

Визуализация данных на python

Мария Мансурова,
аналитик

Цель занятия

- › рассмотрим основные типы визуализаций и научимся выделять подходящую
- › рассмотрим основные инструменты python для создания графиков



План

1. Кто такие аналитики и чем они занимаются?
2. Что такое визуализация и зачем она нужна?
3. Теория визуализации: visual encodings, типы графиков и задачи визуализации
4. Инструменты



Аналитика и аналитики





Кто такие аналитики?

Аналитики в Яндексе

- › Помогают менеджерам принимать решения
- › Умеют писать adhoc код
- › Чаще всего используют python, R, SQL, но могут и другое

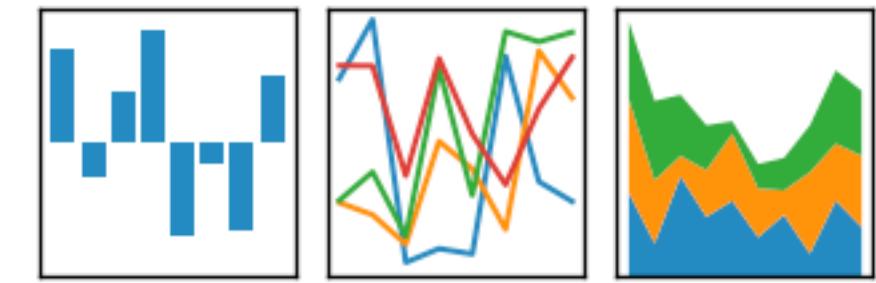


Аналитики в Яндексе



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



python



Jenkins



ClickHouse



Примеры задач

- › Сколько серверов нам нужно заказать на следующие 2 года, чтобы сервис работал?
- › Не сломали ли мы что-то новой версией нашей мобильной SDK?
- › Как наши пользователи используют date picker в Метрике?
- › Как нам узнавать о потерях данных как можно раньше?



Визуализация

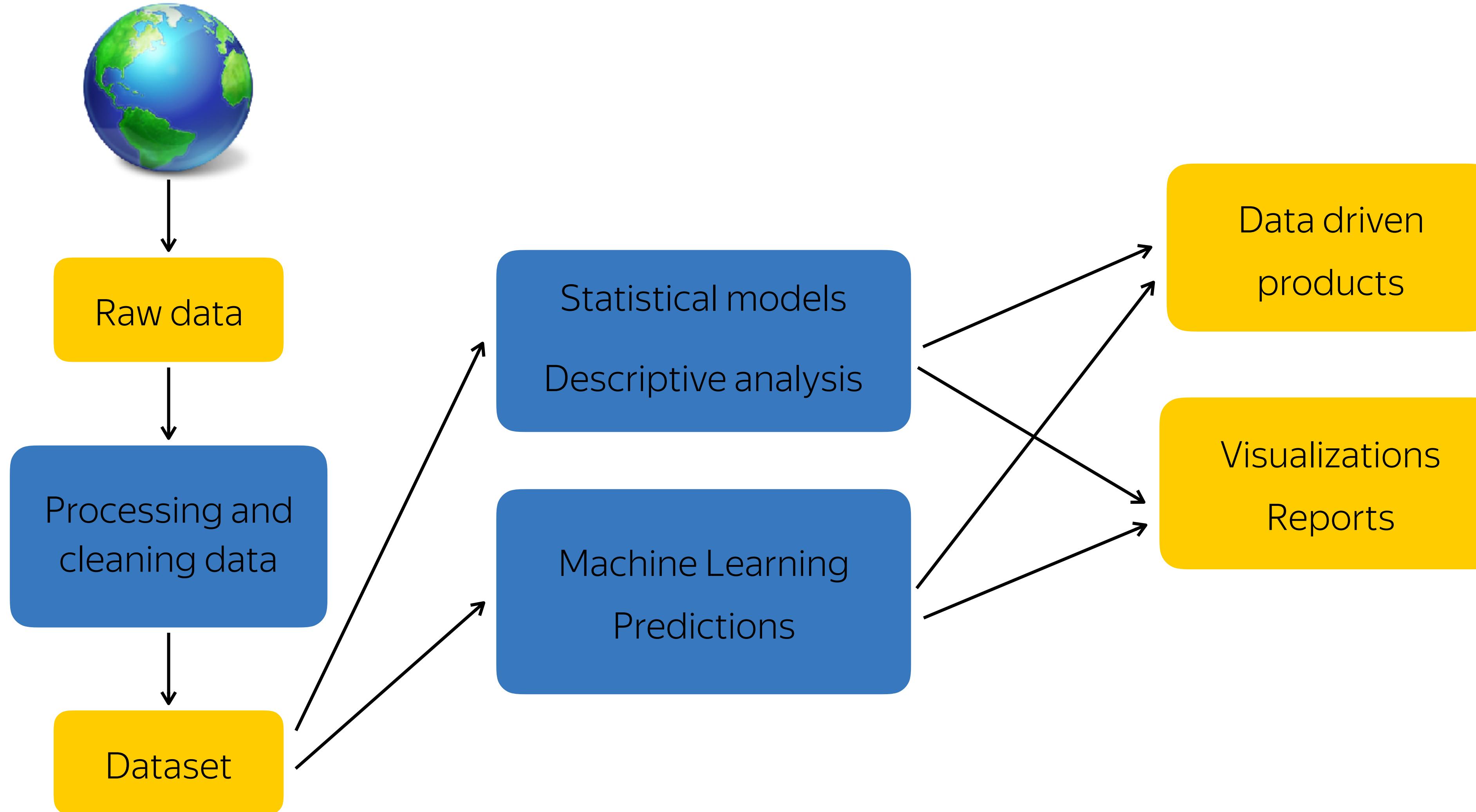


Что такое визуализация данных

Визуализация данных — это представление данных в виде, который обеспечивает наиболее эффективную работу человека по их изучению.



Работа с данными



Роль визуализации

- › **exploratory** - «разговор наедине с данными»
- › **explanatory** - раскрыть и донести свою мысль





А нужна ли
визуализация вообще?

Пример выборок

все статистики 4х выборок
одинаковы

- › mean $x = 9$
- › sample variance of $x = 11$
- › mean $y = 11.5$
- › sample variance of $y = 4.125$
- › correlation between x and $y = 0.816$

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

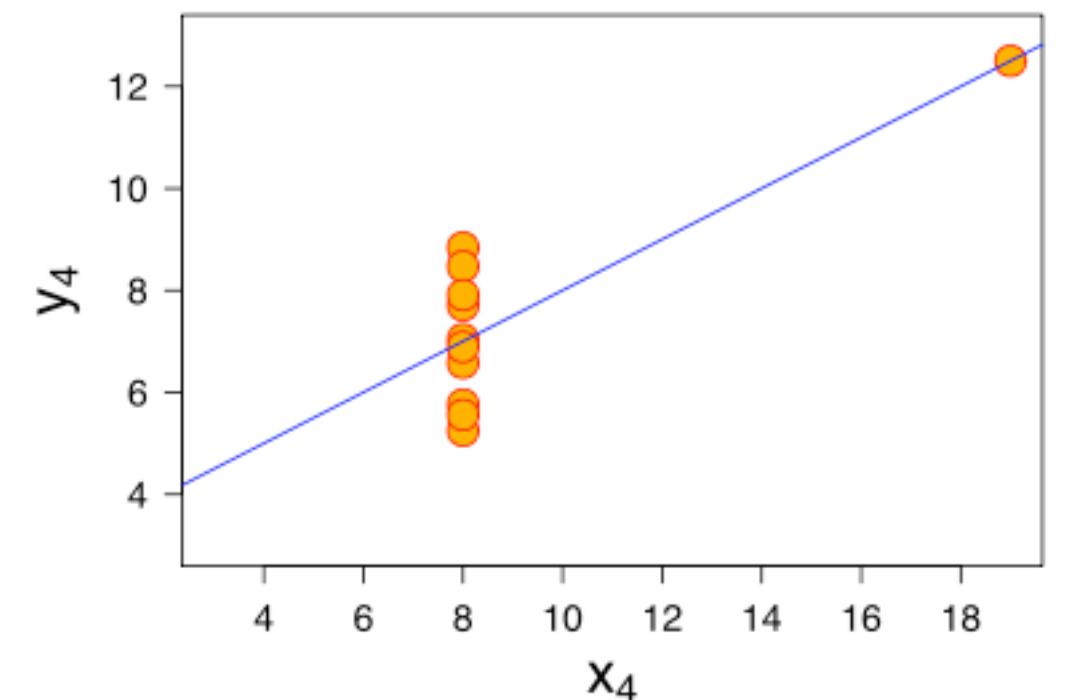
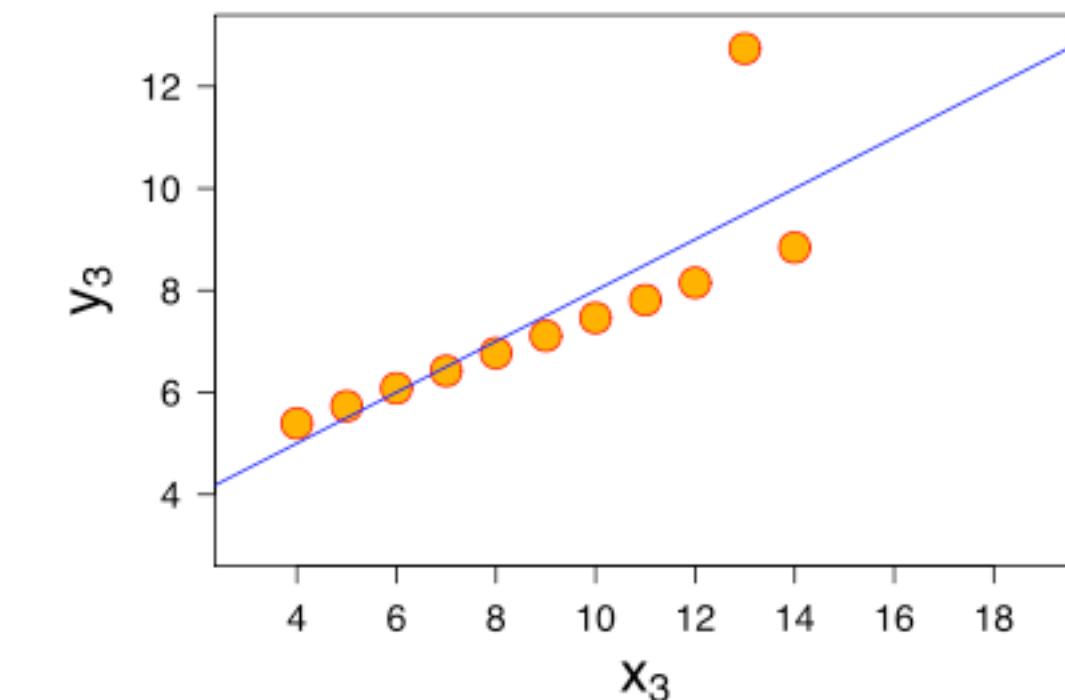
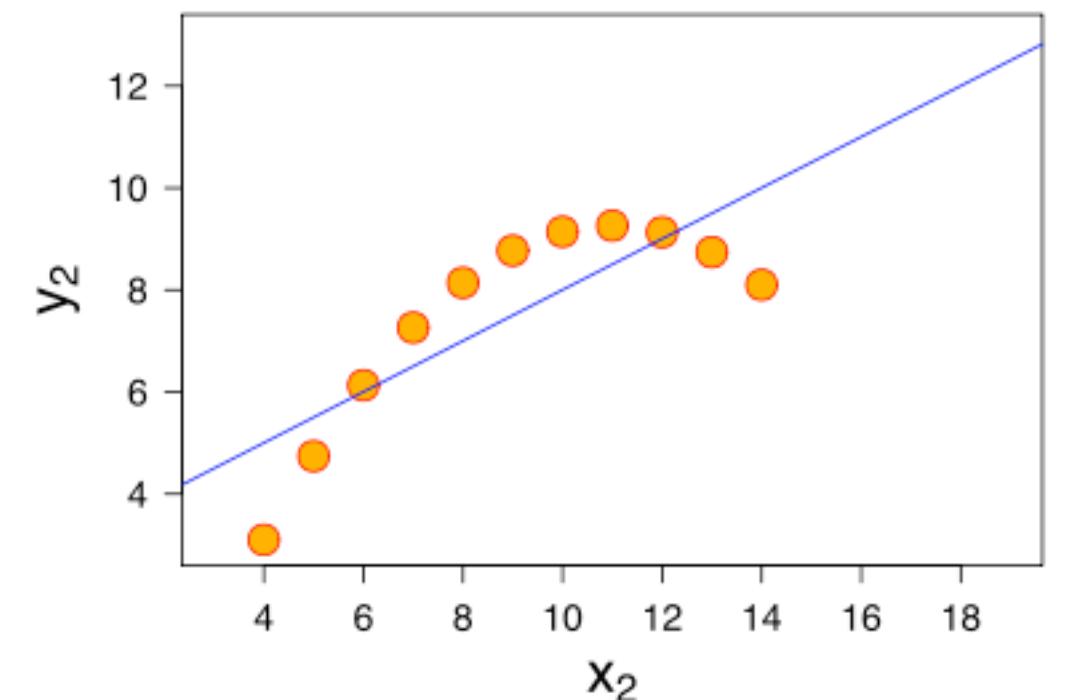
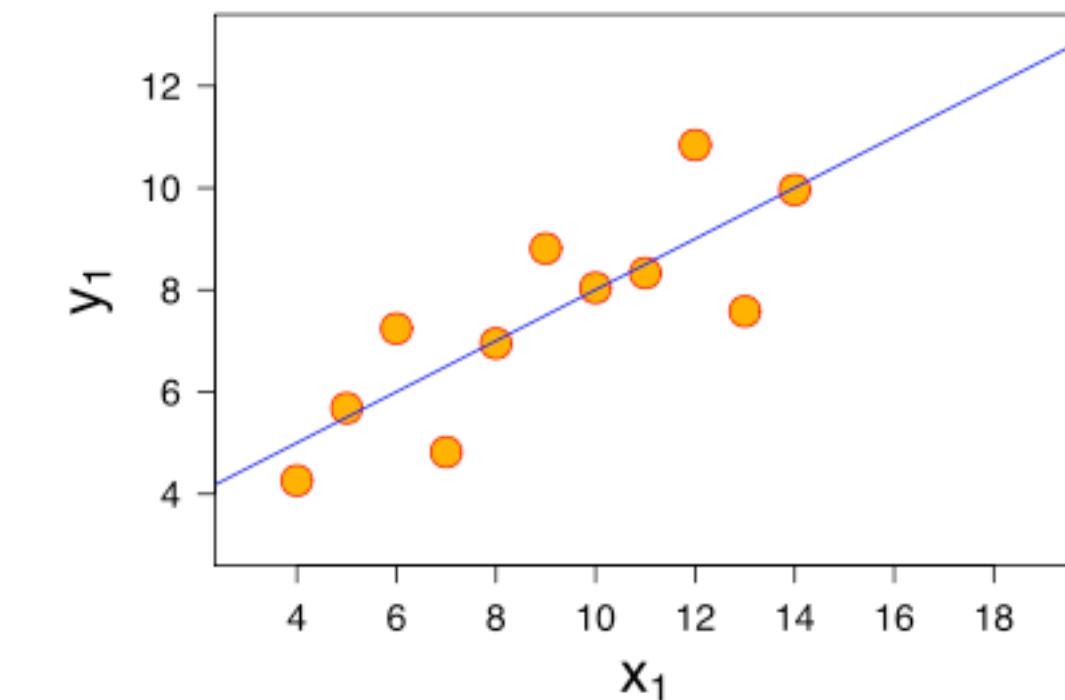


Выборки одинаковые?

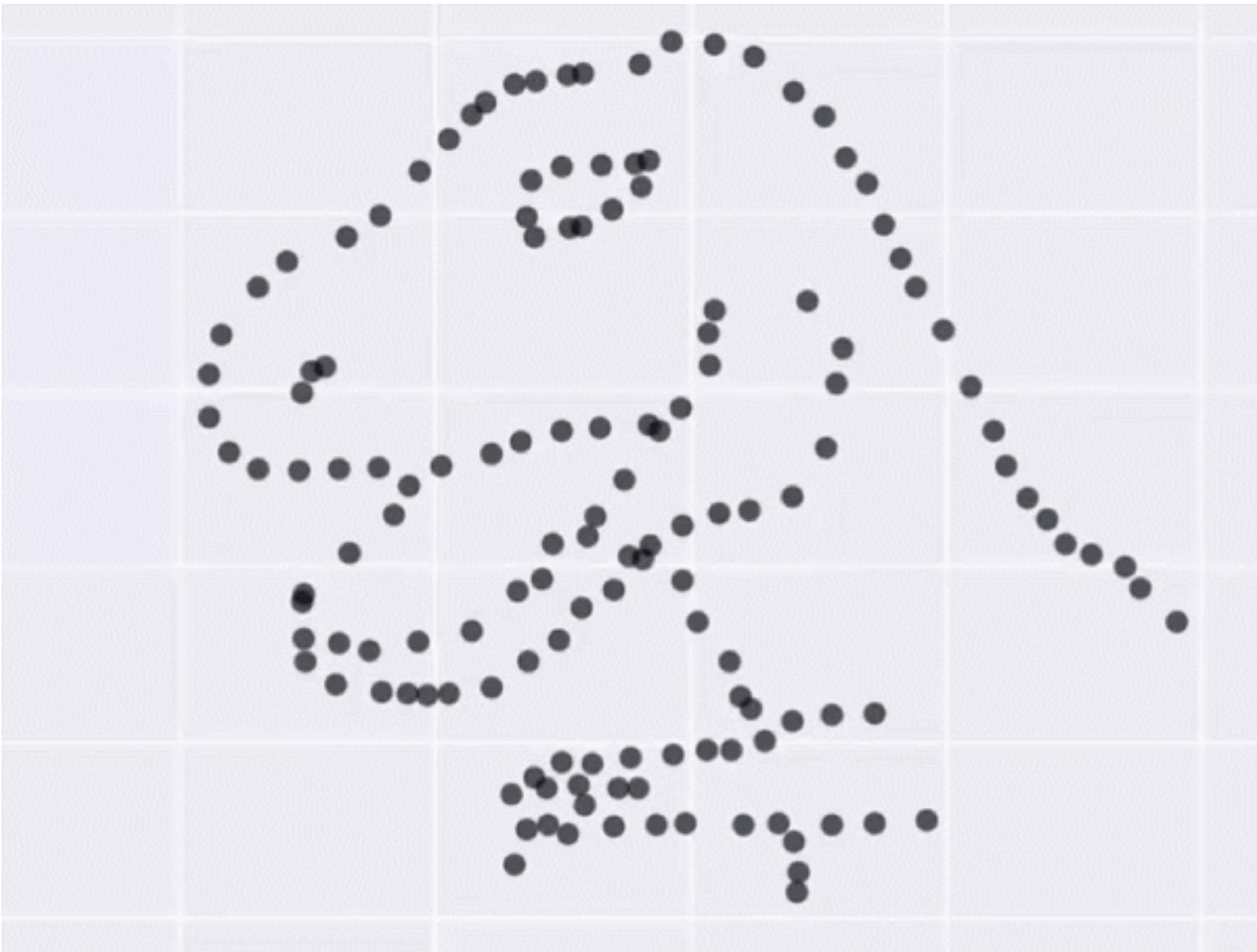
Квартет Энскомба

пример был придуман
статистиком Фрэнсисом
Энскомбом в 1973 году

- › важность визуализаций для анализа данных
- › влияние выбросов (outliers) на статистические показатели



И другие вариации...



Немного теории



Данные

числовые

- › дискретные/непрерывные

категориальные

- › nominal/ordered



Формы выражения (Visual Encodings)

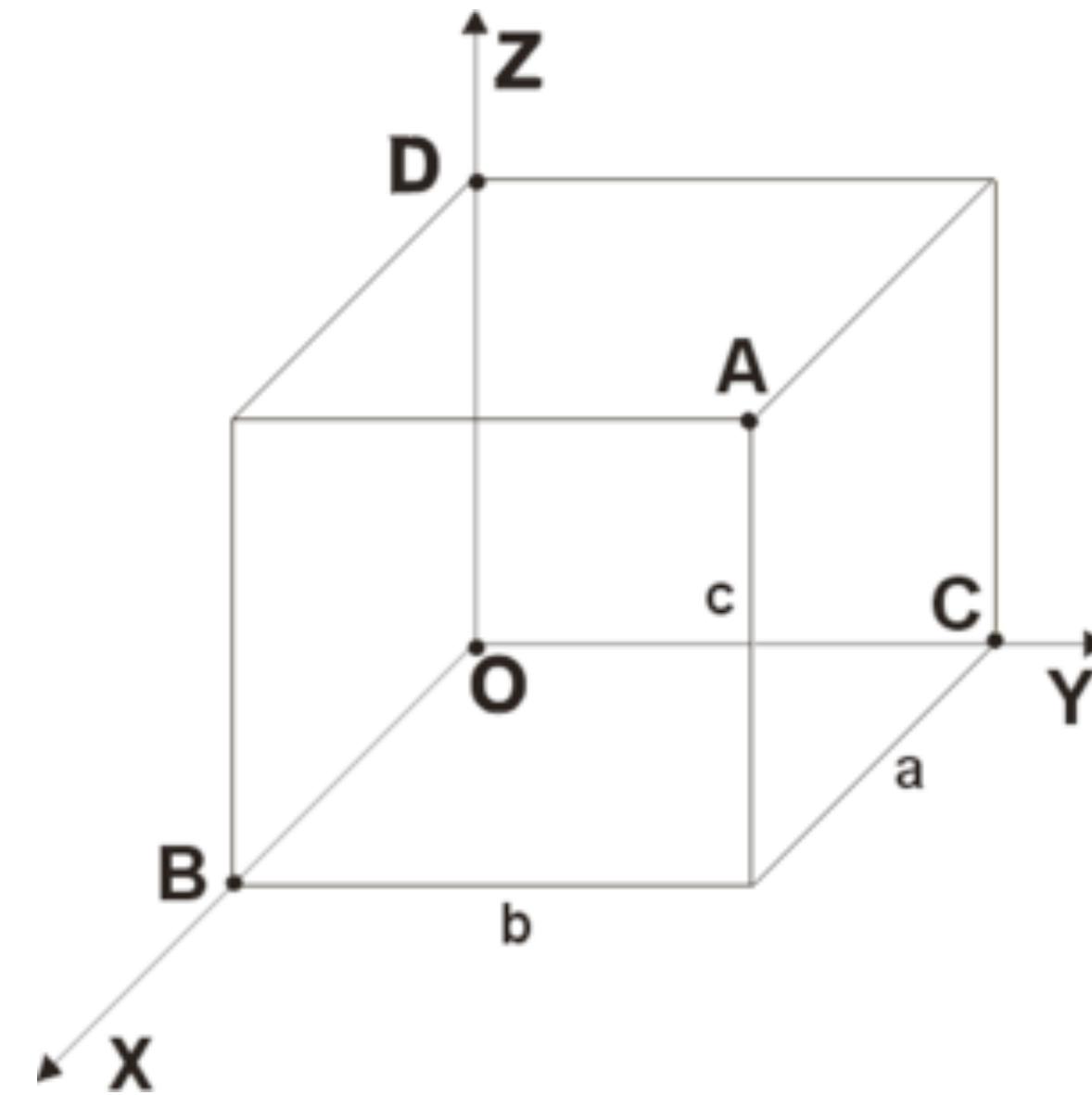
данные -> отображение их на графике

- › позиция
- › размер
- › цвет, оттенок цвета
- › ориентация, наклон
- › форма, текстура
- › движение, анимация



Позиция

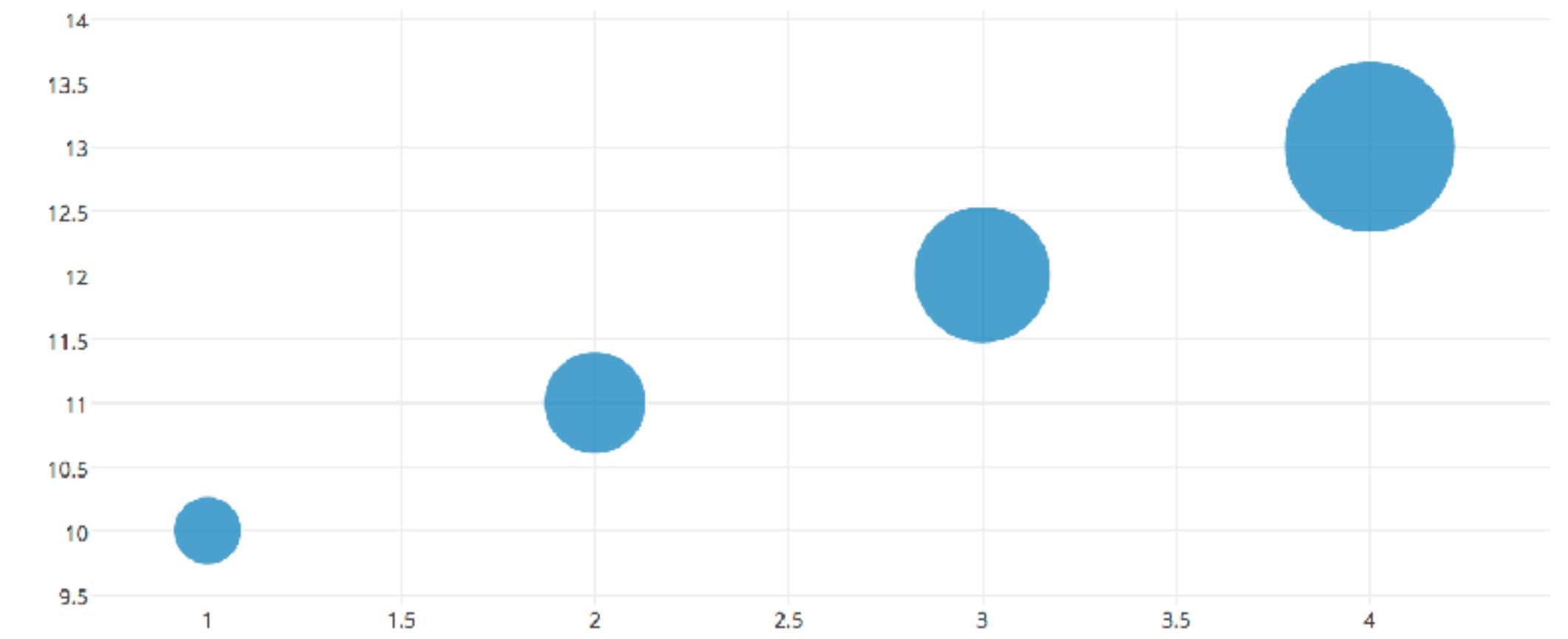
- › легко интерпретируется человеком
- › позволяет отследить корреляции
- › только 2D, максимум 3D с потерей точности



Размер (длина, площадь, объем)

длина

- › хорошо считывается людьми, но позволяет отобразить не более 2D

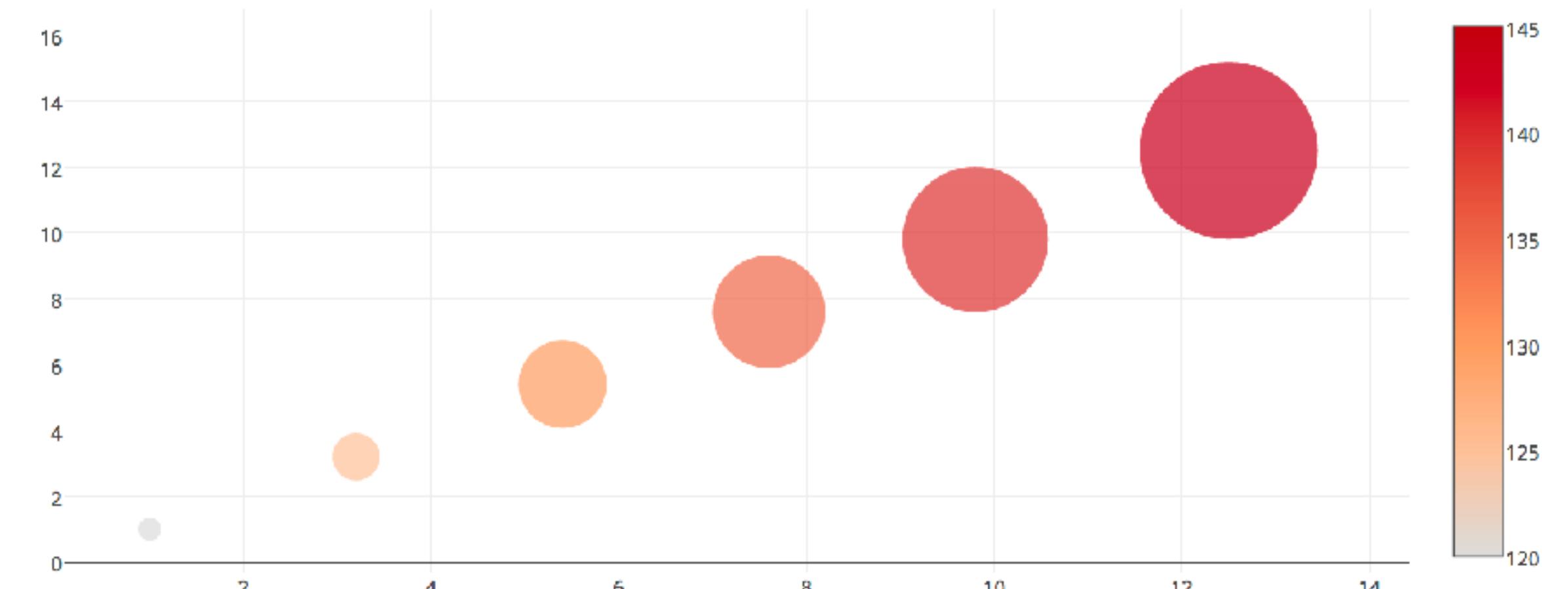
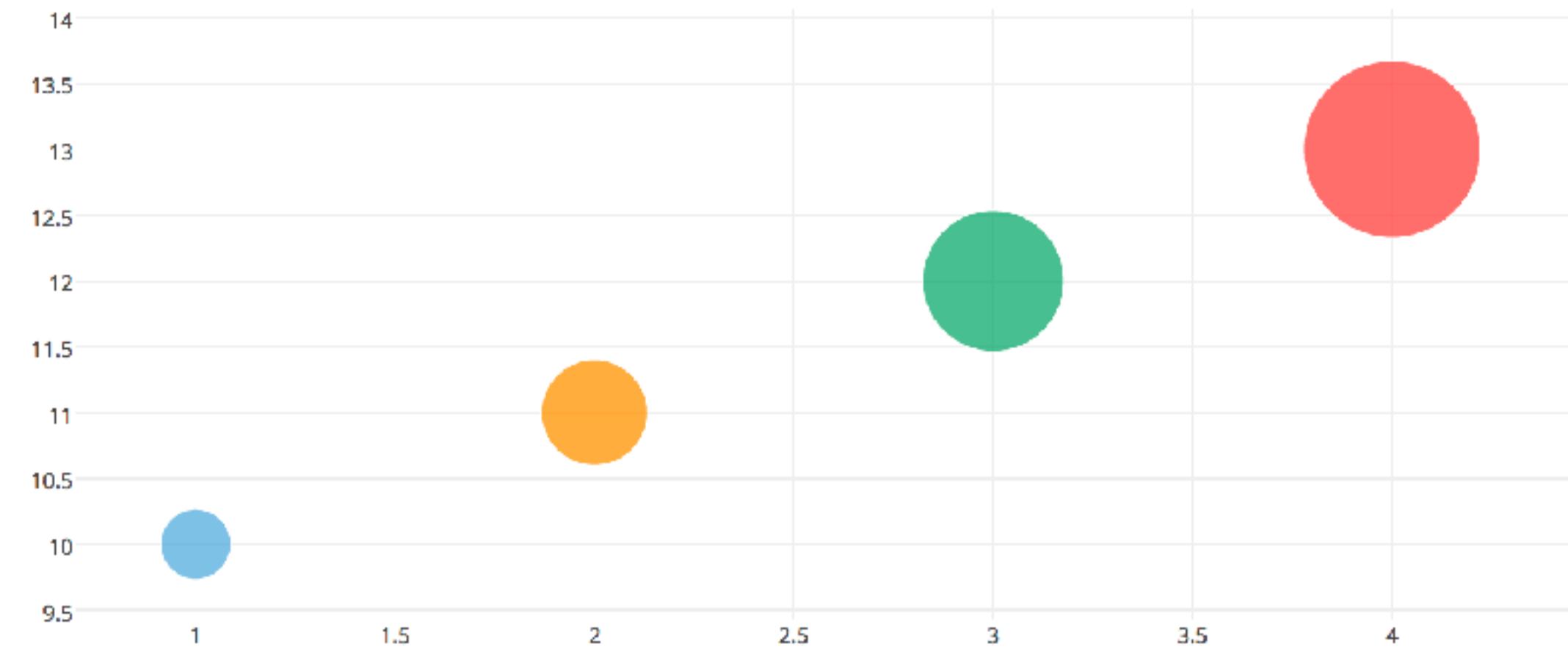


площадь, объем

- › лучше всего подходит для ordered data
- › сложно понять точные отличия в переменных

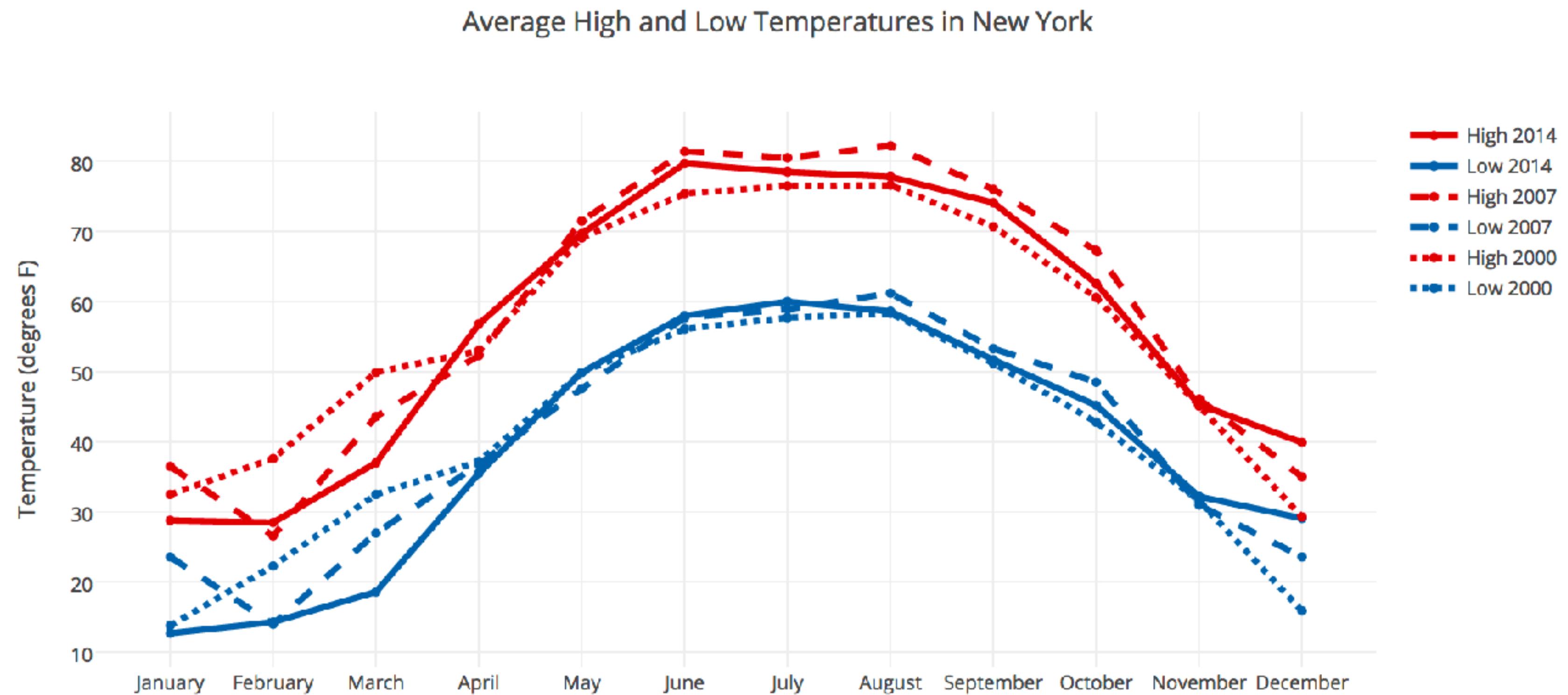
Цвет (hue/saturation)

- › hue подходит для категориальных признаков
- › saturation - для ordered data



И другие

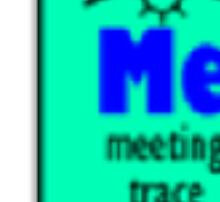
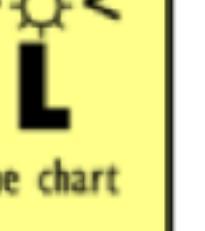
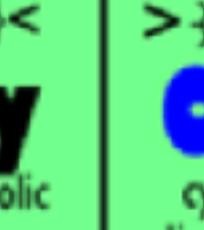
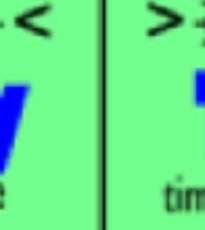
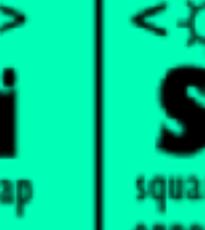
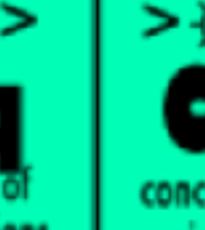
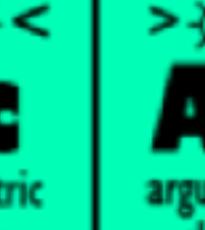
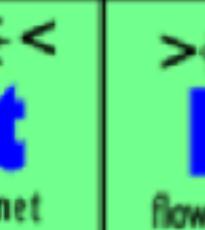
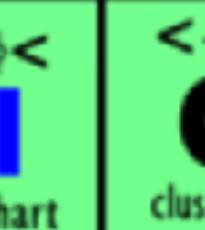
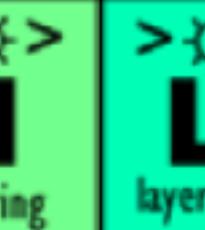
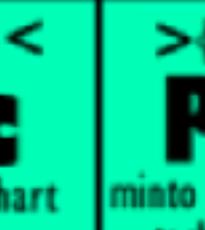
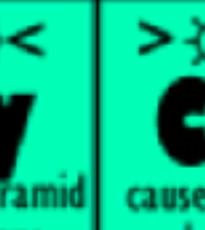
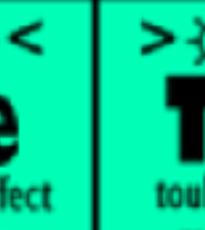
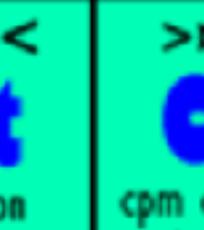
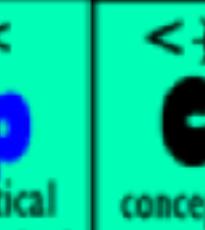
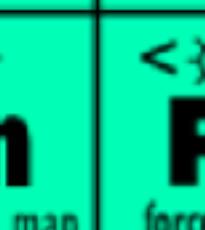
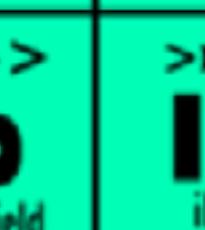
- › тип линии
- › текстура
- › форма markers





Какие типы графиков вы
знаете?

A PERIODIC TABLE OF VISUALIZATION METHODS

 continuum	 Data Visualization Visual representations of quantitative data in schematic form (either with or without axes)												 Strategy Visualization The systematic use of complementary visual representations in the analysis, development, formulation, communication, and implementation of strategies in organizations.				 graphic facilitation
 Tb table	 Ca cartesian coordinates	 Information Visualization The use of interactive visual representations of data to amplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by users as they proceed working with it				 Metaphor Visualization Visual Metaphors position information graphically to organize and structure information. They also convey an insight about the represented information through the key characteristics of the metaphor that is employed	 Compound Visualization The complementary use of different graphic representation formats in one single schema or frame				 Me meeting trace	 Mm metro map	 Tm temple	 St story template	 Tr tree	 Ct cartoon	
 Pi pie chart	 L line chart	 Concept Visualization Methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses.				 Co communication diagram	 Fp flight plan	 Cs concept skeleton	 Br bridge	 Fu funnel	 Ri rich picture						
 B bar chart	 Ac area chart	 R radar chart cobweb	 Pa parallel coordinates	 Hy hyperbolic tree	 Cy cycle diagram	 T timeline	 Ve venn. diagram	 Mi mindmap	 Sq square of oppositions	 Cc concentric circles	 Ar argument slide	 Sw swim lane diagram	 Gc gantt chart	 Pm perspectives diagram	 D dilemma diagram	 Pr parameter ruler	 Kn knowledge map
 Hi histogram	 Sc scatterplot	 Sa sankey diagram	 In information lense	 E entity relationship diagram	 Pt petri net	 Fl flow chart	 Cl clustering	 Lc layer chart	 Py minto pyramid technique	 Ce cause-effect chains	 Tl toulmin map	 Dt decision tree	 Cp cpm critical path method	 Cf concept fan	 Co concept map	 Ic iceberg	 Lm learning map
 Tk tukey box plot	 Sp spectrogram	 Da data map	 Tp treemap	 Cn cone tree	 Sy system dyn./ simulation	 Df data flow diagram	 Se semantic network	 So soft system modeling	 Sn synergy map	 Fo force field diagram	 Ib ibis argumentation map	 Pr process event chains	 Pe pert chart	 Ev evocative knowledge map	 V vee diagram	 Hh heaven 'n' hell chart	 I infomural

Graphical Perception, 1984

- › Позиция на графике (scatter plot)
- › Несколько одинаковых графиков рядом (несколько scatter plots)
- › Длина (bar chart)
- › Угол и наклон (pie chart)
- › Площадь (bubbles)
- › Объем, плотность, насыщенность цвета (heatmap)
- › Цвет



Выбираем график

- › Простое сравнение (Nominal comparison)
- › Динамика во времени (Time series)
- › Ранжирование (Ranking)
- › Часть от целого (Part-to-hole)
- › Отклонение (Deviation)
- › Частотное распределение
(Frequency distribution)
- › Кореляция (Correlation)



Данные о продажах и оценках игр

	Name	Platform	Year_of_Release	Genre	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Rating
0	Wii Sports	Wii	2006	Sports	82.53	76.0	51	8.0	322	E
2	Mario Kart Wii	Wii	2008	Racing	35.52	82.0	73	8.3	709	E
3	Wii Sports Resort	Wii	2009	Sports	32.77	80.0	73	8.0	192	E
6	New Super Mario Bros.	DS	2006	Platform	29.80	89.0	65	8.5	431	E
7	Wii Play	Wii	2006	Misc	28.92	58.0	41	6.6	129	E
8	New Super Mario Bros. Wii	Wii	2009	Platform	28.32	87.0	80	8.4	594	E
11	Mario Kart DS	DS	2005	Racing	23.21	91.0	64	8.6	464	E
13	Wii Fit	Wii	2007	Sports	22.70	80.0	63	7.7	146	E
14	Kinect Adventures!	X360	2010	Misc	21.81	61.0	45	6.3	106	E
15	Wii Fit Plus	Wii	2009	Sports	21.79	80.0	33	7.4	52	E

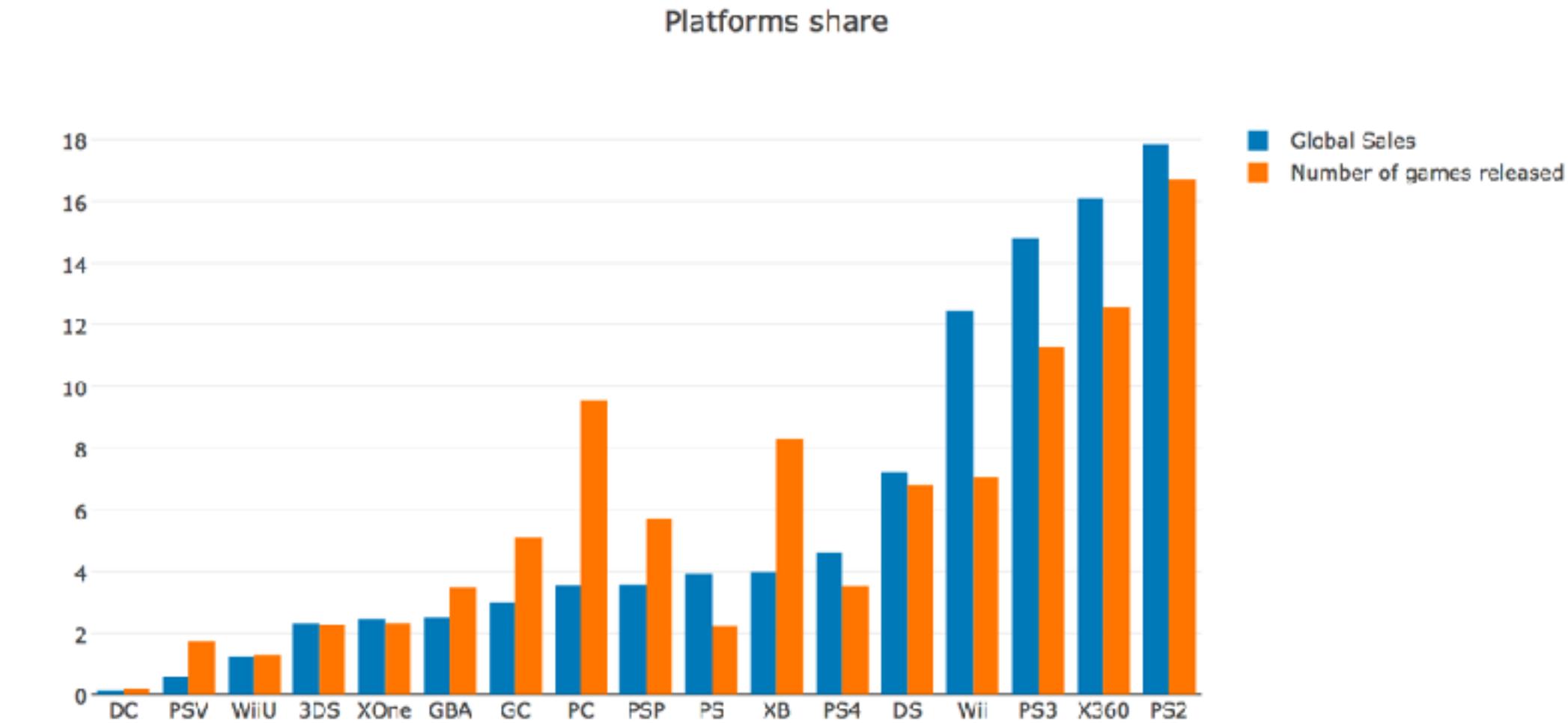
Обычное сравнение (Nominal comparison)

- | **Nominal comparison** - простое сравнение одной или нескольких метрик по категориям без определенного порядка
- | **Задача** - сравнить игровые платформы по числу выпущенных и проданных игр

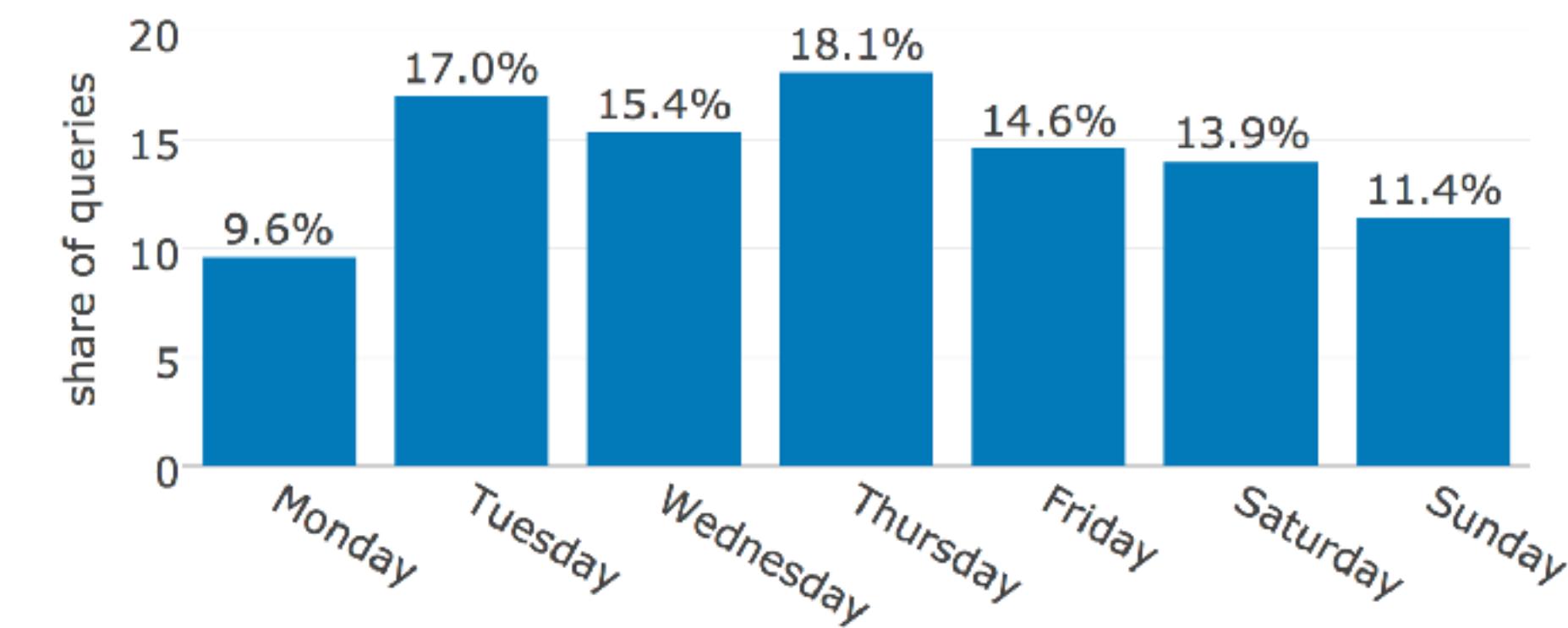
Обычное сравнение (Nominal comparison)

| Nominal comparison - простое сравнение одной или нескольких метрик по категориям без определенного порядка

- › Горизонтальный или вертикальный bar chart



Запросы по дням недели



Time Series

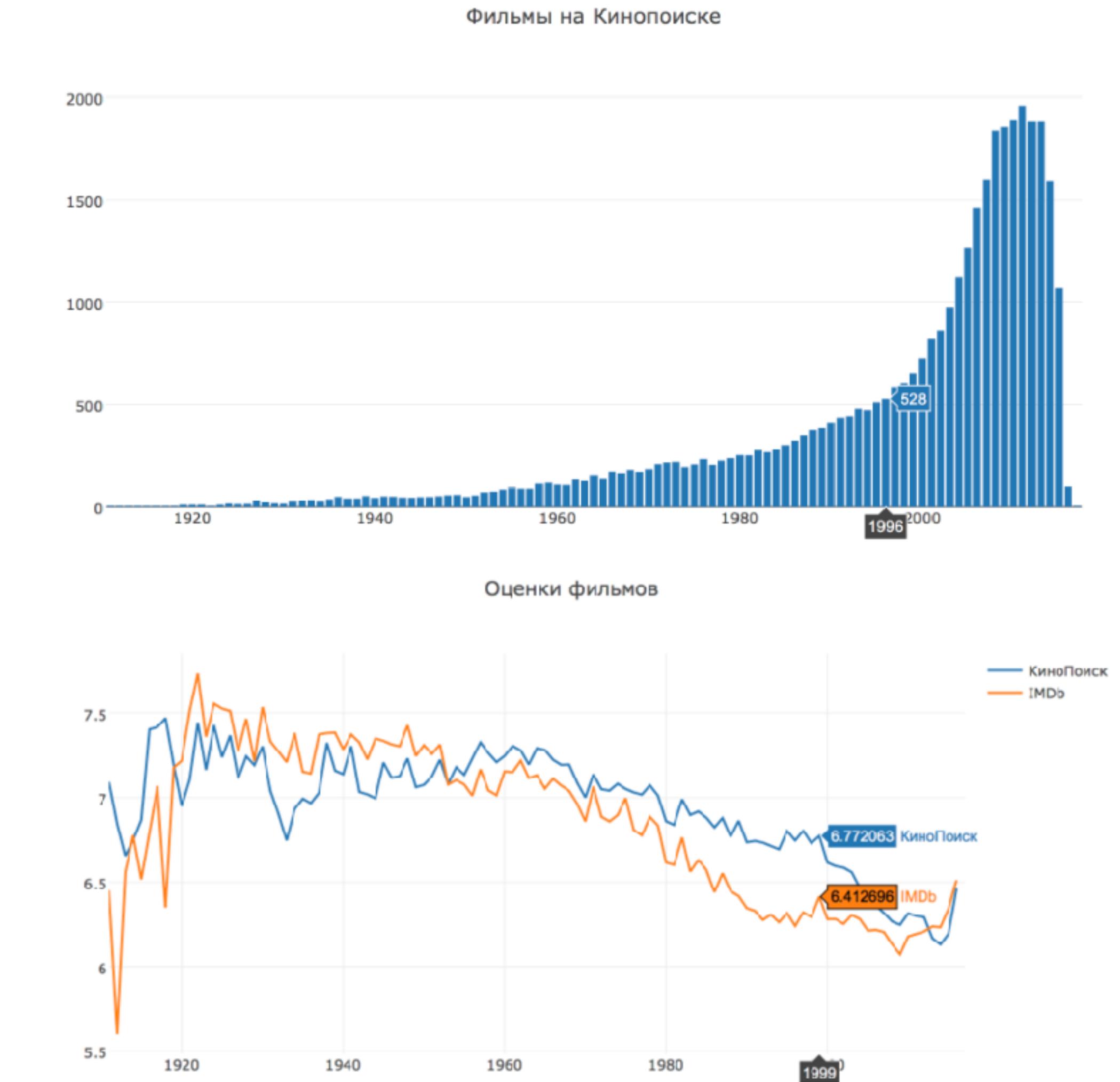
| Time Series - изменение одной или нескольких метрик во времени

| Задача - отобразить динамику числа проданных компьютерных игр в мире

Time Series

| Time Series - изменение одной или нескольких метрик во времени

- › Line chart, чтобы подчеркнуть тренд
- › Bar chart, чтобы выделить отдельные значение
- › Временная переменная должна располагаться на оси X



Ranking

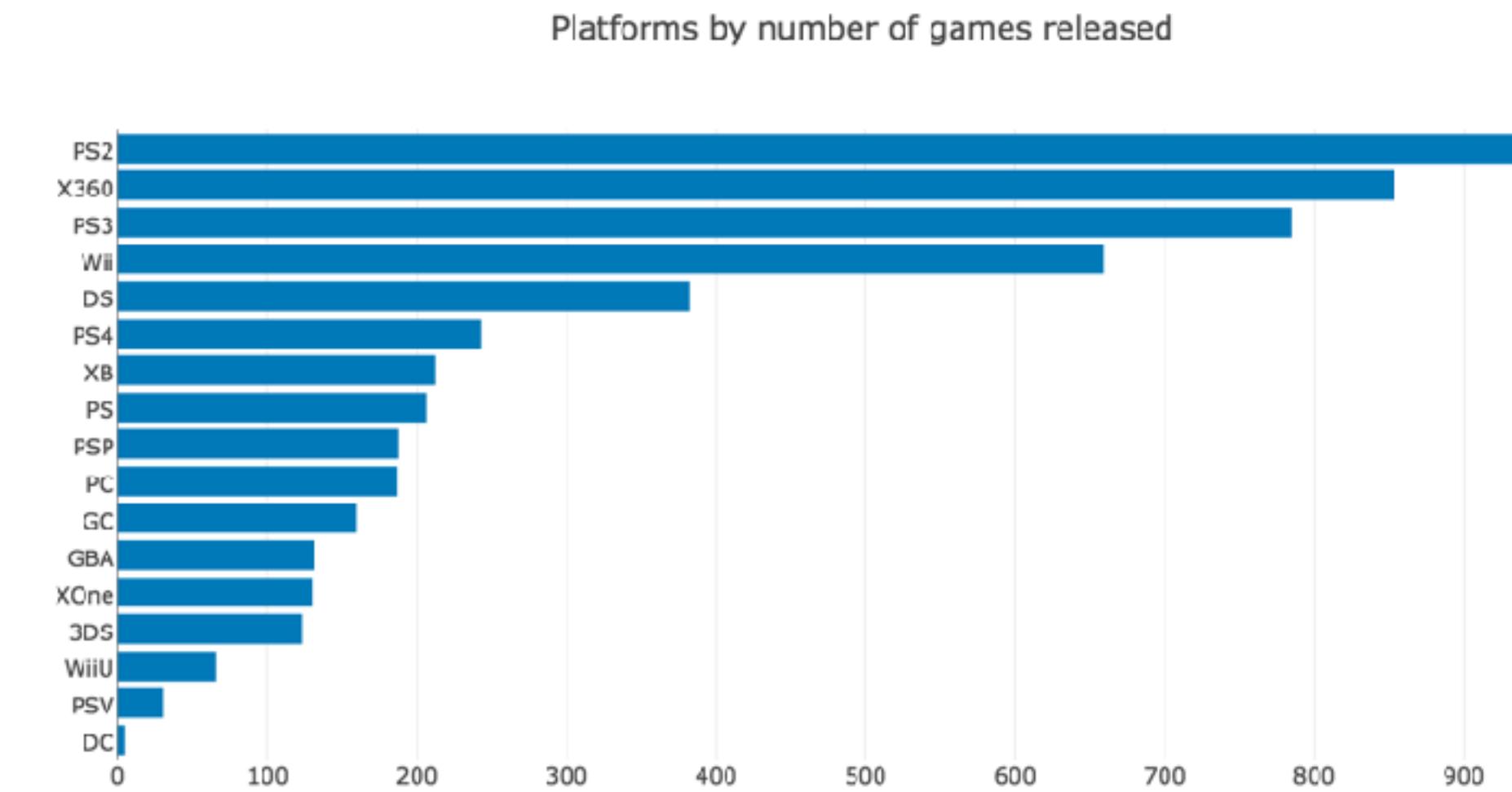
Ranking - значения метрики для категорий, упорядоченные по размеру

Пример - показать, на каких платформах было выпущено большего всего игр

Ranking

Ranking - значения метрики для категорий, упорядоченные по размеру

- › вертикальный или горизонтальный bar chart
- › чтобы выделить большие значения - нужно сортировать по убывания и наоборот



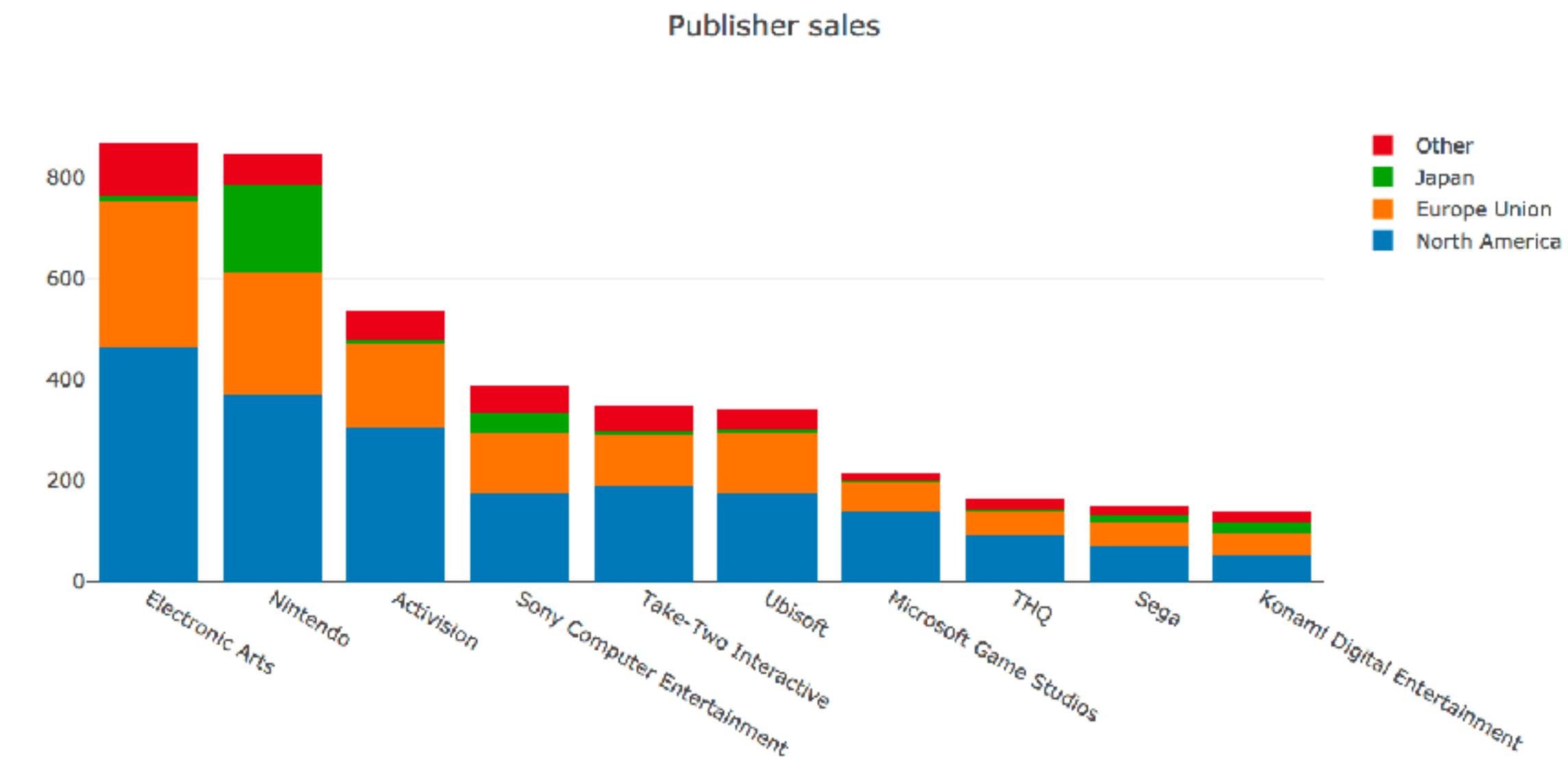
Part-to-hole

- | Part-to-hole - доли отдельных категорий от целого
- | Пример - показать, какие доходы у разных игровых компаний и как они распределяются по рынкам (США, Европа и т.д.)

Part-to-hole

Part-to-hole - доли отдельных категорий от целого

- › вертикальный или горизонтальный bar chart
- › stacked bar chart, только если нужно отобразить суммарное значение



Deviation

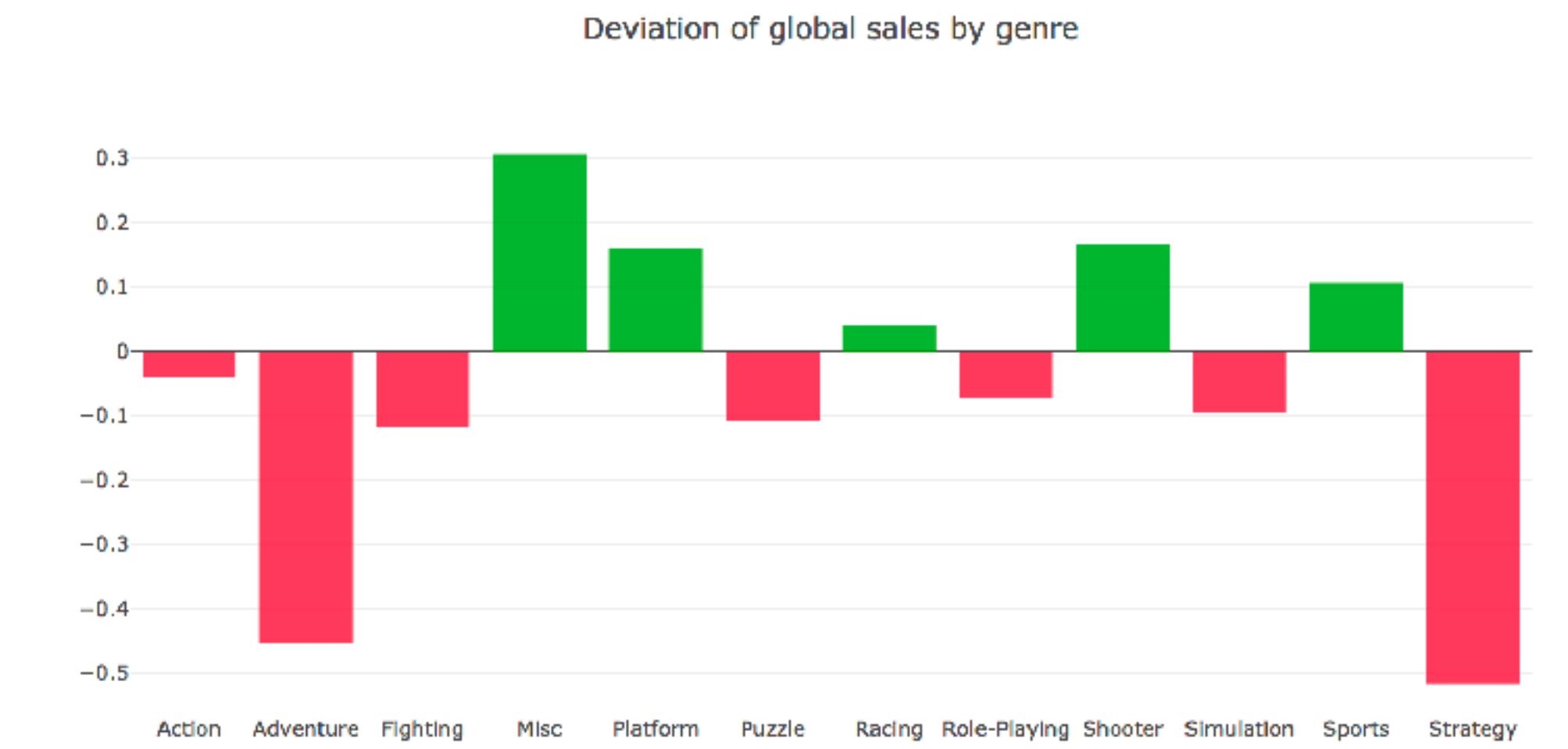
| Deviation - сравнение показателей для категорий с baseline

| Задача - посмотреть, как отличаются средние прожажи для разных жанров

Deviation

Deviation - сравнение показателей для категорий с baseline

- › bar chart, чтобы подчеркнуть отдельные значения



Frequency Distribution

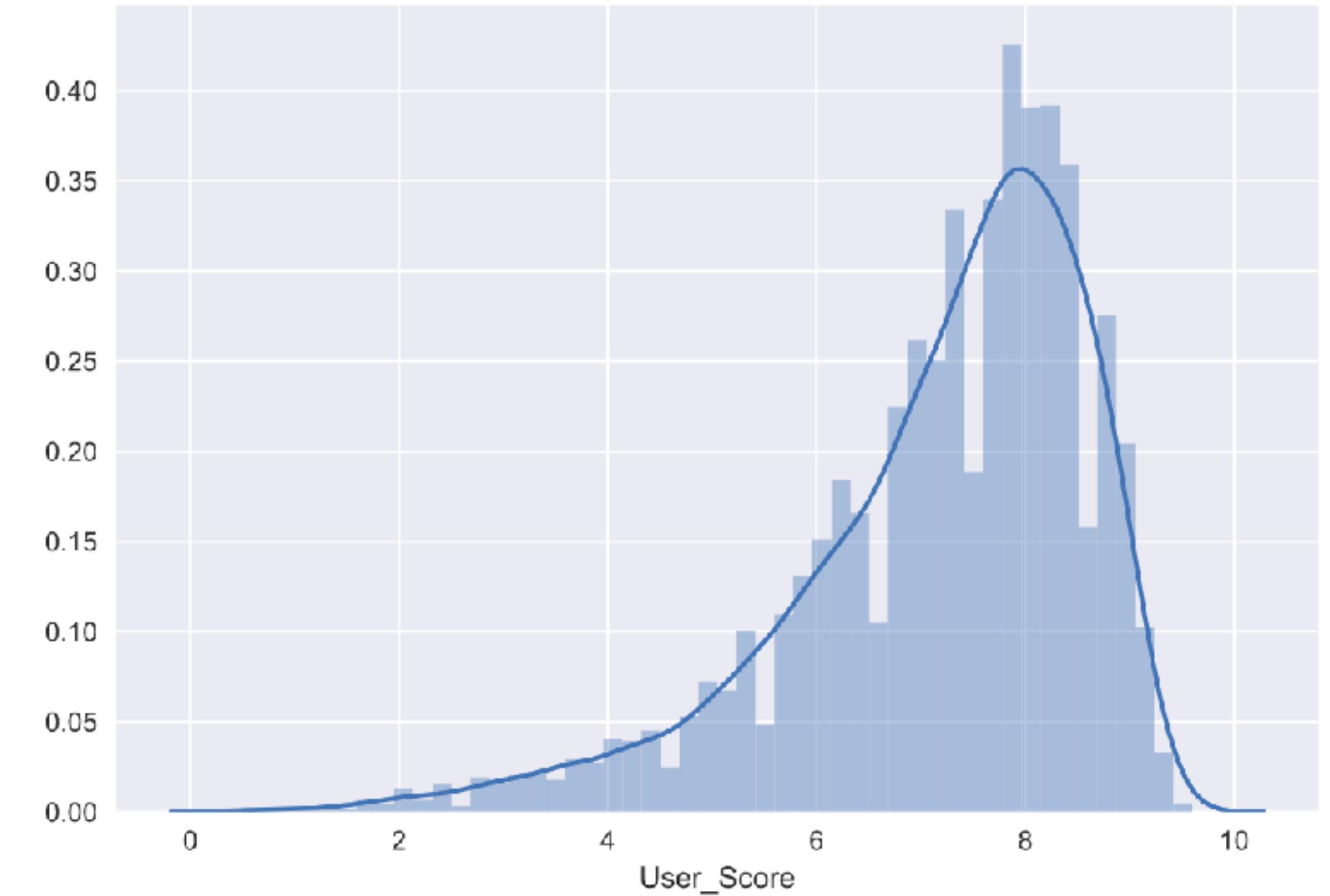
| Frequency Distribution -
распределение величины (может
быть нормированным)

| Задача - показать распределение
пользовательских оценок игр

Frequency Distribution

Frequency Distribution -
распределение величины (может
быть нормированным)

- › vertical bar chart, чтобы выделить
отдельные величины (histogram)
- › line chart, чтобы показать общий
pattern (frequency polygon)



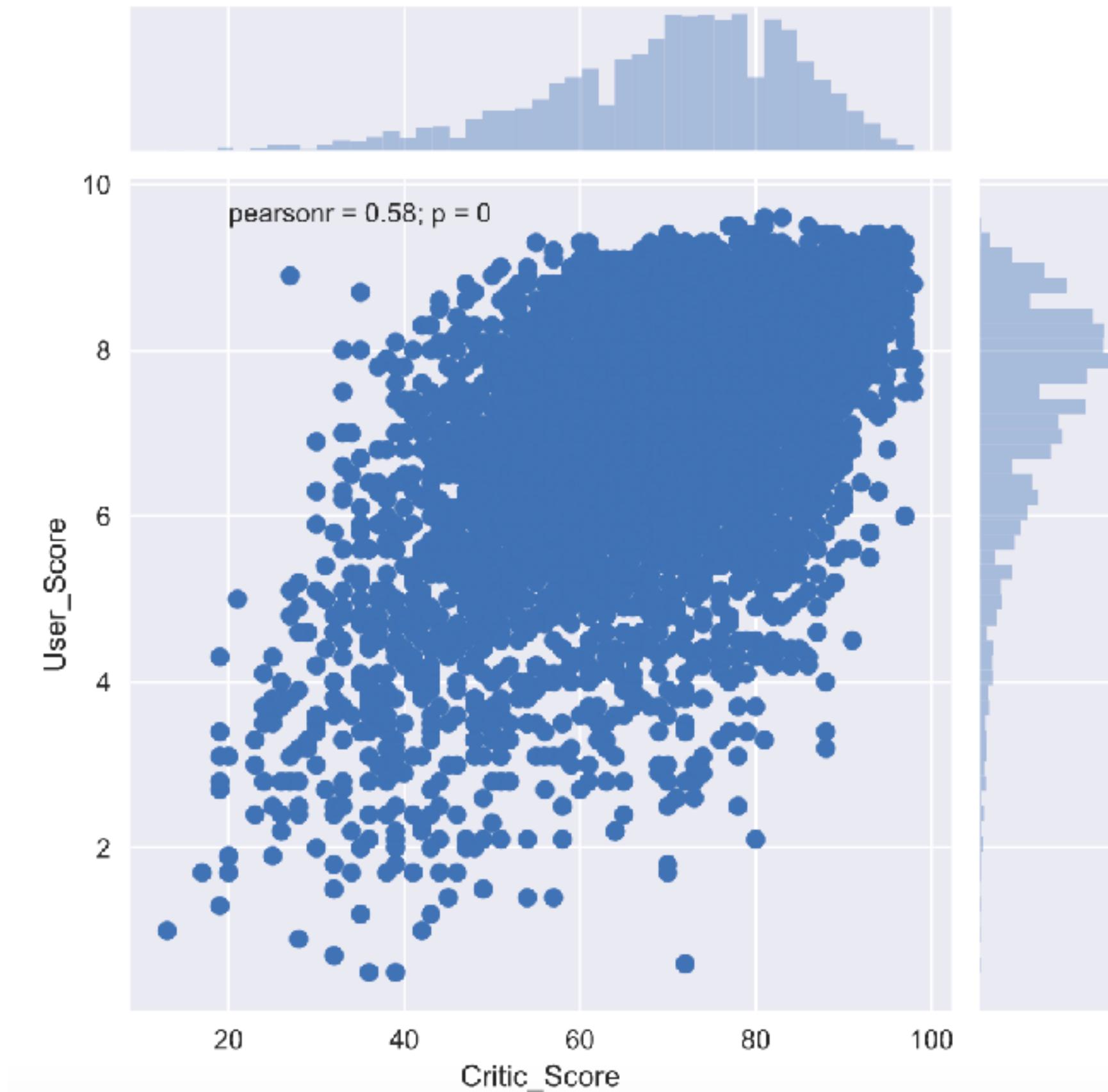
Correlation

| Correlation - кореляция между двумя численными величинами

| Задача - показать, как связаны между собой оценки пользователей и критиков

Correlation

| Correlation - кореляция между
двумя численными величинами
› scatter plot и линия тренда



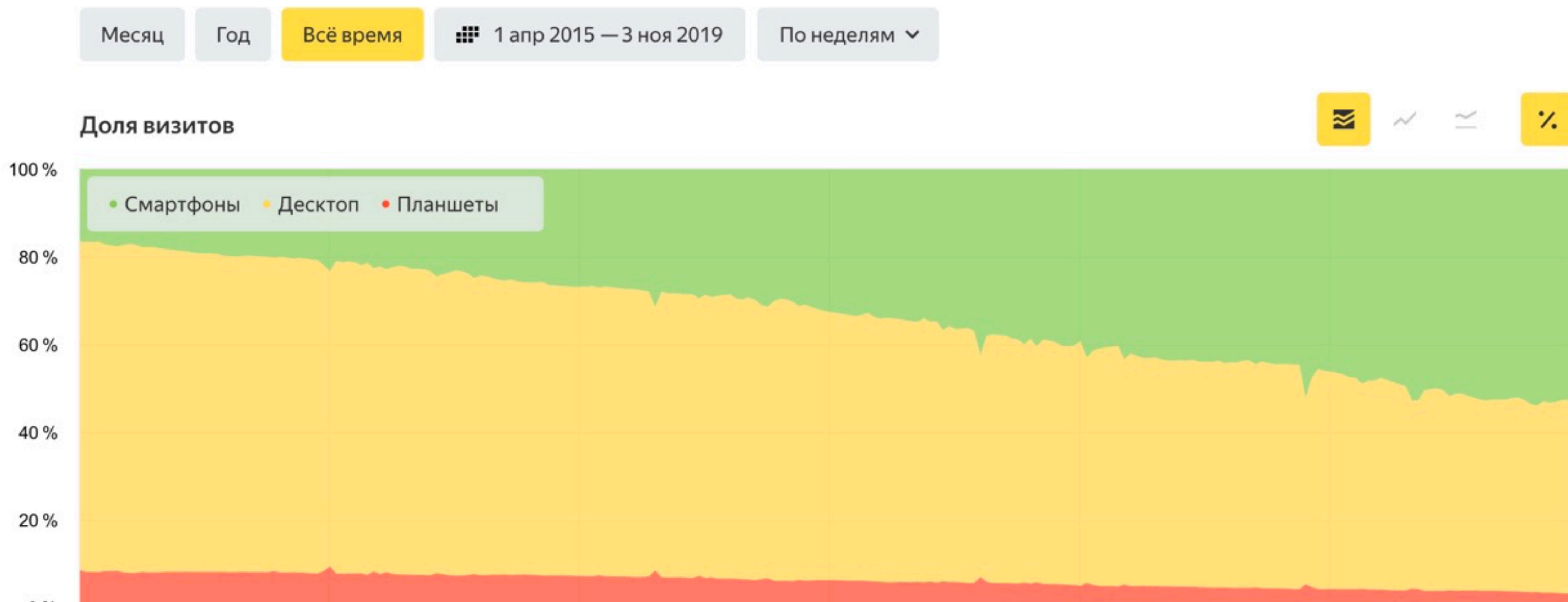
А этого достаточно?



Как показать изменение
разбивки трафика
по устройствам
со временем?

Но бывают и комбинированные варианты

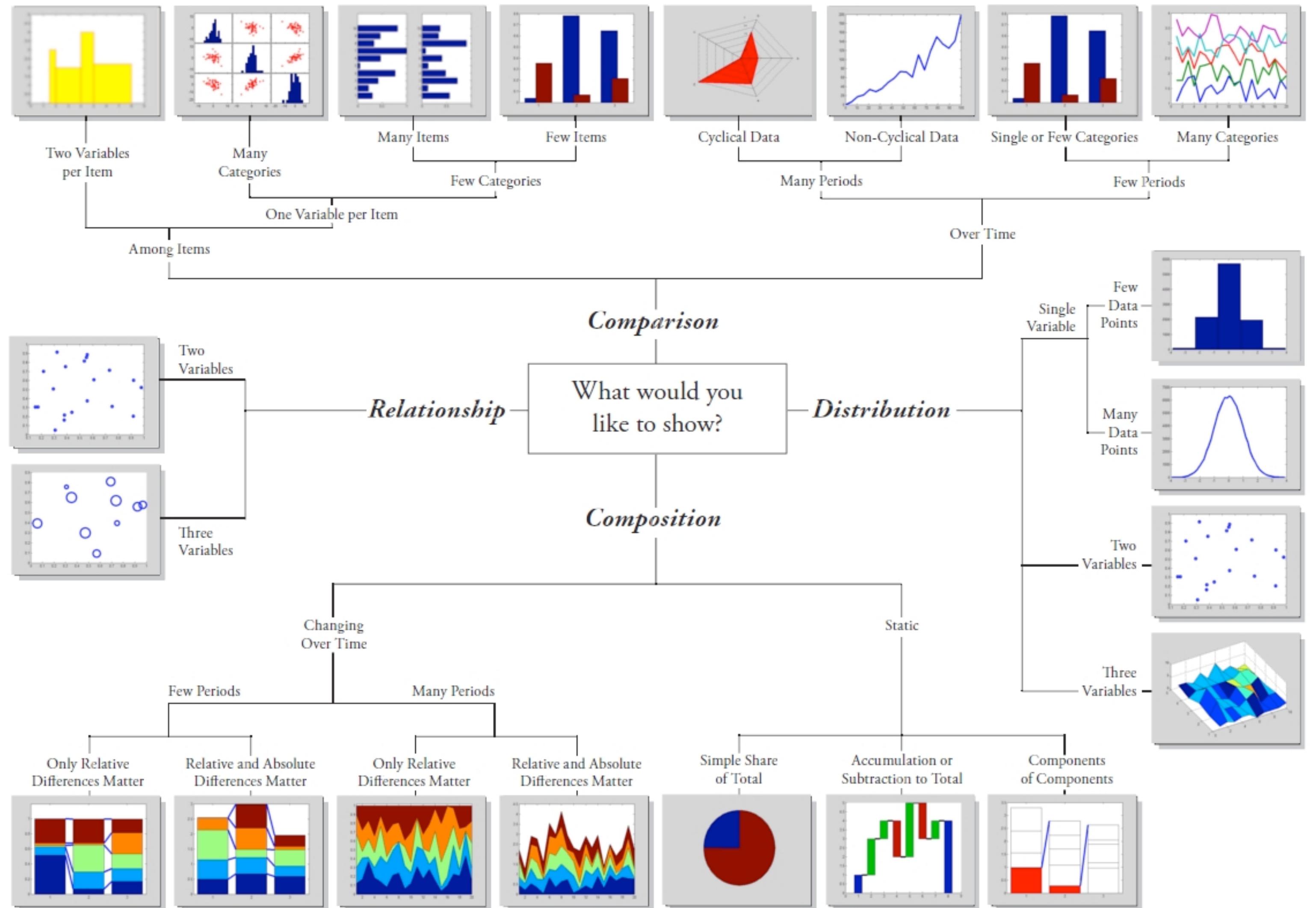
по данным Яндекс.Метрики с задержкой 7 дней





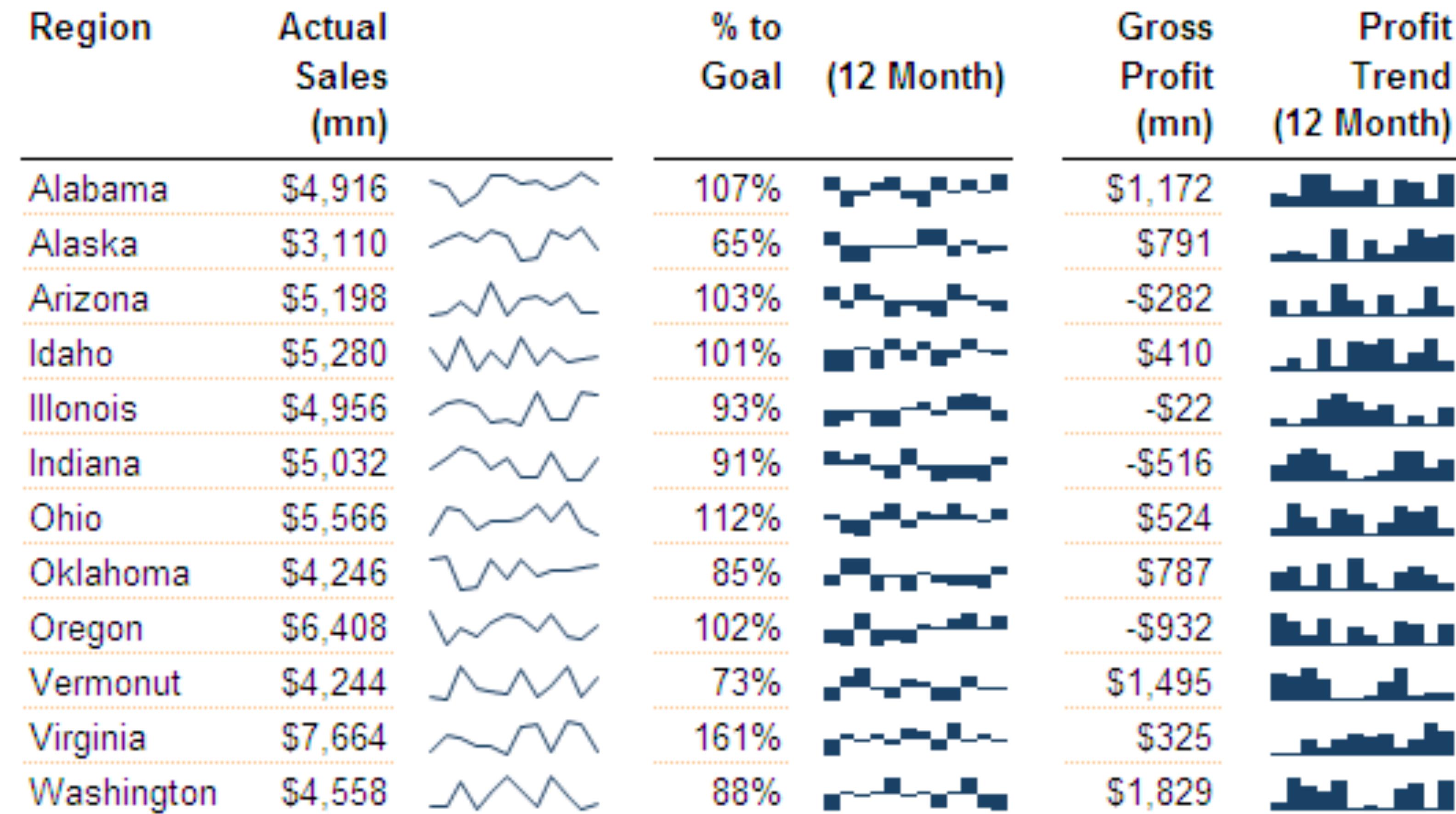
Есть и другие
визуализации...

Chart Suggestions—A Thought-Starter

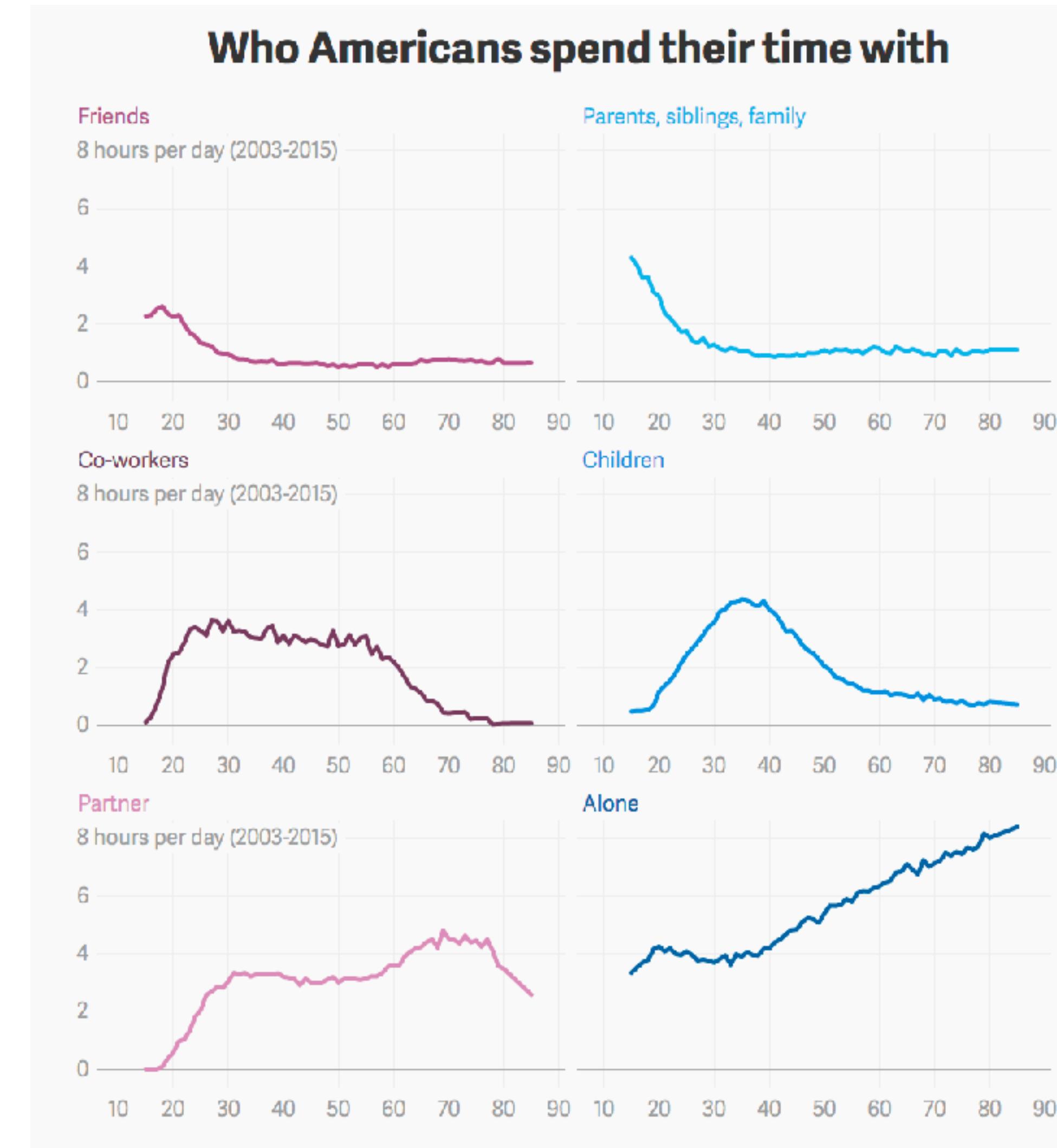


Modified with permission -Doug Hull
blogs.mathworks.com/videos © 2009 A. Abela — a.v.abela@gmail.com
hull@mathworks.com 2009

Table

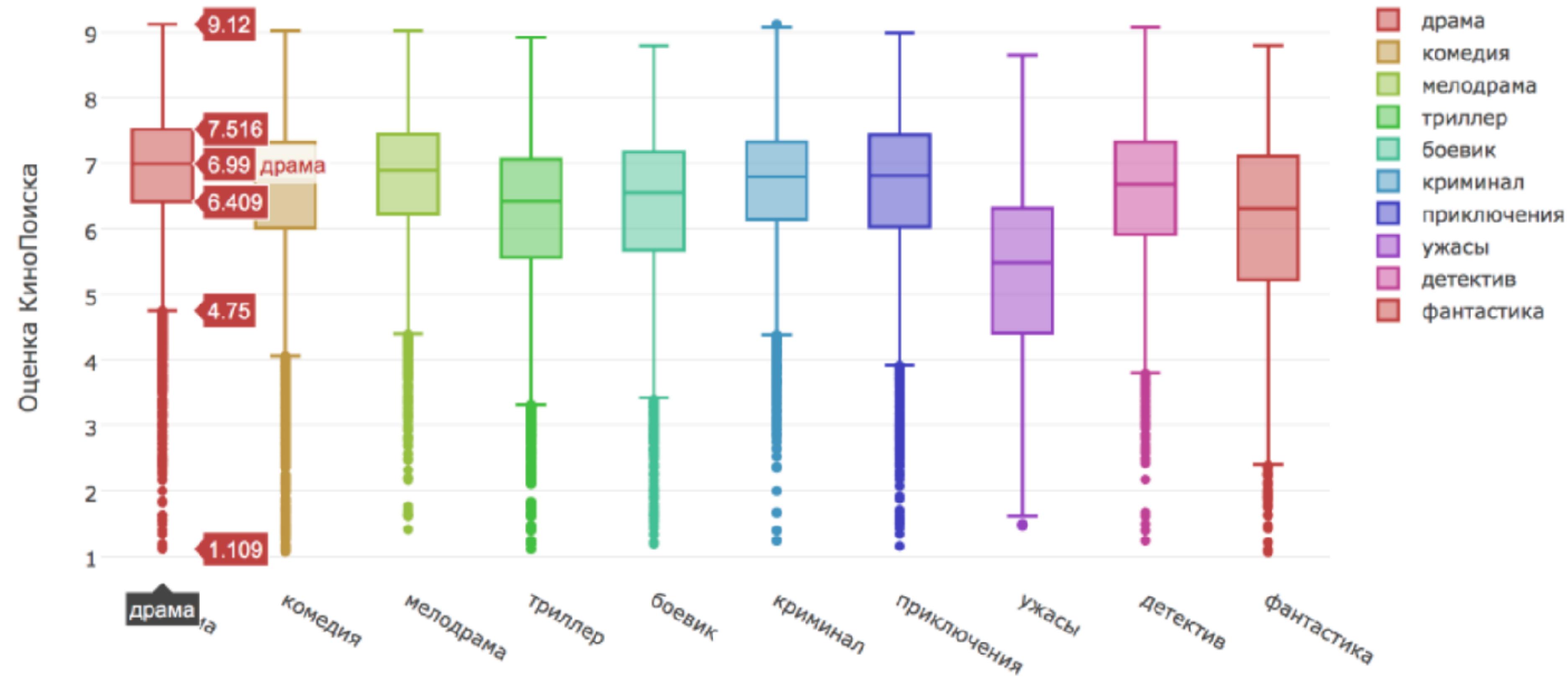


Subplots



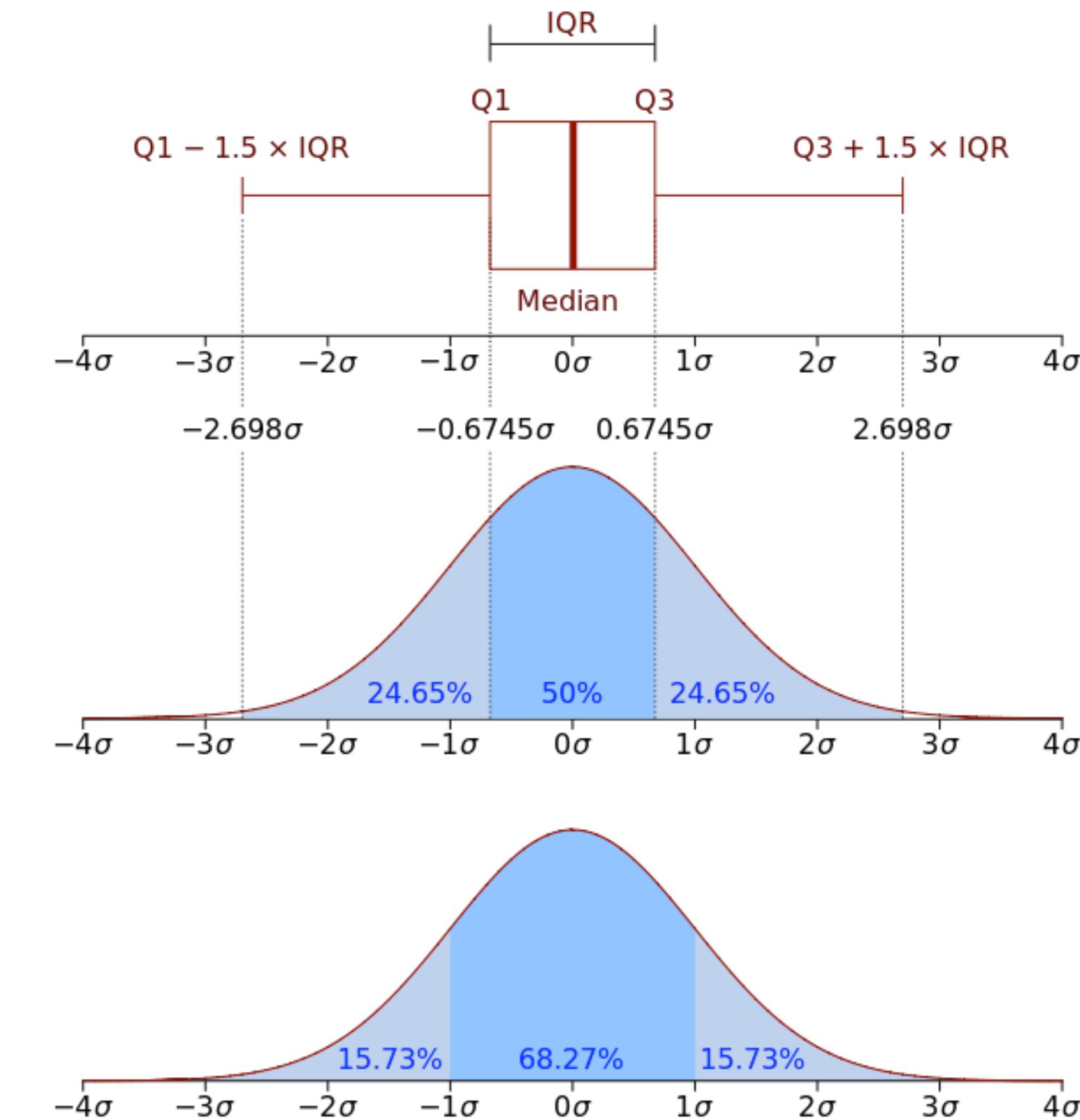
Box plot

Оценки фильмов



Box plot uncovered

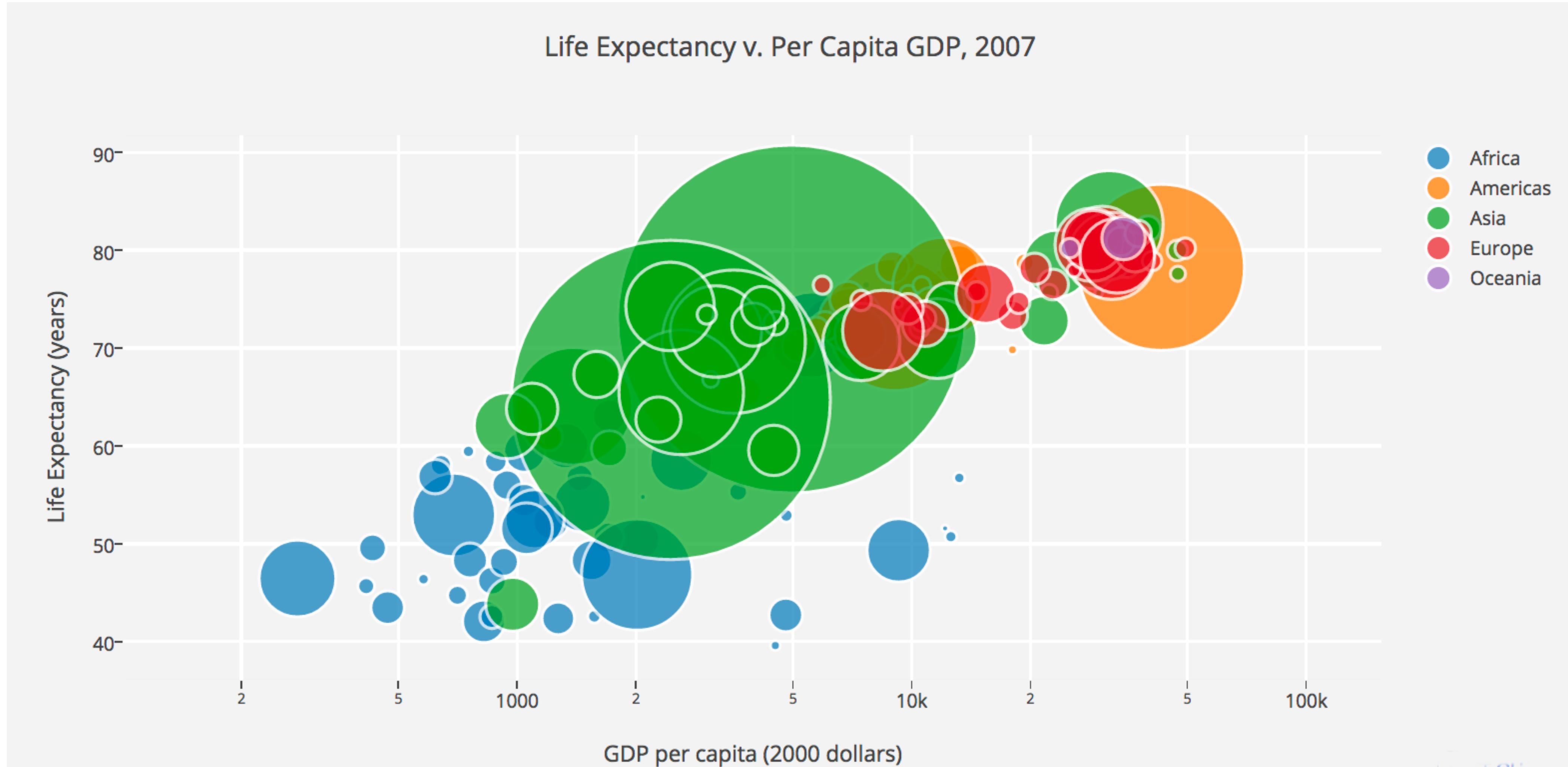
- › линия - медиана
- › коробка - IQR
- › усы - $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$
- › точки - outliers



Heatmap

	Уборка и клининг	Туризм, спорт и отдых	Сантехника и климат	Свет и электрика	Товары для дома	Уборка и клининг				
Автотовары	0.6	1.7	3.1	1.4	1.1	1.6	0.8	1.0	1.1	0.9
Всё для сада	0.7	0.8	1.9	0.8	0.4	0.6	0.3	0.7	0.9	0.6
Всё для строительства	3.5	4.3	8.5	6.5	3.7	4.5	4.0	5.1		8.4
Другое	4.4	5.9	11.3	4.9	4.4	6.3	5.2		5.1	6.2
Инструменты и оборудование	2.9	4.6	7.1	2.7	3.5	4.3		5.5	4.2	3.0
Расходные материалы	6.8	14.4	33.0	14.5	18.6		5.8	9.1	6.4	7.4
Сантехника и климат	14.2	16.6	47.3	18.6		43.0	11.1	14.6	12.3	12.7
Свет и электрика	0.8	0.9	2.0		0.9	1.7	0.4	0.8	1.1	1.2
Товары для дома	1.6	3.5		2.4	2.9	4.7	1.4	2.4	1.8	3.5
Уборка и клининг	3.0		9.6	3.0	2.8	5.6	2.5	3.3	2.4	4.1
Автотовары		3.3	4.9	3.0	2.7	2.9	1.7	2.7	2.2	3.9

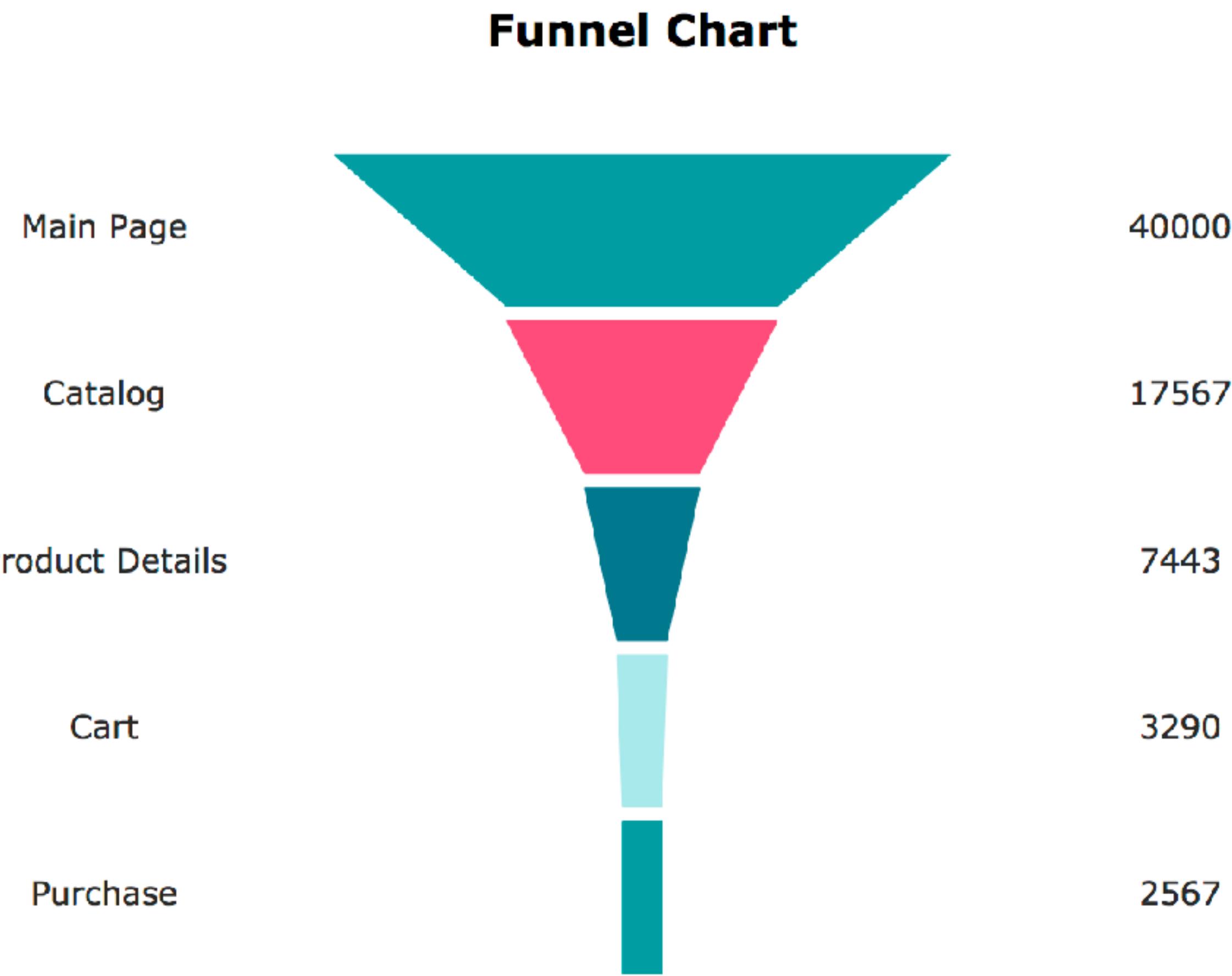
Bubble chart



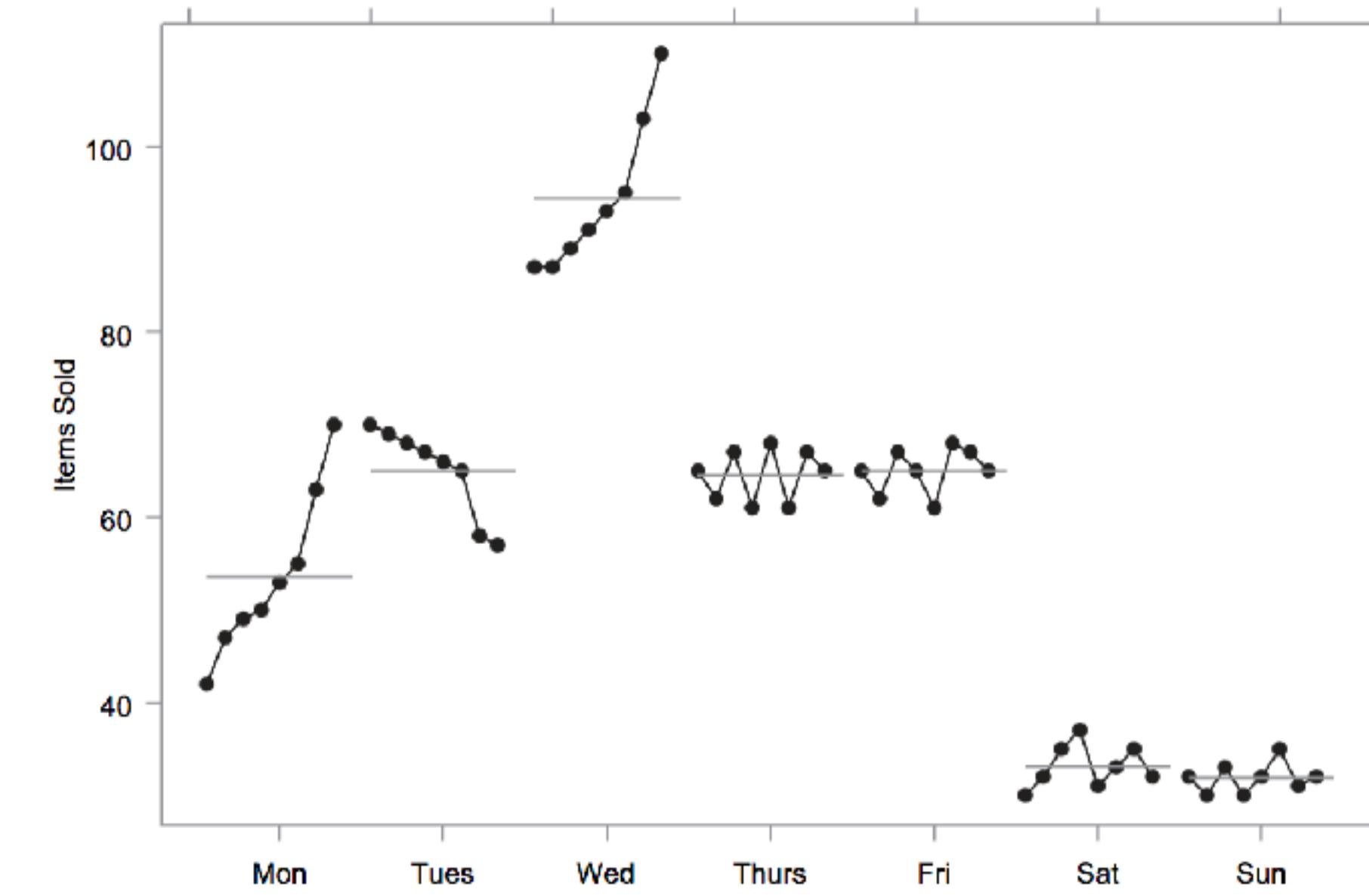
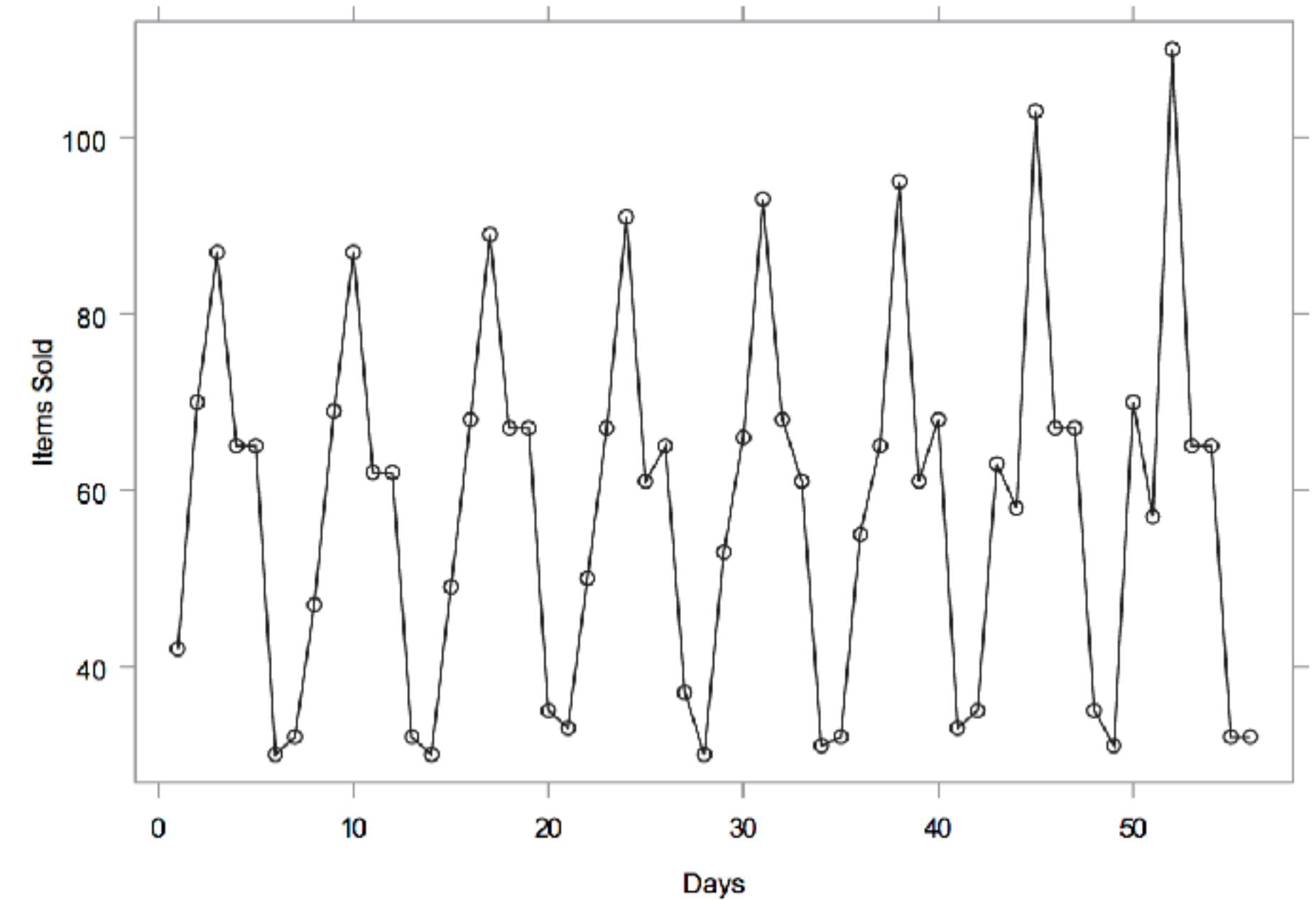


И даже такие...

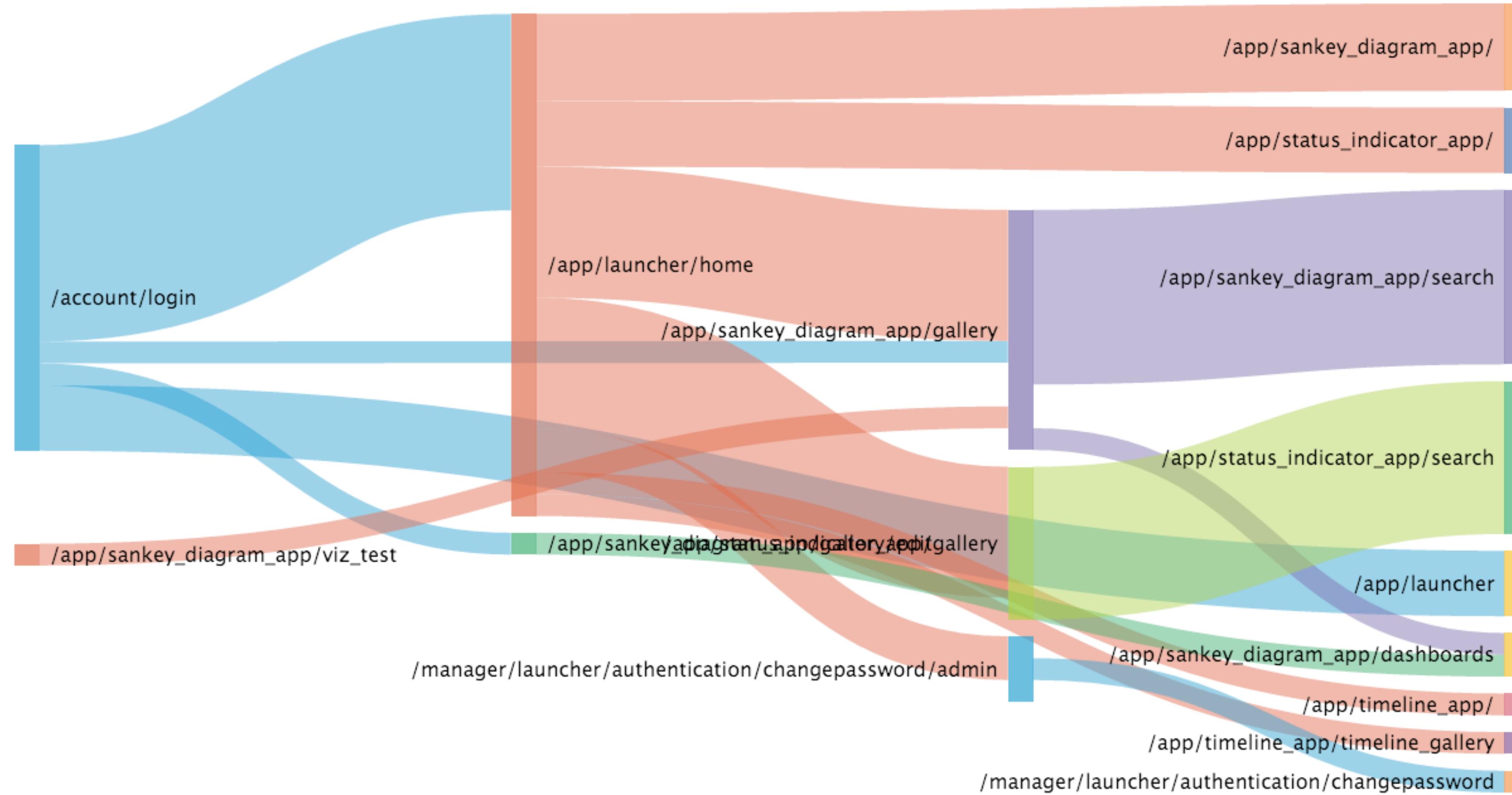
Funnel chart



Cycle Plot



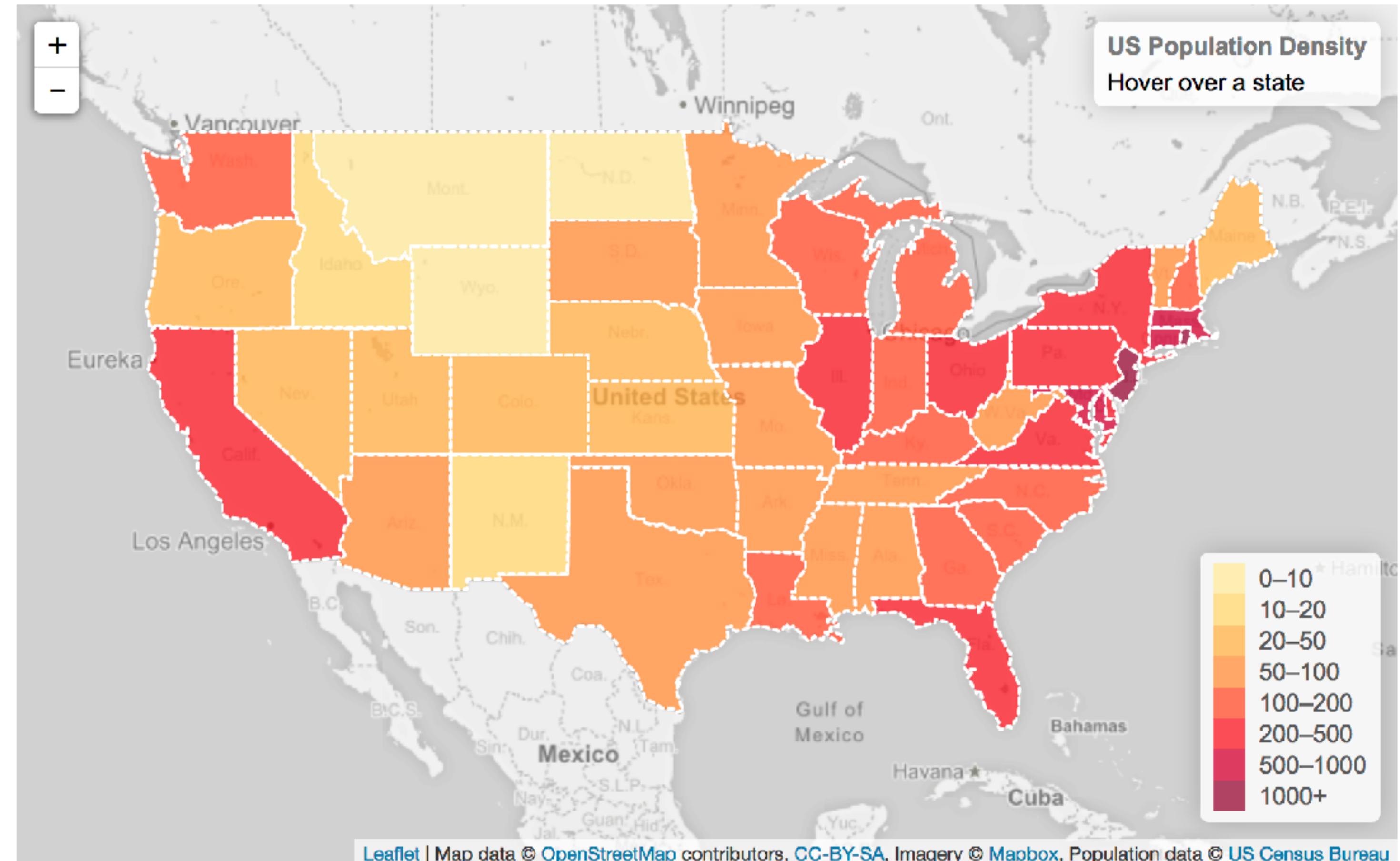
Sankey diagram



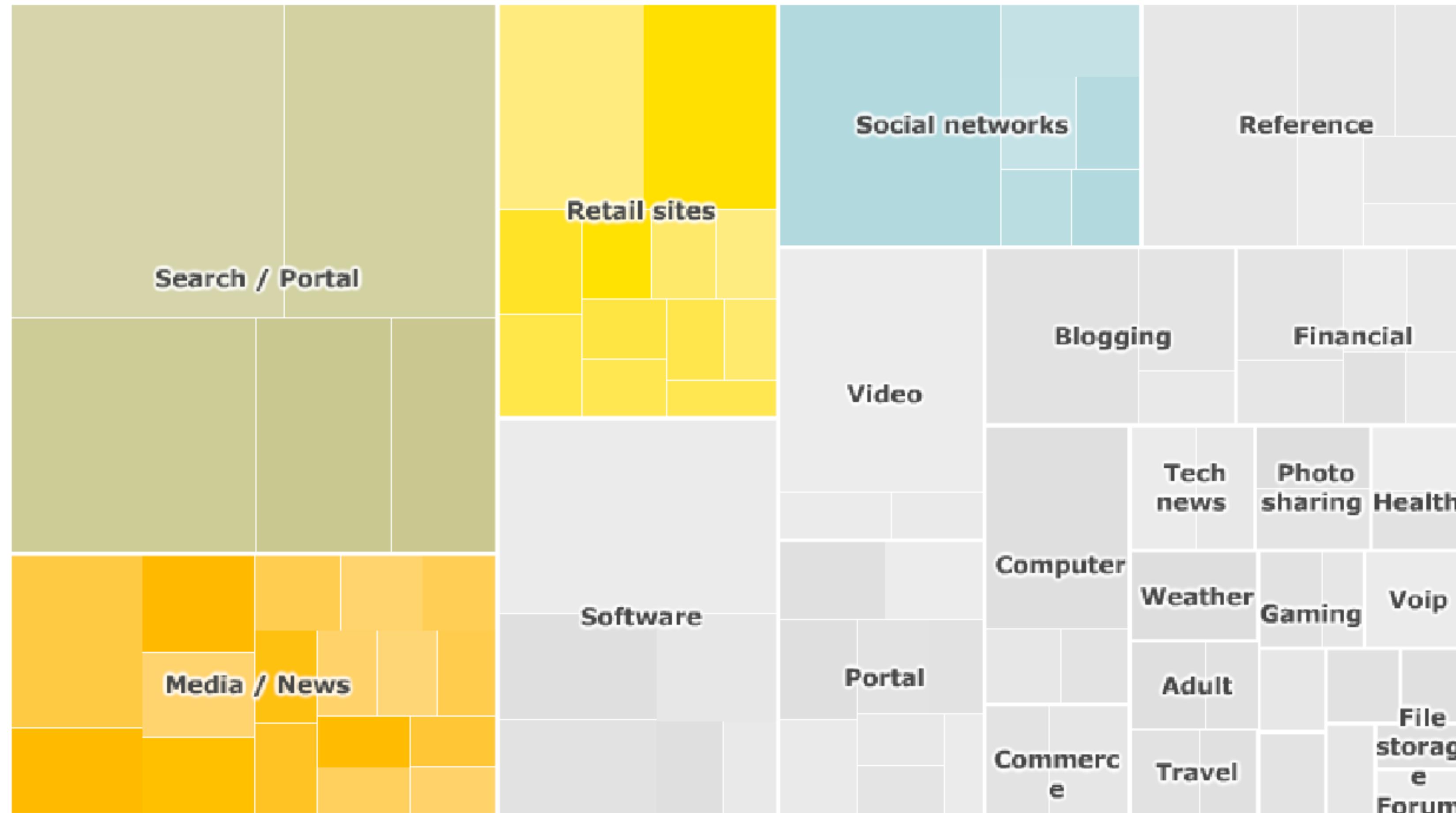
Area Chart + Line Chart



Гео-данные (Choropleth)



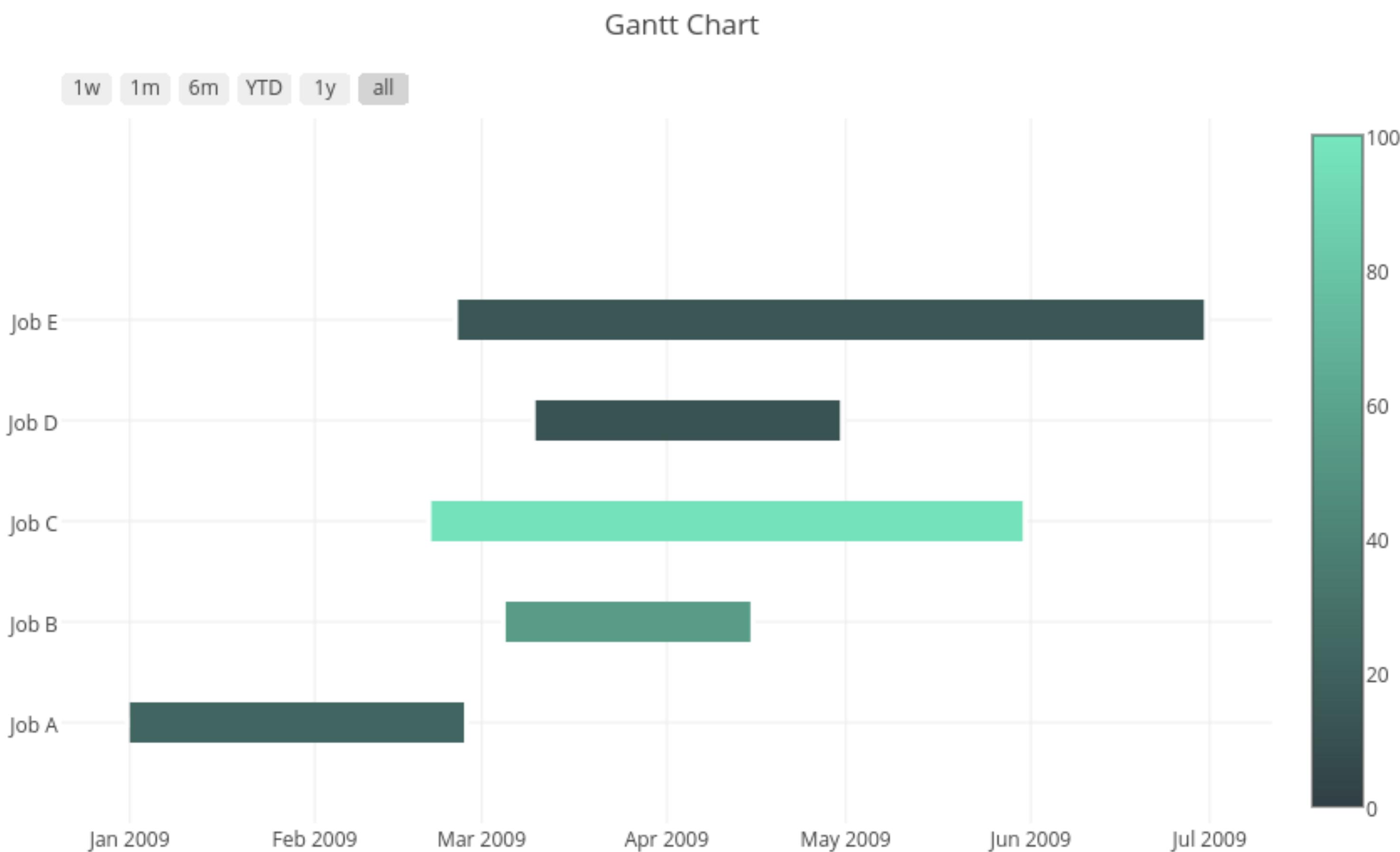
TreeMap



Network Visualization



Gantt Chart



И даже word cloud





The greatest value of a picture
is when it forces us to notice
what we never expected to see.

John Tukey



Excellence in statistical graphics
consists of complex ideas
communicated with clarity,
precision and efficiency.

Edward Tufty

Edward Tufty «The Visual Display of Quantitative Information»

| Visualization should...

- › show the data
- › avoid distorting what the data has to say
- › present many numbers in a small space
- › encourage the eye to compare different pieces of data
- › reveal the data at several levels of detail, from a broad overview to the fine structure

Инструменты визуализации



Python библиотеки

- › matplotlib
- › seaborn
- › plotly
- › ggplot
- › bokeh
- › pygal
- › и т.д.

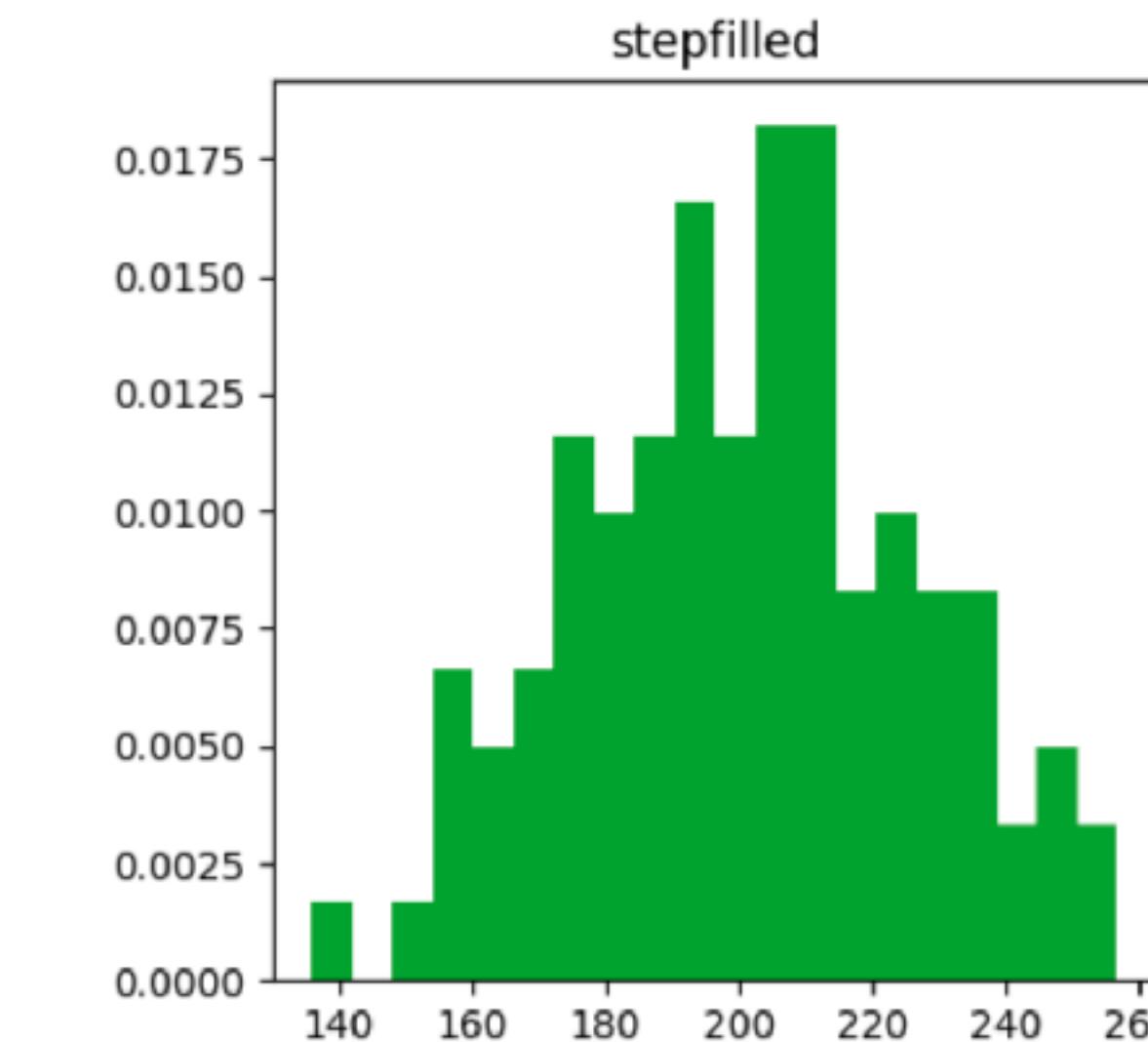


bokeh

Pygal

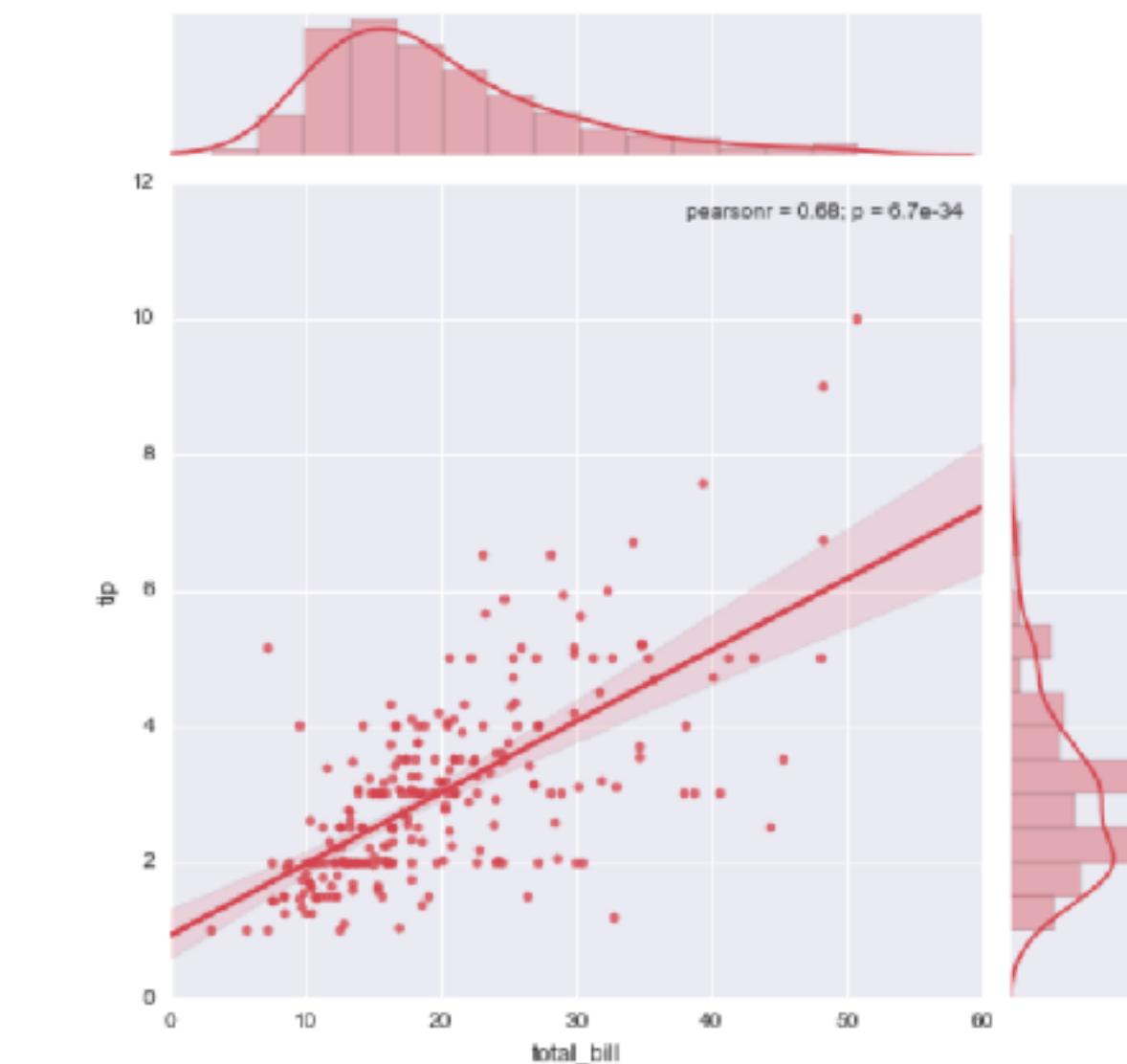
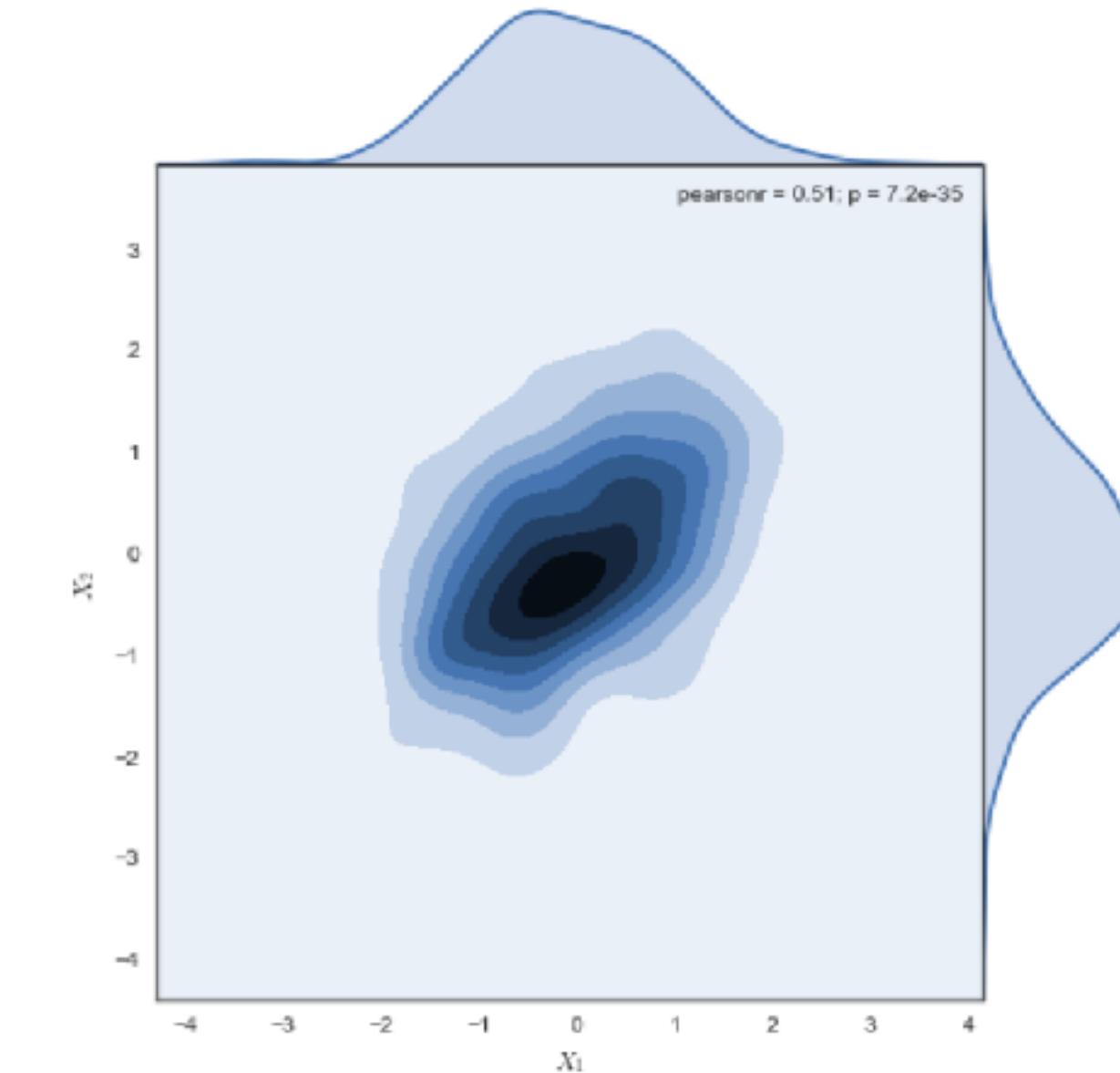
matplotlib

- › первая библиотека на python для визуализации
- › очень гибкая, но и монструозная при этом
- › стили родом из 90х
- › wrappers - pandas, seaborn



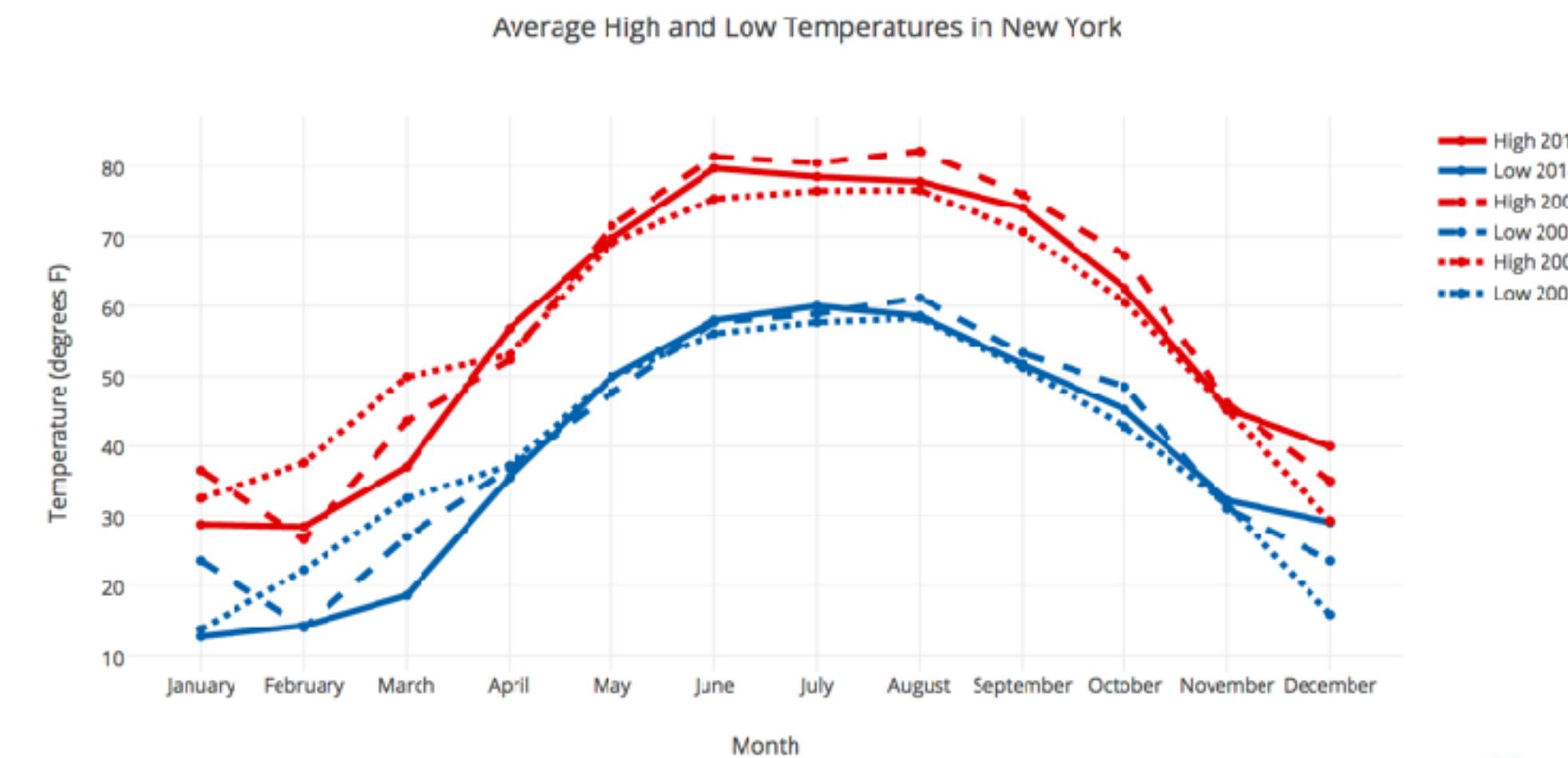
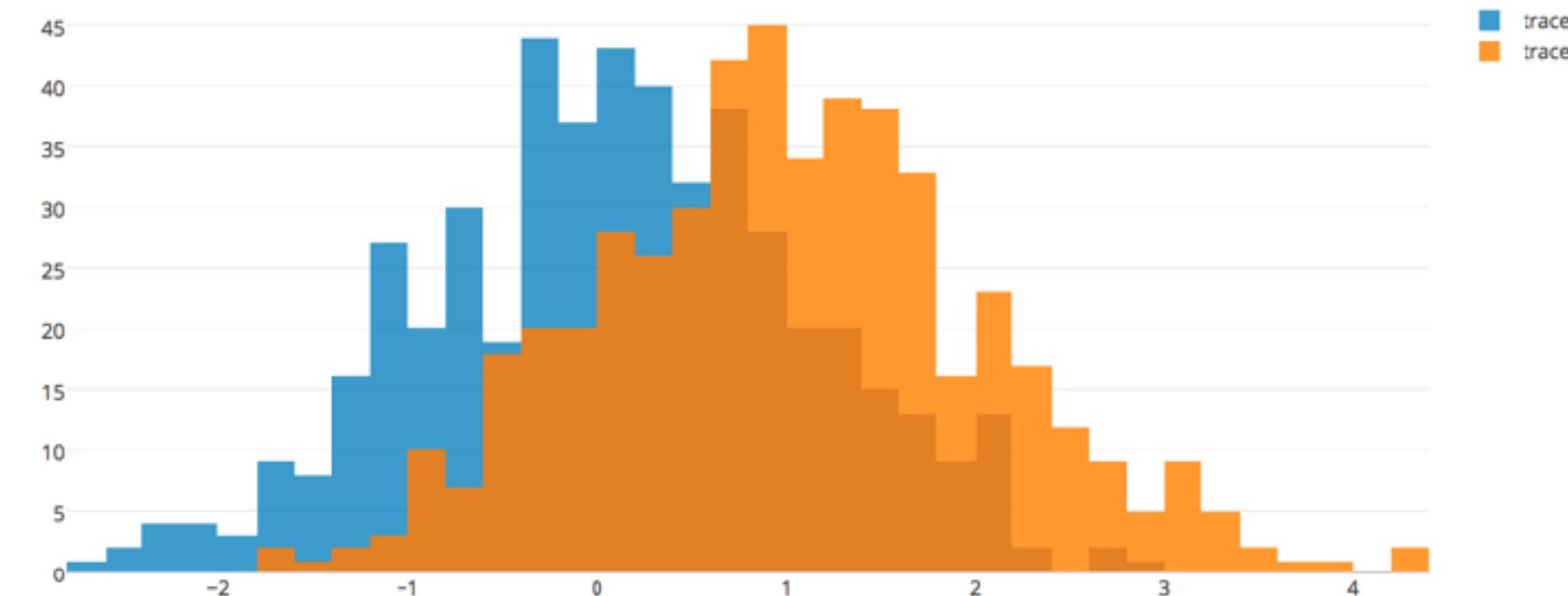
seaborn

- › на основе matplotlib
- › сложные графики за пару строк кода
- › симпатичные default стили
- › для изменения мелочей нужно лезть в дебри matplotlib



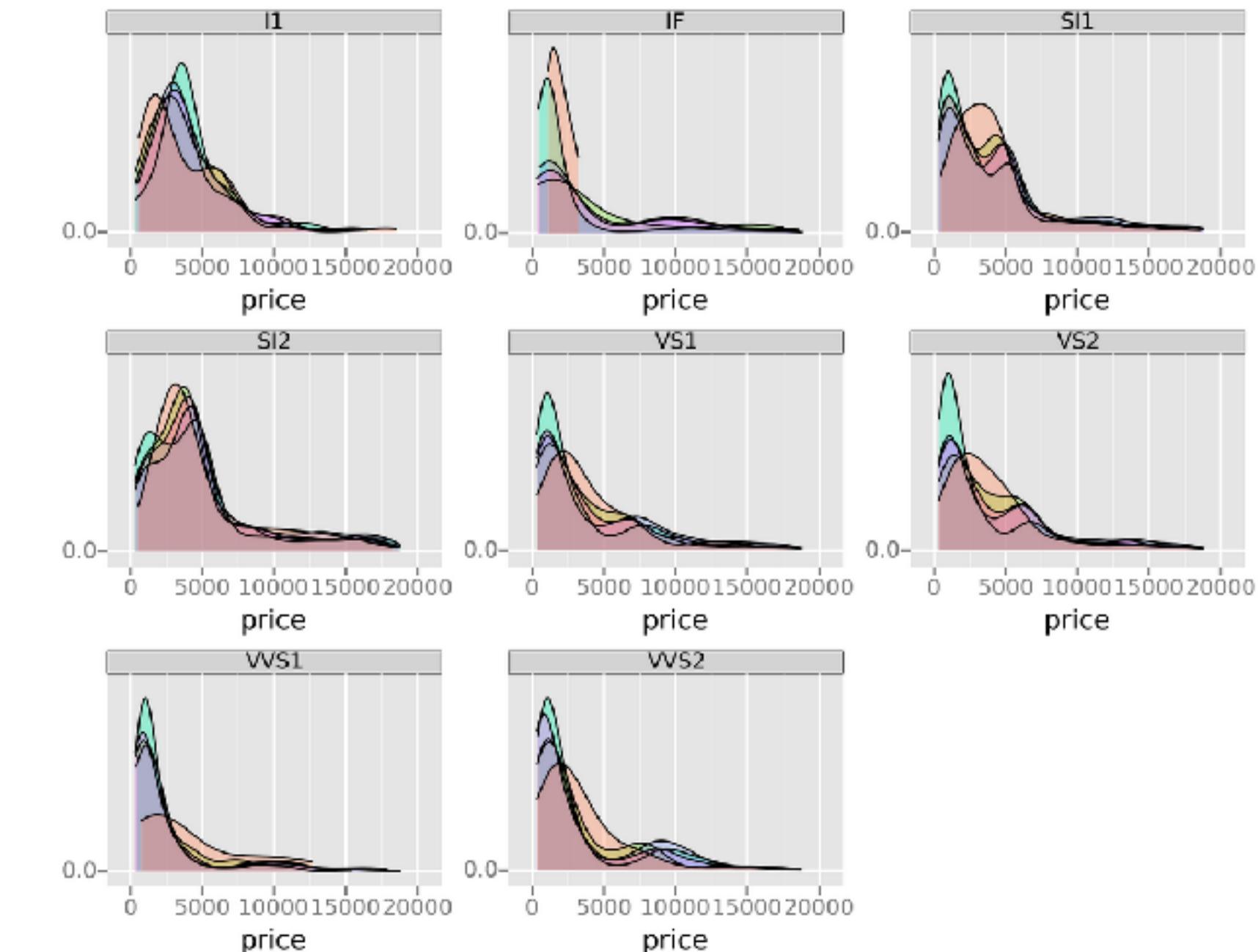
plot.ly + dash

- › интерактивные графики
- › простой API, но есть возможность настройки (тоже придется покопаться в документации)
- › удачные default'ы
- › dash - для полноценных web apps



ggplot

- › на базе ggplot2 в R
- › идеология The Grammar of Graphics: слои компонент (точки, линии, оси)
- › проще matplotlib, но менее гибкий

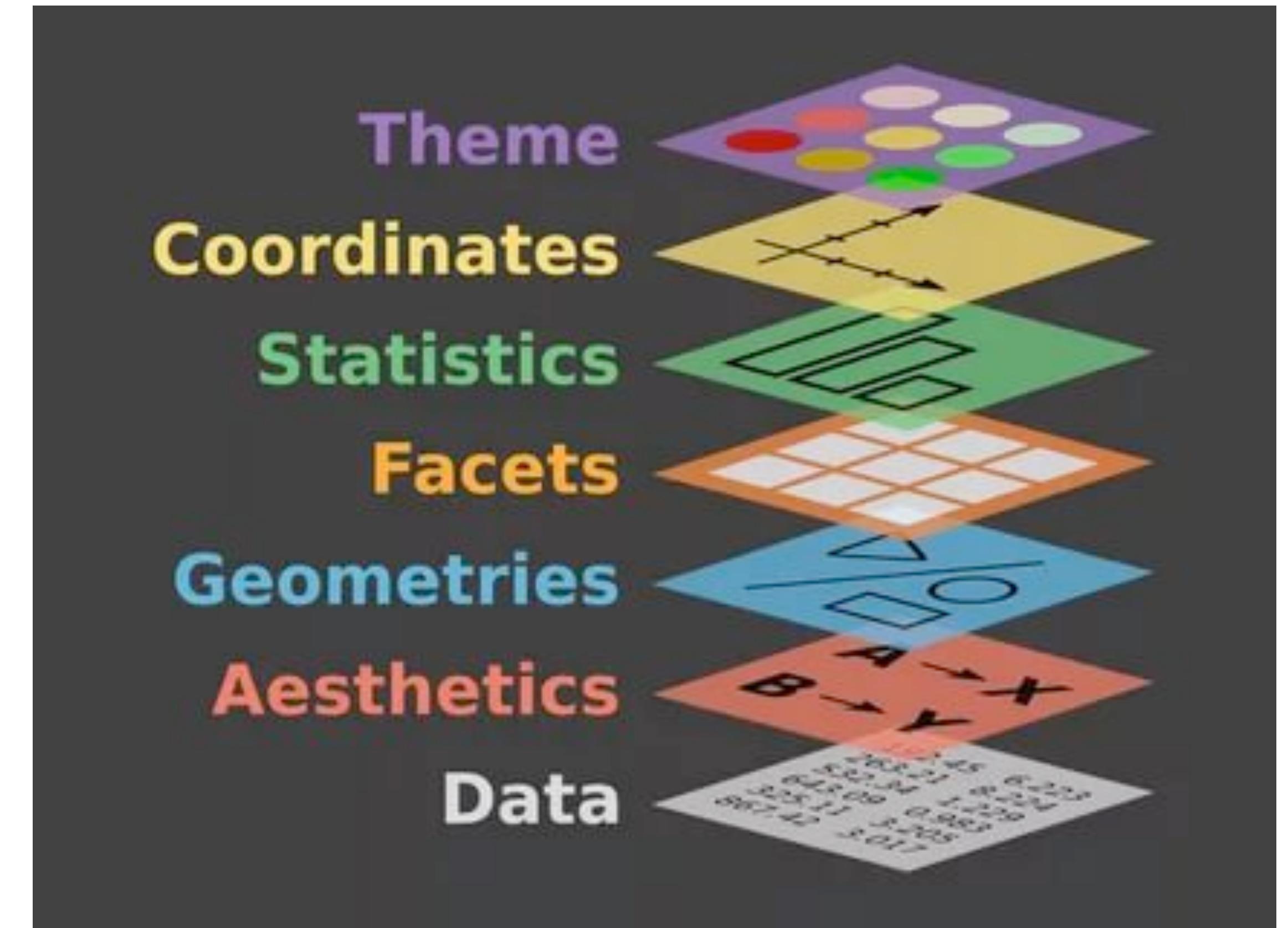


The Grammar of Graphics

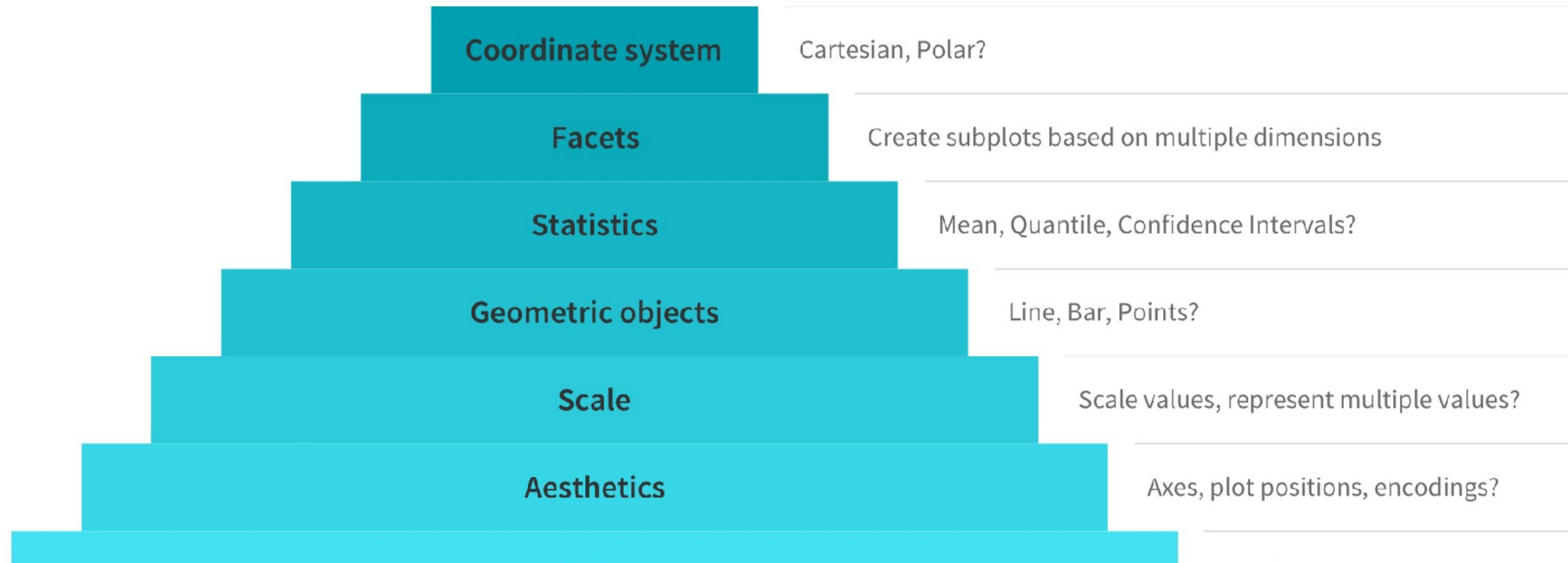
› Leland Wilkinson, 1999

Принципы

- › Отделяем данные data от представления aesthetic
- › Определяем основные элементы и графики
- › Комбинируем их

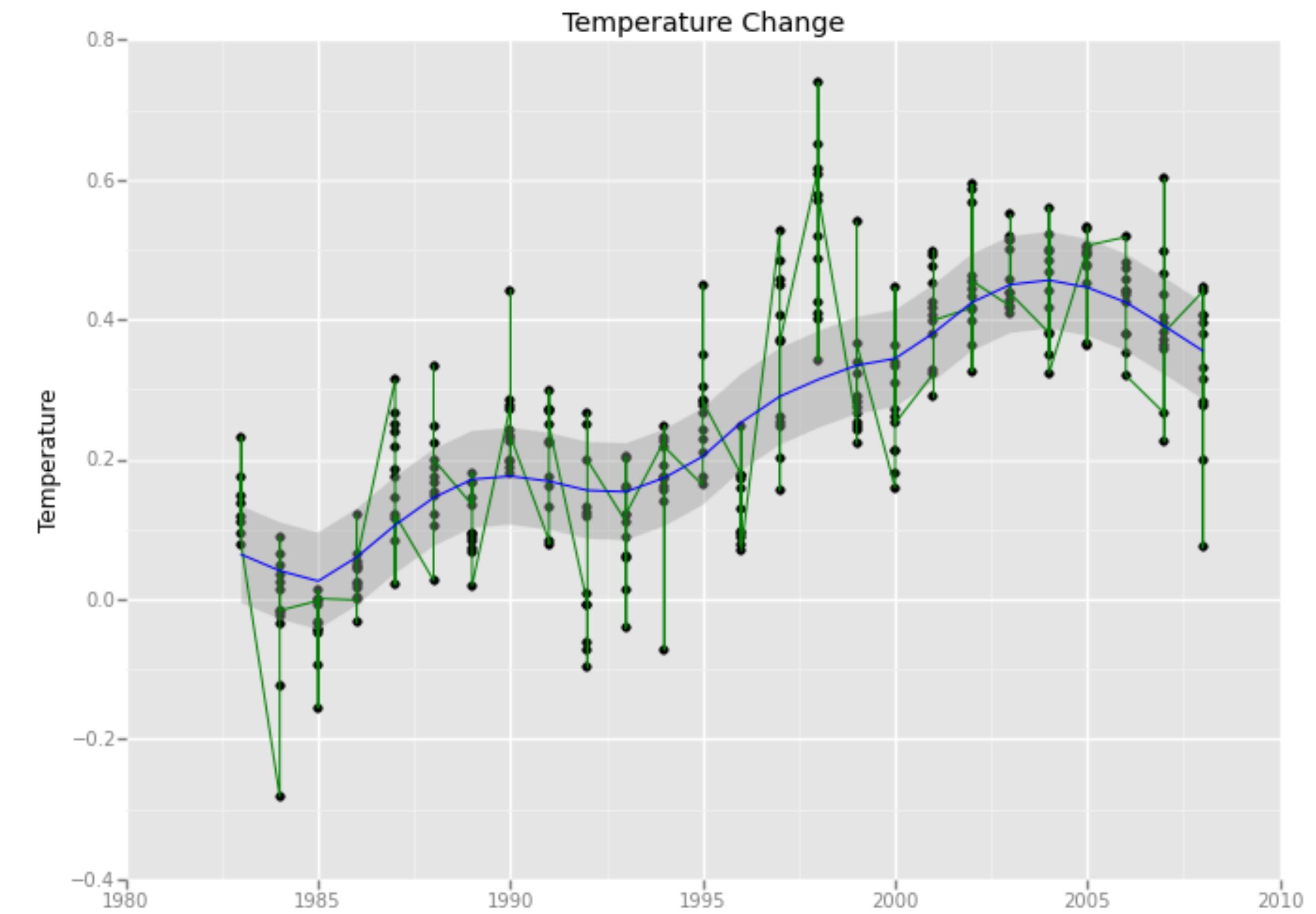


Основные компоненты



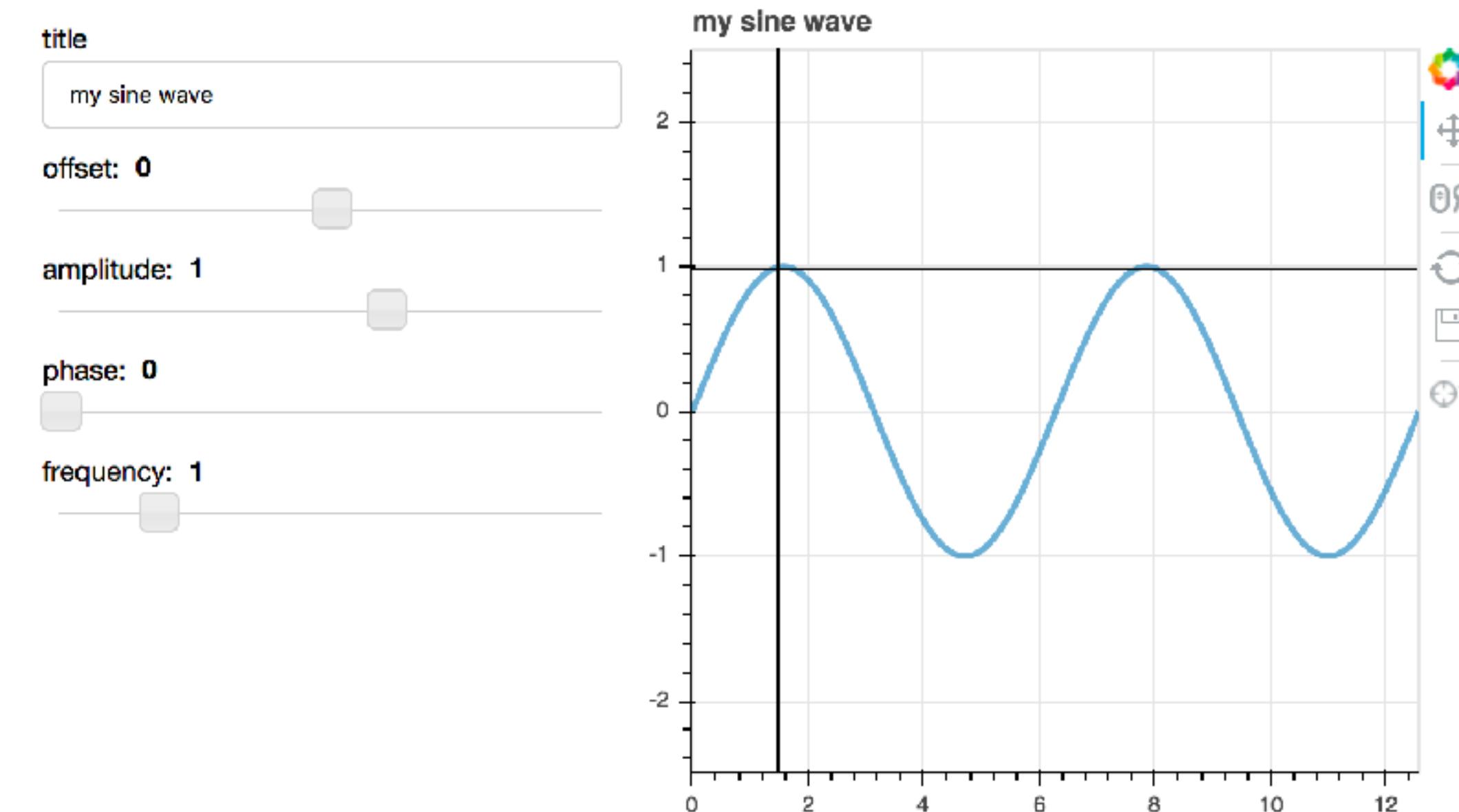
Пример

```
ggplot(clim, aes('Year', 'Temp'))  
+geom_line(color='green')  
+geom_point()  
+ggtitle('Temperature Change')  
+xlab("")+ylab('Temperature')  
+stat_smooth(colour='blue', span=0.2)
```



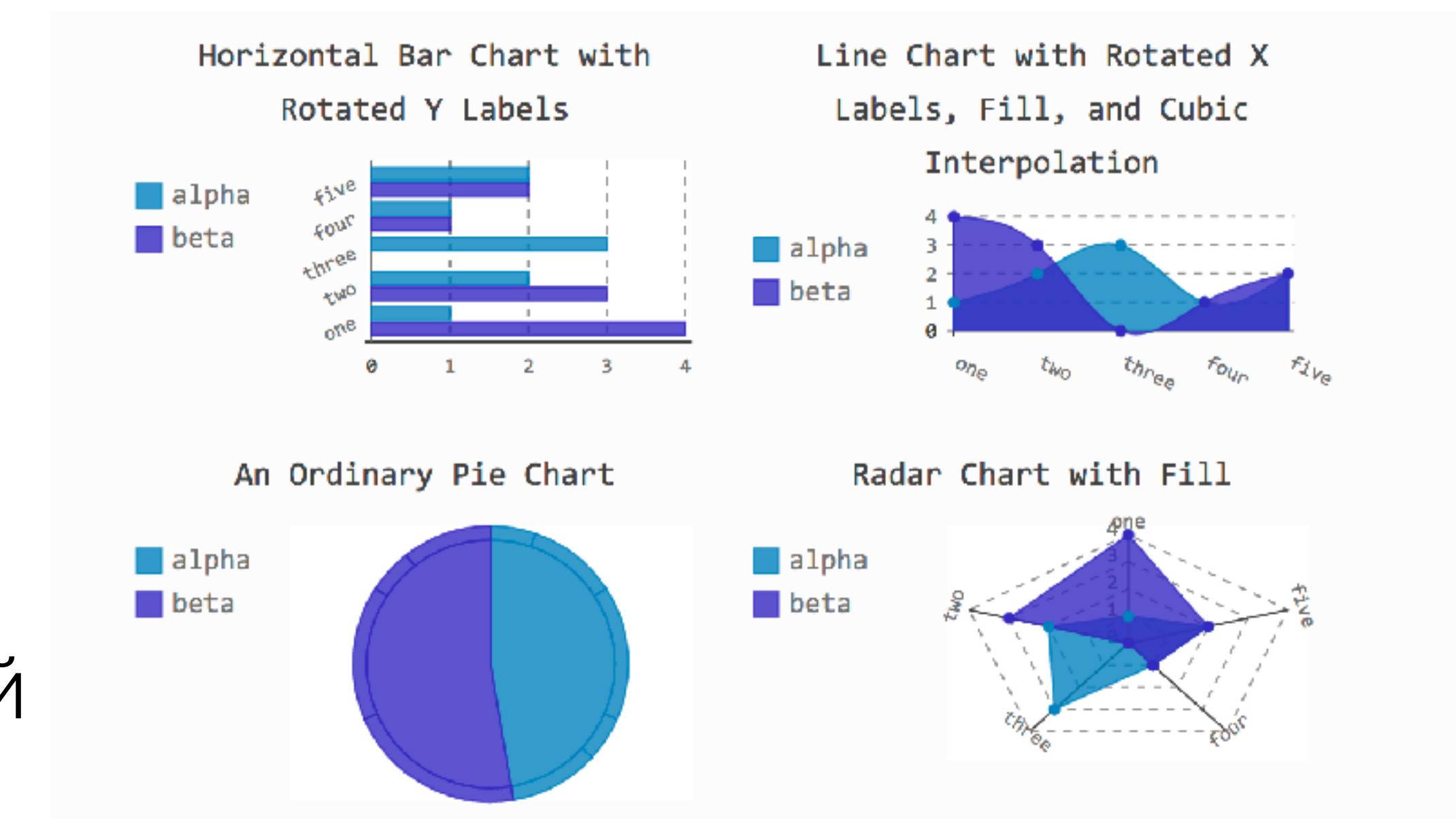
bokeh

- › идеология The Grammar of Graphics
- › интерактивные графики
- › 3 уровня сложности API



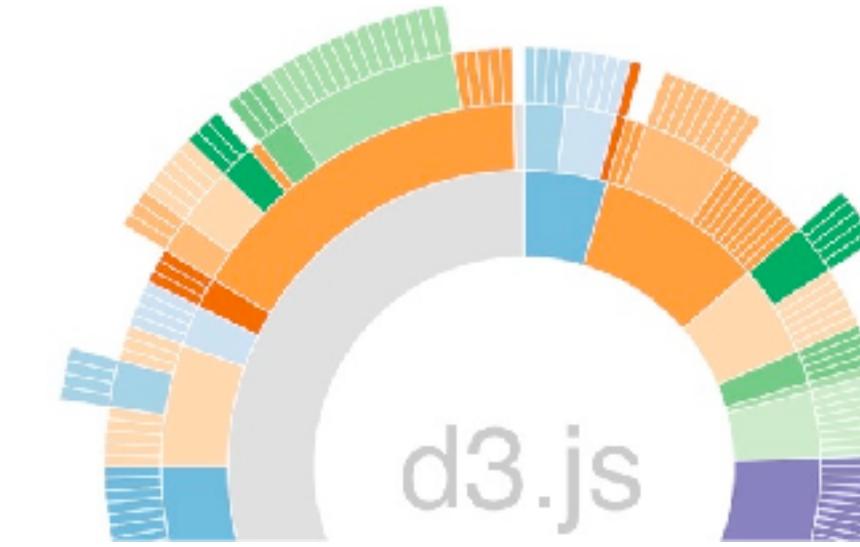
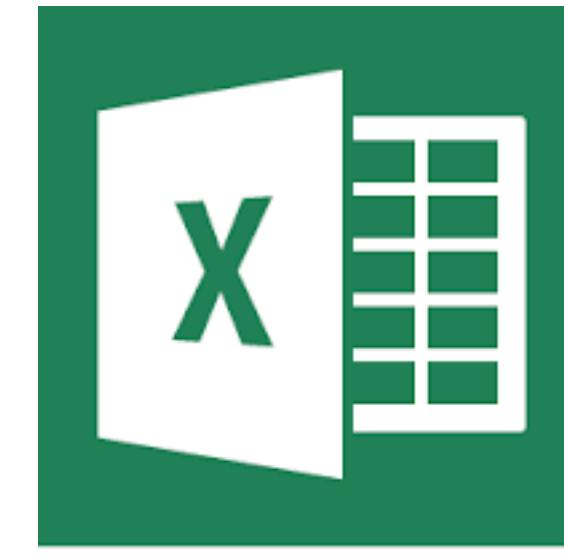
pygal

- › интерактивные графики
- › графики в формате SVGs (не подходит для больших датасетов)
- › симпатичные графики и простой API

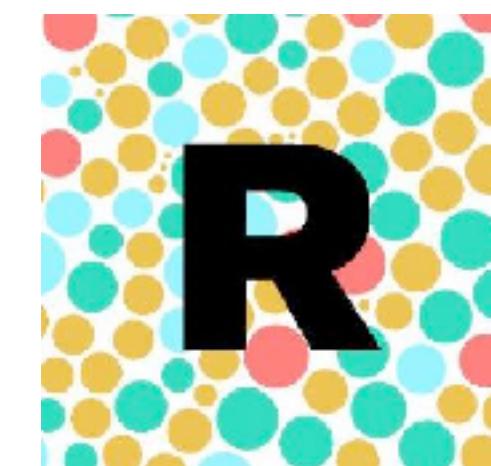


Что еще

- › Excel :)
- › javascript frameworks (самый популярный d3.js, более простой - dimple.js, leaflet.js - для геоданных)
- › online-сервисы (RAWGraphs, Datawrapper)
- › BI системы (Tableau, Power BI)



dimple.js



Dashboard на «коленке»

- › http server (python SimpleHttpServer, nginx, etc.) + static html
- › Flask или Django для более сложных задач, требующих интерактивности



Practice makes perfect :)



Выбор библиотеки визуализации

- › для быстрых графиков - [matplotlib](#)
- › для красивых и интерактивных графиков для менеджеров - [plot.ly](#)
- › [seaborn](#) - pairplot всегда и галерея для вдохновения
- › если вы скучаете по R - [ggplot](#)



Wrapping it up

Сегодня мы

- › познакомились с основными типами визуализаций
- › разобрались, какие есть инструменты и на практике построили графики с помощью библиотек `matplotlib`, `seaborn` и `plotly`

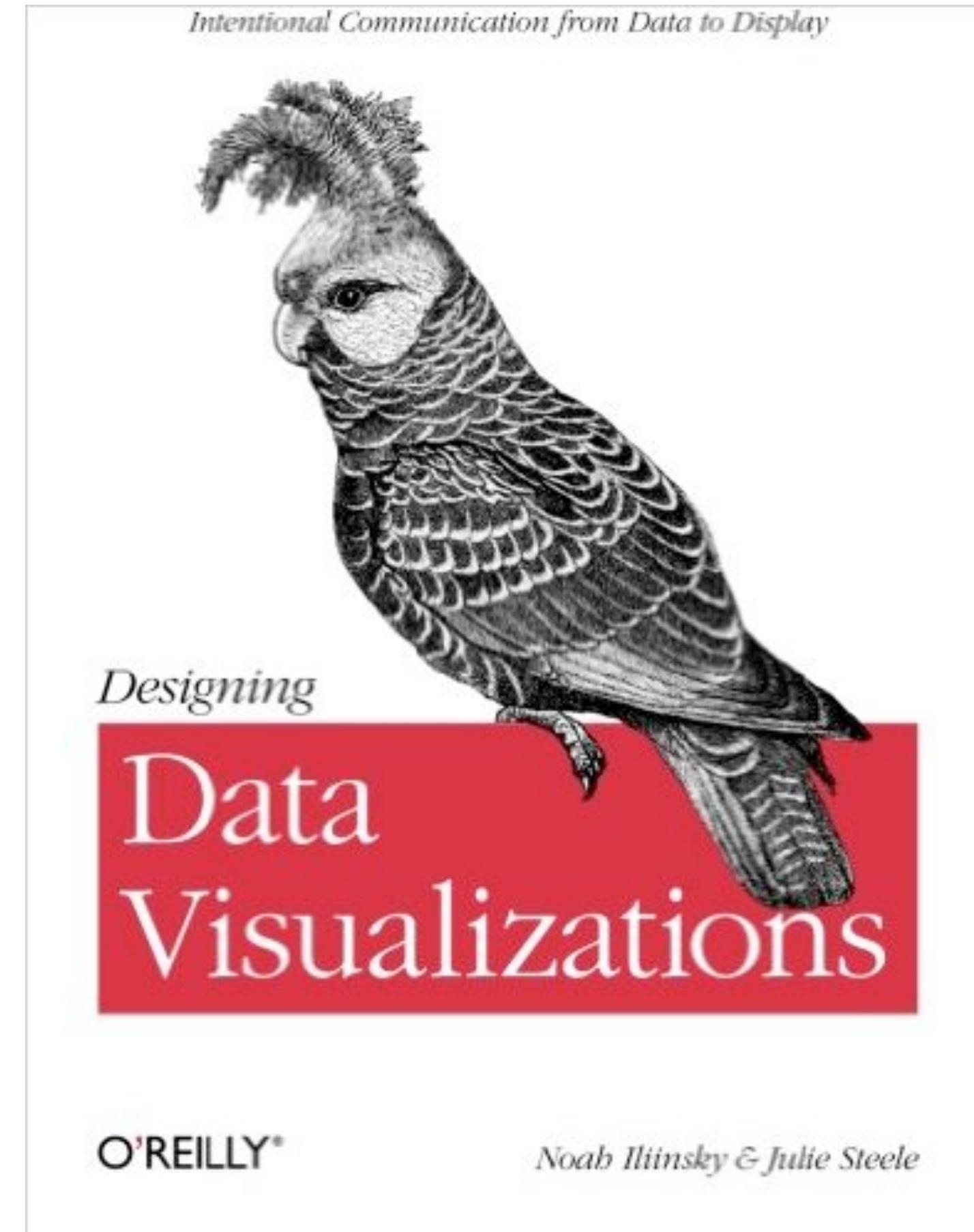


What's next?



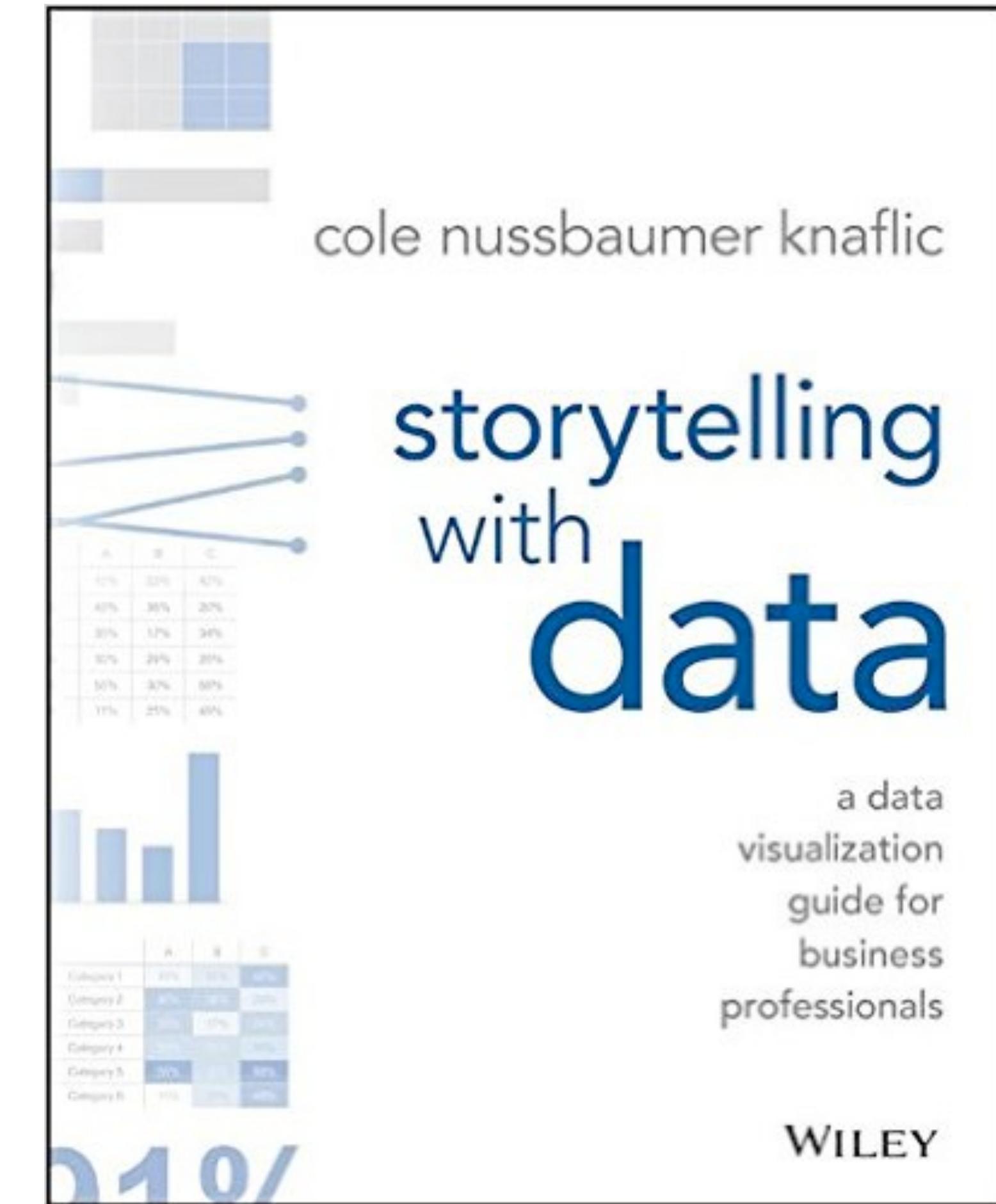
Designing Data Visualizations

- › основные типы визуализаций
- › выбор средств выражения для донесения своих мыслей



Storytelling with Data

- › как сделать из графиков историю



Interactive Data Visualization

- › ОСНОВЫ HTML, JS, SVG, DOM
- › ИСПОЛЬЗОВАНИЕ D3.js

