GPN Intelligence Cup

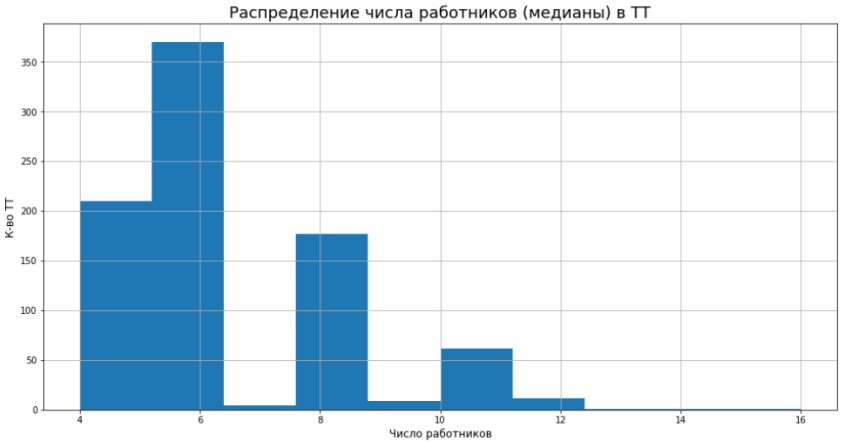
КЛАСТЕРИЗАЦИЯ ТОРГОВЫХ ТОЧЕК

Апроцкий М.В.

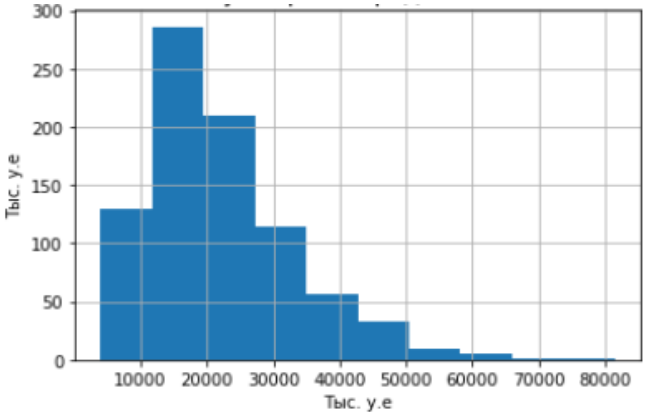
15.11.2020

ДАННЫЕ. SALES

- Не открывалось новых ТТ в анализируемом периоде
- Большинство ТТ **не прекращало** работу, исключения принимаем за перерыв на ремонт/карантин/региональные выходные и праздники
- ТТ, которые прекратили продажи до 01.01.2148 также участвуют а анализе (id 178, 179, 180, 181, 182)
- 157 ТТ меняли собственника в анализируемом периоде
- Для определения числа работников в ТТ используется медиана:



ДАННЫЕ. SALES

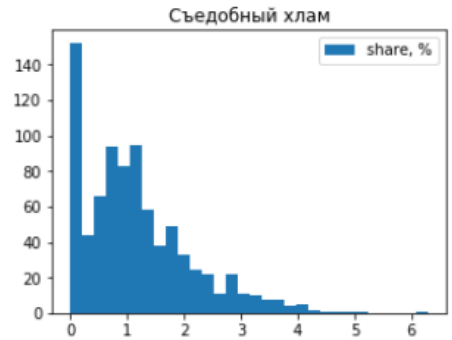


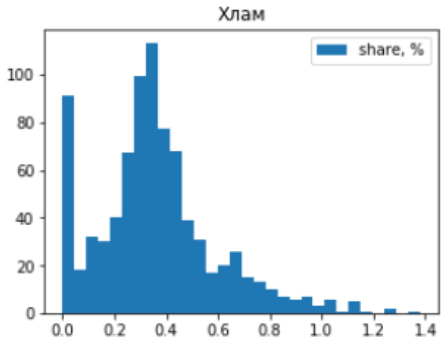
- Распределение торговых точек по совокупной выручке
- Значительная часть ТТ демонстрирует выручку ниже 50000 тыс. у.е.

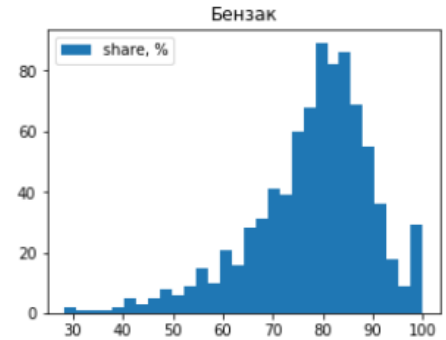
- Категория товаров может быть выведена из ассортиментной матрицы ТТ в течении анализируемого периода.
- Будем считать, если 95% квантиль продаж по категории >0, значит категория в матрице данной ТТ.

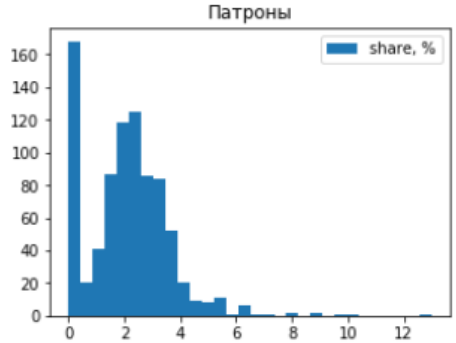
ДАННЫЕ. SALES.

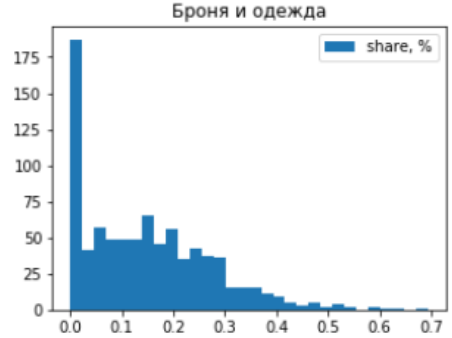
ДОЛЯ КАТЕГОРИИ В ПРОДАЖАХ ТТ

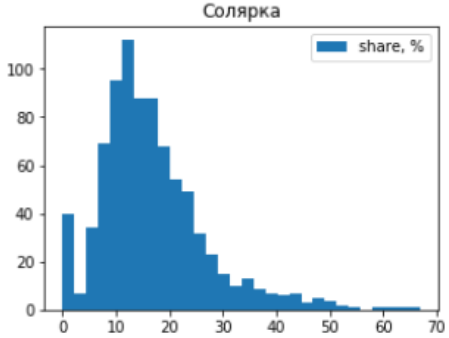


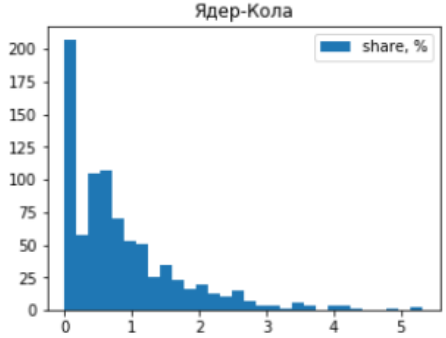


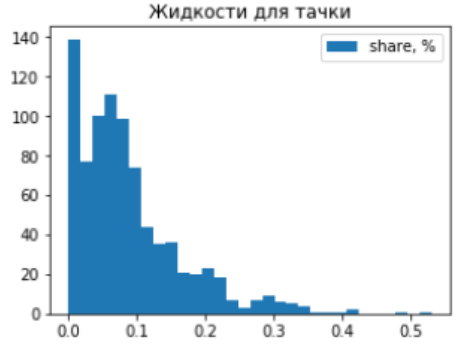


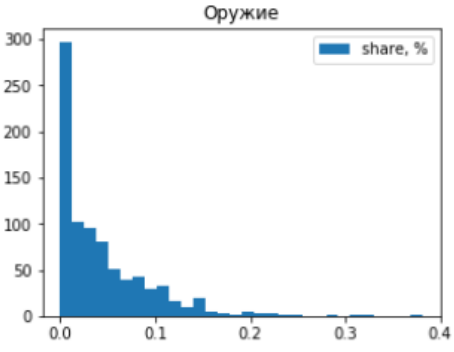


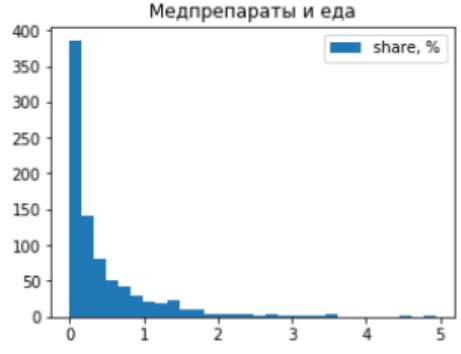






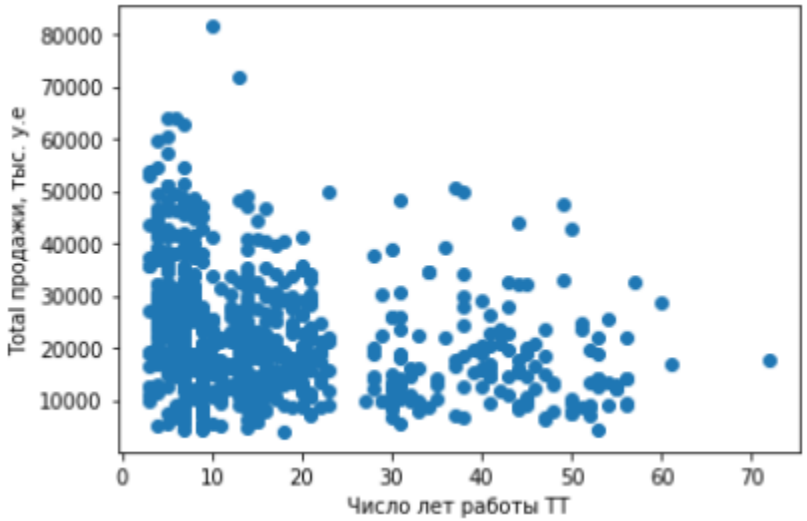






ДАННЫЕ. SHOPS

- 5 переменных с **пропущенными** значениями: 3 бинарных (is_on_the_road, is_with_the_well, is_with_additional_services), 2 категориальных (shop_type, city).
- большая часть ТТ находится **"В центре".** Для дальнейшего анализа преобразуем переменную 'neighborhood' в бинарную, где **1** - ТТ "В центре", **0** - ТТ "Не в центре"
- **15** городов, **3** региона
- Дата открытия магазина определена не для всех ТТ
- **К-т корреляции** между количеством лет работы с момента открытия ТТ и суммарными продажами ТТ ≈ -0.25. Переменную 'year_opened' использовать в дальнейшем не будем

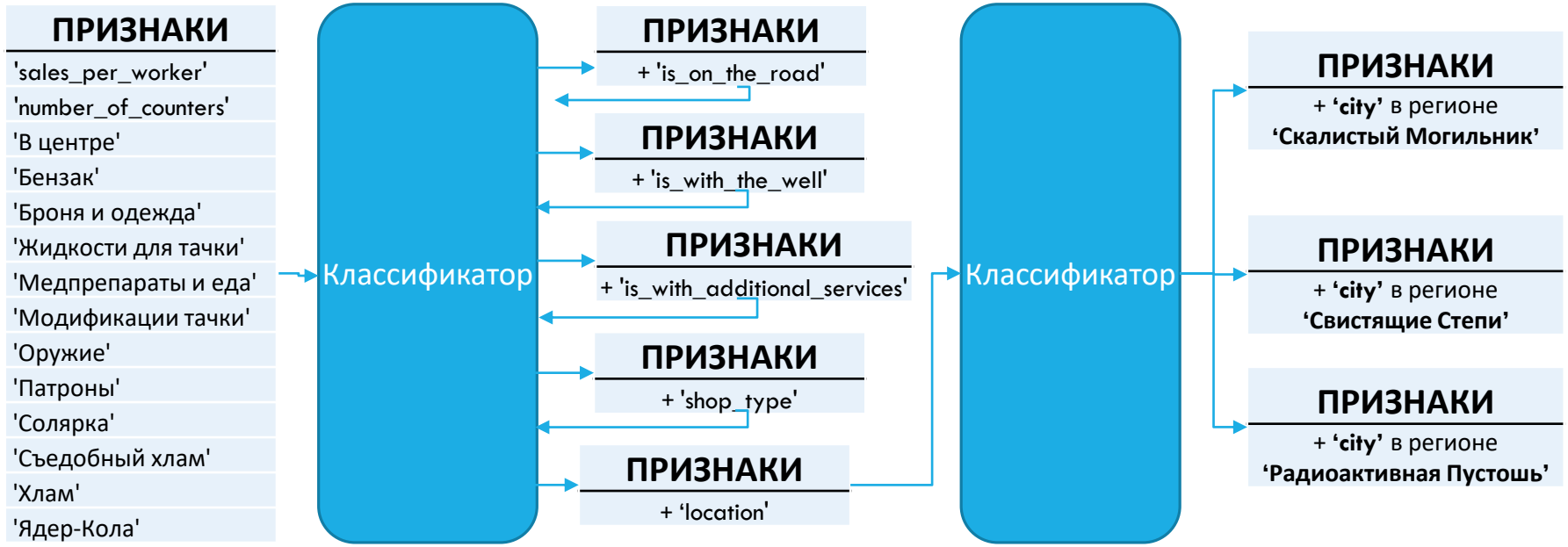


ДАННЫЕ. ПРОПУСКИ ∅

- **Создадим классификатор** (на вход подадим: продажи на 1 сотрудника, число сотрудников, «В центре», доли категорий в продажах для каждой ТТ), с помощью которого заменим пропущенные значения в переменных в следующем порядке (сначала бинарные по возрастанию числа пропусков, затем категориальную 'shop_type'):

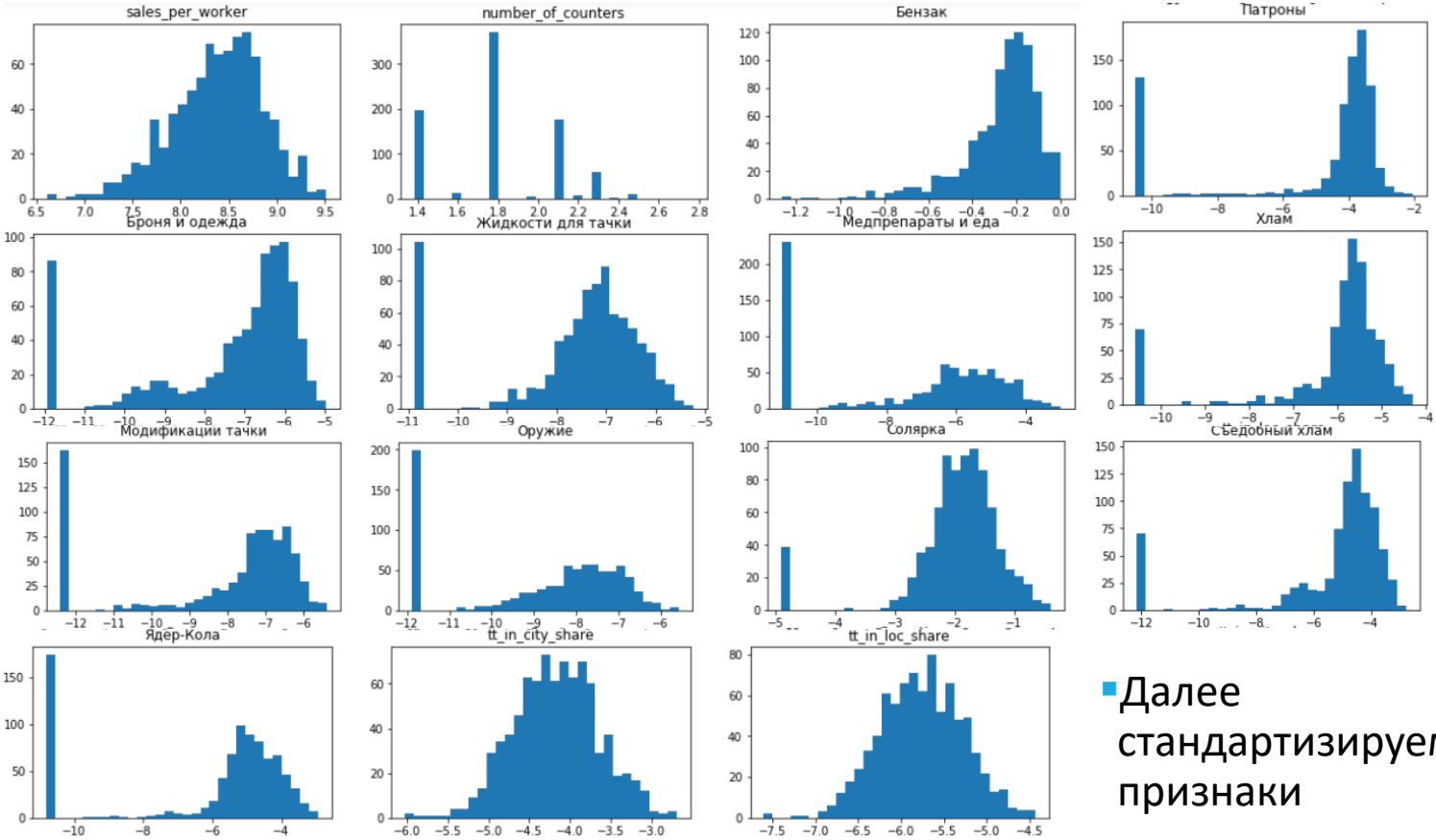
  'is_on_the_road', 'is_with_the_well', 'is_with_additional_services', 'shop_type'

- Для **повышения точности** классификации 'city' сначала классифицируем ТТ по регионам ('location'), после чего произведем классификацию по городам **внутри регионов.**



ДАННЫЕ. PREPROCESSING

- В числовых признаках могут встречаться "выбросы", как в значении 0 (доли категорий в ТТ) и в "правом хвосте" (доли ТТ в городе/регионе) – возьмем натуральный логагифм от числовых признаков*:

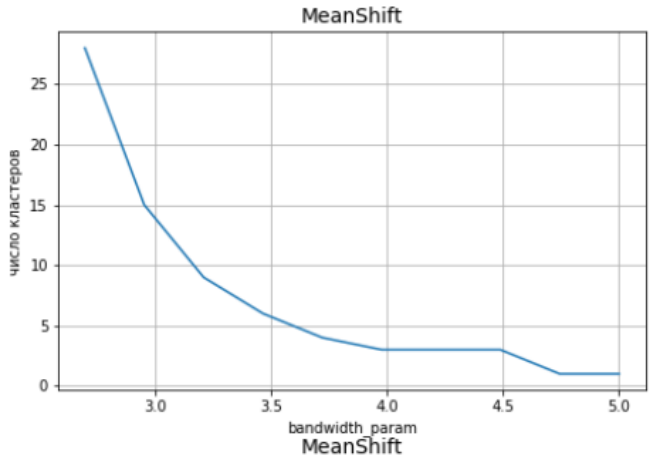


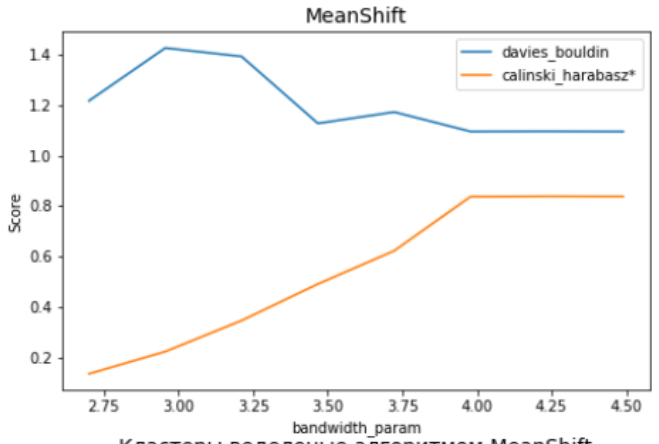
- Далее стандартизируем признаки

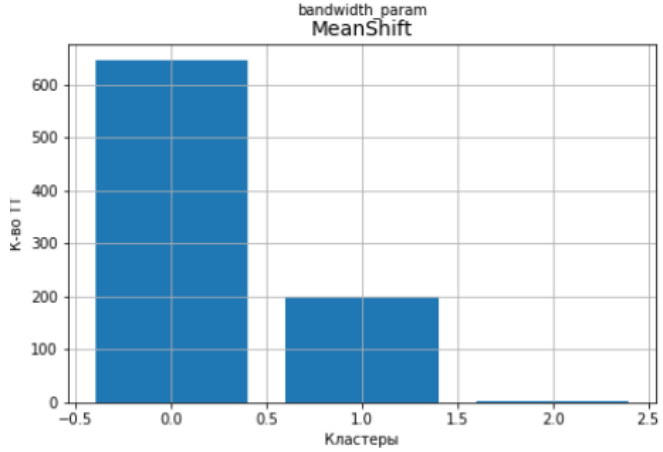
*подробнее в notebook

ДАННЫЕ. КЛАСТЕРИЗАЦИЯ

- **MeanShift.** Для калибровки используем метрики **Davies–Bouldin index*** и **Calinski-Harabasz Index****



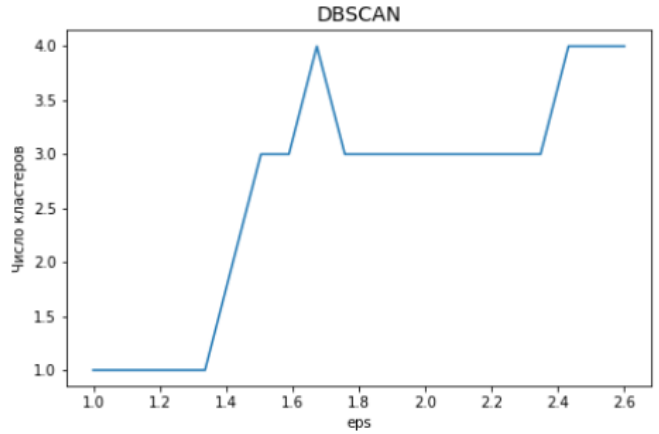


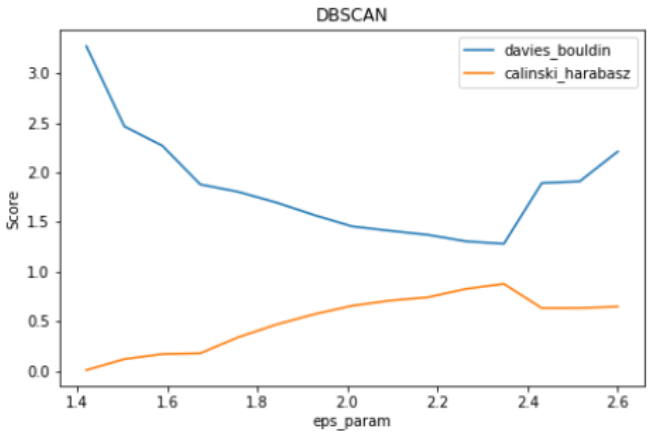


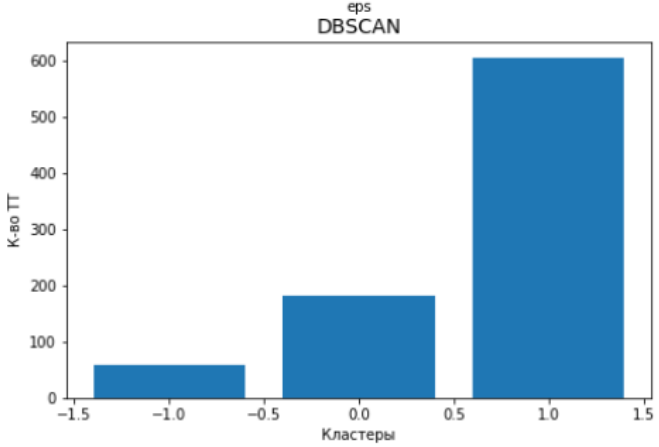
* https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index
** https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

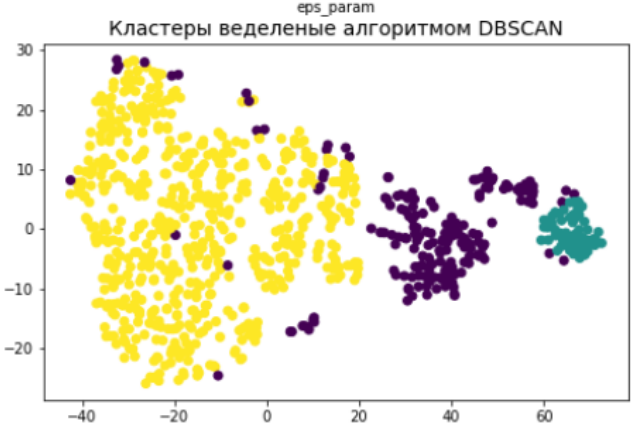
ДАННЫЕ. КЛАСТЕРИЗАЦИЯ

- **DBSCAN.** Для калибровки используем метрики **Davies–Bouldin index*** и **Calinski-Harabasz Index****





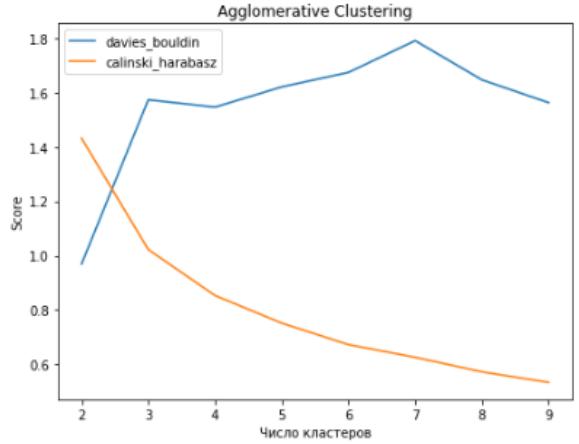


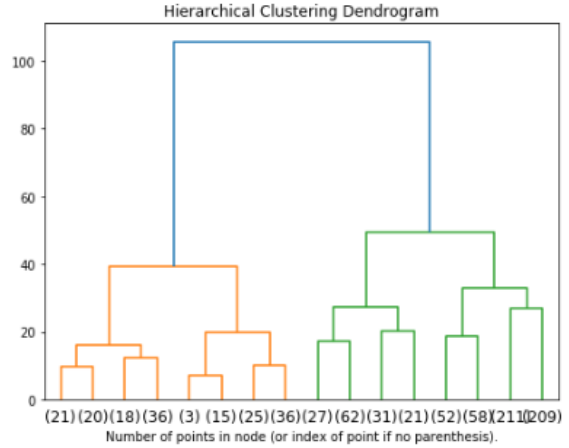


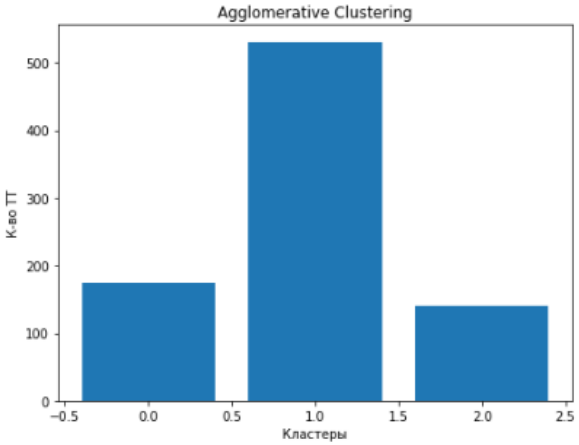
* https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index
** https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

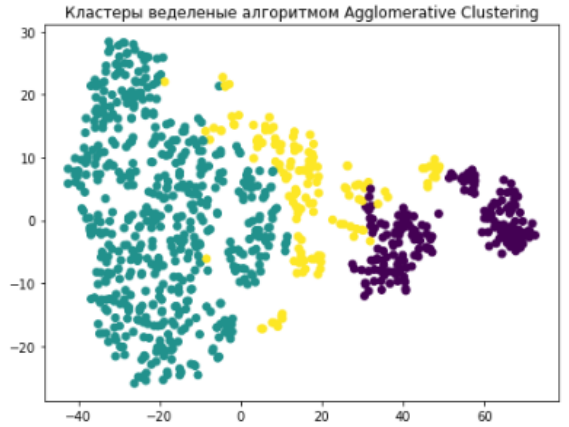
ДАННЫЕ. КЛАСТЕРИЗАЦИЯ

- **Agglomerative Clustering.** Для калибровки используем метрики **Davies–Bouldin index*** и **Calinski-Harabasz Index****





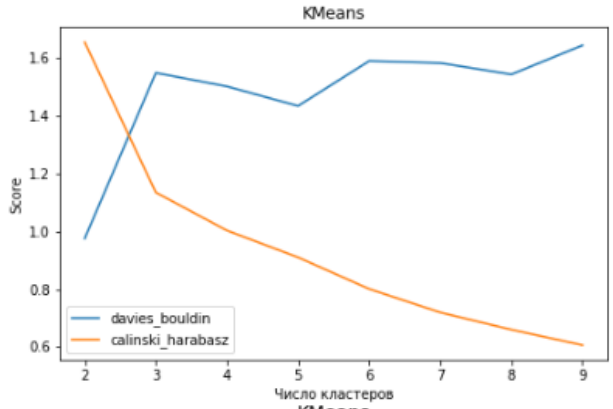


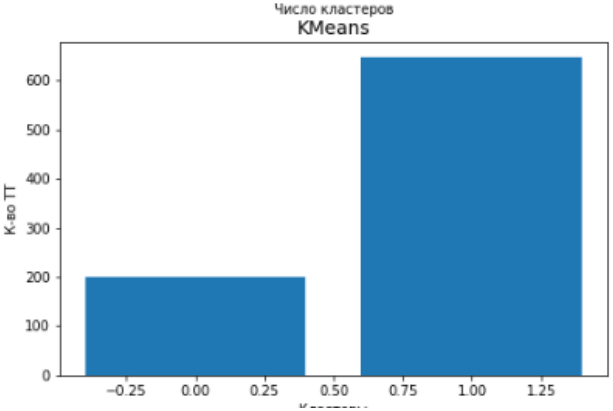


* https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index
** https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

ДАННЫЕ. КЛАСТЕРИЗАЦИЯ

- **Kmeans.** Для калибровки используем метрики **Davies–Bouldin index*** и **Calinski-Harabasz Index****





* https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index
** https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

ДАННЫЕ. КЛАСТЕРИЗАЦИЯ

- Лучшие метрики показывают **Аггломеративная кластеризация** и **KMeans** при обучении для 2 кластеров. Предположим, что наша выборка содержит **как минимум 3 группы ТТ**, содержательно отличающихся между собой. Тогда:

	davies_bouldin*	**calinski_harabasz****
DBSCAN	1,28	351
MeanShift	**1,1**	335
Agglomerative Clustering	1,57	409
KMeans	1,55	**454**

- Как видим, лучшим по криетрию **'davies_bouldin index'** признается **MeanShift** кластеризация. Заметим, что при использовании данного метода кластеризации, наблюдаем знаительный **дисбаланс** по кол-ву элементов в кластерах - только 2 ТТ попадают в класстер 2.

- Вторым по качеству критерия **'davies_bouldin index'** является метод **DBSCAN**, который также превосходит метод MeanShift по индексу **'calinski_harabasz'**. Определим метод **DBSCAN**, как итоговый метод кластеризации.

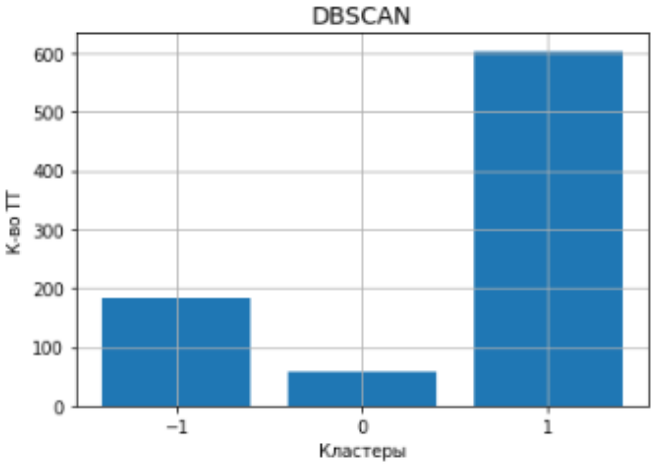
* https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index
** https://www.tandfonline.com/doi/abs/10.1080/03610927408827101

КЛАСТЕРЫ. ИНТЕРПРЕТАЦИЯ

'-1' - **ТТ с низкой операционной эффективностью**. 183 ТТ; преимущественно малого и среднего размера (4-6 сотрудников); самый низкий показатель продаж на 1 работника; как правило, занимают наименьшую долю продаж в своих Городах. Высокая доля **топливных категорий** (Бензак, Солярка), но в матрице присутствуют и нетопливные продукты (в большей степени Съедобный Хлам, Хлам, Патроны, Жидкости для Тачки).

'0' - **Топливноориентированные ТТ**. 58 ТТ; торговые точки "малого" формата - преимущественно по 4 сотрудника, демонстрируют самые высокие продажи на 1 сотрудника, занимают вторую долю по объемам продаж в своих Городах. Чаще остальных ТТ расположены "В центре", практически не встречаются "у дороги". Высокая доля **топливных категорий** (Бензак, Солярка), нетопливыне продукты практически не присутствуют в матрицах данных ТТ. В торговых точках практически всегда отсутствуют "дополнительные сервисы и услуги".

'1' - **ТТ с широким ассортиментом товаров и услуг**. 604 ТТ; как правило наиболее крупные ТТ - в основном по 6-8 сотрудников, занимают второе место по выручке на 1 работника; лидируют по доле продаж в своих Городах и Регионах. Значительно чаще остальных встречаются "У дороги". ТТ лидируют по доле **нетопливных категорий** (в топливном сегменте доля Солярки выше, чем у остальных типов ТТ), ассортиментная матрица у данного типа ТТ максимально разнообразна. Чаще всего в ТТ присутствуют "дополнительные сервисы и услуги"



Различия между кластерами по всем переменным статистически значимы*.

* Подробнее в notebook

КЛАСТЕРЫ. ИСПОЛЬЗОВАНИЕ

Город	Кластер 0	Кластер 1	Кластер -1
Аэропорт		27	7
Буровая Скважина	2	27	10
Газтаун	4	40	1
Дизельные Жилы	2	40	20
Крепость Джита	6	39	25
Лагерь	1	68	3
Нефтеперегонный Завод	4	40	13
Пасть	1	30	41
Равнина Маяка	2	64	10
Свинцовая Ферма	22	21	2
Суховей	7	23	19
Темница	4	36	6
Убежище Жестянщика	2	63	10
Храм Фритюра		63	11
Цитадель	1	23	5

- Предлагаемый метод кластеризации может повысить качество проводимой **категорийной** и **маркетинговой** политики для отдельных ТТ.
- Использование классификации ТТ только по Городам может привести к снижению эффективности в управлении отдельными ТТ. Например, в Свинцовой Ферме большую долю ТТ занимают **Топливноориентированные** магазины., а в Суховее, Крепости Джита и Пасти достаточно высока доля **Низкоэффективных** ТТ – инвестиции в промо по категориям «Бензак», «Солярка» могут не дать ожидаемого эффекта в данных городах.

Узкие места предложенного метода кластеризации:

- Может давать непредсказуемые результаты для недавно открывшихся ТТ, доли продаж по категориям в которых еще не «устоялись» и могут измениться со временем
- При наличии слишком большого (или наоборот, слишком низкого) **кол-ва ТТ в Городе** объекты могут неверно кластеризованы неверно из-за заниженных/завышенных долей собственных продаж в Городе.