УДК 004.8

МЕТОДЫ КЛАСТЕРИЗАЦИИ ДЛЯ СЕГМЕНТАЦИИ ТОВАРОВ В ЗАДАЧАХ РОЗНИЧНОЙ АНАЛИТИКИ

Алексюк Михаил Владимирович[1], Терешкин Дмитрий Александрович[2], Шабуров Антон Андреевич[3], Токарева Оксана Максимовна[4], Пряхина Дарья Игоревна[5]

[1]Студент;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: amv.21@uni-dubna.ru.

[2]Студент;
Университет ИТМО;
Россия, 197101, г. Санкт-Петербург, Кронверкский пр., 49, лит. А.;
e-mail: alexrumling2000@gmail.com.

[3]Студент;
Университет ИТМО;
Россия, 197101, г. Санкт-Петербург, Кронверкский пр., 49, лит. А.;
e-mail: shaburov8242@gmail.com.

[4]Аспирант;
МГУ имени М.В. Ломоносова;
Россия, 119991, г. Москва, Ленинские горы, 1;
e-mail: oxana.tokareva13@yandex.ru.

[5]Кандидат технических наук, доцент;
Государственный университет «Дубна»;
Россия, 141980, Московская обл., г. Дубна, ул. Университетская, 19;
e-mail: pryadi.09@uni-dubna.ru.

Статья посвящена обзору и сравнительному анализу методов кластеризации, применяемых для сегментации товарного ассортимента в сфере розничной торговли. Актуальность исследования обусловлена необходимостью автоматизации процессов анализа больших объемов транзакционных данных и перехода от субъективных экспертных оценок к объективным моделям управления ассортиментом и ценообразованием. В работе рассматриваются классические и современные подходы к кластеризации, включая иерархические методы, алгоритмы разделения, методы на основе плотности, а также модели на основе гауссовых смесей. Особое внимание уделяется вопросам предобработки данных, выбора оптимального числа кластеров с использованием метрик оценки качества и интерпретации полученных сегментов в контексте ценовых категорий и паттернов потребительского спроса. Проводится сравнительная оценка методов по критериям масштабируемости, устойчивости к выбросам и интерпретируемости результатов. Результаты работы могут быть использованы при построении автоматизированных аналитических систем поддержки принятия решений в розничной торговле.

Ключевые слова: кластеризация, методы разбиения, иерархические методы, методы на основе плотности, вероятностная модель, силуэтный коэффициент, сегментация товаров, ценовые сегменты, розничная аналитика, управление ассортиментом, розничная торговля.

Для цитирования:

Методы кластеризации для сегментации товаров в задачах розничной аналитики / М. В. Алексюк, Д. А. Терешкин, А. А. Шабуров [и др.]// Системный анализ в науке и образовании. 2026. № 3. С. 95-105. EDN: GFEMPB. URL: https://sanse.ru/index.php/sanse/article/view/711.

Статья находится в открытом доступе и распространяется в соответствии с лицензией Creative Commons «Attribution» («Атрибуция») 4.0 Всемирная (CC BY 4.0) https://creativecommons.org/licenses/by/4.0/deed.ru

CLUSTERING METHODS FOR PRODUCT SEGMENTATION IN RETAIL ANALYTICS PROBLEMS

Aleksiuk Mikhail V.[1], Tereshkin Dmitry A.[2], Shaburov Anton A.[3], Tokareva Oksana M.[4], Priakhina Daria I.[5]

[1]Student;
Dubna State University;
19 Universitetskaya st., Dubna, Moscow region, 141982, Russia;
e-mail: alexiuk.mischa@gmail.com.

[2]Student;
ITMO University;
Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia;
e-mail: alexrumling2000@gmail.com.

[3]Student;
ITMO University;
49 Kronverksky Pr., bldg. A, St. Petersburg, 197101, Russia;
e-mail: shaburov8242@gmail.com.

[4]PhD student;
Moscow State University named after M.V. Lomonosov, Russia;
1 Leninskie Gory , Moscow, 119991, Russia;
e-mail: oxana.tokareva13@yandex.ru.

[5]PhD in Engineering Sciences, associate professor;
Dubna State University;
19 Universitetskaya st., Dubna, Moscow region, 141982, Russia;
e-mail: pryadi.09@uni-dubna.ru.

This article provides a review and comparative analysis of clustering methods used for product assortment segmentation in retail. The relevance of the study stems from the need to automate the analysis of large volumes of transactional data and the transition from subjective expert assessments to objective models of assortment and pricing management. The paper examines classical and modern approaches to clustering, including hierarchical methods, partitioning algorithms, density-based methods, and Gaussian mixture models. Particular attention is paid to data preprocessing, selecting the optimal number of clusters using validation metrics, and interpreting the resulting segments in the context of price categories and consumer demand patterns. A comparative evaluation of the methods is conducted based on scalability, robustness to outliers, and interpretability of results. The results of the study can be used in the development of automated analytical decision support systems in retail.

Keywords: clustering, partitioning methods, hierarchical methods, density-based methods, probabilistic model, assortment management, silhouette coefficient, product segmentation, price segments, retail analytics, clustering validation, retail.

For citation:

Aleksiuk M.V., et al. Clustering methods for product segmentation in retail analytics problems. *System analysis in science and education,* 2026;(3):95–105 (in Russ). EDN: GFEMPB. URL: https://sanse.ru/index.php/sanse/article/view/711.

Введение

В современной розничной торговле управление ассортиментом, ценообразованием и прогнозированием спроса сопровождается необходимостью анализа значительных объемов структурированных и динамических данных. Это могут быть исторические данные о продажах и ценах, информация о товарных категориях и брендах, а также внешние факторы, такие как сезонность и экономические циклы. Такие данные часто характеризуются сложными взаимосвязями,

нелинейными зависимостями и скрытыми паттернами, что создает серьезные трудности для их ручного анализа и интерпретации. Особую сложность представляет задача динамической сегментации товаров, требующая регулярного пересмотра границ ценовых кластеров [1].

Традиционный подход – ручное назначение границ сегментов категорийным менеджером – обладает рядом недостатков: субъективностью, высокой трудоемкостью при пересмотре границ в динамике и невозможностью учета статистической структуры данных. Методы автоматической кластеризации позволяют преодолеть эти ограничения. В работе рассматриваются различные методы, которые охватывают все основные семейства алгоритмов кластеризации, для решения задачи сегментации товаров из набора исходных данных.

1. Методы кластеризации

1.1. Статистическое правило: ABC-анализ

Методы на основе статистических правил представляют собой наиболее простую и традиционную группу подходов к сегментации товарного ассортимента. Их суть заключается в применении заранее определенных эмпирических правил и пороговых значений, основанных на накопленном опыте и статистических закономерностях распределения данных (в первую очередь, по выручке или цене), без использования процедур обучения модели. Главными преимуществами таких методов являются высокая интерпретируемость, детерминированность результатов, минимальные вычислительные затраты и понятность для бизнес-пользователей. Они не требуют нормализации данных, подбора гиперпараметров или значительных вычислительных ресурсов, что делает их удобным инструментом для быстрой первичной сегментации [2].

Наиболее известным и широко применяемым представителем этой группы является *ABC*-анализ. Алгоритм реализует принцип Парето, на основе которого товары разделяются на категории:

- категория *A* – примерно 20% товаров, дающих 80% выручки;
- категория *B* – 30% товаров, дающих 15% выручки;
- категория *C* – 50% товаров, дающих 5% выручки.

В модифицированном варианте, использованном в работе, после присвоения категорий *A*/*B*/*C* по выручке вычисляются ценовые пороги перехода между категориями, которые затем служат границами ценовых сегментов. К достоинствам метода относятся детерминированность, отсутствие необходимости в обучении, масштабируемость и прозрачность для бизнеса, тогда как недостатками выступают фиксированное число сегментов и чувствительность к выбросам.

1.2. Методы разбиения: K-means и K-medoids

Методы разбиения – это одна из наиболее распространенных групп алгоритмов кластеризации. Их суть заключается в разбиении набора объектов на заранее заданное число кластеров путем оптимизации выбранной целевой функции. Основная особенность данных методов состоит в итеративном поиске оптимального расположения центров кластеров и перераспределении объектов между ними. Они отличаются высокой скоростью работы, простотой реализации и хорошей масштабируемостью, однако обычно требуют предварительного задания количества кластеров и чувствительны к инициализации и форме распределения данных [3, 4].

K-means – наиболее популярный и широко распространенный алгоритм среди методов разбиения. Центроидом каждого кластера выступает среднее арифметическое всех точек, входящих в этот кластер. Целевой функцией алгоритма является внутрикластерная дисперсия (сумма квадратов расстояний от точек до центроидов). Оптимизация (минимизация) этой функции осуществляется итеративно: на каждом шаге пересчитываются положения центроидов как среднее арифметическое точек, входящих в кластер, после чего точки перераспределяются к ближайшему центроиду. Алгоритм обеспечивает высокую масштабируемость и эффективно работает с данными, имеющими примерно сферическую форму кластеров [3].

K-medoids – модификация метода *K-means*, в которой центром кластера является реальный объект данных (медоид), а не среднее. Целевой функцией, подлежащей минимизации, является сумма расстояний от всех точек кластера до медоида. Такая формулировка целевой функции делает метод

существенно устойчивее к выбросам. Хотя *K-medoids* обладает лучшей робастностью и обеспечивает более высокую интерпретируемость центров кластеров (поскольку медоид – это реальный объект), его вычислительная сложность заметно выше по сравнению с *K-means*, что ограничивает применение алгоритма на очень больших выборках [4].

1.3. Иерархический метод: Agglomerative Clustering

Иерархические методы кластеризации объединяют объекты в кластеры путем последовательного объединения или разделения, формируя древовидную структуру. Их ключевая особенность заключается в построении дендрограммы – наглядного иерархического представления взаимосвязей между объектами, которое позволяет визуально оценивать структуру данных на разных уровнях детализации. В отличие от методов разбиения, иерархические алгоритмы не требуют предварительного задания числа кластеров и дают возможность аналитику выбрать оптимальный уровень разбиения уже после построения модели [5].

Наиболее распространенным представителем этой группы является агломеративная иерархическая кластеризация (*Agglomerative Clustering*). Она работает по принципу «снизу-вверх»: каждый объект изначально рассматривается как отдельный кластер. На каждом последующем шаге алгоритм объединяет два ближайших кластера до тех пор, пока все объекты не окажутся в одном кластере. Требуемое число кластеров получают путем «среза» дендрограммы на соответствующем уровне.

Среди наиболее распространенных методов связи в агломеративной кластеризации выделяют:

- *Single linkage* (ближайшего соседа) – объединяет кластеры по минимальному расстоянию между точками;
- *Complete linkage* (дальнего соседа) – по максимальному расстоянию;
- *Average linkage* – по среднему расстоянию между всеми парами точек;
- *Ward* – по критерию минимизации прироста дисперсии.

В работе используется метод связи *Ward*. Данный метод на каждом шаге выбирает те два кластера, объединение которых приводит к наименьшему приросту общей внутрикластерной дисперсии (суммы квадратов расстояний от точек до центроидов) [6].

Метод *Ward* лучше других формирует компактные, сферические и хорошо разделимые кластеры примерно сопоставимого размера. Это особенно важно для задачи ценовой сегментации, где требуется четкое разделение товаров на интерпретируемые сегменты.

1.4. Методы на основе плотности: DBSCAN и OPTICS

Методы на основе плотности представляют принципиально иной подход к кластеризации по сравнению с методами разбиения и иерархическими алгоритмами. Их основная идея заключается в том, что кластеры определяются как области высокой плотности объектов, разделенные областями низкой плотности или разреженными пространствами. Ключевыми особенностями этой группы методов являются способность выявлять кластеры произвольной формы, автоматическое определение числа кластеров, а также встроенная устойчивость к шуму и выбросам, которые помечаются как отдельный класс «шум». Эти методы не требуют предварительного задания количества кластеров и хорошо работают с данными неэллиптической структуры [7].

DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) – базовый и наиболее известный алгоритм плотностной кластеризации. Он выделяет кластеры на основе двух параметров: радиуса окрестности и минимального числа точек в этой окрестности. Точки, для которых в окрестности недостаточно других точек, считаются шумом. Алгоритм эффективен при относительно равномерной плотности данных [7].

OPTICS (*Ordering Points To Identify the Clustering Structure*) является развитием идей метода *DBSCAN*. Вместо построения конечных кластеров алгоритм создает упорядоченную последовательность точек по плотности, что позволяет анализировать структуру данных на разных уровнях масштаба. Благодаря этому *OPTICS* лучше справляется с данными переменной плотности и значительно упрощает задачу подбора параметров по сравнению с *DBSCAN* [7].

К преимуществам плотностных методов для решения задачи сегментации товаров относятся высокая устойчивость к выбросам и возможность обнаружения нестандартных ценовых ниш. Основным ограничением является чувствительность к выбору параметров и более высокая вычислительная сложность на очень больших выборках [7].

1.5. Вероятностная модель: Gaussian Mixture Models (GMM)

Вероятностные методы кластеризации рассматривают данные как результат генерации из некоторой вероятностной модели. В отличие от предыдущих методов, где каждый объект однозначно относится только к одному кластеру, вероятностные подходы реализуют алгоритм, где объекты получают вероятность принадлежности к каждому из кластеров, причем сумма этих вероятностей равна единице. Основная суть таких методов заключается в оценке параметров вероятностной модели, которая наиболее правдоподобно описывает наблюдаемые данные. Ключевыми преимуществами вероятностных моделей являются гибкость, возможность учета неопределенности, богатая статистическая интерпретация результатов и способность моделировать кластеры различной формы, размера и ориентации [8].

Наиболее популярной и эффективной реализацией вероятностного подхода является метод *Gaussian Mixture Models* (*GMM*) – модель гауссовых смесей. Модель предполагает, что все данные генерируются из смеси конечного числа многомерных нормальных (гауссовых) распределений. Каждый кластер описывается тремя параметрами: вектором математического ожидания (центром), ковариационной матрицей (определяющей форму и ориентацию кластера) и весом смеси (априорной вероятностью кластера). Важным аналитическим преимуществом *GMM* в задачах ценовой сегментации является возможность выявления товаров, находящихся на границе между сегментами. Если вероятности принадлежности товара к двум соседним сегментам являются сопоставимыми, это означает, что товар обладает переходными характеристиками и может быть чувствителен к небольшим изменениям цены [8].

2. Вычислительный эксперимент

2.1. Исходные данные и методика эксперимента

В исследовании использовались реальные транзакционные данные крупной розничной сети по формату магазинов «Гипермаркеты» за период 2022–2024 гг. с помесячной детализацией. Исходный набор данных включает следующие поля:

- *sku* – уникальный идентификатор товара;
- *art* – артикул;
- *brand* – бренд;
- *format* – формат магазина (гипермаркеты, супермаркеты, магазины у дома, минимаркеты, хард-дискаунтеры, *e-commerce*);
- *year*, *month* – год и месяц продаж;
- *price* – цена единицы товара в рублях;
- *RUB* – выручка в рублях;
- *PACKS* – количество проданных упаковок;
- *KG* – вес в килограммах;
- *FLAVOUR* / *INGREDIENTS* – вкус/состав;
- *ITEM WEIGHT ACTUAL*, *WEIGHT* 2 – весовые характеристики.

Исходная выборка содержит 5773 уникальных товара. Единицей анализа выступает уникальный *SKU* в разрезе выбранного формата магазина, года и месяца (или в агрегации по всем периодам). Для каждого *SKU* рассчитываются два ключевых признака: средняя цена и медианная цена. Именно эти два признака используются в качестве основных характеристик для последующей кластеризации и определения ценовых сегментов («Эконом», «Средний-», «Средний», «Средний+», «Премиум»).

Все расчеты проводились на языке программирования *Python* 3. Для реализации основных этапов исследования использовался следующие технологии:

- библиотеки *pandas* и *numpy* – для загрузки и агрегации транзакционных данных;
- библиотека *scikit-learn* – основная библиотека для кластеризации (*K-means*, *Agglomerative Clustering*, *DBSCAN*, *OPTICS*, *Gaussian Mixture Models*), а также для вычисления внутренних метрик качества кластеризации;
- библиотека *scipy* – для реализации алгоритма *K-medoids*;
- библиотека *Plotly* – для визуализации.

Для всех методов кластеризации, требующих предварительного задания числа кластеров (*K-means*, *K-medoids*, *Agglomerative*, *GMM*), было установлено 5 ценовых сегментов: «Эконом», «Средний-», «Средний», «Средний+», «Премиум». Для *DBSCAN* и *OPTICS* число кластеров определялось автоматически, после чего выполнялось приведение к 5 сегментам с сохранением иерархии средних цен (шумовые точки относились к ближайшему кластеру).

2.2. Критерии сравнения методов

Для оценки качества кластеризации используются три общепризнанные внутренние метрики [9], не требующие внешней разметки данных.

Коэффициент силуэта измеряет, насколько хорошо каждый объект размещен в своем кластере по сравнению с ближайшим соседним кластером. Значение коэффициента находится в диапазоне от -1 до 1. Чем ближе значение к 1, тем лучше объект размещен в своем кластере, а кластеры хорошо отделены друг от друга. Коэффициент рассчитывается по формуле (1):

$$S \in [-1,1]\ S = \frac{b-a}{max(a,b)}, \tag{1}$$

где a – среднее расстояние от товара до всех остальных товаров в его же кластере;

b – среднее расстояние от товара до всех товаров ближайшего соседнего кластера.

Индекс Калински–Харабаса оценивает качество кластеризации через отношение дисперсии между кластерами к дисперсии внутри кластеров. Более высокие значения индекса указывают на компактные и хорошо разделимые кластеры. Индекс рассчитывается по формуле (2):

$$CH = \frac{\left[\frac{SS_B}{k-1}\right]}{\left[\frac{SS_w}{n-k}\right]}, \tag{2}$$

где SS_B – сумма квадратов расстояний от центра каждого кластера до общего центра всех данных, взвешенная на размер кластера;

SS_w – сумма квадратов расстояний от каждой точки до центра ее кластера;

k – число кластеров;

n – общее число товаров.

Индекс Дэвиса–Болдина рассчитывается как среднее соотношение внутрикластерного разброса к расстоянию между центрами кластеров. Чем ниже значение индекса, тем лучше качество кластеризации, поскольку это свидетельствует о компактных кластерах и значительном расстоянии между ними. Индекс рассчитывается по формуле (3):

$$DB = \frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{S_i + S_j}{M_{ij}}\right), \tag{3}$$

где S_i – средний разброс внутри кластера i;

S_j – средний разброс внутри кластера j;

M_{ij} – расстояние между центрами кластеров i и j;

k – число кластеров (5).

2.3. Результаты эксперимента

На рисунке 1 представлен результат применения модифицированного (количество конечных сегментов расширено до 5) *ABC*-анализа к исходным данным. График демонстрирует распределение товаров в пространстве «медианная цена – средняя цена». Вертикальные пунктирные линии соответствуют автоматически рассчитанным ценовым порогам перехода между сегментами. Хорошо видно, как *ABC*-анализ формирует относительно линейное разделение, однако в верхнем диапазоне цен (сегмент «Премиум») наблюдается значительный разброс, а некоторые товары с высокой средней ценой попадают в категорию «Шум/Выбросы».

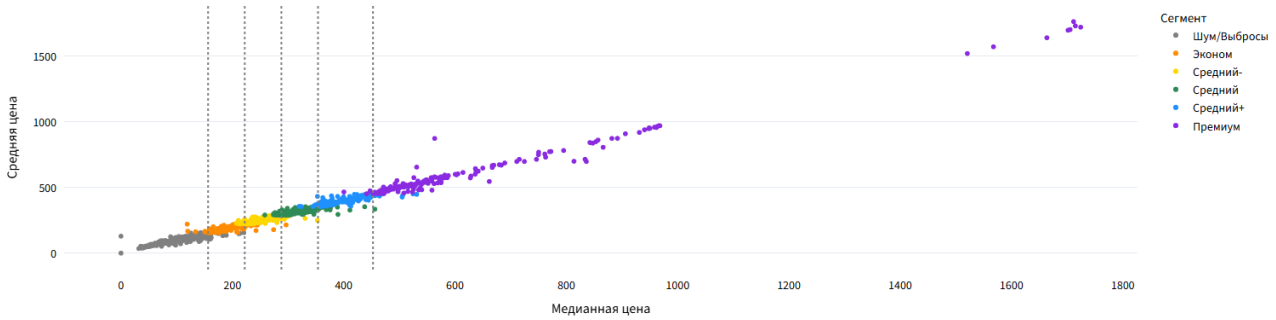


Рис. 1. Пример применения метода ABC-анализа к исходным данным

На рисунке 2 представлен результат кластеризации с помощью метода *K-means*. Алгоритм четко разделил товары на пять ценовых сегментов. Видно, что сформированы достаточно компактные и линейно разделимые кластеры, особенно в диапазоне низких и средних цен. В премиум-сегменте наблюдается небольшое количество точек с высокой дисперсией, что характерно для товаров высокой ценовой категории.

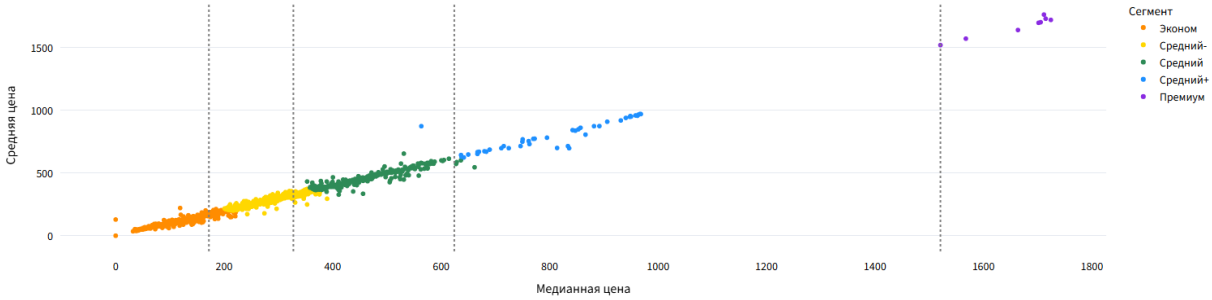


Рис. 2. Пример применения метода K-means к исходным данным

На рисунке 3 представлен результат кластеризации с использованием алгоритма *K-medoids*. График демонстрирует четкое разделение товаров на пять ценовых сегментов. Метод показал хорошую устойчивость к выбросам: крайние точки в сегменте «Премиум» распределены более аккуратно, чем в других алгоритмах. Сегменты сохраняют логичную последовательность по возрастанию цены, однако границы между ними в средней ценовой зоне выглядят чуть более «размытыми» по сравнению с результатами применения метода *K-means*. Это объясняется большей робастностью алгоритма к аномальным значениям цен.

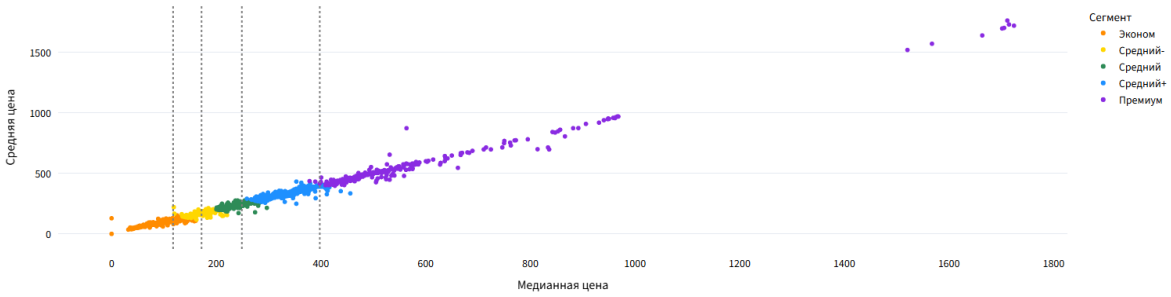


Рис. 3. Пример применения метода K-medoids к исходным данным

На рисунке 4 представлен результат работы агломеративной иерархической кластеризации с методом связи *Ward*. Алгоритм последовательно объединял наиболее похожие объекты, формируя иерархическую структуру. Благодаря использованию метода *Ward* сегменты получились компактными и хорошо сбалансированными. Особенно хорошо видна четкая граница между средними и высокими ценовыми категориями. Распределение точек в сегменте «Премиум» выглядит структурированным и однородным. Агломеративная кластеризация с методом *Ward* показала одно из лучших визуальных разделений.

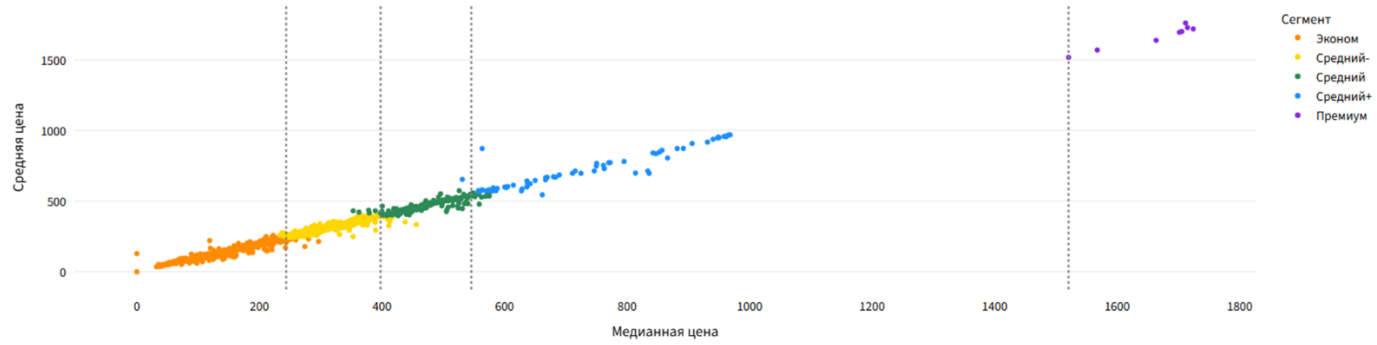


Рис. 4. Пример применения агломеративной кластеризации к исходным данным

На рисунке 5 представлен результат кластеризации с помощью алгоритма *DBSCAN*. На графике видно характерное для этого алгоритма поведение: основной массив *SKU* (преимущественно средний ценовой диапазон) был объединен в один крупный кластер. Кластеры «Эконом» и «Средний-» включают относительно небольшое количество *SKU*. Самым большим по численности оказался сегмент «Средний», что отражает реальную структуру ассортимента – преобладание товаров массового спроса в средней ценовой категории. В сегменты «Премиум» и особенно «Средний+» попало всего несколько *SKU*. Такое распределение демонстрирует способность *DBSCAN* выявлять плотностные структуры данных, но также подчеркивает его склонность объединять большую часть товаров в один доминирующий кластер при наличии выраженной основной плотности. Выбросы и редкие высокодоходные товары были выделены в отдельные небольшие кластеры.

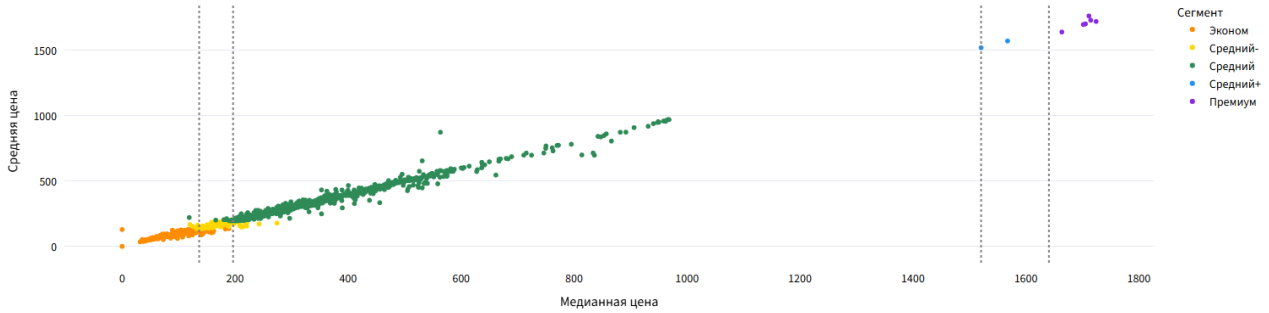


Рис. 5. Пример применения метода DBSCAN к исходным данным

На рисунке 6 представлен результат кластеризации с помощью метода *OPTICS*. На графике видно более сбалансированное и детальное разделение товаров по сравнению с результатом применения метода *DBSCAN*. Однако результат оказался достаточно необычным: сегмент «Премиум» стал самым многочисленным. Большая часть товаров была отнесена именно к высокой ценовой категории. При этом сегменты «Эконом» и «Средний-» получились относительно небольшими, а «Средний» и «Средний+» заняли промежуточное положение. Такое распределение может быть связано с особенностями внутренней структуры данных и выбранными параметрами алгоритма, которые привели к тому, что метод объединил значительную часть товаров в один крупный плотный кластер в верхнем ценовом диапазоне.

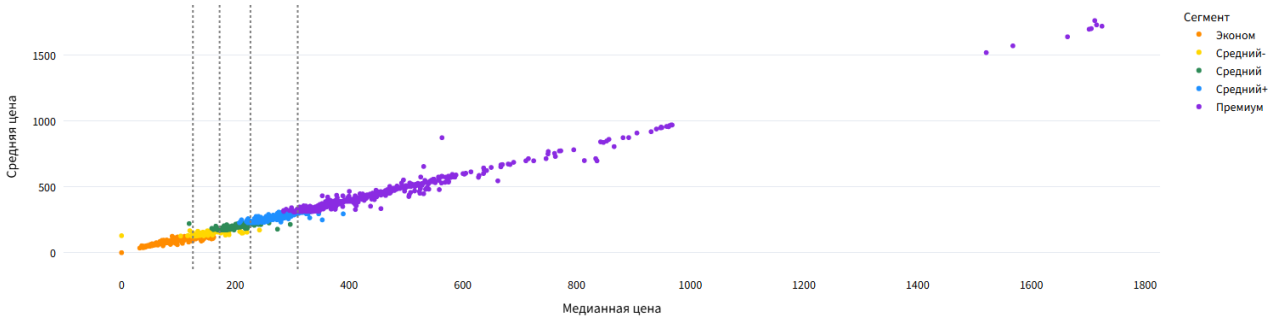


Рис. 6. Пример применения метода OPTICS к исходным данным

На рисунке 7 представлен результат кластеризации с помощью вероятностной модели *Gaussian Mixture Models* (*GMM*). Алгоритм *GMM* реализовал «мягкую» кластеризацию, однако характерной особенностью данного разбиения стало крайне малое количество товаров в низких ценовых сегментах: «Эконом» и «Средний-» практически отсутствуют. Основная масса товаров сконцентрировалась в сегментах «Средний», «Средний+» и «Премиум». Это может указывать на относительно низкую пригодность данного метода для жесткой ценовой сегментации в условиях сильной асимметрии распределения цен в данных.

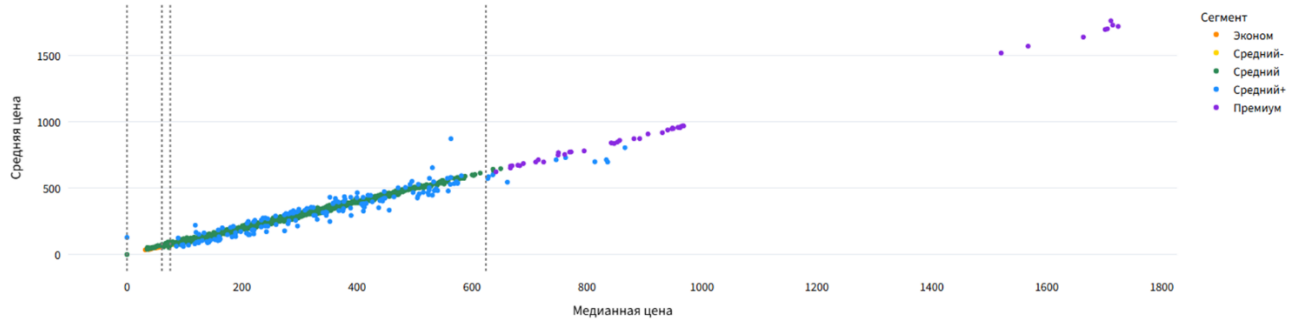


Рис. 7. Пример применения GMM к исходным данным

Результаты оценки качества проведенной кластеризации товаров из исходных данных представлены в таблице 1.

Примечание: для *ABC*-анализа и *GMM* метрики внутренней валидации не применимы по определению.

Табл. 1. Метрики оценки методов кластеризации по результатам эксперимента

Метод	Коэффициент силуэта	Индекс Калински-Харабаса	Индекс Дэвиса-Болдина
K-means	0.573	6 991.3	0.449
Agglomerative Clustering	0.568	5 875.3	0.499
K-medoids	0.483	2 433.4	0.608
OPTICS	0.396	1 512.2	0.658
DBSCAN	0.327	1 416.4	0.502
GMM	—	—	—
ABC	—	—	—

2.4. Интерпретация результатов

Метод *K-means* продемонстрировал наилучшие результаты по всем трем метрикам. Высокий коэффициент силуэта (0,573) указывает на достаточно хорошую отделимость кластеров, а максимальный индекс Калински–Харабаса (почти 7 000) подтверждает компактность кластеров при сильной межкластерной дифференциации. Низкий индекс Дэвиса–Болдина (0,449) также свидетельствует в пользу качества. Основное объяснение – ценовые данные после стандартизации приобретают форму, близкую к сферическим кластерам, для которых применение метода *K-means* является оптимальным.

Agglomerative clustering с методом связи *ward* показал практически идентичный с методом *K-means* результат (силуэт 0,568 против 0,573). Это важно, поскольку агломеративный метод детерминирован и не зависит от случайной инициализации центроидов. При достаточных вычислительных ресурсах он может быть рекомендован как альтернатива метода *K-means*.

Метод *K-medoids* ожидаемо уступил по метрикам, так как его целевая функция менее чувствительна к компактности кластеров, чем в методе *K-means*. Тем не менее, показатель силуэта 0,483 остается приемлемым. Преимущества метода *K-medoids* (устойчивость к выбросам, интерпретируемые центры) могут перевесить снижение метрик в задачах с зашумленными данными.

Алгоритмы *DBSCAN* и *OPTICS* показали более низкие значения метрик, что объясняется принципиальным несоответствием между формой кластеров, выделяемых плотностными методами, и геометрическими предположениями, заложенными в метриках. Однако это не означает, что плотностные методы хуже: они эффективно выявляют нестандартные ценовые ниши и маркируют выбросы. В эксперименте *DBSCAN* идентифицировал 47 *SKU* как шум (≈1% выборки), что имеет самостоятельную аналитическую ценность.

Метод *GMM* не сравнивался напрямую по метрикам, поскольку его вероятностная природа предполагает иные критерии оценки.

Метод *ABC*-анализ формально не оценивался метриками кластеризации, однако его практическая полезность для быстрой первичной сегментации остается высокой.

Заключение

В работе проведен сравнительный анализ методов кластеризации применительно к задаче сегментации товаров по ценам в розничной торговле. На основе вычислительного эксперимента на реальных данных и теоретического анализа свойств алгоритмов сформулированы следующие основные выводы:

- методы *K-means* и *Agglomerative clustering* (*ward*) демонстрируют наилучшее качество по метрикам силуэта, Калински–Харабаса и Дэвиса–Болдина, что делает их предпочтительным выбором для стандартных задач ценовой сегментации;
- метод *K-medoids* является наилучшим выбором в условиях зашумленных данных, уступая по компактности кластеров, но выигрывая в устойчивости к выбросам и интерпретируемости центров;
- методы *DBSCAN* и *OPTICS* незаменимы для выявления нестандартных ценовых ниш и аномальных товаров;
- метод *GMM* предоставляет уникальную возможность «мягкой» кластеризации, выявляя товары на границе ценовых сегментов, что имеет высокую практическую ценность для задач гибкого ценообразования;
- метод *ABC*-анализ сохраняет свою ценность как быстрый и воспроизводимый вариант, особенно на этапе предварительного изучения данных.

Результаты исследования могут быть использованы при построении автоматизированных аналитических систем поддержки принятия решений в розничной торговле для прогнозирования цен и продаж товаров с учетом сезонности спроса.

Список источников

1. Özmen E. S. Time Series Performance and Limitations with SARIMAX: An Application with Retail Store Data // Electronic Turkish Studies. – 2021. – Vol. 16. – Iss. 5. – P.1583-1592. – http:// doi.org/10.7827/TurkishStudies.49699
2. Silaen B. R., Nasution M., Muti'ah R. Implementation of the ABC Analysis to the Inventory Management // International Journal of Science, Technology & Management. – 2024. – Vol. 5. – Iss. 4. – P. 816-825. – https://doi.org/10.46729/ijstm.v5i4.1144.
3. Time-varying price elasticity of demand estimation for demand-side smart dynamic pricing // / J. Ruan, G. Liu, J. Qiu [et al.] // Applied Energy. – 2022. – Vol. 322. – P. 119520. – DOI: 10.1016/j.apenergy.2022.119520.
4. Yahya A., Kurniawan R. Implementasi Algoritma K-Means untuk Pengelompokan Data Penjualan Berdasarkan Pola Penjualan: Implementation of K-Means Algorithm for Clustering Sales Data Based on Sales Patterns // MALCOM: Indonesian Journal of Machine Learning and Computer Science. – 2025. – Vol. 5. – Iss. 1. – P. 350-358. – https://doi.org/10.57152/malcom.v5i1.1773
5. Бова В. В., Кравченко Ю. А., Родзин С. И. Методы и алгоритмы кластеризации текстовых данных (обзор) // Известия ЮФУ. Технические науки. – 2022. – №. 4. – С. 122-143. – https://doi.org/10.18522/2311-3103-2022-4-122-143.
6. Lanza K. E. Comparison of Linkages for Hierarchical Clustering of Directed Unweighted Graph // Louis: open scholarship and creative works. Honors Capstone Projects and Theses. – University of Alabama, 2024. – 54 p. – URL: https://louis.uah.edu/honors-capstones/921/.
7. Xu H. C., Pham N. Scalable DBSCAN with random projections // Advances in Neural Information Processing Systems. – 2024. – Vol. 37. – P. 27978-28008. – https://doi.org/10.52202/079017-0877.
8. Wang F., Li L., Fränti P. // Eurasip Journal on Audio Speech and Music Processing. – 2025. – Iss. 1. – P. 2. – https://doi.org/10.1007/s41060-025-00912-x
9. The silhouette coefficient and the davies-bouldin index are more informative than dunn index, calinski-harabasz index, shannon entropy, and gap statistic for unsupervised clustering internal evaluation of two convex clusters / D. Chicco, A. Campagner, A. Spagnolo [et al.] // PeerJ Computer Science. – 2025. – Vol. 11. – P. e3309. – https://doi.org/10.7717/peerj-cs.3309.