САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО

Отчет по лабораторной работе №7
по курсу «Современные инструменты анализа данных»
Тема: **Кластеризация и использование метода DBScan**

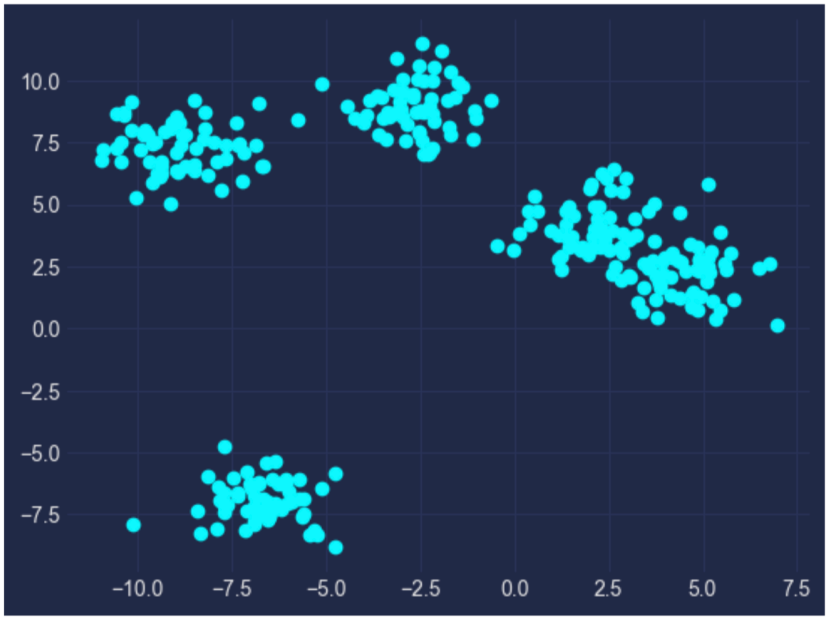
Выполнили:
-

Проверила:
-

Санкт-Петербург
2023 г.

Задание 1. Работа с синтетическими данными

Мы сгенерировали синтетические данные в количестве 300 случайных точек вокруг 4-ех центроид



Синтетические данные

Далее мы тестировали метод K-Means для выбора оптимального числа кластеров. Для этого мы построили график «Локтя», попробовав циклом различное число кластеров для разбиения от 2 до 9

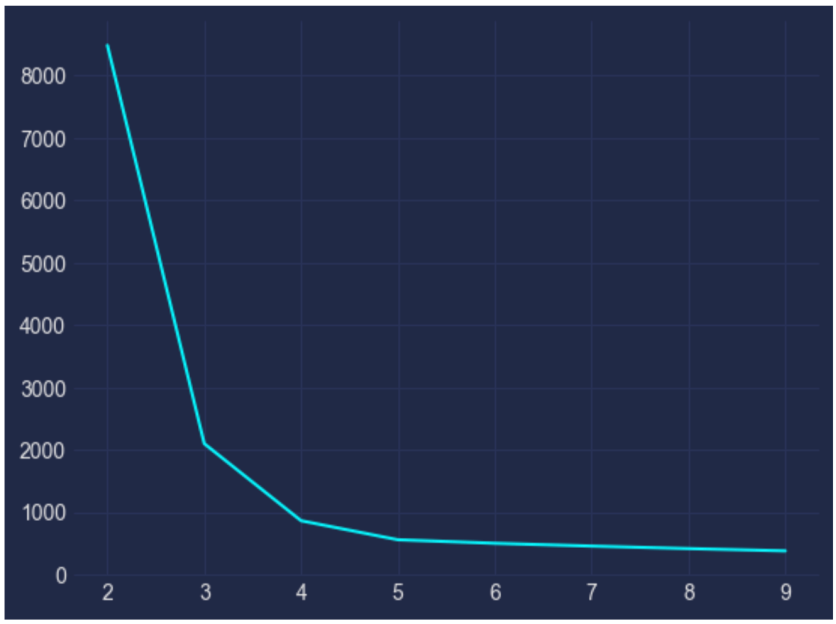
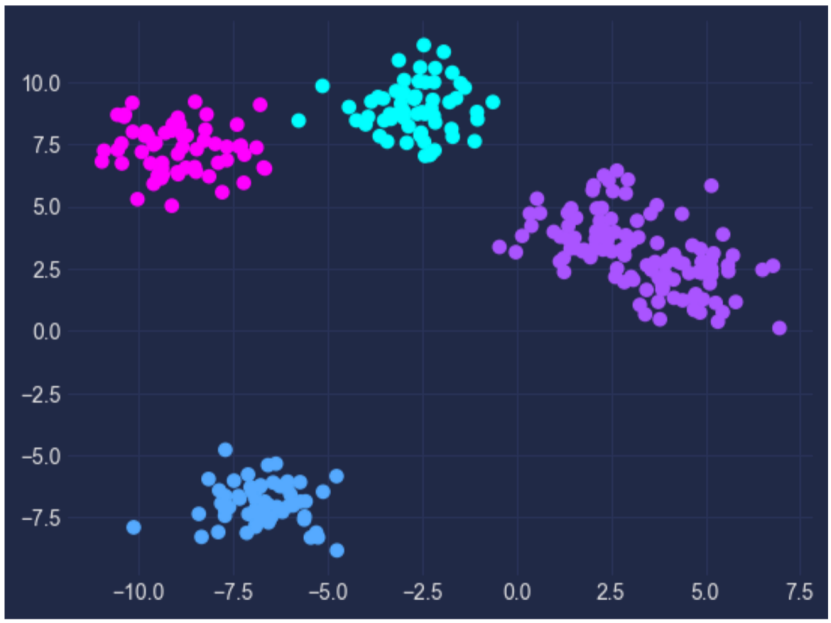


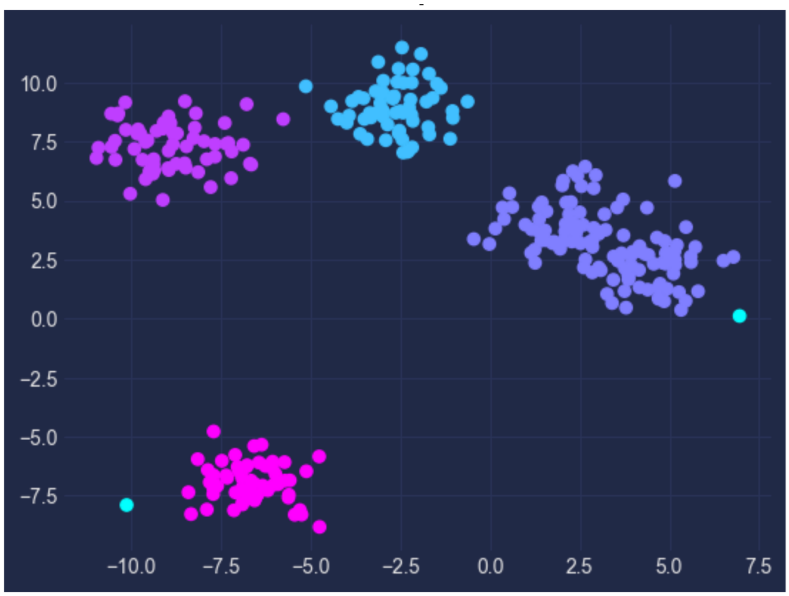
График «Локтя»

Для оптимального числа кластеров необходимо найти точку, после которой график упадает с наименьшей скоростью, но мы решили остановиться на 4-ех кластерах



K-Means на 4 кластера

После мы решали ту же задачу, но уже с использованием DBScan. Алгоритм за счет внутренних эвристик самостоятельно находит оптимальное число кластеров, пользователю же нужно указать только параметр эпсилон, который отвечает за максимальное расстояние между двумя соседями в рамках одного кластера



DBScan на 4 кластера

Можно отметить, что DBScan находит аналогичное количество кластеров, но при этом отмечает некоторые «выбросы» в виде точек, сильно удаленных от центроид

Задание 2. Работа с настоящими данными

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Пример загруженного датасет с Kaggle

С Kaggle загрузили данные и провели первичную предобработку, убрав колонку с индексами и закодировав категориальный признак пола. Для визуализации признаков с помощью T-SNE сократили количество размерностей до двух, сначала нормализовав их с помощью StandartScaler. После этого с единичным шагом начали проверять разбиение на различное количество кластеров, отобразив сравнение на графике «Локтя»

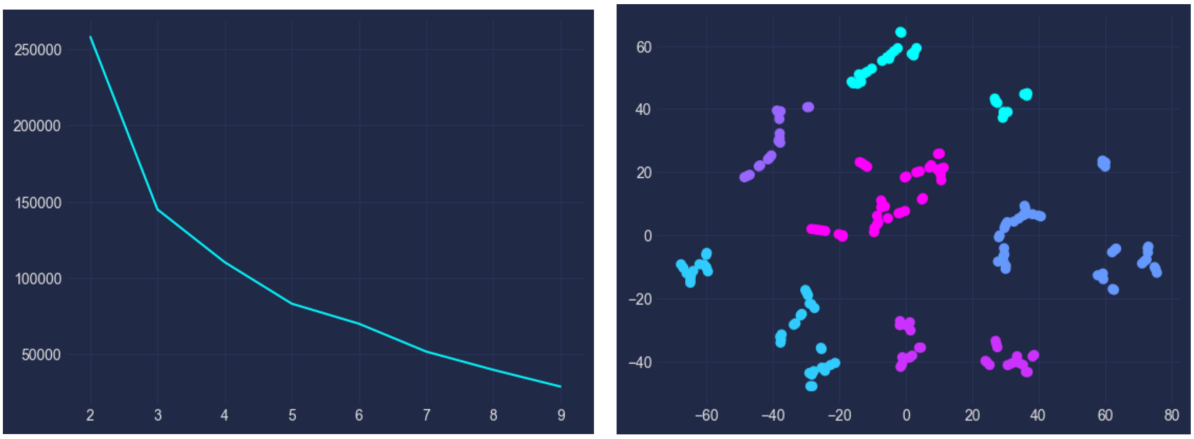
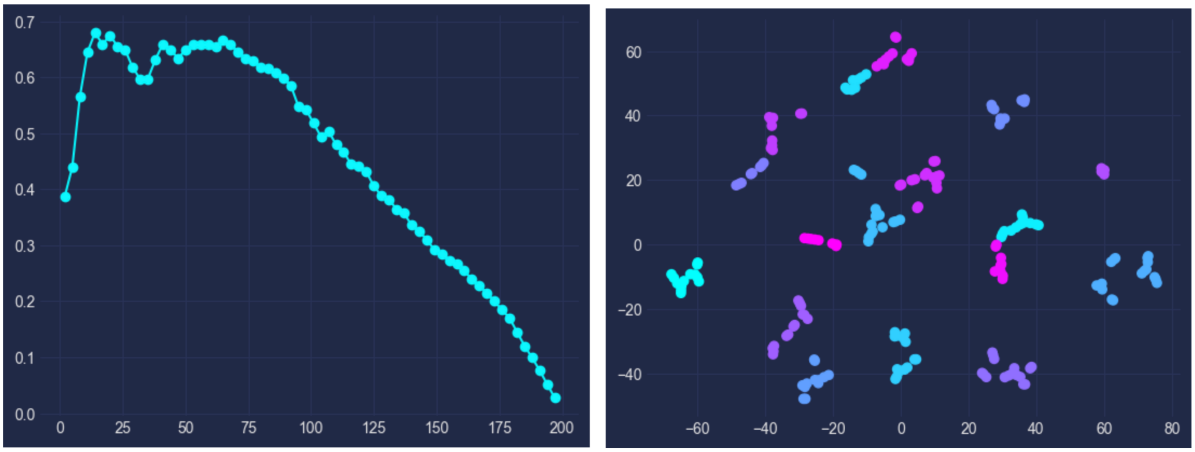


График «Локтя» и соответствующе разбиение на 6 кластеров

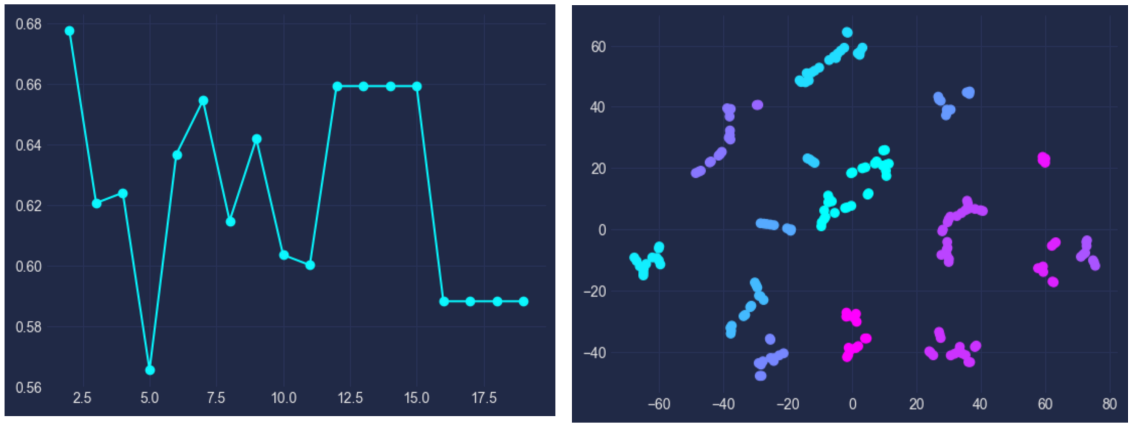
По графику решили остановиться на 6 кластерах, хотя есть точная тенденция к убыванию и возможно стоило попробовать разбиение на другое количество кластеров

Более того мы решили сравнить разбиение кластеров, используя метрику Silhouette, оценивающую качество разбиение кластеров, этот метод обеспечивает краткое графическое представление того, насколько хорошо был классифицирован каждый объект. Чем выше метрика, тем лучше разбиение, т.к. мы проходились с шагом 3 до 2 кластеров до размера всей выборки, по графику делаем вывод, что оптимальное разбиение лежит в районе 17 кластеров, на что также может косвенно указывать предыдущий график «Локтя», который с каждым разбиением на кластеры только продолжает убывать



Silhouette score и соответствующе разбиение на 17 кластеров

Последний шаг мы повторили и для DBScan алгоритма, только используя уже перебор эпсилон значения с шагом один от 1 до 20, получив самое оптимальное 8, исходя из наивысшего показателя метрики Silhouette



Silhouette score и разбиение с наилучшим eps=8 (16 кластеров)

Можно сделать вывод, что при грамотном переборе гиперпараметров, наилучший результат и визуально, и по сторонним метрикам показывается DBScan алгоритм