Phân tích và Phân khúc Khách hàng dựa trên Dữ liệu Marketing

Nhóm 01: Đỗ Kiến Hưng (23133030) Phan Trọng Quí (23133061) Phan Trọng Phú (23133056) Nguyên Văn Quang Duy (23110086)

Table of Contents

```
# --- NẠP CÁC THƯ VIỆN CẦN THIẾT CHO TOÀN BỘ DỰ ÁN ---

# Kiểm tra và cài đặt gói 'pacman' nếu chưa có
# 'pacman' giúp quản lý (kiểm tra, cài đặt, nạp) các gói khác dễ dàng hơn
if (!require("pacman")) install.packages("pacman")
## Loading required package: pacman
## Warning: package 'pacman' was built under R version 4.4.3
# Sử dụng pacman để nạp (và cài đặt nếu cần) các gói
pacman::p_load(
  # Gói cho thao tác dữ liệu
  dplyr,        # Công cụ thao tác dữ liệu mạnh mẽ (filter, mutate, select, group_by, summarise)
  tidyr,        # Giúp làm sạch và định hình lại dữ liệu (pivot_longer, pivot_wider)
  lubridate,    # Xử lý dữ liệu ngày tháng

  # Gói cho trực quan hóa
  ggplot2,      # Hệ thống vẽ đồ thị mạnh mẽ và linh hoạt
  patchwork,    # Ghép nhiều biểu đồ ggplot lại với nhau
  corrplot,     # Vẽ ma trận tương quan
  GGally,       # Chứa hàm ggpairs cho ma trận scatter plot và tương quan

  # Gói cho phân cụm
  cluster,      # Chứa các thuật toán phân cụm như kmeans, silhouette
  factoextra,   # Trực quan hóa kết quả phân cụm, xác định số cụm tối ưu

  # Gói cho mô hình hóa và đánh giá (Học có giám sát)
  caret,        # Công cụ cho chia dữ liệu, tiền xử lý, huấn luyện và đánh giá mô hình
  pROC,         # Vẽ đường cong ROC và tính AUC
```

```
  car,          # Chứa hàm vif() để kiểm tra đa cộng tuyến trong hồi quy

  # Gói cho trình bày bảng biểu đẹp (tùy chọn)
  knitr,        # Hỗ trợ render R Markdown, có hàm kable()
  kableExtra    # Tùy chỉnh bảng kable đẹp hơn
)

# Thiết lập tùy chọn chung cho các chunk R (nếu muốn)
knitr::opts_chunk$set(
  echo = TRUE,      # Hiển thị code R trong output (có thể đổi thành FALSE ở từng chunk nếu cần)
  message = FALSE,  # Ẩn các thông báo (messages)
  warning = FALSE,  # Ẩn các cảnh báo (warnings)
  fig.align = "center" # Căn giữa hình ảnh
)
```

1. Tóm tắt (Abstract)

Dự án này phân tích dữ liệu khách hàng từ bộ "marketing_campaign.csv" (Kaggle) bằng ngôn ngữ R nhằm khám phá hành vi mua sắm và phân khúc khách hàng. Sau khi tiền xử lý và thực hiện phân tích dữ liệu khám phá (EDA), ba mô hình chính được áp dụng: Phân cụm K-Means giúp xác định 4 phân khúc khách hàng với các đặc điểm riêng biệt. Hồi quy Logistic được sử dụng để dự đoán khả năng khách hàng chấp nhận ưu đãi marketing (biến Response), cho thấy các yếu tố như lịch sử tương tác và đặc điểm gia đình có ảnh hưởng đáng kể. Cuối cùng, mô hình Hồi quy Tuyến tính Đa biến được xây dựng để dự đoán tổng chi tiêu của khách hàng (dưới dạng logarit), làm nổi bật vai trò của thu nhập và các kênh mua hàng cụ thể. Các kết quả này cung cấp những hiểu biết giá trị, hỗ trợ doanh nghiệp xây dựng chiến lược marketing cá nhân hóa và hiệu quả hơn.

2. Giới thiệu (Introduction)

Trong môi trường kinh doanh cạnh tranh hiện nay, việc hiểu rõ khách hàng là yếu tố then chốt để thành công. Phân tích hành vi và phân khúc khách hàng cho phép doanh nghiệp tối ưu hóa chiến lược marketing, phát triển sản phẩm phù hợp và nâng cao sự hài lòng của khách hàng. Bộ dữ liệu "Customer Personality Analysis" từ Kaggle, với thông tin đa dạng về nhân khẩu học, lịch sử mua sắm và tương tác marketing của khách hàng, cung cấp một cơ hội quý giá để khám phá các khía cạnh này.

Nghiên cứu này được thực hiện với các mục tiêu chính sau:

- Xác định và mô tả các nhóm khách hàng (phân khúc) có đặc điểm tương đồng trong tập dữ liệu.
- Tìm hiểu các yếu tố ảnh hưởng đến quyết định chấp nhận ưu đãi marketing của khách hàng (biến Response).
- Xây dựng mô hình dự đoán tổng chi tiêu của khách hàng và xác định các yếu tố tác động đến mức chi tiêu này.

Để đạt được các mục tiêu trên, dự án sẽ sử dụng ngôn ngữ R để thực hiện các bước: Tiền xử lý dữ liệu (làm sạch, tạo biến mới) Phân tích dữ liệu khám phá thông qua trực quan hóa Triển khai ba mô hình học máy: Phân cụm K-Means, Hồi quy Logistic, và Hồi quy Tuyến tính Đa biến.

Báo cáo sẽ trình bày quy trình, kết quả phân tích và các đề xuất ứng dụng.

3. Dữ liệu (Data)

Phần này mô tả bộ dữ liệu được sử dụng và các bước tiền xử lý cần thiết để chuẩn bị cho quá trình phân tích và mô hình hóa.

3.1 Nguồn dữ liệu

Dữ liệu được sử dụng trong nghiên cứu này là bộ dữ liệu "Customer Personality Analysis" được công bố công khai trên nền tảng Kaggle. Bộ dữ liệu gốc có thể được truy cập tại [https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis]. File dữ liệu chính là marketing_campaign.csv, chứa thông tin về nhân khẩu học, lịch sử mua hàng và phản hồi chiến dịch của 2240 khách hàng.

```
## Bộ dữ liệu gốc có 2240 quan sát và 29 biến.
## Một vài biến chính ban đầu:
##   Year_Birth  Education Income MntWines Response
## 1       1957 Graduation  58138      635        1
## 2       1954 Graduation  46344       11        0
## 3       1965 Graduation  71613      426        0
```

3.2 Mô tả dữ liệu

Bộ dữ liệu gốc bao gồm r ncol(customers_raw) biến và 2240 quan sát. Các biến số cung cấp thông tin đa dạng về khách hàng, có thể được nhóm thành các loại chính như sau:

- **Nhân khẩu học:** Year_Birth, Education, Marital_Status, Income, Kidhome, Teenhome.
- **Quan hệ với công ty:** Dt_Customer, Recency, Complain.
- **Chi tiêu sản phẩm (2 năm qua):** MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds.
- **Tương tác khuyến mãi:** NumDealsPurchases, AcceptedCmp1 - AcceptedCmp5, Response (biến mục tiêu chính cho phân loại).
- **Kênh mua hàng:** NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth.
- **Biến không sử dụng:** ID, Z_CostContact, Z_Revenue.

```
## Rows: 2,240
## Columns: 29
## $ ID                 <int> 5524, 2174, 4141, 6182, 5324, 7446, 965, 6177, 485…
## $ Year_Birth         <int> 1957, 1954, 1965, 1984, 1981, 1967, 1971, 1985, 19…
## $ Education          <chr> "Graduation", "Graduation", "Graduation", "Graduat…
## $ Marital_Status     <chr> "Single", "Single", "Together", "Together", "Marri…
## $ Income             <int> 58138, 46344, 71613, 26646, 58293, 62513, 55635, 3…
## $ Kidhome            <int> 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1,…
## $ Teenhome           <int> 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,…
## $ Dt_Customer        <chr> "04-09-2012", "08-03-2014", "21-08-2013", "10-02-2…
## $ Recency            <int> 58, 38, 26, 26, 94, 16, 34, 32, 19, 68, 11, 59, 82…
## $ MntWines           <int> 635, 11, 426, 11, 173, 520, 235, 76, 14, 28, 5, 6,…
## $ MntFruits          <int> 88, 1, 49, 4, 43, 42, 65, 10, 0, 0, 5, 16, 61, 2, …
## $ MntMeatProducts    <int> 546, 6, 127, 20, 118, 98, 164, 56, 24, 6, 6, 11, 4…
## $ MntFishProducts    <int> 172, 2, 111, 10, 46, 0, 50, 3, 3, 1, 0, 11, 225, 3…
## $ MntSweetProducts   <int> 88, 1, 21, 3, 27, 42, 49, 1, 3, 1, 2, 1, 112, 5, 1…
## $ MntGoldProds       <int> 88, 6, 42, 5, 15, 14, 27, 23, 2, 13, 1, 16, 30, 14…
## $ NumDealsPurchases  <int> 3, 2, 1, 2, 5, 2, 4, 2, 1, 1, 1, 1, 1, 3, 1, 1, 3,…
## $ NumWebPurchases    <int> 8, 1, 8, 2, 5, 6, 7, 4, 3, 1, 1, 2, 3, 6, 1, 7, 3,…
## $ NumCatalogPurchases <int> 10, 1, 2, 0, 3, 4, 3, 0, 0, 0, 0, 0, 4, 1, 0, 6, 0…
## $ NumStorePurchases  <int> 4, 2, 10, 4, 6, 10, 7, 4, 2, 0, 2, 3, 8, 5, 3, 12,…
## $ NumWebVisitsMonth  <int> 7, 5, 4, 6, 5, 6, 6, 8, 9, 20, 7, 8, 2, 6, 8, 3, 8…
## $ AcceptedCmp3       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,…
## $ AcceptedCmp4       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ AcceptedCmp5       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,…
## $ AcceptedCmp1       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,…
## $ AcceptedCmp2       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ Complain           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
```

```
## $ Z_CostContact      <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,…
## $ Z_Revenue          <int> 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11…
## $ Response           <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,…
```

3.3 Tiền xử lý dữ liệu (Data Preprocessing)

Để đảm bảo chất lượng và tính phù hợp của dữ liệu cho cả ba mô hình, các bước tiền xử lý sau được thực hiện:

3.3.1 Làm sạch dữ liệu (Xử lý NA)

- **Giá trị thiếu (NA):** Kiểm tra cho thấy cột Income có 24 giá trị NA. Do chiếm tỷ lệ nhỏ, các hàng chứa giá trị NA này đã bị loại bỏ.
- **Giá trị ngoại lệ (Outliers):** Sử dụng biểu đồ hộp, các giá trị ngoại lệ trong Year_Birth dẫn đến tuổi > 100 và < 18 cùng với Income > 600000 đã được xác định và loại bỏ khỏi bộ dữ liệu.

```
## Số quan sát còn lại:
## [1] 2212
```

3.3.2 Kỹ thuật đặc trưng (Feature Engineering)

Các biến mới được tạo ra từ dữ liệu gốc để làm giàu thông tin và phục vụ tốt hơn cho các mô hình:

- **Age:** Tuổi của khách hàng (tính đến 2014).
- **total_spent:** Tổng chi tiêu cho 6 loại sản phẩm chính.
- **log_total_spent:** Logarit tự nhiên của (total_spent + 1). Đây sẽ là biến mục tiêu cho mô hình hồi quy tuyến tính.
- **Child_Total:** Tổng số con cái.
- **AcceptedCmp_Total:** Tổng số chiến dịch (1-5) khách hàng đã chấp nhận.
- **Days_Customer:** Số ngày kể từ khi khách hàng đăng ký.

Loại bỏ biến: Các biến không cần thiết hoặc đã được tổng hợp như ID, Year_Birth, Dt_Customer, Kidhome, Teenhome, các biến AcceptedCmp riêng lẻ, Z_CostContact, Z_Revenue được loại bỏ. Các biến chi tiêu thành phần (Mnt…) cũng được loại bỏ khỏi tập dữ liệu cuối cùng sau khi đã tính total_spent và log_total_spent, để tránh rò rỉ thông tin khi dự đoán tổng chi tiêu. Các biến ký tự được chuyển thành kiểu factor.

```
## Bộ dữ liệu cuối cùng (customers_final) có 2212 quan sát và 17 biến.
## Các biến chính trong bộ dữ liệu cuối cùng bao gồm:
```

```
##  [1] "Education"          "Marital_Status"     "Income"
##  [4] "Recency"            "Complain"           "Age"
##  [7] "Child_Total"        "AcceptedCmp_Total"  "Days_Customer"
## [10] "NumDealsPurchases"  "NumWebPurchases"    "NumCatalogPurchases"
## [13] "NumStorePurchases"  "NumWebVisitsMonth"  "Response"
## [16] "log_total_spent"    "total_spent"
```

3.4 Dữ liệu cuối cùng cho phân tích

Sau tiền xử lý, bộ dữ liệu customers_final gồm r nrow(customers_final) khách hàng và r ncol(customers_final) biến đã được chuẩn bị. Dữ liệu này bao gồm các thông tin nhân khẩu học, hành vi mua sắm, tương tác chiến dịch, tổng chi tiêu (gốc và logarit), và biến phản hồi chiến dịch, sẵn sàng cho các bước tiếp theo. Tùy theo yêu cầu của từng mô hình, các bước chuẩn hóa (scaling) hoặc tạo biến giả (dummy variables) sẽ được thực hiện thêm.

```
# Xem cấu trúc cuối cùng
glimpse(customers_final)
## Rows: 2,212
## Columns: 17
## $ Education          <fct> Graduation, Graduation, Graduation, Graduation, Ph…
## $ Marital_Status     <fct> Single, Single, Together, Together, Married, Toget…
## $ Income             <int> 58138, 46344, 71613, 26646, 58293, 62513, 55635, 3…
## $ Recency            <int> 58, 38, 26, 26, 94, 16, 34, 32, 19, 68, 59, 82, 53…
## $ Complain           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ Age                <dbl> 57, 60, 49, 30, 33, 47, 43, 29, 40, 64, 38, 55, 62…
## $ Child_Total        <int> 0, 2, 0, 1, 1, 1, 1, 1, 1, 2, 0, 0, 2, 0, 0, 2, 0,…
## $ AcceptedCmp_Total  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0,…
## $ Days_Customer      <dbl> 663, 113, 312, 139, 161, 293, 593, 417, 388, 108, …
## $ NumDealsPurchases  <int> 3, 2, 1, 2, 5, 2, 4, 2, 1, 1, 1, 1, 3, 1, 1, 3, 2,…
## $ NumWebPurchases    <int> 8, 1, 8, 2, 5, 6, 7, 4, 3, 1, 2, 3, 6, 1, 7, 3, 4,…
## $ NumCatalogPurchases <int> 10, 1, 2, 0, 3, 4, 3, 0, 0, 0, 0, 4, 1, 0, 6, 0, 1…
## $ NumStorePurchases  <int> 4, 2, 10, 4, 6, 10, 7, 4, 2, 0, 3, 8, 5, 3, 12, 3,…
## $ NumWebVisitsMonth  <int> 7, 5, 4, 6, 5, 6, 6, 8, 9, 20, 8, 2, 6, 8, 3, 8, 7…
## $ Response           <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,…
## $ log_total_spent    <dbl> 7.388946, 3.332205, 6.655440, 3.988984, 6.047372, …
## $ total_spent        <int> 1617, 27, 776, 53, 422, 716, 590, 169, 46, 49, 61,…
```

4. Trực quan hóa dữ liệu (Data Visualization / EDA)

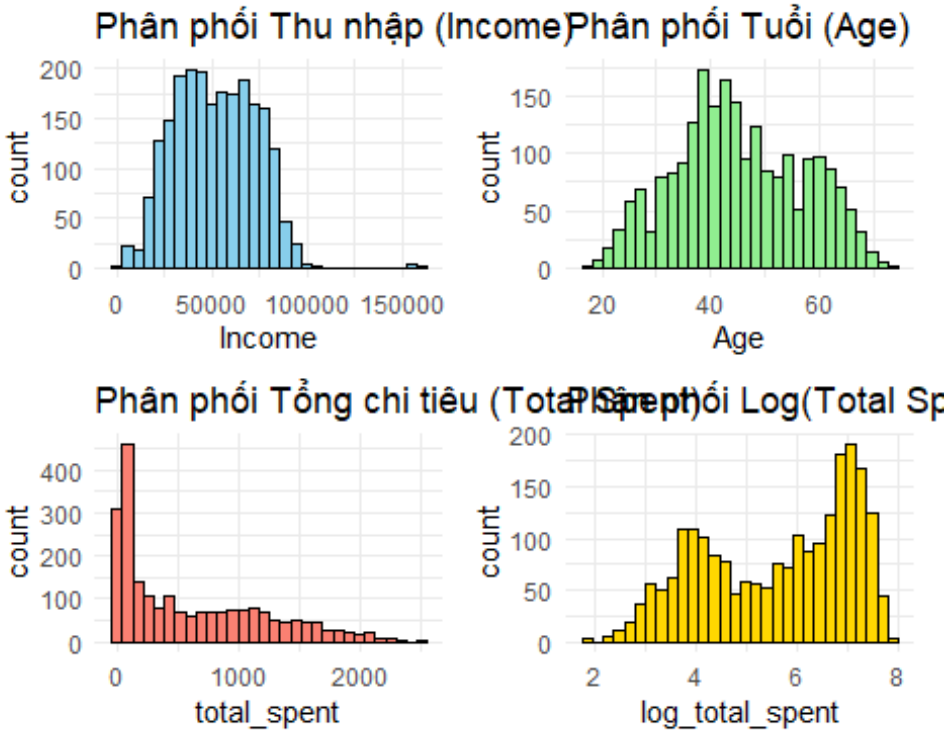
Sau khi tiền xử lý, Phân tích Dữ liệu Khám phá (EDA) được thực hiện thông qua trực quan hóa để hiểu rõ hơn về phân phối của các biến và mối quan hệ giữa chúng, từ đó rút ra những hiểu biết ban đầu làm tiền đề cho việc xây dựng mô hình.

4.1 Phân tích đơn biến

Phân tích này tập trung vào việc xem xét phân phối của từng biến riêng lẻ.

4.1.1 Biến số lượng (Numerical Variables)

Chúng ta sẽ kiểm tra phân phối của các biến số lượng chính như Thu nhập, Tuổi, Tổng chi tiêu (cả gốc và logarit).



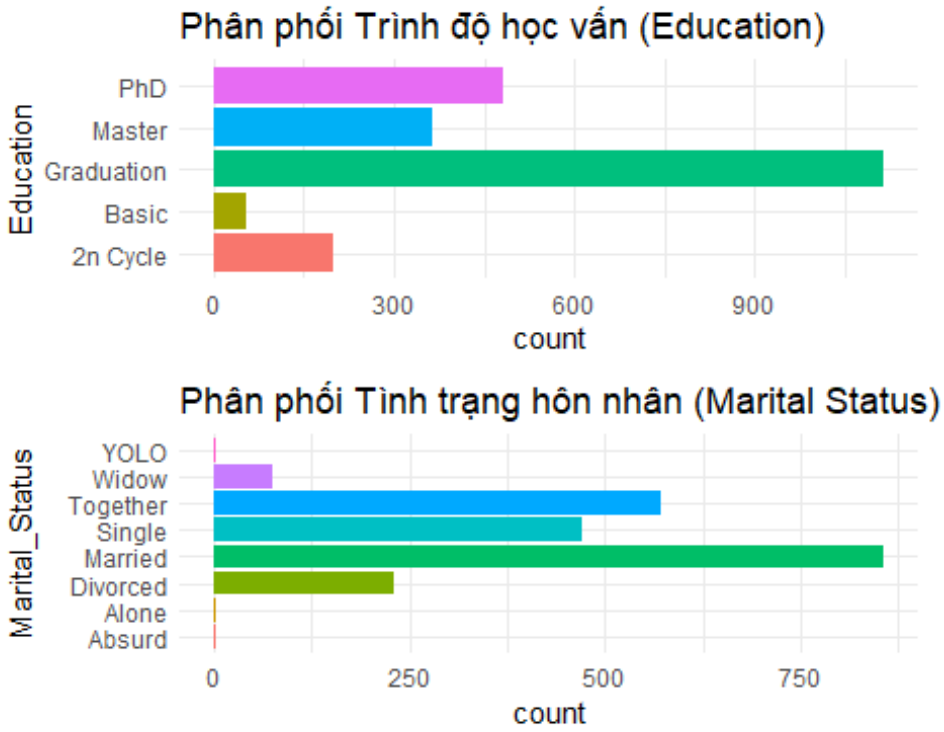
Phân phối của Thu nhập, Tuổi, Tổng chi tiêu(gốc và Logarit)

Nhận xét:

- Income (Thu nhập) và total_spent (Tổng chi tiêu) có phân phối lệch phải, cho thấy phần lớn khách hàng có thu nhập và chi tiêu ở mức thấp hơn. Việc sử dụng log_total_spent giúp phân phối cân đối hơn, phù hợp cho mô hình hồi quy.
- Age có phân phối tương đối rộng, tập trung chủ yếu ở độ tuổi trung niên.

4.1.2 Biến phân loại (Categorical Variables)

Xem xét tỷ lệ các nhóm trong biến Học vấn và Tình trạng hôn nhân.



Phân phối Trình độ học vấn và Tình trạng hôn nhân

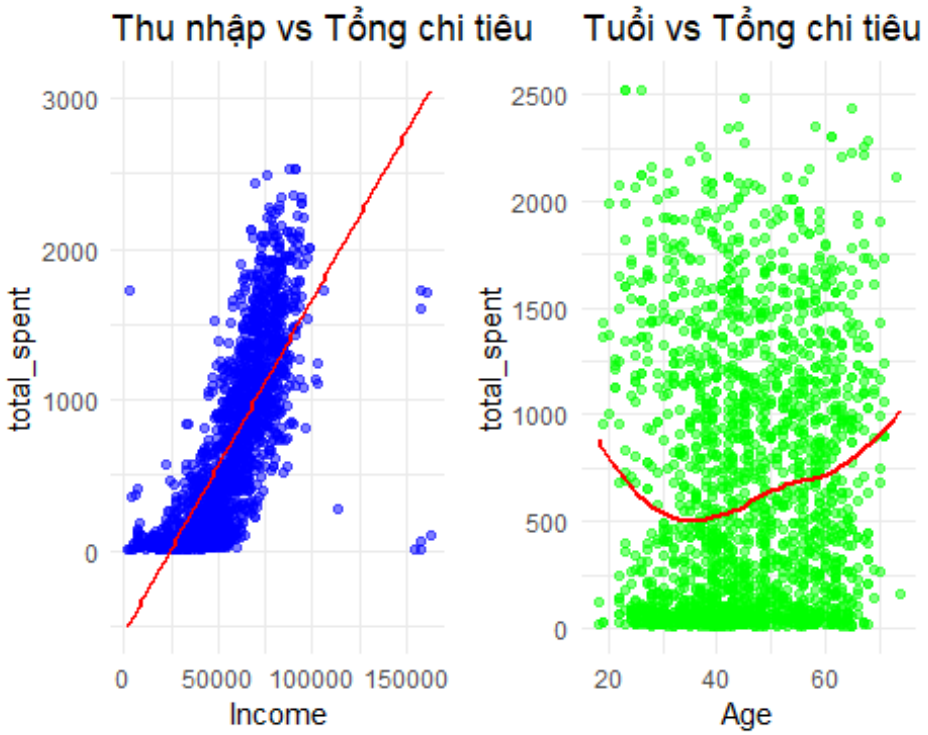
Nhận xét:

- Trình độ học vấn Graduation chiếm đa số, tiếp theo là PhD và Master. Các nhóm 2n Cycle và Basic có số lượng ít hơn đáng kể.
- Về tình trạng hôn nhân, nhóm Married và Together (sống chung) chiếm tỷ lệ lớn nhất, tiếp theo là Single và Divorced/Widow. Các nhóm Alone, Absurd, YOLO có số lượng rất nhỏ.

4.2 Phân tích đa biến

Khám phá mối quan hệ giữa các biến.

4.2.1 Mối quan hệ giữa Thu nhập, Tuổi và Tổng chi tiêu (các biến số lượng)

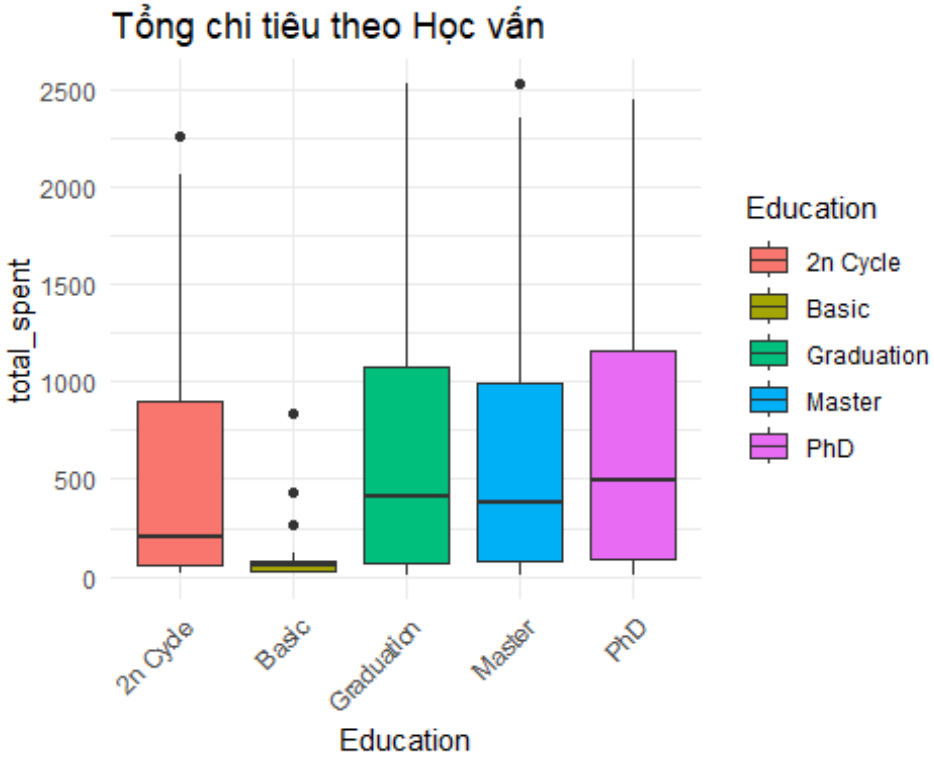


Mối quan hệ giữa Thu nhập/Tuổi và Tuổi/Tổng chi tiêu

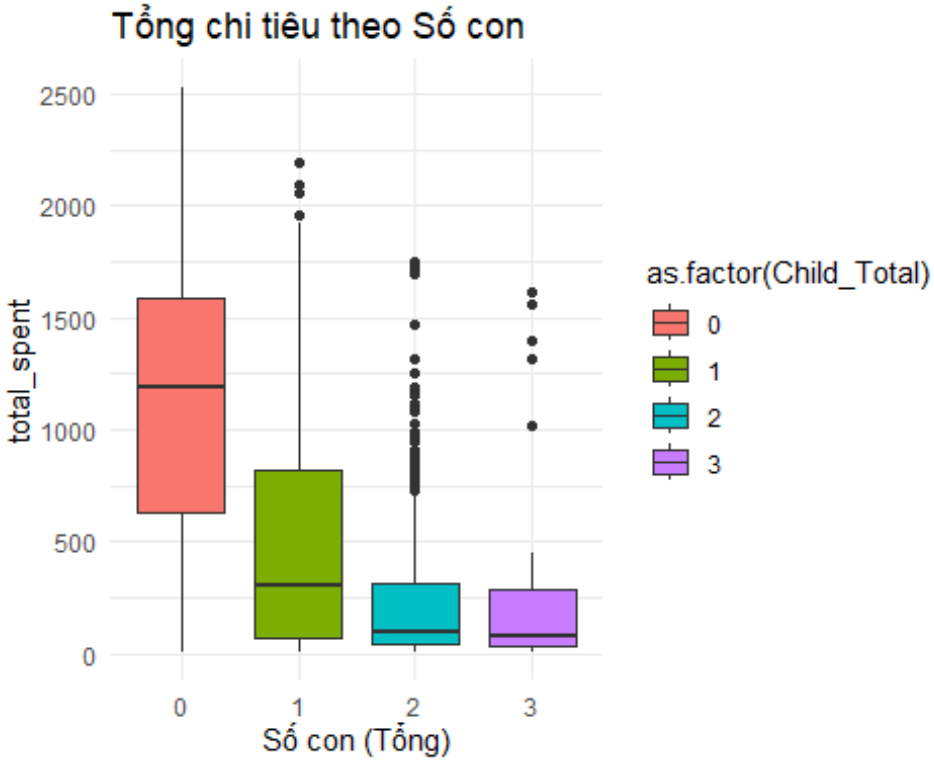
Nhận xét:

- Có mối tương quan dương khá rõ ràng giữa Income và total_spent. Khách hàng có thu nhập cao hơn có xu hướng chi tiêu nhiều hơn.
- Mối quan hệ giữa Age và total_spent ít rõ ràng hơn, chi tiêu có xu hướng tăng nhẹ ở tuổi trung niên.

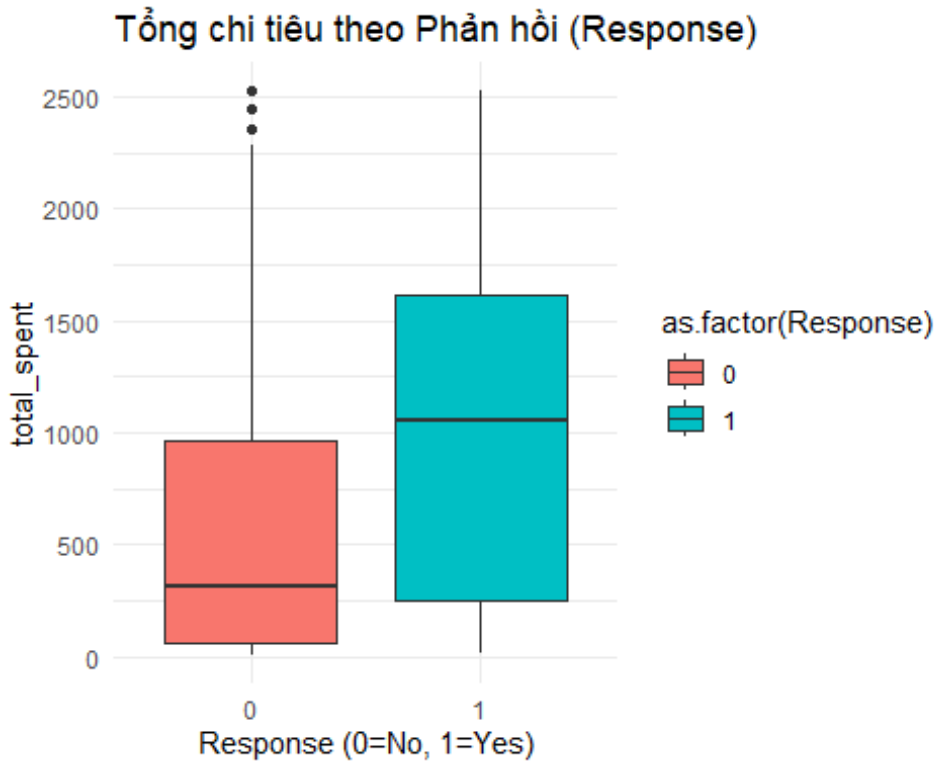
4.2.2 Ảnh hưởng của Học vấn, Số con và Phản hồi chiến dịch đến Tổng chi tiêu



Tổng chi tiêu theo Học vấn



Tổng chi tiêu theo số con



Tổng chi tiêu theo Phản hồi

Nhận xét:

Education: Nhóm có trình độ PhD và Master có xu hướng thu nhập và chi tiêu trung bình cao hơn so với nhóm Graduation và đặc biệt là Basic.

Child_Total: Có xu hướng rõ ràng: số lượng con càng nhiều, tổng chi tiêu trung bình càng giảm.

Response: Khách hàng chấp nhận ưu đãi (Response=1) có xu hướng chi tiêu trung bình cao hơn đáng kể.

4.3 Kết luận sơ bộ từ EDA

Qua phân tích dữ liệu khám phá, một số điểm chính có thể rút ra:

- Các biến Income và total_spent có phân phối lệch phải, việc sử dụng phép biến đổi log cho total_spent trong mô hình hồi quy là hợp lý.
- Có mối quan hệ tương quan dương giữa Income và total_spent.
- Education và Child_Total là các yếu tố nhân khẩu học có vẻ ảnh hưởng rõ rệt đến mức chi tiêu.
- Khách hàng có Response = 1 thường có mức chi tiêu cao hơn, cho thấy tiềm năng của việc phân tích biến Response.

Những hiểu biết này là cơ sở quan trọng cho việc lựa chọn biến và diễn giải kết quả các mô hình ở phần tiếp theo.

5. Mô hình hóa dữ liệu (Data Modeling)

Sau khi khám phá dữ liệu, phần này tập trung vào việc xây dựng các mô hình học máy để giải quyết các câu hỏi nghiên cứu. Ba phương pháp mô hình hóa chính được lựa chọn, bao gồm học không giám sát (phân cụm) và học có giám sát (phân loại và hồi quy), nhằm cung cấp cái nhìn đa chiều về hành vi và phân khúc khách hàng.

5.1 Giới thiệu các mô hình

Dựa trên mục tiêu dự án và đặc điểm dữ liệu, ba mô hình sau được lựa chọn:

1. **Phân cụm K-Means (K-Means Clustering):**
    - **Loại:** Học không giám sát (Unsupervised Learning).
    - **Mục đích:** Phân chia khách hàng thành các nhóm (phân khúc) riêng biệt dựa trên sự tương đồng về các đặc điểm nhân khẩu học và hành vi mua sắm. Điều này giúp doanh nghiệp hiểu rõ hơn cấu trúc khách hàng của mình.
2. **Hồi quy Logistic (Logistic Regression):**
    - **Loại:** Học có giám sát - Phân loại (Supervised Learning - Classification).
    - **Mục đích:** Dự đoán khả năng khách hàng chấp nhận ưu đãi trong chiến dịch marketing cuối cùng (biến Response). Mô hình này giúp xác định các yếu tố thúc đẩy khách hàng phản hồi tích cực.
3. **Hồi quy Tuyến tính Đa biến (Multiple Linear Regression)**
    - **Loại:** Học có giám sát - Hồi quy (Supervised Learning - Regression).
    - **Mục đích:** Dự đoán tổng chi tiêu của khách hàng (sử dụng biến log_total_spent) và xác định các yếu tố ảnh hưởng đến mức chi tiêu. Mô hình này giúp hiểu các động lực kinh tế cơ bản.

Việc áp dụng đồng thời ba mô hình này cho phép không chỉ phân khúc khách hàng mà còn dự đoán hành vi cụ thể và giá trị kinh tế của họ, đồng thời xác định các yếu tố thúc đẩy đằng sau đó.

5.2 Mô hình 1: Phân cụm K-Means (K-Means Clustering)

Mô hình K-Means được sử dụng để tự động phân nhóm các khách hàng có đặc điểm tương đồng vào cùng một cụm, giúp khám phá các phân khúc khách hàng tiềm ẩn trong dữ liệu.

5.2.1 Mục tiêu và Phương pháp

- **Mục tiêu:** Phân khúc khách hàng dựa trên các đặc điểm chính như tuổi (Age), thu nhập (Income), tổng chi tiêu (total_spent), số lần mua hàng gần đây (Recency), số con (Child_Total), và thời gian gắn bó (Days_Customer), cùng các biến về hành vi mua hàng khác.
- **Phương pháp K-Means:**
    - Thuật toán này tìm cách chia N khách hàng thành k cụm sao cho tổng phương sai trong mỗi cụm là nhỏ nhất. Nó hoạt động bằng cách gán mỗi khách hàng vào cụm có tâm (điểm trung bình của cụm) gần nhất, sau đó cập nhật lại vị trí các tâm cụm. Quá trình này lặp lại cho đến khi các cụm ổn định.
- **Chuẩn bị dữ liệu:** Các biến số lượng dùng để phân cụm đã được chuẩn hóa (scaling) để đảm bảo các biến có thang đo khác nhau không ảnh hưởng đến kết quả một cách không công bằng.
- **Xác định số cụm tối ưu (k):** Số lượng cụm k được xác định bằng cách sử dụng kết hợp phương pháp Elbow (quan sát điểm "gãy" trên đồ thị tổng bình phương sai số trong cụm - WCSS) và phương pháp Silhouette (tìm giá trị Silhouette trung bình cao nhất).
    - **Phương pháp Elbow (Elbow Method):** Vẽ biểu đồ tổng bình phương sai số trong cụm (Within-Cluster Sum of Squares - WCSS) theo số lượng cụm k. Chọn giá trị k tại "khuỷu tay" của đồ thị, nơi mà việc tăng thêm k không còn làm giảm WCSS một cách đáng kể.
    - **Phân tích Silhouette (Silhouette Analysis):** Tính toán chỉ số Silhouette trung bình cho các giá trị k khác nhau. Chỉ số Silhouette đo lường mức độ tương đồng của một điểm dữ liệu với cụm của chính nó so với các cụm khác. Giá trị Silhouette trung bình cao hơn cho thấy cấu trúc cụm tốt hơn. Chọn k tương ứng với giá trị Silhouette trung bình cao nhất.

5.2.2 Triển khai

Quá trình triển khai mô hình K-Means trong R bao gồm các bước chính:

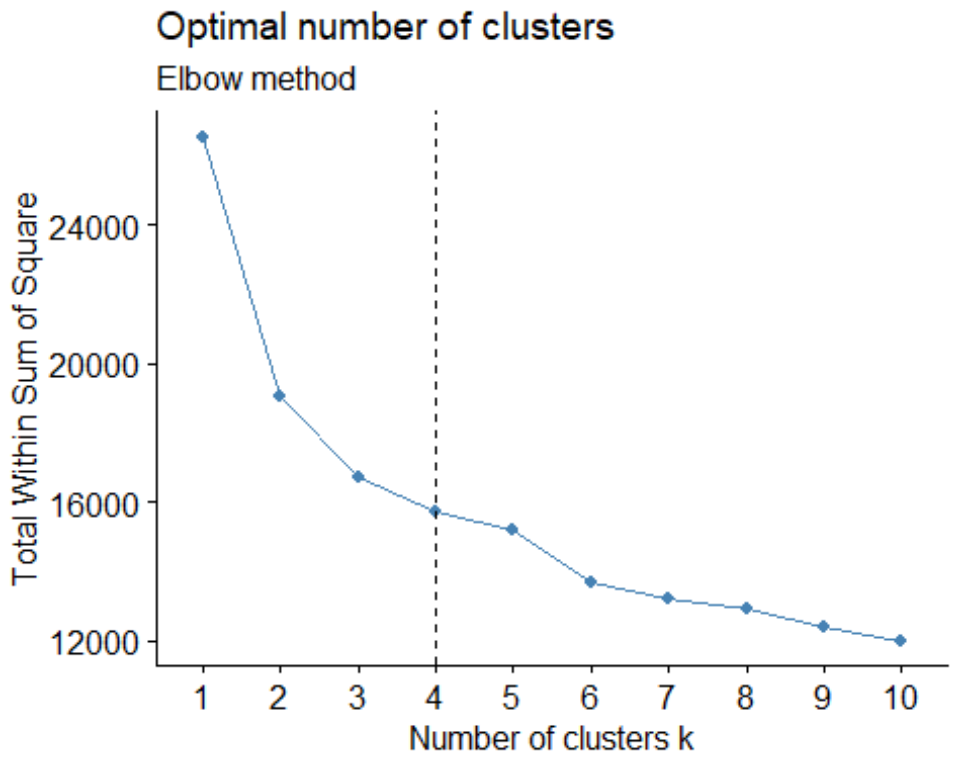
- **Lựa chọn và chuẩn hóa biến:** Các biến số lượng phù hợp từ bộ dữ liệu customers_final (như Income, Age, total_spent, v.v.) được chọn và chuẩn hóa bằng hàm scale().
- **Xác định số cụm k tối ưu:** Dựa trên phân tích Elbow và Silhouette (trình bày chi tiết ở Mục 6.1.1), số cụm tối ưu được chọn là **optimal_k**.
- **Chạy thuật toán K-Means:** Sử dụng hàm kmeans() với số cụm k đã chọn và dữ liệu đã chuẩn hóa. Tham số nstart được thiết lập ở giá trị 50 để tăng độ ổn định của kết quả.

- **Gán nhãn cụm:** Kết quả phân cụm (nhãn của từng khách hàng thuộc về cụm nào) được thêm vào lại dataframe customers_final dưới dạng một cột mới (Cluster_KMeans) để phục vụ cho việc phân tích đặc điểm cụm ở phần sau.

1. **Lựa chọn và chuẩn hóa biến:**
2. **Chuẩn hóa dữ liệu**
3. **Xác định số cụm tối ưu (k)**

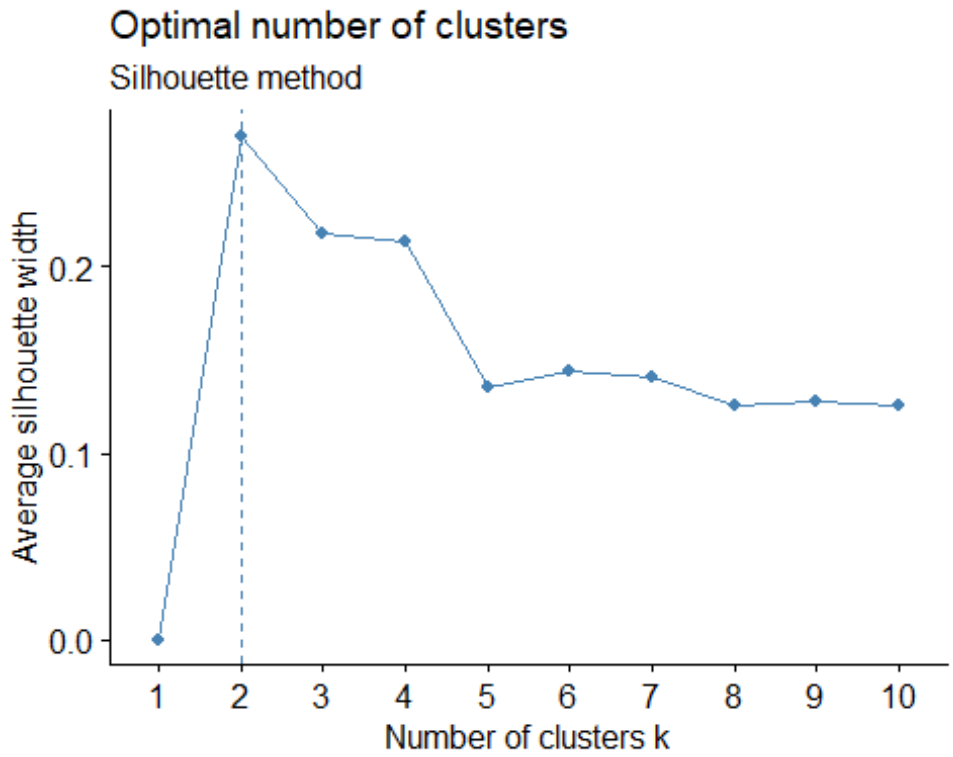
Sử dụng phương pháp Elbow

```
set.seed(123)
fviz_nbclust(customers_scaled, kmeans, method = "wss") +
  #wss = within sum square
  geom_vline(xintercept = 4, linetype = 2) + # Ví dụ chọn k=4
  labs(subtitle = "Elbow method")
```



Sử dụng phương pháp Silhouette

```
set.seed(123)
fviz_nbclust(customers_scaled, kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette method")
```




```
# => Giả sử sau khi xem xét, chúng ta chọn k = 4
optimal_k <- 4
```

4. **Chạy thuật toán K-Means**
5. **Thêm thông tin cụm vào dataframe gốc**

5.3 Mô hình 2: Hồi quy Logistic (Logistic Regression)

Mô hình Hồi quy Logistic được áp dụng để dự đoán khả năng khách hàng chấp nhận ưu đãi trong chiến dịch marketing cuối cùng, dựa trên các đặc điểm và hành vi của họ.

5.3.1 Mục tiêu và Phương pháp

- **Mục tiêu:** Xây dựng mô hình dự đoán biến nhị phân Response (0 = không chấp nhận, 1 = chấp nhận) dựa trên các yếu tố như nhân khẩu học, thu nhập, lịch sử mua sắm và tương tác với các chiến dịch trước.
- **Phương pháp Hồi quy Logistic:**
    - Mô hình không dự đoán trực tiếp giá trị 0 hay 1 mà dự đoán **xác suất** để biến mục tiêu nhận giá trị 1 (trong trường hợp này là xác suất khách hàng chấp nhận ưu đãi, P(Response=1)).
    - Mô hình sử dụng hàm liên kết logit (logit link function) để biến đổi xác suất (nằm trong khoảng [0, 1]) thành một giá trị tuyến tính có thể chạy từ $-\infty$ đến $+\infty$, dựa

trên tổ hợp tuyến tính của các biến độc lập: $logit\big(P(Response = 1)\big) = ln\left(\frac{P(Response=1)}{1-P(Response=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

  - Các hệ số của mô hình cho biết mức độ ảnh hưởng của từng biến độc lập đến khả năng (log-odds) xảy ra sự kiện.

- **Chuẩn bị dữ liệu:**
  - **Chia dữ liệu:** Bộ dữ liệu customers_final được chia thành tập huấn luyện (để xây dựng mô hình) và tập kiểm tra (để đánh giá hiệu suất mô hình). Việc chia này đảm bảo tỷ lệ của biến Response được duy trì trong cả hai tập.
  - **Xử lý biến phân loại:** Các biến phân loại (factor) được tự động xử lý bằng cách tạo biến giả trong quá trình xây dựng mô hình.

5.3.2 Triển khai

Quá trình xây dựng mô hình Hồi quy Logistic bao gồm các bước sau:

- **Chuẩn bị và Chia dữ liệu:** Lựa chọn các biến độc lập phù hợp từ customers_final (loại bỏ các biến liên quan đến tổng chi tiêu như log_total_spent, total_spent để tránh rò rỉ thông tin không liên quan trực tiếp đến Response). Biến Response được chuyển thành kiểu factor. Sử dụng hàm createDataPartition từ gói caret để chia dữ liệu thành tập huấn luyện (75%) và tập kiểm tra (25%).
- **Xây dựng mô hình:** Sử dụng hàm glm() trong R với family = binomial(link = "logit") trên tập huấn luyện.
- **Dự đoán và Đánh giá mô hình:** Mô hình được dùng để dự đoán xác suất trên tập kiểm tra. Dựa trên một ngưỡng (thường là 0.5), xác suất này được chuyển thành dự đoán lớp (0 hoặc 1). Các chỉ số như Ma trận nhầm lẫn, Độ chính xác, Độ nhạy, Độ đặc hiệu, F1-score, và AUC (từ đường cong ROC) được sử dụng để đánh giá hiệu suất.
- **Diễn giải kết quả:** Phân tích ý nghĩa thống kê và giá trị của các hệ số hồi quy (Odds Ratios) để hiểu rõ các yếu tố ảnh hưởng.

1. Chuẩn bị dữ liệu cho Logistic Regression

2. Chia dữ liệu thành tập huấn luyện và kiểm tra (75%/25%)

3. Xây dựng mô hình Logistic Regression trên tập huấn luyện

4. Dự đoán trên tập kiểm tra

5.4 Mô hình 3: Hồi quy Tuyến tính Đa biến (Multiple Linear Regression)

Mô hình Hồi quy Tuyến tính Đa biến được xây dựng để tìm hiểu mối quan hệ giữa tổng chi tiêu của khách hàng (biến log_total_spent) và các yếu tố dự đoán khác, đồng thời dự đoán mức chi tiêu này.

5.4.1 Mục tiêu và Phương pháp

- **Mục tiêu:** Xây dựng một mô hình tuyến tính để dự đoán giá trị log_total_spent dựa trên các đặc điểm nhân khẩu học và hành vi mua sắm. Đồng thời, xác định những yếu tố có ảnh hưởng ý nghĩa thống kê đến tổng chi tiêu.
- **Phương pháp Hồi quy Tuyến tính:**
    - Mô hình giả định rằng giá trị trung bình của biến phụ thuộc (log_total_spent) là một tổ hợp tuyến tính của các biến độc lập: $E[log\_total\_spent] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$
    - Mô hình ước lượng các hệ số β bằng phương pháp Bình phương tối thiểu thông thường (Ordinary Least Squares - OLS).
    - Các hệ số β_i cho biết sự thay đổi trung bình trong log_total_spent khi biến X_i tăng một đơn vị, giữ các biến khác không đổi.
- **Chuẩn bị dữ liệu:**
    - Bộ dữ liệu customers_final được chia thành tập huấn luyện và tập kiểm tra riêng cho mô hình này.
    - Biến mục tiêu là log_total_spent.
    - Các biến phân loại được tự động xử lý thông qua dummy coding khi sử dụng hàm lm().
- **Kiểm tra giả định:** Sau khi xây dựng, các giả định quan trọng của hồi quy tuyến tính (như tính tuyến tính, phương sai không đổi, phân phối chuẩn của phần dư) sẽ được kiểm tra.

5.4.2 Triển khai (Implementation)

Các bước chính để triển khai mô hình Hồi quy Tuyến tính bao gồm:

- **Chuẩn bị và chia dữ liệu:**
- Lựa chọn các biến độc lập phù hợp từ customers_final (loại bỏ biến Response và total_spent gốc).

- Chia dữ liệu đã chọn thành tập huấn luyện (75%) và tập kiểm tra (25%) bằng hàm createDataPartition.
- **Xây dựng mô hình:** Sử dụng hàm lm() trong R trên tập huấn luyện để xây dựng mô hình.
- **Dự đoán và Đánh giá:** Mô hình được dùng để dự đoán log_total_spent trên tập kiểm tra. Các chỉ số như R-squared (R^2), Adjusted R-squared, và Root Mean Squared Error (RMSE) được tính toán để đánh giá hiệu suất.
- **Kiểm tra giả định và Diễn giải:** Sử dụng các biểu đồ chẩn đoán để kiểm tra giả định và phân tích ý nghĩa thống kê của các hệ số hồi quy.

1. Chuẩn bị dữ liệu cho Hồi quy Tuyến tính

2. Chia dữ liệu thành tập huấn luyện và kiểm tra MỚI cho mô hình tuyến tính

3. Xây dựng mô hình Linear Regression trên tập huấn luyện MỚI (train_set_lm)

4. Dự đoán trên tập kiểm tra MỚI (test_set_lm)

```
predictions_lm <- predict(linear_model, newdata = test_set_lm)
```

5. Đánh giá mô hình trên tập kiểm tra MỚI

```
# Ví dụ tính RMSE:
actual_values_lm <- test_set_lm$log_total_spent
rmse_lm <- sqrt(mean((actual_values_lm - predictions_lm)^2))
print(paste("RMSE for Linear Regression:", rmse_lm))
## [1] "RMSE for Linear Regression: 0.506594966510391"
```

6. Thực nghiệm, kết quả, và thảo luận (Experiments, Results, and Discussion)

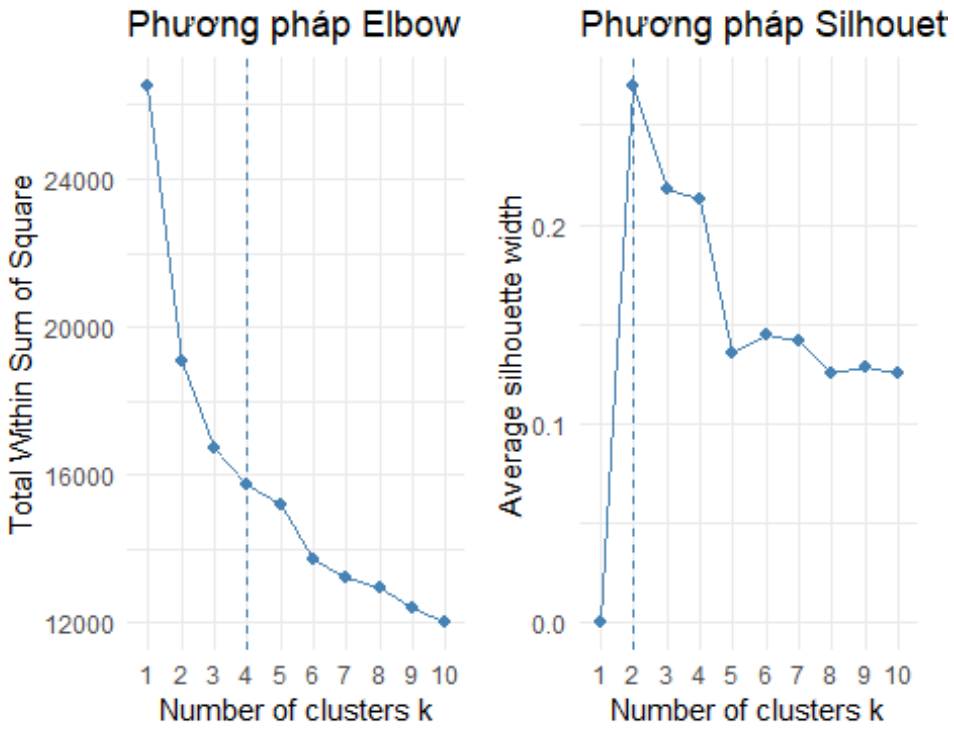
Sau khi xác định các mô hình, phần này trình bày chi tiết kết quả thực nghiệm từ mỗi mô hình, bao gồm việc đánh giá hiệu suất và diễn giải các phát hiện quan trọng về hành vi và phân khúc khách hàng.

6.1 Kết quả Mô hình 1: K-Means

Mô hình K-Means được triển khai để phân nhóm khách hàng thành các phân khúc dựa trên các đặc điểm tương đồng.

6.1.1 Xác định số cụm tối ưu

Số lượng cụm (k) tối ưu được xác định dựa trên phương pháp Elbow và phân tích Silhouette.



Xác định số cụm tối ưu bằng phương pháp Elbow và Silhouette

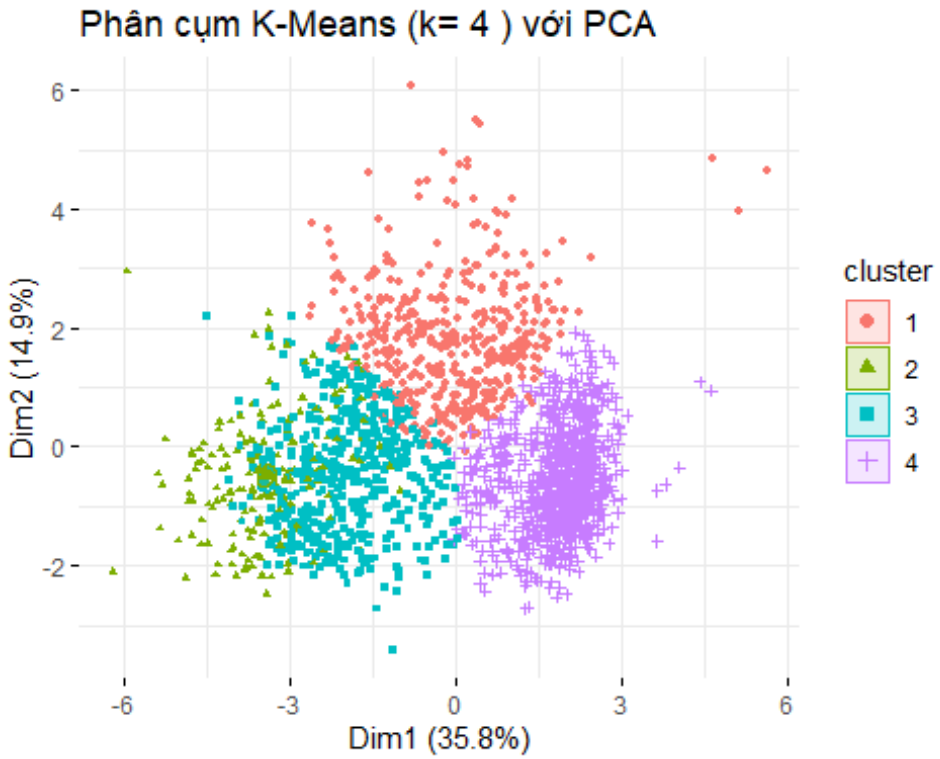
Nhận xét:

- **Phương pháp Elbow:** Quan sát biểu đồ WCSS, ta thấy đường cong bắt đầu "gãy" hoặc giảm độ dốc không còn nhiều tại điểm k = 4. Điều này gợi ý rằng việc tăng thêm số cụm sau điểm này không mang lại hiệu quả giảm WCSS đáng kể.
- **Phương pháp Silhouette**: Biểu đồ Silhouette trung bình cho thấy giá trị cao nhất đạt được tại k = 2.

Dựa trên phương pháp Elbow, điểm "gãy" của đồ thị WCSS xuất hiện tại k=4, cho thấy đây có thể là một lựa chọn tốt cho số lượng cụm. Trong khi đó, phương pháp Silhouette cho thấy giá trị Silhouette trung bình cao nhất tại k=2. Tuy nhiên, để có được các phân khúc chi tiết và mang tính hành động hơn cho mục tiêu marketing, nhóm quyết định lựa chọn k=4 để tiếp tục phân tích.

6.1.2 Trực quan hóa cụm

Để trực quan hóa các cụm đã được hình thành, kỹ thuật Phân tích Thành phần Chính (PCA) được sử dụng để giảm số chiều của dữ liệu xuống còn 2 chiều (hai thành phần chính đầu tiên), sau đó vẽ biểu đồ các điểm dữ liệu với màu sắc tương ứng với cụm của chúng



Phân cụm K-Means (k=4) với PCA

Nhận xét: Biểu đồ PCA cho thấy 4 cụm khách hàng có sự phân tách tương đối, mặc dù có một số vùng chồng lấn. Hai thành phần chính đầu tiên giải thích được khoảng 50.7% phương sai của dữ liệu.

6.1.3 Phân tích và Diễn giải đặc điểm cụm

Bảng dưới đây tóm tắt giá trị trung bình của một số biến số lượng chính theo từng cụm:

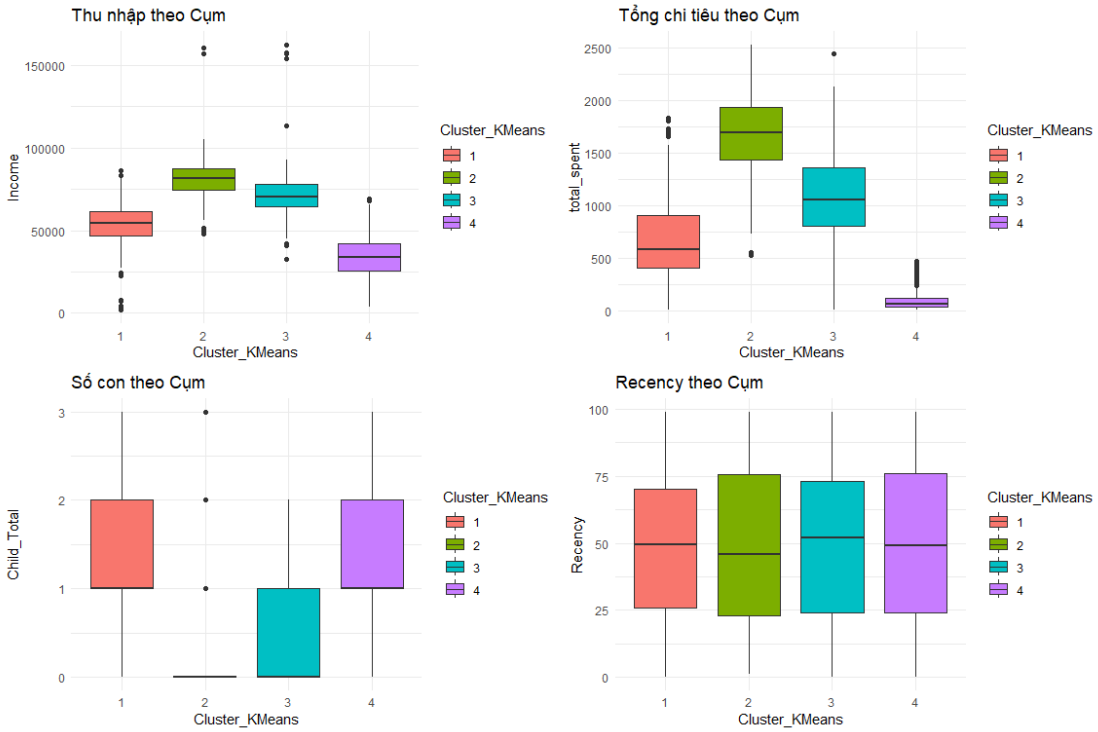
Đặc điểm trung bình các cụm K-Means

Cluster_KMeans	Avg_Income	Avg_Age	Avg_Total_Spent	Avg_Recency	Avg_Child_Total	Avg_AcceptedCmp_Total	Count
1	53672.23	48.31013	677.36076	48.74051	1.3122363	0.2320675	474
2	81449.29	42.85128	1662.91282	48.62564	0.2153846	1.9641026	195
3	71215.	48.04	1076.98940	49.3091	0.4222615	0.1448763	56

Cluster_KMeans	Avg_Income	Avg_Age	Avg_Total_Spent	Avg_Recency	Avg_Child_Total	Avg_AcceptedCmp_Total	Count
	02	594		9			6
4	34085.91	42.25384	90.44422	49.06551	1.2210850	0.0870010	977

Diễn giải và Đặt tên cụm:

- **Cụm 2: "Khách hàng Vàng"** *(Count: 195)* Đặc điểm: Thu nhập và tổng chi tiêu cao nhất, ít con nhất, chấp nhận nhiều chiến dịch nhất. Độ tuổi trung bình trẻ hơn. Hành vi: Ít bị thu hút bởi giảm giá, ít truy cập web. Nhận xét: Nhóm giá trị cao, phản hồi tốt với marketing.
- **Cụm 3: "Khách hàng Bạc"** *(Count: 566)* Đặc điểm: Thu nhập và tổng chi tiêu cao thứ hai, ít con. Hành vi: Ít chấp nhận chiến dịch cũ, ít truy cập web nhất. Thời gian gắn bó ngắn nhất. Nhận xét: Có tiềm năng chi tiêu tốt, cần chiến lược thu hút và tăng tương tác
- **Cụm 1: "Khách hàng Trung thành"** *(Count: 474)* Đặc điểm: Thu nhập và chi tiêu ở mức trung bình, nhiều con nhất, thời gian gắn bó dài nhất. Hành vi: Mua hàng giảm giá nhiều nhất, truy cập web nhiều. Nhận xét: Nhóm gia đình, trung thành, nhạy cảm với ưu đãi.
- **Cụm 4: "Khách hàng Phổ thông** *(Count: 977)* Đặc điểm: Thu nhập và tổng chi tiêu thấp nhất, nhiều con, chấp nhận ít chiến dịch nhất. Hành vi: Truy cập web khá nhiều. Nhận xét: Nhóm đông đảo nhất, chi tiêu thấp, cần các sản phẩm/ưu đãi phù hợp ngân sách.



So sánh một số biến chính giữa các cụm K-Means

Kết luận sơ bộ từ K-Means: Phân tích K-Means đã phân chia khách hàng thành 4 nhóm với đặc điểm nhân khẩu học và hành vi chi tiêu khác biệt, cung cấp cơ sở cho các chiến lược marketing mục tiêu.

6.2 Kết quả Mô hình 2: Hồi quy Logistic

Mô hình Hồi quy Logistic được xây dựng để dự đoán khả năng khách hàng chấp nhận ưu đãi trong chiến dịch marketing cuối cùng (biến Response).

6.2.1 Đánh giá mô hình

Hiệu suất của mô hình được đánh giá trên tập kiểm tra bằng cách sử dụng ngưỡng xác suất 0.5.

```
## [1] "Accuracy: 0.8895"
## [1] "Sensitivity (Recall for X1): 0.506"
## [1] "Specificity (Recall for X0): 0.9574"
## [1] "AUC: 0.8983"
```

Nhận xét kết quả đánh giá:

- **Ma trận nhầm lẫn (Confusion Matrix):**
    - **Độ chính xác (Accuracy):** 0.8895 (88.95%). Mô hình dự đoán đúng tổng thể 88.95% các trường hợp trong tập kiểm tra.

- **Độ nhạy (Sensitivity/Recall - tỷ lệ dự đoán đúng các trường hợp Response="X1"):** 0.50602. Mô hình xác định đúng được khoảng 50.6% số khách hàng thực sự chấp nhận ưu đãi.
- **Độ đặc hiệu (Specificity - tỷ lệ dự đoán đúng các trường hợp Response="X0"):** 0.95736. Mô hình rất tốt trong việc xác định những khách hàng sẽ không chấp nhận ưu đãi (dự đoán đúng 95.7%).

- **Đường cong ROC và AUC:**
  - **Giá trị AUC (Area Under the Curve):** 0.8983. Cho thấy mô hình có khả năng phân biệt rất tốt giữa hai lớp khách hàng (chấp nhận và không chấp nhận ưu đãi).

Thảo luận về hiệu suất mô hình: Nhìn chung, mô hình có độ chính xác tổng thể và AUC tốt. Tuy nhiên, độ nhạy (khả năng phát hiện khách hàng chấp nhận ưu đãi) còn ở mức trung bình, có thể do sự mất cân bằng trong dữ liệu (chỉ khoảng 15% khách hàng thực sự Response=1).

6.2.2 Diễn giải hệ số

Phân tích các hệ số có ý nghĩa thống kê (p-value < 0.05) từ mô hình để hiểu yếu tố nào ảnh hưởng đến khả năng Response.

```
summary_coeffs_log <- summary(logistic_model)$coefficients
significant_coeffs_log <- summary_coeffs_log[summary_coeffs_log[, "Pr(>|z|)"] < 0.05, ]


odds_ratios_sig <- exp(significant_coeffs_log[, "Estimate"])
kable(data.frame(Estimate = significant_coeffs_log[, "Estimate"],
            Std.Error = significant_coeffs_log[, "Std. Error"],
            z.value = significant_coeffs_log[, "z value"],
            P_value = significant_coeffs_log[, "Pr(>|z|)"],
            Odds_Ratio = odds_ratios_sig),
    caption = "Các hệ số có ý nghĩa thống kê trong Mô hình Logistic", digits = 3)
```

Các hệ số có ý nghĩa thống kê trong Mô hình Logistic

	Estimate	Std.Error	z.value	P_value	Odds_Ratio
EducationMaster	0.905	0.389	2.329	0.020	2.472
EducationPhD	1.320	0.371	3.559	0.000	3.744
Recency	-0.032	0.003	-9.290	0.000	0.969
Child_Total	-0.598	0.181	-3.298	0.001	0.550
AcceptedCmp_Total	1.600	0.175	9.129	0.000	4.953
Days_Customer	0.004	0.001	7.724	0.000	1.004

	Estimate	Std.Error	z.value	P_value	Odds_Ratio
NumDealsPurchases	0.206	0.068	3.045	0.002	1.229
NumCatalogPurchases	0.127	0.047	2.710	0.007	1.135
NumStorePurchases	-0.249	0.042	-5.958	0.000	0.779
NumWebVisitsMonth	0.127	0.059	2.135	0.033	1.135

Biến	**Ước lượng (Estimate)**	**Odds Ratio**	**Diễn giải Ảnh hưởng đến Response (X1)**
Recency	-0.0318	0.969	Mỗi ngày Recency giảm đi, odds chấp nhận ưu đãi tăng nhẹ
EducationMaster	0.905	2.472	Có bằng Thạc sĩ làm tăng odds chấp nhận ưu đãi ~2.47 lần (so với "2n Cycle").
EducationPhD	1.320	3.744	Có bằng Tiến sĩ làm tăng odds chấp nhận ưu đãi ~3.74 lần (so với "2n Cycle").
Child_Total	-0.598	0.550	Mỗi đứa con tăng thêm làm giảm odds chấp nhận ưu đãi khoảng 45%.
AcceptedCmp_Total	1.600	4.953	Mỗi chiến dịch trước được chấp nhận làm tăng odds chấp nhận ưu đãi hiện tại gần 5 lần.
NumDealsPurchases	0.206	1.229	Số lần mua hàng giảm giá tăng làm tăng nhẹ odds chấp nhận ưu đãi.
NumCatalogPurchases	0.127	1.135	Số lần mua qua catalog tăng làm tăng

			nhẹ odds chấp nhận ưu đãi.
NumStorePurchases	-0.250	0.779	Số lần mua tại cửa hàng tăng làm giảm odds chấp nhận ưu đãi khoảng 22%.
NumWebVisitsMonth	0.127	1.135	Số lượt truy cập web/tháng tăng làm tăng nhẹ odds chấp nhận ưu đãi.
Cluster_KMeans3	0.713	2.040	Thuộc Cụm 3 (so với Cụm 1) làm tăng odds chấp nhận ưu đãi hơn gấp đôi.
(Lưu ý: Recency, Age có p-value ~0.06, ý nghĩa biên)			

Phân tích các yếu tố ảnh hưởng đến Response:

- **Tích cực:** Việc khách hàng đã từng chấp nhận các chiến dịch trước (AcceptedCmp_Total) là yếu tố thúc đẩy mạnh mẽ nhất. Trình độ học vấn cao (Thạc sĩ, Tiến sĩ) và thuộc phân khúc K-Means 3 cũng làm tăng đáng kể khả năng chấp nhận ưu đãi. Số lần mua hàng giảm giá, mua qua catalog và số lượt truy cập web/tháng cũng có ảnh hưởng tích cực nhỏ.
- **Tiêu cực:** Có nhiều con (Child_Total) và mua hàng nhiều tại cửa hàng (NumStorePurchases) làm giảm khả năng chấp nhận ưu đãi.
- **Không có ý nghĩa thống kê rõ rệt (p > 0.05):** Income, Complain, Days_Customer, NumWebPurchases, EducationBasic, EducationGraduation, và hầu hết các mức của Marital_Status (khi so với mức cơ sở "Absurd" có ít mẫu). Biến Cluster_KMeans2 và Cluster_KMeans4 cũng không cho thấy ảnh hưởng rõ ràng đến Response trong mô hình này.

Kết luận từ Hồi quy Logistic: Mô hình cho thấy hiệu quả dự đoán tổng thể tốt. Lịch sử tương tác tích cực với các chiến dịch trước, trình độ học vấn cao và thuộc một số phân khúc K-Means nhất định là những yếu tố quan trọng làm tăng khả năng khách hàng phản hồi. Ngược lại, số lượng con cái và việc mua sắm nhiều tại cửa hàng có thể làm giảm khả năng này.

6.3 Kết quả Mô hình 3: Hồi quy Tuyến tính Đa biến

Mô hình Hồi quy Tuyến tính Đa biến được xây dựng để dự đoán tổng chi tiêu của khách hàng (sử dụng log_total_spent) và xác định các yếu tố ảnh hưởng.

6.3.1 Đánh giá mô hình

Hiệu suất của mô hình được đánh giá trên tập kiểm tra (test_set_lm).

```
## [1] "R-squared trên tập huấn luyện: 0.8942"
## [1] "Adjusted R-squared trên tập huấn luyện: 0.8925"
## [1] "R-squared trên tập kiểm tra: 0.8825"
## [1] "RMSE trên tập kiểm tra (cho log_total_spent): 0.5066"
## [1] "RMSE trên tập kiểm tra (thang đo gốc của total_spent): 1621.35"
```

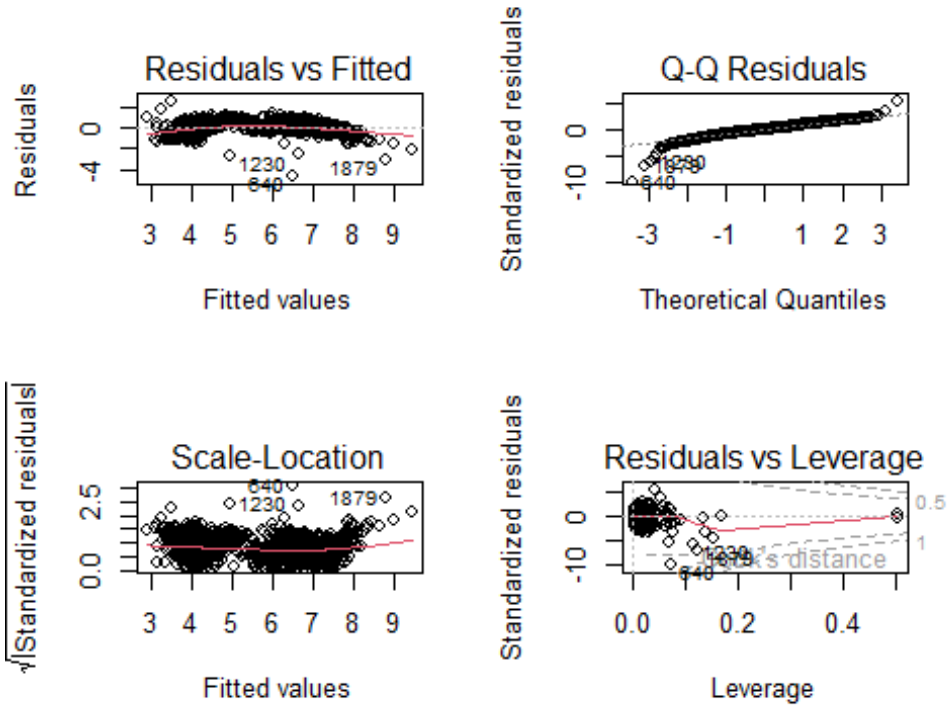
Nhận xét kết quả đánh giá:

- **R-squared (trên tập kiểm tra):** 0.8825. Mô hình giải thích được khoảng 88.25% sự biến thiên của log_total_spent trên dữ liệu mới, cho thấy mức độ phù hợp tốt. Giá trị này cũng khá gần với R-squared trên tập huấn luyện (0.8942), cho thấy mô hình không có dấu hiệu overfitting rõ rệt.
- **RMSE (trên tập kiểm tra, cho log_total_spent):** 0.5066. Đây là độ lệch chuẩn trung bình của sai số dự đoán trên thang đo logarit.
- **RMSE (trên tập kiểm tra, thang đo gốc của total_spent):** 1621.35. Sai số dự đoán trung bình cho tổng chi tiêu thực tế của khách hàng là khoảng 1621.35 đơn vị tiền tệ.

Nhìn chung, mô hình Hồi quy Tuyến tính cho thấy khả năng giải thích và dự đoán tốt về tổng chi tiêu của khách hàng.

6.3.2 Kiểm tra các giả định của mô hình

Các giả định của hồi quy tuyến tính được kiểm tra qua biểu đồ chẩn đoán.



Biểu đồ chẩn đoán cho mô hình Hồi quy Tuyến tính

Nhận xét các biểu đồ chẩn đoán:

1. **Residuals vs Fitted:**
    - Phần dư phân bố tương đối ngẫu nhiên quanh đường 0, không có hình mẫu rõ ràng, ủng hộ giả định tuyến tính và phương sai không đổi.
2. **Normal Q-Q (Quantile-Quantile):**
    - Các điểm phần dư bám khá sát đường thẳng, cho thấy phần dư xấp xỉ phân phối chuẩn, mặc dù có chút lệch ở hai đuôi.
3. **Scale-Location (hoặc Spread-Location):**
    - Đường xu hướng tương đối bằng phẳng, ủng hộ giả định phương sai không đổi.
4. **Residuals vs Leverage:**
    - Hầu hết các điểm có leverage thấp và không có điểm nào có vẻ là outlier ảnh hưởng quá lớn (nằm ngoài đường Cook's distance).

Kết luận về các giả định: Nhìn chung, các giả định của mô hình được đáp ứng ở mức chấp nhận được, cho phép diễn giải các hệ số.

6.3.3 Diễn giải hệ số

Phân tích các hệ số có ý nghĩa thống kê (Pr(>|t|) < 0.05) để xác định các yếu tố ảnh hưởng đến log_total_spent.

Các hệ số có ý nghĩa thống kê trong Mô hình Hồi quy Tuyến tính

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.355	0.496	6.765	0.000
Income	0.000	0.000	13.954	0.000
Complain	-0.321	0.118	-2.724	0.007
Child_Total	-0.280	0.022	-12.628	0.000
AcceptedCmp_Total	0.063	0.028	2.266	0.024
Days_Customer	0.000	0.000	7.241	0.000
NumDealsPurchases	0.086	0.009	9.336	0.000
NumWebPurchases	0.108	0.006	16.759	0.000
NumCatalogPurchases	0.081	0.007	12.394	0.000
NumStorePurchases	0.073	0.006	13.068	0.000
Cluster_KMeans2	0.164	0.080	2.054	0.040
Cluster_KMeans3	0.143	0.052	2.738	0.006
Cluster_KMeans4	-0.623	0.056	-11.153	0.000

Biến	**Ước lượng (Estimate)**	**Diễn giải Ảnh hưởng đến log_total_spent**
(Intercept)	3.355	Giá trị cơ sở của log_total_spent khi các biến độc lập khác bằng 0/mức tham chiếu.
Income	1.610e-05	Thu nhập tăng làm tăng log_total_spent (ảnh hưởng rất mạnh, p < 2e-16).
Complain	-0.321	Nếu khách hàng phàn nàn (Complain=1), log_total_spent giảm khoảng 0.321 đơn vị.
Child_Total	-0.280	Mỗi đứa con tăng thêm làm giảm log_total_spent khoảng 0.280 đơn vị (ảnh hưởng mạnh).
AcceptedCmp_Total	0.063	Mỗi chiến dịch trước được

		chấp nhận làm tăng log_total_spent khoảng 0.063 đơn vị.
Days_Customer	4.795e-04	Mỗi ngày gắn bó tăng thêm làm tăng nhẹ log_total_spent.
NumDealsPurchases	0.086	Số lần mua hàng giảm giá tăng làm tăng log_total_spent khoảng 0.086 đơn vị.
NumWebPurchases	0.108	Số lần mua qua web tăng làm tăng log_total_spent khoảng 0.108 đơn vị.
NumCatalogPurchases	0.081	Số lần mua qua catalog tăng làm tăng log_total_spent khoảng 0.081 đơn vị (ảnh hưởng mạnh).
NumStorePurchases	0.073	Số lần mua tại cửa hàng tăng làm tăng log_total_spent khoảng 0.073 đơn vị (ảnh hưởng mạnh).
Cluster_KMeans2	0.164	Thuộc Cụm 2 (so với Cụm 1) làm tăng log_total_spent khoảng 0.164 đơn vị.
Cluster_KMeans3	0.143	Thuộc Cụm 3 (so với Cụm 1) làm tăng log_total_spent khoảng 0.143 đơn vị.
Cluster_KMeans4	-0.623	Thuộc Cụm 4 (so với Cụm 1) làm giảm log_total_spent đáng kể khoảng 0.623 đơn vị (ảnh hưởng mạnh).
(Lưu ý: Recency, Education, Marital Status, Age, NumWebVisitsMonth không có ý nghĩa thống kê rõ rệt trong mô hình này)		

Phân tích các yếu tố ảnh hưởng đến log_total_spent:

- **Ảnh hưởng mạnh nhất đến tăng chi tiêu:** Income (Thu nhập).
- **Các yếu tố làm tăng chi tiêu đáng kể khác:** Mua hàng qua các kênh (NumWebPurchases, NumCatalogPurchases, NumStorePurchases), số lần mua hàng giảm giá (NumDealsPurchases), tổng số chiến dịch đã chấp nhận (AcceptedCmp_Total), thời gian gắn bó lâu hơn (Days_Customer), và thuộc các Phân khúc K-Means 2 và 3 (so với Cụm 1).
- **Các yếu tố làm giảm chi tiêu đáng kể:** Có nhiều con (Child_Total), việc khách hàng phàn nàn (Complain), và thuộc Phân khúc K-Means 4 (so với Cụm 1). Các yếu tố như Age, Recency, các mức độ Education và Marital_Status (so với mức tham chiếu "Absurd" và "2n Cycle") không cho thấy ảnh hưởng có ý nghĩa thống kê rõ rệt đến log_total_spent trong mô hình này sau khi đã kiểm soát các yếu tố khác.

Kết luận từ Hồi quy Tuyến tính: Mô hình Hồi quy Tuyến tính cho thấy thu nhập là yếu tố quan trọng nhất quyết định tổng chi tiêu của khách hàng. Bên cạnh đó, các hành vi mua sắm tích cực qua nhiều kênh, lịch sử tương tác tốt với chiến dịch, và thuộc các phân khúc khách hàng "Vàng" hoặc "Bạc" cũng góp phần làm tăng chi tiêu. Ngược lại, việc có nhiều con cái, từng phàn nàn, và thuộc phân khúc "Phổ thông" thường liên quan đến mức chi tiêu thấp hơn.

6.4 Thảo luận chung

Kết hợp kết quả từ Phân cụm K-Means, Hồi quy Logistic và Hồi quy Tuyến tính Đa biến mang lại một cái nhìn toàn diện hơn về khách hàng và các yếu tố ảnh hưởng đến hành vi của họ.

Phân khúc và Hành vi Dự đoán: Mô hình K-Means đã xác định 4 phân khúc khách hàng riêng biệt. Điều này được củng cố khi các biến đại diện cho việc thuộc các cụm này cho thấy ảnh hưởng có ý nghĩa thống kê đến cả khả năng chấp nhận ưu đãi (Response) và tổng chi tiêu (log_total_spent) trong các mô hình dự đoán. Cụ thể, khách hàng thuộc **Cụm 2 ("Vàng")** và **Cụm 3 ("Bạc")** không chỉ có thu nhập và chi tiêu cao (từ K-Means) mà còn có xu hướng chấp nhận ưu đãi và chi tiêu tổng thể cao hơn rõ rệt so với Cụm 1 (cụm tham chiếu). Ngược lại, **Cụm 4 ("Phổ thông")** có mức chi tiêu thấp nhất. Điều này cho thấy các phân khúc được xác định ban đầu thực sự phản ánh những khác biệt quan trọng trong hành vi và giá trị khách hàng.

Các yếu tố ảnh hưởng chính:

Lịch sử tương tác (AcceptedCmp_Total): Là yếu tố quan trọng, tác động tích cực đến cả khả năng phản hồi chiến dịch và tổng chi tiêu.

Đặc điểm gia đình (Child_Total): Số lượng con cái có ảnh hưởng tiêu cực nhất quán đến cả hai khía cạnh trên.

Thu nhập (Income): Quyết định mạnh mẽ đến tổng chi tiêu nhưng không có ảnh hưởng rõ rệt đến việc chấp nhận một ưu đãi cụ thể trong mô hình Logistic của nghiên cứu này.

Hạn chế chính: Phân tích này dựa trên dữ liệu có sẵn, có thể chưa bao gồm tất cả các yếu tố ảnh hưởng đến hành vi khách hàng. Một số nhóm nhỏ trong biến Marital_Status có thể ảnh hưởng đến độ ổn định của một số ước lượng. Ngoài ra, các mô hình thống kê luôn có những giả định nhất định và kết quả chỉ ra mối liên hệ chứ không khẳng định quan hệ nhân quả.

Tóm lại, việc kết hợp các phương pháp phân tích giúp vẽ nên một bức tranh khách hàng phong phú hơn, cung cấp những định hướng giá trị cho các chiến lược kinh doanh và marketing.

7. Kết luận (Conclusions)

Dự án này đã thực hiện một phân tích toàn diện về bộ dữ liệu khách hàng từ chiến dịch marketing, sử dụng ngôn ngữ R và áp dụng các kỹ thuật tiền xử lý dữ liệu, phân tích dữ liệu khám phá, cùng ba mô hình học máy chính: Phân cụm K-Means, Hồi quy Logistic và Hồi quy Tuyến tính Đa biến.

7.1 Tóm tắt kết quả chính

Qua quá trình phân tích dữ liệu khách hàng bằng ngôn ngữ R, dự án đã rút ra các kết quả nổi bật sau từ ba mô hình được xây dựng:

1. **Xác định 4 Phân khúc Khách hàng chính (K-Means):**
    - **“Khách hàng Vàng”:** Nhóm nhỏ nhất, thu nhập và chi tiêu cao nhất, ít con, tích cực phản hồi chiến dịch.
    - **“Khách hàng Bạc”:** Thu nhập và chi tiêu khá, ít con, nhưng ít tương tác với các chiến dịch cũ.
    - **“Khách hàng Trung thành”:** Thu nhập và chi tiêu trung bình, nhiều con, gắn bó lâu dài, nhạy cảm với giảm giá.
    - **“Khách hàng Phổ thông”:** Nhóm đông đảo nhất, thu nhập và chi tiêu thấp nhất, nhiều con, ít phản hồi chiến dịch.
2. **Các yếu tố dự đoán Khả năng Phản hồi Chiến dịch (Response - Hồi quy Logistic):**
    - Mô hình có khả năng phân biệt tốt (AUC ≈ 0.898).

- **Yếu tố tăng khả năng phản hồi:** Lịch sử chấp nhận các chiến dịch trước (AcceptedCmp_Total), trình độ học vấn cao (Thạc sĩ, Tiến sĩ), việc mua hàng gần đây (Recency thấp), và thuộc Phân khúc K-Means 3.
- **Yếu tố giảm khả năng phản hồi:** Số lượng con cái nhiều (Child_Total), tuổi tác cao hơn.

3. **Các yếu tố dự đoán Tổng Chi tiêu (log_total_spent - Hồi quy Tuyến tính):**
   - Mô hình giải thích tốt sự biến thiên của chi tiêu (R-squared trên tập kiểm tra ≈ 0.883).
   - **Yếu tố tăng chi tiêu:** Thu nhập (Income - ảnh hưởng mạnh nhất), lịch sử chấp nhận chiến dịch (AcceptedCmp_Total), thời gian gắn bó (Days_Customer), số lần mua qua các kênh (đặc biệt là catalog và tại cửa hàng), và thuộc Phân khúc K-Means 2 và 3.
   - **Yếu tố giảm chi tiêu:** Số lượng con cái nhiều (Child_Total), việc khách hàng từng phàn nàn (Complain), và thuộc Phân khúc K-Means 4.

Những kết quả này cung cấp một bức tranh chi tiết về các nhóm khách hàng khác nhau và các yếu tố chính điều khiển hành vi mua sắm cũng như phản hồi của họ đối với các hoạt động marketing.

7.2 Ý nghĩa và đề xuất

Các kết quả phân tích mang lại nhiều ý nghĩa thực tiễn, từ đó đưa ra các đề xuất cụ thể nhằm tối ưu hóa chiến lược marketing và kinh doanh:

1. **Cá nhân hóa chiến lược Marketing theo từng Phân khúc Khách hàng:**
   - **“Khách hàng Vàng”:** Ưu tiên các sản phẩm/dịch vụ cao cấp, chương trình chăm sóc đặc biệt, hạn chế ưu đãi giảm giá sâu để duy trì hình ảnh thương hiệu và giá trị khách hàng.
   - **“Khách hàng Bạc”:** Tập trung nuôi dưỡng mối quan hệ, giới thiệu sản phẩm mới phù hợp với mức chi tiêu khá, khuyến khích tương tác qua các kênh hiệu quả (ví dụ: catalog nếu họ mua nhiều qua kênh này).
   - **“Khách hàng Trung thành”:** Tiếp tục các chương trình giảm giá, ưu đãi cho gia đình, và tận dụng kênh web để thông báo khuyến mãi do họ có xu hướng mua hàng giảm giá và truy cập web nhiều.
   - **“Khách hàng Phổ thông”:** Cung cấp các sản phẩm giá cả phải chăng, gói combo tiết kiệm. Tập trung vào việc xây dựng nhận diện thương hiệu và các ưu đãi cơ bản.

2. **Tối ưu hóa Mục tiêu Chiến dịch dựa trên các Yếu tố Dự đoán Response:**
    - **Ưu tiên tiếp cận:** Những khách hàng đã từng chấp nhận các chiến dịch trước (AcceptedCmp_Total), có lịch sử mua hàng gần đây (Recency thấp), và thuộc các phân khúc có khả năng phản hồi cao (ví dụ: Phân khúc K-Means 3).
    - **Điều chỉnh thông điệp:** Cá nhân hóa thông điệp cho nhóm khách hàng có trình độ học vấn cao. Cân nhắc các ưu đãi khác nhau cho nhóm có nhiều con.
3. **Khai thác Tiềm năng Chi tiêu từ các Nhóm Khách hàng:**
    - **Thúc đẩy các kênh hiệu quả:** Tăng cường các kênh mua sắm có liên quan mạnh đến tổng chi tiêu cao như mua qua catalog (NumCatalogPurchases) và tại cửa hàng (NumStorePurchases).
    - **Xử lý phàn nàn:** Giải quyết tốt các phàn nàn (Complain) có thể giúp giữ chân và tăng chi tiêu của khách hàng.
    - **Chăm sóc khách hàng lâu năm:** Ghi nhận và có chính sách ưu đãi cho những khách hàng đã gắn bó lâu dài (Days_Customer).

Các đề xuất này, khi được triển khai một cách hợp lý, có thể giúp doanh nghiệp tăng cường hiệu quả marketing, cải thiện mối quan hệ với khách hàng và tối đa hóa doanh thu từ các phân khúc khác nhau.

7.3 Hướng phát triển tương lai

Để tiếp tục khai thác giá trị từ dữ liệu khách hàng và nâng cao hiệu quả kinh doanh, một số hướng phát triển sau đây có thể được xem xét:

1. **Thử nghiệm các mô hình nâng cao hơn:** Áp dụng các thuật toán học máy phức tạp hơn như Gradient Boosting hoặc Support Vector Machines để có thể cải thiện độ chính xác dự đoán cho cả khả năng phản hồi (Response) và tổng chi tiêu (log_total_spent).
2. **Phân tích hành vi theo kênh mua hàng:** Nghiên cứu sâu hơn về cách khách hàng tương tác và mua sắm trên từng kênh cụ thể (web, catalog, cửa hàng) và sự ảnh hưởng qua lại giữa các kênh này.
3. **Xây dựng mô hình Giá trị Vòng đời Khách hàng (CLV):** Dự đoán tổng giá trị mà một khách hàng dự kiến sẽ mang lại cho doanh nghiệp trong suốt thời gian họ còn là khách hàng, giúp ưu tiên nguồn lực chăm sóc hiệu quả.
4. **Phân tích Giỏ hàng (Market Basket Analysis):** Xác định các sản phẩm thường được mua cùng nhau để đưa ra các gợi ý bán chéo (cross-selling) và bán thêm (up-selling) phù hợp.

5. **Bổ sung và tích hợp thêm dữ liệu:** Thu thập thêm các thông tin chi tiết hơn về sở thích cá nhân, phản hồi về sản phẩm/dịch vụ, hoặc dữ liệu tương tác trên các nền tảng trực tuyến khác để có cái nhìn toàn diện hơn về khách hàng.
6. **Tối ưu hóa ngưỡng cho mô hình phân loại:** Điều chỉnh ngưỡng xác suất (thay vì mặc định 0.5) cho mô hình Hồi quy Logistic để cải thiện các chỉ số quan trọng như tỷ lệ phát hiện đúng khách hàng tiềm năng (Recall), tùy theo mục tiêu cụ thể của từng chiến dịch.

Việc theo đuổi các hướng này sẽ giúp doanh nghiệp liên tục cải tiến hiểu biết về khách hàng và đưa ra các quyết định kinh doanh ngày càng chính xác và hiệu quả hơn.

8. Phụ lục (Appendices)

9. Đóng góp (Contributions)

- **Đỗ Kiến Hưng (23133030):**
    - Chịu trách nhiệm chính trong việc tổng hợp nội dung và viết báo cáo hoàn chỉnh, bao gồm các phần lý thuyết nền tảng, mô tả dữ liệu, phân tích dữ liệu khám phá (EDA), và các phần kết luận, đề xuất.
    - Đảm bảo tính logic, nhất quán và hoàn thiện chung của toàn bộ tài liệu đồ án.
- **Nguyễn Văn Quang Duy (23110086):**
    - Phụ trách nghiên cứu, triển khai mã nguồn R, phân tích kết quả và diễn giải các hệ số cho mô hình Hồi quy Logistic (dự đoán biến Response).
    - Đóng góp vào việc đánh giá hiệu suất của mô hình này.
- **Phan Trọng Phú (23133056):**
    - Phụ trách nghiên cứu, triển khai mã nguồn R, phân tích kết quả và diễn giải các hệ số cho mô hình Hồi quy Tuyến tính Đa biến (dự đoán biến log_total_spent).
    - Đóng góp vào việc kiểm tra các giả định của mô hình này.
- **Phan Trọng Quí (23133061):**
    - Phụ trách nghiên cứu, triển khai mã nguồn R và phân tích kết quả cho mô hình Phân cụm K-Means.
    - Đóng góp vào việc xác định số cụm tối ưu và mô tả đặc điểm của từng phân khúc khách hàng.

10. Tham khảo (References)

- **Bộ dữ liệu:**
  - Makash, A. (2022). *Customer Personality Analysis*. Kaggle. Truy cập tại: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis
- **Sách tham khảo chính:**
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.
  - Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.
- **Tài liệu môn học:**
  - Tài liệu hướng dẫn môn học "Lập trình R cho phân tích" (HCMUTE)
- **Các gói R đã sử dụng (Ví dụ một số gói chính, không cần liệt kê tất cả nếu quá nhiều):**
  - dplyr (Wickham H. et al.) - Cho thao tác dữ liệu.
  - ggplot2 (Wickham H.) - Cho trực quan hóa dữ liệu.
  - caret (Kuhn M.) - Cho chia dữ liệu và một số tiện ích mô hình hóa.
  - factoextra (Kassambara A. & Mundt F.) - Hỗ trợ phân cụm và trực quan hóa.
  - cluster (Maechler M. et al.) - Chứa thuật toán phân cụm.
  - pROC (Robin X. et al.) - Cho phân tích đường cong ROC.
  - car (Fox J. & Weisberg S.) - Hỗ trợ kiểm tra giả định hồi quy (ví dụ: VIF).

11. Peer Assessment

Đánh giá chung về quá trình làm việc nhóm: Nhóm đã có sự phối hợp tốt trong suốt quá trình thực hiện đồ án. Các buổi họp nhóm diễn ra đều đặn và hiệu quả trong việc trao đổi ý tưởng, giải quyết vấn đề và phân chia công việc. Nhìn chung, các thành viên đều nỗ lực hoàn thành phần việc được giao.

Đánh giá từng thành viên:

1. **Đỗ Kiến Hưng:**
   - **Nội dung triển khai được:** Tổng hợp và viết báo cáo chính, bao gồm các phần lý thuyết, mô tả dữ liệu, EDA, kết luận, và các phần phụ.
   - **Mức độ hoàn thành:** Tốt

- **Ưu điểm:** Khả năng tổng hợp thông tin tốt, đảm bảo tiến độ chung của báo cáo.

2. **Nguyễn Văn Quang Duy:**
    - **Nội dung triển khai được:** Xây dựng và phân tích mô hình Hồi quy Logistic.
    - **Mức độ hoàn thành:** Tốt
    - **Ưu điểm:** Nắm vững kiến thức về Hồi quy Logistic, thực hiện code cẩn thận, đánh giá mô hình chi tiết
3. **Phan Trọng Phú:**
    - **Nội dung triển khai được:** Xây dựng và phân tích mô hình Hồi quy Tuyến tính Đa biến.
    - **Mức độ hoàn thành:** Tốt
    - **Ưu điểm:** Tìm hiểu kỹ về các giả định của mô hình, diễn giải hệ số rõ ràng, đóng góp tích cực vào thảo luận chung
4. **Phan Trọng Quí:**
    - **Nội dung triển khai được:** Xây dựng và phân tích mô hình Phân cụm K-Means.
    - **Mức độ hoàn thành:** Tốt
    - **Ưu điểm:** Nghiên cứu kỹ các phương pháp xác định số cụm, phân tích đặc điểm cụm sâu sắc, trình bày kết quả trực quan