

Python для анализа данных

МИРЭК | 4 модуль

Почему так важна визуализация статистических данных?

Квартет Энскомба (1973)

Четыре набора числовых данных, у которых простые статистические свойства идентичны, но их графики существенно отличаются. Каждый набор состоит из 11 пар чисел. Квартет был составлен математиком Ф. Дж. Энскомбом для иллюстрации важности применения графиков для статистического анализа, и влияния выбросов значений на свойства всего набора данных.

“A computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.”

– Francis Anscombe, 1973

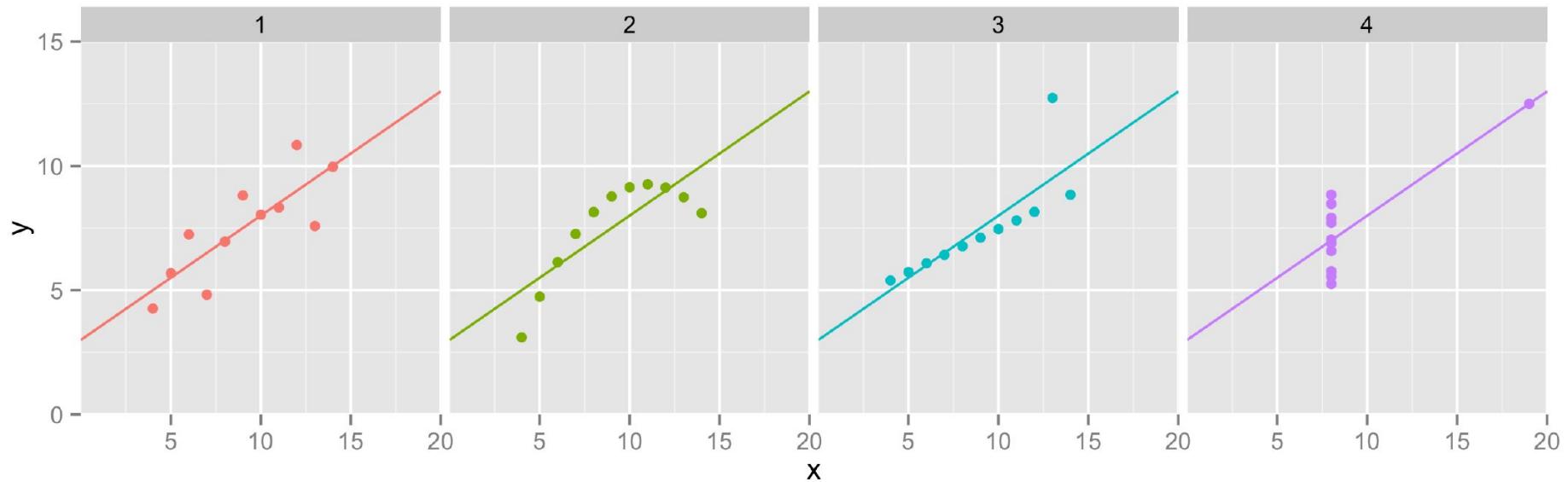
Квартет Энскомба: данные

Group 1		Group 2		Group 3		Group 4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.56
9.0	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

Квартет Энскомба: описательные статистики

	Group 1	Group 2	Group 3	Group 4
mean(x)	9.00	9.00	9.00	9.00
mean(y)	7.50	7.50	7.50	7.50
var(x)	11.00	11.00	11.00	11.00
var(y)	4.13	4.13	4.12	4.12
correlation	0.82	0.82	0.82	0.82
lm intercept	3.00	3.00	3.00	3.00
lm x effect	0.50	0.50	0.50	0.50

Квартет Энскомба: визуализация



Зачастую,
визуализация –
единственный
способ увидеть, что
происходит
в данных

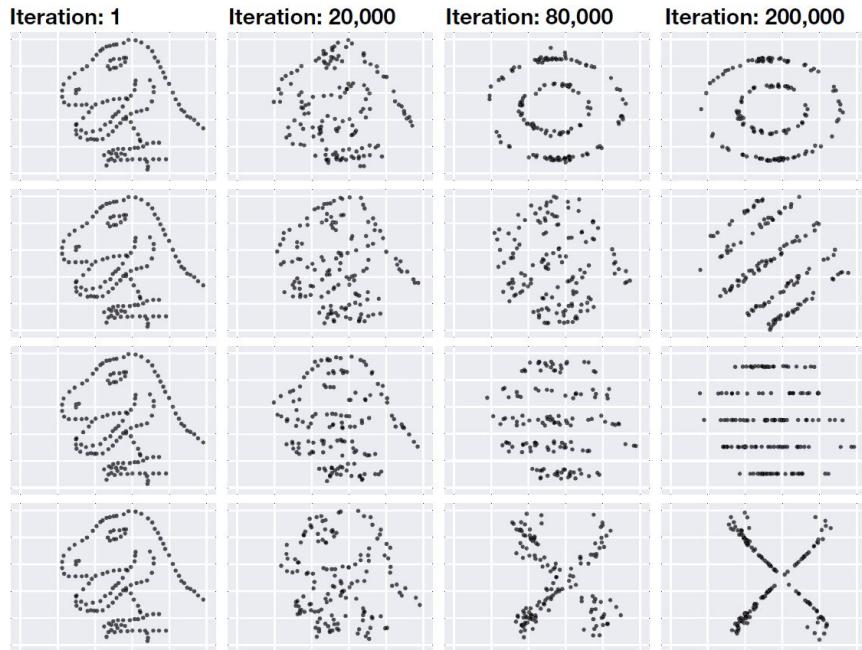


Figure 6. Creating a collection of datasets based on the “dinosaurus” dataset. Each dataset has the same summary statistics to two decimal places: ($\bar{x} = 54.26$, $\bar{y} = 47.83$, $s_{\bar{x}} = 16.76$, $s_{\bar{y}} = 26.93$, Pearson’s $r = -0.06$).

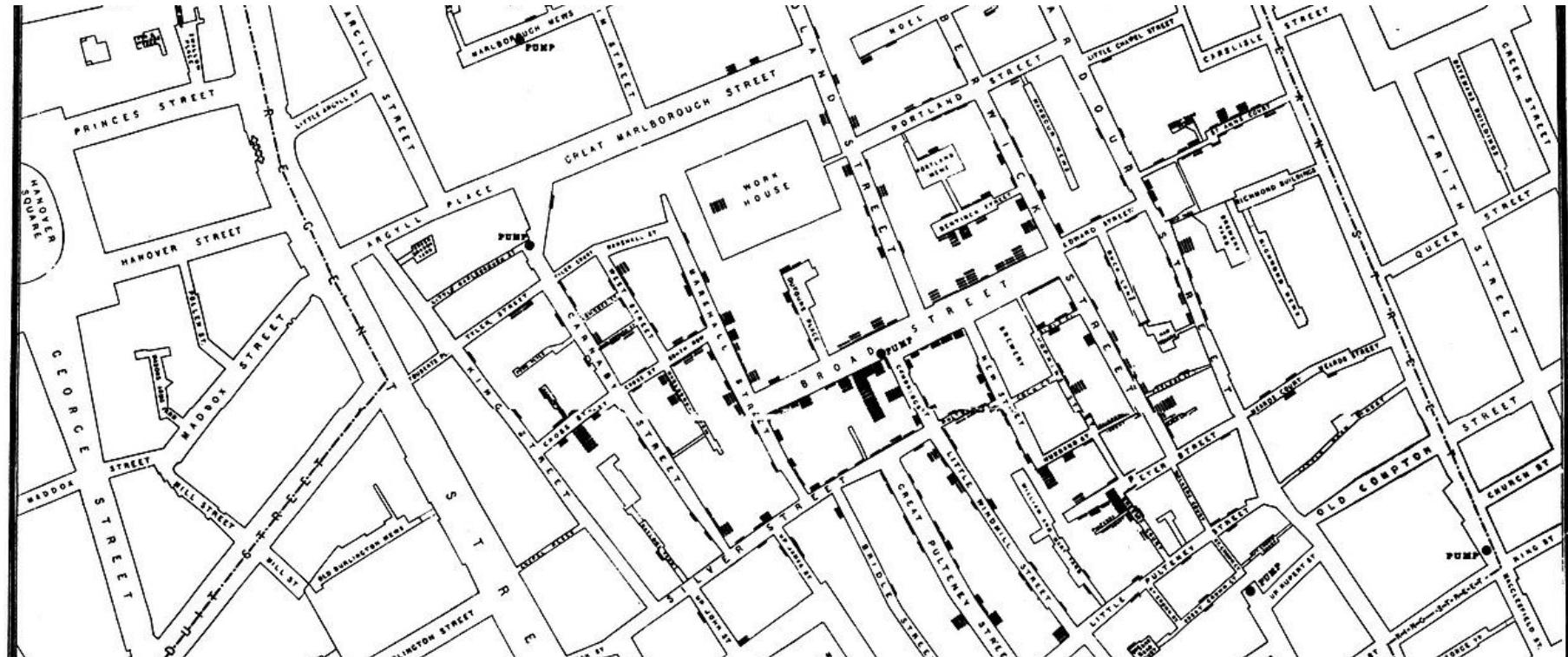
[https://www.autodeskresearch.com/publications/
samestats](https://www.autodeskresearch.com/publications/samestats)

Первые визуализации данных

Вспышка холеры в 1854

- Сделал простой график на карте: адреса пострадавших и расположение колонок с водой

<http://www.ph.ucla.edu/epi/snow/snowcricketarticle.html>



<https://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg>



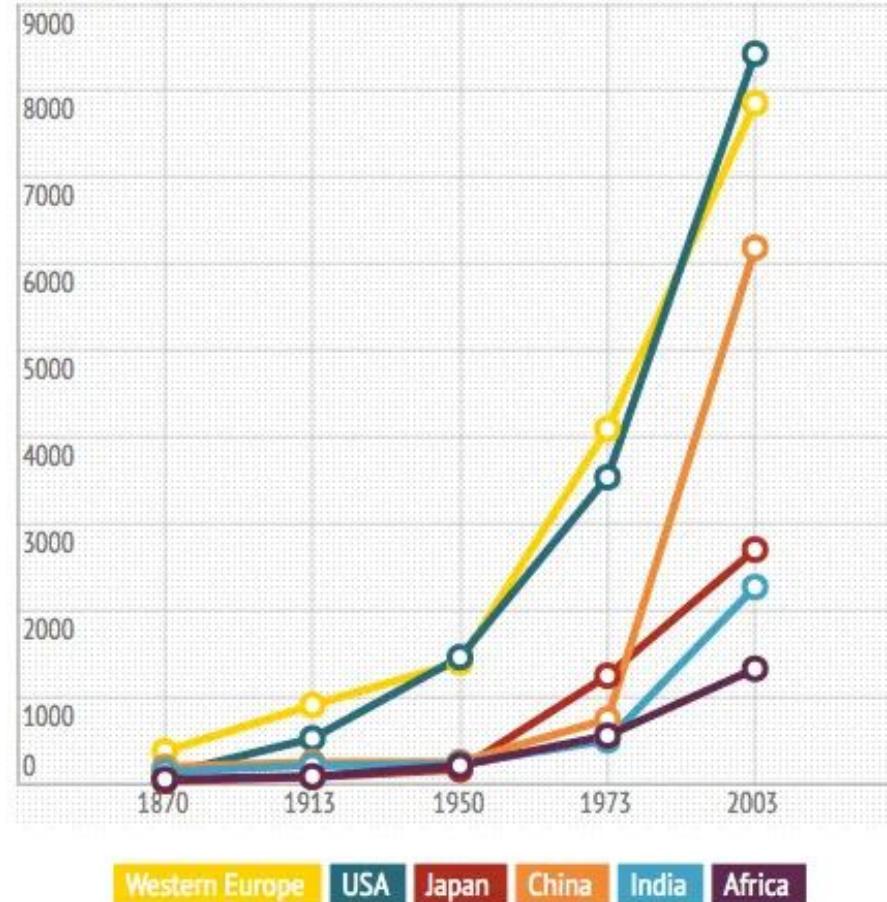
<https://www.theguardian.com/news/datablog/2013/mar/15/john-snow-cholera-map>

Зачем визуализировать данные?

	A	B	C	D	E	F
1	Past GDP	1870	1913	1950	1973	2003
2	Western Europe	367	902	1396	4096	7857
3	USA	98	517	1455	3536	8430
4	Japan	25	71	160	1242	2699
5	China	189	241	244	739	6187
6	India	134	204	222	494	2267
7	Africa	45	79	203	549	1322

Зачем визуализировать данные?

<http://www.mulinblog.com/data-visualization-matters/>

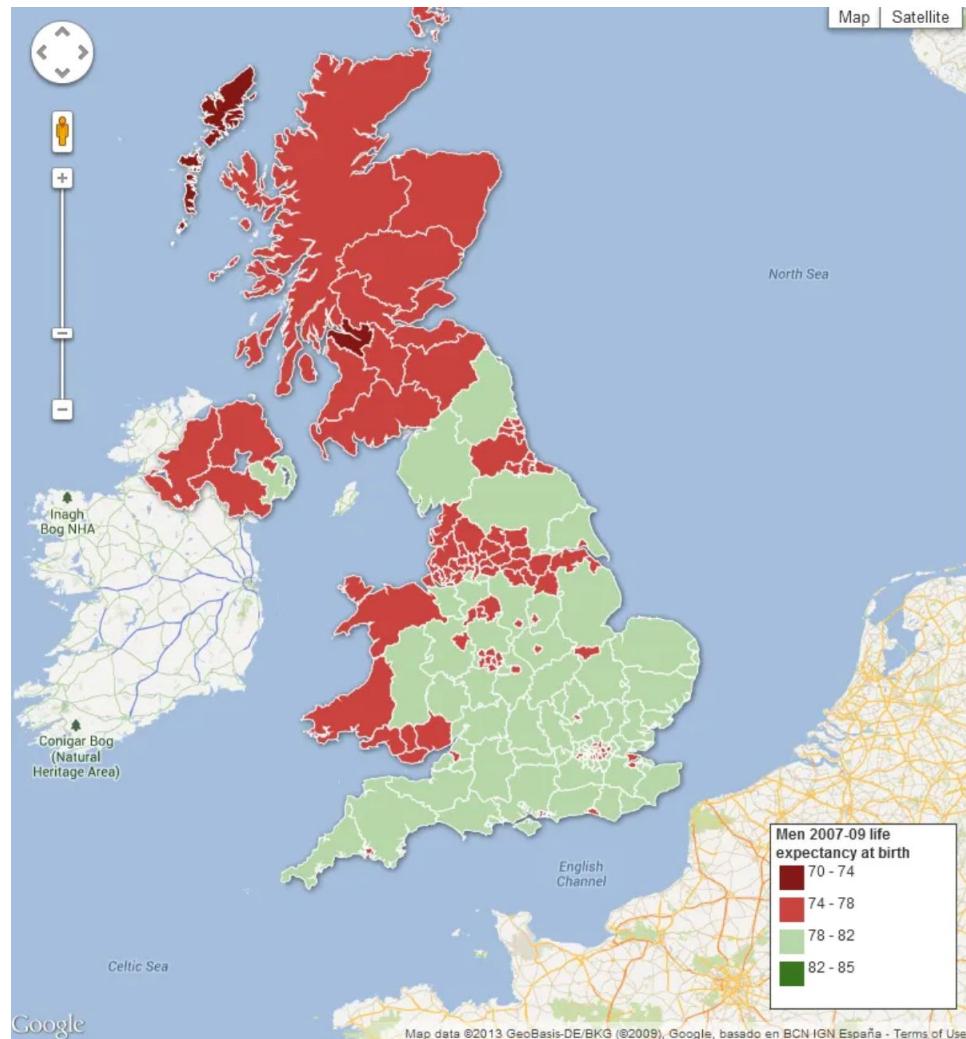


Зачем визуализировать данные?

Организация	81.0	1.5	76.4	1.9
Hillingdon	83.4	2.0	78.6	2.1
Hounslow	82.1	2.3	77.8	2.5
Hull Teaching	80.0	0.6	75.2	1.5
Hywel Dda	82.1	1.7	77.9	2.2
Isle of Wight National Health Service	83.2	1.5	79.1	2.1
Islington	81.2	1.7	75.4	2.1
Kensington and Chelsea	89.0	4.1	84.4	4.8
Kingston	83.7	2.4	80.7	3.0
Kirklees	80.9	1.6	76.7	1.7
Knowsley	79.8	1.8	75.9	2.5
Lambeth	81.1	1.3	76.4	2.9
Lanarkshire	79.2	1.2	74.4	1.4
Leeds	82.0	1.3	77.7	2.2
Leicester City	80.0	1.0	75.4	1.2
Leicestershire County and Rutland	83.4	1.7	79.7	1.7

Зачем визуализировать данные?

[http://www.mulinblog.com/data-visualizati
on-matters/](http://www.mulinblog.com/data-visualizati
on-matters/)

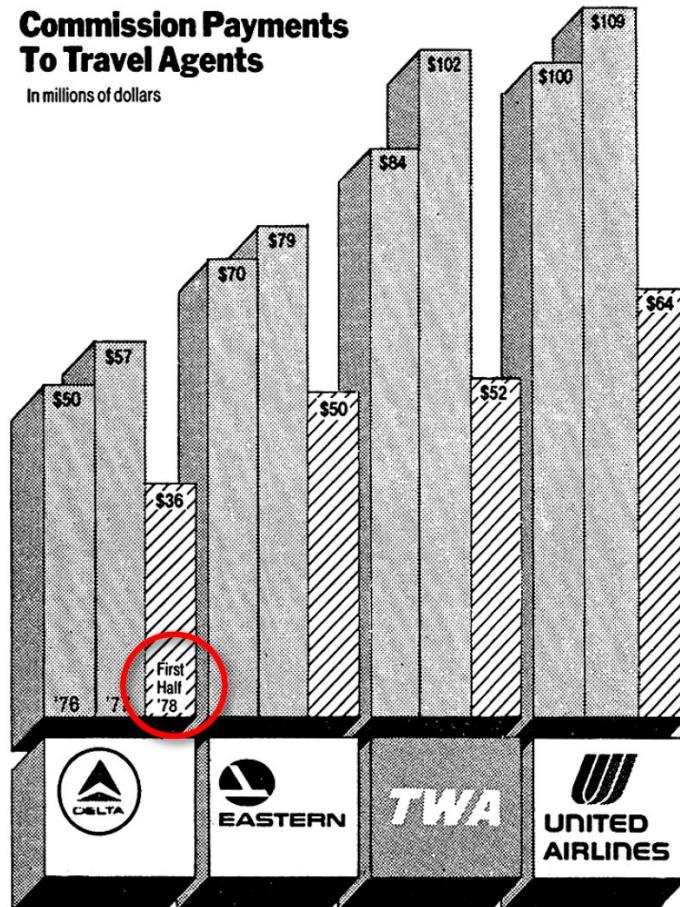


Честность данных (Э. Тафти)

- Отсутствующие шкалы и лэйблы
- Отсутствующий контекст
- Искаженные шкалы / искажающий дизайн

Commission Payments To Travel Agents

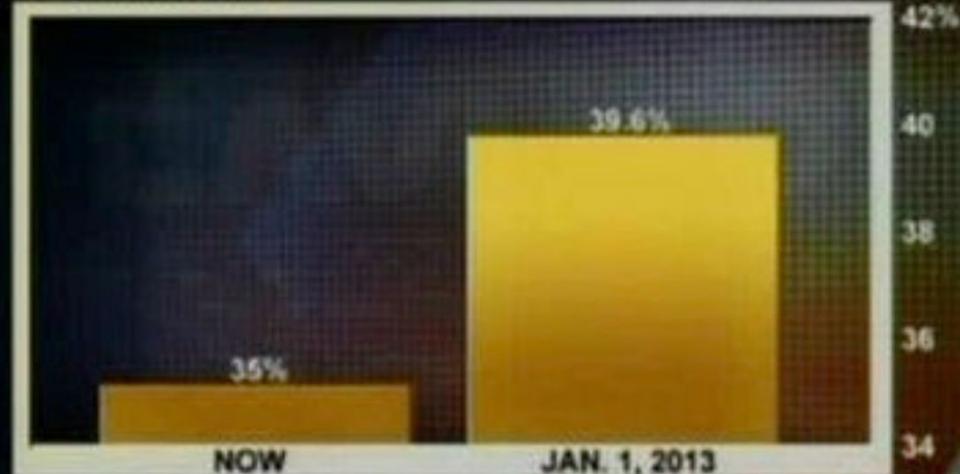
In millions of dollars



- Псевдо-снижение
- Полные годы сравниваются с полугодием

IF BUSH TAX CUTS EXPIRE

TOP TAX RATE



- Искажение шкалы (не с нуля)

8:01p ET

FOX
BUSINESS

TOP STORIES

TECHNOLOGY

CONSUMER

WITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T

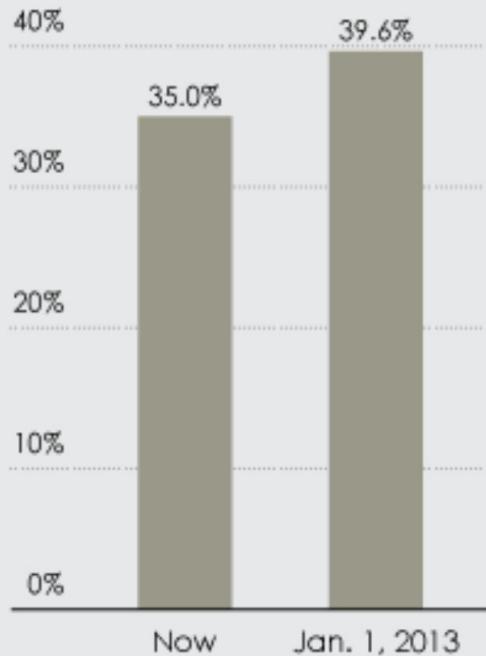
DOW 13008.68 □ 64.33

S&P 1379.32 □ 5.98

NASDAQ 2939.52 □ 6.32

If Bush tax cuts expire...

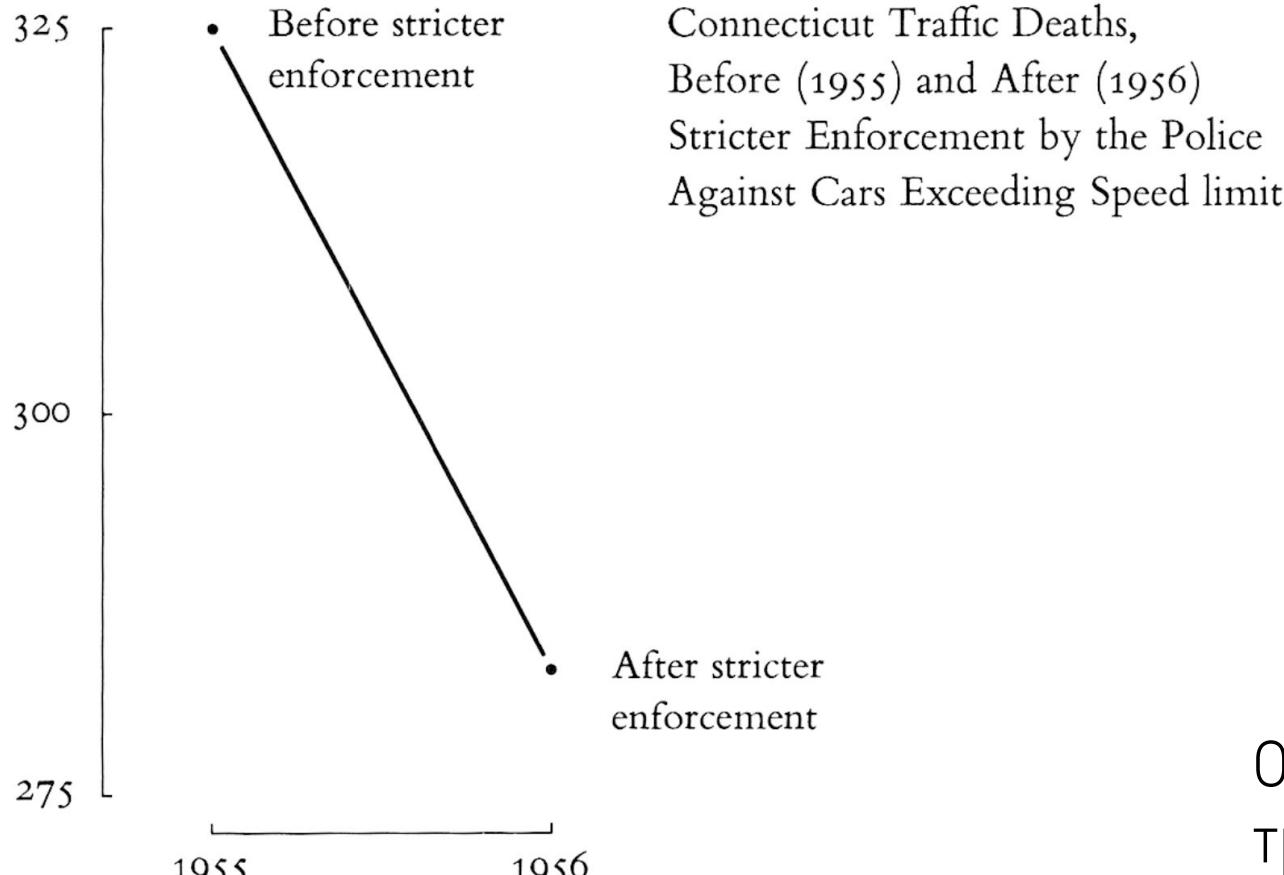
Top tax rate



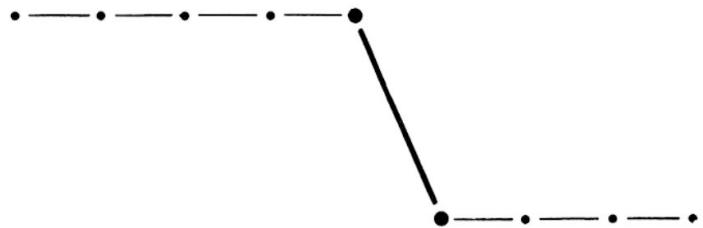
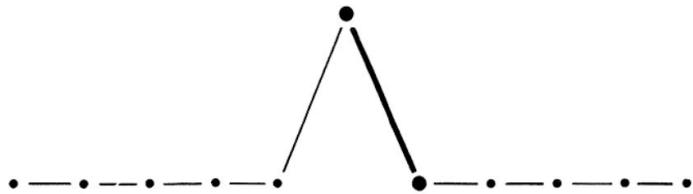
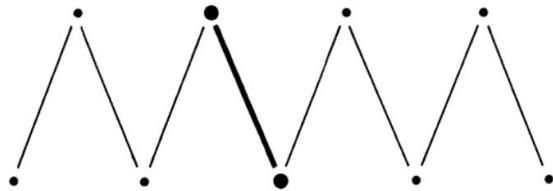
- Правдивый тренд (разница не такая значительная)

Честность данных

- Отсутствующие шкалы и лэйблы
- Отсутствующий контекст
- Искаженные шкалы / искажающий дизайн

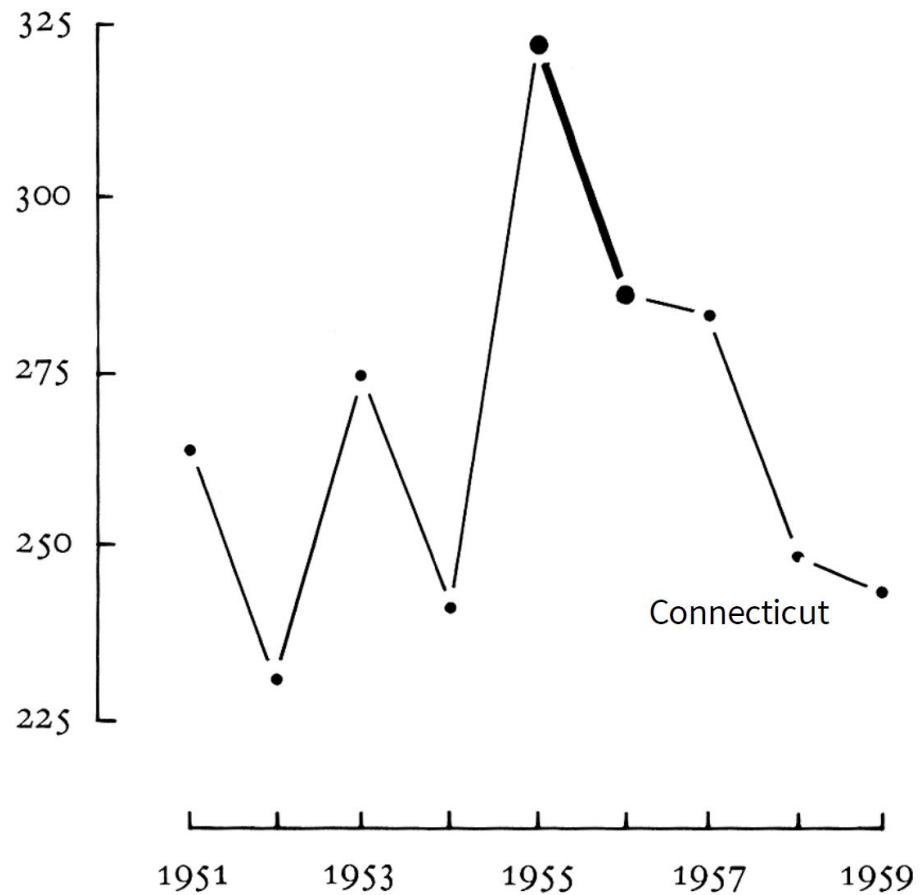


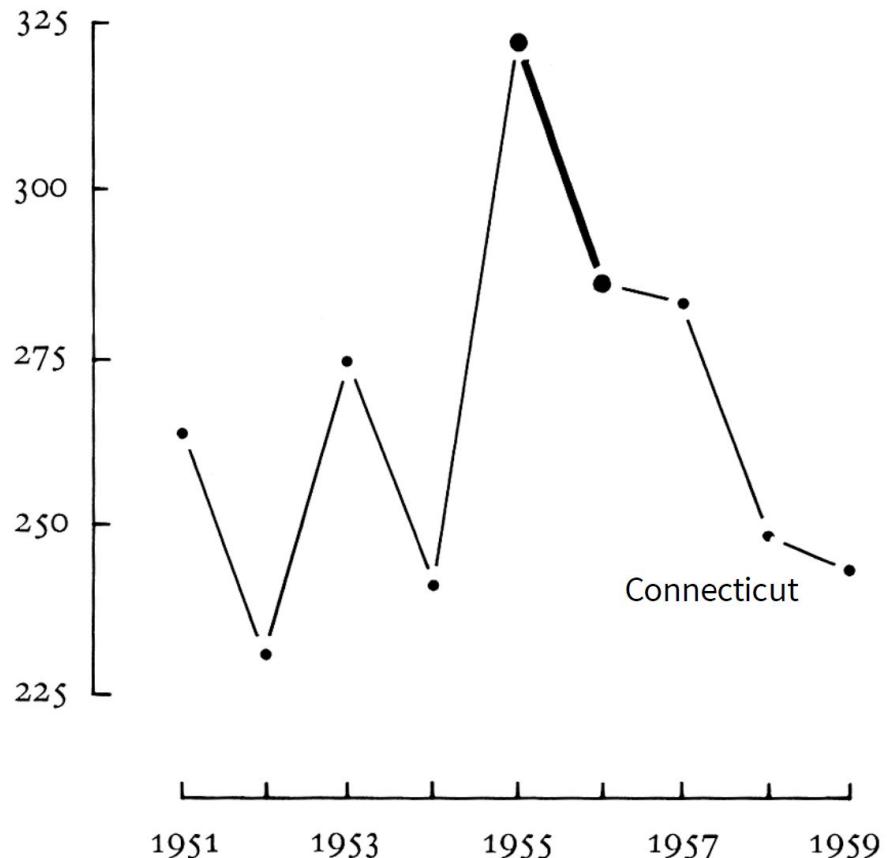
Опишите
тренд.



Откуда взяли этот
сегмент?

Откуда угодно!





VDQI Example (p74)



Честность данных

- Отсутствующие шкалы и лэйблы
- Отсутствующий контекст
- Исаженные шкалы / исажающий дизайн

**Comparative Annual Cost per Capita for care of Insane in
Pittsburgh City Homes and Pennsylvania State Hospitals.**



\$147

South Mountain



\$172

Pittsburgh



\$198

Harrisburg



\$213

Norristown



\$214

Warren

Pittsburgh Civic Commission, *Report on
Expenditures of the Department of Charities*
(Pittsburgh, 1911), p. 7.

Данные-к-чернилам (Data-Ink Ratio)

Пропорция данных к чернилам =

данные/общее количество чернил,

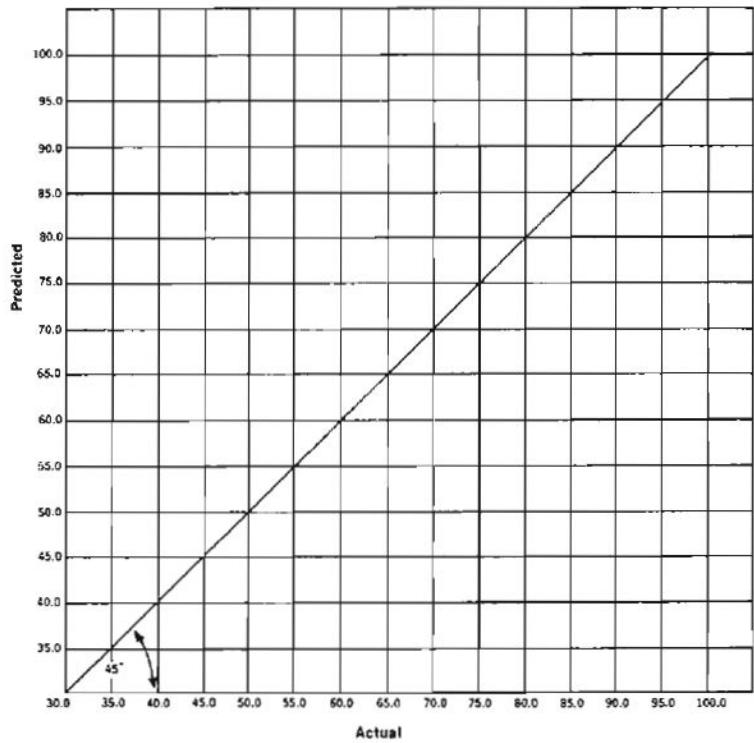
использованных в графике =

пропорция чернил в графике, потраченных на
неизбыточное отображение данных =

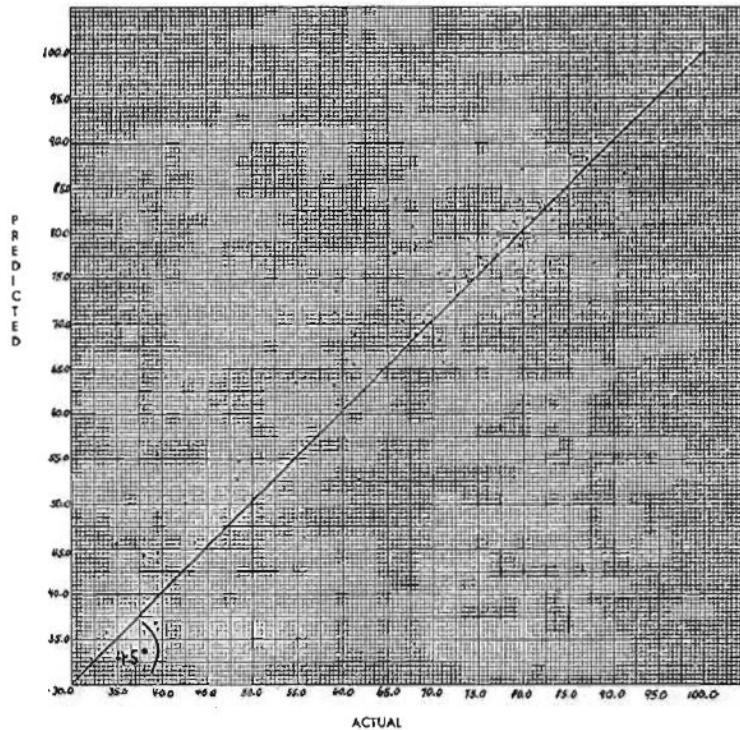
1.0 – пропорция графика, которую можно стереть без
потери данных

Данные-к-чернилам: низкая пропорция

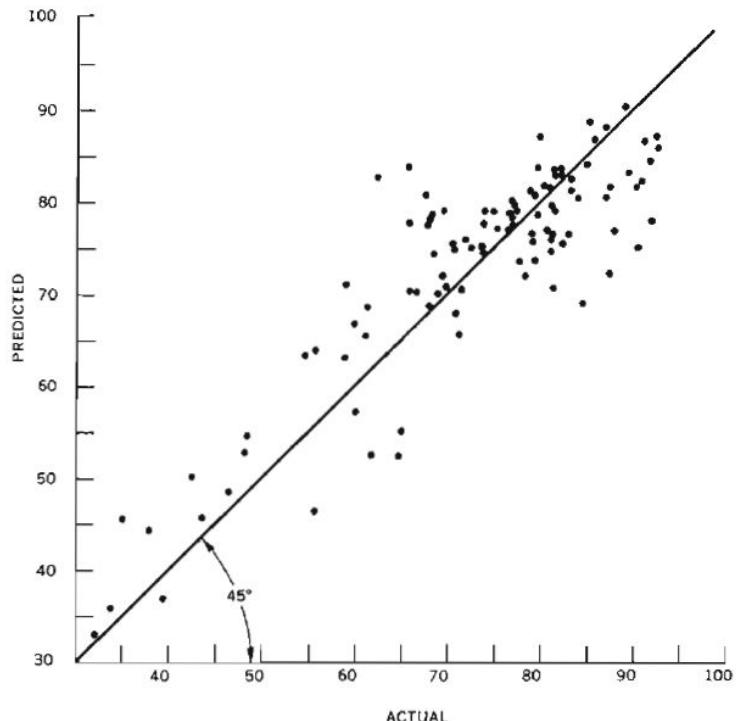
Figure 19.1 Relationship of Actual Rates of Registration to Predicted Rates
(104 cities, 1960)



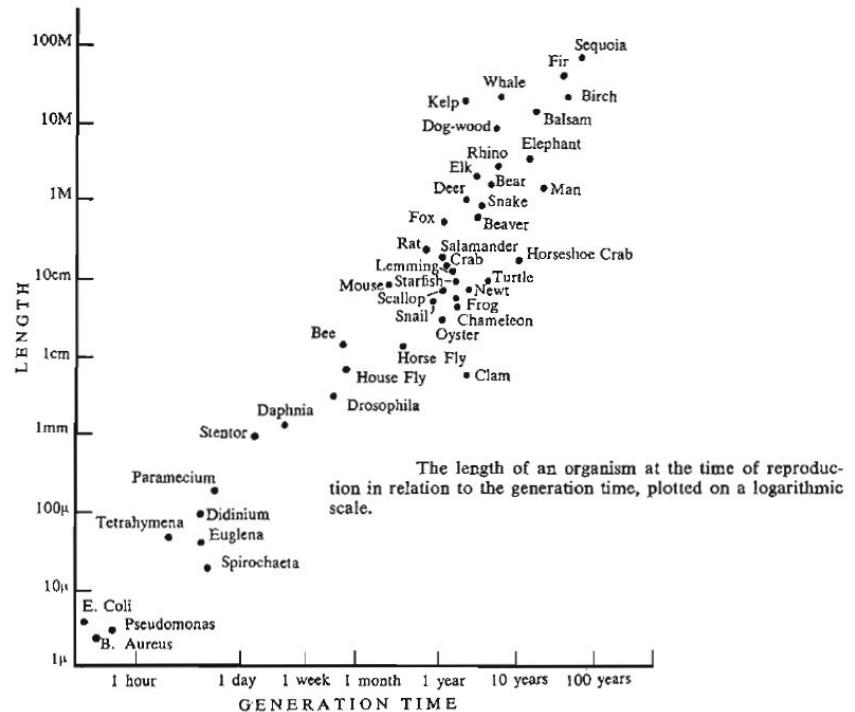
Relationship of Actual Rates of Registration to Predicted Rates
(104 cities 1960).



Данные-к-чернилам: высокая пропорция



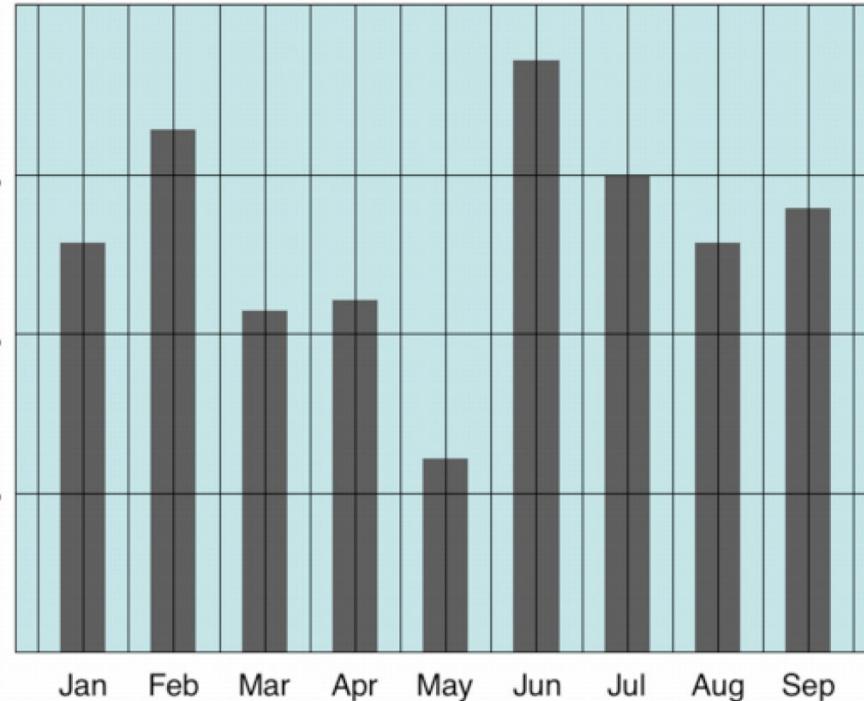
Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).



Редизайн графики

- Максимизируйте соотношение данных-к-чернилам в разумных пределах.
- Убирайте чернила-без-данных, в разумных пределах.
Немного орнамента, маркировка осей и т.д. вполне допустимы.
- Убирайте избыточные чернила-данные, в разумных пределах.
Некоторая избыточность полезна.
- Проверяйте и правьте.

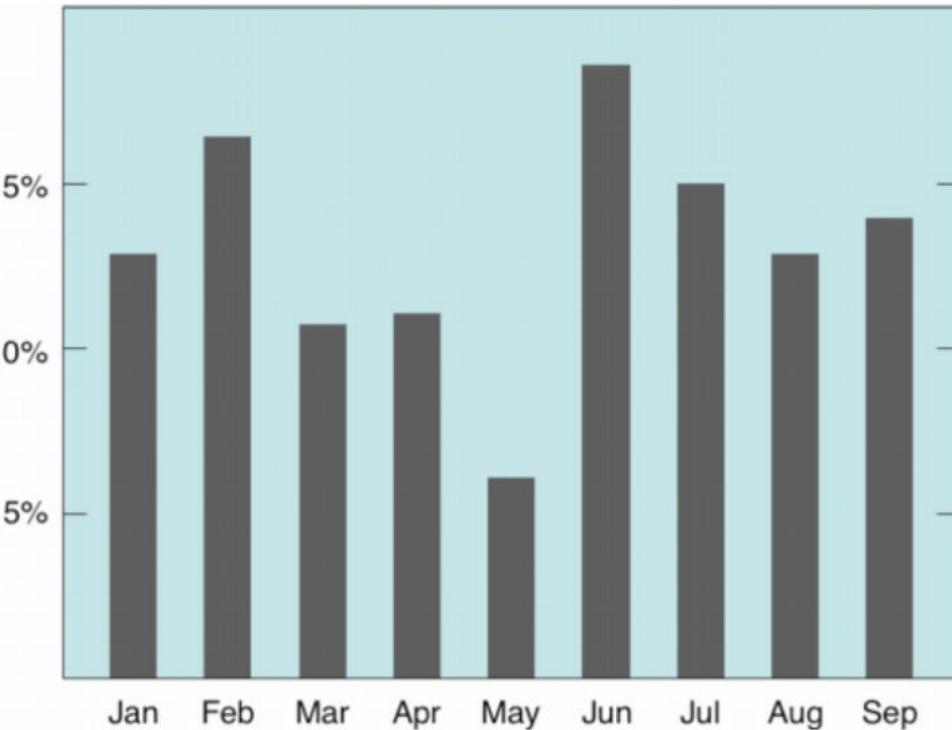
Избегайте мусора в графике



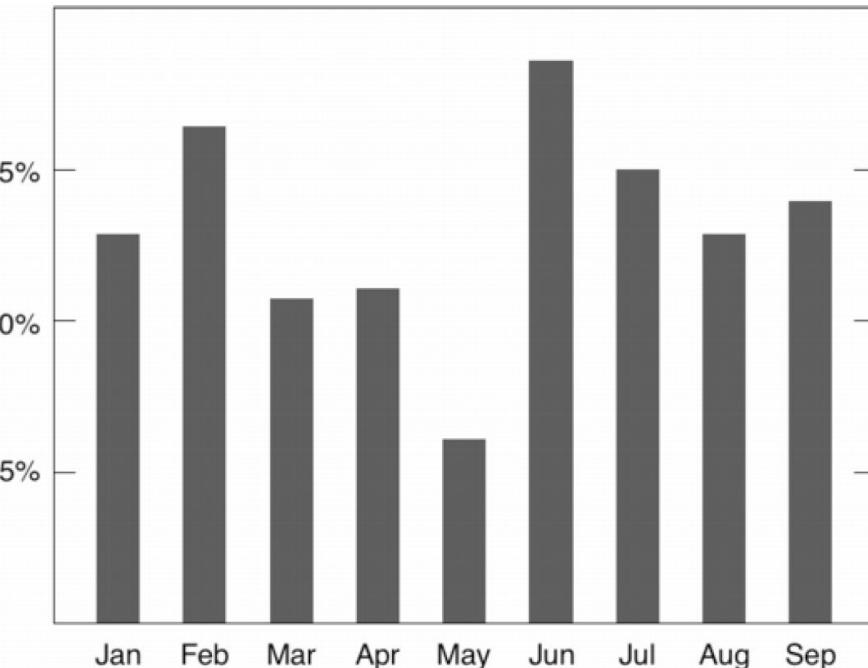
Посторонние визуальные элементы, которые отвлекают от послания

Упражнение: давайте упростим график, но сохраним информацию.

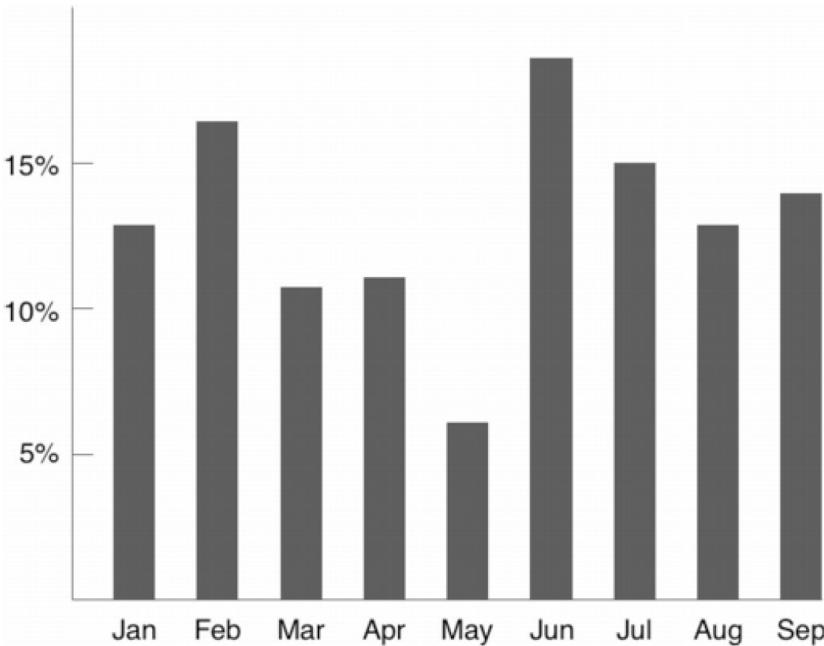
Избегайте мусора в графике



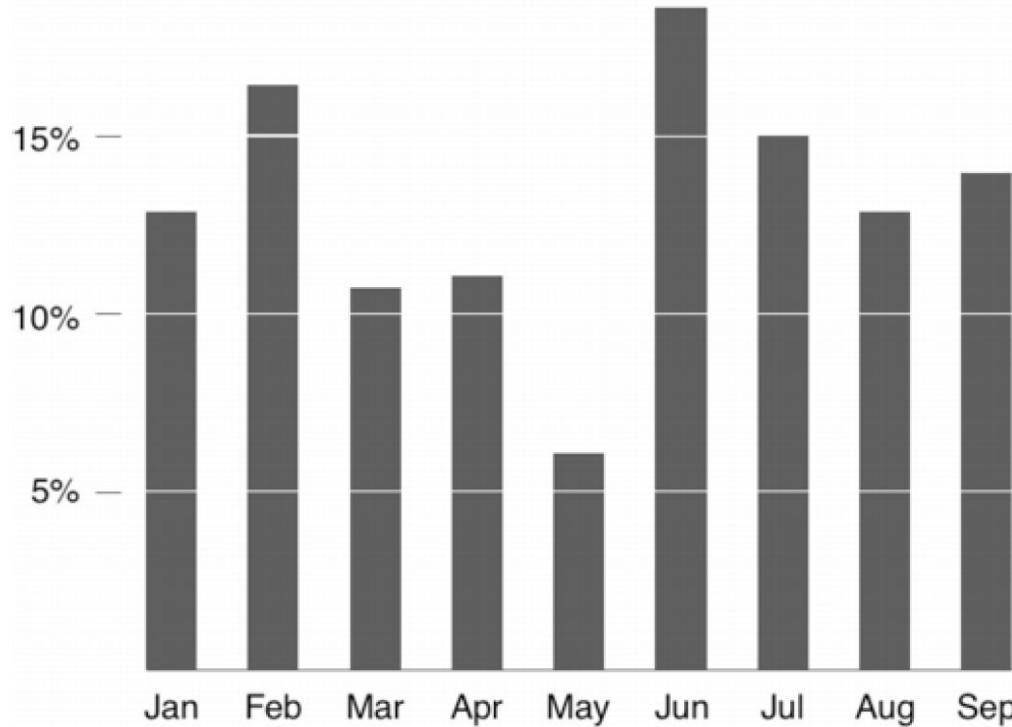
Избегайте мусора в графике



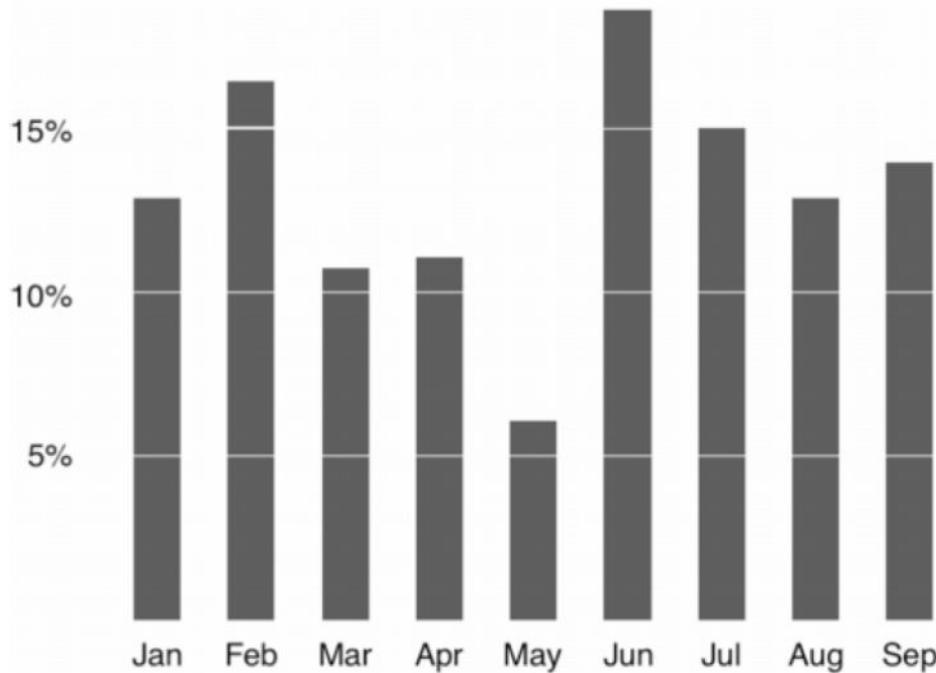
Избегайте мусора в графике



Избегайте мусора в графике



Избегайте мусора в графике



Удалили все лишние элементы/заливки без ущерба информативности.

Основные виды визуализаций



Основные виды визуализаций

- Гистограмма (Column Chart)
- Столбчатый график (Bar Graph)
- Линейный график (Line Graph)
- Графики с двумя шкалами (Dual Axis Chart)
- Диаграмма с областями (Area Chart)
- Составной столбчатый график (Stacked Bar Graph)
- Круговая диаграмма (Pie Chart)
- График рассеяния (Scatter Plot Chart)
- Пузырьковый график (Bubble Chart)
- Каскадный график (Waterfall Chart)
- Воронкообразный график (Funnel Chart)
- Тепловая карта (Heat Map)
- Древовидный график (Treemap Chart)

<https://datavizcatalogue.com/RU/>

Основные виды визуализаций



Дуговая
диаграмма



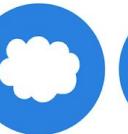
Диаграмма с
областями



Столбиковая
диаграмма



Диаграмма
размаха ('ящик с
усами')



Мозговой штурм



Пузырьковая
диаграмма



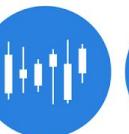
Пузырьковая
карта



Пулевая
диаграмма



Календарь



Свечной график



Хордовая
диаграмма



Фоновая
картограмма
(хороплет)



Укладка круга



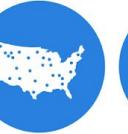
Карта
взаимосвязей



График плотности



Кольцевая
диаграмма

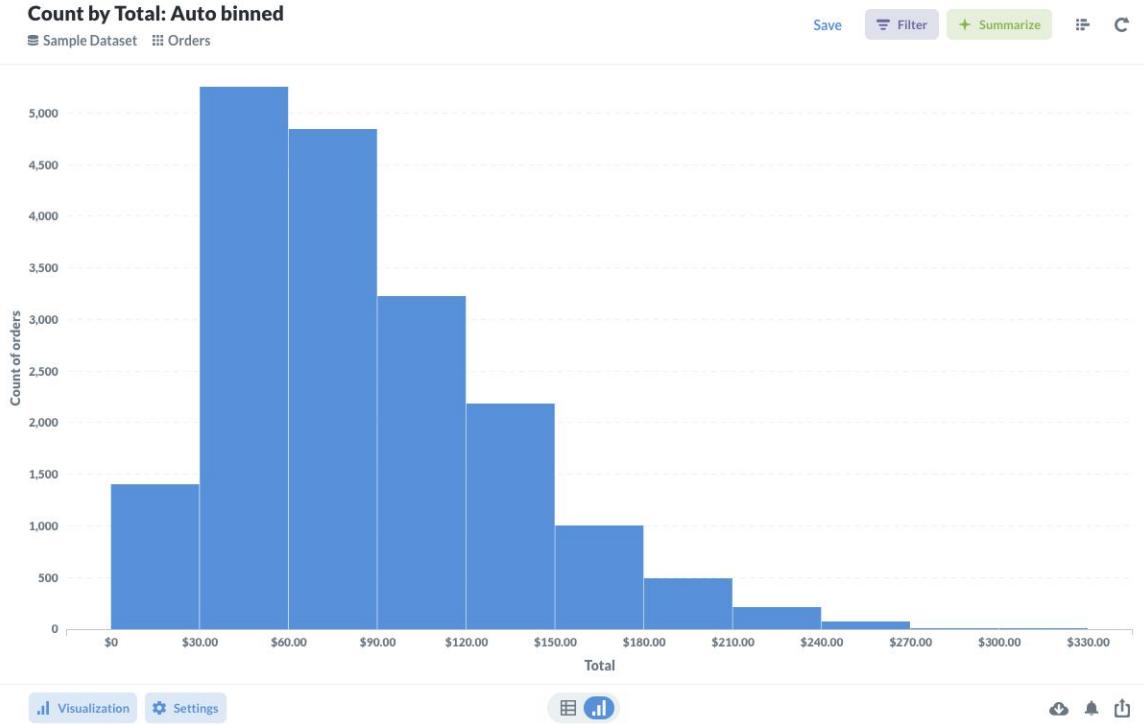


Точечная карта



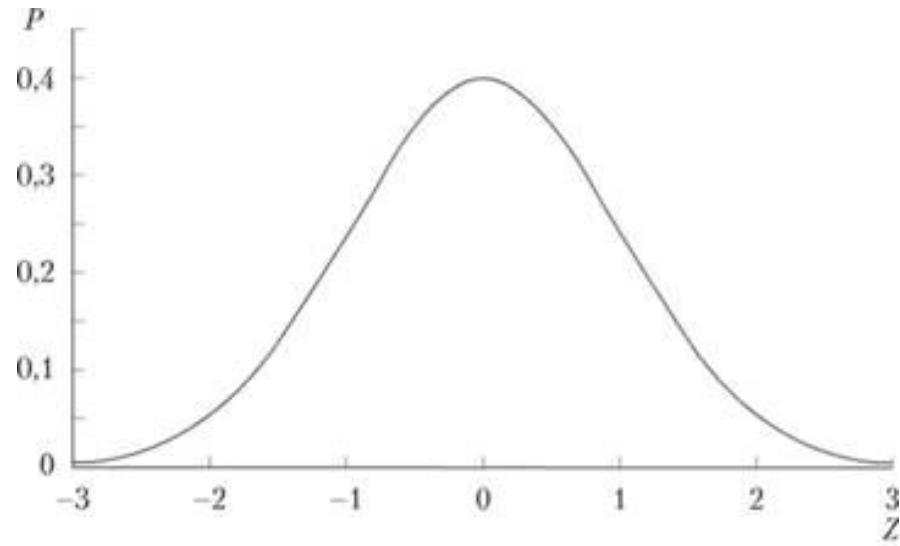
Точечная
матричная
диаграмма

Гистограмма

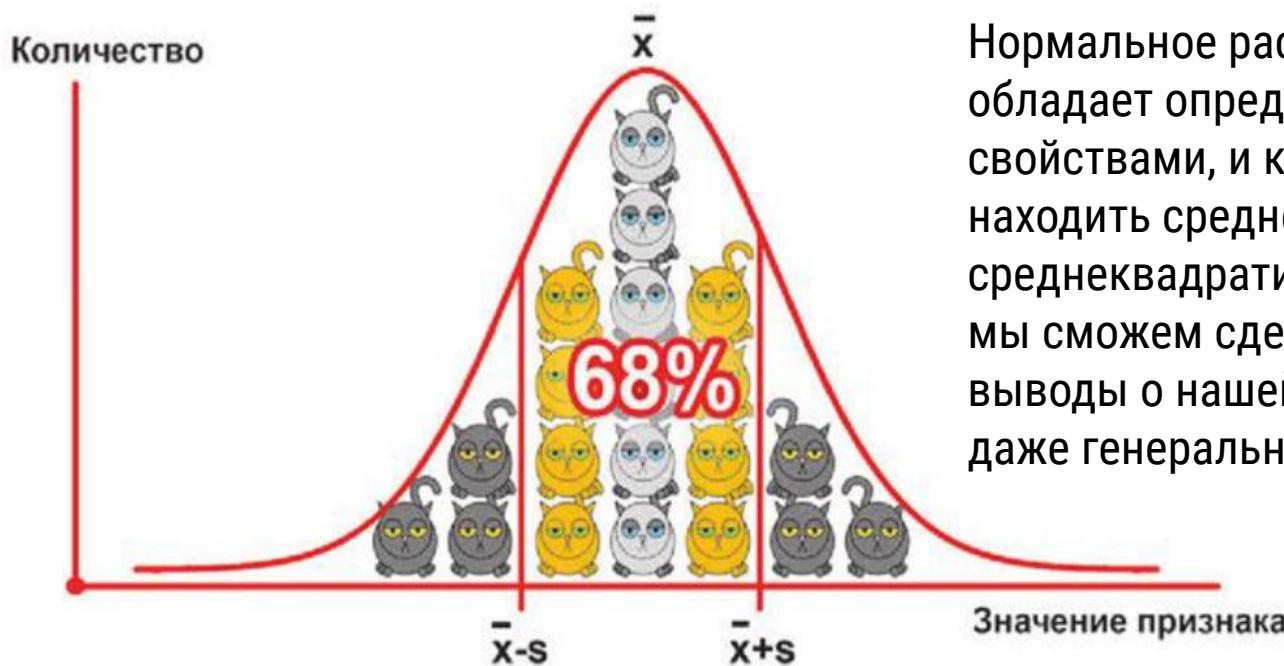


Распределение количественной переменной, искусственно разбитой на категории (bins)

Нормальное распределение

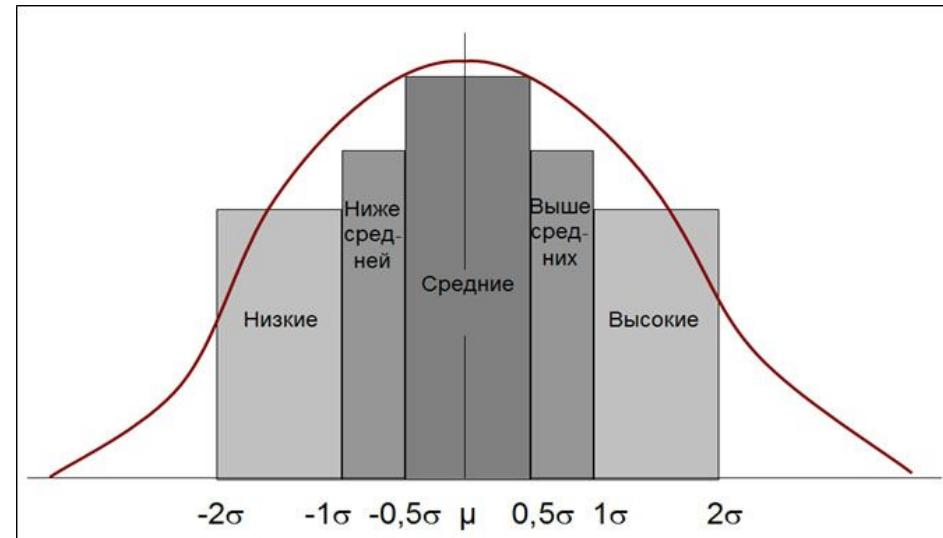
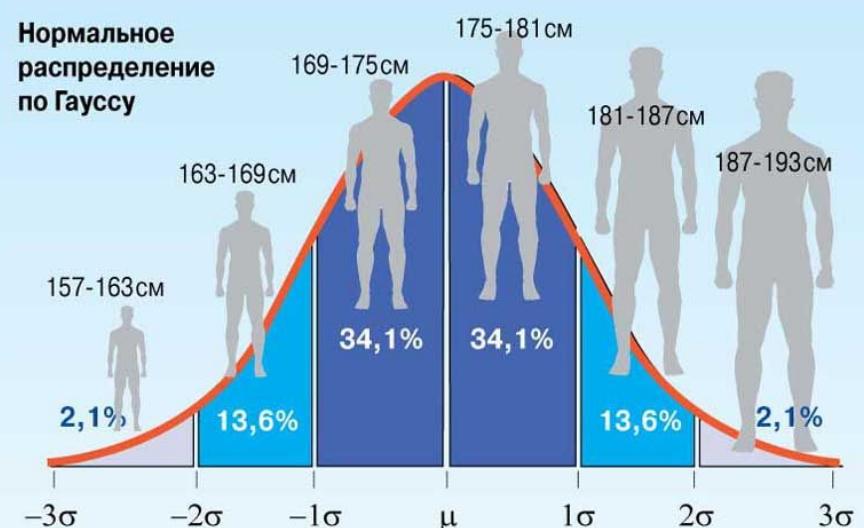


Нормальное распределение

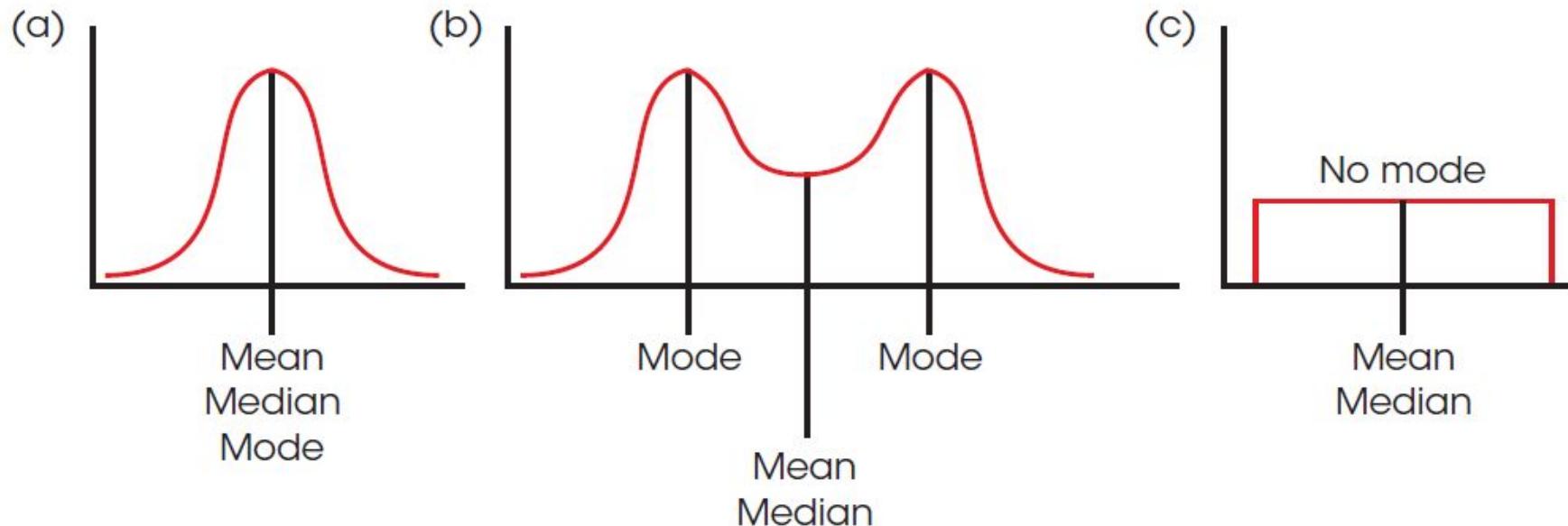


Нормальное распределение обладает определенными свойствами, и когда мы научимся находить среднее и среднеквадратичное отклонение, мы сможем сделать некоторые выводы о нашей выборке (или даже генеральной совокупности!).

Нормальное распределение



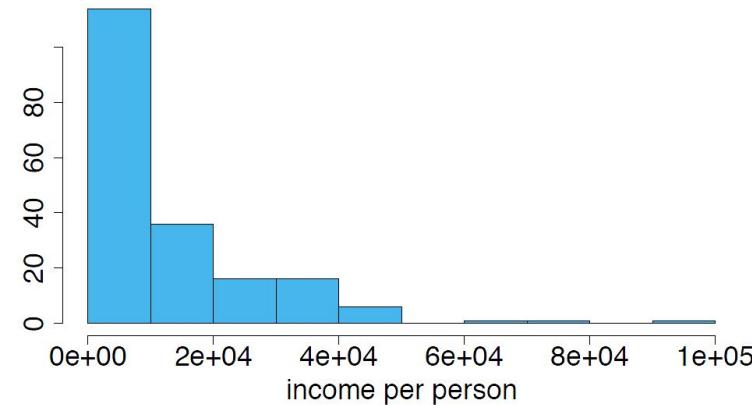
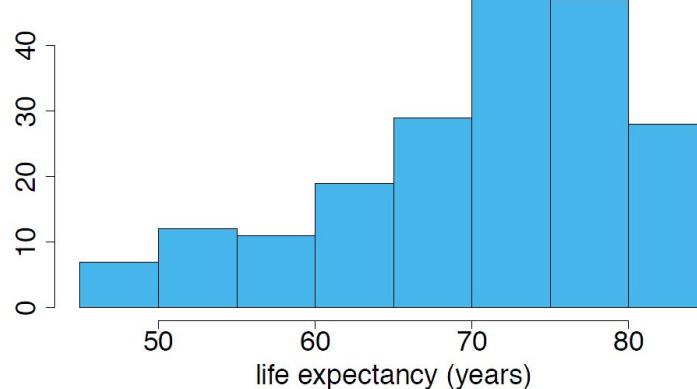
Центральная тенденция и форма распределения



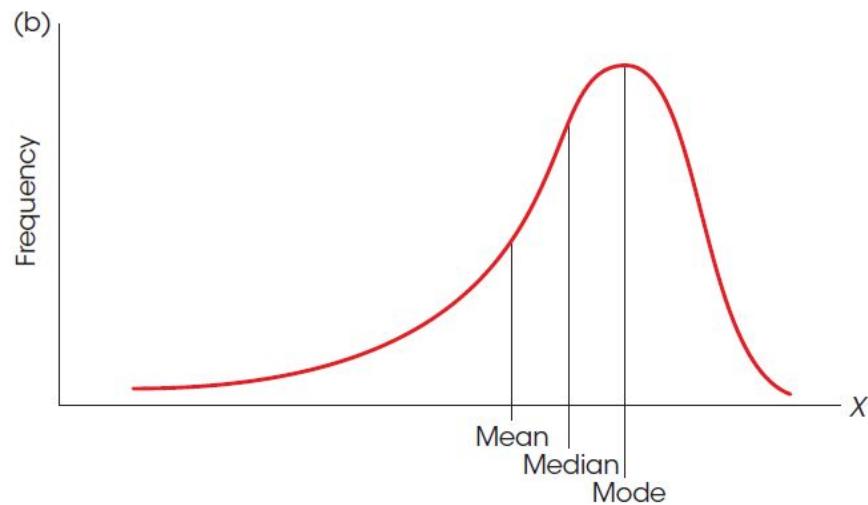
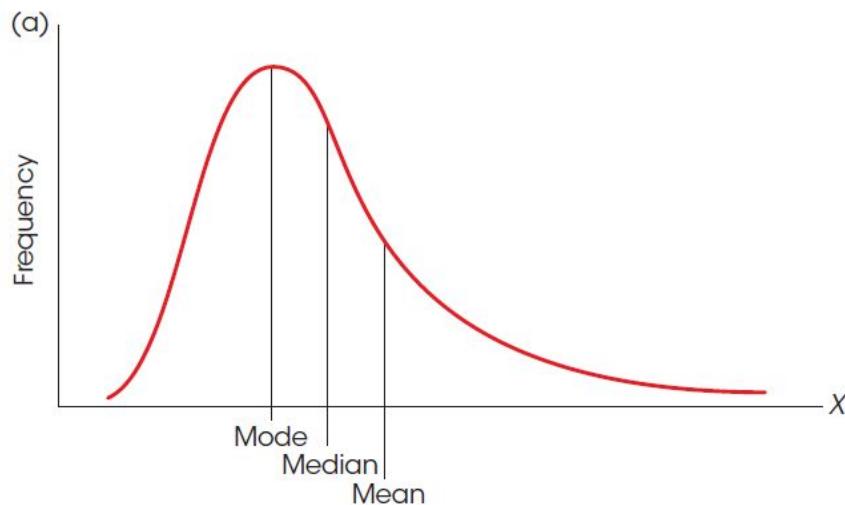
Скошенное распределение

Некоторые вещи в нашей жизни имеют скошенное распределение.

Например, уровень доходов – у нас очень много людей с низким и средним уровнем доходом и длинный-длинный хвост очень богатых людей.

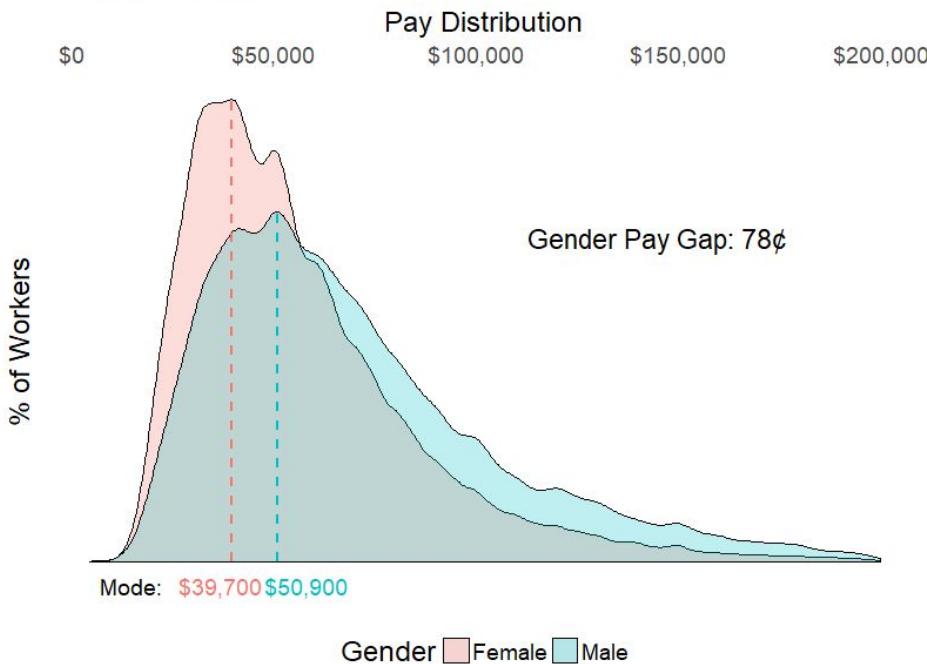


Центральная тенденция и форма распределения



Центральная тенденция и форма распределения

Pay by Gender - with Modes



<https://www.payscale.com/data/average-mean-median-mode>