RWANDA POLYTECHNIC

NGOMA COLLEGE

P.O. Box 35 KIBUNGO - RWANDA
Tel: +250 785 – 883 - 746
Email: info@iprcngoma.rp.ac.rw
www.iprcngoma.rp.ac.rw

Module Detail		Trainee's Detail		
SECTOR:	ICT	Reg No:	1.23RP00498 2.	
SUB-SECTOR:	Information Technology	Class:	Level 8 Information Technology	
		Trainer's Detail		
CERTIFICATE:	Bachelors of Technology	Name:	Eng. NYIRIMANA J.M Vianney	
MODULE (Code &Title):	ITLDM801 – DATA MINING AND DATA WAREHOUSE	Additional info		
Competence:	Apply Data Mining and Warehousing	Duration:		
		Due date:	30 March, 2025	
Training Centre:	RP Ngoma College	Signature:		
Scored marks:		Decision:	Competent	
			Not Yet Competent	

Store Books ETL and Data Warehouse Project Report

1. Summary

This project aims to transform the existing Store Books Sales relational database into a data warehouse to enable advanced analytics, reporting, and business intelligence. The process involves Extract, Transform, Load **(ETL)** operations to migrate data from **Microsoft SQL Server 2022** into a structured warehouse model. The warehouse will support historical data analysis, trend identification, and decision-making for book sales management.

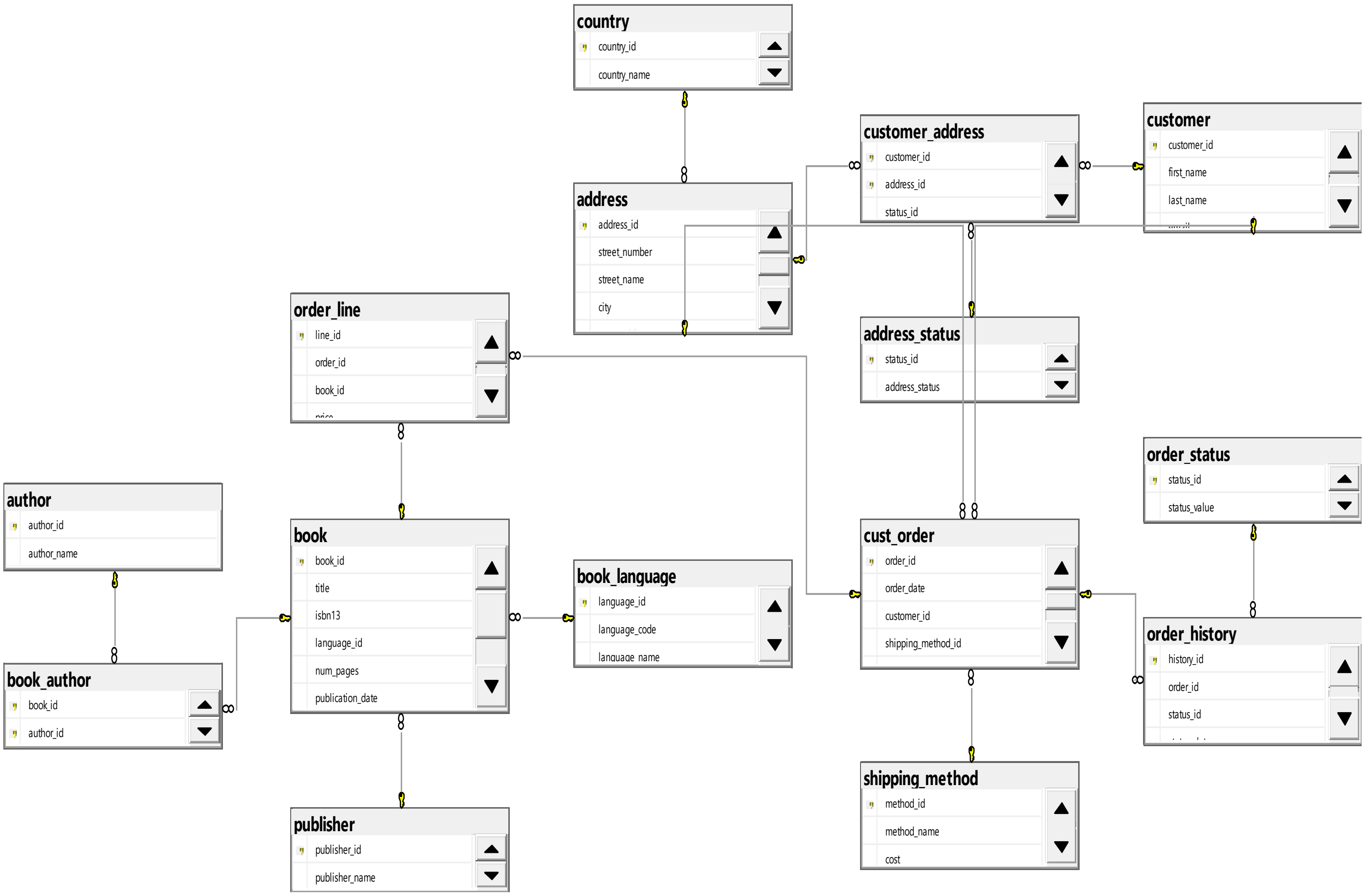
2. Source Database Overview

The current database is set up as a relational database using **Microsoft SQL Server 2022**. It is managed with **SQL Server Management Studio** (SSMS), which helps in organizing, querying, and optimizing the data. The database includes several key tables that store information about authors, publishers, book languages, and books. These tables are connected to each other through relationships using foreign keys.

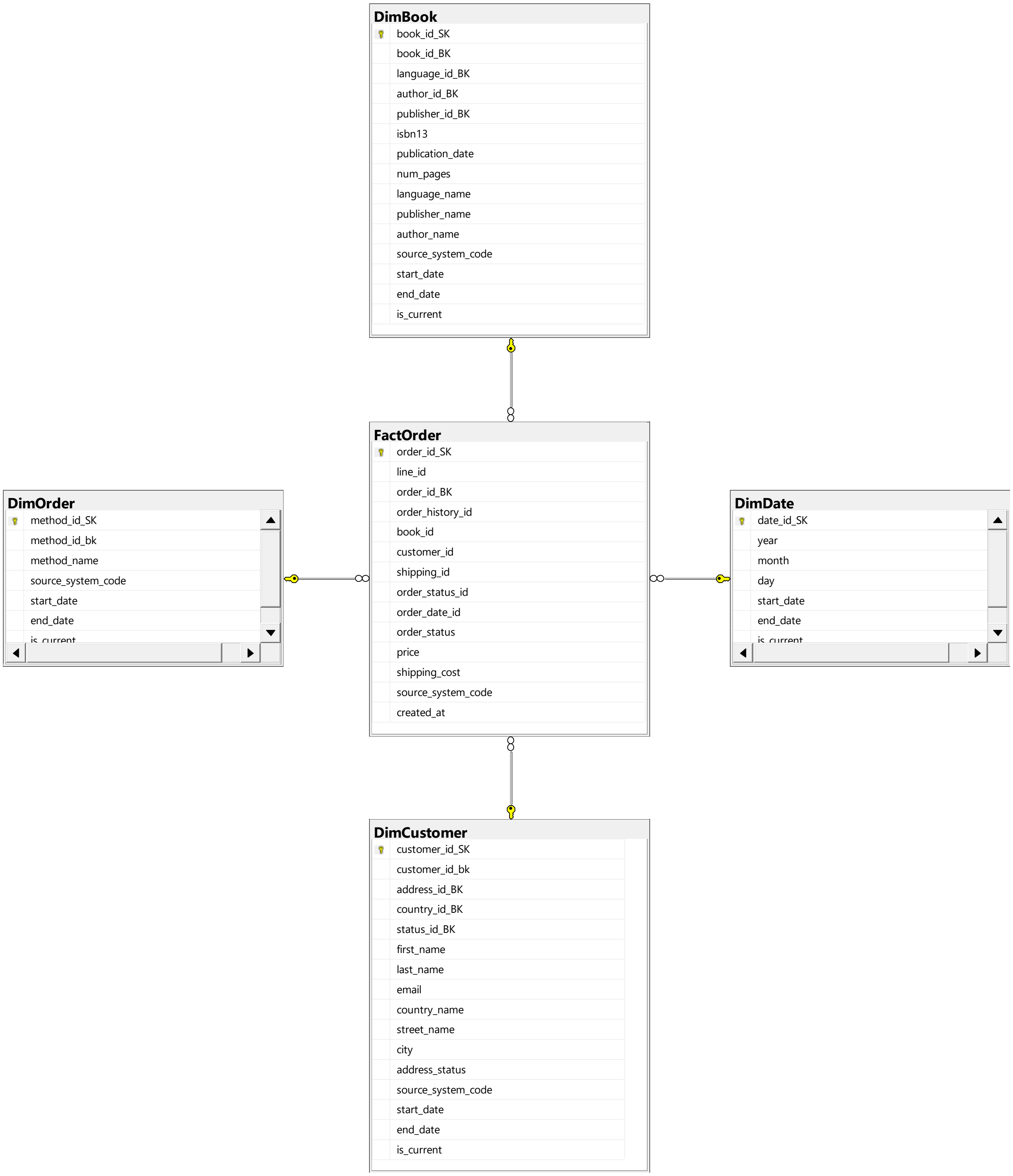
Here are the main tables in the database:

- **Author Table**: This table contains information about the authors, with a unique ID for each author.
- **Publisher Table**: Stores information about the publishers, like the publisher’s name and ID.
- **Book Language Table**: Holds details about the languages of books, such as language codes and names.
- **Book Table:** This table has information about books, such as the title, ISBN number, number of pages, and publication date. It also connects to the Author, Publisher, and Language tables through foreign keys.
- **Book Author Table:** This table shows the many-to-many relationship between books and authors, linking each book to one or more authors.

The following below is the overview of the Store Books Sales source database schema structure which shows us an illustration of tables, relationships, and key entities



Data Warehouse Schema (StoreBooksDW)



DimBook

- book_id_SK
- book_id_BK
- language_id_BK
- author_id_BK
- publisher_id_BK
- isbn13
- publication_date
- num_pages
- language_name
- publisher_name
- author_name
- source_system_code
- start_date
- end_date
- is_current

FactOrder

- order_id_SK
- line_id
- order_id_BK
- order_history_id
- book_id
- customer_id
- shipping_id
- order_status_id
- order_date_id
- order_status
- price
- shipping_cost
- source_system_code
- created_at

DimOrder

- method_id_SK
- method_id_bk
- method_name
- source_system_code
- start_date
- end_date
- is_current

DimDate

- date_id_SK
- year
- month
- day
- start_date
- end_date
- is_current

DimCustomer

- customer_id_SK
- customer_id_bk
- address_id_BK
- country_id_BK
- status_id_BK
- first_name
- last_name
- email
- country_name
- street_name
- city
- address_status
- source_system_code
- start_date
- end_date
- is_current

3. Installation of Libraries

! pip install sqlalchemy pyodbc

!pip install pandas sqlalchemy psycopg2

!pip install sqlalchemy pyodbc

4. Import Necessary Libraries

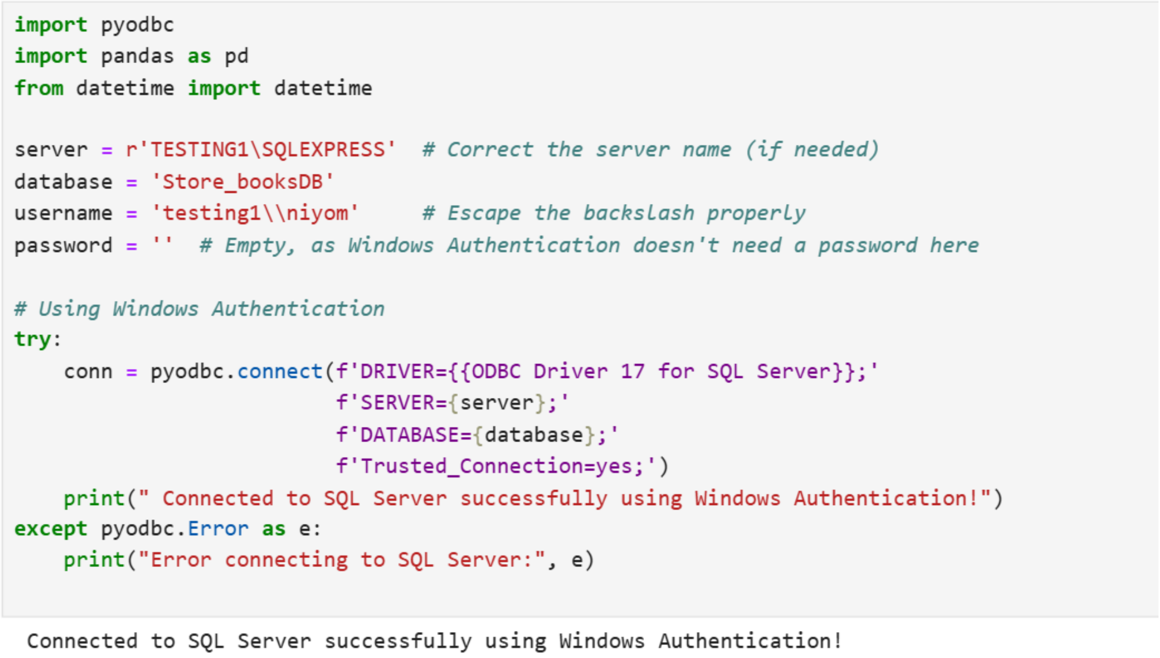
import pyodbc

import pandas as pd

from datetime import datetime

5. Database connection

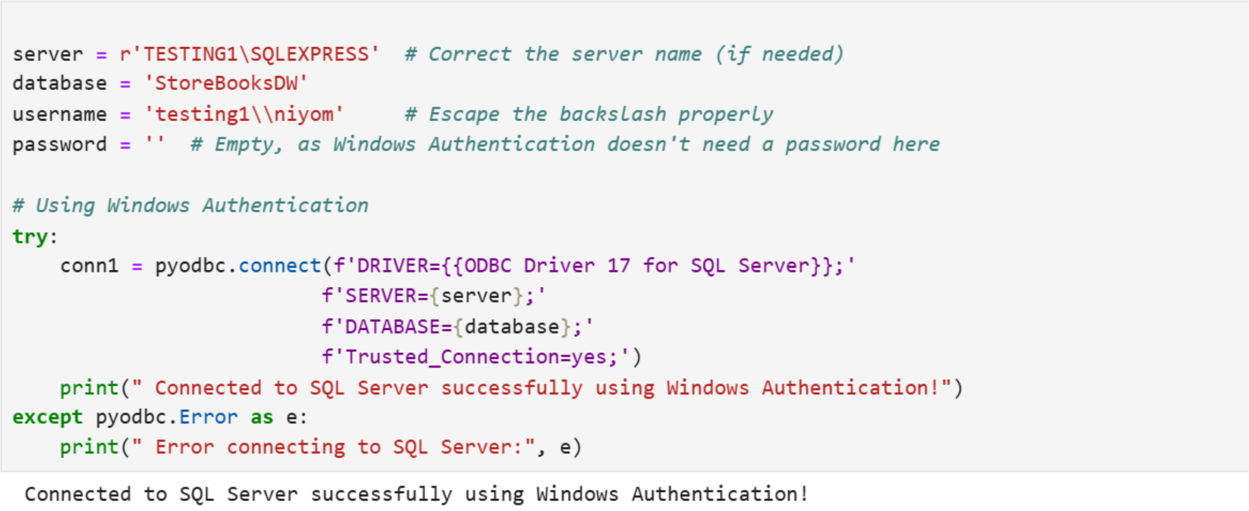
➢ Database Connection



```
import pyodbc
import pandas as pd
from datetime import datetime

server = r'TESTING1\SQLEXPRESS'  # Correct the server name (if needed)
database = 'Store_booksDB'
username = 'testing1\\niyom'     # Escape the backslash properly
password = ''  # Empty, as Windows Authentication doesn't need a password here

# Using Windows Authentication
try:
    conn = pyodbc.connect(f'DRIVER={{ODBC Driver 17 for SQL Server}};'
                          f'SERVER={server};'
                          f'DATABASE={database};'
                          f'Trusted_Connection=yes;')
    print(" Connected to SQL Server successfully using Windows Authentication!")
except pyodbc.Error as e:
    print("Error connecting to SQL Server:", e)
```

```
 Connected to SQL Server successfully using Windows Authentication!
```

➢ Data warehouse Connection



```
server = r'TESTING1\SQLEXPRESS'  # Correct the server name (if needed)
database = 'StoreBooksDW'
username = 'testing1\\niyom'     # Escape the backslash properly
password = ''  # Empty, as Windows Authentication doesn't need a password here

# Using Windows Authentication
try:
    conn1 = pyodbc.connect(f'DRIVER={{ODBC Driver 17 for SQL Server}};'
                          f'SERVER={server};'
                          f'DATABASE={database};'
                          f'Trusted_Connection=yes;')
    print(" Connected to SQL Server successfully using Windows Authentication!")
except pyodbc.Error as e:
    print(" Error connecting to SQL Server:", e)
```

```
 Connected to SQL Server successfully using Windows Authentication!
```

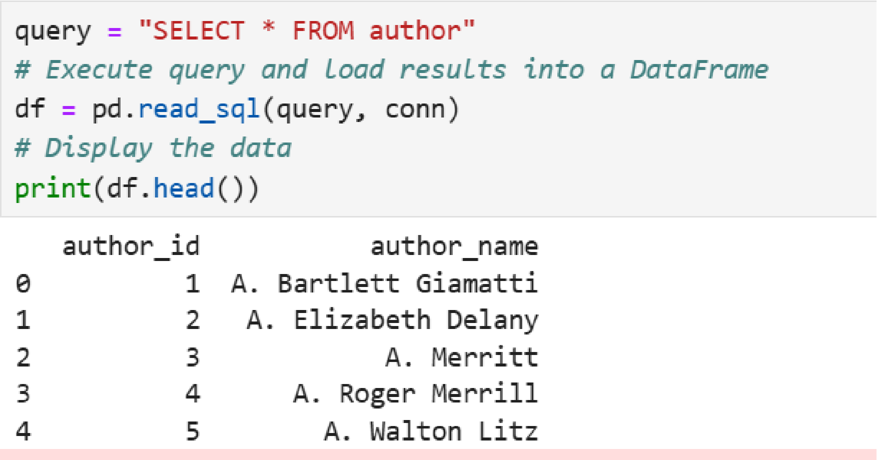
6. ETL Process Overview

Extraction Process

- ➢ Extract data from the existing source database (Store_booksDB) into staging tables.

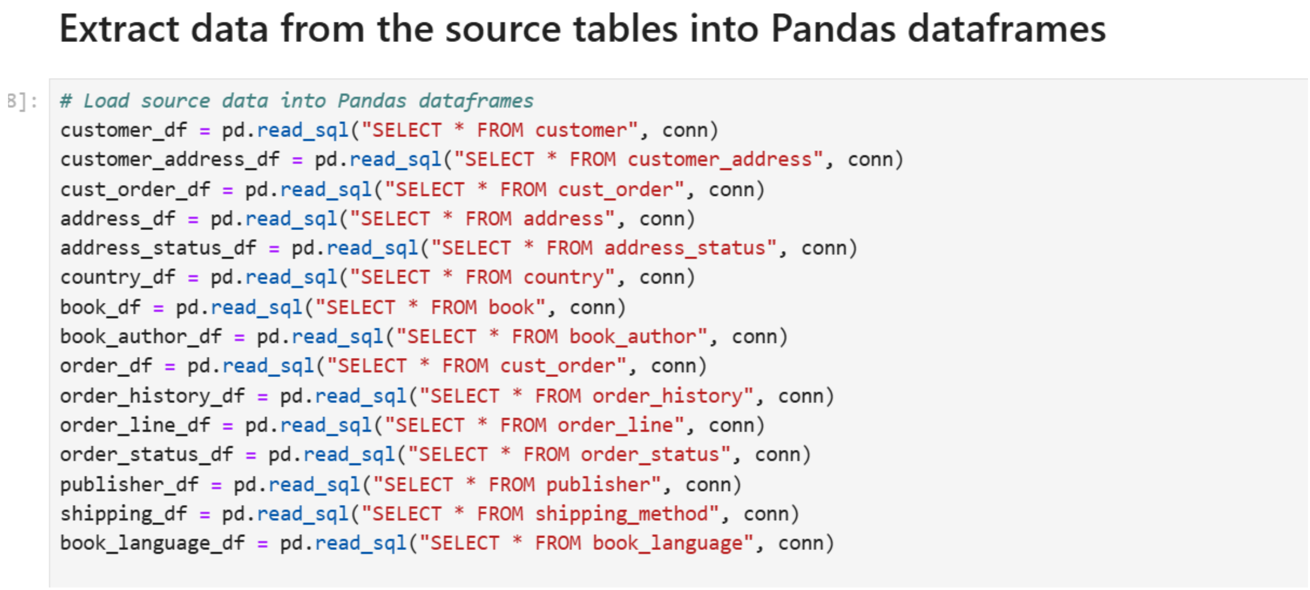
- ➢ Prepare data for transformation and loading into the data warehouse.

1. Extract data from author table



```
query = "SELECT * FROM author"
# Execute query and load results into a DataFrame
df = pd.read_sql(query, conn)
# Display the data
print(df.head())
```

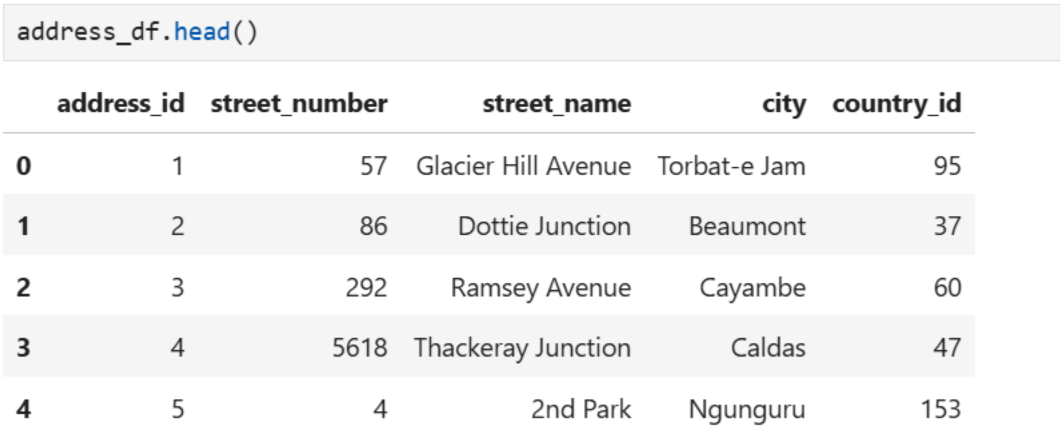
```
   author_id            author_name
0          1  A. Bartlett Giamatti
1          2   A. Elizabeth Delany
2          3            A. Merritt
3          4       A. Roger Merrill
4          5         A. Walton Litz
```



Extract data from the source tables into Pandas dataframes

```python
# Load source data into Pandas dataframes
customer_df = pd.read_sql("SELECT * FROM customer", conn)
customer_address_df = pd.read_sql("SELECT * FROM customer_address", conn)
cust_order_df = pd.read_sql("SELECT * FROM cust_order", conn)
address_df = pd.read_sql("SELECT * FROM address", conn)
address_status_df = pd.read_sql("SELECT * FROM address_status", conn)
country_df = pd.read_sql("SELECT * FROM country", conn)
book_df = pd.read_sql("SELECT * FROM book", conn)
book_author_df = pd.read_sql("SELECT * FROM book_author", conn)
order_df = pd.read_sql("SELECT * FROM cust_order", conn)
order_history_df = pd.read_sql("SELECT * FROM order_history", conn)
order_line_df = pd.read_sql("SELECT * FROM order_line", conn)
order_status_df = pd.read_sql("SELECT * FROM order_status", conn)
publisher_df = pd.read_sql("SELECT * FROM publisher", conn)
shipping_df = pd.read_sql("SELECT * FROM shipping_method", conn)
book_language_df = pd.read_sql("SELECT * FROM book_language", conn)
```

2. **Reading data from address table**



```python
address_df.head()
```

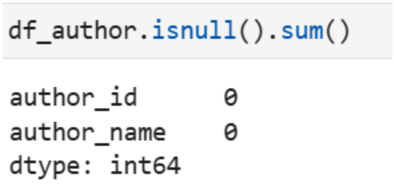
	address_id	street_number	street_name	city	country_id
0	1	57	Glacier Hill Avenue	Torbat-e Jam	95
1	2	86	Dottie Junction	Beaumont	37
2	3	292	Ramsey Avenue	Cayambe	60
3	4	5618	Thackeray Junction	Caldas	47
4	5	4	2nd Park	Ngunguru	153

Transformation

- Data Cleaning: Handle any inconsistencies or missing values in the source data.

- Data Mapping: Map source tables to data warehouse structures.

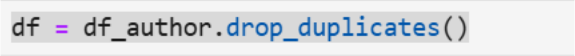
1. **Checking the null values**



```python
df_author.isnull().sum()
```

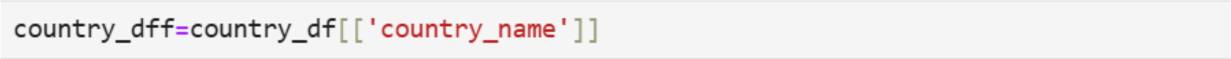
```
author_id      0
author_name    0
dtype: int64
```

2. drop duplicates values

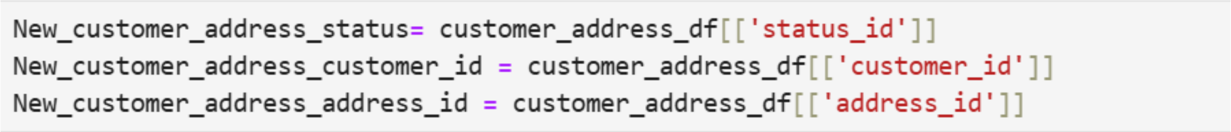


```
df = df_author.drop_duplicates()
```

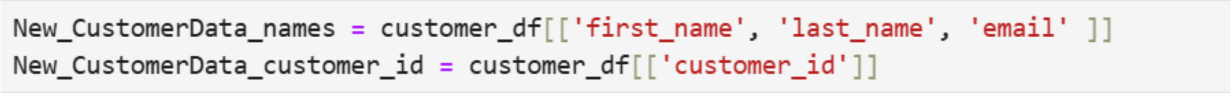
3. Selecting columns from pandas data frame with transformation



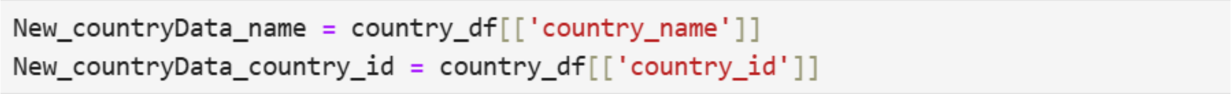
```
country_dff=country_df[['country_name']]
```



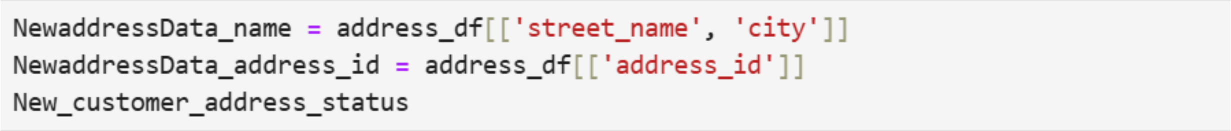
```
New_customer_address_status= customer_address_df[['status_id']]
New_customer_address_customer_id = customer_address_df[['customer_id']]
New_customer_address_address_id = customer_address_df[['address_id']]
```



```
New_CustomerData_names = customer_df[['first_name', 'last_name', 'email' ]]
New_CustomerData_customer_id = customer_df[['customer_id']]
```

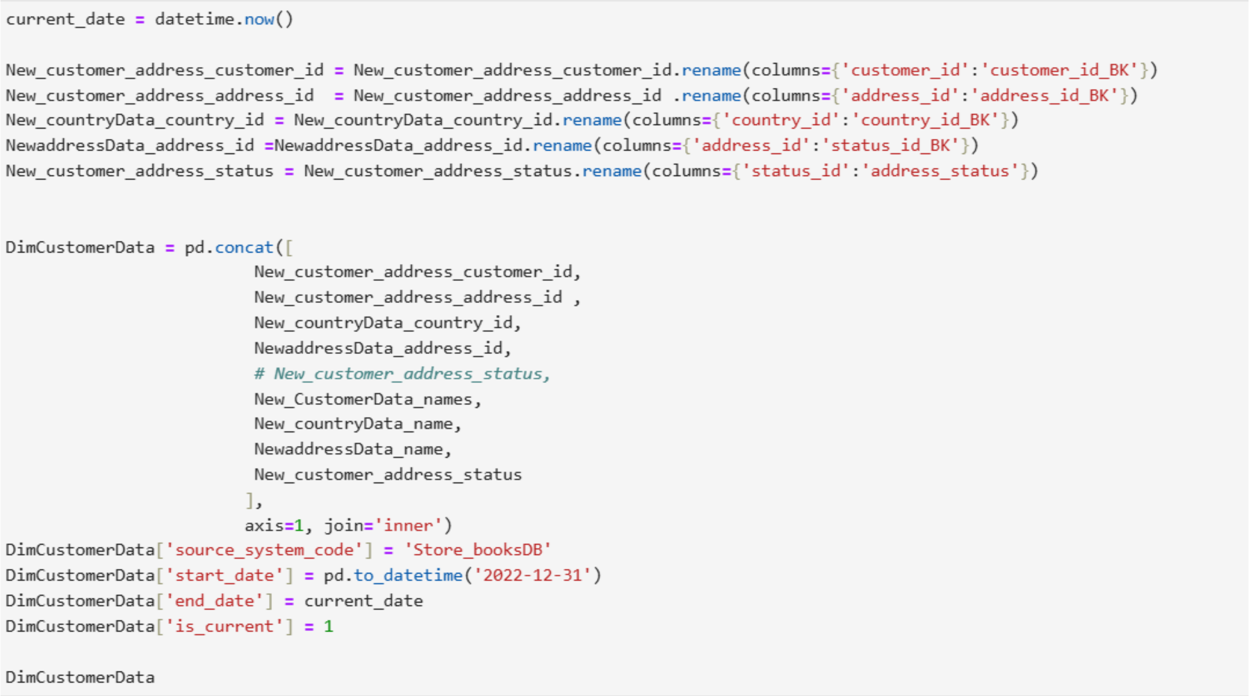


```
New_countryData_name = country_df[['country_name']]
New_countryData_country_id = country_df[['country_id']]
```



```
NewaddressData_name = address_df[['street_name', 'city']]
NewaddressData_address_id = address_df[['address_id']]
New_customer_address_status
```

Rename and concatenate columns of DimCustomerData

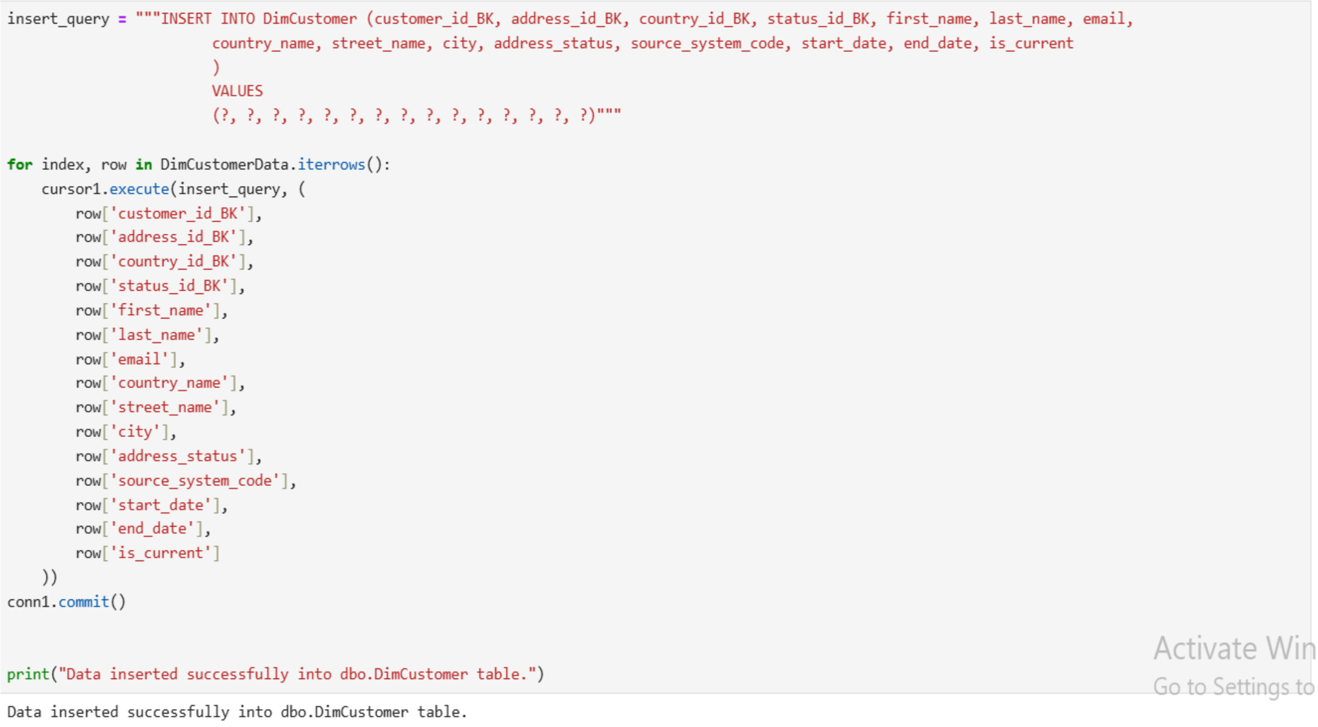


```
current_date = datetime.now()

New_customer_address_customer_id = New_customer_address_customer_id.rename(columns={'customer_id':'customer_id_BK'})
New_customer_address_address_id  = New_customer_address_address_id .rename(columns={'address_id':'address_id_BK'})
New_countryData_country_id = New_countryData_country_id.rename(columns={'country_id':'country_id_BK'})
NewaddressData_address_id =NewaddressData_address_id.rename(columns={'address_id':'status_id_BK'})
New_customer_address_status = New_customer_address_status.rename(columns={'status_id':'address_status'})


DimCustomerData = pd.concat([
                         New_customer_address_customer_id,
                         New_customer_address_address_id ,
                         New_countryData_country_id,
                         NewaddressData_address_id,
                         # New_customer_address_status,
                         New_CustomerData_names,
                         New_countryData_name,
                         NewaddressData_name,
                         New_customer_address_status
                        ],
                        axis=1, join='inner')
DimCustomerData['source_system_code'] = 'Store_booksDB'
DimCustomerData['start_date'] = pd.to_datetime('2022-12-31')
DimCustomerData['end_date'] = current_date
DimCustomerData['is_current'] = 1

DimCustomerData
```

Loading

- Load the transformed data into the data warehouse tables (StoreBooksDW).

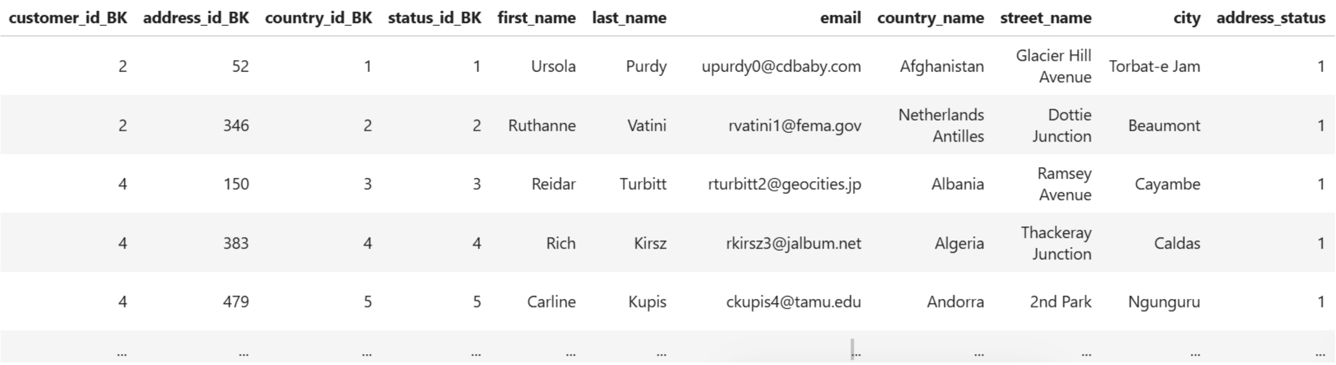
Loading Data to DimCustomer Table



```
insert_query = """INSERT INTO DimCustomer (customer_id_BK, address_id_BK, country_id_BK, status_id_BK, first_name, last_name, email,
                        country_name, street_name, city, address_status, source_system_code, start_date, end_date, is_current
                        )
                        VALUES
                        (?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?)"""

for index, row in DimCustomerData.iterrows():
    cursor1.execute(insert_query, (
        row['customer_id_BK'],
        row['address_id_BK'],
        row['country_id_BK'],
        row['status_id_BK'],
        row['first_name'],
        row['last_name'],
        row['email'],
        row['country_name'],
        row['street_name'],
        row['city'],
        row['address_status'],
        row['source_system_code'],
        row['start_date'],
        row['end_date'],
        row['is_current']
    ))
conn1.commit()


print("Data inserted successfully into dbo.DimCustomer table.")
```

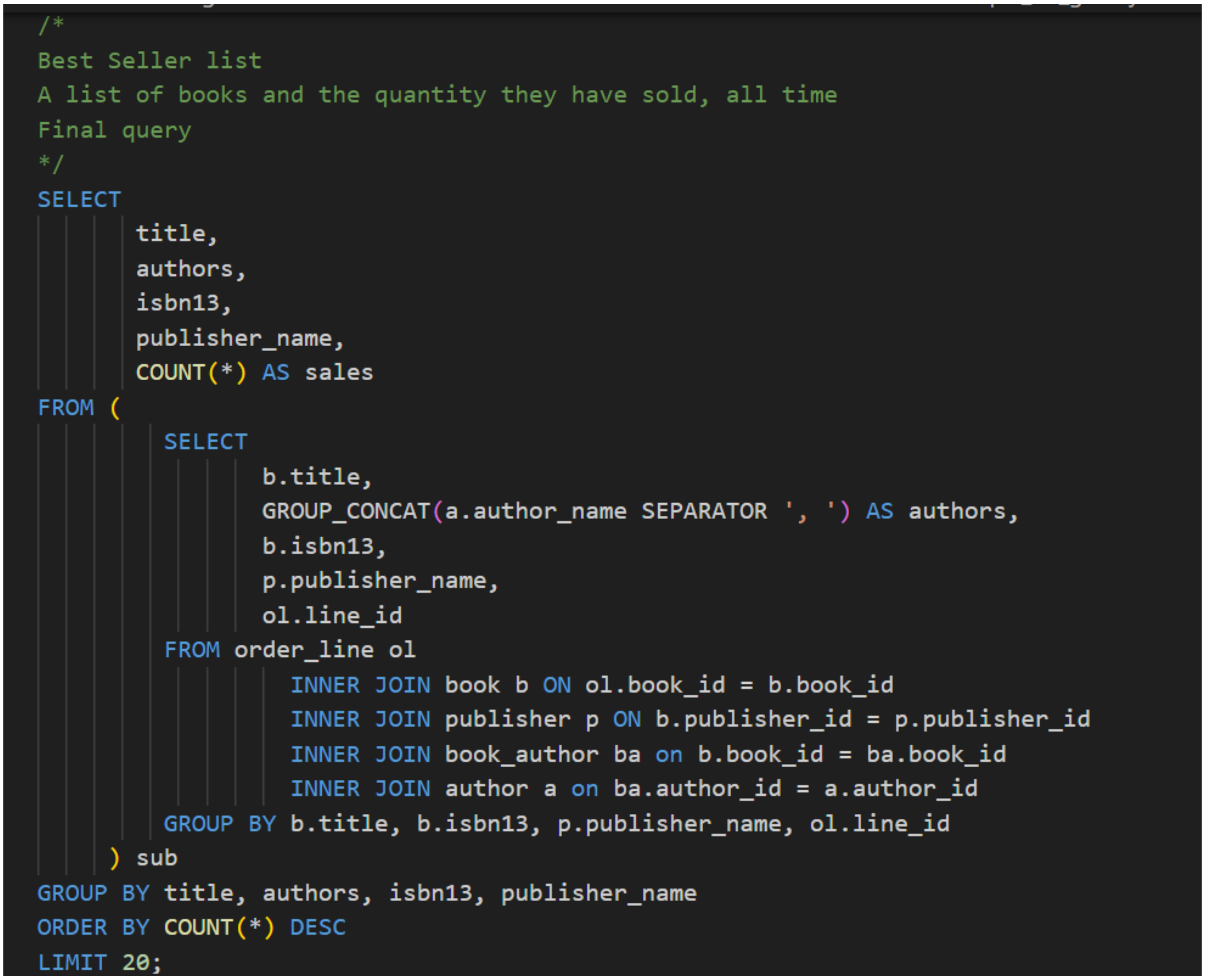
```
Data inserted successfully into dbo.DimCustomer table.
```

OUTPUT



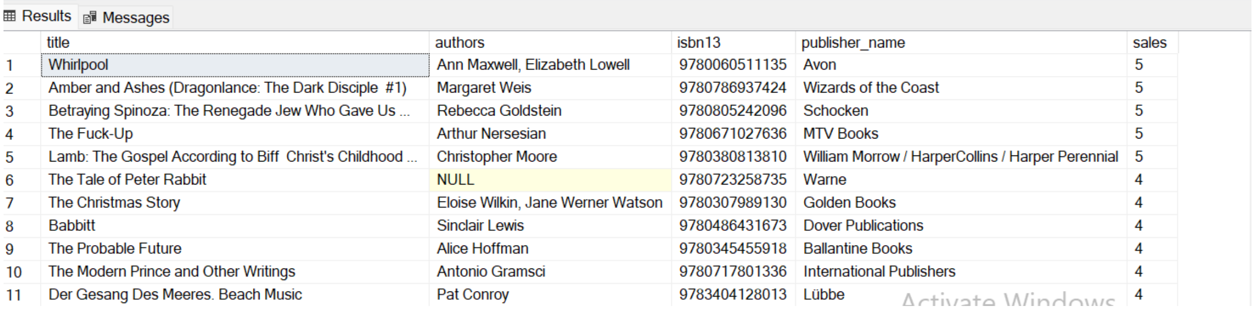
customer_id_BK	address_id_BK	country_id_BK	status_id_BK	first_name	last_name	email	country_name	street_name	city	address_status
2	52	1	1	Ursola	Purdy	upurdy0@cdbaby.com	Afghanistan	Glacier Hill Avenue	Torbat-e Jam	1
2	346	2	2	Ruthanne	Vatini	rvatini1@fema.gov	Netherlands Antilles	Dottie Junction	Beaumont	1
4	150	3	3	Reidar	Turbitt	rturbitt2@geocities.jp	Albania	Ramsey Avenue	Cayambe	1
4	383	4	4	Rich	Kirsz	rkirsz3@jalbum.net	Algeria	Thackeray Junction	Caldas	1
4	479	5	5	Carline	Kupis	ckupis4@tamu.edu	Andorra	2nd Park	Ngunguru	1
...	...	...	...	...	...	...	...	...	...	...

Final query

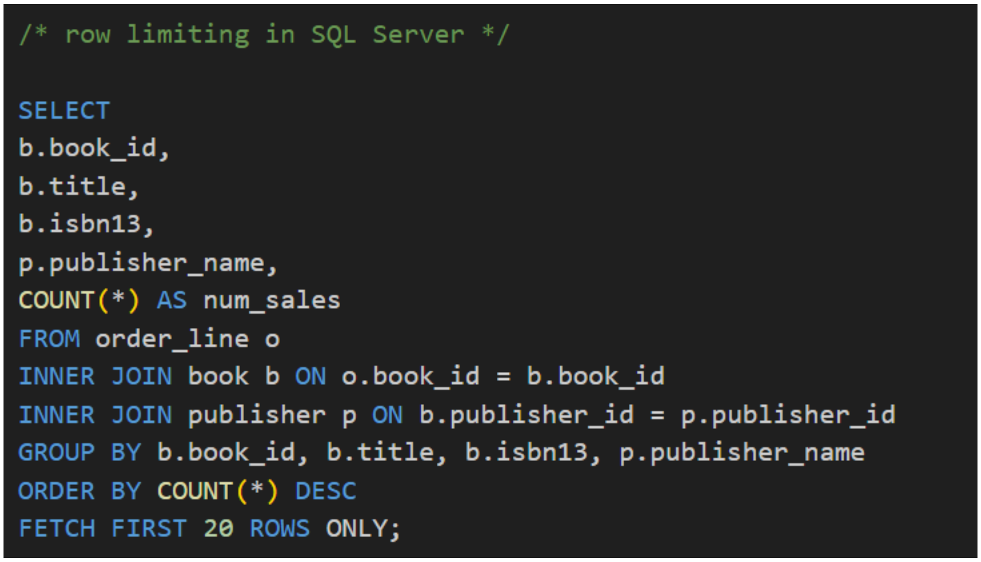


```sql
/*
Best Seller list
A list of books and the quantity they have sold, all time
Final query
*/
SELECT
       title,
       authors,
       isbn13,
       publisher_name,
       COUNT(*) AS sales
FROM (
         SELECT
                b.title,
                GROUP_CONCAT(a.author_name SEPARATOR ', ') AS authors,
                b.isbn13,
                p.publisher_name,
                ol.line_id
         FROM order_line ol
                  INNER JOIN book b ON ol.book_id = b.book_id
                  INNER JOIN publisher p ON b.publisher_id = p.publisher_id
                  INNER JOIN book_author ba on b.book_id = ba.book_id
                  INNER JOIN author a on ba.author_id = a.author_id
         GROUP BY b.title, b.isbn13, p.publisher_name, ol.line_id
      ) sub
GROUP BY title, authors, isbn13, publisher_name
ORDER BY COUNT(*) DESC
LIMIT 20;
```

Output

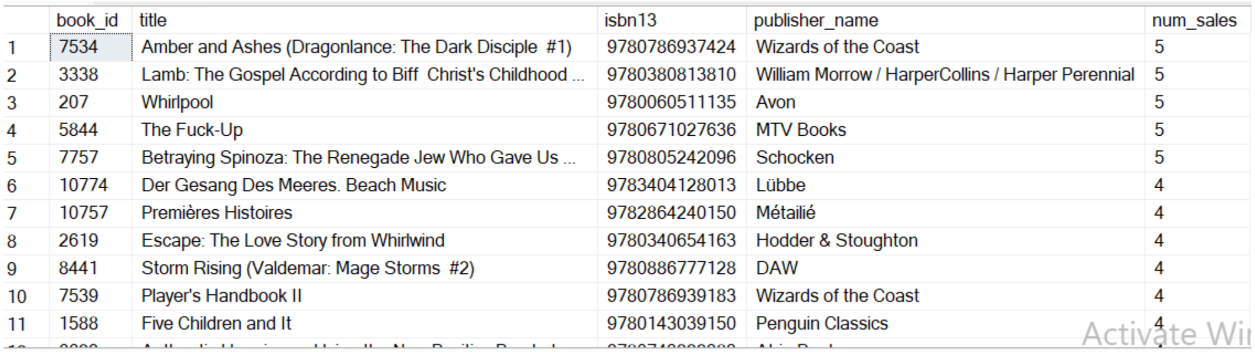


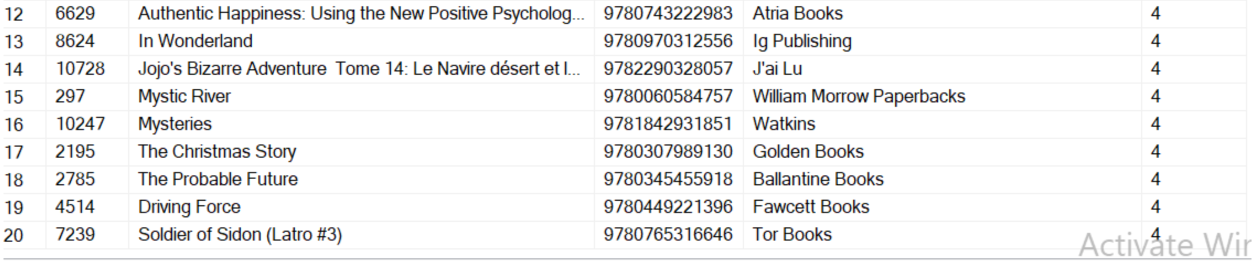
	title	authors	isbn13	publisher_name	sales
1	Whirlpool	Ann Maxwell, Elizabeth Lowell	9780060511135	Avon	5
2	Amber and Ashes (Dragonlance: The Dark Disciple #1)	Margaret Weis	9780786937424	Wizards of the Coast	5
3	Betraying Spinoza: The Renegade Jew Who Gave Us ...	Rebecca Goldstein	9780805242096	Schocken	5
4	The Fuck-Up	Arthur Nersesian	9780671027636	MTV Books	5
5	Lamb: The Gospel According to Biff Christ's Childhood ...	Christopher Moore	9780380813810	William Morrow / HarperCollins / Harper Perennial	5
6	The Tale of Peter Rabbit	NULL	9780723258735	Warne	4
7	The Christmas Story	Eloise Wilkin, Jane Werner Watson	9780307989130	Golden Books	4
8	Babbitt	Sinclair Lewis	9780486431673	Dover Publications	4
9	The Probable Future	Alice Hoffman	9780345455918	Ballantine Books	4
10	The Modern Prince and Other Writings	Antonio Gramsci	9780717801336	International Publishers	4
11	Der Gesang Des Meeres. Beach Music	Pat Conroy	9783404128013	Lübbe	4



```sql
/* row limiting in SQL Server */

SELECT
b.book_id,
b.title,
b.isbn13,
p.publisher_name,
COUNT(*) AS num_sales
FROM order_line o
INNER JOIN book b ON o.book_id = b.book_id
INNER JOIN publisher p ON b.publisher_id = p.publisher_id
GROUP BY b.book_id, b.title, b.isbn13, p.publisher_name
ORDER BY COUNT(*) DESC
FETCH FIRST 20 ROWS ONLY;
```

Output

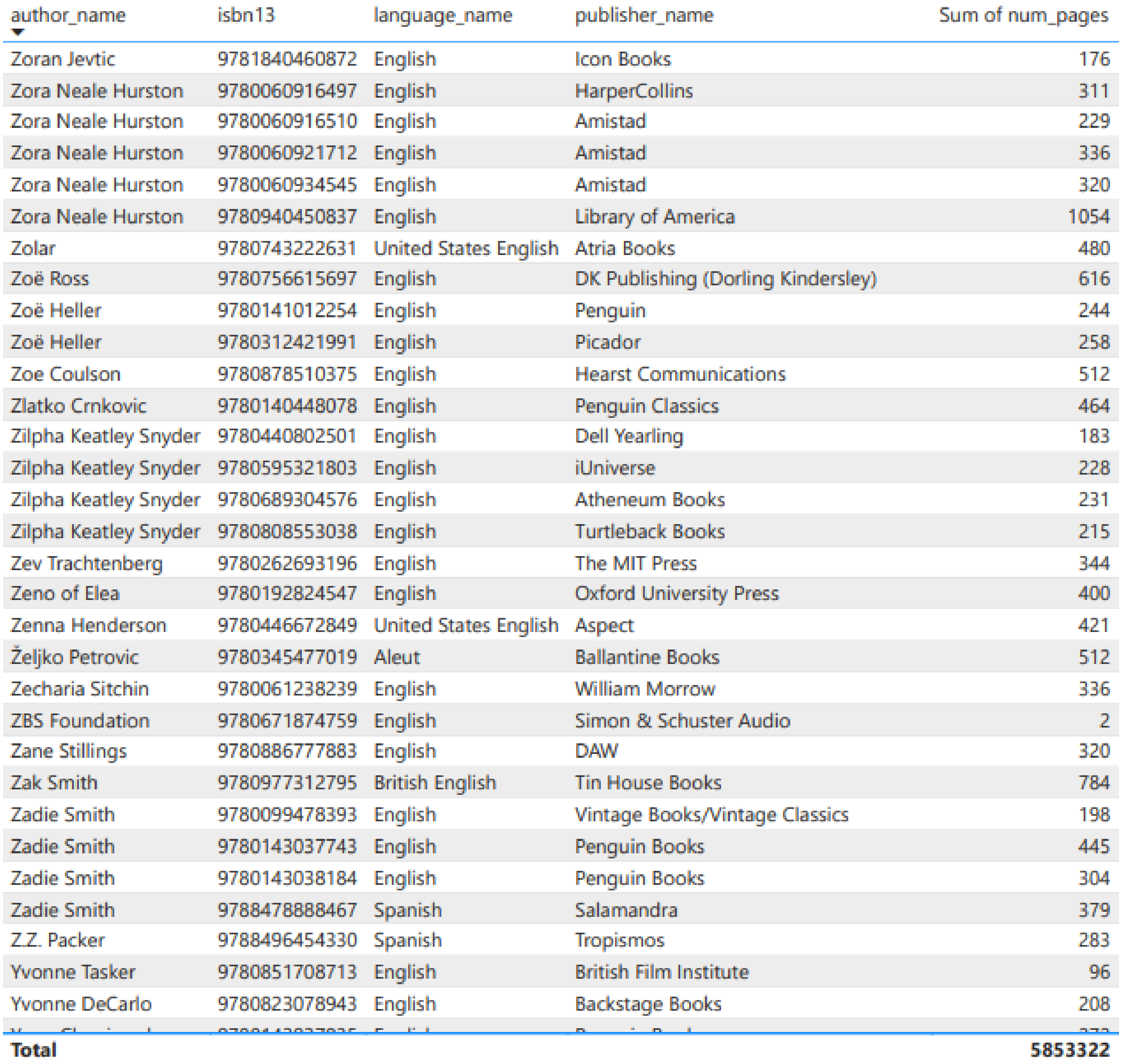




	book_id	title	isbn13	publisher_name	num_sales
1	7534	Amber and Ashes (Dragonlance: The Dark Disciple #1)	9780786937424	Wizards of the Coast	5
2	3338	Lamb: The Gospel According to Biff Christ's Childhood ...	9780380813810	William Morrow / HarperCollins / Harper Perennial	5
3	207	Whirlpool	9780060511135	Avon	5
4	5844	The Fuck-Up	9780671027636	MTV Books	5
5	7757	Betraying Spinoza: The Renegade Jew Who Gave Us ...	9780805242096	Schocken	5
6	10774	Der Gesang Des Meeres. Beach Music	9783404128013	Lübbe	4
7	10757	Premières Histoires	9782864240150	Métailié	4
8	2619	Escape: The Love Story from Whirlwind	9780340654163	Hodder & Stoughton	4
9	8441	Storm Rising (Valdemar: Mage Storms #2)	9780886777128	DAW	4
10	7539	Player's Handbook II	9780786939183	Wizards of the Coast	4
11	1588	Five Children and It	9780143039150	Penguin Classics	4
12	6629	Authentic Happiness: Using the New Positive Psycholog...	9780743222983	Atria Books	4
13	8624	In Wonderland	9780970312556	Ig Publishing	4
14	10728	Jojo's Bizarre Adventure Tome 14: Le Navire désert et l...	9782290328057	J'ai Lu	4
15	297	Mystic River	9780060584757	William Morrow Paperbacks	4
16	10247	Mysteries	9781842931851	Watkins	4
17	2195	The Christmas Story	9780307989130	Golden Books	4
18	2785	The Probable Future	9780345455918	Ballantine Books	4
19	4514	Driving Force	9780449221396	Fawcett Books	4
20	7239	Soldier of Sidon (Latro #3)	9780765316646	Tor Books	4

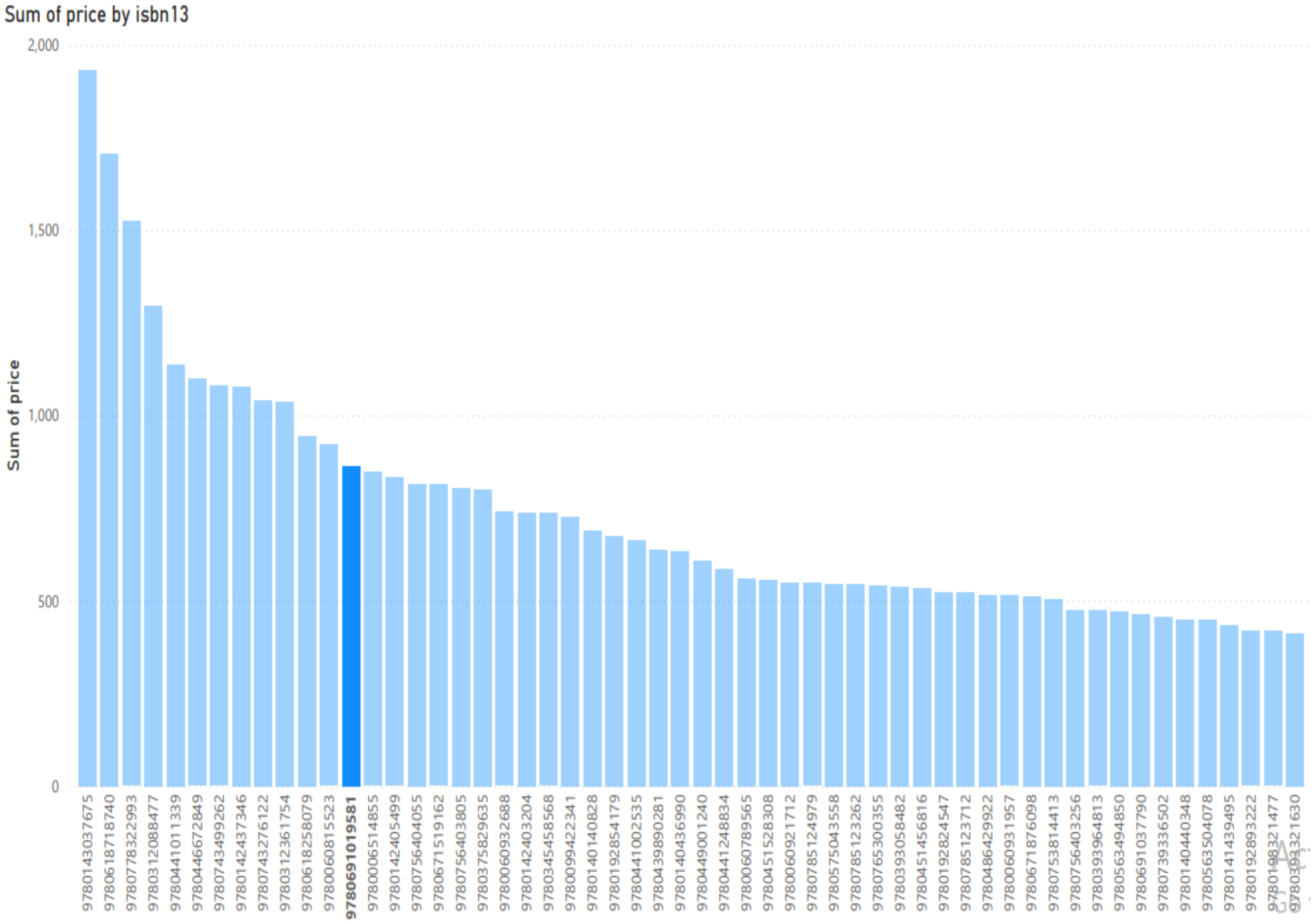
Power BI tool analysis

1. Book details

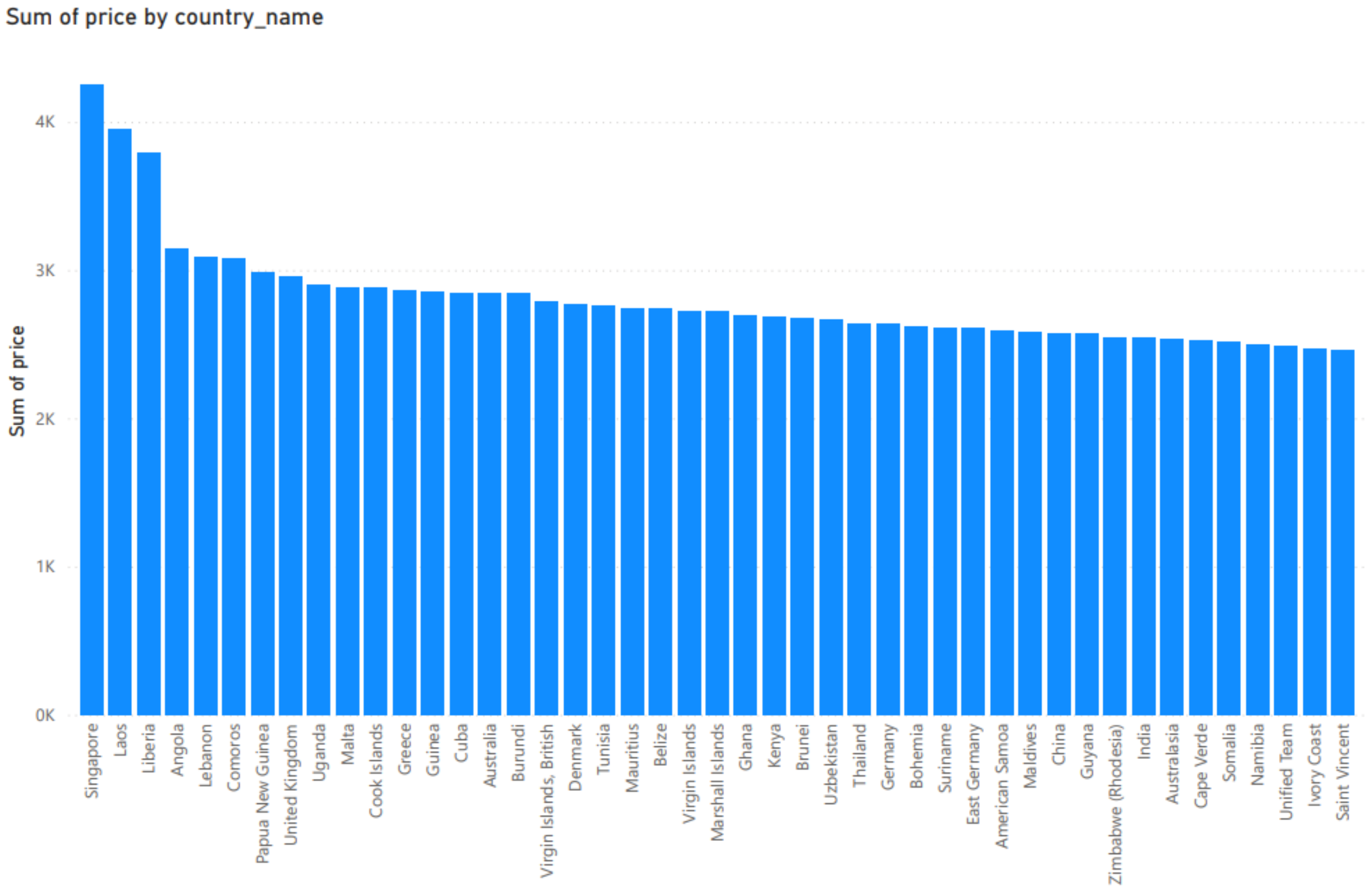


author_name	isbn13	language_name	publisher_name	Sum of num_pages
Zoran Jevtic	9781840460872	English	Icon Books	176
Zora Neale Hurston	9780060916497	English	HarperCollins	311
Zora Neale Hurston	9780060916510	English	Amistad	229
Zora Neale Hurston	9780060921712	English	Amistad	336
Zora Neale Hurston	9780060934545	English	Amistad	320
Zora Neale Hurston	9780940450837	English	Library of America	1054
Zolar	9780743222631	United States English	Atria Books	480
Zoë Ross	9780756615697	English	DK Publishing (Dorling Kindersley)	616
Zoë Heller	9780141012254	English	Penguin	244
Zoë Heller	9780312421991	English	Picador	258
Zoe Coulson	9780878510375	English	Hearst Communications	512
Zlatko Crnkovic	9780140448078	English	Penguin Classics	464
Zilpha Keatley Snyder	9780440802501	English	Dell Yearling	183
Zilpha Keatley Snyder	9780595321803	English	iUniverse	228
Zilpha Keatley Snyder	9780689304576	English	Atheneum Books	231
Zilpha Keatley Snyder	9780808553038	English	Turtleback Books	215
Zev Trachtenberg	9780262693196	English	The MIT Press	344
Zeno of Elea	9780192824547	English	Oxford University Press	400
Zenna Henderson	9780446672849	United States English	Aspect	421
Željko Petrovic	9780345477019	Aleut	Ballantine Books	512
Zecharia Sitchin	9780061238239	English	William Morrow	336
ZBS Foundation	9780671874759	English	Simon & Schuster Audio	2
Zane Stillings	9780886777883	English	DAW	320
Zak Smith	9780977312795	British English	Tin House Books	784
Zadie Smith	9780099478393	English	Vintage Books/Vintage Classics	198
Zadie Smith	9780143037743	English	Penguin Books	445
Zadie Smith	9780143038184	English	Penguin Books	304
Zadie Smith	9788478888467	Spanish	Salamandra	379
Z.Z. Packer	9788496454330	Spanish	Tropismos	283
Yvonne Tasker	9780851708713	English	British Film Institute	96
Yvonne DeCarlo	9780823078943	English	Backstage Books	208
Total				**5853322**

2. Top most book



3. Country vs Price


Sum of price by country_name
Sum of price
4K
3K
2K
1K
0K
Singapore
Laos
Liberia
Angola
Lebanon
Comoros
Papua New Guinea
United Kingdom
Uganda
Malta
Cook Islands
Greece
Guinea
Cuba
Australia
Burundi
Virgin Islands, British
Denmark
Tunisia
Mauritius
Belize
Virgin Islands
Marshall Islands
Ghana
Kenya
Brunei
Uzbekistan
Thailand
Germany
Bohemia
Suriname
East Germany
American Samoa
Maldives
China
Guyana
Zimbabwe (Rhodesia)
India
Australasia
Cape Verde
Somalia
Namibia
Unified Team
Ivory Coast
Saint Vincent