

1주차: 데이터를 보다 (데이터 시각화)

서울시립대학교 통계학과
전종준

수업소개

수업목표

- 고차원 구조 데이터의 속성을 요약하고 전달하기 위한 통계적 방법을 소개
- 전반부에서는 통계적 시각화 도구와 시각화를 구현하기 위한 프로그래밍 방법을 학습
- 후반부에서는 고차원 데이터를 저차원 공간에서 표현하여 그것을 시각화하는 방법

수업 참고사항

- 수업 내용들은 고차원 데이터의 비선형축약과 잠재공간에서 데이터 표현형의 학습 방법론을 포함
- 이 수업은 학부 3, 4학년을 위한 과목으로 파이썬 기초, 선형대수(행렬대수)에 대한 사전 지식이 필요

수업소개

수업 기대효과

- 데이터가 가진 정보를 요약하여 유용한 그래프를 그릴 수 있는 능력 배양
- 머리속에 그리기 어려운 고차원 데이터를 저차원 공간에서 표현하는 통계학의 원리와 시각화 방법론을 배움 (예, t-SNE)

학업상담

- 시간: 개별적으로 사전에 조율
- 장소: 미래관 712호 (교수 연구실) 혹은 709호 (회의실)

수업소개

평가

- 출석 (10%), 중간고사 (40%), 기말고사 (50%)
 - * 결석 5회 (학점 F), 지각 (감점 없으나, 매시 정각에서 10분사이에 교실 출입을 당부)
- 중간고사 혹은 기말고사 미응시자 (학점 F)
 - * 미응시 사유 인정 (유고계, 해외학회 발표를 위한 학회참석) -> 중간/기말고사 점수 중 하나로 평가 후 최종성적은 한 단계 낮추어 조정

데이터 세상 밖으로 나오다

전 세계에 있는 항공편은 어떻게 연결되어 있을까?

데이터 세상 밖으로 나오다

항공편 경로 자료

- 공항 정보

도시	국가	공항	위도	경도	노선수
Tokyo	Japan	NRT	35.76472	140.3864	100
Matsumoto	Japan	MMJ	36.16676	137.9227	7
Ibaraki	Japan	IBR	36.18108	140.4154	4
Kwangju	South Korea	KWJ	35.12639	126.8089	8
Yeosu	South Korea	RSU	34.84233	127.6169	6
Cheju	South Korea	CJU	33.51131	126.493	70
Busan	South Korea	PUS	35.17953	128.9382	46
Ulsan	South Korea	USN	35.59349	129.3517	6

- 항공 노선 정보

공항	위도	경도	연결그룹
ICN	37.46908	126.4505	11
NRT	35.76472	140.3864	11
KIX	34.43472	135.2442	12
ICN	37.46908	126.4505	12
NGO	34.85841	136.8054	13
GMP	37.55831	126.7906	13
NGO	34.85841	136.8054	14



Fuji Dream 134

FDA134 / JH134

[Upgrade account to see tail number](#)

EN ROUTE

Arriving in 1 hour 10 minutes

Activity Log

EN ROUTE FLIGHT				
Date	Departure	Arrival	Aircraft	Duration
Sunday 29-Nov-2020	03:37PM JST Kagoshima - KOJ	05:05PM JST Shizuoka - FSZ	E170	1h 28m
PAST FLIGHTS				
Date	Departure	Arrival	Aircraft	Duration
Saturday 28-Nov-2020	03:35PM JST Kagoshima - KOJ	04:15PM JST (?) Shizuoka - FSZ	E170	0h 40m
Friday 27-Nov-2020	03:39PM JST Kagoshima - KOJ	05:05PM JST Shizuoka - FSZ	E170	1h 26m
Thursday 26-Nov-2020	03:38PM JST Kagoshima - KOJ	05:00PM JST Shizuoka - FSZ	E170	1h 22m
Wednesday 25-Nov-2020	03:35PM JST Kagoshima - KOJ	04:56PM JST Shizuoka - FSZ	E170	1h 21m
Tuesday 24-Nov-2020	03:37PM JST Kagoshima - KOJ	04:55PM JST Shizuoka - FSZ	E170	1h 18m
Monday 23-Nov-2020	03:39PM JST Kagoshima - KOJ	04:59PM JST Shizuoka - FSZ	E170	1h 20m
Friday 20-Nov-2020	04:02PM JST Kagoshima - KOJ	05:19PM JST Shizuoka - FSZ	E170	1h 17m
Thursday 19-Nov-2020	03:38PM JST Kagoshima - KOJ	04:18PM JST (?) Shizuoka - FSZ	E170	0h 40m
Wednesday 18-Nov-2020	03:35PM JST Kagoshima - KOJ	04:15PM JST (?) Shizuoka - FSZ	E170	0h 40m
Tuesday 17-Nov-2020	03:50PM JST Kagoshima - KOJ	04:38PM JST (?) Shizuoka - FSZ	E170	0h 48m



Emirates 9854

UAE9854 / EK9854

Upgrade account to see tail number

Activity Log

EN ROUTE FLIGHT

Date	Departure	Arrival	Aircraft	Duration
Sunday 29-Nov-2020	10:49AM +04 Dubai Int'l - DXB		-	0h 07m

PAST FLIGHTS

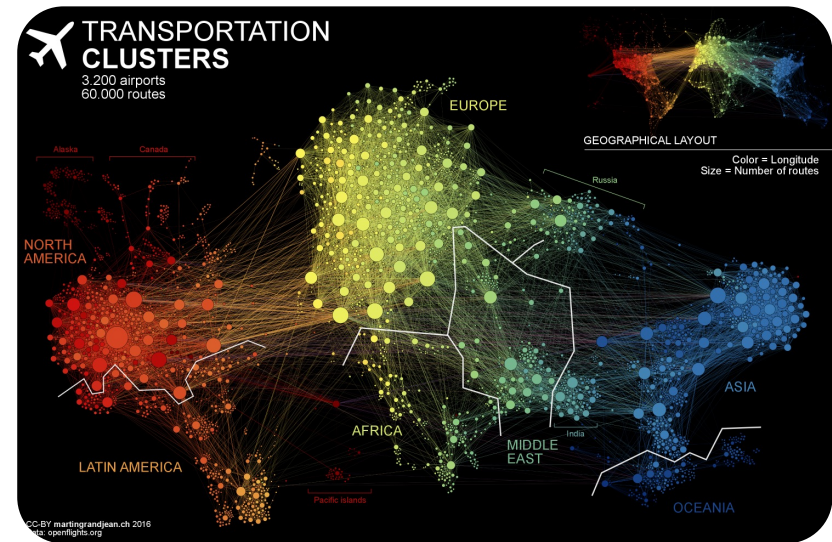
Date	Departure	Arrival	Aircraft	Duration
Sunday 22-Nov-2020	<i>First seen at 10:56AM +04</i> near Dubai, United Arab Emirates	<i>Last seen at 07:38PM +07</i> near Hanoi, Vietnam	-	5h 42m
Sunday 15-Nov-2020	<i>First seen at 11:43AM +04</i> near Dubai, United Arab Emirates	<i>Last seen at 08:32PM +07</i> near Hanoi, Vietnam	-	5h 49m

데이터 세상 밖으로 나오다

- 항공편 경로 자료



출처: <http://www.visualcapitalist.com>



출처: <http://www.visualcapitalist.com>

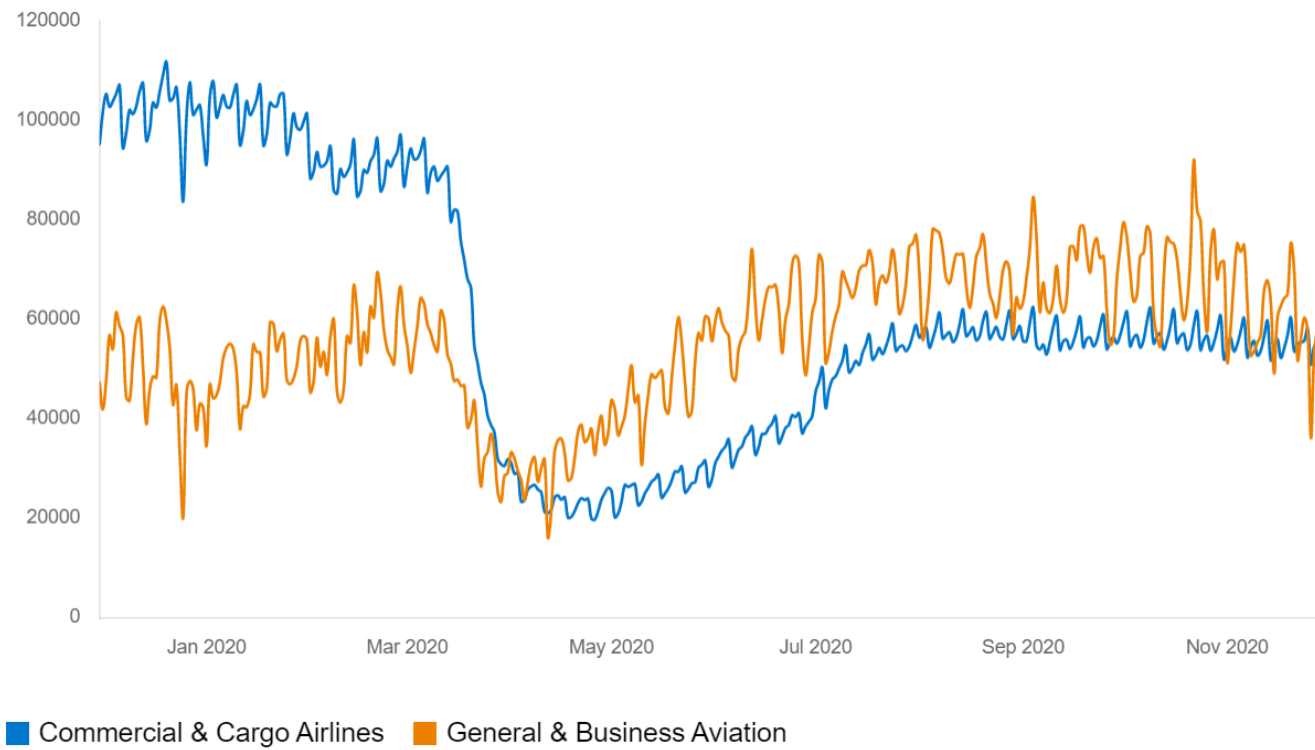


데이터 세상 밖으로 나오다

1년간 전 세계에 있는 항공편의 운항 수 어떻게 변했나?

데이터 세상 밖으로 나오다

GLOBAL ARRIVALS PER DAY OVER THE LAST [1 YEAR](#) ▾



Last updated at 03:37PM KST

데이터 세상 밖으로 나오다

데이터의 메시지를 어떤 방법으로 전달할 것인가?

목차

1. 기록하다, 그리다
2. 범주형 자료분석, 뭐가 달라?
3. 숫자처럼 보이는 범주형 자료
4. 범주형 자료 분석

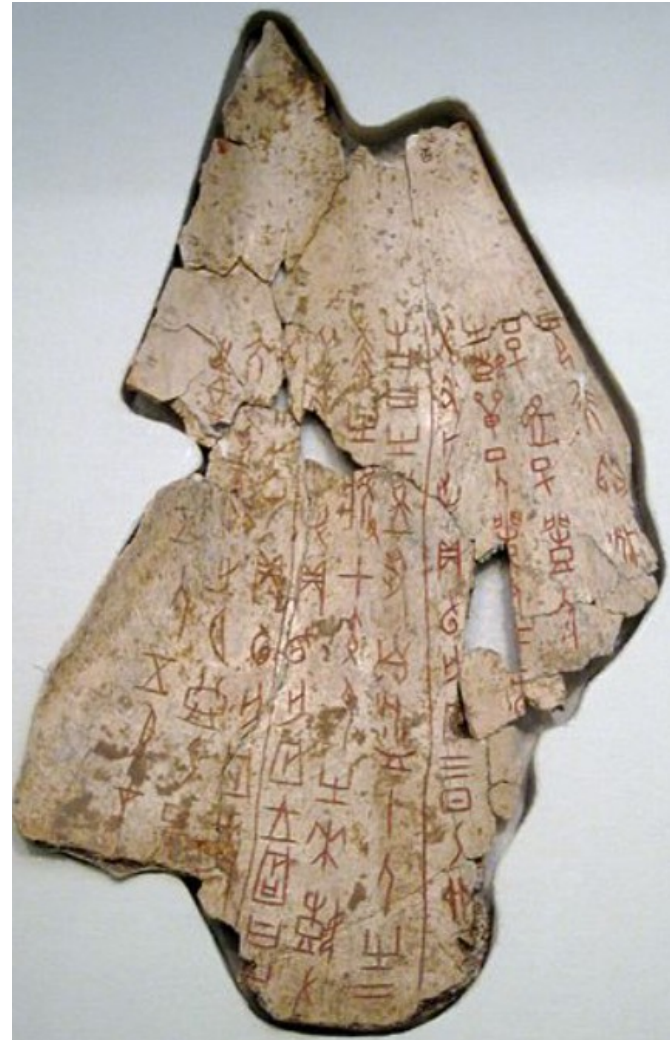
1. 기록하다, 그리다

문자와 기록

상형문자



(이집트)



(중국)

지도와 정보

지도



대동여지도



천상분야열차지도

회화와 기록



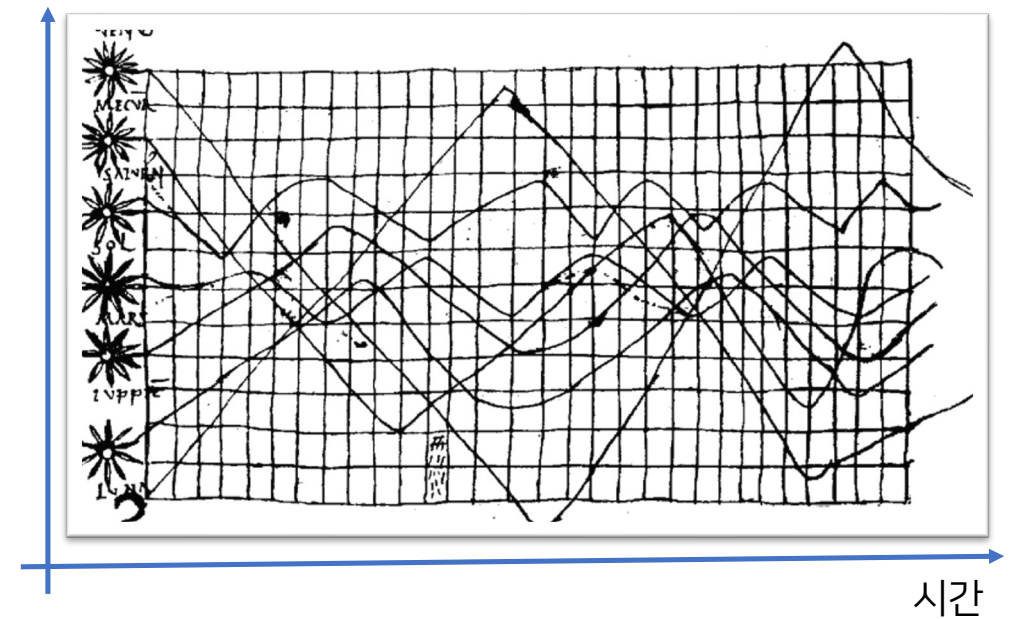
영조 가례도감의궤(1759)반차도 제41~42면-영조의 가마 행렬

정보의 표현

천문도

- 행성의 움직임을 표현하기 위해 사용된 시각화 방법 (약 10세기)
- 시간의 흐름에 따른 천계 위도 표시

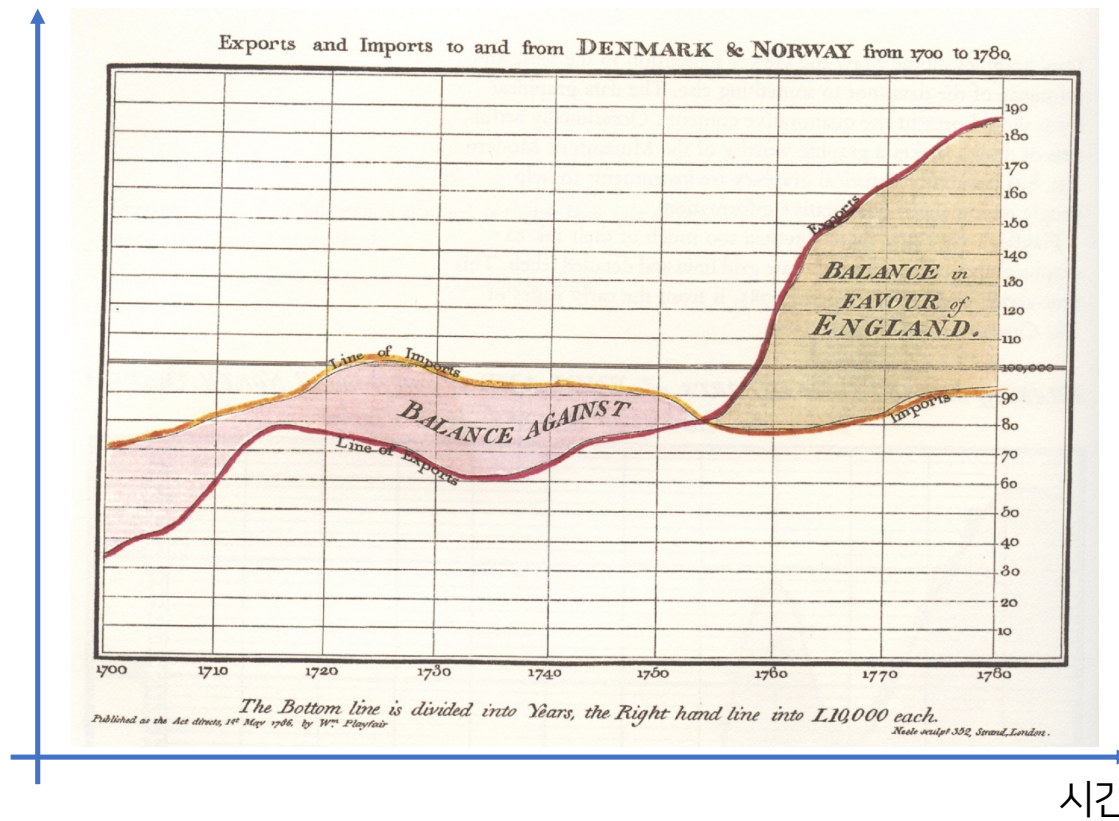
고도



정보의 표현

수출입 시계열 표

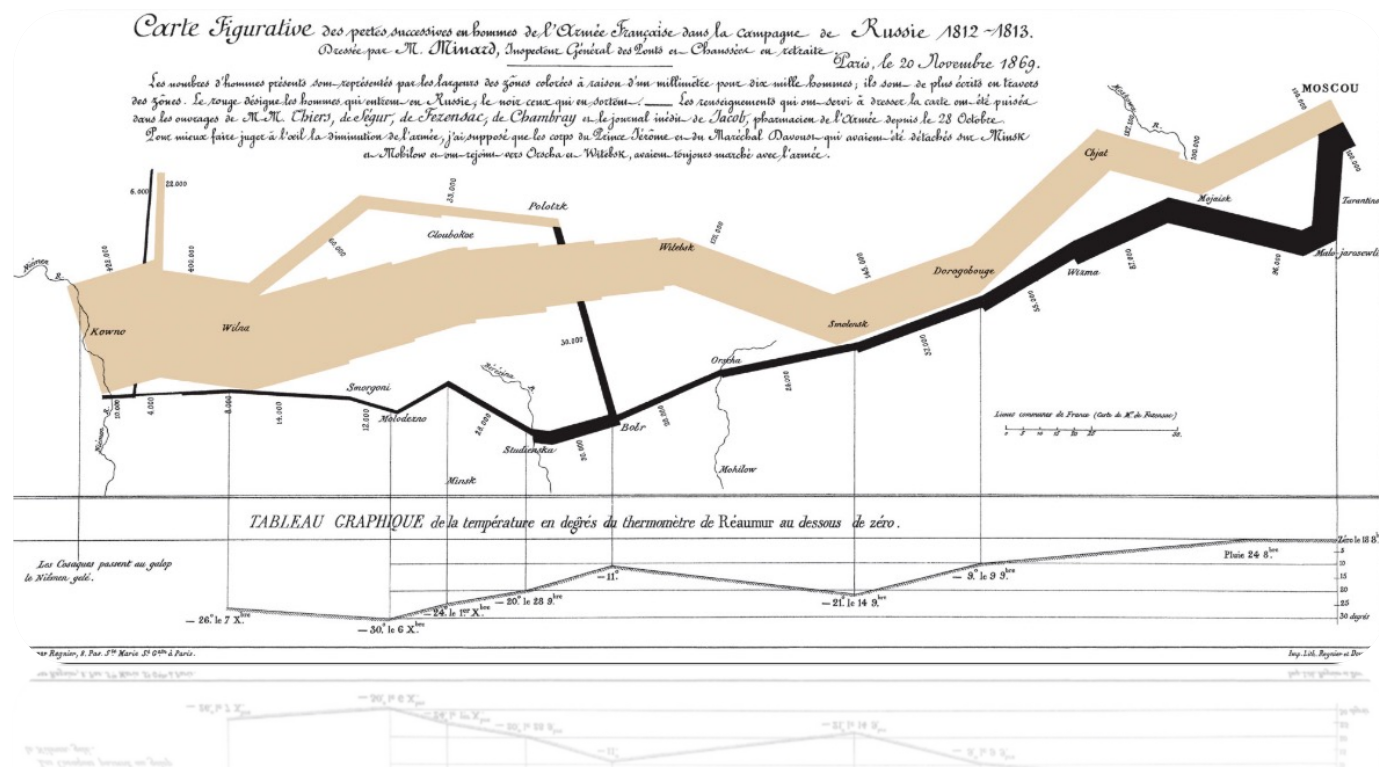
고도



정보의 표현

Minard diagram

1812-1813년에 나폴레옹 군대가 겪었던 손실을 시각화한 그림 (1869)



정보의 표현

나폴레옹의 러시아 원정

- 프랑스 제국시대 유럽에 영향력을 강화하는 시기
- 영국의 경제적 고립을 위해 대륙봉쇄령 실시
- 영국 무역에 크게 의존하고 있었던 러시아의 경제적 타격으로 대륙봉쇄령 위반 (1810)
- 나폴레옹의 러시아 통제를 위한 전쟁 발발 (나폴레옹의 러시아 원정 시작, 1812. 06.24)

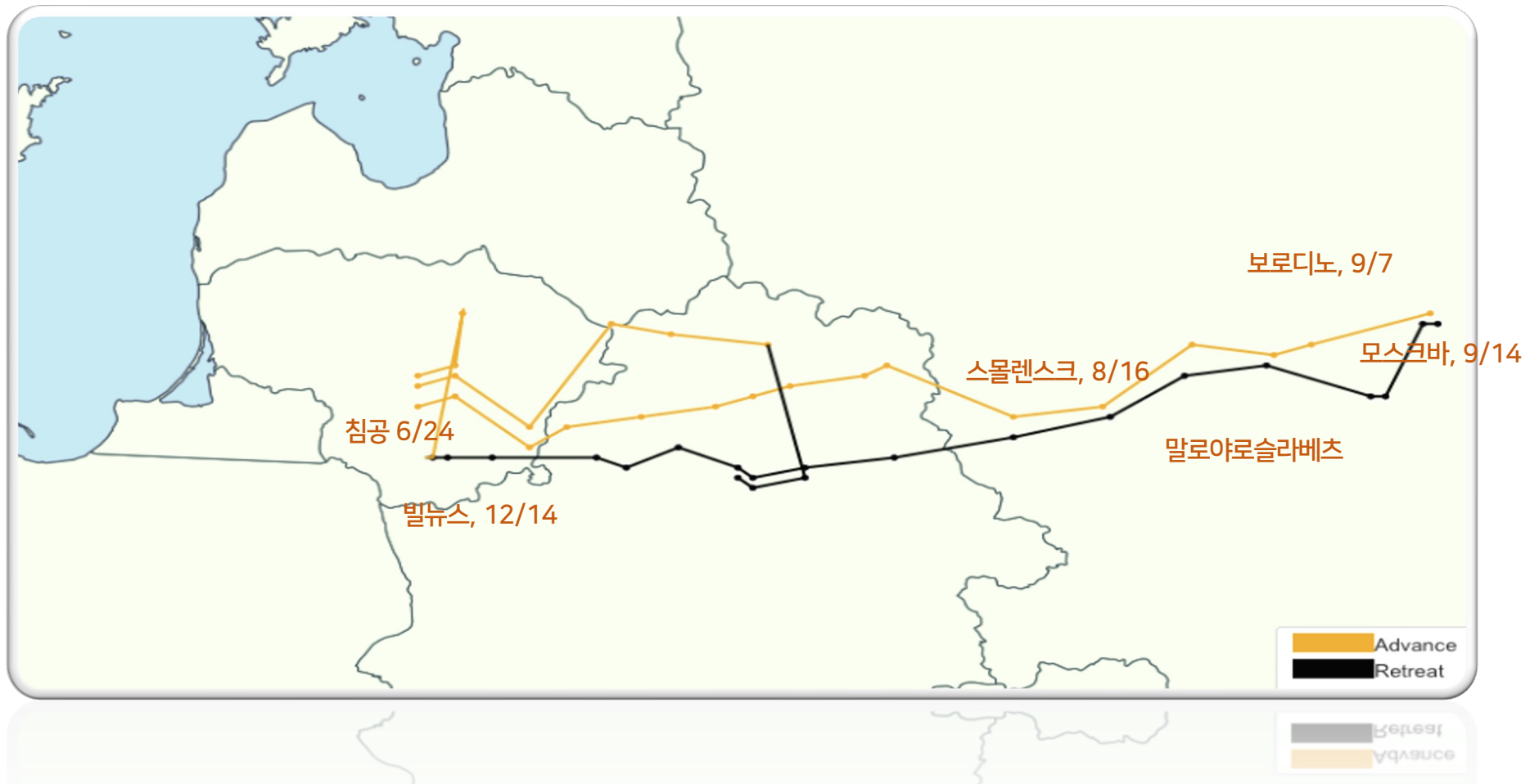
정보의 표현

나폴레옹의 러시아 원정

- (1812. 06. 24) 러시아 원정 개시, 425,000명
- (1812. 08. 16-18) 스몰렌스크 전투, 145,000명
- (1812. 09. 07) 보로디노 전투, 100,000명
- (1812. 09. 14) 나폴레옹의 모스크바 점령
- (1812.10) 나폴레옹의 후퇴와 말로슬라야베츠 전투 96,000명
- (1812.12. 14) 빌뉴스로 후퇴, 8000명

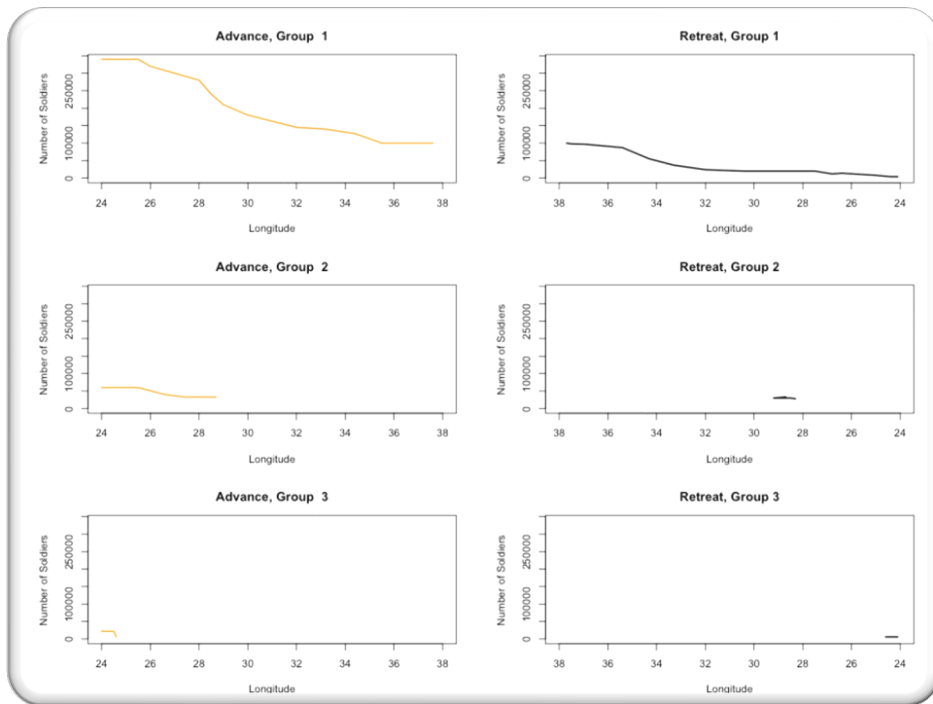
정보의 표현

이동경로

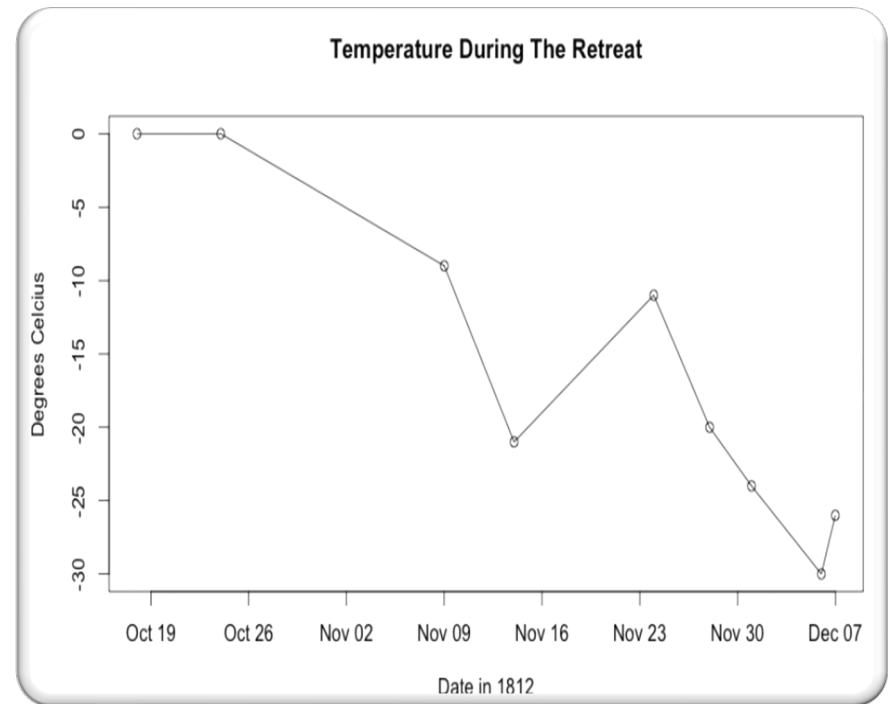


정보의 표현

부대 인원

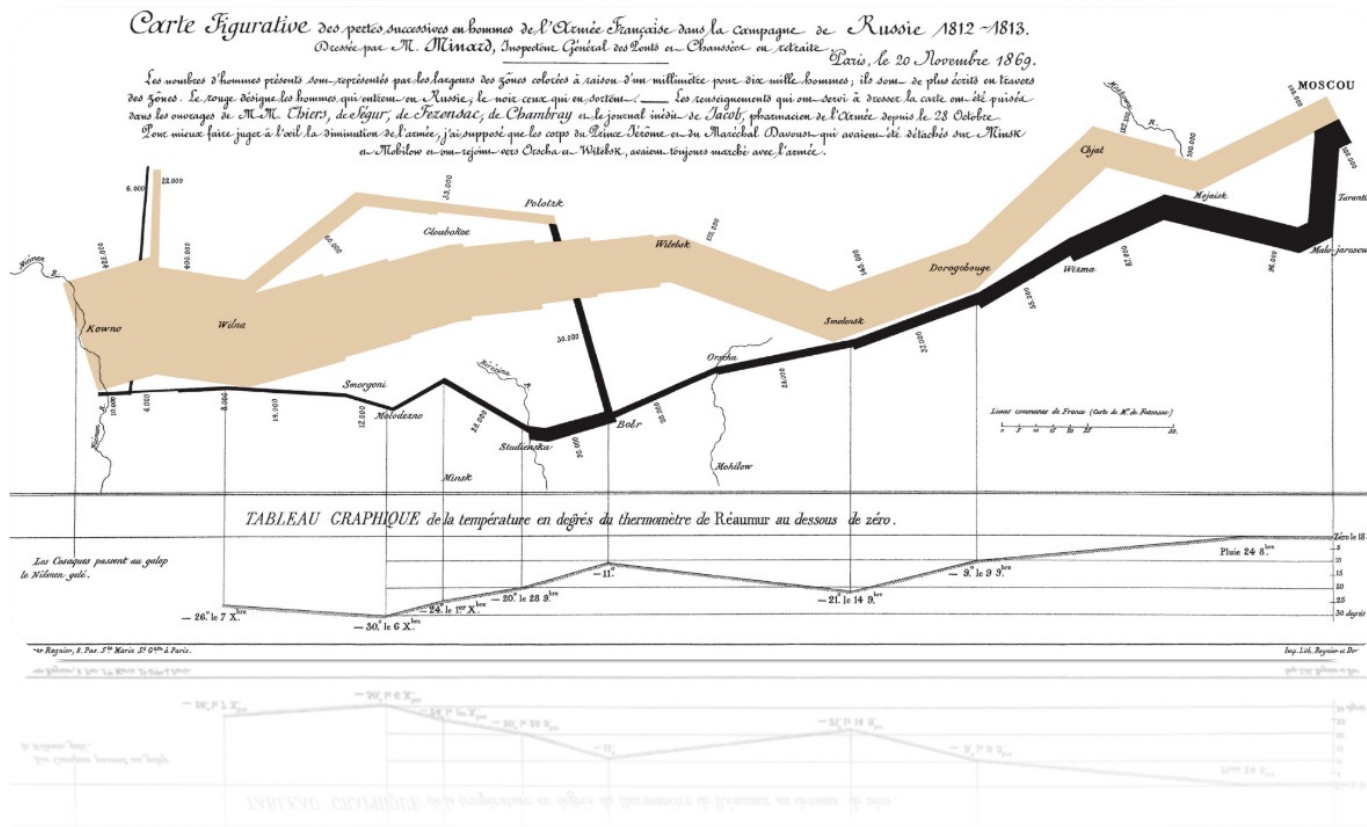


기온



정보의 표현

Minard diagram

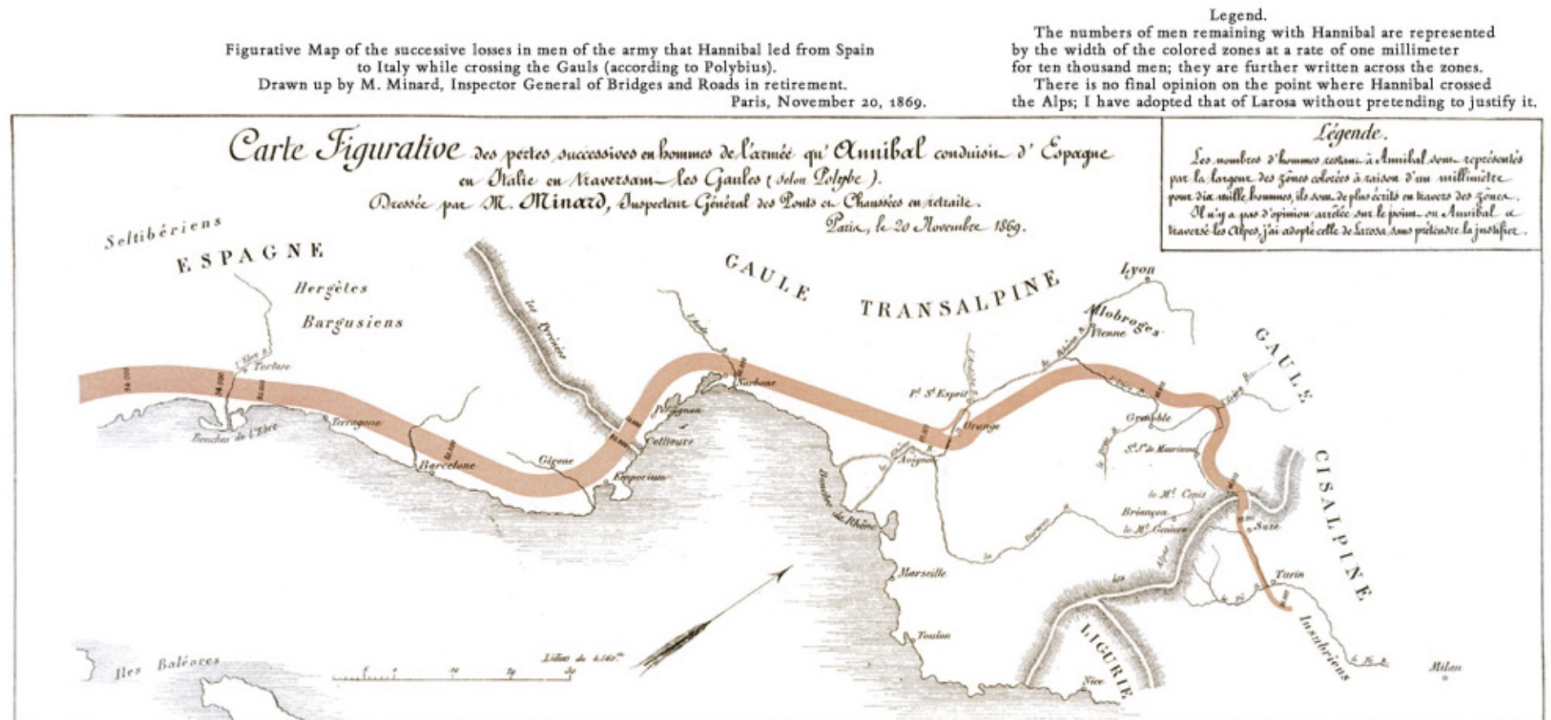


- 군의 크기, 위치, 방향, 온도 변수를 시각화 함

- 진격과 후퇴에 대한 상태 정보를 색깔로 표시하여, 병력의 손실을 쉽게 파악할 수 있음

정보의 표현

- Minard diagram (한니발 전쟁)





Charles Joseph Minard (1781-1870)

- 토목 공학자
- 51개의 시각화 자료를 남김
- 나폴레옹 도표는 1869년 출간



미나드 도표에 대한 단상

“한니발과 나폴레옹, 두 사람이 이끄는 군대는 마치 물처럼 처음에는 세차게 솟아 나왔다가 점차 스러지며 흘러 갔다.

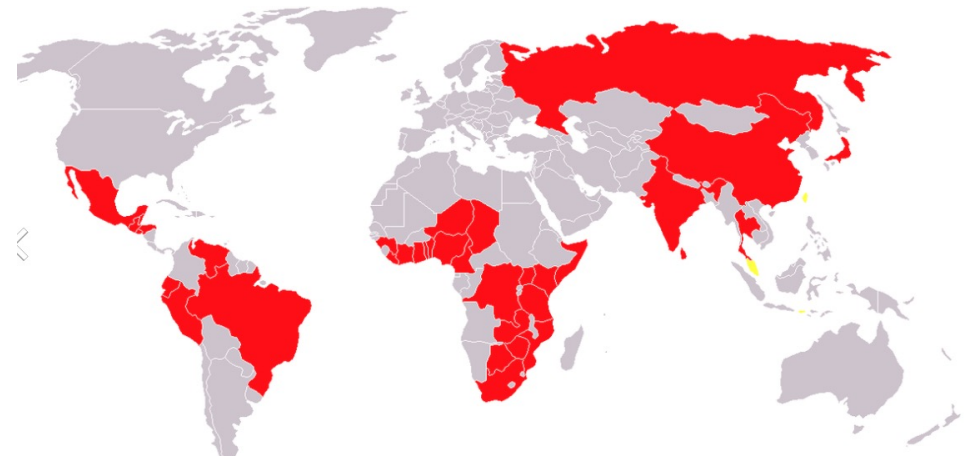
인명 희생에 초점을 맞춘 미나르의 도표는, 군사적 영광을 무자비하게 추구하는 지배자들의 광기가 인간을 얼마나 희생시키는지 잘 보여 준다.”

2. 시각화와 설득

콜레라와 시각화

콜레라

- 전염병의 일종으로 감염되면 설사와 탈수증세를 보이며 치료하지 않는 경우 사망률이 30~50%
- 7차례 세계적인 유행성 콜레라가 발생하였음 (7차 유행, 2009년, 짐바브웨)
- 1853년~1854년 런던에서 발병 (사망자 1만명 이상)
1854~1855년 스페인에서 발병(사망자 23만명 이상)

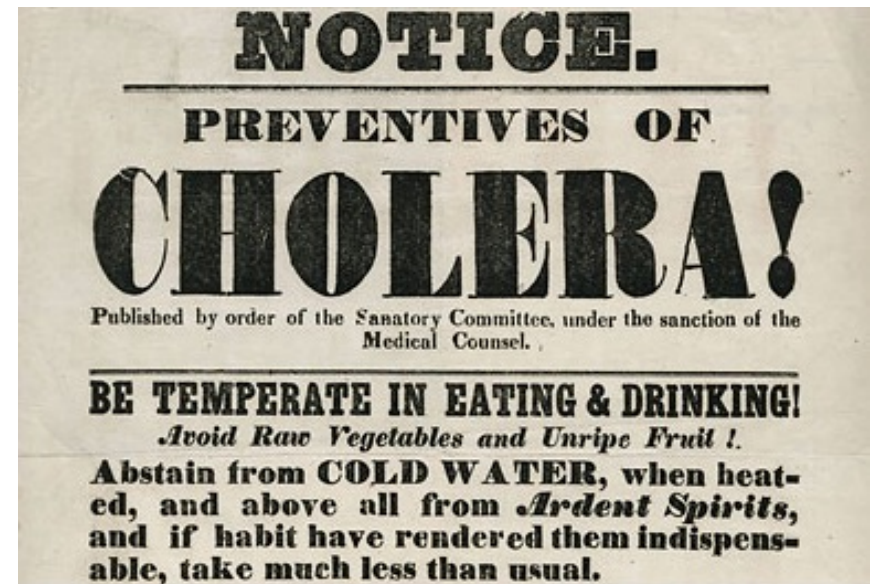


콜레라와 시각화

당시 콜레라에 대한 예방

- 더워도 찬물 마시지 말기
- 독한 증류주 마시지 말기
- 익히지 않은 채소와 덜 익은 과일 먹지 말기

미아즈마 (miasma): 나쁜 공기 혹은 냄새로 전염된다고
믿었기 때문에 콜레라 예방법을 몰랐음



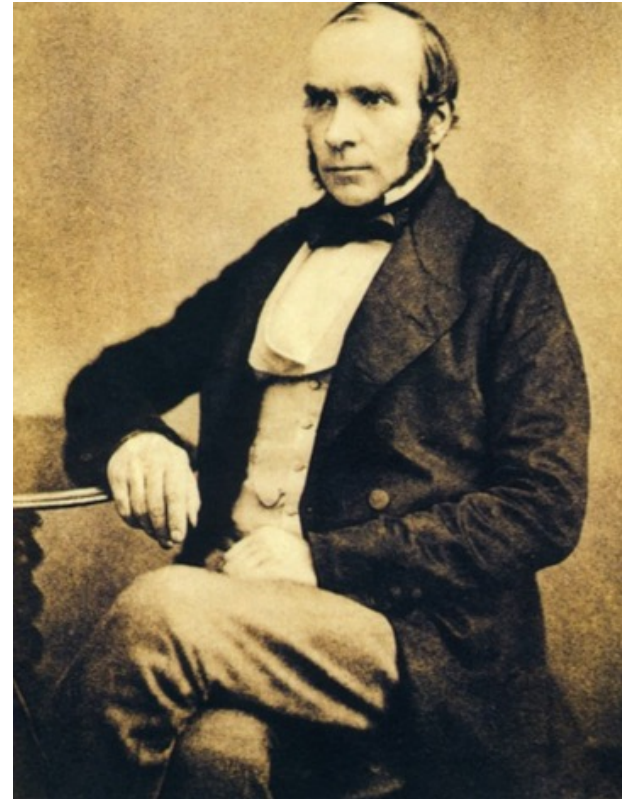
콜레라와 시각화

존 스노우 (1813-1858)

- 영국의 마취통증의학과 의사
- 콜레라의 원인이 오염된 물이라고 믿었음
(소수의견)

1800년대 런던 도시로 상하수체계

- 사설업체가 운용하는 상수도관 이용
- 10개 구역을 나누어 사설업체가 상수도망을 관리
- 템즈강으로 하수방류량 증가로 도시에 악취문제가 심각했음



콜레라와 시각화

콜레라로 인한 사망자 수를 지도에 표시

- 콜레라 사망자가 나온 집을 돌아다니면서 역학조사를 실시하여 지도에 표시
- 물맛 좋다고 소문난(?) 펌프가 문제라고 확인하였음!
(이 펌프의 물을 마신 다른 지역사람도 콜레라에 감염됨)

상수도 펌프

사망자 수



콜레라와 시각화

상수도 펌프의 폐쇄

- 당시 미아즈마 이론하에서는 물을 콜레라의 원인이라 생각하고 조치를 취하기는 어려웠음
- 존 스노우가 주장하는 물펌프의 폐쇄는 당시로서는 받아들이기 힘든 대응방법이었음.
- 존 스노우는 역학조사결과를 토대로 지방 의회에 콜레라와 물공급간의 관련성을 주장하였고, 결국 지방 의회는 상수도 펌프를 폐쇄 결정을 내렸음.



콜레라와 시각화

이후 조치

- 템즈강 하구에서 물을 취수하여 공급한 사우스워크
복스홀 상수도회사의 물을 마신 지역사람들이
타지역의 사람들에 약 14배에 달하는 콜레라
감염률을 보였음
- 이후 런던시의 하수도 정비 사업을 실시하여
공중보건체계를 정비하였고, 이후 콜레라
유행이후에는 물을 반드시 끓여 마시도록
시민들에게 권고하였음



런던 브로드가의 콜레라 펌프

나이팅게일의 장미도표

보건행정가 나이팅게일

- 크림전쟁에서 부상병의 간호체계를 점검
 - 많은 부상병이 전투의 부상이 심각해져 사망하는 것을 목격하고 사망률 원인을 조사하여 정리하였음
 - 나이팅게일 조치
 - 위생상태의 개선
 - 간호지원체계의 강화 필요물자의 요청
- * 객관적인 자료 작성하여 영국의회를 설득하고자 함.



나이팅게일 (1820-1910)

나이팅게일의 장미도표

사망원인표 작성의 원칙

- 다양한 원인들을 나열하는 것을 피하고 주장하고 싶은 방향으로 사망원인 체계를 간단하게 정리함.
- 간호방법의 변화에 따른 효과를 보여주고자 시간 따른에 사망자 변화를 극적으로 보여주고함. 리하였음
- 설득의 대상이 과학자가 아니라 정치인임을 중요하게 생각하였음



나이팅게일의 장미도표

나이팅게일의 장미도표

- 사망원인
파란색: 전염병, 빨간색: 부상악화, 갈색: 다른 원인
- 사망자 수
해당 색깔 영역의 면적
- 시간의 표현
여러 개의 부채꼴을 이어 붙임

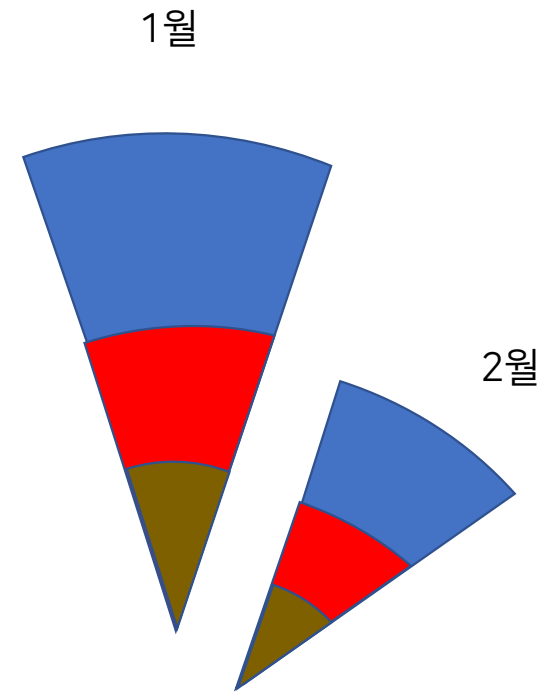
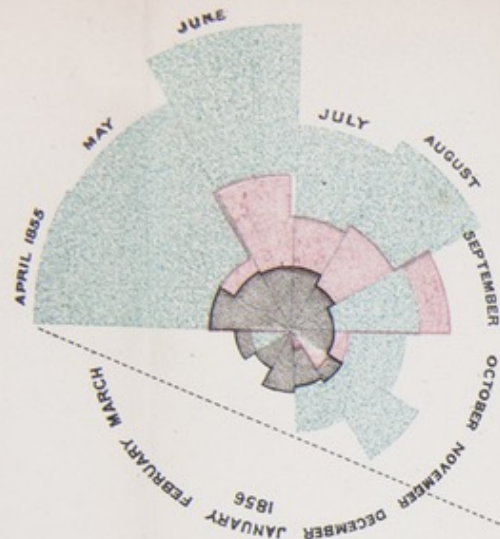
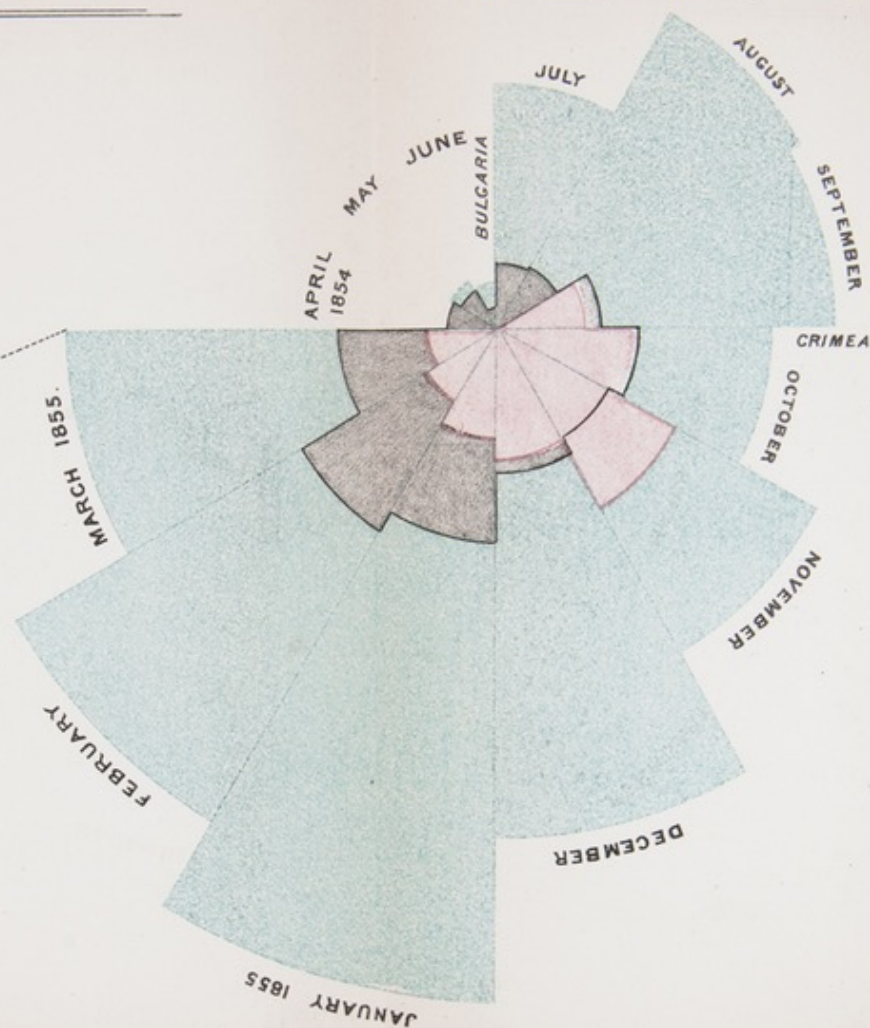


DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

2.
APRIL 1855 TO MARCH 1856.

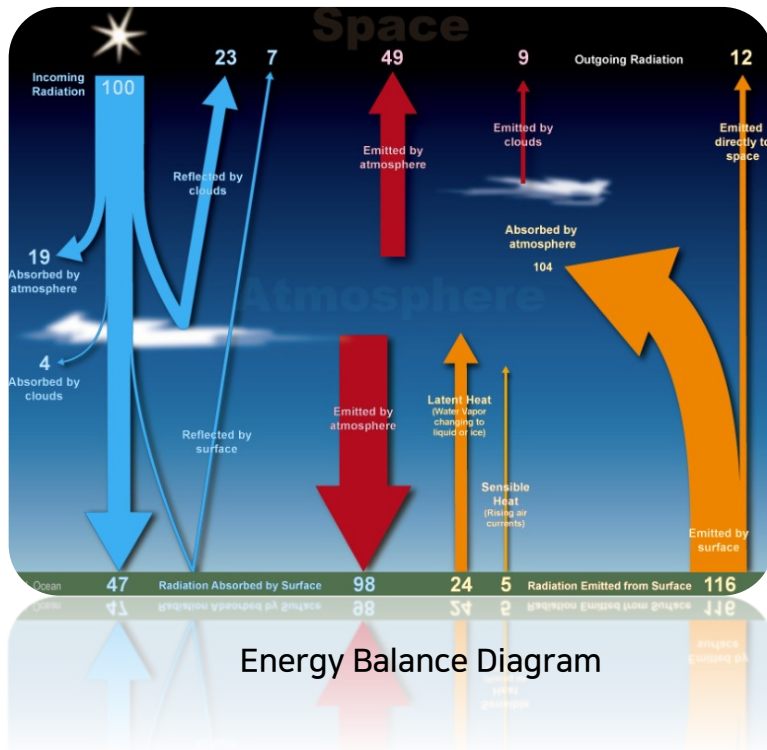


1.
APRIL 1854 TO MARCH 1855.

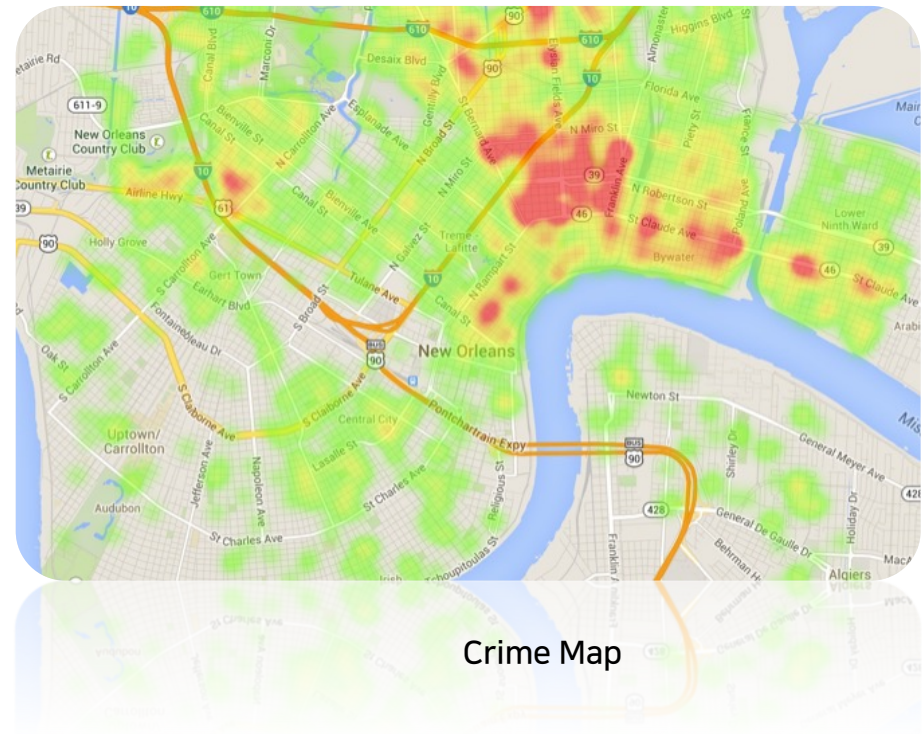


The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov' 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red, in January & February 1855, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

시각화



Energy Balance Diagram



Crime Map

3. 스토리텔링

스토리텔링

스토리텔링

빨간 두건 이야기

- 어린 소녀 빨간 두건은 아프신 할머니에게 드릴 음식을 갖고 숲 속을 지나간다. 사나운 늑대는 빨간 두건을 잡아 먹고 싶지만
...(중략)...이후 사냥꾼이 달려와 늑대를 잡고 배를 갈라 할머니와 빨간 두건을 구한다.



LITTLE RED-CAP

스토리텔링

빨간 두건 이야기

- 빨간 두건(Red Riding Hood: RRH) 집(A) 에서 할머니 집 (B)로 걸어서 0.54마일 이동한다.
- RRH 는 B로 이동중에 늑대를 만났다. (중략).. RRH는 2시 경에 B지점에 도착했고 세가지 질문을 한다.
- 문제점 : 세번째 질문 후에 늑대가 RRH 먹었다.
- 해결책: vendor(사냥꾼) 도끼를 사용한다.
- 예상결과: 할머니와 RRH는 살지만, 늑대는 죽는다.



스토리텔링

빨간 두건 이야기

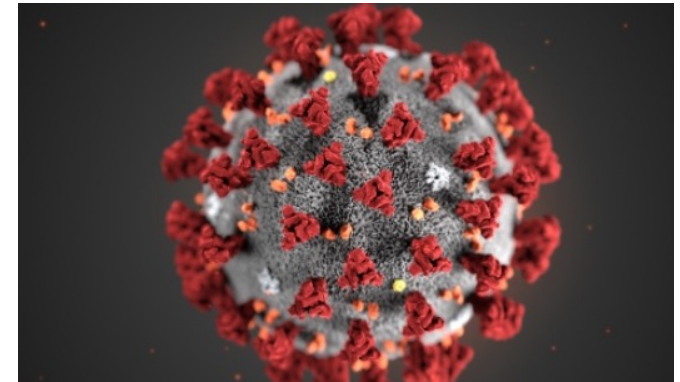
- 데이터와 통계는 어떠한가?
- 규칙에 맞게 나열된 데이터에 감성과 활력을 어떻게 불어 넣을 것인가?

[illegible]

시각화와 스토리텔링

데이터와 통계 이야기

- 무엇을 말하고 싶은가?
- 규칙에 맞게 나열된 데이터에 감성과 활력을 어떻게 불어넣을 것인가?
- 어떻게 이야기를 끌고 갈 것인가?



코로나-19

시각화와 스토리텔링

스토리의 구조 (미국 사례)

1. 미국에서 코로나의 사망자의 증가
2. 코로나 감염이 초기에 대도시를 중심으로 확산되었음
3. 코로나 감염이 지속되면서 감염지역이 소도시 지역으로 확산되고 있음
4. 미국 전역에서 코로나 사망자가 나오고 있는 상태임

시각화와 스토리텔링

스토리의 구조 (미국 사례)

1. 미국에서 코로나의 사망자의 증가

데이터: 일별 코로나 사망자 수를 나타내는 데이터

- 시간에 따라서 코로나 사망자가 어떻게 변하고 있는가를 보여주는 것이 효과적일 것임

시간	사망자 수
2020-03-01	
2020-03-02	
2020-03-03	
...	

시각화와 스토리텔링

스토리의 구조 (미국 사례)

2. 초기의 사망자는 도시를 중심으로 발생하였음

- “도시를 중심으로 발생”한다는 것은 비교의 대상이 필요

도시 vs 도시 외 지역

- 사망자 수, 비율의 선택

- “초기”에서 시간별 구분이 필요함

- 데이터: 코로나 사망자 수를 지역별, 시간별로 정리해야함.

시간	사망자 수	지역
2020-03-01		뉴욕
2020-03-01		히달고
2020-03-01		시카고
...		

시각화와 스토리텔링

스토리의 구조 (미국 사례)

3. 코로나의 감염이 소도시로 확산되고 있음
4. 미국전역에서 사망자가 나오고 있는 상태임
 - "소도시 확산", "미국 전역 " 이라는 말에서 공간적인 의미를 강조할 것인지 결정
 - 데이터: 코로나 사망자 수를 지역별, 시간별로 정리해야 하며, 표기한 지역의 위치정보를 가지고 있어야 함.

시간	사망자 수	지역
2020-01		뉴욕
2020-01		히달고
2020-01		시카고
...		

지역	위도	경도
뉴욕		
히달고		
시카고		

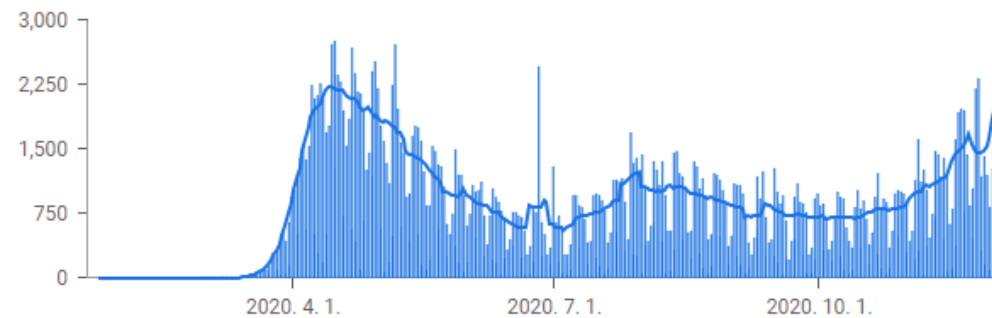
시각화와 스토리텔링

미국에서 코로나의 사망자의 증가

사망자 추세

미국

사망자: ☒ 신규 ☐ 총계

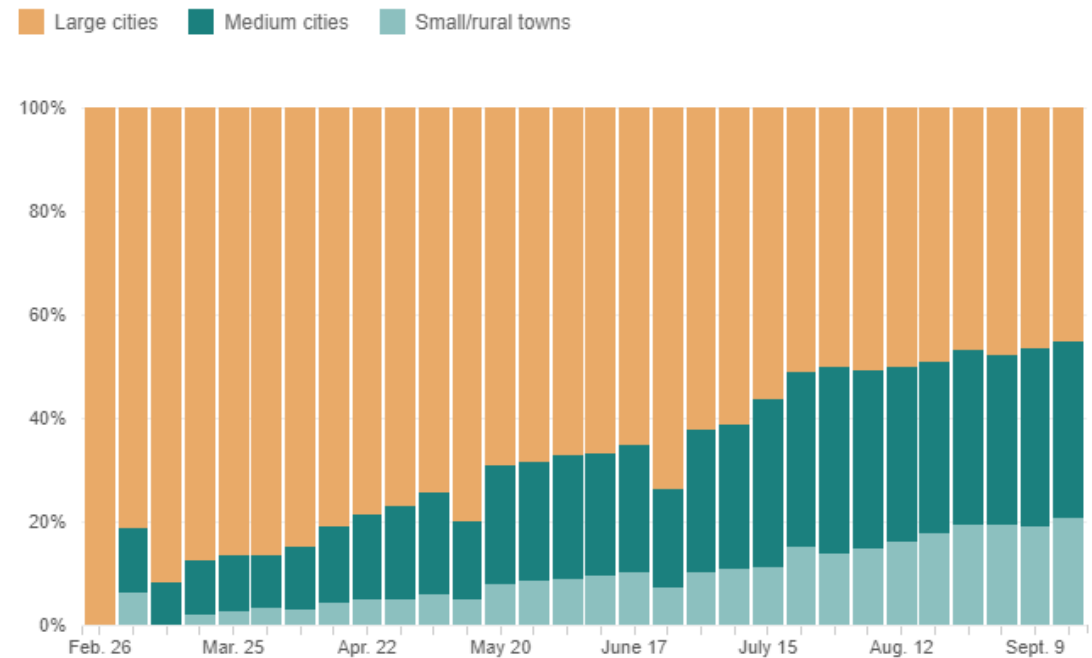


지난 업데이트 시간: 9시간 이내 • 출처: [뉴욕 타임스](#)

시각화와 스토리텔링

코로나 사망자의 도시별 비교

- 시간에 따른 비율을 비교하자
- 도시별로 색깔을 다르게 하여 잘 보이도록 그려보자

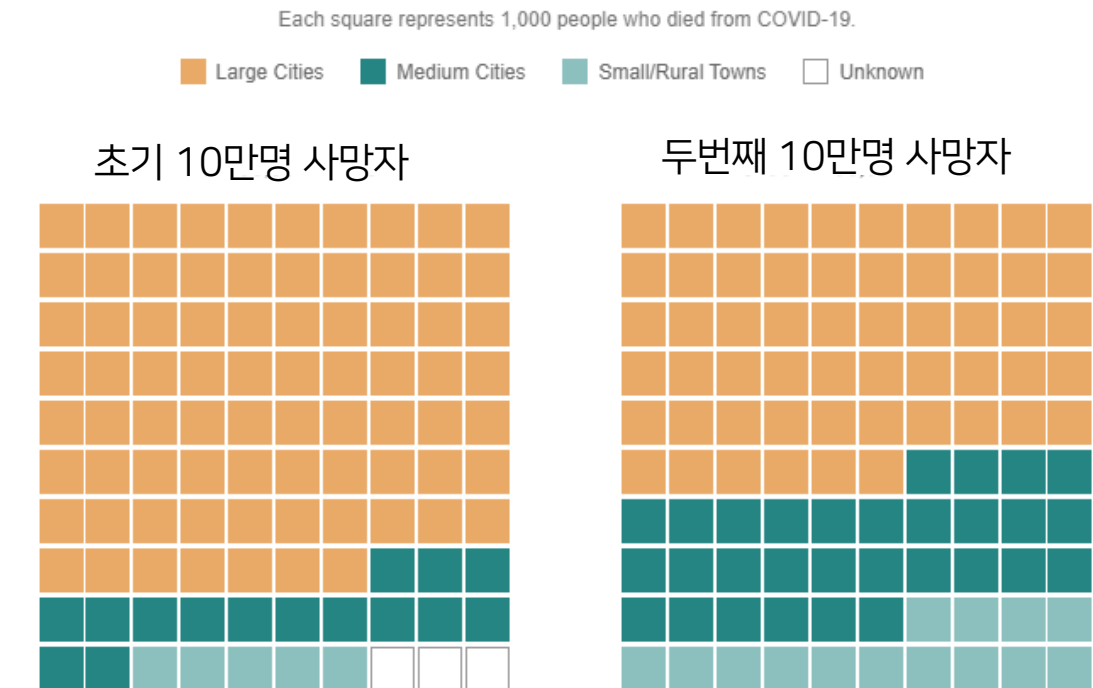


Source: Center for Systems Science and Engineering at Johns Hopkins University, Centers for Disease Control and Prevention
Credit: Jess Eng/NPR

시각화와 스토리텔링

코로나 사망자의 도시별 비교

- 초기 10만명의 사망자와 이후 10만명의 사망자를 나누어 비교하자
- 작은 사각형 하나가 1,000명을 나타냄
- 사각형의 색깔을 다르게 표시



Notes

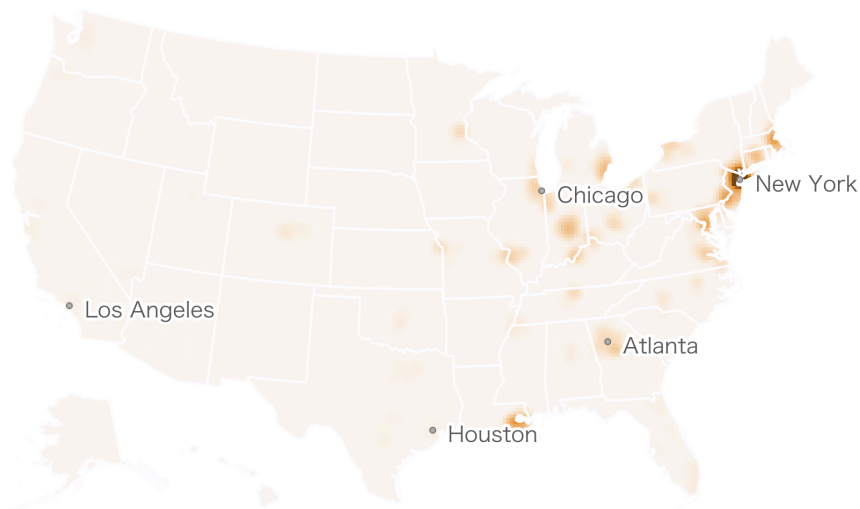
Deaths marked as unknown were uncategorized as of the time the U.S. crossed 100,000 deaths.

Source: Center for Systems Science and Engineering at Johns Hopkins University, Centers for Disease Control and Prevention

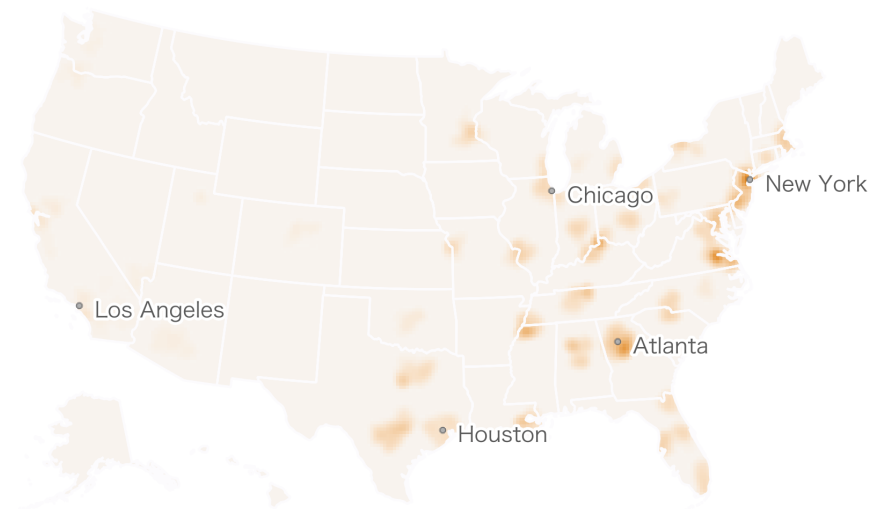
Credit: Sean McMinn/NPR

시각화와 스토리텔링

코로나 사망자의 지역 비교 (대도시)



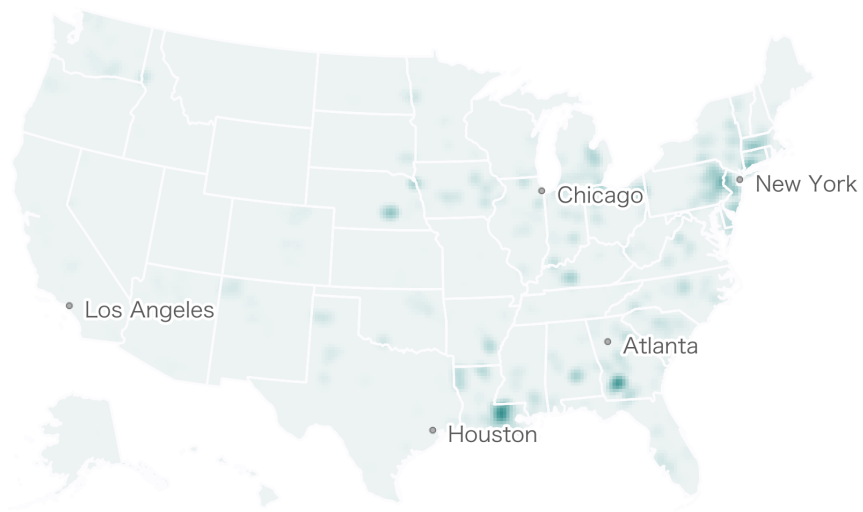
초기 단계



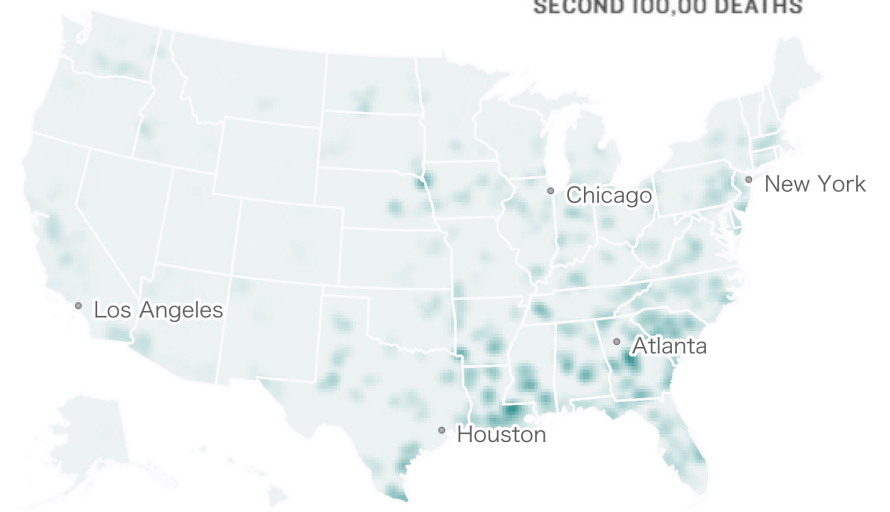
이후 단계

시각화와 스토리텔링

코로나 사망자의 지역 비교 (중도시)



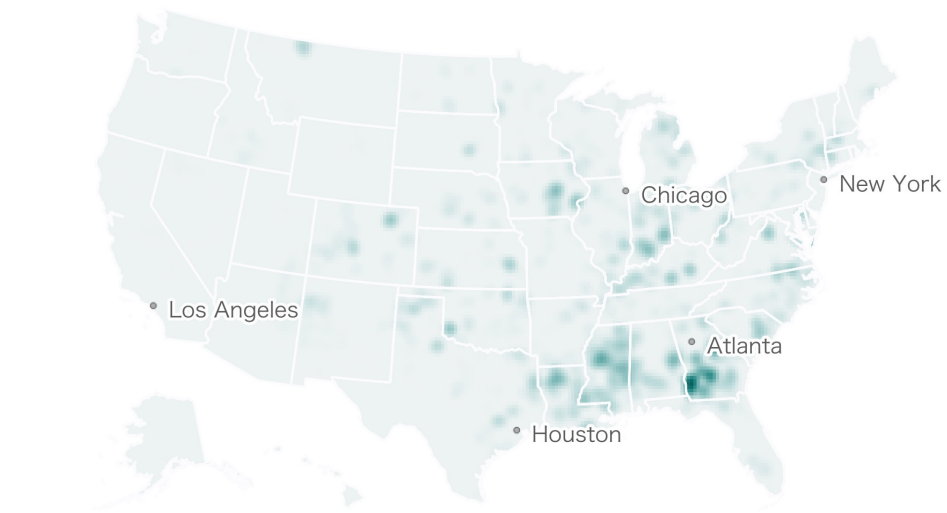
초기 단계



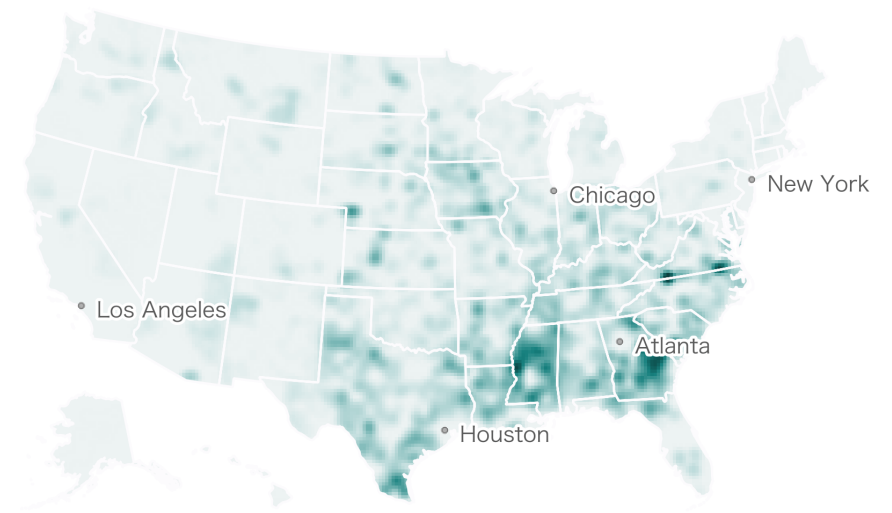
이후 단계

시각화와 스토리텔링

코로나 사망자의 지역 비교 (소도시)



초기 단계



이후 단계

시각화와 스토리텔링

여성의 임신과 출산

1. 여성의 임신은 어떤 결과로 이어지는가?
2. 나이에 따른 임신의 결과는 달라지는가?
3. 결혼한 이후 첫 아기를 몇 년만에 가지게 되는가?

여성의 임신과 출산

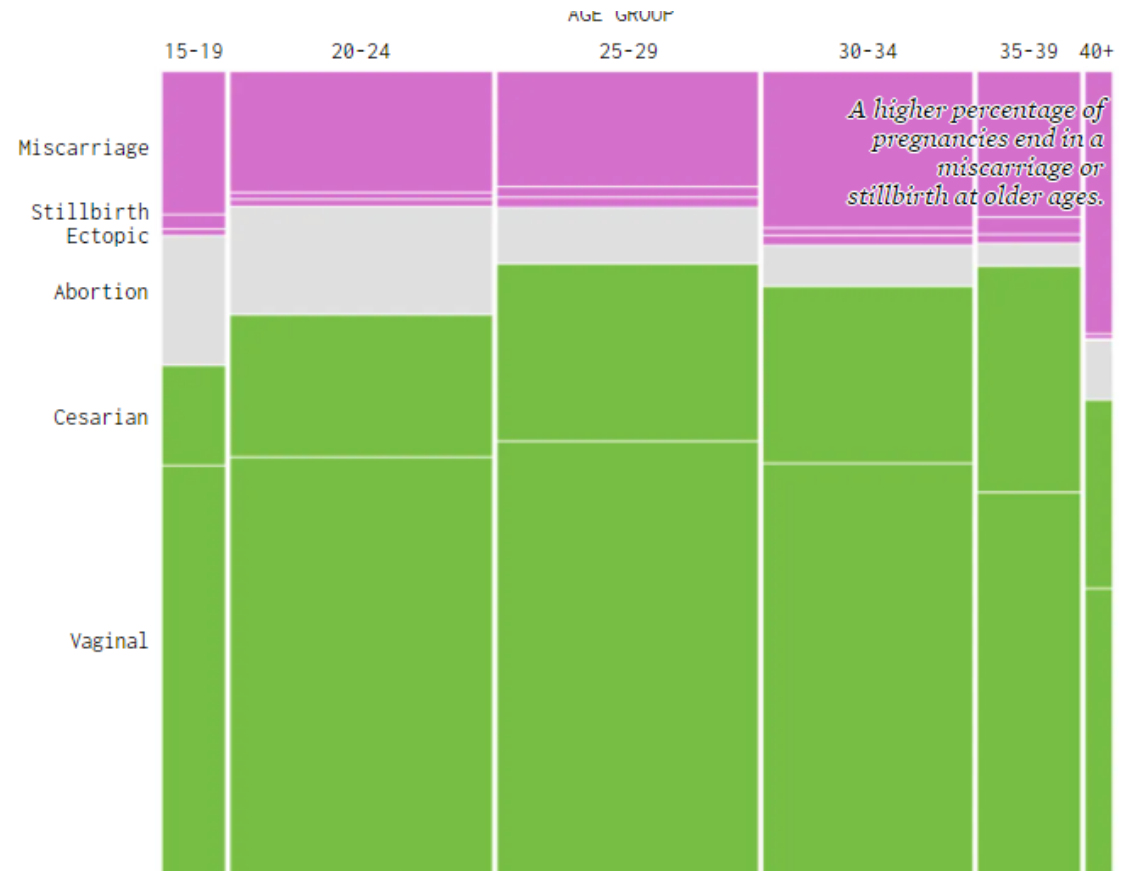
1. 여성의 임신은 어떤 결과로 이어지는가?
 - 임신의 결과로는 유산, 사산, 낙태, 제왕절개, 자연분만으로 나누어 볼 수 있다.
2. 나이에 따른 임신의 결과는 달라지는가?
 - 나이에 따라 임신의 결과는 달라지며 나이가 많은 여성의 경우 임신이 출산으로 이어질 확률이 낮아진다.

시각화와 스토리텔링

스토리텔링

가로축: 여성의 연령대로 사각형 가로길이의 비율이 연령대 사람의 비중임

세로축: 임신의 결과로 세로축의 비중이 비율을 나타냄
(녹색: 출산, 그외: 출산실패)



시각화와 스토리텔링

스토리텔링

가로축: 여성의 연령대로 사각형 가로길이의 비율이 연령대 사람의 비중임

세로축: 임신의 결과로 세로축의 비중이 비율을 나타냄
(녹색: 출산, 그외: 출산실패)



시각화와 스토리텔링

스토리텔링

가로축: 여성의 연령대로 사각형 가로길이의 비율이 연령대 사람의 비중임

세로축: 임신의 결과로 세로축의 비중이 비율을 나타냄

(녹색: 출산, 그외: 출산실패)



시각화와 스토리텔링

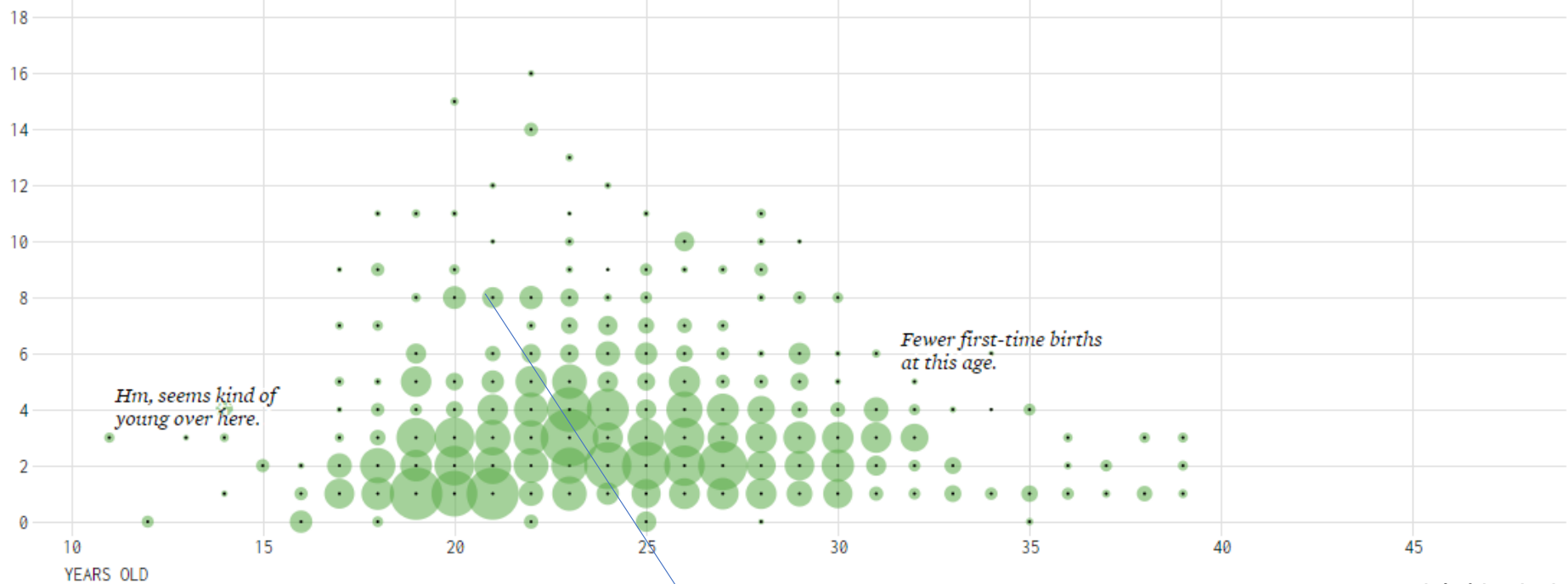
첫 출생시간

SELECT SEX

Female

Male

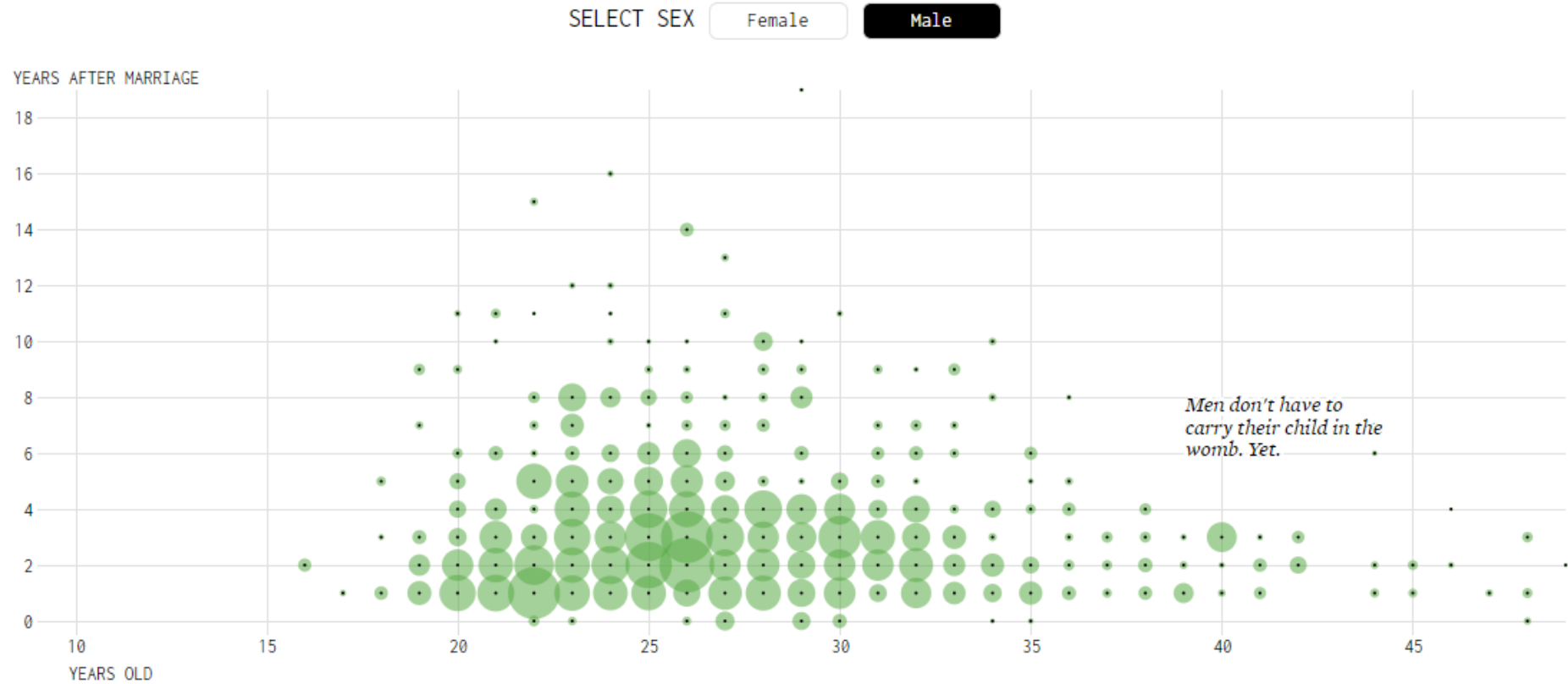
YEARS AFTER MARRIAGE



21세에 결혼한 여성, 8년후 출산
원의 크기가 비중을 나타냄

결혼한 나이

시각화와 스토리텔링



시각화와 스토리텔링

어떤 이야기들을 다루나?

소설

1. 주제: 사랑, 운명, 가족 등
2. 서술방식: 1인칭 시점, 3인칭 전지적 작가시점, 3인칭 관찰자 시점
3. 구성 3요소: 인물, 사건, 배경

어떤 이야기들을 다루나?

데이터 시각화

1. 주제:
2. 서술방식
 - 시간에 따른 변화
 - 순서에 따른 나열
 - 관계의 표현
 - 위치의 표시
3. 시각화 요소: 점, 선, 면, 색

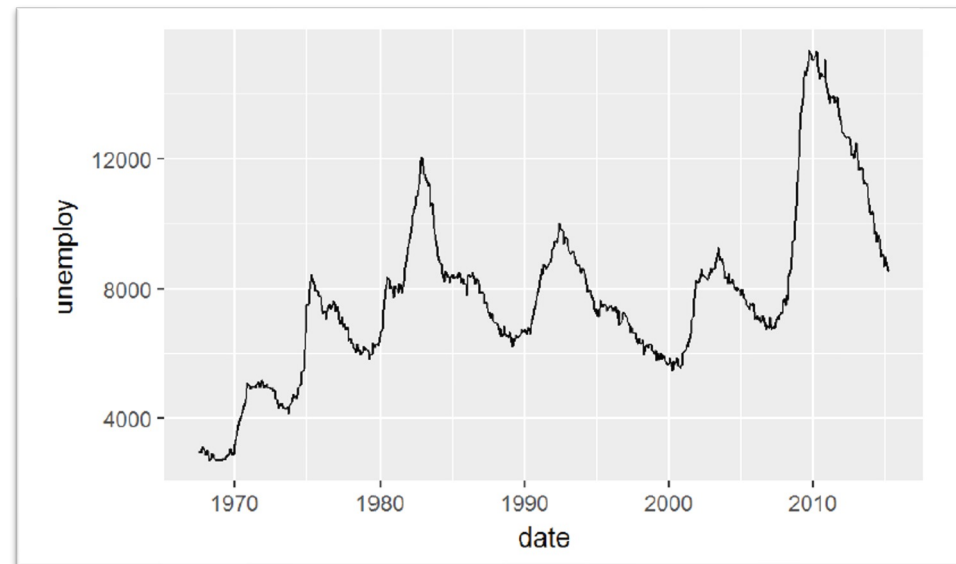
시간에 따른 변화

시계열자료

시간의 흐름에 따라서 자료의 값, 혹은
관심 있는 값의 변화를 보여주는 것이
중요함

- 만약 실업자의 수를 보여주고 싶다면
세로축을 실업자의 수를 보여줄 수 있음

값의 변화



시간의 흐름

순서에 따른 나열

순위자료

순위정보의 시각화: 순서가 중요한 정보가 되며, 순위가 매겨진 부가적인 정보를 함께 제공할 수 있음

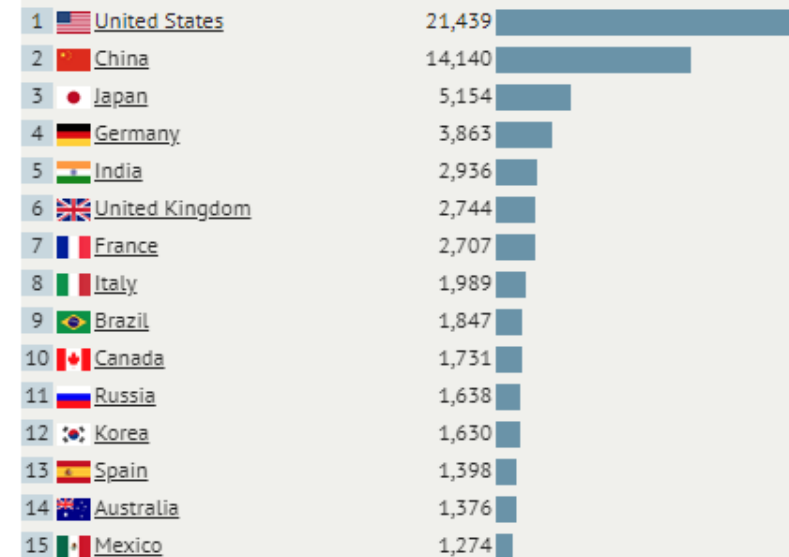
Gross domestic product 2018

	Ranking	Economy	(millions of US dollars)
USA	1	United States	20,494,100
CHN	2	China	13,608,152
JPN	3	Japan	4,970,916
DEU	4	Germany	3,996,759
GBR	5	United Kingdom	2,825,208
FRA	6	France	2,777,535
IND	7	India	2,726,323
ITA	8	Italy	2,073,902
BRA	9	Brazil	1,868,626
CAN	10	Canada	1,709,327
RUS	11	Russian Federation	1,657,554 ^a
KOR	12	Korea, Rep.	1,619,424
AUS	13	Australia	1,432,195
ESP	14	Spain	1,426,189
MEX	15	Mexico	1,223,809

랭킹

GDP in Current Prices

Billion USD



부가정보

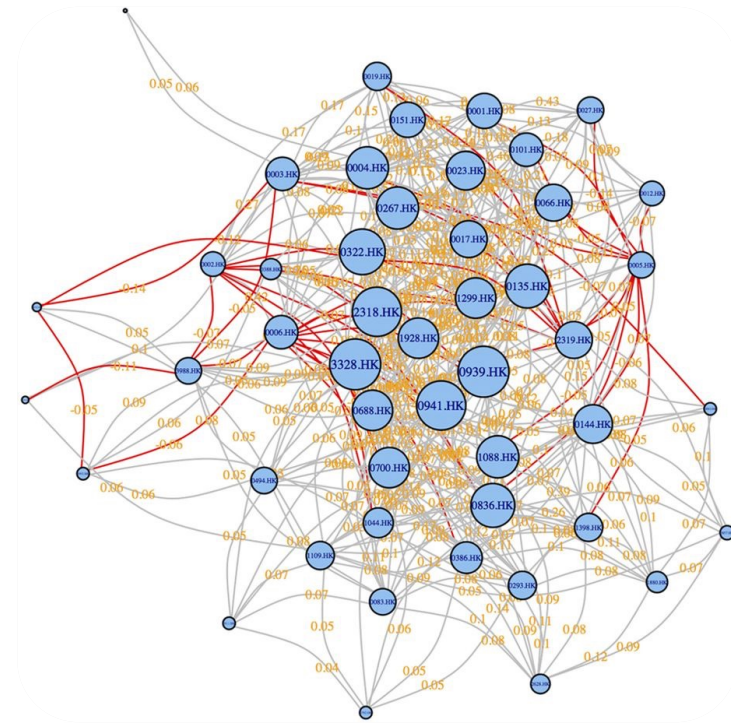
관계의 시각화

연결관계

변량, 데이터 간의 상관성 혹은 연결관계를 보여주고자 함

- 노드 (원): 데이터
- 연결된 선: 관계
 - * 연결 유무
 - * 선의 특성

예) 개별 주가의 동반 상승 및 하락



Network visualization of 48 components of Hang Sang Index ,
Scientific Reports volume7, Article number: 41379 (2017)

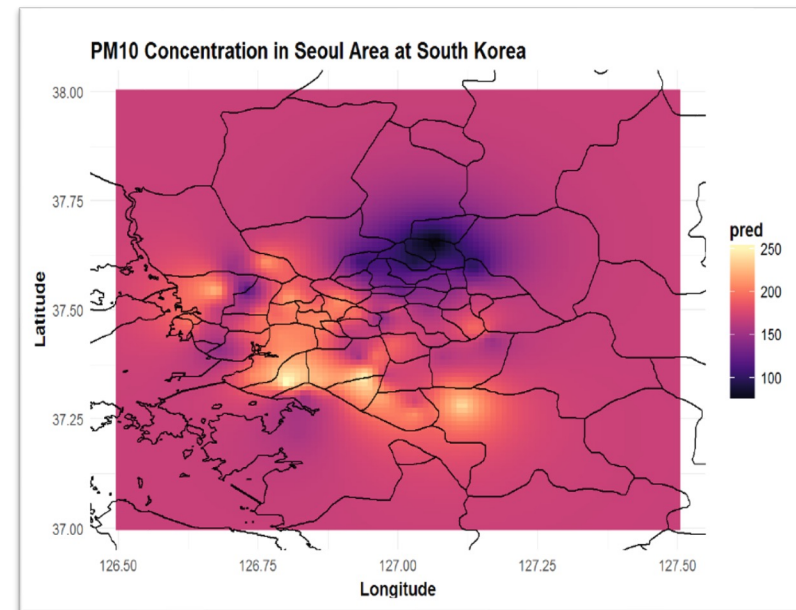
위치의 시각화

연결관계

공간위에서 양적 자료를 시각화

- 지도위에서 작은사각형으로 위치를 표시
- 사각형의 색을 이용하여 양적정보를 표시

예) 서울지역 미세먼지(PM10)의 농도



시각화와 스토리텔링

좋은 그래프를 그린다는 것은...

1. 스토리가 있다는 것, 해야할 말이 있고 그것에 줄거리가 있다는 것임
2. 하고 싶은 말이 있다면 데이터로 그것을 표현할 수 있어야 함.
3. 이야기에 생명을 불어 넣는 작업으로써 시각화가 사용될 수 있음

시각화의 구성요소를 다룰수 있어야 함

4. 시각화의 요소

시각화와 스토리텔링

시각정보 객체

1. 점 (point)

- 주어진 축 위에서 위치를 통해 자료의 값을 표현할 수 있음

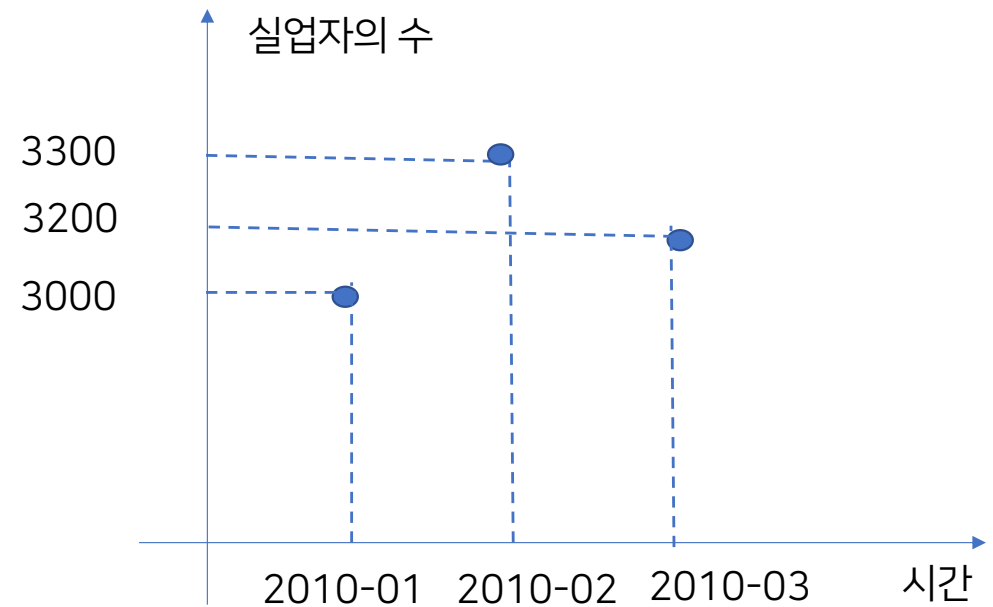
- 시간에 따른 실업자 수의 변화

(시간, 실업자 수)

(2010-01, 3000)

(2010-02, 3500)

(2010-03, 3200)



시각화와 스토리텔링

시각정보 객체

1. 점 (point)

- 주어진 축 위에서 위치를 통해 자료의 값을 표현할 수 있음

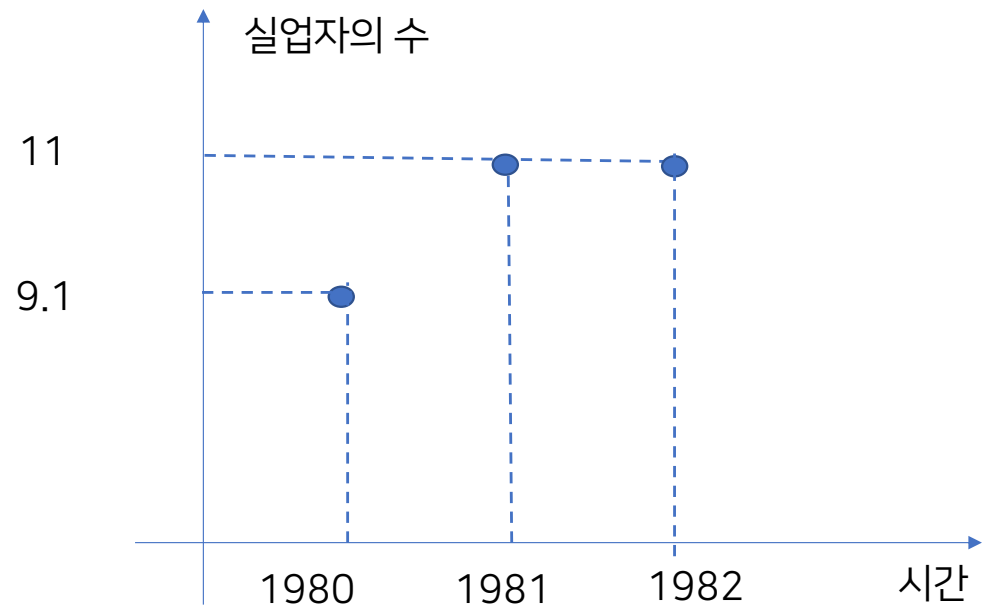
- 핫도그 먹기 대회 결과

(연도, 핫도그 수)

(1980, 9.1)

(1981, 11)

(1982, 11)



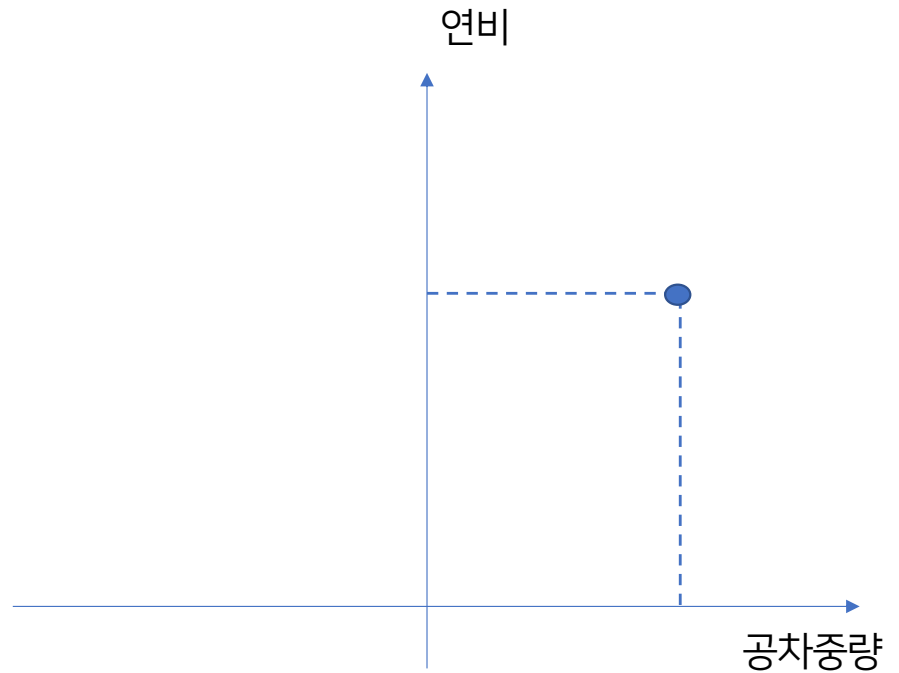
시각화와 스토리텔링

시각정보 객체

1. 점 (point)

- 주어진 축 위에서 위치를 통해 자료의 값을 표현할 수 있음

- 두 변수간의 관계
(공차중량, 연비)

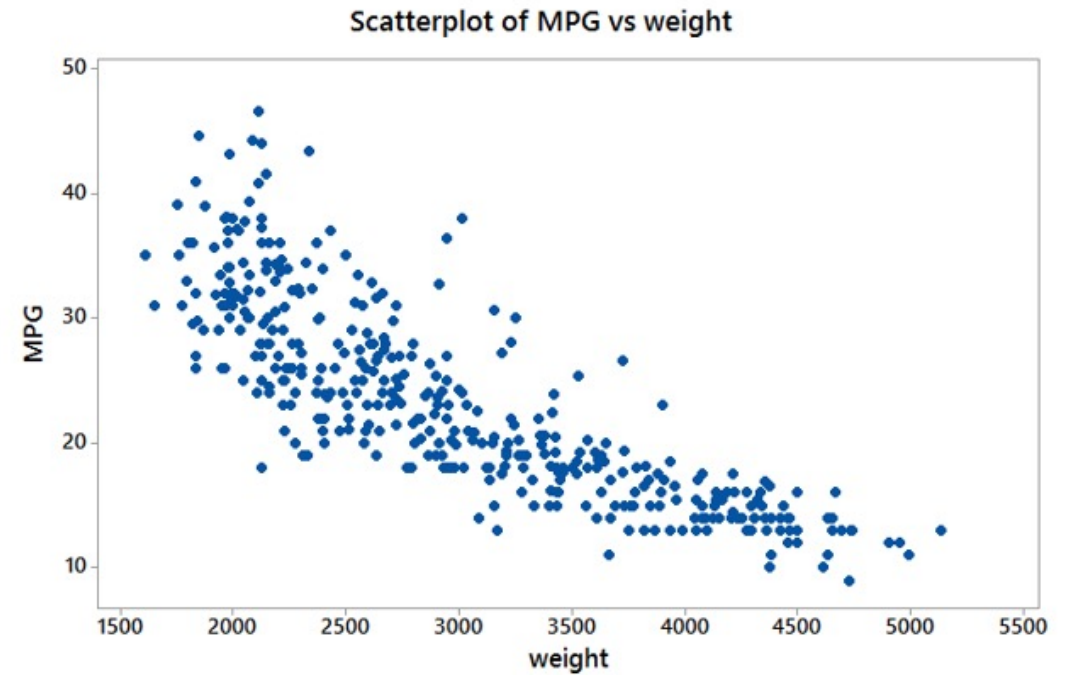


시각화와 스토리텔링

시각정보 객체

1. 점 (point)

- 산점도는 점을 이용한 대표적인 시각화 방법
- 점 하나가 하나의 관측치(차량 1대)를 의미
- 두 변수 (공차중량, 연비) 간의 관계를 확인할 수 있음



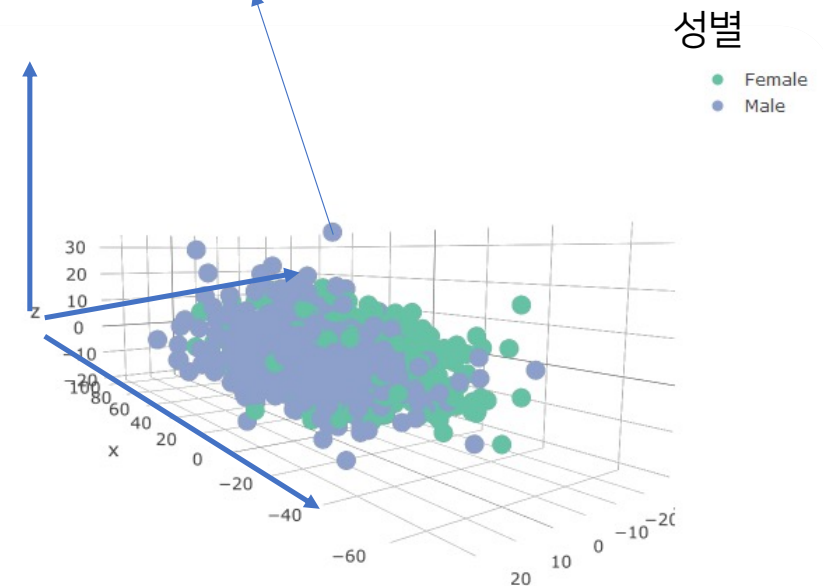
시각화와 스토리텔링

시각정보 객체

1. 점 (point)

- 2차원 평면에 객체를 표시하는 것이 일반적이라 차원이 높은 자료를 표시하는 것에 한계가 있음
- (성별, 키, 몸무게, 연령) 의 자료를 표시하기 위해서는 4개의 축이 필요함
- 좌표축외 다른 시각 요소를 이용하여 데이터를 표현함

(키, 몸무게, 연령) 좌표로 표시
색상을 통해 성별을 구분



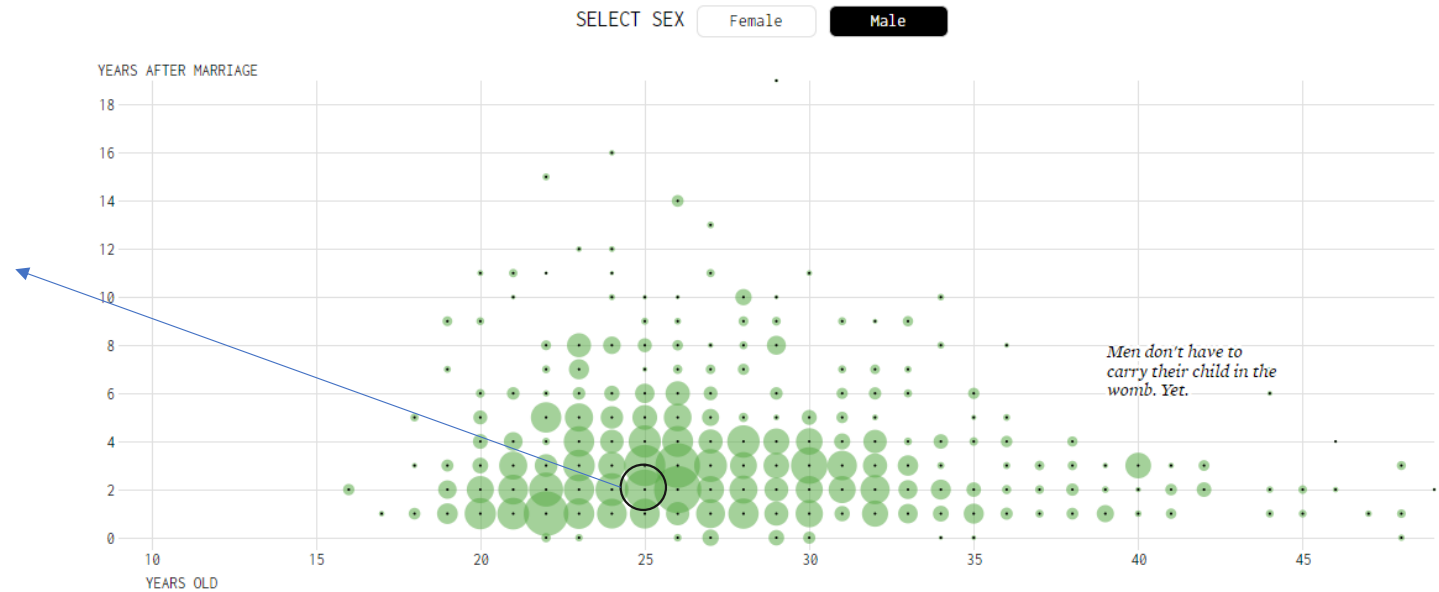
시각화와 스토리텔링

시각정보 객체

1. 점 (point)

- 점의 속성을 이용해서 추가적인 정보를 표시하기도 함
(점의 크기, 점의 색상)

원의 크기: 25세 결혼하여
2년만에 첫아기를 가진
남성그룹의 수

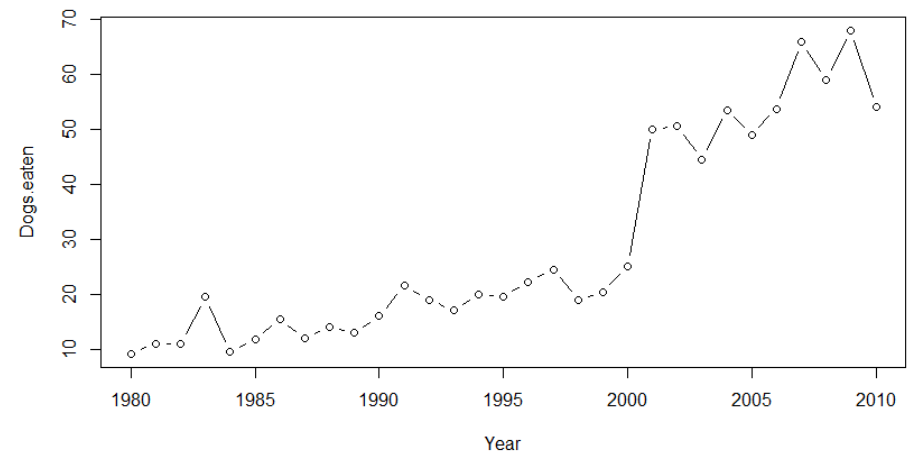
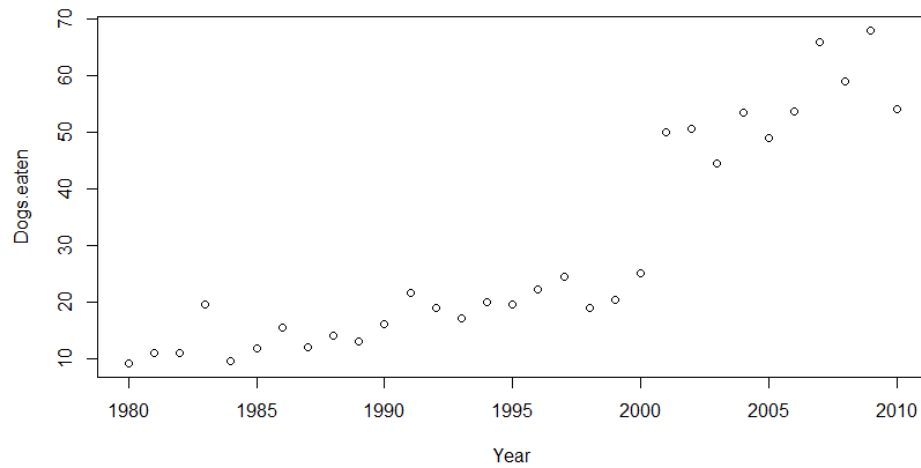


시각화와 스토리텔링

시각정보 객체

2. 선 (line)

- 선: 연결을 통해 값의 변화나 두 개 이상의 관계를 표현

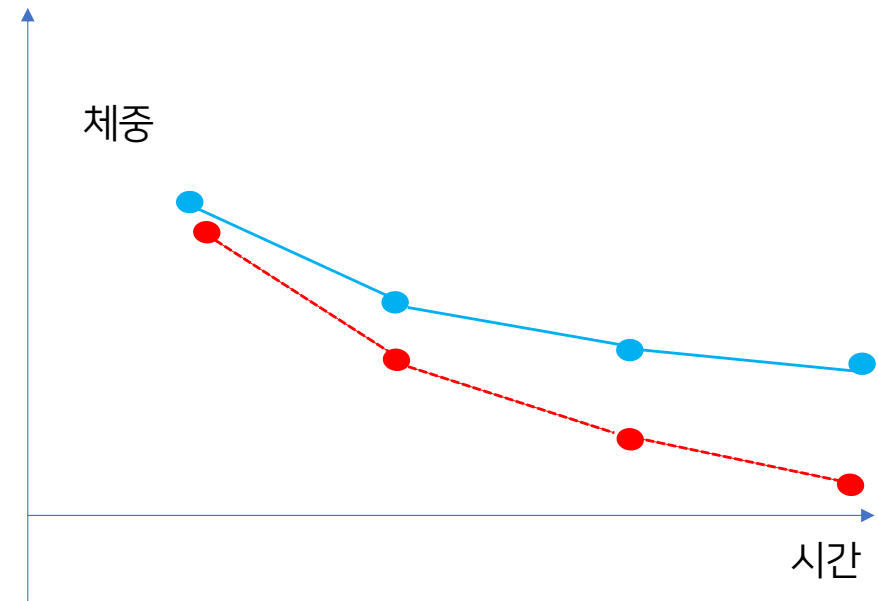


시각화와 스토리텔링

시각정보 객체

2. 선 (line)

- 선의 속성을 변경하여 두 개이상의 계열을 표시할 수 있음



시각화와 스토리텔링

시각정보 객체

2. 선 (line)

- 다차원 점을 선을 연결하여 하나의 객인 선으로 표시할 수 있음

가로축: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species

세로축: 각 변수의 값

색상: Species

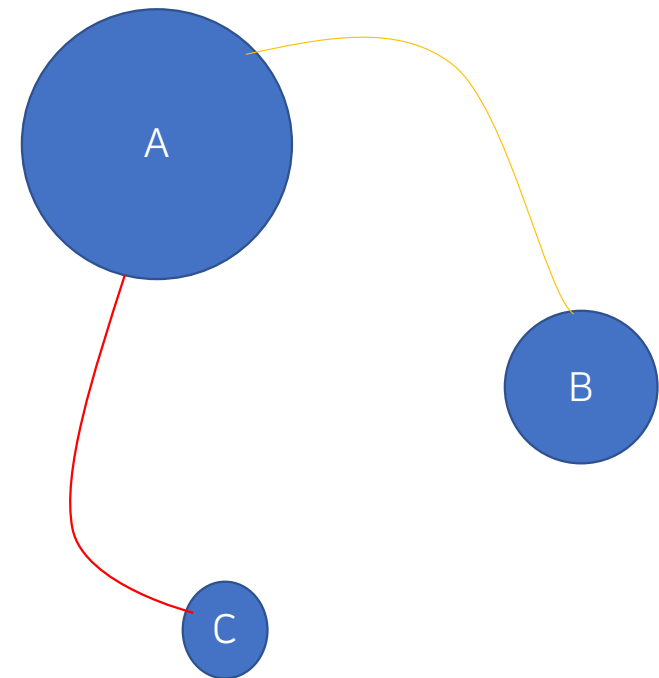


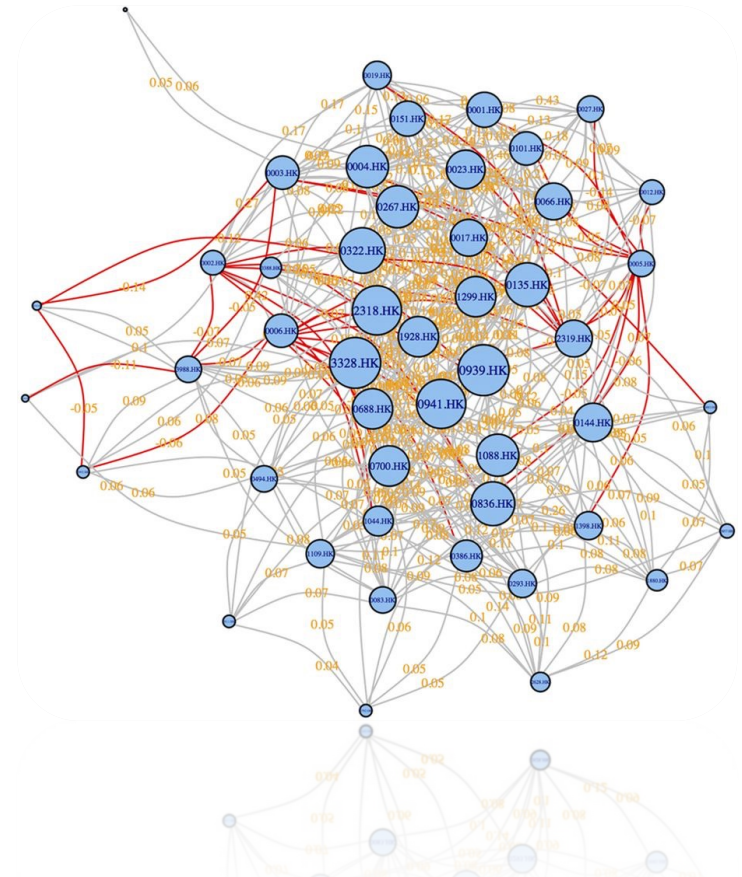
시각화와 스토리텔링

시각정보 객체

2. 선 (line)

- 점(원): 개별주식 하나를 나타냄, 점의 위치는 의미없음
- 점(원)의 크기: 주식의 거래량
- 연결된 선: 연결된 두개의 주식 가격이 유의미한 상관관계가 있음을 표현
- 선의 색깔: 노란색을 가격이 같은 방향으로, 빨간색은 서로 가격이 반대로 움직임을 의미함



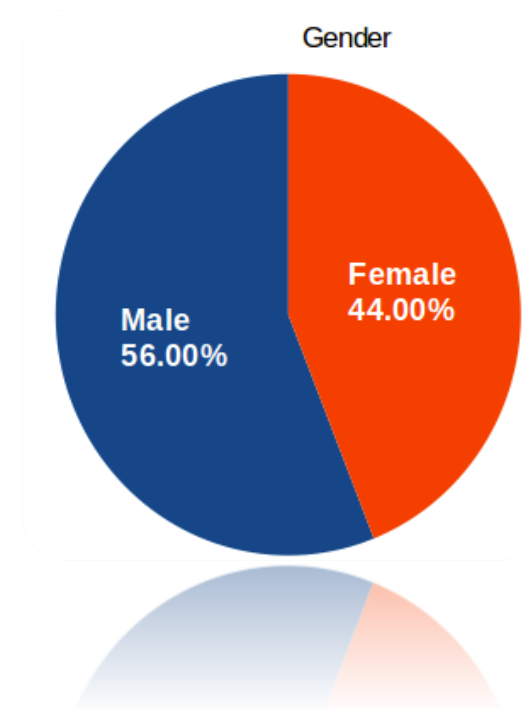


시각화와 스토리텔링

시각정보 객체

3. 면 (region)

- 양적자료에서 양, 크기, 비율을 표시



시각화와 스토리텔링

시각정보 객체

3. 색상

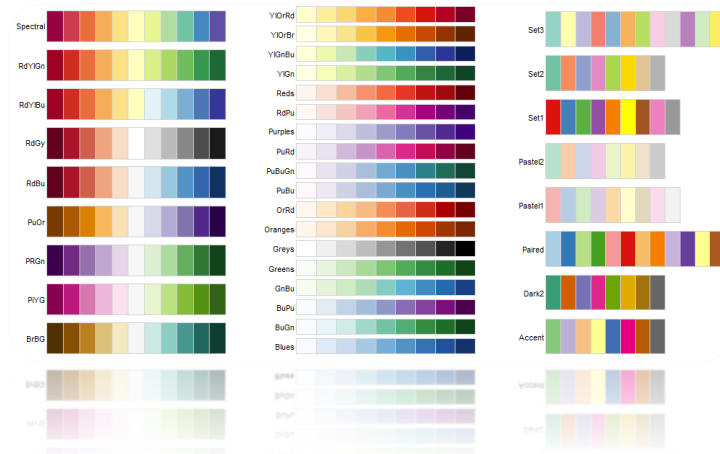
- 시각요소의 속성으로 양적 정보를 효과적으로 전달할 수 있음

- 팔레트: 색상의 순서형 데이터

- 팔레트의 생성: 색상의 생성 체계의 이용

RGB color system

HSV color system

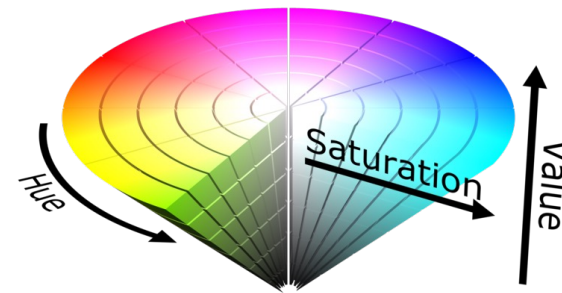


시각화와 스토리텔링

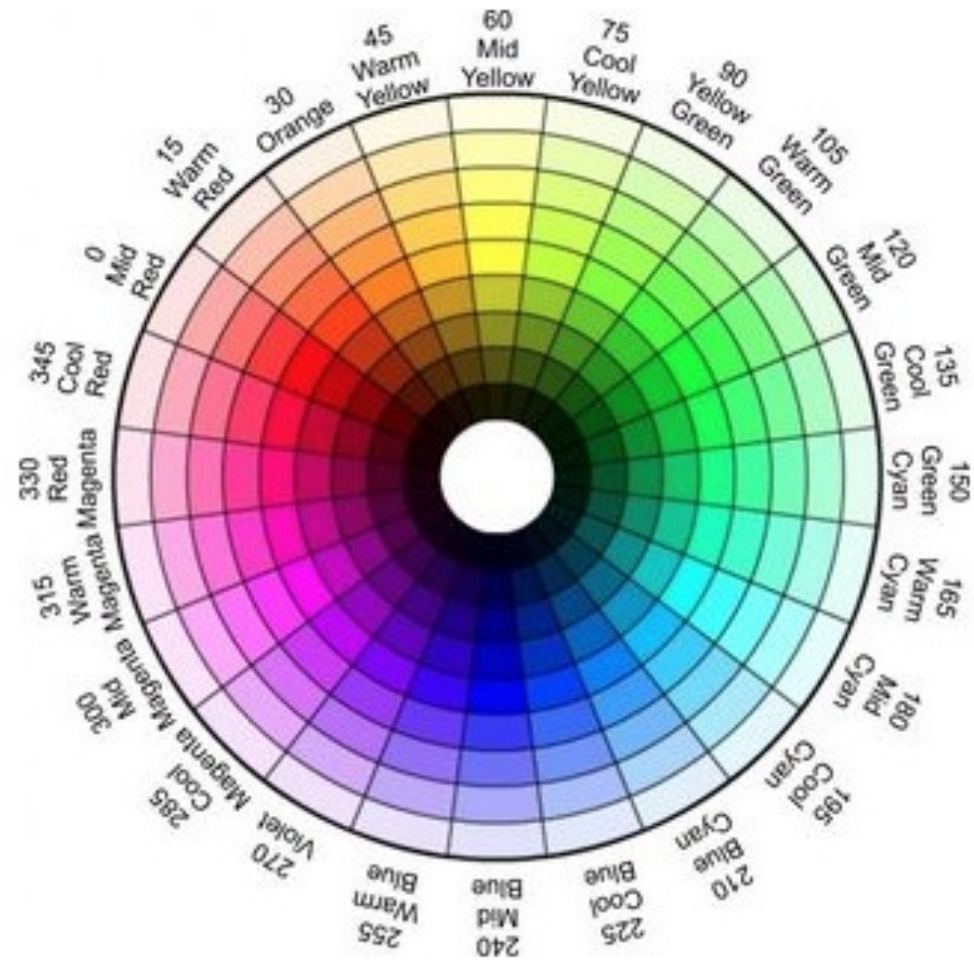
RGB

- Red(00~FF: $256 = 2^8$ scale),
Green(00~FF)
Blue(00~FF)
Opacity(00~FF)
- 총 2^{32} scale의 색상 (32bit color system)

HSV



시각화와 스토리텔링



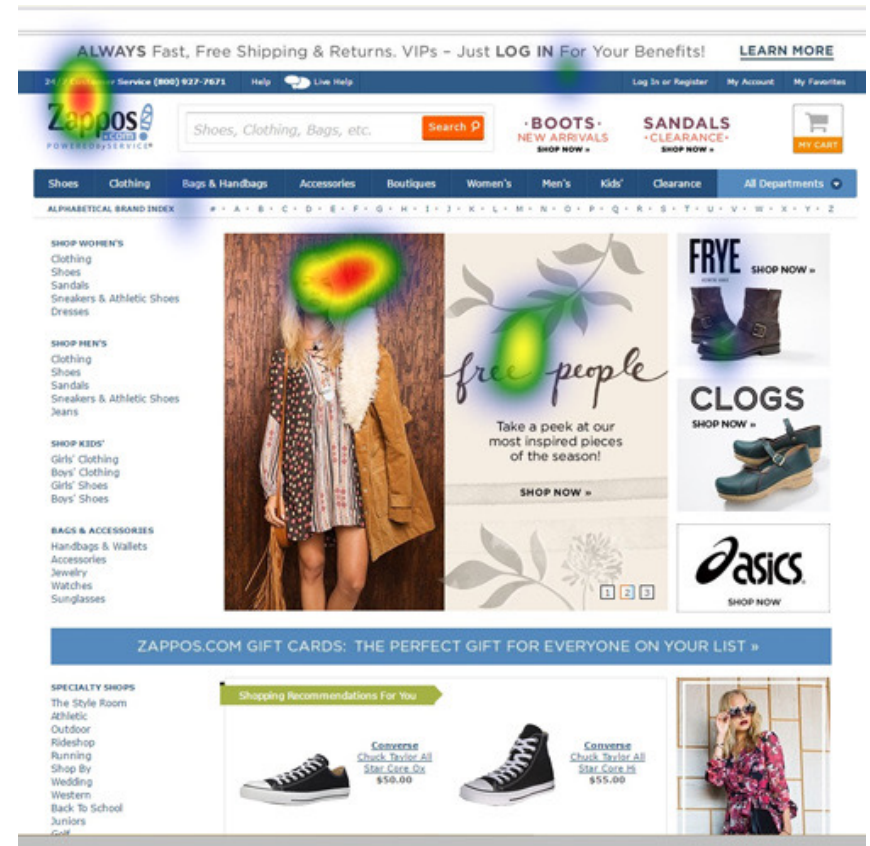
시각화와 스토리텔링

- 특정 공간에서 머무른 시간의 길이를 색상으로 표시함
- 녹색에서 빨간색으로 가는 팔레트를 선택해서 머무른 시간이 길수록 빨간색 쪽 색상으로 표시함



Eye tracker 착용 예시

<https://brainsigns.com/es/science/s2/technologies/eyetracker>

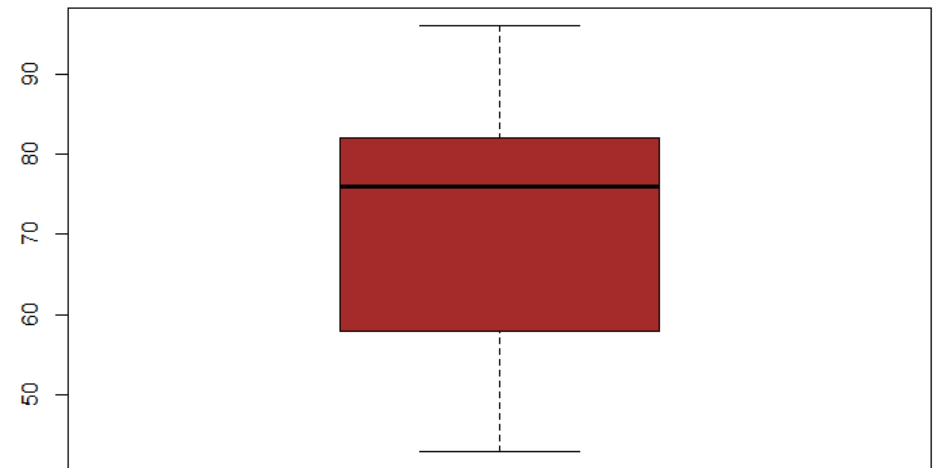


통계량과 시각화

시각화 도구

1. 박스플롯 (boxplot)

- 양적자료의 중요한 정보를 요약하여 표시하는 시각화 도구
- 가운데 선: 자료의 중심
- 상자 윗변, 아랫변: 상위 25%, 75%
(상자의 크기가 분산의 의미를 가짐)
- 최상단, 최하단 선분: 극치 판정 기준

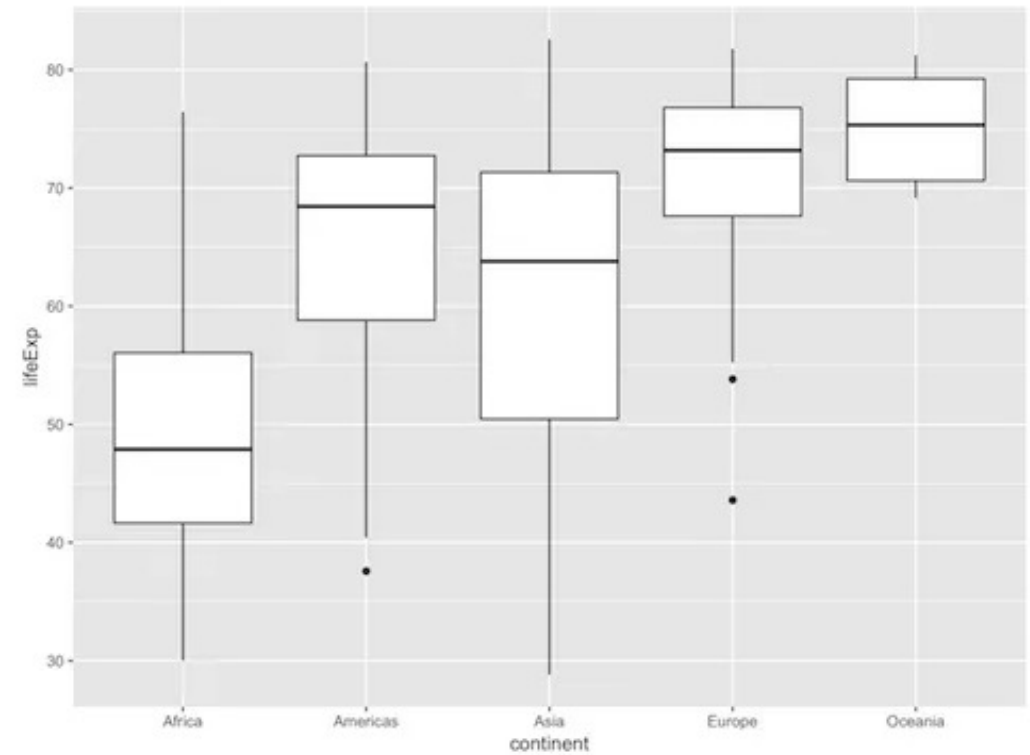


통계량과 시각화

시각화 도구

1. 박스플롯 (boxplot)

- 여러 변량들을 동시에 비교할 수 있음
- 왼쪽그림은 5개 데이터 셋의 요약치를 비교하고 있음



Grouped Boxplot: First Try

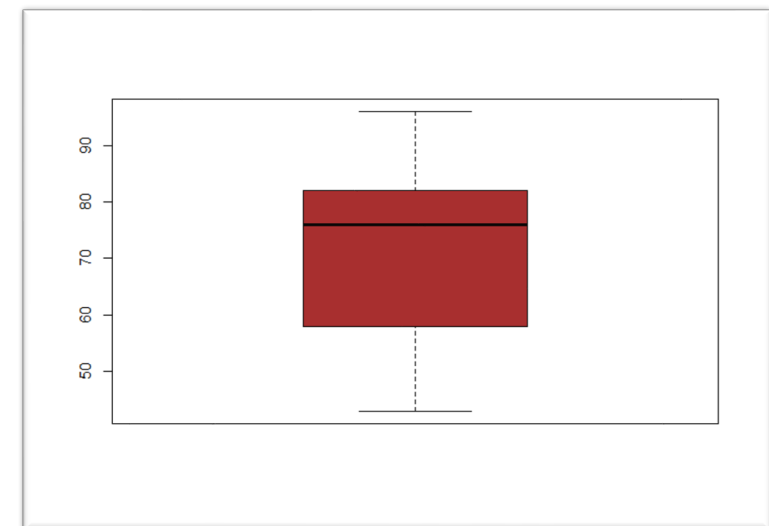
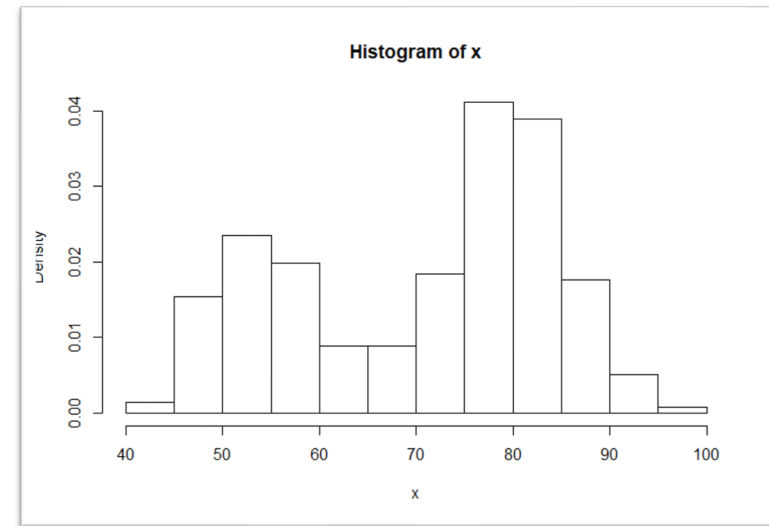
통계량과 시각화

시각화 도구

2. 히스토그램

- 양적자료의 분포의 형태를 보여줌

- 전체적인 자료의 분포 개형을 파악할 수 있어
상자그림에서 확인하지 못하는 자료의 패턴을 탐
지할 수 있음
(자료가 집중된 곳이 두 곳임, 혼합분포)

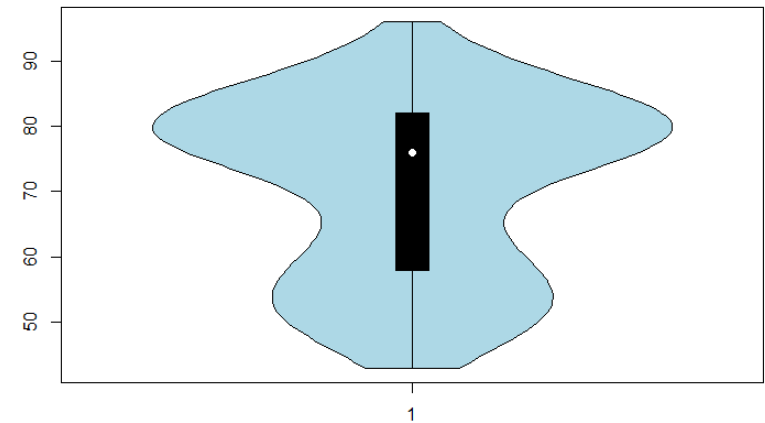


통계량과 시각화

시각화 도구

3. 바이올린플롯

- 히스토그램과 상자그림을 합쳐놓은 시각화 방법

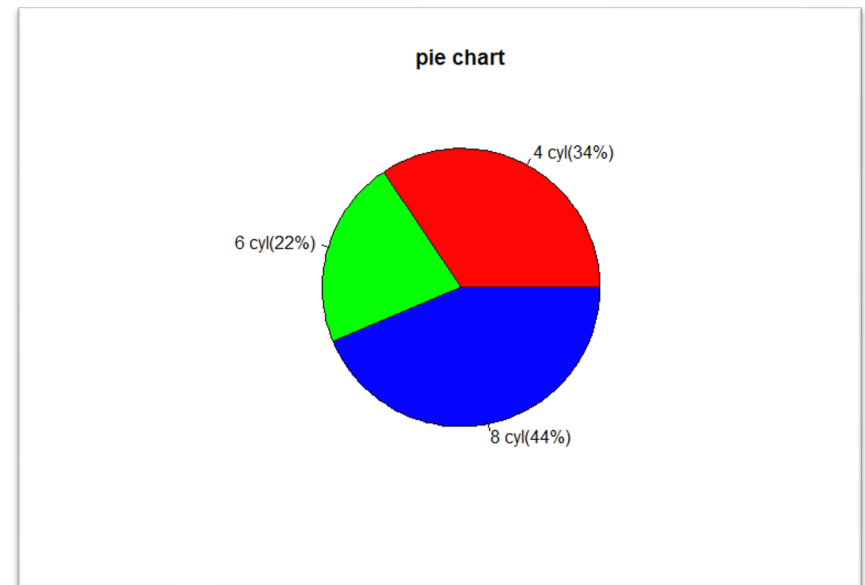
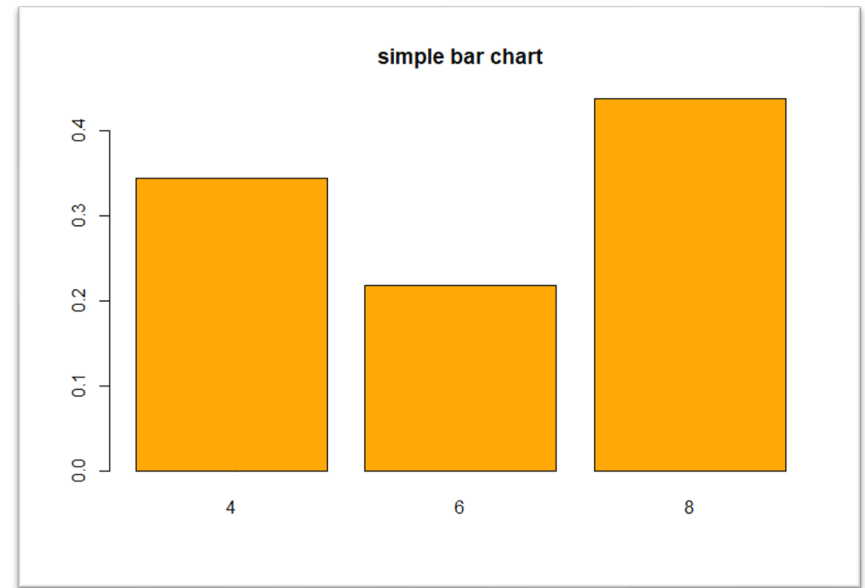


통계량과 시각화

시각화 도구

4. 막대그래프/원도표

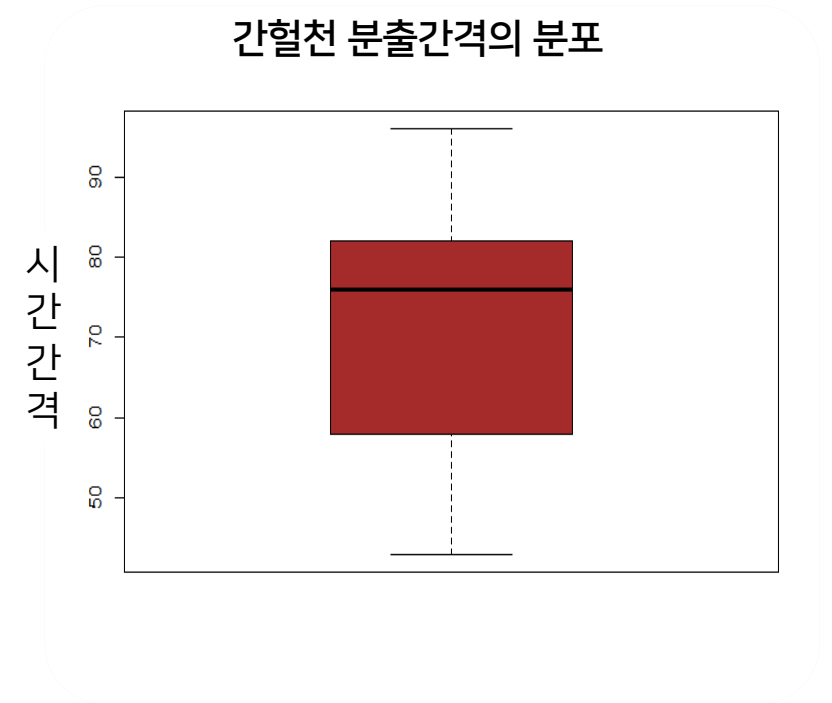
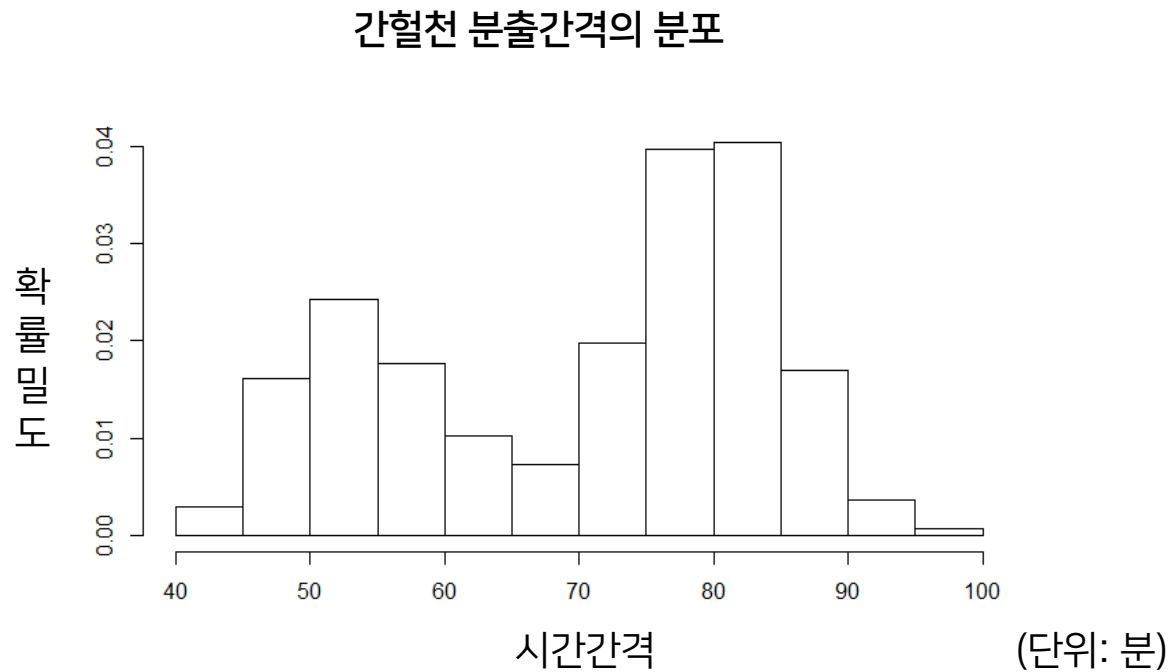
- 자료의 빈도나 비율을 나타내기 위한 그림



5. 시각화를 이용한 스토리 만들기

시각화와 스토리텔링

그림의 제목, 축의 이름 표시



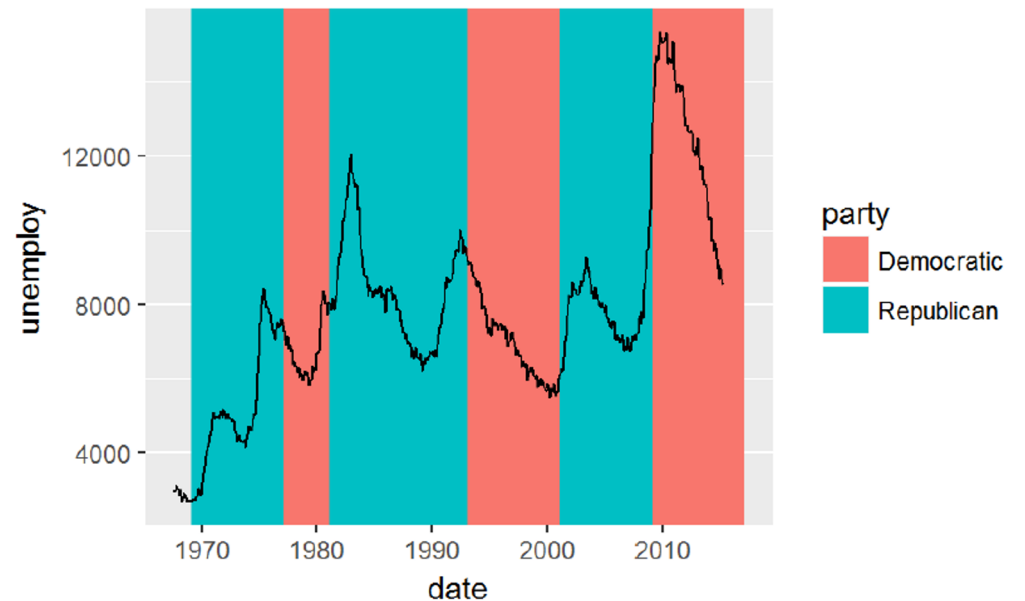
통계량과 시각화

시계열 자료

1. 이벤트 표시

- 시계열 자료에서 어떤 사건이 벌어진 이후에 패턴이 변화하는 것을 말하고 싶은 경우
- 어떤 이벤트 (사건)를 색상을 가진 면으로 시계열 자료 위에 시각화함

집권당과 실업자 수의 추이



통계량과 시각화

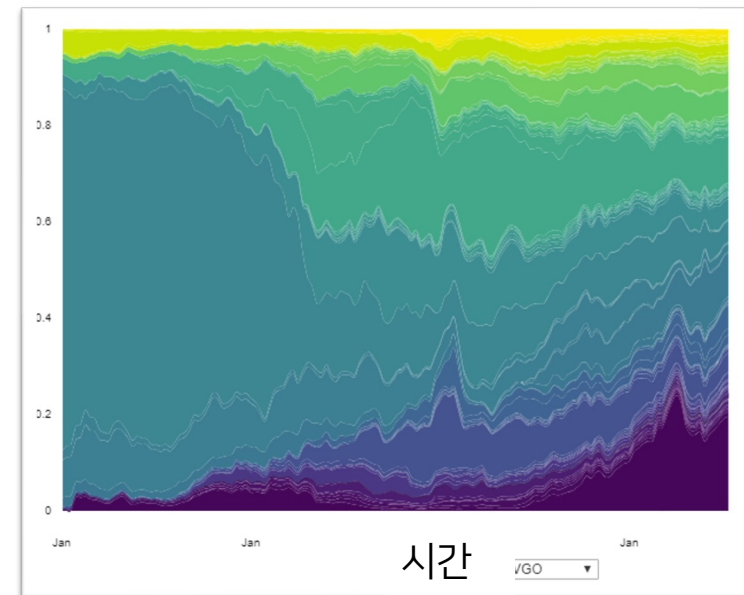
시계열 자료

2. 영역의 표시 (Streamgraph)

- 여러 개의 시계열 자료가 있을 때 시간에 따른 변화를 동시에 비교하고 싶은 경우
- 각 시점에서 상대적인 비중이 중요한 경우 사용
(예: 시가총액에서 비중)

상위10대 기업의 시가총액 비중의 변화

비중



시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

1. 분포의 비교

- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

예) 생태학 연구: 개체의 몸무게와 수면시간의 관계를 밝히
고자 한다. 이 두 변수의 관계는 해당 개체가 어떤 형태의 섭
식을 하는지에 영향을 받을 것이다.

- 자료 (총 84관측치)

bodywt : 몸무게 (kg)

sleep_total : 하루 당 총 수면 시간 (hours/day)

vore : 섭식행태 (Carni: 육식, Herbi:채식, Omni, 잡식,
Insecti: 곤충류)

시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

1. 분포의 비교

- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

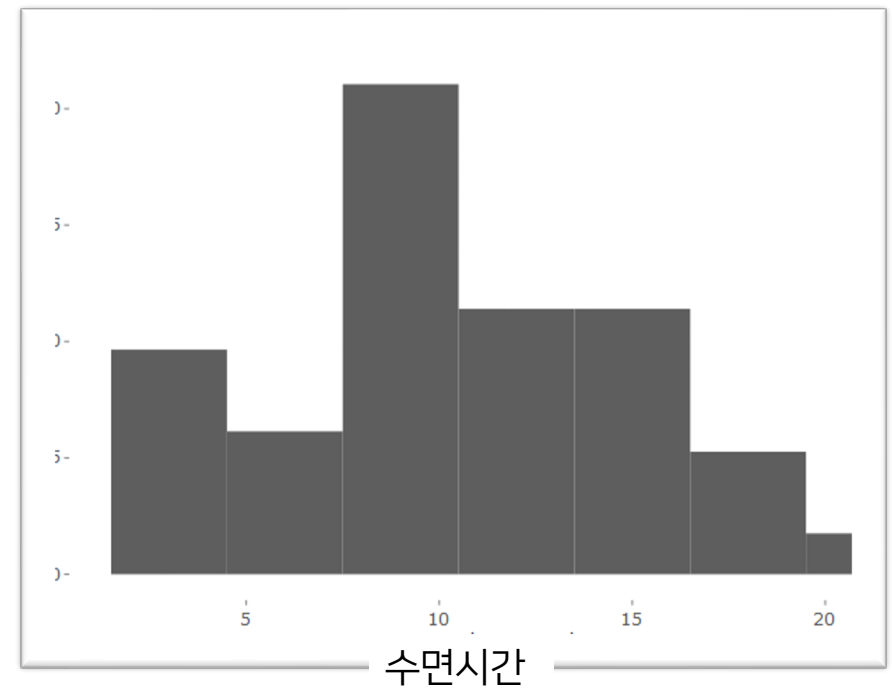
1) 먼저 전체 개체의 수면시간의 분포를 이야기 해보자

평균과 분산:

데이터의 분포:

수면시간의 분포

빈도



시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

1. 분포의 비교

- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

2) 섭식형태로 구분된 그룹에 대해서 각각 데이터의 분포를 확인해본다.

평균과 분산은 각각 어떻게 되는가?

각 집단의 히스토그램은 어떻게 보이는가? 데이터의 개수가 작은 경우 히스토그램의 형태를 파악하기가 어려움



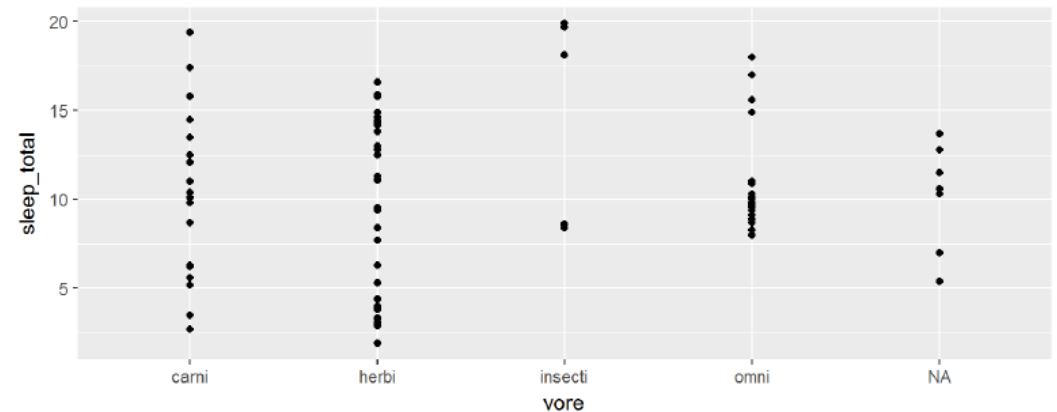
시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

1. 분포의 비교

- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

2) 섭식형태로 구분된 그룹에 대해서 각각 데이터의 분포를 확인해본다.



* 점이 많이 겹쳐서 데이터가 많은 곳끼리 비교가 잘 안된다.

* 점을 좌우로 조금만 흔들면 비교가 쉬울 것 같다.

시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

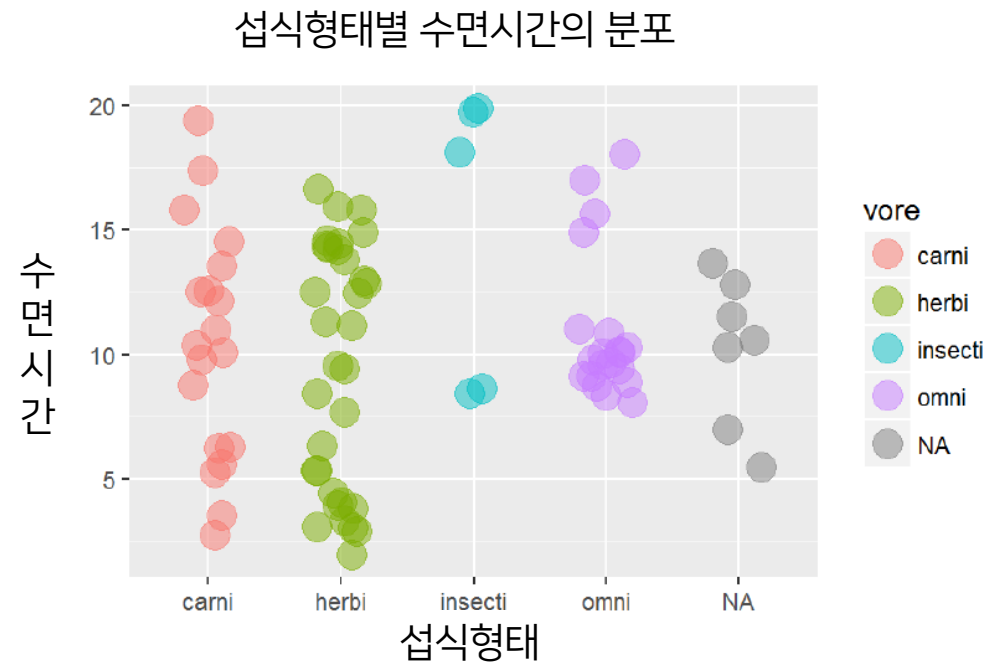
1. 분포의 비교

- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

2) 섭식형태로 구분된 그룹에 대해서 각각 데이터의 분포를 확인해본다.

각 집단의 산점도는 어떻게 보이는가?

(가로축: 섭식형태, 세로축: 수면시간)



시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

1. 분포의 비교

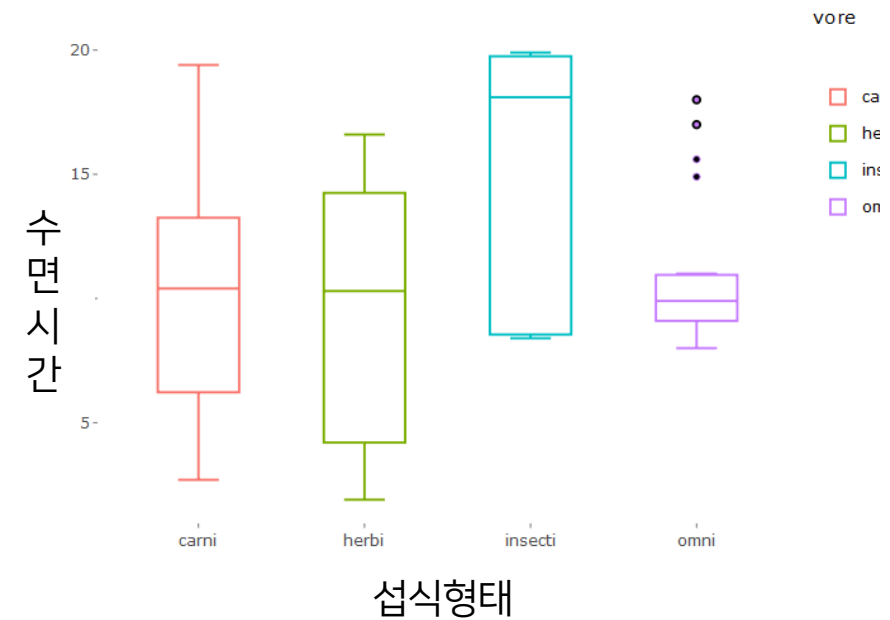
- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

2) 섭식형태로 구분된 그룹에 대해서 각각 데이터의 분포를 확인해본다.

각 집단의 상자그림을 비교한다

(가로축: 섭식형태, 세로축: 수면시간)

섭식형태별 수면시간의 분포



시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

1. 분포의 비교

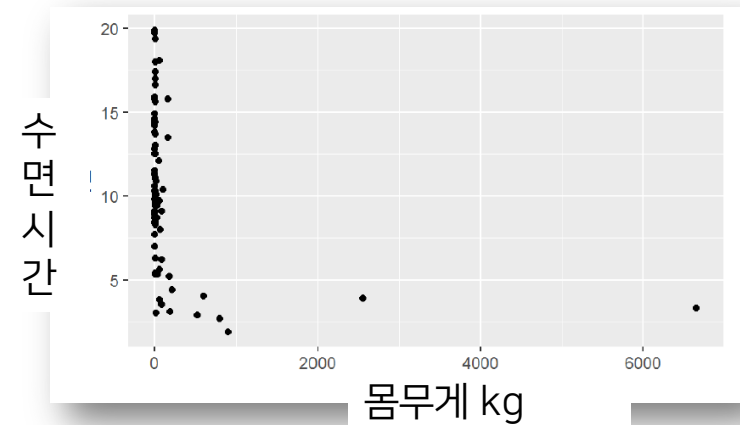
- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

3) 수면시간과 몸무게의 관계를 살펴본다.

(산점도 결과확인: 두 변수간의 관계를 파악하기 어려움)

- 몸무게가 매우 큰 개체가 있어 두 변수간의 관계가 왜곡되어 보이는가? -> 몸무게 100kg이하만 분석해보자

몸무게와 수면시간의 산점도



시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

1. 분포의 비교

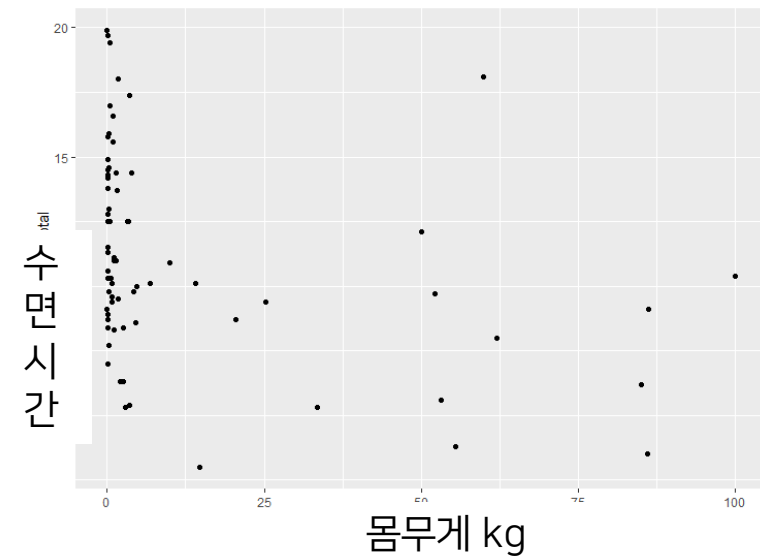
- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

3) 수면시간과 몸무게의 관계를 살펴본다.

(산점도 결과확인: 두 변수간의 관계를 파악하기 어려움)

- 몸무게가 매우 큰 개체가 있어 두 변수간의 관계가 왜곡되어 보이는가? -> 몸무게 100kg이하만 분석해보자

몸무게와 수면시간의 산점도



시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

1. 분포의 비교

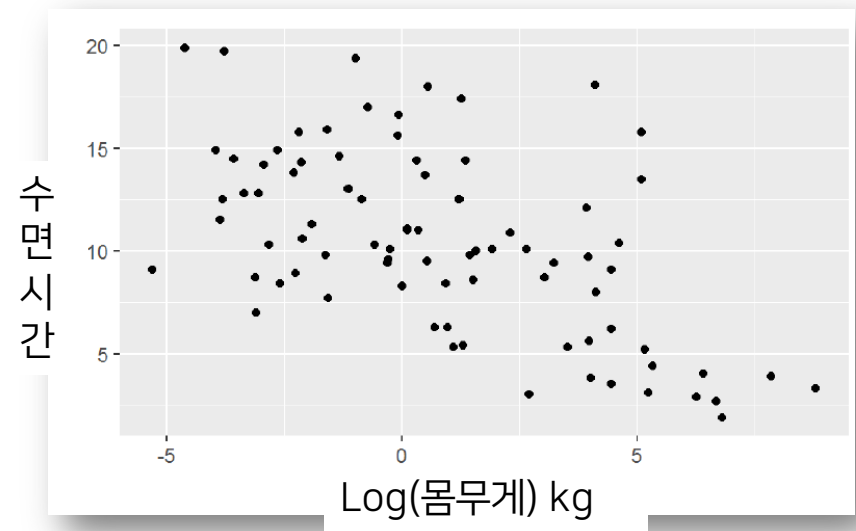
- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

3) 수면시간과 몸무게의 관계를 살펴본다.

- 단위가 큰 변수에 로그변환을 해본다.

* 단위가 큰 변수를 같이 비율로 줄이는 변화는 효과가 없음!

Log(몸무게)와 수면시간의 산점도



시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

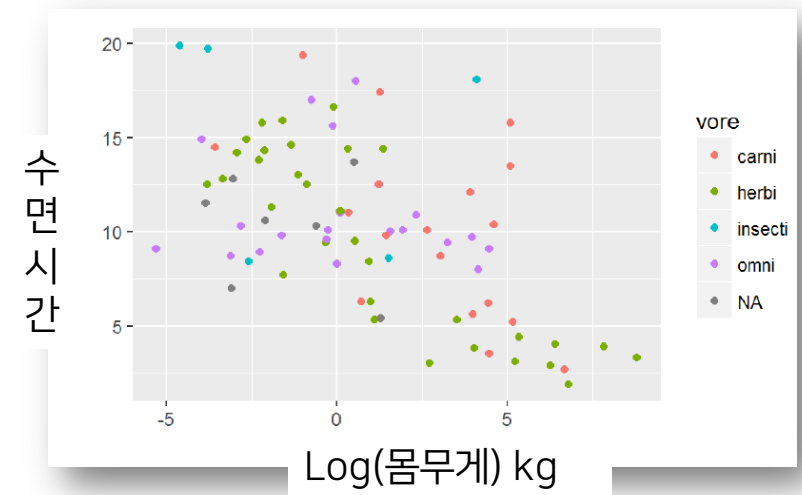
1. 분포의 비교

- 양적자료가 몇 개의 그룹으로 나뉘어진 경우 그 그룹의 차이를 이야기하고 싶다.

4) 섭식형태에 따른 수면시간과 log(몸무게)의 관계를 살펴본다.

(섭식 형태를 점의 색깔로 구분!)

Log(몸무게)와 수면시간의 산점도



시각화와 스토리텔링

몇 개의 그룹으로 나누어진 연속형 자료

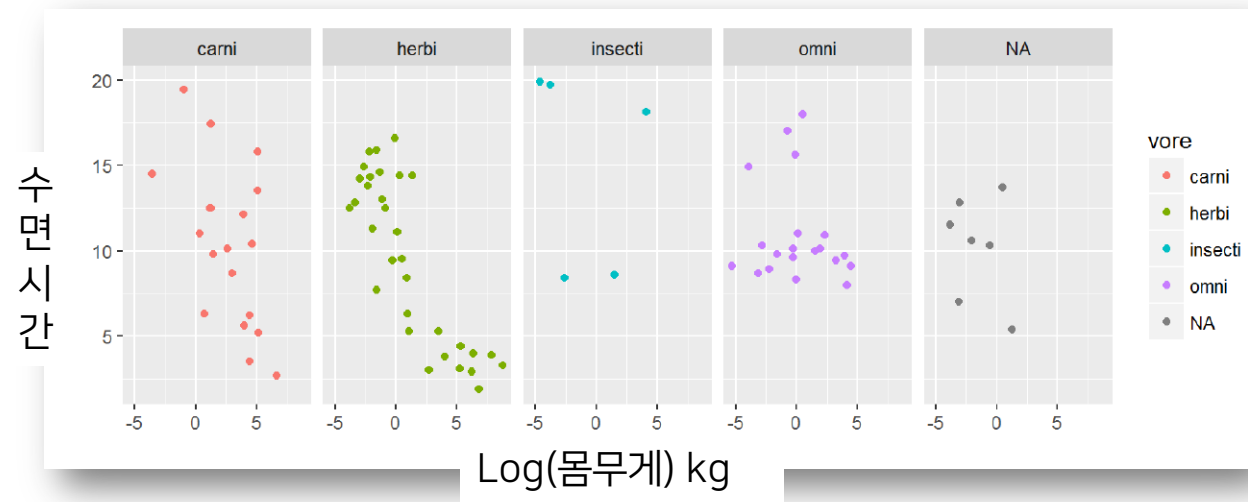
1. 분포의 비교

- 양적자료가 몇 개의 그룹으로 나뉘어진 경우
그 그룹의 차이를 이야기하고 싶다.

4) 섭식형태에 따른 수면시간과 log(몸무게)의
관계를 살펴본다.
(섭식 형태를 점의 색깔로 구분하여 따로 그리
자)

- Carni, Herbi 그룹의 경우에는 log(몸무게)
가 커지면서 수면시간이 줄어드는 경향이 있다.
다른 그룹에서는 그 경향이 나타나지 않았다.

Log(몸무게)와 수면시간의 산점도



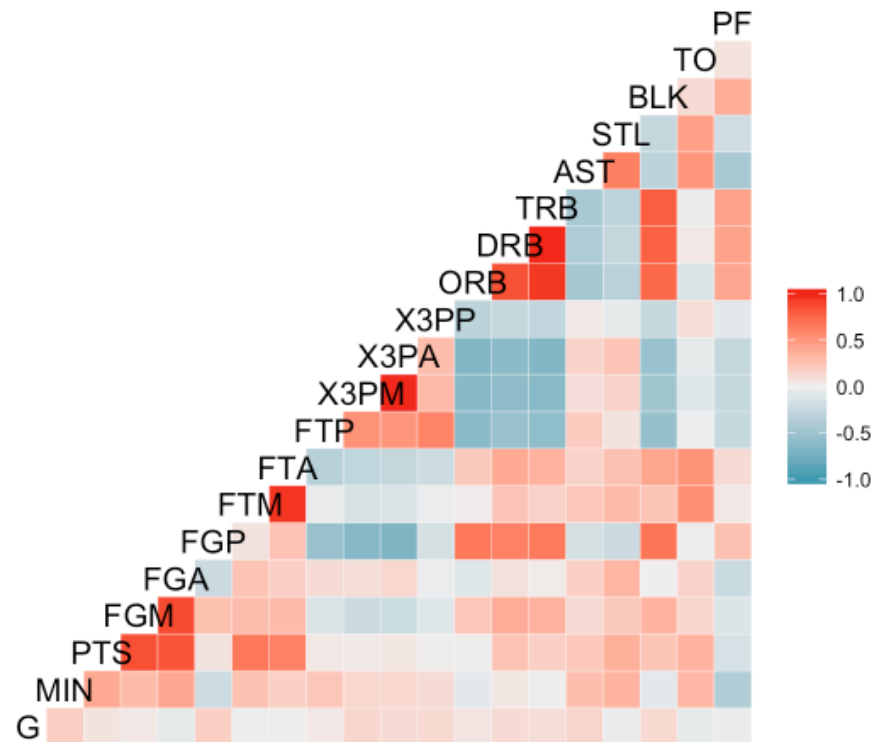
시각화와 스토리텔링

여러 개의 변수를 가지는 연속형 자료

2. 상관계수의 시각화

- 어떤 변수들이 관계가 큰 것인지 말하고 싶다.

예) 농구경기에서 여러 개의 선수 성적의 기록은 어떤 관련이 있는가?



정보가 너무 많아 말하고자 하는 바를 파악하기 어려움!

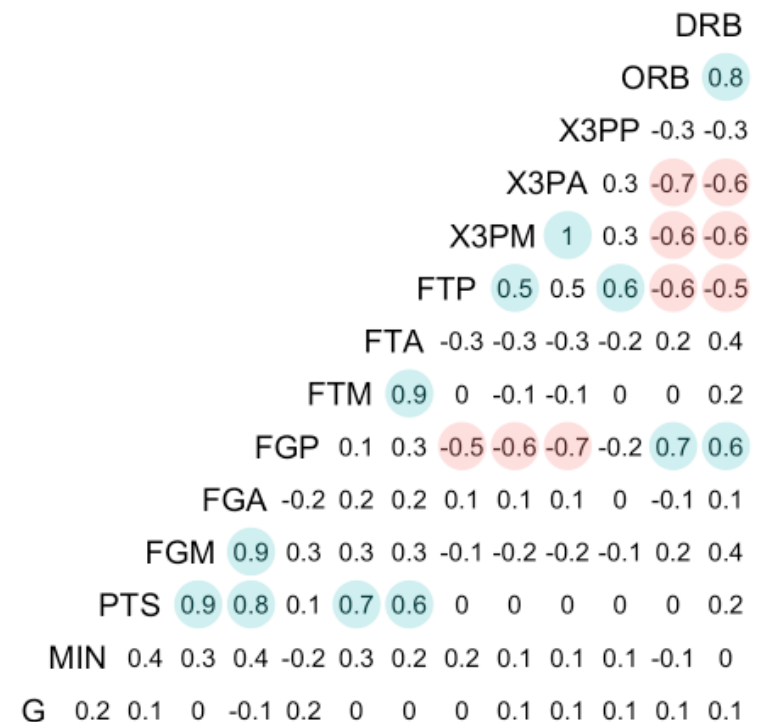
시각화와 스토리텔링

여러 개의 변수를 가지는 연속형 자료

2. 상관계수의 시각화

- 어떤 변수들이 관계가 큰 것인지 말하고 싶다.

예) 농구경기에서 여러 개의 선수 성적의 기록은 어떤 관련이 있는가?

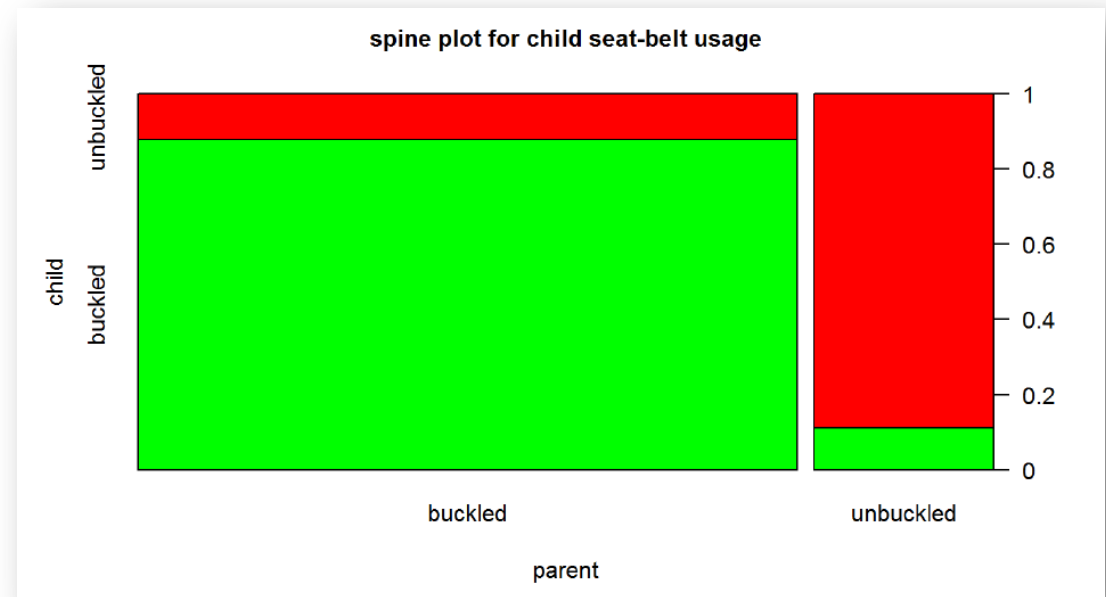


시각화와 스토리텔링

범주형 자료간의 관계의 분석

3. 집단간 범주형 변수의 분포가 다름을 보여주고 싶을 때 (모자이크 플롯)

- 부모님의 안전벨트 착용 유무가 아이들의 안전벨트 착용에 어떤 영향을 주는가?



가로 길이: 부모님의 벨트 착용비

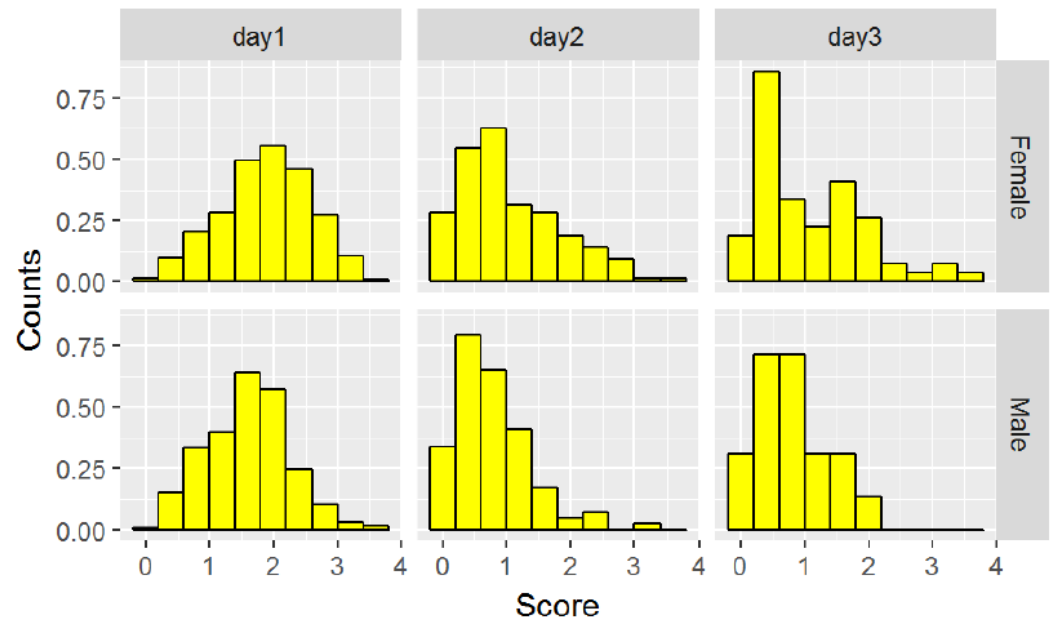
세로 길이: 아이들의 벨트 착용비

시각화와 스토리텔링

범주형 자료와 연속형 자료의 관계

4. 시간에 따라서 두 집단간 연속형 자료의 분포가 달라짐을 보여줌

- 3일동안 열리는 콘서트가 열리는데 날마다 관객이 점수를 매길 수 있다. 날짜에 따라 남성과 여성의 점수 분포가 달라진다.

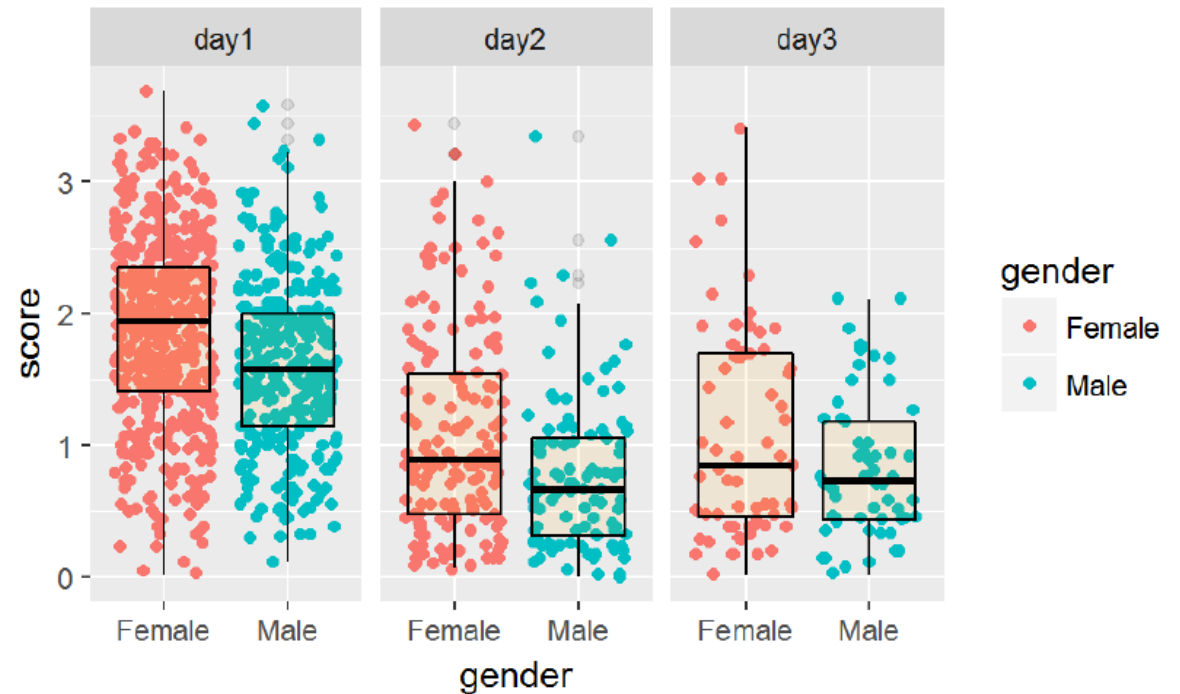


시각화와 스토리텔링

범주형 자료와 연속형 자료의 관계

4. 시간에 따라서 두 집단간 연속형 자료의 분포가 달라짐을 보여줌

- 3일동안 열리는 콘서트가 열리는데 날마다 관객이 점수를 매길 수 있다. 날짜에 따라 남성과 여성의 점수 분포가 달라진다.



시각화와 스토리텔링

범주형 자료와 연속형 자료의 관계

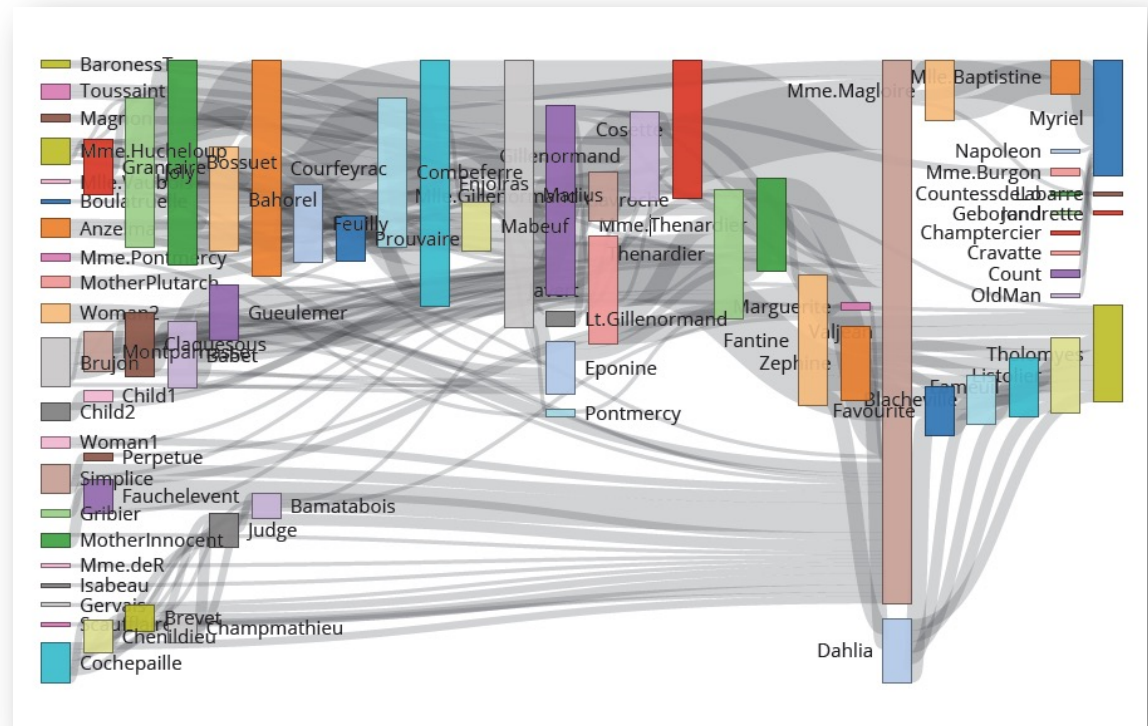
5. 여러 변수간의 관계(네트워크)의 시각화

- 레미제라블의 대사

사각형의 크기: 대본의 대사량

사각형의 연결: 대화의 상대방을 표시

연결된 선의 크기: 쌍방의 대화량



시각화와 스토리텔링

범주형 자료와 연속형 자료의 관계

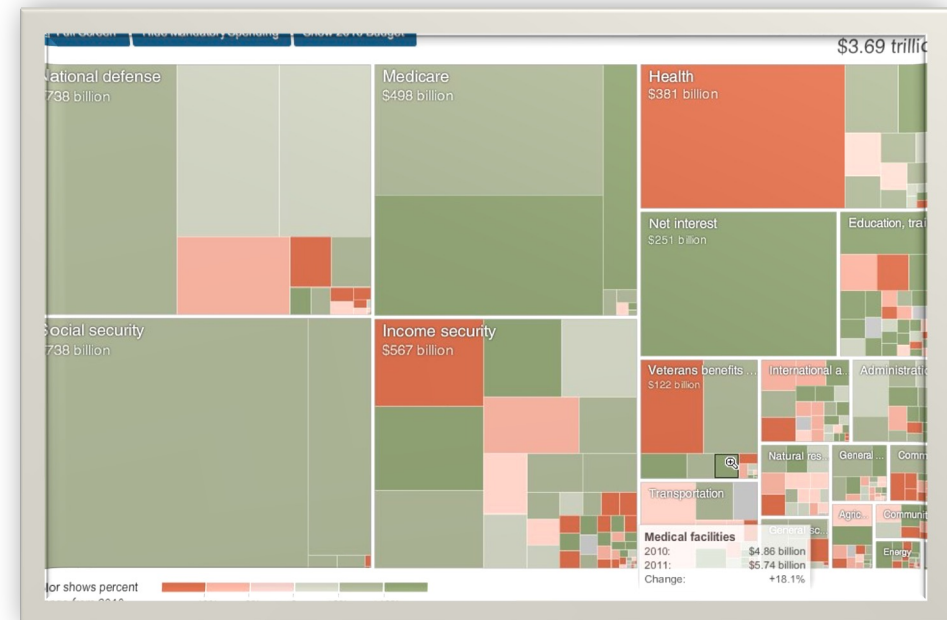
6. 위계가 있는 비율의 표시

(Treemap)

- 예산의 편성 비율 정보

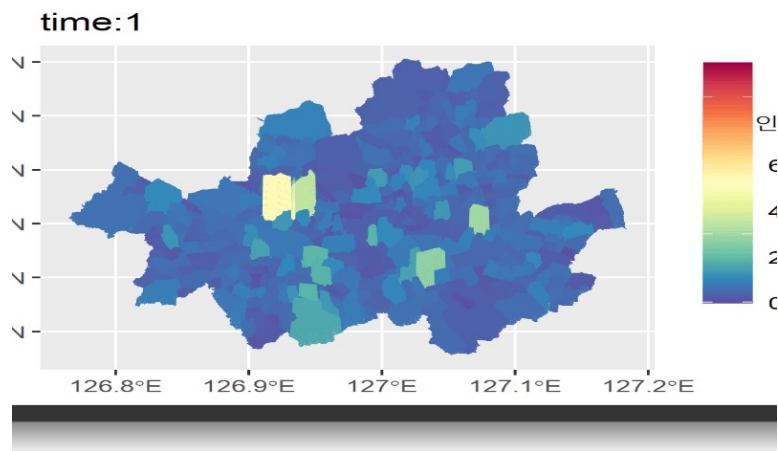
- 예산의 위계가 있어 상위항목이 하위항목들을 가지고 있음

- 직사각형의 면적이 상대적인 비에 근사하도록 그래프를 작성

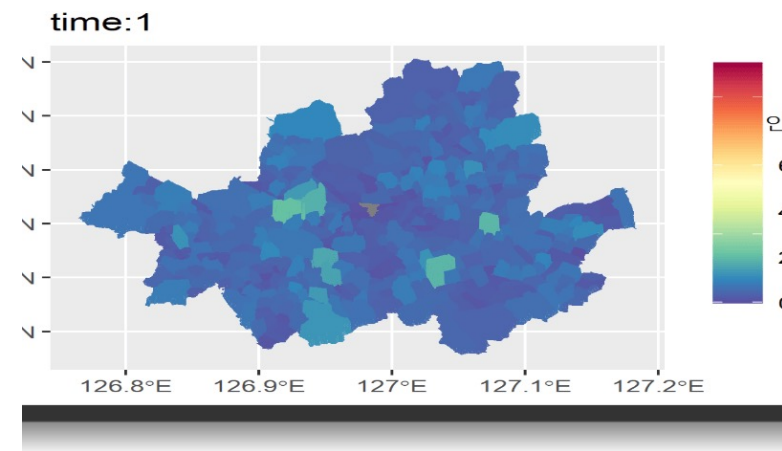


시각화와 스토리텔링

3차원 모형화



Confirmed Cases =
0

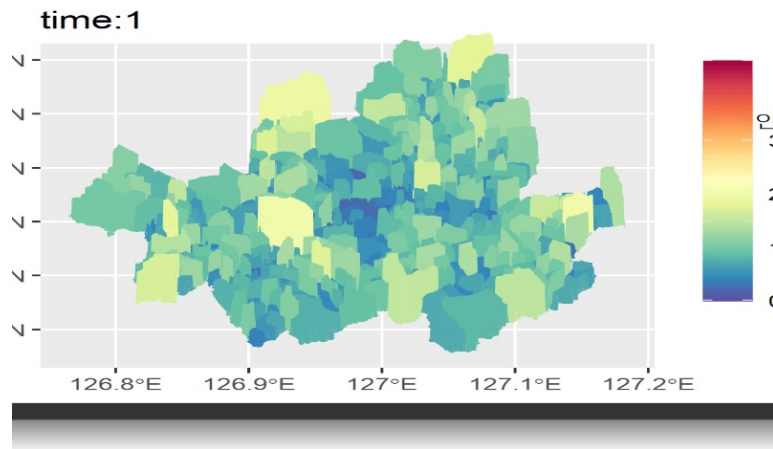


Confirmed Cases =
50

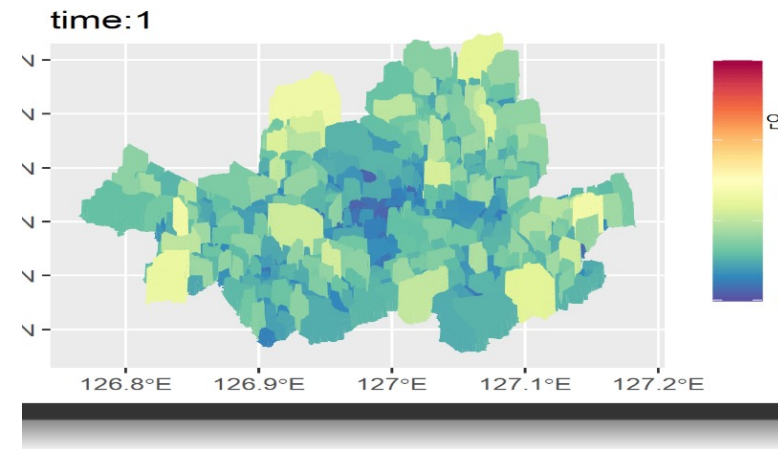
시각화와 스토리텔링

3차원 모형화

Ages: older than 60



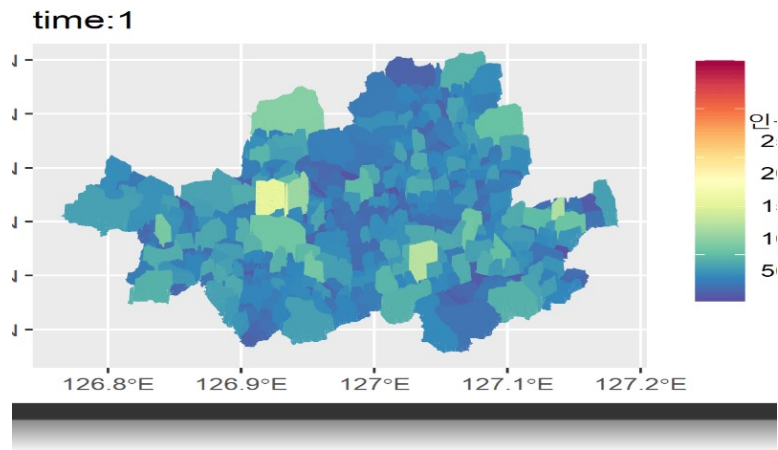
Confirmed Cases =
0



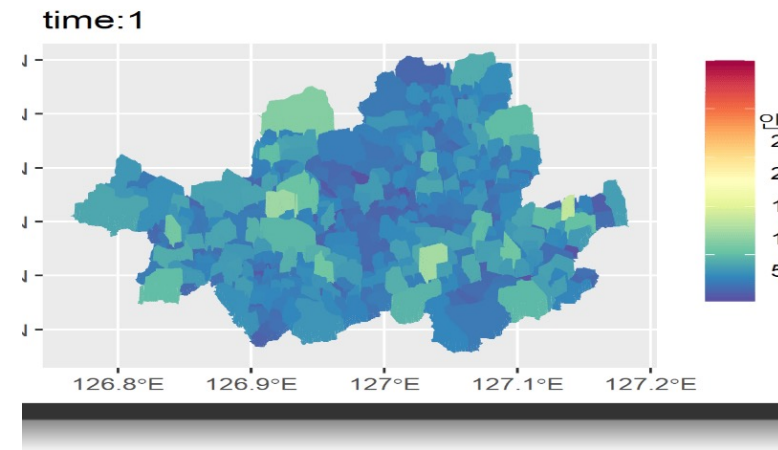
Confirmed Cases =
50

시각화와 스토리텔링

Ages: twenties



Confirmed Cases =
0



Confirmed Cases =
50

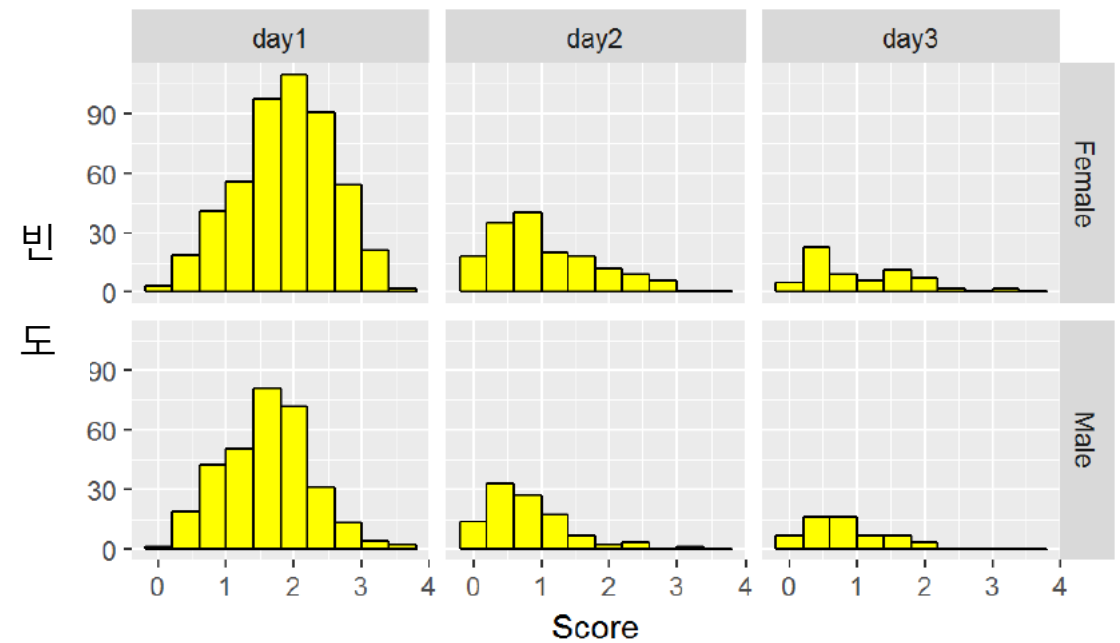
6. 그래프 고치기

그래프 고치기

무엇을 발견하고 싶은가? 무엇을 말하고 싶은가?

(빈도와 분포)

- 3일동안 열리는 콘서트가 열리는데 날마다 관객이 점수를 매길 수 있다. 날짜에 따라 남성과 여성의 점수 분포가 달라진다.

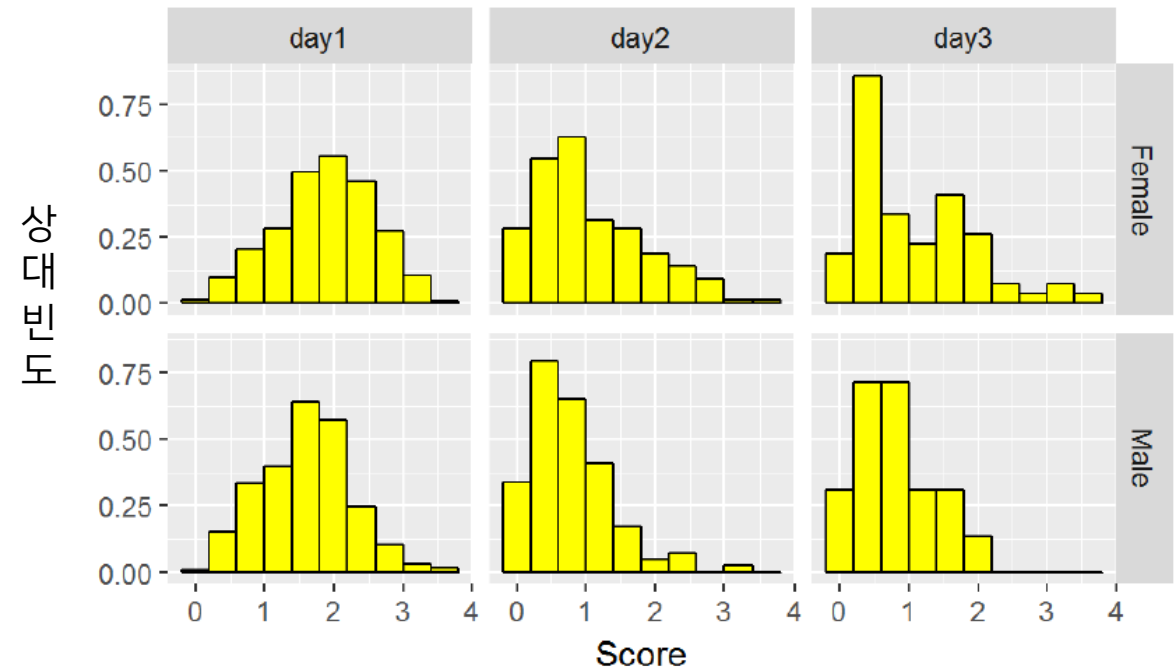


그래프 고치기

무엇을 발견하고 싶은가? 무엇을 말하고 싶은가?

(빈도와 분포)

- 3일동안 열리는 콘서트가 열리는데 날마다 관객이 점수를 매길 수 있다. 날짜에 따라 남성과 여성의 점수 분포가 달라진다.



그래프 고치기

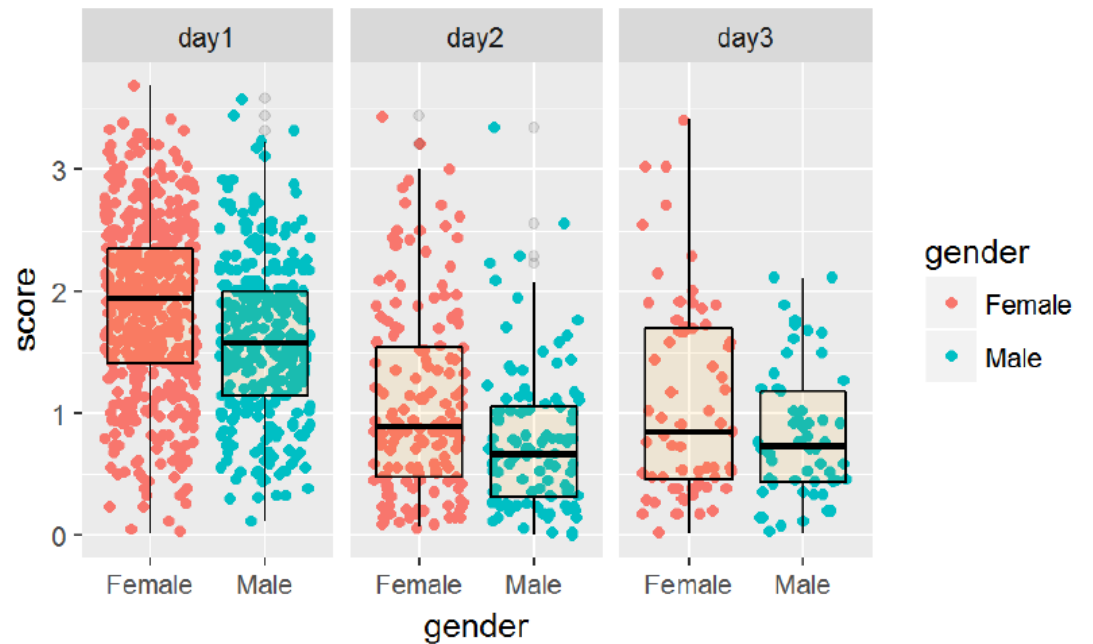
무엇을 발견하고 싶은가? 무엇을 말하고 싶은가?

(빈도와 분포)

- 3일동안 열리는 콘서트가 열리는데 날마다 관객이 점수를 매길 수 있다. 날짜에 따라 남성과 여성의 점수 분포가 달라진다.

그래프의 명확성:

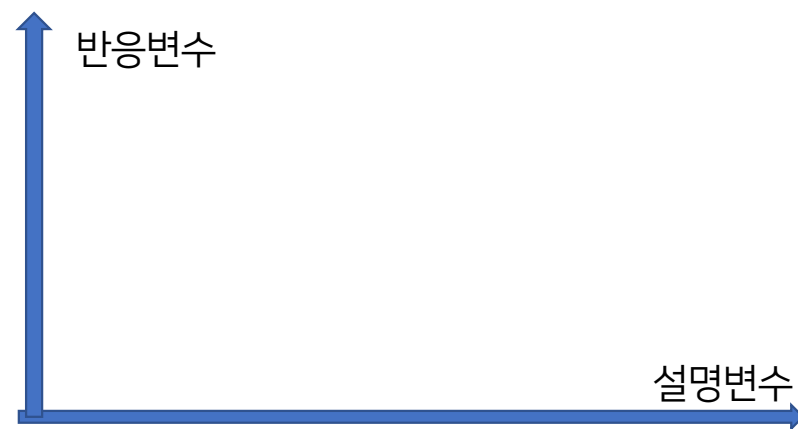
말하고 싶은 것만을 최대한 효율적으로 전달



그래프 고치기

관계를 어떤방식으로 설명하고 싶은가?

- 원인과 결과, 입력과 출력, 설명과 반응

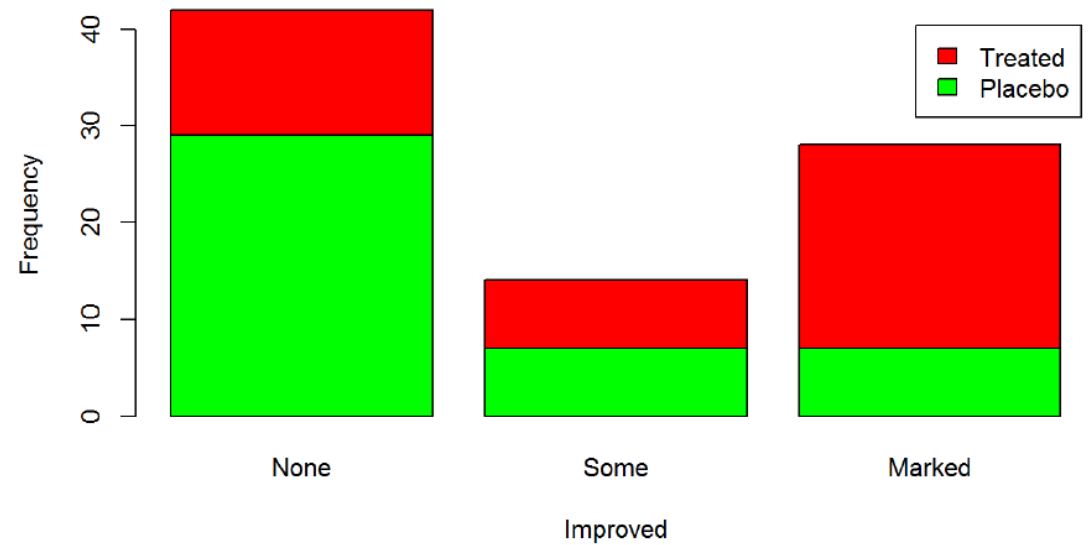


그래프 고치기

관계를 어떤방식으로 설명하고 싶은가?

- 원인과 결과, 입력과 출력, 설명과 반응
- 약의 투여와 관절염의 개선정도를 시각화

* 병증이 개선된 쪽이 진짜 약을 먹은 비율이 높았다!

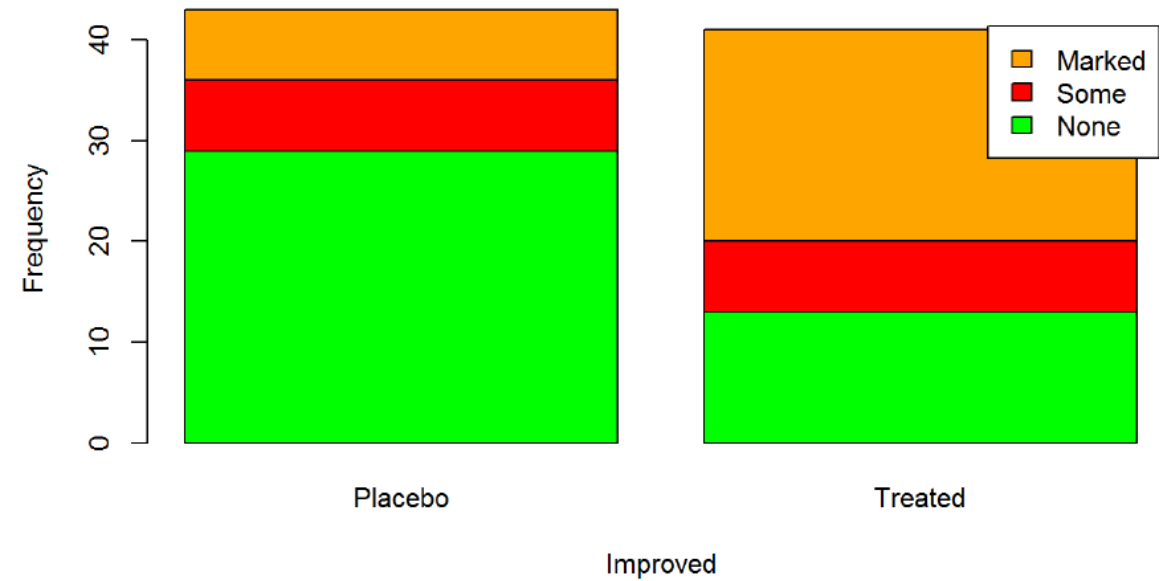


그래프 고치기

관계를 어떤방식으로 설명하고 싶은가?

- 원인과 결과, 입력과 출력, 설명과 반응
- 약의 투여와 관절염의 개선정도를 시각화

* 진짜약을 먹은쪽이 개선된 비율이 높다!

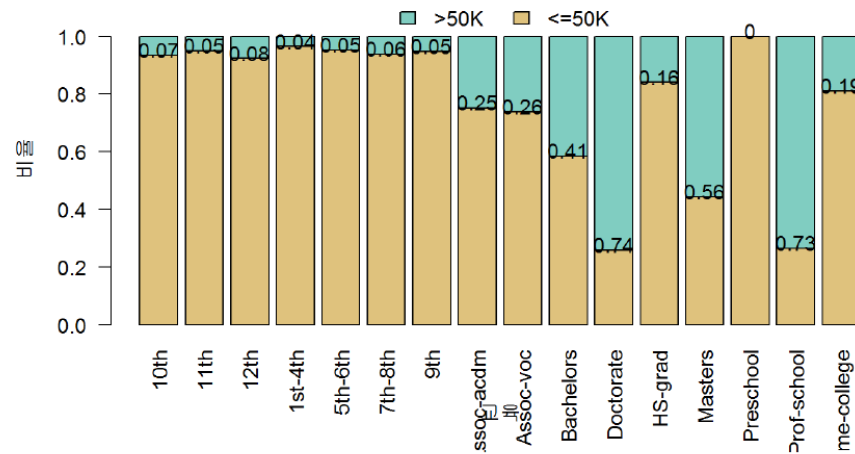


그래프 고치기

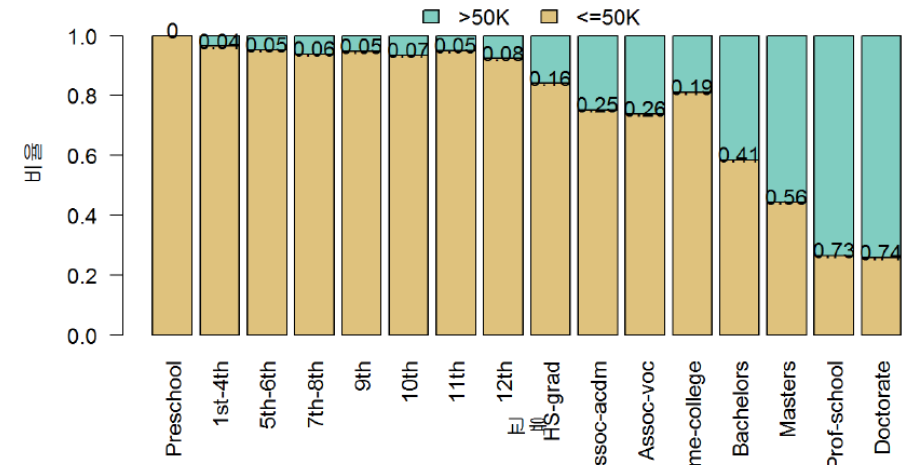
관계를 어떤방식으로 설명하고 싶은가?

- 교육수준에 따른 5만달러 이상 소득자의 비율
- 변수의 재배열

교육과 수입 관계



교육과 수입 관계



그래프 고치기

관계를 어떤방식으로 설명하고 싶은가?

- 범지도

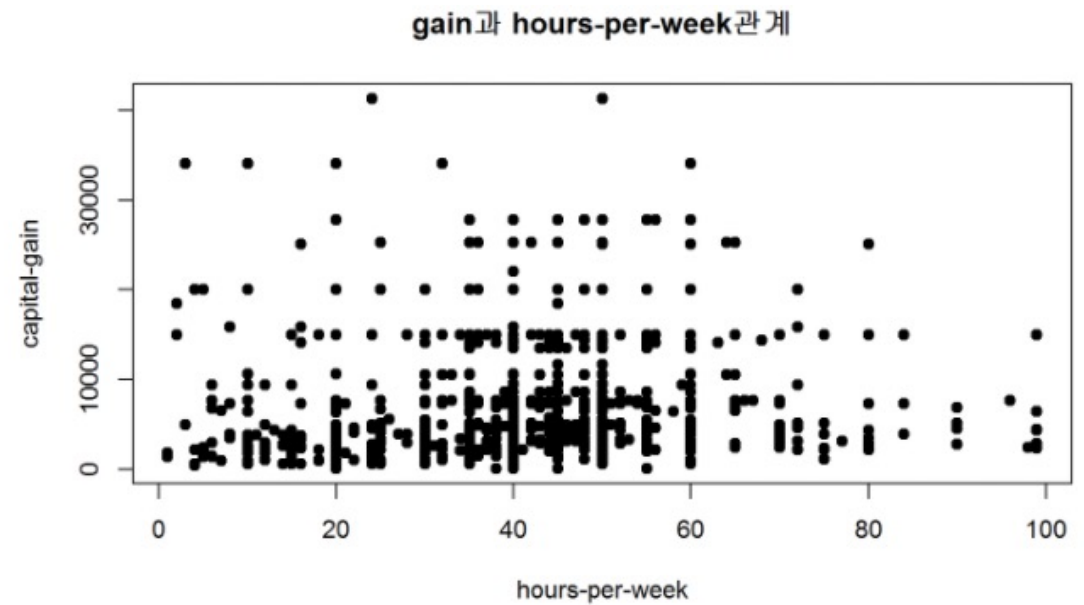
원의 크기를 통해 범주의 심각성을 나타내고 있음



그래프 고치기

투명도의 사용

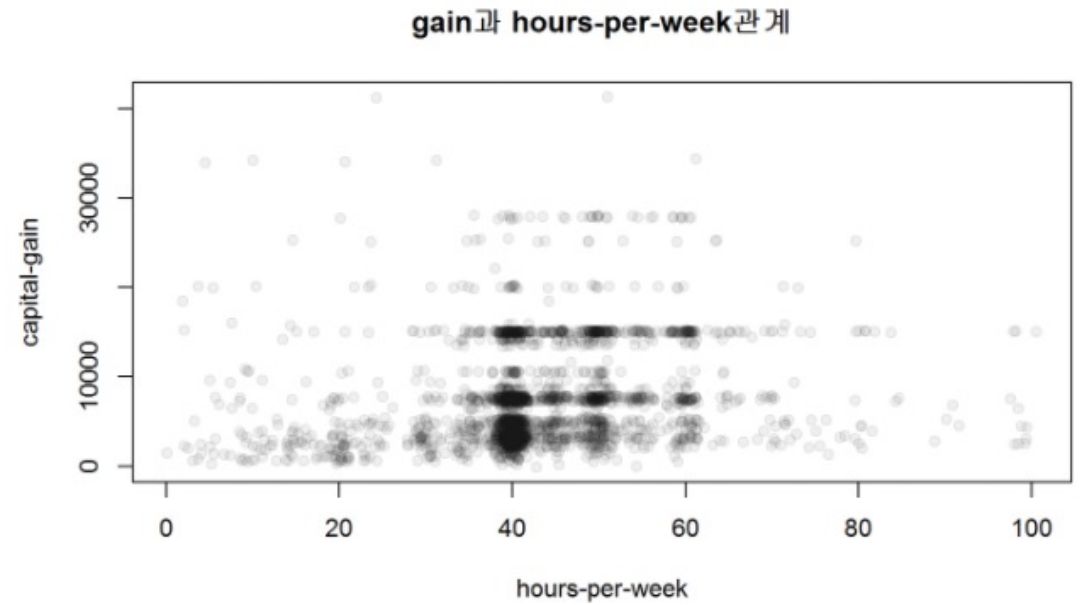
- 주당 근로시간과 자본소득간의 관계를 산점도로 표시한 것임
- 같은 값을 가지는 데이터가 많은 경우, 데이터가 겹쳐서 표시됨
- 데이터의 밀도가 올바르게 표시되지 않음



그래프 고치기

투명도의 사용

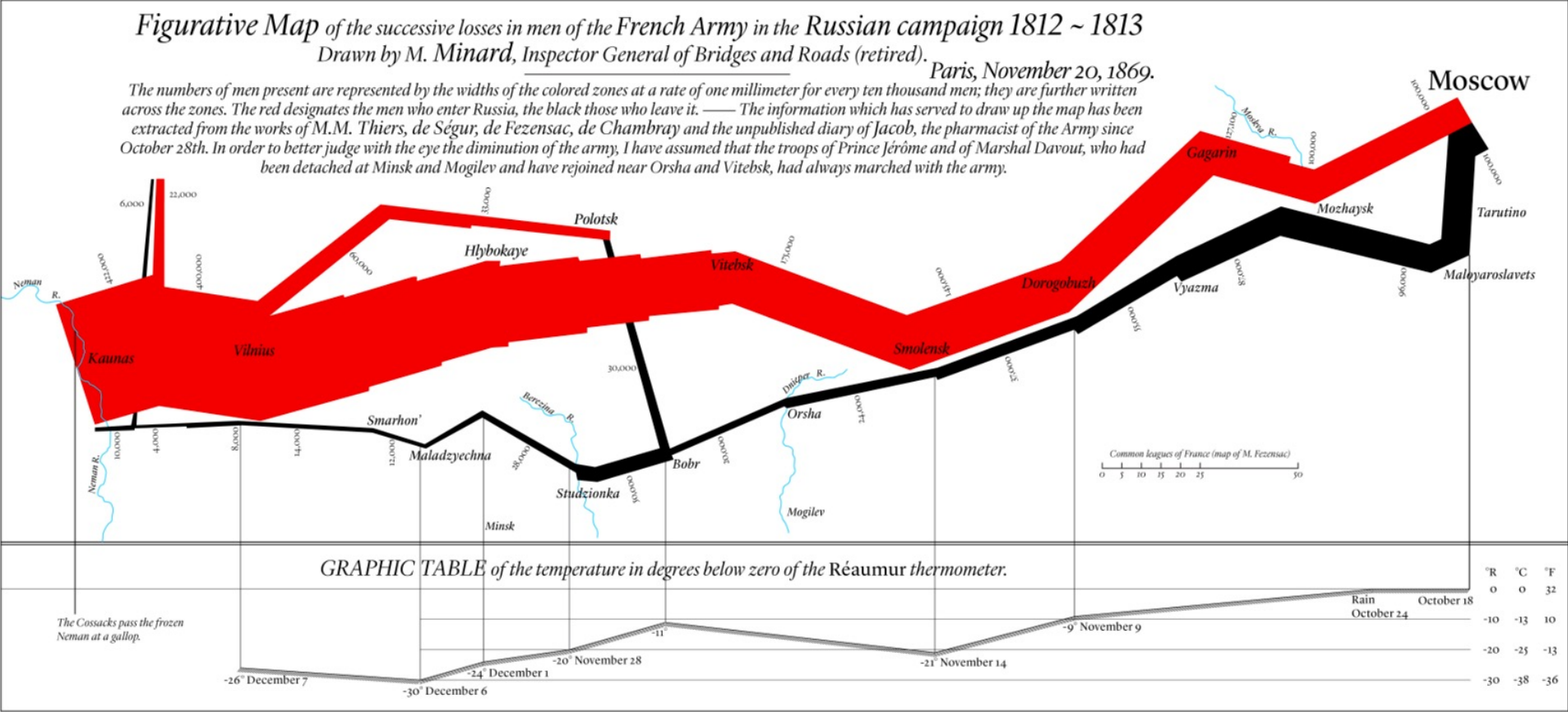
- 점의 속성에 투명도를 설정함
- 점의 위치를 가로 축으로 조금 흔들어 놓음
- 평균적인 근로시간의 위치와 자본소득의 중심점이 관찰됨



7. 맺는말

1. 중요한 것은 스토리다.
2. 스토리는 데이터를 보면서 만들어 갈 수 있다.
3. 그림을 통해 만들어진 스토리에 생명력을 불어 넣을 수도 있다.

그래프 고치기



-

End of Document

-

Thank You

-