4차 6기 세미프로젝트

|  |
| --- |
| 서울시 도서관 별  대출 데이터 분석을 통한  아동 도서(ISBN 기준)  재배치 방안 도출 |

**2019년 12월 20일**

**서비스 산업 데이터를 활용한 머신러닝 분석 과정(A)**

6조

김사무엘

김외솔

서승우

**목 차**

**1. 프로젝트 개요**  **1-2**

1.1 프로젝트 기획 배경 및 목표 1

1.2 구성원 및 역할 1

1.3 프로젝트 추진 일정 2

**2. 프로젝트 결과**  **3-7**

2.1 데이터 수집 3

2.2 데이터 분석 3-6

2.3 데이터 분석 결과 6-7

**3. 기대 효과**  **7**

**4. 향후 발전 방향**  **7-8**

**5. 개발 후기**  **8-9**

**1. 프로젝트 개요**

1.1 프로젝트 기획 배경 및 목표

1. 배경: ‘서비스산업 데이터를 활용한 머신러닝 분석’ 과정 중에서 배운 통계 분석 도구 R을 활용하여, 서울시 공공도서관 아동도서(ISBN 기준) 재배치 방안을 강구.

* 이를 통해 이용자들이 원하는 도서가 없어 번거로운 타관도서 대출을 신청하거나 대출을 포기하는 상황을 방지하고, 미대출 장서의 이용률을 제고하고자 함

1. 목표: 서울시 공공도서관 도서 대출 데이터 및 서울시 인구 데이터를 분석해, 도서관 별로 ‘다른 도서관에 보내야 할 도서 목록’ 도출

1.2 구성원 및 역할

|  |  |  |  |
| --- | --- | --- | --- |
| 이름 | 전공 | 역할 | 구현 부분 |
| 김사무엘 | 철학과 | 팀장 | 도서관 대출 데이터 크롤링  크롤링 데이터 전처리 및 리스트 통합  데이터 시각화  PPT 발표 |
| 김외솔 | 경제학부 | 팀원 | 데이터 전처리  탐색적 데이터 분석  관련 도메인 리서치 |
| 서승우 | 경제학과 | 팀원 | 데이터 전처리  탐색적 데이터 분석  데이터 시각화 |

1.3 프로젝트 추진 일정

|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 기간 | 활동 | 비고 |
| 사전  기획 | 11.18 | 프로젝트 기획 및 팀 구성 |  |
| 11.18~19 | PJT주제 선정, 팀(PM/팀원) 구성 | 3인/팀 |
| 11.23 | 프로젝트 멘토링  [프로젝트 방향 설정 및 현업프로젝트 소개] | 현업 멘토 참여 |
| PJT  수행  /  완료 | 11.25~12.20 | 프로젝트 수행 |  |
| 11.25~12.17 | 프로젝트 설계 |  |
| 11.23 | 프로젝트 멘토링  [프로젝트 점검 및 기술자문 ] | 현업 멘토 참여 |
| 12.18~19 | 구현 및 테스트 |  |
| 12.20 | 팀별 최종 발표 (구축 완료 보고) | 최우수 한 팀 선발  멘토 평가 |

**2. 프로젝트 개발 결과**

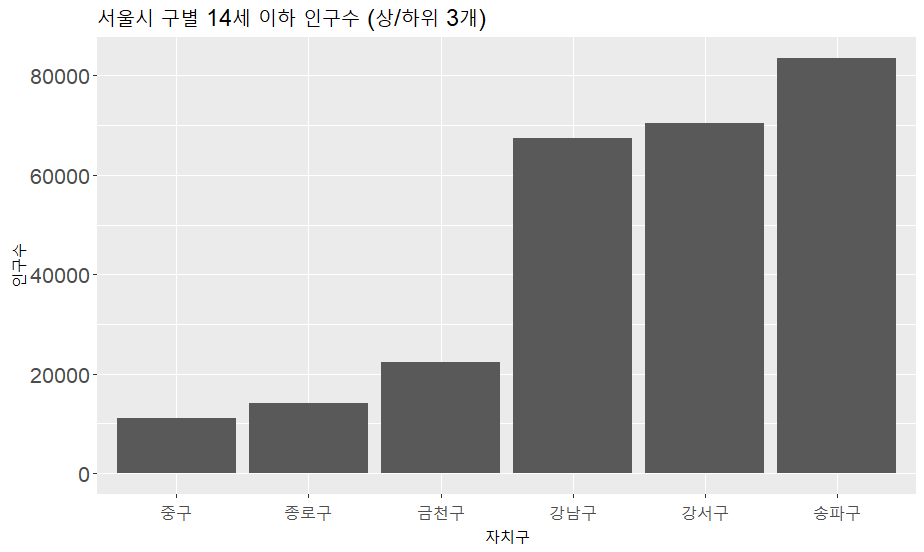
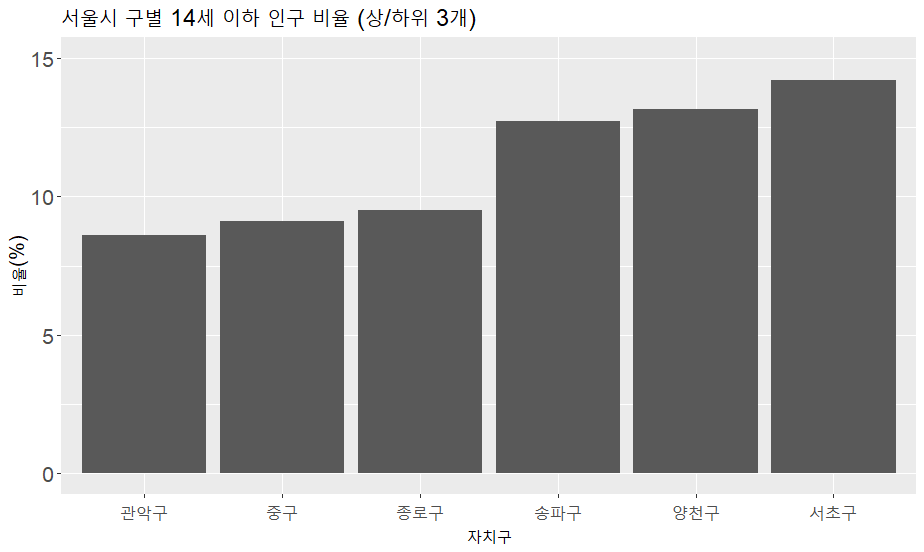
2.1 데이터 수집 (크롤링, API)

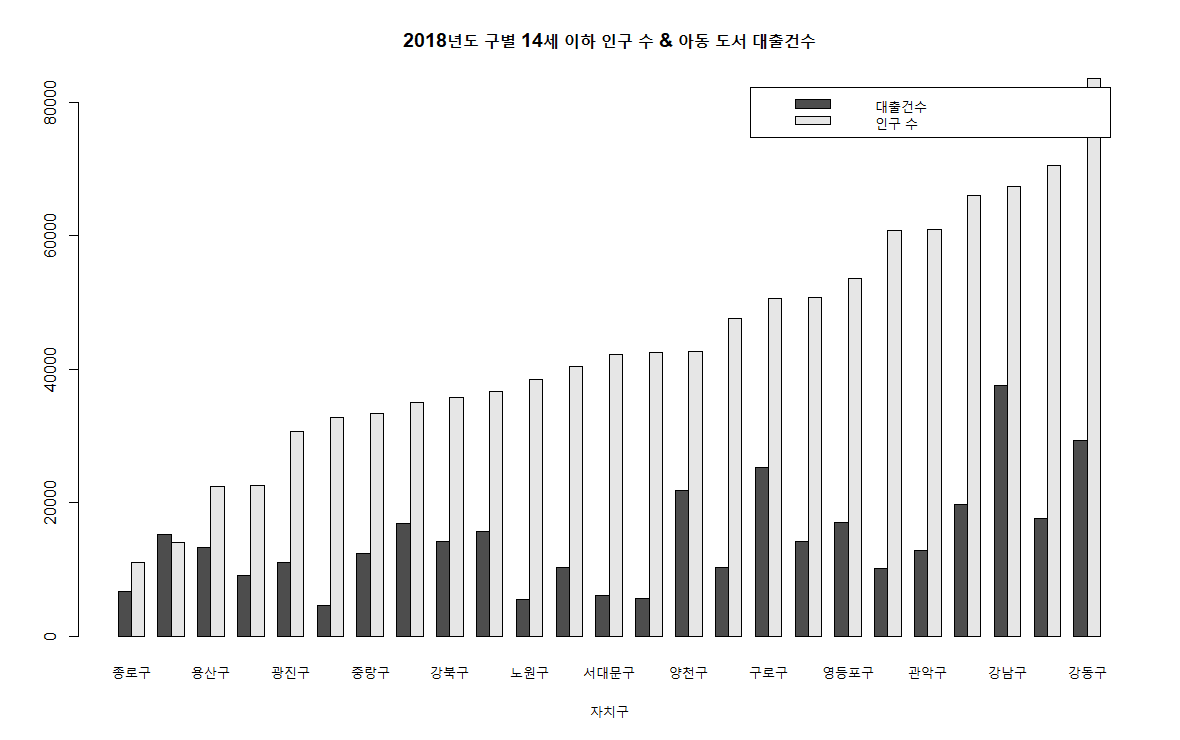
* 2016~2018년 서울시 구별 아동도서(ISBN 기준) 인기대출도서 데이터 (API 제공)
* 2018년 서울시 성북구/도봉구/강북구 도서관별 장서/대출 데이터 (csv 파일 크롤링)
* 서울시 구별 주민등록연앙인구(연령별) 통계 (csv 파일 다운로드)
* 서울시 구별 세대원수(세대원수별) 통계 (csv 파일 다운로드)

2.2 데이터 분석

2.2.1. EDA 및 시각화

* 2018년 서울시 구별 아동 인구비율 및 인구수 시각화 → 서울시 구별로 아동 인구비율 및 인구수가 상이하므로, 아동 관련 도서 대출 패턴도 상이할 것이라고 판단





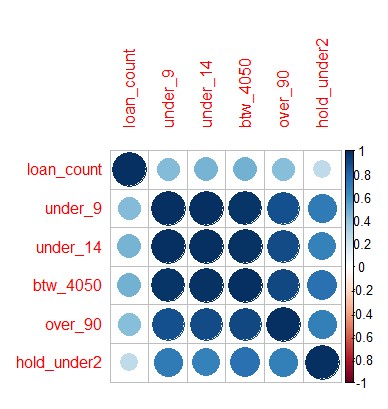
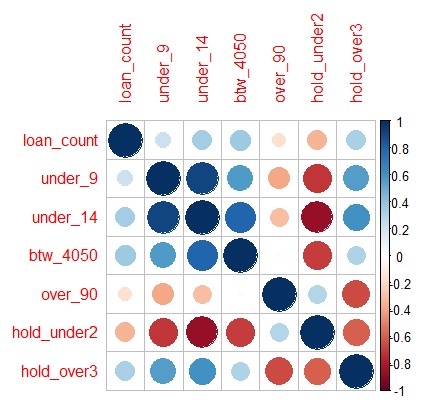
2.2.2. 상관관계 분석 및 시각화

2.2.2.1. 변수 설정

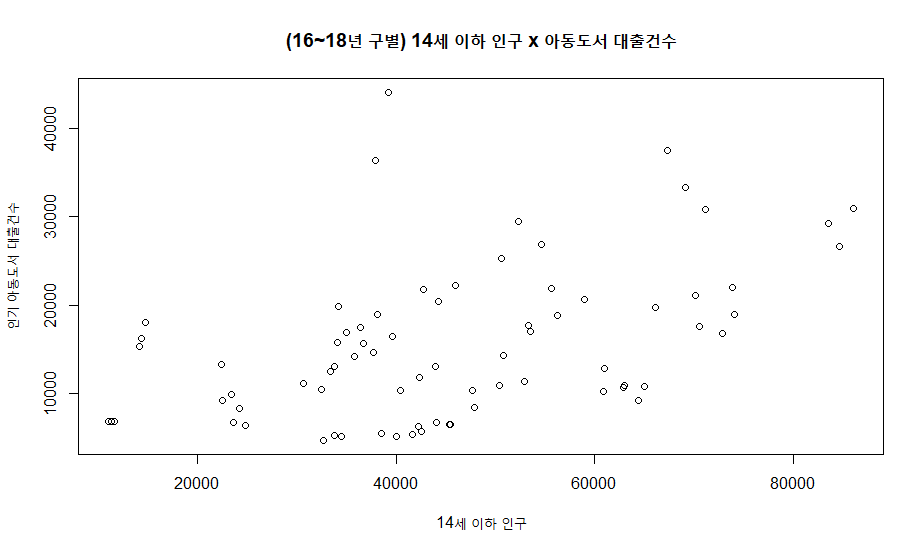
* <종속변수> 2016~2018년 서울시 구별 아동도서(ISBN 기준) 인기대출도서들의 총 대출 건수
* <독립변수 1> 서울시 구별 주민등록연앙인구(연령별) 중 9세 이하 인구 수 및 비율
* <독립변수 2> 서울시 구별 주민등록연앙인구(연령별) 중 14세 이하 인구 수 및 비율
* <독립변수 3> 서울시 구별 주민등록연앙인구(연령별) 중 40대 인구 수 및 비율
* <독립변수 4> 서울시 구별 주민등록연앙인구(연령별) 중 90세 이상 인구 수 및 비율
* <독립변수 5> 서울시 구별 세대원수(세대원수별) 중 2인 이하 가구 수 및 비율
* <독립변수 6> 서울시 구별 세대원수(세대원수별) 중 3인 이상 가구 비율

2.2.2.2. 결과 시각화

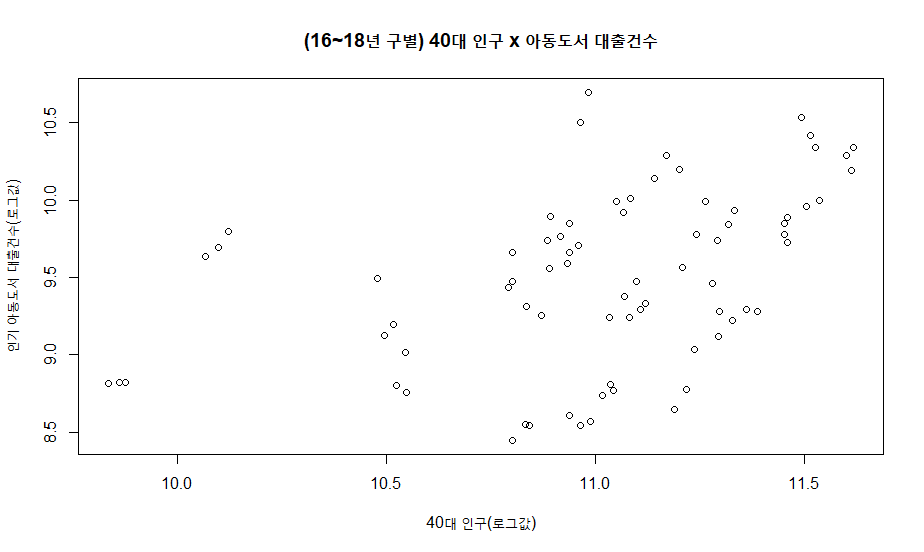
* 상관관계 그래프 시각화 ( 좌측 인구 절대수치, 우측 인구 상대비율 기준 )

* ‘14세 이하 인구 수’와 ‘아동 도서 대출 건수’ 간에 양의 상관관계 보임



* ‘40대 인구 수’와 ‘아동 도서 대출 건수’ 간에 양의 상관관계 보임



2.3 데이터 분석 결과 - 재배치 도서 선정

* 미대출 장서 비율 높은 도서관(둘리, 북서울중학교)에서, 송파구로 재배치할 도서 목록 도출
  + 기준 : 개관 이후 누적 대출건수 5건 미만이면서, 2018년 송파구 아동 인기도서 목록에 등재된 도서



* 14세 미만 인구 수가 적은 구(중구)에서, 송파구로 재배치할 도서 목록 도출
  + 기준 : 개관 이후 누적 대출건수 5건 미만이면서, 2018년 송파구 아동 인기도서 목록에 등재된 도서

****

**3. 기대 효과**

* 공공도서관 사용자 관점에서는 도서관 이용의 불편함(원하는 도서가 없어 번거로운 타관도서 대출을 신청하거나 대출 포기)을 줄이고, 운영자 관점에서는 미대출 장서의 이용률을 향상하며 타관도서 이동을 최적화할 수 있을 것으로 기대됨

**4. 향후 발전 방향**

* 분석 범위 세분화
  + 범위를 세분화해 데이터 왜곡 완화
  + 기간: ‘연(Year)’->‘월(Month)’
  + 공간: ‘구’ -> ‘동’
* 대출건수 예측 모델링
  + 인구 통계량 이외에 다양한 변수를 추가해, 대출건수를 예측하는 기계학습 모델 개발
  + ex) ‘월별 공휴일 수’, ‘해당 지역 초등학교 수’ 등의 변수 추가

**5. 개발 후기**

|  |
| --- |
|  |

|  |  |
| --- | --- |
| 성명 | 후기 |
| 김사무엘 | * 데이터 크롤링 과정에서 예상치 못한 에러가 많이 발생해, 익숙하지 않더라도 크롤링보다는 API를 적극 이용하는 것이 효율적인 프로젝트 진행을 위해 중요하다는 사실을 깨닫게 되었음. * 데이터 전처리 과정에서 결측치를 꼼꼼히 검토하지 않아 불필요한 시간과 노력을 낭비했음. 데이터 수집 및 분석 과정에서 결측치 다루기(Handling NAs)가 매우 중요하다는 사실을 절실히 느낄 수 있었음. * 크고 멋있는 과제를 설정하는 것보다, 작아 보이더라도 아주 구체적인 목표와 계획을 세우는 것이 매우 중요하다고 느꼈음. 데이터 상태가 좋다고 느끼다 보니 초반에 이것저것 하고 싶은 것이 많았고, 이후 시간이 지나면서 방향을 잃는 일이 발생했음.. 명확한 목표와 계획 설정의 중요성을 다시금 깨달았음 |
| 김외솔 | * 기존에 배운 프로그래밍 언어와 분석 방법을 활용하여 실제 빅데이터를 다뤄볼 수 있어서 좋았음. * 현재 본인이 무엇을 알고 있는지, 그리고 앞으로 더 알아가야 할 부분이 무엇인지 깨닫게 되는 유익한 시간이었음. * 해당 프로젝트가 향후 커리어 방향을 설정하는데 도움이 되었다고 생각함. |
| 서승우 | * 프로젝트 진행 기간 동안 데이터 전처리를 진행하면서 자연스레 함수의 활용법을 익힐 수 있었음. * 데이터의 구조와 성격을 먼저 파악하지 않고 무턱대고 데이터 전처리 작업을 진행하다보니 실수가 잦아져 작업 속도가 느려짐. 이에 summary() 및 class(), str() 등의 함수를 사용하여 결측치 혹은 데이터 성격을 파악하는 습관을 들여 작업 속도를 개선하는데 큰 도움이 되는 것을 느낌. |