Dokumentasi Book Analytics Dashboard

Nama : Hanif Aditia Sofian

Brief explanation of methodology

1. Data Extraction & Cleaning

- Membaca books.json dan mengubah dalam bentuk pandas dataframe.
- Menghapus missing value (title atau author).
- standarisasi string formatting untuk title, author, dan genre (mengubah menjadi string, format penulisan judul, menghapus ekstra spasi.)
- Menghapus buku duplikat.

2. Data Simulation

- Membuat 10 sample user yang melakukan interaksi random (view, reading, complete).
- Setiap interaksi sesuai timeline (action “complete” terjadi setelah action “reading”)
- Membuat reading progress data (pages_read, total_pages, completion_rate)

3. Data Transformation

- Menghitung average completion rate untuk setiap user.
- mengestimasi reading speed dari awal action “reading” sampai action “complete”.
- Segments users menjadi 3 kategori:
    - Bookworm Reader → rata-rata completion rate ≥ 80%
    - Moderate Reader → rata-rata completion rate 40–79%
    - Casual Reader → rata-rata completion rate < 40%

4. Aggregation

Membuat user summary yang berisi:

- Average completion rate dan reading segment
- Total pages read

- Total books completed
- Average reading speed
- Most-read genre

5. Data Model (SQL)

- Terdapat 4 tabel (books, user_interactions, reading_progress, user_segments.)
- Hubungan antar tabel :
    - books – user_interactions = One-to-many (buku dapat memiliki banyak interaksi user)
    - books – reading_progress = One-to-many (buku dapat muncul di banyak user reading progress)

6. Visualization

- Menggunakan python dengan library plotly.
- Di buat di python notebook.
- Dashboard di buat dengan library streamlit

Key findings and recommendations

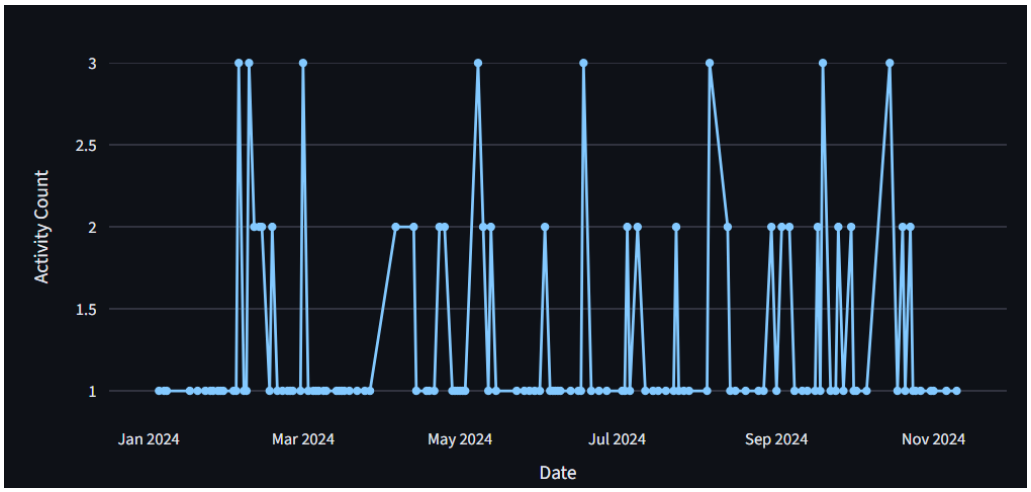
- Genre buku terbanyak adalah fantasy dengan total 2 buku.

- Jumlah completion rate tertinggi ada pada kisaran 0.85-0.94.

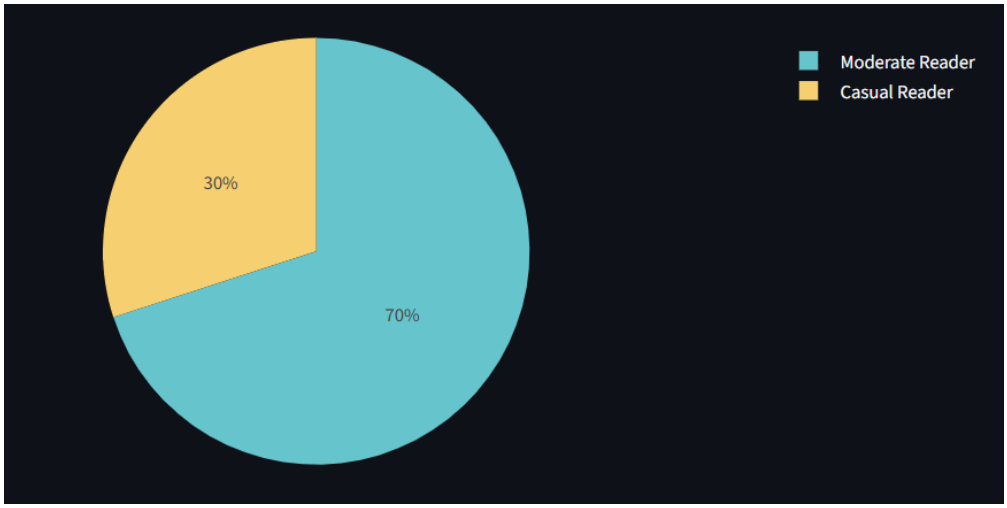
- Buku dengan rating tertinggi adalah The Lord Of The Rings dengan nilai rating 4.9.

title	author	rating
The Lord Of The Rings	J.R.R. Tolkien	4.9
To Kill A Mockingbird	Harper Lee	4.8
Harry Potter And The Sorcerer'S Stone	J.K. Rowling	4.7
1984	George Orwell	4.6
Pride And Prejudice	Jane Austen	4.4
Dune	Frank Herbert	4.3
The Great Gatsby	F. Scott Fitzgerald	4.2
The Catcher In The Rye	J.D. Salinger	3.8

- User interaction tertinggi adalah 3 (Karena data merupakan beberapa sample yang dibuat secara random, User interaction tidak menunjukkan jumlah interaksi yang tinggi pada tiap harinya)



- 70% user dikategorikan sebagai casual reader (rata-rata completion rate < 40%) dan 30% user dikategorikan sebagai moderate reader (rata-rata completion rate 40–79%). Tidak ada user yang termasuk dalam kategori bookworm reader (rata-rata completion rate ≥ 80%)



- Recommendation: Menggunakan data interaksi user yang sesuai dengan real world scenario sehingga analisis akan lebih aktual dan akurat, analisis lain seperti genre paling popular, buku paling banyak dilihat, buku dengan pembaca terbanyak, dll.

Assumptions and limitation

- Data di hasilkan secara random sehingga tidak merepresentasikan real world scenario.
- Reading speed bisa nan ketika tidak ada action “complete” (sehingga tidak ada end date) dan Ketika action “reading” dan action “complete” berada pada hari yang sama (hal ini karena perhitungan reading speed adalah per hari sehingga jika 0 maka reading speed akan Nan/inf.)
- Hanya terdapat 3 interaksi user yaitu view, reading, dan complete.
- Reading speed menggunakan dihitung dengan cara banyakna lembar yang dibaca setiap hari (pages/day).
- Setiap Data Preparation & ETL (part1.py) di run, maka data akan berubah karena generate random data.