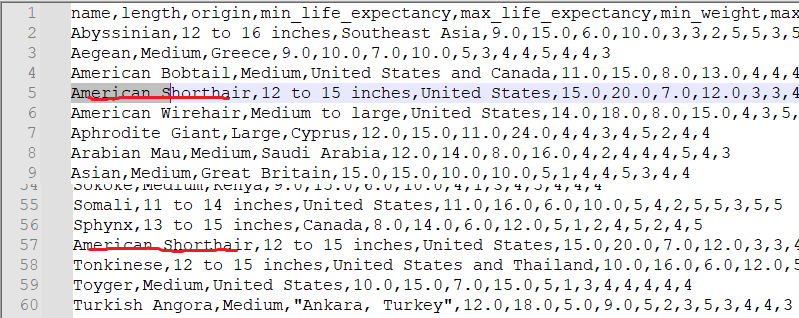
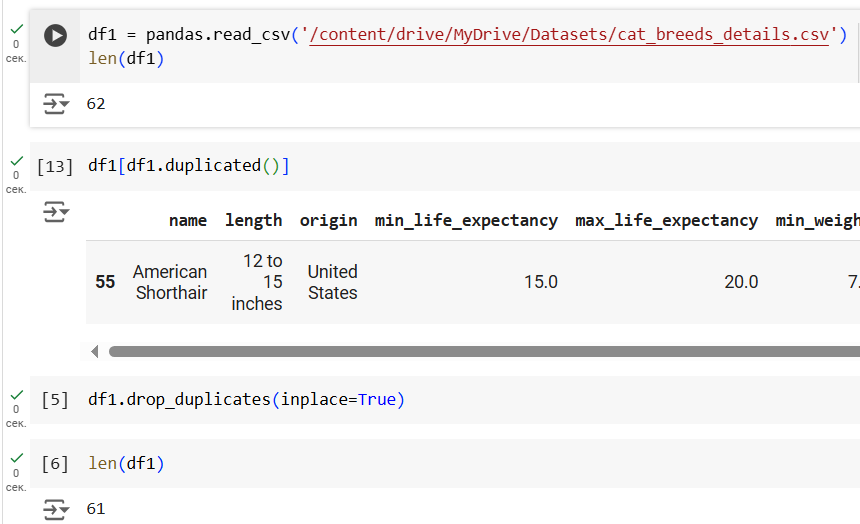
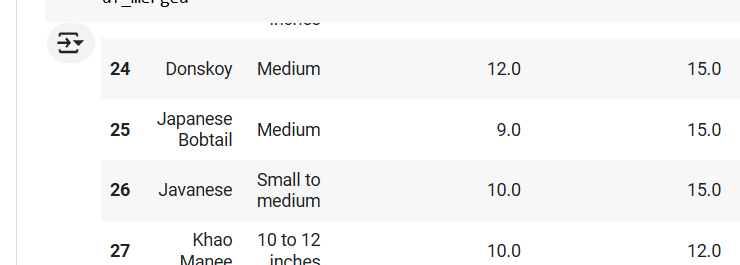
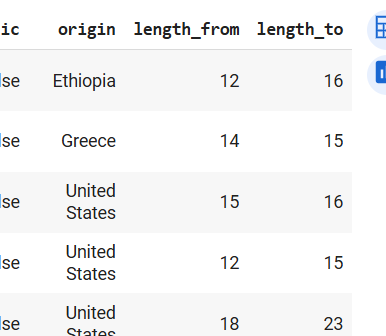
Для выполнения работы был выбран набор данных уникальных пород котов и их характеристик из сервиса Kaggle <https://www.kaggle.com/datasets/warcoder/cat-breeds-details>. Как будет сказано ниже, набор будет дополнен несколькими свойствами из других данных.

Для удобства работы с данными их необходимо подготовить.

Имеются дубликаты, поэтому их нужно удалить

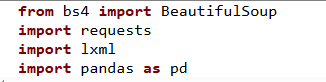


Также можно увидеть, что длины котов представлены в виде текста и в разных форматах. Была проведена «нормализация», с разбиением на два столбца length\_from и length\_to.

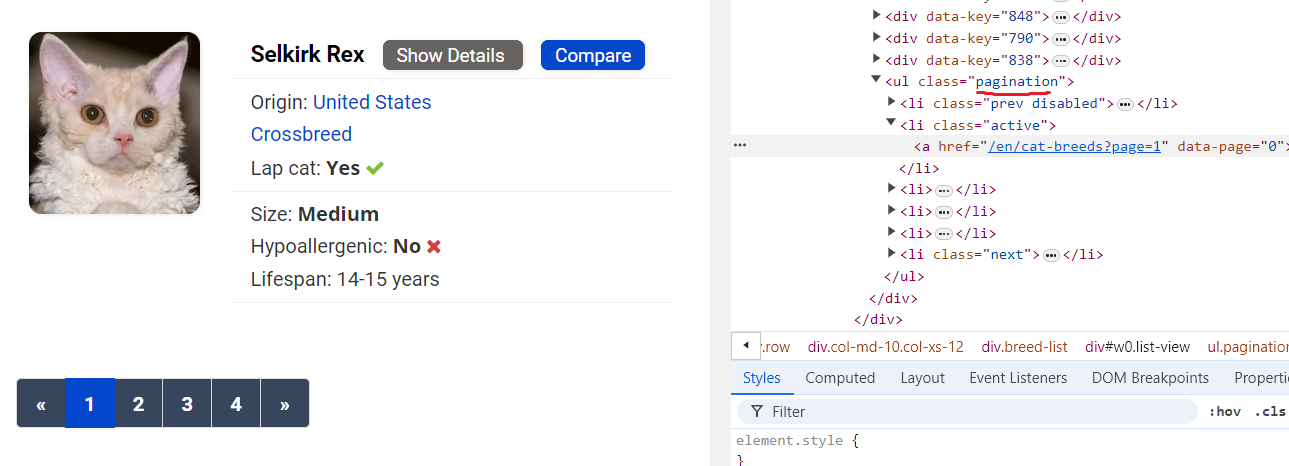
 

Имеющийся набор данных дополним с помощью анализа сайта пород кошек <https://dogell.com/en/cat-breeds>, внешний вид которого показан на рисунках ниже.

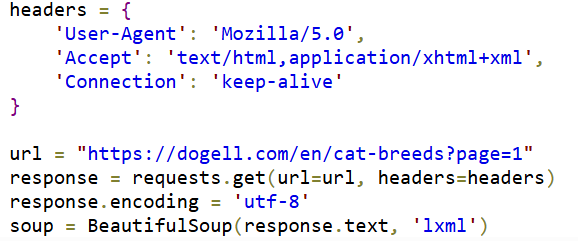
Для получения данных использовались библиотеки для веб-парсинга, такие как Beautifulsoup4 и Requests:



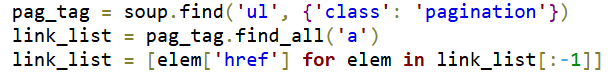
Страница имеет пагинацию, поэтому сначала необходимо определить все ссылки для анализа. Элемент пагинации на сайте имеет следующий вид:



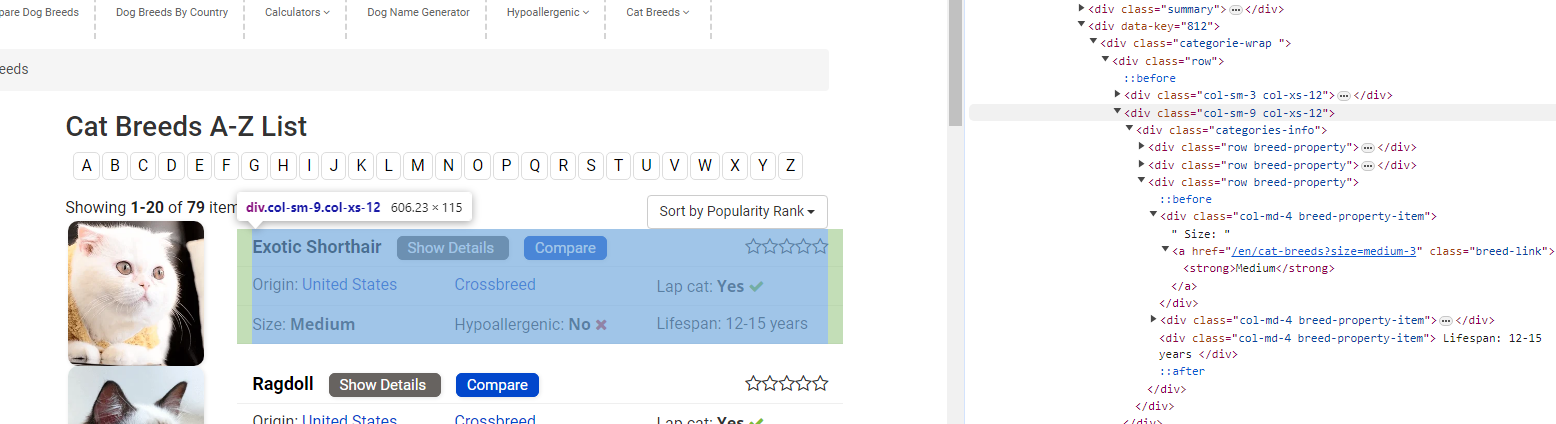
Определим страницы парсинга данных с помощью следующего отрывка кода. Сначала получим страницу, настроив заголовок запроса и преобразовав кодировку:

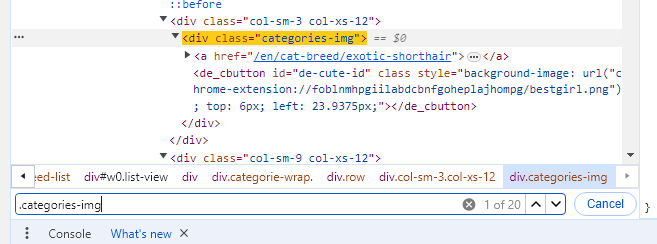


После чего можно найти набор страниц. Последний элемент исключен, т.к. относится к «стрелке» и указывает на следующую страницу:

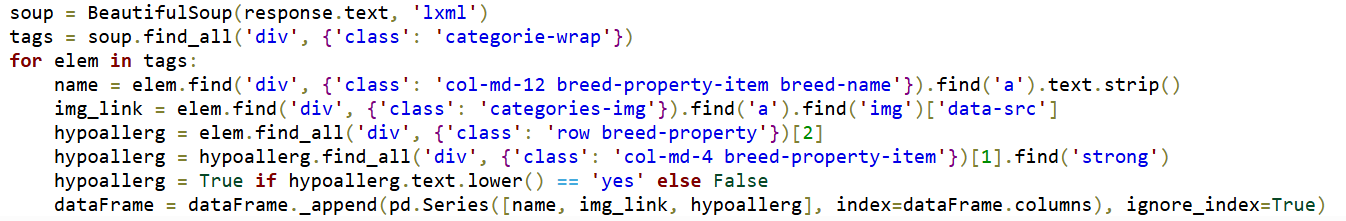


Получив все ссылки, можно проанализировать каждую страницу для определения дополнительных параметров – ссылки на изображение породы, регион и новое свойство «гипоаллергенности». Расположение ссылки и структура блоков с информацией о породе показаны на рисунке ниже:





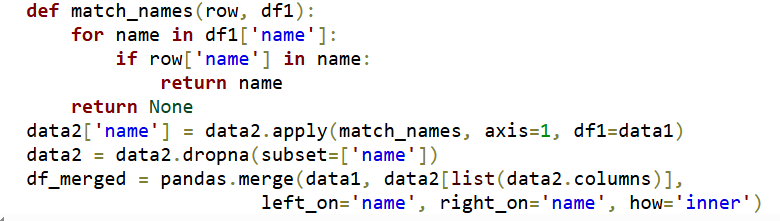
С помощью следующего скрипта можно получить необходимые свойства, которые будут сохранены в dataframe (для сохранения в csv).



В результате был получен следующий набор данных:



После этого было проведено «объединение» данных. Для этого названия порог приводится к одному виду и затем в результирующий dataframe сохраняется пересечение данных:



Результирующий набор данных представлен в файле mainData2DB.csv.

Далее, для более удобного импорта данных в таблицы базы данных сразу из csv-файлов, единый датафрейм pandas был разбит на несколько наборов, соответствующих таблицам БД.



Заключение

В результате выполнения работы были освоены и применены инструмента для обработки данных (pandas, numpy). Были использованы такие инструменты как BeautifulSoup4 и библиотека requests для парсинга информации, а также применены полученные знания по работе с базами данных и обработки с помощью языка SQL.

ПРИЛОЖЕНИЕ

Импорт сгенерированных csv файлов в базу данных.



