Метод k-средних. Пример

Итерационные методы кластерного анализа

Novosibirsk State University
*THE REAL SCIENCE

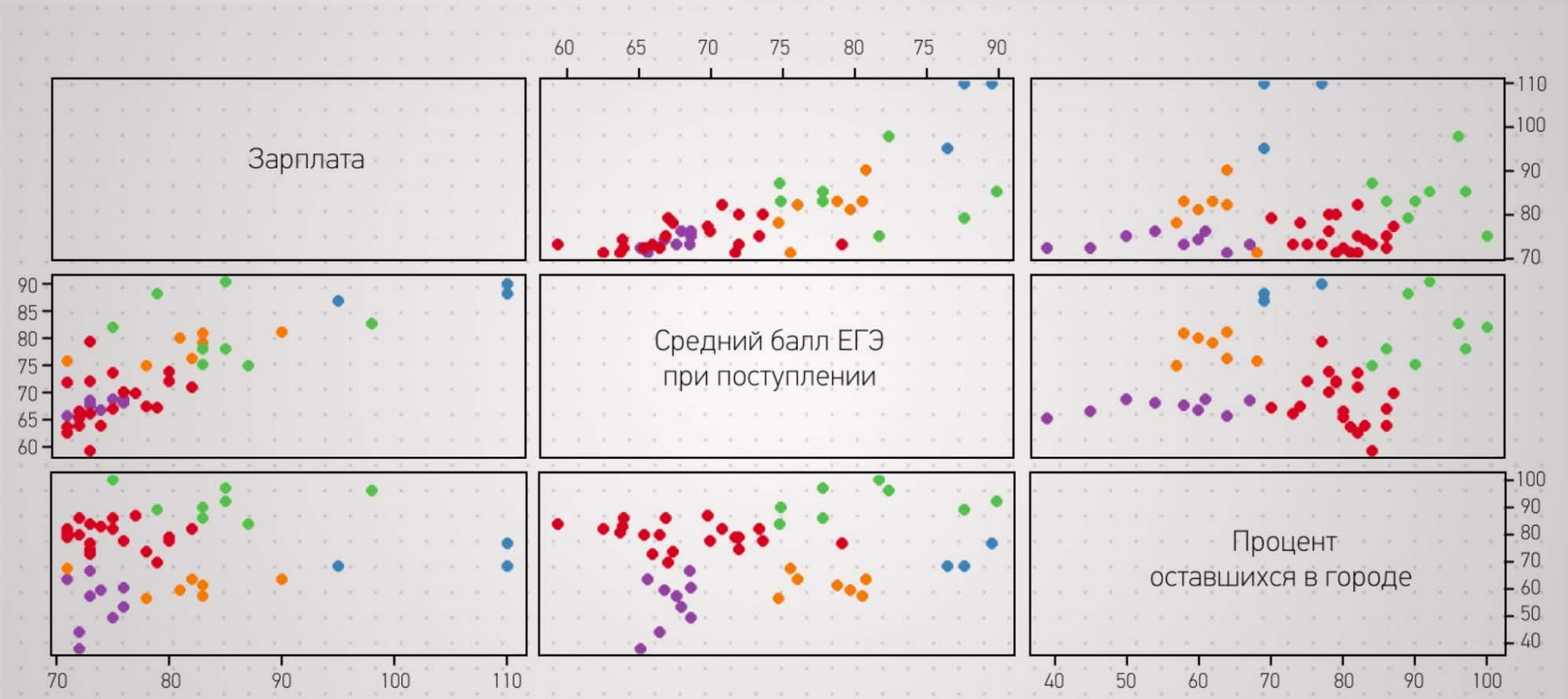
2GIS

Данные рейтинга вузов.

Рейтинг вузов по уровню зарплат выпускников, построенный порталом Superjob в 2017 году.

- Зарплата выпускников в Москве.
- Средний балл ЕГЭ при поступлении.
- Процент остающихся в городе обучения.

Результаты кластерного анализа. Метод k-средних



Алгоритм выбора оптимального количества кластеров

1. Выбираем количество кластеров.
2. Разбиваем данные на это количество кластеров.
3. Считаем меру качества.
4. Повторяем алгоритм для другого количества кластеров.
5. Смотрим на изменения меры качества.
6. Находим оптимальную границу.

Скорректированный R-квадрат

$$R^2_{adj} = 1 - \frac{WSS \cdot (n-1)}{TWSS \cdot (n-k)}$$

WSS — сумма внутрикластерных сумм квадратов расстояний

TWSS — общая сумма квадратов расстояний по выборке

n — объём выборки

k — количество кластеров