**Тестовое задание: кластеризация клиентов брокерской компании**

**Задача:** Кластеризовать клиентов брокерской компании для определения сегментов пользователей с целью дальнейшего таргетинга маркетинговых кампаний.

**Данные:** Вам предоставлен датасет df\_test\_middle.csv с информацией о клиентах брокерской компании.

Датасет содержит следующие столбцы:

* id: уникальный идентификатор клиента
* birth\_date: дата рождения клиента
* gender: пол клиента (Male/Female)
* kval\_flag: флаг квалифицированного инвестора (0 или 1)
* premial\_flag: флаг премиального клиента (0 или 1)
* account\_sum: сумма на счету клиента
* remains\_instrument\_cnt: количество оставшихся инструментов
* remains\_sec: остаток ценных бумаг
* remains\_bonds: остаток облигаций
* remains\_der: остаток деривативов
* remains\_curr: остаток валюты
* input\_cnt: количество пополнений на счет клиента
* input\_sum: сумма пополнений на счет клиента
* yields\_365: доходность за 365 дней
* yields\_90: доходность за 90 дней
* deals\_sum\_month: сумма сделок за месяц
* trancnt\_month: количество транзакций за месяц
* deals\_sec: были ли сделки с ценными бумагами (0 или 1)
* deals\_bonds: были ли сделки с облигациями (0 или 1)
* deals\_der: были ли сделки с деривативами (0 или 1)
* deals\_curr: были ли сделки с валютой (0 или 1)

**Требования:**

1. Провести исследовательский анализ данных (EDA).
2. Очистить и обработать данные.
3. Применить различные алгоритмы кластеризации для сегментации клиентов.
4. Оценить качество кластеризации и выбрать наилучший метод.
5. Визуализировать полученные кластеры и интерпретировать результаты.
6. Сохранить результаты в файл clusters.csv, который должен содержать два столбца: id и cluster\_label.

**Описание работы:**

1. **Исследовательский анализ данных (EDA):**
   * Рассмотреть основные статистики и распределения переменных.
   * Проанализировать наличие пропущенных значений и выбросов.
   * Визуализировать данные: распределения, матрица корреляций и парные графики.
2. **Очистка и предобработка данных:**
   * Обработать пропущенные значения и выбросы.
   * Преобразовать категориальные переменные (например, пол клиента) в числовые.
   * Нормализовать или стандартизировать числовые переменные, если это необходимо.
3. **Кластеризация:**
   * Применить несколько алгоритмов кластеризации, например, K-средних (K-Means), DBSCAN, иерархическую кластеризацию.
   * Оценить качество кластеров с использованием подходящих метрик (например, силуэтный показатель, инерция, индексы Дэвиса-Болдина).
   * Определить оптимальное количество кластеров для каждого метода (например, используя метод локтя для K-средних).
4. **Визуализация и интерпретация результатов:**
   * Визуализировать кластеры с использованием 2D или 3D графиков (например, метод главных компонент (PCA) или t-SNE).
   * Проанализировать характеристики каждого кластера и дать рекомендации по каждому сегменту.
5. **Сохранение результатов:**
   * Сохранить результаты кластеризации в файл clusters.csv.

Также следует представить отчет в формате Jupyter Notebook, включающий:

* Описание и результаты EDA.
* Шаги и обоснования для предобработки данных.
* Описание и результаты кластеризации, включая визуализации.
* Итоговые выводы и рекомендации по каждому сегменту.

**Критерии оценки:**

1. Полнота и корректность анализа данных (EDA).
2. Эффективность и обоснованность методов предобработки данных.
3. Качество и интерпретация кластеризации.
4. Визуализация и интерпретация полученных кластеров.
5. Понятность и структура предоставленного отчета.