서울대학교 중앙도서관 대출서비스

이용 현황 및 개선책

Index

- Data
- Data Exploration
- Problem
- Model
- Result
- Plan

Raw Data

- **데이터 종류** : 1) 사건별 데이터 (사건 : 대출/반납/예약/연장)
  2) 연체 데이터
- **기간** : 1년(2013년 11월 ~ 2014년 10월)
- **변수**

  1) 사건별 데이터
     - 도서 관련 정보 : 등록번호, 서명, 저자, 출판사, 출판년도, 청구기호
     - 사건 관련 정보 : 년월일, 시분, 초, 구분
     - 이용자 관련 정보 : 이용자, 신분, 학번, 대학
  2) 연체 데이터 : 등록번호, 서명, 저자, 출판사, 출판년도, 청구기호, 이용자, 신분, 학번, 대학, 대출일, 반납예정일, 반납일, 연체일
- **양** : 사건별 데이터 약 105만 건, 연체 데이터 약 7만 건

Data

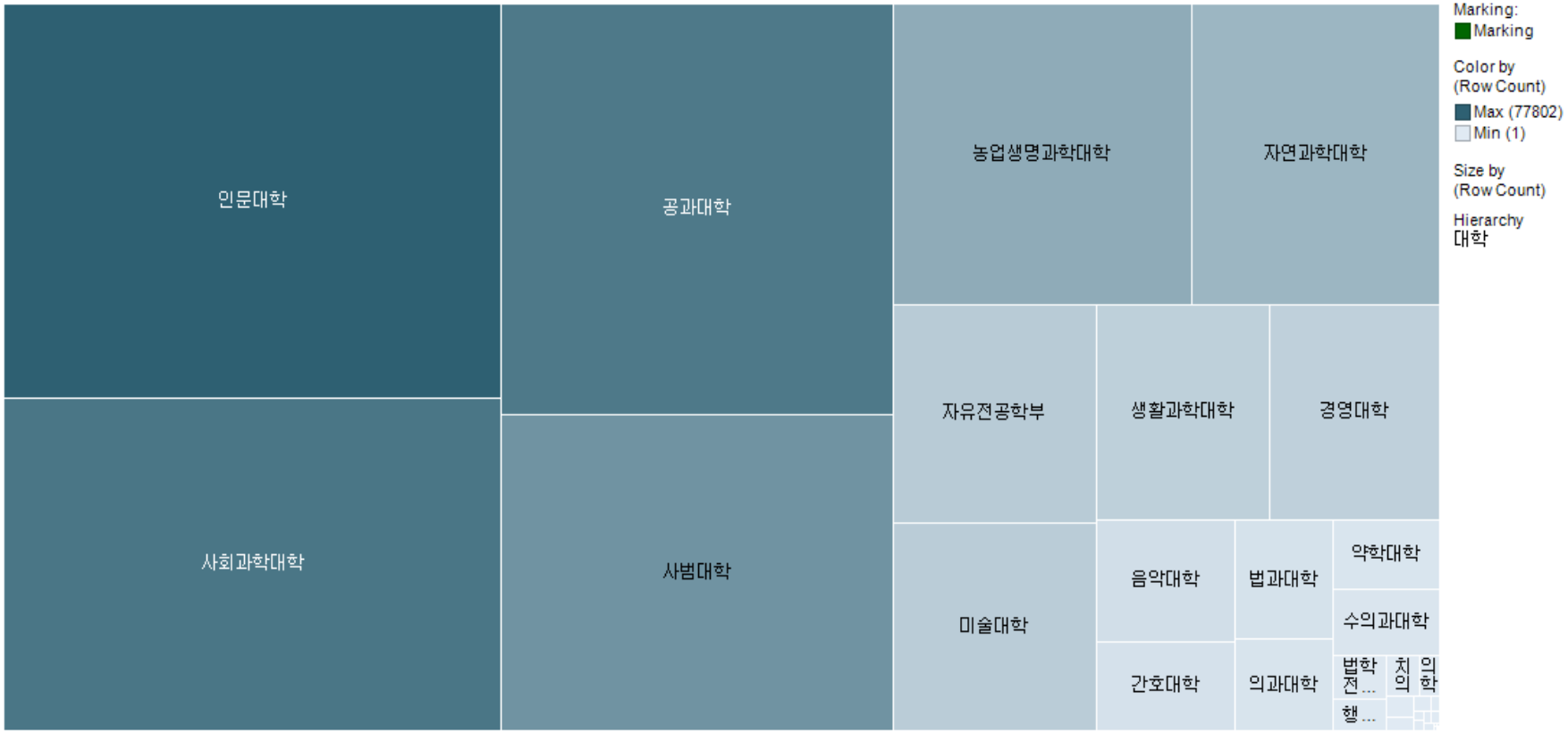
- **데이터 종류** : 1) 대출-반납 사건 기록
  2) 예약 횟수 기록
- **기간** : 1년(2013년 11월 ~ 2014년 10월)
- **변수**

1) 대출-반납 사건 기록
   - 도서 관련 정보 : 서명, 출판년도, 대분류, 중분류
   - 사건 관련 정보 : 대출년월일, 반납년월일, 연체여부, 연체일수
   - 이용자 관련 정보 : 이용자, 신분, 학번, 대학
2) 책별 정보 및 예약 횟수 기록
   - 서명, 저자, 출판사, 출판년도, 대분류, 중분류, 예약 횟수

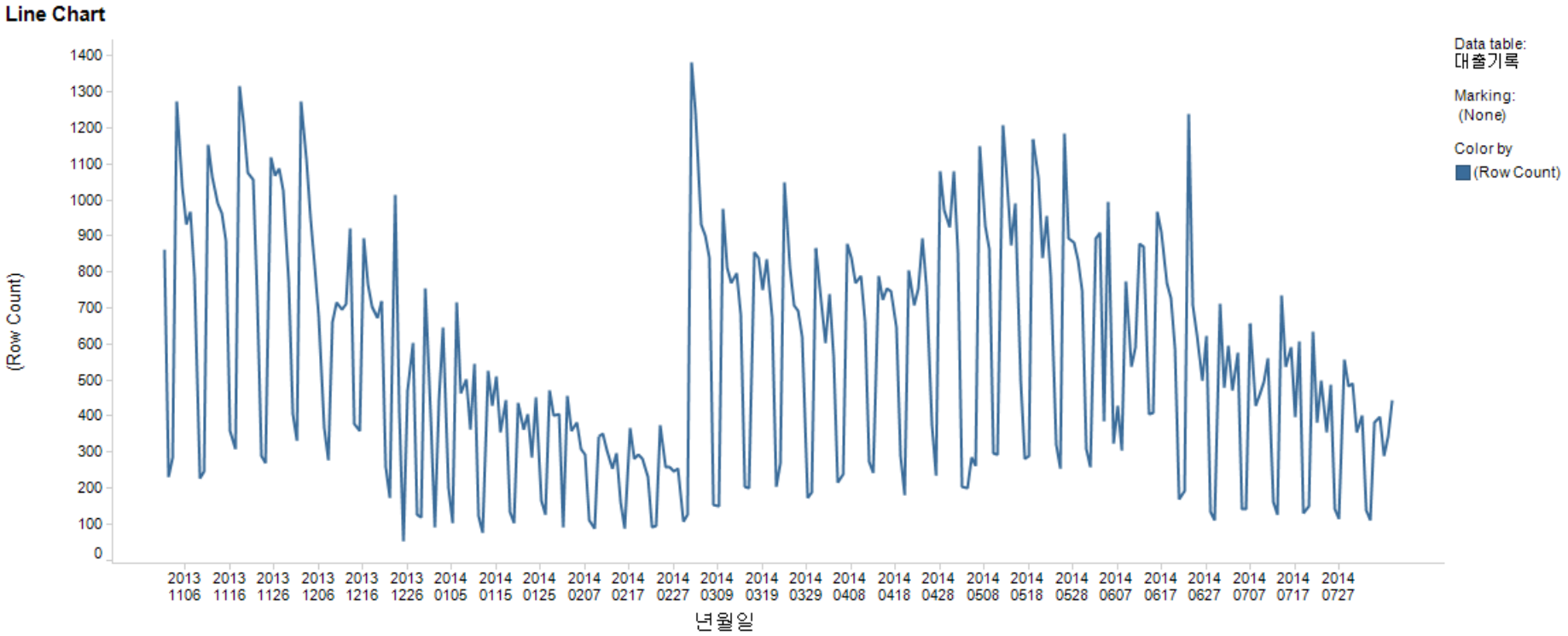
- **양** : 대출-반납 기록 약 48만 건, 예약 횟수 기록 약 1만 건

Data Exploration

Treemap

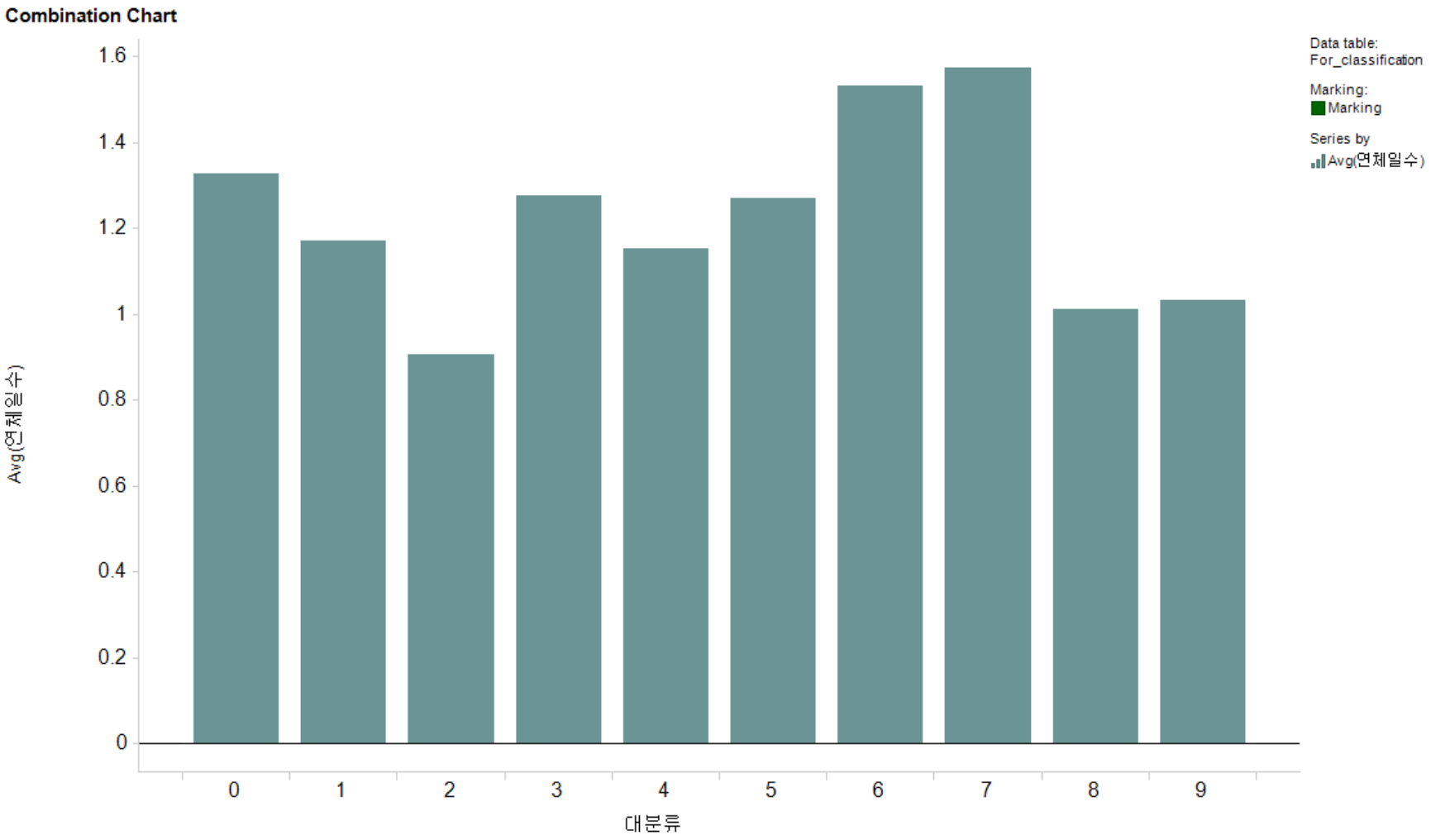


Data Exploration

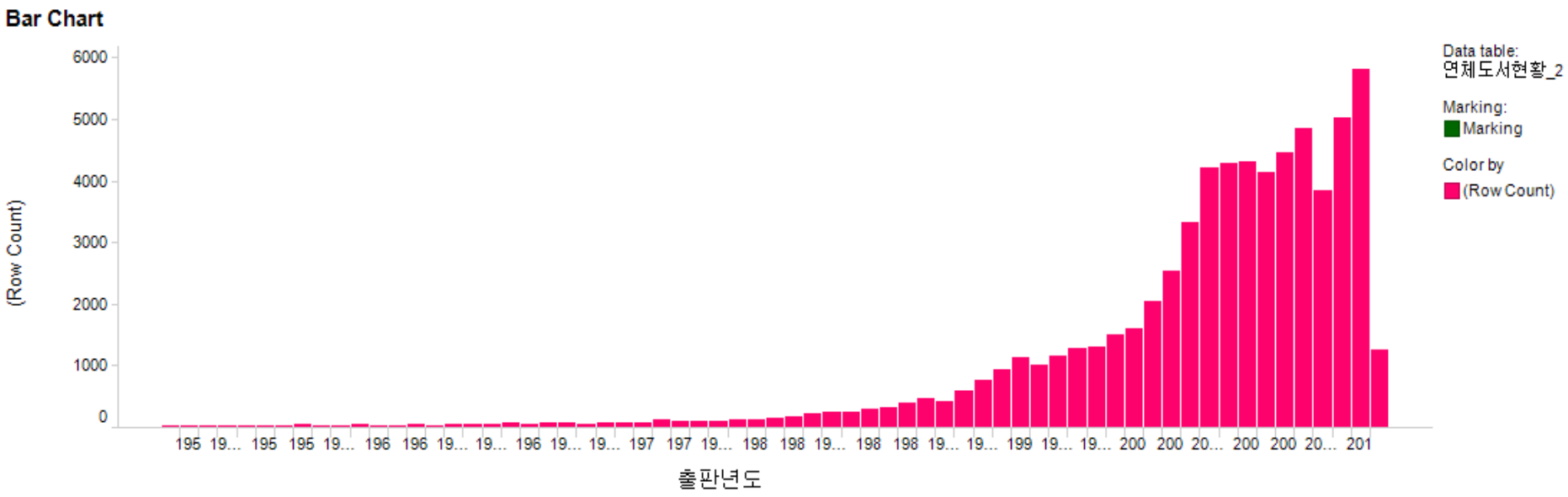


주기성 : 정규학기/계절학기/방학 (큰 주기), 요일(작은 주기)

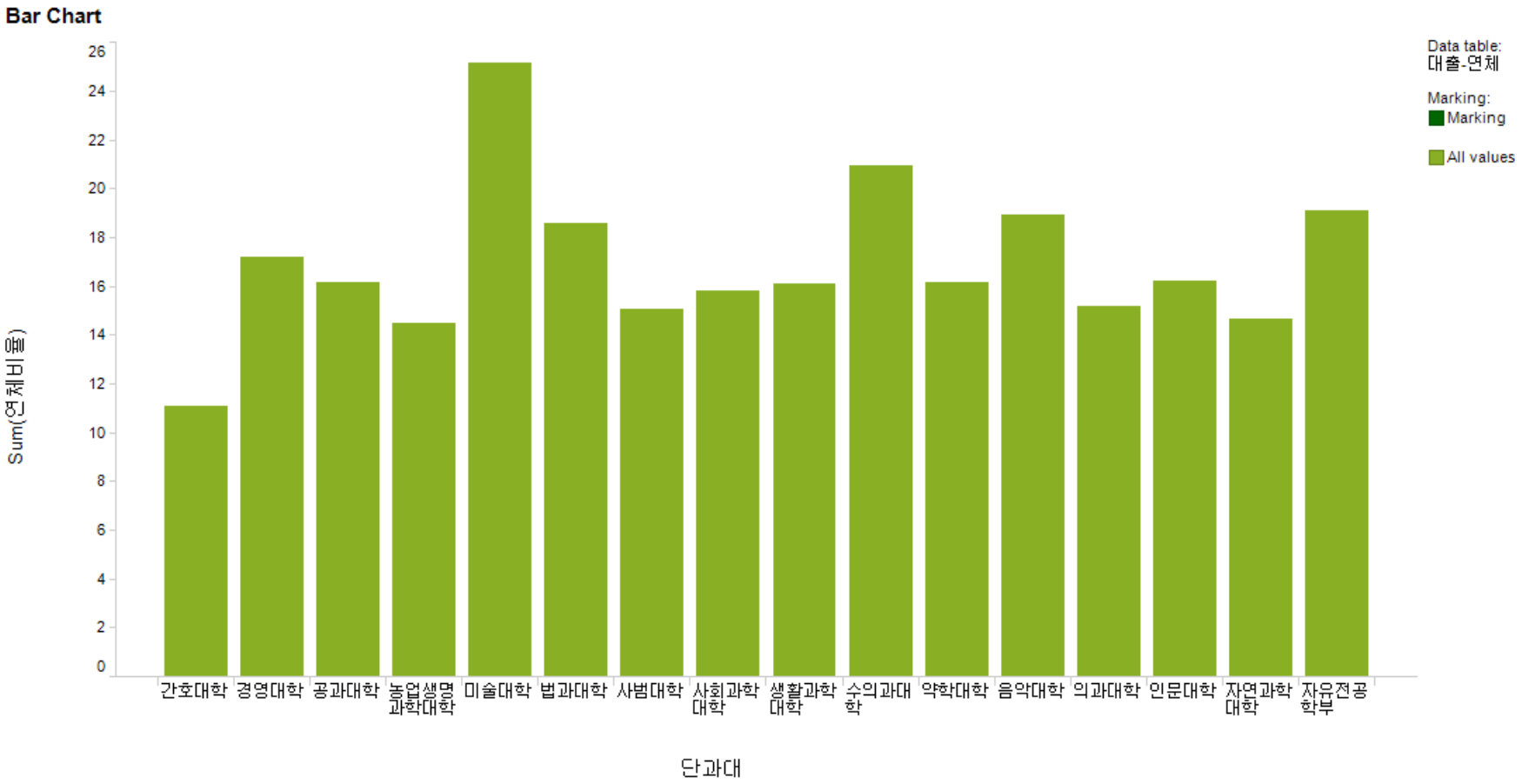
Data Exploration



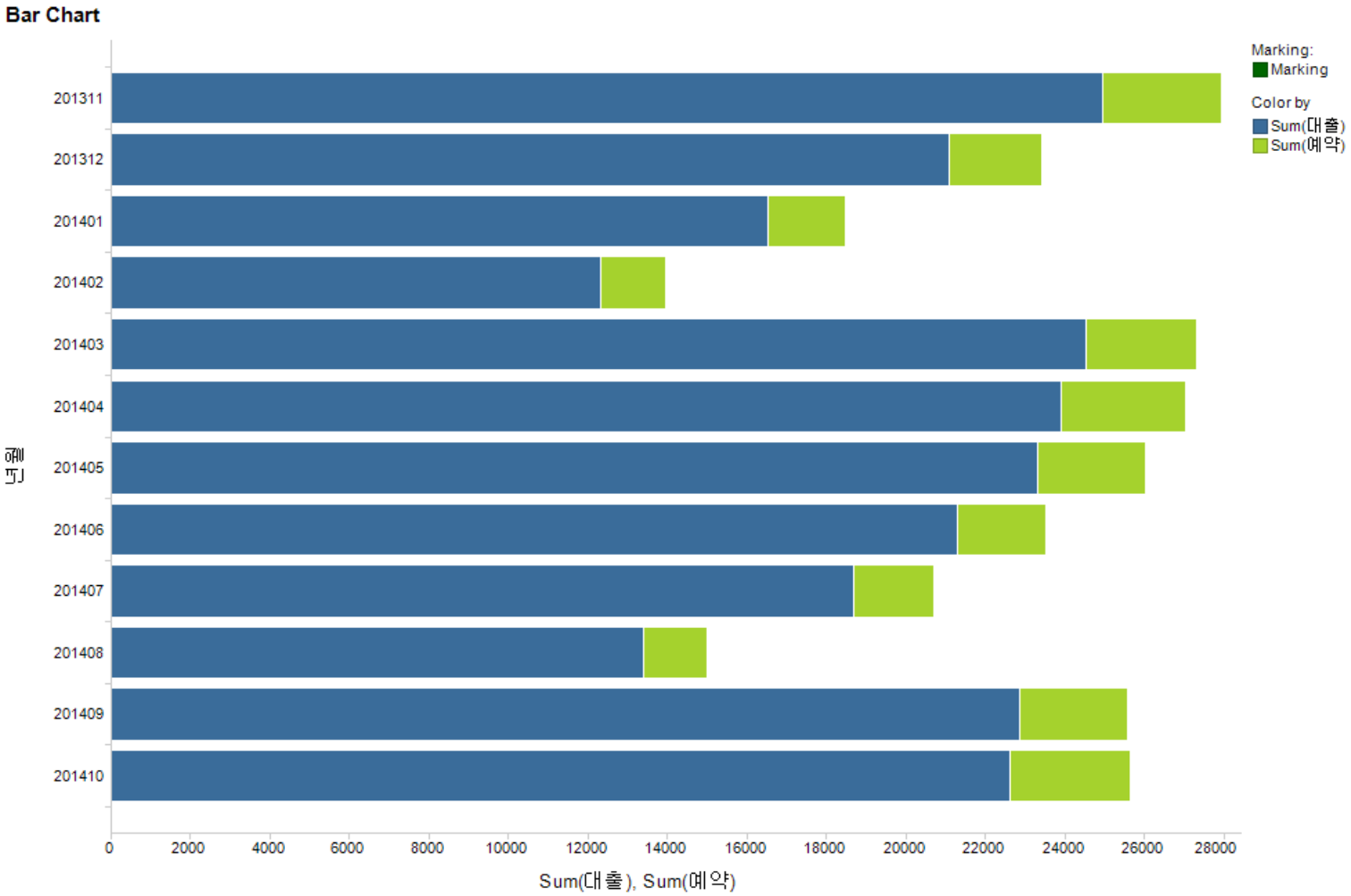
Data Exploration



Data Exploration



Data Exploration



많은 수의 예약이 항상 발생함 → 대여가능도서 수가 부족함

Datamining Problem

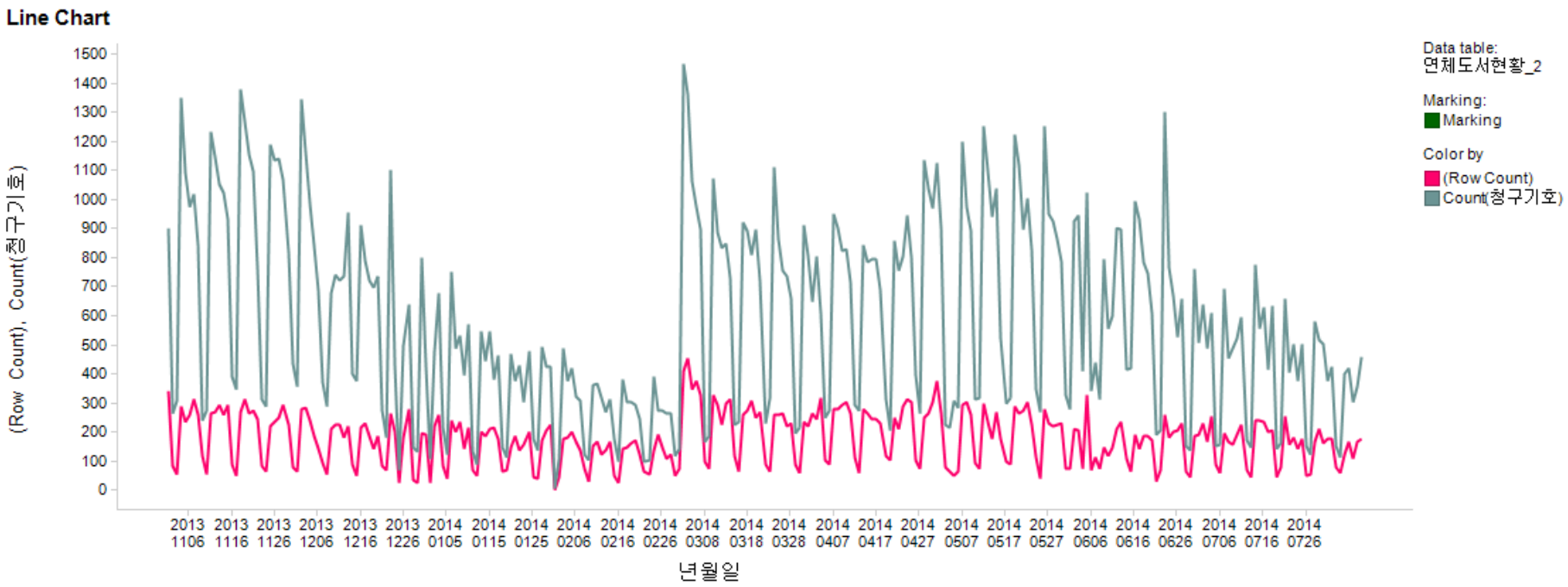
• 많이빌려본 책 목록 1위~20위 [2013-11-15 ~ 2014-07-15]

순위	서명	저자	출판사	발행년도	이용횟수
1	총, 균, 쇠	Diamond, Jared M	문학사상	2007	76
2	아이스퀼로스 비극	Aeschylus	단국대학교출판부	1998	67
3	에우리피데스 비극	Euripides	단국대학교출판부	2004	66
4	아리스토파네스 희극	Aristophanes	단국대학교출판부	2000	59
5	감시와 처벌 2005	Foucault, Michel	나남	2007	57
6	왜 세계의 절반은 굶주리는가?	Ziegler, Jean, 1934-	갈라파고스	2007	55
7	정의란 무엇인가 2010	Sandel, Michael J	김영사	2010	54
8	두근두근 내 인생 2011	김애란, 1980-	창비	2011	53
9	생각의 탄생	Root-Bernstein, Robert Scott	에코의서재	2007	53
10	오이디푸스 왕 ; 안티고네 / 아가멤논 ; 코에포로이 2006	Sophokles	문예	2006	52
11	(젊은이를 위한) 인간관계의 심리학	권석만, 1958-	학지사	2004	52
12	서양사 강의	배영수, 1953-	한울	2000	49
13	불안	De Botton, Alain, 1969-	이레	2005	47
14	소포클레스 비극	Sophocles	단국대학교출판부	2004	47
15	(삼미 슈퍼스타즈의)마지막 팬클럽	박민규	한겨레신문사	2003	46
16	파우스트 v.21	Goethe, Johann Wolfgang von, 1749-1832	민음사	1999	46
17	서양미술사	Gombrich, E. H. (Ernst Hans), 1909-2001	예경	1997	45
18	동양미술사 v.2	한정희, 1952-	미진사	2007	44
19	일반물리학 v.1	Halliday, David, 1916-	범한서적	2009	43
20	이기적 유전자	Dawkins, Richard, 1941-	을유문화사	2007	43

원인 1 : 보유장서에 비해 수요가 매우 높은 도서들이 있음

→도서 적정 권수 구비 필요

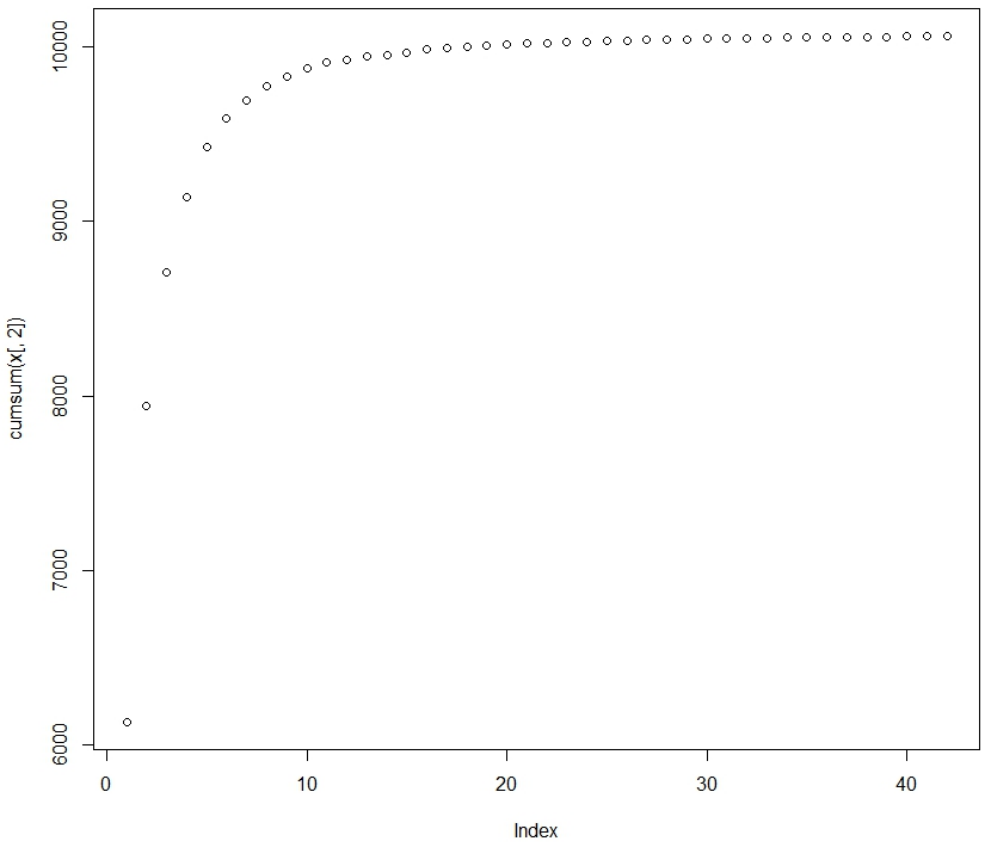
Datamining Problem



원인 2 : 연체가 꾸준히 발생

→연체를 줄이기 위한 정책 필요

Model 1 : 신규도서 구입량 결정



- 예약횟수 3회 이상인 책은 13.5%이지만 예약횟수의 38% 차지

  → 수요가 높은 도서를 파악하여 처음부터 많이 구매해야 함 !

- 예약횟수 = 0인 Dataset(도서관의 전체 장서 목록)이 도착하지 않아 분류모델은 아직 세울 수 없음

- Multiple Regression

  X : 출판연도, 대분류 (혹은 중분류)

  Y : 예약 횟수

  → 결정계수 : 둘 다 0.05

Model 2 : 연체 위험자 분류

Input variable	Estimate	Std. Error	p-value	odds
Constant term	-2.0783088	0.0141352	< 2e-16	*
대분류1	0.2375425	0.0246042	< 2e-16	1.268
대분류2	0.1208054	0.0203095	2.71e-09	1.128
대분류3	-0.0849556	0.0318268	0.0076	0.919
대분류4	0.1739433	0.0168861	< 2e-16	1.190
대분류5	0.1308483	0.0256793	3.48e-07	1.140
대분류6	0.2000273	0.0206871	< 2e-16	1.221
대분류7	0.2446726	0.0200943	< 2e-16	1.277
대분류8	0.3221305	0.0198476	< 2e-16	1.380
대분류9	0.0001896	0.0178535	0.9915	1.000

- ✓ Logistic Regression

  X : 대분류

  Y : 연체가능성

- ✓ 대분류9를 제외하고 유의미

- ✓ Odds를 이용하여 연체가능성 예측

  If 대분류 = 3

  → 연체가능성 높음

  If 대분류 = 8

  → 연체가능성 낮음

- ✓ 분류가 세분화되지 못함

  → 계산량의 한계

Model 2 : 연체 위험자 분류

Input variable	Estimate	Std. Error	p-value	odds
Constant term	-2.055725	0.265518	9.76e-15	*
계절학기	1.220403	0.332154	0.000239	3.388553
교수	-1.621075	0.277931	5.46e-09	0.197686
동문회원(회원제)	0.535999	0.266740	0.044490	1.709155
명예교수	-1.057790	0.332672	0.001474	0.347222
법인직원	-0.876271	0.271142	0.001230	0.416333
보건직	1.544899	0.451479	0.000622	4.687498
부교수	-1.079432	0.280337	0.000118	0.339788
비전임교수	-1.378598	0.285048	1.32e-06	0.251932
석사과정(장애우)	-2.071409	1.042415	0.046908	0.126008
연구생(장애우)	0.840702	0.344276	0.014608	2.317994
예탁금예치자(석박통합과정)	3.665163	1.127165	0.001147	39.0625
조교수	-1.189760	0.286659	3.32e-05	0.304294
초빙교원	-2.579004	0.568267	5.67e-06	0.07585
특별수강생	-1.410011	0.572986	0.013862	0.244141
특별회원(회원제)	-1.246627	0.362613	0.000586	0.287473
협약대학교	-0.825514	0.299502	0.005846	0.43801

- ✓ Logistic Regression
  X : 신분
  Y : 연체가능성
- ✓ 유의미한 신분들만 출력
  (총 56개 중 17개)
- ✓ Odds를 이용하여 연체가능성 예측
  If 신분 = 교수
  → 연체가능성 높음
  If 신분 = 예탁금예치자(석박통합과정)
  → 연체가능성 낮음

Plan

- **Model 1 : 신규도서 구입량 결정**
  - Logistic Regression으로 예약발생여부를 classifying
  - 모델이 효과적인 경우 추가적으로 예약을 고예약/저예약으로까지 분류하는 모델 수립

- **Model 2 : 연체 위험자 분류**
  - 장르와 신분, 출판년도, 학번, 대출월 등을 X로 하는 Logisitic 모델 수립
  - 장르와 신분 변수의 경우 유의미한 항목들만 남기고 나머지를 '기타'로 처리하여 사용할 예정
  - 모델이 효과적인 경우 추가적으로 연체를 장기연체/단기연체로까지 분류하는 모델 수립

Thank You