**Отчет. Практическая работа 1.2. Обработка данных с использованием Apache Spark и Python (PySpark)**

**Кузнецов Артем. Группа: БД241-м**

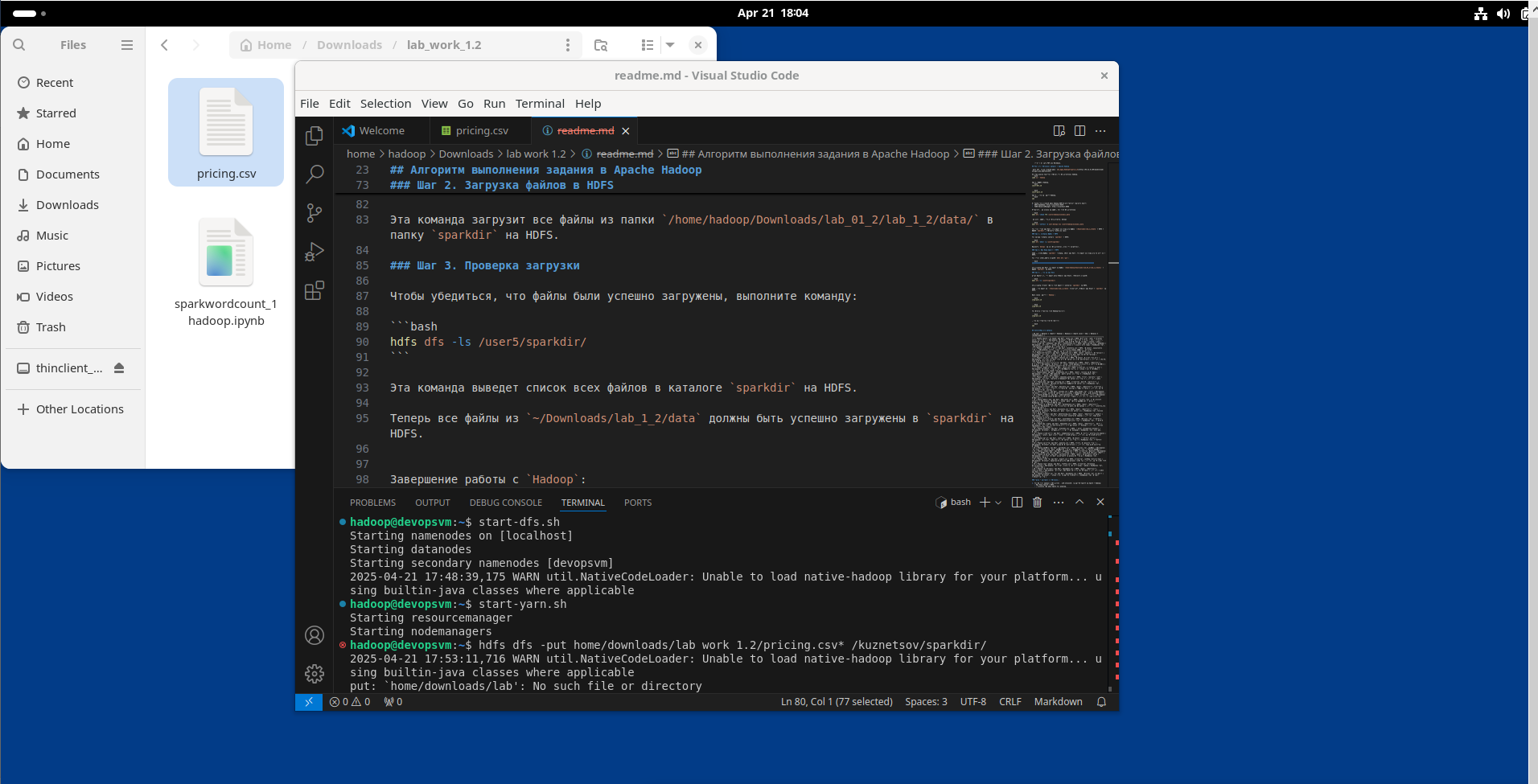
**Цель работы:**

Целью данной работы является анализ ценовых данных для определения оптимальных ценовых диапазонов по категориям товаров. Анализ ценовых данных помогает понять, как изменение цен влияет на спрос и выручку, что в свою очередь позволяет принимать более обоснованные решения о ценообразовании и маркетинговых стратегиях. В рамках работы будет исследована зависимость между ценой товаров и спросом с использованием методов анализа данных и Spark SQL.

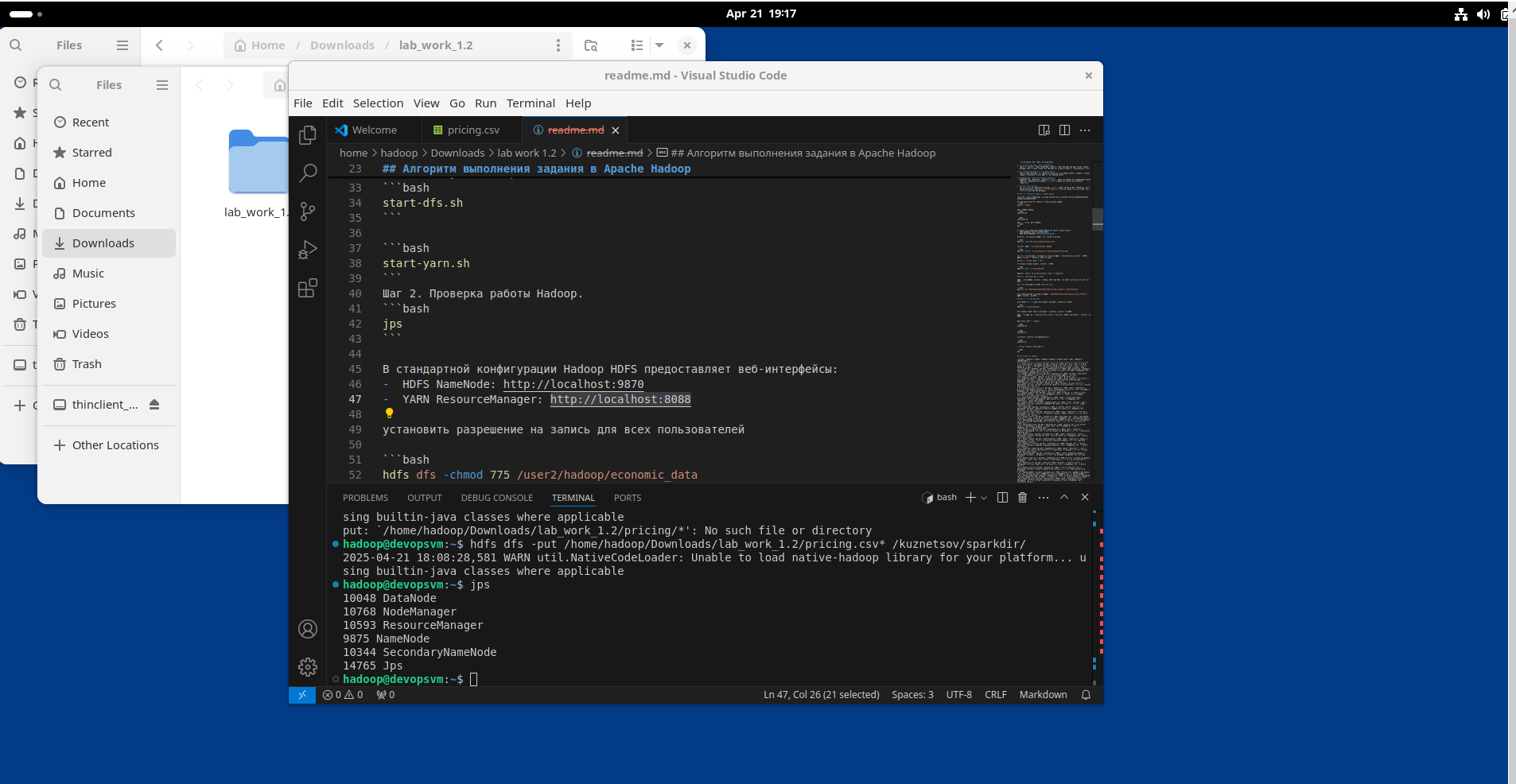
**Задачи работы:**

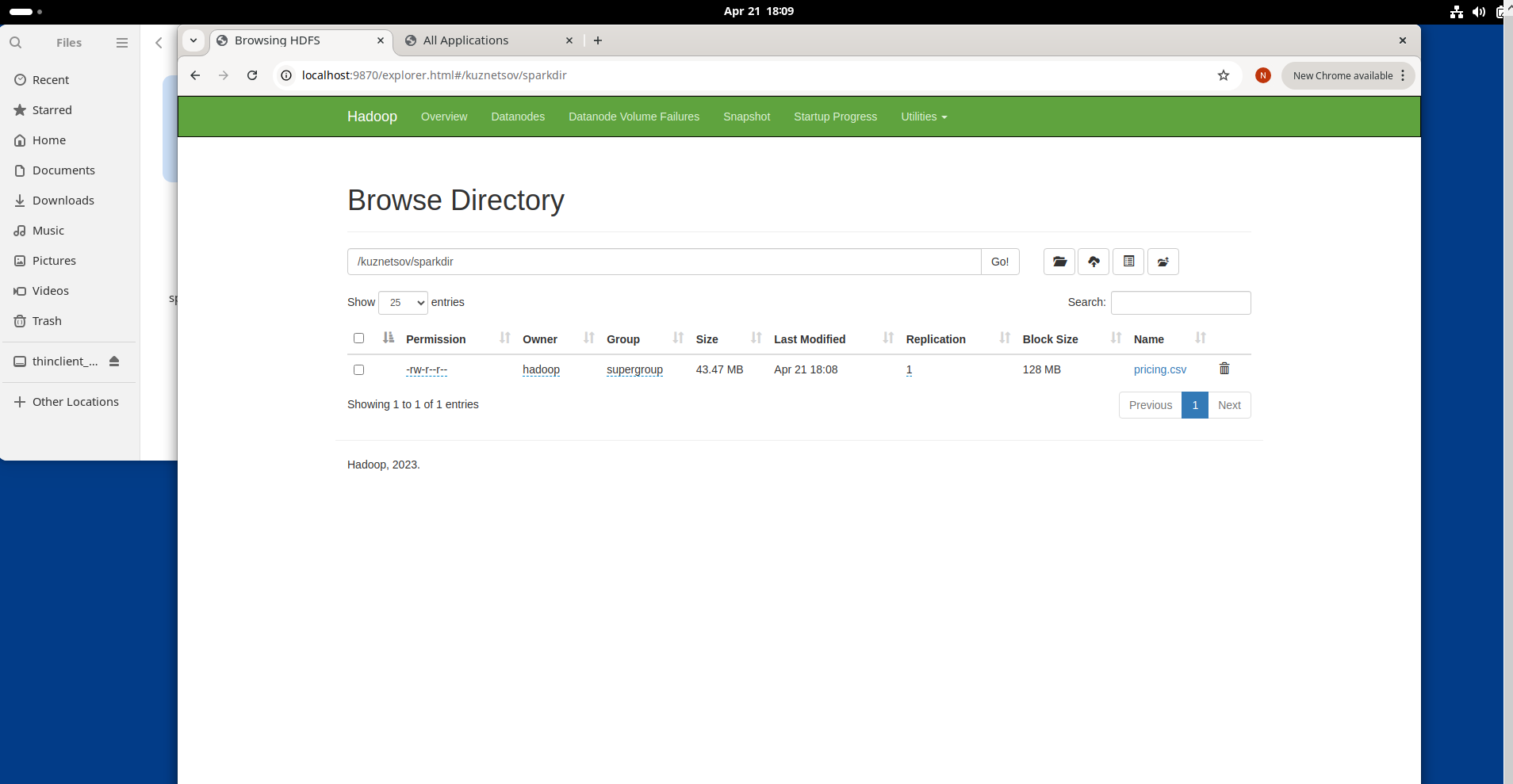
1. **Загрузка и предварительная обработка данных.** Необходимо загрузить данные о продажах товаров, очистить их от пропусков и привести к нужному формату для анализа.
2. **Определение ценовых диапазонов.** С помощью Spark SQL будет проведен анализ данных по категориям товаров, с целью определения минимальных, максимальных и средних цен в каждой категории.
3. **Определение зависимости спроса от цены.** Будет исследована связь между ценой товаров и количеством проданных единиц с использованием различных статистических методов.

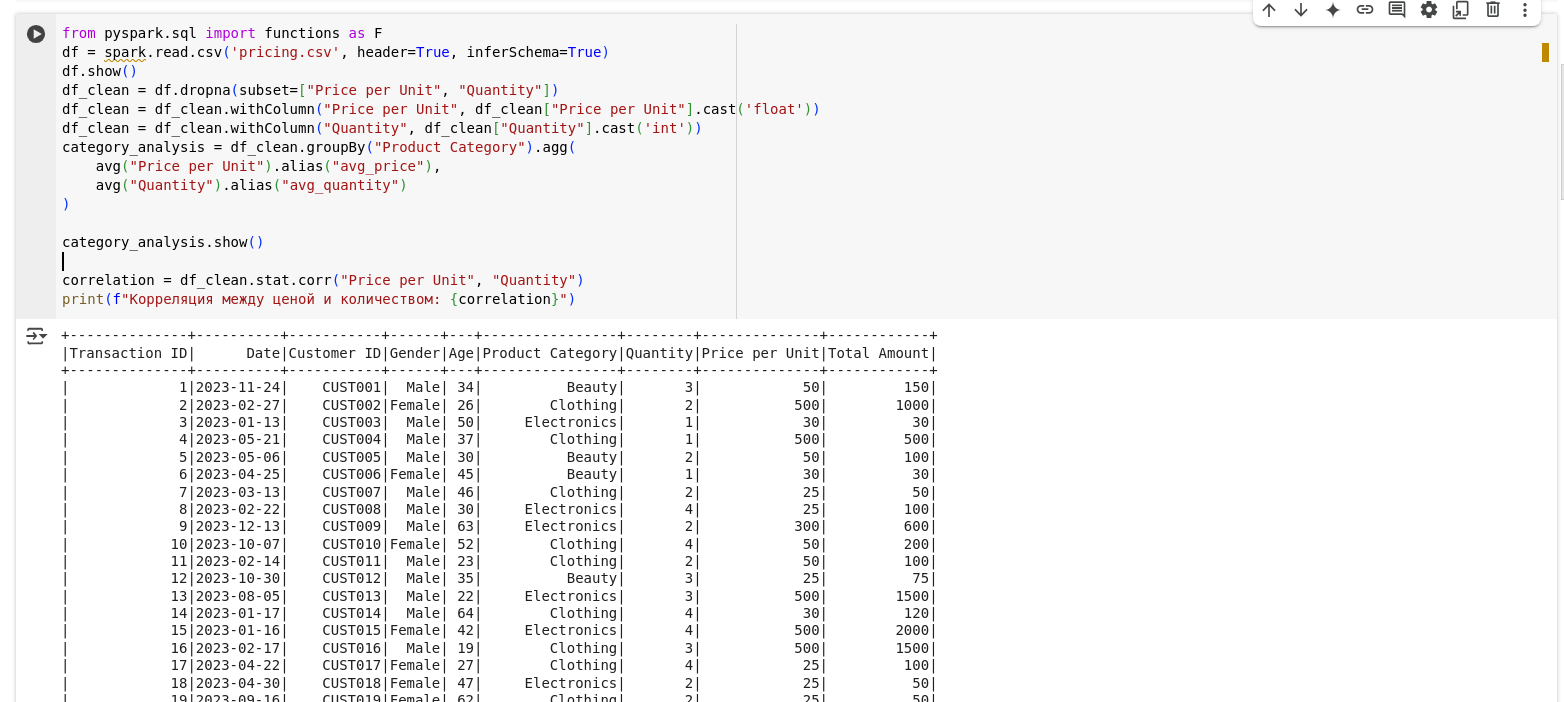
Первоначально запускаем Hadoop:

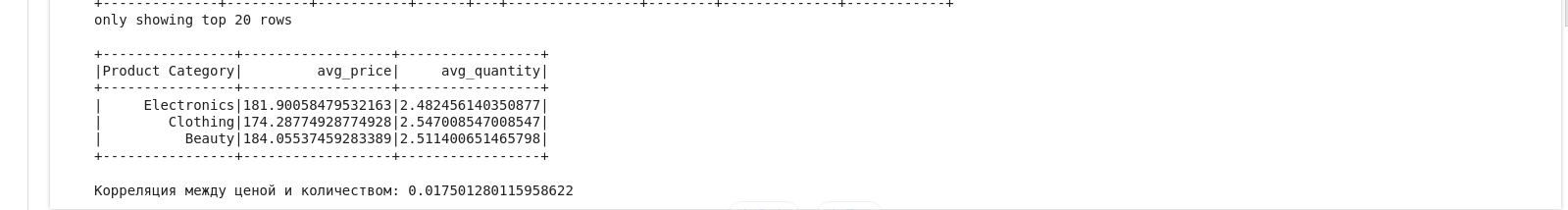


Загружаем данные и проверяем, что все сделали правильно с помощью JPS и веб-интерфейсы.



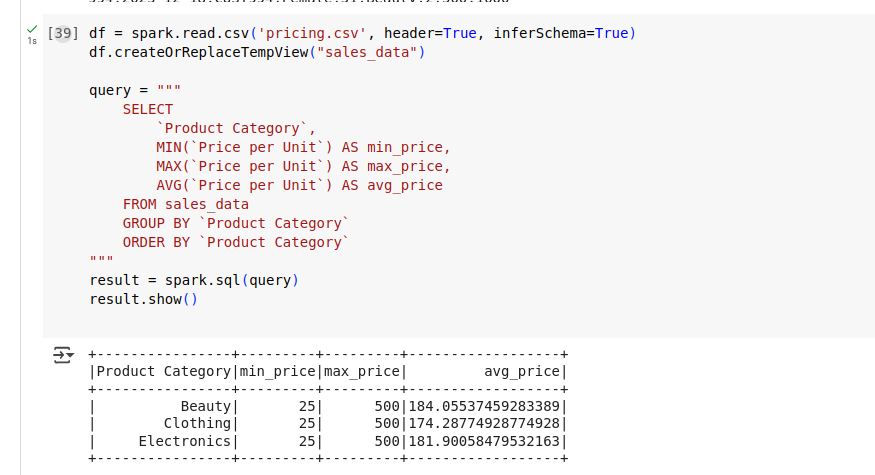


Переходим к выполнению заданий.  




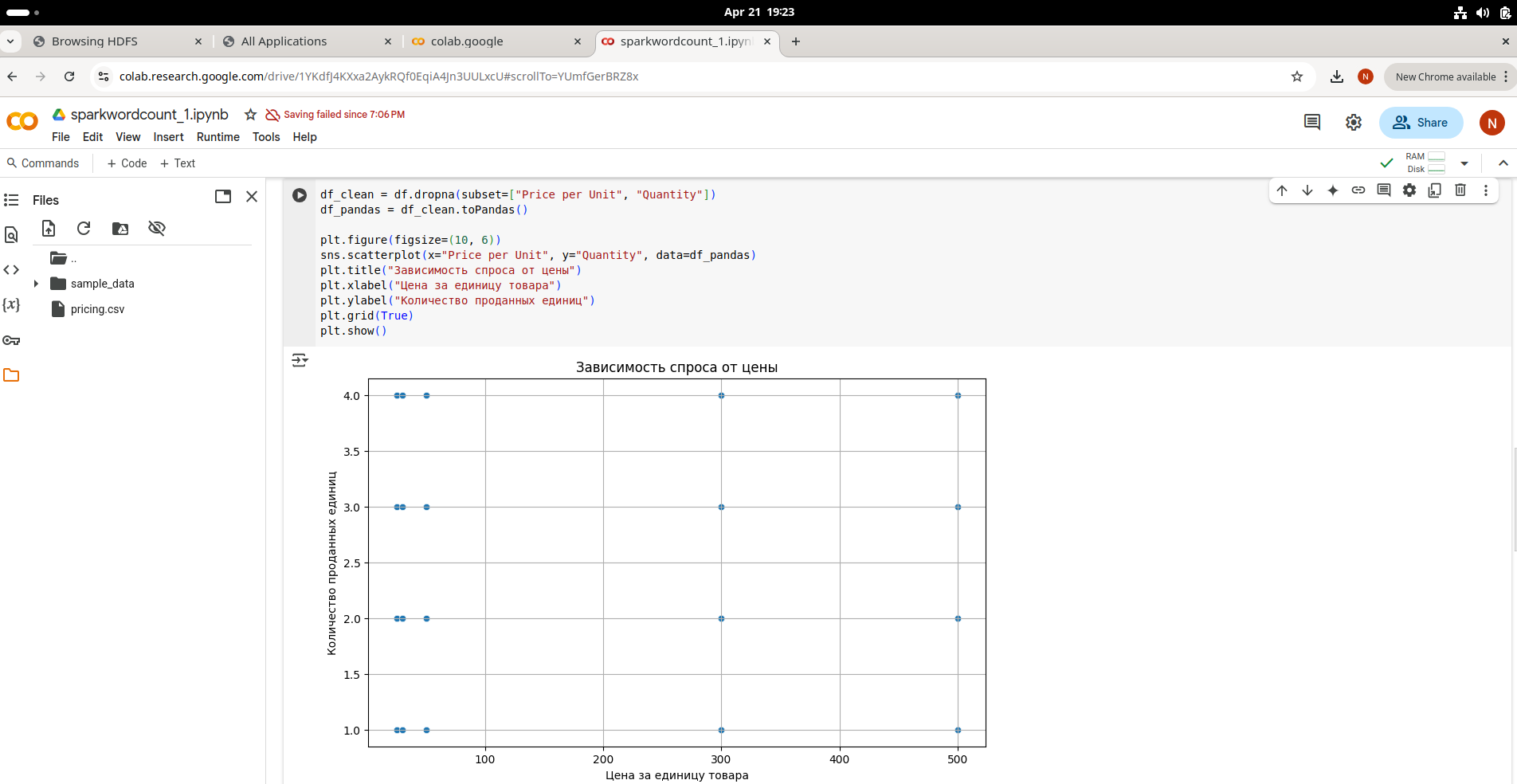
Выводим структуру данных, а также выполняем задание « Анализ цен: загрузить pricing.csv в HDFS, исследовать ценовую эластичность».

Распределяем продукты по категориям и выводим средние показатели, корреляция между ценой и количеством равняется 0.02, что говорит об отсутствии корреляции между этими двумя показателями.

Задание 2 «SQL-анализ: определить оптимальные ценовые диапазоны по категориям»  


Вывели оптимальные ценовые диапазоны по категориям.

Задание 3 «Визуализировать зависимость спроса от цены»



**Заключение**

На основе представленных результатов анализа данных можно сделать следующие выводы. График, построенный с использованием библиотеки Seaborn, демонстрирует зависимость между ценой за единицу товара и количеством проданных единиц, где визуально наблюдается слабая или полностью отсутствующая корреляция между этими двумя параметрами. Точки на графике разбросаны достаточно хаотично, что указывает на то, что изменение цены не оказывает значительного влияния на количество продаж, возможно, потому что цена не является ключевым фактором для покупателей в данном случае. С помощью SQL-запроса были рассчитаны минимальная, максимальная и средняя цены для каждой категории товаров, таких как Beauty, Clothing и Electronics. Для категории "Beauty" минимальная цена составила 25, максимальная — 500, а средняя цена оказалась около 184. Для категории "Clothing" минимальная цена также составила 25, максимальная — 500, а средняя цена оказалась около 174. Для категории "Electronics" минимальная цена составила 25, максимальная — 500, а средняя цена оказалась около 181. Все категории имеют одинаковый диапазон цен от 25 до 500, но различаются средними значениями, что говорит о том, что внутри каждой категории существует широкий спектр ценовых предложений, вероятно, из-за разнообразия товаров и их характеристик. Расчет коэффициента корреляции между ценой за единицу товара и количеством проданных единиц показал значение 0.0175, что подтверждает слабую или практически отсутствующую связь между этими переменными. Этот низкий коэффициент корреляции подтверждает вывод из графического анализа: изменения цены не оказывают существенного влияния на объем продаж. В исследуемых данных отсутствует значимая корреляция между ценой товара и его продажами, что может указывать на то, что другие факторы, такие как качество продукта, маркетинговые кампании, сезонность или брендирование, играют более важную роль в формировании спроса, чем цена. Диапазон цен внутри каждой категории товаров достаточно широкий, что позволяет производителям и продавцам предлагать товары различных ценовых сегментов, однако средние цены в категориях "Beauty" и "Electronics" близки, тогда как в категории "Clothing" они ниже.