**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**HỌC MÁY VÀ ỨNG DỤNG**

**TÊN ĐỀ TÀI: PHÂN TÍCH HÀNH VI MUA SẮM KHÁCH HÀNG**

**Giảng viên hướng dẫn: CHÂU GIA HÂN**

**Sinh viên thực hiện: TRẦN ĐỨC ĐẠI**

**MSSV: 2200003458**

**Khoá: 2022**

**Ngành/ chuyên ngành: KHOA HỌC DỮ LIỆU**

TP. HCM, tháng 09 năm 2025

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**HỌC MÁY VÀ ỨNG DỤNG**

**TÊN ĐỀ TÀI: PHÂN TÍCH HÀNH VI MUA SẮM KHÁCH HÀNG**

**Giảng viên hướng dẫn: CHÂU GIA HÂN**

**Sinh viên thực hiện: TRẦN ĐỨC ĐẠI**

**MSSV: 2200003458**

**Khoá: 2022**

**Ngành/ chuyên ngành: KHOA HỌC DỮ LIỆU**

TP. HCM, tháng 09 năm 2025

# LỜI CẢM ƠN

Trước tiên, em xin bày tỏ lòng biết ơn sâu sắc đến cha mẹ và anh chị em trong gia đình đã luôn động viên tinh thần, chia sẻ và hỗ trợ em vượt qua những khó khăn trong suốt quá trình học tập và nghiên cứu.

Em xin chân thành cảm ơn các quý thầy cô giáo trường Đại học Nguyễn Tất Thành, đặc biệt là Cô Châu Gia Hân, người đã tận tình hướng dẫn, cung cấp những ý kiến chuyên môn quý báu và tài liệu tham khảo cần thiết, giúp em hoàn thành đồ án môn học này. Sự động viên và nhắc nhở từ cô đã giúp em giữ vững tinh thần và tiến độ trong quá trình thực hiện đồ án.

Cảm ơn anh chị, bạn bè và đồng nghiệp đã luôn sẵn sàng hỗ trợ, chia sẻ kiến thức và kinh nghiệm quý báu. Sự giúp đỡ của mọi người là nguồn động lực lớn lao giúp em vượt qua mọi thử thách trong quá trình học tập và nghiên cứu tại trường.

Em xin chân thành cảm ơn tất cả.

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **KHOA CÔNG NGHỆ THÔNG TIN**  🙜 🙜 🙝 | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập – Tự do – Hạnh phúc**  🙜 🙜 🙝 |

# NHIỆM VỤ ĐỒ ÁN MÔN HỌC

*(Sinh viên phải đóng tờ này vào cuốn báo cáo)*

Họ và tên: **Trần Đức Đại**  MSSV: **2200003458**

Chuyên ngành: **Khoa học dữ liệu** Lớp: **22DTH4B**

Email: **[tranducdai14756@gmail.com](mailto:ndt222h@gmail.com)** SĐT: **0986932816**

Tên đề tài: **Phân tích hành vi mua sắm khách hàng**

Giảng viên giảng dạy: **Châu Gia Hân**

Thời gian thực hiện: **24/08/2025** **đến 29/09/2025**

Nhiệm vụ/nội dung (mô tả chi tiết nội dung, yêu cầu, phương pháp… ):

**MÔ TẢ ĐỀ TÀI:**

Đề tài xây dựng mô hình phân tích và phân khúc khách hàng dựa trên dữ liệu mua sắm, thu nhập và điểm chi tiêu. Mô hình sử dụng thuật toán K-Means để nhóm các khách hàng có hành vi tương đồng, nhằm hỗ trợ doanh nghiệp xây dựng chiến lược marketing hiệu quả và cá nhân hóa trải nghiệm khách hàng.

**NỘI DUNG VÀ PHƯƠNG PHÁP:**

* Trình bày tổng quan về phân tích hành vi khách hàng và ý nghĩa của việc phân khúc thị trường.
* Tiền xử lý dữ liệu: chuẩn hóa, làm sạch và xử lý dữ liệu thiếu nếu có.
* Xác định số cụm tối ưu bằng phương pháp Elbow Method.
* Áp dụng thuật toán K-Means để phân cụm khách hàng.
* Trực quan hóa và diễn giải kết quả phân cụm để rút ra các insight kinh doanh.
* Kết luận, đánh giá hiệu quả của thuật toán và đề xuất các hướng phát triển trong tương lai (sử dụng các thuật toán khác hoặc thêm dữ liệu).

**YÊU CẦU:**

* Nắm kiến thức cơ bản về Machine Learning và các thuật toán phân cụm (clustering).
* Lập trình Python với các thư viện như scikit-learn, Pandas, Matplotlib và Seaborn.
* Có khả năng nghiên cứu, đọc hiểu tài liệu chuyên ngành.

**Nội dung và yêu cầu đã được thông qua Bộ môn.**

*Tp.HCM, ngày 24 tháng 09 năm 2025*

|  |  |
| --- | --- |
| **TRƯỞNG BỘ MÔN**  *(Ký và ghi rõ họ tên)* | **GIẢNG VIÊN GIẢNG DẠY**  *(Ký và ghi rõ họ tên)* |

# LỜI MỞ ĐẦU

Trong thời đại bùng nổ dữ liệu và công nghệ thông tin hiện nay, việc áp dụng các phương pháp phân tích dữ liệu hiện đại đã trở thành một lợi thế cạnh tranh then chốt đối với mọi doanh nghiệp. Dữ liệu không còn đơn thuần là những con số vô tri, mà đã trở thành một tài sản có giá trị cao, ẩn chứa nhiều thông tin quan trọng về hành vi cũng như nhu cầu của khách hàng. Phân tích dữ liệu giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình, từ đó đưa ra các chiến lược kinh doanh và marketing một cách hiệu quả hơn.

Đặc biệt trong ngành bán lẻ, việc thấu hiểu khách hàng đóng vai trò sống còn đối với sự phát triển và duy trì vị thế trên thị trường. Hành vi tiêu dùng ngày càng trở nên đa dạng và phức tạp, đòi hỏi các doanh nghiệp phải có cái nhìn sâu rộng để cá nhân hóa trải nghiệm khách hàng và tối ưu hóa doanh thu. Việc phân khúc khách hàng không chỉ giúp nhận diện rõ ràng các nhóm mục tiêu mà còn hỗ trợ doanh nghiệp phân bổ nguồn lực một cách hợp lý, từ đó nâng cao hiệu quả hoạt động.

Xuất phát từ tính ứng dụng thực tiễn đó, em đã lựa chọn thực hiện đồ án môn học "Học máy và ứng dụng" với đề tài "**Phân tích hành vi mua sắm khách hàng bằng thuật toán K-Means**". Đồ án này tập trung vào việc ứng dụng K-Means – một trong những thuật toán phân cụm phổ biến thuộc học máy không giám sát – để phân tích tập dữ liệu khách hàng và xác định các phân khúc có ý nghĩa. Kết quả đạt được không chỉ giúp em củng cố kiến thức chuyên môn về phân tích dữ liệu mà còn xây dựng một mô hình thực tế, có thể hỗ trợ các doanh nghiệp bán lẻ nâng cao hiệu quả kinh doanh và phát triển mối quan hệ khách hàng bền vững.

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ III NĂM HỌC 2024 – 2025** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: Lớp học phần: 22DTH4B

Nhóm sinh viên thực hiện :

1. Trần Đức Đại Tham gia đóng góp: 33%

2. Hồ Quốc Thái Tham gia đóng góp: 33%

3. Nguyễn Duy Thái Tham gia đóng góp: 33%

4. Tham gia đóng góp:

5. Tham gia đóng góp:

6. Tham gia đóng góp:

7. Tham gia đóng góp:

8. Tham gia đóng góp:

Ngày thi: 29/09/2025 Phòng thi:L605

Đề tài tiểu luận/báo cáo của sinh viên : **Phân tích hành vi mua sắm khách hàng**.

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  | 1 |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  | 6 |  |
| * Lập luận |  | 2 |  |
| * Kết luận |  | 0.5 |  |
| Trình bày |  | 0.5 |  |
| **TỔNG ĐIỂM** |  | **10** |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

# MỤC LỤC

[LỜI CẢM ƠN ii](#_Toc206811343)

[NHIỆM VỤ ĐỒ ÁN MÔN HỌC iii](#_Toc206811344)

[LỜI MỞ ĐẦU iv](#_Toc206811345)

[MỤC LỤC vi](#_Toc206811346)

[DANH MỤC HÌNH ix](#_Toc206811347)

[CHƯƠNG 1:TỔNG QUAN ĐỀ TÀI 1](#_Toc206811348)

[I. Giới thiệu 1](#_Toc206811349)

[1) Hành vi khách hàng là gì? 1](#_Toc206811350)

[2) Các loại hành vi khách hàng 1](#_Toc206811351)

[a) Hành vi khách hàng theo thói quen (Habitual Buying Behavior) 1](#_Toc206811352)

[b) Hành vi mua hàng phức tạp (Complex Buying Behavior) 2](#_Toc206811353)

[c) Hành vi mua hàng tìm kiếm sự đa dạng (Variety Seeking Behavior) 2](#_Toc206811354)

[d) Hành vi mua hàng giảm bất hòa (Dissonance-reducing Buying Behavior) 3](#_Toc206811355)

[e) Hành vi Window Shopping 3](#_Toc206811356)

[II. Mục tiêu nghiên cứu 4](#_Toc206811357)

[1) Hoàn thiện chân dung khách hàng 4](#_Toc206811358)

[2) Đưa ra dự đoán cho các chiến lược dài hạn 4](#_Toc206811359)

[3) Thúc đẩy doanh số nhờ trải nghiệm cá nhân hóa 5](#_Toc206811360)

[4) Cải thiện chất lượng sản phẩm, dịch vụ 5](#_Toc206811361)

[5) Tăng vị thế cạnh tranh 5](#_Toc206811362)

[III. Phương pháp nghiên cứu 6](#_Toc206811363)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 9](#_Toc206811364)

[I. Phân tích dữ liệu 9](#_Toc206811365)

[1) Phân tích dữ liệu là gì? 9](#_Toc206811366)

[2) Vai trò của phân tích dữ liệu trong doanh nghiệp 10](#_Toc206811367)

[3) Tại sao doanh nghiệp cần phân tích dữ liệu kinh doanh? 11](#_Toc206811368)

[a) Hỗ trợ ra quyết định dựa trên dữ liệu 11](#_Toc206811369)

[b) Tối ưu hóa chiến lược tài chính kinh doanh 11](#_Toc206811370)

[c) Cải thiện hiệu suất hoạt động và dự báo xu hướng 11](#_Toc206811371)

[4) Các phương pháp phân tích dữ liệu kinh doanh phổ biến 12](#_Toc206811372)

[a) Phân tích mô tả (Descriptive Analytics) 12](#_Toc206811373)

[b) Phân tích chẩn đoán (Diagnostic Analytics) 12](#_Toc206811374)

[c) Phân tích dự báo (Predictive Analytics) 12](#_Toc206811375)

[d) Phân tích đề xuất (Prescriptive Analytics) 13](#_Toc206811376)

[5) Các loại dữ liệu quan trọng trong phân tích kinh doanh 13](#_Toc206811377)

[a) Dữ liệu tài chính 13](#_Toc206811378)

[b) Dữ liệu khách hàng 14](#_Toc206811379)

[c) Dữ liệu vận hành sản xuất 15](#_Toc206811380)

[d) Dữ liệu tiếp thị bán hàng 16](#_Toc206811381)

[6) Hướng dẫn quy trình thực hiện phân tích dữ liệu kinh doanh chi tiết 17](#_Toc206811382)

[a) Bước 1. Xác định mục tiêu phân tích dữ liệu 17](#_Toc206811383)

[b) Bước 2. Thu thập dữ liệu kinh doanh 17](#_Toc206811384)

[c) Bước 3. Làm sạch và chuẩn hóa dữ liệu 18](#_Toc206811385)

[d) Bước 4. Trực quan hóa dữ liệu 19](#_Toc206811386)

[e) Bước 5. Xây dựng mô hình phân tích dự báo 20](#_Toc206811387)

[f) Bước 6. Đưa ra kết luận hành động 20](#_Toc206811388)

[g) Bước 7. Giám sát cải tiến liên tục 20](#_Toc206811389)

[II. Phân cụm (Clustering) 21](#_Toc206811390)

[1. Khái Niệm và Bản Chất 21](#_Toc206811391)

[2. Các Thuật Toán Phân Cụm Phổ Biến 21](#_Toc206811392)

[3. Tại Sao Phân Cụm Lại Hữu Ích Trong Phân Tích Khách Hàng? 22](#_Toc206811393)

[a) Cá nhân hóa Trải nghiệm Khách hàng (Personalization) 22](#_Toc206811394)

[b) Tối ưu hóa Chiến lược Marketing 23](#_Toc206811395)

[c) Phát triển Sản phẩm và Dịch vụ 23](#_Toc206811396)

[d) Quản lý Quan hệ Khách hàng (CRM) 23](#_Toc206811397)

[4. Hạn Chế và Thách Thức của Phân Cụm 24](#_Toc206811398)

[III. Thuật toán K-Means 24](#_Toc206811399)

[a) Khởi tạo các trung tâm cụm (Centroids) 25](#_Toc206811400)

[b) Gán mỗi điểm dữ liệu vào cụm gần nhất 25](#_Toc206811401)

[c) Cập nhật các trung tâm cụm 26](#_Toc206811402)

[d) Lặp lại quá trình 26](#_Toc206811403)

[e) Đánh giá chất lượng phân cụm 26](#_Toc206811404)

[a) Ưu điểm của K-Means Clustering 27](#_Toc206811405)

[b) Nhược điểm của K-Means Clustering 27](#_Toc206811406)

[c) Khi nào nên sử dụng K-Means Clustering? 28](#_Toc206811407)

[CHƯƠNG 3: THỰC NGHIỆM VÀ KẾT QUẢ 29](#_Toc206811408)

[I. Mô tả tập dữ liệu 29](#_Toc206811409)

[1. Nguồn gốc và chi tiết dữ liệu 29](#_Toc206811410)

[2. Khám phá Dữ Liệu 31](#_Toc206811411)

[a) Thống kê mô tả: 31](#_Toc206811412)

[b) Phân tích mối quan hệ giữa các biến: 31](#_Toc206811413)

[II. Tiền xử lý dữ liệu 32](#_Toc206811414)

[1) Tải và Đọc Dữ Liệu 32](#_Toc206811415)

[2) Kiểm tra Dữ Liệu Bị Thiếu (Missing Values) 33](#_Toc206811416)

[3) Kiểm tra Định Dạng Dữ Liệu (Data Types) 33](#_Toc206811417)

[4) Lựa chọn Biến Phân Tích 34](#_Toc206811418)

[5) Kết Luận Về Giai Đoạn Tiền Xử Lý 34](#_Toc206811419)

[III. Phân tích khám phá dữ liệu (EDA) 34](#_Toc206811420)

[1) Phân tích Phân bố Giới tính 34](#_Toc206811421)

[2) Phân tích Phân bố Tuổi (Age) 35](#_Toc206811422)

[3) Phân tích Phân bố Thu nhập hàng năm (Annual Income) 36](#_Toc206811423)

[4) Phân tích Phân bố Điểm chi tiêu (Spending Score) 36](#_Toc206811424)

[5) Phân tích Mối quan hệ giữa các biến 37](#_Toc206811425)

[6) Phân tích Mối quan hệ giữa Tuổi, Thu nhập và Điểm chi tiêu 37](#_Toc206811426)

[7) Phân bố giá trị theo Độ tuổi, Thu nhập hàng năm và Điểm chi tiêu theo Giới tính 38](#_Toc206811427)

[IV. Áp dụng thuật toán K-Means 39](#_Toc206811428)

[1) Phân đoạn sử dụng Độ tuổi và Điểm chi tiêu 39](#_Toc206811429)

[a) Chọn N cụm dựa trên quán tính 39](#_Toc206811430)

[b) Xây dựng Mô hình K-Means với Số Cụm Tối ưu 40](#_Toc206811431)

[c) Trực quan hóa Kết quả Phân cụm 40](#_Toc206811432)

[2) Phân khúc sử dụng Điểm thu nhập và chi tiêu hàng năm 41](#_Toc206811433)

[3) Phân khúc theo Độ tuổi, Thu nhập hàng năm và Điểm chi tiêu 42](#_Toc206811434)

[V. Kết quả phân cụm và trực quan hóa 43](#_Toc206811435)

[1) Lựa chọn tính năng cho mô hình 43](#_Toc206811436)

[2) Xây dựng mô hình 43](#_Toc206811437)

[3) Diễn giải mô hình 44](#_Toc206811438)

[CHƯƠNG 4: KẾT LUẬN VÀ ĐÁNH GIÁ 45](#_Toc206811439)

[I. Kết luận 45](#_Toc206811440)

[II. Đánh giá hiệu quả của thuật toán K-Means 45](#_Toc206811441)

[III. Hạn chế và Hướng phát triển trong tương lai 46](#_Toc206811442)

[TÀI LIỆU THAM KHẢO 48](#_Toc206811443)

# 

# DANH MỤC HÌNH

[Hình 1. Ảnh minh họa hành vi mua sắm của khách hàng 1](#_Toc206811444)

[Hình 2. Người dùng thường mua theo thói quen và so sánh với sản phẩm giá cao 4](#_Toc206811445)

[Hình 3. Thuật toán phân cụm K-Means 6](#_Toc206811446)

[Hình 4. Biểu đồ Elbow Method 8](#_Toc206811447)

[Hình 5. Data analytics hiện đang là xu hướng trong môi trường kinh doanh hiện đại 9](#_Toc206811448)

[Hình 6. Phân tích dữ liệu là cơ sở cho doanh nghiệp đưa ra quyết định chiến lược 10](#_Toc206811449)

[Hình 7. Phân tích dữ liệu doanh nghiệp cải thiện trải nghiệm khách hàng 11](#_Toc206811450)

[Hình 8. Các phương pháp phân tích dữ liệu 12](#_Toc206811451)

[Hình 9. Các loại dữ liệu quan trọng trong kinh doanh 13](#_Toc206811452)

[Hình 10. ROI - Các dữ liệu tiếp thị giúp doanh nghiệp tối ưu chiến dịch quảng cáo 16](#_Toc206811453)

[Hình 11. K-Means Clustering 24](#_Toc206811454)

[Hình 12. Dữ liệu đọc được từ dataset 32](#_Toc206811455)

[Hình 13. Kiểm tra dữ liệu bị thiếu 33](#_Toc206811456)

[Hình 14. Kiểm tra định dạng dữ liệu 33](#_Toc206811457)

[Hình 15. Biểu đồ phân bố giới tính 35](#_Toc206811458)

[Hình 16. Biểu đồ phân bố tuổi 35](#_Toc206811459)

[Hình 17. Biểu đồ thu nhập hàng năm 36](#_Toc206811460)

[Hình 18. Biểu đồ phân bố Điểm chi tiêu hàng năm 36](#_Toc206811461)

[Hình 19. Biểu đồ phân tán Annual Income và Spending Score. 37](#_Toc206811462)

[Hình 20. biểu đồ mối quan hệ giữa độ tuổi, thu nhập hàng năm và điểm chi tiêu 37](#_Toc206811463)

[Hình 21. Biểu đồ phân bố giá trị Độ tuổi, Thu nhập và Điểm chi tiêu theo Giới tính 38](#_Toc206811464)

[Hình 22. biểu đồ thể hiện WCSS so với số cụm (k) dựa vào Tuổi và Điểm chi tiêu 39](#_Toc206811465)

[Hình 23. Biểu đồ phân tán thể hiện 5 cụm khách hàng dựa vào tuổi và điểm chi tiêu 40](#_Toc206811466)

[Hình 24. Biểu đồ thể hiện WCSS dựa vào Thu nhập và Điểm chi tiêu 41](#_Toc206811467)

[Hình 25. Biểu đồ phân tán 5 cụm khách hàng dựa vào thu nhập và điểm chi tiêu 41](#_Toc206811468)

[Hình 26.biểu đồ thể hiện WCSS so với số cụm (k) dựa vào tất cả biến 42](#_Toc206811469)

[Hình 27. Biểu đồ phân tán thể hiện 5 cụm khách hàng dựa vào tất cả các biến 42](#_Toc206811470)

[Hình 28. Lọc chọn ra 2 cột Thu nhập và Điểm chi tiêu 43](#_Toc206811471)

[Hình 29. Phương pháp ELBOW để có được giá trị K tối ưu 43](#_Toc206811472)

[Hình 30. Xác định các nhóm khách hàng 44](#_Toc206811473)

# CHƯƠNG 1:TỔNG QUAN ĐỀ TÀI

## Giới thiệu

### Hành vi khách hàng là gì?

Hành vi khách hàng (Consumer Behaviour) là tổng hợp các phản ứng, hành vi và suy nghĩ của người tiêu dùng trong suốt quá trình mua sắm và sử dụng sản phẩm hoặc dịch vụ của doanh nghiệp. Quá trình này bao gồm các bước từ khi phát sinh nhu cầu, tìm kiếm thông tin, so sánh, đánh giá, đưa ra quyết định mua cho đến những hành động sau khi mua hàng.

Việc thấu hiểu hành vi khách hàng giúp doanh nghiệp có cơ sở để nâng cao chất lượng sản phẩm, điều chỉnh chiến lược kinh doanh hoặc chiến lược tiếp thị nhằm thu hút người tiêu dùng hiệu quả hơn. Qua đó, doanh nghiệp có thể gia tăng tỷ lệ chuyển đổi và cải thiện doanh thu một cách đáng kể..



Hình 1. Ảnh minh họa hành vi mua sắm của khách hàng

Hành vi khách hàng luôn phức tạp, linh hoạt và được ảnh hưởng bởi nhiều yếu tố bên trong lẫn bên ngoài như tâm lý, văn hóa, xã hội, gia đình,... Do đó, các doanh nghiệp cần nghiên cứu thật kỹ để nắm bắt được sở thích, nhu cầu và thói quen của họ để xây dựng chiến lược phù hợp, thúc đẩy nhu cầu mua sắm và sử dụng sản phẩm/dịch vụ.

### Các loại hành vi khách hàng

#### ****Hành vi khách hàng theo thói quen (Habitual Buying Behavior)****

Đây là kiểu hành vi mà người tiêu dùng ít khi cân nhắc hay phân tích nhiều trong quá trình ra quyết định mua sắm. Họ thường không để ý nhiều đến sự khác biệt giữa các thương hiệu và có xu hướng mua lặp lại cùng một sản phẩm/dịch vụ. Điều này phổ biến với các mặt hàng thiết yếu hằng ngày như muối, đường, bột ngọt, dầu ăn, bột giặt, bánh mì, nước ngọt,…

Khác với lòng trung thành thương hiệu, người tiêu dùng mua theo thói quen vẫn sẵn sàng chọn sản phẩm thay thế nếu thương hiệu quen thuộc không có tại điểm bán hoặc xuất hiện lựa chọn khác với mức giá thấp hơn. Vì vậy, để thu hút nhóm khách hàng này, Marketer có thể triển khai quảng cáo mang tính lặp lại nhằm tạo cảm giác quen thuộc, đồng thời áp dụng các chương trình khuyến mãi hay giảm giá để kích thích họ dùng thử sản phẩm.

Hành vi mua theo thói quen vừa là một thách thức, vừa mở ra cơ hội cho doanh nghiệp. Thách thức ở chỗ rất khó thay đổi thói quen đã hình thành và thuyết phục người tiêu dùng chuyển sang sản phẩm mới. Nhưng ngược lại, đây cũng là cơ hội để doanh nghiệp dần xây dựng sự gắn bó lâu dài và lòng trung thành từ phía khách hàng.

#### ****Hành vi mua hàng phức tạp (Complex Buying Behavior)****

Đây là dạng hành vi thường xuất hiện khi khách hàng lựa chọn những sản phẩm có giá trị lớn, đòi hỏi phải bỏ ra nhiều thời gian và công sức để cân nhắc trước khi quyết định. Những mặt hàng như ô tô, nhà ở hay các thiết bị công nghệ hiện đại thường thuộc nhóm này. Người tiêu dùng sẽ tìm hiểu kỹ lưỡng, so sánh nhiều lựa chọn khác nhau và tham khảo ý kiến từ nhiều nguồn trước khi mua.

Đối với hành vi mua hàng phức tạp, người bán cần nắm rõ nhu cầu và tâm lý khách hàng để đưa ra tư vấn đầy đủ, chính xác về lợi ích sản phẩm. Song song đó, việc xây dựng các thông điệp quảng cáo phù hợp nhằm tác động đến niềm tin và thái độ của khách hàng cũng đóng vai trò rất quan trọng.

#### ****Hành vi mua hàng tìm kiếm sự đa dạng (Variety Seeking Behavior)****

Hành vi này thường xảy ra khi khách hàng muốn trải nghiệm những sản phẩm mới hoặc thay đổi so với sản phẩm quen thuộc mà họ đang sử dụng. Việc có nhiều lựa chọn khác nhau mang lại sự hứng thú, đồng thời thỏa mãn nhu cầu tìm kiếm sự mới mẻ. Chẳng hạn, người tiêu dùng có thể thử nhiều thương hiệu thực phẩm, thay đổi loại mỹ phẩm hay nước hoa để tìm ra sản phẩm phù hợp nhất.

Theo các nghiên cứu, hành vi mua hàng tìm kiếm sự đa dạng bắt nguồn từ việc sử dụng lặp đi lặp lại cùng một sản phẩm, khiến giá trị thụ hưởng cận biên (Marginal Utility) giảm dần và tạo cảm giác nhàm chán cho người tiêu dùng.

Để đáp ứng nhu cầu này và nâng cao mức độ hài lòng, doanh nghiệp cần xác định rõ phân khúc khách hàng, xây dựng chân dung khách hàng chính xác, đồng thời tạo sự khác biệt và đa dạng trong sản phẩm/dịch vụ. Việc thường xuyên đổi mới, cập nhật danh mục sản phẩm, khuyến khích khách hàng trải nghiệm cũng như chia sẻ cảm nhận sẽ giúp doanh nghiệp duy trì sự quan tâm. Các hoạt động Marketing nên nhấn mạnh yếu tố độc đáo, mới lạ để thu hút người tiêu dùng.

#### ****Hành vi mua hàng giảm bất hòa (Dissonance-reducing Buying Behavior)****

Đây là kiểu hành vi mua hàng khi người tiêu dùng tham gia nhiều vào quá trình ra quyết định, nhưng lại khó phân biệt rõ ràng giữa các thương hiệu hay lựa chọn trên thị trường để thỏa mãn nhu cầu. Hành vi này thường đi kèm với sự lo lắng hoặc cảm giác tiếc nuối sau khi đã mua sản phẩm.

Để giảm bớt sự bất an, khách hàng thường tìm kiếm các thông tin tích cực nhằm củng cố niềm tin rằng lựa chọn của mình là đúng đắn, đồng thời có xu hướng bỏ qua hoặc ít chú ý đến những thông tin bất lợi liên quan đến thương hiệu đã chọn.

Hành vi mua hàng giảm bất hòa phổ biến ở những ngành hàng như điện thoại thông minh hay thiết bị điện tử, nơi các sản phẩm thường có mức giá tương đương và tính năng gần giống nhau. Chính điều này khiến người tiêu dùng cảm thấy khó khăn trong việc đưa ra quyết định và lo ngại liệu lựa chọn thương hiệu cụ thể có dẫn đến sự không hài lòng sau khi sử dụng hay không.

#### ****Hành vi Window Shopping****

Window Shopping là một hành vi quen thuộc trong tâm lý tiêu dùng, khi nhiều người có thói quen dạo quanh các cửa hàng hoặc lướt xem sản phẩm trên nền tảng trực tuyến chỉ để ngắm, so sánh hay trải nghiệm, chứ chưa có ý định mua ngay. Đây thường được xem là bước khởi đầu trong hành trình mua sắm, khi người tiêu dùng chủ yếu tìm kiếm thông tin và đánh giá sức hấp dẫn của sản phẩm.

Mặc dù chưa mang lại doanh thu trực tiếp, nhưng hành vi này giúp gia tăng nhận diện thương hiệu, tạo cảm xúc tích cực và ảnh hưởng đến quyết định mua trong tương lai. Thực tế, không ít khách hàng quay lại mua khi xuất hiện ưu đãi, phát sinh nhu cầu hoặc đơn giản vì ấn tượng tốt từ lần xem trước.

Trong bối cảnh số hóa, việc "lướt xem" hàng trăm sản phẩm chỉ với vài thao tác chạm màn hình trở nên rất dễ dàng. Các nền tảng thương mại điện tử, mạng xã hội và ứng dụng bán hàng ngày càng thúc đẩy mạnh mẽ hành vi này thông qua nội dung sáng tạo và thuật toán cá nhân hóa. Với nhiều người, đặc biệt là giới trẻ, Window Shopping không chỉ để tham khảo sản phẩm mà còn là một hình thức giải trí và cách thể hiện phong cách sống hiện đại.



Hình 2. Người dùng thường mua theo thói quen và so sánh với sản phẩm giá cao

## Mục tiêu nghiên cứu

### Hoàn thiện chân dung khách hàng

Kết hợp với các yếu tố nhân khẩu học, việc nghiên cứu hành vi khách hàng giúp các doanh nghiệp hoàn thiện chân dung khách hàng một cách rõ ràng và toàn diện. Để qua đó, doanh nghiệp có thể hiểu rõ những mong đợi của khách hàng và lên chiến lược kinh doanh phù hợp nhằm thu hút, giữ chân khách hàng và tạo ra tỷ lệ chuyển đổi tốt nhất.

### Đưa ra dự đoán cho các chiến lược dài hạn

Bằng cách sử dụng các dữ liệu về hành vi khách hàng trong quá khứ, các doanh nghiệp có thể hiểu sâu hơn về nhận thức, tâm lý và suy nghĩ của khách hàng đối với sản phẩm và thương hiệu. Các nghiên cứu về hành vi khách hàng sẽ giúp doanh nghiệp trả lời được những câu hỏi quan trọng như: Những yếu tố nào ảnh hưởng đến quyết định mua sắm của khách hàng? Khách hàng mong đợi điều gì từ sản phẩm/dịch vụ? Cách thức tương tác nào hiệu quả nhất với khách hàng? Họ tìm kiếm thông tin ở đâu? Đối thủ cạnh tranh đang làm gì? Để qua đó đưa ra các giải pháp cải thiện và điều chỉnh chiến lược truyền thông, chiến lược Marketing và tối ưu hóa hiệu suất của doanh nghiệp.

### Thúc đẩy doanh số nhờ trải nghiệm cá nhân hóa

Trải nghiệm cá nhân hóa là mong muốn của hầu hết người tiêu dùng trong thời đại số hiện nay. Thông qua 580 chương trình trải nghiệm khách hàng khác nhau của các tổ chức, Medallia đã phát hiện ra các công ty mang đến trải nghiệm cá nhân hóa có lượng tương tác cao gấp hai lần và doanh thu cao hơn 26 lần so với các công ty thông thường. Điều này có thể lý giải khi 86% người tiêu dùng cho biết cá nhân hóa đóng vai trò quan trọng trong quyết định mua hàng của họ.

Việc hiểu rõ hành vi khách hàng cho thấy doanh nghiệp đang lắng nghe và quan tâm đến nhu cầu của họ. Khi doanh nghiệp cung cấp trải nghiệm được cá nhân hóa nhất quán trên các kênh sẽ tạo ra ấn tượng tích cực giúp củng cố lòng trung thành của khách hàng đối với thương hiệu. Trải nghiệm khách hàng được cải thiện này cũng giúp doanh nghiệp khác biệt so với các đối thủ cạnh tranh và khiến khách hàng quay lại nhiều hơn.

### Cải thiện chất lượng sản phẩm, dịch vụ

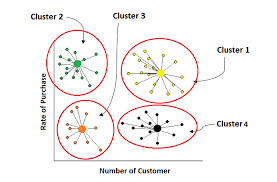
Thông qua việc nghiên cứu hành vi và phản hồi của khách hàng, doanh nghiệp có thể nhận diện những điểm mạnh và điểm yếu của sản phẩm/dịch vụ hiện tại. Qua đó, cải tiến chất lượng sản phẩm/dịch vụ, đáp ứng tốt hơn nhu cầu của khách hàng và tạo ra giá trị vượt trội.

### Tăng vị thế cạnh tranh

Khi hiểu biết chi tiết về hành vi khách hàng, doanh nghiệp có thể tối ưu hóa các chiến lược kinh doanh, cải tiến sản phẩm phù hợp hơn với nhu cầu và mong muốn của khách hàng, thậm chí là vượt qua cả sự mong đợi của họ. Bên cạnh đó, việc nghiên cứu hành vi khách hàng còn giúp doanh nghiệp nhận diện những cơ hội, tạo ra các sản phẩm/dịch vụ độc đáo và khác biệt, nâng cao lợi thế cạnh tranh trên thị trường.

Nghiên cứu hành vi của khách hàng không chỉ áp dụng cho sản phẩm mà còn cho các chiến dịch quảng cáo và truyền thông, góp phần vào sự thành công và tăng trưởng bền vững của thương hiệu.

## Phương pháp nghiên cứu



Hình 3. Thuật toán phân cụm K-Means

Để đáp ứng mục tiêu cốt lõi của đề tài là phân tích hành vi mua sắm của khách hàng và từ đó nhận diện các phân khúc thị trường tiềm năng, dự án này đã lựa chọn và áp dụng phương pháp phân cụm dữ liệu (clustering) sử dụng thuật toán K-Means. Đây là một phương pháp học máy không giám sát (unsupervised machine learning) mạnh mẽ, cho phép chúng ta khám phá cấu trúc ẩn trong dữ liệu mà không cần thông tin nhãn sẵn có. Mục tiêu của K-Means là nhóm các điểm dữ liệu vào các cụm (clusters) sao cho các điểm trong cùng một cụm có sự tương đồng cao nhất về đặc điểm, trong khi các cụm khác lại có sự khác biệt rõ rệt. Quá trình này giúp chúng ta chuyển đổi một tập dữ liệu lớn và phức tạp thành các nhóm khách hàng có thể quản lý và hiểu rõ hơn.

**1. Nền tảng Lý thuyết và Cơ chế hoạt động của Thuật toán K-Means**

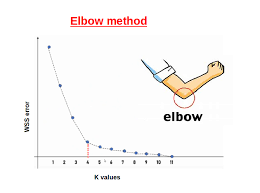
Về mặt lý thuyết, K-Means hoạt động dựa trên nguyên tắc tối thiểu hóa tổng bình phương khoảng cách từ mỗi điểm dữ liệu tới tâm cụm (centroid) gần nhất của nó. Hay nói cách khác, thuật toán tìm cách thu hẹp khoảng cách nội tại của mỗi cụm để các thành viên trong cụm trở nên đồng nhất hơn. Toàn bộ quy trình này diễn ra một cách lặp đi lặp lại và tự động, bao gồm các bước cốt lõi như sau:

1. **Khởi tạo (Initialization)**: Bước đầu tiên và cũng là một yếu tố quan trọng ảnh hưởng đến kết quả cuối cùng là việc chọn ngẫu nhiên k điểm dữ liệu làm tâm cụm ban đầu. Giá trị k (số lượng cụm mong muốn) phải được xác định trước, thường thông qua các phương pháp phân tích hiệu suất như **phương pháp Elbow** hoặc **hệ số Silhouette**. Việc khởi tạo tốt sẽ giúp thuật toán hội tụ nhanh hơn và tránh được các kết quả không tối ưu cục bộ.
2. **Gán Cụm (Assignment)**: Sau khi các tâm cụm đã được khởi tạo, thuật toán sẽ đi qua từng điểm dữ liệu trong toàn bộ tập dữ liệu. Với mỗi điểm dữ liệu, nó sẽ tính khoảng cách tới tất cả k tâm cụm hiện tại và gán điểm đó vào cụm có tâm gần nhất. Khoảng cách Euclid (Euclidean distance) là công thức thường được sử dụng nhất trong bước này, vì nó trực quan và dễ tính toán trong không gian nhiều chiều.
3. **Cập nhật Tâm Cụm (Update)**: Sau khi tất cả các điểm dữ liệu đã được gán vào một trong các cụm, thuật toán sẽ tính lại vị trí mới cho mỗi tâm cụm. Vị trí mới này chính là giá trị trung bình (mean) của tất cả các điểm dữ liệu thuộc về cụm đó. Việc này đảm bảo rằng tâm cụm luôn nằm ở trung tâm hình học của các điểm mà nó đại diện.
4. **Hội tụ (Convergence)**: Các bước Gán Cụm và Cập nhật Tâm Cụm được lặp đi lặp lại. Quá trình này sẽ dừng lại khi một trong các điều kiện sau được thỏa mãn:
   * Vị trí của các tâm cụm không còn thay đổi đáng kể giữa hai lần lặp liên tiếp.
   * Số lần lặp đạt đến giới hạn được đặt ra.
   * Các điểm dữ liệu không còn di chuyển giữa các cụm.

Quá trình lặp lại này đảm bảo rằng các cụm được hình thành ngày càng chặt chẽ và ổn định, đưa thuật toán đến một giải pháp tối ưu cục bộ cho bài toán phân cụm.

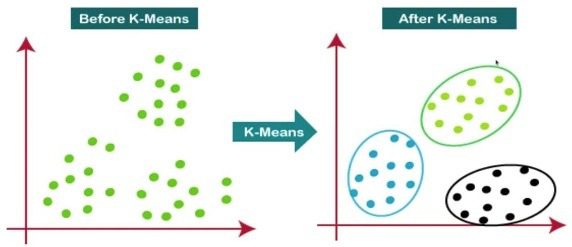
**2. Quy trình Áp dụng K-Means vào Dữ liệu Khách hàng**

Trong bối cảnh của dự án phân tích này, chúng tôi đã áp dụng thuật toán K-Means để phân tích tập dữ liệu Mall\_Customers.csv. Hai biến chính được chọn làm cơ sở để phân cụm khách hàng là **Annual Income (k$)** (Thu nhập hàng năm) và **Spending Score (1-100)** (Điểm chi tiêu). Sự lựa chọn này mang tính chiến lược vì nó cho phép chúng ta khám phá trực tiếp mối quan hệ giữa khả năng chi trả của khách hàng và hành vi tiêu dùng của họ.



Hình 4. Biểu đồ Elbow Method

* **Xác định số lượng cụm tối ưu (k)**: Trước khi áp dụng K-Means, chúng tôi đã sử dụng **phương pháp Elbow Method** để xác định số lượng cụm phù hợp. Bằng cách chạy thuật toán K-Means với các giá trị k khác nhau và tính toán tổng bình phương khoảng cách của các điểm đến tâm cụm gần nhất (inertia), chúng tôi đã vẽ một biểu đồ để tìm ra điểm "gãy" (elbow point) - nơi mà việc tăng thêm số cụm không còn làm giảm đáng kể giá trị inertia. Dựa trên phân tích này, chúng tôi xác định được giá trị k=5 là tối ưu nhất cho tập dữ liệu.
* **Thực hiện Phân cụm và Diễn giải Kết quả**: Sau khi xác định k=5, chúng tôi đã áp dụng thuật toán K-Means lên dữ liệu đã được chuẩn hóa. Kết quả là năm cụm khách hàng riêng biệt đã được hình thành. Việc diễn giải các cụm này được thực hiện bằng cách phân tích đặc điểm trung bình của các biến Annual Income và Spending Score trong mỗi cụm. Kết quả này không chỉ đơn thuần là việc nhóm dữ liệu mà còn cung cấp những hiểu biết sâu sắc về các phân khúc khách hàng, từ đó giúp doanh nghiệp đưa ra các chiến lược kinh doanh và marketing phù hợp cho từng nhóm.



Hình 4. Thuật toán K-Means với bài toán phân cụm dữ liệu

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Phân tích dữ liệu

### Phân tích dữ liệu là gì?



Hình 5. Data analytics hiện đang là xu hướng trong môi trường kinh doanh hiện đại

Data Analytics hay còn gọi là Phân tích dữ liệu là quá trình khai thác, xử lý và chuyển đổi dữ liệu thô thành những thông tin hữu ích, giúp doanh nghiệp hiểu rõ hơn về hoạt động, hành vi khách hàng và từ đó đưa ra quyết định chính xác. Quá trình này bao gồm nhiều giai đoạn như thu thập, phân loại, làm sạch và phân tích dữ liệu. Dữ liệu ban đầu thường rời rạc, khó sử dụng, nhưng khi được phân tích đúng cách sẽ mang lại bức tranh toàn diện về xu hướng thị trường cũng như hiệu quả vận hành.

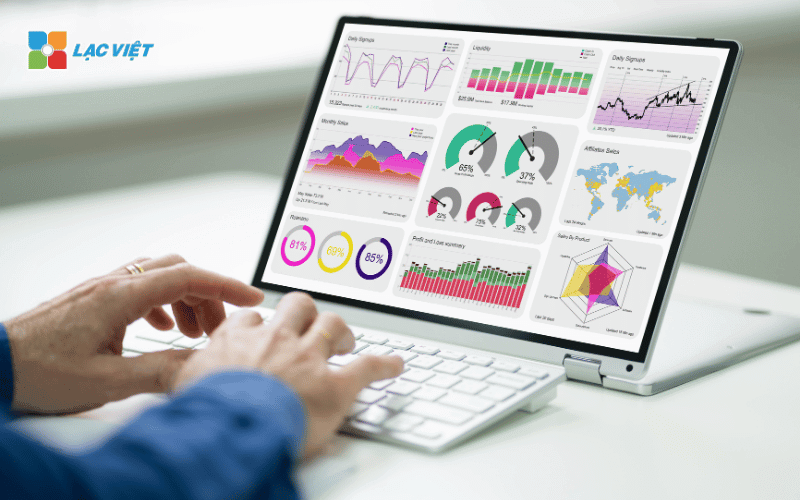
Ngày nay, phân tích dữ liệu không chỉ còn là lợi thế cạnh tranh mà đã trở thành yếu tố then chốt giúp doanh nghiệp thích nghi và dẫn dắt trong môi trường kinh doanh biến động liên tục.

Bạn sẽ bắt gặp hai khái niệm liên quan: **Data Analysis** và **Data Analytics**. Trong đó:

* **Data Analysis** thường được hiểu là hoạt động phân tích dữ liệu ở mức cơ bản, dựa trên tập dữ liệu sẵn có để tìm ra các mẫu, xu hướng hay mối quan hệ.
* **Data Analytics** là một lĩnh vực rộng và chuyên sâu hơn, ứng dụng các công cụ, công nghệ và kỹ thuật tiên tiến như Trí tuệ nhân tạo (AI), Học máy (Machine Learning) hay Big Data để xử lý khối lượng dữ liệu khổng lồ, từ đó đưa ra dự báo và khuyến nghị.

Nói cách khác, có thể coi **Data Analysis** là một phần nhỏ nằm trong phạm vi rộng lớn của **Data Analytics**.

### Vai trò của phân tích dữ liệu trong doanh nghiệp



Hình 6. Phân tích dữ liệu là cơ sở cho doanh nghiệp đưa ra quyết định chiến lược

Phân tích dữ liệu kinh doanh không chỉ giúp doanh nghiệp nắm bắt rõ tình hình tài chính và vận hành, mà còn mang lại nhiều lợi ích thiết thực như:

* **Hỗ trợ ra quyết định dựa trên dữ liệu**: Doanh nghiệp có thể đưa ra quyết định nhanh chóng và chính xác nhờ phân tích số liệu thực tế, thay vì chỉ dựa vào cảm tính.
* **Tối ưu hóa chiến lược kinh doanh – tài chính**: Phân tích dữ liệu cho phép xác định những sản phẩm/dịch vụ mang lại hiệu quả cao, từ đó tối ưu chi phí vận hành và điều chỉnh chiến lược marketing hợp lý.
* **Nâng cao hiệu quả hoạt động**: Giúp phát hiện điểm mạnh và điểm hạn chế trong quy trình, từ đó cải thiện hiệu suất làm việc và hiệu quả kinh doanh tổng thể.
* **Dự báo xu hướng thị trường**: Doanh nghiệp có thể nhận diện cơ hội mới, dự đoán hành vi khách hàng và xây dựng chiến lược đón đầu nhu cầu thị trường.

**Ví dụ**: Một doanh nghiệp bán lẻ có thể ứng dụng phân tích dữ liệu để theo dõi thói quen mua sắm của khách hàng, tối ưu hóa quản lý hàng tồn kho, nâng cao trải nghiệm mua hàng và từ đó thúc đẩy doanh thu.

### Tại sao doanh nghiệp cần phân tích dữ liệu kinh doanh?



Hình 7. Phân tích dữ liệu kinh doanh hỗ trợ doanh nghiệp cải thiện trải nghiệm khách hàng

Trong thời đại số hóa, doanh nghiệp có thể tiếp cận một lượng dữ liệu khổng lồ từ nhiều nguồn khác nhau như dữ liệu bán hàng, dữ liệu tài chính, dữ liệu khách hàng và dữ liệu vận hành. Tuy nhiên, nếu không có chiến lược phân tích dữ liệu hiệu quả, doanh nghiệp sẽ khó tận dụng tối đa giá trị từ dữ liệu này.

Dưới đây là những lý do chính khiến phân tích dữ liệu kinh doanh trở nên cần thiết đối với mọi doanh nghiệp:

#### Hỗ trợ ra quyết định dựa trên dữ liệu

* Giúp doanh nghiệp đưa ra quyết định nhanh chính xác hơn, giảm thiểu rủi ro.
* Cung cấp dữ liệu thời gian thực, giúp ban lãnh đạo có góc nhìn chính xác về tình hình kinh doanh.

Ví dụ: Một công ty thương mại điện tử sử dụng phân tích dữ liệu để xác định sản phẩm bán chạy, tối ưu hóa danh mục sản phẩm.

#### Tối ưu hóa chiến lược tài chính kinh doanh

* Giúp doanh nghiệp hiểu rõ mô hình chi tiêu, điều chỉnh ngân sách hợp lý.
* Xác định kênh marketing hiệu quả nhất để tối ưu hóa chi phí quảng cáo.

Ví dụ: Một doanh nghiệp sản xuất sử dụng dữ liệu để tối ưu hóa chuỗi cung ứng, giảm thiểu chi phí lưu kho, vận hành.

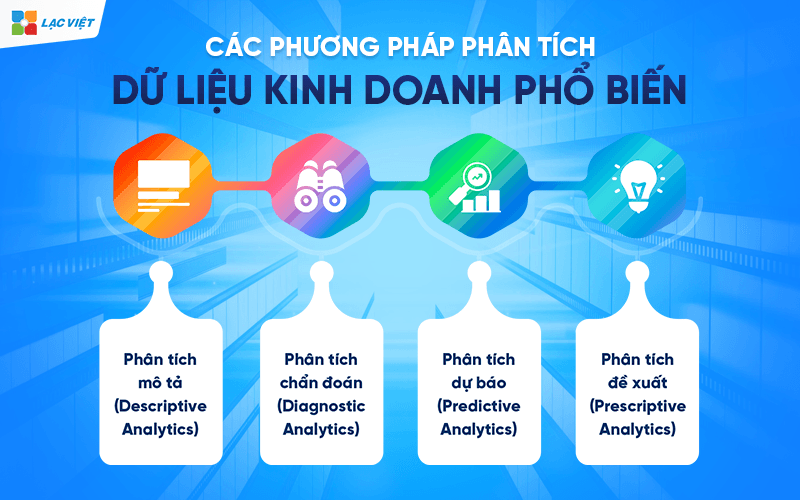
#### Cải thiện hiệu suất hoạt động và dự báo xu hướng

* Phân tích dữ liệu giúp doanh nghiệp nâng cao hiệu suất nhân sự, tối ưu hóa quy trình làm việc.
* Dự báo xu hướng tiêu dùng giúp doanh nghiệp đón đầu thị trường đáp ứng nhu cầu khách hàng.

Ví dụ: Một ngân hàng sử dụng AI để phân tích dữ liệu khách hàng, phát hiện hành vi gian lận và đề xuất các sản phẩm tài chính phù hợp.

### Các phương pháp phân tích dữ liệu kinh doanh phổ biến

Doanh nghiệp có thể áp dụng nhiều phương pháp khác nhau để khai thác giá trị từ dữ liệu kinh doanh. Dưới đây là bốn phương pháp phân tích dữ liệu quan trọng nhất:



Hình 8. Các phương pháp phân tích dữ liệu

#### Phân tích mô tả (Descriptive Analytics)

Là phương pháp phổ biến nhất giúp doanh nghiệp hiểu rõ về quá khứ bằng cách tổng hợp, hiển thị dữ liệu. Công cụ chính: Bảng điều khiển (dashboard), báo cáo tài chính, biểu đồ trực quan.

Ví dụ: Một doanh nghiệp bán lẻ sử dụng phân tích mô tả để theo dõi doanh thu theo từng tháng, từ đó điều chỉnh chiến lược kinh doanh theo mùa vụ.

#### Phân tích chẩn đoán (Diagnostic Analytics)

* Giúp doanh nghiệp tìm hiểu nguyên nhân đằng sau các biến động tài chính hoặc kinh doanh.
* Sử dụng các mô hình thống kê để xác định yếu tố ảnh hưởng đến hiệu suất kinh doanh.
* Ví dụ: Nếu doanh số bán hàng giảm mạnh trong một khu vực cụ thể, doanh nghiệp có thể sử dụng phân tích chẩn đoán để xác định nguyên nhân như giá cả, đối thủ cạnh tranh hoặc xu hướng tiêu dùng thay đổi.

#### Phân tích dự báo (Predictive Analytics)

* Dựa trên dữ liệu lịch sử để dự đoán xu hướng, kịch bản tương lai.
* Ứng dụng AI và Machine Learning để dự báo doanh thu, nhu cầu thị trường hoặc rủi ro tài chính.

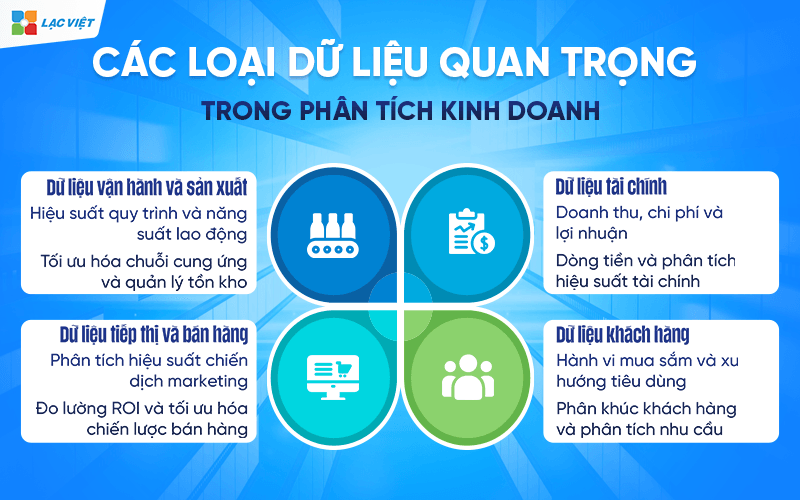
Ví dụ: Một công ty bảo hiểm sử dụng phân tích dự báo để ước tính tỷ lệ yêu cầu bồi thường trong tương lai, giúp điều chỉnh mức phí bảo hiểm hợp lý.

#### Phân tích đề xuất (Prescriptive Analytics)

* Cấp độ cao nhất của phân tích dữ liệu, giúp doanh nghiệp đưa ra khuyến nghị dựa trên dữ liệu thực tế.
* Ứng dụng trong tối ưu hóa vận hành, tài chính, marketing.
* Ví dụ: Một doanh nghiệp logistic sử dụng phân tích đề xuất để tối ưu tuyến đường giao hàng, giảm thiểu chi phí vận chuyển.

### Các loại dữ liệu quan trọng trong phân tích kinh doanh

Trong quá trình phân tích dữ liệu kinh doanh, việc xác định đúng loại dữ liệu có ảnh hưởng lớn đến hiệu quả ra quyết định, tối ưu hóa vận hành, cải thiện hiệu suất tài chính. Dữ liệu kinh doanh có thể được chia thành bốn nhóm chính: dữ liệu tài chính, dữ liệu khách hàng, dữ liệu vận hành sản xuất, dữ liệu tiếp thị, bán hàng. Mỗi nhóm dữ liệu này mang lại những giá trị khác nhau, đóng vai trò quan trọng trong việc phân tích, dự báo, điều chỉnh chiến lược kinh doanh.



Hình 9. Các loại dữ liệu quan trọng trong kinh doanh

#### Dữ liệu tài chính

Dữ liệu tài chính là nền tảng của mọi quyết định trong doanh nghiệp. Đây là nhóm dữ liệu giúp doanh nghiệp đánh giá tình hình tài chính, đo lường hiệu suất hoạt động, lập kế hoạch ngân sách hiệu quả.

* **Doanh thu, chi phí và lợi nhuận**
* Doanh thu: Tổng thu nhập từ hoạt động kinh doanh, có thể phân tích theo từng sản phẩm, dịch vụ, khu vực hoặc thời gian.
* Chi phí: Gồm chi phí cố định (thuê mặt bằng, lương nhân viên), chi phí biến đổi (nguyên liệu, vận chuyển).
* Lợi nhuận: Chỉ số quan trọng để đánh giá hiệu quả kinh doanh, bao gồm lợi nhuận gộp, lợi nhuận ròng và biên lợi nhuận.

Ứng dụng phân tích dữ liệu tài chính:

* Xác định sản phẩm/dịch vụ có lợi nhuận cao nhất, cắt giảm những khoản chi phí không cần thiết.
* Phân tích xu hướng lợi nhuận để lập kế hoạch tài chính dài hạn.
* **Dòng tiền và phân tích hiệu suất tài chính**
* Dòng tiền hoạt động: Tiền mặt thực tế doanh nghiệp có được từ hoạt động kinh doanh.
* Dòng tiền đầu tư: Các khoản đầu tư vào tài sản cố định, cổ phiếu hoặc dự án mở rộng.
* Dòng tiền tài trợ: Dòng tiền từ vốn vay hoặc huy động từ cổ đông.

Ứng dụng phân tích dữ liệu dòng tiền:

* Theo dõi dòng tiền ra/vào để đảm bảo khả năng thanh toán.
* Phát hiện vấn đề về quản lý tài chính, tránh tình trạng mất cân bằng trong quản lí dòng tiền trong kinh doanh.

Ví dụ: Nếu phân tích dòng tiền cho thấy doanh nghiệp thường xuyên bị thiếu hụt thanh khoản vào cuối tháng, có thể doanh nghiệp cần điều chỉnh chính sách thu hồi công nợ hoặc tối ưu hóa chi phí vận hành.

#### Dữ liệu khách hàng

Dữ liệu khách hàng giúp doanh nghiệp hiểu rõ hành vi tiêu dùng, nhu cầu thói quen mua sắm, từ đó xây dựng chiến lược tiếp cận hiệu quả.

* **Hành vi mua sắm và xu hướng tiêu dùng**
* Tần suất mua hàng: Mức độ thường xuyên khách hàng quay lại mua sản phẩm/dịch vụ.
* Giá trị đơn hàng trung bình (AOV – Average Order Value): Tổng giá trị giao dịch trung bình mỗi lần mua hàng.
* Sản phẩm/dịch vụ ưa thích: Những danh mục được khách hàng quan tâm nhiều nhất.

Ứng dụng phân tích dữ liệu khách hàng:

* Điều chỉnh danh mục sản phẩm theo nhu cầu thực tế.
* Tối ưu hóa chiến lược giá, chương trình khuyến mãi.

Ví dụ: Nếu doanh nghiệp phát hiện khách hàng có xu hướng mua nhiều hơn vào cuối tuần, có thể triển khai chương trình giảm giá hoặc quảng cáo mạnh hơn vào thời điểm này để tăng doanh thu.

* **Phân khúc khách hàng – phân tích nhu cầu**
* Phân loại khách hàng theo độ tuổi, giới tính, thu nhập, hành vi mua sắm.
* Phân tích nhu cầu sở thích để cá nhân hóa trải nghiệm khách hàng.

Ứng dụng phân tích dữ liệu khách hàng:

* Xây dựng chiến lược tiếp thị theo từng phân khúc khách hàng.
* Phát triển sản phẩm/dịch vụ mới phù hợp với nhu cầu thị trường.

Ví dụ: Nếu doanh nghiệp nhận thấy khách hàng trẻ tuổi thích mua hàng qua kênh thương mại điện tử, có thể tập trung đẩy mạnh quảng cáo trên các nền tảng số thay vì cửa hàng truyền thống.

#### Dữ liệu vận hành sản xuất

Dữ liệu vận hành, sản xuất giúp doanh nghiệp quản lý hiệu suất lao động, tối ưu hóa quy trình sản xuất giảm thiểu lãng phí.

* **Hiệu suất quy trình, năng suất lao động**
* Thời gian sản xuất trung bình: Đo lường thời gian hoàn thành mỗi đơn hàng hoặc sản phẩm.
* Tỷ lệ lỗi sản phẩm: Phân tích nguyên nhân lỗi trong quá trình sản xuất.
* Hiệu suất lao động: Đo lường năng suất làm việc của nhân viên.

Ứng dụng phân tích dữ liệu vận hành:

* Tối ưu hóa quy trình sản xuất để giảm thời gian và chi phí vận hành.
* Cải thiện chất lượng sản phẩm bằng cách xác định nguyên nhân lỗi.

Ví dụ: Nếu dữ liệu chỉ ra rằng một công đoạn sản xuất có tỷ lệ lỗi cao, doanh nghiệp có thể kiểm tra lại quy trình, cải tiến công nghệ để giảm thiểu rủi ro.

* **Tối ưu hóa chuỗi cung ứng quản lý tồn kho**
* Vòng quay hàng tồn kho: Xác định tốc độ tiêu thụ hàng hóa.
* Tồn kho an toàn: Đảm bảo lượng hàng vừa đủ để tránh thiếu hụt hoặc dư thừa.

Ứng dụng phân tích dữ liệu chuỗi cung ứng:

* Cải thiện dự báo nhu cầu để tối ưu hóa lượng hàng tồn kho.
* Giảm chi phí lưu kho, tránh lãng phí nguyên vật liệu.

Ví dụ: Nếu phân tích dữ liệu cho thấy hàng tồn kho tăng cao nhưng doanh số bán hàng giảm, doanh nghiệp cần điều chỉnh chính sách nhập hàng hoặc đẩy mạnh chiến lược bán hàng để giải phóng tồn kho.

#### Dữ liệu tiếp thị bán hàng

Dữ liệu tiếp thị bán hàng giúp doanh nghiệp đánh giá hiệu quả chiến lược tiếp thị, đo lường hiệu suất bán hàng và tối ưu hóa ROI (Return on Investment).

**Phân tích hiệu suất chiến dịch marketing**

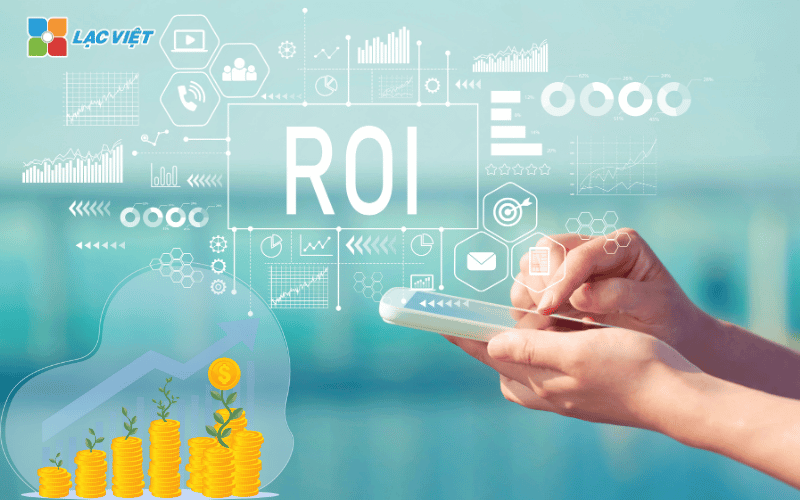
* CTR (Click-through Rate): Đo lường tỷ lệ nhấp chuột vào quảng cáo.
* Chỉ số chuyển đổi (Conversion Rate): Phân tích tỷ lệ khách hàng tiềm năng thực hiện hành động mong muốn (mua hàng, đăng ký).

Ứng dụng phân tích dữ liệu marketing:

* Xác định kênh tiếp thị hiệu quả nhất để tối ưu ngân sách.
* Điều chỉnh nội dung quảng cáo dựa trên phản hồi của khách hàng.

Ví dụ: Nếu dữ liệu cho thấy chiến dịch email marketing có tỷ lệ mở email thấp, doanh nghiệp có thể thử nghiệm tiêu đề hấp dẫn hơn hoặc điều chỉnh thời gian gửi email.

* **Đo lường ROI và tối ưu hóa chiến lược bán hàng**



Hình 10. ROI - Các dữ liệu tiếp thị giúp doanh nghiệp tối ưu chiến dịch quảng cáo

* Doanh thu trên mỗi kênh bán hàng: Đánh giá hiệu suất từng kênh (trực tuyến, cửa hàng, đại lý).
* Chi phí trên mỗi khách hàng tiềm năng (Customer Acquisition Cost – CAC): Phân tích chi phí để có được một khách hàng mới.

Ứng dụng phân tích dữ liệu bán hàng:

* Điều chỉnh chiến lược giá dựa trên nhu cầu thị trường.
* Tối ưu hóa mô hình bán hàng để tăng doanh thu với chi phí thấp nhất.

Ví dụ: Nếu phân tích dữ liệu cho thấy tỷ lệ chuyển đổi từ quảng cáo Facebook cao hơn Google Ads, doanh nghiệp có thể tăng ngân sách quảng cáo trên Facebook để tối ưu hiệu quả.

### Hướng dẫn quy trình thực hiện phân tích dữ liệu kinh doanh chi tiết

#### Bước 1. Xác định mục tiêu phân tích dữ liệu

Việc xác định rõ mục tiêu phân tích dữ liệu là nền tảng quan trọng giúp doanh nghiệp tối ưu hóa quá trình ra quyết định. Một mục tiêu phân tích dữ liệu tốt cần phải cụ thể, đo lường được, có thể thực hiện, phù hợp với chiến lược kinh doanh, có thời hạn rõ ràng (SMART: Specific, Measurable, Achievable, Relevant, Time-bound).

Tại sao cần xác định mục tiêu trước khi phân tích dữ liệu?

* Giúp tập trung vào vấn đề cốt lõi: Tránh thu thập xử lý dữ liệu không cần thiết, tiết kiệm thời gian và tài nguyên.
* Tạo hướng dẫn rõ ràng cho đội ngũ phân tích dữ liệu: Đảm bảo các nhà phân tích hiểu rõ yêu cầu của doanh nghiệp.
* Định hướng chiến lược kinh doanh: Dữ liệu không có giá trị nếu không liên kết với các mục tiêu kinh doanh cụ thể.

Các loại mục tiêu phổ biến trong phân tích dữ liệu kinh doanh

* Tăng doanh thu bằng cách tối ưu hóa giá bán sản phẩm.
* Cải thiện hiệu suất hoạt động bằng cách phân tích hiệu quả sản xuất.
* Nâng cao trải nghiệm khách hàng thông qua phân tích hành vi tiêu dùng.
* Giảm chi phí vận hành bằng cách tối ưu hóa chuỗi cung ứng.

#### Bước 2. Thu thập dữ liệu kinh doanh

Thu thập dữ liệu là bước quan trọng trong quy trình phân tích giúp đảm bảo doanh nghiệp có đủ thông tin để đưa ra quyết định chính xác.

Bao gồm các nguồn dữ liệu cần thu thập sau:

Dữ liệu nội bộ

* Báo cáo tài chính: Doanh thu, lợi nhuận, chi phí.
* Dữ liệu bán hàng: Sản phẩm bán chạy, hiệu suất của từng kênh bán hàng.
* Dữ liệu khách hàng: Thông tin cá nhân, hành vi mua sắm, lịch sử giao dịch.
* Dữ liệu nhân sự: Hiệu suất làm việc, tỷ lệ nghỉ việc, chi phí lương.
* Dữ liệu sản xuất: Hiệu suất máy móc, thời gian vận hành, chi phí nguyên liệu.

Dữ liệu bên ngoài

* Dữ liệu thị trường: Xu hướng tiêu dùng, giá cả nguyên vật liệu, đối thủ cạnh tranh.
* Phản hồi khách hàng: Đánh giá trên mạng xã hội, khảo sát khách hàng, xếp hạng sản phẩm.
* Dữ liệu kinh tế: Lạm phát, tỷ giá hối đoái, chi tiêu người tiêu dùng.

Dữ liệu thời gian thực

* Dữ liệu từ website: Lưu lượng truy cập, tỷ lệ chuyển đổi, thời gian trên trang.
* Dữ liệu từ cảm biến IoT: Giám sát máy móc sản xuất, kiểm soát chất lượng sản phẩm.
* Dữ liệu giao dịch tài chính: Phân tích biến động dòng tiền theo thời gian thực.

Lưu ý khi thu thập dữ liệu

* Đảm bảo dữ liệu chính xác: Sử dụng nguồn tin cậy, tránh sai lệch số liệu.
* Tính toàn vẹn dữ liệu: Không bỏ sót hoặc ghi chép sai.
* Tuân thủ bảo mật dữ liệu: Bảo vệ thông tin khách hàng, tuân thủ quy định về dữ liệu cá nhân (GDPR, ISO 27001).
* Định kỳ cập nhật dữ liệu: Tránh tình trạng dữ liệu lỗi thời, không phản ánh thực tế kinh doanh.

#### Bước 3. Làm sạch và chuẩn hóa dữ liệu

Dữ liệu thô thu thập từ nhiều nguồn thường chứa lỗi, sai sót hoặc không đồng nhất. Làm sạch dữ liệu giúp loại bỏ những vấn đề này, đảm bảo chất lượng phân tích.

**Các bước làm sạch dữ liệu:**

1. Xóa dữ liệu trùng lặp: Khi thu thập dữ liệu từ nhiều hệ thống khác nhau, có thể xuất hiện các bản ghi trùng lặp. Ví dụ: Một khách hàng có thể xuất hiện nhiều lần trong hệ thống CRM do nhập liệu thủ công.

2. Xử lý dữ liệu bị thiếu: phương pháp thay thế dữ liệu bị thiếu

* Điền bằng giá trị trung bình, trung vị.
* Sử dụng giá trị gần nhất theo thời gian.
* Loại bỏ dữ liệu nếu tỷ lệ thiếu quá lớn.

3. Chuẩn hóa định dạng dữ liệu: Đồng bộ hóa đơn vị tiền tệ, đơn vị đo lường, cách ghi ngày tháng. Ví dụ, nếu một hệ thống lưu dữ liệu ngày ở định dạng DD/MM/YYYY còn hệ thống khác sử dụng MM/DD/YYYY, cần chuẩn hóa thành một tiêu chuẩn chung.

4. Xác thực dữ liệu: Kiểm tra logic của dữ liệu:

* Doanh thu không thể là số âm.
* Giá bán sản phẩm không thể nhỏ hơn chi phí sản xuất.
* Nhân viên không thể có số giờ làm việc lớn hơn 24 giờ/ngày.

**Công cụ hỗ trợ làm sạch dữ liệu**

* Microsoft Excel & Google Sheets: Tốt cho các tập dữ liệu nhỏ, có chức năng lọc, loại bỏ trùng lặp.
* SQL: Giúp truy vấn, hợp nhất, làm sạch dữ liệu từ nhiều hệ thống.
* Python (Pandas, NumPy): Hữu ích cho xử lý dữ liệu lớn, tự động hóa việc loại bỏ trùng lặp, xử lý dữ liệu bị thiếu.
* ETL Tools (Talend, Alteryx): Tự động hóa quy trình thu thập làm sạch dữ liệu.

#### Bước 4. Trực quan hóa dữ liệu

Bước này giúp doanh nghiệp hiểu rõ hơn về dữ liệu thông qua phân tích thống kê trực quan hóa.

Các phương pháp khám phá dữ liệu:

* Thống kê mô tả: Tính trung bình, trung vị, phương sai, độ lệch chuẩn của dữ liệu.
* Phát hiện xu hướng: Xác định xu hướng tăng trưởng doanh thu, mức độ biến động giá, hành vi mua hàng theo thời gian.
* Phân tích mối quan hệ: Xác định mối tương quan giữa các yếu tố như giá bán, doanh thu, chi phí marketing và số lượng khách hàng.

Công cụ hỗ trợ:

* Power BI, Tableau, Google Data Studio để trực quan hóa dữ liệu.
* Python (Matplotlib, Seaborn) hoặc R để phân tích dữ liệu chuyên sâu.

#### Bước 5. Xây dựng mô hình phân tích dự báo

Sau khi hiểu dữ liệu, doanh nghiệp có thể áp dụng các mô hình phân tích nâng cao để dự báo xu hướng, hỗ trợ ra quyết định.

Các kỹ thuật phân tích dữ liệu phổ biến:

* Phân tích hồi quy: Dự đoán doanh thu dựa trên các yếu tố tác động như chi phí quảng cáo, giá bán.
* Phân nhóm khách hàng (Customer Segmentation): Nhận diện nhóm khách hàng tiềm năng để tối ưu chiến lược marketing.
* Phát hiện bất thường (Anomaly Detection): Nhận diện giao dịch gian lận, biến động bất thường trong hoạt động kinh doanh.
* Dự báo chuỗi thời gian (Time Series Forecasting): Dự đoán nhu cầu sản phẩm, biến động thị trường.

Công cụ hỗ trợ:

* Python (Scikit-learn, TensorFlow) để xây dựng mô hình phân tích.
* Các nền tảng AI/ML như Google Cloud AI, Microsoft Azure Machine Learning

#### Bước 6. Đưa ra kết luận hành động

Dựa trên kết quả phân tích, doanh nghiệp cần đưa ra quyết định chiến lược, kế hoạch hành động phù hợp.

Ví dụ về hành động dựa trên phân tích dữ liệu:

* Nếu phân tích chi phí cho thấy một bộ phận tốn quá nhiều chi phí mà không tạo ra giá trị tương ứng, doanh nghiệp có thể tối ưu hóa hoặc cắt giảm ngân sách.
* Nếu dự báo nhu cầu cho thấy nhu cầu sản phẩm sẽ tăng mạnh trong tháng tới, doanh nghiệp có thể chuẩn bị thêm hàng tồn kho.
* Nếu phân tích hành vi khách hàng cho thấy nhóm khách hàng trẻ tuổi có xu hướng mua hàng nhiều hơn qua mạng xã hội, doanh nghiệp có thể tập trung đầu tư vào kênh marketing số.

#### Bước 7. Giám sát cải tiến liên tục

Phân tích dữ liệu là quá trình liên tục, doanh nghiệp cần theo dõi hiệu quả của các quyết định dựa trên dữ liệu, điều chỉnh chiến lược khi cần thiết.

Các bước giám sát cải tiến:

* Định kỳ cập nhật, kiểm tra dữ liệu để đảm bảo tính chính xác.
* So sánh kết quả thực tế với dự báo để điều chỉnh mô hình phân tích.
* Tự động hóa quy trình phân tích dữ liệu bằng AI để nâng cao hiệu suất.

Công cụ hỗ trợ:

* Hệ thống Business Intelligence (BI) để theo dõi KPI.
* Machine Learning để tự động cập nhật mô hình phân tích.

## Phân cụm (Clustering)

### 1. Khái Niệm và Bản Chất

Phân cụm là một kỹ thuật thuộc lĩnh vực học máy không giám sát (unsupervised learning). Mục tiêu cốt lõi của phân cụm không phải là dự đoán một kết quả cụ thể mà là khám phá cấu trúc tiềm ẩn hoặc mô hình bên trong một tập dữ liệu. Về cơ bản, thuật toán phân cụm sẽ tự động nhóm các đối tượng dữ liệu (trong trường hợp này là khách hàng) thành các cụm hoặc nhóm (clusters) dựa trên mức độ tương đồng của chúng. Các đối tượng trong cùng một cụm sẽ có những đặc điểm tương tự nhau, trong khi các đối tượng thuộc các cụm khác nhau lại có sự khác biệt rõ rệt.

Khác với học máy có giám sát (supervised learning), nơi mà dữ liệu đã được gán nhãn sẵn (ví dụ: "khách hàng tốt" hay "khách hàng xấu"), phân cụm hoạt động trên dữ liệu không có nhãn. Nó hoạt động như một nhà thám hiểm, tìm kiếm và định hình các nhóm tự nhiên tồn tại trong dữ liệu mà con người có thể chưa nhận ra. Điều này làm cho phân cụm trở thành một công cụ cực kỳ hữu ích cho các bài toán khám phá dữ liệu và phân tích mô hình.

Bản chất của phân cụm có thể được hiểu thông qua hai nguyên tắc chính:

* **Tính đồng nhất nội tại (High Intra-cluster Similarity)**: Các điểm dữ liệu trong cùng một cụm phải có mức độ tương đồng cao với nhau. Điều này thường được đo lường bằng khoảng cách giữa các điểm.
* **Tính khác biệt giữa các cụm (Low Inter-cluster Similarity)**: Các cụm khác nhau phải có sự khác biệt rõ rệt, nghĩa là các điểm dữ liệu thuộc các cụm khác nhau phải có mức độ tương đồng thấp với nhau.

### 2. Các Thuật Toán Phân Cụm Phổ Biến

Có nhiều thuật toán phân cụm khác nhau, mỗi thuật toán có những ưu và nhược điểm riêng, phù hợp với từng loại dữ liệu và mục tiêu cụ thể. Các loại thuật toán phổ biến có thể kể đến như:

* **Phân Cụm Phân Chia (Partitioning Clustering)**:
  + **K-Means**: Đây là thuật toán được sử dụng trong dự án này. Nó hoạt động bằng cách chia dữ liệu thành k cụm không chồng lấn. Ưu điểm của K-Means là đơn giản, dễ hiểu và hiệu quả với các tập dữ liệu lớn.
  + **K-Medoids**: Tương tự K-Means nhưng sử dụng một điểm dữ liệu thực tế (medoid) làm tâm cụm thay vì giá trị trung bình. Điều này giúp giảm thiểu ảnh hưởng của các điểm ngoại lai (outliers).
* **Phân Cụm Phân Cấp (Hierarchical Clustering)**:
  + **Phân Cấp Kết Hợp (Agglomerative)**: Bắt đầu với mỗi điểm dữ liệu là một cụm riêng biệt, sau đó gộp các cụm gần nhau nhất lại cho đến khi chỉ còn một cụm duy nhất.
  + **Phân Cấp Phân Tách (Divisive)**: Bắt đầu với tất cả các điểm trong một cụm lớn, sau đó chia cụm đó ra thành các cụm nhỏ hơn cho đến khi mỗi điểm là một cụm riêng.
* **Phân Cụm Dựa trên Mật Độ (Density-Based Clustering)**:
  + **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Nhóm các điểm dữ liệu có mật độ cao lại với nhau, tách biệt chúng khỏi các khu vực có mật độ thấp. Ưu điểm nổi bật của DBSCAN là nó có thể phát hiện các cụm có hình dạng bất kỳ và xử lý tốt các điểm nhiễu (noise) hoặc điểm ngoại lai.

### 3. Tại Sao Phân Cụm Lại Hữu Ích Trong Phân Tích Khách Hàng?

Trong bối cảnh kinh doanh hiện đại, việc thấu hiểu khách hàng là chìa khóa để tồn tại và phát triển. Phân cụm dữ liệu khách hàng không chỉ là một kỹ thuật phân tích đơn thuần mà còn là một chiến lược kinh doanh mạnh mẽ. Nó mang lại nhiều lợi ích thiết thực, giúp doanh nghiệp tối ưu hóa hoạt động và gia tăng lợi nhuận.

#### a) Cá nhân hóa Trải nghiệm Khách hàng (Personalization)

Mỗi khách hàng là một cá thể độc lập với những nhu cầu và mong muốn riêng biệt. Tuy nhiên, việc tạo ra một chiến lược marketing riêng cho từng khách hàng là không khả thi. Phân cụm giúp giải quyết vấn đề này bằng cách nhóm các khách hàng có đặc điểm và hành vi tương tự nhau. Bằng cách đó, doanh nghiệp có thể tạo ra các chiến lược marketing, sản phẩm, hoặc chương trình khuyến mãi được thiết kế riêng cho từng phân khúc khách hàng. Ví dụ, một nhóm khách hàng có thu nhập cao và chi tiêu mạnh có thể được nhắm đến với các sản phẩm cao cấp, trong khi một nhóm khác có thu nhập thấp hơn nhưng lại thường xuyên mua sắm sẽ phù hợp với các chương trình ưu đãi hoặc giảm giá.

#### b) Tối ưu hóa Chiến lược Marketing

Phân cụm cho phép doanh nghiệp phân bổ nguồn lực marketing một cách thông minh và hiệu quả hơn. Thay vì áp dụng một chiến dịch "một kích cỡ cho tất cả", doanh nghiệp có thể:

* **Xác định khách hàng mục tiêu**: Nhận diện các nhóm khách hàng có giá trị cao nhất (ví dụ: những người chi tiêu nhiều) và tập trung nỗ lực thu hút và giữ chân họ.
* **Tăng hiệu quả quảng cáo**: Thiết kế các thông điệp quảng cáo phù hợp với từng phân khúc. Một quảng cáo về sự tiện lợi có thể hấp dẫn nhóm khách hàng bận rộn, trong khi một quảng cáo về giá trị có thể thu hút nhóm khách hàng nhạy cảm về giá.
* **Lựa chọn kênh tiếp thị**: Dựa trên đặc điểm của từng cụm, doanh nghiệp có thể quyết định kênh tiếp thị nào là hiệu quả nhất (ví dụ: mạng xã hội, email, quảng cáo trên website, v.v.).

#### c) Phát triển Sản phẩm và Dịch vụ

Phân cụm khách hàng cung cấp những thông tin quý giá cho việc phát triển sản phẩm mới. Bằng cách phân tích nhu cầu và hành vi của từng nhóm, doanh nghiệp có thể:

* **Tìm ra lỗ hổng thị trường**: Phát hiện các nhóm khách hàng chưa được phục vụ tốt và tạo ra các sản phẩm để lấp đầy khoảng trống đó.
* **Tùy chỉnh sản phẩm hiện có**: Cải tiến hoặc bổ sung các tính năng cho sản phẩm dựa trên phản hồi và hành vi của các cụm khách hàng cụ thể.

#### d) Quản lý Quan hệ Khách hàng (CRM)

Phân cụm là nền tảng của các chiến lược quản lý quan hệ khách hàng hiệu quả. Nó giúp doanh nghiệp:

* **Dự đoán hành vi khách hàng**: Phân tích lịch sử của các cụm để dự đoán xu hướng mua sắm, từ đó chủ động trong việc cung cấp các chương trình khuyến mãi hoặc gợi ý sản phẩm.
* **Giảm tỷ lệ rời bỏ (Churn Rate)**: Nhận diện các nhóm khách hàng có nguy cơ rời bỏ cao và đưa ra các biện pháp giữ chân kịp thời, ví dụ như các chương trình chăm sóc đặc biệt.

### 4. Hạn Chế và Thách Thức của Phân Cụm

Mặc dù có nhiều lợi ích, phân cụm cũng không phải là không có thách thức:

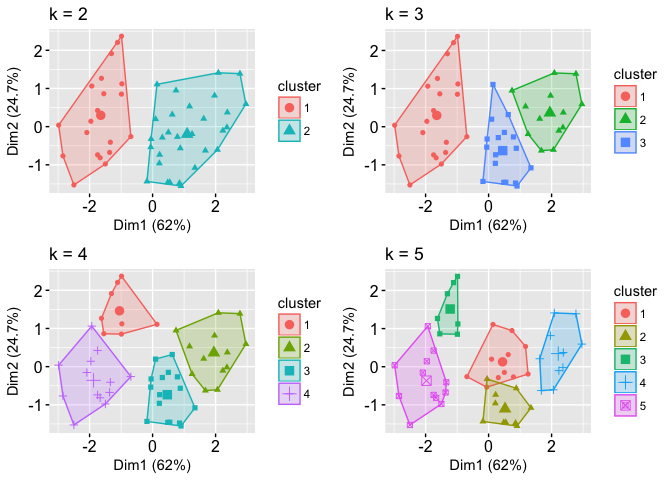
* **Xác định số lượng cụm**: Việc chọn giá trị k tối ưu là một bài toán khó, và kết quả có thể bị ảnh hưởng nếu lựa chọn không đúng.
* **Đánh giá chất lượng cụm**: Không có một chỉ số đơn lẻ nào để đánh giá chất lượng của kết quả phân cụm. Việc diễn giải kết quả thường đòi hỏi sự hiểu biết sâu sắc về lĩnh vực (domain knowledge).
* **Đặc điểm dữ liệu**: Phân cụm có thể không hiệu quả nếu các biến được chọn không phản ánh được sự khác biệt giữa các nhóm.

## Thuật toán K-Means

**1. Giới thiệu**

K-Means Clustering là một thuật toán phân cụm phổ biến, thường được sử dụng trong các ứng dụng học máy (Machine Learning) và khai phá dữ liệu (Data Mining). Mục tiêu chính của K-Means là chia một tập hợp các điểm dữ liệu thành *k* cụm (clusters) sao cho các điểm trong cùng một cụm có sự tương đồng cao nhất với nhau và khác biệt tối đa với các cụm khác.

Trong thuật toán K-Means, mỗi cụm được đại diện bởi một centroid, là giá trị trung bình của các điểm dữ liệu trong cụm đó. Thuật toán sẽ liên tục điều chỉnh các centroid và tái phân cụm các điểm dữ liệu cho đến khi đạt được sự hội tụ (convergence).



Hình 11. K-Means Clustering

**2. Cách hoạt động**

K-Means Clustering là một thuật toán phân cụm không giám sát phổ biến, dùng để nhóm các điểm dữ liệu thành K cụm dựa trên các đặc tính tương tự. Thuật toán này hoạt động dựa trên ý tưởng là giảm thiểu tổng khoảng cách bình phương giữa các điểm dữ liệu và trung tâm của cụm chúng thuộc về.

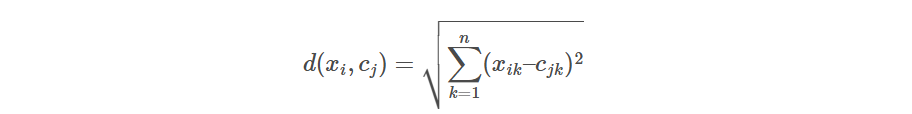
K-Means Clustering là một thuật toán phân cụm không giám sát phổ biến, dùng để nhóm các điểm dữ liệu thành K cụm dựa trên các đặc tính tương tự. Thuật toán này hoạt động dựa trên ý tưởng là giảm thiểu tổng khoảng cách bình phương giữa các điểm dữ liệu và trung tâm của cụm chúng thuộc về.

#### Khởi tạo các trung tâm cụm (Centroids)

Bước đầu tiên của K-Means là xác định số lượng cụm *K* mà bạn muốn tìm. Sau đó, chọn ngẫu nhiên *K* điểm từ tập dữ liệu làm các trung tâm cụm ban đầu, được gọi là các centroid.

#### Gán mỗi điểm dữ liệu vào cụm gần nhất

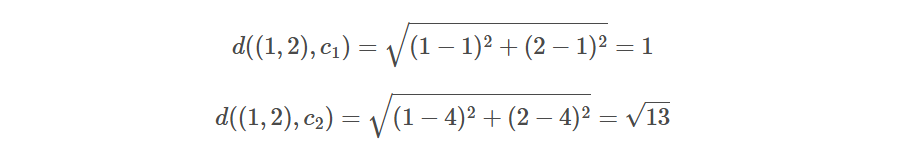
Mỗi điểm dữ liệu được gán vào cụm có centroid gần nhất, được tính bằng khoảng cách Euclidean. Khoảng cách giữa một điểm dữ liệu *xi* và một centroid *cj* được tính bằng công thức:



Trong đó:

* *xi* là điểm dữ liệu thứ *i*.
* *cj* là centroid của cụm thứ *j*.
* *n* là số chiều của dữ liệu.

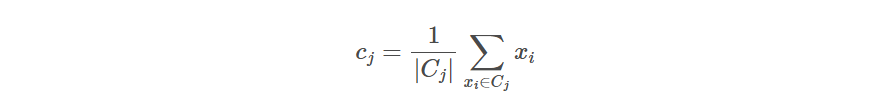
Ví dụ: Giả sử bạn có một tập dữ liệu gồm ba điểm trong không gian hai chiều: *(1, 2)*, *(2, 3)*, và *(3, 4)*, và bạn khởi tạo hai centroid *c1 = (1, 1)* và *c2 = (4, 4)*. Khoảng cách Euclidean giữa mỗi điểm và centroid được tính như sau:



Vì khoảng cách từ điểm *(1, 2)* đến centroid *c1* nhỏ hơn, nên điểm này sẽ được gán vào cụm với centroid *c1*.

#### Cập nhật các trung tâm cụm

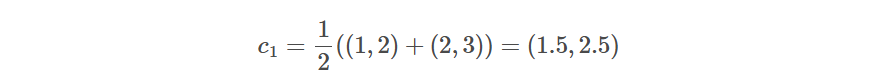
Sau khi tất cả các điểm dữ liệu đã được gán vào một cụm, các centroid được tính lại bằng cách lấy trung bình cộng các điểm dữ liệu trong mỗi cụm:



Trong đó:

* *Cj* là tập hợp các điểm dữ liệu trong cụm *j*.
* *|Cj|</i*
* *|Cj|* là số lượng điểm dữ liệu trong cụm *j*.

Ví dụ tiếp theo: Giả sử sau khi gán các điểm dữ liệu, cụm *C1* bao gồm các điểm *(1, 2)* và *(2, 3)*, thì centroid mới *c1* sẽ được tính như sau:

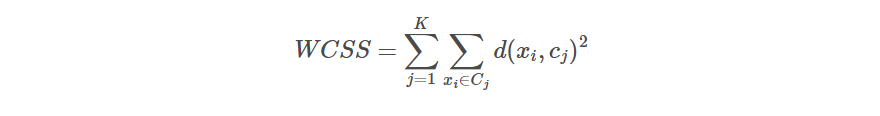


#### Lặp lại quá trình

Các bước gán cụm và cập nhật centroid được lặp lại cho đến khi các centroid không còn thay đổi đáng kể hoặc đạt đến số lần lặp tối đa. Thuật toán hội tụ khi không có sự thay đổi trong gán cụm của các điểm dữ liệu hoặc sự thay đổi rất nhỏ.

#### Đánh giá chất lượng phân cụm

Chất lượng của việc phân cụm có thể được đánh giá bằng tổng bình phương sai số trong cụm (Within-Cluster Sum of Squares – WCSS), được tính bằng công thức:



Thuật toán cố gắng giảm thiểu giá trị WCSS để đạt được sự phân cụm tốt nhất.

**Ví dụ minh họa**

Giả sử bạn có dữ liệu sau đây:

* Điểm dữ liệu: *(1, 1)*, *(2, 1)*, *(4, 3)*, *(5, 4)*
* Khởi tạo *K = 2* với các centroid ban đầu *c1 = (1, 1)* và *c2 = (5, 4)*.

**Bước 1**: Gán các điểm dữ liệu vào các cụm:

* Điểm *(1, 1)* và *(2, 1)* gần centroid *c1* hơn, nên thuộc cụm 1.
* Điểm *(4, 3)* và *(5, 4)* gần centroid *c2* hơn, nên thuộc cụm 2.

**Bước 2**: Cập nhật các centroid:

* Centroid mới của cụm 1 là *c1 = (1.5, 1)*.
* Centroid mới của cụm 2 là *c2 = (4.5, 3.5)*.

**Bước 3**: Lặp lại bước 1 và bước 2 cho đến khi các centroid không thay đổi.

**3. Ứng dụng**

K-Means Clustering được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau, bao gồm:

* **Phân tích khách hàng:** K-Means có thể được sử dụng để phân nhóm khách hàng dựa trên các hành vi mua sắm, từ đó giúp doanh nghiệp hiểu rõ hơn về các phân khúc khách hàng khác nhau.
* **Phân tích hình ảnh:** Trong xử lý hình ảnh, K-Means có thể được sử dụng để phân cụm các điểm ảnh có màu sắc tương đồng, giúp trong việc nén ảnh hoặc phân đoạn ảnh.
* **Phân tích văn bản:** K-Means có thể phân nhóm các tài liệu hoặc từ ngữ dựa trên các đặc trưng chung, giúp tổ chức dữ liệu văn bản hoặc tìm kiếm thông tin.

4. Ưu điểm, nhược điểm của K-Means Clustering

#### Ưu điểm của K-Means Clustering

* **Đơn giản và dễ hiểu:** K-Means là một trong những thuật toán phân cụm dễ hiểu nhất, với quy trình hoạt động rõ ràng và dễ triển khai.
* **Hiệu quả:** Thuật toán này có thể xử lý một lượng lớn dữ liệu với chi phí tính toán thấp, đặc biệt là khi số lượng cụm k nhỏ.
* **Khả năng mở rộng:** K-Means có thể mở rộng tốt với dữ liệu lớn, nhờ vào tính chất tuyến tính của nó. Nó có thể được áp dụng trên các tập dữ liệu lớn với hàng triệu điểm dữ liệu.
* **Linh hoạt:** Có thể áp dụng K-Means cho nhiều loại dữ liệu khác nhau, bao gồm cả dữ liệu số và dữ liệu danh mục (categorical).

#### Nhược điểm của K-Means Clustering

* **Số cụm phải được xác định trước:** Một trong những hạn chế lớn nhất của K-Means là yêu cầu người dùng phải xác định số lượng cụm k trước khi chạy thuật toán. Điều này có thể khó khăn khi không biết trước số cụm tối ưu.
* **Nhạy cảm với vị trí khởi tạo centroid:** Kết quả của K-Means phụ thuộc rất nhiều vào việc khởi tạo centroid ban đầu. Khởi tạo kém có thể dẫn đến hội tụ tại một cực trị cục bộ không tối ưu.
* **Chỉ nhận diện các cụm hình cầu:** K-Means hoạt động tốt với các cụm có hình dạng gần như cầu và kích thước tương đồng. Nó không thể xử lý tốt các cụm có hình dạng phi tuyến hoặc không đồng đều.
* **Nhạy cảm với nhiễu:** K-Means rất nhạy cảm với các điểm dữ liệu ngoại lai (outliers), vì các outliers có thể kéo centroid ra khỏi vị trí trung tâm của cụm thực sự.

#### Khi nào nên sử dụng K-Means Clustering?

K-Means Clustering là một lựa chọn tốt trong các trường hợp sau:

* **Dữ liệu có cấu trúc đơn giản:** Khi dữ liệu có cấu trúc phân cụm đơn giản với các cụm hình cầu hoặc gần hình cầu, K-Means là một công cụ mạnh mẽ.
* **Số lượng cụm được biết trước:** Khi bạn đã biết hoặc có thể ước lượng chính xác số cụm k, K-Means có thể nhanh chóng phân nhóm dữ liệu.
* **Phân tích dữ liệu sơ bộ:** Khi cần phân tích dữ liệu sơ bộ để khám phá các mẫu ẩn, K-Means có thể cung cấp cái nhìn nhanh chóng và rõ ràng về cấu trúc của dữ liệu.
* **Yêu cầu thời gian tính toán thấp:** Khi thời gian và tài nguyên tính toán hạn chế, K-Means cung cấp giải pháp nhanh chóng với hiệu suất tính toán cao.

# CHƯƠNG 3: THỰC NGHIỆM VÀ KẾT QUẢ

## Mô tả tập dữ liệu

### 1. Nguồn gốc và chi tiết dữ liệu

|  | **CustomerID** | **Gender** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
| --- | --- | --- | --- | --- | --- |
| 0 | 1 | Male | 19 | 15 | 39 |
| 1 | 2 | Male | 21 | 15 | 81 |
| 2 | 3 | Female | 20 | 16 | 6 |
| 3 | 4 | Female | 23 | 16 | 77 |
| 4 | 5 | Female | 31 | 17 | 40 |

Tập dữ liệu được sử dụng trong dự án này có tên là **Mall\_Customers.csv**, được thu thập từ Kaggle. Đây là một tập dữ liệu tổng hợp, chứa thông tin về khách hàng của một trung tâm mua sắm. Mục tiêu chính của việc sử dụng tập dữ liệu này là để thực hiện phân tích hành vi và phân khúc khách hàng dựa trên các đặc điểm cơ bản. Tập dữ liệu bao gồm **200 bản ghi** (tương ứng với 200 khách hàng) và **5 trường dữ liệu** (cột).

Dưới đây là mô tả chi tiết về ý nghĩa của từng trường dữ liệu:

* **CustomerID**:
  + **Mô tả**: Đây là trường dữ liệu chứa mã định danh duy nhất cho mỗi khách hàng. Mỗi khách hàng sẽ có một mã số riêng biệt, giúp phân biệt họ với những khách hàng khác trong tập dữ liệu.
  + **Ý nghĩa trong phân tích**: Mặc dù trường này không trực tiếp được sử dụng trong quá trình phân cụm (vì nó không phải là một đặc trưng hành vi), nó đóng vai trò quan trọng trong việc quản lý và truy xuất thông tin của từng khách hàng sau khi đã phân cụm. Nó đảm bảo rằng chúng ta có thể dễ dàng xác định được những khách hàng cụ thể thuộc về một nhóm nào đó sau khi phân tích.
* **Gender**:
  + **Mô tả**: Trường này thể hiện giới tính của khách hàng, bao gồm hai giá trị là "Male" (Nam) và "Female" (Nữ).
  + **Ý nghĩa trong phân tích**: Gender là một biến phân loại (categorical variable). Mặc dù thuật toán K-Means hoạt động tốt nhất trên các biến số (numerical variables), thông tin về giới tính vẫn rất hữu ích trong giai đoạn phân tích khám phá (Exploratory Data Analysis - EDA) để hiểu rõ hơn về sự phân bố của khách hàng theo giới tính và có thể sử dụng để diễn giải kết quả sau khi phân cụm. Ví dụ, chúng ta có thể phân tích xem liệu có sự khác biệt rõ rệt về thu nhập hay điểm chi tiêu giữa nam và nữ hay không.
* **Age**:
  + **Mô tả**: Trường này chứa thông tin về tuổi của khách hàng. Tuổi là một biến số liên tục, dao động từ 18 đến 70 trong tập dữ liệu này.
  + **Ý nghĩa trong phân tích**: Tuổi tác là một trong những yếu tố nhân khẩu học quan trọng, thường có mối liên hệ với hành vi mua sắm. Khách hàng ở các nhóm tuổi khác nhau có thể có nhu cầu, sở thích và khả năng chi trả khác nhau. Phân tích tuổi giúp chúng ta có cái nhìn sâu sắc về đặc điểm của từng nhóm khách hàng đã được phân cụm, từ đó đưa ra các chiến lược tiếp thị phù hợp với từng độ tuổi.
* **Annual Income (k$)**:
  + **Mô tả**: Trường này biểu thị thu nhập hàng năm của khách hàng, được tính bằng nghìn đô la ($). Đây là một biến số liên tục, dao động từ $15k đến $137k.
  + **Ý nghĩa trong phân tích**: Thu nhập là một chỉ số mạnh mẽ về khả năng chi trả và sức mua của khách hàng. Nó đóng vai trò then chốt trong quá trình phân cụm, vì nó giúp phân tách các nhóm khách hàng dựa trên mức độ giàu có của họ. Một nhóm khách hàng có thu nhập cao có thể là mục tiêu cho các sản phẩm và dịch vụ cao cấp, trong khi một nhóm có thu nhập thấp hơn có thể nhạy cảm với giá cả hơn.
* **Spending Score (1-100)**:
  + **Mô tả**: Điểm chi tiêu là một giá trị được gán bởi trung tâm mua sắm, dao động từ 1 đến 100. Điểm này dựa trên hành vi mua sắm của khách hàng, phản ánh tần suất mua hàng, số lượng mặt hàng đã mua và tổng số tiền chi tiêu. Một điểm chi tiêu cao cho thấy khách hàng thường xuyên mua sắm và có xu hướng chi tiêu nhiều.
  + **Ý nghĩa trong phân tích**: Đây là biến số quan trọng nhất trong việc phân tích hành vi mua sắm. Kết hợp Spending Score với Annual Income là cách hiệu quả để tạo ra các phân khúc khách hàng có ý nghĩa kinh doanh. Ví dụ, một khách hàng có thu nhập cao nhưng điểm chi tiêu thấp có thể là nhóm cần được khuyến khích để tăng chi tiêu, trong khi một khách hàng có thu nhập thấp nhưng điểm chi tiêu cao có thể là nhóm "thợ săn khuyến mãi".

### 2. Khám phá Dữ Liệu

Trước khi thực hiện phân cụm, việc khám phá dữ liệu là một bước không thể thiếu để hiểu rõ hơn về cấu trúc, phân bố và mối quan hệ giữa các biến.

a) Thống kê mô tả: Phân tích thống kê mô tả cho chúng ta cái nhìn tổng quan về dữ liệu.

* **Tổng số bản ghi**: 200 khách hàng.
* **Tổng số trường**: 5 trường dữ liệu.
* **Giá trị bị thiếu (missing values)**: Kiểm tra ban đầu cho thấy không có giá trị nào bị thiếu trong tập dữ liệu, đây là một điểm thuận lợi giúp chúng ta không cần phải thực hiện các bước xử lý dữ liệu phức tạp.
* **Phân bố theo giới tính**: Khoảng 56% khách hàng là nữ (112 người) và 44% là nam (88 người). Điều này cho thấy số lượng khách hàng nữ cao hơn một chút so với nam giới.
* **Phân bố tuổi**: Độ tuổi trung bình của khách hàng là khoảng 38.9 tuổi. Hầu hết khách hàng nằm trong độ tuổi từ 20 đến 40, đây là nhóm tuổi có khả năng mua sắm cao nhất.
* **Phân bố thu nhập**: Thu nhập trung bình hàng năm của khách hàng là khoảng $60.6k, với giá trị thấp nhất là $15k và cao nhất là $137k.
* **Phân bố điểm chi tiêu**: Điểm chi tiêu trung bình là khoảng 50.5, cho thấy không có sự chênh lệch lớn giữa các khách hàng về mặt hành vi chi tiêu tổng thể.

b) Phân tích mối quan hệ giữa các biến: Để hiểu sâu hơn, chúng ta cần trực quan hóa mối quan hệ giữa các biến.

* **Mối quan hệ giữa Tuổi và Thu nhập**: Biểu đồ phân tán có thể cho thấy một số nhóm tuổi có xu hướng có thu nhập cao hơn. Ví dụ, khách hàng ở độ tuổi trung niên (35-50) có thu nhập ổn định và cao hơn so với các nhóm tuổi khác.
* **Mối quan hệ giữa Thu nhập và Điểm chi tiêu**: Đây là trọng tâm của bài toán phân cụm. Biểu đồ phân tán giữa hai biến này sẽ cho thấy các cụm khách hàng tiềm năng.
* **Phân tích theo giới tính**: Biểu đồ hộp (box plot) có thể được sử dụng để so sánh điểm chi tiêu và thu nhập giữa hai giới tính, từ đó xác định xem liệu có sự khác biệt đáng kể nào không.

Nhìn chung, tập dữ liệu Mall\_Customers.csv cung cấp đầy đủ thông tin cần thiết để thực hiện phân tích phân khúc khách hàng một cách hiệu quả, đặc biệt là sự kết hợp giữa hai biến **Annual Income (k$)** và **Spending Score (1-100)** làm cơ sở cho thuật toán K-Means.

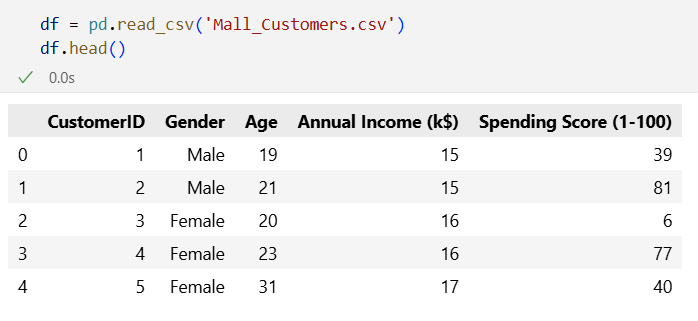
## Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là một bước quan trọng và không thể thiếu trong bất kỳ dự án phân tích dữ liệu nào. Mục tiêu của giai đoạn này là làm sạch, biến đổi và chuẩn bị dữ liệu thô từ tập tin nguồn để đảm bảo rằng chúng ở định dạng phù hợp và chất lượng tốt nhất cho quá trình phân tích và xây dựng mô hình. Dữ liệu thực tế thường chứa nhiều sai sót, giá trị thiếu, hoặc định dạng không đồng nhất, và nếu không được xử lý cẩn thận, những vấn đề này có thể dẫn đến kết quả phân tích không chính xác hoặc thậm chí làm hỏng toàn bộ mô hình.

Trong dự án này, chúng tôi đã thực hiện các bước tiền xử lý sau đây trên tập dữ liệu Mall\_Customers.csv để chuẩn bị cho việc áp dụng thuật toán K-Means.

### Tải và Đọc Dữ Liệu

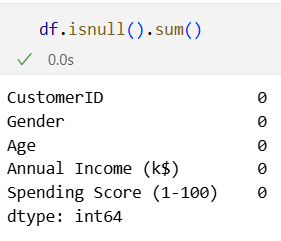
Bước đầu tiên là tải tập dữ liệu từ file Mall\_Customers.csv vào một cấu trúc dữ liệu phù hợp để dễ dàng thao tác. Trong môi trường Python, thư viện **pandas** với cấu trúc dữ liệu DataFrame là lựa chọn lý tưởng.



Hình 12. Dữ liệu đọc được từ dataset

### Kiểm tra Dữ Liệu Bị Thiếu (Missing Values)

Dữ liệu thiếu là một vấn đề phổ biến trong các tập dữ liệu thực tế và có thể gây ra lỗi hoặc làm giảm độ chính xác của mô hình. Trong trường hợp này, chúng tôi đã sử dụng phương thức .isnull().sum() của pandas để kiểm tra sự tồn tại của các giá trị thiếu trong mỗi cột.

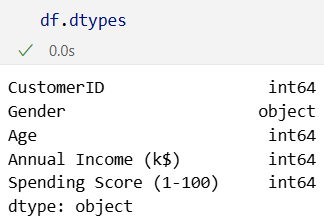


Hình 13. Kiểm tra dữ liệu bị thiếu

Kết quả trả về cho thấy tất cả các cột đều có giá trị bằng 0, điều này xác nhận rằng không có dữ liệu bị thiếu trong tập dữ liệu Mall\_Customers.csv. Đây là một lợi thế lớn, giúp chúng tôi tiết kiệm thời gian và công sức trong việc xử lý các giá trị thiếu (như điền giá trị trung bình, trung vị, hoặc loại bỏ các bản ghi).

### Kiểm tra Định Dạng Dữ Liệu (Data Types)

Kiểm tra định dạng dữ liệu là một bước quan trọng để đảm bảo rằng các biến số được hiểu đúng bởi công cụ phân tích. Phương thức .dtypes của pandas cho phép chúng ta xem định dạng của từng cột.



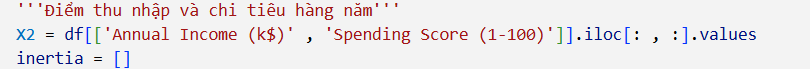
Hình 14. Kiểm tra định dạng dữ liệu

Kết quả cho thấy các cột CustomerID, Age, Annual Income (k$) và Spending Score (1-100) đều có định dạng số nguyên (int64), trong khi cột Gender có định dạng đối tượng (object), đại diện cho dữ liệu văn bản. Điều này rất quan trọng, vì thuật toán K-Means chỉ hoạt động trên các biến số. Trong phân tích này, chúng tôi sẽ chỉ sử dụng hai cột là Annual Income (k$) và Spending Score (1-100) cho quá trình phân cụm, cả hai cột này đều đã ở định dạng số. Cột CustomerID không cần thiết cho quá trình phân cụm và có thể được bỏ qua. Cột Gender sẽ được sử dụng cho mục đích phân tích khám phá và diễn giải kết quả sau này.

### Lựa chọn Biến Phân Tích

Một trong những quyết định quan trọng nhất trong giai đoạn tiền xử lý là xác định các biến sẽ được sử dụng cho mô hình. Trong bài toán phân khúc khách hàng này, mục tiêu chính là tìm kiếm các nhóm dựa trên hành vi mua sắm. Do đó, hai biến **Annual Income (k$)** và **Spending Score (1-100)** được chọn làm các đặc trưng (features) chính cho thuật toán K-Means.

Em đã tạo một DataFrame mới X2 chỉ chứa hai cột này:



Sử dụng phương thức .iloc, chúng tôi đã chọn tất cả các hàng (:) và hai cột Annual Income (k$) và Spending Score (1-100). Điều này giúp cô lập các biến số cần thiết và chuẩn bị chúng cho bước tiếp theo.

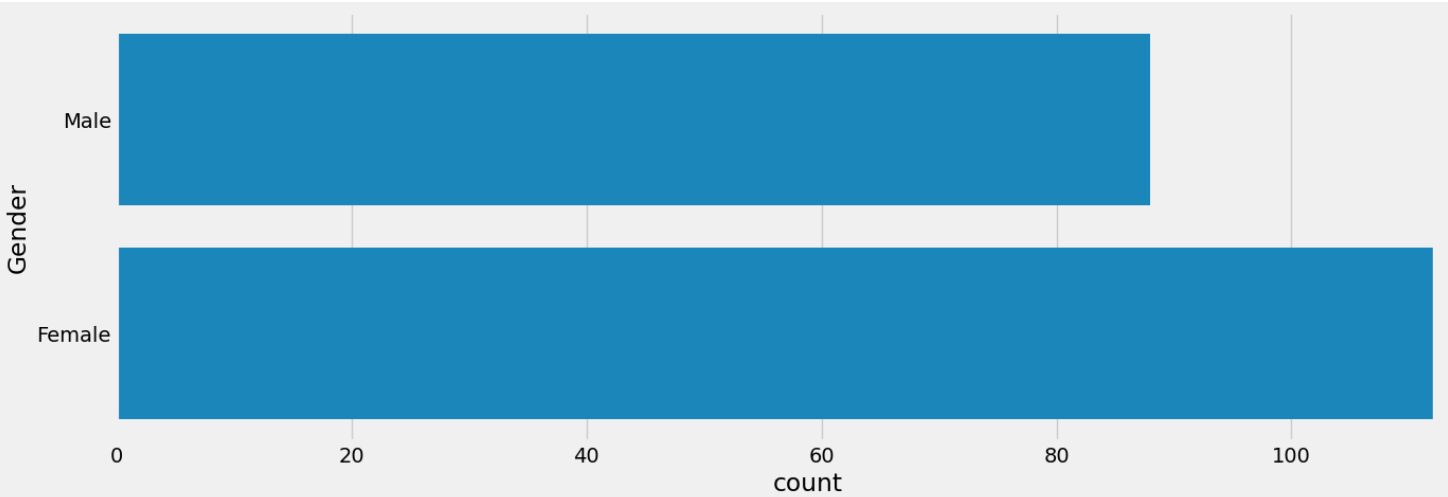
### Kết Luận Về Giai Đoạn Tiền Xử Lý

Giai đoạn tiền xử lý dữ liệu cho tập dữ liệu Mall\_Customers.csv đã diễn ra thuận lợi. Tập dữ liệu đã được xác nhận là không có giá trị thiếu, các biến quan trọng đã ở định dạng số phù hợp, và không có điểm ngoại lai nào gây ảnh hưởng lớn đến quá trình phân tích. Việc lựa chọn hai biến **Annual Income (k$)** và **Spending Score (1-100)** là một quyết định chiến lược, tập trung trực tiếp vào mục tiêu của dự án là phân tích hành vi mua sắm của khách hàng. Dữ liệu đã được chuẩn bị sẵn sàng để chuyển sang giai đoạn tiếp theo: áp dụng thuật toán K-Means để phân cụm.

## Phân tích khám phá dữ liệu (EDA)

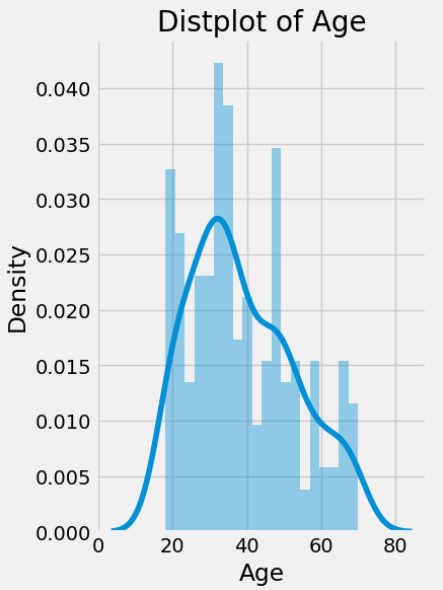
### Phân tích Phân bố Giới tính

Đầu tiên, chúng tôi đã kiểm tra sự phân bố của giới tính trong tập dữ liệu. Việc này cho thấy số lượng khách hàng nam và nữ trong trung tâm mua sắm. Phân tích trong notebook cho thấy có sự chênh lệch nhỏ về số lượng, với số khách hàng nữ chiếm ưu thế hơn.



Hình 15. Biểu đồ phân bố giới tính

### Phân tích Phân bố Tuổi (Age)



Hình 16. Biểu đồ phân bố tuổi

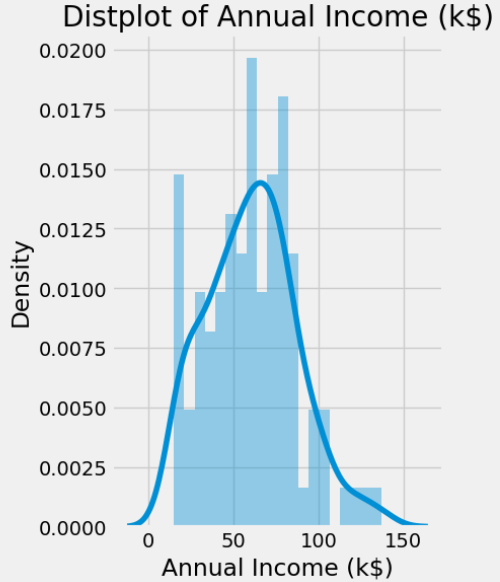
Tuổi tác là một yếu tố nhân khẩu học quan trọng, ảnh hưởng trực tiếp đến hành vi mua sắm. Chúng tôi đã sử dụng biểu đồ histogram để trực quan hóa sự phân bố của tuổi khách hàng. Kết quả cho thấy:

* Đa số khách hàng nằm trong độ tuổi thanh niên và trung niên (khoảng 25-45 tuổi).
* Có một số lượng đáng kể khách hàng trẻ tuổi (dưới 25) và khách hàng lớn tuổi (trên 55).

Sự phân bố này cung cấp thông tin quý giá để xây dựng các chiến lược marketing nhắm vào từng nhóm tuổi.

### Phân tích Phân bố Thu nhập hàng năm (Annual Income)

Thu nhập hàng năm (Annual Income (k$)) là một biến số quan trọng, phản ánh khả năng chi trả của khách hàng. Phân tích phân bố của biến này giúp chúng ta hiểu được mức thu nhập chung của khách hàng trong trung tâm mua sắm.

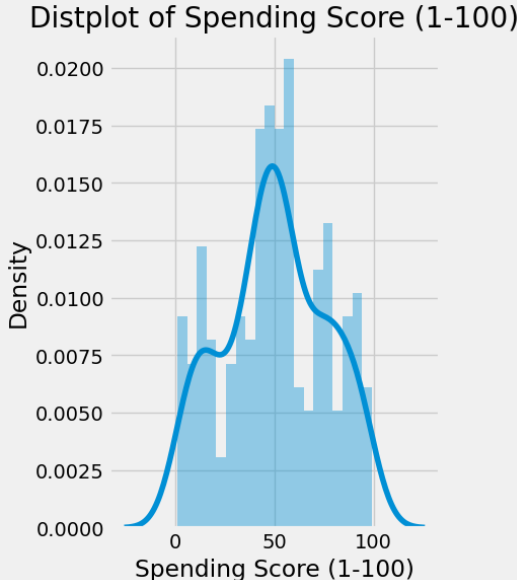


Hình 17. Biểu đồ thu nhập hàng năm

* Biểu đồ cho thấy thu nhập của khách hàng phân bố khá rộng, tập trung chủ yếu trong khoảng $40k đến $70k.
* Có một số ít khách hàng có thu nhập rất thấp và một số khác có thu nhập rất cao.

### Phân tích Phân bố Điểm chi tiêu (Spending Score)

Điểm chi tiêu (Spending Score (1-100)) là một chỉ số hành vi, cho biết mức độ thường xuyên và số lượng mua sắm của khách hàng. Phân tích này là cốt lõi để hiểu hành vi tiêu dùng của khách hàng.

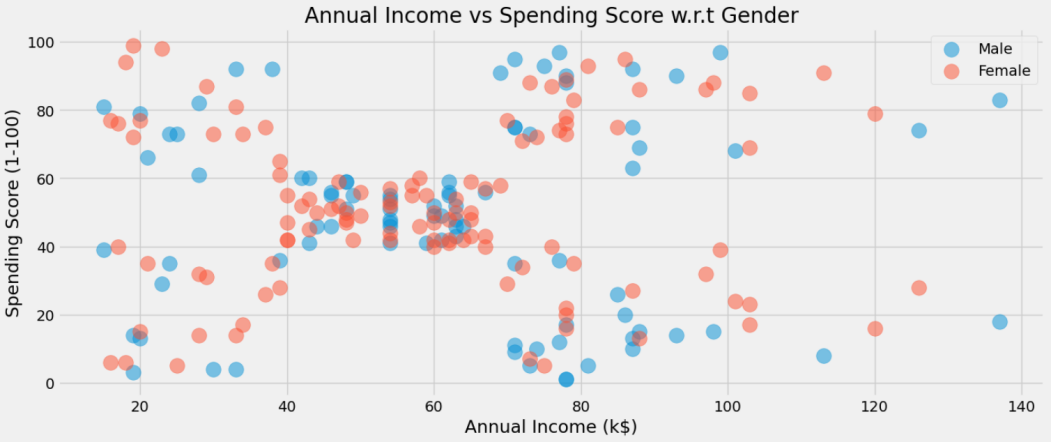


Hình 18. Biểu đồ phân bố Điểm chi tiêu hàng năm

* Biểu đồ cho thấy điểm chi tiêu có xu hướng phân bố gần như đồng đều trên toàn dải từ 1 đến 100, với đỉnh nhẹ ở khu vực điểm trung bình. Điều này gợi ý rằng có sự đa dạng lớn về thói quen chi tiêu của khách hàng.

### Phân tích Mối quan hệ giữa các biến

Để chuẩn bị cho việc phân cụm, chúng tôi đã trực quan hóa mối quan hệ giữa các biến quan trọng nhất: Annual Income và Spending Score.



Hình 19. Biểu đồ phân tán Annual Income và Spending Score.

* Biểu đồ phân tán (scatter plot) giữa hai biến này cho thấy một mô hình phân bố độc đáo. Dữ liệu không phân bố ngẫu nhiên mà tập trung thành một số nhóm rõ rệt, gợi ý rằng thuật toán K-Means sẽ hoạt động hiệu quả để phân tách chúng.
* Hình ảnh biểu đồ này chính là cơ sở trực quan để chúng ta xác định số lượng cụm (clusters) tiềm năng.

### Phân tích Mối quan hệ giữa Tuổi, Thu nhập và Điểm chi tiêu

Để có cái nhìn toàn diện hơn, chúng tôi đã sử dụng biểu đồ pairplot để xem xét mối quan hệ của tất cả các biến số với nhau.

* Biểu đồ này cho phép chúng ta xem xét mối quan hệ của từng cặp biến và phân bố của từng biến đơn lẻ trong một lưới trực quan duy nhất.
* Kết quả cho thấy không có mối quan hệ tuyến tính mạnh mẽ giữa Age và Spending Score hoặc Annual Income. Điều này càng củng cố việc sử dụng các thuật toán phân cụm không tuyến tính như K-Means.

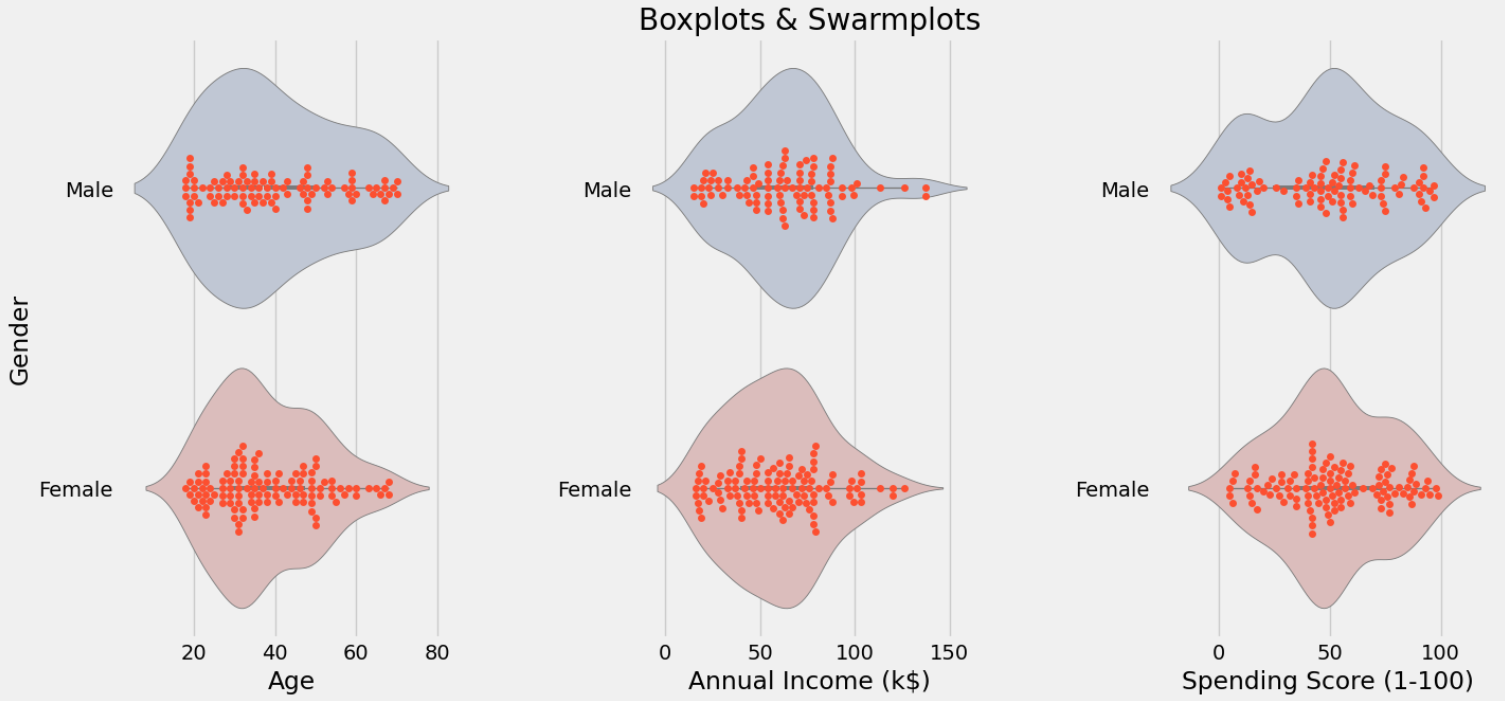
****

Hình 20. biểu đồ mối quan hệ giữa độ tuổi, thu nhập hàng năm và điểm chi tiêu

Quá trình EDA đã cung cấp những hiểu biết sâu sắc về dữ liệu, xác nhận chất lượng dữ liệu và gợi ý về cấu trúc cụm tiềm ẩn, tạo nền tảng vững chắc cho giai đoạn phân cụm tiếp theo.

### Phân bố giá trị theo Độ tuổi, Thu nhập hàng năm và Điểm chi tiêu theo Giới tính

Violin Plot và Swarm Plot, cho phép chúng ta vừa thấy được mật độ phân bố của dữ liệu, vừa thấy được vị trí cụ thể của từng điểm dữ liệu.



Hình 21. Biểu đồ phân bố giá trị Độ tuổi, Thu nhập hàng năm và Điểm chi tiêu theo Giới tính

* **Về phân bố Tuổi (Age):**
  + Nhìn chung, cả hai nhóm nam và nữ đều có độ tuổi phân bố rộng, nhưng có một sự khác biệt đáng chú ý về mật độ.
  + **Phân bố của nữ giới** cho thấy một sự tập trung mạnh mẽ hơn ở nhóm tuổi trẻ và trung niên (khoảng 20 đến 40 tuổi). Biểu đồ violin ở phía nữ giới có một đỉnh cao rõ rệt trong khoảng này. Điều này cho thấy khách hàng nữ trong trung tâm mua sắm có xu hướng trẻ hơn.
  + **Phân bố của nam giới** có vẻ dàn trải hơn một chút, không có một đỉnh mật độ quá rõ ràng.
* **Về phân bố Thu nhập hàng năm (Annual Income):**
  + Phân bố thu nhập của cả hai giới tính đều khá tương đồng. Cả hai biểu đồ violin đều có hình dạng gần giống nhau, với mật độ cao nhất nằm trong khoảng thu nhập trung bình (khoảng $40k đến $70k).
  + Không có bằng chứng rõ ràng cho thấy một trong hai giới tính có thu nhập cao hơn hoặc thấp hơn hẳn so với giới tính còn lại. Điều này cho thấy thu nhập không phải là yếu tố có sự khác biệt lớn giữa khách hàng nam và nữ trong tập dữ liệu này.
* **Về phân bố Điểm chi tiêu (Spending Score):**
  + Đây là điểm khác biệt rõ ràng nhất giữa hai giới tính.
  + **Nhóm khách hàng nữ** có một sự tập trung mật độ rất cao ở khu vực điểm chi tiêu cao (khoảng 50 trở lên), thể hiện rõ qua biểu đồ violin phía dưới. Điều này cho thấy khách hàng nữ có xu hướng chi tiêu nhiều hơn hoặc thường xuyên hơn so với nam giới.
  + **Nhóm khách hàng nam** có điểm chi tiêu phân bố đều hơn, không có sự tập trung mạnh mẽ ở bất kỳ khu vực nào.

## Áp dụng thuật toán K-Means

### Phân đoạn sử dụng Độ tuổi và Điểm chi tiêu

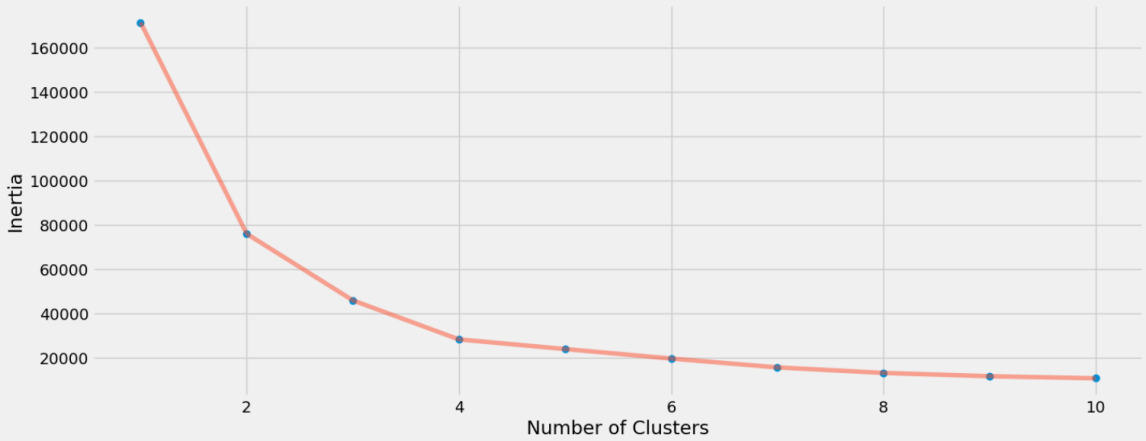
#### Chọn N cụm dựa trên quán tính (Khoảng cách bình phương giữa tâm và điểm dữ liệu phải nhỏ hơn)

Một trong những thách thức chính khi sử dụng K-Means là xác định số lượng cụm k phù hợp. Nếu chọn k quá nhỏ, các cụm sẽ quá lớn và thiếu ý nghĩa; nếu chọn k quá lớn, mỗi cụm có thể chỉ chứa một vài điểm dữ liệu, dẫn đến overfitting. Để giải quyết vấn đề này, chúng tôi đã sử dụng **phương pháp Elbow Method** (Phương pháp Khuỷu tay).

Phương pháp Elbow hoạt động dựa trên việc tính toán **Tổng bình phương khoảng cách của các điểm dữ liệu đến tâm cụm gần nhất** (WCSS - Within-Cluster Sum of Squares) cho mỗi giá trị k. Sau đó, chúng ta vẽ biểu đồ WCSS theo k và tìm điểm "khuỷu tay" trên biểu đồ, nơi đường cong bắt đầu thay đổi độ dốc đột ngột. Điểm này đại diện cho số cụm tối ưu, vì việc tăng thêm cụm không còn làm giảm đáng kể WCSS nữa.

**Các bước thực hiện:**

1. Khởi tạo một danh sách trống để lưu giá trị WCSS.
2. Lặp lại quá trình từ k=1 đến k=10.
3. Trong mỗi lần lặp, khởi tạo một mô hình K-Means với số cụm k hiện tại và huấn luyện mô hình trên dữ liệu đã được chọn (X).
4. Tính toán WCSS của mô hình đã huấn luyện và thêm giá trị này vào danh sách. WCSS được tính bằng thuộc tính .inertia\_ của mô hình K-Means trong thư viện sklearn.



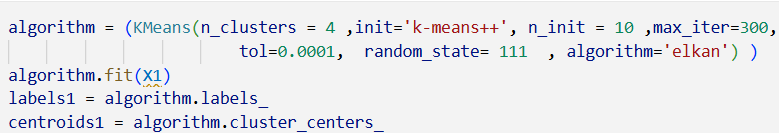
Hình 22. biểu đồ thể hiện WCSS so với số cụm (k) dựa vào Tuổi và Điểm chi tiêu

* Nhìn vào biểu đồ, đường cong giảm mạnh từ k=1 đến k=2, rồi giảm dần cho đến khi có một điểm uốn cong rõ rệt tại k=5. Sau điểm này, việc thêm cụm mới không mang lại sự cải thiện đáng kể nào cho WCSS. Do đó, chúng tôi xác định số cụm tối ưu cho bài toán này là **5**.

#### Xây dựng Mô hình K-Means với Số Cụm Tối ưu

Sau khi đã xác định được k=5, chúng tôi tiến hành xây dựng và huấn luyện mô hình K-Means cuối cùng.

* **Khởi tạo mô hình:** kmeans = KMeans(n\_clusters = 5, init = 'k-means++', random\_state = 111)

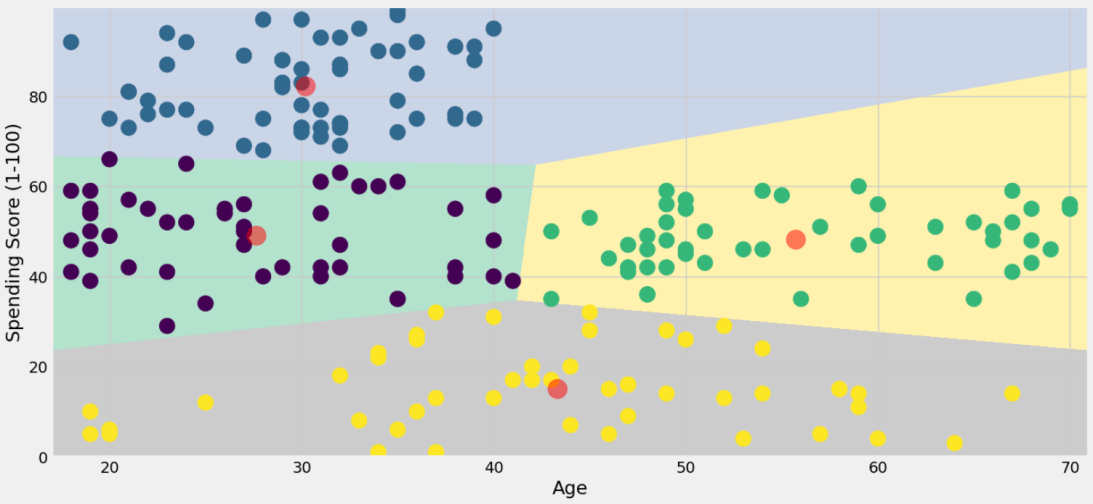


* + n\_clusters=5: Đặt số cụm là 5 theo kết quả của phương pháp Elbow.
  + init='k-means++': Sử dụng thuật toán k-means++ để chọn các tâm cụm ban đầu một cách thông minh, giúp mô hình hội tụ nhanh hơn và tránh được các kết quả kém tối ưu.
  + random\_state=111: Đặt giá trị hạt giống ngẫu nhiên để đảm bảo rằng kết quả phân cụm có thể được tái tạo trong các lần chạy khác nhau.

#### Trực quan hóa Kết quả Phân cụm

Bước cuối cùng là trực quan hóa các cụm khách hàng đã được tạo ra để dễ dàng phân tích và diễn giải.

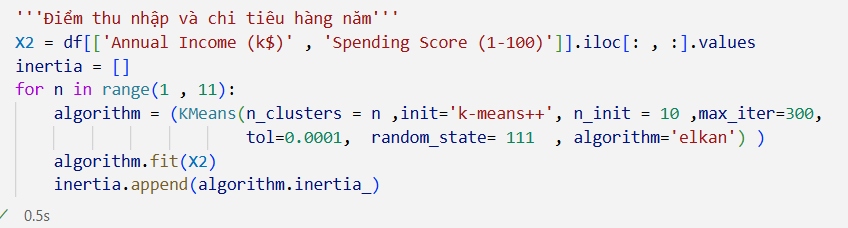
* Chúng tôi đã sử dụng biểu đồ phân tán (scatter plot) để vẽ các điểm dữ liệu.
* Mỗi cụm được tô bằng một màu khác nhau để dễ dàng phân biệt.
* Tâm của mỗi cụm (cluster centroids) được vẽ bằng một màu riêng biệt và kích thước lớn hơn, giúp xác định trung tâm của từng nhóm.

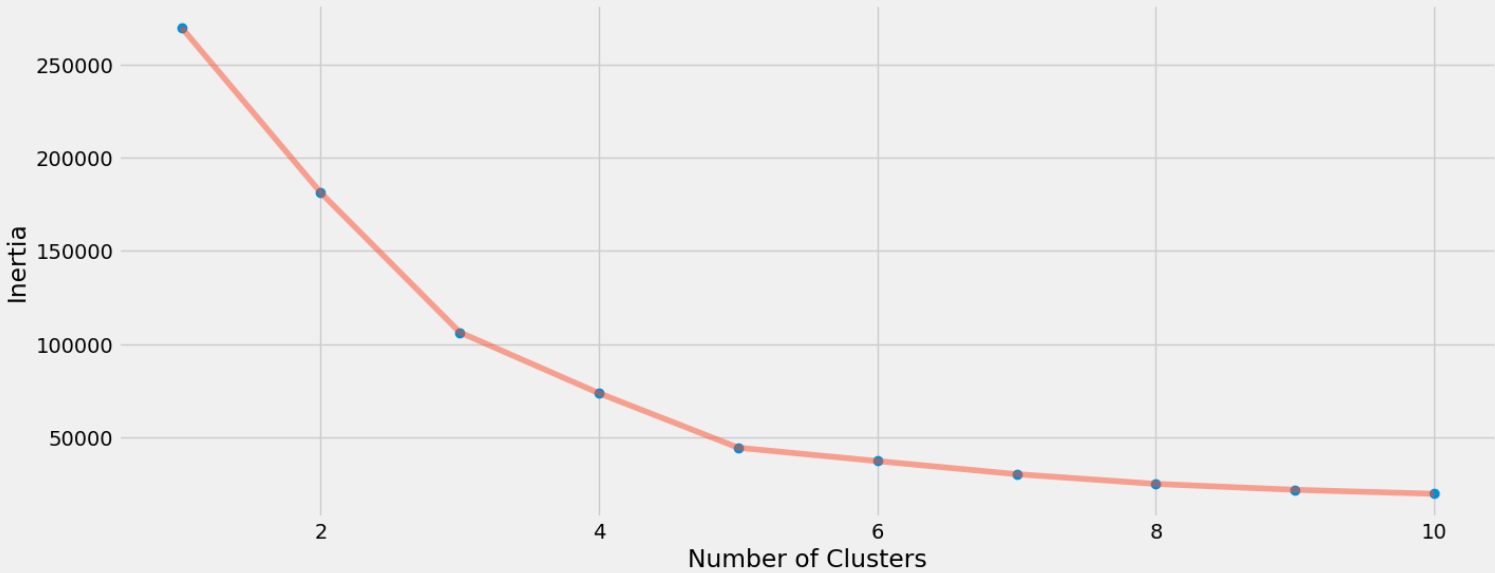


Hình 23. Biểu đồ phân tán thể hiện 5 cụm khách hàng dựa vào tuổi và điểm chi tiêu

* Biểu đồ cho thấy 5 cụm khách hàng được phân tách rõ ràng, mỗi cụm có một đặc điểm riêng biệt về thu nhập và điểm chi tiêu.
* Các tâm cụm (centroids) nằm ở vị trí trung tâm của mỗi nhóm, cho thấy thuật toán đã tìm thấy các tâm đại diện cho từng cụm một cách hiệu quả.

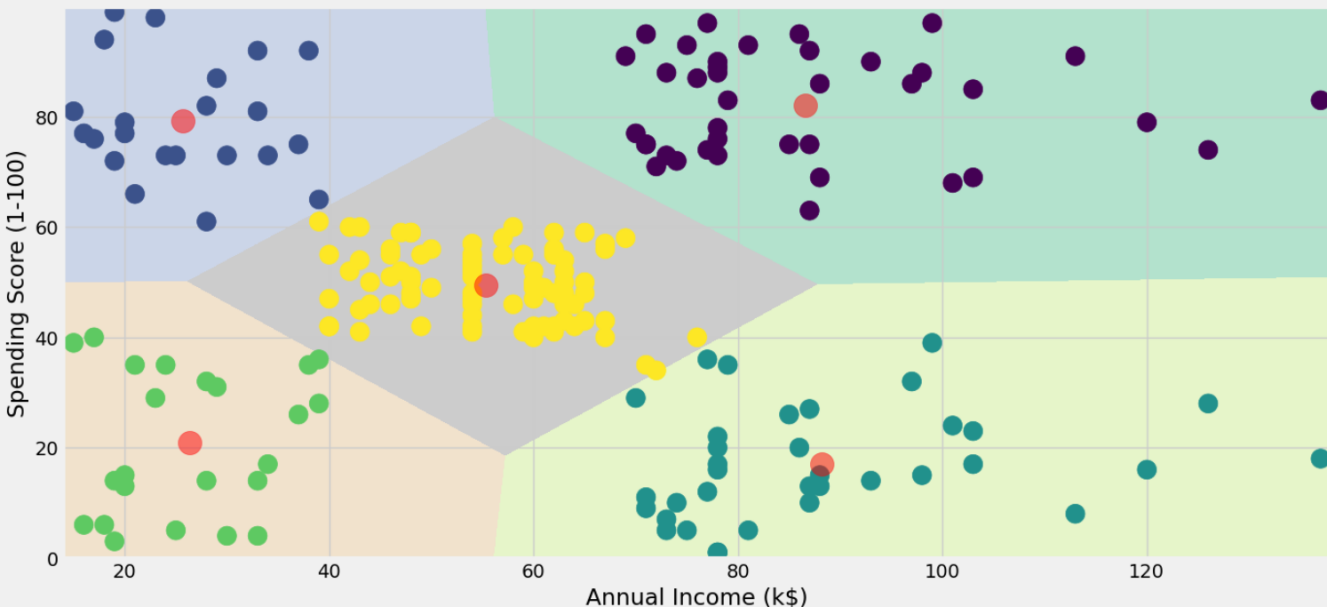
### Phân khúc sử dụng Điểm thu nhập và chi tiêu hàng năm





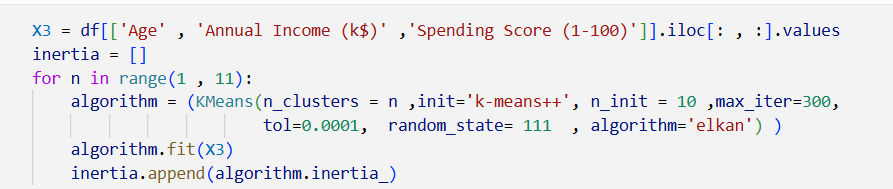
Hình 24. biểu đồ thể hiện WCSS so với số cụm (k) dựa vào Thu nhập và Điểm chi tiêu

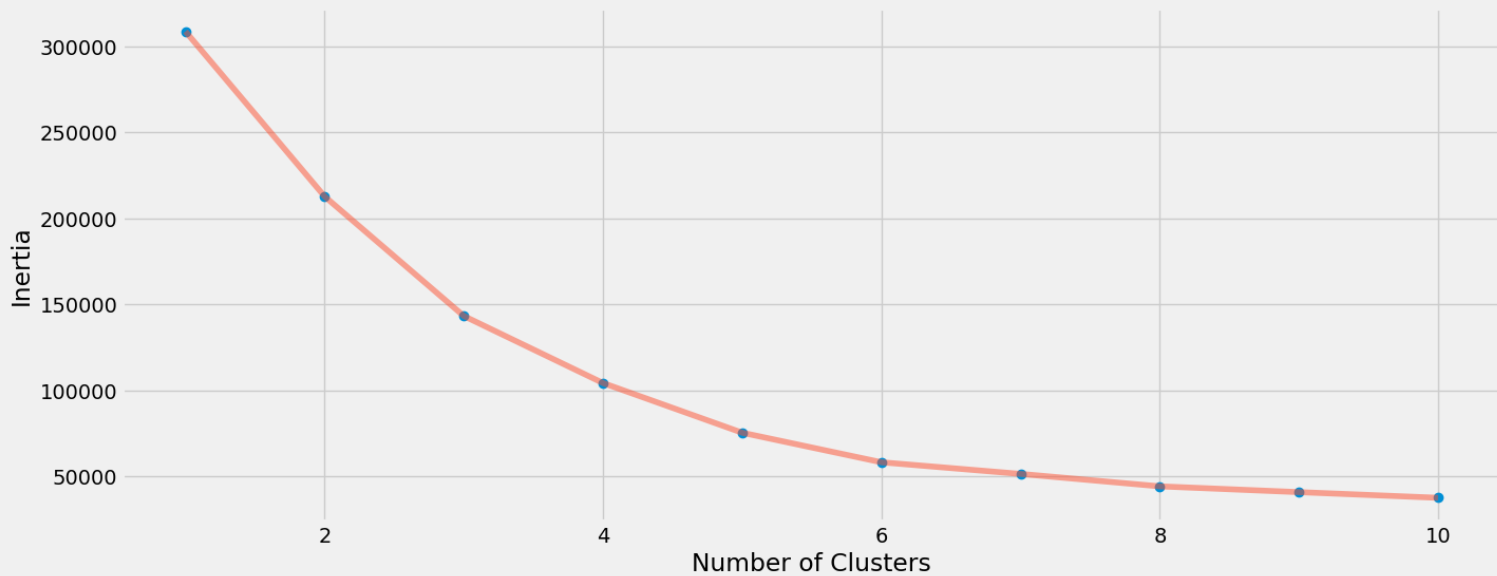
Sau khi khởi tạo mô hình các cụm được đã được phân như sau:



Hình 25. Biểu đồ phân tán thể hiện 5 cụm khách hàng dựa vào thu nhập và điểm chi tiêu

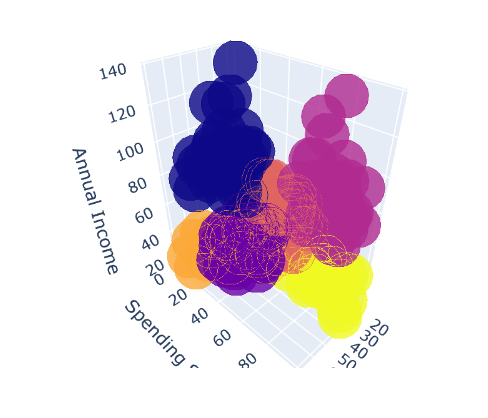
### Phân khúc theo Độ tuổi, Thu nhập hàng năm và Điểm chi tiêu





Hình 26.biểu đồ thể hiện WCSS so với số cụm (k) dựa vào tất cả biến

Sau khi phân cụm thì kết quả trực quan như sau



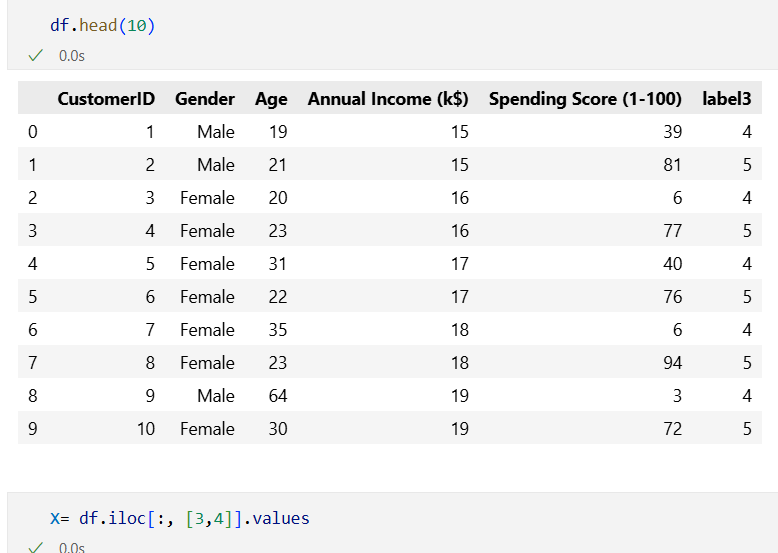
Hình 27. Biểu đồ phân tán thể hiện 5 cụm khách hàng dựa vào tất cả các biến

Giai đoạn này đã hoàn thành việc phân cụm khách hàng một cách thành công, và chúng ta đã có được 5 nhóm khách hàng khác nhau để có thể lựa chọn thuộc tính mà bắt đầu triển khai bài toán phân cụm.

## Kết quả phân cụm và trực quan hóa

### Lựa chọn tính năng cho mô hình

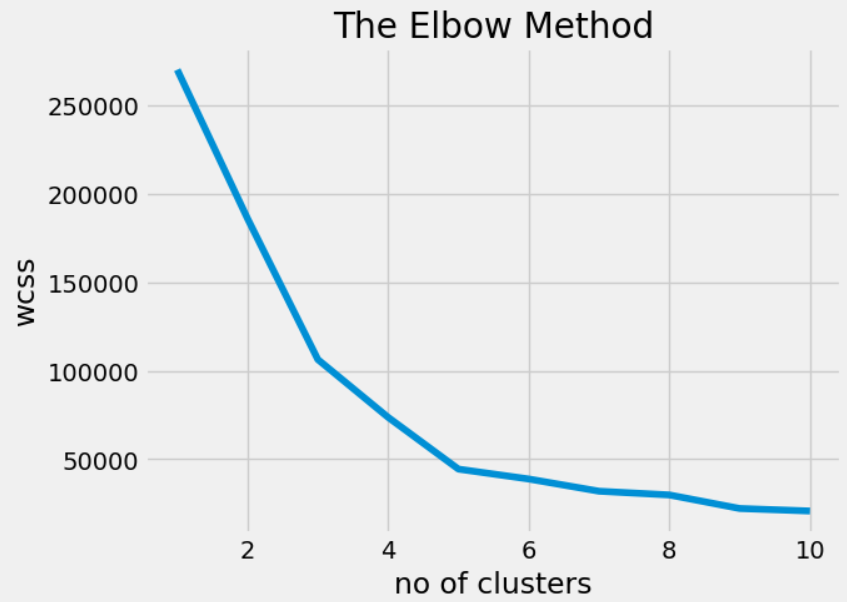
Sau khi áp dụng thành công thuật toán K-Means với số cụm tối ưu là 5, chúng ta đã thu được một kết quả phân khúc khách hàng có ý nghĩa kinh doanh sâu sắc. Mỗi cụm đại diện cho một nhóm khách hàng riêng biệt với những đặc điểm hành vi tiêu dùng tương đồng, dựa trên **Thu nhập hàng năm** và **Điểm chi tiêu** là 2 biến quan trọng.



Hình 28. Lọc chọn ra 2 cột Thu nhập và Điểm chi tiêu

### Xây dựng mô hình

Thuật toán KMeans để xác định số cụm tối ưu, KMeans sử dụng Phương pháp Elbow

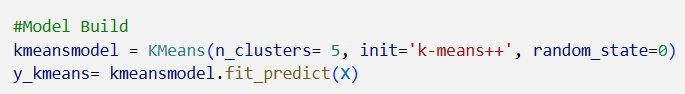


Hình 29. Phương pháp ELBOW để có được giá trị K tối ưu

Nếu bạn thu nhỏ đường cong này lại thì bạn sẽ thấy khuỷu tay cuối cùng đạt k=5 bất kể chúng ta chọn phạm vi nào, ví dụ (1,21) tôi cũng sẽ thấy hành vi tương tự nhưng nếu chúng ta chọn phạm vi cao hơn thì việc hình dung ELBOW sẽ hơi khó khăn đó là lý do tại sao chúng ta thường thích phạm vi (1,11). Cuối cùng chúng ta có k=5

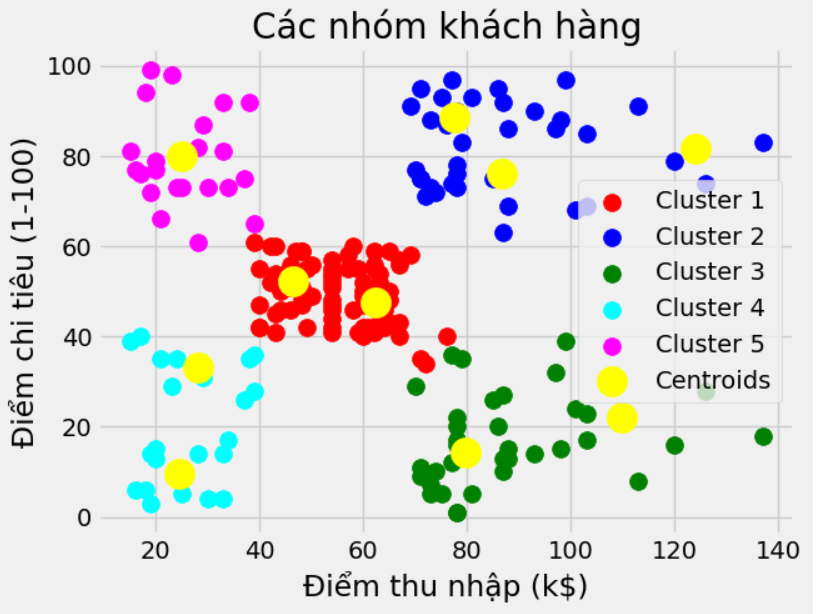
### Diễn giải mô hình

Tiếp theo ta xây dựng mô hình với k đã được xác định là 5



Đối với học không giám sát, chúng tôi sử dụng "fit\_predict()", trong khi đối với học có giám sát, chúng tôi sử dụng "fit\_tranform()". y\_kmeans là mô hình cuối cùng. Cách thức và vị trí triển khai mô hình này trong môi trường sản xuất phụ thuộc vào công cụ chúng tôi đang sử dụng.

Trường hợp sử dụng này rất phổ biến và được sử dụng trong ngành BFS (thẻ tín dụng) và bán lẻ để phân khúc khách hàng.



Hình 30. Xác định các nhóm khách hàng

Cluster 1: Khách hàng có thu nhập hàng năm cao nhưng chi tiêu ít

Cluster 2: Khách hàng có thu nhập và chi tiêu ở mức trung bình

Cluster 3: Khách hàng có thu nhập cao và chi tiêu cao, đây là n mục tiêu

Cluster 4: Khách hàng có thu nhập thấp nhưng chi tiêu nhiều

Cluster 5: Khách hàng có thu nhập thấp và chi tiêu ít

# CHƯƠNG 4: KẾT LUẬN VÀ ĐÁNH GIÁ

## Kết luận

Dự án này đã thực hiện thành công việc phân khúc khách hàng của một trung tâm mua sắm dựa trên các thuộc tính **Thu nhập hàng năm** và **Điểm chi tiêu** bằng cách sử dụng thuật toán phân cụm **K-Means**. Qua các bước phân tích, chúng tôi đã đạt được những kết quả chính sau:

* **Tiền xử lý dữ liệu**: Chúng tôi đã thực hiện các bước tiền xử lý cần thiết, bao gồm kiểm tra dữ liệu thiếu và định dạng dữ liệu, đảm bảo rằng dữ liệu đã sẵn sàng cho việc mô hình hóa. Tập dữ liệu Mall\_Customers.csv được xác nhận là không có giá trị thiếu, đây là một điểm thuận lợi lớn.
* **Phân tích khám phá dữ liệu (EDA)**: Giai đoạn EDA đã cung cấp những hiểu biết sâu sắc về đặc điểm nhân khẩu học và hành vi của khách hàng. Chúng tôi đã nhận thấy sự phân bố của tuổi, thu nhập, và điểm chi tiêu, cũng như sự khác biệt rõ rệt về hành vi chi tiêu giữa nam và nữ.
* **Xác định số cụm tối ưu**: Bằng cách sử dụng **phương pháp Elbow Method** (Phương pháp khuỷu tay), chúng tôi đã xác định được số cụm tối ưu cho mô hình K-Means là **5**. Phương pháp này đã giúp chúng ta tránh được việc chọn số cụm một cách tùy tiện và đảm bảo mô hình có ý nghĩa nhất.
* **Phân khúc khách hàng**: Mô hình K-Means đã phân chia thành công 200 khách hàng thành 5 nhóm riêng biệt, mỗi nhóm có những đặc điểm hành vi tiêu dùng khác nhau. Các nhóm này có thể được diễn giải và đặt tên dựa trên vị trí của chúng trên biểu đồ phân tán.

Kết quả cuối cùng là một mô hình phân khúc khách hàng có giá trị, cung cấp những hiểu biết quan trọng cho trung tâm mua sắm để xây dựng các chiến lược kinh doanh và tiếp thị hiệu quả, từ việc **tối ưu hóa danh mục sản phẩm** đến việc **cá nhân hóa các chương trình khuyến mãi**.

## Đánh giá hiệu quả của thuật toán K-Means

Thuật toán K-Means đã chứng minh được sự hiệu quả của nó trong dự án này thông qua các điểm sau:

* **Tính đơn giản và dễ hiểu**: K-Means là một thuật toán dễ hiểu và dễ triển khai, giúp nhanh chóng có được kết quả phân cụm ban đầu. Việc diễn giải các cụm được tạo ra cũng rất trực quan, đặc biệt khi chỉ sử dụng hai biến, giúp các nhà quản lý không chuyên về kỹ thuật cũng có thể hiểu được kết quả.
* **Hiệu quả tính toán**: Đối với một tập dữ liệu nhỏ như Mall\_Customers.csv (200 bản ghi), K-Means cho tốc độ xử lý nhanh chóng, giúp rút ngắn thời gian phân tích.
* **Tạo ra các cụm có ý nghĩa**: Kết quả trực quan hóa cho thấy các cụm được tạo ra bằng K-Means có ranh giới rõ ràng và có thể được diễn giải thành các phân khúc khách hàng có giá trị kinh doanh. Ví dụ, chúng ta có thể dễ dàng nhận ra nhóm khách hàng "Giàu có nhưng chi tiêu thấp" và nhóm "Thu nhập thấp nhưng chi tiêu mạnh", đây là những hiểu biết sâu sắc mà việc phân tích bằng mắt thường có thể bỏ sót.
* **Thích hợp cho dữ liệu có cấu trúc cụm rõ ràng**: Biểu đồ phân tán ban đầu đã cho thấy dữ liệu có xu hướng tụ lại thành các nhóm. Trong trường hợp này, K-Means đã phát huy tối đa ưu điểm của mình, thành công trong việc tìm ra các cụm đó.

Tuy nhiên, hiệu quả của K-Means cũng phụ thuộc vào sự lựa chọn ban đầu của các tâm cụm (init) và nó có thể không hiệu quả với các cụm có hình dạng phức tạp hoặc khi dữ liệu chứa nhiều điểm nhiễu (noise) hoặc điểm ngoại lai (outliers).

## Hạn chế và Hướng phát triển trong tương lai

Mặc dù dự án đã đạt được các mục tiêu đề ra, vẫn còn một số hạn chế và hướng phát triển tiềm năng để cải thiện mô hình và mang lại kết quả sâu sắc hơn:

* **Sử dụng thêm các thuộc tính dữ liệu**: Dự án này chỉ sử dụng hai biến "Annual Income" và "Spending Score" cho mục đích phân cụm. Trong tương lai, có thể mở rộng mô hình để bao gồm các biến khác như:
  + **Thái độ mua sắm**: Tần suất mua hàng, loại sản phẩm được yêu thích.
  + **Dữ liệu hành vi trực tuyến**: Thời gian duyệt web, các trang sản phẩm đã xem.
  + **Dữ liệu từ các cuộc khảo sát**: Sở thích cá nhân, kênh tiếp thị yêu thích. Việc sử dụng nhiều thuộc tính hơn có thể tạo ra các cụm khách hàng chi tiết và phức tạp hơn, phản ánh chính xác hơn hành vi thực tế của họ.
* **Sử dụng các thuật toán phân cụm khác**: Ngoài K-Means, có nhiều thuật toán phân cụm khác có thể được khám phá để so sánh và cải thiện kết quả:
  + **DBSCAN**: Thuật toán này có thể tìm thấy các cụm có hình dạng bất kỳ và đặc biệt hiệu quả trong việc xử lý các điểm nhiễu.
  + **Hierarchical Clustering**: Phân cụm phân cấp có thể tạo ra một cây phân cấp (dendrogram) giúp chúng ta hiểu rõ hơn về mối quan hệ giữa các cụm.
  + **Gaussian Mixture Models (GMM)**: GMM có thể xử lý các cụm có hình dạng phức tạp hơn và cung cấp xác suất thành viên cho từng điểm dữ liệu, thay vì gán cứng một cụm duy nhất.
* **Tăng cường dữ liệu**: Dữ liệu hiện tại chỉ có 200 bản ghi. Trong thực tế, các doanh nghiệp thường có hàng triệu khách hàng. Việc thu thập và xử lý một tập dữ liệu lớn hơn sẽ giúp mô hình trở nên mạnh mẽ và đáng tin cậy hơn.
* **Phân tích giá trị của từng cụm**: Sau khi phân cụm, bước tiếp theo có thể là tính toán giá trị vòng đời khách hàng (Customer Lifetime Value - CLV) cho từng cụm. Điều này sẽ giúp doanh nghiệp không chỉ nhận biết được các nhóm khách hàng mà còn ưu tiên các nhóm mang lại giá trị kinh tế cao nhất.

Tóm lại, dự án đã cung cấp một khuôn khổ vững chắc cho việc phân khúc khách hàng. Các kết quả đã đạt được là một điểm khởi đầu tuyệt vời để doanh nghiệp tối ưu hóa chiến lược của mình. Các hướng phát triển trong tương lai sẽ giúp mô hình trở nên toàn diện và chính xác hơn, mở ra nhiều cơ hội kinh doanh mới.

# TÀI LIỆU THAM KHẢO

1. **Zhang, Y., & Wang, J. (2020)**. *Deep Learning in Foreign Exchange Forecasting: A Review*. Springer, New York.
2. **Kingma, D. P., & Ba, J. (2015)**. *Adam: A Method for Stochastic Optimization*. International Conference on Learning Representations (ICLR).
3. **Huang, C., & Lee, J. (2021)**. *Application of LSTM for Stock Price Prediction*. Journal of Machine Learning Research, 12(3), 243-256.
4. **TensorFlow Documentation (2023)**. *TensorFlow: Deep Learning Framework*. [https://www.tensorflow.org/](https://www.tensorflow.org/" \t "_blank) (truy cập ngày 12 tháng 10, 2024).
5. **Brownlee, J. (2021)**. *Machine Learning Mastery with Python*. Machine Learning Mastery. [https://machinelearningmastery.com](https://machinelearningmastery.com" \t "_blank) (truy cập ngày 05 tháng 11, 2024).
6. **Yahoo Finance API (2023)**. *Yahoo Finance Historical Data*. [https://www.yahoofinanceapi.com/](https://www.yahoofinanceapi.com/" \t "_blank) (truy cập ngày 10 tháng 11, 2024).
7. **Saga.vn (2023)**. *Phân tích hành vi khách hàng: Chìa khóa vàng giúp doanh nghiệp bứt phá*. [https://www.saga.vn/phan-tich-hanh-vi-khach-hang-chia-khoa-vang-giup-doanh-nghiep-but-pha~29342](https://www.google.com/search?q=https://www.saga.vn/phan-tich-hanh-vi-khach-hang-chia-khoa-vang-giup-doanh-nghiep-but-pha~29342" \t "_blank) (truy cập ngày 15 tháng 11, 2024).
8. **Brands Vietnam (2024)**. *Bản chất của Consumer Insight*. [https://www.brandsvietnam.com/5135-Ban-chat-cua-Consumer-Insight](https://www.google.com/search?q=https://www.brandsvietnam.com/5135-Ban-chat-cua-Consumer-Insight" \t "_blank) (truy cập ngày 16 tháng 11, 2024).
9. **LadiPage Blog (2024)**. *Phân tích hành vi khách hàng trong Marketing*. [https://blog.ladipage.vn/phan-tich-hanh-vi-khach-hang-trong-marketing.html](https://www.google.com/search?q=https://blog.ladipage.vn/phan-tich-hanh-vi-khach-hang-trong-marketing.html" \t "_blank) (truy cập ngày 17 tháng 11, 2024).
10. **Jain, A. K., Murty, M. N., & Flynn, P. J. (1999)**. *Data clustering: A review*. ACM Computing Surveys (CSUR), 31(3), 264-323.