



# BÁO CÁO KHO VÀ KHAI PHÁ DỮ LIỆU

CHỦ ĐỀ:

QUẢN TRỊ NHÂN SỰ  
RỜI BỎ DOANH NGHIỆP

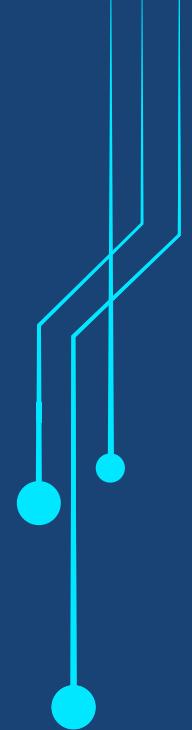
NHÓM:

10 - MIS3009\_48K29.1

GVHD:

TS. PHAN ĐÌNH VĂN

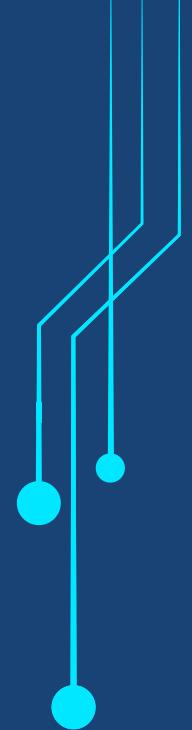
# THÀNH VIÊN NHÓM



TÊN THÀNH VIÊN	MỨC ĐÓNG GÓP
NGUYỄN MẠNH THỊNH	25%
LÊ THUÝ QUỲNH	25%
NGUYỄN THỊ DIÊM LY	25%
NGUYỄN PHƯƠNG THẢO	25%

NGUYỄN MẠNH THỊNH

# THÀNH VIÊN NHÓM



TÊN THÀNH VIÊN	Công việc chính
NGUYỄN MẠNH THỊNH	Nghiên cứu, thực hiện thuật toán phân lớp, phân cụm
LÊ THUÝ QUỲNH	Nghiên cứu, thực hiện thuật toán luật kết hợp, phân lớp
NGUYỄN THỊ DIỄM LY	Nghiên cứu, thực hiện thuật toán luật kết hợp, phân cụm
NGUYỄN PHƯƠNG THẢO	Nghiên cứu, thực hiện thuật toán luật kết hợp, phân lớp

NGUYỄN MẠNH THỊNH

# TABLE OF CONTENT

1

GIỚI THIỆU ĐỀ TÀI

2

TIỀN XỬ LÍ DỮ LIỆU & EDA

3

PHÂN LỚP

4

PHÂN CỤM

5

LUẬT KẾT HỢP

6

ĐÁNH GIÁ & KẾT LUẬN

NGUYỄN MẠNH THỊNH

# ĐẶT VĂN ĐỀ

**68%**  
of employees leave  
due to Controllable  
factors

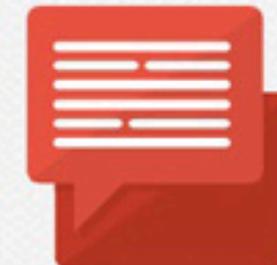


**72%** of employees stated that  
they were unhappy with  
the quality of retention  
effort

  
**37%**  
of employees leave  
because of their  
immediate  
managers



**3 OUT OF 10**  
Candidates do not join  
the company in spite of  
accepting the offer



**61%**

of Managers stated that they need  
coaching on areas of appraisal,  
promotion and salary related  
communication

Get in touch with us to improve the  
quality of retention at your  
workplace!

[www.AceNgage.com](http://www.AceNgage.com)

 tellmemore@acengage.com  
+919901998587

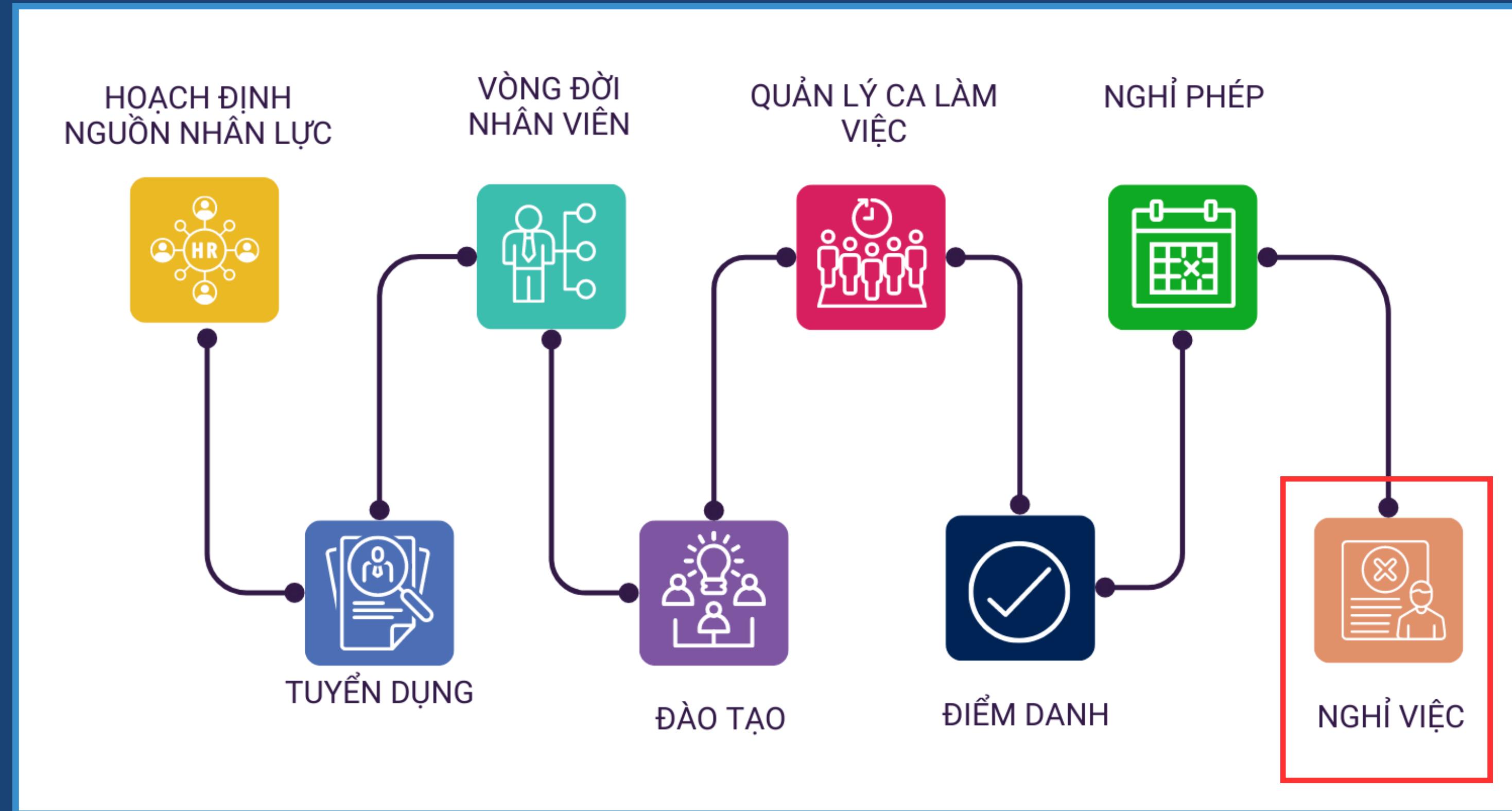
\*based on feedback from over 150,000 exit interviews



Mỗi **quyết định rời bỏ** của nhân viên đều  
bị ảnh hưởng bởi nhiều **yếu tố** như: **mức  
độ hài lòng công việc, điều kiện làm  
việc, cơ hội phát triển**, hay các **yếu tố cá  
nhân khác**.

NGUYỄN MẠNH THỊNH

# QUY TRÌNH QUẢN TRỊ NHÂN SỰ



NGUYỄN MẠNH THỊNH

# MỤC TIÊU



Tập trung vào việc **khai thác dữ liệu** liên quan đến tình trạng nhân viên rời bỏ

→ Tìm hiểu được **nguyên nhân** dẫn tới **nghỉ việc** và có thể **ra quyết định** và triển khai các **chiến lược phù hợp** để **giữ chân nhân viên**.

➤ Đánh giá các tác động đến tình trạng nhân viên rời bỏ

➤ Phân nhóm các nhân viên có đặc điểm chung đặc biệt

➤ Tìm ra các mối liên hệ có thể dẫn tới quyết định rời bỏ

➤ Đề xuất các giải pháp cải thiện tình trạng rời bỏ doanh nghiệp

# MÔ TẢ DỮ LIỆU

Gồm 13423 dòng và 39 cột

	EmployeeID	JoiningYear	Age	BusinessTravel	DailyRate	Department	DistanceFromHome	EducationField	EmployeeCount	EmployeeNumber	...	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion
0	100000	2005	57	Travel_Rarely	164	Corporate Functions	22	Doctorate	1	217	...	14	4	5
1	100001	2006	52	Travel_Rarely	265	Corporate Functions	19	Doctorate	1	519	...	12	4	5
2	100002	2006	53	Travel_Rarely	607	Corporate Functions	2	Doctorate	1	1572	...	12	3	1
3	100003	2006	54	Travel_Rarely	215	Corporate Functions	19	Diploma	1	309	...	13	4	5
4	100004	2007	57	Travel_Rarely	285	Marketing	2	Diploma	1	828	...	11	5	5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
13418	113418	2021	36	Travel_Frequently	1266	Corporate Functions	9	Diploma	1	240	...	1	1	1
13419	113419	2021	31	Travel_Rarely	424	Corporate Functions	6	Bachelors	1	1919	...	1	1	1
13420	113420	2021	36	Travel_Rarely	927	Product	5	Diploma	1	108	...	1	1	1
13421	113421	2021	58	Travel_Rarely	1200	Sales	14	Bachelors	1	1836	...	1	1	1
13422	113422	2021	40	Travel_Rarely	734	Sales	16	Doctorate	1	832	...	1	1	1

13423 rows x 39 columns

NGUYỄN MẠNH THỊNH

# MÔ TẢ DỮ LIỆU

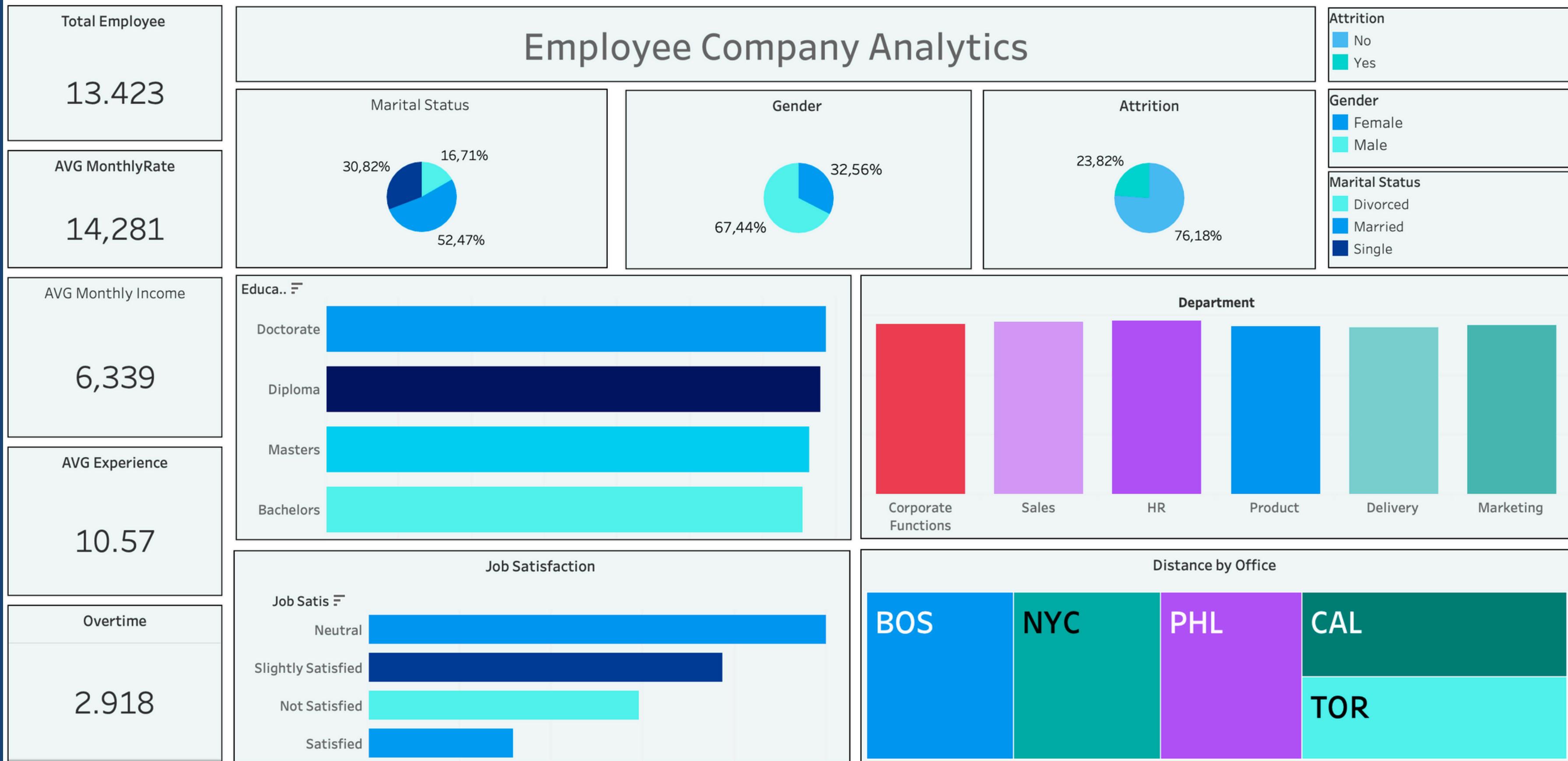
Tên cột	Loại	Định nghĩa	Ghi chú
EmployeeID	Numeric	Mã số nhân viên, duy nhất cho mỗi nhân viên.	
JoiningYear	Numeric	Năm nhân viên gia nhập tổ chức.	
Age	Numeric	Tuổi của nhân viên.	
BusinessTravel	Categorical	Thông tin về mức độ đi công tác của nhân viên.	Travel_Rarely - Travel_Frequently -Other
DailyRate	Numeric	Mức lương hàng ngày của nhân viên.	
Department	Categorical	Phòng ban mà nhân viên làm việc.	Department names
DistanceFromHome	Numeric	Khoảng cách từ nhà đến nơi làm việc của nhân viên.	
EducationField	Categorical	Lĩnh vực học vấn của nhân viên.	

Tên cột	Loại	Định nghĩa	Ghi chú
<b>EmployeeCount</b>	<b>Numeric</b>	Số lượng nhân viên (có thể là hằng số).	
<b>EmployeeNumber</b>	<b>Numeric</b>	Một mã số duy nhất khác cho mỗi nhân viên.	
<b>EnvironmentSatisfaction</b>	<b>Categorical</b>	Sự hài lòng của nhân viên về môi trường làm việc.	1, 2, 3, 4, 5
<b>Gender</b>	<b>Categorical</b>	Giới tính của nhân viên.	Male, Female
<b>HourlyRate</b>	<b>Numeric</b>	Mức lương theo giờ của nhân viên.	
<b>JobInvolvement</b>	<b>Categorical</b>	Mức độ tham gia công việc của nhân viên.	
<b>JobSatisfaction</b>	<b>Categorical</b>	Mức độ hài lòng với công việc.	1, 2, 3, 4, 5
<b>MaritalStatus</b>	<b>Categorical</b>	Tình trạng hôn nhân của nhân viên.	Yes, No
<b>MonthlyIncome</b>	<b>Numeric</b>	Mức lương hàng tháng của nhân viên.	
<b>MonthlyRate</b>	<b>Numeric</b>	Mức lương hàng tháng của nhân viên.	
<b>NumCompaniesWorked</b>	<b>Numeric</b>	Số công ty mà nhân viên đã làm việc.	

Tên cột	Loại	Định nghĩa	Ghi chú
<b>Over18</b>	<b>Numeric</b>	Chỉ rõ nhân viên có trên 18 tuổi hay không.	
<b>OverTime</b>	<b>Categorical</b>	Chỉ rõ nhân viên có làm thêm giờ hay không.	Yes/ No
<b>PercentSalaryHike</b>	<b>Numeric</b>	Tăng lương theo tỷ lệ phần trăm của nhân viên.	
<b>PerformanceRating</b>	<b>Categorical</b>	Đánh giá hiệu suất công việc của nhân viên.	1, 2, 3, 4, 5
<b>RelationshipSatisfaction</b>	<b>Categorical</b>	Mức độ hài lòng với các mối quan hệ công việc.	1, 2, 3, 4, 5
<b>StandardHours</b>	<b>Numeric</b>	Số giờ làm việc tiêu chuẩn.	
<b>StockOptionLevel</b>	<b>Categorical</b>	Mức độ quyền chọn cổ phiếu của nhân viên.	
<b>TotalWorkingYears</b>	<b>Numeric</b>	Tổng số năm làm việc của nhân viên.	
<b>TrainingTimesLastYear</b>	<b>Numeric</b>	Số buổi đào tạo mà nhân viên tham gia trong năm trước.	
<b>WorkLifeBalance</b>	<b>Categorical</b>	Sự cân bằng giữa công việc và cuộc sống cá nhân.	1, 2, 3, 4, 5

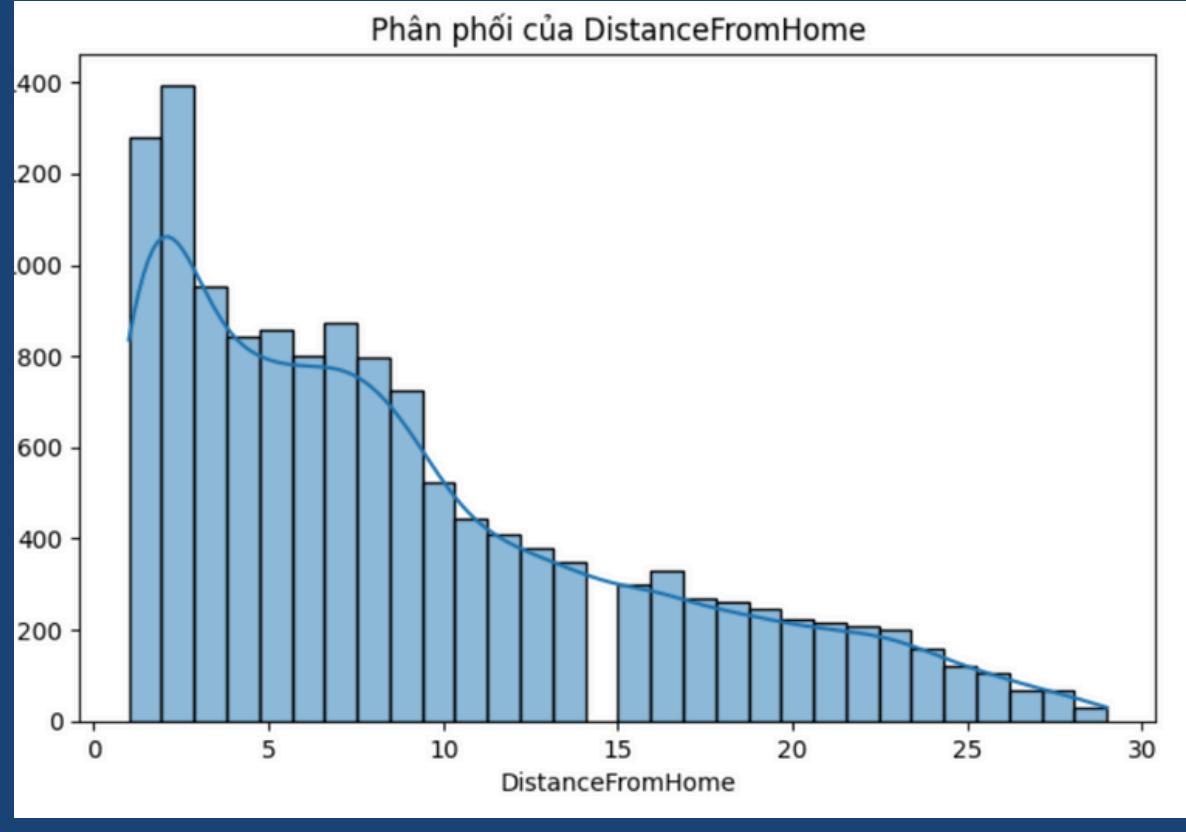
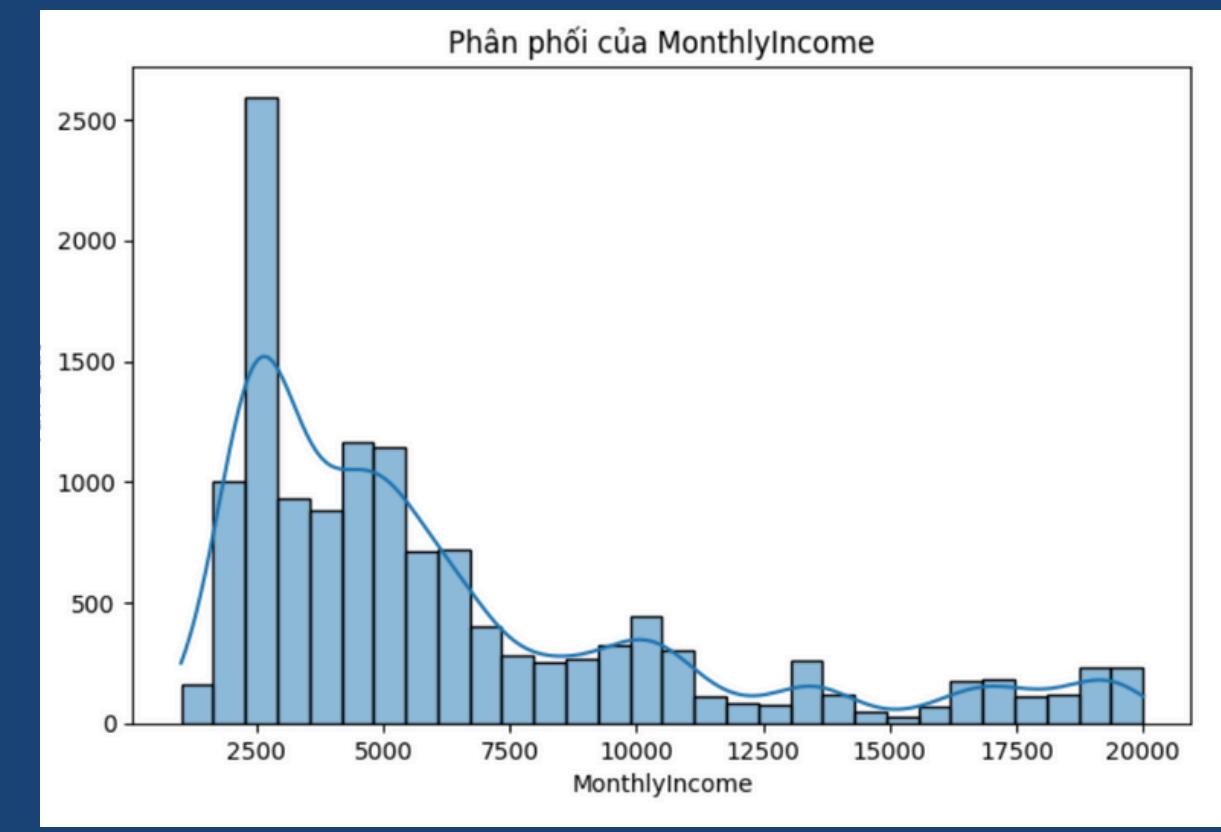
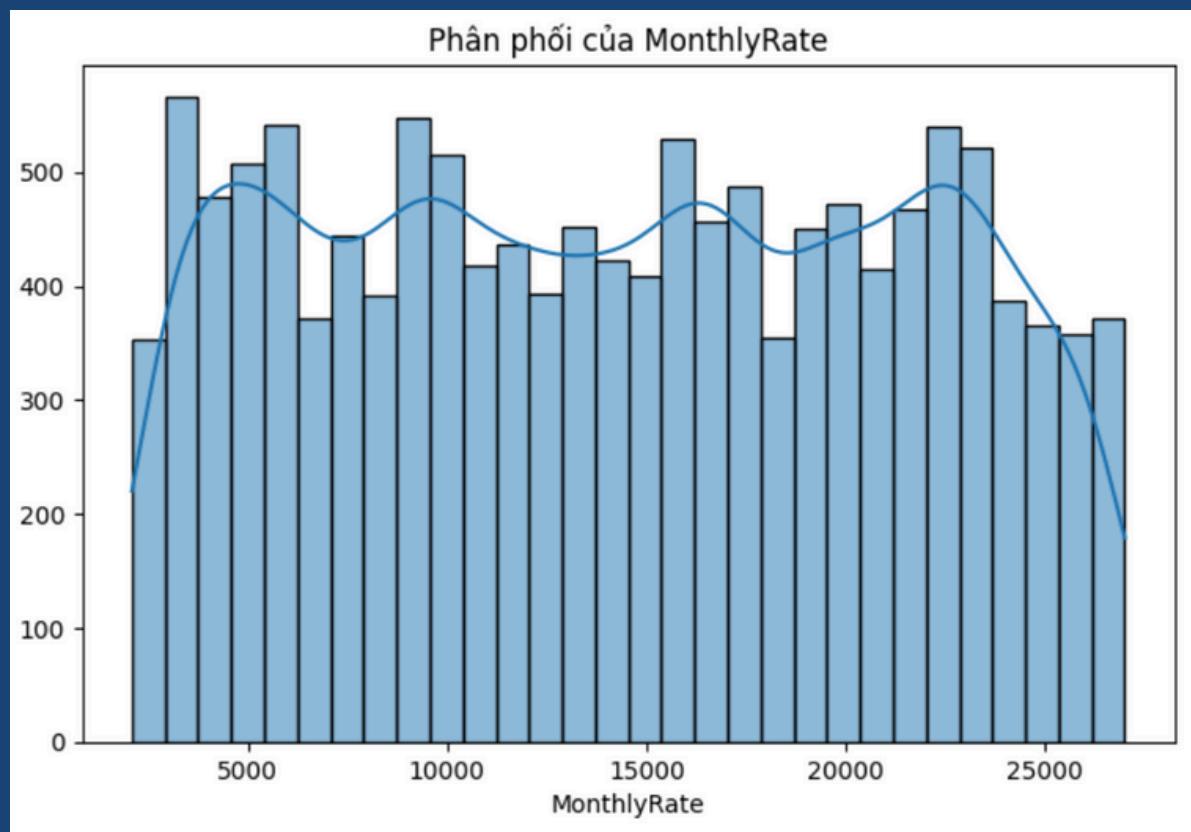
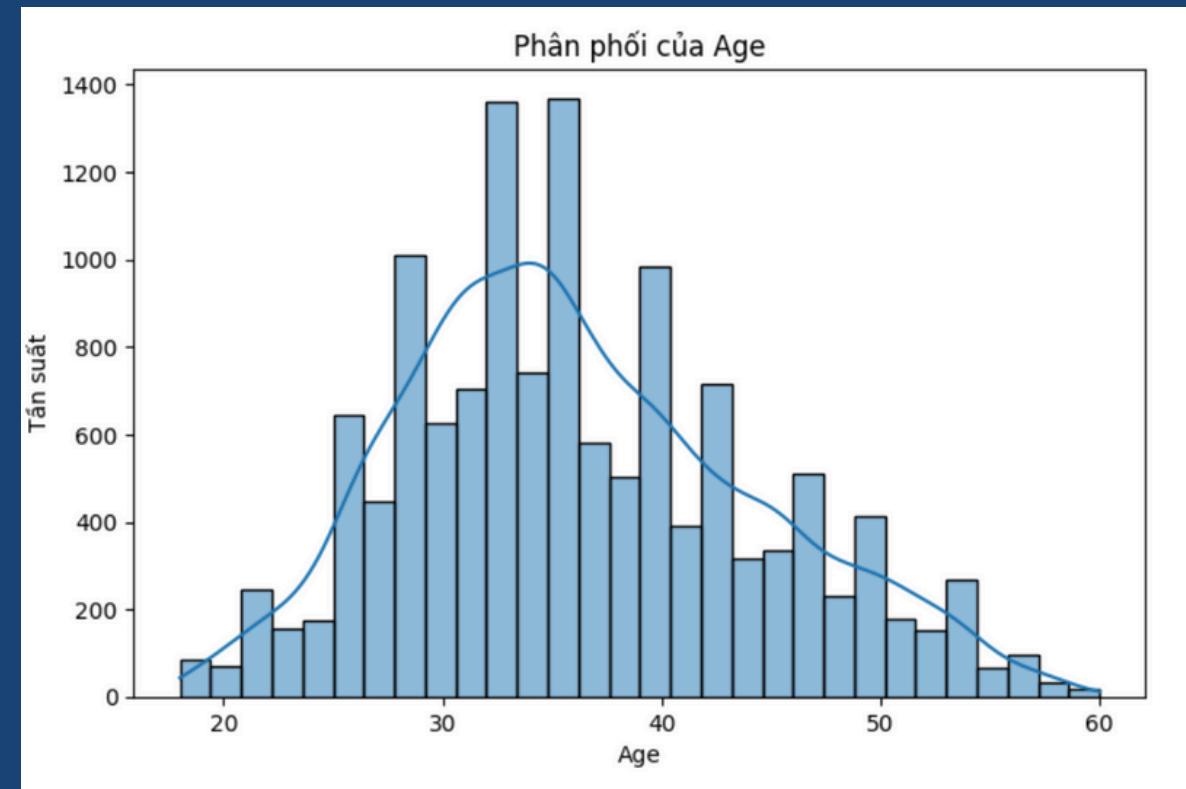
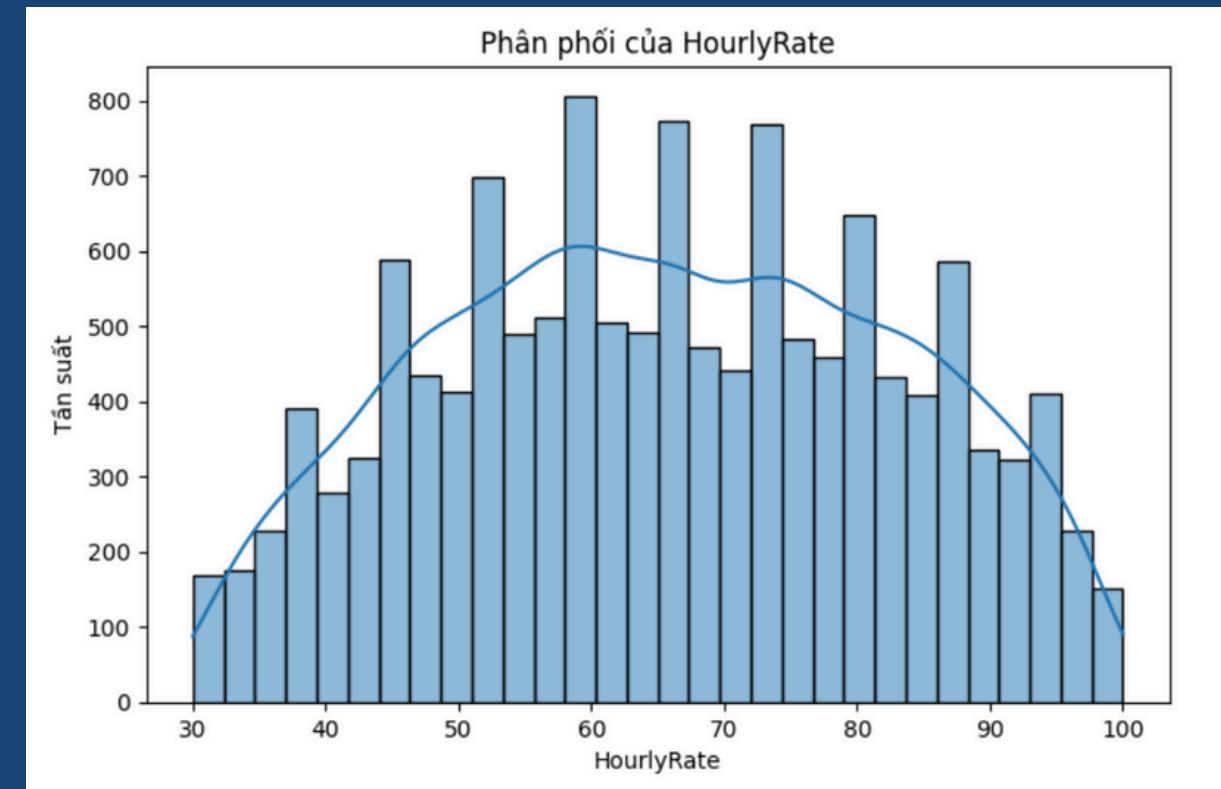
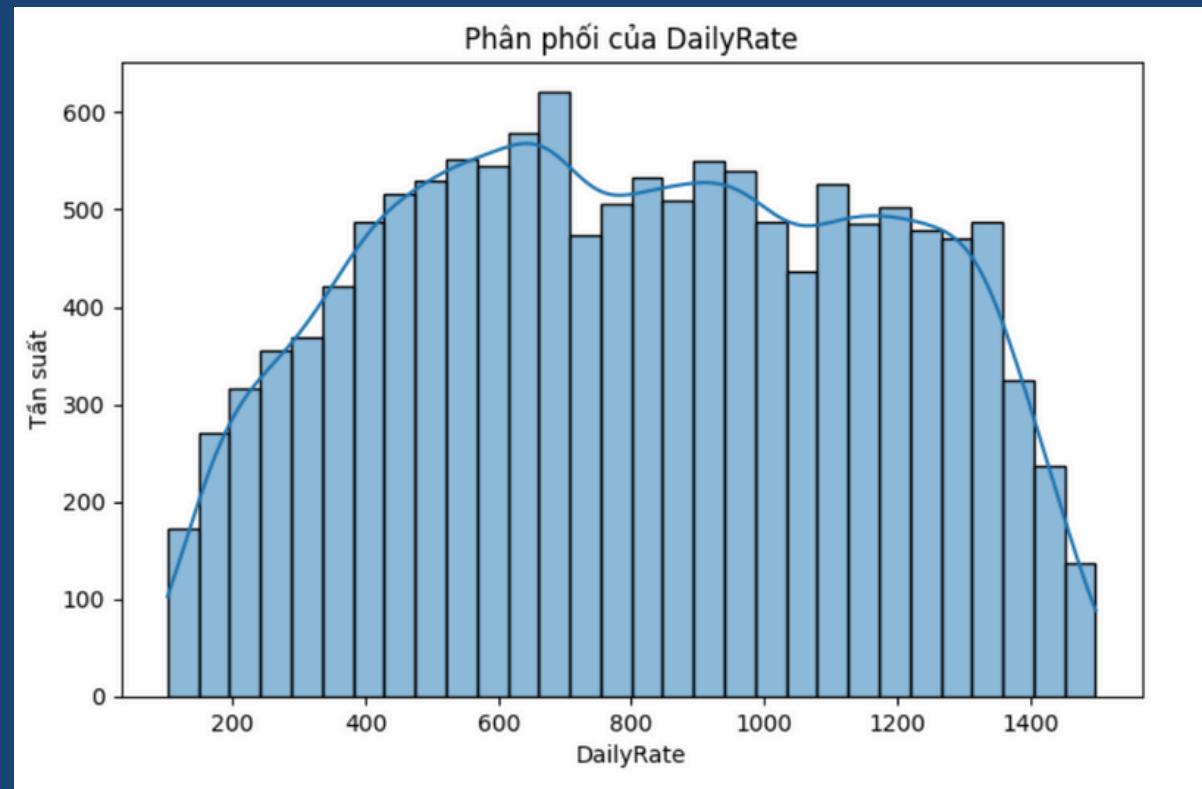
Tên cột	Loại	Định nghĩa	Ghi chú
<b>YearsAtCompany</b>	<b>Numeric</b>	Số năm làm việc tại công ty.	
<b>YearsInCurrentRole</b>	<b>Numeric</b>	Số năm làm việc trong vai trò hiện tại.	
<b>YearsSinceLastPromotion</b>	<b>Numeric</b>	Số năm kể từ lần thăng chức gần nhất.	
<b>YearsWithCurrManager</b>	<b>Numeric</b>	Số năm làm việc với quản lý hiện tại.	
<b>Attrition</b>	<b>Categorical</b>	Chỉ ra liệu nhân viên đã rời công ty hay chưa.	Yes / No
<b>LeavingYear</b>	<b>Numeric</b>	Năm nhân viên rời công ty.	
<b>Reason</b>	<b>Categorical</b>	Lý do nhân viên rời công ty.	
<b>RelievingStatus</b>	<b>Categorical</b>	Tình trạng giải quyết khi nhân viên rời công ty.	
<b>office_code</b>	<b>Categorical</b>	Mã văn phòng.	
<b>JobLevel_updated</b>	<b>Categorical</b>	Cập nhật cấp bậc công việc.	

# DASHBOARD



# EDA

NGUYỄN MẠNH THỊNH





# MÔ HÌNH PHÂN LOẠI

NGUYỄN MẠNH THỊNH

# MỤC TIÊU PHÂN LỚP

Xác định xem một nhân viên có khả năng rời bỏ công ty hay không dựa trên đặc điểm của nhân viên, như mức lương, thâm niên công tác, sự hài lòng với công việc, hoặc các yếu tố cá nhân khác.

Giúp doanh nghiệp chủ động phát hiện những nhân viên có nguy cơ cao và can thiệp kịp thời.



NGUYỄN MẠNH THỊNH

# TIỀN XỬ LÍ DỮ LIỆU

Dropping  
Encryption data  
Data normalisation  
Imbalanced Data Handling

NGUYỄN MẠNH THỊNH

# Dropping Columns

**Loại bỏ các cột không cần thiết:** EmployeeID, EmployeeNumber, EmployeeCount, StandardHours, Over18, office\_code, RelievingStatus, Reason, JobLevel\_updated, JoiningYear, LeavingYear

# Encryption data

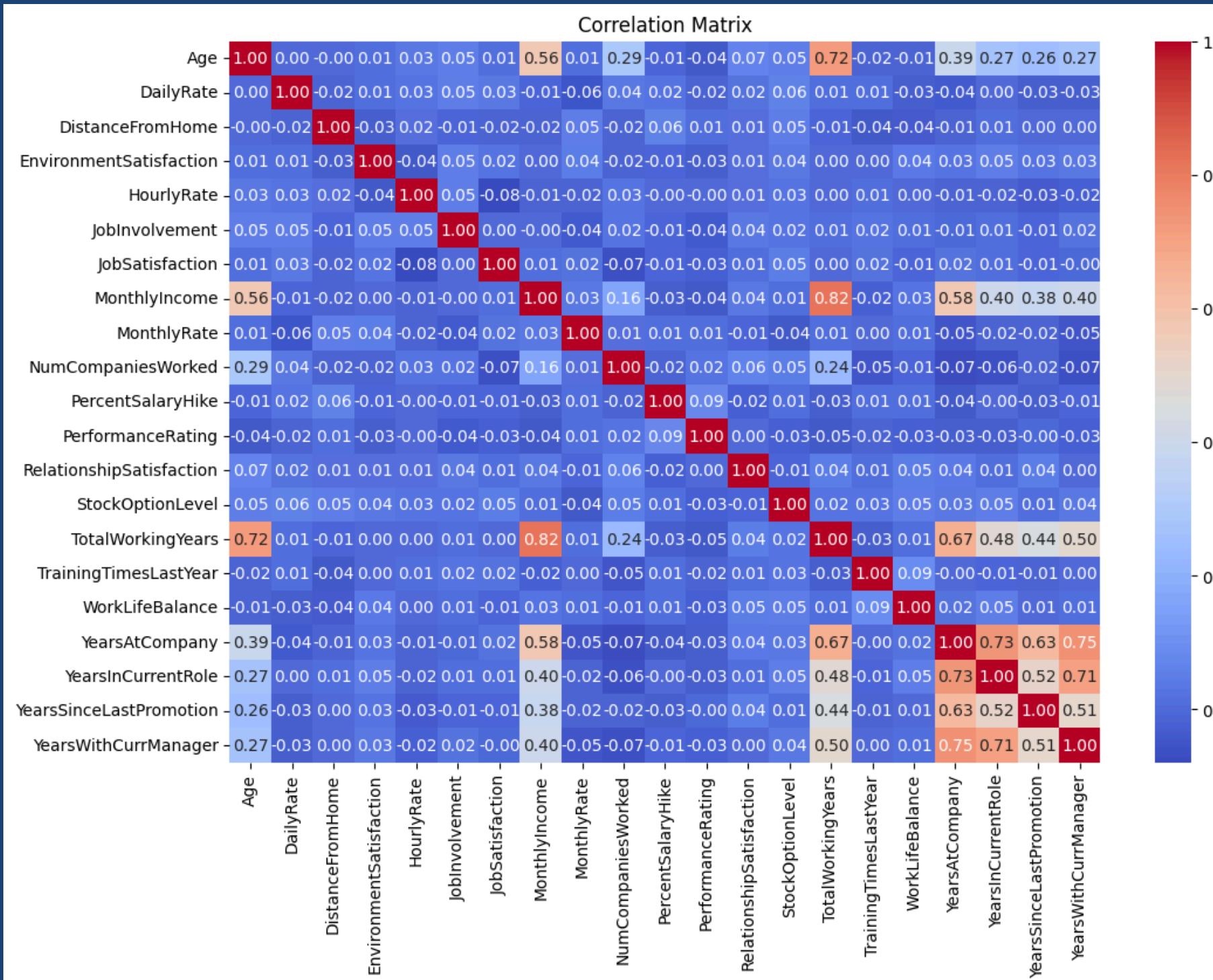
Tiến hành mã hoá dữ liệu đối với các biến phân loại (**categorical**) bằng thư viện **Label Encoding**

**Bao gồm các cột:** Attrition, Gender, OverTime, BusinessTravel, MaritalStatus, Department, EducationField

Ví dụ: Attrition : Yes - No => 0 - 1

BusinessTravel: Travel\_Rarely - Travel\_Frequently - Other => 1 - 2 - 3

# Dropping columns

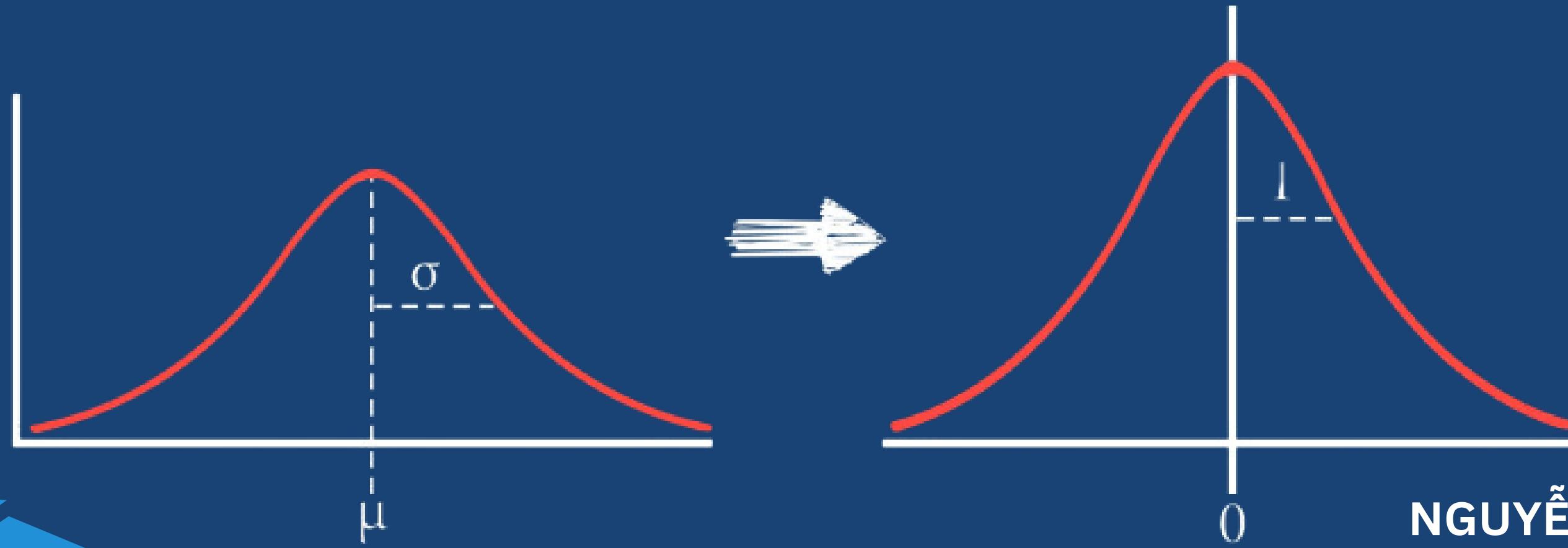


Cân nhắc bỏ các cột có chỉ số tương quan cao với nhau để xử lý hiện tượng **đa cộng tuyến**

Giữ lại cột: **TotalWorkingYear**  
Bỏ cột: '**MonthlyIncome**, **YearAtCompany**,  
**YearWithCurrManager**, **YearsInCurrentRole**,  
**YearsSinceLastPromotion**, **Age**'

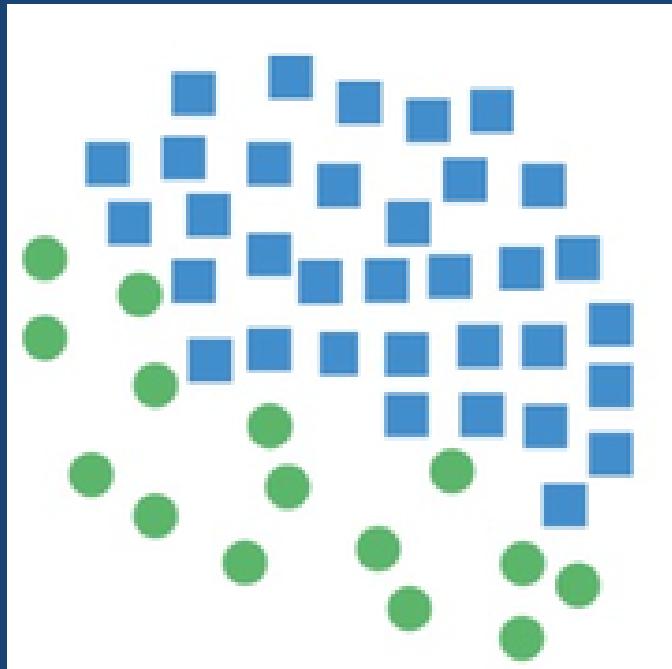
# Data scaling

- Sử dụng phương pháp phân phối chuẩn (**Z-score Scaling**) nhằm chuẩn hóa dữ liệu
  - Đảm bảo rằng dữ liệu có phân phối chuẩn với giá trị trung bình bằng **0** và độ lệch chuẩn bằng **1**
- > Nhằm giúp mô hình phân loại hội tụ nhanh hơn và ổn định hơn trên tập dữ liệu

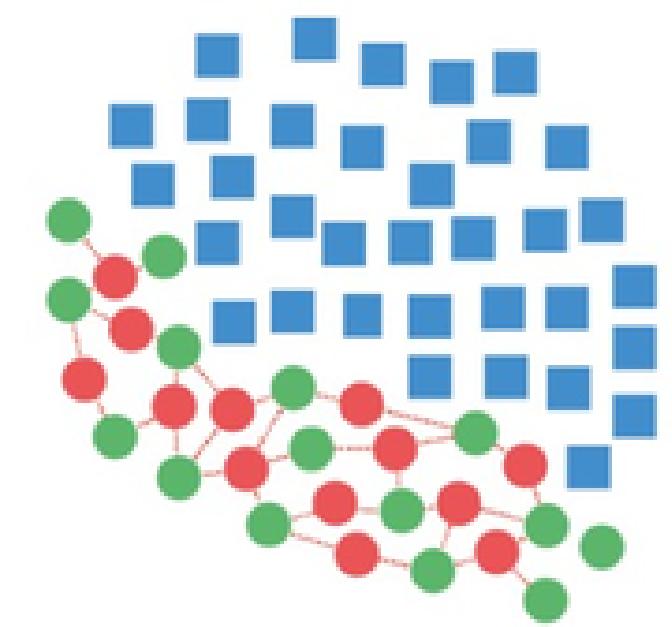


# Handling imbalanced data

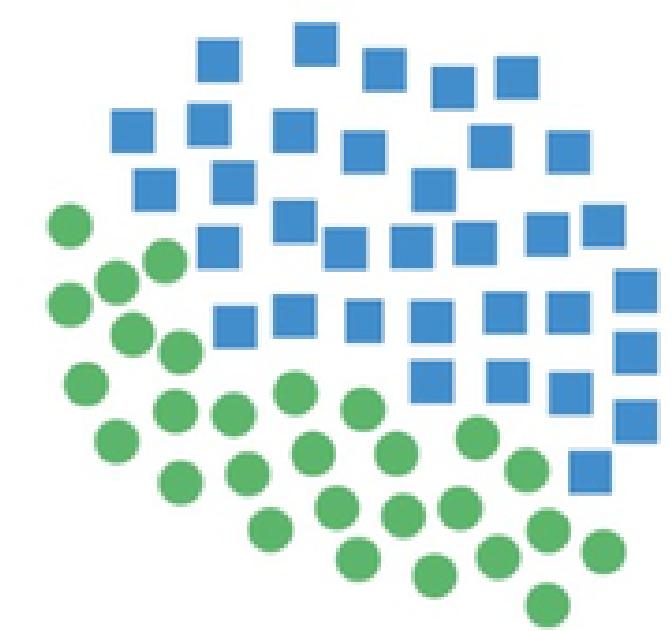
- Sử dụng phương pháp xử lí mất cân bằng **SMOTE oversampling data**



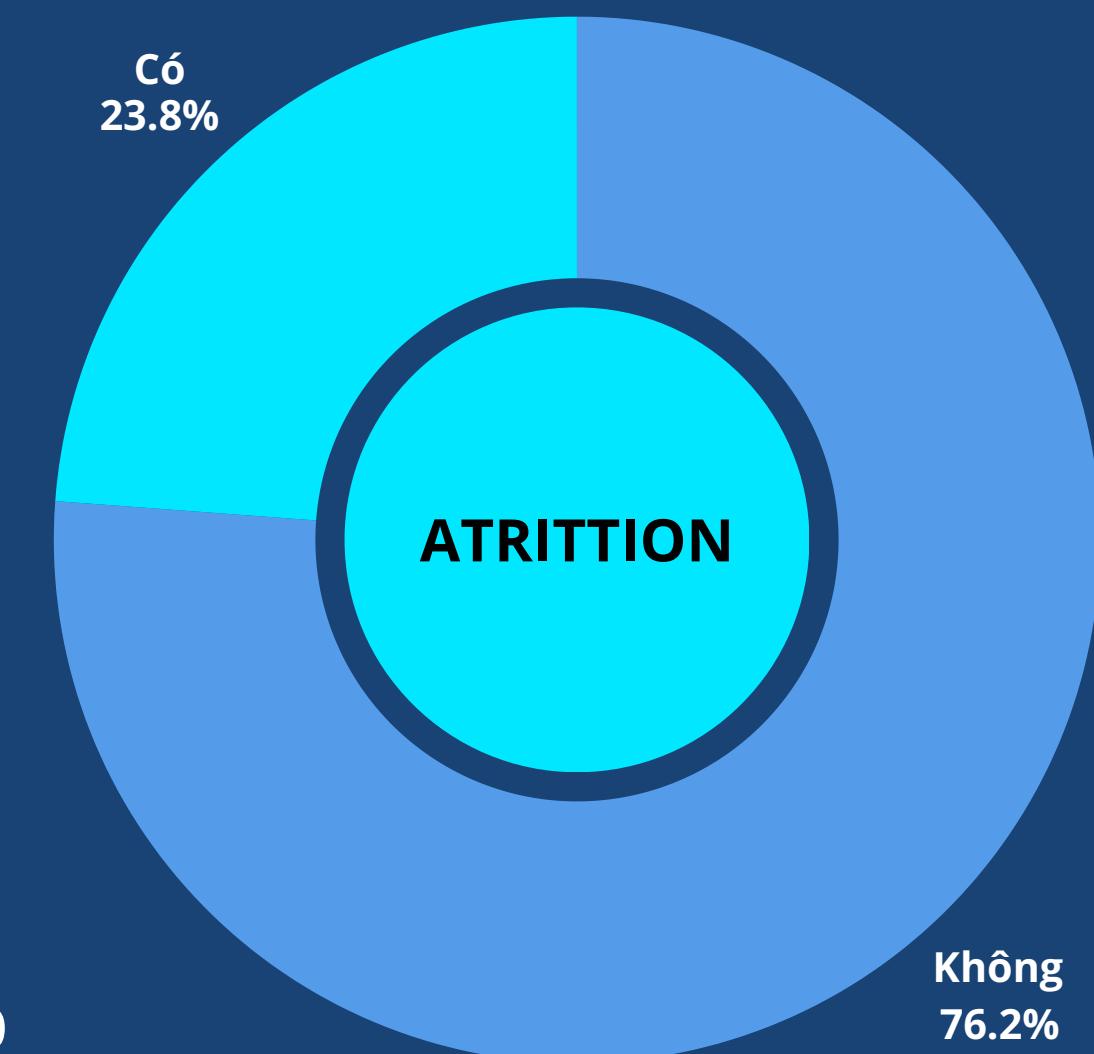
Imbalanced Dataset



Generating new synthetic samples



SMOTE Dataset



=> **SMOTE** giúp giải quyết vấn đề mất cân bằng dữ liệu giữa hai lớp bằng cách tăng cường mẫu cho lớp thiểu số

# Data Splitting

Để đạt được mục tiêu dự đoán phân lớp **Attrition**, ứng dụng phương pháp **Data Spliting** chia 2 tập **Train - Test**  
**Nhằm đánh giá khả năng tổng quát hóa của mô hình**

Xác định biến độc lập (**features**) và biến mục tiêu (**target**)

- **X = df.drop(columns=['Attrition'])** - Loại bỏ cột 'Attrition' vì nó là biến mục tiêu
- **y = df['Attrition']** - 'Attrition' là biến mục tiêu

**Chia Train/ Test theo phương pháp (80/20)**

**X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.20, random\_state=42)**

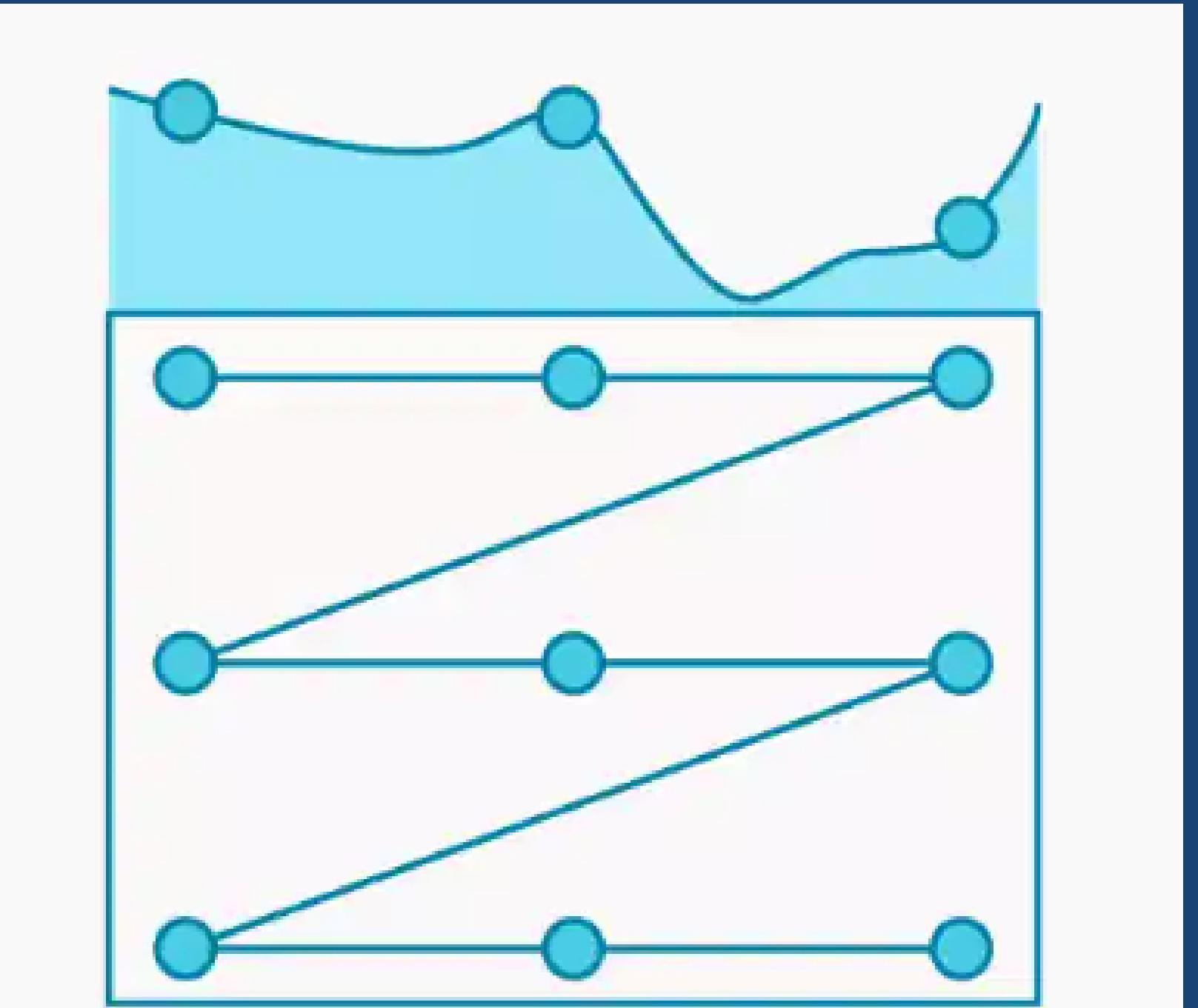
————→ Sau đó tiến hành thực hiện phương pháp **Oversampling** là **SMOTE** để xử lí mất cân bằng dữ liệu

# GRIDSEARCH

**Tìm kiếm siêu tham số tốt nhất:**

Duyệt qua toàn bộ không gian các siêu tham số mà bạn chỉ định, thử nghiệm mọi tổ hợp để tìm ra bộ siêu tham số tối ưu nhất giúp mô hình đạt hiệu suất cao nhất.

Đảm bảo tìm được siêu tham số  
tốt nhất trong mô hình cần tìm



# THUẬT TOÁN

1

LOGISTIC  
REGRESSION

2

RANDOM  
FOREST

3

DECISION  
TREE

4

K-NEAREST  
NEIGHBORS

5

XGBOOST

Lê Thúy Quỳnh

# LOGISTIC REGRESSION (IMBALANCED - SMOTE)

- Sử dụng LogisticRegression từ thư viện **sklearn.linear\_model** để khởi tạo mô hình hồi quy logistic
- Phương pháp **fit()** với các dữ liệu huấn luyện để huấn luyện mô hình.

Predicted		
Actual	0	1
0	1944	94
1	280	367

Predicted		
Actual	0	1
0	1719	319
1	151	496

	precision	recall	f1 -score	support
0	0.87	0.95	0.91	2038
1	0.80	0.57	0.66	647
accuracy			0.86	2685
macro avg	0.84	0.76	0.79	2685
weighted avg	0.86	0.86	0.85	2685

	precision	recall	f1 -score	support
0	0.92	0.84	0.88	2038
1	0.61	0.77	0.68	647
accuracy			0.82	2685
macro avg	0.76	0.81	0.78	2685
weighted avg	0.84	0.82	0.83	2685

# RANDOM FOREST (IMBALANCED - SMOTE)

Predicted

Actual	2031	7
73	574	

	precision	recall	f1 - score	support
0	0.97	1.00	0.98	2038
1	0.99	0.89	0.93	647
accuracy			<b>0.97</b>	2685
macro avg	0.98	0.94	0.96	2685
weighted avg	0.97	0.97	0.97	2685

Bộ siêu tham số (**GridSearchCV**)

	No Balance	SMOTE
max_depth	None	None
min_samples_leaf	1	1
min_samples_split	2	2
n_estimators	300	200

Predicted

Actual	2026	12
70	577	

	precision	recall	f1 - score	support
0	0.97	0.99	0.98	2038
1	0.98	0.89	0.93	647
accuracy			<b>0.97</b>	2685
macro avg	0.97	0.94	0.96	2685
weighted avg	0.97	0.97	0.97	2685

# DECISION TREE (IMBALANCED - SMOTE)

**Predicted**

Actual	2000	38		
120	527			
	precision	recall	f1-score	support
0	0.94	0.98	0.96	2038
1	0.93	0.81	0.87	647
accuracy			<b>0.94</b>	2685
macro avg	0.94	0.90	0.92	2685
weighted avg	0.94	0.94	0.94	2685

Bộ siêu tham số (**GridSearchCV**)

	No Balance	SMOTE
criterion	gini	entropy
min_samples_leaf	2	1
min_samples_split	5	2
max_depth	10	20

**Predicted**

Actual	1977	61		
82	565			
	precision	recall	f1-score	support
0	0.96	0.97	0.97	2038
1	0.90	0.87	0.89	647
accuracy			<b>0.95</b>	2685
macro avg	0.93	0.92	0.93	2685
weighted avg	0.95	0.95	0.95	2685

# KNN (IMBALANCED - SMOTE)

Predicted	
Actual	Predicted
2032	6
41	606

Bộ siêu tham số (**GridSearchCV**)

	precision	recall	f1 - score	support
0	0.98	1.00	0.99	2038
1	0.99	0.94	0.96	647
accuracy			<b>0.98</b>	2685
macro avg	0.99	0.97	0.98	2685
weighted avg	0.98	0.98	0.98	2685

	No Balance/ SMOTE
<b>metric</b>	manhattan
<b>n_neighbors</b>	3
<b>weights</b>	distance

	precision	recall	f1 - score	support
0	<b>0.99</b>	<b>0.99</b>	<b>0.99</b>	2038
1	0.97	0.97	0.97	647
accuracy			<b>0.98</b>	2685
macro avg	0.98	0.98	0.98	2685
weighted avg	0.98	0.98	0.98	2685

# XGBOOST (IMBALANCED - SMOTE)

Predicted

	2034	4
Actual	63	584

	precision	recall	f1 - score	support
0	0.97	1.00	0.98	2038
1	0.99	0.90	0.95	647
accuracy			<b>0.98</b>	2685
macro avg	0.99	0.95	0.96	2685
weighted avg	0.98	0.98	0.97	2685

Bộ siêu tham số (**GridSearchCV**)

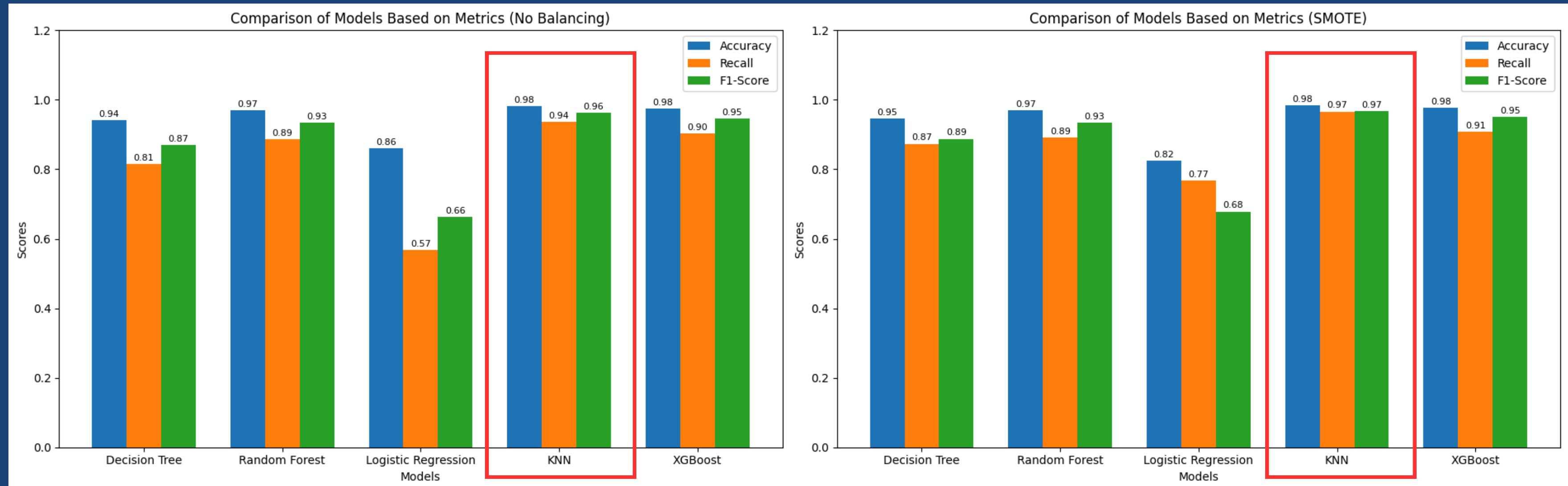
	No Balance/ SMOTE
<b>colsample_bytree</b>	0.8
<b>learning_rate</b>	0.2
<b>max_depth</b>	7
<b>n_estimators</b>	150
<b>subsample</b>	0.8

Predicted

Actual	2035	3
Actual	59	588

	precision	recall	f1 - score	support
0	<b>0.97</b>	1.00	<b>0.98</b>	2038
1	0.99	0.91	0.95	647
accuracy			<b>0.98</b>	2685
macro avg	0.98	0.95	0.97	2685
weighted avg	0.98	0.98	0.98	2685

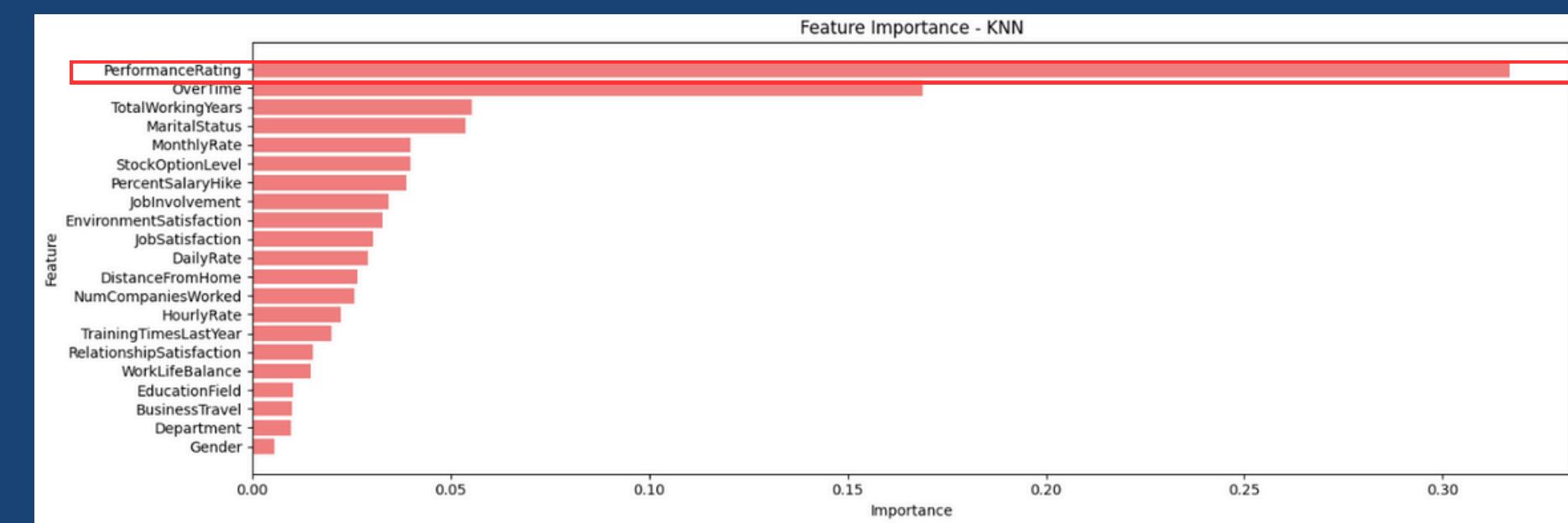
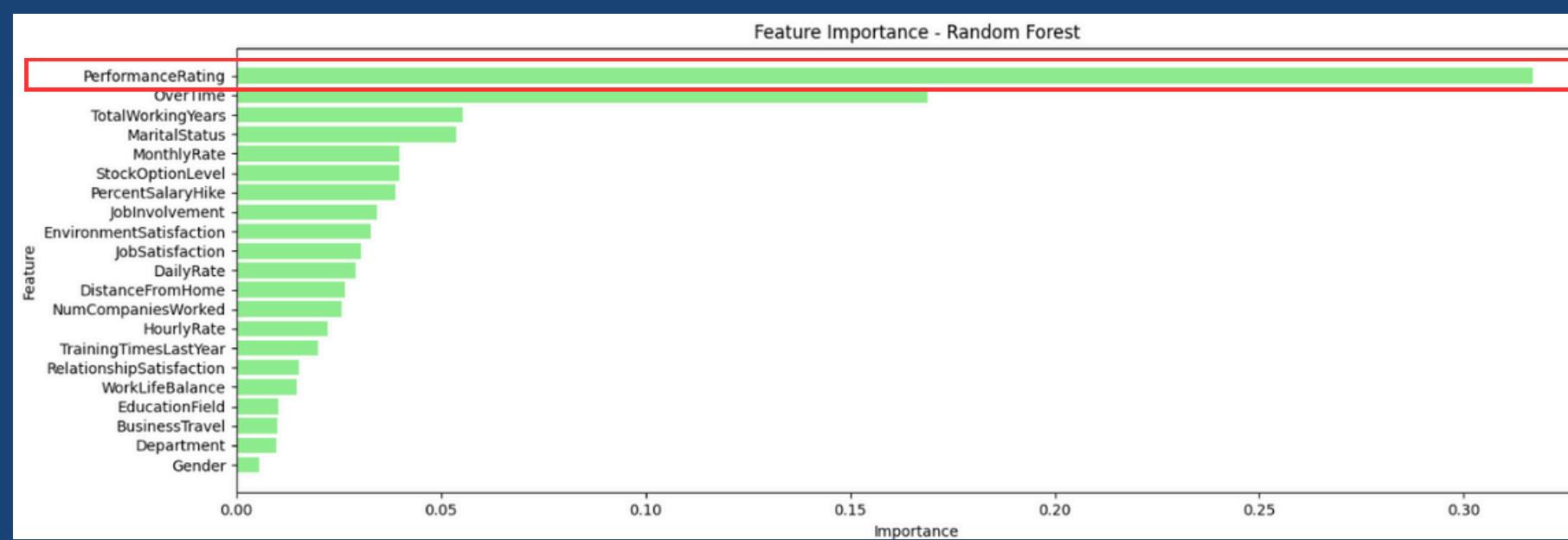
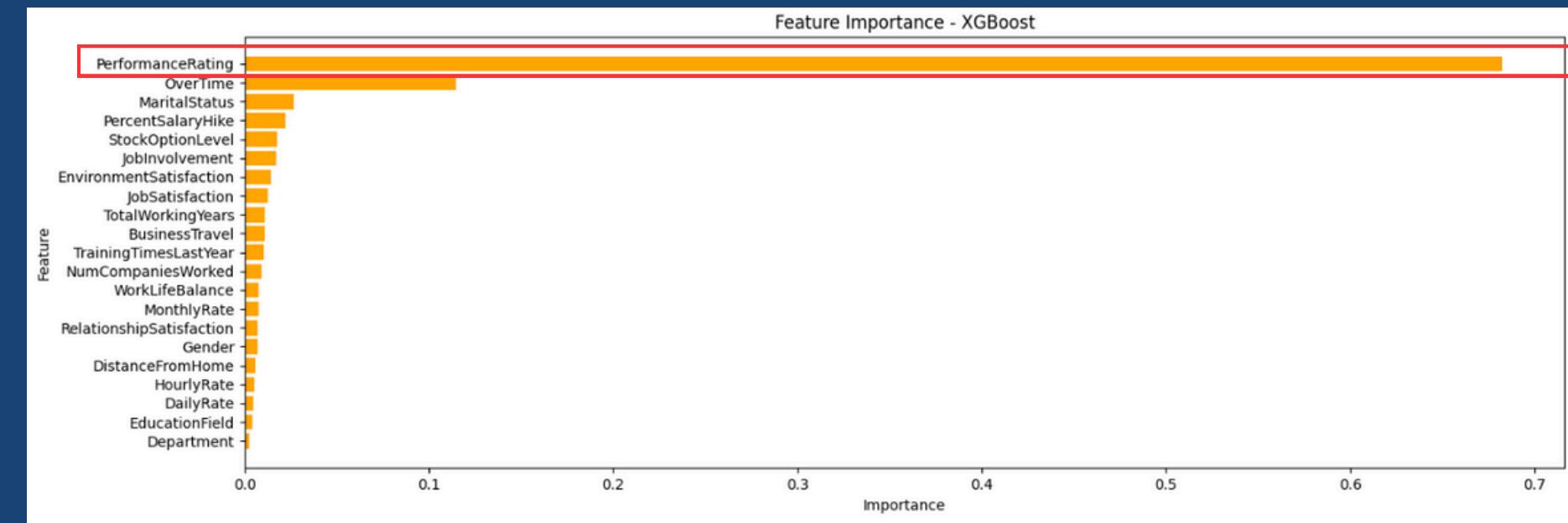
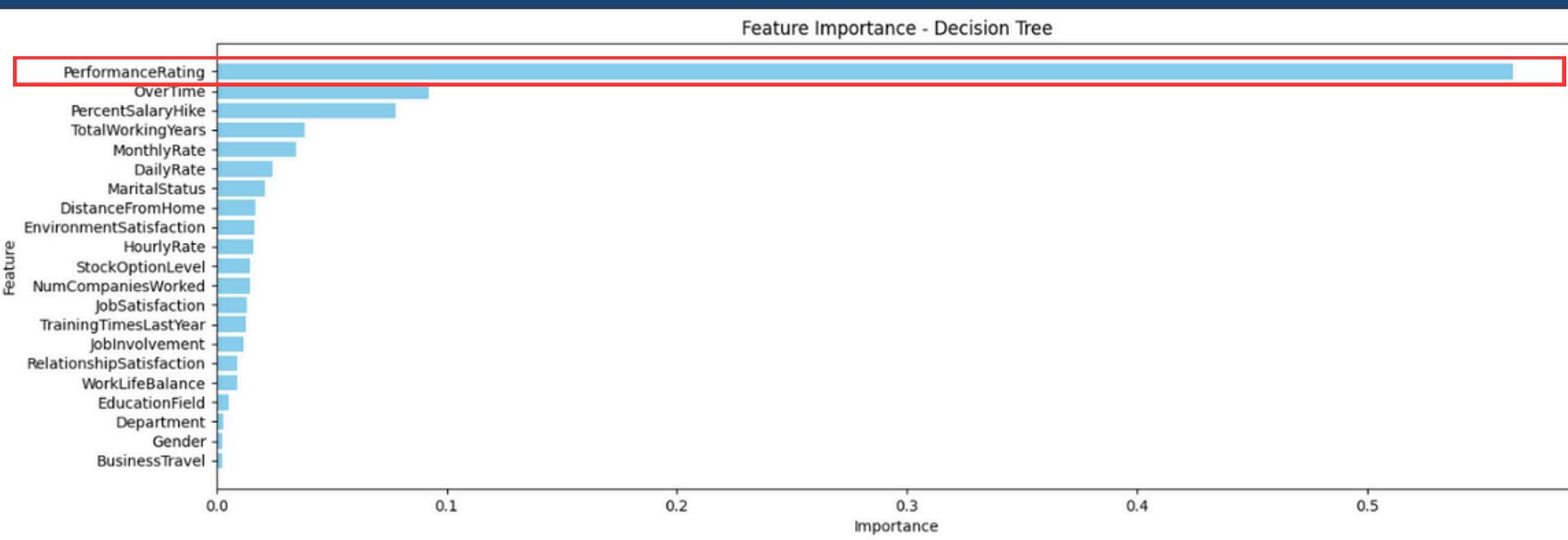
# SO SÁNH CHỈ SỐ CÁC MÔ HÌNH



Thông qua cả 2 phương pháp **xử lí mất cân bằng** và **chưa xử lí mất cân bằng**, nhận thấy được **KNN** trả về chỉ số đánh giá cao nhất trong các mô hình còn lại

# TRÍCH XUẤT ĐẶC TRƯNG QUAN TRỌNG

Bảng phương pháp **Mean Decrease in Impurity**

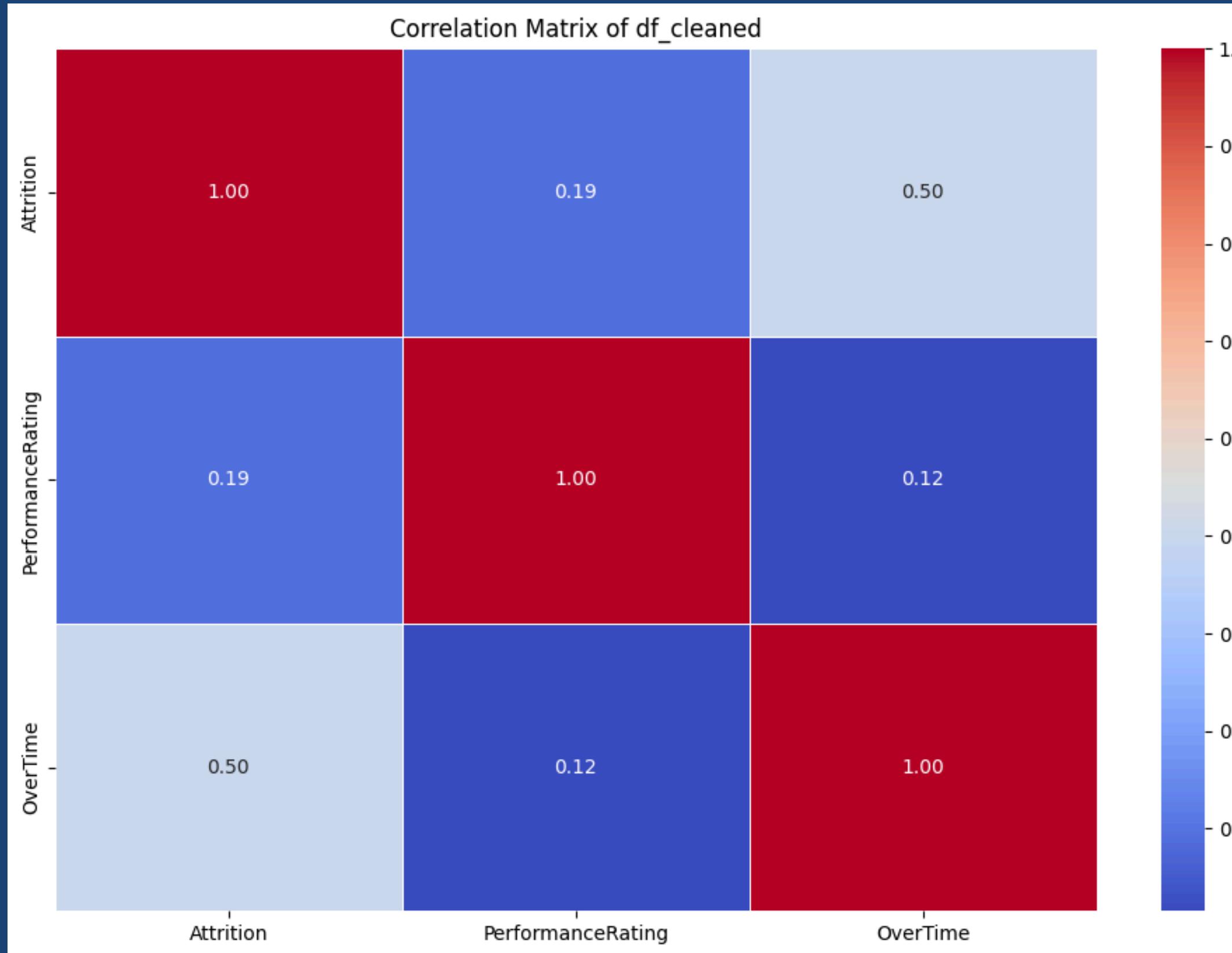


=> **PerformanceRating** (Đánh giá hiệu suất): Là yếu tố quan trọng nhất ảnh hưởng trong cả 4 mô hình.

=> Sau đó là **Overtime**: là yếu tố xếp quan trọng thứ hai ảnh hưởng tới 4 mô hình trên

# TRÍCH XUẤT ĐẶC TRƯNG QUAN TRỌNG

Đánh giá tương quan giữa hai thuộc tính **PerformanceRating**, **Overtime** với **Attrition**



3 biến **PerformanceRating**, **Overtime** và **Attrition** có độ tương quan lần lượt là 0.19, 0.5  
=> tuy rất yếu nhưng vẫn là **tương quan thuận**

- **Chứng tỏ**, các nhân viên càng có **Performance Rating** cao thì **càng có xu hướng nghỉ việc** hơn so với các thuộc tính còn lại
- **Các nhân viên** đang thường xuyên **làm việc tăng ca** thường **càng** có xu hướng nghỉ việc ở **công ty**

# ĐỀ XUẤT THAY ĐỔI

## Performance Rating

**Đảm bảo chính sách khen thưởng tương xứng, minh bạch:**

- Xây dựng cơ chế **khen thưởng công khai, minh bạch** dựa trên kết quả đánh giá hiệu suất.

**Tạo cơ hội phát triển và thăng tiến rõ ràng:**

- Xây dựng **lộ trình thăng tiến minh bạch**: cho thấy các tiêu chí và thời gian cần đạt được để lên vị trí cao hơn.
- Giao thêm các **dự án thử thách** giúp nhân viên hiệu suất cao **phát triển kỹ năng mới** và **giảm bớt sự nhảm chán** trong công việc.

**Giảm áp lực công việc:**

- **Điều chỉnh kỳ vọng công việc hợp lý** và tránh tình trạng giao quá nhiều nhiệm vụ cho nhân viên có hiệu suất cao.

## OverTime

**Triển khai chính sách giới hạn giờ tăng ca:**

- Quy định rõ số giờ làm thêm tối đa trong **ngày/tuần** (ví dụ: không quá 48 tiếng/tuần).

**Cân đối và phân bổ lại khối lượng công việc:**

- **Rà soát định kỳ** để xác định **phòng ban** nào có khối lượng **công việc quá tải** và **phân bổ** lại nguồn lực.

**Khuyến khích làm việc hiệu quả trong giờ hành chính:**

- **Đào tạo** nhân viên về **kỹ năng quản lý thời gian và công việc** hiệu quả.



# CLUSTERING

Nguyễn Thị Diễm Ly

# MỤC TIÊU PHÂN CỤM

Xác định các nhóm nhân viên có đặc điểm tương đồng, qua đó giúp doanh nghiệp hiểu rõ tình hình nhân sự và xây dựng các giải pháp phù hợp để nâng cao hiệu suất làm việc, đồng thời giảm thiểu tình trạng rời bỏ công ty.



Nguyễn Thị Diễm Ly

# THUẬT TOÁN

1

K-MEANS  
CLUSTERING

2

HIERARCHICAL  
CLUSTERING

3

DBSCAN  
CLUSTERING

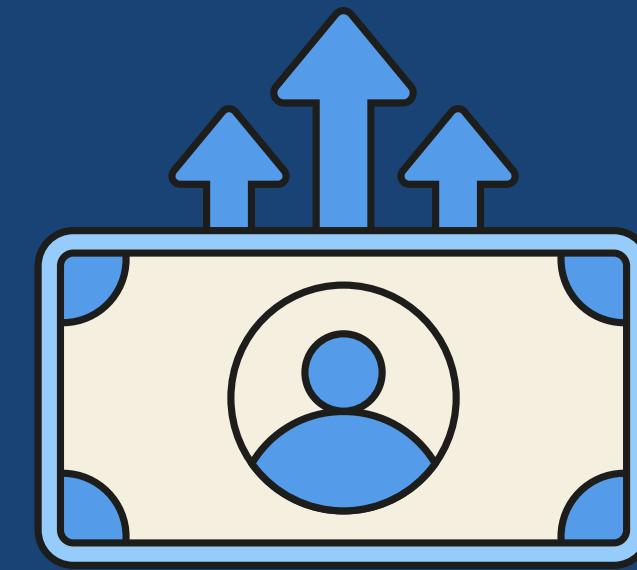
# LỰA CHỌN ĐẶC TRƯNG



KINH NGHIỆM  
LÀM VIỆC



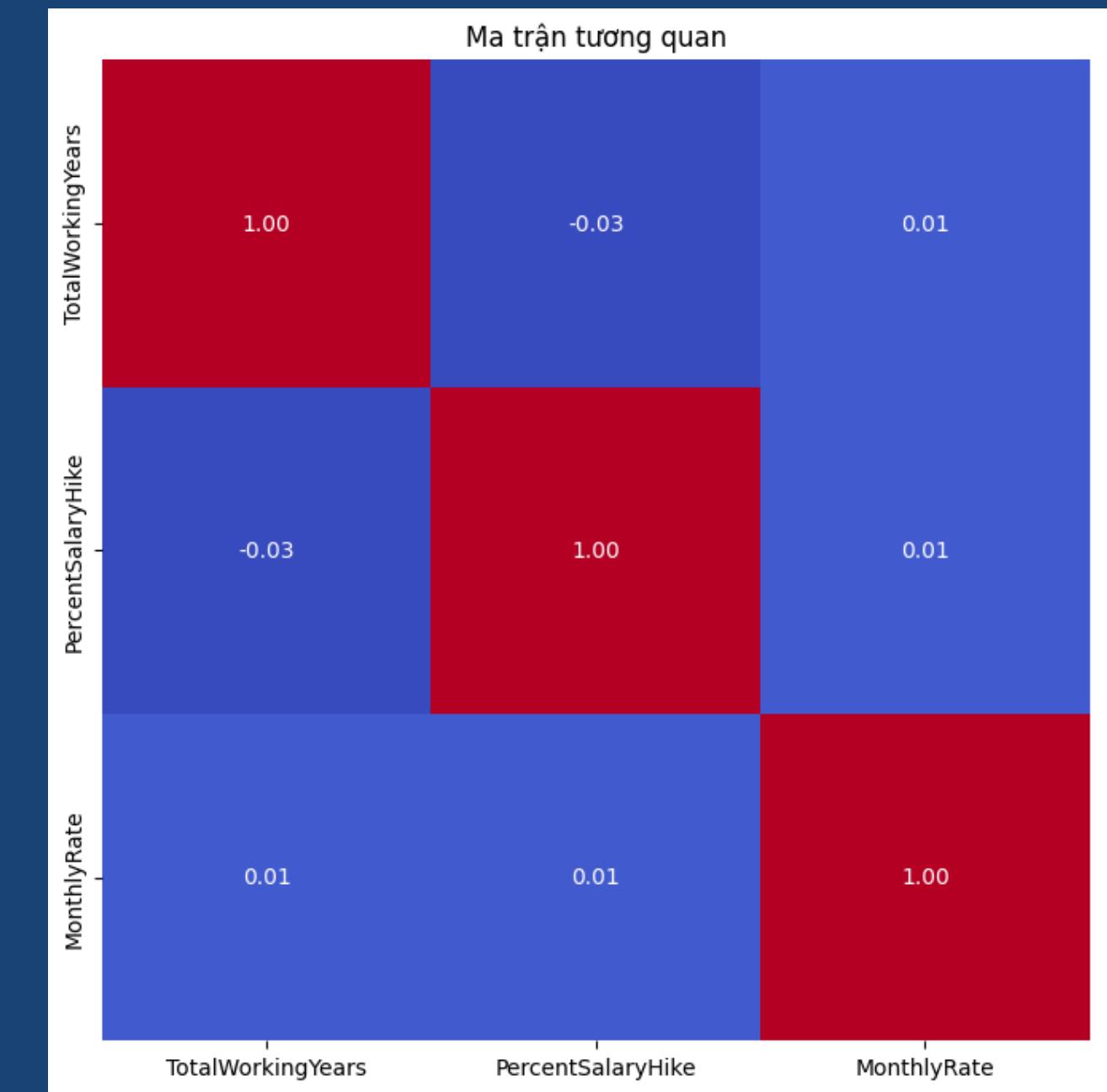
PHẦN TRĂM  
TĂNG LƯƠNG



MỨC ĐÓNG GÓP  
THEO THÁNG

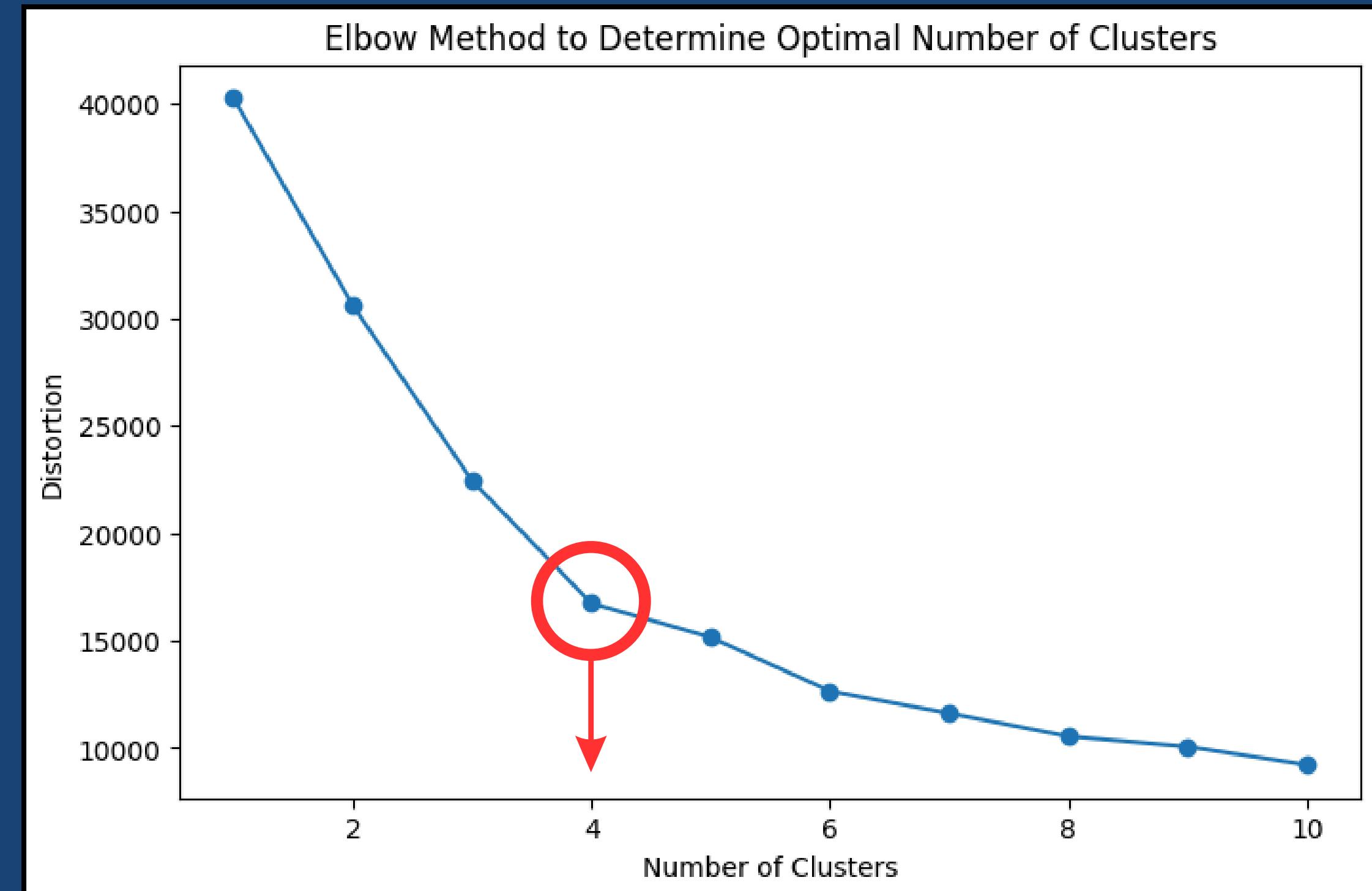
# LỰA CHỌN ĐẶC TRƯNG

```
# 1.Chọn đặc trưng để phân cụm  
features = ['TotalWorkingYears', 'PercentSalaryHike', 'MonthlyRate']  
df_features = df[features]  
  
# 2.Ma trận tương quan  
correlation_matrix = df_features.corr()  
plt.figure(figsize=(10, 8))  
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')  
plt.title('Ma trận tương quan')  
plt.show()  
  
# 3.Chuẩn hóa dữ liệu  
scaler = StandardScaler()  
features_scaled = scaler.fit_transform(df_features)
```



# K-MEANS

XÁC ĐỊNH  
SỐ CỤM TỐI ƯU



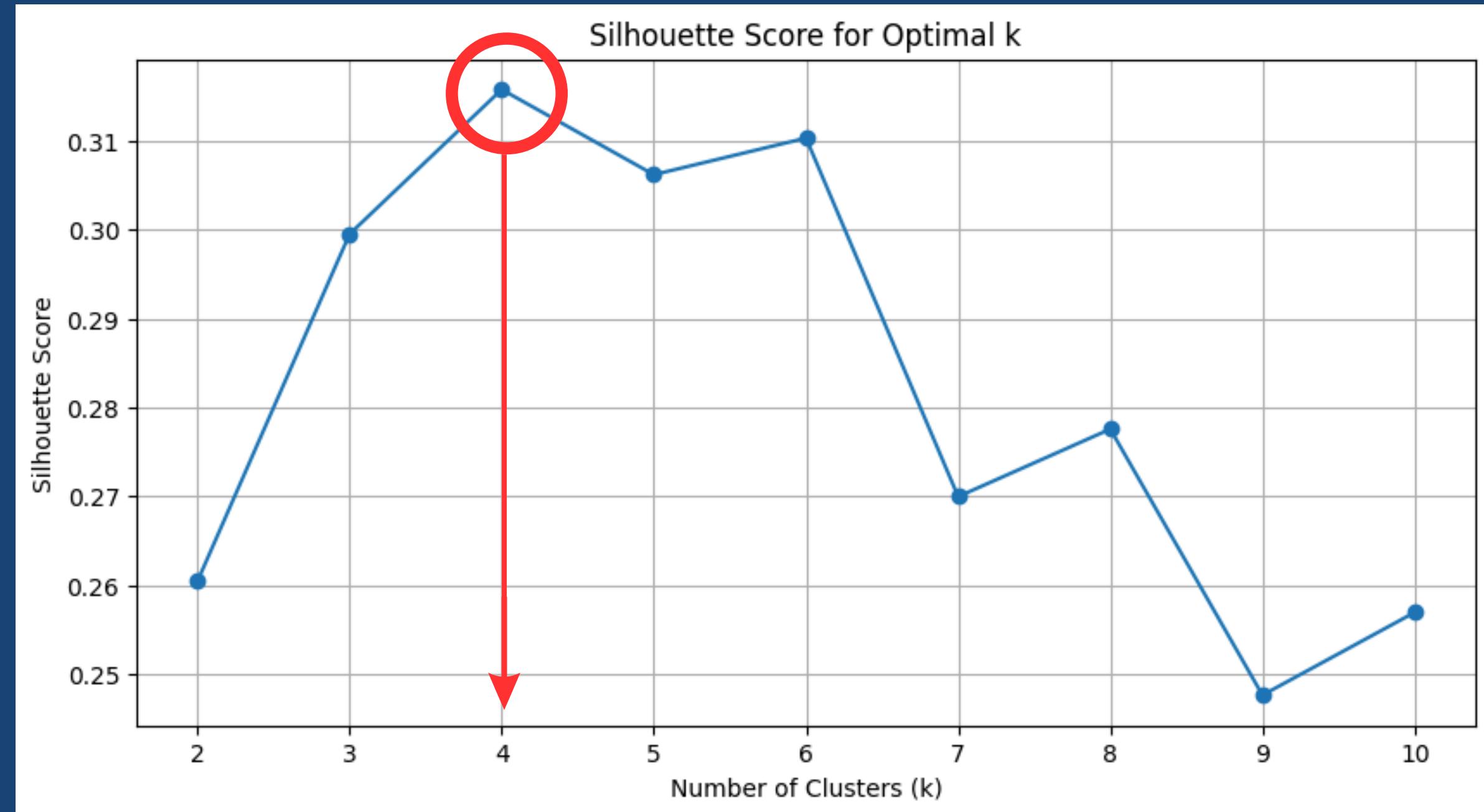
Phương pháp Elbow

Nguyễn Thị Diễm Ly

# K-MEANS

XÁC ĐỊNH  
SỐ CỤM TỐI ƯU

Số cụm tối ưu là 4

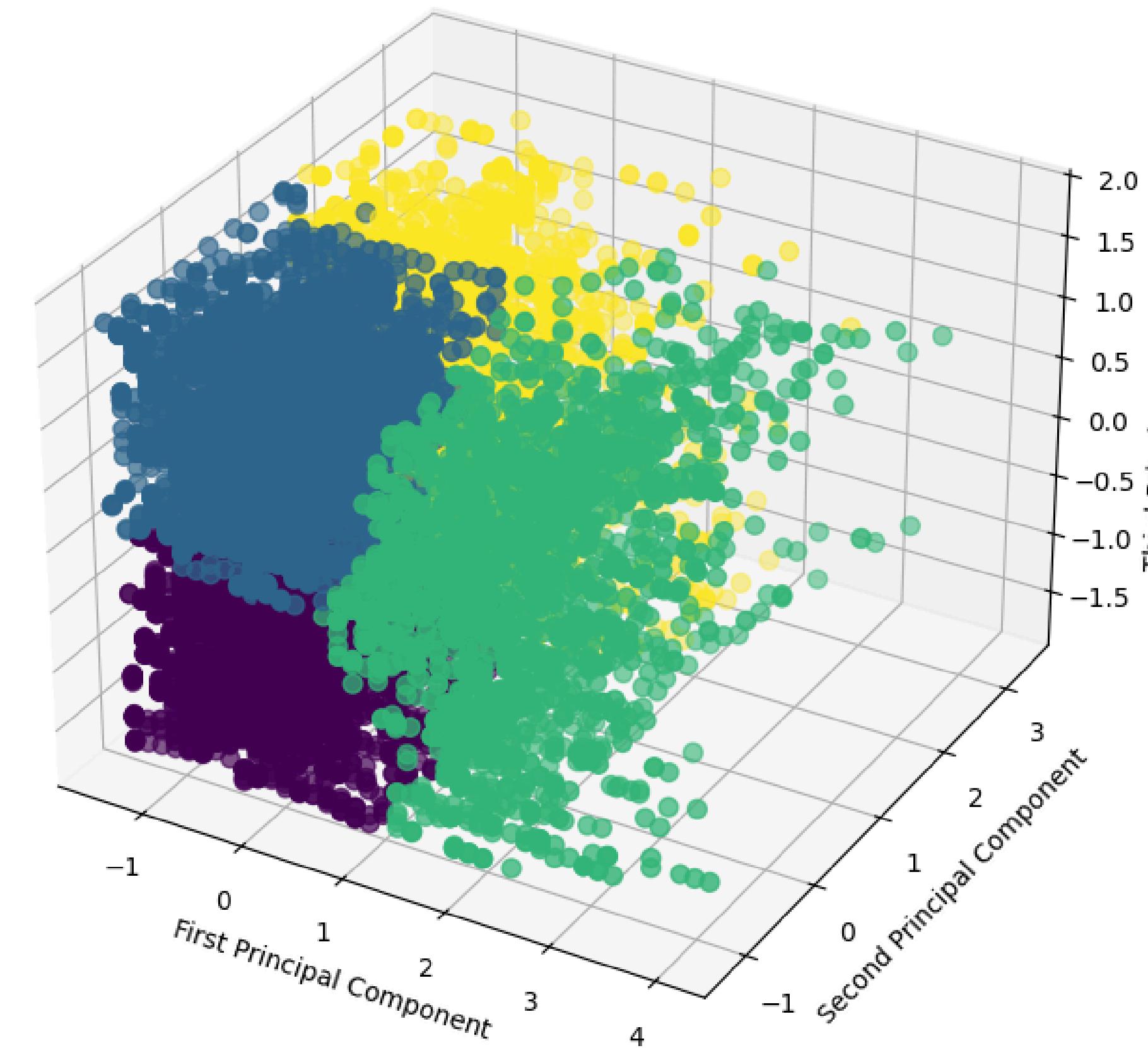


Nguyễn Thị Diễm Ly

# KẾT QUẢ

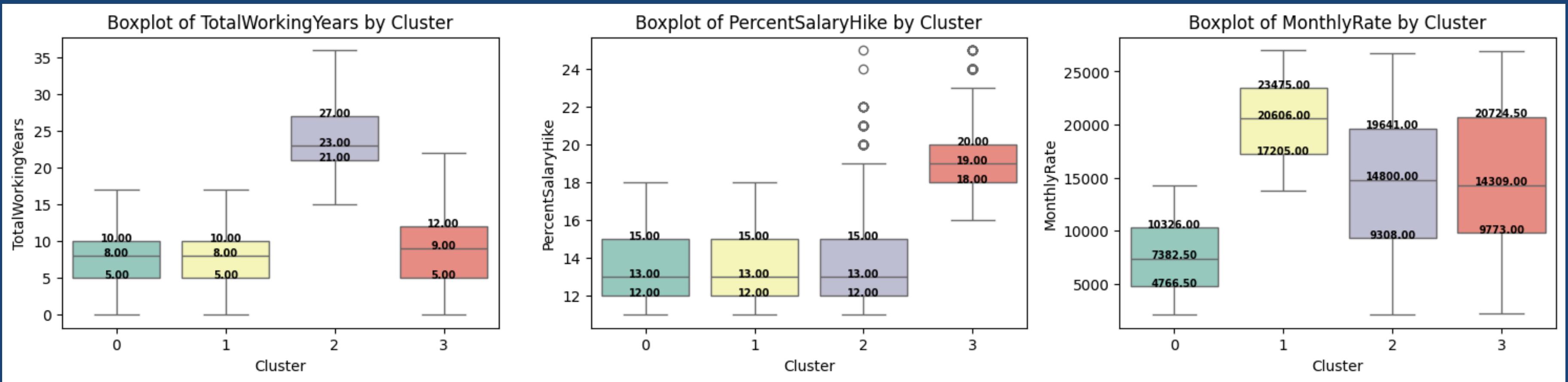
Silhouette Score for K - Means: 0.32

K-means Clustering (K=4)



Nguyễn Thị Diễm Ly

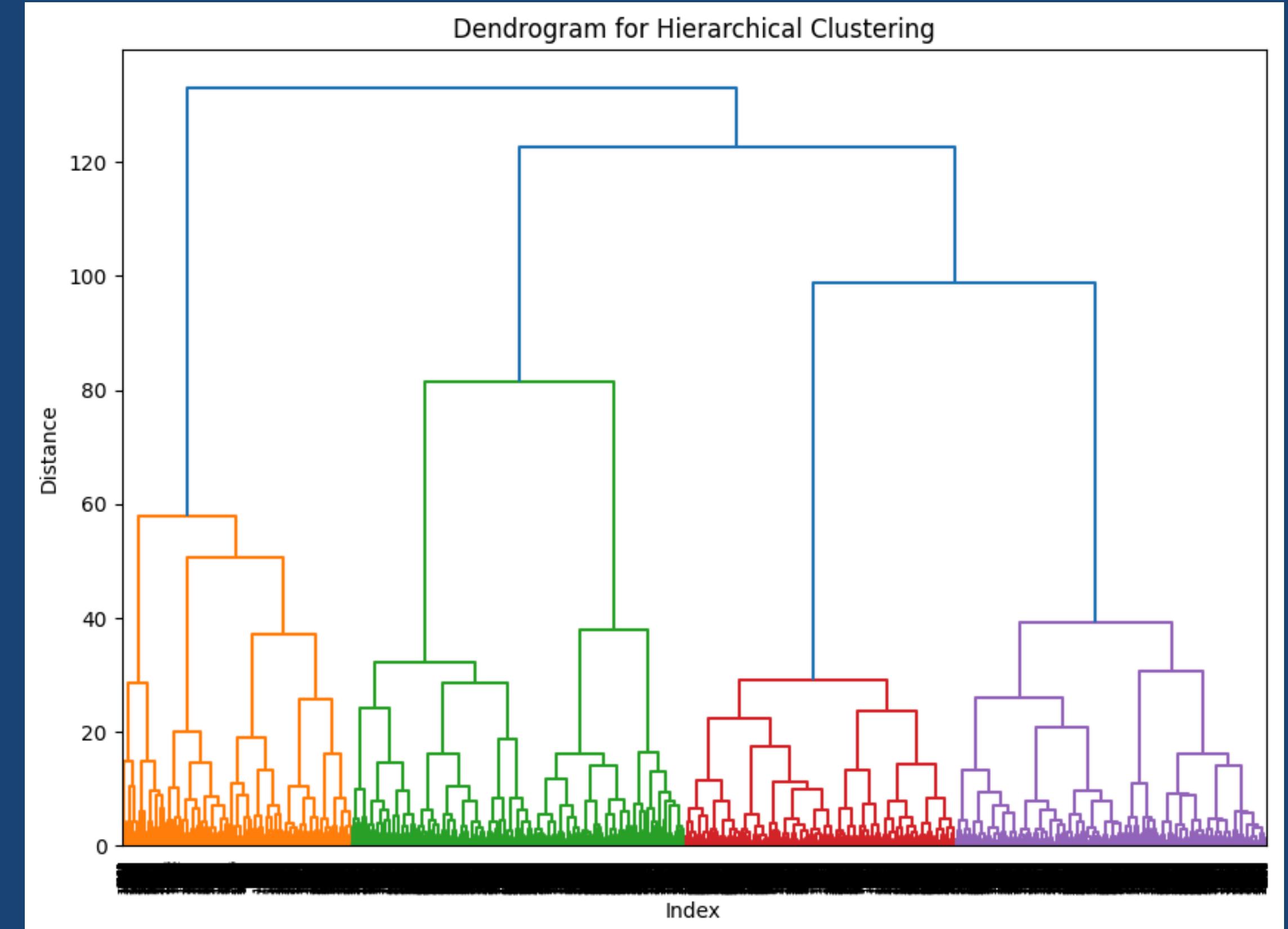
# BIỂU ĐỒ PHÂN PHỐI CÁC CỤM



Nguyễn Thị Diễm Ly

# HIERARCHICAL

XÁC ĐỊNH  
SỐ CỤM TỐI ƯU

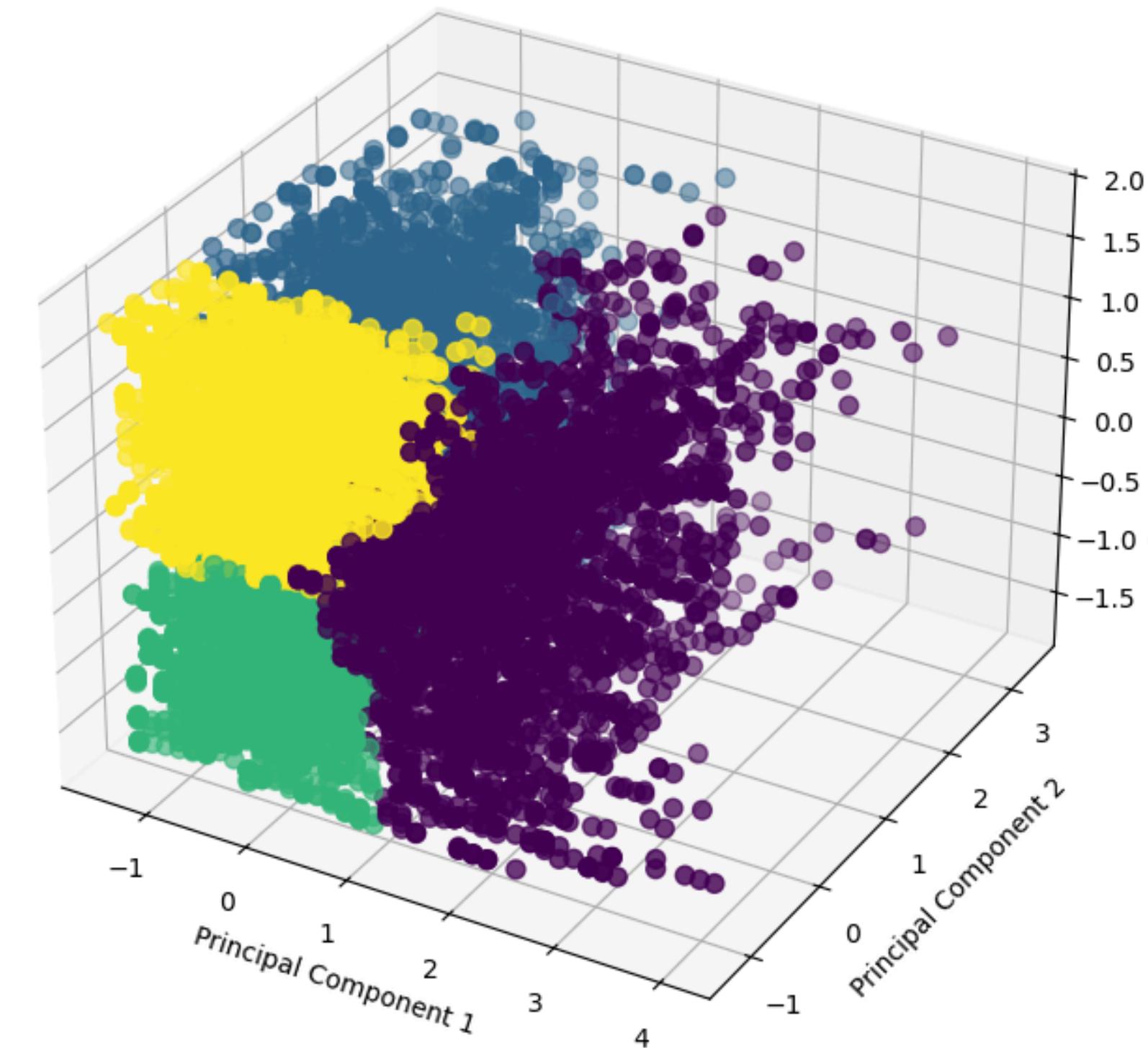


Dendrogram

Nguyễn Thị Diễm Ly

# KẾT QUẢ

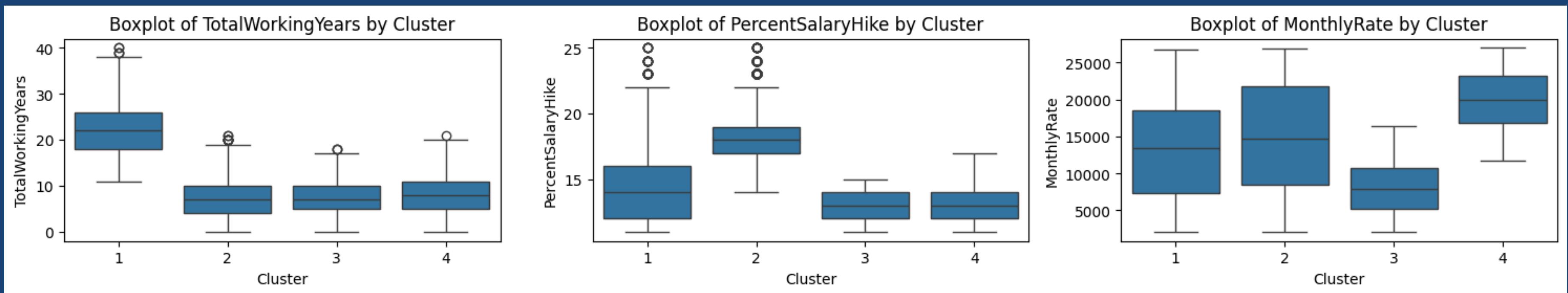
Hierarchical Clustering Results



Silhouette Score for Hierarchical Clustering: 0.25

Nguyễn Thị Diễm Ly

# BIỂU ĐỒ PHÂN PHỐI CÁC CỤM

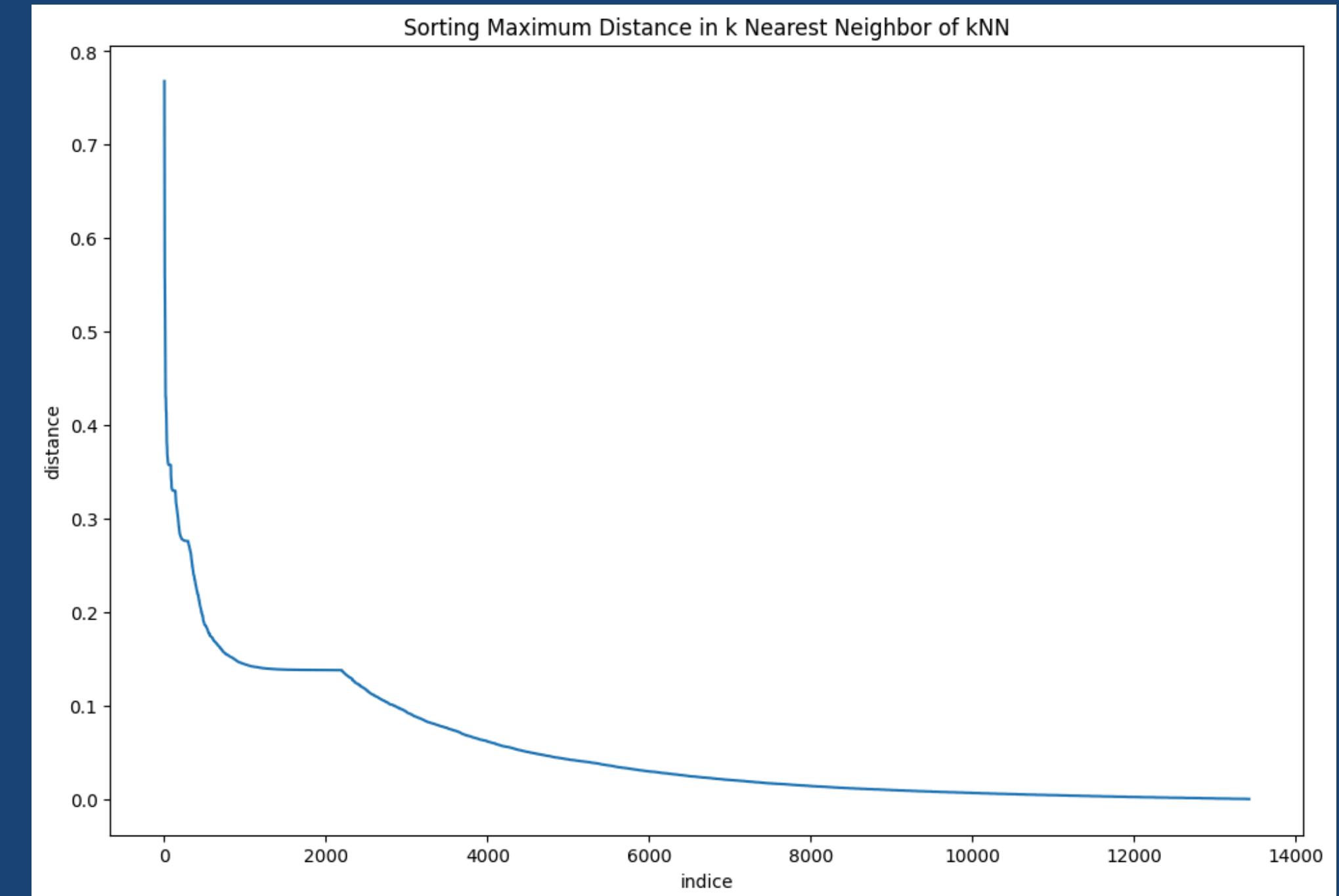


# DBSCAN

## XÁC ĐỊNH EPSILON TỐI ƯU

- MinPts = 4
- K = 3

Epsilon tối ưu nằm trong  
khoảng [ 0.1 : 0.2 ]



# DBSCAN

```
eps: 0.10, Silhouette Score: 0.41  
eps: 0.11, Silhouette Score: 0.34  
eps: 0.12, Silhouette Score: 0.29  
eps: 0.13, Silhouette Score: 0.22  
eps: 0.14, Silhouette Score: -0.21  
eps: 0.15, Silhouette Score: -0.24  
eps: 0.16, Silhouette Score: -0.26  
eps: 0.17, Silhouette Score: -0.27  
eps: 0.18, Silhouette Score: -0.27  
eps: 0.19, Silhouette Score: -0.27  
eps: 0.20, Silhouette Score: -0.26
```

Best eps: 0.10, Best Silhouette Score: 0.41  
Số cụm: 981, Số điểm nhiễu (outliers): 3808



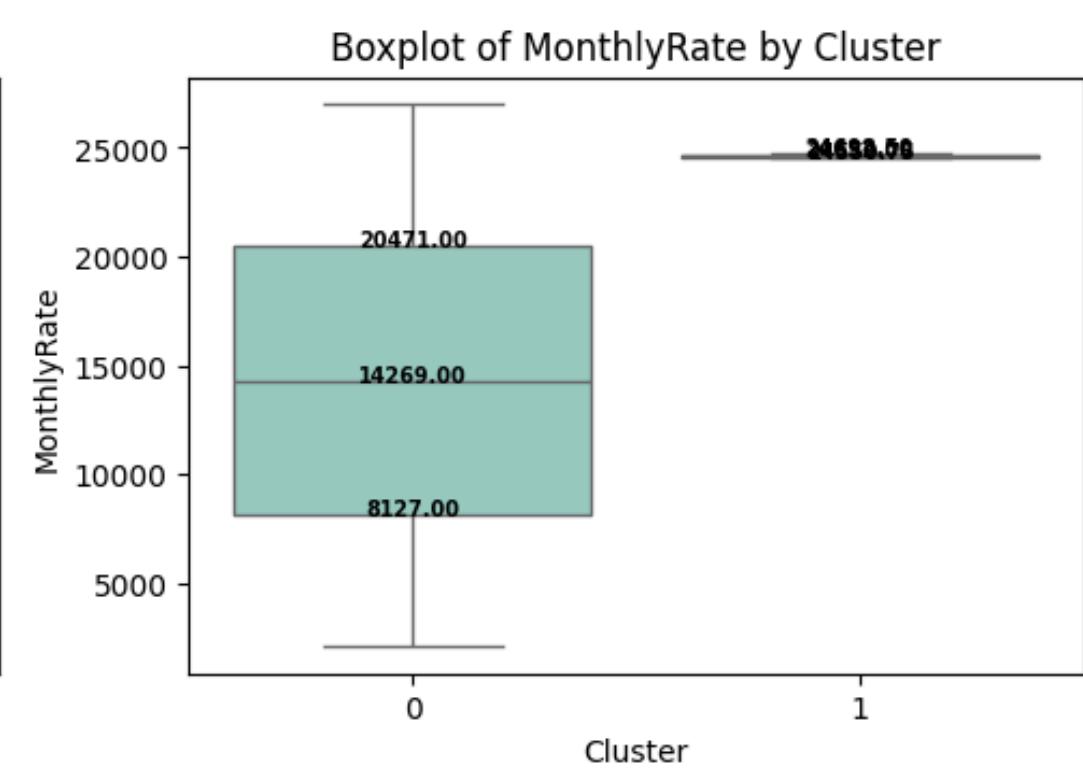
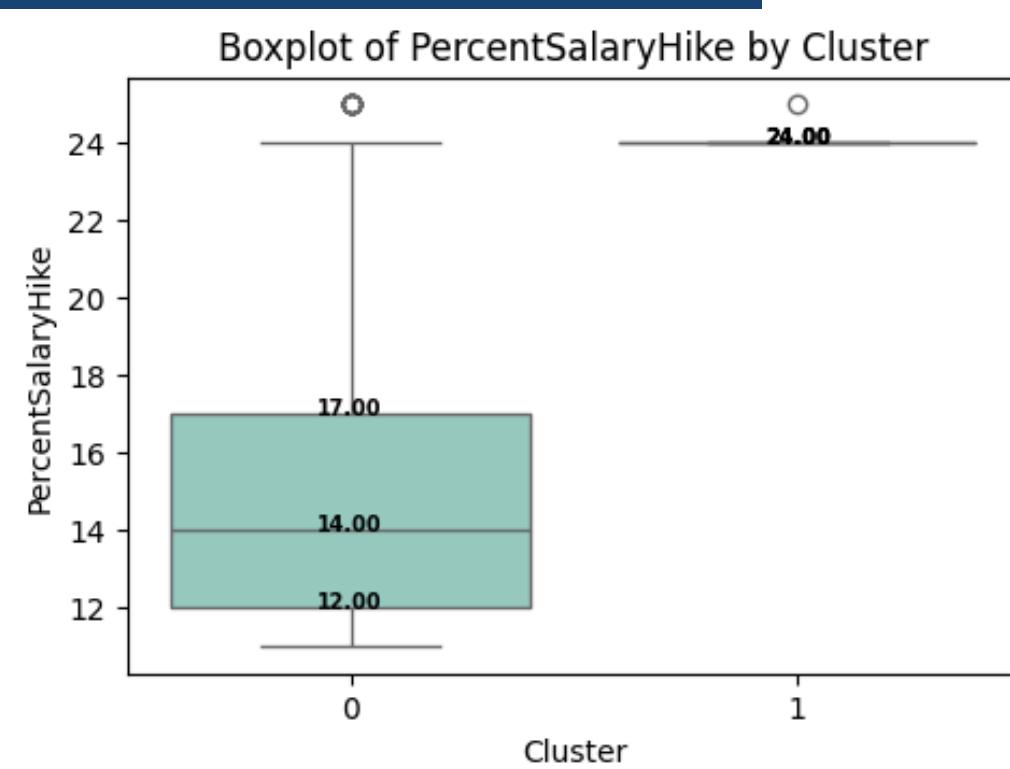
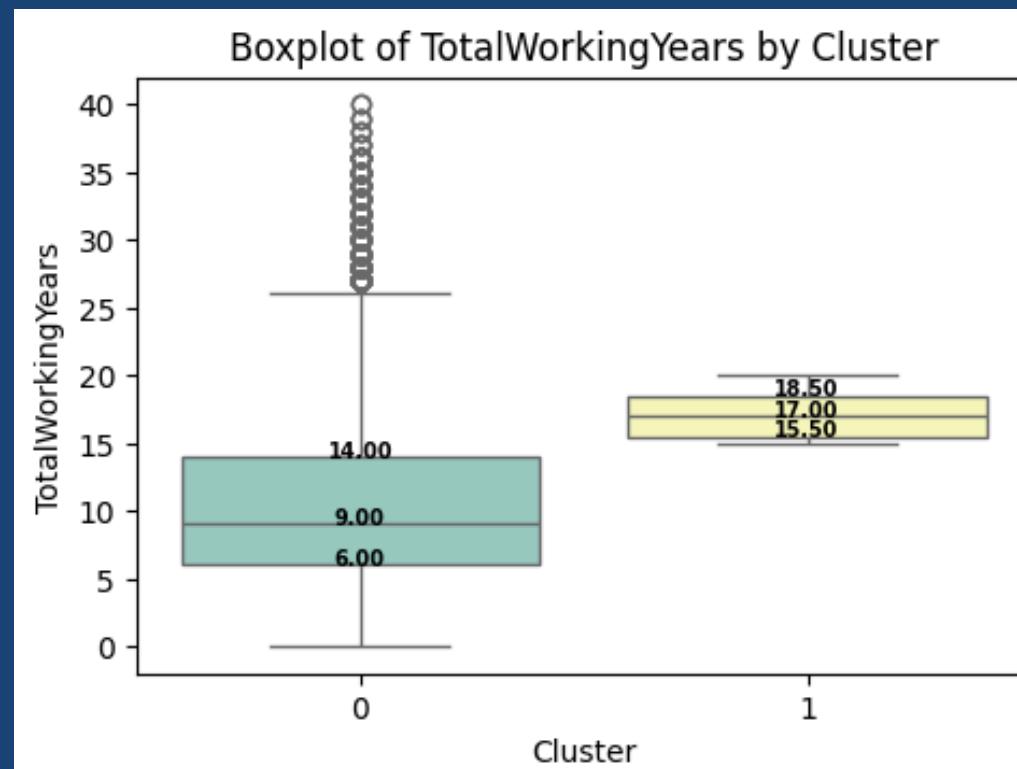
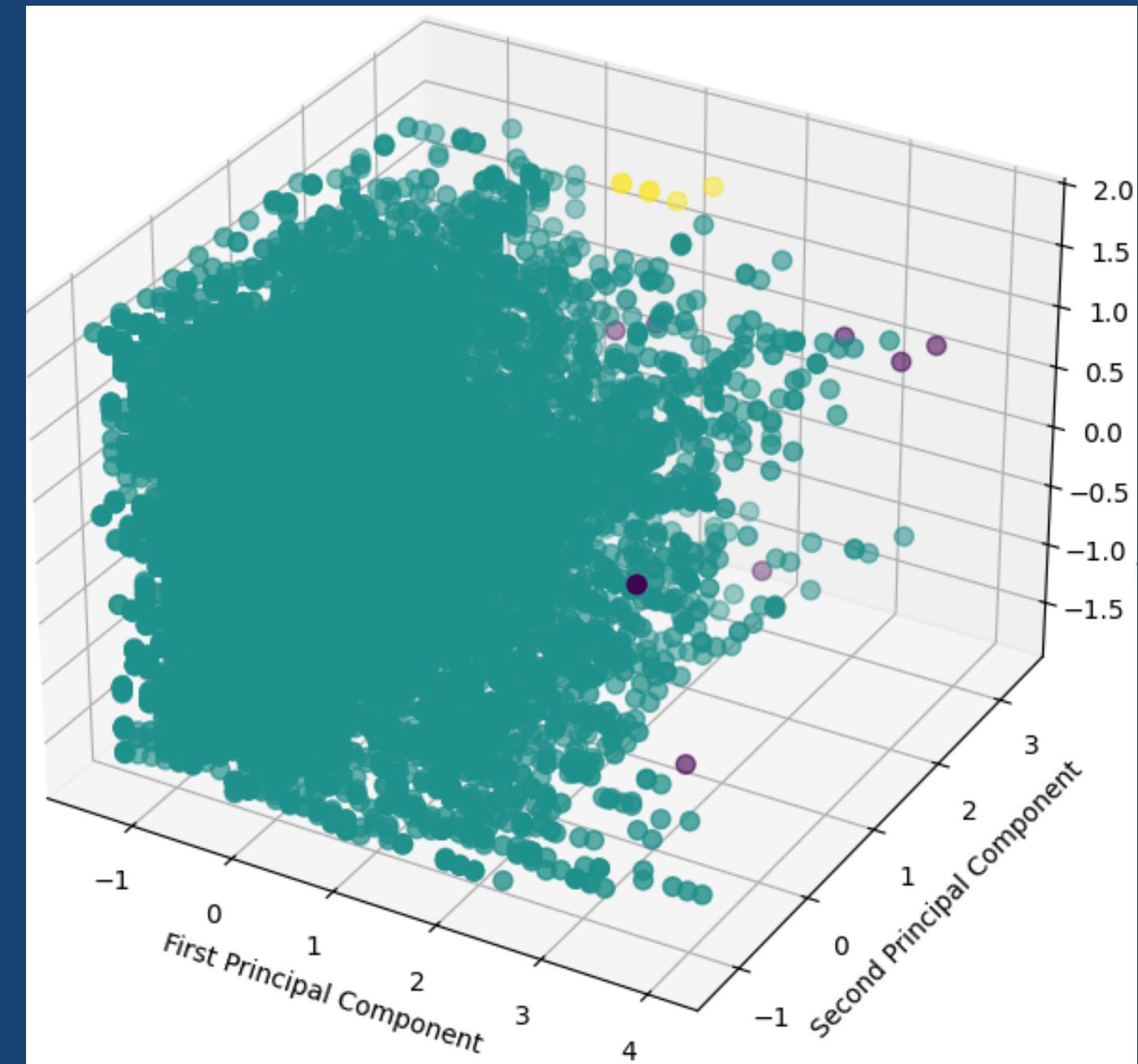
## ĐIỀU CHỈNH THAM SỐ

- MinPts = 4
- Epsilon = 0.5



Số cụm: 2, Số điểm nhiễu (outliers): 8  
Silhouette Score for DBSCAN: 0.38

# KẾT QUẢ



# CLUSTERS



## Cụm 0: Cơ bản

- Kinh nghiệm: 5 - 10
- Đóng góp: 5K - 10K
- % tăng lương: 12% - 15%



## Cụm 1: Ngôi sao

- Kinh nghiệm: 5 - 10
- Đóng góp: 17K - 23K
- % tăng lương: 12% - 15%



## Cụm 2: Chuyên gia

- Kinh nghiệm: 21 - 27
- Đóng góp: 9K - 19K
- % tăng lương: 12% - 15%



## Cụm 3: Tiềm năng

- Kinh nghiệm: 5 - 12
- Đóng góp: 10K - 21K
- % tăng lương: 18% - 20%

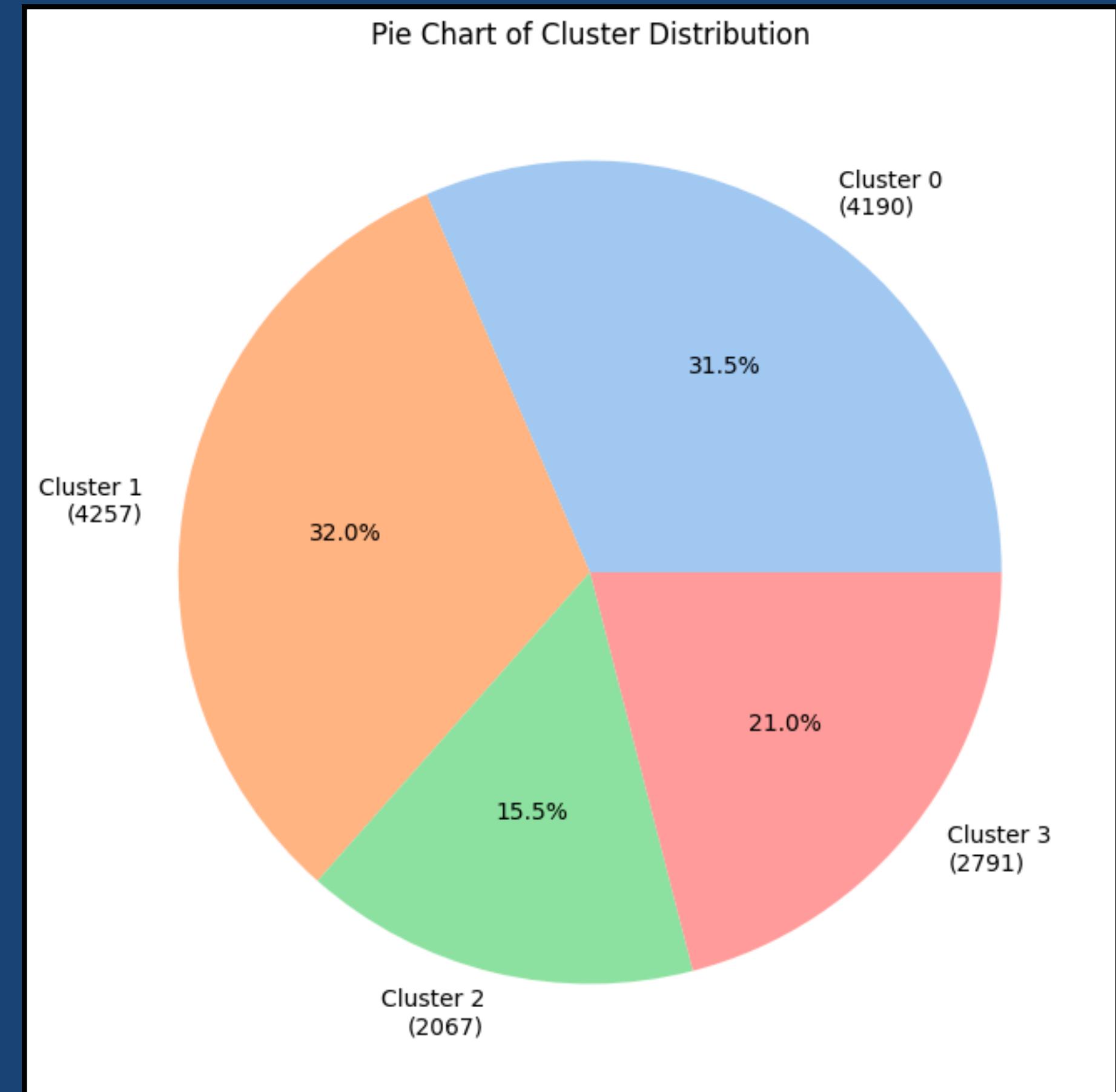


# EDA IN 4 CLUSTERS

Nguyễn Thị Diễm Ly

# PHÂN PHỐI CỤM

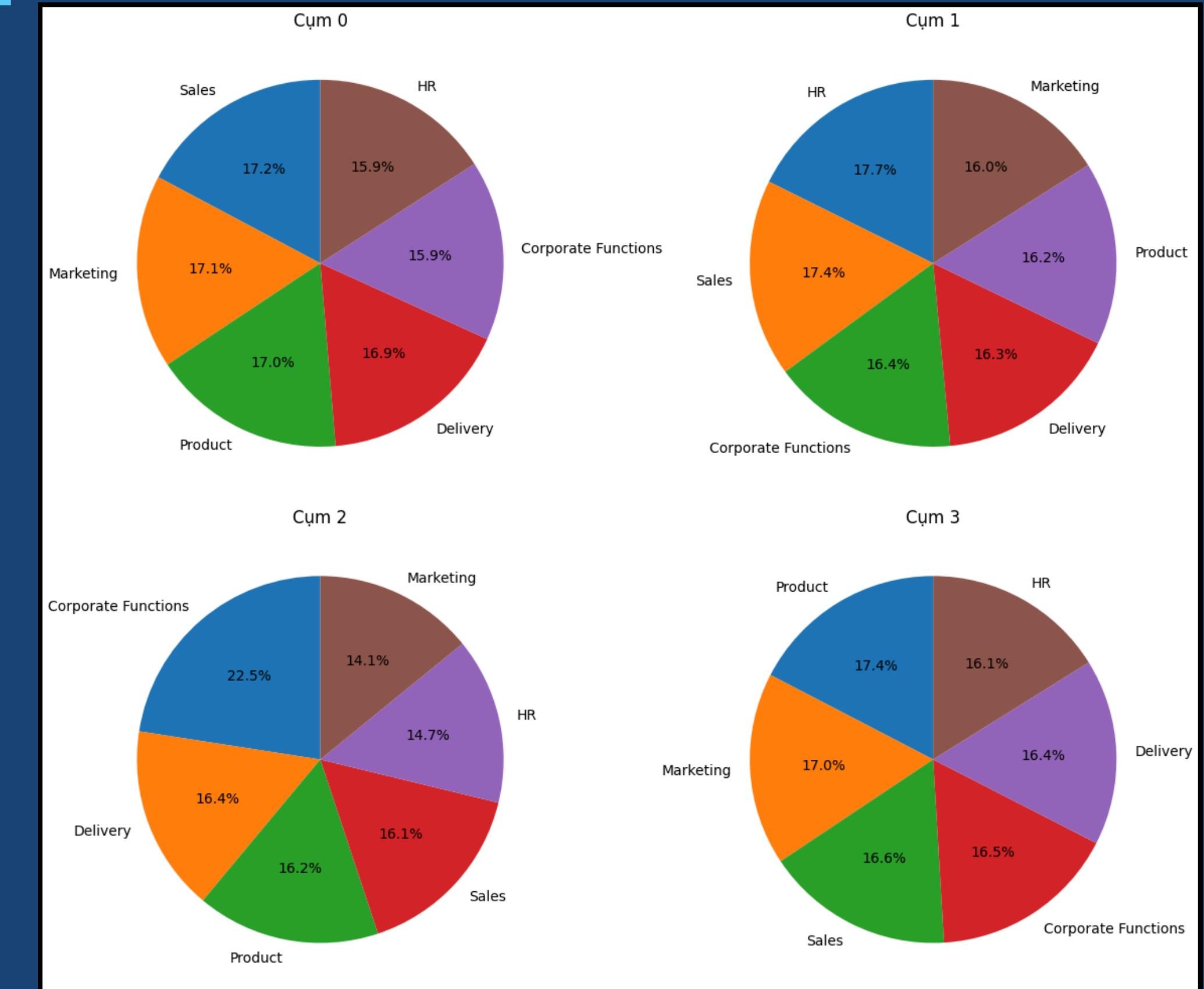
- Nhóm nhân viên “**Ngôi sao**” và “**Cơ bản**” chiếm tỷ lệ cao nhất
- Nhóm “**Chuyên gia**” thấp nhất với 15.5%



Nguyễn Thị Diễm Ly

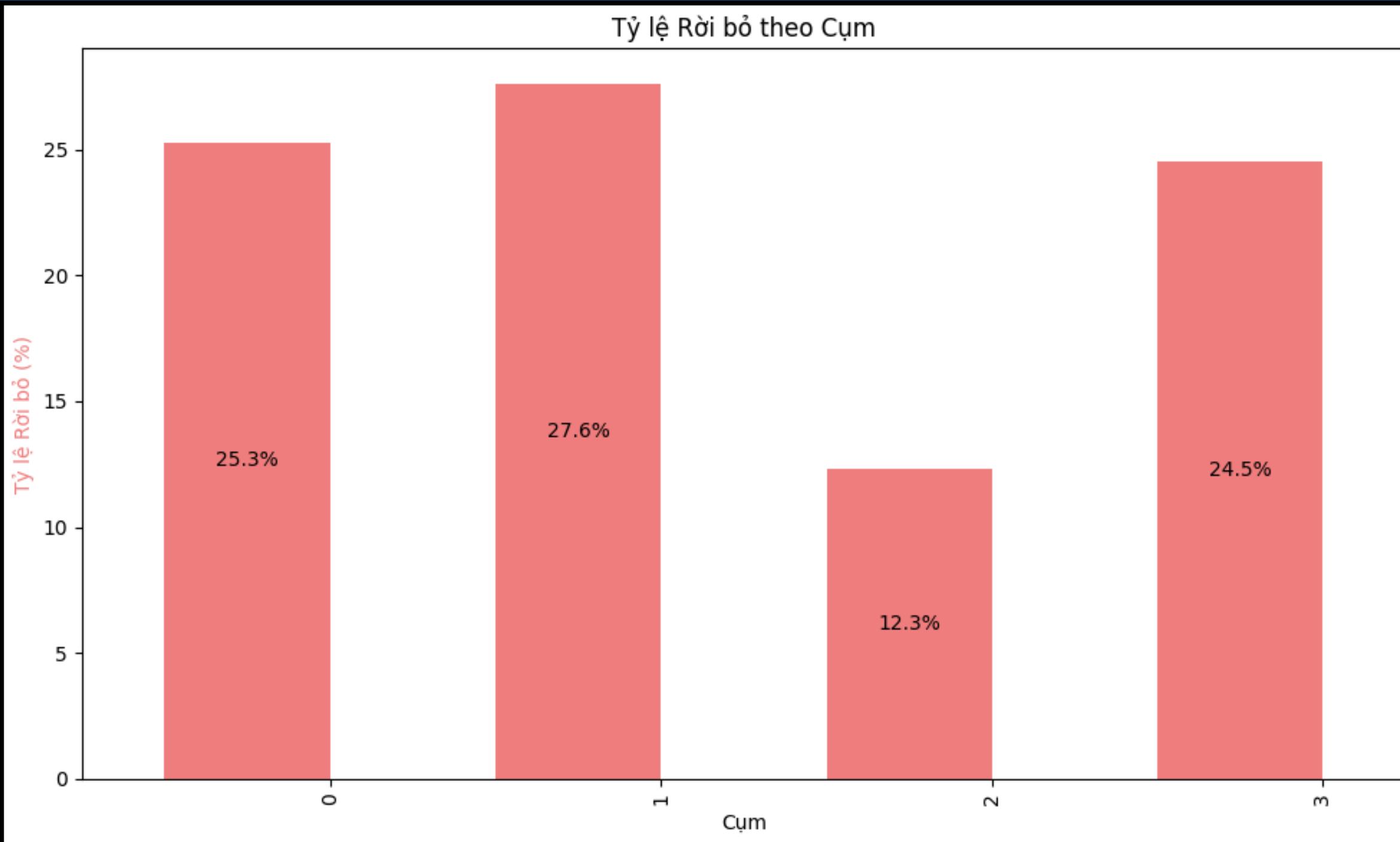
# PHÒNG BAN THEO CỤM

- Nhân viên ở các phòng ban được phân phối **khá đều nhau**
- Cụm “**Chuyên gia**” có đến **22.5%** nhân viên thuộc bộ phận **Corporate Functions**



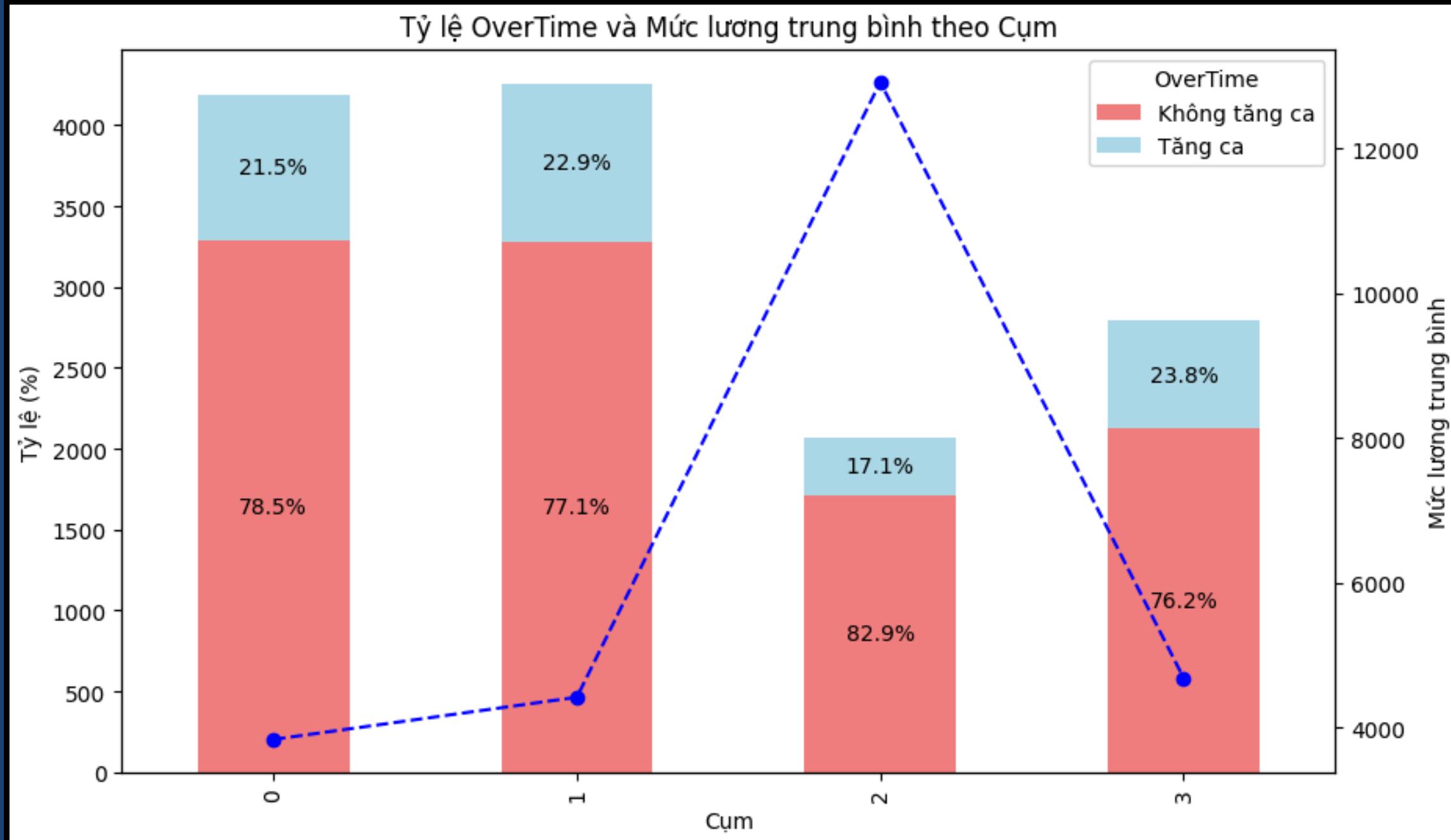
# TỶ LỆ RỜI BỎ THEO CỤM

Tỷ lệ Rời bỏ theo Cụm



- Cụm “**Ngôi sao**” chiếm tỷ lệ **cao nhất** về số nhân viên rời bỏ
- Trong khi đó, cụm “**Chuyên gia**” có lượng nhân viên rời bỏ **ít nhất**

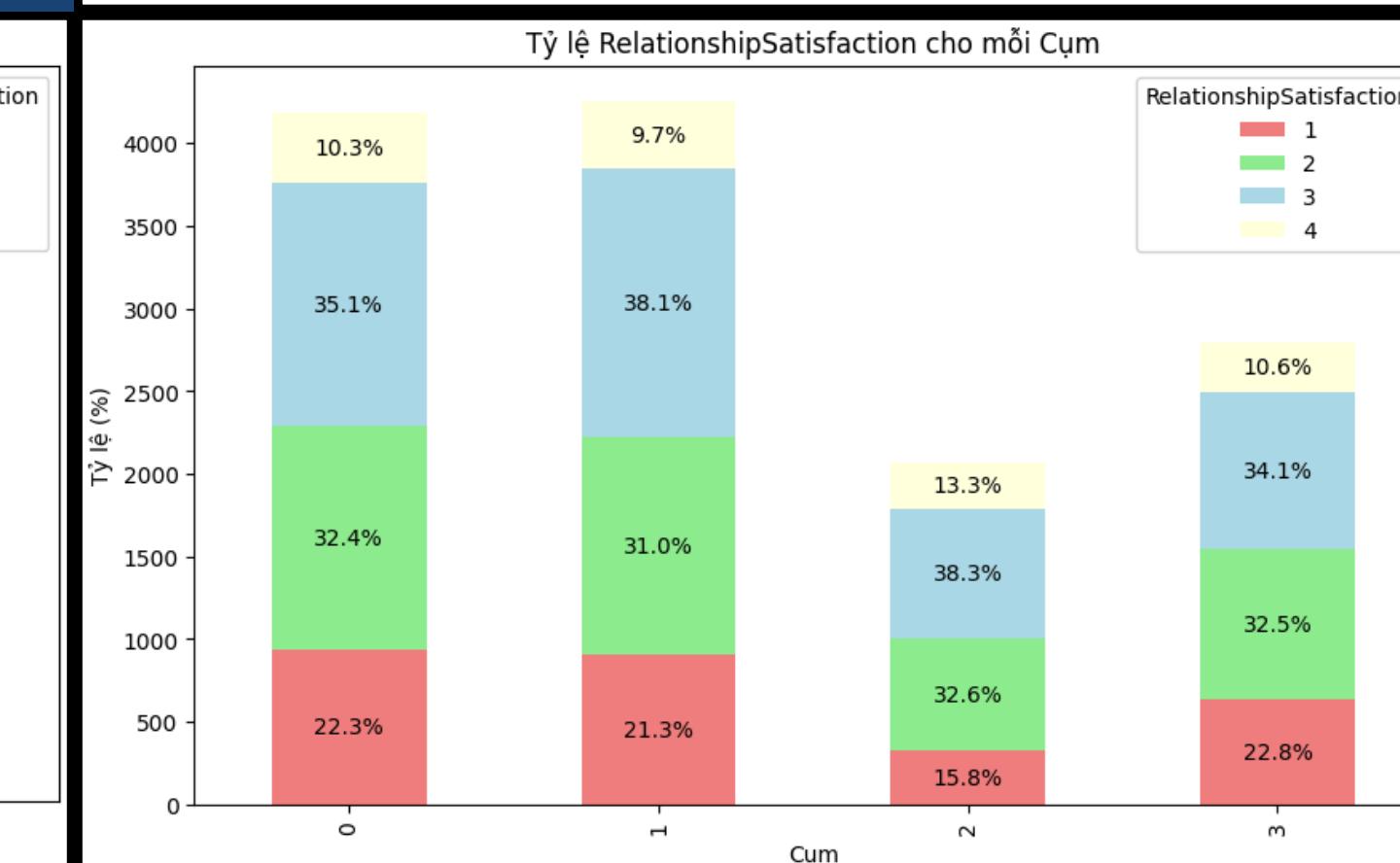
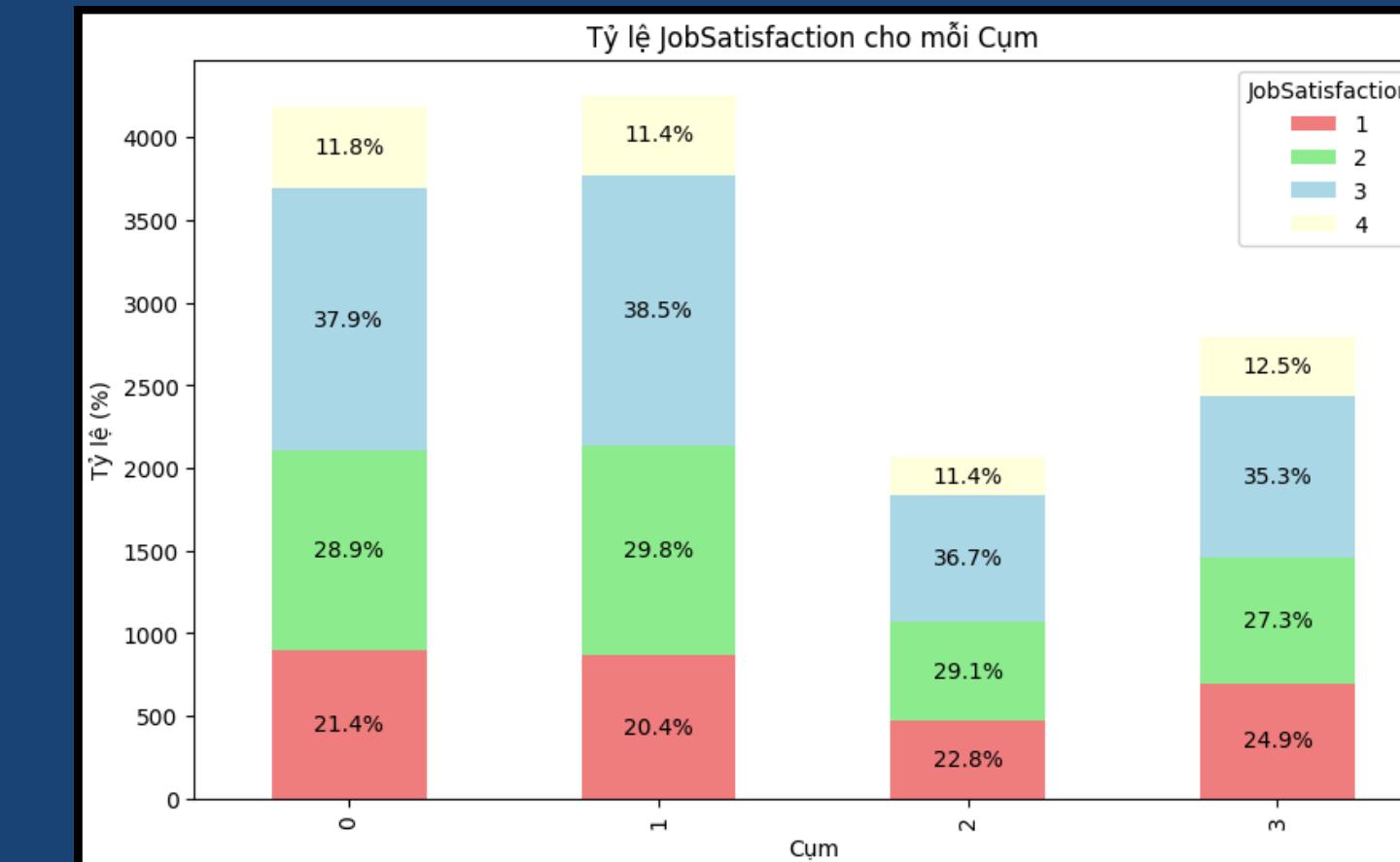
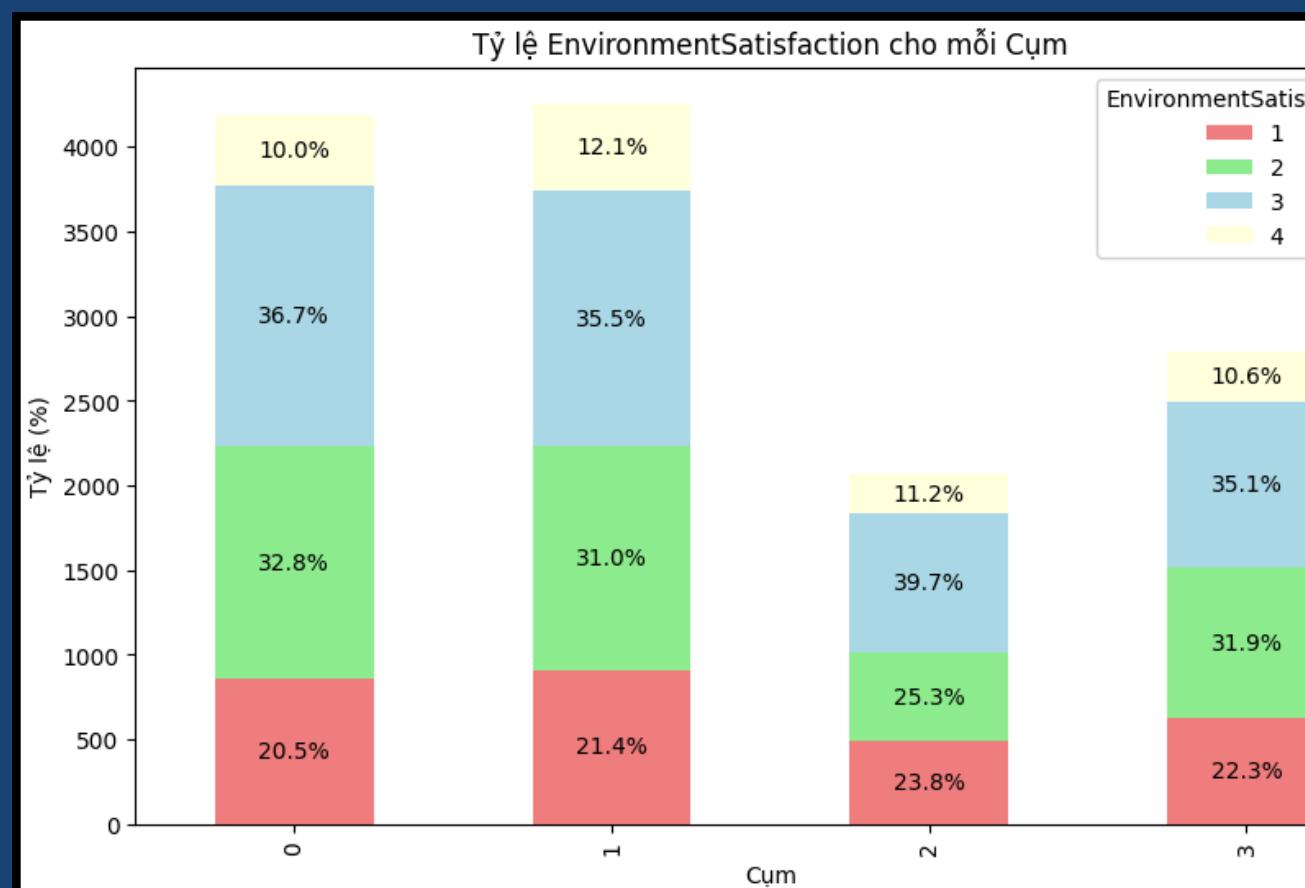
# TĂNG CA VÀ LƯƠNG TRUNG BÌNH THEO CỤM



- Cụm “**Chuyên gia**” chỉ có **17.1%** các nhân viên làm việc tăng ca nhưng mức lương trung bình **cao nhất** lên đến **11K USD**
- Trong khi đó, các cụm còn lại có tỷ lệ nhân viên tăng ca **cao hơn** nhưng mức lương trung bình đều dưới **5K USD**

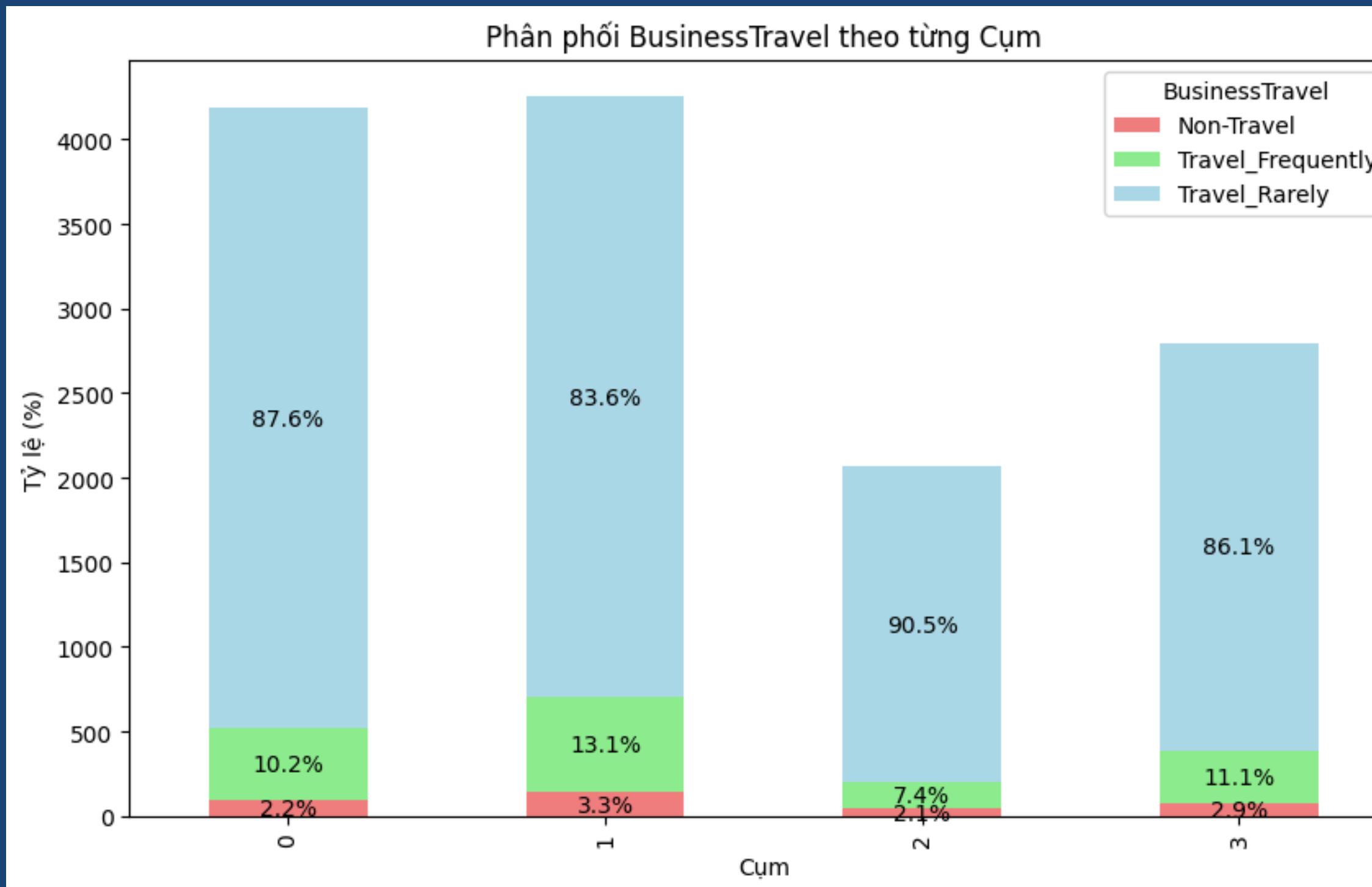
# MỨC HÀI LÒNG THEO CỤM

- Nhìn chung không có sự chênh lệch đáng kể về các mức độ hài lòng giữa các nhóm.
- Tuy nhiên, mức **hài lòng thấp** (Mức 1 và 2) ở cả 4 nhóm đều **chiếm tỷ lệ cao (>50%)**



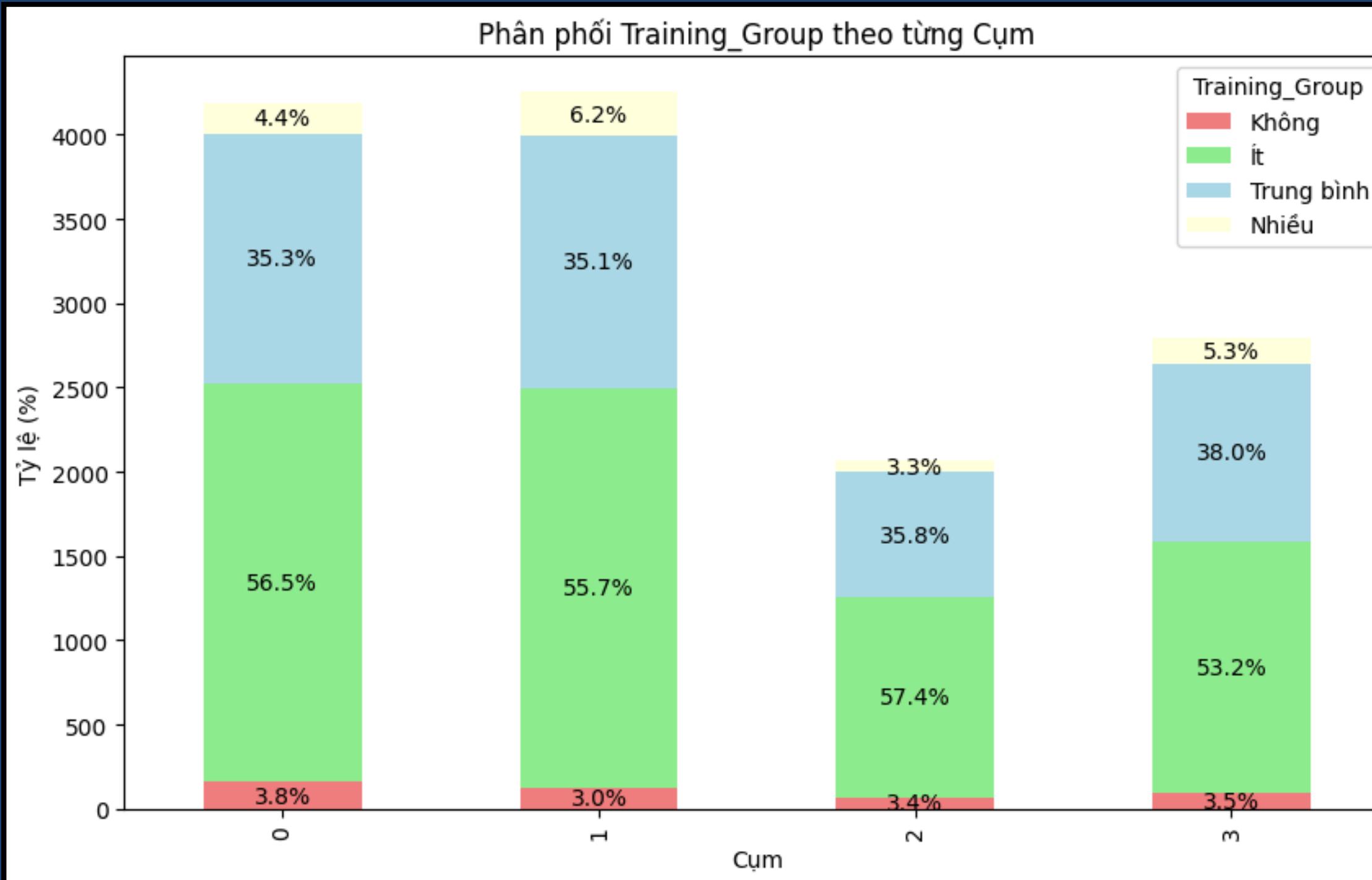
Điểm Ly

# TẦN SUẤT CÔNG TÁC THEO CỤM



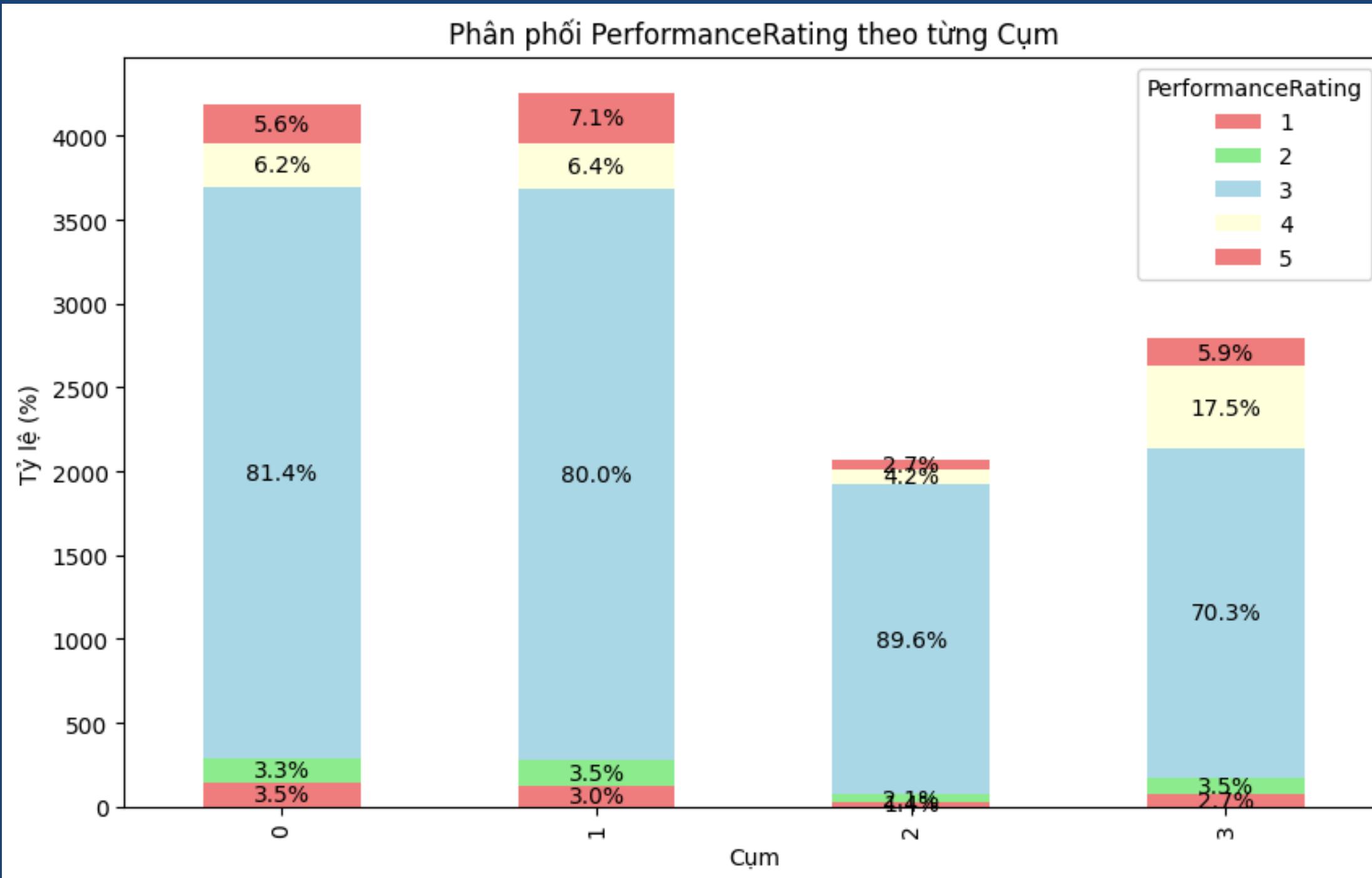
- Phần lớn các nhân viên của cả 4 nhóm đều **hiếm khi** đi công tác (**>86%**)
- Số lượng nhân viên **thường xuyên** đi công tác chiếm **>10% mỗi nhóm**

# TẦN SUẤT TRAINING THEO CỤM



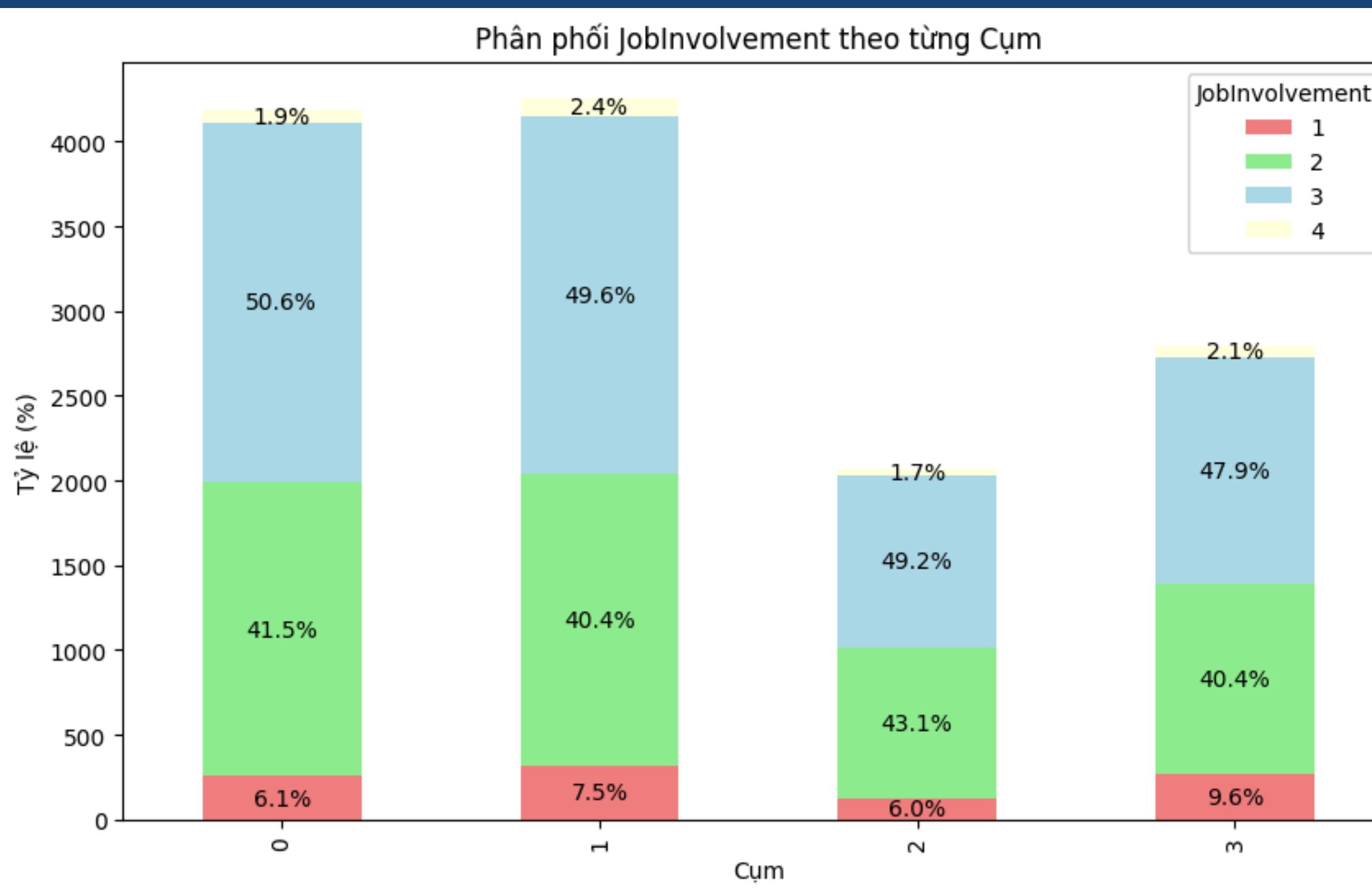
- Ở cả 4 nhóm, **gần 60%** nhân viên **không tham gia và ít** tham gia training.
- Các nhân viên **thường xuyên tham gia** các khóa training chỉ **chiếm phần nhỏ**

# ĐÁNH GIÁ HIỆU SUẤT THEO CỤM



- Cụm “**Tiềm năng**” và “**Ngôi Sao**” được đánh giá hiệu suất ở **mức tốt (4 và 5)** chiếm tỷ lệ **cao nhất** lần lượt là **23.4% và 13.5%**.
- Vẫn còn tồn tại những nhân viên bị **đánh giá kém** về hiệu suất

# MỨC ĐÓNG GÓP TRONG DỰ ÁN THEO CỤM



- Cụm “**Tiềm năng**” có đến **9.6%** các nhân viên **ít đóng góp** trong dự án

# ĐỀ XUẤT

Nhóm “**Cơ bản**” còn **ít kinh nghiệm** và **hiệu suất chưa ổn định**, có tỷ lệ nhân viên **rời bỏ cao**.

1

Việc ít tham gia các chương trình đào tạo hạn chế cơ hội phát triển năng lực.

- Xây dựng chương trình đào tạo chuyên biệt theo nhu cầu từng vị trí.
- Thiết lập chính sách lương thưởng gắn liền với KPI để tạo động lực.
- Tạo cơ hội tham gia các dự án nhỏ nhằm nâng cao kỹ năng thực tế.

2

Nhóm “**Ngôi Sao**” có **hiệu suất xuất sắc**, đóng góp vượt trội nhưng lại nhận **mức lương và tăng lương thấp**. **Tỷ lệ rời bỏ cao** cho thấy sự thiếu quan tâm từ công ty.

- Ưu tiên tăng lương và thưởng thường xuyên hơn, gắn liền với thành tích.
- Phát triển các chương trình khen thưởng phi tài chính như thăng tiến nhanh, công nhận đóng góp.

# ĐỀ XUẤT

3

Nhóm “**Tiềm năng**” kinh nghiệm chưa nhiều, nhưng có **mức đóng góp tốt, ổn định** cho công ty, họ có **phần trăm tăng lương cao nhất** tuy mức lương vẫn còn thấp và **tỷ lệ nghỉ việc cao**.

- Tăng cường hỗ trợ phát triển sự nghiệp như đào tạo chuyên môn cao hơn.
- Ưu tiên tăng lương và thưởng thường xuyên hơn, gắn liền với thành tích.
- Tổ chức khảo sát để nắm rõ tâm lý nhân viên, lý do nghỉ việc, từ đó cải thiện chính sách giữ chân.

4

Nhóm “**Chuyên gia**” là những nhân viên kỳ cựu của công ty, nắm những chức vụ quan trọng, họ **đóng góp tốt và ổn định** cho công ty.

- Xây dựng chương trình mentor, trong đó chuyên gia dẫn dắt nhân viên trẻ.
- Đảm bảo chính sách đãi ngộ cạnh tranh để duy trì sự gắn bó.



# LUẬT KẾT HỢP

- Apriori
- FP-Growth

# MỤC TIÊU

- Tìm ra mối quan hệ giữa các đặc trưng quan trọng
- Tìm hiểu yếu tố ảnh hưởng đến quyết định nghỉ việc
- Hỗ trợ ra quyết định quản lý nhân sự



# XỬ LÝ DỮ LIỆU

**Biến đổi dữ liệu:** Chuyển đổi những dữ liệu liên tục thành nhóm/ khoảng giá trị

- Tuổi (Age): '18-30', '31-40', '41-50', '51-60'
- Thu nhập (MonthlyIncome): 'Thấp', 'Trung bình', 'Cao', 'Rất cao'
- Khoảng cách (DistanceFromHome): 'Rất gần', 'Gần', 'Xa', 'Rất xa' [10, 15, 20, 25]
- Tăng lương (PercentSalaryHike): 'Thấp', 'Trung bình', 'Cao' [10, 15, 20, 25]
- Gắn bó (YearsAtCompany): '1-5', '6-10', '11-15'
- Training (TrainingTimesLastYear): 'Không', 'Ít', 'Trung bình', 'Nhiều' [-1, 0, 2, 4, 6]
- Kinh nghiệm (TotalWorkingYears): 'Chưa có kinh nghiệm', 'Có kinh nghiệm', 'Thành thạo', 'Chuyên gia', 'Lão làng' [-1, 0, 5, 15, 25, 40]

**Mã hóa:** One-hot encoding

# APRIORI



1

2

3

## THƯ VIỆN

mlxtend.frequent\_patterns.apriori,  
.association\_rules

## ITEMSETS

Min support =0.13  
(1750 lần)

## RULES

Min confidence=0.1

# Apriori

1

## Tổng quan

- Số lượng: 189862
- Thời gian: 4 phút 52 giây

2

## Luật về quyết định rời bỏ của nhân viên

- Lọc các luật liên quan đến cột Attrition từ cột consequents (vết phải của luật)

```
rules_apr['consequents'].apply(lambda x: 'Attrition_Yes' in x or 'Attrition_No' in x)
```

- Số lượng: 55697 (30%)

# Apriori

## Attrition\_Yes

STT	Về trái	Về phải	Confidence	Lift
1	(OverTime_Yes, YearsAtCompany_Group_1-5)	(Attrition_Yes)	0.66	2.77
2	(OverTime_Yes)	(Attrition_Yes)	0.64	2.69
3	(OverTime_Yes)	(Attrition_Yes, YearsAtCompany_Group_1-5)	0.61	2.72

# Apriori

## Attrition\_Yes

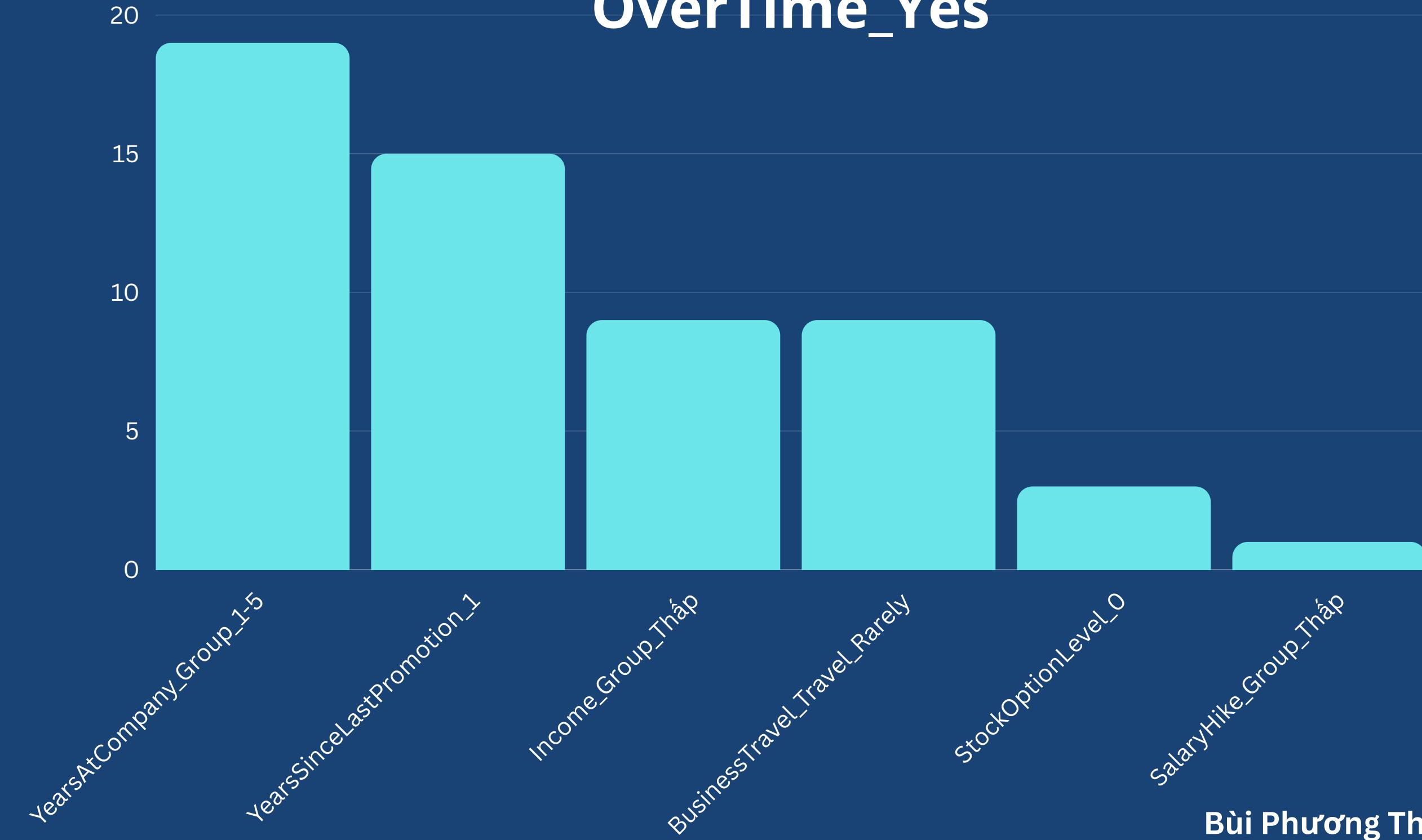


- Chỉ có 3 luật liên quan đến việc nhân viên rời bỏ công ty
- Tất cả itemset đều xuất hiện OverTime\_Yes: nhân viên có tăng ca
- Vẽ phải có xuất hiện YearsAtCompany\_Group\_1-5: gắn bó với công ty 1-5 năm

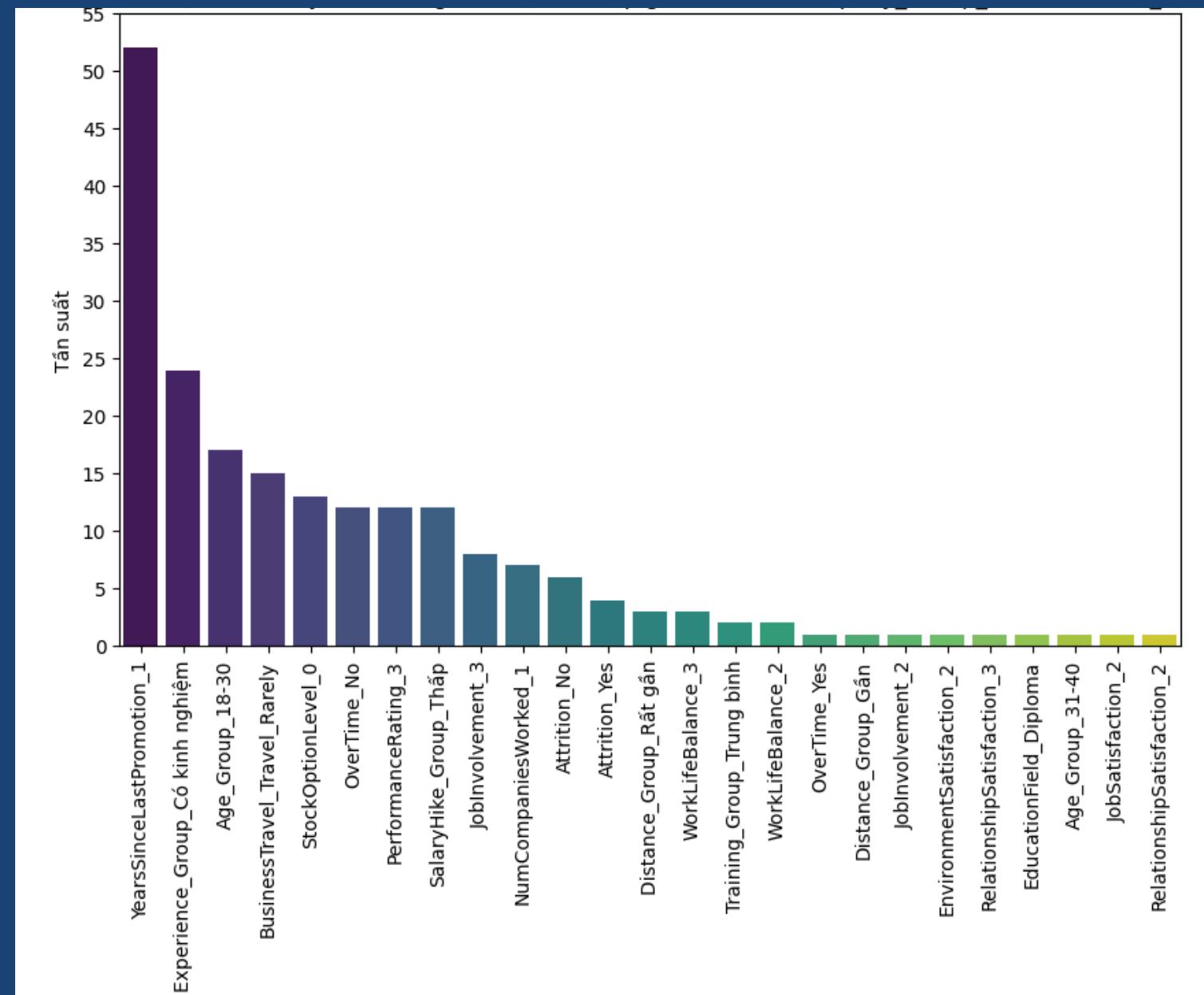
# Apriori

- 42 luật
- Confidence <0,61

OverTime\_Yes



# Apriori



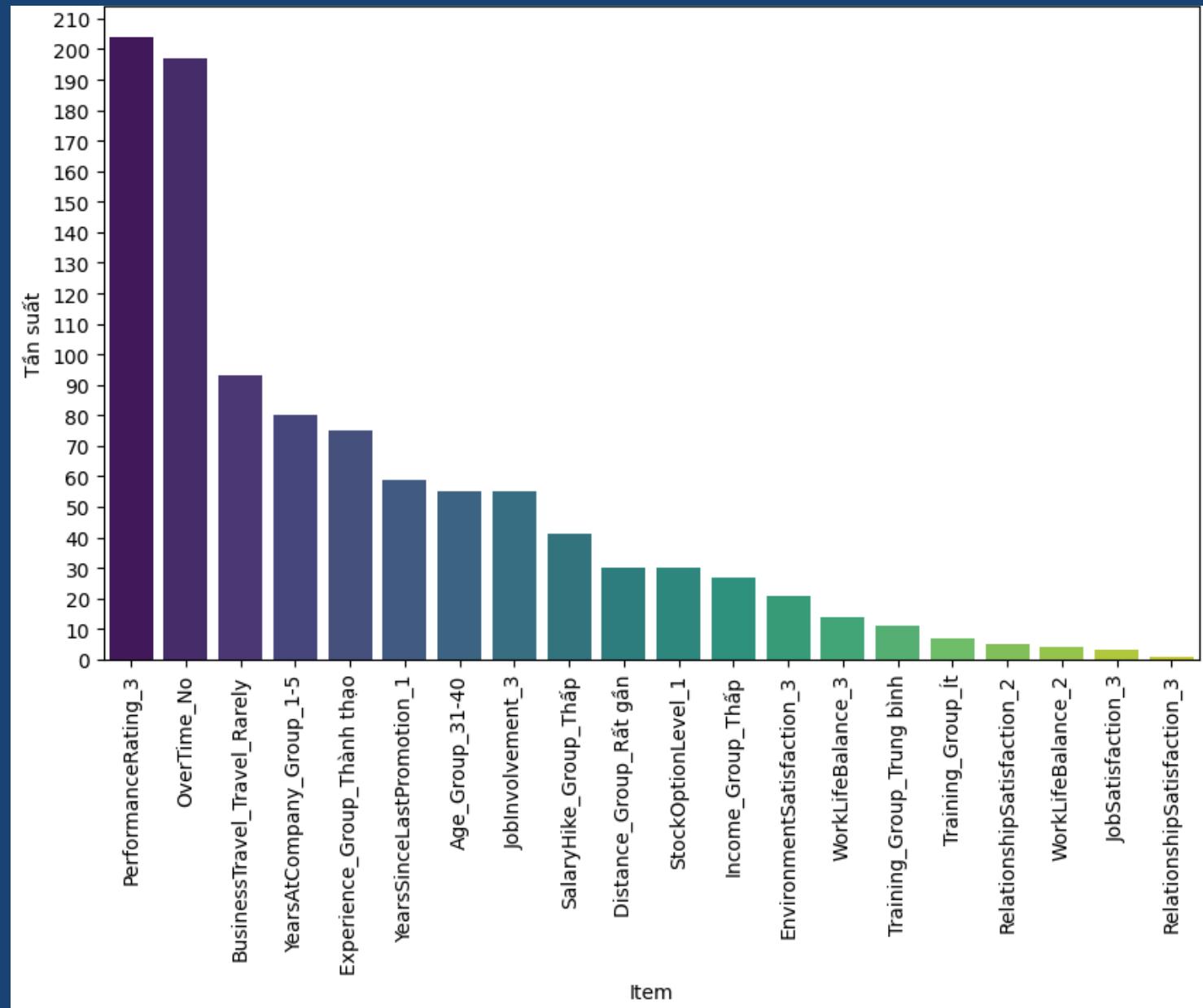
'YearsAtCompany\_Group\_1-5'

và 'Income\_Group\_Thấp'

- 78 luật, **confidence >0,7**
- Là những nhân viên **có kinh nghiệm** nhưng còn **trẻ**, khối lượng công việc ít

# Apriori

## Attrition\_No



- Giữ lại luật có min confidence > 0,97: 204 luật
- Là những nhân viên không mấy áp lực về công việc: có hiệu suất khá, khối lượng và thời gian khá ít, có kinh nghiệm
- 3 yếu tố phổ biến: PerformanceRating\_3, Overtime\_No, BusinessTravel\_Rarely



# FP-GROWTH

1

2

3

## THƯ VIỆN

fpgrowth, association\_rules

## ITEMSETS

Min support =0.1  
(1340 lần)

## RULES

Min confidence=0.6

# FP-GROWTH



## Tổng quan

- Số lượng: 117010 luật
- Thời gian: 17 phút



## Luật liên quan đến quyết định rời bỏ

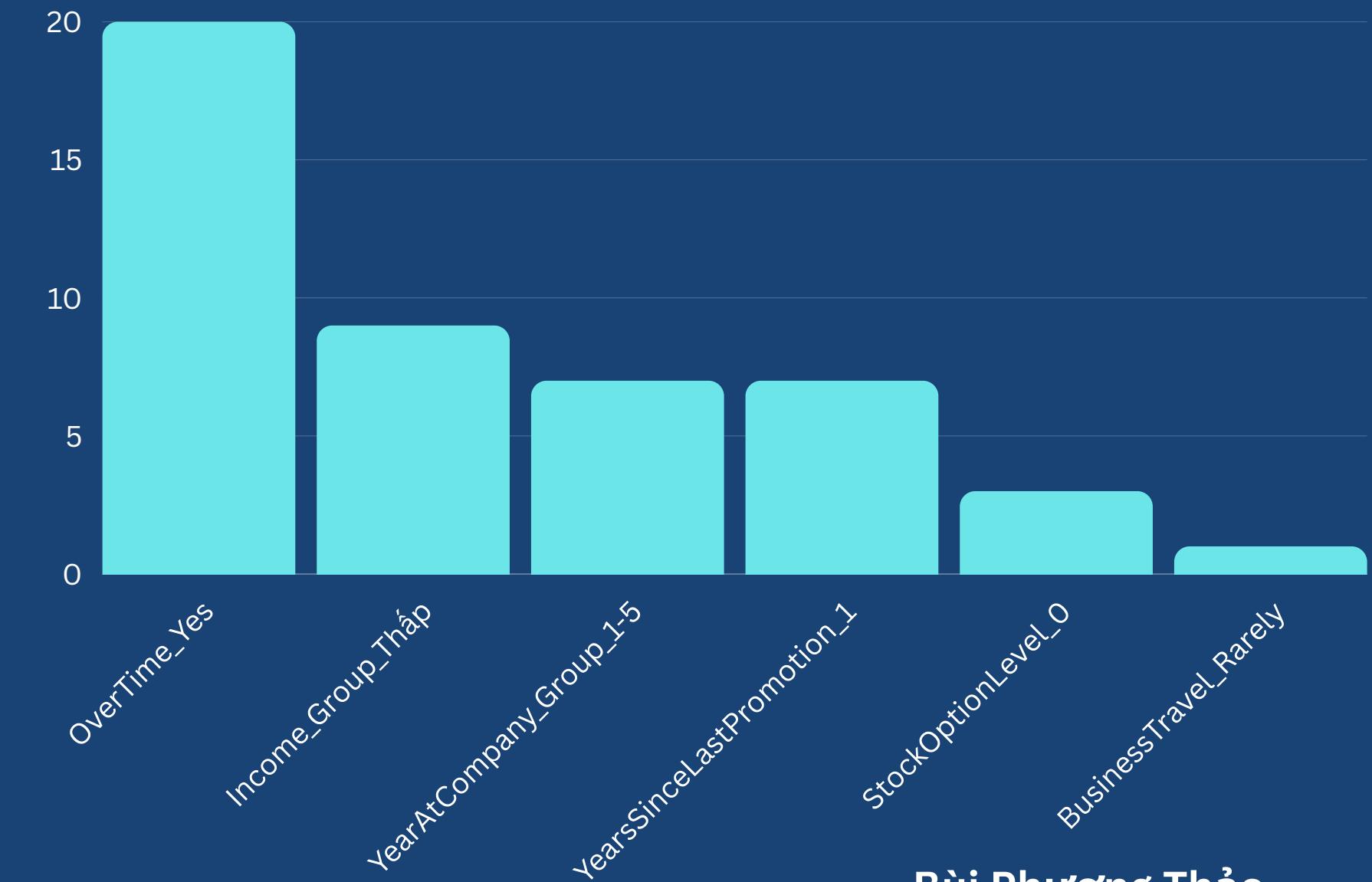
- Số lượng: 33350
- Min confidence= 0.6 (28%)

# FP-GROWTH



## Attrition\_Yes

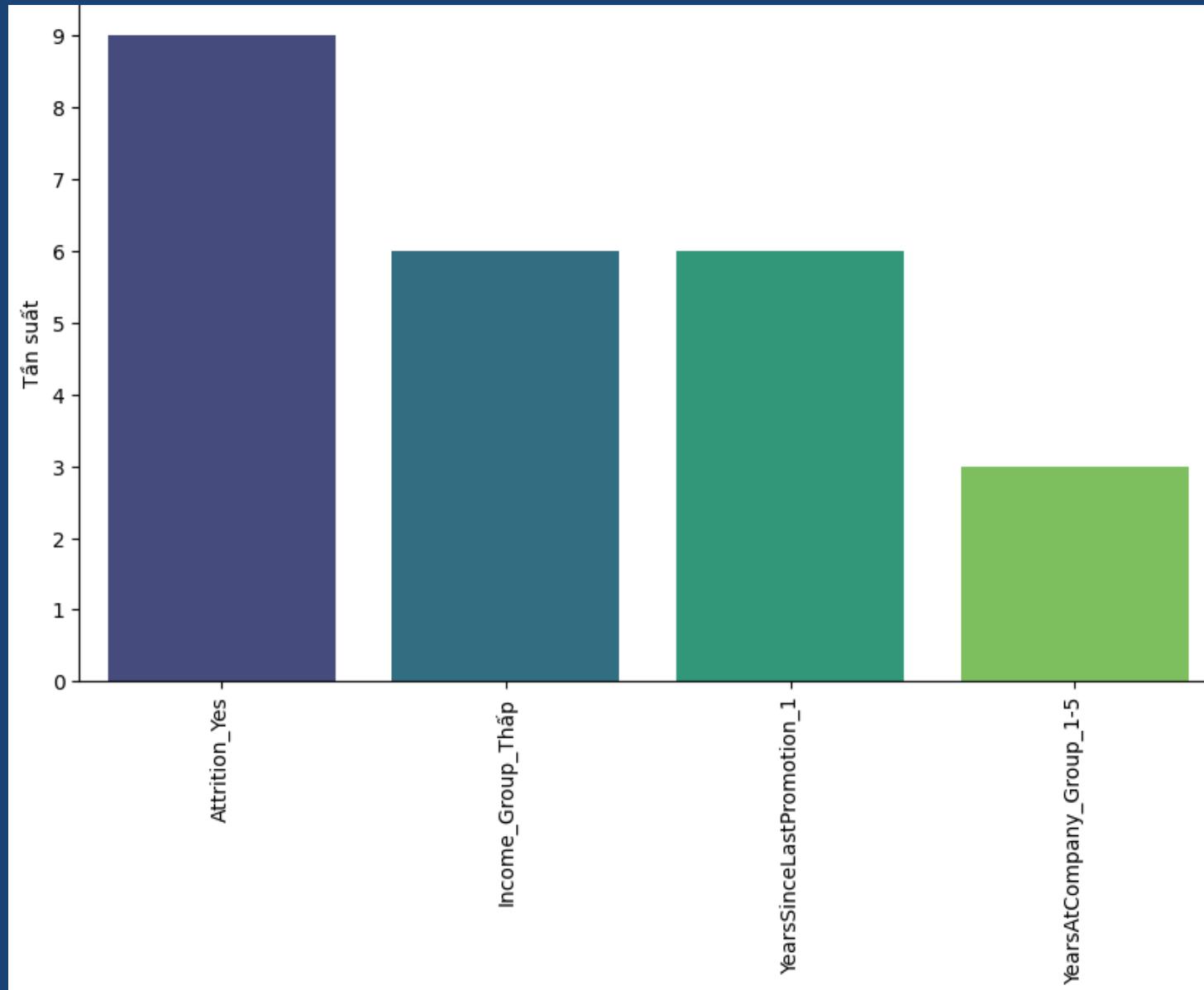
- 20 luật với confidence <0,77
- Lift >2.6
- 100% nhân viên rời bỏ phải tăng ca
- Là những nhân viên lương thấp, mới gắn bó với công ty
- Vẽ phải xuất hiện YearsAtCompany\_Group\_1-5: gắn bó với công ty 1-5 năm



# FP-GROWTH

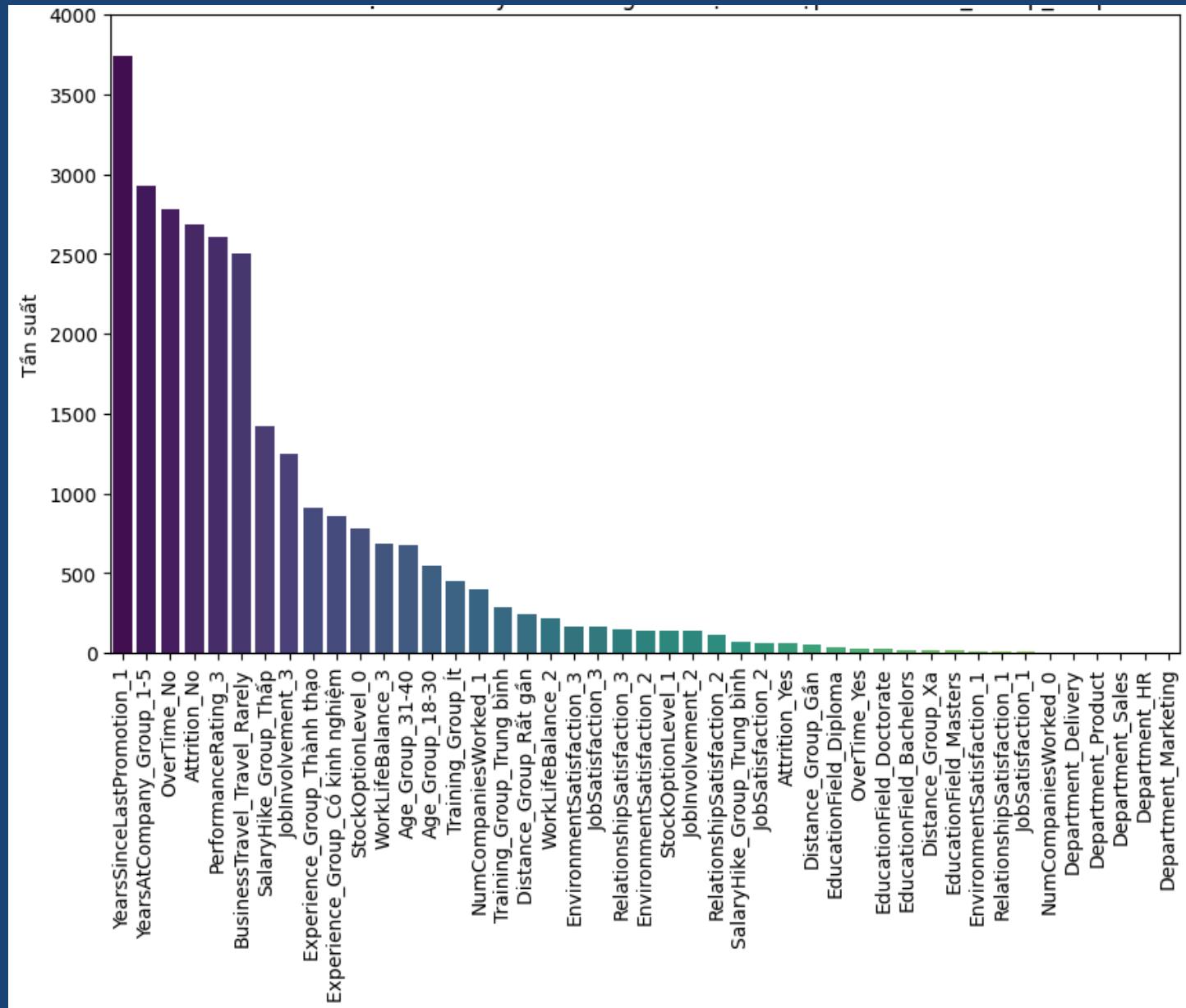


## OverTime\_Yes



- 9 luật với confidence <0.64, lift>2.7
- 100% nhân viên rời bỏ phải tăng ca
- Là những nhân viên **lương thấp, mới gắn bó với công ty**

# FP-GROWTH



## Income\_Group\_Thấp

- 7037 luật
- Là những nhân viên có kinh nghiệm nhưng còn trẻ, khối lượng công việc ít
- Gần 1000 nhân viên có kinh nghiệm cao (Thành thạo) nhưng mức lương thấp

# FP-GROWTH



## Attrition\_No

- 4421 luật với confidence >0,97
- Tương tự apriori: đây là những nhân viên không mấy áp lực về công việc
- 3 yếu tố phổ biến: PerformanceRating\_3 (3400), Overtime\_No (2600), BusinessTravel\_Rarely (2000)

# SO SÁNH

	Apriori (0.1)	FP-Growth (0.6)
Thời gian	5 phút	17 phút
Số lượng luật	189862	117010
Số lượng luật liên quan đến rời bỏ	3	20
Chất lượng (Trung bình Lift)	1.13	1.15

Với min confidence cao hơn (0,6) nhưng thuật toán FP-Growth đưa ra được nhiều luật có ý nghĩa hơn.

# KẾT LUẬN

- Yếu tố quyết định đến quyết định rời bỏ công ty của nhân viên là **tăng ca (OverTime\_Yes)**
- Nhân viên có khối lượng công việc ít, có kinh nghiệm, không áp lực nên **gắn bó** với công ty
- Mức lương khá hợp lý, ngoại trừ vài trường hợp

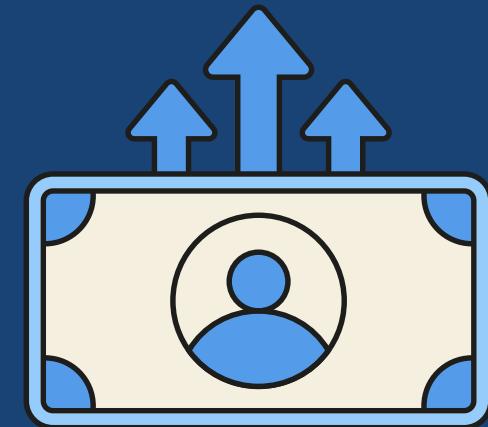


# GIẢI PHÁP



## Tăng ca

- Điều chỉnh lại thời gian tăng ca
- Phân chia tăng ca phù hợp với mỗi nhân viên, nhất là đối với nhân viên mới, trẻ



## Tăng lương

- Nên điều chỉnh lương, tăng lương cho nhân viên có kinh nghiệm.
- Thêm phúc lợi cho nhân viên làm việc năng suất.



## Phát triển kỹ năng

- Thúc đẩy hiệu suất làm việc
- Phân chia công việc đồng đều, phù hợp với năng lực

# THANK YOU