



Seaborn



Seaborn

- Seaborn – это библиотека по построению статистических графиков, которая отлично работает с Pandas DataFrames для создания часто используемых графиков
- Seaborn построен на основе Matplotlib, но в нём многие графики можно создать “одной строкой кода”



Seaborn

- Seaborn – это баланс между возможностями кастомизации графиков и лёгкостью работы
- Однако, поскольку seaborn построен на основе Matplotlib, мы можем вызывать методы plt для изменения свойств графиков seaborn



Seaborn

- Типичный график в seaborn можно создать с помощью одной строки кода:
 - `sns.scatterplot(x='salary',y='sales',data=df)`
- Seaborn принимает на вход объект DataFrame, и далее нужно указать названия колонок в виде строковых значений (разные параметры для разных графиков)



Seaborn

- В этом разделе мы изучим области применения для каждого графика, и синтаксис seaborn
- Онлайн-документация:
 - <https://seaborn.pydata.org>
- Частый вопрос: как выбрать тип графика?



Seaborn

- Ответ на этот вопрос зависит от того, какие связи в данных мы хотим исследовать
- Попробуйте поискать в Google фразу “choosing a plot visualization” – Вы найдёте много вариантов путеводителей по выбору типа графика
- В конце этого раздела у Вас появится понимание того, в каких случаях какие графики полезны.



Seaborn

- В этом разделе мы изучим следующие графики:
 - Scatter Plots – диаграммы рассеяния
 - Distribution Plots – диаграммы распределения вероятностей
 - Categorical Plots – графики по категориям
 - Comparison Plots – графики сравнения
 - Seaborn Grids – наборы графиков в Seaborn
 - Matrix Plots – матричные графики



Давайте начнём!



Scatter Plots – диаграммы рассеяния



Scatter Plots – диаграммы рассеяния

- Графики Scatter Plots показывают связь между двумя непрерывными переменными
- “Непрерывная” переменная – это числовая переменная, которая может принимать любые промежуточные значения
- Примеры: возраст, высота, зарплата, температура



Scatter Plots – диаграммы рассеяния

- В непрерывной переменной между двумя значениями всегда найдётся какое-то третье значение
- Это сильно отличается от категориальных переменных, где каждое значение – отдельная категория. Примеры: цвета, имена, формы и т.д.



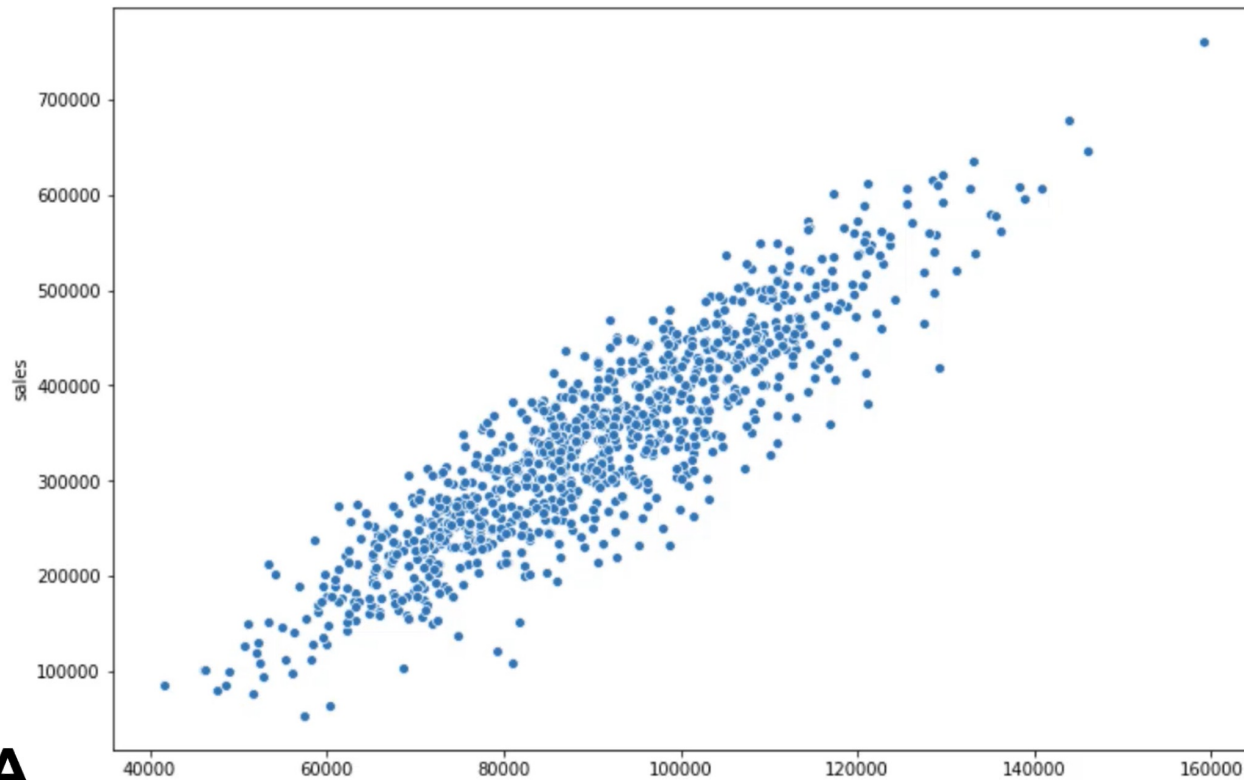
Scatter Plots – диаграммы рассеяния

- График Scatter Plots берёт две непрерывные переменные, и наносит их на две оси графика
- В качестве примера представим сотрудников, у которых есть зарплата, и объём выполненных ими продаж за год. Мы можем исследовать связь между этими двумя переменными - зарплата (salary) и объём продаж (sales).



Scatter Plots – диаграммы рассеяния

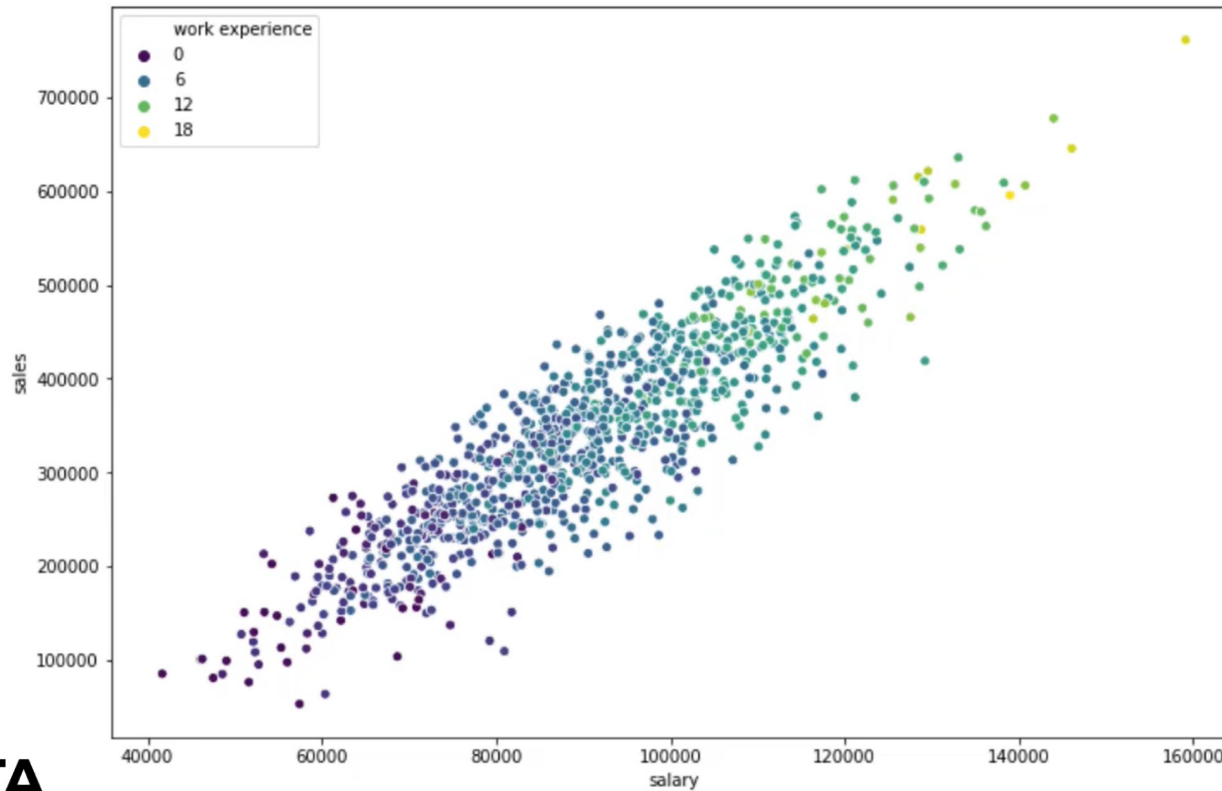
- Наносим координаты (x,y) на график в виде точек





Scatter Plots – диаграммы рассеяния

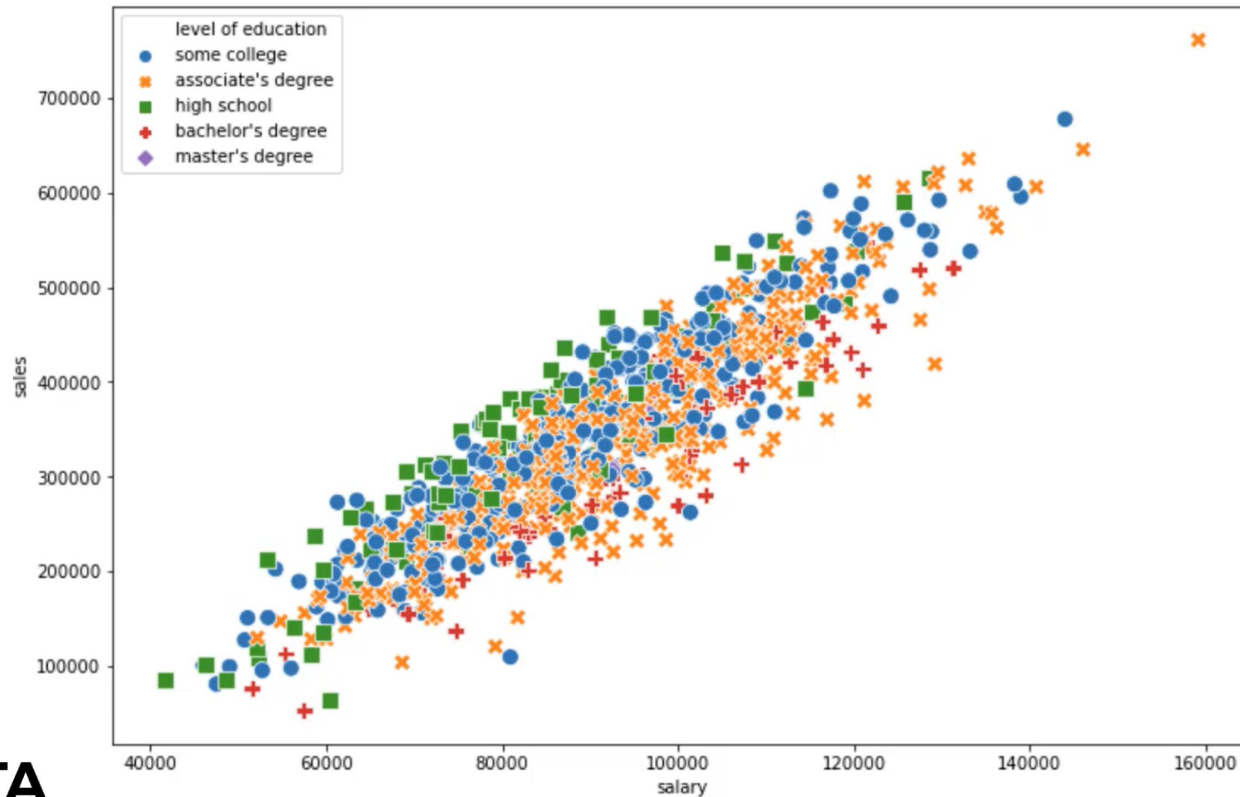
- Далее можно добавить цветовую раскраску





Scatter Plots – диаграммы рассеяния

- В Seaborn можно настроить цвета и форму точек





Scatter Plots – диаграммы рассеяния

- Давайте построим графики Scatter Plots!



Distribution Plots

Часть 1 - виды графиков



Distribution Plots – графики распределения вероятностей

- Графики Distribution Plots отображают одну непрерывную переменную, и помогают визуализировать атрибуты этой переменной, например среднее значение или дисперсию.
- Три основных типа графиков:
 - Rug Plot
 - Histogram
 - KDE Plot



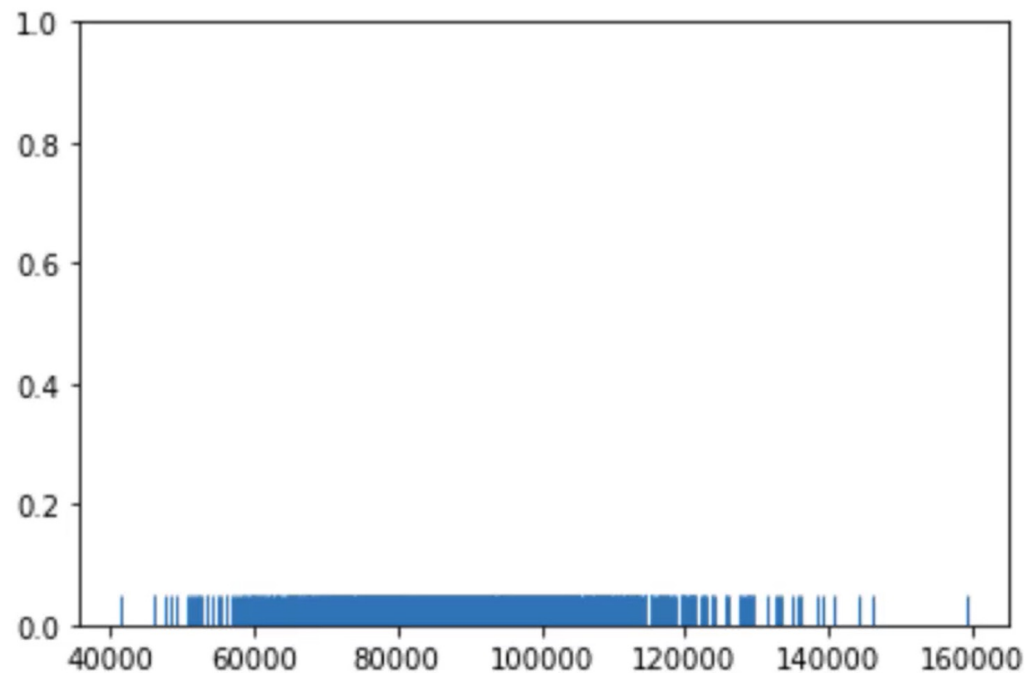
Distribution Plots – графики распределения вероятностей

- Давайте исследуем распределение зарплат сотрудников
- Это можно сделать с помощью графика rug plot.
- График rug plot – простейший из графиков распределения, в нём для каждого сотрудника мы ставим точку (чёрточку) на горизонтальной оси
- Вертикальная ось по сути не используется



Distribution Plots – графики распределения вероятностей

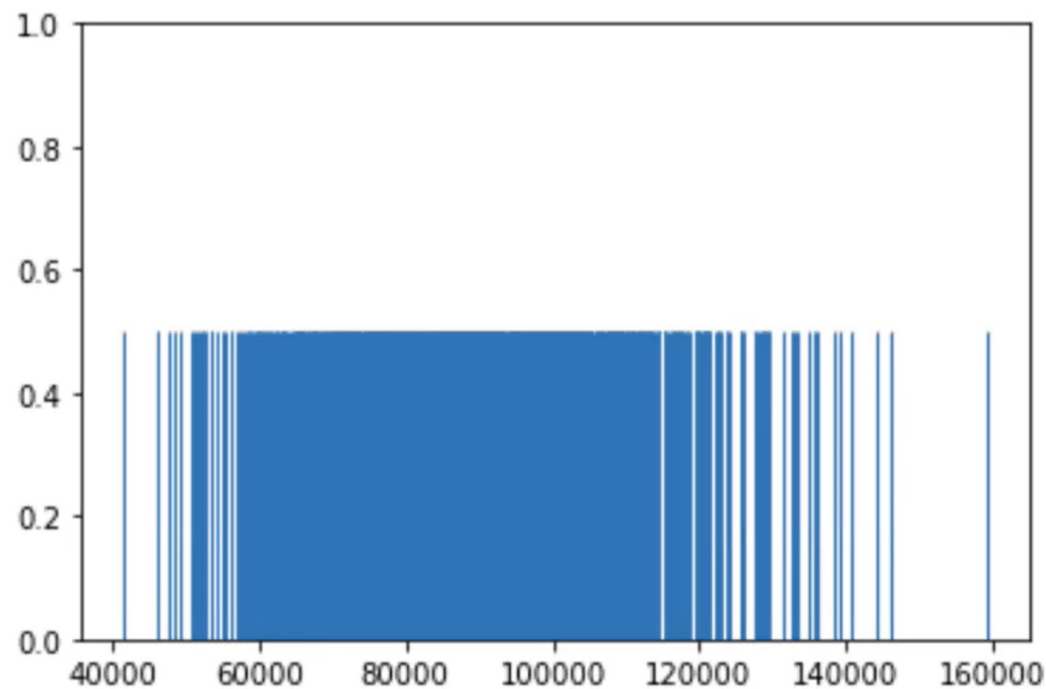
- Rug plot - для каждого сотрудника мы ставим чёрточку на горизонтальной оси





Distribution Plots – графики распределения вероятностей

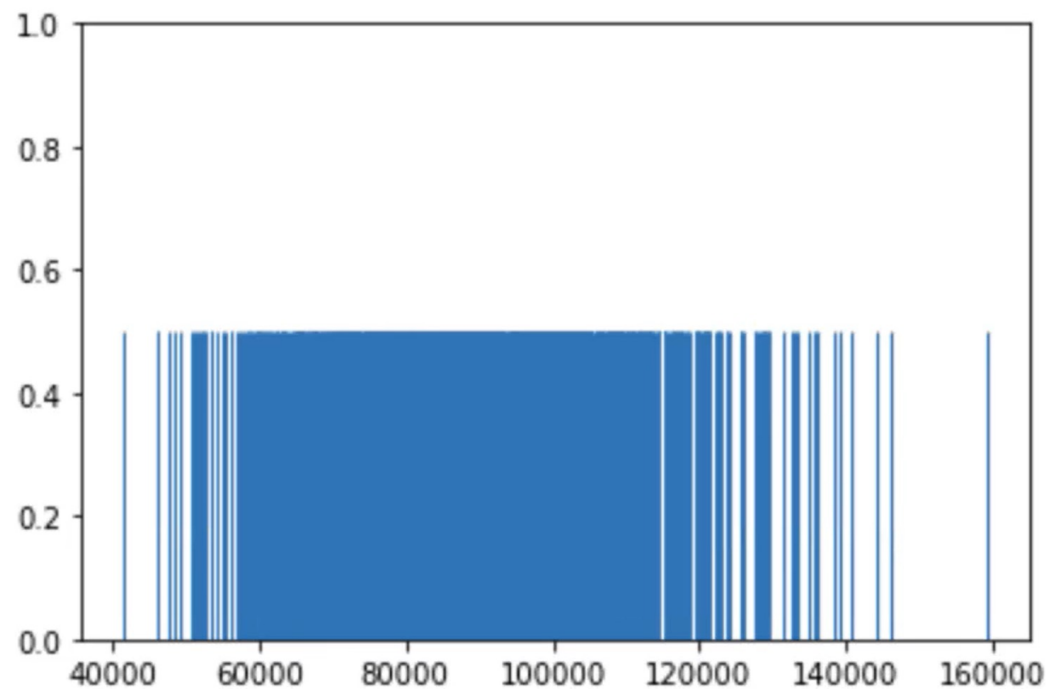
- При желании можно увеличить высоту чёрточек
- Но вертикальная ось по-прежнему не используется





Distribution Plots – графики распределения вероятностей

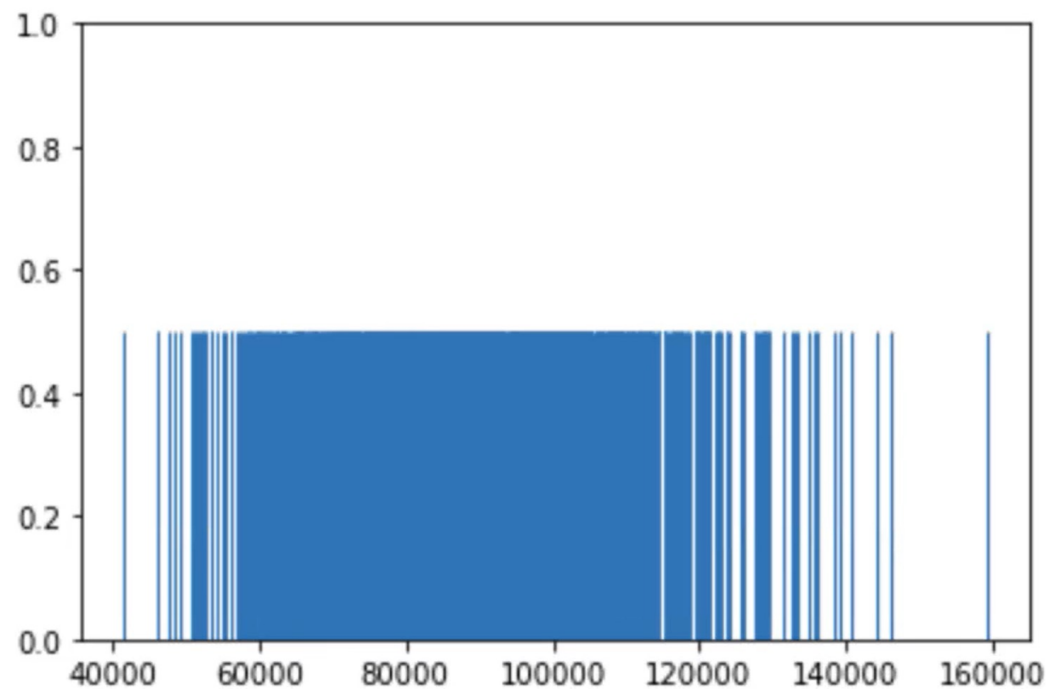
- Наибольшая зарплата – \$160k.
- В среднем – между \$60k и \$120k





Distribution Plots – графики распределения вероятностей

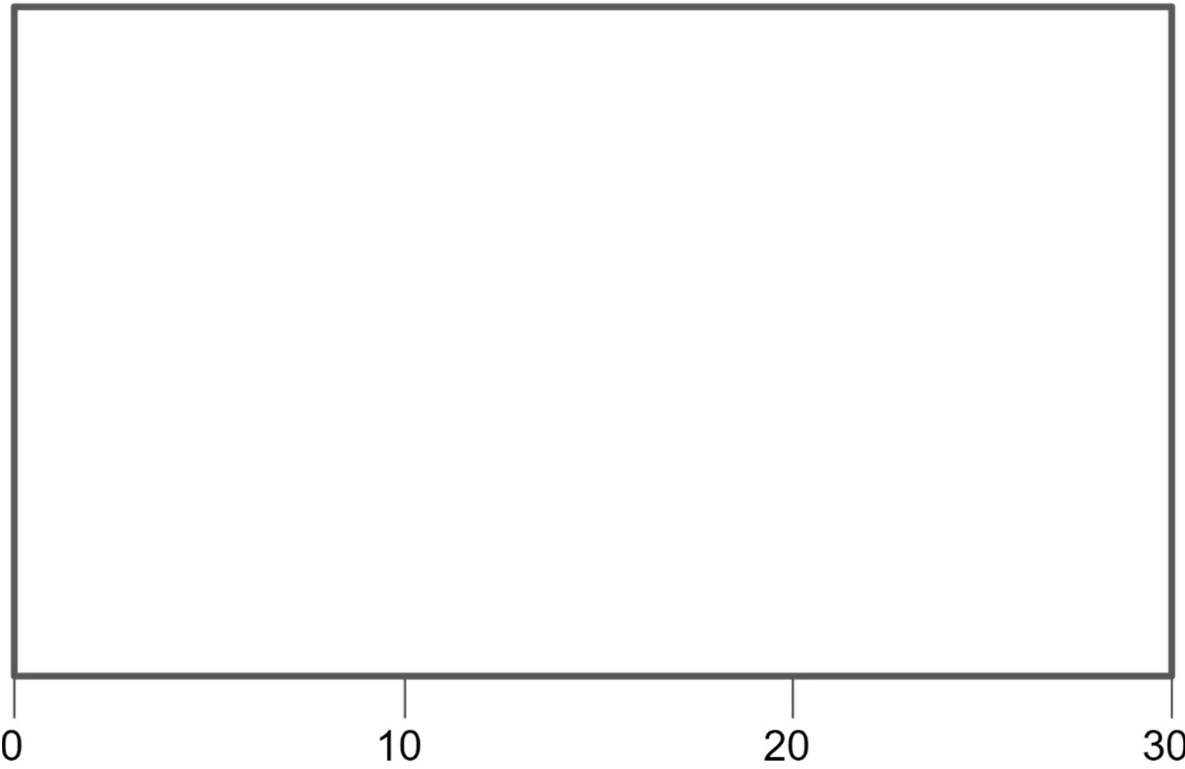
- Мы не видим, где много или мало чёточек
- Если посчитаем чёточки, то получим гистограмму





Distribution Plots – графики распределения вероятностей

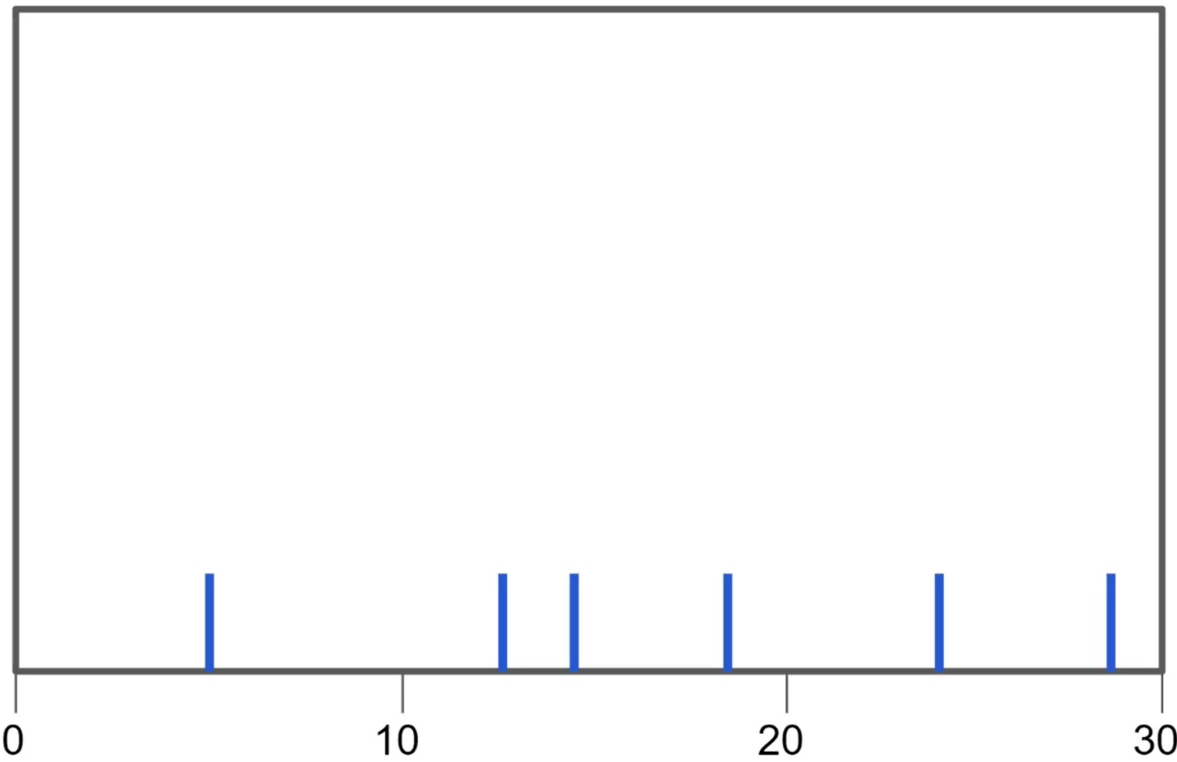
- Рассмотрим простой пример





Distribution Plots – графики распределения вероятностей

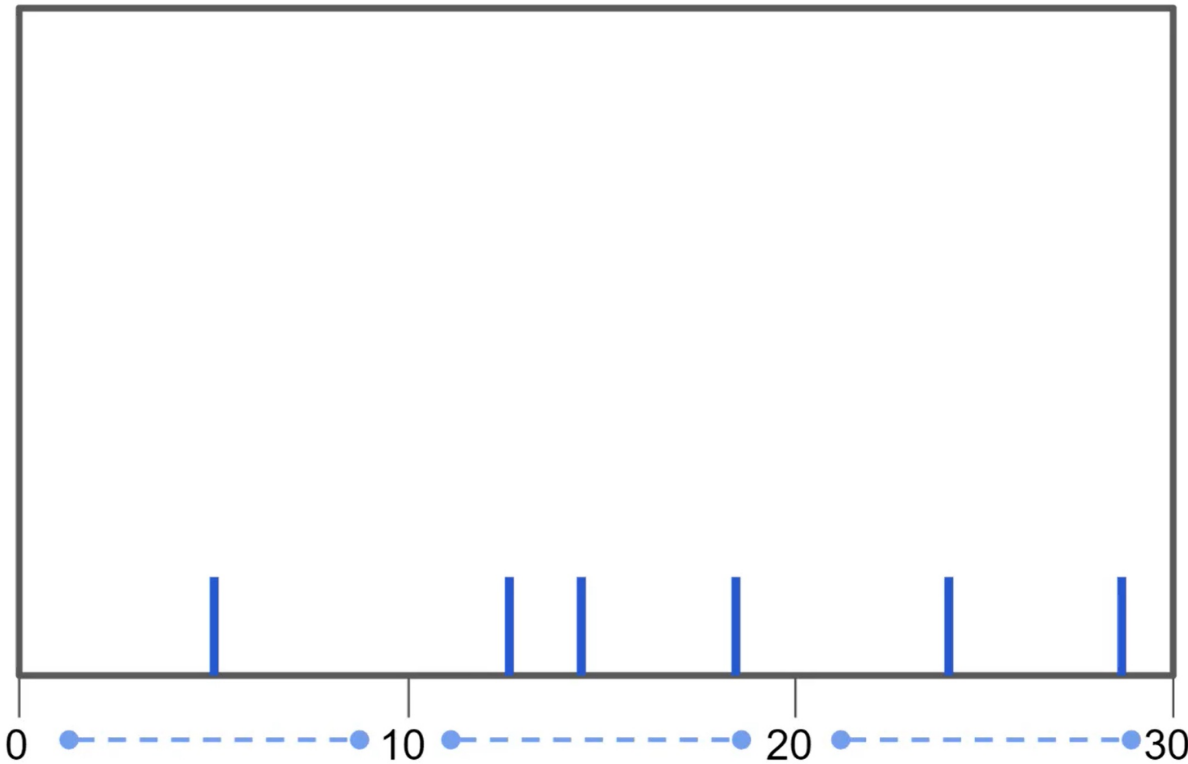
- Нанесём сюда чёточки из графика rug plot





Distribution Plots – графики распределения вероятностей

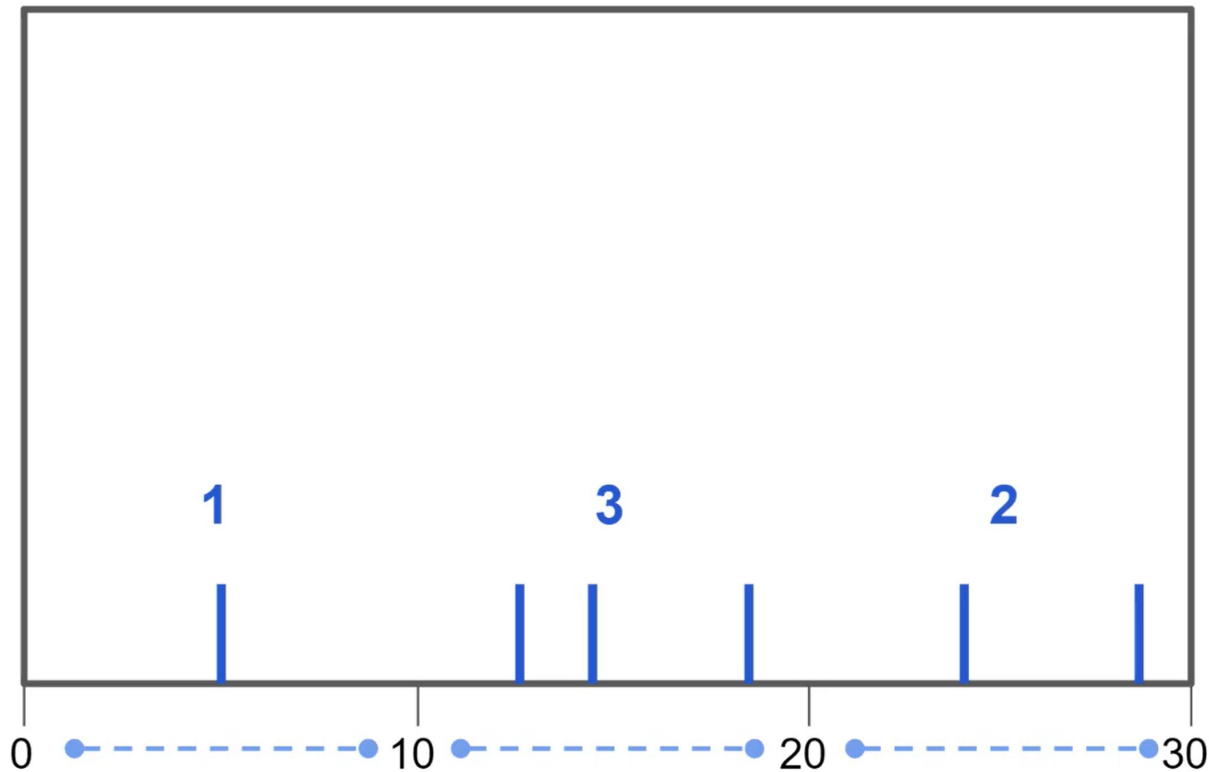
- Выберем количество интервалов (bins), равное 3





Distribution Plots – графики распределения вероятностей

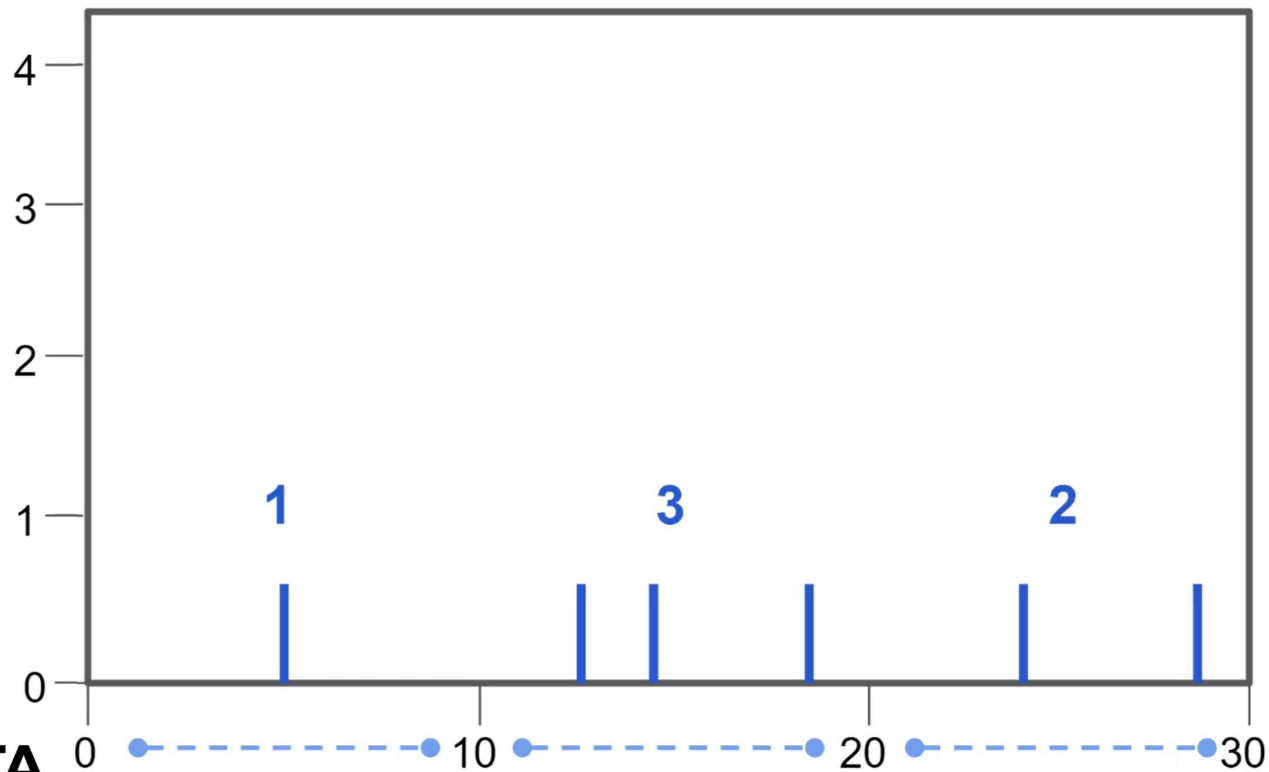
- Считаем количество чёточек в каждом интервале





Distribution Plots – графики распределения вероятностей

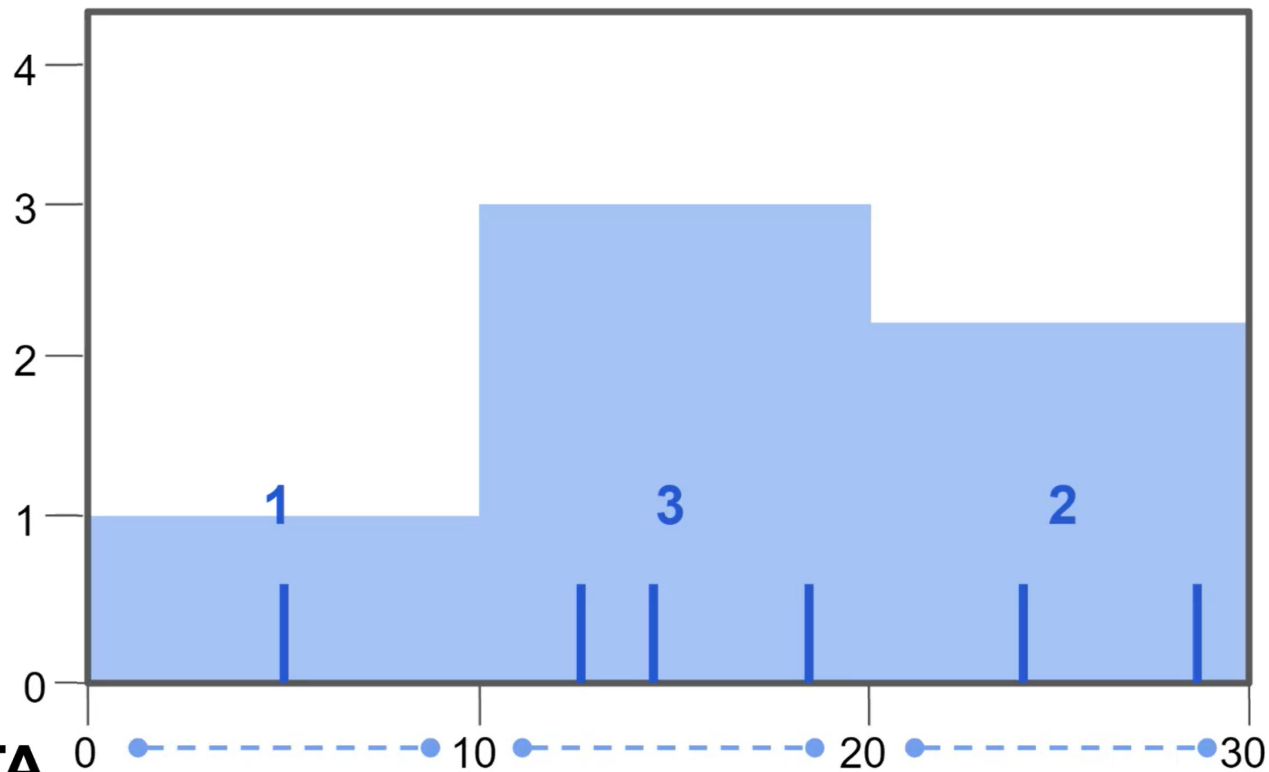
- Создаём вертикальную шкалу





Distribution Plots – графики распределения вероятностей

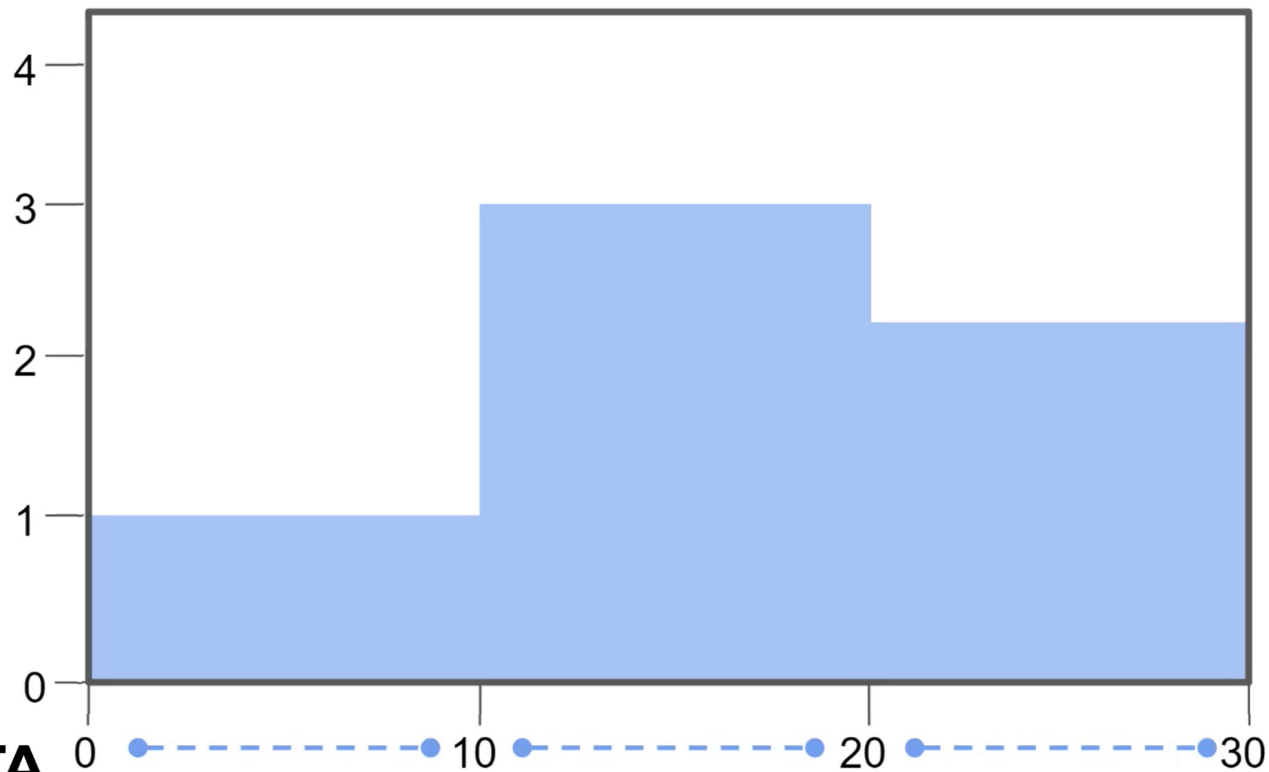
- Создаём вертикальную шкалу





Distribution Plots – графики распределения вероятностей

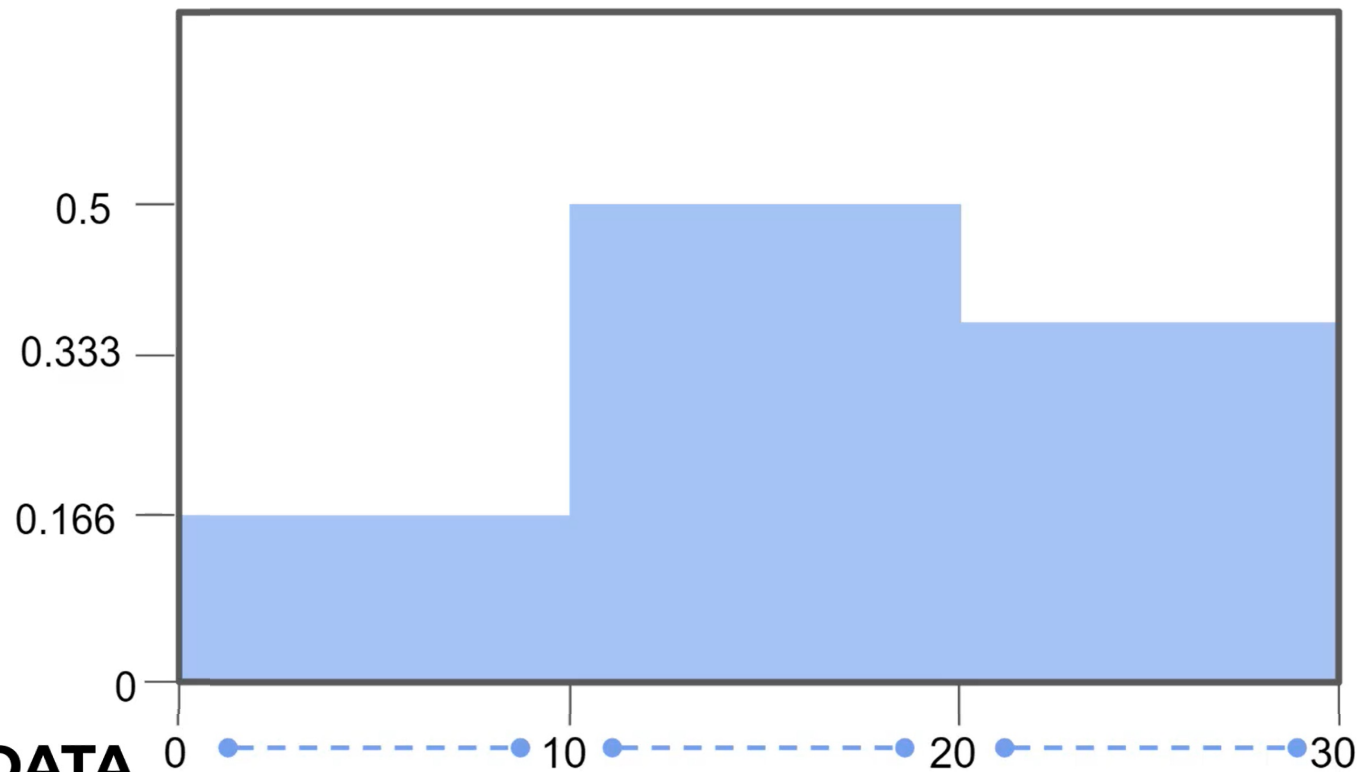
- Гистограмма готова!





Distribution Plots – графики распределения вероятностей

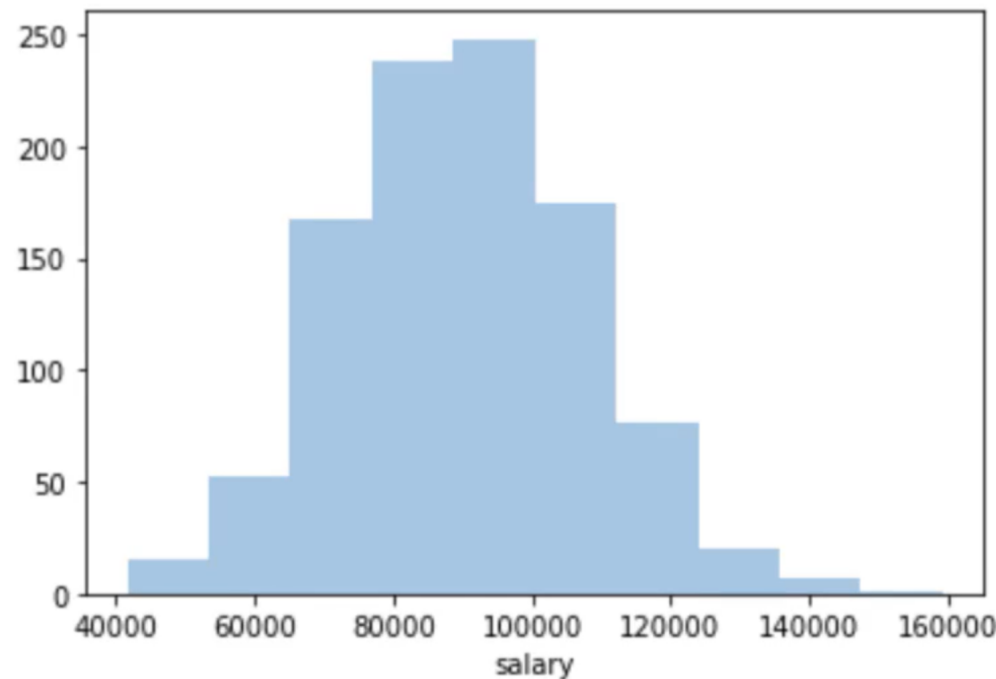
- Ещё можно нормализовать значения по оси y





Distribution Plots – графики распределения вероятностей

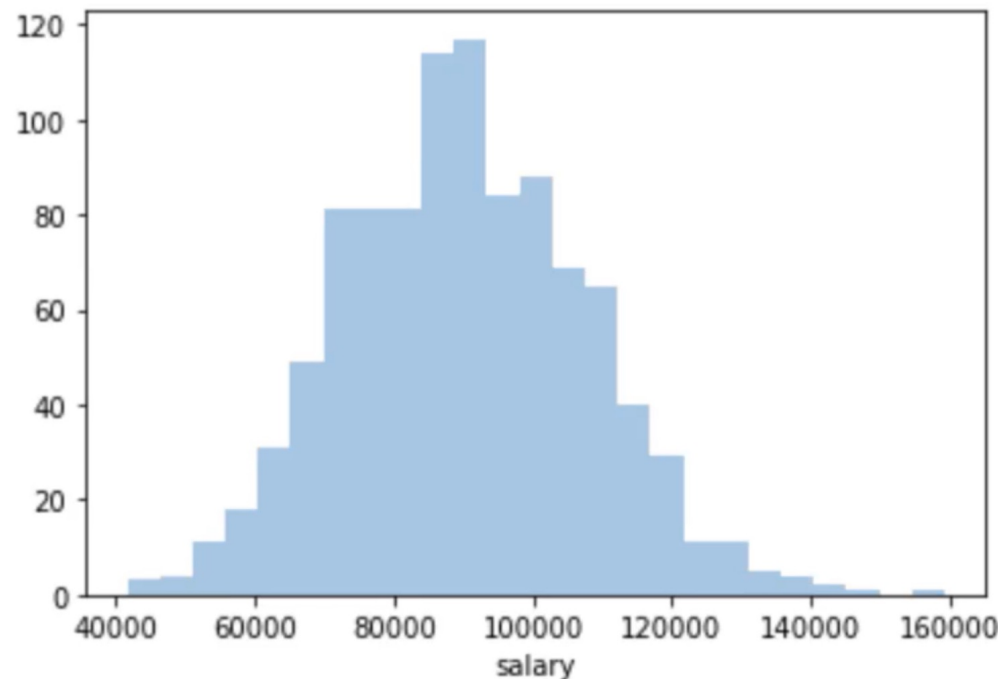
- Увеличивая количество интервалов (bins), получаем более детальный график распределения





Distribution Plots – графики распределения вероятностей

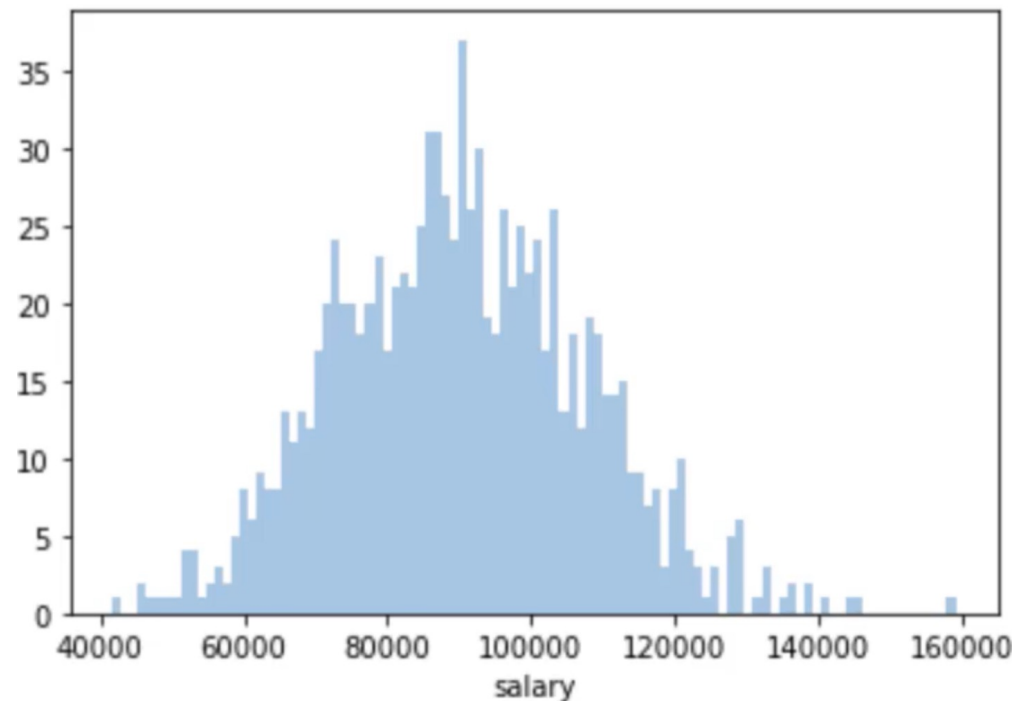
- Увеличивая количество интервалов (bins), получаем более детальный график распределения





Distribution Plots – графики распределения вероятностей

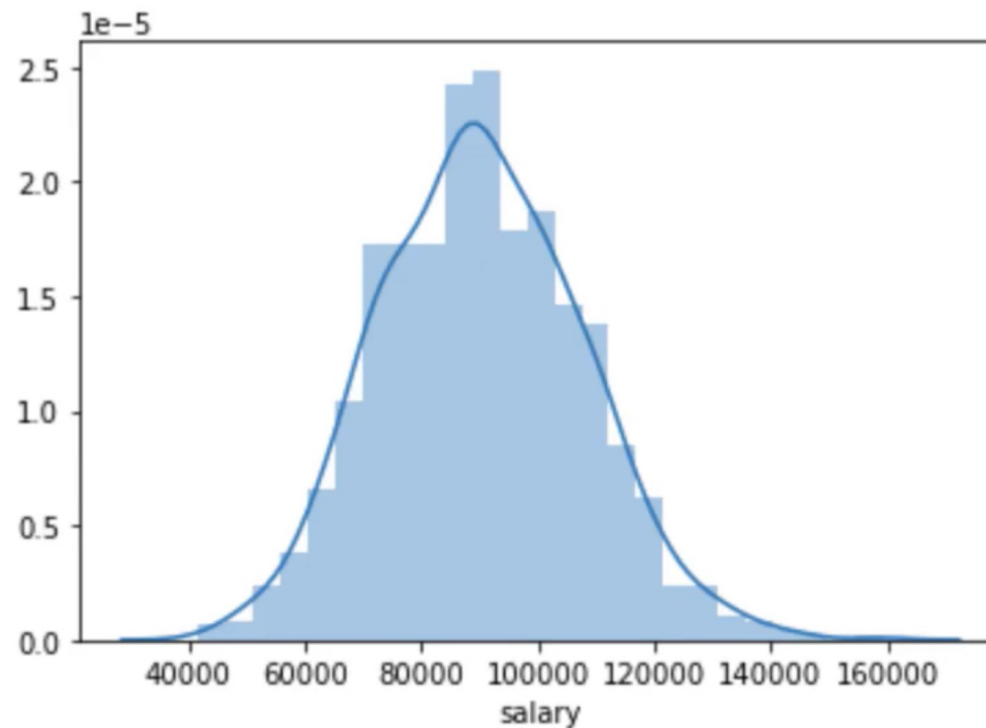
- Увеличивая количество интервалов (bins), получаем более детальный график распределения





Distribution Plots – графики распределения вероятностей

- Seaborn позволяет добавить поверх гистограммы график KDE





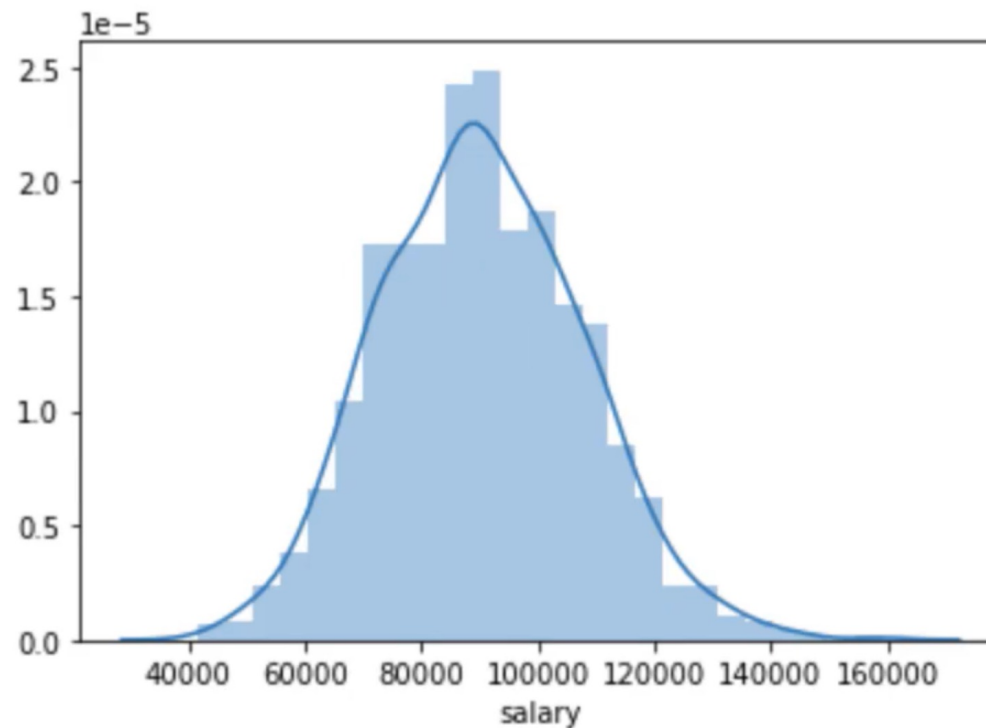
Distribution Plots – графики распределения вероятностей

- KDE расшифровывается как Kernel Density Estimation – ядерная оценка плотности.
- Это способ оценки функции распределения случайной величины
- Простыми словами – это способ построения непрерывного графика по некоторому набору дискретных данных.



Distribution Plots – графики распределения вероятностей

- Давайте посмотрим, как строится график KDE



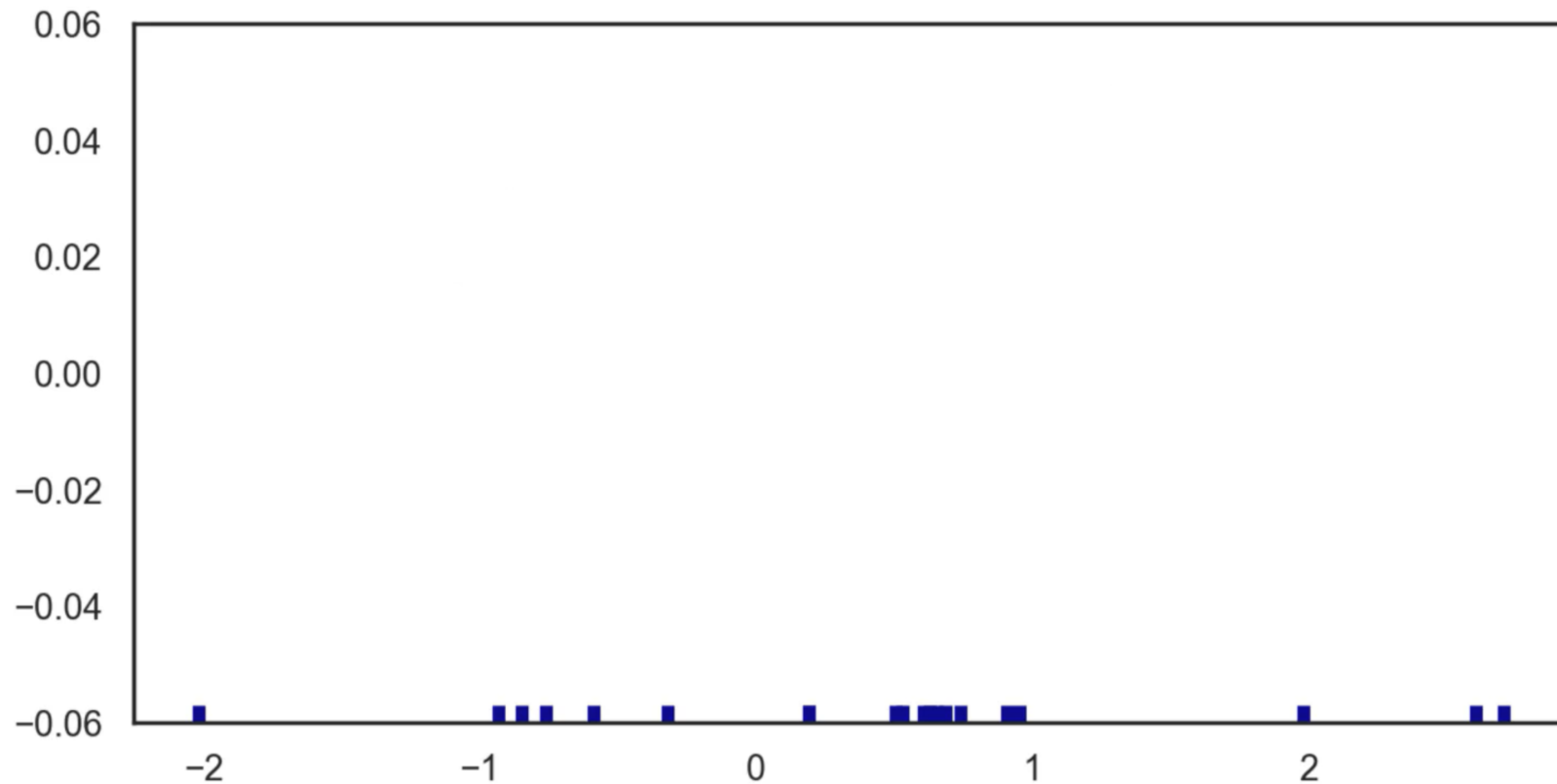


Distribution Plots – графики распределения вероятностей

- Проще всего понять графики KDE, визуализируя процесс их построения
- Давайте начнём с графика rug plot...

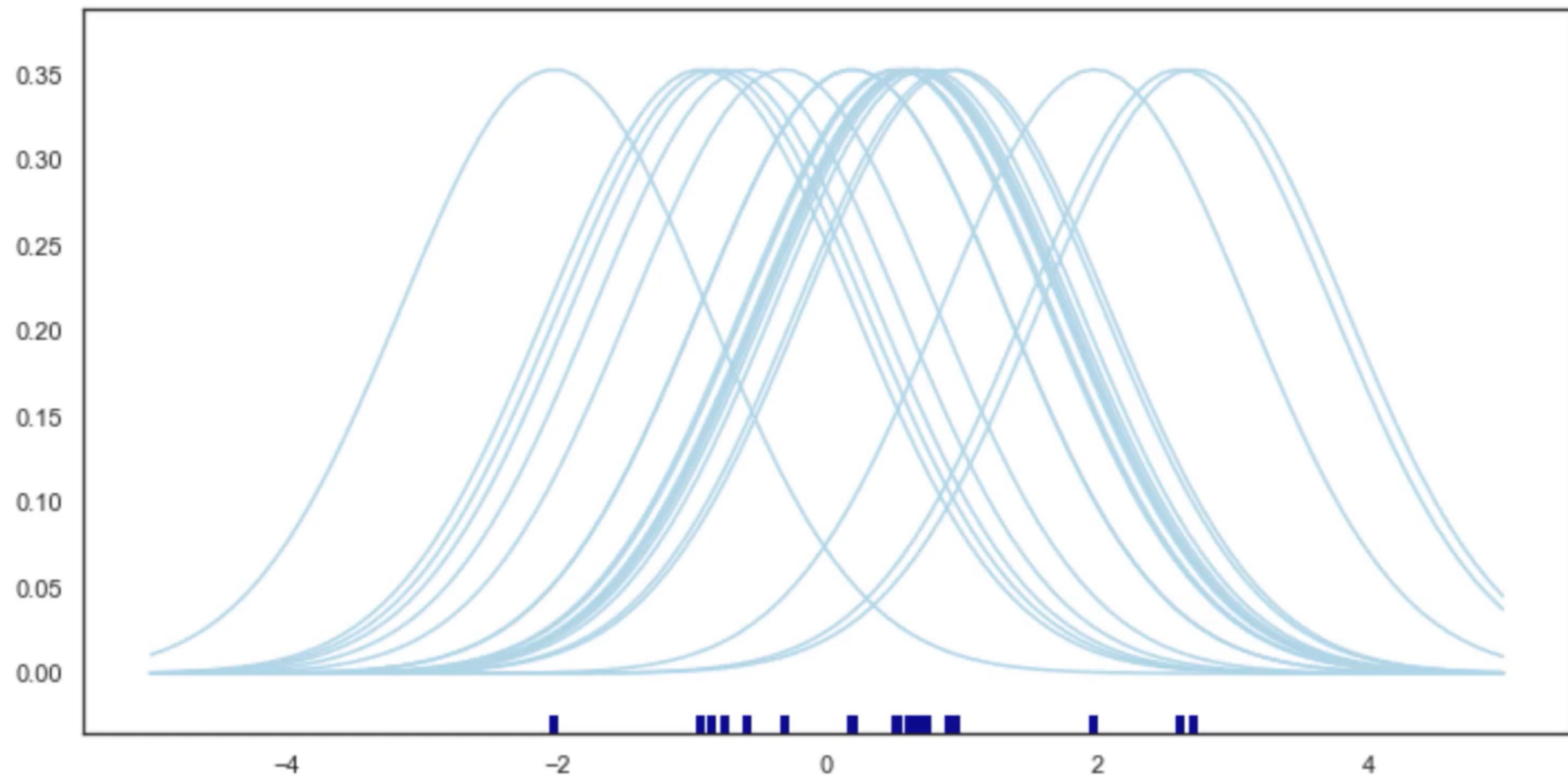


Distribution Plots – графики распределения вероятностей



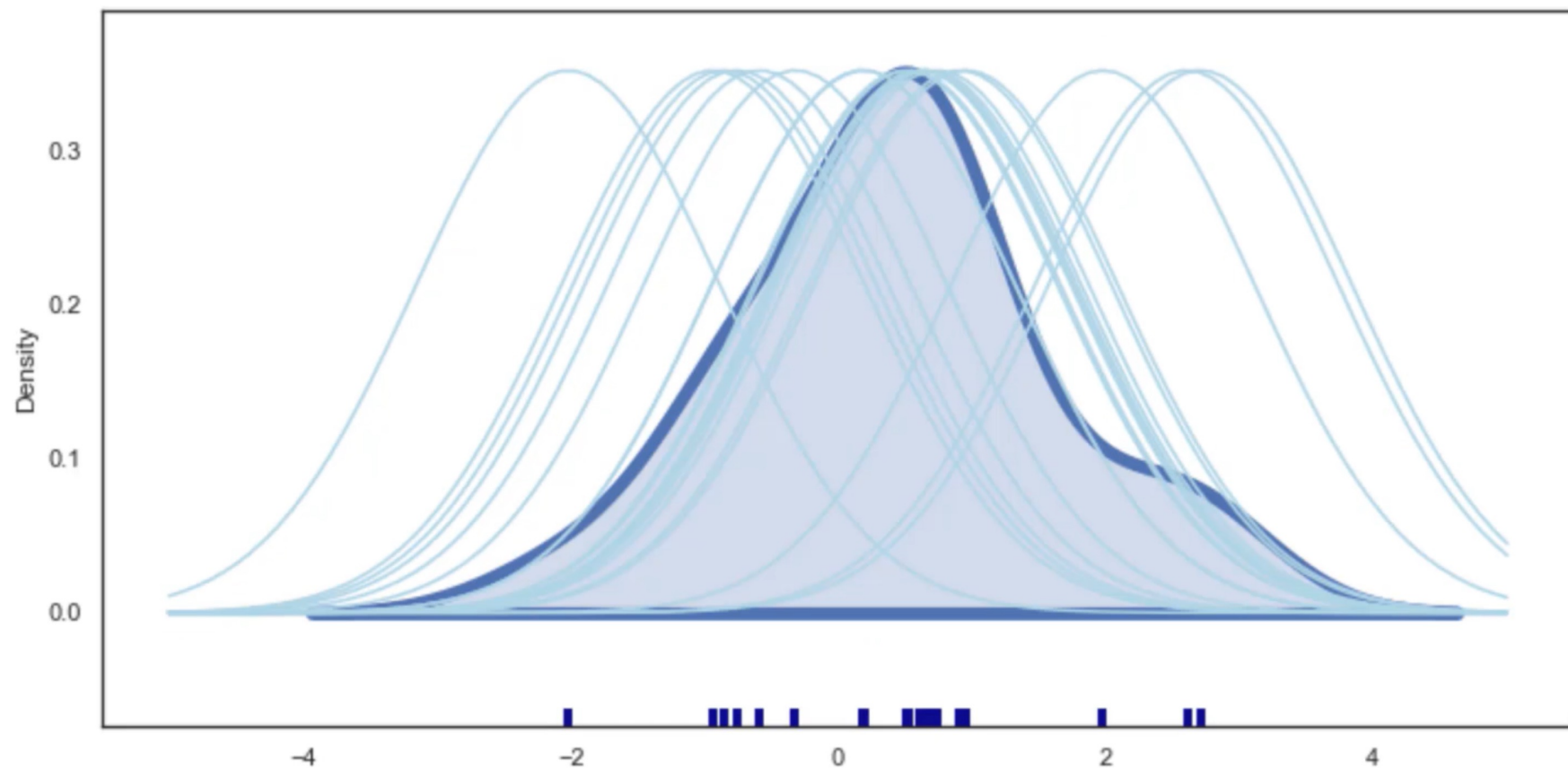


Distribution Plots – графики распределения вероятностей



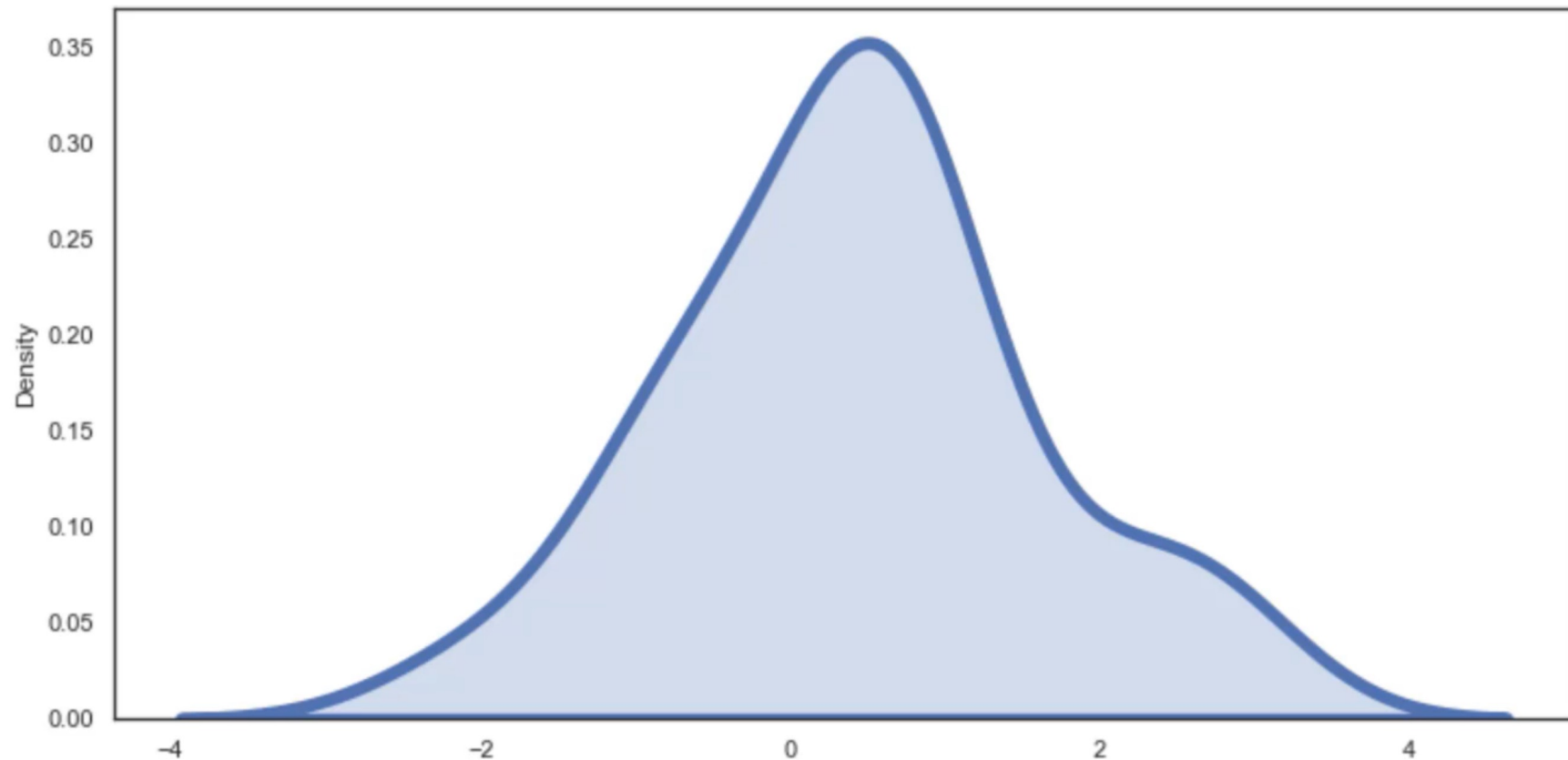


Distribution Plots – графики распределения вероятностей



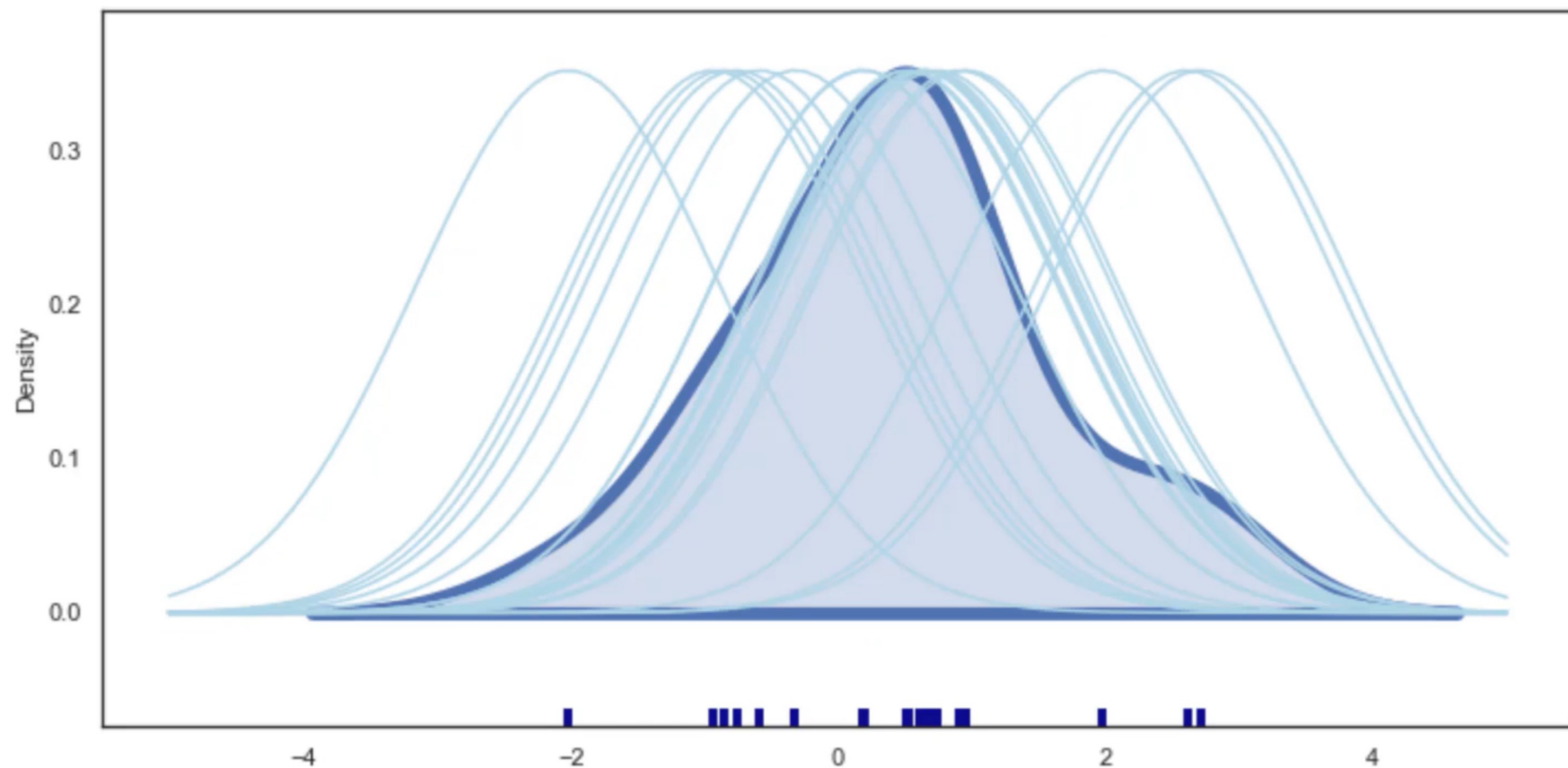


Distribution Plots – графики распределения вероятностей



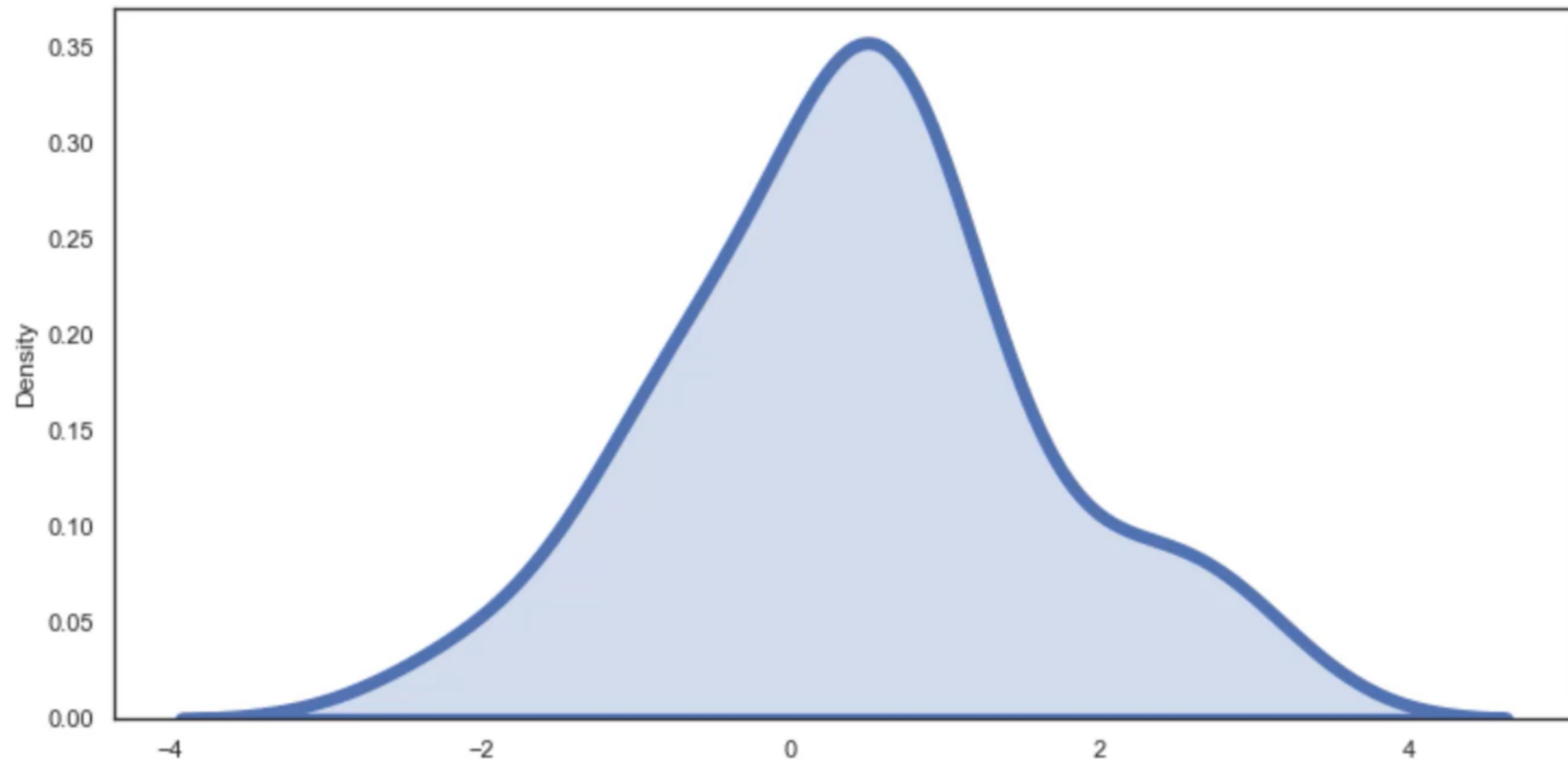


Distribution Plots – графики распределения вероятностей





Distribution Plots – графики распределения вероятностей





Distribution Plots – графики распределения вероятностей

- В графике KDE можно поменять такие параметры, как функцию ядра (kernel) и ширину полосы (bandwidth), чтобы изобразить больше или меньше вариации, содержащейся в данных
- В следующей лекции мы увидим, как построить такие графики с помощью python и seaborn!



Distribution Plots

Часть 2 – код в seaborn



Categorical Plots

Часть 1 – что это за графики



Categorical Plots – графики по категориям

- Графики Categorical Plots, которые мы будем здесь обсуждать, отображают статистические метрики для каждой из категорий переменной
- Например, среднее значение переменной для каждой категории, или количество строк для каждой категории
- Это эквивалентно вызову функции `groupby()`



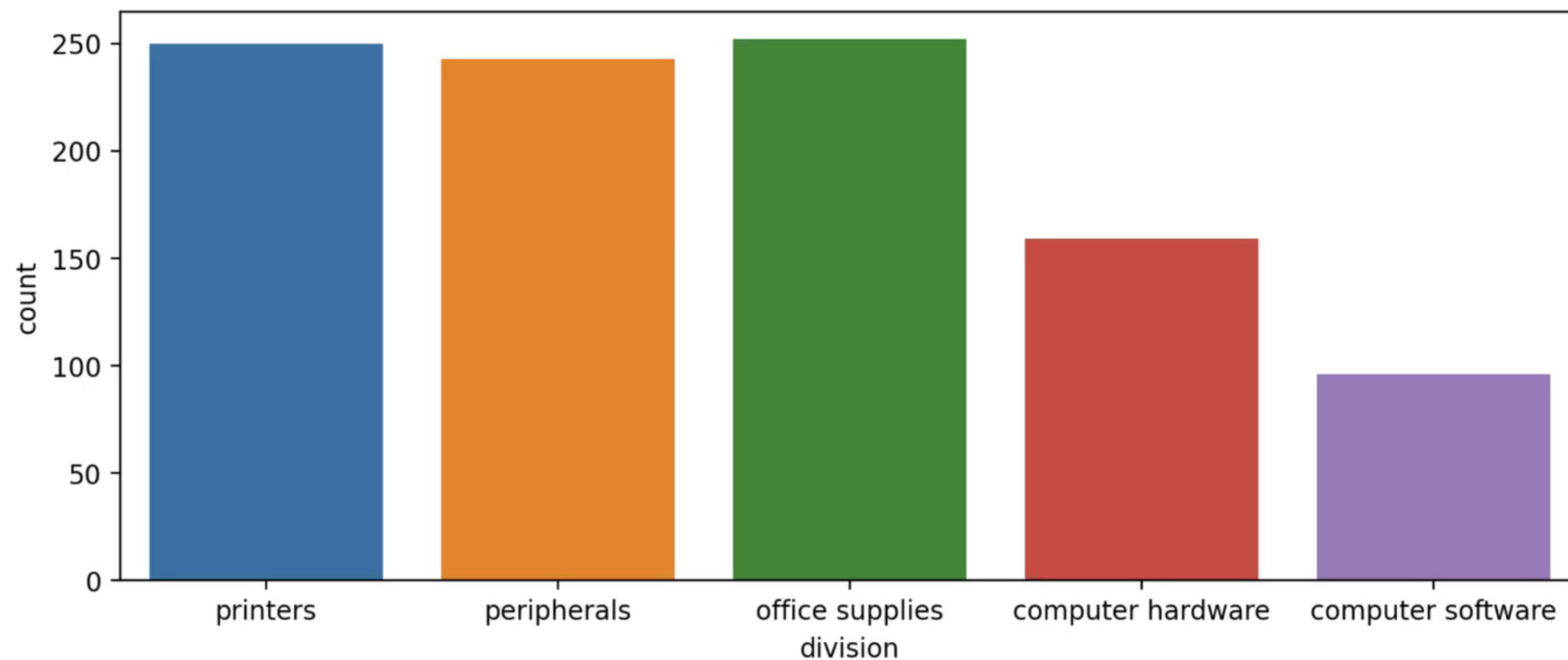
Categorical Plots – графики по категориям

- Основные два вида графиков:
 - `countplot()`
 - подсчитывает количество строк по категориям
 - `barplot()`
 - общая форма отображения любой метрики по категориям



Categorical Plots – графики по категориям

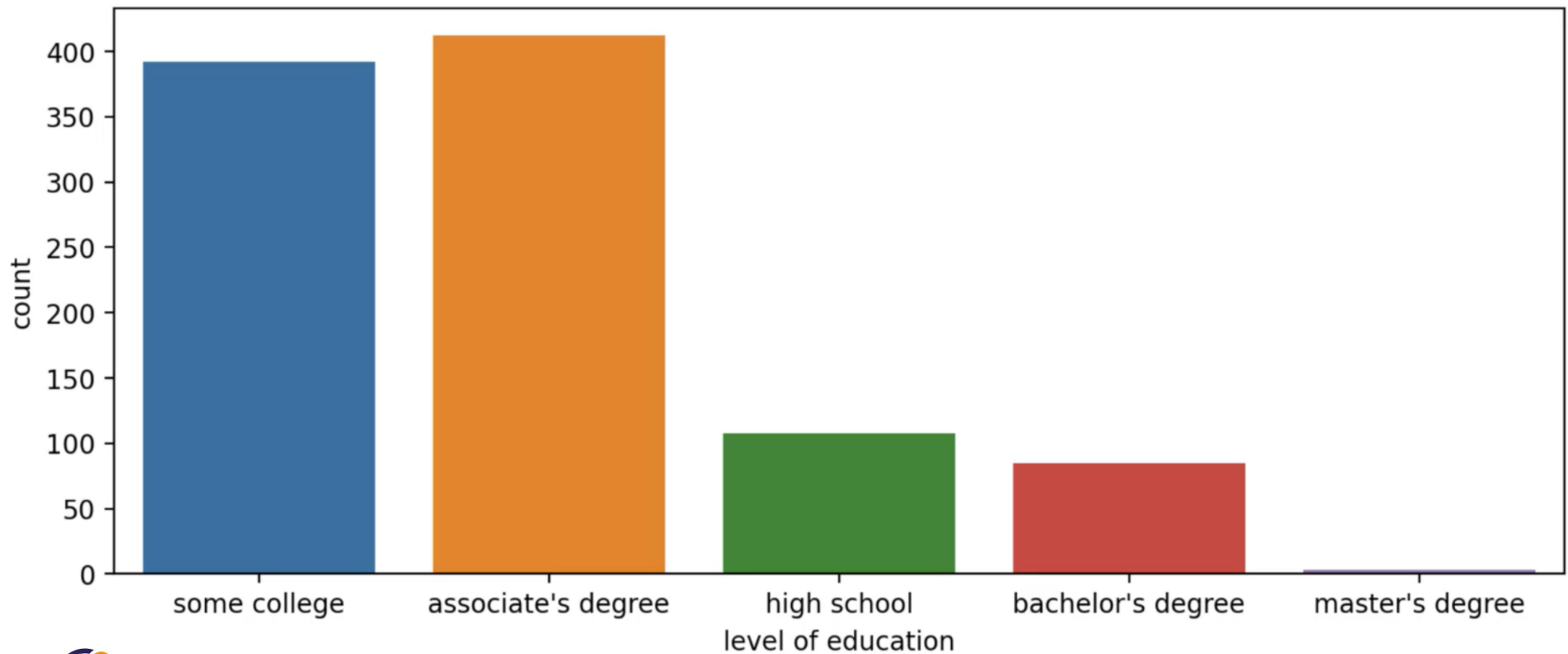
- Пример графика `countplot()` по подразделениям





Categorical Plots – графики по категориям

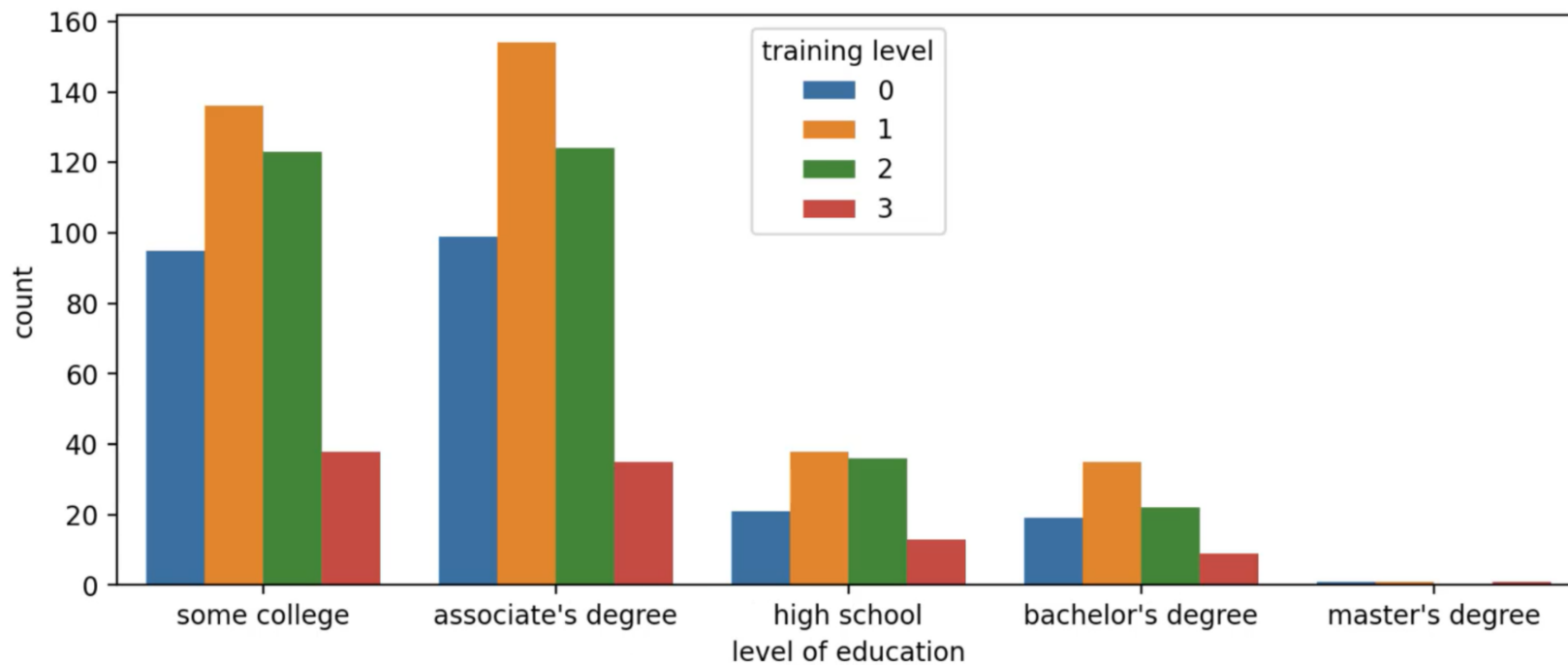
- `countplot()` по уровню образования





Categorical Plots – графики по категориям

- `countplot()` + дополнительное деление цветом (hue)





Categorical Plots – графики по категориям

- График `barplot` позволяет Вам выбрать любую метрику для оси y
- Например, можем нарисовать среднее значение или среднеквадратичное отклонение для каждой категории



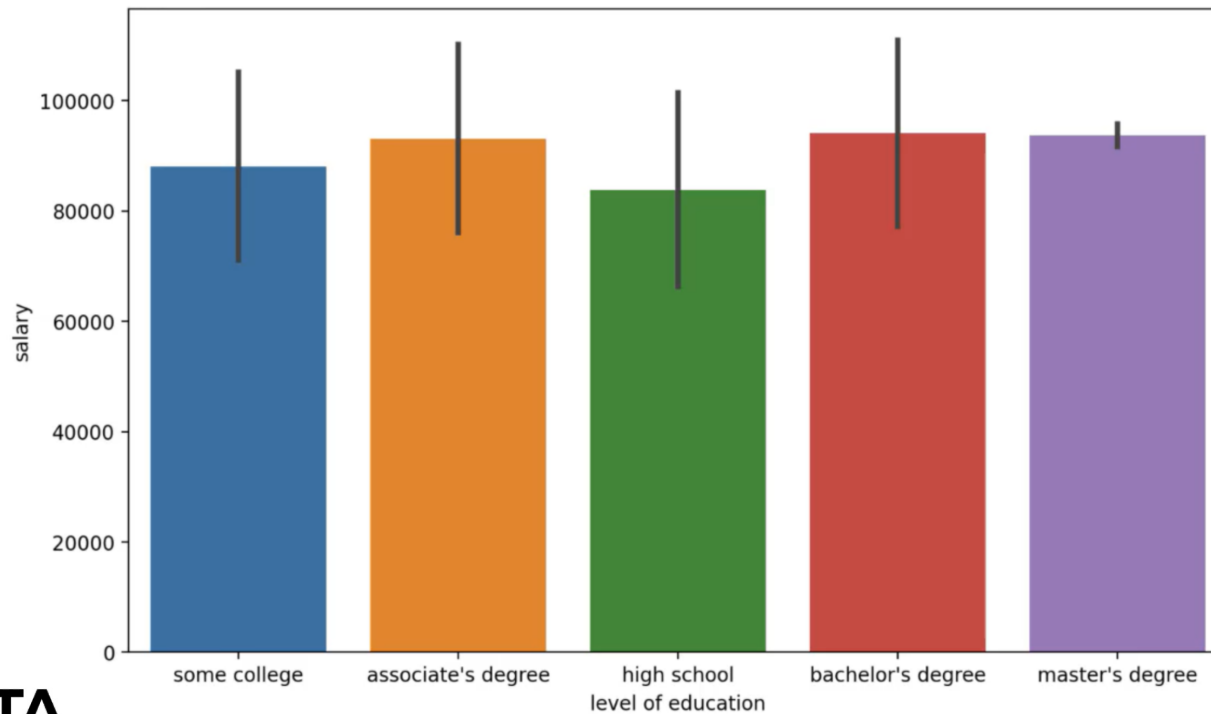
Categorical Plots – графики по категориям

- Важное замечание! Будьте аккуратны с графиками barplot – поскольку столбцы на них выглядят как непрерывные столбцы, то смотрящий на график человек может подумать, что ось y является непрерывной, хотя это не так!
- Всегда добавляйте описание для этих графиков.



Categorical Plots – графики по категориям

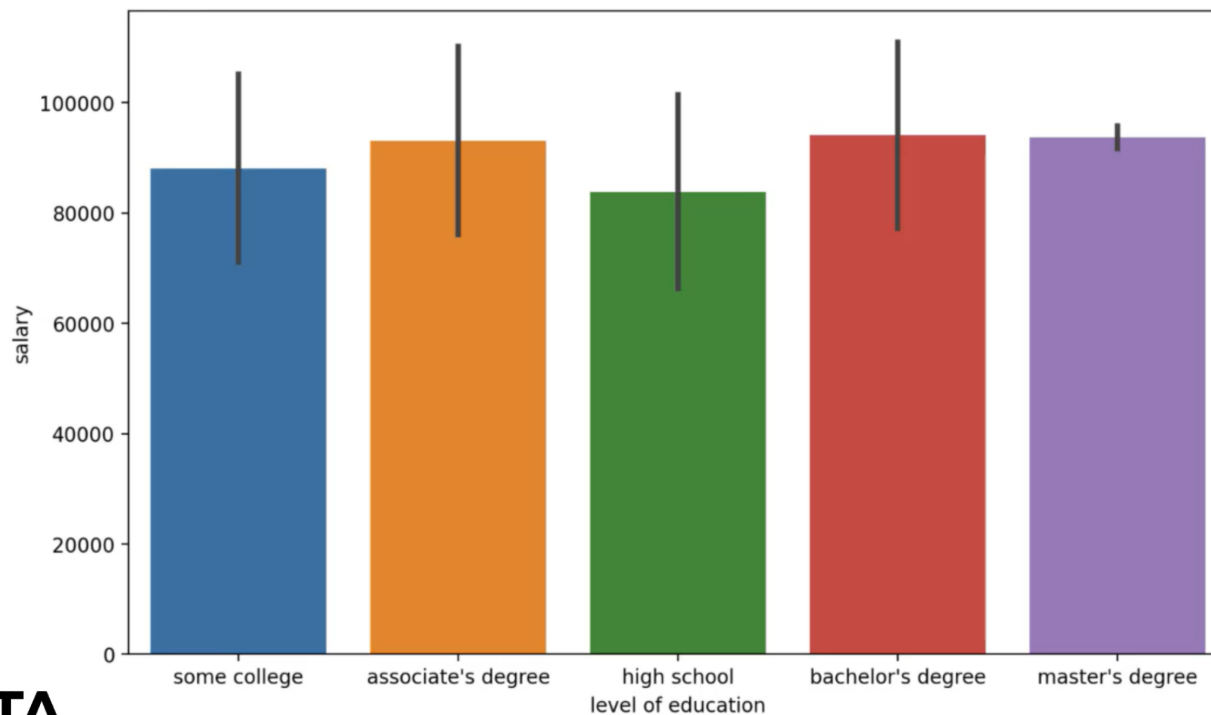
- Barplot со средним значением и среднеквадратичным отклонением





Categorical Plots – графики по категориям

- Является ли Barplot лучшим способом отобразить эти значения?
- Вполне возможно что нет! Это просто единичные значения.





Categorical Plots – графики по категориям

- Вероятно их проще отобразить в виде обычной таблицы:

	mean	std
level of education		
associate's degree	93156.41	17066.06
bachelor's degree	94133.76	17007.09
high school	83887.35	17674.44
master's degree	93718.00	2497.63
some college	88115.84	17076.28



Categorical Plots – графики по категориям

- В следующей лекции мы посмотрим, как нарисовать эти графики в seaborn!



Categorical Plots

Часть 2 – пишем код



Categorical Plots

Часть 3



Categorical Plots – графики по категориям

- Мы посмотрели графики distribution plots для одной переменной. Но что, если мы хотим сравнить распределения для разных категорий?
- Например, вместо общего распределения зарплат, мы хотим сравнить распределение зарплат отдельно для каждого уровня образования?



Categorical Plots – графики по категориям

- Распределение по категориям:
 - Boxplot
 - Violinplot
 - Swarmplot
 - Boxenplot (Letter-Value Plot)
- Посмотрим эти графики для примера с зарплатами



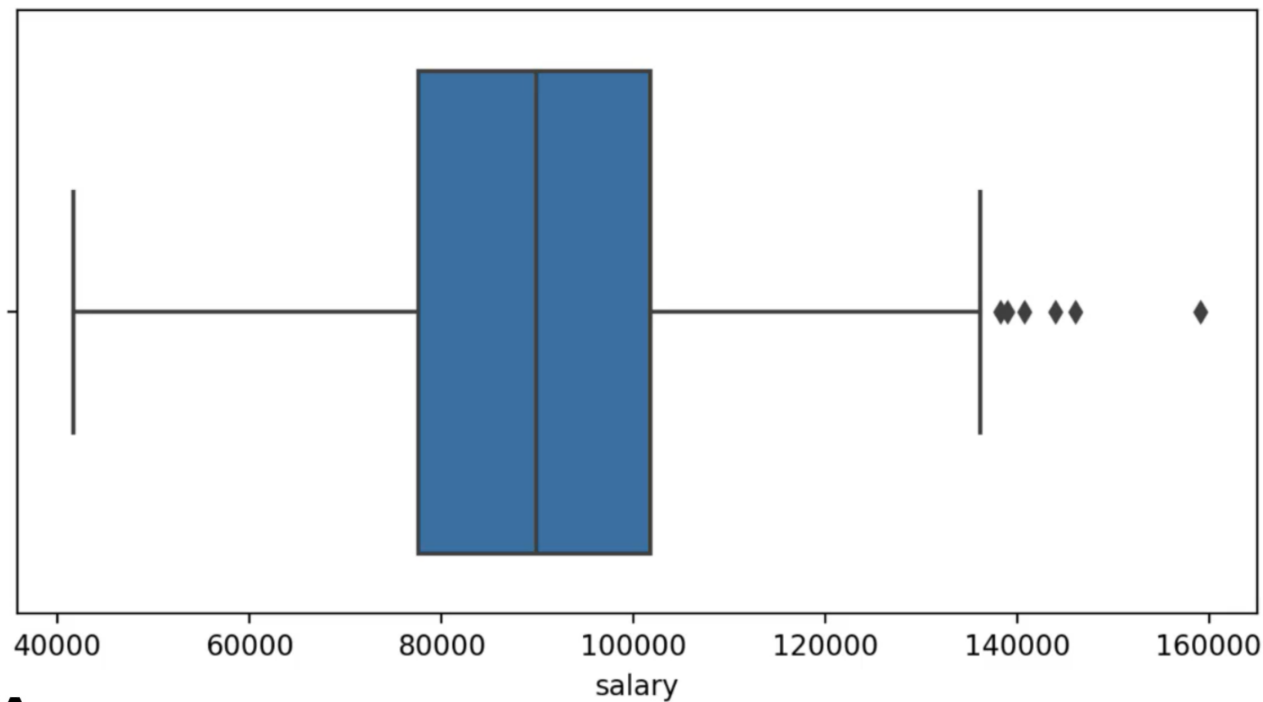
Categorical Plots – графики по категориям

- Boxplot отображает распределение непрерывной переменной
- Он использует квартили
- Квартили разделяют все данные на 4 равных интервала по количеству точек:
 - 25% точек в нижнем квартиле
 - 50-й процентиль (Q2) это медиана



Categorical Plots – графики по категориям

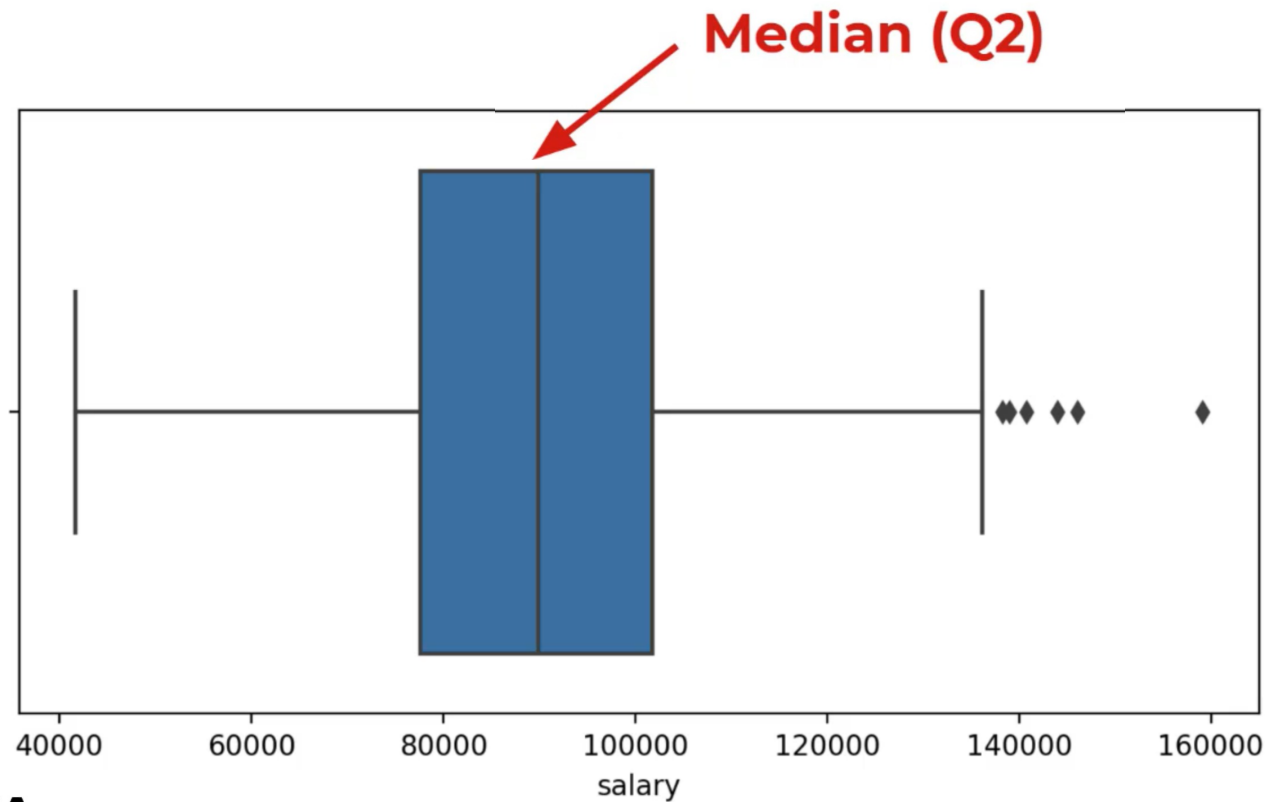
- Boxplot для одной переменной





Categorical Plots – графики по категориям

- Медиана разбивает данные пополам

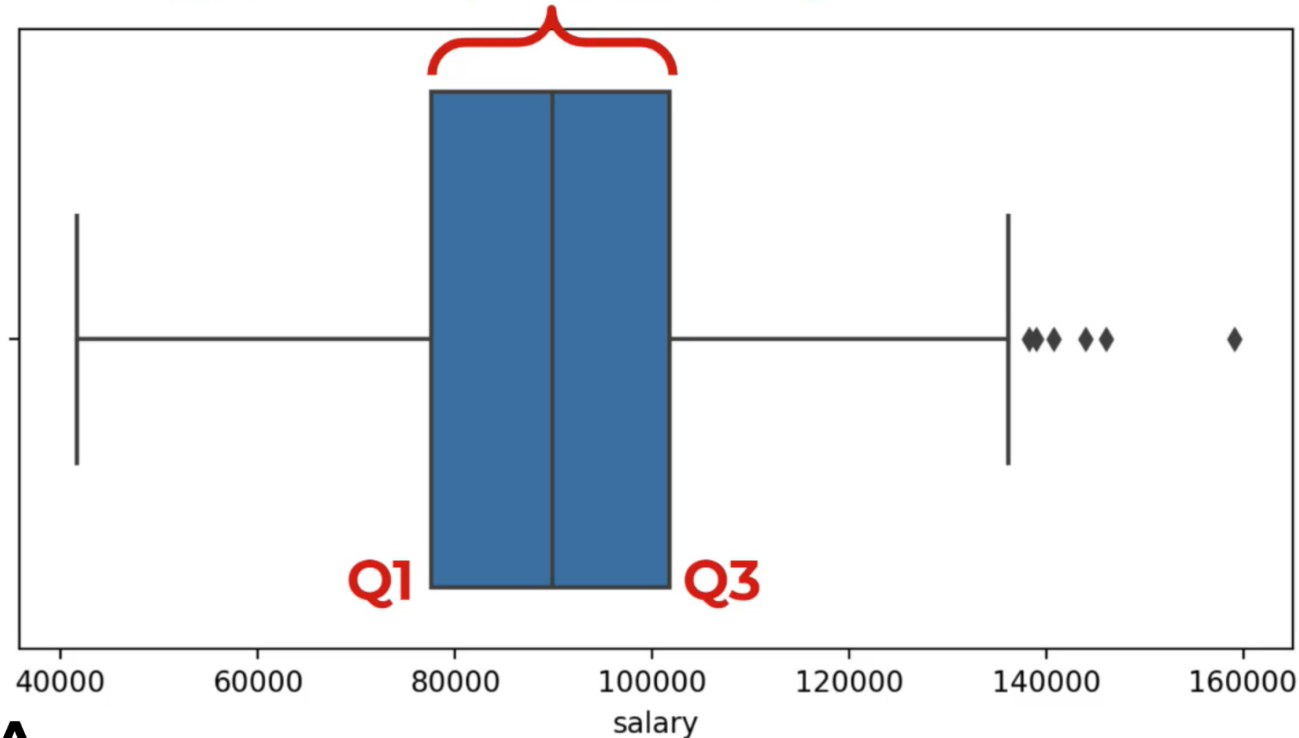




Categorical Plots – графики по категориям

Интерквартильный размах – от Q1 до Q3. Это 50% точек.

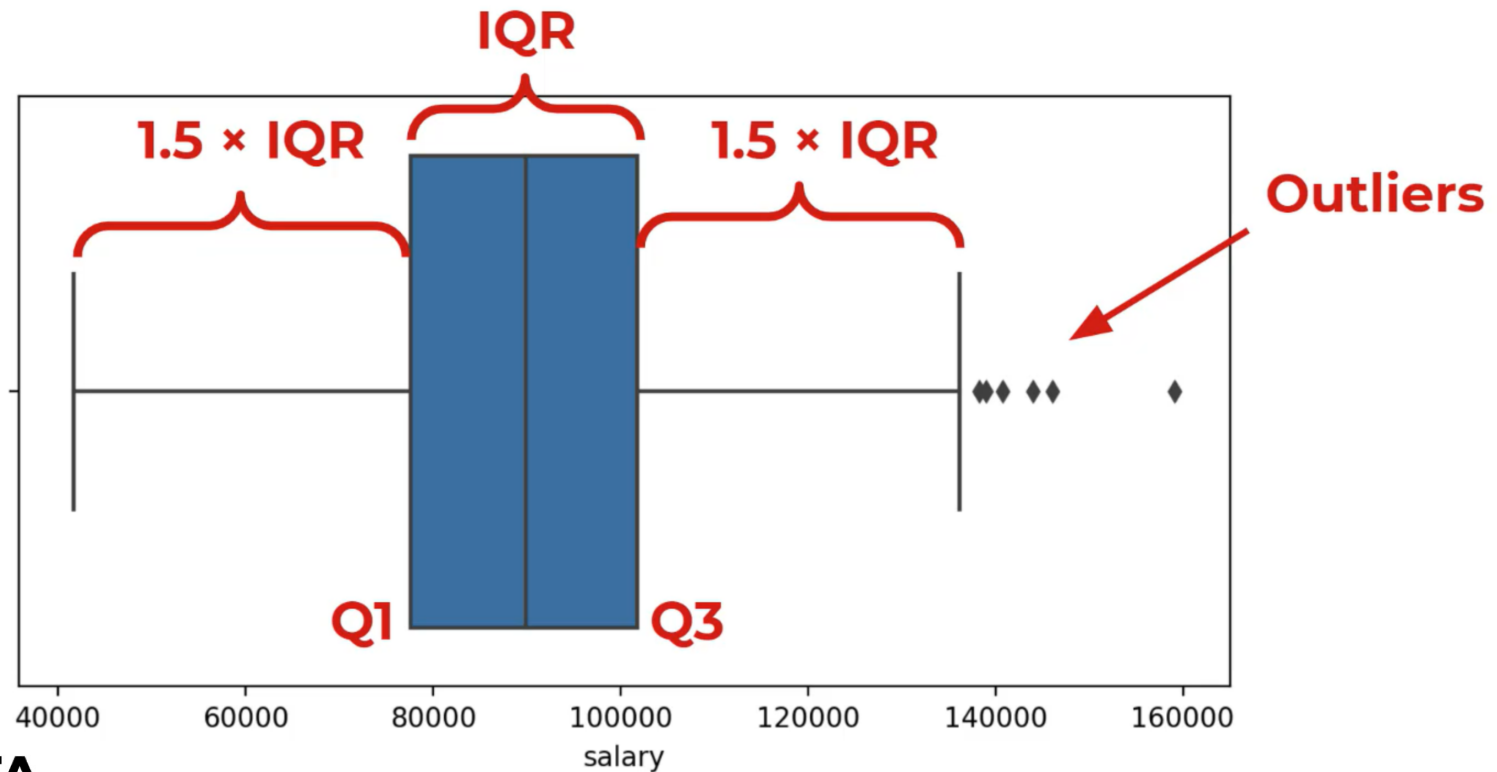
IQR - Interquartile Range





Categorical Plots – графики по категориям

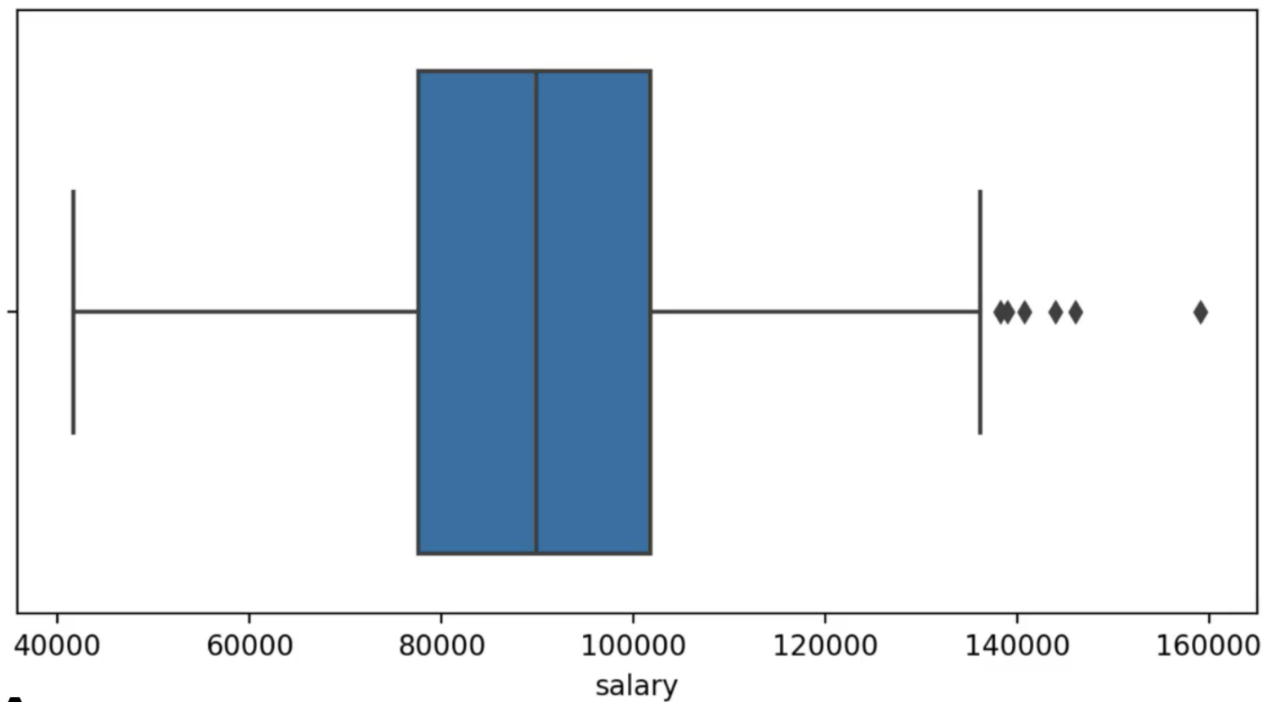
“Усы” (whiskers) определяются как $1.5 * IQR$.
Значения вне этого интервала – выбросы (outliers).





Categorical Plots – графики по категориям

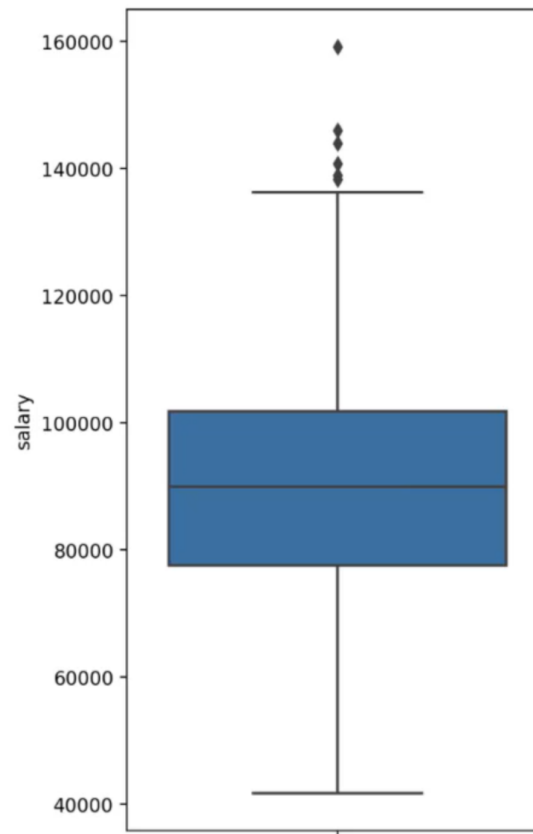
- Boxplot визуально отображает основные статистические характеристики распределения





Categorical Plots – графики по категориям

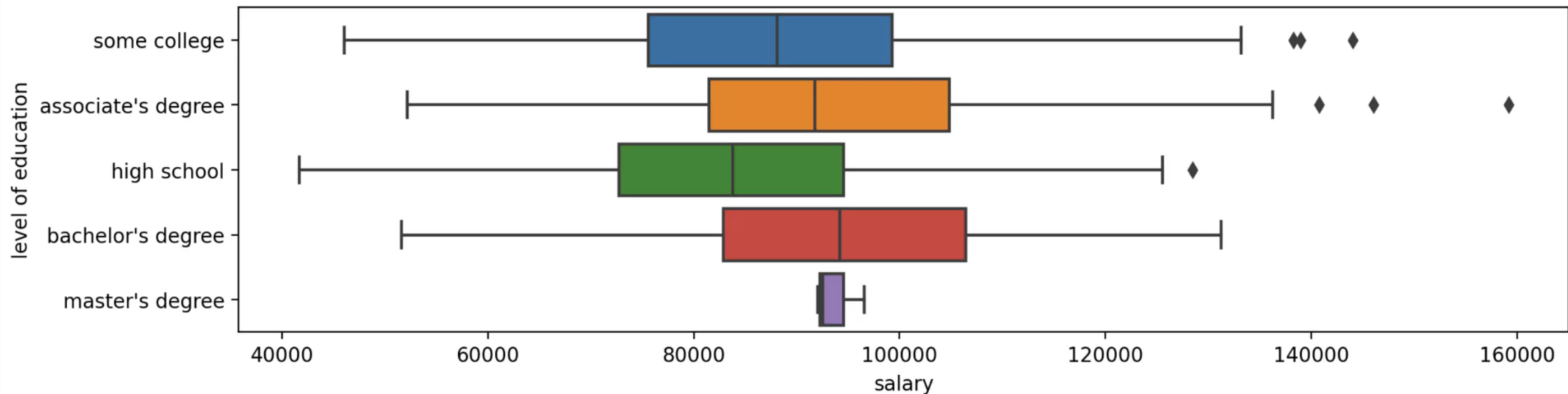
- Boxplot может располагаться по горизонтали и по вертикали





Categorical Plots – графики по категориям

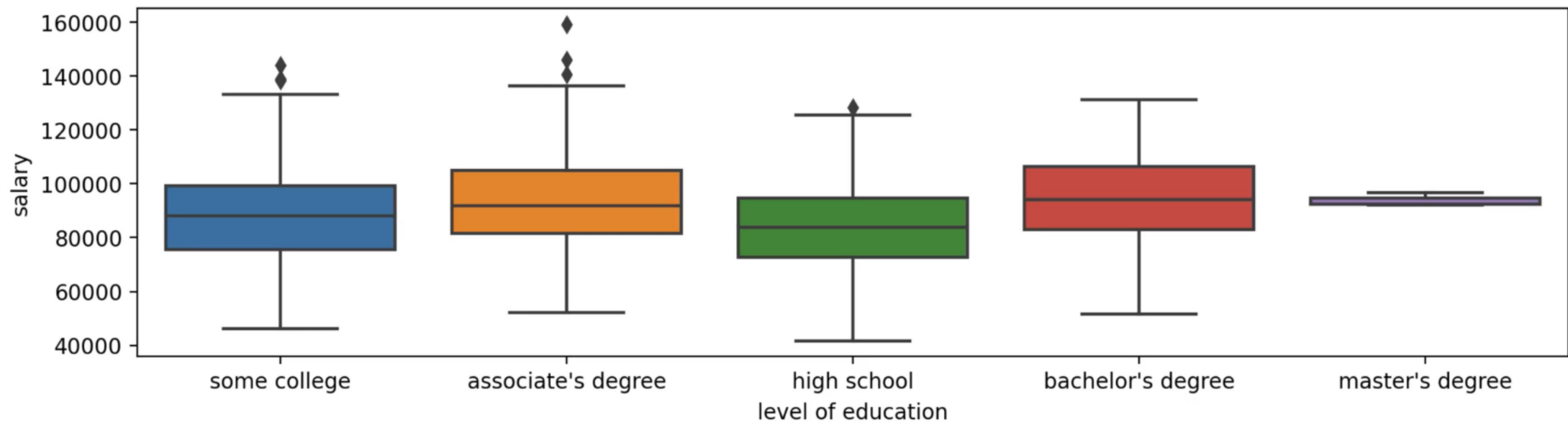
- Мы можем нарисовать Boxplot'ы по категориям!





Categorical Plots – графики по категориям

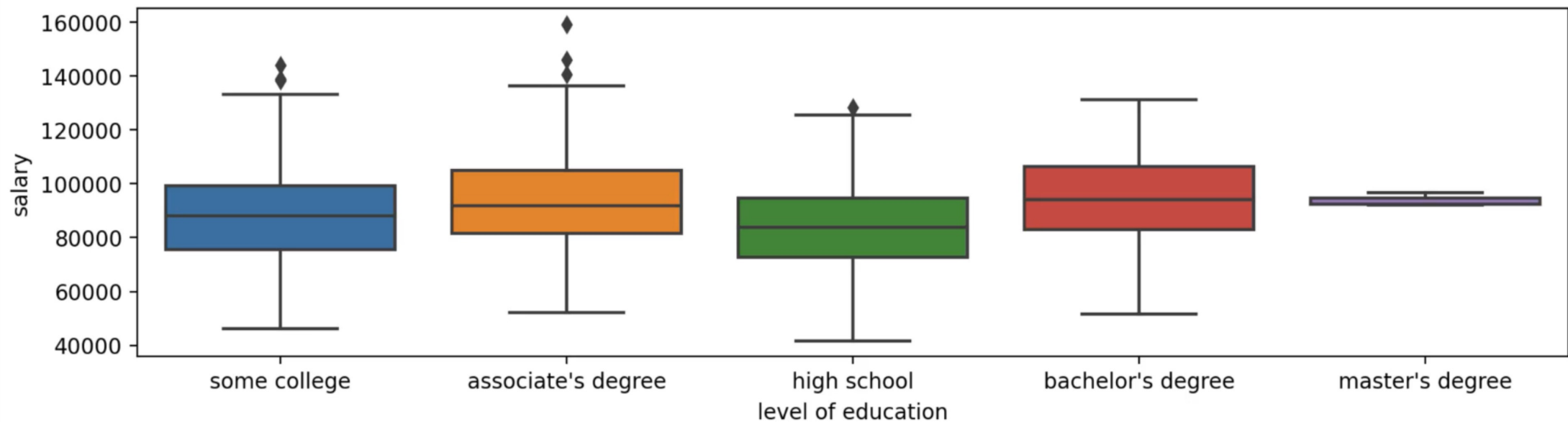
- Мы можем нарисовать Boxplot'ы по категориям!





Categorical Plots – графики по категориям

- Мы можем нарисовать Boxplot'ы по категориям!





Categorical Plots – графики по категориям

- Violinplot очень похож на Boxplot
- Он отображает плотность распределения по данным с помощью KDE
- Его можно представить себе как график KDE, отражённый в зеркале



Categorical Plots – графики по категориям

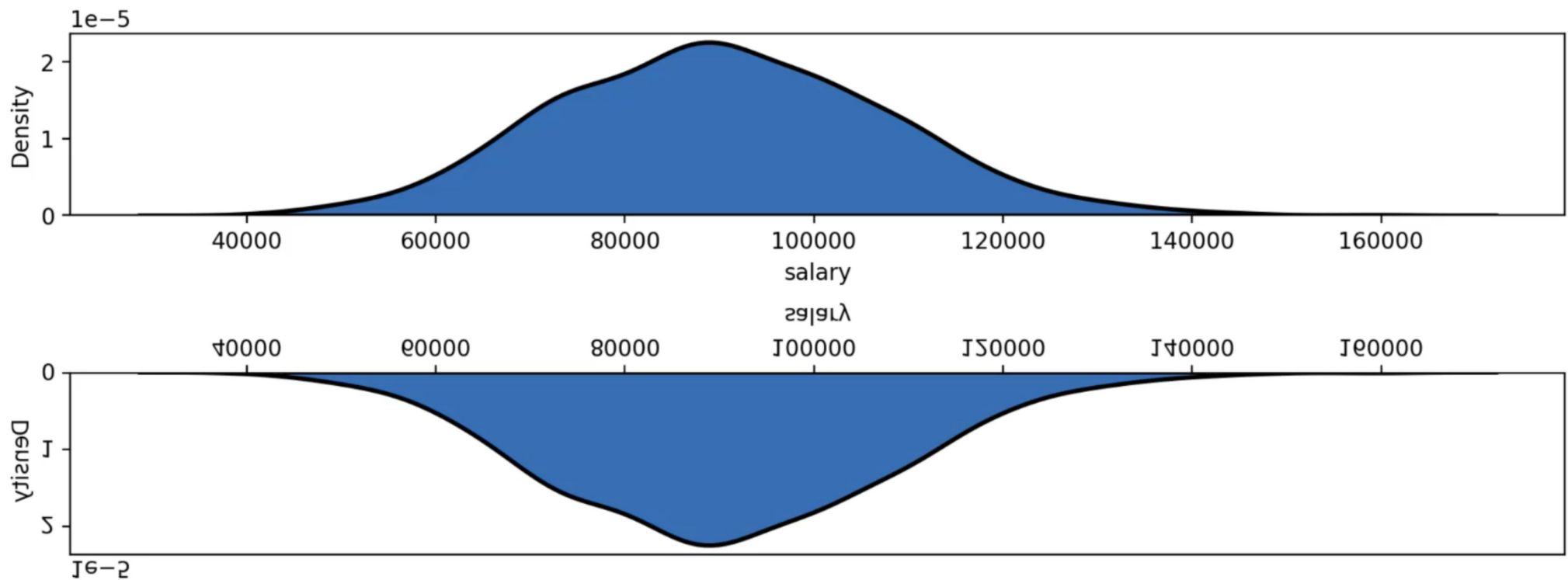
- Мы берём KDE одной переменной:





Categorical Plots – графики по категориям

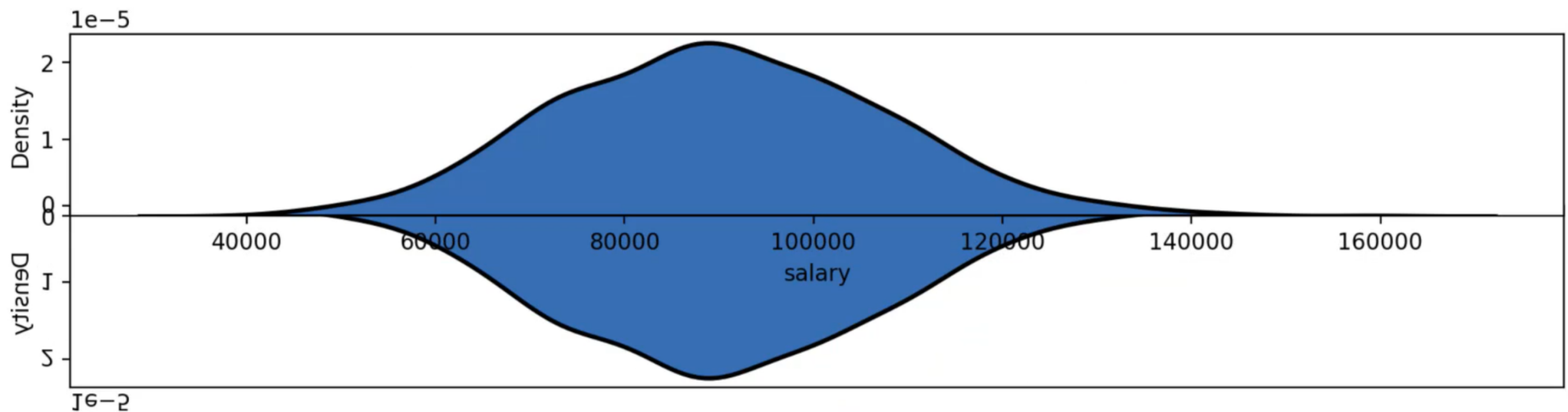
- Далее делаем зеркальное отражение:





Categorical Plots – графики по категориям

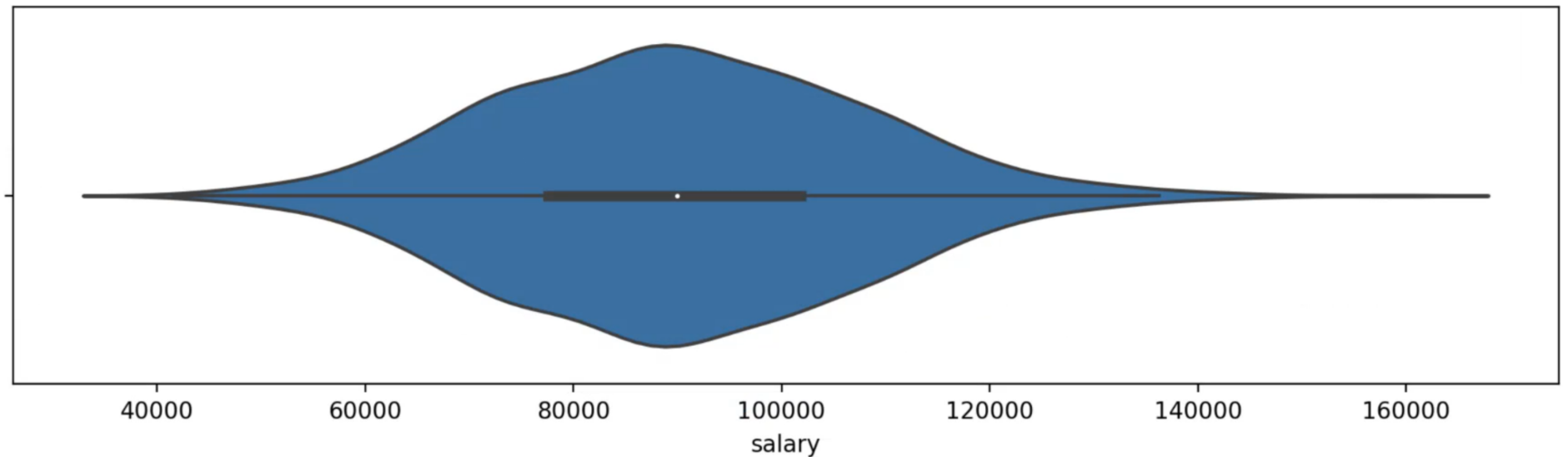
- Объединяем их – получаем Violinplot:





Categorical Plots – графики по категориям

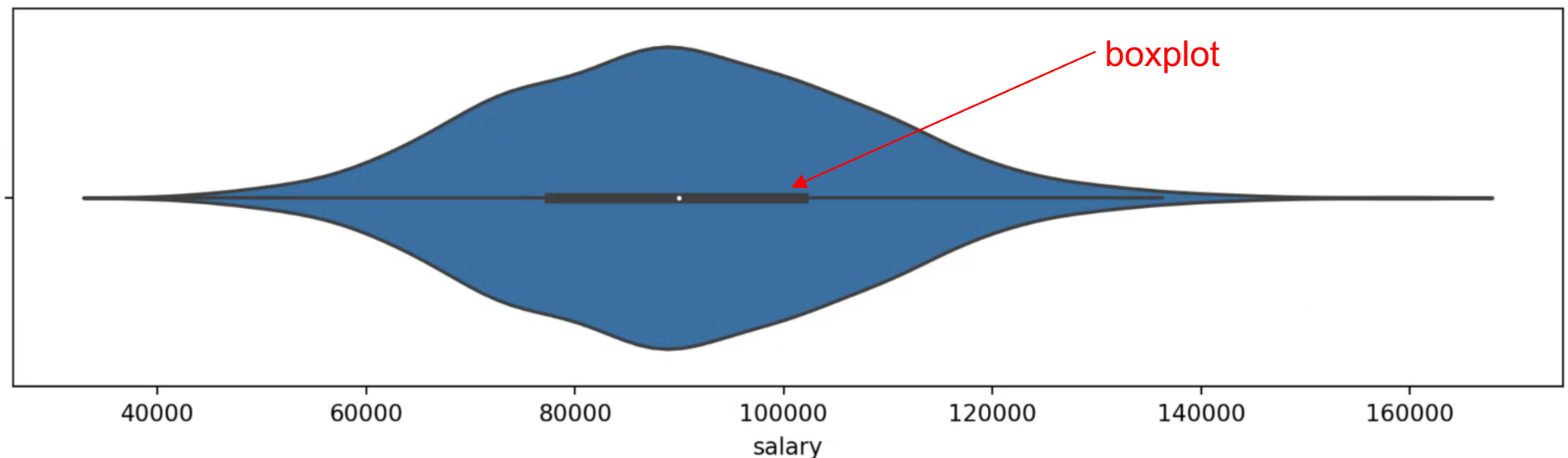
- Объединяем их – получаем Violinplot:





Categorical Plots – графики по категориям

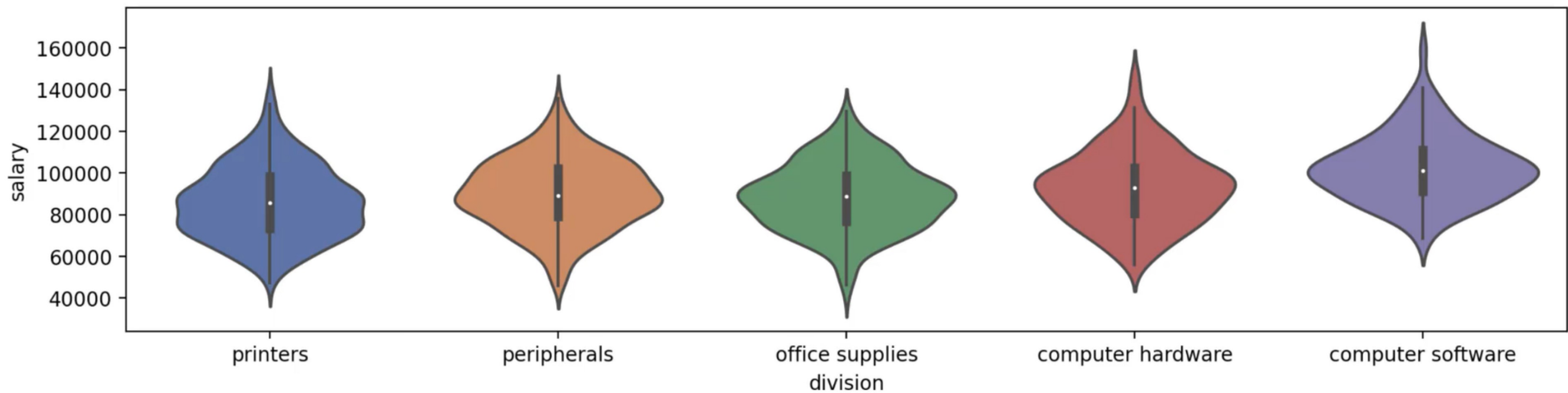
- Внутри violinplot может располагаться миниатюрный boxplot (за это отвечает параметр `inner`):





Categorical Plots – графики по категориям

- Violinplot можно нарисовать по категориям:





Categorical Plots – графики по категориям

- Намного реже применяются графики `swarmplot` и `boxenplot`
- Давайте быстро на них тоже посмотрим...



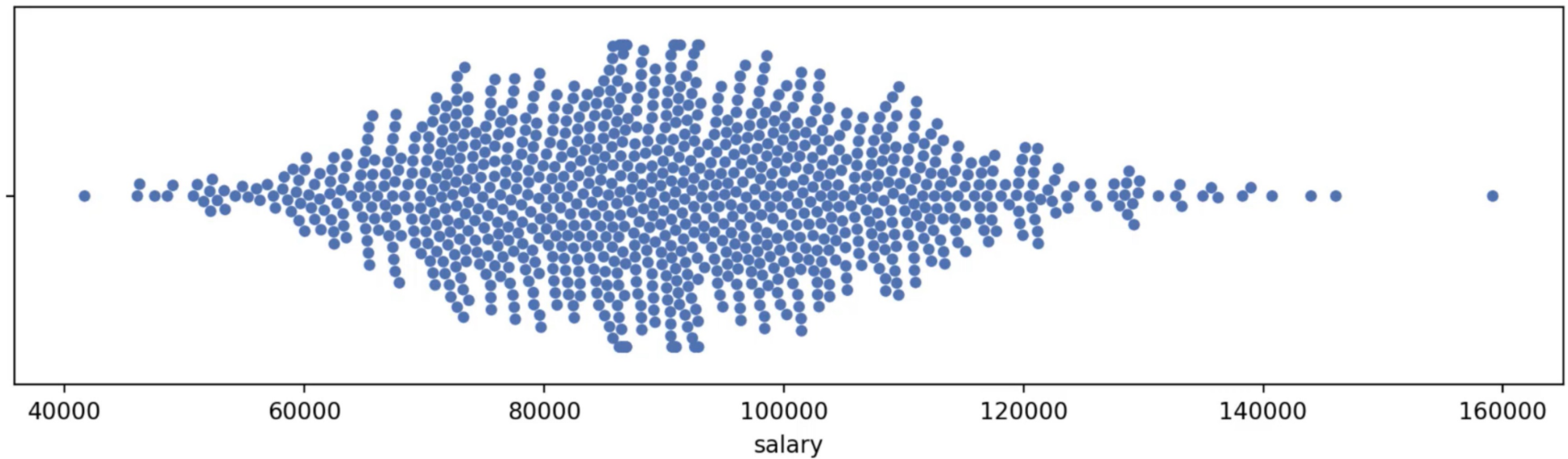
Categorical Plots – графики по категориям

- swarmplot просто отображает все точки в распределении
- Для очень больших наборов данных он показывает не все точки, а общее их распределение



Categorical Plots – графики по категориям

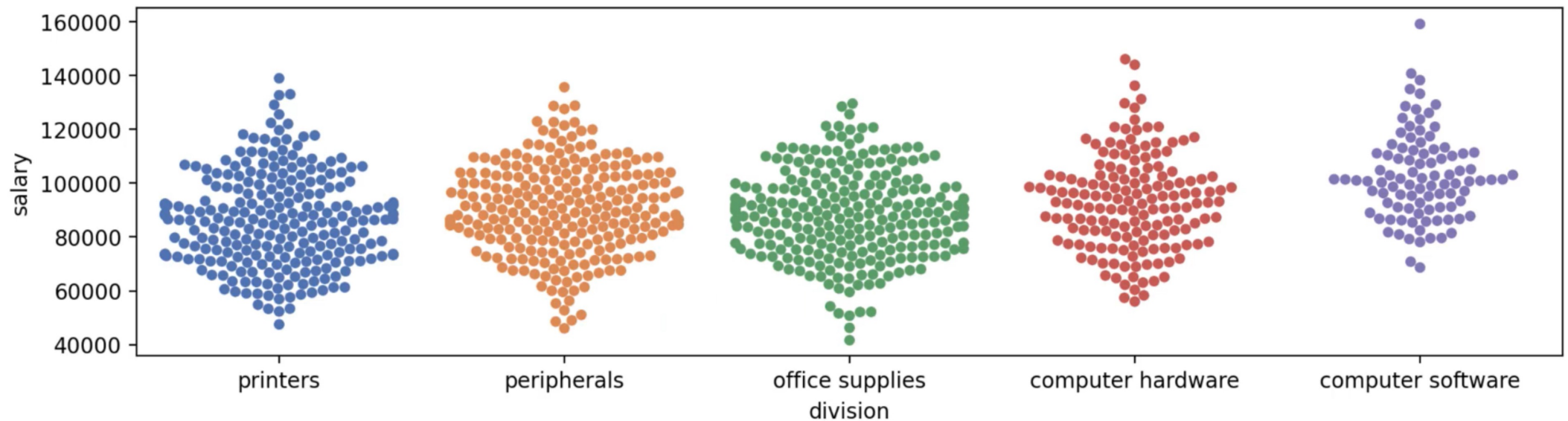
- swarmplot





Categorical Plots – графики по категориям

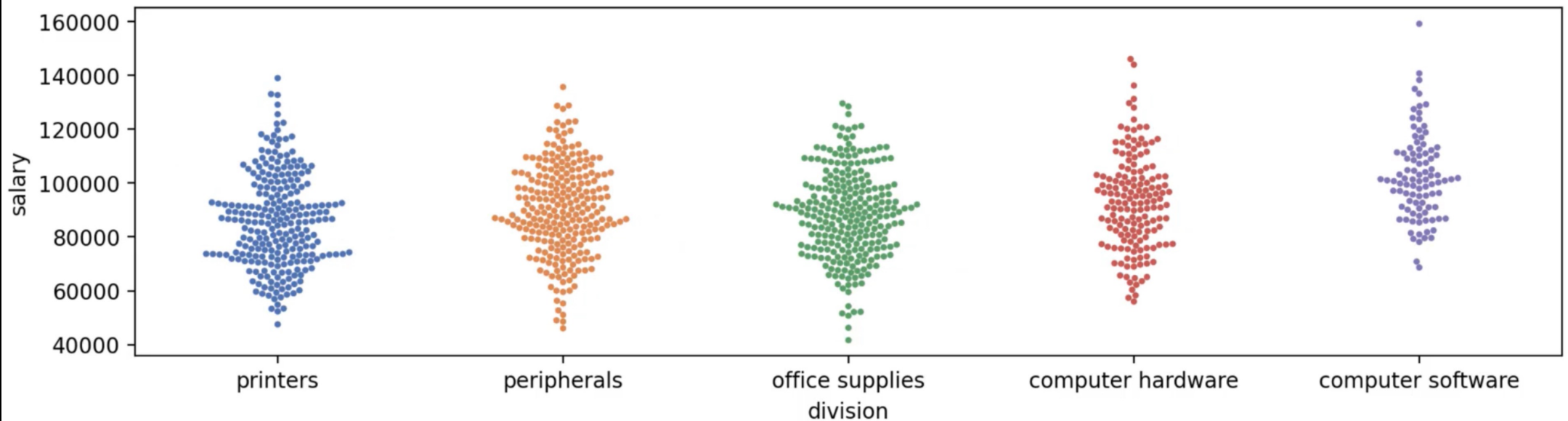
- swarmplot по категориям





Categorical Plots – графики по категориям

- Можно поменять размер точек





Categorical Plots – графики по категориям

- boxenplot (Letter-value plot) был разработан как расширение обычного boxplot
- Если будете использовать этот график, то прочитайте статью о нём, ссылка на которую есть в блокноте для этой лекции



Categorical Plots – графики по категориям

- boxenplot используется достаточно редко, и если искать его в Google, то он скажет что скорее всего Вы имели ввиду boxplot
- Используйте этот график только в случае, если Ваша аудитория знакома с этим графиком
- Давайте кратко обсудим boxenplot и его преимущества



Categorical Plots – графики по категориям

- Используя набор буква-значение, мы можем вместо строгих квартилей использовать несколько квантилей

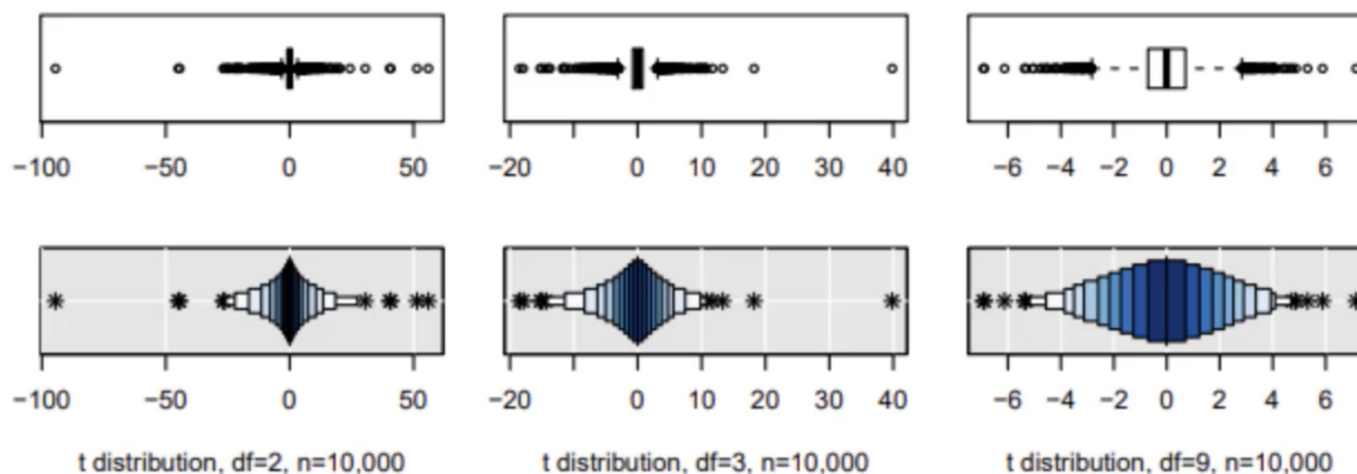
LV	ideal tail area	rough %	odds (2^i)	SEfactor	n-equiv*
M	.50	50.0%	2	1.253314	
F	.25	25.0%	4	1.36	1.0
E	.125	12.5%	8	1.60	1.4
D	.0625	6.25%	16	1.96	2.1
C	.03125	3.13%	32	2.47	3.3
B	.015625	1.56%	64	3.16	5.4
A	.0078125	0.8%	128	4.10	9.1
Z	.00390625	0.4%	256	5.37	15.6
Y	.001953125	0.2%	512	7.11	27.3
X	.0009765625	0.1%	1,024	9.48	48.4
W	.00048828125	0.05%	2,048	12.70	87.0
V	.000244140625	0.024%	4,096	17.11	157.7
U	.0001220703125	0.012%	8,192	23.14	288.5
T	.00006103515625	0.006%	16,384	31.40	531.3
S	.000030517578125	0.003%	32,768	42.75	984.4
R	.0000152587890625	0.0015%	65,536	58.34	1833.5
Q	.00000762939453125	0.0008%	131,072	79.80	3430.5
P	.000003814697265625	0.0004%	252,144	109.38	6444.3
O	.0000019073486328125	0.0002%	504,288	150.19	12149.2
N	.00000095367431640625	0.0001%	1,008,576	206.55	22977.6

Table 1: First 20 letter values. Ideal tail area is 2^{-i} , $i = 1, \dots, 20$. rough% rounds $2^{-i} \times 100\%$ to the first 1 or 2 nonzero digits. odds expresses tail area as 1 in 2^i . SEfactor gives the factor for the asymptotic standard error of the order statistic (from a Gaussian population, variance σ^2) corresponding to tail area, i.e., $SE(LV) \approx SEfactor \times \sigma / \sqrt{n}$, where $SEfactor = \sqrt{p_i(1-p_i)/\phi(\Phi^{-1}(p_i))}$, $p_i = \text{tail area} = 2^{-i}$. n-equiv = $(SEfactor/1.362633)^2$ which gives the factor of increase in sample size for the uncertainty in that letter value to be the same as that for the fourth; e.g., need $1.4n$ (respectively, $2.1n$) observations for the eighth (respectively, sixteenth) to have the same uncertainty as that of a fourth from a sample of size n .



Categorical Plots – графики по категориям

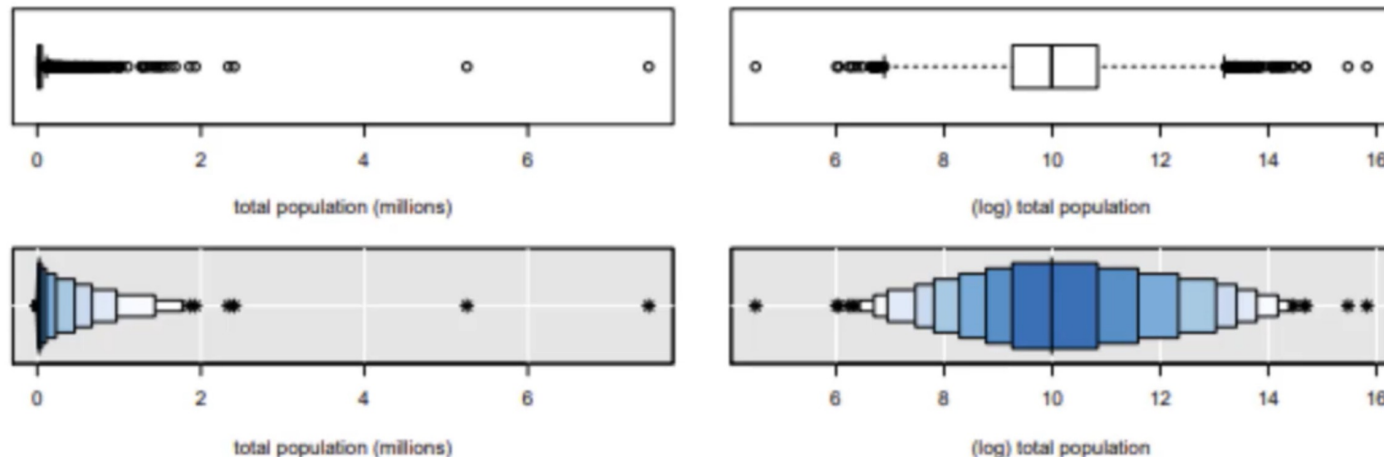
- Вохеплот отображает квантили буква-значение на стандартном boxplot:





Categorical Plots – графики по категориям

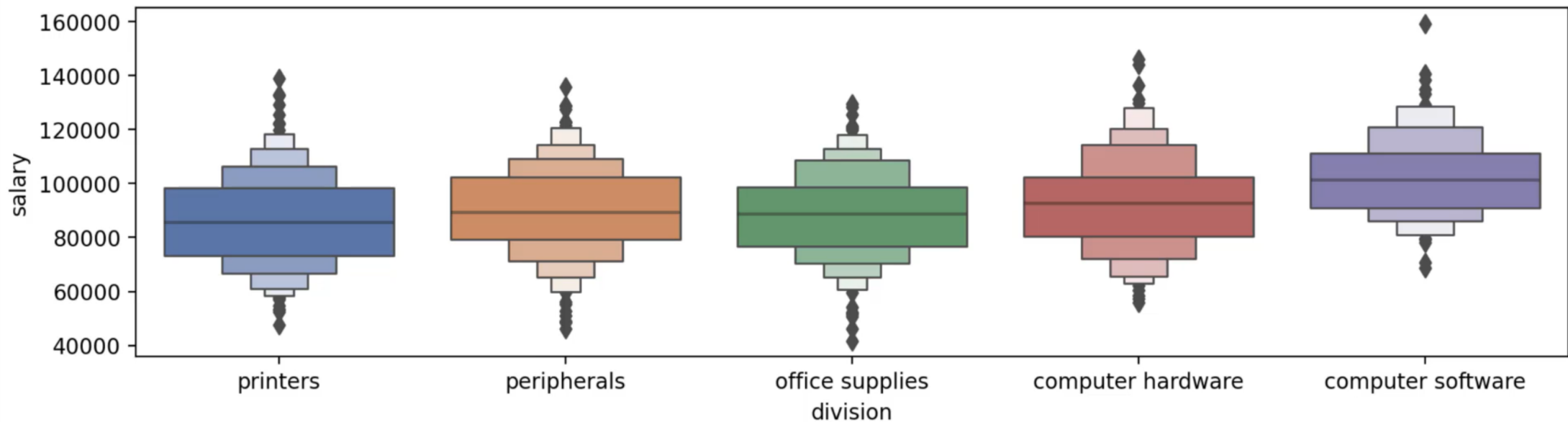
- Вохеплот отображает квантили буква-значение на стандартном boxplot:





Categorical Plots – графики по категориям

- Boxenplot в seaborn:





Categorical Plots – графики по категориям

- Помните, главная задача любого графика – отобразить нужную информацию, а не запутать или впечатлить кого-то редкими графиками!
- В следующей лекции мы посмотрим синтаксис для этих графиков.



Categorical Plots

Часть 2 – код в seaborn



Comparison Plots

часть 1



Comparison Plots – графики сравнения

- Графики сравнения – это двухмерные версии тех графиков, которые мы уже видели
- Мы обсудим два типа графиков:
 - `jointplot()`
 - `pairplot()`



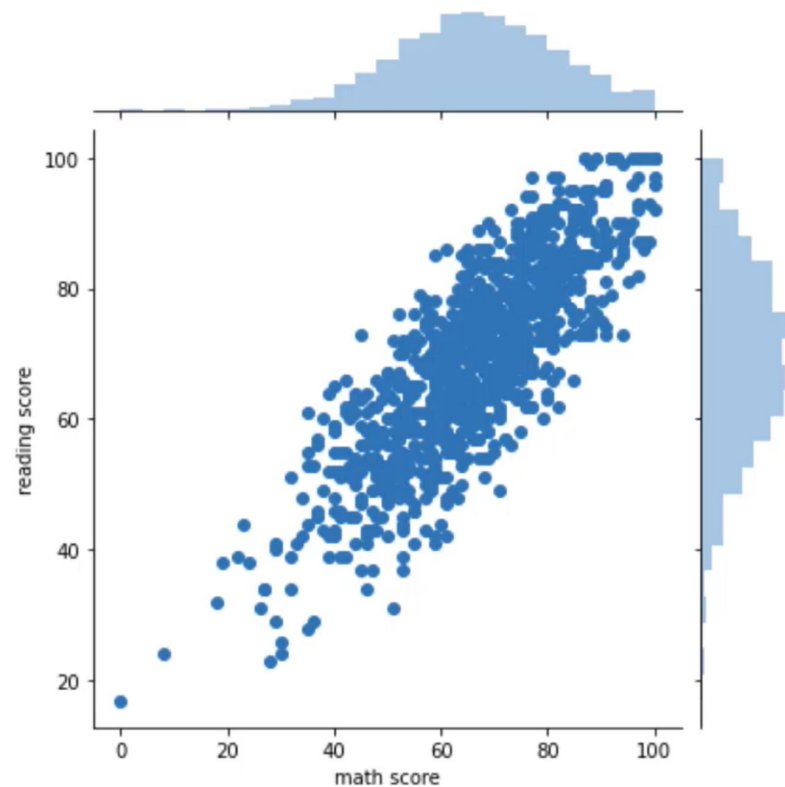
Comparison Plots – графики сравнения

- `jointplot()`:
 - Мы можем взять гистограммы для каждой переменной графика `scatterplot`, чтобы уточнить распределение данных по каждой переменной
 - Кроме `scatterplot` можно также использовать `hex plot` или двумерный `KDE plot`



Comparison Plots – графики сравнения

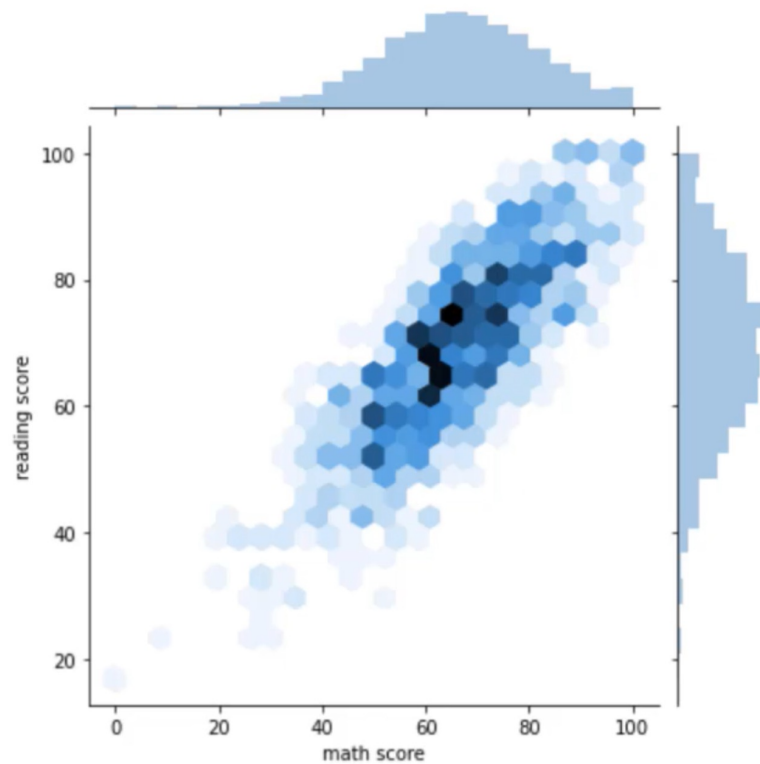
- Гистограммы и scatterplot





Comparison Plots – графики сравнения

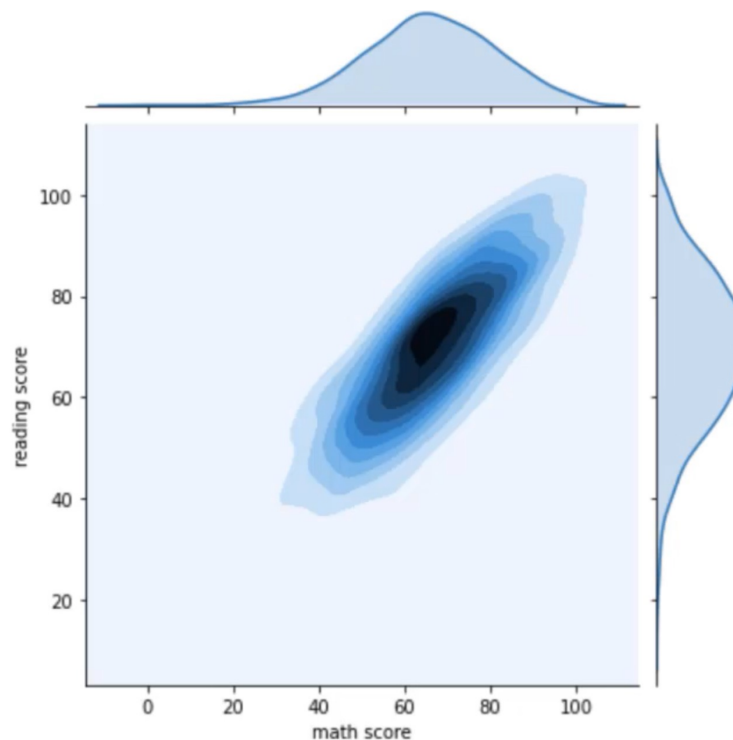
- Гистограммы и hex plot (hexagon plot)
- Шестиугольники темнее там, где больше точек





Comparison Plots – графики сравнения

- Двумерные KDE-графики показывают с помощью теней распределение между двумя KDE





Comparison Plots – графики сравнения

- `pairplot()`:
 - Это быстрый способ сравнить между собой все числовые переменные в наборе данных
 - Автоматически создаёт гистограмму для каждой переменной, и сравнительные графики `scatterplot` для всех комбинаций переменных



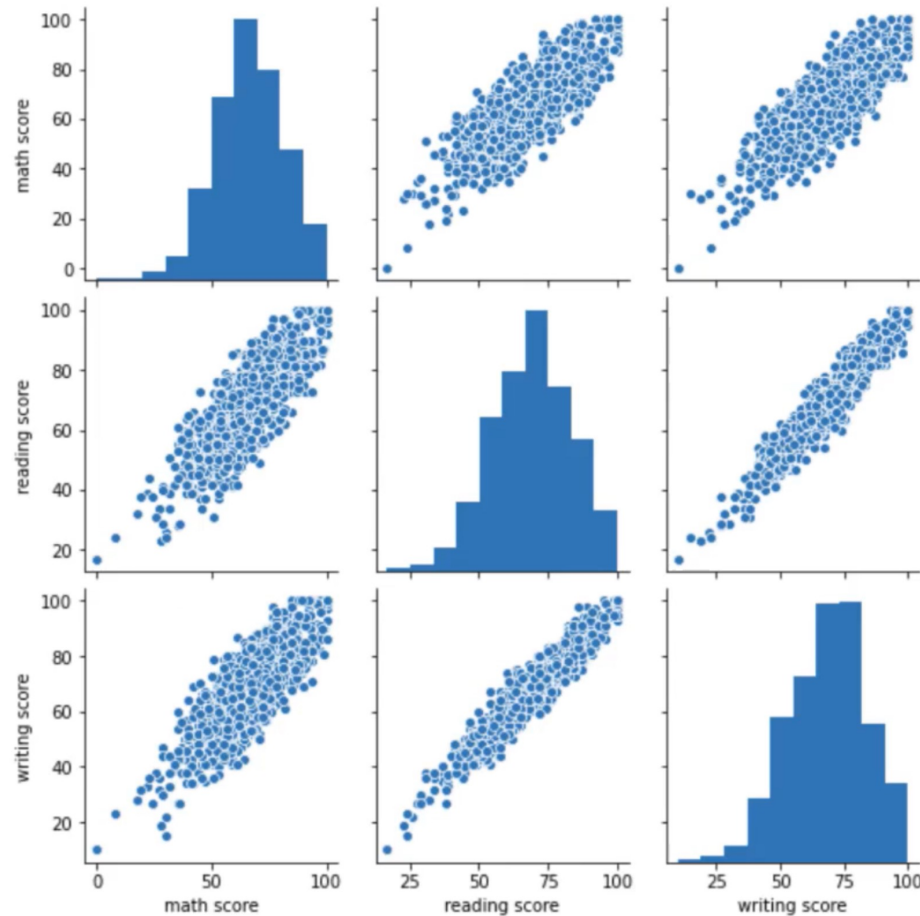
Comparison Plots – графики сравнения

- `pairplot()`:
 - Внимание!! `pairplot()` может потреблять много CPU и памяти в случае больших объёмов данных и большого количества колонок
 - Хорошая идея – сначала оставить только те колонки, которые нужно исследовать



Comparison Plots – графики сравнения

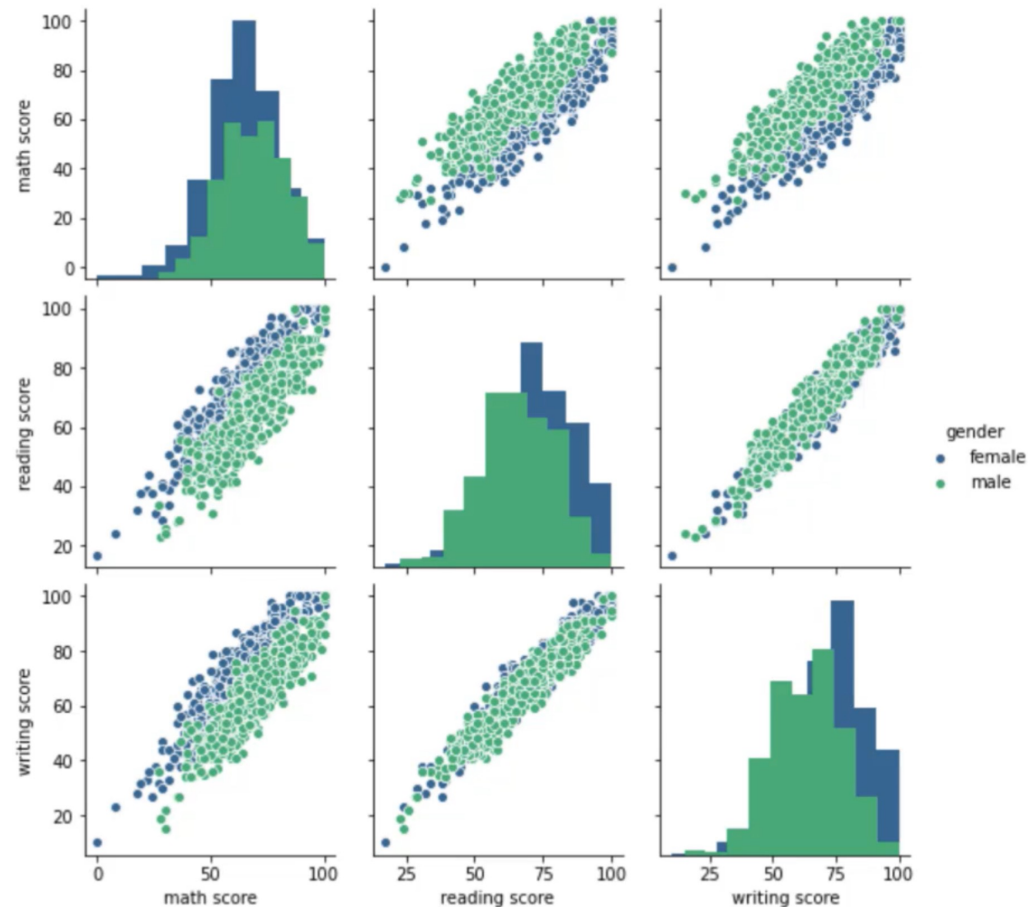
- `pairplot()`





Comparison Plots – графики сравнения

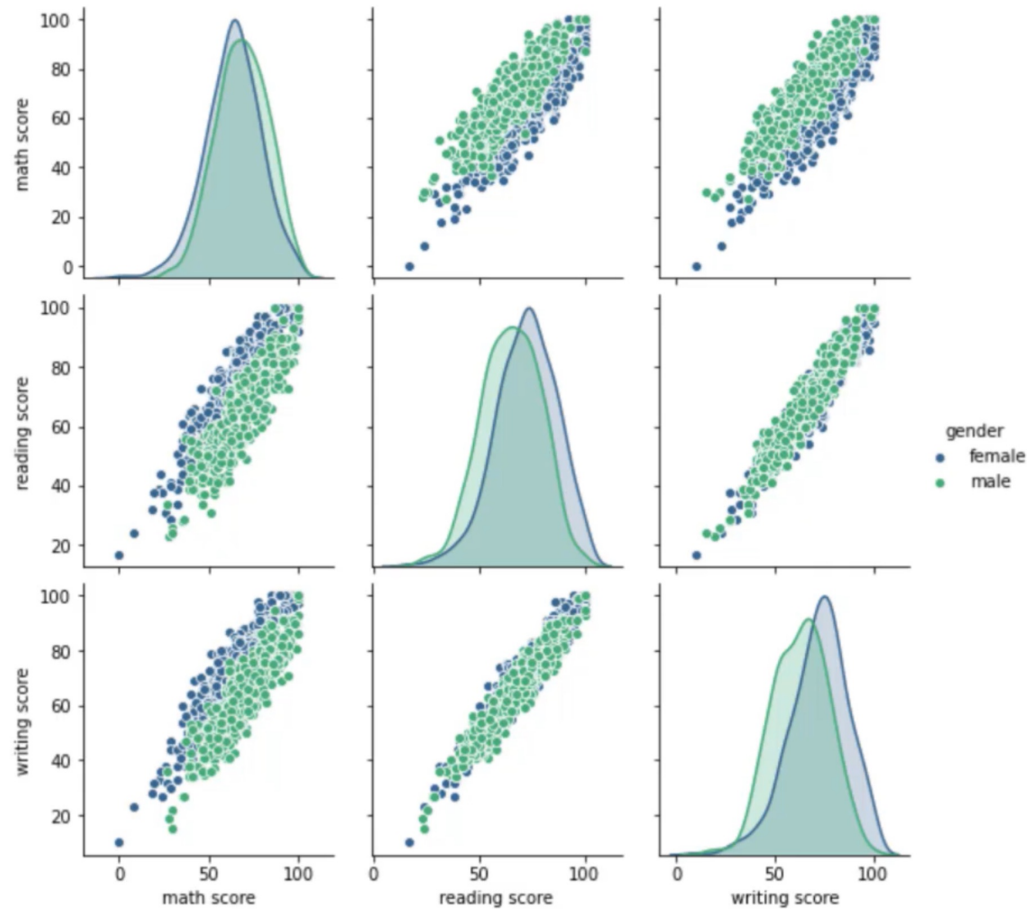
- `pairplot()`





Comparison Plots – графики сравнения

- `pairplot()`





Comparison Plots – графики сравнения

- В следующей лекции мы посмотрим код для этих графиков!



Comparison Plots

часть 2 – пишем код



Seaborn grids



Seaborn grids

- Многие встроенные методы seaborn внутри себя используют seaborn grid
- Если вызывать его напрямую, то мы получим возможность настроить различные параметры графиков.



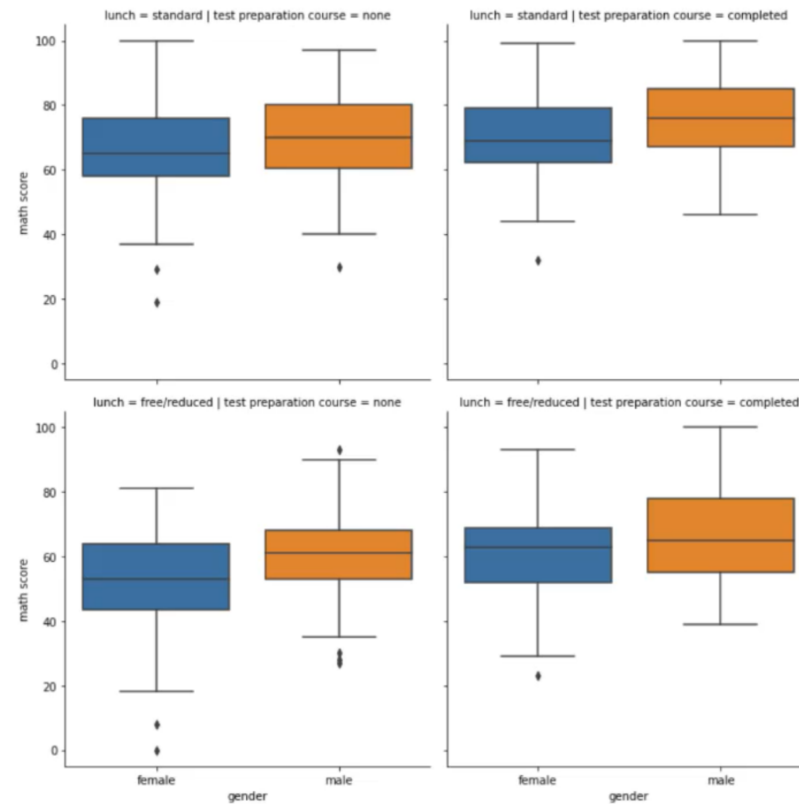
Seaborn grids

- Seaborn grid – это способ построения набора графиков рядом друг с другом. В основе лежит `subplots()` из Matplotlib, с помощью которого автоматически создаётся набор графиков для категориальной переменной
- Указывать количество колонок и строк не нужно, вместо этого мы указываем название колонки, и seaborn автоматически создаёт набор графиков (`subplot`)



Seaborn grids

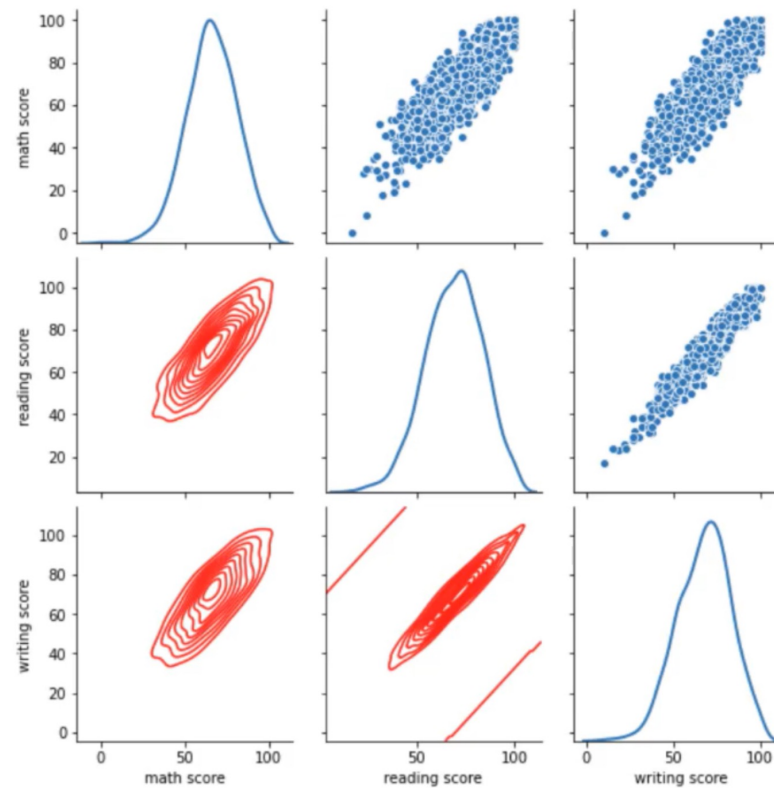
- Создание набора графиков на основе grid:





Seaborn grids

- Графики на основе pairplot() grid:





Seaborn grids

- Это проще понять изучением кода, поэтому давайте перейдём в блокнот!



Matrix plots



Matrix plots - матричные графики

- Матричные графики это визуальный аналог отображения сводной таблицы
- Матричный график показывает все данные, визуализируя все числовые значения в датафрейме
- Замечание! Не каждый датафрейм хорошо подходит для матричных графиков, например для тепловой карты heatmap



Matrix plots - матричные графики

- Два основных типа матричных графиков:
 - `heatmap()` – тепловая карта – изображает распределение значений по ячейкам с помощью цветов
 - `clustermat()` – карта кластеров – похожа на `heatmap`, но сначала выполняет иерархическую кластеризацию, объединяя данные в группы



Matrix plots - матричные графики

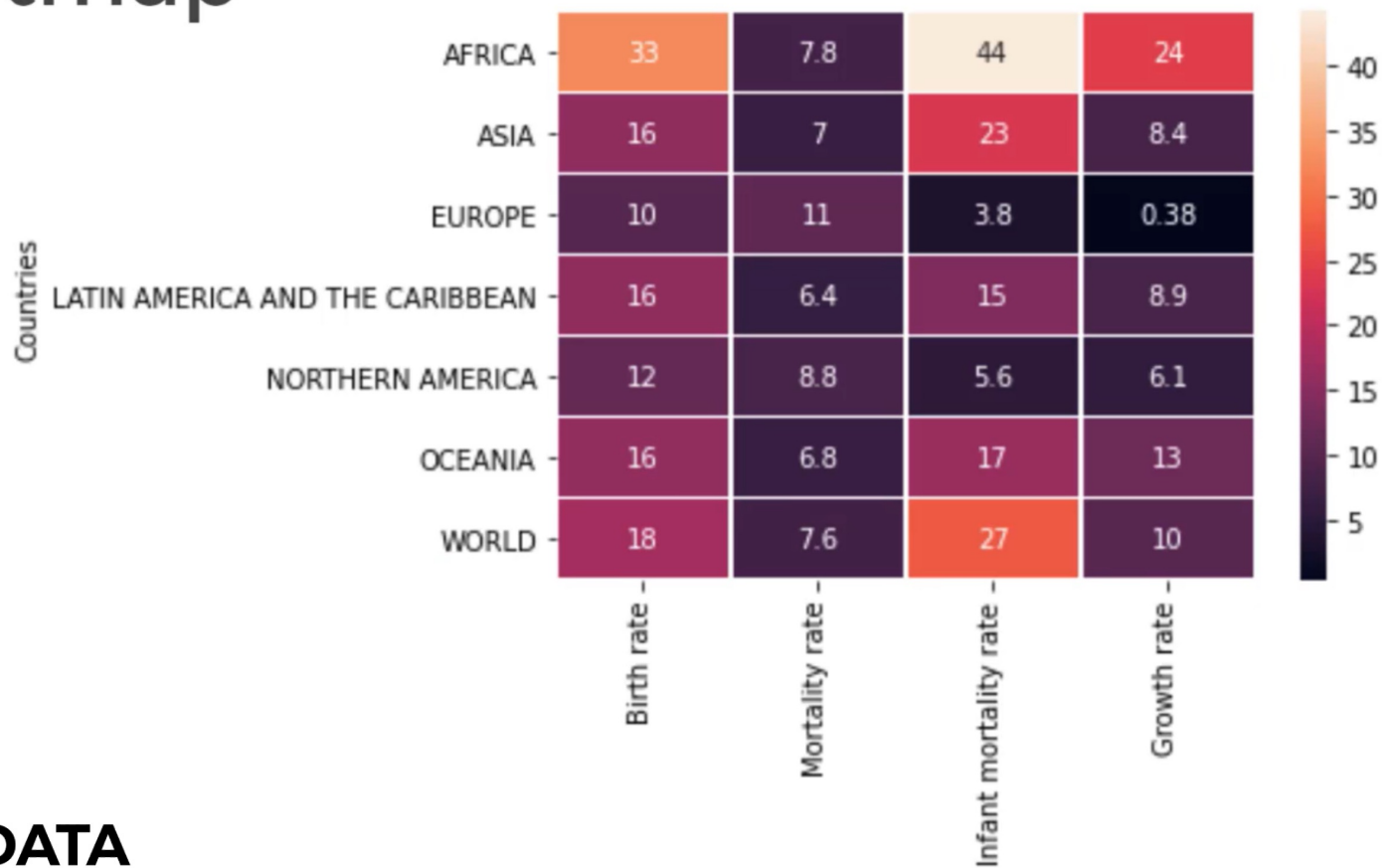
- Heatmap

	Birth rate	Mortality rate	Life expectancy	Infant mortality rate	Growth rate
Countries					
AFRICA	32.577	7.837	63.472	44.215	24.40
ASIA	15.796	7.030	73.787	23.185	8.44
EUROPE	10.118	11.163	78.740	3.750	0.38
LATIN AMERICA AND THE CARIBBEAN	15.886	6.444	75.649	14.570	8.89
NORTHERN AMERICA	11.780	8.833	79.269	5.563	6.11
OCEANIA	16.235	6.788	78.880	16.939	12.79
WORLD	17.963	7.601	72.766	27.492	10.36



Matrix plots - матричные графики

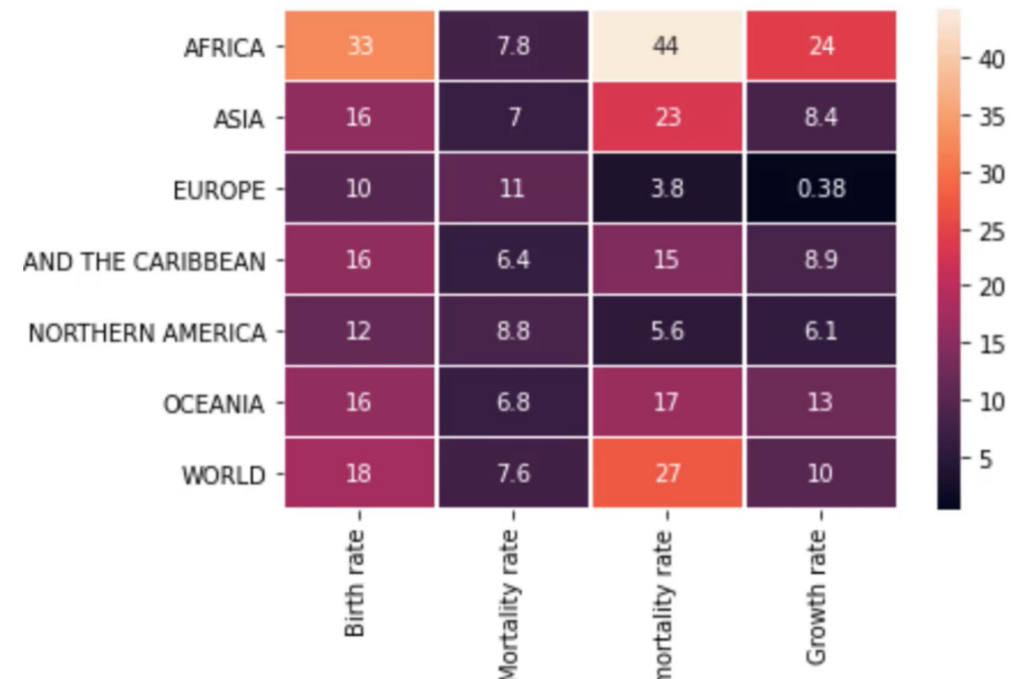
- Heatmap





Matrix plots - матричные графики

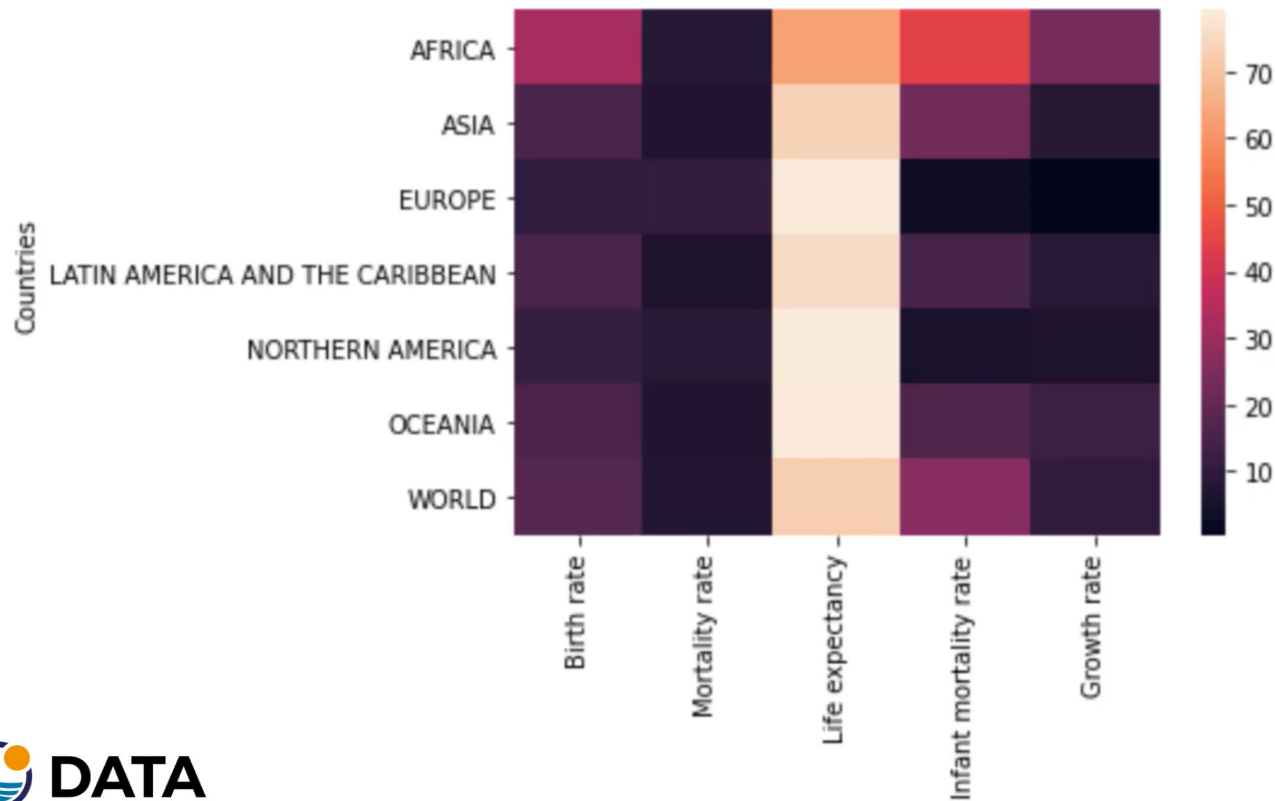
- heatmap выглядит лучше всего, если все переменные измеряются в одних и тех же единицах – тогда цвета имеют один и тот же смысл для всех переменных
- В нашем примере показатели “... rate” измерялись в процентах прироста





Matrix plots - матричные графики

- Представьте, что мы добавили ещё возраст (life expectancy)





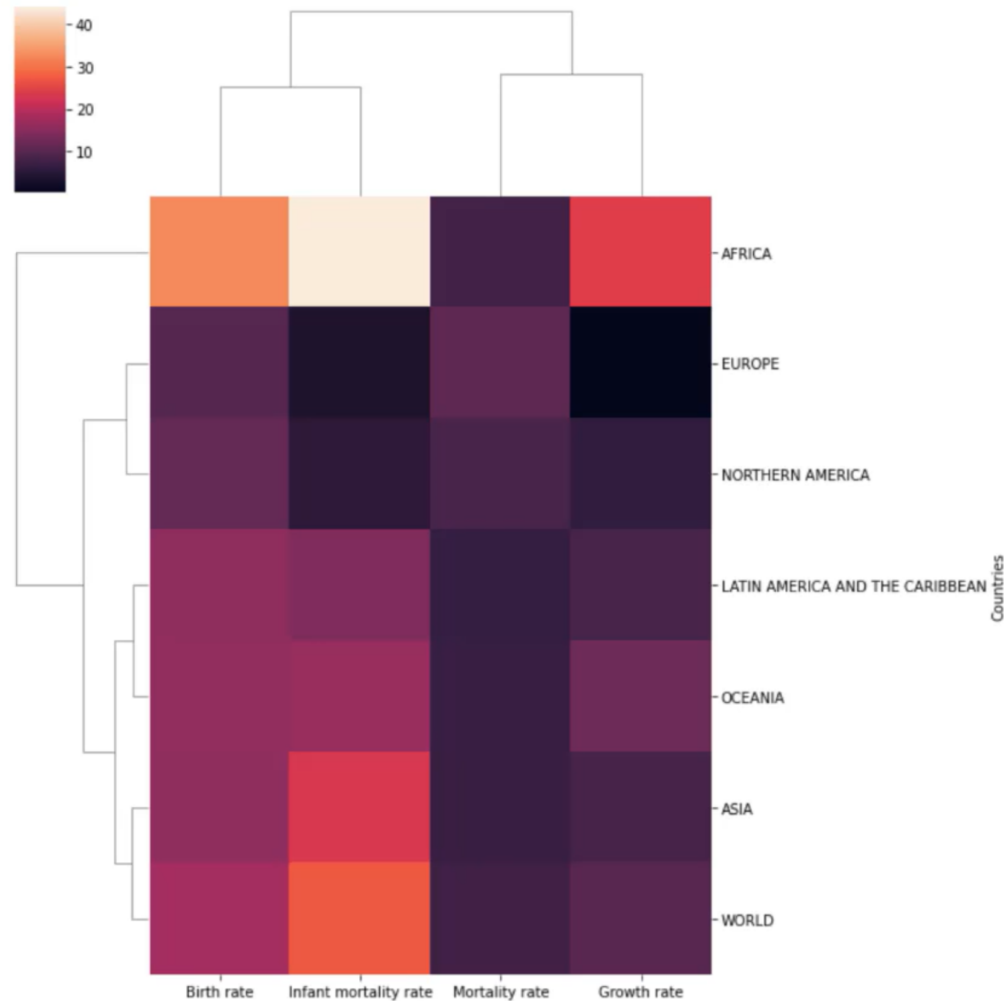
Matrix plots - матричные графики

- В seaborn есть возможность группировать в кластеры похожие значения
- Позже в разделе про машинное обучение мы обсудим, как выполняется разбиение данных на кластеры



Matrix plots - матричные графики

- Кластеризация





Matrix plots - матричные графики

- Давайте напишем код для этих графиков!



Упражнения по seaborn



Упражнения по seaborn

- Основная цель seaborn – применение более простого синтаксиса для более быстрого создания информативных графиков
- Поэтому сложно проверять навыки владения seaborn, поскольку большинство графиков требуют только передать данные, а также выбрать x и y



Упражнения по seaborn

- В этих упражнениях примеры графиков, которые Вам нужно будет воссоздать, приведены в виде картинок JPG
- Не волнуйтесь, если цвета и размеры у Вас получатся другими. Главное – исследовать данные, создав похожие графики.



Упражнения по seaborn

- Внимательно читайте описание графиков!
- Многие графики потребуют обработки и фильтрации данных в pandas перед построением графика в seaborn



Упражнения по seaborn

РЕШЕНИЯ