

1 데이터시각화 분석

[Open in Colab](#)

1) 배경:

"데이터 분석에 이 울레들이 어떻게 쓰이는지 정리 해 보면.."

데이터 분석	목적	대응
데이터 시각화	데이터가 어떻게 생겼는지 알고 싶다	(1) 전체 데이터를 한눈에 확인 - 점그림, 선그림, 영역그림 - 막대그림, 등고선그림, 분포그림(히스토그램) → 통계 사용 (2) 데이터를 뿌리고 통계를 계산하기 위해 데이터 값 하나하나를 표현 → 확률/컴퓨터 사용
기술적 분석	데이터가 어떻게 생겼는지 알고 싶다	(1) 전체 데이터 특성을 몇 개의 숫자들로 확인 → 통계 사용 (2) 통계를 계산하기 위해 데이터 값 하나하나를 표현 → 확률/통계 사용
상관관계/인과관계	여러종류 데이터끼리의 관계를 알고 싶다	(1) 각 데이터를 몇 개의 숫자들로 표현 → 통계 사용 (2) 표현된 숫자들을 비교 → 확률/통계 사용
통계추론	일부 데이터로 전체를 알고 싶다	(1) 일부 데이터의 특성을 확인 → 통계 사용 (2) 반복적으로 실험 진행 및 통계치 재확인 → 컴퓨터 사용 (3) 전체 특성을 추론 → 통계 사용
알고리즘학습	전체 데이터로 미래를 알고 싶다	(1) 데이터의 관계를 수학적으로 표현 → 확률/통계/함수/컴퓨터 사용 (2) 미래를 예측한 후 정확성 확인 → 확률/통계/컴퓨터 사용
가설검정(A/B Test)	원가 진실과 가까운 의사결정을 하고 싶다	(1) 기존 데이터와 새로운 데이터 비교를 위해 숫자들로 표현 → 통계 사용 (2) 표현된 숫자들을 비교 → 확률/통계/컴퓨터 사용

2) 목적: 데이터 시각화(Data Visualization)는 효과적으로 정보를 전달하는 수단

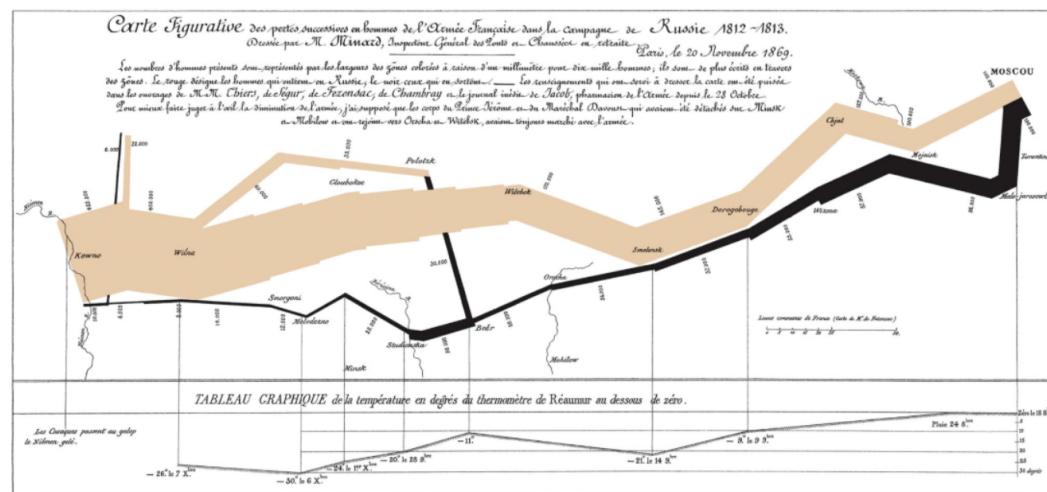
- 인간은 눈으로 입력된 정보가 뇌에 전달될 때 정보의 의미를 파악하도록 진화된 동물
- 인간의 감각 중 대부분은 시각에 의존하는데 약 77%
- 빅데이터 시대에 데이터를 단순히 눈으로 볼 수 있는 한계를 넘어섬
- 데이터 시각화 분석은 빅데이터를 일일이 보지 않고 인사이트 도출 과정

1.1 역사상 최고로 꼽히는 데이터 시각화

"뛰어나고 아름다운 데이터 시각화는 분석 기술 뿐만 아니라 그래픽 디자인과 스토리텔링 창의성 필요"

1) 1812년 나폴레옹의 러시아 진군 맵

- "왜 나폴레옹은 겨우 1만명만 돌아왔을까?"

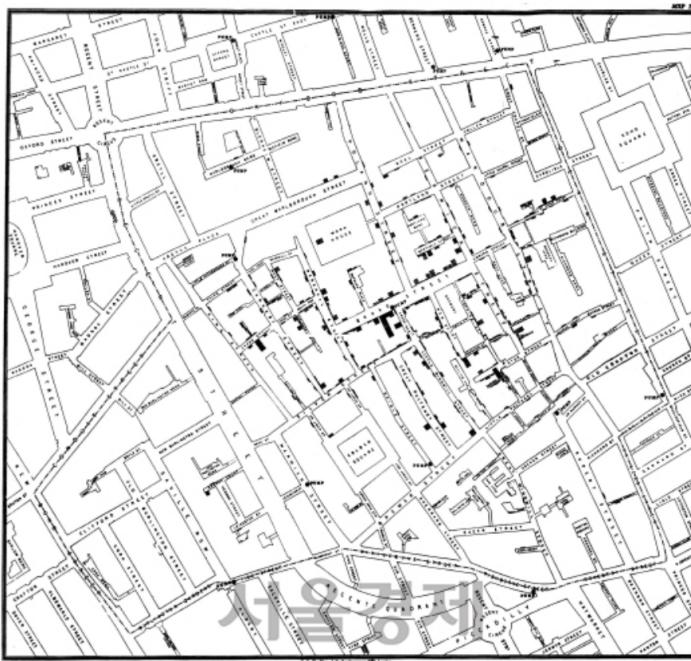


비주얼리제이션 작성자: Charles Joseph Minard

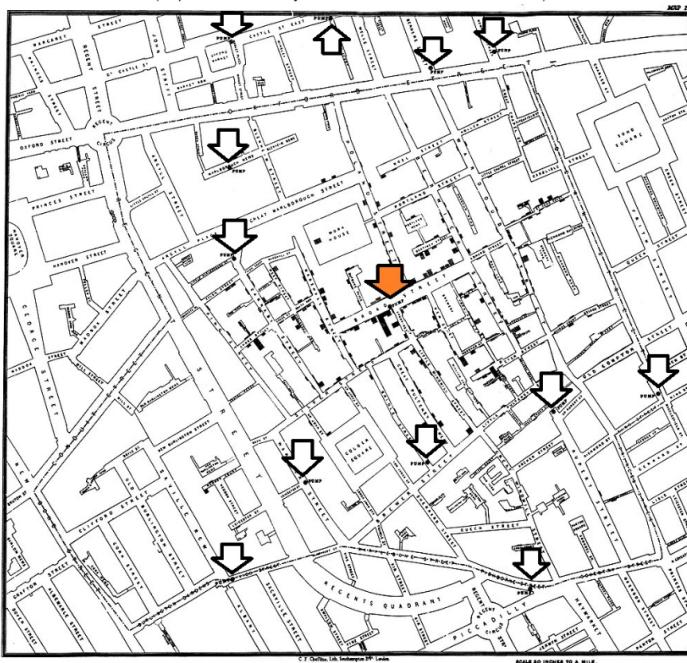
- (1) 나폴레옹 군대는 47만명이 출정하였으나 1만명만 복귀
- (2) 이동에 따른 급격한 겨울 기온으로 인해 피해가 극대화
- (3) 패배 이후에도 끝난게 아닌 전쟁의 현실

2) 1854년 Broad가의 콜레라 발병 맵

- "왜 콜레라 사망 추세가 다른 곳보다 높을까?"



(<https://www.sedaily.com/NewsView/1OC2UDNWZV>)



(<https://www.scientetimes.co.kr/news/%EC%97%AD%ED%95%99%EC%9D%98-%EC%97%AD%EC%82%AC%EB%A5%BC-%EC%97%B0-%EC%8A%A4%EB%85%B8%EC%9D%98-%EC%BD%9C%EB%A0%88%EB%9D%BC-%EC%A7%80%EB%8F%842/>)

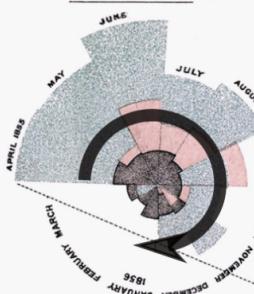
- (1) 콜레라 피해를 가장 많이 입은 세대들은 모두 같은 우물을 식수로 사용
- (2) 콜레라와 오염된 우물 간의 상관관계를 분석으로 입증
- (3) 우물을 오염되지 않게 보호하는 것이 콜레라 예방 솔루션으로 정책적 반영

3) 1850년대 크림전쟁 사망 원인 맵

- "전쟁 중 사망 원인은 무엇일까?"

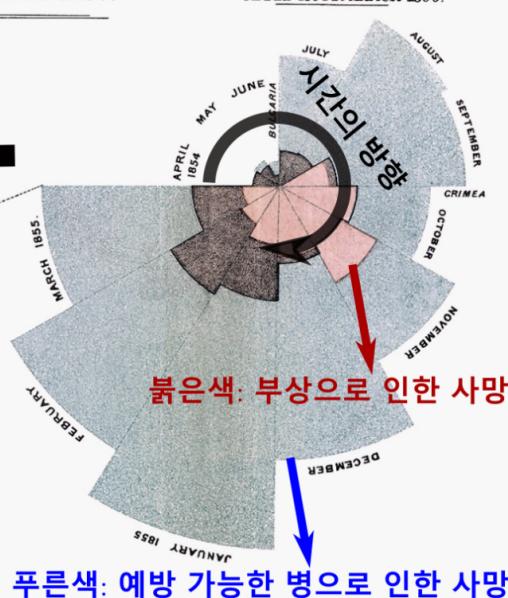
1855년 4월 - 1856년 3월

APRIL 1855 TO MARCH 1856.



1854년 4월 - 1855년 3월

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 Preventable 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.

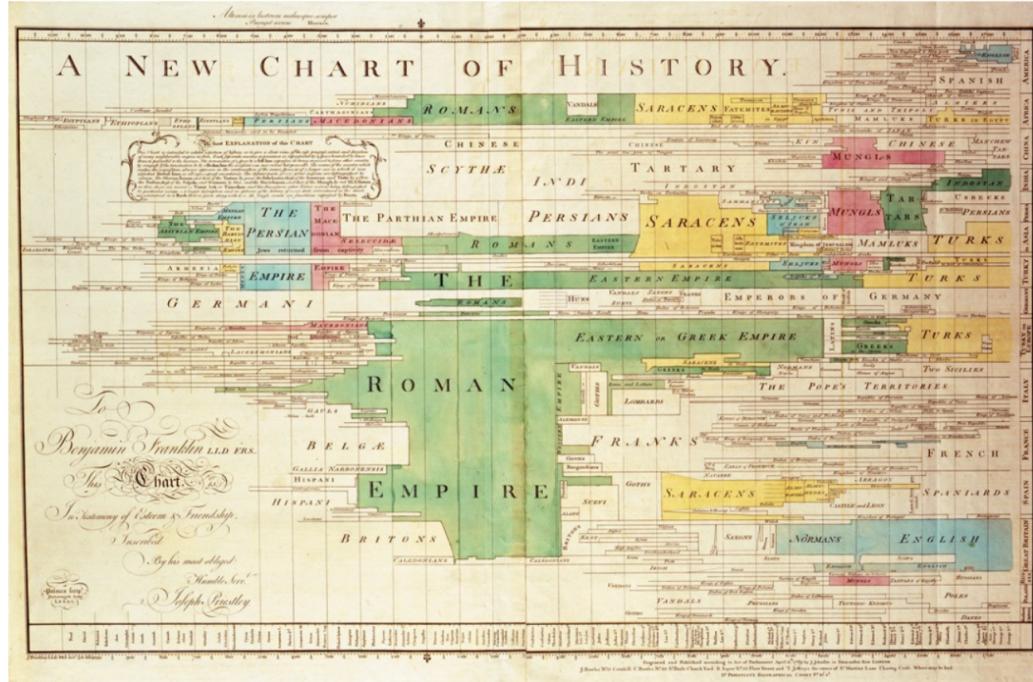
(1) 크림 전쟁 중에 군인 사망률은 계속 증가세

(2) 간호사였던 Florence Nightingale은 대부분 사망이 전투가 아닌 열악한 병원 상태임을 파악

(3) 부상 이외의 병원환경 개선정책으로 사망자 수 감소가 가능할 것이라는 인사이트

4) 역사적 진화 차트

- “동시대에 존재한 주요 제국과 문화의 영향은?”



비주얼리제이션 작성자: Joseph Priestley

(1) 색상과 크기로 영향력의 정도를 표현

(2) X축은 시간정보를 Y축은 창의적인 위치 표현

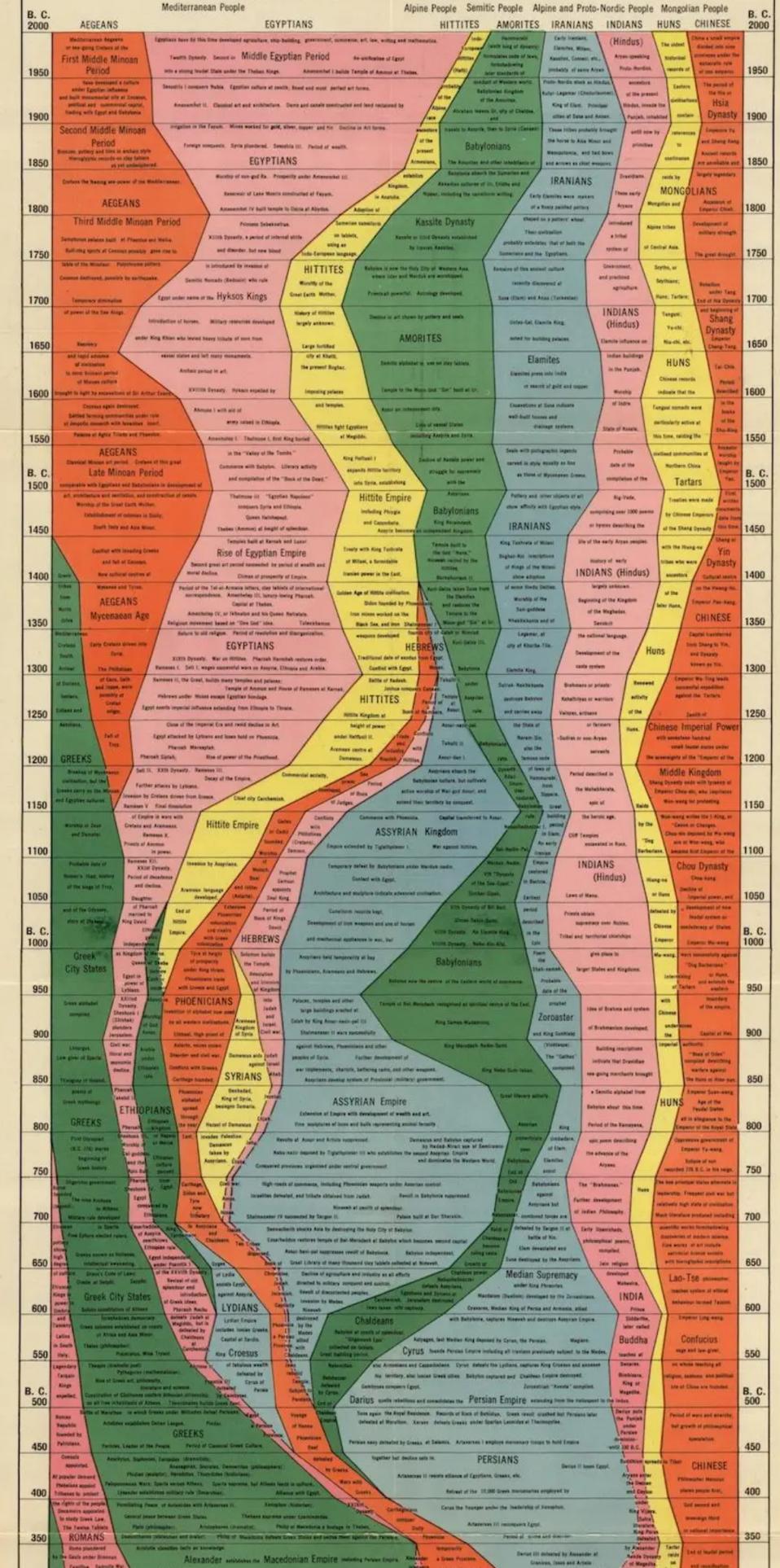
- “4000년의 역사”

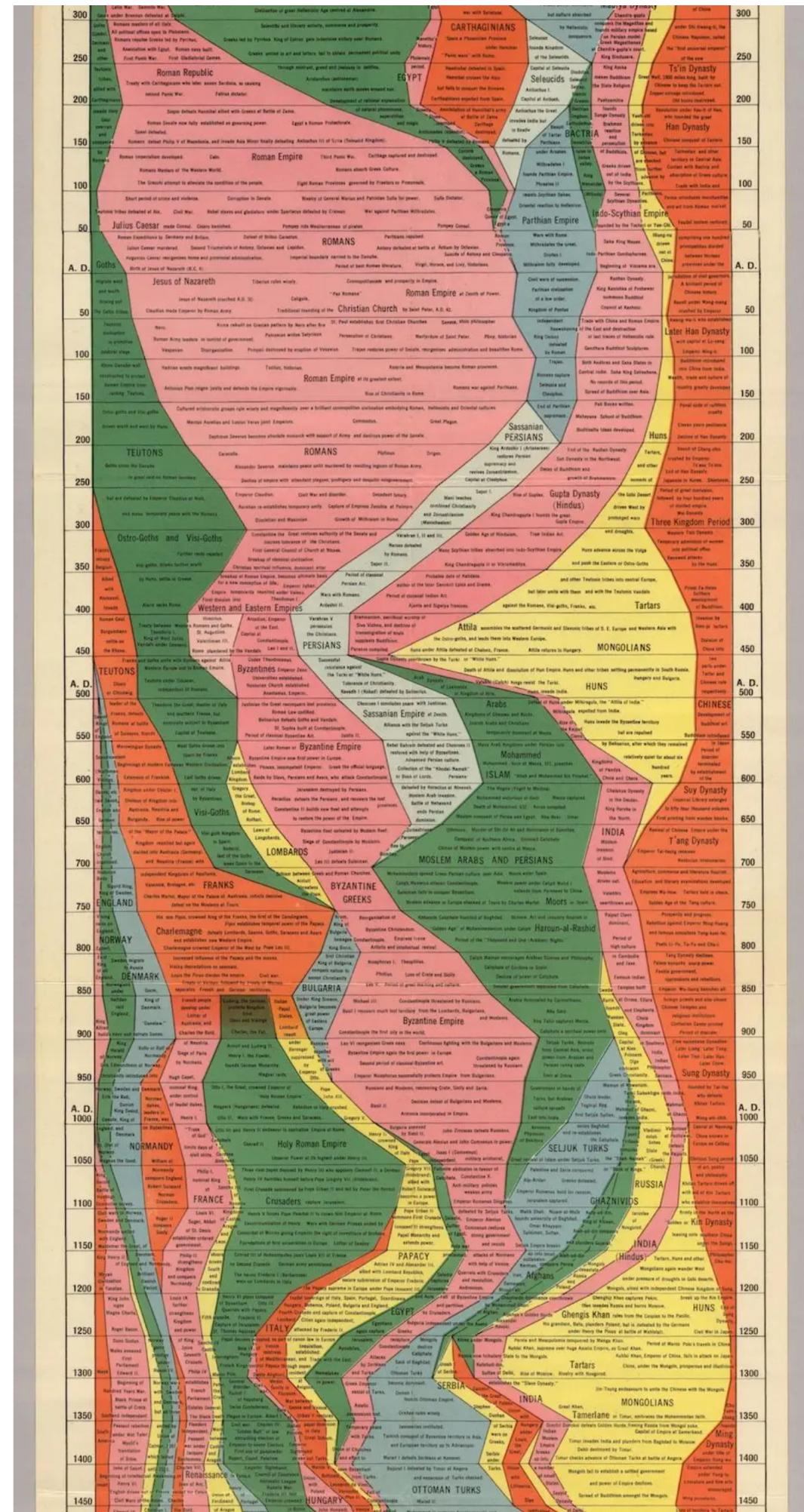


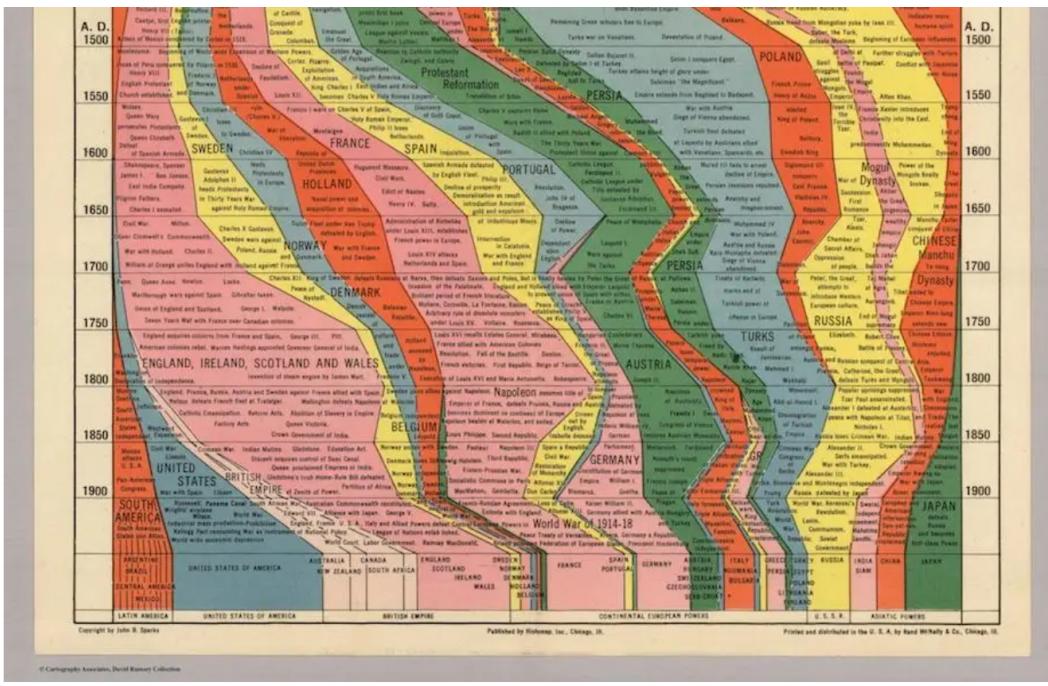
THE HISTORYMAP

FOUR THOUSAND YEARS OF WORLD HISTORY

RELATIVE POWER OF CONTEMPORARY STATES, NATIONS AND EMPIRES







(<https://www.businessinsider.com/one-chart-shows-all-of-world-history-2016-8>)

5) 2016년 오바마 정부의 예산 맵



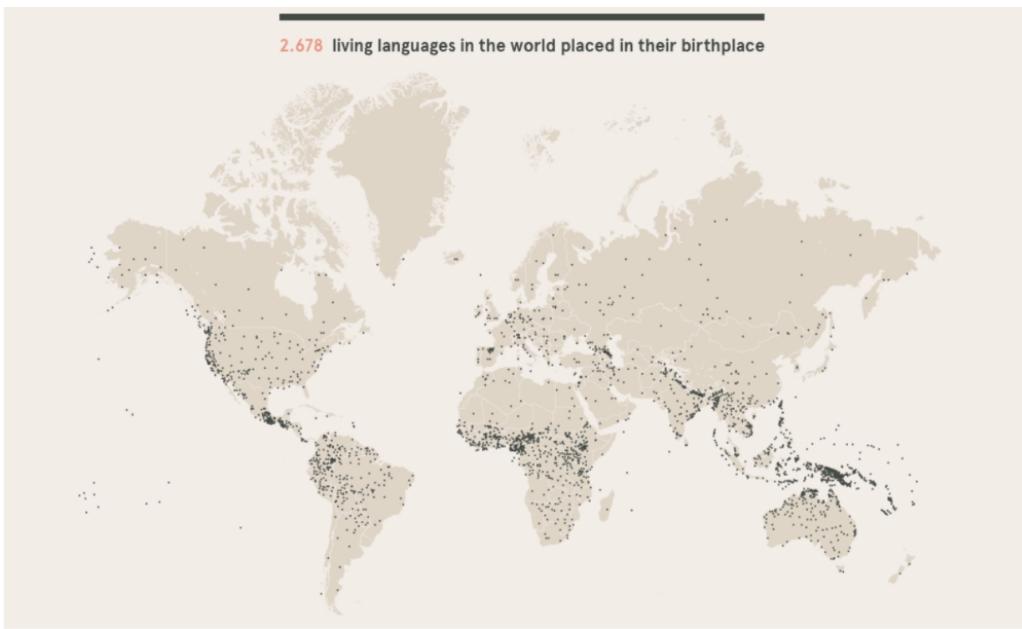
비주얼리제이션 작성자: 미국 관리예산실(2016)

- (1) 세계적 주요 강대국에서 납세자와 세금이 어디로 가는지 시민들에게 의사소통하고자 시각화 사용
- (2) 차트 형식은 혁신적인 것도 아니었고 일반적이지만, 복잡하고 모호한 주제를 말없이 간단하고 명확히 제시

6) 언어의 확산 과정 맵

- “세상의 언어 수는 몇개고 어떻게 퍼지고 영향을 줄까?”

LANGUAGES IN THE WORLD

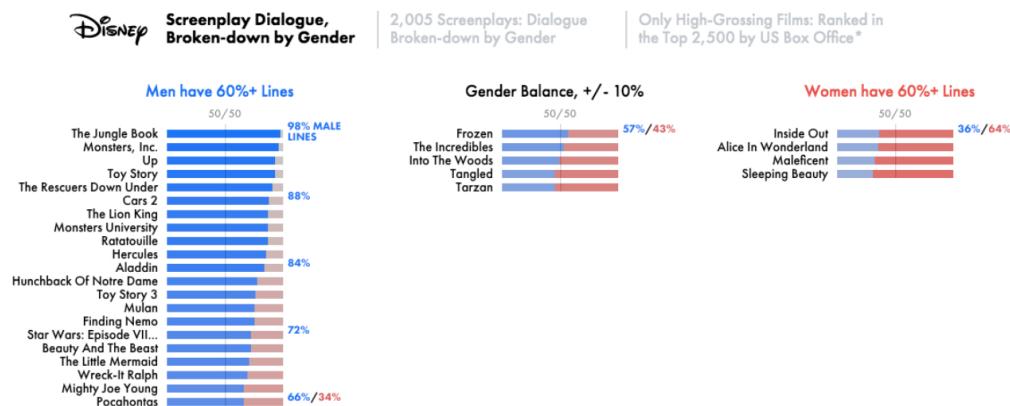


비주얼리제이션 작성자: Density Design Lab

- (1) 세계의 언어를 대화형 맵과 그래프로 표시
- (2) 총 2678개의 언어의 기원, 인구수, 다른 언어와의 관계 표현

7) 영화라는 대중문화에서의 성별 분류 맵

- "영화에서의 남성과 여성의 비중은?"



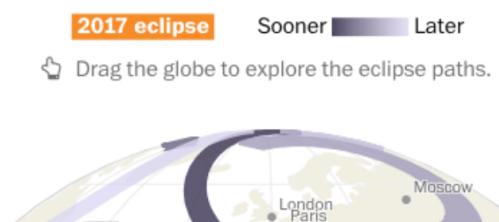
In January 2016, researchers reported that men are increasingly speaking more in movies.

비주얼리제이션 작성자: Hanah Anderson, Matt Daniels

- (1) 영화 역사상 가장 뛰어난 영화 2000편의 남성과 여성 대사를 분류하여 성별 격차 시작화
- (2) 모든 장르에서 나타나는 성별의 절대적이고 극명한 불균형 존재 파악
- (3) 디지털 영화에 대한 분류, 사용자가 영화를 검색 및 필터 기능, 남성과 여성 역할에 대한 연령 편견 등 포함
- (4) 단순히 2000개의 대본 분석이 이슈가 아니라, 솔직한 투명성 기반 전달력이 핵심

8) 다가오는 모든 일식 맵

- "내 평생 남은 일식은 얼마나 될까?"



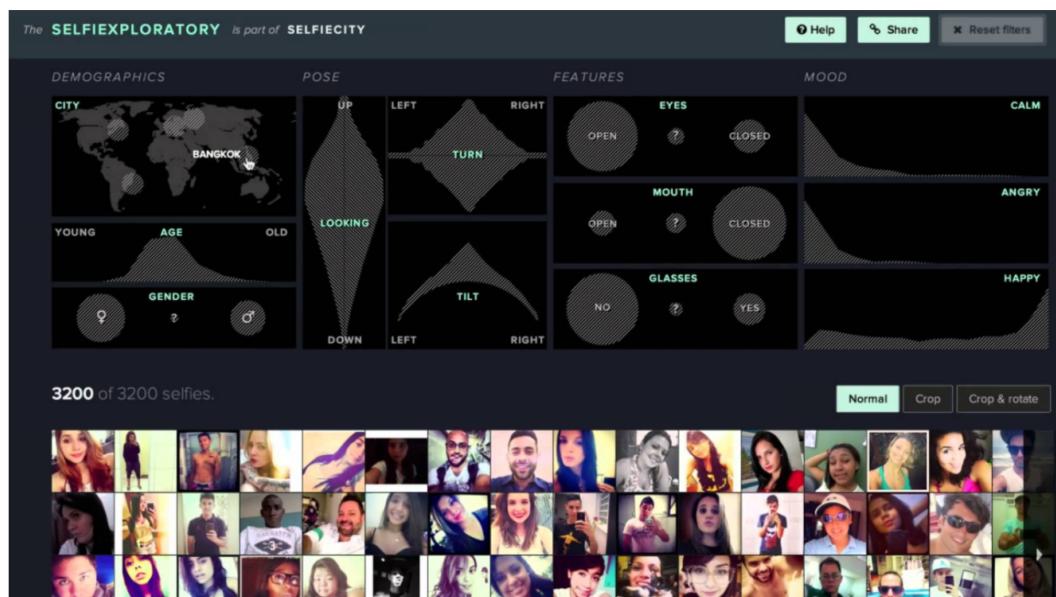


- (1) 일식의 경로와 2080년까지의 모든 미래 일식 경로 표현
 (2) 개기일식의 경로, 일어날 지점과 시점, 일생에 몇 번의 일식이 남았는지 표현

1.2 데이터의 인텔리전트 시각화

1) 셀카 트랜트 맵

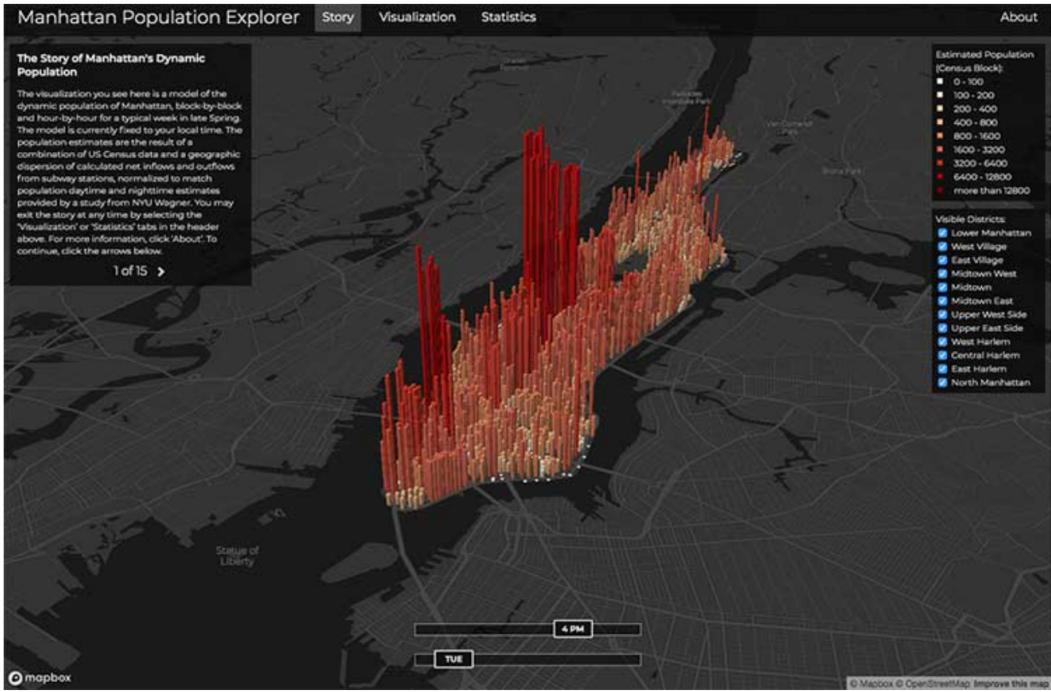
- "사람들은 어떻게 셀카를 찍을까?"



- (1) 셀카 데이터의 광범위한 뷰를 하나의 조국가적 컨텍스트로 표현
 (2) 사람이 어떻게 셀카를 찍는지 세계 곳곳의 12만장 셀카 사진 분석
 (3) 도시별 머리 기율이는 각도, 포즈, 연령 및 성별에 따른 웃는 빈도 트랜드 파악 가능
 (4) 셀카는 젊은 사람들이 찍는다는 건 별로 놀라운 사실이 아님
 (5-1) 상파울루의 여성들은 다른 지역에 비해 머리를 훨씬 많이 기울임
 (5-2) 방콕에서는 모두 웃으며 찍음
 (5-3) 셀카는 글로벌 현상이 아닌 특정 지역의 문화 산물로 널리 퍼져 있지 않음

2) 뉴욕시의 보이지 않는 심장 박동 맵

• "맨해튼의 상권은 어떨까?"



(1) 맨해튼은 미국에서 가장 인구 밀도가 높은 지역으로 주중 유입인구 2백만명

(2) 하루동안 일어나는 맨해튼의 이민 축소판 분석으로 한주에 걸친 시간변화 표현

3) 골휴일의 갑사함 맵

• "오늘은 글로벌 축제?"

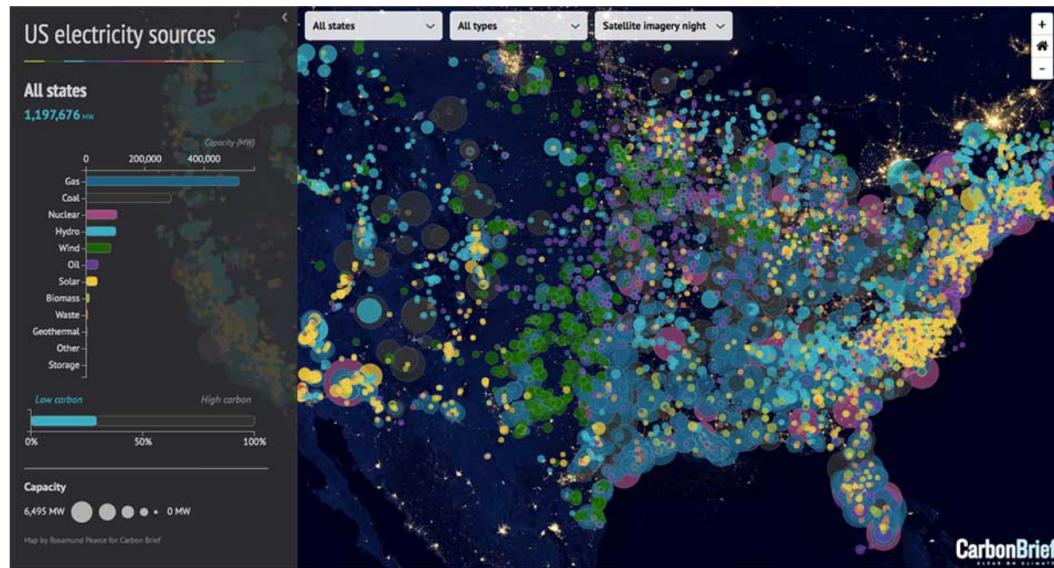


(1) 세계의 공휴일을 분석하여 연중 매일 다른 공유일 표현

(2) 각 나라별 공휴일의 갯수와 특정 날짜의 글로벌 경축정도 표현

4) 미국의 전력원 맵

- "내가 쓰는 전기는 어디서 오는걸까?"

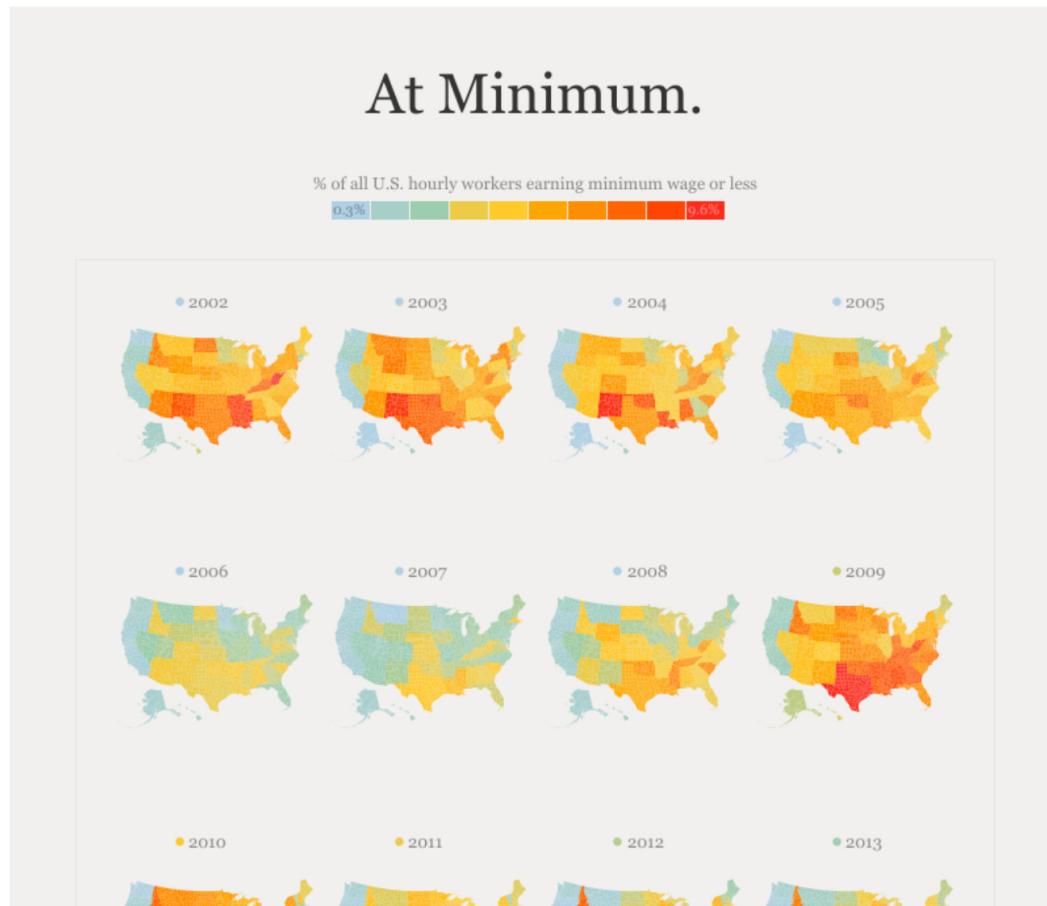


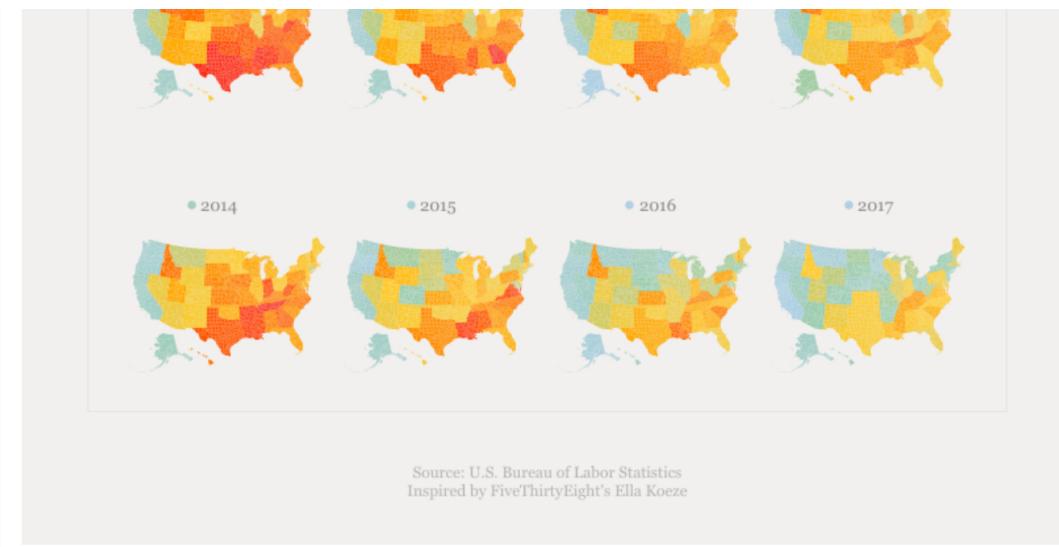
(1) 미국의 전기가 정확히 어디서 오는지 그리고 얼마나 많은 에너지가 생성되는지 분석

(2) 각각의 원은 개별 전력원, 유형별로 색상이 지정되어 있고, 원의 크기는 그 전력원에서 생성하는 전력 출력량

5) 최저임금 이하의 근로자 맵

- "미국의 경기는 좋아지고 있을까?"

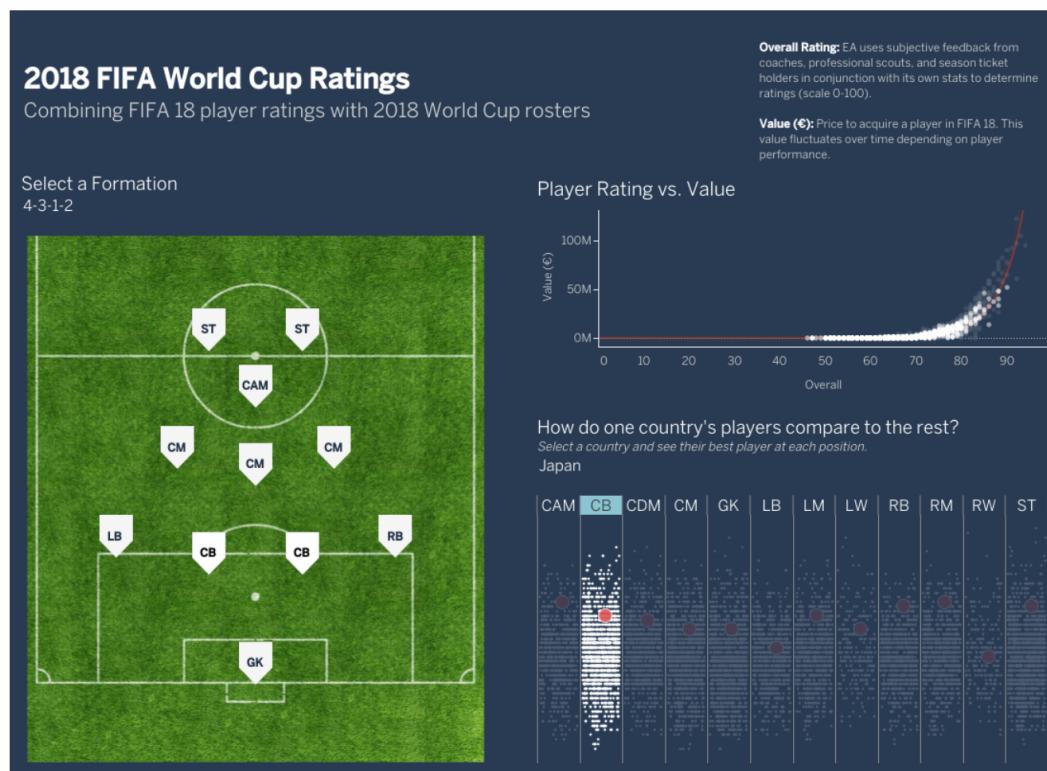




- (1) 최저 임금 시급 이하를 받는 모든 근로자의 백분율을 표시
- (2) 2002년까지 거슬러 올라가서 최소 임금 근로자의 연도별 스냅샷을 제공
- (3) 하나의 주 위에 마우스오버하면 그 지역의 근로자 백분율이 표시되고, 연도별, 주별로 비율이 상향 혹은 하향 추세인지 분석

6) FIFA 월드컵 팀과 선수정보 맵

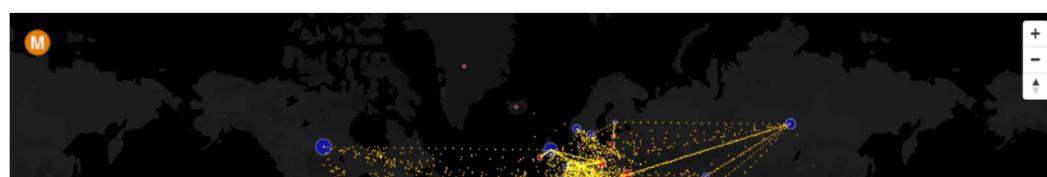
- "우리팀의 가치와 선수별 가치는?"

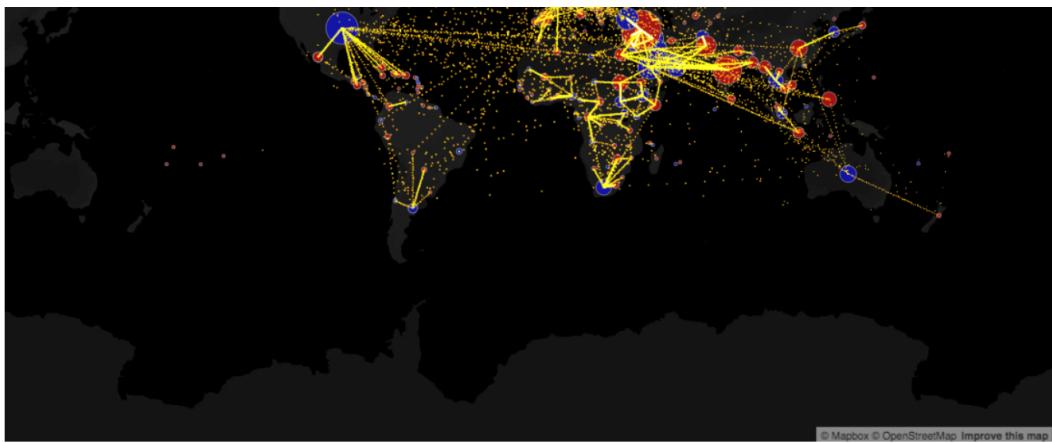


- (1) 2018 월드컵 토너먼트의 팀과 선수의 혼합 정보를 FIFA 18 선수 순위와 결합하여 표현
- (2) 각각의 포지션에 해당하는 선수들이 경쟁 선수들과 비교하여 어떤 위치에 있는지 분석

7) 세계 국가간 인구 이동 맵

- "우리나라의 인구는 정말 감소하는가?"

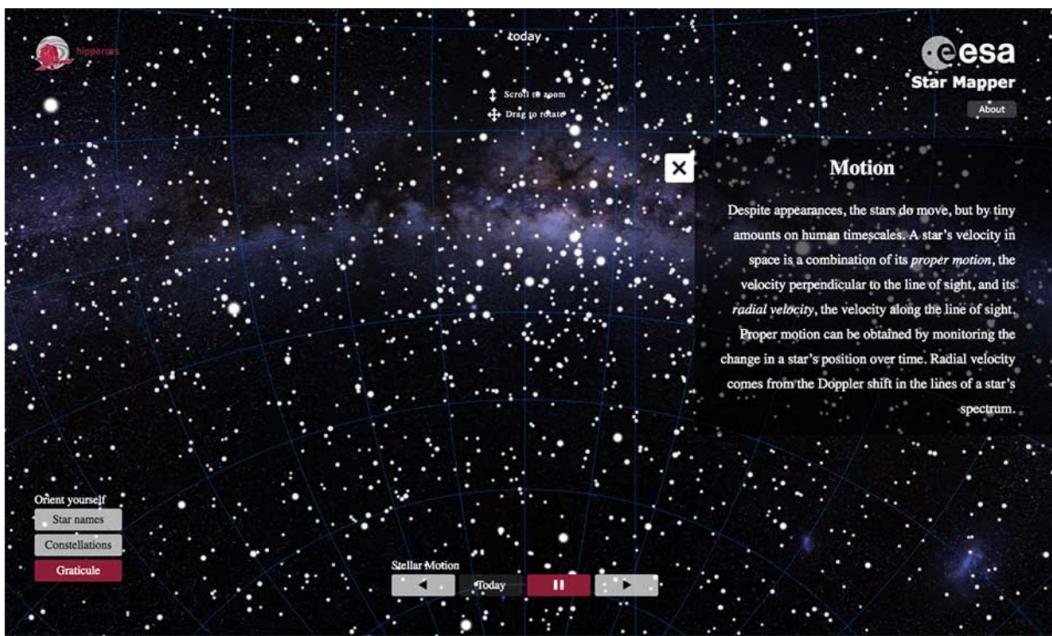




(1) 푸른색 원은 유입 순 이동 인구수를 양수로, 붉은 원은 유출 인구수를 음수로 표현

8) 별 지도

- "하늘을 못본다면 PC로라도?"



(1) 별자리, 별의 밝기, 은하수 내 위치 등 별자리의 과학적 가치 분석

(2) 59,921개의 별을 보여주며, 밤하늘을 탐색하기 위해 이동도 가능

9) 예술가들의 작품 정리 맵

- "자기|PR도 예술가처럼?"

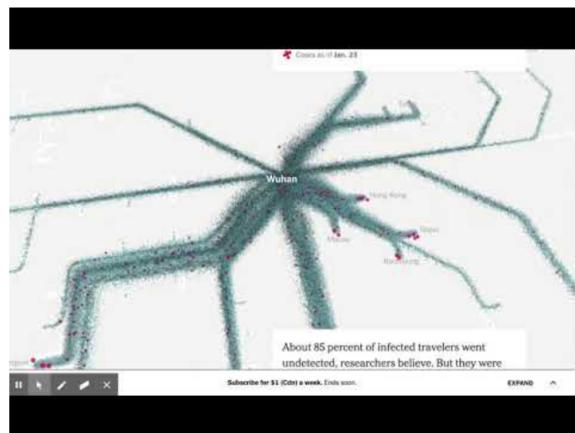




- (1) 특정 도시에서 작품 위치를 안내하고, 작품이 만들어진 스토리와 함께 작품 감상 가능
- (2) 기본 맵에 맵과 데이터 간의 사용자 지정 레이블과 색상의 조화를 사용하여 비주얼리제이션을 실감 나게 표현

10) 코로나 정리 맵

- "자기PR도 코로나처럼?"



- 사회적 거리두기 효용

- References:

- <https://bigxdata.io/18/?q=YToxOntzOjEyOiJrZXI3b3JkX3R5cGUIo3M6MzoiYWxsIjg&bmode=view&idx=4211696&t=board>
- <https://www.tableau.com/ko-kr/learn/articles/interactive-map-and-data-visualization-examples>

1.3 기술통계 분석의 한계

"데이터분석에 이 굴레들이 어떻게 쓰이는지 정리 해 보면.."

데이터 분석	목적	대응
데이터 시각화	데이터가 어떻게 생겼는지 알고 싶다	<ul style="list-style-type: none"> (1) 전체 데이터를 한눈에 확인 <ul style="list-style-type: none"> - 점그림, 선그림, 영역그림 - 막대그림, 등고선그림, 분포그림(히스토그램) → 통계 사용 (2) 데이터를 뿌리고 통계를 계산하기 위해 데이터 값 하나하나를 표현 <ul style="list-style-type: none"> → 확률/컴퓨터 사용
기술적 분석	데이터가 어떻게 생겼는지 알고 싶다	<ul style="list-style-type: none"> (1) 전체 데이터 특성을 몇 개의 숫자들로 확인 <ul style="list-style-type: none"> → 통계 사용 (2) 통계를 계산하기 위해 데이터 값 하나하나를 표현 <ul style="list-style-type: none"> → 확률/컴퓨터 사용
상관관계/인과관계	여러종류 데이터끼리의 관계를 알고 싶다	<ul style="list-style-type: none"> (1) 각 데이터를 몇 개의 숫자들로 표현 <ul style="list-style-type: none"> → 통계 사용 (2) 표현된 숫자들을 비교 <ul style="list-style-type: none"> → 확률/통계 사용
통계추론	일부 데이터로 전체를 알고 싶다	<ul style="list-style-type: none"> (1) 일부 데이터의 특성을 확인 <ul style="list-style-type: none"> → 통계 사용 (2) 반복적으로 실험 진행 및 통계치 재확인 <ul style="list-style-type: none"> → 컴퓨터 사용 (3) 전체 특성을 추론 <ul style="list-style-type: none"> → 통계 사용
알고리즘학습	전체 데이터로 미래를 알고 싶다	<ul style="list-style-type: none"> (1) 데이터의 관계를 수학적으로 표현 <ul style="list-style-type: none"> → 확률/통계/함수/컴퓨터 사용 (2) 미래를 예측한 후 정확성 확인 <ul style="list-style-type: none"> → 확률/통계/컴퓨터 사용
가설검정(A/B Test)	뭔가 진실과 가까운 의사결정을 하고 싶다	<ul style="list-style-type: none"> (1) 기존 데이터와 새로운 데이터 비교를 위해 숫자들로 표현 <ul style="list-style-type: none"> → 통계 사용 (2) 표현된 숫자들을 비교 <ul style="list-style-type: none"> → 확률/통계/컴퓨터 사용

1) 같은 통계량 다른 그래프



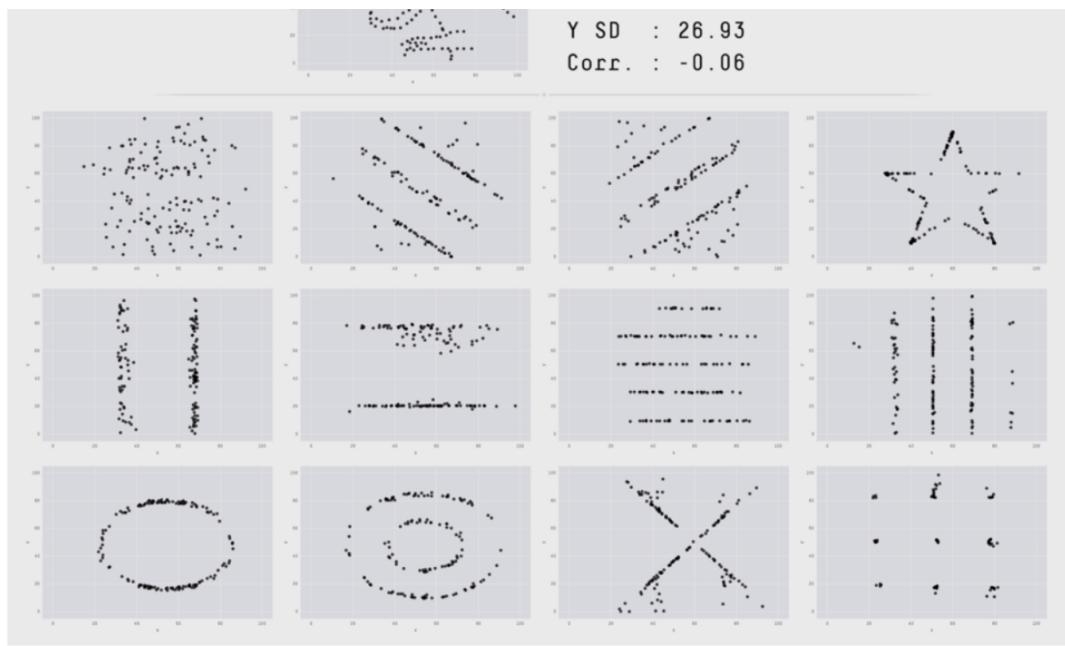


그림 1.23 데이터 공룡(Datasaurus Dozen) – 데이터셋은 소수점 두 자릿수 기준으로 같은 요약 통계값(평균, 표준편차, 상관계수)을 보여주지만, 시각적 패턴은 모두 다르다
(<http://newsjel.ly/archives/newsjelly-report/visualization-report/12282>)

2) 앤스콤 시작화

I	II	III	IV
10 8.04	10 9.14	10 7.46	8 6.58
8 6.95	8 8.14	8 6.77	8 5.76
13 7.58	13 8.74	13 12.74	8 7.71
9 8.81	9 8.77	9 7.11	8 8.84
11 8.33	11 9.26	11 7.81	8 8.47
14 9.96	14 8.1	14 8.84	8 7.04
6 7.24	6 6.13	6 6.08	8 5.25
4 4.26	4 3.1	4 5.39	19 12.5
12 10.84	12 9.13	12 8.15	8 5.56
7 4.82	7 7.26	7 6.42	8 7.91
5 5.68	5 4.74	5 5.73	8 6.89

Mean of X	11.0	Correlation between X and Y	0.875
Variance of X	10.0	Linear regression	$y=3.0+0.5x$
Mean of Y	7.5		
Variance of Y	3.75		

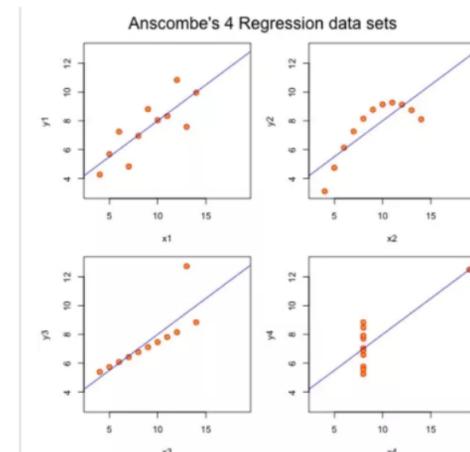


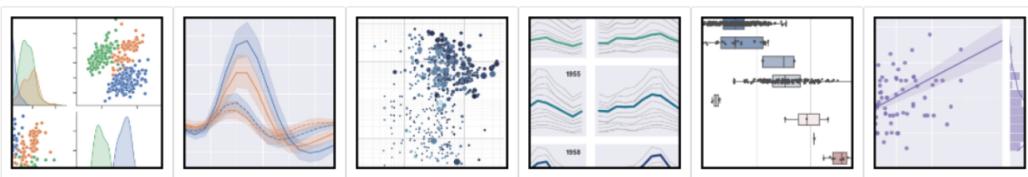
그림 1. 22 동일한 요약 통계를 가진 데이터셋 4개(왼쪽), 왼쪽 데이터셋 4개를 시각화한 결과(오른쪽)
(<http://newsjel.ly/archives/newsjelly-report/visualization-report/12282>)

"데이터의 정확한 이해를 위해 데이터 분석 과정에서 시각화를 필수적으로 활용해야!"

2 데이터시각화 종류와 도구

• 시각화 활용 목적:

- 변수 내 숫자(Number) 값의 분포(Distribution) 파악
- 변수 내 항목(Category) 별 값의 분포(Distribution) 파악
- 변수 값의 시간(Time Series) 변화에 따른 예측(Regression) 파악
- 변수들 간의 관계(Relation) 파악
- 변수들 간의 비교(Multiple) 파악



Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the introductory notes or the [full documentation](http://seaborn.pydata.org/introduction.html).

Contents

- Introduction

Features

- Relational: API | Tutorial

the paper. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see some of the things that you can do with seaborn, and then check out the tutorial or API reference to find out how.

To see the code or report a bug, please visit the GitHub repository. General support questions are most at home on stackoverflow or discourse, which have dedicated channels for seaborn.

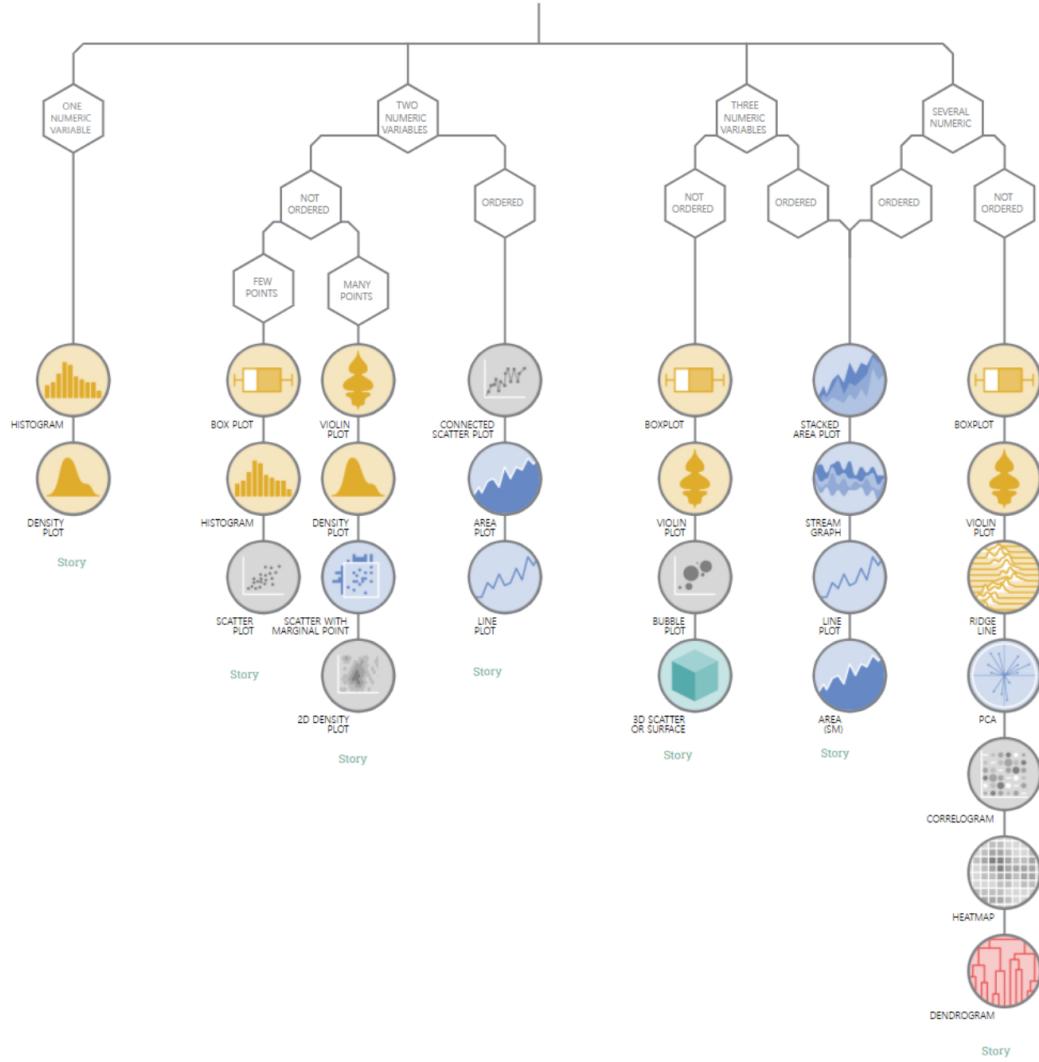
- Release notes
- Installing
- Example gallery
- Tutorial
- API reference

- DISTRIBUTION: API | Tutorial
- Categorical: API | Tutorial
- Regression: API | Tutorial
- Multiples: API | Tutorial
- Style: API | Tutorial
- Color: API | Tutorial

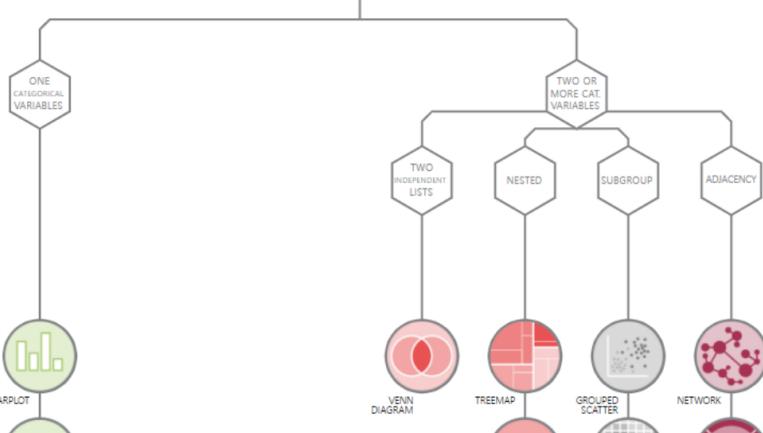
(<https://seaborn.pydata.org/>)

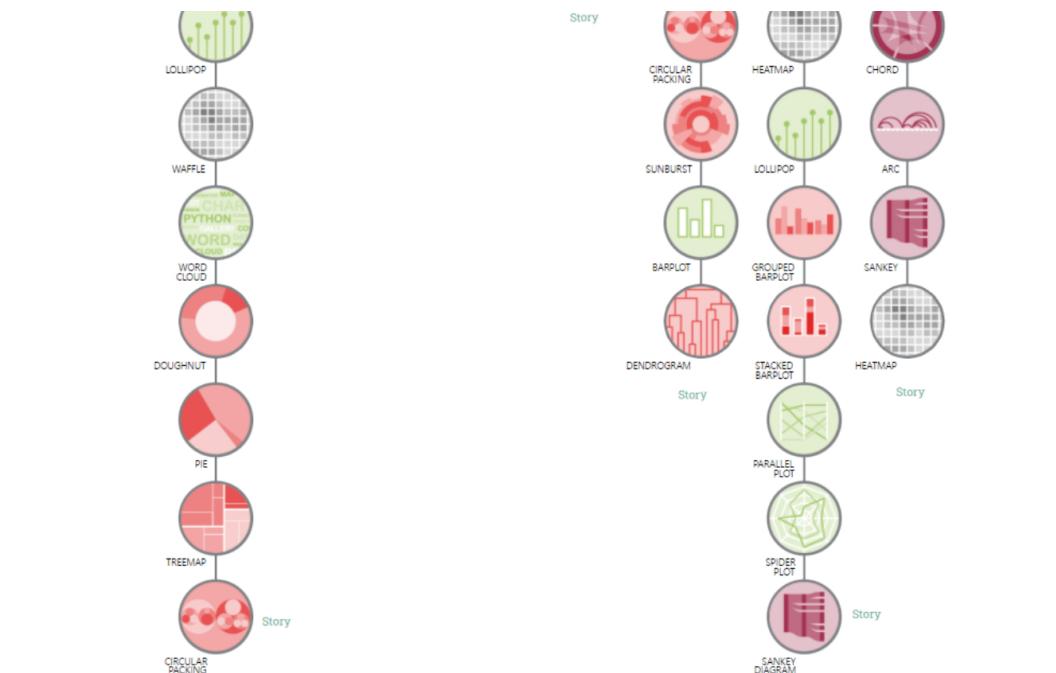
• 시작화 종류:

Numeric Categoric Num & Cat Maps Network Time series

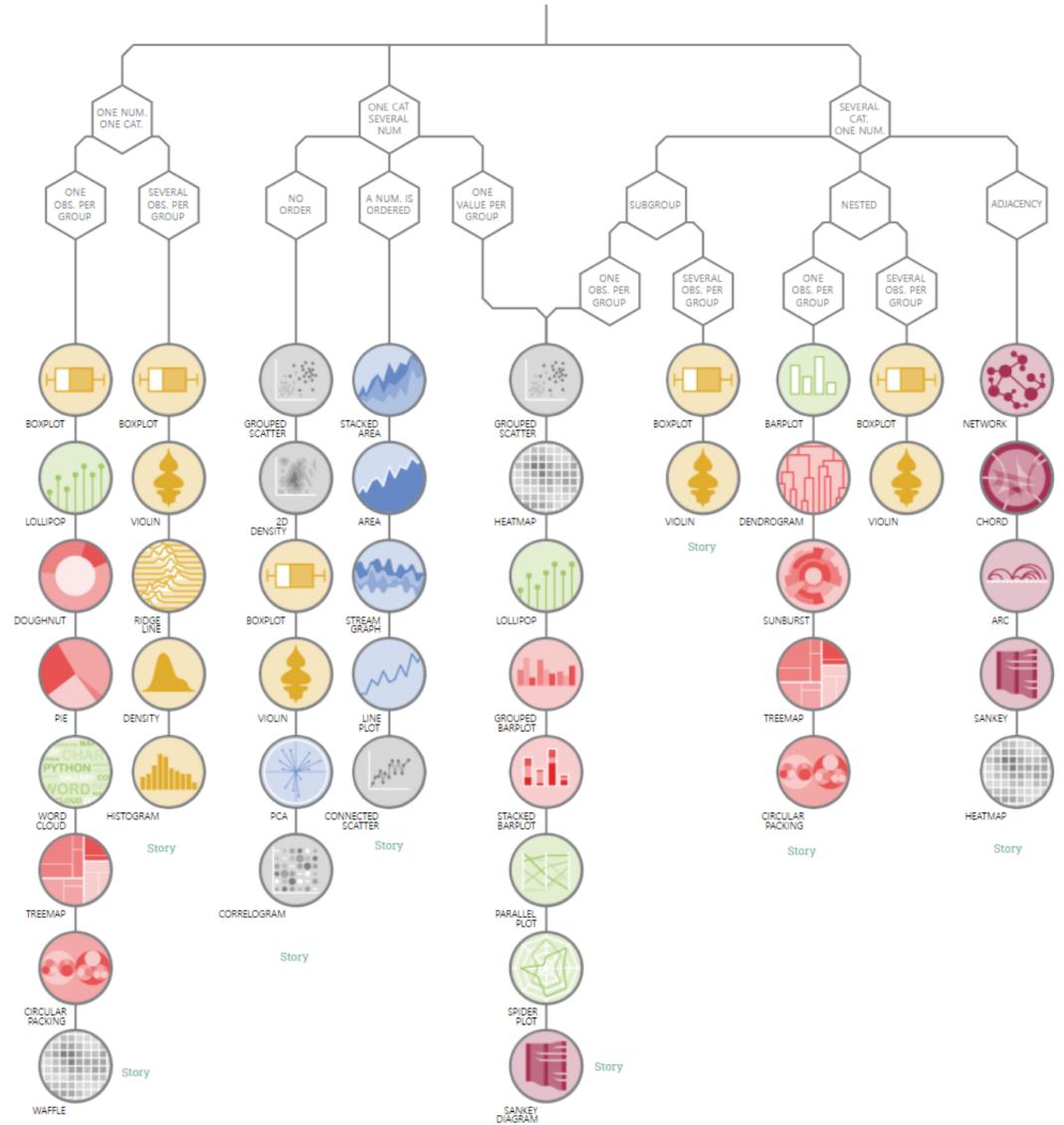


Numeric Categoric Num & Cat Maps Network Time series

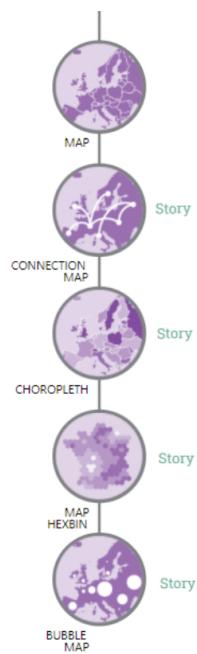




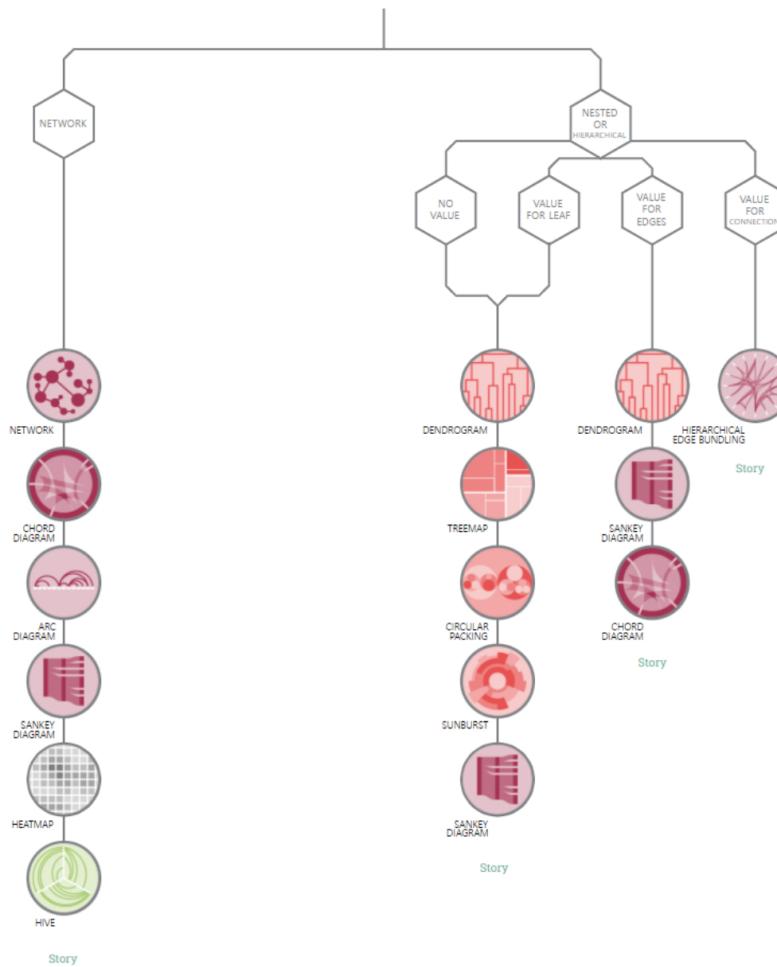
Numeric Categoric Num & Cat Maps Network Time series



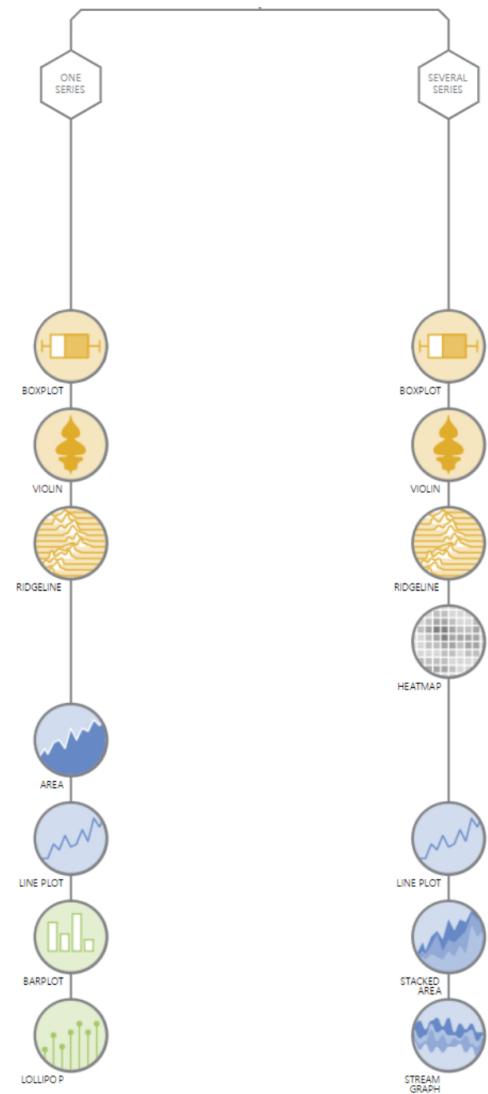
Numeric Categoric Num & Cat Maps Network Time series



Numeric Categoric Num & Cat Maps **Network** Time series



Numeric Categoric Num & Cat Maps Network **Time series**



(<https://www.data-to-viz.com/>)

- Python 시각화 도구(라이브러리/패키지):

Library	Description
pandas	Anaconda 설치시 기본 설치 라이브러리로 별도의 라이브러리 불러오기 없이 사용 가능
matplotlib	pandas 시각화와 함께 가장 많이 쓰이는 라이브러리로 pandas의 데이터 프레임을 시각화 세부 설정시 함께 사용하는 편
seaborn	matplotlib을 기반으로 다양한 색 테마, 차트 기능을 추가한 라이브러리
plotnine	R의 ggplot2에 기반해 그래프를 작성하는 라이브러리
folium	지도 데이터(Open Street Map)에 leaflet.js를 이용해 위치정보를 시각화하는 라이브러리
pyecharts	중국 바이두에서 데이터 시각화를 위해 만든 Echarts.js의 파이썬
plotly	인터랙티브 그래프를 그려주는 라이브러리로 R, Scala, Python, Java Script, Matlab 등에서 사용 가능