데이터 시각화 ggplot 뛰어넘기

김영진

SK텔레콤

발표 자료 웹 버전 보기

데이터 시각화

ggplot2 뛰어넘기

김 영 진

김영진

- 사회학 박사 (2015)
- SKTelecom
- Data Analytics Group

데이터 시각화?

이 자료는 RUCK에 공개될 예정입니다.

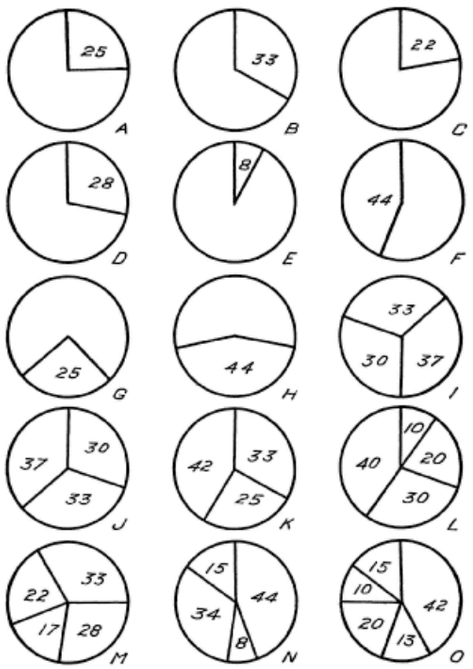
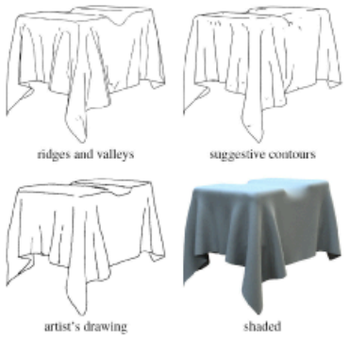
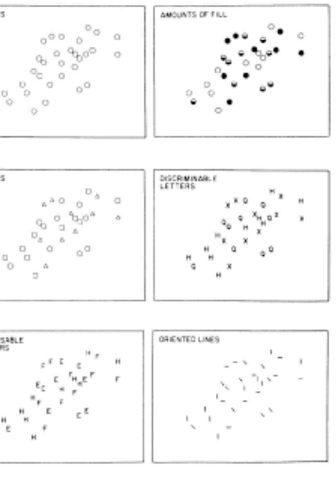
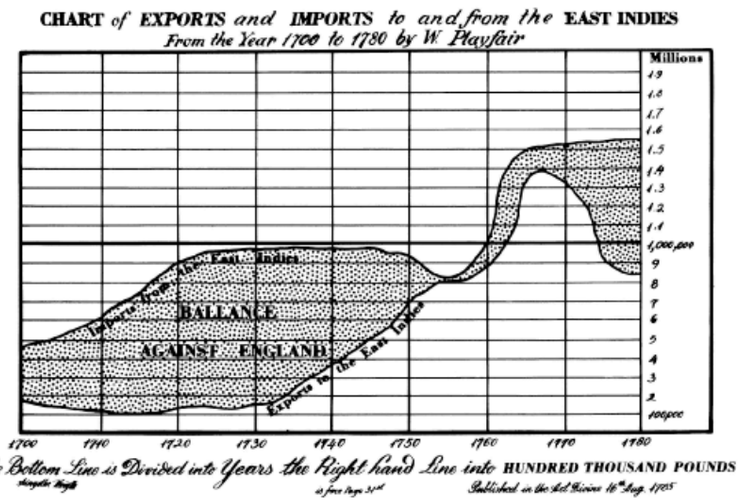


FIGURE I







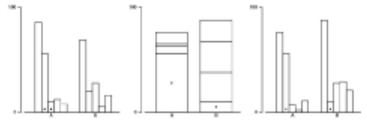


Figure 1: Stimuli for judgment tasks T1, T2 & T3. Subjects estimated percent differences between elements.

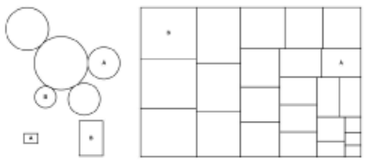
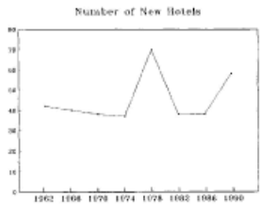


Figure 2: Area judgment stimuli. Top left: Bubble chart (T7), Bottom left: Center-aligned rectangles (T8), Right: Treemap (T9).



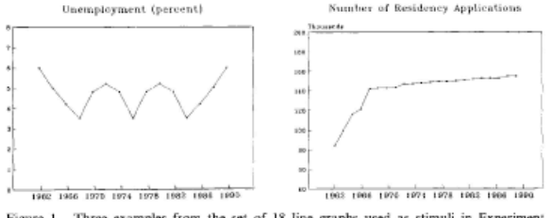
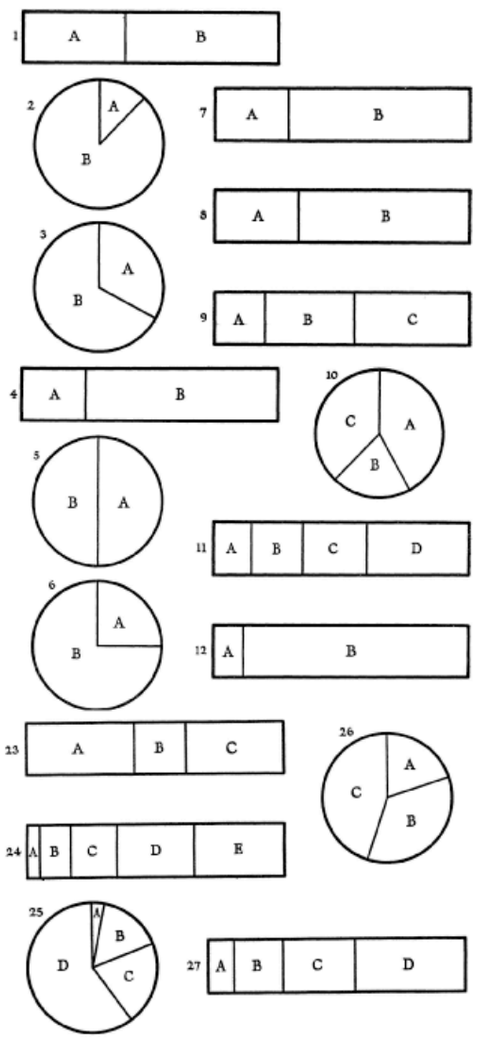
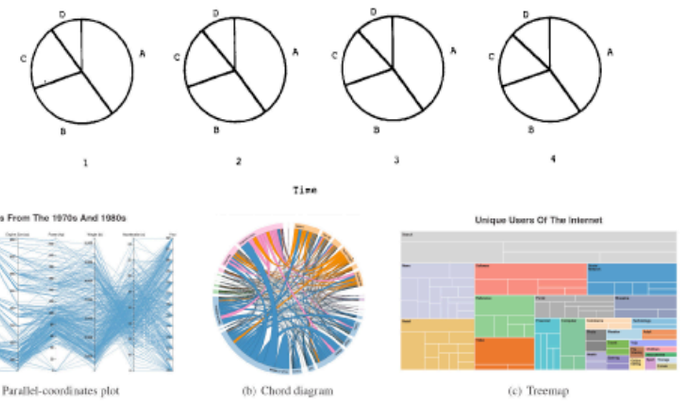


Figure 1. Three examples from the set of 18 line graphs used as stimuli in Experiment 1. The 'new hotels' graph in the upper left is an asymmetric graph with 8 data points and 3 trend reversals. The 'residency' graph in the upper right is asymmetric, has 22 data points, and no trend reversals. The bottom graph is symmetric, has 15 data points, and no 5 trend reversals





(a) Parallel-coordinates plot (b) Chord diagram (c) Treemap

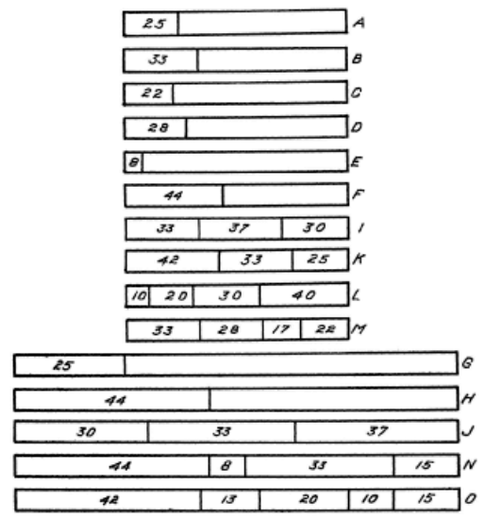


FIGURE II

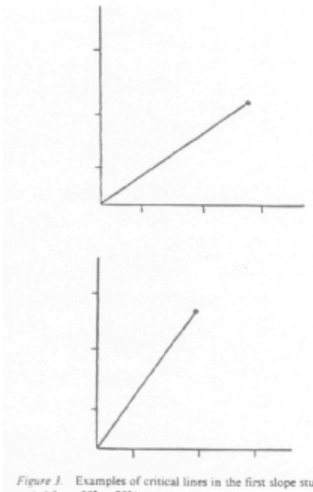


Figure 3. Examples of critical lines in the first slope study. (Angles varied from 20° to 70°.)

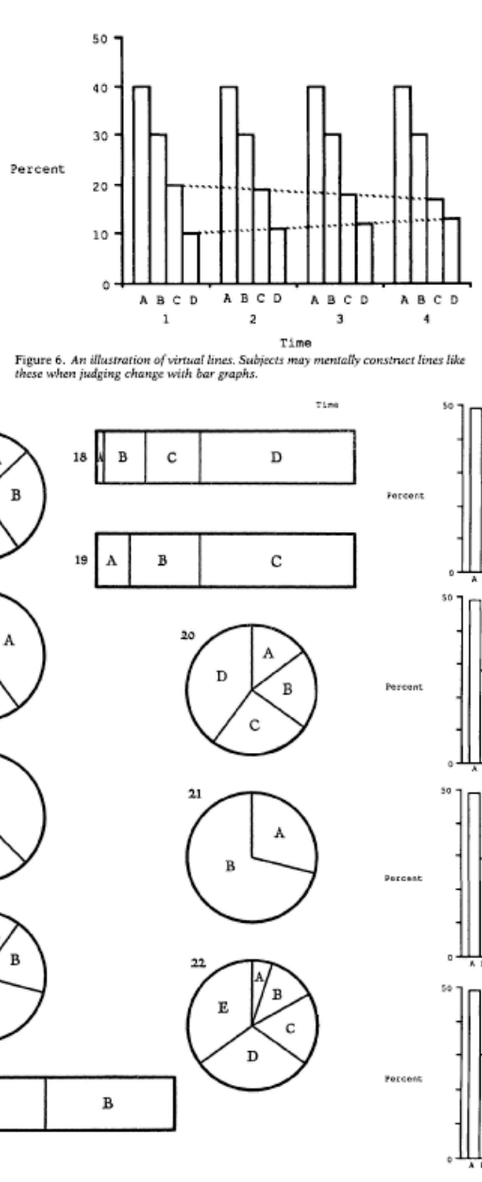
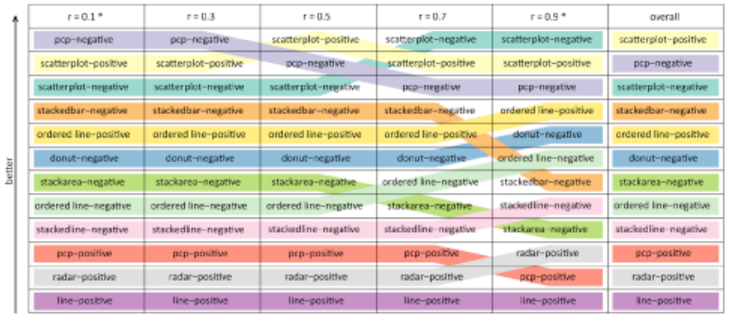


Figure 6. *An illustration of virtual lines. Subjects may mentally construct lines like these when judging change with bar graphs.*



정보를 보기 좋게 표현하는 것?

MS 4551
Account of grain products, bread, beer, butter oil. Sumer, 32nd c. BC

John Tukey (1915-2000)

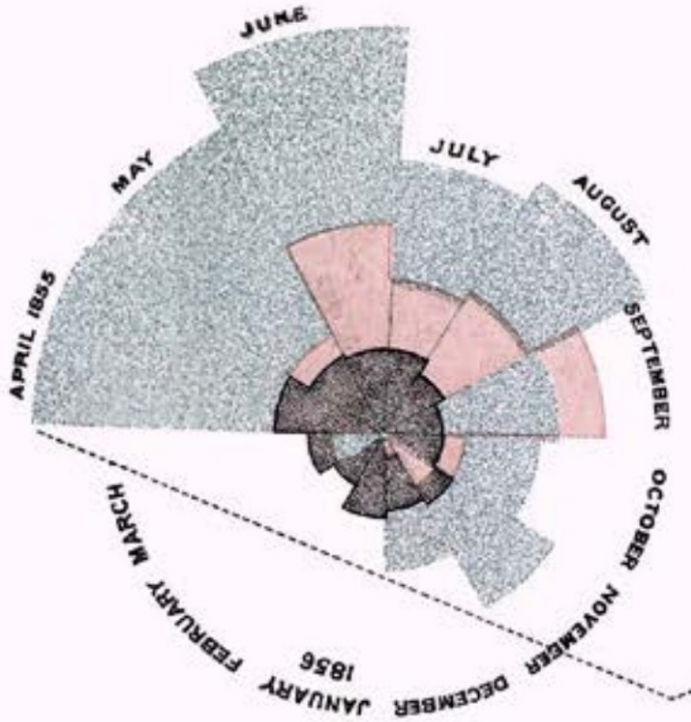
Exploratory Data Analysis
탐색적 데이터 분석

"데이터를 **탐색**하면서,
질문과 **가설**을 만들고,
분석의 **뼈대**를 만드는 **과정**"

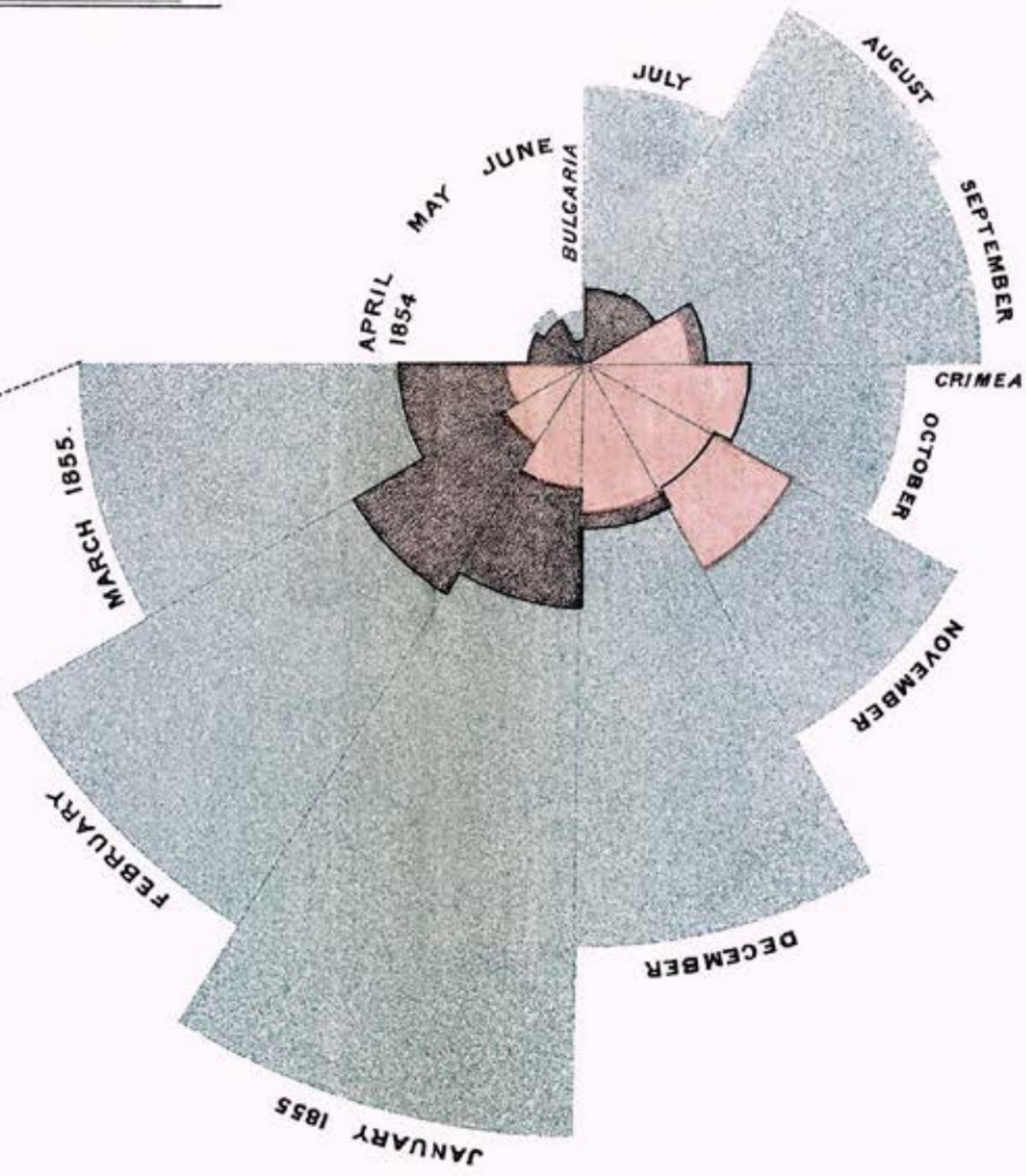
"형사가 **수사**를 할 때
여러 가지 **단서**들로
가설을 세우고
증거를 확립하는 것과 유사"

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.r 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 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red & the black lines enclosing them.

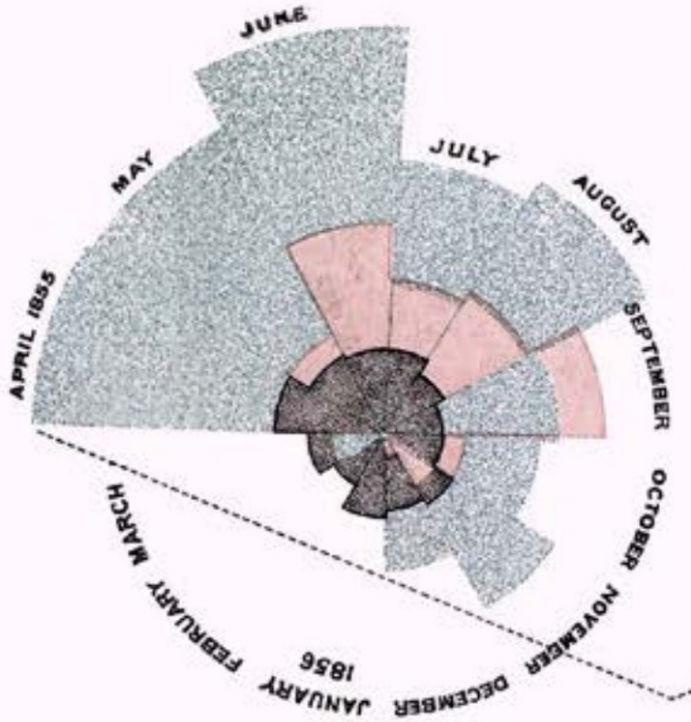
Florence Nightingale

플로렌스 나이팅게일
(1820~1910)

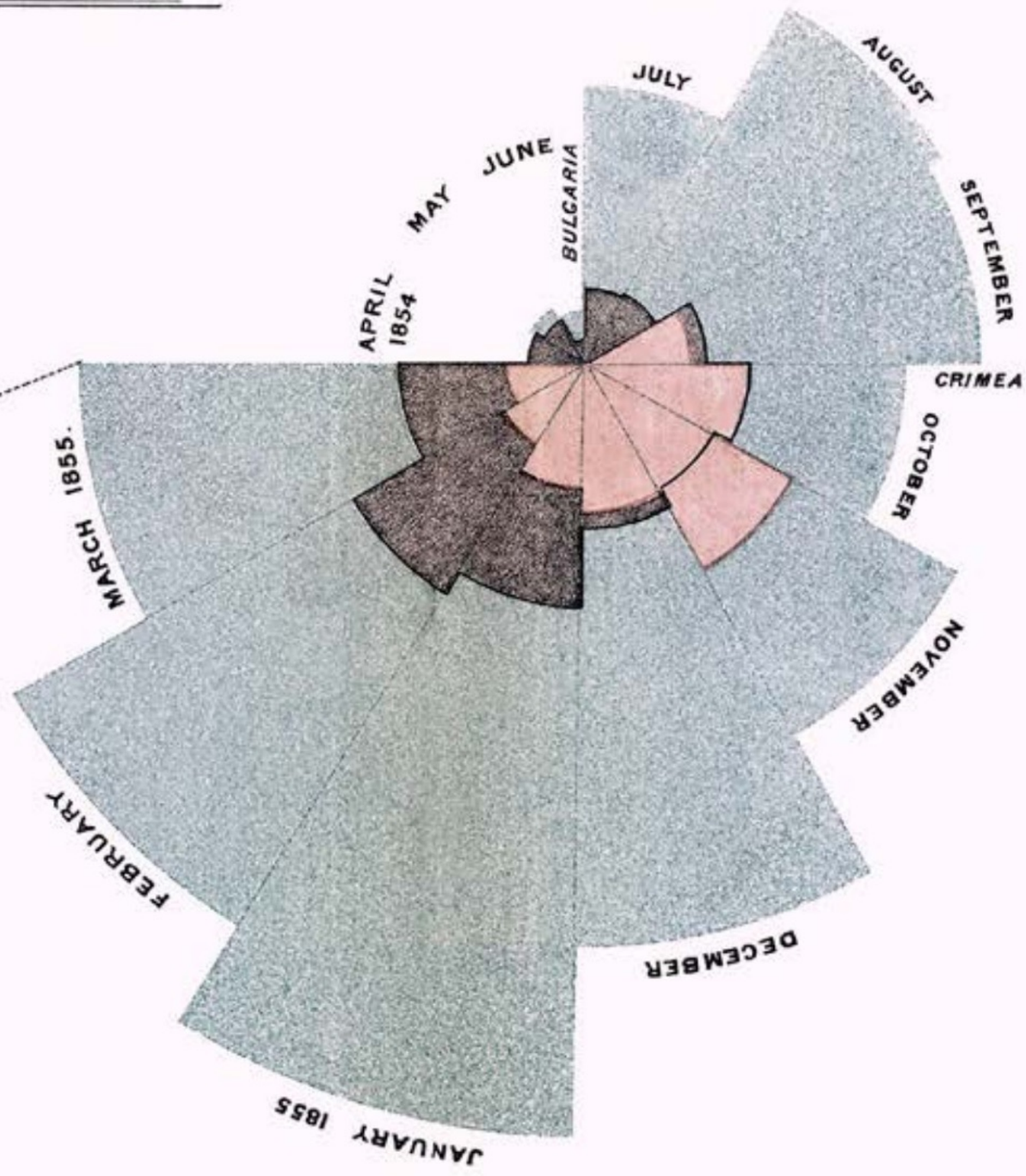
"나는 **간호**를 받는 사람들의
안녕을 위해 **헌신**하겠습니다."

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.r 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 1856, the blue coincides with the black.

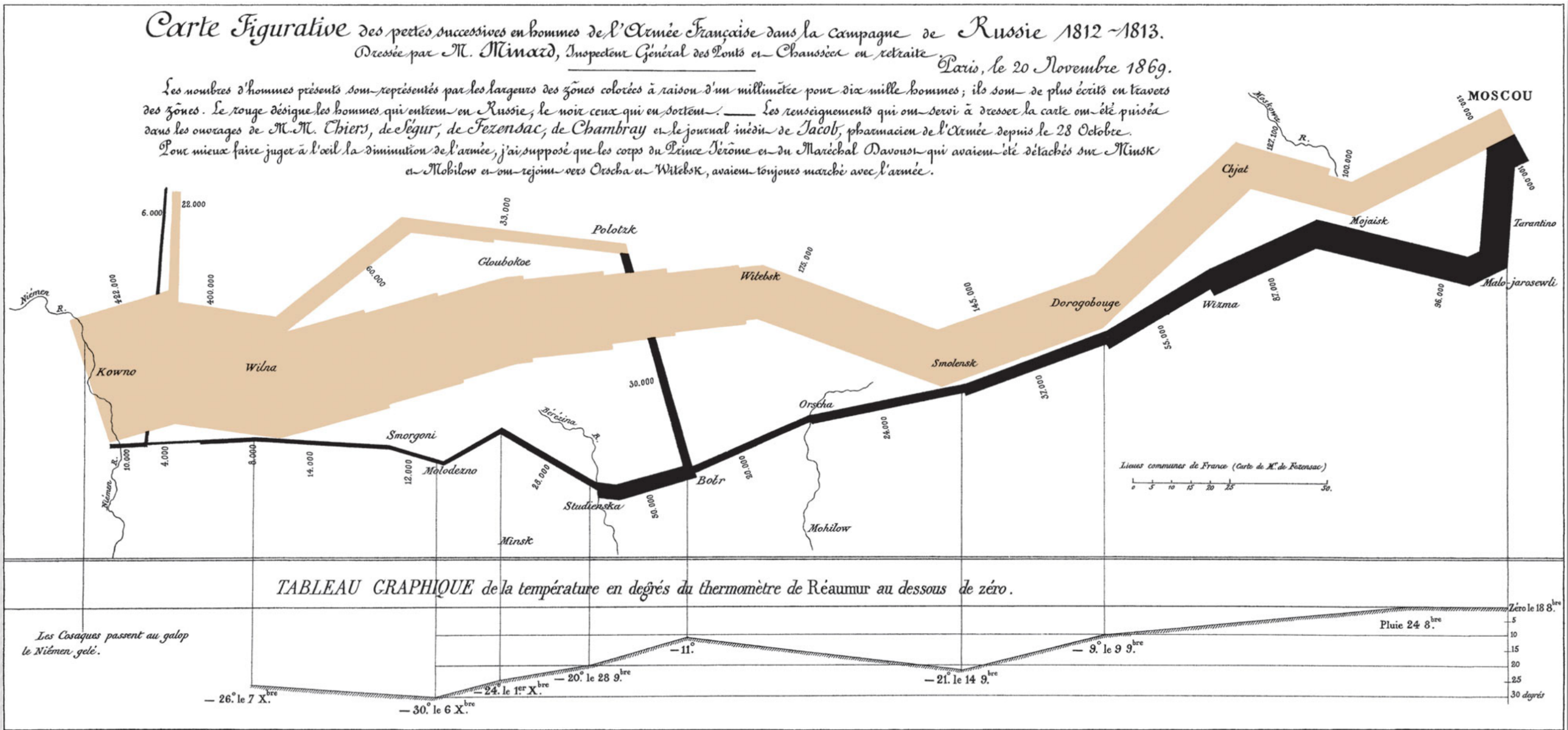
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

이 도표 한 장으로,
영국 군대는 전투력을 위해 전장에서
위생을 최우선으로 변경하였음

百聞而不如一見

백문이불여일견

March on Moscow - Charles Minard



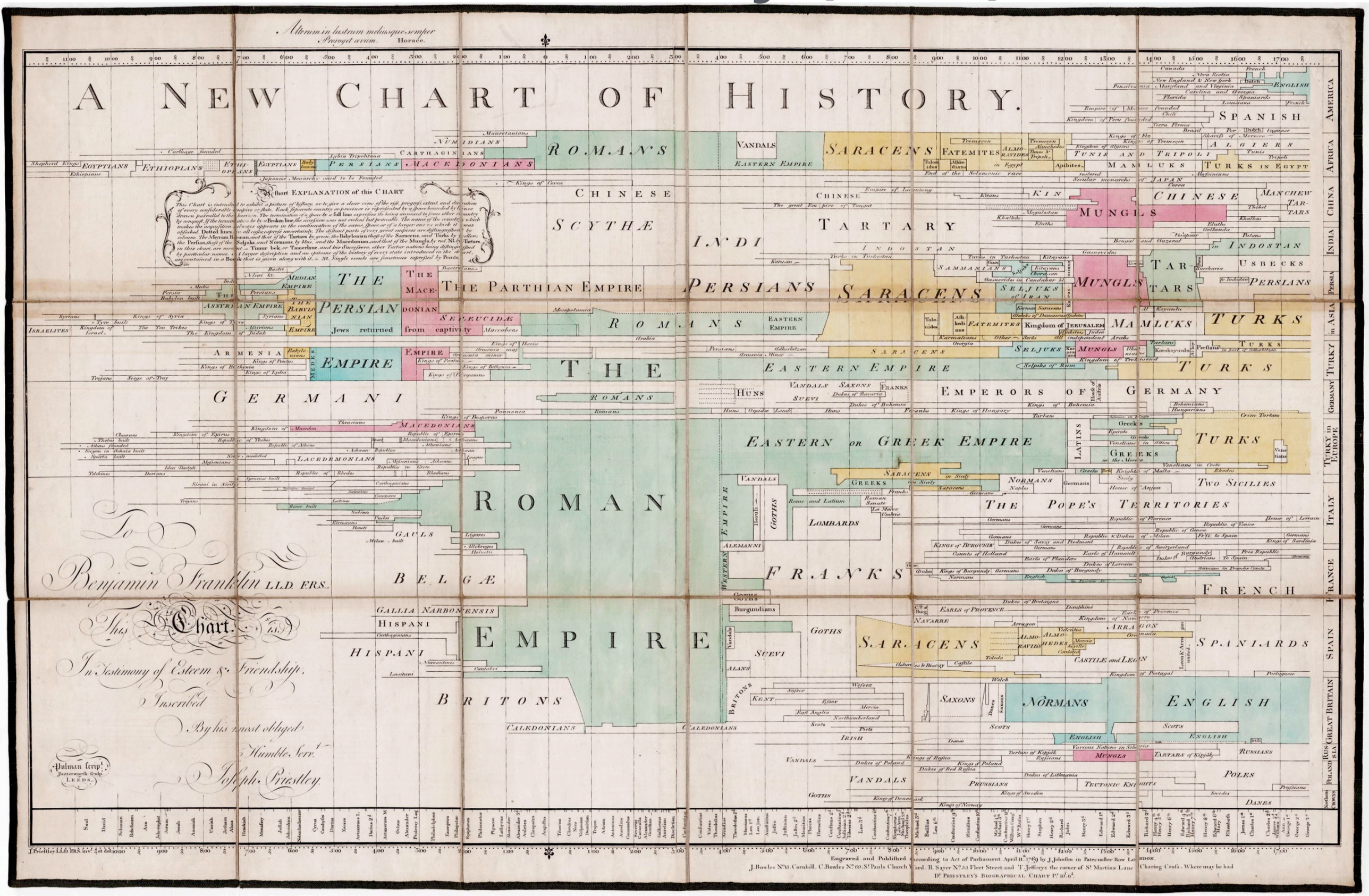
http://ecowest.org/tag/edward-tufte/

https://www.edwardtufte.com/tufte/posters

하나의 **그래프가** 한 권의 책보다
더 나은 **설명**과 **이해**

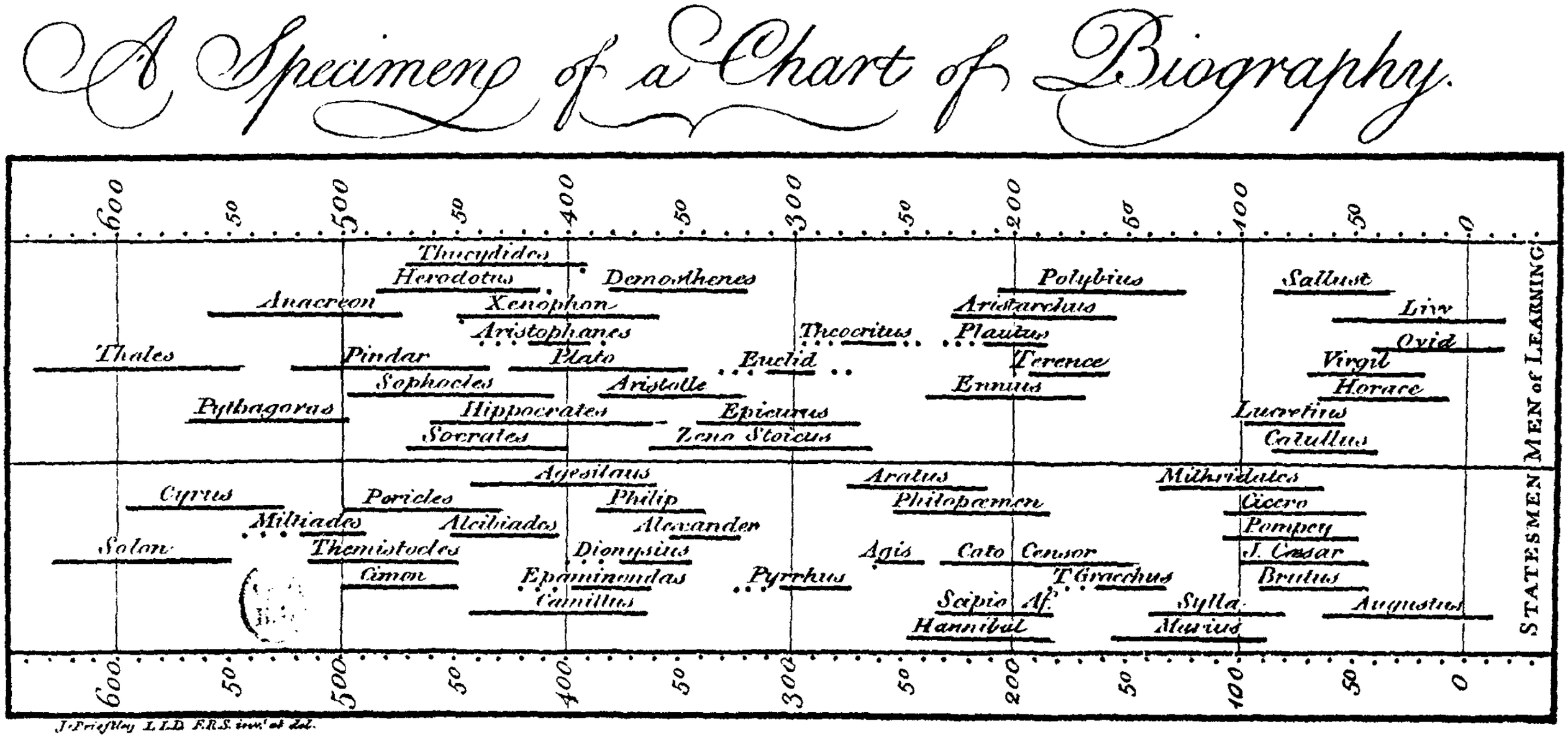
나폴레옹의 행진은 독창적 디자인?

A New Chart of History (1769)

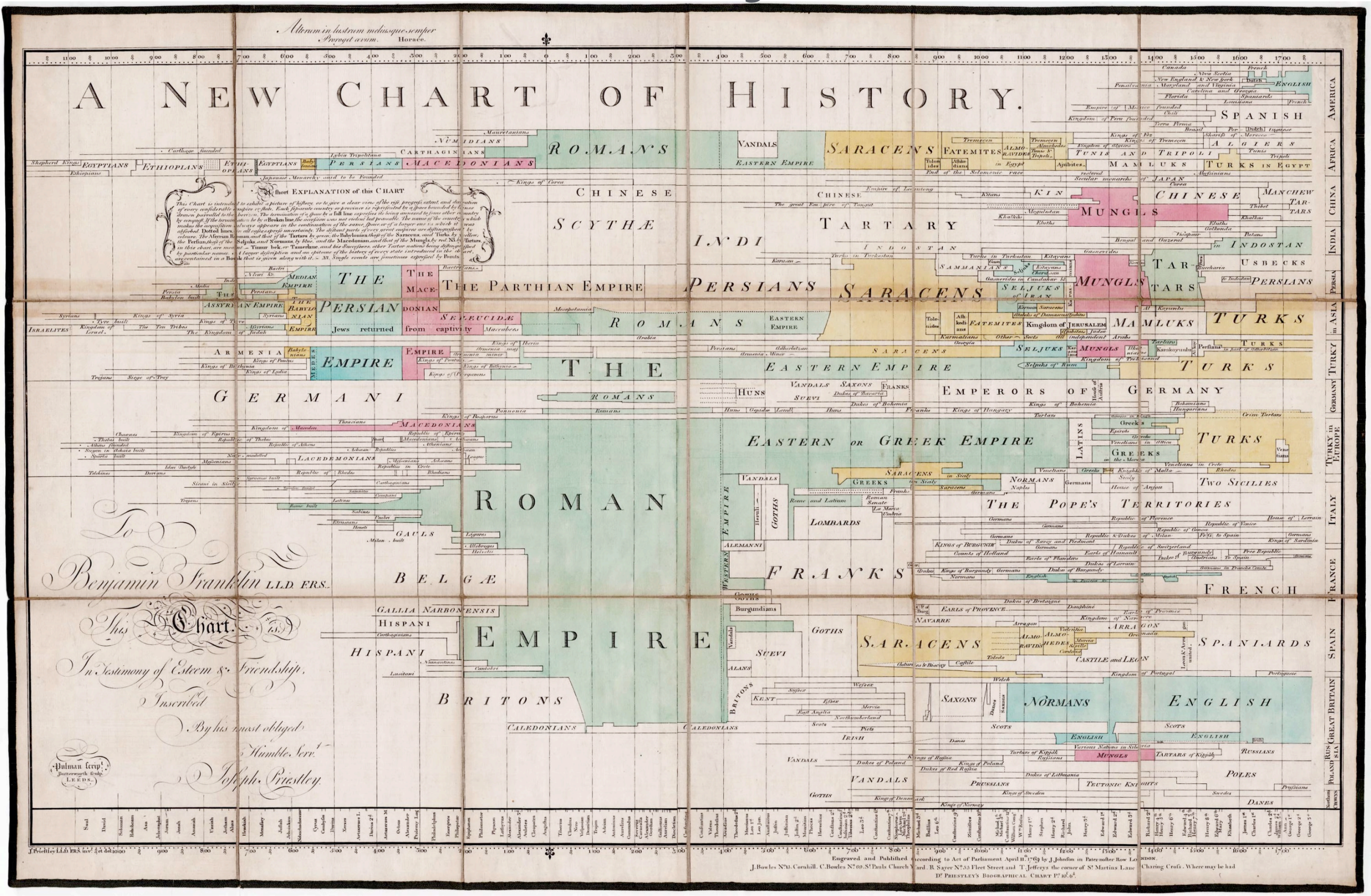


https://bostonraremaps.com/inventory/joseph-priestley-chart-of-history/

A Chart of Biography (1765)

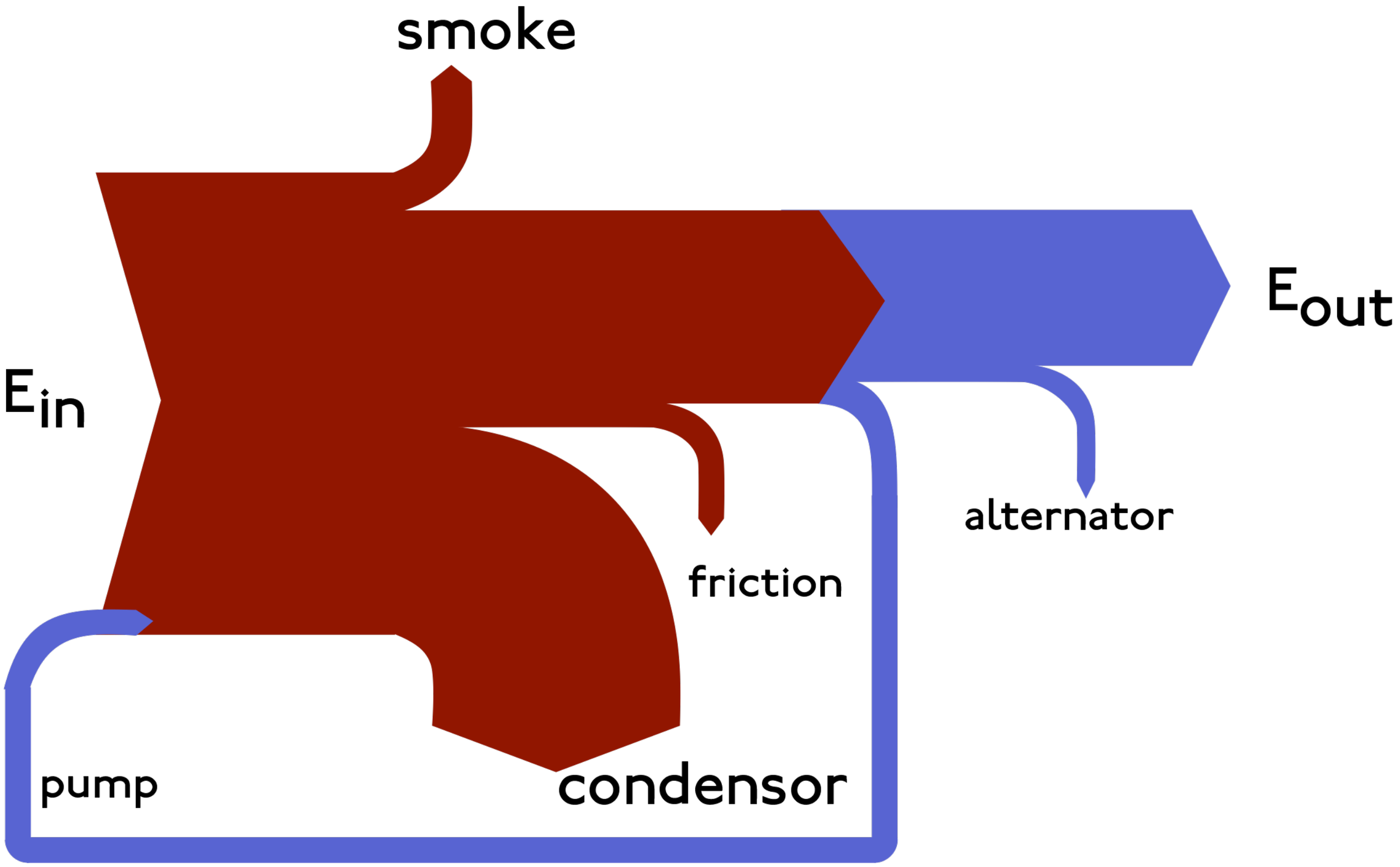


A New Chart of History (1769)

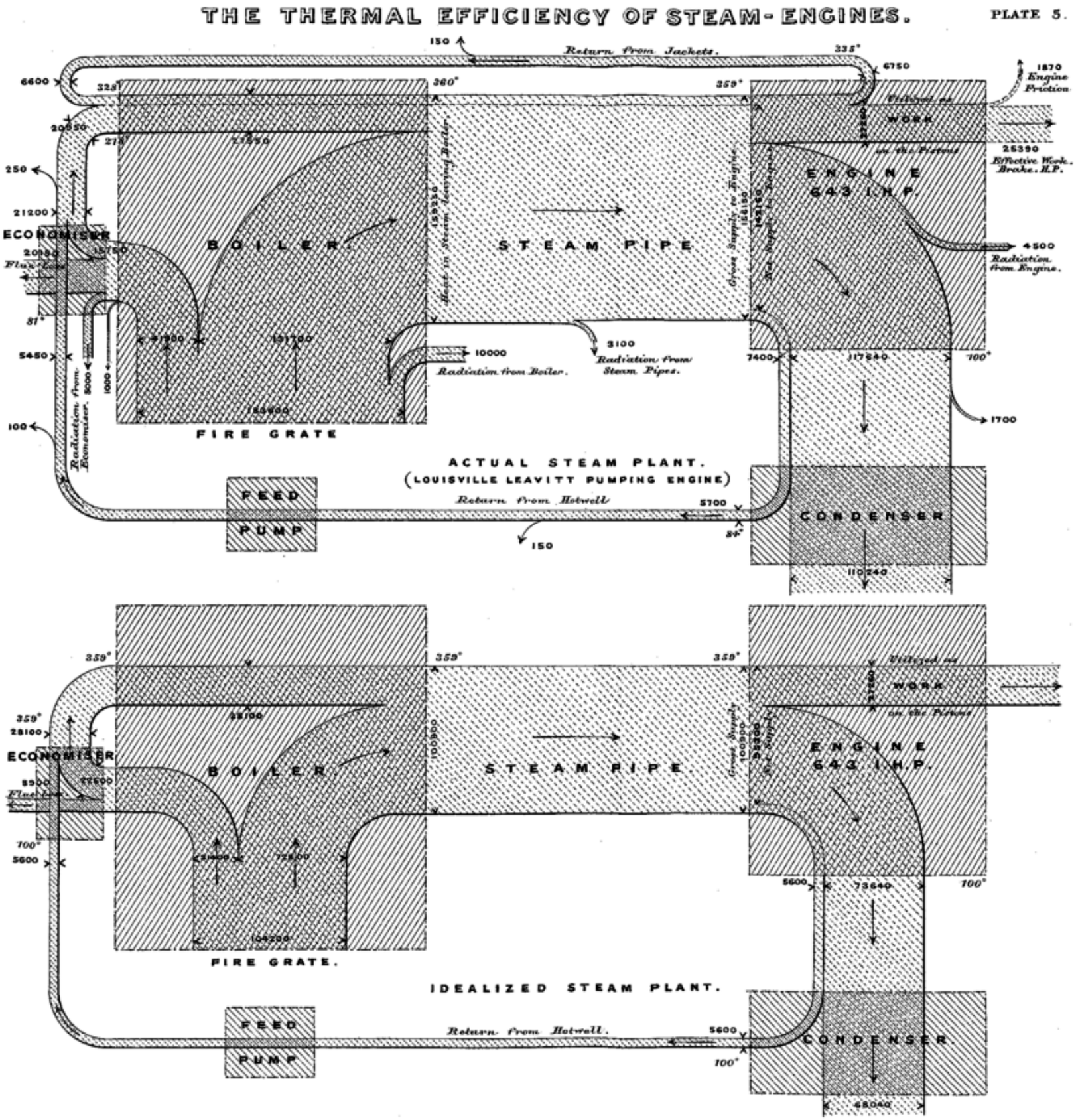


https://bostonraremaps.com/inventory/joseph-priestley-chart-of-history/

Sankey diagram

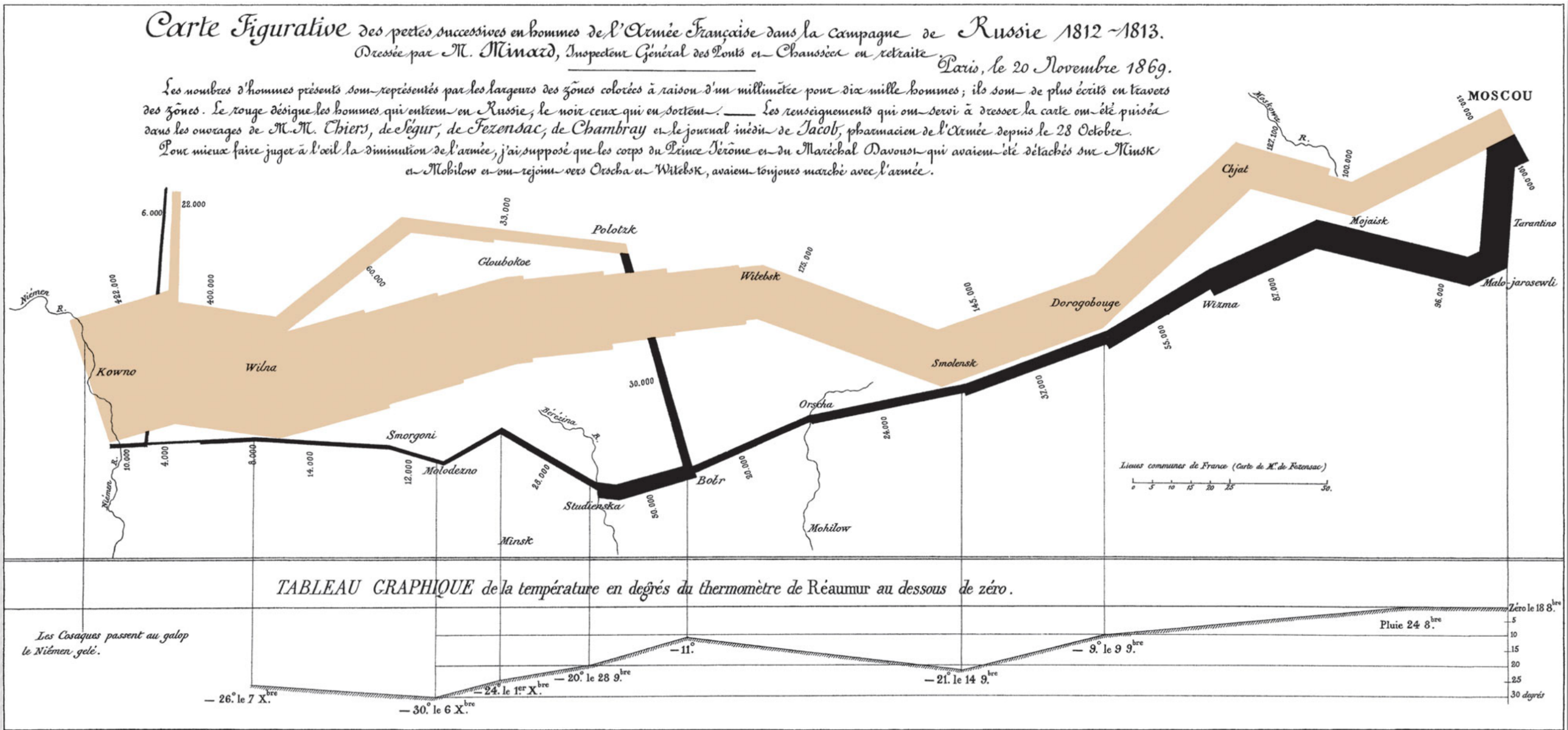


Sankey's original 1898



Minutes of Proceedings of The Institution of Civil Engineers Vol. CXXXIV, Session 1897-98 Part IV

March on Moscow (1869)- Charles Minard



http://ecowest.org/tag/edward-tufte/

https://www.edwardtufte.com/tufte/posters

시간의 흐름과 이벤트의 강도를 표현하고자 한 그래프

Streamgraph

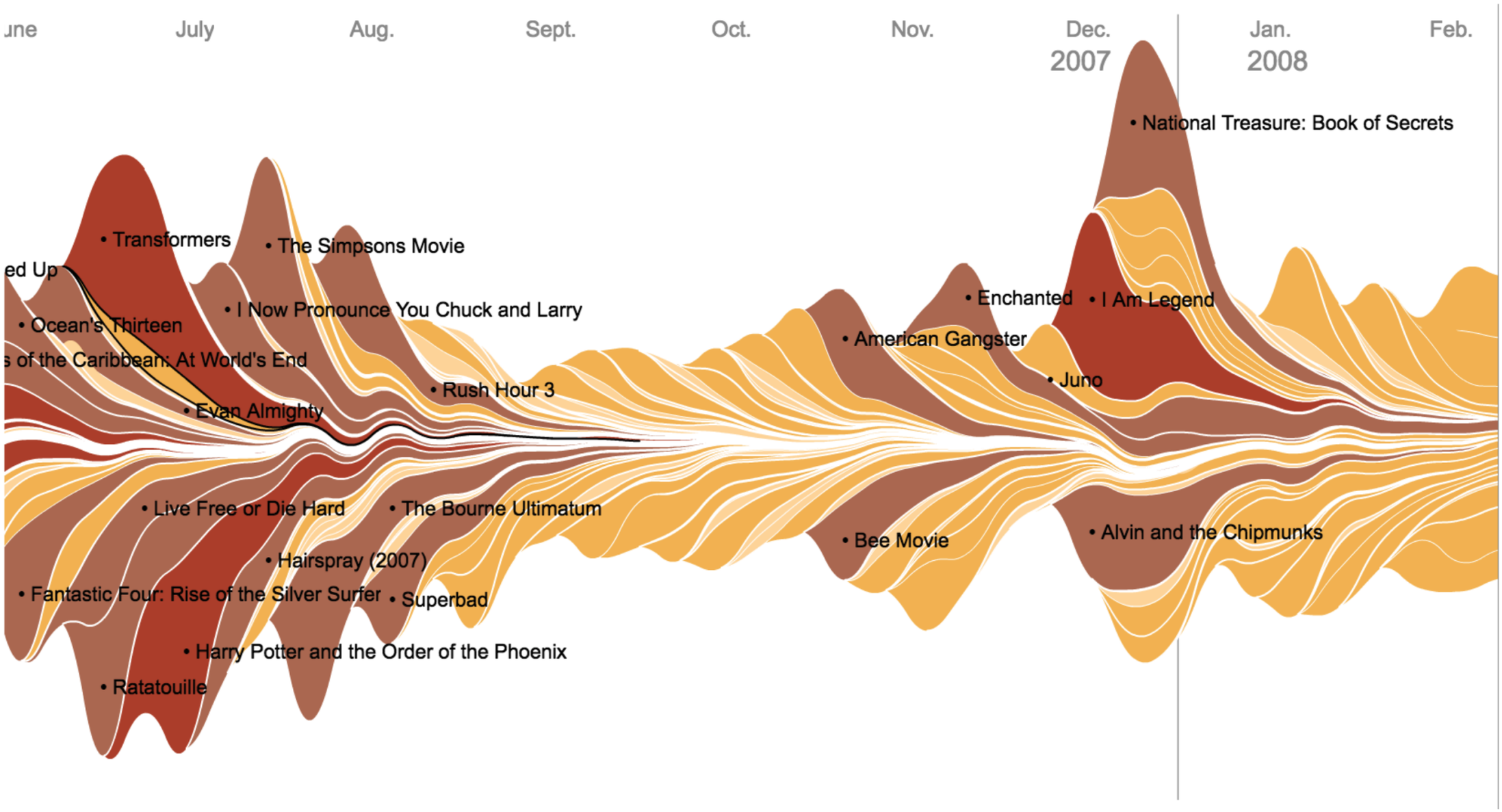
February 23, 2008

FEEDBACK

The Ebb and Flow of Movies: Box Office Receipts 1986 — 2008

Summer blockbusters and holiday hits make up the bulk of box office revenue each year, while contenders for the Oscars tend to attract smaller audiences that build over time. Here's a look at how movies have fared at the box office, after adjusting for inflation.

Find Movie Go



Each shape shows how one film did at the box office.

Height shows weekly box office revenue

Width shows longevity

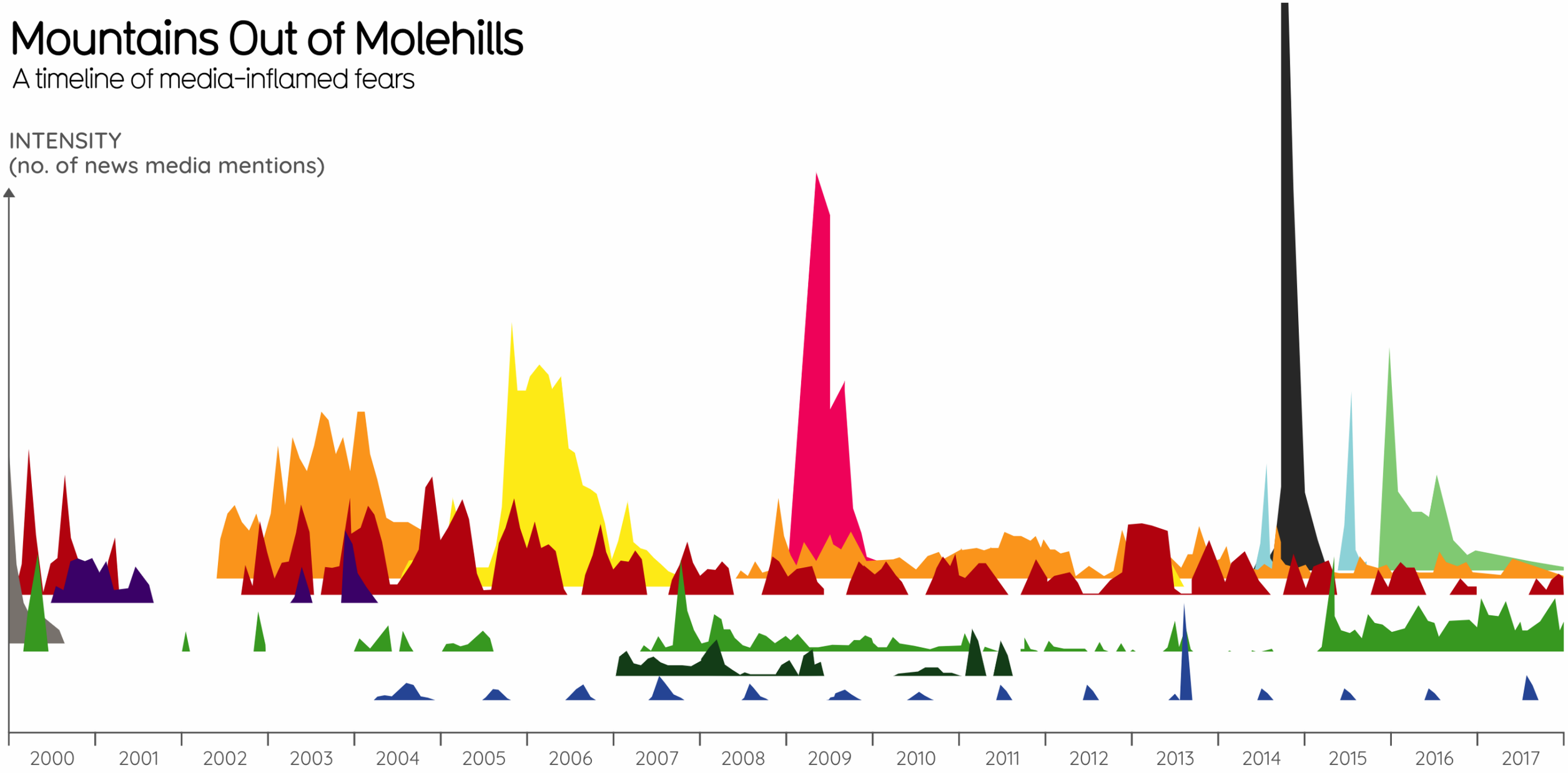
The **area** of the shape (and its **color**) corresponds to the film's total domestic gross, through Feb. 21

$862 milllion
250
100
25
1

layered density graph

Mountains Out of Molehills

A timeline of media-inflamed fears

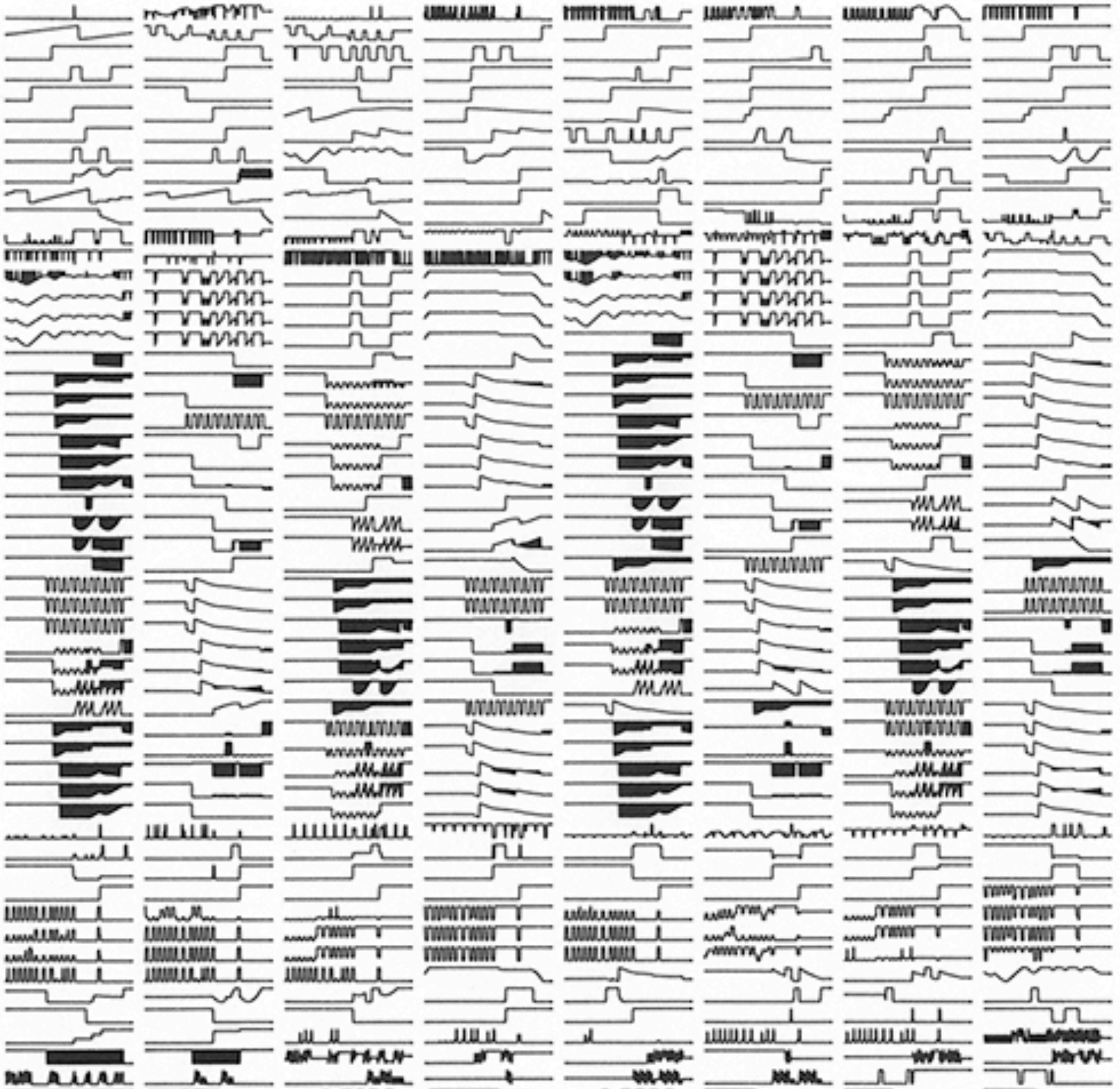


ASTEROIDS | BIRD FLU | EBOLA | KILLER WASPS | MAD COW DISEASE | MERS | MILLENIUM BUG | SARS | SWINE FLU | VACCINES & AUTISM | VIOLENT VIDEO GAMES | ZIKA

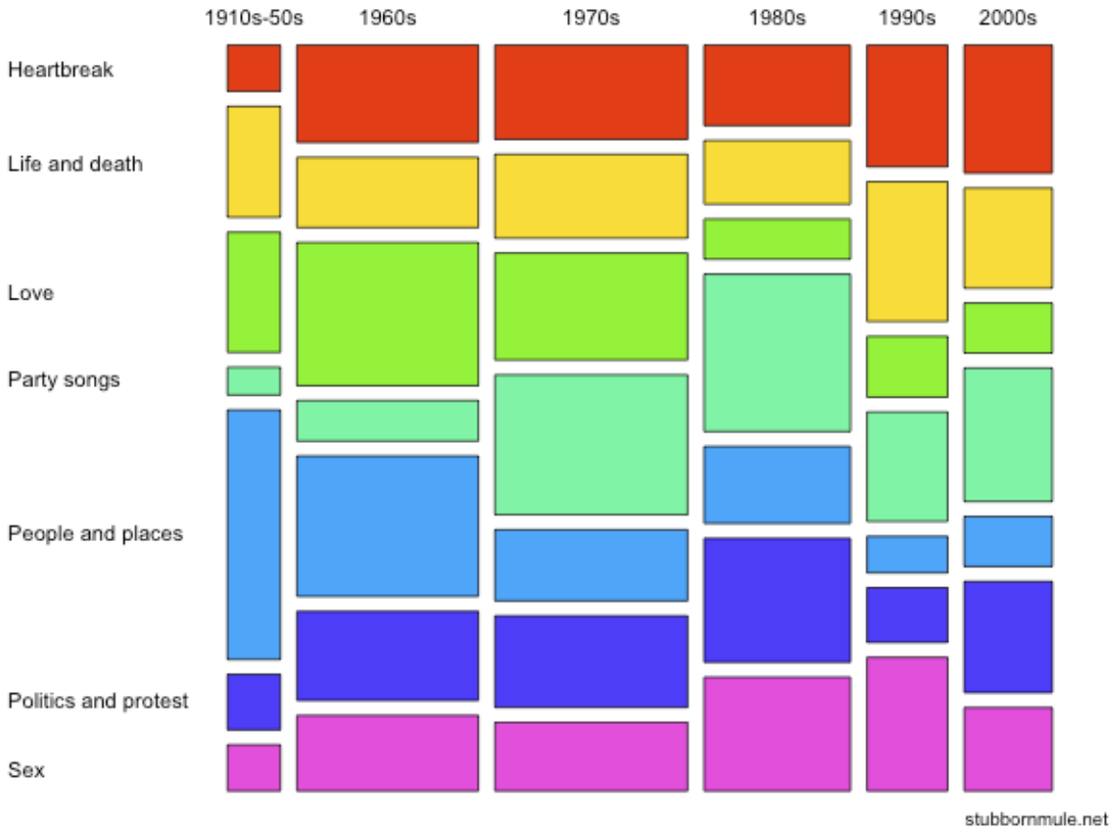
Concept & Design: David McCandless // Design & Code: Fabio Bergamaschi

align to baseline | scale to fit Ebola | scale to deaths

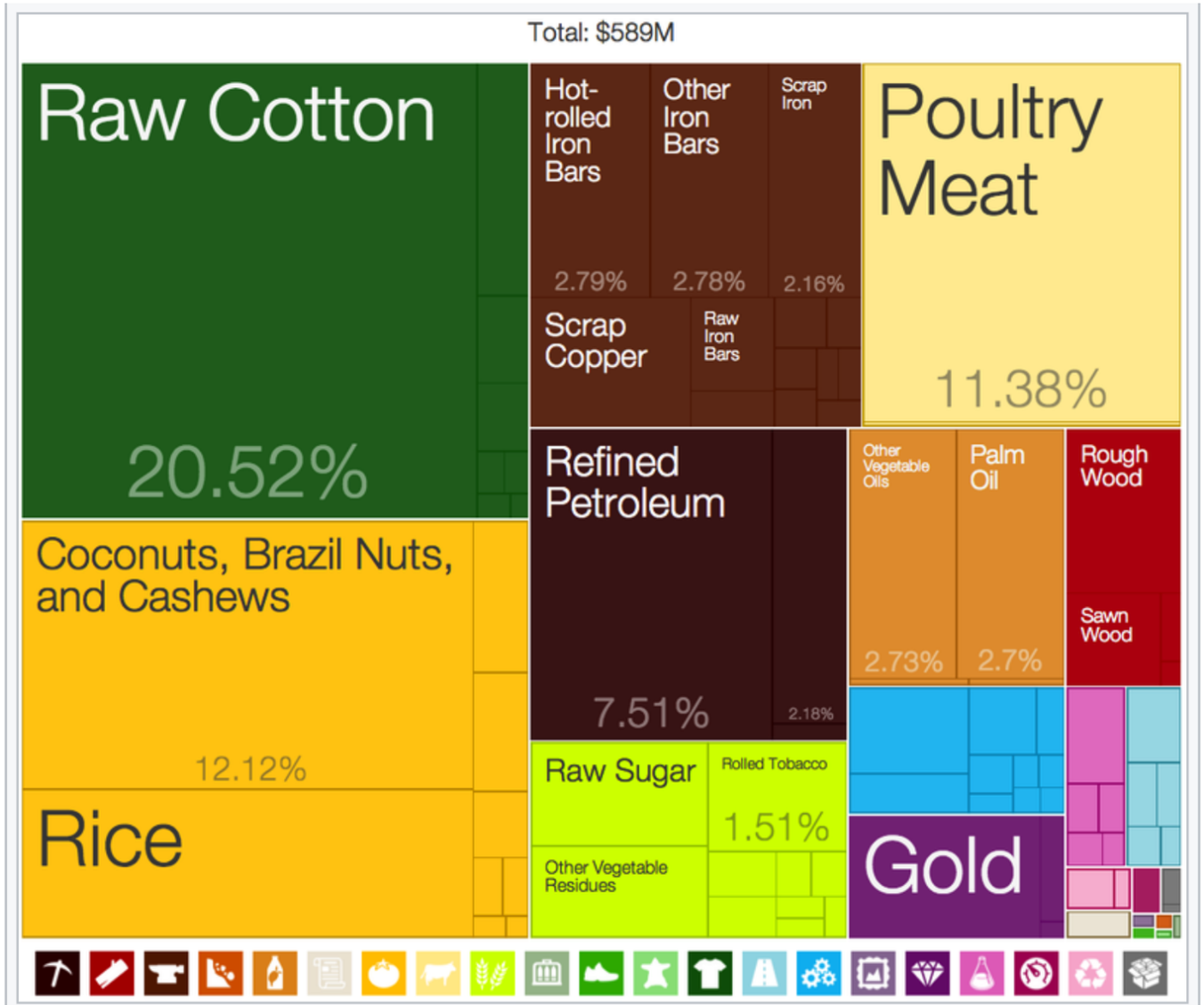
Spikelines



Mosaic plot



TreeMap

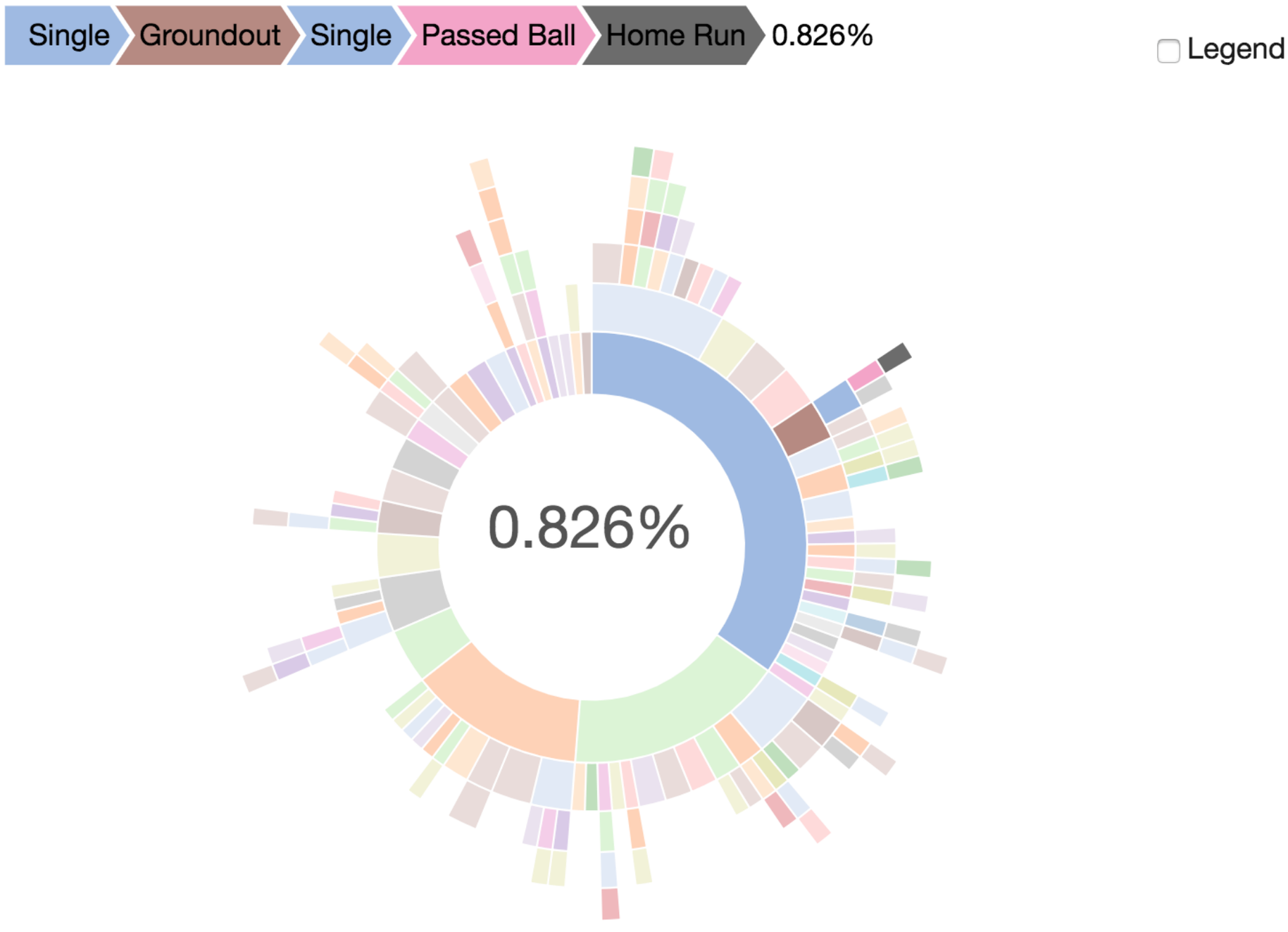


Treemap of Benin's exports by product category, 2009. The Product Exports Treemaps are one of the most recent applications of these kind of visualizations, developed by the Harvard-MIT Observatory of Economic Complexity

Marimekko

Sunburst Chart

Multi-level Pie Chart, Belt Chart, Radial Treemap.



진짜 데이터로 R을 이용하여 위에서 이야기한 Graph 그려보기

아파트 실거래가 데이터

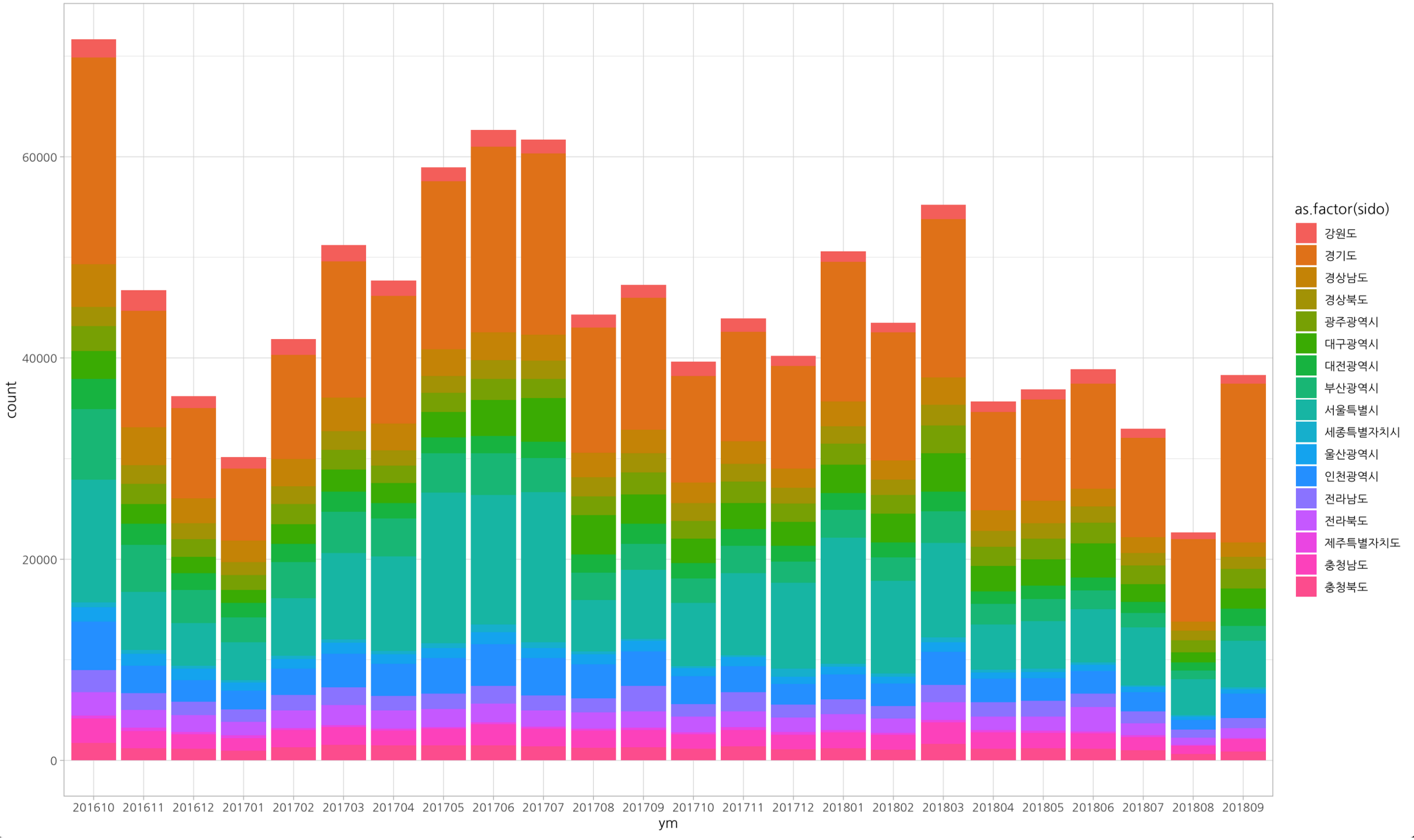
http://rtdown.molit.go.kr/

2016. 10 ~ 2018. 9

1,078,890 rows

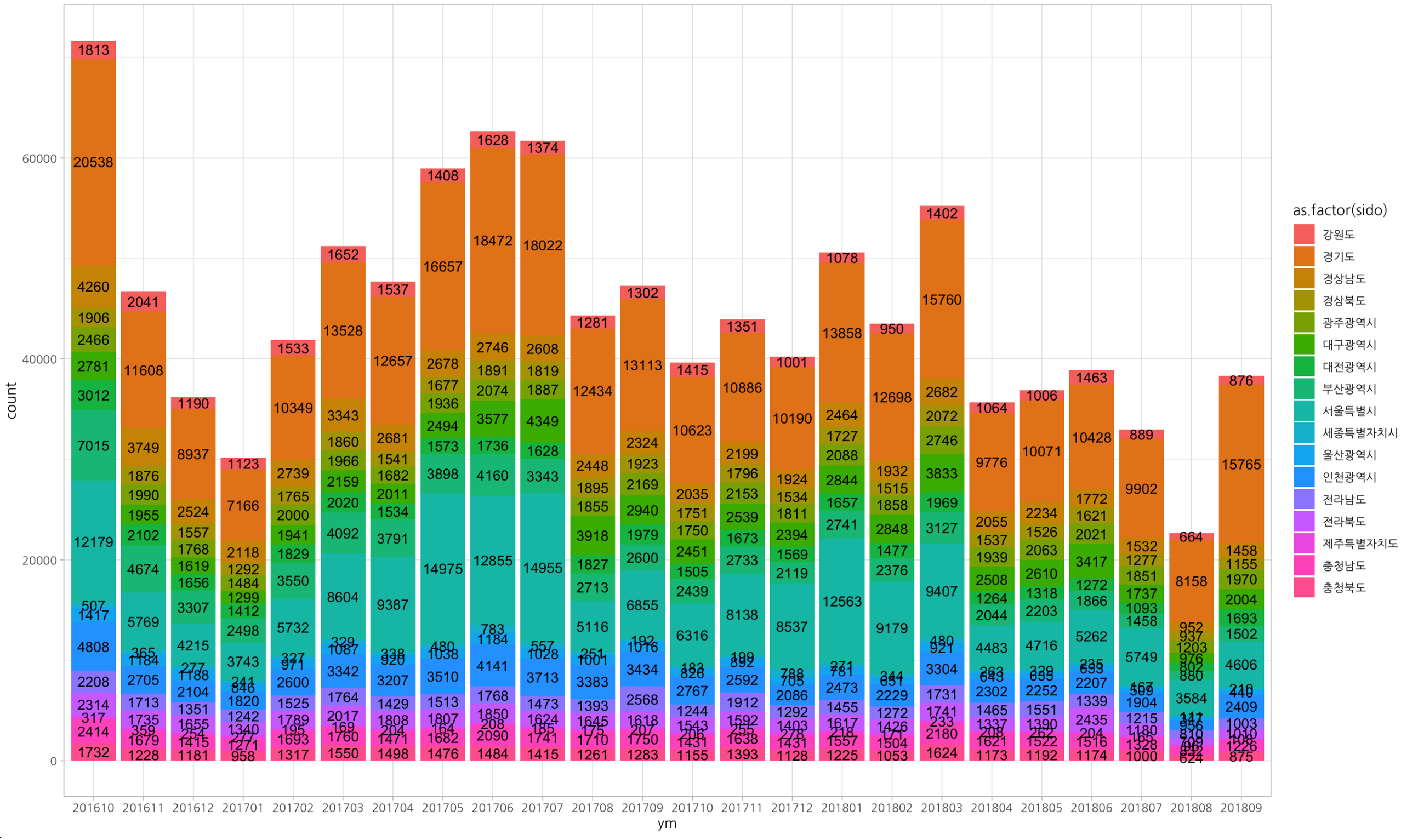
http://rstudio-pubs-static.s3.amazonaws.com/246453_fe28ab9dcf644558849595d1f570c693.html

ggplot2 geom_bar()



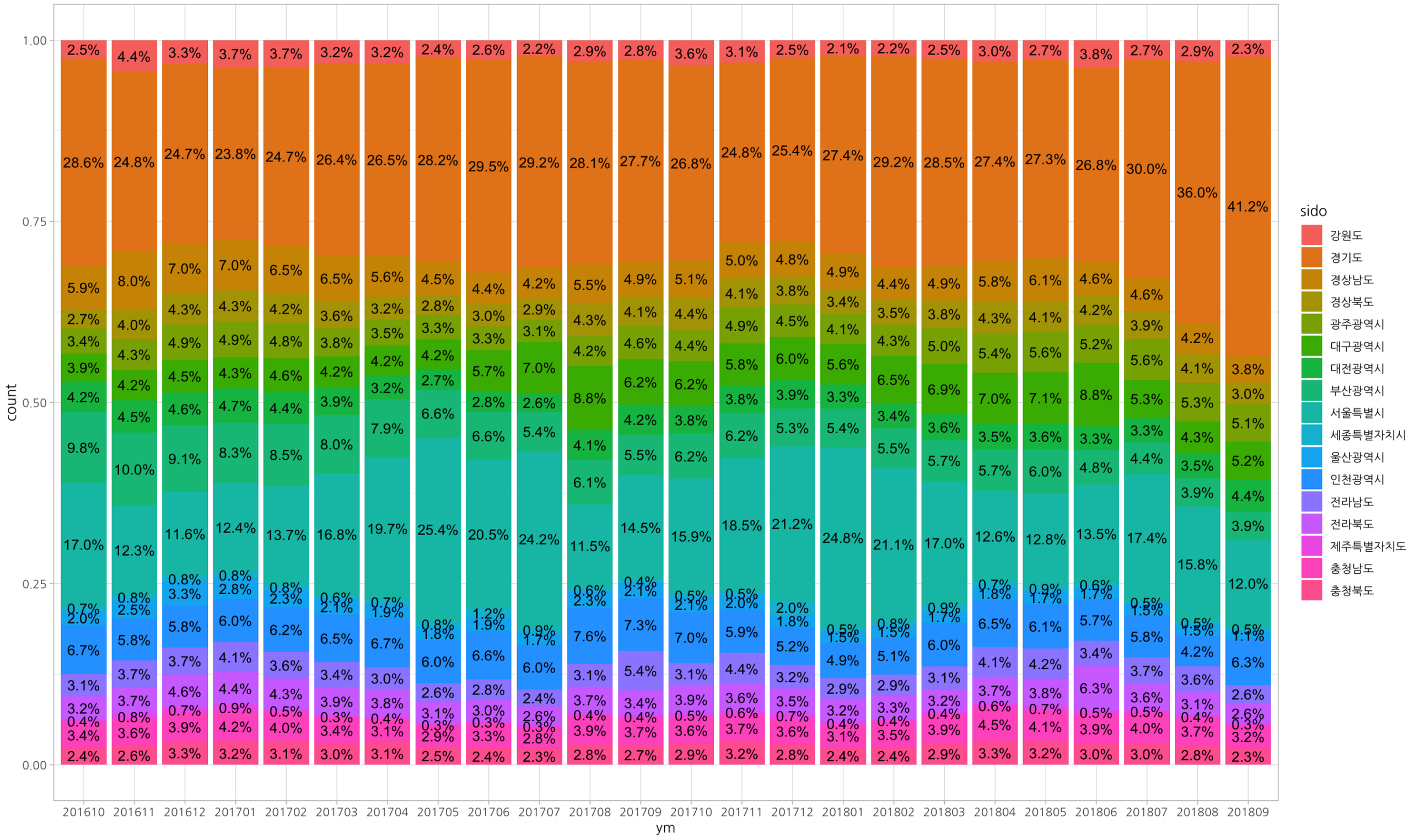
ggplot(apt, aes(x= ym, fill = as.factor(sido))) + geom_bar()

ggplot2 geom_bar() + geom_text()



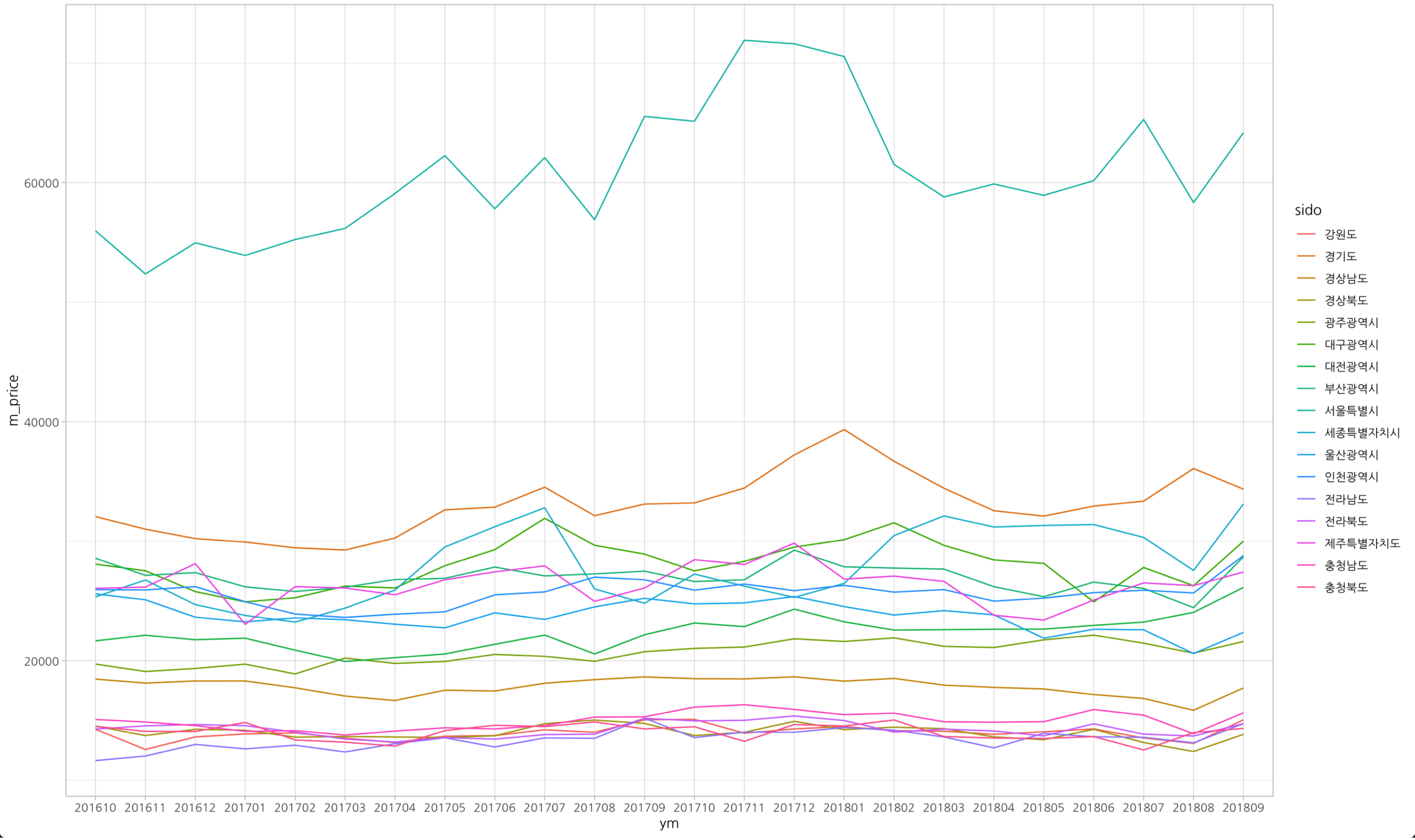

```
ggplot(apt, aes(x= ym, fill = as.factor(sido))) + geom_bar() +
geom_text(aes(label=..count..),stat="count",position=position_stack(0.5))
```

ggplot2 geom_bar() + geom_text() position_fill

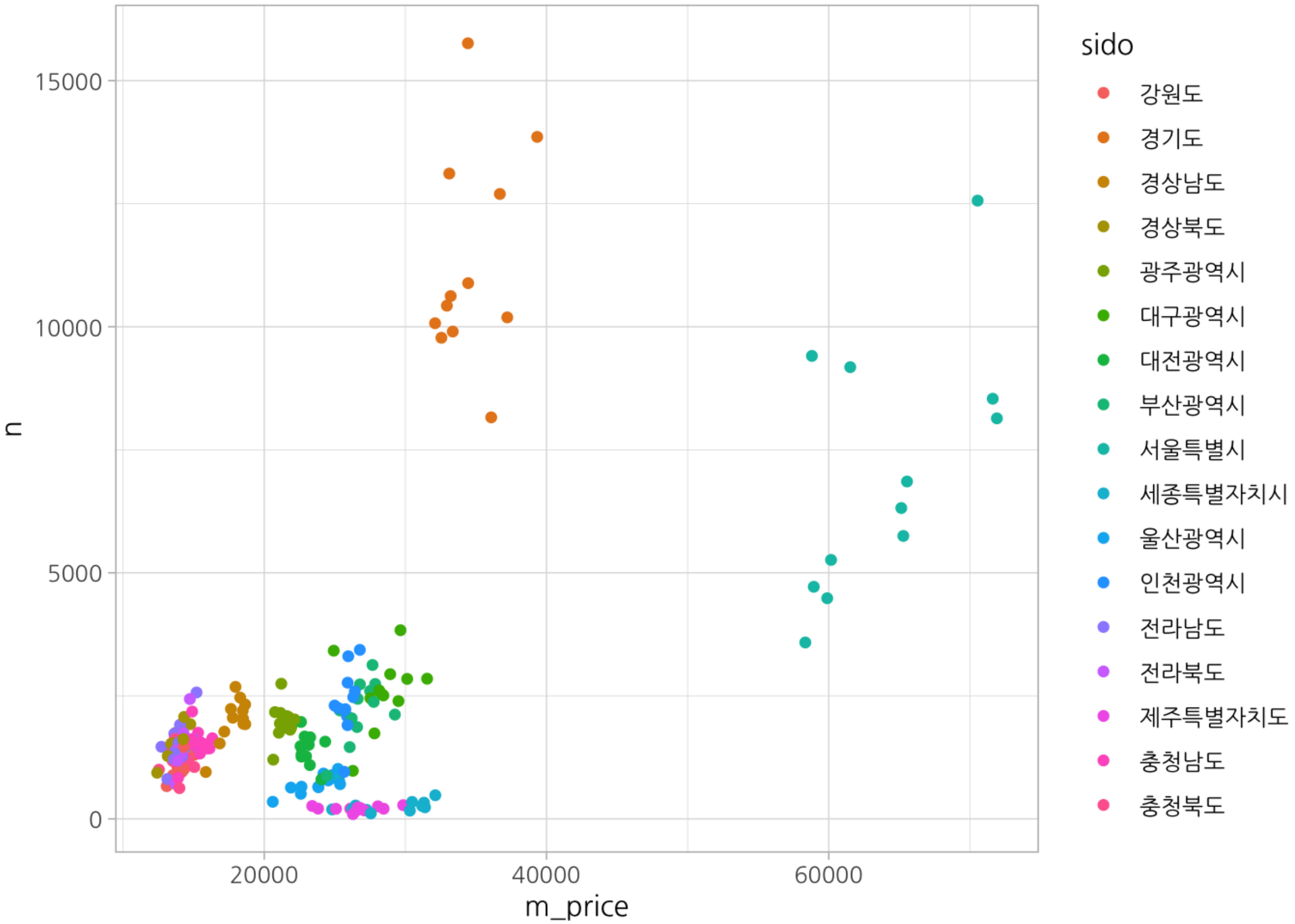



```
ggplot(apt,aes(x=ym,fill=sido)) + geom_bar(position="fill")+
geom_text(data=percentData, aes(y=n, label=ratio), position=position_fill(vjust=0.5))
```

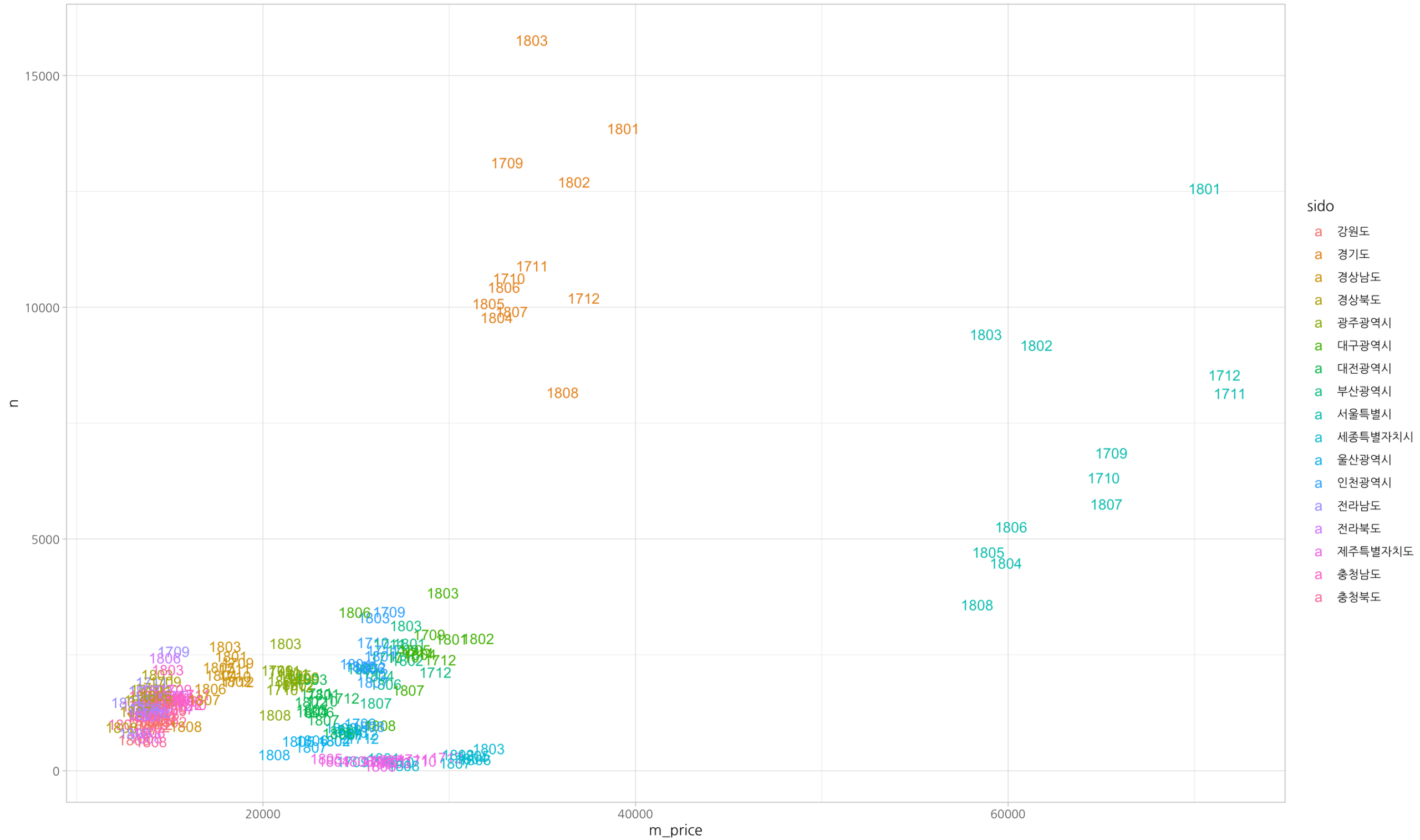
ggplot2 geom_line()



ggplot(percentData, aes(x= ym, y= m_price, group=sido, color = sido)) + geom_line()

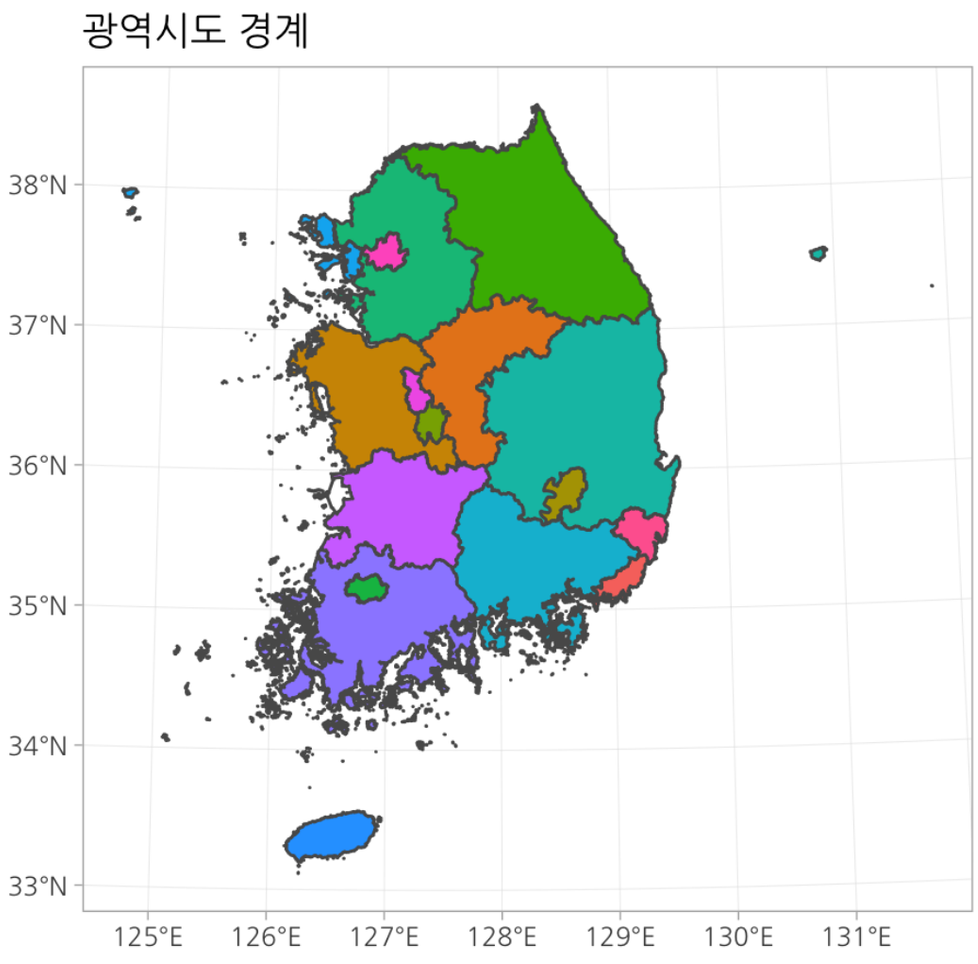


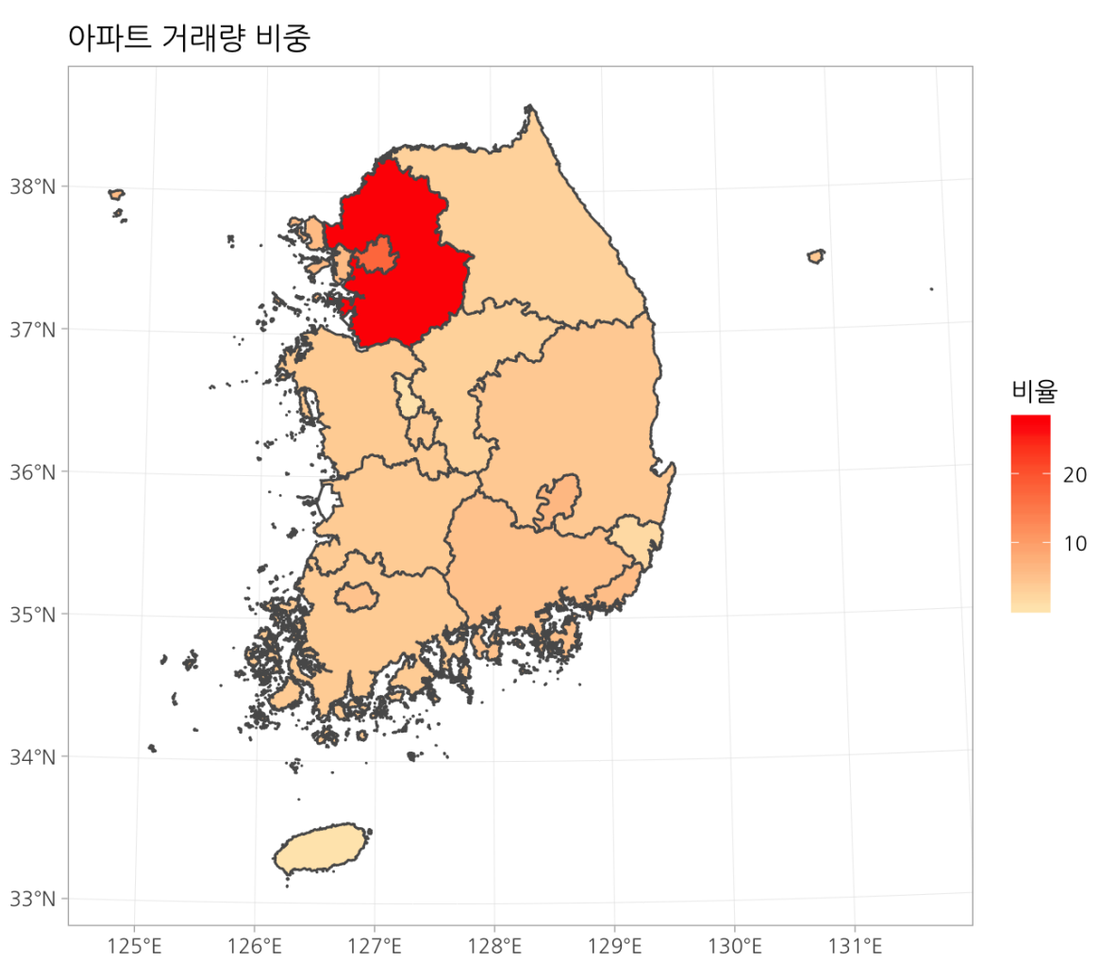

```
ggplot(percentData,aes(x=m_price,y = n, color=sido))+
  geom_point()
```



ggplot(percentData,aes(x=m_price,y = n, color=sido))+
geom_text(aes(label=substr(ym, 3, 6)))

ggplot2 geom_sf()






```
ggplot(data=sido_simp_shp, aes(fill=ratio)) +
  geom_sf() +
  labs(title="아파트 거래량 비중") +
  theme(legend.position = "right") +
  scale_fill_gradient(low = "wheat1", high = "red", name = "비율", labels = scales::comma)
```

ggplot2 limitation

1. 구분 항목이 많아질 경우 2차원 그래프 상에서 색상 표현 및 연속적인 구분이 어렵다.
2. 그래프만 가지고 표현하려고 할 경우 누락되는 정보가 많다.

ggplot2 alternatives

1. 구분 항목이 많아질 경우 2차원 그래프 상에서 색상 표현 및 연속적인 구분이 어렵다.
2. 그래프만 가지고 표현하려고 할 경우 누락되는 정보가 많다.

1. 연속적인 흐름을 표현할 수 있도록 그래프를 그린다.
2. 데이터가 갖고 있는 정보들을 표현할 수 있도록 한다.

서울특별시: 12855

60,000

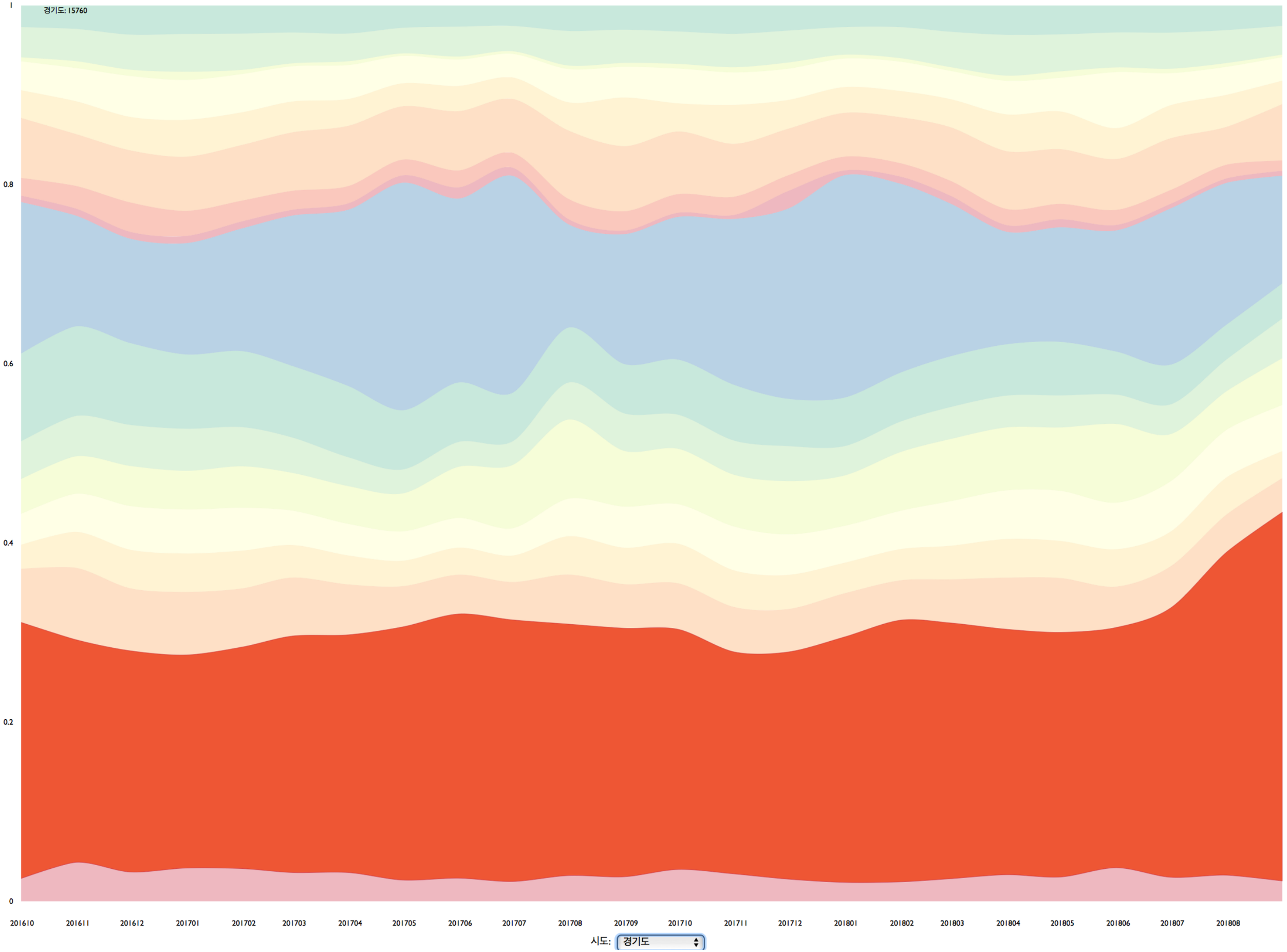
40,000

20,000

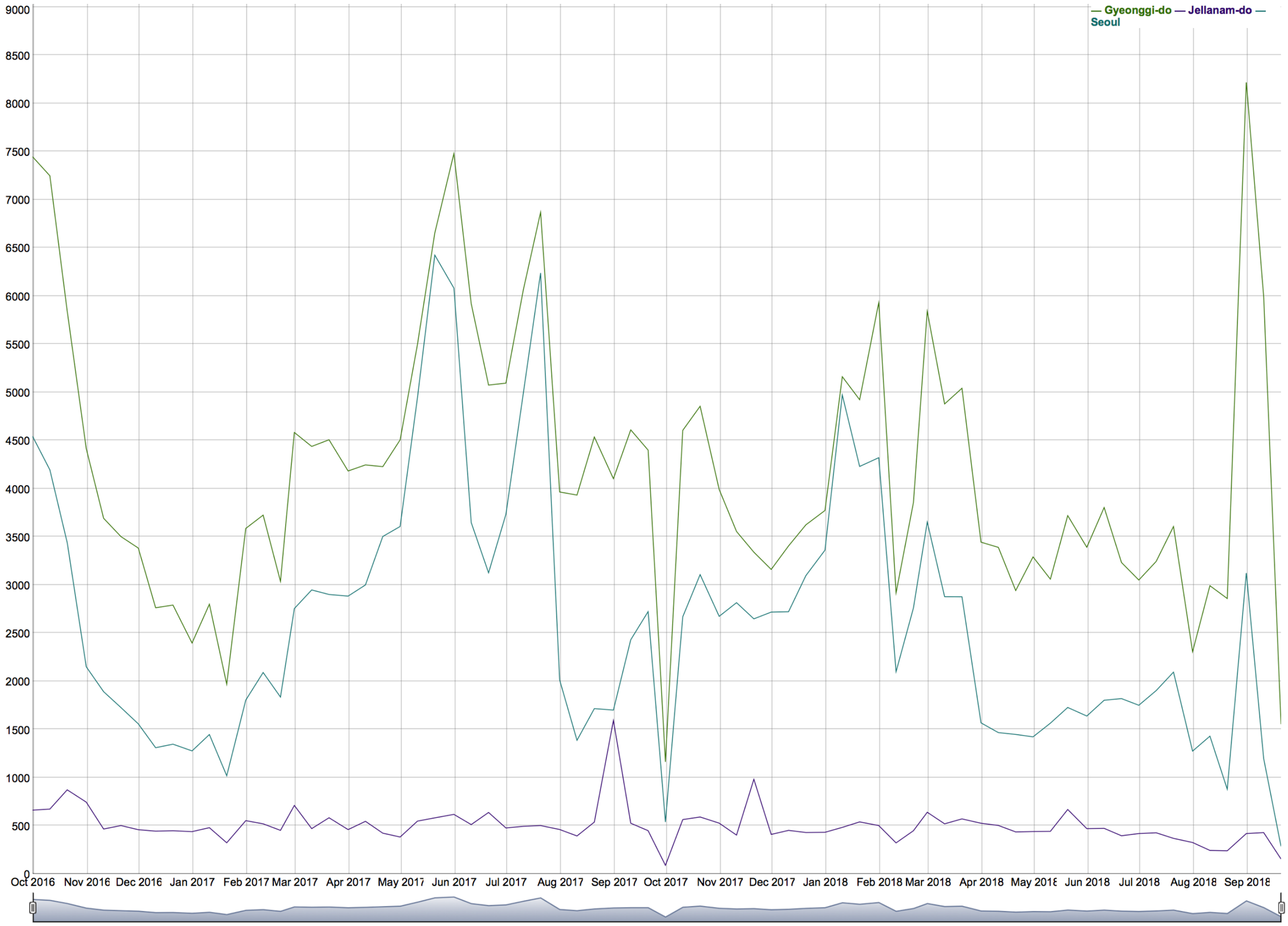
0

201610 201611 201612 201701 201702 201703 201704 201705 201706 201707 201708 201709 201710 201711 201712 201801 201802 201803 201804 201805 201806 201807 201808

시도: --- Select ---


경기도: 15760
1
0.8
0.6
0.4
0.2
0
201610
201611
201612
201701
201702
201703
201704
201705
201706
201707
201708
201709
201710
201711
201712
201801
201802
201803
201804
201805
201806
201807
201808
시도:
경기도

dygraphs for R



sido_nm		n	
Gangwon-do	1813		876
Gyeonggi-do	20538		15765
Gyeongsangnam-do	4260		1458
Gyeongsangbuk-do	1906		1155
Gwangju	2466		1970
Daegu	2781		2004
Daejeon	3012		1693
Busan	7015		1502
Seoul	12179		4606
Sejong-si	507		210
Ulsan	1417		440
Incheon	4808		2409
Jellanam-do	2208		1003
Jeollabuk-do	2314		1010
Jeju-do	317		108
Chungcheongnam-do	2414		1226
Chungcheongbuk-do	1732		875

plotSparklineTable(select(data.frame(percentData), sido_nm, n, ym), row.var = 'sido_nm', col.vars = 'n')

sido

경기도

301,596

서울특별시

부산광역시

대구광역시

광주광역시

전라북도

충청남도

인천광역시

경상남도

대전광역시

전라남도

충청북도

경상북도

강원도

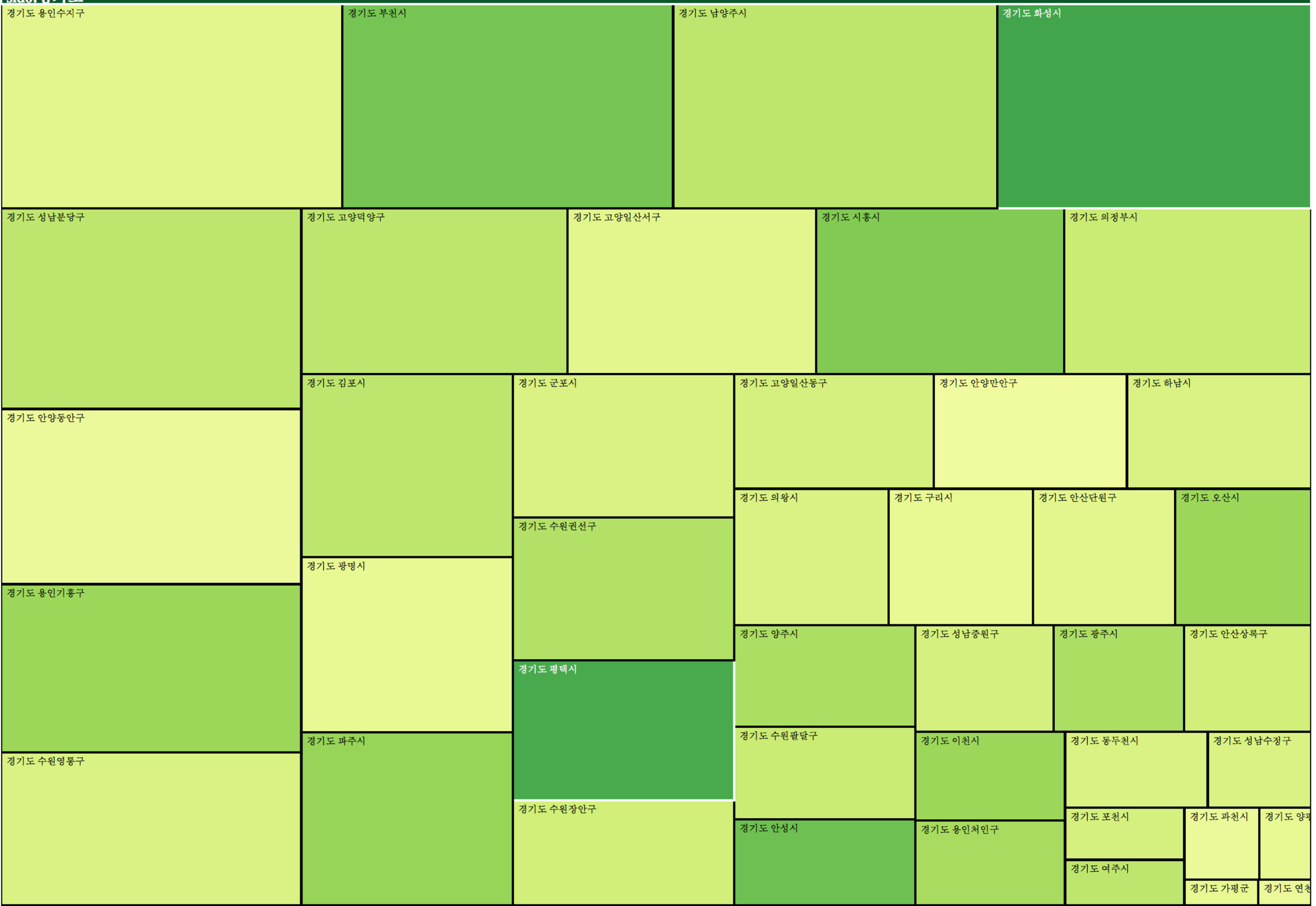
울산광역시

세종특별자치시

제주특별자치도

0 5 10 15 20 25 30 35 40 45

sido.경기도

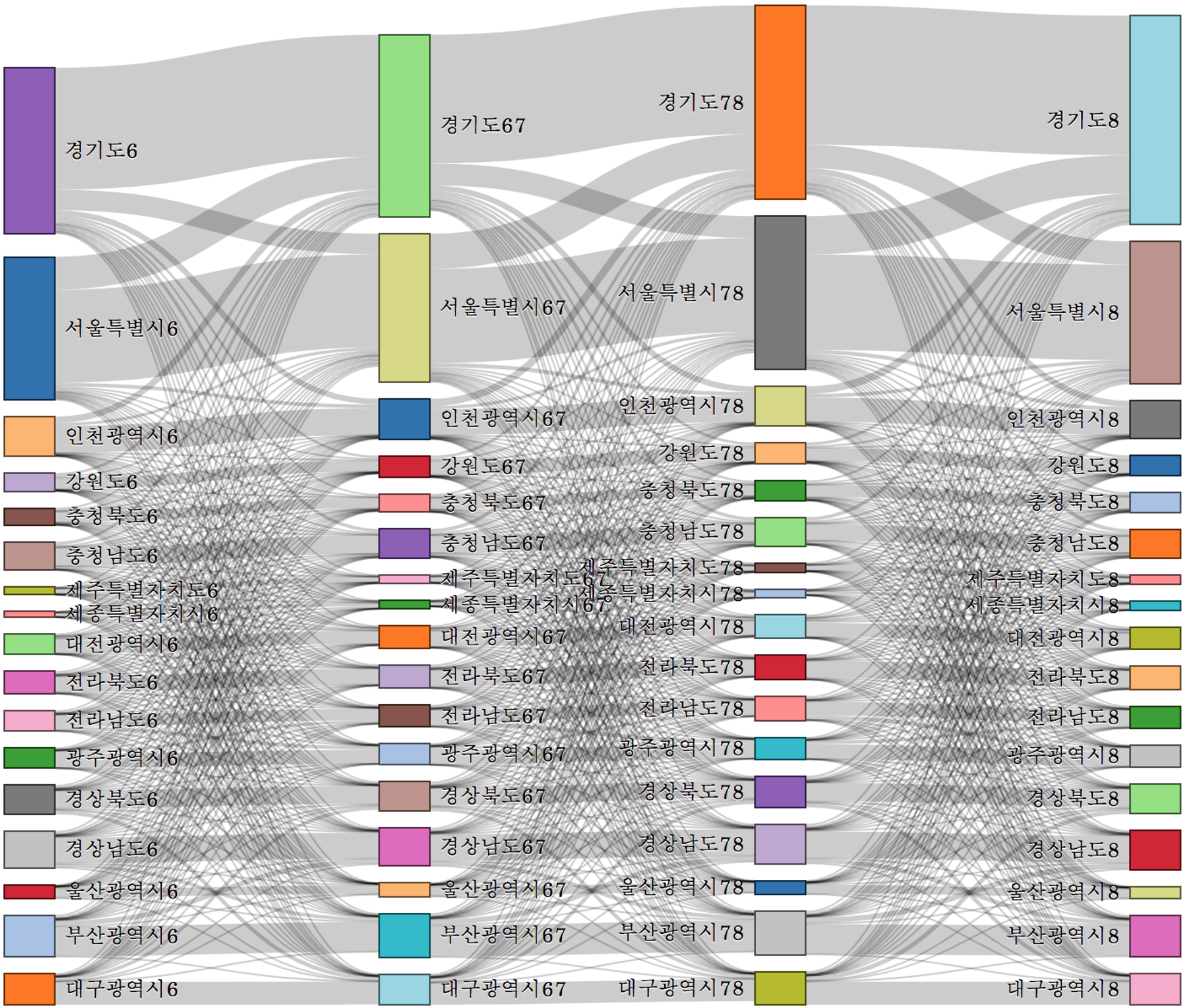

경기도 용인수지구
경기도 부천시
경기도 남양주시
경기도 화성시
경기도 성남분당구
경기도 고양덕양구
경기도 고양일산서구
경기도 시흥시
경기도 의정부시
경기도 김포시
경기도 군포시
경기도 고양일산동구
경기도 안양만안구
경기도 하남시
경기도 안양동안구
경기도 수원권선구
경기도 의왕시
경기도 구리시
경기도 안산단원구
경기도 오산시
경기도 광명시
경기도 용인기흥구
경기도 평택시
경기도 양주시
경기도 성남중원구
경기도 광주시
경기도 안산상록구
경기도 파주시
경기도 수원팔달구
경기도 이천시
경기도 동두천시
경기도 성남수정구
경기도 수원영통구
경기도 수원장안구
경기도 안성시
경기도 용인처인구
경기도 포천시
경기도 과천시
경기도 양평
경기도 여주시
경기도 가평군
경기도 연천
0
5
10
15
20
25
30
35
40
45

인구 이동

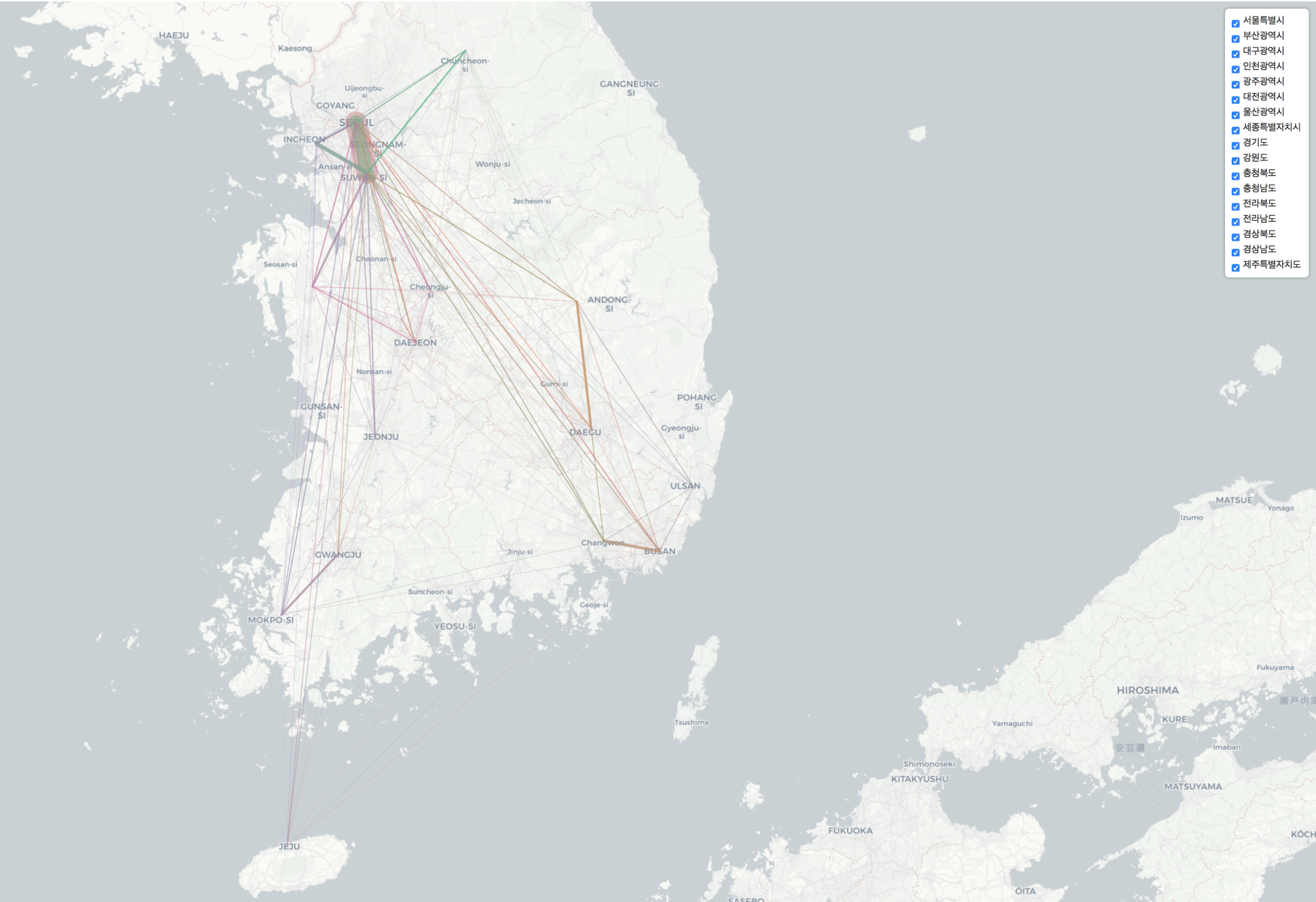
1) 시군구별 이동자수

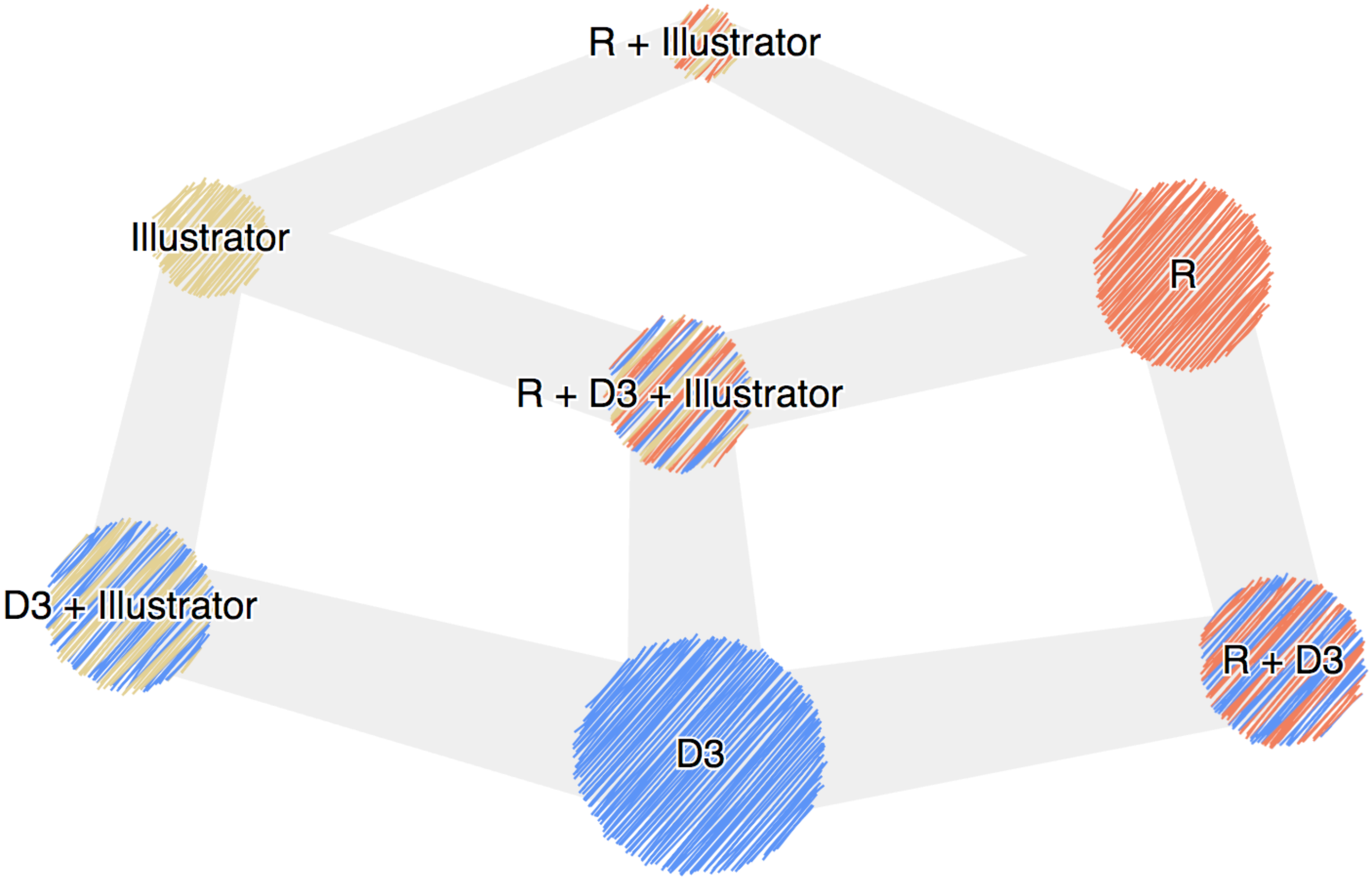
http://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B26001_A01

인구이동 Sankey


경기도6
서울특별시6
인천광역시6
강원도6
충청북도6
충청남도6
제주특별자치도6
세종특별자치시6
대전광역시6
전라북도6
전라남도6
광주광역시6
경상북도6
경상남도6
울산광역시6
부산광역시6
대구광역시6
경기도67
서울특별시67
인천광역시67
강원도67
충청북도67
충청남도67
제주특별자치도67
세종특별자치시67
대전광역시67
전라북도67
전라남도67
광주광역시67
경상북도67
경상남도67
울산광역시67
부산광역시67
대구광역시67
경기도78
서울특별시78
인천광역시78
강원도78
충청북도78
충청남도78
제주특별자치도78
세종특별자치시78
대전광역시78
전라북도78
전라남도78
광주광역시78
경상북도78
경상남도78
울산광역시78
부산광역시78
대구광역시78
경기도8
서울특별시8
인천광역시8
강원도8
충청북도8
충청남도8
제주특별자치도8
세종특별자치시8
대전광역시8
전라북도8
전라남도8
광주광역시8
경상북도8
경상남도8
울산광역시8
부산광역시8
대구광역시8

인구이동 leaflet




R + Illustrator
Illustrator
R
R + D3 + Illustrator
D3 + Illustrator
D3
R + D3

참고사이트

- http://gallery.htmlwidgets.org/

- https://rmarkdown.rstudio.com/flexdashboard/

- http://personal.tcu.edu/kylewalker/interactive-flow-visualization-in-r.html