Отчёт по заданию 1 «Кластерный анализ»

Выполнил студент 4 курса 7.1 группы Путин Павел

Вариант 8 «Сегментация покупателей»

1 Описание данных:

Для подготовки рекламной компании необходимо провести сегментацию постоянных покупателей. Данные предоставлены российской компанией, продающей кожгалантерею (Эдминс). Рекламный бюджет компании ограничен.

Описание переменных:

— Пол

— Возраст

— Предпочитаемые телеканалы

— Читаемая пресса

Имеется 184 наблюдения, пропусков нет.

2 Решение

В качестве теоретической основы использовалось учебное пособие Демидовой Л.А. «Кластерный анализ». В соответствии с рекомендациями из пособия использовались библиотеки scipy для иерархической кластеризации и построения дендрограммы и sklearn для реализации кластеризации методом k-средних и многомерного шкалирования.

2.1 Иерархическая кластеризация

Иерархическая кластеризация не вызвала трудностей и была решена стандартным методом, представленным в книге (см. Рисунок 1):

```
def make_hierarciacal_clustering(df):
    scaler = preprocessing.MinMaxScaler().fit(df.to_numpy())
    scaled_data = scaler.transform(df.to_numpy())
    return linkage(scaled_data,
                   method=CLUSTERS_DISTANCE_TYPE,
                   metric=OBJECTS_DISTANCE_TYPE)
```

Рисунок 1 - Метод иерархической кластеризации

На вход метод принимает данные в виде объекта pandas.DataFrame и возвращает матрицу расстояний между объектами. CLUSTERS_DISTANCE_TYPE имеет значение ward – метод расчёта расстояний Уорда. Выбор производился между методом Уорда и методом медианы. И хотя по одному из признаков, полу, наблюдалось явное преобладание женщин (155 к 29), по остальным признакам значения не были столь явно разделены, из-за чего предпочтение было отдано методу Уорда. Метрика расстояния была выбрана евклидова (Euclidean), потому что метод Уорда предполагает её использование.

Результат иерархической кластеризации можно увидеть на дендрограмме (см. Рисунок 2):

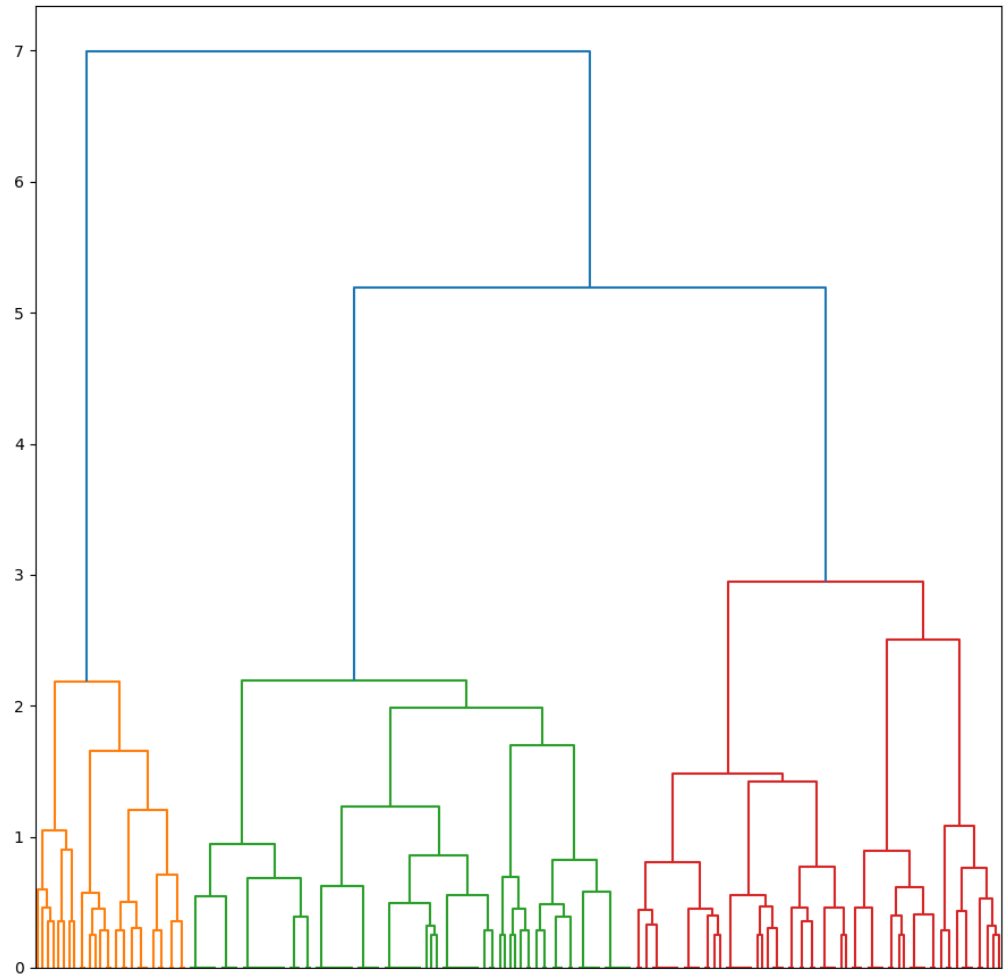

Рисунок 2 - Дендрограмма кластеров покупателей

На дендрограмме отчётливо выделяются 3 кластера (оранжевый, зелёный и красный). Их анализ будет произведён далее.

2.2 Кластеризация методом k-средних

При выполнение кластеризации методом k-средних возникла проблема определения количества кластеров на основе диаграммы локтя: присутствовало два потенциальных изгиба (см. Рисунок 3):

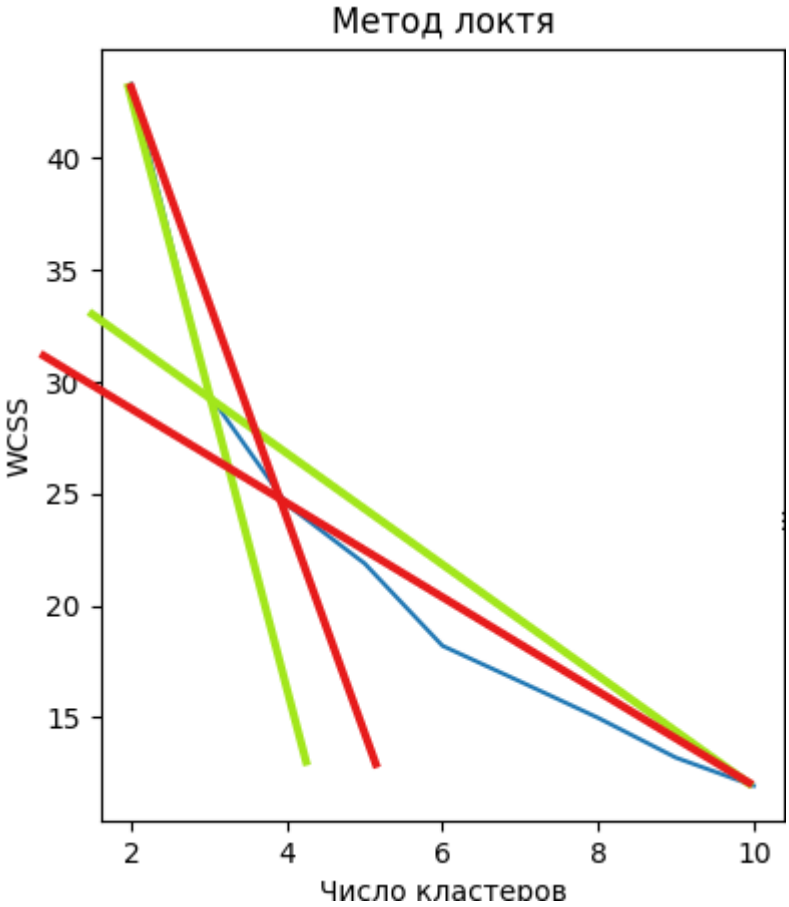

Рисунок 3 - Диаграмма локтя. Зелёный и красные изгибы почти одинаковы

Поэтому было принято решение построить диаграмму силуэтов кластеров, из которой видно, что индекс силуэта при трёх кластерах выше, чем при 4 (см.). На этой основе было принято решение о разбиении на 4 кластера.

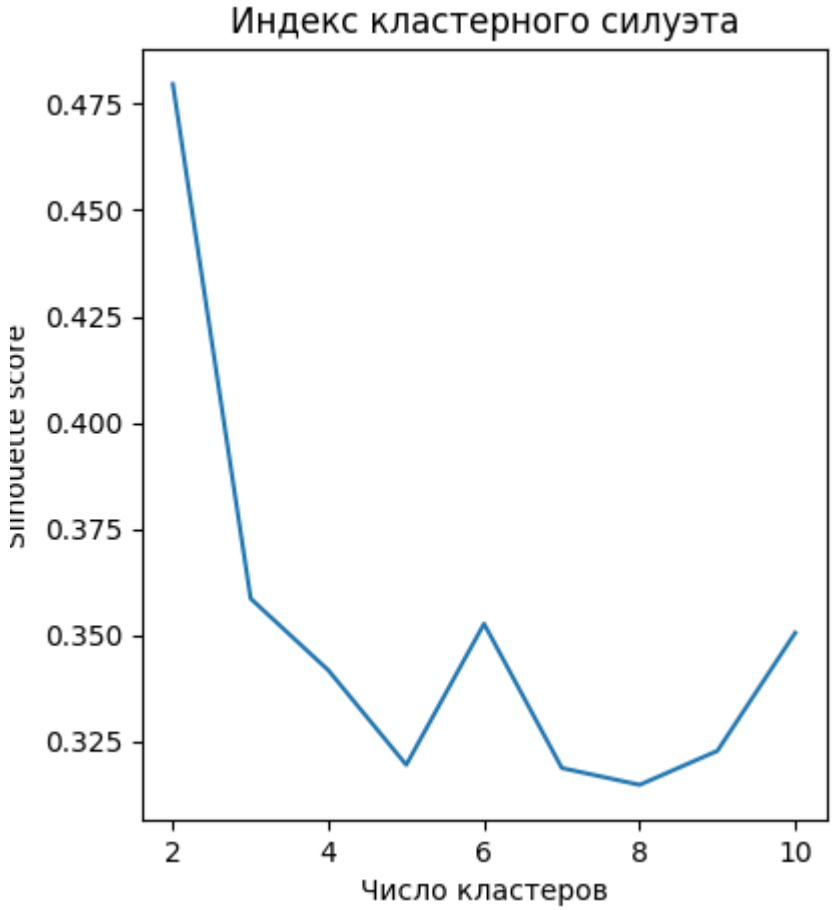

Рисунок 4 - График индекса кластерного силуэта

На основе кода, приведённого в пособии Демидовой Л.А., был написан метод make_k_means_clustering, принимающий на вход данные в виде объекта pandas.DataFrame и возвращающего масштабированные данные, значения индекса локтя и силуэта, а также кластеризацию на 3 категории (см. Рисунок 5).

```python
def make_k_means_clustering(df):
    scaler = preprocessing.MinMaxScaler().fit(df.to_numpy())
    scaled_data = scaler.transform(df.to_numpy())

    # Сумма квадратов внутрикластерных расстойний WCSS
    elbow = []
    silhouette = []
    for i in range(2, 11):
      kmeans = KMeans(n_clusters=i, init='k-means++',
                      max_iter=300, n_init=10, random_state=0)
      kmeans.fit(scaled_data)
      elbow.append(kmeans.inertia_)
      silhouette.append(silhouette_score(scaled_data,
                                         kmeans.fit_predict(scaled_data),
                                         metric='euclidean'))

    result = KMeans(n_clusters=K_MEANS_CLUSTERS, init='k-means++',
                    max_iter=300, n_init=10, random_state=0)
    result.fit(scaled_data)
    df[K_MEANS_CLUSTER_LABELS] = result.fit_predict(scaled_data)

    return scaled_data, elbow, silhouette, result
```

Рисунок 5 - Метод кластеризации k-средних

Результат кластеризации, отрисованный с использованием многомерного шкалирования показан на рисунке 6:

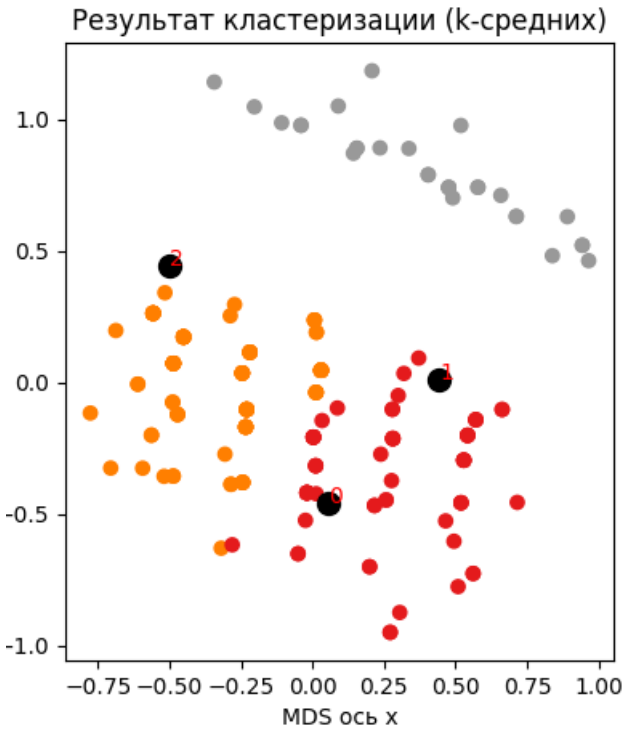

Рисунок 6 - Результат кластеризации методом k-средних

3 Выводы

Данные были разбиты на 3 кластера, информация о которых представлена в таблице 1:

Таблица 1 - Информация о кластерах (50 процентиль)

	Кластер 1	Кластер 2	Кластер 3
Размер	64 (35%)	91 (49%)	29 (16%)
Пол	Женский	Женский	Мужской
Возраст	От 35 до 44	От 25 до 34	От 35 до 44
Телеканал	познавательный (дискавери, Культура)	государственный (ОРТ, РТР)	частный (НТВ)
Пресса	глянцевые журналы	глянцевые журналы	деловая

Из этих данных можно сделать следующие рекомендации по проведению маркетинговой компании: большую часть бюджета необходимо направить на рекламу в глянцевых журнал с ориентиром на женскую молодую аудиторию. Так же часть средств необходимо направить на покупку рекламы на государственных и познавательных телеканалах.