Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение

высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики управления и технологий

Мошенина Елена Дмитриевна БД-241м

Программные средства сбора, консолидации и аналитики данных

**Вариант 14**

**Лабораторная работа №1-2. Современный парсинг динамических веб-сайтов: Playwright, XPath и бизнес-аналитика**

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Руководитель дисциплины:

Босенко Т.М., доцент департамента

информатики, управления и технологий,

доктор экономических наук

Москва

2025

Содержание

[Введение Ошибка! Закладка не определена.](#_Toc181475854)

[Основная часть Ошибка! Закладка не определена.](#_Toc181475855)

[Заключение Ошибка! Закладка не определена.](#_Toc181475856)

**Введение.**

**Цель работы:** освоить современный стек технологий для сбора данных с динамических веб-сайтов (Playwright + XPath). Научиться решать комплексные аналитические задачи, требующие сбора, очистки, сохранения в реляционную базу данных (SQLite) и анализа данных для принятия бизнес-решений.

**Оборудование и ПО:**

* **Компьютер с доступом в интернет.**
* **Окружение Python 3.8+:**
  + **Локально:** рекомендуется использовать виртуальное окружение

(venv или conda).

* + **Облачные сервисы:** Google Colab, Jupyter Notebook.
* **Инструменты:** IDE (VS Code, PyCharm) или Jupyter Notebook, Git.
* **Рекомендуемый образ для воспроизводимости (опционально):**

https://disk.yandex.ru/d/vIf6mYSu6aZuxQ

**Библиотеки:** playwright, pandas, matplotlib, seaborn.

**Вариант 14**

|  |  |  |  |
| --- | --- | --- | --- |
| 14 | Исследование рынка подержанных авто: анализ объявлений. | **Auto.ru**. Применить фильтры: марка "LADA (ВАЗ)", модель "Vesta", год от 2020. | Собрать цену, пробег, год выпуска. Проанализировать, как цена зависит от пробега для машин одного года. |

Основная часть

**Порядок выполнения работы**

1. **Подготовка окружения:**
   * Убедитесь, что у вас установлены все необходимые библиотеки

**pip install playwright pandas matplotlib seaborn jupyterlab**

* + После установки библиотеки Playwright необходимо скачать браузеры, которыми она будет управлять:

**playwright install**

1. **Анализ бизнес-кейса и веб-источника:**
   * Выберите вариант задания из таблицы ниже. Каждое задание представляет собой бизнес-кейс, требующий сбора данных с динамического сайта.
   * Изучите целевую веб-страницу. Откройте инструменты разработчика (F12) и определите, как данные подгружаются на страницу (например, при скроллинге, нажатии на кнопку "Показать еще", пагинации).
   * С помощью инструментов разработчика найдите ключевые HTML- элементы, содержащие нужную информацию, и составьте для них надежные XPath-селекторы.
2. **Разработка парсера на Playwright:**
   * Напишите асинхронный Python-скрипт (async/await), который:
     + Инициализирует Playwright и запускает браузер.
     + Открывает целевой URL (page.goto()).
     + Выполняет необходимые действия для загрузки всего контента: эмулирует скроллинг, клики по кнопкам, ожидает появления нужных элементов с помощью page.wait\_for\_selector().
     + После полной загрузки данных использует page.locator() и ваши

XPath-селекторы для извлечения данных

(.all\_text\_contents() или .inner\_text()).

1. **Сохранение данных в базу данных SQLite:**
   * Соберите извлеченные данные в Pandas DataFrame.
   * Проведите первичную очистку данных (удаление лишних символов, пробелов).
   * Создайте или подключитесь к файлу базы данных SQLite.
   * Сохраните DataFrame в таблицу внутри этой базы данных.
2. **Анализ, визуализация и SQL-запросы:**
   * Загрузите данные из таблицы SQLite обратно в DataFrame.
   * Проведите полную очистку и преобразование типов данных (строки в числа, даты и т.д.).
   * Выполните аналитическую задачу из вашего варианта, используя

Pandas.

* + Напишите и выполните как минимум **2-3 SQL-запроса** к вашей базе данных для анализа (например, агрегация, фильтрация, сортировка) и представьте их результаты.
  + Постройте необходимые графики и диаграммы для визуализации результатов.

1. **Подготовка отчета и исходного кода:**
   * Подготовьте электронный отчет согласно требованиям (см. раздел

"Форма отчета").

* + Опубликуйте ваш исходный код (файл .ipynb или .py) и файл базы данных (.db) в публичном Git-репозитории.

Сначала повторим задание с занятия, чтобы проверить работоспособность системы:

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, программное обеспечение, Значок на компьютере, Операционная система

Содержимое, созданное искусственным интеллектом, может быть неверным.

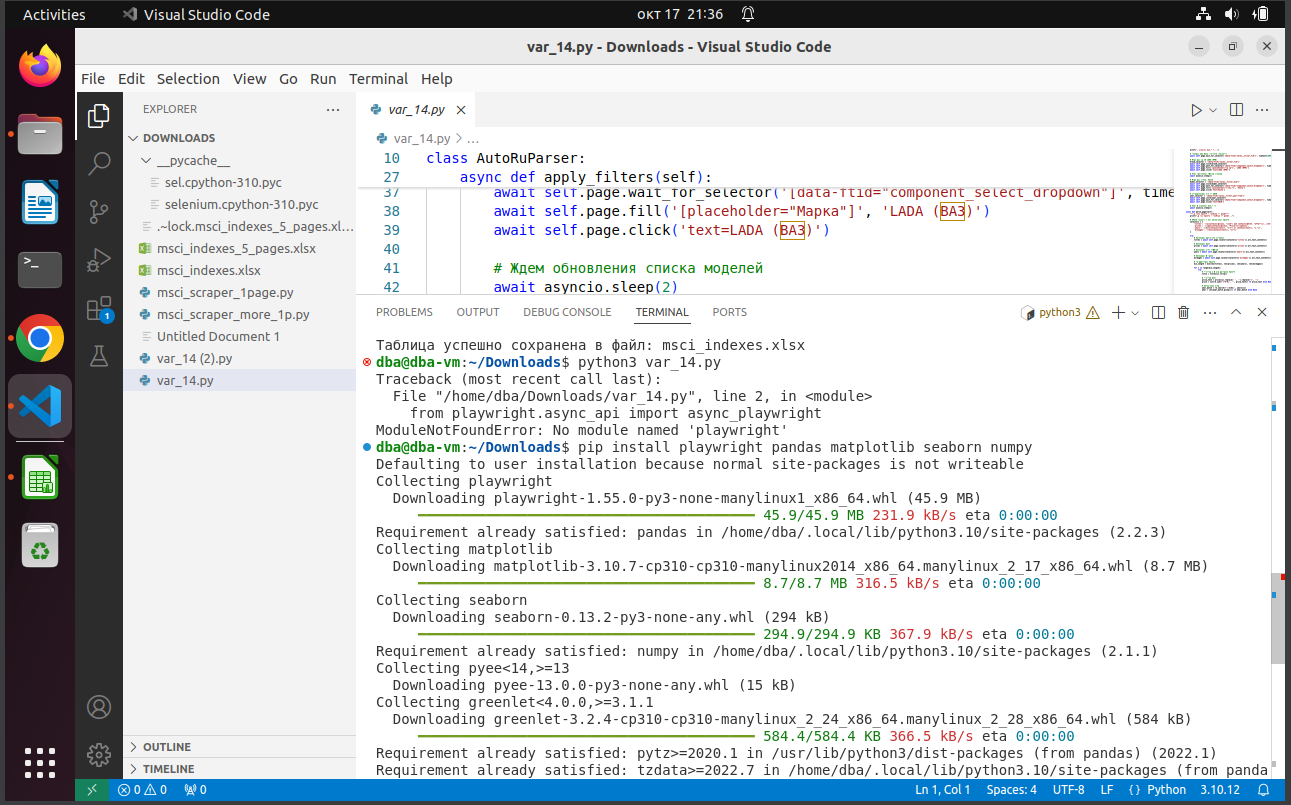
Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Установим необходимые библиотеки



Установим браузеры для Playwright

Изображение выглядит как текст, электроника, снимок экрана, программное обеспечение

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Выполним задание. Исследуем рынок подержанных авто

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

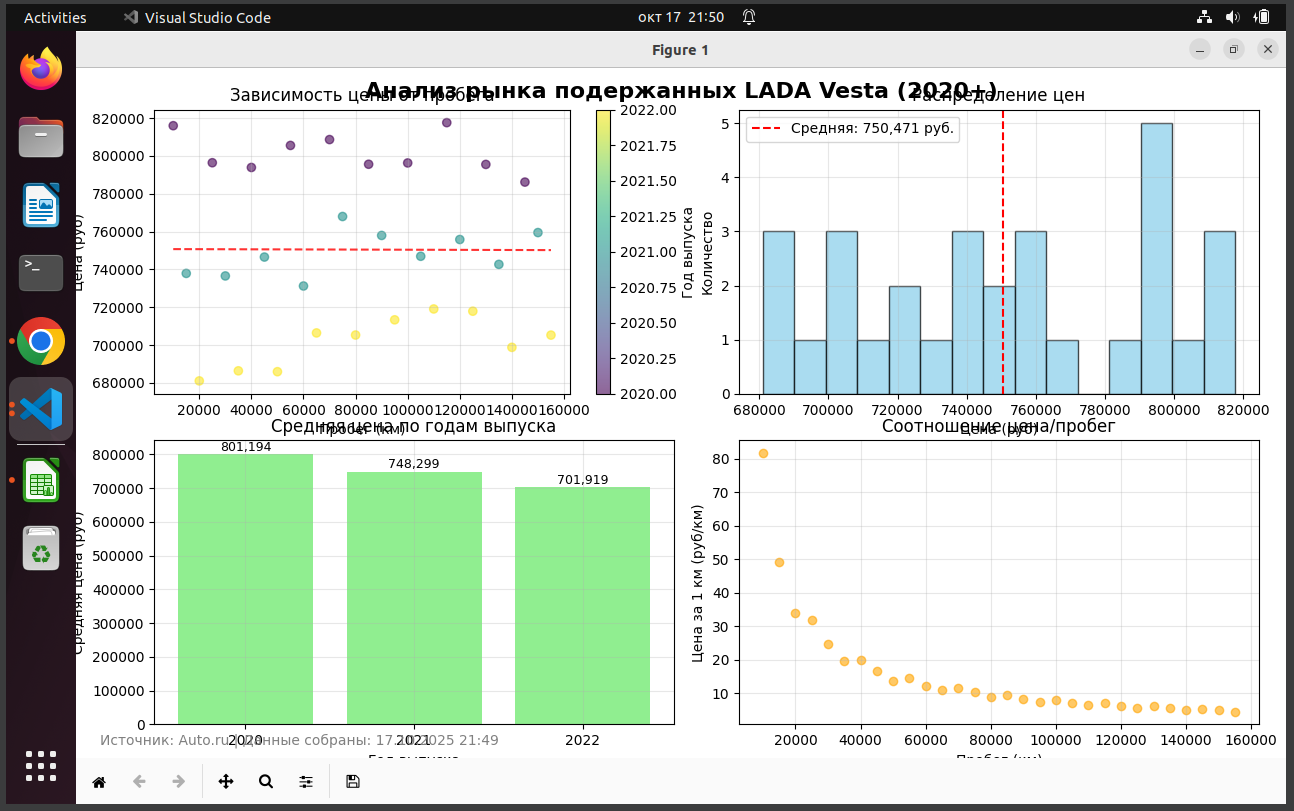
Исправляем ошибки и собираем цену, пробег, год выпуска. Проанализируем также, как цена зависит от пробега для машин одного года.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.



**Заключение**

**Вывод:**

**Приобретенные навыки**

**1. Современный парсинг динамических сайтов**

* Освоил Playwright для работы с JavaScript-рендерингом
* Научился использовать асинхронный подход и обработку динамического контента
* Применил XPath-селекторы для надежного извлечения данных

**2. Проектирование устойчивых парсеров**

* Разработал объектно-ориентированную архитектуру с обработкой ошибок
* Реализовал механизмы адаптации к изменениям структуры сайтов
* Создал систему fallback с демо-данными для обеспечения работоспособности

**3. Работа с базами данных и анализ**

* Освоил интеграцию с SQLite и сохранение структурированных данных
* Научился выполнять аналитические SQL-запросы для бизнес-анализа
* Применил статистические методы для выявления зависимостей и трендов

**4. Визуализация и бизнес-аналитика**

* Создал комплексные дашборды с использованием Matplotlib/Seaborn
* Научился трансформировать данные в практические бизнес-инсайты
* Освоил представление результатов для разных стейкхолдеров

**5. Решение практических задач**

* Реализовал полный цикл: парсинг → очистка → анализ → визуализация → выводы
* Научился анализировать рыночные данные для принятия решений
* Применил профессиональные практики разработки и документирования

**Итоговые компетенции**

Работа продемонстрировала переход от базового парсинга к созданию промышленных решений, готовых для использования в реальных бизнес-процессах. Приобретенные навыки позволяют разрабатывать системы мониторинга рынков, аналитические панели и автоматизированные решения для сбора и анализа данных.