# Guidelist

1. **Business Understanding**:  
   * **Mục tiêu**: Xây dựng hệ thống phân cụm khách hàng dựa trên hành vi mua sắm và các sản phẩm họ quan tâm. Từ đó, giúp cửa hàng X phát triển các chiến lược marketing hiệu quả, chăm sóc khách hàng, và tăng doanh thu.
2. **Data Understanding**:  
   * **Transactions.csv**: Chứa thông tin giao dịch của khách hàng (Member\_number, Date, productId, items).
   * **Products\_with\_Categories.csv**: Chứa thông tin về sản phẩm (productId, productName, price, Category).  
     .
3. **Tiền xử lý Dữ liệu**:  
   * **Làm sạch dữ liệu**: Kiểm tra và xử lý giá trị thiếu, các giao dịch không hợp lệ, loại bỏ dữ liệu trùng lặp hoặc ngoại lệ.
   * **Tạo các đặc trưng phù hợp**:  
     + Tính tổng chi tiêu của khách hàng (theo productId, items, và price).
     + Tính tần suất mua hàng (số lần khách hàng tham gia giao dịch).
     + Tính giá trị trung bình của mỗi đơn hàng (dựa trên số lượng sản phẩm và giá trị giao dịch).
   * **Kết nối dữ liệu**: Kết hợp dữ liệu giao dịch và sản phẩm qua productId để có thông tin đầy đủ về sản phẩm khách hàng mua.
4. **Khám phá và Phân tích Dữ liệu**:

**Phân tích hành vi mua hàng**:

* **Thời gian mua sắm**: Chúng ta phân tích **ngày trong tuần** và **tháng** mà khách hàng thực hiện giao dịch bằng cách trích xuất các thông tin từ cột **Date**.
* **Tỷ lệ khách hàng tái mua**: Tính toán tỷ lệ khách hàng quay lại mua hàng sau lần mua đầu tiên. Khoảng cách thời gian giữa các lần mua sẽ giúp bạn xác định nhóm khách hàng tái mua.

**Phân tích các sản phẩm bán chạy và lợi nhuận**:

* **Sản phẩm bán chạy nhất**: Tính tổng số lượng sản phẩm bán được và hiển thị các sản phẩm bán chạy nhất.
* **Sản phẩm có doanh thu cao nhất**: Tính doanh thu cho mỗi sản phẩm và tìm các sản phẩm có doanh thu cao nhất.

**Phân tích dòng tiền**:

* **Doanh thu theo tháng**: Phân tích doanh thu theo từng tháng trong năm để tìm các xu hướng theo mùa.
* **Sản phẩm bán chạy nhất trong từng tháng**: Xác định sản phẩm bán chạy nhất trong mỗi tháng.

1. **Xây dựng Mô hình Phân đoạn Khách hàng**:  
   * **Sử dụng mô hình phân cụm**:  
     + **RFM (Recency, Frequency, Monetary)**: Sử dụng các đặc trưng như tần suất giao dịch (Frequency), tổng chi tiêu (Monetary), và thời gian gần nhất mua sắm (Recency), AOV (Giá trị trung bình của mỗi đơn hàng)
     + **Thuật toán phân cụm**:  
       - K-means (sử dụng scikit-learn).
       - Hierarchical Clustering.
       - Sử dụng PySpark nếu dữ liệu quá lớn.
   * **Xác định các nhóm khách hàng tiềm năng**: Chia khách hàng thành các nhóm có hành vi tương tự (ví dụ: nhóm khách hàng trung thành, nhóm khách hàng chi tiêu cao, nhóm khách hàng ít tương tác).
2. **Đánh giá Kết quả Phân đoạn**:  
   * **Phân tích các nhóm khách hàng**: Xem xét các đặc điểm nhóm khách hàng (Ví dụ: khách hàng thường xuyên mua sản phẩm tươi sống, khách hàng chỉ mua sữa và bánh mì).
   * **Kiểm tra tính hợp lý**: Đánh giá mức độ hợp lý của các phân đoạn từ góc độ chiến lược marketing và khả năng áp dụng thực tế.
3. **Chiến lược Marketing và Quảng cáo**:  
   * **Nhắm mục tiêu chính xác**:  
     + Dựa trên các nhóm khách hàng, xây dựng các chiến lược marketing đặc thù (Ví dụ: gửi các chương trình khuyến mãi cho nhóm khách hàng mua thường xuyên, cung cấp ưu đãi cho khách hàng chi tiêu cao).
   * **Tùy chỉnh sản phẩm**:  
     + Dựa trên hành vi mua hàng, cung cấp các sản phẩm phù hợp với nhu cầu của từng nhóm khách hàng (Ví dụ: nhóm khách hàng thích sản phẩm tươi sống có thể được giới thiệu thêm các sản phẩm trái cây hoặc thực phẩm tươi).