



Факультет Экономических Наук

Программа «Экономика и  
экономическая политика»

Москва  
2022

# H&M Personalized Fashion Recommendations

Проект выполнили:  
Денисова Кристина  
Шелованова Татьяна  
Команда Economizers



## Описательная статистика данных

Период с 20.09.2018 по 22.09.2020

Три типа данных:

- **Articles:** данные, определяющие товары H&M. Всего есть 25 параметров, но мы выбрали 9 из них
- **Customers:** данные о покупателях H&M. Мы посмотрели распределение по 1) возрасту, 2) является ли участником клуба, 3) подписан ли на новостную рассылку
- **Transactions:** информация о покупке: id клиента, id товара, цена и источник

1 362 281

Уникальных id покупателей

104 547

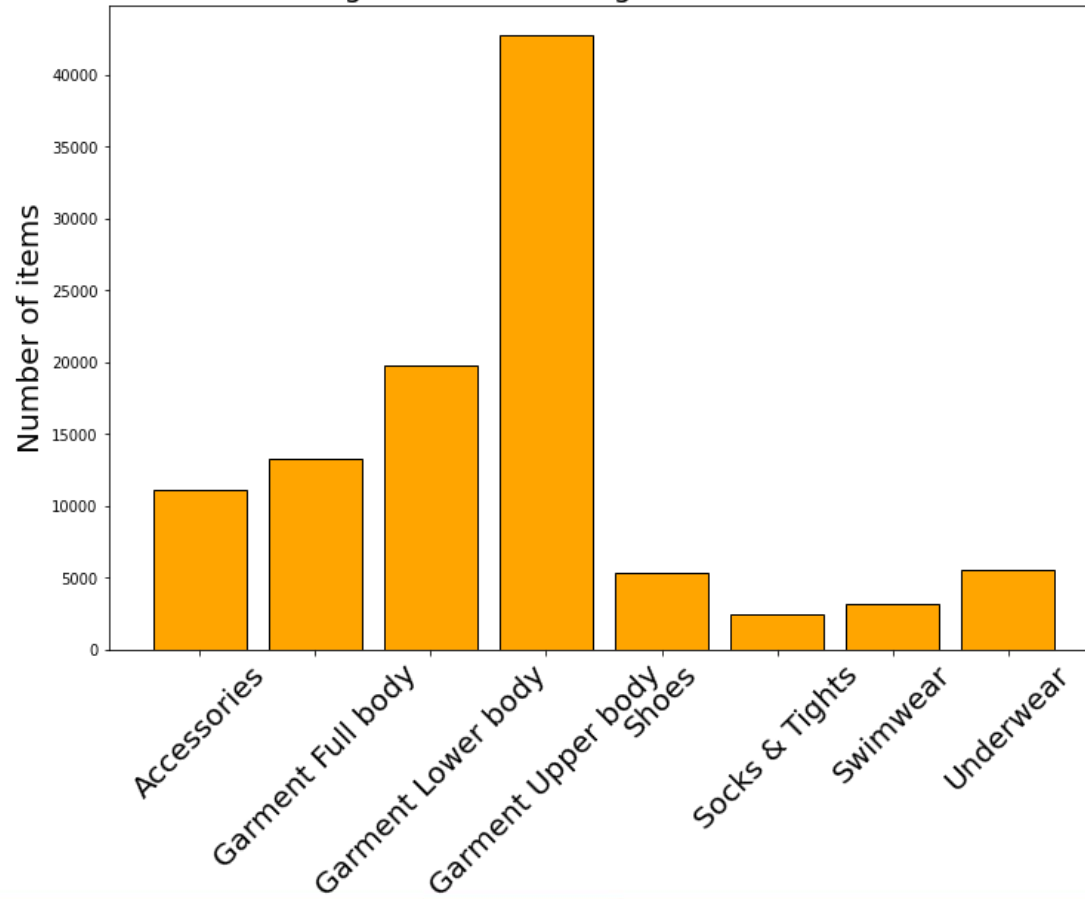
Уникальных id товаров

**Задача:** сформировать рекомендации покупателям на основе базы данных по покупках

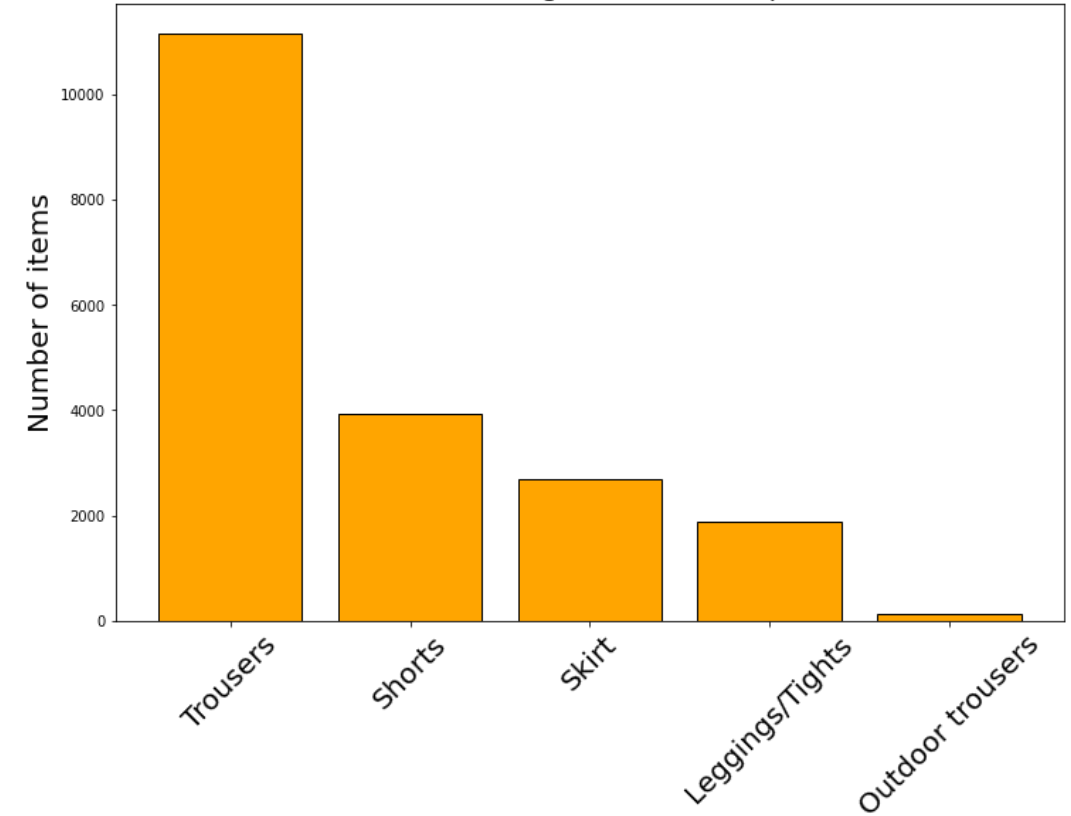


## Описательная статистика данных. Articles

Categories with the largest number of items

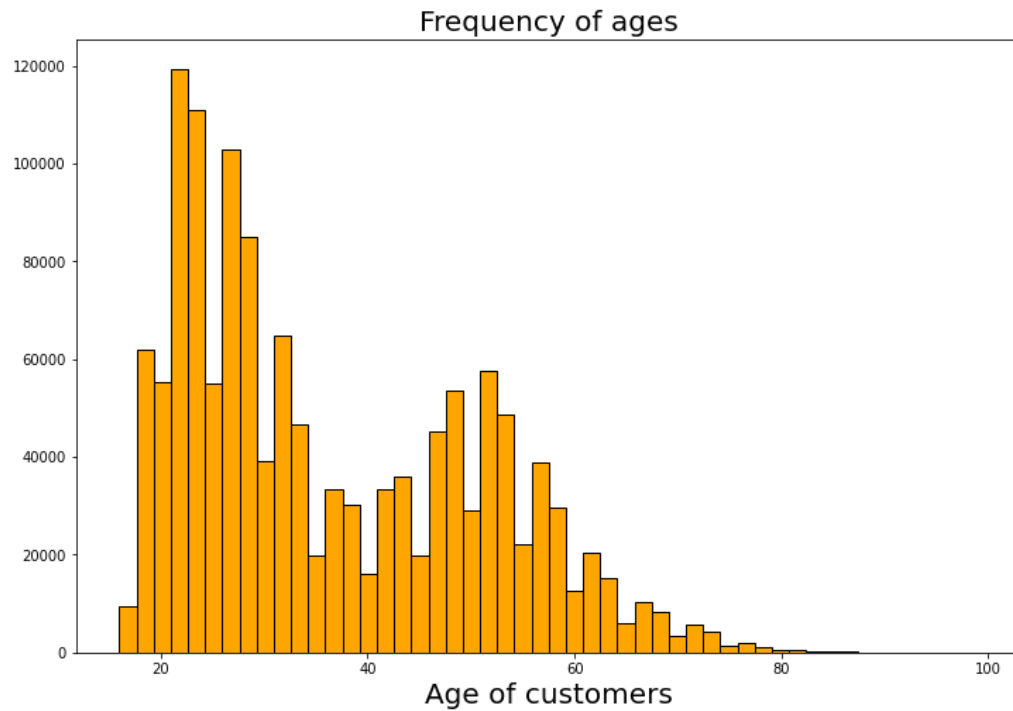


Articles with the largest number of purchases





## Описательная статистика данных. Customers



Proportion of club members

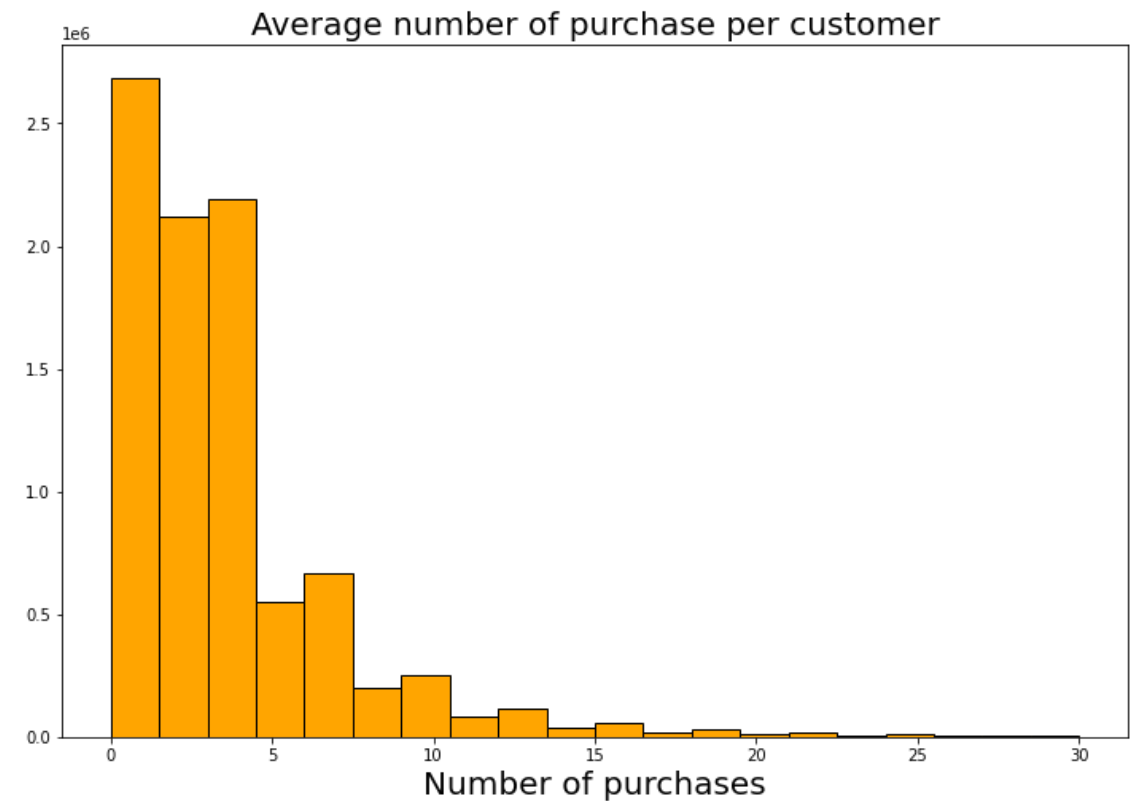
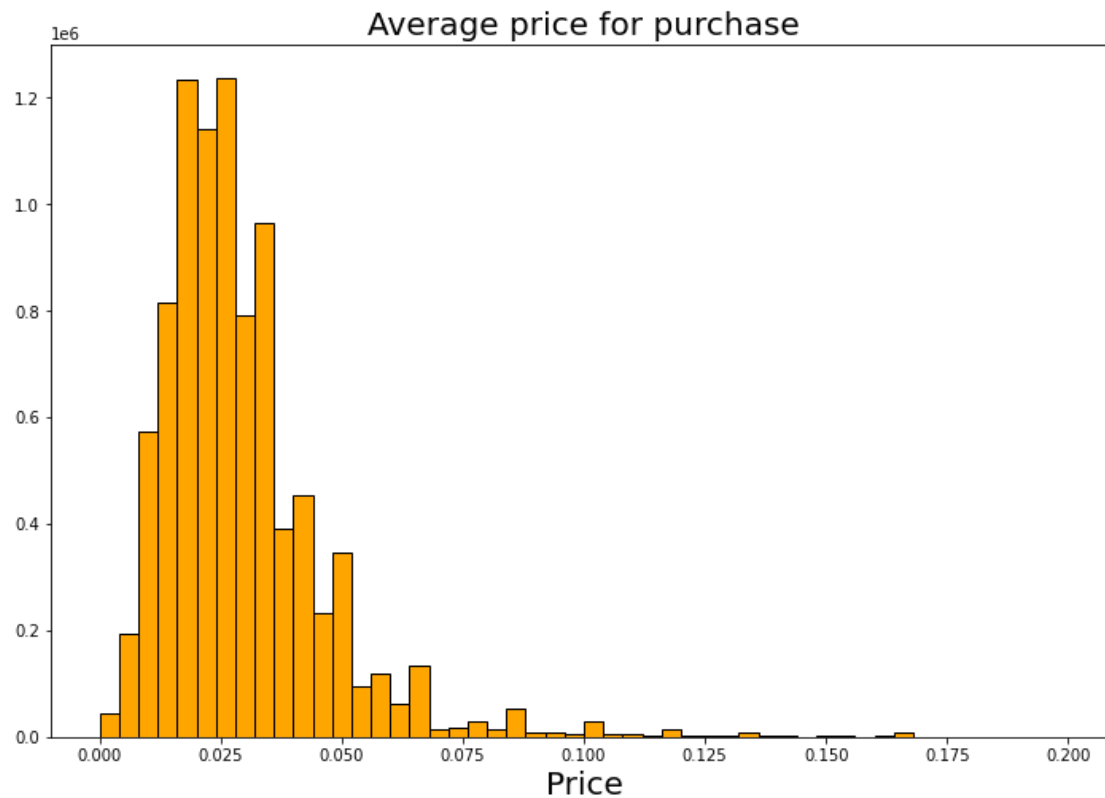


Proportion of news subscribers



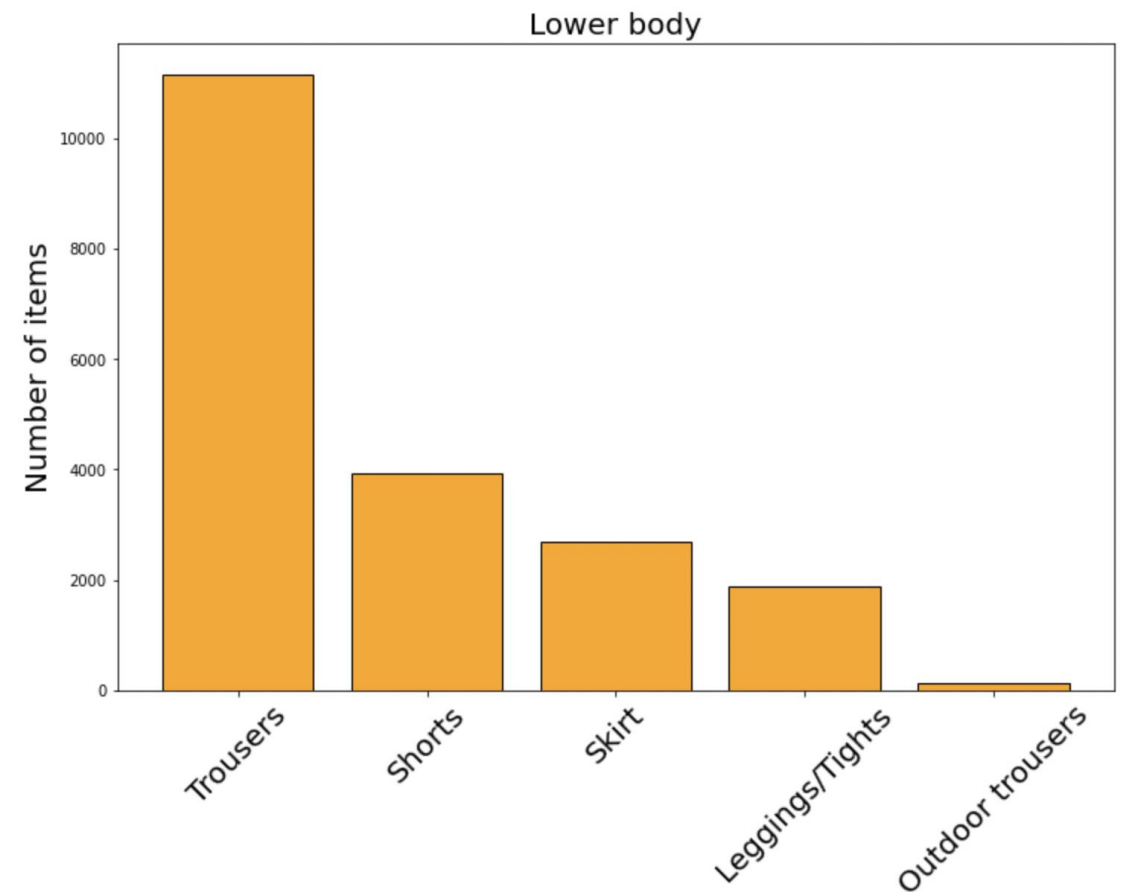
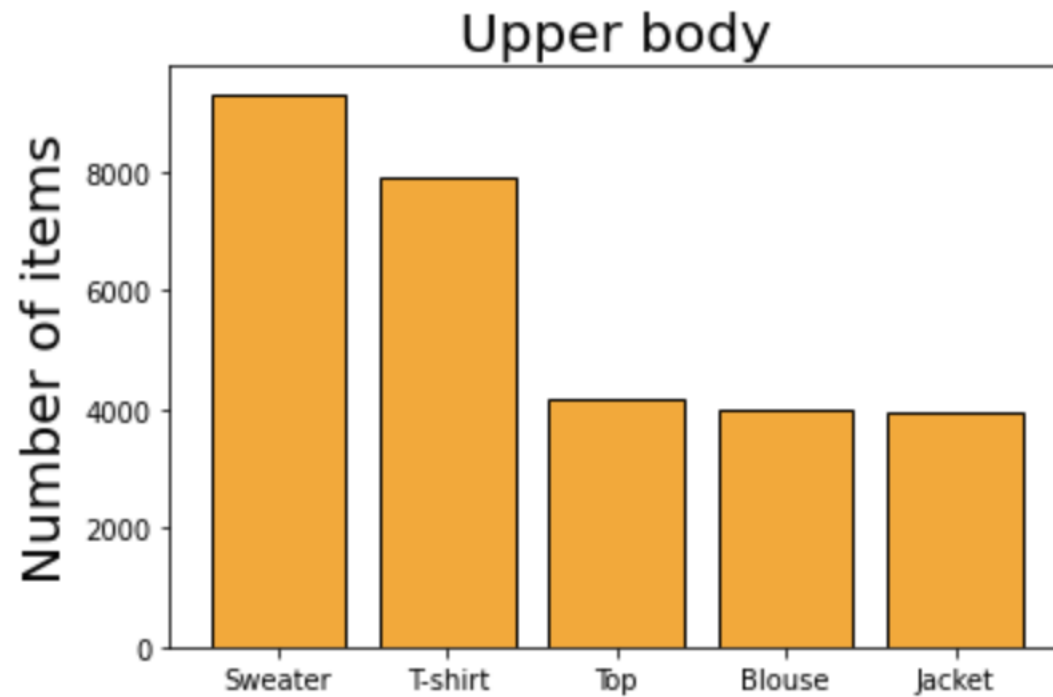


## Описательная статистика данных. Transactions





## Описательная статистика данных. Самые популярные типы проданной одежды



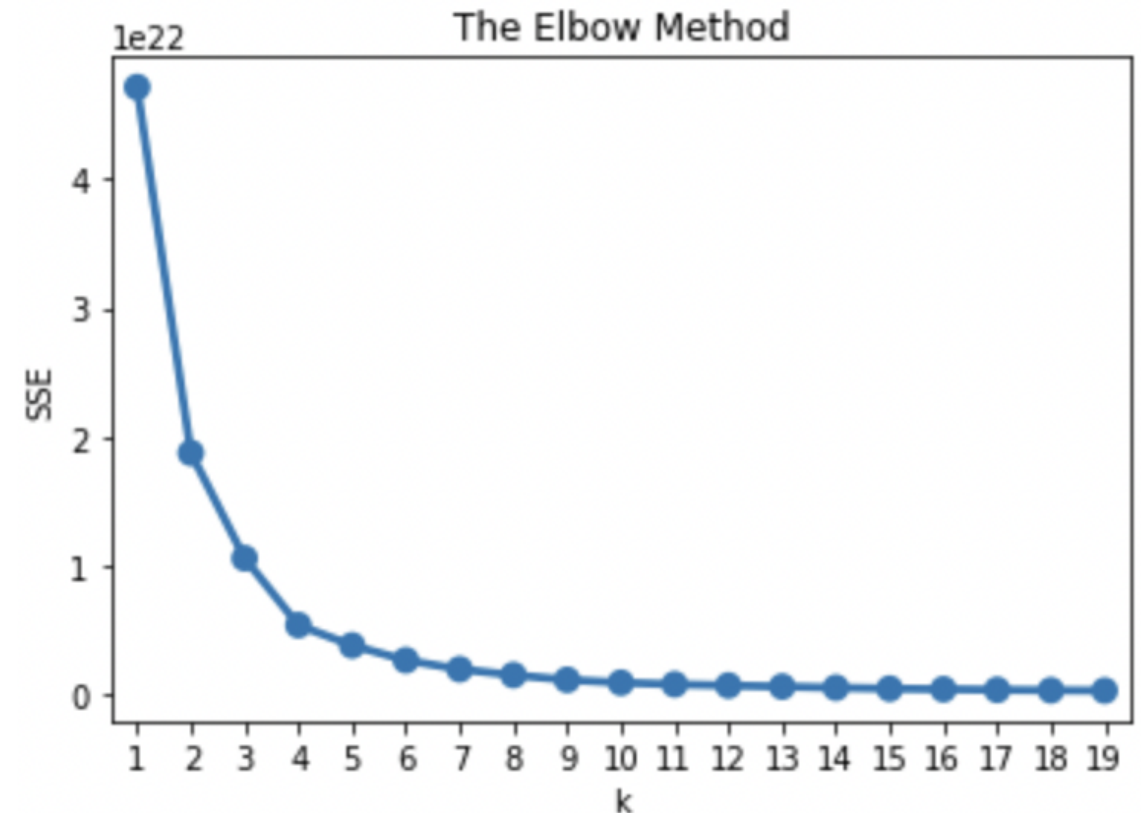


## Kmeans

### Данные для кластеризации

Название	Значение
price	Цена
product_code	Название продукта
product_type_no	Тип продукта
colour_group_code	Цвет

### Метод локтя





## Базовая модель

### Рекомендации по популярности

- Создание дополнительной переменной, отвечающей за частоту покупки данного товара у конкретного пользователя
- Создание нормированных показателей по данной переменной
- Оценивание score каждого потребителя относительно самого популярного товара среди потребителей

#### Недостатки:

- Не учитываем особенности пользователей
- Насколько релевантен показатель «частота покупки»?
- Проблема холодного старта

Название	Тип	Категория	Score
Long Leggings	Leggings/Tights	Garment Lower body	0.15
Delphine Shirt	Shirt	Garment Upper body	0.14
Billie	T-shirt	Garment Upper body	0.13
Ronny Reg T-shirt	T-shirt	Garment Upper body	0.12
Nicky long	Vest top	Garment Upper body	0.11

Mean Precision: 0.017





## Item-item collaborative filtering

Рекомендуем товары относительно тех, что похожи на данный

- Строим коэффициент показатель схожести (через косинус) между товарами
- Строим рекомендации, основанные на товарах, которые понравились или потреблялись пользователем

### Недостатки:

- Проблема холодного старта
- Большие размеры данных
- Нет рекомендаций для нетипичных товаров
- Рекомендация одинаковых товаров

Название	Тип	Категория	Score
Jade HW Skinny Denim TRS	Trousers	Garment Lower body	0.10
Skinny Ankle R.W Brooklyn	Trousers	Garment Lower body	0.08
Jade HW Skinny Denim TRS	Trousers	Garment Lower body	0.07
Jade HW Skinny Denim TRS	Trousers	Garment Lower body	0.06
The Low Line Highwaist	Trousers	Garment Lower body	0.05

Mean Precision: 0.11



## Collaborative filtering with KNN

Объединяем схожих покупателей

- Смотрим на сходство между пользователями
- Найдем тех, что больше всего похожи на конкретного пользователя
- Рекомендуем товары на основе их предпочтений

### Недостатки:

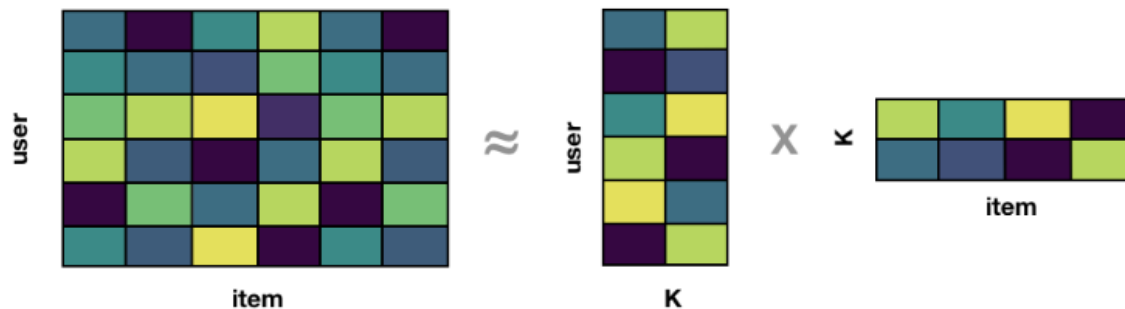
- Проблема холодного старта
- Хранение большого объема данных в матрице

Название	Тип	Категория	Score
Hazelnut Push Melbourne	Bra	Underwear	
Shenzi LP	Top	Garment Upper body	0.71
Long leggings update	Leggings	Garment Lower Body	0.73
Clarence Push Wireless	Bra	Underwear	0.82
Therese tee	T-shirt	Garment Upper body	0.86

## ALS

- Возьмем большую матрицу и разложим её на некоторое меньшее представление исходной матрицы.
- Меньшее представление возможно благодаря выделению скрытых общих факторов между потребителями и товарами

$$X_{mn} \approx P_{mk} \times Q_{nk}^T = \hat{X}$$



Этапы построения модели:

- Шаг 1: оставляем неизменной user-factor matrix и решаем для item-factor matrix
- Шаг 2: оставляем неизменной item-factor matrix и решаем для user-factor matrix
- Чередуем первый и второй шаг до тех пор, пока произведение матриц не будет равно исходной user-item matrix



## ALS

### Преимущества:

- Решают проблему холодного старта
- Дают адекватные оценки сходства покупателей и товаров
- Сокращают объем хранимых данных (за счет сжатия матрицы)

Название	Тип	Категория	Score
Tara top	Sweater	Garment Upper body	
Sheinzi LP	Top	Garment Upper body	0.45
Hazelnut Push Melbourne	Bra	Underwear	0.36
Twenty HW taperd	Trousers	Garment Lower body	0.27
Noa skinny trouser	Trousers	Garment Lower body	0.25

Mean Precision: 0.33



## Light FM – гибридная модель коллаборативной фильтрации

Light FM позволяет учитывать фичи продуктов и юзеров

Метод: Стохастический градиентного спуск

Функция потерь: WARP: Weighted Approximate-Rank Pairwise 2 loss

Метрика качества: ROC AUC

	Model without features	Model with features
AUC	0.70	0.29



## Критерии выбора модели

Рекомендательные системы можно охарактеризовать с помощью нескольких критериев:

- Разнообразие: учитывает количество различных категорий
- Новизна: количество новых товаров, ранее не рассматриваемых потребителем
- Покрытие: доля объектов, которую рекомендует наша система
- Догадливость: способность учесть нетривиальные предпочтения

