



Group 4



# MID-TERM REPORT



# INTRODUCTION

---



# TABLE OF CONTENT

---

01

Giới thiệu đề tài và  
bộ dữ liệu

02

Phân tích khám  
phá dữ liệu (EDA)

03

Xây dựng đặc  
trưng

04

Huấn luyện mô  
hình

05

Đánh giá và cải  
thiện mô hình

06

Phân cụm khách  
hang

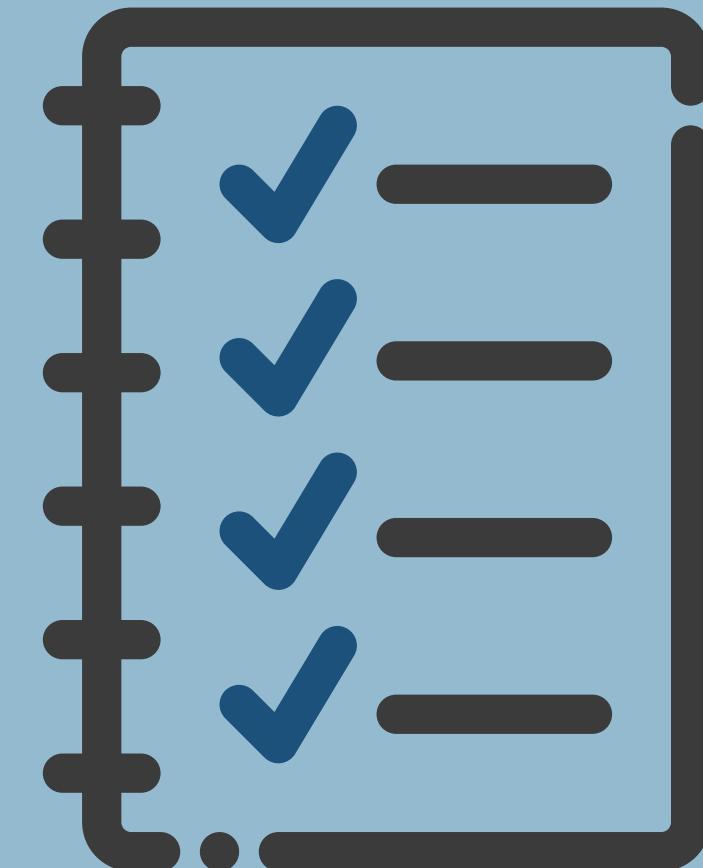
07

Phân tích khám phá  
các cụm khách hàng

08

Kết luận & Đề  
xuất

CONTENT



# MEMBER'S CONTRIBUTION

---

| THÀNH VIÊN                      | CÔNG VIỆC  | % ĐÓNG GÓP |
|---------------------------------|--|------------|
| Trần Quốc Toản<br>(Nhóm trưởng) | <ul style="list-style-type: none"><li>Bao quát công việc chung, phân công, và góp ý</li><li>Đóng góp phân cụm khách hàng</li></ul>         | 25         |
| Huỳnh Phương Anh                | Phân tích khám phá dữ liệu và hỗ trợ đánh giá mô hình  | 25         |
| Nguyễn Thị Diễm Ly              | Xây dựng đặc trưng, huấn luyện mô hình và đánh giá   | 25         |
| Phan Trường Huy                 | <ul style="list-style-type: none"><li>Xây dựng hướng đi cho bài toán</li><li>Phân cụm khách hàng, rút insights và đưa ra đề xuất</li></ul> | 25         |

# OVERVIEW PROJECT

---

## ỨNG DỤNG MÔ HÌNH HỌC MÁY TRONG CHIẾN DỊCH DIGITAL MARKETING

- Dự đoán khả năng chuyển đổi của khách hàng
- Phân đoạn khách hàng



# OVERVIEW DATASET

---

- Đây là bộ dữ liệu cung cấp góc nhìn về hành vi tương tác của khách hàng trong chiến dịch tiếp thị kỹ thuật số

- Thông tin tổng quan

- Dữ liệu từ 8000 khách hàng
- Thông tin về nhân khẩu học
- Hành vi tương tác của khách hàng
- Hiệu quả chiến dịch tiếp thị



# OVERVIEW DATASET

| THÔNG TIN KHÁCH HÀNG |              |  |
|----------------------|--------------|--|
| Tên trường           | Kiểu dữ liệu | Mô Tả                                    |
| CustomerID           | String       | Mã định danh duy nhất cho mỗi khách hàng |
| Age                  | Integer      | Tuổi của khách hàng                      |
| Gender               | String       | Giới tính của khách hàng (Nam/Nữ)        |
| Income               | Float        | Thu nhập hàng năm của khách hàng (USD)   |
| PreviousPurchases    | Integer      | Số lần mua hàng trước đây của khách hàng |
| LoyaltyPoints        | Float        | Số điểm tích lũy của khách hàng          |

# OVERVIEW DATASET

---

| THÔNG TIN CHIẾN DỊCH |              |   |
|----------------------|--------------|---|
| Tên trường           | Kiểu dữ liệu | Mô Tả                                     |
| CampaignType         | String       | Loại chiến dịch tiếp thị                  |
| CampaignChannel      | String       | Kênh thực hiện chiến dịch tiếp thị        |
| AdSpend              | Float        | Số tiền chi cho chiến dịch tiếp thị (USD) |
| AdvertisingTool      |              | Thông tin được bảo mật                    |
| AdvertisingPlatform  |              | Thông tin được bảo mật                    |

# OVERVIEW DATASET

---

| HÀNH VI TƯƠNG TÁC |              |   |
|-------------------|--------------|---|
| Tên trường        | Kiểu dữ liệu | Mô Tả   |
| ClickThroughRate  | Float        | Tỷ lệ nhấp chuột vào nội dung tiếp thị              |
| ConversionRate    | Float        | Tỷ lệ chuyển đổi từ nhấp chuột thành hành động      |
| WebsiteVisits     | Integer      | Số lần truy cập vào trang web                       |
| PagePerVisits     | Float        | Số trang trung bình được xem trong mỗi lần truy cập |
| TimeOnSite        | Float        | Thời gian trung bình dành cho mỗi lần truy cập      |

# OVERVIEW DATASET

---

## HÀNH VI TƯƠNG TÁC

| Tên trường   | Kiểu dữ liệu | Mô Tả                                 |
|--------------|--------------|---------------------------------------|
| SocialShares | Float        | Số lần nội dung tiếp thị được chia sẻ |
| EmailOpens   | Float        | Số lần email tiếp thị được mở         |
| EmailClicks  | Integer      | Số lần nhấp vào liên kết trong email  |

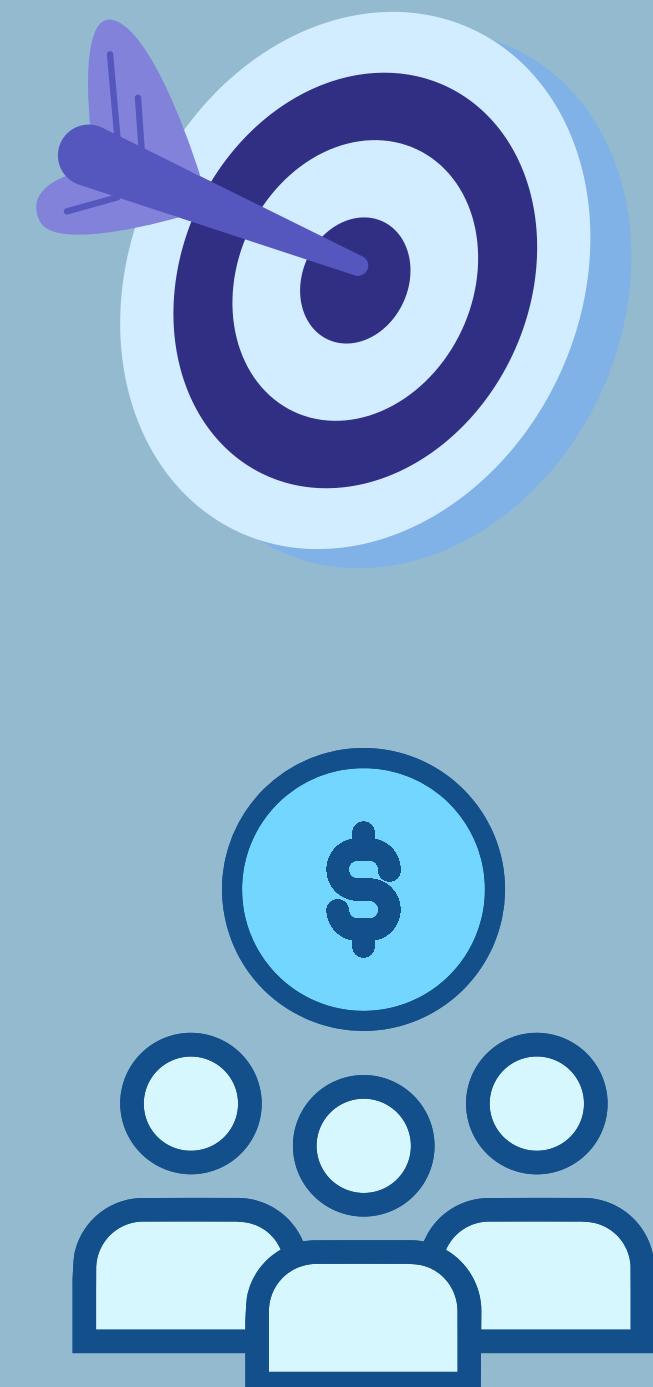
## BIẾN MỤC TIÊU

|            |     |                               |
|------------|-----|-------------------------------|
| Conversion | Bit | Chuyển đổi (1) hoặc không (0) |
|------------|-----|-------------------------------|

# MAIN OBJECTIVES

---

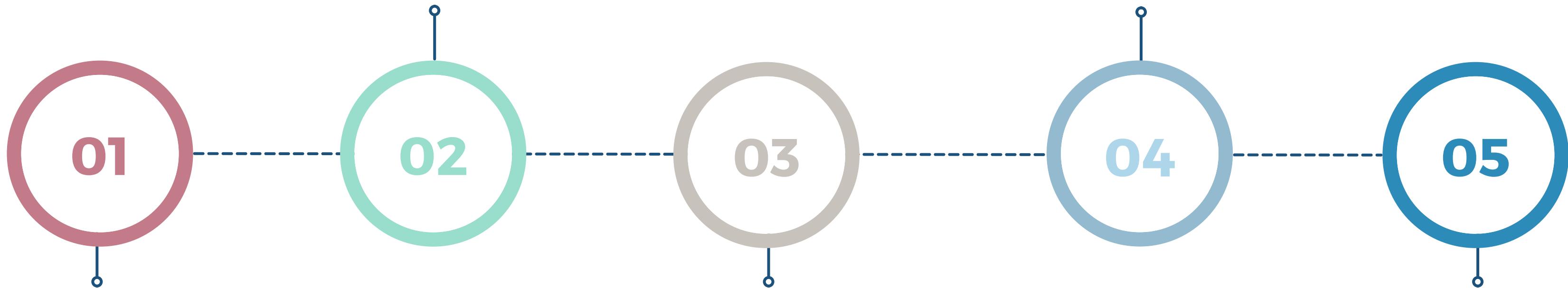
1. Xây dựng mô hình dự đoán tỷ lệ chuyển đổi
2. Phân tích yếu tố ảnh hưởng đến chuyển đổi
3. Xây dựng mô hình phân cụm khách hàng
4. Đánh giá hiệu quả các chiến dịch tiếp thị
5. Đề xuất chiến lược tiếp thị cho doanh nghiệp



- **Làm sạch** dữ liệu
- **Mã hóa** các biến phân loại

- **Đánh giá**, so sánh **hiệu quả** từng thuật toán
- **Cải thiện** (nếu có)

## FEATURE ENGINEERING



## EDA

- Thống kê mô tả
- Khám phá **mối quan hệ** giữa các biến

## MODELING & CLUSTERING

### 1. Modeling:

- Chuẩn bị **dữ liệu huấn luyện** và **kiểm tra**
- Huấn luyện mô hình (**Decision Tree**, **Random Forest**, **LightGBM**, **XGBoost**)

### 2. Clustering

- Chọn bộ đặc trưng
- Phân cụm với **K - Means** & **Hierachical**

## EVALUATION & IMPROVE

## CONCLUSION

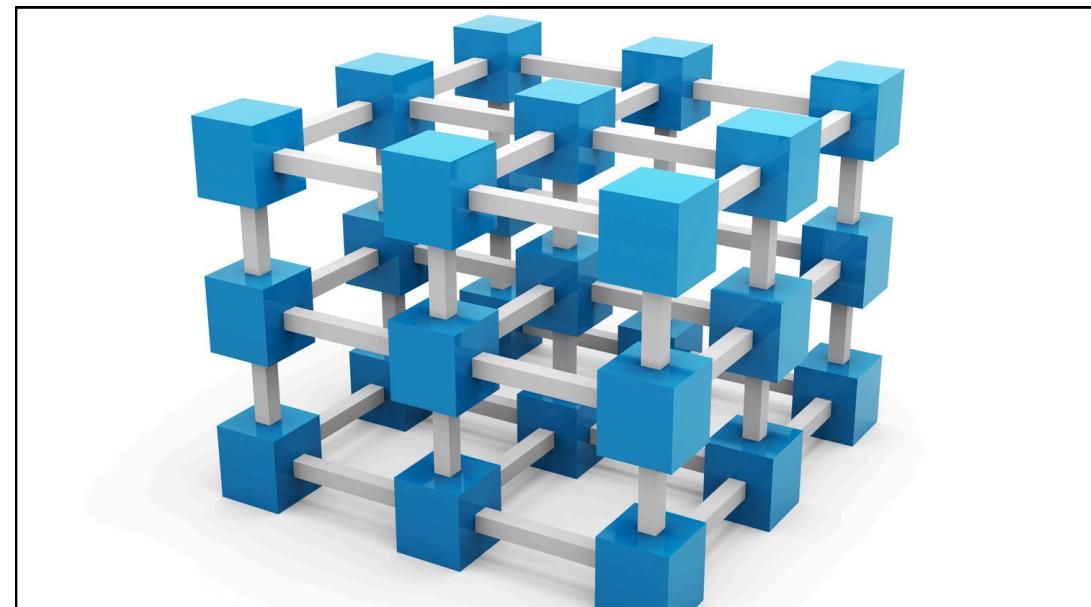
- Lựa chọn những mô hình tối ưu nhất
- Rút ra **insights**, đưa ra **khuyến nghị** cho doanh nghiệp

# EDA

## Exploratory Data Analysis



# GOAL



Hiểu cấu trúc dữ liệu



Tìm kiếm và phát hiện mối  
quan hệ giữa các biến



Đưa ra định hướng

# CHECK AND STATISTICS

## CHECK

01

Kích thước dữ liệu

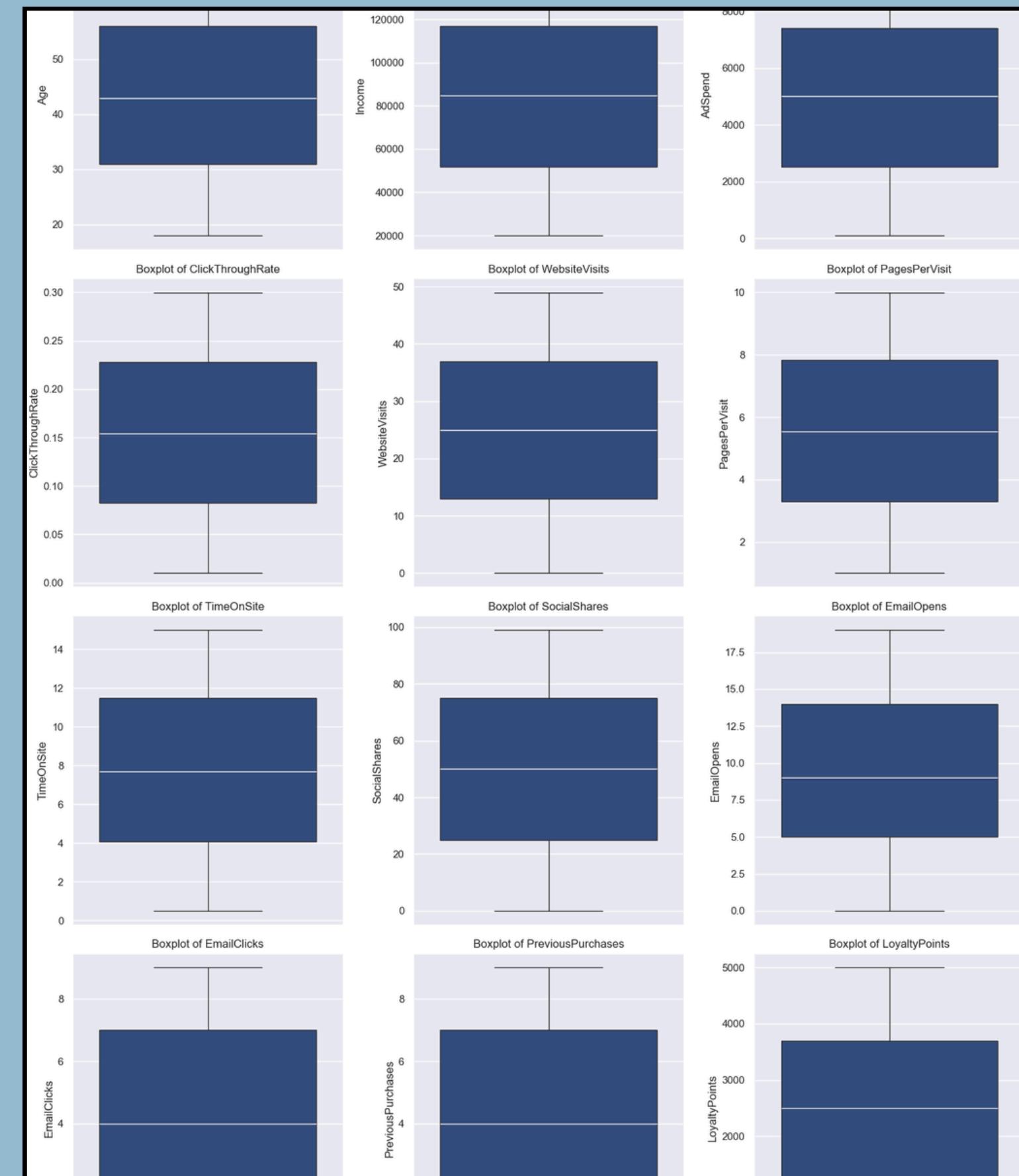
02

Số lượng và kiểu của từng biến

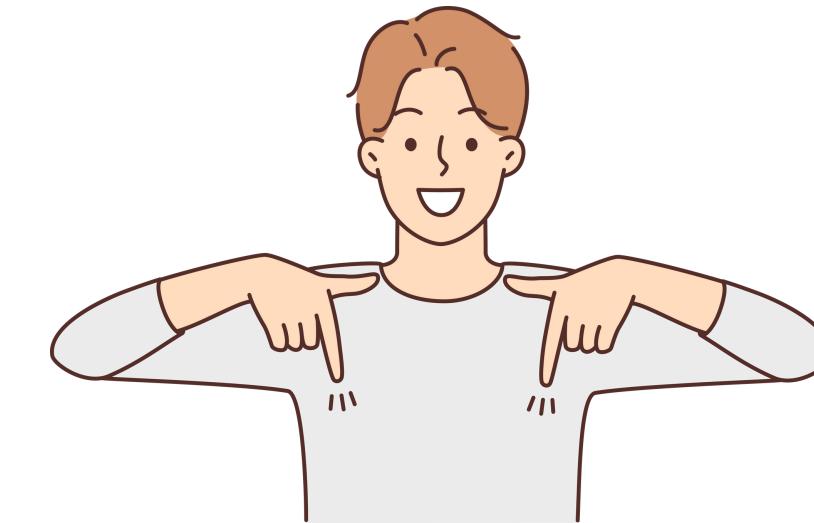
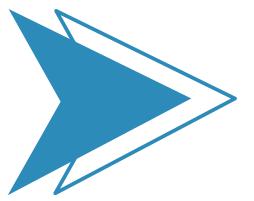
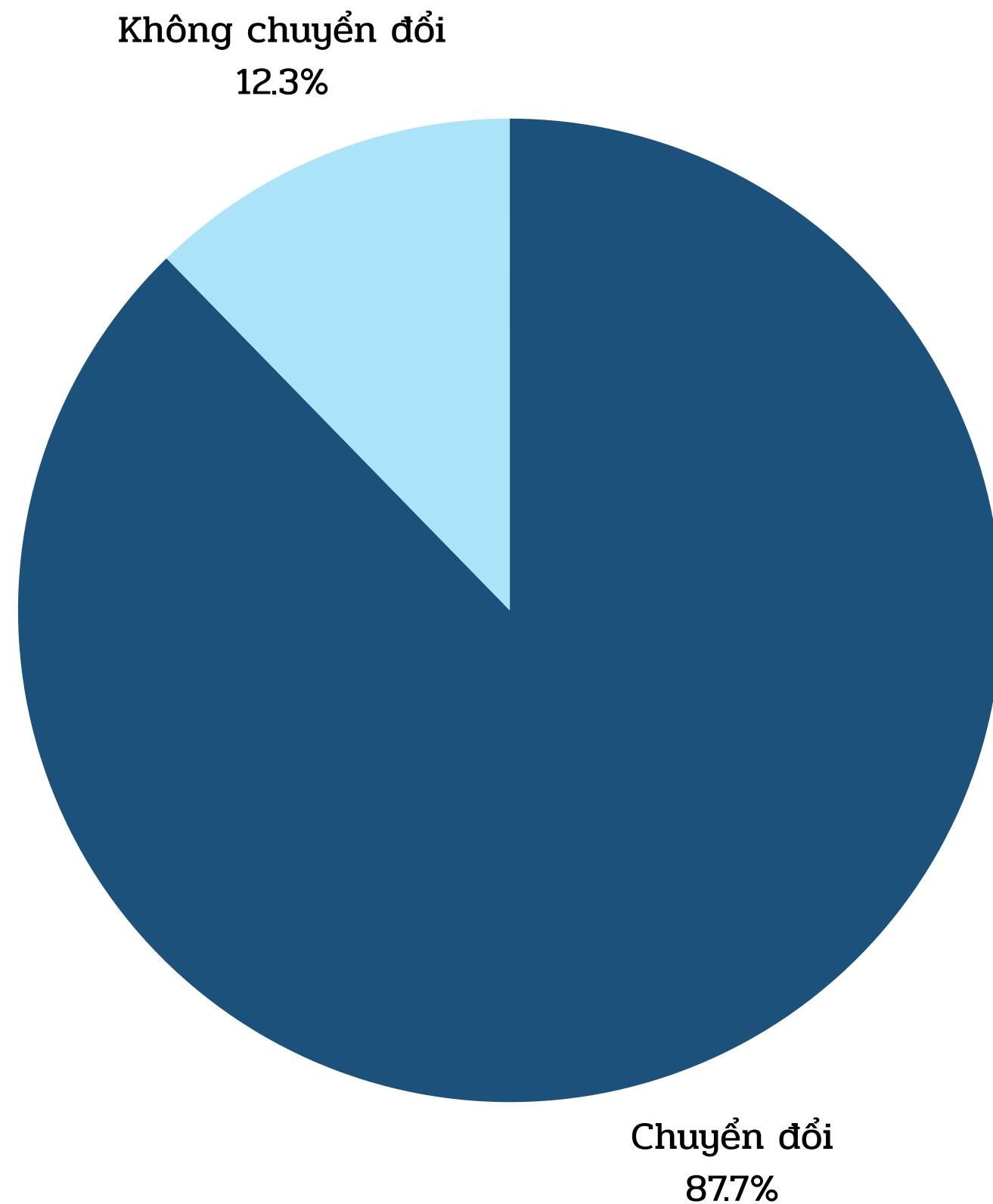
03

Các vấn đề của dữ liệu:

- duplicate
- missing
- outliers
- anomalies
- imbalanced
- incorrect data type
- inconsistent data



# CHECK AND STATISTICS



**Dấu hiệu mất cân bằng**

# CHECK AND STATISTICS

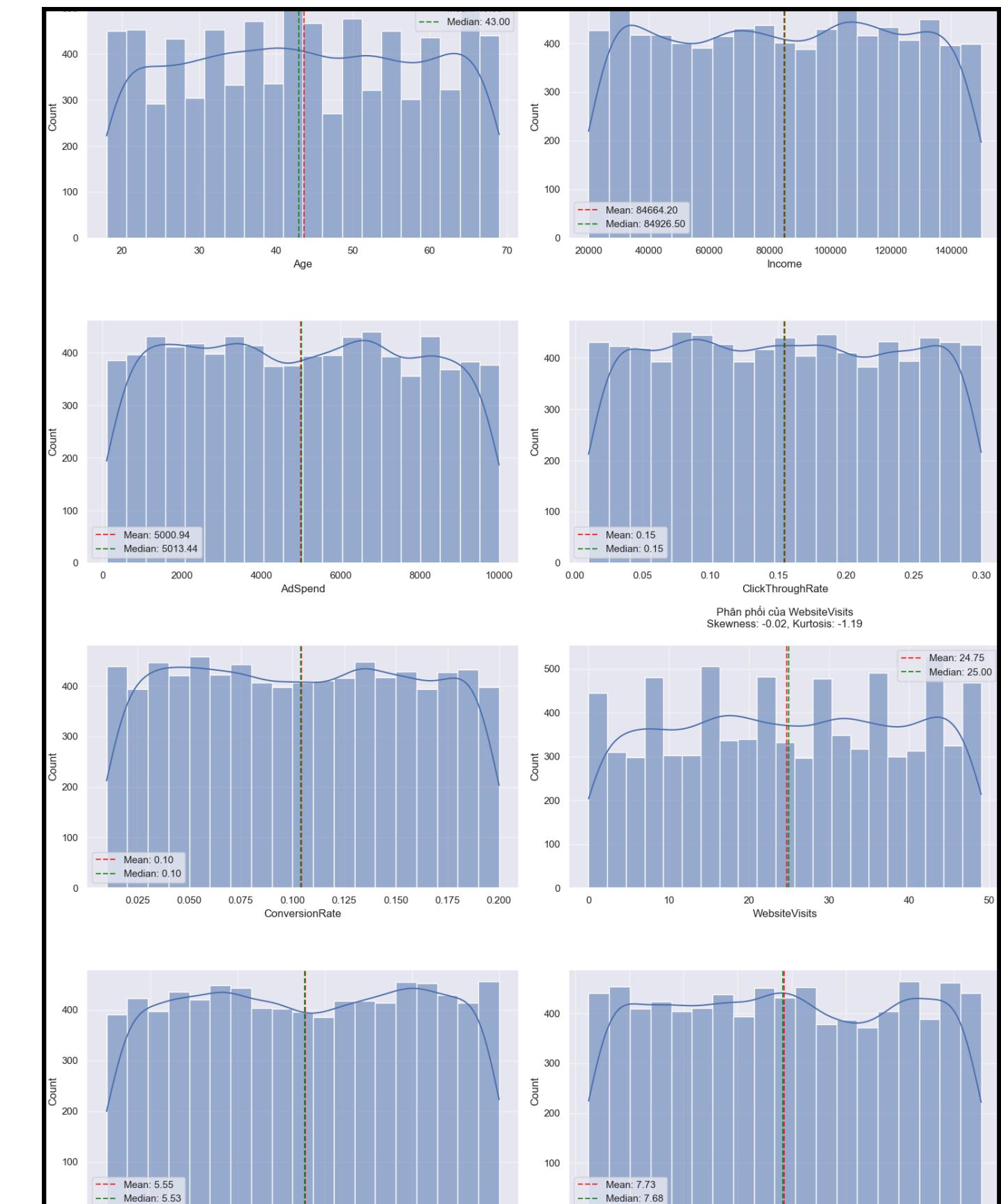
## STATISTICS

01

Các thống kê cơ bản: mean, min, max, phân vị

02

Các biến định lượng

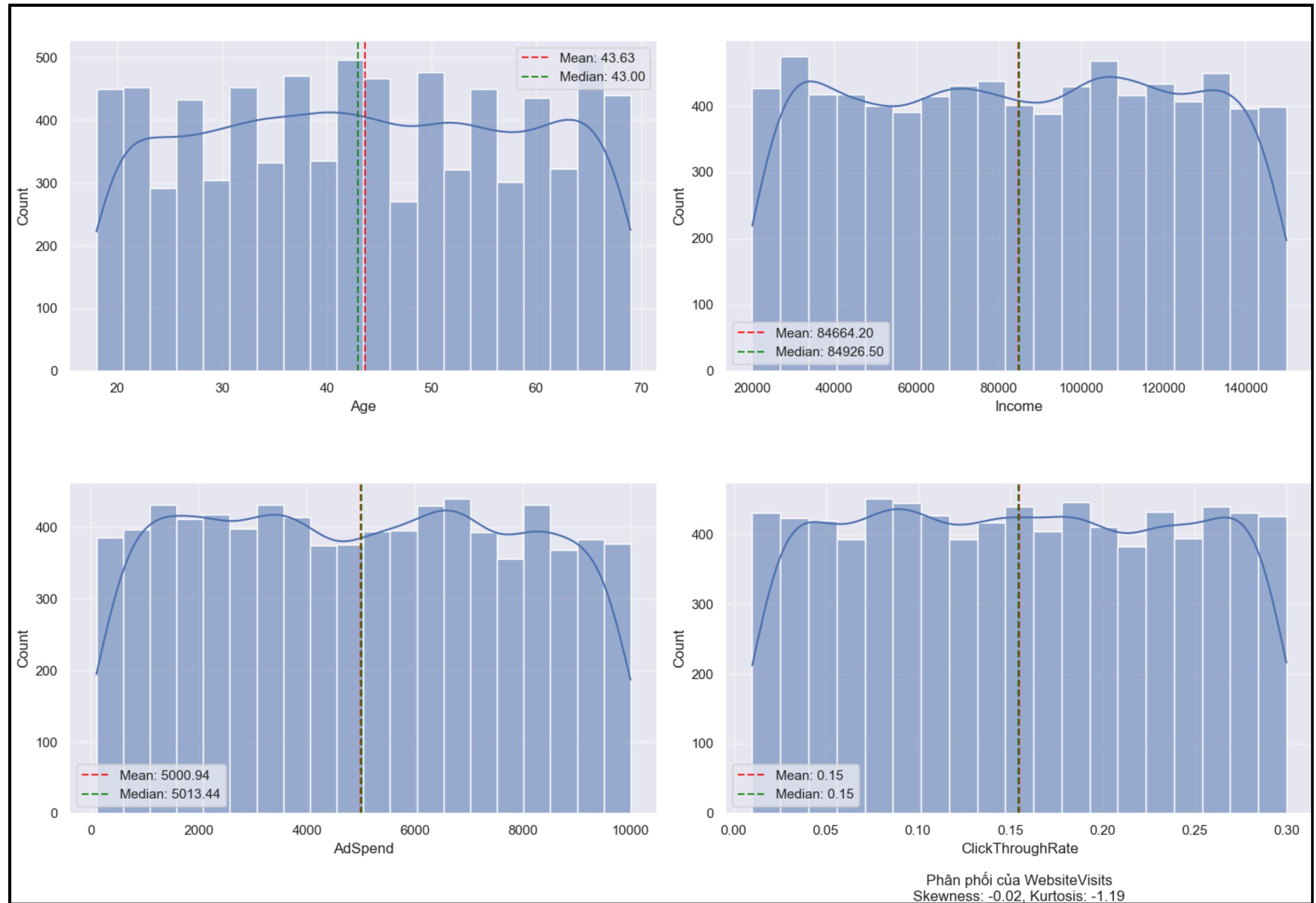


**Age**

**Income**

**AdSpend**

**ClickThroughRate**

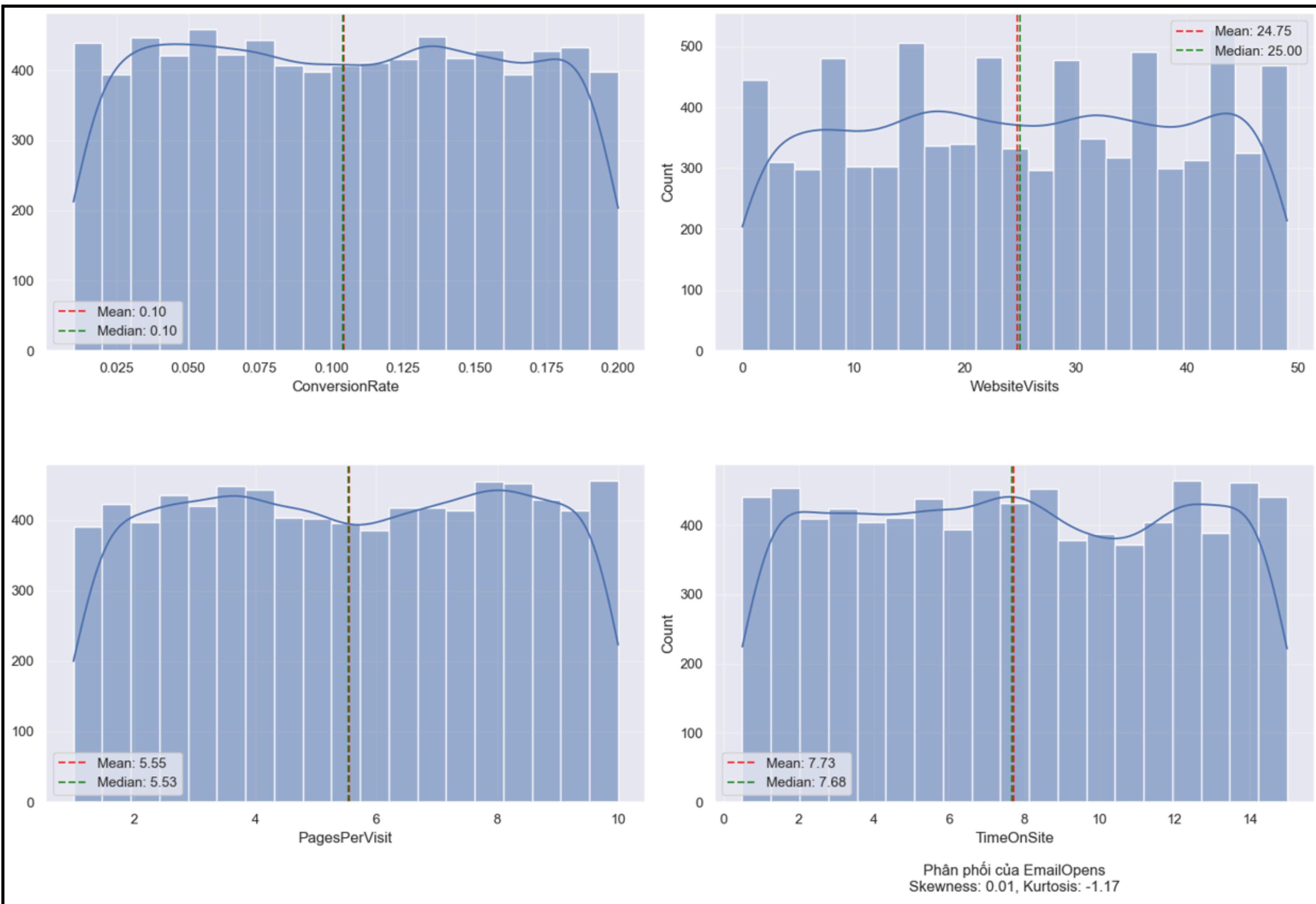


**ConversionRate**

**WebsiteVisits**

**PagesPerVisit**

**TimeOnSite**

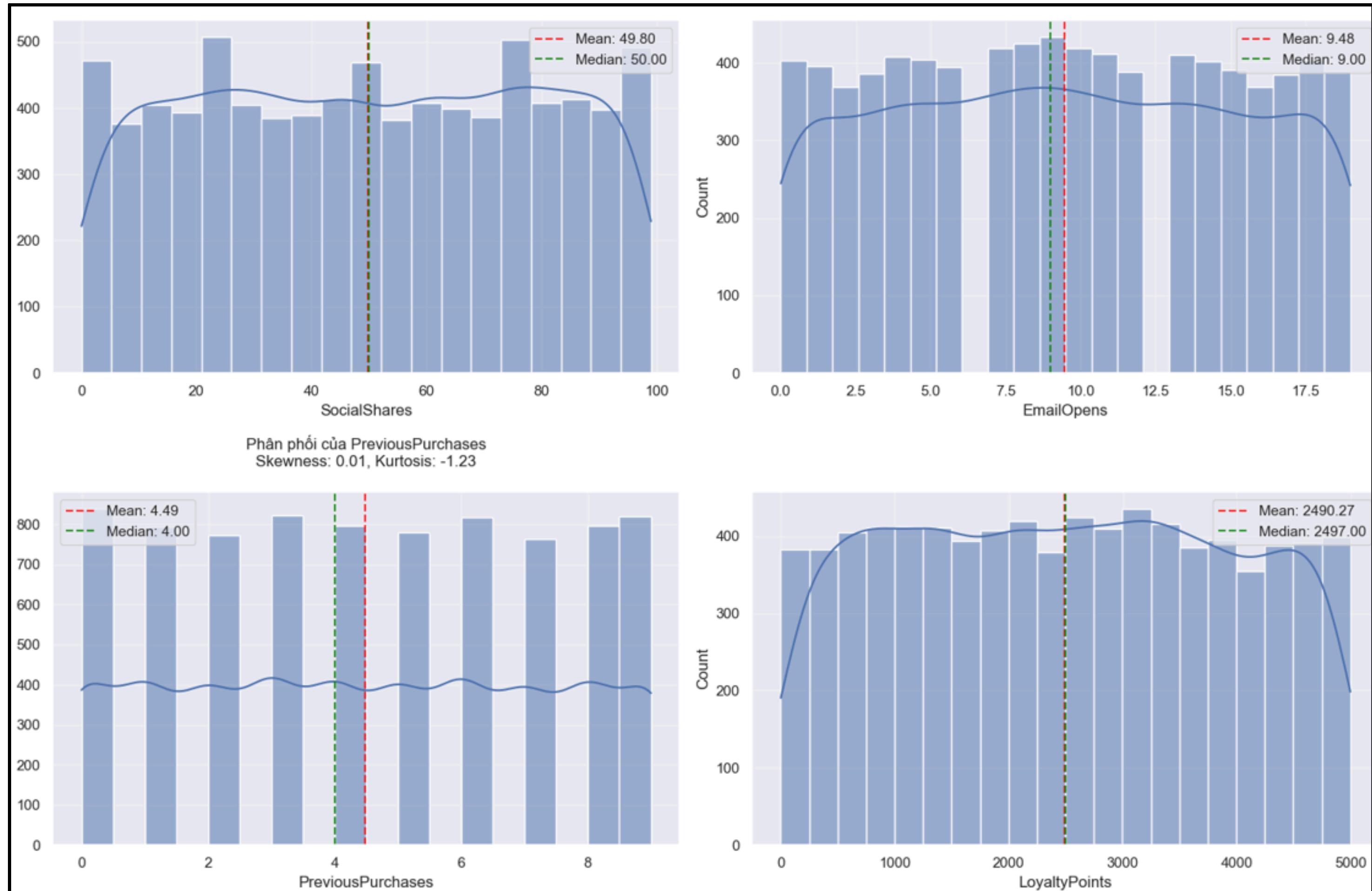


# SocialShares

# EmailOpens

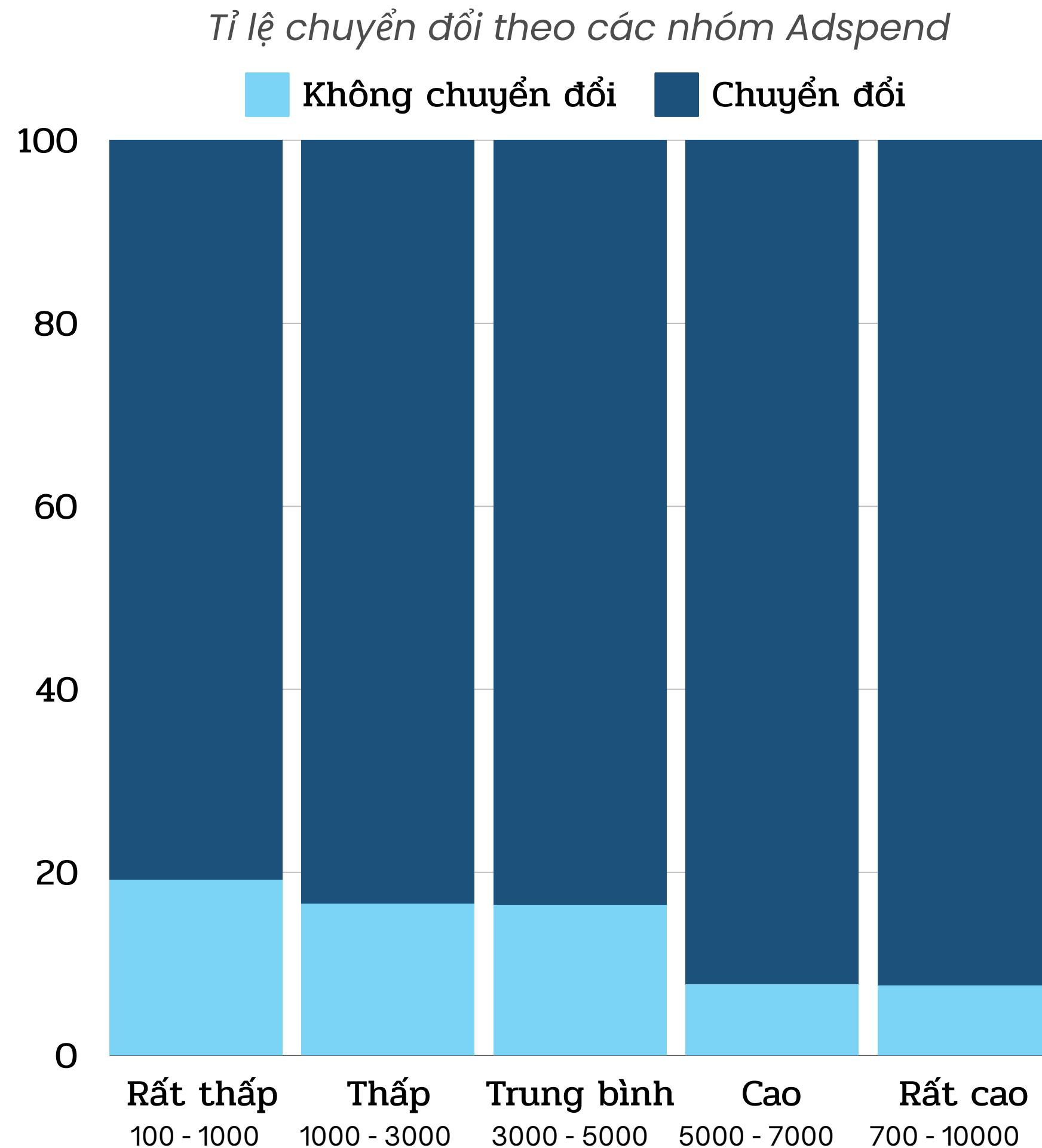
# PreviousPurchases

# LoyaltyPoints



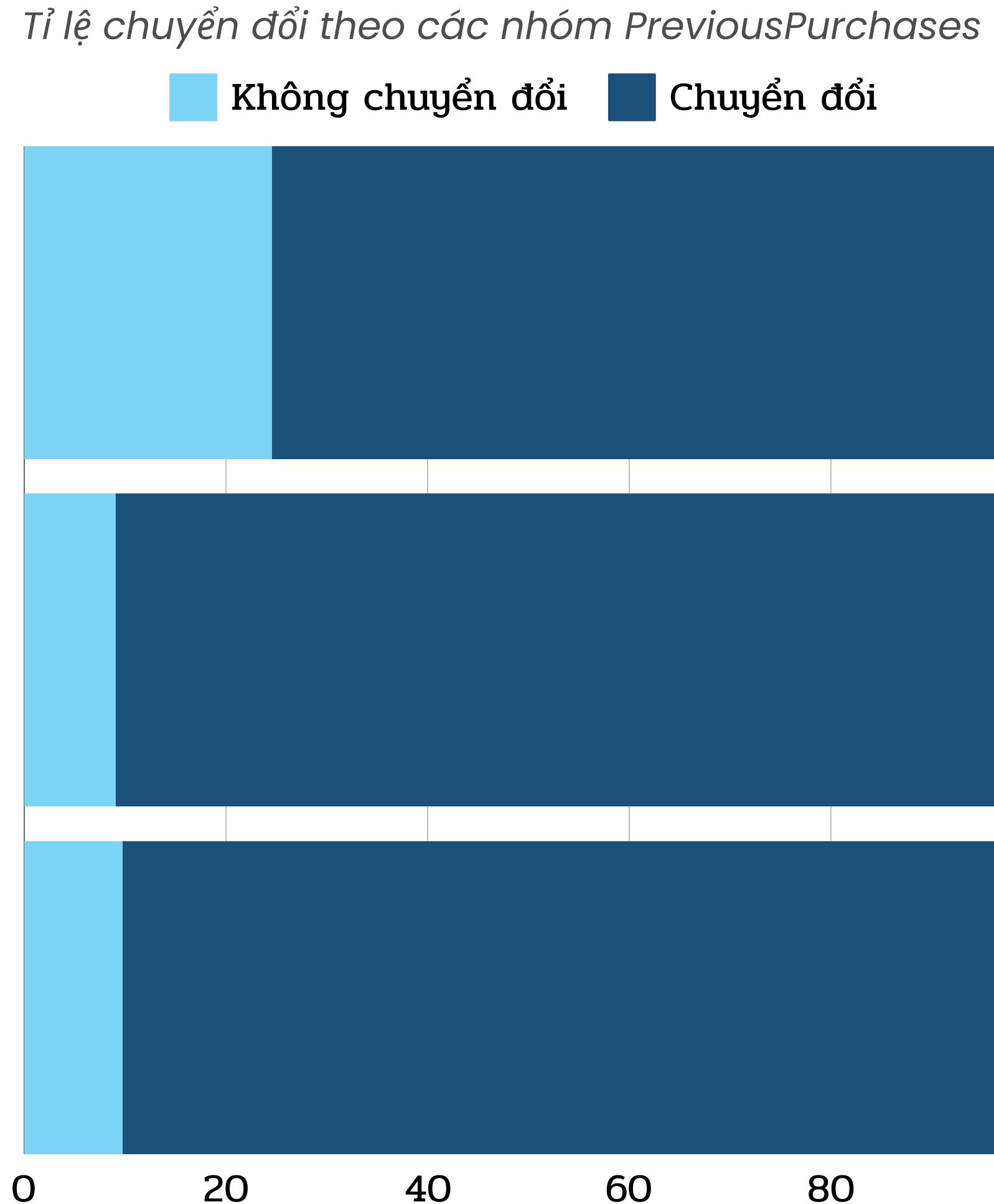
# RELATIONSHIP CATEGORICAL & TARGET

- ❖ Tăng dần theo từng nhóm chi phí
- ❖ Chi phí quảng bá càng cao tỉ lệ chuyển đổi càng lớn
- Adspend là một đặc trưng quan trọng



# RELATIONSHIP CATEGORICAL & TARGET

- ❖ Tăng dần theo từng nhóm mua hàng
  - ❖ Số lần mua càng nhiều, tỉ lệ chuyển đổi càng cao
- PreviousPurchases là một đặc trưng quan trọng**



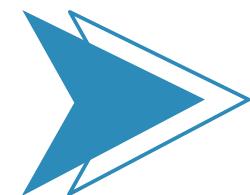
# RELATIONSHIP CATEGORICAL & TARGET



Tăng dần theo từng nhóm điểm



Tỉ lệ chuyển đổi tỉ lệ thuận với số  
điểm trung thành



LoyaltyPoints là một đặc trưng  
quan trọng

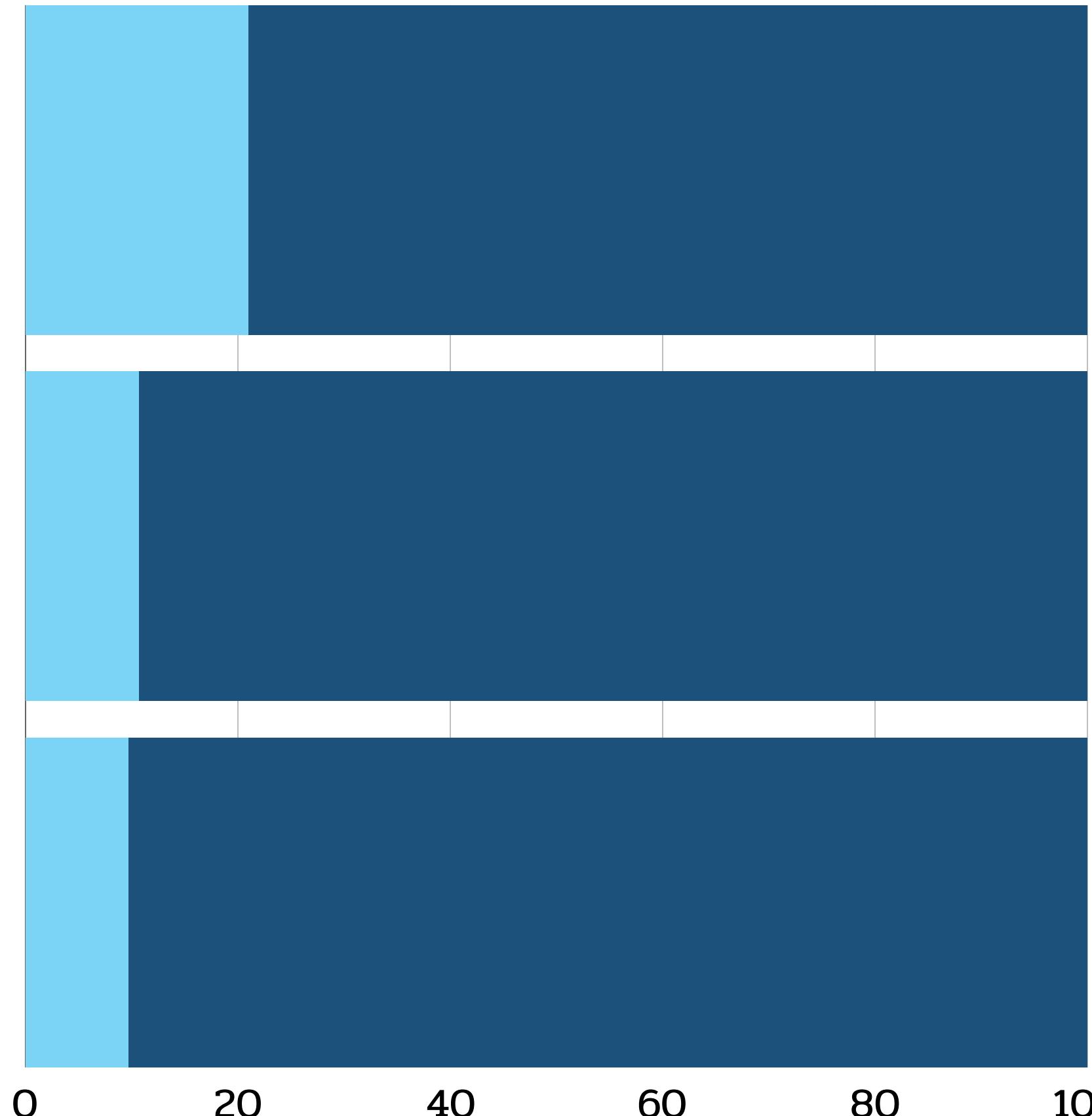
Tỉ lệ chuyển đổi theo các nhóm LoyaltyPoints



Không chuyển đổi

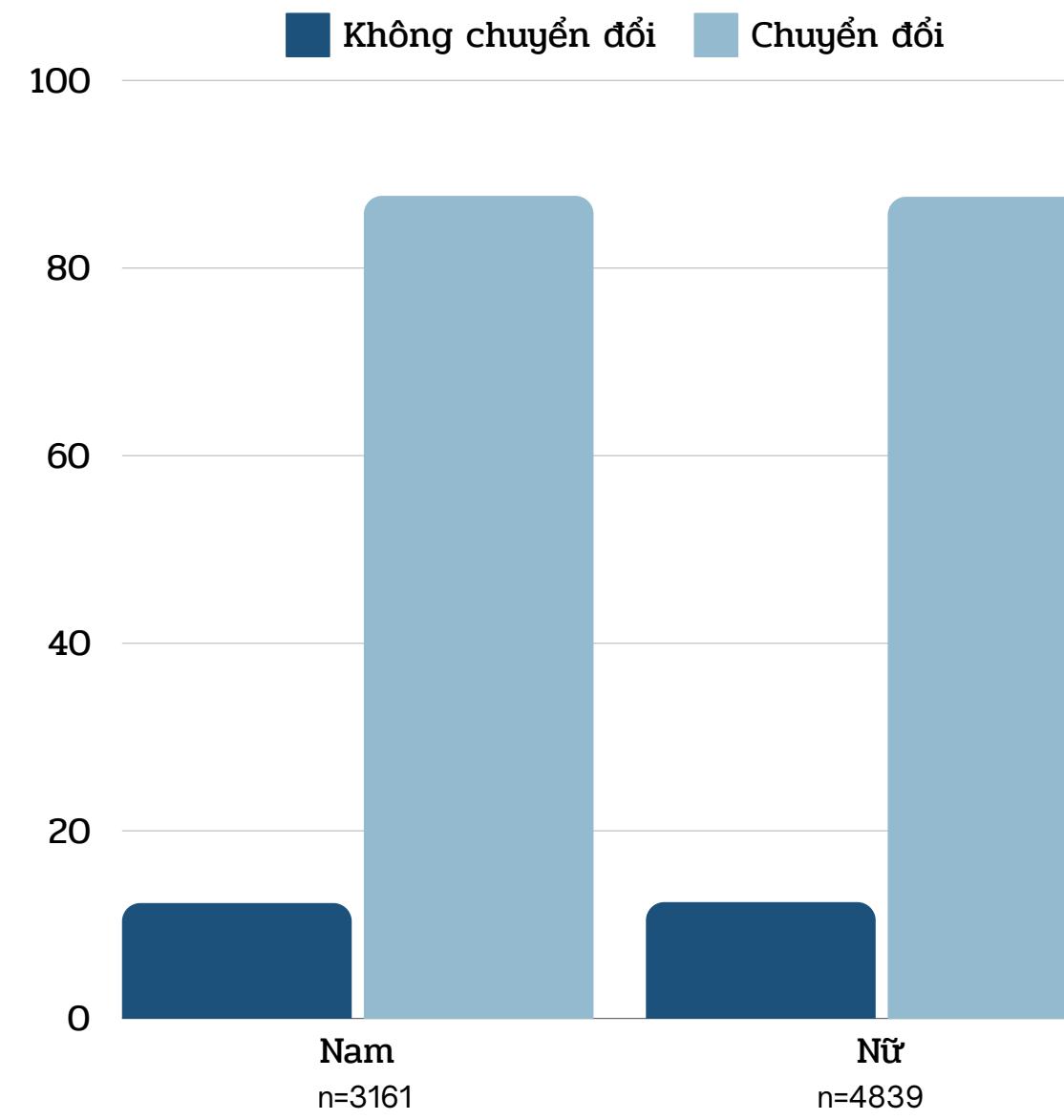


Chuyển đổi

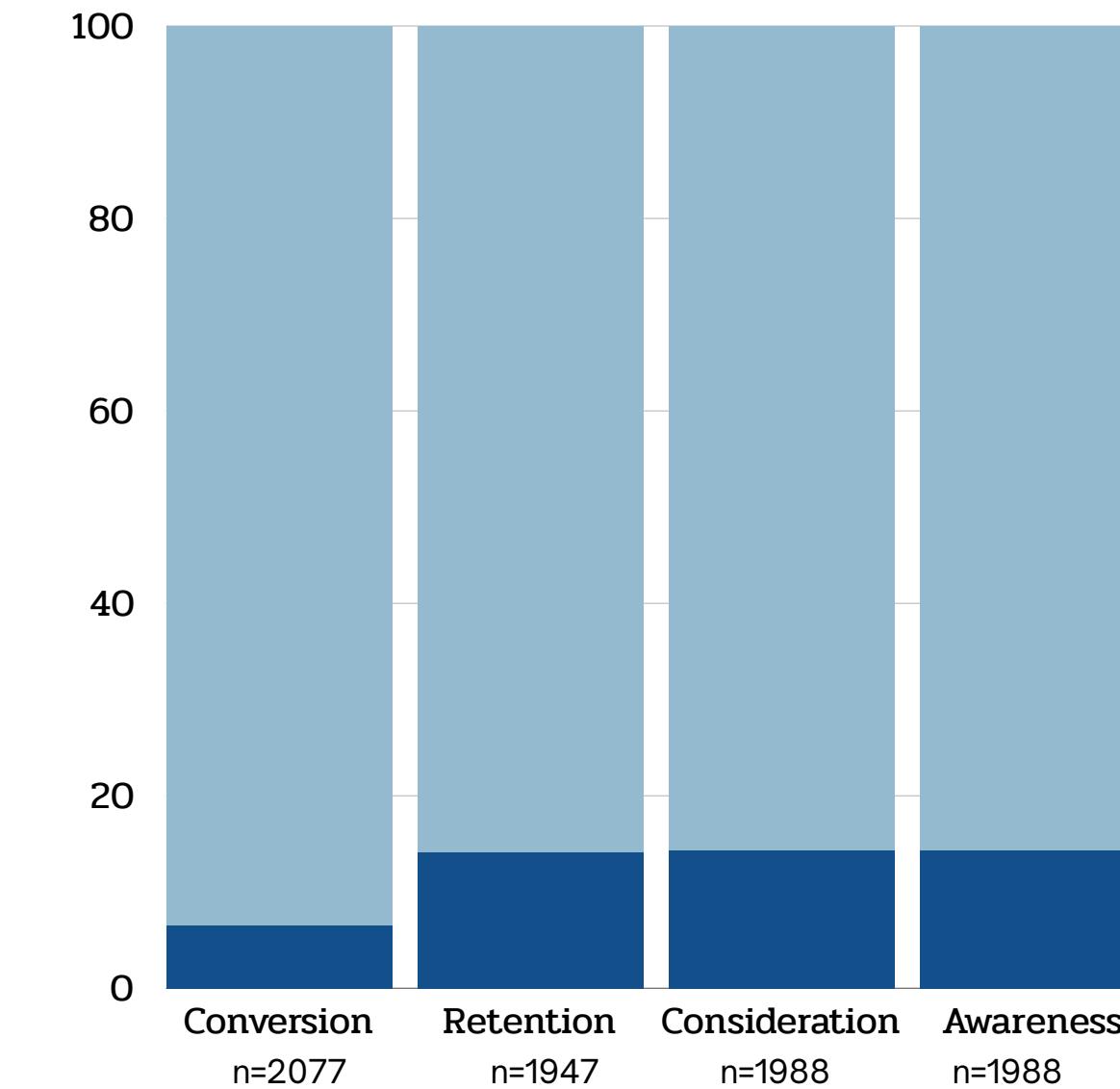


# RELATIONSHIP CATEGORICAL & TARGET

Tỉ lệ chuyển đổi theo giới tính



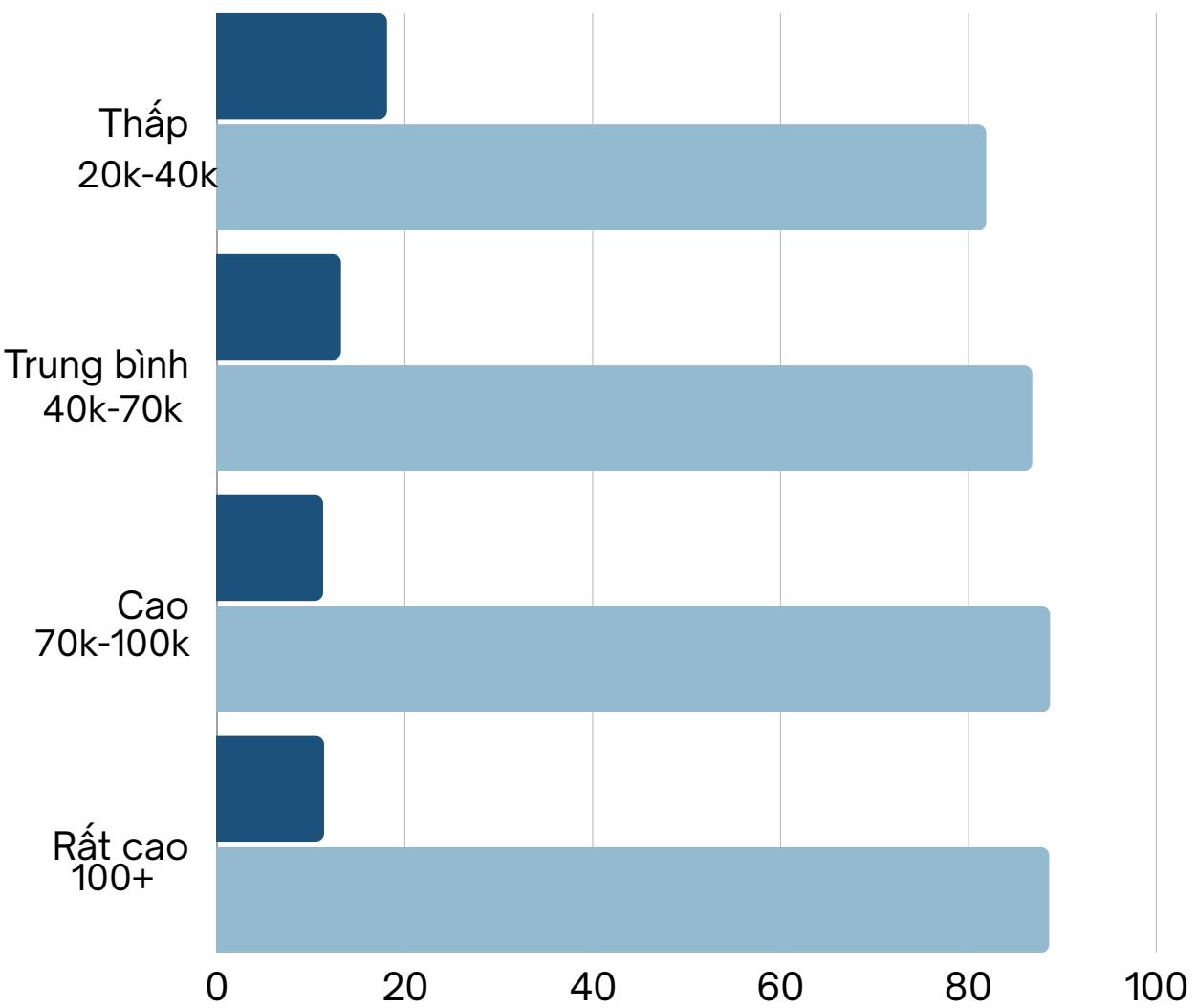
Tỉ lệ chuyển đổi theo loại chiến dịch



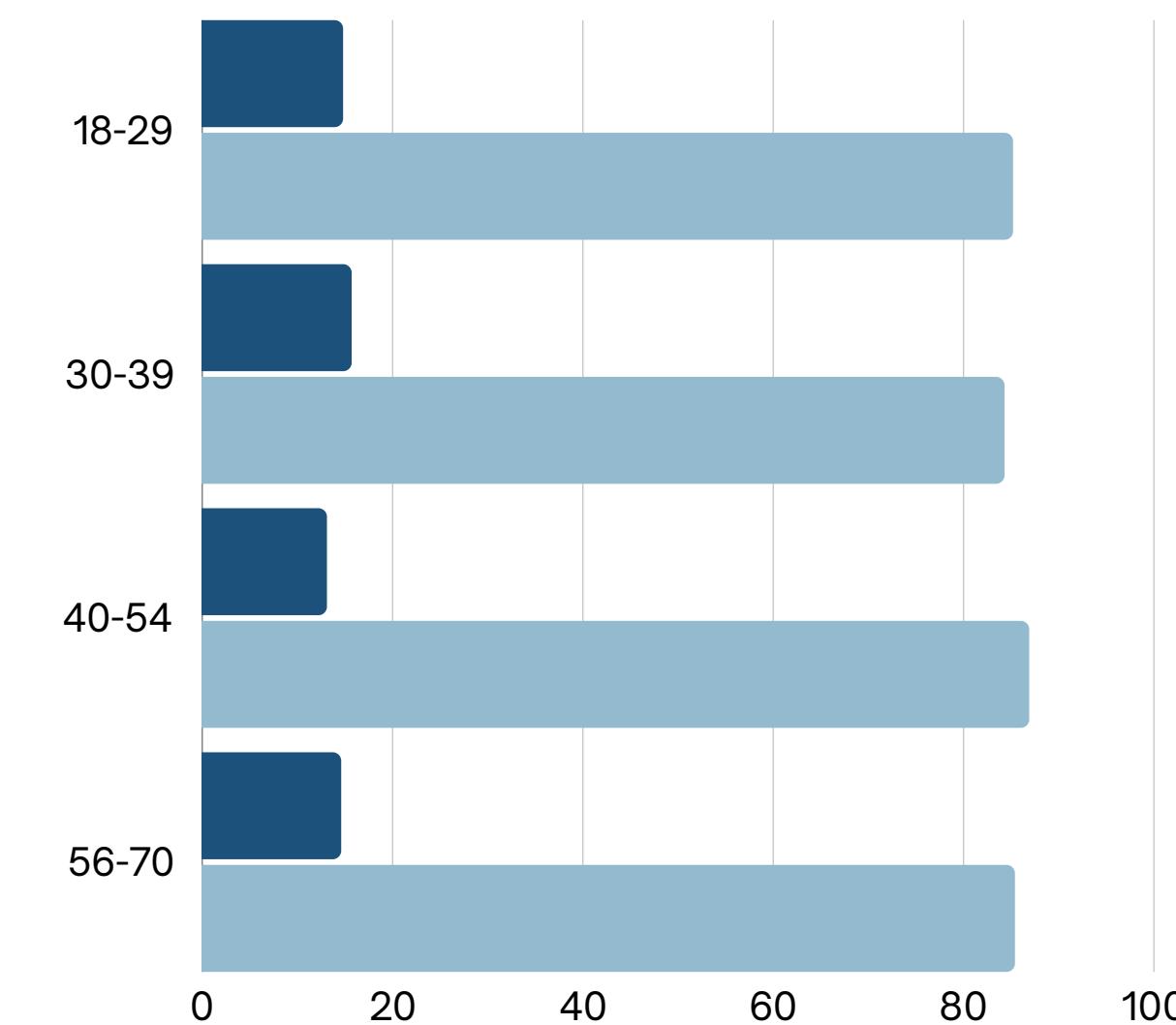
**Không có sự chênh lệch lớn về tỉ lệ chuyển đổi ở các biến phân loại**

# RELATIONSHIP CATEGORICAL & TARGET

Tỉ lệ chuyển đổi theo nhóm thu nhập cho kênh Email



Tỉ lệ chuyển đổi theo nhóm tuổi cho Awareness

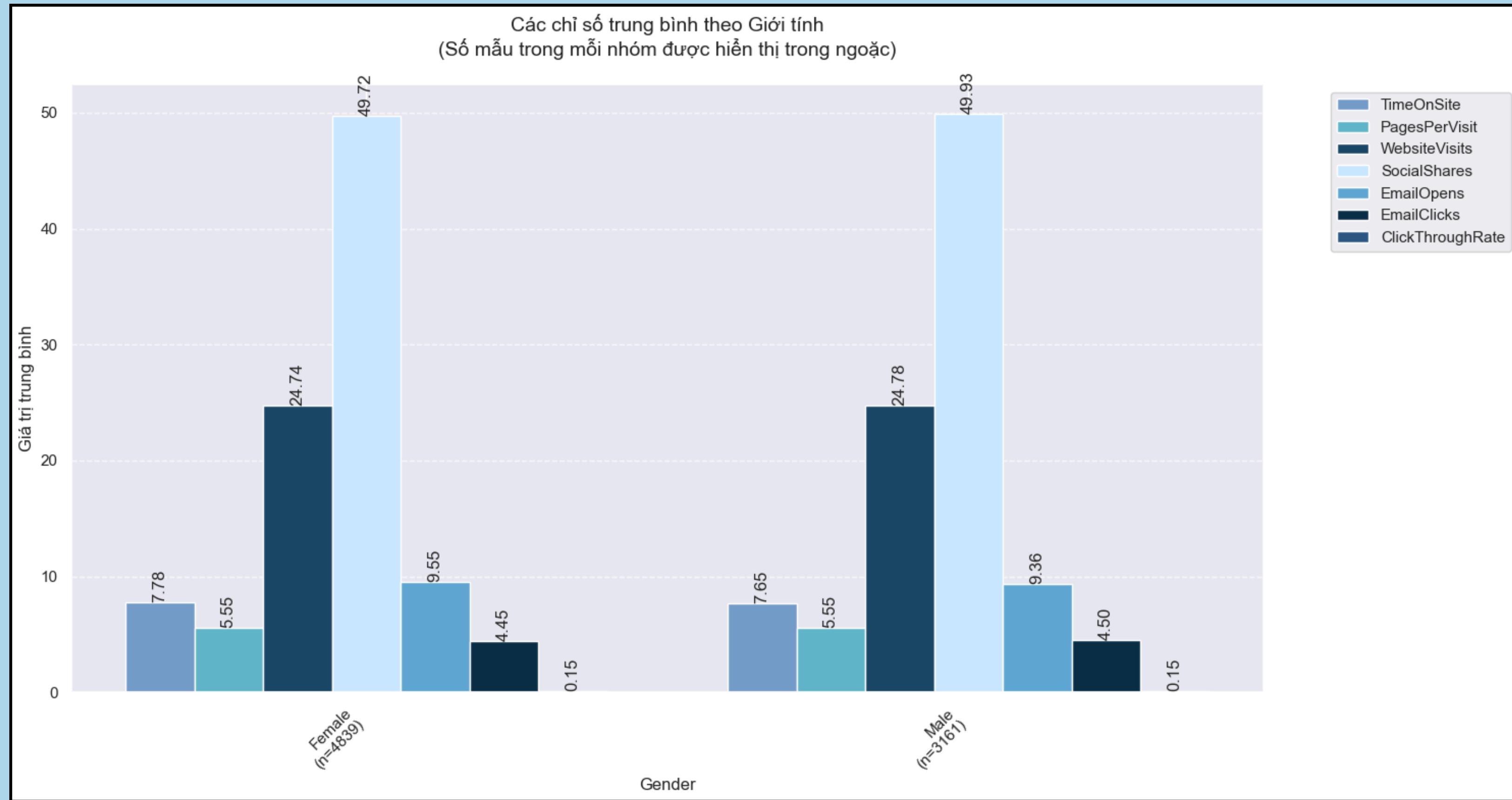


❖ Thậm chí khi kết hợp 2 biến phân loại với nhau cũng không tạo ra sự khác biệt

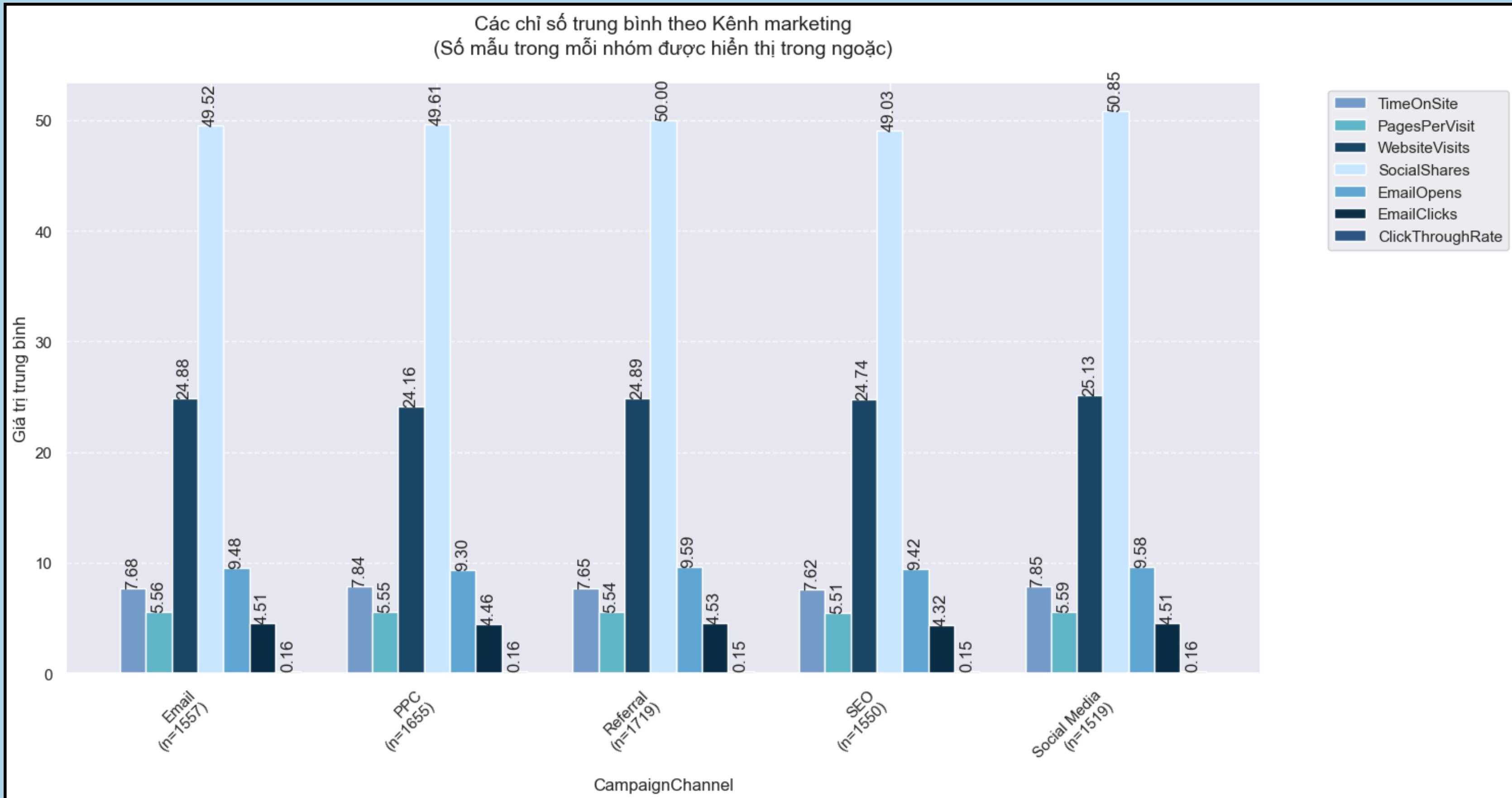
# RELATIONSHIP INTERACTIVE BEHAVIOR & CATEGORICAL



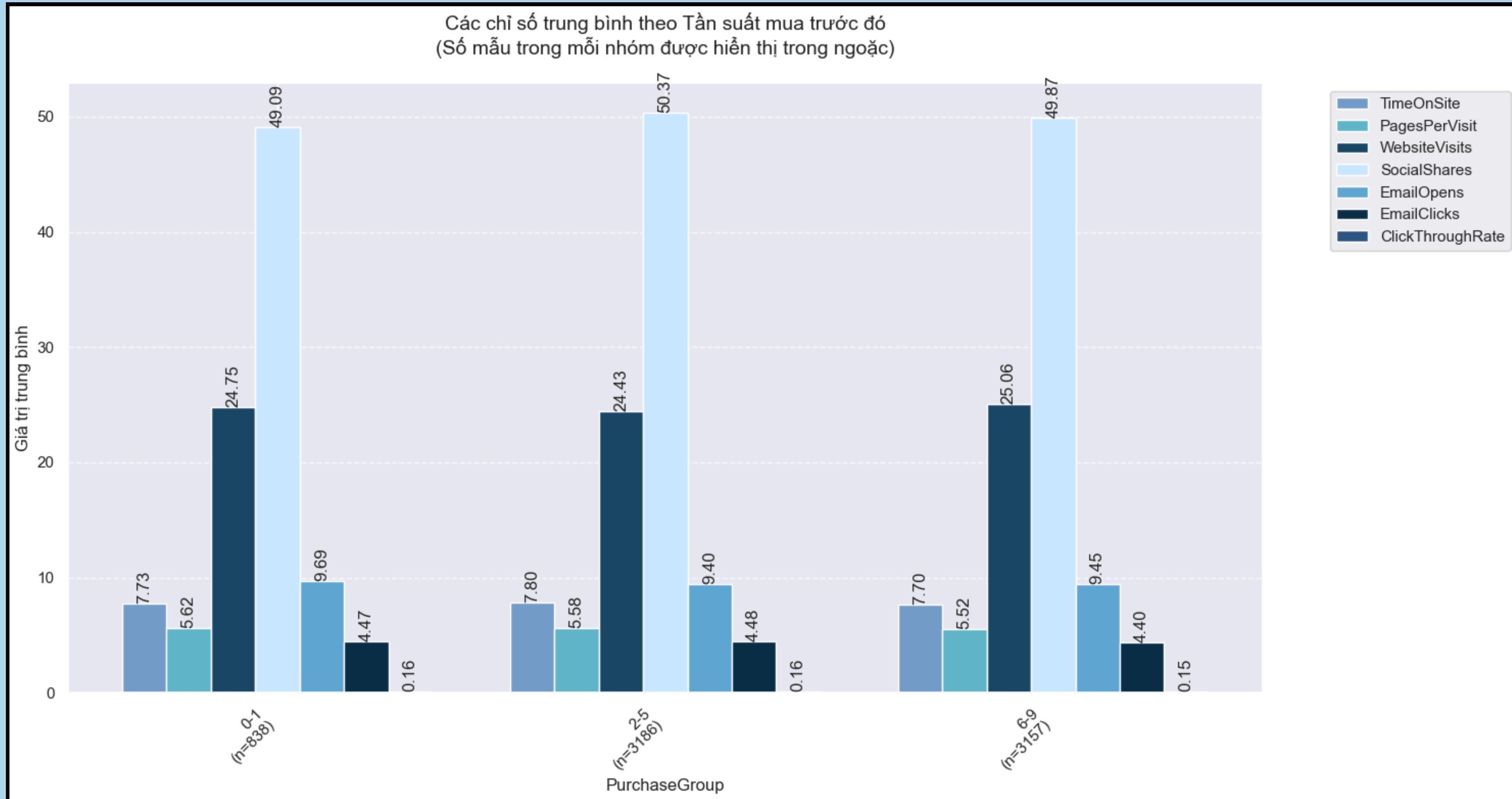
# INTERATIVE - GENDER



# INTERATIVE - CHANNEL

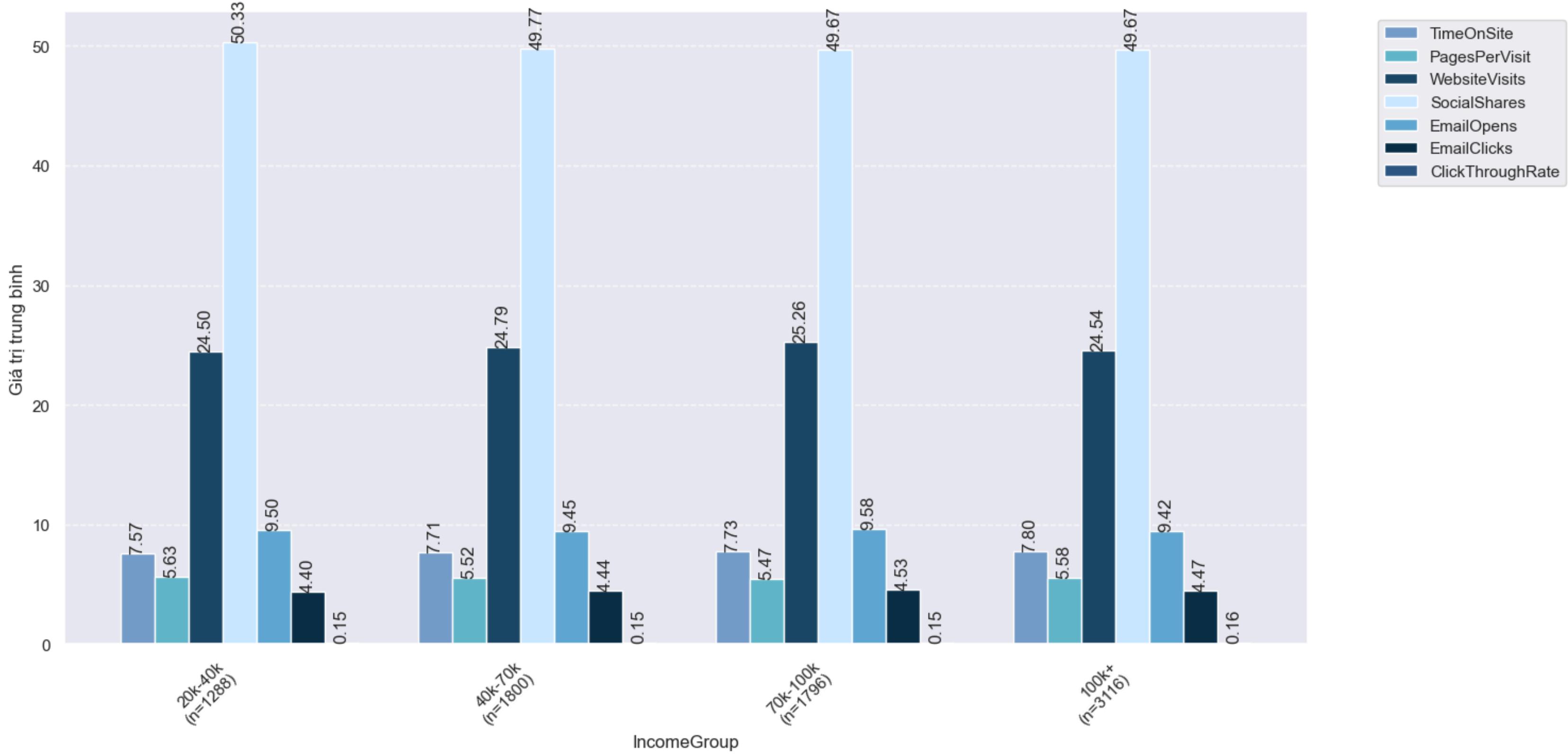


# INTERATIVE - PREVIOUSPURCHASES



# INTERATIVE - INCOME

Các chỉ số trung bình theo Nhóm thu nhập  
(Số mẫu trong mỗi nhóm được hiển thị trong ngoặc)

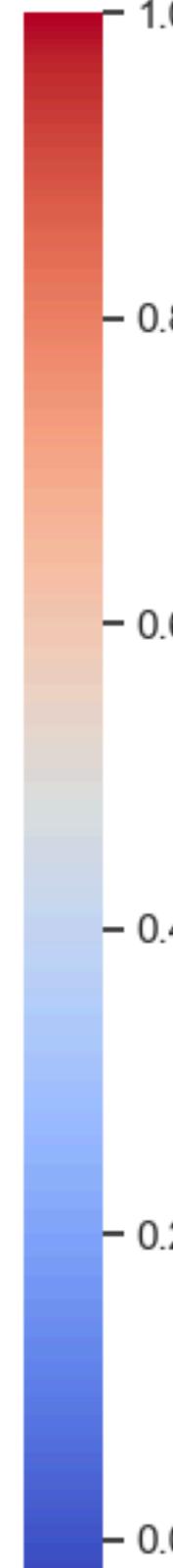


# CORRELATION MATRIX



# CORRELATION MATRIX

|                   |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
|-------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Age               | 1.00  | 0.01  | -0.00 | 0.01  | 0.02  | -0.00 | -0.01 | -0.01 | -0.01 | 0.02  | 0.01  | -0.01 | 0.01  | 0.01  | 0.00  |
| Income            | 0.01  | 1.00  | 0.00  | 0.01  | 0.02  | -0.00 | 0.00  | 0.02  | -0.01 | -0.00 | 0.01  | -0.01 | -0.01 | -0.01 | 0.01  |
| AdSpend           | -0.00 | 0.00  | 1.00  | -0.01 | -0.02 | 0.01  | -0.01 | -0.00 | -0.02 | 0.02  | 0.00  | 0.00  | 0.00  | 0.00  | 0.12  |
| ClickThroughRate  | 0.01  | 0.01  | -0.01 | 1.00  | -0.01 | -0.02 | 0.00  | -0.01 | -0.01 | -0.01 | -0.01 | -0.00 | -0.02 | -0.02 | 0.12  |
| ConversionRate    | 0.02  | 0.02  | -0.02 | -0.01 | 1.00  | -0.01 | 0.02  | 0.01  | 0.01  | 0.01  | 0.01  | -0.02 | -0.00 | 0.09  |       |
| WebsiteVisits     | -0.00 | -0.00 | 0.01  | -0.02 | -0.01 | 1.00  | -0.01 | -0.02 | 0.00  | 0.01  | 0.00  | 0.01  | 0.00  | 0.00  | 0.08  |
| PagesPerVisit     | -0.01 | 0.00  | -0.01 | 0.00  | 0.02  | -0.01 | 1.00  | 0.02  | 0.01  | 0.00  | -0.00 | -0.02 | -0.01 | -0.01 | 0.10  |
| TimeOnSite        | -0.01 | 0.02  | -0.00 | -0.01 | 0.01  | -0.02 | 0.02  | 1.00  | 0.00  | -0.00 | 0.00  | -0.01 | -0.01 | -0.01 | 0.13  |
| SocialShares      | -0.01 | -0.01 | -0.02 | -0.01 | 0.01  | 0.00  | 0.01  | 0.00  | 1.00  | -0.01 | 0.00  | -0.01 | -0.00 | -0.00 | -0.01 |
| EmailOpens        | 0.02  | -0.00 | 0.02  | -0.01 | 0.01  | 0.01  | 0.00  | -0.00 | -0.01 | 1.00  | 0.00  | 0.00  | -0.00 | -0.00 | 0.12  |
| EmailClicks       | 0.01  | 0.01  | 0.00  | -0.01 | 0.01  | 0.00  | -0.00 | 0.00  | 0.00  | 0.00  | 1.00  | 0.00  | -0.00 | -0.00 | 0.13  |
| PreviousPurchases | -0.01 | -0.01 | 0.00  | -0.00 | -0.02 | 0.01  | -0.02 | -0.01 | -0.01 | 0.00  | 0.00  | 1.00  | 0.01  | 0.01  | 0.11  |
| LoyaltyPoints     | 0.01  | -0.01 | 0.00  | -0.02 | -0.00 | 0.00  | -0.01 | -0.01 | -0.00 | -0.00 | 0.01  | 1.00  | 0.01  | 0.10  |       |
| Conversion        | 0.00  | 0.01  | 0.12  | 0.12  | 0.09  | 0.08  | 0.10  | 0.13  | -0.01 | 0.12  | 0.13  | 0.11  | 0.10  | 0.10  | 1.00  |



Mức độ tương quan giữa các đặc trưng và biến mục tiêu yếu (<0.2)



# CONCLUSION

- ❖ Dataset không có ngoại lai (Outliers) hay dị thường (Anomalies).
- ❖ Dataset có dấu hiệu mất cân bằng ở các lớp với tỉ lệ 1:7
- ❖ Các features đều sẽ có ảnh hưởng nhất định đến biến mục tiêu
- ❖ Đảm bảo không có hiện tượng đa cộng tuyến
- ❖ Sử dụng các mô hình phi tuyến như Random Forest, Decision Trees, XGBoost, LightGBM để xây dựng model.

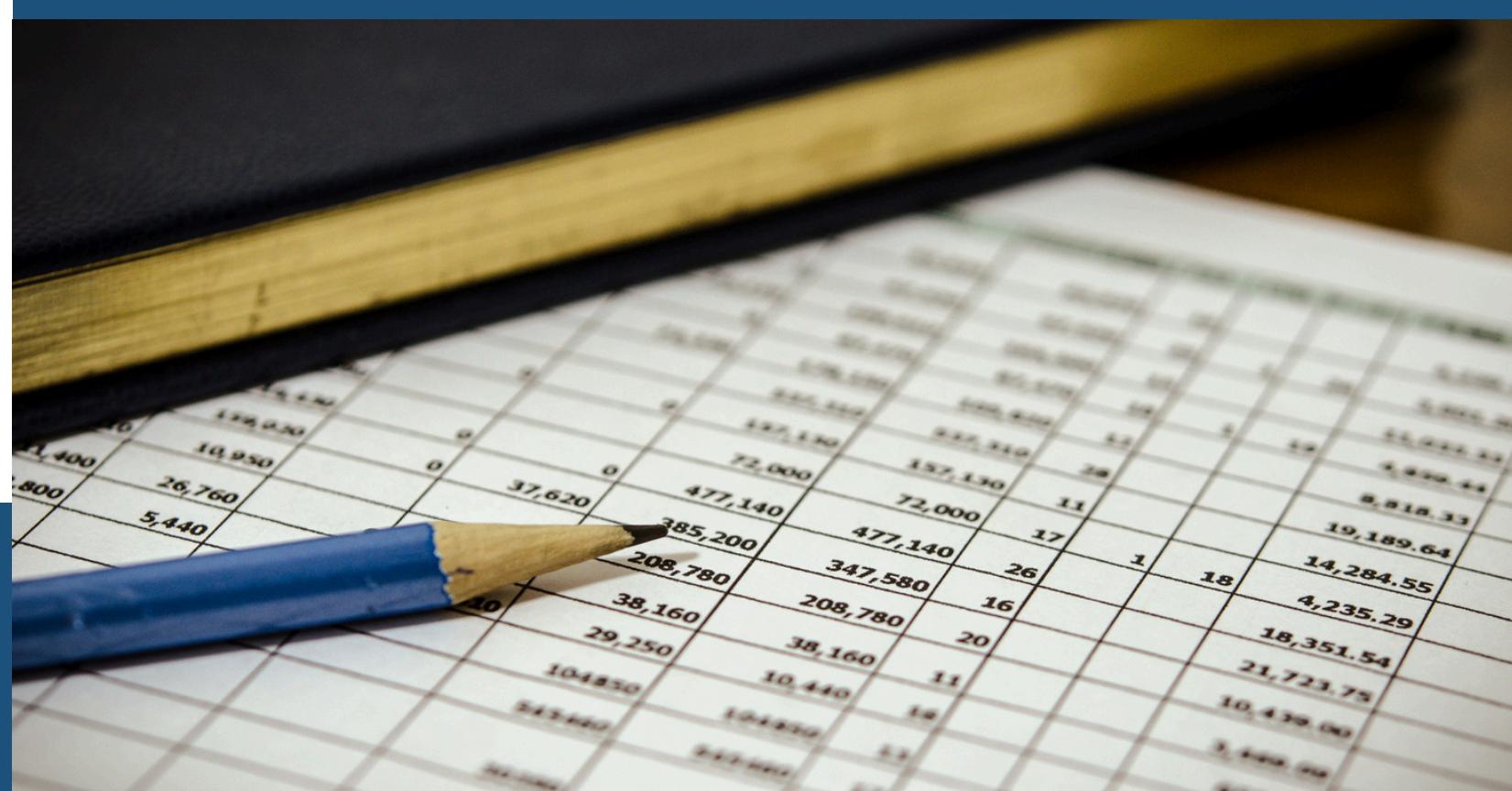
# DATA PREPARATION

---



# DROP COLUMNS

---



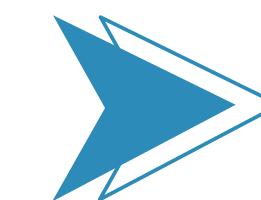
CustomerID



AdvertisingPlatform



AdvertisingTool



Bộ dữ liệu còn lại 17 cột

# FEATURE ENCODING



## Gender

```
df['Gender'] = df['Gender'].replace(  
    |   |   |   {'Male': 1, 'Female': 0}).astype(int)
```



## CampaignType

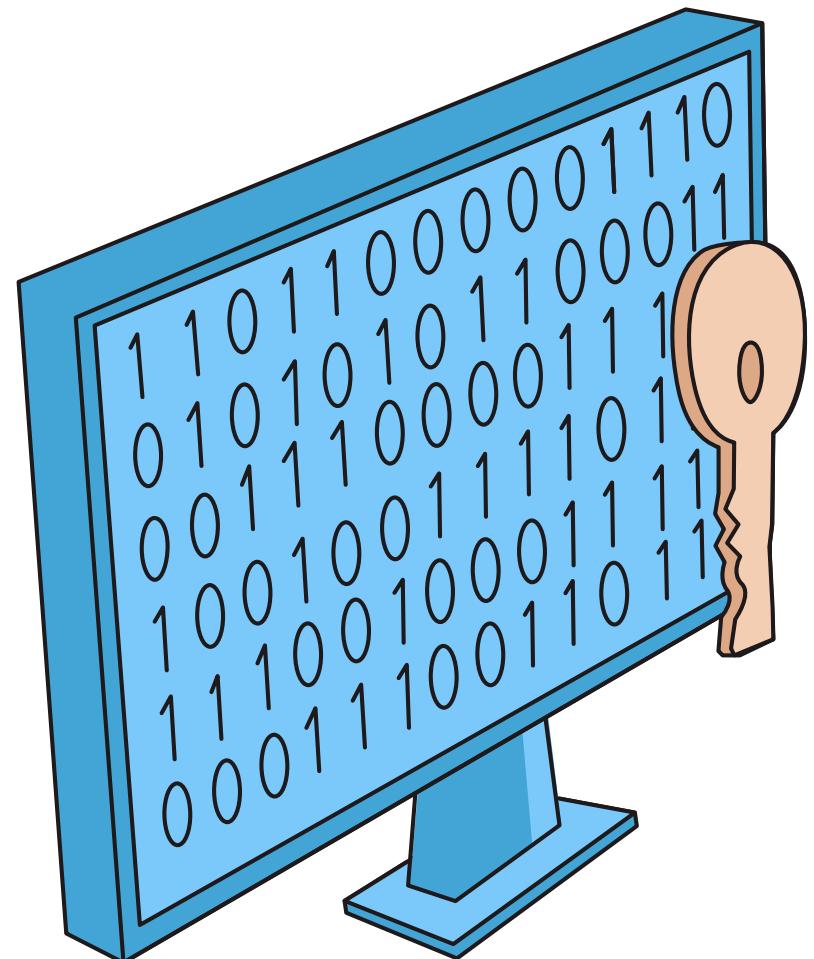
```
from sklearn.preprocessing import OneHotEncoder
```



## CampaignChannel



# FEATURE ENCODING



## Gender

- Male --> 1
- Female --> 0



## CampaignType

- Retention --> 1
- Conversion --> 10
- Consideration --> 100
- Awareness --> 1000



## CampaignChannel

- Social Media --> 1
- SEO --> 10
- Referral --> 100
- PPC --> 1000
- Email --> 10000

# BUILDING CLASSIFICATION MODELS



# DATA PREPARATION

- ❖ Xác định biến đầu vào và biến mục tiêu

```
X=df.drop('Conversion', axis=1)  
y= df['Conversion']
```

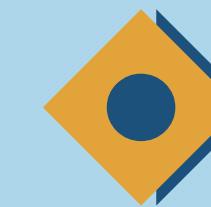


- ❖ Chuẩn bị dữ liệu train và test

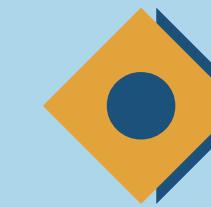
```
from sklearn.model_selection import StratifiedKFold  
kf = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
```

# TRAIN MODELS

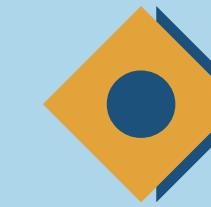
---



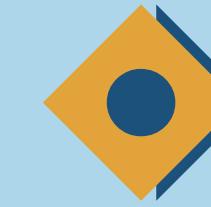
**Decision Tree**



**Random Forest**



**XGBoost**



**LightGBM**

# EVALUATION

| Decision Tree | Train | Test |
|---------------|-------|------|
| Accuracy      | 1.00  | 0.83 |
| Precision     | 1.00  | 0.91 |
| Recall        | 1.00  | 0.90 |
| F1-score      | 1.00  | 0.90 |
| AUC           | 1.00  | 0.63 |

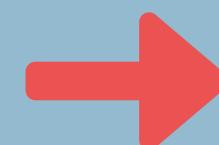
| Random Forest | Train | Test |
|---------------|-------|------|
| Accuracy      | 1.00  | 0.90 |
| Precision     | 1.00  | 0.90 |
| Recall        | 1.00  | 1.00 |
| F1-score      | 1.00  | 0.94 |
| AUC           | 1.00  | 0.82 |

# EVALUATION

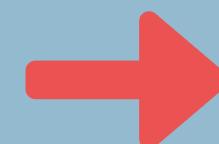
| XGBoost   | Train | Test |
|-----------|-------|------|
| Accuracy  | 1.00  | 0.92 |
| Precision | 1.00  | 0.92 |
| Recall    | 1.00  | 0.99 |
| F1-score  | 1.00  | 0.95 |
| AUC       | 1.00  | 0.81 |

| LightGBM  | Train | Test |
|-----------|-------|------|
| Accuracy  | 0.97  | 0.92 |
| Precision | 0.97  | 0.92 |
| Recall    | 1.00  | 0.99 |
| F1-score  | 0.98  | 0.96 |
| AUC       | 1.00  | 0.82 |

# EVALUATION



| Decision Tree | Precision | Recall | F1-score |
|---------------|-----------|--------|----------|
| Lớp 0         | 0.33      | 0.36   | 0.34     |
| Lớp 1         | 0.91      | 0.90   | 0.90     |



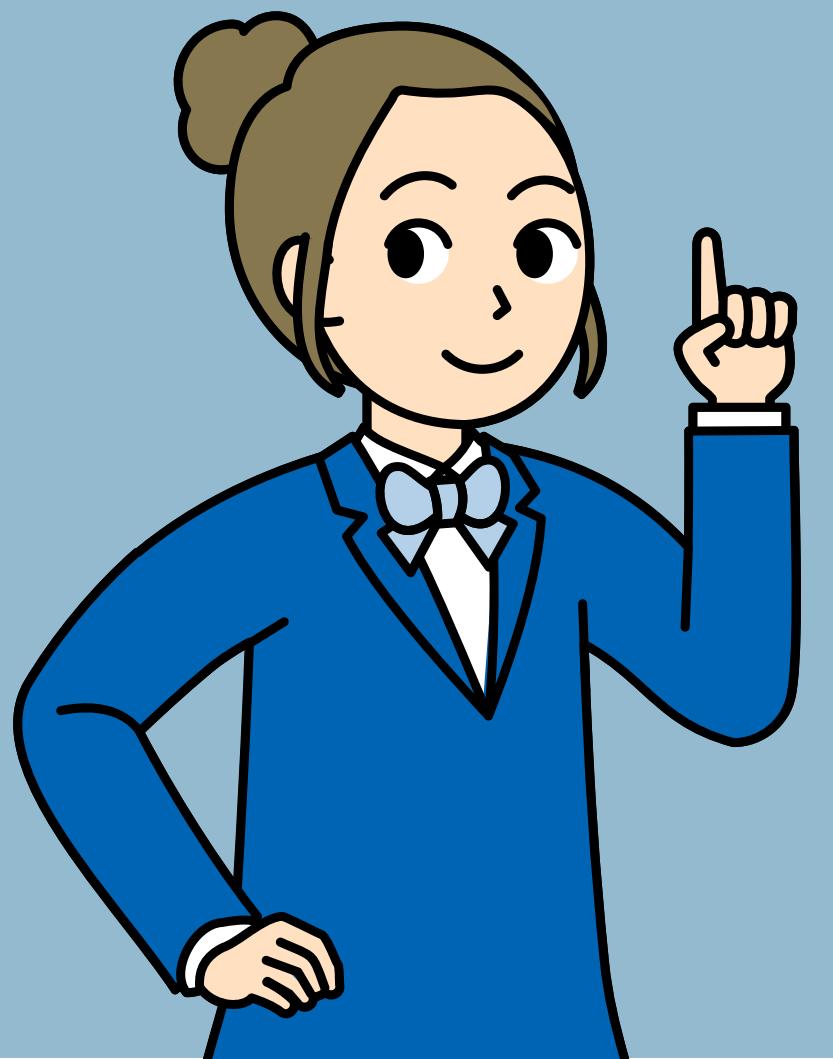
| Random Forest | Precision | Recall | F1-score |
|---------------|-----------|--------|----------|
| Lớp 0         | 0.87      | 0.20   | 0.33     |
| Lớp 1         | 0.90      | 1.00   | 0.94     |

# EVALUATION

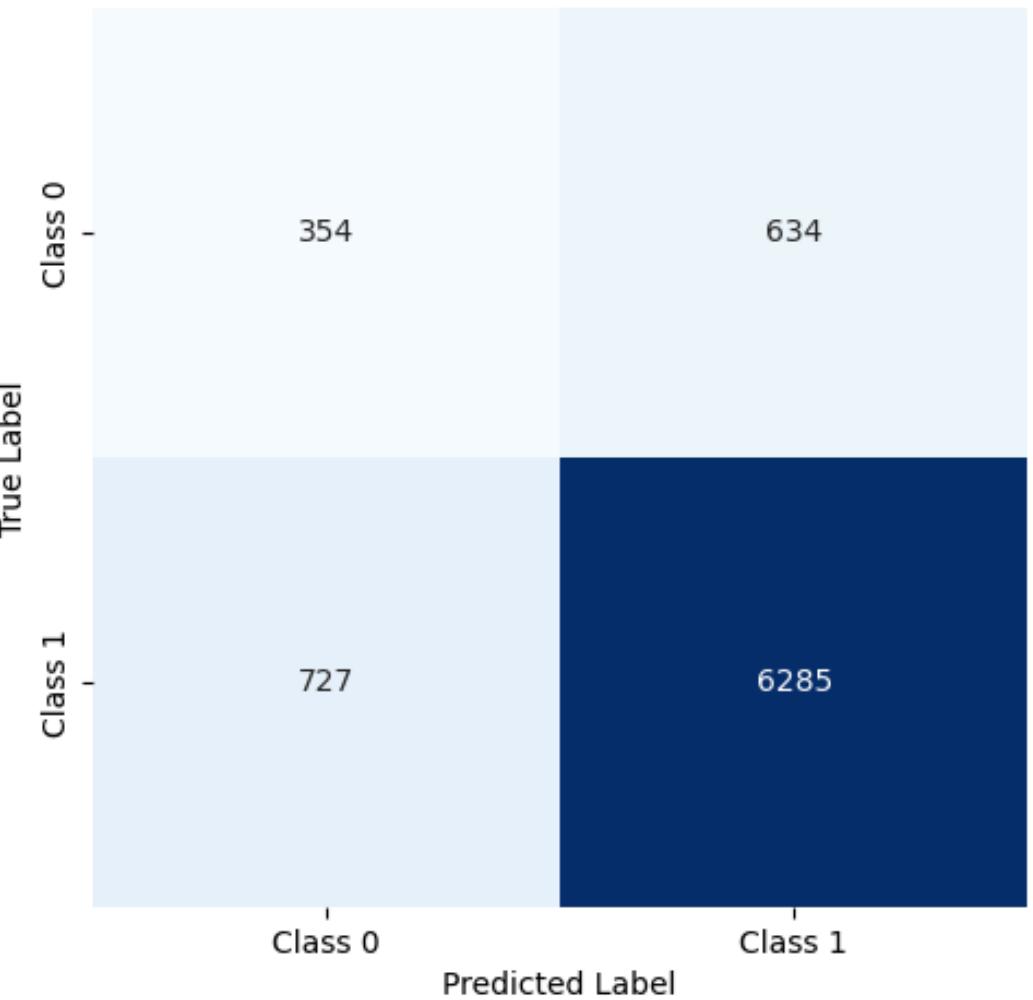
| XGBoost | Precision | Recall | F1-score |
|---------|-----------|--------|----------|
| Lớp 0   | 0.83      | 0.39   | 0.53     |
| Lớp 1   | 0.92      | 0.99   | 0.95     |

| LightGBM | Precision | Recall | F1-score |
|----------|-----------|--------|----------|
| Lớp 0    | 0.90      | 0.40   | 0.55     |
| Lớp 1    | 0.92      | 0.99   | 0.96     |

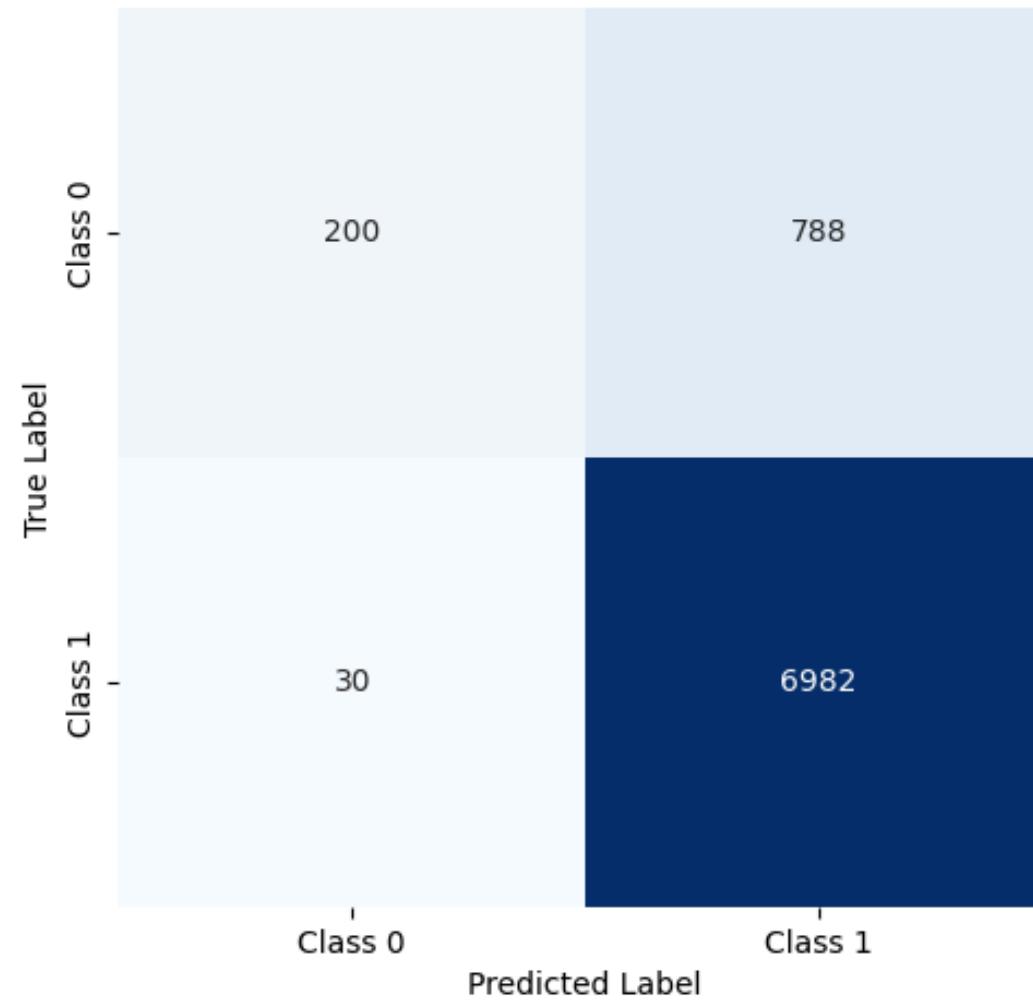
# CONFUSION MATRIX



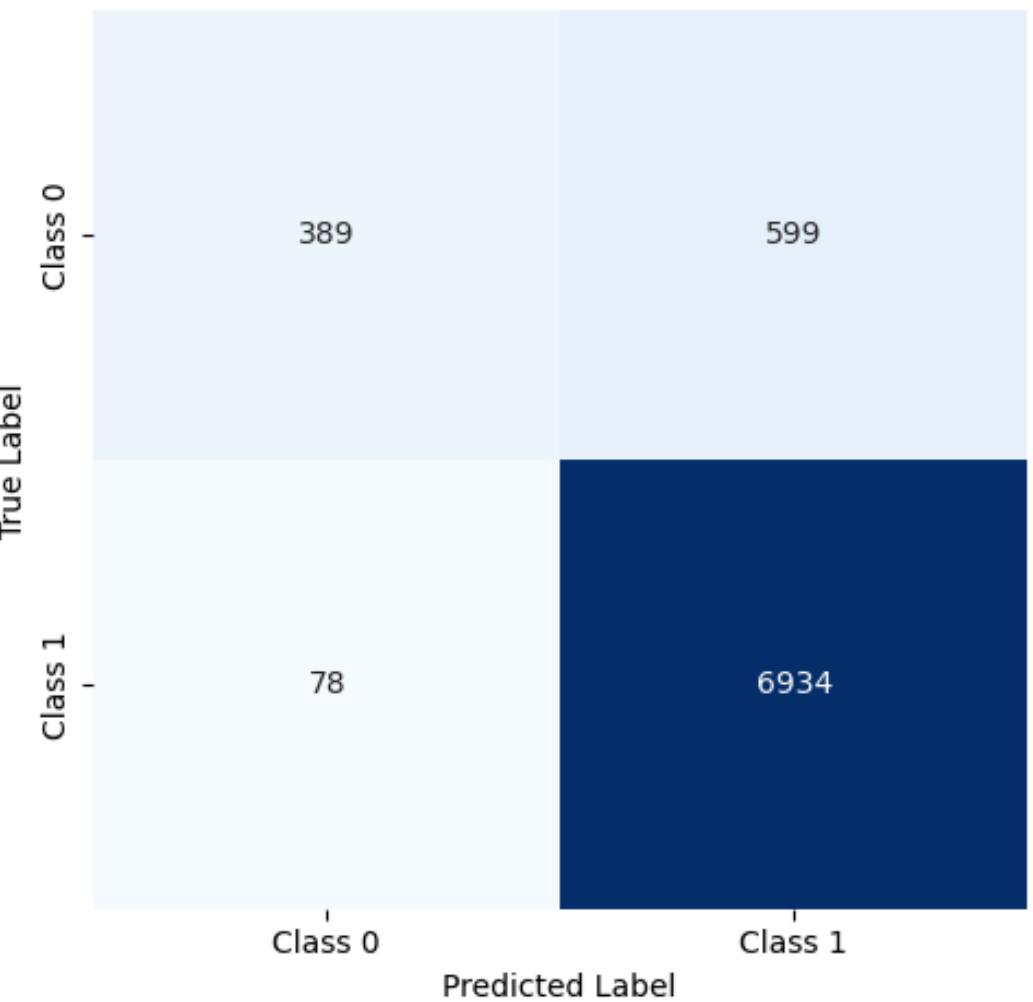
Confusion Matrix for Decision Tree



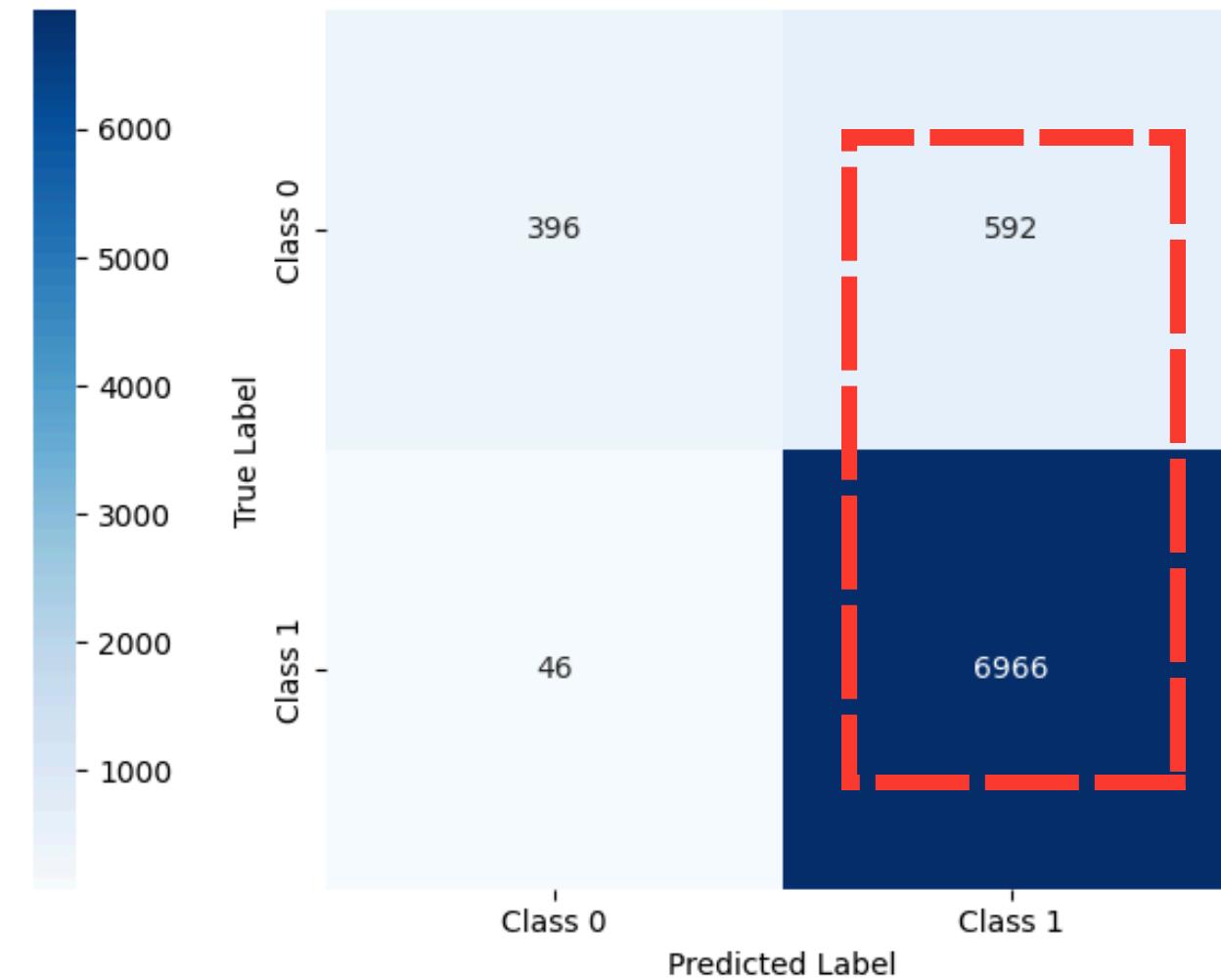
Confusion Matrix for Random Forest



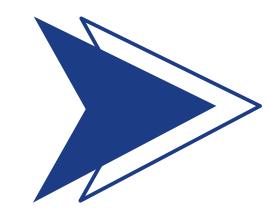
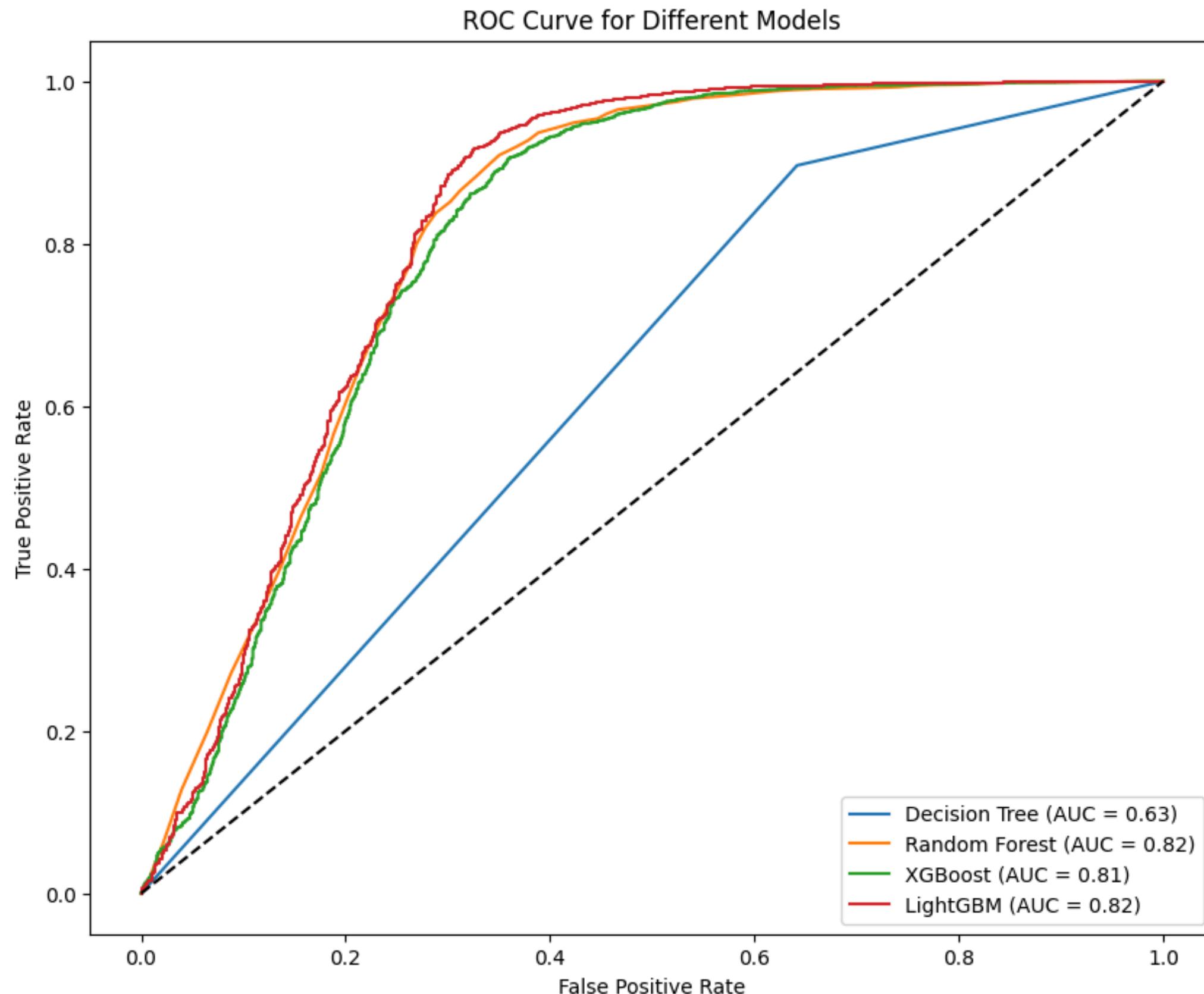
Confusion Matrix for XGBoost



Confusion Matrix for LightGBM



# EVALUATION



Mô hình **LightGBM** có hiệu quả cao trong việc xác định các khách hàng có khả năng chuyển đổi, nhưng phân loại sai lượng lớn khách hàng không chuyển đổi

# IMPROVE MODELS

---



# METHOD



## SMOTEENN

- **SMOTE:** tạo ra các mẫu tổng hợp cho lớp thiểu số
- **ENN:** loại bỏ những điểm dữ liệu không đáng tin cậy trong tập dữ liệu



## Feature Importance

- Tính toán độ quan trọng của các đặc trưng
- Top 5 đặc trưng quan trọng nhất mỗi mô hình



# EVALUATION

| Decision Tree | Train | Test |
|---------------|-------|------|
| Accuracy      | 0.88  | 0.74 |
| Precision     | 0.98  | 0.91 |
| Recall        | 0.88  | 0.78 |
| F1-score      | 0.93  | 0.84 |
| AUC           | 0.88  | 0.61 |

| Random Forest | Train | Test |
|---------------|-------|------|
| Accuracy      | 0.92  | 0.83 |
| Precision     | 0.98  | 0.92 |
| Recall        | 0.92  | 0.88 |
| F1-score      | 0.95  | 0.90 |
| AUC           | 0.96  | 0.75 |

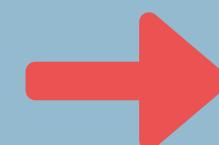
# EVALUATION

---

| XGBoost   | Train | Test |
|-----------|-------|------|
| Accuracy  | 0.94  | 0.87 |
| Precision | 0.98  | 0.93 |
| Recall    | 0.95  | 0.92 |
| F1-score  | 0.97  | 0.92 |
| AUC       | 0.96  | 0.78 |

| LightGBM  | Train | Test |
|-----------|-------|------|
| Accuracy  | 0.92  | 0.87 |
| Precision | 0.97  | 0.94 |
| Recall    | 0.94  | 0.91 |
| F1-score  | 0.96  | 0.92 |
| AUC       | 0.94  | 0.79 |

# EVALUATION



| Decision Tree | Precision | Recall | F1-score |
|---------------|-----------|--------|----------|
| Lớp 0         | 0.22      | 0.44   | 0.55     |
| Lớp 1         | 0.91      | 0.78   | 0.84     |



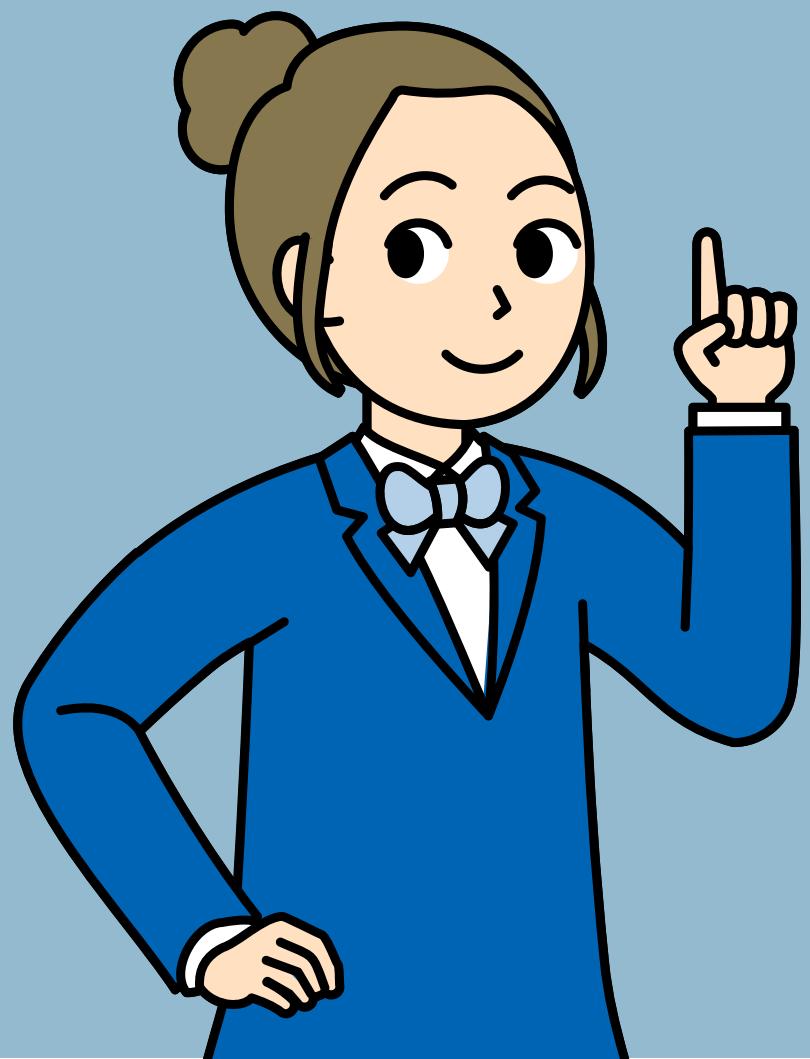
| Random Forest | Precision | Recall | F1-score |
|---------------|-----------|--------|----------|
| Lớp 0         | 0.35      | 0.47   | 0.40     |
| Lớp 1         | 0.92      | 0.88   | 0.90     |

# EVALUATION

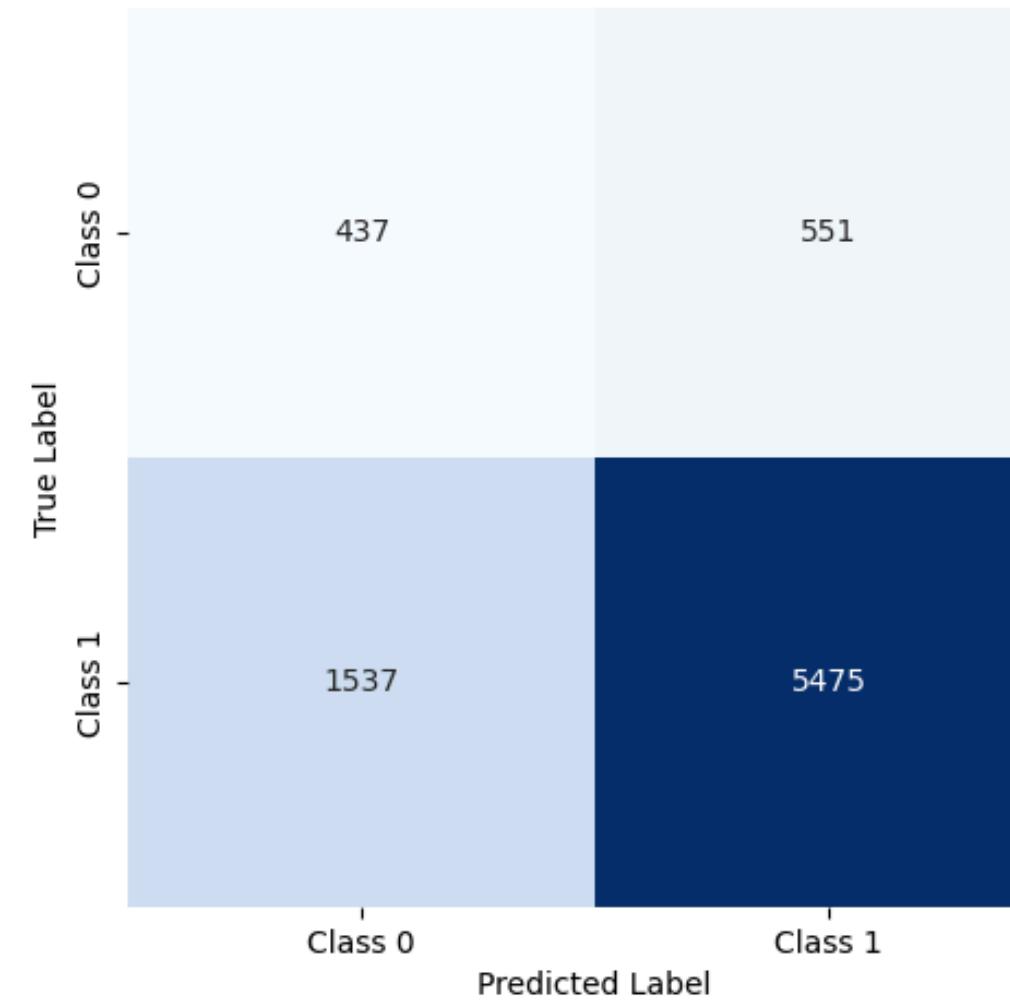
| XGBoost | Precision | Recall   | F1-score |
|---------|-----------|----------|----------|
| Lớp 0   | 0.46      | [ 0.50 ] | 0.48     |
| Lớp 1   | 0.93      | 0.92     | 0.92     |

| LightGBM | Precision | Recall   | F1-score |
|----------|-----------|----------|----------|
| Lớp 0    | 0.47      | [ 0.56 ] | 0.51     |
| Lớp 1    | 0.94      | 0.91     | 0.92     |

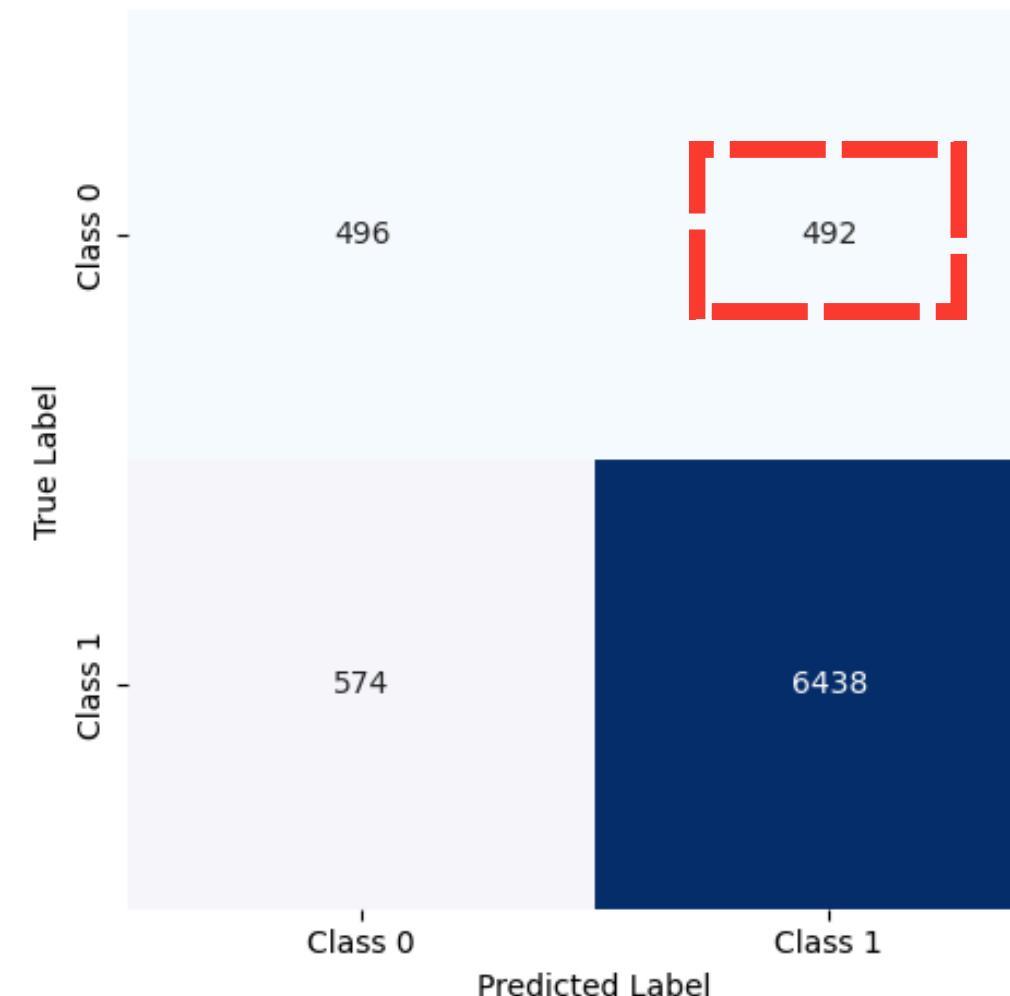
# CONFUSION MATRIX



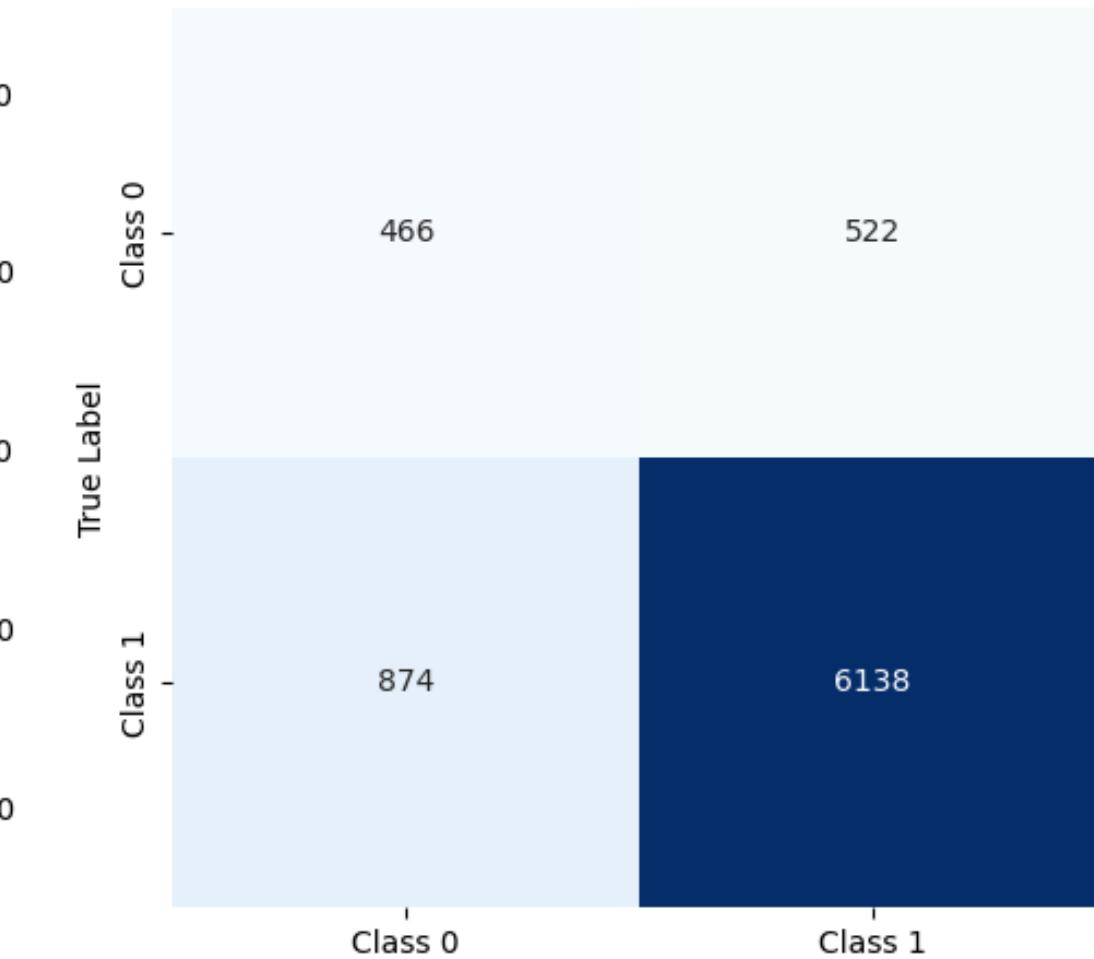
Confusion Matrix for Decision Tree



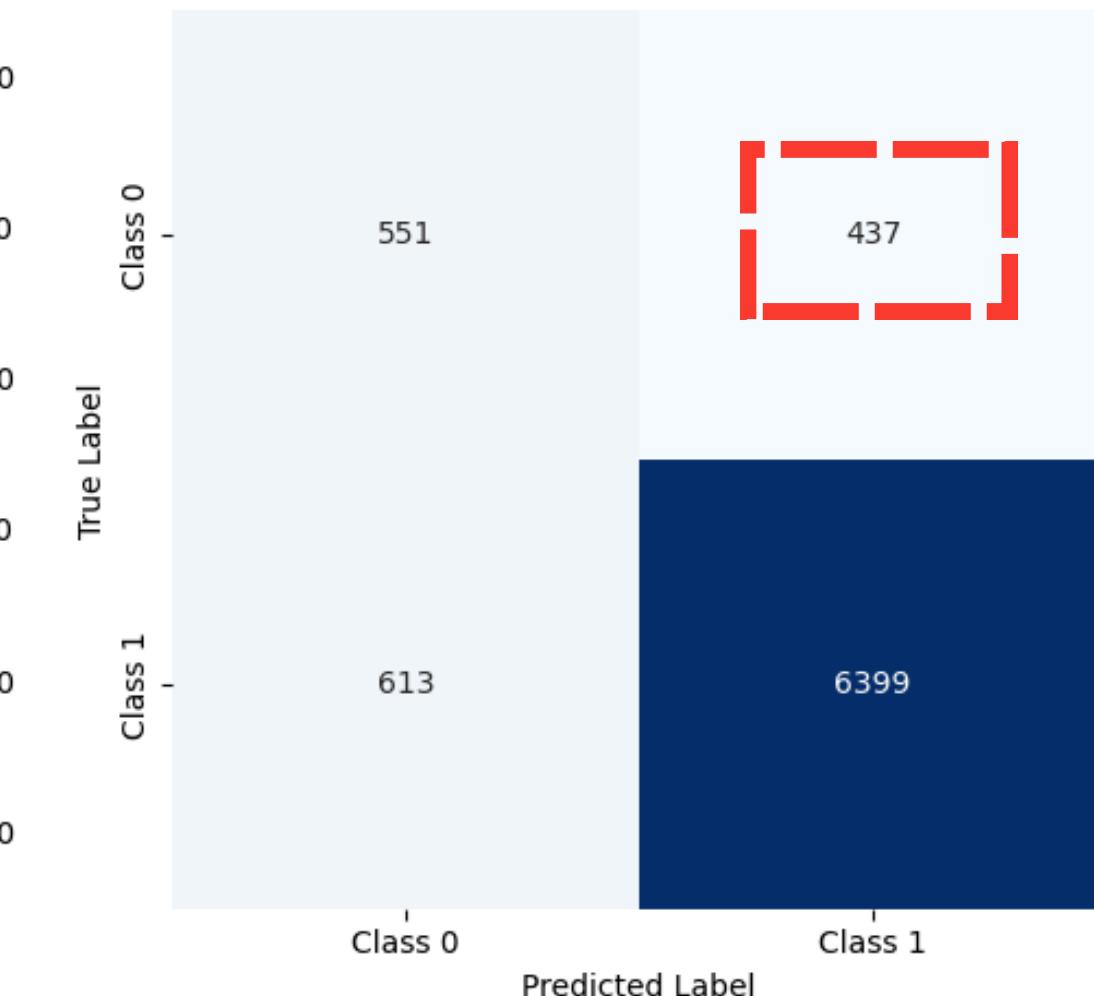
Confusion Matrix for XGBoost



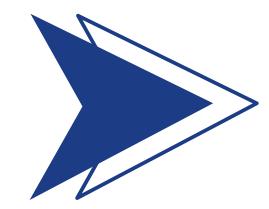
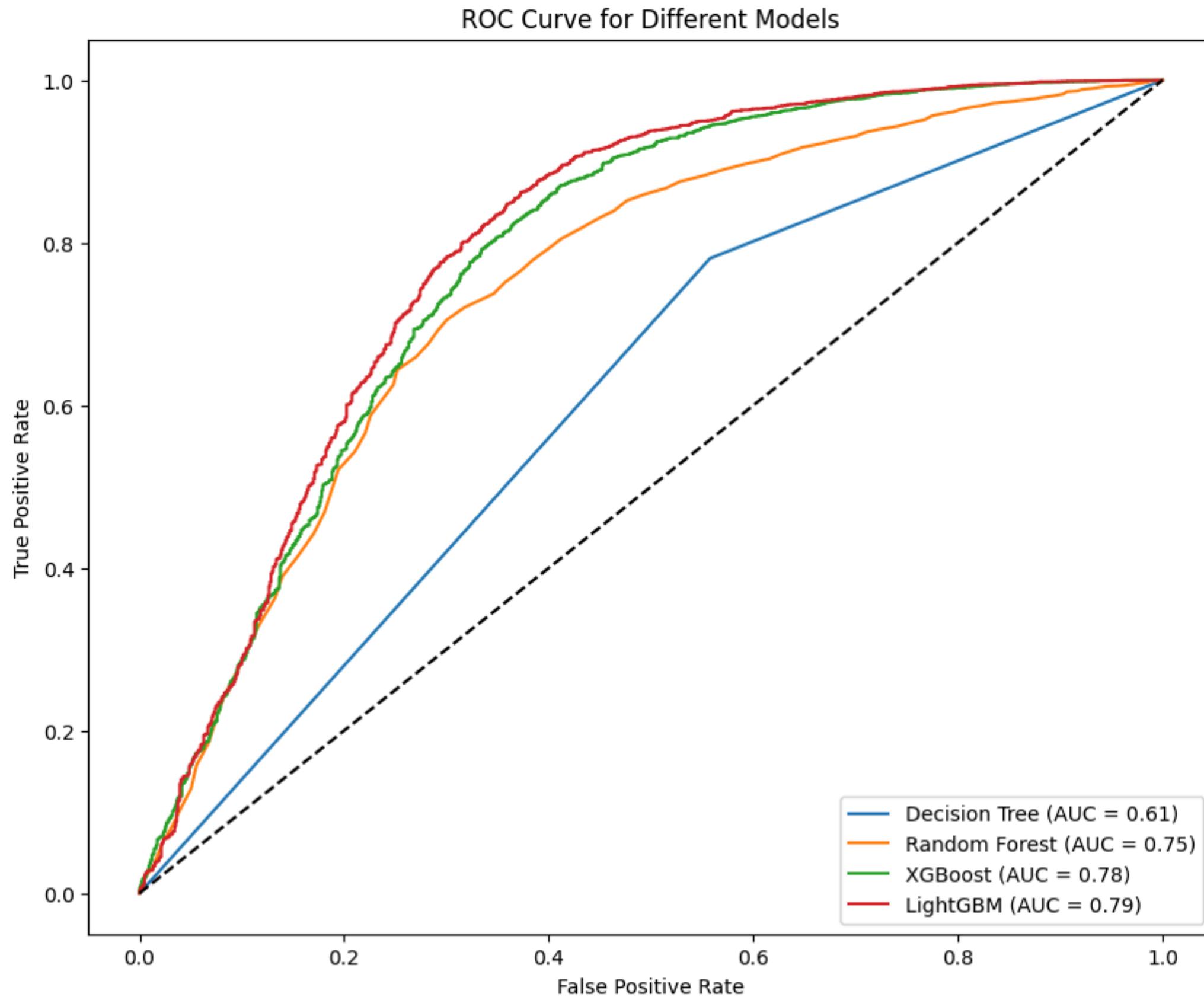
Confusion Matrix for Random Forest



Confusion Matrix for LightGBM

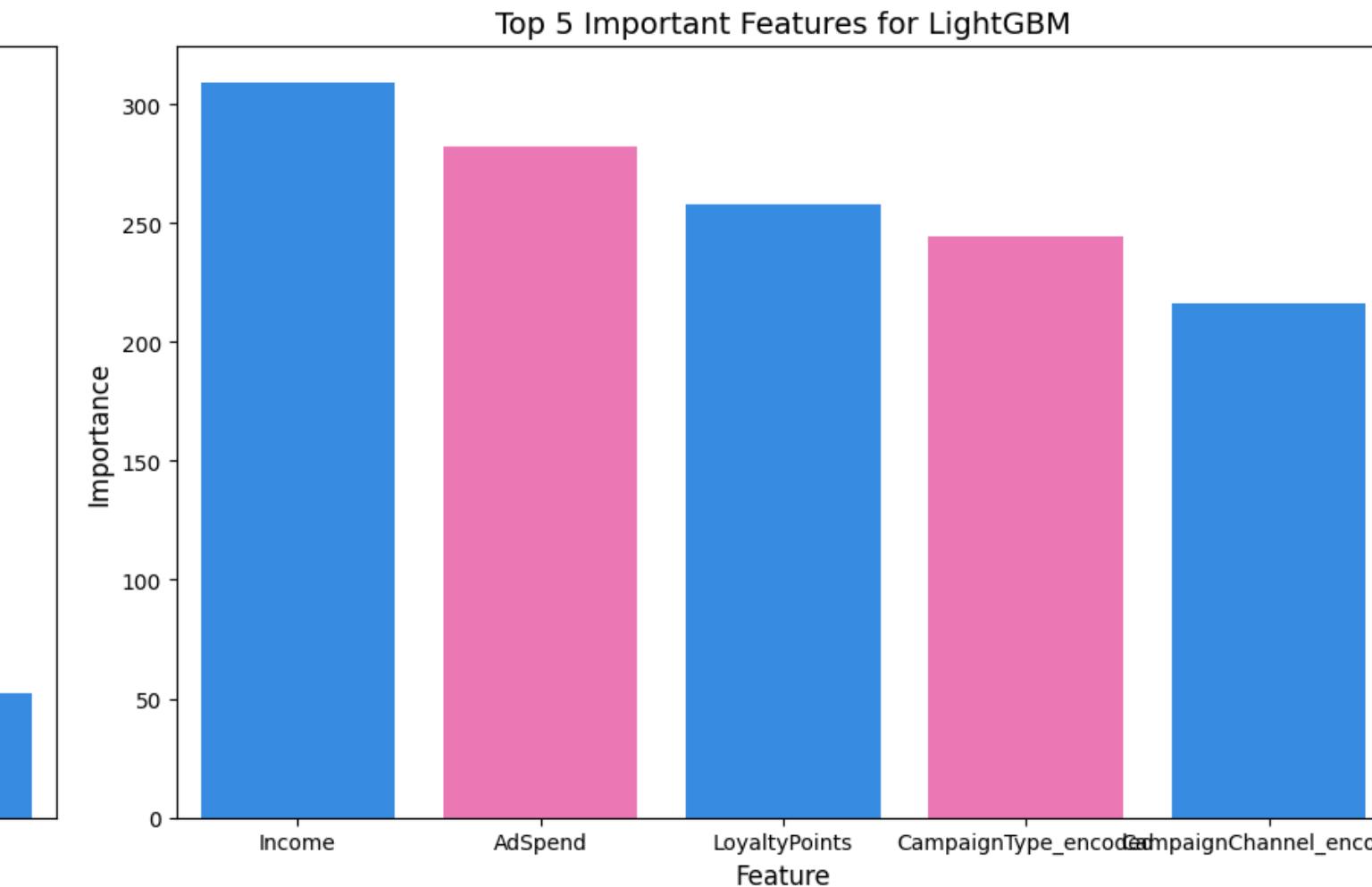
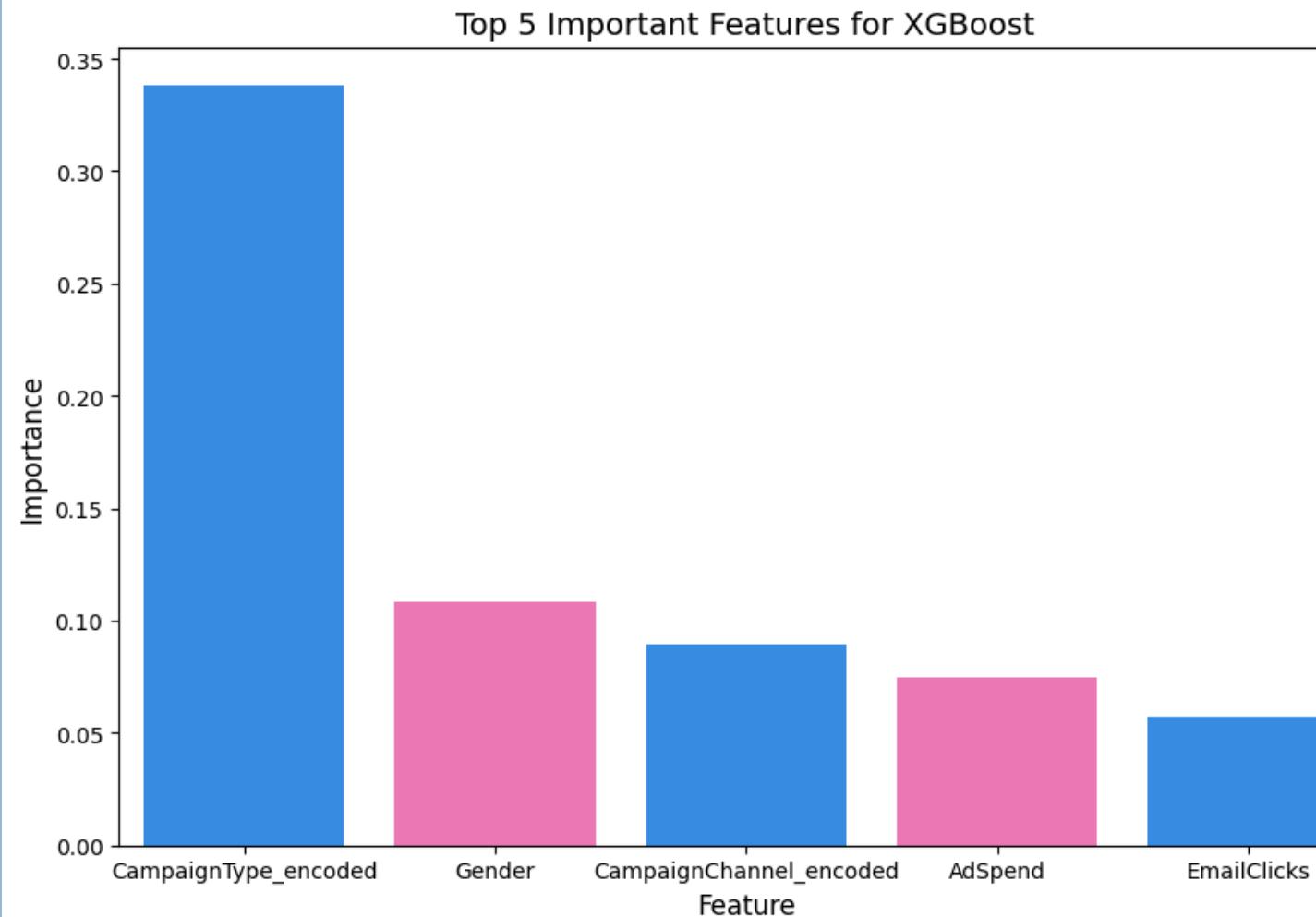
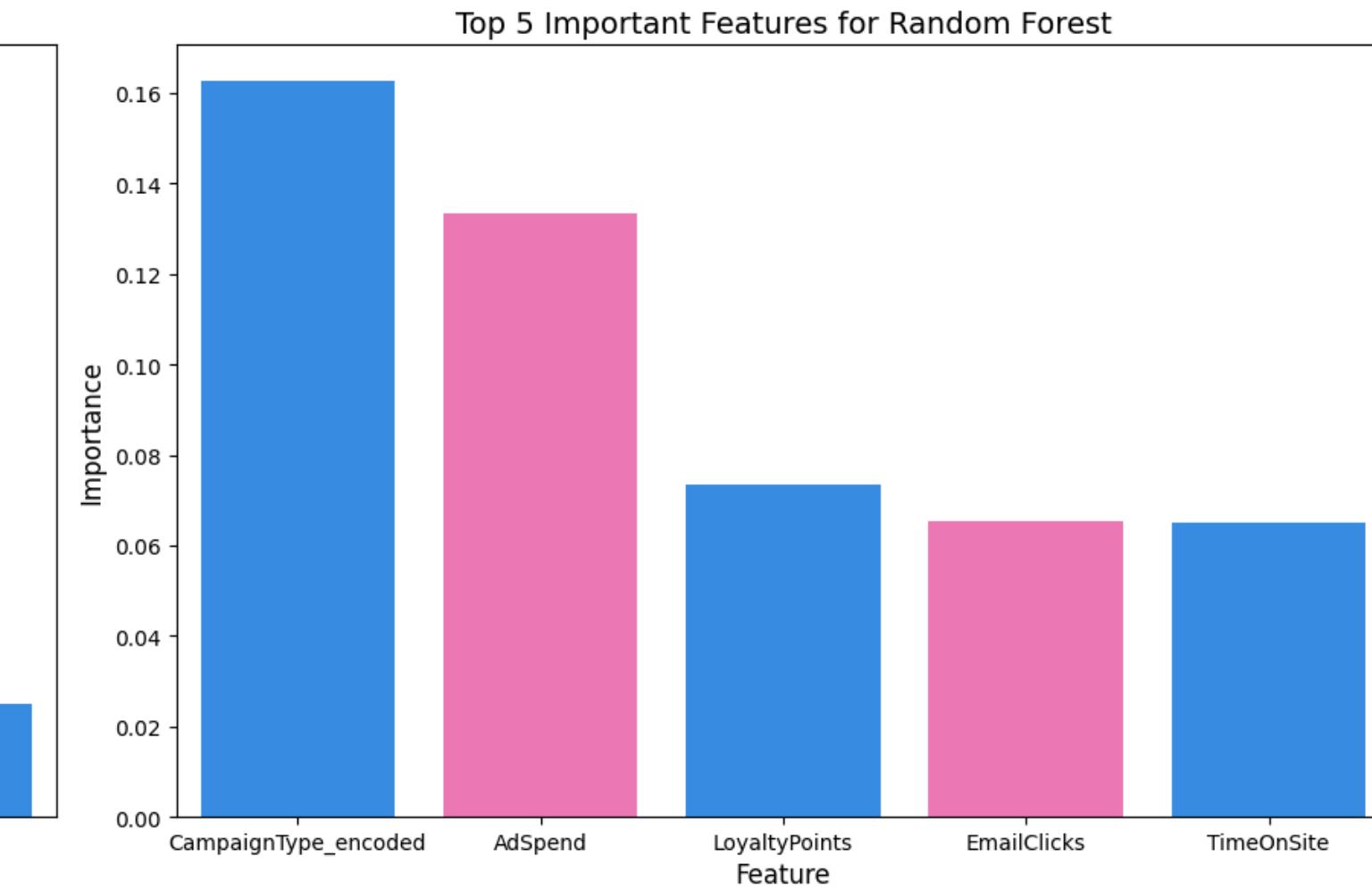
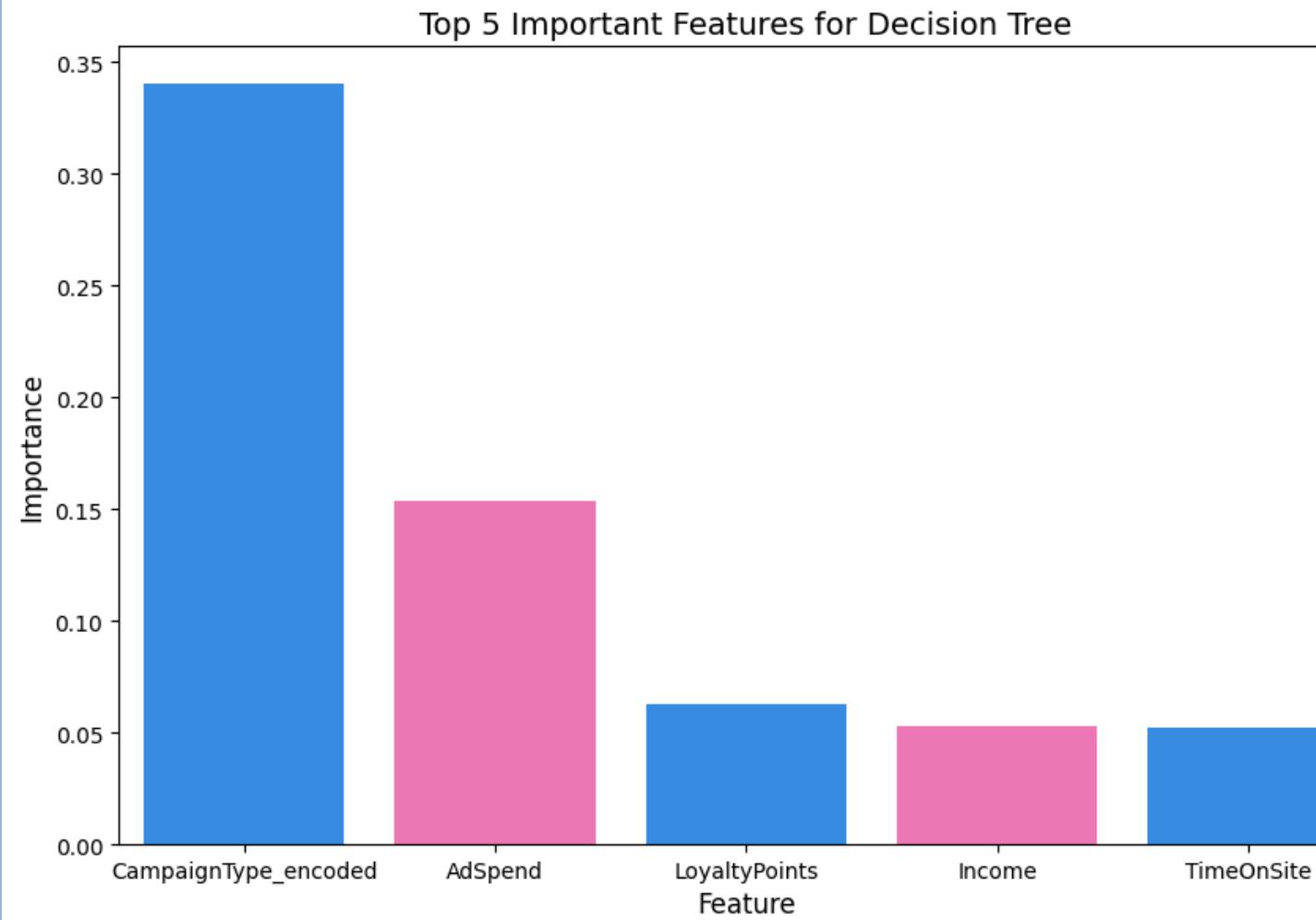
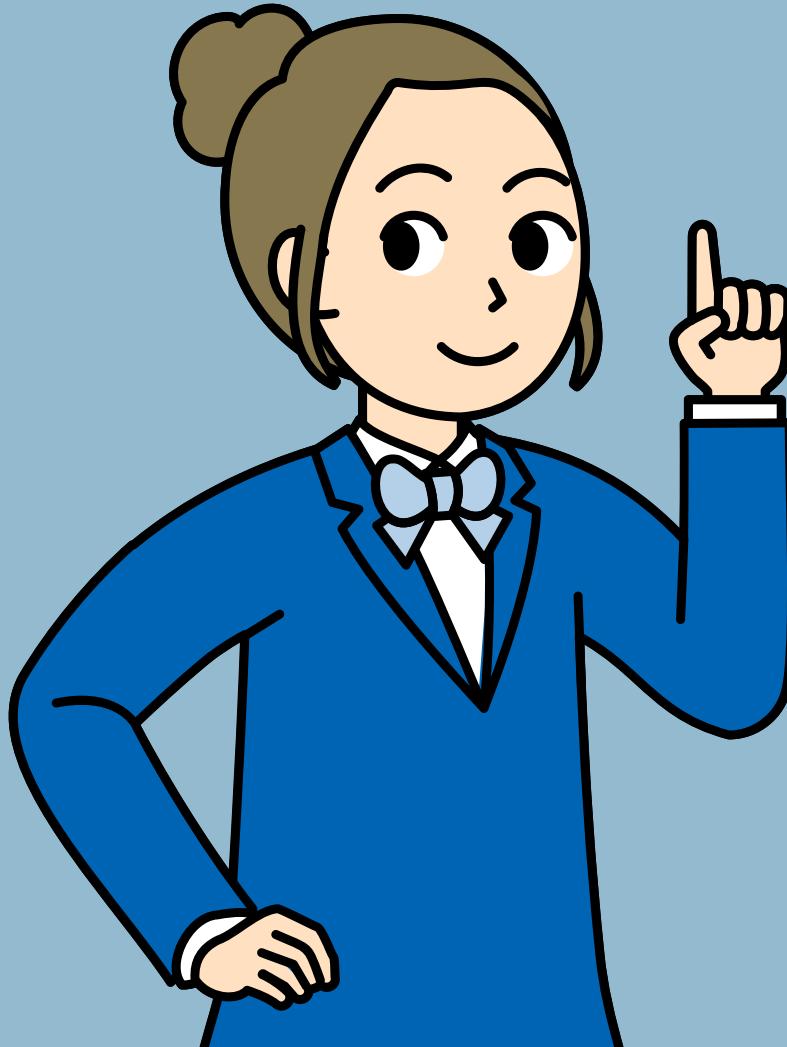


# EVALUATION



Mô hình **LightGBM** vẫn đứng đầu  
về hiệu quả phân biệt lớp

# TOP 5 FEATURES IMPORTANCE



# CONCLUSION

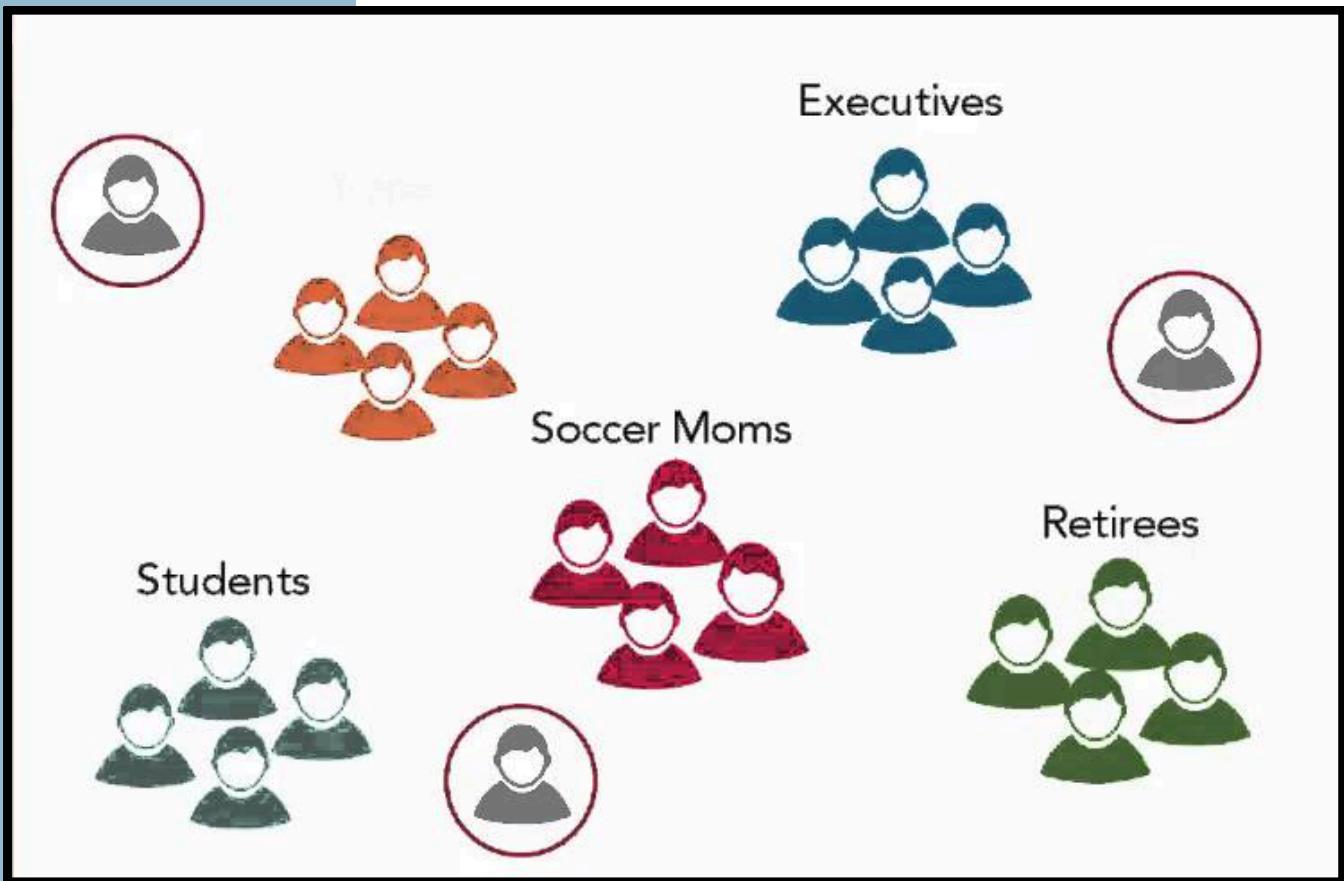
---

- ❖ Mô hình **LightGBM** sau khi cải thiện **phù hợp** và **tối ưu nhất** cho mục tiêu ban đầu, nhưng vẫn cần **cải thiện thêm**
- ❖ Các đặc trưng bao gồm: **Income, Adspend, Loyalty Points, CampaignType, CampaignChannel** có ảnh hưởng lớn đến sự chuyển đổi của khách hàng

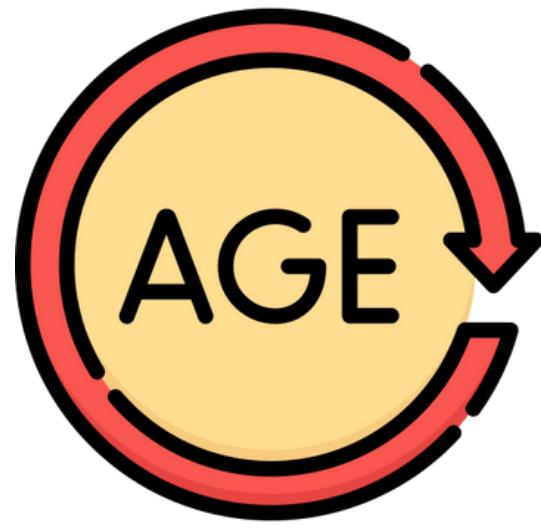


# CLUSTERING

by demographics & behaviour



# FEATURES SELECTED TO CLUSTER



TUỔI TÁC



THU NHẬP

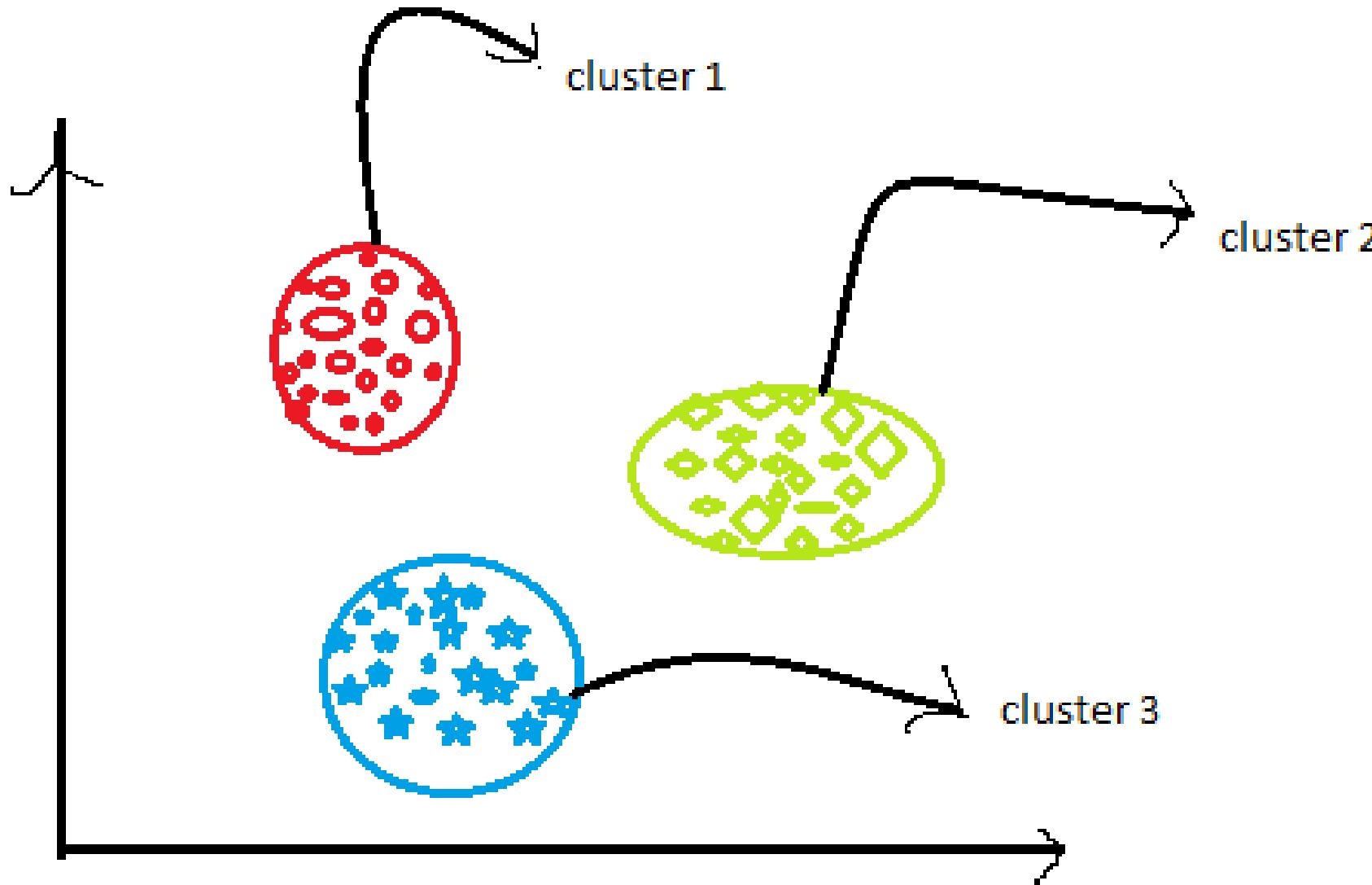


ĐIỂM TRUNG THÀNH

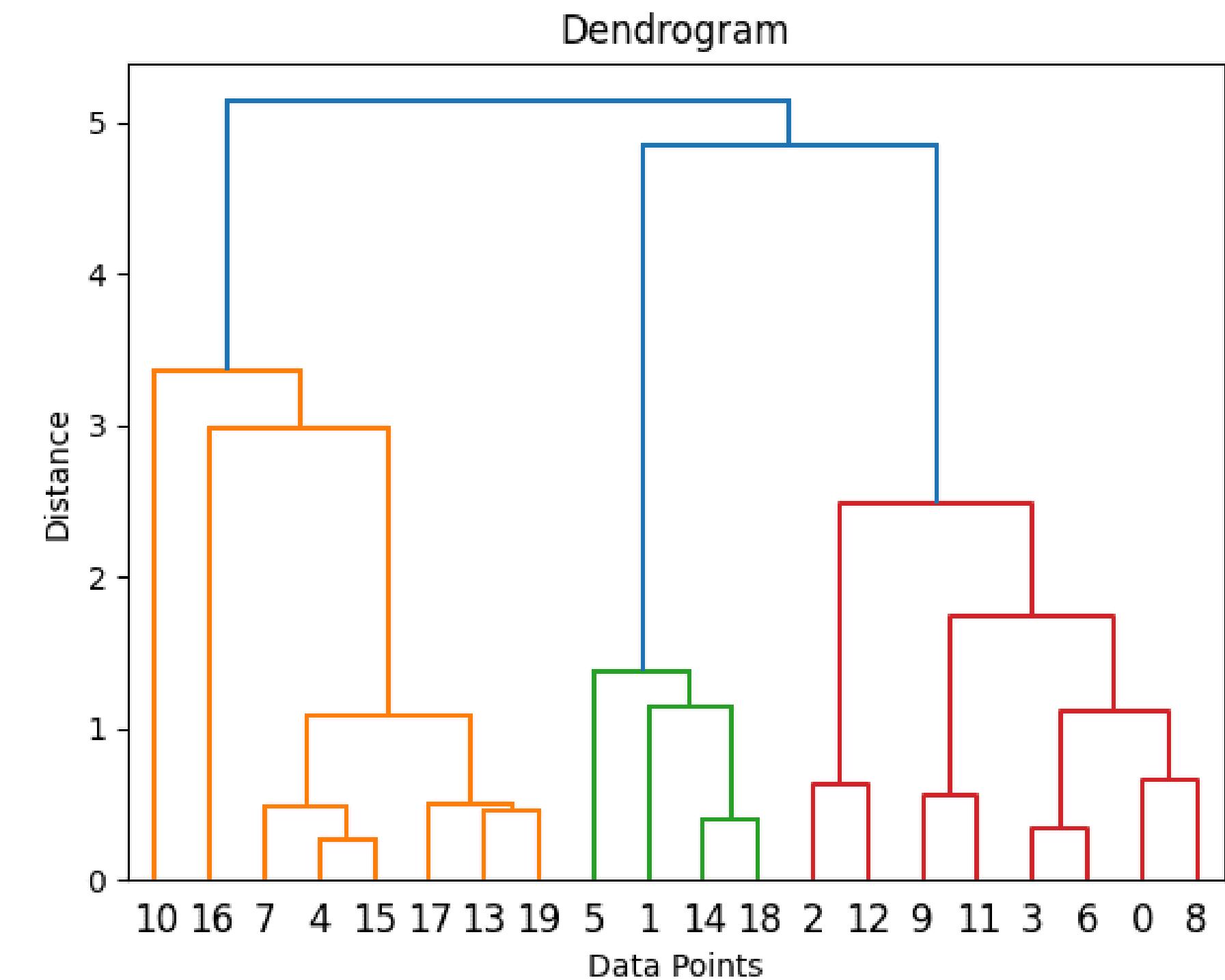
Giúp phân cụm khách hàng tốt hơn dựa trên cả đặc điểm về  
nhân khẩu học và hành vi

# ALGORITHM TO CLUSTER

## K - MEANS CLUSTERING



## HIERARCHICAL CLUSTERING

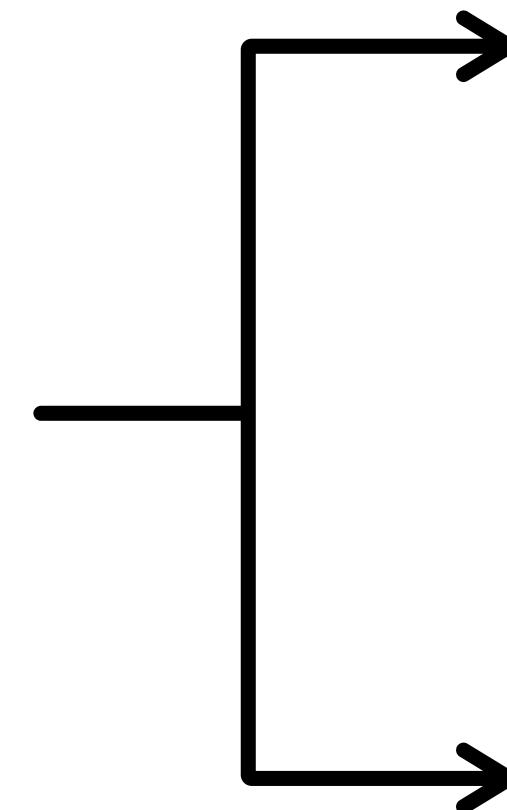


# ALGORITHM TO CLUSTER

## K - MEANS CLUSTERING

### ƯU ĐIỂM

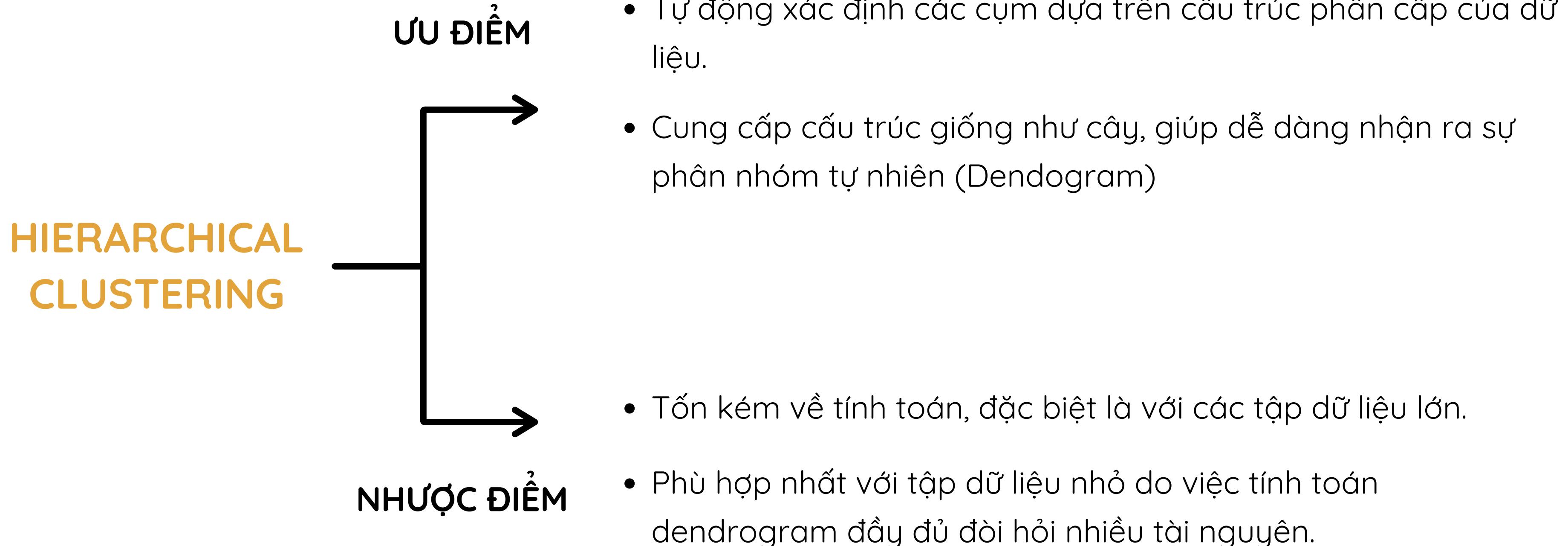
- Làm việc tốt với các tập dữ liệu lớn và có hiệu suất tính toán cao.



### NHƯỢC ĐIỂM

- Tạo ra các cụm với các trung tâm rõ ràng, giúp việc diễn giải kết quả trở nên đơn giản hơn.
- Các lần khởi tạo khác nhau có thể cho kết quả khác nhau
- Giả định các cụm có dạng cầu, điều này có thể không phản ánh đúng phân phối thực tế.
- Nhạy cảm với outlier

# ALGORITHM TO CLUSTER



# EVALUATION BY SILHOUETTE SCORE

- Đo lường mức độ tương đồng của một điểm với cụm của nó so với các cụm khác, thích hợp để đánh giá chất lượng của cụm trong cả hai thuật toán **K-Means** và **Hierarchical Clustering**.
- Ý nghĩa:
  - Silhouette Score gần +1 cho thấy mẫu dữ liệu phù hợp với cụm của nó.
  - Điểm số khoảng 0 cho thấy mẫu nằm ở ranh giới của hai cụm.
  - Điểm âm cho thấy mẫu có thể đã được gán sai cụm.

# RESULTS

Biểu đồ

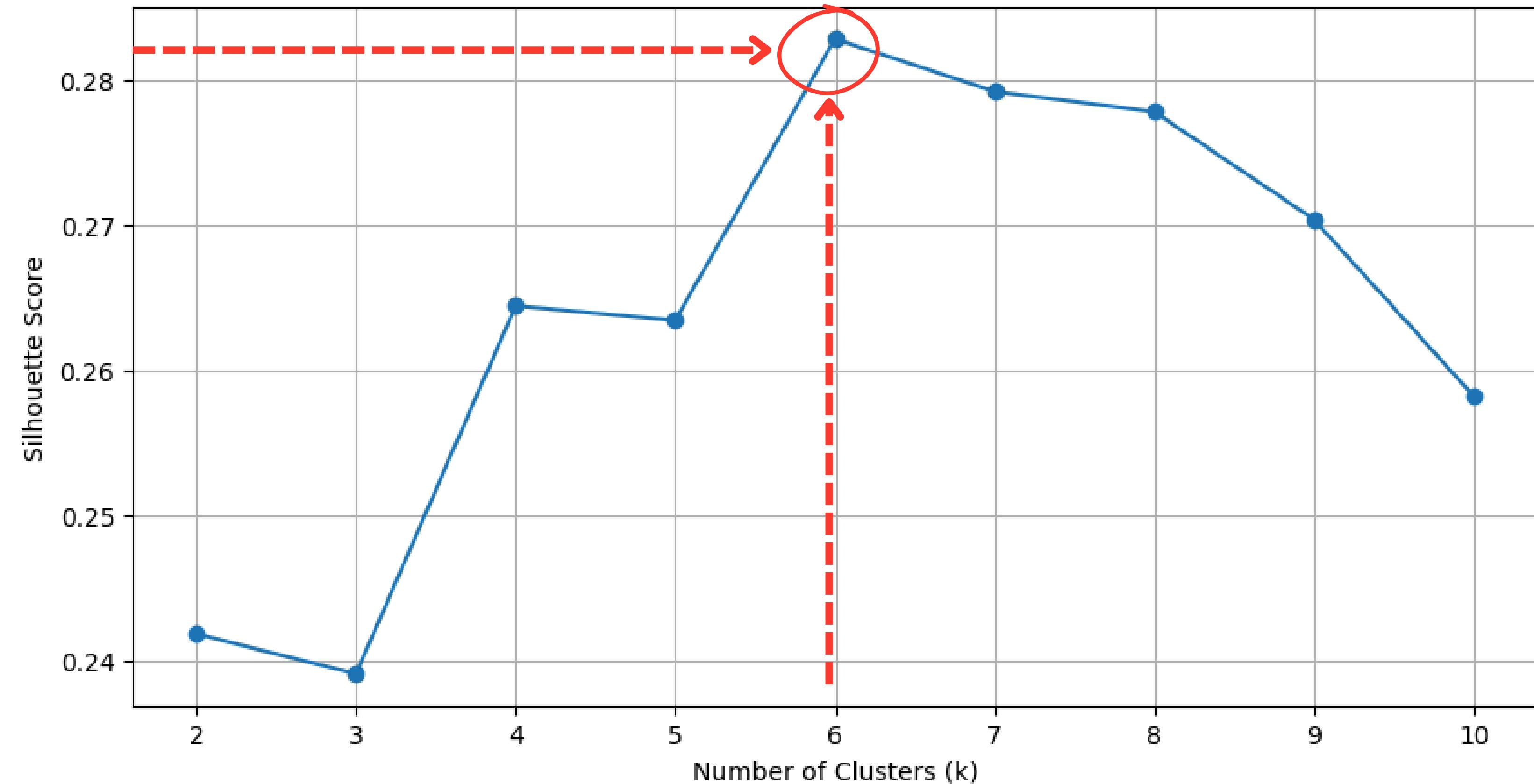
Đánh giá

Silhouette

Score

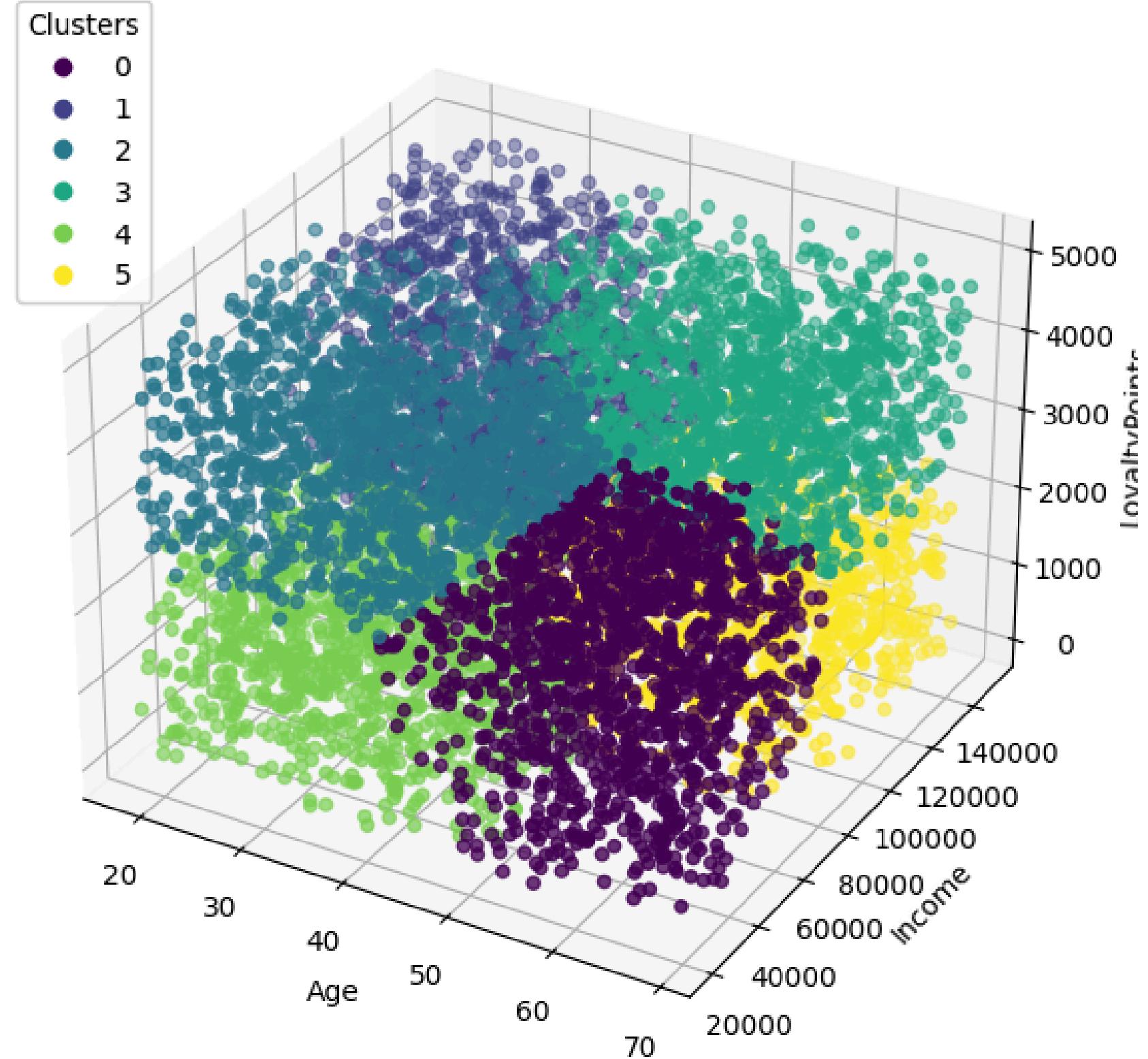
theo số

lượng cụm

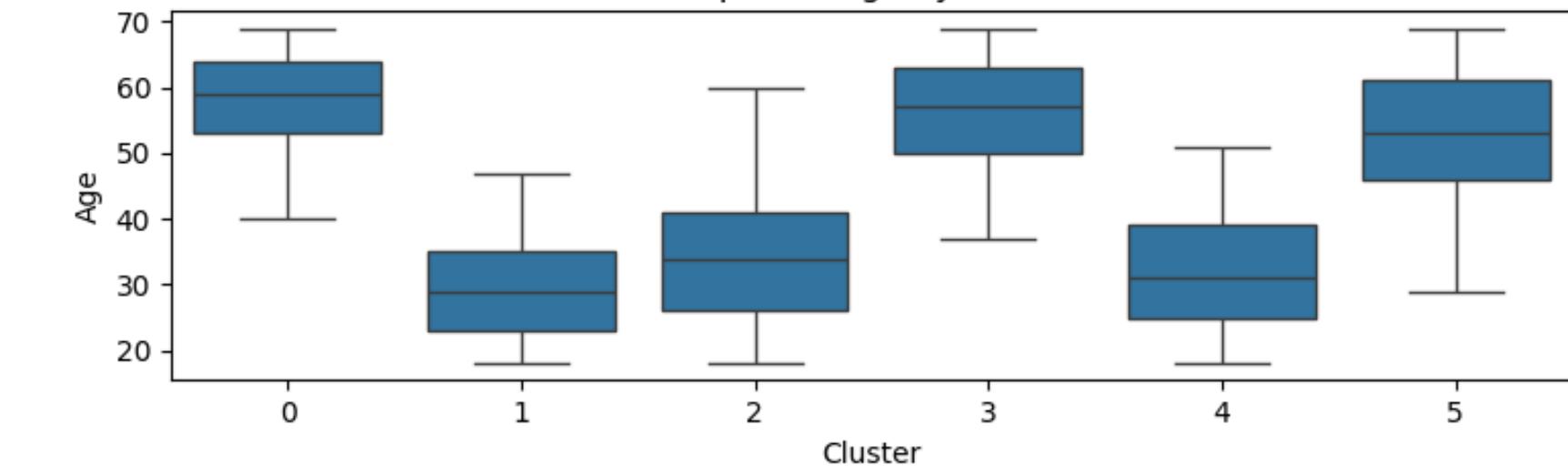


# RESULTS

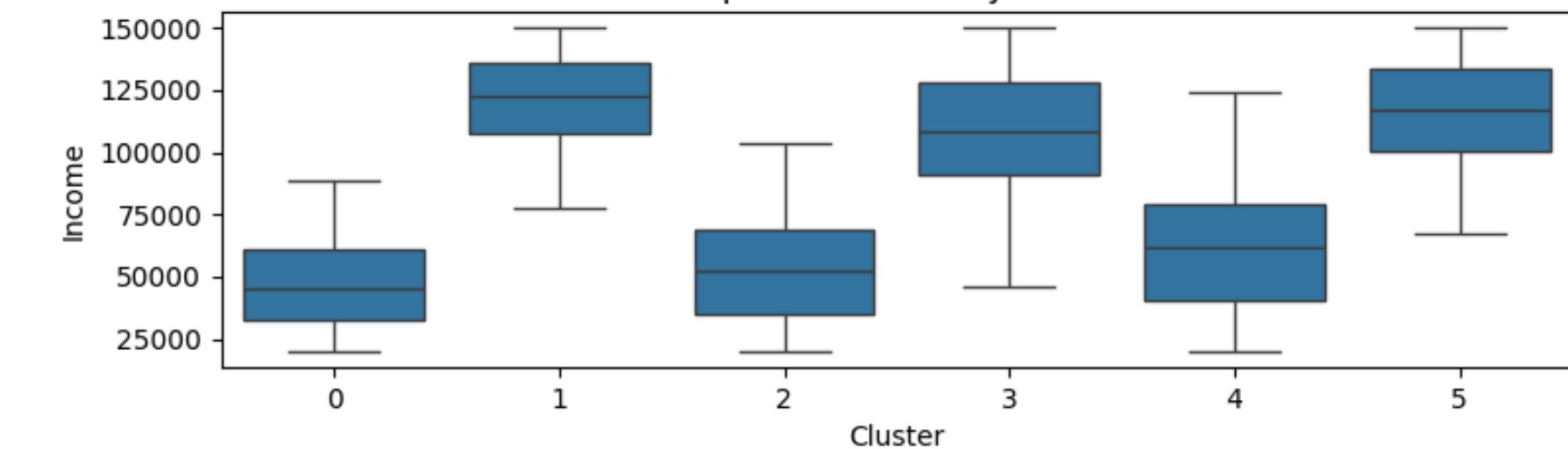
3D Visualization of Clusters



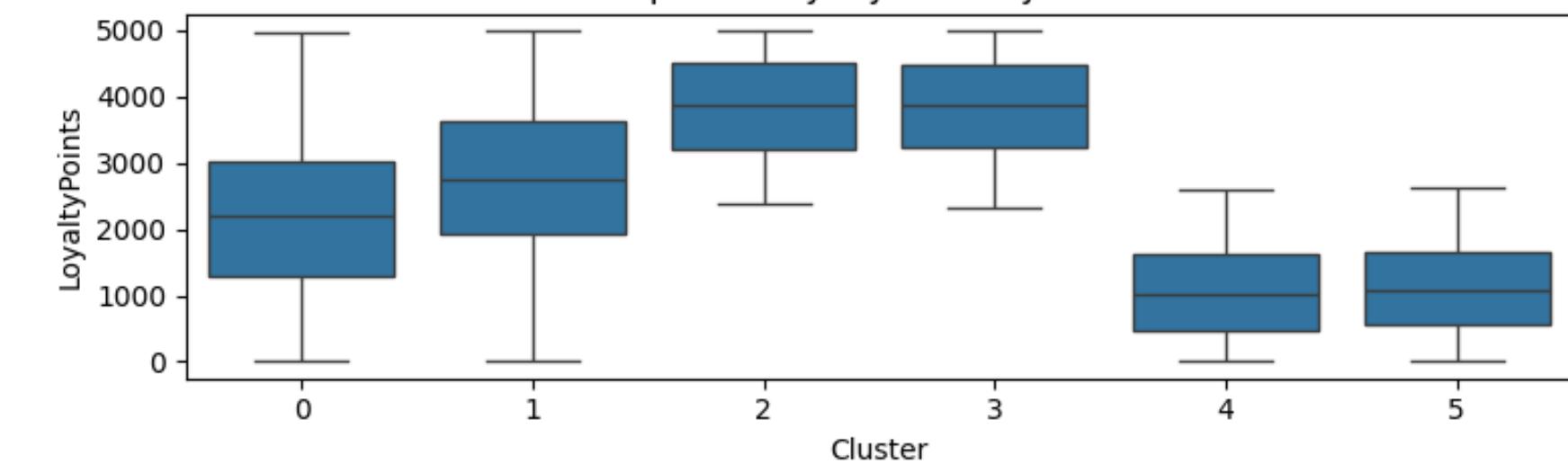
Boxplot of Age by Cluster



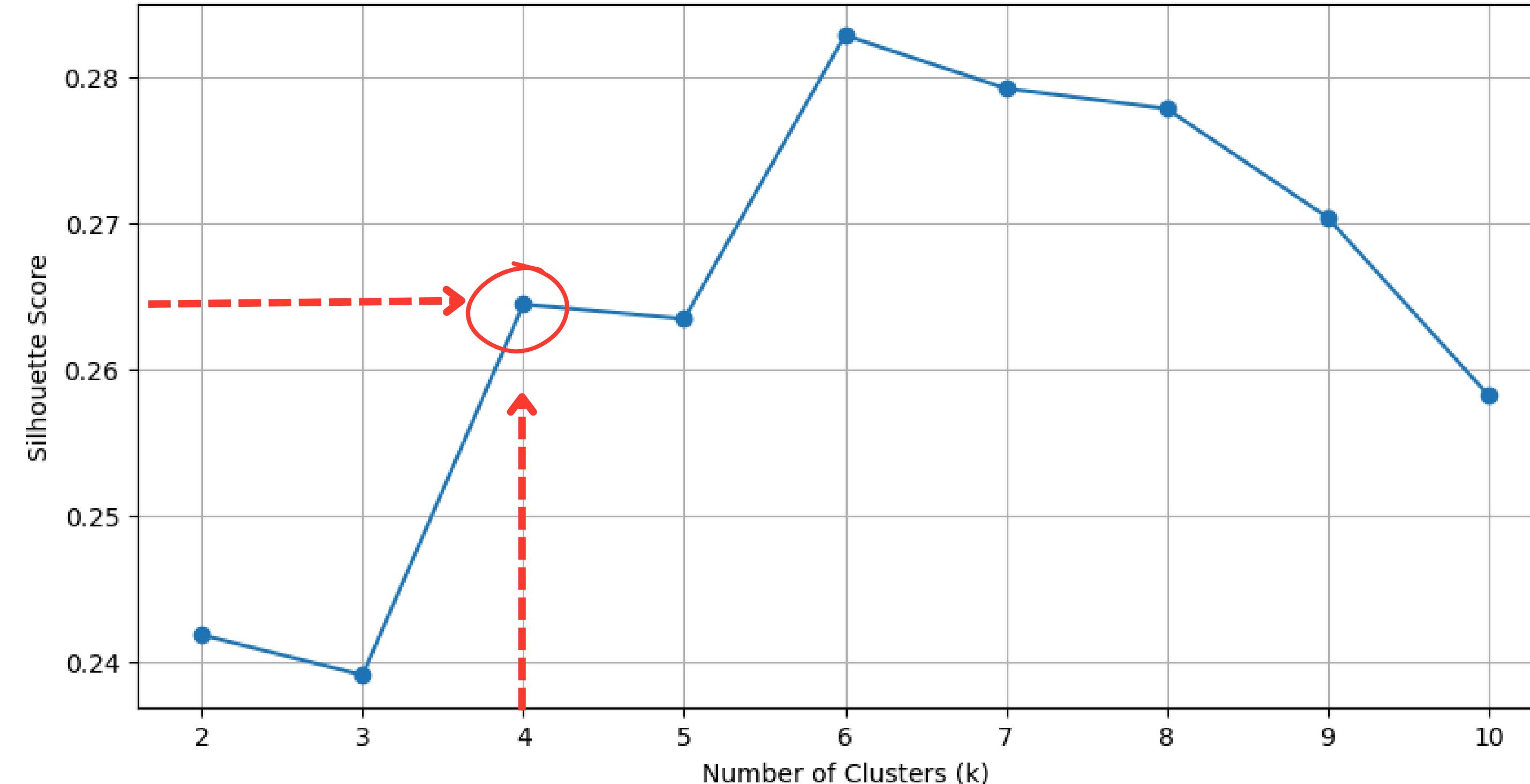
Boxplot of Income by Cluster



Boxplot of LoyaltyPoints by Cluster



# ASSESSMENT



Chọn  $K = 4$  là phù hợp

# COMPARE SILHOUETTE SCORE AT K = 4

```
# 3. Phân cụm với k cố định là 4
num_clusters = 4
kmeans = KMeans(n_clusters=num_clusters, random_state=42)
df['K-means_Cluster'] = kmeans.fit_predict(scaled_features)

# 4. In ra Silhouette Score
silhouette_avg = silhouette_score(scaled_features, df['K-means_Cluster'])
print(f'Silhouette Score for K - Means: {silhouette_avg:.4f}')

✓ 0.95
```

Silhouette Score for K - Means: 0.2644

Chọn K - Means vì có

Silhouette Score lớn hơn

```
# 5. Phân cụm dựa trên số cụm đã chọn (giả sử chọn 4 cụm từ dendrogram)
num_clusters = 4
df['Hierarchical_Cluster'] = fcluster(Z, num_clusters, criterion='maxclust')
```

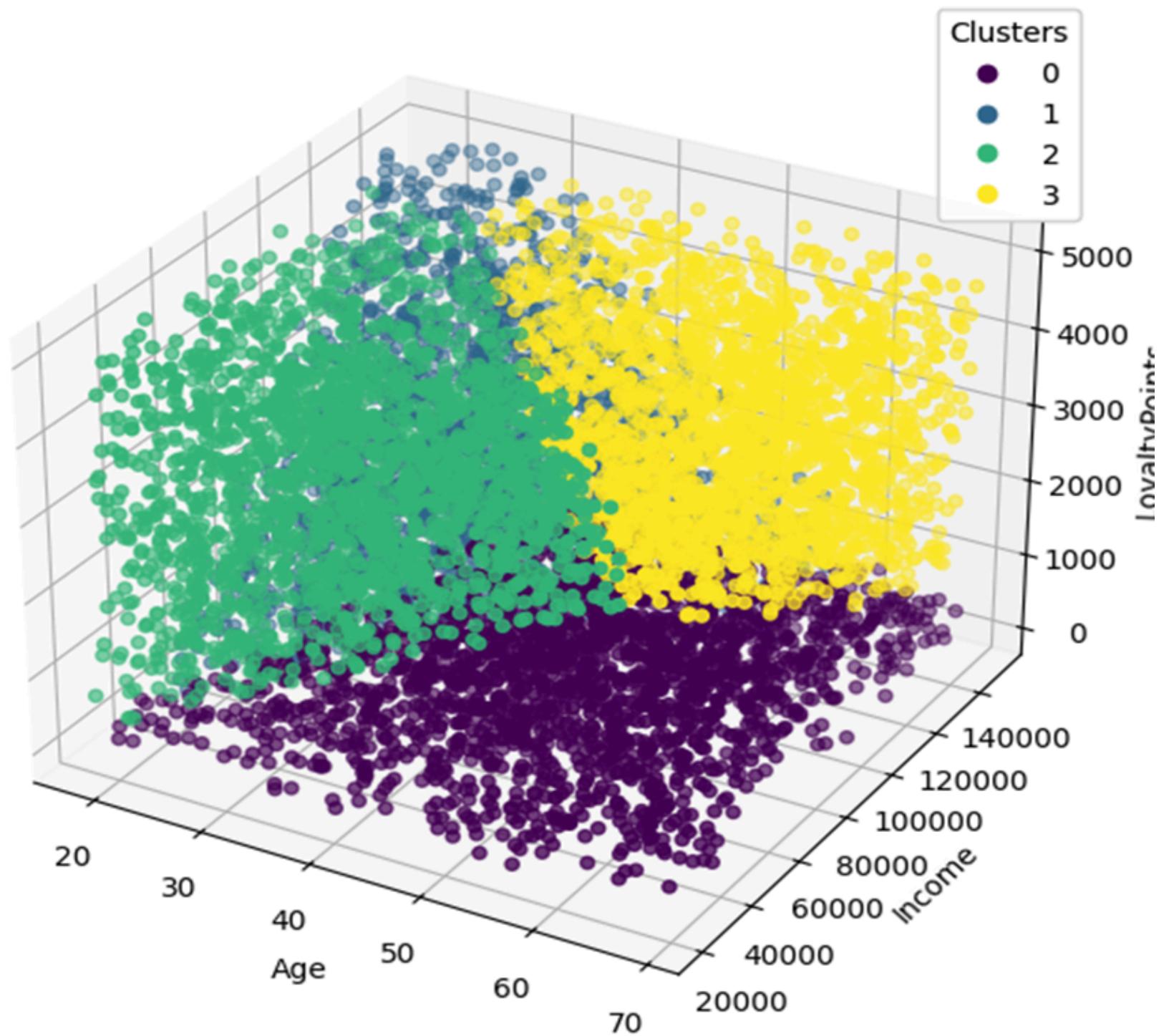
```
# 6. Tính Silhouette Score cho các cụm phân cấp
silhouette_avg = silhouette_score(scaled_features, df['Hierarchical_Cluster'])
print(f"Silhouette Score for Hierarchical Clustering: {silhouette_avg}")

✓ 2.5s
```

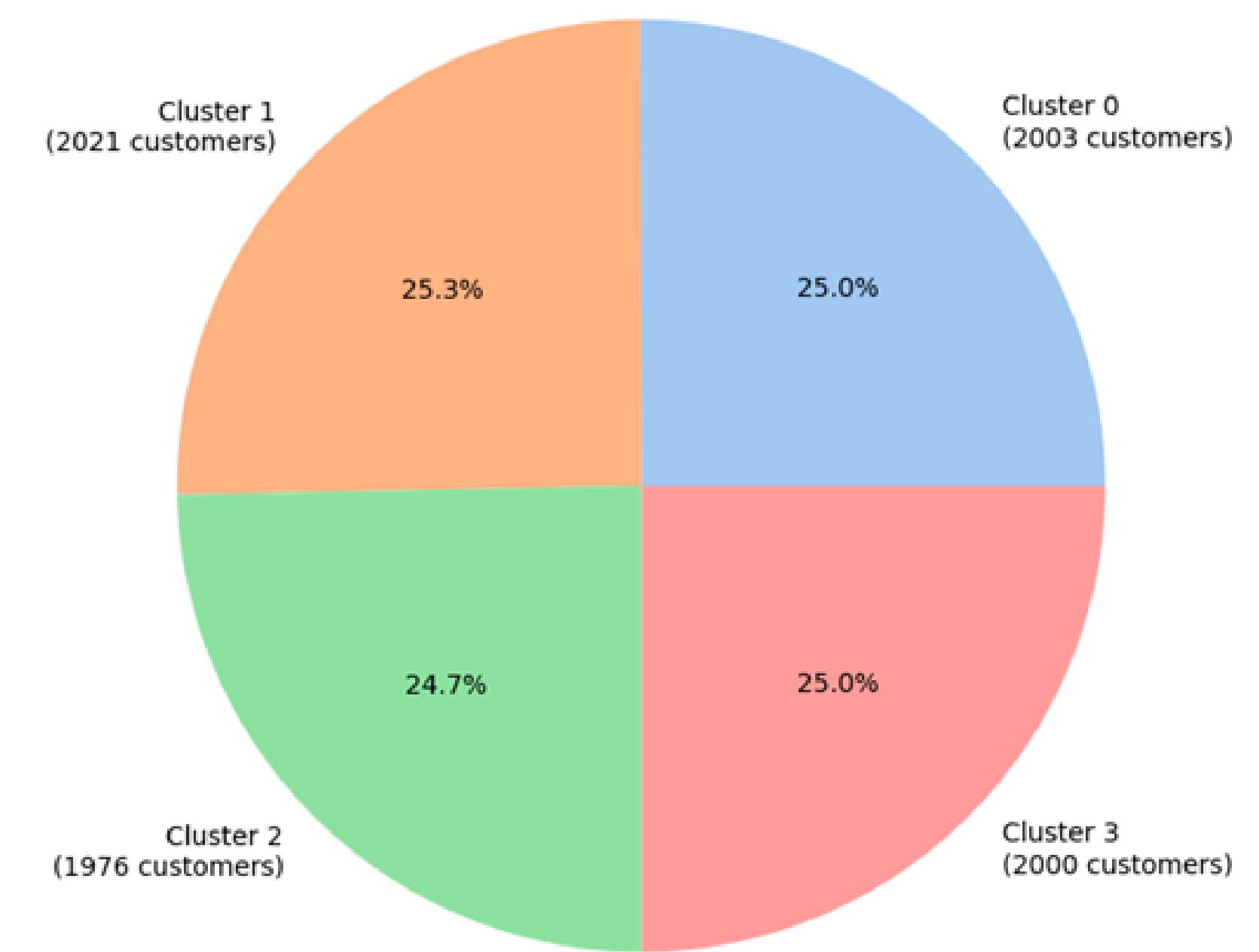
Silhouette Score for Hierarchical Clustering: 0.21

# VISUALIZE 4 CLUSTERS

3D Visualization of Clusters



Pie Chart of Cluster Distribution



# BOX PLOT OF 4 CLUSTER



# NAME CLUSTERS

## Cụm 0: Lớn tuổi - Truyền thống

- Độ tuổi: 45 - 61
- Thu nhập: 40K - 86K
- Điểm trung thành: 480 - 1,600



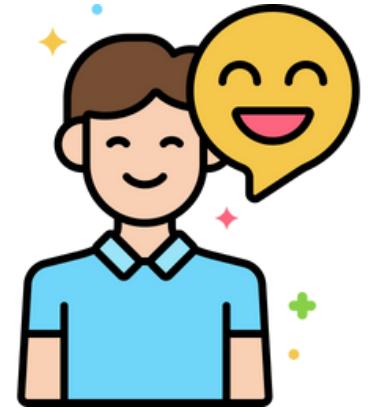
## Cụm 1: Trẻ tuổi - tiềm năng

- Độ tuổi: 24 - 38
- Thu nhập: 101K - 135K
- Điểm trung thành: 960 - 2800



## Cụm 2: Trẻ tuổi - Trung thành

- Độ tuổi: 25 - 42
- Thu nhập: 33K - 68K
- Điểm trung thành: 2,800 - 4,400



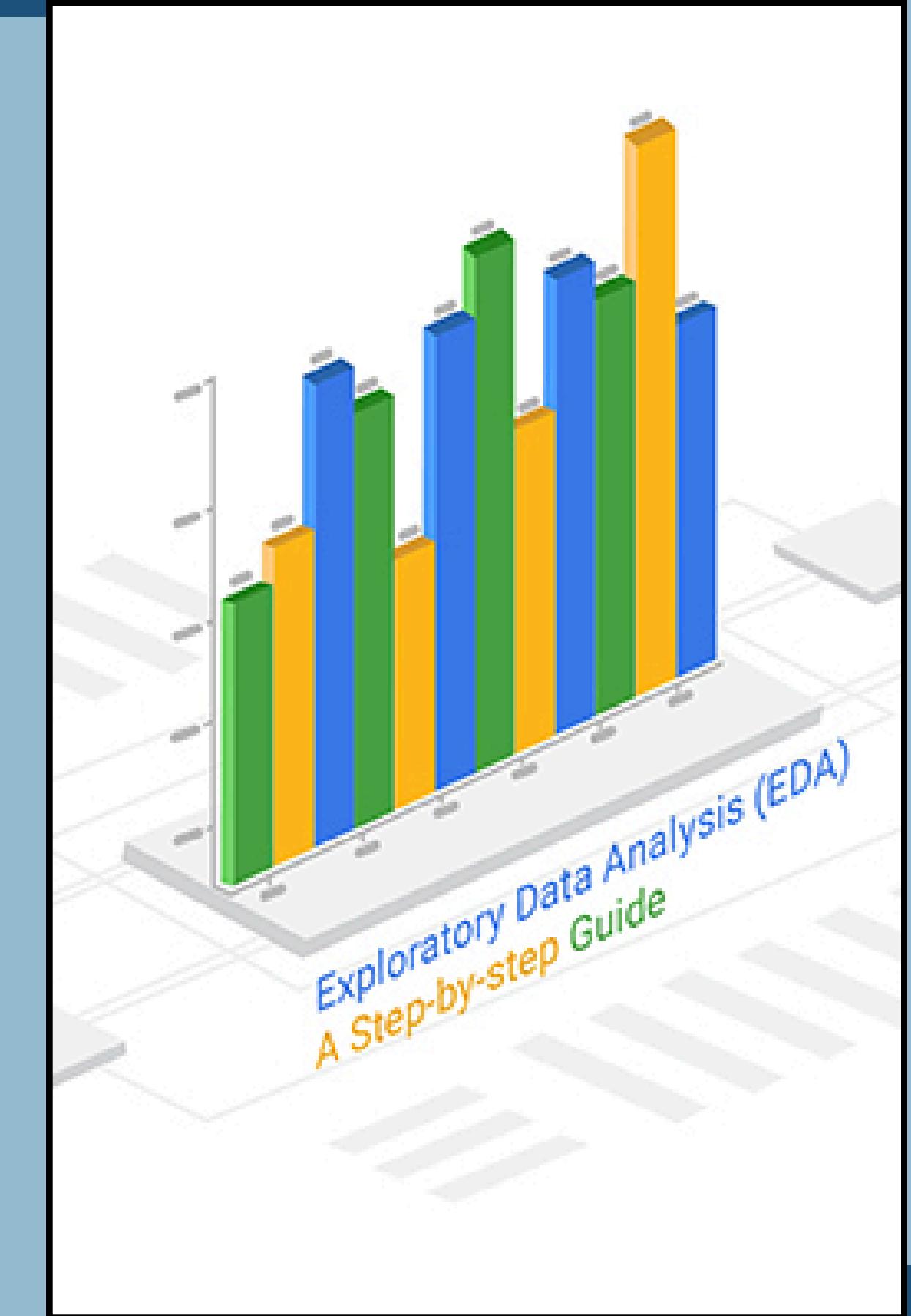
## Cụm 3: VIP - Lâu năm

- Độ tuổi: 50 - 64
- Thu nhập: 83K - 129K
- Điểm trung thành: 2,900 - 4,400

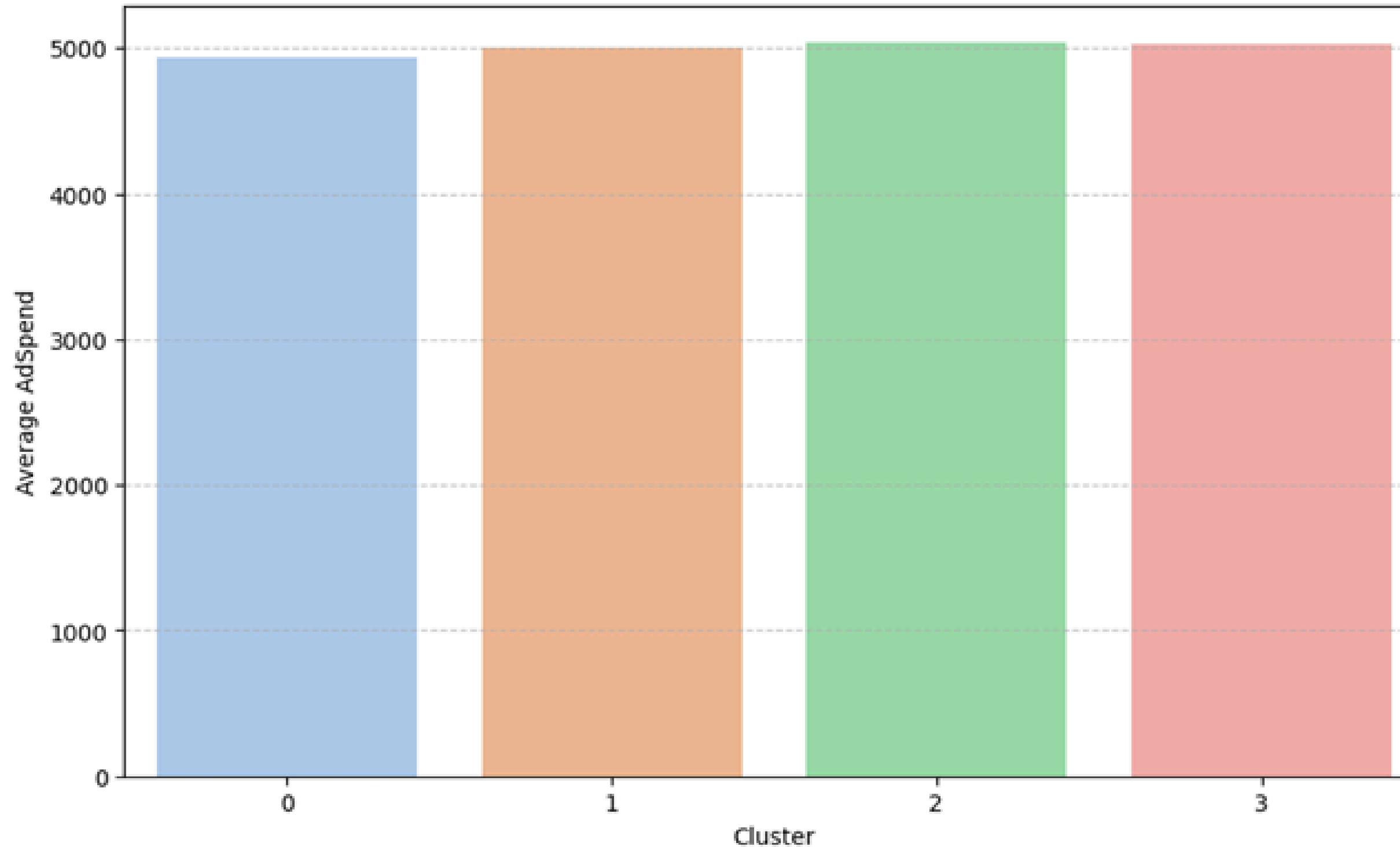


# EDA

## In 4 Clusters

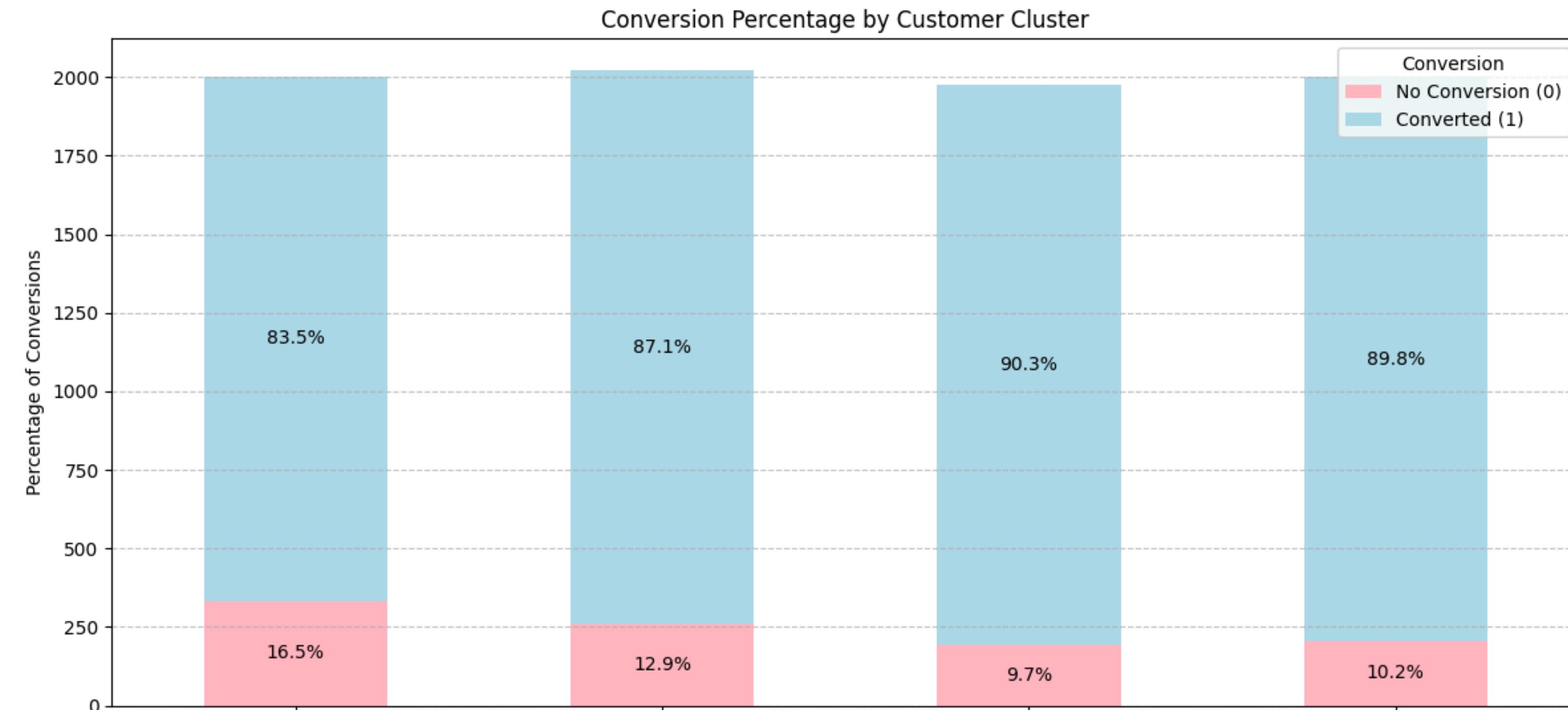


# 1. AVG ADSPEND BY CLUSTER



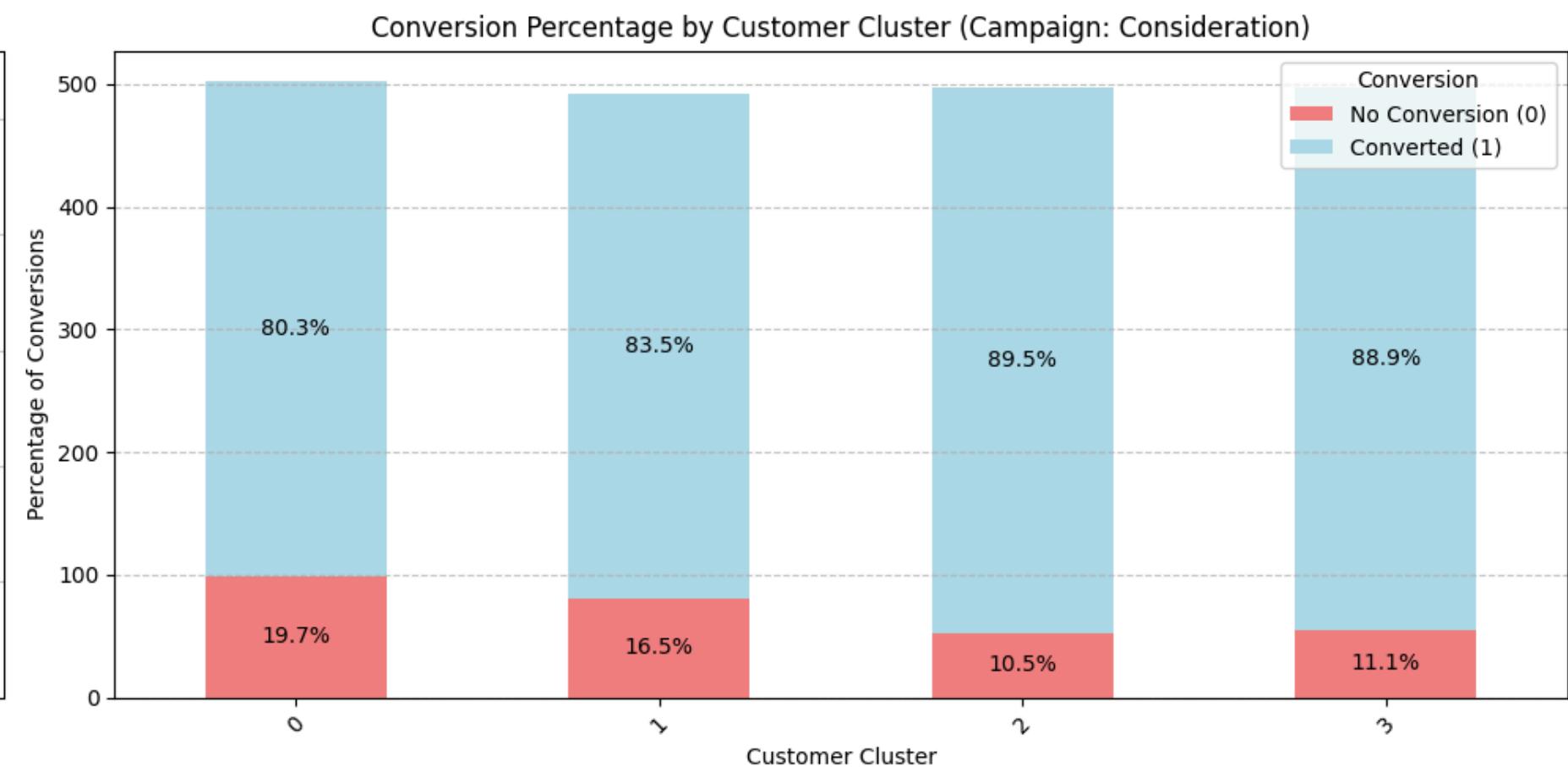
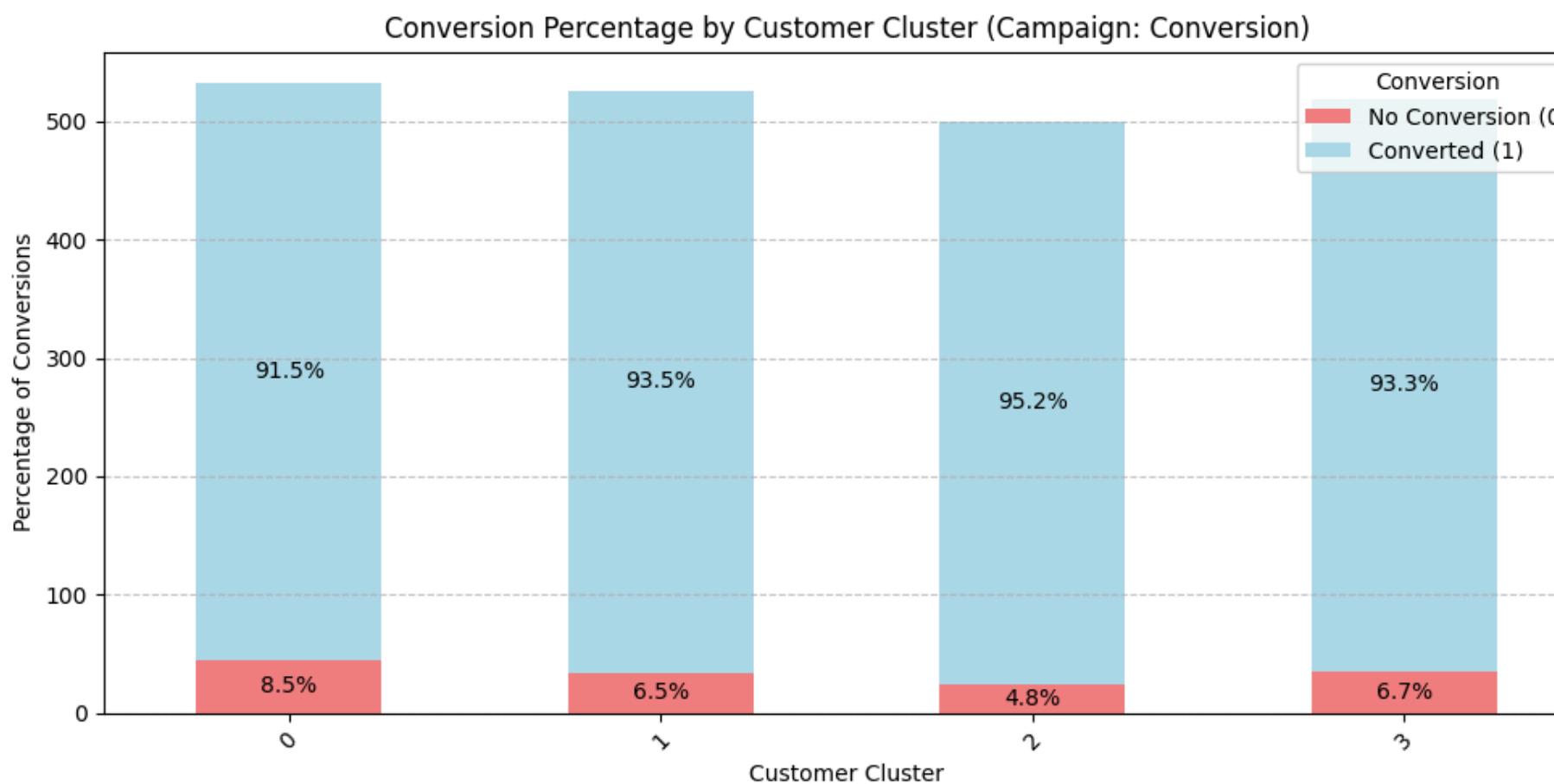
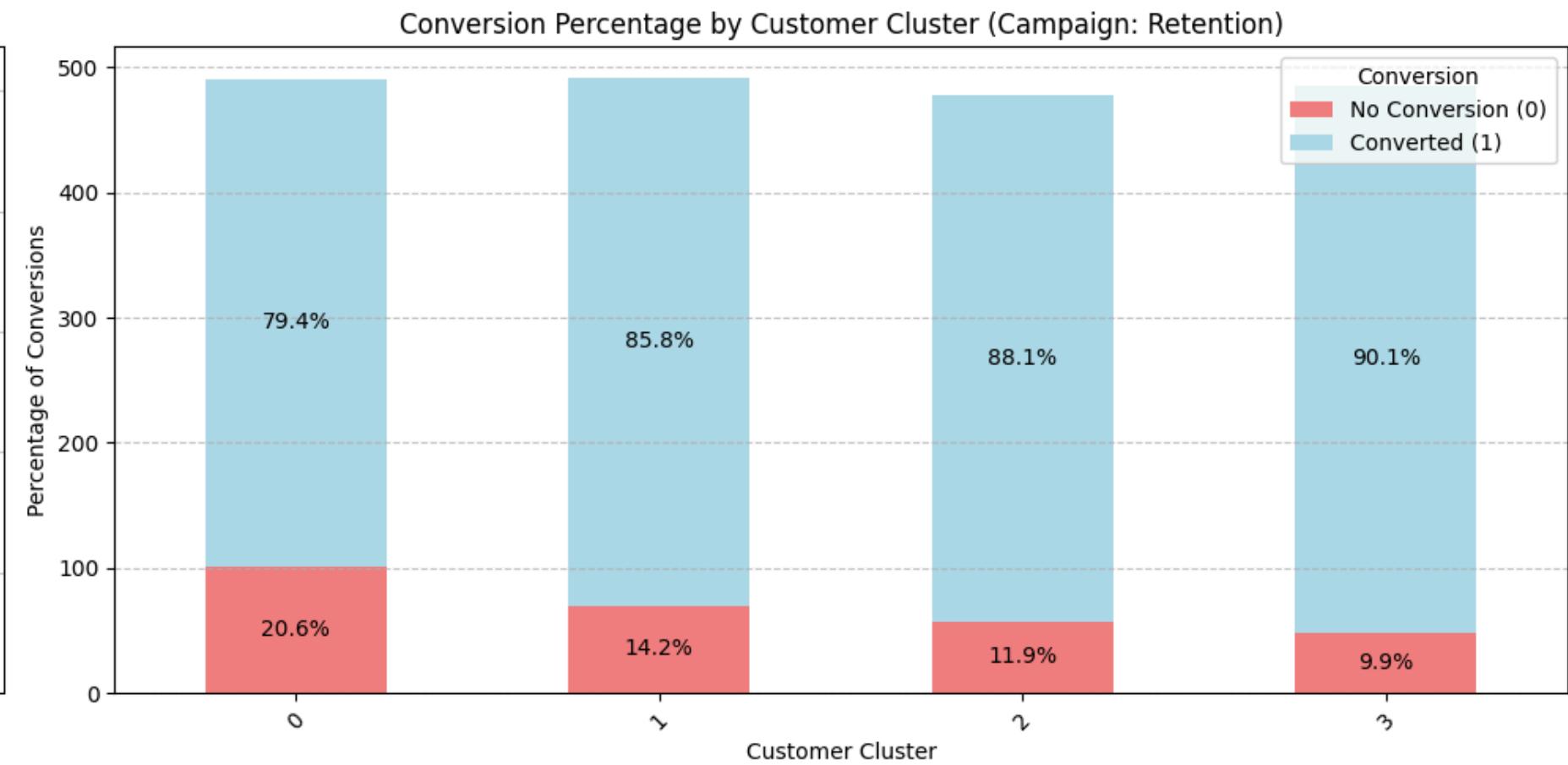
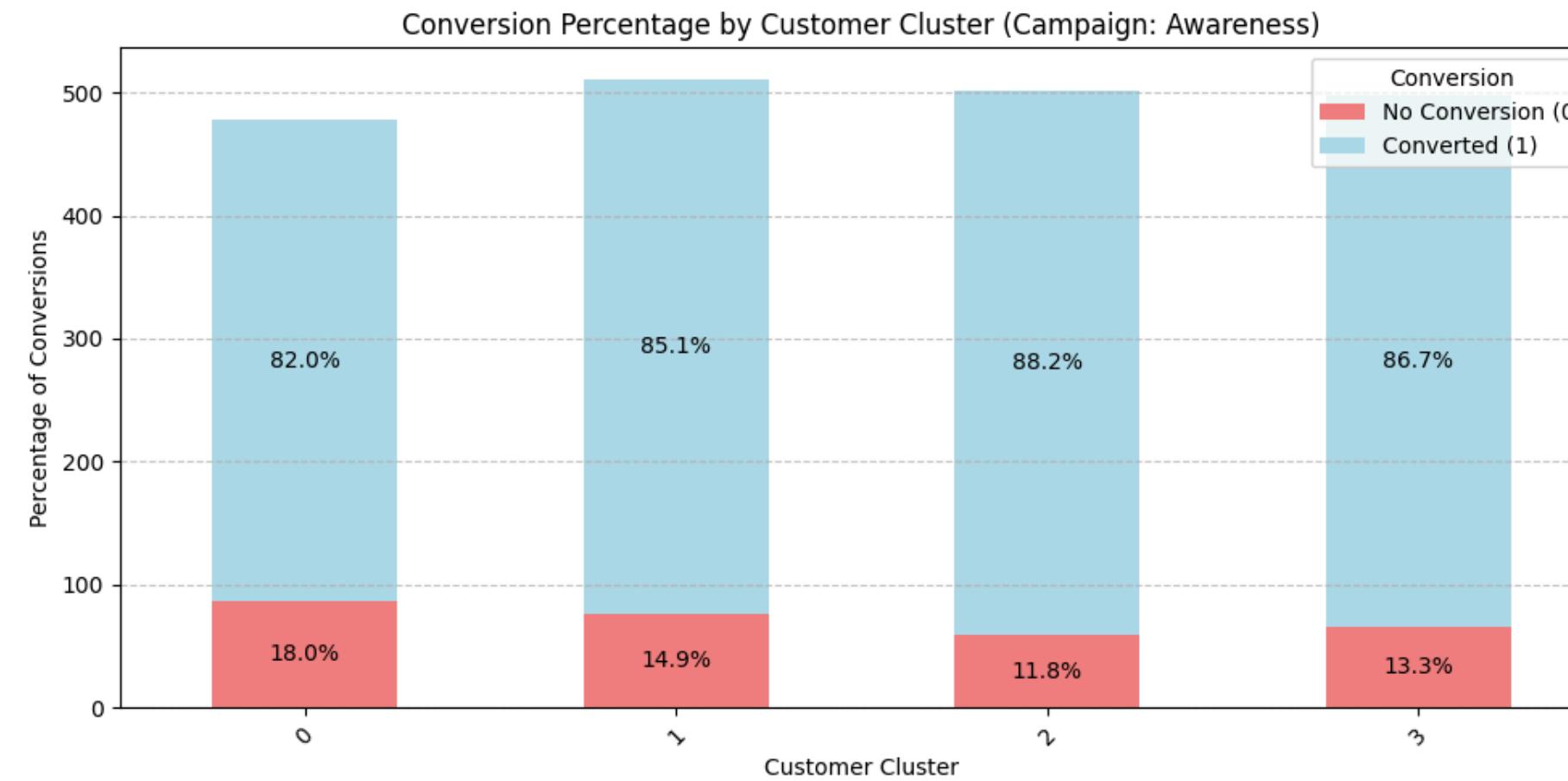
Doanh nghiệp chi tiền quảng  
 cáo cho 4 cụm là **như nhau**

## 2. CONVERSION RATE BY CLUSTER



Tệp khách hàng “VIP - Lâu năm” và “Trẻ tuổi - Trung thành” có tỉ lệ chuyển đổi **cao nhất**  
Trong khi đó, tệp “Truyền thống” có tỉ lệ chuyển đổi **thấp nhất**

# 3. CONVERSION RATE BY CAMPAIGN

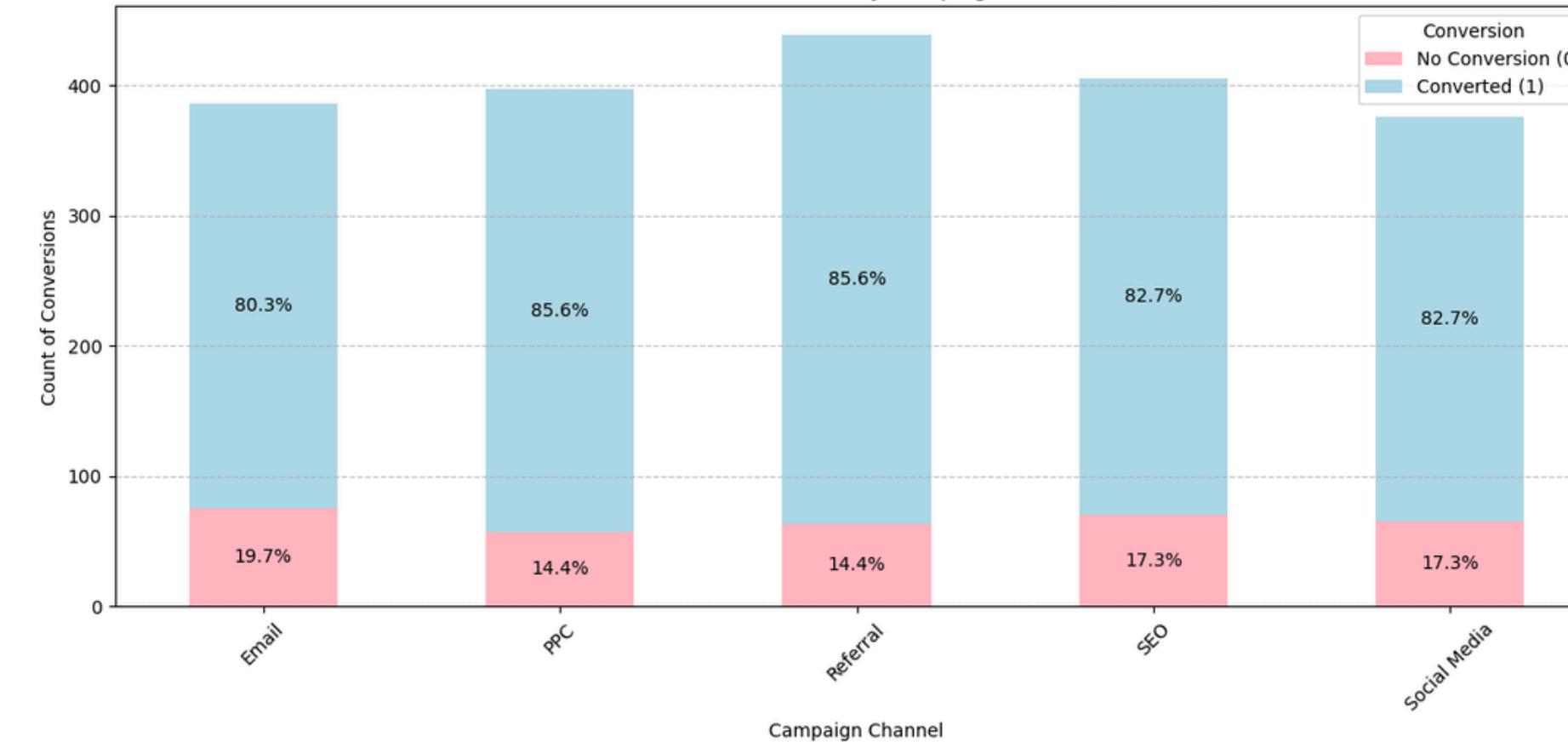


### 3. CONVERSION RATE BY CAMPAIGN

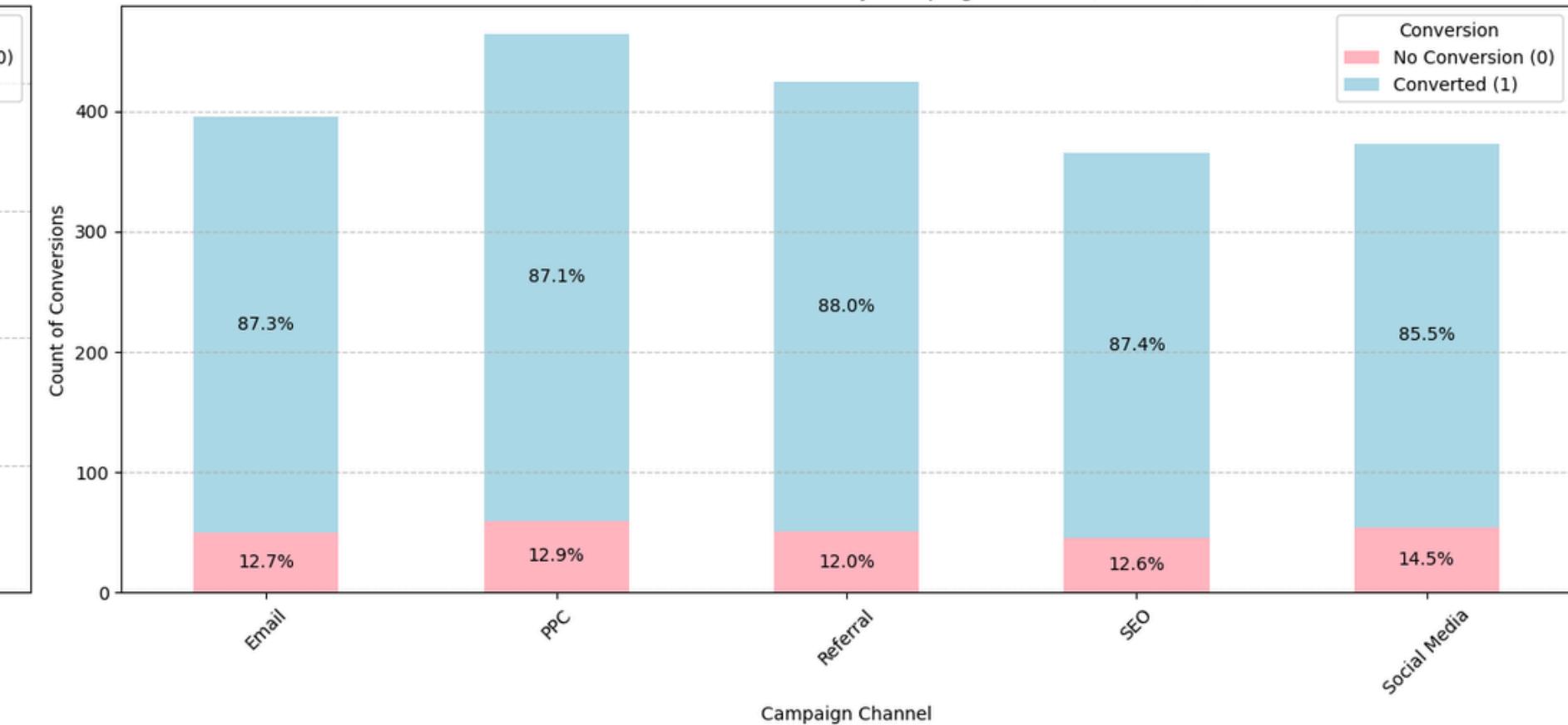
- Chiến dịch Conversion có tỉ lệ chuyển đổi cao nhất, trên 91.5%.
- 3/4 nhóm đều phản ứng tốt với các chiến dịch, đặc biệt là nhóm “VIP - Lâu năm” và “Trẻ tuổi - Trung thành” cho thấy đây là tệp khách hàng có giá trị cao
- Tệp khách hàng “Lớn tuổi - Truyền thống” có tỉ lệ không chuyển đổi cao nhất, với sự chênh lệch rõ rệt so với các nhóm khác, nhất là trong chiến dịch Retention (10%) và Consideration (3,3%).

# 4. CONVERSION RATE BY CHANNEL

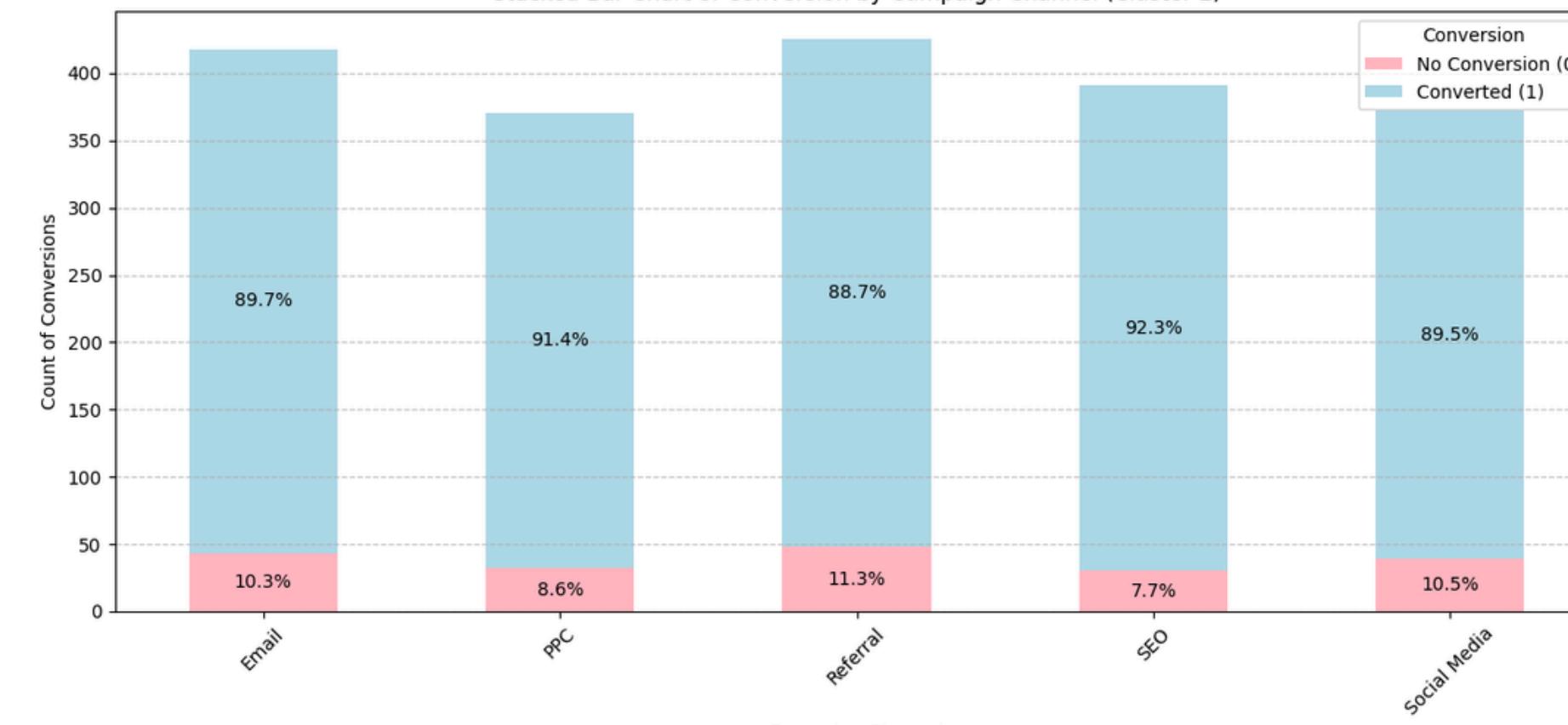
Stacked Bar Chart of Conversion by Campaign Channel (Cluster 0)



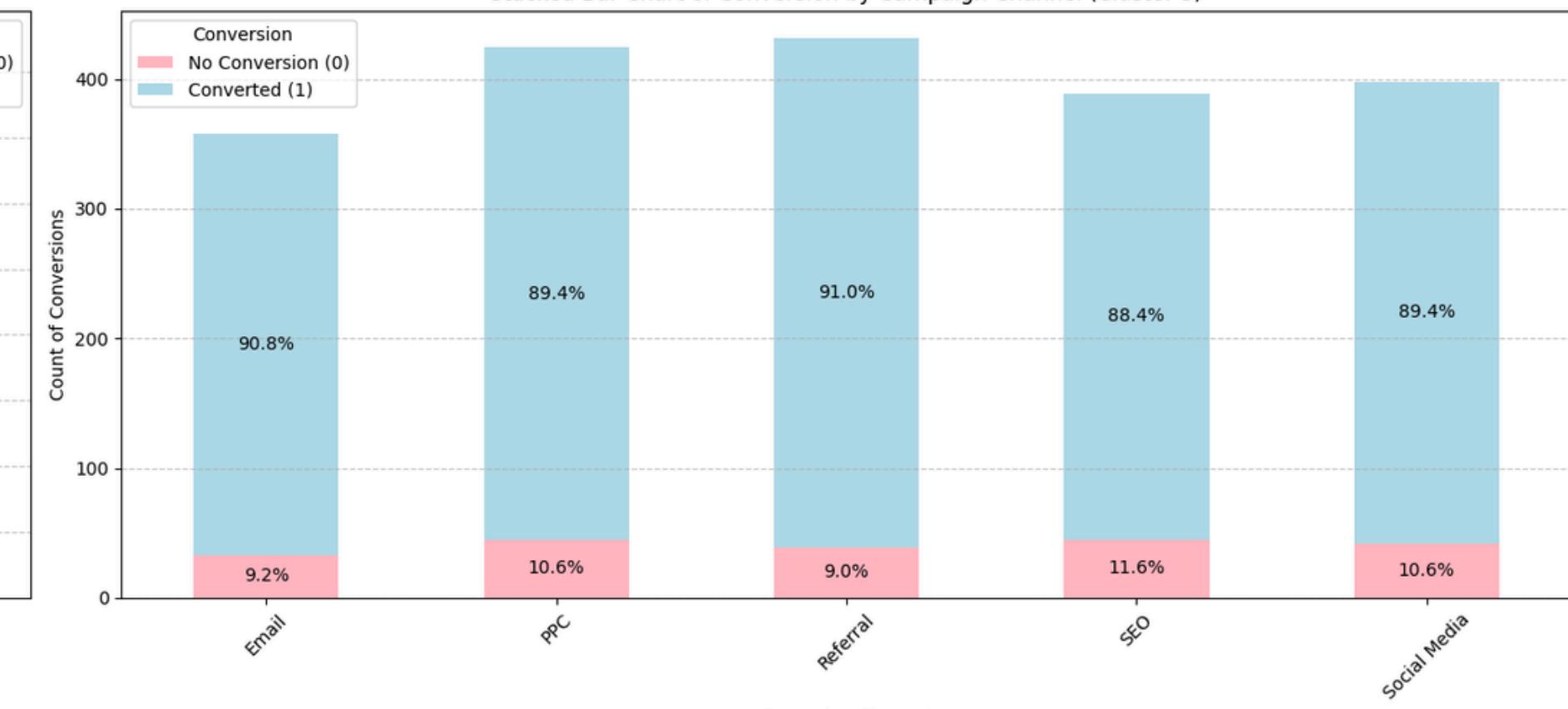
Stacked Bar Chart of Conversion by Campaign Channel (Cluster 1)



Stacked Bar Chart of Conversion by Campaign Channel (Cluster 2)



Stacked Bar Chart of Conversion by Campaign Channel (Cluster 3)



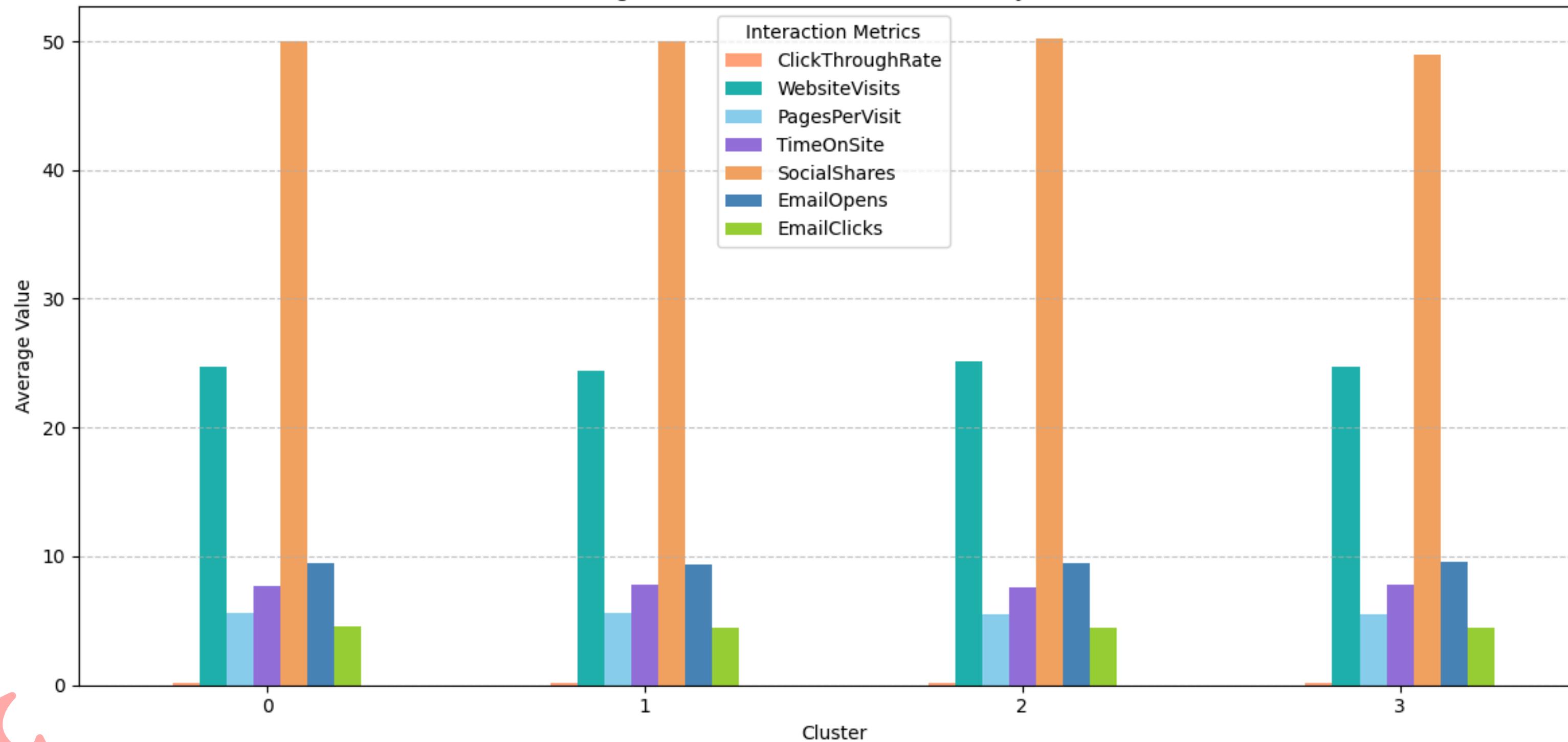
# 4. CONVERSION RATE BY CHANNEL

Các tệp khách hàng **đều phản ứng** như nhau với các kênh tiếp thị, không có sự khác biệt lớn trong hiệu quả chuyển đổi của các kênh.



Tuy nhiên các khách hàng thuộc cụm “**Lớn tuổi - Truyền thống**” có **xu hướng chuyển đổi cao hơn** khi tiếp xúc với kênh **PPC và Referral**.

# 5. CLUSTER'S INTERACTION



GẦN NHƯ KHÔNG CÓ SỰ KHÁC BIỆT LỚN TRONG HÀNH VI GIỮA CÁC CỤM

# INSIGHTS & RECOMMENDATION

01

Nhóm "Lớn tuổi - Truyền thống" có tỷ lệ chuyển đổi **thấp** và **ít quan tâm** đến các chiến dịch marketing.

=> Doanh nghiệp nên **cân nhắc cắt giảm chi phí quảng cáo** cho nhóm này và dành tài nguyên cho kênh **PPC, Referral**

02

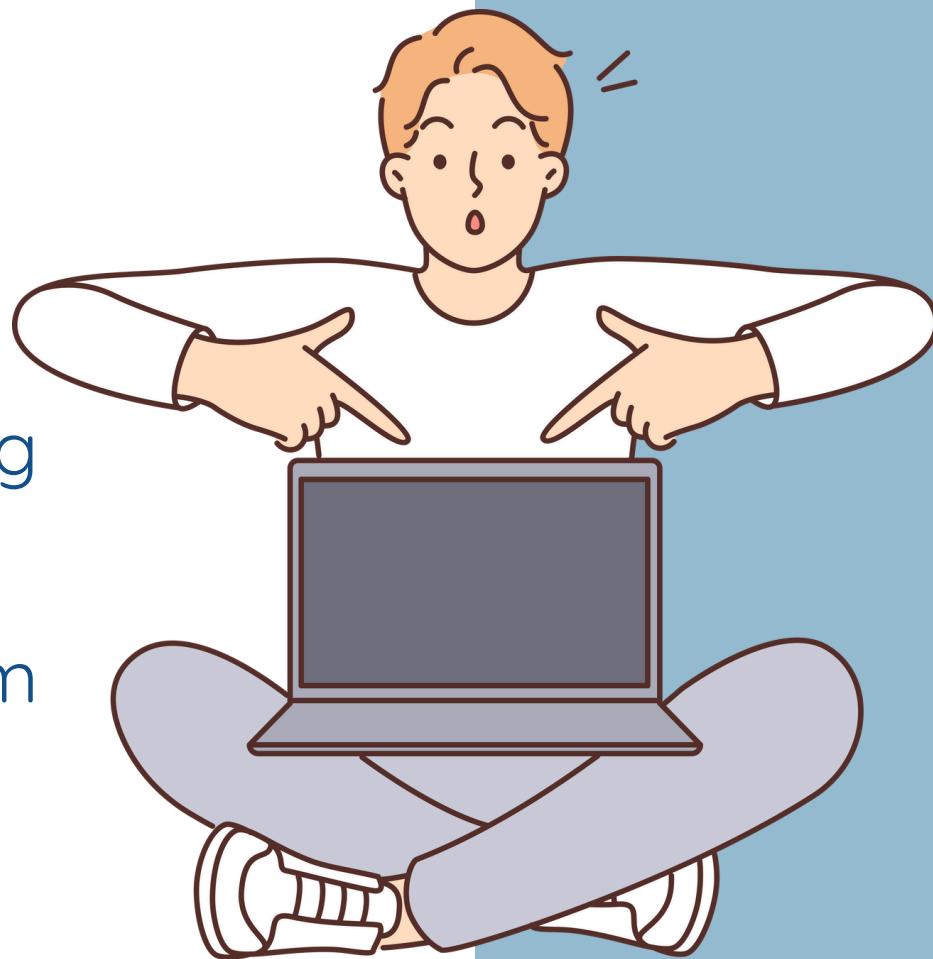
Nhóm "VIP - Lâu năm" & "Trẻ trung - trung thành" là nhóm khách hàng **có giá trị nhất**, mang lại doanh thu cao và có lòng trung thành.

=> Doanh nghiệp **cần tiếp tục duy trì** các chiến dịch đang **hiệu quả** nhằm giữ mối quan hệ tốt với nhóm này.

03

Nhóm "Trẻ tuổi - Tiềm năng" có **tiềm năng tăng trưởng cao**.

=> Cần những chiến dịch **ưu đãi, giảm giá** nhằm thu hút, chuyển đổi sự tiềm năng thành **lợi nhuận to lớn** cho doanh nghiệp trong tương lai





GROUP 4

# THANK YOU

● for your nice attention

