빅데이터 분석과 데이터 시각화 1

글. 김범준 EBS 뉴미디어프로젝트팀 연구원

빅데이터란

빅데이터는 '규모(Volume)', 정형/비정형의 '다양성(Variety)', 수집과 처리에서의 '속도(Velocity)'의 특성을 가진 데이터이며 최근에는 데이터의 품질 여부에 따른 '신뢰성(Veracity)', 잠재적 '가치(Value)' 등 특성의 영역이 확장되고 있다. 넓은 관점에서의 빅데이터는 처리와 분석 기술의 변화, '데이터 사이언티스트' 등 새로운 유형의 인재와 '데이터 중심 조직'까지 포괄하며, 산업 현장에서의 고객 데이터 축적, 학계의 거대 데이터를 다룰 수 있는 기술 아키텍처 및 통계 도구들의 발전, 디지털화, 저장 기술의 발달, 클라우드 컴퓨팅 등 기술의 발달이 빅데이터 등장의 주요 원인으로 꼽힌다. 즉, 갑자기 등장한 개념이 아니라, 기존의 데이터 속성, 처리방식, 다루는 사람과 조직 차원에서 일어나고 있는 변화를 바탕으로 빅데이터 현상이 출현했고, 오늘날 기업, 정부, 개인에게 큰 영향을 미치며 사회 전반에 스며들어 가고 있다.

빅데이터 분석

싱싱한 식재료는 레시피와 요리사의 손을 거쳐야만 맛있는 요리가 된다. 빅데이터가 확보된 후에는 분석을 통해 가치를 창출해야 한다. 빅데이터 분석과 데이터 기반 의사결정의 문화가 효과적으로 기업 내에 정착하기 위해선 이를 체계화한 절차와 방법을 포괄하는 빅데이터 분석 방법론이 필수적이다. 과거 정형 데이터만의 빅데이터 분석 방법론에서 시작하여 오늘날 비정형 데이터의 처리 방법을 포괄하도록 발전한 빅데이터 분석 방법론은 1. 분석 기획, 2. 데이터 준비, 3. 데이터 분석, 4. 시스템 구현, 5. 평가 및 전개의 단계로 구성되어 있다.

1단계 '분석 기획'은 비즈니스 도메인과 문제점을 인식하고 분석 계획 및 프로젝트 수행계획을 수립하는 단계이고, 2단계의 '데이터 준비'는 비즈니스 요구사항과 데이터 분석에 필요한 원천 데이터를 정의하고 준비하는 단계이다. 3단계의 '데이터 분석'에선 원천 데이터를 분석용 데이터 셋으로 편성하고 다양한 분석 기법과 알고리즘을 이용하여 분석을 진행한다. 분석 단계를 수행하는 과정에서 추가적인 데이터 확보가 필요할 경우 데이터 준비 단계로 피드백하여 두 단계를 반복 진행한다. 4단계의 '시스템 구현'은 분석 기획에 맞는 모델을 도출하고 이를 운영 중인 가동 시스템에 적용하거나 시스템 개발을 위한 사전 검증으로 프로토타입 시스템을 구현하는 단계이다. 마지막 5단계 '평가 및 전개'에선 일련의 프로젝트의 성과를 평가하고 정리하거나 모델의 발전 계획을 수립하여 차기 분석 기획으로 전달하고 프로젝트를 종료하게 된다.

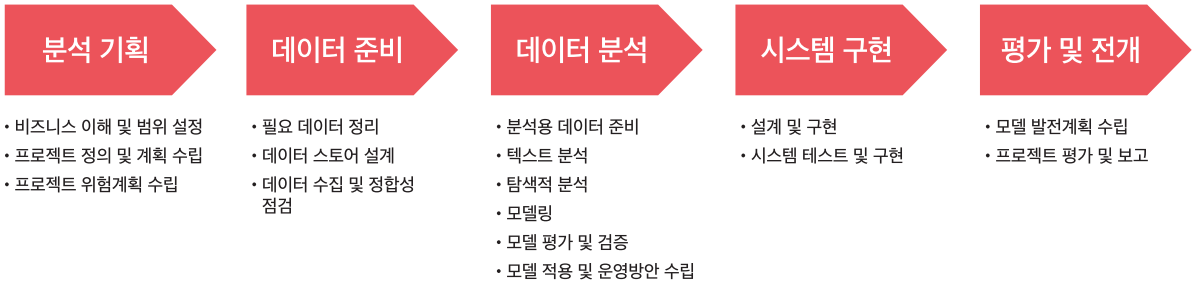


그림 1. 빅데이터 분석 방법론 / 출처 : 한국데이터베이스진흥원. 데이터 분석 전문가 가이드

참고로 정형 데이터에 적용되는 대표적인 분석 방법론은 조직의 성격, 데이터의 속성, 분석의 시작점 등에 따라 여러 가지가 존재한다. 기술과 데이터를 중심으로 인사이트 발굴을 위한 절차와 단계를 정의한 데이터 마이닝 기법인 KDD(Knowledge Discovery in Database), 비즈니스 이해 → 데이터 이해 → 데이터 전처리 → 데이터 분석 → 데이터 평가 → 분석 적용의 6단계를 거치며 공공 분야에서 많이 사용되는 CRISP-DM(Cross Industry Standard Process for Data Mining), 문제의 정의 자체가 어려운 경우 데이터를 기반으로 문제의 재정의 및 해결방안을 탐색하여 개선하는 방식으로서 제조업 공정 분야에서 주로 활용되는 SEMMA(Sampling Exploration Modification Modeling Assessment) 등이 있다.

빅데이터 분석 방법론의 각 절차는 각자의 전문화, 구체화와 함께 서로 연관되어 있지만, 데이터 분석가 고유의 영역이자 흔히 데이터 분석이라 하면 연상되는 통계와 데이터마이닝, 여러 인공지능 알고리즘이 적용되어 핵심 분석이 이루어지는 단계는 3단계의 '데이터 분석' 단계이다. 데이터 분석가는 분석요구사항을 파악한 후 분석 모델링을 진행한다. 그 후 분석 모델을 검증하고 마지막으로 결과해석과 전달을 진행한다. 앞서 언급한 통계와 분석 알고리즘의 역량, 비즈니스 분야에 대한 배경지식과 함께 분석가가 갖추어야 하는 필수요소가 바로 '데이터 시각화'이다.

데이터 시각화란

(빅)데이터 시각화는 광범위하게 분산된 방대한 양의 자료를 분석해 한눈에 볼 수 있도록 도표나 차트 등으로 정리하는 것이다. 시각화를 통해 자료로부터 정보를 습득하는 시간 절감으로 즉각적인 상황 판단이 가능하며, 자료를 습득하는 사람들의 흥미를 유발하고 정보의 빠른 확산을 촉진할 수도 있다. 당연히 자료를 기억하는 데에도 도움이 된다. 데이터 패턴, 주요 변수, 특징 등 원 데이터를 파악하고 탐색하는 탐색적 데이터 분석(Exploratory Data Analysis)과 분석 모델의 설계 및 분석 결과의 표현/전달을 위한 시각적 스토리텔링 등 데이터 분석의 여러 세부 과정에서 데이터 시각화가 중요한 역할을 한다.

데이터 시각화는 데이터를 기반으로 객관적 표현에 초점을 맞추고 또한 데이터가 내포하는 의미를 담은 정보형 메시지를 전달하기 위해 사용된다. 데이터 시각화는 표현이 다차원적이어야 하며 통계적 차원의 시각화 방법과 함께 이에 따른 시각적 요소나 원리의 병행이 중요 요소이다. 즉, 보기 좋은 떡이 먹기 좋다는 속담을 그대로 따른다는 것이다.

'정보의 구조화', '정보의 시각화', '정보 시각표현'의 3단계 과정으로 데이터 시각화는 진행된다. '정보의 구조화'는 데이터 수집 및 탐색, 데이터 분류, 데이터 배열 및 재배열의 단계이고 엑셀, 구글차트, R, 파이썬, Q-GIS, D3JS, Many eyes 등 다양한 데이터 시각화 툴이 사용된다. '정보의 시각화'는 분석을 진행하며 다양한 데이터를 시각화하는 단계이다. 다양한 그래프를 어떤 이유로, 왜 쓰는지, 어떻게 표현하는지를 고려해 그래프를 효율적으로 이용하는 것이 중요하다. 마지막 '정보의 시각표현'에선 만든 그래프를 시각적으로 다듬거나 시각 표현을 극대화하는 방안을 실험하면서 시각적인 완성을 달성한다.

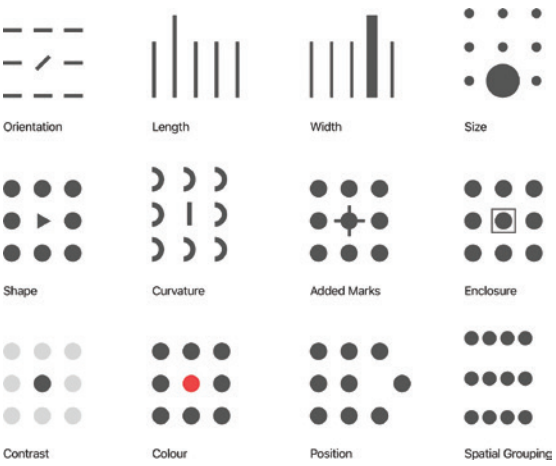


그림 2. 데이터의 전주의적 속성 예시
/ 출처 : TowardsDataScience. Lies, Damn Lies, and Data Visualisation.

데이터 시각화를 잘 활용하기 위해서는 '전주의적(Pre-attentive) 속성'과 '데이터 속성'에 대한 이해가 필요하다. 먼저 '전주의적 속성'은 우리가 의식적으로 노력하지 않아도 한눈에 패턴을 알아채는 시각적 속성이며 Color(색상, 색조), Form(방향, 길이, 너비, 크기, 형태, 곡률, 표시추가, 둘러싸기), Spatial Positioning(공간적 그룹핑), Movement(이동) 등의 요소가 포함된다. '데이터 속성'은 데이터가 연속형, 이산형, 순서형, 명목형 중 어떤 데이터 타입에 속하는지를 파악하는 것과 그래프(차트)가 점, 선, 면의 몇 차원으로 표현되는지에 대한 시각적 요소이며 효과적으로 활용해야 한다.

데이터 시각화 분류

데이터 시각화의 방법에 따라 시간, 분포, 관계, 비교, 공간 시각화로 나뉘는데 분석과 함께 제공되는 시각화 툴에 의해 결정되는 경향이 강하다. 차트와 분석의 내용을 반영하기 위해 어떤 방식으로 써야 하는지 그 쓰임새를 익히고, 적절한 데이터와 정보 시각화를 하기 위한 수단으로 사용해야 한다. 시각화 도구에 한정된 그래프로만 구현하다 보면, 분석적 사고를 효율적으로 보여주기보다는 단지 멋져 보이는 그래프를 선택하기 쉬워 분석 내용을 어떻게 효율적으로 전달할 것인가를 고려해 그래프를 선택하는 것도 중요하다.

시간 시각화

시간 시각화는 시간에 걸쳐 진행되는 변화 또는 트렌드 추적에 사용된다. 시간에 따른 데이터의 변화를 표현하며, 시계열 데이터의 가장 특징적인 요소는 트렌드(경향성)이다. 막대그래프, 누적 막대그래프, 연결된 선 그래프 등이 시간 시각화에 사용될 수 있다.

막대그래프

시간 시각화에서 막대그래프는 분절형 시간에 대해 값들이 뚜렷한 차이를 보이는 경우 사용한다. 막대 값들의 차이가 미미하거나 표시할 값(막대)의 수가 많은 경우, 비교가 쉽지 않아 유의해야 한다.

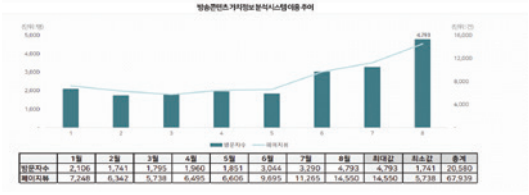


그림 3. 방송콘텐츠 가치정보 분석시스템 이용 추이
/ 출처 : 한국방송광고진흥공사, RACOI 인터넷 이용행태 분석보고서(2020년)

누적 막대그래프

일반적인 막대그래프와 거의 유사하지만 한 구간에 해당하는 막대가 누적인 그래프이다. 한 구간이 몇 개의 세부항목으로 나뉘면서도 전체의 합이 의미가 있을 때 누적 막대그래프를 사용한다. 한 구간의 각 세부사항은 질감 또는 색상으로 구분한다. 세부항목이 너무 많은 경우 세부항목에 대한 의미를 발견하기가 어려워 사용에 유의해야 한다.

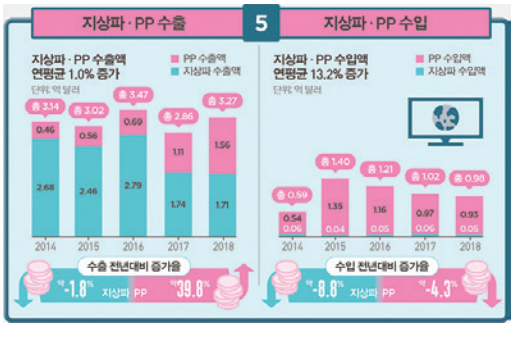


그림 4. 지상파, PP 수출입 누적 막대그래프
/ 출처 : MEDIASTAT, 2019년 방송산업 인포그래픽

연결된 선 그래프

그래프상에 점을 찍고 점 사이를 선으로 이은 그래프이며 변수의 변화, 트렌드, 변화율 정보가 중요한 경우 사용된다.

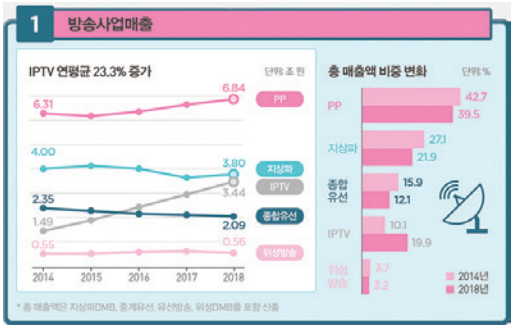


그림 5. 지상파, PP 수출입 누적 막대그래프
/ 출처 : MEDIASTAT, 2019년 방송산업 인포그래픽

분포 시각화

분포 시각화는 비율 시각화라고도 하며 전체의 관점에서 부분 간의 관계, 즉 데이터의 분포 정도를 보여주는 시각화이다. 분포 데이터는 부분을 전부 합치면 1 또는 100%가 되어야 한다. 전체와 부분의 관계뿐 아니라 전체에서 각 부분 요소가 차지하는 비율, 최대/최소 영향 요소를 도출할 수 있다. 파이 그래프, 도넛 차트, 트리맵 등이 분포 시각화에 사용될 수 있다.

파이 그래프

부분과 부분 간의 비율을 알아보는 데 사용하며 분포의 정도를 총합 100%로 나타내서 부분 간의 관계를 보여준다. 육안으로 조각의 면적을 가늠 후 각도를 비교한다. 최대한 구성 요소를 제한하고 내용을 설명하기 위한 텍스트와 퍼센티지를 포함하는 것이 좋다.

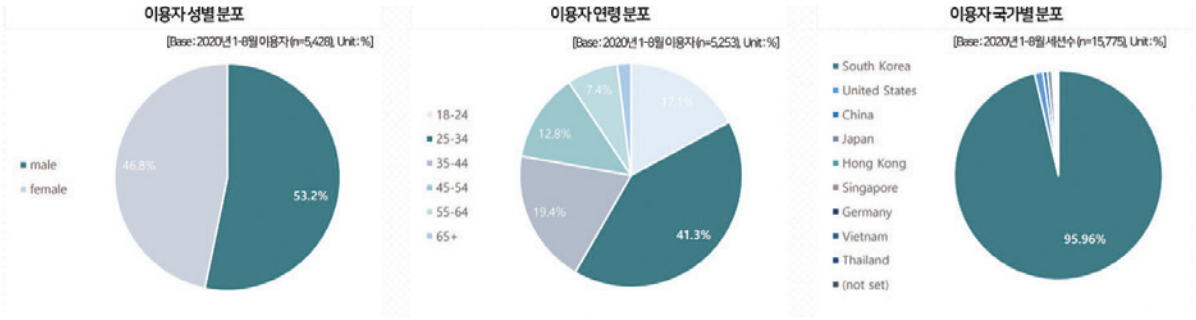


그림 6. 방송콘텐츠 가치정보 분석시스템 이용자별 인구통계학적 분포 / 출처 : 한국방송광고진흥공사, RACOI 인터넷 이용행태 분석보고서 (2020년)

도넛 차트

파이 그래프와 마찬가지로 수치를 각도로 표현하며 중심부를 잘라냈다는 차이가 있다. 중심의 구멍 때문에 조각에 해당하는 수치는 '조각의 면적'이 아니라 '조각의 길이'로 표현된다는 점을 유의해야 한다.

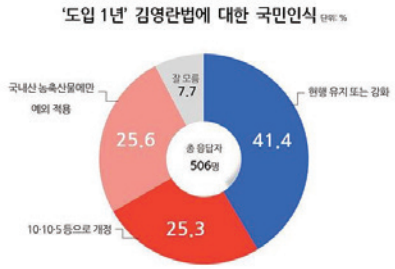


그림 7. '도입 1년' 김영란법에 대한 국민 인식 도넛 차트 / 출처 : 2017.9 리얼미터 여론조사

그림 8. 식품안전나라 인기 검색어 트리맵
/ 출처 : 식품의약품안전처 식품안전나라 시각화 서비스

트리맵 차트

트리맵 차트는 영역 기반의 시각화이다. 각 사각형의 크기가 수치를 나타내며 한 사각형을 포함하고 있는 바깥의 영역은 그 사각형이 포함된 대분류, 내부의 사각형은 내부적인 세부 분류를 의미한다. 위계 구조 기반 데이터나 트리구조 데이터 표현에 유용하며 단순 분류별 분포 시각화에도 사용 가능하다.

관계 시각화

관계 시각화를 통해 데이터의 상관관계를 분석할 수 있다. 변수 간의 상관관계를 이용한 수치의 변화를 모니터링하여 다른 수치의 변화 예측에 용이하다. 그 종류로는 산점도, 버블 차트, 히스토그램 등이 있다.

산점도 그래프

두 데이터 항목의 공통 변이를 나타내는 2차원 도표로서 가로축과 세로축의 변수값에 대응하는 점을 좌표에 배치하여 상관관계를 확인할 수 있다. 점들의 배치가 우상향 추세이면 양의 상관관계가, 우하향 추세이면 하강 추세를 의미하고 점들의 배치에 패턴이 보이지 않는다면 상관관계가 없다고 판단한다. 데이터 분포에 존재하는 패턴의 신속한 식별이 가능하고 데이터 포인터가 많을 때 특히 유용하다. 데이터 포인트의 수가 적은 경우에는 막대그래프나 일반 표가 효과적이다.

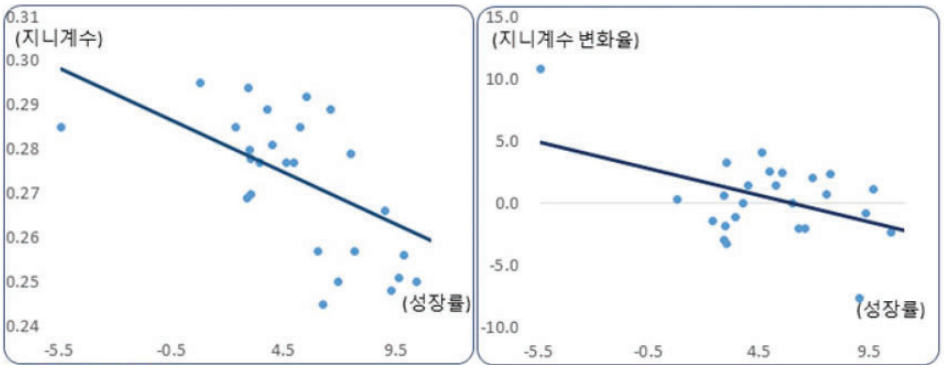


그림 9. 경제성장률과 가처분소득 지니계수 간 관계 / 출처 : 한국경제연구원 경제성장이 소득 불평등에 미치는 영향 분석

버블 차트

버블 차트는 가로축, 세로축, 버블의 크기를 통해 3가지 요소의 상관관계를 표현한다. 수십 또는 수백 개의 값을 갖거나 값들이 몇 자릿수씩 차이가 나는 데이터 셋에 유용하다. 특정 값들을 다양한 크기의 버블로 시각적인 표현을 하고자 할 때도 이 방식이 사용 가능하다.

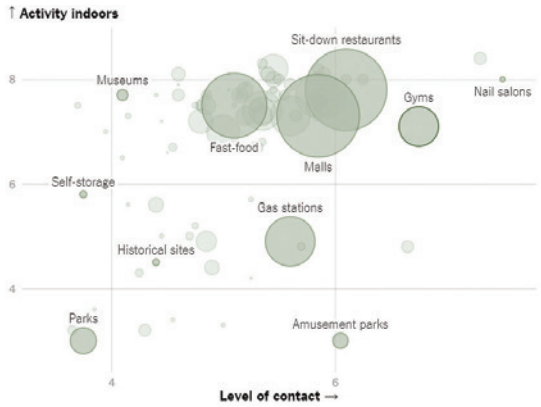
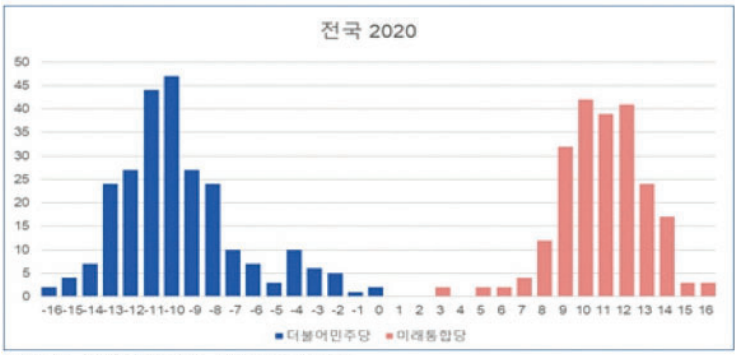


그림 10. 공공장소별 대인접촉과 활동의 관계 버블 차트. 버블의 크기는 비즈니스 분야별 주 평균 방문자 수
/ 출처 : The New York Times. What's Going On in This Graph?



※ 가로축 : (당일선거 득표율 - 사전선거 득표율) %
세로축 : 구간 내 지역구 빈도수

그림 11. 20대 총선결과 지역구별 득표율 차 히스토그램
/ 출처 : 미래한국. [전문가진단] 4·15 총선 결과의 통계적 특이점들

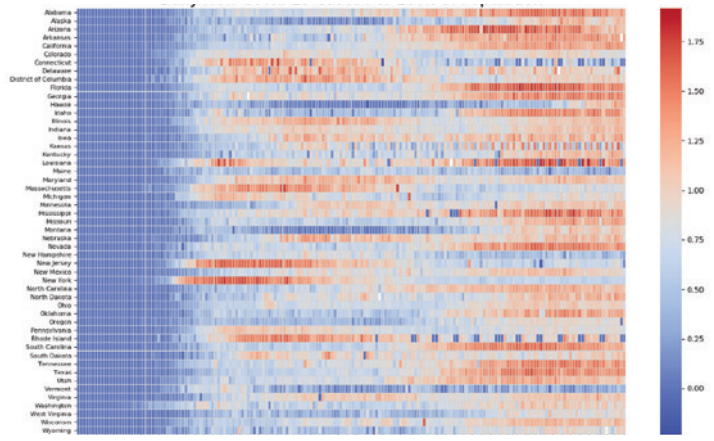


그림 12. Heat map of new COVID-19 cases per 100K of population, by day
/ 출처 : TowardsDataScience. Visualization Of COVID-19 New Cases Over Time In Python.

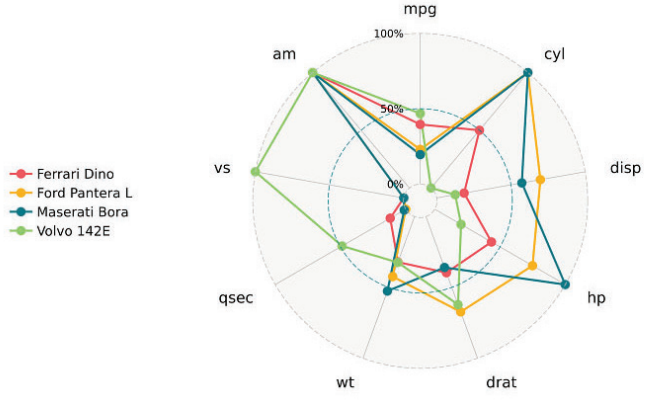


그림 13. R 데이터 시각화 패키지인 ggplot2를 활용한 방사형 차트 예시
/ 출처 : ggplot2 extensions. ggradar

히스토그램

히스토그램은 평균값을 중심으로 양옆으로 점진적인 감소 모양을 보이는 종 곡선이다. 계급을 가로축에 도수를 세로축에 나타낸 뒤, 각 계급의 크기를 가로의 길이로, 도수를 세로의 길이로 하는 직사각형을 차례로 그려서 표현한다. 가로축과 세로축은 연속적이다.

비교 시각화

여러 개의 변수를 다뤄야 할 때 마주하는 첫 번째 난관은 시작점을 찾는 것이다. 빅데이터의 많은 변수와 세부 분류에 압도되면 분석이 어려워진다. 비교 시각화에선 데이터를 전체적으로 한눈에 볼 수 있도록 값을 색으로 나타내는 방법으로 시각화한다. 이를 통해 모든 데이터를 한 번에 훑어본 다음 흥미로운 점을 짚고 다른 점을 찾아가는 방향을 제시해줄 수 있으며 여러 변수의 비교가 가능하다. 히트맵, 방사형 차트, 평행 좌표계 등이 대표적인 비교 시각화 기법이다.

히트맵

시각화 전체를 통틀어서 가장 많이 사용되는 그래프 중 하나이며 한 칸의 색상으로 데이터값을 표현함으로써 여러 가지 변수를 비교할 수 있다. 하나의 대상에 해당하는 한 행/열을 한 방향으로 보면서 모든 변수를 파악한다. 데이터가 지나치게 많을 경우 혼란스러울 수 있으며 적당한 색상 선택, 약간의 정렬 과정이 수반되어야 한다.

방사형 차트

거미줄 차트, 또는 레이더 차트라고도 불리며 중앙에서 외부까지 이어지는 몇 개의 축을 그리고, 전체 공간에서 하나의 변수마다 축 위의 중앙으로부터의 거리를 통해 수치를 나타낸다. 중심점은 축이 나타내는 값의 최솟값을, 가장 먼 끝의 점은 변수의 최댓값을 나타낸다. 각 변수 간의 비율과 균형, 경향을 종합적이고 직관적으로 파악할 수 있다.

평행 좌표계

평행 좌표계는 대상이 많은 데이터에서 집단적인 경향성을 쉽게 알아볼 수 있게 해준다. 여러 축을 평행으로 배치해서 만들며 Y축에서 윗부분은 변수값 범위의 최댓값을, 아래는 변수값 범위의 최솟값을 의미한다. 측정 대상은 변수값에 따라 위아래로 이어지는 연결선으로 그려진다. 데이터 분석의 초기 단계에 많은 변수 중 변수 간의 경향을 찾을 때 유용하다. 디자인을 적용하여 경향성을 시각적으로 설득력 있게 전달할 시에도 활용된다.

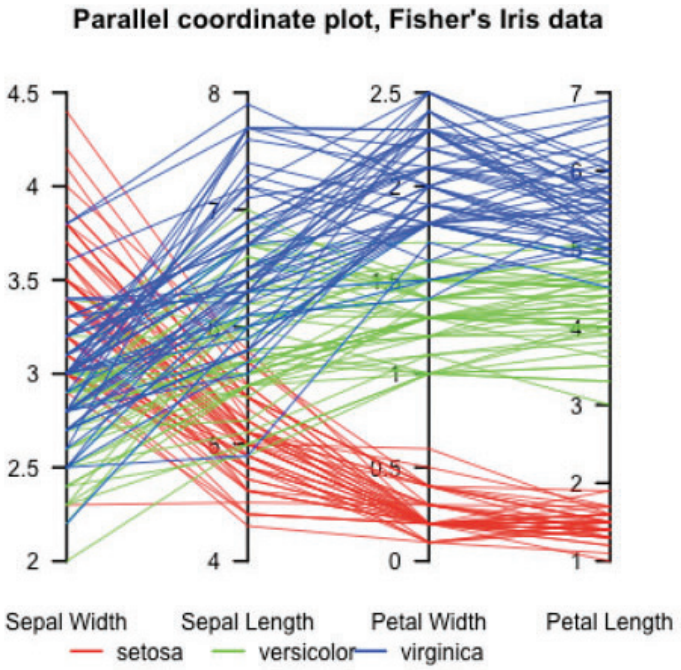


그림 14. Iris 데이터를 활용한 평행 좌표계 / 출처 : 위키피디아. 평행 좌표계

공간 시각화

데이터가 좌표나 주소로 주어진 데이터를 지도상에 매핑시켜 시각화하는 시각화 기법이다. 데이터를 단순히 지도상에 매핑할 뿐 아니라 점들을 격자화하거나 버블 차트의 모양으로 만드는 등 다양한 응용이 가능하다. 데이터별로 설정된 좌표계를 매핑할 지도의 좌표계와 맞추는 것이 중요하다.

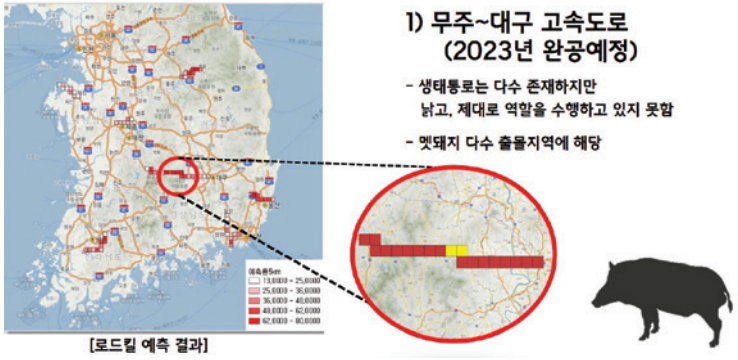


그림 15. Q-GIS를 이용한 고속도로 내 로드킬 발생 예측 시각화 / 출처 : 공공데이터 포털 로드킬 데이터 활용

마무리

지금까지 빅데이터 분석 방법론, 데이터 시각화에 대해 알아보았다. 오늘날 많은 조직이 빅데이터를 활용하여 산업과 기술의 영역을 확장하고 데이터 기반 의사결정을 통해 운영되는 조직을 지향하며 변화하고 있다. 비단 데이터 분석을 업으로 삼는 전문가가 아니더라도 4차 산업혁명 시대를 살아가며 빅데이터를 다루는 전체 흐름을 파악하는 역량을 갖추는 것은 조직의 소통과 업무 효율에 큰 도움이 될 것이다. 또한 데이터 시각화와 그래프, 차트에 대해 이해하는 것은 데이터 인사이트 파악, 논리적이고 직관적인 의사 전달과 스토리텔링의 측면에서 데이터 시각화 결과물의 생산자와 소비자 모두에게, 즉 4차 산업혁명과 빅데이터 시대를 살아가는 현대인들에게 기본적이자 필수로 갖춰야 하는 중요한 역량이 되어가고 있다.