ФКН ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

НИС «Анализ данных в Python»

Меликян Алиса Валерьевна
кандидат наук, доцент ФКН НИУ ВШЭ

Яндекс

О дисциплине

- Цель дисциплины – изучение методов и технологий анализа данных и развитие навыков реализации аналитических проектов с помощью языка программирования Python.
- Занятия проходят в онлайн формате в 1 – 3 модулях.

Чему научитесь

- Осуществлять автоматизированный сбор данных.
- Выбирать подходящие методы исследования в зависимости от задачи и специфики имеющихся данных.
- Осуществлять предобработку и модификацию данных.
- Проводить анализа данных и интерпретировать полученные результаты.
- Представлять результаты анализа данных.

Содержание дисциплины

1. Введение в Python
2. Работа с данными
3. Описательный анализ данных
4. Графический анализ данных
5. Анализ взаимосвязей переменных
6. Регрессионный анализ данных
7. Кластерный анализ данных
8. Анализ временных рядов
9. Сбор данных в сети Интернет

Темы дисциплины

- Тема 1. Введение в Python.

Синтаксис языка. Типы данных и переменные. Структуры данных и их свойства. Работа с функциями.

- Тема 2. Работа с данными

Работа с таблицами. Обработка и проверка данных перед анализом. Внесение изменений в данные. Объединение таблиц. Группировка и агрегирование данных. Библиотека pandas.

- Тема 3. Описательный анализ данных

Частотный анализ данных. Описательная статистика. Подготовка данных для создания сводных таблиц. Нормальное распределение, Z-стандартизация, тест Колмогорова-Смирнова.

Темы дисциплины

- Тема 4. Графический анализ данных
- Тема 5. Анализ взаимосвязей переменных

Таблица сопряжённости. Формулировка гипотез. Этапы проверки гипотез. Уровень значимости и ошибка первого рода. Корреляционный анализ данных.

- Тема 6. Регрессионный анализ данных

Линейная регрессия. Бинарные модели. Оценка качества модели. Диагностика регрессионной модели.

Темы дисциплины

- Тема 7. Кластерный анализ данных

Иерархический кластерный анализ. Кластерный анализ методом к-средних. DBSCAN. Поиск оптимального кластерного решения. Содержательная характеристика кластеров.

- Тема 8. Анализ временных рядов

Модели ARIMA. Восстановление пропущенных значений. Прогнозирование значений на будущие периоды.

- Тема 9. Сбор данных в сети Интернет

Извлечение данных из веб-страниц. Библиотека beautifulsoup. API.

Оценивание

- Практические задания – 20%
- Исследовательский проект – 20%
- Контрольная работа 1 – 15%
- Контрольная работа 2 – 15%
- Письменный экзамен – 30%

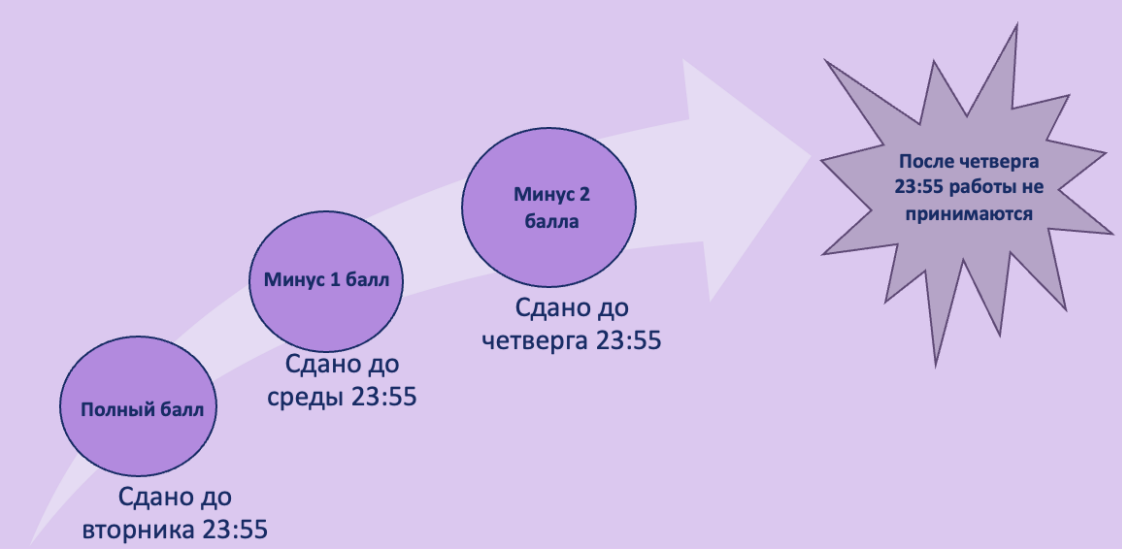
При условии выполнения более 50% заданий текущего контроля студент может быть освобождён от сдачи экзамена. В этом случае формула расчёта итоговой оценки выглядит следующим образом:

*Итоговая оценка = 0.3 * ИП+ 0.2 * КР1 + 0.2 * КР 2 + 0.3 * ПЗ*

Квизы и система гибких дедлайнов

- Почти на каждом занятии проводятся квизы по пройденным материалам. Победители квизов получают дополнительные баллы, которые добавляются к их оценкам за практические задания. Это необязательная форма контроля, пропуск квиза не влияет на оценки.
- Используется система гибких дедлайнов.

WAYGROUND
formerly Quizizz



Как получить 9 или 10 за практические задания?

- Выполнить не только задачи в рамках полученного задания, но и расширить анализ, применить дополнительные приемы обработки и анализа данных, которые позволят получить новые знания об изучаемых объектах или выявить интересные тенденции в рассматриваемых данных.
- Выступить с докладом на инициативную тему.

  Возможные темы на ближайшее занятие: интересные библиотеки и функции, необычные графики/интерактивные графики, создание и оформление профиля на GitHub.
- Стать победителем в квизе (проводятся на каждом занятии).

Инструменты

- Дистрибутив Anaconda содержит интерпретатор языка Python и несколько сред разработки. Мы будем использовать интерактивную среду разработки Jupyter Notebook, работающую с Python

  https://www.anaconda.com/

- Google Colab (интерактивная облачная среда для работы с кодом)

  https://colab.research.google.com/

Jupyter Notebook

Интерактивная веб-среда разработки, со следующим особенностями:

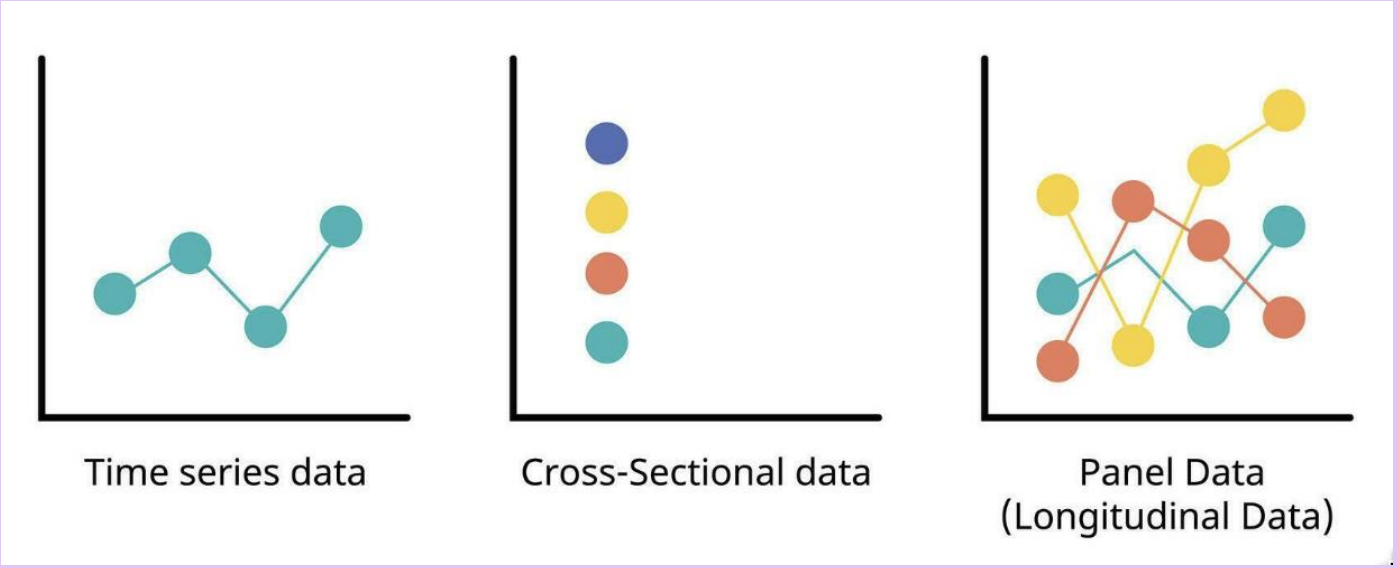
- можно сразу увидеть результат выполнения всего кода или отдельных его фрагментов, при этом код можно разбить на куски и выполнять их в произвольном порядке;
- предусмотрен вывод результата сразу после фрагмента кода;
- поля для ввода кода чередуются с полями, в которые можно добавлять текст, ссылки, изображения.

Полезные ресурсы

- Kaggle
- Google Dataset Search
- Harvard Dataverse
- Eurobarometer
- Росстат
- Портал открытых данных правительства Москвы

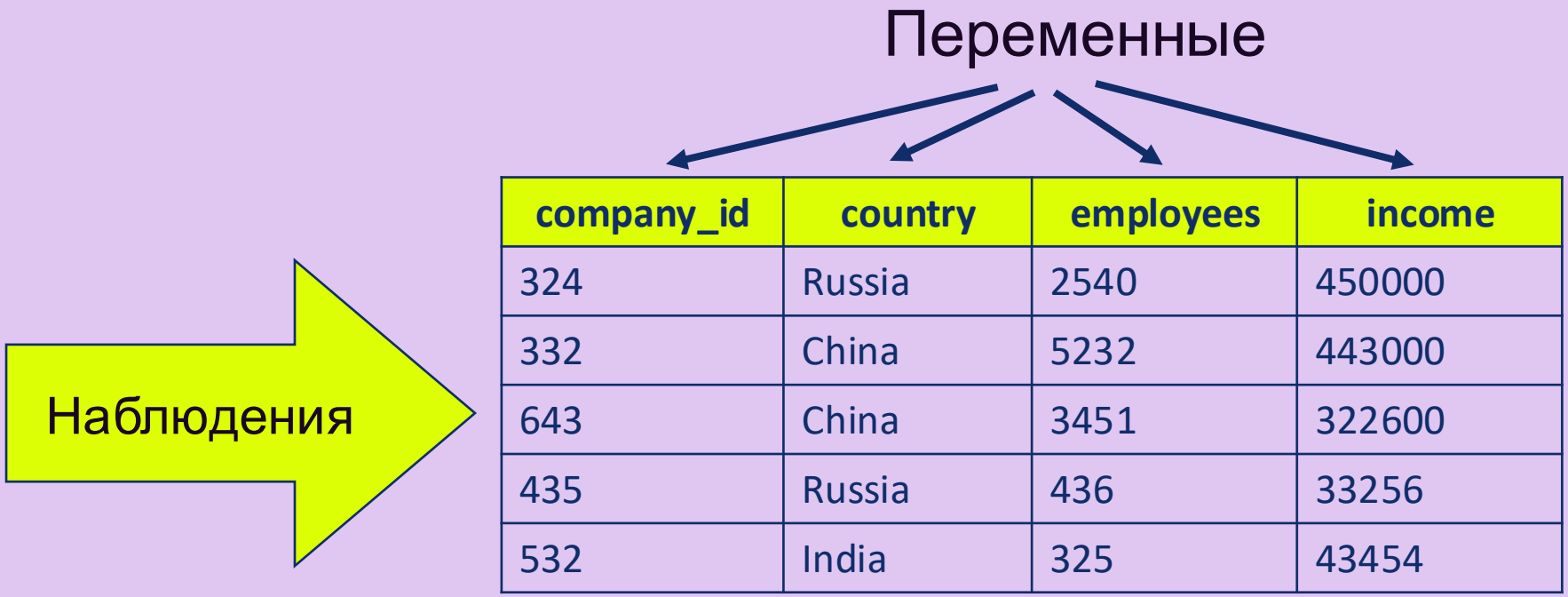
Типы данных и их структура

- Перекрёстные данные (cross-sectional data)
- Временные ряды (time series data)
- Панельные данные (panel data)



Перекрестные данные

Тип данных, собранный путем наблюдения за многими объектами в один и тот же период времени.

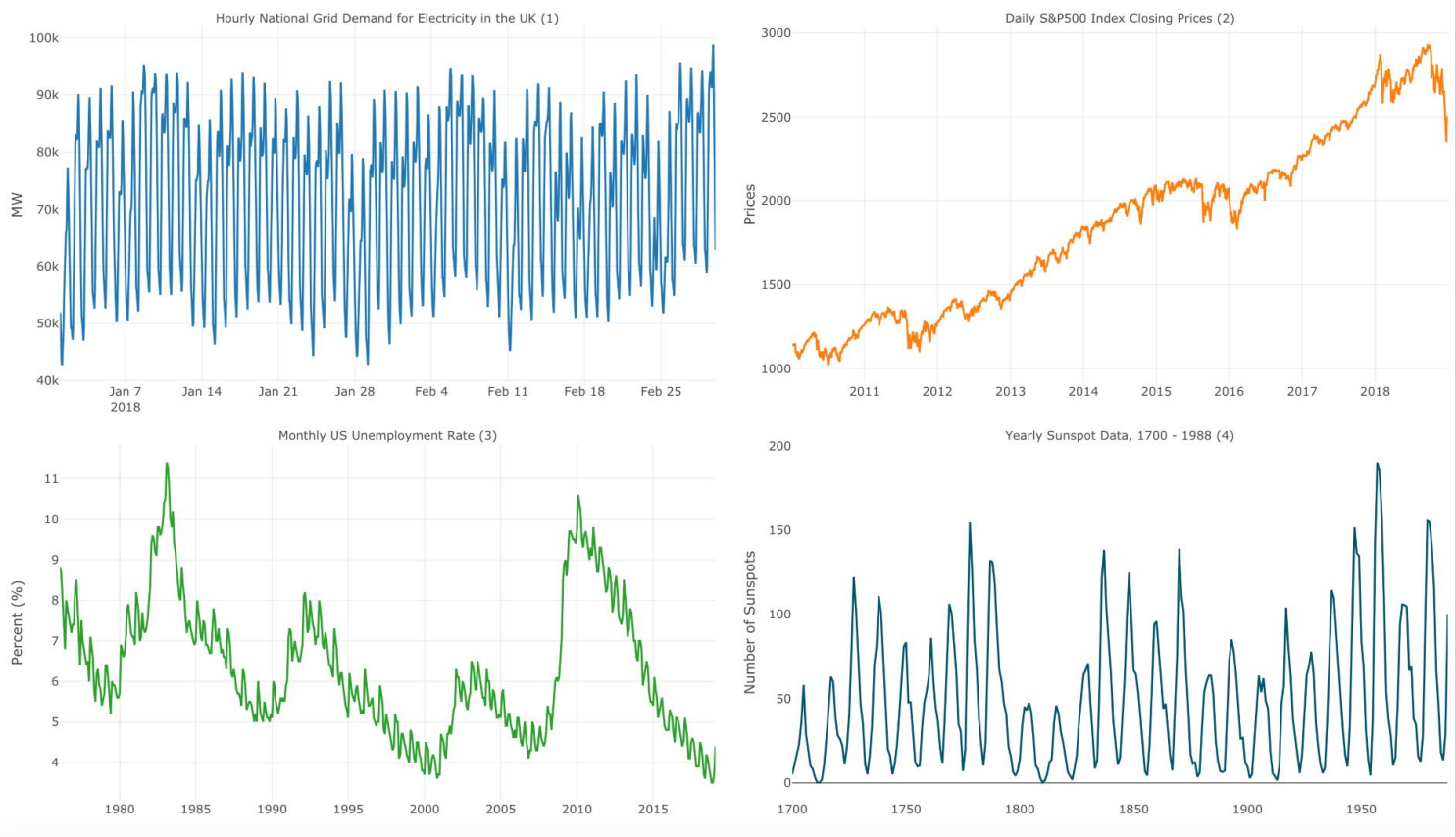


company_id	country	employees	income
324	Russia	2540	450000
332	China	5232	443000
643	China	3451	322600
435	Russia	436	33256
532	India	325	43454

Структура перекрестных данных

- Каждая строка представляет собой отдельное наблюдение. Это единица анализа – элементарная, единичная часть объекта исследования (респондент, организация, город, страна).
- Каждый столбец представляет отдельную переменную. Переменная – элементарный показатель, признак, характеризующий одно из изучаемых свойств единицы анализа (пол, возраст, заработная плата респондента).
- Ячейки содержат значения (числовые, текстовые, даты). Каждая ячейка содержит одно значение конкретной переменной для определённого наблюдения.

Временной ряд



Временной ряд (ряд динамики) – значения признака, измеренные через постоянные временные интервалы, например, ежедневные курсы валют, средняя дневная цена акции компании.

Панельные данные

- Каждая строка представляет собой отдельное наблюдение. Это единица анализа – элементарная, единичная часть объекта исследования (респондент, организация, город, страна).
- Каждый столбец представляет отдельную переменную. Переменная – элементарный показатель, признак, характеризующий одно из изучаемых свойств единицы анализа (пол, возраст, заработная плата респондента).
- Ячейки содержат значения (числовые, текстовые, даты). Каждая ячейка содержит одно значение конкретной переменной для определённого наблюдения.

Структура панельных данных

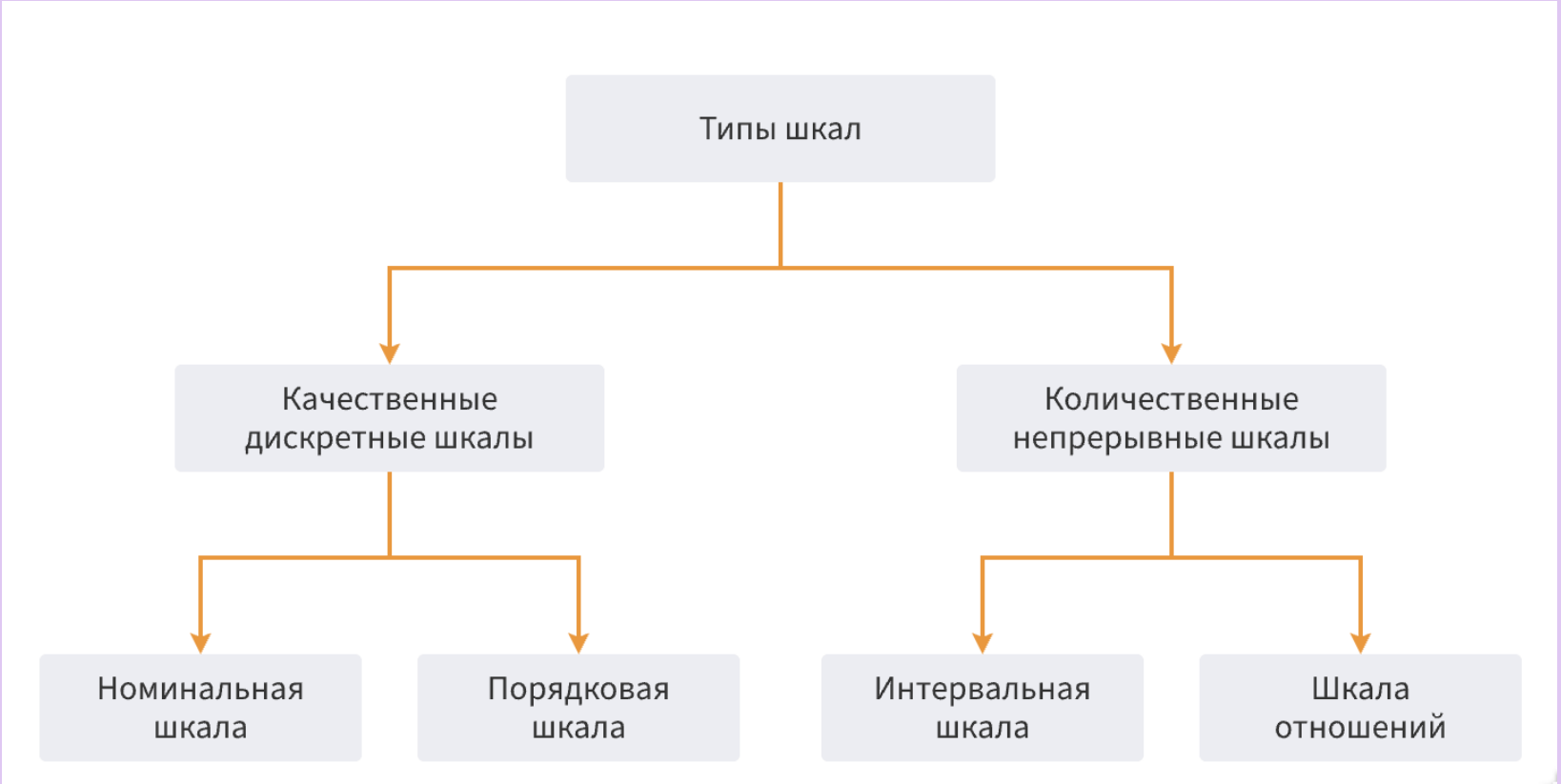
Панельные данные (Panel Data or Longitudinal Data) состоят из наблюдений одних и тех же единиц, которые осуществляются в последовательные периоды времени. Они насчитывают три измерения:

1) признаки (переменные),
2) объекты,
3) время.

Пример панельных данных

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4

Шкалы измерения переменных



Номинальная шкала

- Номинальные (категориальные) переменные используются для качественной классификации. То есть мы определяем принадлежность к определённому классу, отличающемуся от других но при этом классы не подлежат упорядочиванию (например, национальность, цвет, город).
- Дихотомические переменные – имеют два варианта ответа (например, пол, да/нет).

Порядковая шкала

Порядковые (ординальные) переменные позволяют ранжировать (упорядочить) объекты, указав какие из них в большей или меньшей степени обладают качеством, выраженным данной переменной. Но не позволяют сказать «на сколько больше» или «на сколько меньше» (например, социально-экономический статус семьи). Разница между уровнями низкий, средний и высокий существует, но её нельзя измерить.

Интервальная шкала

Интервальная шкала позволяет не только упорядочивать объекты измерения, но и численно выразить и сравнить различия между ними, например, температура по Цельсию. Так, температура воды в море утром – 18 градусов, вечером – 24, т.е. вечерняя на 5 градусов выше, но нельзя сказать, что она в 1.33 раз выше. 0 градусов – это не отсутствие температуры.

Абсолютная шкала или шкала отношений

В абсолютной шкале действует отношение "во столько-то раз больше". Это единственная из четырёх шкал имеющая абсолютный ноль, характеризующий отсутствие измеряемого качества. Например, ежемесячный доход. Здесь за точку отсчета можно взять «ноль» рублей.

Сравнение шкал

Свойства \ Тип шкалы	Номинальная	Порядковая	Интервальная	Отношений
Идентифицируемость	●	●	●	●
Величина (магнитуда)		●	●	●
Равенство интервалов			●	●
Абсолютный ноль				●

Дихотомный метод кодировки

Какими языками Вы владеете?
(можно выбрать любое число ответов)
1) русский
2) английский
3) французский
4) итальянский
5) немецкий

Ответы:
Марк: английский, немецкий
Анна: русский
Жан: английский, французский, итальянский, немецкий
Ольга: русский, немецкий

Name	Lan_en	Lan_ru	Lan_fr	Lan_it	Lan_ge
Марк	1	0	0	0	1
Анна	0	1	0	0	0
Жан	1	0	1	1	1
Ольга	0	1	0	0	1

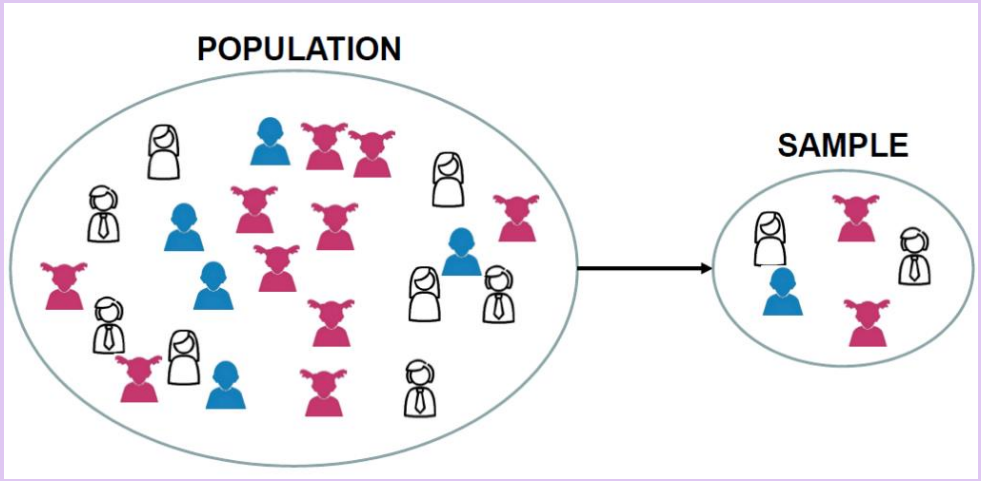
Категориальный метод кодировки

Какими языками Вы владеете?
(можно выбрать любое число ответов/
выберите не более 2 вариантов/
расположите языки в порядке
владения ими)
1) русский
2) английский
3) французский
4) итальянский
5) немецкий

Ответы:
Марк: английский, немецкий
Анна: русский
Жан: английский, французский, итальянский, немецкий
Ольга: русский, немецкий

Name	Lan_1	Lan_2	Lan_3	Lan_4
Марк	2	5		
Анна	1			
Жан	2	3	4	5
Ольга	1	5		

Выборка и генеральная совокупность



- Генеральная совокупность — это совокупность всех объектов, относительно которых исследователь намерен делать выводы при решении задачи.
- Объем генеральной совокупности может быть настолько велик, что на практике рассмотреть все ее элементы не представляется возможным. Поэтому обычно из генеральной совокупности извлекаются **выборки**, на основе анализа которых исследователь делает вывод о свойствах всей совокупности, скрытых в ней закономерностях, действующих правилах и т.д. При этом выборки должны быть репрезентативными.

Репрезентативность выборки

Под репрезентативностью выборки понимается соответствие её структурных характеристик характеристикам генеральной совокупности, из которой она сформирована. Репрезентативность определяет, насколько возможно обобщать результаты исследования с использованием выборки на всю исходную совокупность. Также репрезентативность можно определить как свойство выборочной совокупности представлять параметры генеральной совокупности, значимые с точки зрения задач исследования.

Подходы к формированию выборки

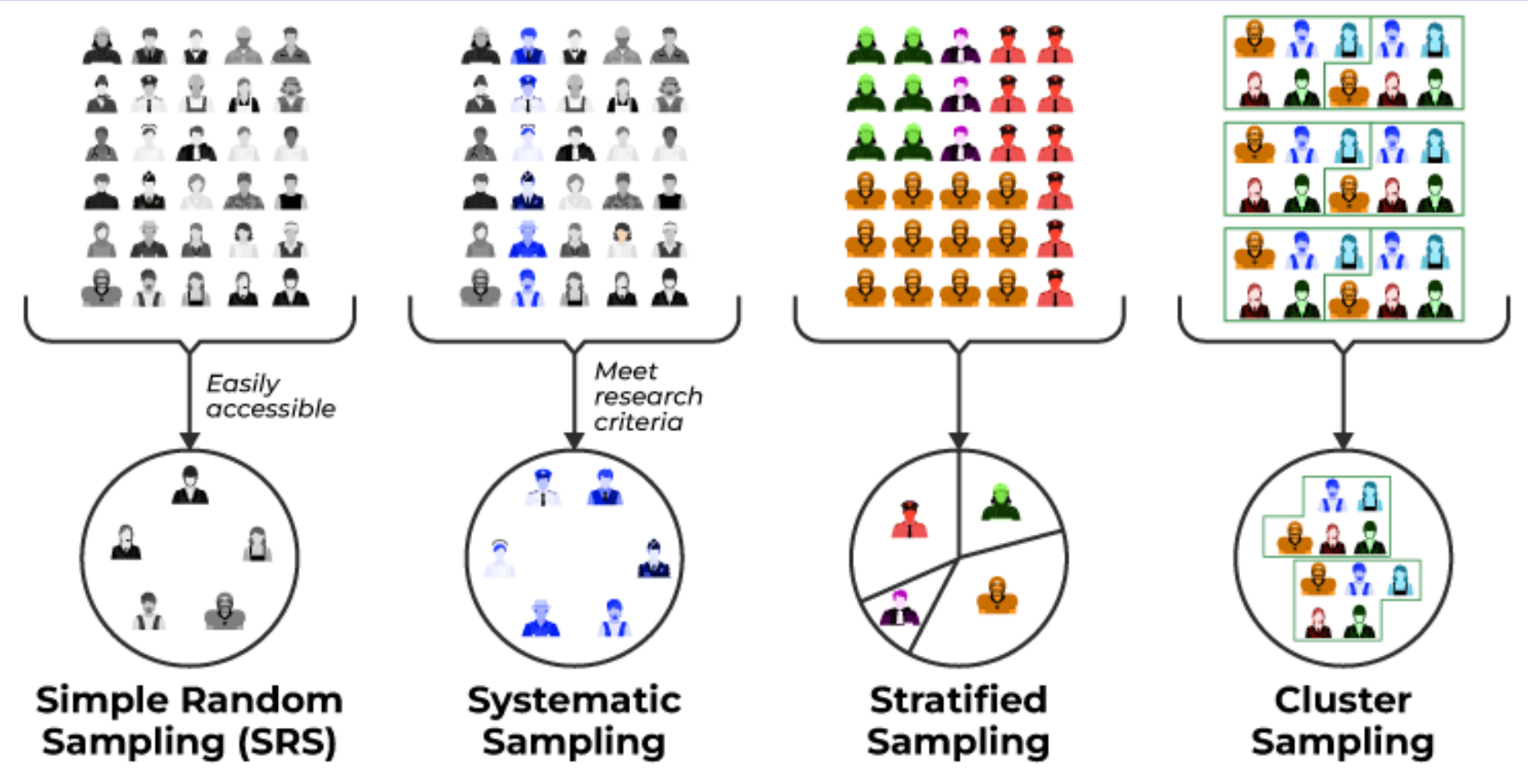
1. Простая случайная выборка (Simple Random Sampling)
Каждый элемент имеет равную вероятность быть выбранным, при этом выбор одного элемента не влияет на выбор других.

2. Систематическая выборка (Systematic Sampling)
Предполагает отбор каждого n-го элемента из генеральной совокупности. Например, если у вас есть список студентов и вы выбираете каждого 10-го студента.

3. Стратифицированная выборка (Stratified Sampling)
Генеральная совокупность делится на подгруппы (страты) по определенным характеристикам (например, возраст, пол), а затем из каждой страты случайным образом отбираются элементы. Этот метод обеспечивает репрезентативность каждой подгруппы.

4. Кластерная выборка (Cluster Sampling)
Генеральная совокупность делится на кластеры или группы, часто по географическому принципу. Случайным образом отбираются кластеры, и затем все элементы внутри выбранных кластеров включаются в выборку.

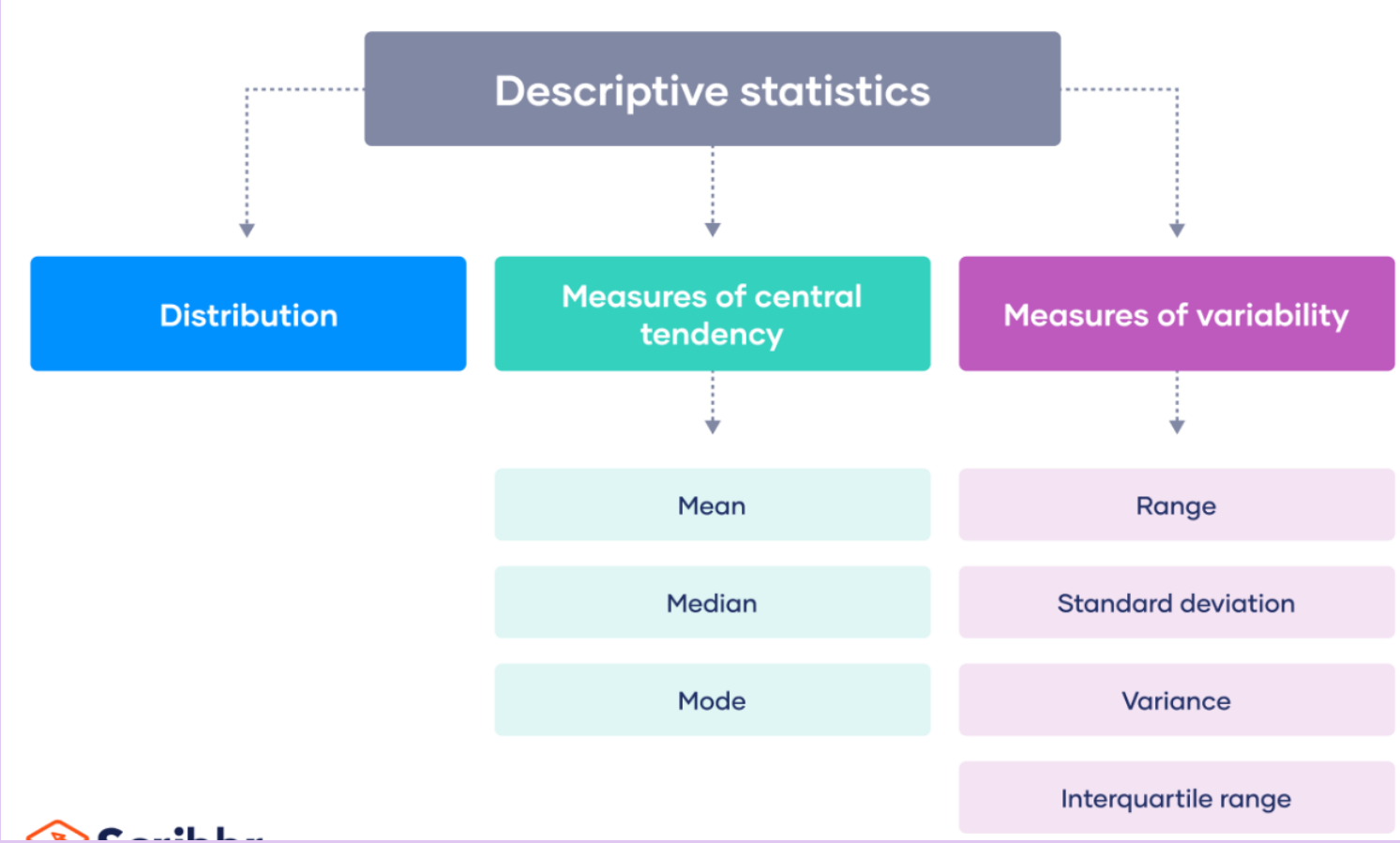
Подходы к формированию выборки



Описательная статистика

Описательная или дескриптивная статистика (descriptive statistics) решает задачу сжатия исходной информации, компактного её представления для дальнейшего осмысления. Данные наглядно представляются в форме графиков или таблиц, распределение их значений описывается посредством ряда статистических показателей.

Описательная статистика

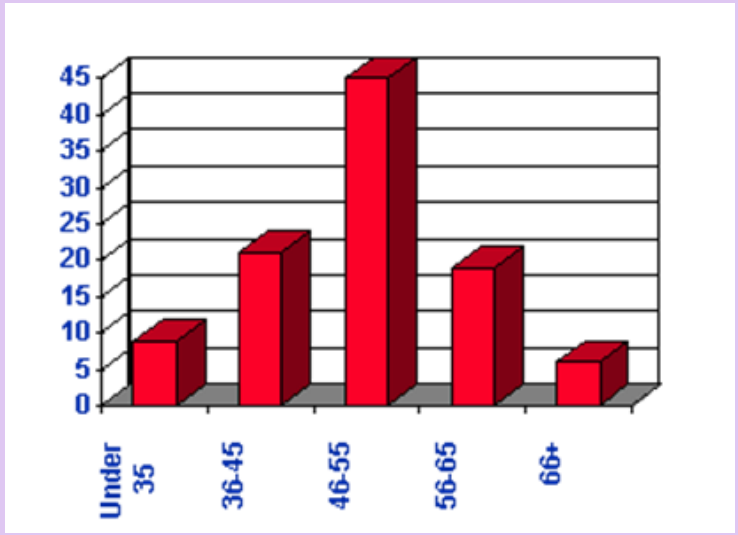


Частотный анализ

Анализ частотных распределений позволяет получить общее представление об изучаемой выборке. Чаще применяется при анализе значений категориальных переменных.

Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

Частотная таблица



Столбиковая диаграмма

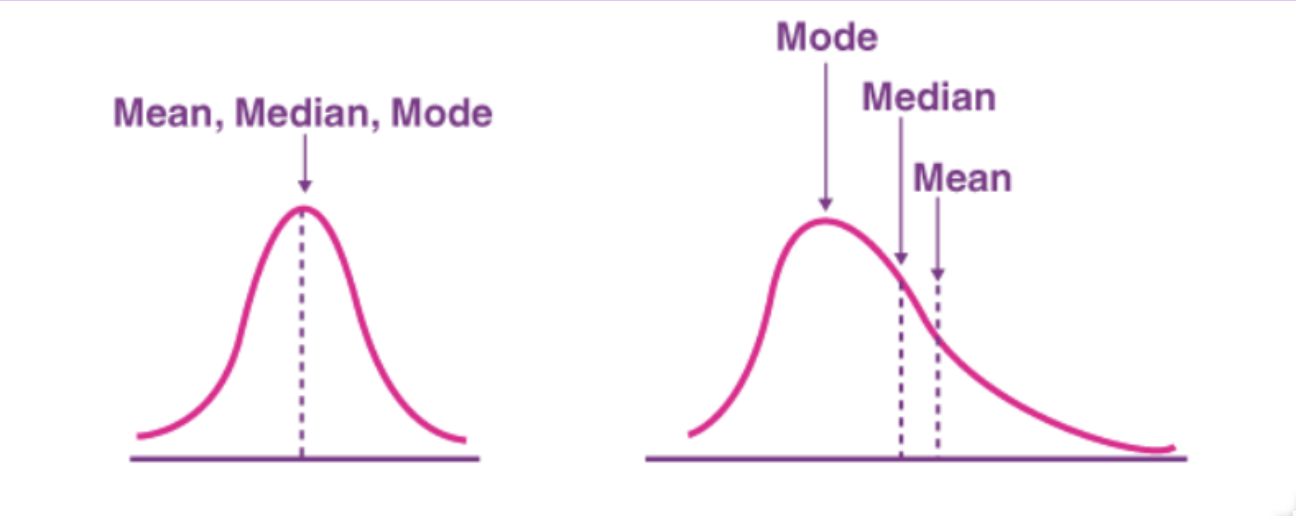
Меры центральной тенденции

Позволяют охарактеризовать множество значений признака, измеренного на выборке, одним числом. Показывают концентрацию группы значений на числовой шкале.

Шкала измерения переменной	**Допустимые меры центральной тенденции**
номинальная	мода
порядковая	мода, медиана
метрическая	мода, медиана, среднее арифметическое

Мода

Мода — это значение, которое наиболее часто встречается в выборке. Если одна и та же наибольшая частота встречается у нескольких значений, то выбирается наименьшее из них.



Среднее арифметическое

Рассчитывается как сумма элементов ряда делённая на их количество. Информация о том, что среднее арифметическое оценок студентов 2-го курса по результатам сдачи дисциплины «Анализ данных» составляет 4,2 с учетом того, что оценки варьируются от 2 до 5, позволяет сделать вывод о том, что в среднем студенты 3-го курса сдали этот экзамен хорошо.

$$\bar{X} = \frac{\sum x_i}{n}$$

Медиана (2-ой квартиль/50-й перцентиль)

Медиана — это точка на шкале измеренных значений, выше и ниже которой лежит по половине всех измеренных значений. Например, если измеренные значения таковы:

3 7 8 5 4 6 3 9 2 8 4

то сначала их нужно расположить в порядке возрастания:

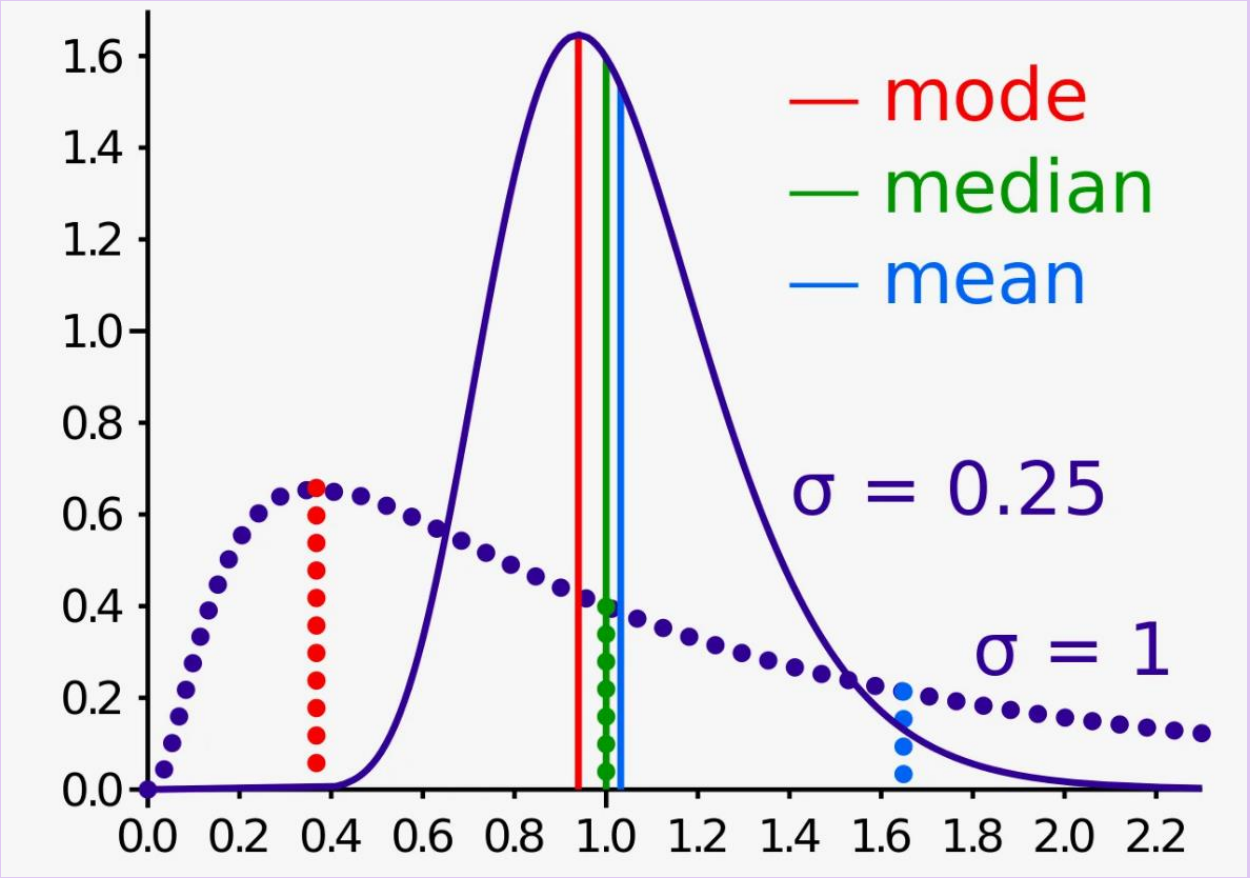
2 3 3 4 4 **5** 6 7 8 8 9

В данном случае медианой будет значение 5. Всего у нас 11 измеренных значений, следовательно, медианой является шестое значение. Выше него располагается 5 значений, и ниже — тоже 5. При нечетном количестве значений медиана всегда будет совпадать с одним из измеренных значений. При четном количестве медиана будет средним арифметическим двух соседних значений. Например, если имеются следующие измеренные значения:

3 4 4 5 **6 7** 8 8 9 9

то медиана будет равна: (6 + 7) / 2 = 6,5.

Меры центральной тенденции



Меры разброса

Статистические показатели, которые описывают, насколько данные отклоняются от центра распределения. Они помогают оценить вариативность и однородность данных.

Мера центральной тенденции	Подходящие меры разброса	Область применения
Среднее арифметическое	Стандартное отклонение, Дисперсия	Нормальное распределение, симметричные данные без выбросов
Медиана	Межквартильный размах (IQR), Межквартильная широта, Децильное отношение Размах (max – min)	Асимметричные распределения, данные с выбросами, порядковые шкалы
Мода		Категориальные данные, номинальные шкалы

Дисперсия

Дисперсия может использоваться в качестве показателя точности модели. Особенность дисперсии в том, что она отражает значения, возведенные в квадрат. Поэтому чаше в качестве показателя, отражающего точность модели, используется стнадратное отклонение, представляющее собой корень квадратный из дисперсии.

$$\sigma^2 = \frac{\sum(x - \bar{X})^2}{n}$$

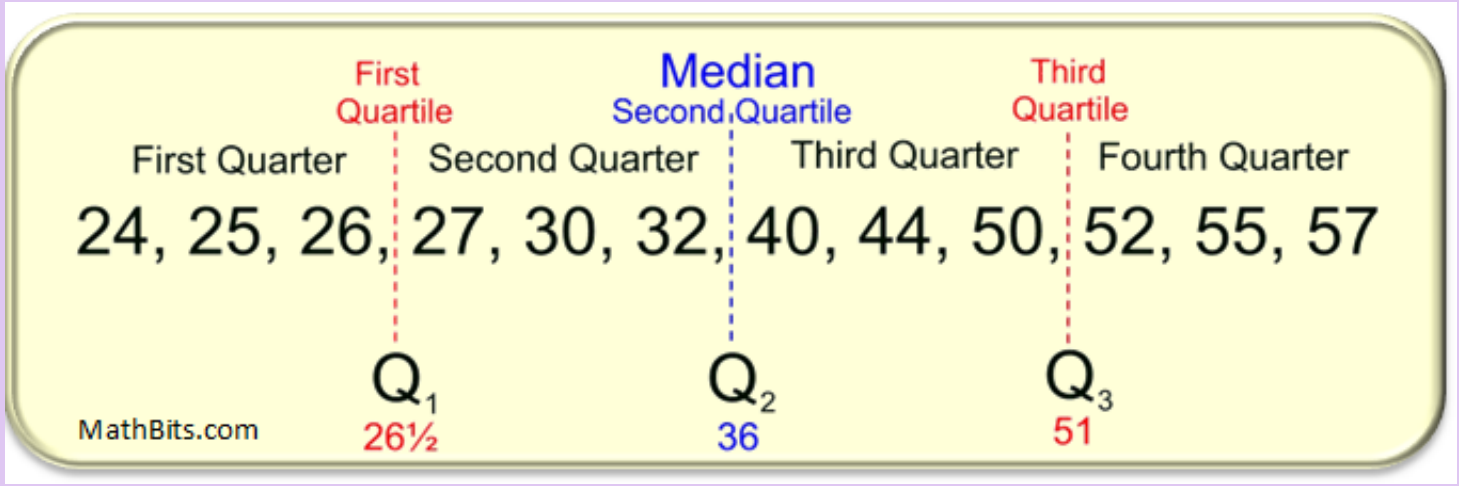
Стандартное отклонение

Дисперсия может использоваться в качестве показателя точности модели. Особенность дисперсии в том, что она отражает значения, возведенные в квадрат. Поэтому чаше в качестве показателя, отражающего точность модели, используется стандартное отклонение, представляющее собой корень квадратный из дисперсии.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x - \bar{X})^2}{n}}$$

Квартили

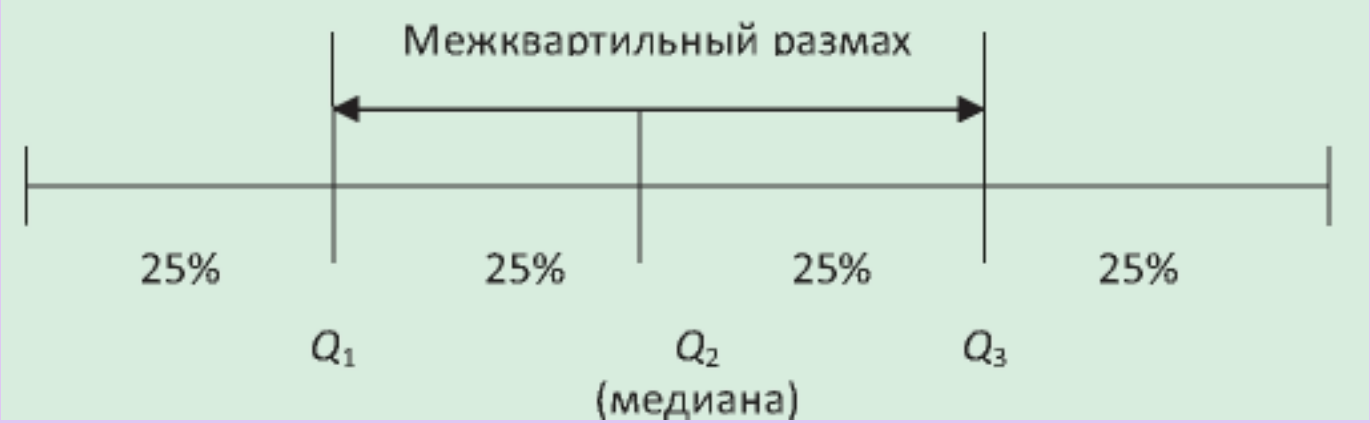
- **Первый квартиль** (25-й перцентиль) – точка на шкале значений переменной, ниже значения которой находятся 25% значений переменной.
- **Второй квартиль** (медиана) – точка на шкале значений переменной, ниже значения которой находятся 50% значений переменной.
- **Третий квартиль** (75-й перцентиль) – точка на шкале значений переменной, ниже значения которой находятся 75% значений переменной.



Межквартильный размах

Дисперсия может использоваться в качестве показателя точности модели. Особенность дисперсии в том, что она отражает значения, возведенные в квадрат. Поэтому чаше в качестве показателя, отражающего точность модели, используется стандартное отклонение, представляющее собой корень квадратный из дисперсии.

$$IRQ = Q_3 - Q_1$$



Межквартильная широта

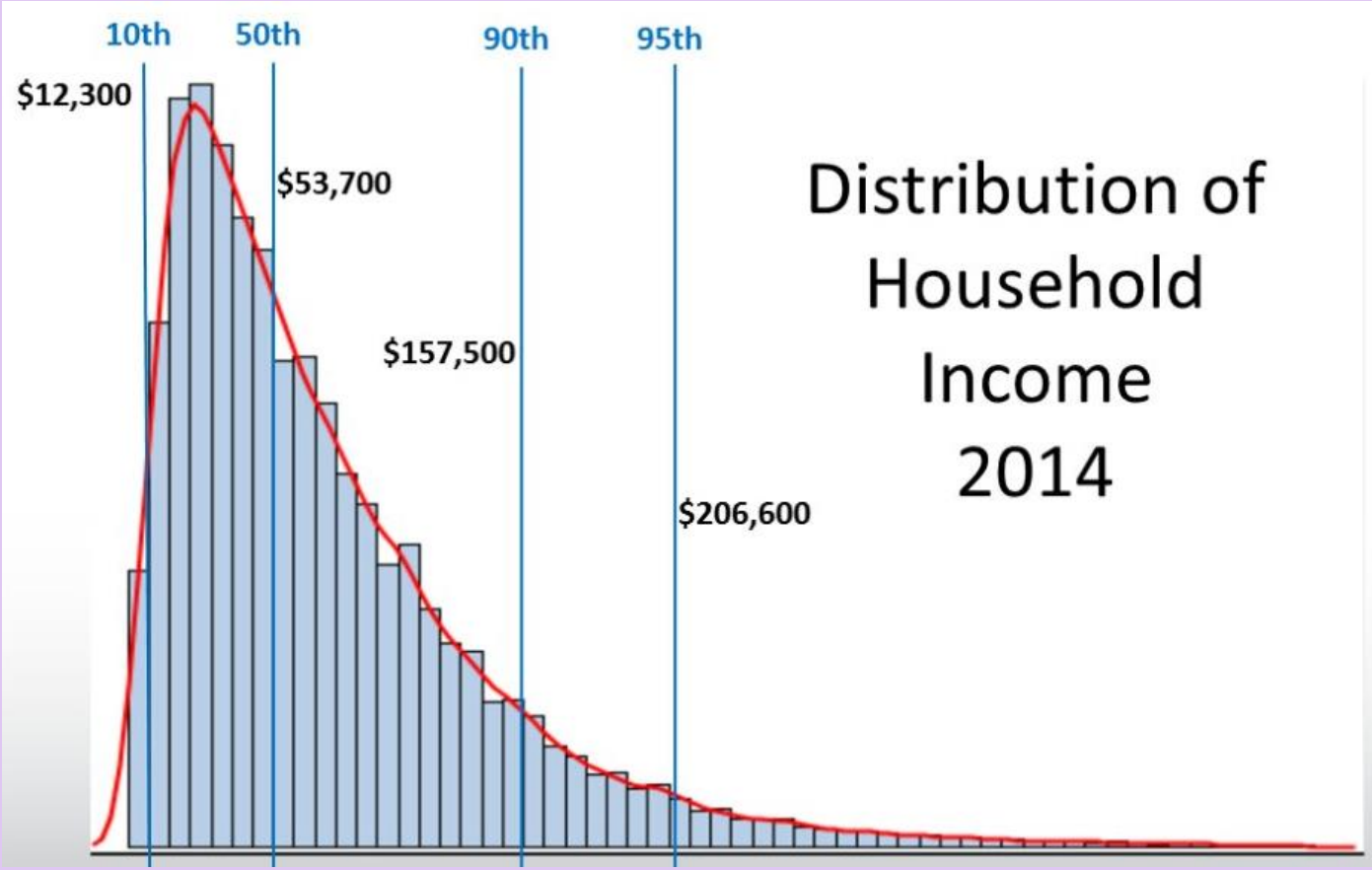
Для оценки меры разброса порядковых переменных может подсчитываться межквартильная широта (квартильное отклонение).

$$\text{Межквартильная широта} = \frac{(\text{Третий квартиль} - \text{Первый квартиль})}{2}$$

Если порядковая переменная имеет 7 значений и квартильное отклонение равно 3, то можно сделать вывод, что мера центральной тенденции (медиана) не точно хорошо характеризует распределение значений переменной, т.к. много респондентов имеют значения, отличающиеся от медианы.

Децильное отношение

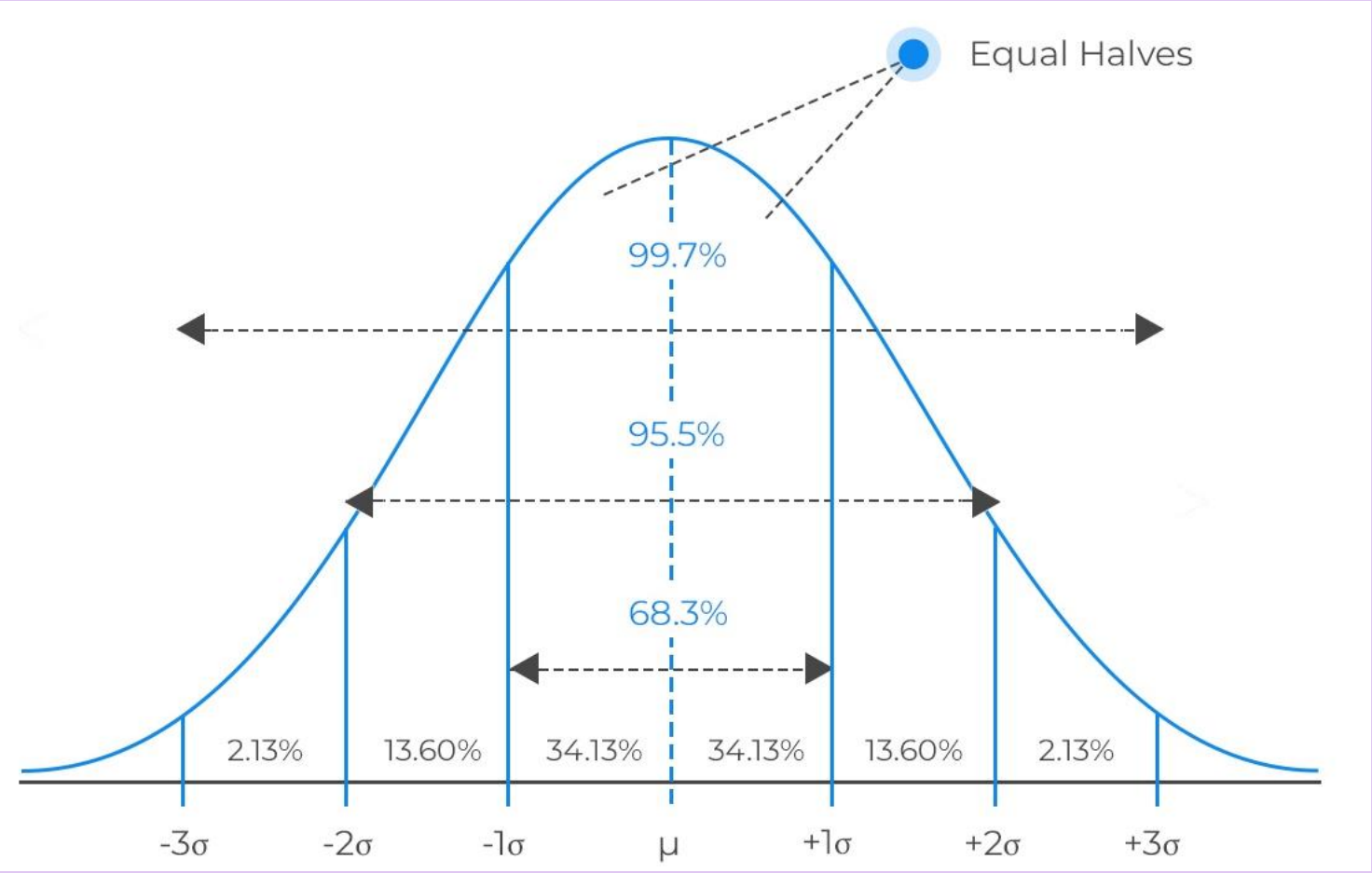
Децильное отношение – это отношение границы 10-го дециля к границе 1-го дециля. Данный показатель может, например, демонстрировать насколько больше получают 10 % высокооплачиваемых респондентов в сравнении с 10 % наименее оплачиваемых, что позволит оценить степень неоднородности доходов.



Нормальное распределение

Определенные методы анализа могут применяться если значения метрических переменных подчиняются нормальному распределению. При таком распределении большая часть значений группируется около среднего значения, по обе стороны от которого частота наблюдений равномерно снижается. Если провести вертикальную линию через центр гистограммы, то график будет выглядеть идентично по обе стороны от вертикальной линии. Линия, соединяющая вершины составляющих гистограмму столбиков, будет иметь вид симметричного купола. На практике не встречаются выборки, строго подчиняющиеся нормальному распределению. Однако есть распределения, которые статистически значимо не отличаются от нормального распределения, Перед тем, как применять методы анализа необходимо выяснить можно ли распределение считать нормальным и насколько сильно оно отличается от нормального.

Нормальное распределение



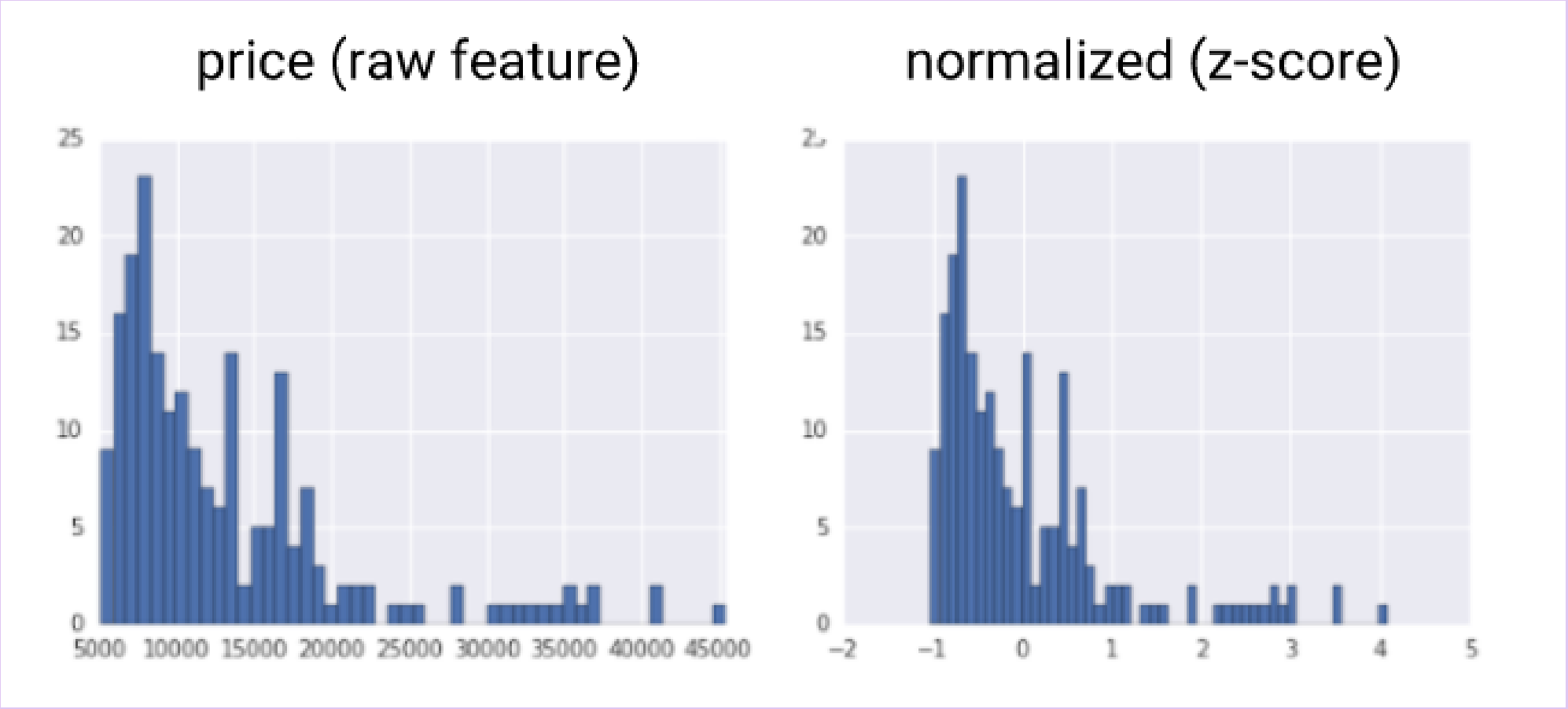
Z-стандартизация (нормирование)

В нормальном распределении среднее арифметическое равно 0, а стандартное отклонение – 1. Любые данные могут быть преобразованы таким образом, чтобы среднее арифметическое было равно 0, а стандартное отклонение – 1. Процедура называется стандартизацией данных. Для этого мы из каждого значения переменной вычитаем среднее значение и делим эту разницу на стандартное отклонение. Получаемые значения называются z-значения.

$$z_i = \frac{x_i - \overline{X}}{\sigma}$$

Для наблюдений, имеющих значение переменной выше её среднего значения z-значение положительно, ниже среднего – отрицательно, равно среднему – ноль.

Z-стандартизация (нормирование)



Оценка отличия распределения от нормального

Частотное распределение может отличаться от нормального в двух основных направлениях:

1) Симметричность распределения (Skewness);

2) Заострённость распределения (Kurtosis).

В случае нормального распределения значение показателя симметричности равно 0, а показателя эксцесса равно 3.

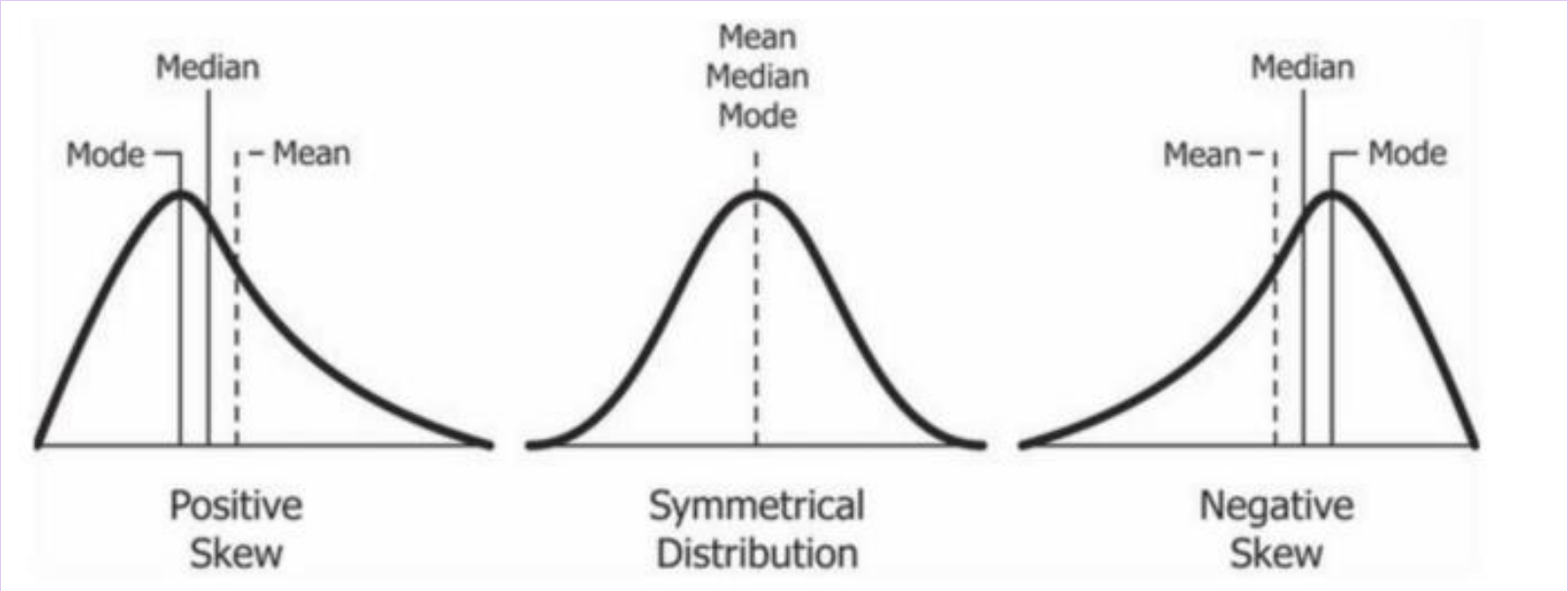
Симметричность распределения

В случае асимметричного распределения (skewed distribution) наиболее часто встречающиеся значения (самые высокие столбики гистограммы) расположены большей частью на одной из сторон шкалы. Т.е. на одной стороне шкалы сосредоточены самые высокие столбики, а при приближении к другой стороне шкалы высота столбиков снижается. Несимметричные распределения могут иметь как положительную асимметрию (перекос в сторону меньших значений), так и отрицательную (перекос в сторону больших значений).

Симметричность распределения

$$skewness = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{(N-1)s^3}$$

При положительной асимметрии в распределении чаще встречаются более низкие значения признака.



При отрицательной асимметрии в распределении чаще встречаются более высокие значения признака.

Заостренность распределения

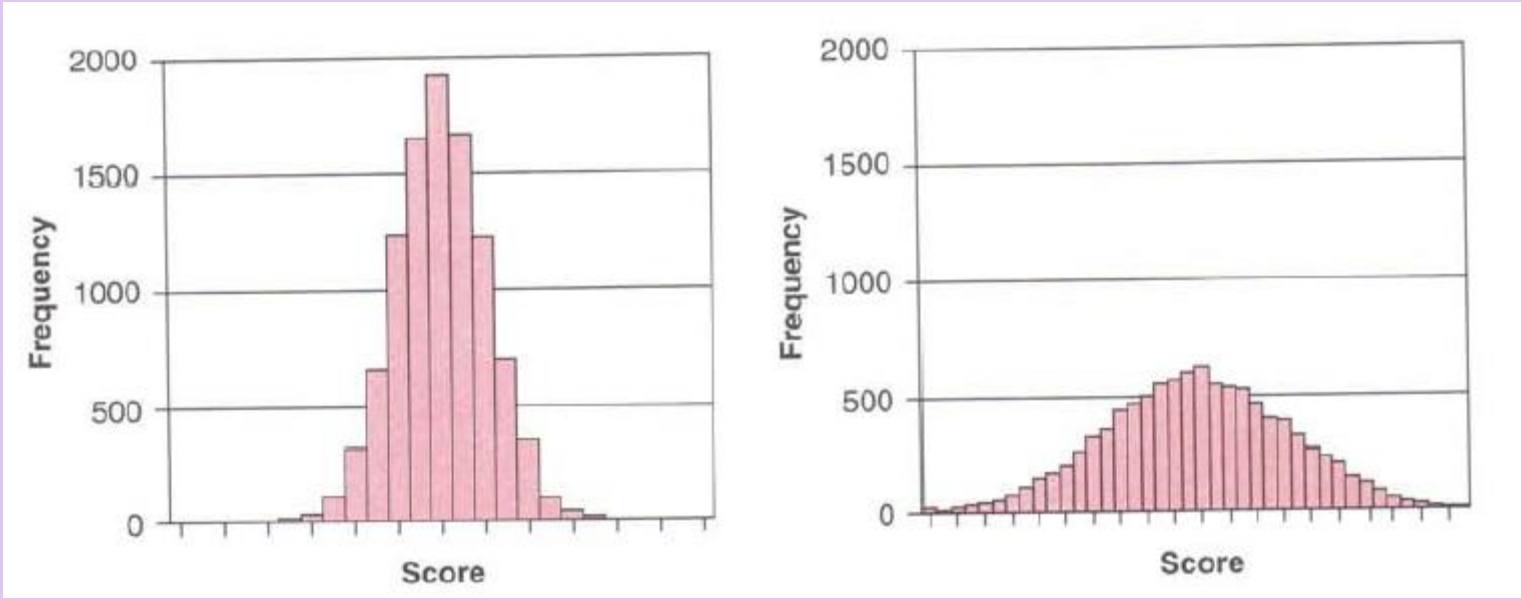
Распределения также могут различаться по степени заострённости/пологости (эксцесс). Показывает степень кластеризации распределения. Пологое распределение имеет «длинные хвосты».

Островершинное распределение (распределение с эксцессом выше нормального) предполагает значение показателя эксцесса, превышающее 3.

Плосковершинное распределение (распределение с эксцессом ниже нормального) предполагает значение показателя эксцесса меньше 3.

Заостренность распределения

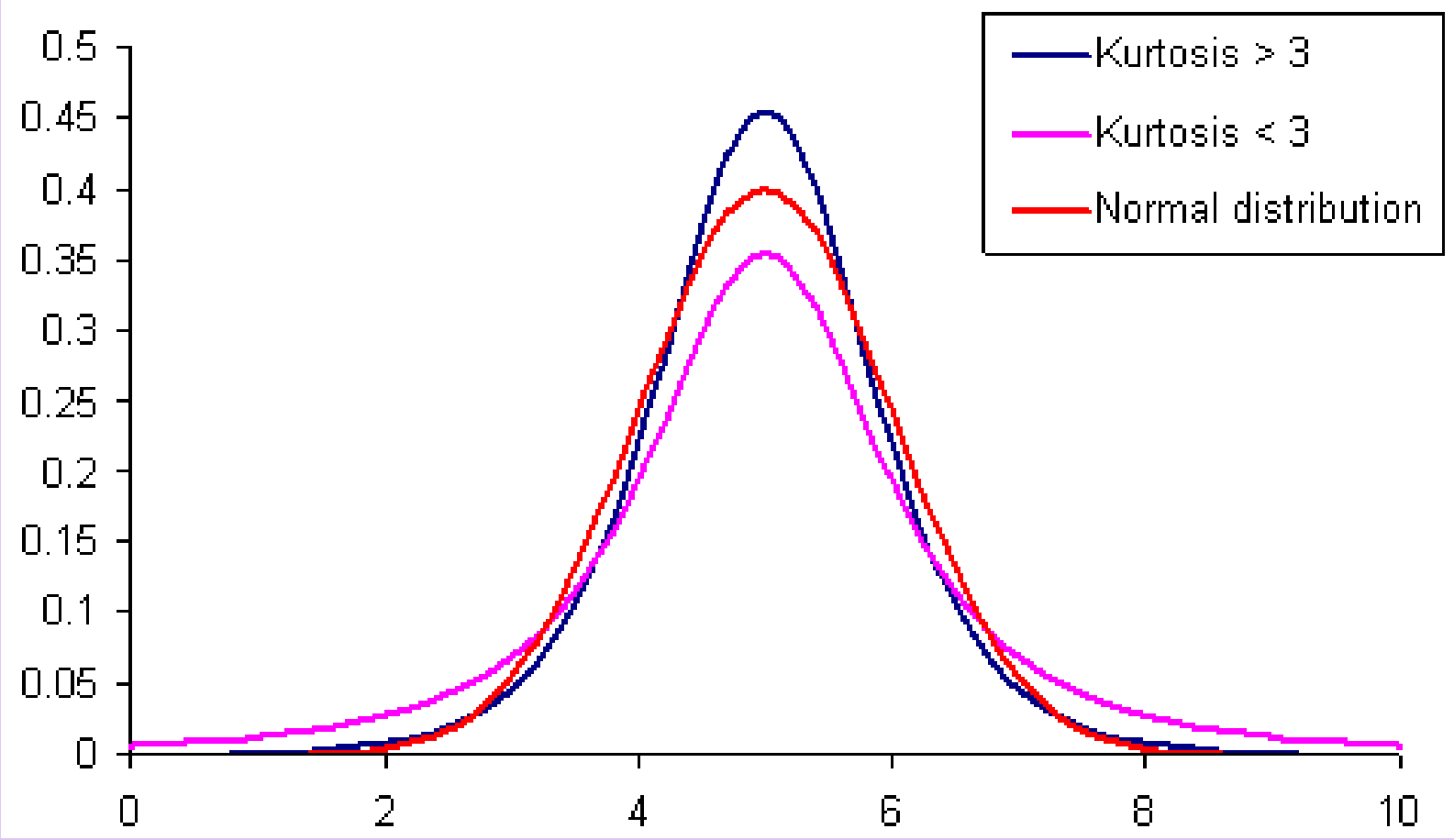
$$kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$



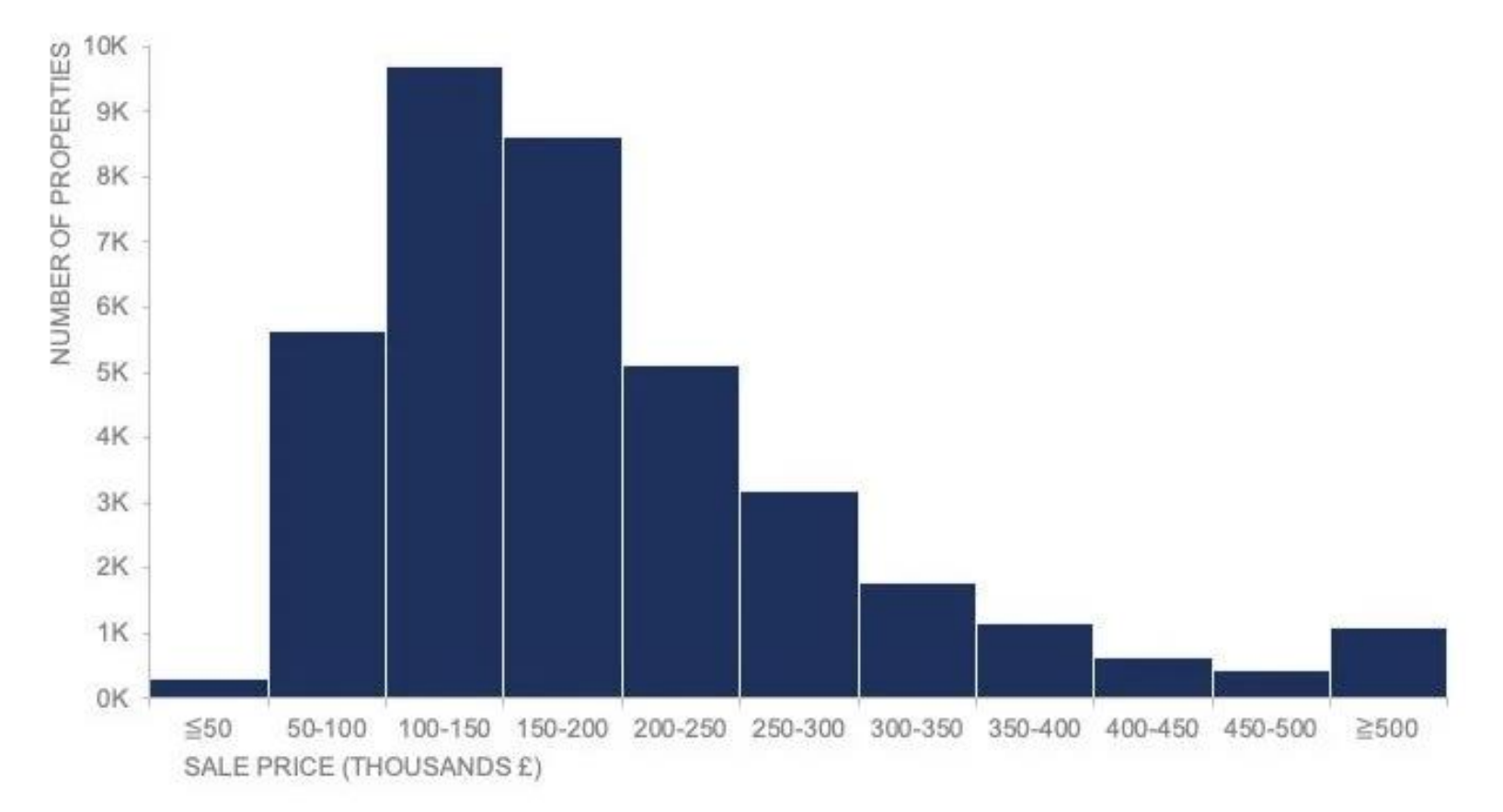
Пиковое распределение
kurtosis > 3

Пологое распределение
kurtosis <3

Заостренность распределения

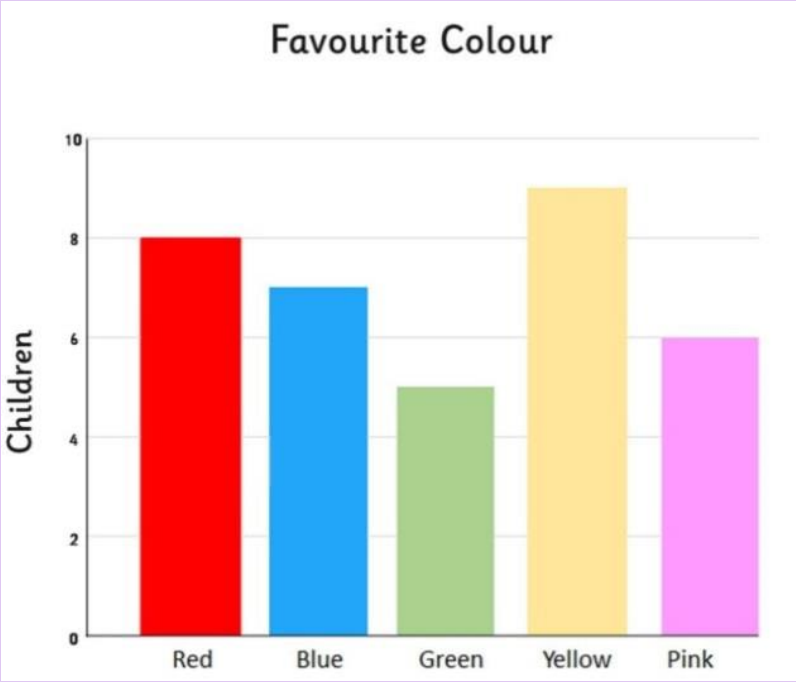


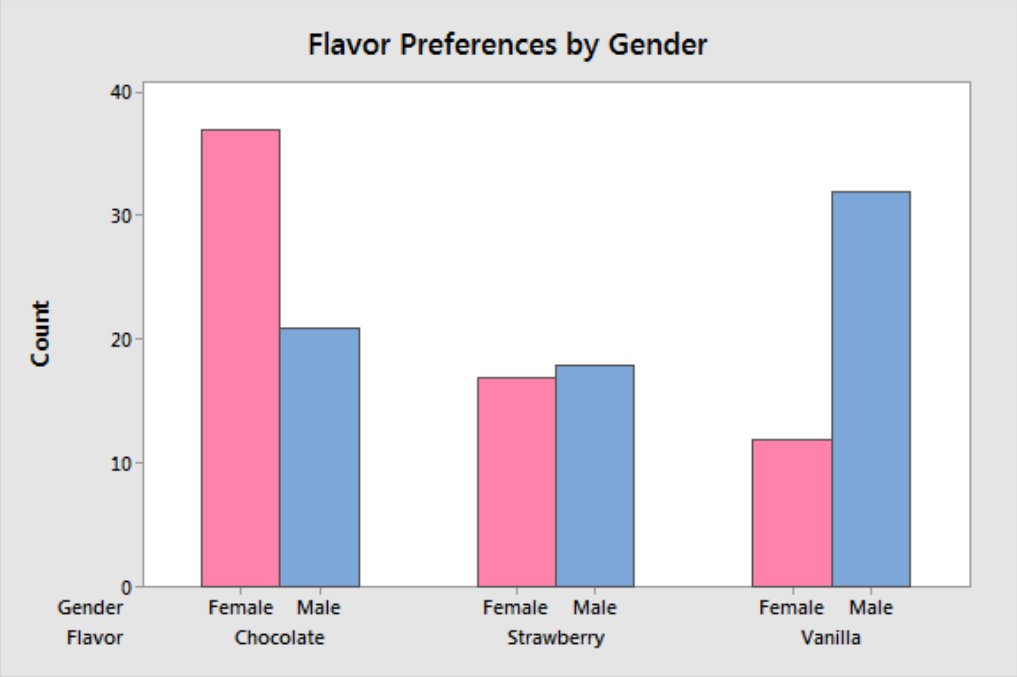
Гистограмма



Подходит для визуализации распределения значений метрических переменных

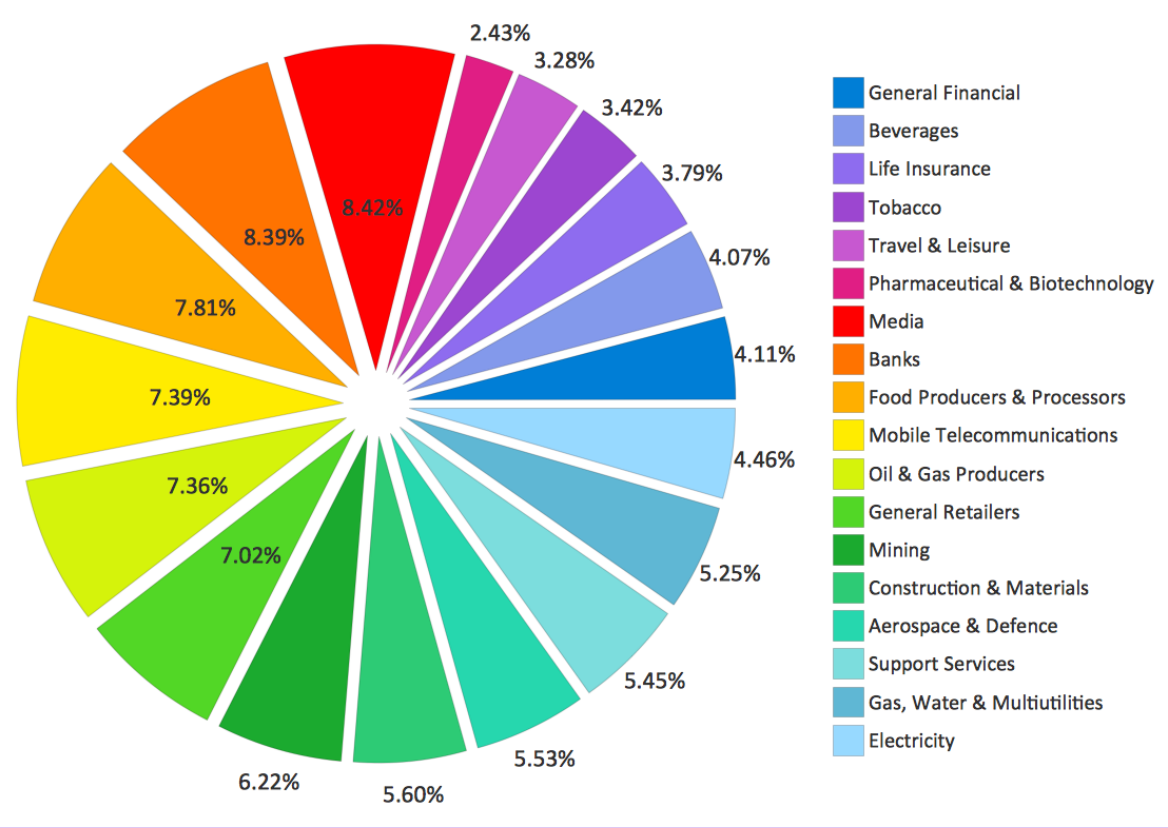
Столбиковая диаграмма





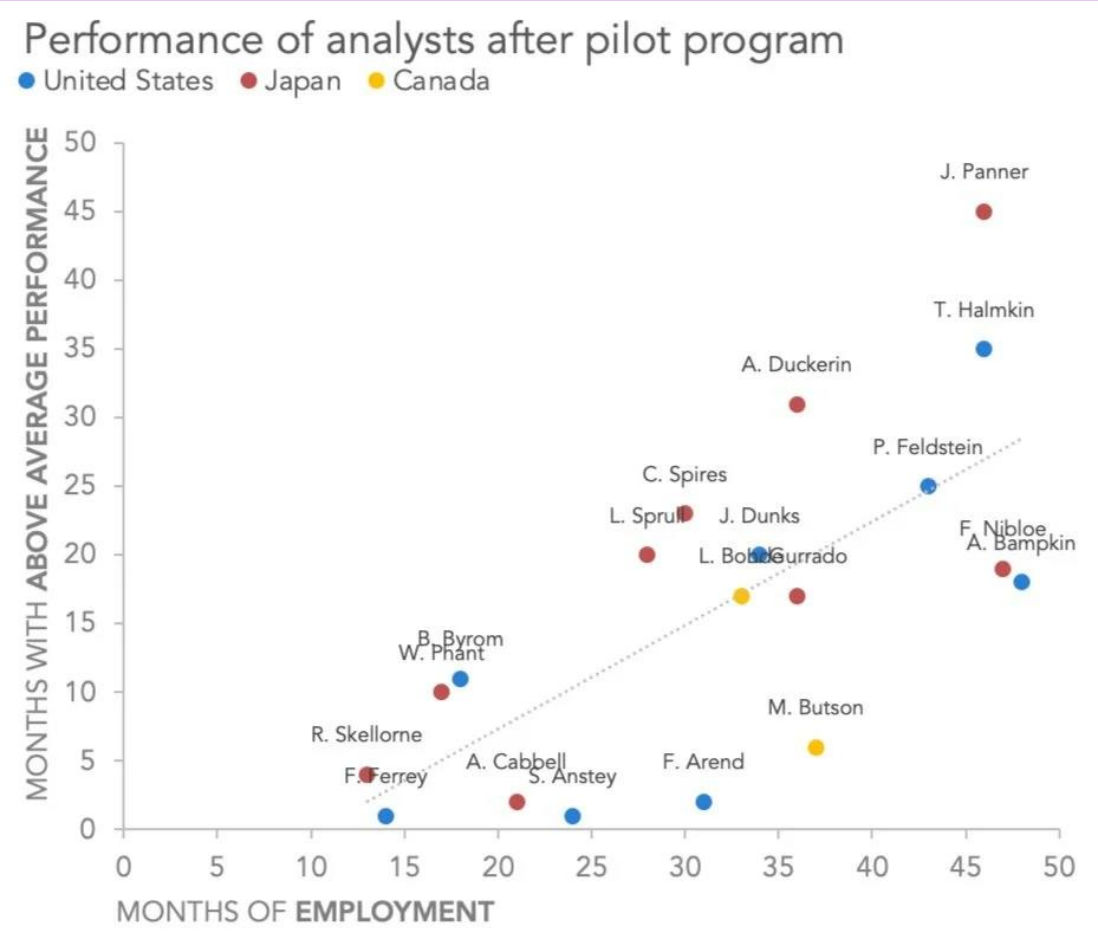
Подходит для визуализации распределения значений категориальных переменных

Круговая диаграмма



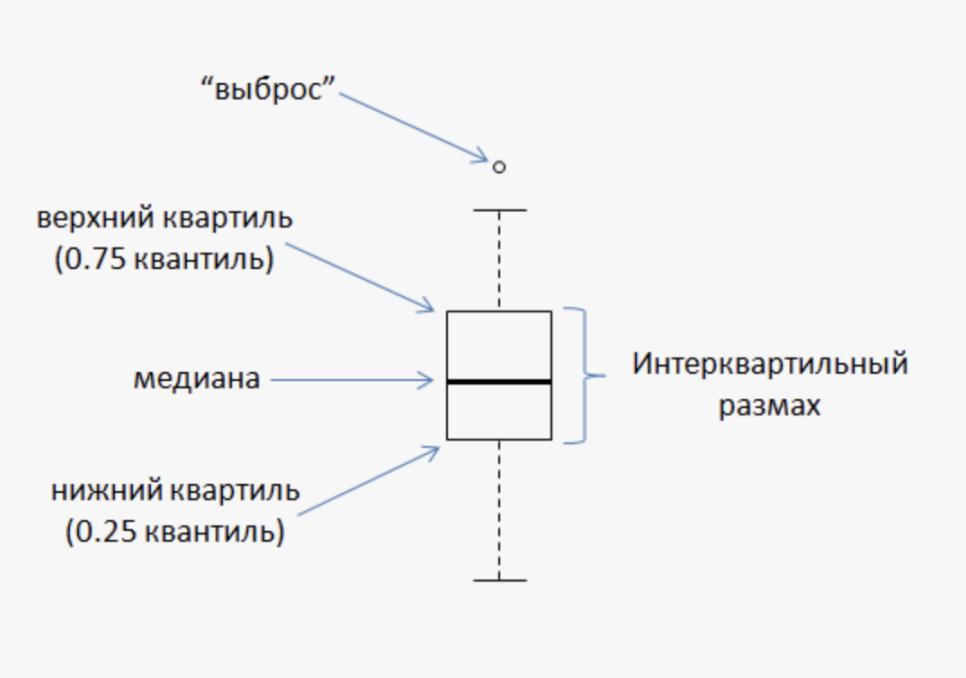
Подходит для визуализации распределения значений категориальных переменных

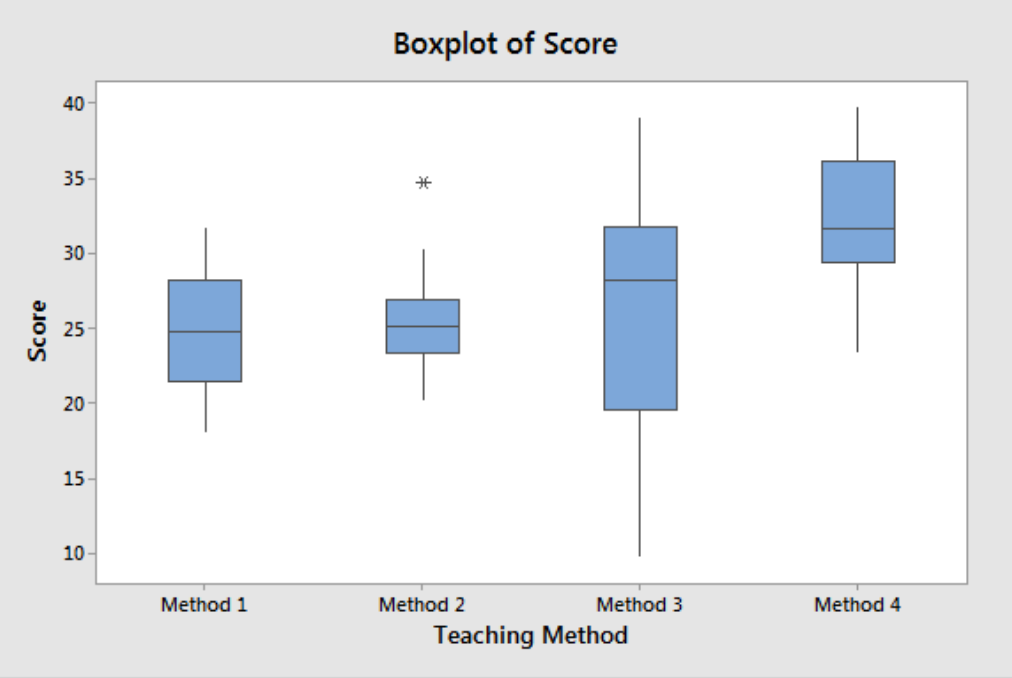
Диаграмма рассеяния



Подходит для визуализации взаимосвязи между метрическими переменными

Ящичковая диаграмма





Подходит для визуализации распределения значений метрических и порядковых переменных

Графики в Python

- https://python-graph-gallery.com/
- https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/