Содержание

Визуализация данных

Зачем нужна визуализация?

Как можно визуализировать данные?
- График / line chart
- График с площадью под ним / area chart
- Гистограмма / histogram
- Столбчатая диаграмма / bar chart
- Круговая диаграмма / pie chart
- Ящик с усами (Коробчатая диаграмма) / box plot
- Диаграмма рассеяния / scatter plot
- Тепловая карта / heat map
- Стилизация таблиц

Правила визуализации данных

Визуализация данных

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

> **Простой график привнёс больше информации в сознание аналитика данных, чем любое устройство**
>
> Джон Тьюки

Зачем нужна визуализация?

Визуализация — один из инструментов EDA / Exploratory Data Analysis, который позволяет:

- определить распределения данных
- найти аномалии
- найти зависимости
- сделать первичное выдвижение гипотез и пр.

Одним из лучших примеров, который объясняет необходимость визуализации — Квартет Энскомба. Что он из себя представляет?

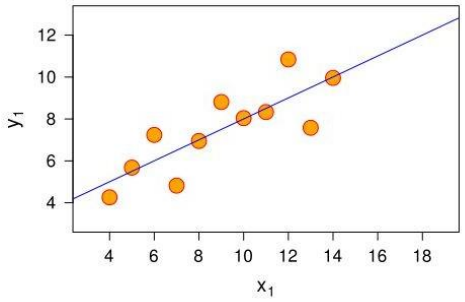
I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0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

Это четыре набора числовых данных (датасетов). В каждом наборе 11 пар чисел.

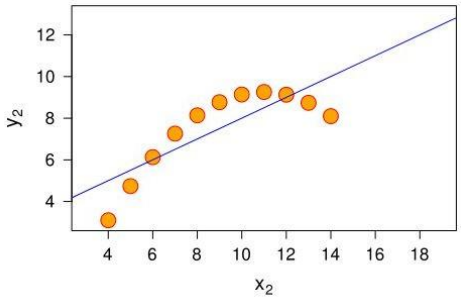
Характеристика	Значение
Среднее значение переменной x	9.0
Дисперсия переменной x	10.0
Среднее значение переменной y	7.5
Дисперсия переменной y	3.75
Корреляция между переменными x и y	0.816
Прямая линейной регрессии	y = 3 + 0.5x
Коэффициент детерминации линейной регрессии	0.67

В датасетах содержатся разные числа, при этом простые статистические свойства у них идентичны.

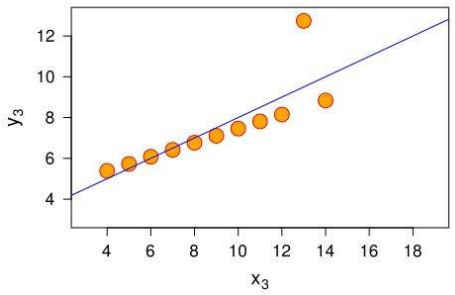
Но если те же самые данные представить в виде графиков, сразу станет понятно, что датасеты представляют разные зависимости.



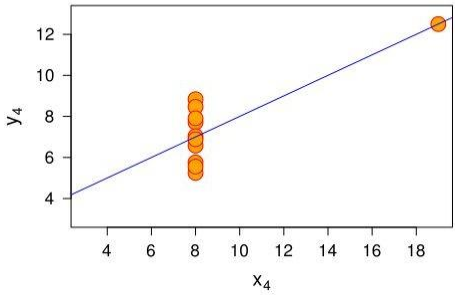
y1: приближается как линейная зависимость, без сильных выбросов



y2: нелинейная зависимость без сильных выбросов



y3: приближается линейно, но с наличием одного сильного выброса



y4: константа с сильным выбросом, причём константой является x4 по отношению к y4

Источник иллюстраций

Как можно визуализировать данные?

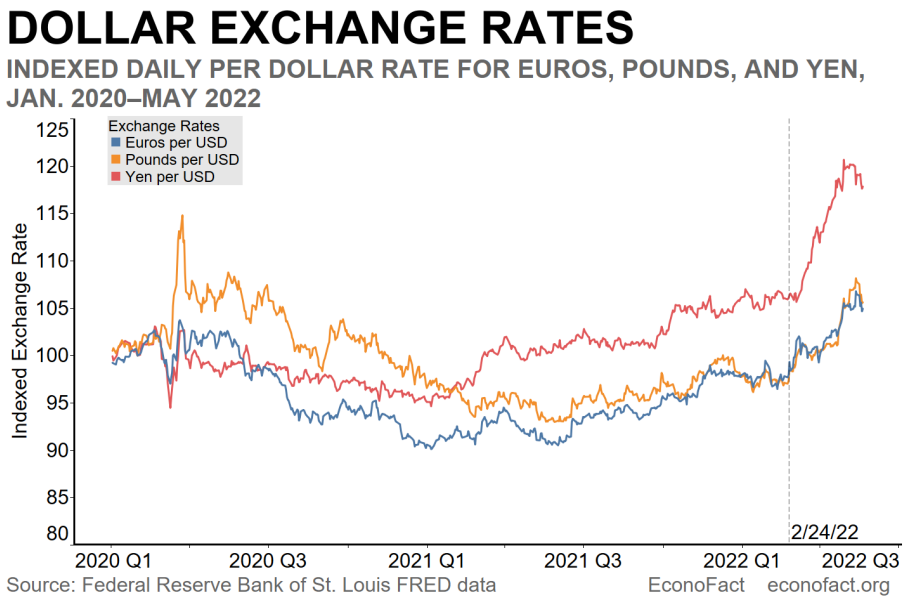
Эффективность визуализации зависит от правильности её применения. Особенно важно грамотно выбрать тип графика или диаграммы, который ты будешь использовать.

➔ График / line chart

Один из наиболее часто используемых типов визуализаций.

Используй, если:

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

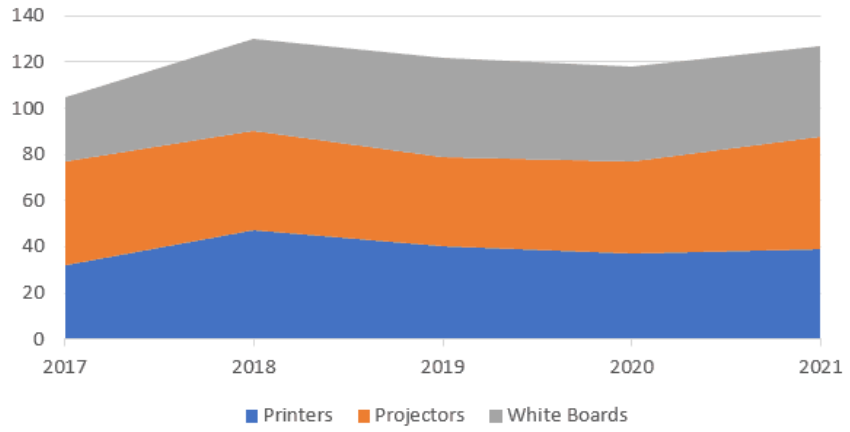


Источник иллюстрации

➔ График с площадью под ним / area chart

Пространство между осью X и линией графика заполняется цветом.

Используй, чтобы показать изменения в составе комплексной величины с течением времени. Не используйте, если категорий больше 3–5.



Источник иллюстрации

➔ Гистограмма / histogram

Часто используется для анализа изображений.

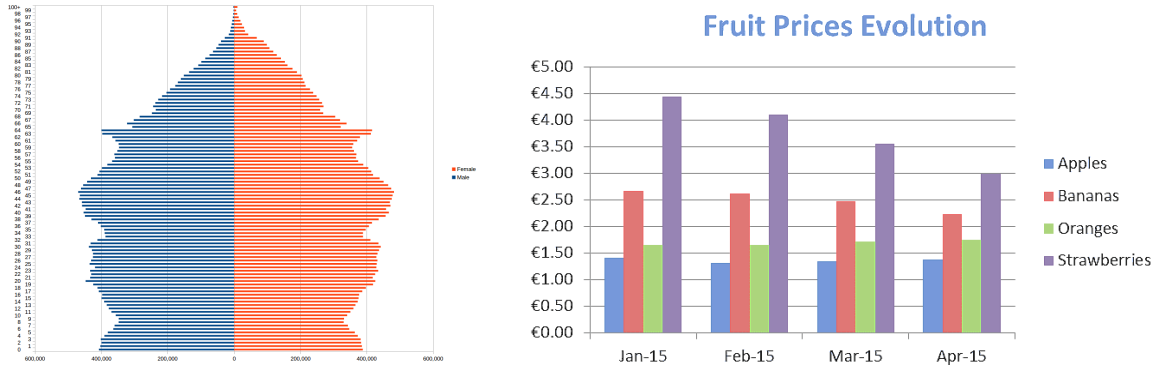
По форме гистограммы можно оценить закон распределения данных. По горизонтальной оси откладывается диапазон значений, разбитый на 10–15 интервалов. По вертикальной — вероятность или частота её попадания в каждый интервал.

Источник иллюстрации

➔ Столбчатая диаграмма / bar chart

Используй для сравнения показателей.

Горизонтальные столбчатые диаграммы – для сравнения показателей между собой
Вертикальные – для демонстрации изменения показателя в разные периоды.

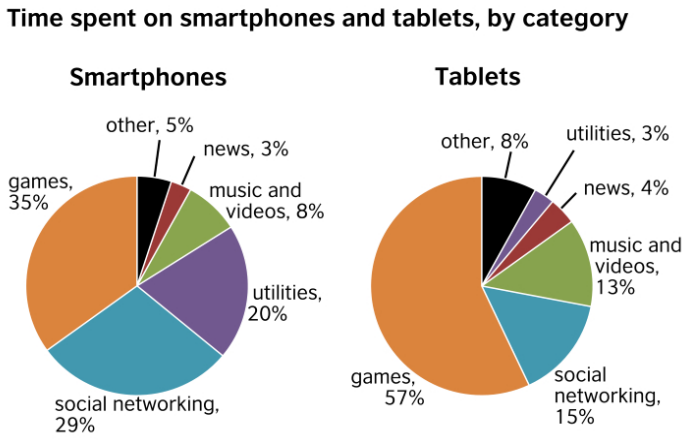


Источники иллюстрации 1, Источник иллюстрации 2

➔ Круговая диаграмма / pie chart

Способ показать, какую часть от общего количества составляют отдельные значения. Круговые диаграммы не предназначены для сравнения категорий друг с другом.

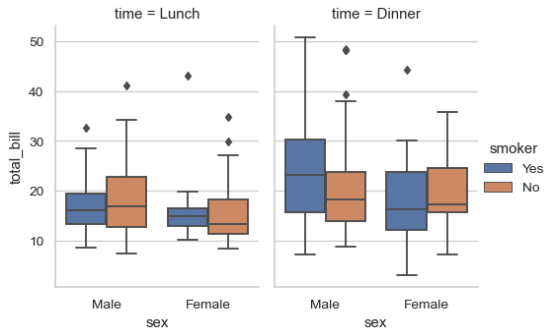
Лучше не используй в работе, потому что углы и площади сложны для восприятия.



Источник иллюстрации

➔ Ящик с усами (Коробчатая диаграмма) / box plot

Компактно изображает распределение величин. Одна из немногих визуализаций, позволяющих показать выбросы.

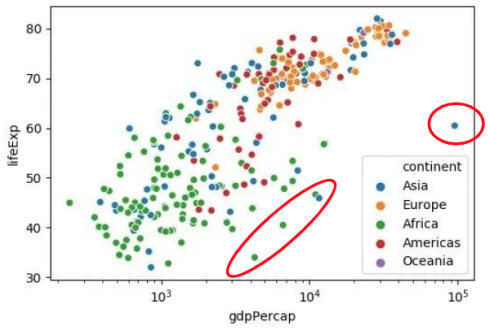


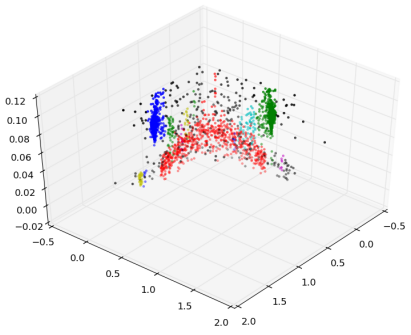
Источники иллюстрации 1

➔ Диаграмма рассеяния / scatter plot

Каждому наблюдению на графике соответствует точка. Координаты точек равны значениям двух параметров наблюдений.

Используются для изучения взаимосвязи между двумя переменными. Помогают выявлять выбросы.



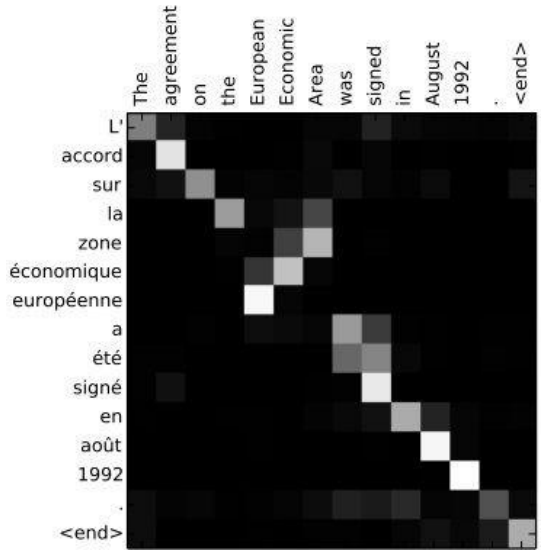


Источники иллюстрации 1, Источник иллюстрации 2

➔ Тепловая карта / heat map

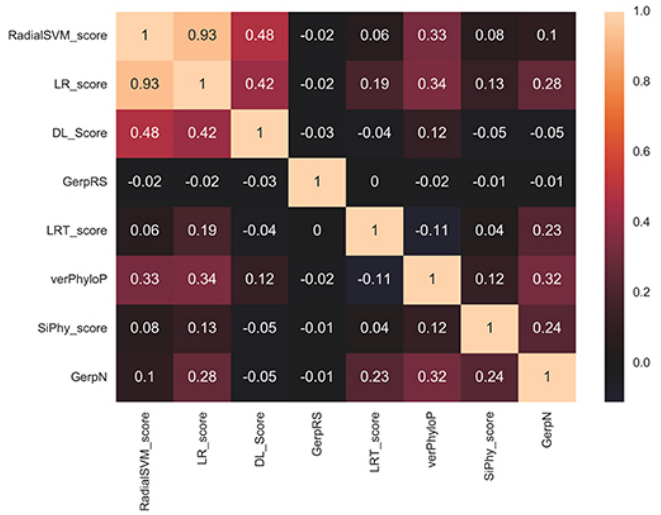
Матричное представление данных. Каждое значение на карте отображается при помощи определённого цвета.

Тепловые карты хорошо показывают связи нескольких переменных.



Источник иллюстрации

Этот вид визуализации часто используется для того, чтобы подсветить «куда смотрит нейросеть». Он помогает визуализировать матрицы корреляции, которые накопились
в весах сети при обучении.



Источник иллюстрации

➔ Стилизация таблиц

Когда построение визуализации неоправданно, можно стилизовать табличную структуру при помощи встроенных инструментов Pandas.

```
[24]: df.style.bar(subset=['A', 'B'], align='mid', color=['#d65f5f', '#5fba7d'])
```

[24]:

	A	B	C	D	E
0	1	1.32921	nan	-0.31628	-0.99081
1	2	-1.07082	-1.43871	0.564417	0.295722
2	3	-1.6264	0.219565	0.678805	1.88927
3	4	0.961538	0.104011	-0.481165	0.850229
4	5	1.45342	1.05774	0.165562	0.515018
5	6	-1.33694	0.562861	1.39285	-0.063328
6	7	0.121668	1.2076	-0.00204021	1.6278
7	8	0.354493	1.03753	-0.385684	0.519818
8	9	1.68658	-1.32596	1.42898	-2.08935
9	10	-0.12982	0.631523	-0.586538	0.29072

Источники иллюстраций

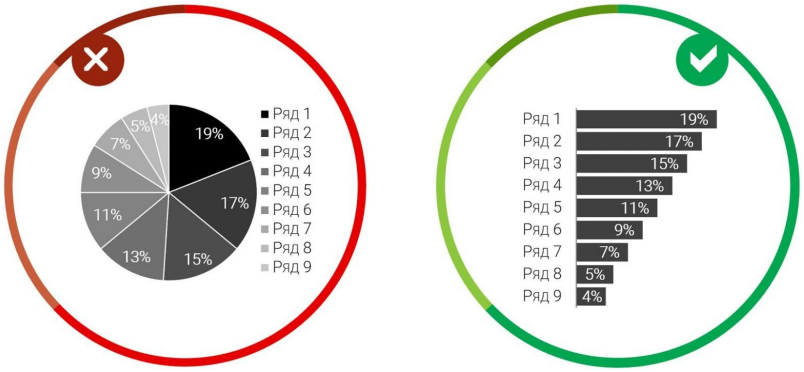
Также нужно добавить акценты для лучшего восприятия.

Правила визуализации данных

Визуализация данных также помогает структурировать их и сделать более доступными для восприятия. Но чтобы достичь этой цели, важно следовать определённым правилам.

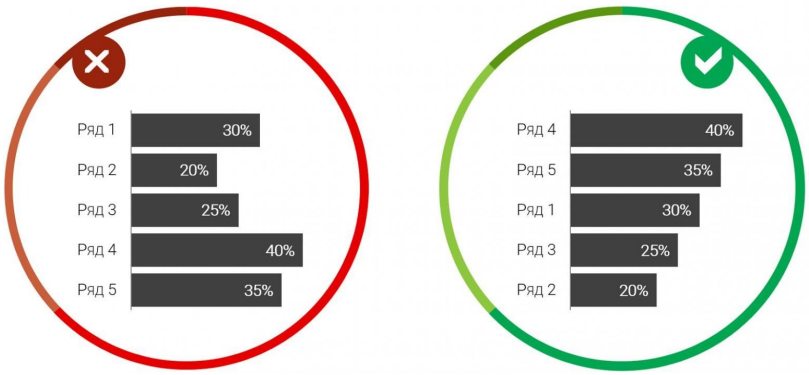
Вот основные из них:

1. Покажи сравнение
2. Подсвети причины и построй гипотезы
3. Покажи многомерные данные
4. Опиши и задокументируй график
5. Используй правильный тип графика
6. Визуализируй так, чтобы было легко сравнивать (не pie-чарты)



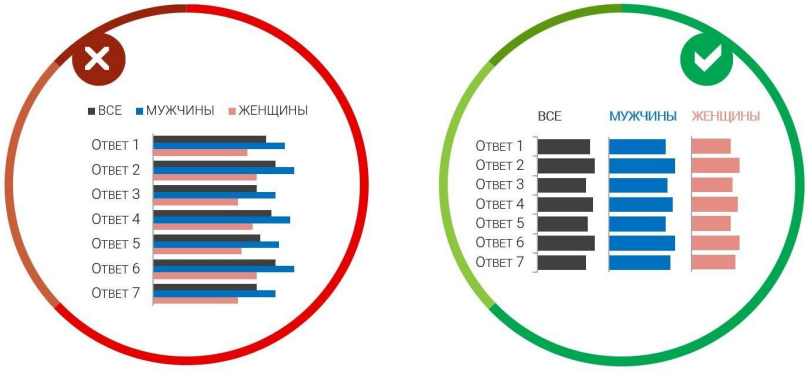
Источник иллюстрации

7. Используй простой дизайн
8. Выстрой удобоваримый порядок



Источник иллюстрации

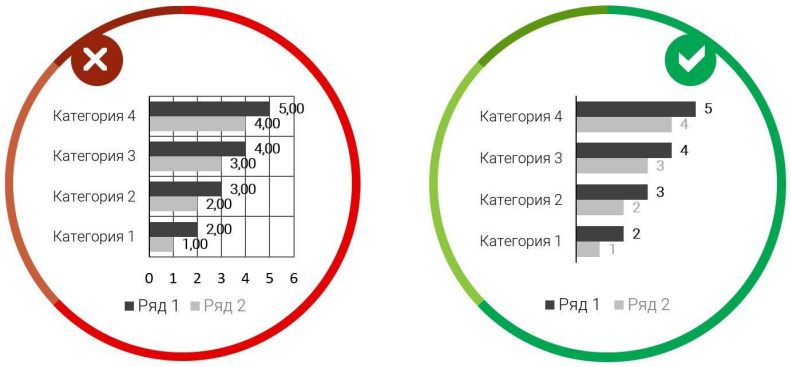
9. Используй необходимый минимум элементов и надписей
10. Указывай название и полную легенду



Источник иллюстрации: orgcomnet.ru

11. Не перегружай информацией

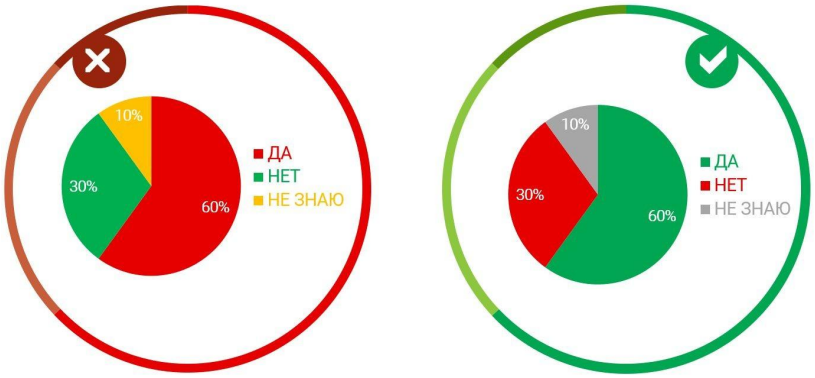
12. Используй легко читаемый формат чисел



Источник иллюстрации

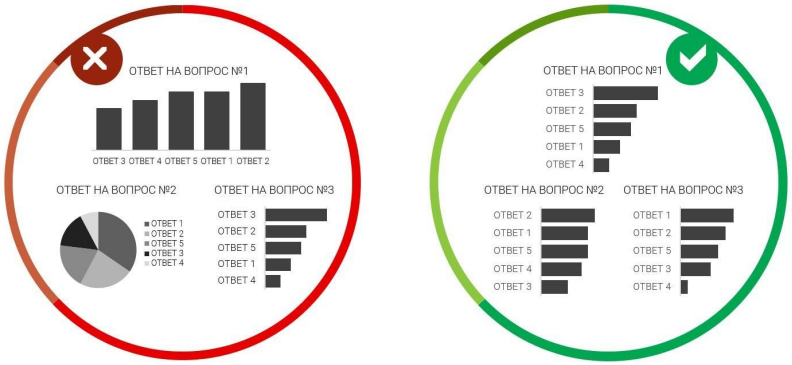
13. Используй общепринятые цветовые решения

14. Делай всё в единой цветовой палитре



Источник иллюстрации

15. Не используй много разных типов диаграмм



Источник иллюстрации