집계 말고 개별 레코드 수준의 시각화를 통해
데이터에 대한 평균적 이해 뛰어넘기

EDA(Exploratory Data Analysis) = DESCRIBE (기술 분석) + EXPLORE (탐험 분석)

EDA, 데이터와 함께 떠나는 창의적 여행

- inspect data structure
- data quality
- summarize
- visualize data
- hypothesis generation
- != modeling

Source: Booz Allen Hamilton

데이터에 대해 사실적으로 묘사하는 법

Description
요약

변수의 대표값과
모양이 어떻나?

개별 변수(Y)의
통계값과 분포 확인

Comparison
비교

변수값의 차이가
어디서 얼마나 나나?

서로 다른 범주(X) 간 Y의
특성·모양 비교

Relationship
관계

변수(Y)의 변화와 관계를
갖는 다른 변수(X)는?

X와 Y 사이의
상관관계 파악

The Problems with Average: Not Robust!

상관관계 - Correlation and Scatter Plot (source: WHY)

1 기온(X)과 상인숫자(Y)

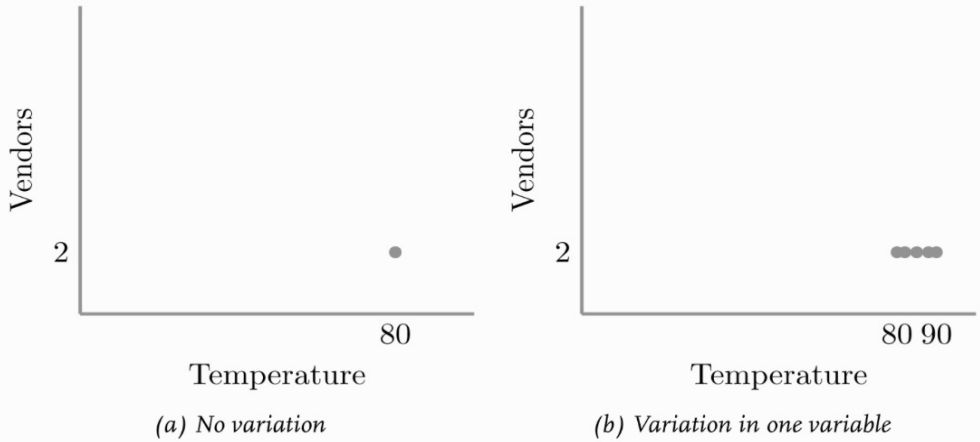


Figure 3-2. Without variation in both variables, we cannot find a correlation.

3

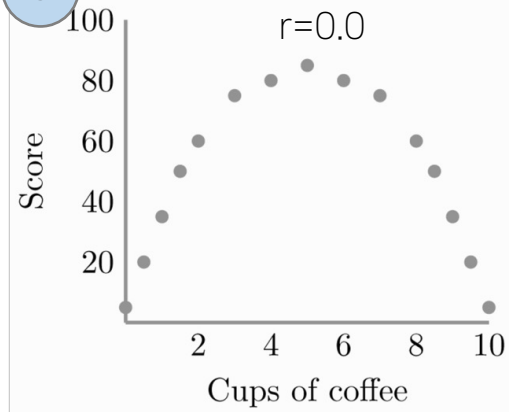


Figure 3-4. Nonlinear relationship (r=0.000).

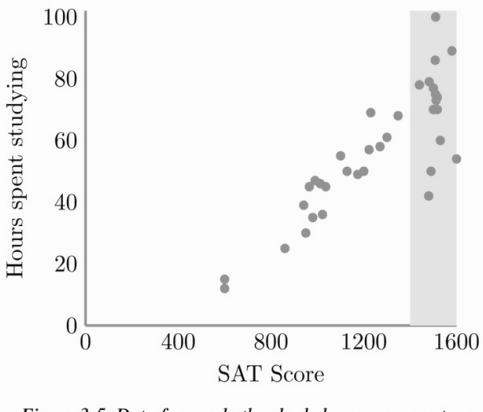
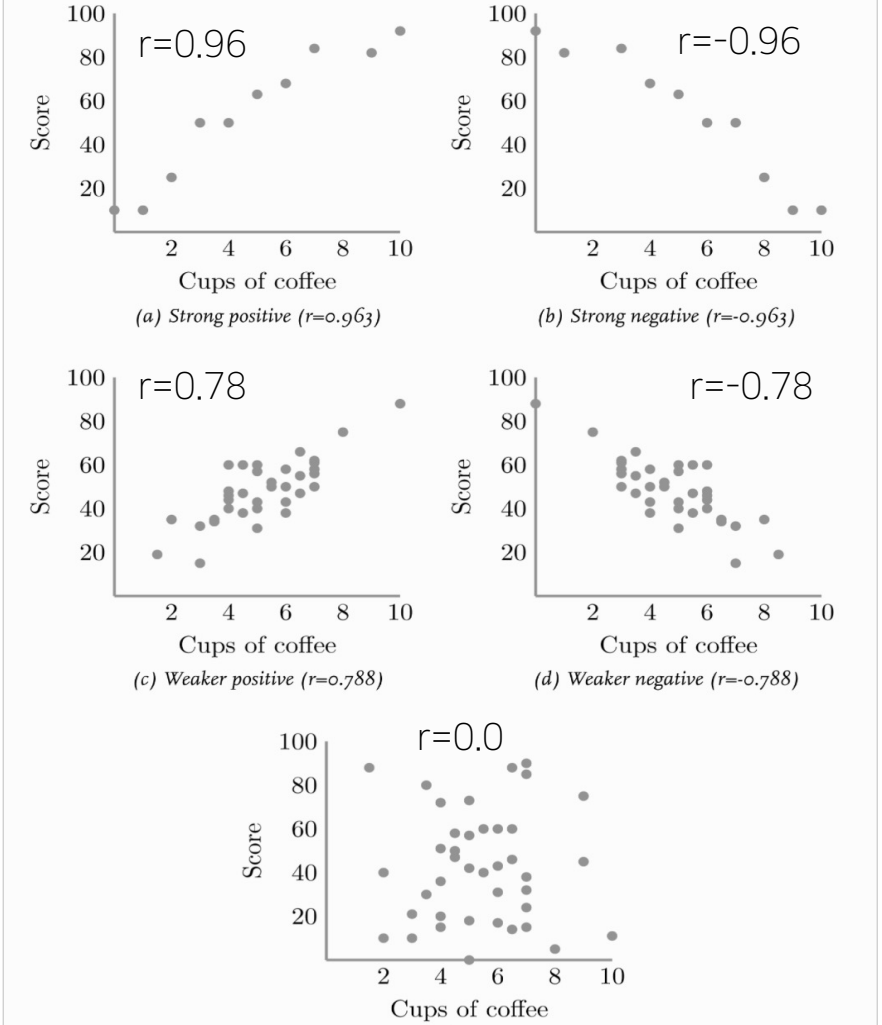


Figure 3-5. Data from only the shaded area represents a restricted range.

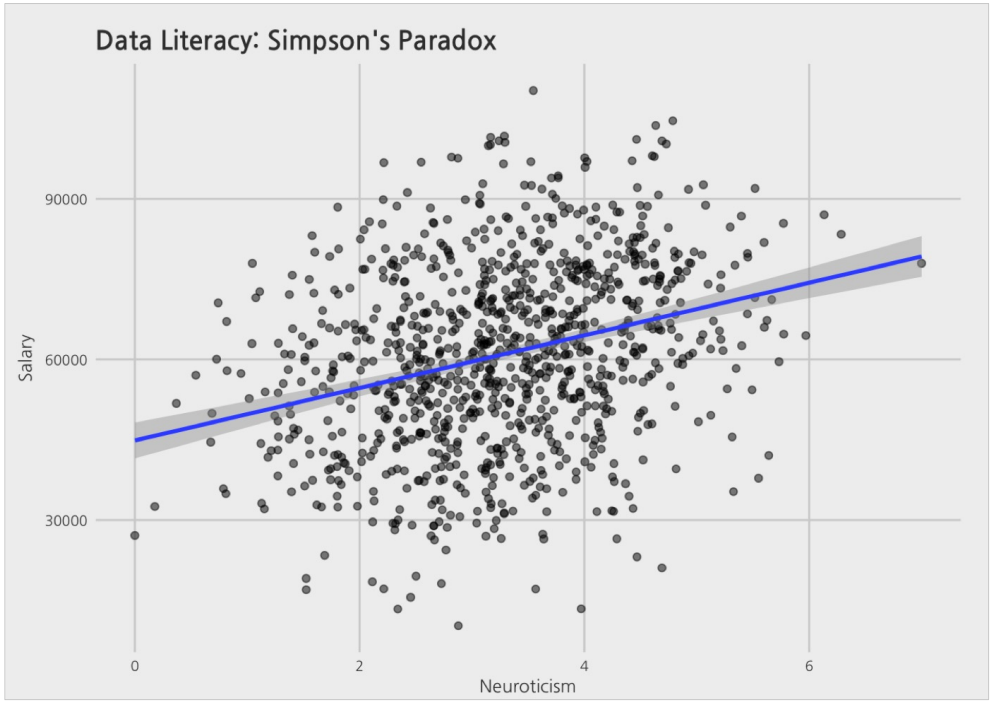
2 커피(X)과 시험점수(Y)

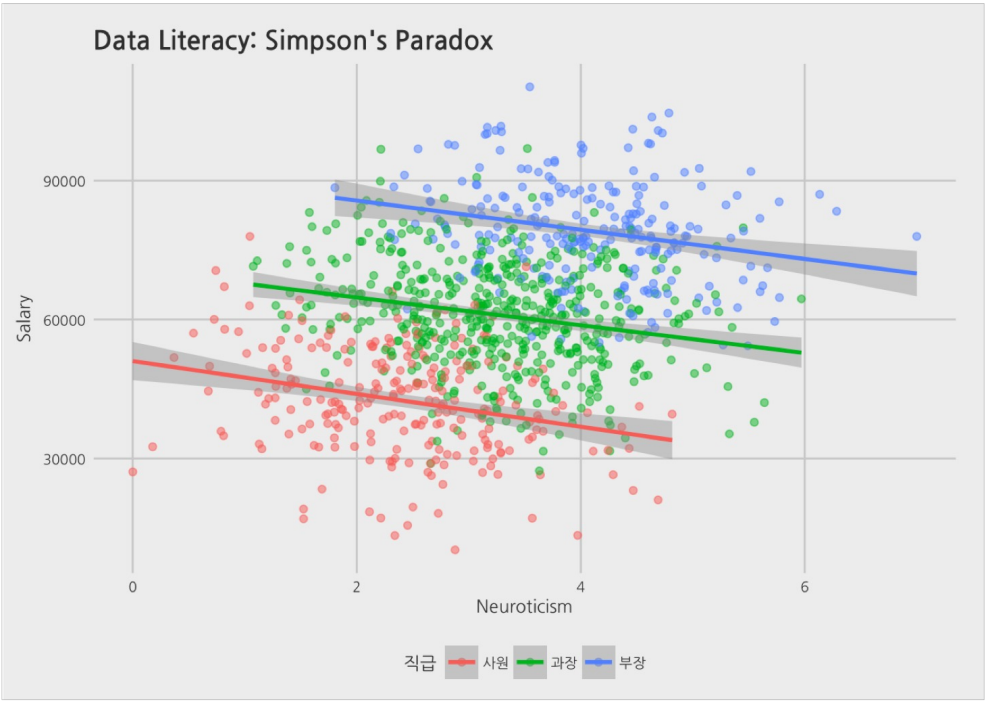


뻔하지 않은 결과 - 제한된·익숙한 관점 탈피

새로운 관점(Dimension) 추가

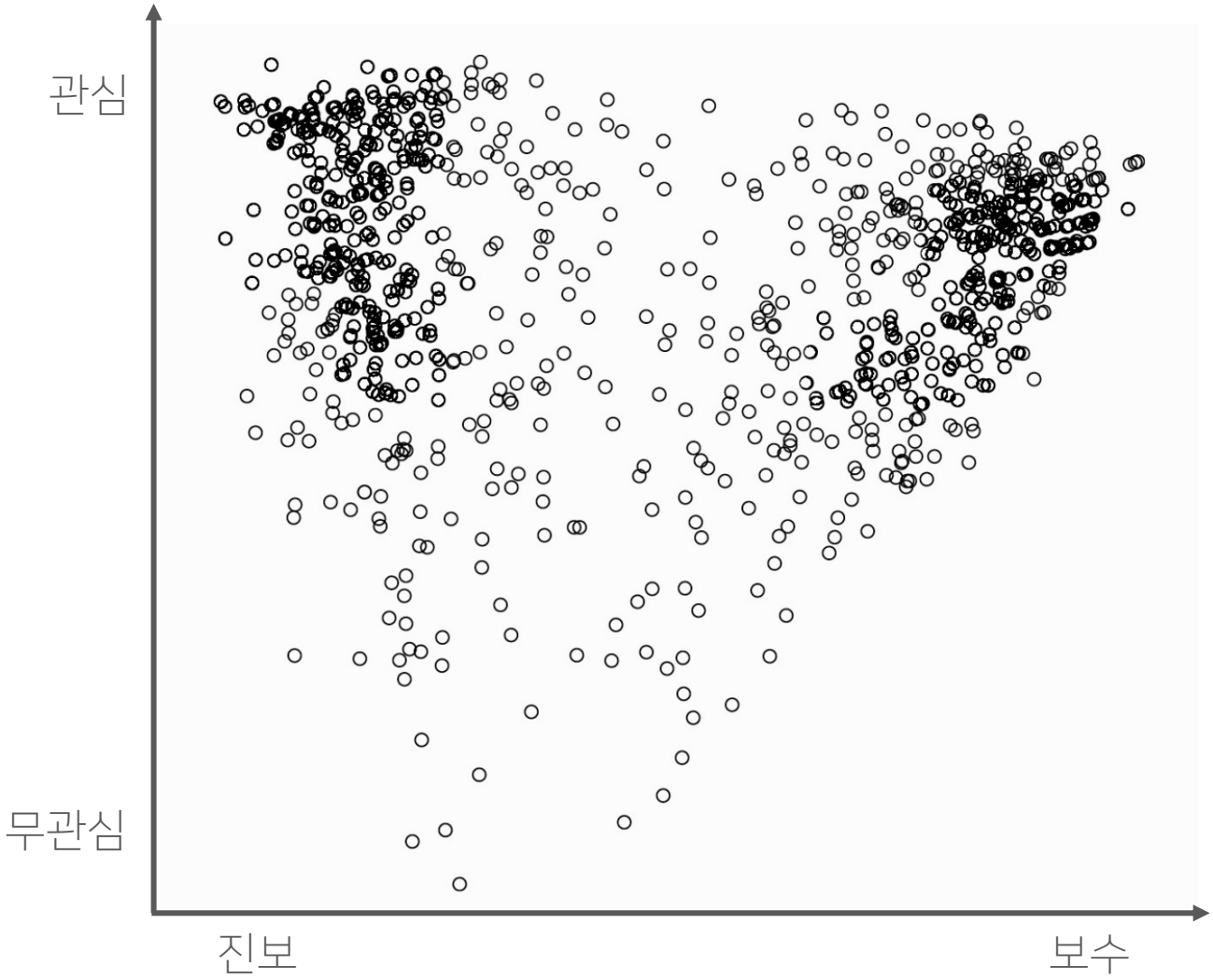
연봉과 까칠함과의 관계 → 직급별 연봉과 까칠함과의 관계





Uncorrelated Doesn't Mean Unrelated

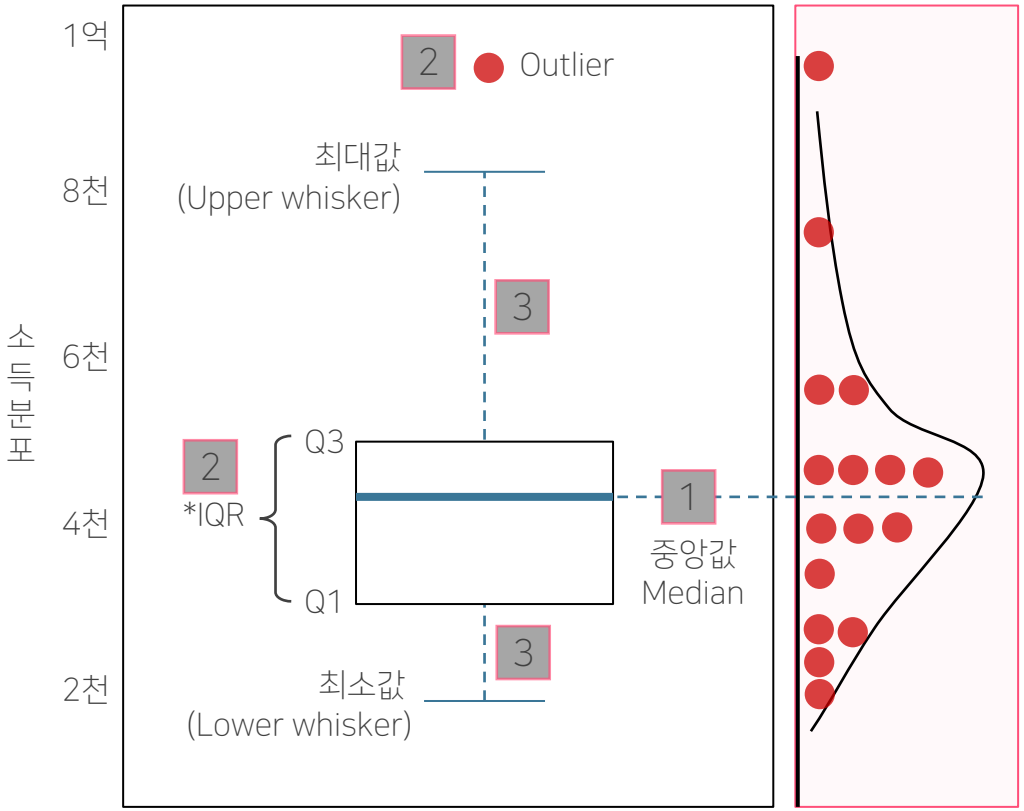
"정치적 관심도"와 "정치적 성향(보수-진보)" 간 관계 해석



(source: how not to be wrong)

Distribution (boxplot)

Box-and-whiskers plot (예시)



Box-and-whiskers plot 해석

1 중심 경향

- 중앙값(median) 파악

2 특이값(Outlier)

- *IQR(Inter Quartile Range)의 1.5배 이상 벗어나 있는 값들을 Outlier로 정의

3 대칭성 및 분포

- 대칭성(symmetry): 최대값과 최소값까지의 수염 길이 비교

데이터 요약 & 시각화

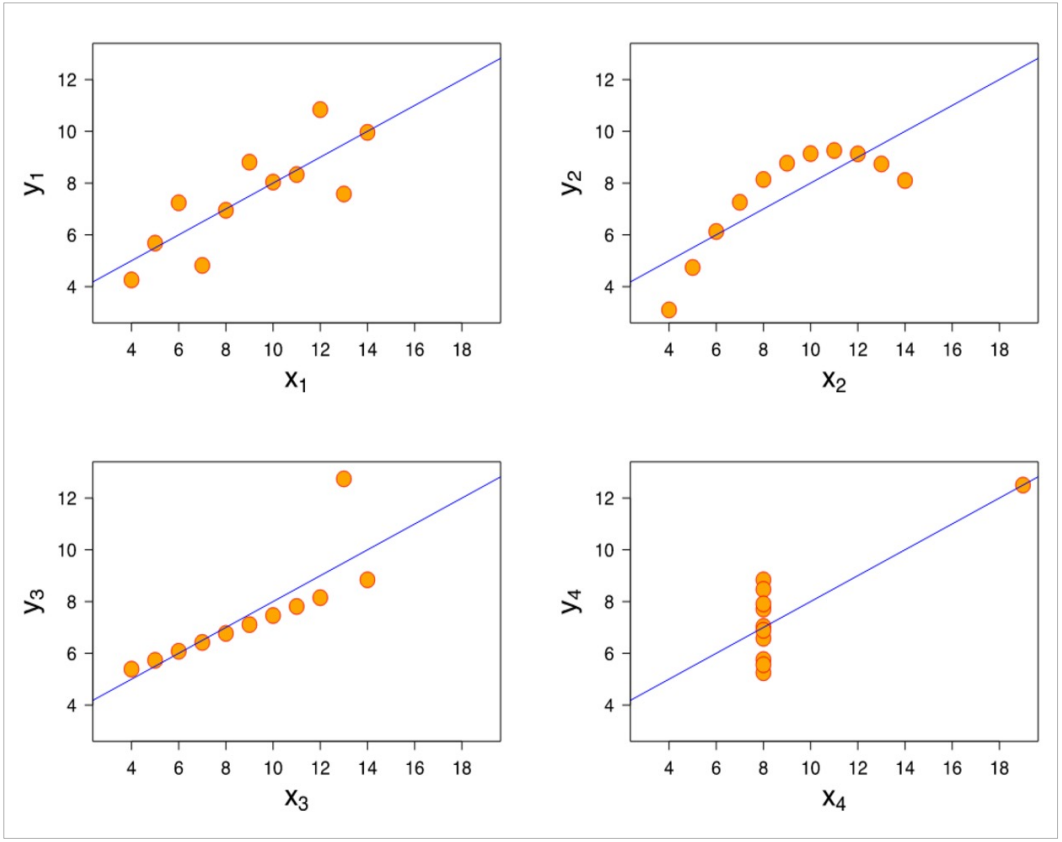
평균: 평균적 이해; 시각화: 차이에 대한 이해

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	

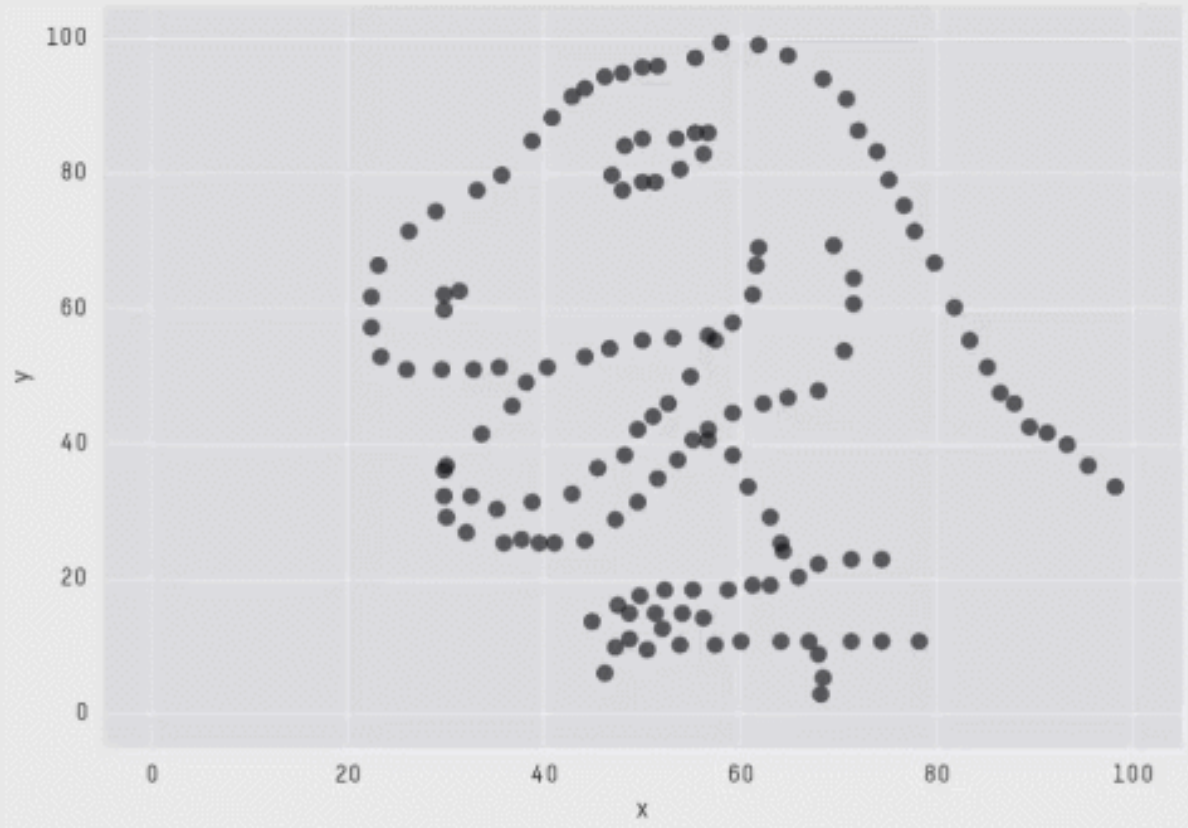
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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



Data Visualization - Seeing is Believing

통계치와 시각화 결과를 함께 확인



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

Secret to Success: Finding the Right Angle

Analysts pinpointed the range of 25-35 degrees as the sweet spot for home runs, when paired with an exit velocity of 95 mph or greater.