BIÊN BẢN HỌP LẦN 3
LẬP TRÌNH KHOA HỌC DỮ LIỆU

1. **Thời gian:** từ 9g00 đến 11g00 ngày 12/12/2024.
2. **Hình thức:** trực tuyến qua Google Meet.
3. **Nội dung :**
    a. **Xem lại các phần trả lời câu hỏi.**
    b. **Bổ sung các phần trình bày.**
    c. **Báo cáo kết quả.**
4. **Kết quả:**
    a. **Xem lại các phần trả lời câu hỏi:** các câu hỏi còn vấn đề, nên xem xét và sửa đổi lại.
    b. **Bổ sung các phần trình bày:**
        i. Phải có comment trong mỗi cell code.
        ii. Phải có Table of Content trong mỗi file notebook.
    c. **Nội dung báo cáo cần bổ sung:**
        i. Nguồn dữ liệu.
        ii. Khó khăn, cách giải quyết và kiến thức học được:

- 22120003:
    - Khó khăn và cách giải quyết:
        - Một phần lớn dữ liệu có Nhà xuất bản là "Không rõ", gây khó khăn trong việc phân loại và phân tích cụ thể. → Phân tích các xu hướng chung trên dữ liệu có thông tin để đại diện cho nhóm có nhãn "Không rõ".
        - Khó khăn trong việc xác định ngưỡng đánh giá (ví dụ như số lượng đánh giá tối thiểu) đòi hỏi sự cân nhắc kỹ lưỡng giữa dữ liệu thực tế và các tiêu chuẩn thống kê. → Sử dụng phân phối thống kê thực tế để xác định ngưỡng hợp lý.
        - Dữ liệu phân phối sách trên nhiều thể loại khác nhau khiến việc xác định thế mạnh của từng Nhà xuất bản phức tạp. → Nhóm dữ liệu theo manufacturer và category, sau đó tìm thể loại nổi bật nhất của từng Nhà xuất bản.
    - Kiến thức học được:
        - Kết quả rút ra được không nhất thiết phải trực quan ở dạng biểu đồ, mà có thể in ra dưới dạng DataFrame.
        - Cải thiện kỹ năng phân tích dữ liệu.
        - Dựa vào các chỉ số thống kê để chọn ra ngưỡng phù hợp với dữ liệu.

        - Chất lượng dữ liệu là yếu tố then chốt để phân tích nên cần chuẩn bị và làm sạch kỹ lưỡng trước khi phân tích.

- 22120008:
    - Khó khăn và cách giải quyết
        - Dữ liệu có 1 outlier vượt trội dẫn đến khó khăn trong việc phân tích dữ liệu → Loại bỏ ngoại lệ để phân tích.
        - Khi thể hiện biểu đồ phân vân việc lựa chọn top để thể hiện → Thống nhất 1 số lượng top và chỉ phân tích theo số đó.
    - Kiến thức học được:
        - Học được thêm các tham số khác trong các hàm vẽ biểu đồ.
        - Biết cách thêm giá trị của các dữ liệu lên trên biểu đồ.
        - Kỹ năng phân tích biểu đồ được cải thiện.

- 22120009:
    - Khó khăn và cách giải quyết:
        - Ban đầu, các phân tích còn rời rạc chưa có định hướng rõ ràng → tập trung vào mục tiêu, xác định thực hiện phân tích theo góc nhìn của Tiki: tối ưu hóa doanh thu.
        - Số lượng tác giả quá lớn → tìm tiêu chí (doanh thu, số lượng bán ra) để chọn ra các tác giả xu hướng và phân tích các tác giả này.
        - Các biểu đồ chưa hiển thị như mong muốn → tìm kiếm thông tin trong các thư viện để điều chỉnh cho phù hợp.
        - Khi tính doanh thu, con số này khá lớn, khi vẽ lên biểu đồ không thể hiện hết → chuẩn hóa về đơn vị tỷ VNĐ.
    - Kiến thức học được:
        - Cách tùy chỉnh các tham số trong biểu đồ.
        - Cách sử dụng jupyter Notebook TOC.
        - Lựa chọn màu sắc của biểu đồ phù hợp.
        - Kỹ năng phân tích được cải thiện.
        - Cách tổ chức công việc trên git.
        - Kỹ năng tiền xử lý dữ liệu

d. Nếu có thêm thời gian, nhóm sẽ làm gì?

- Tinh chỉnh thêm mô hình.
- Phân tích thêm nhiều khía cạnh.
- Tổ chức thành 1 dashboard.
- Thu thập thêm dữ liệu.