

Практическое применение методов анализа данных на примере датасета онлайн-продаж

Лапушкина Елена Павловна

Приамурский государственный университет им. Шолом-Алейхема

Студент

Научный руководитель:

Глаголев Владимир Александрович

Приамурский государственный университет им. Шолом-Алейхема

К.г.н., доцент, доцент кафедры информационных систем и защиты информации

Аннотация

В данной статье описан пошаговый процесс подготовки данных и анализа продаж онлайн-магазина с использованием среды разработки Visual Studio и языка Python. В интегрированной среде Jupyter Notebook, установлены основные библиотеки pandas, numpy, matplotlib, seaborn. В исследовании особое внимание уделено вопросам комплексной обработки массивов данных средствами Python, охватывающей этапы предобработки, очистки, преобразования и визуализации.

Ключевые слова: Visual Studio, Jupyter Notebook, Python, анализ.

Practical application of data analysis methods using the example of an online sales dataset

Lapushkina Elena Pavlovna

Sholom-Aleichem Priamursky State University

Student

Scientific supervisor:

Glagolev Vladimir Alexandrovich

Sholom-Aleichem Priamursky State University

PhD, Associate Professor, Associate Professor of the Department of Information Systems and Information Security

Abstract

This article describes a step-by-step process for preparing data and analyzing online store sales using the Visual Studio development environment and Python. In the Jupyter Notebook integrated environment, the main libraries pandas, numpy, matplotlib, seaborn are installed. The study focuses on the issues of complex processing of data arrays using Python tools, covering the stages of preprocessing, cleaning, transformation and visualization.

Keywords: Visual Studio, Jupyter Notebook, Python, analysis.

1 Введение

1.1 Актуальность

Современная аналитика продаж онлайн-ритейла требует эффективных методов подготовки данных, моделирования и визуализации для извлечения бизнес-инсайтов. Это включает очистку и унификацию данных, обработку пропусков и дубликатов, а также создание признаков, отражающих сезонность, акции и поведение клиентов. Эффективная визуализация позволяет быстро идентифицировать тренды, а также сравнивать сегменты клиентов, каналы продаж и регионы.

1.2 Обзор исследований

Авторы О.К. Фарайдунов, З.А. Амонбекова описали основные подходы и инструменты Python для анализа данных в экономических задачах, включая обработку, визуализацию, статистический анализ и интерпретацию результатов, а также представили практический пример анализа продаж с использованием Pandas и Matplotlib [1]. Н.А. Моисеева, Т.А. Полякова исследовали критическую роль методов корреляционно-регрессионного анализа в условиях цифровой экономики как инструмента для обработки больших данных, прогнозирования производственных процессов и оптимизации управленческих решений в промышленности и строительстве. [2]. А.С. Ремнев, А.Р. Муртазина доказали преимущества применения языка Python над традиционными инструментами (такими как Excel) для анализа больших данных, выделив его способность эффективно обрабатывать крупномасштабные датасеты а также интегрироваться с технологиями машинного обучения и облачными системами, а также минимизировать ошибки за счёт автоматизации процессов [3]. Г. Капбар, Е. Ким описали методику обучения Python с использованием современных инструментов (Jupyter, Git, облачные платформы) и практико-ориентированного подхода для подготовки специалистов, отвечающих требованиям IT-индустрии [4]. Д.С. Пономарев рассмотрел в своей работе разработку инструментов нормализации и стандартизации данных при помощи библиотек Python [5].

1.3 Цель исследования

Цель исследования — систематизировать практические подходы к обработке данных интернет-магазина с использованием библиотек Python и выявить ключевые факторы, влияющие на объём продаж.

2 Материалы и методы

Для анализа данных выбрана среда разработки Visual Studio, язык программирования Python.

Датасет скачан с официального сайта Kaggle.com <https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset>.

3 Результаты и дискуссии

Для работы необходимо установить Visual Studio, скачать его можно на официальном сайте <https://code.visualstudio.com/download>, в нем выбрать среду разработки Jupyter Notebook (рис. 1).

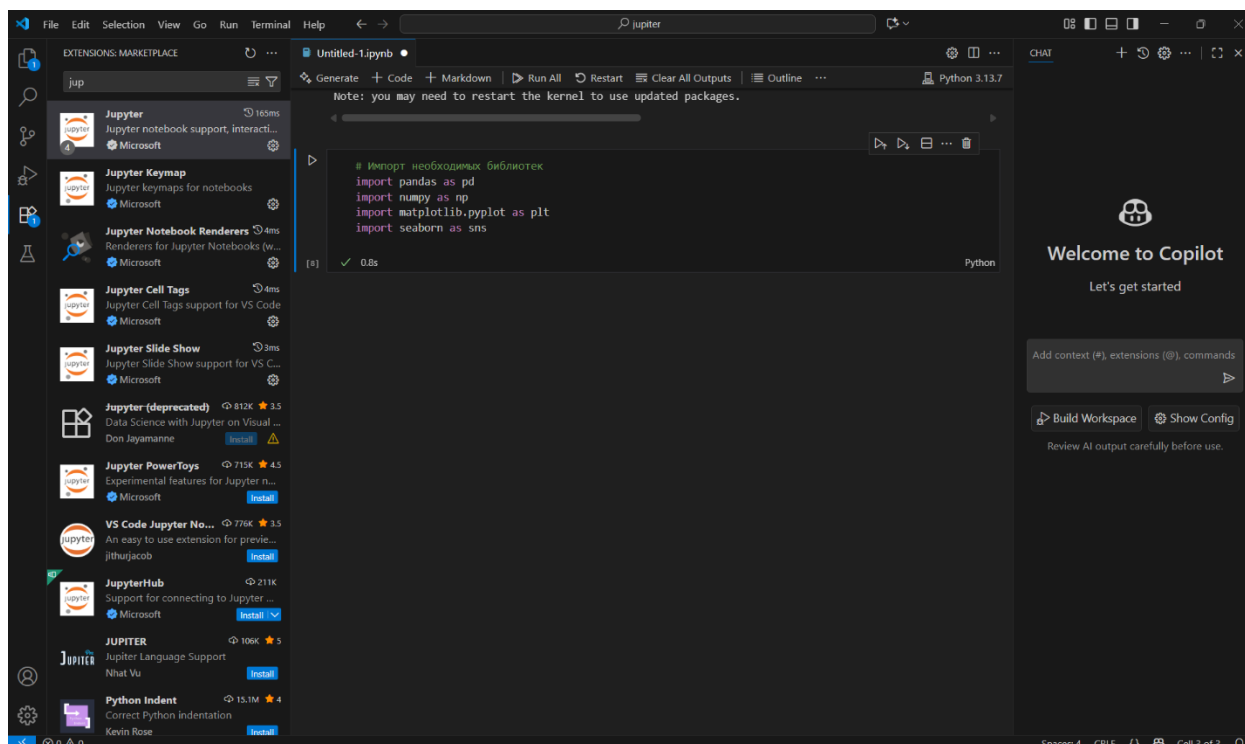


Рисунок 1- Установка библиотек

В качестве набора данных используется "Online Retail Dataset" из репозитория UCI, содержащий транзакции онлайн-магазина (рис. 2).

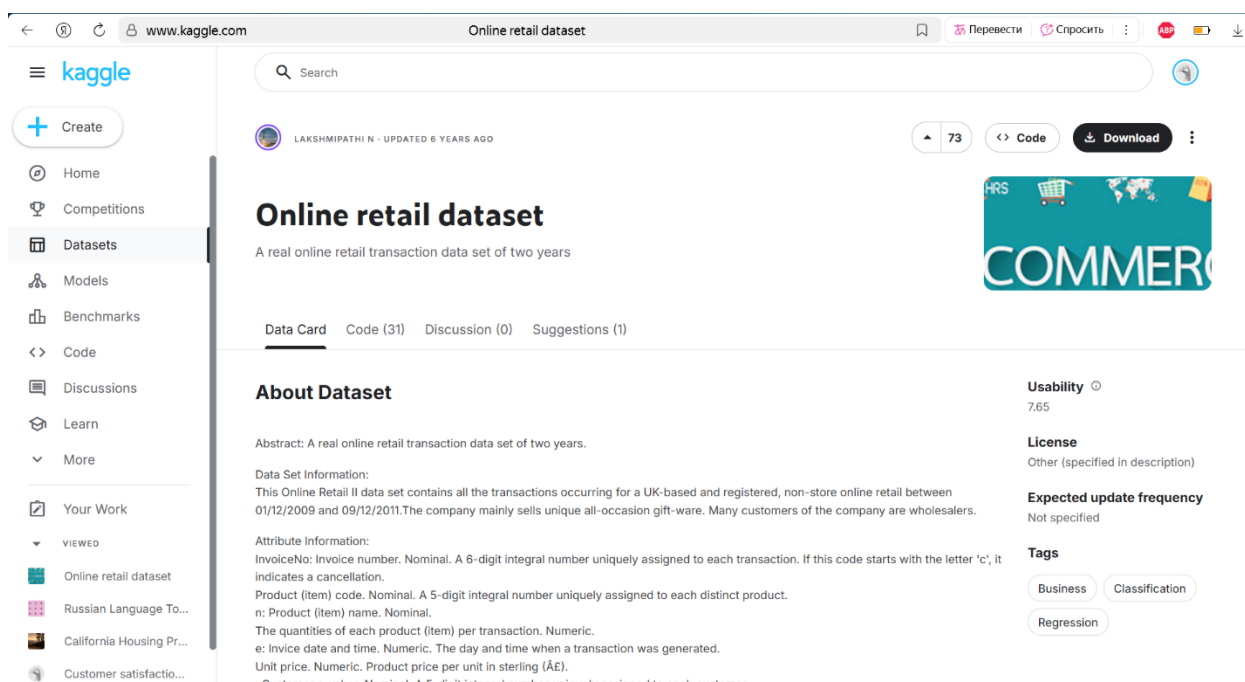


Рисунок 2- Официальная страница Kaggle.com

Для последующего анализа производится загрузка датасета в проект (рис.3).

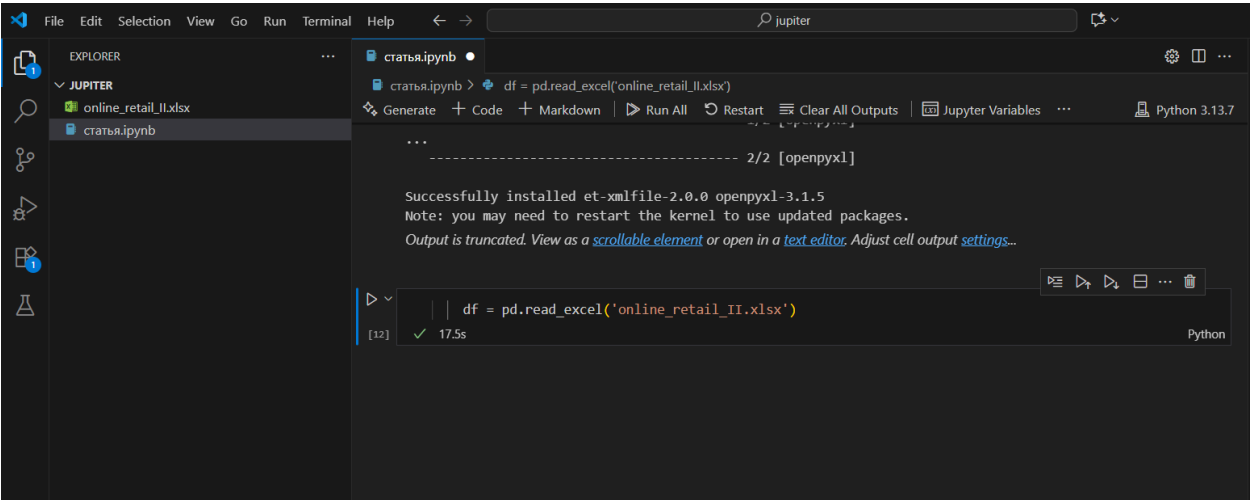


Рисунок 3- Загрузка датасета

Для первичного ознакомления со структурой и содержанием данных выведен фрагмент таблицы, содержащих первых пять строк (рис. 4).

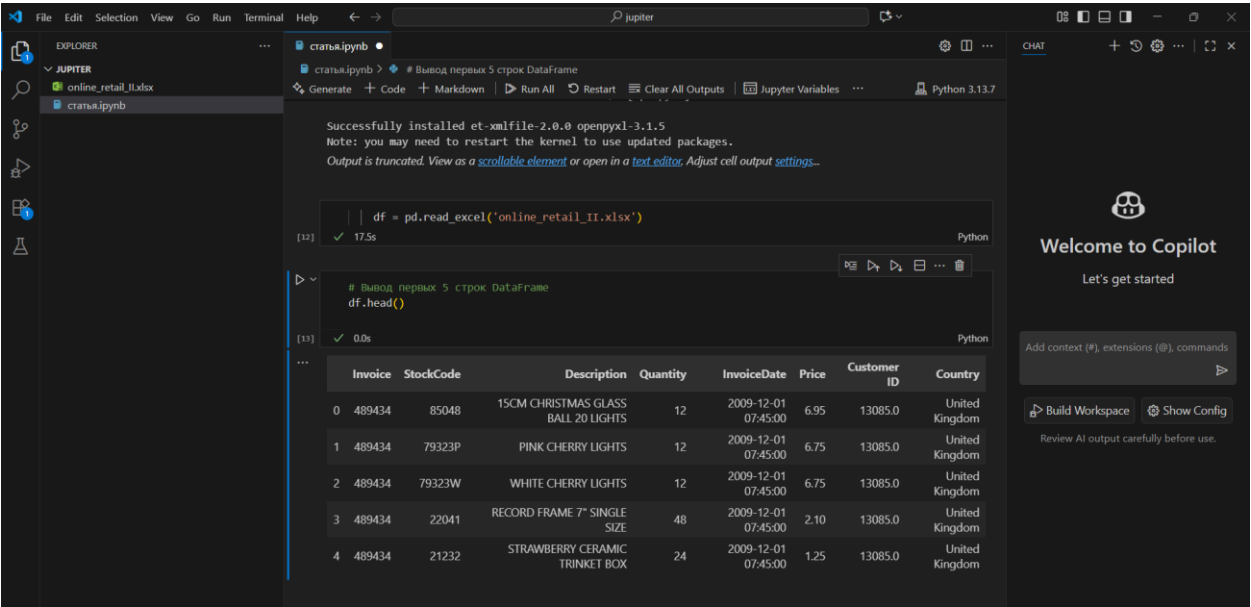


Рисунок 4- Обзор первых пяти строк

Для того чтобы выявить столбцы с пропусками, просматривается информация о названиях столбцов, их типах данных и количестве нулевых значений (рис. 5).

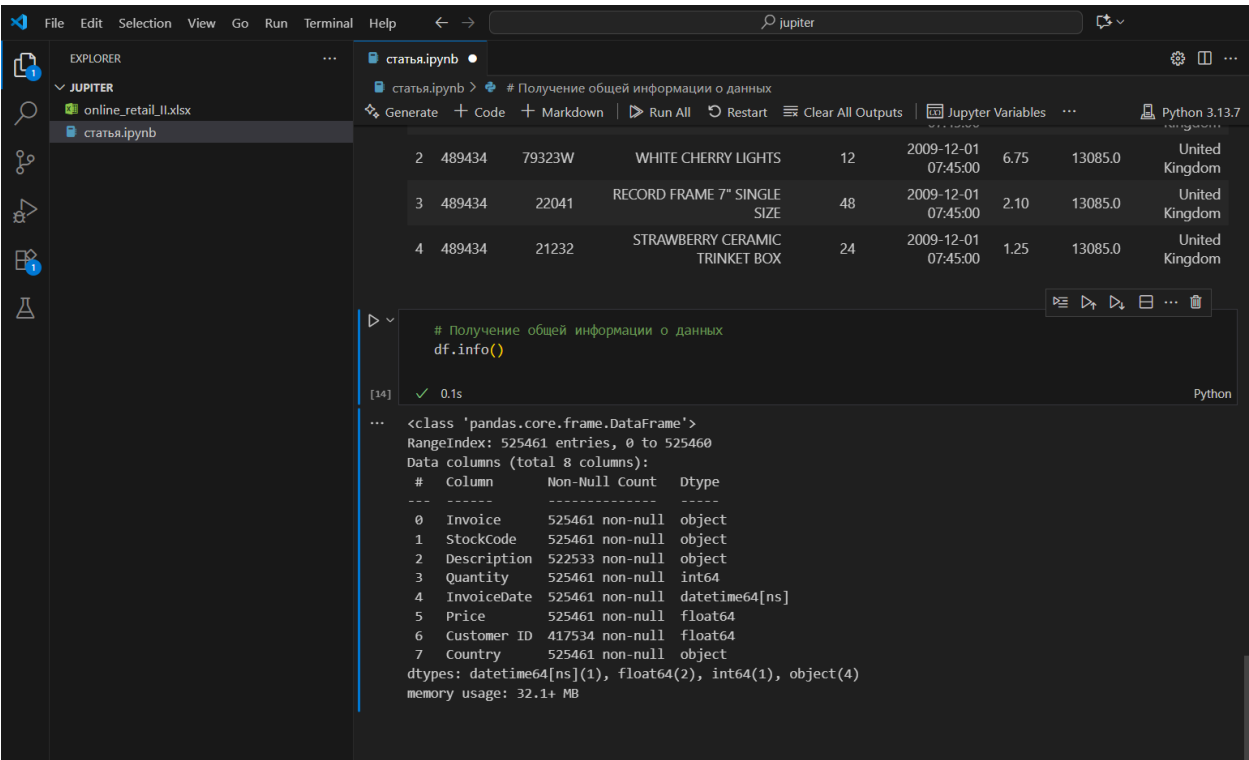


Рисунок 5- Получение общей информации о данных

Для расчета основных статистических показателей (среднее, медиана, стандартное отклонение и др.) для числовых столбцов, следует применить метод `describe` (Рис.6).

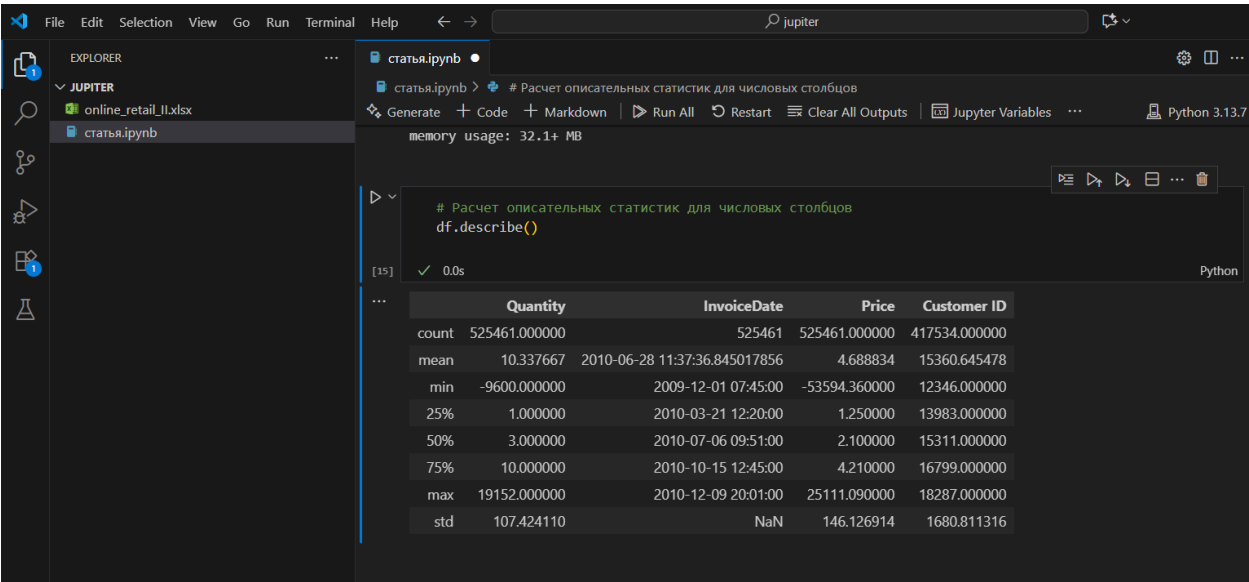


Рисунок 6 – Расчет описательных статистик для числовых столбцов

Следующим необходимым шагом является подсчет количества пропусков в каждом столбце (рис.7).

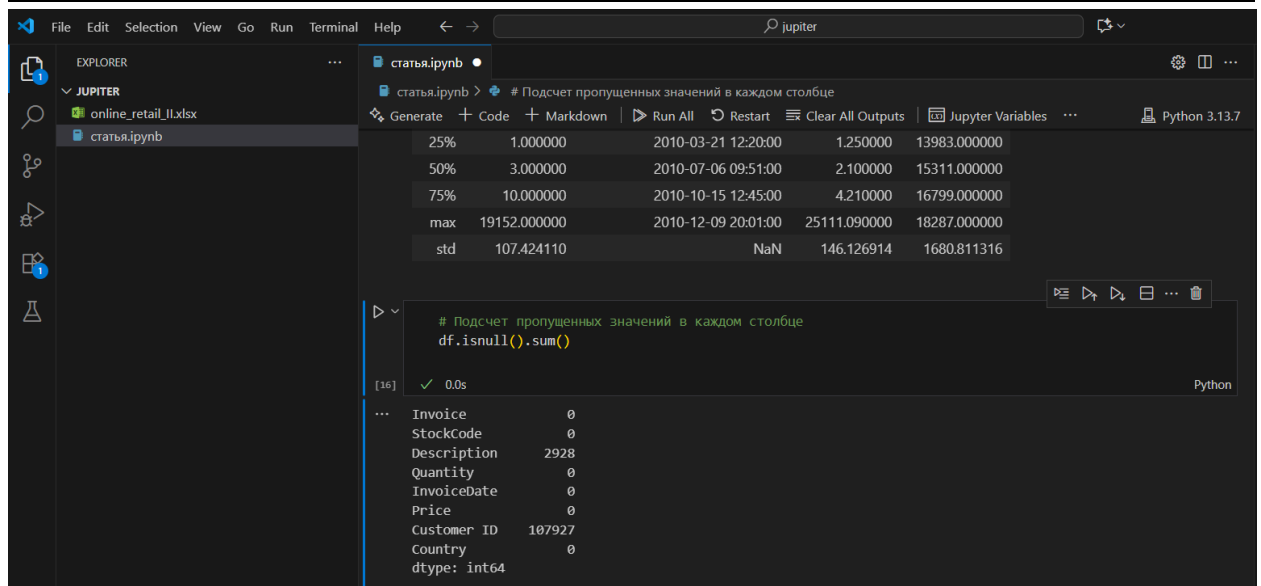


Рисунок 7 – Подсчет пропущенных значений

Для корректности типа данных необходимо проверить Customer ID является строкой (или целым числом, но без дробной части), а InvoiceDate — датой (рис. 8).

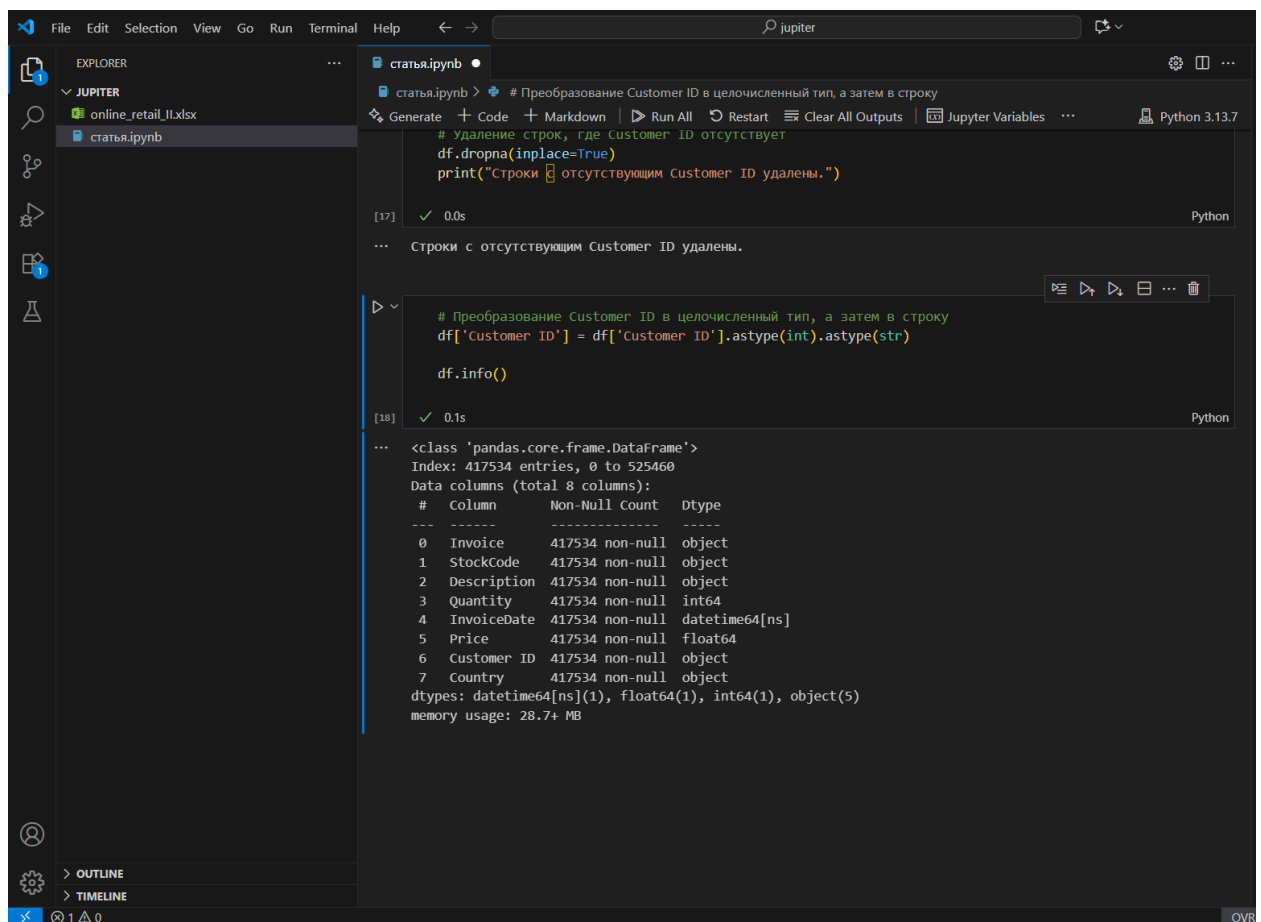


Рисунок 8- Преобразование Customer ID в целочисленный тип, затем в строку

Для хранения общей суммы: умножения цены на количество, создать столбец, переименовав его в TotalPrice (рис. 10).

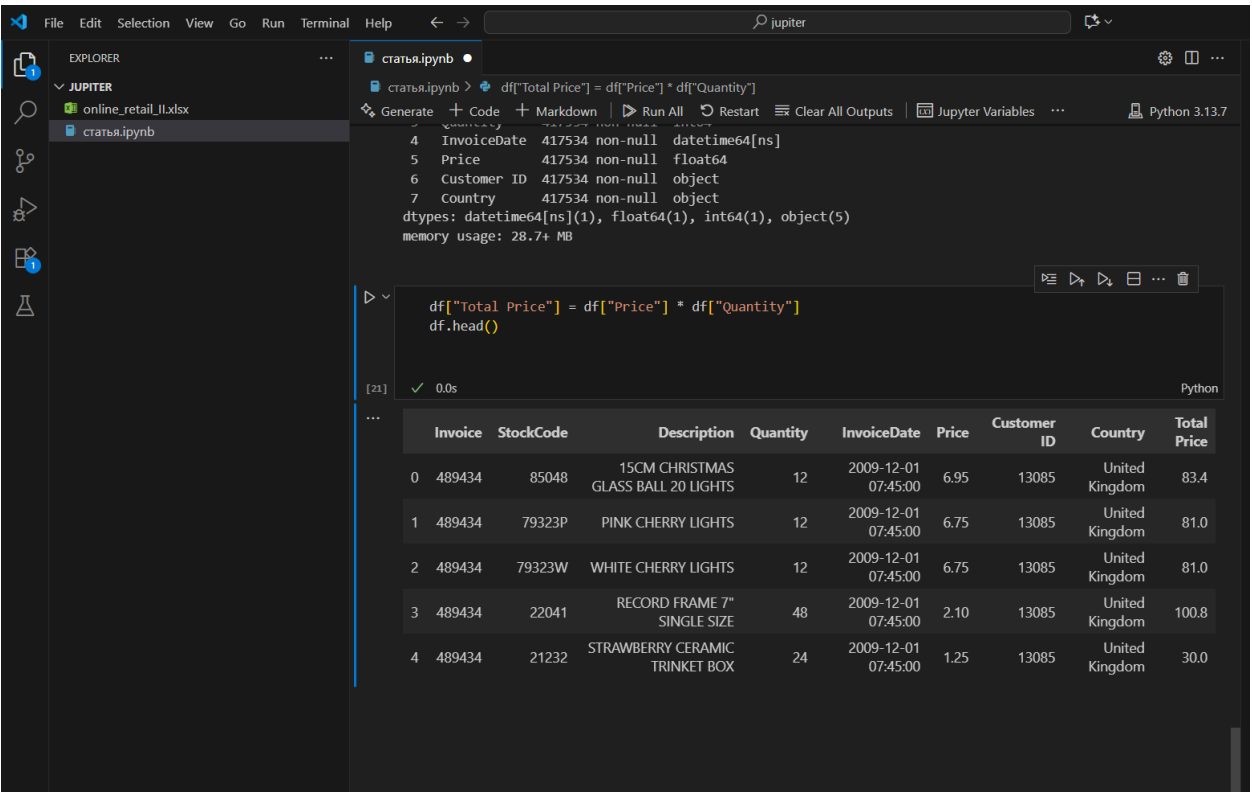


Рисунок 10- Создание новых признаков

Колонка Total Price суммирует значение для каждой группы, так же сортирует по возрастанию/убыванию и выводит топ-10 стран по выручке (рис. 11).

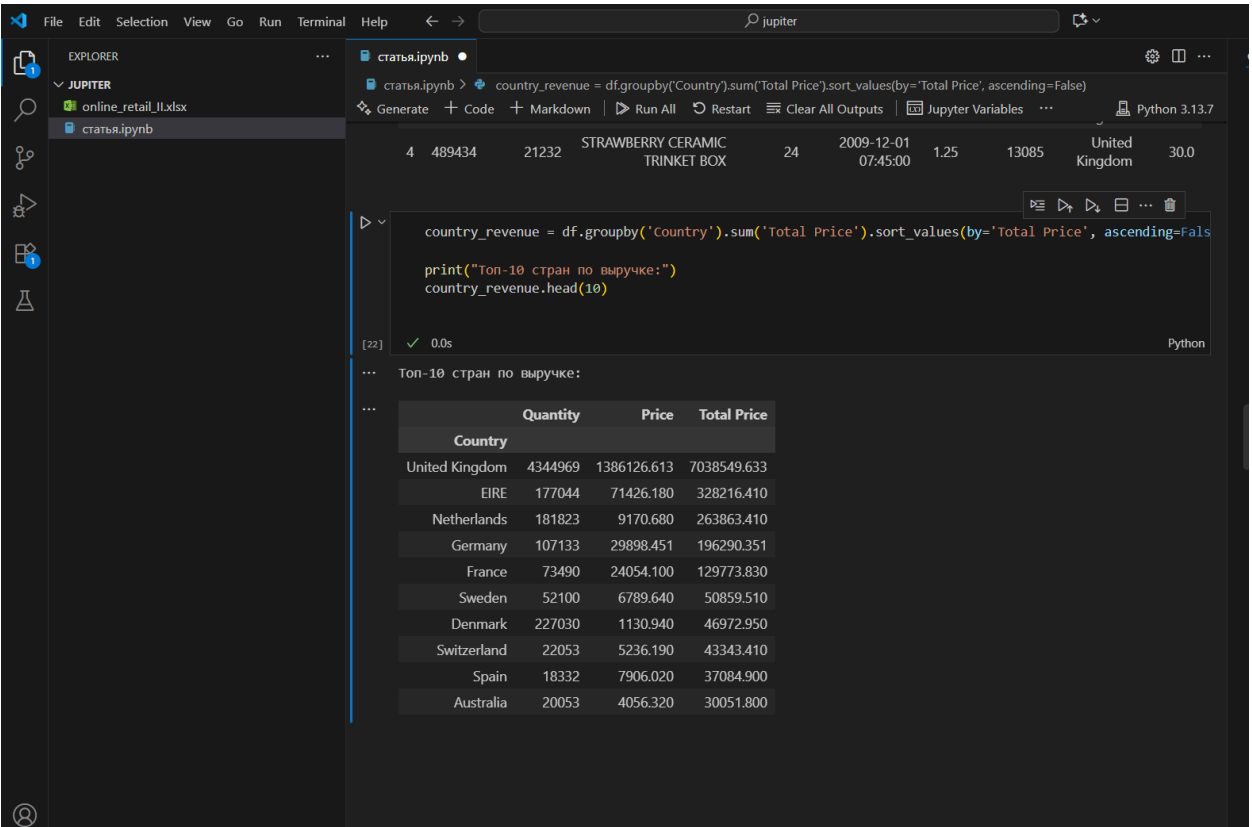


Рисунок 11- Группировка данных по странам и расчет суммарной выручки

Далее следует визуализировать топ-10 стран по выручке с помощью столбчатой гистограммы (рис. 12).

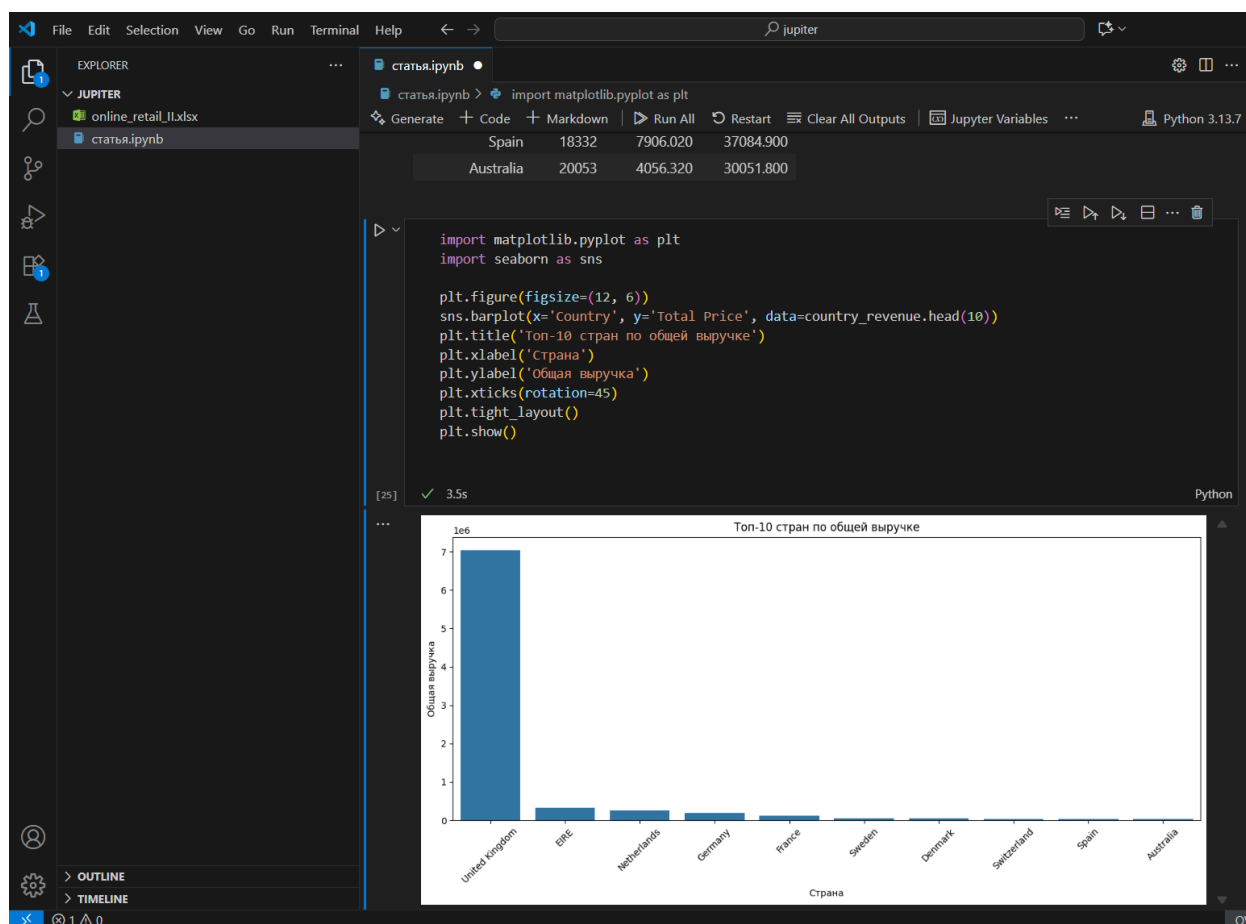


Рисунок 12- Построение гистограммы

Далее следует сделать обзор динамики продаж во времени, с помощью которой можно отследить динамику выручки по месяцам, самый высокий пик приходится на декабрь месяц (рис.13).

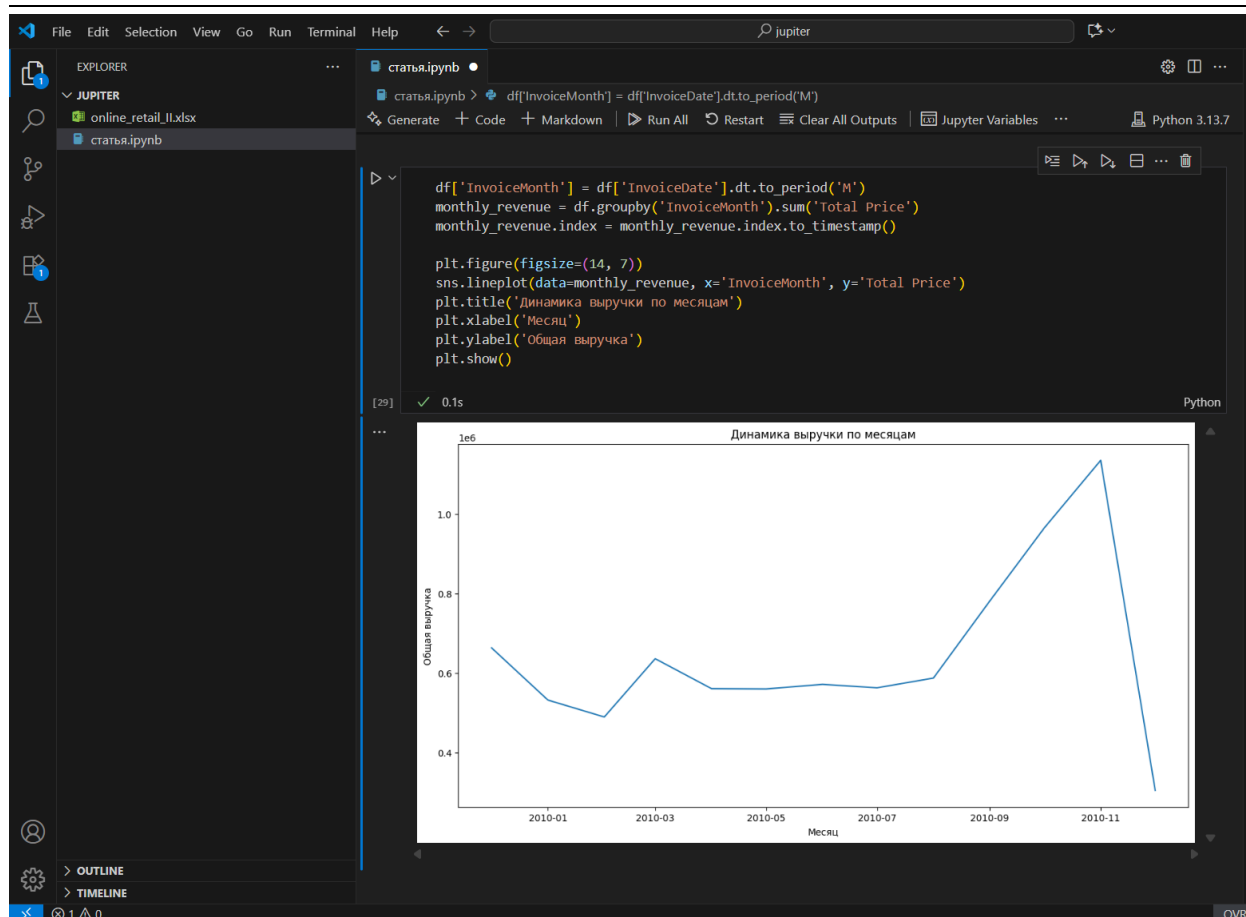


Рисунок 13- Обзор анализа продаж во времени

4 Выводы

В ходе исследования была продемонстрирована работа современных методов анализа данных для вывода информации о транзакциях онлайн-ритейла. Комплексная обработка массивов данных, включающая этапы предобработки, очистки, преобразования и визуализации, позволила выявить ключевые закономерности потребительского поведения и рыночные тенденции. Перспективы дальнейших исследований лежат в применении методов машинного обучения для прогнозирования продаж, сегментации клиентской базы и выявления скрытых паттернов поведения, что позволит перейти от ретроспективной аналитики к предиктивной.

Библиографический список

1. Фарайдунов О. К., Амонбекова З. А. Анализ данных в среде языка программирования PYTHON для решения экономических задач // Финансово-экономический вестник. 2024. №4.1. С. 91-98. URL: <https://cyberleninka.ru/article/n/analiz-dannyh-v-srede-yazyka-programmirovaniya-python-dlya-resheniya-ekonomicheskikh-zadach>
2. Моисеева Н. А., Полякова Т. А. Регрессионный анализ данных в технических исследованиях // Математическое и компьютерное моделирование: сборник. 2024. С. 107. URL:

- https://fctk.omsu.ru/nauka/Conf/2024/MCM_XI.pdf#page=107
3. Ремнев А. С., Муртазина А. Р. Использование языка Python для анализа и визуализации данных // Инновационное развитие техники и технологий в промышленности: сборник материалов Всероссийской научной конференции молодых исследователей с международным участием. Часть 5. М.: ФГБОУ ВО «РГУ им. А.Н. Косыгина», 2025. С. 125. URL: <https://rguk.ru/upload/medialibrary/d50/1d37nnw13nusen4je2iyrrlpqlbkgcm7d/ИНТЕКС-2025%20Часть%205.pdf#page=125>
 4. Капбар Г., Ким Е. Использование современных инструментов и технологий в обучении Python // ММІТ Proceedings. 2024. С. 62-66. URL: <https://mmit.tiue.uz/index.php/journal/article/view/137>
 5. Пономарев Д. С. Разработка web-приложения для предобработки данных с использованием библиотек Python // Инженерный вестник Дона. 2024. №8. С. 11. URL: <https://cyberleninka.ru/article/n/razrabotka-web-prilozheniya-dlya-predobrabotki-dannyh-s-ispolzovaniem-bibliotek-python>