

Визуализация данных

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ

Вопросы занятия

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

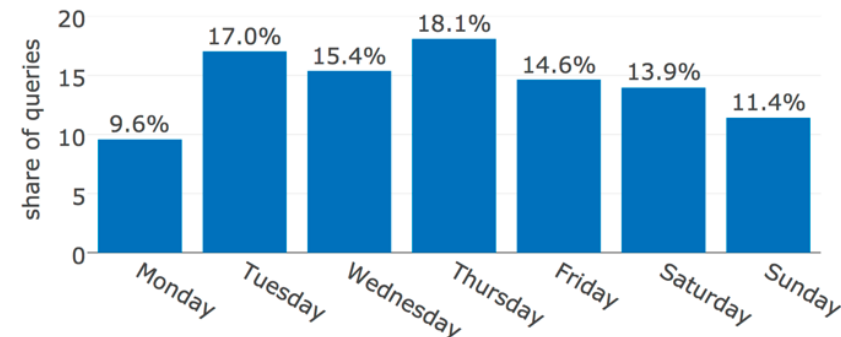
В конце занятия научимся:

1. Рассмотрим основные типы визуализаций и научимся выделять подходящую
2. Рассмотрим основные инструменты python для создания графиков

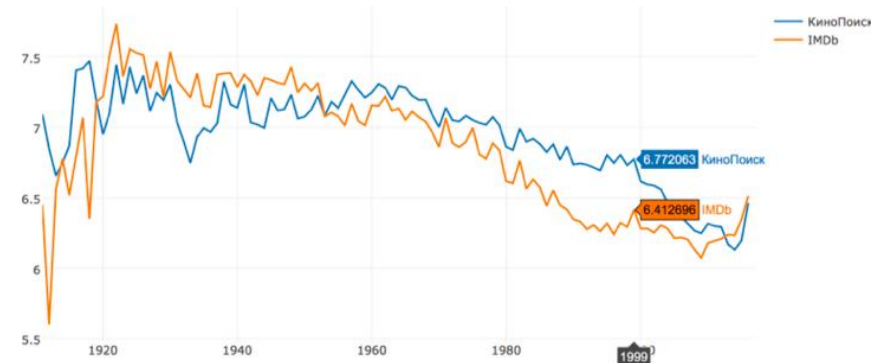
ЧТО ТАКОЕ ВИЗУАЛИЗАЦИЯ

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

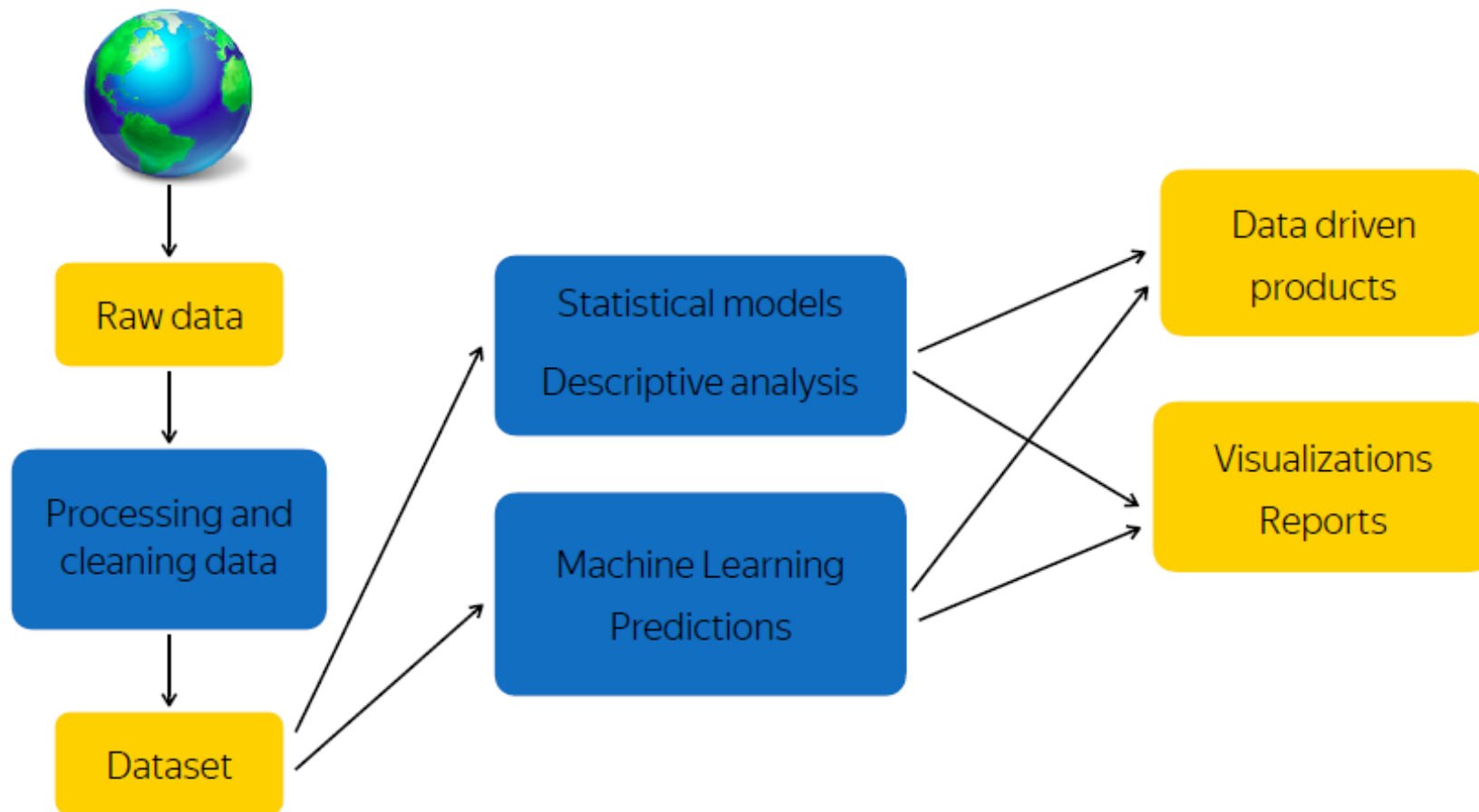
Запросы по дням недели



Оценки фильмов



Работа с данными



РОЛЬ ВИЗУАЛИЗАЦИИ

- *exploratory* - «разговор наедине с данными»
- *explanatory* - раскрыть и донести СВОЮ МЫСЛЬ



Визуализация нужна вообще?

Выборки одинаковые?

Пример выборок

все статистики 4х выборок
одинаковы

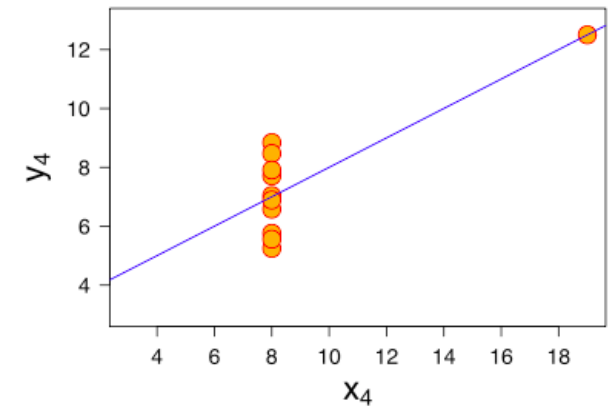
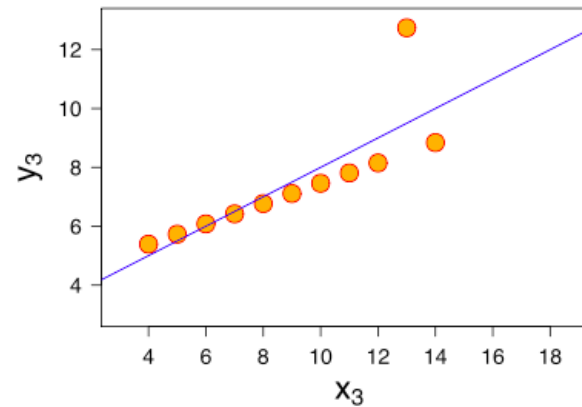
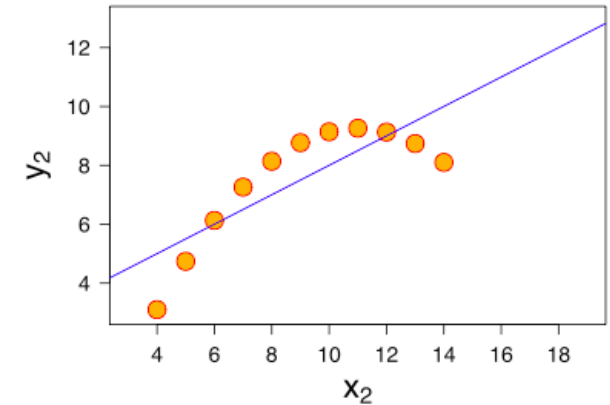
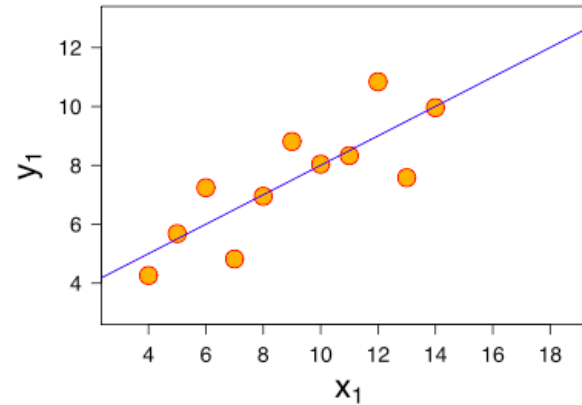
- › $\text{mean } x = 9$
- › $\text{sample variance of } x = 11$
- › $\text{mean } y = 11.5$
- › $\text{sample variance of } y = 4.125$
- › $\text{correlation between } x \text{ and } y = 0.816$

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

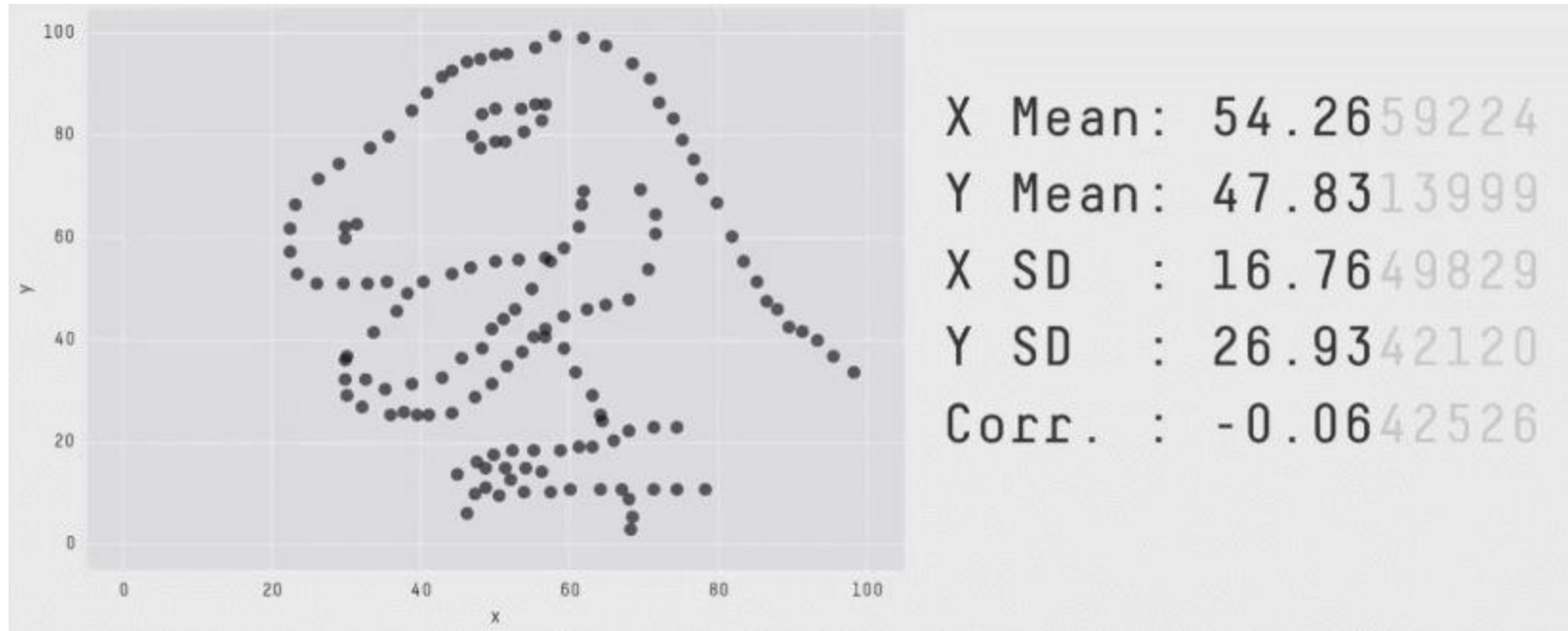
Квартет Энскомба

пример был придуман
статистиком Фрэнсисом
Энскомбом в 1973 году

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



И другие варианты...



<https://www.autodeskresearch.com/publications/samestats>

Данные

числовые

› дискретные/непрерывные

категориальные

› nominal/ordered



Формы выражения (*Visual Encodings*)

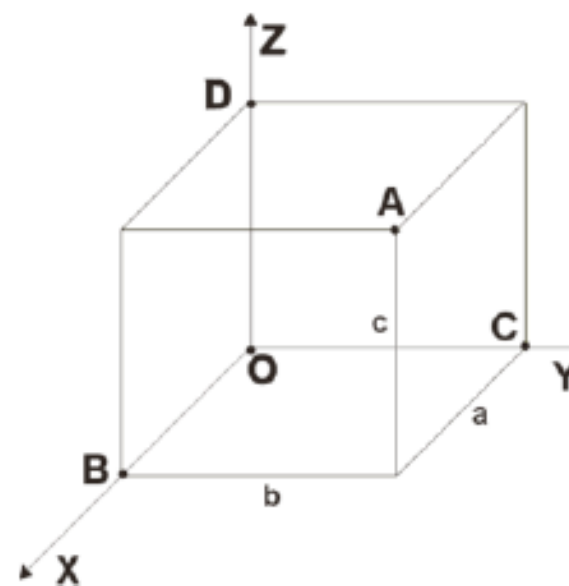
данные -> отображение их на графике

- › 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



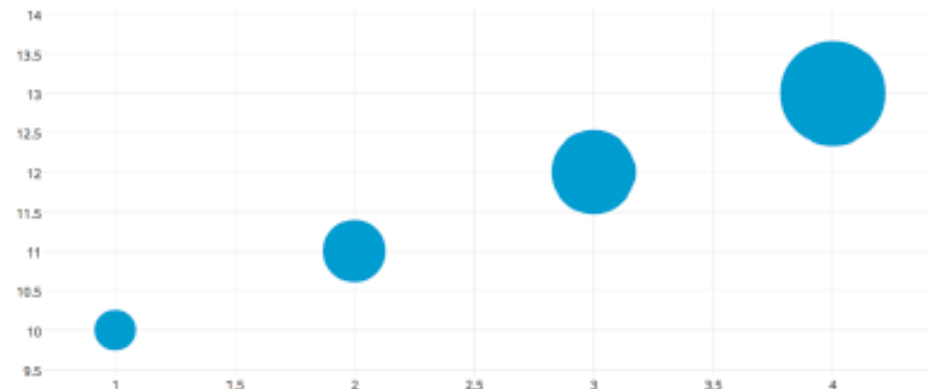
Размер (длина, площадь, объем)

длина

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

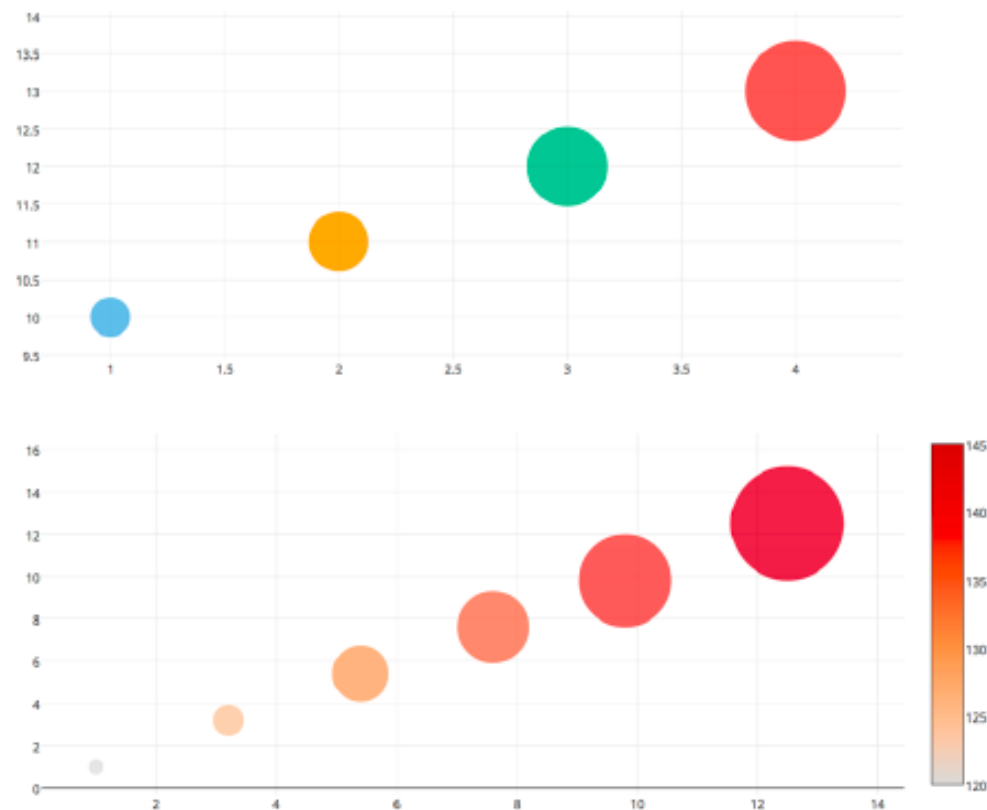
площадь, объем

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



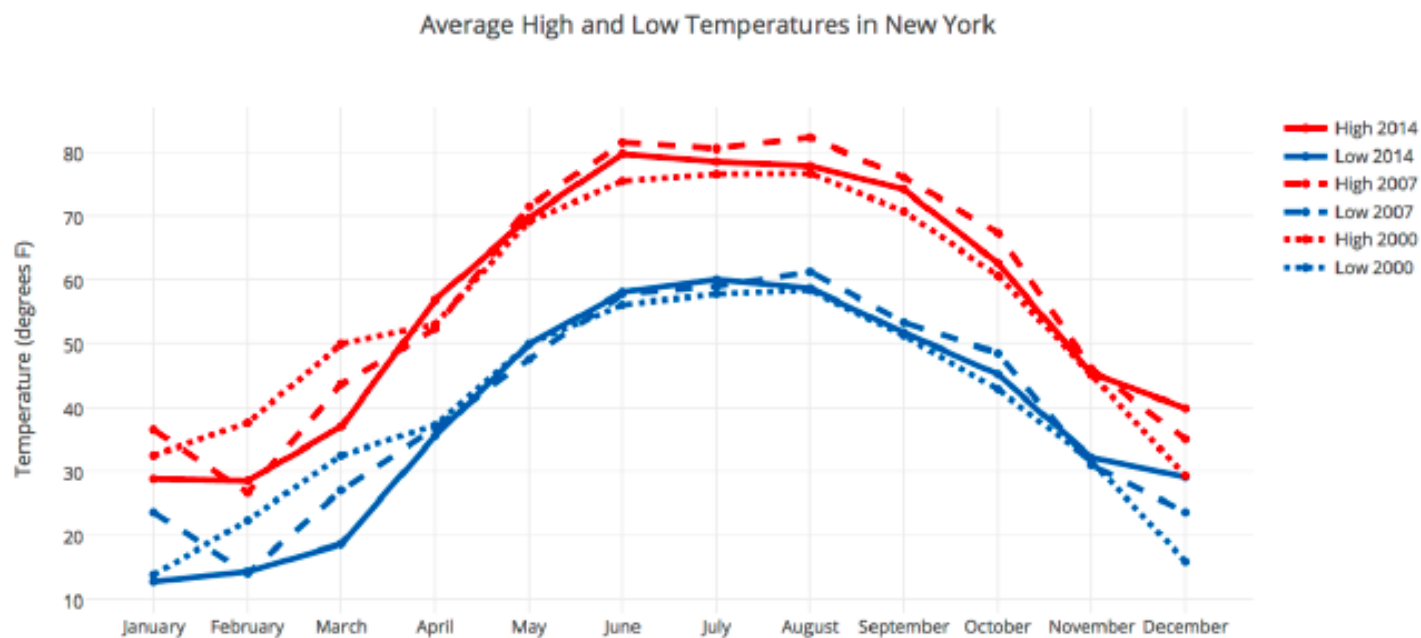
Цвет (hue/saturation)

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



И многие другие...

- › тип линии
- › текстура
- › форма markers



Наиболее восприимчивые графики для людей

- › 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



<http://flowingdata.com/2010/03/20/graphical-perception-learn-the-fundamentals-first/>

Выбор графика

- › 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

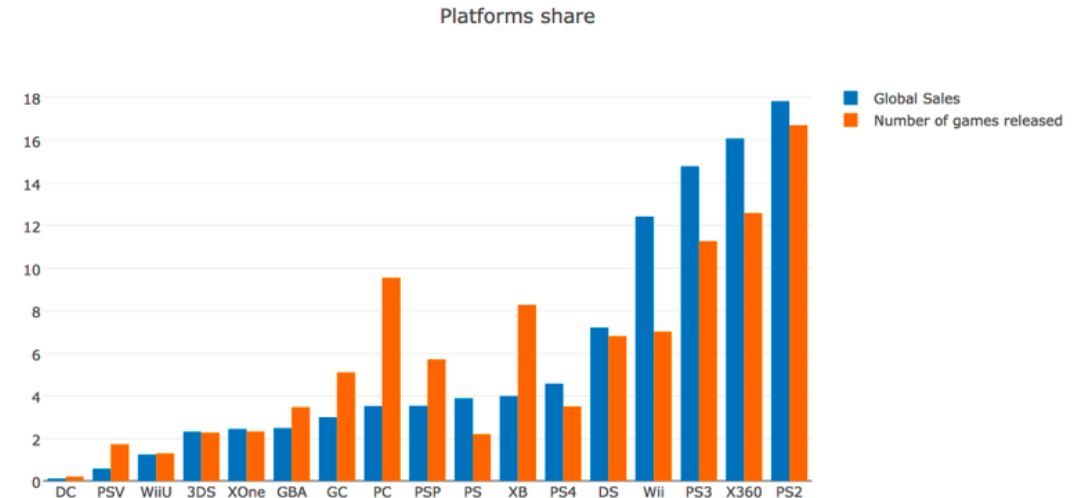
<https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings>

Обычное сравнение (*Nominal comparison*)

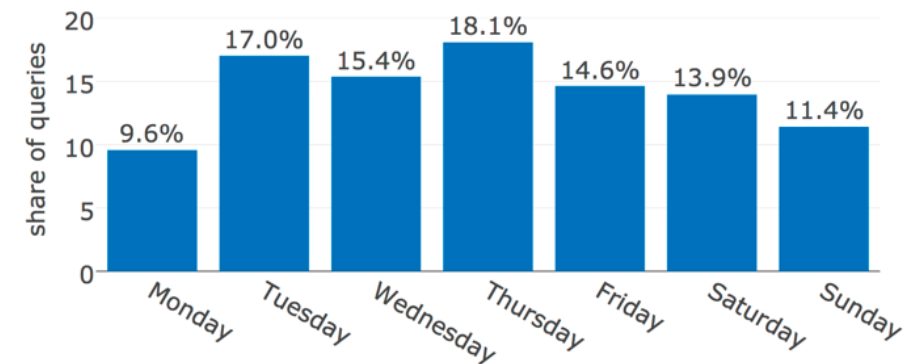
Nominal comparison – простое сравнение одной или нескольких метрик по категориям без определенного порядка

Задача - сравнить игровые платформы по числу выпущенных и проданных игр

› Горизонтальный или вертикальный bar chart



Запросы по дням недели



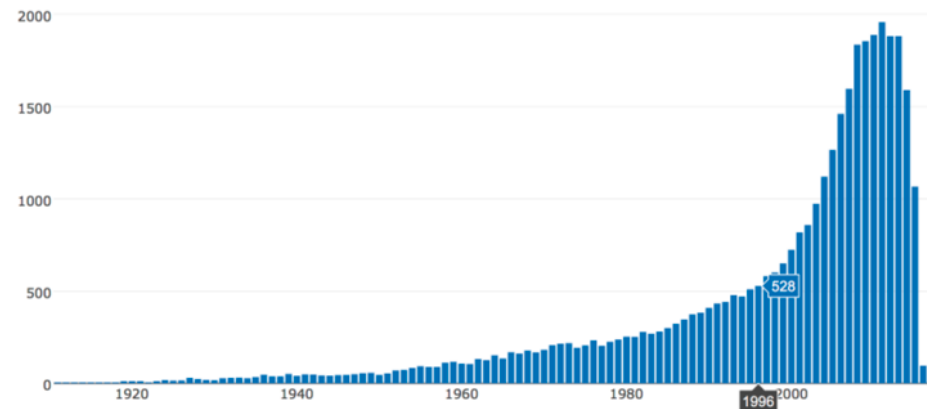
Time Series

Time Series - изменение одной или нескольких метрик во времени

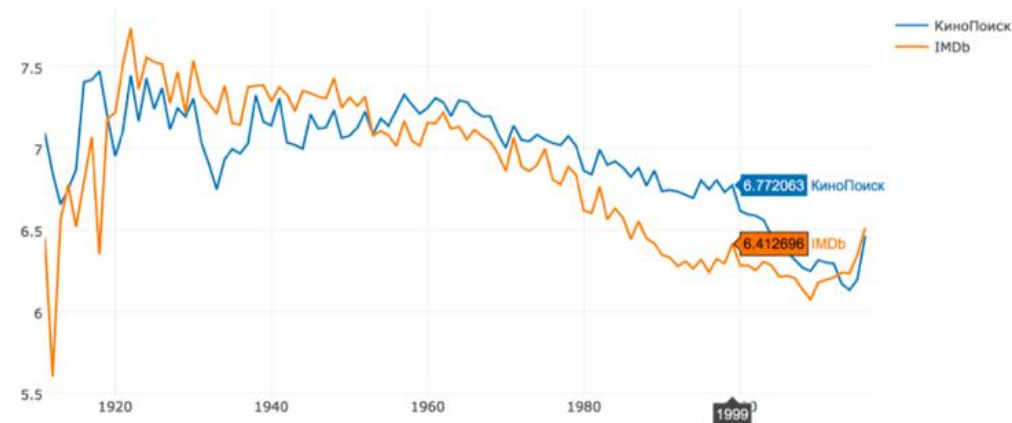
- › Line chart, чтобы подчеркнуть тренд
- › Bar chart, чтобы выделить отдельные значения
- › Временная переменная должна располагаться на оси X

Задача - отобразить динамику числа проданных компьютерных игр в мире

Фильмы на Кинопоиске



Оценки фильмов

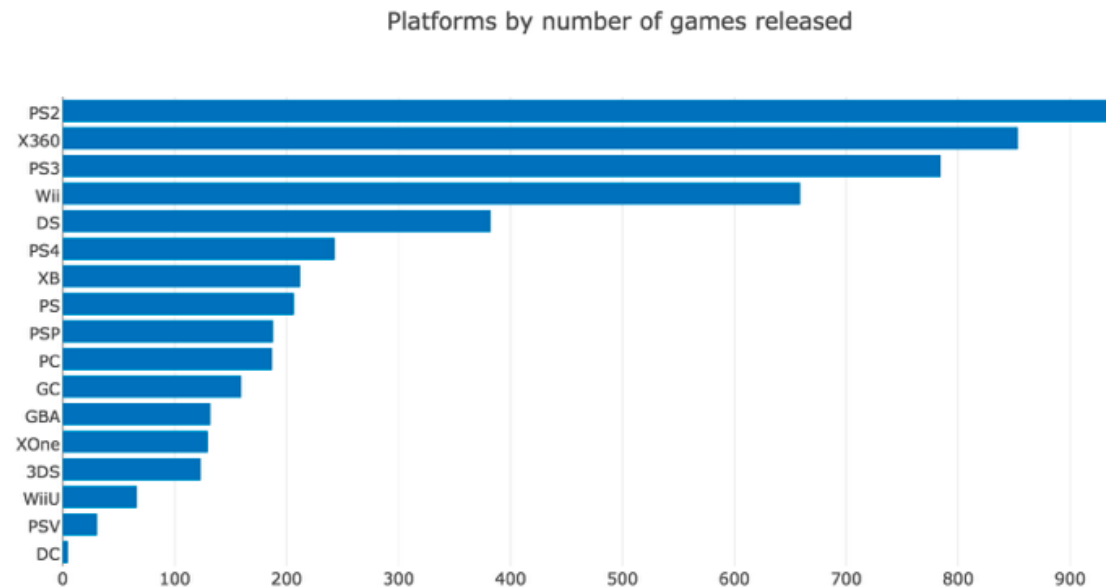


Ranking

Ranking - значения метрики для категорий, упорядоченные по размеру

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

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

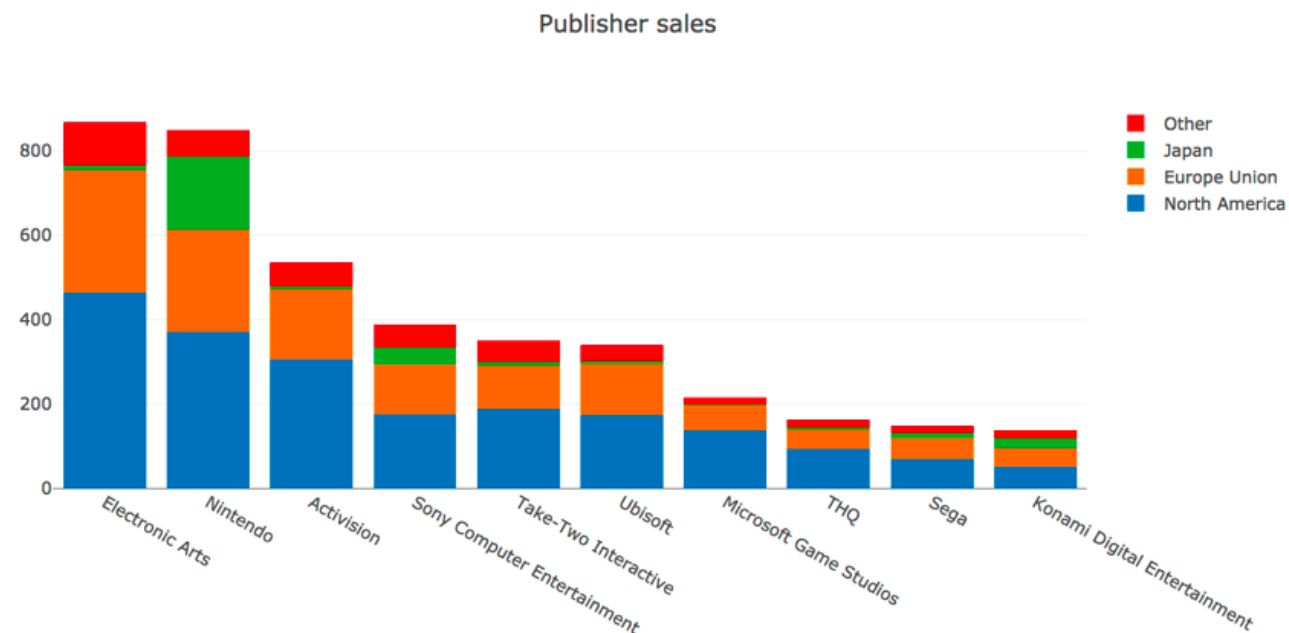


Part-to-hole

Part-to-hole - доли отдельных категорий от целого

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

Пример - показать, какие доходы у разных игровых компаний и как они распределяются по рынкам (США, Европа и т.д.)

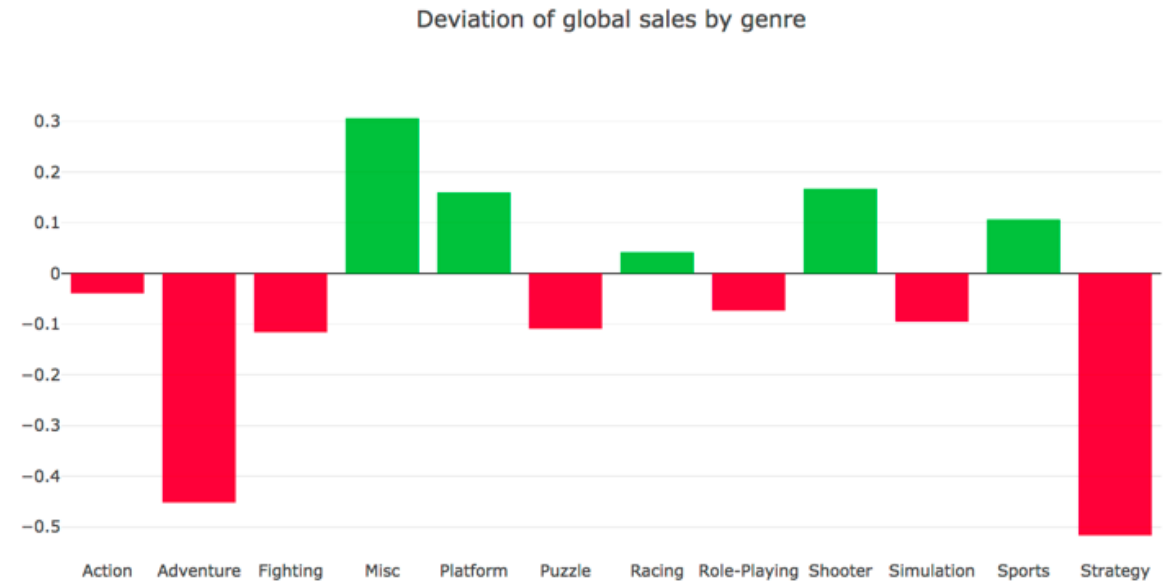


Deviation

Deviation - сравнение показателей для категорий с baseline

› bar chart, чтобы подчеркнуть отдельные значение

Задача — посмотреть, как отличаются средние продажи для разных жанров



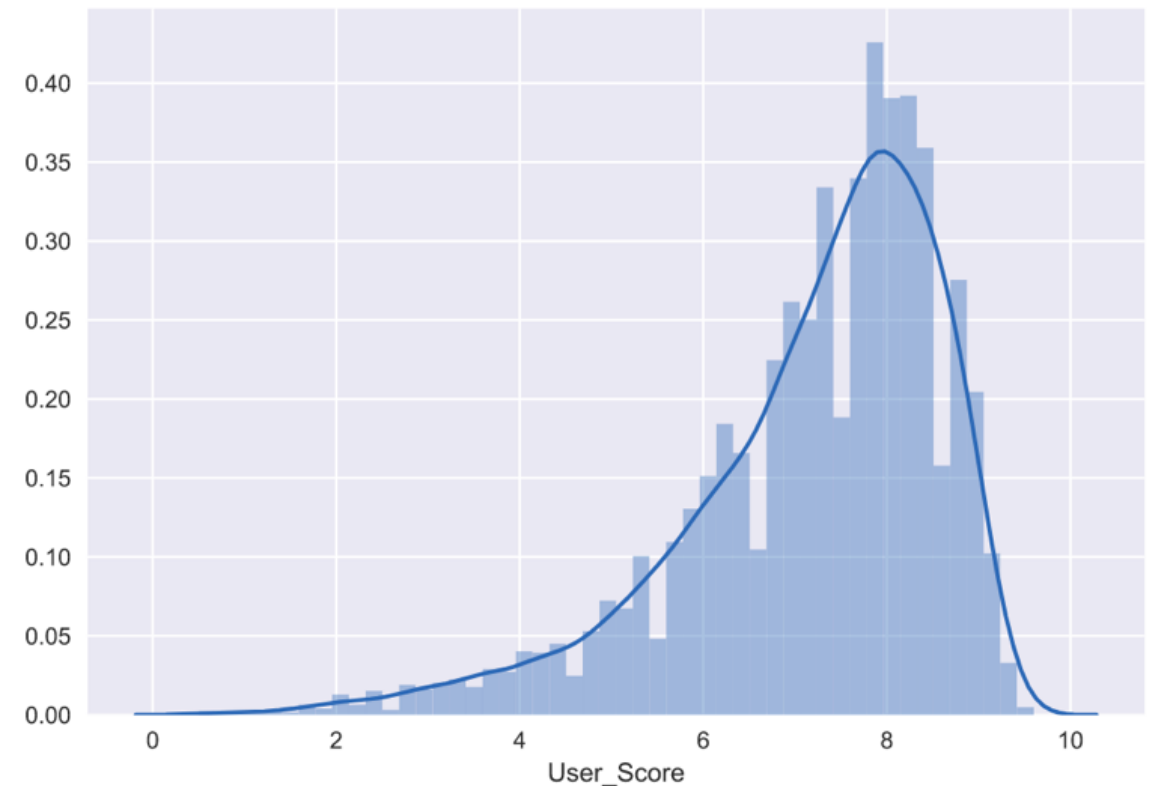
Frequency Distribution

Frequency Distribution -

распределение величины (может быть нормированным)

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

Задача - показать распределение пользовательских оценок игр

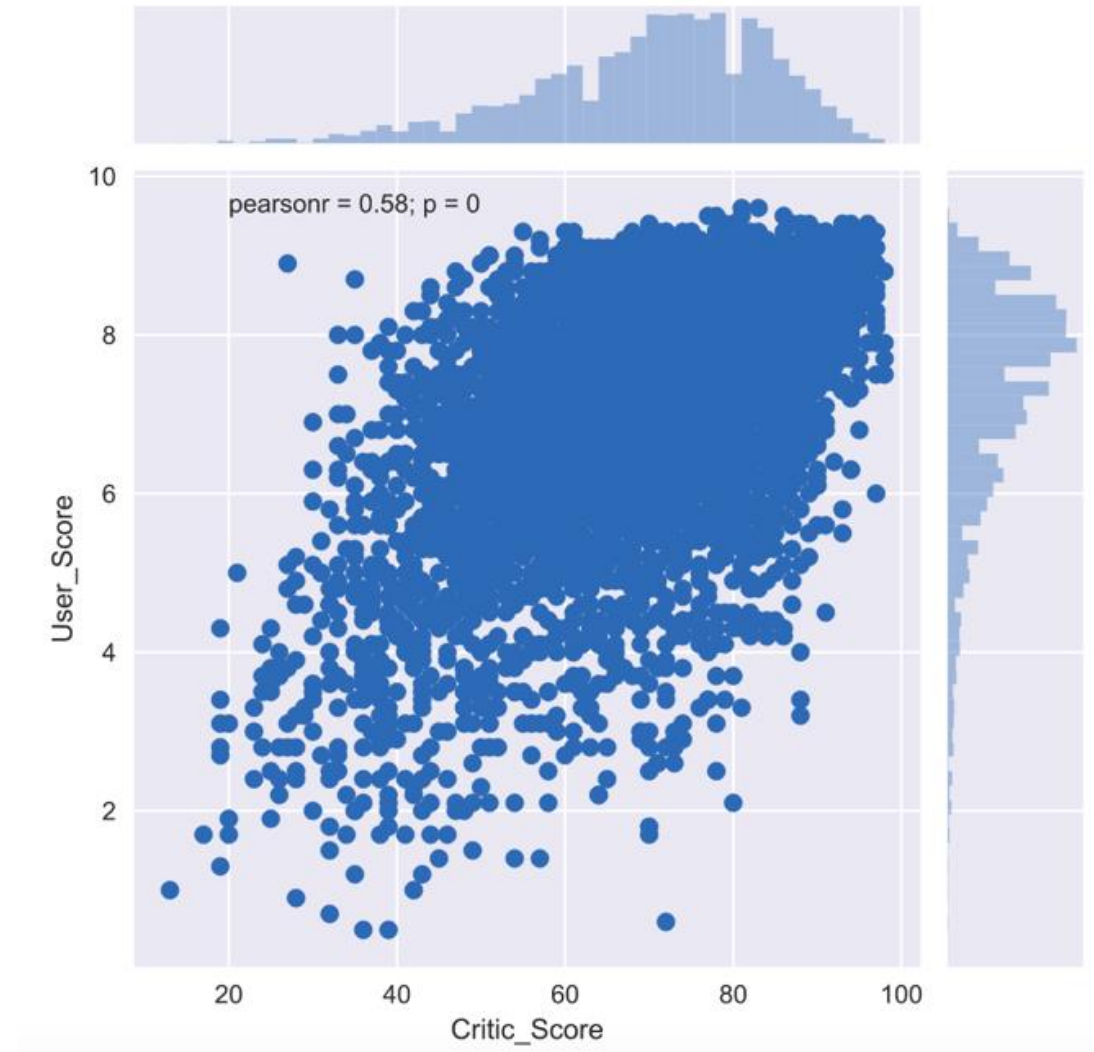


Correlation

Correlation - корреляция между двумя численными величинами





































› scatter plot и линия тренда

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

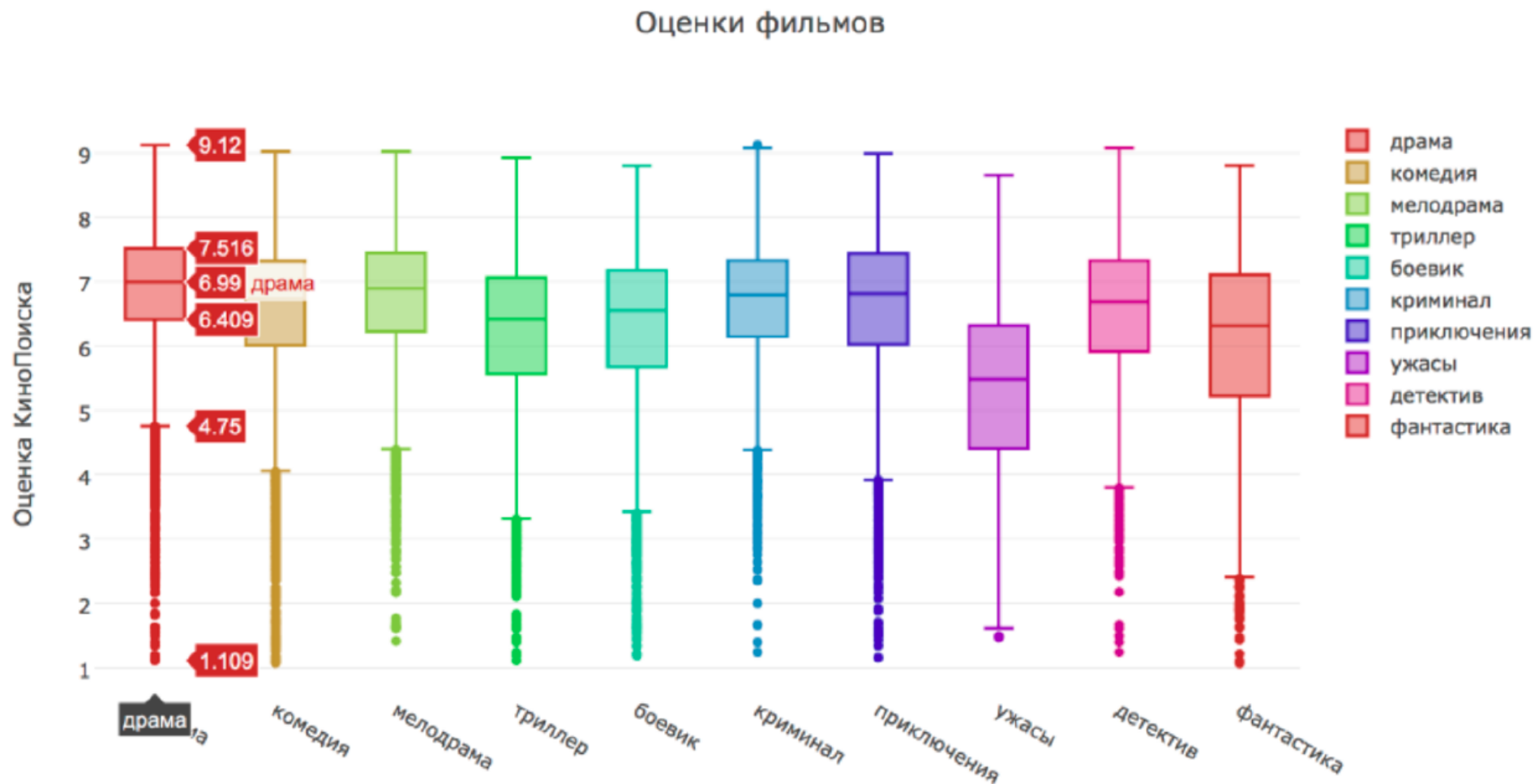


Есть и другие визуализации...

Table

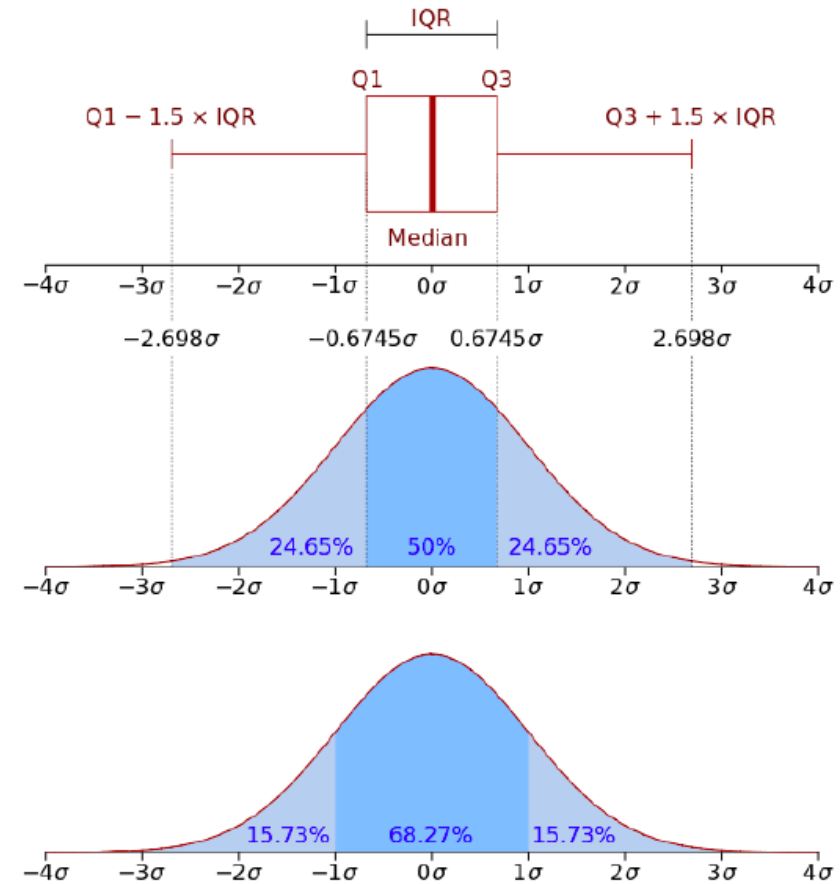
Region	Actual Sales (mn)		% to Goal	(12 Month)	Gross Profit (mn)	Profit Trend (12 Month)
Alabama	\$4,916		107%		\$1,172	
Alaska	\$3,110		65%		\$791	
Arizona	\$5,198		103%		-\$282	
Idaho	\$5,280		101%		\$410	
Illonois	\$4,956		93%		-\$22	
Indiana	\$5,032		91%		-\$516	
Ohio	\$5,566		112%		\$524	
Oklahoma	\$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

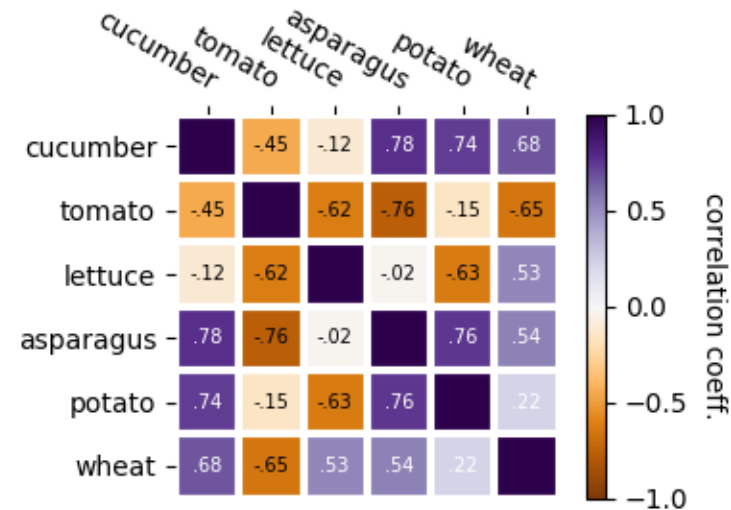
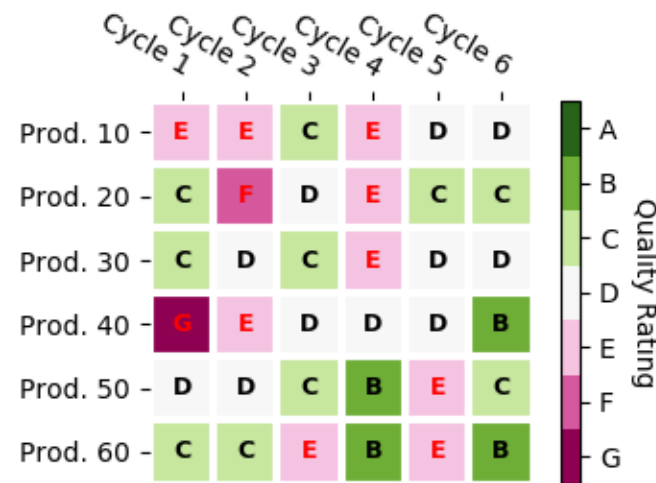
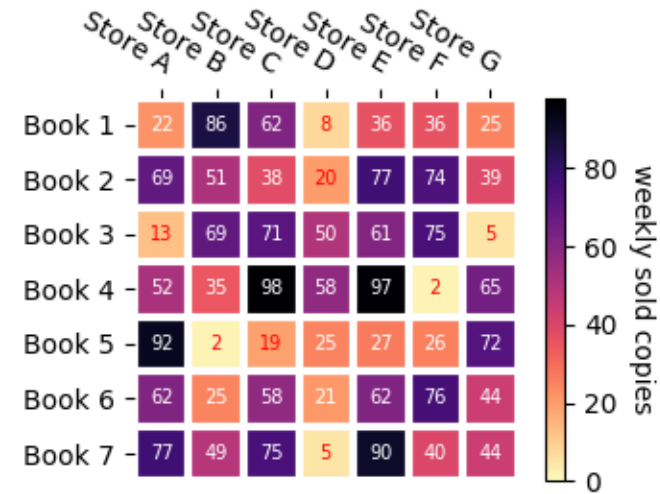
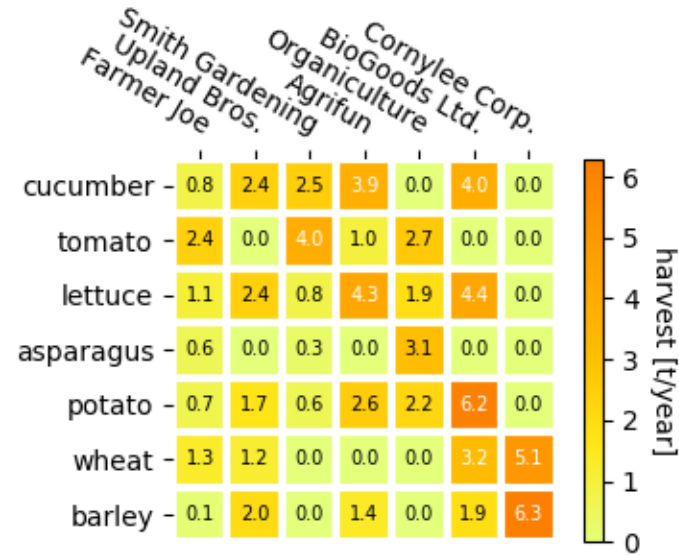


Box plot uncovered

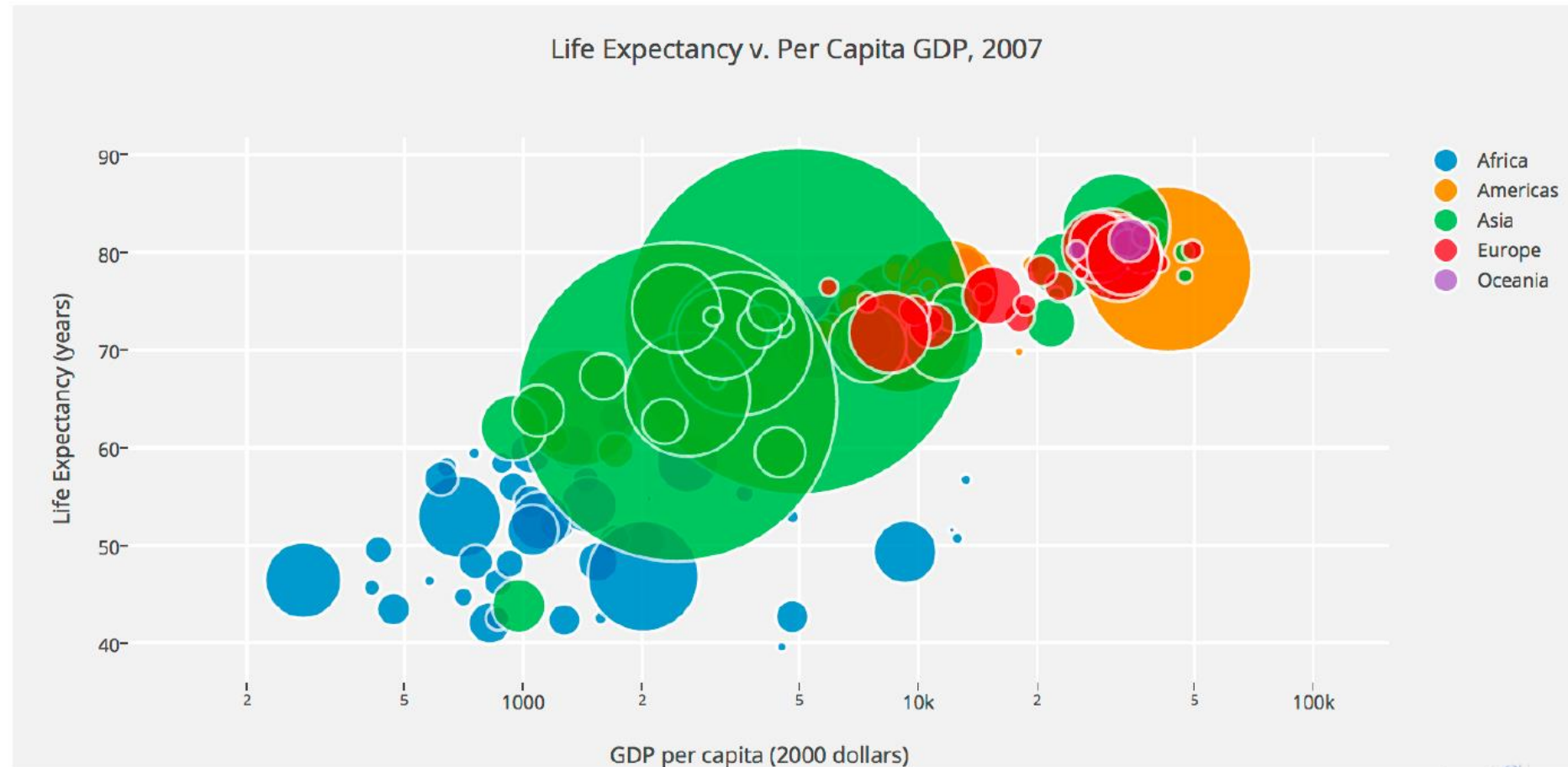
- › линия - медиана
- › коробка - IQR
- › усы - $[Q1 - 1.5IQR, Q3 + 1.5IQR]$
- › точки - outliers



Heatmap

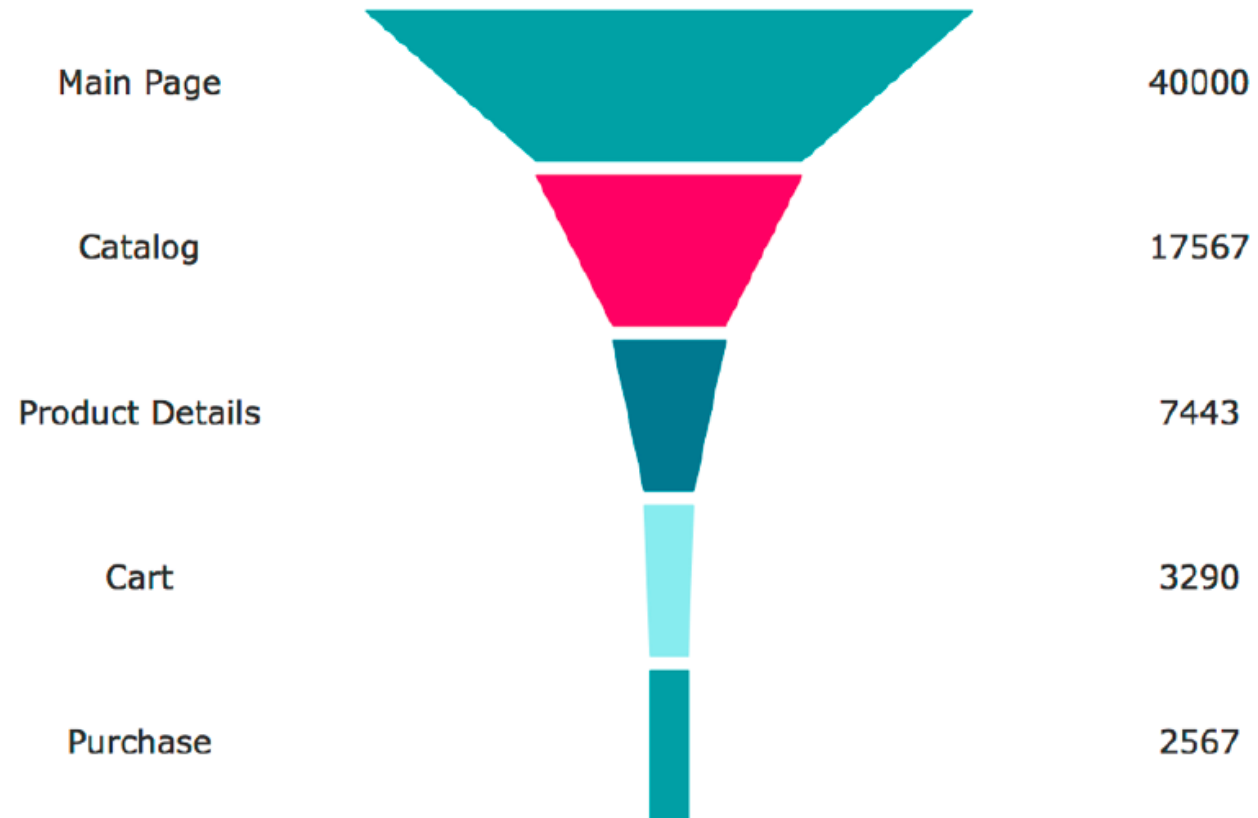


Bubble chart



Funnel chart

Funnel Chart



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

matplotlib

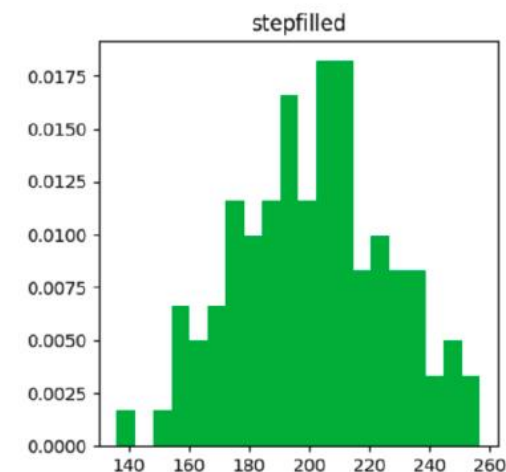
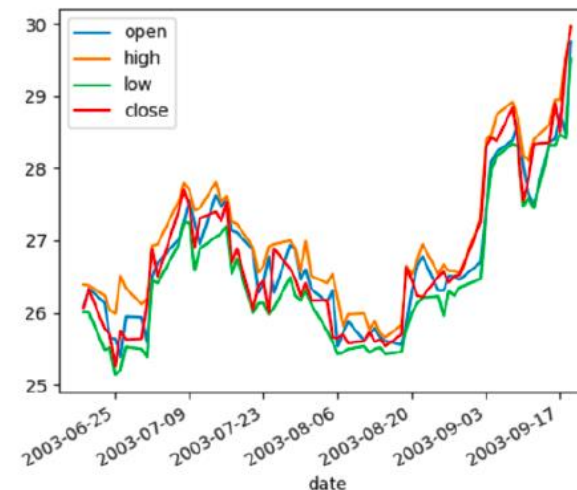


bokeh

Pygal

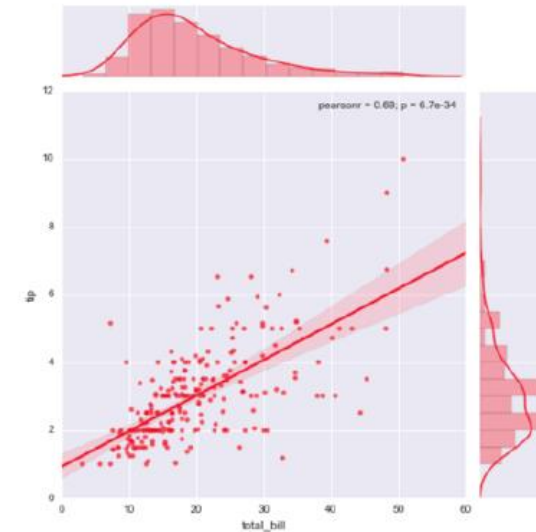
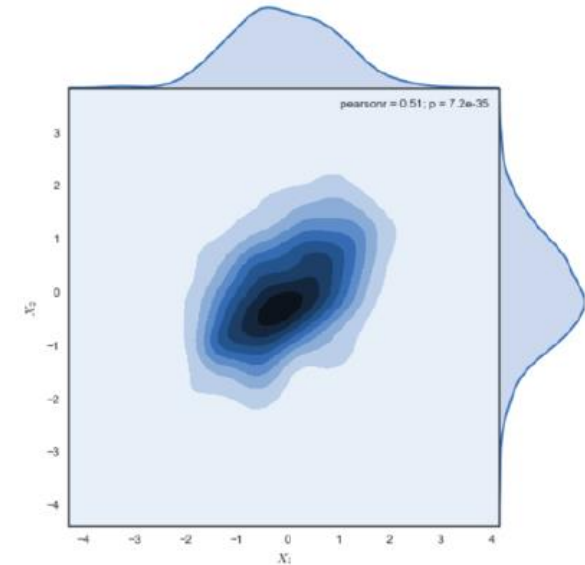
Matplotlib

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

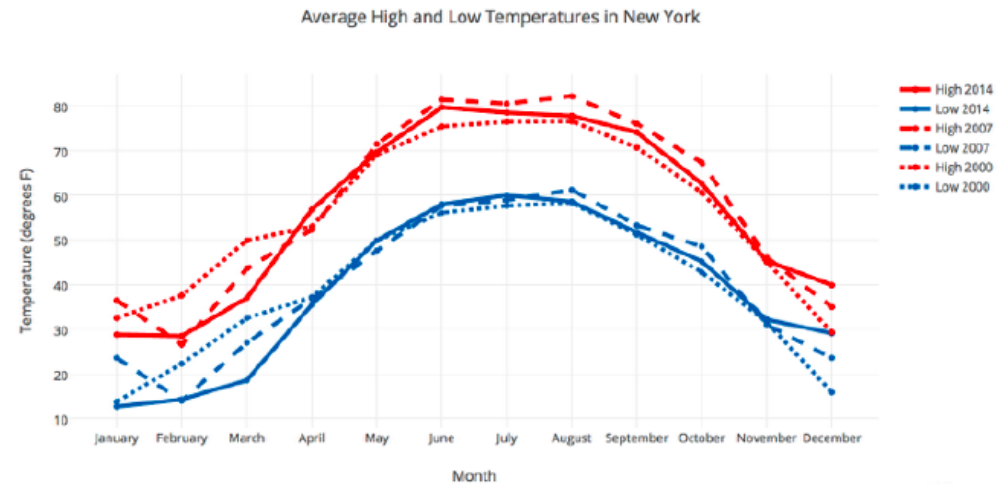
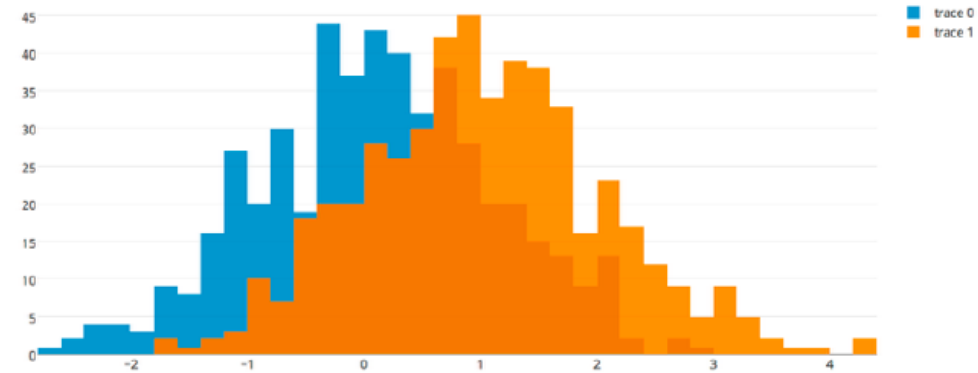


Seaborn

- › на основе matplotlib
- › сложные графики за пару строк кода
- › симпатичные default стили
- › для мелких изменений нужно изменять настройки matplotlib



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



Визуализация данных

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ