Проект: анализ базы данных сервиса чтения книг по подписке (SQL)

1 Описание проекта

Заказчик:

новый владелец крупного сервиса для чтения книг по подписке.

Цель проекта:

- сформулировать ценностное предложение для нового продукта

Задача проекта:

- исследовать исходные данные;
- посчитать, сколько книг вышло после 1 января 2000 года;
- для каждой книги посчитать количество обзоров и среднюю оценку;
- определить издательство, которое выпустило наибольшее число книг толще 50 страниц (так можно исключить из анализа брошюры);
- определить автора с самой высокой средней оценкой книг (учитываем только книги с 50 и более оценками);
- посчитать среднее количество обзоров от пользователей, которые поставили больше 48 оценок;
- сформулировать выводы;
- сформулировать предложение для нового продукта.

Источник данных:

Реляционная база данных postgresql (5 таблиц)

Описание данных:

1. Таблица `books`

Содержит данные о книгах:

- `book_id` — идентификатор книги;
- `author_id` — идентификатор автора;
- `title` — название книги;
- `num_pages` — количество страниц;
- `publication_date` — дата публикации книги;
- `publisher_id` — идентификатор издателя.

2. Таблица `authors`

Содержит данные об авторах:

- `author_id` — идентификатор автора;
- `author` — имя автора.

3. Таблица `publishers`

Содержит данные об издательствах:

- `publisher_id` — идентификатор издательства;
- `publisher` — название издательства.

4. Таблица `ratings`

Содержит данные о пользовательских оценках книг:

- `rating_id` — идентификатор оценки;
- `book_id` — идентификатор книги;
- `username` — имя пользователя, оставившего оценку;
- `rating` — оценка книги.

5. Таблица `reviews`

Содержит данные о пользовательских обзорах на книги:

- `review_id` — идентификатор обзора;
- `book_id` — идентификатор книги;

- `username` — имя пользователя, написавшего обзор;
- `text` — текст обзора.

К выводам

2 Подготовительный этап анализа

2.1 Загрузка библиотек

In [1]:
```
# загрузим библиотеки
import pandas as pd
from sqlalchemy import text, create_engine
```

2.2 Подключение к базе данных

In [2]:
```
# установим параметры
db_config = {'user': 'praktikum_student', # имя пользователя
'pwd': 'Sdf4$2;d-d30pp', # пароль
'host': 'rc1b-wcoijxj3yxfsf3fs.mdb.yandexcloud.net',
'port': 6432, # порт подключения
'db': 'data-analyst-final-project-db'} # название базы данных
connection_string = 'postgresql://{user}:{pwd}@{host}:{port}/{db}'.format(**db_config)

# сохраним коннектор
engine = create_engine(connection_string, connect_args={'sslmode':'require'})

con=engine.connect()
```

2.3 Исследование таблиц

2.3.1 Таблица books

In [3]:
```
# выполним SQL-запрос в Pandas
query = '''SELECT * FROM books'''
```

In [4]:
```
# выведем таблицу на экран
pd.io.sql.read_sql(sql=text(query), con = con)
```

Out[4]:

	book_id	author_id	title	num_pages	publication_date	publisher_id
0	1	546	'Salem's Lot	594	2005-11-01	93
1	2	465	1 000 Places to See Before You Die	992	2003-05-22	336
2	3	407	13 Little Blue Envelopes (Little Blue Envelope...	322	2010-12-21	135
3	4	82	1491: New Revelations of the Americas Before C...	541	2006-10-10	309
4	5	125	1776	386	2006-07-04	268
...	...	...	...	...	...	...
995	996	571	Wyrd Sisters (Discworld #6; Witches #2)	265	2001-02-06	147
996	997	454	Xenocide (Ender's Saga #3)	592	1996-07-15	297
997	998	201	Year of Wonders	358	2002-04-30	212
998	999	94	You Suck (A Love Story #2)	328	2007-01-16	331
999	1000	509	Zen and the Art of Motorcycle Maintenance: An ...	540	2006-04-25	143

1000 rows × 6 columns

Вывод:

- таблица books содержит 6 колонок и 1000 строк;
- названия столбцов:
  - book_id — идентификатор книги;
  - author_id — идентификатор автора;
  - title — название книги;
  - num_pages — количество страниц;
  - publication_date — дата публикации книги;

- publisher_id — идентификатор издателя.

2.3.2 Таблица authors

In [5]:
```
# выполним SQL-запрос в Pandas
query2 = '''SELECT * FROM authors'''
```

In [6]:
```
# выведем таблицу на экран
pd.io.sql.read_sql(sql=text(query2), con = con)
```

Out[6]:

	author_id	author
0	1	A.S. Byatt
1	2	Aesop/Laura Harris/Laura Gibbs
2	3	Agatha Christie
3	4	Alan Brennert
4	5	Alan Moore/David Lloyd
...	...	...
631	632	William Strunk Jr./E.B. White
632	633	Zadie Smith
633	634	Zilpha Keatley Snyder
634	635	Zora Neale Hurston
635	636	Åsne Seierstad/Ingrid Christopherson

636 rows × 2 columns

Вывод:

- таблица authors содержит 2 колонки и 636 строк;
- названия столбцов:
  - `author_id` — идентификатор автора;

- `author` — имя автора.

2.3.3 Таблица `publishers`

In [7]:
```
# выполним SQL-запрос в Pandas
query3 = '''SELECT * FROM publishers'''
```

In [8]:
```
# выведем таблицу на экран
pd.io.sql.read_sql(sql=text(query3), con = con)
```

Out[8]:

	publisher_id	publisher
0	1	Ace
1	2	Ace Book
2	3	Ace Books
3	4	Ace Hardcover
4	5	Addison Wesley Publishing Company
...	...	...
335	336	Workman Publishing Company
336	337	Wyatt Book
337	338	Yale University Press
338	339	Yearling
339	340	Yearling Books

340 rows × 2 columns

Вывод:

- таблица publishers содержит 2 колонки и 340 строк;
- названия столбцов:

- `publisher_id` — идентификатор издательства;
- `publisher` — название издательства.

2.3.4 Таблица `ratings`

In [9]:
```
# выполним SQL-запрос в Pandas
query4 = '''SELECT * FROM ratings'''
```

In [10]:
```
# выведем таблицу на экран
pd.io.sql.read_sql(sql=text(query4), con = con)
```

Out[10]:

	rating_id	book_id	username	rating
0	1	1	ryanfranco	4
1	2	1	grantpatricia	2
2	3	1	brandtandrea	5
3	4	2	lorichen	3
4	5	2	mariokeller	2
...	...	...	...	...
6451	6452	1000	carolrodriguez	4
6452	6453	1000	wendy18	4
6453	6454	1000	jarvispaul	5
6454	6455	1000	zross	2
6455	6456	1000	fharris	5

6456 rows × 4 columns

Вывод:

- таблица ratings содержит 4 колонки и 6456 строк;

- названия столбцов:
  - `rating_id` — идентификатор оценки;
  - `book_id` — идентификатор книги;
  - `username` — имя пользователя, оставившего оценку;
  - `rating` — оценка книги.

2.3.5 Таблица `reviews`

In [11]:
```
# выполним SQL-запрос в Pandas
query5 = '''SELECT * FROM reviews'''
```

In [12]:
```
# выведем таблицу на экран
pd.io.sql.read_sql(sql=text(query5), con = con)
```

Out[12]:

	review_id	book_id	username	text
0	1	1	brandtandrea	Mention society tell send professor analysis. ...
1	2	1	ryanfranco	Foot glass pretty audience hit themselves. Amo...
2	3	2	lorichen	Listen treat keep worry. Miss husband tax but ...
3	4	3	johnsonamanda	Finally month interesting blue could nature cu...
4	5	3	scotttamara	Nation purpose heavy give wait song will. List...
...	...	...	...	...
2788	2789	999	martinadam	Later hospital turn easy community. Fact same ...
2789	2790	1000	wknight	Change lose answer close pressure. Spend so now.
2790	2791	1000	carolrodriguez	Authority go who television entire hair guy po...
2791	2792	1000	wendy18	Or western offer wonder ask. More hear phone f...
2792	2793	1000	jarvispaul	Republican staff bit eat material measure plan...

2793 rows × 4 columns

Вывод:

- таблица reviews содержит 2 колонки и 2793 строк;
- названия столбцов: - `review_id` — идентификатор обзора;
  - `book_id` — идентификатор книги;
  - `username` — имя пользователя, написавшего обзор;
  - `text` — текст обзора.

3 Исследовательский этап анализа

3.1 Задача 1

Формулировка задачи:

- Посчитать, сколько книг вышло после 1 января 2000 года

In [13]:
```
# выполним SQL-запрос в Pandas и выведем результат на экран
query6 = '''
SELECT
    COUNT(DISTINCT book_id) AS count_books
FROM
    books
WHERE
    publication_date > '2000-01-01'
'''
pd.io.sql.read_sql(query6, con = engine)
```

Out[13]:

	count_books
0	819

Вывод:

- в базе данных сервиса для чтения книг хранится 819 книг, вышедших после 01.01.2000 г.

3.2 Задача 2

Формулировка задачи:

- Для каждой книги посчитать количество обзоров и среднюю оценку

In [14]:
```
# выполним SQL-запрос в Pandas и выведем результат на экран
query7 = '''
SELECT
    b.title AS book_name,
    COUNT(DISTINCT rv.review_id) AS count_reviews,
    AVG(rt.rating) AS avg_rating
FROM books AS b
LEFT JOIN reviews AS rv USING(book_id)
LEFT JOIN ratings AS rt USING(book_id)
GROUP BY b.book_id, 1
ORDER BY 3 DESC, 2 DESC
'''
pd.io.sql.read_sql(query7, con = engine)
```

Out[14]:

	book_name	count_reviews	avg_rating
0	A Dirty Job (Grim Reaper #1)	4	5.00
1	School's Out—Forever (Maximum Ride #2)	3	5.00
2	Moneyball: The Art of Winning an Unfair Game	3	5.00
3	Arrows of the Queen (Heralds of Valdemar #1)	2	5.00
4	Wherever You Go There You Are: Mindfulness Me...	2	5.00
...	...	...	...
995	The World Is Flat: A Brief History of the Twen...	3	2.25
996	Drowning Ruth	3	2.00
997	His Excellency: George Washington	2	2.00
998	Junky	2	2.00
999	Harvesting the Heart	2	1.50

1000 rows × 3 columns

Вывод:

- максимальный средний рейтинг у книг - 5;

- минимальный средний рейтинг - 1.5;
- максимальное кол-во обзоров - 8;
- минимальное кол-во обзоров - 0;
- топ-3 лучших по рейтингу книг с максимальным кол-вом обзоров:
  - A Dirty Job (Grim Reaper #1);
  - Moneyball: The Art of Winning an Unfair Game
  - School's Out—Forever (Maximum Ride #2)
- анти топ-3 (min средний рейтинг + min кол-во обзоров):
  - Harvesting the Heart;
  - Junky;
  - His Excellency: George Washington

3.3 Задача 3

Формулировка задачи:

- определить издательство, которое выпустило наибольшее число книг толще 50 страниц (так можно исключить из анализа брошюры).

In [15]:
```sql
# выполним SQL-запрос в Pandas и выведем результат на экран
query8 = '''
WITH book_publisher AS (
    SELECT
        b.book_id,
        p.publisher
    FROM books AS b
    LEFT JOIN publishers AS p USING(publisher_id)
    WHERE b.num_pages>50
    ),

    publisher_count_books AS (
    SELECT
    publisher,
    COUNT(book_id) AS count_books
    FROM book_publisher
    GROUP BY 1)

SELECT
    publisher,
    count_books
FROM publisher_count_books
WHERE count_books = (SELECT MAX(count_books) FROM publisher_count_books)
'''
pd.io.sql.read_sql(query8, con = engine)
```

Out[15]:

	publisher	count_books
0	Penguin Books	42

Вывод:

- наибольшее число книг толще 50 страниц выпустило издательство Penguin Books - 42 книги.

3.4 Задача 4

Формулировка задачи:

- определить автора с самой высокой средней оценкой книг (учитываем только книги с 50 и более оценками).

In [16]:
```
query9 = '''
WITH df1 AS (
             SELECT
                   b.book_id,
                   a.author,
                   r.rating
             FROM books AS b
             LEFT JOIN authors AS a USING(author_id)
             LEFT JOIN ratings AS r USING(book_id)
             WHERE b.book_id IN (SELECT
                                       book_id
                                 FROM ratings
                                 GROUP BY book_id
                                 HAVING COUNT(rating)>=50)
                                 )
SELECT author,
       AVG(rating)
FROM df1
GROUP BY 1
ORDER BY 2 DESC
LIMIT 1
'''
pd.io.sql.read_sql(query9, con = engine)
```

Out[16]:

	author	avg
0	J.K. Rowling/Mary GrandPré	4.287097

Вывод:

- автор с самой высокой средней оценкой книг (среди книг с 50 и более оценками) - J.K. Rowling/Mary GrandPré, ее книги имеют среднюю оценку 4.287097.

3.5 Задача 5

Формулировка задачи:

- посчитаем среднее количество обзоров от пользователей, которые поставили больше 48 оценок.

In [17]:
```
# выполним SQL-запрос в Pandas и выведем результат на экран
query10 = '''
WITH df AS (
              SELECT
                    username,
                    COUNT(review_id) AS count_review
              FROM reviews
              WHERE username IN (SELECT
                                       username
                                 FROM
                                       ratings
                                 GROUP BY 1
                                 HAVING COUNT(rating)>48
                                  )
              GROUP BY 1
)

SELECT AVG(count_review)
FROM df
'''
pd.io.sql.read_sql(query10, con = engine)
```

Out[17]:

	avg
0	24.0

Вывод:

- среднее количество обзоров от пользователей, которые поставили больше 48 оценок, равно 24.

4 Общий вывод

К началу

- в базе сервиса для чтения книг по подписке - 1000 книг;
- в базе представлены 636 авторов, 340 издательств;
- пользователями сервиса выставлено 6456 оценок и написано 2793 обзоров;
- 819 книг были выпущены в свет после 01.01.2000 г.;
- максимальный средний рейтинг у книг - 5, минимальный - 1.5;
- книги имеют максимальное кол-во обзоров - 8. минимальное - 0;
- издательство - лидер по количеству выпущенных книг Penguin Books, им выпущено 42 книги;
- автор с самой высокой средней оценкой книг J.K. Rowling/Mary GrandPré;
- пользователи, которые активно ставят оценки, также активно пишут обзоры.

Рекомендация:

- в базе сервиса мало книг, ее нужно активно пополнять;
- 636 авторов, а книг всего 1000, думаю можно найти популярных у аудитории авторов и дополнить базу другими их произведениями;
- пользователям интересно не только читать сами книги, но и читать и писать обзоры, поэтому можно разработать бонусную программу для активных в плане написания обзоров клиентов.