The Korean Journal of Applied Statistics
2024, Vol. 37, No. 1, 87–102

DOI: https://doi.org/10.5351/KJAS.2024.37.1.087
Print ISSN1225-066X / Online ISSN 2383-5818

A study on rethinking EDA in digital transformation era

Seoung-gon Ko[1,a]

[a]Department of Applied Statistics, Gachon University

Abstract

Digital transformation refers to the process by which a company or organization changes or innovates its existing business model or sales activities using digital technology. This requires the use of various digital technologies – cloud computing, IoT, artificial intelligence, etc. – to strengthen competitiveness in the market, improve customer experience, and discover new businesses. In addition, in order to derive knowledge and insight about the market, customers, and production environment, it is necessary to select the right data, preprocess the data to an analyzable state, and establish the right process for systematic analysis suitable for the purpose. The usefulness of such digital data is determined by the importance of pre-processing and the correct application of exploratory data analysis (EDA), which is useful for information and hypothesis exploration and visualization of knowledge and insights. In this paper, we reexamine the philosophy and basic concepts of EDA and discuss key visualization information, information expression methods based on the grammar of graphics, and the ACCENT principle, which is the final visualization review standard, for effective visualization.

Keywords: digital transformation, data analytics, exploratory data analysis, grammar of graphics, ACCENT principle

1. 서론

디지털 전환(digital transformation; DX)이라는 용어는 이제 정치, 경제, 사회 그리고 일상에서조차 흔한 단어가 되었다. 이는 우리 사회의 모든 조직과 개인이 급속한 변화의 중심에 있음을 의미한다. 누구나 디지털(digital)을 원한다. 따라서 DX를 서로 다른 관점 – 기술, 자동화, 네트워킹, 인공 지능 등 –으로 설명하고자 한다 (Ebert와 Duarte, 2018). 하지만 디지털 전환은 기본적으로 데이터(data) 전환을 기반으로 한다 (Forbes, 2017). 데이터는 장치-사물-사람 간의 초연결을 달성할 수 있는 유일한 프로토콜(protocol)이기 때문이다 (Ko, 2023). 빅 데이터의 특징 – 4V's – 을 갖는 디지털 데이터는 전처리(preprocessing)의 적정성 여부가 그 유용성을 결정하며, 적합한 분석 방법들의 체계적 적용이 그 효과성 – 올바른 정보와 지식의 도출 –을 결정한다 (Ko, 2022). 또한 DX 환경에서 고려되는 데이터는 특정 목적에 따라 계획되고 수집되는 통상적인 실험/조사에 기초하는 경우보다는 주로 시장, 고객 또는 생산 환경에서 소급 또는 관찰되는 특징을 갖는다. 따라서 단순 집계를 통한 현상 파악을 목적으로 하는 경우를 제외하고는 좀 더 체계적인 분석 절차가 필요하며, 올바른 정보의 진단과 도출된 지식을 공유하기 위한 다양한 시각화(visualization) 방법의 활용이 요구된다 (Grolemund와 Wickham, 2014).

다양한 영역과 분야에서 다양한 형태의 데이터가 급격히 증가함에 따라 이와 관련된 분석 방법 역시 미시적인 데이터 수집-분석-해석-적용의 구조로부터 실무적 응용력을 확대하는 방향으로 발전되어 왔다 (Provost

[1]Department of Applied Statistics, Gachon University, San 65 Bokjung-Dong, Soojung-Gu, Sungnam 461-701, Korea.
E-mail : sgk@gachon.ac.kr

Published 29 February 2024 / journal homepage: http://kjas.or.kr

Table 1: Terminology related to data analytics

용어	설명
고급 분석론 Advanced Analytics	- 데이터에 대한 심층적인 이해와 실행 가능한 통찰 발견을 가능하게 하는 이론, 기술, 도구 및 프로세스 - 기존의 데이터 분석 및 처리 방식으로는 달성할 수 없는 빅 데이터 환경에서 실행 가능한 통찰을 심층적으로 이해하고 발견할 수 있게 해주는 이론, 기술, 도구 및 프로세스
데이터 분석론 Data Analytics	- 데이터에 대한 심층적인 이해와 실행 가능한 통찰 발견을 가능하게 하는 이론, 기술, 도구, 프로세스 - 기술적(descriptive) 분석론, 예측적(predictive) 분석론, 그리고 처방적(prescriptive) 분석론으로 구성
명시적 분석론 Explicit Analytics	- 보고, 설명, 경고 및 예측을 통한 기술적 분석론 중심의 이론, 기술, 도구 및 프로세스
암시적 분석론 Implicit Analytics	- 심층 분석에 중점을 두며, 예측 모델링, 최적화, 처방적 분석론 및 실행 가능한 지식 전달 중심의 이론, 기술, 도구 및 프로세스
심층 분석론 Deep Analytics	- 어떤 일이 왜, 어떻게 일어났고, 일어나고 있으며, 일어날 것인지에 대한 심층적인 이해를 돕기 위한 이론, 기술, 도구 및 프로세스

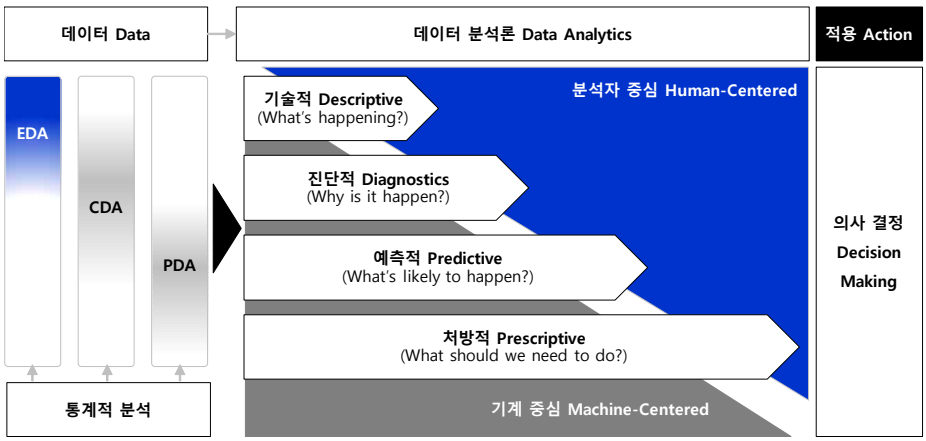

Figure 1: *Key stages of data analytics and useful statistical methods at each stage.*

와 Fawcett, 2013). 1997년 J. Wu 교수는 통계학의 데이터 과학(data science)으로 명명해야 한다고 제안하면서, 통계학이 관심을 가져할 부분이 '데이터 수집, 모델링, 분석, 문제 이해/해결, 의사 결정'으로부터 '대규모/복잡한 데이터, 경험적-물리적 접근, 지식의 표현 및 활용'에 대한 미래 방향으로 전환해야 한다고 주장하였다 (Wu, 1997). 2001년 Brieman 교수는 기존의 해석 중심의 확률적 모델링(stochastic modeling) 이외에 컴퓨터에 기반한 예측 중심의 알고리즘 모델링(algorithmic modeling)에 대한 다양한 사례를 제시하면서 확률적 모델링에 대한 의존도를 낮추고 메커니즘을 알 수 없는 것으로 취급하던 알고리즘 모델링과 같은 다양한 도구를 채택해야 한다고 제안하였다 (Breiman, 2001). 이러한 데이터 수집 환경 변화와 개념적 확장은 여러 분야에서 Table 1과 같이 서로 다른 용어들로 소개되고 있으며, 이에 대한 이론적 근거와 적용 사례에 관한 연구가 통계학습(statistical learning), 예측 모델링(predictive modeling), 기계 학습(machine learning) 분야에서 독창적이지만 공통성을 기반으로 활발하게 진행되고 있다 (Cao, 2017).

Gartner는 데이터 분석론에 기초하여 DX 환경에서의 데이터로부터 현상을 이해하고 실행 가능한 통찰에 대한 발견에 대한 접근 방법을 Figure 1과 같이 제시하였다. 이를 통하여 데이터와 이의 분석은 단순히 적용하는 것이 아니라 현대적인 사업 운영을 주도해야 함을 강조하였고, 분석적 측면에서 (1) 사용 가능한 분석 기능의 범위가 기존의 데이터 보고 및 분석과 비교할 때 매우 광범위해 졌다는 점, (2) 조직에서의 분석 노력

및 예산은 기술(descriptive) 또는 진단(diagnostics) 분석론의 적용 단계에서 가장 많이 소요된다는 점 그리고 (3) 더 많은 비즈니스 역량 강화를 위하여 개별 사업자들의 실용적인 접근 방식의 정립과 함께 이러한 분석 활동이 전체 조직 간의 상호 작용으로 전환되어야 한다는 점을 강조하였다 (Gartner, 2016).

데이터 분석론은 특정 분석 방법을 지칭하는 것이 아니라 데이터에 대한 심층적인 이해와 실행 가능한 통찰 발견을 가능하게 하는 이론, 기술, 도구 또는 프로세스를 의미한다. 따라서 각 단계별 분석론에 적합한 방법은 다양한 분석 이론과 도구의 조합일 수 있다. Figure 1에서 통계적 방법을 EDA (exploratory data analysis), CDA (confirmatory data analysis) 그리고 PDA (predictive data analysis)로 구분하고 이러한 방법들이 어떤 단계별 분석론에서 유용한가를 표현하였다. 예를 들어, 기술적 데이터 분석론에서는 EDA의 통계 방법 그리고 진단적 데이터 분석론에서는 CDA에서 제공하는 통계적 방법이 유용하게 사용된다. 또한 기술적/진단적 분석론에서는 다양한 시행 착오(trian-and-error)에 기초한 지식과 정보 도출의 목적하에서 분석자 중심의 활동으로 구성된다. 그리고 예측적/처방적 분석론은 기술적/진단적 단계를 통하여 추출된 정보와 통찰을 구현하는 단계로 기본적으로 컴퓨터 프로그래밍에 기초한 기계 중심의 작업으로 구성된다.

기술적(descriptive) 분석론에서는 디지털 데이터로부터 전반적인 현상에 대한 요약과 정확하고 정밀한 기준을 도출하는 것으로 숫자와 문자에 대한 요약 그리고 중요 결과에 대한 시각화(visualization)가 강조된다. 진단적(diagnostics) 분석론에서는 발생한 또는 발생이 예상되는 특정 현상에 대하여 그 원인을 파악하고 분리하는 정보 도출을 목적으로 하며, 새로운 가설을 탐색하고 이를 확증하는 작업이 강조된다. 이러한 작업은 진단 팀별 역량에 기초하며, 유능한 데이터 과학자(data scientists)들과의 협업이 필수적이다. 왜냐하면, 실무적으로 실험 데이터가 아닌 경우의 원인 파악은 다양하고 심도 있는 통계적 방법의 적용이 필요하기 때문이다 (Pearl, 2009). 예측(predictive) 분석론에서는 이전 단계에서 확인된 가용 데이터, 전처리 결과와 예상 패턴 그리고 적합한 통계 학습(statistical learning)방법의 적용을 통하여, 관심 사건에 대한 예측 모델을 정의하고 이를 자동화하는 작업으로 구성된다 (Hastie 등, 2009). 이때, 작업들은 전체 조직에 대한 일반화보다는 특정 사업/조직/제품 별로 적용되어야 하며, 지속적인 관리와 보완 활동을 통하여 예측력을 향상할 수 있어야 한다. 마지막으로 처방적(prescriptive) 분석론에서는 특정 사업/조직/제품별로 고객과 시장의 상황을 고려하여 유용한 의사 결정을 추천하는 방법으로 기계 학습(machine learning) 방법과 컴퓨터 프로그래밍에 의한 자동화 작업으로 구성되며, 데이터 관련 작업 이외에 사업 환경 그리고 사회/문화적 요소를 동시에 반영할 수 있어야 한다.

본 논문에서는 1962년 J. Tukey 교수에 의하여 제안한 새로운 데이터 접근 방법이자 DX 현장에서 고객 조사 및 새로운 비즈니스 창출을 위한 데이터 관련 작업에서 중요한 위치를 차지하는 탐색적 데이터 분석 (exploratory data analysis; 이하 EDA)의 철학과 핵심 개념에 대한 재고찰을 통하여 DX 환경에서 수집되는 데이터의 올바른 분석 방향과 시각화를 위한 핵심 정보의 분류, 그래프의 문법(grammar of graphs)에 기초한 다양한 정보 표현 방법 그리고 시각화된 그래프의 검토 기준을 소개한다. 제2장에서는 EDA의 철학과 사상에 대하여 알아보고, 제3장에서는 디지털 데이터의 기술/진단 단계에서 유용한 시각화의 기준에 대하여 소개한다. 마지막으로 제4장에서는 결론 및 추가 연구 방향에 대하여 논의한다.

2. EDA의 철학과 4R

데이터 분석(data analysis)은 다양한 응용 분야와 통계학의 중요한 연결 고리이며, 원시 데이터로부터 새로운 사실, 실행 가능한 통찰 그리고 유용한 지식으로 전환하기 위한 도전적인 활동이다. 이는 「데이터 분석은 본질적으로 경험적 과학」이라는 새로운 관점에서 출발하였으며, 근본적으로 정보 처리 (Morrell, 1968)와 탐색적 데이터 분석(EDA)을 중심으로 실행된다 (Tukey, 1962, 1977). 여기서 EDA는 프린스턴 대학(Princeton University)의 John W. Tukey(이하 JWT) 교수에 의하여 정립된 개념과 방법으로 1962년 “The Future of Data Analysis”에서 다음과 같은 목적으로 제안되었다.

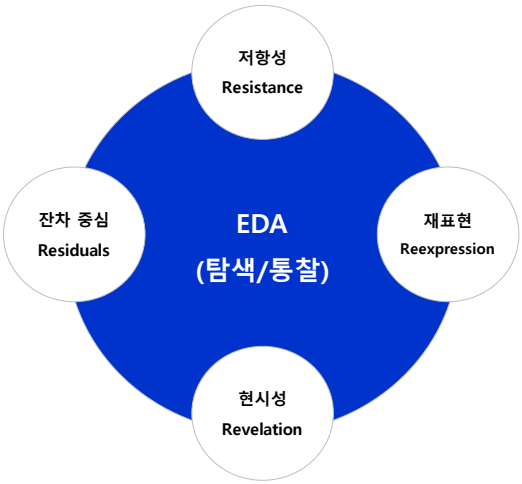

Figure 2: *Key concepts of EDA – 4Rs.*

EDA는 당시 주류 통계 분석 방법이었던 확률 모형 그리고 임의성이 보장된 데이터에 기반한 통계적 추론 중심의 확증적 데이터 분석(confirmatory data analysis; CDA)을 보완하기 위하여 제안되었다. 즉, EDA는 데이터에 내재한 기본 패턴과 구조를 탐색하고 기존의 확증적 데이터 분석(CDA)을 통하여 실증해야 할 정보와 가설을 도출하는 것을 목적으로 적용되며, 간단한 통계 계산과 다양한 도표 표현(graphical displays)을 통하여 데이터에 대한 이해도와 실무 적용도를 높이는 것을 목적으로 한다. 이는 JWT 교수의 중요 저서들과 다양한 논문을 통하여 이론적 근거와 활용 방법이 제시되었고, 다음과 같은 점에서 CDA와 구별된다 (Tukey, 1962, 1977; Mosteller와 Tukey, 1977; Hoaglin 등, 1983, 2011).

– 모집단(population)에 대한 가정을 전제하지 않는다.
– 주어진 데이터로 부터 정보 도출에 집중한다.
– 추론적인 목표를 설정하지 않는다.

또한, 수학과 통계 및 확률보다는 해당 연구에서의 데이터 자체의 분석을 강조하였고, 통계 분석에서도 수학의 특성이 아닌 과학의 특성, 특히 다음의 실행적(practical) 요소를 강조하였다 (Tukey, 1962).

– 데이터 분석의 시작은 적용 범위(scope)와 유용성(usefulness)을 추구해야 한다.
– 데이터 분석의 과정은 불충분한 증거가 더 자주(shall more often) 올바른 답을 제시할 수 있도록 적당히 자주 오류를 범할 용의(willing to err moderately often)가 있어야 한다.
– 데이터 분석의 해석에서 수리적 결과는 입증의 근거가 아닌 판단의 근거로 사용해 야 한다.

이러한 EDA 철학을 구현하기 위하여 JWT 교수는 (1) 결과를 직접 보는것(숫자 또는 그림) (2) 단순하고 직관적인 계산 (3) 적합성과 함께 잔차(residuals)에 대한 심도 있는 고찰 (4) 순차적인 단계에 의한 보다 개선된 이해와 해석 그리고 (5) 공식적이던 비공시적이던 단지 확률(probability)만으로 판단하지 않아야 함을 강조하면서 Figure 2의 4 가지 핵심 개념인 4R – Resistance, Residuals, Reexpression, Revelation – 을 제시하였다 (Hoaglin 등, 1983).

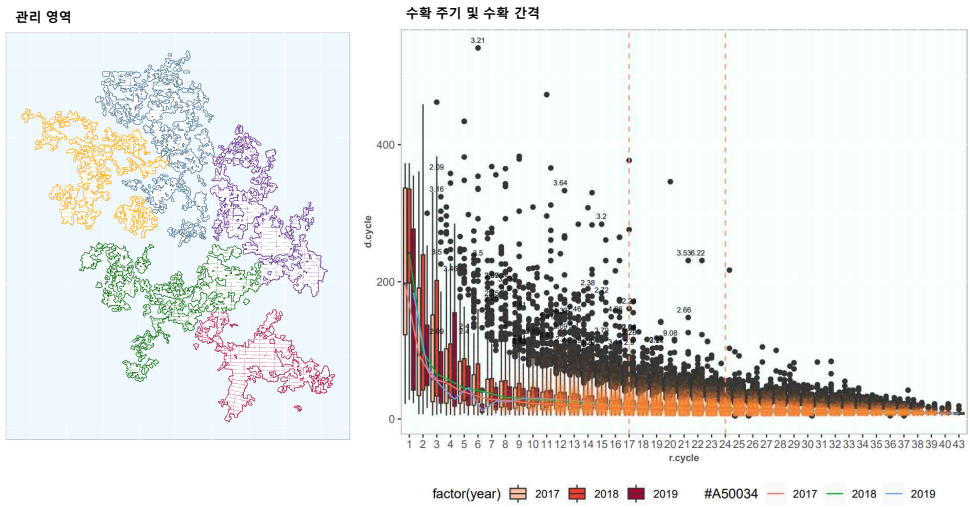

Figure 3: *EDA work for yield prediction: harvest cycle (r.cycle) by region (block) vs. Harvest interval (d.cycle).*

저항성(resistance)이란 분석의 결과가 정확하지 못한 관찰 또는 이상점(outliers)에 영향을 적게 받아야 한다는 개념이다. 실제로 실험적으로 통제된 환경과 임의성이 보장된 표본이 아닌 경우, 일정한 수의 이상점 또는 군집을 포함하게 된다. 이 경우, 기존의 일반적으로 사용되는 통계량 또는 추정량을 사용한다면, 데이터 분석의 결과가 왜곡될 수 있다. 따라서 EDA는 이를 실현하는 방법으로 일반적인 평균(mean)과 표준편차(standard deviation) 대신에 순서 통계량(order statistic)의 순위(rank)에 기반한 중앙값(median) 또는 사분위수(quartiles) 등을 사용하는 것을 추천한다. 이러한 통계량들은 잘못 측정된 값 그리고/또는 이상점들의 포함 여부에 대하여 영향을 덜 받게 되므로 그 분석 결과가 비정상적인 값에 대하여 저항성(resistance)을 가질 뿐만 아니라 데이터를 객관적으로 설명하는데 더 유용할 수 있다는 점을 강조한다. 이러한 개념은 강건 통계(Robust statistics) 방법의 발전에 크게 영향을 주었다 (Huber, 2004).

저항성의 개념은 고전적인 통계 방법들과 함께 디지털 빅 데이터의 분석에 다양하게 활용되고 있다 (Avella와 Ronchetti, 2015; Zoubir 등, 2018). 또한 실무적으로 탐색된 비정상적인 값들에 대한 제거 또는 수정 등의 결정을 할 수 없는 상황에서 일반성을 갖는 결과 도출에 유용할 수 있다. Figure 3는 해외에서 운영 중인 농장에서 약 20,000(Ha.s)의 식재 면적에 대한 수확량 예측 모델 구축을 위한 EDA 작업으로 3개년의 수확 데이터를 년도별 수확 주기(r.cycle)과 수확 간격(d.cycle)를 탐색한 그림이다 (Ko, 2019). 실제로 수확량은 다양한 요소 – 품종, 식재 연령, 수확 시기, 토양, 강수량, 작업 방법, 작업 인력 수급 여부 등 – 의 영향을 받으며, 실제 수확량은 운송 트럭 전체에 대한 무게에서 트럭 무게를 제외하고 측정되므로, 측정 오류는 심하지 않다. EDA의 결과는 거의 모든 관리 단위에서 이상점과 군집이 발생하지만, 이에 대한 원인 규명은 완전하지 않으며, 이상점이라고 할지라도 실무적으로 의미가 없는 데이터가 아니므로 기록 오류를 제외하고는 제거하거나 재측정할 수 없는 경우에 해당한다.

잔차(residuals)란 특정 모형(model)하에서 적합(fitting)된 부분을 제외한 나머지를 의미한다.

$$\text{데이터} = \text{적합된 값} + \text{잔차} = \text{주 요약} + \text{잔차},$$

여기서 적합된 값들이란 데이터의 주 경향성을 파악하는 값으로 간단하고 유용한 정보를 제공한다는 장점을

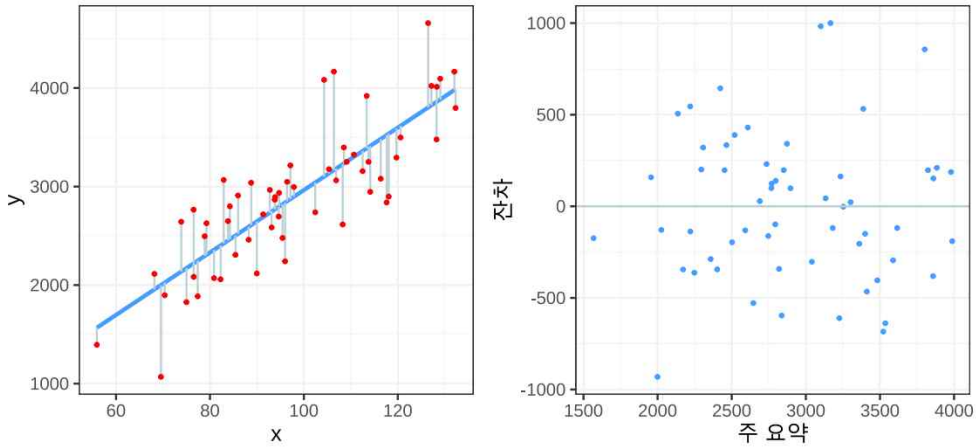

Figure 4: *Fits and residuals between variables with simple linear relationships.*

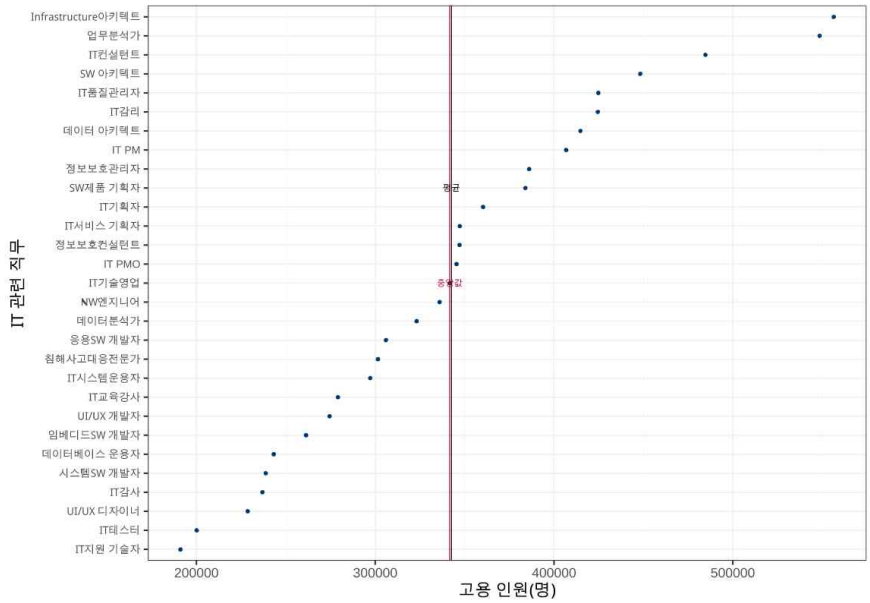

Figure 5: *Number of employees by job in the domestic IT industry in 2021 <Data source: Ministry of Employment and Labor <https://www.moel.go.kr >.*

갖지만, Figure 4에서와 같이 이러한 정보는 데이터의 부분적인 요약이라는 단점을 갖는다. 특성 대상에 대한 연구 또는 일정 패턴이 요구되는 공학적 실험에서의 관심 사항은 주 요약이며 잔차는 일반적으로 잡음(noise)

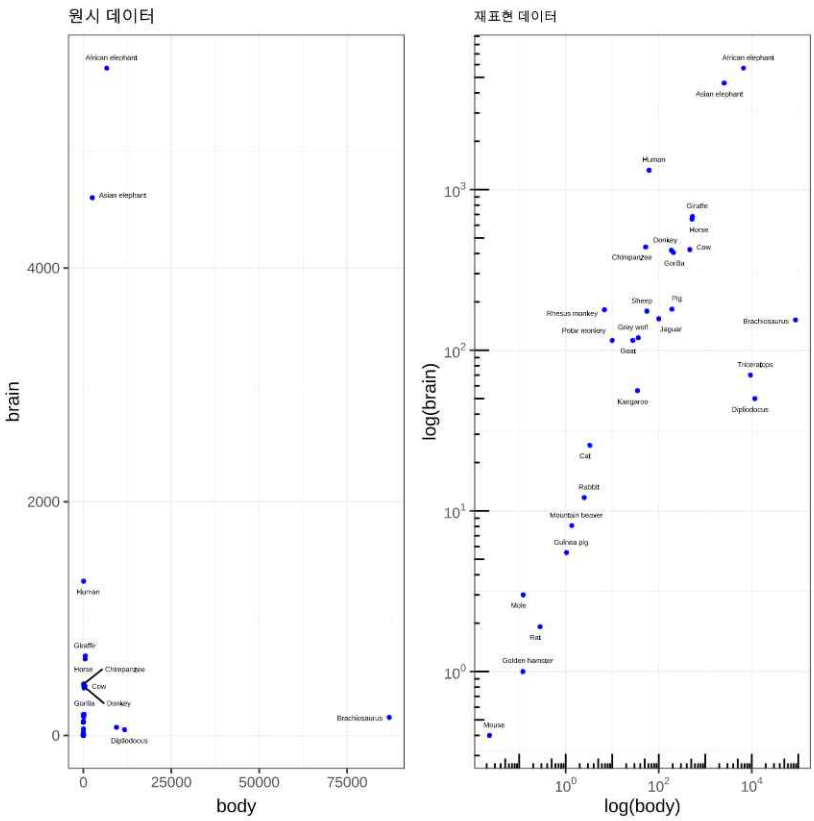

Figure 6: *Body weight and brain weight of 28 land-dwelling animals: Re-expression.*

으로 가정한다. 하지만 시장 또는 고객 관련 데이터에서는 주 요약 보다 개별 또는 집단별 잔차 정보가 더 의미 있는 경우가 많다. 주 요약에서 설명되지 못한 대부분의 중요한 근거는「잔차 = 데이터 – 주요약」에 포함되기 때문이다. 잔차에 대한 심도 있는 고찰은 의도하지 않았던 새로운 사실을 확인할 수 있는 근거를 제공한다. 예를 들어, Figure 5는 2021년 국내 IT 업계의 직무별 고용 인원(단위:명)을 표현한 것이다 (Ministry of Employment and Labor, 2022). 직무 별 고용 인원에 대한 통계적 모형은 다음과 같다.

$$y_i = \mu + \epsilon_i, \quad i = 1, 2, \ldots, 29,$$

여기서 전체 29개 직무에 대한 주 요약은 전체 평균 또는 중앙값으로 나타낼 수 있다. 이는 IT 업계의 직무별 특성을 설명하고 다른 업계와 비교하기 위한 중요한 정보일 수 있지만, 실제 직무에 관심이 있는 구직 준비자들에게는 이러한 주 요약 보다는 잔차로 표현되는 직무 별 고용 인원에 더 관심이 있을 수 있다. 만일 잔차의 크기를 심도 있게 고찰한다면 직무의 선택에 대한 새로운 사실을 도출할 수 있을 것이다.

재표현(re-expression)이란 더 쉽고 명확한 분석과 해석을 위하여 원 데이터(raw data)에 대한 적절한 변환을 의미한다. 원 데이터에 사용된 단위(unit) 또는 척도(scale)는 실무적 측정 용이성에 기초한다. 하지만 이러한 단위/척도가 데이터를 요약하거나 분석할 때도 적합하다는 보장은 없다. 실제로 많은 경우에 분석과 해석을 위하여 원 데이터를 로그(logs) 또는 제곱근(square roots)을 이용하여 새로운 척도로 재표현하여

Table 2: Purpose, goals and design principles of visualization (Ko, 2023, pp 266)

목적	목표	설계 원칙
분석 Analysis	탐색 Exploration	탐지 Detection
	수색 Reconnaissance	인지 Perception
	진단 Diagnosis	비교 Comparison
소개 Presentation	예증 To simulate	미학 Aesthetics
	설득 To persuade	수사 Rhetoric
	알림 To inform	설득 Exposition

Table 3: Information Guideline for data visualization: 10-Key-Info (Ko, 2023, pp 267)

정보	설명
크기 Magnitude	빈도(frequency, counts) 또는 양(size)의 요약 및 비교
순서 Order	크기의 오름(ascending) 또는 내림(descending)의 순위(rank) 또는 그 차이
부분/비율 Proportion	전체 중에서 특정 항목들의 집계 수 또는 비중(fraction)
편차 Deviation	기준값(reference value(s))과의 차이 : 양(+) 또는 음(−)
분포 Distribution	전체 데이터의 구조(structure), 형태(shape), 위치(location), 산포(scale) 등
흐름 Flow	기준값의 변화에 따른 데이터 시퀀스(sequence)의 패턴(patterns)
변화 Change	시간의 변화에 따른 추세(trends)의 형태 또는 변화량
상관 Correlation	2개의 양적 변수 간의 상관성(correlations) 유무와 정도
연관 Association	개체 간의 관계(network) 또는 밀접도(associations)
공간 Space	공간적(spatial) 위치 또는 지리학적(geographical)으로 표현되는 값, 빈도, 또는 변화(changes) 패턴과 크기

사용한다. 이러한 재표현은 (1) 분포의 대칭성 (2) 분산의 안정성 (3) 관계의 선형성 그리고 (4) 관련 변수의 가법성을 확보하여 쉬운 분석과 유용한 해석을 도울 수 있다. Figure 6는 28종의 육상 동물의 체중(Kg) 두뇌 무게(g)를 조사하여 도시한 것이다 (Venables과 Ripley, 1999).

이때, 원 데이터에서는 일정한 관계를 파악할 수 없지만, 로그 함수를 이용하여 재표현된 $\log_{10}(\text{brain}_i) = \beta_0 + \beta_1 \log_{10}(\text{body}) + \epsilon_i, i = 1, 2, \ldots, 28$의 경우에는 일부 관찰값을 제외하고 선형 관계가 밀접함을 확인할 수 있다 (Hur와 Moon, 2010).

현시성(revelation)이란 정보의 진단 또는 획득된 정보의 시각화(visualization)를 강조하는 개념으로 JWT 교수는 숫자와 문자 그리고 빈도 등에 대하여 데이터를 쉽게 해석하기 위한 다양한 아이디어 – Tally Diagram, Box-and-whisker plots, Back-to-Back STEM-and-Leaf 등 –를 제공하였다 (Tukey, 1977). 또한, 패턴을 찾거나 적합도를 표시하기 위하여「 (심지어 정신적으로도) 잊고 있었던 질문을 떠올리게 하는 데 그림보다 더 좋은 것은 없습니다. — J. W. Tukey (1985) 」라고 주장하면서, 그래프는 숫자만으로는 불가능한 다양한 정보를 표현하고 전달할 수 있음을 강조하였다. 실제로 그래프는 새로운 지식과 통찰을 탐구하기 위하여 데이터에 내재한 정보를 효율적으로 검토할 수 있는 유용한 도구임과 동시에 데이터 분석론을 통하여 도출된 결과에 대한 의사소통을 도울 수 있는 유일한 도구임은 경험적으로 증명되었다.

이러한 4R의 핵심 개념은 다양한 데이터 분석 방법의 발전에 기여하였고 (Camacho, 2014; Firouzi 등, 2018), 특히, 현시성의 개념은 컴퓨터의 발전과 함께 데이터 시각화(visualization)를 위한 SNAP/IEDA, EXPACK, Data Desk, SYSSTAT 등의 다양한 동적(dynamic) 그리고 대화형(interactive) 그래픽 소프트웨어의 성장에 지대한 영향을 주었다 (Friendly, 2008). 다시 말해, 새로운 그래픽 소프트웨어의 개발로 인하여 EDA가 중요해진 것이 아니라, EDA에서 제시된 개념과 방법들을 쉽고 다양하게 적용할 수 있도록 데이터 시각화 소프트웨어가 발전하고 있다는 것이다.

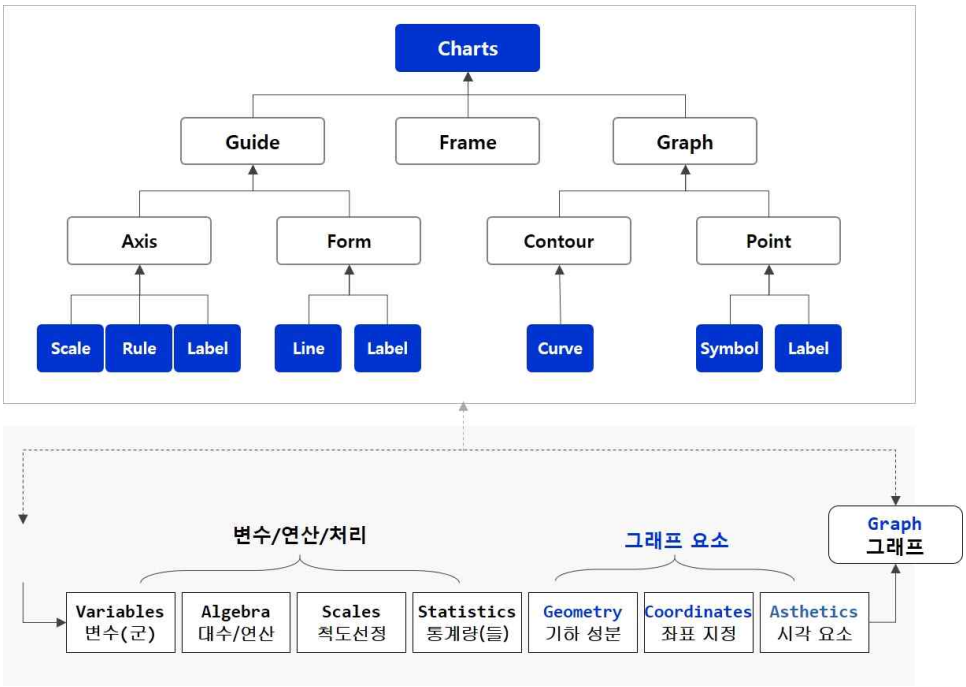

Figure 7: *Wilkinson's graph grammar (Wilkinson, 1999).*

3. 시각화를 위한 고려 사항

EDA의 핵심 개념 중의 하나인 현시성(revelation)은 효과적인 데이터 탐색을 위하여 그래픽 요약(graphical summary)을 활용하여 의도된 메시지의 생생함, 불가피성 그리고 숫자들만으로 표현할 수 없는 정보들을 표현할 수 있어야 한다는 점을 강조한다 (Tukey, 1977). 다양한 영역에서 데이터의 측정 능력의 향상과 함께 정보 교류의 필요성이 증대하면서 Table 2과 같이 다양한 목적하에서 시각화를 강조하고 있다 (Friendly, 2015).

일반적인 시각화의 경우에는 소개(presentation)를 목적으로 하며, 특정 사건 또는 현상의 예증, 설득 그리고 알림을 목표로 실행된다. 이때, 목표에 적절하도록 설계 원칙을 정의하고, 적절한 구도와 색상 등의 미학 그리고 전체 흐름을 무리 없이 보여 줄수 있는 수사 그리고 장단점과 함께 유용한 점 등을 빈틈없이 구성하는 설득의 설계 원칙을 준수해야 한다.

데이터의 분석 목적으로 시각화를 실행하는 경우에는 탐색, 수색 그리고 진단을 목표로 한다. 특히, EDA의 경우에는「분석-탐색-탐지」를 기준으로 적용되며, 그 유용성 여부는 (1) 데이터 소비자의 요구 사항 (2) 적합한 데이터의 확보 (3) 분석자의 데이터 이해도와 시각화 방법에 대한 역량에 따라 결정된다. 따라서, 올바른 시각화를 위해서는 먼저 데이터 소비자의 요구 사항 또는 분석자가 공유하고자 하는 정보를 명확히 정의할 수 있어야 하며, 이러한 정보를 표현할 수 있는 적합한 시점과 범위의 데이터를 확보한 후, 적절한 시각화 방법을 선택하고 적용할 수 있어야 한다. Table 3는 단일 변수 기준으로 시각화 가능 중요 정보(10-Key-Info)를 정의한 것이다 (Ko, 2023). 이를 기준으로 데이터 소비자의 요구 사항과 분석자가 공유하고자 하는 정보를 구분하여 적용할 수 있다. 만일 요구 사항이 2가지 이상의 개별 정보를 포함한다면, 이를 시각화 결과에 효과적으로

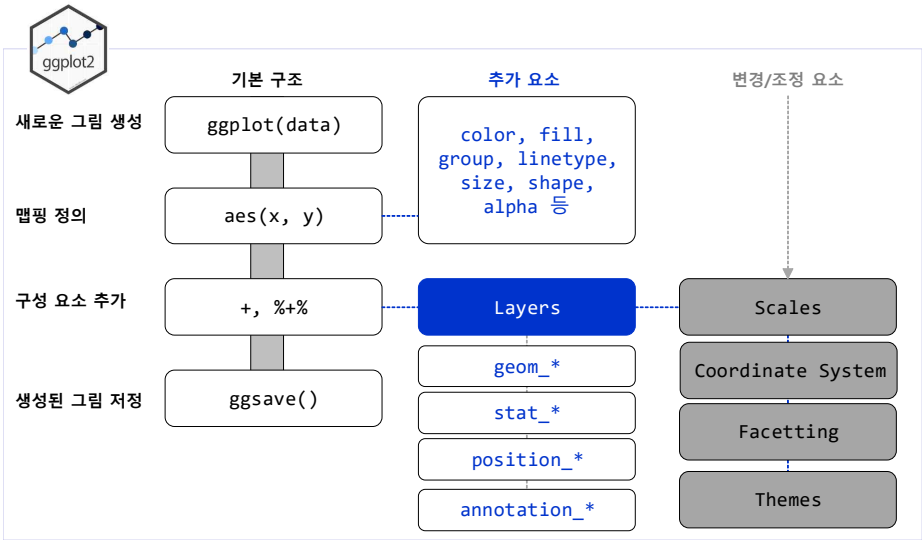

Figure 8: *R package ggplot2 based on the grammar of graphics.*

반영할 수 있어야 한다. 최근의 시각화 소프트웨어들은 이러한 요구 사항을 쉽게 반영할 수 있도록 발전하고 있다.

실무적으로 시각화 작업이 정보의 종류와 조합에 따라서 다양한 방법을 시도하고, 평가하고, 선택하는 시행착오(trial-and-error) 적인 방법으로 적용된다는 점을 상기할 때, 적절한 시각화 소프트웨어를 선택하는 것은 성공적인 시각화를 위한 중요한 요소라고 할 수 있다 (Ali 등, 2016).

1980년대에 소개된 개인용 PC와 이에 적용된 새로운 그래픽 방식은 다양한 그래프들을 더욱 쉽고 많이 그릴 수 있게 되었다. 널리 알려진 SPSS 및 SAS와 같은 통계 소프트웨어는 고전적인 통계 분석 –기술 통계, ANOVA, 회귀 분석 등 – 을 쉽게 실행할 수 있어서 일찍부터 널리 알려져 왔지만, EDA의 다양한 시각화의 구현은 컴퓨팅 성능과 소프트웨어 인프라가 발전함에 따라 뒤늦게 발전하였다 (Friendly, 2008). 그 중에서 현대 계산 그래픽 방법을 위한 Figure 7의 그래픽 문법(grammar of graphics) 아이디어는 Bertin에서 시작되어 Wilkinson에 의하여 체계적으로 정리되었다 (Bertin, 1983; Wilkinson, 1999). 그래프 문법은 그래프의 구성 요소를 명시적으로 분류하고, 데이터로부터 변수의 선택, 필요 연산, 척도 선정과 통계량 선택 등의 기본 단계와 그래프의 기하 성분, 좌표 지정, 시각 요소 등으로 구분하여 순차적으로 적용할 수 있는 구조를 정의함으로써, 그래프 작성을 위한 다양한 요구 사항을 거의 제한 없이 반영하는 방법을 제시하였다 (Ko, 2023).

또한, 1980년대 Bell Labs에서 데이터 분석을 위한 언어 S가 개방형 소프트웨어로 소개되고 (Chambers 등, 1983; Becker와 Chambers 1984), 1995년 S를 기초로 하는 R 언어가 소개됨에 따라 (Ihaka와 Gentleman, 1996), R에서 그래프 문법을 명시적으로 구현한 ggplot2 패키지가 개발되면서 (Wickham 2009), 오늘날 거의 모든 데이터 시각화 소프트웨어에 영향을 미치고 있다 (Friendly, 2021). Figure 8는 ggplot2의 기본 구조를 보여 준다.

ggplot2 패키지는 그래프 맵핑(mapping)을 정의한 다음에 원하는 구성 요소를 제한 없이 추가할 수 있으며, 최종 생성된 그래프는 원하는 형식으로 저장할 수 있다. 또한 기본 구조에 다양한 심미적 요소와 레이어

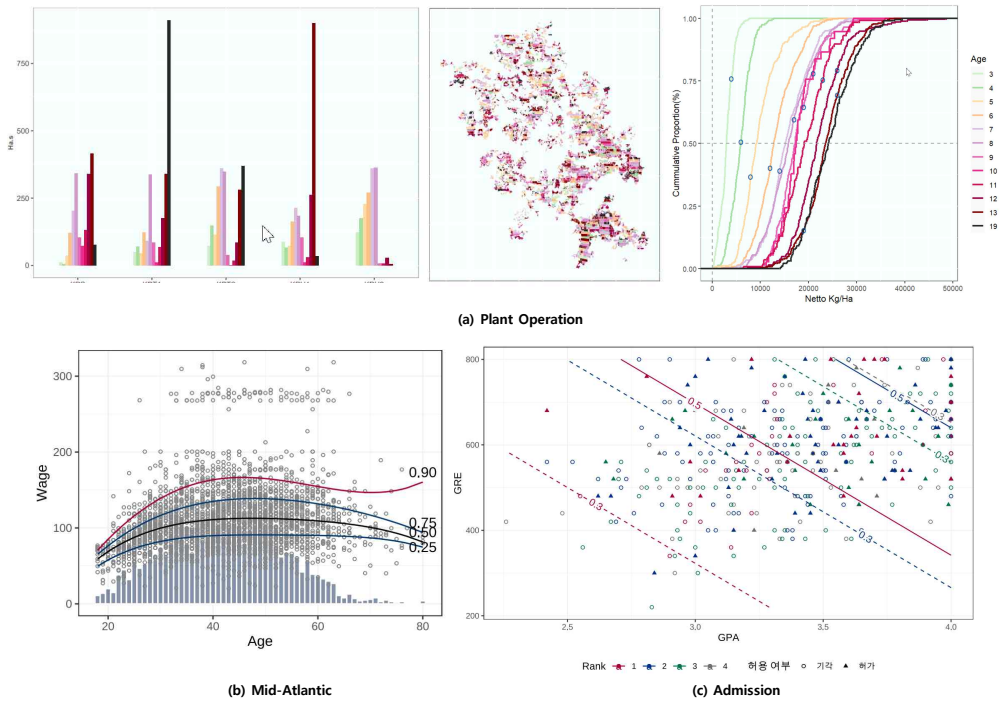

Figure 9: *Representation of multiple information in a visualization (example).*

(layer)를 추가할 수 있으며, 다양한 방법으로 이를 변경 또는 조정하는 방법을 제공한다 (Wickham, 2011). 이러한 그래프 생성 구조는 데이터 소비자의 다양한 요구 사항을 쉽게 반영할 수 있고, 시각화의 기본 기능인 진단뿐만 아니라 원 데이터와 함께 다양한 분석 결과를 반영할 수 있는 유용한 방법을 제공한다.

Figure 9은 ggplot2 패키지를 이용하여 다중 정보를 시각화한 예제이다. Figure 9(a)는 농장 운영 대시보드를 작성하기 위하여 EDA의 결과를 동적 그래프(dynamic graphs)로 전환하는 과정에서 생성된 그래프의 일부이며 (Ko, 2019), (b)와 (c)는 관찰된 정보와 추정된 정보를 함께 표시한 시각화 예제이다.

Figure 9(a)는 데이터 소비자의 요구에 따라 탐색된 정보를 농장의 관리 기준별로 식재 나무의 연령(age)에 따른 면적(Ha)의 "크기"를 표현하고, "공간" 정보를 이용한 식재 연령의 "분포"를 표현한 것이다. 그리고 연령별 연간 수확량(Netto)을 Ha 당 Kg으로 변환하여 이의 누적 "비율"을 표시한 것이다. S-형태가 늦게 시작하고 완만할수록 수확량이 많고 지속적으로 수확량을 확보할 수 있다는 것을 나타내며, 해당 국가의 농무부에서 제공하는 지역의 표준 생산량을 원(o)으로 추가하여 목표 설정과 관리 기준과 비교하도록 하였다. 만일 0.5% 미만이라면 해당 연령의 생산량이 부족한 것으로 판단하고, 그렇지 않다면, 해당 연령의 생산량이 초과한다는 의미로 해석할 수 있다. 이러한 정보를 제공해야 하는 이유는 식재 나무의 연령 별 표준 생산량을 초과한다는 것은 과잉 생산을 의미하며, 이를 통하여 수확 주기를 조정하는 기준을 변경할 수 있기 때문이다. 이러한 EDA의 결과는 R의 동적 패키지를 이용하여 선택(select)−확인(view)의 구조로 제공될 수 있다.

Figure 9(b)는 미국 Mid-Atlantic의 연령별 평균 급여와 함께 다양한 정보를 포함한 데이터 (Miller, 2011)를 이용하여, age와 wage의 "관계"는 산점도로 표현되었고, age의 "빈도"는 히스토그램 그리고 다항 분위수 회귀분석을 이용하여 "추정"된 사분위 수(Quartile)는 개별 실선으로 표현되었다. (c)의 경우는 미국의 특정 대학

원에 응시한 학생들을 GPA (grade point average)와 GRE (greaduate record examination) 점수의 관계 그리고 4가지로 구분된 졸업 대학교 등급(Tier)에 따라 합격 여부를 표시한 것이다. GRE와 GPA의 "관계"는 산점도로 표현되었고, 이때 삼각형(▲)은 합격을 의미하고 원(o)은 불합격을 의미한다. 여기에 로지스틱(logistic) 회귀분석를 이용하여 "추정"된 예측 합격 경계(0.3, 0.5) 등고선을 Tier1과 Tier2에 대하여 함께 도시한 것이다.

단순 정보가 아닌 다중 정보를 반영하는 시각화에서 너무 많은 정보를 포함하는 경우, 정보 전달이 효과적이지 않을 수 있다. 따라서 시각화된 그래프는 의도된 메시지의 생생함, 불가피성 그리고 숫자들만으로 표현할 수 없는 정보들을 표현할 수 있어야 한다는 기준 하에서 검토되어야 하고, 필요한 경우 수정되어야 한다.

시각화 결과의 검토 기준은 먼저 데이터 소비자의 요구 사항을 올바로 반영했는가이어야 하며, 데이터 분석을 통하여 얻어진 정보와 통찰을 간단하고 명확하게 전달할 수 있어야 한다 (Boukhelifa와 Duke, 2009; Hepworth, 2017). Table 4은 이러한 활동에 도움이 되는 검토 기준으로 시각화의 과정에서 적용할 수 있다 (Burn, 1993).

시각화는 복잡한 데이터를 쉽게 이해하고 탐색된 정보를 명확하게 공유하기 위한 도구이다. DX 환경에서 수집되는 다양한 데이터들 역시 의사 결정을 위한 정보 공유와 향상된 협업을 위하여 다양한 시각화 방법을 적용하고 있다 (Ali 등, 2016). 이의 효과성을 높이기 위하여 데이터 소비자 또는 목적에 적합한 정보를 구별하고, 다양한 형태를 표현할 수 있는 소프트웨어를 선택하고 시도하여 정보를 표현할 수 있어야 한다. 또한 최종 시각화된 그래프는「ACCENT 원칙」을 기준으로 검토하고 수정하여 간결하면서도 명확한 그래프를 제시할 수 있어야 한다. 또한 데이터의 출처와 수집 및 시각화는 복잡한 데이터를 쉽게 이해하고 탐색된 정보를 명확하게 공유하기 위한 도구이다. DX 환경에서 수집되는 다양한 데이터들 역시 의사 결정을 위한 정보 공유와 향상된 협업을 위하여 다양한 시각화 방법을 적용하고 있다 (Ali 등, 2016). 이의 효과성을 높이기 위하여 데이터 소비자 또는 목적에 적합한 정보를 구별하고, 다양한 형태를 표현할 수 있는 소프트웨어를 선택하고 시도하여 정보를 표현할 수 있어야 한다. 또한 최종 시각화된 그래프는「ACCENT 원칙」을 기준으로 검토하고 수정하여 간결하면서도 명확한 그래프를 제시할 수 있어야 한다. 또한 데이터의 출처와 수집 및 처리 방법을 명확히 제시하여 데이터 시각화에 의한 과잉 해석을 피할 수 있어야 한다 (Glazer, 2011).

4. 결론

DX은 기업이나 조직이 디지털 기술을 활용하여 비즈니스 프로세스, 문화, 고객 경험 등을 혁신하는 과정이다. 이러한 과정에서 시장과 고객의 이해, 시장 트랜드 및 소비자 행동에 기반한 의사 결정 그리고 비즈니스 혁신 모델 발굴 등에서 데이터는 중요한 역할을 한다. 본 논문에서는 Gartner가 제시한 데이터 분석론(data analytics)의 4가지 단계를 소개하고, 이 중에서 대부분의 노력과 시간이 소요되는 기술(descriptive)과 진단(diagnostics) 단계에 유용한 EDA에 대하여 재고찰하였다. EDA는 일반적인 가설 검정 또는 통계 모델을 결정하기 이전에 데이터의 구조, 패턴, 이상점/군집 여부 등을 파악하는 초기 단계에 적용되는 방법으로 일종의 "탐정 활동(detective works)"이라고 할 수 있다 (Tukey, 1977). 다음 명제는 DX 환경에서 수집되는 데이터에 왜 EDA가 필요한지를 잘 설명해 준다.

> Far better an approximate answer to the right question, which is often vague,
>
> than an exact answer to the wrong question which can always be made precise.
>
> — J. W. Tukey (1962)

EDA는 4가지 핵심 개념(4R's) – 저항성(Resistance), 잔차(Residuals) 중심, 재표현(Reexpression), 현시성(Revelation) – 을 중심으로 데이터의 검토와 진단에 필요한 다양한 아이디어를 제시하였다. 이러한 방법들은 수작업(pencil-and-paper)을 목적으로 제안되었지만, 그 개념과 방법은 DX 환경에서의 데이터의 탐색에서도

Table 4: ACCENT principle in visualization

항목	설명
A Apprehension, 인지	고려한 변수들을 올바로 반영할 방법 선택 – 그래프가 고려한 변수들의 관계를 이해하는데 최적화되어 있는가?
C Clarity, 명확	그래프의 모든 요소를 시각적으로 구분 – 가장 중요한 요소나 관계가 시각적으로 가장 잘 나타나는가?
C Consistency, 일치	이전 그래프와의 유사성을 기반으로 그래프를 작성 – 그래프의 요소, 기호, 모양 그리고 색상이 이전에 사용했던 그래프와 일치하는가?
E Efficiency, 효율	복잡할 수 있는 관계를 가능한 한 단순하게 묘사 – 그래프의 요소가 경제적으로 사용되었는가 그리고 그래프는 쉽게 해석할 수 있는가?
N Necessity, 필요	꼭 필요한 그래프 그리고 그래프의 요소들의 선택 – 그래프가 표 또는 요약값 등의 대안보다 데이터를 설명하는데 더 유용한가? – 전달하고자 하는 관계를 표현하는데 현재의 모든 그래프 요소 가 필요한가?
T Truthfulness, 진실	암묵적 또는 명시적 척도들을 올바로 반영하는 그래프 요소들의 척도 결정 – 그래프의 요소들이 정확하게 배치되고 올바른 척도로 조정 되었는가?

여전히 유용할 뿐만 아니라 데이터로부터 올바른 정보와 실행 가능한 통찰 확보라는 근본적인 목적을 보다 효율적으로 달성하기 위한 필요조건이라고 할 수 있다. 실무적으로 DX 환경에서 수집되는 데이터는 분석을 위하여 정제되지 않았으며, 이상점 그리고 정확하지 못한 관찰값을 포함하고 있을 뿐만 아니라, 명확한 가설하에서 설계되고 수집되지 않는 것이 일반적이기 때문이다. 이러한 DX 환경에서 다양한 시각화 소프트웨어의 발전은 데이터의 탐색과 분석의 효율성과 효과성 향상이라는 측면에서 모두 긍정적이다. 하지만 데이터 소비자의 요구 사항이 다양하고 복잡한 만큼, 시각화에서 유용한 정보를 명확히 구분할 수 있어야 하고 이를 올바로 반영할 수 있는 적합한 소프트웨어의 선택 역시 개인 또는 조직의 중요한 역량에 해당한다. 이를 위하여 (1) 시각화를 위한 10-Key-Info (2)그래프 문법(grammar of graphics)에 기초한 R의 ggplot2 패키지의 유용성과 확장성 그리고 다중 정보 표현 예제 (3) 최종 시각화 그래프의 검토 방법인 ACCENT 원칙을 제시하였다.

References

Ali SM, Gupta N, Nayak GK, and Lenka RK (2016). Big data visualization: Tools and challenges. In *Proceedings of 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, Greater Noida, India, 656–660.

Avella Medina M and Ronchetti E (2015). Robust statistics: A selective overview and new directions, *Wiley Interdisciplinary Reviews: Computational Statistics*, **7**, 372–393.

Becker RA and Chambers JM (1984). *S: An Interactive Environment for Data Analysis and Graphics*, Chapman and Hall/CRC, London.

Boukhelifa N and Duke DJ (2009). Uncertainty visualization: Why might it fail?. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems* (pp. 4051-4056), Boston, United States.

Breiman L (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author), *Statistical Science*, **16**, 199–231.

Bertin J (1983). *Semiology of Graphics*, University of Wisconsin Press, Madison, WI.

Burn DA (1993). Designing effective statistical graphs. Chapter 22 in Handbook of Statistics, Vol. 9 [C.R. Rao ed.], Available from: https://www.sciencedirect.com/science/article/abs/pii/S0169716105801464

Camacho J (2014). Visualizing big data with compressed score plots: Approach and research challenges, *Chemometrics and Intelligent Laboratory Systems*, **135**, 110–125.

Cao L (2017). Data science: A comprehensive overview, *ACM Computing Surveys (CSUR)*, **50**, 1–42.

Chambers JM, Cleveland WS, Kleiner B, and Tukey PA (1983). *Graphical Methods for Data Analysis*, Wadsworth, Belmont CA.

Ebert C and Duarte CHC (2018). Digital transformation, *IEEE Software*, **35**, 16–21.

Firouzi F, Farahani B, Ibrahim M, and Chakrabarty K (2018). Keynote paper: From EDA to IoT eHealth: promises, challenges, and solutions, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, **37**, 2965–2978.

Fobes (2017). Available from: <https://www.forbes.com/sites/peterbendorsamuel/>

Friendly M (2008). A brief history of data visualization. In *Handbook of Data Visualization* (pp. 15–56), Springer, Berlin, Heidelberg.

Friendly M and Meyer D (2015). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data (Vol. 120)*, CRC Press.

Friendly M (2021). *A History of Data Visualization and Graphic Communication*, Harvard University Press, Cambridge, MA.

Gartner (2016). 2017 Planning Guide for Data and Analytics, Available from: https://www.gartner.com/en/documents/3471553

Glazer N (2011). Challenges with graph interpretation: A review of the literature, *Studies in Science Education*, **47**, 183–210.

Grolemund G and Wickham H (2014). A cognitive interpretation of data analysis, *International Statistical Review*, **82**, 184–204.

Hastie T, Tibshirani R, Friedman JH, and Friedman JH (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Hepworth K (2017). Big data visualization: Promises & pitfalls, *Communication Design Quarterly Review*, **4**, 7–19.

Hoaglin DC, Mosteller F, and Tukey JW (1983). *Understanding Robust and Exploratory Data Anlysis. Wiley Series in Probability and Mathematical Statistics*, John Wiley & Sons, New York.

Hoaglin DC, Mosteller F, and Tukey JW (Eds) (2011). *Exploring Data Tables, Trends, and Shapes*, John Wiley & Sons, New York.

Huber PJ (2004). *Robust Statistics*, John Wiley & Sons, New York.

Hur M and Moon S (2010). *Exploratory Data Analysis*, FREEACADEMY, Seoul.

Ihaka R and Gentleman R (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Ko S (2019). Data-driven Analytics for early achievement of DX – Focusing on Palm business, unpublished seminar, LG Internationals

Ko S (2022). A study on guidance of data preprocessing for process analysis in digital transformation era, *Journal of Creativity and Innovation*, **15**, 247–278.

Ko S (2023). *Exploratory Data Analysis using R and Statistical Methods*, KYOWOO, Seoul.

Miller S (2011). Inquidia Consulting(formerly Open BI). From the March 2011 Supplement to Current Population Survey data, Available from: https://www.re3data.org/repository/r3d100011860

Ministry of Employment and Labor (2022). Number of employees by job in the domestic IT industry in 2021,

Available from: https://www.moel.go.kr

Morrell AJH (Ed) (1968). Information Processing 68: Proceedings of IFIP Congress 1968, Edinburgh, 5-10 August 1968, Amsterdam: North-Holland Publishing Company.

Mosteller F and Tukey JW (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley Series in Behavioral Science: Quantitative Methods. Reading, MA: Addison-Wesley.

Pearl J (2009). Causal inference in statistics: An overview, *Statistical Surveys*, **3**, 96–146.

Provost F and Fawcett T (2013). Data science and its relationship to big data and data-driven decision making, *Big Data*, **1**, 51–59.

Rousseeuw PJ and Leroy AM (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

Tufte ER (2001). *The Visual Display of Quantitative Information (Vol. 2, p. 9)*, Graphics Press, Cheshire, CT.

Tukey JW (1962). The future of data analysis, *The Annals of Mathematical Statistics*, **33**, 1–67.

Tukey JW (1977). *Exploratory Data Analysis*, Reading, Massachusetts, USA.

Venables WN and Ripley BD (1999). *Modern Applied Statistics with S-PLUS* (3rd ed), Springer, New York.

Tukey and Tukey (1985). Computer graphics and exploratory data analysis: An introduction. In *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics85*, Fairfax, VA: National Computer Graphics Association, III:773–85.

Wickham H and Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis*, Springer, New York.

Wickham H (2010). A layered grammar of graphics, *Journal of Computational and Graphical Statistics*, **19**, 3–28.

Wilkinson L (1999). *The Grammar of Graphics*, Springer, New York.

Wu J (1997). Statistics = data science, Available from: http://www2.isye. gatech. edu/ jeffwu/presentations/data-science. pdf

Zoubir AM, Koivunen V, Ollila E, and Muma M (2018). *Robust Statistics for Signal Processing*, Cambridge University Press, New York.

Received November 2, 2023; Revised November 24, 2023; Accepted November 27, 2023

DX 전환 환경에서 EDA에 대한 재고찰

고승곤[1,a]

[a]가천대학교 응용통계학과

요 약

디지털 전환(digital transformation)이란 기업이나 조직이 기존의 비즈니스 모델이나 영업 활동을 디지털 기술을 활용하여 변화시키거나 새롭게 혁신하는 과정을 말한다. 이는 시장에서의 경쟁력 강화, 고객 경험 개선 그리고 새로운 사업의 발굴 등을 위하여 다양한 디지털 기술들 – 클라우드 컴퓨팅, IoT, 인공 지능 등 – 의 활용이 요구된다. 또한 시장, 고객 그리고 생산 환경에 대한 지식과 통찰을 도출할 수 있도록 올바른 데이터의 선택, 분석 가능한 상태로의 데이터 전처리(preprocessing) 그리고 목적에 적합한 체계적인 분석들에 대한 올바른 프로세스 정립을 필요로 한다. 이러한 디지털 빅 데이터의 유용성은 적합한 전처리와 함께 정보 및 가설 탐색 그리고 지식과 통찰의 시각화를 위한 탐색적 데이터 분석(exploratory data analysis; EDA)의 올바른 적용이 결정한다. 본 논문에서는 EDA의 철학과 기본 개념에 대하여 재고찰과 함께 효과적인 시각화를 위하여 시각화 핵심 정보, 그래프 문법(grammar of graphics)에 기초한 정보 표현 방법 그리고 최종 시각화 검토 기준인 ACCENT 원칙을 논의한다.

주요용어: 디지털 전환, 데이터 분석론, 탐색적 데이터 분석, 그래프 문법, ACCENT 원칙

[1](461-701) 경기도 성남시 수정구 복정동 산 65번지, 가천대학교 응용통계학과. E-mail: sgk@gachon.ac.kr