# Отчет по результатам работы 3-го дня команды «Excellent’s»

## Парсер:

Наша команда применила парсер найденный в открытых источниках, 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>

! ПЕРВОЕ ПРИБЛИЖЕНИЕ К ДАННЫМ

Командой были доработаны функции с учетом особенностей продукта + добавлено описание функций и комментарии к коду, а также дополнили код, который переводит данные в формат excel.

## результат отработки парсера

В итоге мы получили 9207 строк по категории новогодние елки. Мы сразу обратили внимание, что ценность представляют следующие столбцы (хотя как потом оказалось все они представляют интерес):



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 непосредственно характеристики елок

Очевидно, что из описания можно «достать» как минимум высоту елки и другие характеристики. Плюс команде захотелось посмотреть и другую статистику: сколько раз купили, вес, упаковка, сопутствующие товары. Тогда мы нашли и доработали еще один парсер, который получает данные непосредственно со страниц сайта.

|  |
| --- |
| 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 |

[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)

А страницы у нас уже есть – мы добыли их первым парсером. Однако задача оказалось не из легких, простой код (все парсеры уже выложены в репозитарии) работал очень долго, а начать анализировать на неполных данных как-то не хотелось. Тогда мы решили пойти на эксперимент и отработать датасет с помощью горячо любимого всеми нами Excel, и как оказалось это оправдано, мы увидели те проблемы, с которыми столкнемся потом при отработке кодом уже большого датасета.

## задачи по очистке данных и ее методология на EXCEL

Развертывать очистку и поиск характеристик мы начали с отработки названия (ему больше подходит наименование - описание) «высота». 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:



Работа заняла 3 часа, но это данные, с которыми уже можно работать.

## Второй парсер

Тем временем второй парсер скачал данные по 2000 страницами, и мы увидели, что он дает неплохие фичи, ночь впереди осталось еще 7000 и будет интересно). Цифры – это количество непустых значений.

Изображение выглядит как текст

Автоматически созданное описание

## Разработка репозитария на Git-hub

Для сдачи отчетности нам необходимо было сделать репозиторий на GitHub и README и мы это сделали и загрузили в папки код парсеров и предварительные данные.

## Задача на завтра

* Сформировать датасет
* Очистить данные
* Проанализировать
* Визуализировать выводы
* Реализовать возможности ML