I. 데이터 시각화의 기초

최근 우리는 많은 데이터 사이에서 살고 있다. 매일 아침 핸드폰의 알람이 울려서 시작되는 하루는 아침 기상시간부터 데이터로 기록된다. 출근과 등교를 위해 길을 나서자마자 우리를 찍는 CCTV, 버스나 지하철을 타면서 사용하는 교통카드의 탑승 기록, 회사와 학교에 도착하면서 찍는 출근 카드, 컴퓨터를 켜면서 접속하는 각종 로그 기록들... 우리는 우리의 존재 자체가 데이터로 기록되는 시대에 살고 있다. 알게 모르게 발생되는 데이터들은 모두 기록되고 분석되어 다시 우리의 생활 속으로 들어온다. 쇼핑 패턴이 분석되어 날아오는 마케팅 메일들, 자주 보는 동영상들의 패턴이 분석되어 추천되는 유튜브 추천 영상들, 네비게이션의 경로 추천 등이 대표적이 예이다. 최근 한 컨설팅 기관에서는 하루에 디지털 공간에서만 생산되는 데이터의 양이 2ZB(제타바이트, 10^{21})를 넘는다는 추정치를 내놓았고 데이터를 현대의 금광이라는 표현까지 쓰고 있다. 그렇다면 당신에게 데이터가 주어진다면 이 데이터를 어떻게 사용하겠는가? 과연 당신은 데이터와 얼마나 대화가 가능한가? 이렇게 세상의 모든것이 데이터로 기록되는 세상에서 우리는 이 데이터를 얼마나 사용하고 있는가?

수많은 데이터를 사용하여 의사결정까지 가기 위한 단계는 다음의 그림과 같다. 이와 같이 데이터를 다루고, 모델링하고, 분석하고, 예측하는 일련의 과정을 데이터 과학이라고 한다. 데이터 과학의 단계에서는 먼저 데이터를 수집하고 요약하고 클리닝하는 과정을 거치게 된다. 이후 탐색적 데이터 분석 단계와 모델링을 통한 분석 및 예측을 진행하고, 분석 결과를 문서화하기 위한 시각화 단계로 마무리 된다. 이 과정에서 데이터 시각화는 탐색적 데이터 분석의 단계와 마지막 문서화를 위한 단계에서 필요하다.

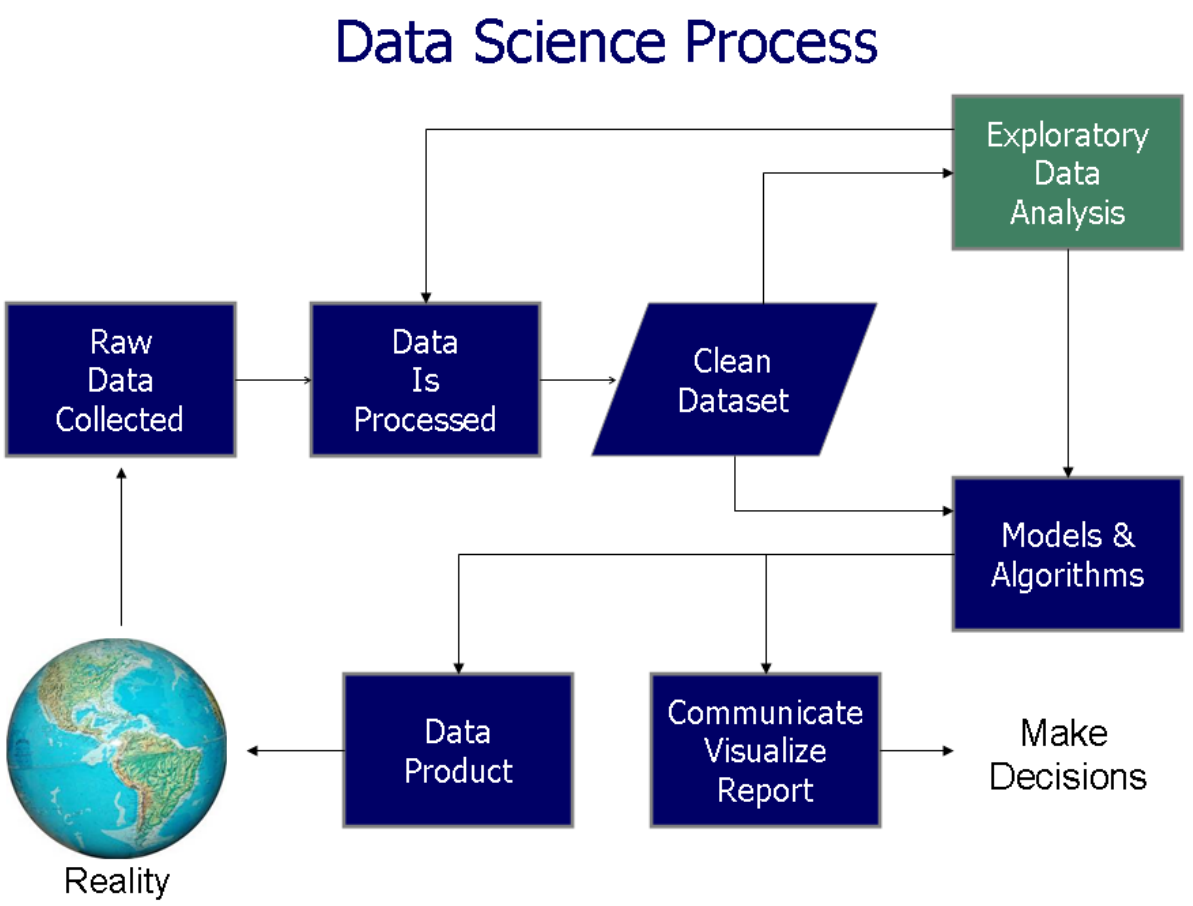


그림 1-1 데이터 프로세스 단계(출처: https://en.wikipedia.org/wiki/Data_visualization#/media/File:Data_visualization_process_v1.png)

효율적인 의사결정을 위해서는 수많은 데이터 사이에서 당신이 발견한 무언가를 일목요연하게 정리하여 누군가에게 이야기하고 설득할 수 있도록 정리해야 한다. 이렇게 데이터를 요약하고 분석하여 청중이 알기 쉽게 정리한 결과를 적절한 양식으로 작성하여 전달하는 능력을 데이터 리터러시라고 한다. 데이터 리터러시는 빅데이터 시대를 살아야하는 현대인에게는 매우 중요한 소양이 되었다.

빅데이터 시대의 기업들은 비지니스 과정에서 데이터를 수집하고 해석하여 데이터가 알려주는 통찰력(Insight)을 만들어내기 위해 비지니스의 첫번째 과정부터 데이터를 활용할 수 있는 능력을 지닌 사람들을 필요로 한다. 이 역량은 누구나 배양할 수 있는 역량이며, 기업도

직원들이 이 역량을 강화할 수 있도록 많은 지원을 하고 있다. 결국 이는 데이터 기반의 의사결정이 비지니스 성과를 크게 향상시키는 중요한 요인으로 작용하고 있다.[1]

[1] https://hbr.org/2020/02/boost-your-teams-data-literacy

표 I - 1. 데이터 리터러시 하위 역량(출처: https://dbr.donga.com/article/view/1201/article_no/8184)

하위 역량	내용
데이터 수집	수행하려고 하는 과제에 필요한 데이터를 알고, 데이터 유형별 장단점을 이해한다. 업무에 필요한 데이터를 빠른 시간 내에 검색, 확인을 통해 확보한다. 비정형적인 데이터 소스로부터 필요한 데이터를 선별적으로 추출한다.
데이터 관리	원시 데이터에서 노이즈(오류 및 방해요소)를 제거해 분석이 가능한 형태로 전환한다. 데이터 소스로부터 형성된 신규 데이터가 주기적으로 입력되도록 자동화한다. 형태가 다른 데이터 세트에서 추출, 필터링, 조정등을 통해 하나의 데이터 세트로 통합한다.
가공 및 분석	확보, 정리된 데이터를 분석의 목적에 맞는 데이터 세트로 가공한다. 다양한 수준의 데이터 분석을 실시하고 복잡한 통계 처리를 위한 쿼리(질문)을 설계한다. 정량적, 정성적으로 분석된 결과를 해석하여 설득력 있는 보고서로 작성한다.
데이터 시각화	그래프, 차트, 인포그래픽 등 직관적이고 효과적인 방식으로 데이터를 표현한다. 방대한 데이터에서 빠르게 특징적인 패턴이나 특이 사항을 추출한다. 변수의 추이에 따라 결과값이 시계열적으로 바뀌어 표현하도록 프로그래밍한다.
데이터 기획	업무 프로세스 별로 생성되는 데이터 종류와 양을 파악한다. 다양한 데이터 간의 관계를 유추하고 분석/활용 방법을 도출한다. 데이터를 수집, 관리, 분석하는 일련의 과정을 계획, 실행, 개선한다.

이렇게 데이터 리터러시가 비지니스 성과를 가르는데 중요한 요소임에도 불구하고 이 역량이 충분한 기업들은 많지 않다. 글로벌 기업의 21% 직원들만이 데이터 리터러시를 완전히 활용할

수 있다는 설문 결과[2]가 보고되고 있으며 최고 경영진의 75%가 비지니스 의사 결정에 중요하지만 30%만이 코로나 19 로 인한 팬데믹 이후 데이터의 활용이 늘었다고 답한 결과도 보고되었다.[3]

이와 같이 기업 비지니스에는 데이터가 점점 중요해졌지만 데이터를 잘 활용할 수 있는 능력을 지닌 전문가는 아직도 부족한 상태이다. 이는 데이터를 분석하고 해석하는 능력을 배양하는 것은 아직 늦지 않았다는 의미이다.

1. 데이터 시각화란?

연어와 문자는 우리가 의사를 전달하는 가장 명확한 수단이다. 지금까지의 수천 년동안 인류는 문자를 통해 자신의 생각을 언어와 문자로 상대방에게 전달하였다. 과거에는 데이터가 충분하지 않았기 때문에 자신의 생각을 뒷받침하는 소량의 데이터를 문자로 전달하는 것이 가능했다. 하지만 지금과 같이 수많은 데이터가 자신의 생각을 설득하는 근거로 사용되는 시대에서는 짧은 문자나 언어로 이를 표현할 수 없다. 게다가 인간은 시각적 자극에 매우 민감하다. 따라서 자신의 주장에 근거가 되는 데이터를 시각적으로 표현하는 것은 상대방을 설득하는데 매우 효과적인 도구로 사용된다.

데이터 시각화는 데이터 또는 정보를 시각적으로 전달하기 위한 표현 기법을 의미한다. 데이터 시각화의 목표는 데이터 또는 정보가 가지는 의미를 상대에게 명확하고 효과적으로 전달하여 상대방을 설득하는 것이다. 일반적으로 데이터는 표, 플롯, 그래프로 구성된 차트, 인포그래픽, 다이어그램 또는 지도의 형태로 시각화된다. 예를 들어 제품의 판매량과 날씨와의 상관관계, 지역별 판매량의 추세, 1 년 중 판매가 집중되는 기간 등을 그래프나 표로 표현하여 의사결정자들이 비지니스 의사결정을 하는데 도움을 줄 수 있다.

최근에는 데이터 시각화가 단순히 정보의 전달 차원을 넘어서 데이터를 사용한 스토리텔링(Story Telling)이 구현되고 있으며 예술의 경지에까지 이르고 있다. 결국 데이터

---

[2] https://www.accenture.com/_acnmedia/PDF-115/Accenture-Human-Impact-Data-Literacy-Latest.pdf

[3] https://www.cityam.com/exclusive-pandemic-has-exposed-lack-of-data-literacy-amid-growing-digital-skills-gap/

시각화는 데이터 기반의 커뮤니케이션 기능이 가장 우선이지만 창의적이고 보기 좋게 구성하는 것이 더 중요해지고 있다는 의미이다.

하지만 무엇보다도 데이터 시각화는 데이터의 시각적 표현을 통해 데이터 안에 내재되어 있는 패턴과 상관관계를 쉽게 파악할 수 있어 비지니스의사 결정 프로세스에 긍정적인 영향을 준다. 비지니스 환경에서 데이터 분석과 시각화가 가지는 의미는 다음과 같다. [4]

- **데이터 간의 상관 관계(Correlation):** 데이터 시각화는 독립 변수와 종속 변수 간의 상관 관계를 쉽게 식별할 수 있다. 이와 같은 상관관계의 파악은 비지니스 의사결정에 큰 도움이 된다.
- **시간 경과에 따른 추세(Trend):** 추세의 시각화는 데이터 시각화의 가장 효율적인 응용분야이다. 과거와 현재의 정보 없이 예측을 한다는 것은 불가능하기 때문에 시간 경과에 따른 추세는 우리가 어디에 있었고 잠재적으로 어디로 갈 수 있는지 알려줄 수 있는 중요한 정보이다.
- **빈도(Frequency):** 빈도는 시간 경과에 따른 추세와 밀접하게 관련된 요소이다. 고객이 구매하는 비율 또는 빈도와 구매 시점을 조사하면 잠재적인 신규 고객이 다양한 마케팅 및 고객 확보 전략에 어떻게 반응하고 반응할지 더 잘 알 수 있다.
- **시장 조사:** 데이터 시각화는 다양한 시장의 정보를 사용하여 관심을 집중해야 하는 대상과 멀리해야 하는 대상에 대한 통찰력을 제공한다. 대상별로 정리된 판매량에 데이터를 다양한 차트와 그래프에 표시함으로써 해당 시장 내의 기회에 대해 정확히 전망할 수 있다.
- **위험 및 보상:** 데이터 시각화 없이는 복잡한 스프레드시트와 숫자를 해석해야 하므로 가치 및 위험 메트릭을 살펴보는 데 전문 지식이 필요하다. 정보가 시각화되면 위험에 따른 조치가 필요할 영역과 위험의 대응이 필요하지 않는 영역을 정확히 찾아낼 수 있다.

---

[4] https://analytiks.co/importance-of-data-visualization/

- **시장 대응:** 대시보드를 통해 구현된 대화형 데이터 시각화는 변화하는 데이터에서 빠르고 쉽게 정보를 얻을 수 있다. 이는 기업이 변화하는 환경에 신속하게 대응할 수 있으며 실수를 방지할 수 있는 정보를 제공한다.

인간의 두뇌는 보통 7 가지 이상의 데이터를 처리할 수 없다고 한다.[5] 정리되지 않은 데이터, 이해하기 쉽게 표현되지 않은 데이터는 우리에게 단순한 문자와 숫자들의 집합이라는 의미 외에는 별다른 의미가 없다. 이것이 데이터 시각화가 필요한 가장 큰 이유이다.

이제 의사결정을 위해 수천 행의 스프레드시트를 샅샅이 뒤지던 시대는 끝났다.

2. 왜 데이터 시각화를 해야하는가?

미국에는 '천 마디의 말보다 한번 보는게 낫다'(A picture is worth a thousand words)라는 격언이 있다. 이 말은 20 세기 초 미국의 한 신문광고에서 게재되면서 유명해졌는데 사실 공자의 화의능달만언(畵意能達萬言)에서 유래됐다고 전해지기도 한다. 이 문구는 데이터의 시각화가 왜 필요한지 가장 잘 말해주고 있다. 결국 데이터 시각화를 해야하는 이유는 공자 시대부터 강조되어 왔다는 것이다.

[5] George A. Miller, The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information, Psychological Review, Vol. 101, No. 2, 343-352, 1955

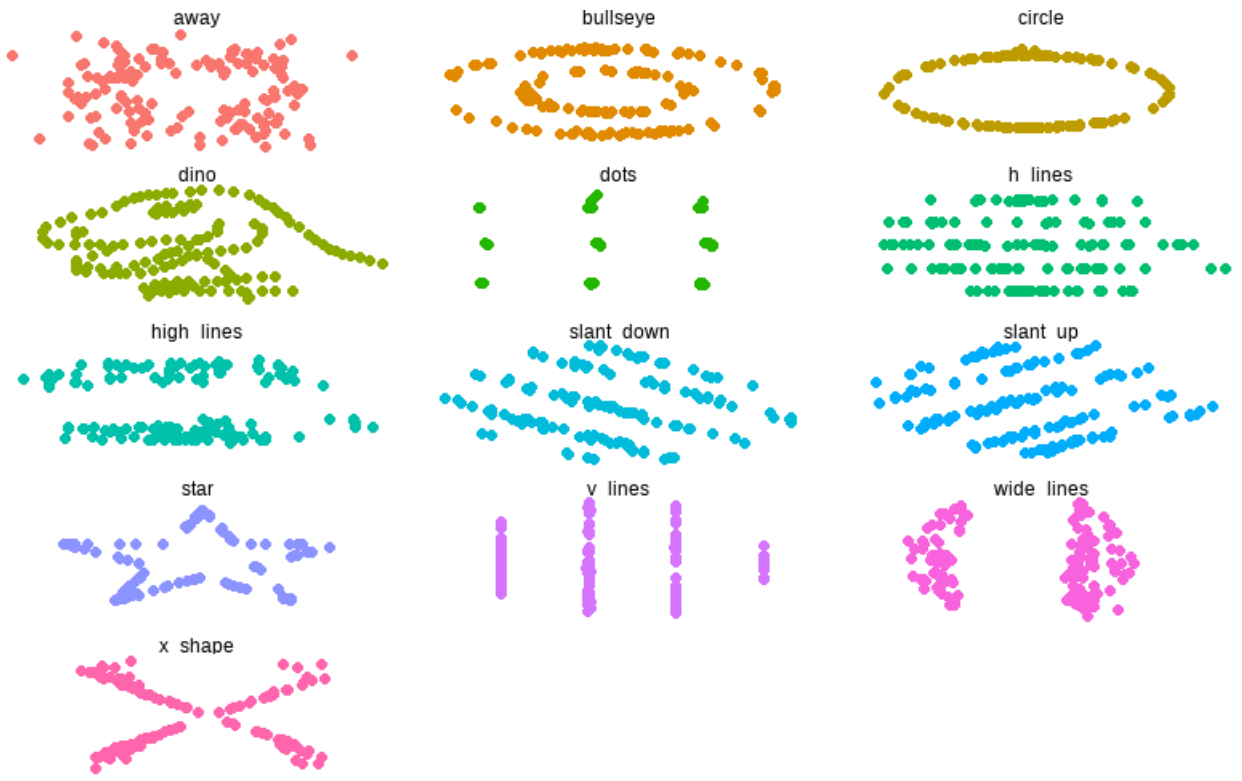
그림 1-2 시각화 신문광고(출처: https://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words)

데이터 시각화는 데이터 클리닝, 데이터 구조의 탐색, 이상값 및 비정상적인 그룹의 탐지, 추세 및 클러스터 식별, 로컬 패턴 발견, 모델링 출력, 평가 및 결과를 확인하는데 효과적인 도구이다. 특히 데이터 과학의 첫번째 단계로 제시되는 탐색적 데이터 분석 과정에서 데이터 시각화는 반드시 수행해야 하는 과정이다.

탐색적 데이터 분석과정에서 데이터 시각화는 데이터의 비정상 분포, 결측값, 이상값 등 통계 및 모델로는 알아내기 어려운 부분을 쉽게 파악할 수 있다. 다음의 그림은 모두 동일한 평균, 표준편차, 상관계수를 가지는 데이터들이다. 수치로만 확인하면 13 가지 데이터 셋은 모두 같은 분포를 가지는 것으로 생각된다. 하지만 이들 데이터를 시각화하면 그 형태가 각각 다르다는 것을 쉽게 알 수 있다.

```
## # A tibble: 13 x 6
##    dataset    mean_x mean_y std_dev_x std_dev_y corr_x_y
##    <chr>       <dbl>  <dbl>     <dbl>     <dbl>    <dbl>
##  1 away         54.3   47.8      16.8      26.9  -0.0641
##  2 bullseye     54.3   47.8      16.8      26.9  -0.0686
##  3 circle       54.3   47.8      16.8      26.9  -0.0683
##  4 dino         54.3   47.8      16.8      26.9  -0.0645
##  5 dots         54.3   47.8      16.8      26.9  -0.0603
##  6 h_lines      54.3   47.8      16.8      26.9  -0.0617
##  7 high_lines   54.3   47.8      16.8      26.9  -0.0685
##  8 slant_down   54.3   47.8      16.8      26.9  -0.0690
##  9 slant_up     54.3   47.8      16.8      26.9  -0.0686
## 10 star         54.3   47.8      16.8      26.9  -0.0630
## 11 v_lines      54.3   47.8      16.8      26.9  -0.0694
```

```
## 12 wide_lines    54.3   47.8      16.8      26.9  -0.0666
## 13 x_shape       54.3   47.8      16.8      26.9  -0.0656
```



실행결과 1-1 그림 1-3 동일한 통계치의 다른 분포

데이터 시각화를 통해 얻을 수 있는 이점은 다음과 같다.

- 데이터 이해가 편리

  데이터 시각화는 우리가 수많은 데이터를 헤매면서 데이터를 해석하지 않도록 도와준다. 데이터의 전반적인 분포, 상관, 패턴을 쉽게 파악할 수 있어 데이터를 이해하고 해석하기가 쉽다. 이러한 이유로 각종 보고서에 데이터 시각화를 많이 사용하게 되고 이를 통해 보고서 작성자의 의견을 쉽게 전달할 수 있다. 따라서 판매 보고서든 마케팅 전략이든 데이터 시각화는 기업이 더 나은 분석 의사 결정을 유도하고 이는 결국 기업의 수익성을 높이는 데 도움이 된다.

- 빠른 의사결정

인간은 문자보다 시각적 이미지를 60,000 배 빠르게 인식한다고 한다.[6] 따라서 차트, 플롯, 그래프 등으로 전달되는 데이터 시각화 이미지는 문자로 그 내용을 확인하고 인식하는 것보다 훨씬 이해하는데 빠르고 쉽다. 한 연구에서 시각적 이미지와 문자를 혼합한 발표자는 문자만 사용하는 발표자보다 17% 더 설득력이 있는 것으로 나타났다. 미네소타 대학의 또 다른 연구에 따르면 시각적 이미지가 있는 발표자는 청중을 설득하는 데 43% 더 효과적이다고 발표하였다.[7] 또 와튼 비지니스 스쿨에에 따르면 데이터 시각화는 비즈니스 회의를 최대 24%까지 단축할 수 있다고 발표하였다.[8] 이러한 시각적 데이터를 쉽게 해석할 수 있는 능력 덕분에 데이터 시각화는 의사 결정 프로세스의 속도를 크게 향상시킬 수 있는 핵심 요소가 되었다.

- 청중의 주의 집중력 향상

  데이터의 분석과정에서 사용되는 다양한 지표들은 통계적 지식이 부족한 청중들에게는 이해하기가 매우 어려운 일이다. 이러한 지표를 많이 사용한다면 데이터에 대한 설명을 듣는 청중들의 주의력과 집중력이 저하되어 발표자의 주장에 쉽게 설득되지 않는다. 따라서 통계적 지식이 없는 청중들도 데이터를 쉽게 이해할 수 있는 도구가 필요하고 이에 가장 좋은 도구가 데이터 시각화이다. 아름답게 디자인되고 알아보기기 쉬운 데이터 시각화는 청중의 주의력과 집중력을 향상시켜 발표자가 말하고자 하는 내용을 쉽고 빠르게 이해할 수 있다.

- 데이터 패턴의 파악

  데이터들은 패턴이라고 하는 의미를 지니고 있다. 패턴이 강한 데이터일수록 데이터의 예측능력이 높아지기 때문에 비지니스 환경에서 매우 좋은 데이터가 된다. 이 때문에 데이터가 가지는 패턴을 쉽게 찾아내기 위해 많은 방법이 제시되었다. 사실 데이터 시각화는 데이터 계산 외에 그래픽카드를 통한 이미지 계산까지 해야하므로 컴퓨팅 파워가 지금과 같이 높지 않을 때는 통계적 방법만을 활용하여 대량의 데이터에서 패턴을 찾아내었다. 하지만 앞선 그림에서도 보듯이 데이터 시각화는 통계적 분석이

---

[6] https://oit.williams.edu/files/2010/02/using-images-effectively.pdf

[7] https://www.amanet.org/articles/using-visual-language-to-create-the-case-for-change/

[8] https://blog.datumize.com/top-five-advantages-of-data-visualization

찾아낼 수 없는 패턴까지 찾아낼 수 있고 이 패턴은 매우 직관적으로 확인할 수 있어 통계적 지식이 없는 사람들도 쉽게 데이터를 이해할 수 있게 도와준다.

- 오류 및 이상치 검출

  데이터 시각화는 앞선 패턴 파악과 유사하게 통계적 방법에서 찾아낼 수 없고, 통계적 방법에서 효과적으로 전달할 수 없는 이상치와 오류값 들을 명확하고 직관적으로 제시할 수 있다.

- 데이터 스토리텔링(Story Telling)

  데이터 스토리텔링은 최근 매우 각광받는 데이터 시각화의 방법이다. 보통 스토리텔링은 대쉬보드(DashBoard)를 통해서 전달된다. 발표자가 자신이 데이터를 전달하는 과정을 시각적 자료로 구성함으로써 청중이 데이터를 차근차근 이해하도록 이끌어 가는 과정을 말한다. 스토리텔링에 사용되는 시각 자료는 가급적 단순하게 작성되는게 일반적이다. 스토리를 통해 전달하는 데이터는 개별적으로 생성된 데이터 시각화 여러 개보다 훨씬 효과적이다.

- 비지니스 인사이트

  경쟁적인 비즈니스 환경에서 최고경영자의 적절한 비지니스 의사결정은 기업의 생존을 위해 가장 중요한 능력일 것이다. 과거의 경영진들은 이러한 의사결정을 직관이나 감과 같은 주관적 의지에 의해 결정하였다. 하지만 최근의 경영진들은 이러한 의사결정을 감이나 직관에서 벗어나 많은 데이터를 탐색하고 분석하는 과정에서 얻어지는 객관적 데이터를 중요하게 생각한다. 이 과정에서 데이터 시각화는 매우 중요한 자료로 활용할 수 있다.

3. 데이터 시각화로 무엇을 표현하는가?

데이터 시각화는 수많은 데이터의 특성을 한눈에 볼 수 있도록 만드는 과정이다. 그렇다면 데이터 시각화는 데이터의 무엇을 표현하고 어떤 특성들을 나타낼 수 있는가? 우리는 초등학교 시절부터 데이터 시각화를 배워왔다. 지금도 초등학교 4 학년 1 학기 수학 교과서에 막대 그래프를 사용하여 데이터를 표현하는 방법을 배운다. 그래서인지 데이터 시각화에서 가장 많이 쓰이는 방법은 막대 그래프이며 경우에 따라서 선 그래프, 파이 차트 등이 흔하게

사용되고 있다. 데이터 시각화를 쓰기 전에 먼저 데이터 시각화로 무엇을 표현하고 어떤 데이터 시각화 방법을 쓸지를 결정해야 한다. 일반적으로 데이터 시각화를 통해 데이터의 분포(Distribution), 비교(Comparison), 추세(Trend), 구성(Composition)을 표현하는데 효과적으로 활용된다.

- 분포(distribution)

  분포는 데이터들이 전반적으로 어떻게, 얼마나 흩어져 있는지를 파악하기 위해 사용된다. 보통 분포를 나타내기 위해 사용되는 데이터는 요약된 데이터가 아닌 원 데이터 차원를 사용하는 것이 일반적이다. 분포 시각화를 통해 수치로 보여지는 평균, 중위수, 범위, 분산, 표준편차들을 보다 직관적으로 파악할 수 있다는 장점을 가진다. 전체 데이터 스펙트럼을 보고 관련되거나 관련되지 않은 데이터 포인트를 시각화하는 데 도움이 될 수 있다.

- 비교(Comparison)

  비교는 특정 변수의 변화에 따른 값들의 차이를 확인하기 위해 사용되는 데이터 시각화 방법이다. 비교에 사용되는 변수나 범주가 적을 경우는 막대 그래프나 선 그래프 등으로 표현이 가능하지만 비교에 사용되는 변수의 양이 많아지면 히트맵과 같은 방식이 사용된다.

- 추세(Trend)

  추세는 일반적으로 시간의 흐름에 따른 값의 변화를 표현하는 방법이다. 보통 선 그래프를 사용하여 표현하는데 막대 그래프를 사용하는 경우도 많이 있다. 추세의 시각화를 통해 데이터의 전반적인 증감 현황을 쉽게 파악할 수 있고 계절성과 같은 부가적인 정보도 파악할 수 있다.

- 구성(Composition)

  구성은 전체에 대한 비율(백분율)을 시각적으로 표현하는데 사용된다. 보통 원형 차트 또는 막대 차트를 통해 표현되는데 시장 점유율과 같은 비율의 비교를 표시할 수 있다.

- 상관(Correlation)

상관은 종속 변수와 독립 변수로 표시되는 데이터를 시각화 함으로써 전체적인 분포가 X 값과 Y 값에 대하여 어떤 관계를 가지는지를 보여주는 시각화이다. 상관 시각화는 일반적으로 산점도를 통해 표현된다.

- 지리(Geographic)

  데이터를 지도나 약도 위에 표현함으로써 지역간의 데이터의 차이를 보여주는 시각화 방법이다.

4. 유명한 데이터 시각화의 사례

4.1. 나폴레옹의 러시아 원정 데이터 시각화(Minard 다이어그램)

Minard 다이어그램은 나폴레옹의 러시아 원정 과정(1812~1813)에서 프랑스 군대가 입은 손실을 시각화하고 있다. 이 시각화는 폴란드에서 러시아 국경으로 출정하는 나폴레옹의 프랑스 군대의 손실을 이동거리, 온도, 위경도, 진격 방향 등의 6 가지 변수를 통해 매우 효과적으로 표현하고 있다. 이 시각화는 나폴레옹의 프랑스 군대가 얼마나 비참한 패배를 했는지 한눈에 표현된다. 예일대학의 유명한 통계학 교수인 Edward Tuft 는 "지금까지 그려진 것 중 최고의 통계 그래픽일 것"이라고 했다.[9]

[9] Tufte, Edward (1983). *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphics Press. ISBN 0-9613921-4-2.

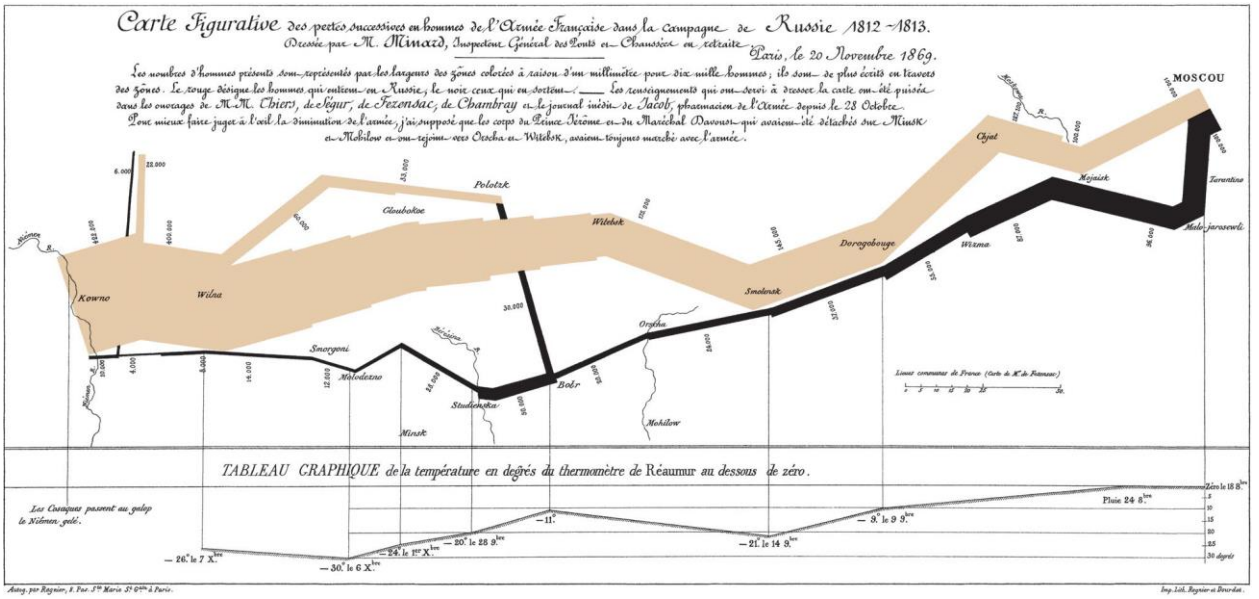
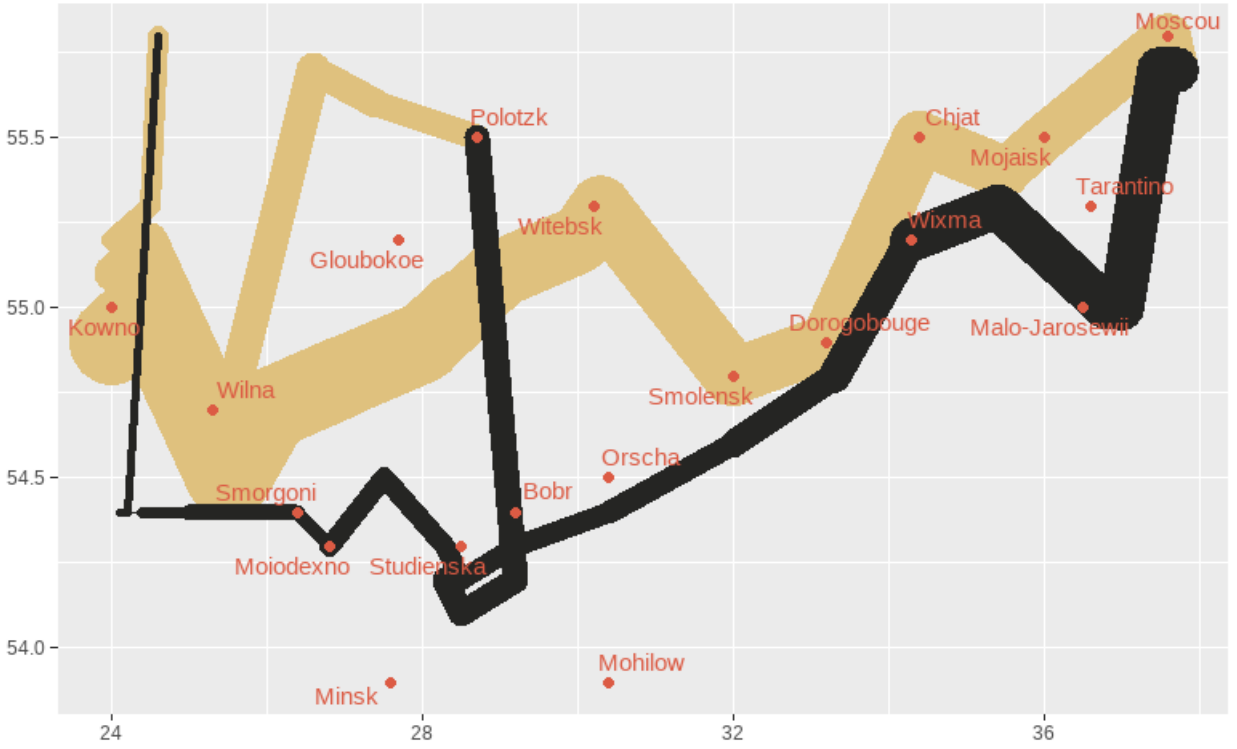


그림 1-4 나폴레옹의 러시아 원정 시각화(출처: https://en.wikipedia.org/wiki/Charles_Joseph_Minard)



실행결과 1-2 그림 1-5 R 로 구현한 Minard 다이어그램

4.2. 나이팅게일의 로즈 다이어그램(Rose Diagram)

흔히 나이팅게일은 우리에게 '나이팅게일 선서'로 유명한 간호사이다. 하지만 나이팅게일을 검색하면 간호사와 함께 통계학자라고 검색된다. 나이팅게일은 1853 년에 발발한 크림전쟁에 간호사로 참전한다. 전쟁의 부상자를 치료하는 과정에서 숨진 많은 병사들이 총상으로 인한 사망이 아닌 열악한 위생 환경으로 인한 전염병으로 인한 사망이라는 것을 알아냈다. 이 결과를 다양한 통계와 Rose 다이어그램이라는 데이터 시각화를 통해 영국 정부를 설득하여 40%에 이르던 영국군 부상자의 사망률을 2%까지 낮추데 큰 역할을 하였다.

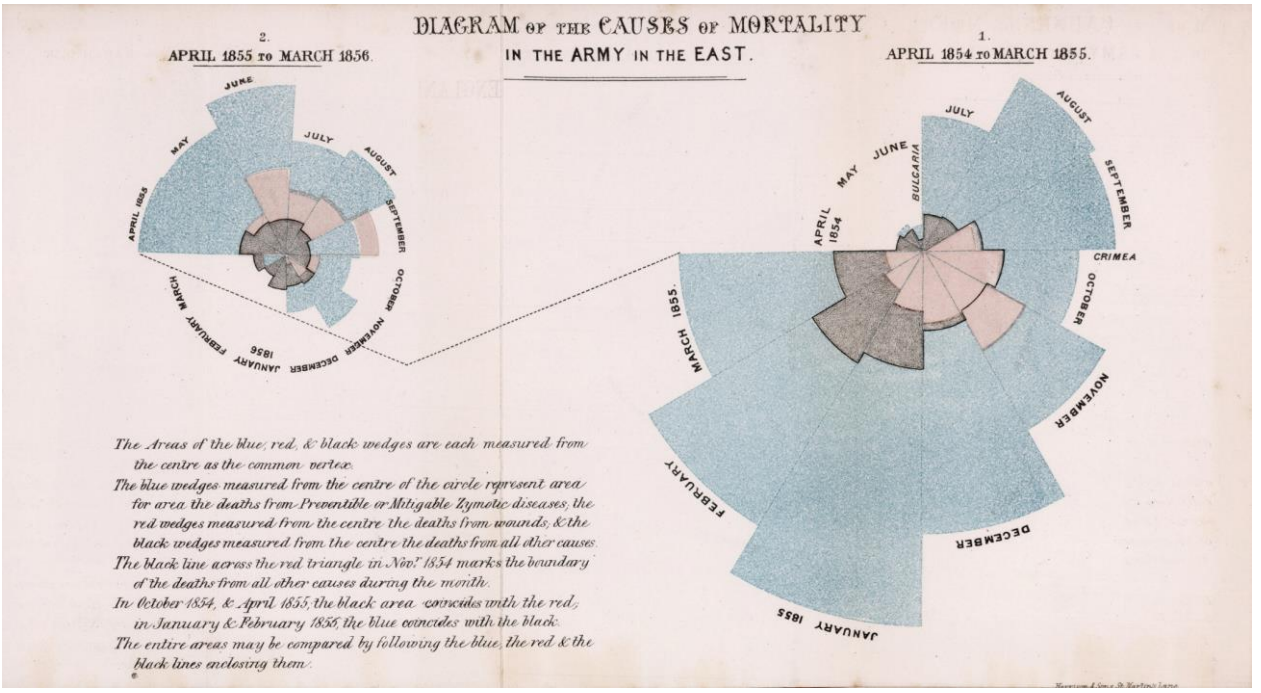


그림 1-6 나이팅게일의 로즈 다이어그램(출처: https://en.wikipedia.org/wiki/Florence_Nightingale)

4.3. 영국의 콜레라 발병 지도

1854 년에 영국의 런던 웨스트민스터의 소호(Soho) 지역에서 대규모 콜레라가 발병했다. 이 당시 런던의 소호 지역은 산업화로 인해 인구가 급격히 늘었지만 사회 인프라의 미비로 위생시설이 적절히 갖추어지지 못했다. 특히 하수도 시스템의 부재는 사람들의 생활 환경을 매우 오염시킨 원인이 되었다. 이런 상황에서 1832 년과 1849 년에 발생한 발병한 콜레라로 영국에서 총 14,137 명이 사망했다. 당시에는 세균이 병의 원인이 될 수 있다는 사실을 몰랐던 시기였기 때문에 콜레라의 원인을 엉뚱한 곳에서 찾고 있었다. 이 때 영국의 의사였던 존

스노우는 콜레라가 특정 상수도 펌프를 중심으로 발병한다는 사실을 파악하고 수인성 전염병이라는 사실을 증명하기 위해 상수도 주변의 사망자 수를 표기한 지도를 만들었다. 이 지도를 통해 지역 이사회를 설득하였고 콜레라의 주 원인으로 지목된 상수도 펌프를 폐쇄하는 성과를 이룸으로써 콜레라를 효과적으로 차단하였다.

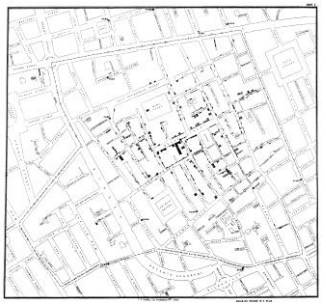


그림 1-7 영국의 콜레라 발병 지도(출처: https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak)

4.4. 새로운 역사 차트(New Chart of History)

새로운 역사 차트는 18 세기 영국의 수학자인 Joseph Priestley 가 *역사와 일반 정책에 관한 강의(Lectures on History and General Policy)* 에 대한 보충 자료로 1769 년에 만든 역사의 타임라인에 대한 시각화이다. 새로운 역사 차트에는 두가지 타임라인을 제공한다. 첫 번째 타임라인은 역사상 주요 인물들을 중심으로 한 700 년간의 일대기 차트(Chart of Biography)이며 두 번째 타임라인은 역사에 걸쳐 동시대에 존재한 주요 제국과 문화의 영향에 중점을 두는 타임라인이다.

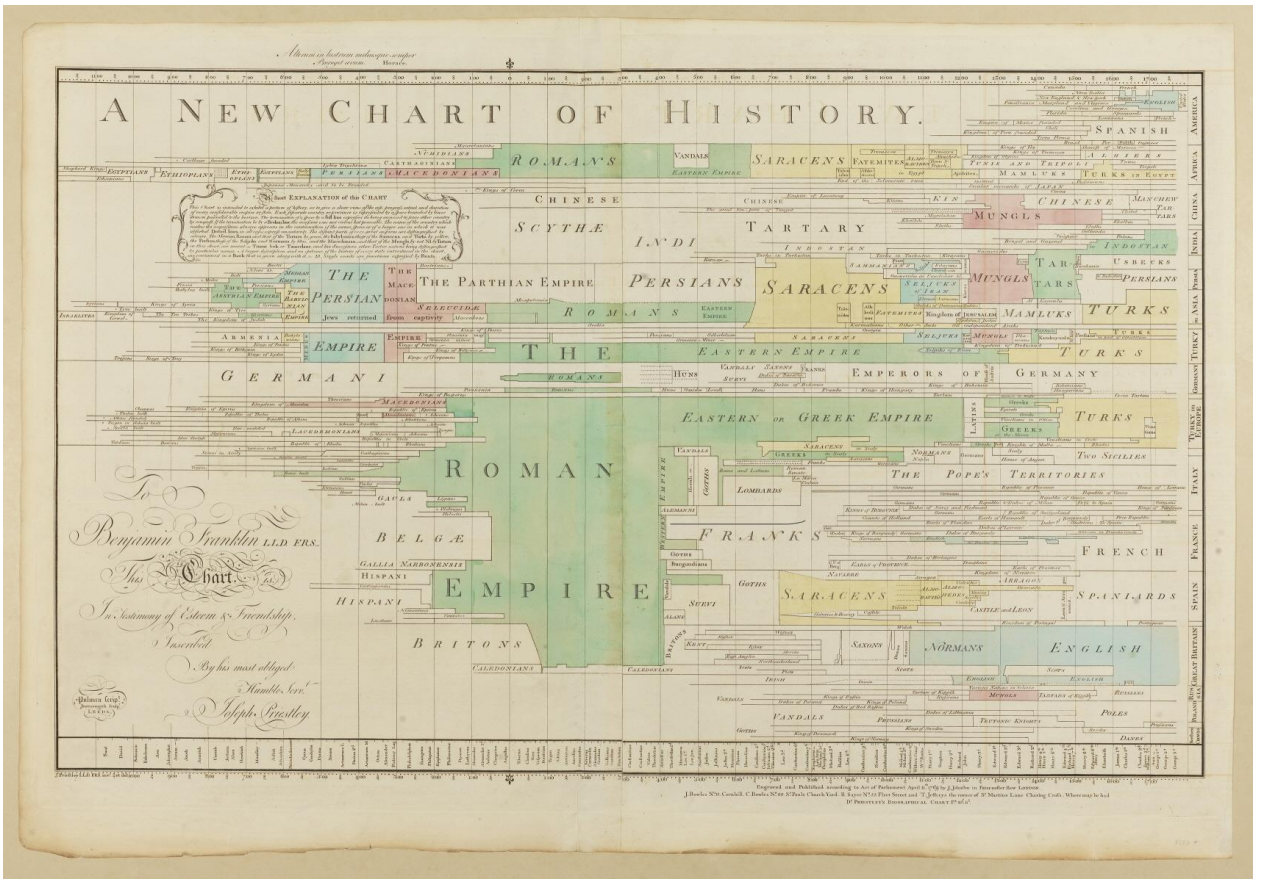


그림 1-8 새로운 역사 차트(출처 : https://en.wikipedia.org/wiki/A_New_Chart_of_History)

5. 차트(Chart), 플롯(Plot), 그래프(Graph)

시각화를 수행하다 보면 흔히 만나는 용어가 차트, 플롯, 그래프이다. 일반적으로 차트는 파이 차트, 도넛 차트 등, 플롯은 박스 플롯, 바이올린 플롯 등, 그래프는 선 그래프, 막대 그래프 등이 있다. 하지만 막대 그래프, 막대 차트와 같이 혼용해서 쓰는 경우도 있다. 그렇다면 차트와 플롯, 그래프의 차이는 무엇일까?

사실 어디에도 이 세가지 용어의 차이를 명확하게 구분해 놓은 정의는 찾기가 어렵다. 혹시 정확한 정의를 찾는다면 필자에게도 알려주시길 바란다. 그리고 그 범위나 종류도 명확히

구분되고 있지 못한 듯 하다. 이 책에서도 정확히 구분하지는 못한다. 하지만 나름대로 다음과 같이 정의해보도록 하겠다.[10]

차트는 특정 문제에 대해 여러 청중을 대상으로 브리핑하기 위해 문자, 숫자, 그래프, 플롯 등을 활용하여 만든 자료를 뜻한다. 컴퓨터를 활용한 프리젠테이션이 시행되기 이전에 큰 종이에 사람이 직접 손으로 작성해서 만들었던 것을 차트라고 했다. 이 차트에는 문제를 설명하기 위해 필요한 다양한 정보들이 표현되어 있다. 따라서 차트는 시각화의 가장 큰 정의라고 할 수 있겠다.

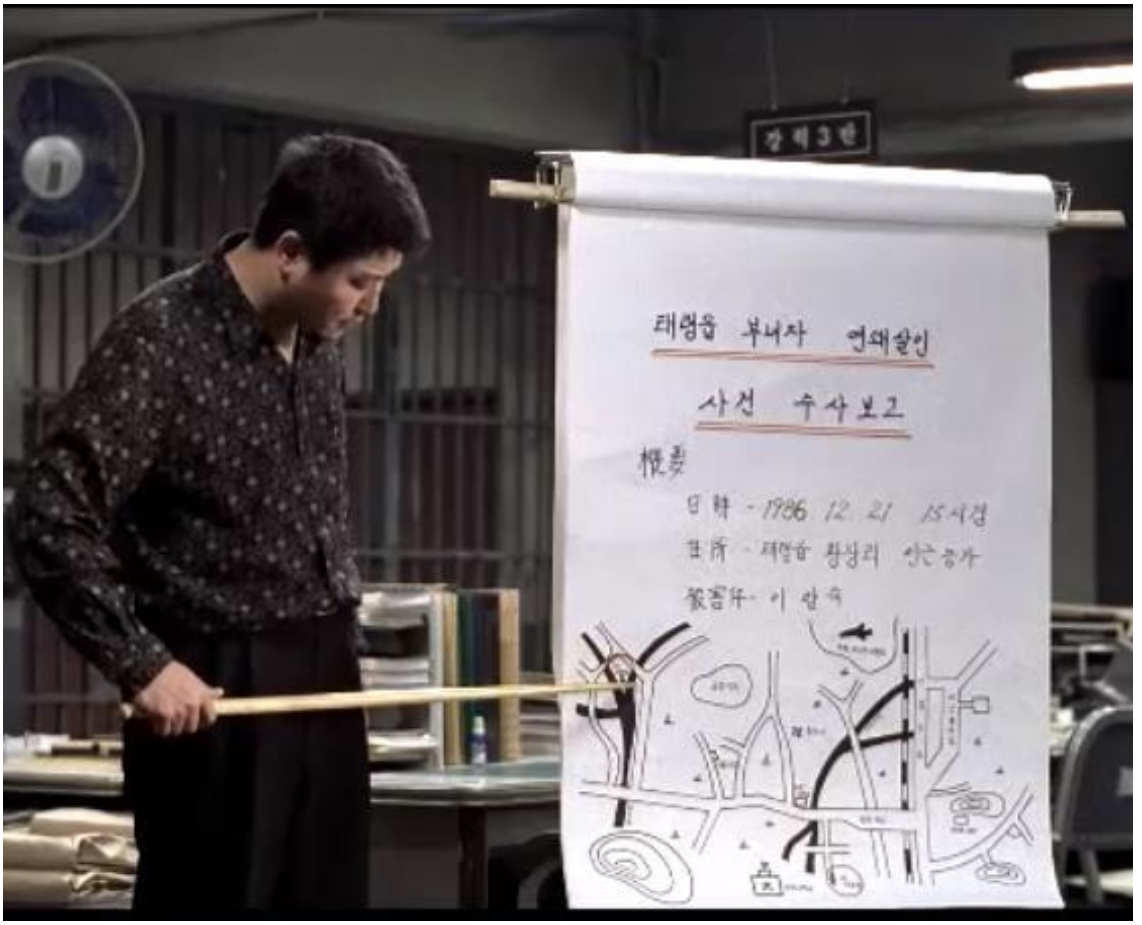


그림 1-9 과거 차트의 사용(출처: 영화 살인의 추억)

[10] 이 정의는 필자의 경험에 의해 내린 자체 정의로 학문적 검증이 되지 않은 것임을 밝힌다. 그래서 본 도서에서도 차트, 플롯, 그래프 등의 용어가 혼용되어 사용된다.

플롯은 주로 데이터의 위치를 직접 표현하는 시각화 방법이다. 사실 영어로 Plot 이라는 단어의 뜻 중에는 '종이나 지도에 어떤 것의 위치나 움직임에 대한 표시'라는 의미가 있다. 결국 데이터에 대한 점을 찍는 시각화를 의미하는 것으로 데이터 자체에 대한 특별한 통계 처리가 없이 데이터가 가진 성질을 그대로 표현하는 시각화 방법이다. 플롯으로 가장 대표적인것이 산점도라 불리는 스캐터 플롯이다. 박스플롯의 경우도 박스의 형태로 표현했지만 사실 그 표현 대상이 각 이산형 변수에 속한 데이터의 위치를 박스로 표현한 것이기 때문에 플롯으로 표기하는 것으로 보인다.

그래프는 수학적 이론에서 그 정의를 찾을 수 있다. 수학적 정의상 그래프는 일부 객체들의 쌍들이 서로 연관된 객체의 집합을 이루는 구조로 점과 점들을 잇는 선으로 구성된 구조라고 정의된다.[11] 이 정의에 가장 가까운 시각화가 선 그래프이다. 사실 막대 그래프도 선이 막대로 표현되었다고 본다면 역시 데이터 점과 축까지 선(막대)으로 이어진 그래프라고 볼 수 있을 것이다.

또 시각화에 많이 사용되는 용어가 다이어그램이다. 다이어그램의 사전적 정의는 정보를 조율, 묘사, 상징화 하여 2 차원 기하학 모델로 시각화하는 기술[12]로 정의되고 있다. 이 다이어그램은 다이어그램 일러스트레이션과 같은 용어로도 표현되는데 그래픽, 기술적 드로잉, 표 정보 등의 기술적 유형의 집합과 데이터의 성질을 표현하기 위해 선, 화살표 등의 시각적 고리들로 연결된 형태의 유형을 포괄한다. 주요 다이어그램의 유형으로 순서도, 벤다이어그램, 챠트, 표 등이 이에 속한다.

---

11 https://ko.wikipedia.org/wiki/%EA%B7%B8%EB%9E%98%ED%94%84_(%EC%88%98%ED%95%99)

12 https://ko.wikipedia.org/wiki/%EB%8B%A4%EC%9D%B4%EC%96%B4%EA%B7%B8%EB%9E%A8