Основы визуализации данных в Python

библиотеки matplotlib и seaborn

Олег Булыгин

Python для анализа данных

Олег Булыгин

Lead Data scientist / Data analyst / developer, IT-тренер.

Аккаунты в соц.сетях

obulygin91@ya.ru

linkedin.com/in/obulygin

vk.com/obulygin91

@obulygin91

Визуализация данных

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

Очень важный инструмент в рамках EDA, который облегчает определение распределений, поиск аномалий, зависимостей, первичное выдвижение гипотез и пр.

А нужна ли визуализация вообще?

Квартет Энскомба

Квартет был составлен в 1973 году английским математиком Ф. Дж. Энскомбом для иллюстрации важности применения визуализации для статистического анализа и влияния выбросов на свойства набора данных.

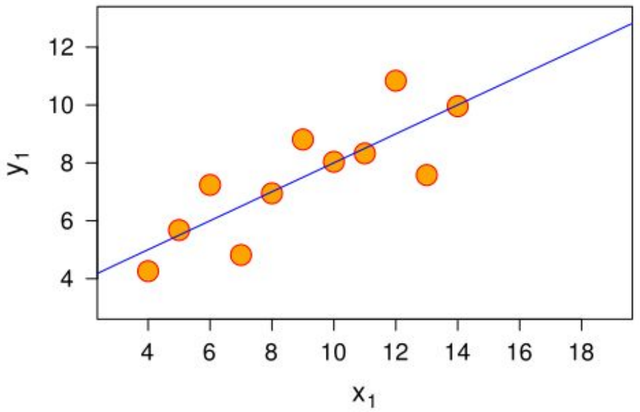
I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	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

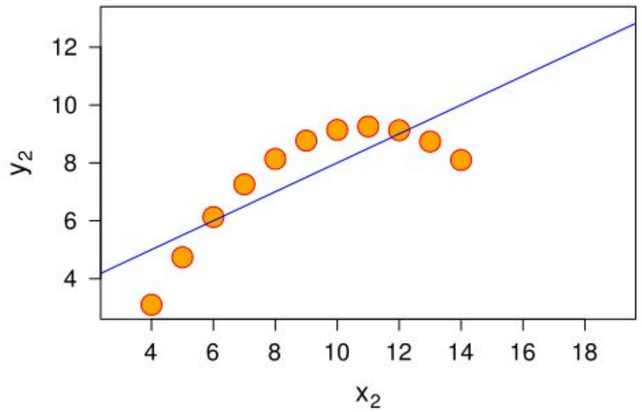
Характеристика	Значение
Среднее значение переменной x	9.0
Дисперсия переменной x	10,0
Среднее значение переменной y	7,5
Дисперсия переменной y	3,75
Корреляция между переменными x и y	0,816
Прямая линейной регрессии	$y = 3 + 0{,}5x$
Коэффициент детерминации линейной регрессии	0,67

Квартет Энскомба

Не смотря на то, что статистические свойства данных идентичны, их графики существенно отличаются.

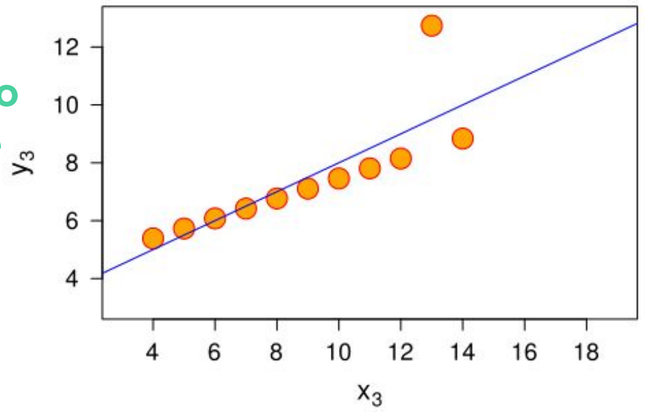
В первом наборе мы четко видим **прямую взаимосвязь** между **x** и **y**.

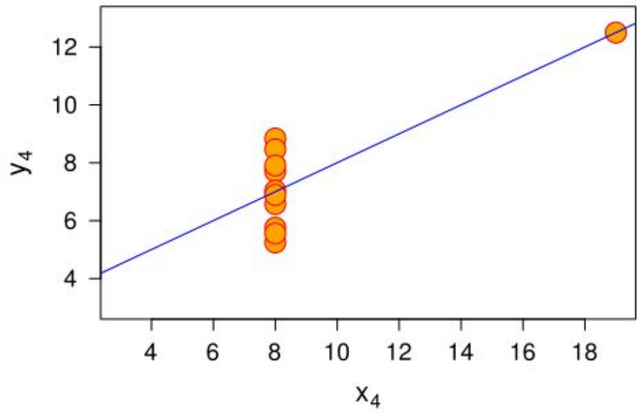




Во втором наборе мы можем сказать, что **линейной зависимости** в данных **нет**.

В третьем наборе мы видим **функциональную зависимость** и **наличие** одного **выброса**, который сильно искажает статистики.





В четвертом наборе **x** является **константой**, но присутствует **выброс**, который также влияет на все статистики.

Библиотеки Python для визуализации

matplotlib

Seaborn

plotly

Altair

bokeh

Pygal

и другие

Библиотека matplotlib

Matplotlib – одна из самых популярных библиотек Python для визуализации данных.

Импорт: **import matplotlib.pyplot as plt**

Документация: https://matplotlib.org/contents.html

Метод **.plot()** в pandas основан на matplotlib. Если мы работаем с датафреймами, то это, как правило, более удобный вариант, чем использование matplotlib самой по себе. Мы сконцентрируемся на этом варианте.

Документация:
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.html

Библиотека seaborn

Seaborn

Seaborn – популярная библиотека готовых шаблонов для статистической визуализации, написанная на основе matplotlib.

Имеет выразительный высокоуровневый интерфейс (построение большинства простых графиков происходит в одну строчку кода), а встроенные в нее стили более приятны.

Импорт: **import seaborn as sns** (библиотека названа в честь Сэмюела Нормана Сиборна (S.N.S) – героя сериала The West Wing, который очень любил автор библиотеки).

Документация: **https://seaborn.pydata.org/tutorial.html**

Какие типы визуализации вы знаете?

Метод .plot() в pandas

Аргумент **kind**, позволяет задать тип графика. Исчерпывающий список типов визуализации:

bar – столбчатая диаграмма;

barh – горизонтальная столбчатая диаграмма;

hist – гистограмма;

box – “ящик с усами”;

kde – ядерная оценка плотности;

area – диаграмма с областями накопления;

pie – круговая диаграмма

scatter – точечная диаграмма;

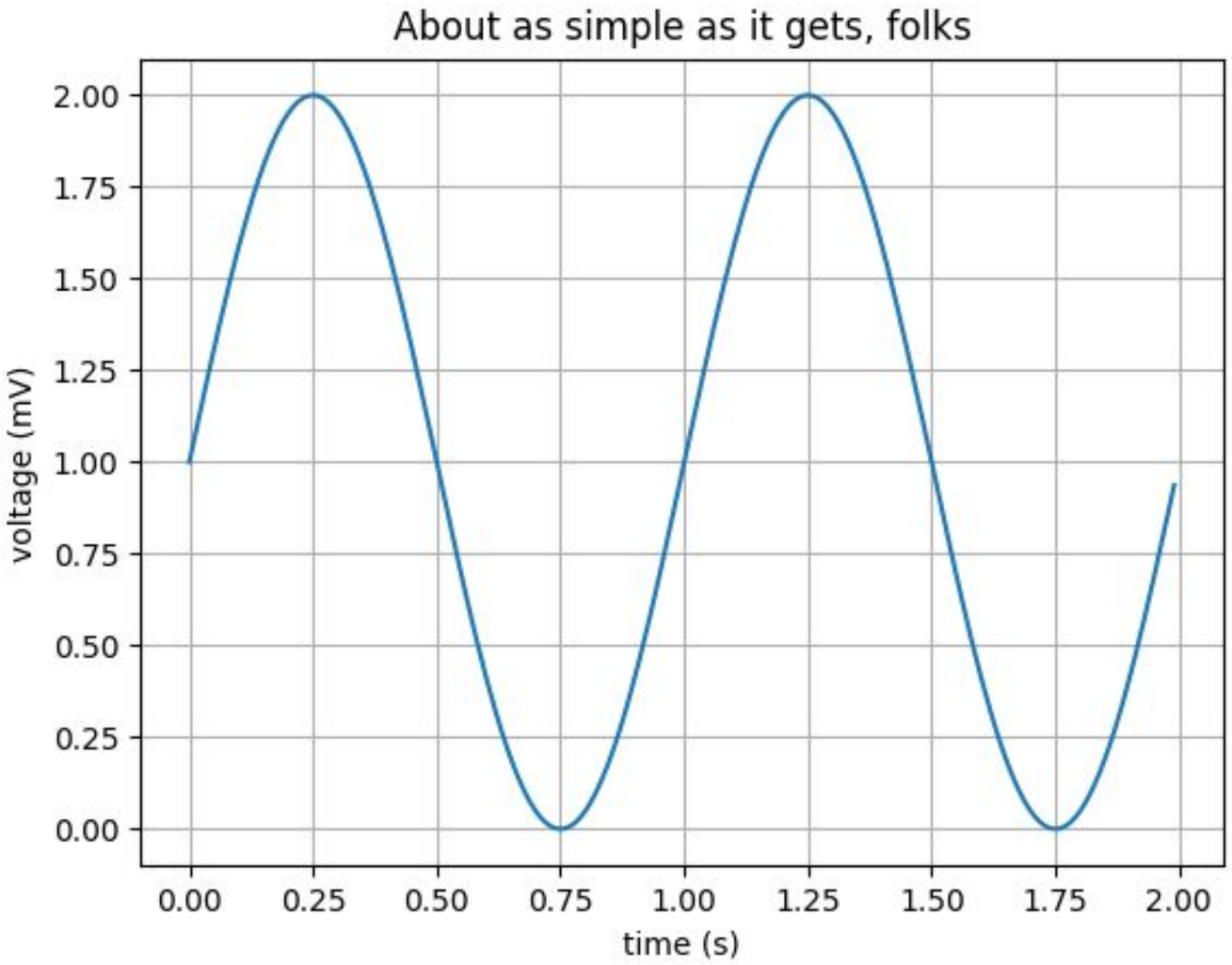
hexbin – гексагональная диаграмма.

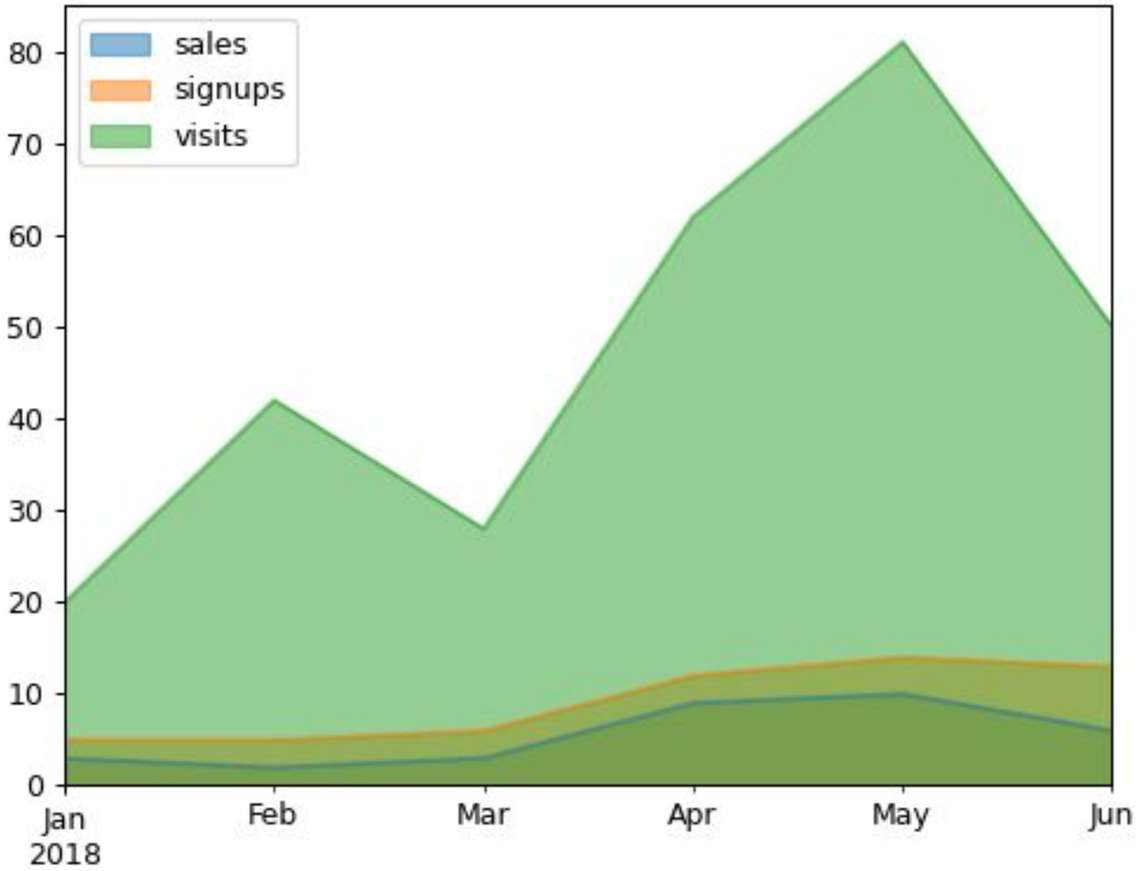
График
Line chart

– один из наиболее часто используемых типов визуализаций.

Отлично подходит, если:

- набор данных непрерывен;
- количество значений больше 20;
- необходимо выявить тенденцию.





Area chart

аналогична графику, но пространство между осью X и линией графика заполняется цветом или рисунком.

Лучше всего подходит для отображения изменений в составе комплексной величины с течением времени.

Если категорий больше 3-5, то ее использование вряд ли будет оправдано.

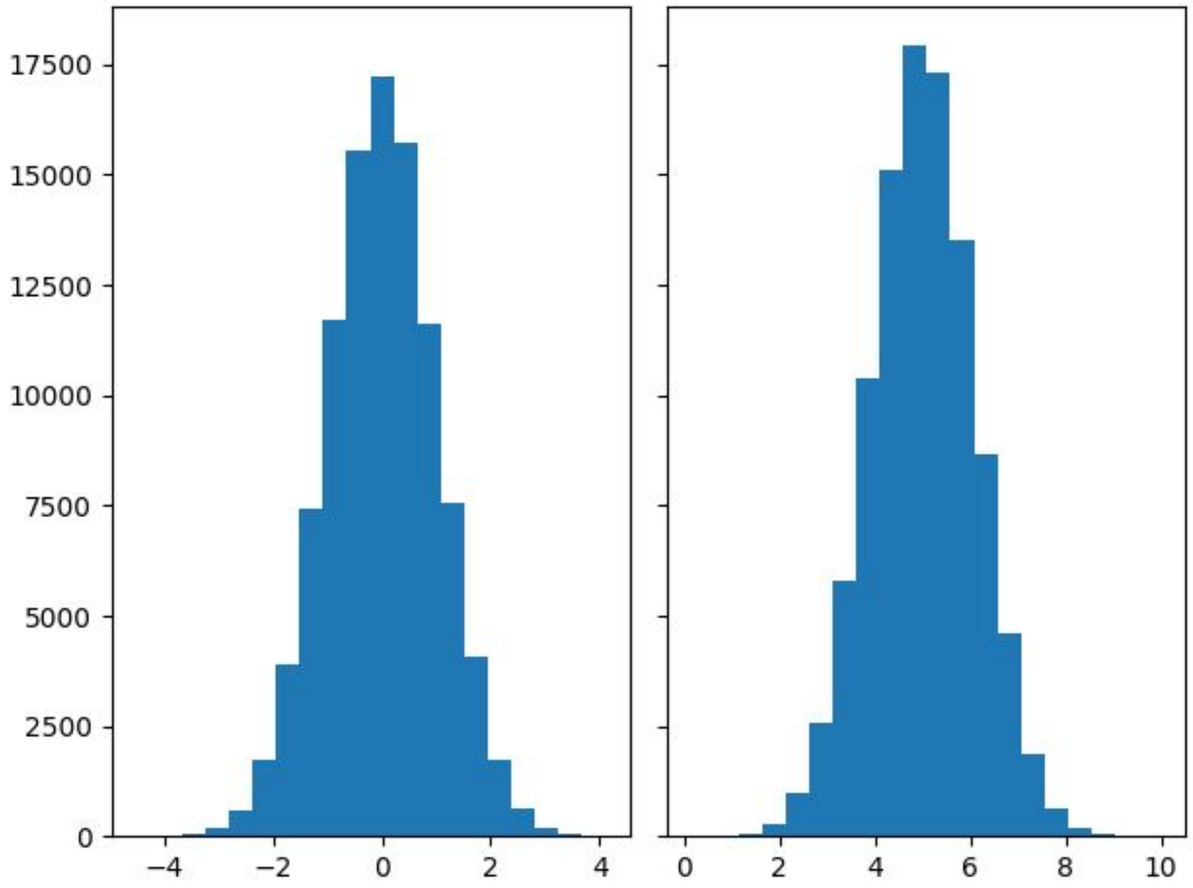
Гистограмма

Histogram

используется в статистике для представления распределения величины.

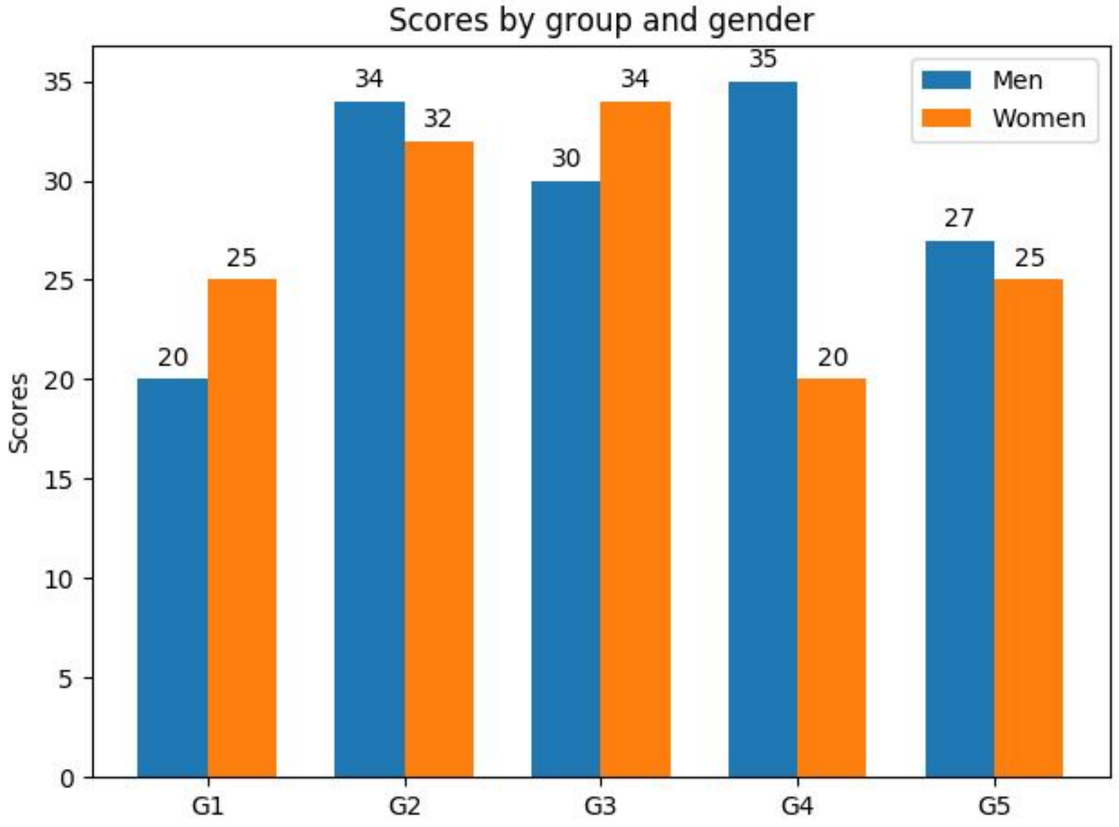
По горизонтальной оси откладывается диапазон наблюдаемых значений, разбитый на определенное число (обычно 10-15) интервалов, а по вертикальной – вероятность или частота ее попадания в каждый интервал.

По форме гистограммы аналитик может оценить, какому статистическому закону распределения подчиняется величина.



Столбчатая диаграмма

Bar chart



идеально подходит для сравнения показателей.

Горизонтальные столбчатые диаграммы обычно используют, когда нужно сравнить показатели между собой.

А вертикальные вариант хорошо подходит для демонстрации изменения показателя в разные периоды.

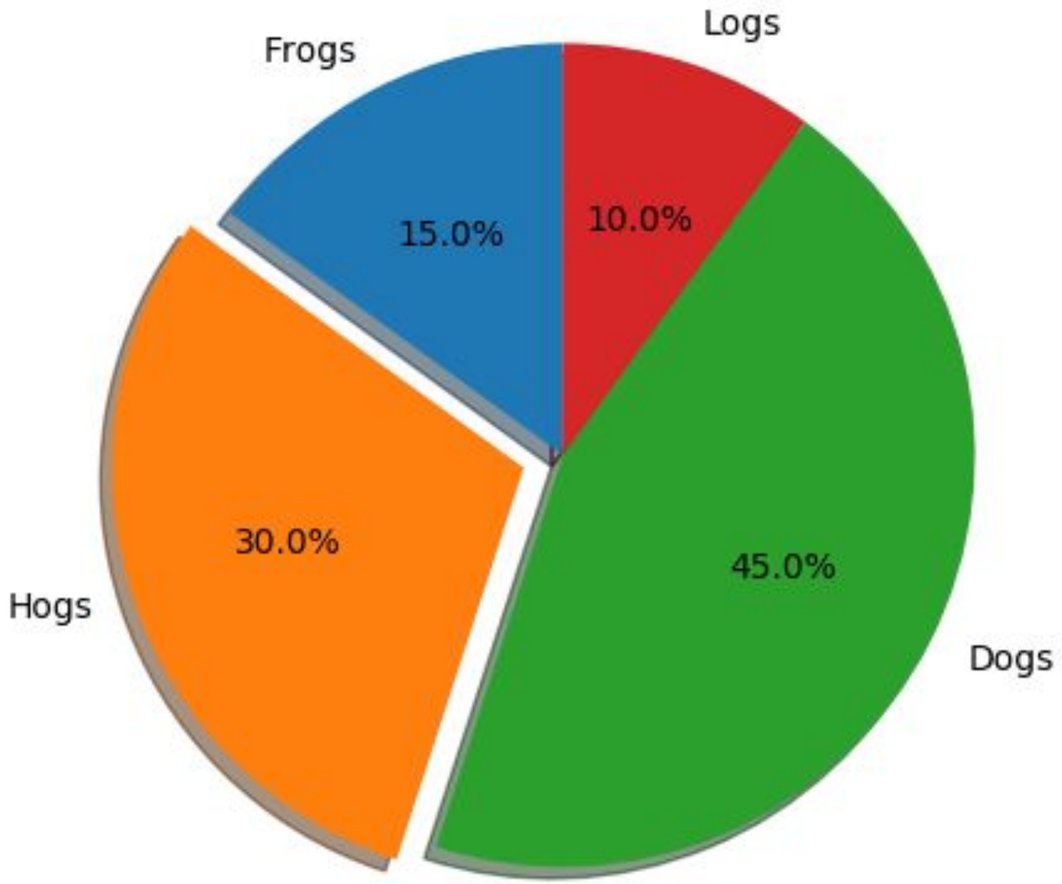
! Не путать с гистограммой

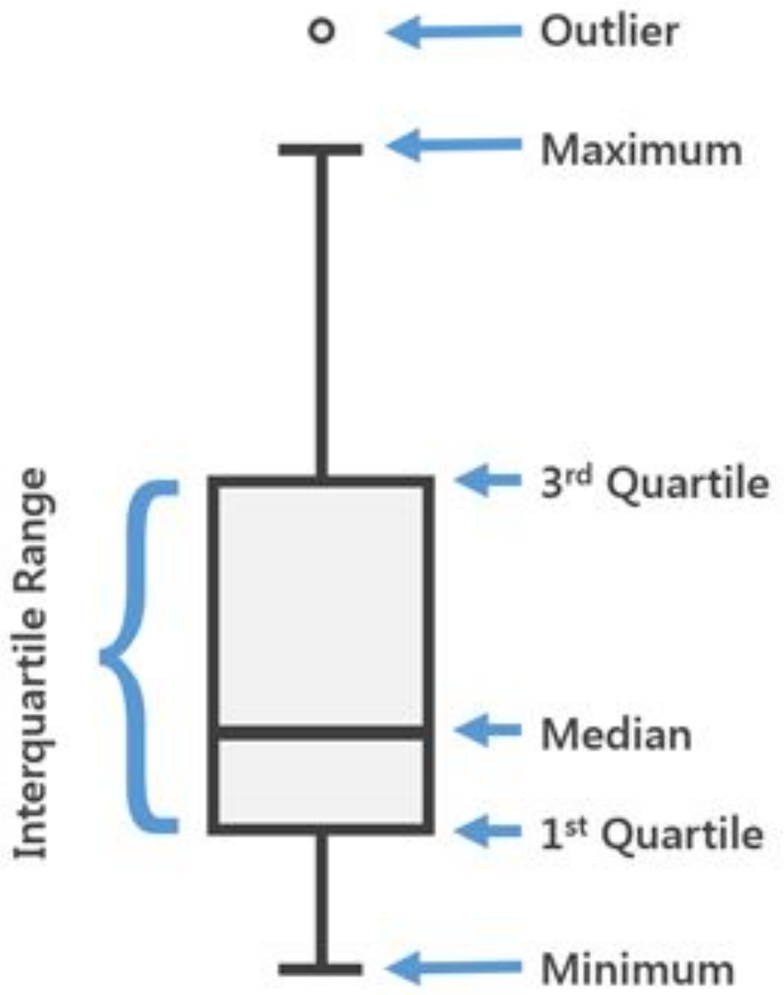
Круговая диаграмма
Pie chart

~~худшая диаграмма на свете~~
распространенный способ показать структуру набора данных (какую часть от общего количества составляют отдельные значения).

Круговые диаграммы не предназначены для сравнения отдельных категорий друг с другом.

По возможности избегайте их. Мы хорошо воспринимаем длины и размеры, но углы и площади нам воспринимать и сравнивать тяжело.





Ящик с усами, диаграмма размаха

Box plot

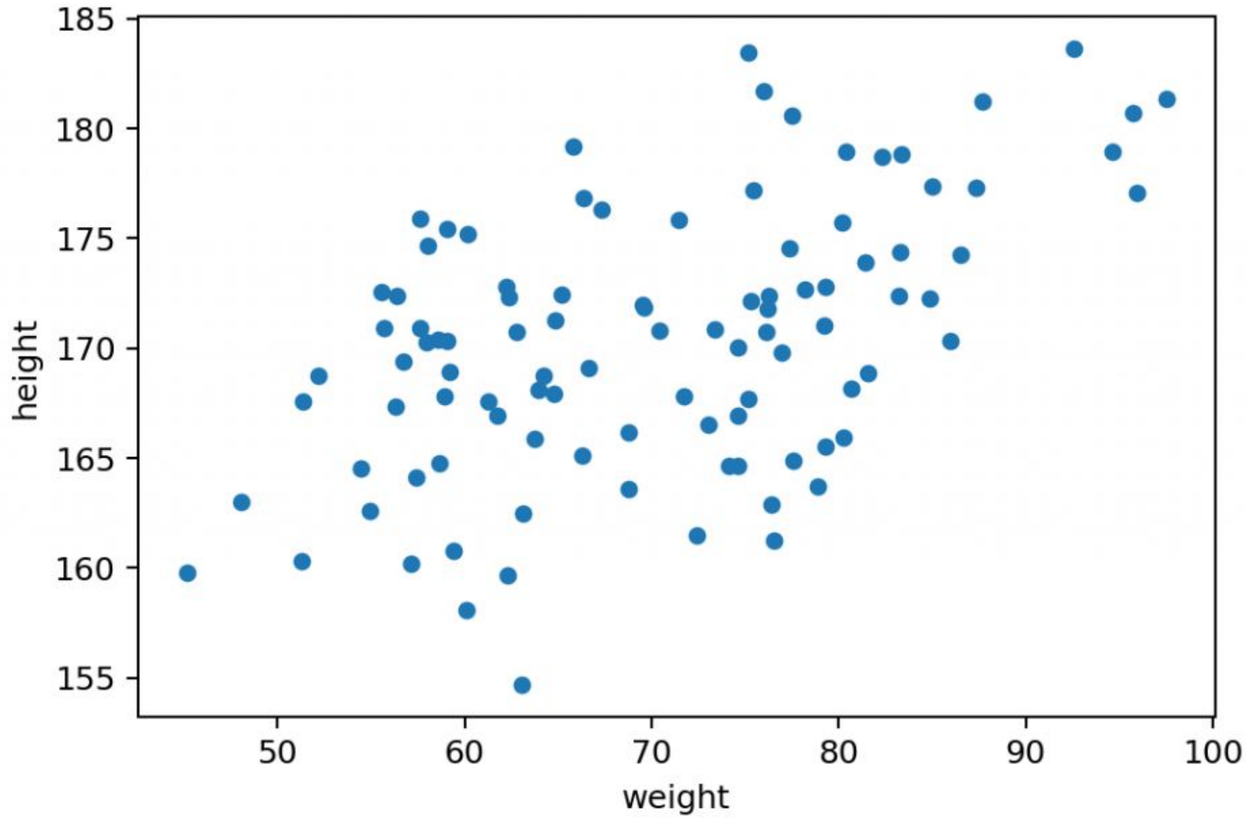
используется в описательной статистике, компактно изображает распределение величин. Одна из немногих визуализаций, позволяющая показать выбросы.

Точечная диаграмма

Scatter Plot / диаграмма рассеяния

Каждому наблюдению соответствует точка, координаты которой равны значениям двух параметров этого наблюдения.

Используются для изучения взаимосвязи между двумя переменными. Также помогают выявлять выбросы.

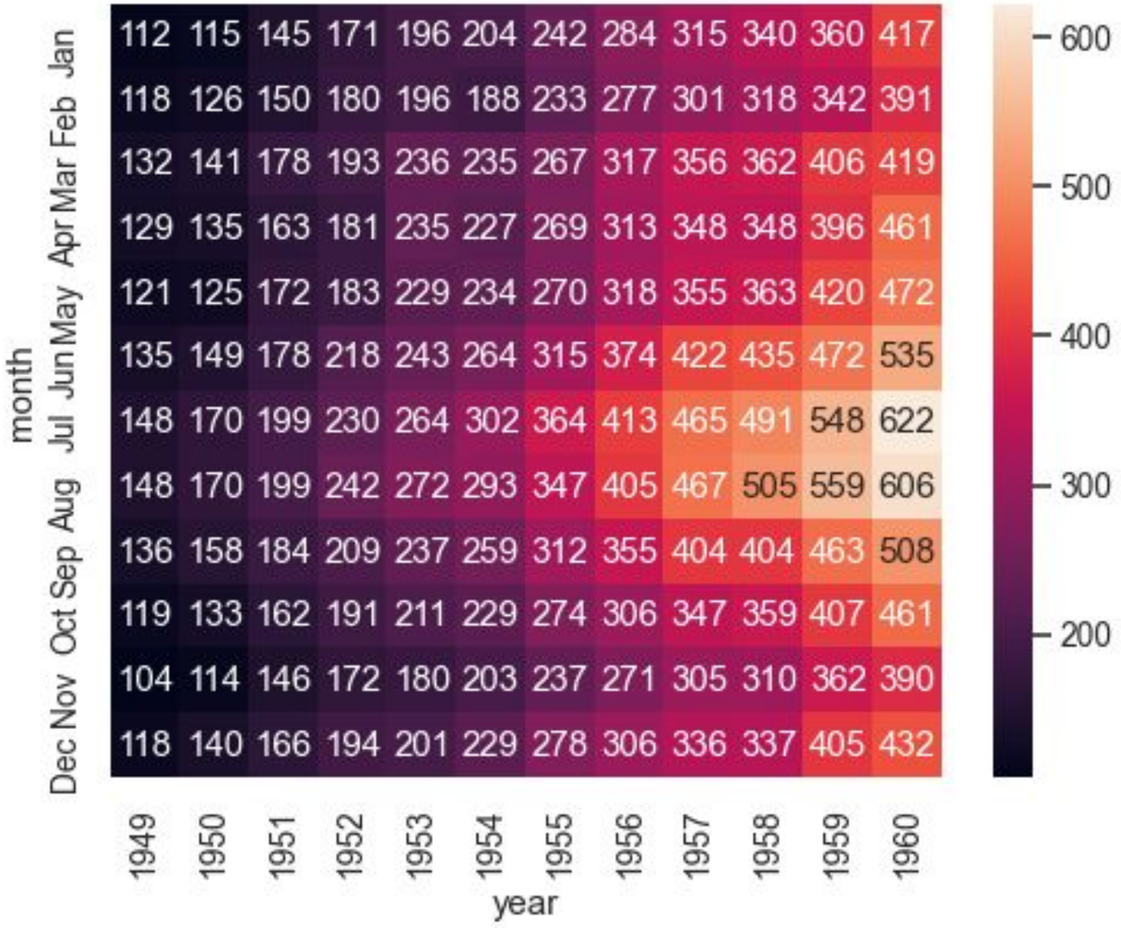


Стилизация таблиц

Когда построение визуализации неоправданно, можно стилизовать табличную структуру при помощи встроенных инструментов pandas, добавив нужные акценты для лучшего ее восприятия.

Документация:
https://pandas.pydata.org/pandas-docs/stable/user_guide/style.html

	A	B	C	D	E
0	1.000000	1.329212	nan	-0.316280	-0.990810
1	2.000000	-1.070816	-1.438713	0.564417	0.295722
2	3.000000	-1.626404	0.219565	0.678805	1.889273
3	4.000000	0.961538	0.104011	nan	0.850229
4	5.000000	1.453425	1.057737	0.165562	0.515018
5	6.000000	-1.336936	0.562861	1.392855	-0.063328
6	7.000000	0.121668	1.207603	-0.002040	1.627796
7	8.000000	0.354493	1.037528	-0.385684	0.519818
8	9.000000	1.686583	-1.325963	1.428984	-2.089354
9	10.000000	-0.129820	0.631523	-0.586538	0.290720



Тепловая карта

Heat Map

– это матричное представление данных, в котором каждое значение отображается при помощи определенного цвета.

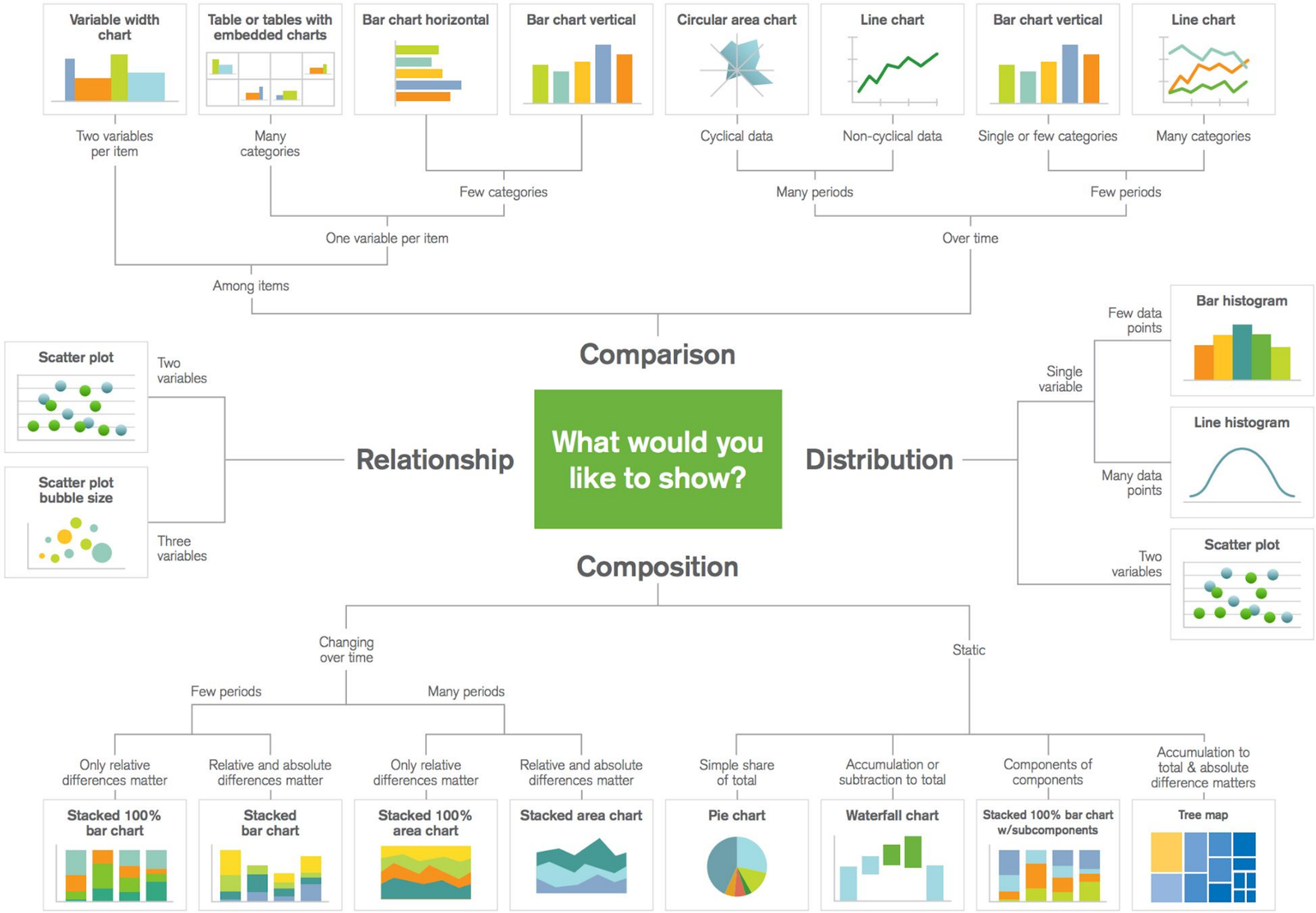
Хорошо показывают связи нескольких переменных между собой.

Как выбирать тип визуализации?

Различных визуализаций очень много, как выбрать подходящую?

Источники, которые помогают выбрать тип визуализации

[DataVizCatalogue](#)
[ExtremePresentation](#)



А так делать не нужно:
https://t.me/awfulcharts

Основы визуализации данных в Python

Вопросы?

Олег Булыгин

Python для анализа данных

нетология