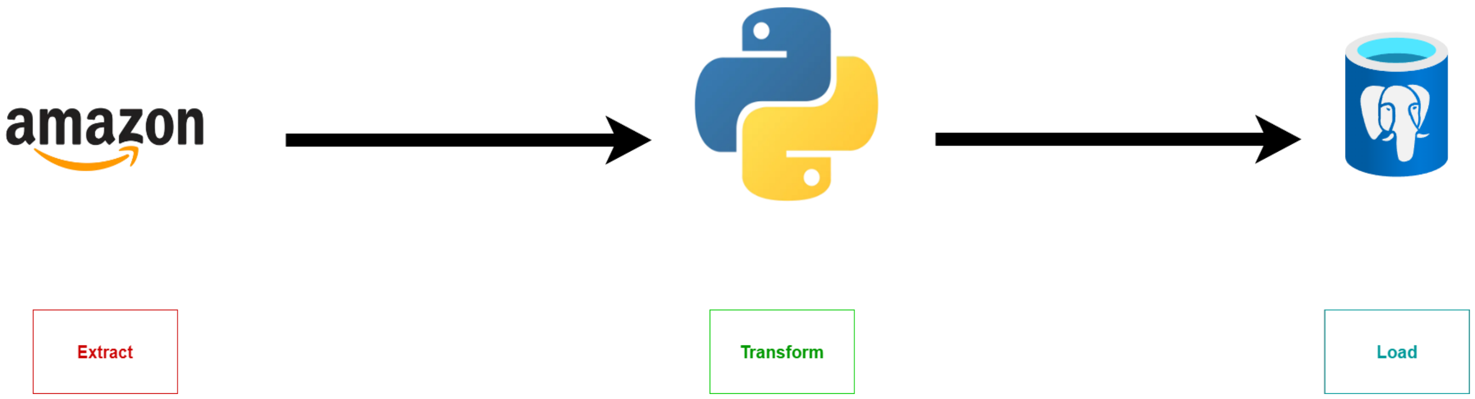
Amazon Books Data Pipeline



Extract

Transform

Load

Project Objective :

In this project, I created a system to collect book data from Amazon using Web Scraping and ETL (Extract, Transform, Load) techniques, with the aim of extracting data about books in the category of "Data Engineering Books" and storing them in a PostgreSQL database. This system aims to provide a tool to collect and analyze book information such as title, author, price, and rating, automatically and periodically.

Technologies and Tools Used

1-Airflow:

- Apache Airflow was used to organize and execute the workflow (DAG) which includes three main tasks: extracting data from Amazon, cleaning data, and loading data into the database.
- Airflow helped in scheduling the tasks and ensuring that they are executed periodically and provided a powerful mechanism for monitoring tasks and handling errors.

2-Web Scraping:

- BeautifulSoup and Requests were used to collect data from Amazon search results pages.
- HTML parsing techniques were used to extract accurate information such as title, author, price, and rating from web pages.

3-PostgreSQL:

- PostgreSQL was used as a database to store the extracted data. A database containing a table was designed to store the book data.
- PostgresOperator was used in Airflow to interact with the database and enter the collected data.

General Project Structure:

The general project structure is to implement a DAG in Airflow which consists of three main tasks:

Extract:

- Book data is collected from Amazon using Web Scraping.
- The number of books to be collected is determined (e.g. 300 books), and the data collection process begins across multiple pages on the Amazon website.

Transform:

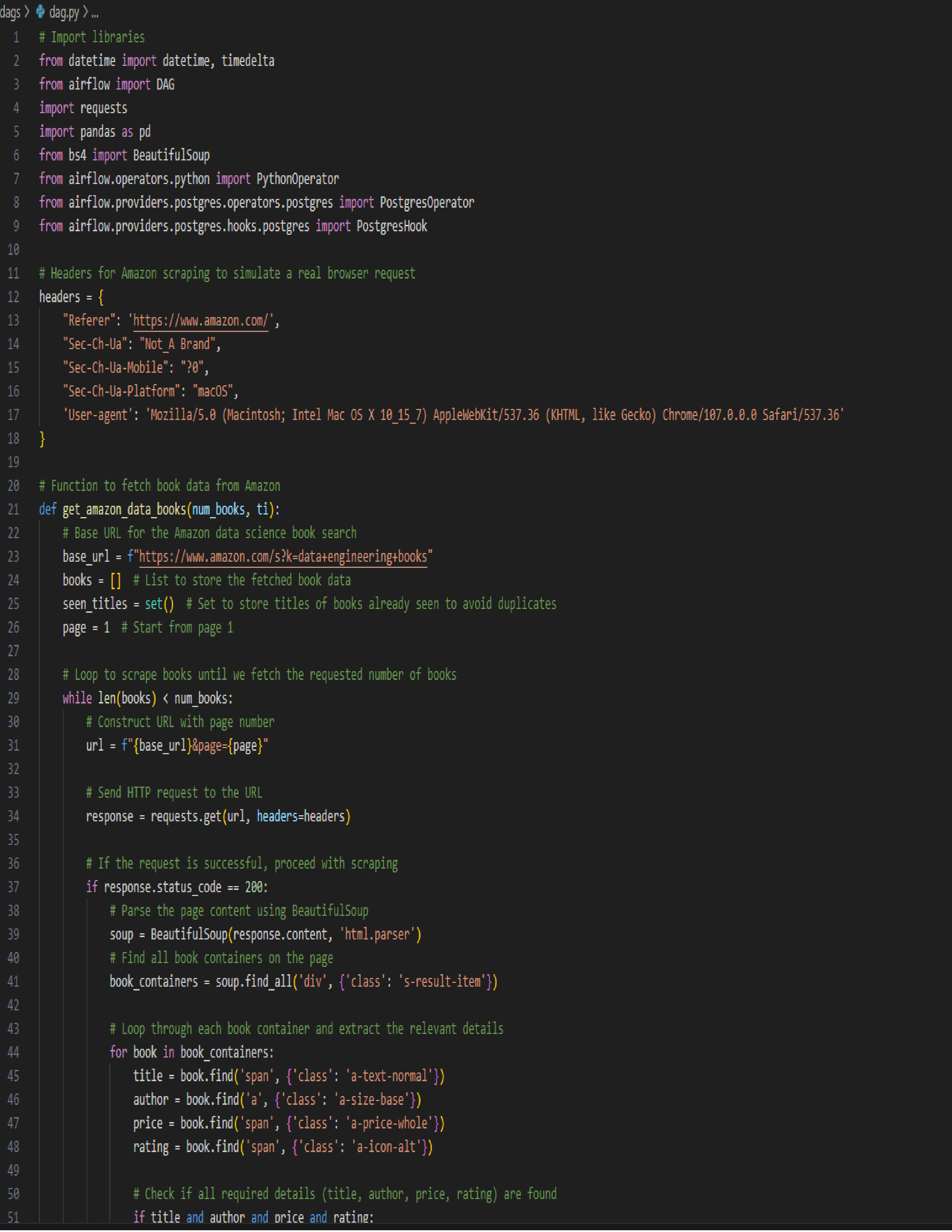
- After the data is extracted, it is cleaned using Pandas. At this stage, we:
- Remove duplicates.
- Ensure that the data is consistent with the required format.

Load:

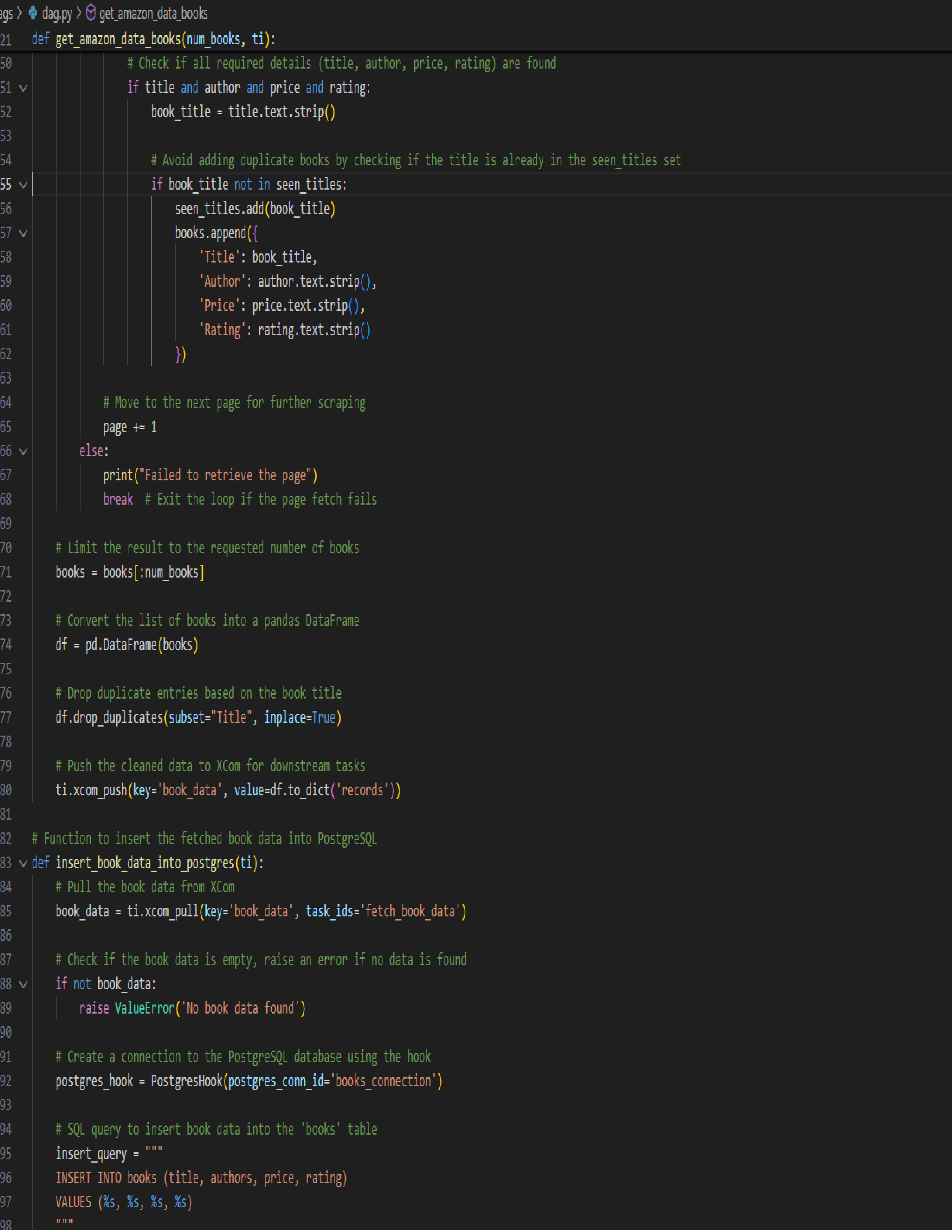
- After the data is cleaned, it is entered into the PostgreSQL database using the PostgresOperator in Airflow.

let's see code

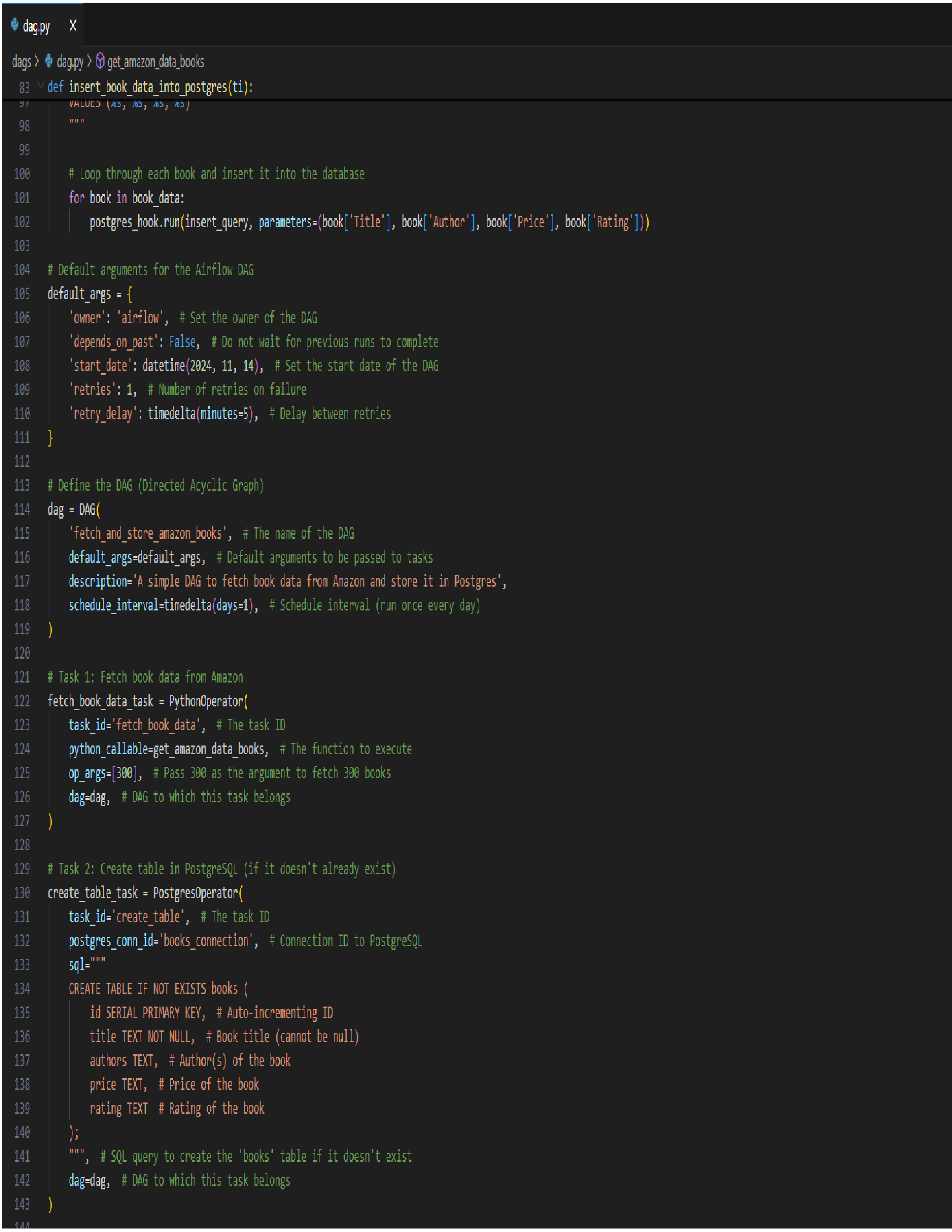
1-Airflow Setup for Scraping Amazon Data



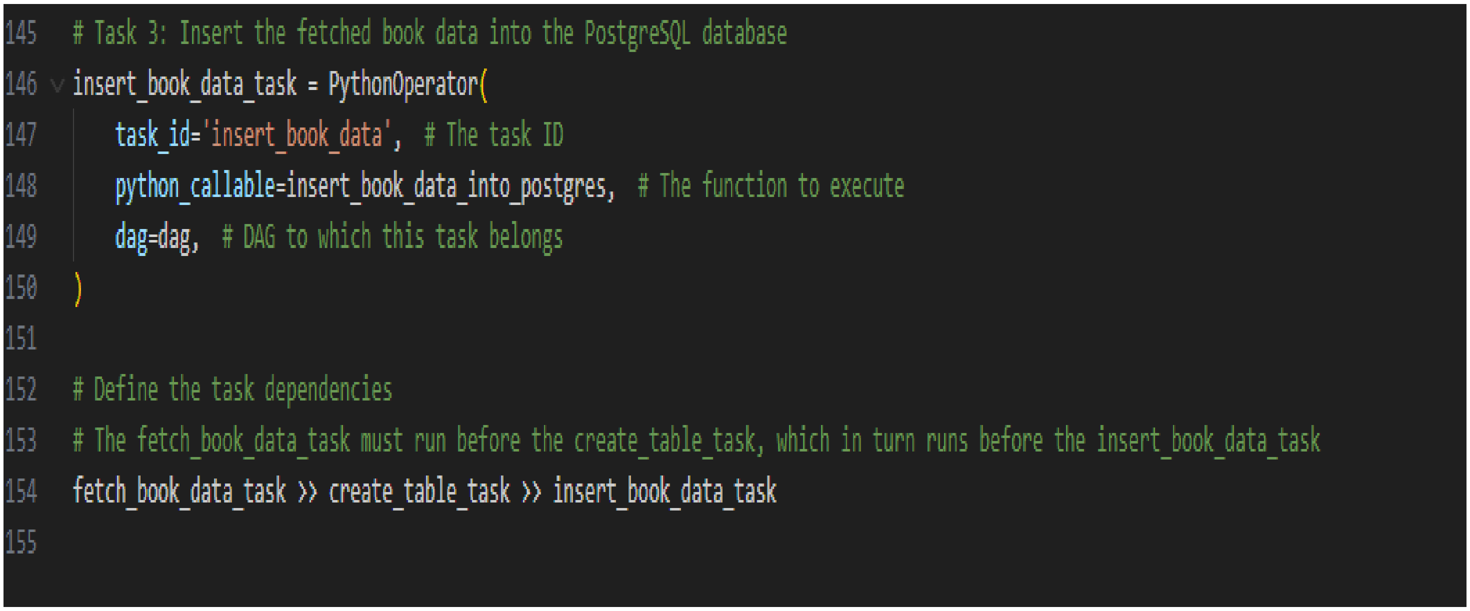
```python
# Import libraries
from datetime import datetime, timedelta
from airflow import DAG
import requests
import pandas as pd
from bs4 import BeautifulSoup
from airflow.operators.python import PythonOperator
from airflow.providers.postgres.operators.postgres import PostgresOperator
from airflow.providers.postgres.hooks.postgres import PostgresHook

# Headers for Amazon scraping to simulate a real browser request
headers = {
    "Referer": 'https://www.amazon.com/',
    "Sec-Ch-Ua": "Not_A Brand",
    "Sec-Ch-Ua-Mobile": "?0",
    "Sec-Ch-Ua-Platform": "macOS",
    'User-agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36'
}

# Function to fetch book data from Amazon
def get_amazon_data_books(num_books, ti):
    # Base URL for the Amazon data science book search
    base_url = f"https://www.amazon.com/s?k=data+engineering+books"
    books = []  # List to store the fetched book data
    seen_titles = set()  # Set to store titles of books already seen to avoid duplicates
    page = 1  # Start from page 1

    # Loop to scrape books until we fetch the requested number of books
    while len(books) < num_books:
        # Construct URL with page number
        url = f"{base_url}&page={page}"

        # Send HTTP request to the URL
        response = requests.get(url, headers=headers)

        # If the request is successful, proceed with scraping
        if response.status_code == 200:
            # Parse the page content using BeautifulSoup
            soup = BeautifulSoup(response.content, 'html.parser')
            # Find all book containers on the page
            book_containers = soup.find_all('div', {'class': 's-result-item'})

            # Loop through each book container and extract the relevant details
            for book in book_containers:
                title = book.find('span', {'class': 'a-text-normal'})
                author = book.find('a', {'class': 'a-size-base'})
                price = book.find('span', {'class': 'a-price-whole'})
                rating = book.find('span', {'class': 'a-icon-alt'})

                # Check if all required details (title, author, price, rating) are found
                if title and author and price and rating:
```



```python
def get_amazon_data_books(num_books, ti):
                # Check if all required details (title, author, price, rating) are found
                if title and author and price and rating:
                    book_title = title.text.strip()

                    # Avoid adding duplicate books by checking if the title is already in the seen_titles set
                    if book_title not in seen_titles:
                        seen_titles.add(book_title)
                        books.append({
                            'Title': book_title,
                            'Author': author.text.strip(),
                            'Price': price.text.strip(),
                            'Rating': rating.text.strip()
                        })

            # Move to the next page for further scraping
            page += 1
        else:
            print("Failed to retrieve the page")
            break  # Exit the loop if the page fetch fails

    # Limit the result to the requested number of books
    books = books[:num_books]

    # Convert the list of books into a pandas DataFrame
    df = pd.DataFrame(books)

    # Drop duplicate entries based on the book title
    df.drop_duplicates(subset="Title", inplace=True)

    # Push the cleaned data to XCom for downstream tasks
    ti.xcom_push(key='book_data', value=df.to_dict('records'))

# Function to insert the fetched book data into PostgreSQL
def insert_book_data_into_postgres(ti):
    # Pull the book data from XCom
    book_data = ti.xcom_pull(key='book_data', task_ids='fetch_book_data')

    # Check if the book data is empty, raise an error if no data is found
    if not book_data:
        raise ValueError('No book data found')

    # Create a connection to the PostgreSQL database using the hook
    postgres_hook = PostgresHook(postgres_conn_id='books_connection')

    # SQL query to insert book data into the 'books' table
    insert_query = """
    INSERT INTO books (title, authors, price, rating)
    VALUES (%s, %s, %s, %s)
    """
```



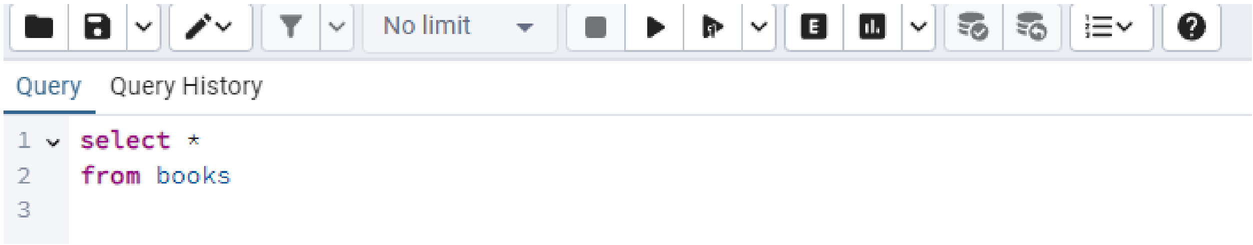
```python
def insert_book_data_into_postgres(ti):
    VALUES (%s, %s, %s, %s)
    """


    # Loop through each book and insert it into the database
    for book in book_data:
        postgres_hook.run(insert_query, parameters=(book['Title'], book['Author'], book['Price'], book['Rating']))

# Default arguments for the Airflow DAG
default_args = {
    'owner': 'airflow',  # Set the owner of the DAG
    'depends_on_past': False,  # Do not wait for previous runs to complete
    'start_date': datetime(2024, 11, 14),  # Set the start date of the DAG
    'retries': 1,  # Number of retries on failure
    'retry_delay': timedelta(minutes=5),  # Delay between retries
}

# Define the DAG (Directed Acyclic Graph)
dag = DAG(
    'fetch_and_store_amazon_books',  # The name of the DAG
    default_args=default_args,  # Default arguments to be passed to tasks
    description='A simple DAG to fetch book data from Amazon and store it in Postgres',
    schedule_interval=timedelta(days=1),  # Schedule interval (run once every day)
)

# Task 1: Fetch book data from Amazon
fetch_book_data_task = PythonOperator(
    task_id='fetch_book_data',  # The task ID
    python_callable=get_amazon_data_books,  # The function to execute
    op_args=[300],  # Pass 300 as the argument to fetch 300 books
    dag=dag,  # DAG to which this task belongs
)

# Task 2: Create table in PostgreSQL (if it doesn't already exist)
create_table_task = PostgresOperator(
    task_id='create_table',  # The task ID
    postgres_conn_id='books_connection',  # Connection ID to PostgreSQL
    sql="""
    CREATE TABLE IF NOT EXISTS books (
        id SERIAL PRIMARY KEY,  # Auto-incrementing ID
        title TEXT NOT NULL,  # Book title (cannot be null)
        authors TEXT,  # Author(s) of the book
        price TEXT,  # Price of the book
        rating TEXT  # Rating of the book
    );
    """,  # SQL query to create the 'books' table if it doesn't exist
    dag=dag,  # DAG to which this task belongs
)
```



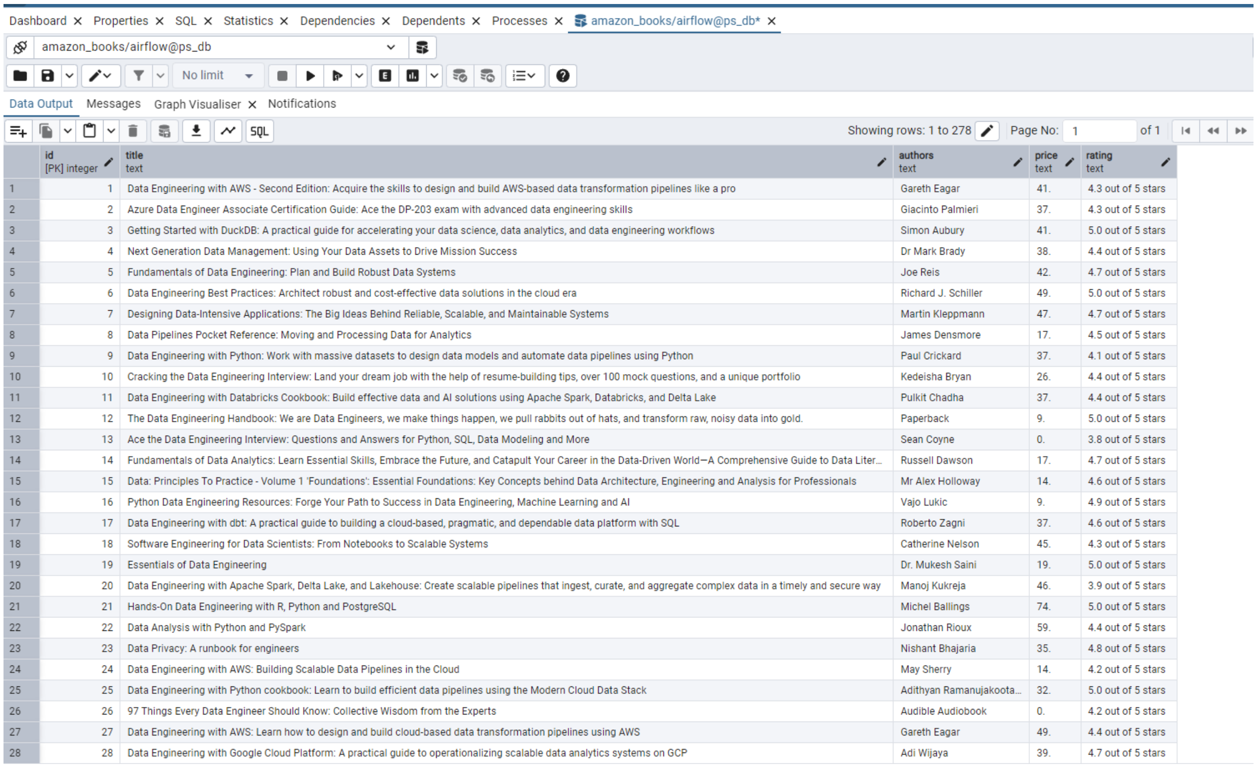
```python
# Task 3: Insert the fetched book data into the PostgreSQL database
insert_book_data_task = PythonOperator(
    task_id='insert_book_data',  # The task ID
    python_callable=insert_book_data_into_postgres,  # The function to execute
    dag=dag,  # DAG to which this task belongs
)

# Define the task dependencies
# The fetch_book_data_task must run before the create_table_task, which in turn runs before the insert_book_data_task
fetch_book_data_task >> create_table_task >> insert_book_data_task
```

2-PostgreSQL Data Querying for Book Analysis

1. View all books



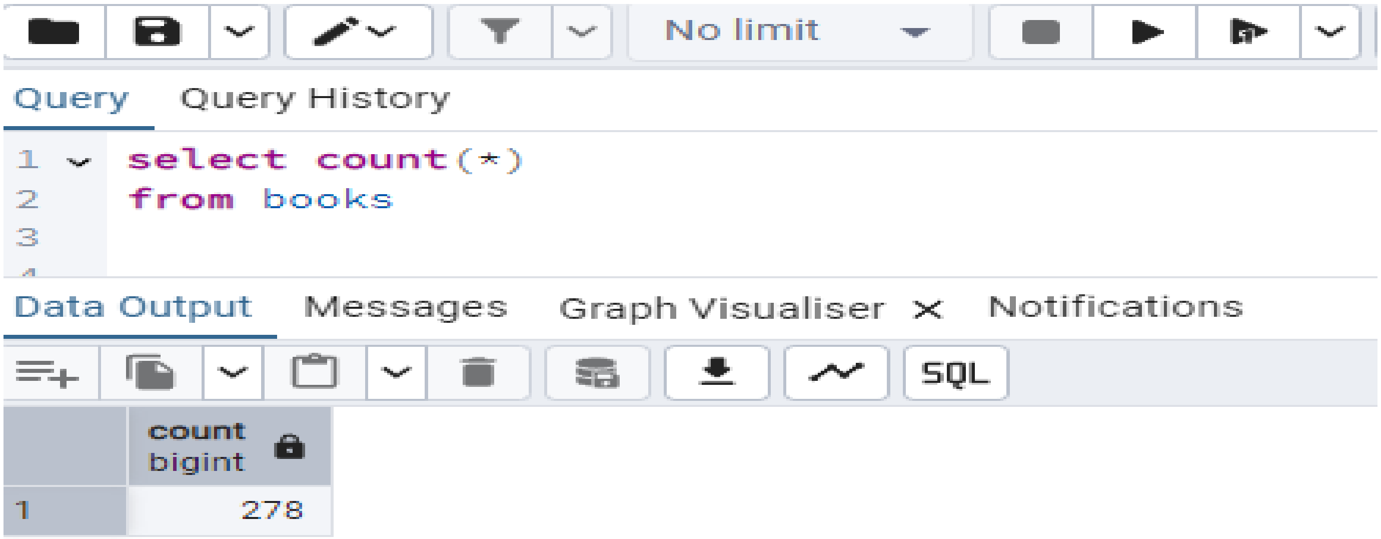
```
select *
from books
```


Dashboard ✕ Properties ✕ SQL ✕ Statistics ✕ Dependencies ✕ Dependents ✕ Processes ✕ amazon_books/airflow@ps_db* ✕
amazon_books/airflow@ps_db
No limit
Data Output
Messages
Graph Visualiser ✕
Notifications
SQL

Showing rows: 1 to 278 Page No: 1 of 1

	id [PK] integer	title text	authors text	price text	rating text
1	1	Data Engineering with AWS - Second Edition: Acquire the skills to design and build AWS-based data transformation pipelines like a pro	Gareth Eagar	41.	4.3 out of 5 stars
2	2	Azure Data Engineer Associate Certification Guide: Ace the DP-203 exam with advanced data engineering skills	Giacinto Palmieri	37.	4.3 out of 5 stars
3	3	Getting Started with DuckDB: A practical guide for accelerating your data science, data analytics, and data engineering workflows	Simon Aubury	41.	5.0 out of 5 stars
4	4	Next Generation Data Management: Using Your Data Assets to Drive Mission Success	Dr Mark Brady	38.	4.4 out of 5 stars
5	5	Fundamentals of Data Engineering: Plan and Build Robust Data Systems	Joe Reis	42.	4.7 out of 5 stars
6	6	Data Engineering Best Practices: Architect robust and cost-effective data solutions in the cloud era	Richard J. Schiller	49.	5.0 out of 5 stars
7	7	Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems	Martin Kleppmann	47.	4.7 out of 5 stars
8	8	Data Pipelines Pocket Reference: Moving and Processing Data for Analytics	James Densmore	17.	4.5 out of 5 stars
9	9	Data Engineering with Python: Work with massive datasets to design data models and automate data pipelines using Python	Paul Crickard	37.	4.1 out of 5 stars
10	10	Cracking the Data Engineering Interview: Land your dream job with the help of resume-building tips, over 100 mock questions, and a unique portfolio	Kedeisha Bryan	26.	4.4 out of 5 stars
11	11	Data Engineering with Databricks Cookbook: Build effective data and AI solutions using Apache Spark, Databricks, and Delta Lake	Pulkit Chadha	37.	4.4 out of 5 stars
12	12	The Data Engineering Handbook: We are Data Engineers, we make things happen, we pull rabbits out of hats, and transform raw, noisy data into gold.	Paperback	9.	5.0 out of 5 stars
13	13	Ace the Data Engineering Interview: Questions and Answers for Python, SQL, Data Modeling and More	Sean Coyne	0.	3.8 out of 5 stars
14	14	Fundamentals of Data Analytics: Learn Essential Skills, Embrace the Future, and Catapult Your Career in the Data-Driven World—A Comprehensive Guide to Data Liter...	Russell Dawson	17.	4.7 out of 5 stars
15	15	Data: Principles To Practice - Volume 1 'Foundations': Essential Foundations: Key Concepts behind Data Architecture, Engineering and Analysis for Professionals	Mr Alex Holloway	14.	4.6 out of 5 stars
16	16	Python Data Engineering Resources: Forge Your Path to Success in Data Engineering, Machine Learning and AI	Vajo Lukic	9.	4.9 out of 5 stars
17	17	Data Engineering with dbt: A practical guide to building a cloud-based, pragmatic, and dependable data platform with SQL	Roberto Zagni	37.	4.6 out of 5 stars
18	18	Software Engineering for Data Scientists: From Notebooks to Scalable Systems	Catherine Nelson	45.	4.3 out of 5 stars
19	19	Essentials of Data Engineering	Dr. Mukesh Saini	19.	5.0 out of 5 stars
20	20	Data Engineering with Apache Spark, Delta Lake, and Lakehouse: Create scalable pipelines that ingest, curate, and aggregate complex data in a timely and secure way	Manoj Kukreja	46.	3.9 out of 5 stars
21	21	Hands-On Data Engineering with R, Python and PostgreSQL	Michel Ballings	74.	5.0 out of 5 stars
22	22	Data Analysis with Python and PySpark	Jonathan Rioux	59.	4.4 out of 5 stars
23	23	Data Privacy: A runbook for engineers	Nishant Bhajaria	35.	4.8 out of 5 stars
24	24	Data Engineering with AWS: Building Scalable Data Pipelines in the Cloud	May Sherry	14.	4.2 out of 5 stars
25	25	Data Engineering with Python cookbook: Learn to build efficient data pipelines using the Modern Cloud Data Stack	Adithyan Ramanujakoota...	32.	5.0 out of 5 stars
26	26	97 Things Every Data Engineer Should Know: Collective Wisdom from the Experts	Audible Audiobook	0.	4.2 out of 5 stars
27	27	Data Engineering with AWS: Learn how to design and build cloud-based data transformation pipelines using AWS	Gareth Eagar	49.	4.4 out of 5 stars
28	28	Data Engineering with Google Cloud Platform: A practical guide to operationalizing scalable data analytics systems on GCP	Adi Wijaya	39.	4.7 out of 5 stars

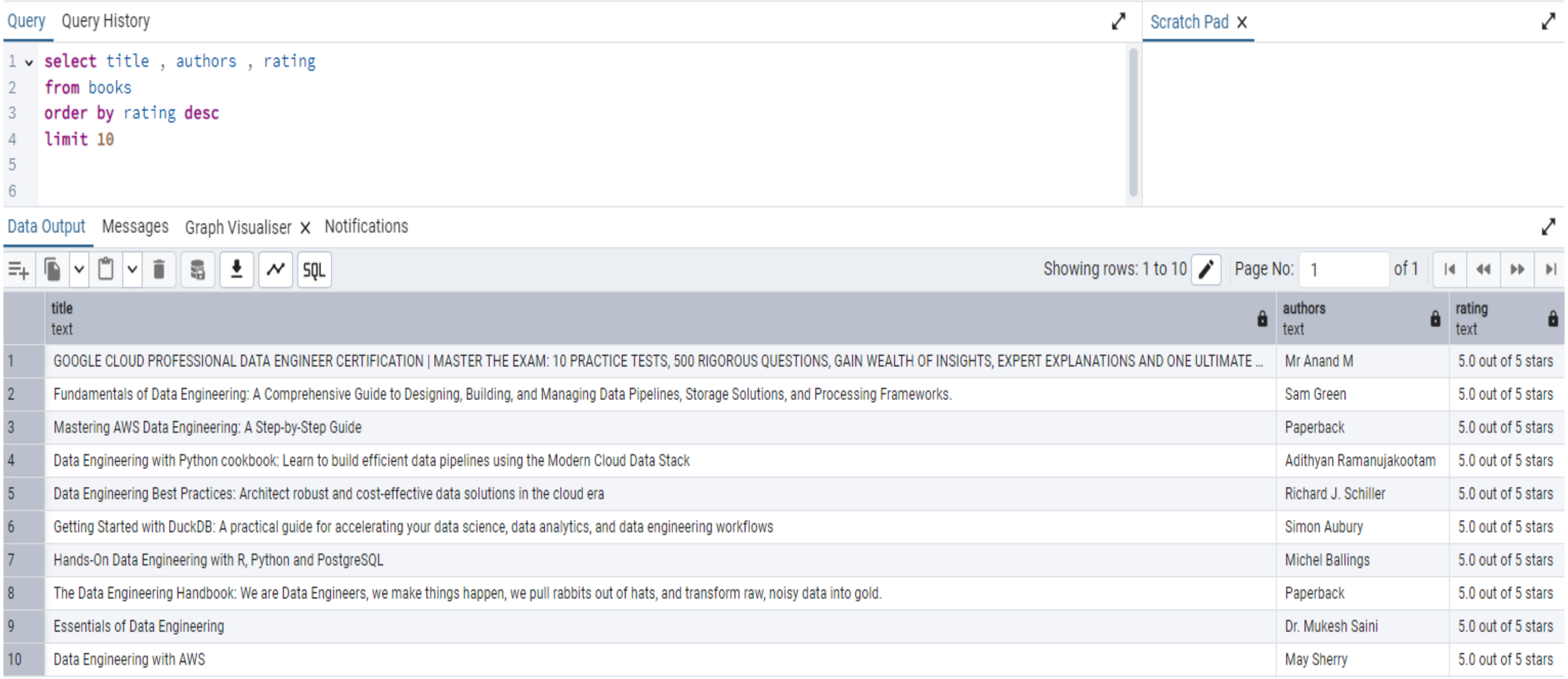
2. Total number of books



```sql
select count(*)
from books
```

	count bigint
1	278

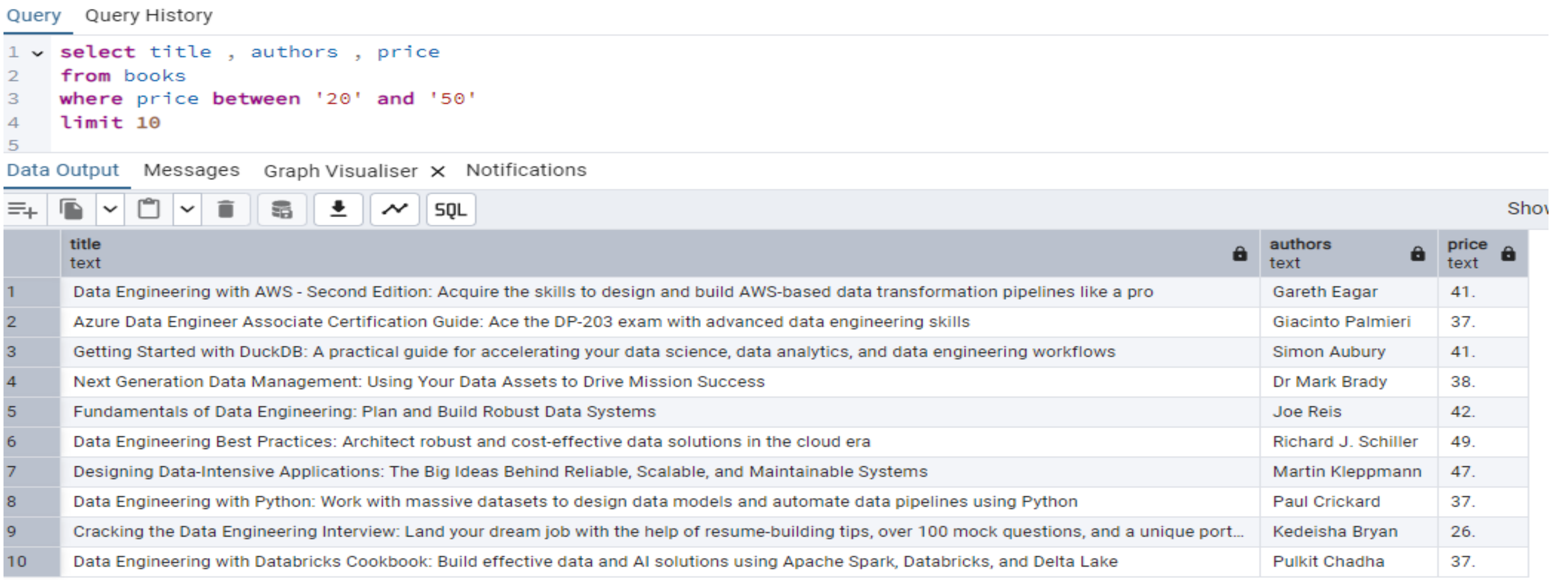
3. Most Popular Books (Highest Rated)



```sql
select title , authors , rating
from books
order by rating desc
limit 10
```

	title text	authors text	rating text
1	GOOGLE CLOUD PROFESSIONAL DATA ENGINEER CERTIFICATION | MASTER THE EXAM: 10 PRACTICE TESTS, 500 RIGOROUS QUESTIONS, GAIN WEALTH OF INSIGHTS, EXPERT EXPLANATIONS AND ONE ULTIMATE ...	Mr Anand M	5.0 out of 5 stars
2	Fundamentals of Data Engineering: A Comprehensive Guide to Designing, Building, and Managing Data Pipelines, Storage Solutions, and Processing Frameworks.	Sam Green	5.0 out of 5 stars
3	Mastering AWS Data Engineering: A Step-by-Step Guide	Paperback	5.0 out of 5 stars
4	Data Engineering with Python cookbook: Learn to build efficient data pipelines using the Modern Cloud Data Stack	Adithyan Ramanujakootam	5.0 out of 5 stars
5	Data Engineering Best Practices: Architect robust and cost-effective data solutions in the cloud era	Richard J. Schiller	5.0 out of 5 stars
6	Getting Started with DuckDB: A practical guide for accelerating your data science, data analytics, and data engineering workflows	Simon Aubury	5.0 out of 5 stars
7	Hands-On Data Engineering with R, Python and PostgreSQL	Michel Ballings	5.0 out of 5 stars
8	The Data Engineering Handbook: We are Data Engineers, we make things happen, we pull rabbits out of hats, and transform raw, noisy data into gold.	Paperback	5.0 out of 5 stars
9	Essentials of Data Engineering	Dr. Mukesh Saini	5.0 out of 5 stars
10	Data Engineering with AWS	May Sherry	5.0 out of 5 stars

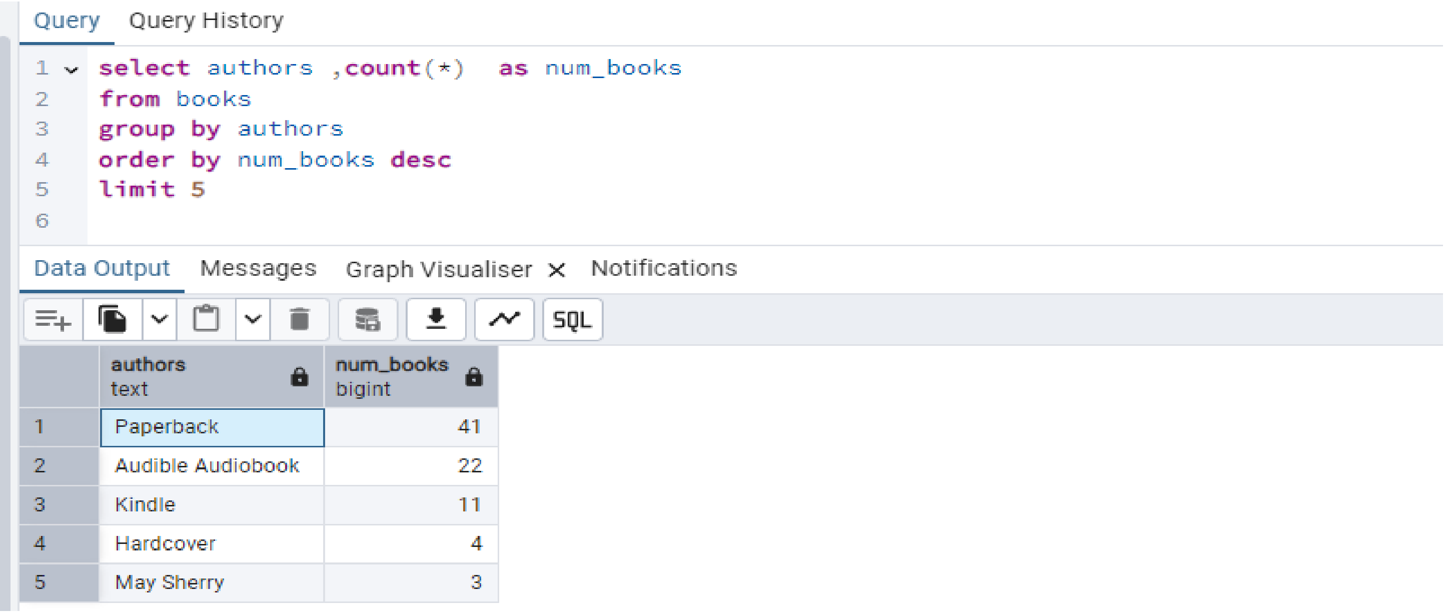
4. Books by price



```sql
select title , authors , price
from books
where price between '20' and '50'
limit 10
```

	title text	authors text	price text
1	Data Engineering with AWS - Second Edition: Acquire the skills to design and build AWS-based data transformation pipelines like a pro	Gareth Eagar	41.
2	Azure Data Engineer Associate Certification Guide: Ace the DP-203 exam with advanced data engineering skills	Giacinto Palmieri	37.
3	Getting Started with DuckDB: A practical guide for accelerating your data science, data analytics, and data engineering workflows	Simon Aubury	41.
4	Next Generation Data Management: Using Your Data Assets to Drive Mission Success	Dr Mark Brady	38.
5	Fundamentals of Data Engineering: Plan and Build Robust Data Systems	Joe Reis	42.
6	Data Engineering Best Practices: Architect robust and cost-effective data solutions in the cloud era	Richard J. Schiller	49.
7	Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems	Martin Kleppmann	47.
8	Data Engineering with Python: Work with massive datasets to design data models and automate data pipelines using Python	Paul Crickard	37.
9	Cracking the Data Engineering Interview: Land your dream job with the help of resume-building tips, over 100 mock questions, and a unique port...	Kedeisha Bryan	26.
10	Data Engineering with Databricks Cookbook: Build effective data and AI solutions using Apache Spark, Databricks, and Delta Lake	Pulkit Chadha	37.

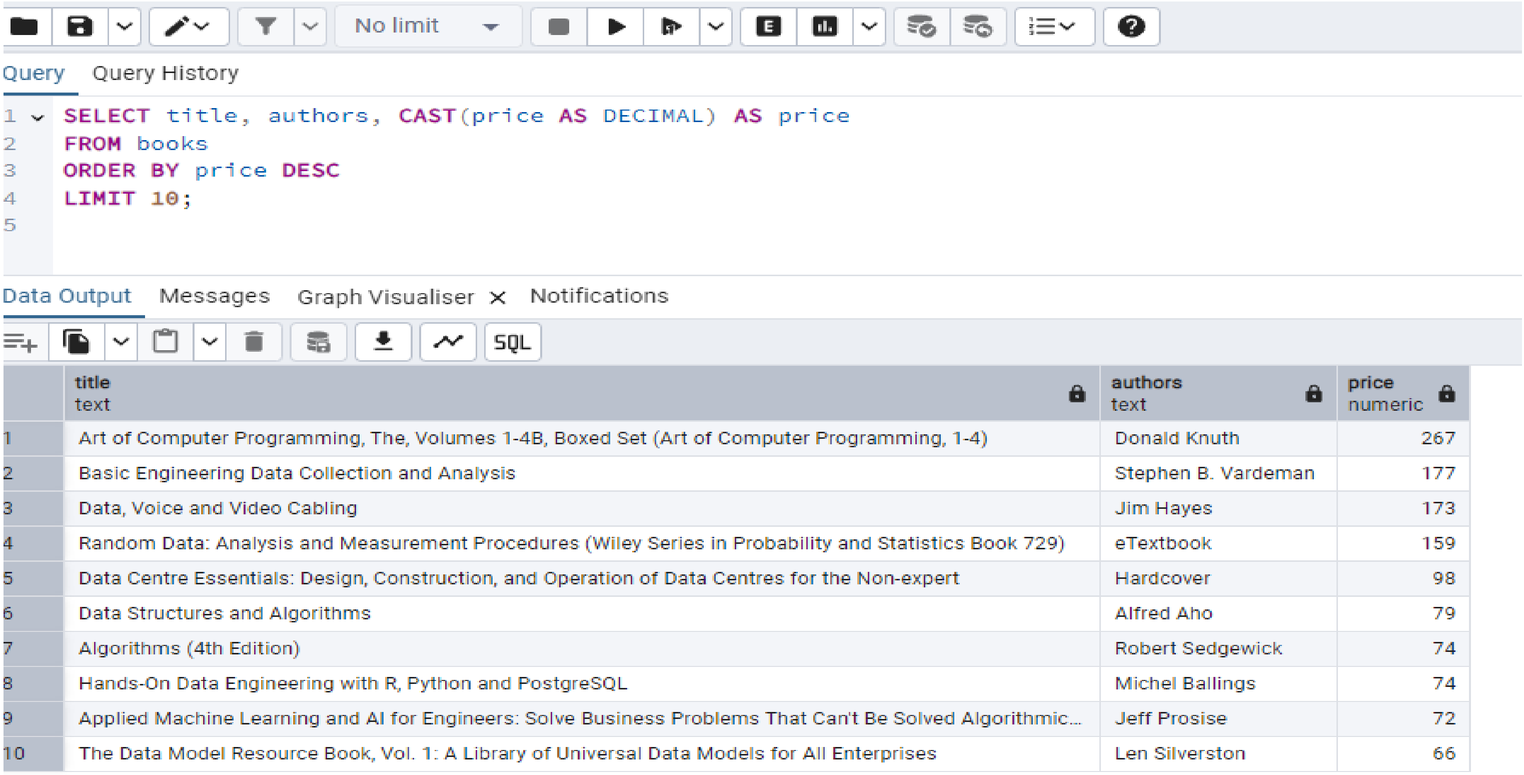
5. Top 5 Authors by Number of Books



```sql
select authors ,count(*)  as num_books
from books
group by authors
order by num_books desc
limit 5
```

	authors text	num_books bigint
1	Paperback	41
2	Audible Audiobook	22
3	Kindle	11
4	Hardcover	4
5	May Sherry	3

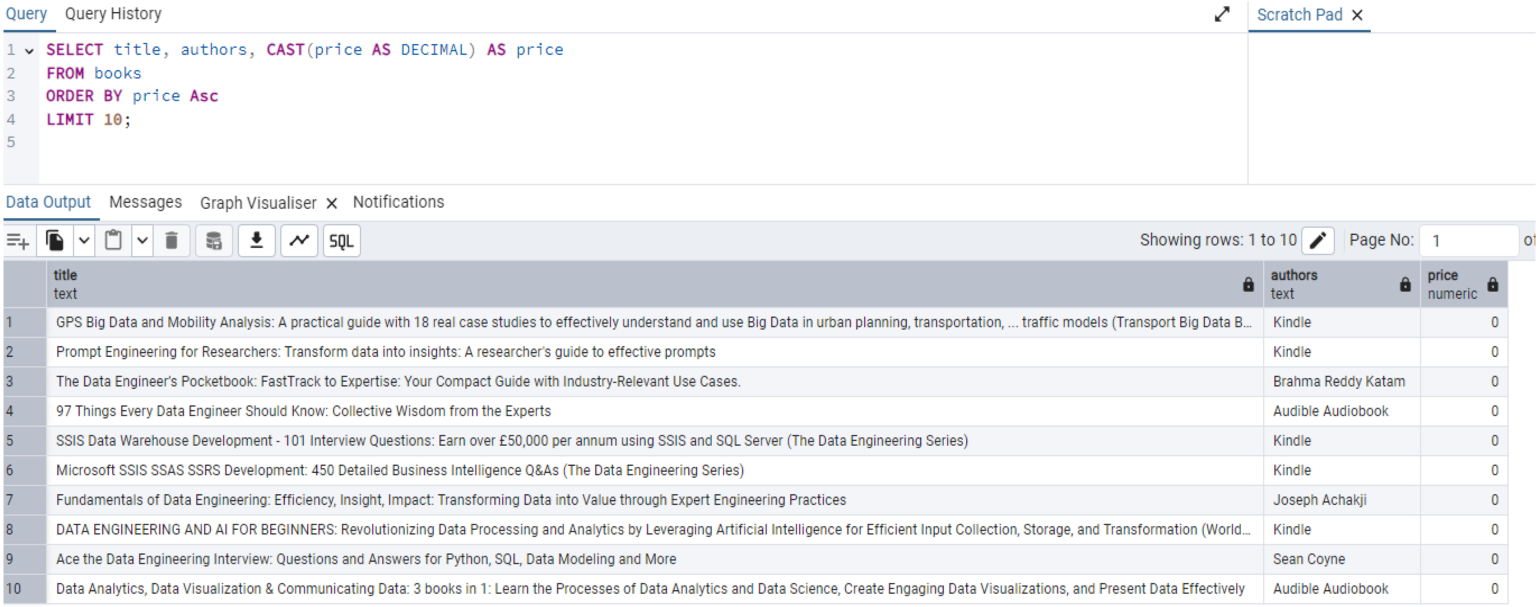
6. The most expensive books



```sql
SELECT title, authors, CAST(price AS DECIMAL) AS price
FROM books
ORDER BY price DESC
LIMIT 10;
```

	title text	authors text	price numeric
1	Art of Computer Programming, The, Volumes 1-4B, Boxed Set (Art of Computer Programming, 1-4)	Donald Knuth	267
2	Basic Engineering Data Collection and Analysis	Stephen B. Vardeman	177
3	Data, Voice and Video Cabling	Jim Hayes	173
4	Random Data: Analysis and Measurement Procedures (Wiley Series in Probability and Statistics Book 729)	eTextbook	159
5	Data Centre Essentials: Design, Construction, and Operation of Data Centres for the Non-expert	Hardcover	98
6	Data Structures and Algorithms	Alfred Aho	79
7	Algorithms (4th Edition)	Robert Sedgewick	74
8	Hands-On Data Engineering with R, Python and PostgreSQL	Michel Ballings	74
9	Applied Machine Learning and AI for Engineers: Solve Business Problems That Can't Be Solved Algorithmic...	Jeff Prosise	72
10	The Data Model Resource Book, Vol. 1: A Library of Universal Data Models for All Enterprises	Len Silverston	66

7. Less expensive books

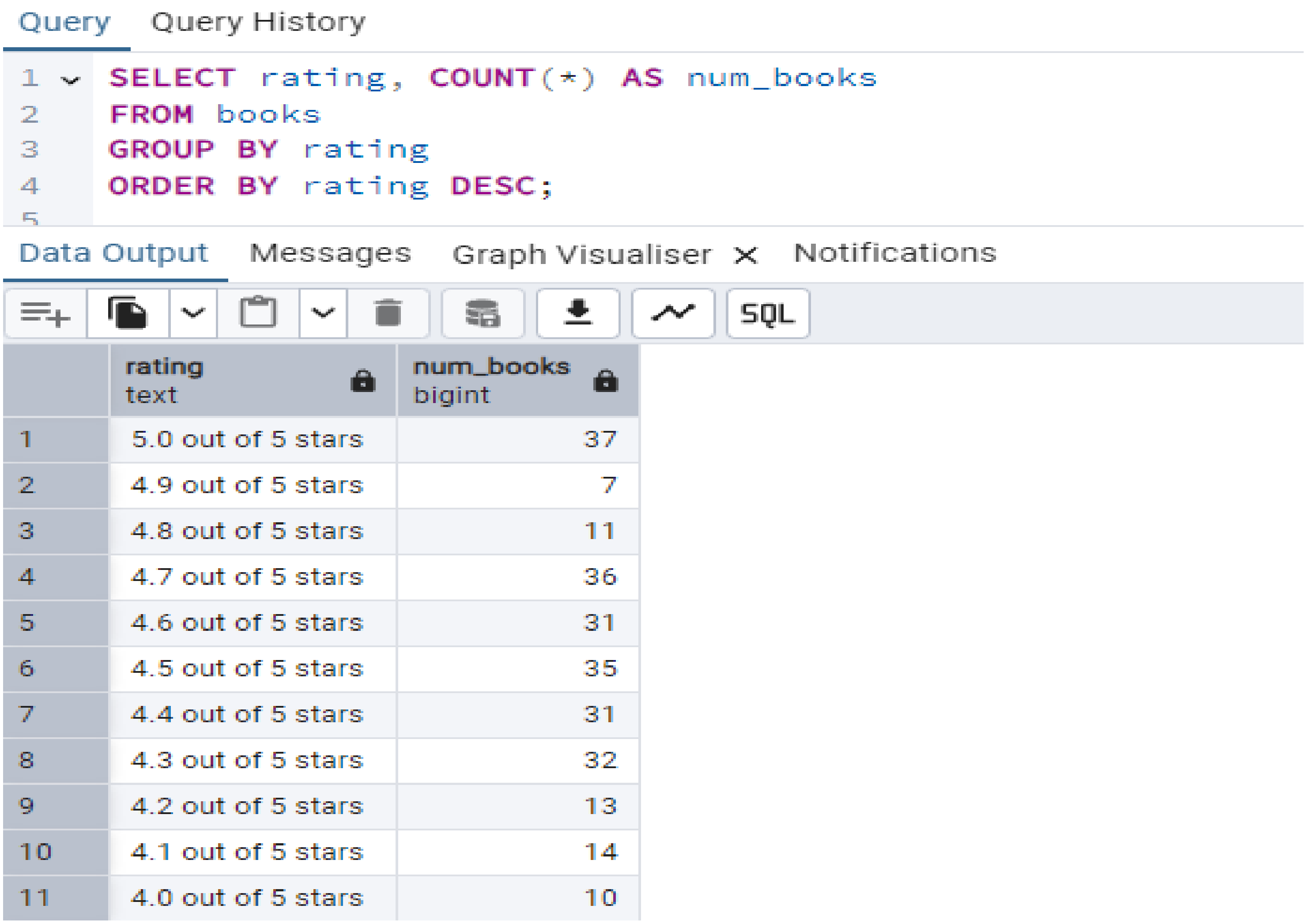


```sql
SELECT title, authors, CAST(price AS DECIMAL) AS price
FROM books
ORDER BY price Asc
LIMIT 10;
```

Showing rows: 1 to 10 Page No: 1 of

	title text	authors text	price numeric
1	GPS Big Data and Mobility Analysis: A practical guide with 18 real case studies to effectively understand and use Big Data in urban planning, transportation, ... traffic models (Transport Big Data B...	Kindle	0
2	Prompt Engineering for Researchers: Transform data into insights: A researcher's guide to effective prompts	Kindle	0
3	The Data Engineer's Pocketbook: FastTrack to Expertise: Your Compact Guide with Industry-Relevant Use Cases.	Brahma Reddy Katam	0
4	97 Things Every Data Engineer Should Know: Collective Wisdom from the Experts	Audible Audiobook	0
5	SSIS Data Warehouse Development - 101 Interview Questions: Earn over £50,000 per annum using SSIS and SQL Server (The Data Engineering Series)	Kindle	0
6	Microsoft SSIS SSAS SSRS Development: 450 Detailed Business Intelligence Q&As (The Data Engineering Series)	Kindle	0
7	Fundamentals of Data Engineering: Efficiency, Insight, Impact: Transforming Data into Value through Expert Engineering Practices	Joseph Achakji	0
8	DATA ENGINEERING AND AI FOR BEGINNERS: Revolutionizing Data Processing and Analytics by Leveraging Artificial Intelligence for Efficient Input Collection, Storage, and Transformation (World...	Kindle	0
9	Ace the Data Engineering Interview: Questions and Answers for Python, SQL, Data Modeling and More	Sean Coyne	0
10	Data Analytics, Data Visualization & Communicating Data: 3 books in 1: Learn the Processes of Data Analytics and Data Science, Create Engaging Data Visualizations, and Present Data Effectively	Audible Audiobook	0

8. Ratings Statistics



```sql
SELECT rating, COUNT(*) AS num_books
FROM books
GROUP BY rating
ORDER BY rating DESC;
```

	rating text	num_books bigint
1	5.0 out of 5 stars	37
2	4.9 out of 5 stars	7
3	4.8 out of 5 stars	11
4	4.7 out of 5 stars	36
5	4.6 out of 5 stars	31
6	4.5 out of 5 stars	35
7	4.4 out of 5 stars	31
8	4.3 out of 5 stars	32
9	4.2 out of 5 stars	13
10	4.1 out of 5 stars	14
11	4.0 out of 5 stars	10