

UBND THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC SÀI GÒN



Báo Cáo
THUẬT TOÁN PHÂN CỤM
Khoa Toán - Ứng Dụng

Họ Và Tên	MSSV
Lê Phước Thành	3123580045
Nguyễn Hoàng Long	3123580022
Đường Minh Đức	3123580010
Hà Tuấn Duy	3123580006

*Giảng Viên Hướng Dẫn: **Đỗ Như Tài***

THÀNH PHỐ HỒ CHÍ MINH

Mục lục

Bảng phân công.....	2
BÁO CÁO PHÂN CỤM DỮ LIỆU CHIM CÁNH CỤT	3
1. Giới Thiệu	3
2. Dữ Liệu	3
3. Quy Trình Phân Tích.....	3
3.1. Tiền Xử Lý Dữ Liệu.....	3
3.2. Giảm Chiều Dữ Liệu với PCA.....	4
3.3. Phân Cụm K-means	4
4. Kết Quả	5
4.1. Kết Quả PCA	5
4.2. Kết Quả Phân Cụm	5
5. Ưu Điểm và Hạn Chế.....	5
Ưu điểm:	5
Hạn chế:	5
6. Kết Luận.....	5
7. Hướng Phát Triển.....	6
BÁO CÁO PHÂN CỤM DỮ LIỆU BÁN HÀNG TRỰC TUYẾN.....	6
1. Giới Thiệu	6
2. Dữ Liệu	6
3. Quy Trình Phân Tích.....	6
3.1. Tiền Xử Lý Dữ Liệu.....	6
3.2. Xác Định Số Cụm Tối Ưu - Phương Pháp Elbow	7
3.3. Thực Hiện Phân Cụm K-means.....	8
4. Kết Quả	9
4.1. Phân Đoạn Khách Hàng.....	9
4.2. Biểu Đồ Phân Tán (Scatter Plot).....	9
5. Ý Nghĩa Kinh Doanh	10
5.1. Ứng Dụng Thực Tế	10
5.2. KPI và Đo Lường.....	10
6. Ưu Điểm và Hạn Chế.....	11

Ưu điểm:	11
Hạn chế:	11
7. Kết Luận.....	11
8. Hướng Phát Triển.....	11
8.1. Cải Tiến Mô Hình	11
8.2. Ứng Dụng Thực Tế	12
PHÂN TÍCH PHÂN CỤM ĐA CẤP DỮ LIỆU PENGUINS	13
1. Tổng quan Dự án.....	13
2. Tiền xử lý dữ liệu (Data Cleaning & Preprocessing).....	13
3. Mô hình hóa (Clustering Modeling)	13
3.1. Agglomerative Clustering	14
3.2. K-Means Clustering	14
4. Đánh giá Hiệu suất Mô hình	14
5. So sánh với Phân tích Online Retail (Lưu ý kỹ thuật).....	15

Bảng phân công

Thành viên	Công việc
Lê Phước Thành (Nhóm trưởng)	Phân cụm K-means_tập dữ liệu mua sắm, báo cáo word, báo cáo
Nguyễn Hoàng Long	Phân cụm K-means_tập dữ liệu chim cánh cụt, báo cáo word
Đường Minh Đức	Phân cụm đa cấp_tập dữ liệu mua sắm, báo cáo word
Hà Tuấn Duy	Phân cụm đa cấp_tập dữ liệu chim cánh cụt, báo cáo word, tổng hợp bài, báo cáo word

BÁO CÁO PHÂN CỤM DỮ LIỆU CHIM CÁNH CỤT

1. Giới Thiệu

Báo cáo này trình bày quy trình phân tích và phân cụm dữ liệu chim cánh cụt (penguins) sử dụng thuật toán K-means kết hợp với phương pháp giảm chiều dữ liệu PCA (Principal Component Analysis). Mục tiêu là nhóm các chim cánh cụt thành các cụm dựa trên đặc điểm sinh học của chúng.

2. Dữ Liệu

Dữ liệu được sử dụng là bộ dữ liệu `penguins.csv` chứa thông tin về các đặc điểm của chim cánh cụt, có thể bao gồm các thuộc tính như chiều dài vây, khối lượng cơ thể, giới tính và loài.

3. Quy Trình Phân Tích

3.1. Tiền Xử Lý Dữ Liệu

Khám phá dữ liệu ban đầu:

- Sử dụng `head()` để xem 5 dòng đầu tiên
- Vẽ biểu đồ boxplot để phát hiện các giá trị ngoại lai

Xử lý dữ liệu thiếu:

- Loại bỏ các dòng có giá trị null bằng `dropna()`

Phát hiện và xử lý ngoại lai:

- Kiểm tra các giá trị bất thường: chiều dài vây (`flipper_length_mm`) $> 4000\text{mm}$ hoặc $< 0\text{mm}$
- Phát hiện và loại bỏ 2 dòng dữ liệu ngoại lai tại index 9 và 14

Mã hóa dữ liệu:

- Sử dụng `pd.get_dummies()` để chuyển đổi biến phân loại (categorical) thành dạng số
- Loại bỏ cột `"sex_"` (có thể là giá trị giới tính thiếu)

Chuẩn hóa dữ liệu:

- Áp dụng StandardScaler để chuẩn hóa tất cả các đặc trưng về cùng một thang đo
- Điều này quan trọng để đảm bảo các thuộc tính có đơn vị khác nhau không ảnh hưởng không cân đối đến kết quả phân cụm

3.2. Giảm Chiều Dữ Liệu với PCA

Phân tích thành phần chính:

- Thực hiện PCA với tất cả các thành phần để xác định mức độ giải thích phương sai
- Phân tích `explained_variance_ratio_` để đánh giá tầm quan trọng của từng thành phần

Lựa chọn số chiều:

- Chọn các thành phần có tỷ lệ phương sai giải thích $> 10\%$ (0.1)
- Số thành phần được chọn tự động dựa trên ngưỡng này
- Kết quả cho thấy số thành phần tối ưu được lưu trong biến `n_components`

Biến đổi dữ liệu:

- Áp dụng PCA với số thành phần đã chọn để giảm chiều dữ liệu
- Tạo tập dữ liệu mới `penguins_PCA` với số chiều ít hơn nhưng vẫn giữ được hầu hết thông tin

3.3. Phân Cụm K-means

Xác định số cụm tối ưu - Phương pháp Elbow:

- Thử nghiệm với số cụm k từ 1 đến 9
- Tính toán độ trơ (inertia) cho mỗi giá trị k
- Inertia đo tổng khoảng cách bình phương từ các điểm đến tâm cụm gần nhất
- Vẽ biểu đồ Elbow để xác định điểm "khuỷu tay" - nơi độ giảm inertia bắt đầu chậm lại
- Dựa vào biểu đồ, lựa chọn $k = 4$ cụm

Thực hiện phân cụm:

- Áp dụng thuật toán K-means với $k = 4$
- Sử dụng `random_state=42` để đảm bảo kết quả có thể tái tạo
- Gán nhãn cụm cho từng điểm dữ liệu

Trực quan hóa kết quả:

- Vẽ scatter plot với 2 thành phần chính đầu tiên làm trục tọa độ
- Mỗi điểm được tô màu theo cụm được gán
- Sử dụng colormap "viridis" để phân biệt các cụm

4. Kết Quả

4.1. Kết Quả PCA

Sau khi áp dụng PCA, số chiều dữ liệu được giảm xuống còn `n_components` thành phần chính, giúp loại bỏ nhiễu và giảm độ phức tạp tính toán trong khi vẫn giữ được trên 90% thông tin gốc.

4.2. Kết Quả Phân Cụm

Dữ liệu chim cánh cụt được phân thành 4 cụm riêng biệt. Biểu đồ scatter plot cho thấy:

- Sự phân tách rõ ràng giữa các cụm trong không gian 2 chiều của PCA
- Mỗi cụm có thể đại diện cho một nhóm chim cánh cụt có đặc điểm tương đồng
- Các cụm này có thể tương ứng với các loài khác nhau hoặc các nhóm theo giới tính và đặc điểm sinh học

5. Ưu Điểm và Hạn Chế

Ưu điểm:

- Quy trình xử lý dữ liệu chi tiết và khoa học
- Sử dụng PCA giúp giảm chiều và loại bỏ nhiễu
- Phương pháp Elbow giúp xác định số cụm hợp lý
- Kết quả có thể tái tạo nhờ `random_state`

Hạn chế:

- K-means giả định các cụm có hình dạng cầu và kích thước tương đương
- Thuật toán nhạy cảm với khởi tạo ban đầu
- Việc chọn $k = 4$ dựa trên phương pháp Elbow có thể mang tính chủ quan
- Chỉ trực quan hóa trên 2 thành phần chính đầu tiên, có thể bỏ sót thông tin từ các chiều khác

6. Kết Luận

Thuật toán K-means kết hợp với PCA đã được áp dụng thành công để phân cụm dữ liệu chim cánh cụt thành 4 nhóm riêng biệt. Quy trình phân tích bao gồm các bước tiền xử lý dữ liệu cẩn thận, giảm chiều với PCA và phân cụm với K-means. Kết quả cho thấy các cụm được phân tách tương đối rõ ràng, phản ánh sự khác biệt về đặc điểm sinh học giữa các nhóm chim cánh cụt.

7. Hướng Phát Triển

Để cải thiện phân tích, có thể:

- Thử nghiệm với các thuật toán phân cụm khác như DBSCAN, Hierarchical Clustering
- Sử dụng các chỉ số đánh giá như Silhouette Score, Davies-Bouldin Index
- Phân tích sâu hơn về ý nghĩa của từng cụm bằng cách kiểm tra đặc điểm trung bình

So sánh kết quả phân cụm với nhãn thực tế (nếu có) để đánh giá độ chính xác

BÁO CÁO PHÂN CỤM DỮ LIỆU BÁN HÀNG TRỰC TUYẾN

1. Giới Thiệu

Báo cáo này trình bày quy trình phân tích và phân cụm dữ liệu bán hàng trực tuyến (Online Retail) sử dụng thuật toán K-means. Mục tiêu là phân đoạn khách hàng dựa trên hành vi mua sắm, cụ thể là số lượng sản phẩm mua và giá đơn vị, để hỗ trợ các quyết định kinh doanh và marketing.

2. Dữ Liệu

Dữ liệu được sử dụng là bộ dữ liệu OnlineRetail.csv chứa thông tin về các giao dịch bán hàng trực tuyến, bao gồm:

- **InvoiceNo:** Mã hóa đơn
- **StockCode:** Mã sản phẩm
- **Description:** Mô tả sản phẩm
- **Quantity:** Số lượng sản phẩm mua
- **InvoiceDate:** Ngày giao dịch
- **UnitPrice:** Giá đơn vị sản phẩm
- **CustomerID:** Mã khách hàng
- **Country:** Quốc gia

3. Quy Trình Phân Tích

3.1. Tiền Xử Lý Dữ Liệu

Đọc và khám phá dữ liệu:

- Đọc file CSV với encoding 'ISO-8859-1' để xử lý đúng các ký tự đặc biệt và ngôn ngữ
- Hiển thị DataFrame để kiểm tra cấu trúc dữ liệu

Phát hiện và xử lý dữ liệu thiếu:

- Sử dụng `isnull().sum()` để đếm số lượng giá trị null trong từng cột
- Phát hiện có giá trị thiếu trong các cột `Description` và `CustomerID`
- Điền giá trị 0 cho các trường thiếu:
 - `Description`: Điền 0 cho mô tả sản phẩm bị thiếu
 - `CustomerID`: Điền 0 cho mã khách hàng bị thiếu
- Kiểm tra lại bằng `isnull().sum()` để xác nhận không còn giá trị null

Phân tích khám phá dữ liệu (EDA):

- Sử dụng `value_counts()` để xem phân bố các giá trị
- Tính tổng số lượng sản phẩm theo từng khách hàng bằng `groupby('CustomerID')['Quantity'].sum()`
- Giúp hiểu được quy mô mua sắm của từng khách hàng

Lựa chọn đặc trưng cho phân cụm:

- Chọn 2 đặc trưng chính: `Quantity` (số lượng) và `UnitPrice` (giá đơn vị)
- Tạo DataFrame mới x chỉ chứa 2 cột này
- Loại bỏ các dòng có giá trị thiếu bằng `dropna()`

Chuẩn hóa dữ liệu:

- Áp dụng `StandardScaler` để chuẩn hóa dữ liệu
- Chuyển đổi `Quantity` và `UnitPrice` về cùng thang đo (`mean=0, std=1`)
- Tạo tập dữ liệu `x_scaled` đã được chuẩn hóa
- Điều này quan trọng vì `Quantity` và `UnitPrice` có đơn vị và phạm vi giá trị khác nhau

3.2. Xác Định Số Cụm Tối Ưu - Phương Pháp Elbow

Thiết lập phạm vi thử nghiệm:

- Thử nghiệm với số cụm k từ 1 đến 10
- Tạo danh sách `k_range` để lặp qua các giá trị k

Tính toán Inertia:

- Với mỗi giá trị k:
 - Khởi tạo mô hình K-means với `n_clusters=k`

- Sử dụng phương pháp khởi tạo `k-means++` để cải thiện tốc độ hội tụ và chất lượng
- Đặt `random_state=42` để đảm bảo kết quả có thể tái tạo
- Fit mô hình trên dữ liệu đã chuẩn hóa `x_scaled`
- Lưu giá trị inertia (tổng bình phương khoảng cách từ các điểm đến tâm cụm gần nhất)

Trực quan hóa và lựa chọn k:

- Vẽ biểu đồ đường với:
 - Trục x: Số cụm k (từ 1 đến 10)
 - Trục y: Giá trị inertia
 - Marker 'o' để đánh dấu các điểm dữ liệu
- Quan sát biểu đồ để tìm điểm "khuỷu tay" (elbow point)
- Điểm này là nơi độ giảm inertia bắt đầu chậm lại đáng kể
- Dựa trên biểu đồ Elbow, quyết định chọn `k = 4` cụm

3.3. Thực Hiện Phân Cụm K-means

Khởi tạo và huấn luyện mô hình:

- Khởi tạo K-means với các tham số:
 - `n_clusters=4`: Số cụm đã chọn
 - `random_state=42`: Đảm bảo tính nhất quán
 - `init='k-means++'`: Phương pháp khởi tạo thông minh
- Fit mô hình trên dữ liệu đã chuẩn hóa `x_scaled`

Gán nhãn cụm:

- Lấy nhãn cụm từ `kmeans.labels_`
- Thêm cột mới `cluster` vào DataFrame gốc df
- Mỗi giao dịch được gán vào 1 trong 4 cụm (0, 1, 2, 3)
- Hiển thị 5 dòng đầu với `head()` để kiểm tra kết quả

Trực quan hóa kết quả:

- Tạo scatter plot với kích thước 8x6 inch
- Các thông số:
 - Trục x: Quantity (số lượng sản phẩm)
 - Trục y: UnitPrice (giá đơn vị)
 - Màu sắc: Cluster (cụm được gán)
 - Colormap: 'viridis' (gradient màu xanh-vàng)
 - Alpha: 0.6 (độ trong suốt 60%)

- Thêm colorbar để chú thích màu sắc cho từng cụm
- Thêm tiêu đề và nhãn trục rõ ràng

4. Kết Quả

4.1. Phân Đoạn Khách Hàng

Dữ liệu được phân thành 4 cụm dựa trên số lượng mua và giá sản phẩm:

Cụm 0 - Khách hàng thông thường:

- Đặc điểm: Mua số lượng trung bình với giá trung bình
- Hành vi: Mua sắm ổn định, không quá nhiều cũng không quá ít
- Chiếm tỷ lệ: Phần lớn giao dịch
- Giá trị: Đây là nhóm khách hàng cốt lõi, đem lại doanh thu ổn định

Cụm 1 - Khách hàng mua sỉ/số lượng lớn:

- Đặc điểm: Mua số lượng rất lớn
- Hành vi: Đặt hàng với khối lượng nhiều, có thể để bán lại
- Giá trị: Doanh thu cao mỗi đơn hàng
- Phân khúc: Có thể là đại lý, nhà bán lẻ hoặc doanh nghiệp

Cụm 2 - Khách hàng cao cấp:

- Đặc điểm: Mua sản phẩm giá cao
- Hành vi: Quan tâm đến chất lượng, sẵn sàng trả giá cao
- Số lượng: Có thể ít hoặc trung bình
- Giá trị: Margin cao, cần chăm sóc đặc biệt

Cụm 3 - Khách hàng mua lẻ nhỏ:

- Đặc điểm: Mua số lượng ít, giá thấp đến trung bình
- Hành vi: Mua thử, mua lẻ, không thường xuyên
- Giá trị: Doanh thu thấp mỗi giao dịch
- Tiềm năng: Có thể chuyển đổi thành khách hàng thường xuyên

4.2. Biểu Đồ Phân Tán (Scatter Plot)

Biểu đồ cho thấy:

- Sự phân tách rõ ràng giữa các cụm trong không gian 2 chiều
- Các điểm cùng màu (cùng cụm) có xu hướng tập trung gần nhau
- Có sự chồng lấn nhỏ giữa một số cụm, phản ánh ranh giới mờ giữa các phân khúc khách hàng
- Colorbar giúp dễ dàng nhận biết từng cụm

5. Ý Nghĩa Kinh Doanh

5.1. Ứng Dụng Thực Tế

Marketing và Quảng cáo:

- Tạo chiến dịch marketing riêng biệt cho từng phân khúc
- Cụm 0: Chương trình khách hàng thân thiết, ưu đãi định kỳ
- Cụm 1: Chính sách giá sỉ, hỗ trợ logistics
- Cụm 2: Marketing sản phẩm cao cấp, dịch vụ VIP
- Cụm 3: Ưu đãi lần mua đầu, khuyến khích tăng giá trị đơn hàng

Chiến lược giá:

- Điều chỉnh giá và khuyến mãi phù hợp với từng nhóm
- Chương trình giảm giá theo số lượng cho khách mua sỉ
- Giữ giá ổn định cho khách cao cấp, tập trung vào giá trị

Quản lý tồn kho:

- Dự đoán nhu cầu dựa trên hành vi mua của từng cụm
- Chuẩn bị hàng số lượng lớn cho cụm mua sỉ
- Đa dạng hóa sản phẩm cao cấp cho cụm 2

Chăm sóc khách hàng:

- Ưu tiên nguồn lực cho các cụm có giá trị cao (cụm 1, 2)
- Phát triển kênh liên lạc phù hợp cho từng nhóm
- Tạo trải nghiệm mua sắm cá nhân hóa

Phát triển sản phẩm:

- Thiết kế sản phẩm/gói dịch vụ theo nhu cầu từng phân khúc
- Bundle sản phẩm cho khách mua số lượng lớn
- Sản phẩm premium cho khách cao cấp

5.2. KPI và Đo Lường

Theo dõi các chỉ số cho từng cụm:

- Giá trị đơn hàng trung bình (Average Order Value - AOV)
- Tần suất mua hàng (Purchase Frequency)
- Tỷ lệ giữ chân khách hàng (Retention Rate)
- Giá trị vòng đời khách hàng (Customer Lifetime Value - CLV)

6. Ưu Điểm và Hạn Chế

Ưu điểm:

- Đơn giản, dễ hiểu và dễ triển khai
- Kết quả dễ giải thích cho nhà quản lý kinh doanh
- Sử dụng k-means++ cho khởi tạo tốt hơn, hội tụ nhanh hơn
- Phân cụm trực tiếp trên 2 đặc trưng quan trọng nhất
- Có thể áp dụng ngay vào thực tế kinh doanh

Hạn chế:

- Chỉ sử dụng 2 đặc trưng, có thể bỏ sót thông tin từ các thuộc tính khác
- Không xử lý tương quan giữa các biến
- K-means giả định các cụm có hình dạng cầu và kích thước tương tự
- Nhạy cảm với outliers (giá trị ngoại lai)
- Việc điền giá trị 0 cho CustomerID thiếu có thể ảnh hưởng đến kết quả
- Số cụm k=4 được chọn có tính chủ quan dựa trên biểu đồ Elbow

7. Kết Luận

Thuật toán K-means đã được áp dụng thành công để phân đoạn khách hàng từ dữ liệu bán hàng trực tuyến thành 4 nhóm riêng biệt dựa trên số lượng mua và giá sản phẩm. Quy trình phân tích tuân thủ các bước chuẩn của khai phá dữ liệu: tiền xử lý, chuẩn hóa, xác định số cụm tối ưu, và trực quan hóa kết quả.

Kết quả phân cụm cung cấp cái nhìn có giá trị về các phân khúc khách hàng khác nhau, giúp doanh nghiệp:

- Hiểu rõ hơn về hành vi mua sắm của khách hàng
- Tối ưu hóa chiến lược marketing và bán hàng
- Cá nhân hóa trải nghiệm khách hàng
- Tăng hiệu quả kinh doanh và lợi nhuận

8. Hướng Phát Triển

8.1. Cải Tiến Mô Hình

Thêm đặc trưng:

- Tích hợp thêm các đặc trưng như:
 - Tần suất mua hàng (Frequency)
 - Thời gian từ lần mua gần nhất (Recency)
 - Tổng giá trị đơn hàng (Monetary)
 - Thực hiện phân tích RFM (Recency, Frequency, Monetary) hoàn chỉnh

- Thêm thông tin về sản phẩm, ngày trong tuần, mùa vụ

Thử nghiệm thuật toán khác:

- DBSCAN: Xử lý tốt với outliers và cụm không đồng nhất
- Hierarchical Clustering: Tạo cây phân cấp các cụm
- Gaussian Mixture Models: Cho phép cụm có hình dạng ellipse

Đánh giá chất lượng:

- Silhouette Score: Đo độ gắn kết và phân tách của cụm
- Davies-Bouldin Index: Đánh giá độ tương đồng giữa các cụm
- Calinski-Harabasz Score: Tỷ lệ phương sai giữa cụm và trong cụm

8.2. Ứng Dụng Thực Tế

Hệ thống tự động:

- Xây dựng pipeline tự động phân loại khách hàng mới
- Cập nhật mô hình định kỳ với dữ liệu mới
- Tích hợp vào hệ thống CRM

Theo dõi động:

- Theo dõi sự thay đổi cụm của khách hàng theo thời gian
- Phát hiện khách hàng có nguy cơ rời bỏ (churn)
- Đánh giá hiệu quả của các chiến dịch marketing

A/B Testing:

- Thử nghiệm các chiến lược khác nhau cho từng cụm
- Đo lường ROI của từng phân khúc khách hàng
- Tối ưu hóa ngân sách marketing

PHÂN TÍCH PHÂN CỤM ĐA CẤP DỮ LIỆU PENGUINS

1. Tổng quan Dự án

Mục tiêu của notebook Penguins.ipynb là áp dụng các kỹ thuật phân cụm không giám sát để nhóm các loài chim cánh cụt dựa trên các đặc điểm hình thái học.

- **Dữ liệu đầu vào:** Tập dữ liệu penguins.csv chứa thông tin về 344 chú chim cánh cụt.
- **Thư viện sử dụng:** Pandas, Matplotlib, Scikit-learn (PCA, KMeans, StandardScaler, silhouette_score, AgglomerativeClustering).

2. Tiền xử lý dữ liệu (Data Cleaning & Preprocessing)

Trước khi đưa vào mô hình, dữ liệu đã được xử lý như sau:

- **Tải và kiểm tra dữ liệu:** Dữ liệu có 5 cột: `culmen_length_mm`, `culmen_depth_mm`, `flipper_length_mm`, `body_mass_g`, và `sex`. Có một số giá trị thiếu (null).
- **Xử lý giá trị thiếu:** Loại bỏ các dòng chứa giá trị thiếu bằng `dropna()`.
- **Lựa chọn đặc trưng:** Chỉ giữ lại các cột số (`culmen_length_mm`, `culmen_depth_mm`, `flipper_length_mm`, `body_mass_g`) để phân cụm.
- **Chuẩn hóa dữ liệu (Scaling):** Sử dụng `StandardScaler` để đưa các đặc trưng về cùng một thang đo (`mean=0`, `std=1`), giúp thuật toán phân cụm hoạt động hiệu quả hơn.

3. Mô hình hóa (Clustering Modeling)

Dữ liệu sau khi chuẩn hóa được phân cụm bằng hai thuật toán:

3.1. Agglomerative Clustering

- **Cấu hình:** Sử dụng liên kết trung bình (`linkage='average'`) và số cụm `n_clusters=3`.
- **Trực quan hóa:** Biểu đồ phân tán (Scatter plot) cho thấy sự phân bố của 3 cụm dựa trên hai đặc trưng đầu tiên (`culmen_length_mm` và `culmen_depth_mm`).

3.2. K-Means Clustering

- **Cấu hình:** Số cụm `n_clusters=3`, `random_state=42`.
- **Trực quan hóa:** Tương tự, biểu đồ phân tán hiển thị kết quả phân cụm K-Means.

4. Đánh giá Hiệu suất Mô hình

Sử dụng chỉ số **Silhouette Score** để đánh giá chất lượng phân cụm:

Mô hình	Silhouette Score
K-Means Clustering	~0.4963
Agglomerative Clustering	~0.4950

Nhận xét:

- Cả hai thuật toán đều cho kết quả tương đương nhau với Silhouette Score xấp xỉ 0.5.
- Điểm số này cho thấy các cụm có độ tách biệt trung bình, không quá rõ rệt nhưng vẫn chấp nhận được.
- K-Means có điểm số nhỉnh hơn một chút so với Agglomerative Clustering.

5. So sánh với Phân tích Online Retail (Lưu ý kỹ thuật)

So với notebook trước (`OnlineRetail.ipynb`), notebook này có một số điểm cải thiện và khác biệt:

- 1. Chuẩn hóa dữ liệu:** Notebook này **đã thực hiện** bước chuẩn hóa dữ liệu (`StandardScaler`), khắc phục được vấn đề chênh lệch thang đo giữa các biến (ví dụ: khối lượng cơ thể hàng nghìn gram so với độ dài mỏ chỉ vài chục mm). Đây là một bước quan trọng giúp kết quả phân cụm chính xác hơn về mặt toán học.
- 2. Xử lý dữ liệu thiếu:** Tương tự, notebook này cũng loại bỏ các giá trị thiếu trước khi xử lý.
- 3. Trực quan hóa:** Các biểu đồ đã được gán nhãn trục và tiêu đề chính xác hơn, phản ánh đúng dữ liệu chim cánh cụt.
- 4. Kết quả Silhouette:** Mặc dù điểm số thấp hơn so với notebook trước (~ 0.5 so với > 0.9), nhưng kết quả này **đáng tin cậy hơn** vì dữ liệu đã được chuẩn hóa. Điểm số quá cao trong notebook trước có thể là do sự chi phối của biến `Monetary` chưa được chuẩn hóa.

PHÂN TÍCH PHÂN KHÚC KHÁCH HÀNG (ONLINE RETAIL)

1. Tổng quan Dự án

Mục tiêu của notebook là thực hiện phân tích dữ liệu giao dịch bán lẻ trực tuyến (Online Retail) để phân khúc khách hàng. Quy trình bao gồm làm sạch dữ liệu, trích xuất đặc trưng hành vi tiêu dùng (RFM) và áp dụng các thuật toán học máy (Clustering) để gom nhóm khách hàng.

- Dữ liệu đầu vào: Tập dữ liệu OnlineRetail.csv chứa 541,909 dòng dữ liệu giao dịch.
- Thư viện sử dụng: Pandas, NumPy, Matplotlib, Seaborn, Missingno, Scikit-learn.

2. Tiền xử lý dữ liệu (Data Cleaning)

Trước khi phân tích, dữ liệu thô đã trải qua các bước làm sạch sau:

- Kiểm tra dữ liệu thiếu: Ban đầu xác định cột CustomerID có lượng dữ liệu thiếu lớn.
- Xử lý dữ liệu thiếu:
 - Số lượng dòng thiếu CustomerID: 135,080 dòng.
 - Hành động: Loại bỏ toàn bộ các dòng thiếu CustomerID để đảm bảo tính chính xác khi định danh khách hàng (Sử dụng `df.dropna()`).
- Định dạng dữ liệu: Chuyển đổi cột InvoiceDate từ dạng chuỗi sang dạng thời gian (datetime) để phục vụ tính toán độ gần của giao dịch.

3. Kỹ thuật Đặc trưng (Feature Engineering) - Mô hình RFM

Dữ liệu giao dịch được chuyển đổi thành dữ liệu cấp độ khách hàng dựa trên mô hình RFM (Recency, Frequency, Monetary).

Chỉ số	Định nghĩa trong Notebook	Phương pháp tính
Monetary (M)	Tổng số tiền chi tiêu	Quantity * UnitPrice, sau đó tính tổng (sum) theo CustomerID.

Frequency (F)	Tần suất mua hàng	Đếm số lượng mã hóa đơn (InvoiceNo) theo CustomerID.
Recency (R)	Thời gian kể từ lần mua cuối	Lấy ngày giao dịch gần nhất trong tập dữ liệu trừ đi ngày giao dịch cuối cùng của từng khách hàng.

Kết quả là một DataFrame mới (df_new) chứa thông tin tổng hợp cho từng khách hàng duy nhất.

4. Mô hình hóa (Clustering Modeling)

Dữ liệu sau khi xử lý được đưa vào hai thuật toán phân cụm không giám sát để tìm ra các nhóm khách hàng tương đồng.

4.1. Thuật toán sử dụng

- K-Means Clustering: Phân chia dữ liệu thành \$K\$ cụm (trong notebook chọn \$K=3\$).
- Agglomerative Clustering: Phân cụm phân cấp (Hierarchical Clustering) với liên kết trung bình (linkage='average'), cũng chọn \$3\$ cụm.

4.2. Trực quan hóa

Notebook đã hiển thị biểu đồ phân tán (Scatter plot) cho kết quả của cả hai mô hình để quan sát sự phân tách giữa các nhóm.

5. Đánh giá Hiệu suất Mô hình

Sử dụng chỉ số **Silhouette Score** để đánh giá độ tách biệt giữa các cụm. Giá trị càng gần 1 thì phân cụm càng tốt.

Kết quả từ Notebook:

- **K-Means Clustering:** ~0.9245
- **Agglomerative Clustering:** ~0.9727

Kết luận: Dựa trên chỉ số Silhouette, mô hình **Agglomerative Clustering** cho hiệu quả phân chia cụm tốt hơn một chút so với K-Means trên tập dữ liệu này.

6. Hướng Phát triển và Cải thiện (Future Work)

Dựa trên kết quả hiện tại và mã nguồn đã triển khai, các hướng phát triển tiếp theo được đề xuất nhằm nâng cao độ chính xác của mô hình và tính ứng dụng thực tiễn:

6.1. Ứng dụng Kinh doanh và Phân tích Mở rộng

- **Định danh cụm (Cluster Profiling):** Sau khi phân cụm, cần tính toán giá trị trung bình RFM của từng nhóm để gán nhãn cụm (Ví dụ: "Khách hàng VIP", "Khách hàng nguy cơ rời bỏ", "Khách hàng mới tiềm năng").
- **Phân tích theo thời gian (Cohort Analysis):** Theo dõi sự dịch chuyển của khách hàng giữa các cụm theo thời gian để đánh giá hiệu quả của các chiến dịch marketing giữ chân khách hàng.