

ПРОЕКТ: РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА для женской одежды с использованием кластеризации.

Чубарова Татьяна, 332 группа

Датсет состоит из:

- **Unnamed: 0** – индекс или идентификатор строки, добавленный при импорте данных.
- **Clothing ID** – идентификатор конкретного товара. Каждый уникальный ID соответствует отдельной позиции одежды.
- **Age** – возраст пользователя.
- **Title** – заголовок отзыва.
- **Review Text** – текст отзыва.
- **Rating** – рейтинг, поставленный пользователем от 1 до 5.
- **Recommended IND** – бинарная переменная (1 рекомендует товар, а 0 – не порекомендует).
- **Positive Feedback Count** – количество пользователей, которые оценили отзыв как полезный.
- **Division Name** – категории, к которому принадлежит товар (женская одежда, аксессуары).
- **Department Name** – отдел, к которому относится товар (платья, футболки, джинсы).
- **Class Name** – подкатегории товара (casual wear, formal wear)

Предварительная обработка данных. Изучение структуры

- Все числовые столбцы (Clothing ID, Age, Rating, Recommended IND, Positive Feedback Count) имеют тип int64.
- Текстовые столбцы (Title, Review Text, Division Name, Department Name, Class Name) имеют тип object.

Unnamed: 0	int64
Clothing ID	int64
Age	int64
Title	object
Review Text	object
Rating	int64
Recommended IND	int64
Positive Feedback Count	int64
Division Name	object
Department Name	object
Class Name	object

- Title: 3810 пропущенных значений (16.2%).
- Review Text: 845 пропущенных значений (3.6%).
- Division Name, Department Name, Class Name: по 14 пропущенных значений (менее 0.1%).

Unnamed: 0	0
Clothing ID	0
Age	0
Title	3810
Review Text	845
Rating	0
Recommended IND	0
Positive Feedback Count	0
Division Name	14
Department Name	14
Class Name	14

Предварительная обработка данных. Изучение структуры

- Обработка пропущенных значений (замена).
 - Заполняем пропущенные значения в столбце Title значением 'Без заголовка'.
 - Заполняем пропущенные значения в столбце Review Text заполнить пропущенные значения значением 'Без отзыва'.
- Проверка на пропущенные значения.

Unnamed:	0	0
Clothing ID		0
Age		0
Title		0
Review Text		0
Rating		0
Recommended IND		0
Positive Feedback Count		0
Division Name		0
Department Name		0
Class Name		0

Предварительная обработка данных. Анализ признаков

- Числовые признаки Clothing ID, Age, Rating, Recommended IND, Positive Feedback Count можно использовать напрямую.
- Категориальные признаки Division Name, Department Name, Class Name обрабатываю с помощью **Target Encoding**, причем , не потеряв возможность восстановить оригинальные названия (словарь).
- Текстовые признаки:
 - Title заменяю на новый признак Title Length, который показывает количество слов в заголовке.
 - К признаку Review Text применяем анализ настроений с использованием **VADER (Valence Aware Dictionary and sEntiment Reasoner)**.

Предварительная обработка данных. Target Encoding

Target Encoding – это метод кодирования категориальных признаков, который преобразовывает категориальные данные в числовые, что делает их более пригодными для использования в моделях машинного обучения.

Суть метода заключается в том, чтобы для каждой категории в категориальном признаке вычислить среднее значение целевой переменной (Rating), связанное с каждой категорией. Затем это среднее значение используется в качестве нового представления категории.

Также я создаю обратный словарь, чтобы суметь вернуться к первоначальным значениям.

Проверка: Обратный словарь: Division Name

Average Rating (key)	Names (value)
4.18	General
4.21	General Petite
4.29	Initmates

Проверка: словарь: Department Name

Average Rating (key)	Names (value)
4.29	Bottoms
4.15	Dresses
4.28	Intimate
4.26	Jackets
4.17	Tops
3.82	Trend

Предварительная обработка данных. VADER

Valence Aware Dictionary and sEntiment Reasoner - это инструмент для анализа

(Словарь с учетом валентности и анализатор настроений)

настроений текста, который используется в обработке естественного языка (NLP).

VADER является одним из популярных инструментов для извлечения настроения из текстов, таких как **отзывы, твитты, сообщения в социальных сетях** и другие короткие тексты.

Как работает VADER?

VADER использует словарь, в котором каждому слову присваивается базовое значение настроения, причем он обращает внимание:

1. Слова-усилители и слова-ослабители
2. Эмодзи

Предварительная обработка данных. Итоговый результат

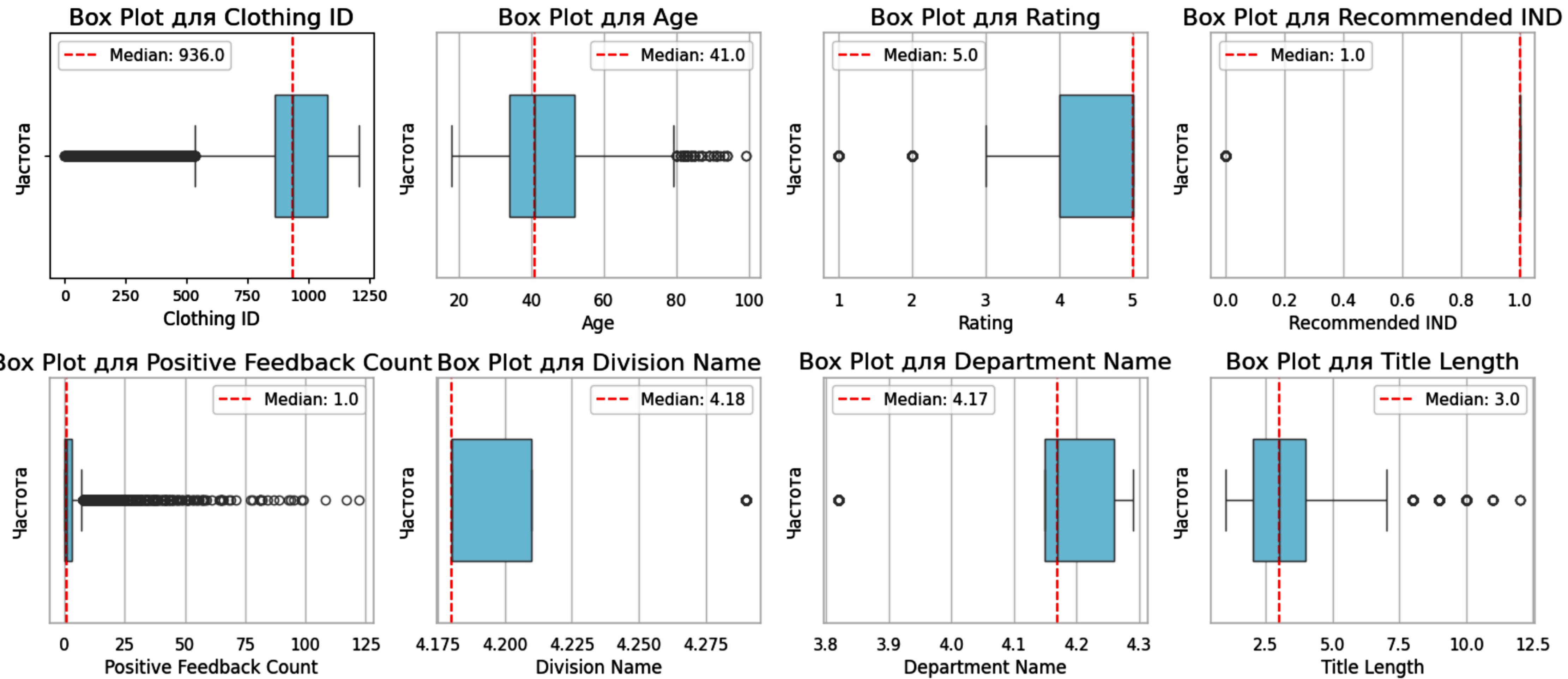
(23472, 10)

```
Index(['Clothing ID', 'Age', 'Rating', 'Recommended IND',
       'Positive Feedback Count', 'Division Name', 'Department Name',
       'Title Length', 'Sentiment', 'Sentiment_Label'],
      dtype='object')
```

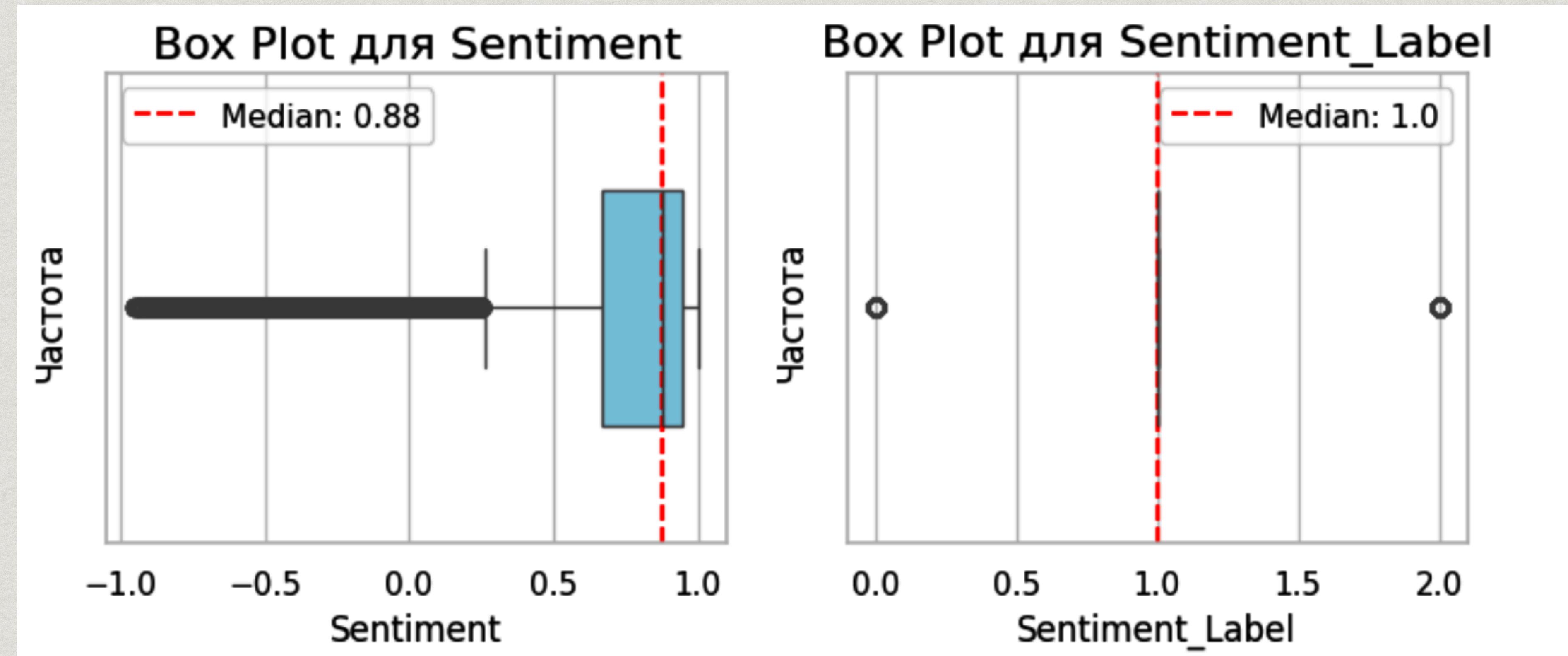
```
Clothing ID      0
Age              0
Rating            0
Recommended IND   0
Positive Feedback Count  0
Division Name     0
Department Name    0
Title Length       0
Sentiment          0
Sentiment_Label    0
dtype: int64
```

	Clothing ID	Age	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Title Length	Sentiment	Sentiment_Label
0	767	33	4	1	0	4.29	4.28	2	0.89	1
1	1080	34	5	1	4	4.18	4.15	2	0.97	1
2	1077	60	3	0	0	4.18	4.15	4	0.92	1
3	1049	50	5	1	0	4.21	4.29	3	0.57	1
4	847	47	5	1	6	4.18	4.17	2	0.93	1

Анализ данных и визуализация. Распределение признаков



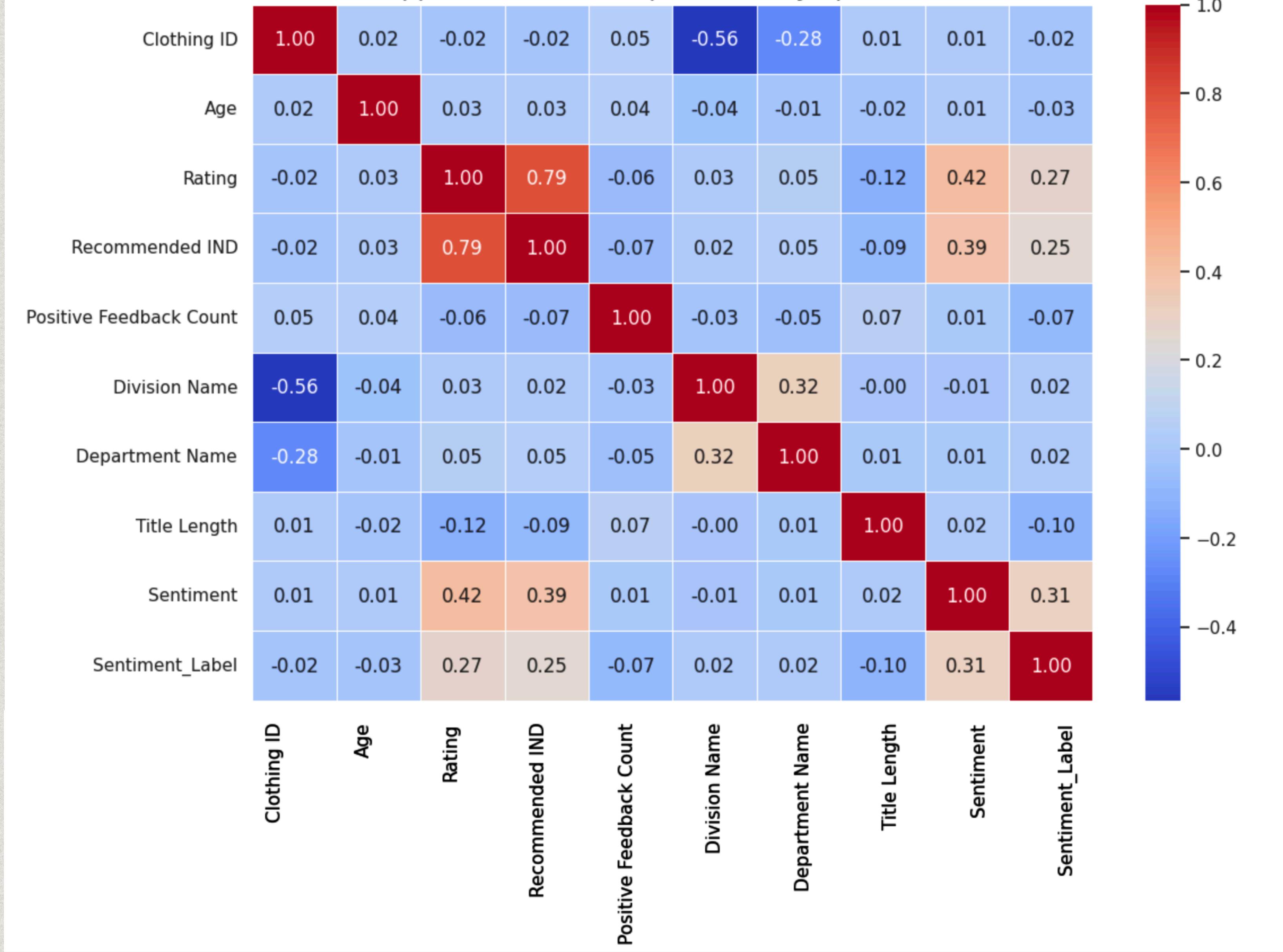
Анализ данных и визуализация. Распределение признаков



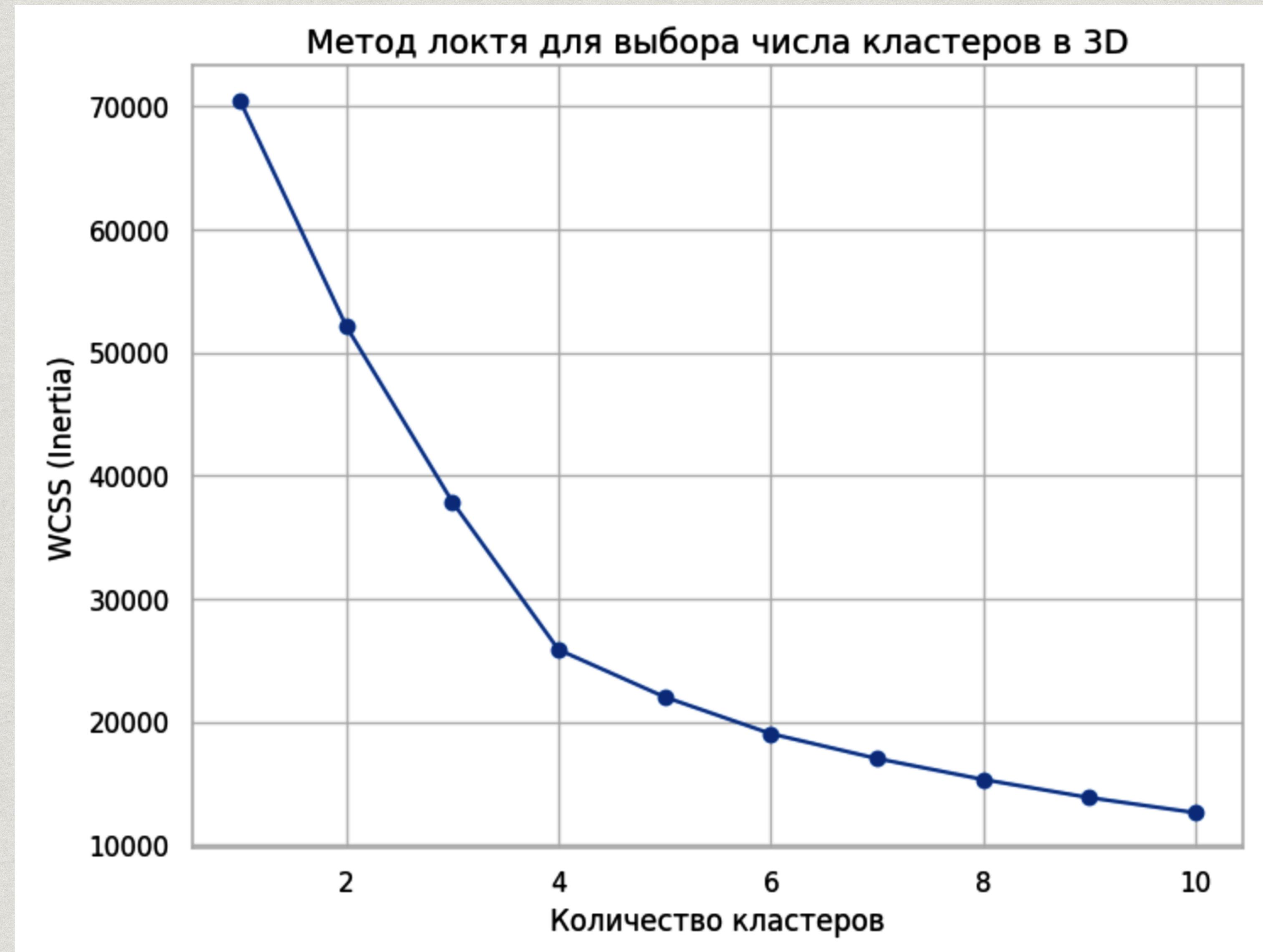
Sentiment представляет собой числовое значение, которое обычно вычисляется с помощью инструментов анализа настроений ($-1 < \text{Sentiment} < 1$).

Sentiment_Label – это категориальная переменная, которая часто используется для упрощённой классификации текста по настроению ($\text{Sentiment_Label} = 1$ (полож. настроение) или -1 (отриц.) или 0 (нейтральное)).

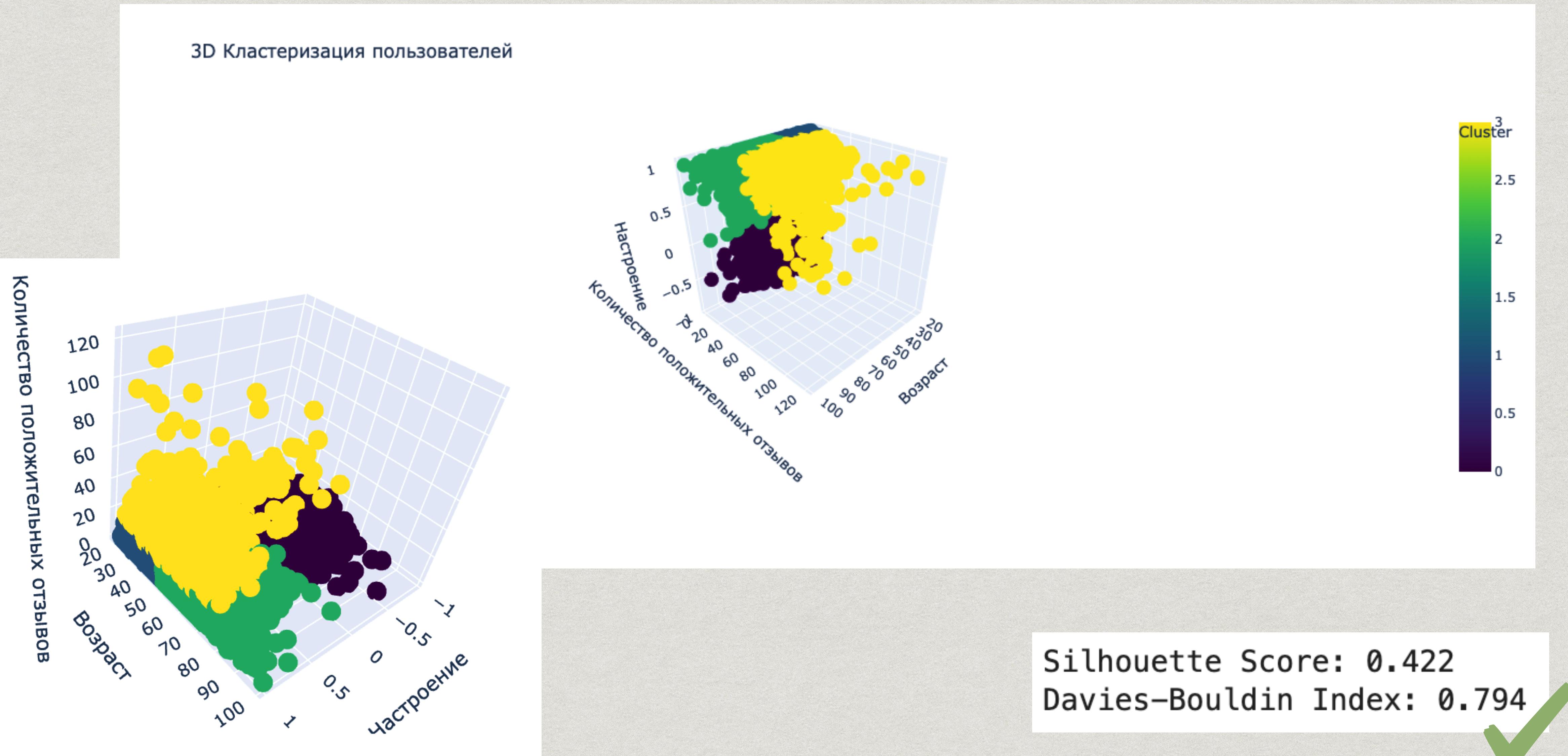
Корреляционная матрица между признаками



КЛАСТЕРИЗАЦИЯ. МЕТОД ЛОКТА



3D КЛАСТЕРИЗАЦИЯ ПОЛЬЗОВАТЕЛЕЙ.



КЛАСТЕРИЗАЦИЯ. КЛАСТЕРЫ.

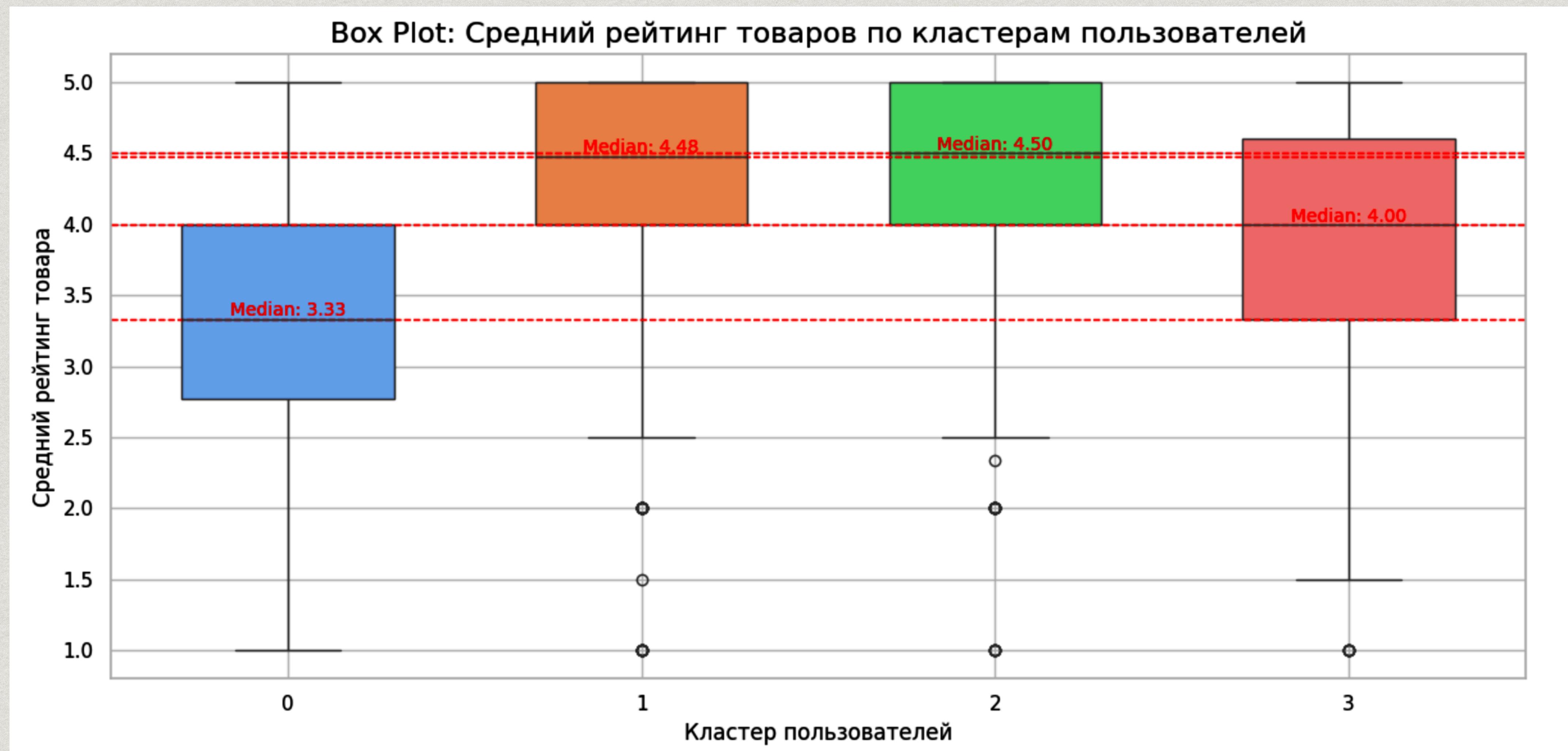
Присвоили каждому пользователю кластер:

Clothing ID	Age	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Title Length	Sentiment	Sentiment_Label	Cluster
0	767	33	4	1	0	4.29	4.28	2	0.89	1 1
1	1080	34	5	1	4	4.18	4.15	2	0.97	1 1
2	1077	60	3	0	0	4.18	4.15	4	0.92	1 2
3	1049	50	5	1	0	4.21	4.29	3	0.57	1 2
4	847	47	5	1	6	4.18	4.17	2	0.93	1 2

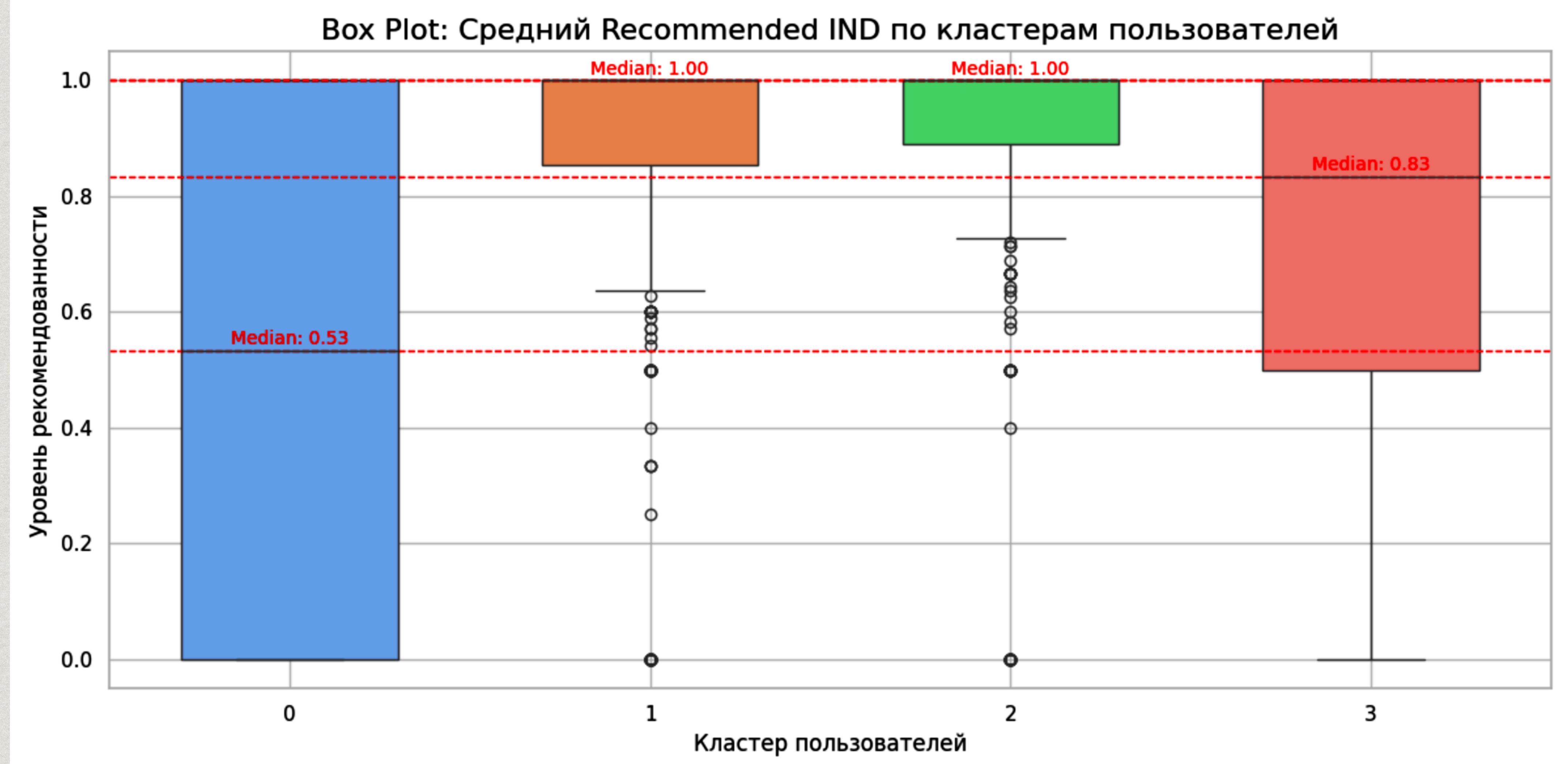
Получаем следующие аудитории:

	Cluster	Age	Positive Feedback Count	Sentiment
0	0.00	42.14	1.57	-0.13
1	1.00	35.41	1.60	0.85
2	2.00	56.86	1.98	0.84
3	3.00	44.56	24.52	0.76

Задача 1: Оценить, как пользователи с различными характеристиками взаимодействуют с продуктами.



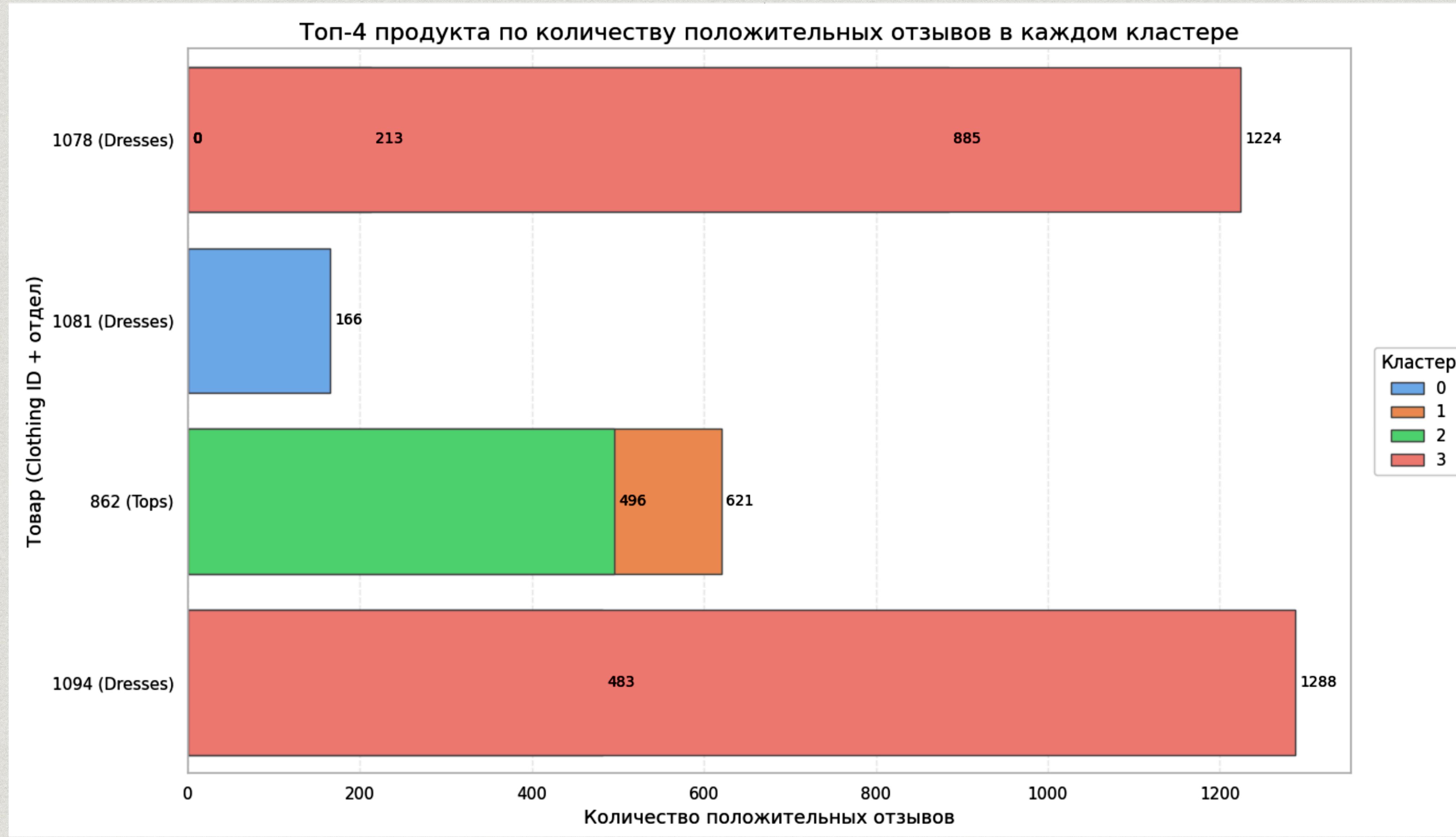
Задача 1: Оценить, как пользователи с различными характеристиками взаимодействуют с продуктами.



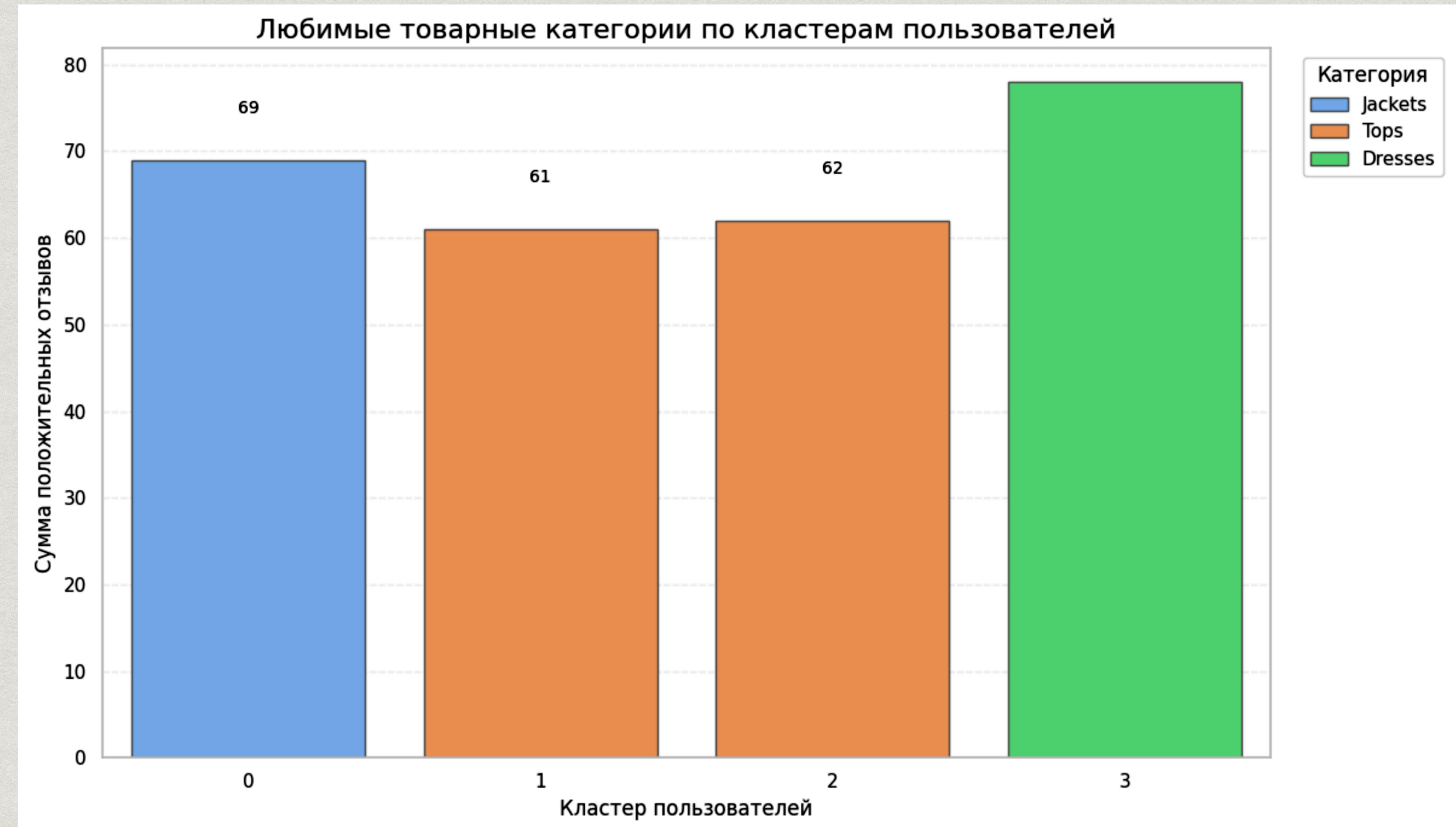
Задача 2: Оценить, какие товары наиболее привлекательны для разных групп пользователей.

Cluster	Clothing ID	feedback_count	avg_rating	Division Name	Department Name
0	0	1078	213	3.523438	General Dresses
1	0	1078	213	3.523438	General Petite Dresses
2	0	1081	166	3.519481	General Dresses
3	0	1081	166	3.519481	General Petite Dresses
4	1	1078	885	4.271739	General Petite Dresses
5	1	1078	885	4.271739	General Dresses
6	1	862	621	4.371981	General Tops
7	1	862	621	4.371981	General Petite Tops
8	2	862	496	4.388235	General Tops
9	2	862	496	4.388235	General Petite Tops
10	2	1094	483	4.208738	General Petite Dresses
11	2	1094	483	4.208738	General Dresses
12	3	1094	1288	4.212766	General Dresses
13	3	1094	1288	4.212766	General Petite Dresses
14	3	1078	1224	4.409091	General Petite Dresses
15	3	1078	1224	4.409091	General Dresses

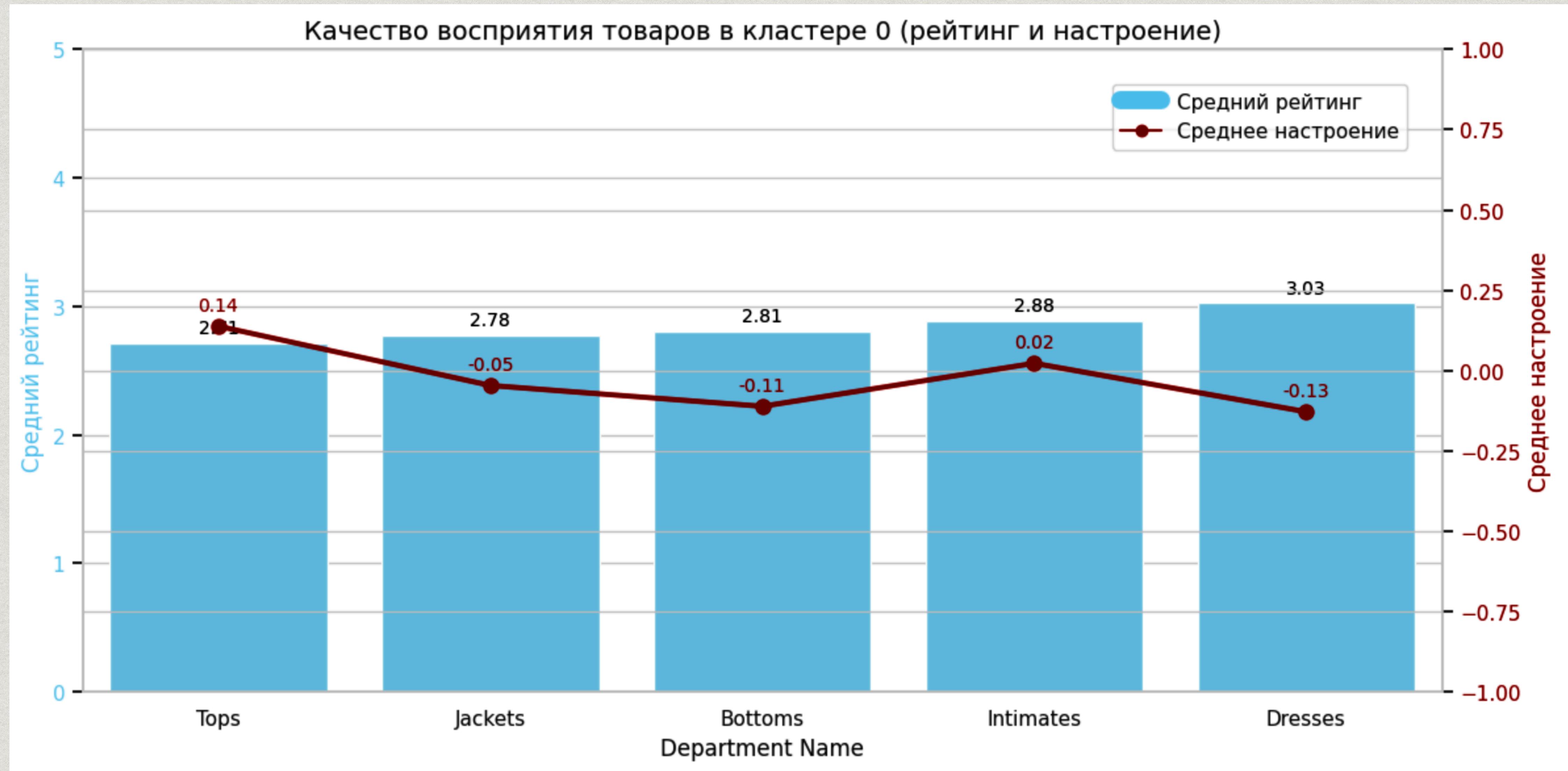
Задача 2: Оценить, какие товары наиболее привлекательны для разных групп пользователей.



Задача 3. Определим любимую товарную категорию `Department Name` для каждого кластера

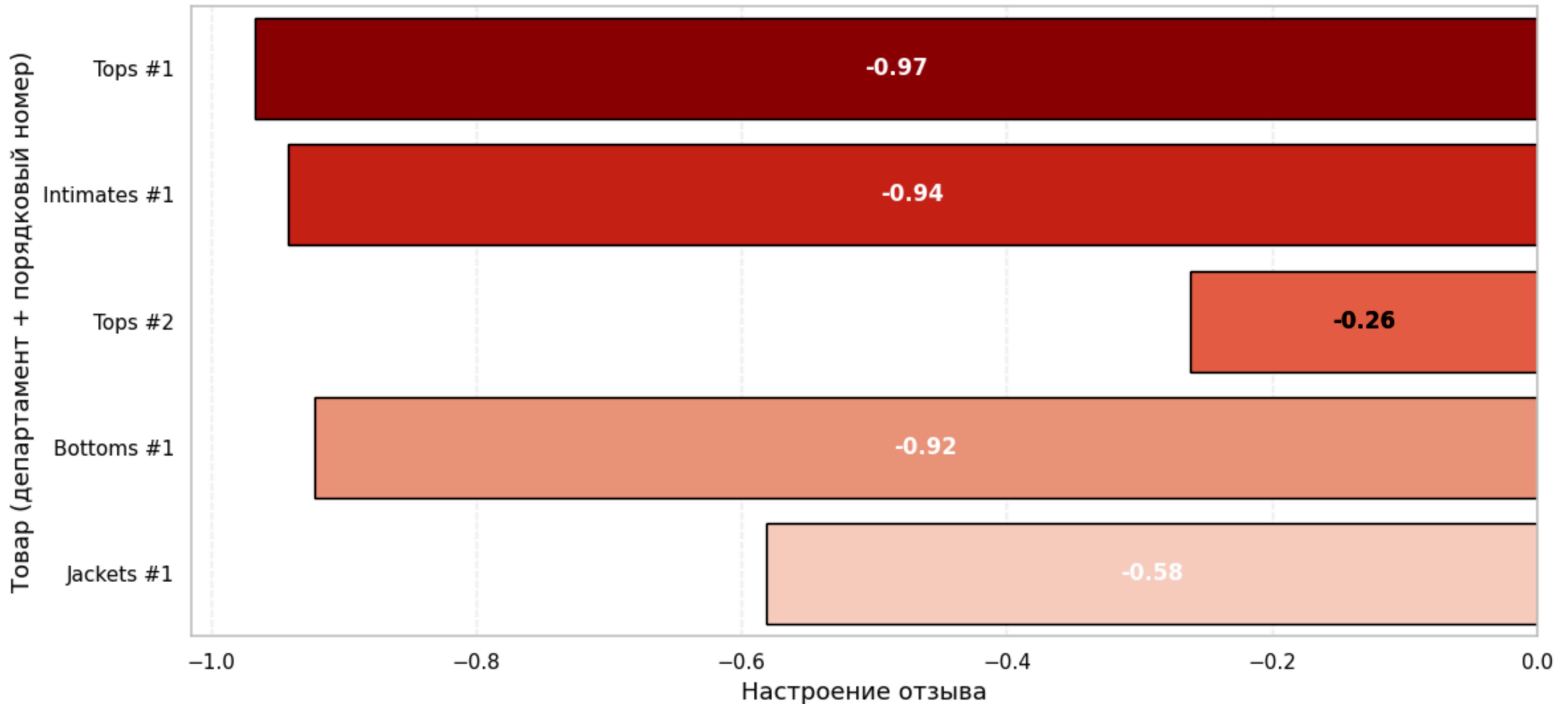


Задача 3. Определим любимую товарную категорию `Department Name` для каждого кластера



Дополнительно к задаче 3. Определить конкретные товары из кластера 0, которые вызывают негатив.

Топ-5 самых негативных товаров в кластере 0 (по настроению)

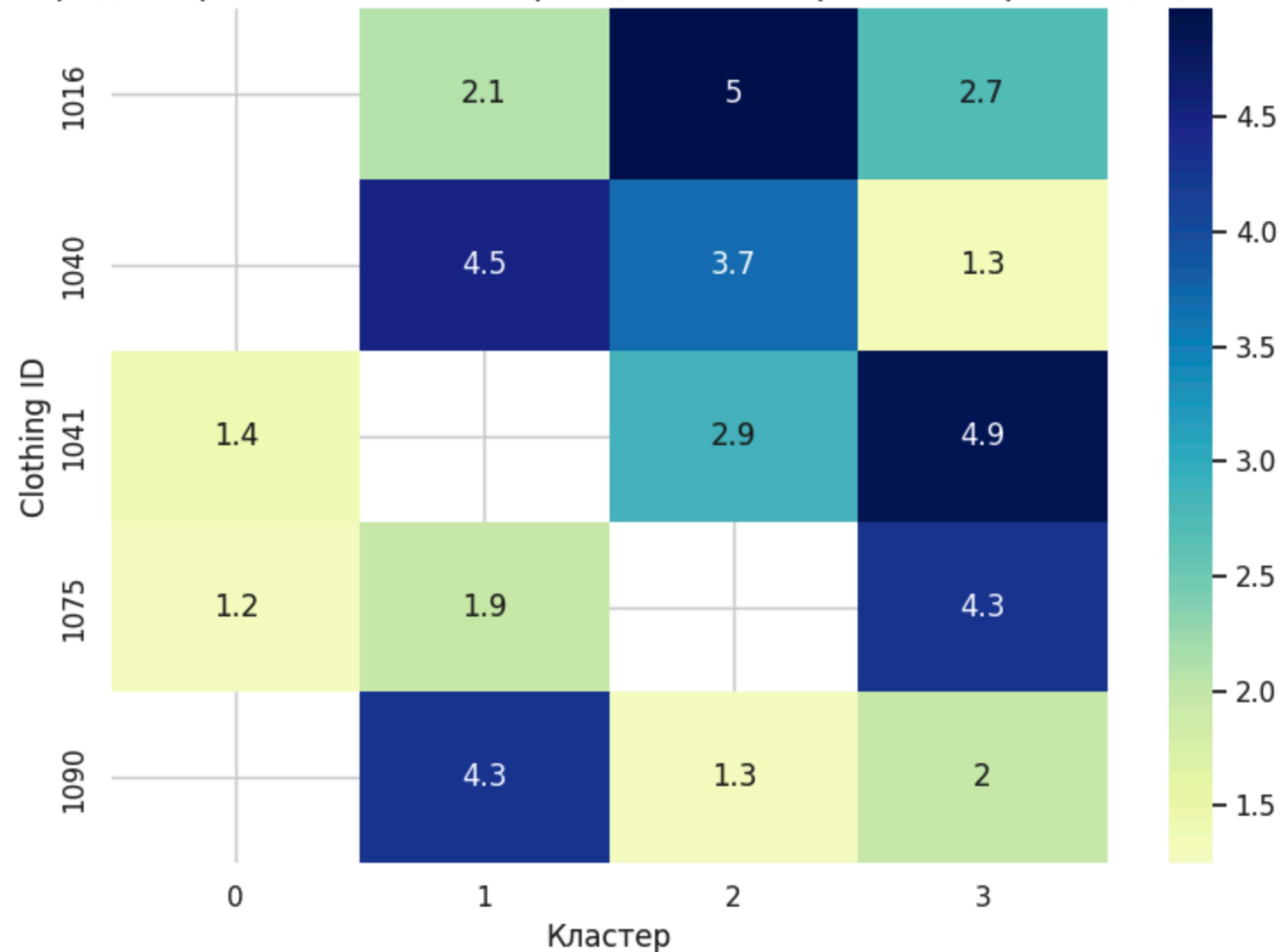


Задание 4. Выделите товары с наибольшей разницей между кластерами по количеству отзывов или рейтингу

Clothing ID	max_min_diff_feedback	max_min_diff_rating
17	1017	14.0
78	1078	11.0
41	1041	9.0
54	1054	9.0
31	1031	9.0

Задание 4.

Средний рейтинг по кластерам (топ-5 товаров с max разницей)



МОДЕЛЬ CatBoostClassifier

CatBoostClassifier - это метод градиентного бустинга, оптимизированный для работы с категориальными признаками.



Создаём рекомендательную модель, которая предсказывает, стоит ли рекомендовать/не рекомендовать определённый товар конкретному сегменту пользователей (кластеру).

Знаем, что

- * пользователи из разных кластеров по-разному воспринимают одни и те же товары;
- * товары с хорошей оценкой в одном сегменте могут быть токсичны в другом.

И также возьмем во внимание:

- * отзывов пользователя,
- * частоту взаимодействия с определёнными категориями,

МОДЕЛЬ CatBoostClassifier

Label	Precision	Recall	F1-score	Support
0	0.74	0.24	0.36	83
1	0.85	0.98	0.91	374
accuracy	0.85	0.85	0.85	0.846827
macro avg	0.8	0.61	0.64	457
weighted avg	0.83	0.85	0.81	457

Улучшение модели CatBoostClassifier

1. Найдем лучшие гиперпараметры с помощью RandomizedSearchCV:

```
Best Params: {'learning_rate': 0.1, 'l2_leaf_reg': 1, 'iterations': 200, 'depth': 6}
```

2. Улучшаем таргет, так чтобы:

```
recommend
1    0.817784
0    0.182216
Name: proportion, dtype: float64
```



```
recommend
0    0.746386
1    0.253614
Name: proportion, dtype: float64
```

Улучшение модели CatBoostClassifier

	precision	recall	f1-score	support
0	0.98	0.96	0.97	341
1	0.89	0.93	0.91	116
accuracy	0.95	0.95	0.95	0.95
macro avg	0.93	0.94	0.94	457
weighted avg	0.95	0.95	0.95	457

Получаем модель с высокой точностью поочностию: $\text{`recall'} = \frac{TP}{TP + FN}$. Для класса 1: $\text{`recall'} \approx 0.93$ -- модель почти всегда находит то, что действительно стоит рекомендовать.

И полнота: $\text{`precision'} = \frac{TP}{TP + FP}$. Для класса 1: $\text{`precision'} \approx 0.89$ -- среди рекомендованных мало лишнего.

МОДЕЛЬ CatBoostClassifier. Дополнительная проверка

	Предсказано 1	Предсказано 0
Истинно 1	TP = 108	FN = 8
Истинно 0	FP = 14	TN = 327

- 1) Только 14 товаров ошибочно рекомендованы (FP) – это допустимо.
- 2) И всего 8 реально хороших товаров не были рекомендованы (FN) – тоже допустимо, особенно если продукт осторожен.

Таким образом, **Модель обучена отлично**. Она сбалансирована: и не "перекомендует", и не упустит важное.

Такую модель можно использовать в реальной системе рекомендаций.

**СПАСИБО ЗА
ВНИМАНИЕ!**

