**PHÂN TÍCH DỮ LIỆU MARKETING CAMPAIGN CÙNG AUTOVIZ**

**1. TỔNG QUAN**

**1.1. Mục tiêu**

Phân tích hành vi khách hàng và hiệu quả của các chiến dịch marketing để đưa ra insights và recommendations cho chiến lược marketing trong tương lai.

**1.2. Nguồn dữ liệu**

* **Dataset**: Customer Personality Analysis
* **Nguồn**: Kaggle (https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis)
* **Kích thước**: 2,240 khách hàng với 29 thuộc tính
* **Thời gian**: Dữ liệu từ năm 2012-2014

**2. QUY TRÌNH XỬ LÝ DỮ LIỆU**

**2.1. Làm sạch dữ liệu**

* **Xử lý giá trị null**: Phát hiện 24 giá trị null trong cột Income, được điền bằng giá trị median
* **Loại bỏ outliers**: Loại bỏ các khách hàng có tuổi bất thường (< 18 hoặc > 100 tuổi)
* **Chuẩn hóa định dạng**: Chuyển đổi cột Dt\_Customer sang định dạng datetime

**2.2. Feature Engineering**

Tạo các biến mới để phân tích sâu hơn:

| **Feature** | **Công thức** | **Mục đích** |
| --- | --- | --- |
| Age | 2014 - Year\_Birth | Tính tuổi khách hàng |
| Customer\_Days | Ngày hiện tại - Dt\_Customer | Số ngày là khách hàng |
| Total\_Purchases | Tổng các kênh mua hàng | Đo lường tần suất mua |
| Total\_Amount\_Spent | Tổng chi tiêu tất cả sản phẩm | Giá trị khách hàng |
| Total\_Accepted\_Campaigns | Tổng các chiến dịch chấp nhận | Mức độ tương tác |
| Total\_Children | Kidhome + Teenhome | Số con trong gia đình |
| Spending\_Segment | Phân khúc theo chi tiêu | Phân loại khách hàng |

**3. PHÂN TÍCH KHÁM PHÁ DỮ LIỆU (EDA)**

**3.1. Đặc điểm Khách hàng**

**3.1.1. Nhân khẩu học**

* **Độ tuổi trung bình**: ~52 tuổi
* **Thu nhập trung bình**: ~52,000 USD/năm
* **Trình độ học vấn**:
  + Graduation: ~50%
  + PhD: ~21%
  + Master: ~17%
  + 2n Cycle: ~8%
  + Basic: ~4%

**3.1.2. Tình trạng gia đình**

* **Married**: Chiếm tỷ lệ cao nhất (~38%)
* **Together**: ~26%
* **Single**: ~21%
* **Divorced**: ~11%
* **Widow**: ~4%

**3.2. Hành vi Mua hàng**

**3.2.1. Kênh mua hàng ưa thích**

* **Store Purchases** (Mua tại cửa hàng): Kênh phổ biến nhất
* **Web Purchases** (Mua online): Kênh thứ hai
* **Catalog Purchases**: Ít phổ biến hơn
* **Deals Purchases**: Tần suất thấp nhất

**3.2.2. Danh mục sản phẩm**

Chi tiêu theo danh mục (theo tỷ lệ %):

1. **Wines**: ~40-45% tổng chi tiêu
2. **Meat Products**: ~25-30%
3. **Gold Products**: ~10-15%
4. **Fish Products**: ~8-12%
5. **Sweet Products**: ~3-5%
6. **Fruits**: ~2-4%

**Insight**: Rượu và thịt là 2 danh mục sản phẩm chủ lực, chiếm ~70% doanh thu.

**3.3. Phân khúc Khách hàng**

**Theo mức chi tiêu (Spending\_Segment):**

* **Low** (< $100): Khách hàng ít chi tiêu
* **Medium** ($100-$500): Phân khúc lớn nhất
* **High** ($500-$1,000): Khách hàng có giá trị
* **Premium** (> $1,000): Khách hàng VIP, cần chăm sóc đặc biệt

**4. PHÂN TÍCH HIỆU QUẢ CHIẾN DỊCH**

**4.1. Tỷ lệ chấp nhận chiến dịch**

| **Chiến dịch** | **Tỷ lệ chấp nhận** | **Đánh giá** |
| --- | --- | --- |
| Campaign 1 | ~6.4% | Trung bình |
| Campaign 2 | ~1.3% | Thấp |
| Campaign 3 | ~7.3% | Cao nhất |
| Campaign 4 | ~7.5% | Cao nhất |
| Campaign 5 | ~7.3% | Cao |
| Last Campaign | ~15% | Rất cao |

**Phát hiện quan trọng**:

* Campaign 2 có hiệu quả kém nhất (1.3%)
* Last Campaign (chiến dịch gần nhất) có tỷ lệ chấp nhận cao nhất (15%)
* Campaigns 3, 4, 5 có hiệu quả tương đương nhau (~7%)

**4.2. Yếu tố ảnh hưởng đến chấp nhận chiến dịch**

Phân tích correlation cho thấy:

* **Thu nhập cao** → Tỷ lệ chấp nhận cao hơn
* **Số con nhiều** → Tỷ lệ chấp nhận thấp hơn
* **Chi tiêu cao** → Xu hướng chấp nhận nhiều hơn
* **Recency thấp** (mua hàng gần đây) → Tỷ lệ chấp nhận cao hơn

**5. PHÂN TÍCH TƯƠNG QUAN**

**5.1. Các yếu tố ảnh hưởng đến Total\_Amount\_Spent**

**Tương quan dương mạnh**:

* **Income** (r ≈ 0.7-0.8): Thu nhập là yếu tố quan trọng nhất
* **Total\_Purchases** (r ≈ 0.6-0.7): Tần suất mua cao → Chi tiêu cao

**Tương quan âm**:

* **Total\_Children** (r ≈ -0.2 đến -0.3): Có con → Chi tiêu giảm
* **NumWebVisitsMonth**: Truy cập web nhiều nhưng không mua → Chi tiêu thấp

**Không tương quan mạnh**:

* **Age**: Tuổi tác không ảnh hưởng nhiều đến chi tiêu
* **Recency**: Thời gian mua hàng gần đây có ảnh hưởng nhỏ

**6. INSIGHTS QUAN TRỌNG**

**6.1. Về Khách hàng**

1. **Khách hàng có giá trị cao**:
   * Thu nhập > 70,000 USD
   * Không có hoặc ít con
   * Học vấn cao (PhD, Master, Graduation)
   * Tình trạng: Married hoặc Together
2. **Khách hàng tiềm năng**:
   * Thu nhập trung bình (40,000-60,000 USD)
   * Mua hàng thường xuyên tại store
   * Ít truy cập web nhưng conversion cao
3. **Khách hàng cần chăm sóc**:
   * Nhiều con, thu nhập trung bình
   * Quan tâm đến deals và promotions
   * Nhạy cảm với giá

**6.2. Về Sản phẩm**

1. **Wines** là danh mục chủ lực, cần:
   * Đầu tư marketing mạnh
   * Phát triển dòng sản phẩm cao cấp
   * Chương trình loyalty cho wine lovers
2. **Meat Products** là danh mục thứ hai:
   * Cross-selling với Wines
   * Bundle promotions
3. **Fruits & Sweets** có doanh thu thấp:
   * Cần đánh giá lại chiến lược
   * Xem xét ngưng hoặc điều chỉnh

**6.3. Về Chiến dịch Marketing**

1. **Campaign 2 thất bại** - Cần phân tích:
   * Thông điệp không phù hợp?
   * Timing không đúng?
   * Targeting sai đối tượng?
2. **Last Campaign thành công** - Học hỏi:
   * Personalization tốt hơn?
   * Offer hấp dẫn hơn?
   * Timing phù hợp hơn?
3. **Multi-channel approach** cần thiết:
   * Store vẫn là kênh chính
   * Web đang phát triển
   * Catalog đang giảm

**7. KHUYẾN NGHỊ CHIẾN LƯỢC**

**7.1. Phân khúc & Targeting**

**Khuyến nghị 1: Focus vào Premium Segment**

* Tập trung 80% ngân sách marketing vào 20% khách hàng chi tiêu cao
* Tạo chương trình VIP với exclusive benefits
* Personalized communication

**Khuyến nghị 2: Re-activate Low Spenders**

* Chiến dịch win-back với special offers
* Survey để hiểu rõ barriers
* Entry-level products với giá hấp dẫn

**7.2. Product Strategy**

**Khuyến nghị 3: Tăng cường Wine Category**

* Wine tasting events
* Wine subscription program
* Educational content về wines

**Khuyến nghị 4: Cross-selling thông minh**

* Bundle: Wine + Meat Products
* Recipe suggestions
* Meal planning content

**Khuyến nghị 5: Đánh giá lại Fruits & Sweets**

* A/B testing với promotions
* Nếu không cải thiện → Consider phasing out
* Reallocate resources sang categories có ROI cao hơn

**7.3. Campaign Optimization**

**Khuyến nghị 6: Áp dụng learnings từ Last Campaign**

* Replicate success factors
* Personalization dựa trên purchase history
* Timing: Sau 30-45 ngày kể từ last purchase

**Khuyến nghị 7: Cải thiện Campaign Acceptance Rate**

* Current rate: ~7-15%
* Target: 20-25%
* Tactics:
  + Better segmentation
  + Personalized offers
  + A/B testing messages
  + Optimal timing

**Khuyến nghị 8: Multi-channel Integration**

* Online-to-Offline strategy
* Click & Collect
* Unified customer experience
* Retargeting web visitors

**7.4. Customer Retention**

**Khuyến nghị 9: Loyalty Program**

* Points-based rewards
* Tiered membership (Silver, Gold, Platinum)
* Exclusive access to new products
* Birthday & anniversary offers

**Khuyến nghị 10: Reduce Churn**

* Identify at-risk customers (high Recency, low purchases)
* Proactive outreach
* Win-back campaigns với incentives

**8. KẾT LUẬN**

**8.1. Tóm tắt**

Phân tích dữ liệu Marketing Campaign đã cung cấp insights sâu sắc về:

* Đặc điểm và hành vi khách hàng
* Hiệu quả của các chiến dịch marketing
* Cơ hội tối ưu hóa doanh thu

**8.2. Key Takeaways**

1. **Thu nhập là yếu tố quyết định** chi tiêu khách hàng
2. **Wines & Meat Products** là 2 danh mục chủ lực
3. **Store channel** vẫn dominant nhưng web đang phát triển
4. **Last Campaign** thành công nhất → cần replicate
5. **Premium segment** (>$1,000) cần chăm sóc đặc biệt

**8.3. Next Steps**

1. **Ngắn hạn** (1-3 tháng):
   * Implement VIP program cho Premium customers
   * Launch improved campaigns dựa trên Last Campaign learnings
   * A/B testing cho Fruits & Sweets category
2. **Trung hạn** (3-6 tháng):
   * Phát triển wine subscription program
   * Tối ưu hóa multi-channel experience
   * Build predictive model cho campaign acceptance
3. **Dài hạn** (6-12 tháng):
   * Customer Lifetime Value (CLV) modeling
   * Advanced personalization engine
   * Automated marketing campaigns

**9. SỬ DỤNG AUTOVIZ ĐỂ PHÂN TÍCH DỮ LIỆU**

**1. Giới thiệu về AutoViz**

AutoViz là một thư viện Python mã nguồn mở giúp tự động hóa quá trình trực quan hóa dữ liệu chỉ bằng một dòng mã. Đây là công cụ hữu ích cho các nhà khoa học dữ liệu và nhà phân tích để nhanh chóng khám phá và hiểu rõ hơn về bộ dữ liệu của họ.

**Các tính năng chính**

* Tự động hóa: AutoViz tự động chọn các biểu đồ phù hợp nhất để hiển thị mối quanah hệ giữa các biến trong dữ liệu của bạn. Điều này giúp tiết kiệm thời gian và công sức so với việc tạo từng biểu đồ một cách thủ công.
* Toàn diện: Thư viện này có thể tạo ra nhiều loại biểu đồ khác nhau, bao gồm biểu đồ phân tán, biểu đồ tần suất, biểu đồ thanh, biểu đồ hộp, và bản đồ nhiệt. Điều này cho phép bạn có được cái nhìn tổng quan về dữ liệu từ nhiều góc độ.
* Đơn giản: Chỉ với một hàm duy nhất, AutoViz\_Class().AutoViz(), bạn có thể tạo ra tất cả các biểu đồ cần thiết.

**Lợi ích khi sử dụng AutoViz**

* Tốc độ: Nhanh chóng tạo ra nhiều biểu đồ để khám phá dữ liệu mà không cần viết nhiều mã.
* Dễ sử dụng: Giao diện đơn giản giúp người dùng, ngay cả những người mới bắt đầu, có thể dễ dàng sử dụng.
* Hiểu biết sâu sắc: Bằng cách trực quan hóa dữ liệu, AutoViz giúp phát hiện các mẫu, các mối tương quan, và các giá trị ngoại lai mà có thể khó nhận thấy khi chỉ nhìn vào dữ liệu thô.

**Cách cài đặt và sử dụng**

Bạn có thể cài đặt AutoViz một cách dễ dàng bằng pip:

**Bash**

*pip install autoviz*

Sau khi cài đặt, bạn có thể sử dụng thư viện này để phân tích bộ dữ liệu của mình. Dưới đây là một ví dụ đơn giản:

**Python**

*from autoviz.AutoViz\_Class import AutoViz\_Class*

*import pandas as pd*

*# Tải bộ dữ liệu mẫu*

*df = pd.read\_csv('your\_dataset.csv')*

*# Khởi tạo lớp AutoViz*

*AV = AutoViz\_Class()*

*# Chạy AutoViz*

*report = AV.AutoViz(df)*

Sau khi chạy đoạn mã trên, AutoViz sẽ tự động tạo và hiển thị một loạt các biểu đồ trực quan hóa dữ liệu từ tệp your\_dataset.csv của bạn.

**2. Sử dụng AutoViz để phân tích dữ liệu marketing campaign**

**Kết quả:**

**Đầu tiên, AutoViz đã đọc và xử lí dữ liệu từ file đã cung cấp**

**BƯỚC 1: ĐỌC VÀ XỬ LÝ DỮ LIỆU**

✓ Đã đọc dữ liệu: 2240 dòng, 29 cột

Các cột trong dataset:

1. ID

2. Year\_Birth

3. Education

4. Marital\_Status

5. Income

6. Kidhome

7. Teenhome

8. Dt\_Customer

9. Recency

10. MntWines

11. MntFruits

12. MntMeatProducts

13. MntFishProducts

14. MntSweetProducts

15. MntGoldProds

16. NumDealsPurchases

17. NumWebPurchases

18. NumCatalogPurchases

19. NumStorePurchases

20. NumWebVisitsMonth

21. AcceptedCmp3

22. AcceptedCmp4

23. AcceptedCmp5

24. AcceptedCmp1

25. AcceptedCmp2

26. Complain

27. Z\_CostContact

28. Z\_Revenue

29. Response

**Sau đó tiến hành xử lí dữ liệu**

Đã điền 24 giá trị null trong Income bằng median: $51,381.50

* Cột “Income” thường có vài giá trị bị thiếu.
* Dùng **giá trị trung vị (median)** để thay thế.  
  => **Mục đích:** Loại bỏ lỗi do dữ liệu thiếu, tránh sai số khi vẽ biểu đồ hoặc tính toán.

Đã chuyển đổi Dt\_Customer và tạo Customer\_Days, Customer\_Years

* Chuyển cột ngày thành dạng **datetime**.
* Tạo thêm 2 biến: số **ngày** và **năm** mà khách hàng đã gắn bó.  
  => **Mục đích:** Dễ phân tích hành vi theo thời gian (ví dụ khách hàng trung thành lâu năm chi tiêu nhiều hơn).

Đã tính tuổi từ Year\_Birth

* Năm dữ liệu kết thúc là 2014, nên lấy 2014 - năm sinh.  
  => **Mục đích:** Có biến tuổi để phân tích theo độ tuổi.

Đã loại bỏ 5 outliers trong Age

* Loại bỏ khách hàng quá nhỏ tuổi hoặc quá cao tuổi (dữ liệu lỗi).  
  => **Mục đích:** Làm sạch dữ liệu, tăng độ tin cậy khi trực quan hóa.

Đã tạo các feature: Total\_Purchases, Total\_Amount\_Spent, Total\_Accepted\_Campaigns, Total\_Children

* **Mục đích:** Gom nhiều cột chi tiết thành chỉ số tổng quan:
* Tổng số lần mua hàng.
* Tổng số tiền chi tiêu.
* Tổng số chiến dịch khách hàng đã phản hồi.
* Tổng số con (để phân tích theo quy mô gia đình).

Đã tạo Spending\_Segment (Low, Medium, High, Premium)

* **Mục đích:** Tạo nhãn nhóm chi tiêu (Low → Premium) để dễ phân tích hành vi từng nhóm khách hàng

Kích thước dữ liệu sau xử lý: 2235 dòng, 37 cột

Đã lưu dữ liệu đã xử lý vào 'marketing\_campaign\_processed.csv'

**BƯỚC 2: SỬ DỤNG AUTOVIZ ĐỂ PHÂN TÍCH DỮ LIỆU**

AutoViz sẽ tự động:

• Phân tích tất cả các biến

• Tạo biểu đồ phù hợp cho từng loại dữ liệu

• Tìm mối quan hệ với biến mục tiêu (Total\_Amount\_Spent)

Giải thích các biểu đồ mà AutoViz tạo ra:  
AutoViz tự động tạo nhiều loại biểu đồ khác nhau, được tổ chức theo các nhóm:

**1. SCATTER PLOTS (Biểu đồ phân tán)**

Mục đích:

• Hiển thị mối quan hệ giữa 2 biến liên tục

• Xác định correlation (tương quan)

• Phát hiện outliers và patterns

Cách đọc:

• Mỗi điểm = 1 khách hàng

• Trục X: Biến độc lập (ví dụ: Income)

• Trục Y: Total\_Amount\_Spent (biến mục tiêu)

• Màu sắc: Phân biệt theo nhóm (nếu có)

Ví dụ biểu đồ AutoViz tạo:

• Income vs Total\_Amount\_Spent

  → Nếu các điểm tạo thành đường thẳng đi lên: Correlation dương

  → Income cao → Chi tiêu cao

• Age vs Total\_Amount\_Spent

  → Nếu các điểm rải rác: Không có correlation rõ ràng

  → Tuổi không ảnh hưởng nhiều đến chi tiêu

• Total\_Purchases vs Total\_Amount\_Spent

  → Mua nhiều lần → Chi tiêu tổng cao

  → Quan trọng cho chiến lược loyalty

💡 INSIGHT: Tìm các biến có correlation mạnh để:

   - Targeting customers

   - Predictive modeling

   - Segmentation

**2. DISTRIBUTION PLOTS (Biểu đồ phân phối)**

Mục đích:

• Hiển thị cách dữ liệu được phân bố

• Xác định skewness (độ lệch)

• Tìm outliers

Loại biểu đồ:

a) Histogram:

   • Trục X: Giá trị của biến

   • Trục Y: Tần suất (số lượng)

   • Hình dạng cho biết distribution type

b) KDE Plot (Kernel Density Estimate):

   • Đường cong mượt mà thể hiện density

   • Peak = giá trị phổ biến nhất

Ví dụ với Total\_Amount\_Spent:

• Hình dạng: Right-skewed (lệch phải)

  → Phần lớn khách chi tiêu thấp

  → Một số ít chi tiêu rất cao (VIP)

• Mean > Median

  → Có outliers ở phía cao

  → Top customers kéo trung bình lên

INSIGHT:

   - Nếu right-skewed: Focus vào top customers

   - Nếu normal distribution: Mass market strategy

   - Outliers = VIP customers cần retention

**3. VIOLIN PLOTS (Biểu đồ Violin)**

Mục đích:

• So sánh phân phối giữa các nhóm categorical

• Kết hợp box plot và density plot

Cách đọc:

• Chiều rộng = Density (nhiều data points)

• Chiều dài = Phạm vi giá trị

• Đường giữa = Median

Ví dụ: Total\_Amount\_Spent theo Education

• PhD: Violin hẹp ở trên → Chi tiêu tập trung ở mức cao

• Basic: Violin rộng ở dưới → Chi tiêu tập trung ở mức thấp

• Graduation: Violin rộng cả 2 đầu → Đa dạng

💡 INSIGHT:

   - Nhóm nào có spending cao và ổn định?

   - Nhóm nào có potential để upsell?

   - Nên target nhóm nào?

**4. HEATMAP - MA TRẬN TƯƠNG QUAN**

Mục đích:

• Hiển thị correlation giữa TẤT CẢ các biến số

• Tìm multicollinearity

• Xác định biến quan trọng

Cách đọc:

• Màu đỏ đậm: Correlation dương mạnh (+0.7 đến +1.0)

• Màu xanh đậm: Correlation âm mạnh (-0.7 đến -1.0)

• Màu trắng: Không correlation (0)

Ví dụ với Total\_Amount\_Spent:

• Income: +0.8 (đỏ đậm) → Quan trọng nhất

• Total\_Purchases: +0.7 (đỏ) → Rất quan trọng

• Total\_Children: -0.3 (xanh nhạt) → Có con = chi tiêu giảm

• Age: ~0 (trắng) → Không ảnh hưởng

INSIGHT:

   - Focus vào biến có correlation cao

   - Tránh multicollinearity trong modeling

   - Hiểu được drivers của spending

**5. BAR PLOTS (Biểu đồ cột)**

Mục đích:

• So sánh giá trị trung bình giữa các nhóm categorical

• Xác định nhóm nào perform tốt nhất

Ví dụ: Average Total\_Amount\_Spent theo Education

• PhD: Cột cao nhất → Chi tiêu trung bình cao nhất

• Master: Cột cao thứ 2

• Graduation: Cột trung bình

• Basic: Cột thấp nhất

Với Marital\_Status:

• Together/Married: Chi tiêu cao

• Single: Chi tiêu trung bình

• Widow: Chi tiêu thấp

💡 INSIGHT:

   - Segment nào profitable nhất?

   - Nên focus marketing vào đâu?

   - Opportunity để grow các segment thấp?

**6. PAIR PLOTS (Biểu đồ cặp)**

Mục đích:

• Hiển thị mối quan hệ giữa NHIỀU biến cùng lúc

• Ma trận scatter plots

Cách đọc:

• Đường chéo: Distribution của từng biến

• Ô khác: Scatter plot giữa 2 biến

• Màu: Phân biệt theo target variable

Ví dụ 4x4 pair plot với:

• Income, Age, Total\_Purchases, Total\_Amount\_Spent

Nhìn vào pair plot có thể thấy:

• Income-Total\_Amount\_Spent: Strong positive correlation

• Age-Total\_Amount\_Spent: Weak correlation

• Total\_Purchases-Total\_Amount\_Spent: Strong positive

**INSIGHT:**

   - Nhìn tổng quan mối quan hệ

   - Tìm patterns và clusters

   - Feature selection cho modeling