전문도서관에서 데이터 사이언스를 시작하는 방법

김규환
(인천대학교 문헌정보학과 교수)

1 들어가는 말

주위에서 발생하는 다양한 사회현상을 보다 명확하게 이해할 수 있는 방법은 없을까? 또 왜 그런 사회현상이 발생하는지 객관적인 원인들은 어떻게 찾을 수 있을까? 이런 지적 호기심은 필자로 하여금 데이터 기반 계량적 분석방법, 즉 데이터 사이언스(Data Science)에 관심을 가지도록 해 주었습니다. 다행스럽게도 대학원 시절부터 지금까지 학술통계분석, 이용자 서베이, 데이터 마이닝과 관련된 다수의 강의, 연구프로젝트를 경험하게 되었습니다. 최근에는 대학 기획처 빅데이터센터장이라는 임무를 맡아 교육 빅데이터 기반 학생성공 지원서비스 모델 개발에 참여하기도 했습니다.

그러는 와중에 평소 알고 지내던 전문도서관 사서 한 분께서 이제 전문도서관에서도 데이터 분석을 해야 할 것 같은데 어떻게 시작해야 하는지 잘 모르겠다는 말씀을 해 주셨습니다. 분명한 사실은 도서관사서들을 위한 데이터 강좌가 국립중앙도서관과 한국과학기술정보연구원 등에서 진행되고 있다는 것입니다. 그럼에도 많은 전문도서관 사서들께서 해당 데이터 강좌를 신청하고 직접 가서 수강하기에는 현실적 어려움이 많은 것도 분명한 사실인 것 같습니다. 이처럼 전문도서관에서 데이터 분석을 시작하고 싶지만 어디서 어떻게 시작해야 할지 막막해 하시는 분들에게 그동안 습득한 지식과 경험을 여기서 공유하고자 합니다.

2 빅데이터센터장을 하면서 '데이터 사이언스'에 대해서 깨달은 것들

빅데이터센터장을 하면서 센터 직원만이 아니라 대학 교직원의 데이터분석 역량을 높힐 수 있으면 좋겠다는 생각을 했습니다. 그래서 대학 교직원 대상 데이터 분석 강좌를 기획한 적이 있습니다. 일명 '대학 교육서비스 향상을 위한 빅데이터 이해와 활용 교육 프로그램'으로 동계방학 2 주간 진행

하였습니다. 간략히 강좌내용을 말씀드리면 [빅데이터 개요, 빅데이터 수집저장, 빅데이터 기초분석, 빅데이터 집단분석, 상관분석, 예측분석, 분류분석 머신러닝, 클러스터링, 텍스트 마이닝, 딥러닝기초]입니다(<그림 1>참조).

〈그림 1〉 대학 교육서비스 향상을 위한 빅데이터 이해와 활용 교육 프로그램

교육프로그램

		1교시	2교시	3교시	4교시
1차_1. 07.(월)	빅데이터 개요	교육과정 및 강사소개	4차산업혁명 빅데이터	빅데이터 개요	공공데이터 활용분야
2차_1. 08.(화)	빅데이터 수집저장	빅데이터 자료수집	빅데이터 자료저장	주요기업 기술소개	R. Rstudio 설치실행
3차_1. 09.(수)	빅데이터 기초분석	시각화방법	기술통계 분석	분석기초 이론	교차분석
4차_1. 10.(목)	빅데이터 집단분석	두 집단 간 차이분석 1	두 집단 간 차이분석 2	세 집단 간 차이분석 1	세 집단 간 차이분석 2
5차_1. 11.(금)	상관분석 예측분석	연속형 상관분석	범주형 상관분석	예측분석 기본이론	단순회귀 분석
6차_1. 14.(월)	예측분석	다중회귀 분석 1	다중회귀 분석 2	로지스틱 회귀분석	구조방정식 모형분석
7차_1. 15.(화)	분류분석 머신러닝	판별분석	의사결정 트리 분석	랜덤 포레스트	서포트 벡터머신
8차_1. 16.(수)	클러스터링	계층적 분석	비계층적 분석	연관규칙 분석 1	연관규칙 분석 2
9차_1. 17.(목)	텍스트 마이닝	텍스트 마이닝이해	토픽분석	감성분석	자료효율성 분석
10차_1. 18.(금)	딥러닝기초 교육평가	딥러닝의 이해	딥러닝분석 1	딥러닝분석 2	교육평가

출처 : J대학교 빅데이터센터 홍보 자료(2018년)

전반적인 강좌내용을 보면 알겠지만 요즘 핫한 빅데이터와 관련된 새로운 분석기법들을 R 이라는 분석도구를 활용하여 실습해 보도록 구성하였습니다. 강좌가 끝난 후에 수강생 만족도 및 요구조사를 실시하였고 센터직원들과 강좌평가 세미나를 가졌습니다. 강좌평가 세미나를 통해 다음의 사실을 알게 되었습니다. 첫째, 아무리 최신 분석기법과 분석도구라도 지금 당면한 문제 해결에 직접적으로 도움을 줄 수 없다면 의미가 없다는 것입니다. 둘째, 당면한 문제 해결에는 요즘 이야기되는 대용량 빅데이터가 필요한 것이 아니라 업무단위에서 생산되는 적절한 규모의 데이터만으로도 충분하다는 것입니다. 셋째, R 과 같은 프로그래밍 기반 분석도구는 프로그래밍 언어에 익숙하지 않은 실무자에게는 업무에 바로 활용하는데 심적 부담이 크다는 것입니다.

이에 여러분들에게 다음과 같이 제안드리고자 합니다.

① 데이터 수집과 분석은 명확한 문제정의로부터 시작하는 것이 중요합니다.

실무에서 해결할 문제가 명확하게 정의되어질 경우 어떤 데이터를 수집해야 할지, 수집된 데이터를 가지고 어떤 분석기법과 분석도구를 활용해야 할지, 그리고 분석결과를 어떻게 해석하고 적용할지를 결정할 수 있습니다. 예를 들어, 해결 문제가 '성별에 따른 교육만족도 차이 분석'일 경우, 수집할 데이터는 '성별'과 '교육만족도'에 대한 데이터이며 분석기법은 '독립표본 T 검정'으로 정해집니다. 분석도구의 경우, 엑셀에서 T.Test()함수를 활용하여 데이터 분석결과를 도출할 수 있고 SPSS 에서는 [분석]-[평균비교]-[독립표본 T 검정] 메뉴를 통해 데이터 분석결과를 도출할 수 있습니다. 즉, 자신의 업무영역에서 해결할 문제가 명확하게 정의된다면 수집 데이터와 적용할 분석기법과 분석도구도 명확하게 결정됩니다.

② 대용량 데이터보다 문제해결에 필요한 적절한 규모의 데이터 수집이 더 중요합니다.

실제로 문제영역에 빅데이터가 없거나 확보하기 어려운 경우도 많습니다. 따라서 명확한 문제정의를 토대로 입수가능한 데이터부터 수집하면서 점차 데이터 수집범위를 넓혀가는 것이 필요합니다. 이를 위해서는 평상시에 자신의 업무와 관련된 내부 데이터를 수집해 두고 필요한 경우 외부 데이터(공공데이터 등)를 내부 데이터에 통합해 두는 것이 중요합니다.

③ 데이터 분석도구는 자신에게 익숙한 것부터 시작하여 필요한 경우 확장해 나가면 됩니다.

분명 R 이나 Python 과 같은 객체지향 프로그래밍 언어는 많은 장점을 가지고 있습니다. 대용량 데이터를 처리하는데 매우 유익하며 반복작업에 소요되는 많은 시간을 간단한 프로그래밍으로 줄이는 것이 가능합니다. 특히, R 이나 Python 은 데이터 탐색과 시각화, 통계적 분석, 데이터마이닝, 빅데이터 분석, 텍스트마이닝 등 현재 데이터 분석과 관련된 모든 분석범위를 지원해 줍니다. 그러나, 앞서 말씀 드린 것과 같이 해결해야 할 문제의 유형과 범위에 따라 엑셀이나 SPSS 등으로도 충분히 해결할 수 있는 경우가 많습니다. 예를 들어 속성 데이터간의 관계를 집계하는 것이라면 엑셀에서 피봇데이블(pivot table)을 만들어서 바로 해결할 수 있습니다. 또한 통계분석이나 데이터마이닝 작업이 필요할 경우에도 엑셀에서 제공하는 함수

기능을 활용하여 기초적인 통계분석이 가능하며 엑셀에 애드인(add in) 프로그램을 설치할 경우 데이터 마이닝[1]도 수행해 볼 수 있습니다.

3 도서관 분야 빅데이터 분석서비스 사례 및 개선점

지금은 누구나 인정하듯이 '데이터'가 주인공인 시대입니다. 인터넷 포털, 각종 뉴스 채널에서 데이터의 중요성을 다루는 뉴스기사를 이제는 쉽게 접할 수 있고 실제로 많은 기업과 조직에서는 데이터 기반의 과학적 의사결정 체제로 전환했거나 전환을 준비하고 있습니다.

자랑스럽게도 도서관 분야에서도 빅데이터 서비스 사업이 진행되고 있습니다. 2019년 12월 15일 기준, 948개 공공도서관으로부터 도서관 관련 데이터를 수집, 분석, 활용할 수 있는 시스템을 구축하여 운영하고 있습니다(<그림 2> 참조).

〈그림 2〉 공공도서관 빅데이터 분석 시스템 개념도

출처 : 도서관 정보나루 홈페이지(https://www.data4library.kr/)

본 사업에서 제공하는 서비스는 '사서업무지원서비스 솔로몬'과 '도서관 정보나루'로 구분됩니다. '사서업무지원서비스 솔로몬'은 공공도서관 내부의

1) 엑셀에서 '데이터 마이닝'을 하고자 할 경우 XLMiner(엑셀 확장 데이터 마이닝 툴)를 설치하여 사용할 수 있습니다. 참고사이트 〈https://www.solver.com/xlminer-data-mining〉

장서 데이터, 이용자 데이터, 대출 데이터와 함께 온라인 서점 데이터, 공공데이터 등 공공도서관 외부 데이터를 수집·분석하여 공공도서관 운영을 위한 사서 업무를 지원하는 웹 기반의 의사결정지원 서비스입니다. '도서관 정보나루'는 공공도서관 빅데이터 활용에 관심 있는 연구자, 개발자, 도서관을 위해 다양한 공공도서관 데이터를 공유, 활용할 수 있도록 지원하는 서비스로 수집 데이터 통계, 기간별/지역별/성별/연령별 베스트 대출 장서 등 전국 공공도서관 주요 현황을 제공하고 전국의 서지 데이터, 대출 데이터, 대출도서 순위 등 다양한 유형의 도서관 빅데이터 및 분석 데이터를 제공하고 있습니다. 더 주목할 것은 공공도서관 빅데이터를 활용한 우수서비스 사례들이 등장하고 있다는 것입니다. 대표적으로 '합리적인 장서개발 및 수서정책 수립 사례', '도서관 홍보 기반 마련 사례', '장서이용률 향상을 위한 서비스 개발 사례' 등이 있습니다.[2)]

한편, 매년 공공도서관 빅데이터를 활용한 우수서비스 사례가 증가하고는 있지만 아직은 일부 공공도서관에만 국한되어 있는 것으로 보입니다. 여러 가지 이유가 있을 수 있겠지만 많은 공공도서관의 경우 데이터분석역량을 갖춘 사서 인력이 부족하기 때문이 아닐까 생각합니다. 만약 데이터 분석역량을 갖춘 사서 인력을 채용할 수 있는 여건이 되지 않는다면 기존 사서의 데이터 분석역량을 높힐 수 있는 방안을 고민해 봐야 할 것입니다. 국립중앙도서관과 한국과학기술정보연구원 등에서 도서관 실무자를 위한 데이터 강좌를 매년 열고 있으니 여기에 꾸준히 참여함으로써 기본적인 데이터 분석역량을 확보하는 것도 좋을 것입니다. 직접 오프라인 강좌에 참여할 수 없다면, KOCW(Korea OpenCourseWare)[3)]에서 제공하는 무료 온라인 대학 공개 강좌 중 '데이터 분석' 관련 강좌를 검색하여 수강하는 것도 한 가지 방법이 될 수 있을 것입니다.

4

실무에서 데이터 사이언스를 진행하기 위해 알아야 할 기본 지식들

여기서는 실무에서 데이터 사이언스를 진행하기 위해 알아야 할 기본적 지식들을 다루고자 합니다. 데이터 사이언스를 간단하게 말하면 데이터(Data)를 가지고 사이언스(Science)를 하는 것이라고 할 수 있습니다. 먼저 인간이 세상의 복잡한 현상을 이해하기 위해서 고안한 사이언스, 즉 과학(科學)이 무엇인지에 대한 이해가 필요합니다. 다음으로 과학을 수행하기 위해 활용 되는 데이터가 무엇인지를 이해할 필요가 있습니다.

2) 공공도서관 빅데이터 활용 우수사례에 대한 보다 자세한 내용은 〈https://www.data4library.kr/utilizationCase〉에 접근하여 살펴보는 것을 권해 드립니다. 2014년부터 2018년까지 활용사례집을 다운로드받아 볼 수 있습니다.
3) KOCW(고등교육 교수학습자료 공동활용 체제) 〈http://www.kocw.net/home/index.do〉

① 기본 지식 1 : 기계론적 과학관

지금까지 세상의 복잡한 현상을 이해하는 과학적 방식은 기계론적 과학관이 지배적이라고 할 수 있습니다. 처음에는 자연현상에 적용되었던 기계론적 과학관이 그 범위가 넓어져서 사회현상에 적용되기 시작하였습니다. 데이터 사이언스를 제대로 하기 위해서는 기계론적 과학관에 대한 명확한 이해가 매우 중요합니다. 물론 과학적 방식으로써 기계론적 과학관은 한계점도 가지고 있습니다. 이에 전체론적 과학관[4]이 등장하기도 하지만 여전히 상당부분의 과학적 연구는 기계론적 과학관을 전제로 하고 있습니다.

기계론적 과학관은 자연현상을 각각의 구성요소들로 분해하고 구성요소의 내재된 속성 자체를 분석하거나 속성간의 관계를 분석함으로써 자연현상에 대한 이해에 도달할 수 있다는 입장입니다. 이를 사회현상 분석에 적용해 보면 대부분의 사회통계 분석방법이 기계론적 과학관에 토대를 두고 있습니다. 예를 들어보면, '도서관이용자의 서비스만족도에 대한 통계분석 연구'의 경우 해당 도서관이용자 전체가 관심집단(모집단)이 되고, 개별 도서관이용자가 모집단의 구성요소인 기본 개체가 됩니다. 통계분석에서는 기본 개체를 분석 단위라고 합니다. 그리고 분석 단위인 도서관이용자에게 내재된 성별, 나이, 주거지, 서비스만족도 등이 속성이 됩니다.

정리하면 사회현상을 분석하는 데이터 사이언스는 관심대상인 전체집단(모집단)에서 일어나는 사회적 현상을 분석하기 위해서 집단을 구성하는 기본 개체를 선정하고 각 개체들의 속성들을 설문지 등으로 측정하여 데이터화하고 이들 속성 데이터간의 관계를 분석함으로써 전체집단에서 발생하는 사회현상을 이해하고자 합니다.

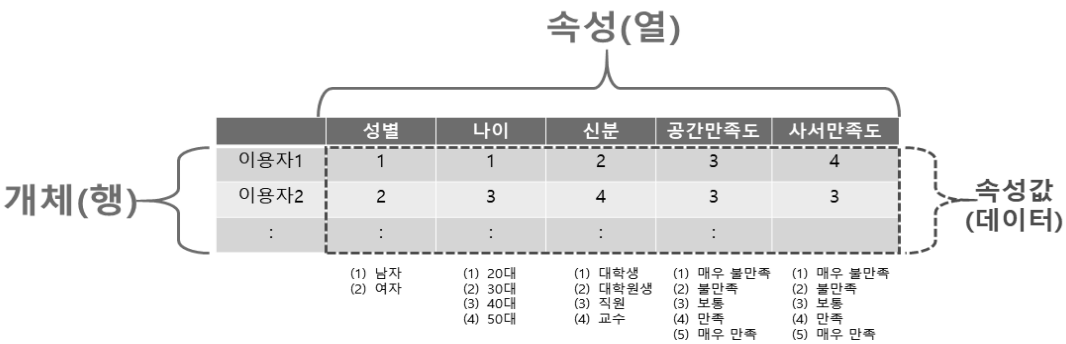
② 기본 지식 2 : 데이터의 이해

데이터 사이언스에서 논의되는 '데이터'는 우리에게 일상적으로 친숙한 [표]의 형태로 구성됩니다(<그림 3> 참조). 데이터의 구성요소를 보면 행에는 기본 개체들이 배열되고 열에는 기본 개체들의 속성들이 배열됩니다. 그리고 행과 열이 만나는 빈칸에는 기본 개체의 속성값이 기재됩니다. 예를 들어 이용자 1(행)의 성별(열)이 만나는 빈칸에는 '남자'라는 속성에 '1'이라는 속성값이 기재되는 식입니다.

---

4) 전체적론 과학관을 전제로 하는 대표적인 분석방법은 네크워크 분석방법이 있습니다. 이에 대한 보다 상세한 내용을 알고자 할 경우에는 '네트워크 분석 방법론(이수상 저, 2012. 논형)'을 참고하시기 바랍니다.

〈그림 3〉 데이터의 구성요소(개체(행), 속성(열), 속성값)



행과 열에 만나는 빈칸에 숫자형태의 속성값이 기재되었다면 분석에 필요한 데이터가 확보되었다고 할 수 있습니다. 이때부터 본격적인 데이터 분석이 가능합니다. 첫째, 개별 속성의 특성을 분석할 수 있습니다. 예를 들어 성별의 빈도와 비율을 구할 수 있습니다. 남자는 몇 명이고 전체에서 몇 퍼센트인지, 반대로 여자는 몇 명, 전체에서 몇 퍼센트인지를 분석할 수 있습니다. 둘째, 속성간의 관계적 특성을 분석할 수 있습니다. 예를 들어 '성별에 따른 공간만족도의 차이', '신분에 따른 사서만족도의 차이' 등을 분석할 수 있습니다.

여기서 기억할 것은 데이터는 개체, 속성, 속성값으로 구성된 [표]형태로 구성되며 [표]의 열에 해당하는 개별 속성 자체나 속성간의 관계가 주요한 분석대상이 된다는 것입니다.

주요한 분석대상인 속성이 취할 수 있는 데이터의 유형은 크게 범주형 데이터와 연속형 데이터로 구분됩니다.

- 범주형 데이터 : 속성에 기재된 숫자가 덧셈, 뺄셈, 곱셈과 같은 사칙연산 등의 일반적 수학적 관계식을 가질 수 없는 데이터를 말합니다. 대표적으로 성별, 학년, 신분 등이 여기에 해당합니다.
- 연속형 데이터 : 속성에 기재된 숫자가 숫자로서 의미를 가지며 사칙연산 등 일반적 수학적 관계식을 가질 수 있는 데이터를 말합니다. 몸무게, 키, 수입 등이 여기에 해당합니다.

데이터분석에서 데이터 유형을 알아두는 것은 매우 중요합니다. 그 이유는 대부분의 데이터분석에서는 데이터 유형에 따라 어떤 분석방법을 적용할 수 있는지가 이미 결정되어 있기 때문입니다.

5

데이터 유형과 분석기법간의 관계

여기서는 전통적인 통계학에서 제시된 데이터 유형과 주요 분석기법간의 관계를 보다 자세히 제시하고자 합니다.[5] <표 1>에서 독립변수는 원인변수를 의미하며 종속변수는 결과변수를 의미합니다. 예를 들어 '성별에 따른 공간만족도의 차이'에서 성별은 독립변수에 해당되며 공간만족도는 종속변수에 해당됩니다.

〈표 1〉 데이터유형에 따른 분석기법

구 분		종속변수	
		범주형 자료	연속형 자료
독립변수	범주형 자료	교차분석	독립표본 T 검증 일원배치분산분석
	연속형 자료	로지스틱 회귀분석	상관관계분석 선형회귀분석

① 범주형 데이터와 범주형 데이터간의 분석

독립변수가 범주형 데이터이고 종속변수도 범주형 데이터인 경우 두 변수간의 관계를 파악하기 위해 '교차분석'이 적용됩니다. 예를 들어 '성별에 따른 선호 도서관공간의 차이' 분석에서 성별과 선호 도서관공간은 모두 범주형 데이터이기 때문에 교차분석을 실시하면 됩니다.

② 범주형 데이터와 연속형 데이터간의 분석

독립변수가 범주형 자료이고 종속변수가 연속형 데이터인 경우 독립변수의 하위집단 수가 2 개이면 독립표본 T 검정, 3 개 이상이면 일원배치 분산분석(ANOVA)이 적용됩니다. 예를 들어 '성별에 따른 공간만족도의 차이' 분석에서 성별의 하위집단 수가 2 개이기 때문에 독립표본 T 검정을 실시하면 됩니다. 한편 '신분에 따른 공간만족도의 차이'분석에서 신분의 하위집단 수가 3 개 이상이기 때문에 일원배치 분산분석(ANOVA)을 실시하면 됩니다.

5) 〈표 1〉에서 제시된 데이터 유형별 분석기법들에 대한 상세한 설명은 지면관계상 자세히 다루지 못한 점을 밝힙니다. 제시된 각각의 분석기법에 대해서는 시중에 판매되고 있는 통계분석 관련 서적이나 동영상 강의를 참고해 주시기 바랍니다.

③ 연속형 데이터와 연속형 데이터간의 분석

독립변수가 연속형 데이터이고 종속변수도 연속형 데이터인 경우, 아래와 같이 상관관계분석과 선형회귀분석 등을 실시하면 됩니다.

- 상관관계분석 예) 학습시간과 성적간의 상관관계
- 선형회귀분석 예) 도서관서비스품질이 이용자만족도에 미치는 영향관계

④ 연속형 데이터와 범주형 데이터간의 분석

독립변수가 연속형 데이터이고 종속변수가 범주형 데이터인 경우 대표적으로 로지스틱 회귀분석이 적용됩니다. 예를 들어, 병원에서 정상인과 암환자를 구분하는 원인 변수가 무엇인지를 파악하기 위해서 이들의 신체관련 속성변수(연령, 키, 몸무게 등) 데이터를 수집할 수 있습니다. 연령, 키, 몸무게 등이 연속형 데이터로 독립변수에 투입되고 정상인과 암환자가 범주형 데이터로 종속변수에 투입됩니다.

여기서는 전통적인 통계학에서 다루는 데이터 유형과 주요 분석기법 간의 관계를 설명하였지만, 기계학습을 중심으로 하는 데이터 마이닝 분야에서도 데이터 유형에 따라 적용할 수 있는 분석기법이 결정되는 것은 마찬가지입니다.

따라서, 실무에서는 문제해결을 위해서 분석할 속성들이 무엇인지, 그리고 각 속성이 어떤 데이터 유형으로 수집될 수 있는지를 파악해야 합니다. 그러고 나서 데이터 유형에 따라 이미 정해진 분석기법을 다양하게 적용해 보면 됩니다.

6 나가면서

지금까지 필자의 지식과 경험을 토대로 데이터 사이언스를 시작하려는 전문도서관 사서분들을 위한 기본적인 지식과 데이터 분석시 고려사항들을 말씀드렸습니다. 그 내용을 요약하여 제시하면 다음과 같습니다.

첫째, 새로운 분석기법이나 분석도구에 집중하기 보다는 자신의 업무에서 분석할 가치가 있는 문제를 정의하는 것이 우선입니다.

둘째, 빅데이터와 같은 대용량 데이터가 확보되어야 데이터 분석이 가능한 것이 아니라 해결할 문제에 필요한 적절한 규모의 데이터를 정기적으로

확보하는 것이 더 중요합니다. 그리고 필요한 경우 데이터 양과 수집 범위를 확대해 나가면 됩니다.

셋째, R 이나 Python 과 같은 프로그래밍 언어가 꼭 필요한 것이 아닙니다. 문제정의와 분석 수준에 적합하면서 자신에게 익숙한 분석도구를 우선적으로 활용하면 됩니다. 아직 익숙하지 않은 프로그래밍 언어를 이해하는데 시간을 보내기 보다는 자신에게 익숙한 분석도구를 활용해서 문제해결을 위한 데이터 분석 자체에 더 집중하는 것이 중요합니다.

넷째, 데이터 분석의 철학적 기반이 되는 기계론적 과학관을 이해해야 합니다. 그리고 데이터가 개체, 속성, 속성값으로 구성되며 대부분의 데이터 분석은 속성간의 관계를 분석한다는 것을 아는 것이 중요합니다. 이를 위해 실무에서 서비스 대상(개체)들의 다양한 속성들을 파악하려는 노력이 필요합니다.

다섯째, 전통적인 통계학뿐만 아니라 데이터 마이닝 분야에서도 속성의 데이터 유형에 따라 분석기법이 정해져 있다는 사실을 아는 것이 중요합니다.

마지막으로 본 내용을 토대로 데이터 분석과 관련된 더 많은 서적과 강의를 접하시기를 권해 드립니다. 다만 필자가 제시한 것들이 전문도서관 사서들께서 실무에서 데이터 분석 업무를 시작하는데 좋은 동기부여가 되었으면 합니다.

참고문헌

김용대, 조광현 (2013). 빅데이터와 통계학. 한국데이터정보과학회지, 24(5), 959-974.

김진영 (2018). (MS 본사 데이터 과학자가 알려주는) 헬로 데이터과학. 한빛미디어.

이수상 (2012). 네트워크 분석 방법론. 논형.

Micheline Kamber, Jiawei Han. (2014). 정사범, 송용근 공역. (2015). 데이터 마이닝 개념과 기법. 에이콘.

그 외 참고한 웹사이트 등은 본문상에 기술하였음