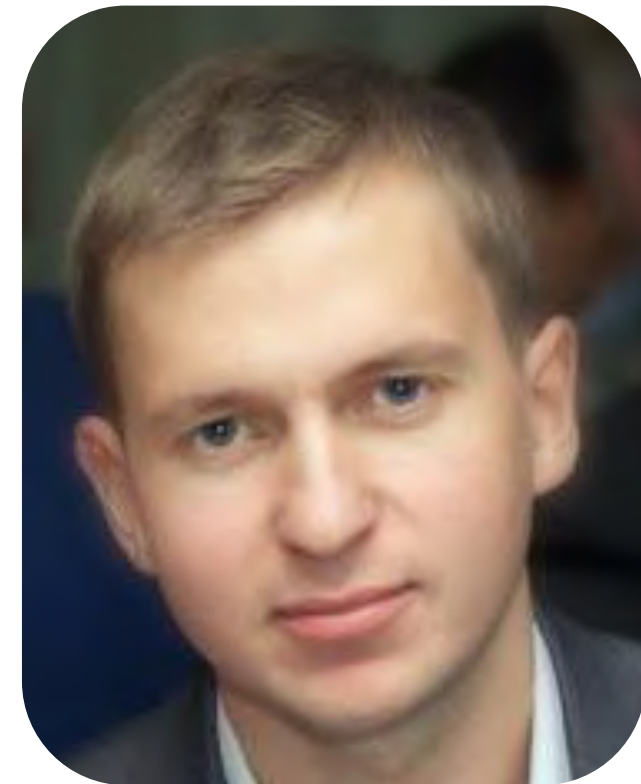




Визуализация данных в Python

Олег Булыгин

Нетология



Булыгин Олег

- Преподаватель курсов по Python в Нетологии
- Начальник бюро планирования и управления в АО “НПО автоматики”



o.bulygin@netology.ru



<https://www.facebook.com/obulygin91>

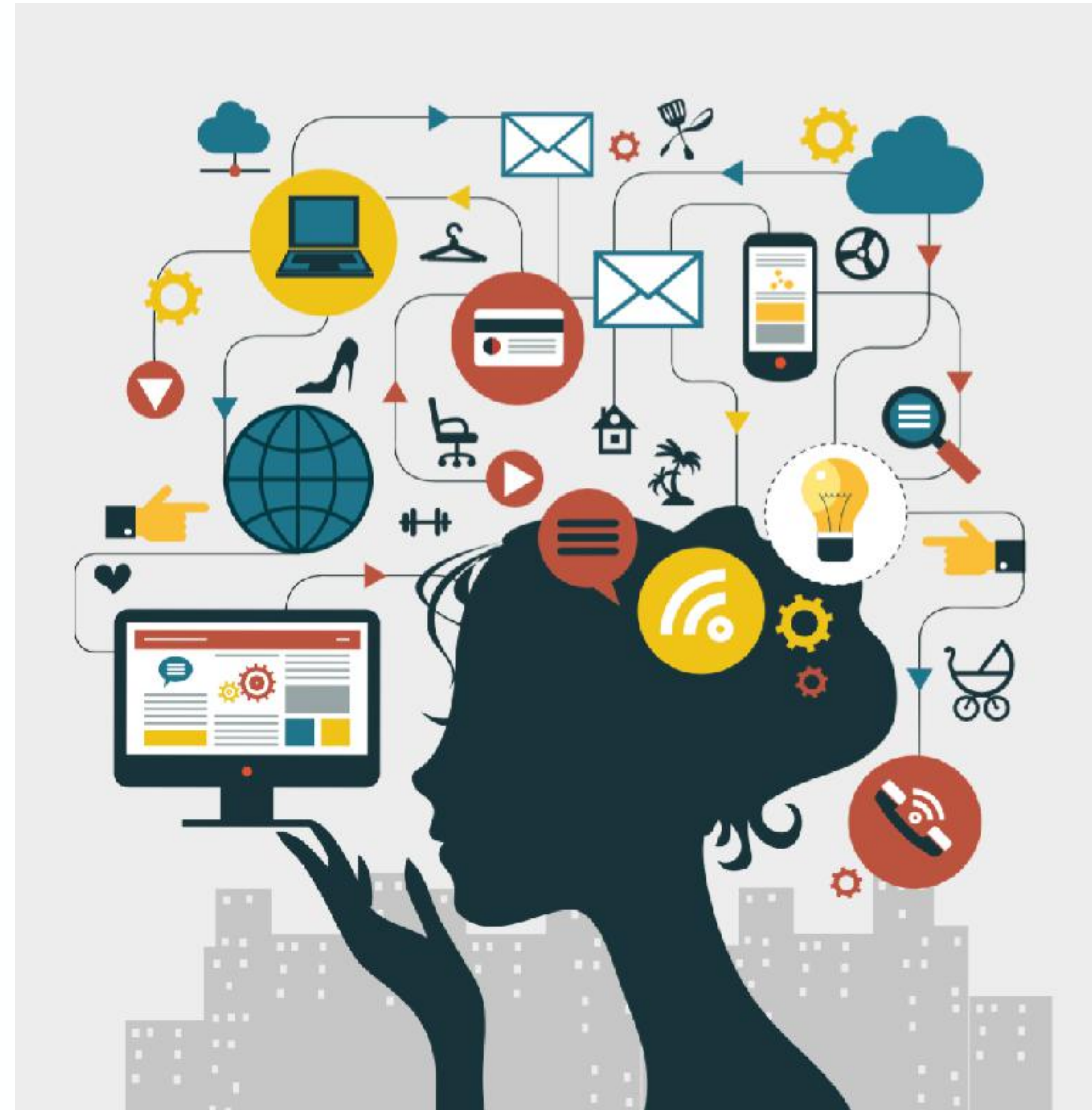
Цель занятия

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



План

1. Что такое визуализация и зачем она нужна?
2. Теория визуализации: visual encodings, типы графиков и задачи визуализации
3. Инструменты

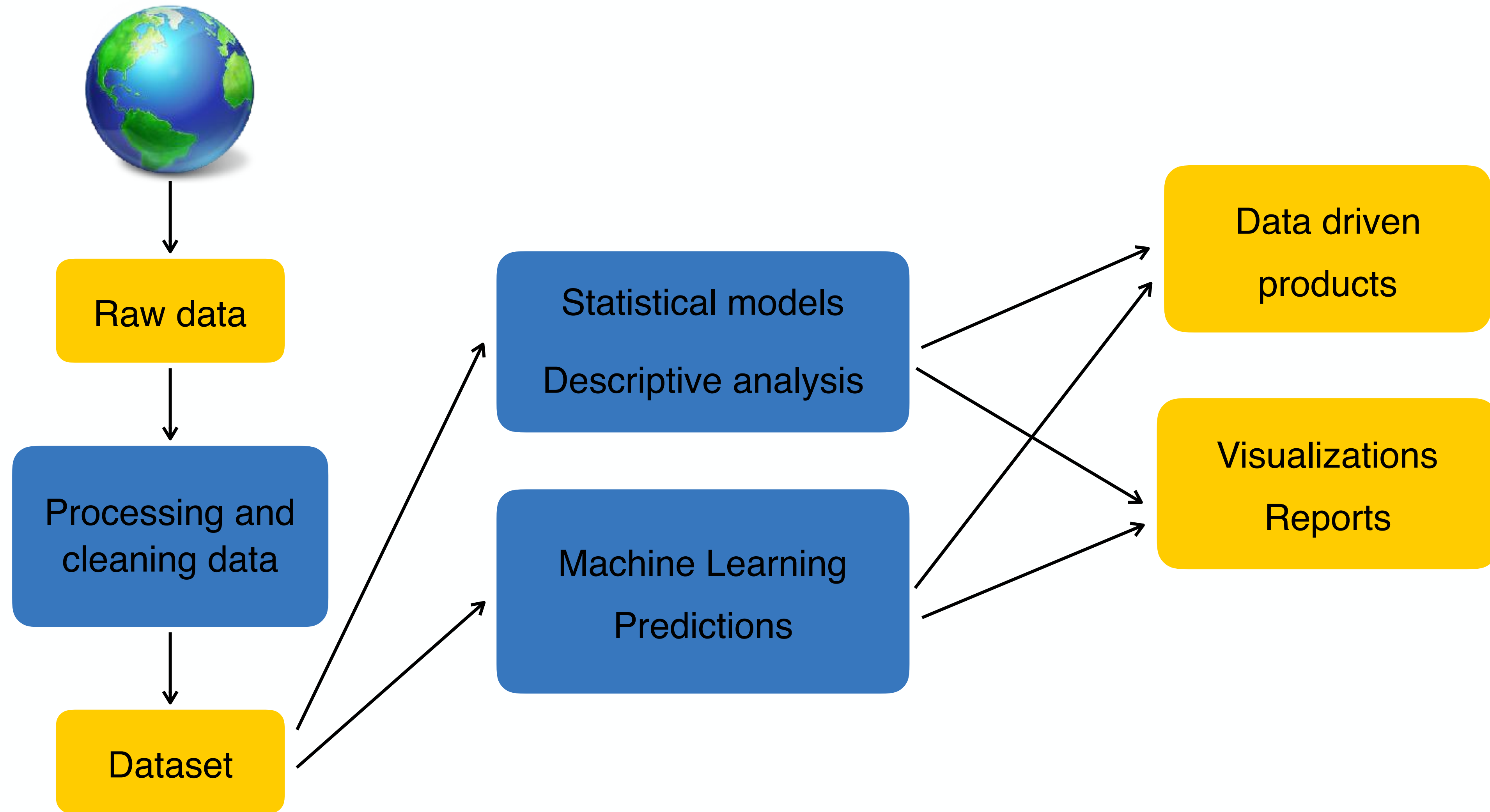


Что такое визуализация данных

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 выборки

все статистики 4х выборок
ОДИНАКОВЫ

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

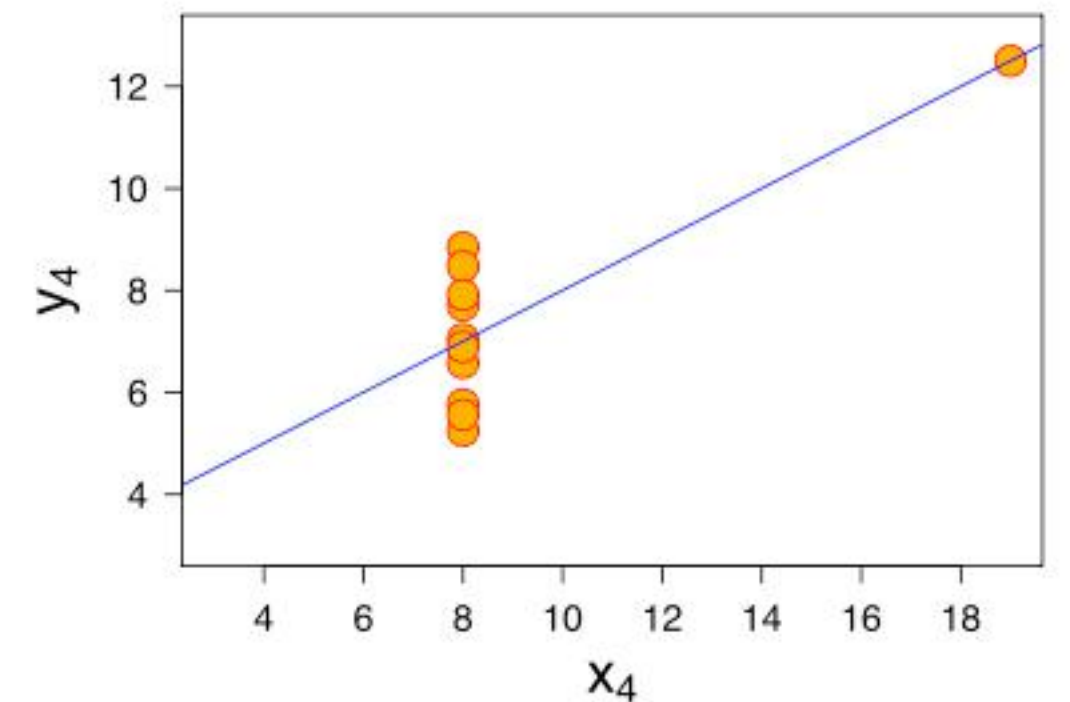
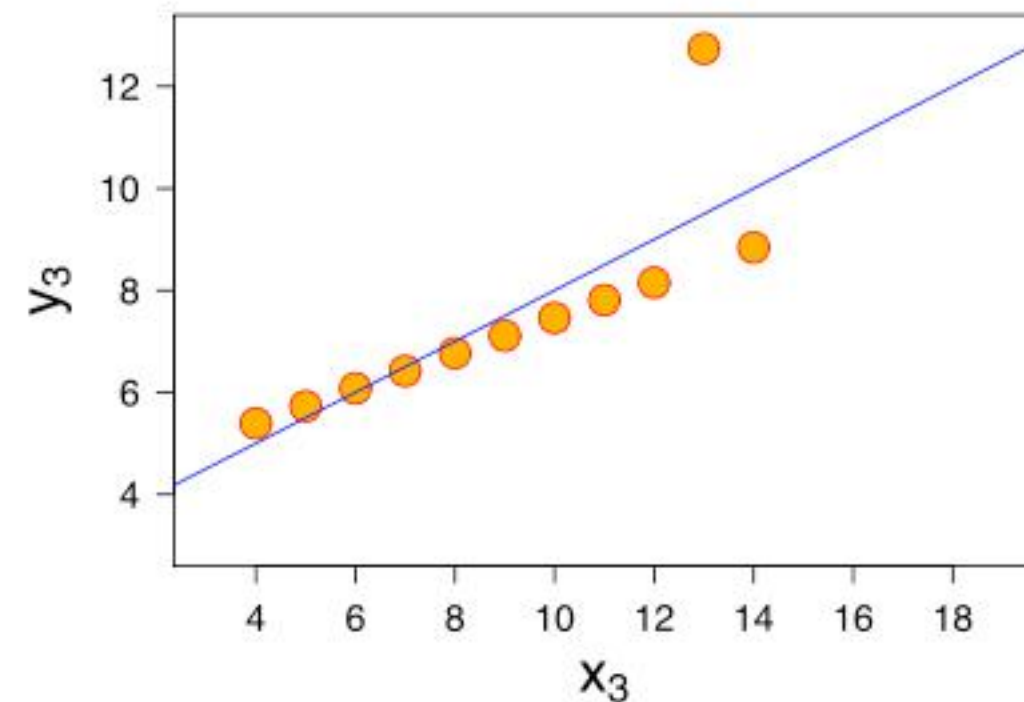
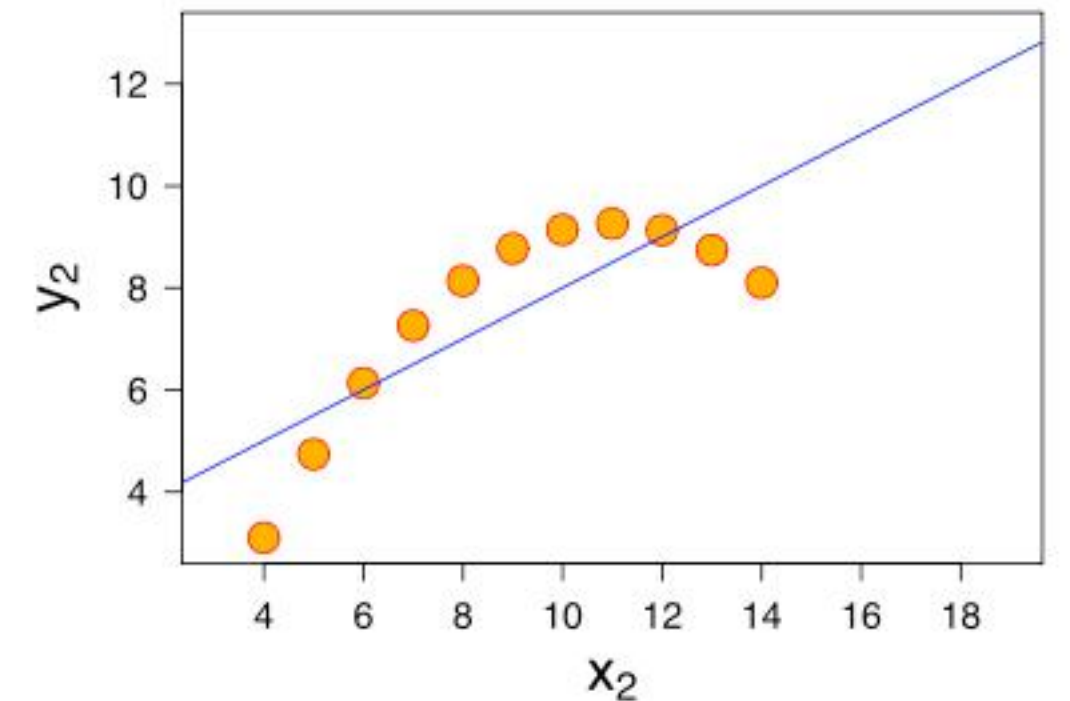
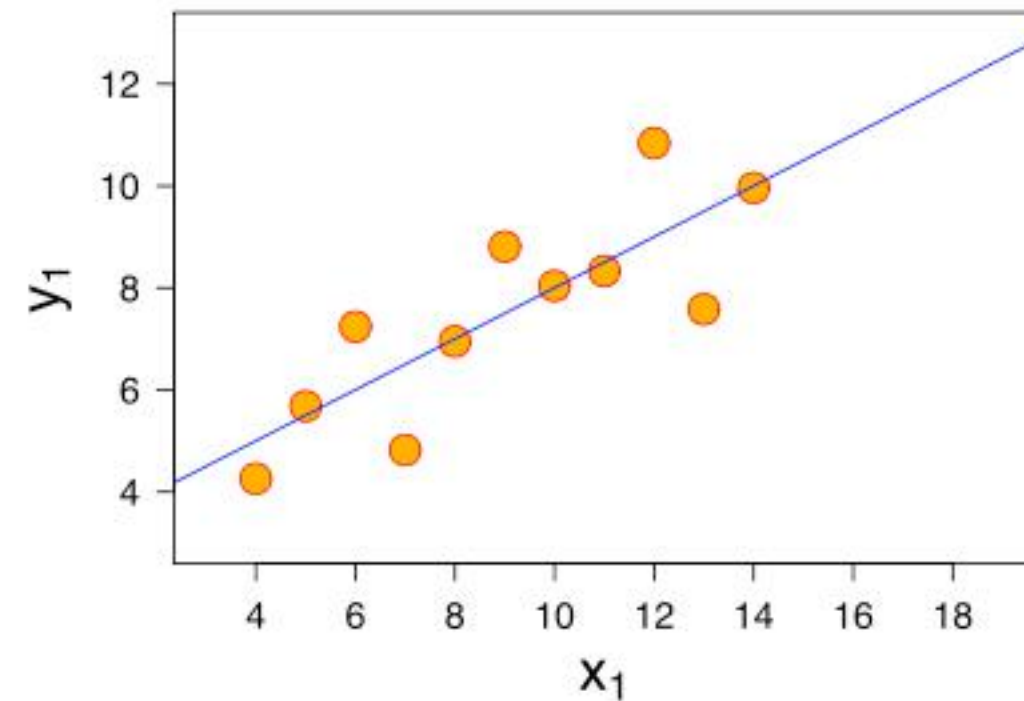
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0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

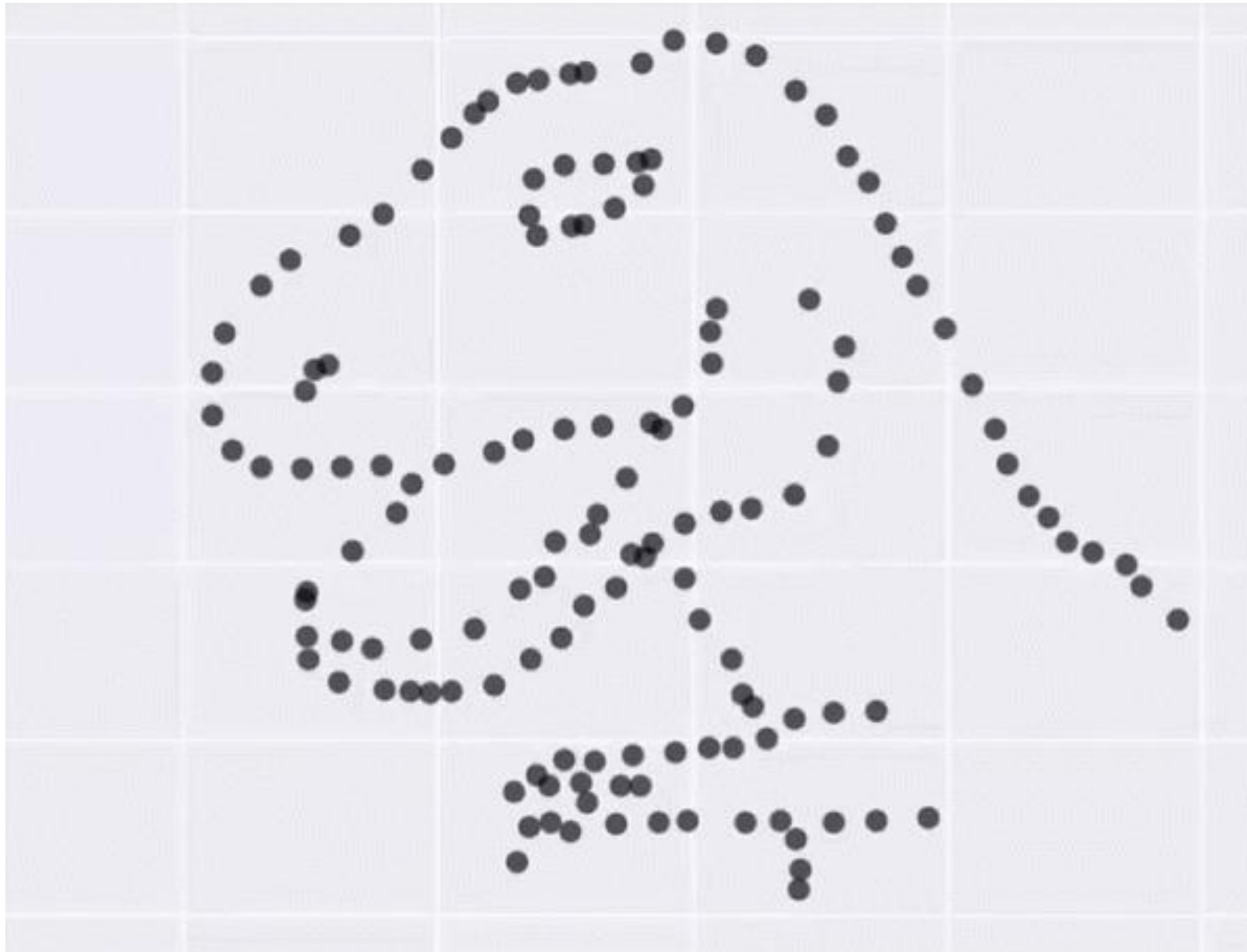
Квартет Энскомба

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)

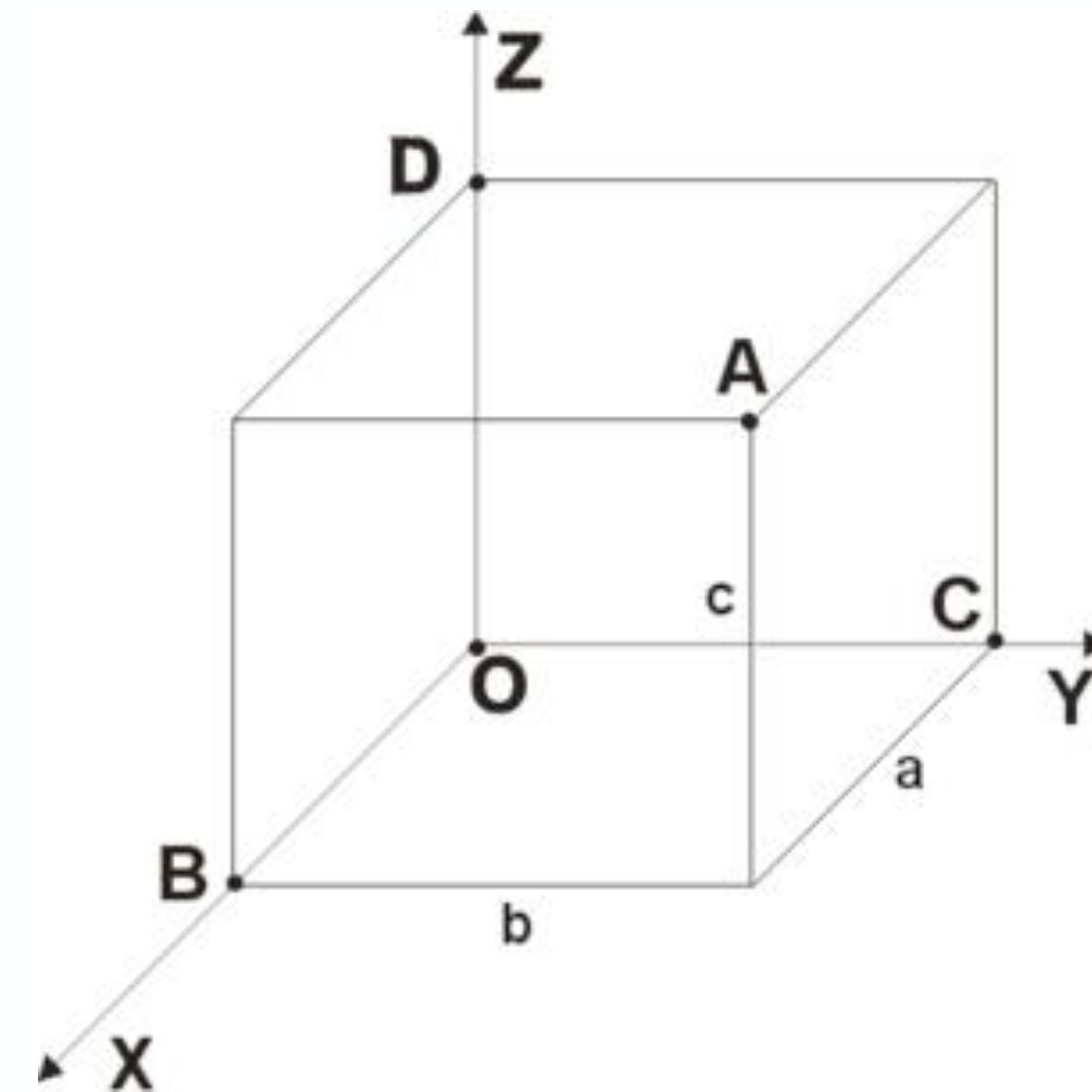
данные -> отображение их на графике

- › 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



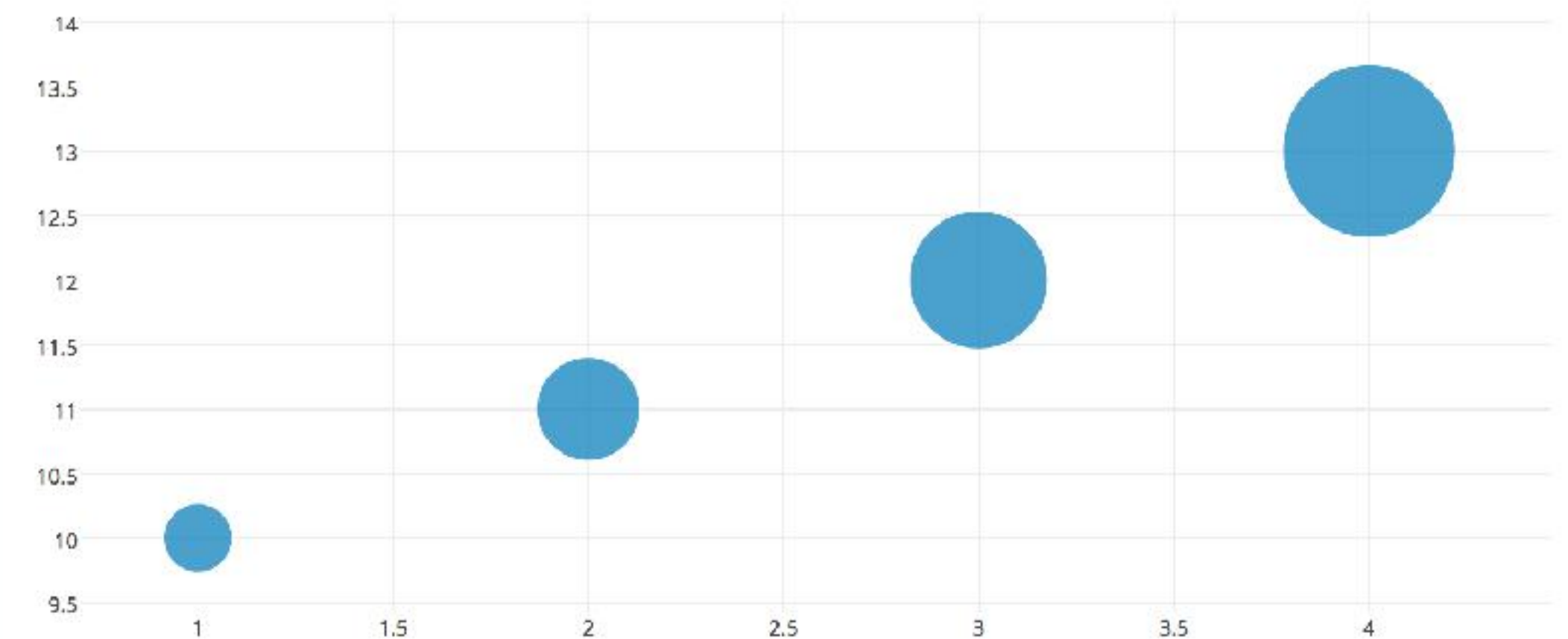
Размер (длина, площадь, объем)

длина

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

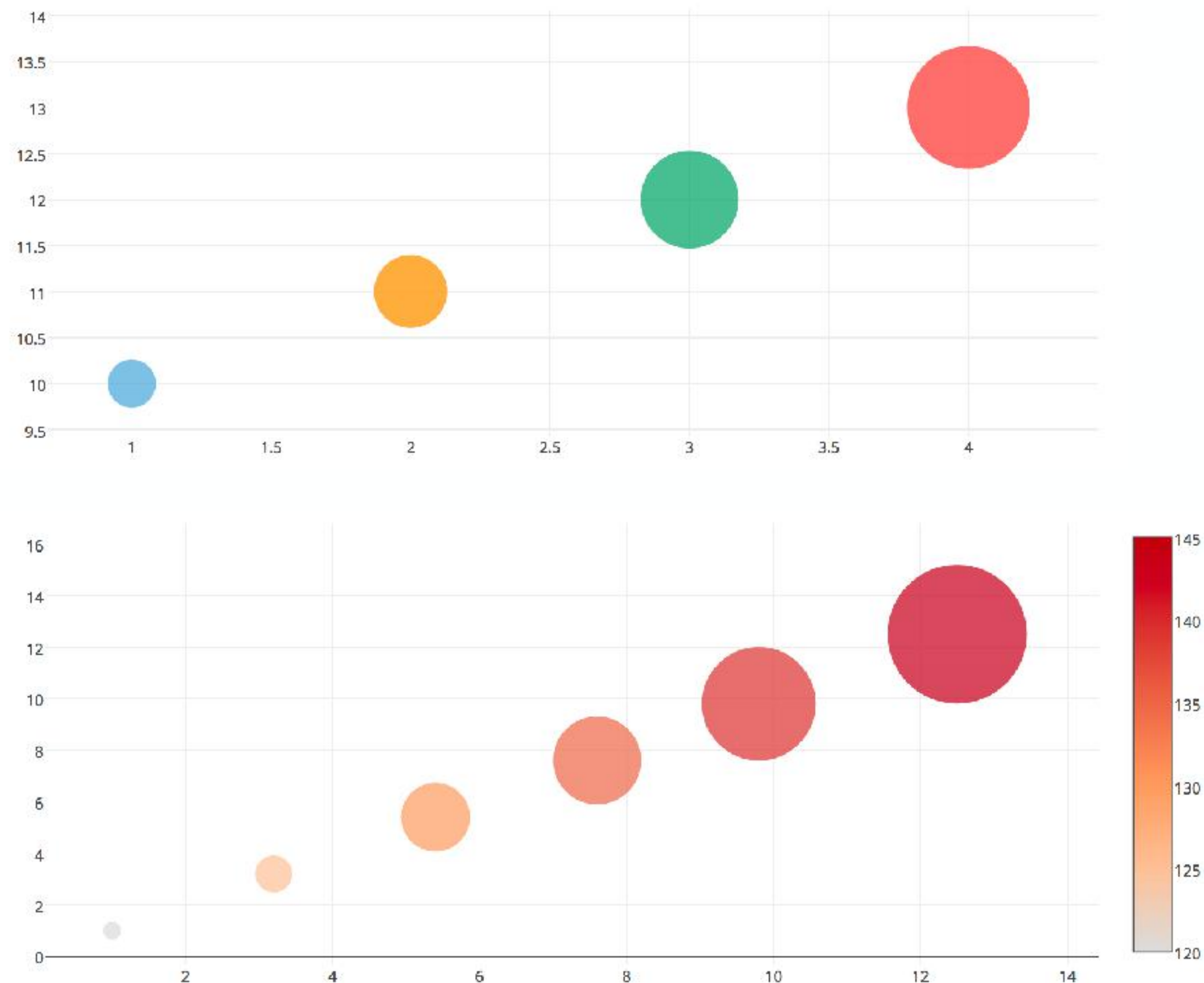
площадь, объем

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



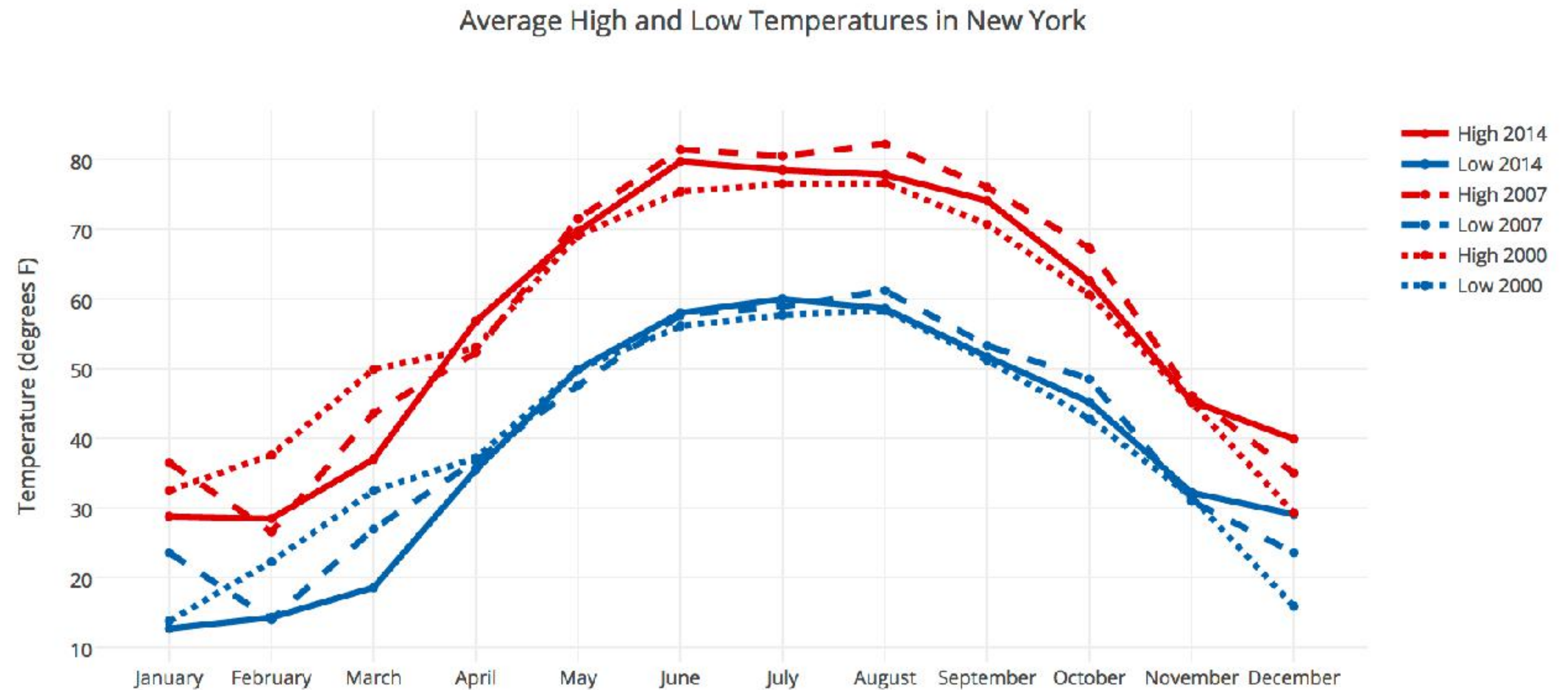
Цвет (hue/saturation)

- › 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

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

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

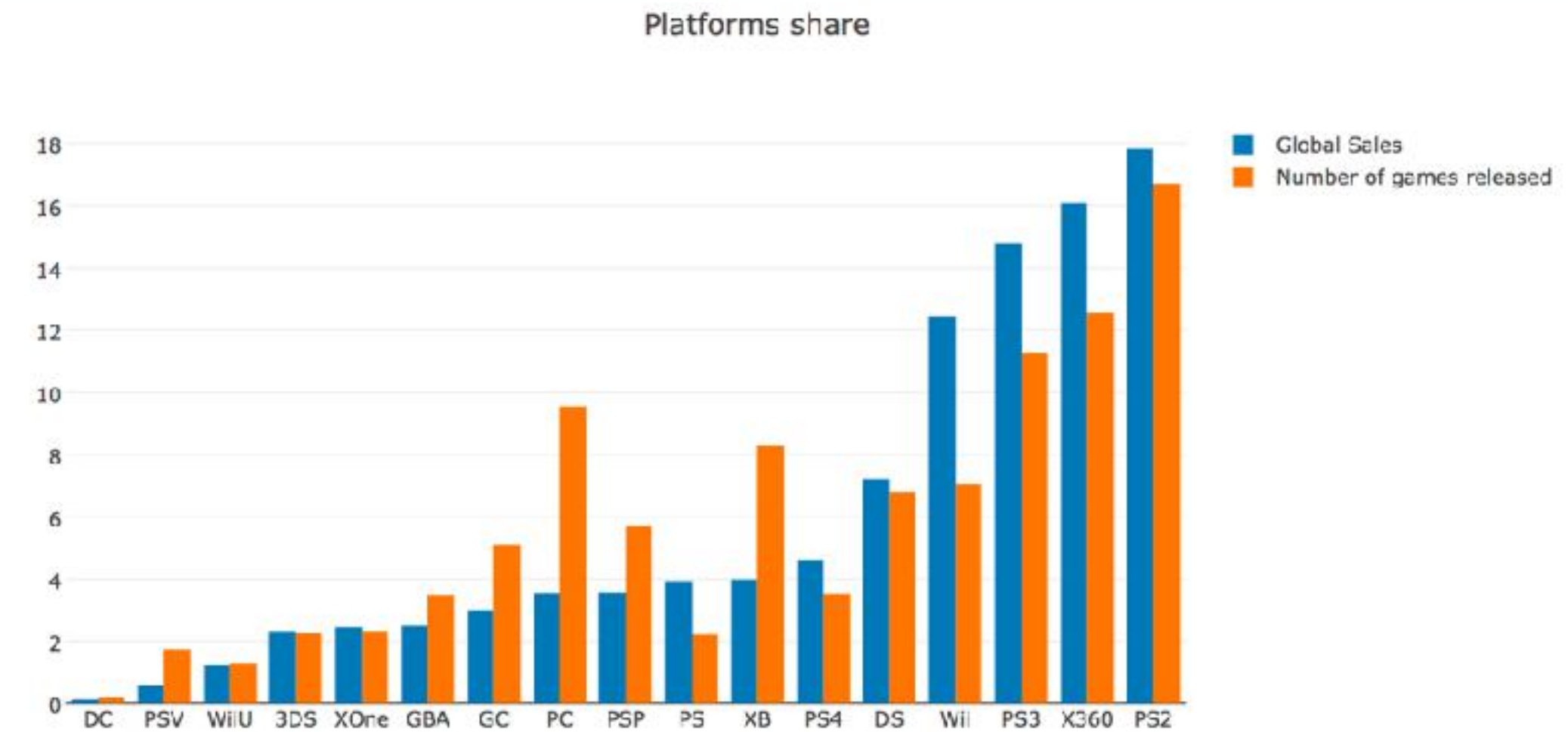
Обычное сравнение (Nominal comparison)

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

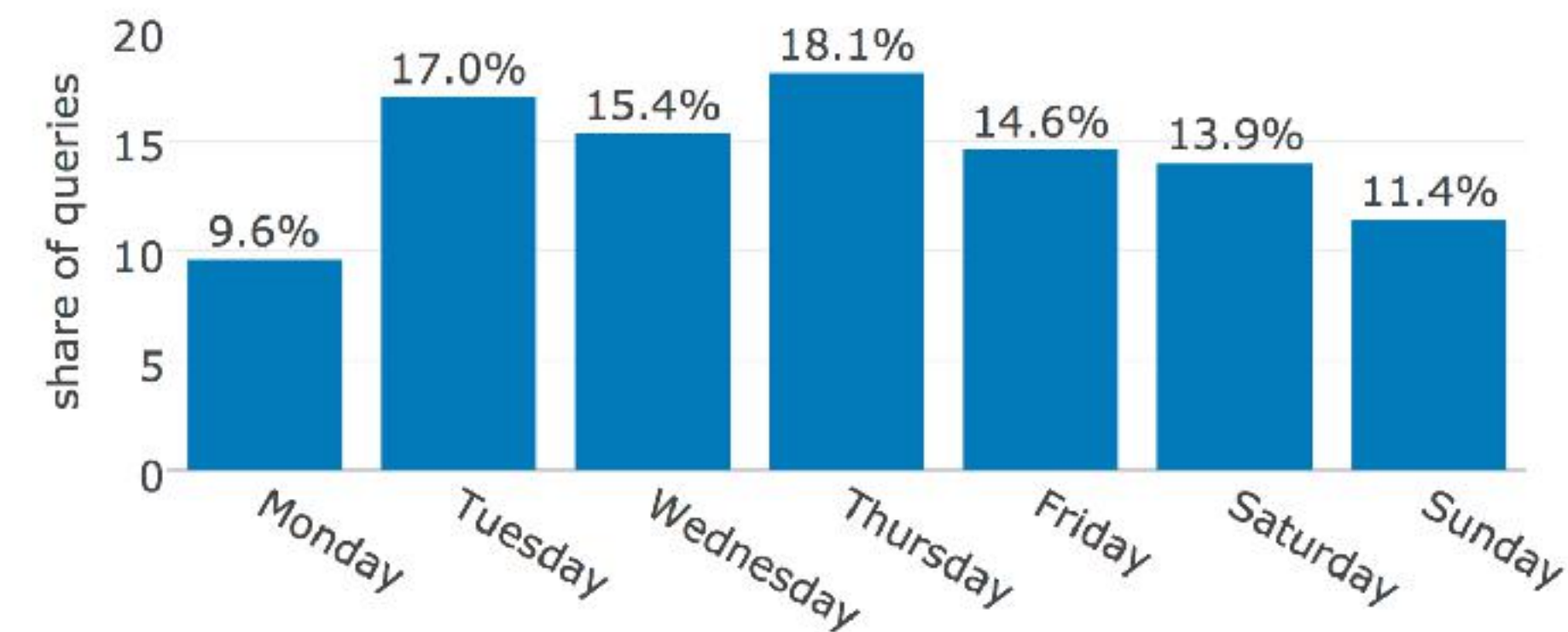
Обычное сравнение (Nominal comparison)

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

› Горизонтальный или вертикальный bar chart



Запросы по дням недели

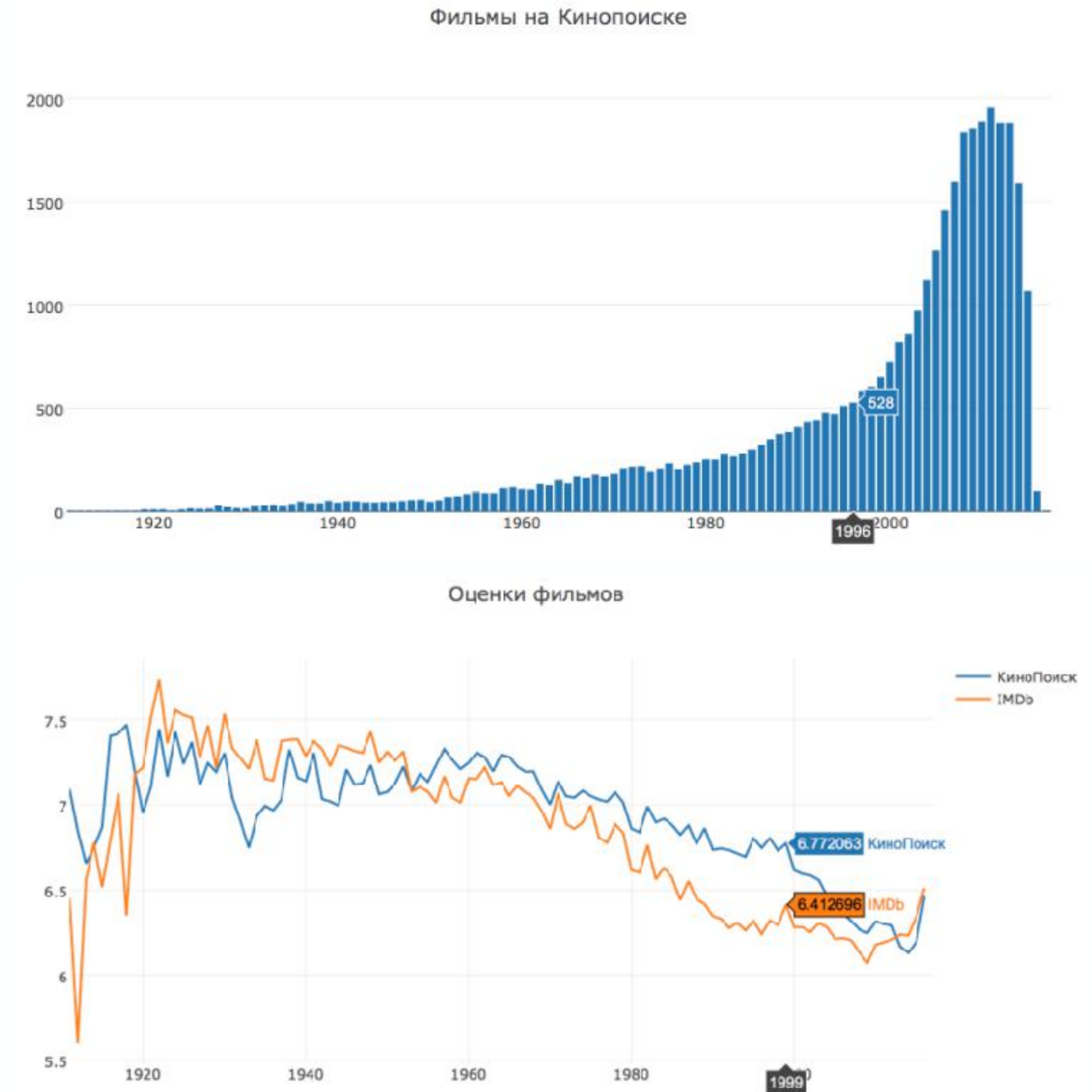


Time Series

- | 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 значения
 - › Временная переменная должна располагаться на оси X

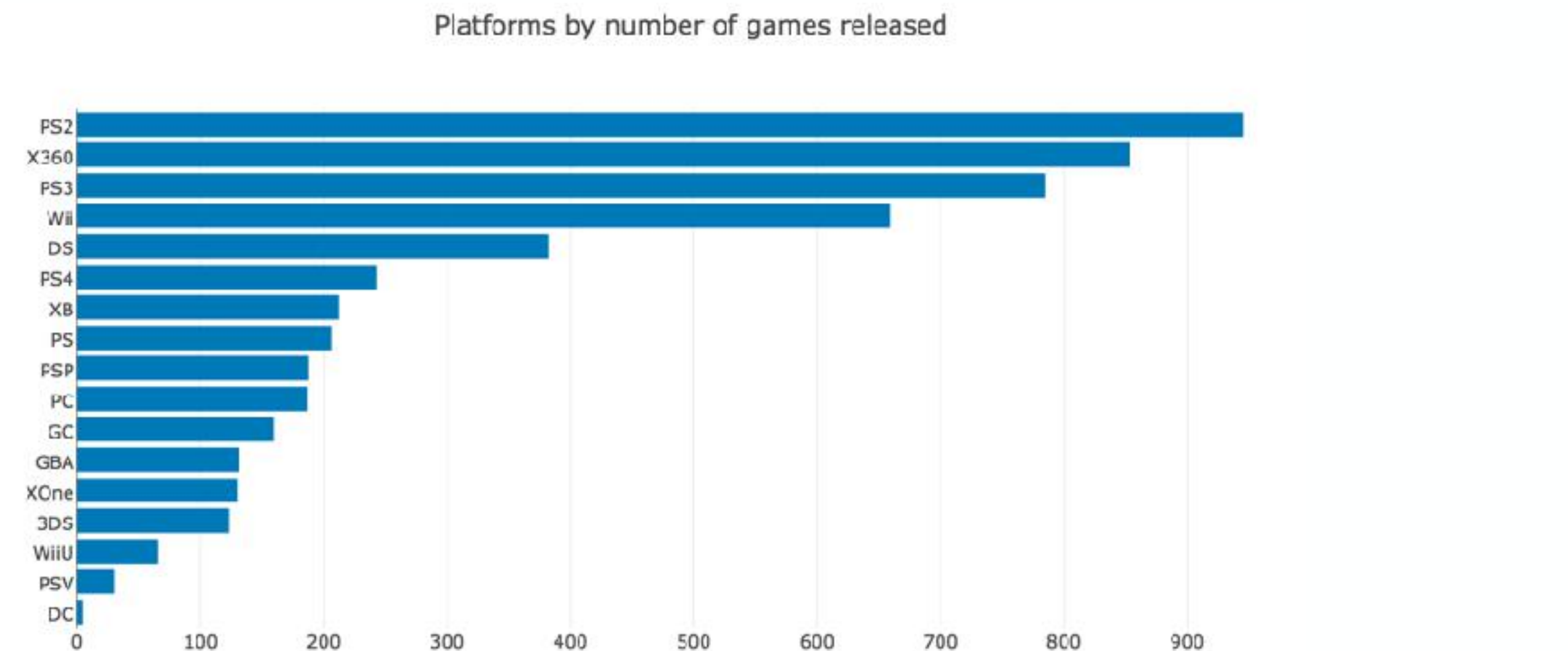


Ranking

- | 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



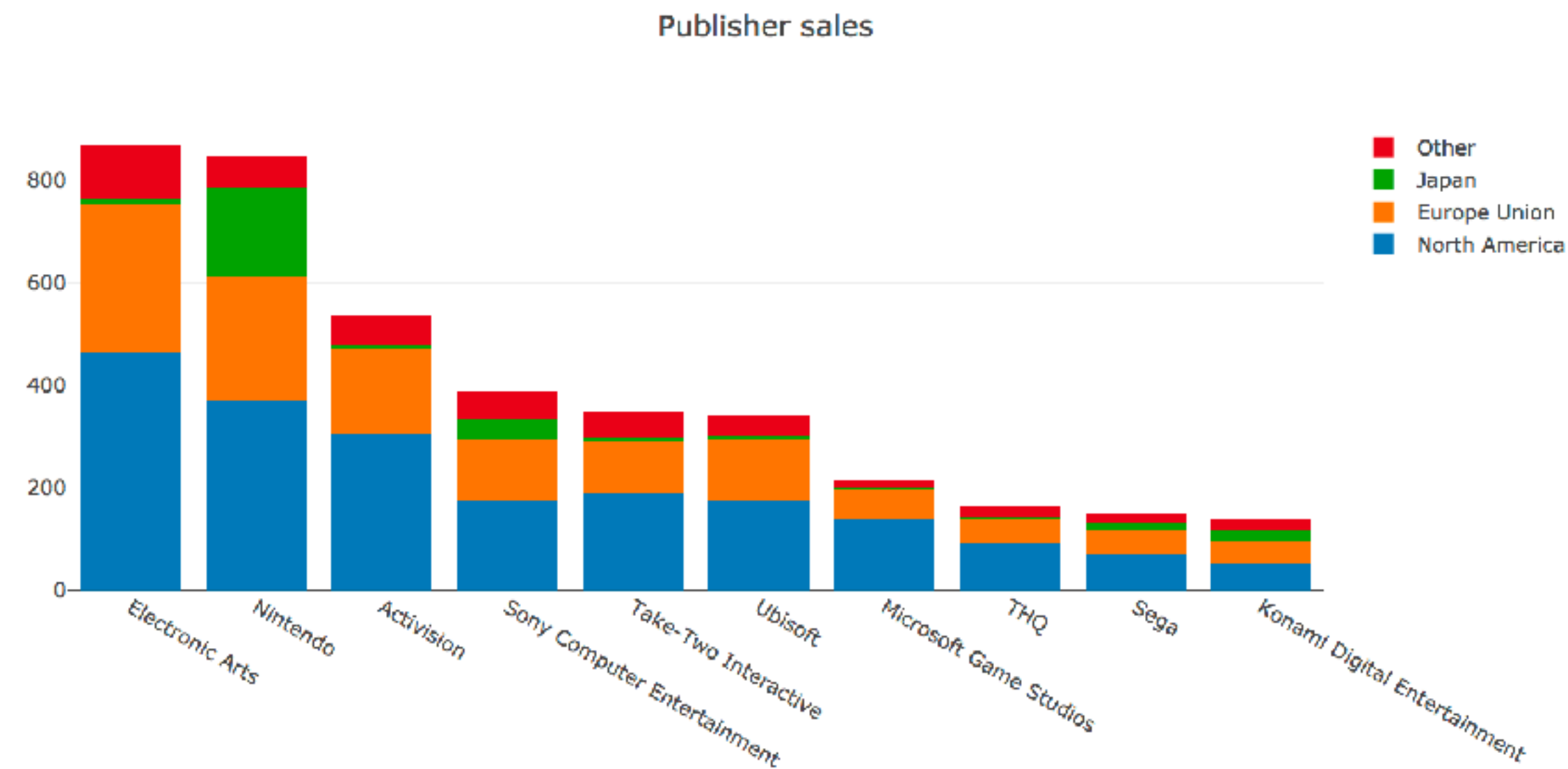
Part-to-hole

- | 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

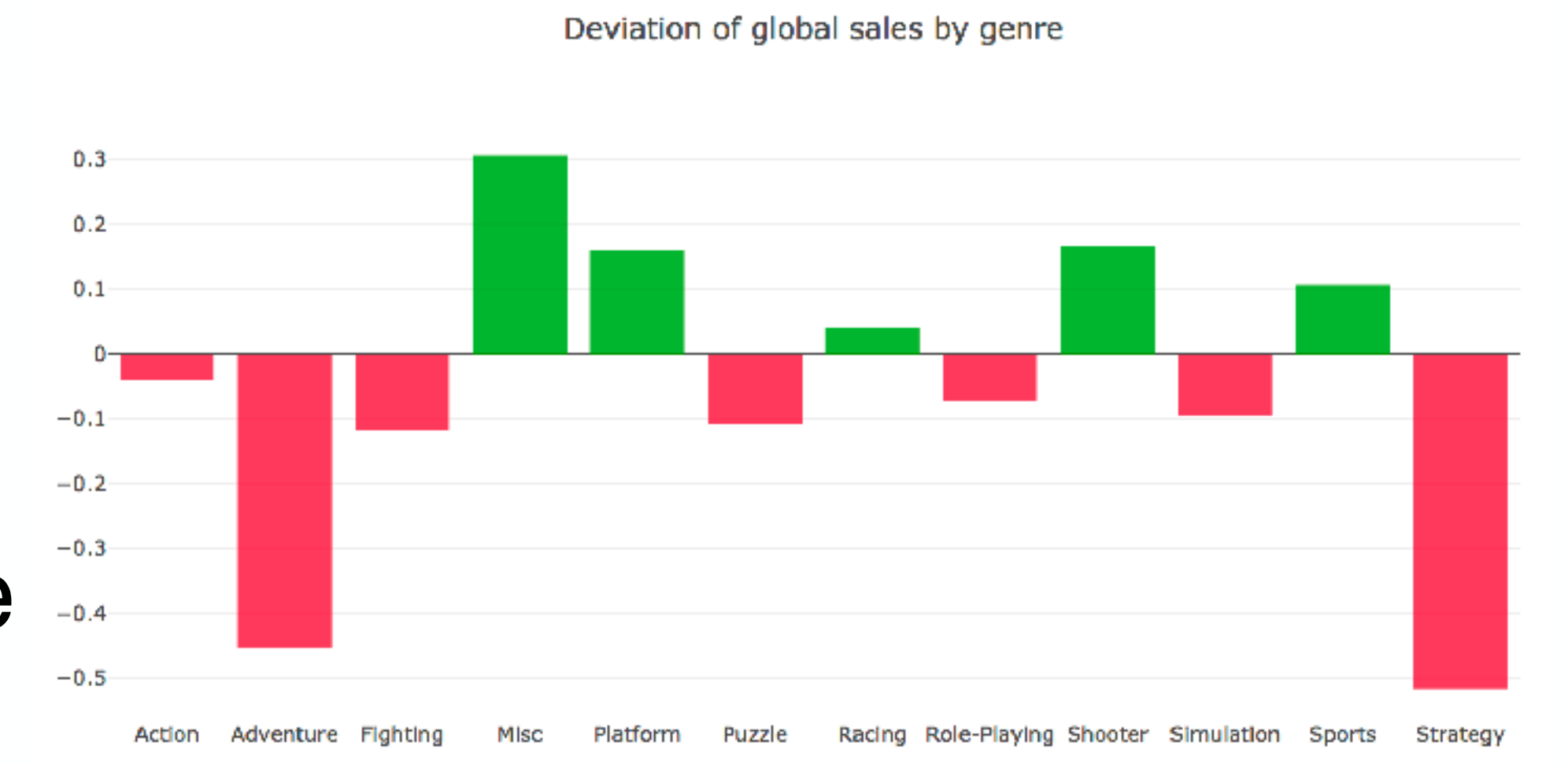


Deviation

- | 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 значение

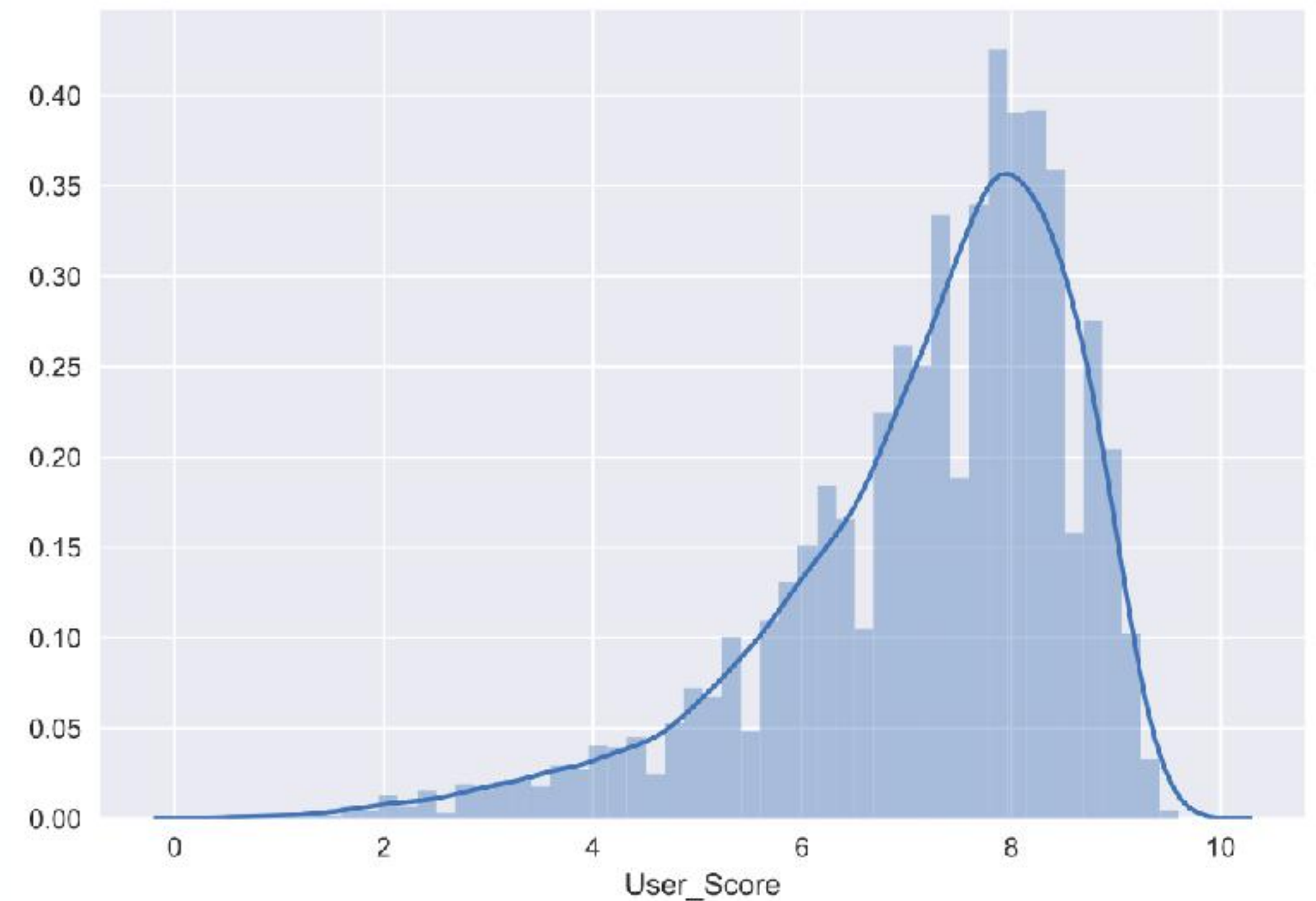


Frequency Distribution

- | 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 - корреляция между двумя численными величинами
- | Задача - показать, как связаны между собой оценки пользователей и критиков

Correlation

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

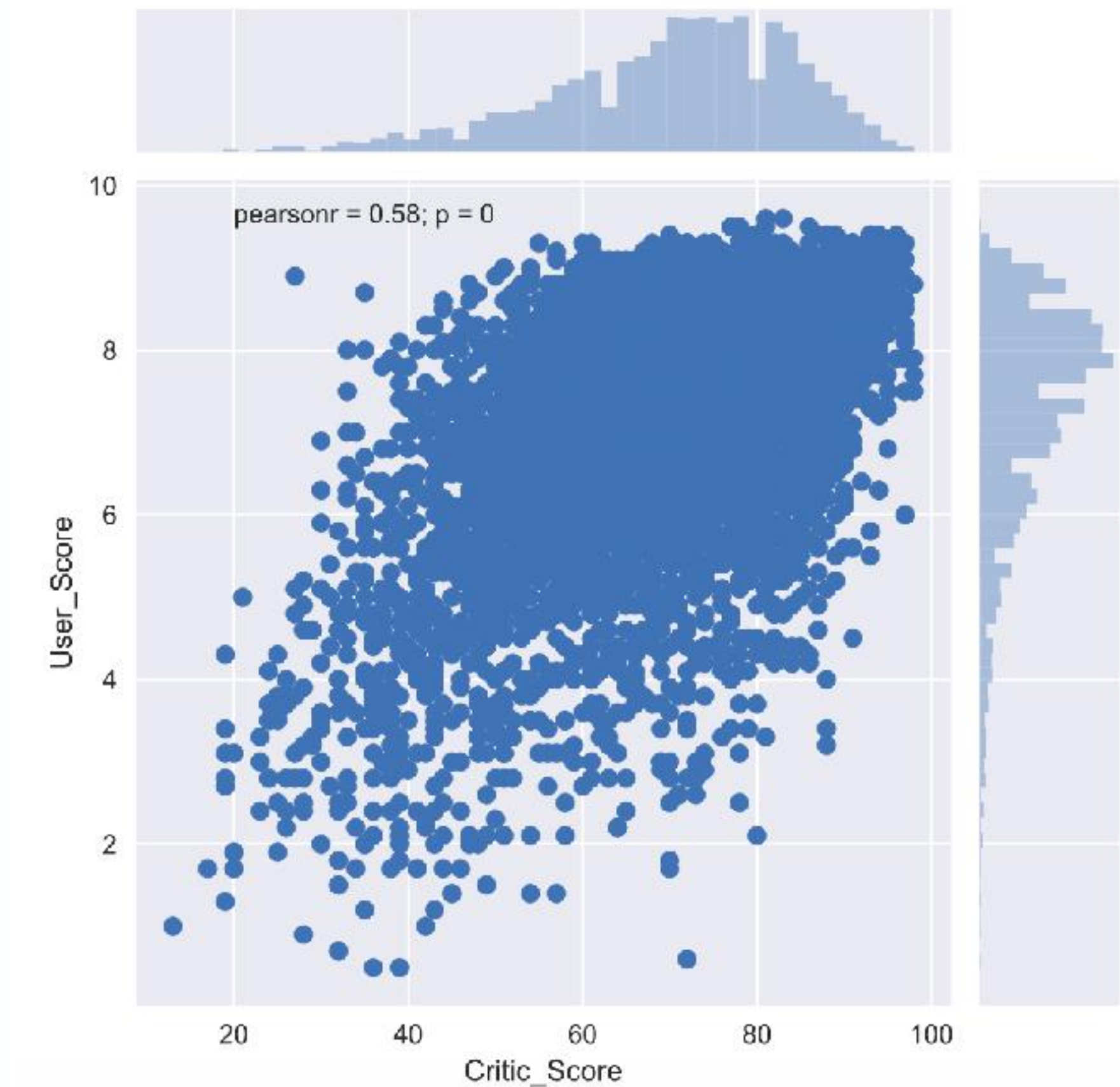
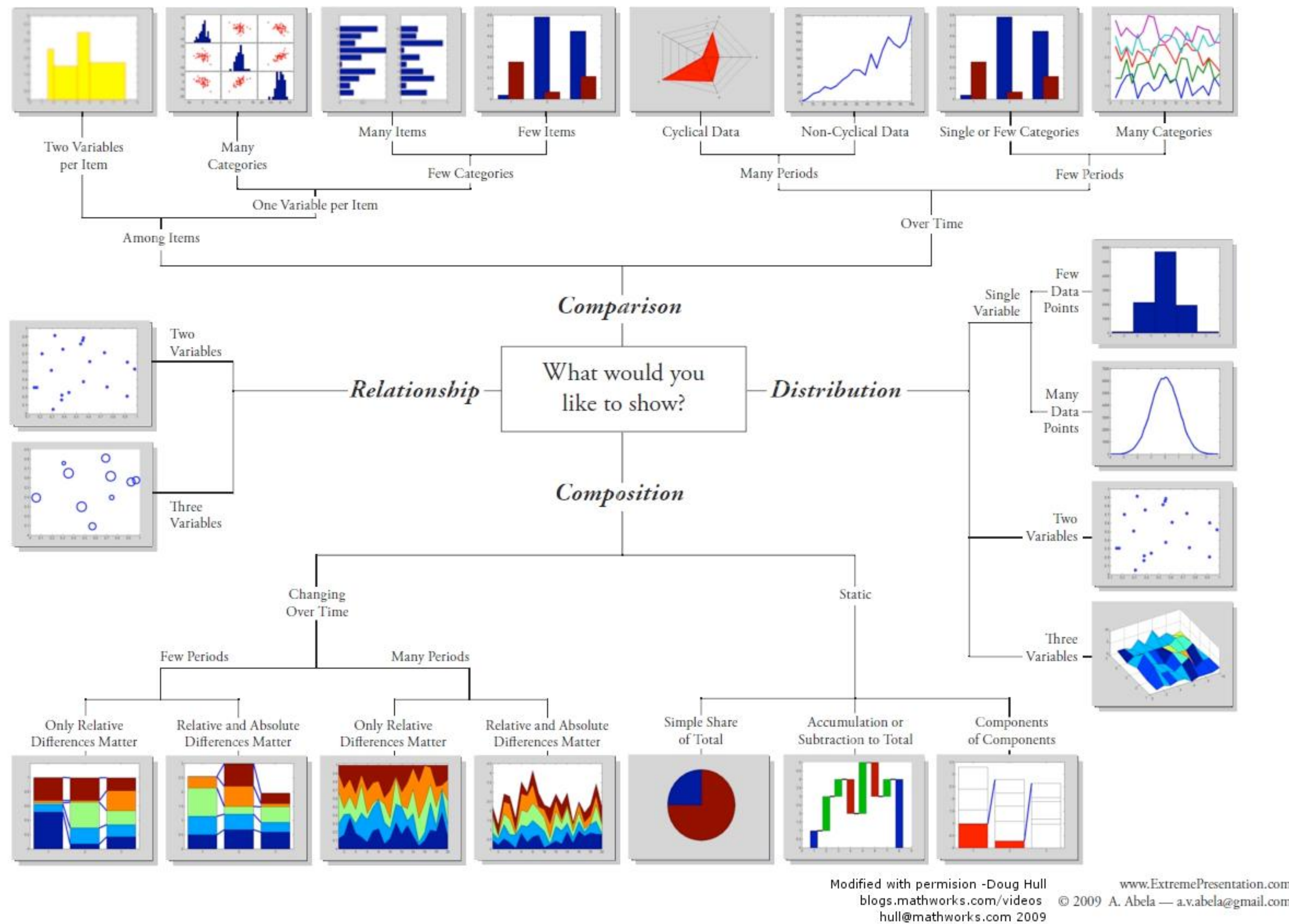






































Chart Suggestions—A Thought-Starter

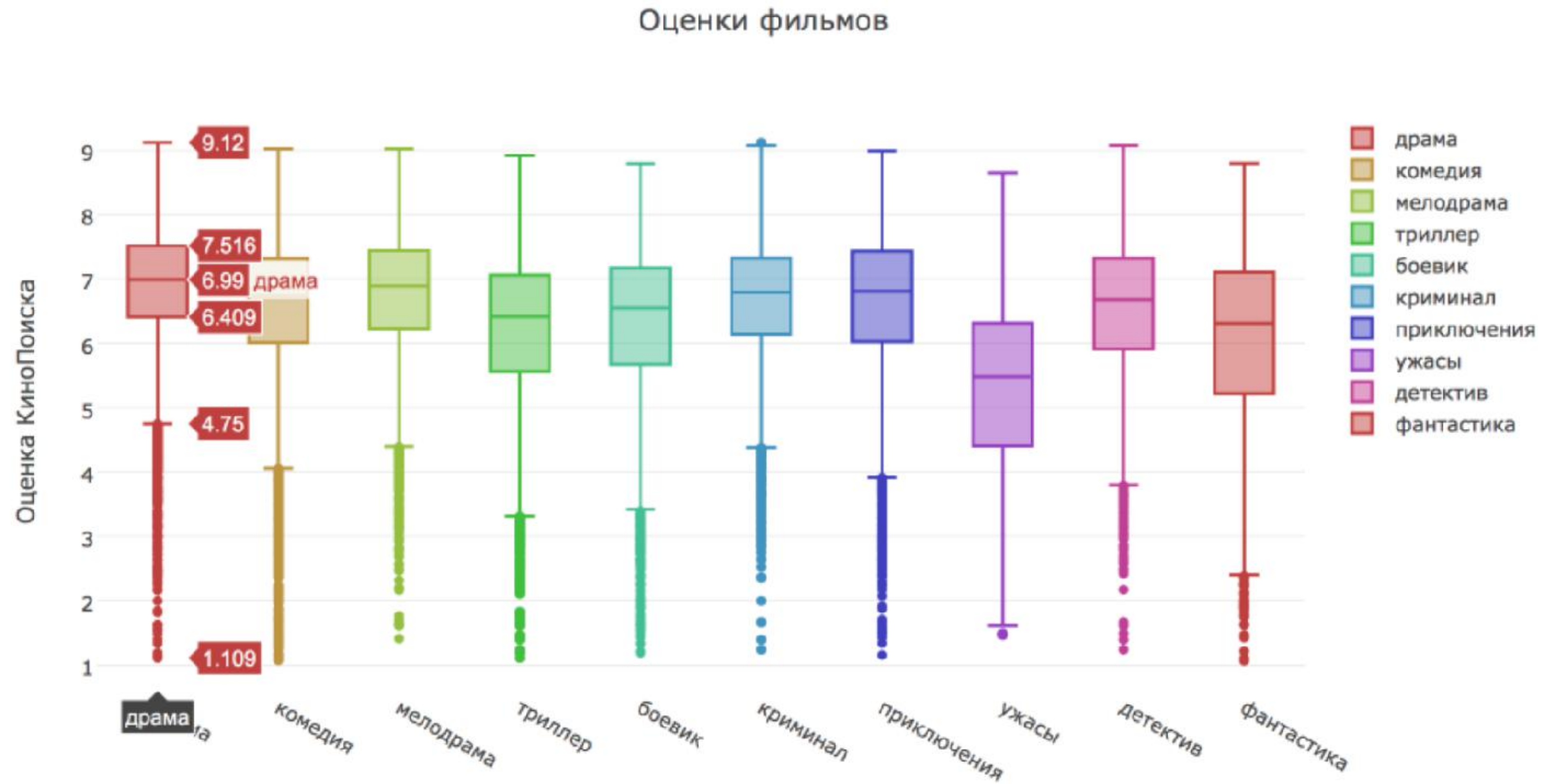


| Есть и другие
визуализации...

Table

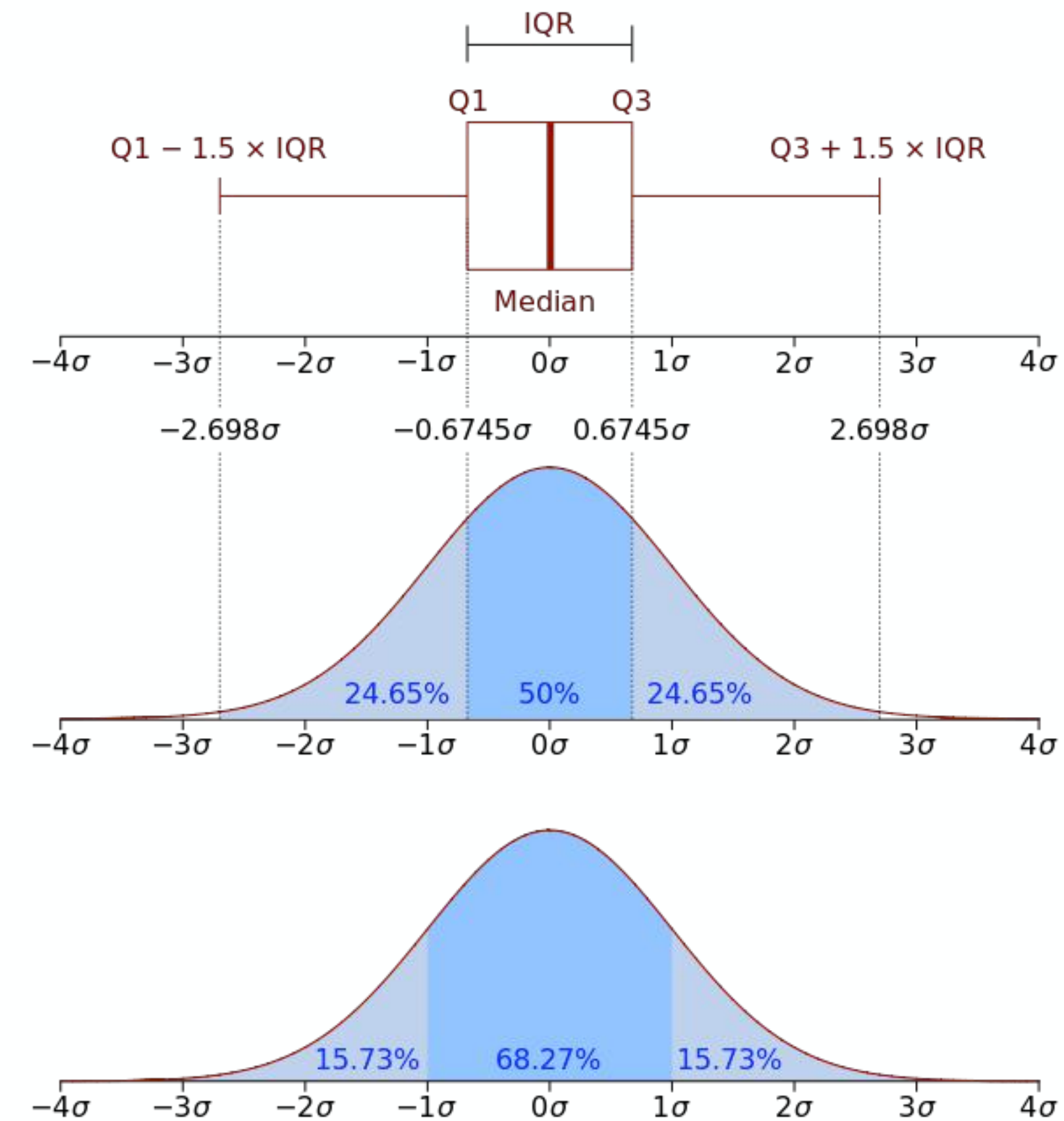
Region	Actual Sales (mn)		% to Goal	(12 Month)	Gross Profit (mn)	Profit Trend (12 Month)
Alabama	\$4,916		107%		\$1,172	
Alaska	\$3,110		65%		\$791	
Arizona	\$5,198		103%		-\$282	
Idaho	\$5,280		101%		\$410	
Illonois	\$4,956		93%		-\$22	
Indiana	\$5,032		91%		-\$516	
Ohio	\$5,566		112%		\$524	
Oklahoma	\$4,246		85%		\$787	
Oregon	\$6,408		102%		-\$932	
Vermonut	\$4,244		73%		\$1,495	
Virginia	\$7,664		161%		\$325	
Washington	\$4,558		88%		\$1,829	

Box plot

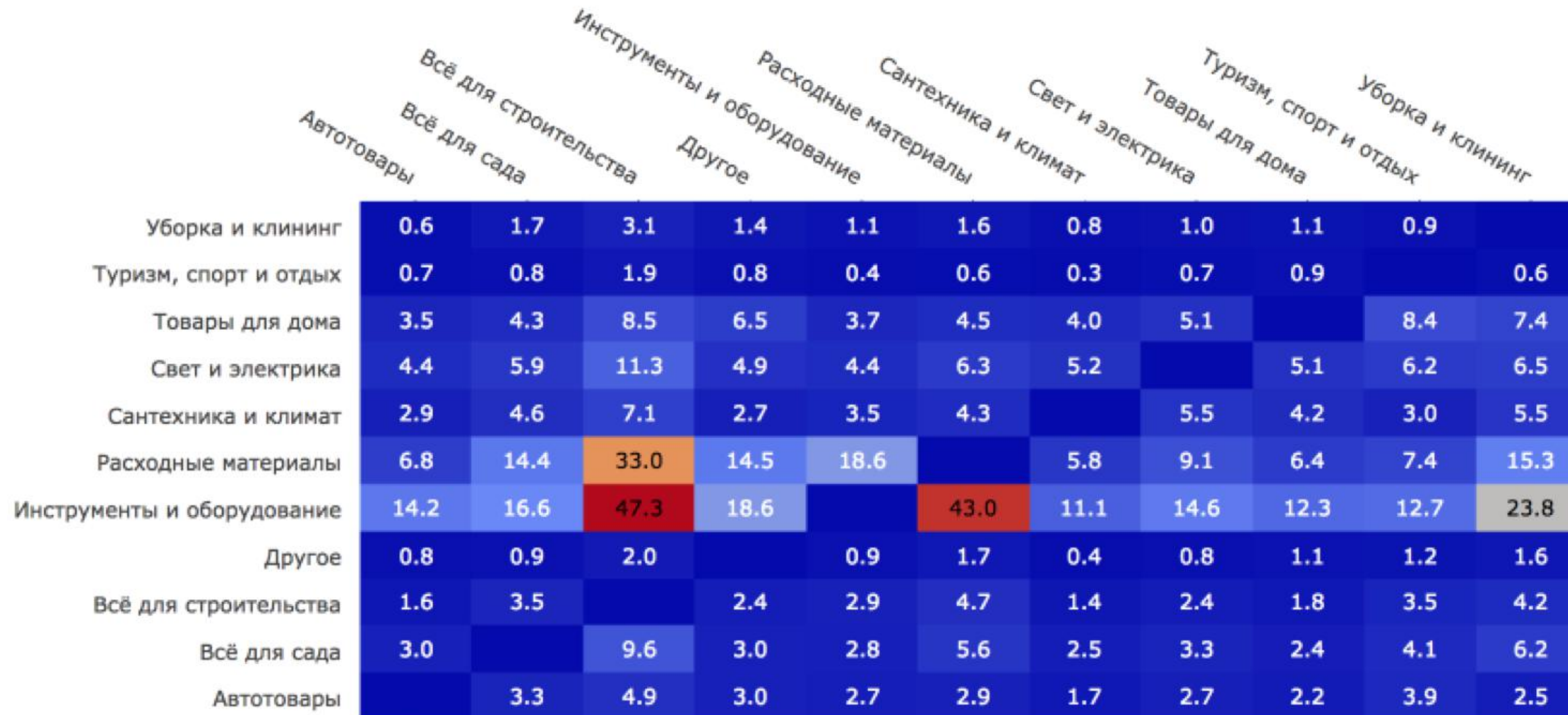


Box plot uncovered

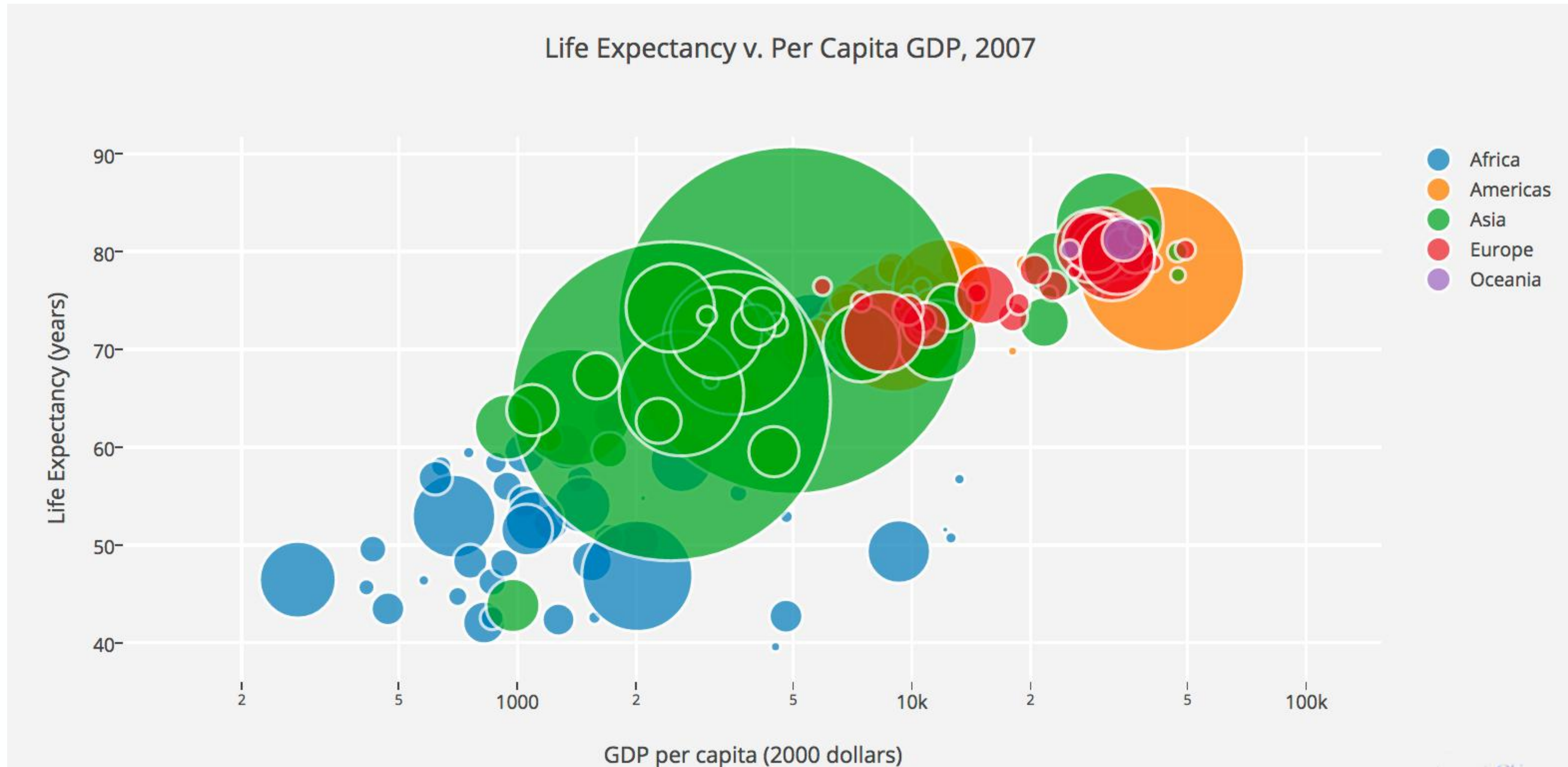
- › линия - медиана
- › коробка - IQR
- › усы - $[Q1 - 1.5IQR, Q3 + 1.5IQR]$
- › точки - outliers




Heatmap

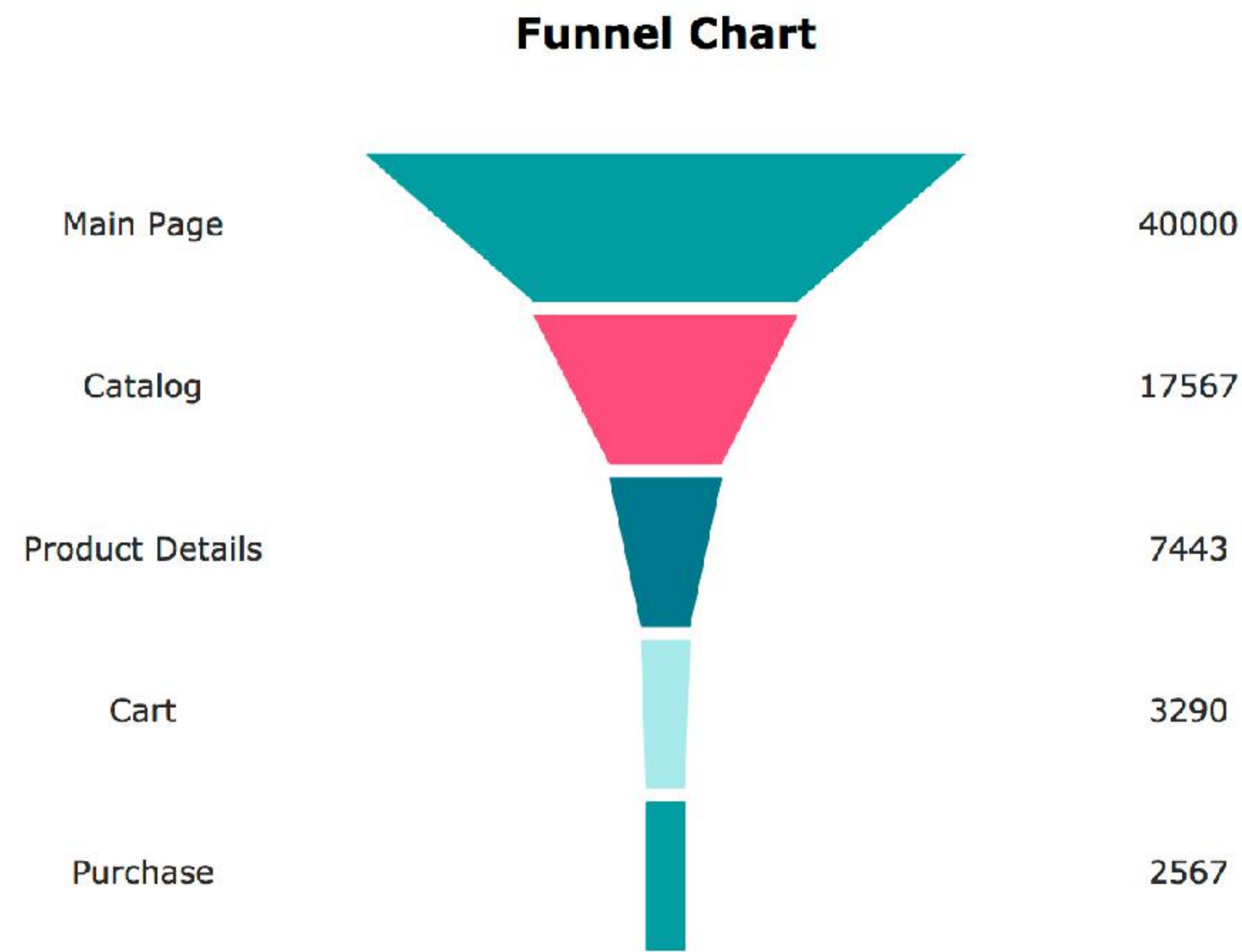


Bubble chart

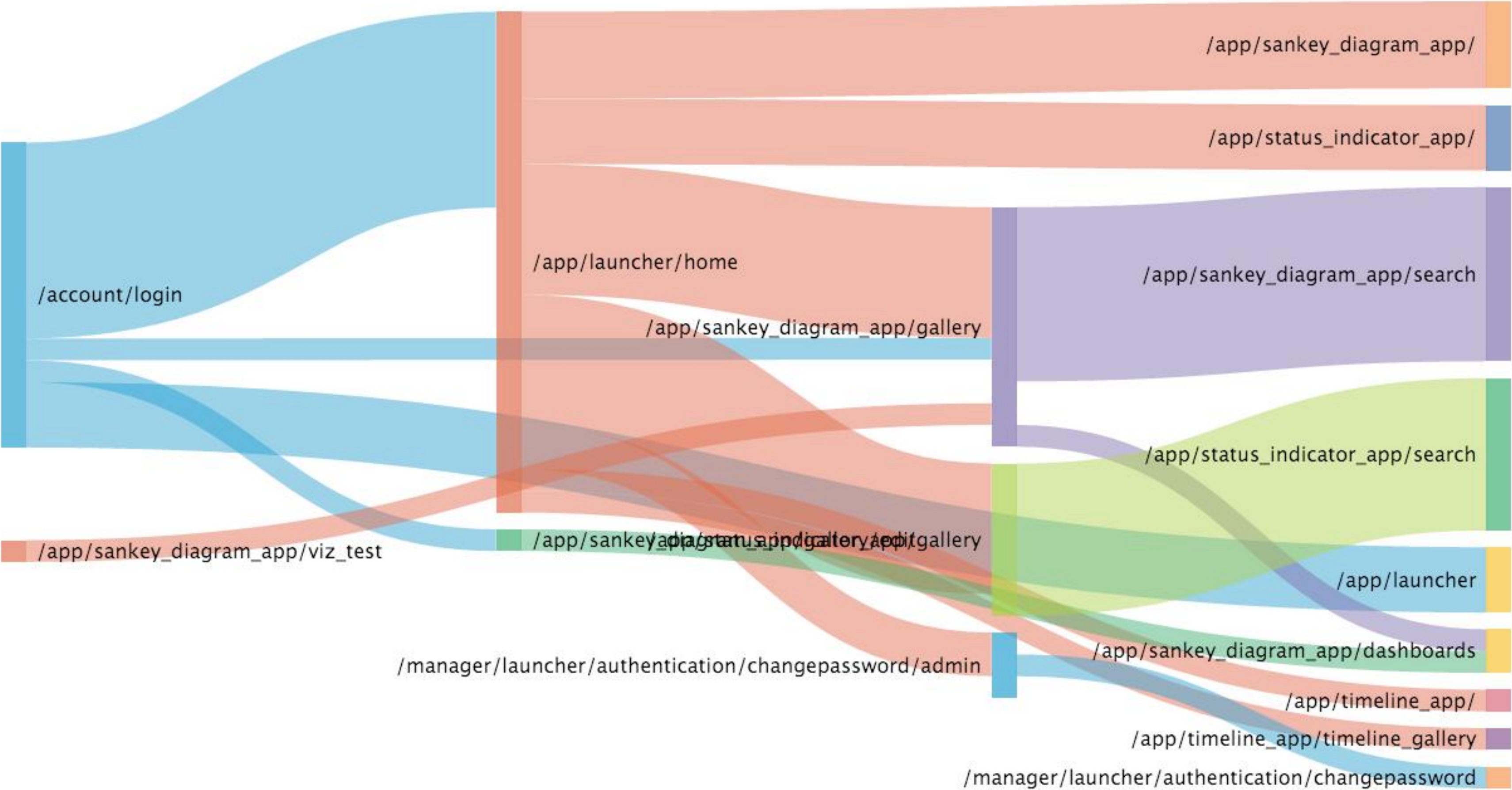


 И даже такие...

Funnel chart



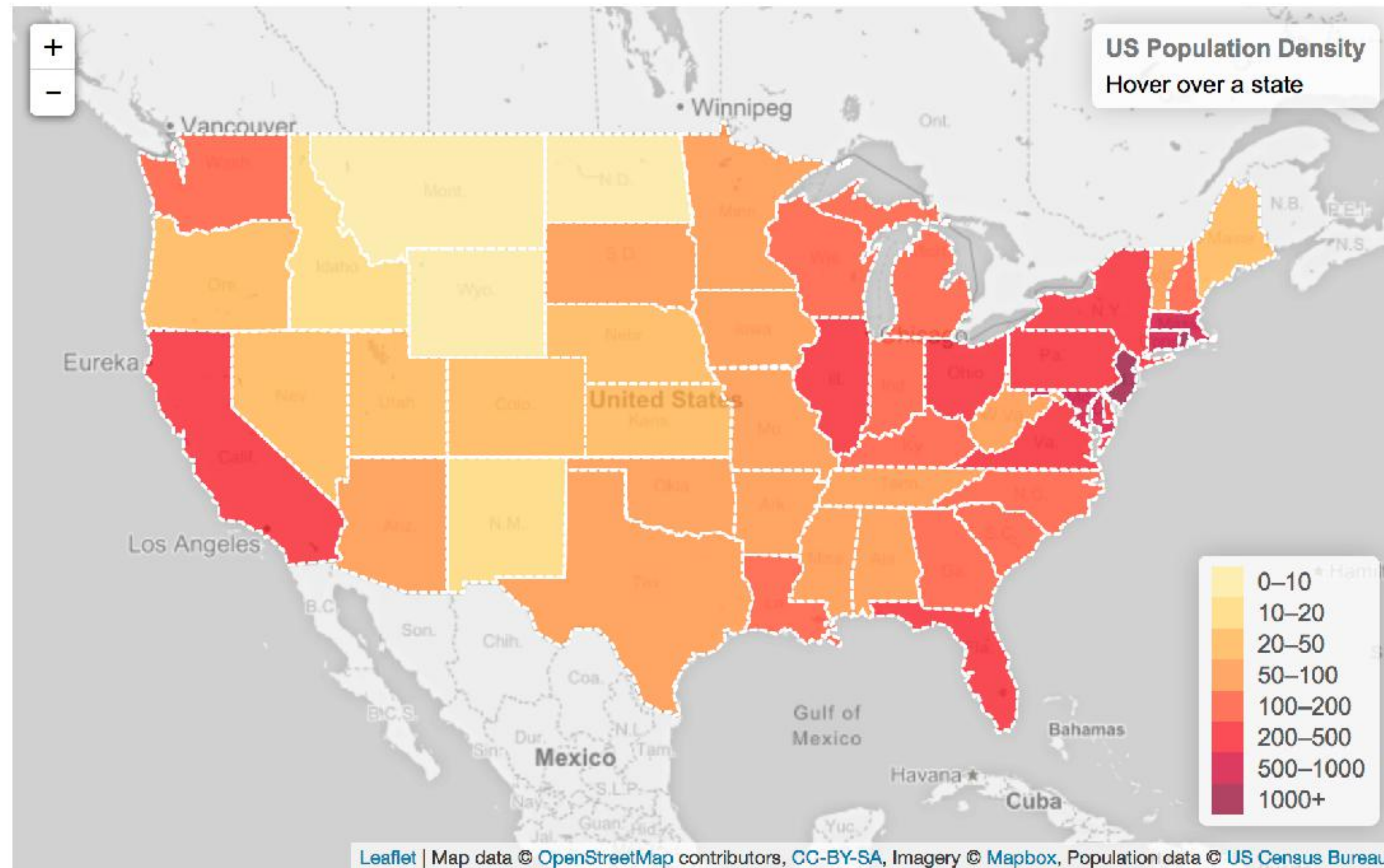
Sankey diagram



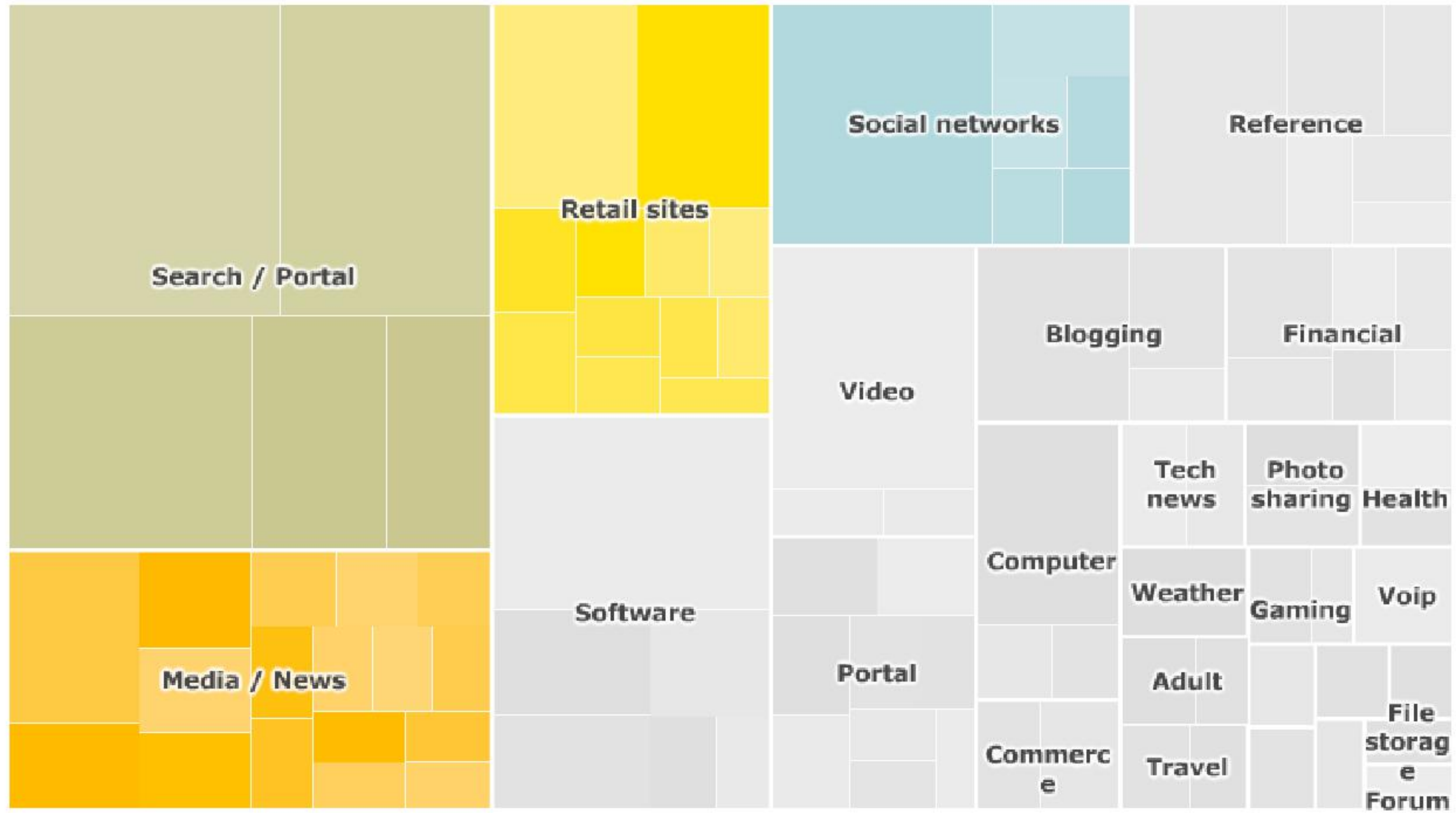
Area Chart + Line Chart



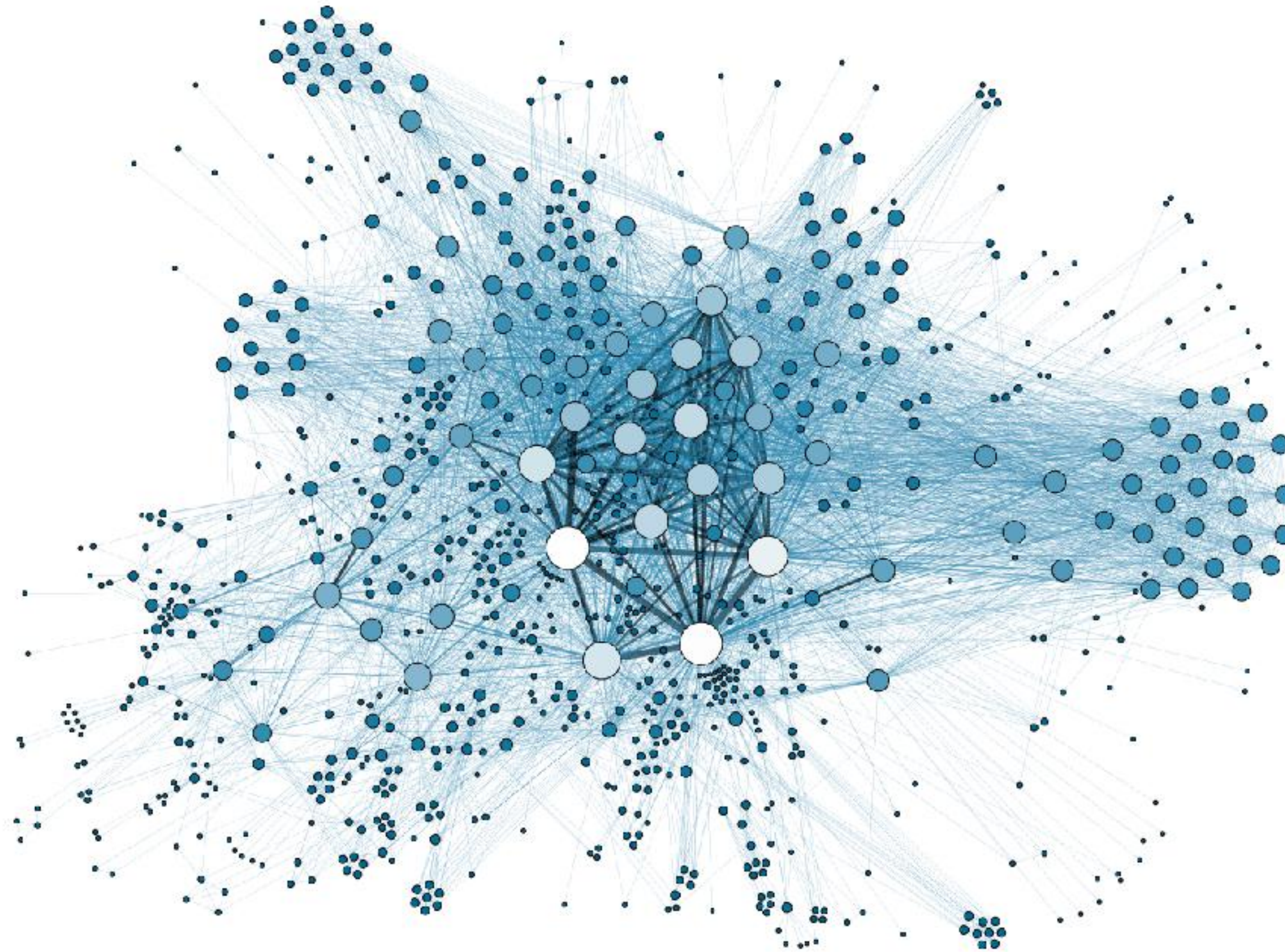
Гео-данные (Choropleth)



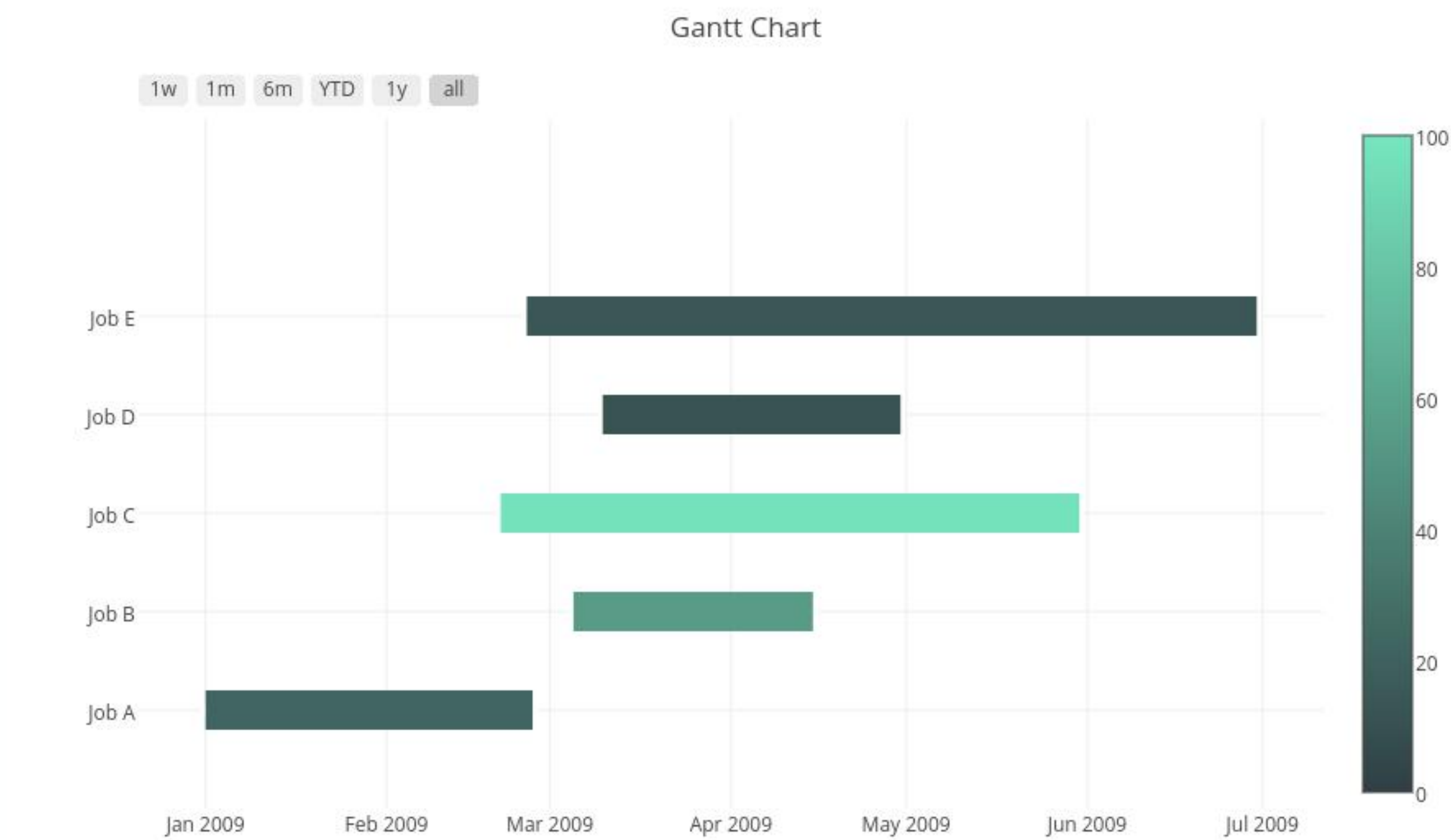
TreeMap



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 Tufte

Edward Tufte «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.

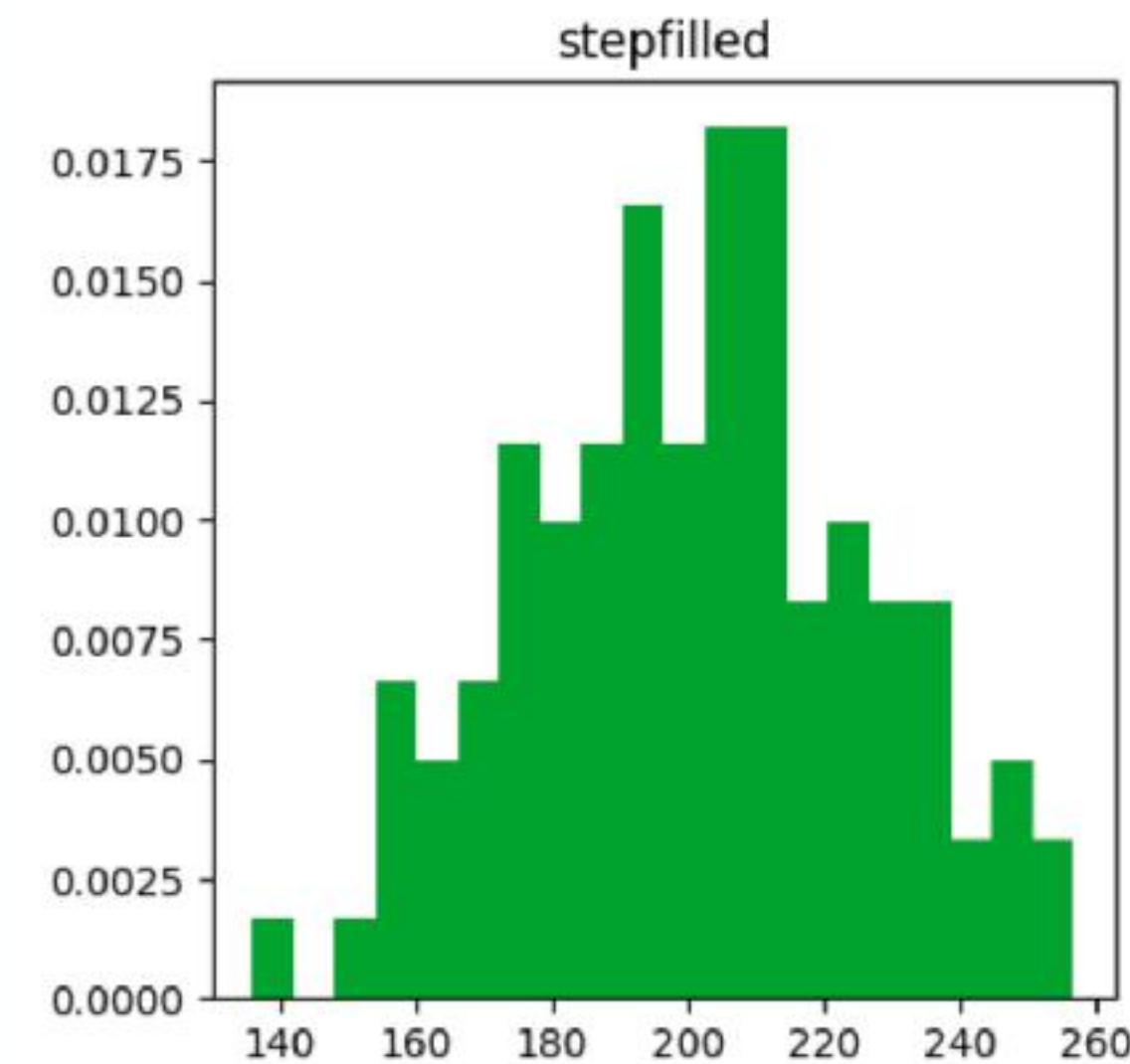
matplotlib



Pygal

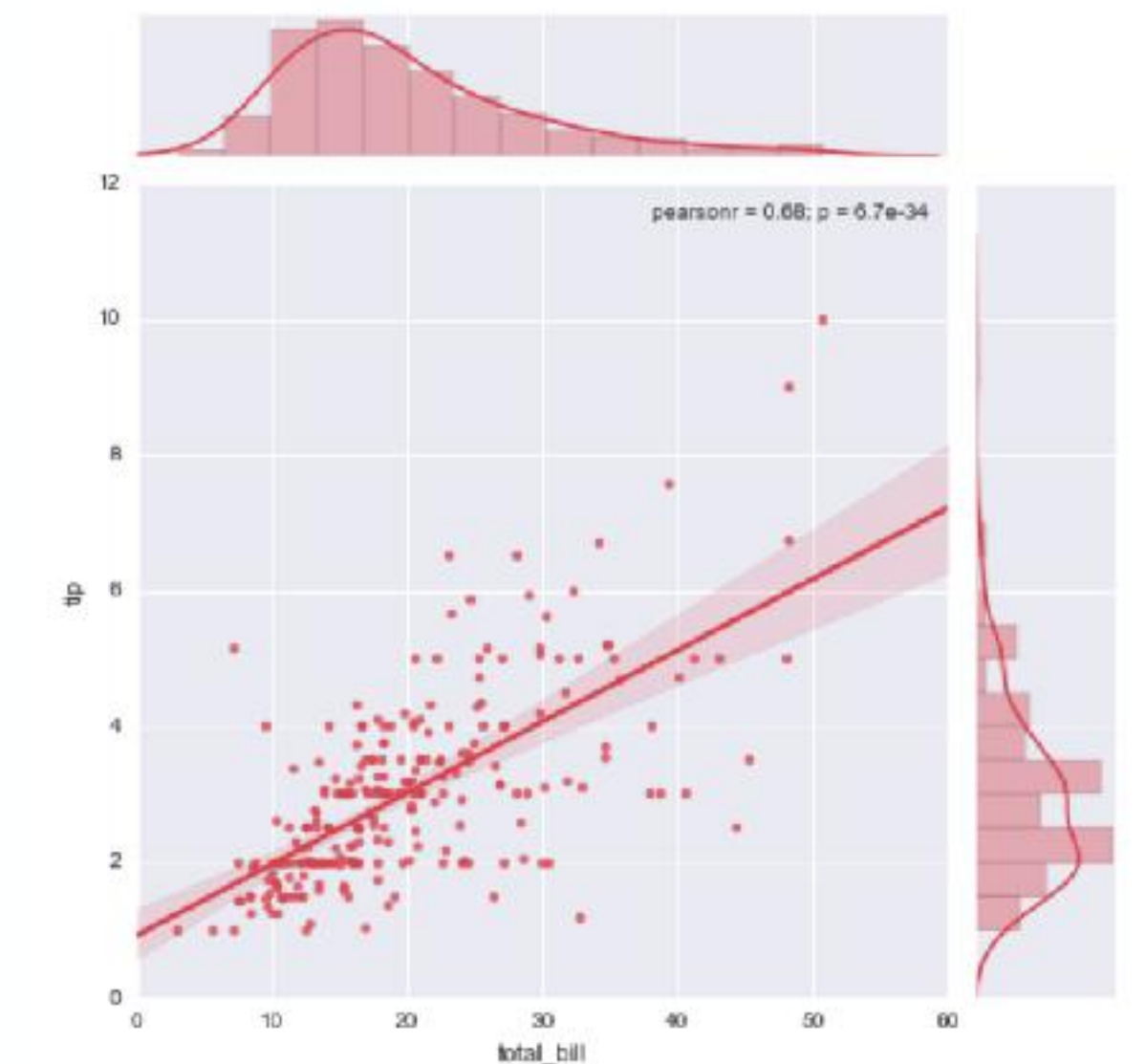
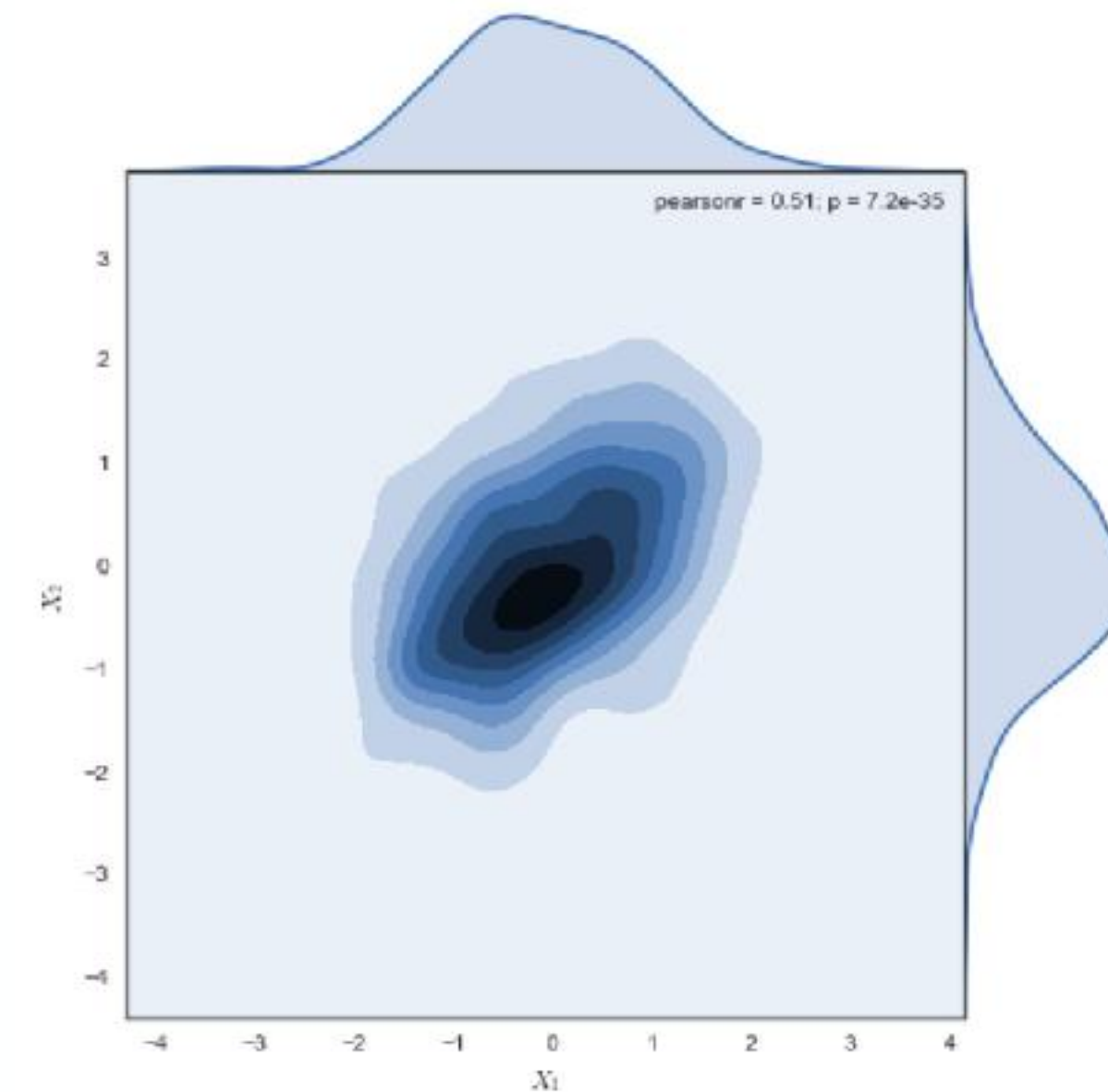
matplotlib

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



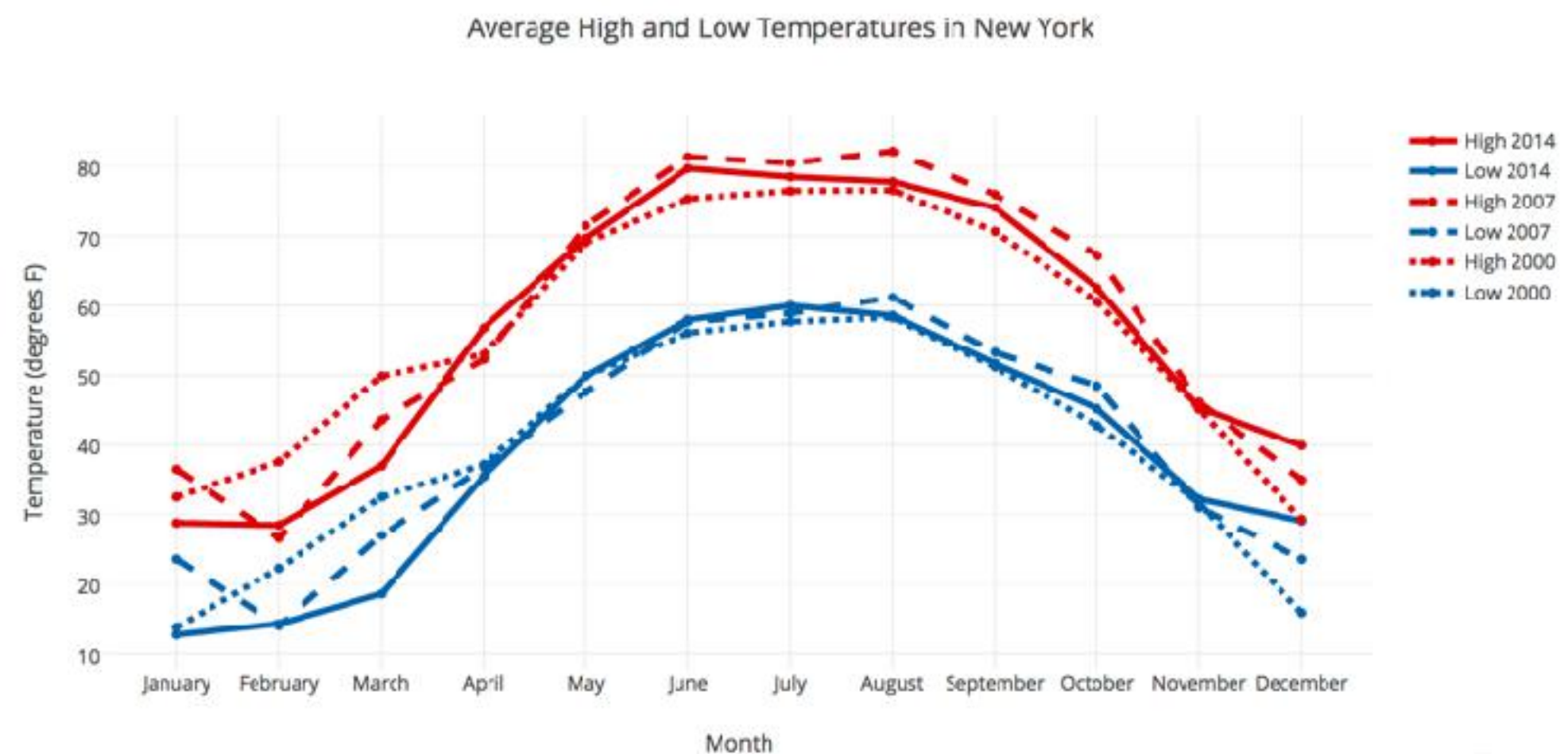
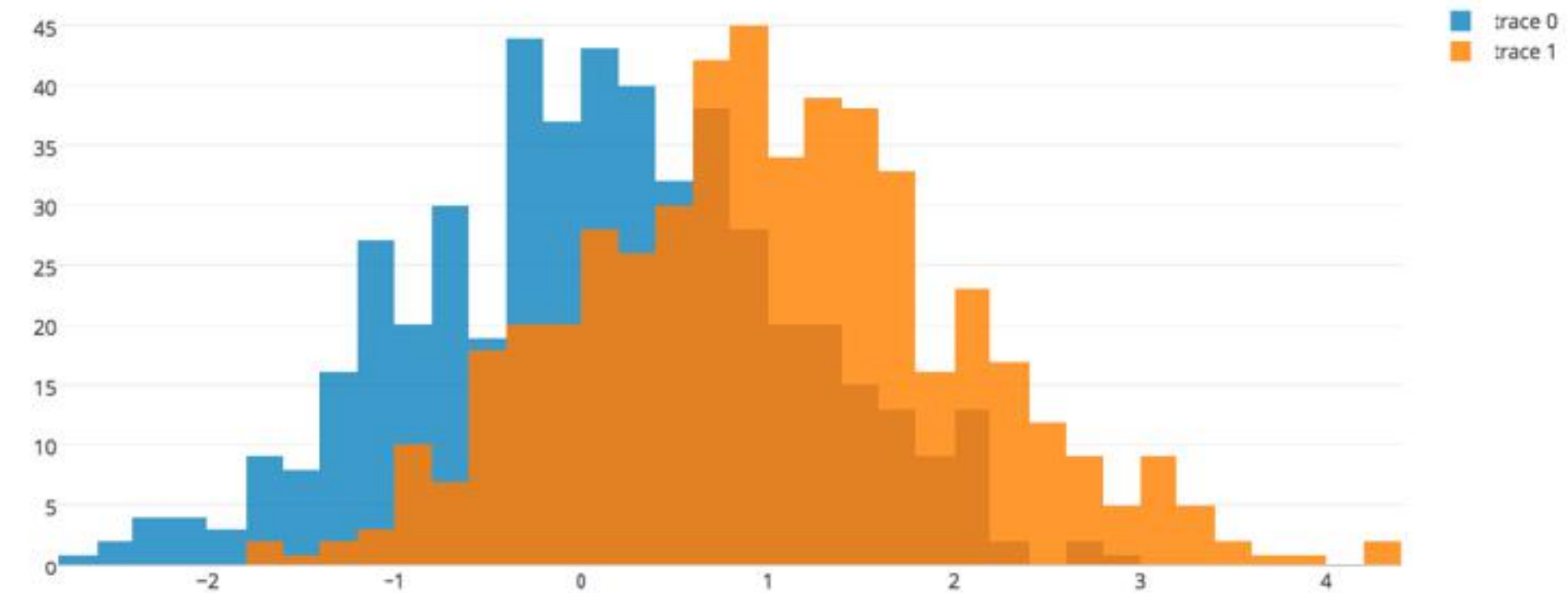
seaborn

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



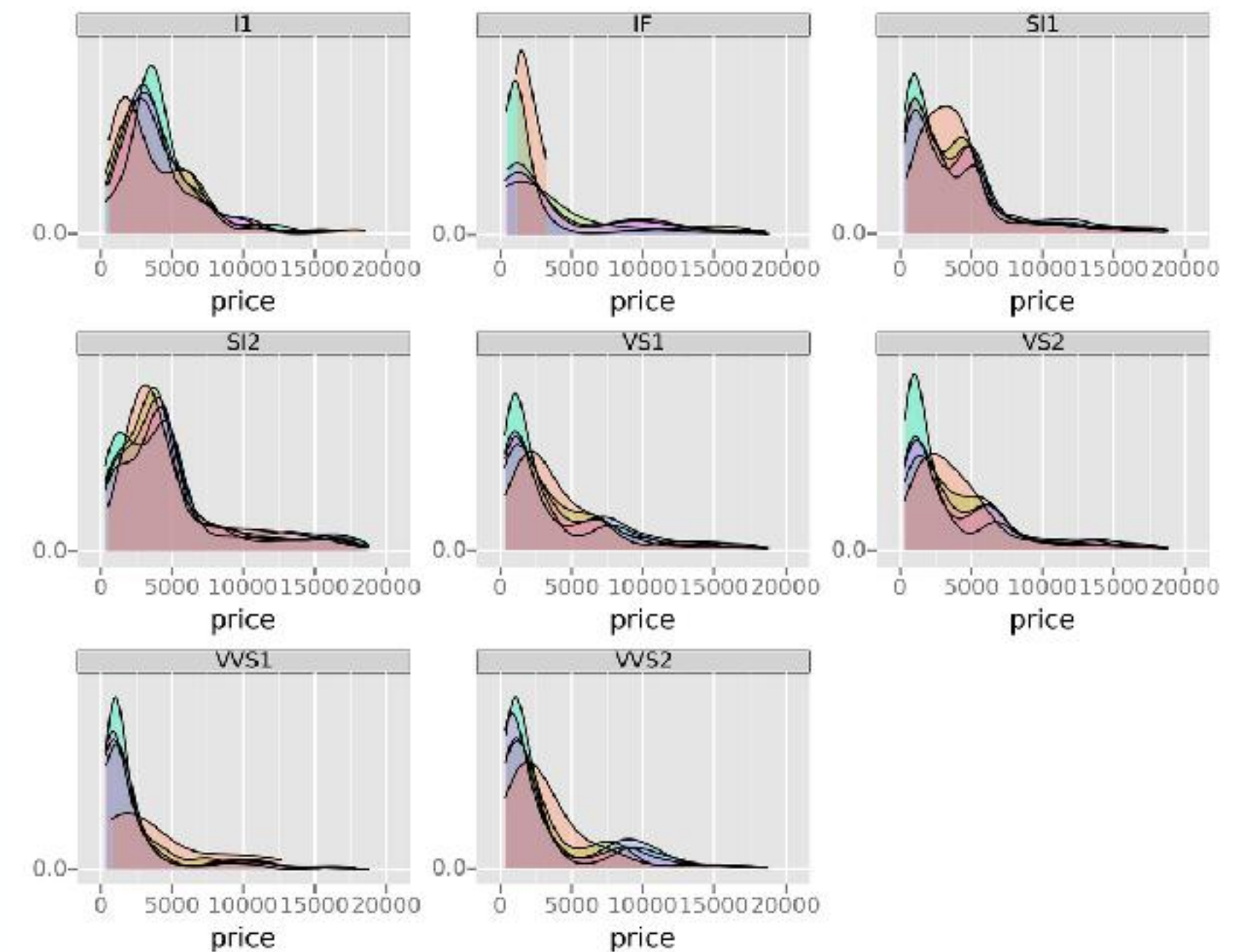
plot.ly + dash

- › 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

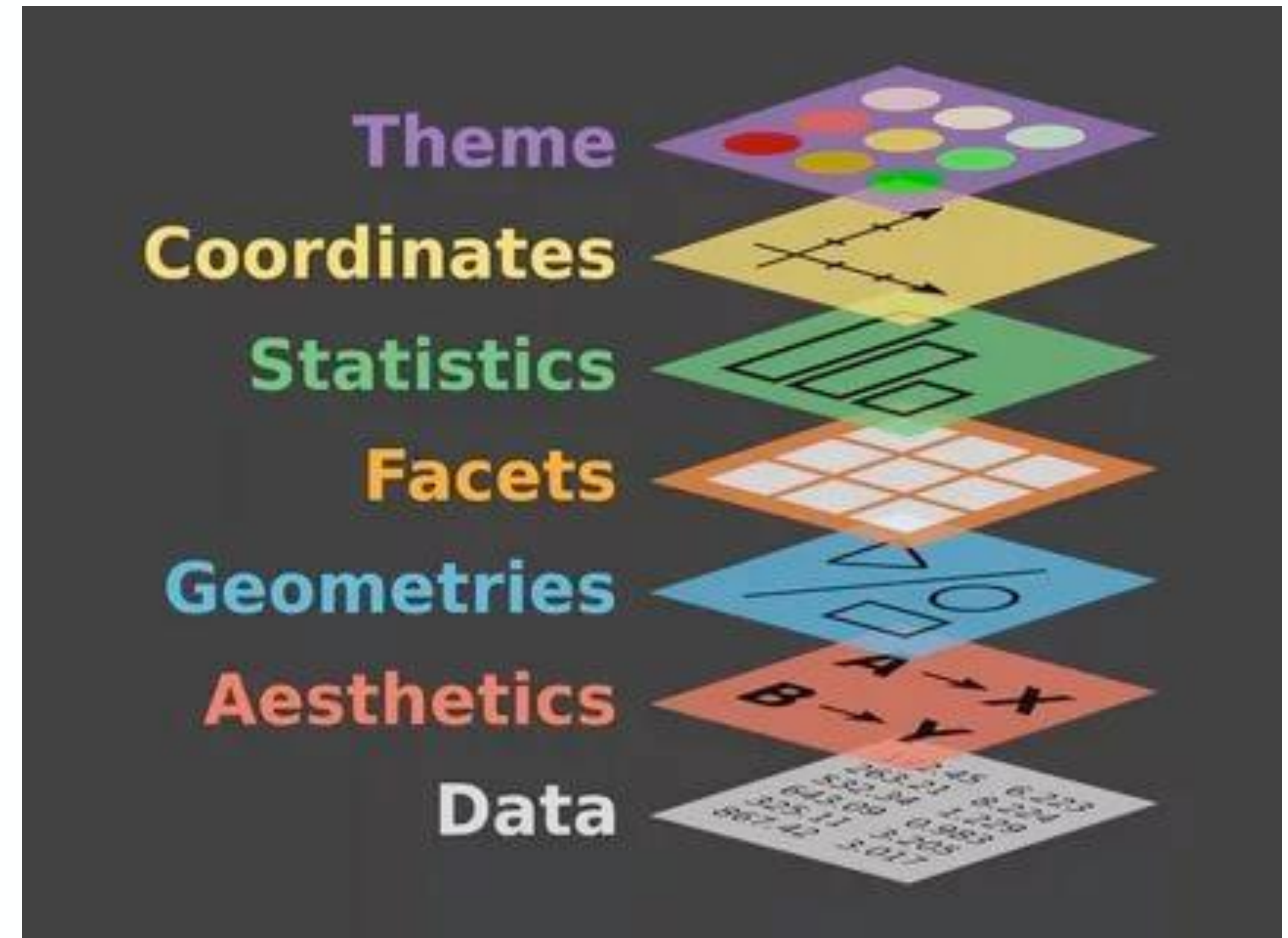


The Grammar of Graphics

- › Leland Wilkinson, 1999

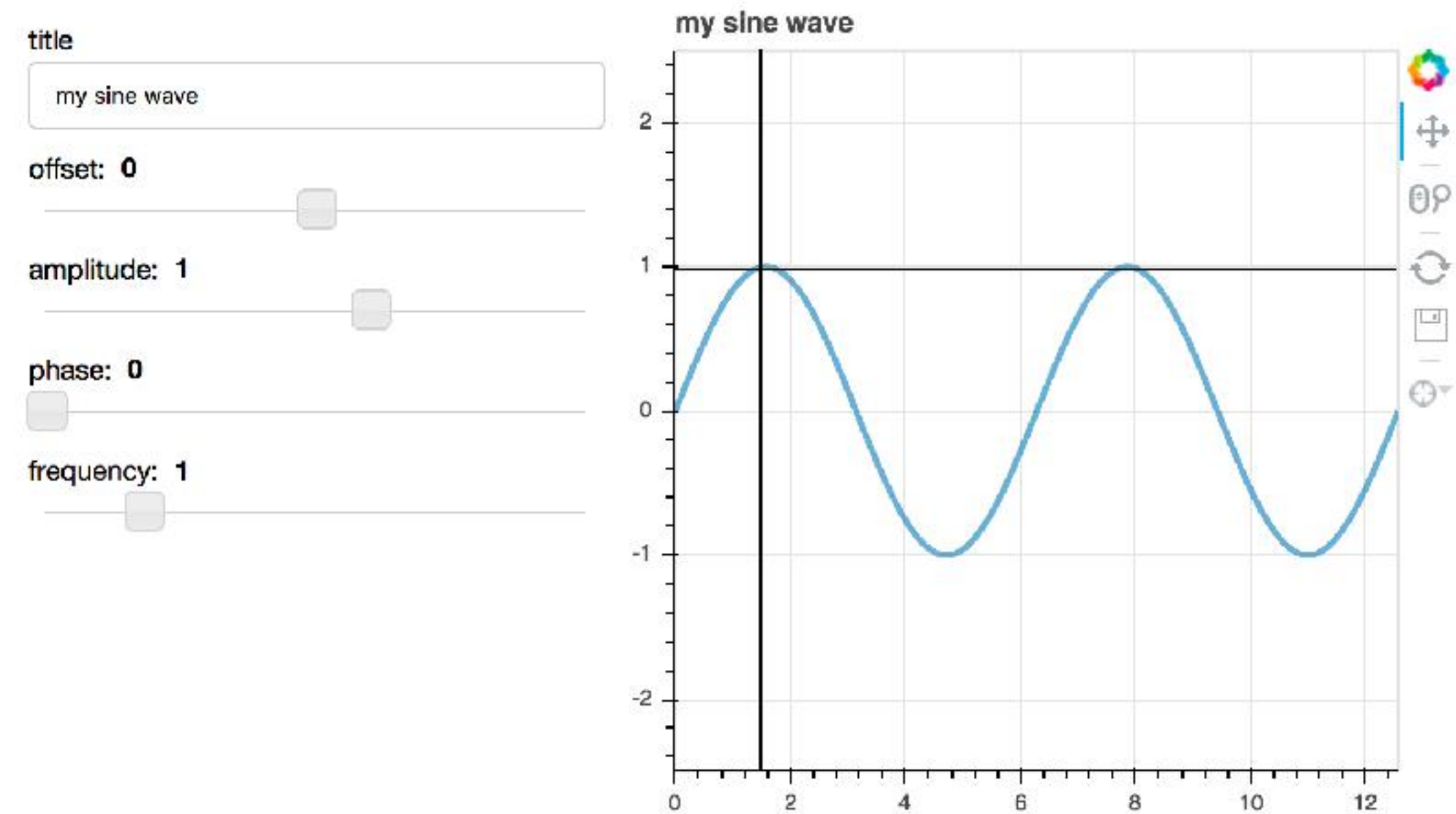
Принципы

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



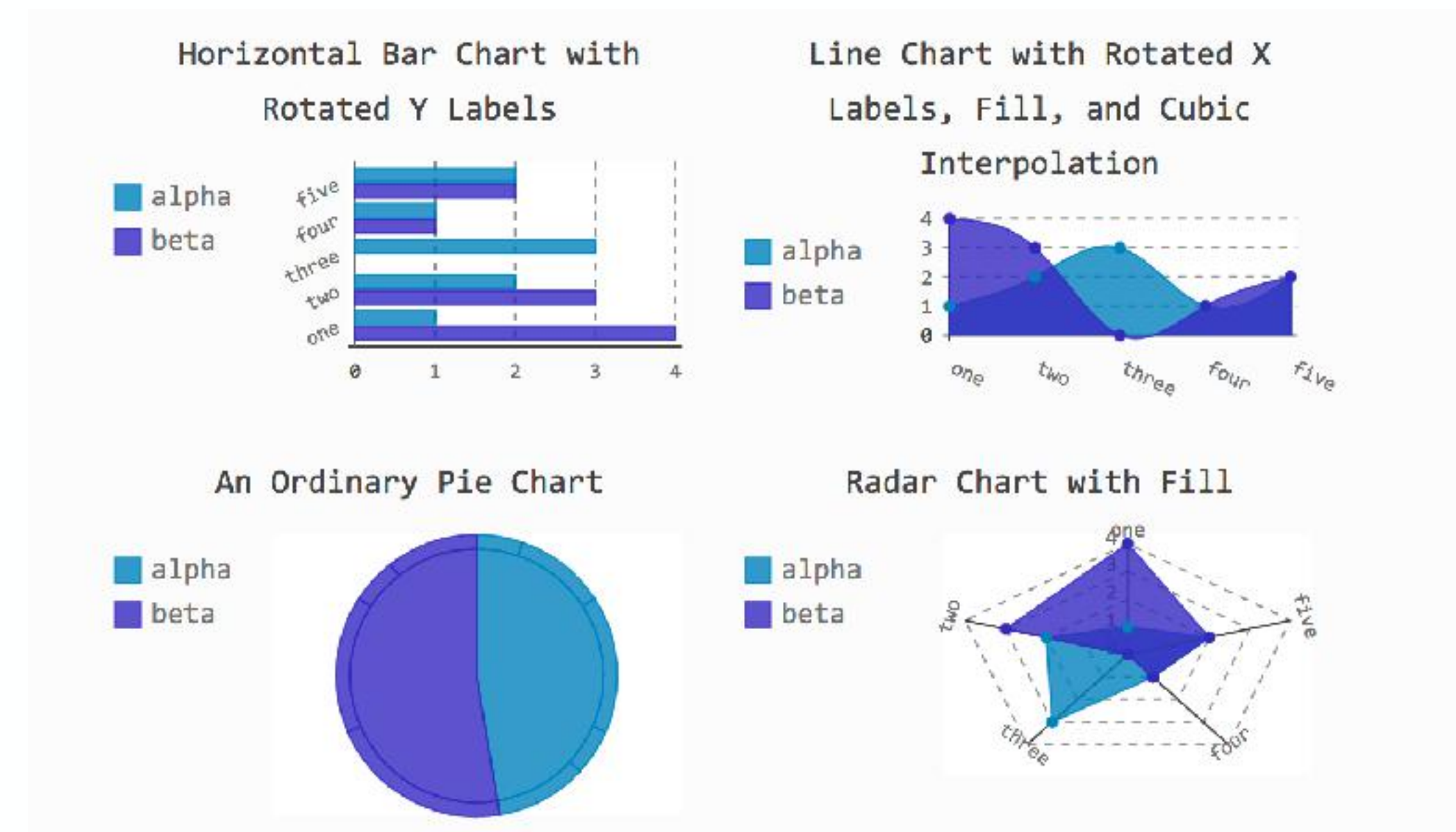
bokeh

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



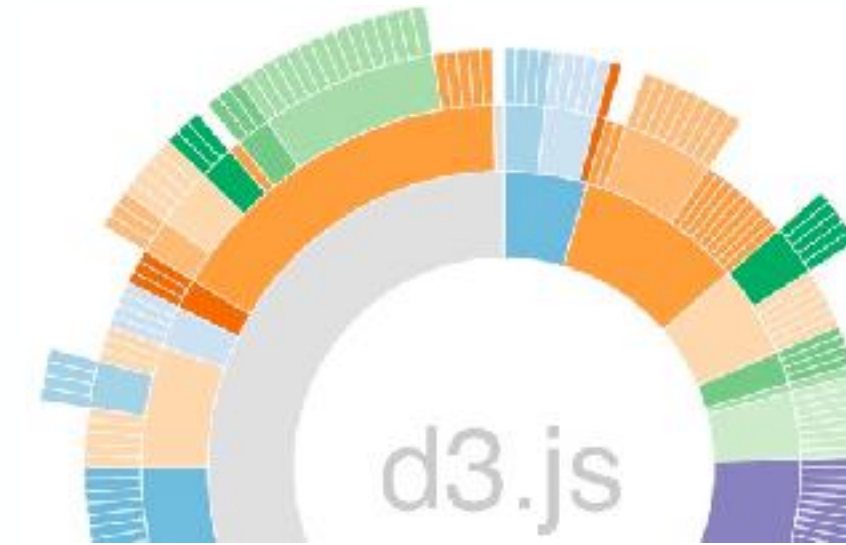
pygal

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

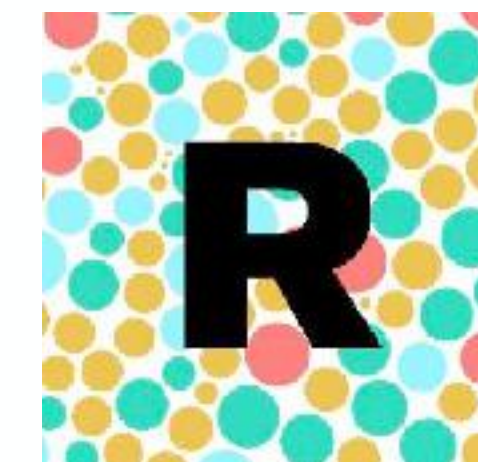


Что еще

- › 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`

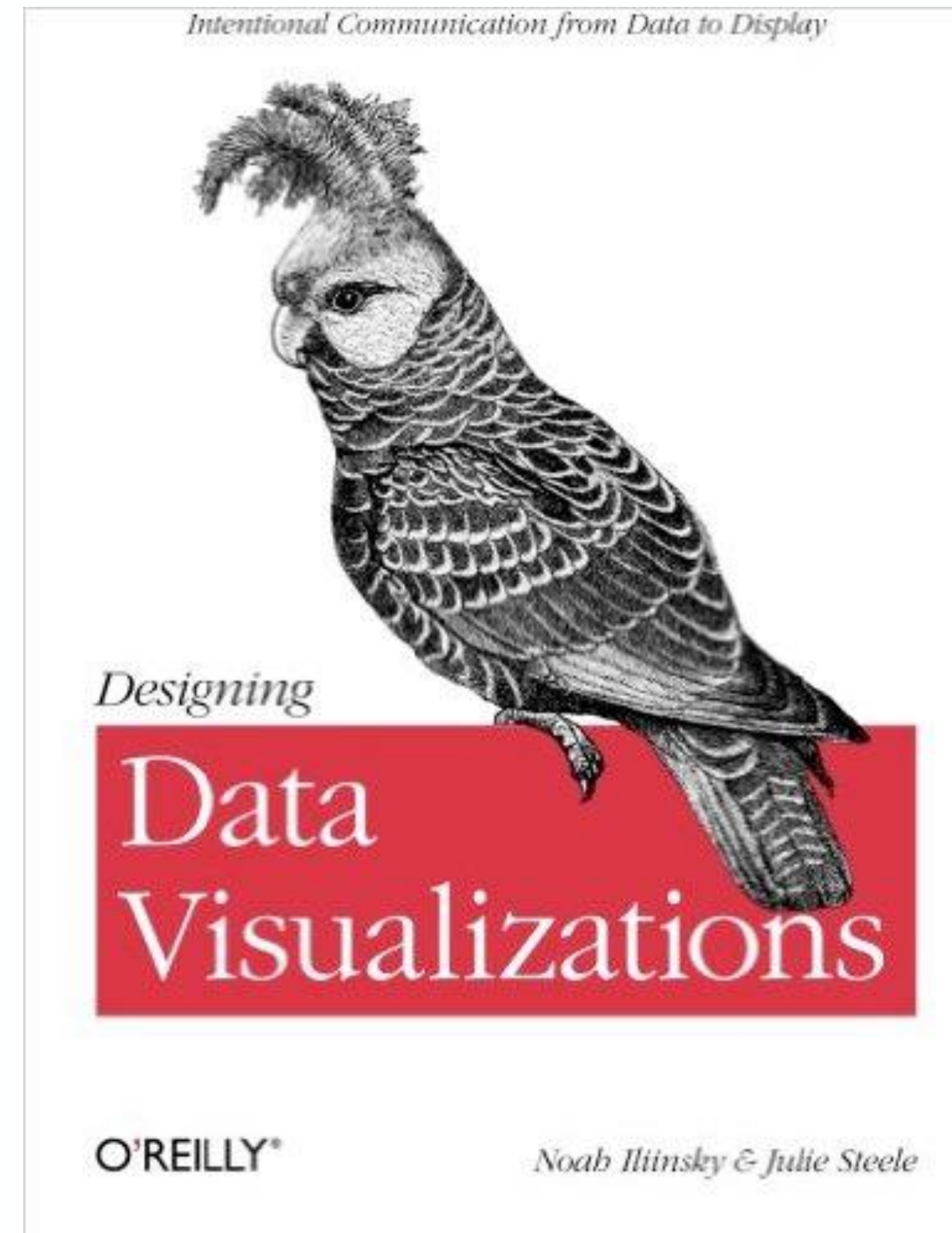


What's next?



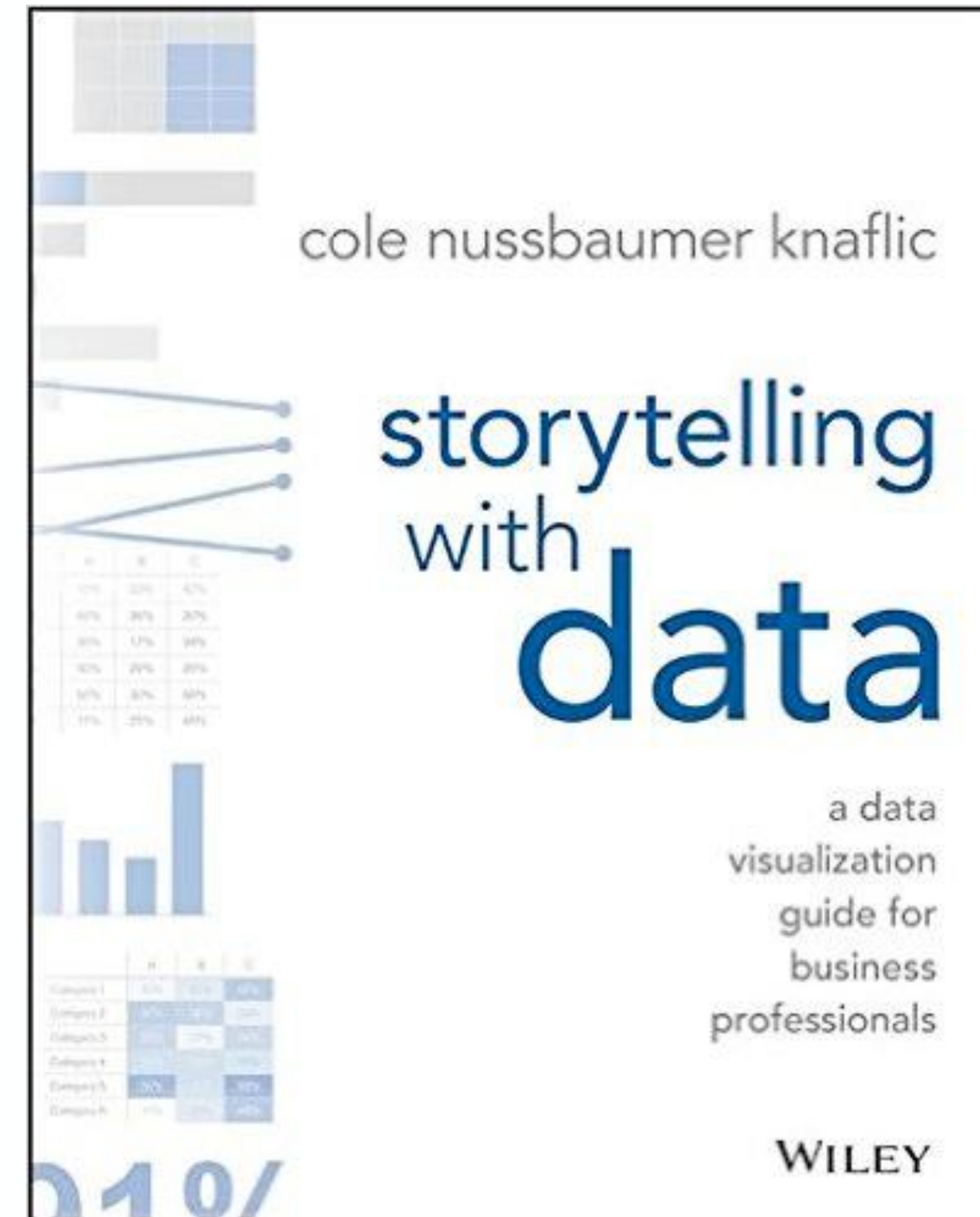
Designing Data Visualizations

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



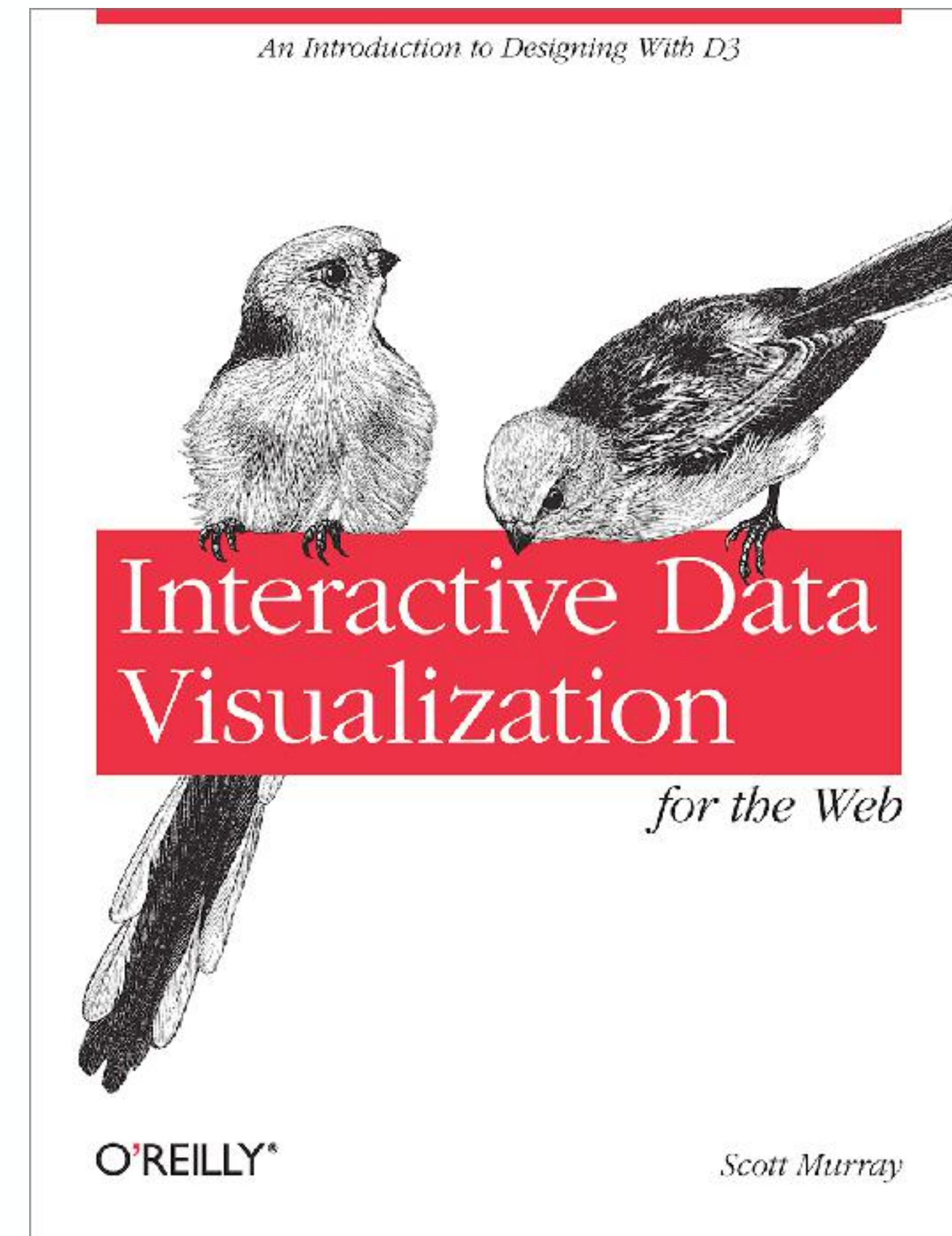
Storytelling with Data

- › как сделать из графиков историю



Interactive Data Visualization

- основы HTML, JS, SVG, DOM
- использование D3.js



Blogs on Data Visualizations

- <http://www.storytellingwithdata.com/>
- <http://junkcharts.typepad.com/>
- <http://www.informationisbeautiful.net/blog/>



Open Data Science

- › крупнейшее русско-язычное сообщество Data Scientist'ов

