# ОТЧЕТ О ПРОДЕЛАННОЙ РАБОТЕ КОМАНДЫ “EXCELLENT’S” НА ДАТАТОНЕ МАГАСТРАТУРЫ МФТИ И SKILLFACTORY

## Идея, постановка задачи и модель итогового результата

- Какой самый важный атрибут новогоднего праздника?

- Ну конечно же ёлка!

Устанавливать в новогодние праздники украшенную елку – хорошая традиция в российских семьях. С каждым годом все больше и больше жителей нашей страны отказываются от покупки срубленных живых деревьев, это не экологично и не практично, когда можно один раз купить хорошую искусственную ель, которая прослужит хозяевам многие годы. Но как выбрать посредствам самую крепкую, качественную, красивую, словом, лучшую? Вопрос сложный и наша команда в рамках дататона магистратуры «Науки о данных» решила помочь россиянам с выбором, проанализировав предложения на одном из самых популярных маркетплейсов в России – в интернет-магазине «Wildberries».

|  |  |  |
| --- | --- | --- |
|  | НА | Изображение выглядит как текст, коллекция картинок  Автоматически созданное описание |

Мы решили выяснить какой размер и материал самые популярные, какой самый продаваемый бренд и изготовитель, ценятся ли на елках шишки или украшения в виде инея и вообще – на что обращают внимание покупатели, заказывая в магазине новогоднюю елку.

Естественно для анализа нам понадобились данные. В нашей науке все крутится вокруг данных, а сами данные конечно получаются не из воздуха, они добываются! И это далеко не просто, поэтому остановимся на этом вопросе подробнее:

## методология получения данных

- один парсер – хорошо, но два гораздо лучше!

Наша команда применила парсер специально разработанный для поиска информации по API “Wildberries”. То есть вот наш первый источник <https://www.wildberries.ru/webapi/menu/main-menu-ru-ru.json>

Командой были написаны функции с учетом особенностей продукта + добавлено описание функций и комментарии к коду, мы решили написать код, который переводит данные в формат excel:

<https://github.com/calabarOlga/dataton_christmas_tree/blob/main/notebooks/Парсер_Елки.ipynb>

В итоге мы получили 9207 строк по категории новогодние елки. Мы сразу обратили внимание, что ценность представляют следующие столбцы, отмеченные на рисунке ниже красными стрелками:



На первый взгляд наиболее интересные в плане аналитики — это цена, количество отзывов и рейтинг, однако этих характеристик очень мало для анализа. И тогда мы решили обратить более пристальное внимание на столбец «наименование», поскольку в нем скрыты характеристики елок, которые мы хотели бы исследовать, но вот «изъять» их было очень сложно. Чисто для эксперимента мы попробовали выделить такую характеристику как высота, пользуясь исключительно любимым всеми членами команды инструментом Excel.

|  |
| --- |
| Ель "Кавказская Люкс" 2,2 м |
| Ёлка искусственная новогодняя 210 180 150 120 см с шишками |
| Елка Искусственная Сосна с инеем |
| Ёлка искусственная новогодняя 210 180 150 120 см с шишками |
| Елка искусственная Елка новогодняя Карпатская 180см |

Простой формулой, удалось вычленить из описания цифры, и потом формулами и фильтром очищать данные. Мы увидели, что в выборку попали гирлянды, подставки, ветки, шарики и прочие – все без сожаления удалили (эксперимент же).

Потом удалили строки, где за одну цену указывались елки разных размеров. В итоге осталось чуть больше половины – 4804 строчки. Работая далее с описанием, мы путем нехитрой комбинации фильтров выделили следующие категории и заполнили пропуски:



Посмотреть этот набор данных можно здесь:

<https://github.com/calabarOlga/dataton_christmas_tree/blob/main/data/елкиExcel.xlsx>

Но этих данных было очень мало для анализа, более 70 % столбцов с характеристиками, которые удалось выделить из описания, были пустыми, очевидно, что их просто необходимо было заполнить реальными данными (а никак ни чем-то средним).

Вот, пожалуй, единственный очевидный график, который мы приведем здесь по данному датасету. Плюс команде захотелось посмотреть и другие статистики: сколько раз купили, вес, упаковка, сопутствующие товары. Тогда мы разработали еще один парсер, который получает данные непосредственно со страниц сайта:

[https://catalog.wb.ru/catalog/new\_year1/catalog?appType=1&curr=rub&dest=-1075831,-7[…]0,69,1,48,22,66,31,40&sort=popular&spp=0&subject=260;7295;3738](https://catalog.wb.ru/catalog/new_year1/catalog?appType=1&curr=rub&dest=-1075831,-77677,-398551,12358499&locale=ru&page=1&priceU=0;1000000000&reg=0&regions=64,83,4,38,80,33,70,82,86,30,69,1,48,22,66,31,40&sort=popular&spp=0&subject=260;7295;3738)

Вот код этого замечательного инструмента:

<https://github.com/calabarOlga/dataton_christmas_tree/blob/main/notebooks/Парсер_Елки_Страницы_Продукта.ipynb>

Парсер оказался очень продуктивным, он работал по ссылкам, полученным в результате отработки первого парсера (последний столбец на рисунке выше озаглавленный «ссылка») и получал данные о количестве покупок и дополнительные характеристики, **УЖЕ** разделенные по категориям (а не так как в первом выходе в виде строки в колонке наименование). Однако задача оказалось не из легких, простой код работал очень долго, пришлось запускать его на нескольких машинах. Не удалось избежать и технических проблем виде перебоя с интернетом, которые вынудили нас запускать код повторно, но мы справились и получили 3 замечательных файла:

df1=pd.read\_excel('ind\_from\_0\_to\_5000.xlsx')

df2=pd.read\_excel('ind\_from\_5001\_to\_7000.xlsx')

df3=pd.read\_excel('ind\_from\_7001\_to\_10000.xlsx')

Посмотреть их можно в репозитарии:

<https://github.com/calabarOlga/dataton_christmas_tree/tree/main/data>

Завершая раздел по получению данных, стоит отметить, просмотр и беглая проработка Excel первых результатов подсветила проблемы, с которыми мы столкнемся в дальнейшем, это и лишние товары типа подставок и неуточненный размер елок (сразу несколько вариантов при одинаковой цене)

## Сбор датасета

Итак, второй этап работы состоял в том, что нам необходимо было соединить в один 4 таблицы с данными и выделить из столбца «описание» значимые признаки. Мы решили сразу исключить те признаки, содержащиеся в дополнительном описании, в которых было менее 1000 значений.

Для начала мы объединили в одну три таблицы, получившиеся в результате отработки второго парсера и соединили его по строкам с первоначальной таблицей (для проверки корректности соединения мы сохранили нумерацию строк, как своего рода id, и естественно проверили глазами). Все получилось очень даже приятно. Далее мы разбили наш столбец описание на категории с учетом оговоренного условия, что признак встречается не менее 1000 раз. В итоге мы получили дополнительные столбцы: ['Конструкция елки', 'Подсветка', 'Страна производства', 'Высота упаковки', 'Длина упаковки', 'Особенности елки', 'Тип елки', 'Высота елки', 'Ширина упаковки', 'Вес с упаковкой (кг)', 'Назначение подарка', 'Материал елки', 'Повод', 'Количество ветвей', 'Комплектация']

Промежуточный «склеенный» датасет получился вот такой:

<https://github.com/calabarOlga/dataton_christmas_tree/blob/main/data/final_data_for_cleaning_and_working_with_gaps.xlsx>

А код, которым мы пользовались, можно посмотреть вот здесь:

<https://github.com/calabarOlga/dataton_christmas_tree/blob/main/notebooks/Сбор_итогового_датасета_и_выделение_дополнительных_фичей.ipynb>

## Очистка данных:

- Нужен закон о том как описывать товар в ИНТЕРНЕТ-МАГАЗИНЕ, иначе DS сойдут с УМА!

Над очисткой данных работали всей командой. В самом первом приближении мы поняли, что строки ‘Назначение подарка’ и ‘Повод’ абсолютно бесполезны, поскольку содержали такие банальные истины как: «подарок в семью», «елка в дом», «дедушке», «в детскую», «на новогоднее торжество», and so on. – Соответственно эти характеристики удалили без жалости.

- Мы его слепили из того что было и получилось замечательно)!

Далее приступили к очистке:

1. Первым делом мы заполнили пропущенные значения в столбце «Страна производитель», а имеющиеся значения привели в строковый формат, и дополнили данные о производителе, теми, что были в наименовании, но были пропущены в анализируемом столбце
2. Далее удалили из данных строки, которые содержали сведения о сопутствующих товарах: таких как подставки, венок, юбка и т. д.
3. Скорректировали данные в столбце подсветка (1- есть, 0 – нет)
4. Огромная работа была проделана по определению высоты елки, данные были записаны в метрах, миллиметрах, сантиметрах, содержали лишние символы и значения. При отсутствии данных подтягивали их колонки «название».
5. Аналогичным образом проведена обработка столбца «высота упаковки» и «ширина упаковки», «вес с упаковкой»
6. Из-за неполноты (менее 70%) мы решили полностью удалить сведения по столбцам «Конструкция елки, количество ветвей, комплектация елки, особенности елки, тип елки и материал елки. Попытка заполнить их из названия была чревата возможным искажением – мы не стали рисковать
7. Пропуски значений по высоте елки заполнили с учетом разбиения елок на категории по высоте, ширине и весу упаковки
8. Удалили 260 оставшихся дубликатов
9. В итоге мы получили 6221 строчку полностью заполненных и готовых для анализа данных

Весь процесс в ноутбуках, ввиду сложностей командной работы очистку проводили в 2 этапа:

1. Первая очистка:

Ноутбук:

<https://github.com/calabarOlga/dataton_christmas_tree/blob/main/notebooks/Очистка_первый_этап.ipynb>

Данные:

<https://github.com/calabarOlga/dataton_christmas_tree/blob/main/data/cleaned_up_transformed_data.xlsx>

1. Вторая очистка:

Ноутбук:

<https://github.com/calabarOlga/dataton_christmas_tree/tree/main/notebooks/final_data.ipynb> - итоговый ноутбук или вот тут:

<https://disk.yandex.ru/d/Z0V6LVp3pDiM2g>

Данные:

<https://github.com/calabarOlga/dataton_christmas_tree/blob/main/data/final_data_for_analysis.xlsx> — вот он итоговый датасет!!!

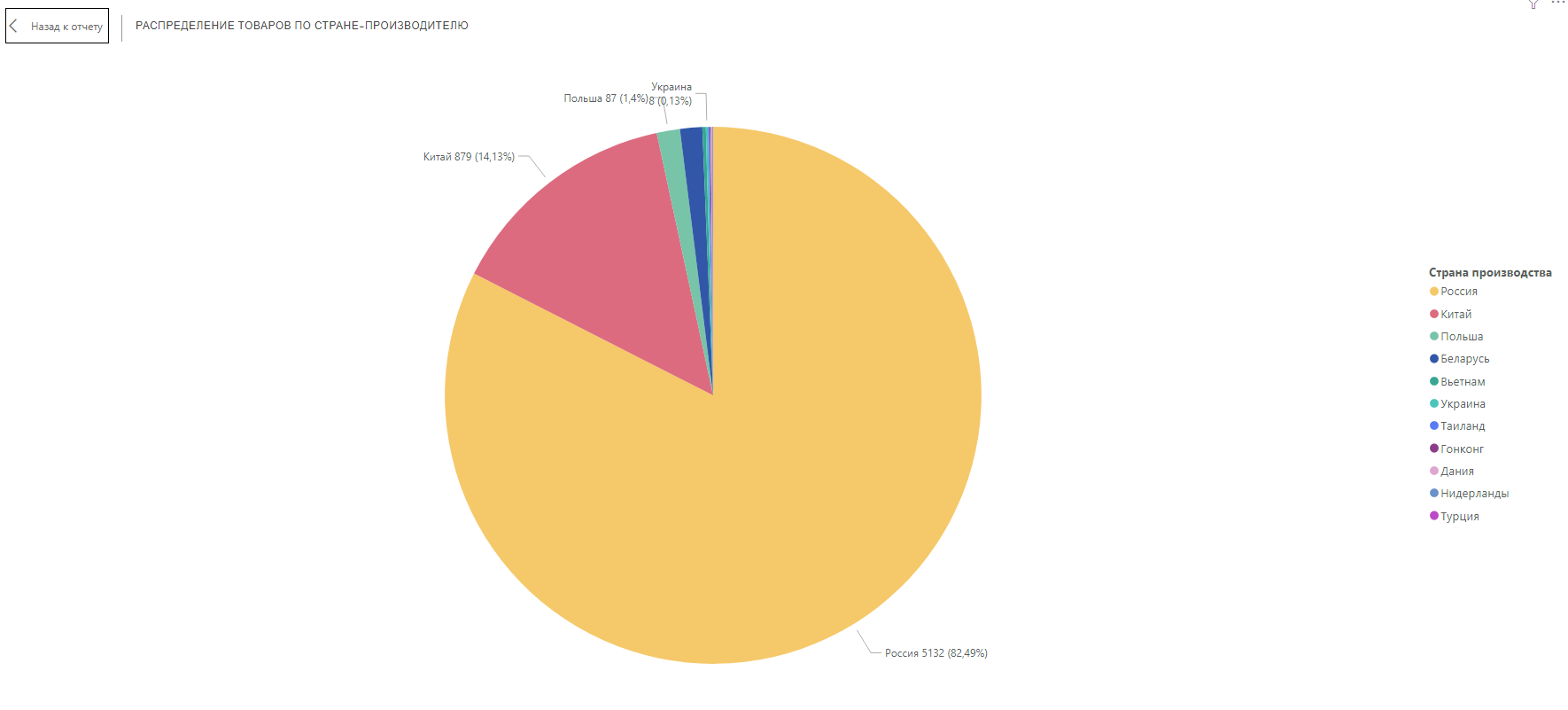
## Анализ данных:

- Какую елочку рекомендовать посетителям сайта «Wildberries»?

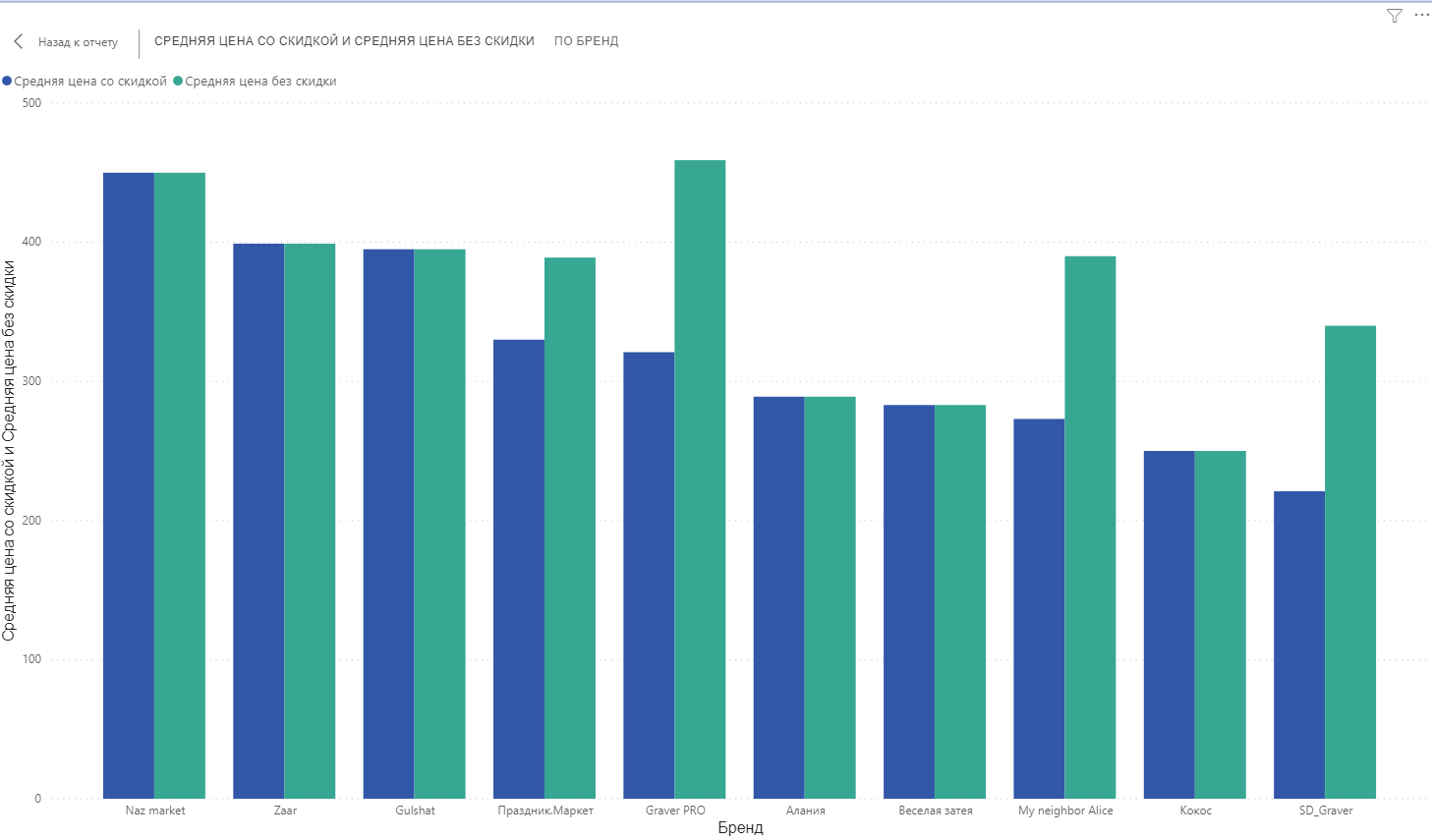
Полученные данные позволяют проводить аналитику, по сочетаниям разных категорий. И это очень интересно. К сожалению, времени у команды было очень не много, поэтому покажем лишь некоторые из возможных аналитических выкладок:

Вот распределение по странам производителям, мы видим, что Российские производители – 83 % рынка заработали на елках более 46 миллионов рублей (цена с учетом скидки)!

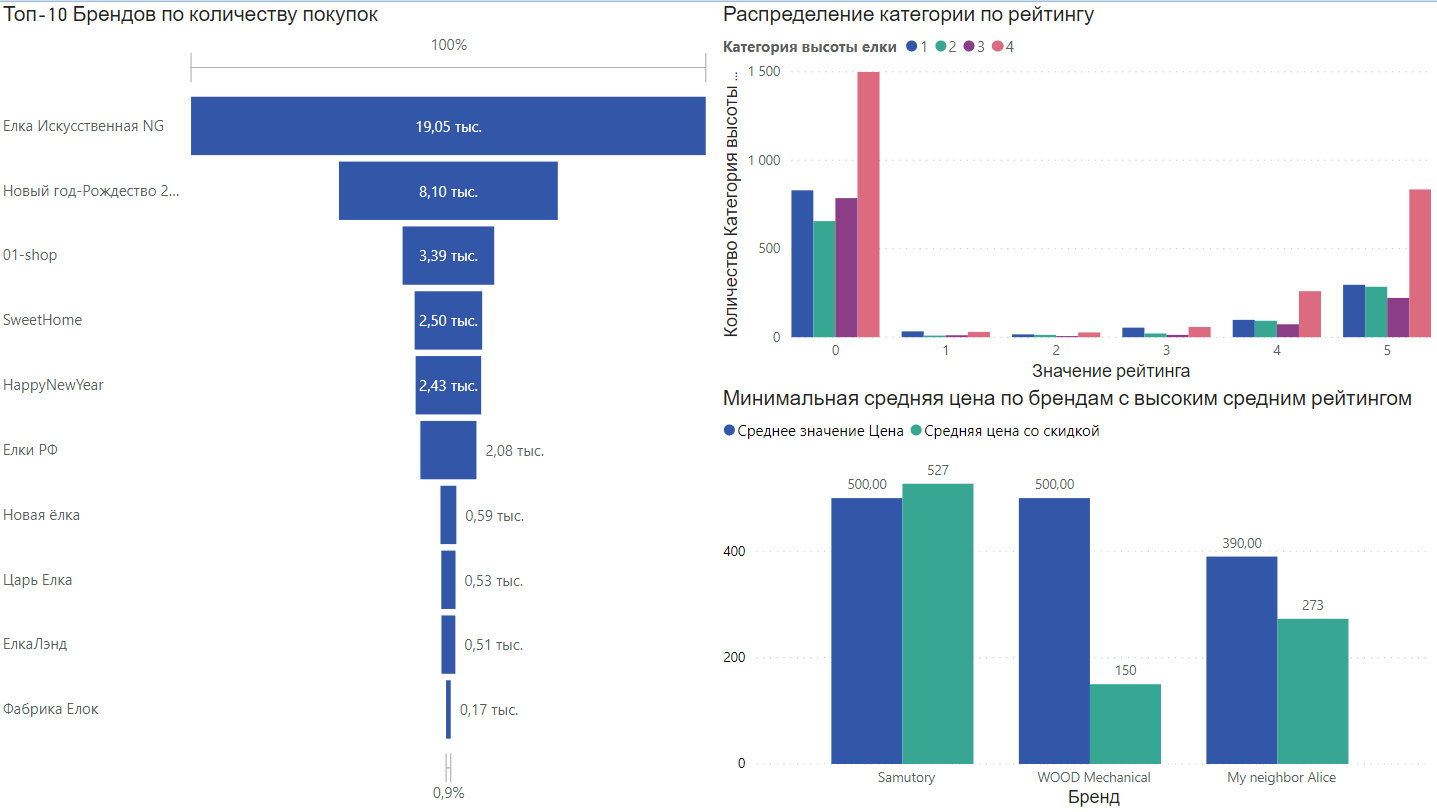




Мы можем посмотреть какие бренды делают максимальные скидки:



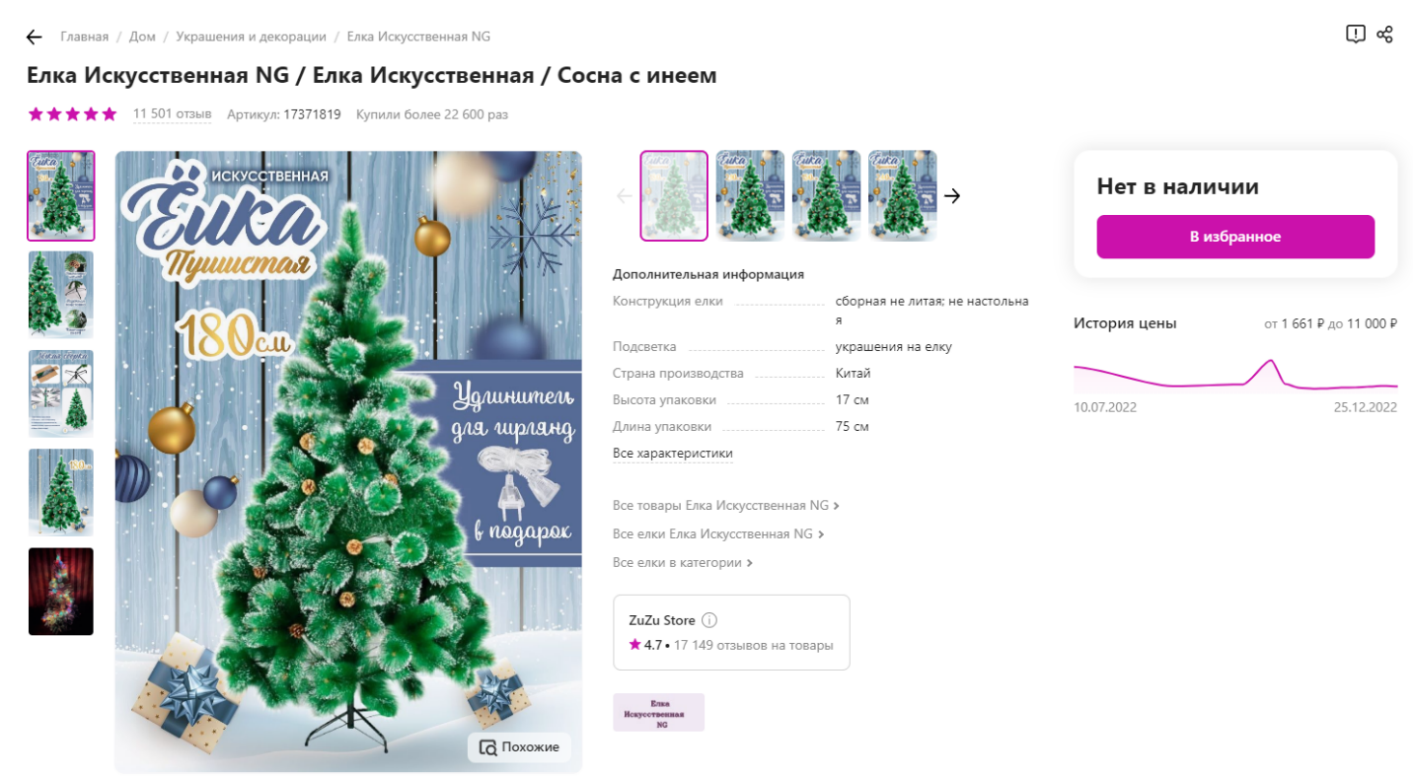
А вот анализ лучших брендов в срезе количества покупок и высоты, мы видим, например, что елки средней категории сильно уступают по количеству маленьким и самым высоким. Если вы ищите маленькую елочку по хороший выбор по соотношению цена – рейтинг будут бренды: Samutory, WOOD Mechanical и My neighor Alisе, ну а наиболее покупаемый бренд на Wildberries – это Китайский производитель «Елка Искусственная NG»



А вот и ответ на вопрос – какая елка лучшая:

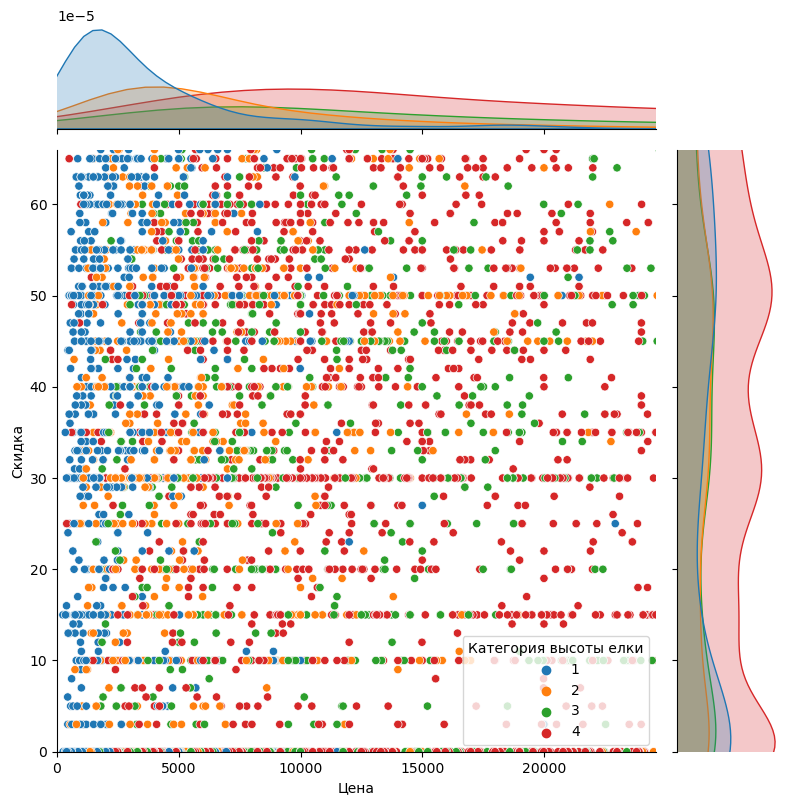




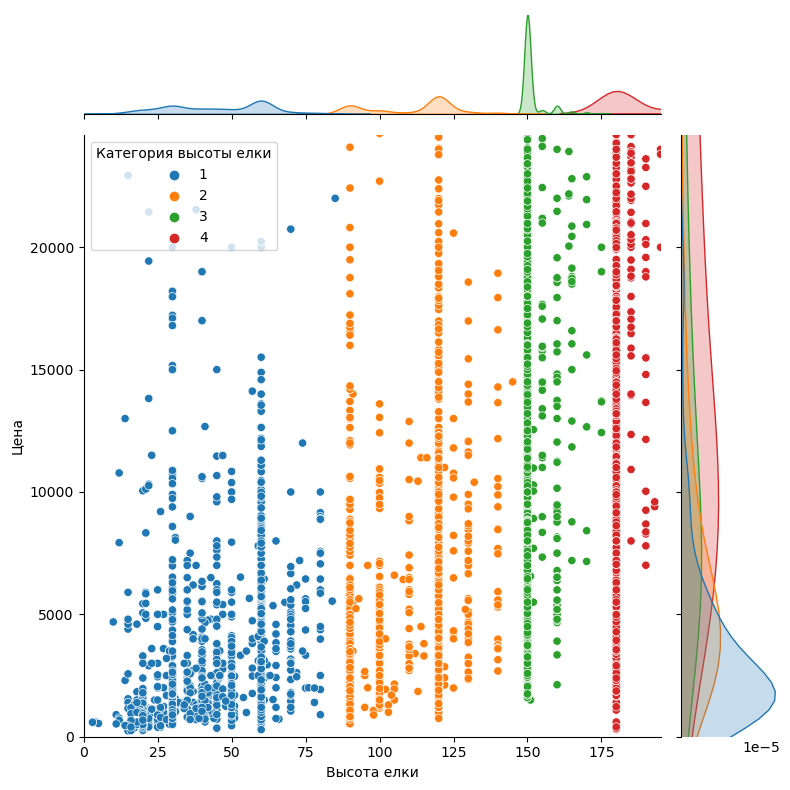


Не будем забывать и про возможности визуализации на python. Приведем несколько удачных визуализаций:

Вот интересный график, на котором можно увидеть закономерность распределения цены и скидок в зависимости от категории елки по высоте, мы видим, что самые маленькие елки наиболее дешевые, самые популярные скидки 15, 20,30, 50, 60 %



Неплохо визуализируется и самая очевидная зависимость, цена от высоты елки, распределение выглядит еще более ярким если добавить раскраску по категории (высота):



Уникальный по красоте и полезности получился график реализованной кластеризации значимых негабаритных числовых признаков:

 'Скидка',

 'Цена',

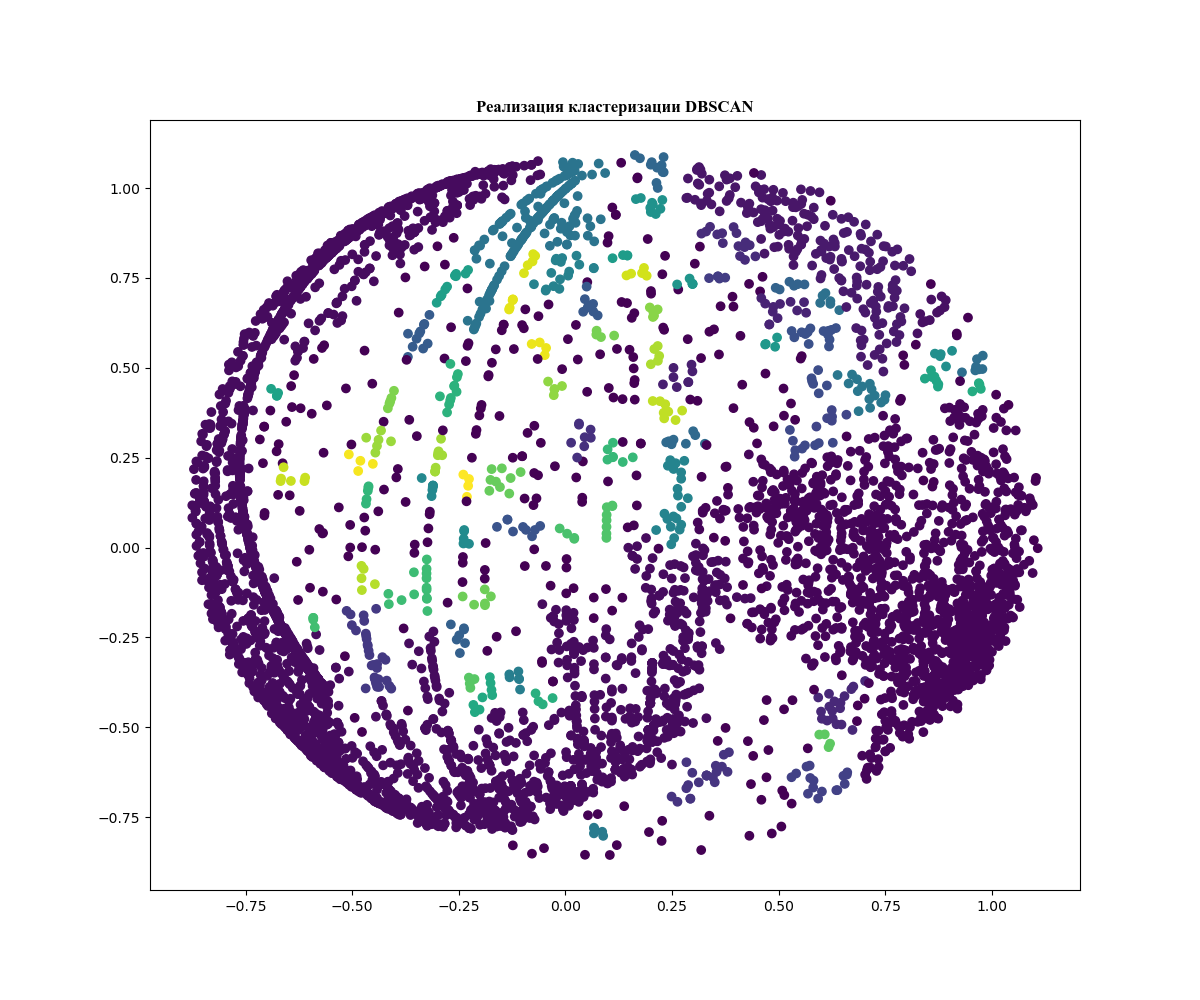
 'Цена со скидкой',

 'feedbacks',

 'rating',

 'Количество раз купили',

Его цель распределить «без учителя» товары на кластеры, рекомендованные к покупке (синий, желтый и зеленый цвета с оттенками) и не рекомендованные (фиолетовый):



Итак, мы готовы рекомендовать покупателям лучшие елки на «Wildberries»!

Это далеко не все возможности собранного датасета, он готов к реализации и других задач машинного обучения, скажем ничего не мешает выявить коэффициенты характеристик, наиболее существенно влияющих на цену, есть вариант определения реальной цены и «фиктивных скидок». А можно выбрать себе елку-мечты просто настроив фильтры столбцов по желаемым характеристикам в Excel. Наш датасет может помочь покупателю, но, наверное, сейчас главное, что он помог нашей команде сработаться, реально погрузиться в профессию и неплохо прокачаться.

С благодарность! Команда «Excellen’s»