МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

ОТЧЕТ
по лабораторной работе №5
Технологии Аналитической обработки информации

Выполнил,
студент группы КЭ-403
Исхаков М.Р.
Проверил:
Гоглачев А.И.

Челябинск, 2025 г.

ФОРМУЛИРОВКА ЗАДАНИЯ

В рамках выполнения лабораторной работы необходимо:

1. Разработайте программу, которая выполняет кластеризацию заданного набора данных с помощью алгоритмов k-Means и k-Medoids. Параметрами программы являются набор данных и число кластеров. Программа должна выдавать координаты точек и назначенные им кластера, а также значение ошибки кластеризации.

2. Проведите эксперименты на наборе данных **customers** (сведения о клиентах банка: скачать zip-архив с данными в формате CSV и описанием).

3. Выполните визуализацию полученных результатов в следующем виде:

– точечный график, на котором цвет точки отражает принадлежность кластеру;

– зависимость ошибки кластеризации от параметра *k*.

4. Доработайте программу, добавив в список ее параметров долю зашумленных объектов набора. Дополнительно к ранее реализованным функциям программа должна вносить шум в набор данных: случайным образом изменить заданную долю объектов набора (изменение может заключаться в добавлении/вычитании случайного числа к/из одной/нескольких координат объекта).

5. Проведите эксперименты на ранее выбранных наборах данных, варьируя долю зашумленных объектов (1%, 3%, 5%, 10%) и используя различные значения параметра (из интервала 3..9).

6. Выполните визуализацию полученных результатов указанным выше способом.

Исходные коды для задания представлены в репозитории: https://github.com/SMarkls/analasys

РИСУНКИ С РЕЗУЛЬТАТАМИ ВИЗУАЛИЗАЦИИ

В ходе выполнения кластеризации указанного набора данных с помощью алгоритмов k-Means и k-Medoids были получены следующие рисунки.

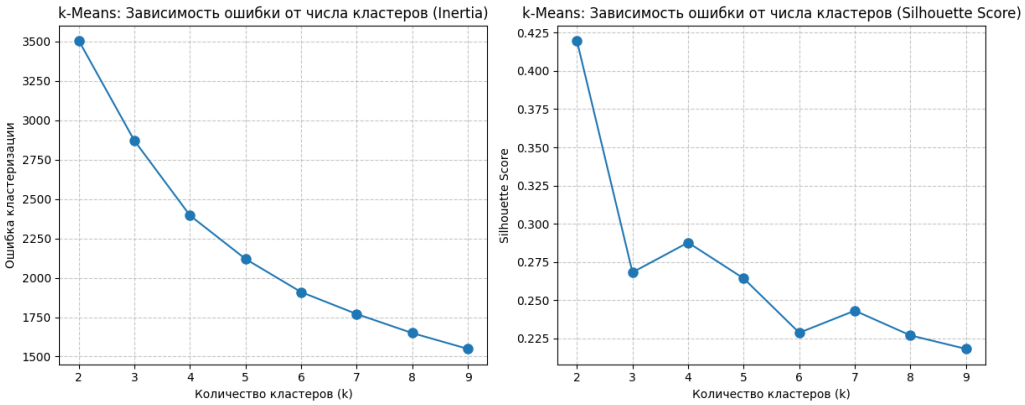

Рисунок 1 - Зависимости ошибки от параметров для метода k-Means

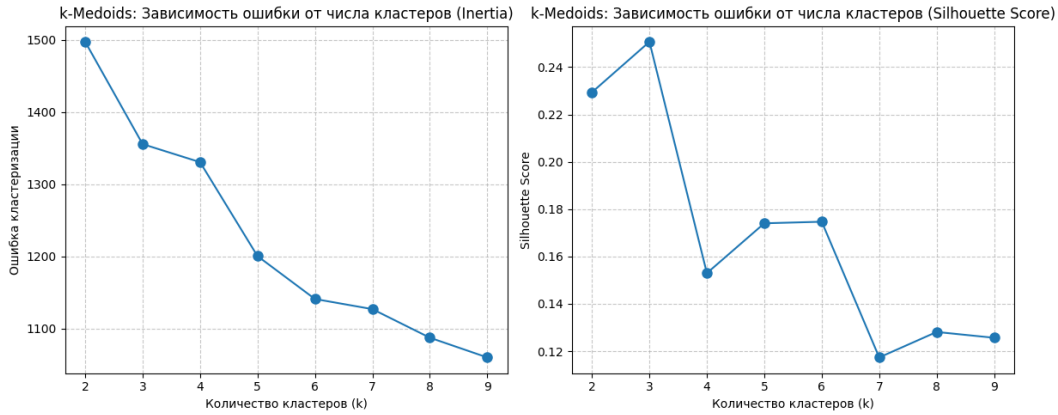

Рисунок 2 - Зависимости ошибки от параметров для метода k-Medoids

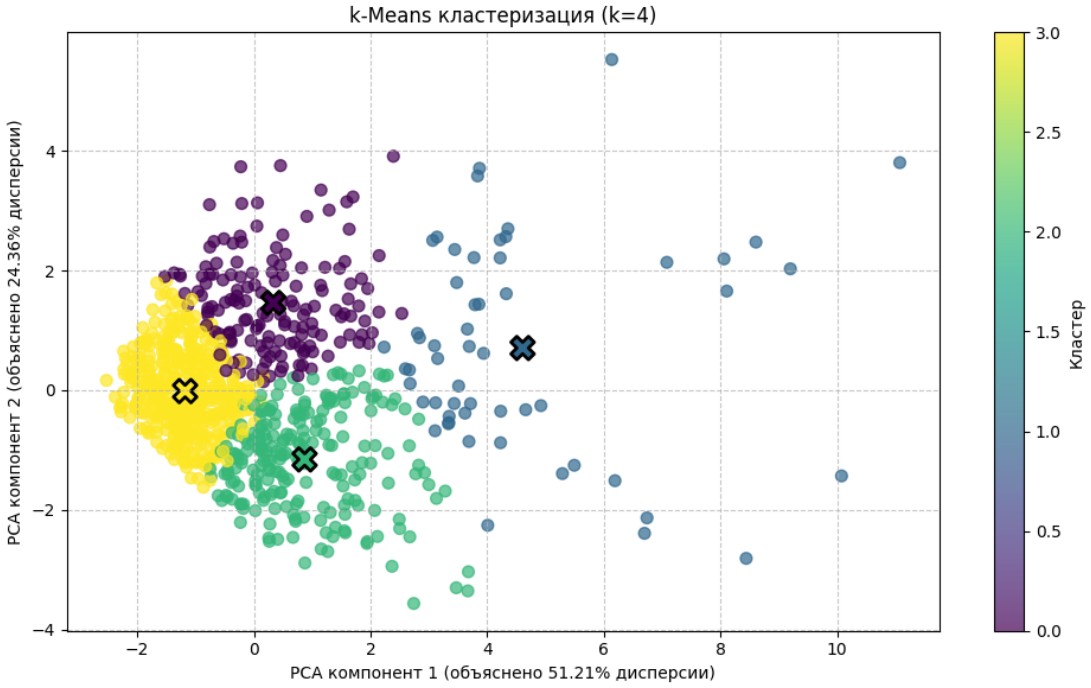

Рисунок 3 - Результат кластеризации методом k-Means при k = 4

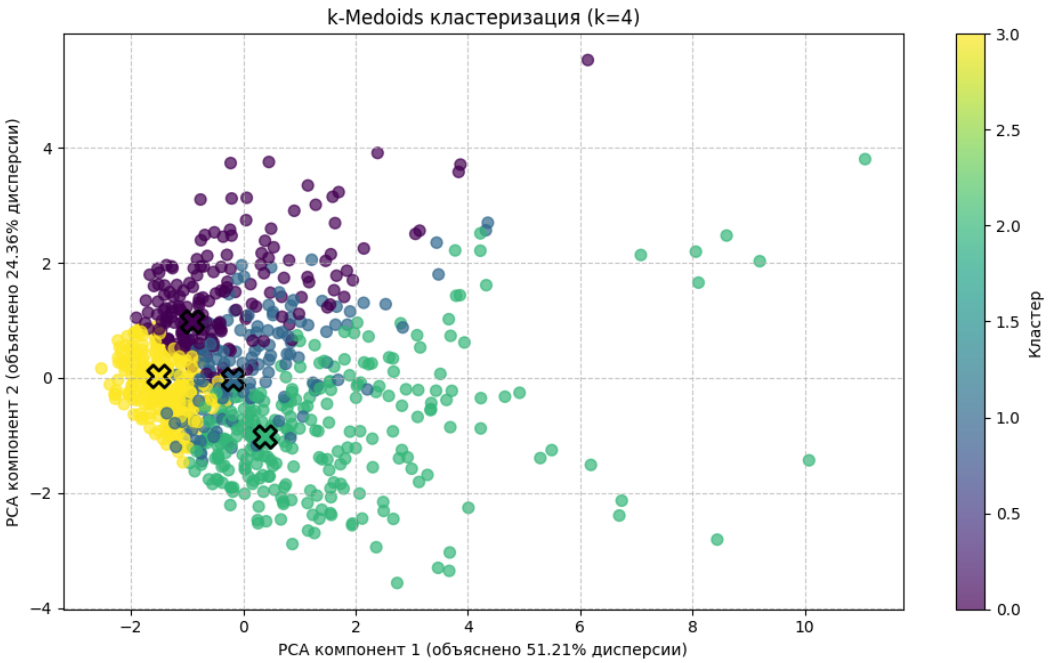

Рисунок 4 - Результат кластеризации методом k-Medoids при k = 4

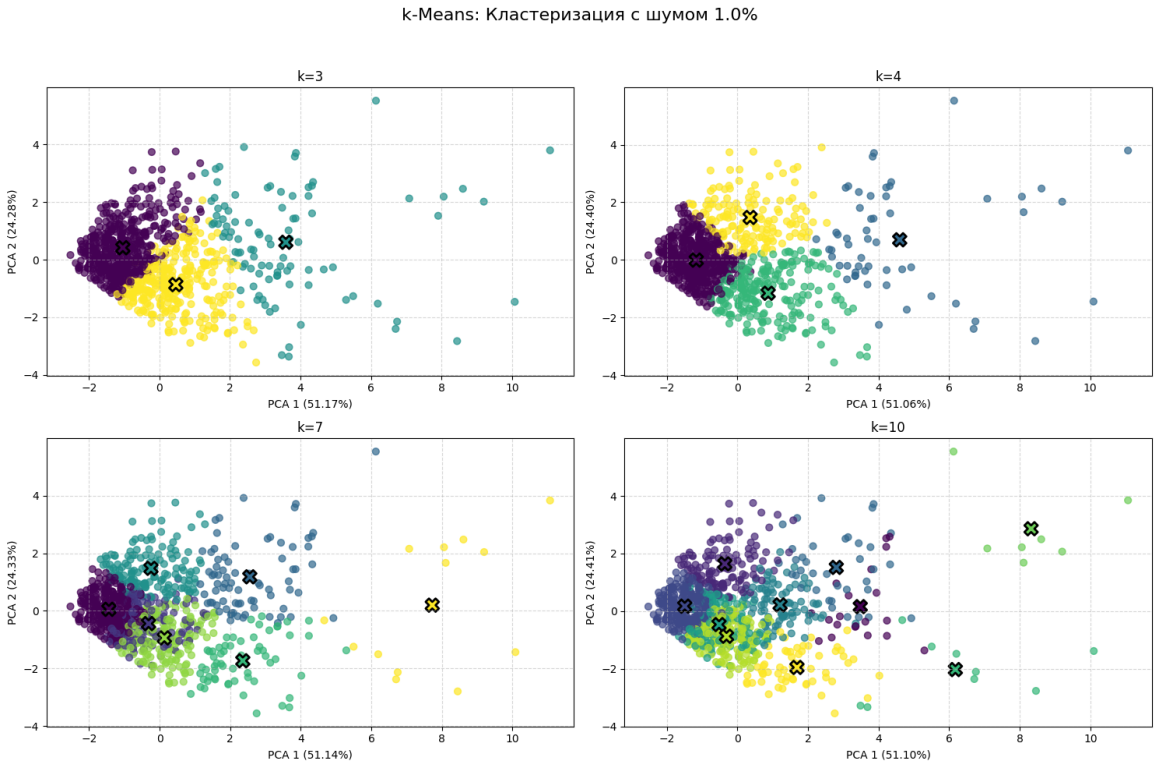

Рисунок 5 - Результат кластеризации методом k-Means с шумом 1%

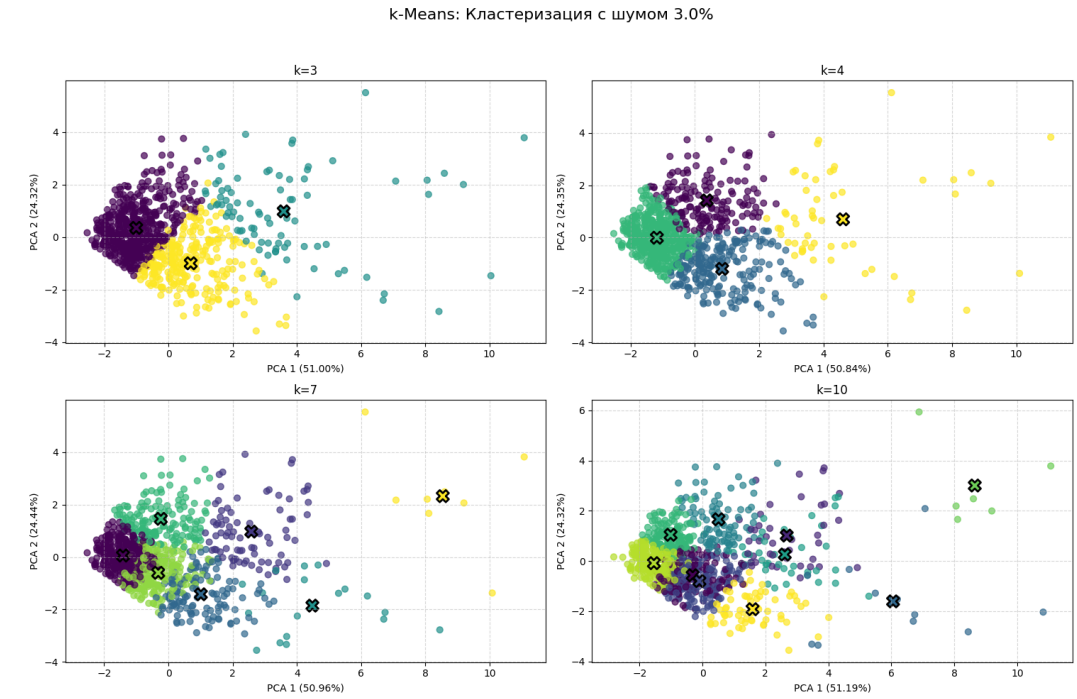

Рисунок 6 - Результат кластеризации методом k-Means с шумом 3%

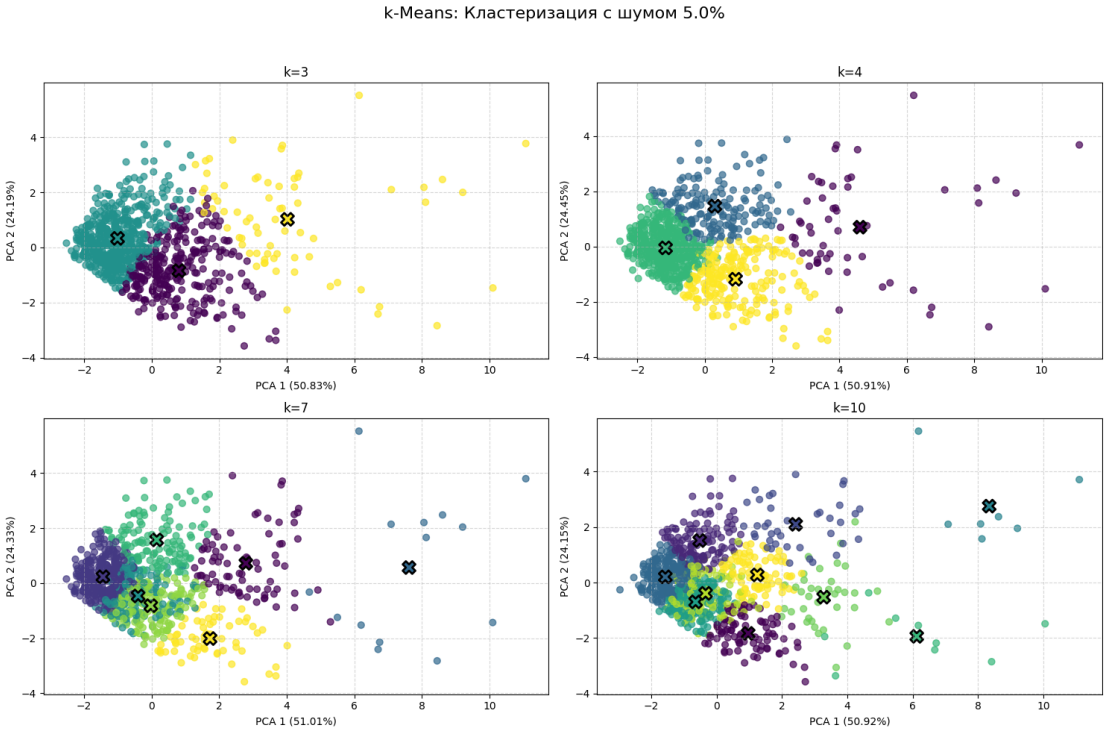

Рисунок 7 - Результат кластеризации методом k-Means с шумом 5%

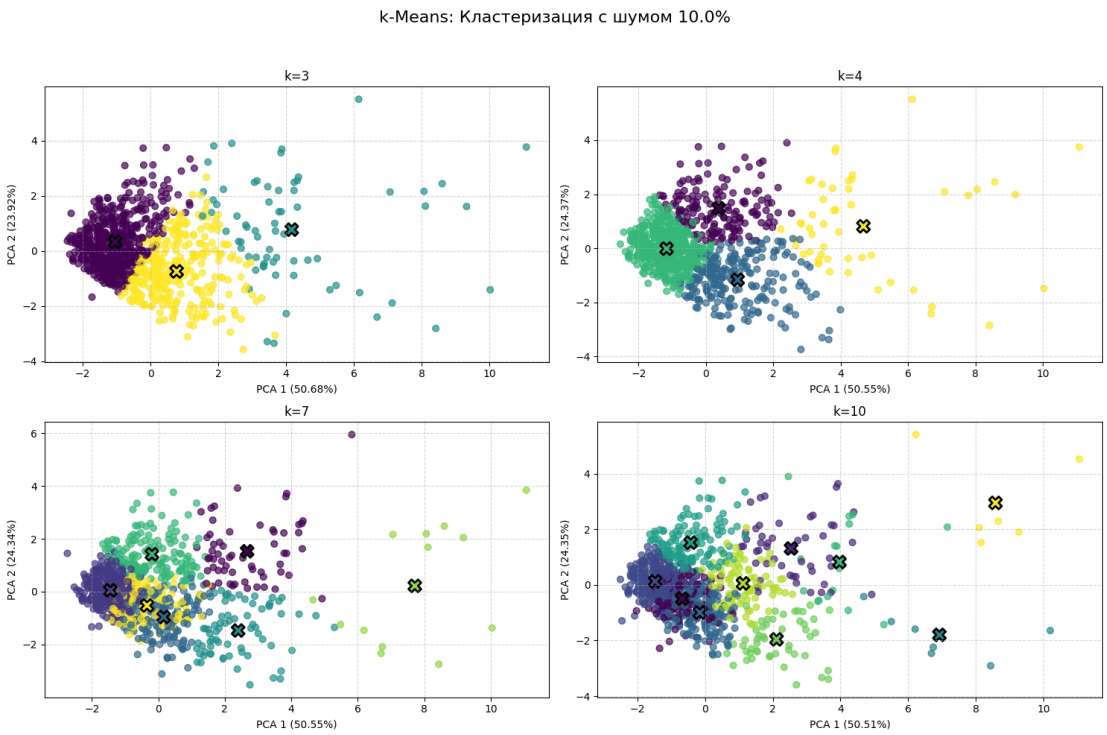

Рисунок 8 - Результат кластеризации методом k-Means с шумом 10%

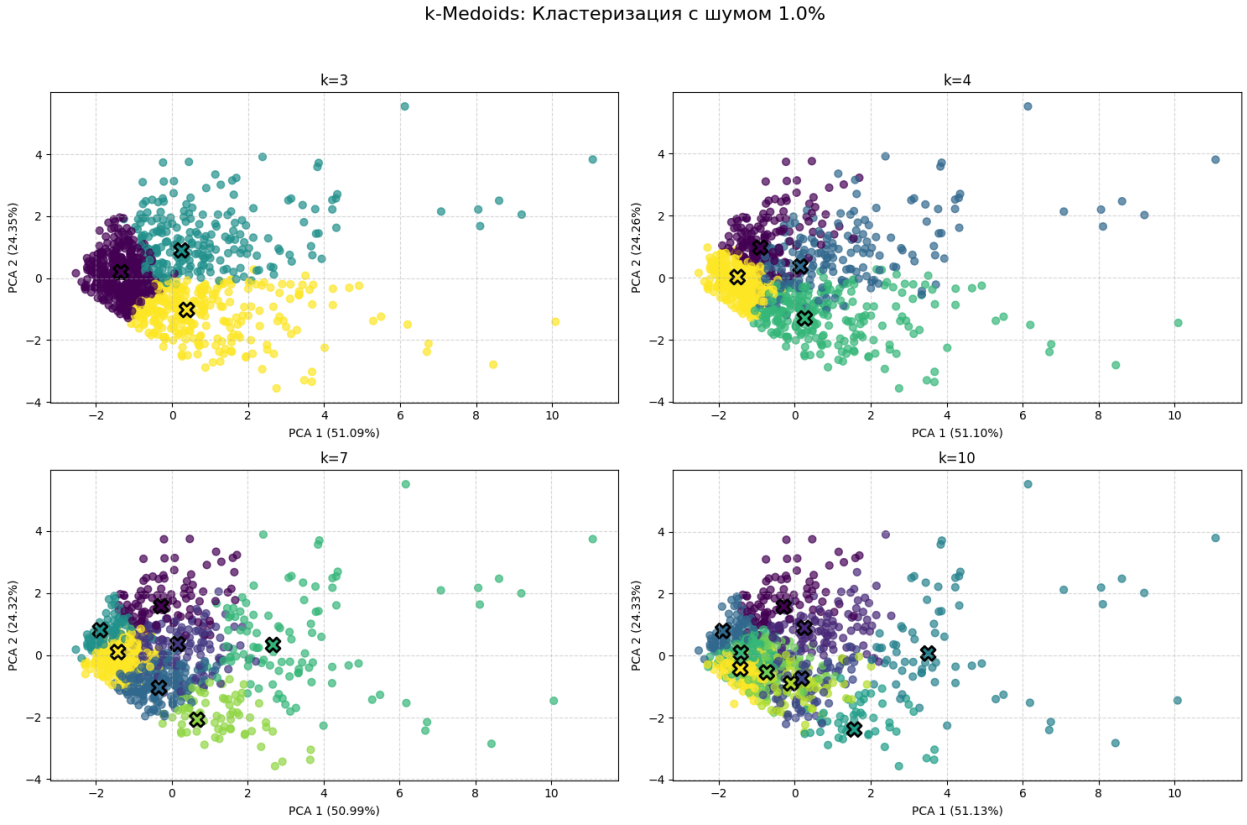

Рисунок 9 - Результат кластеризации методом k-Medoids с шумом 1%

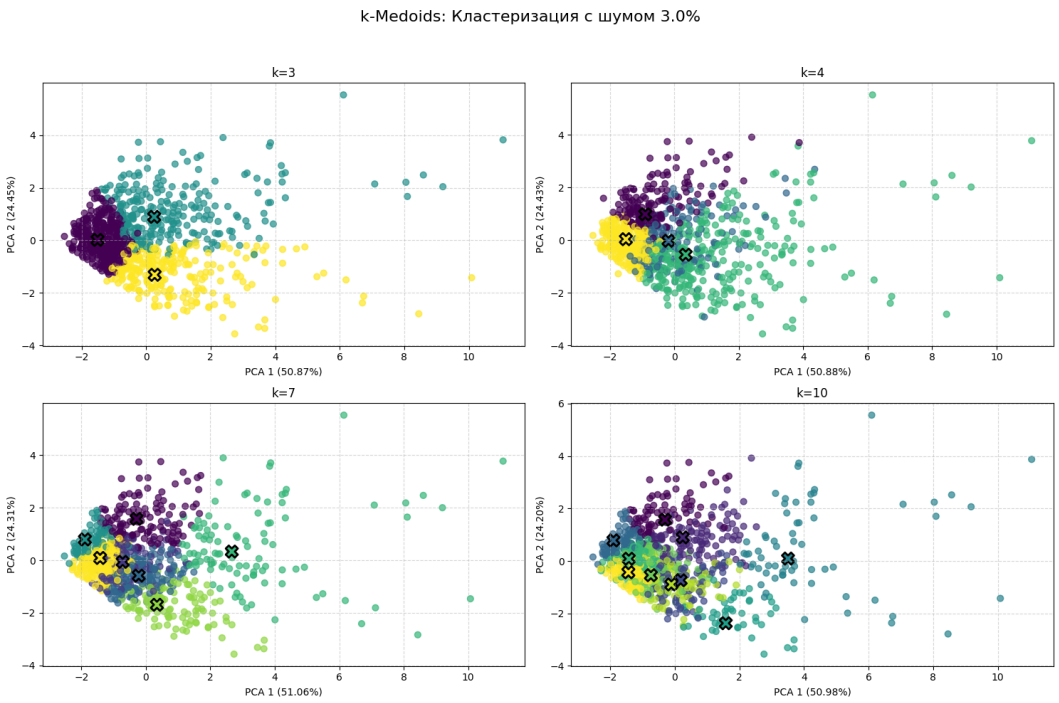

Рисунок 10 - Результат кластеризации с шумом 3%

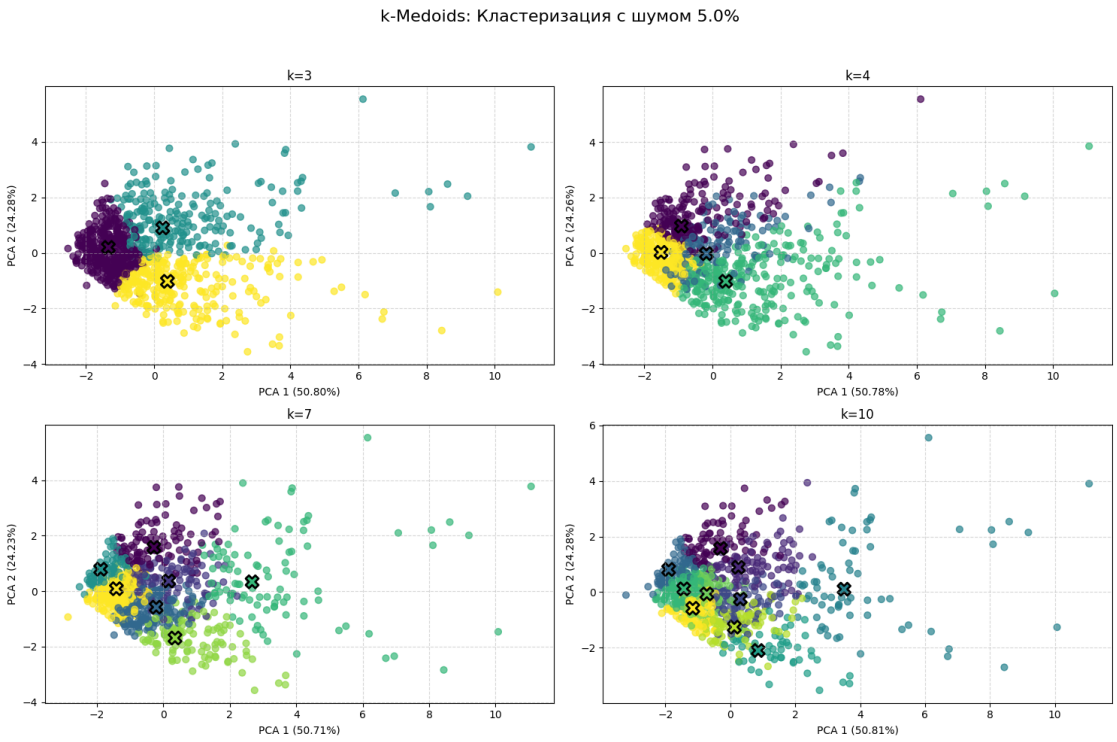

Рисунок 11 - Результат кластеризации с шумом 5%

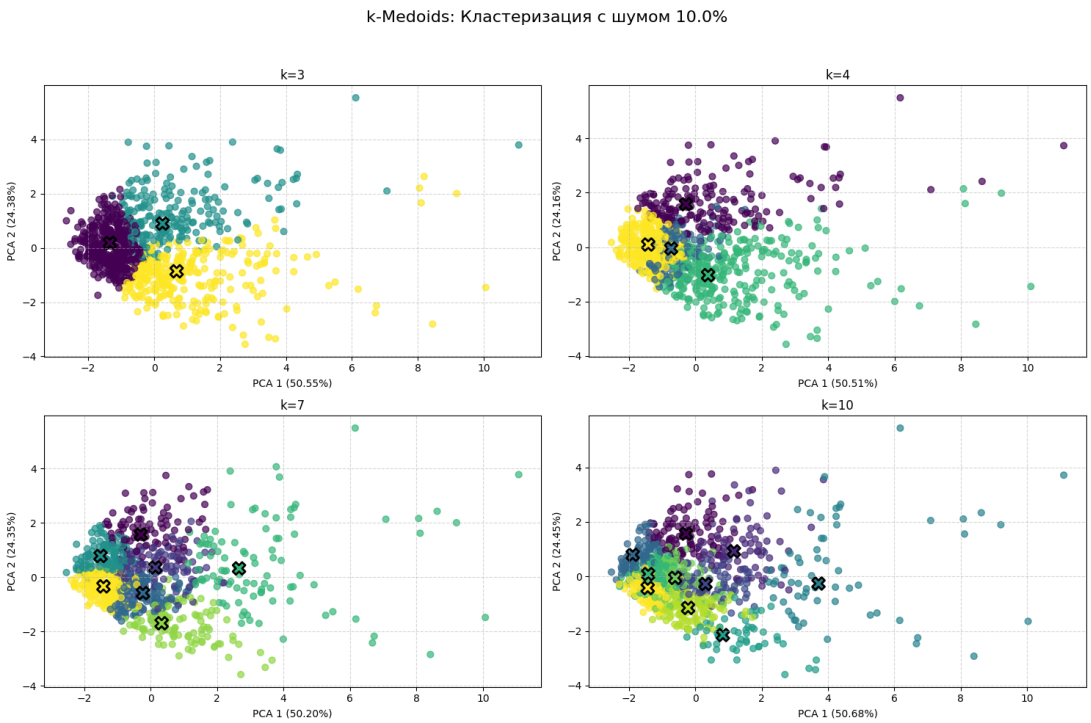

Рисунок 12 - Результат кластеризации с шумом 10%

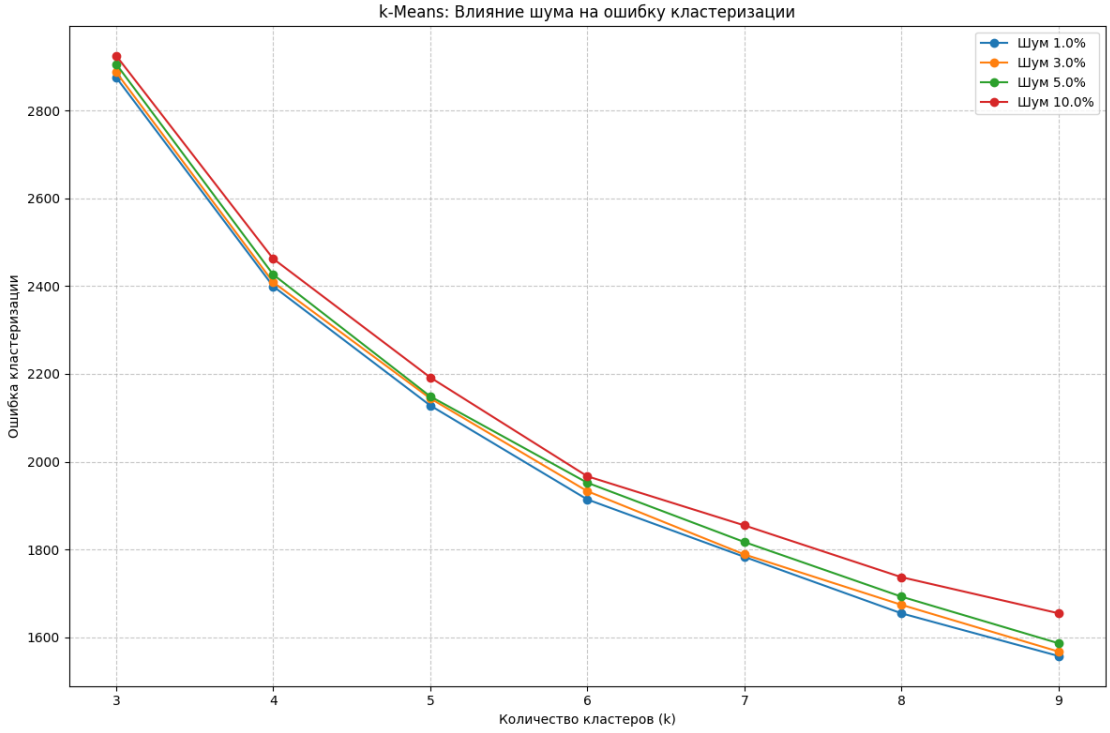

Рисунок 13 - Зависимость ошибки кластеризации от количества кластеров при различном уровне шума метода k-Means

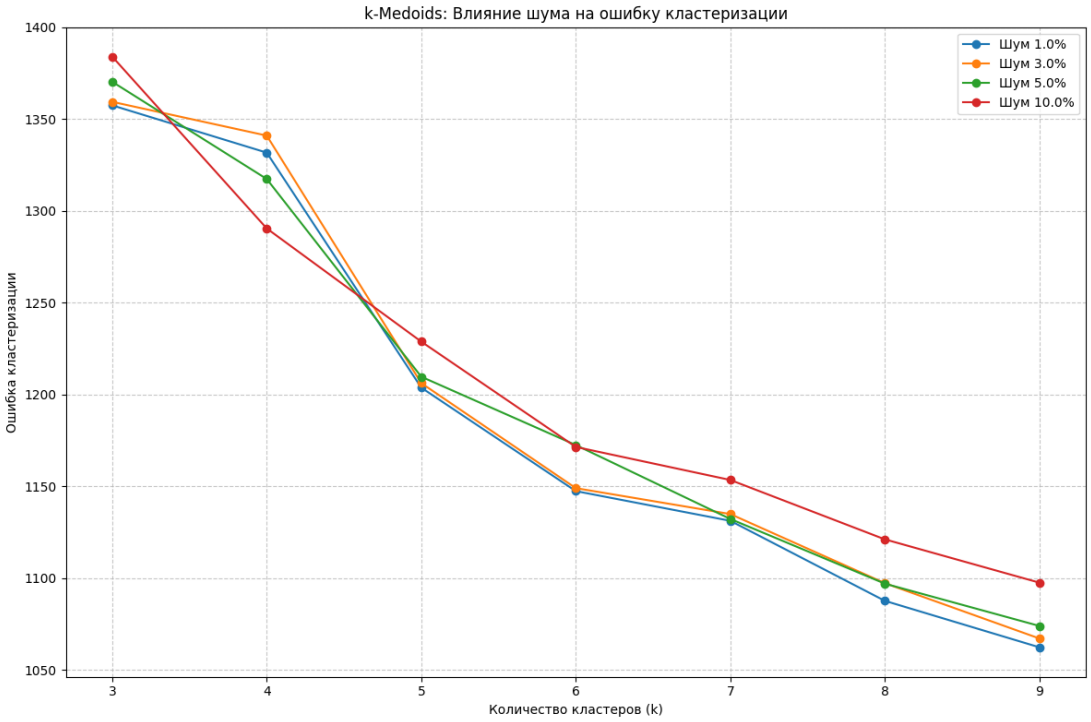

Рисунок 14 - Зависимость ошибки кластеризации метода k-Medoids от количества кластеров при различном уровне шума

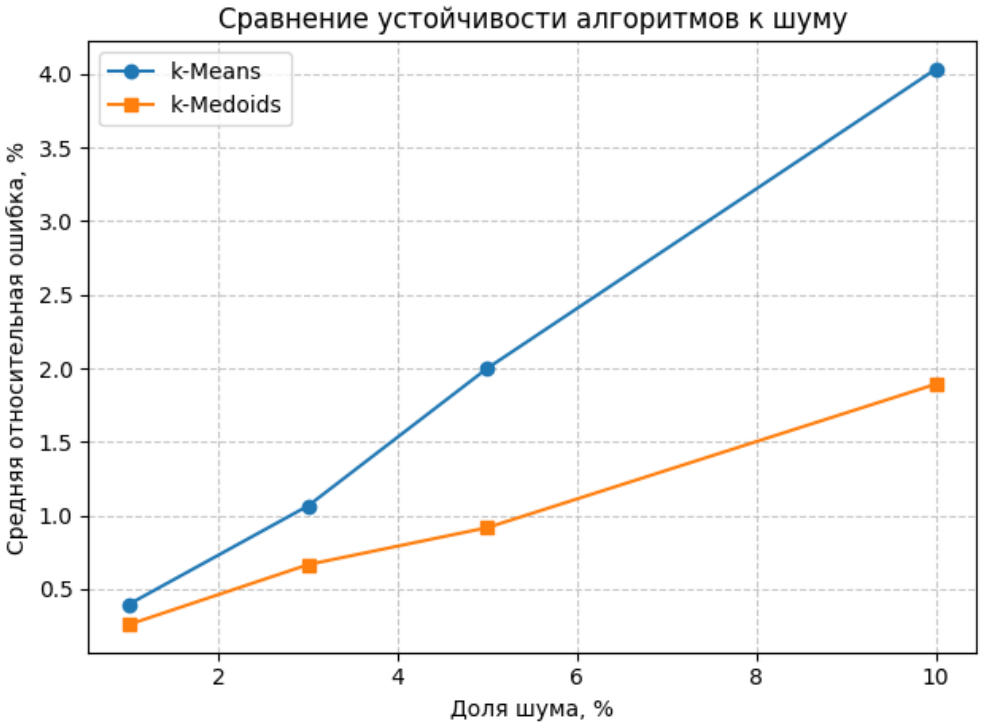

Рисунок 15 - Сравненеи устойчивости алгоритмов к шуму

ВЫВОДЫ ИЗ ВИЗУАЛИЗАЦИИ

Из рисунка 1 видно, что ошибка кластеризации монотонно снижается с увеличением числа кластеров. Silhouette Score показывает наивысшее значение при k = 2, затем падает и имеет локальные пики при k = 4 и k = 7, что может указывать на потенциально хорошие варианты разбиения данных.

Из рисунка 2 видно, что ошибка так же монотонно уменьшается, но локальный пик получается при значении k = 3, что может говорить, что это значение может быть оптимальным выбором для k-Medoids.

Рисунки 3 и 4 внешне похожи друг на друга, положение центроидов и распределение точек немного отличается. Сами кластеры отличаются по размеру и компактности.

Исходя из рисунков 5-12 можно сделать выводы, что шум вносит свои коррективы в результат кластеризации, но он едва различим.

На рисунке 13 видна ожидаемая зависимость. Уровень ошибки тем выше, чем выше уровень шума и тем меньше, чем больше уровень кластеризации.

Из рисунка 15 можно лишь сделать вывод, что метод k-Medioids чуть лучше устойчив к шуму.