NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science

Bachelor's Programme "HSE and University of London Double Degree Programme in Data Science and Business Analytics"

Research project report

Study of clustering methods in market basket analysis

Student:

group БПАД234 Ivan Churilov

Supervisor: Nataliya Titova

Senior lecturer MIEM HSE

Moscow 2025

Contents

1 Annotation

Abstract (English). This research explores the application of clustering methods in market basket analysis for customer segmentation in an online retail environment. Using transaction data, a client-item matrix was constructed and reduced using PCA for visualization and preprocessing. Clustering was performed using KMeans and GaussianMixture (EM algorithm), revealing distinct groups of customers with specific purchasing patterns. Additionally, association rule mining (Apriori and 'association rules') was applied to identify frequently co-purchased items within each cluster. The results highlight the differences between the clusters - from baby products to groceries - allowing personalized recommendations and targeted marketing strategies based on cluster-specific behavior.

Аннотация (Русский). В рамках данного исследования были изучены методы кластеризации и их применение в анализе рыночной корзины для сегментации клиентов онлайн-магазина. На основе данных о заказах клиентов была проведена подготовка матрицы клиент × товар, выполнено снижение размерности с помощью PCA, а также применены алгоритмы KMeans и EM (GaussianMixture) для выявления групп потребителей с похожими предпочтениями. Дополнительно проведен Market Basket Analysis с использованием алгоритмов Apriori и association rules для определения популярных пар товаров внутри каждого кластера. Результаты показали различия в поведении кластеров: от детских товаров до продуктов питания, что позволяет использовать полученную информацию для персонализированных рекомендаций и маркетинговых стратегий.

Key words

Market Basket Analysis, Consumer Behavior, Data Visualization, Transaction Analysis, Feature Selection, Clustering Methods (Future Plans)

GitHub

https://github.com/juzzzzzt/Study-of-clustering-methods-in-market-basket-analysis

2 Introduction

In modern retail and e-commerce, understanding customer behavior is one of the most important tasks for effective marketing, product recommendations, and sales optimization. With the growing quantity of transactional statistics, businesses have get admission to to unique statistics about shopping patterns, frequency, and economic cost — however extracting significant insights from this records requires superior analytical techniques.

Customer segmentation via clustering has grow to be a powerful technique for identifying companies with comparable behavior, alternatives, and shopping behavior. By applying unsupervised learning methods such as KMeans and Expectation-Maximization (EM) algorithms, it becomes possible to divide clients into distinct clusters based on capabilities consisting of purchase frequency, common order cost, and geographic area.

These clusters can then serve as a foundation for targeted advertising strategies, personalized promotions, and optimized stock control. This research focuses on the software of clustering strategies in marketplace basket analysis to uncover how exclusive patron corporations engage with products and what their behavioral styles appear like. The integration of clustering with category-based feature analysis allows deeper personalization and strategic decision making in customer engagement

2.1 Research Goal

The goal of this study is to investigate and apply clustering methods (KMeans, EM/GMM) to segment customers based on their purchasing behavior, using real-world transaction data from an online retail business. The work emphasizes the use of clustering not only as a tool for grouping clients, but also as a basis for further marketing recommendations.

2.2 Research Objectives

Data preparation: Clean and structure raw transactional data for feature engineering.

Feature extraction : Derive key metrics per customer, including: Total spending Number of orders Average check Regional affiliation Category-specific purchase frequency

Dimensionality Reduction : Apply PCA to visualize the customer space and improve clustering performance.

Clustering analysis : Use KMeans and Gaussian Mixture Models (EM) to segment customers into meaningful groups.

Cluster Interpretation : Analyze the behavior, preferences, and RFM characteristics of each group.

Application of Clusters : Connect results with Market Basket Analysis insights to support personalized marketing and cross-selling strategies.

2.3 Hypothesis and key question

Clustering methods are hypothesized to allow the identification of distinct customer segments with consistent inner traits, which may be used to customize marketing movements and optimize product guidelines.

How can clustering strategies assist become aware of meaningful customer segments based on buying conduct, and how do these segments vary in phrases of product choices?

3 Literature review

3.1 Clustering as a Tool for Customer Segmentation

Customer segmentation is one of the most widely used techniques in marketing and data analytics. It allows businesses to identify groups of customers with similar behavior patterns, enabling personalized marketing strategies and targeted promotions. In this study, we focus on two clustering methods: KMeans and Expectation-Maximization (EM) , both of which are unsupervised learning approaches.

KMeans is an iterative algorithm that partitions observations into K clusters based on Euclidean distances. It is known for its simplicity and speed, making it suitable for large datasets [1]. Gaussian Mixture Models (GMM) , which use the EM algorithm , offer probabilistic clustering and can better handle overlapping or non-spherical clusters [2]. Dimensionality reduction techniques such as Principal Component Analysis (PCA) and UMAP were applied prior to clustering to reduce feature space and improve model performance [3].

3.2 Market Basket Analysis (MBA)

Market Basket Analysis is a powerful method for identifying associations between products purchased together [4]. The core idea is to discover frequent itemsets and association rules that help businesses understand customer preferences and optimize product recommendations.

The main algorithms used for MBA include:

Apriori Algorithm : Introduced by Agrawal and Srikant, it uses support and confidence measures to find frequent itemsets. FP-Growth (Frequent Pattern Growth) : A more efficient approach that avoids candidate generation and builds a compact tree structure [5]. Association Rules Mining : Based on metrics such as support , confidence , and lift , this technique identifies statistically significant relationships between items [6]. These algorithms have been successfully applied in retail for:

Cross-selling and upselling Inventory optimization Personalized recommendation systems

3.3 Integration of Clustering and MBA

Recent studies show that combining clustering with MBA improves the accuracy and relevance of market basket insights [7].

Key advantages: Personalization : Different customer segments show different buying habits; MBA per cluster provides deeper insights. Regional targeting : When clusters incorporate geographic information (e.g., 'Центральный ф.о. регион', 'Южный ф.о. регион'), it becomes possible to tailor offers accordingly. RFM + MBA synergy : Using Recency-Frequency-Monetary value clusters helps identify high-value and low-frequency buyers, whose association patterns may vary significantly [8]. Examples of successful integration:

In a large e-commerce dataset, [7] showed that applying KMeans followed by MBA led to a 17 percent increase in cross-selling conversion . It was also demonstrated that EM-based clustering allowed for soft assignment and revealed nuanced buying behaviors, especially in mixed-category environments.

4 Methodology and description of the methods applied

4.1 Introduction to Customer Clustering

Customer segmentation is a critical task in modern retail analytics, allowing businesses to divide their customer base into meaningful groups based on behavioral and demographic features. Clustering — as an unsupervised machine learning technique — enables the discovery of hidden patterns in data without predefined labels.

In this research, we focus on two major clustering algorithms:

- KMeans – a centroid-based hard clustering method

- Gaussian Mixture Model (GMM) with Expectation-Maximization (EM) algorithm – a probabilistic soft clustering approach

Both methods are applied after dimensionality reduction techniques such as:

- Principal Component Analysis (PCA)

- UMAP (Uniform Manifold Approximation and Projection)

These approaches are used to analyze transactional data from an online retail store, aiming to uncover distinct customer profiles that can later be linked to market basket analysis for personalized marketing strategies.

4.1.1 KMeans Clustering: Theory and Parameters

Description: KMeans is one of the most widely used algorithms for clustering tasks due to its simplicity, speed, and interpretability. It partitions a dataset into K clusters by minimizing the sum of squared distances between each data point and its assigned cluster center (centroid).

The algorithm works iteratively:

- Randomly initialize K centroids.

- Assign all data points to the nearest centroid.

- Recalculate the centroid as the mean of the assigned points.

- Repeat until convergence.

Mathematical Formulation:

$$\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

where mu i is the centroid of cluster i, and S i is the set of points assigned to cluster i.

4.1.2 Expectation-Maximization (EM) Algorithm with Gaussian Mixture Models (GMM)

Description: The EM algorithm is an iterative method often used in conjunction with Gaussian Mixture Models (GMM) for probabilistic clustering. Unlike KMeans, GMM assumes that the data points are generated from a mixture of Gaussian distributions, where each distribution represents a cluster.

The EM algorithm alternates between two steps:

- E-step (Expectation): Estimate the probability of each data point belonging to each cluster.

- M-step (Maximization): Update the parameters of the Gaussian distributions (means, covariances, and weights).

This process continues until convergence.

Mathematical Formulation:

$$p(x) = \sum_{i=1}^{K} \alpha_i \cdot \mathcal{N}(x \mid \mu_i, \Sigma_i)$$

where:

- α_i is the weight of component i

- μ_i is the mean vector of component i

- Σ_i is the covariance matrix of component i

Each point receives a probability of belonging to each cluster, rather than being assigned to just one.

4.1.3 Dimensionality Reduction Techniques

High-dimensional datasets frequently suffer from the curse of dimensionality , that could degrade the overall performance of clustering algorithms. To mitigate this, we use dimensionality reduction techniques to mission the facts right into a decrease-dimensional area at the same time as retaining key systems.

Principal Component Analysis (PCA)

Description: PCA is a linear dimensionality reduction technique that finds orthogonal directions (principal components) that capture the maximum variance in the data.

It transforms the original feature space into a new coordinate system where the first axis corresponds to the direction of highest variance, the second to the next, and so on.

Formula: For a given data matrix X, PCA computes the eigenvectors of the covariance matrix XTX, sorted by eigenvalues in descending order.

4.1.4 Quality assessment of clustering

Silhouette Score

$$\text{Silhouette Score} = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where:

- $a(i)$ — average distance between sample i and other samples in the same cluster

- $b(i)$ — average distance to samples in the nearest cluster

Interpretation:

- $+1 \rightarrow$ samples are far away from neighboring clusters

- $0 \rightarrow$ samples lie equally distant to both clusters

- $-1 \rightarrow$ samples assigned to wrong clusters

Davies-Bouldin Index

This metric evaluates the ratio of within-cluster scatter to between-cluster separation:

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

Where:

- σ_i — average distance of points in cluster i

- $d(c_i, c_j)$ — distance between centroids of clusters i and j

Interpretation:

- Lower values indicate better separation between clusters

Calinski-Harabasz Index

$$\text{CHI} = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{n - K}{K - 1}$$

Where:

- $\text{Tr}(B_k)$ — trace of the between-cluster dispersion matrix

- $\text{Tr}(W_k)$ — trace of the within-cluster dispersion matrix

- n — number of data points

- K — number of clusters

Interpretation:

- A higher CHI value indicates better-defined, more separated clusters. This index is particularly useful when comparing different clustering solutions for the same dataset.

5 Course of research

5.1 Data Preparation and Cleaning

The data preparation process began by uploading the source file "!01-02-18-VSE.csv which contained transactional records of customer purchases. The source dataset included 693,710 rows and 39 columns covering various order attributes such as customer information, products, delivery regions, payment methods, and reasons for canceling orders. Despite the richness of the data, it required significant preprocessing to eliminate inconsistencies, omissions, and redundant information before analysis could be performed.

The first step was to investigate the data structure. The main data types were numeric values (e.g. order amounts, number of products) and categorical attributes (e.g. regions, product categories). At the same time, some of the data contained missing values, especially in columns related to product information (Группа2, Группа3, Группа4), where the percentage of missing values reached 21.5 percent. For some numeric fields, such as ЦенаЗакупки and Маржа, skips were also observed, which could significantly affect the quality of clustering. To deal with this problem, it was decided to remove rows with critical missing values, such as region or order number, and fill the numeric fields with the average values of the corresponding columns.

The next step was to remove irrelevant data. There were test orders in the dataset that could distort the analysis results, so they were excluded. Transactions with unreliable statuses, such as “Отменен” or “Не обеспечен”, were also filtered out. Special attention was paid to the reasons for canceling orders, many of which had vague descriptions, such as integration errors with delivery services. Such records were removed to focus only on actual purchases. In addition, all records related to delivery services (Тип = “ИНОЕ”) were excluded from the dataset, as they were not relevant to the product category analysis.

To standardize the categorical data, geographic regions were grouped into federal districts, such as the Central, Volga and Southern Federal Districts. This decision made it possible to simplify data interpretation and reduce the level of detail that could complicate the analysis. Product categories were also aggregated at a higher-order level. For example, instead of many small subcategories like Номенклатура, the data was grouped into broader categories such as Игрушки, Косметика/Гигиена, and Подгузники. This helped to reduce the dimensionality of

the data and improve the interpretability of the results.

Numeric fields required additional cleanup. In many cases, values such as order amount or price contained formatting with commas or special characters that could interfere with mathematical operations. These characters were removed and the data was converted to a numeric format. All numeric fields were then normalized to ensure that each feature had a uniform effect on subsequent clustering algorithms.

The data was then aggregated at the customer level to create a single view of each customer. For each customer, key metrics such as total sum of all orders, total number of products, average check and number of purchases per product category were calculated. New binary attributes were created to indicate a customer's affiliation with a specific federal district. These metrics formed the basis of the final dataset, which contained 548,654 rows and 29 columns.

Several difficulties arose during the data preparation process. One of them was the high dimensionality of the Номенклатура category, which contained too many unique values, making it unsuitable for clustering. This problem was solved by aggregating items into larger categories. Another challenge was the presence of ambiguous reasons for canceled orders, such as technical errors, which could have skewed the analysis. These records were removed. In addition, regional information was initially too detailed, but once aggregated into federal districts, it became more manageable and meaningful.

After all data cleaning and processing steps were completed, a final quality control was performed. Unique values in each column, the number of null values and missing data were checked. All columns were converted to the appropriate data types: numeric fields became integer or real, and categorical fields became string fields. The final dataset contained exactly 548,654 rows and 29 columns, ready for dimensionality reduction and clustering algorithms.

5.2 Calculation of summary metrics

The process of calculating summary metrics began with aggregating data to calculate key metrics such as total revenue, average check, number of unique customers and products, and other important characteristics. To accomplish this, an aggregation function was created that allowed different metrics to be calculated for each group of data. For example, to calculate total revenue, the sum of values in the СуммаСтроки was used, and to determine the number of unique orders, the function of counting unique values in the НомерЗаказаНаСайте column was used. Similarly, average values were calculated, such as the average check, which was calculated as the average value in the СуммаДокумента column.

One of the first tasks was to eliminate problems with formatting numerical data. In the original dataset, the values in the СуммаДокумента and СуммаСтроки columns contained characters such as commas and spaces, which could interfere with the correct execution of mathematical operations. To solve this problem, all values were converted to string format, after which special characters were replaced with dots or removed. The data was then converted back to numeric format, allowing further calculations to be performed without errors.

Two features were used to calculate the overall quantity of products and their average values: one to calculate the sum of all products and the opposite to calculate the average cost. These metrics had been calculated for man or woman product categories, which includes Группа2 and Группа3, in addition to for the complete information set as a whole. In addition, the range of unique gadgets turned into calculated for each product class, which made it feasible to assess the diversity of the product range.

After finishing the calculations, the statistics turned into grouped by order week to analyze the dynamics of indicators through the years. For each week, metrics consisting of general sales, number of particular orders, common check, and number of gadgets were calculated. This made it feasible to become aware of trends in client behavior and determine durations of improved demand. For example, the evaluation confirmed that during some weeks there has been a significant increase in the wide variety of orders, that can be related to seasonal promotions or advertising campaigns.

Special interest changed into paid to calculating metrics for brought orders. To do this, handiest the ones information in which the order reputation turned into indicated as "Доставлен" have been filtered. After filtering, the records become aggregated again the usage of the identical features as earlier than. This allowed us to obtain more correct metrics that reflect real client conduct, as canceled or unfulfilled orders could distort the evaluation results.

There have been lacking values in key columns consisting of СуммаДокумента and Количество. To reduce the effect of these gaps on the calculation consequences, strategies have been used to fill within the gaps with average values or to delete rows with crucial gaps. Another challenge became the excessive dimensionality of the records, in particular in columns related to product classes. To address this, the facts was aggregated at a better level, which decreased the complexity of the analysis and advanced the interpretability of the results.

The final dataset after calculating the summary metrics contained information on key indicators for every week of orders. These metrics fashioned the basis for subsequent analysis, which includes customer clustering and identification of buying behavior styles. The effects of the calculations have been offered in tables, wherein every row corresponded to one week, and the columns contained values for metrics together with general sales, average test, variety of unique

clients, and range of merchandise. This supplied a clean image of commercial enterprise dynamics and organized the records for in addition evaluation.

5.3 Data preparation for clustering

Data preparation for clustering began with the aggregation of transaction records at the customer level to create a unified view of each buyer's behavior. The original dataset contained order information, including attributes such as order number, order amount, region, product categories, and payment methods. However, to perform clustering, this data had to be converted into a format that could effectively describe customer behavior through numerical and categorical features.

The first step was to examine the data structure. The main types of data were numerical values (e.g., order amount, number of items) and categorical features (e.g., regions, product categories). For each customer, summary metrics were calculated, such as total number of orders, total number of products, average check, and number of purchases per product category. These metrics allowed us to create a more compact representation of the data, where each row corresponded to a single customer and the columns reflected their behavior in the store.

One of the important thing duties was to take away troubles with formatting numerical information. In the unique dataset, values in columns along with СуммаЗаказаНаСайте and СуммаДокумента contained characters including commas and spaces, which can intrude with the precise execution of mathematical operations. To remedy this problem, all values have been transformed to numerical format, after which special characters had been changed with dots or removed. This ensured that further calculations may be carried out without mistakes.

For categorical data, such as regions and product categories, new attributes were created that better reflected the data structure. For example, regions were combined into federal districts, such as the Central, Volga, and Southern federal districts. This decision simplified data interpretation and reduced the level of detail that could complicate analysis. Product categories were also aggregated at a higher level. For example, instead of many small subcategories such as Номенклатура, the data was grouped into broader categories such as "Игрушки," "Косметика/Гигиена," and "Подгузники." This helped reduce the dimensionality of the data and improve the interpretability of the results.

After completing the calculations of summary metrics, the data was supplemented with new attributes, such as the number of orders, total number of items, and average check. These metrics formed the basis of a new dataset that contained information about each customer. In

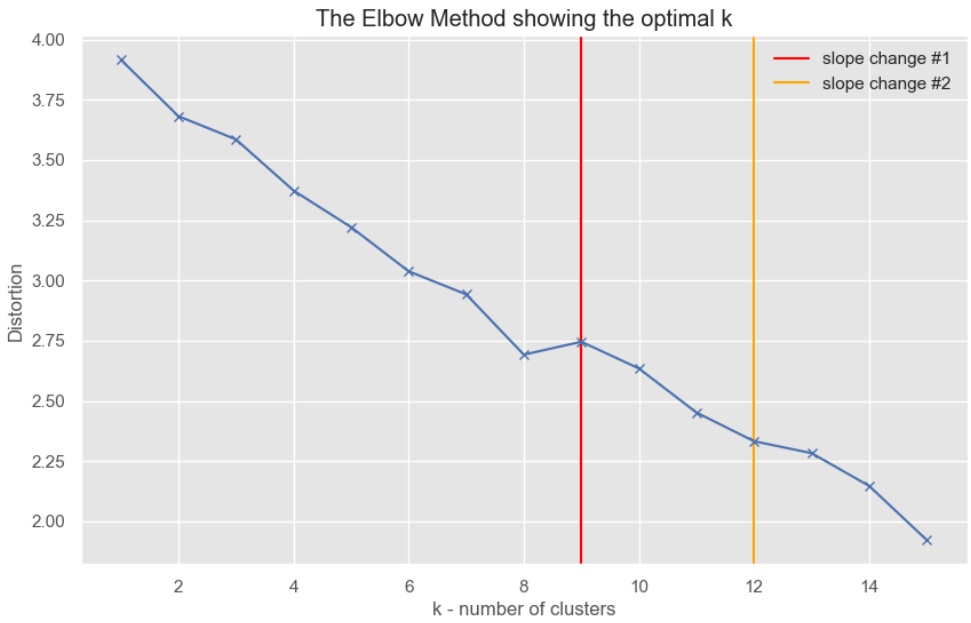
addition, binary flags were created for each customer, indicating their affiliation with a specific federal district. For example, if a customer made purchases in the Central Federal District, the corresponding flag took the value 1, otherwise — 0. This made it possible to take into account regional characteristics of customer behavior in the clustering process.

The very last dataset after guidance contained information on key metrics for each purchaser, such as total quantity of orders, general quantity of items, common check, and range of purchases per product class. These metrics fashioned the premise for subsequent evaluation, including client clustering and identity of purchasing behavior patterns. The outcomes of the calculations had been presented in tables, in which each row corresponded to one customer, and the columns contained values for metrics such as quantity of orders, general quantity of gadgets, and common check.

5.4 K-Means Clustering

Data clustering using the K-Means method became a key level inside the evaluation aimed toward figuring out hidden styles in consumer behavior. The K-Means algorithm changed into used for clustering, which divides facts into a predetermined quantity of clusters (K) by minimizing the gap between information points and cluster centroids. This technique changed into selected for its simplicity, efficiency, and ability to method massive quantities of statistics.

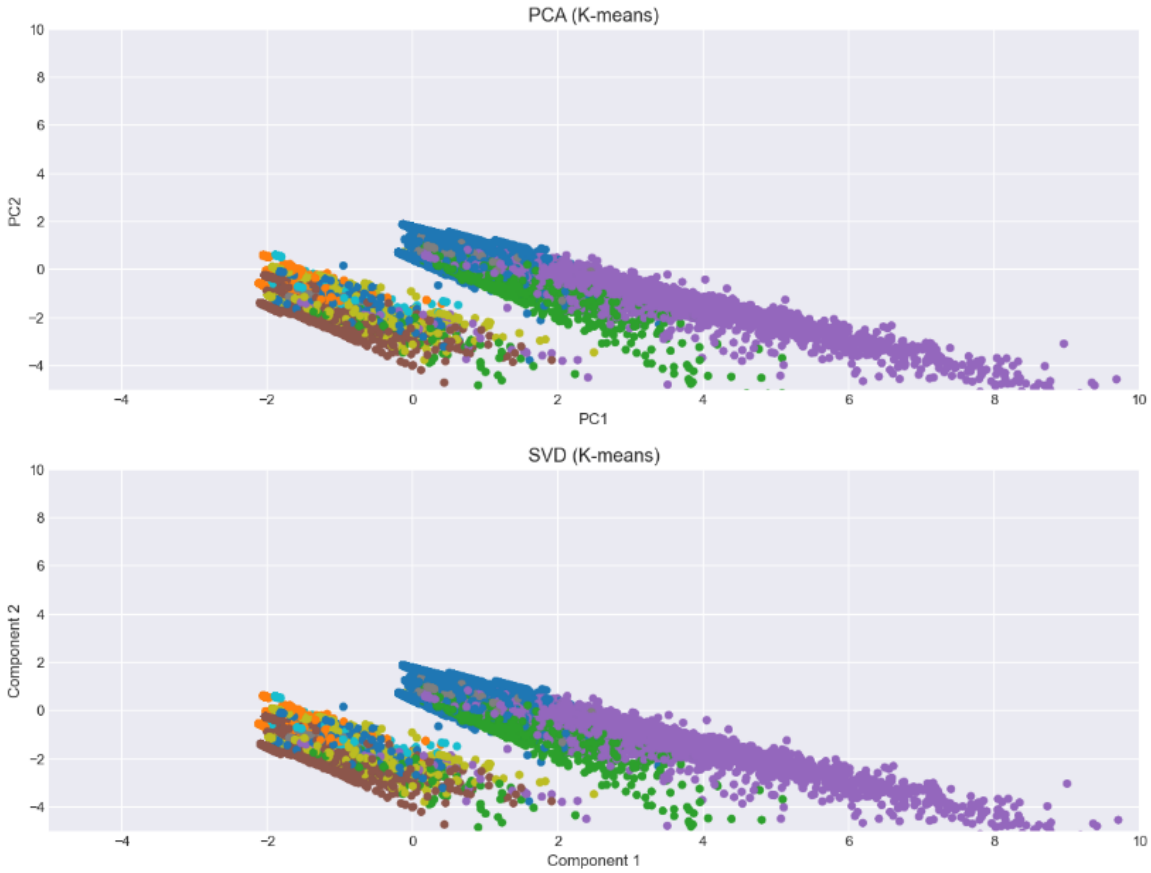
The first step become to decide the finest variety of clusters. To do that, several clustering best assessment techniques have been applied, which include the Elbow Method, Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. The Elbow Method allowed us to visualize the dependence of intracluster dispersion at the variety of clusters, which helped to perceive the inflection point wherein including new clusters ceases to significantly improve the nice of the version. The analysis confirmed that the most efficient variety of clusters for this dataset is within the range of 1 to 14, as we are able to see at the graph.



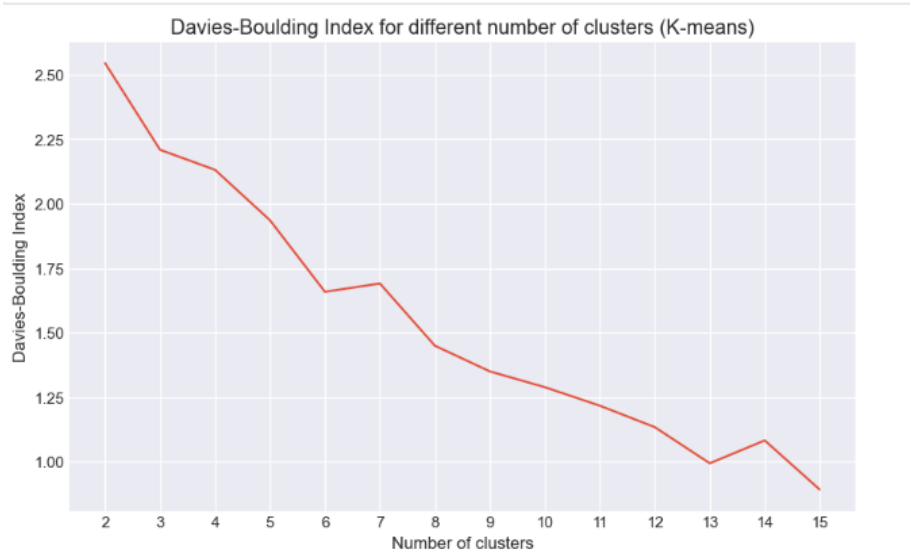
To verify the robustness of the results, additional experiments have been carried out with unique values of K. For example, when K=eight, the clusters showed a clean separation consistent with key metrics consisting of total order quantity, common test, and wide variety of objects. When the number of clusters became improved to ten and 12, smaller segments appeared, which allowed us to pick out unique consumer businesses, consisting of buyers with a excessive order frequency or customers targeted on certain product classes. However, an excessive boom within the number of clusters caused a decrease in the interpretability of the consequences, making the selection of K=eight the maximum affordable.

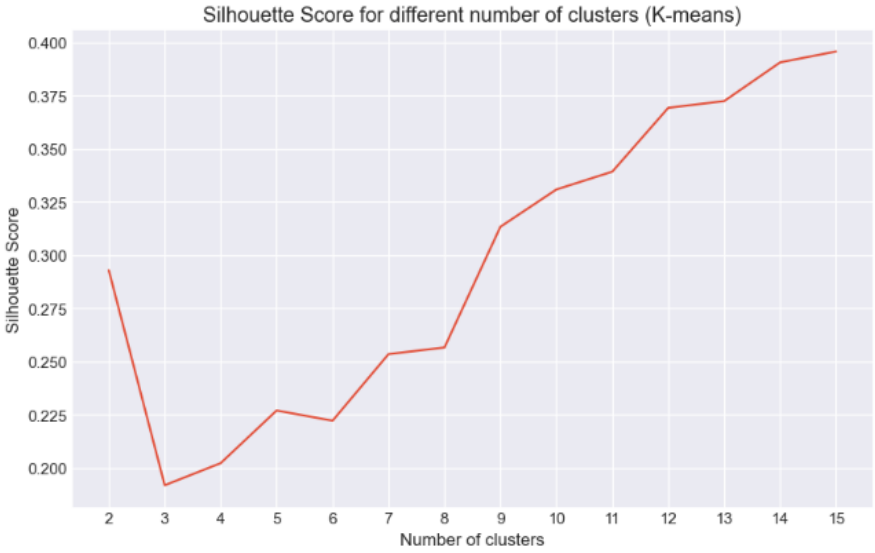
Before strolling the K-Means set of rules, the facts changed into further processed to do away with missing values. For this, the KNN imputation approach become used, which replaces missing values with the average values of the nearest acquaintances. This minimized the impact of missing values at the clustering results. After information imputation, the K-Means set of rules changed into educated on standardized numerical functions which includes general order amount, quantity of objects, and average test. Cluster centroids have been calculated because the common values of the points belonging to every cluster.

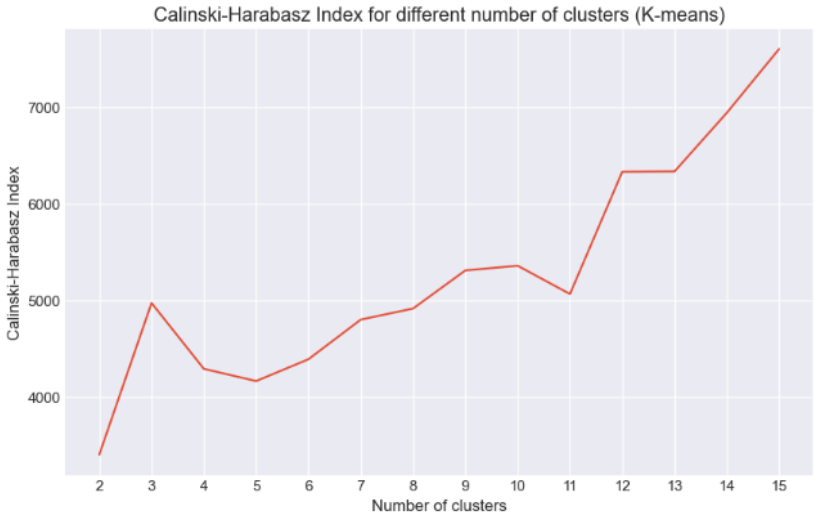
To visualize the clustering effects, size reduction methods inclusive of PCA (Principal Component Analysis) and SVD (Singular Value Decomposition) had been implemented. These methods allowed us to convert multidimensional information right into a two-dimensional space for a clear illustration of the clusters. The graphs show that the clusters are truely separated, even though a number of them have small overlaps, which can be because of noise or outliers within the facts. The visualization additionally confirmed that the clusters differ in key metrics along with local affiliation and possibilities in product categories.



After completing the clustering, quality metrics were calculated for each value of K. The silhouette index showed that at K=8, the metric value reaches its maximum, indicating good cluster separation. The Kalinski-Harabasz index also showed high values for K=8, confirming that the clusters are well separated from each other. The Davis-Boldin index, which tends to minimize at optimal separation, also confirmed that K=8 is the most appropriate choice.







The final dataset after clustering contained information about each customer's membership in one of the clusters. These labels were added to the original dataset, which allowed for a detailed analysis of the characteristics of each cluster. The clustering results were saved in tables, where each row corresponded to one customer, and the columns contained values for metrics such as total order amount, number of items, and cluster label. This provided a clear picture of customer behavior and prepared the data for further analysis.

5.5 Expectation-Maximization clustering

Data clustering using the Expectation-Maximization (EM) method has end up an important step in analysis aimed at figuring out hidden styles in patron behavior. This set of rules differs from K-Means in that it makes use of a probabilistic approach to determine the membership of records

points in clusters. Instead of rigidly dividing factors between clusters, the EM algorithm estimates the opportunity that every factor belongs to a particular cluster, which makes it especially beneficial for operating with heterogeneous facts.

The first step become to lessen the dimensionality of the data using the essential issue analysis (PCA) method. The original dataset contained a big variety of features, that can complicate the translation of the consequences and boom the execution time of the set of rules. To cope with this trouble, PCA was carried out, which allowed the records to be transformed into a area with fewer dimensions at the same time as keeping the maximum variance of the unique information. After applying PCA, the entire defined variance reached one hundred percent, confirming the effectiveness of the approach for this project.

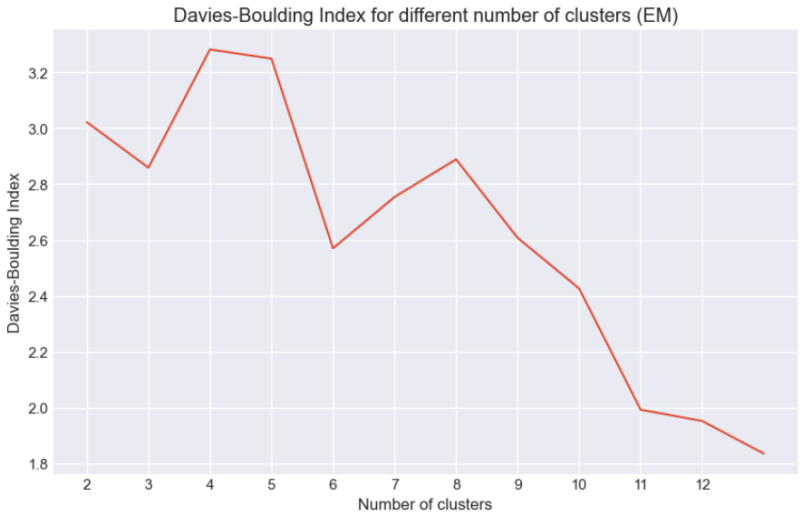
To prepare the records for clustering, a pipeline changed into created that covered 3 essential degrees: filling in missing values, normalizing numerical capabilities, and making use of PCA. Missing values had been crammed the usage of the suggest cost approach, which minimized the effect of lacking values at the analysis consequences. Data normalization turned into performed to make sure that all features had an identical affect on the clustering method. After those steps, the records was transformed the use of PCA, which decreased the dimension to two components for visualization of the consequences.

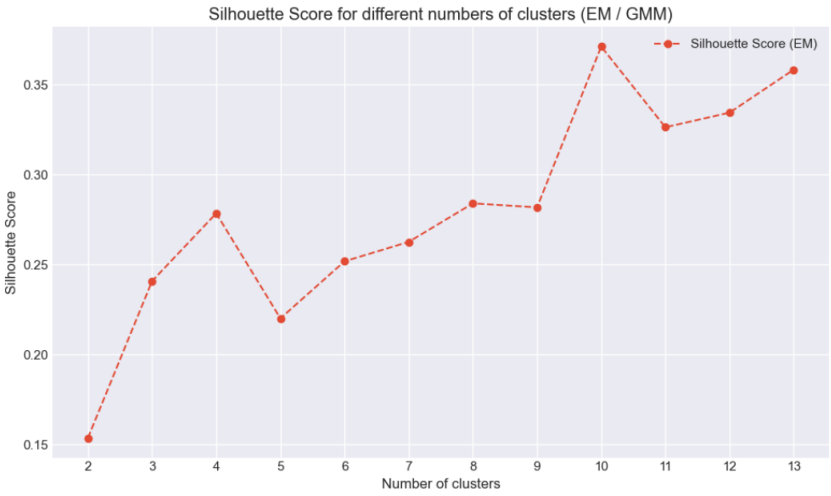
The next step changed into to train the Gaussian Mixture Model (GMM), that is an implementation of the EM algorithm. To do that, numerous values for the variety of clusters had been examined, ranging from 2 to 12. The pleasant of clustering become evaluated the use of metrics such as the silhouette index, the Kalinski-Harabasz index, and the Davis-Boldin index. The evaluation showed that the highest quality number of clusters is inside the range of 8 to 12, similar to the results received the use of K-Means.

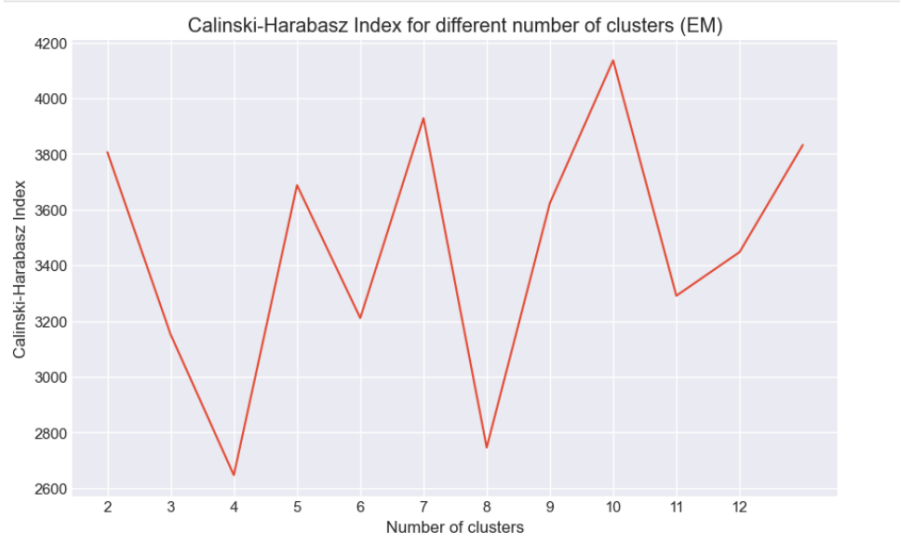
The clustering effects have been visualized the usage of graphs primarily based on the first two foremost additives. The graphs show that the clusters are truly separated, even though a number of them have small overlaps, which can be because of the probabilistic nature of the EM algorithm. Unlike K-Means, in which each factor belongs uniquely to simplest one cluster, the EM algorithm permits each point to have possibilities of belonging to all clusters. This is mainly beneficial for reading clients who showcase mixed conduct styles.

After clustering turned into completed, excellent metrics were calculated for every cluster count price. The silhouette index confirmed that the metric cost reaches its maximum at 8 clusters, indicating correct cluster separation. The Kalinski-Harabasa index additionally showed excessive values for 8 clusters, confirming that the clusters are properly separated from every different. The Davis-Boldin index, which tends to limit at most appropriate separation, additionally confirmed

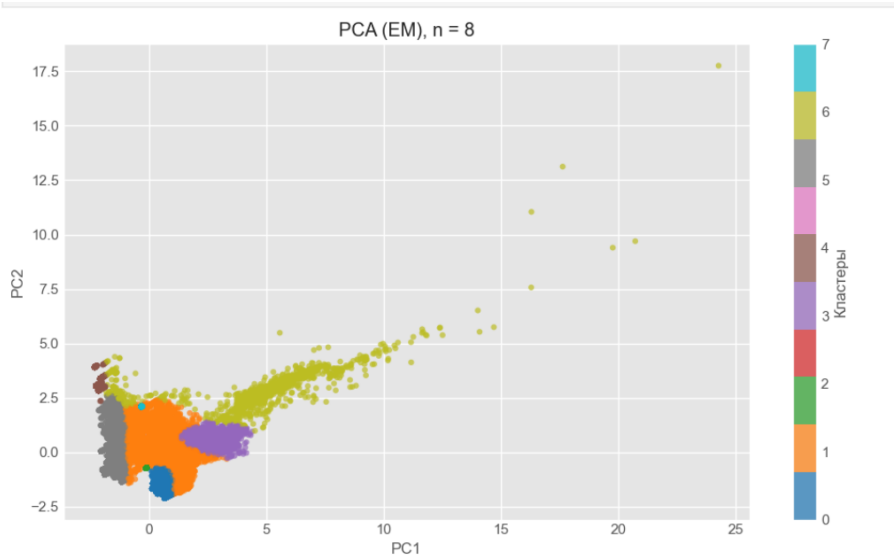
that 8 clusters are the maximum appropriate preference.

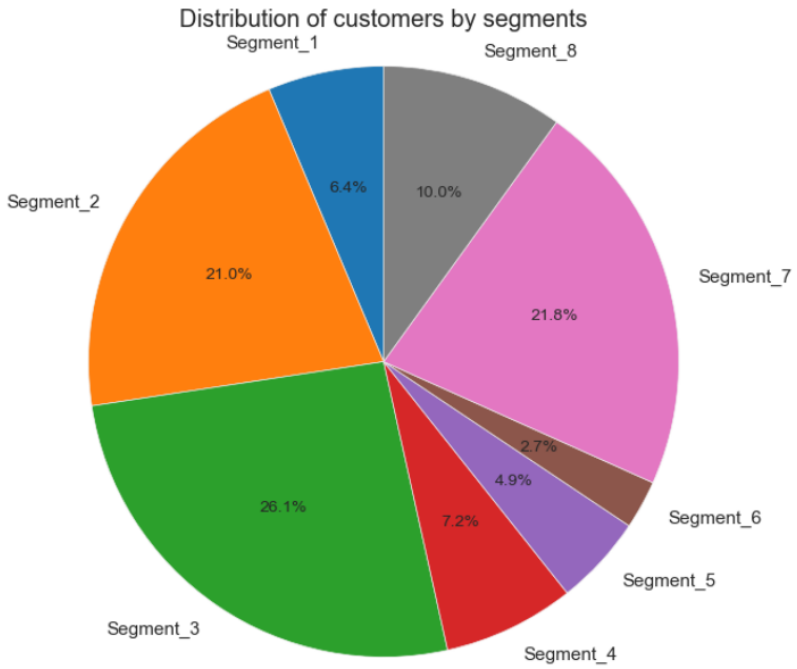


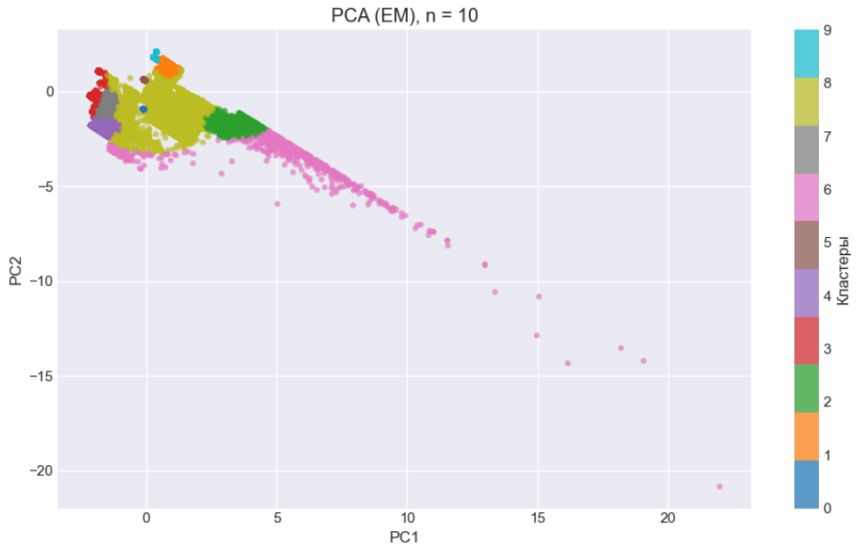


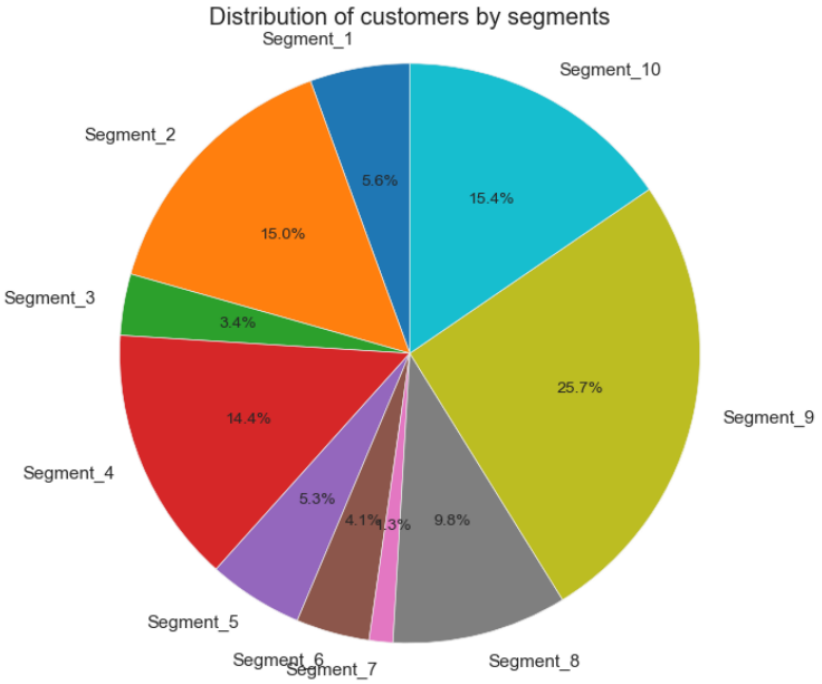


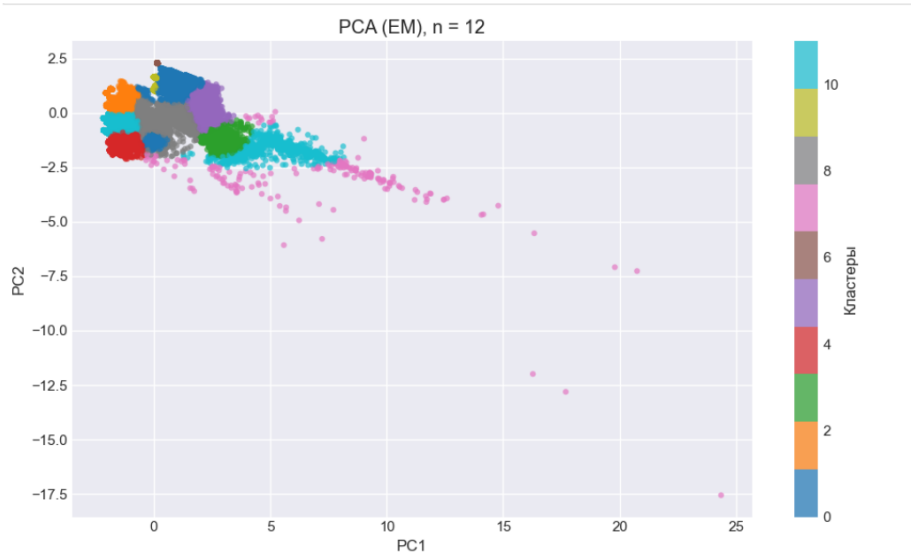
At this stage, a relatively clear visualization of each number of clusters was developed. It includes a pie chart and a PCA graph.

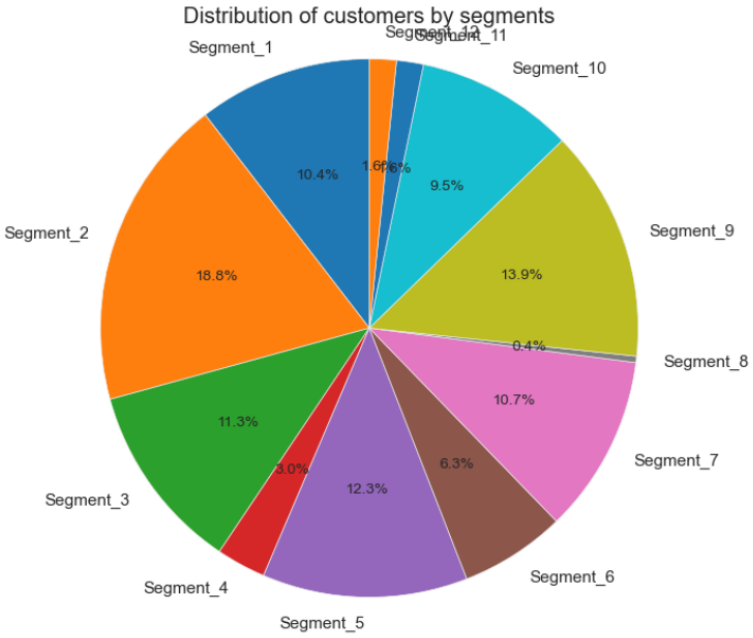





PCA (EM), n = 10
PC2
PC1
Кластеры
0
−5
−10
−15
−20
0
5
10
15
20
9
8
7
6
5
4
3
2
1
0


Distribution of customers by segments
Segment_1
Segment_10
Segment_2
Segment_3
Segment_4
Segment_5
Segment_6
Segment_7
Segment_8
Segment_9
5.6%
15.4%
15.0%
3.4%
25.7%
14.4%
5.3%
4.1%
1.3%
9.8%





The final dataset after clustering contained information about each customer's membership in one of the clusters. These labels were added to the original dataset, which allowed for a detailed analysis of the characteristics of each cluster. The clustering results were saved in tables, where each row corresponded to one customer, and the columns contained values for metrics such as total order amount, number of items, and cluster label. This provided a clear picture of customer behavior and prepared the data for further analysis.

5.6 Market Basket Analysis

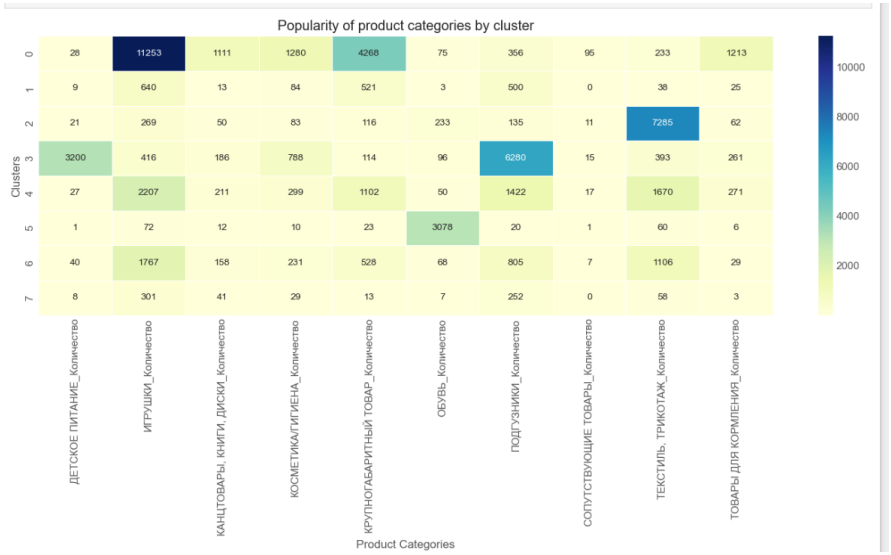
Market basket analysis became an important stage in the research aimed at identifying associative links between products that customers buy together. This method allows not only to understand customer preferences, but also to develop effective marketing strategies, such as cross-selling and recommendation systems. To perform the analysis, a dataset containing order information was used, including attributes such as order number, product categories, number of units sold, and regional affiliation.

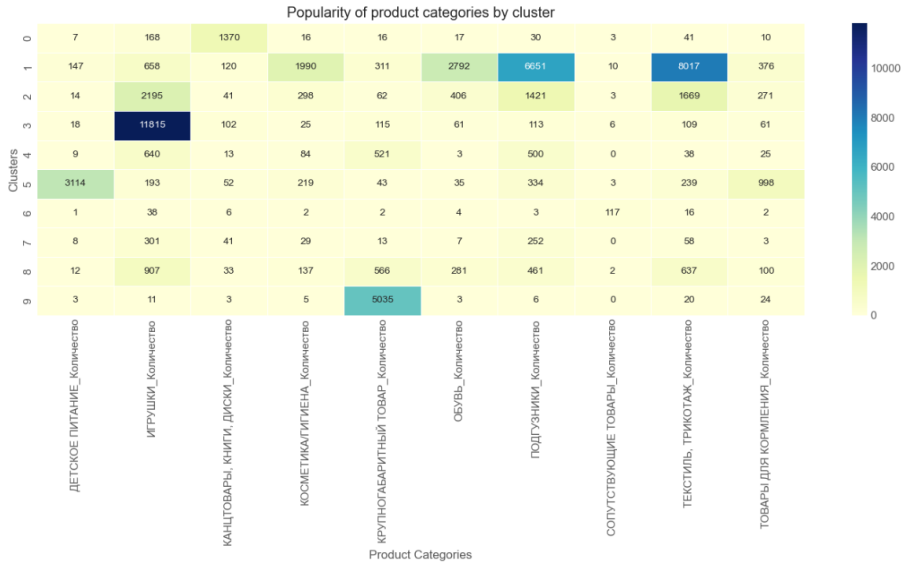
The first step was to examine the data structure. The original dataset contained 223,975 rows and 21 columns covering attributes such as order date, order amount, region, product categories (Группа2, Группа3), and number of units sold. However, some of the data was redundant for market basket analysis, so unnecessary columns, such as СуммаДокумента, which did not affect the analysis results, were removed. After that, the data was converted into an order matrix, where the rows corresponded to unique orders and the columns corresponded to product categories. The values in the matrix reflected the number of units sold for each product within a single order.

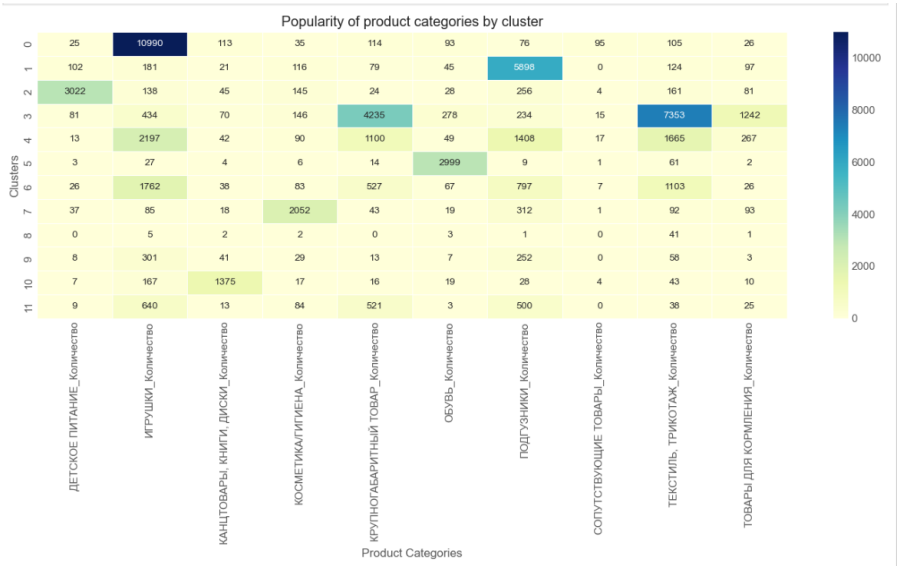
To simplify the analysis, the values in the matrix were converted to binary format: if the product was present in the order, the value was set to 1, otherwise it was set to 0. This conversion allowed us to focus on the fact that the product was present in the order, ignoring its quantity. The resulting matrix became the basis for applying the Apriori algorithm, which is used to identify frequent sets of items. The algorithm was configured with a minimum support (min support) of 0.01, which meant that an item or set of items had to appear in at least 1 percent of all orders to be considered significant.

The analysis results revealed numerous exciting styles. For example, “Diapers” and “Детское питание” were often determined collectively, confirming the logical connection among those categories. Other popular pairs blanketed “Игрушки” and “Косметика/Гигиена,” which could suggest purchases for children and their parents. These findings fashioned the basis for guidelines on growing customized offers for clients, along with discounts on related product categories.

To assess the strength of associative links, metrics such as confidence and lift were calculated. Confidence shows the probability that when one product is purchased, another will be purchased, while lift reflects how much higher the probability of this combination is than random chance. For example, the combination of “Подгузники” and “Детское питание” had a high level of confidence and lift, confirming their strong association. These metrics made it possible to rank associations by their significance and use the strongest ones for marketing campaigns.


Popularity of product categories by cluster
Clusters
0 28 11253 1111 1280 4268 75 356 95 233 1213
1 9 640 13 84 521 3 500 0 38 25
2 21 269 50 83 116 233 135 11 7285 62
3 3200 416 186 788 114 96 6280 15 393 261
4 27 2207 211 299 1102 50 1422 17 1670 271
5 1 72 12 10 23 3078 20 1 60 6
6 40 1767 158 231 528 68 805 7 1106 29
7 8 301 41 29 13 7 252 0 58 3
ДЕТСКОЕ ПИТАНИЕ_Количество
ИГРУШКИ_Количество
КАНЦТОВАРЫ, КНИГИ, ДИСКИ_Количество
КОСМЕТИКА/ГИГИЕНА_Количество
КРУПНОГАБАРИТНЫЙ ТОВАР_Количество
ОБУВЬ_Количество
ПОДГУЗНИКИ_Количество
СОПУТСТВУЮЩИЕ ТОВАРЫ_Количество
ТЕКСТИЛЬ, ТРИКОТАЖ_Количество
ТОВАРЫ ДЛЯ КОРМЛЕНИЯ_Количество
Product Categories
10000
8000
6000
4000
2000


Popularity of product categories by cluster
Clusters
0 7 168 1370 16 16 17 30 3 41 10
1 147 658 120 1990 311 2792 6651 10 8017 376
2 14 2195 41 298 62 406 1421 3 1669 271
3 18 11815 102 25 115 61 113 6 109 61
4 9 640 13 84 521 3 500 0 38 25
5 3114 193 52 219 43 35 334 3 239 998
6 1 38 6 2 2 4 3 117 16 2
7 8 301 41 29 13 7 252 0 58 3
8 12 907 33 137 566 281 461 2 637 100
9 3 11 3 5 5035 3 6 0 20 24
ДЕТСКОЕ ПИТАНИЕ_Количество
ИГРУШКИ_Количество
КАНЦТОВАРЫ, КНИГИ, ДИСКИ_Количество
КОСМЕТИКА/ГИГИЕНА_Количество
КРУПНОГАБАРИТНЫЙ ТОВАР_Количество
ОБУВЬ_Количество
ПОДГУЗНИКИ_Количество
СОПУТСТВУЮЩИЕ ТОВАРЫ_Количество
ТЕКСТИЛЬ, ТРИКОТАЖ_Количество
ТОВАРЫ ДЛЯ КОРМЛЕНИЯ_Количество
Product Categories
10000
8000
6000
4000
2000
0


Popularity of product categories by cluster
Clusters
0 25 10990 113 35 114 93 76 95 105 26
1 102 181 21 116 79 45 5898 0 124 97
2 3022 138 45 145 24 28 256 4 161 81
3 81 434 70 146 4235 278 234 15 7353 1242
4 13 2197 42 90 1100 49 1408 17 1665 267
5 3 27 4 6 14 2999 9 1 61 2
6 26 1762 38 83 527 67 797 7 1103 26
7 37 85 18 2052 43 19 312 1 92 93
8 0 5 2 2 0 3 1 0 41 1
9 8 301 41 29 13 7 252 0 58 3
10 7 167 1375 17 16 19 28 4 43 10
11 9 640 13 84 521 3 500 0 38 25
ДЕТСКОЕ ПИТАНИЕ_Количество
ИГРУШКИ_Количество
КАНЦТОВАРЫ, КНИГИ, ДИСКИ_Количество
КОСМЕТИКА/ГИГИЕНА_Количество
КРУПНОГАБАРИТНЫЙ ТОВАР_Количество
ОБУВЬ_Количество
ПОДГУЗНИКИ_Количество
СОПУТСТВУЮЩИЕ ТОВАРЫ_Количество
ТЕКСТИЛЬ, ТРИКОТАЖ_Количество
ТОВАРЫ ДЛЯ КОРМЛЕНИЯ_Количество
Product Categories
10000
8000
6000
4000
2000
0

5.7 Clustering results

Clustering using the K-Means and EM algorithms allowed us to divide the customer base into eight main groups. Each group is characterized by certain features that help to better understand customer needs and develop personalized marketing strategies. The main conclusions for each cluster can be summarized as follows:

Clusters with a high interest in children's products (e.g., clusters 0, 3, 4, 6): These customers mainly buy baby food and diapers, which indicates their focus on children's products. These are likely to be young families or parents who regularly order products for children. They are less likely to buy other categories of goods, such as toys, books, discs, cosmetics/hygiene products, large items, shoes, related goods, and textiles. For this group, it is recommended to develop special offers for children's goods, such as discounts on baby food or diapers, as well as offers for additional goods related to children's goods.

Clusters with uniform needs across all product categories (e.g., clusters 1, 5, 7): These customers do not show a clear preference for any one product category. They place orders once a month or less, and their choice of products depends on current needs rather than consistent habits. For this group, it is recommended to use situational promotions and personalized recommendations based on an analysis of recent purchases.

Clusters with low interest in all product categories: These customers may be new users of the platform or buyers who place orders very rarely. Their behavior may be related to seasonal or situational factors, such as holiday promotions or special offers. For this group, it is important to stimulate activity through attractive offers, such as discounts on the first order or bonuses for registration.

6 Conclusions and results

The research objective was achieved. Using the example of the dataset, it was possible to establish connections between the elements of the clusters and draw corresponding conclusions about each of them.

7 References

[1] Jain, A. K. Data clustering: 50 years beyond K-means Pattern Recognition Letters 2010

[2] McLachlan, G., Peel, D.. Finite Mixture Models 2000

[3] Abdi, H., Williams, L. J. Principal component analysis 2010

[4] Agrawal, R., Imieliński, T., Swami, A. Mining association rules between sets of items in large databases . Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data 1993

[5] Han, J., Pei, J., Yin, Y. Mining frequent patterns without candidate generation: A frequent-pattern tree approach . Data Mining and Knowledge Discovery 2004

[6] Tan, P. N., Steinbach, M., Kumar, V. (2005). Introduction to Data Mining . Pearson Education

[7] Kotz, D., Gray, R. S., Liu, J., Nekovee, M. Customer Segmentation and Market Basket Analysis in Retail . International Journal of Retail Analytics 2019

[8] Chen, M., Mao, S., Liu, Y. Big data analysis: Applications in retail and consumer behavior . Springer Handbook of Marketing Analytics 2014