



Факультет компьютерных наук

Машинное обучение

Москва 2026

Лекция 2

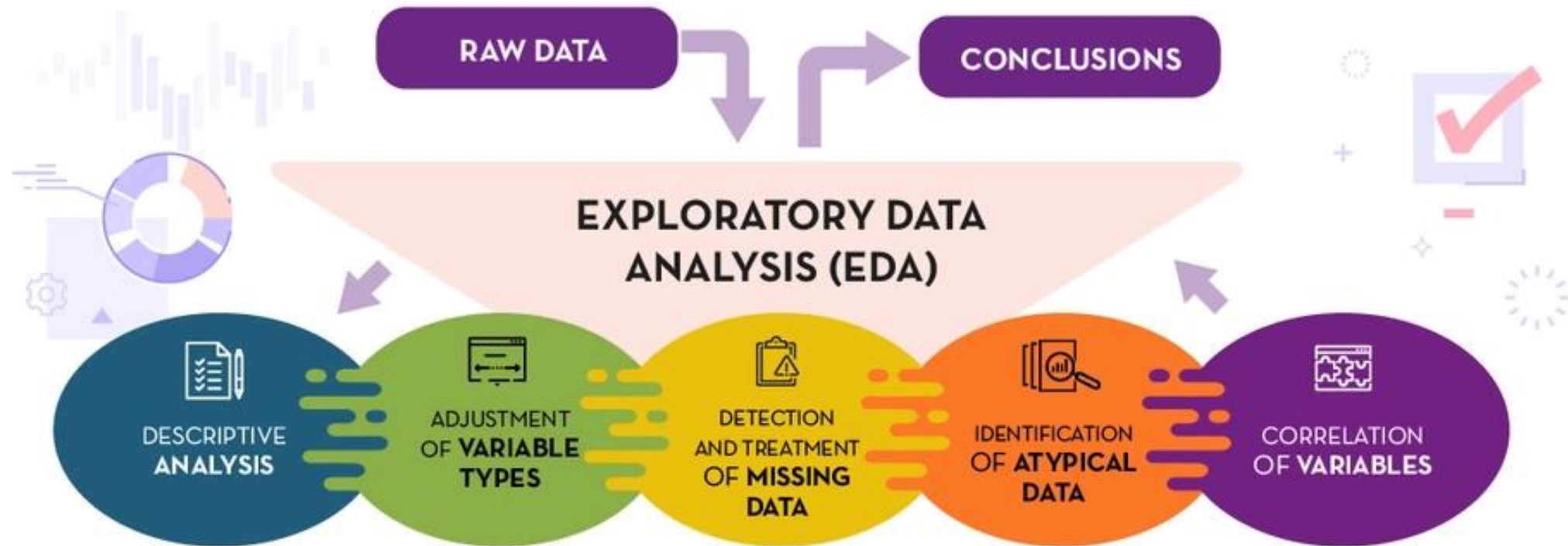
Разведочный анализ данных (EDA)

Преподаватель: Меликян Алиса Валерьевна, amelikyan@hse.ru
кандидат наук, доцент Департамента программной инженерии ФКН НИУ ВШЭ,
академический руководитель магистерской программы
«Продуктовый подход и аналитика данных в HR-менеджменте»

Разведочный анализ данных

Разведочный анализ данных (Exploratory Data Analysis, EDA) – анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий. Это первый и один из самых важных этапов работы с данными до моделирования. Он помогает ответить на вопросы:

- Что содержат данные?
- Какие проблемы в данных?
- Какие признаки важны?
- Как лучше подготовить данные для модели?
- Какие гипотезы можно сформулировать?





Основные шаги EDA

1. Понимание решаемой проблемы и специфики данных (объем и структура данных, типы признаков).
2. Загрузка и изучение данных (поиск пропусков, дубликатов, аномальных значений).
3. Описательный анализ данных (статистики и визуализация).
4. Анализ взаимосвязей между признаками.
5. Формулирование выводов и гипотез, которые будут проверены в ходе построения моделей.



Результаты EDA

- ✓ Понимание специфики данных;
- ✓ Список выявленных проблем;
- ✓ Идеи для feature engineering;
- ✓ Понимание какие алгоритмы подойдут для построения модели;
- ✓ Формулировка бизнес-гипотез;
- ✓ Необходимость сбора дополнительных данных;
- ✓ Приоритизация следующих шагов.



Типы структурированных данных

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

Перекрёстные данные

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

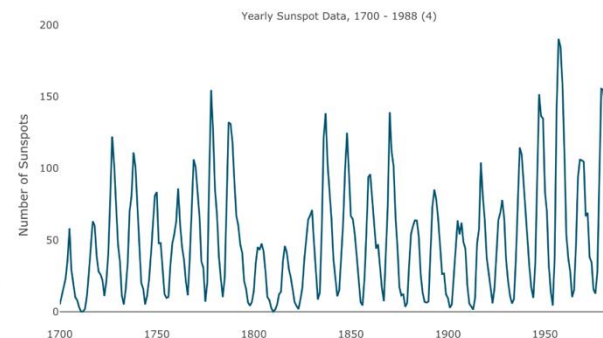
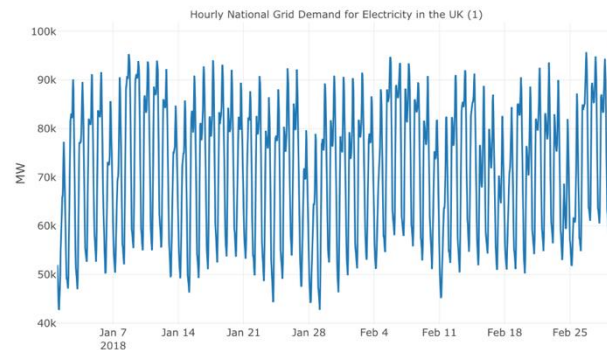


Структура перекрёстных данных

- Каждая строка – отдельный объект (сотрудник, организация, город, страна).
- Каждый столбец – значения признака объекта в конкретный момент времени (пол, возраст, заработная плата сотрудника).
- Ячейки содержат значения (числовые, текстовые, даты). Каждая ячейка содержит одно значение конкретного признака для определённого объекта.

Временной ряд

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



Панельные данные

Панельные данные (Panel Data or Longitudinal Data) состоят из наблюдений одних и тех же объектов, которые осуществляются в последовательные периоды времени. Представляют собой комбинацию перекрёстных данных и временных рядов. Насчитывают три измерения:

- 1) признаки,
- 2) объекты,
- 3) время.



Примеры 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

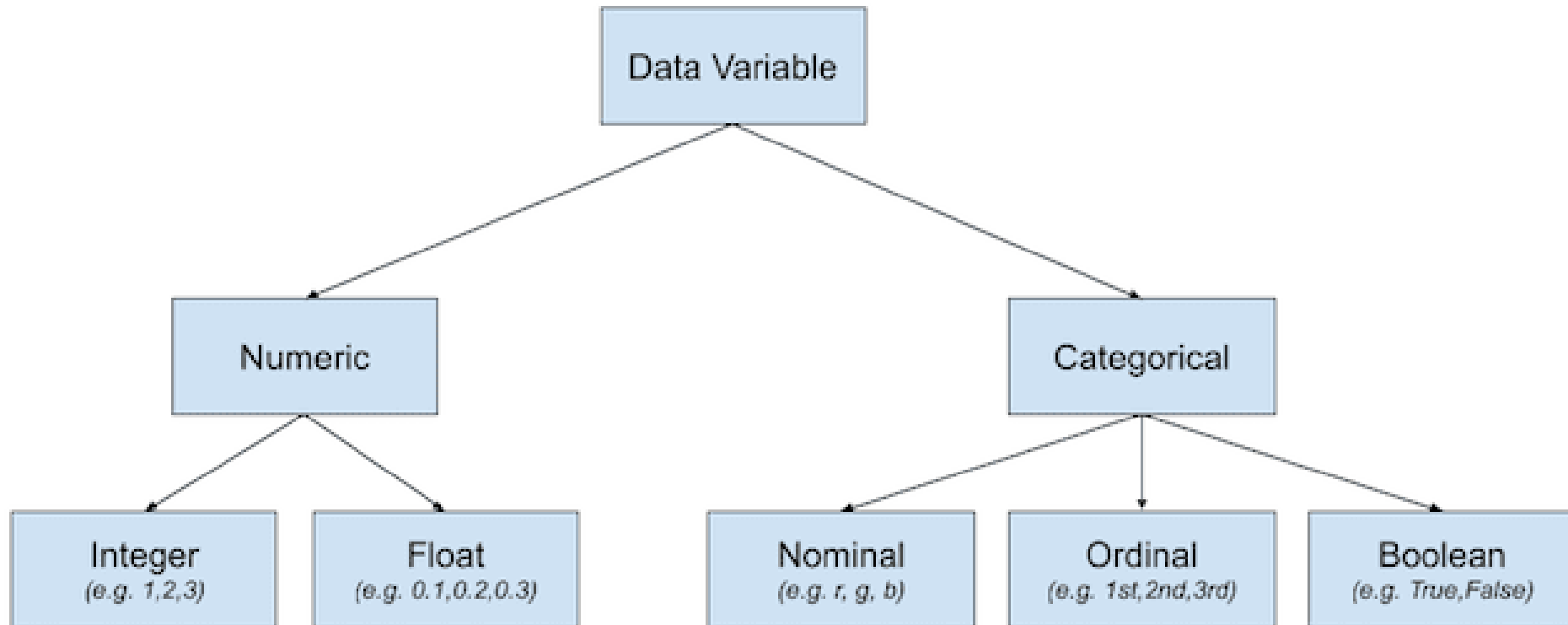
Person ID	Age	Sex	Year	Income
1	27	1	2015	1600
1	28	1	2016	1500
2	42	2	2015	1900
2	43	2	2016	2000
2	44	2	2017	2100
3	34	1	2015	3300

Типы признаков

- **Бинарный признак** принимает два значения (обычно наблюдается свойство объекта или нет): 1 или 0.
- **Номинальный признак** используется для качественной классификации. Определяет принадлежность объекта к определённому классу, отличающемуся от других но при этом классы не подлежат упорядочиванию (например, национальность, цвет, город).
- **Порядковый признак** позволяет ранжировать объекты, указав какие из них в большей или меньшей степени обладают определённым свойством, оцениваемым этим признаком. Разница между значениями признака существует, но её нельзя измерить (уровень образования, степень удовлетворенности).
- **Числовой признак** позволяет не только упорядочивать объекты, но и численно выразить и сравнить различия между ними.



Типы признаков



Определение типа признака

- Возраст
- Пол
- Стаж
- Образование
- Должность
- Отдел
- Уровень квалификации
- Количество опозданий
- Уровень удовлетворенности условиями труда
- Отношение к руководству
- Число прогулов



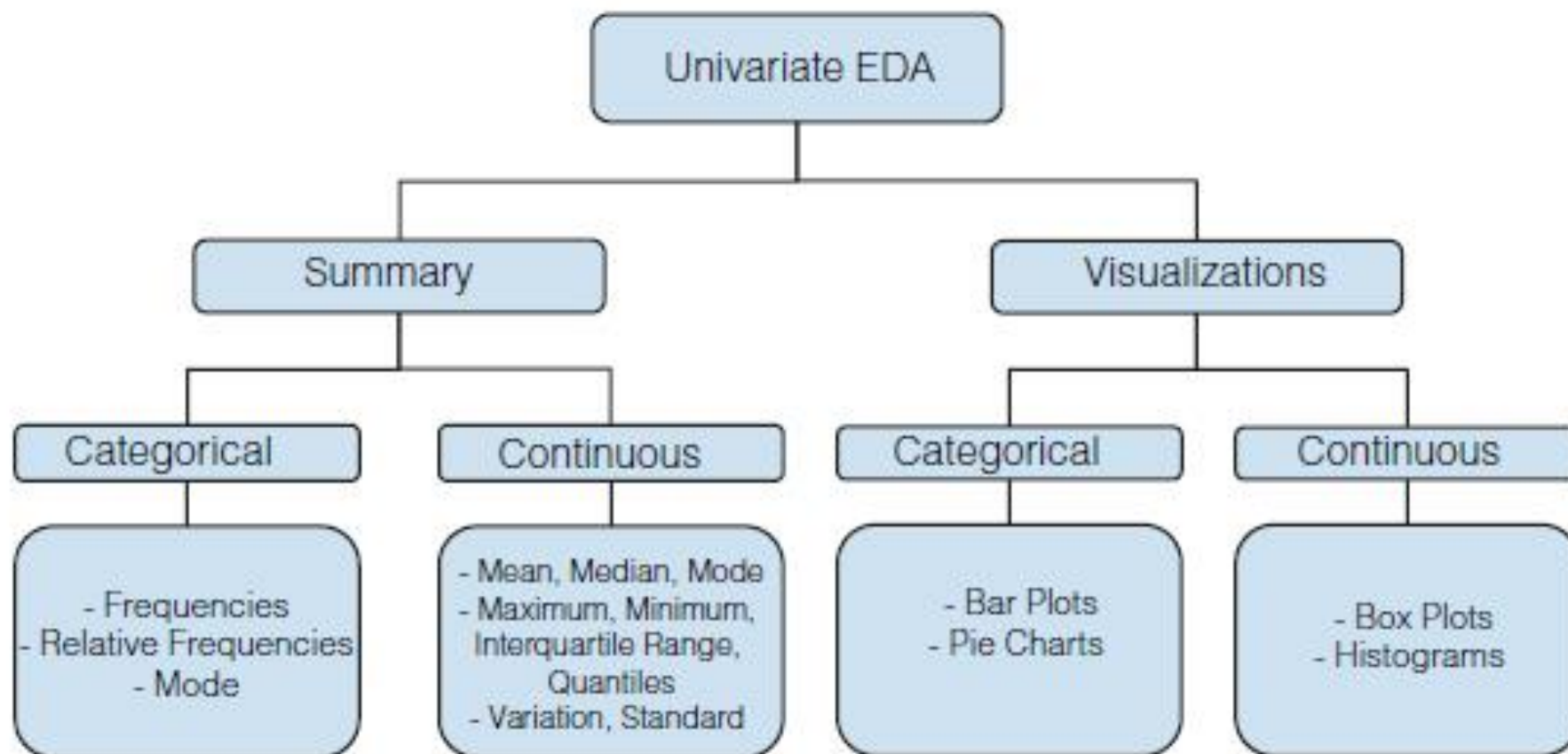
Одномерный EDA

Одномерный анализ (Univariate) – это анализ одного признака без учета его связей с другими. Используется на раннем этапе EDA для определения типа признака и методов, которые можно использовать для его анализа, а также понимания особенностей распределения значений признака, поиска выбросов и аномалий, нахождения пропусков и ошибок в данных. Виды анализа отличаются для категориальных и числовых признаков.

Помогает выбрать способ кодирования признака, принять решение о нормализации его значений или определить необходимость модификации данных.

Без качественного одномерного EDA модель может учиться на ошибках данных, а не на закономерностях.

Одномерный EDA



Одномерный EDA

Для **числовых признаков** используются описательные статистики (меры центральной тенденции и разброса), гистограмма (histogram), ящичковая диаграмма (boxplot), график плотности распределения (density plot). На графиках мы можем оценить нормальность распределения (normality), выявить асимметрию (skewness) или пиковость/пологость (kurtosis), диагностировать выбросы (outliers).

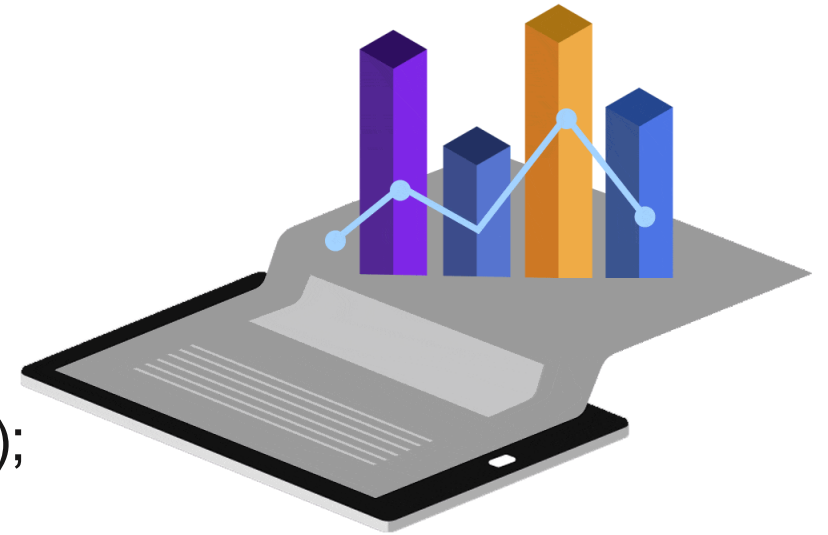
Для **категориальных признаков** применяется частотный анализ (frequency distribution), рассматривается количество уникальных значений (cardinality), строятся столбиковые (bar chart) или круговые диаграммы (pie chart). Анализ направлен на выявление дисбаланса классов, редких категорий, ошибок в названиях категорий, рассмотрения возможностей удаления или объединения некоторых категорий.

Одномерный EDA

Может выявить следующие проблемы:

- отрицательные значения (зарплата < 0);
- невозможные значения (возраст 150 или 99999999);
- скрытые пропуски ("Unknown", "-");
- сильный дисбаланс классов (95% и 5%);
- нерепрезентативность класса (в одном классе мизерное число объектов);
- разные варианты написания идентичных категорий (г. Москва, Москва, город Москва).

Выявленные проблемы станут основой для preprocessing и feature engineering.

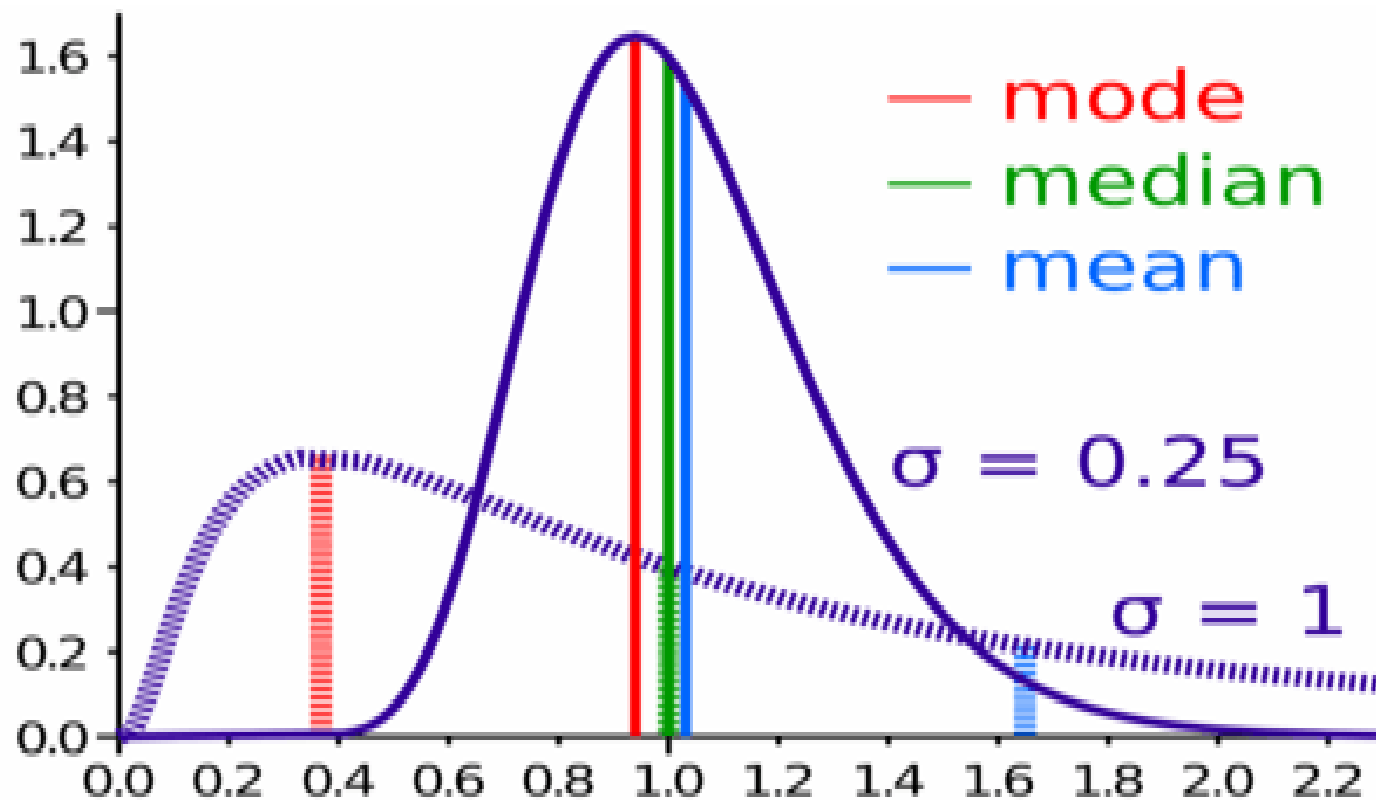


Меры центральной тенденции

Меры центральной тенденции позволяют охарактеризовать множество значений признака, измеренного на выборке, одним числом. Показывают концентрацию группы значений на числовой шкале.

Шкала измерения признака	Допустимые меры центральной тенденции
номинальная	мода
порядковая	мода, медиана
числовая	мода, медиана, среднее арифметическое

Меры центральной тенденции



Если распределение похоже на нормальное, то мода, медиана и среднее арифметическое близки по значениям и можно выбрать любую статистику, чтобы охарактеризовать центральную тенденцию.

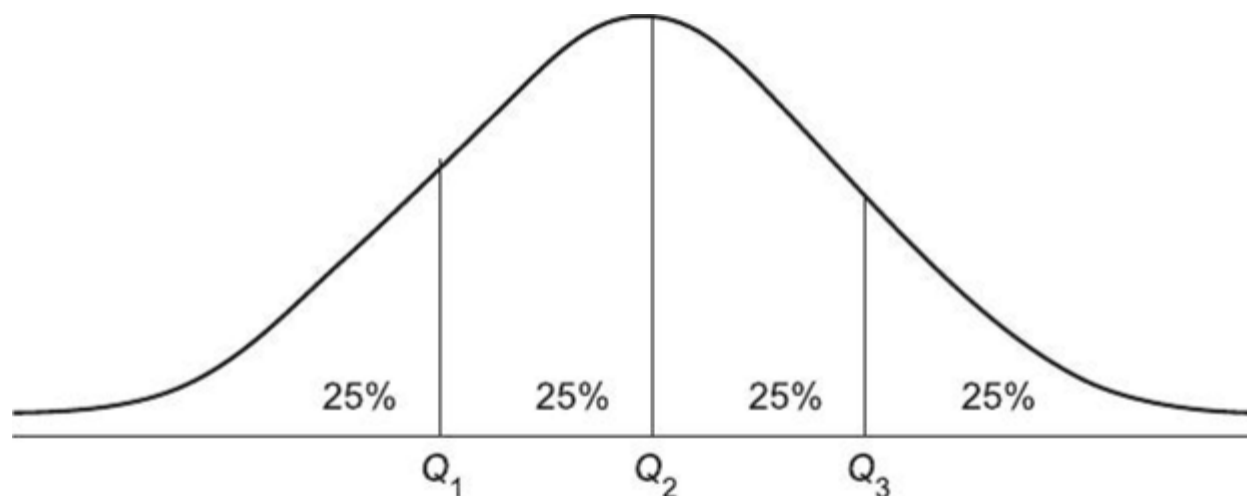
Если распределение отличается от нормального, то лучше выбрать медиану.

Квартиль

Первый квартиль (25-й перцентиль) – точка на шкале значений признака, ниже значения которой находятся 25% значений признака.

Второй квартиль (медиана) – точка на шкале значений признака, ниже значения которой находятся 50% значений признака.

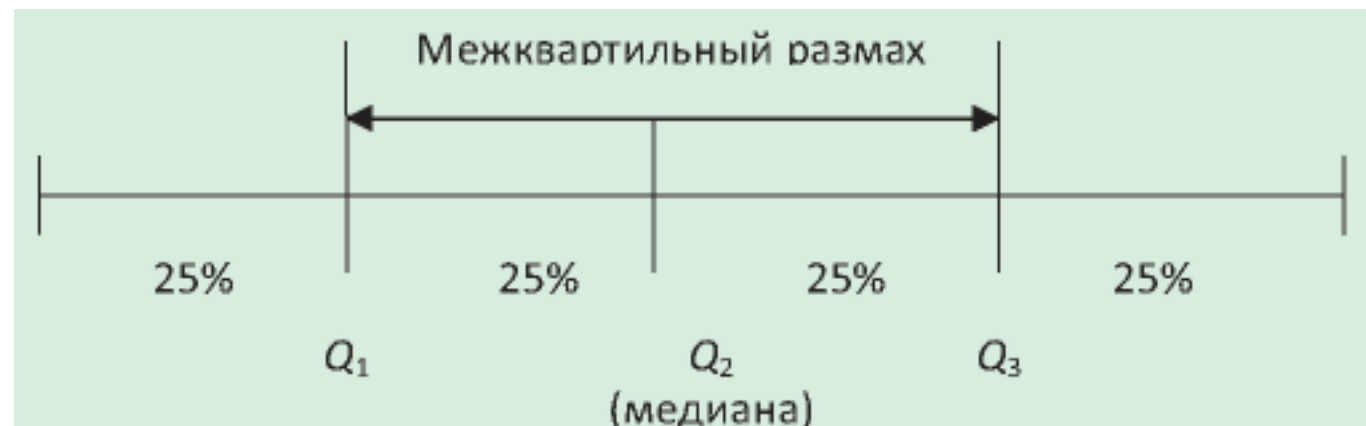
Третий квартиль (75-й перцентиль) – точка на шкале значений признака, ниже значения которой находятся 75% значений признака.



Межквартильный размах

Межквартильный размах (IQR) – это разница между третьим и первым квартилями ($Q_3 - Q_1$), которая показывает диапазон, в котором находятся центральные 50% значений и используется для оценки разброса данных без учета выбросов.

$$IQR = Q_3 - Q_1$$



Межквартильная широта

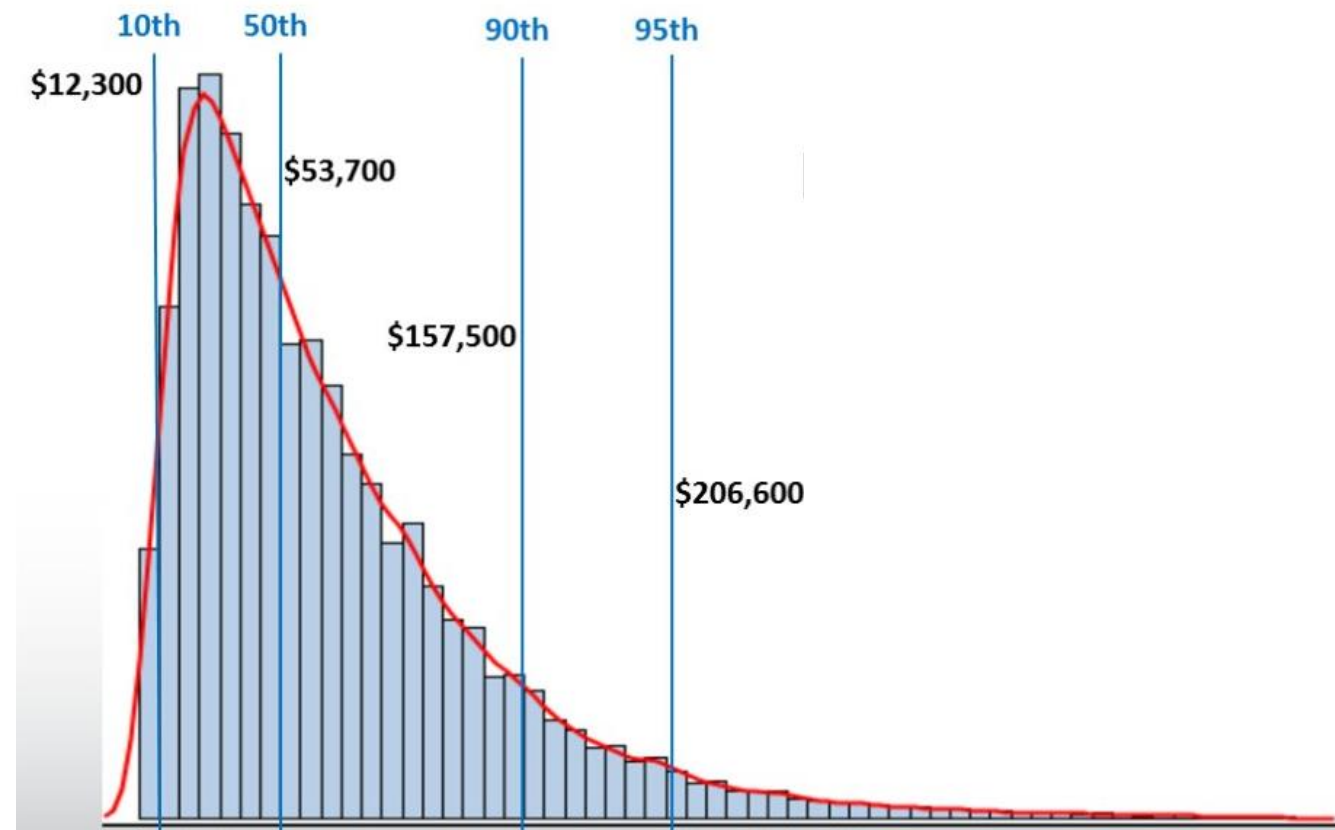
Для оценки меры разброса порядковых признаков может подсчитываться межквартильная широта.

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

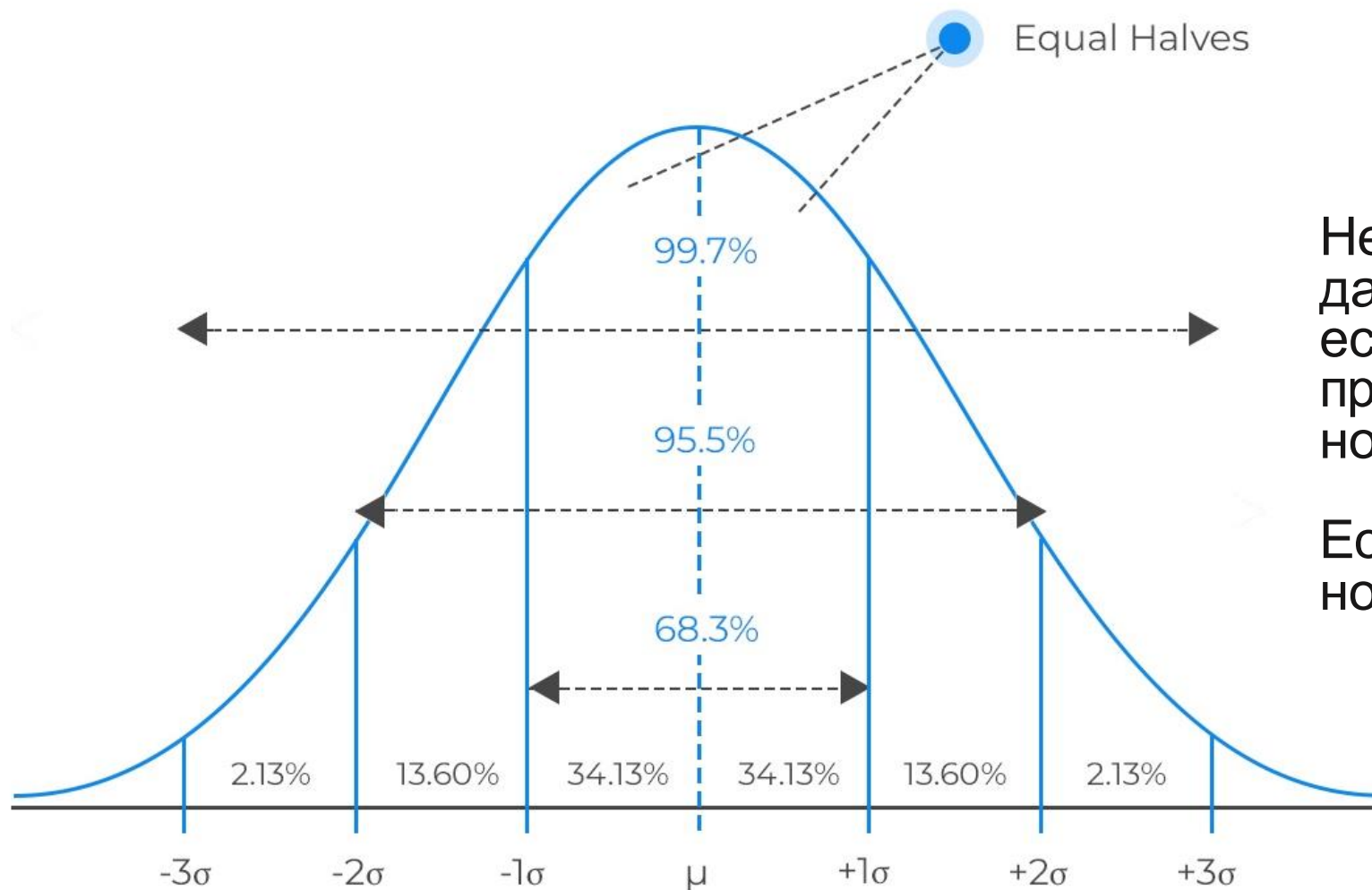
Если порядковый признак принимает 7 различных значений, а межквартильная широта равна 3, это говорит о том, что медиана как показатель центральной тенденции не полностью отражает распределение данных. Большая межквартильная широта указывает на значительное разброс значений вокруг медианы.

Децильное отношение

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



Нормальное распределение



Некоторые методы исследования данных дают корректные оценки, если значения анализируемых признаков подчиняются нормальному распределению.

Есть тесты, проверяющие нормальность распределения.

Оценка отличия распределения от нормального

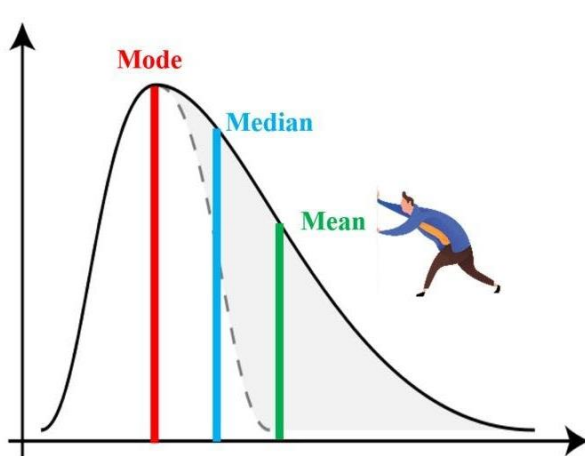
Распределение может отличаться от нормального по двум основным характеристикам:

- 1) симметричности (*skewness*);
- 2) заострённости (*kurtosis*).

Для нормального распределения значение коэффициентов асимметрии и эксцесса равны 0.

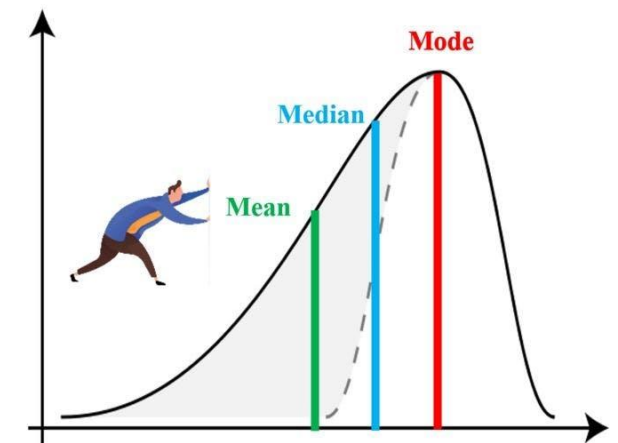
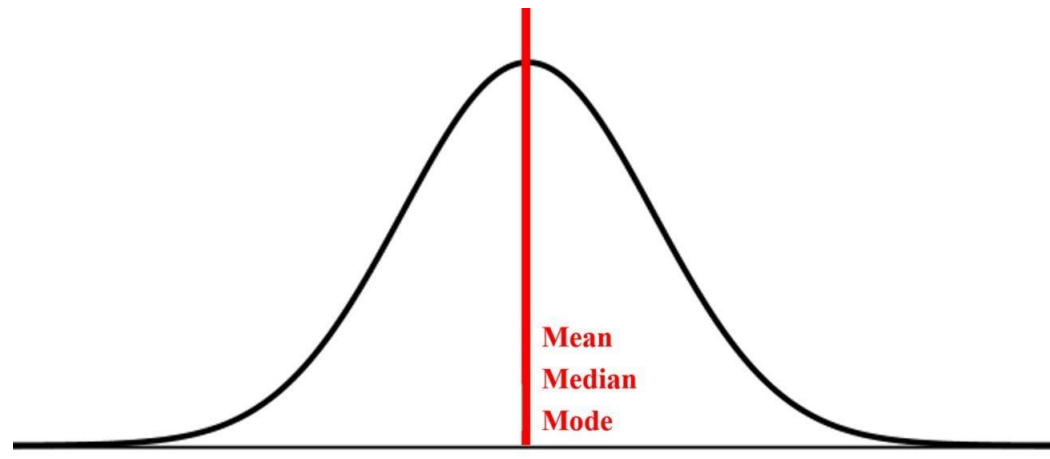
Симметричность распределения

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

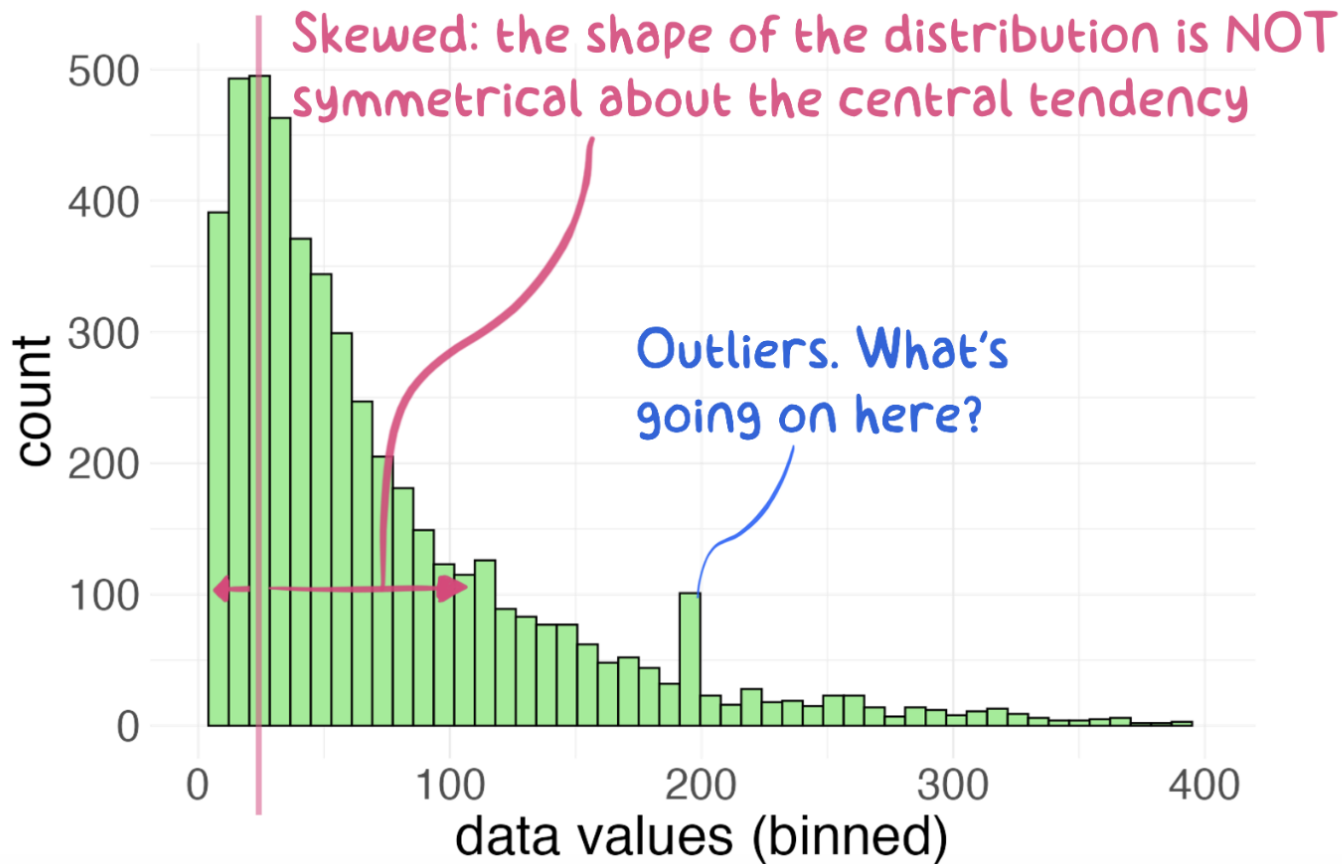


$$\text{Skewness} = \frac{\sum_i^N (X_i - \bar{X})^3}{(N - 1) * \sigma^3}$$

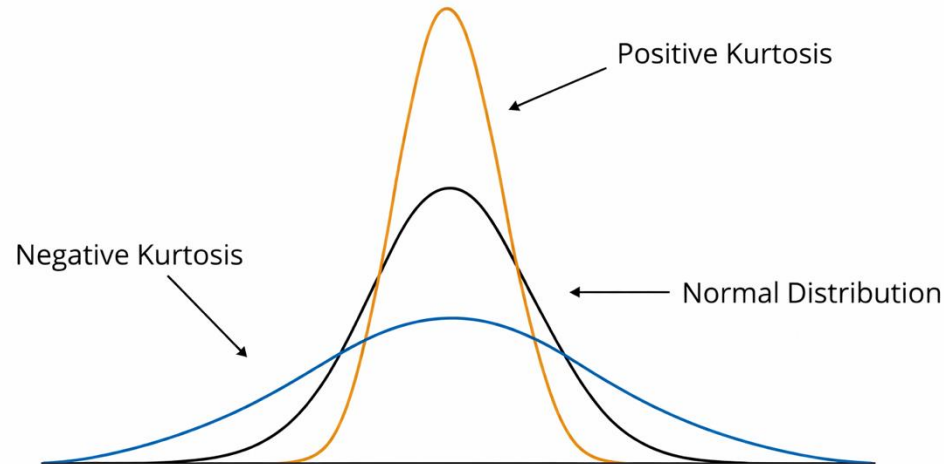
При отрицательной асимметрии в распределении чаще встречаются более высокие значения признака.



Симметричность распределения



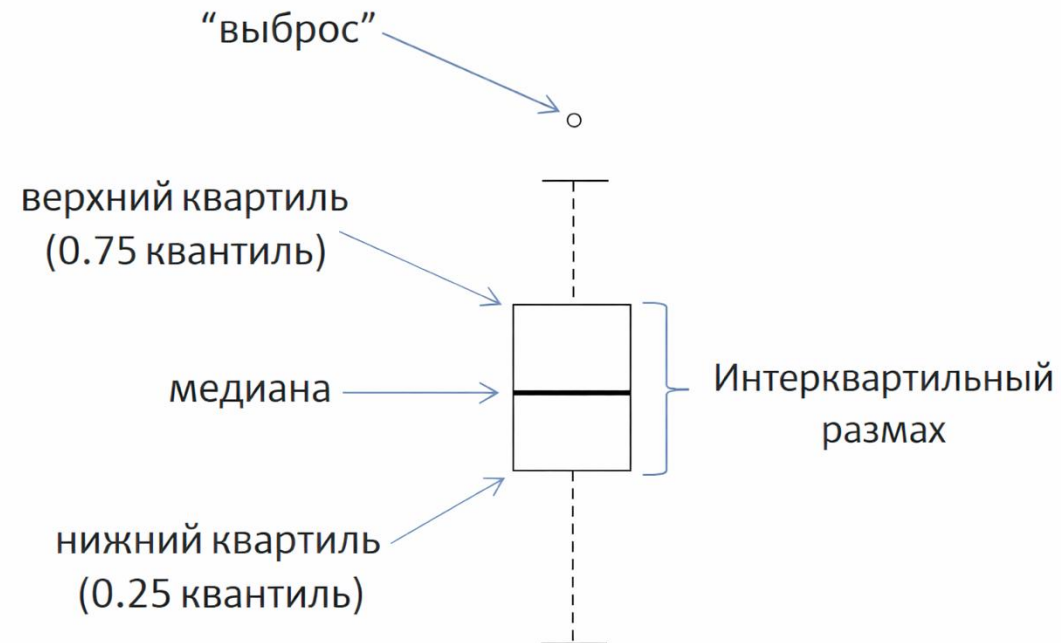
Заострённость распределения



$$\text{Kurtosis} = n * \frac{\sum_i^n (Y_i - \bar{Y})^4}{\sum_i^n (Y_i - \bar{Y})^2^2}$$

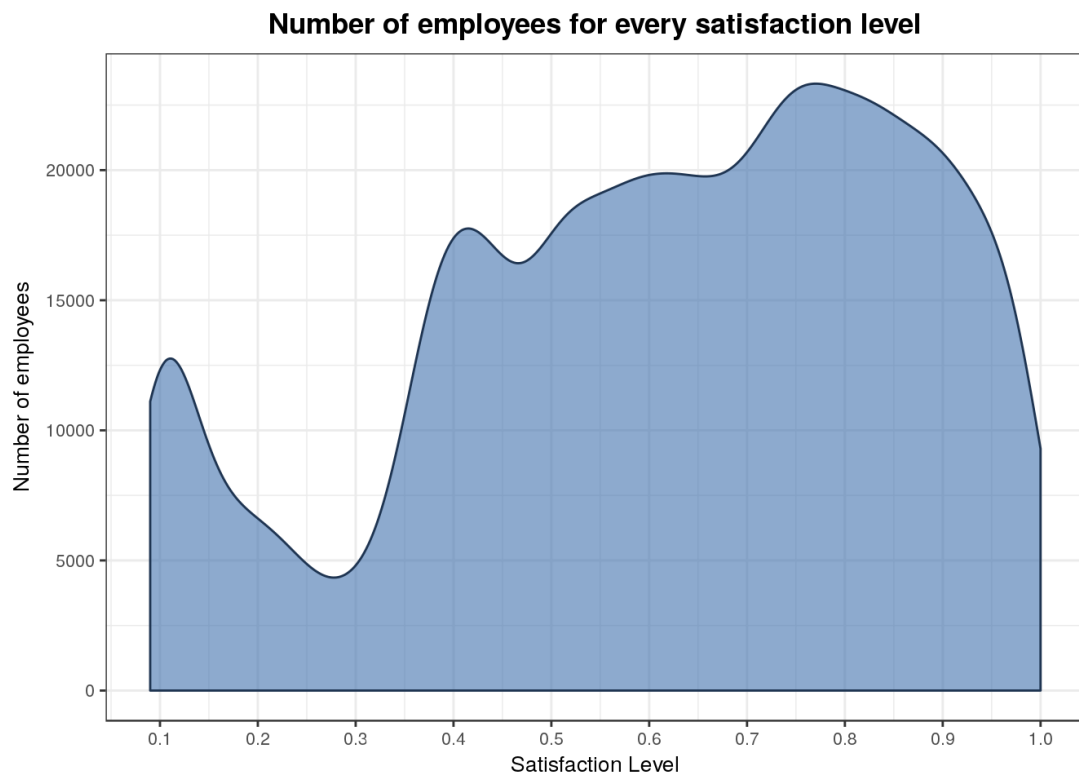
- Пиковое распределение ($\text{kurtosis} > 0$) характеризуется выраженным максимумом: большинство наблюдений сосредоточено в узком диапазоне, а разброс данных относительно невелик.
- Пологое распределение ($\text{kurtosis} < 0$) имеет сглаженную форму без ярко выраженного пика, что указывает на более равномерное распределение значений и большую вариативность данных.

Ящичковая диаграмма



Подходит для визуализации распределения значений числовых и порядковых признаков.

Одномерный EDA

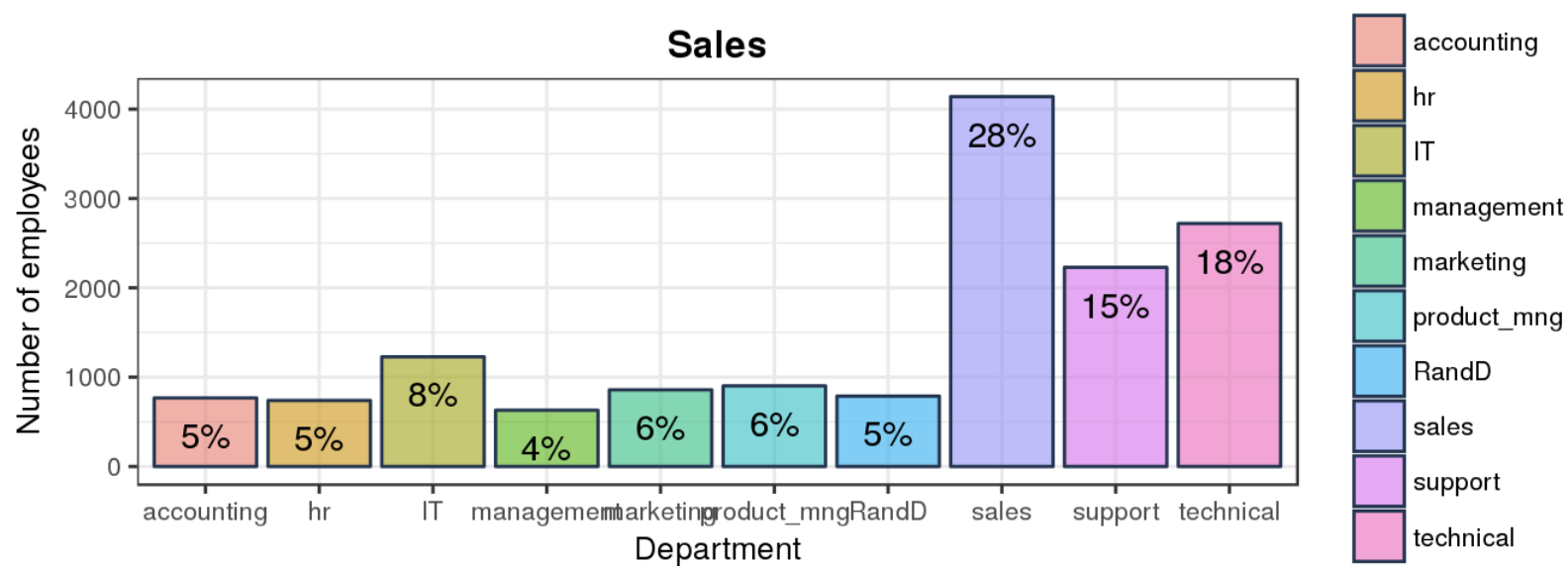


Распределение плотности показателя удовлетворённости бимодальное. Среднее (0.61) близко к медиане (0.64). Межквартильный размах: $0.82 - 0.44 = 0.38$, что характеризует довольно большой разброс значений. Возможно стоит перекодировать в порядковый признак с небольшим числом категорий «низкий», «средний» и «высокий».

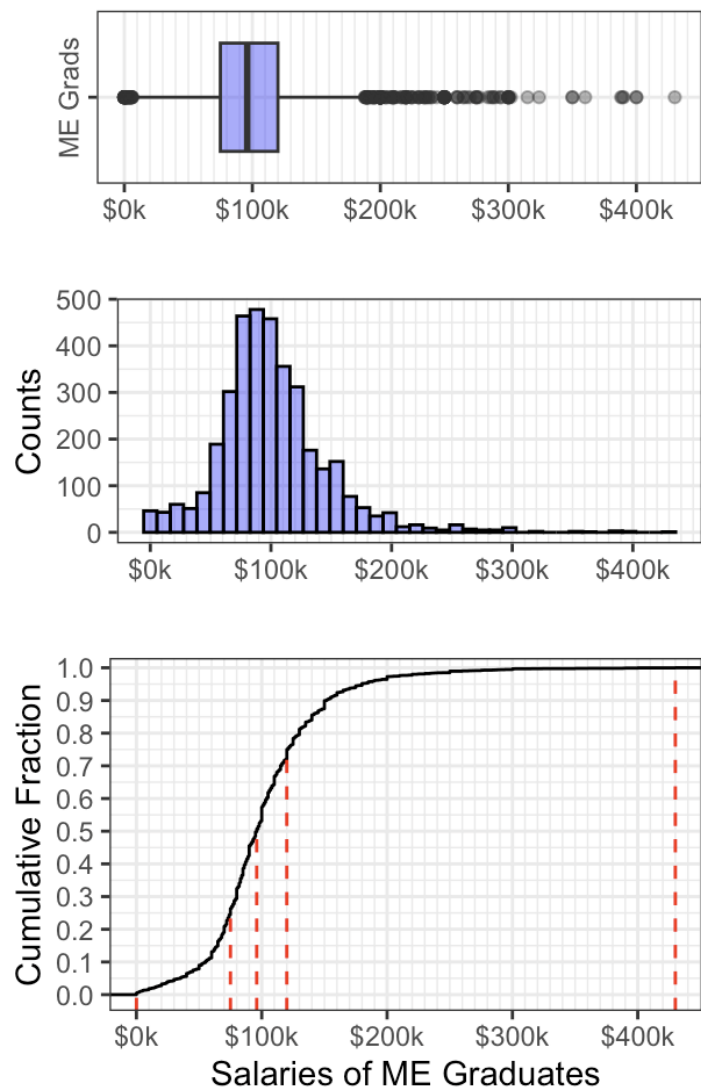
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0900	0.4400	0.6400	0.6128	0.8200	1.0000



Одномерный EDA



Показано распределение численности сотрудников по отделам в долях. Самые крупные отделы: продажи, поддержка, технический. Если в некоторых отделах очень малое число объектов, то для моделирования можно подумать над их объединением, либо по тематике, либо создать группу «другие».



На графиках показано распределение заработных плат выпускников: boxplot и гистограмма указывают на правостороннюю асимметрию распределения с основной концентрацией значений в диапазоне около 80–120 тыс. долларов и наличием выбросов с высокими доходами, достигающими 400 тыс. долларов и выше. Кумулятивная кривая демонстрирует, что большая часть выпускников получает заработную плату ниже 150 тыс. долларов, при этом рост кумулятивной доли замедляется на высоких уровнях дохода, что подтверждает редкость экстремально высоких значений.

Для моделирования может быть применено лог-преобразование, чтобы уменьшить асимметрию и снизить влияние экстремальных значений. При применении линейных моделей это улучшит их стабильность и качество.

Одномерный EDA

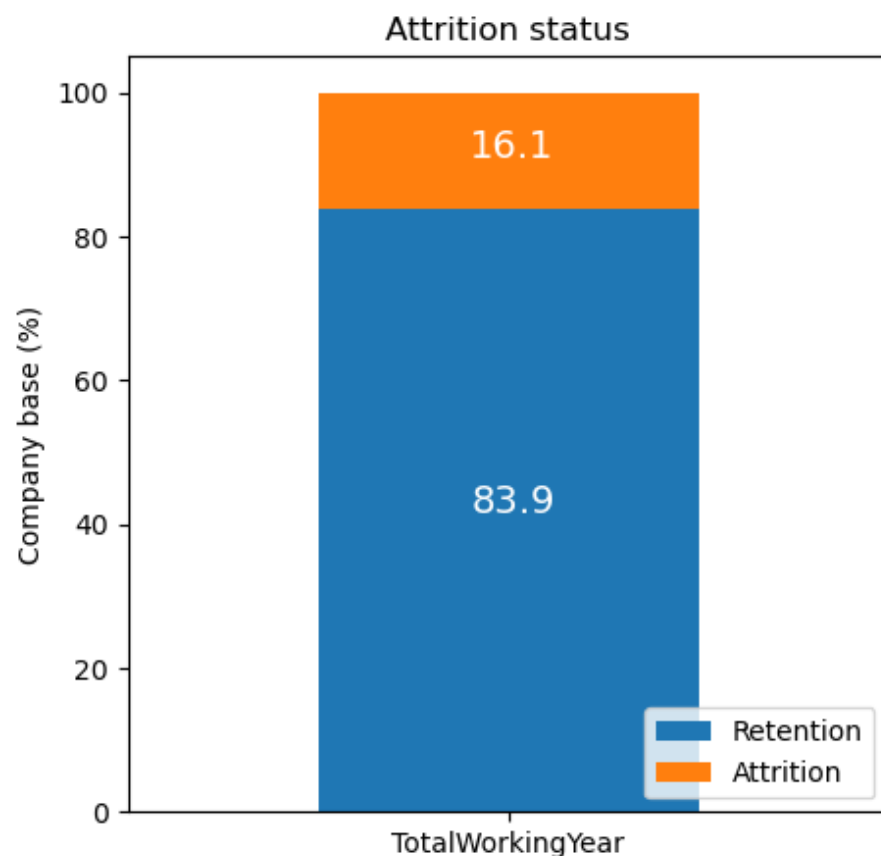


Диаграмма демонстрирует, что 83,9% сотрудников продолжили работу в компании, а 16,1% приняли решение об увольнении, что свидетельствует о приемлемом уровне удержания персонала. С точки зрения дальнейшего построения модели наблюдается выраженная несбалансированность классов. Может потребоваться применить методы обработки несбалансированных данных, настраивать алгоритмы с учётом весов классов либо использовать модели, изначально устойчивые к подобному дисбалансу.

Двумерный EDA

Двумерный анализ (Bivariate) исследует особенности взаимосвязей между двумя признаками.

feature ↔ target

feature ↔ feature

Это позволяет:

- выявить наличие и характер зависимости (линейная, нелинейная, положительная или отрицательная взаимосвязь);
- обнаружить закономерности и аномалии, которые не видны при одновариантном анализе;
- сформировать гипотезы для дальнейшего моделирования;
- выбрать подходящие методы обработки данных и модели, учитывая выявленные взаимосвязи.

Двумерный EDA

- Таблица сопряженности (Contingency Table);
- Гистограмма, ящичковая или столбиковая диаграмма с разделением на группы;
- Диаграмма рассеяния;
- Корреляции;
- Статистические тесты: Хи-квадрат, t-тесты, непараметрические тесты, ANOVA;
- Тепловые карты (heatmaps) корреляций и таблиц сопряженности.

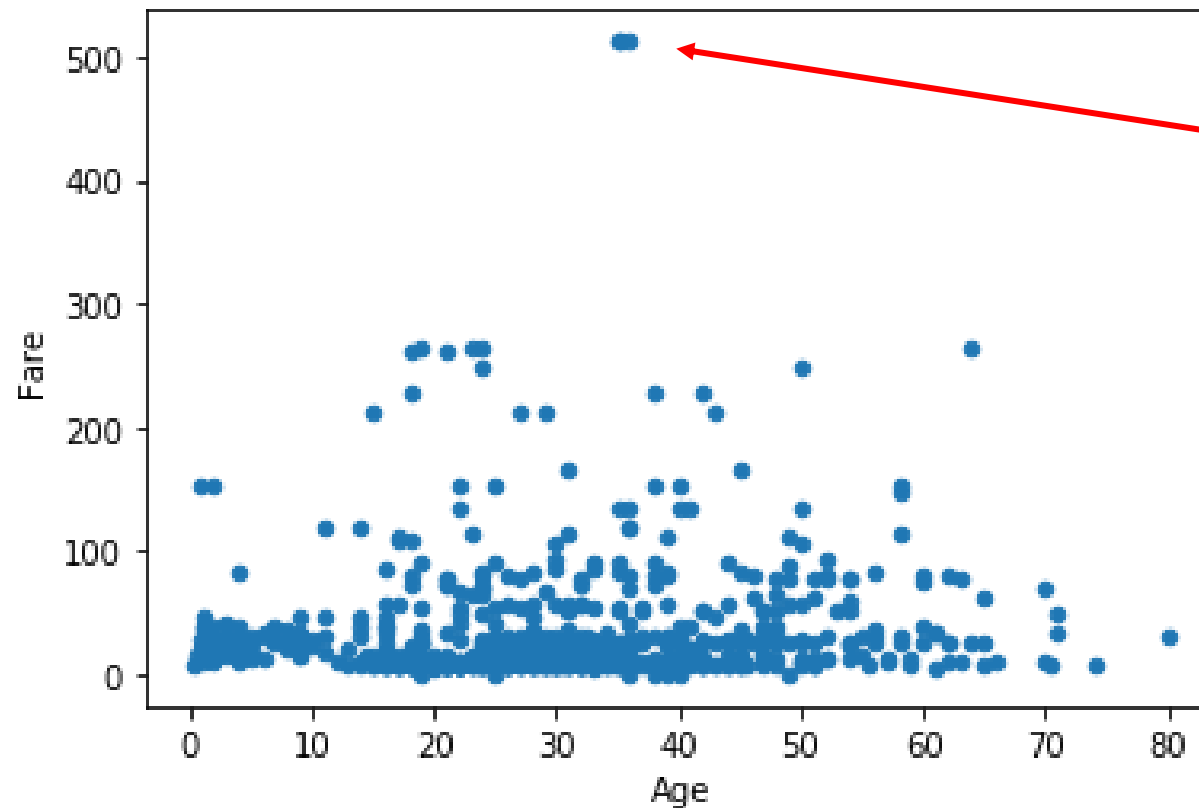
Таблица сопряженности

Связь между категориальными признаками можно исследовать на основе таблицы сопряжённости (crosstab). В ячейках могут быть частоты или проценты. Позволяют увидеть некоторые закономерности в распределении частот. В примере с ростом уровня IT-компетенций респондента возрастает и восприятие их важности среди HR-менеджеров.

	Perceived Importance of IT competencies for HR managers		
	<i>Skeptic</i>	<i>Average</i>	<i>Enthusiast</i>
Level of IT competencies held by respondents	%	%	%
<i>Weak</i>	31.40	37.30	31.40
<i>Fair</i>	13	48	39
<i>Strong</i>	10	39	51

Диаграмма рассеяния

Целевая
переменная



Потенциальные
выбросы

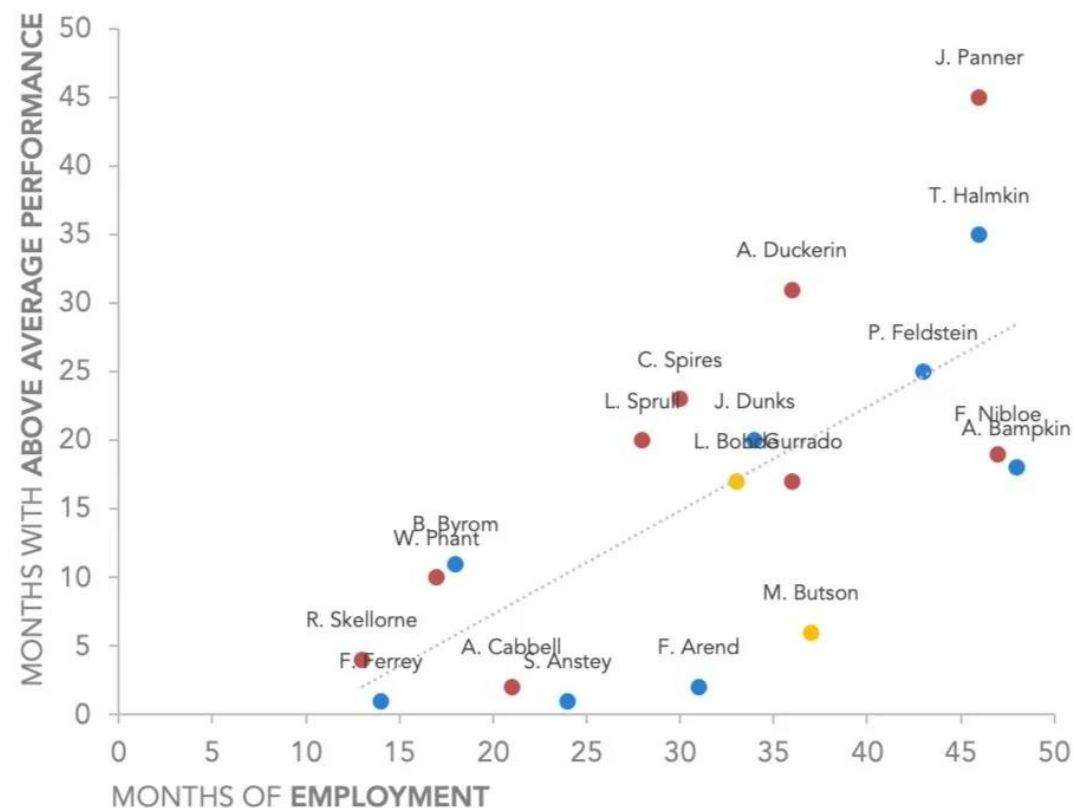
Взаимосвязь
очень слабая

Предиктор

Диаграмма рассеяния

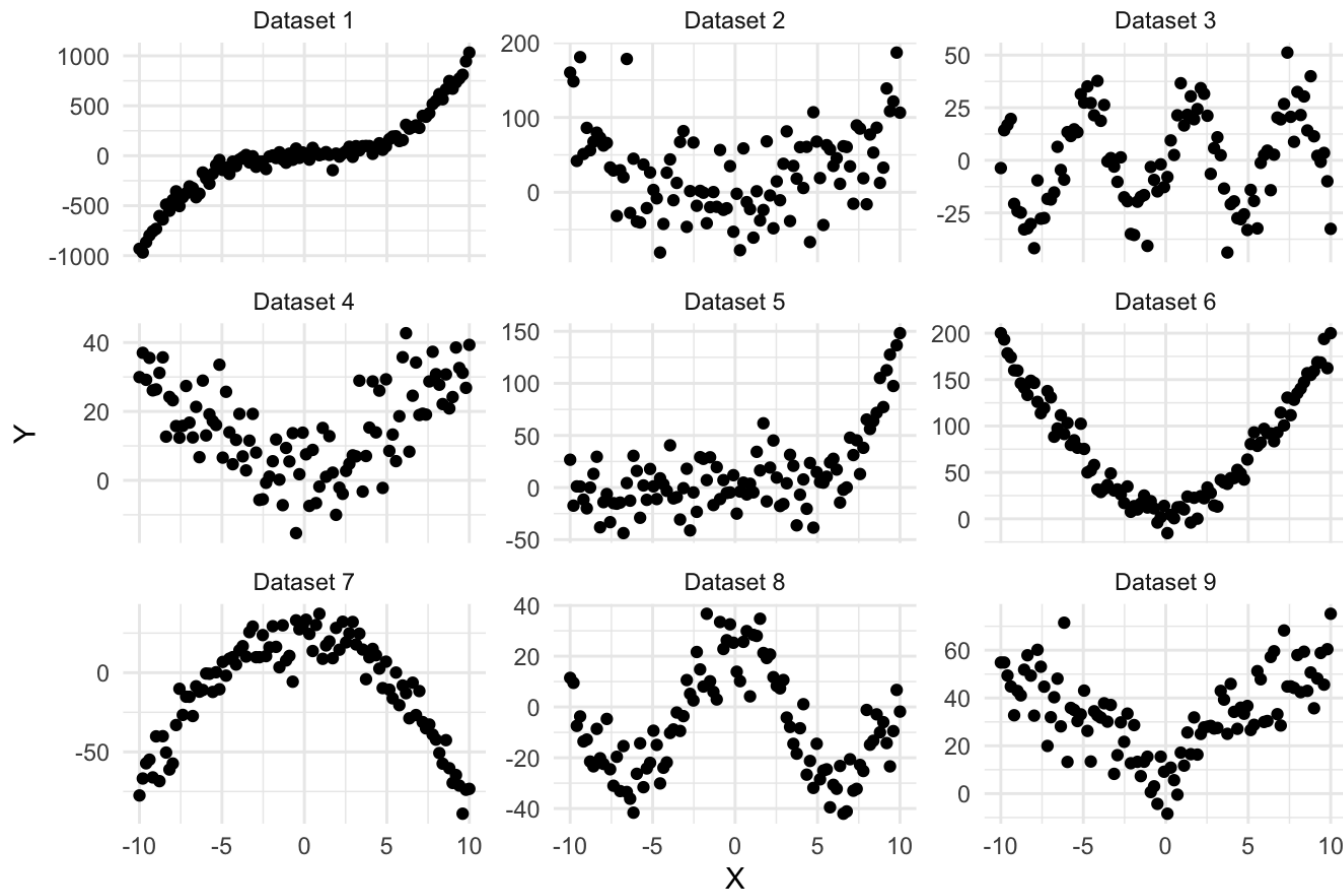
Performance of analysts after pilot program

● United States ● Japan ● Canada



Взаимосвязь
довольно сильная

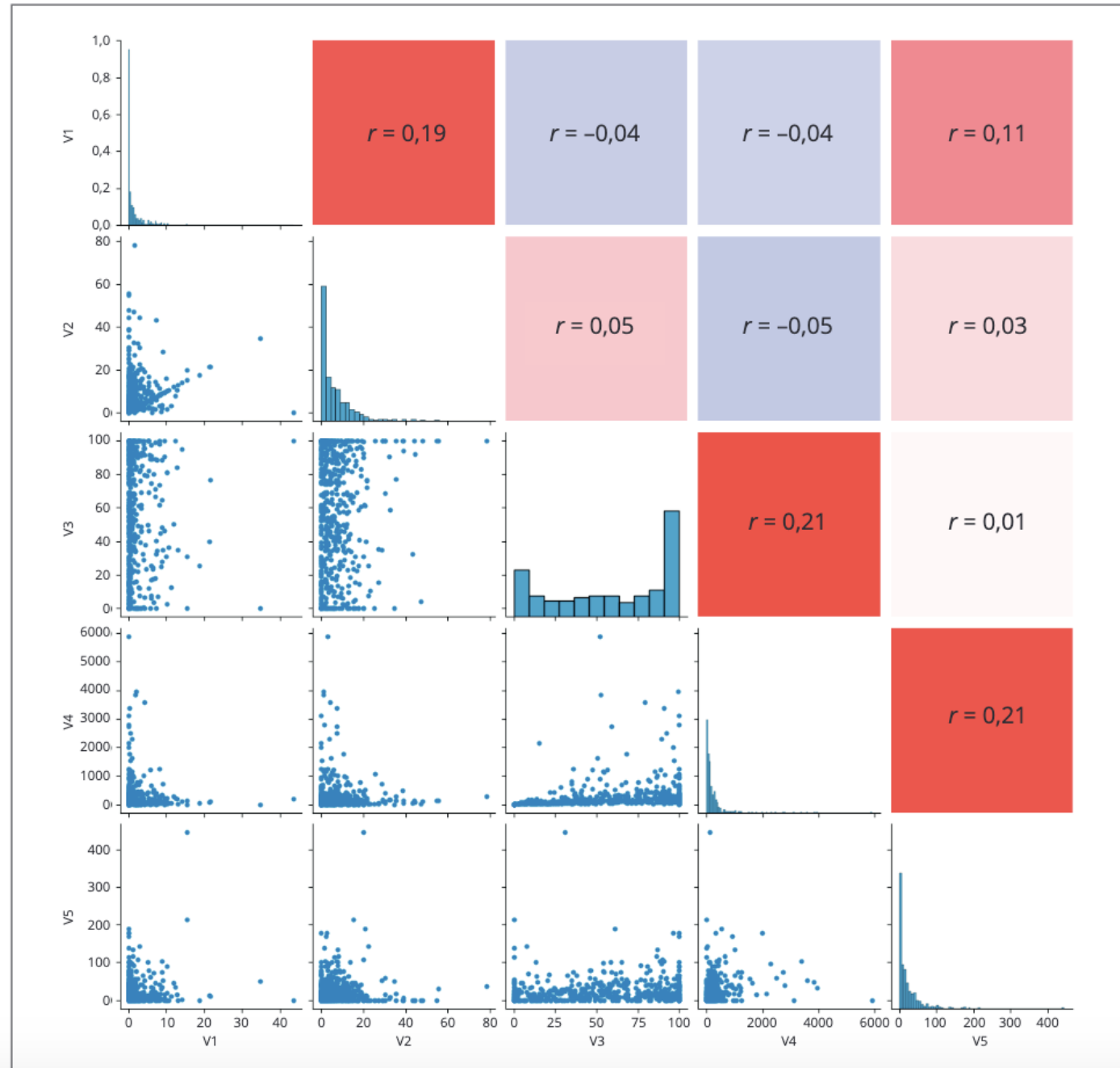
Диаграмма рассеяния



Нелинейные
зависимости



Матрица диаграмм рассеяния в комбинации с тепловой картой корреляций



Кластеризованная ящичковая диаграмма

Медианное значение уровня последней оценки увеличивается по мере роста количества проектов. Вероятно, оценщики выставляют более высокие баллы сотрудникам, задействованным в большем числе проектов, что может отражать более высокую вовлечённость или нагрузку. При этом в последней группе (7 проектов) больше всего выбросов. Красным цветом на графике показано среднее значение уровня последней оценки.

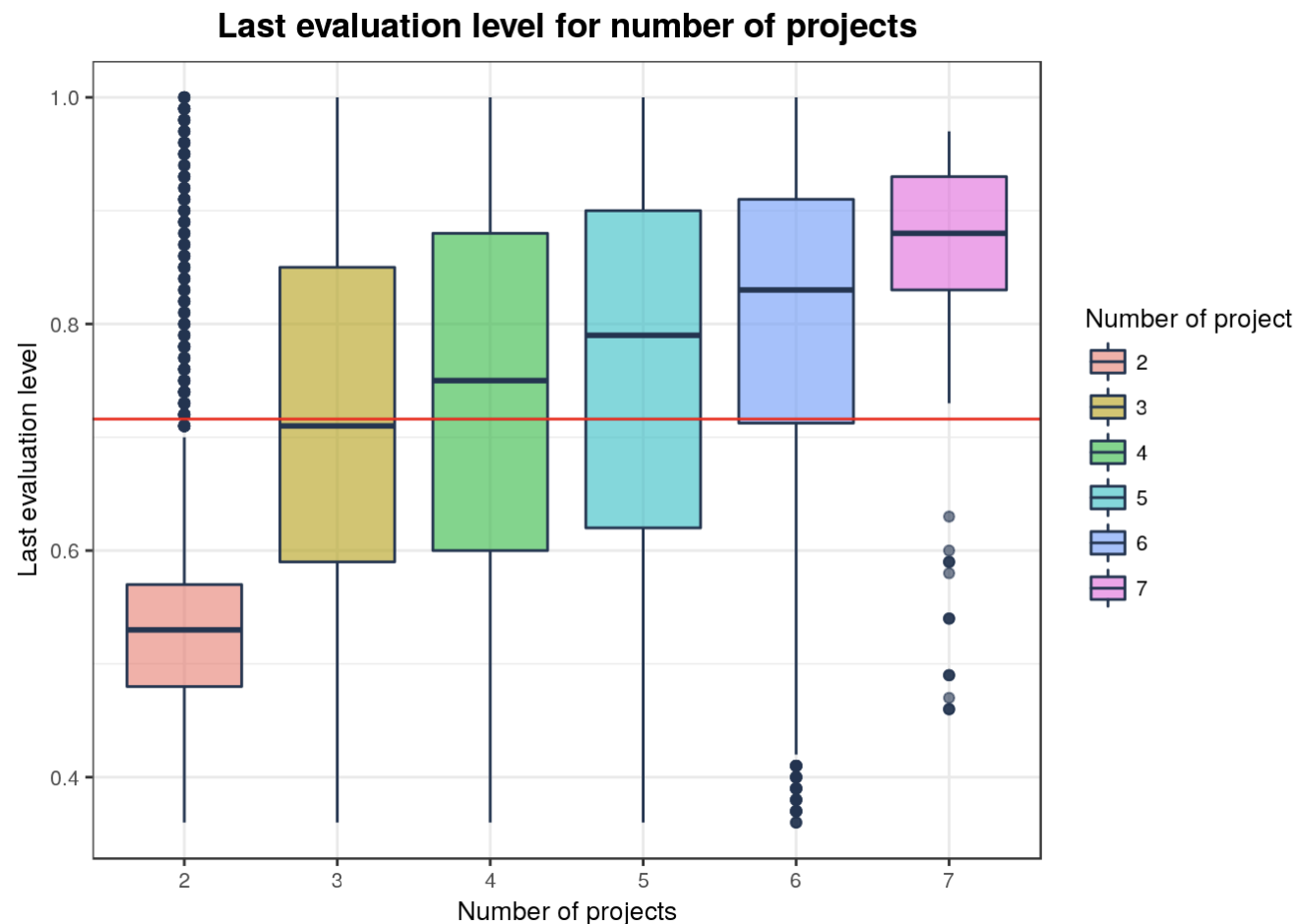
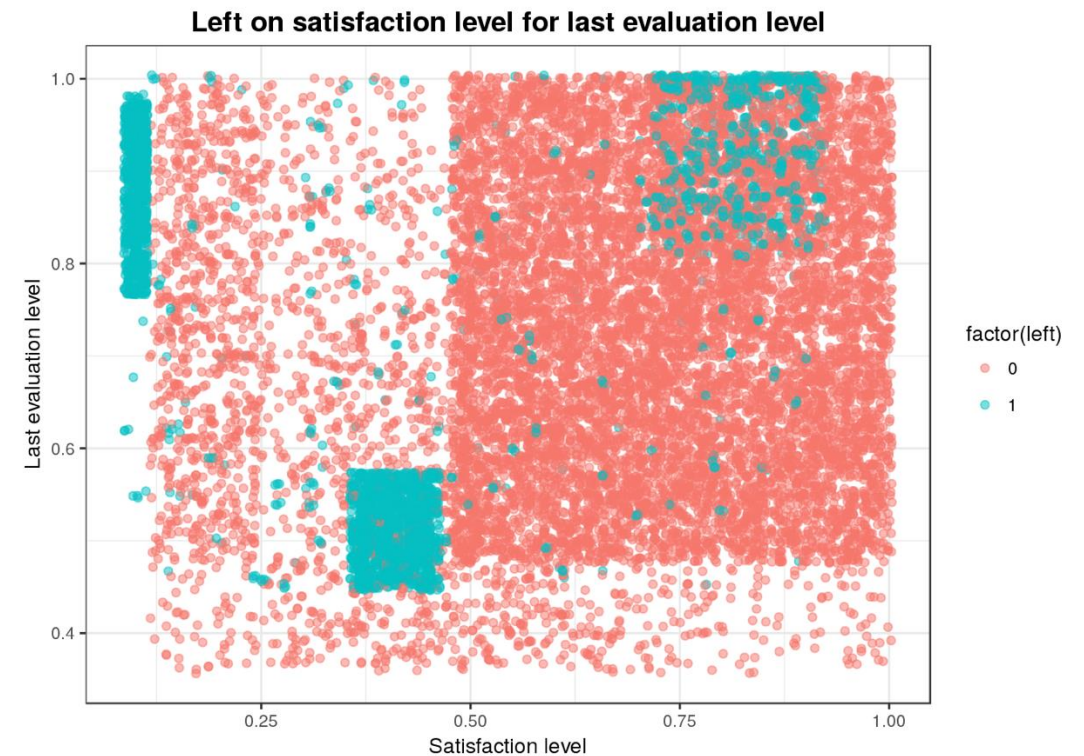


Диаграмма рассеяния с метками по значениям целевой переменной

На графике можно выделить три группы сотрудников, покинувших компанию:

1. сотрудники с высоким уровнем оценки, но низким уровнем удовлетворённости;
2. сотрудники с низким уровнем оценки и низким уровнем удовлетворённости;
3. сотрудники с высоким уровнем удовлетворённости и высоким уровнем оценки.



Этапы корреляционного анализа

1. Построить диаграмму рассеяния, чтобы изучить специфику взаимосвязи;
2. Выбрать подходящий коэффициент корреляции в зависимости от особенностей признаков;
3. Рассчитать коэффициент корреляции и соответствующее ему p-value;
4. Интерпретировать статистическую значимость, силу и направление взаимосвязи.

Коэффициент корреляции Пирсона

Измеряет силу линейной взаимосвязи между двумя признаками.

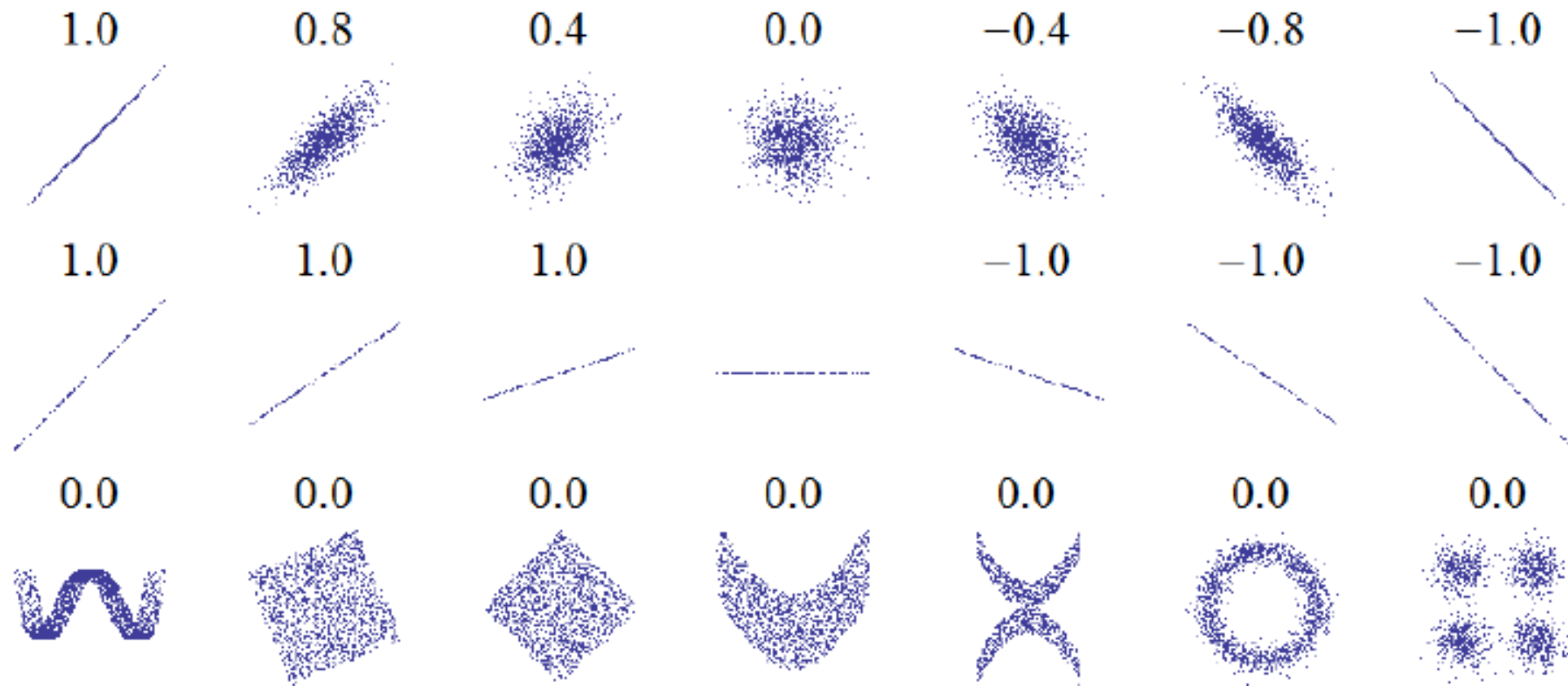
$$r_{xy} = \frac{\sum_{i=1}^n [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{(n-1) \cdot \sigma_x \cdot \sigma_y}$$

- Принимает значения от -1 до +1.
- Может быть подсчитан на основе числовых признаков, значения которых подчиняются закону нормального распределения.
- Измеряет направление (знак) и силу (величина) связи.

Интерпретация значений коэффициента корреляции

Значение коэффициента корреляции	Интерпретация
$0 < r \leq 0,2$	Очень слабая корреляция
$0,2 < r \leq 0,5$	Слабая корреляция
$0,5 < r \leq 0,7$	Средняя корреляция
$0,7 < r \leq 0,9$	Сильная корреляция
$0,9 < r \leq 1$	Очень сильная корреляция

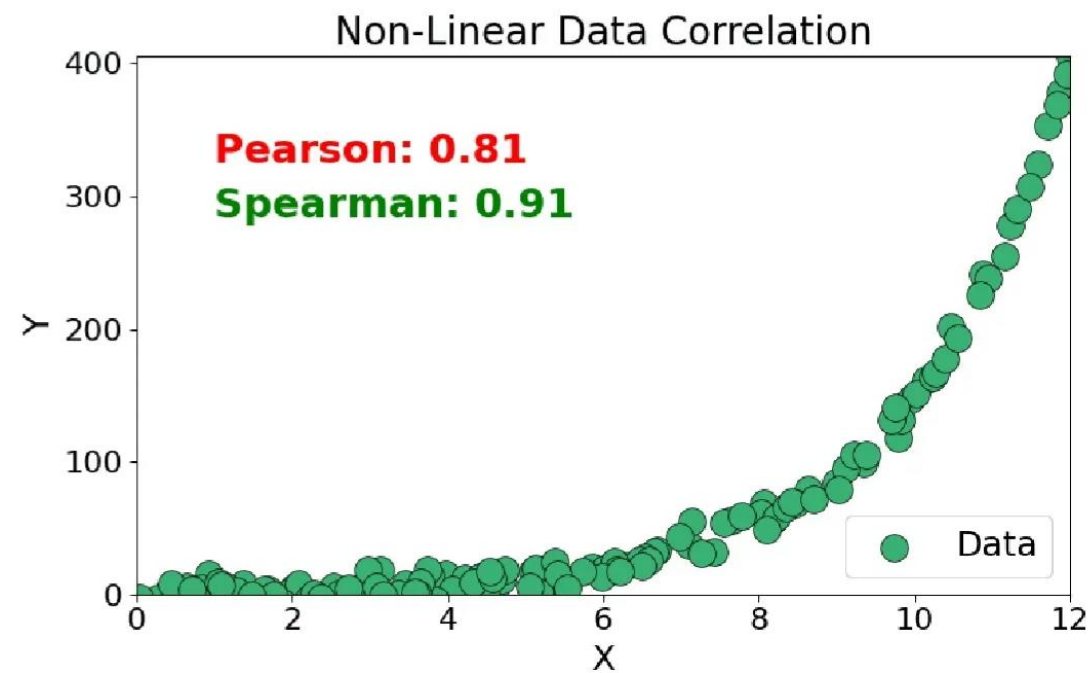
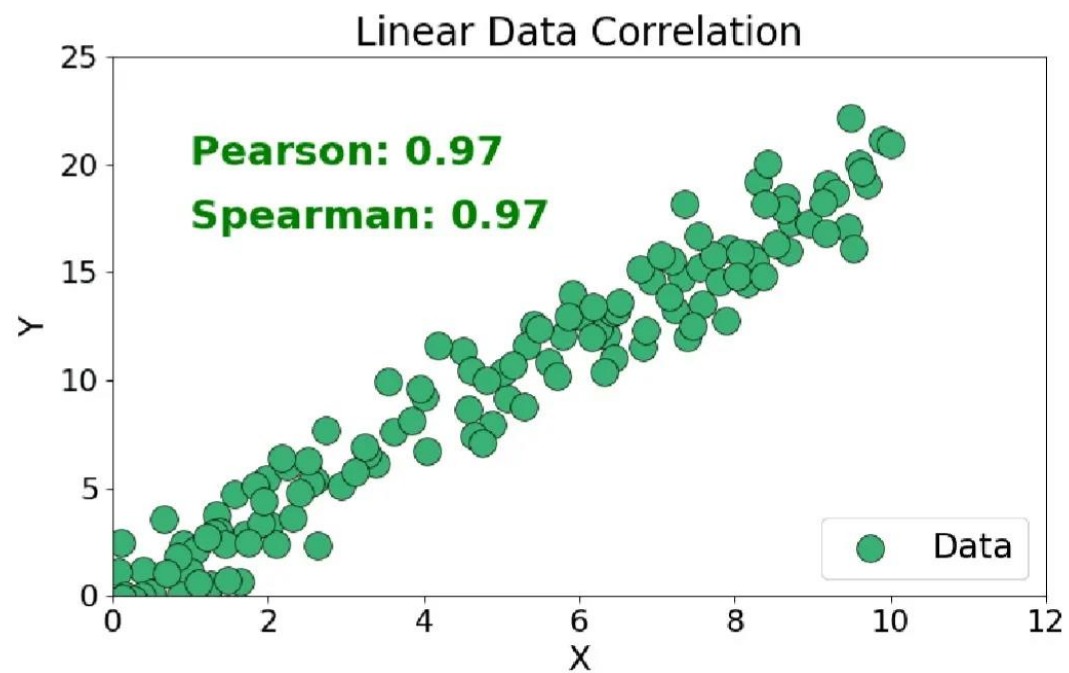
Диаграмма рассеяния для разных значений r Пирсона



Ранговые коэффициенты корреляции

- Применяются для измерения связи между порядковыми или числовыми признаками, распределения которых отличаются от нормального.
- При подсчёте вместо исходных значений используются ранги.
- Наиболее популярные ранговые коэффициенты: Спирмена и Кендалла.
- Измеряют монотонную зависимость, поэтому способны оценивать нелинейные взаимосвязи.

Коэффициент Кендалла подсчитывается при большом числе связанных рангов, при небольшом числе связанных рангов подсчитывается коэффициент Спирмена.



Корреляции с бинарными признаками

Коэффициент корреляции	Признак 1	Признак 2
Точечно-бисериальная корреляция	Бинарный	Числовой
Фи (ϕ) коэффициент корреляции	Бинарный	Бинарный

Коэффициент частной корреляции

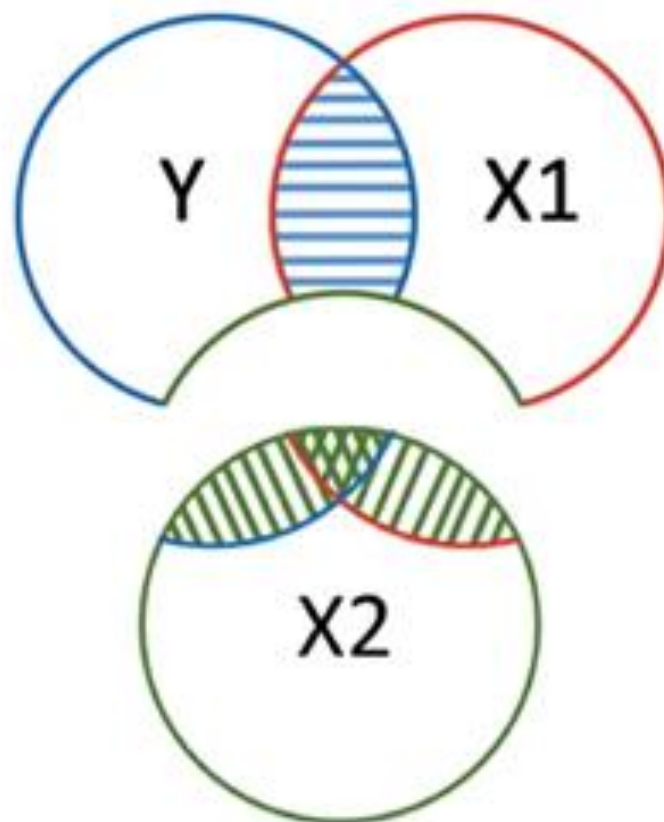
При изучении корреляции двух признаков необходимо учитывать возможное воздействие на них со стороны других признаков. Частная корреляция позволяет измерить связь между парой признаков после удаления линейных воздействий третьего.

Частный коэффициент корреляции (Partial correlation coefficient) – это мера зависимости между двумя признаками при фиксации эффектов одного или нескольких других признаков.

$$r_{xy.z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2} \sqrt{1 - r_{yz}^2}}$$

Частная корреляция

Partial Correlation between Y and X1
controlling for X2



Пример частной корреляции

Поль Брока в 1873 году обнаружил сильную связь между полом и размером мозга: у женщин, в среднем, мозг меньше, чем у мужчин. В то время это использовали как аргумент в пользу того, что женщины интеллектуально уступают мужчинам. Однако очевидная проблема заключалась в том, что эта связь не учитывала размер тела: у людей с большим телом, как правило, и мозг больше, независимо от уровня интеллекта.

В 1981 году Стивен Джей Гулд заново проанализировал данные Брока и показал, что сильная связь между полом и размером мозга исчезает, если учитывать размер тела.

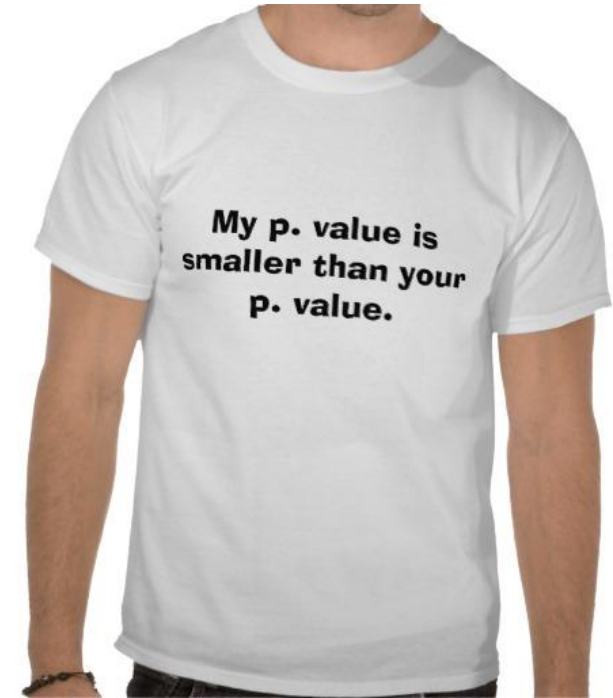
Пример частной корреляции из HR

При изучении влияния мотивации сотрудников на их производительность важно помнить, что возраст может оказывать влияние и на мотивацию, и на производительность. Простая корреляция между мотивацией и результатами работы не учитывает этот скрытый фактор, что может исказить выводы. Чтобы получить более точную и объективную оценку, используется частная корреляция — она позволяет измерить связь между мотивацией и производительностью с учётом влияния возраста. Таким образом, частная корреляция помогает выявить истинное влияние мотивации на эффективность работы, устраняя «шум» от возрастных различий.

P-value

P-value характеризует статистическую значимость результатов исследования и является инструментом для оценки надежности выводов о взаимосвязях между признаками. Чем меньше его значение, тем менее вероятно, что наблюдаемый эффект обусловлен случайными колебаниями. Обычно используется порог 0,05 (уровень значимости).

Полученные экстремально низкие значения p-value сигнализируют о высокой статистической значимости результатов.



	Rate	Log Frequency	Synonyms	Mutual-Information	Imageability	Arousal	Log Senses	Log AgeOfAcq
Rate	-	-.273**	.242*	-.281*	-.254*	.029	-.046	.255*
LogFrequency	-.273**	-	.381***	.221	-.208	-.046	.357***	-.482***
Synonyms	.242*	.381***	-	-.003	-.486***	.195	.592***	-.043
Mutual-Information	-.281*	.221	-.003	-	.506***	-.111	.031	-.367***
Imageability	-.254*	-.208	-.486***	.506***	-	-.071	-.369***	-.238
Arousal	.029	-.046	.195	-.111	-.071	-	.046	.097
LogSenses	-.046	.357***	.592***	.031	-.369***	.046	-	-.178
LogAgeOfAcq	.255*	-.482***	-.043	-.367***	-.238	.097	-.178	-

***: $p < .0001$

** : $p < .01$

* $p < .05$.

doi:10.1371/journal.pone.0147924.t002

Для большей гибкости в интерпретации результатов можно использовать разные пороговые значения уровня значимости.



Графики в Python

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



Факультет компьютерных наук

Машинное обучение

Москва 2026

Спасибо за внимание!