2020
DATA SCIENCE

GPN

Intelligence Cup

Докладчик: Лепёхина Анфиса
Студент 3 курса СПбГЭУ
"Прикладная математика
и информатика"

ЗАДАЧА

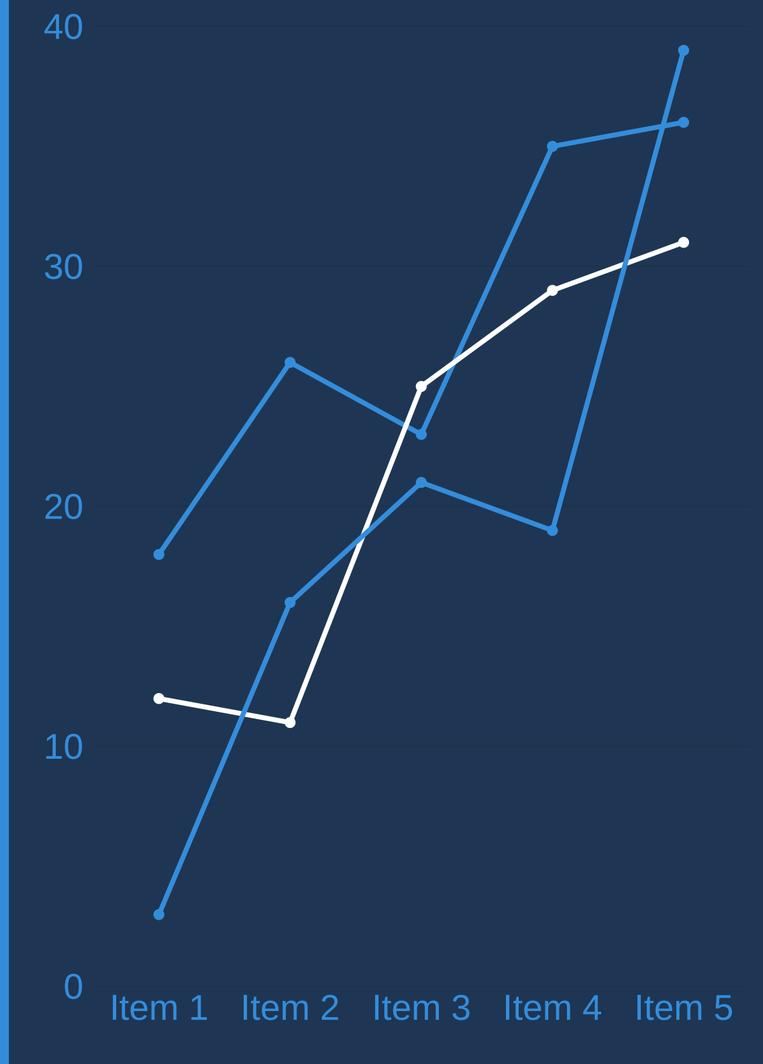
Разбить магазины на кластеры похожих

- улучшение системы управления магазинами
- обеспечение оптимального планирования
- прогнозирование спроса

ВОЗМОЖНЫЕ РЕЗУЛЬТАТЫ

- увеличение товарооборота
- увеличение трафика
- увеличение прибыли
- снижение затрат на логистику
- повышение точности доставки товаров в магазины
- снижение количества остатков в магазинах

ГАЗПРОМ НЕФТЬ
Стремимся к большему



ЭТАПЫ ПРОВЕДЕННОГО КЛАСТЕРНОГО АНАЛИЗА

1. АНАЛИЗ ДАННЫХ
2. ОТБОР ПРИЗНАКОВ
3. ВЫБОР МЕТРИК
4. ВЫБОР АЛГОРИТМОВ
5. ИНТЕРПРИТАЦИЯ РЕЗУЛЬТАТА

1.АНАЛИЗ ДАННЫХ

Знакомство с признаковым описанием объектов

Знакомство с типами данных

Выявление пропусков и аномалий в данных

2.ОТБОР ПРИЗНАКОВ

Аналитический отбор и формирование информативных признков

Обработка информативных признаков:

- обработка пропусков
- нормализация и масштабирование числовых признаков
- dummy-кодирование для работы с категориальными признаками

3.ВЫБОР МЕТРИК КАЧЕСТВА

Стремимся к большему

Среднее внутрикластерное расстояние ->min

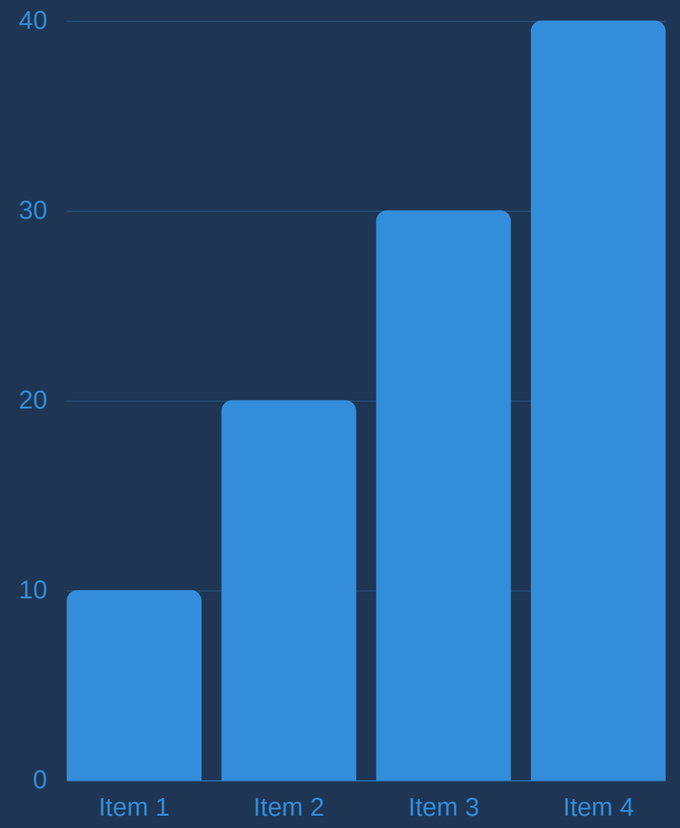
Среднее межкластерное расстояние -> max

Невозможно подобрать количество кластеров

Коэффициент силуэта
$-1 < s < 1$

+ максимизируя силуэт, можно подобрать количество кластеров

+ чем больше силуэт, тем более четко выделены кластеры



4.ВЫБОР АЛГОРИТМОВ

- DBSCAN

  + определяет число кластеров

  - плохо распознает вложенные и близко расположенные кластеры

- Агломеративная кластеризация

  + удобно визуализировать

  - самостоятельный подбор количества кластеров

- K-means

  + простота использования, понятный алгоритм

  - вычислительно сложный при большом количестве признаков

  - самостоятельный подбор количества кластеров

4. ВЫБОР КОЛИЧЕСТВА КЛАСТЕРОВ

- ✓ анализ дендрограмм, построенных по иерархической кластеризации
- ✓ анализ количества кластеров, максимизирующее коэффициент силуэта
- ✓ анализ скорости убывания расстояния между кластерами
- ✓ анализ специфики поставленной задачи

N = 6

Кластер	Общая информация	Специализация магазинов
1	Магазины, расположенные «В центре»: **shop_type1**	**Много:** «Ядер-Колы», «Солярки», «Брони и одежды», «Съедобного хлама», «Модификации тачек», «Жидкости для тачки» **Средние результаты**: «Медпрепараты и еда», «Патроны», «Оружие», «Модификации тачек»
3	Магазины, расположенные «У тоннеля», «У ночлега»: **shop_type1**	**Много:** «Брони и одежды», «Модификации тачек», «Съедобного хлама», «Жидкости для тачки» **Средние результат**: «Ядер-кола», «Медпрепараты и еда», «Солярка», «Патроны», «Оружие», «Хлам»
5	Магазины, расположенные «На отшибе», «В промзоне», «С краю», «У воды»: **shop_type1**	**Лидер**: «Ядер-Кола», «Медпрепараты и еда», «Солярка», «Броня и одежда», «Оружие», «Модификации тачек», «Съедобный хлам», «Жидкости для тачки» **Много**: «Патрон», «Хлама»
2	Все магазины: **shop_type_2**	**Лидер**: «Патроны», «Хлам» **Много**: «Ядер-Колы», «Съедобного хлама», «Жидкости для тачки» **Средние результаты**: «Медпрепараты и еда», «Солярка»
4	Все магазины: **shop_type_3**	**Средние результаты**: «Солярка» **Минимум** по всем остальным показателям!
0	Магазины, расположенные не у воды, **shop_type 4**	**Средние результаты**: «Солярка» **Минимум** по всем остальным показателям!

5.ИНТЕРПРИТАЦИЯ РЕЗУЛЬТАТА

- Получено **6 кластеров**, описание которых представлено в таблице
- Два кластера **(2, 5) развитых** магазинов-лидеров, где помимо "Бензака, на высоком уровне продаются и другие типы товаров
- Два кластера **(1, 3) развивающихся** магазинов, в таких магазинах на хорошем уровне продаются многие типы товаров, однако от развитых магазинов эта группа отстает по продажам
- Два кластера **(4, 0) отстающих** магазинов, в таких магазинах на высоком уровне продаётся только"Бензак", в остальном эта группа значительно проигрывает другим типам кластеров