SkillFactory

8. Искусство визуализации

✍ Познакомившись с потрясающим миром визуализации, вы, должно быть, поняли, что код для построения графиков довольно простой и не требует специализированных навыков, документация ко всем библиотекам максимально понятная и наполнена примерами. Построение графиков главным образом заключается в настройке параметров их отображения.

→ На самом деле всё не так просто. Помимо изучения различных типов графиков, нужно знать, для чего эти графики применяются, на каких данных их правильнее строить и как не потерять информативность, используя их.

В данном юните мы приведём несколько общих рекомендаций по тому, как должны выглядеть графики, а также рекомендации по визуализации для основных типов графиков.

ОБЩИЕ РЕКОМЕНДАЦИИ К СОЗДАНИЮ ВИЗУАЛИЗАЦИИ

1. Первое и самое главное правило — **график должен быть информативным и понятным**. То есть любой человек, взглянув на ваш график, должен понять, что на нём изображено.

   Для этого всегда подписывайте оси графика или сам график, делайте интервалы между отметками на осях, не используйте слишком много графиков на одной координатной плоскости.

2. **Одна плоскость — один вид графика**. Не стоит смешивать типы визуализации, это делает результат нечитабельным.

3. **Принцип минимализма: чем проще график, тем лучше** — не нужно добавлять сглаживающие кривые, многочисленные подписи, лишние отметки на осях, яркие, отвлекающие внимание цвета, если это не помогает вам донести идею.

4. **Если тип значений всего один, легенда не нужна**.

5. **Не используйте сложный дизайн там, где это не требуется**. Прежде чем строить для презентации *3D*-график, подумайте, можно ли обойтись без него.

6. **Если у вас несколько диаграмм, используйте единую цветовую гамму**.

7. **Время всегда указывается по горизонтальной оси слева направо**. Размещайте отметки времени так, чтобы они не сливались друг с другом (например, под углом 45 градусов).

8. **При построении графиков необходимо отталкиваться от их предназначения**: например, не надо строить линейный график на категориальных данных — он не предназначен для этого.

РЕКОМЕНДАЦИИ ПО ИСПОЛЬЗОВАНИЮ ПОПУЛЯРНЫХ ТИПОВ ДИАГРАММ

Навык визуализации предполагает правильный выбор диаграммы и умение её преподнести. Давайте на примерах рассмотрим, как правильно (*и неправильно тоже*) строить основные виды графиков.

СТОЛБЧАТАЯ ДИАГРАММА

Столбчатая диаграмма (*bar plot*) — один из самых популярных видов графиков, однако не все исследователи используют его правильно.

> → Столбчатая диаграмма используется, когда необходимо сравнить какой-то показатель (количество, среднее, медиану) в зависимости от категориального признака (возможен вариант сравнения по датам). Это часто помогает понять, каково соотношение категорий, какая категория доминирующая. Всё это можно учитывать при построении модели, чтобы делать прогнозы более точными.

СОВЕТЫ

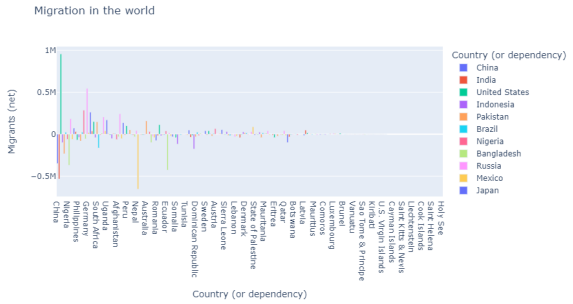
- Не сравнивайте больше десяти категорий друг с другом с помощью столбчатой диаграммы — вам будет сложно ориентироваться в ней, особенно если диаграмма многоуровневая.

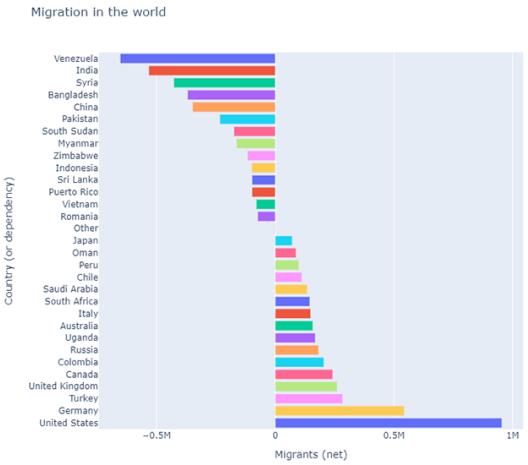
  Если категорий слишком много, можно использовать иерархическую диаграмму или попробовать отразить диаграмму горизонтально (поменять параметры *x* и *y* местами).

  Ещё один вариант — объединить непопулярные категории в общую категорию «другие».

- Начало точки отсчёта для значений показателя всегда 0.

  Если у вас есть значения меньше 0, используйте горизонтальную столбчатую диаграмму.





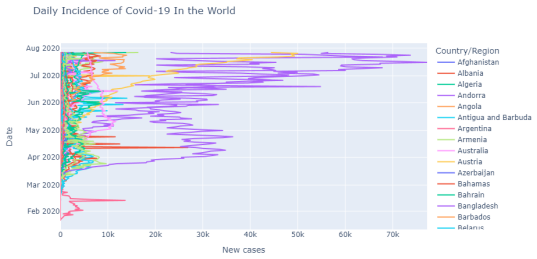
ЛИНЕЙНЫЙ ГРАФИК

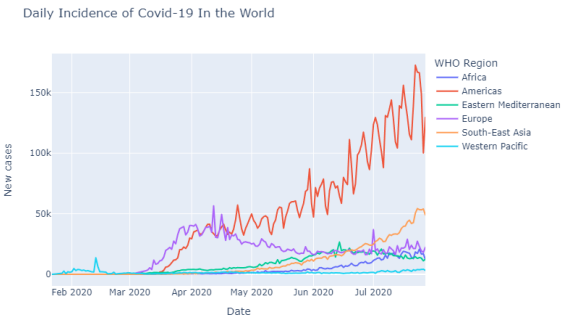
→ Линейный график (*linear plot*) отлично подходит, если набор данных непрерывен (как мы уже видели раньше, обычно это временной ряд). График используется для определения тенденций во временном ряду и сравнения нескольких рядов между собой.

СОВЕТЫ

- Не используйте график, если набор данных дискретный (менее 20 наблюдений) — в таком случае лучше воспользуйтесь столбчатой диаграммой.
- Время всегда отображается по оси абсцисс и разбивается на равные интервалы.
- Если даты сливаются, используйте наклон в 45 градусов.
- Не используйте график для сравнения рядов, если их больше 7-10 — график станет нечитабельным. Попробуйте уменьшить число категорий.

  Например, если вам необходимо сравнить графики продаж по странам (а их у вас около 200), добавьте признак региона (например, *Азия*, *Европа*, *Северная Америка* и т. д.) и сравнивайте временные ряды в его разрезе. Когда вы выявите лидера по продажам среди регионов, можете вернуться к сравнению стран внутри этого региона.





ГИСТОГРАММА

→ Гистограммы (*histogram*) часто применяются для разведывательного анализа данных (*EDA*), так как они дают информацию о распределении признака. С их помощью можно сразу определить диапазон изменения признака, его модальное значение (пик гистограммы), а также найти «пеньки», которые выбиваются от непрерывного распределения гистограммы, — **аномалии**.

СОВЕТЫ

- Не стоит строить гистограмму, если наблюдений мало — распределение окажется далёким от действительного и вы просто сделаете ложные выводы. По статистике, для того, чтобы гистограмма хоть как-то оценивала истинное распределение, нужно как минимум 30 наблюдений (на практике нужно хотя бы 100).

- Попробуйте (*ради эксперимента*) построить пять-семь гистограмм на одном графике для их сравнения по категориям (например, страны).
  ⛔ На практике так делать не нужно. Для сравнения параметров распределений по категориям предназначена коробчатая диаграмма (*boxplot*).

- Если вы всё же хотите сравнить гистограммы между собой, предварительно обязательно приведите признаки к одной шкале (мы делали это, когда сравнивали ежедневную заболеваемость коронавирусом в процентах от населения страны). Если этого не сделать, распределения окажутся несопоставимыми.

  Например, что значит 1 млн заболевших в день для Китая с населением в 1.5 млрд человек? А что значит такая же цифра для Ватикана, население которого меньше 1 тыс. человек? Вероятно, суть ясна.

  После того как вы привели все признаки к одной шкале, лучше используйте цветовую гистограмму, которая показывает частоту интенсивностью цвета (загляните в юнит по Seaborn, если забыли). Да, она не покажет высоту столбцов в цифрах, но зато вы сможете визуально сравнить категории между собой и избежать наложения гистограмм друг на друга.

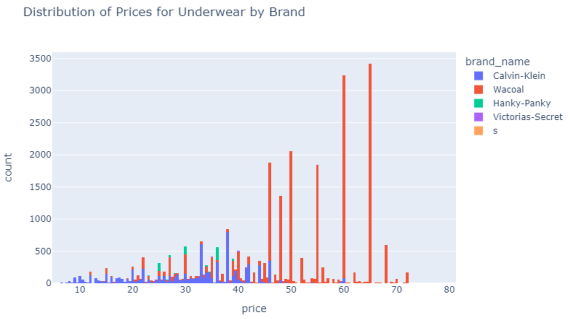


ДИАГРАММА РАССЕЯНИЯ

→ Диаграмма рассеяния (*scatter plot*) и её производные — *jointplot* (гистограммы с рассеянием), kdeplot (диаграмма плотностей) и bubble plot (пузырьковая диаграмма) — предназначены для выявления взаимосвязи между двумя (или в случае *3D* — тремя) признаками.

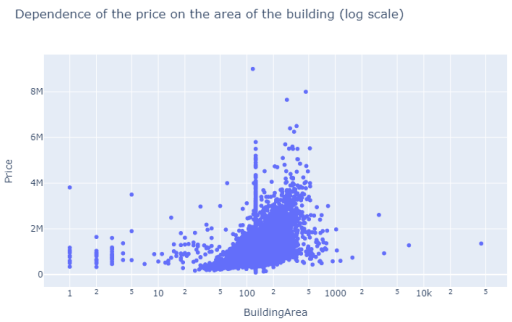
Можно добавлять в график расцветку по одному категориальному признаку, а тип или размер маркеров — по другому. Итого мы сможем наблюдать взаимосвязь нескольких признаков (до пяти).

СОВЕТЫ

- Не используйте диаграммы рассеяния на маленьком наборе данных. Здесь ситуация та же, что и с гистограммами.
- Не стоит использовать расцветку и размер точек для признаков с большим числом уникальных категорий.
- В случае если вы не видите зависимостей в данных, попробуйте использовать логарифмическую шкалу по оси абсцисс (по оси абсцисс и ординат в случае *3D*-графика). Во всех библиотеках в методе есть параметр log, значение которого нужно установить на *True*.

  ? Почему логарифмическая шкала?

  Ответ, конечно же, кроется в математике. Если говорить, не вдаваясь в подробности, функция логарифма, во-первых, отбрасывает отрицательные значения, а во-вторых, «приземляет» более высокие значения — зависимость становится более гладкой, и её становится легче просматривать. При этом логарифмирование не искажает исходную зависимость: то есть если на исходных данных был тренд роста признака А от признака Б, то на логарифмированных данных этот тренд сохранится.



КРУГОВАЯ ДИАГРАММА

> → Круговая диаграмма (*pie chart*) показывает структуру признака, то есть процентную долю каждого из возможных значений признака.

Вы можете столкнуться с ней, например, когда захотите отобразить соотношение классов в данных или вклад отдельных компонентов в общую прибыль.

СОВЕТЫ

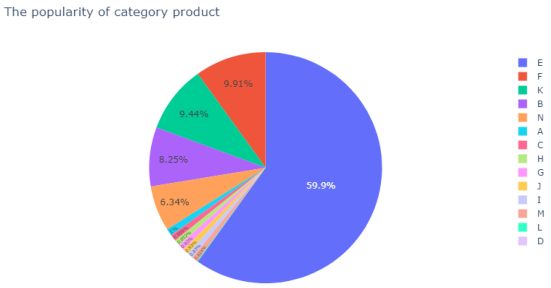
- Сумма значений в круге всегда должна равняться единице, то есть всегда должно быть целое и его части (например, отношение числа заболевших вирусом по странам к общему количеству населения).

  С помощью круговой диаграммы нельзя сравнить средний чек в ресторанах — эти средние не являются частью единого целого. Однако можно сравнить число сотрудников в этих ресторанах, так как они являются частью одной совокупности.

- Не визуализируйте секторы, близкие к 0, — их невозможно сравнить друг с другом.

- Не делайте больше 6-8 секторов — воспринимать информацию будет сложно. Если компонентов больше, выделите ТОП-6-8, а остальные обозначьте как «прочие».

- Всегда отображайте легенду либо подписи категорий внутри секторов.
- Если важно выделить часть графика, «вытащите» его из центра.



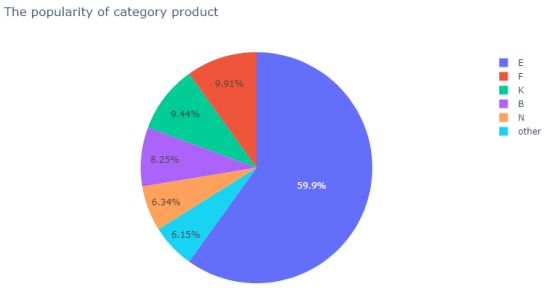


СХЕМА ВЫБОРА ТИПА ВИЗУАЛИЗАЦИИ

Наконец, в завершение темы визуализации предлагаем вам ознакомиться со схемой, на которую вы можете ориентироваться при выборе типа визуализации для своих реальных задач:

Примечание. В основном все перечисленные типы визуализации вам уже знакомы, а те, что не знакомы, являются небольшими модификациями знакомых.

Так, **пончиковая диаграмма** является модификацией круговой и предназначена для отображения многоуровневой группировки.

Диаграмма областей является версией линейного графика и позволяет оценить вклад каждого компонента в рассматриваемый процесс.

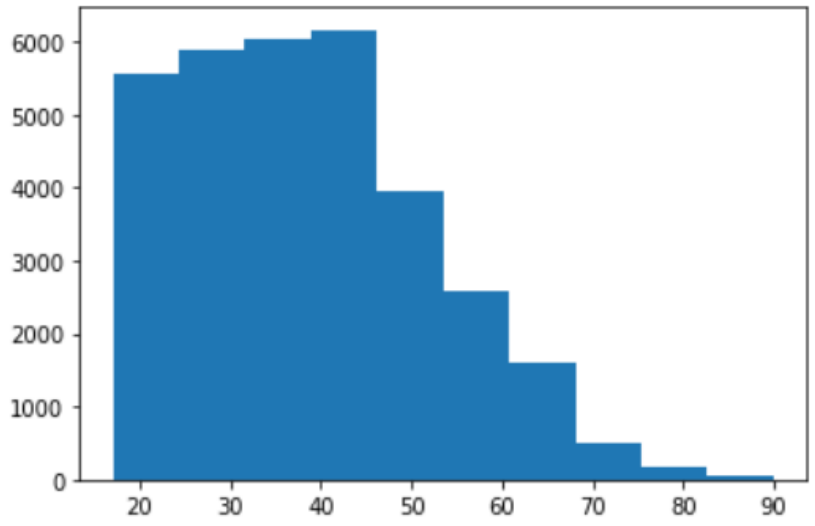
Водопадная диаграмма — это разновидность столбчатой диаграммы. Она используются для понимания, как некоторая группа факторов, например реклама, найм новых сотрудников и скидки, повлияла на начальное значение исследуемого показателя (прибыли).

✍ Теперь вы разбираетесь в тонкостях построения графиков и умеете правильно преподносить их аудитории.

Задание 8.1

1/1 point (graded)

Какое правило визуализации нарушено на представленном ниже графике?



- ◉ График не имеет опознавательных знаков (отсутствуют название графика и подписи осей) ✔

- ○ График нарушает принцип минимализма
- ○ Различная цветовая гамма для графиков
- ○ Излишне сложный дизайн

Отправить

Задание 8.2

1/1 point (graded)

Какие правила визуализации нарушены на представленном ниже графике?

- ☑ **A** График не имеет подписей осей
- ☑ **B** Нарушено правило единой цветовой гаммы для графиков
- ☑ **C** Некорректное отображение дат по временной оси
- ☐ **D** Излишне сложный дизайн

✔

Ответ

Верно:

A Верно.

B Верно.

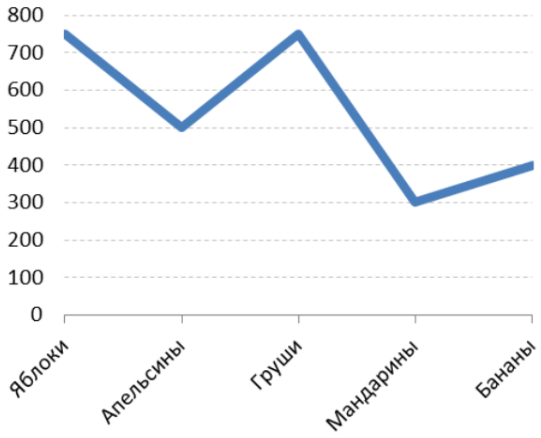
C Верно.

Отправить

Задание 8.3

1/1 point (graded)

Ниже представлен график количества проданных в магазине продуктов. Какое правило визуализации в нём нарушено?



- ○ Нарушен принцип минимализма
- ◉ Неверный выбор графика (верный выбор — круговая или столбчатая диаграмма) ✔
- ○ Неверный выбор графика (верный выбор — гистограмма или коробчатая диаграмма)
- ○ Нарушено правило «одна плоскость — один график»

Отправить

Задание 8.4

1/1 point (graded)

В чём ошибка на представленной ниже круговой диаграмме?

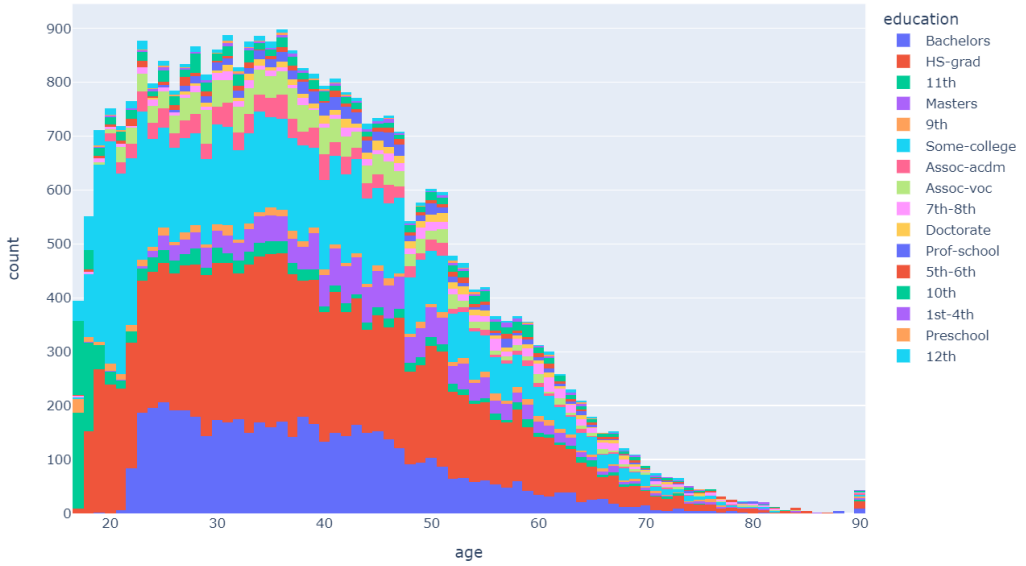
- ○ Визуализированы близкие к 0 сектора
- ○ Не выделена целевая часть круга
- ○ Отсутствует легенда
- ◉ Секторы не являются частью единого целого ✔

Отправить

Задание 8.5

1/1 point (graded)

На представленной ниже гистограмме сравнивается распределение возраста в зависимости от уровня образования. Данное сравнение является неверным. Выберите альтернативный вид визуализации, который позволит корректно сравнить распределения.



- ○ Столбчатая диаграмма
- ○ Круговая диаграмма
- ○ Диаграмма рассеяния
- ◉ Коробчатая диаграмма ✔

Отправить