**Что мы вообще делали и зачем?**

У нас есть **несколько организаций**, про каждую из которых известны такие вещи:

* сколько там работников (X1),
* какая зарплата в сумме выплачена (X2),
* сколько у них контрагентов (X3),
* сколько контрактов они заключили (X4),
* и какая у них прибыль (Y).

**Цель** — понять, можно ли сказать, что **организации делятся на какие-то типы** (например, «богатые», «средние», «бедные») **и связаны ли с этими группами** такие признаки, как число работников, фонд зарплаты и т.д.

**Шаг 1: Кластеризация (группировка по прибыли)**

Мы **разделили все организации на 3 группы по прибыли**.

🔧 Как? С помощью KMeans — это метод, который находит «схожие» по значению Y организации и группирует их. Он сам вычисляет среднюю прибыль в каждой группе и распределяет данные по этим центрам.

🧠 Например:

* Группа 1 — бедные (прибыль ≈ 1 млн),
* Группа 2 — средние (прибыль ≈ 5 млн),
* Группа 3 — богатые (прибыль ≈ 10 млн).

Теперь каждая организация попала в **одну из 3 групп по прибыли**, которую мы назвали Группа\_Y.

**Шаг 2: Категоризация признаков X1–X4**

Признаки вроде числа работников (X1) могут быть слишком «детальными» — одно предприятие имеет 32 работника, другое 127. Так трудно сравнивать.

🔧 Поэтому мы разбили каждый признак на 3 группы:

* «мало»,
* «средне»,
* «много».

🧠 Пример:  
Если X1 (число работников) колеблется от 10 до 300, мы сделали так:

* 10–100 → «мало»
* 101–200 → «средне»
* 201–300 → «много»

Так мы получили X1\_cat, X2\_cat и т.д.

**Шаг 3: Таблицы сопряжённости**

Теперь мы спрашиваем: **есть ли связь между группой по прибыли и уровнем признака?**

Например:  
📊 «Среди организаций с **много** работников — больше ли тех, кто в **богатой группе**?»

Для этого мы построили **таблицы сопряженности** — это просто таблицы, в которых написано, сколько организаций попадает одновременно:

* в такую-то группу прибыли (Группа\_Y)
* и в такой-то уровень признака (X1\_cat).

**Шаг 4: Хи-квадрат тест**

Теперь мы используем **статистический тест хи-квадрат**, чтобы ответить на вопрос:

❓ Случайно ли так получилось, что среди «богатых» так много организаций с «много» работников?

Если **p-значение** маленькое (обычно < 0.05), это означает:  
→ **связь есть**, и не случайная.

**Что нам показали результаты?**

| **Признак** | **Chi-squared** | **p-значение** |
| --- | --- | --- |
| X1 | 9.08 | 0.059 |
| X2 | 9.06 | 0.060 |
| X3 | 5.82 | 0.213 |
| X4 | 11.38 | 0.023 ✅ |

* X4 — единственный признак, у которого **p < 0.05** → значит, количество контрактов **связано с прибылью**.
* X1 и X2 близко к границе → возможна связь, но не уверены.
* X3 — нет явной связи.