HEARTCOUNT

KOREA

LOGIN

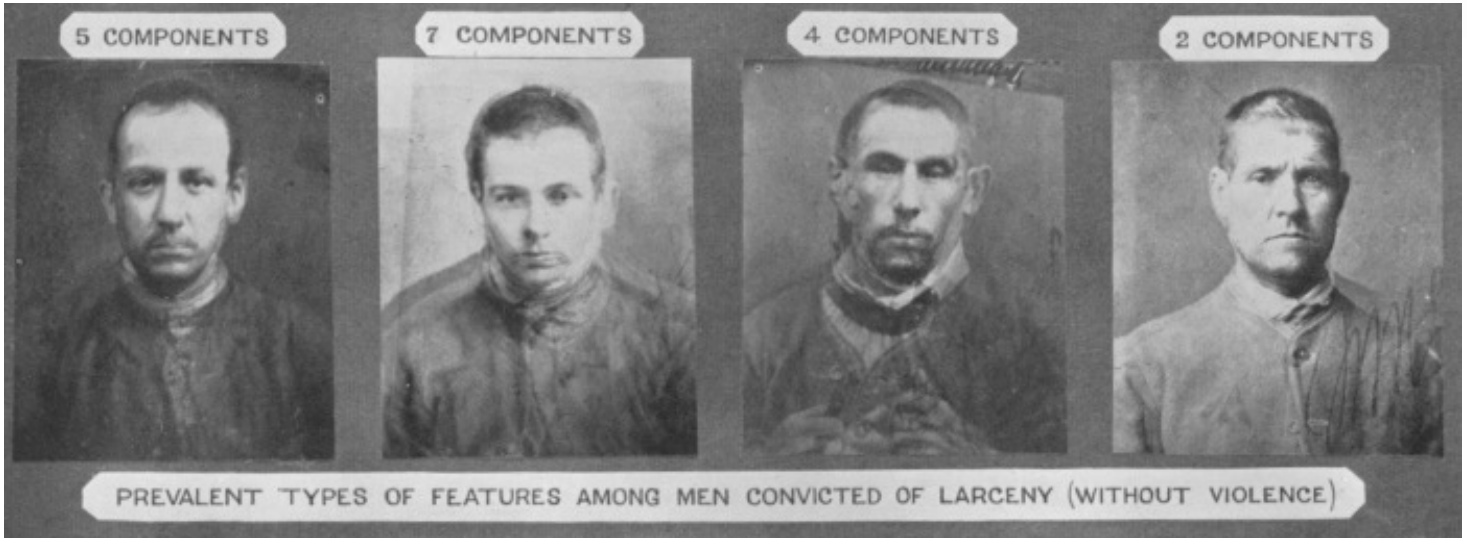
Module II-c
Data Understanding

아이디케이스퀘어드 양승준 / sidney.yang@idk2.co.kr
https://www.heartcount.io

평균의 문제 - Data Aggregation

Average Man
Galton's Composite Portraits
골튼의 합성 초상화

범죄자 사진을
자꾸 포갤수록
범죄자의 특징이
평범함에 묻힌다.



The Problems with Average: Not Robust!

평균의 문제 - Linear Thinking vs. Non-Linear Relationship

친환경 제품을 출시하려 한다.
어떤 세그먼트에 프로모션할까?

잠재 고객 세그먼트	환경 관심도 평균 점수
A	4
B	3

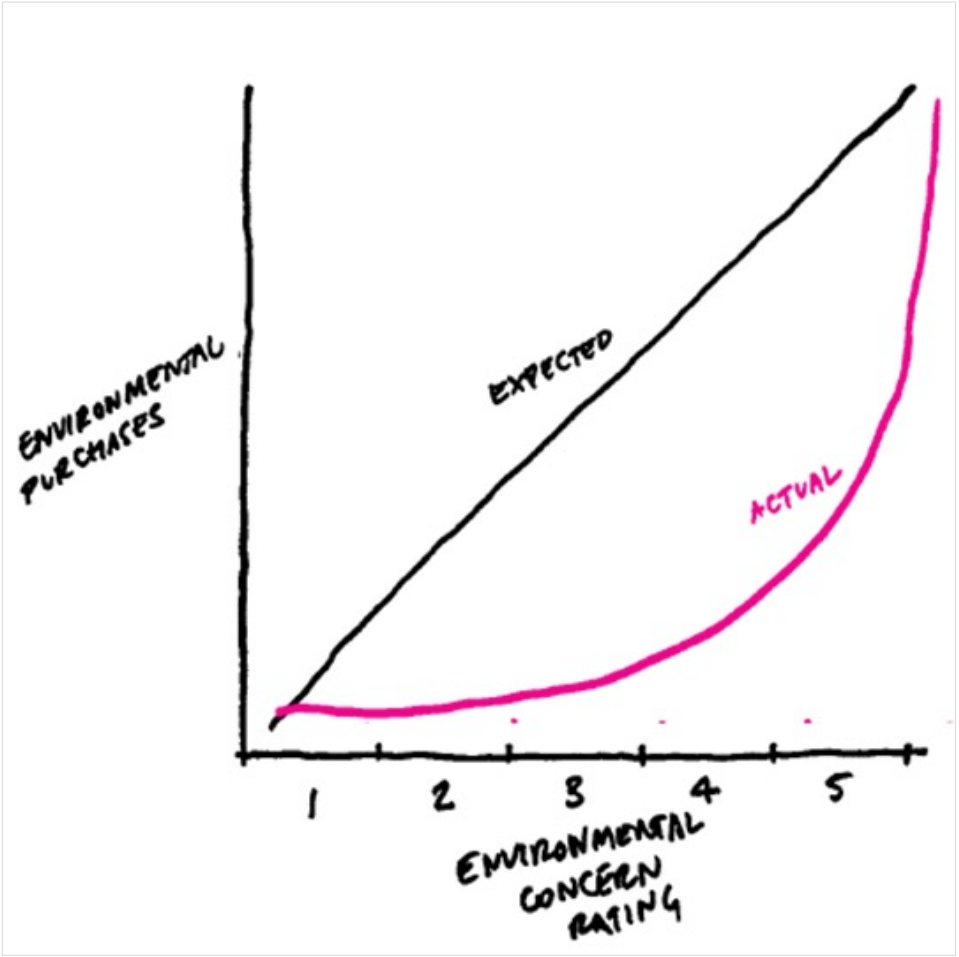
A = [4, 4, 4, 4, 4, 4, 4, 4]
B = [1, 1, 1, 1, 5, 5, 5, 5]

5점에 가중치 2를 부여

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}$$

잠재 고객 세그먼트	환경 관심도 가중 평균 점수
A	4
B	5.5

A = [4, 4, 4, 4, 4, 4, 4, 4]
B = [1, 1, 1, 1, 5, 5, 5, 5, **5, 5, 5, 5**]



제한된·익숙한 관점 - Simpson's Paradox

심슨의 역설

뭉뚱그린 수치는 현실을 왜곡할 수 있음

쪼개보는 일(Segmentation; Drill-Down; Dimensions)의 중요성

남녀 지원자 합격률

	지원자 수	합격자 수	합격률
여자	1,000	150	15%
남자	1,000	250	**25%**

문과대 합결률

	지원자 수	합격자 수	합격률
여자	800	80	**10%**
남자	200	10	5%

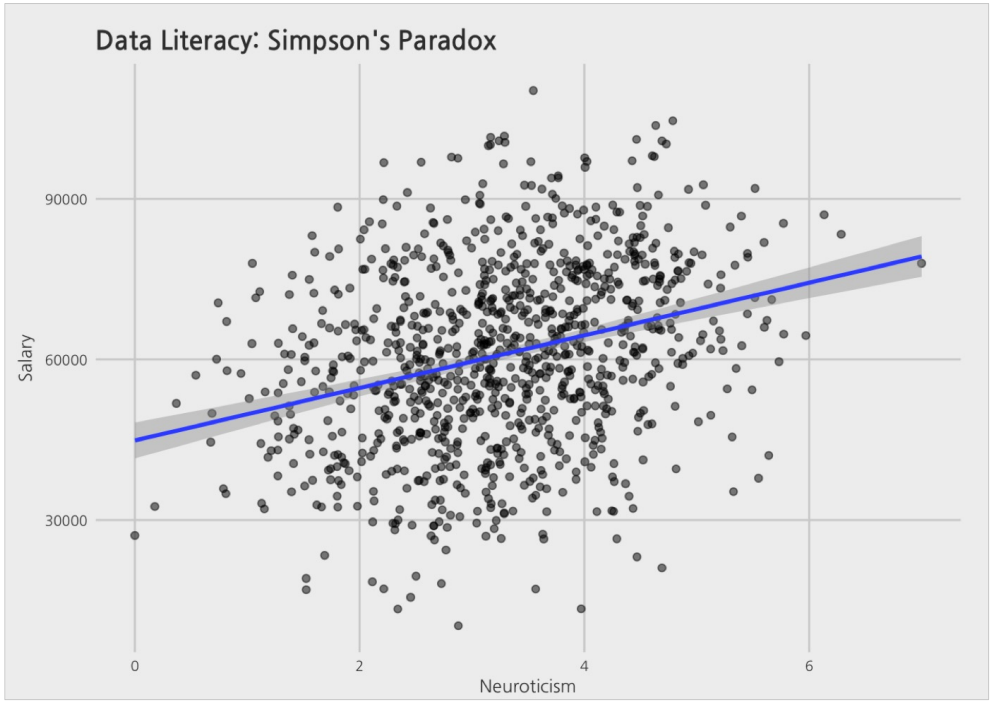
이공대 합격률

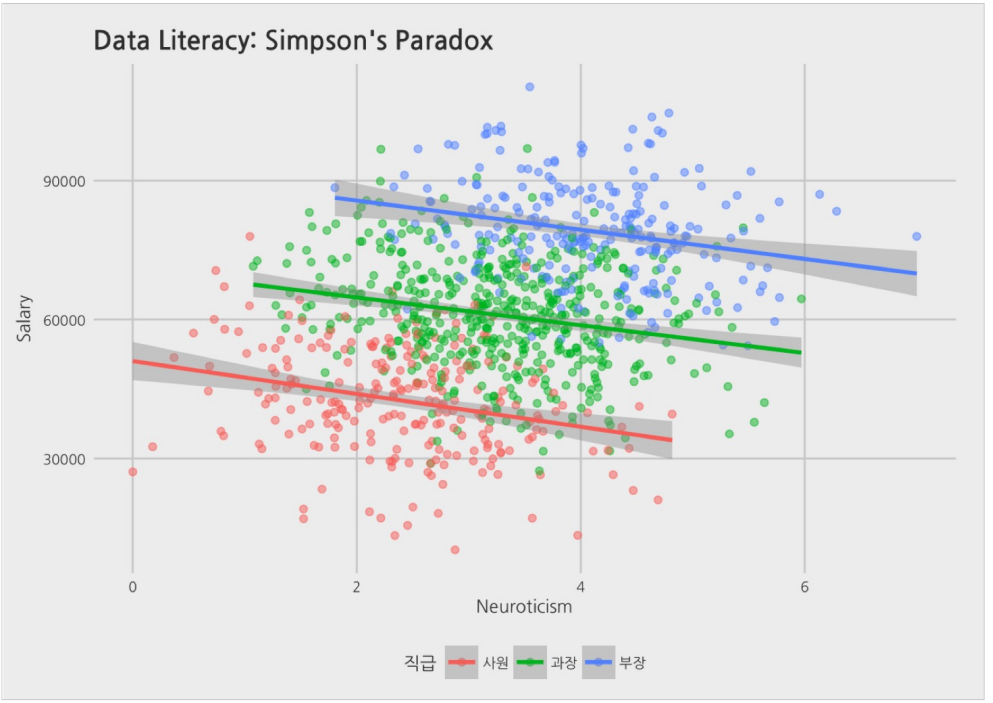
	지원자 수	합격자 수	합격률
여자	200	70	**35%**
남자	800	240	30%

제한된·익숙한 관점 - Simpson's Paradox

새로운 관점(Dimension)

연봉과 까칠함과의 관계 → 직급별 연봉과 까칠함과의 관계





차원의 저주

"If you test enough things, just by random chance,
one of them will be statistically significant."

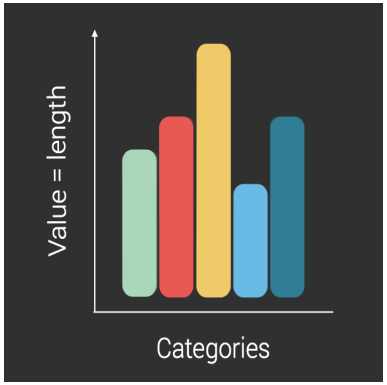
Lucky Coin #1217

- 1년 동안 S&P 상승/하락과 2,000개 동전을 던진 후 결과를 기록
- 95%의 정확도로 앞면이면 S&P지수 상승, 뒷면이면 하락했음

	S&P 지수	Coin 1	Coin 2	Coin 3	Coin 4	Coin 5	.....	.....	Coin 1217	.....	.....	Coin 1999	Coin 2000
1	상승	앞	뒤	앞	앞	뒤			앞			앞	뒤
2	하락	뒤	뒤	앞	앞	뒤			앞			뒤	뒤
3	상승	뒤	앞	뒤	앞	뒤			앞			뒤	앞
4	상승	앞	뒤	앞	뒤	뒤			앞			앞	뒤
5	하락	뒤	앞	뒤	앞	앞			뒤			뒤	뒤
...													
248	상승	앞	뒤	뒤	뒤	앞			앞			뒤	뒤
249	하락	뒤	앞	뒤	뒤	뒤			뒤			뒤	앞
250	하락	앞	뒤	앞	앞	앞			뒤			뒤	뒤

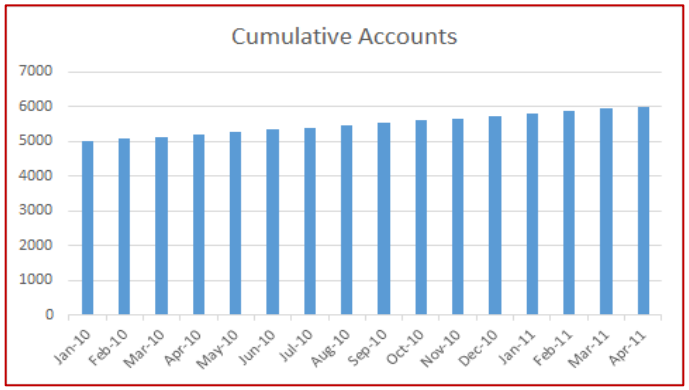
Data Visualization 101 – 1차원

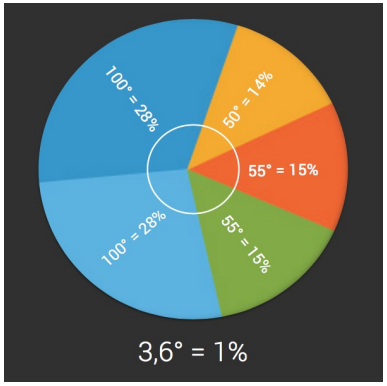
1 Dimension



Bar Chart

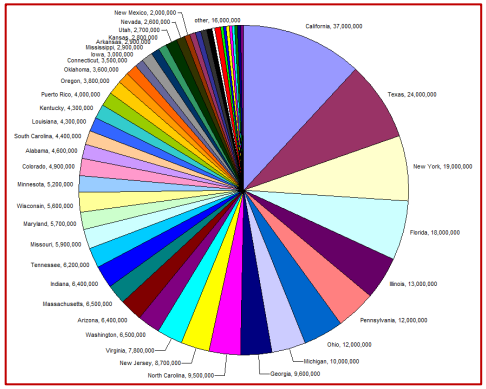
- 서로 다른 범주(사업부)간 평균값(매출 평균)의 차이를 비교하는데 효과적
- 시계열에 따른 변화를 표현하기에는 부적절





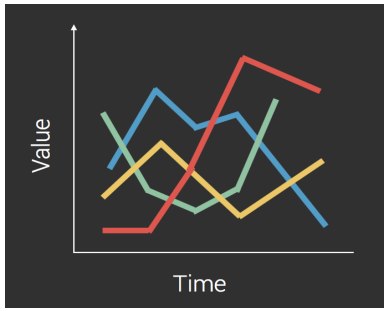
Pie Chart

- 서로 다른 범주가 전체(매출 총합)에서 차지하는 비율을 대비하는데 효과적
- 범주의 갯수가 많거나 비율이 비슷한 경우 부적절



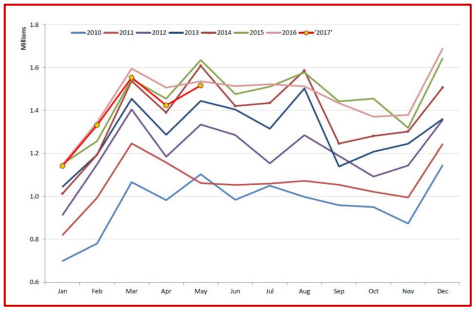
Data Visualization 101 – 2차원

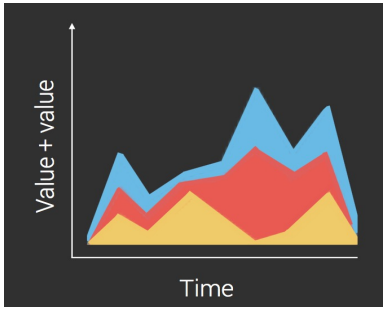
2 Dimensions



Line Chart

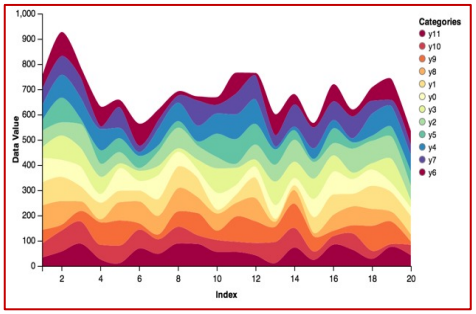
- 시계열에 따른 추세를 보는데 효과적
- 너무 많은 범주(4~5개 이상)가 함께 표현되는 경우 헤깔림

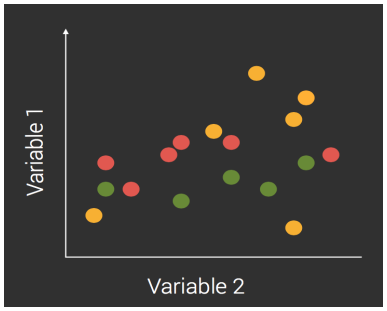




Stacked Area Chart

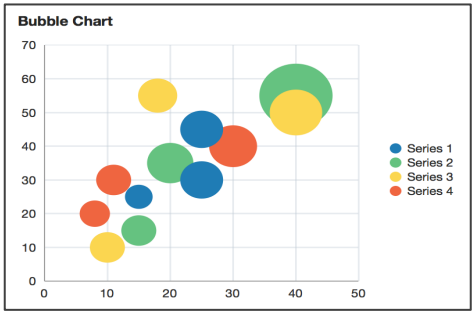
- 시간의 흐름에 따른 개별 범주의 전체 크기 내에서의 상대적 크기 변화를 표현하는데 효과적
- 범주가 너무 많으면 역시 헤깔림





Scatter Plot

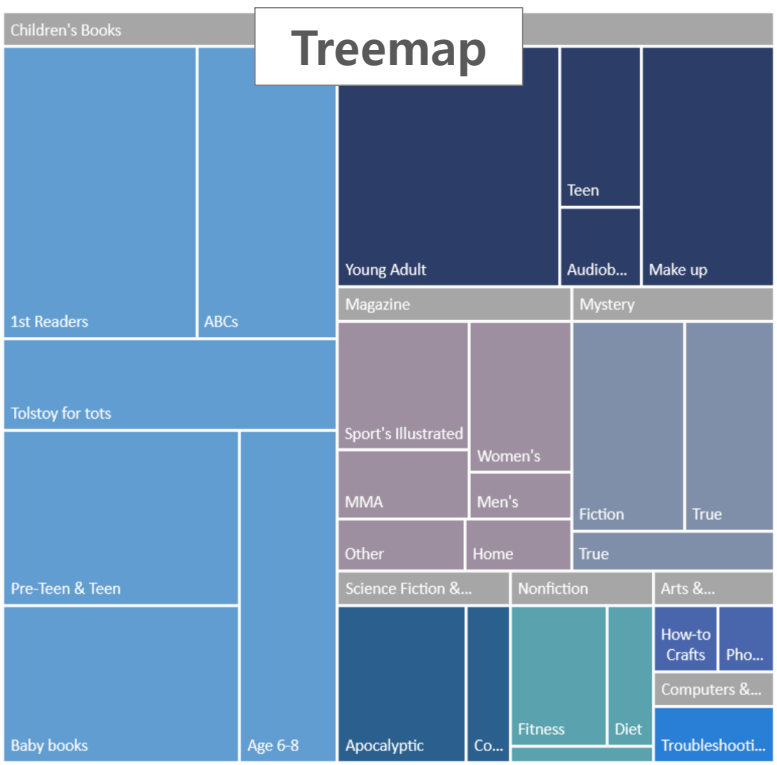
- 두개의 숫자형 변수 간 관계를 파악하는데 효과적
- Outlier를 발견하는데도 효과적
- **원 크기/색상으로 4 Dimensions 표현**

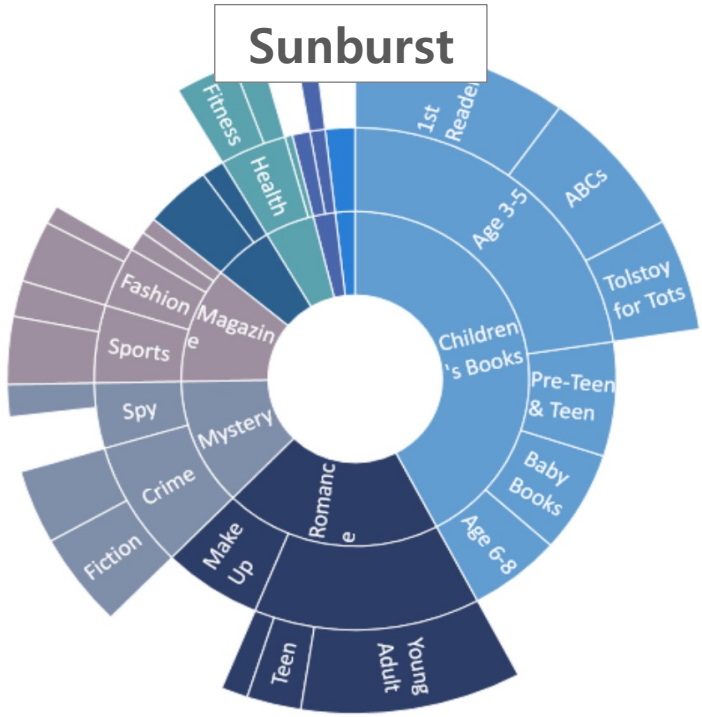


Data Visualization 101 – 계층적 데이터

Hierarchical Data

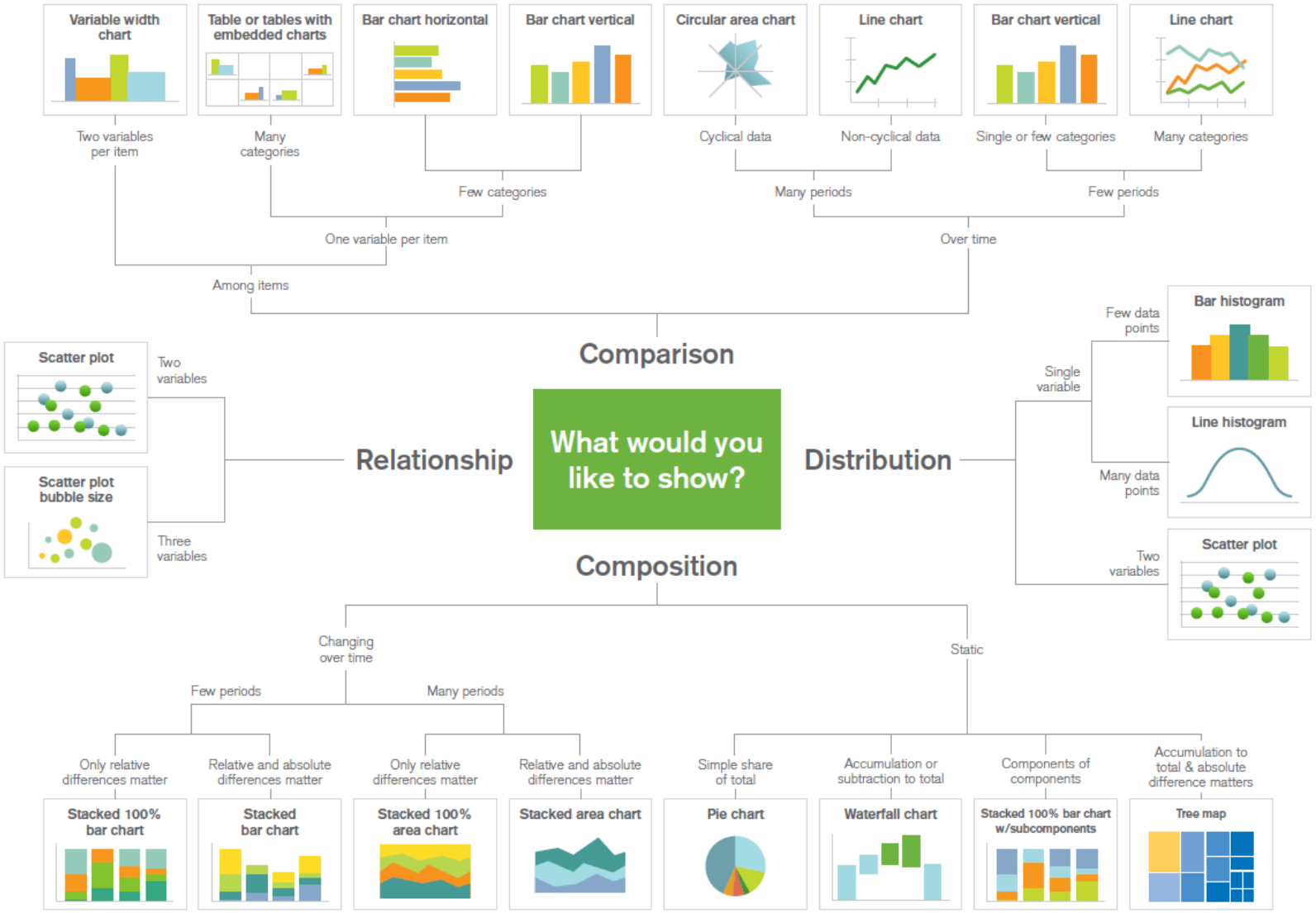
GENRE	SUB-GENRE	TOPIC	REVENUE
ARTS & PHOTOGRAPHY	How-to Crafts	How-to Crafts	$ 2,711
ARTS & PHOTOGRAPHY	Coffee-table	Photography	$ 2,309
CHILDREN'S BOOKS	Baby Books	Baby Books	$ 16,092
CHILDREN'S BOOKS	Age 3-5	1st Readers	$ 24,514
CHILDREN'S BOOKS	Age 3-5	ABCs	$ 17,771
CHILDREN'S BOOKS	Age 3-5	Tolstoy for Tots	$ 13,295
CHILDREN'S BOOKS	Age 6-8	Age 6-8	$ 14,046
CHILDREN'S BOOKS	Pre-Teen & Teen	Pre-Teen & Teen	$ 18,046
COMPUTERS & INTERNET	Troubleshooting	Troubleshooting	$ 4,527
MYSTERY	Crime	Fiction	$ 11,186
MYSTERY	Crime	True Crime	$ 8,790
MYSTERY	Spy	Spy	$ 6,516
MYSTERY	Spy	True Spy	$ 3,809





Data Visualization 101 – 기본 문법 숙지

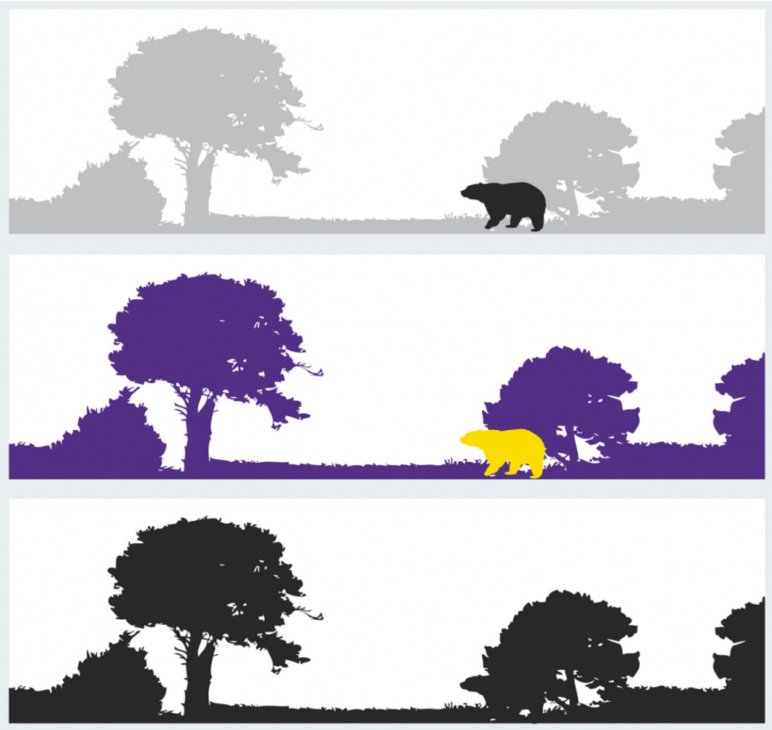
데이터를 통해 알(리)고 싶은 것을 잘 표현할 방법이 무언가?



Data Visualization – Pre-attentive Processing

전주의 처리 (Pre-attentive Processing)

주의를 기울여 복잡한 시각 정보를 처리하기 전에
빠르게 시각 정보를 처리하는 무의식적 과정

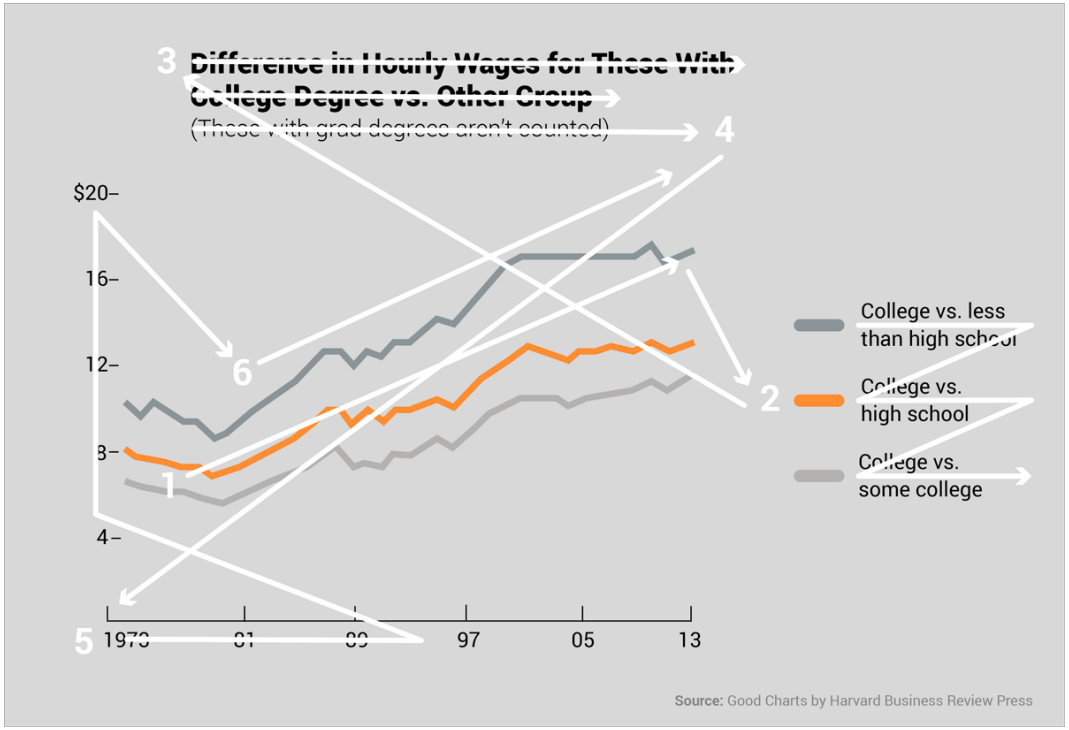


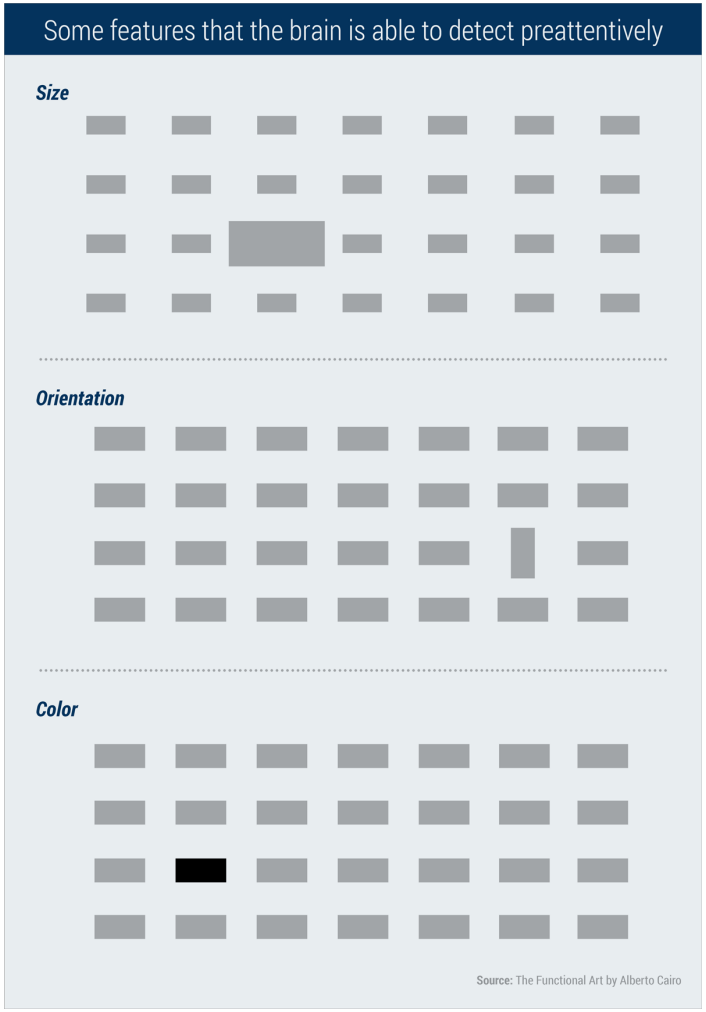
전주의 처리 (Pre-attentive Processing)

핵심 정보가 두드러지도록 적절히 강조

종잡을 수 없는 시선의 흐름

차트를 볼 때 사람의 눈동자가
어떤 순서로 어디로 향할지 알 수 없음



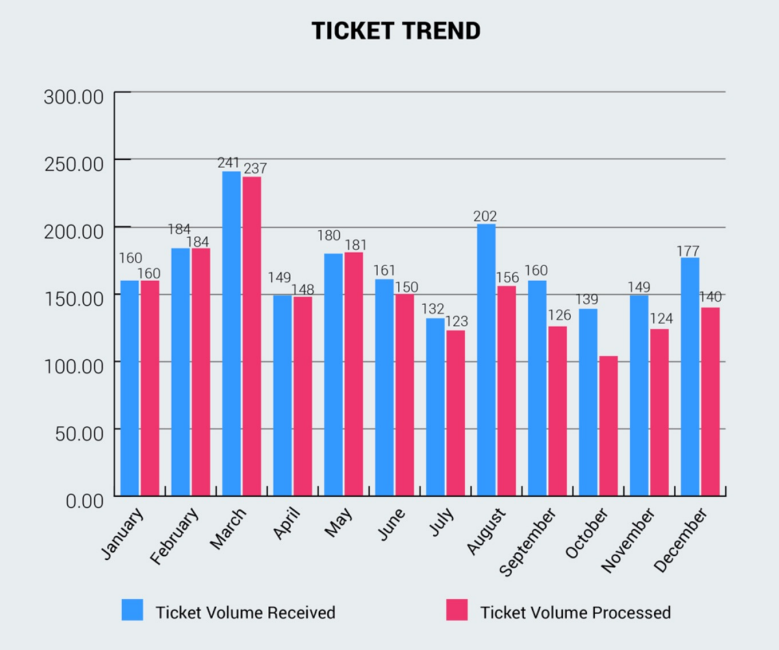


Data Storytelling – Before and After

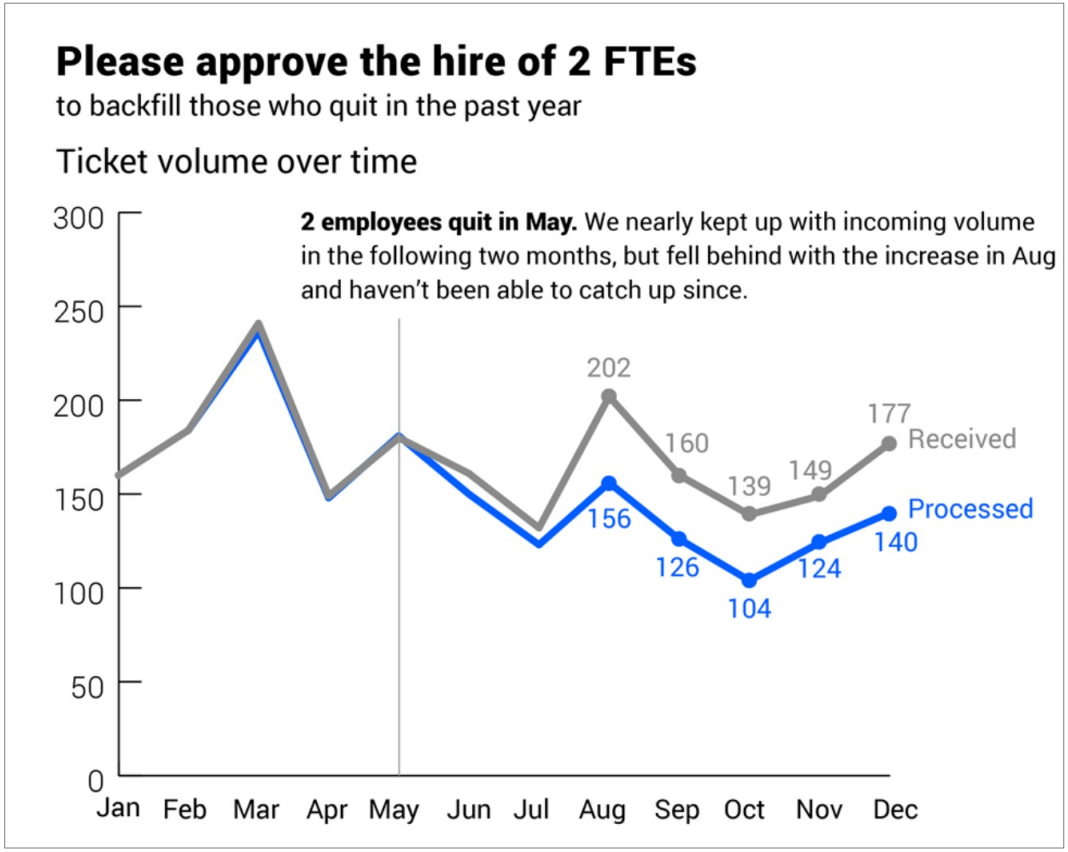
Data Storytelling: 당신의 보고는 *핍진성이 있는가?

*핍진성: 텍스트가 신뢰할 만하고 개연성이 있다고 독자에게 납득시키는 정도.

Before



After: Call to Action

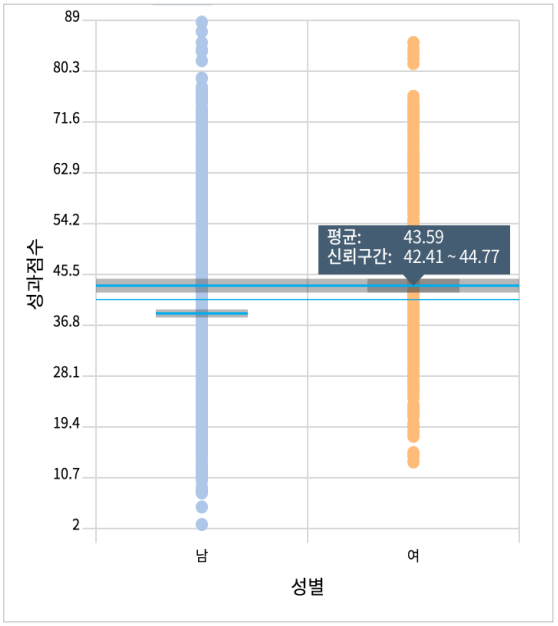


Data Visualization

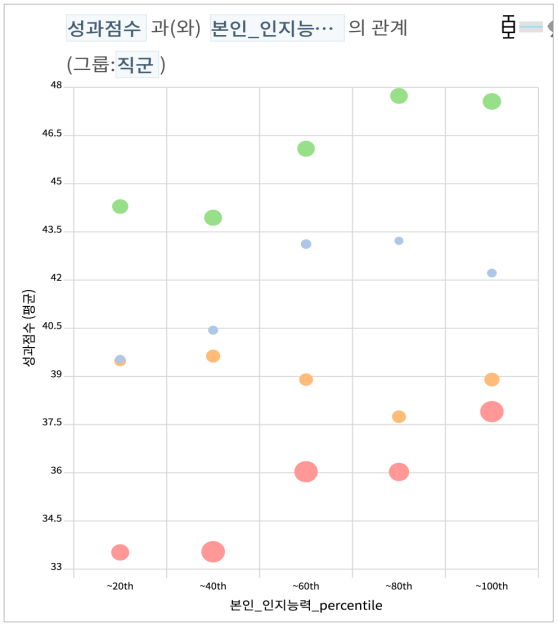
Data: A New Language of Business
Data Viz: A Medium of Data Communication

목적에 맞는 데이터 시각화 방법 선택

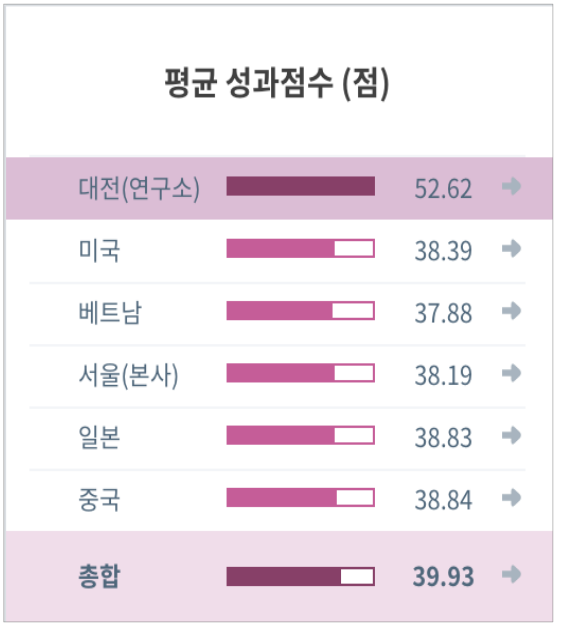
Visual Confirmation
가설을 시각적으로 검증



Visual Exploration
뭐라도 하나 걸렸으면



Visual Affirmation
검증된 사실을 주장/보고



Correlation and Scatter Plot (source: WHY)

1 기온(X)과 상인수(Y)

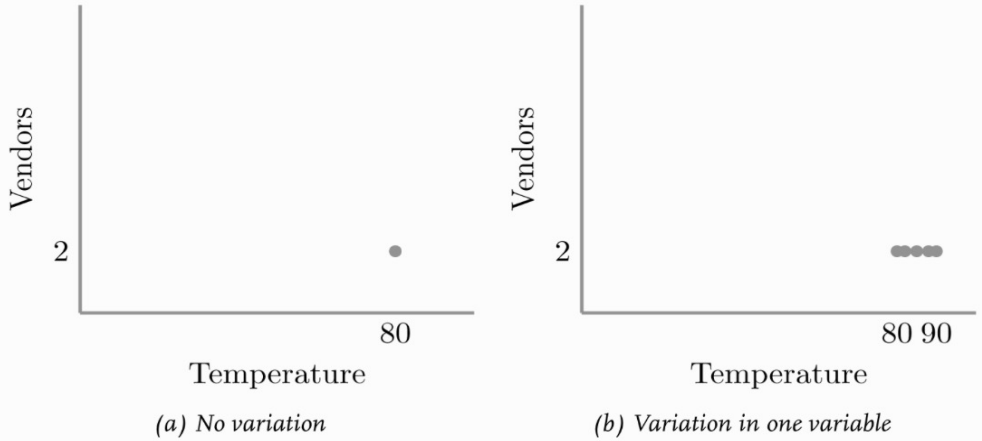
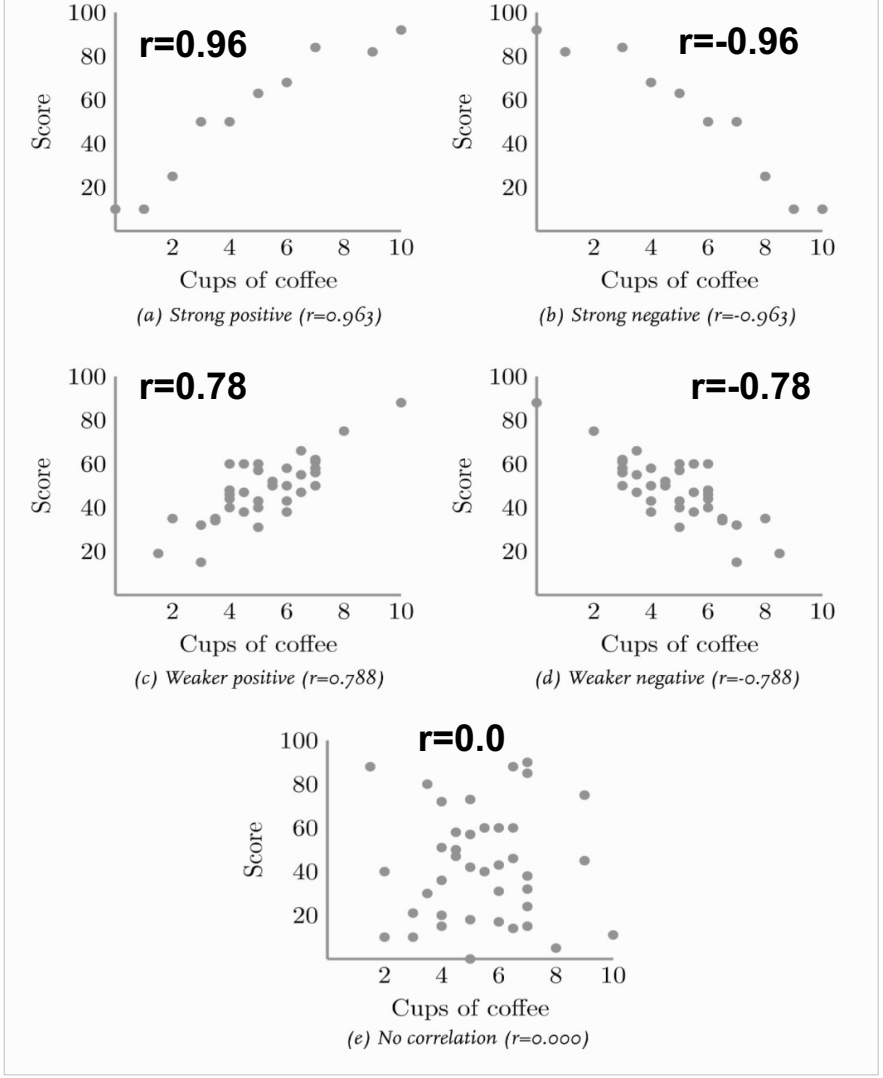


Figure 3-2. Without variation in both variables, we cannot find a correlation.

2 커피음용량(X)과 시험점수(Y)



3

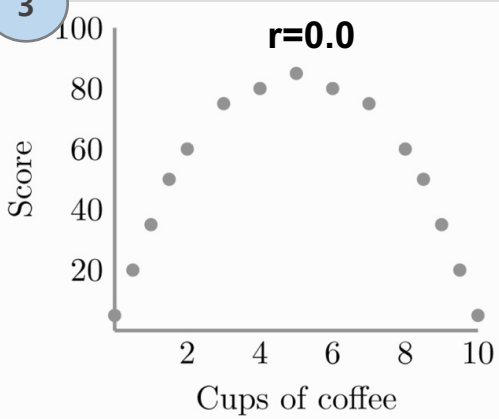


Figure 3-4. Nonlinear relationship (r=0.000).

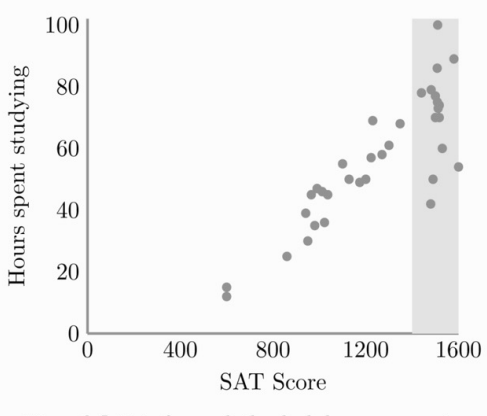
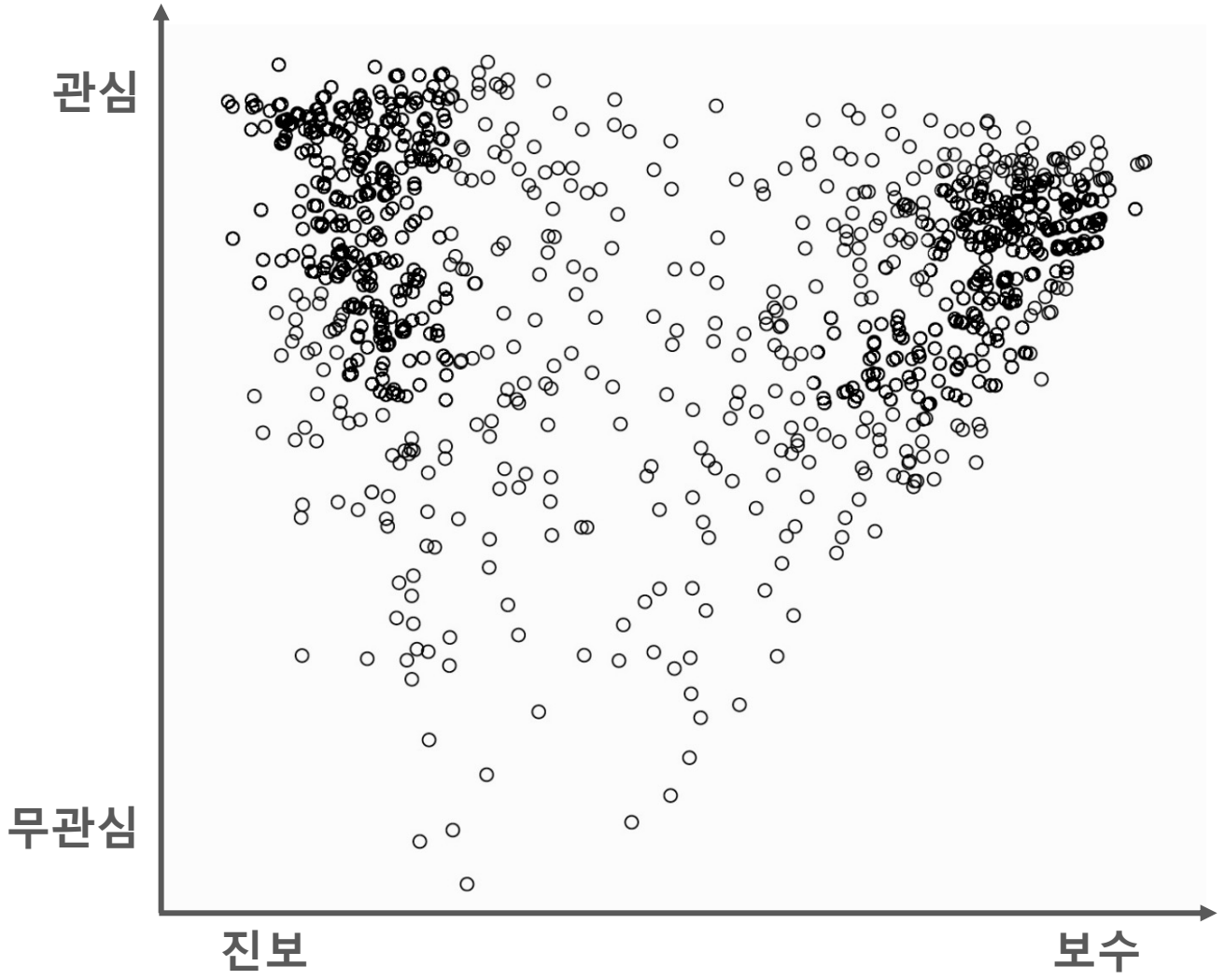


Figure 3-5. Data from only the shaded area represents a restricted range.

Uncorrelated Doesn't Mean Unrelated (source: how not to be wrong)

"정치적 관심도"와 "정칙적 성향(보수-진보)" 간 관계 해석



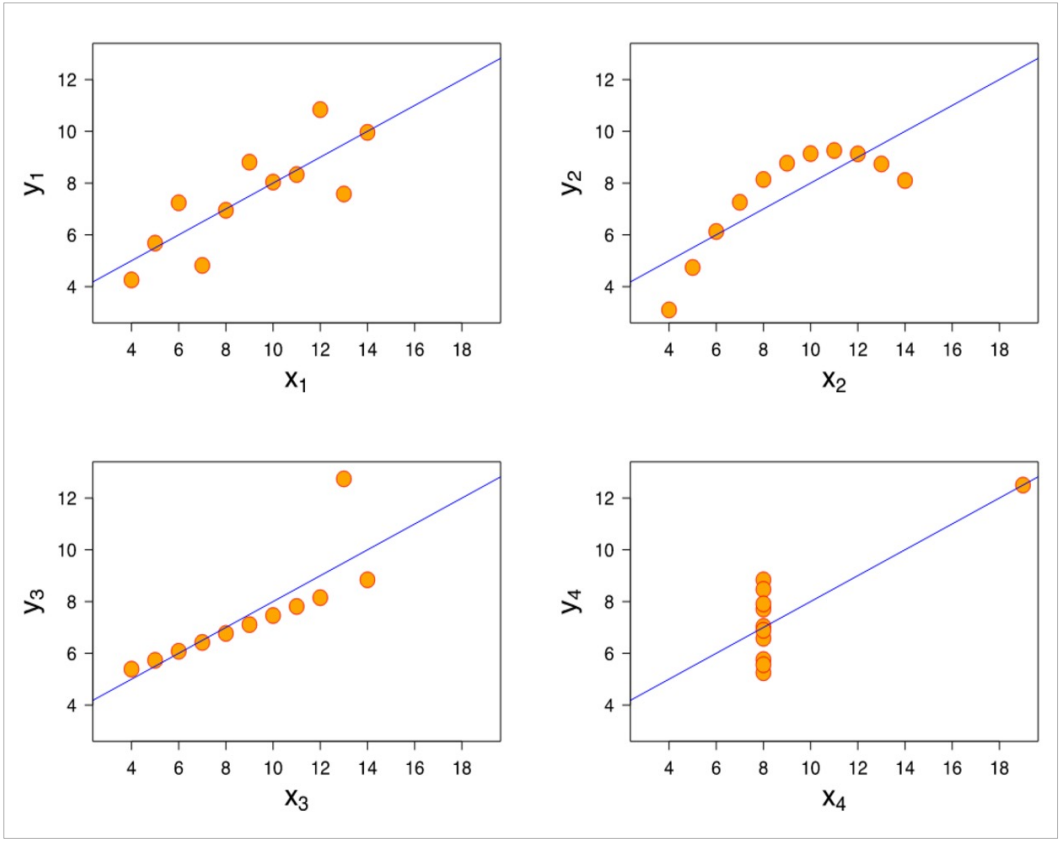
평균: 평균적 이해; 시각화: 차이에 대한 이해

동일한 평균, 분산, 상관계수

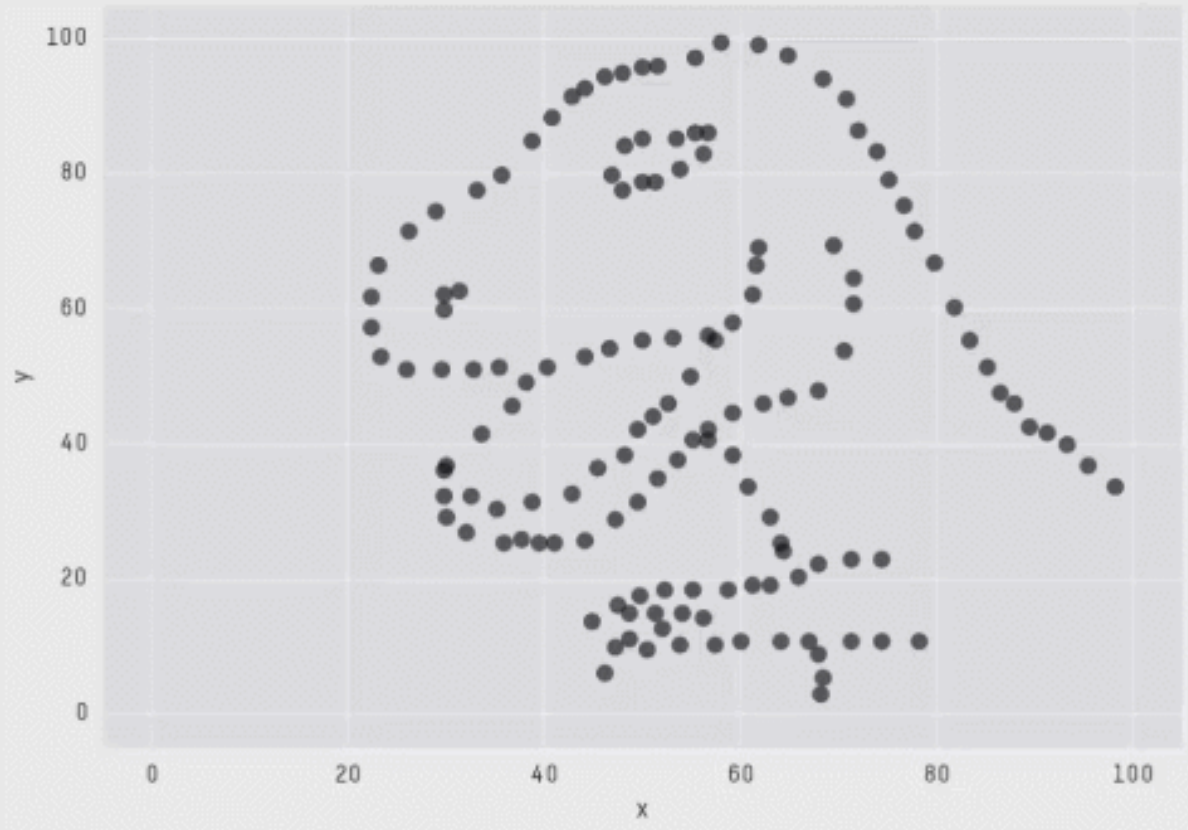
	I		II		III		IV	
	X	Y	X	Y	X	Y	X	Y
평균	9	7.5	9	7.5	9	7.5	9	7.5
분산	11	4.1	11	4.1	11	4.1	11	4.1
상관계수	0.82		0.82		0.82		0.82	

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

시각화를 통해 현실의 복잡성이 드러남



통계치와 시각화 결과를 함께 확인

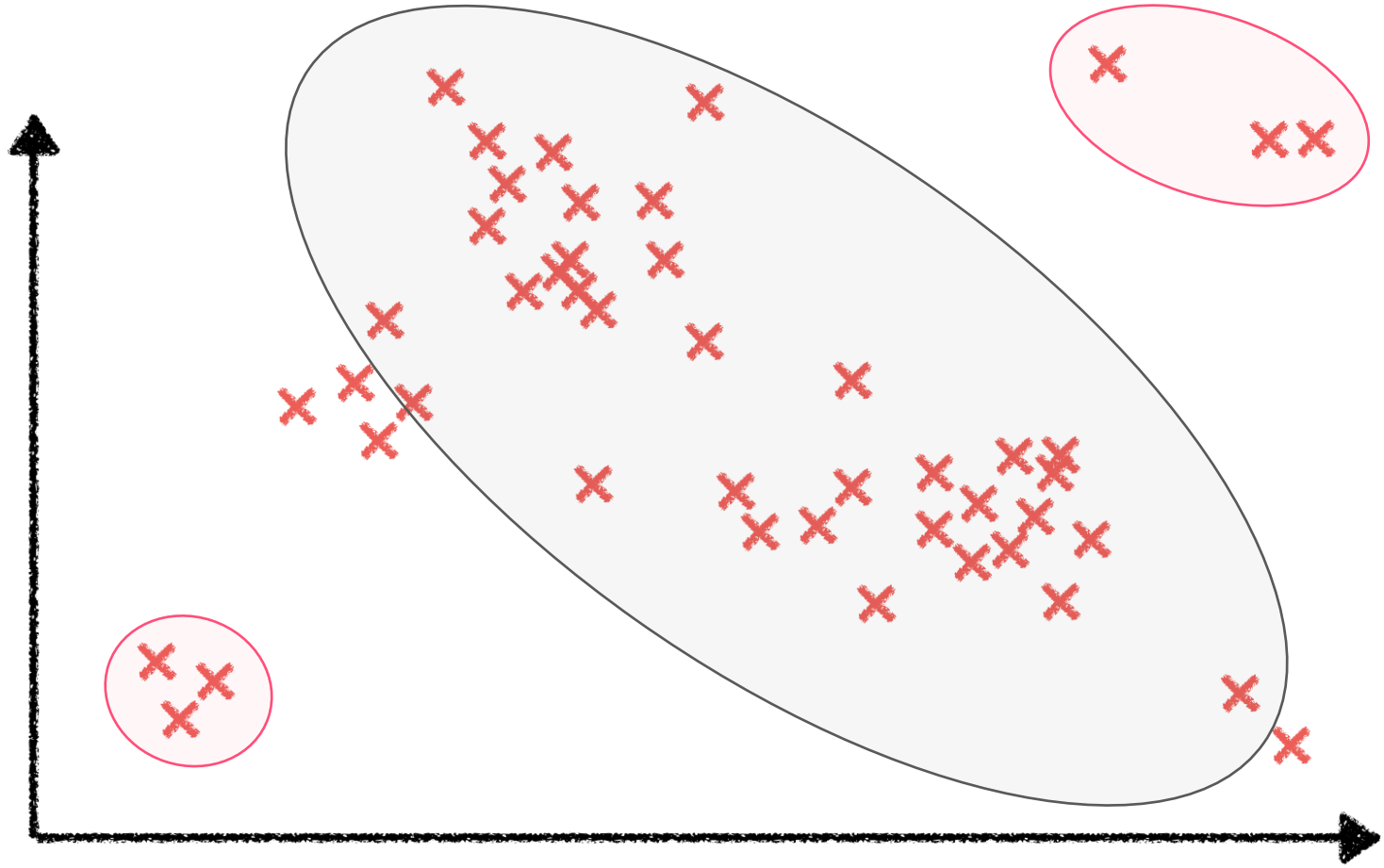


X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Data Viz: Understand and Explain Data

Signal
패턴, 일반적 경향

Noise?
이상한 애? vs. 특별한 애?

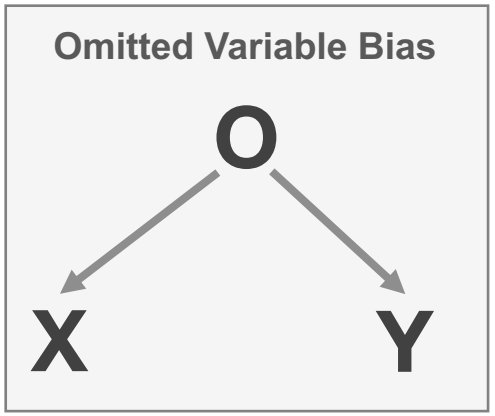


Correlation vs. Causation

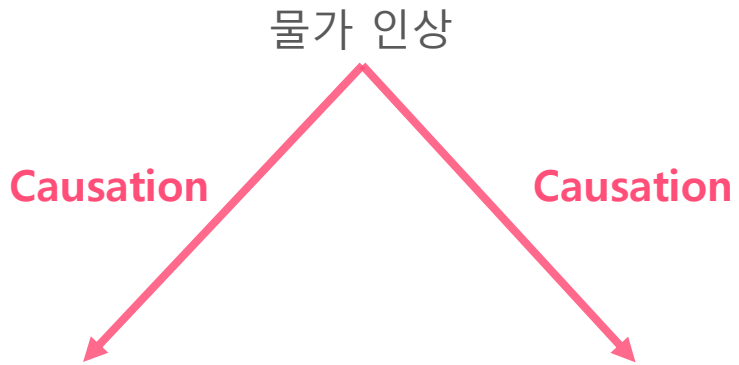
Correlation helps you predict the future;
Causality lets you change the future.

- Correlation: X의 변화로 Y의 변화 예측 가능
- Causation: X에 개입해서 Y를 바꿀 수 있음

대표적인 오류



교통비 인상 ⟷ **Correlation** ⟷ 백반 가격 인상



교통비 인상 ⟷ **Correlation** ⟷ 백반 가격 인상

Correlation을 의사결정에 어떻게 활용하나?

수영장 장난감 광고를 하려고 하는데 누가 수영장을 소유했는지 모름