Сила визуализации

Пример

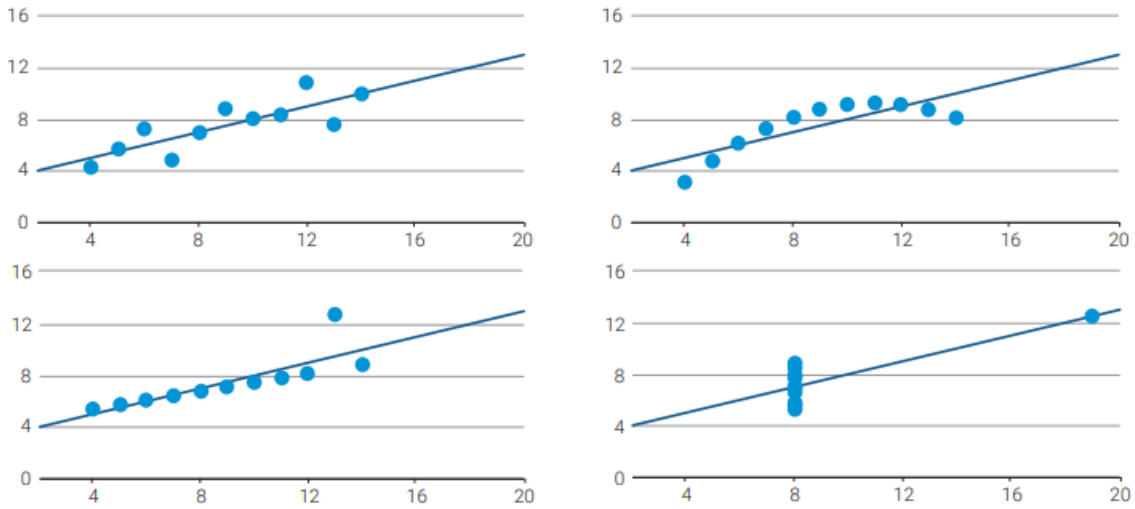
Силу визуализации можно проиллюстрировать на простом примере, известном как квартет Энскомба (https://www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf). Это четыре набора данных, которые почти идентичны по описательным характеристикам, но имеют разное распределение и при графическом представлении дают совершенно разную картину. Каждый набор данных состоит из 11 точек (x, y). Эти наборы данных были разработаны в 1973 году специалистом по статистике Фрэнсисом Энскомбом, чтобы продемонстрировать, как важно перед анализом данных представить их в виде графика. Тем самым он опровергал расхожее представление в среде специалистов по статистике о том, что «числовые расчеты точны, а графики есть лишь грубое приближение».

В таблице 1 представлены наборы данных, разработанные Энскомбом. В первых трех из них значения x одинаковы. В конце таблицы приводится несколько описательных статистик. Основные статистические характеристики одинаковы, поэтому можно предположить, что четыре набора данных похожи, но, если представить их в виде диаграмм (рис. 1), различия становятся очевидны с первого взгляда. Визуализация данных может дать информацию, которую не всегда дают табличные данные и описательные статистики.

Таблица 1.Квартет Энскомба

	Набор данных I		Набор данных II		Набор данных III		Набор данных IV	
Наблюдение	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8,04	10	9,14	10	7,46	8	6,58
2	8	6,95	8	8,14	8	6,77	8	5,76
3	13	7,58	13	8,74	13	12,74	8	7,71
4	9	8,81	9	8,77	9	7,11	8	8,84
5	11	8,33	11	9,26	11	7,81	8	8,47
6	14	9,96	14	8,1	14	8,84	8	7,04
7	6	7,24	6	6,13	6	6,08	8	5,25
8	4	4,26	4	3,1	4	5,39	19	12,5
9	12	10,84	12	9,13	12	8,15	8	5,56
10	7	4,82	7	7,26	7	6,42	8	7,91
11	5	5,68	5	4,74	5	5,73	8	6,89
Сводная статистика								
Число	11	11	11	11	11	11	11	11
Среднее	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
Среднеква-дратичное отклонение	3,16	1,94	3,16	1,94	3,16	1,94	3,16	1,94
Коэффициент корреляции	0,82		0,82		0,82		0,82	

Рис. 1.Визуализация квартета Энскомба



Источник данных Anscombe
(https://www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf)

Без визуального представления информации аудитории может быть трудно понять истинное значение найденных закономерностей. Например, приводя таблицу со значениями ожидаемой продолжительности жизни населения за последние 20 лет, мы не даем аудитории веских причин проецировать эти данные на себя, однако если она увидит график, на котором виден рост или снижение ожидаемой продолжительности жизни – желательно в сравнении с данными других стран, – мы обязательно привлечем ее внимание.