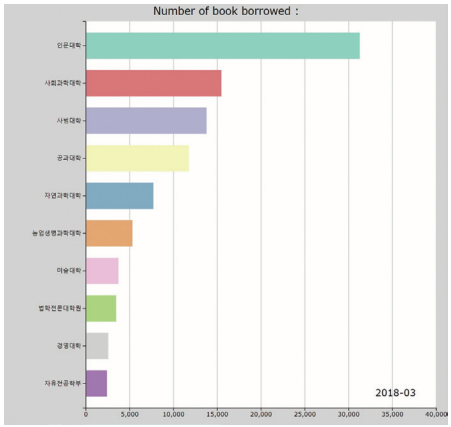
〈2016. 3. ~ 2016. 6.〉



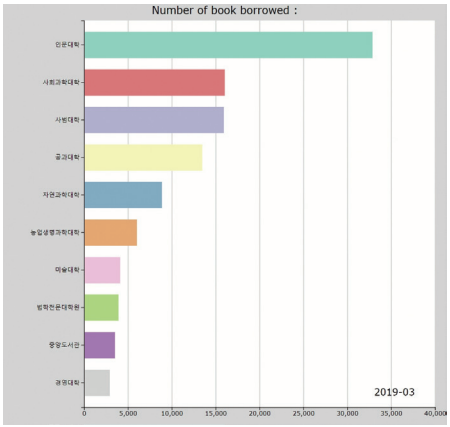
〈2017. 3. ~ 2017. 6.〉



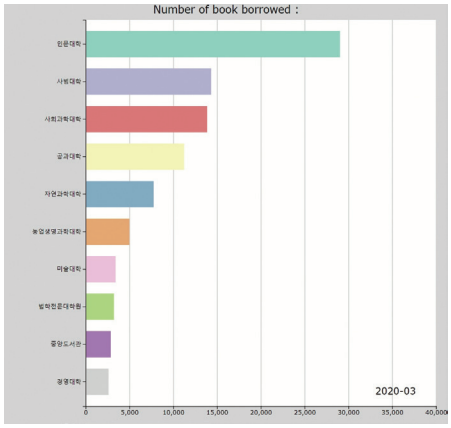
〈2018. 3. ~ 2018. 6.〉



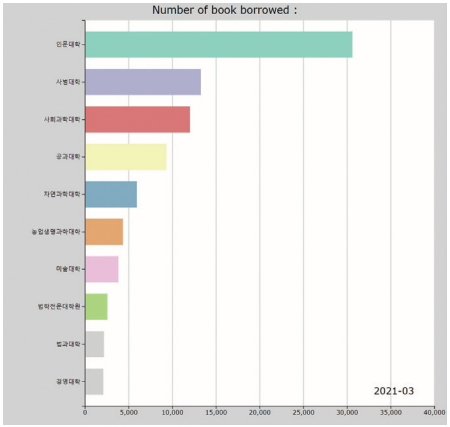
〈2019. 3. ~ 2019. 6.〉



〈2020. 3. ~ 2020. 6.〉

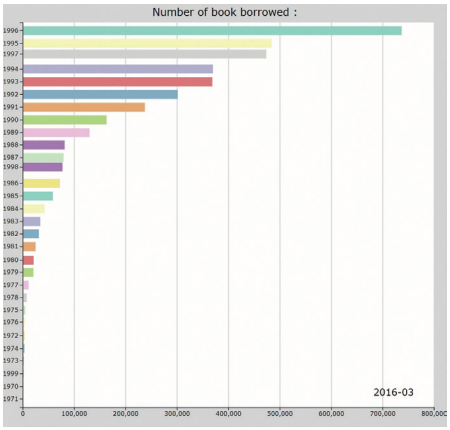
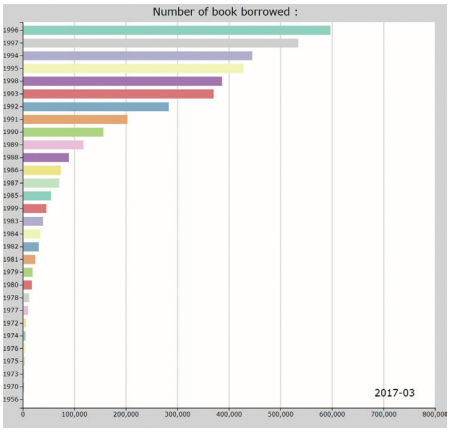
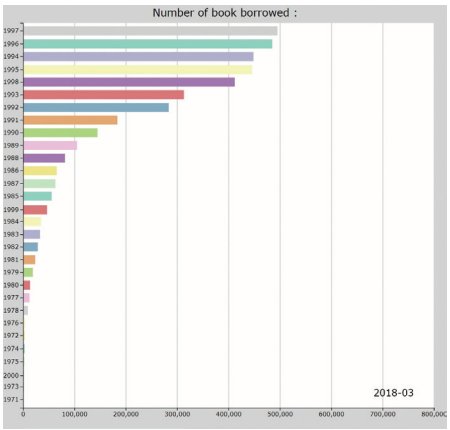
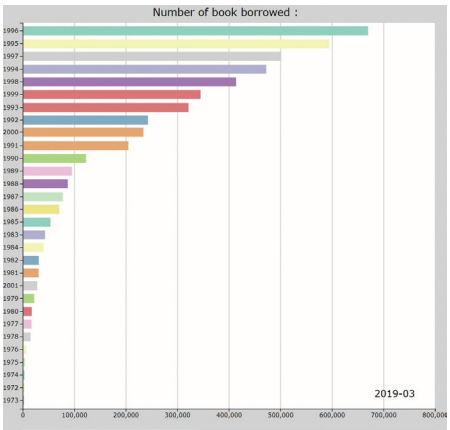


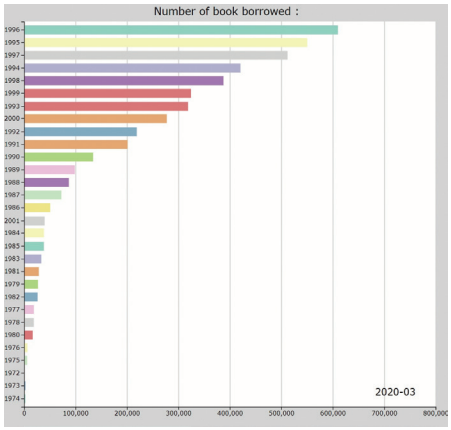
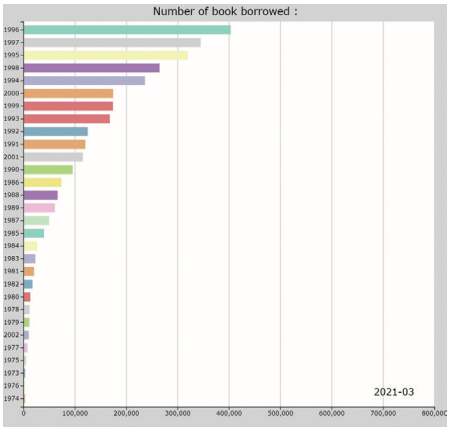
〈2021. 3. ~ 2021. 6.〉



○ 2016년: 인문대학> 공과대학 > 사회과학대학 > 사범대학
○ 2017~2019년: 인문대학 > 사회과학대학 > 사범대학 > 공과대학
○ 2020~2021년: 인문대학 > 사범대학 > 사회과학대학 > 공과대
○ 사범대학의 대출량이 꾸준히 상승하고 있으며, 공과대학은 꾸준한 감소세

○ 나이/출생연도별 순위

〈2016. 3. ~ 2016. 6.〉	〈2017. 3. ~ 2017. 6.〉
	
〈2018. 3. ~ 2018. 6.〉	〈2019. 3. ~ 2019. 6.〉
	

〈2020. 3. ~ 2020. 6.〉	〈2021. 3. ~ 2021. 6.〉
	

○ 2016년: 1996년 〉 1995년 〉 1997년 〉 1994년 〉 1993년 〉 1992년

○ 2017년: 1996년 〉 1997년 〉 1994년 〉 1995년 〉 1998년 〉 1993년

○ 2018년: 1997년 〉 1996년 〉 1994년 〉 1995년 〉 1998년 〉 1993년

○ 2019년: 1996년 〉 1995년 〉 1997년 〉 1994년 〉 1998년 〉 1999년

○ 2020년: 1996년 〉 1995년 〉 1997년 〉 1994년 〉 1998년 〉 1999년

○ 2021년: 1996년 〉 1997년 〉 1995년 〉 1998년 〉 1994년 〉 2000년

○ 1996년생의 도서대출량이 제일 많음

○ 1995년생의 경우 2016년 이후 도서대출량이 꾸준히 감소

○ 1997년생의 경우 2016년 ~2018년까지 대출 순위가 꾸준히 상승 후 2019년부터는 상위 자리매김

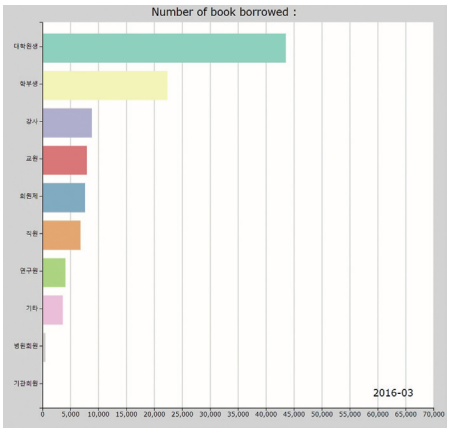
○ 1998년생의 경우 앞선 95, 96, 97년생의 패턴을 따르지 않고 5위권 자리매김

○ 1999년생의 경우에도 1998년생과 마찬가지로 하위권 자리매김

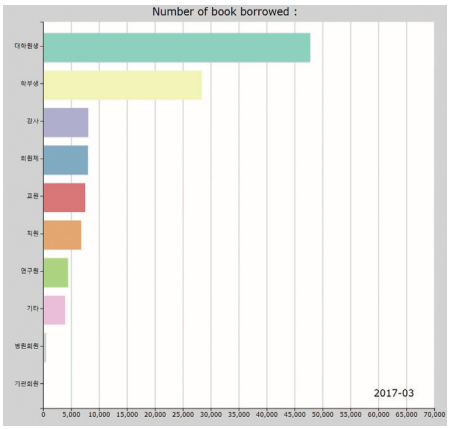
○ 1998, 1999년생의 패턴이 코로나19로 인한 일시적 경향인지, 장기적인 흐름인지는 살펴보아야 함

○ 그룹에 따른 총대출량 순위

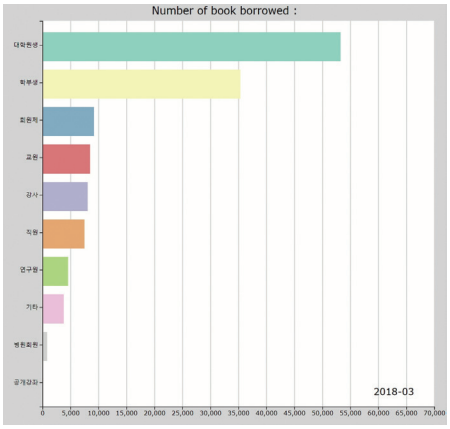
〈2016. 3. ~ 2016. 6.〉



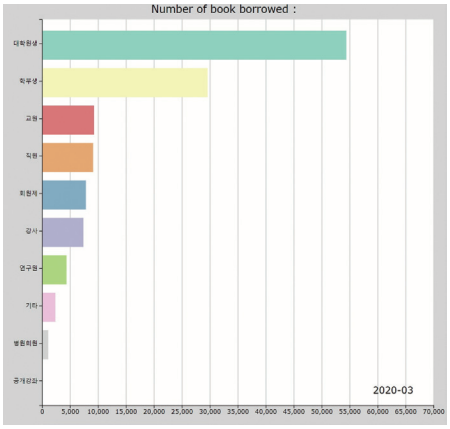
〈2017. 3. ~ 2017. 6.〉



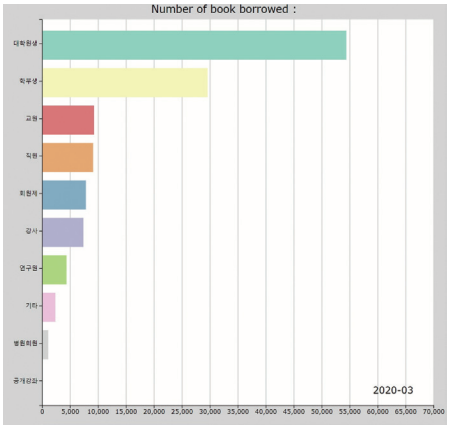
〈2018. 3. ~ 2018. 6.〉



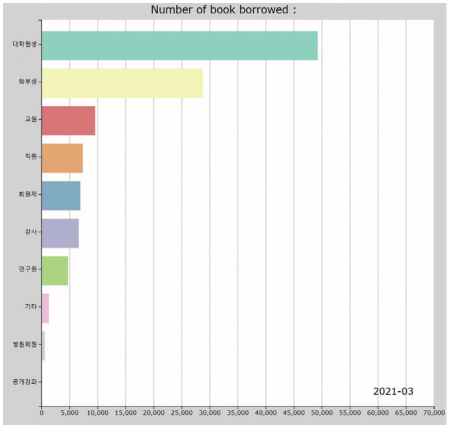
〈2019. 3. ~ 2019. 6.〉



〈2020. 3. ~ 2020. 6.〉



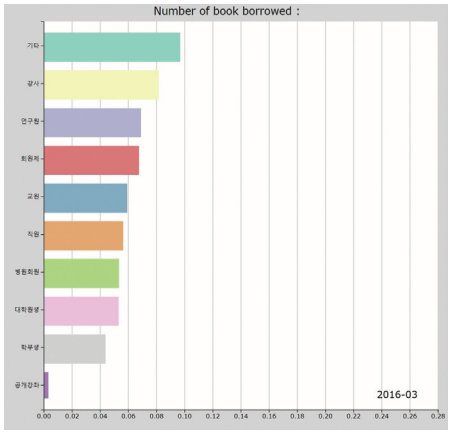
〈2021. 3. ~ 2021. 6.〉



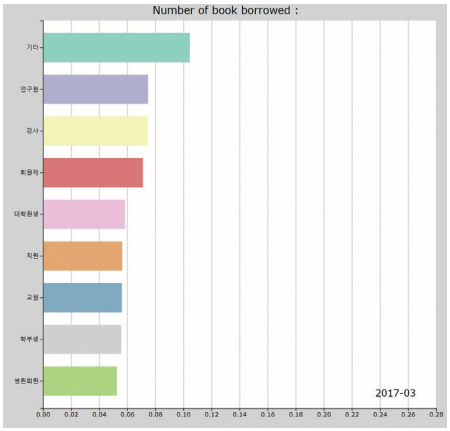
○ 2016년: 대학원생 > 학부생 > 강사 > 교원 > 회원제
○ 2017년: 대학원생 > 학부생 > 강사 > 회원제 > 교원
○ 2018년: 대학원생 > 학부생 > 회원제 > 교원 > 강사
○ 2019년: 대학원생 > 학부생 > 교원 > 회원제 > 직원
○ 2020-21년: 대학원생 > 학부생 > 교원 > 직원 > 회원제
○ 대학원생과 학부생 1, 2위는 변화 없음
○ 강사의 대출 순위가 떨어지는 추세

○ 그룹에 따른 1인당 대출량 순위

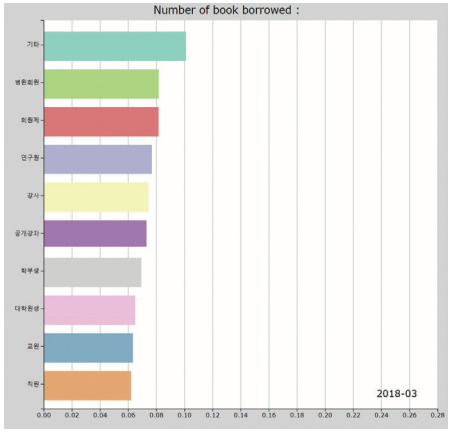
〈2016. 3. ~ 2016. 6.〉



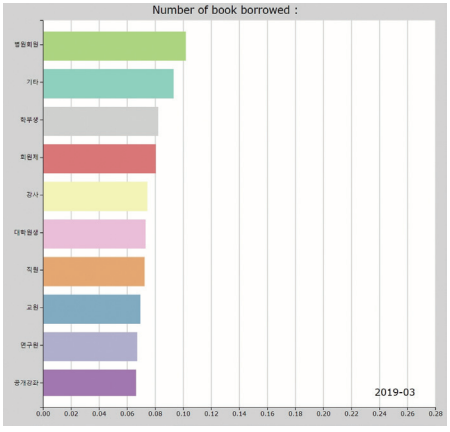
〈2017. 3. ~ 2017. 6.〉

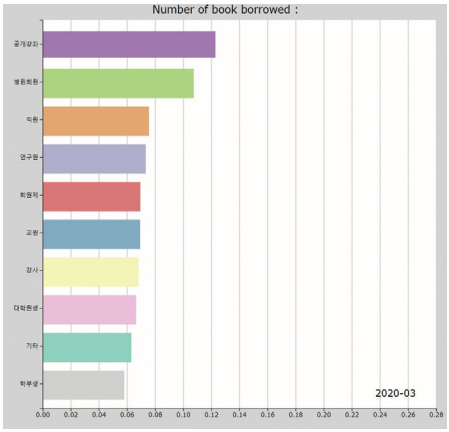
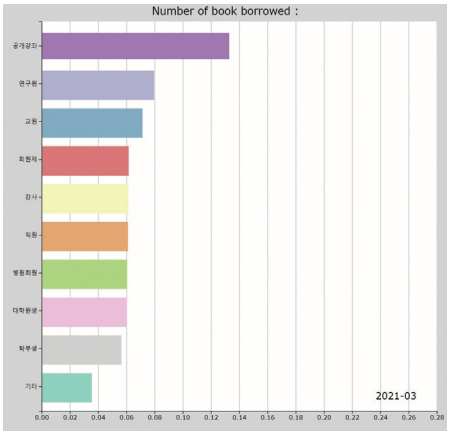


〈2018. 3. ~ 2018. 6.〉



〈2019. 3. ~ 2019. 6.〉



〈2020. 3. ~ 2020. 6.〉	〈2021. 3. ~ 2021. 6.〉
	

○ 2016년: 기타> 강사> 연구원 > 회원제 > 교원 > 직원

○ 2017년: 기타> 연구원 > 강사 > 회원제 > 대학원생

○ 2018년: 기타> 병원회원 > 회원제 > 연구원 > 강사

○ 2019년: 병원회원 > 기타 > 학부생 > 회원제 > 강사

○ 2020년: 공개강좌 > 병원회원 > 직원 > 연구원 > 회원제

○ 2021년: 공개강좌 > 연구원 > 교원 > 회원제 > 강사

○ 1인당 도서대출은 1, 2위를 제외하고는 매년 비슷한 수준

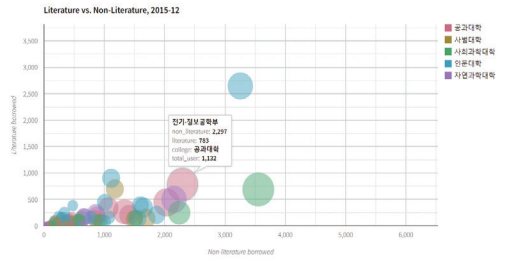
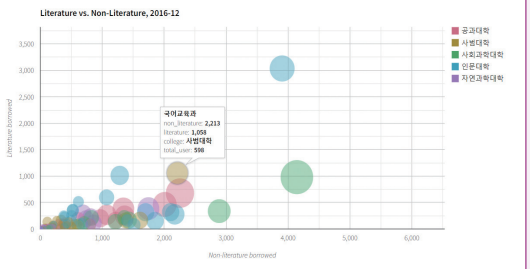
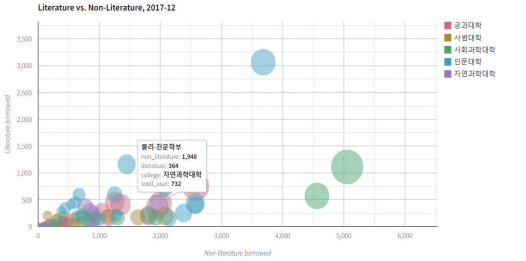
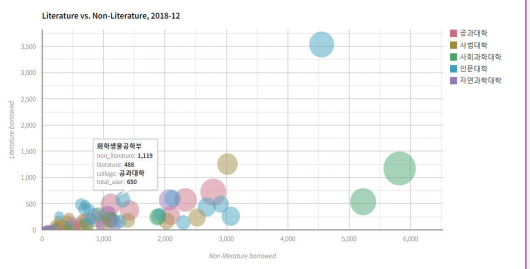
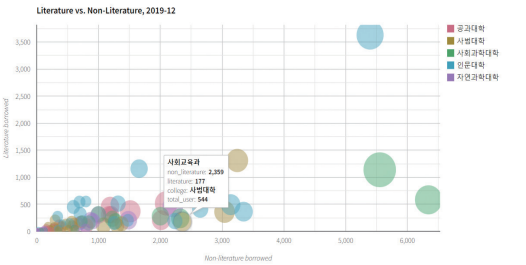
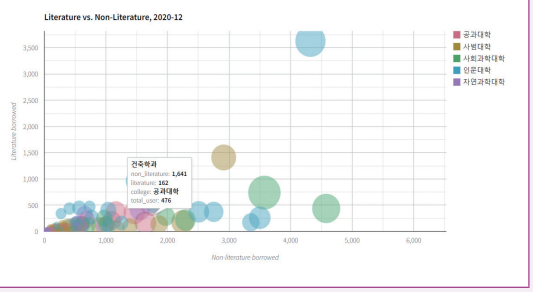
○ 기타회원은 수가 너무 적어서 신뢰할 수 없음

○ 2020년 이후 공개강좌 이용자들의 1인당 평균 대출량이 급증함

○ 학부생 회원의 경우 인원이 많아 총대출량은 많지만 1인당 평균 대출량은 적음

반응형 시각화(Google Chart)

○ 학과별 문학과 비문학 대여 관계: 대출량

〈2015. 6. ~ 2015. 12.〉	〈2016. 6. ~ 2016. 12.〉
	
〈2017. 6. ~ 2017. 12.〉	〈2018. 6. ~ 2018. 12.〉
	
〈2019. 6. ~ 2019. 12.〉	〈2020. 6. ~ 2020. 12.〉
	

○ 대출량 상위 5: 공과대학, 사범대학, 사회과학대학, 인문대학, 자연과학대학

○ 개별 과마다 도서관 이용 인원을 원의 크기로 표현하였으며 같은 단과대학은 같은 색으로 표현

○ 같은 단과대에 속해있더라도 도서대출량이 매우 다르며, 문학과 비문학 대출 비율도 매우 다름

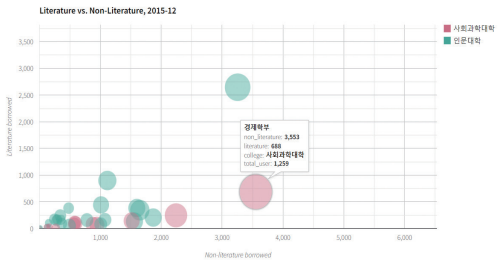
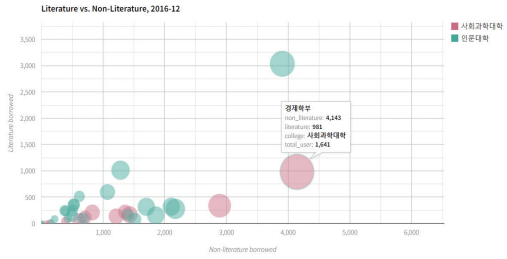
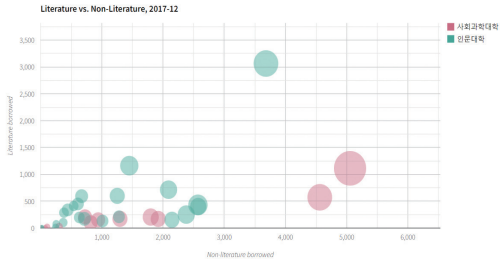
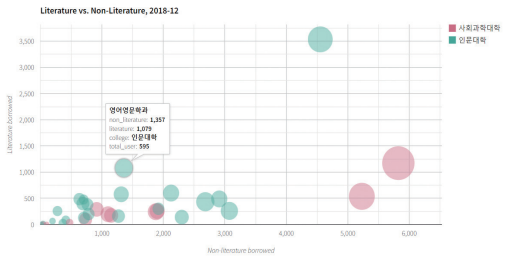
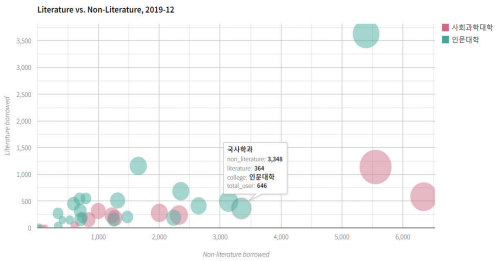
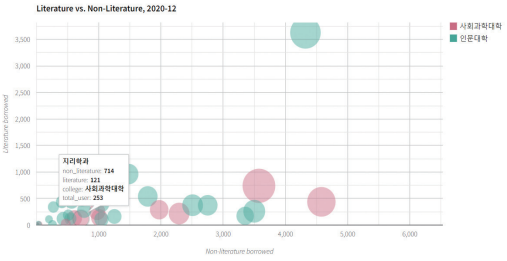
○ 대부분 비문학 도서대출량이 문학 도서대출량보다 현저히 많음

○ 인문대학의 경우 문학과 비문학 비율이 타 대학보다 유의미하게 높은 경향이 있음

○ 사회과학대학원의 경우 문학과 비문학 비율이 타 대학보다 유의미하게 낮은 경향이 있음

○ 시간이 지남에 따라 문학 도서대출량이 증가하였으나 문학/비문학 도서대출 비율은 일정하게 유지되는 경향이 있음

○ 학과별 문학과 비문학 대여 관계: 사회과학대학 vs 인문대학

〈2015. 6. ~ 2015. 12.〉	〈2016. 6. ~ 2016. 12.〉
	
〈2017. 6. ~ 2017. 12.〉	〈2018. 6. ~ 2018. 12.〉
	
〈2019. 6. ~ 2019. 12.〉	〈2020. 6. ~ 2020. 12.〉
	

○ 문학/비문학 도서대출 비율이 다르게 보이는 인문대학과 사회과학대학만 한정하여 분석

○ 인문대학의 경우 문학/비문학 도서대출 비율이 높은 학과와 사회과학대학과 비슷한 비율을 보이는 학과 두 개의 클러스터가 있는 것으로 보임

○ 비율이 높은 학과: 독어독문학과, 불어불문학과, 중어중문학과, 영어영문학과, 국어국문학과

○ 비율이 낮은 학과: 아시아언어문명학부, 서양사학과, 종교학과, 동양사학과, 미학과, 고고미술사학과 국사학

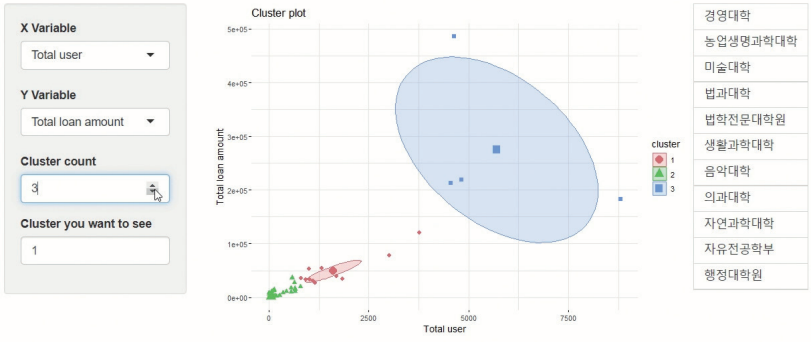
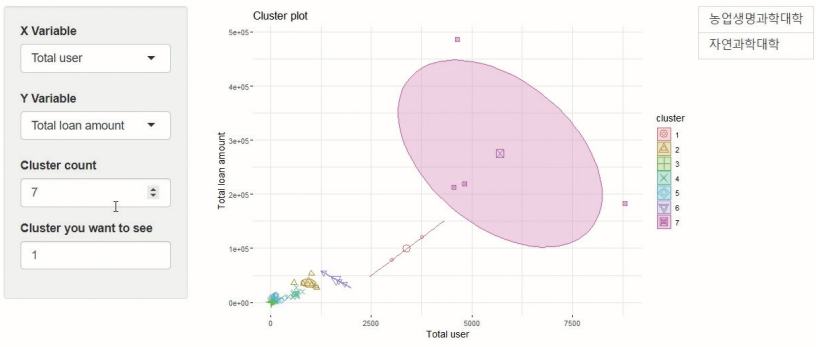
○ 코로나19 이후 문학 도서대출량보다 비문학 도서대출량이 더 많이 감소한 것을 확인유지되는 경향이 있음

반응형 클러스터링

- 두 가지 변수를 이용한 이용자 소속(e.g. 학과) 군집화 분석
- 방법: K-means 클러스터링
- 사용 변수

필드명	설명	비고
Total user	총 이용자 수	이용기준: 도서대출
Reuse rate	재이용률	Repeat user / Total user
Total loan amount	총 도서대출량	
Average age	총 이용자 나이의 평균	
Average rental period	평균 대출 기간	
College	소속대학	

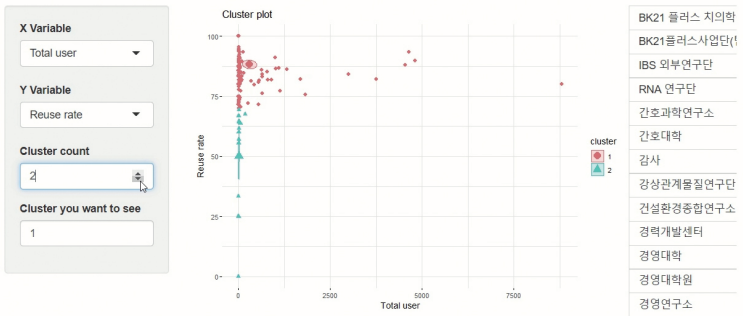
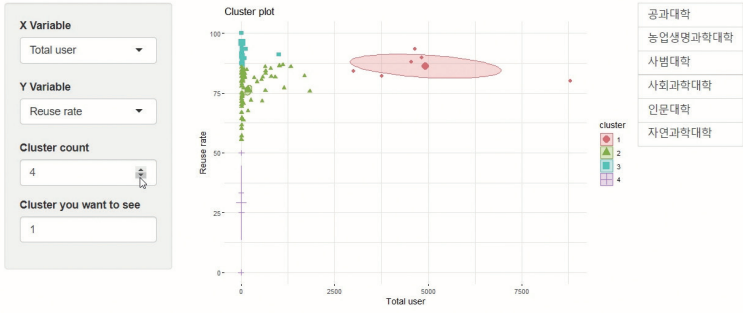
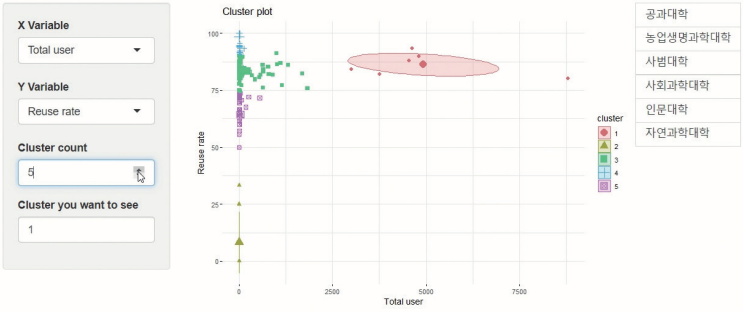
○ 총 이용자 수와 총 도서대출량 변수를 이용한 K-means 클러스터링 결과

클러스터	총이용자 수 vs 총 도서대출량
3	
7	

○ 총 이용자 수와 총 도서대출량이 강한 양의 상관관계를 가지고 있으며, 클러스터링 결과 두 변수 비슷한 수준의 값을 가진 소속대학들끼리 그룹화 함

○ 그룹의 수를 늘려도 비슷한 성격의 그룹만 확인됨. 일부 이 상치의 값을 가진 소속대학들도 관찰됨

○ 총 이용자 수와 재이용률을 이용한 K-means 클러스터링 결과

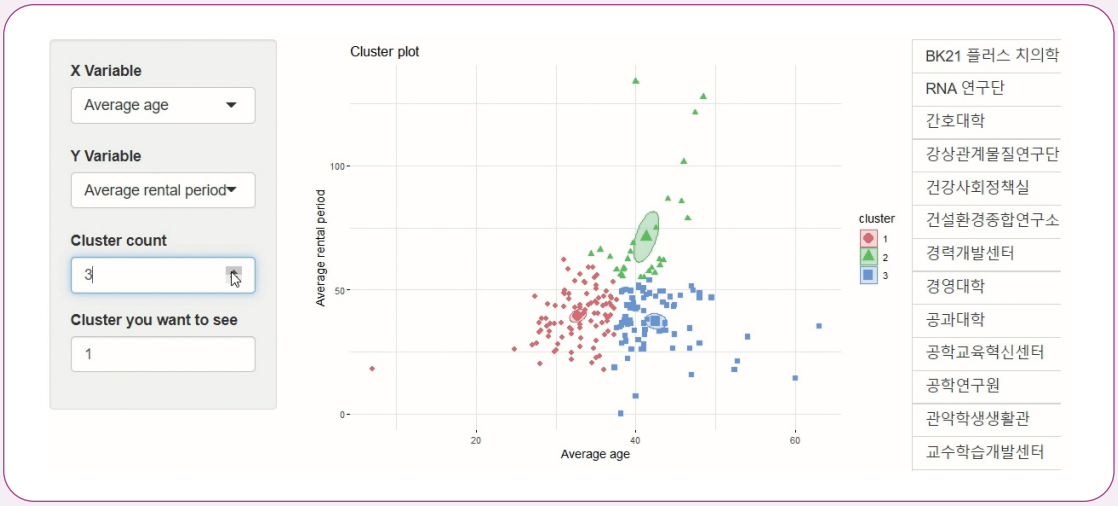
클러스터	총이용자 수 vs 재이용률
2	
3	
4	
5	

○ 총이용자 수가 많은 그룹은 재이용률이 75% 이상으로 높음

○ 하지만 총이용자 수가 작은 그룹의 경우 재이용률의 차이가 큼. 이 경우 소수의 인원이 재이용률에 큰 영향을 미칠 수 있어서 신뢰하기 어려움

○ K-means 클러스터링의 경우 그룹의 수를 바꾸어도 재이용률을 기준으로 그룹이 선택되는 경향이 있음

○ 이용자 나이 평균과 평균 대출 기간을 이용한 K-means 클러스터링 결과



○ 이용자 나이와 대출 기간은 큰 상관관계를 찾기는 어려움

○ 클러스터링 결과 나이가 적은 그룹 1, 나이는 많지만, 평균 대출 기간이 긴 그룹 2, 나이는 많지만, 평균 대출 기간이 짧은 그룹 3으로 나누어짐

향후계획

탐색적 자료 분석을 바탕으로 양질의 데이터 분석 방향 제공

- 시계열 데이터 분석 방법을 이용한 패턴 분석
- 분야별 책 대여량의 코로나19 영향력 파악

- 학과별 책 대여량의 코로나19 영향력 파악
- 패턴 분석을 이용하여 특정과의 도서대출량이 늘어난 혹은 감소한 원인 파악
- 다양한 기준으로 클러스터링 분석

⚙ 더욱 다양한 데이터 병합을 이용한 탐색적 자료 분석

- 현재는 문학과 비문학 간 총대출량의 상관관계를 단과대학/학과별로 살펴보고 있지만 이를 다른 변수들 고려 가능
- 허구와 실화 간 주제별 총대출량의 상관관계를 단과대학/학과별 분석
- 한국문학과 비 한국문학 총대출량의 상관관계를 단과대학/학과별 분석
- 문학과 비문학 간 총대출량의 상관관계를 나이별 분석
- 문학과 비문학 간 총대출량의 상관관계를 이용자 그룹별 분석

⚙ 잠재고객들의 관심을 유발하는 다양한 탐색적 자료 분석

- 빅데이터 기초통계 분석 사례: 배민트렌드 2022

⚙ 개인정보를 유출하지 않도록 데이터 마스킹 기법을 도입하여 도서관 데이터를 공개 데이터로 전환하고 연구에 자유롭게 활용할 수 있는 방법 모색