

# Dự đoán giá và phân cụm ô tô cũ

Phan Thanh Tâm  
Huỳnh Văn Lộc  
Võ Đình Huy

# Tổng quan Bài thuyết trình

## CÁC CHỦ ĐỀ CHÍNH

- Hiệu chỉnh mô hình dự đoán
- Phân cụm xe

# I. Hiệu chỉnh mô hình

## 1. Chuẩn hoá biến mục tiêu

Giá

1050.0

2089.0

295.0

Giá

0.569986

2.069264

-0.519481



# I. Hiệu chỉnh mô hình

## 2. Crawl data test

- Crawl trên web cũ: Bonbanh.com
- Crawl đc gần 3k mẫu mới (25/05/2024)

# I. Hiệu chỉnh mô hình

## 3. So sánh kết quả

Mô hình Linear Regression không có sự thay đổi lớn

Thông số đánh giá	Tập Validate	Tập Test
RMSE (triệu đồng)	751.468	852.050
R2 Score	0.727	0.667

# I. Hiệu chỉnh mô hình

## 3. So sánh kết quả

Mô hình Gradient Boosting trước khi hiệu chỉnh

Thông số đánh giá	Tập Validate	Tập Test
RMSE (triệu đồng)	398.140	631.628
R2 Score	0.923	0.817

# I. Hiệu chỉnh mô hình

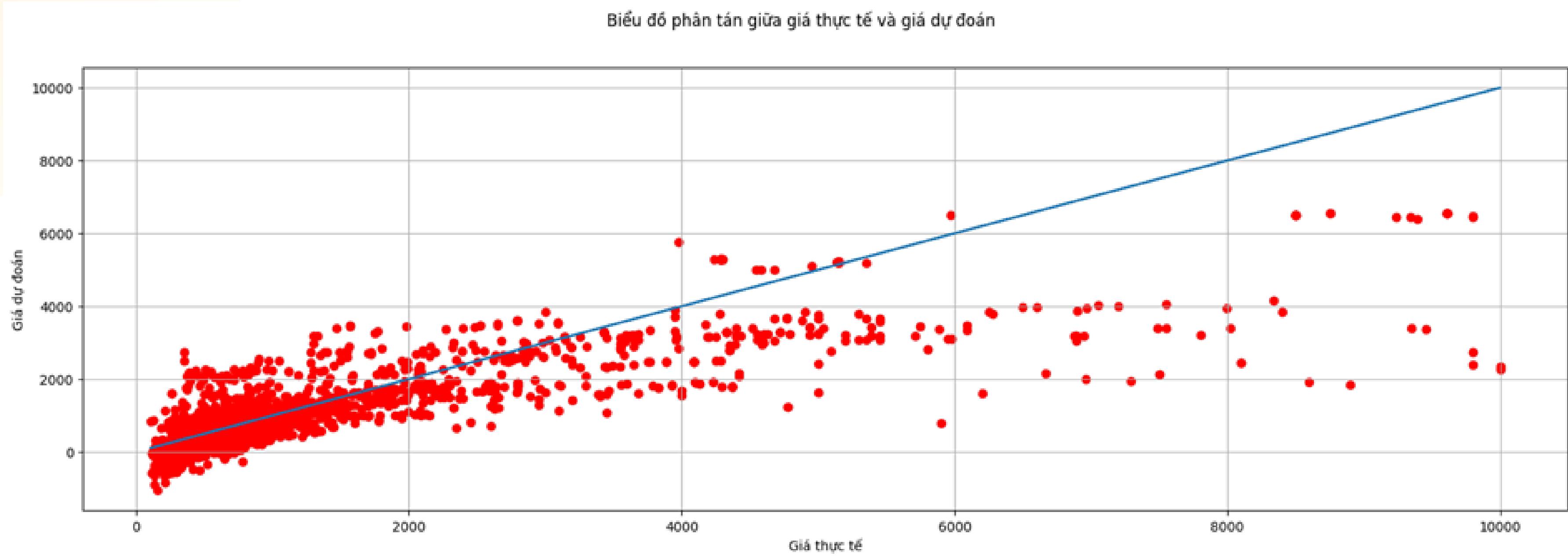
## 3. So sánh kết quả

Mô hình Gradient Boosting sau khi hiệu chỉnh

Thông số đánh giá	Tập Validate	Tập Test
RMSE (triệu đồng)	243.999	637.516
R2 Score	0.971	0.814

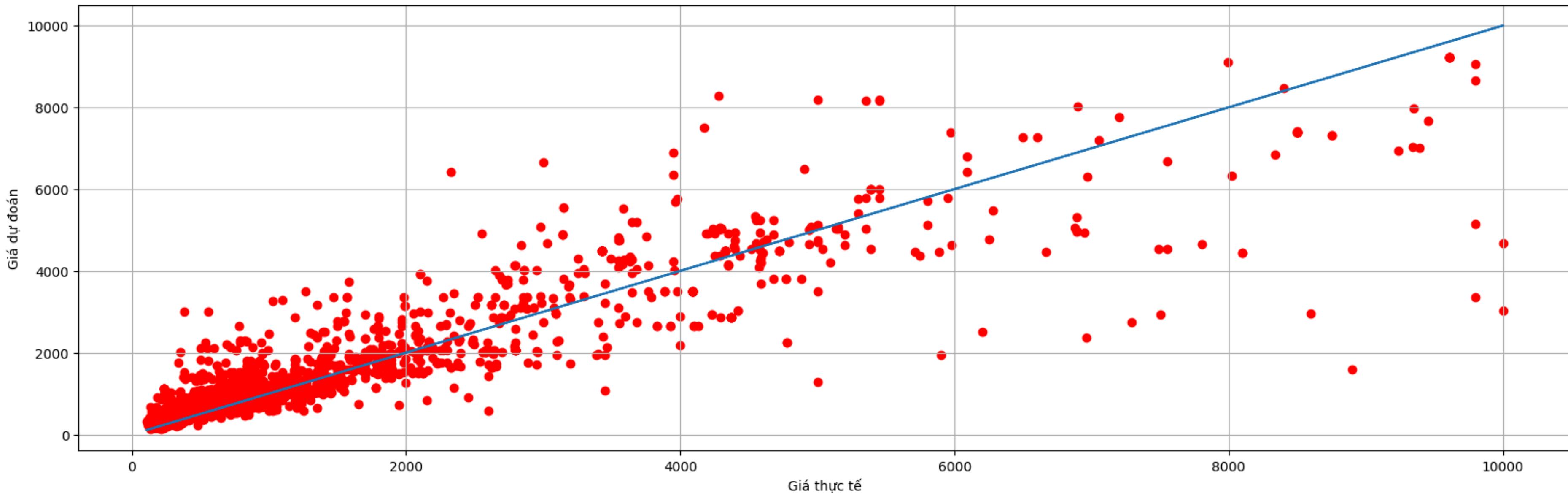
# ĐỒ THỊ LINEAR REGRESSION

Biểu đồ phân tán giữa giá thực tế và giá dự đoán



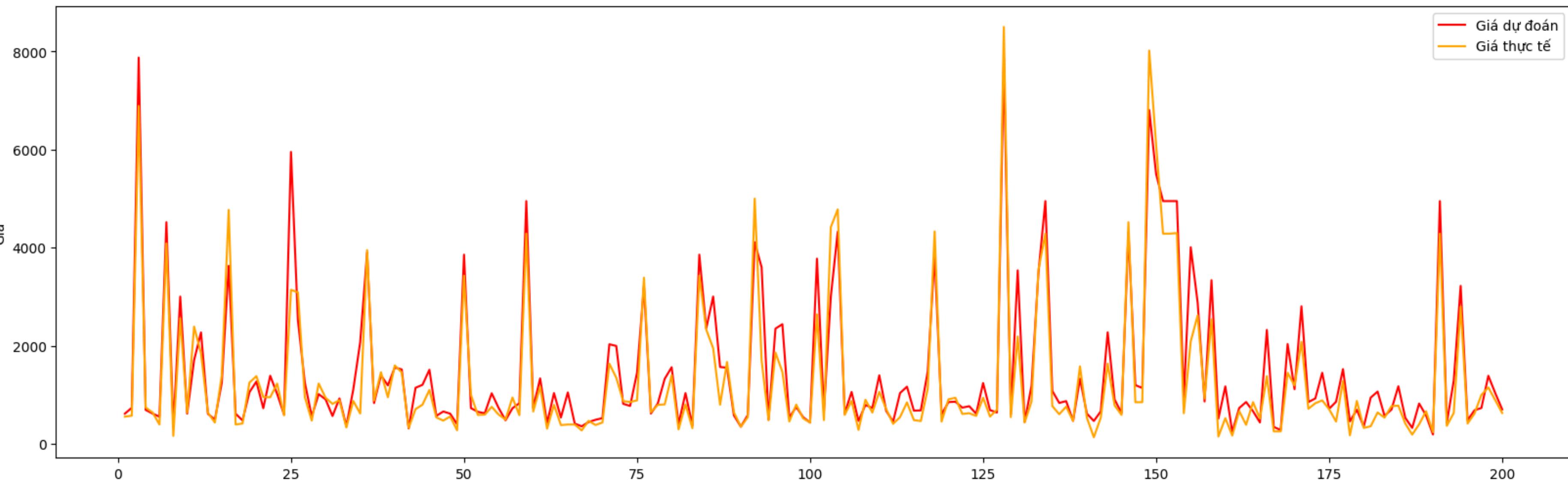
# ĐỒ THỊ GRADIENT BOOSTING

Biểu đồ phân tán giữa giá thực tế và giá dự đoán



# ĐỒ THỊ

Giá dự đoán với giá thực tế



# Phân cụm xe

## 2. Phân cụm

### 1.1 Chuẩn hóa và mã hóa dữ liệu

- Áp dụng dữ liệu thu thập được phần trước tiếp tục thực hiện mã hóa biến danh mục bằng LabelEncoder và chuẩn hóa biến ‘Số Km đã đi’ và ‘Giá’ bằng RobustScaler

	Năm sản xuất	Tình trạng	Kiểu dáng	Số Km đã đi	Xuất xứ	Hãng xe	Dẫn động	Số chỗ ngồi	Địa điểm	Động cơ	Hộp số	Giá
0	2022	1	5	-0.134459	1	18	3	7	17	5	1	0.569986
1	2024	0	5	-0.537837	1	46	3	7	17	45	1	2.069264
2	2007	1	6	0.513257	1	69	2	5	46	53	1	-0.519481
3	2024	0	5	-0.537837	1	46	1	7	17	59	1	6.096681
4	2023	0	5	-0.537837	1	64	2	7	45	28	0	-0.297258

## 2. Phân cụm

### 1.2. Giảm chiều dữ liệu

- Để trực quan hóa và nhìn nhận tính chất phân cụm của dữ liệu. Ta lần lượt kiểm tra 2 phương pháp: PCA và t-sne để lựa chọn phương pháp tối ưu để giảm chiều dữ liệu

Ta tiến hành so sánh trên 2 phương diện

- Đo thời gian thực thi của việc giảm chiều dữ liệu từ 12 đặc trưng về 2 đặc trưng đối với cả 2 phương pháp
- So sánh sự biến đổi của cấu trúc, phân bố khi có sự biến động trên tập gốc

Thời gian thực thi của PCA: 0.033 giây

Thời gian thực thi của t-SNE: 670.072 giây

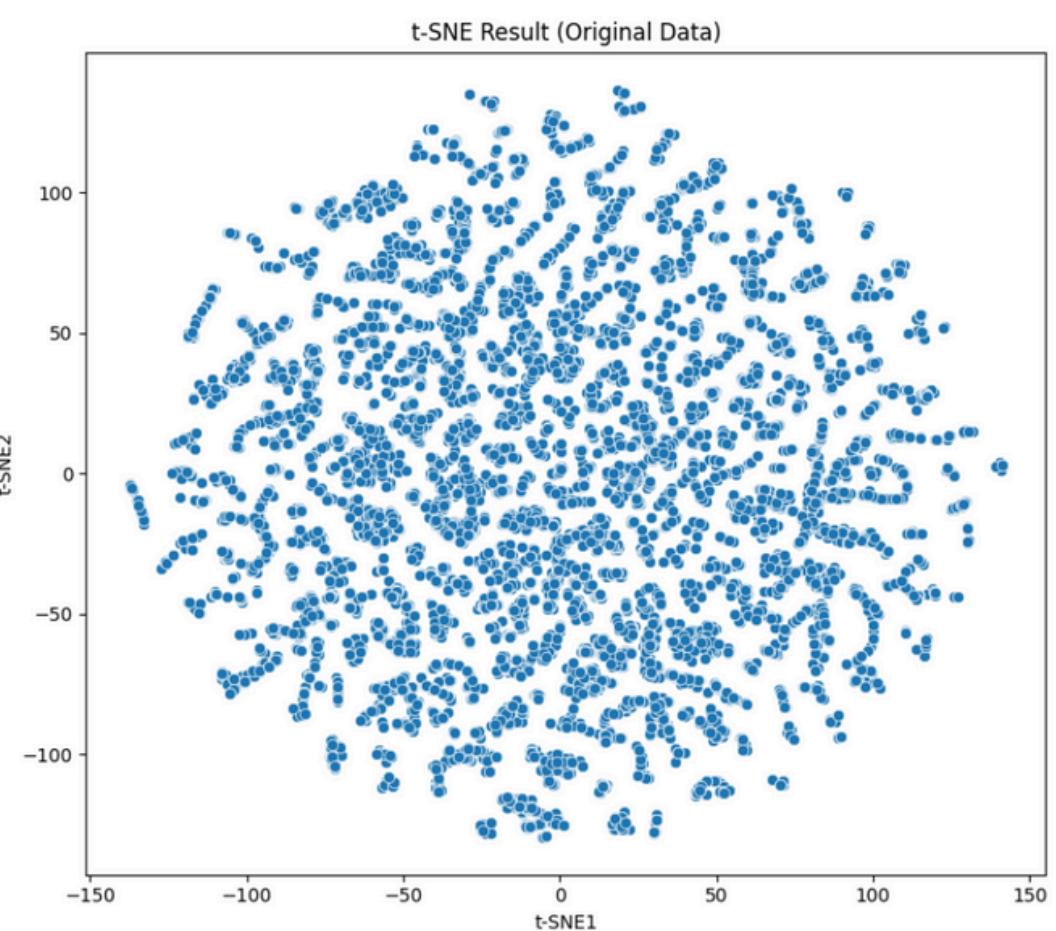
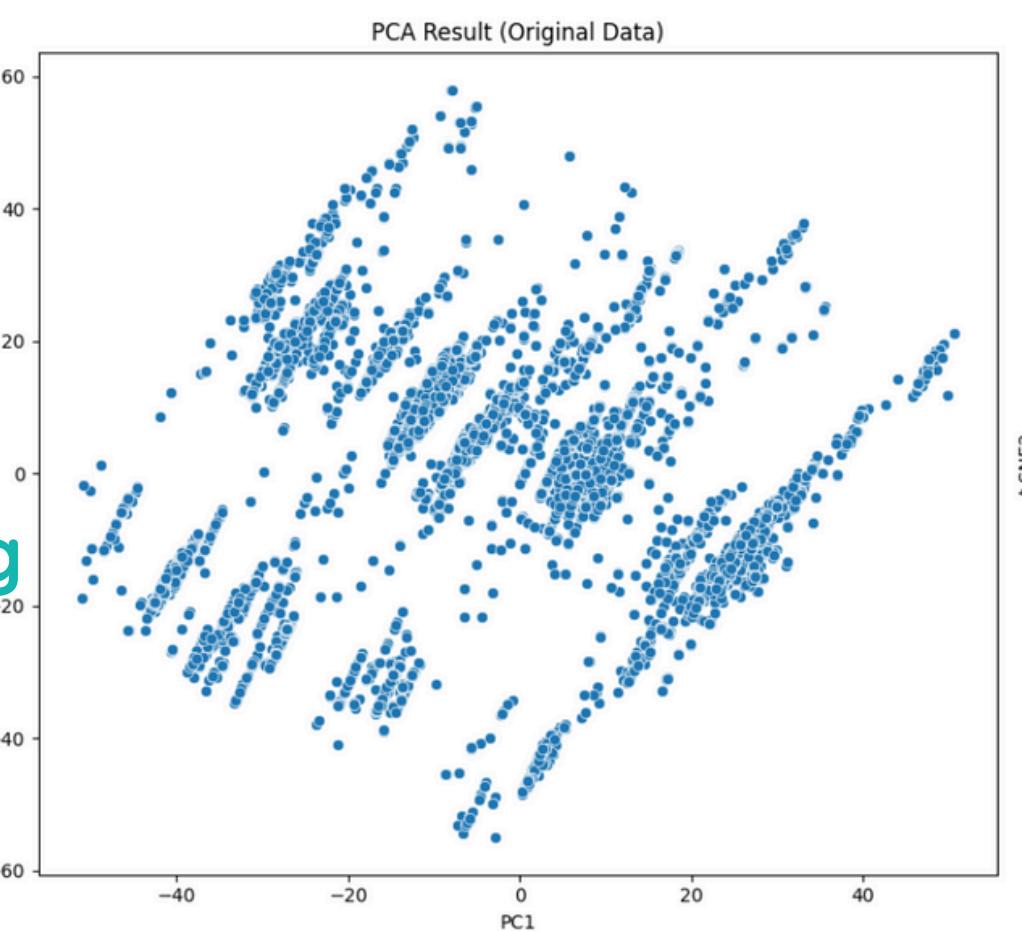
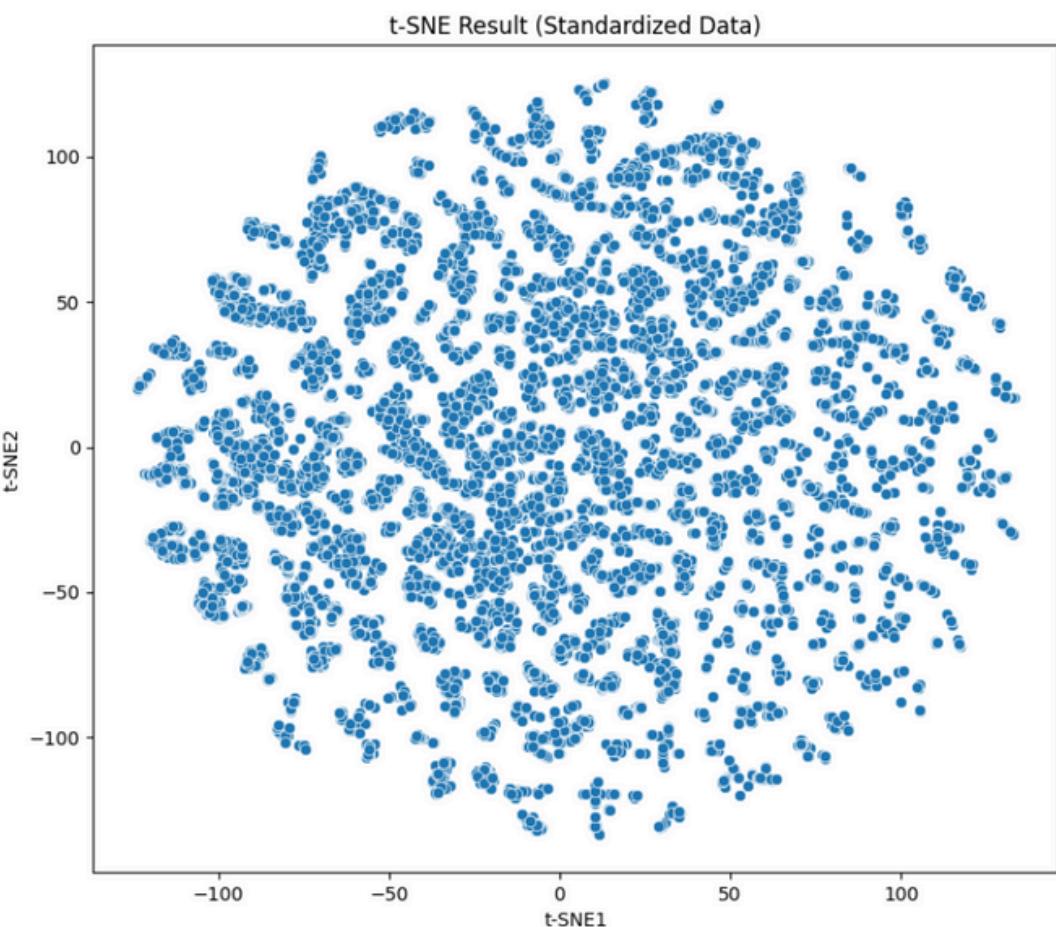
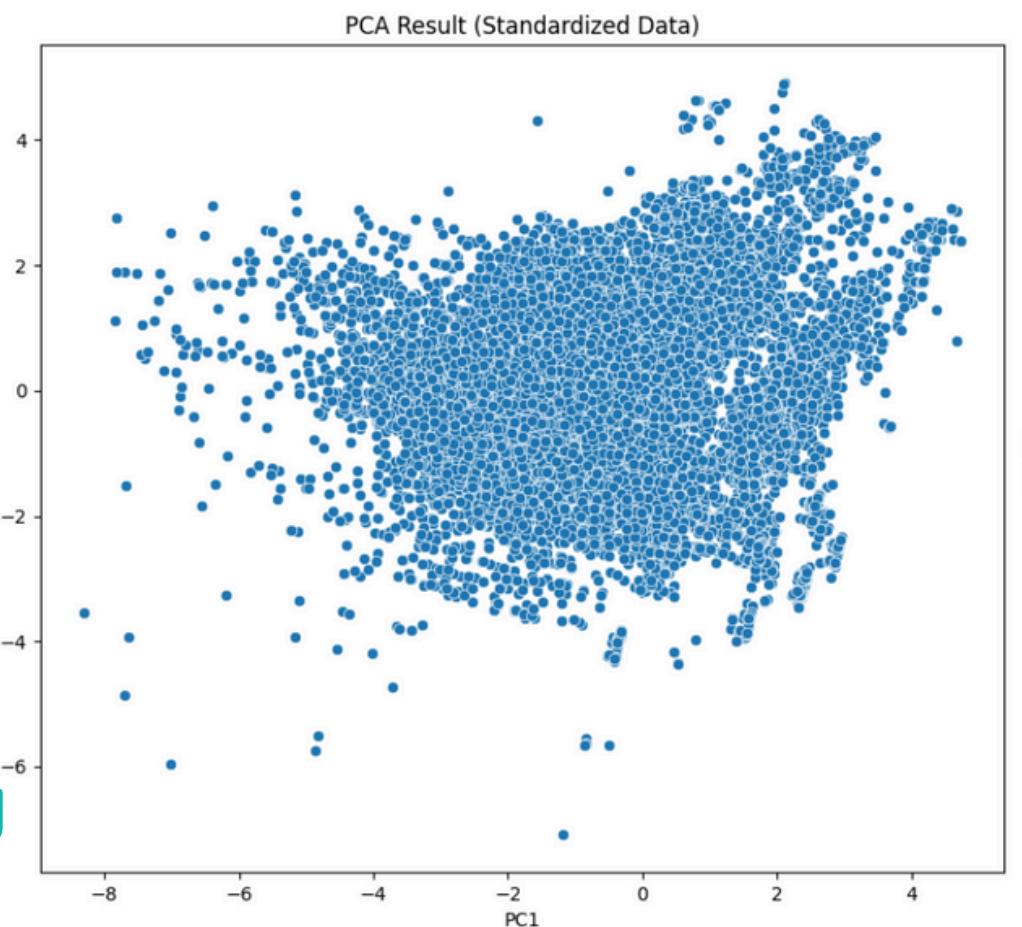
# 2. Phân cụm

## 1.2. Giảm chiều dữ liệu

+ Phương pháp PCA có thời gian thực thi cực kì nhanh hơn vượt trội so với phương pháp t-sne

+ Phương pháp PCA lại bị ảnh hưởng, biến đổi lớn về cấu trúc và phân bố khi dữ liệu gốc có sự thay đổi còn t-sne vẫn còn giữ lại được, dữ liệu ít bị mất mát hơn

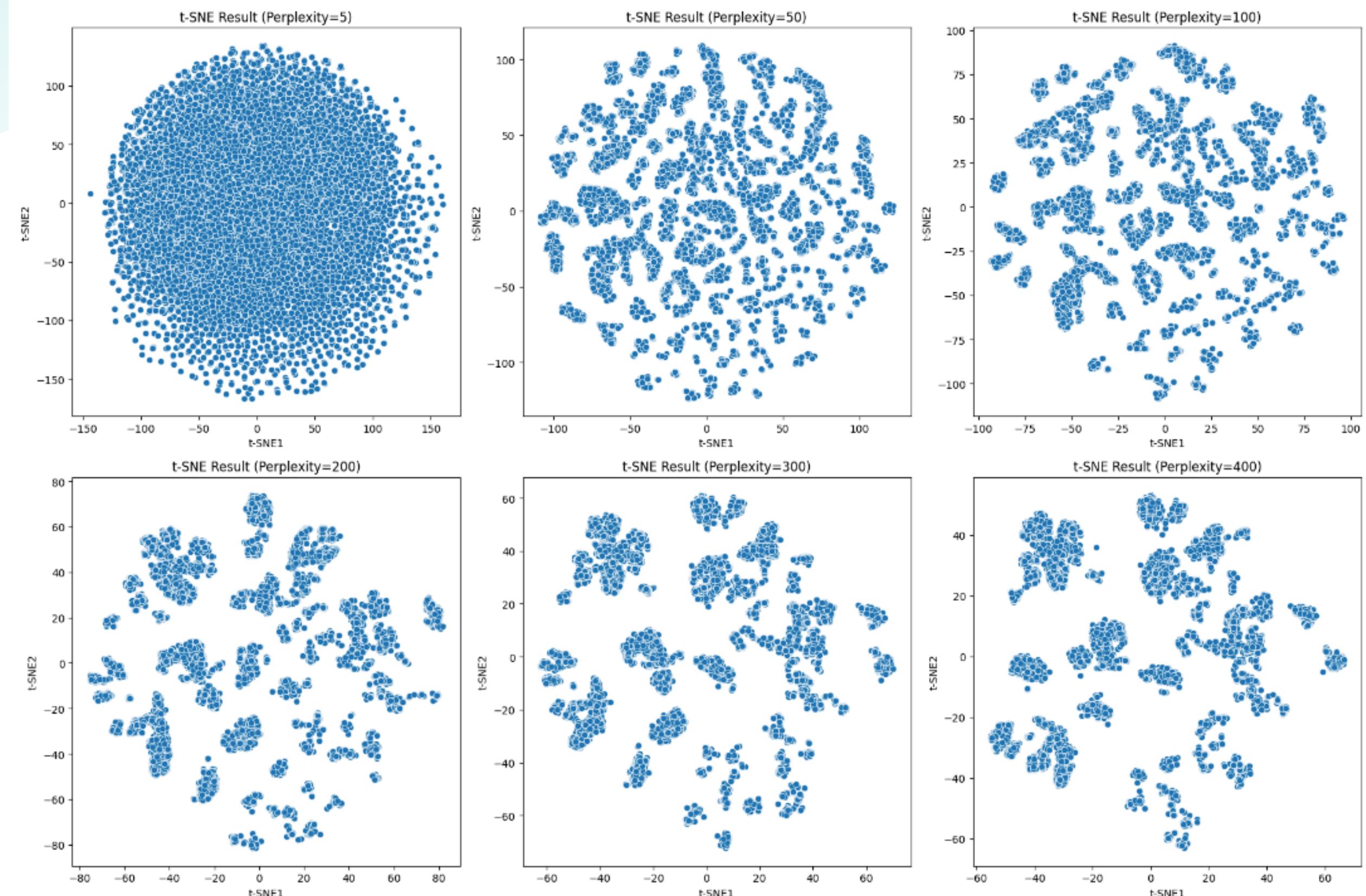
=> Lựa chọn t-sne để giảm chiều và trực quan hóa tuy lâu về mặc thời gian nhưng độ chính xác cao và mất mát dữ liệu thấp



# 2. Phân cụm

## 1.2. Giảm chiều dữ liệu

- **Tiến hành điều chỉnh tham số để phân bố dữ liệu thể hiện tính chất cụm nhất đồng thời hạn chế sự sai lệch và mất mát dữ liệu**
- **Thay đổi giá trị tham số perplexity chạy từ khoảng 5 đến 400 để trực quan hóa**



## 2. Phân cụm

### 2. K means

- K-means là một thuật toán phân cụm (clustering algorithm) phổ biến trong học máy.
- Mục tiêu của thuật toán K-means là phân chia dữ liệu thành K cụm (clusters) sao cho khoảng cách từ mỗi điểm dữ liệu đến trung tâm cụm của nó là nhỏ nhất.

## 2. Phân cụm

### 2. K means

Bước 1: Khởi tạo

- Chọn ngẫu nhiên  $K$  điểm làm trung tâm cụm ban đầu (centroids).

$$\mu_1, \mu_2, \dots, \mu_k.$$

Bước 2: Gán cụm

- Gán mỗi điểm dữ liệu vào cụm có trung tâm gần nhất.

$$c_i = \arg \min_j \|\mathbf{x}_i - \mu_j\|_2^2$$

Bước 3: Cập nhật trung tâm cụm

- Tính toán lại vị trí trung tâm của mỗi cụm dựa trên các điểm dữ liệu đã gán.

$$\mu_j := \frac{\sum_{i=1}^n \mathbf{1}(c_i = j) \mathbf{x}_i}{\sum_{i=1}^n \mathbf{1}(c_i = j)}$$

Bước 4: Lặp lại

- Lặp lại bước 2 và 3 cho đến khi trung tâm cụm không thay đổi hoặc thay đổi rất ít.

## 2. Phân cụm

### 2. KMeans

#### Ưu điểm

- Đơn giản và dễ hiểu.
- Hiệu quả và nhanh chóng cho các tập dữ liệu lớn.
- Dễ dàng triển khai và áp dụng vào các bài toán thực tế.

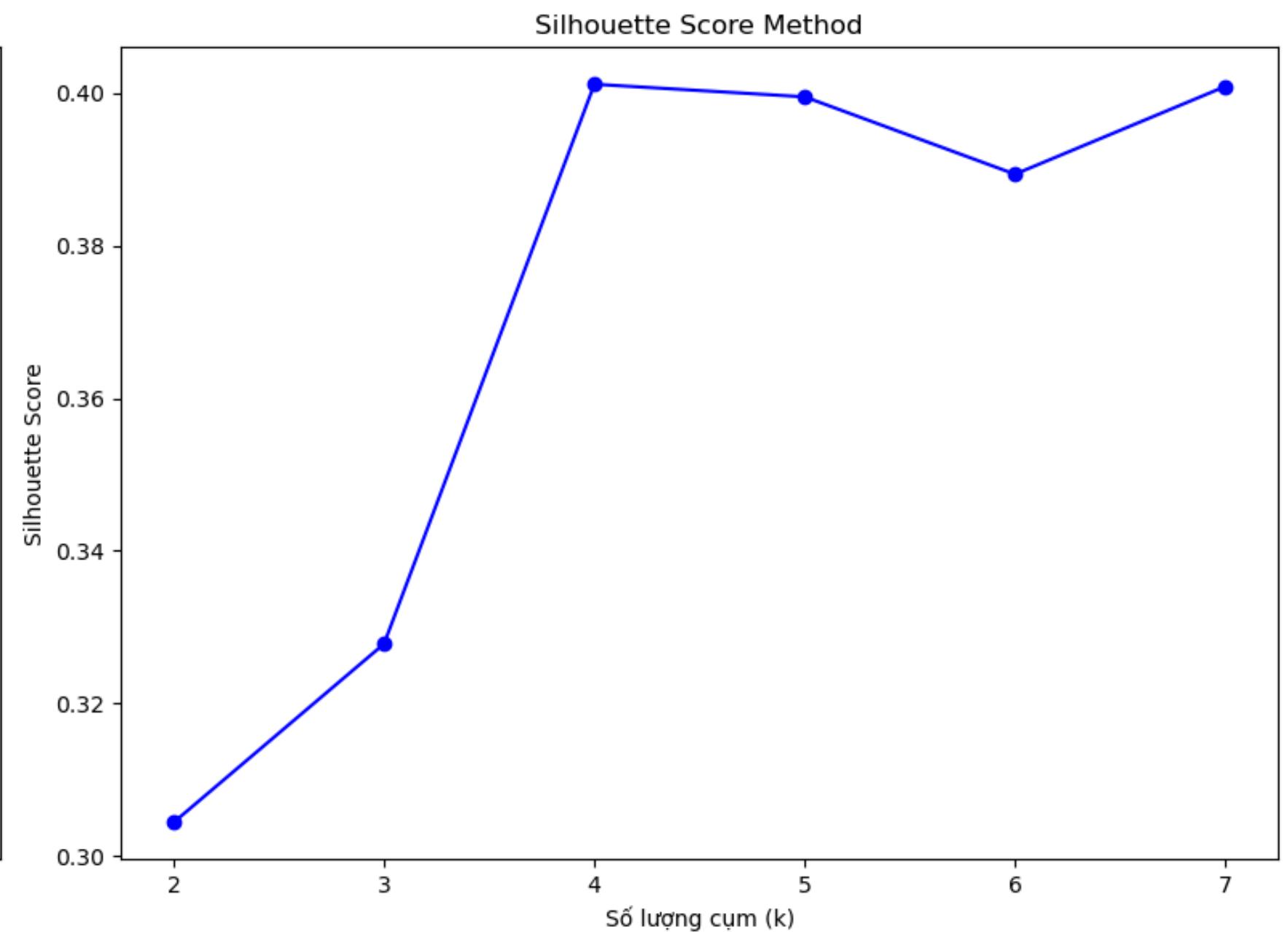
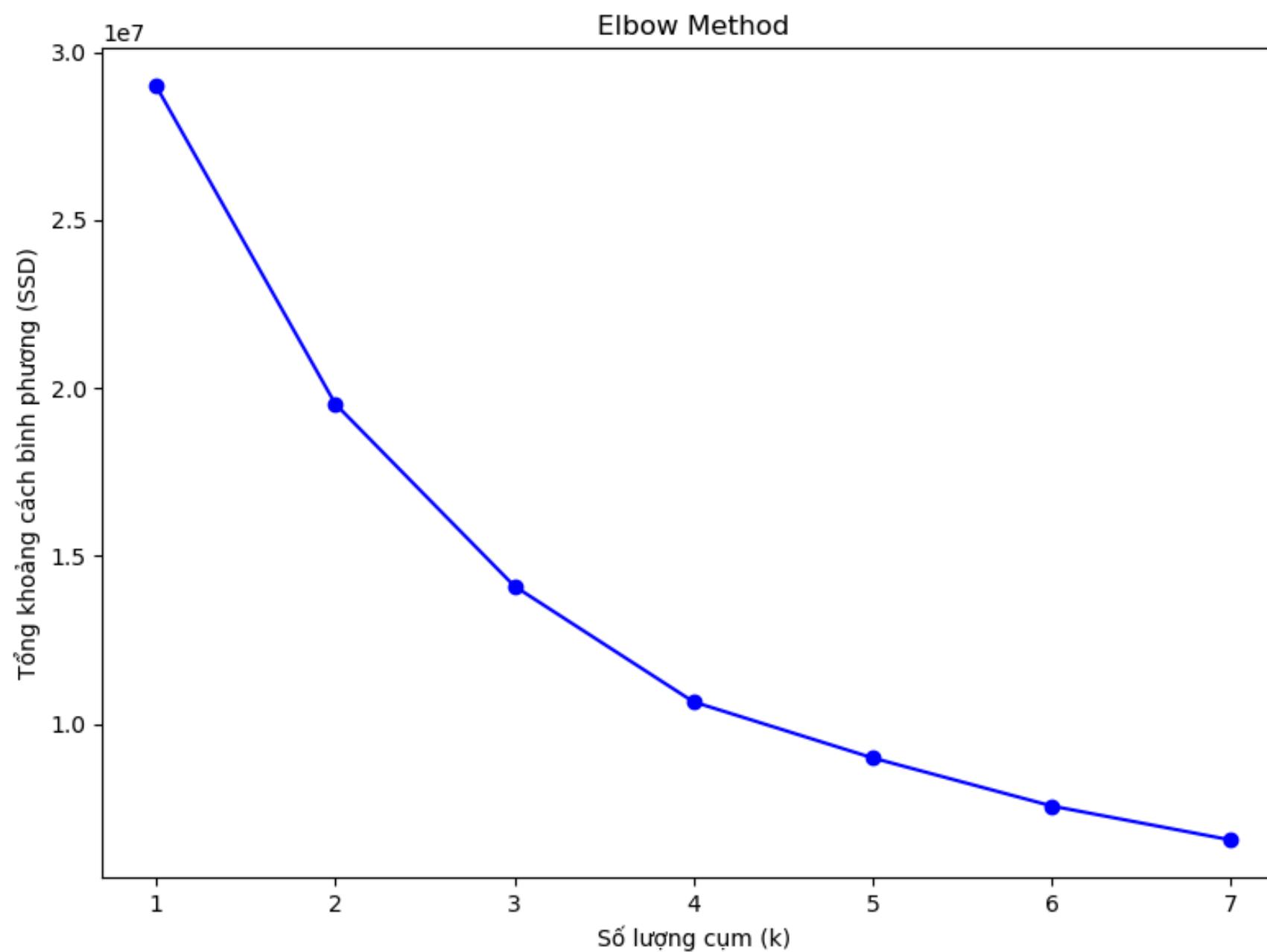
#### Nhược điểm

- Nhạy cảm với giá trị khởi tạo trung tâm cụm ban đầu.
- Không đảm bảo tìm được cụm tối ưu toàn cục (global optimum).
- Phải xác định trước số lượng cụm K.
- Nhạy cảm với outliers và dữ liệu nhiễu.



## 2. KMeans

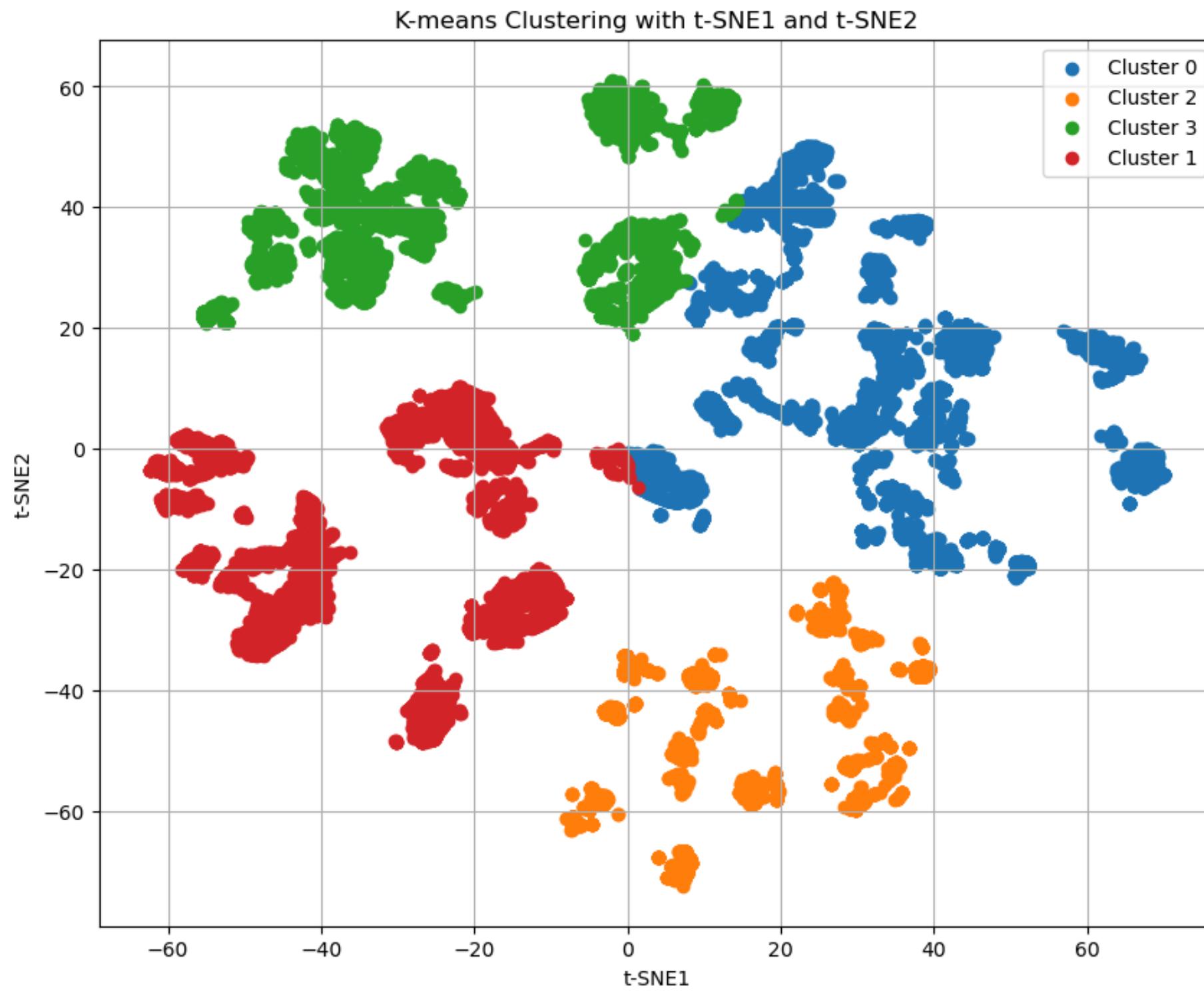
### a. Xác định giá trị K



$k = 4$

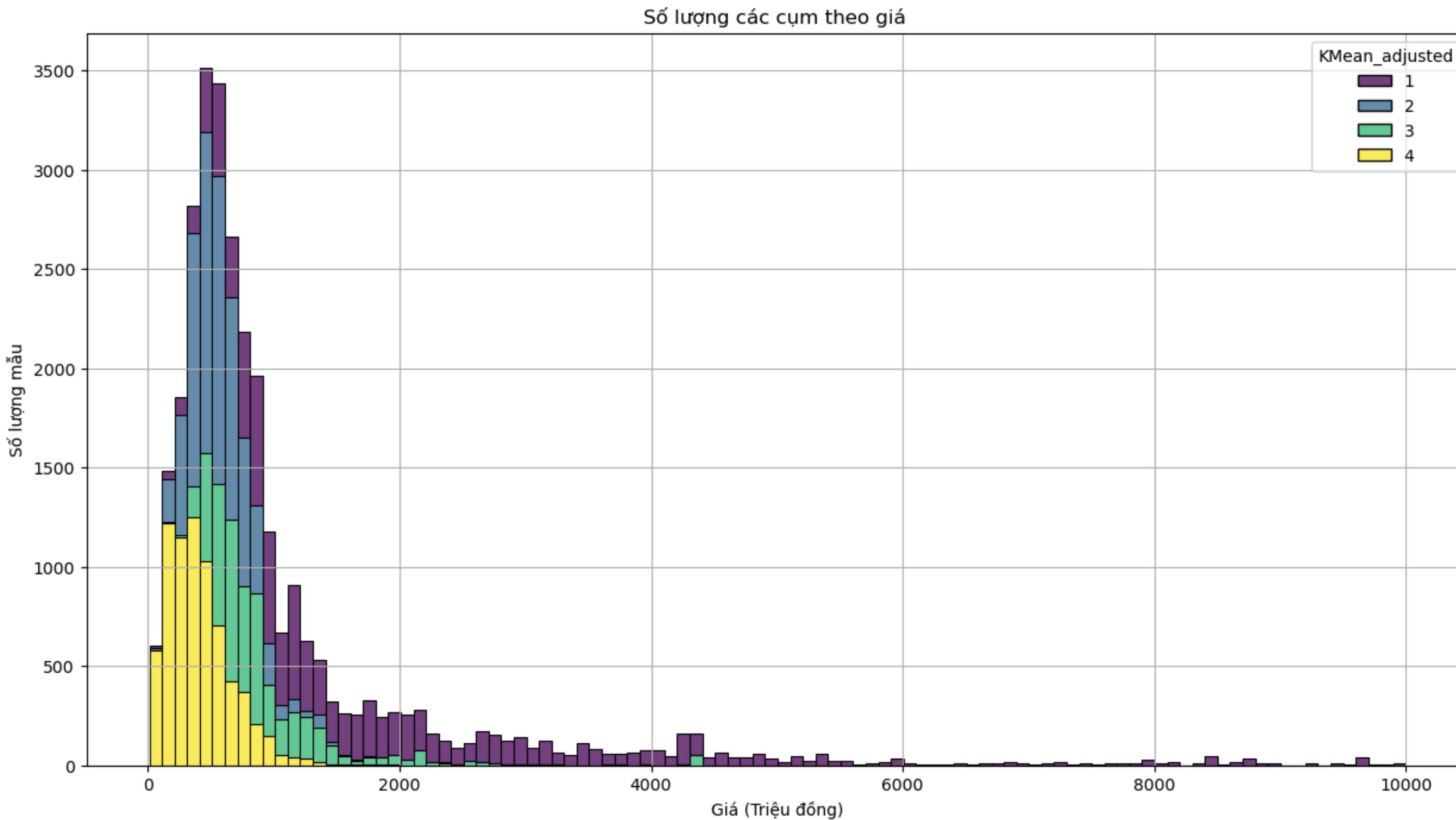
## 2. KMeans

### a. Xác định giá trị K



## 2. KMeans

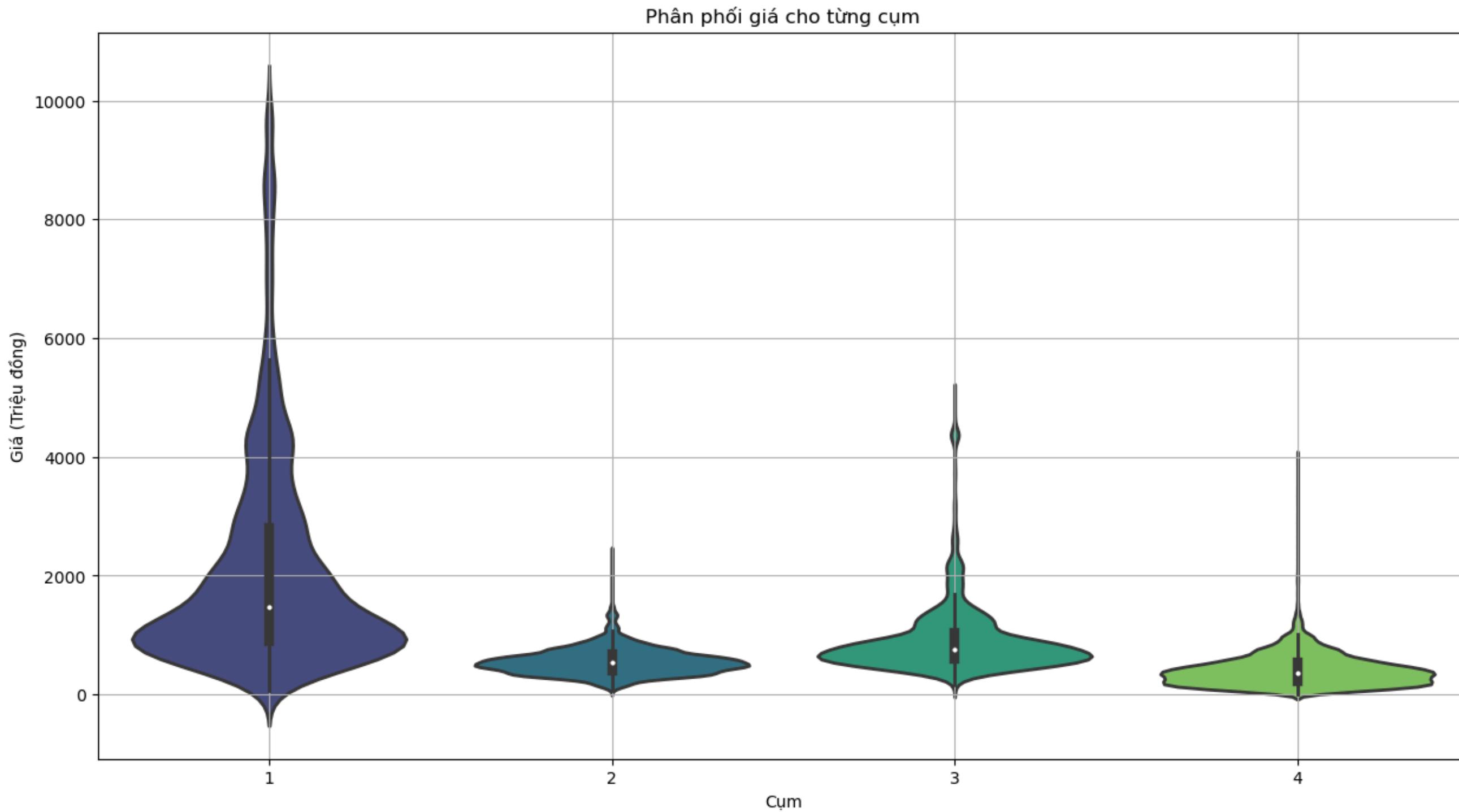
### b. Trực quan hóa



Số lượng mẫu các  
cụm theo giá xe

## 2. KMeans

### b. Trực quan hóa

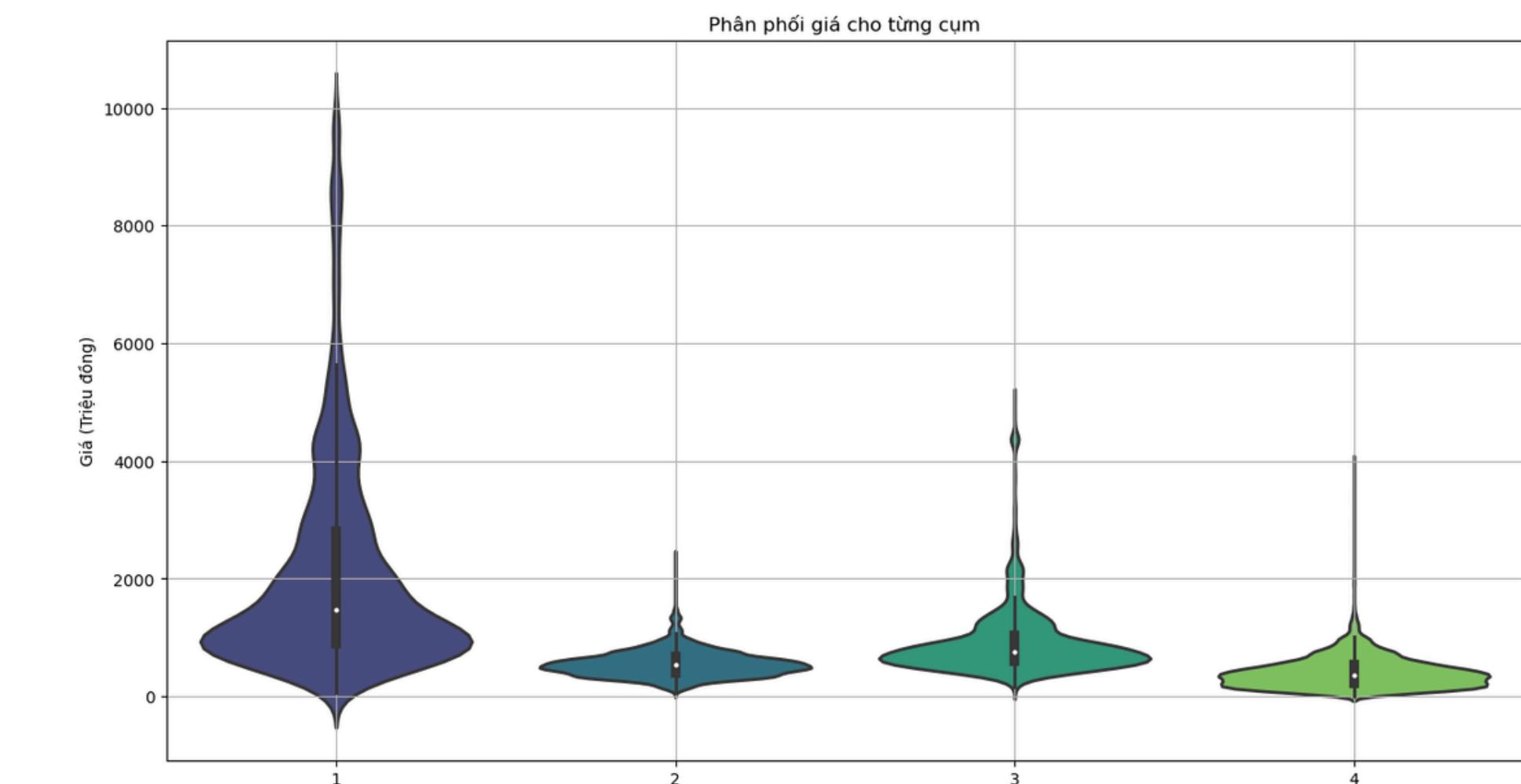
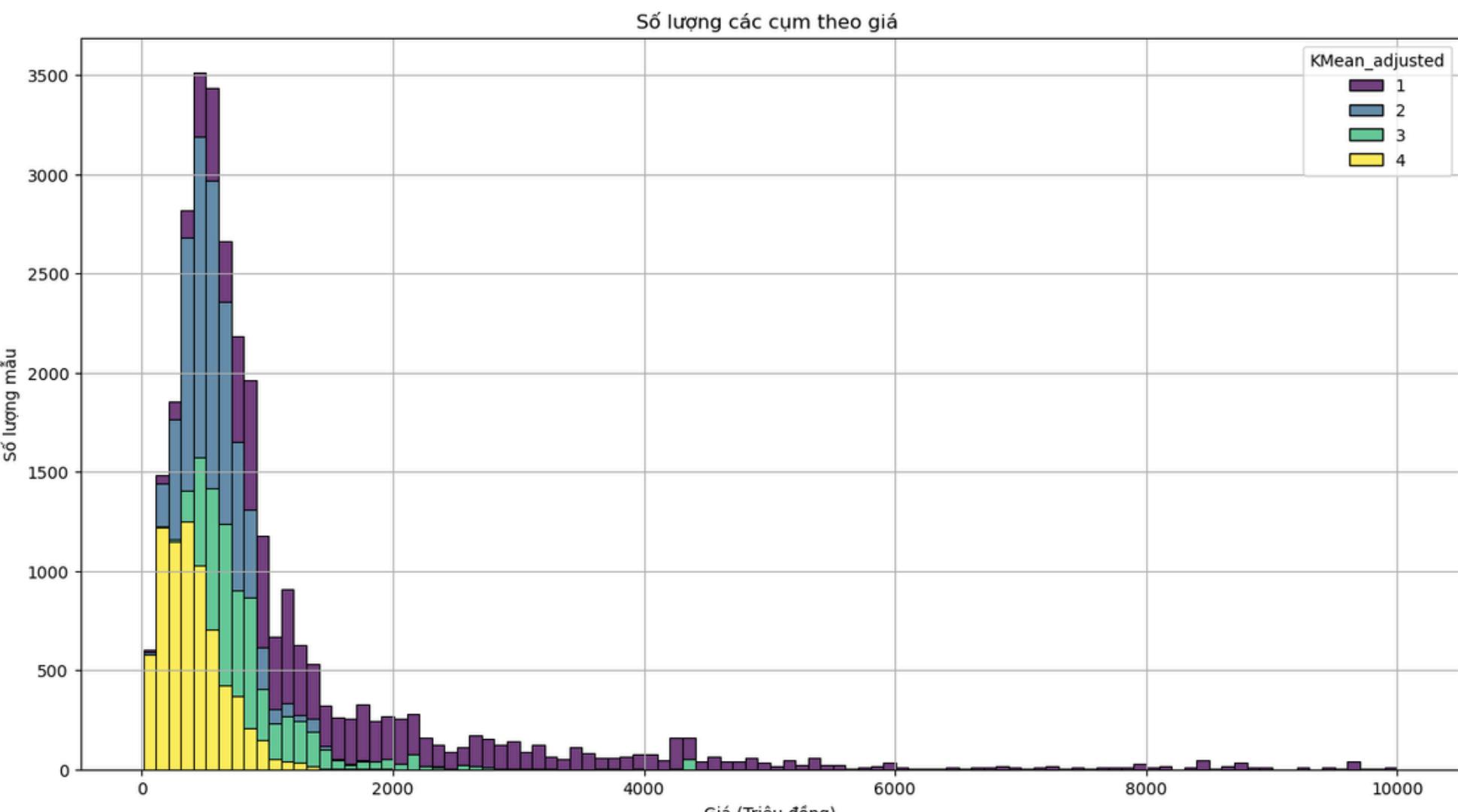


Sự phân bổ của các cụm theo giá

## 2. KMeans

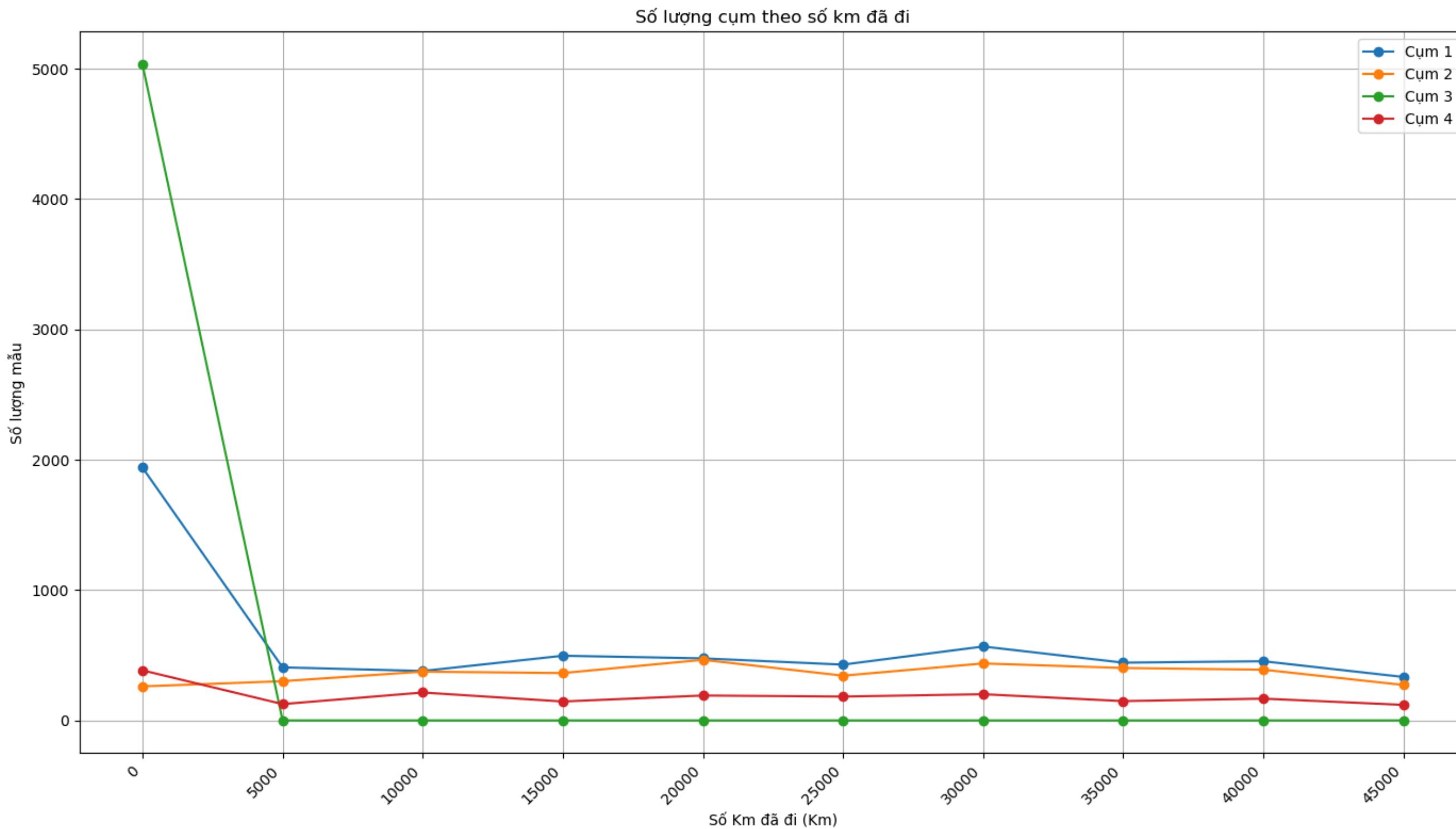
### b. Trực quan hóa

- Cụm 1 chiếm số lượng nhiều nhất và phân bổ trải dài nhất trong 4 cụm.
- Ở khoảng giá 1.5 tỉ trở xuống, số lượng mẫu cụm 2 nhiều hơn cụm 3 nhưng cụm 3 trải dài hơn và trải đến 4 tỉ.
- Cụm 4 nhìn chung có số lượng ít nhất và phân bổ chủ yếu trong khoảng 0 tới 1 tỉ.



## 2. KMeans

### b. Trực quan hóa



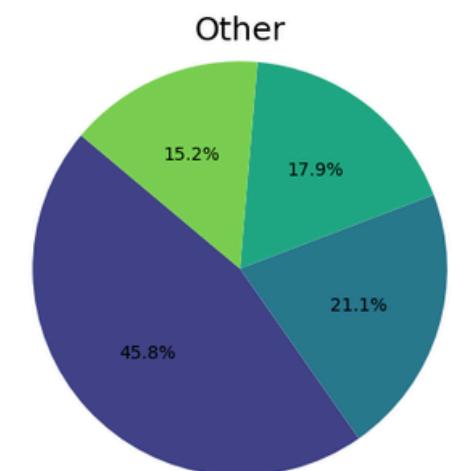
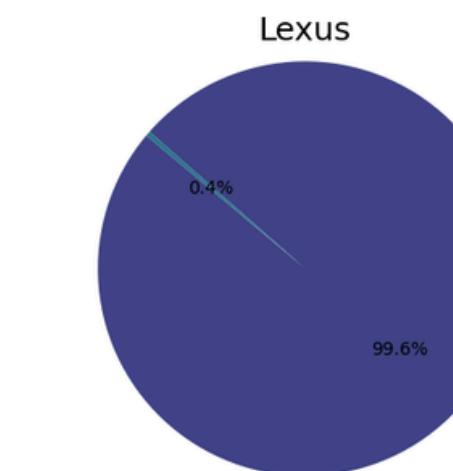
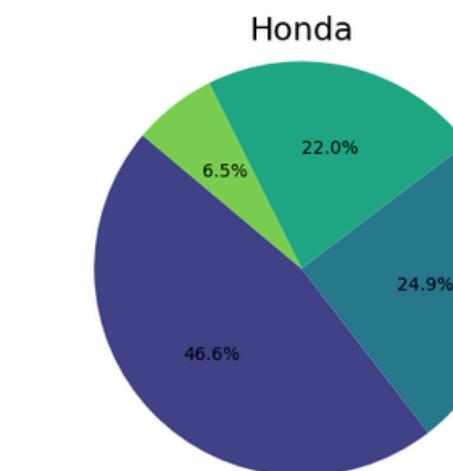
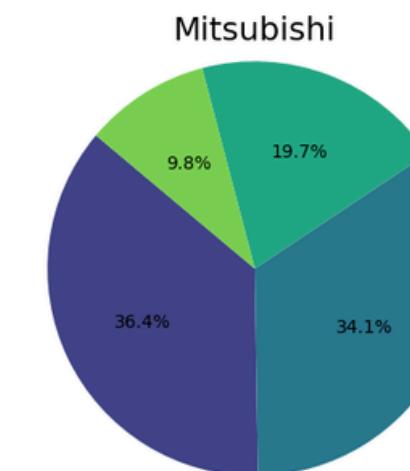
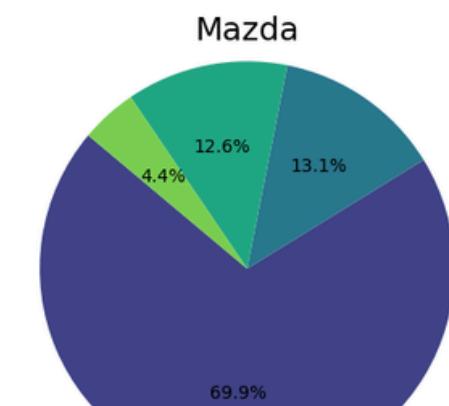
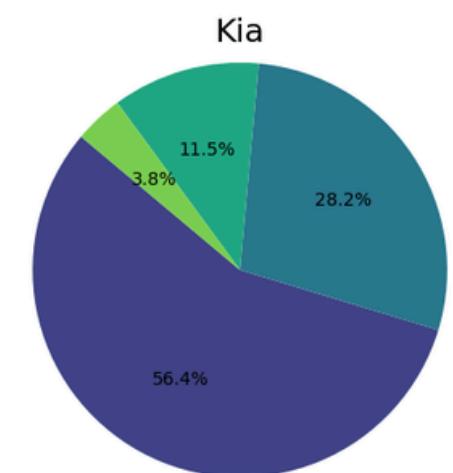
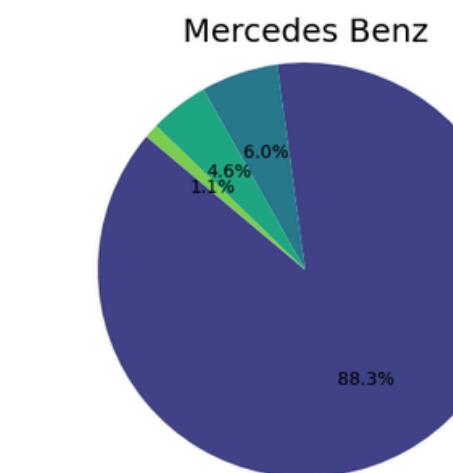
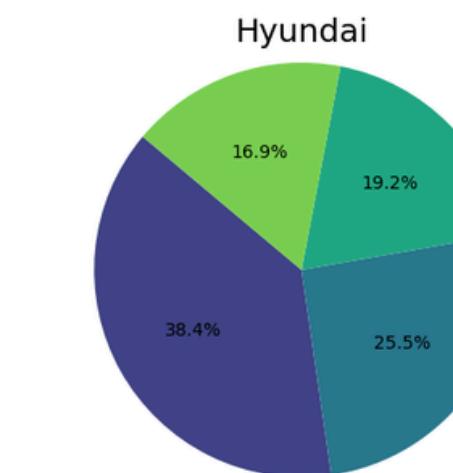
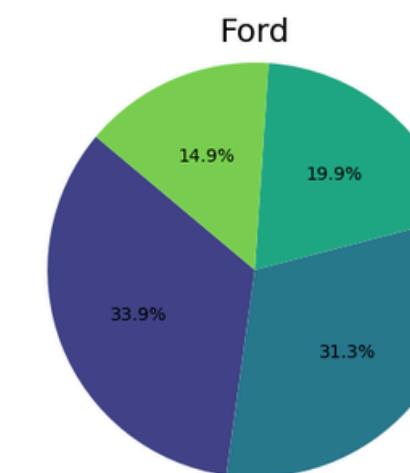
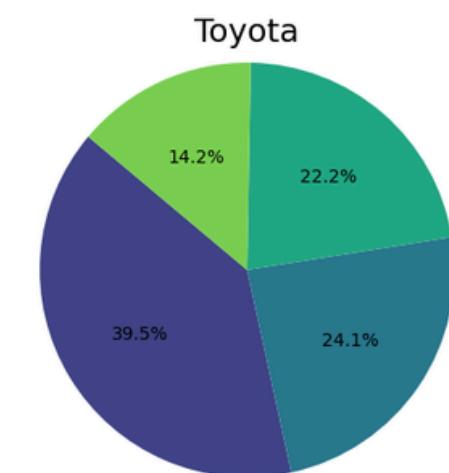
Số lượng cụm  
theo số Km đã đi

## 2. KMeans

### b. Trực quan hóa

Phân bố cụm cho các hãng xe

Phân bố cụm cho các hãng xe

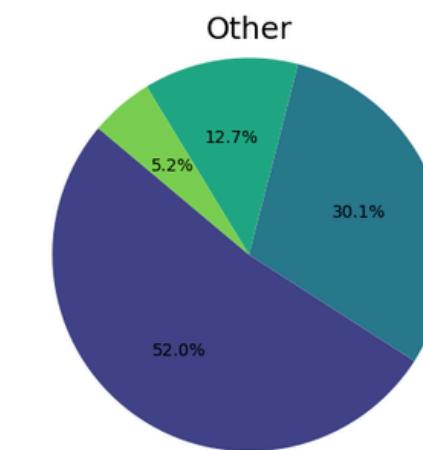
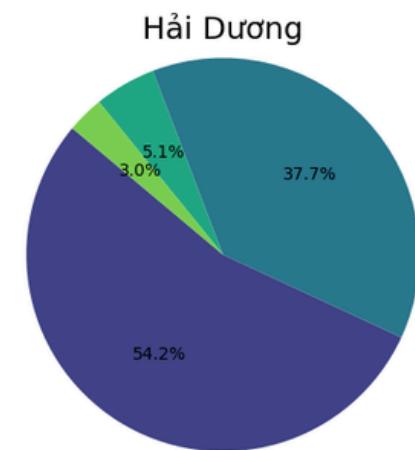
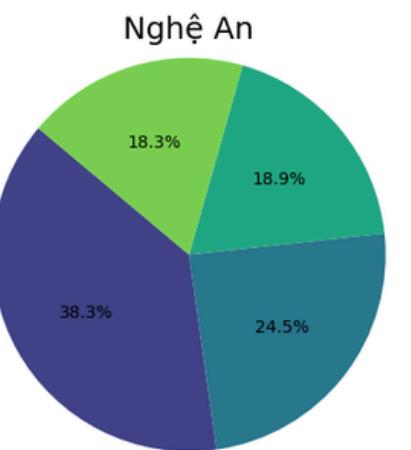
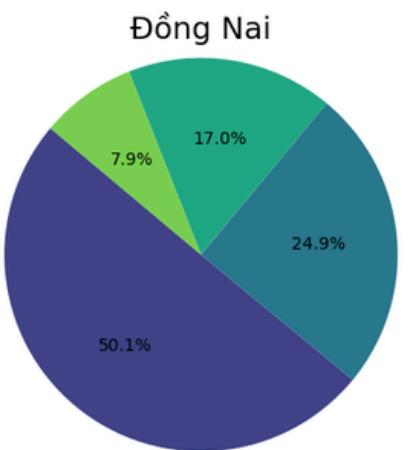
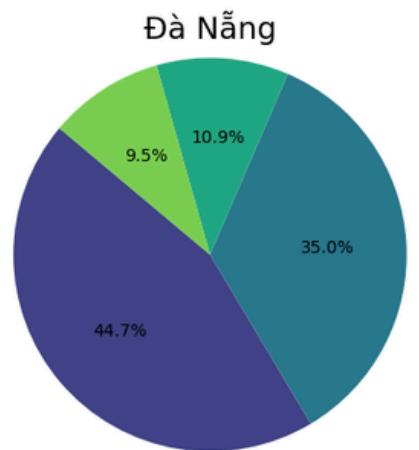
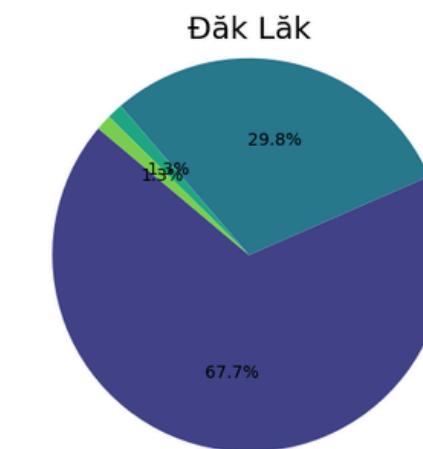
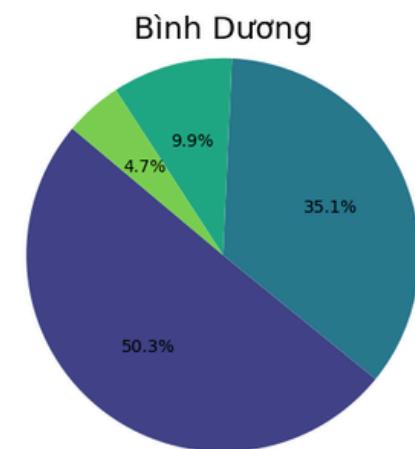
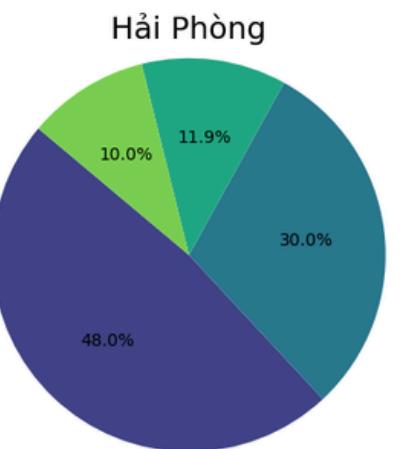
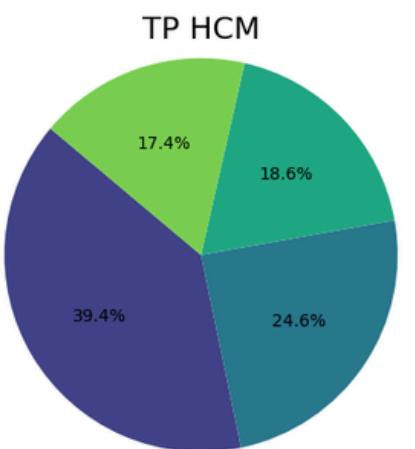
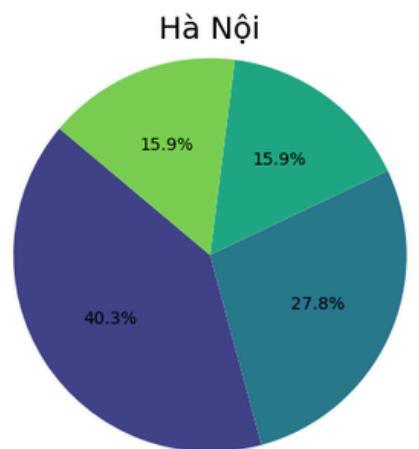


■ Cụm 1 ■ Cụm 2 ■ Cụm 3 ■ Cụm 4

## 2. KMeans

### b. Trực quan hóa

Tỉ lệ số lượng cụm cho các địa điểm

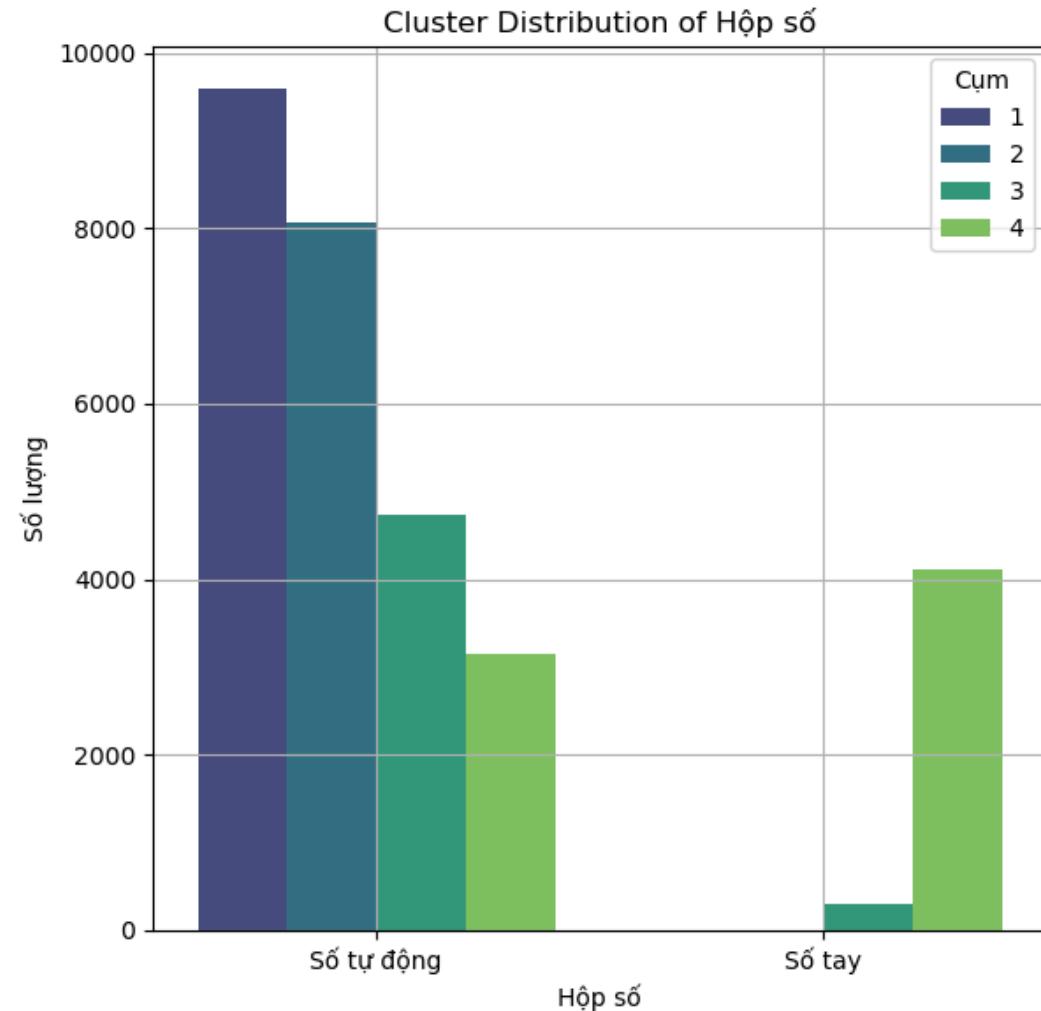
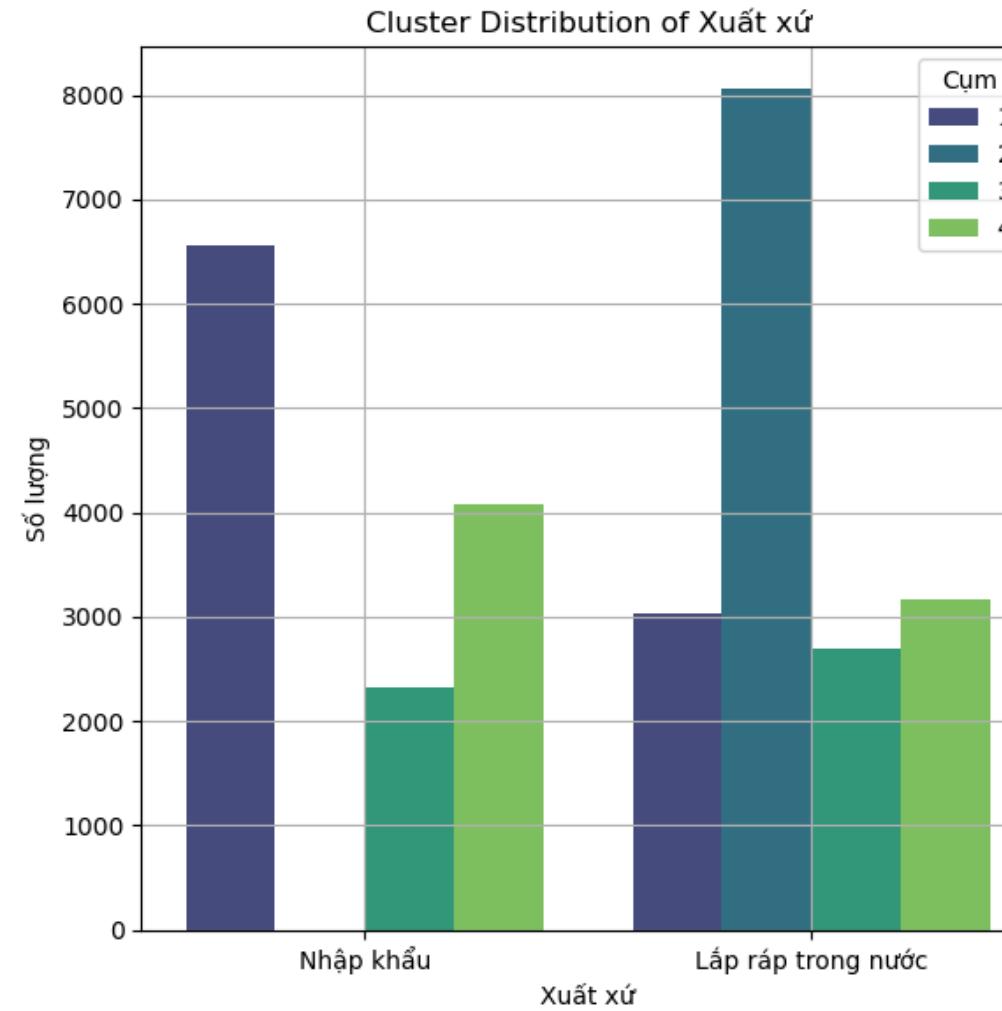
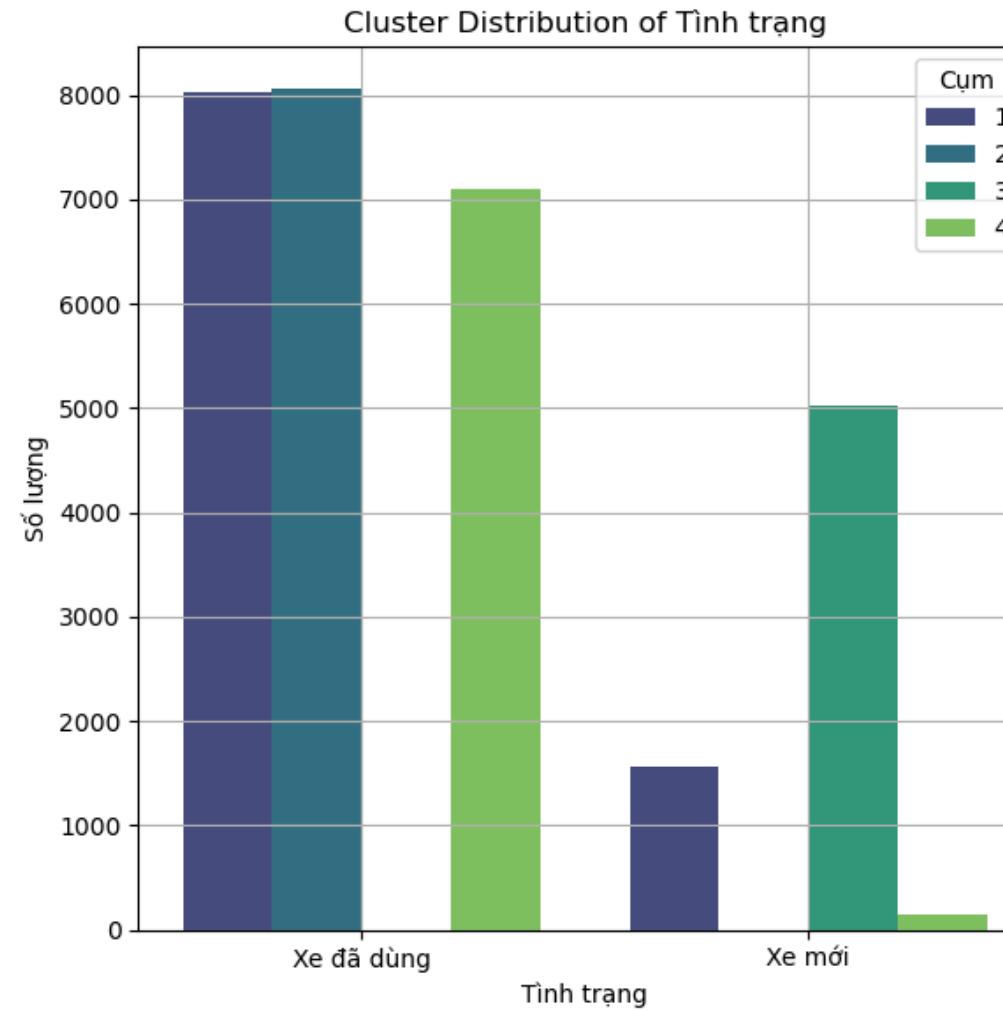


Tỉ lệ số lượng cụm  
của các địa điểm

■ Cụm 1 ■ Cụm 2 ■ Cụm 3 ■ Cụm 4

## 2. KMeans

### b. Trực quan hóa



Số lượng cụm của 1 vài kiểu dữ liệu khác

### 3. GMM

- Gaussian Mixture Model (GMM) là một mô hình thống kê dùng để biểu diễn dữ liệu bao gồm nhiều phân phối Gaussian khác nhau.
- Mô hình này thường được sử dụng để thực hiện phân cụm (clustering) và mô hình hóa phân phối của dữ liệu phức tạp.

### 3. GMM

Thành phần của GMM

- GMM bao gồm nhiều phân phối Gaussian (còn gọi là thành phần) với các tham số riêng biệt.

Bước 1: Khởi tạo

- Khởi tạo các tham số của mô hình (trọng số, giá trị trung bình và phương sai của các phân phối Gaussian).

Bước 2: E-step (Expectation Step)

- Tính toán xác suất một điểm dữ liệu thuộc về mỗi thành phần Gaussian.

Bước 3: M-step (Maximization Step)

- Cập nhật các tham số của mô hình dựa trên các xác suất đã tính toán ở bước E-step.

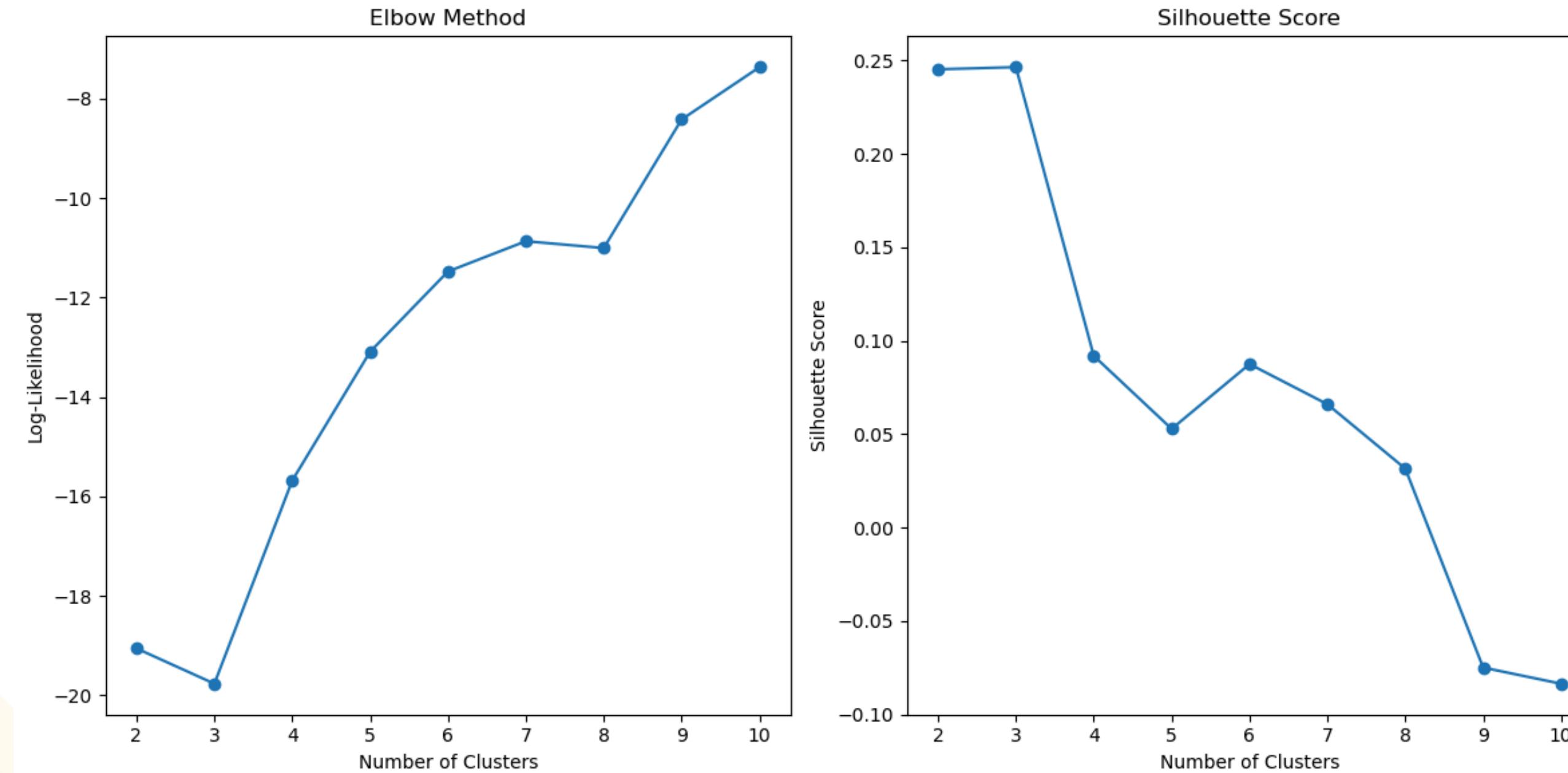
Bước 4: Lặp lại

- Lặp lại bước 2 và 3 cho đến khi hội tụ (các tham số không thay đổi hoặc thay đổi rất ít).

### 3. GMM

#### a. Xác định số cụm

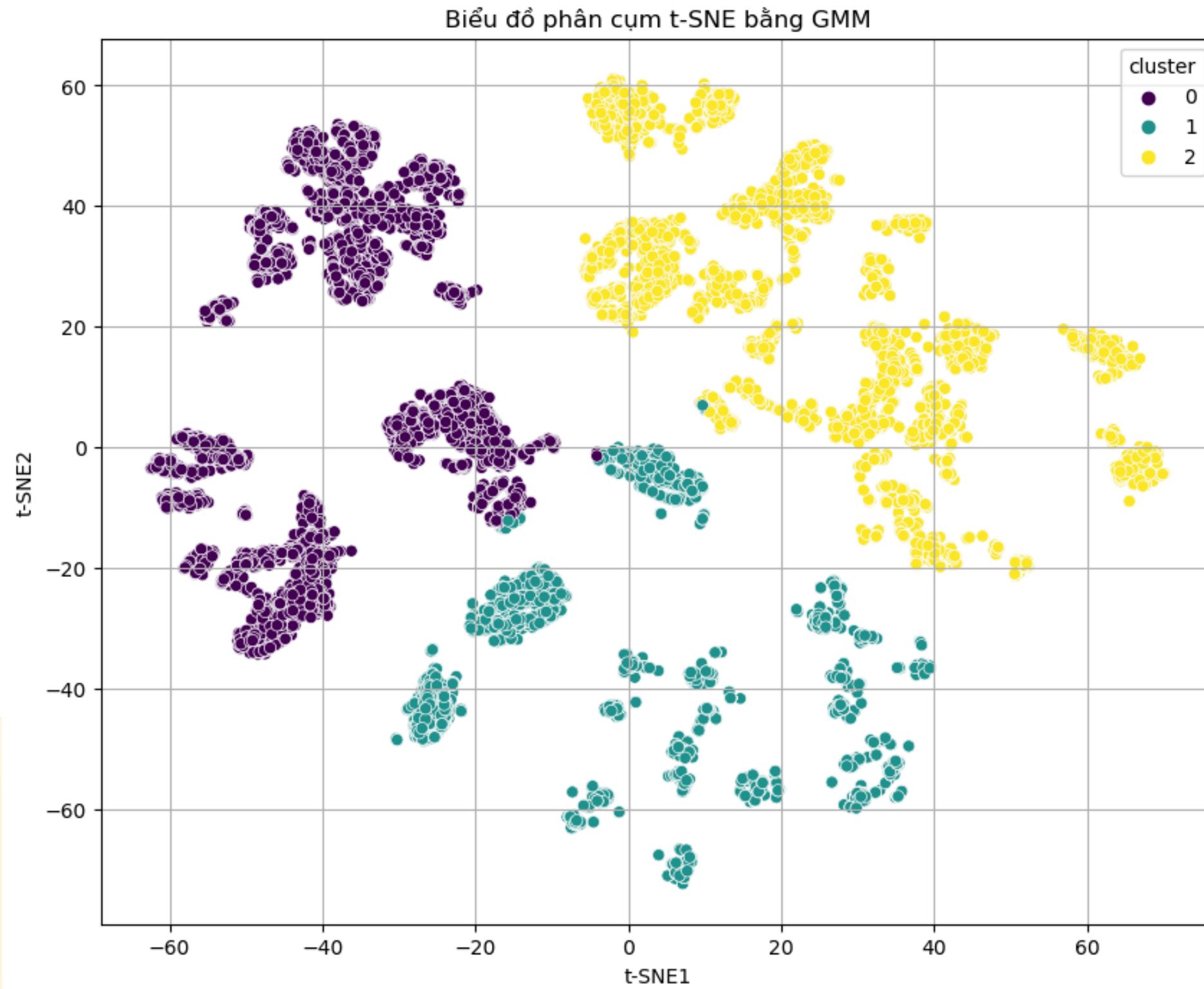
- Sử dụng elbow method và silhouette score



**K = 3**

### 3. GMM

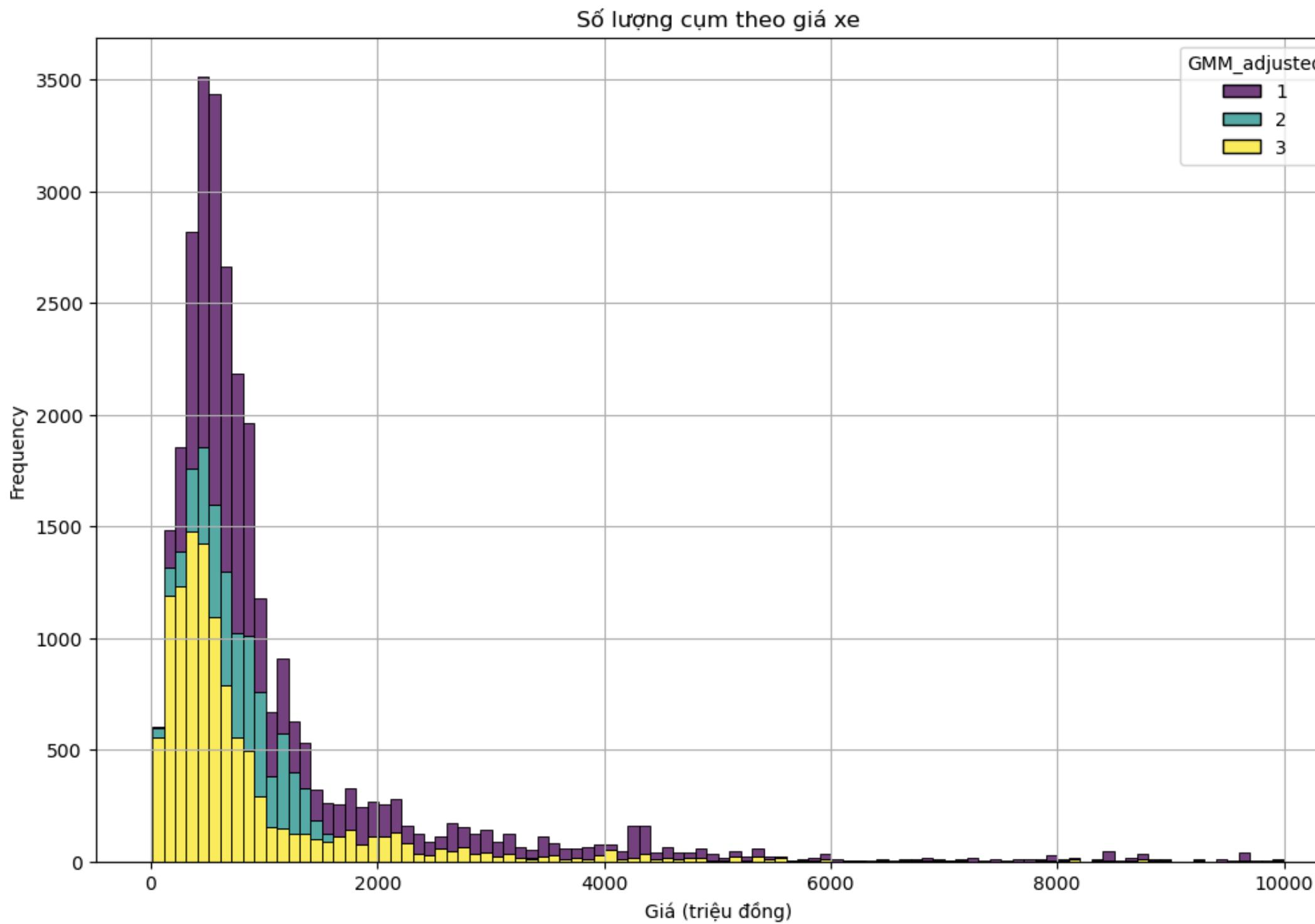
#### a. Xác định số cụm



Biểu đồ phân cụm

### 3. GMM

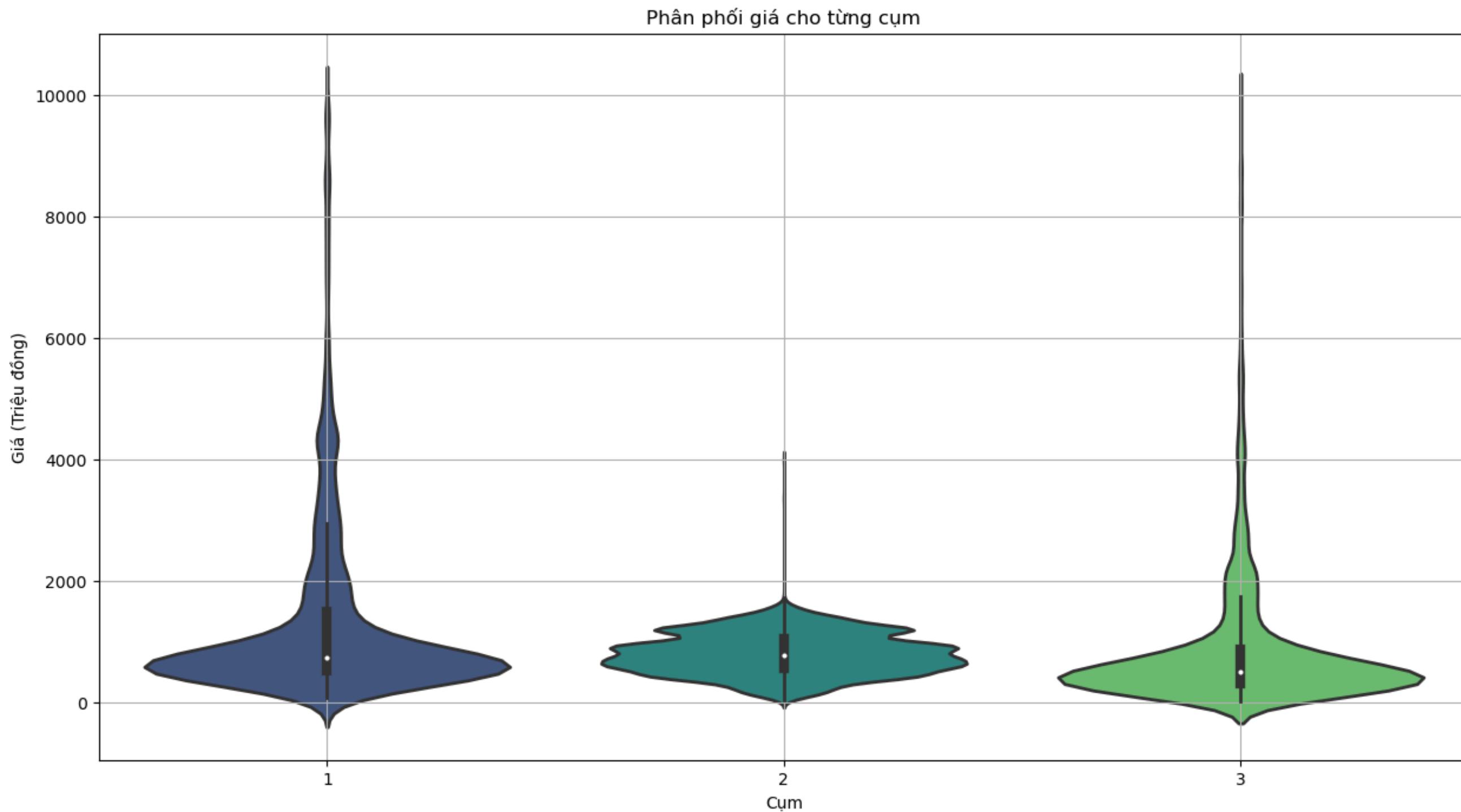
#### a. Xác định số cụm



Số lượng cụm theo giá xe

### 3. GMM

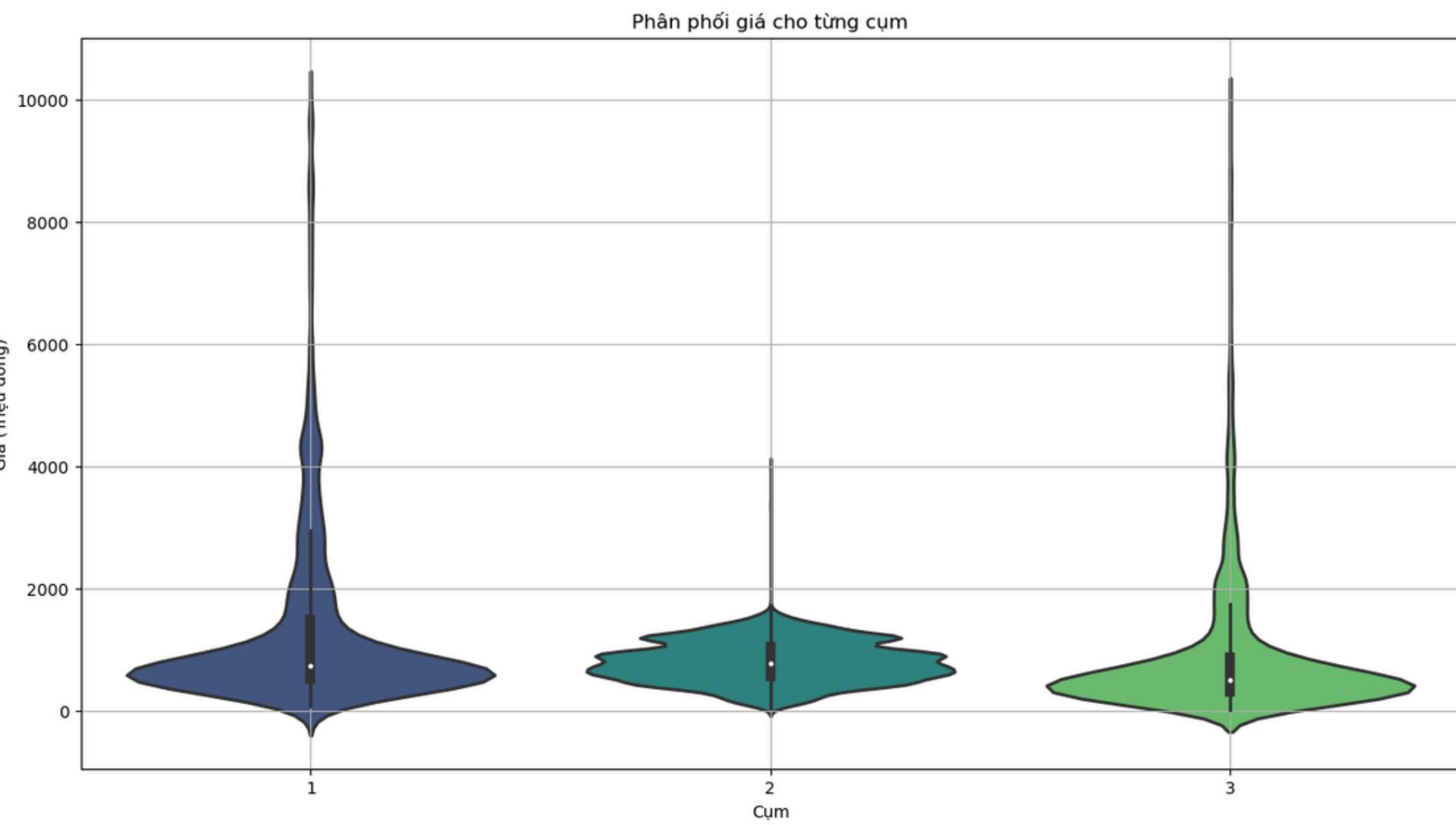
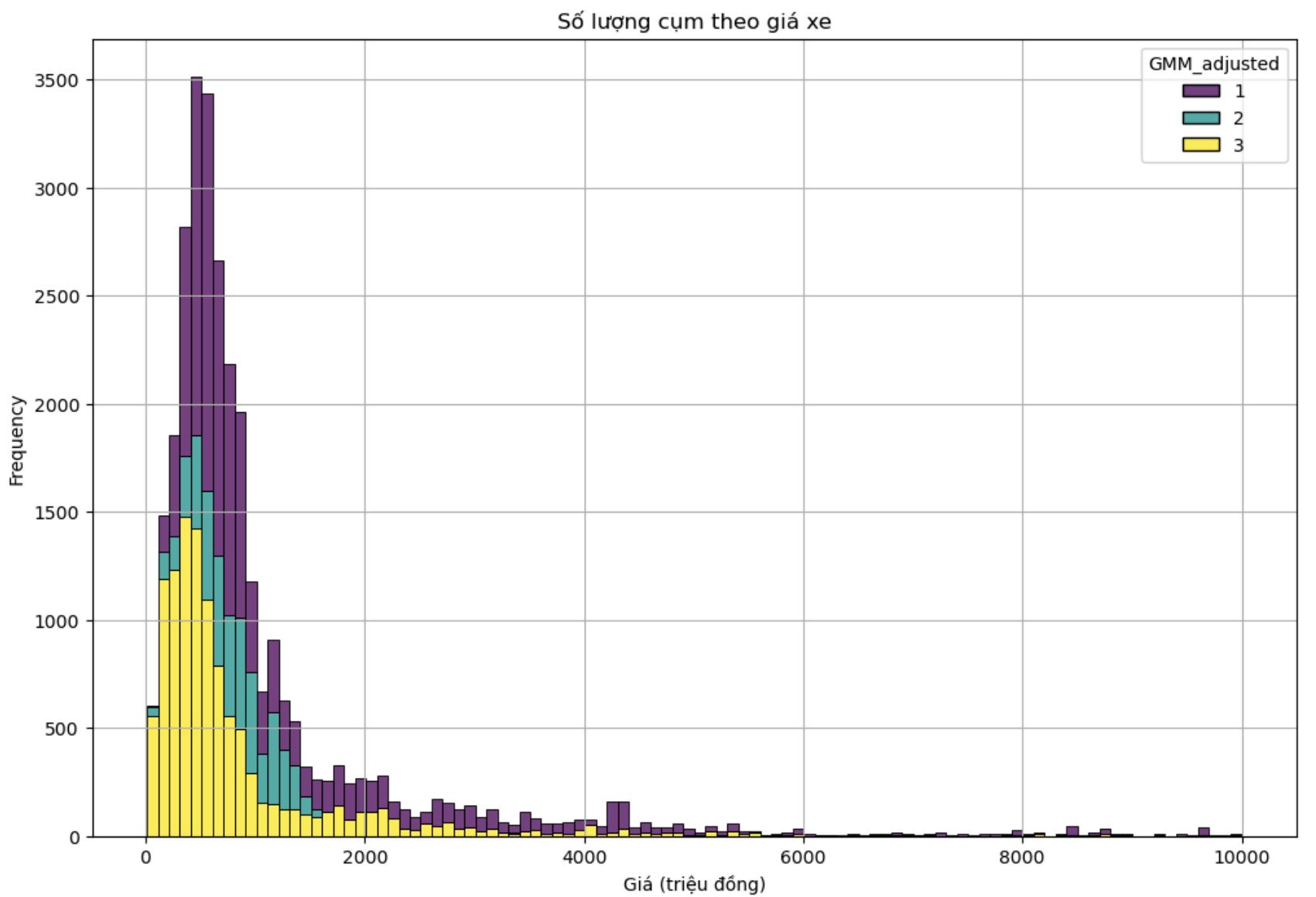
#### a. Xác định số cụm



Phân phối giá cho  
từng cụm

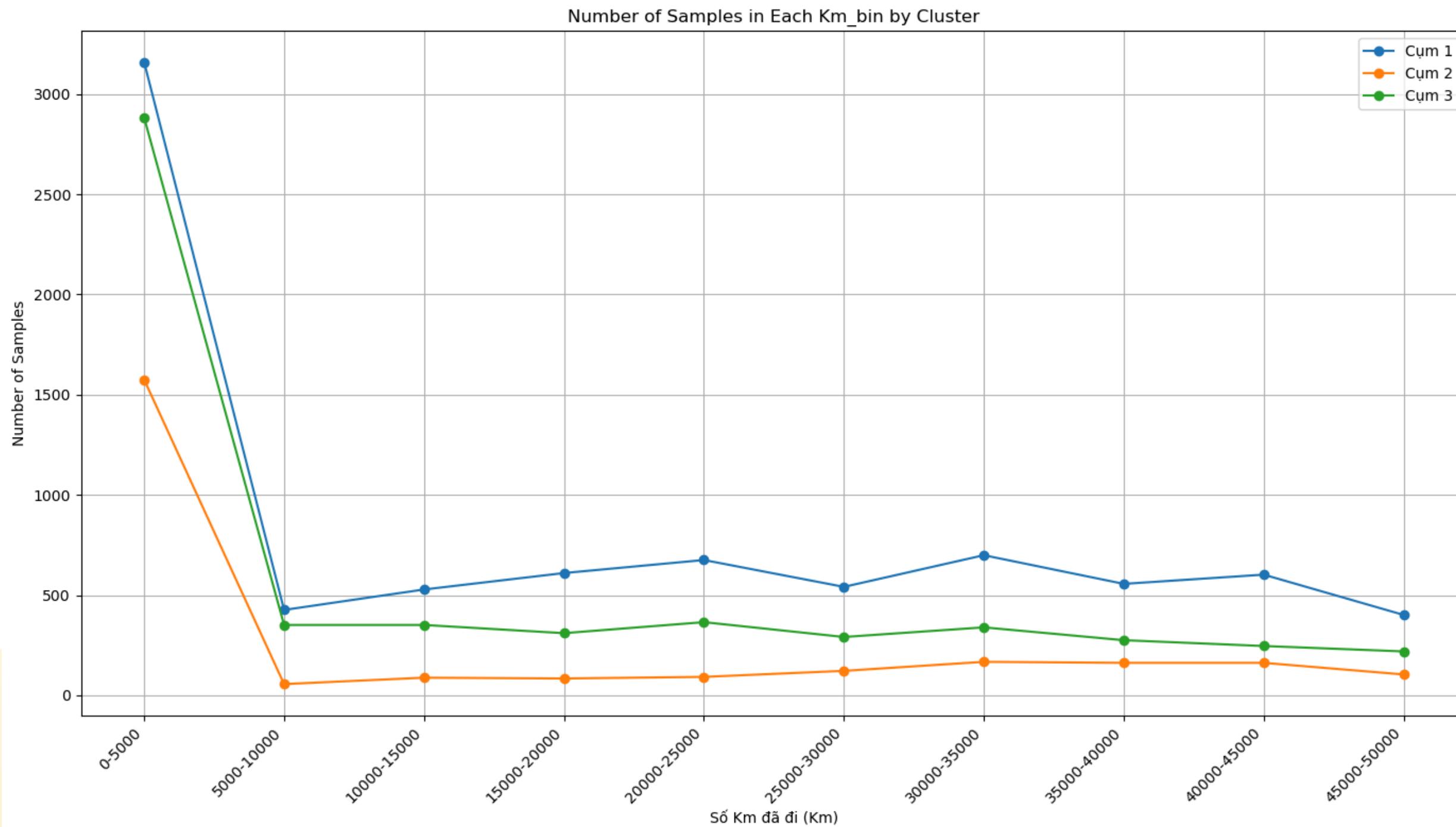
### 3. GMM

#### a. Xác định số cụm



### 3. GMM

#### a. Xác định số cụm

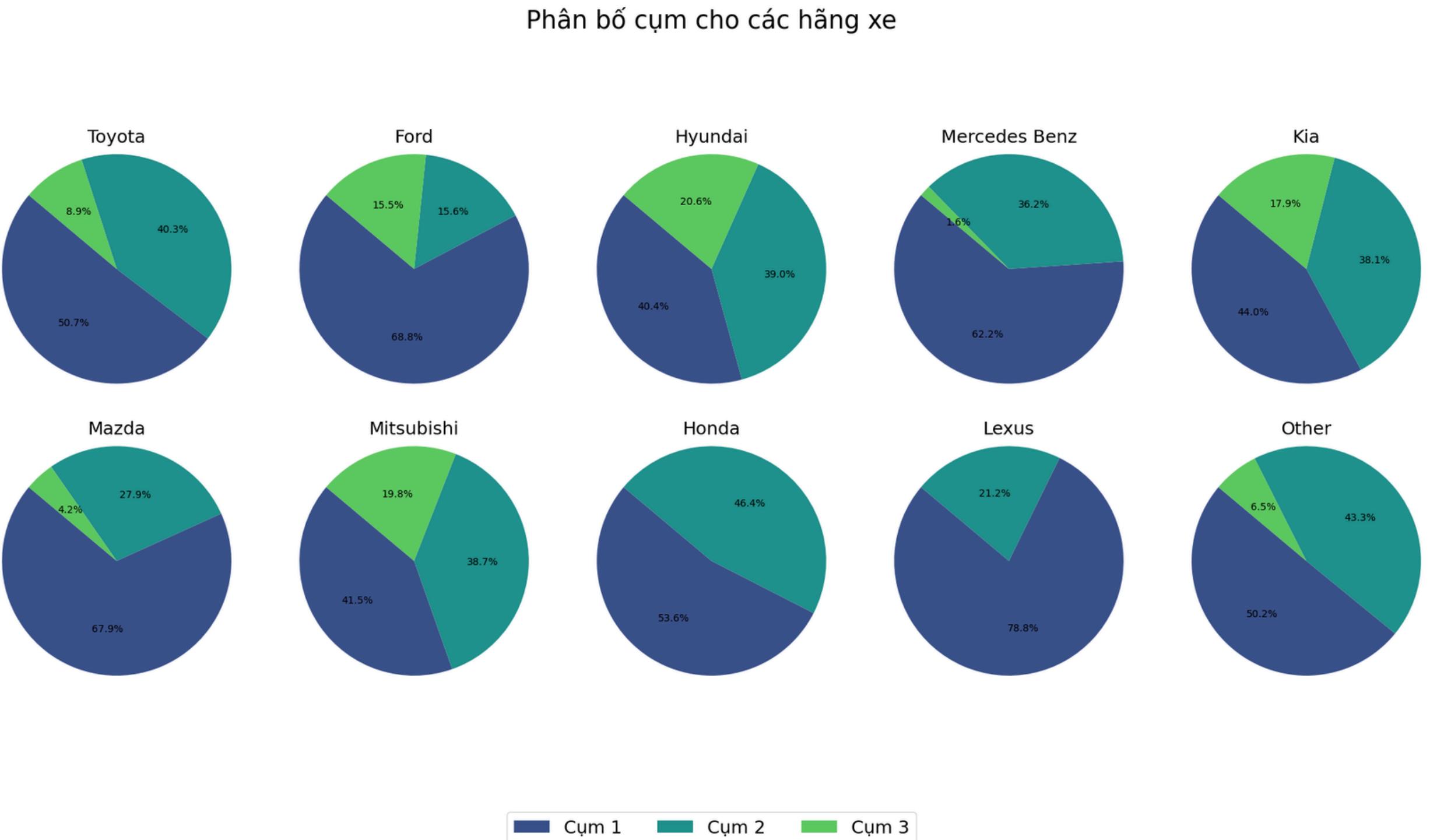


Số lượng phân cụm  
theo số Km đã đi

### 3. GMM

#### a. Xác định số cụm

Tỉ lệ số lượng cụm  
cho các hãng xe

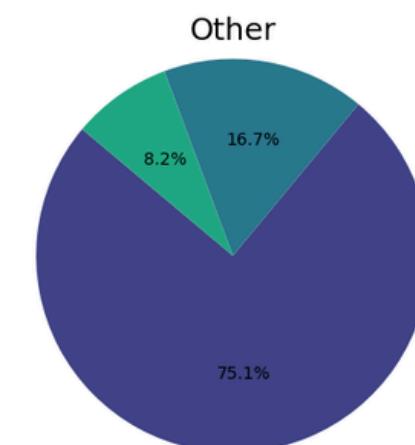
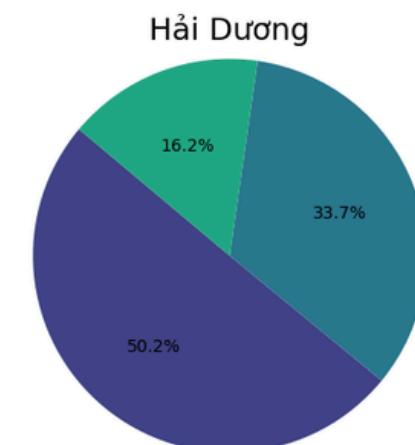
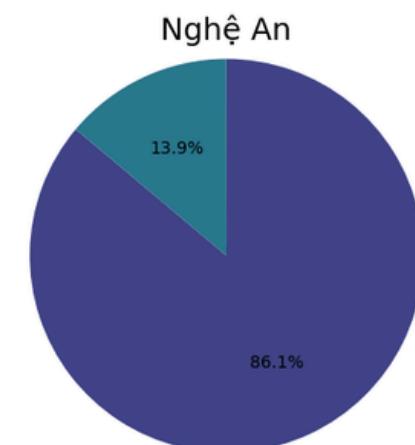
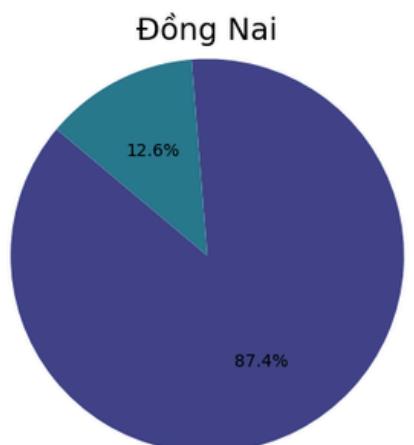
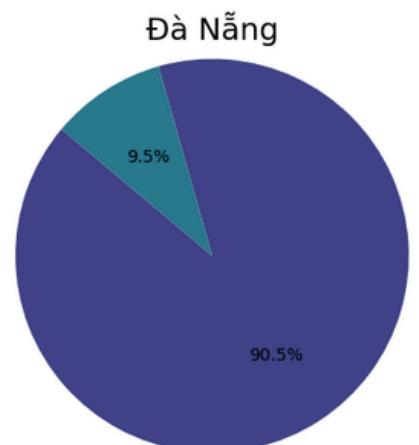
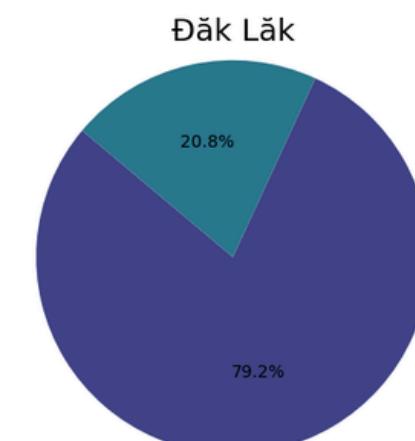
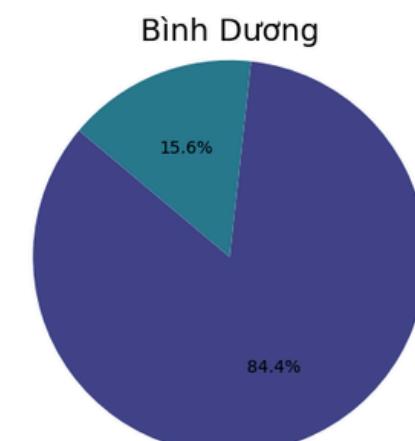
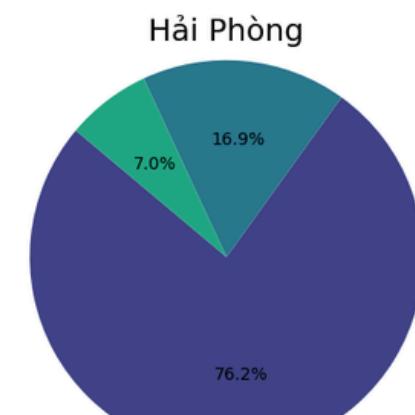
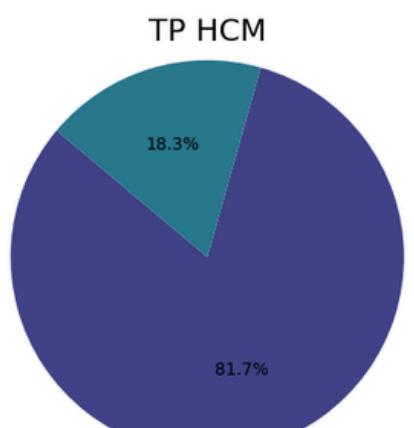
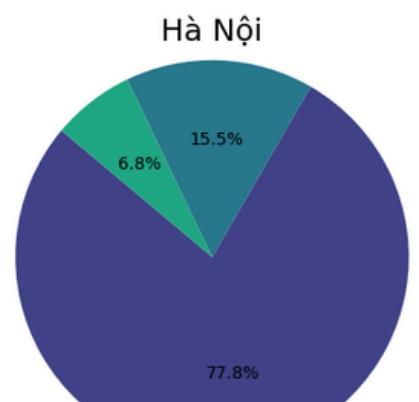


### 3. GMM

#### a. Xác định số cụm

Tỉ lệ số lượng cụm  
cho các địa điểm

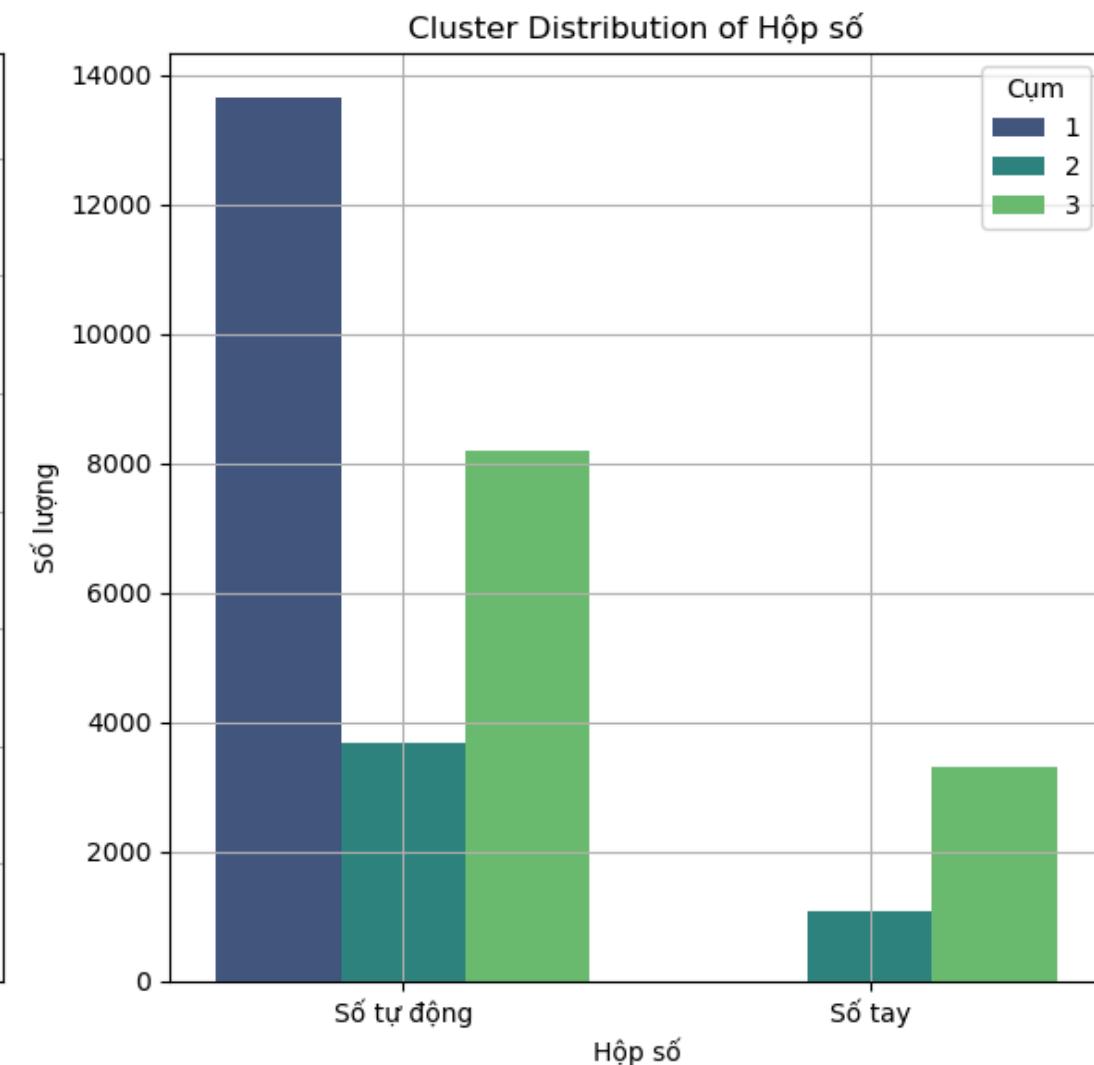
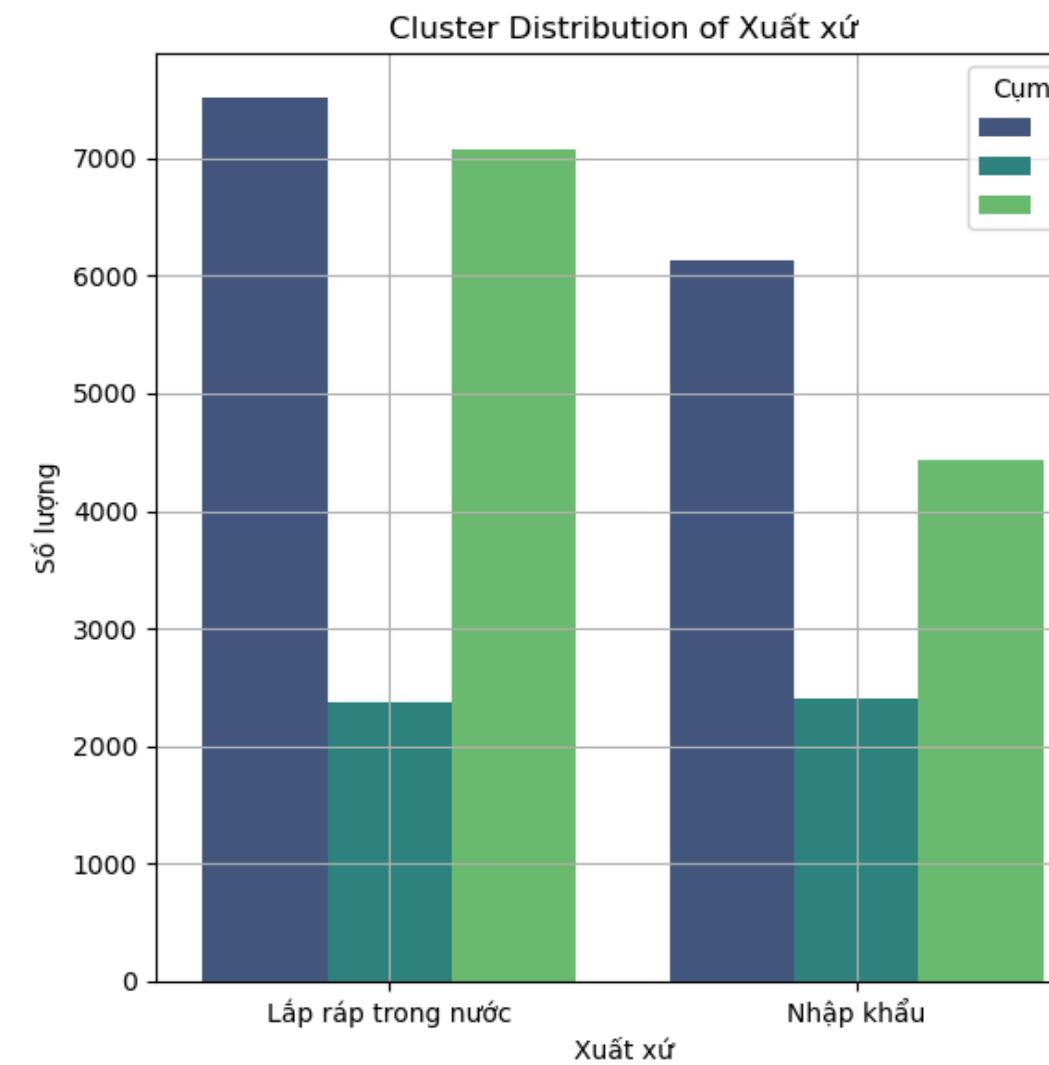
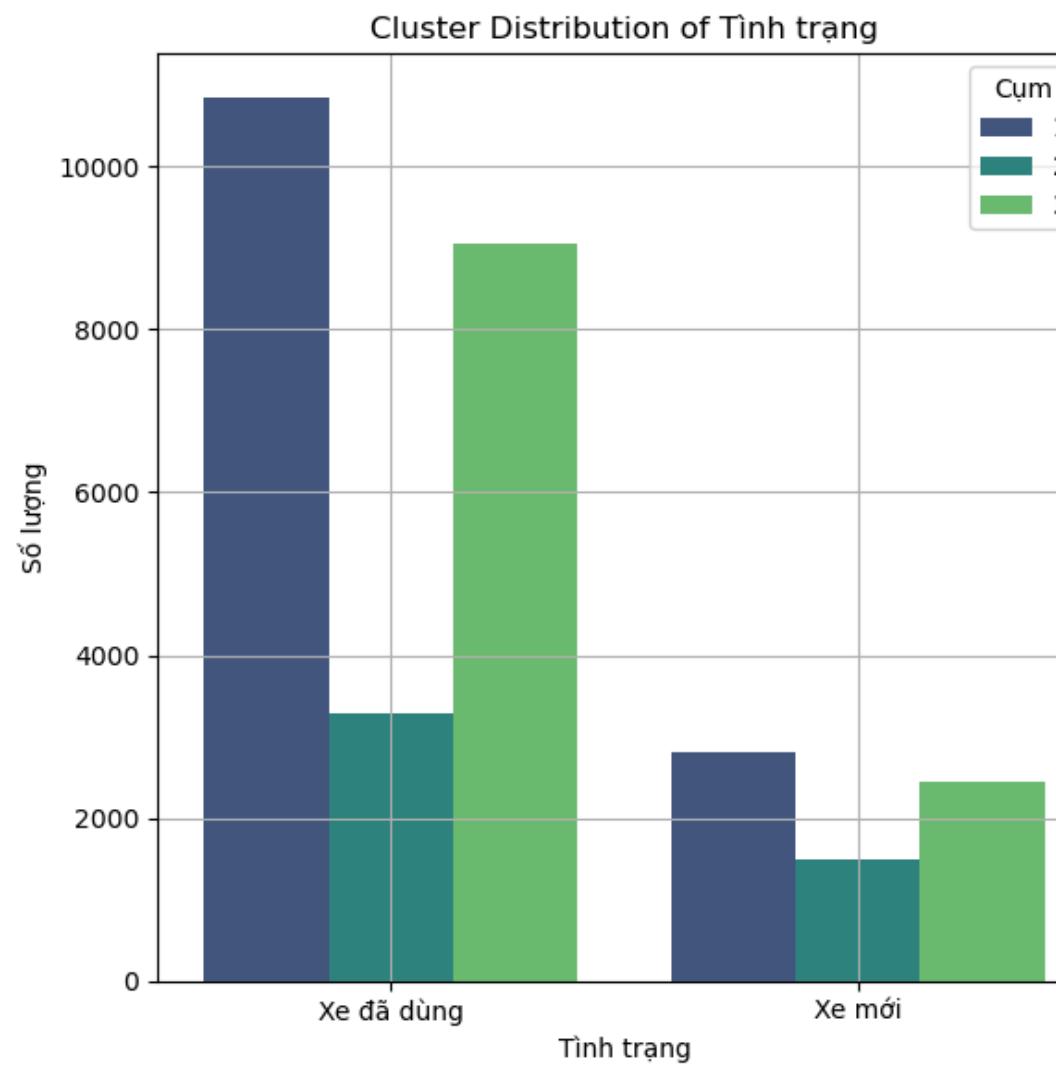
Tỉ lệ số lượng cụm cho các địa điểm



■ Cụm 1 ■ Cụm 2 ■ Cụm 3

### 3. GMM

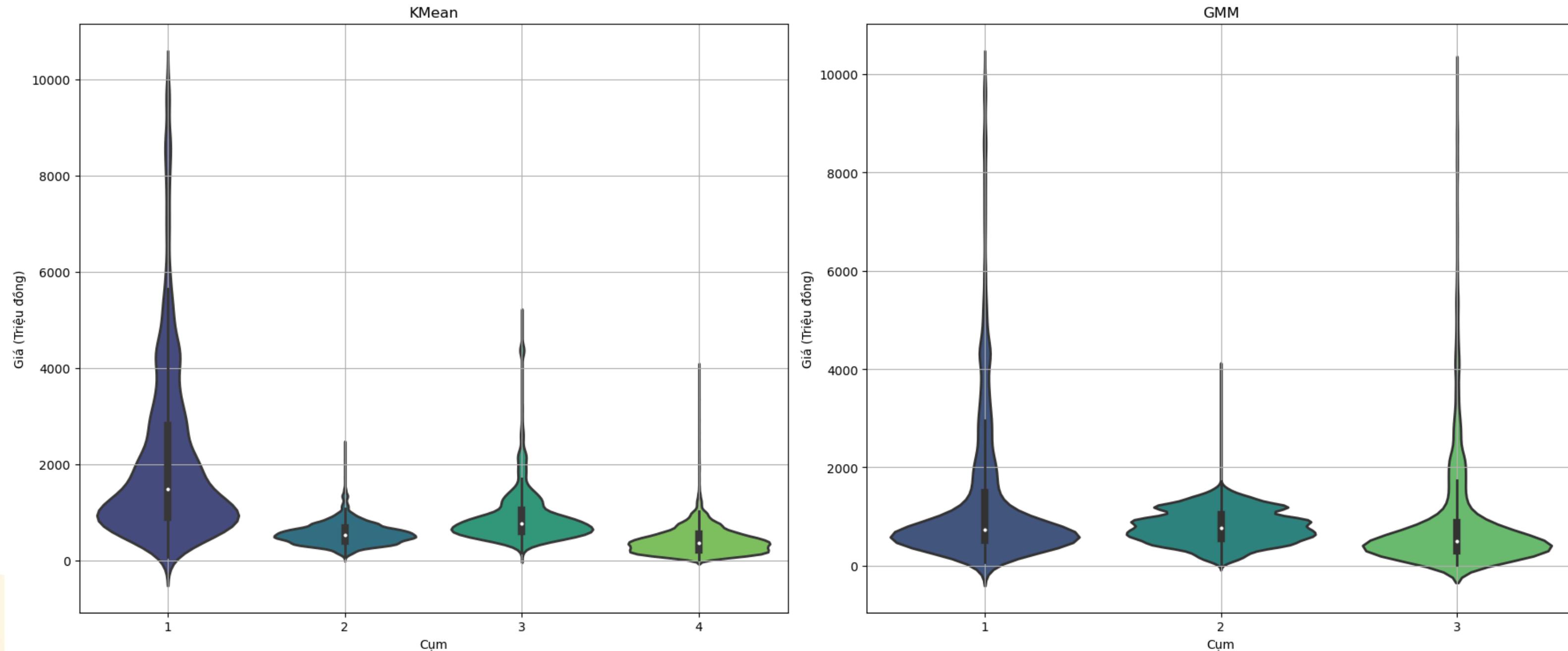
#### a. Xác định số cụm



Số lượng cụm của 1 vài kiểu dữ liệu khác

## 4. So sánh mô hình

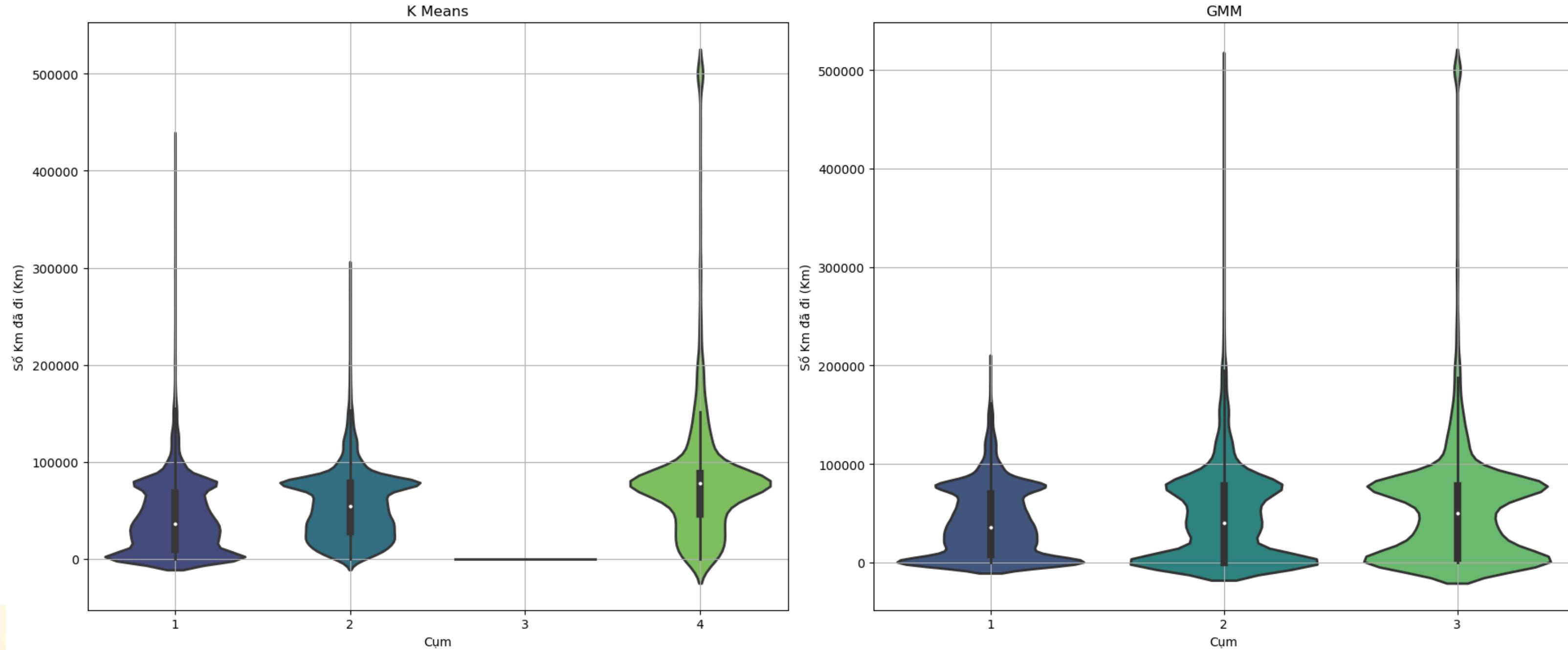
So sánh sự phân bổ cụm theo giá của 2 mô hình



Tỉ lệ số lượng phân bổ cụm của KMeans và GMM  
theo Giá xe

## 4. So sánh mô hình

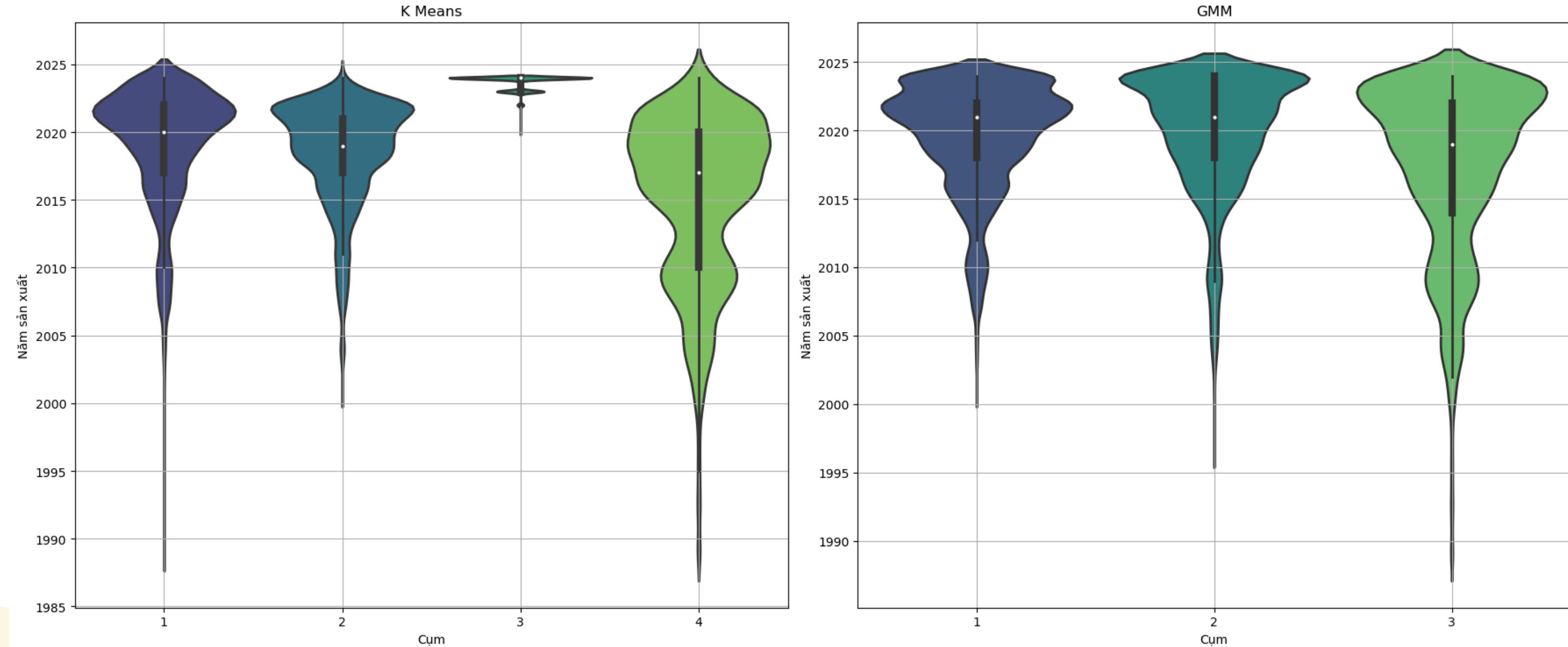
So sánh sự phân bổ cụm theo số km đã đi của 2 mô hình



Tỉ lệ số lượng phân bổ cụm của KMeans và GMM  
theo Số Km đã đi

## 4. So sánh mô hình

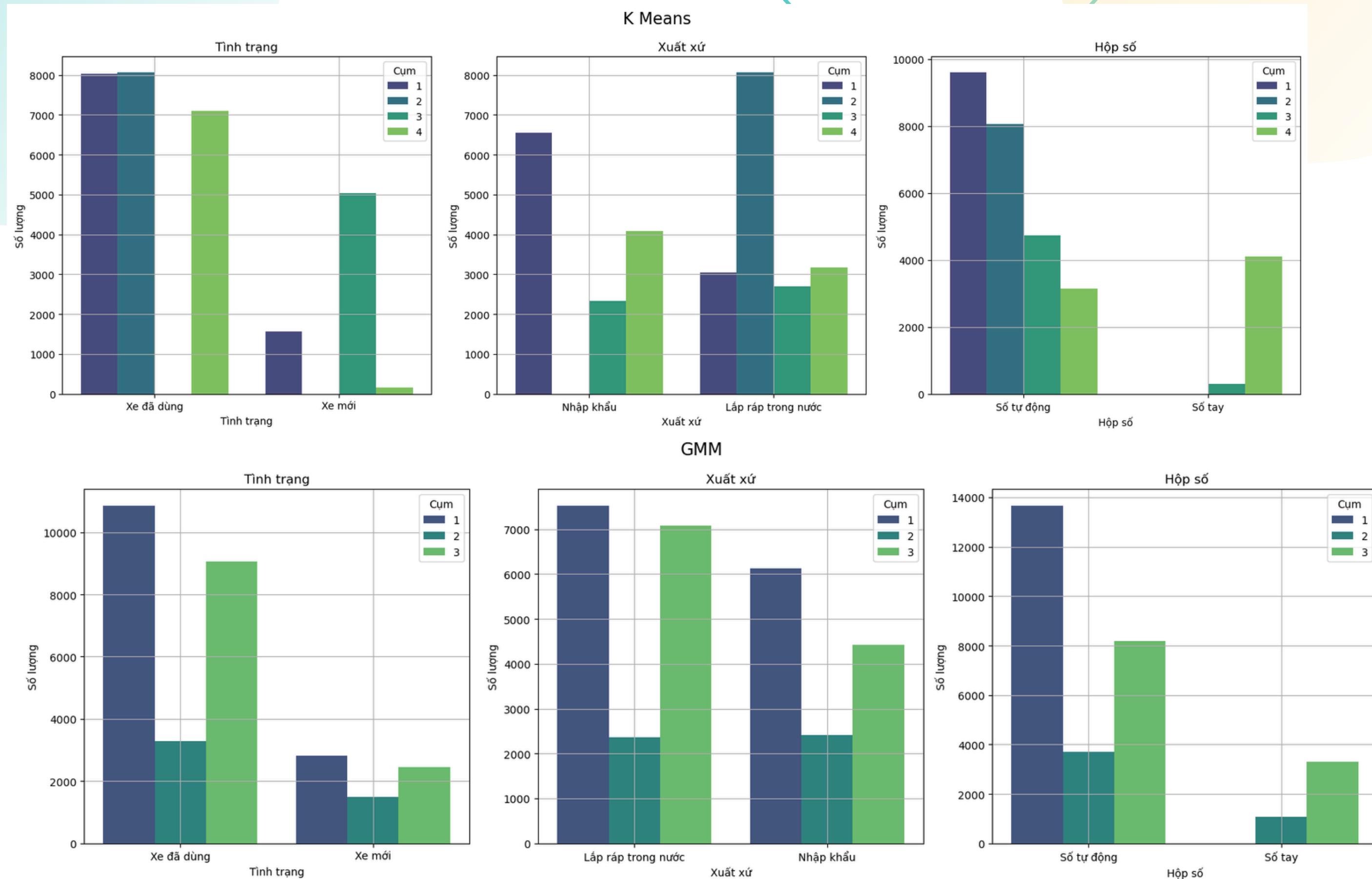
So sánh sự phân bổ cụm theo số km đã đi của 2 mô hình



Tỉ lệ số lượng phân bổ cụm của KMeans và GMM  
theo năm sản xuất

# 4. So sánh mô hình

Tỉ lệ số lượng phân bổ  
cụm của KMeans và GMM  
theo 1 vài đặc trưng khác



## 5. Kết luận

1. Mô hình GMM cho thấy ưu điểm, tốt hơn mô hình K-means. Có thể nhận diện và phân cụm rõ ràng ưu việt hơn.
2. Đối với mô hình GMM dữ liệu phân chia thành 3 cụm là hợp lý nhất, cụ thể: đa số các đặc trưng GMM phân cụm đồng đều và hợp lý, tuy nhiên vẫn có những đặc trưng mô hình phân bổ không đều như tình trạng, hộp số, ...
3. Đối với mô hình K-means dữ liệu được phân chia thành 4 cụm là tối ưu nhất cụ thể: đa số các đặc trưng mô hình dự đoán không đồng đều đáng kể, nhất là các đặc trưng như giá, số Km đã đi, năm sản xuất, ...
4. Những đặc trưng như Giá, Tình trạng, Hộp số cả 2 mô hình đều dự đoán không đồng đều