



TRƯỜNG ĐẠI HỌC QUY NHƠN
KHOA TOÁN - THỐNG KÊ



HR ANALYTICS

NHÓM 1

Nguyễn Quang Nghĩa

Lê Quang Kiệt

Võ Quốc Dũng

Nguyễn Trọng Quý

Mentor: Lê Đ. Việt

TỔNG QUAN VỀ DỮ LIỆU

I.TỔNG QUAN

- Tập data: "HR-Employee-Attrition.csv", chứa thông tin về nhân viên trong công ty, tập trung vào các yếu tố liên quan đến **tình trạng nghỉ việc của nhân viên**.
- Gồm 1470 dòng (thông tin nhân viên) và 35 cột (thuộc tính)
- Bộ dữ liệu HR Employee Attrition cung cấp một nguồn thông tin phong phú để phân tích các yếu tố ảnh hưởng đến tình trạng nghỉ việc của nhân viên. Với sự đa dạng của các biến số, dữ liệu trên mở ra nhiều hướng phân tích thú vị và có giá trị thực tiễn cao trong quản trị nhân sự.

I.TỔNG QUAN

1. Mô tả các trường dữ liệu

- **Age:** Tuổi của nhân viên
- **Attrition:** Tình trạng nghỉ việc
- **Gender:** Giới tính của nhân viên
- **BusinessTravel:** Tần suất nhân viên đi công tác
- **DailyRate:** Tiền lương hàng ngày của nhân viên
- **Department:** Phòng ban nhân viên
- **DistanceFromHome:** Khoảng cách từ nhà đến văn phòng theo km
- **Education:** Trình độ của nhân viên
- **EducationField:** Lĩnh vực giáo dục
- **EmployeeCount:** Số lượng nhân viên
- **EmployeeNumber:** ID nhân viên
- **EnvironmentSatisfaction:** Sự hài lòng về môi trường làm việc
- **HourlyRate:** Mức lương theo giờ của nhân viên

I.TỔNG QUAN

- **JobInvolvement:** Mức độ tham gia công việc
- **JobLevel:** Cấp độ công việc
- **JobRole:** Vai trò công việc của nhân viên
- **JobSatisfaction:** Mức độ hài lòng trong công việc của nhân viên
- **MaritalStatus:** Tình trạng hôn nhân
- **MonthlyIncome:** Thu nhập của nhân viên
- **MonthlyRate:** Mức lương hàng tháng của nhân viên
- **NumCompaniesWorked:** Số công ty đã làm việc
- **Over18:** Độ tuổi trên 18
- **OverTime:** Nhân viên làm thêm giờ
- **PercentSalaryHike:** Tỷ lệ được tăng lương
- **PerformanceRating:** Đánh giá hiệu suất
- **RelationshipSatisfaction:** Mức độ hài lòng trong mối quan hệ
- **StandardHours:** Giờ làm việc tiêu chuẩn mỗi tuần

I.TỔNG QUAN

- **StockOptionLevel:** Quyền chọn cổ phiếu của công ty
- **TotalWorkingYears:** Tổng số năm làm việc
- **TrainingTimesLastYear:** Số lần đào tạo năm ngoái
- **WorkLifeBalance:** Cân bằng giữa công việc và cuộc sống
- **YearsAtCompany:** Tổng số năm làm việc tại công ty
- **YearsInCurrentRole:** Tổng số năm giữ chức vụ hiện tại
- **YearsSinceLastPromotion:** Số năm kể từ lần thăng chức gần nhất
- **YearsWithCurrManager:** Số năm làm việc dưới quyền quản lý hiện tại

Thông tin về dữ liệu

- Thuộc tính mục tiêu (Target): Attrition–có 2 giá trị: Yes(nghỉ việc), No(đang làm)
- Biến số định tính (Categorical): JobRole, Department, BusinessTravel, Gender, MaritalStatus, v.v.
- Biến số định lượng (Numeric): Age, MonthlyIncome, YearsAtCompany, v.v.
- Không có giá trị thiếu (Missing value): Tất cả các cột đều đầy đủ dữ liệu.

I.TỔNG QUAN

1. Tóm tắt thông tin về Values của các features trong data:

- **Thông tin cá nhân của nhân viên**

- + **Age**: 18 - 60

- + **Gender**: Male , Female

- + **MaritalStatus**: Single , Married , Divorced

- + **Education**: 1 , 2 , 3 , 4 , 5

- + **DistanceFromHome**: 1 - 29

- + **TotalWorkingYears**: 0 - 40

- + **NumCompaniesWorked**: 0 - 9

- **Thông tin công ty về nhân viên**:

- + **YearsAtCompany**: 0 - 40

- + **YearsInCurrentRole**: 0 - 18

- + **YearsWithCurrManager**: 0 - 17

- + **YearsSinceLastPromotion**: 0 -15

- + **TrainingTimesLastYear**: 0 - 6

- + **WorkLifeBalance**: 1, 2, 3, 4

I.TỔNG QUAN

- **Thông tin công việc của nhân viên:**
 - + **EducationField:** Human Resources, Life Sciences, Marketing, Medical, Other, Technical Degree
 - + **Department:** Human Resources, Research & Development, Sales
 - + **JobLevel:** 1, 2, 3, 4, 5
 - + **JobRole:** Healthcare Representative, Human Resources, Laboratory Technician, Manager, Manufacturing Director, Research Director, Research Scientist, Sales Executive, Sales Representative
 - + **JobInvolvement:** 1, 2, 3, 4
 - + **OverTime:** Yes, No
 - + **JobSatisfaction:** 1, 2, 3, 4

I.TỔNG QUAN

- **Thông tin về công ty:**

- + **PercentSalaryHike:** 11 - 25

- + **StockOptionLevel:** 0, 1, 2, 3

- + **BusinessTravel:** Non-Travel, Travel_Frequently,
Travel_Rarely

- + **PerformanceRating:** 3, 4

- + **EnvironmentSatisfaction:** 1, 2, 3, 4

- + **RelationshipSatisfaction:** 1, 2, 3, 4

- **Thông tin về lương thưởng:**

- + **MonthlyIncome:** 1k - 20k

- + **HourlyRate:** 30 - 100

- + **DailyRate:** 100 - 1500

- + **MonthlyRate:** 2000 - 27000

II. XỬ LÝ DỮ LIỆU

1. Xử lý dữ liệu trùng lắp

- Trước khi thực hiện phần xử lý dữ liệu, ta thực hiện phân chia dữ liệu thành 2 tập *train-test*

✓ Phân chia tập train - test



```
X = df_attrition.drop('Attrition', axis=1)
y = df_attrition['Attrition']
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

```
[ ] #Concat X data and y data
train_df_attrition = pd.concat([X_train, y_train], axis=1)
test_df_attrition = pd.concat([X_test, y_test], axis=1)
#Save data
train_df_attrition.to_csv(path + '/data/processed/train_data_attrition.csv', index=False)
test_df_attrition.to_csv(path + '/data/processed/test_data_attrition.csv', index=False)
```

II. XỬ LÝ DỮ LIỆU

Một số thông tin về dữ liệu bao gồm: Tên các features, Số lượng giá trị Null, Kiểu dữ liệu

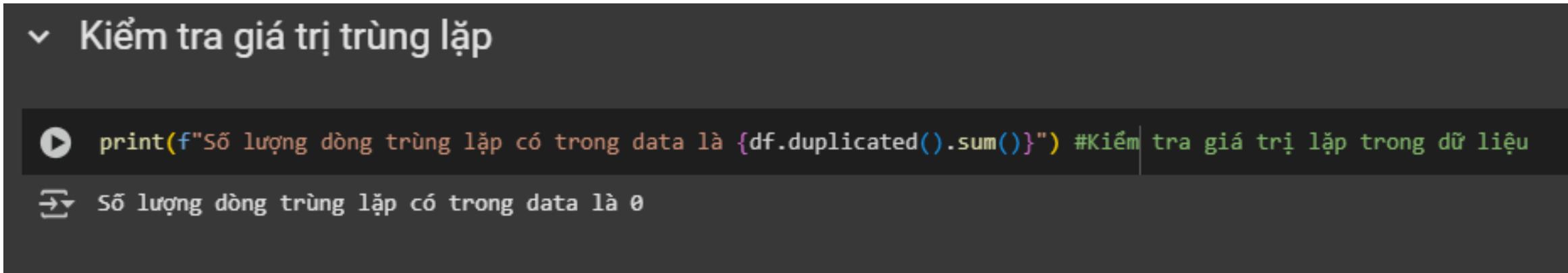
```
df = train_df_attrition
df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 1176 entries, 1097 to 1126
Data columns (total 35 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Age              1176 non-null    int64  
 1   BusinessTravel   1176 non-null    object  
 2   DailyRate        1176 non-null    int64  
 3   Department       1176 non-null    object  
 4   DistanceFromHome 1176 non-null    int64  
 5   Education        1176 non-null    int64  
 6   EducationField   1176 non-null    object  
 7   EmployeeCount   1176 non-null    int64  
 8   EmployeeNumber   1176 non-null    int64  
 9   EnvironmentSatisfaction 1176 non-null    int64  
 10  Gender            1176 non-null    object  
 11  HourlyRate       1176 non-null    int64  
 12  JobInvolvement  1176 non-null    int64  
 13  JobLevel         1176 non-null    int64  
 14  JobRole          1176 non-null    object  
 15  JobSatisfaction 1176 non-null    int64  
 16  MaritalStatus   1176 non-null    object  
 17  MonthlyIncome   1176 non-null    int64  
 18  MonthlyRate     1176 non-null    int64
```

12	JobInvolvement	1176 non-null	int64
13	JobLevel	1176 non-null	int64
14	JobRole	1176 non-null	object
15	JobSatisfaction	1176 non-null	int64
16	MaritalStatus	1176 non-null	object
17	MonthlyIncome	1176 non-null	int64
18	MonthlyRate	1176 non-null	int64
19	NumCompaniesWorked	1176 non-null	int64
20	Over18	1176 non-null	object
21	Overtime	1176 non-null	object
22	PercentSalaryHike	1176 non-null	int64
23	PerformanceRating	1176 non-null	int64
24	RelationshipSatisfaction	1176 non-null	int64
25	StandardHours	1176 non-null	int64
26	StockOptionLevel	1176 non-null	int64
27	TotalWorkingYears	1176 non-null	int64
28	TrainingTimesLastYear	1176 non-null	int64
29	WorkLifeBalance	1176 non-null	int64
30	YearsAtCompany	1176 non-null	int64
31	YearsInCurrentRole	1176 non-null	int64
32	YearsSinceLastPromotion	1176 non-null	int64
33	YearsWithCurrManager	1176 non-null	int64
34	Attrition	1176 non-null	object
dtypes: int64(26), object(9)			
memory usage: 363.0+ KB			

II. XỬ LÝ DỮ LIỆU

- Thực hiện xử lý dữ liệu trùng lặp



The screenshot shows a Jupyter Notebook cell with the following content:

```
▶ print(f"Số lượng dòng trùng lặp có trong data là {df.duplicated().sum()}" ) #Kiểm tra giá trị lặp trong dữ liệu
```

Output:

```
→ Số lượng dòng trùng lặp có trong data là 0
```

- Sau khi kiểm tra các giá trị trùng lặp trong dữ liệu ta có kết quả “ *Số lượng dòng trùng lặp có trong data là 0* ”
- Kết luận: Dữ liệu không có giá trị trùng lặp

II. XỬ LÝ DỮ LIỆU

2. Xử lý dữ liệu trống

✓ Kiểm tra số lượng giá trị null ở mỗi trường

```
▶ print(df.isnull().sum())
```

```
Age           0  
BusinessTravel 0  
DailyRate      0  
Department     0  
DistanceFromHome 0  
Education       0  
EducationField   0  
EmployeeCount    0  
EmployeeNumber    0  
EnvironmentSatisfaction 0  
Gender          0  
HourlyRate       0  
JobInvolvement    0  
JobLevel         0  
JobRole          0  
JobSatisfaction   0  
MaritalStatus     0  
MonthlyIncome     0  
MonthlyRate       0  
NumCompaniesWorked 0  
Over18           0
```

```
Gender          0  
HourlyRate      0  
JobInvolvement    0  
JobLevel         0  
JobRole          0  
JobSatisfaction   0  
MaritalStatus     0  
MonthlyIncome     0  
MonthlyRate       0  
NumCompaniesWorked 0  
Over18           0  
Overtime          0  
PercentSalaryHike 0  
PerformanceRating 0  
RelationshipSatisfaction 0  
StandardHours     0  
StockOptionLevel    0  
TotalWorkingYears   0  
TrainingTimesLastYear 0  
WorkLifeBalance    0  
YearsAtCompany     0  
YearsInCurrentRole 0  
YearsSinceLastPromotion 0  
YearsWithCurrManager 0  
Attrition          0  
dtype: int64
```

- Sau khi kiểm tra, kết luận: **không có** giá trị nào bị bỏ trống

II. XỬ LÝ DỮ LIỆU

3. Thống kê dữ liệu

▼ Hiển thị một số thống kê mô tả

```
▶ #display descriptive statistics  
df.describe(include='int64')
```

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	Job
count	1176.000000	1176.000000	1176.000000	1176.000000	1176.0	1176.000000	1176.000000	1176.000000	1176.000000	1176.000000	1176.000000
mean	36.774660	799.410714	9.260204	2.896259	1.0	1028.283163	2.695578	66.671769	2.721939	2.028061	
std	9.203851	405.727156	8.153860	1.038552	0.0	607.598269	1.088828	20.463674	0.713442	1.103215	
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.000000	1.000000	1.000000	
25%	30.000000	464.000000	2.000000	2.000000	1.0	492.750000	2.000000	48.000000	2.000000	1.000000	
50%	36.000000	796.500000	7.000000	3.000000	1.0	1018.000000	3.000000	67.000000	3.000000	2.000000	
75%	43.000000	1157.250000	14.000000	4.000000	1.0	1569.750000	4.000000	84.000000	3.000000	3.000000	
max	60.000000	1496.000000	29.000000	5.000000	1.0	2068.000000	4.000000	100.000000	4.000000	5.000000	

II. XỬ LÝ DỮ LIỆU

4. Kiểm tra outliers

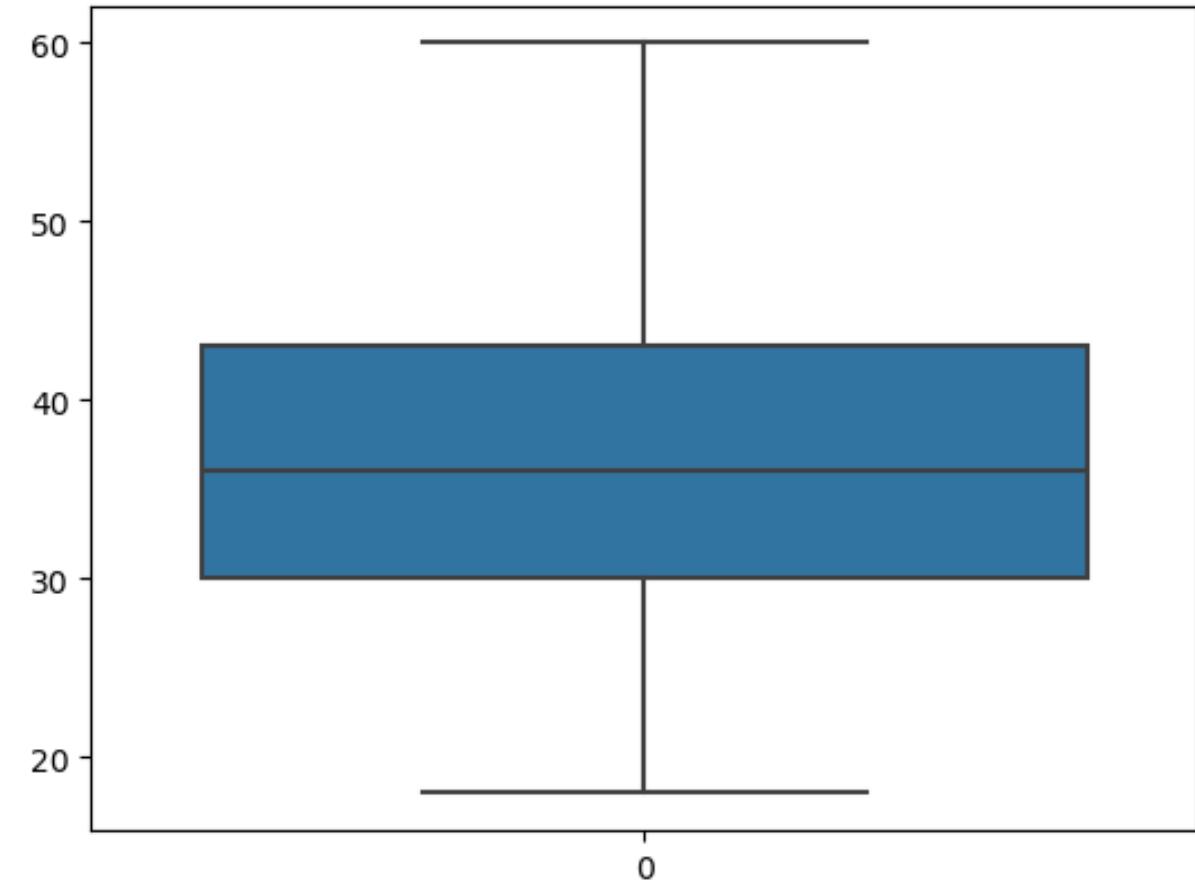
- Trước khi thực hiện việc kiểm tra outliers, ta thực hiện bước lựa chọn các trường *dữ liệu định lượng*
- Và đây là các trường mà ta đã lọc ra để thực hiện việc xử lí outliers

```
#Exclude object data
non_object_columns = df.select_dtypes(exclude='object').columns
print(non_object_columns)

Index(['Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome',
       'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike',
       'PerformanceRating', 'RelationshipSatisfaction', 'StandardHours',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

II. XỬ LÝ DỮ LIỆU

Biểu đồ outliers của cột Age



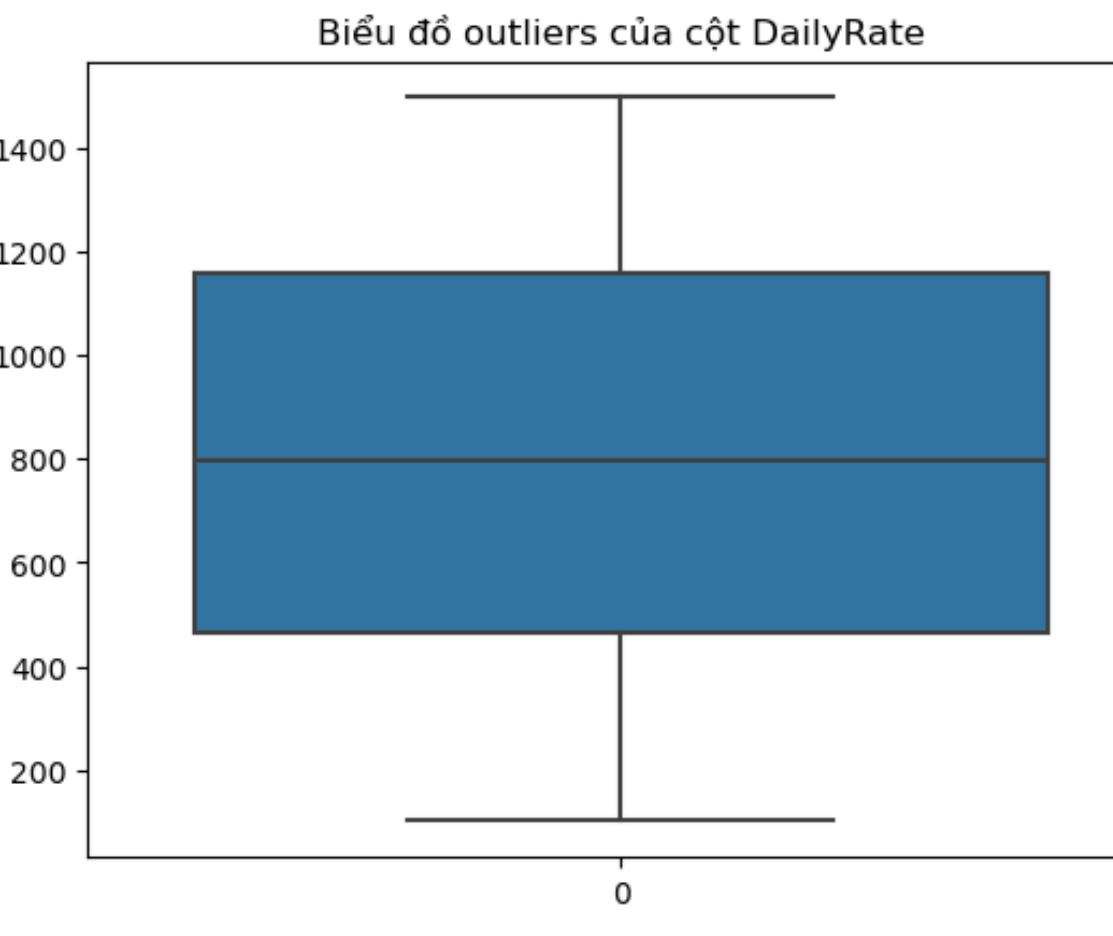
Age:

Phân bố khá đều, không có outlier rõ rệt.

Đao động từ khoảng 18 đến 60 tuổi.

Phân bố hơi nghiêng về nhóm tuổi 30 đến 40.

Biểu đồ outliers của cột DailyRate

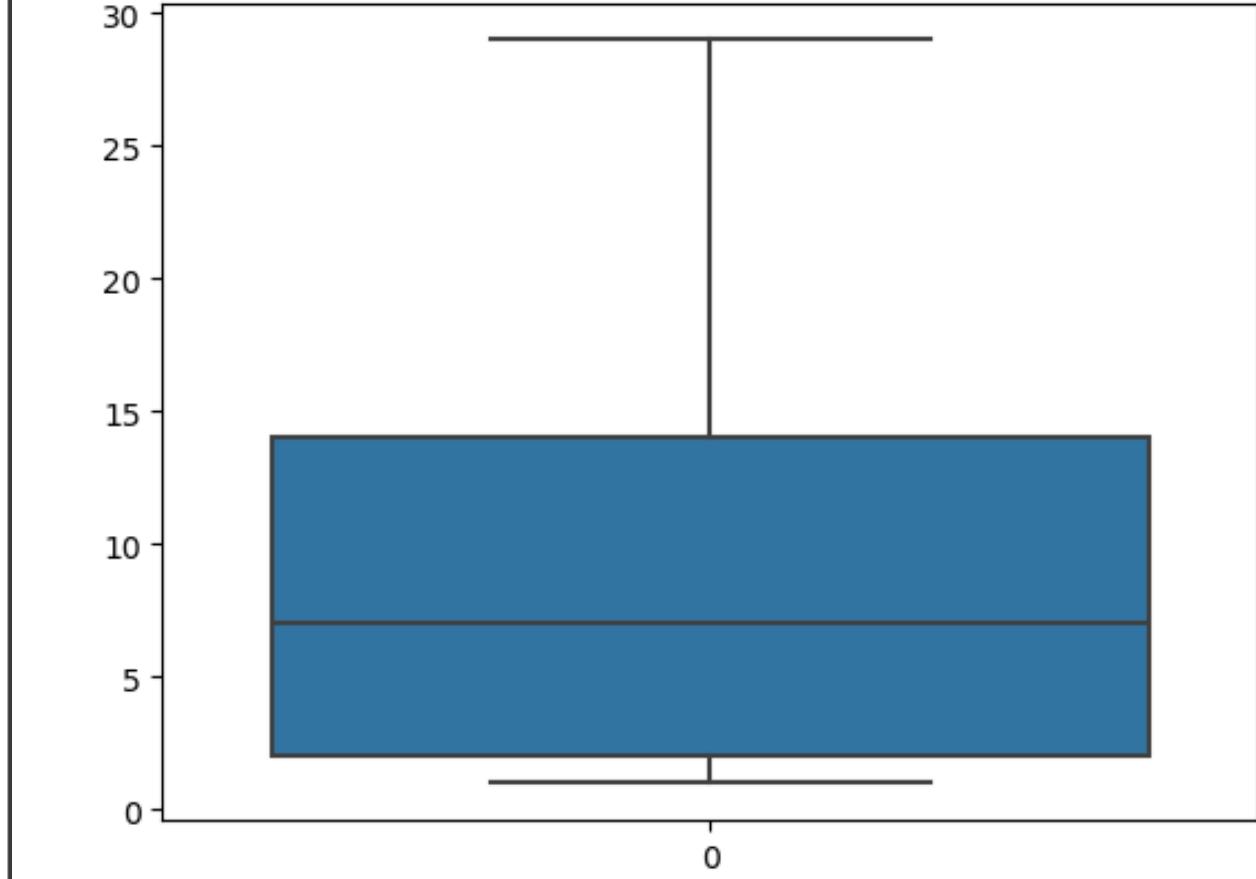


DailyRate:

Có nhiều outliers ở cả hai phía (cao và thấp).

Phân bố rộng, lương theo ngày dao động mạnh.

Biểu đồ outliers của cột DistanceFromHome



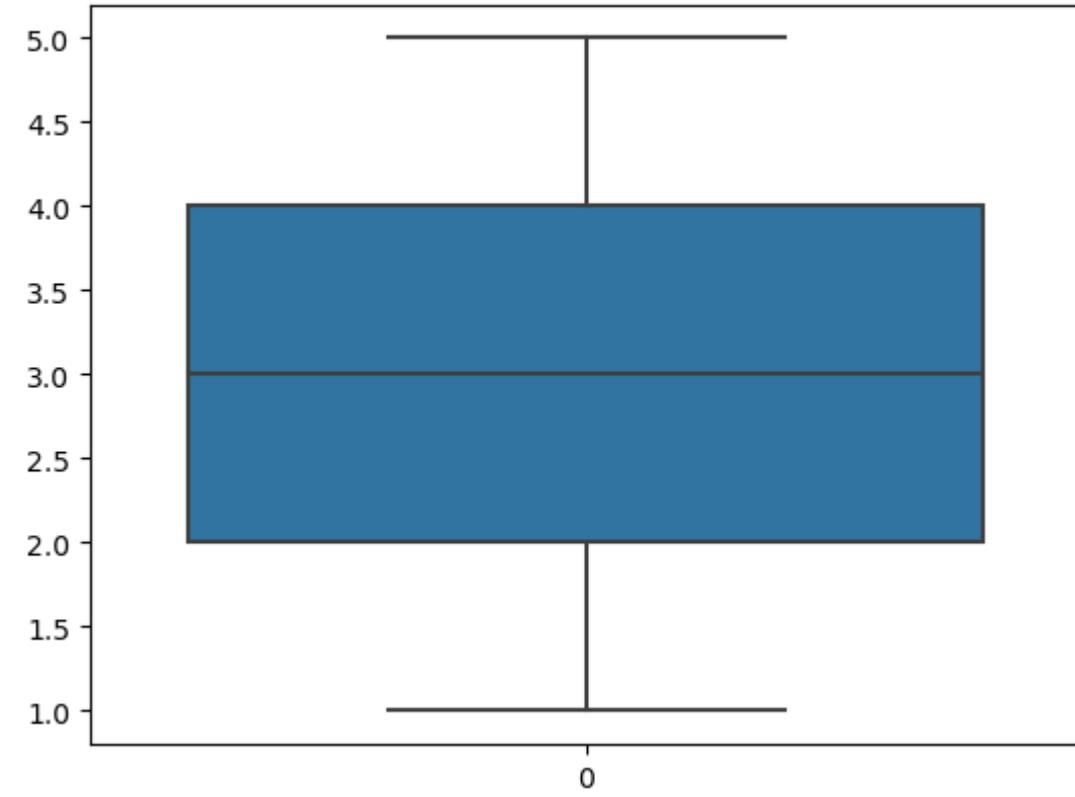
DistanceFromHome:

Có một vài outliers phía trên (> 25km).

Phần lớn nhân viên sống cách nơi làm < 20km.

II. XỬ LÝ DỮ LIỆU

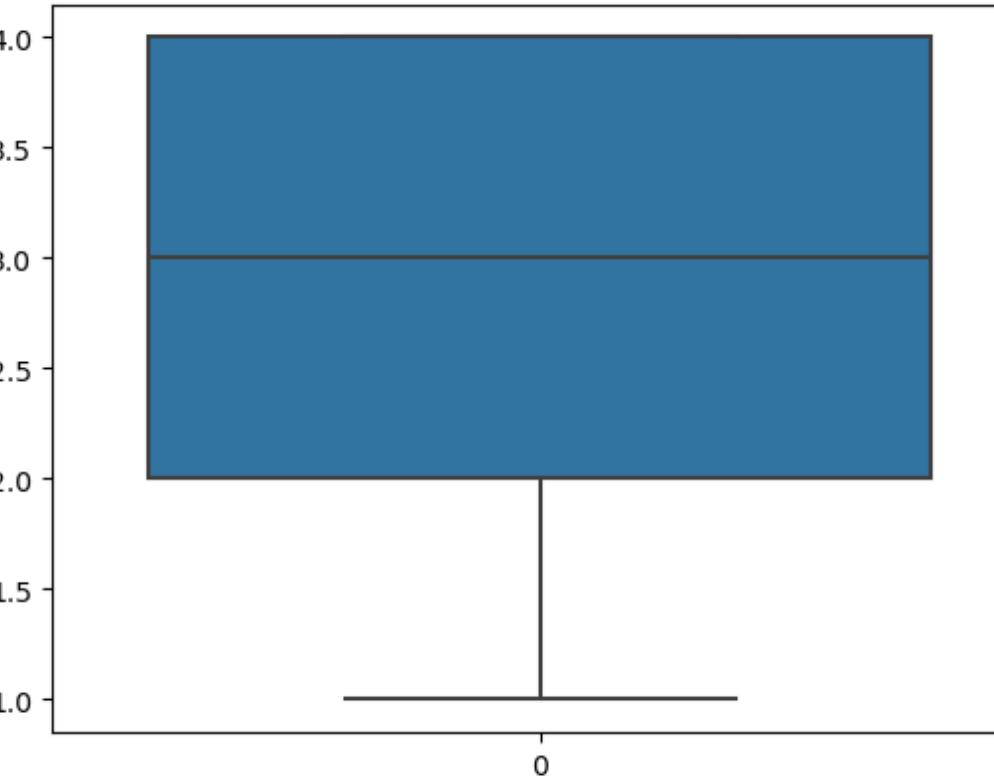
Biểu đồ outliers của cột Education



Education:

Giá trị từ 1 đến 5, không có outliers.

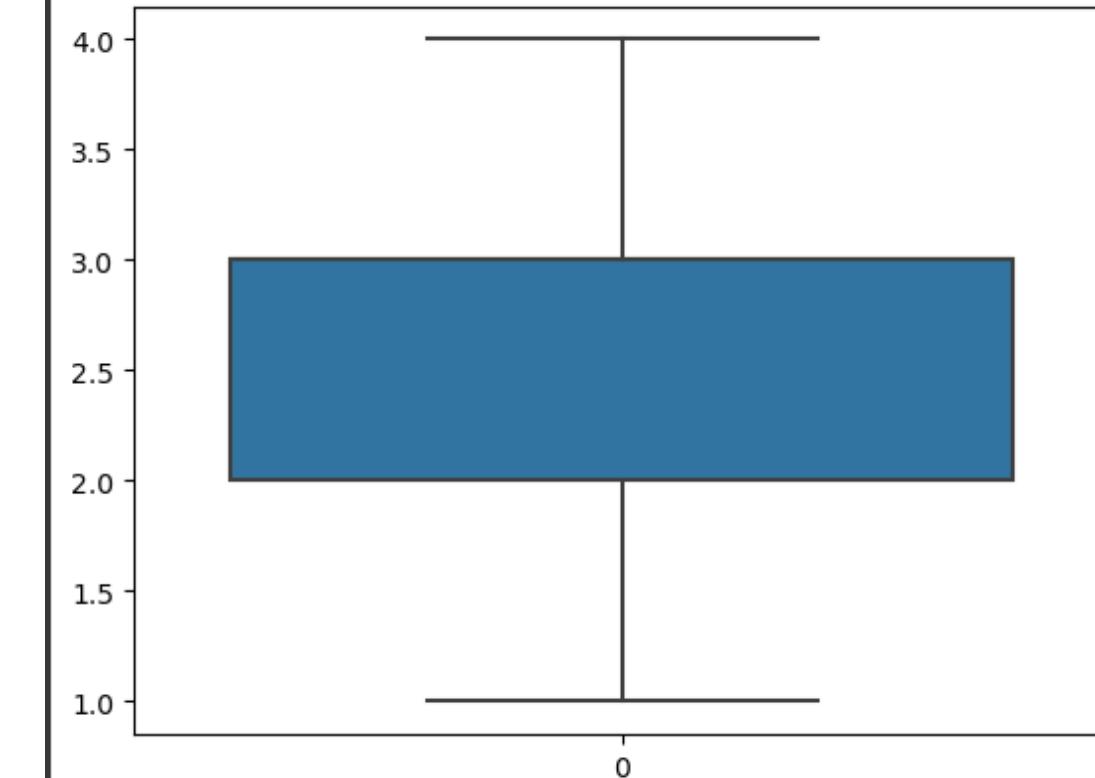
Biểu đồ outliers của cột EnvironmentSatisfaction



EnvironmentSatisfaction:

Giá trị 1 đến 4, không có outliers.

Biểu đồ outliers của cột JobInvolvement

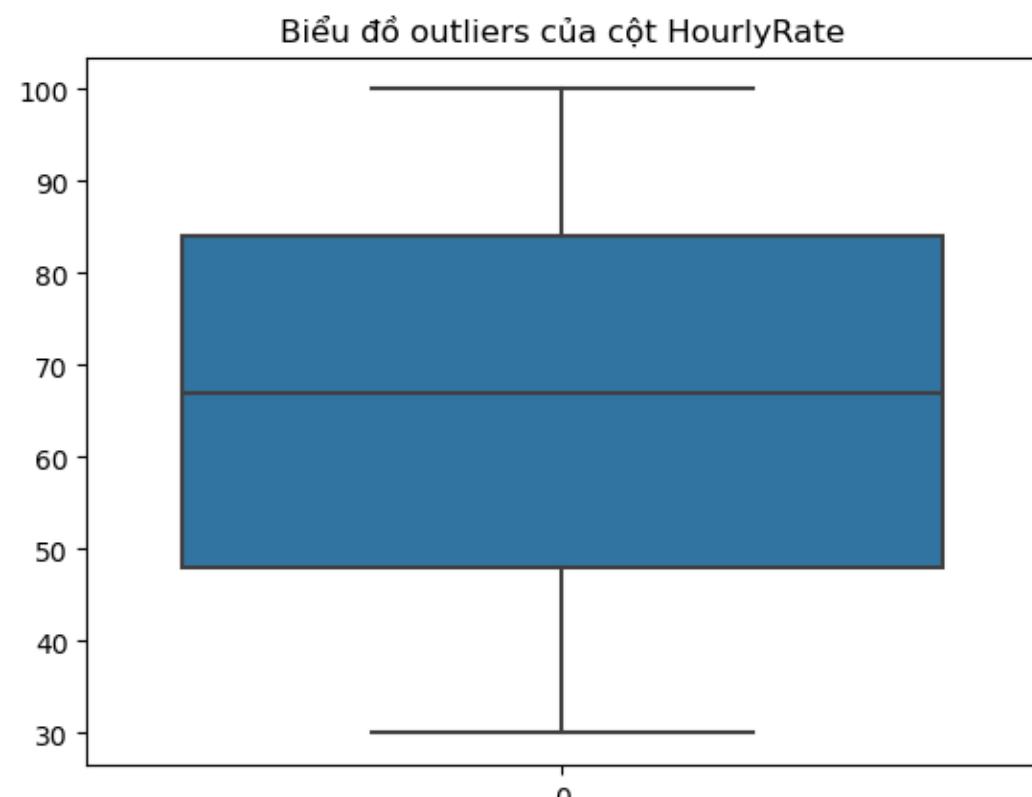


JobInvolvement:

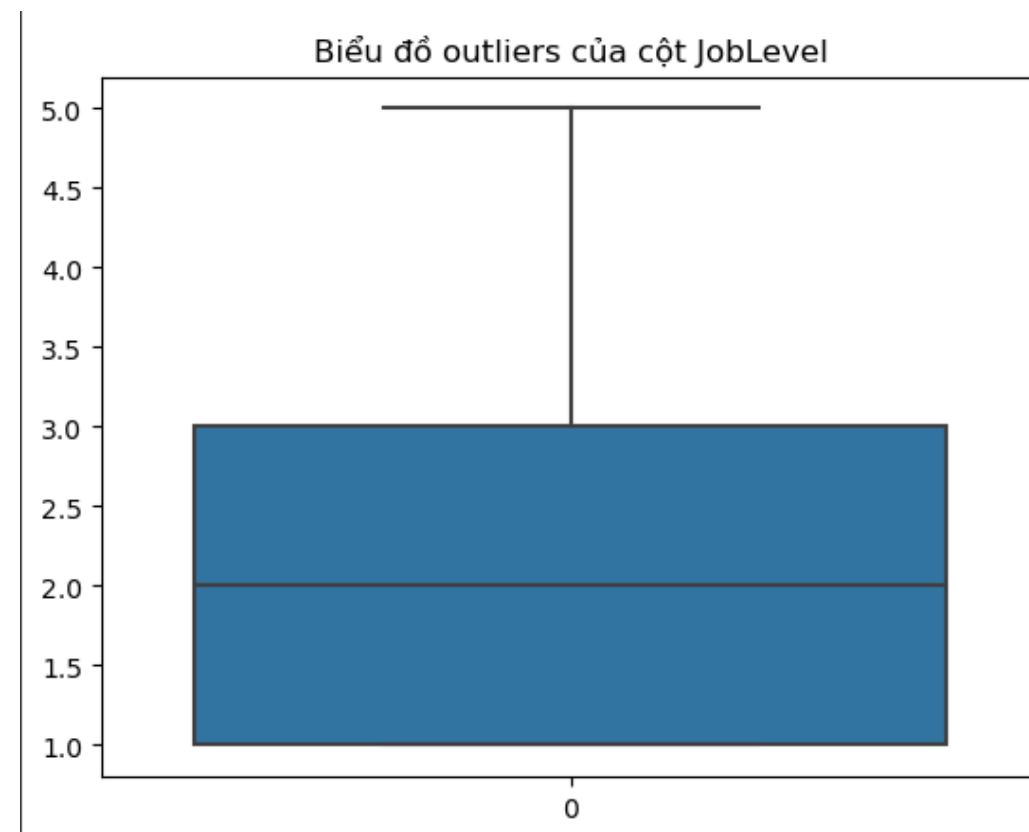
Giá trị 1 đến 4, không có outliers.

Là thang đo mức độ tham gia

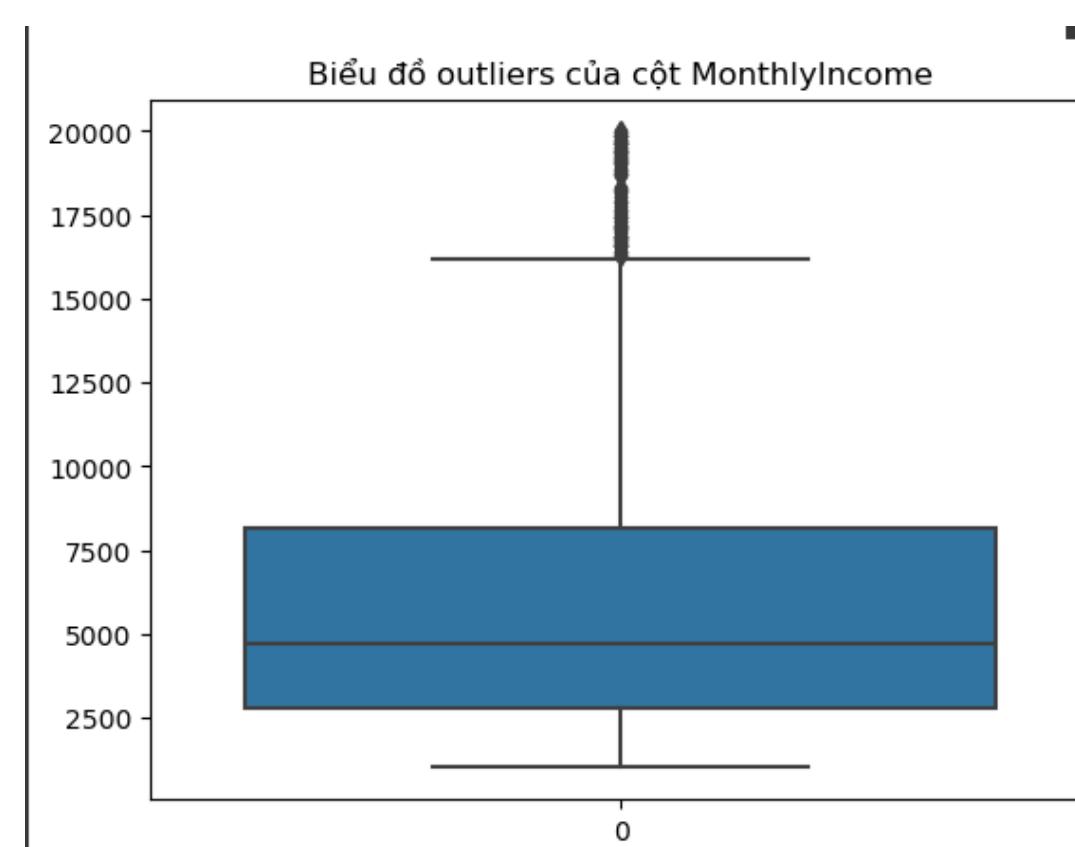
II. XỬ LÝ DỮ LIỆU



HourlyRate:
Có vài outliers nhẹ phía trên và dưới.
Lương theo giờ phân bố không đều, nhưng
không có bất thường nghiêm trọng.

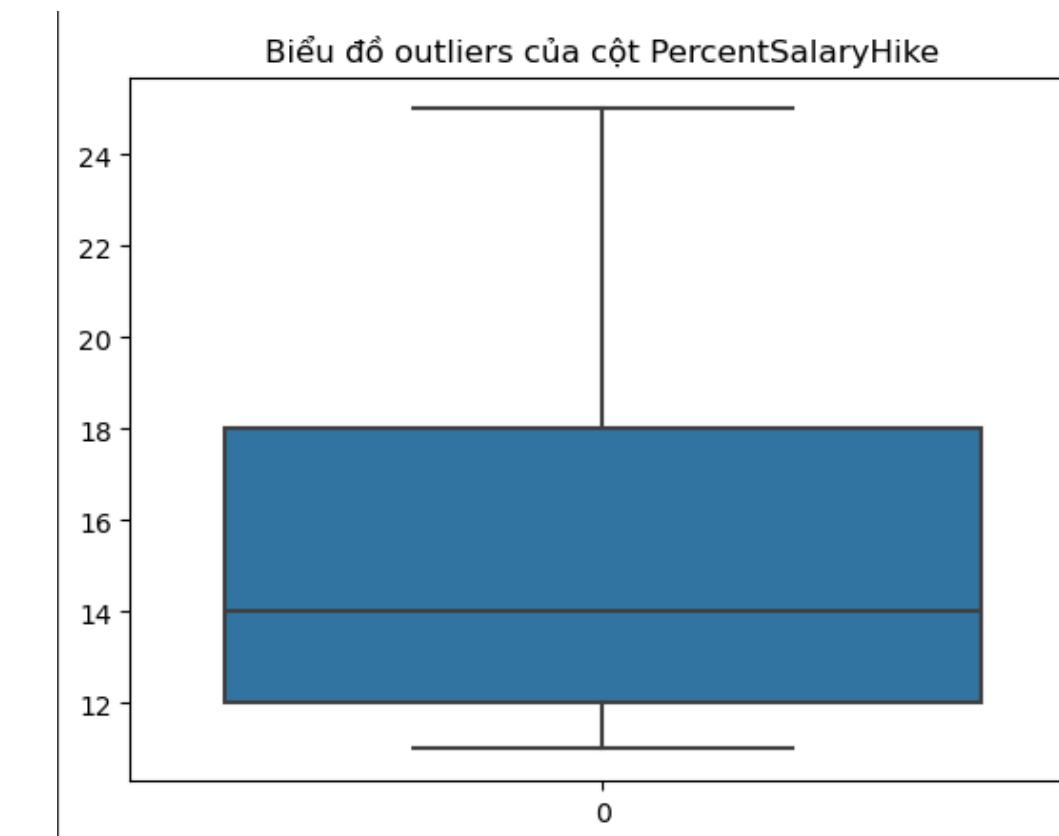
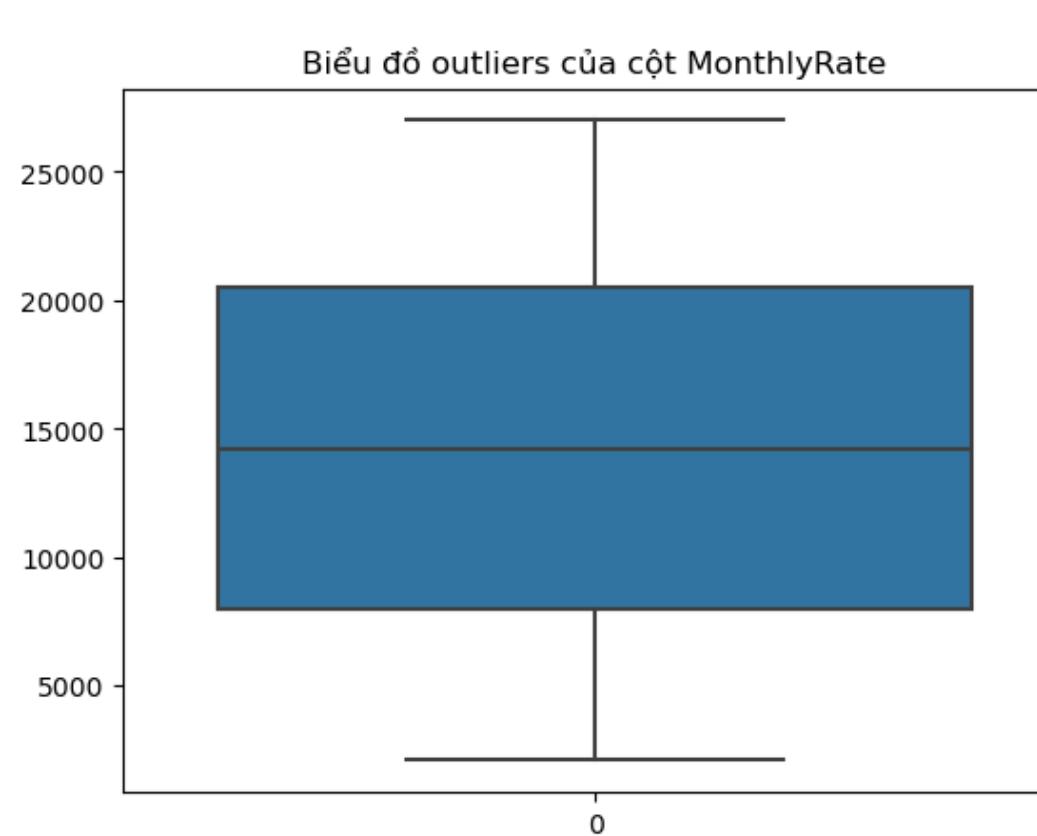
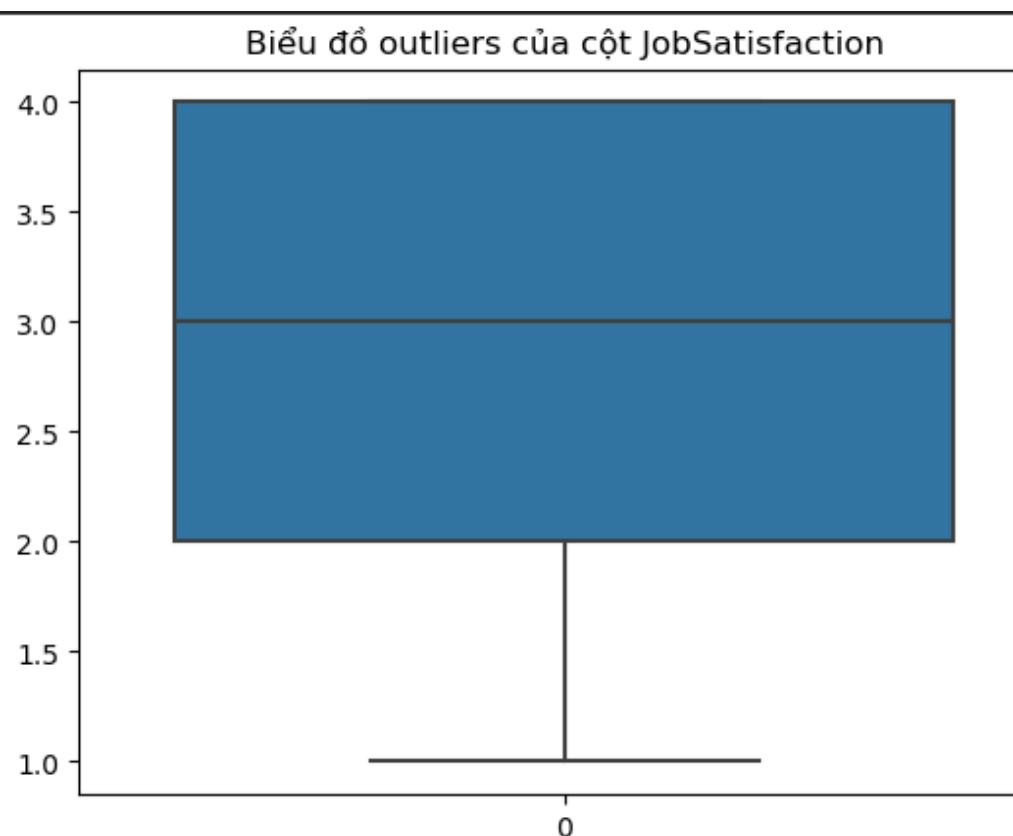


JobLevel:
Phân bố hợp lý trong khoảng 1 đến 2.
Không có outliers.



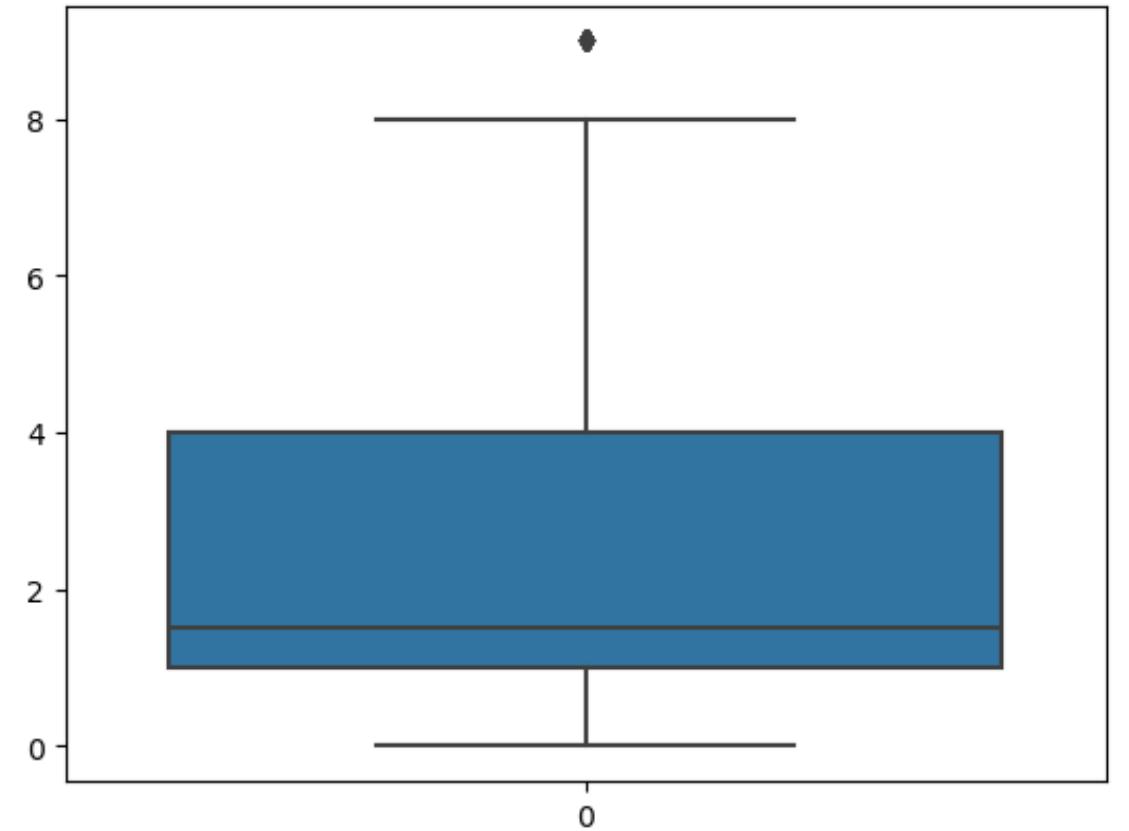
MonthlyIncome:
Có rất nhiều outliers ở phía cao.
Một số nhân viên có thu nhập > 20.000 trong
khi phần lớn < 10.000.

II. XỬ LÝ DỮ LIỆU



II. XỬ LÝ DỮ LIỆU

Biểu đồ outliers của cột NumCompaniesWorked

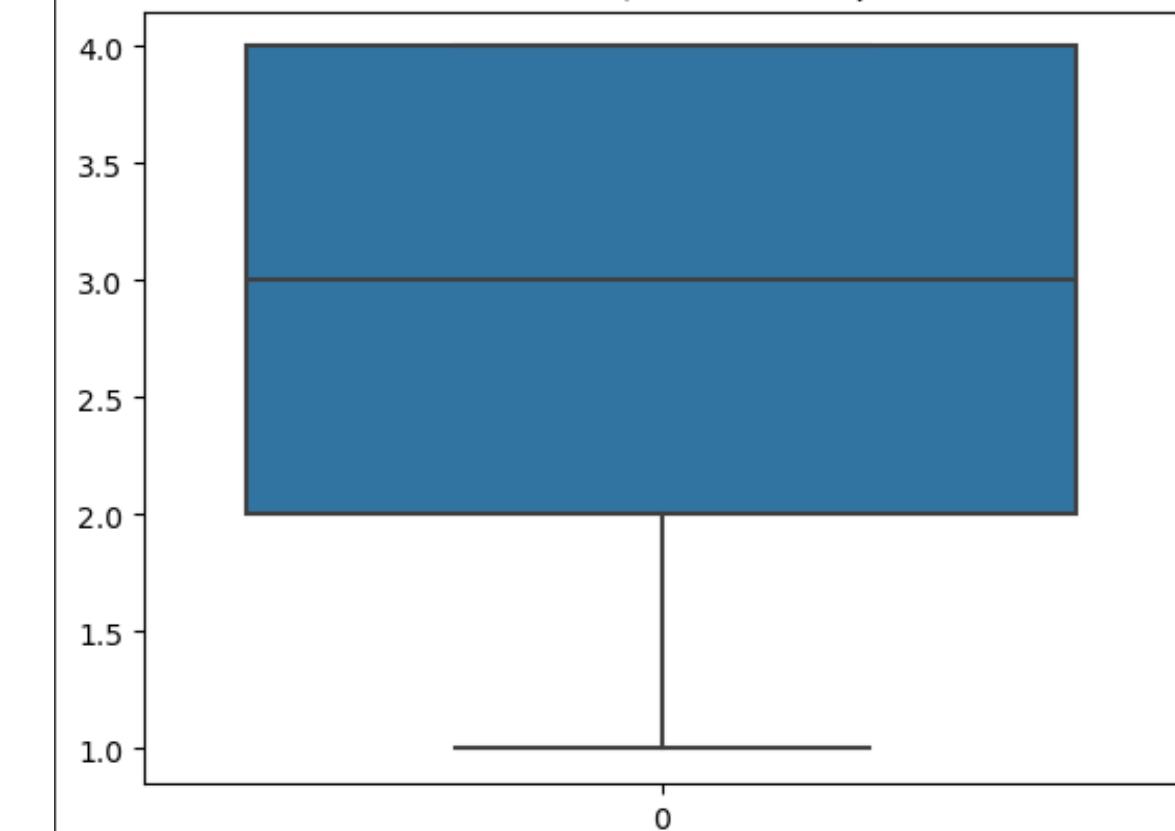


NumCompaniesWorked:

Có outliers (giá trị lớn như 9).

Đa phần nhân viên từng làm < 5 công ty.

Biểu đồ outliers của cột RelationshipSatisfaction

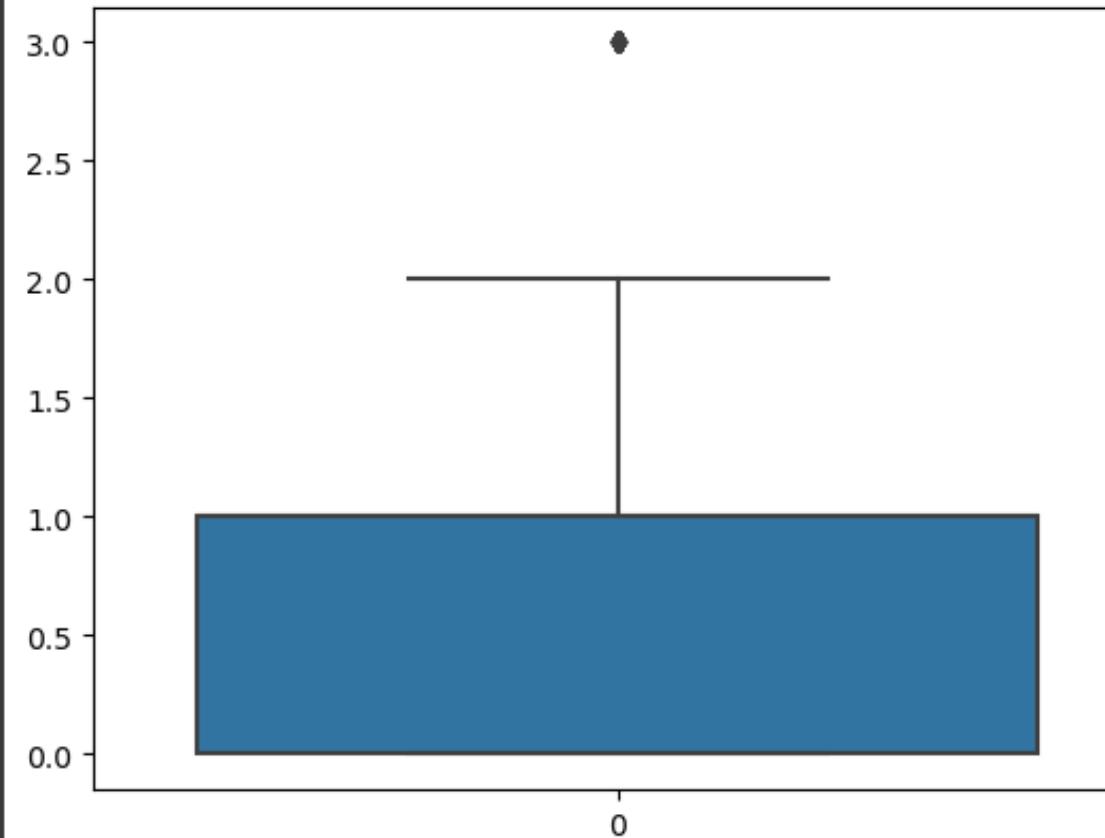


RelationshipSatisfaction:

Giá trị từ 1 đến 4, không có outlier.

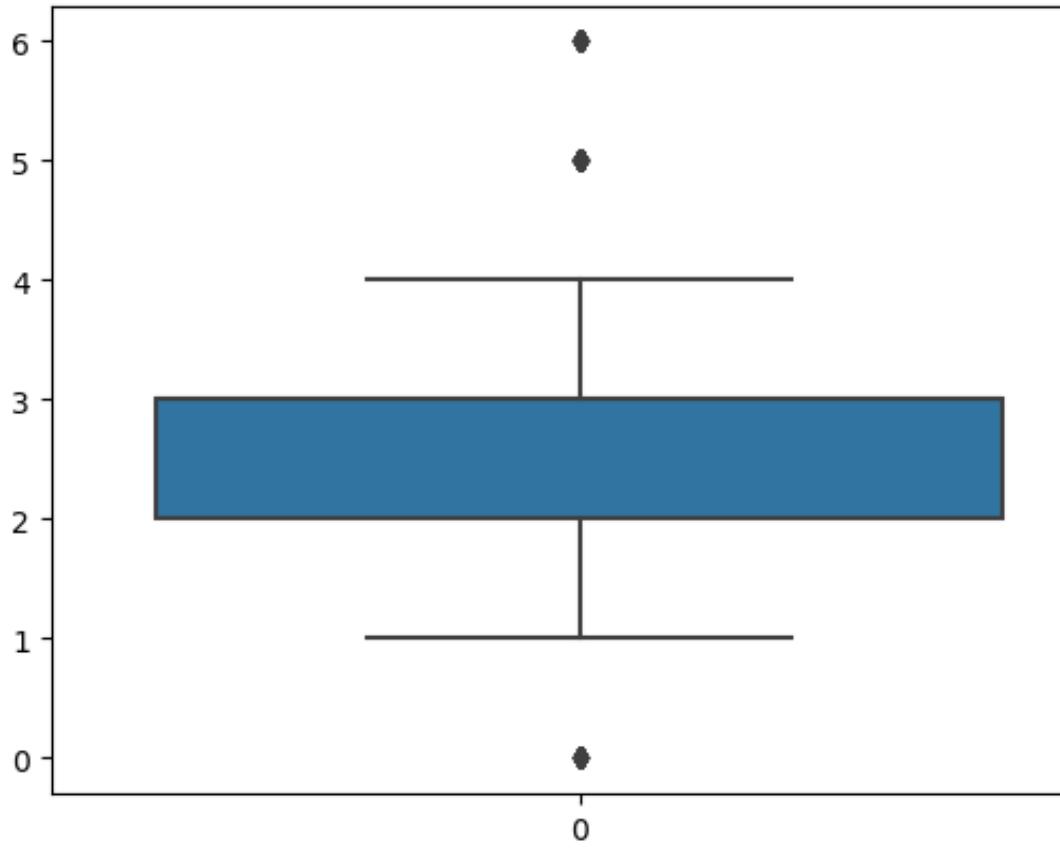
II. XỬ LÝ DỮ LIỆU

Biểu đồ outliers của cột StockOptionLevel



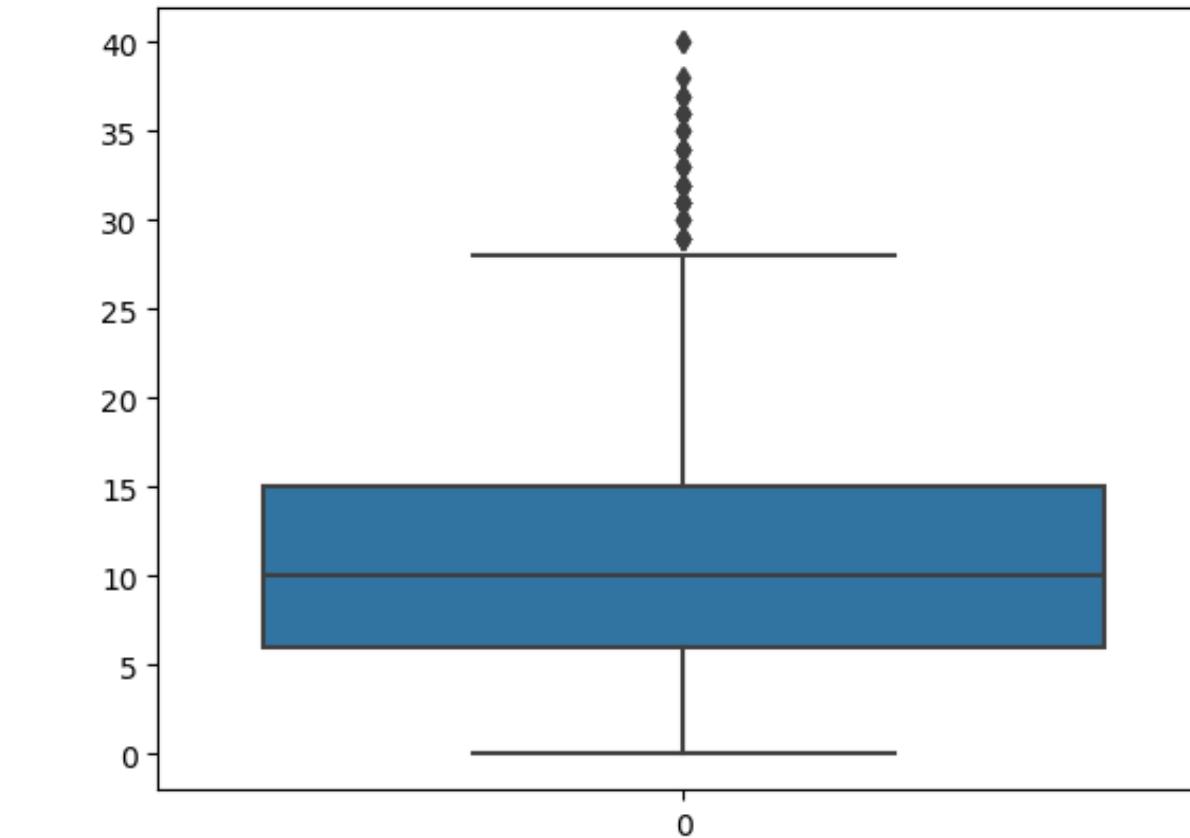
StockOptionLevel:
Giá trị từ 0 đến 3, không có outlier.

Biểu đồ outliers của cột TrainingTimesLastYear



TrainingTimesLastYear:
Giá trị 0 đến 6, có phân bố đều.

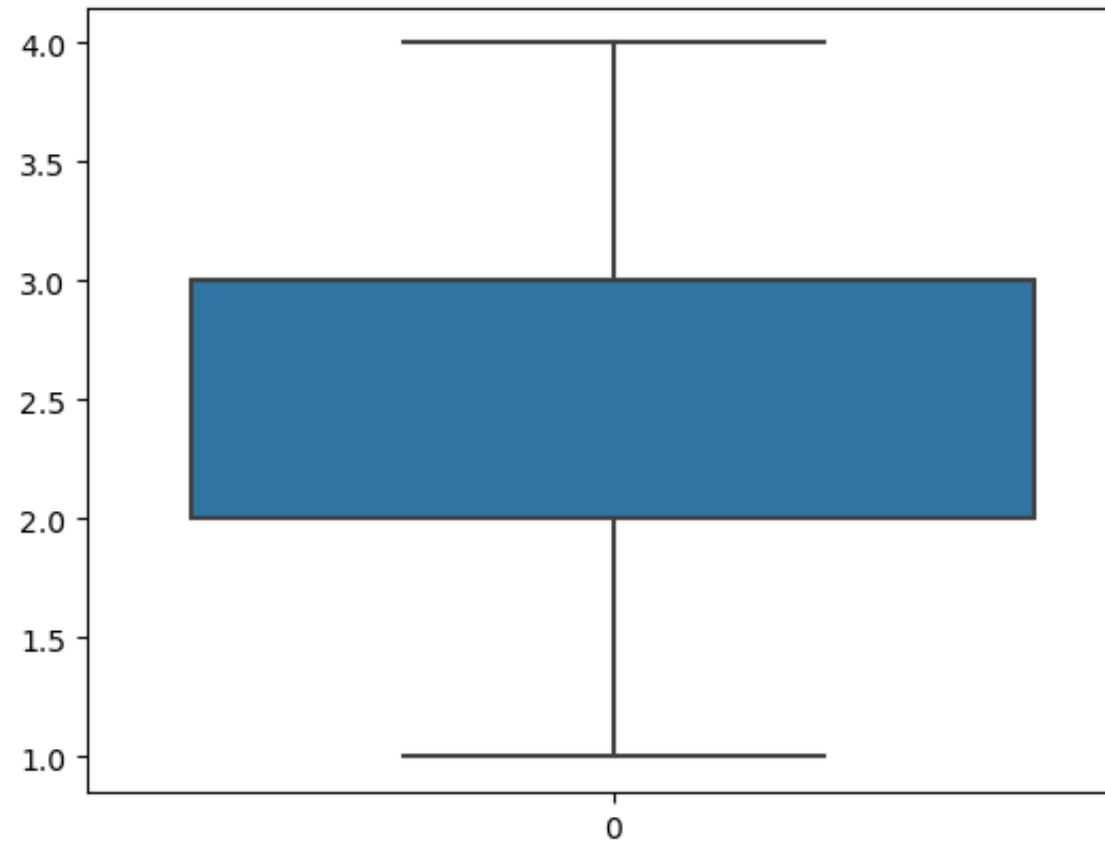
Biểu đồ outliers của cột TotalWorkingYears



TotalWorkingYears:
Có vài outliers > 35 đến 40 năm.
Phần lớn nhân viên có kinh nghiệm < 20 năm.

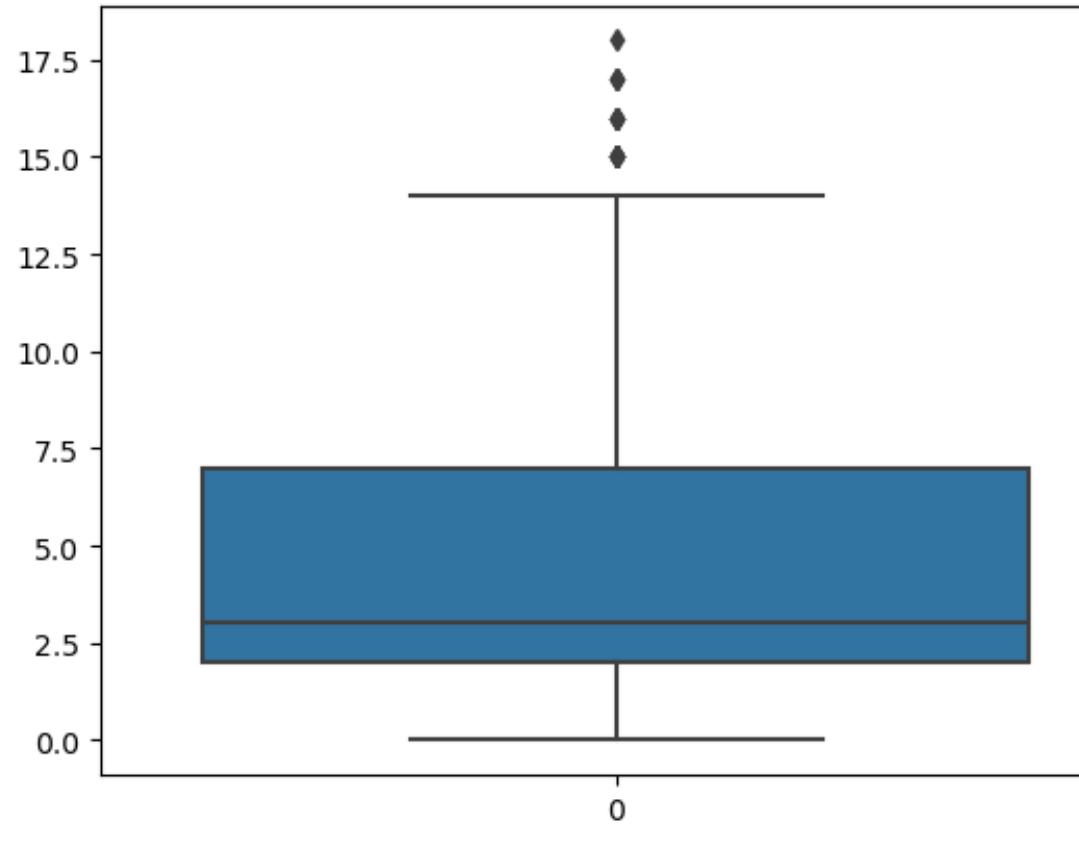
II. XỬ LÝ DỮ LIỆU

Biểu đồ outliers của cột WorkLifeBalance



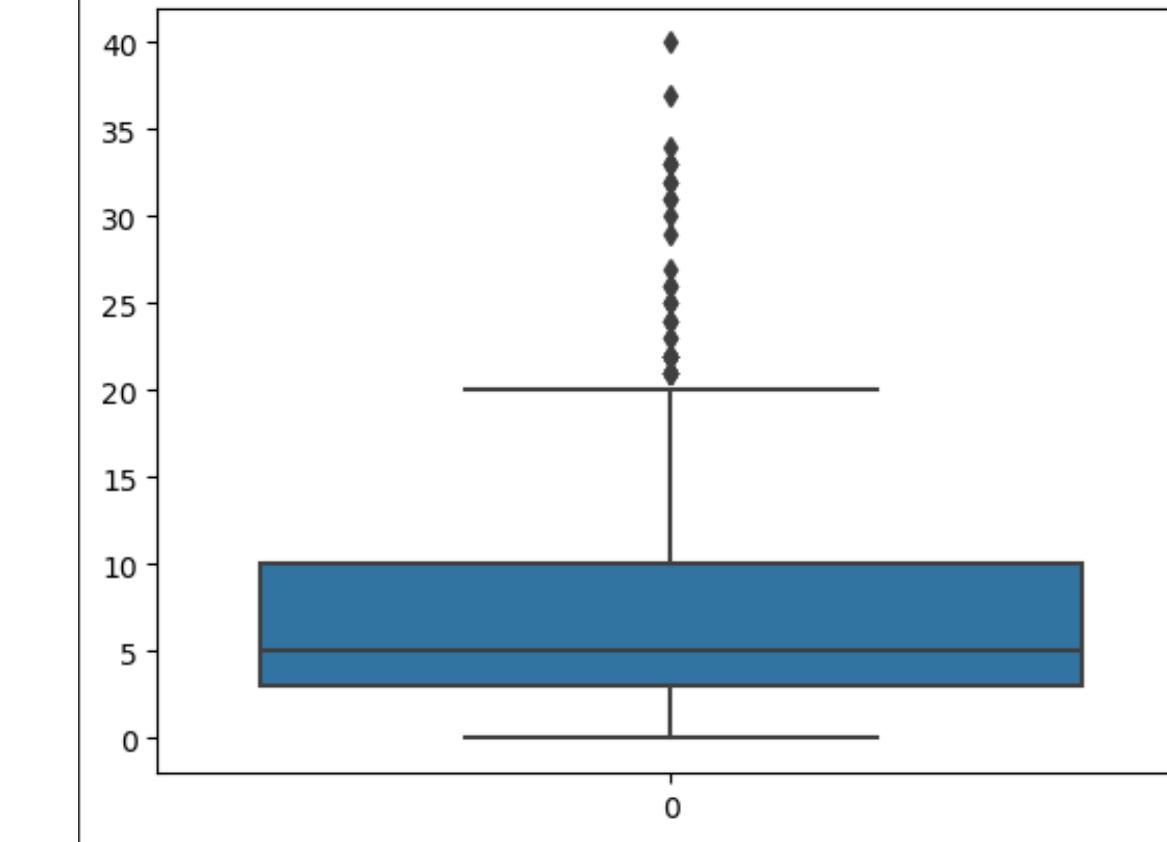
WorkLifeBalance:
Giá trị 1 đến 4, không có outliers.

Biểu đồ outliers của cột YearsInCurrentRole



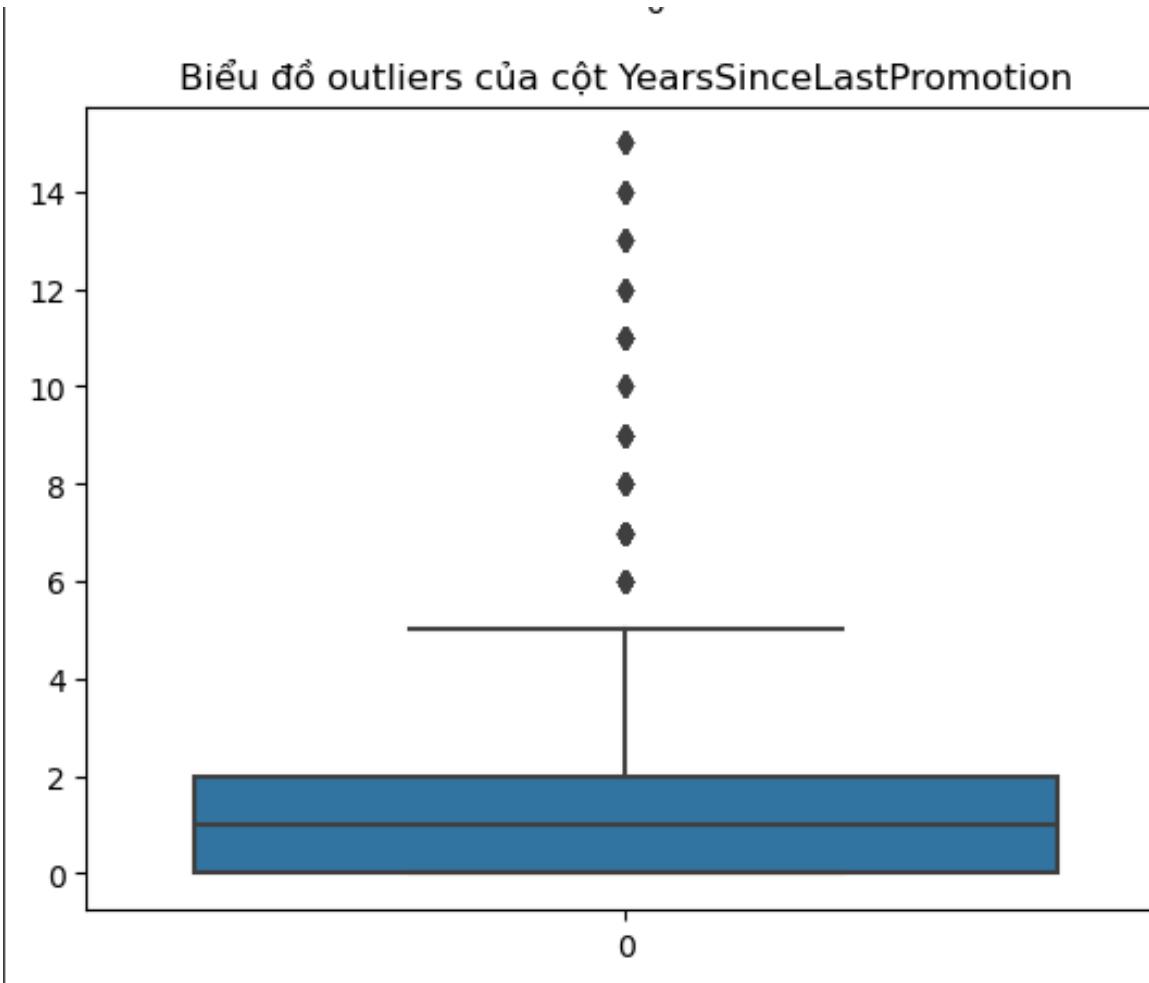
YearsInCurrentRole:
Phân bố hợp lý, có vài outliers ở giá trị cao.

Biểu đồ outliers của cột YearsAtCompany

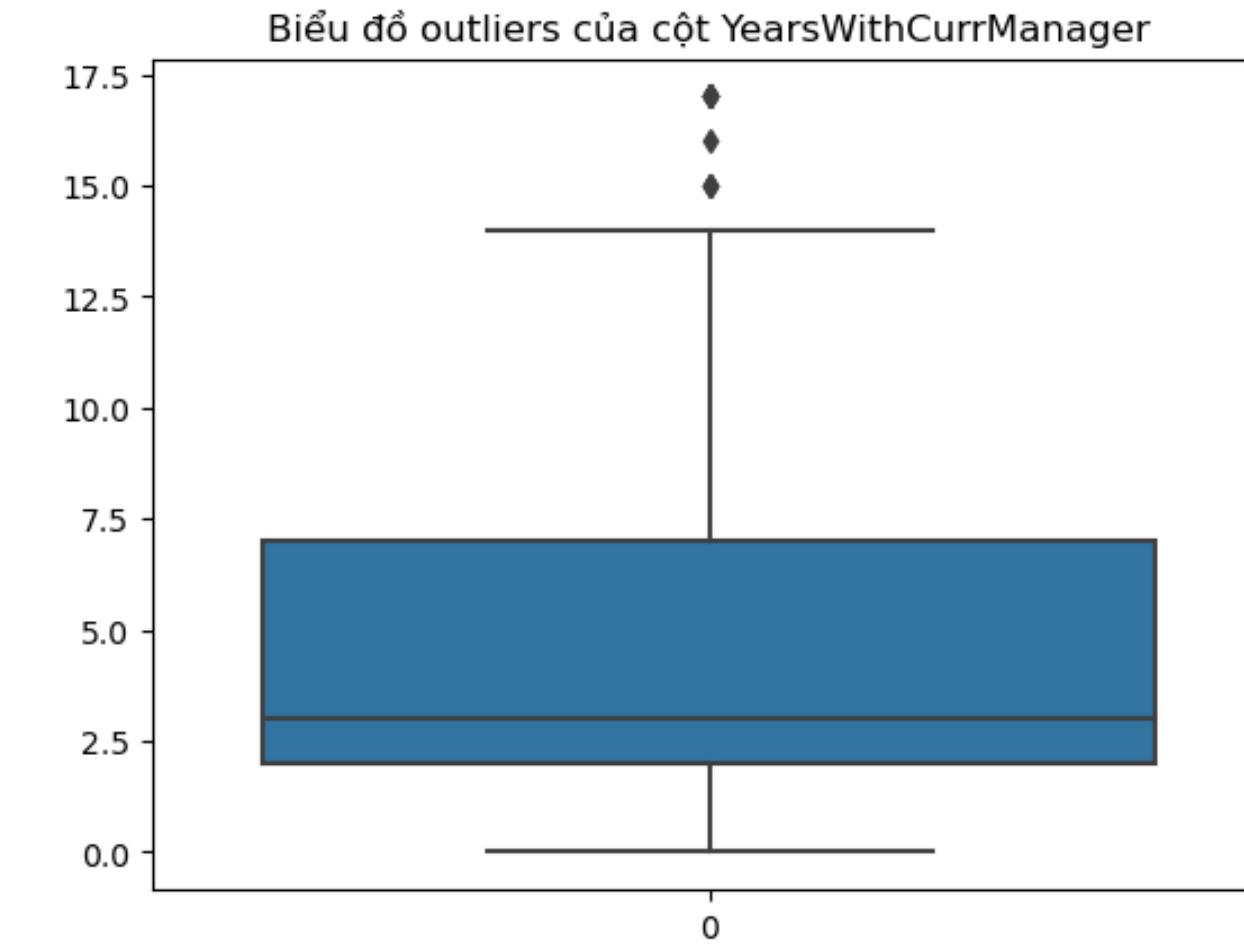


YearsAtCompany:
Có outliers > 30 năm.
Đa số làm dưới 10 năm

II. XỬ LÝ DỮ LIỆU



YearsSinceLastPromotion:
Có outliers rõ (15 năm không được thăng chức).



YearsWithCurrManager:
Có vài giá trị lớn hơn phần còn lại (> 15 năm).

II. XỬ LÝ DỮ LIỆU

- Đánh giá chung: Việc có sự xuất hiện dữ liệu outliers trong bộ dữ liệu thuộc lĩnh vực này nguyên nhân là do đặc thù của các công ty, giả sử ở trường MonthlyIncome, có rất nhiều giá trị outliers vì những người này nắm giữ những vị trí quan trọng trong công ty, và mức lương của họ sẽ có sự chênh lệch đối với nhân viên trong công ty, nên mặt bằng chung các giá trị này **thường trội hơn giá trị trung bình**.
- Đề xuất: Ta sẽ không loại bỏ các giá trị outliers, và ta sẽ thực hiện đánh giá ở bước Data Analysis để đưa ra kết luận

II. XỬ LÝ DỮ LIỆU

- Tiếp tục xử lý các biến định tính
 - Đối với các biến định tính, ta sẽ thực hiện chuyển đổi các giá trị phân loại và thực hiện mã hoá thành các dạng số nguyên

```
[ ] from sklearn.preprocessing import LabelEncoder  
object_columns = df.select_dtypes(include='object').columns  
print(object_columns)  
object_df = df[object_columns]  
  
→ Index(['BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole',  
         'MaritalStatus', 'Over18', 'OverTime', 'Attrition'],  
        dtype='object')
```

- Kết quả sau khi thực hiện Encoder

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender
0	41	1	2	1102	2	1	2	1	1	1	1	2
1	49	0	1	279	1	8	1	1	1	1	2	3
2	37	1	2	1373	1	2	2	4	1	4	4	4
3	33	0	1	1392	1	3	4	1	1	5	4	4
4	27	0	2	591	1	2	1	3	1	7	1	1

Nhận xét: Việc chuyển đổi dữ liệu định tính sang định lượng nhằm mục đích đưa dữ liệu vào model vào trong tính toán dễ dàng hơn.

II. XỬ LÝ DỮ LIỆU

2. Kiểm tra imbalance data

- Xem thử có giá trị bị thiếu hay không (Đã thực hiện ở phần trước), xác định bộ dữ liệu không có giá trị bị thiếu

```
Kiểm tra

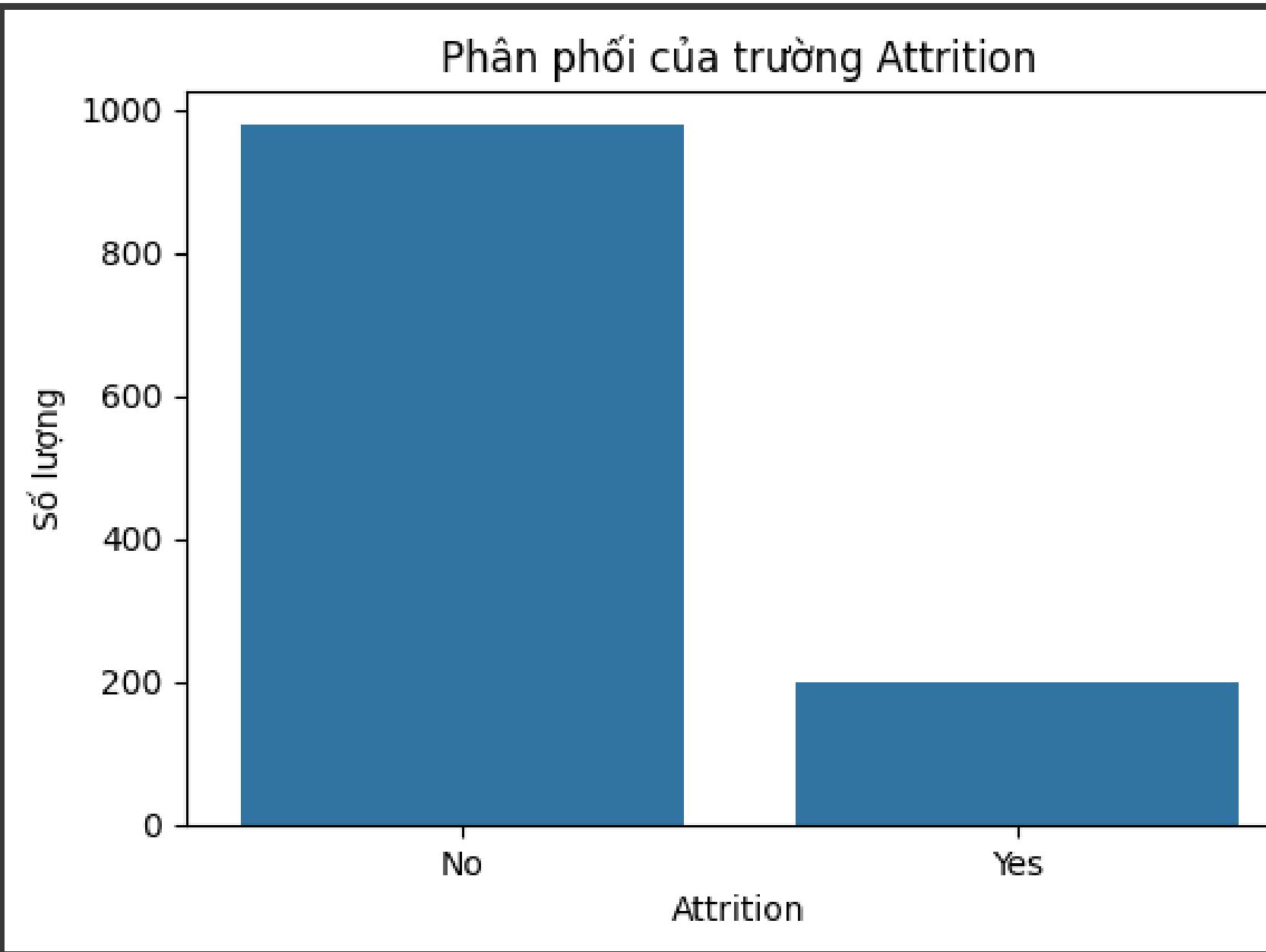
[ ] # Đếm số lượng mẫu theo lớp
attrition_counts = df['Attrition'].value_counts()
print(attrition_counts)

# Tính tỷ lệ phần trăm
attrition_percent = df['Attrition'].value_counts(normalize=True) * 100
print(attrition_percent)

→ Attrition
No      978
Yes     198
Name: count, dtype: int64
Attrition
No      83.163265
Yes     16.836735
Name: proportion, dtype: float64
```

- Kết luận: Trong data này, lớp "No" chiếm 83.2%, lớp "Yes" chiếm 16.8% → Dữ liệu mất cân bằng (lớp "Yes" là lớp thiểu số).

II. XỬ LÝ DỮ LIỆU



Cột "No" cao gấp nhiều lần cột "Yes"

Đánh giá mức độ mất cân bằng

Tỷ lệ mất cân bằng (Imbalance Ratio): Tính tỷ lệ giữa lớp đa số và lớp thiểu số: Ta có 978 mẫu "No" và 198 mẫu "Yes", tỷ lệ là $978 / 198 \approx 4.9:1$.

Ngưỡng đánh giá:

- Tỷ lệ $< 2:1$: Dữ liệu được coi là cân bằng hoặc ít mất cân bằng.
- Tỷ lệ $2:1$ đến $4:1$: Mất cân bằng nhẹ.
- Tỷ lệ $> 4:1$: Mất cân bằng đáng kể.
- Tỷ lệ $> 10:1$: Mất cân bằng nghiêm trọng.

Kết luận: dữ liệu mất cân bằng đáng kể

II. XỬ LÝ DỮ LIỆU

Hướng xử lý : Sử dụng SMOTE để xử lí mất công bằng dữ liệu

```
: categorical_cols = df_train_le.select_dtypes(include=['object', 'category']).columns.tolist()
if 'Attrition' in categorical_cols:
    categorical_cols.remove('Attrition')

df_encoded = pd.get_dummies(df_train_le, columns=categorical_cols, drop_first=True)

X = df_encoded.drop('Attrition', axis=1)
y = df_encoded['Attrition']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

smote = SMOTE(random_state=42)
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)

print(pd.Series(y_train_resampled).value_counts())
df_train_le = pd.concat([X_train_resampled, y_train_resampled], axis=1)
df_train_le
```

Kết quả sau xử lý :

```
Attrition
0      787
1      787
Name: count, dtype: int64
```

II. XỬ LÝ DỮ LIỆU

6. Vài hướng xử lý dữ liệu mất cân bằng (Đề xuất)

6.1 Tái Lấy Mẫu (Resampling)

a. Oversampling (Tăng mẫu lớp thiểu số)

-Mô tả: Tăng số lượng mẫu của lớp thiểu số bằng cách sao chép ngẫu nhiên hoặc tạo mẫu tổng hợp

b. Undersampling (Giảm mẫu lớp đa số)

-Mô tả: Giảm số lượng mẫu của lớp đa số để cân bằng với lớp thiểu số.

.6.2 Sử Dụng Trọng Số Lớp (Class Weighting)

-Mô tả: Gán trọng số cao hơn cho lớp thiểu số trong hàm mất mát của thuật toán để mô hình chú ý hơn đến lớp này.

6.3 Thu Thập Thêm Dữ Liệu

-Mô tả: Thu thập thêm mẫu cho lớp thiểu số nếu có thể.

6.4 Sử Dụng Thuật Toán Phù Hợp

-Mô tả: Một số thuật toán ít nhạy cảm với dữ liệu mất cân bằng hơn

6.5 Sử Dụng Độ Đo Đánh Giá Phù Hợp

6.6 Kết Hợp Nhiều Kỹ Thuật

II. XỬ LÝ DỮ LIỆU

Một số tools sử dụng xử lý

1. Downsampling (Undersampling)

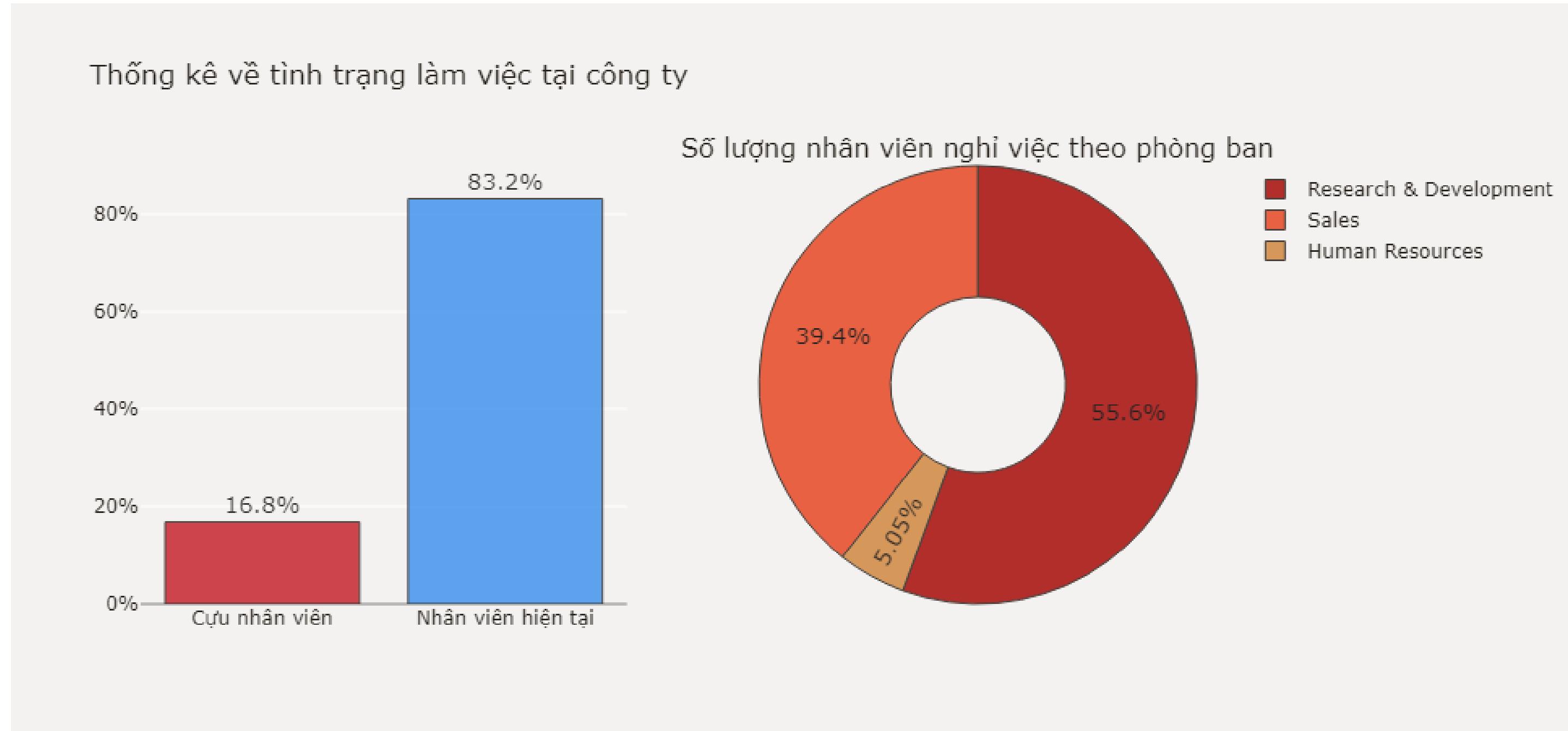
- Mô tả: Giảm số lượng mẫu của lớp đa số (No) để cân bằng với lớp thiểu số (Yes) bằng cách loại bỏ ngẫu nhiên một phần dữ liệu lớp đa số.
- Ưu điểm: Giảm thời gian huấn luyện, đơn giản hóa dữ liệu.
- Nhược điểm: Có thể làm mất thông tin quan trọng nếu loại bỏ quá nhiều mẫu.

1. k-fold Cross-Validation

- Mô tả: Chia dữ liệu thành k tập con (folds), sử dụng k-1 fold để huấn luyện và 1 fold để kiểm tra, lặp lại k lần. Kỹ thuật này không trực tiếp xử lý mất cân bằng nhưng giúp đánh giá hiệu suất mô hình trên dữ liệu mất cân bằng một cách công bằng, đặc biệt khi kết hợp với các kỹ thuật resampling.
- Ưu điểm: Giảm thiểu rủi ro overfitting, cung cấp đánh giá ổn định hơn.
- Nhược điểm: Tăng thời gian tính toán, cần kết hợp với resampling để xử lý mất cân bằng.

PHÂN TÍCH DỮ LIỆU

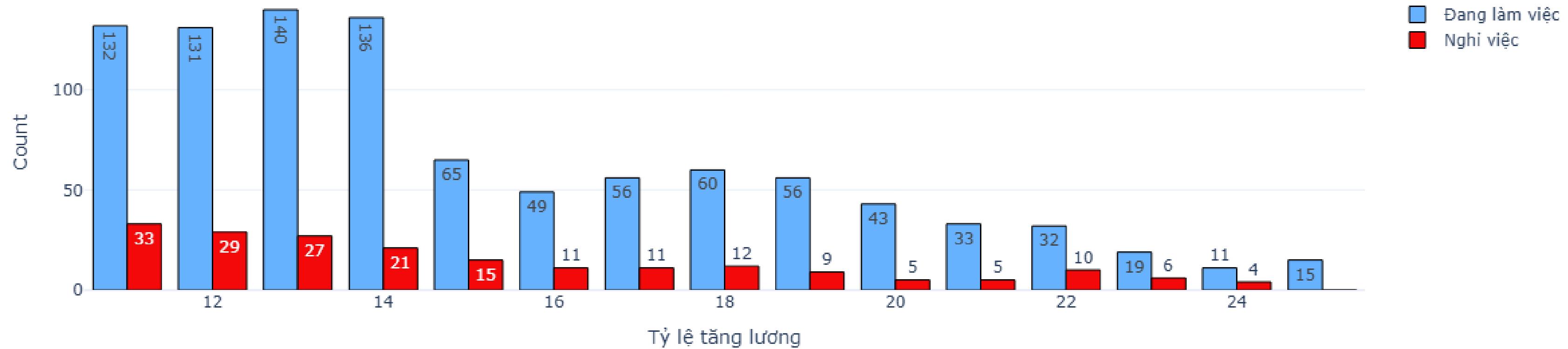
1. PHÂN TÍCH TỔNG QUAN VỀ CÔNG TY



- Dựa vào biểu đồ ta thấy trong bộ dữ liệu bao gồm **16.1% nhân viên đã nghỉ việc** và có **83.9% nhân viên vẫn còn đang làm**
- Trong đó
 - Ở **phòng ban R&D**, số lượng nhân viên nghỉ việc chiếm tỷ lệ cao nhất, khoảng 51.3%
 - Ngay sau đó là **phòng ban Sales**, chiếm khoảng 38.8% tỷ lệ nhân viên nghỉ việc
 - **Phòng ban HR** chiếm tỷ lệ ít nhất, với 2.06%

1. PHÂN TÍCH TỔNG QUAN VỀ CÔNG TY

Biểu đồ thể hiện phần trăm lương tăng theo nhân viên trong công ty

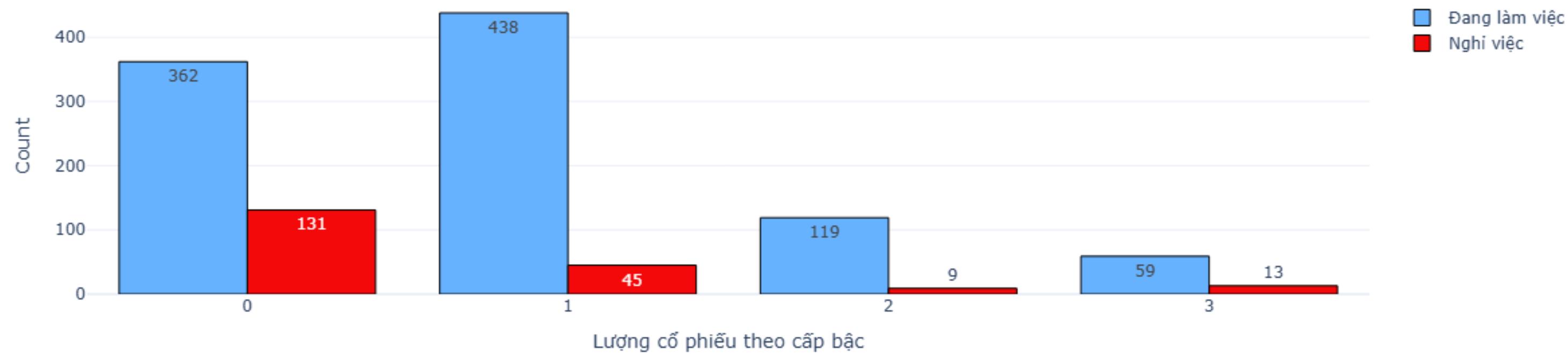


- Nhận xét:

- Ta thấy, nhóm có phần trăm tăng lương (11 - 14%) chiếm số lượng nhân viên nhiều nhất, nhưng đồng thời đây cũng là nhóm mà có số lượng **nhân viên nghỉ việc** nhiều nhất
- Số lượng **nhân viên nghỉ việc** có xu hướng **giảm dần** khi phần trăm lương tăng dần

1. PHÂN TÍCH TỔNG QUAN VỀ CÔNG TY

Tỷ lệ nhân viên nghỉ việc theo Sứ nắm giữ cổ phiếu công ty



- Nhận xét:

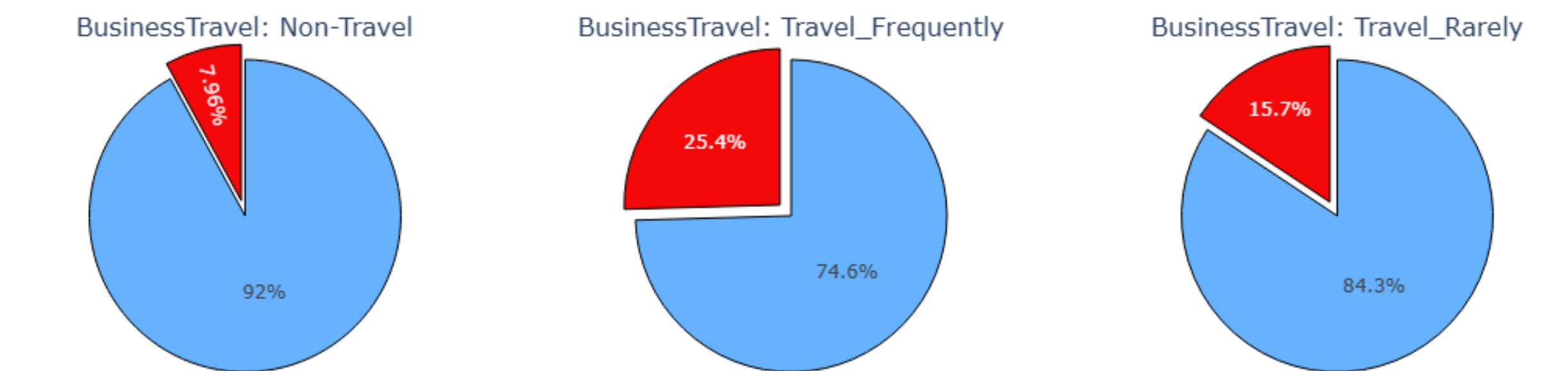
- Xu hướng của biểu đồ này cho ta có cái nhìn tương tự như Tỷ lệ tăng lương
- Ta có thể thấy ở **nhóm 0 và 1** chiếm số lượng lớn nhân viên trong công ty, đây là những nhóm nhân viên thường sẽ **không** có sở hữu cổ phiếu hoặc có nắm giữ nhưng số lượng không đáng kể, vì vậy xu hướng khi họ nghỉ việc thường sẽ không quan tâm tới vấn đề này nên ta có thể thấy được ở 2 nhóm này nhân viên có số lượng nghỉ việc nhiều trong tổng cả 4 nhóm
- Mặc khác ở **nhóm 2 và 3** cũng có nhưng không chiếm số lượng lớn trong tổng cả 4 nhóm (số lượng nhóm 2 + 3 ít hơn rất nhiều so với nhóm 1)

1. PHÂN TÍCH TỔNG QUAN VỀ CÔNG TY

Số lượng nhân viên nghỉ việc theo BusinessTravel



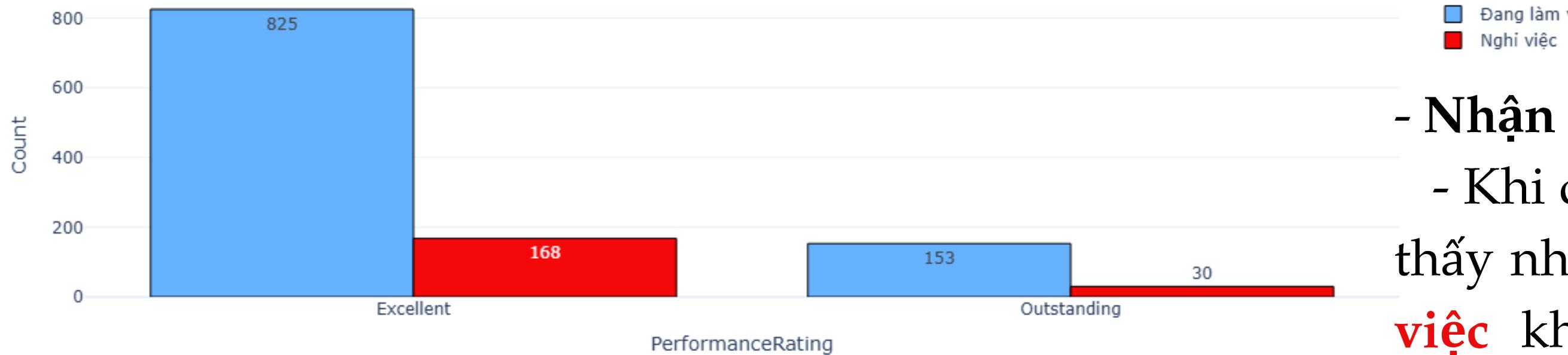
Tỷ lệ nhân viên nghỉ việc theo BusinessTravel



- **Nhận xét:**
 - Nhìn vào biểu đồ **Số lượng nhân viên nghỉ việc** ta thấy những nhân viên có tần suất công tác không thường xuyên chiếm số lượng rất nhiều, tuy nhiên xét về tỷ lệ thì những nhân viên có tần suất công tác thường xuyên lại có tỷ lệ rất cao (**22.4 %**)

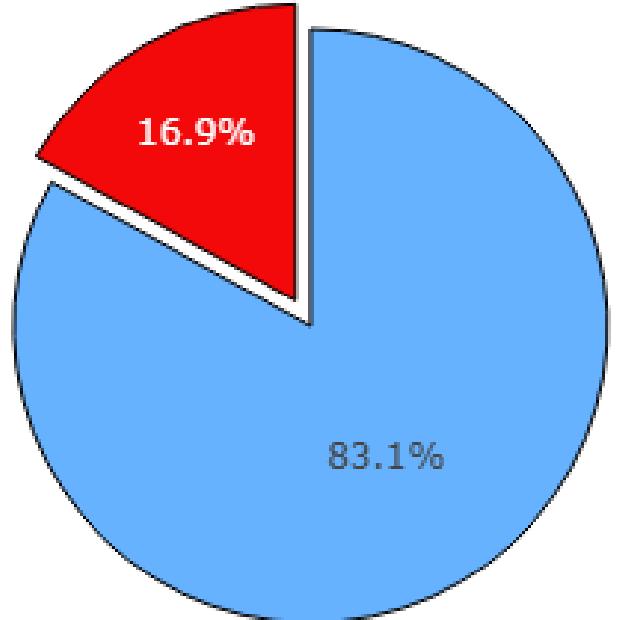
1. PHÂN TÍCH TỔNG QUAN VỀ CÔNG TY

Số lượng nhân viên trong công ty theo Hiệu suất làm việc

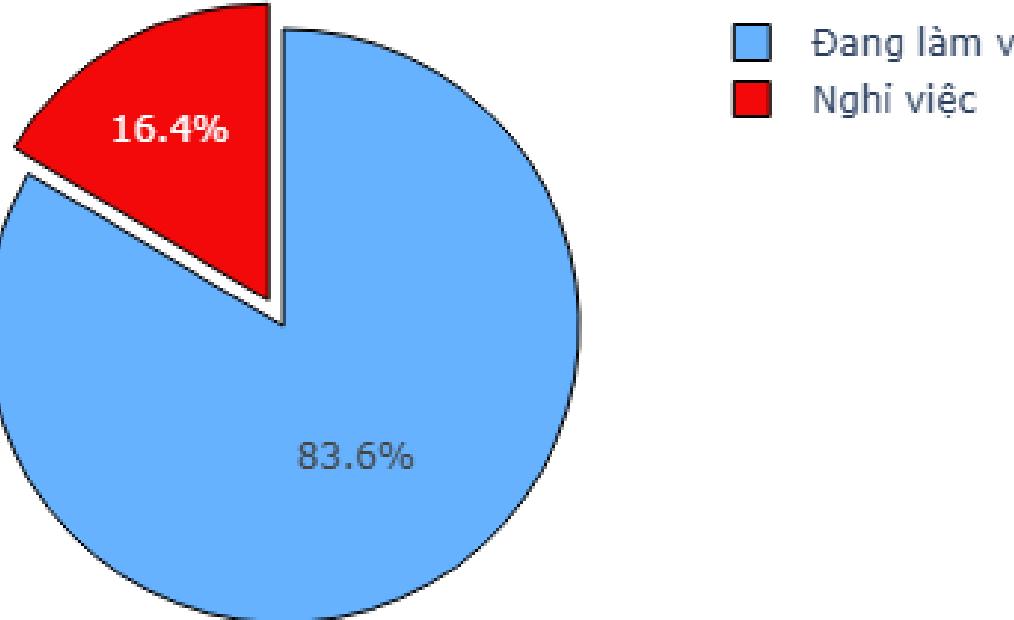


Tỷ lệ nhân viên nghỉ việc theo Hiệu suất làm việc

PerformanceRating: Excellent



PerformanceRating: Outstanding



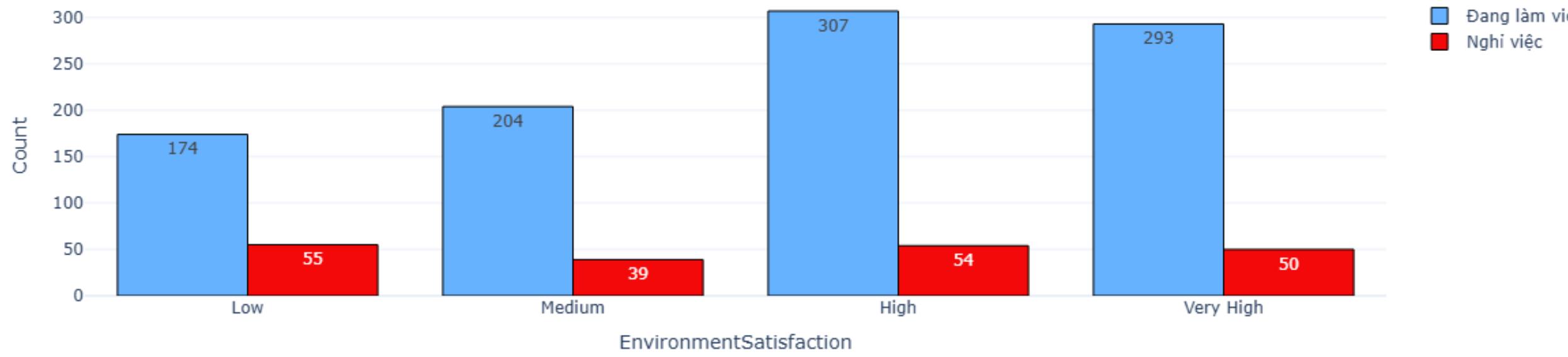
- Nhận xét:

- Khi quan sát ở biểu đồ *số lượng*, ta thấy những **số lượng nhân viên nghỉ việc** khi có hiệu xuất làm việc tốt chiếm số lượng lớn hơn (**168 nhân viên**) khi so sánh với nhân viên nghỉ việc như có hiệu suất nổi bật hơn (30 nhân viên)

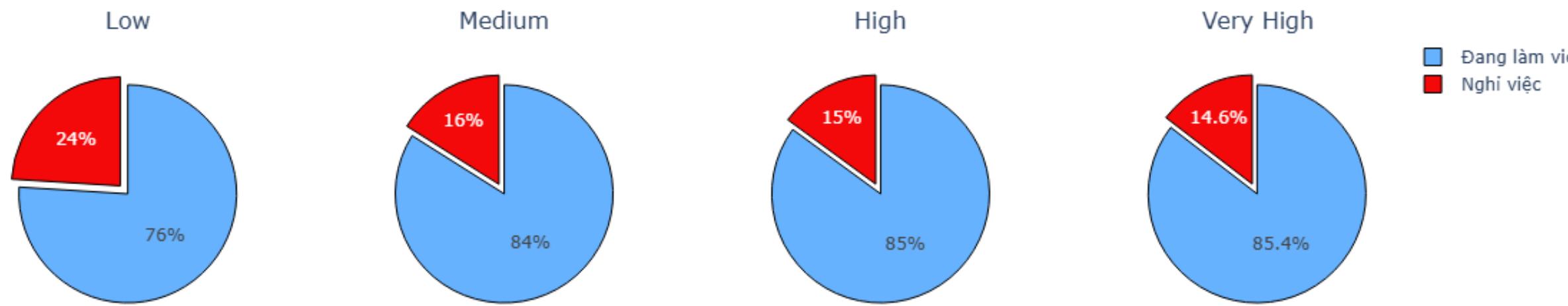
- Tuy nhiên khi xem xét về *tỉ lệ*, ta nhận thấy ở 2 nhóm nhân viên có hiệu suất làm việc này có **tỉ lệ nhân viên nghỉ việc tương đương nhau**

1. PHÂN TÍCH TỔNG QUAN VỀ CÔNG TY

Số lượng nhân viên khi xem xét về Sự hài lòng với môi trường làm việc tại công ty



Tỷ lệ nhân viên nghỉ việc theo Sự hài lòng với môi trường làm việc tại công ty



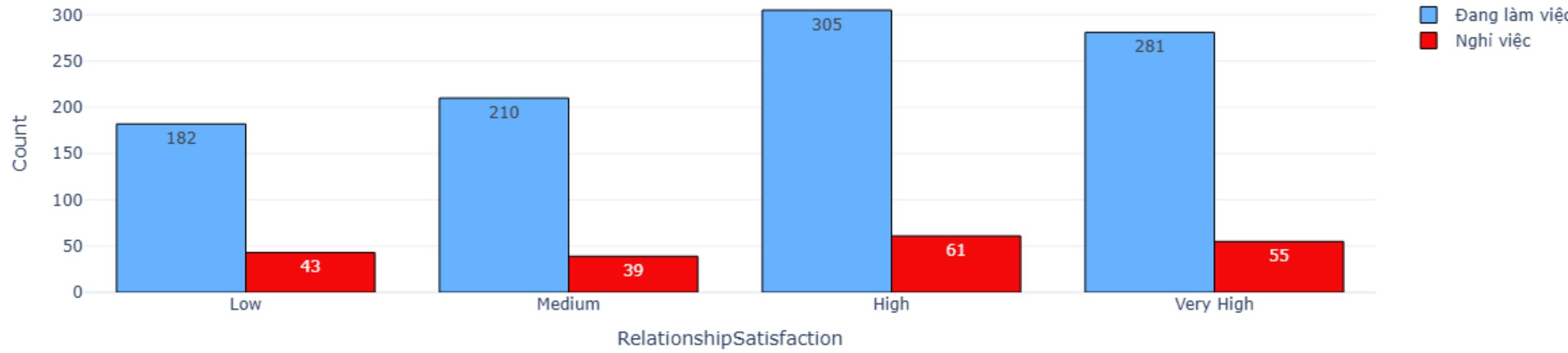
- Nhận xét:

- Số lượng **nhân viên nghỉ việc** ở mỗi mức độ hài lòng đều gần như tương đương nhau

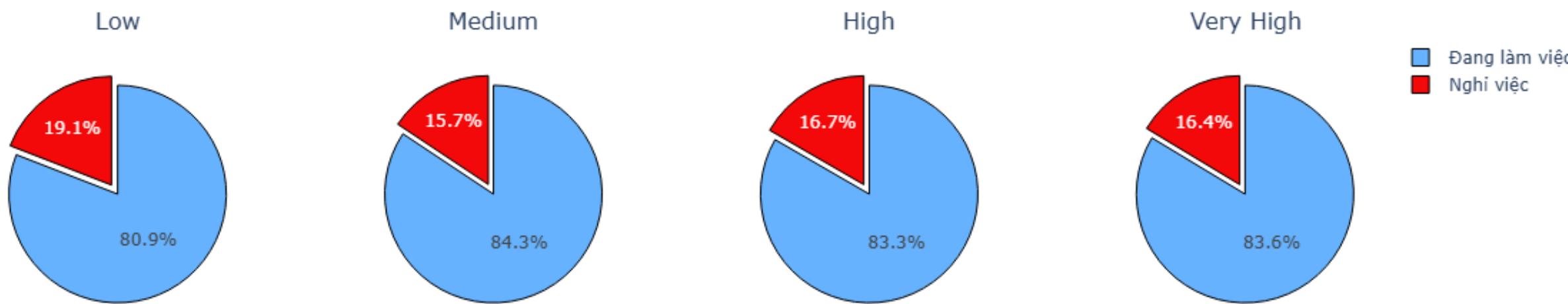
- Tuy nhiên khi xem xét về tỷ lệ ta thấy được những người có sự hài lòng về môi trường công ty càng **thấp** thì họ lại càng có **xu hướng nghỉ việc** và tỷ lệ **giảm dần** khi nhân viên có mức độ hài lòng với công ty **cao**

1. PHÂN TÍCH TỔNG QUAN VỀ CÔNG TY

Số lượng nhân viên theo Sự hài lòng về các mối quan hệ trong công ty



Tỷ lệ nhân viên nghỉ việc theo Sự hài lòng về các mối quan hệ trong công ty

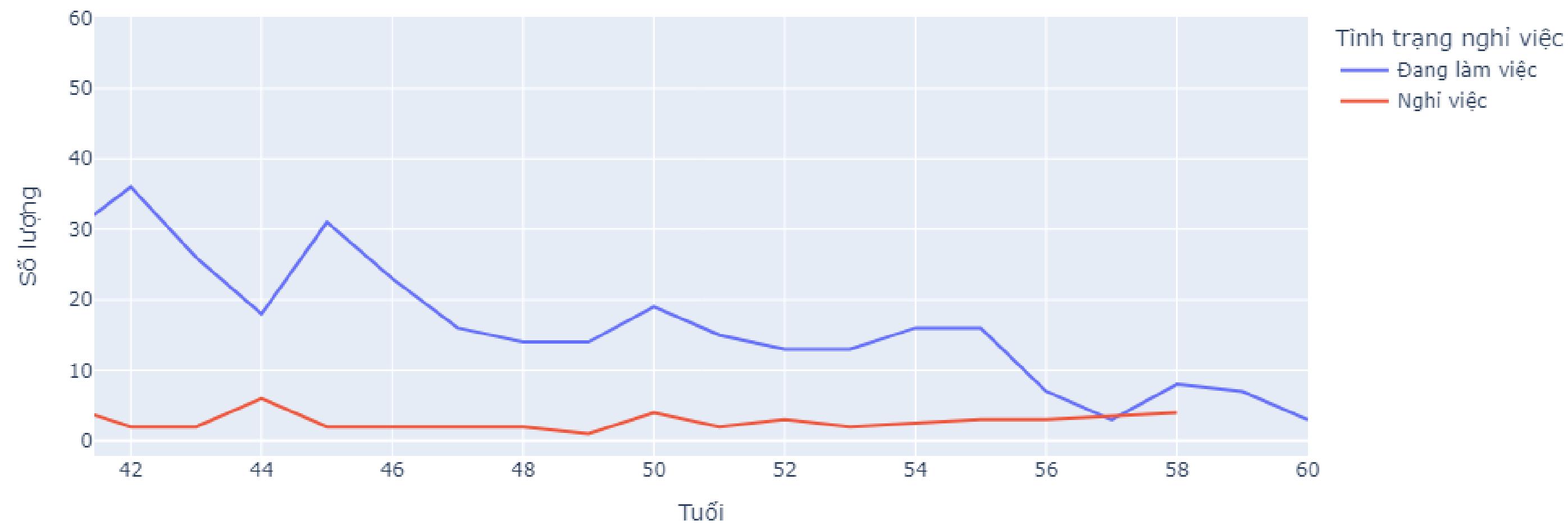


- Nhận xét:

- Ta có thể thấy biểu đồ này thể hiện **khá tương đồng** với biểu đồ về sự hài lòng về môi trường làm việc
- Khi các **mối quan hệ trong công việc** có **xu hướng tốt hơn** thì tỷ lệ nhân viên nghỉ việc lại thấp hơn

2. THÔNG TIN CỦA NHÂN VIÊN

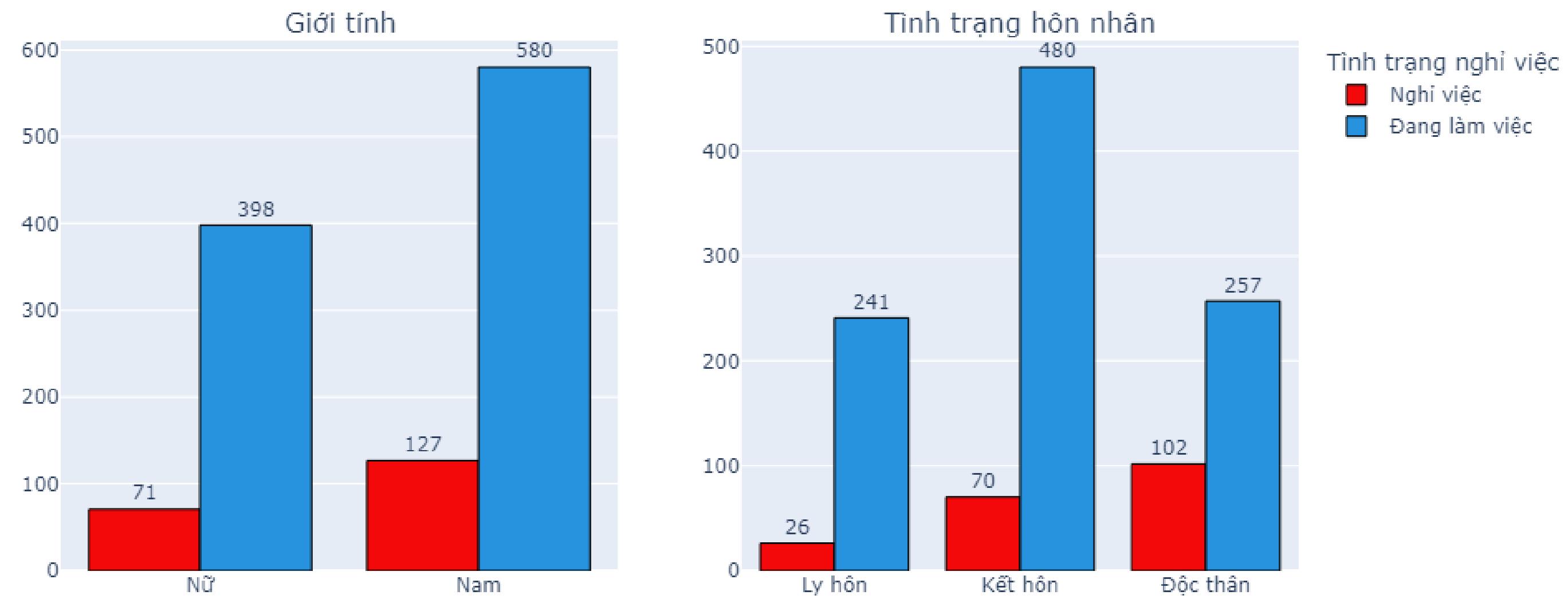
Số lượng nhân viên trong công ty theo độ tuổi



- Ta có thể thấy ở độ tuổi 26 đến 38 số lượng **nhân viên nghỉ việc tăng cao**. Điều này là vì đây là độ tuổi bắt đầu tìm kiếm việc làm nên họ thường có xu hướng muốn khám phá và tìm kiếm nhiều cơ hội mới, và họ yêu cầu trải nghiệm trong công việc nhiều hơn.
- Còn từ **độ tuổi 40 trở về sau** số lượng **nhân viên nghỉ việc càng giảm**. Bởi vì đây là độ tuổi mà họ cần có sự ổn định trong công việc để lo cho cuộc sống và đây cũng là độ tuổi có rủi ro tìm việc cao nhất

2. THÔNG TIN CỦA NHÂN VIÊN

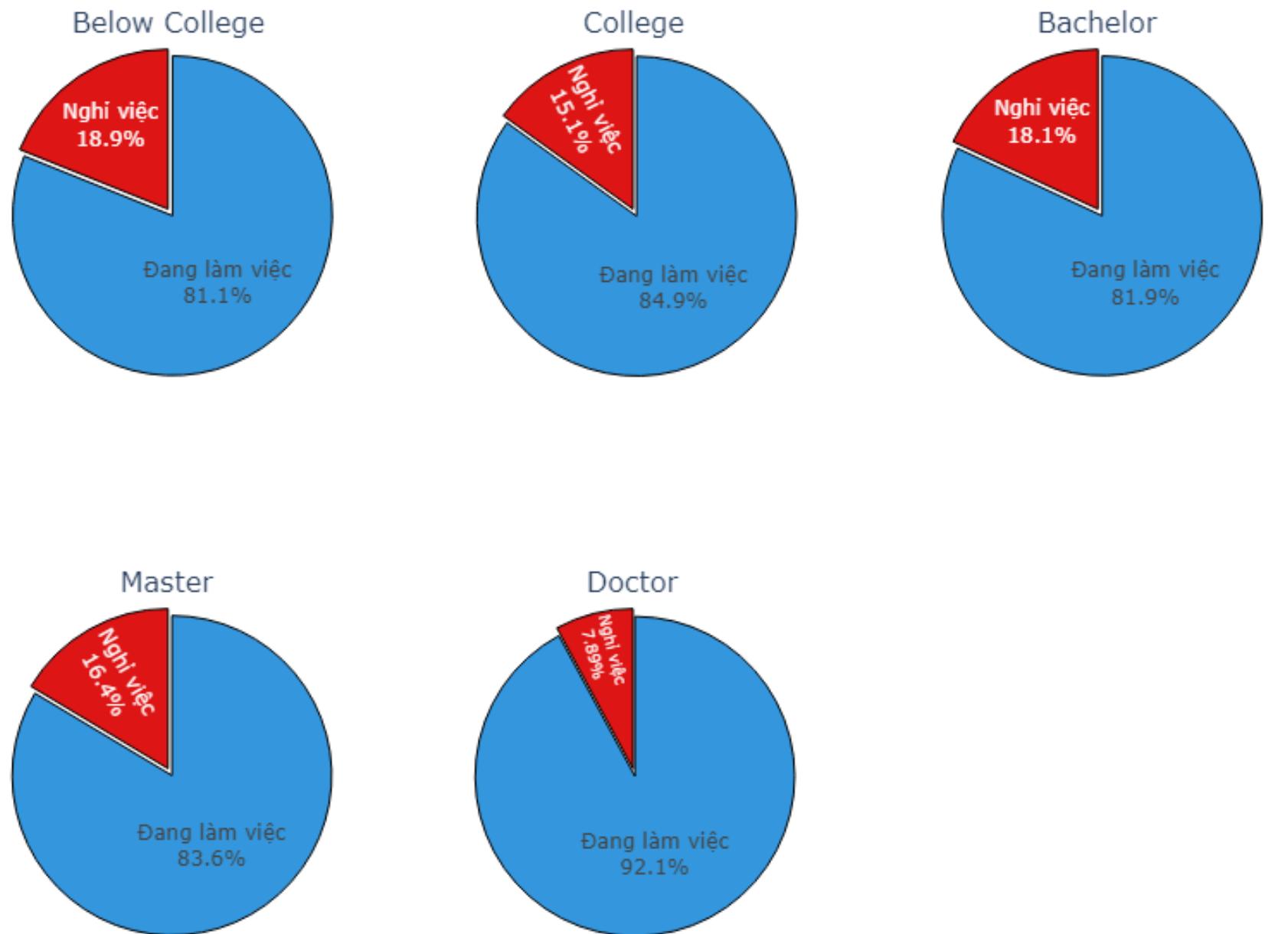
Số lượng nhân viên nghỉ việc theo giới tính và tình trạng hôn nhân



Ta có thể thấy rằng **số lượng nhân viên nam nghỉ việc** cao hơn nhân viên
Số lượng **nhân viên độc thân** nghỉ việc cao nhất (102 người) tiếp đến là **nhân
viên đã kết hôn** (70 người) cuối cùng là **nhân viên đã ly hôn** (26 người)

2. THÔNG TIN CỦA NHÂN VIÊN

Tỷ lệ nghỉ việc theo trình độ học vấn

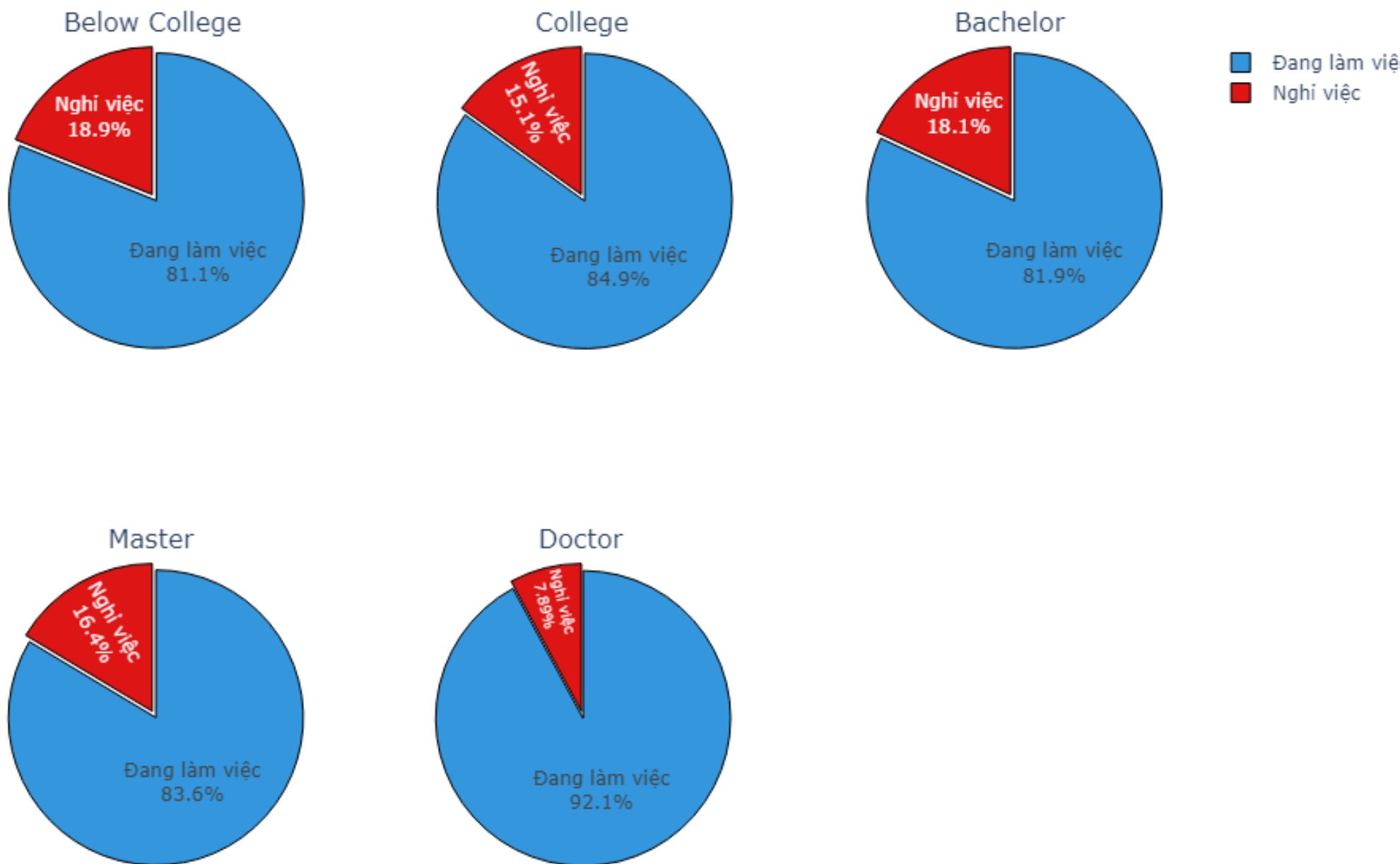


Đang làm việc
Nghỉ việc

- Trình độ "**Bachelor**" có số lượng nhân viên đông nhất cho thấy đây là nhóm phổ biến nhất trong công ty. Tiếp theo là "**Master**", "**College**", "**Below College**" và "**Doctor**" là ít nhất.
- Số lượng **nhân viên nghỉ việc** tỷ lệ thuận với số lượng nhân viên ở mỗi trình độ
- **Below College**: 81.1% đang làm việc, 18.9% nghỉ việc.
- **College**: 84.9% đang làm việc, 15.1% nghỉ việc.
- **Bachelor**: 81.9% đang làm việc, 18.1% nghỉ việc.
- **Master**: 83.6% đang làm việc, 16.4% nghỉ việc.
- **Doctor**: 92.1% đang làm việc, 7.89% nghỉ việc.

2. THÔNG TIN CỦA NHÂN VIÊN

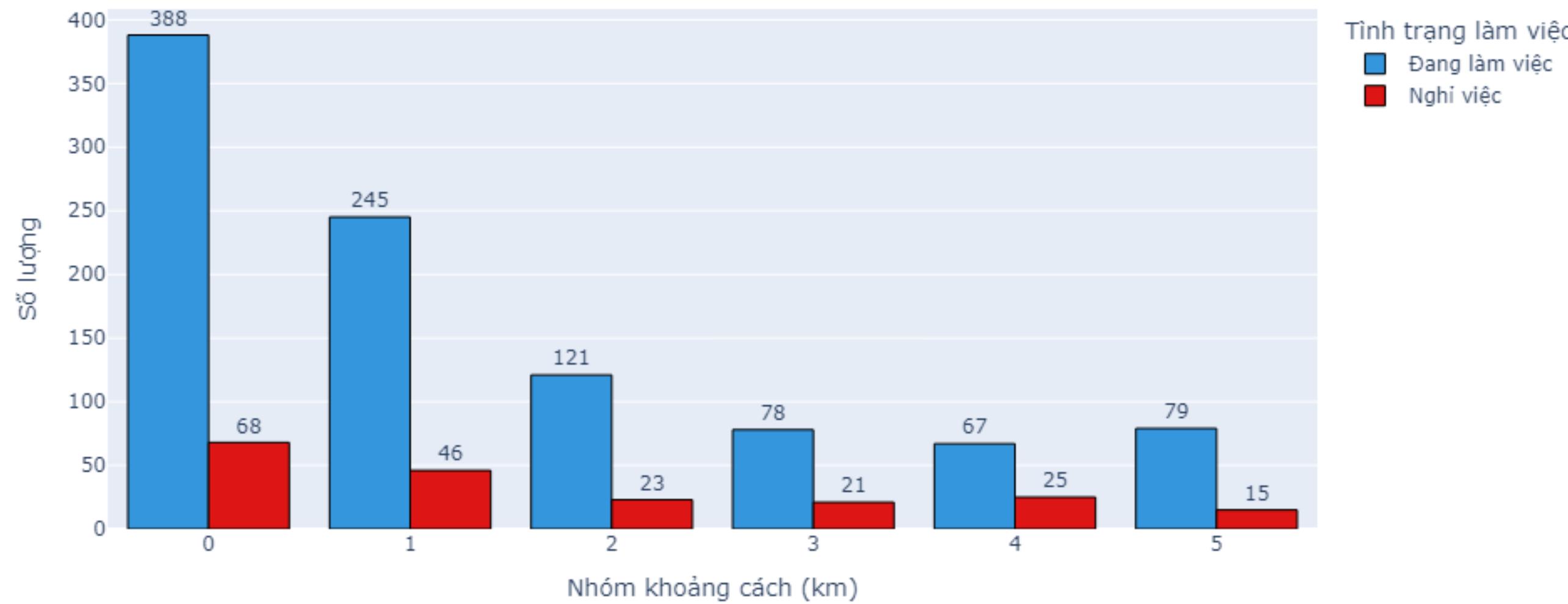
Tỷ lệ nghỉ việc theo trình độ học vấn



=> **Tỷ lệ nghỉ việc** dao động từ **7.89%** **đến 18.9%**, không có sự khác biệt lớn giữa các trình độ học vấn. Nhóm "Below College" có **tỷ lệ nghỉ việc cao nhất** (18.9%), trong khi "Doctor" có **tỷ lệ thấp nhất (7.89%)**.

2. THÔNG TIN CỦA NHÂN VIÊN

Khoảng cách từ nhà đến công ty



Nhóm 0 (0-4 km) có số lượng nhân viên đông nhất (388 người **đang làm việc**, 68 người **nghỉ việc**), cho thấy nhiều người sống gần công ty.

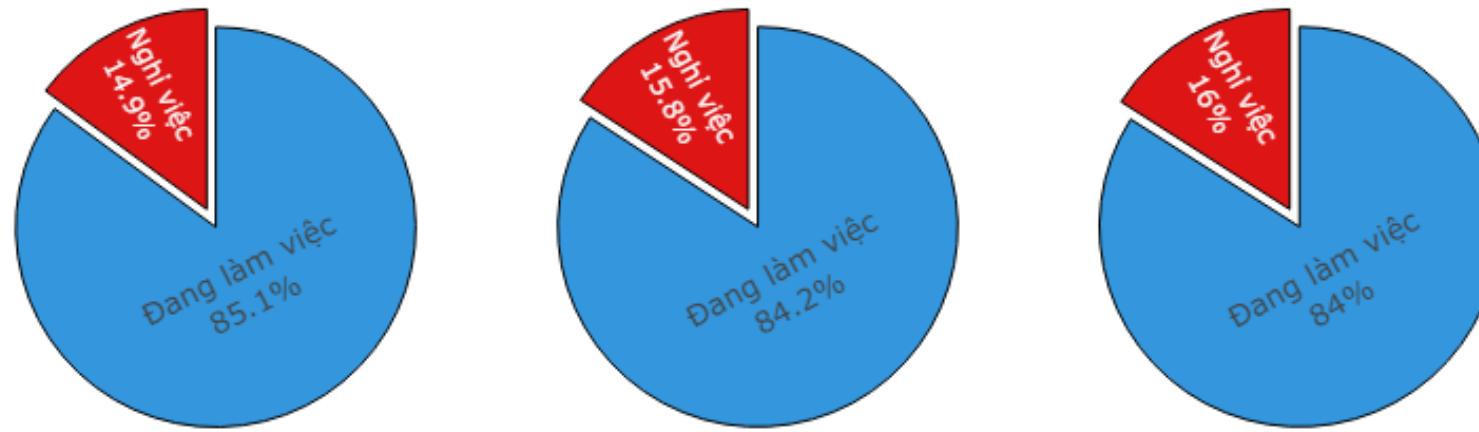
Số lượng nhân viên giảm dần khi khoảng cách tăng: **nhóm 1 (5-9 km)** có 245 người **đang làm việc** và 46 người **nghỉ việc**, **nhóm 2 (10-14 km)** có 121 người **đang làm việc** và 23 người **nghỉ việc**, và tiếp tục giảm ở các nhóm xa hơn.

Tỷ lệ nghỉ việc dường như không tăng rõ rệt theo khoảng cách

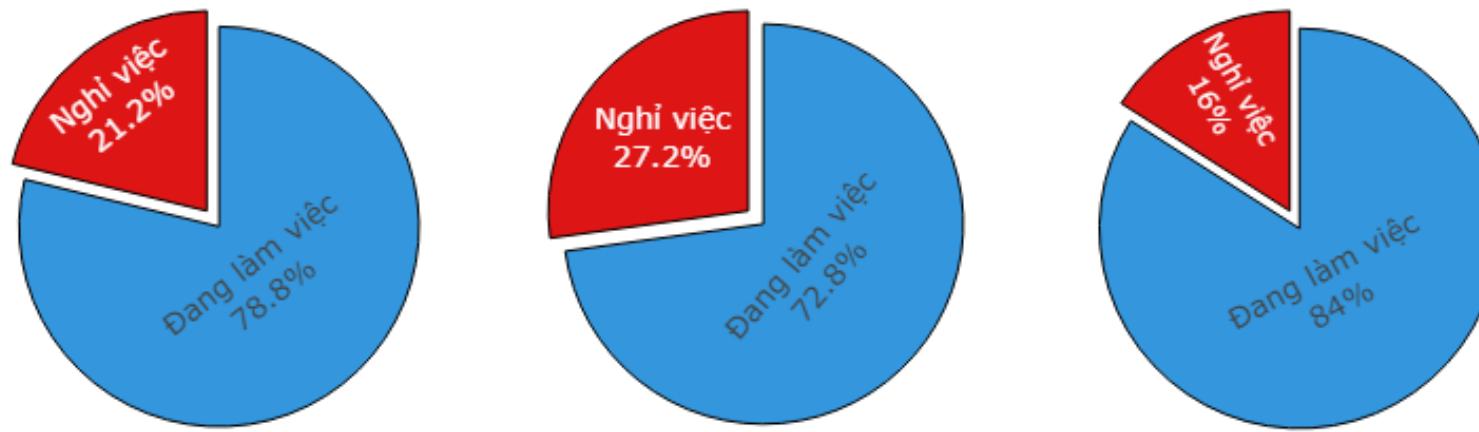
2. THÔNG TIN CỦA NHÂN VIÊN

Tỉ lệ nghỉ việc theo từng nhóm khoảng cách từ nhà đến công ty

Nhóm 0: Khoảng cách 0 - 4 km Nhóm 1: Khoảng cách 5 - 9 km Nhóm 2: Khoảng cách 10 - 14 km



Nhóm 3: Khoảng cách 15 - 19 km Nhóm 4: Khoảng cách 20 - 24 km Nhóm 5: Khoảng cách 25 - 29 km

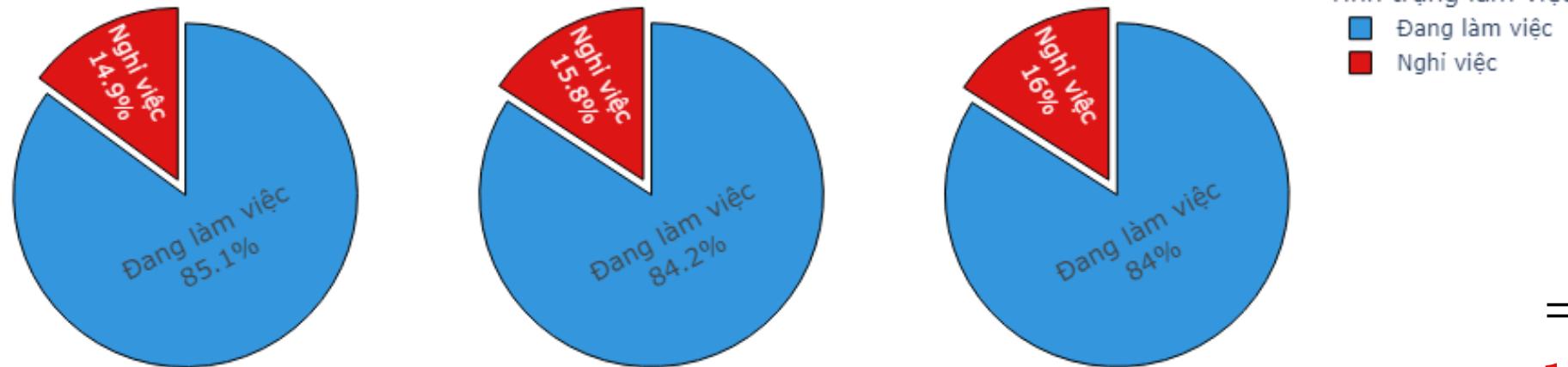


- **Nhóm 0 (0-4 km):** 82.1% đang làm việc, 14.9% nghỉ việc.
- **Nhóm 1 (5-9 km):** 84.2% đang làm việc, 12.8% nghỉ việc.
- **Nhóm 2 (10-14 km):** 84% đang làm việc, 16% nghỉ việc.
- **Nhóm 3 (15-19 km):** 78.8% đang làm việc, 21.2% nghỉ việc.
- **Nhóm 4 (20-24 km):** 72.8% đang làm việc, 27.2% nghỉ việc
- **Nhóm 5 (25-29 km):** 84% đang làm việc, 16% nghỉ việc.

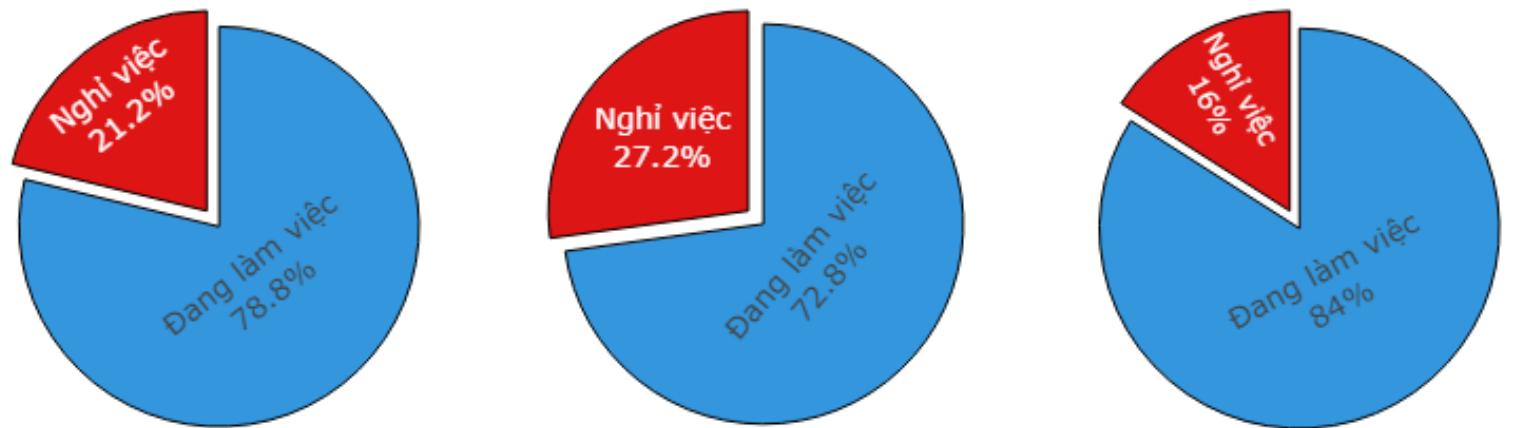
2. THÔNG TIN CỦA NHÂN VIÊN

Tỉ lệ nghỉ việc theo từng nhóm khoảng cách từ nhà đến công ty

Nhóm 0: Khoảng cách 0 - 4 km Nhóm 1: Khoảng cách 5 - 9 km Nhóm 2: Khoảng cách 10 - 14 km



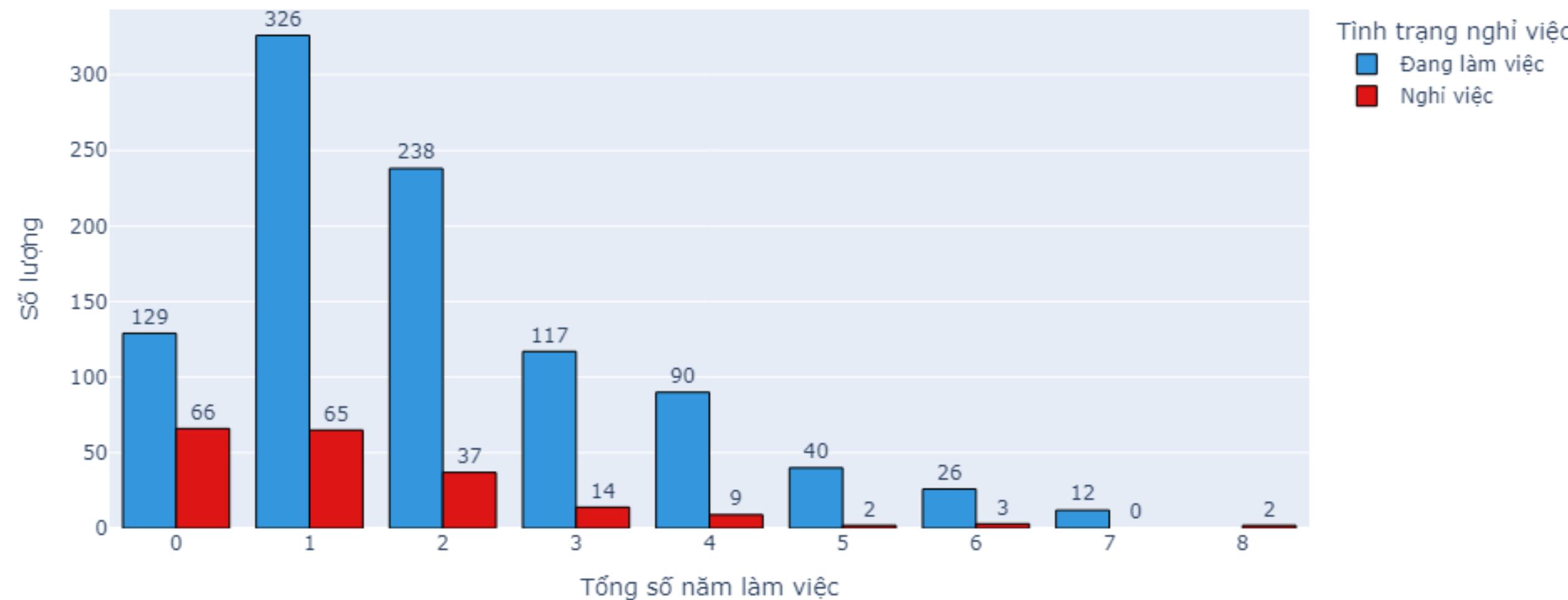
Nhóm 3: Khoảng cách 15 - 19 km Nhóm 4: Khoảng cách 20 - 24 km Nhóm 5: Khoảng cách 25 - 29 km



=> Khoảng cách từ nhà đến công ty có ảnh hưởng đến **tỷ lệ nghỉ việc**: nhân viên sống xa (đặc biệt 20-24 km) có **xu hướng nghỉ việc** cao hơn. Tuy nhiên, mối quan hệ này không hoàn toàn đáng lo, vì nhóm xa nhất (25-29 km) lại có **tỷ lệ nghỉ việc** thấp hơn nhóm 20-24 km. Số lượng nhân viên giảm khi khoảng cách tăng, cho thấy đa số nhân viên có xu hướng sống gần công ty.

2. THÔNG TIN CỦA NHÂN VIÊN

Tổng số năm làm việc của nhân viên

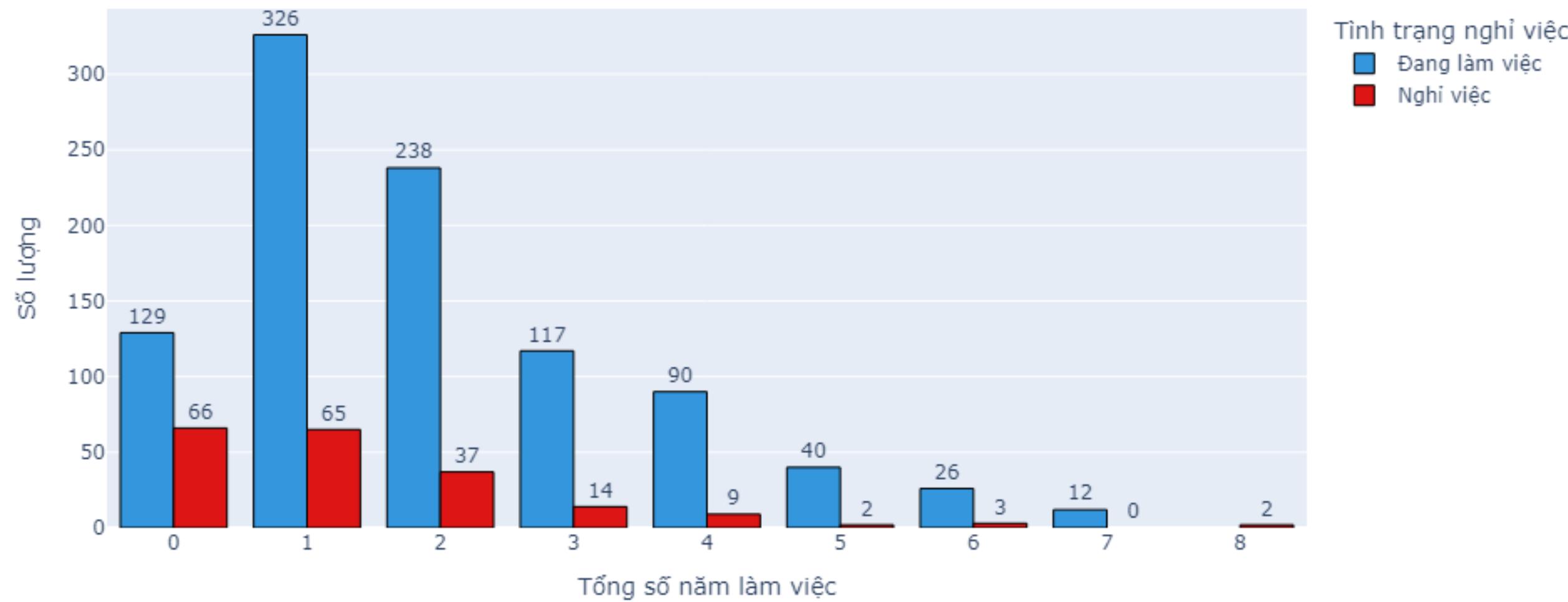


Phân bố số năm làm việc:

- Số lượng nhân viên đông nhất ở khoảng từ nhóm 0 đến nhóm 4, cho thấy phần lớn nhân viên có thâm niên dưới 5 năm.
- Số lượng nhân viên giảm dần khi số năm làm việc tăng.
- Từ nhóm 5 trở lên có rất ít nhân viên, với số lượng dao động từ 2 đến 40 người, cho thấy ít nhân viên ở lại công ty lâu dài.

2. THÔNG TIN CỦA NHÂN VIÊN

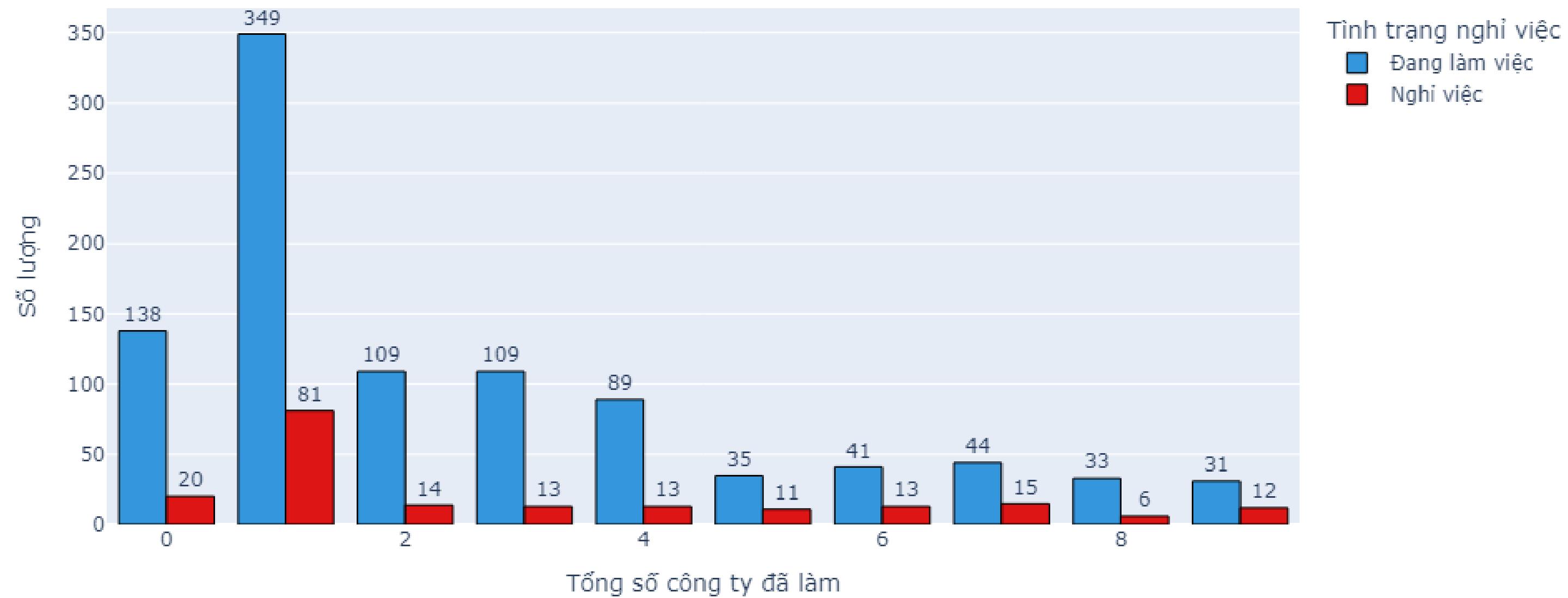
Tổng số năm làm việc của nhân viên



- **Số lượng nghỉ việc** dường như cao hơn ở các nhóm có thâm niên thấp (Điều này là vì đây là năm đầu tiên họ làm việc nên họ có thể nghỉ việc tại công ty nếu cảm thấy bản thân không phù hợp với công ty)
- Ở các nhóm thâm niên cao hơn, **số lượng nghỉ việc** rất thấp (Có thể là do họ đã gắn bó với công ty lâu dài và cảm thấy bản thân phù hợp với công ty)

2. THÔNG TIN CỦA NHÂN VIÊN

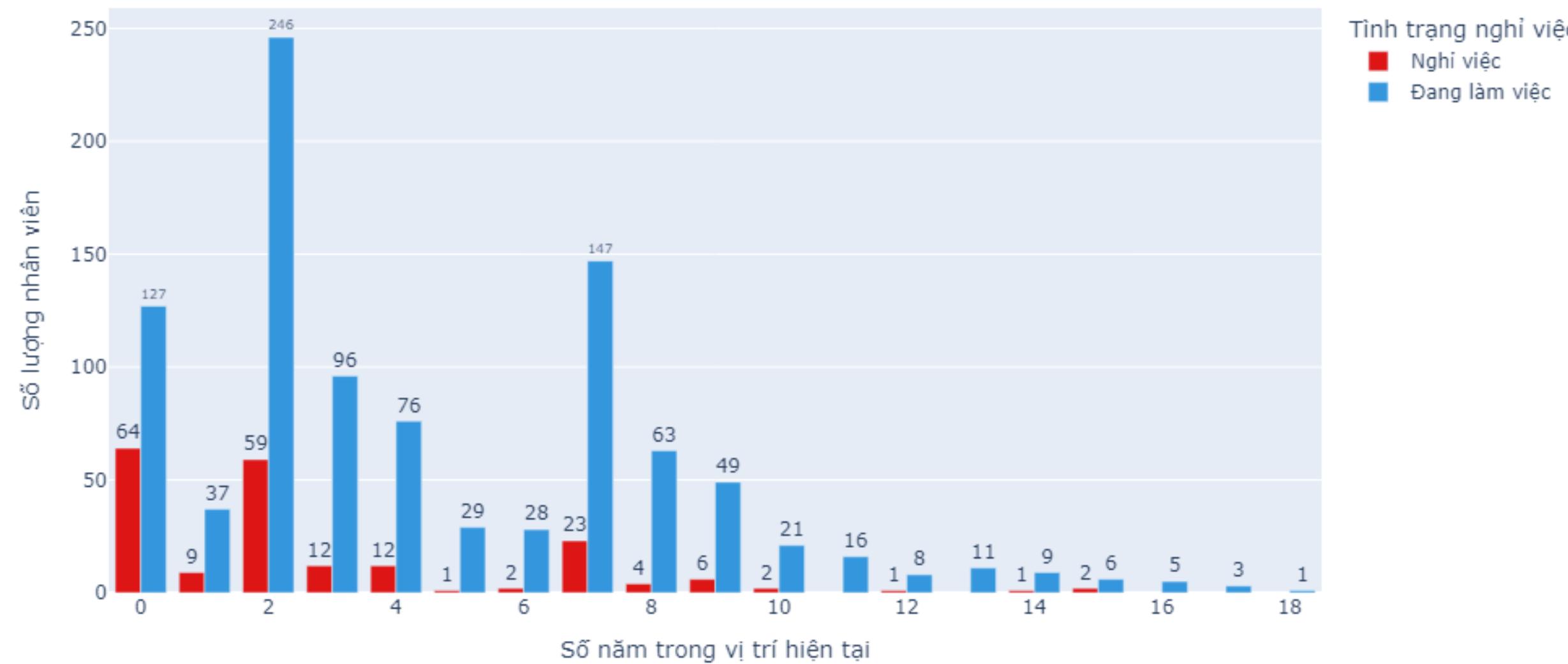
Tổng số công ty đã làm của nhân viên



- Phần lớn nhân viên trong công ty chỉ làm việc cho 1 công ty
- Số lượng **nhân viên nghỉ việc** tập trung nhiều ở nhóm 1 (điều này có thể là vì họ muốn tìm kiếm thêm kinh nghiệm trong công việc ở những năm đầu làm việc)
- Nhân viên làm cho càng nhiều công ty thì **số lượng nghỉ việc** càng thấp (Có thể là vì họ đã có đủ kinh nghiệm làm việc mà họ cần)

2. THÔNG TIN CỦA NHÂN VIÊN

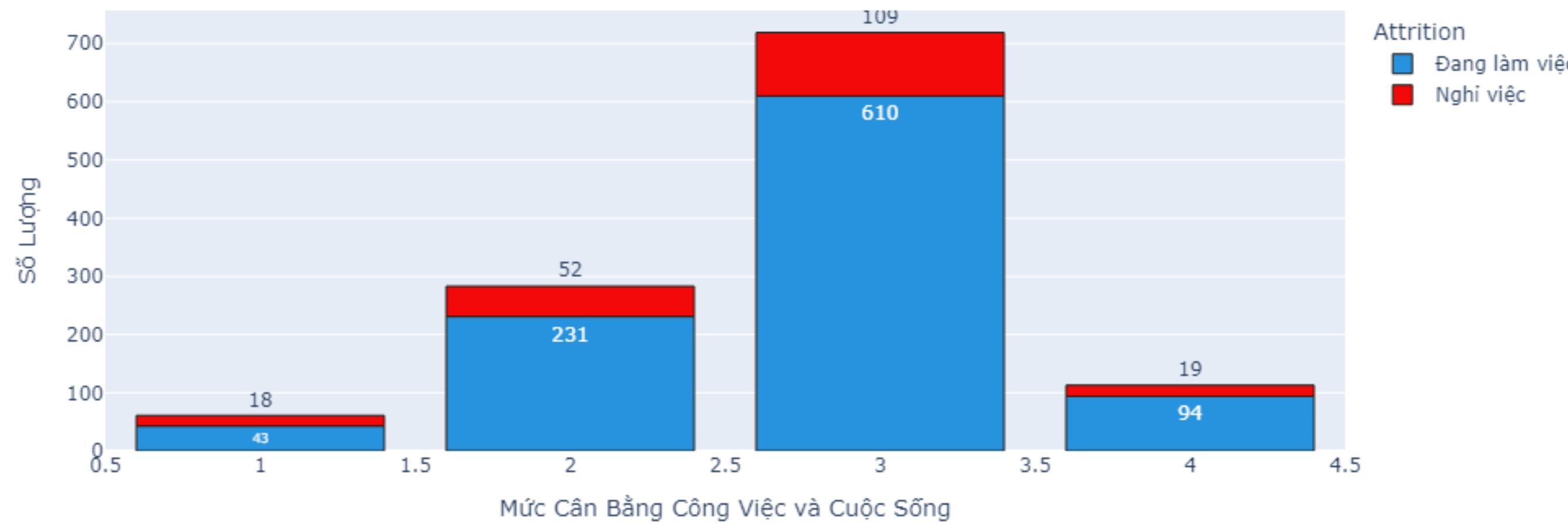
Số năm làm việc trong vị trí hiện tại



- Ta có thể thấy rằng phần lớn **nhân viên nghỉ việc** nằm ở nhóm 0 đến nhóm 2 và tăng nhẹ ở nhóm 7 (Điều này là vì khi họ làm việc ở một vị trí quá lâu mà không được thăng tiến thì họ thường có xu hướng nghỉ việc để tìm cơ hội tốt hơn).
- Ở những nhóm cao thì **số lượng nhân viên nghỉ việc** ít. Điều này chứng tỏ rằng họ hài lòng với vị trí hiện tại của mình

3. THÔNG TIN CÔNG TY VỀ NHÂN VIÊN

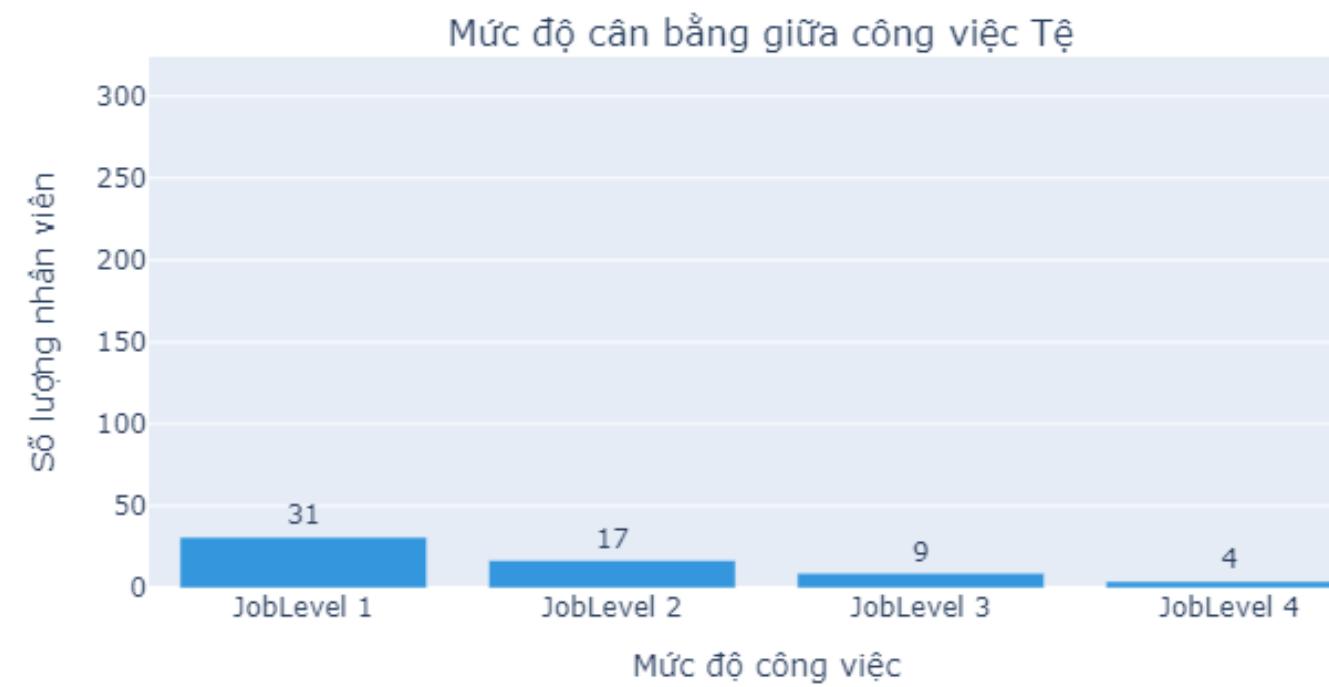
Biểu đồ thể hiện mức độ Cân bằng Công việc và Cuộc sống theo số lượng nhân viên



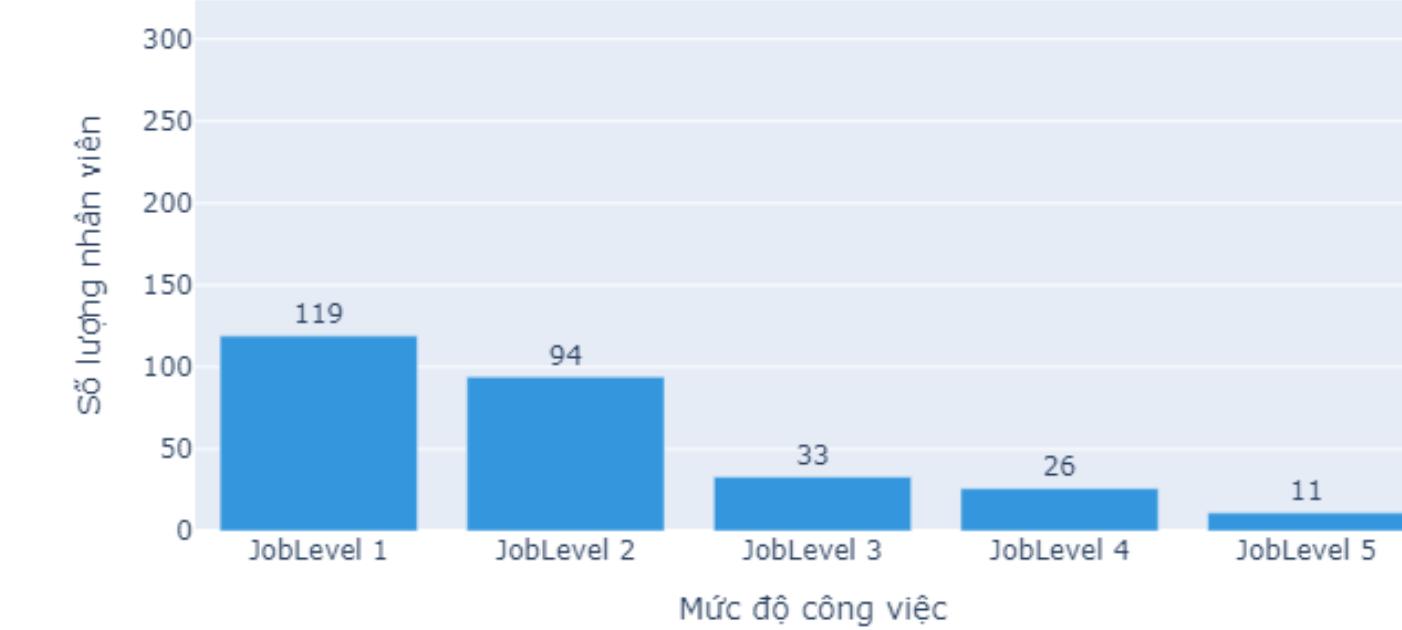
Những người nằm ở mức thấp thường có **xu hướng nghỉ việc**, đổi mới những người ở bậc cao hơn (bậc 3) **xu hướng họ nghỉ việc nhiều hơn**, có thể họ mong muốn tìm thêm cơ hội và đai ngộ xứng đáng hơn, tuy nhiên biểu đồ này cho mức độ cân bằng cuộc sống ở mức 3 trở lên vẫn chiếm số lượng nhiều hơn, thấy nhân viên ở đây có sự hài lòng về công việc hiện tại ở công ty. Để kiểm tra nhận định này, ta sẽ tìm hiểu sự hài lòng của nhân viên đối với công ty như thế nào

3. THÔNG TIN CÔNG TY VỀ NHÂN VIÊN

Phân bố mức độ công việc theo mức độ cân bằng giữa công việc



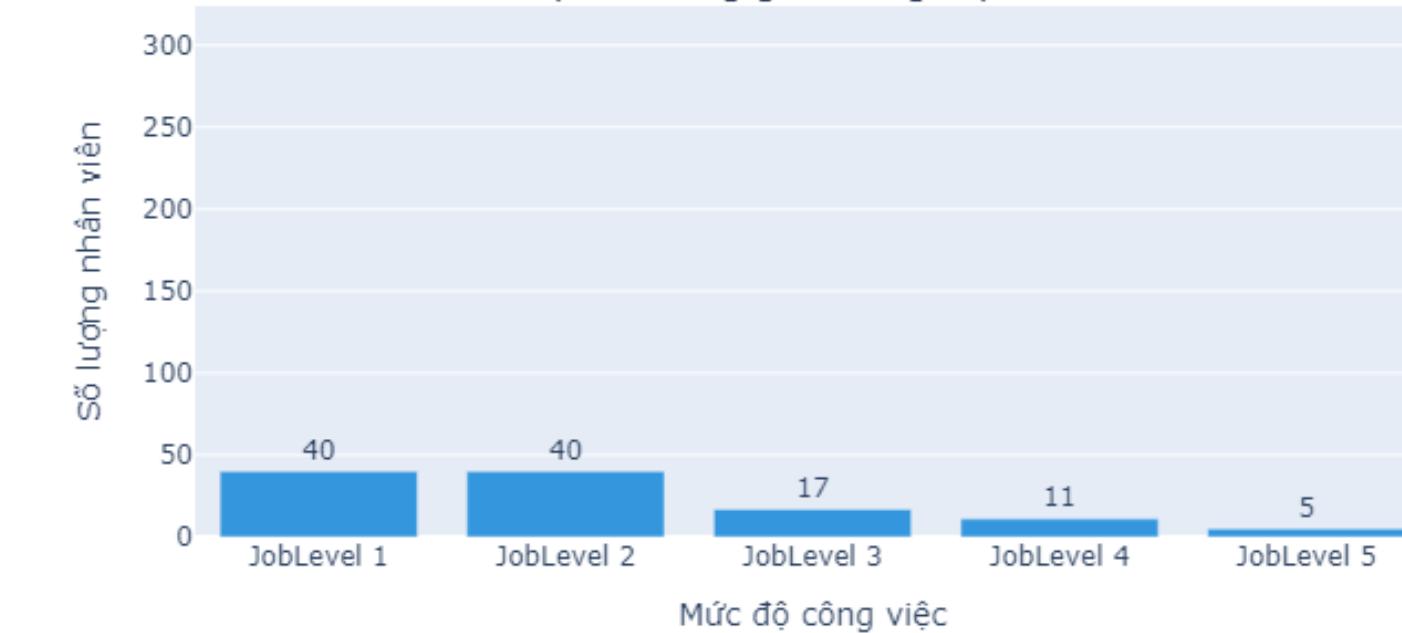
Mức độ cân bằng giữa công việc Khá



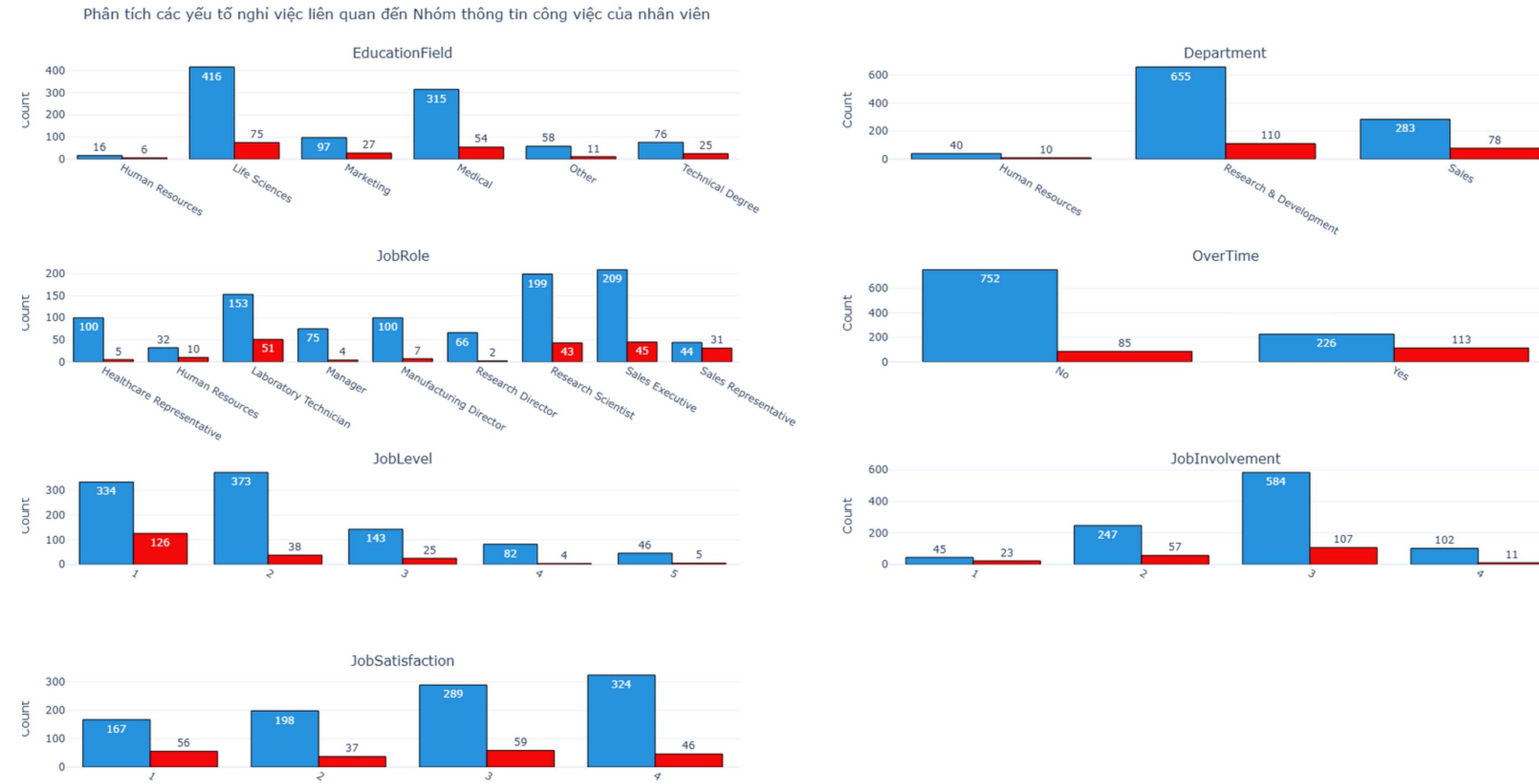
Mức độ cân bằng giữa công việc Tốt



Mức độ cân bằng giữa công việc Tốt nhất



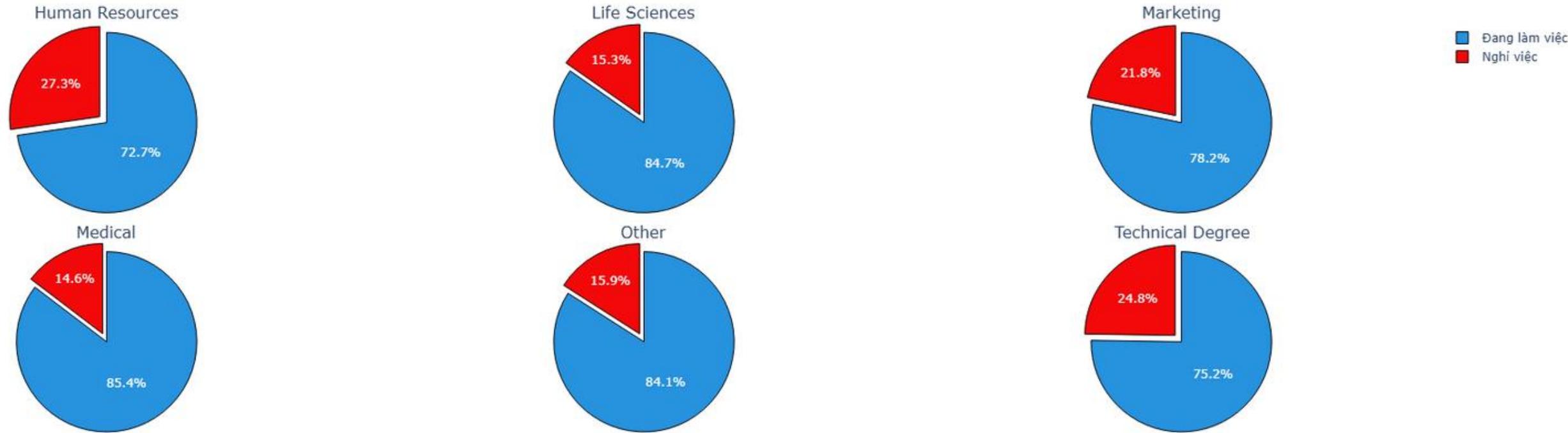
4. PHÂN TÍCH VỀ THÔNG TIN CÔNG VIỆC CỦA NHÂN VIÊN



- Nhận xét chung: Ta nhận thấy đa số những **người còn làm việc** luôn nhiều hơn những **người đã nghỉ việc**. Để làm rõ hơn tỉ lệ của **nhân viên nghỉ việc** chiếm bao nhiêu phần trăm trong từng đặc trưng, ta sẽ tiếp tục làm rõ

4. PHÂN TÍCH VỀ THÔNG TIN CÔNG VIỆC CỦA NHÂN VIÊN

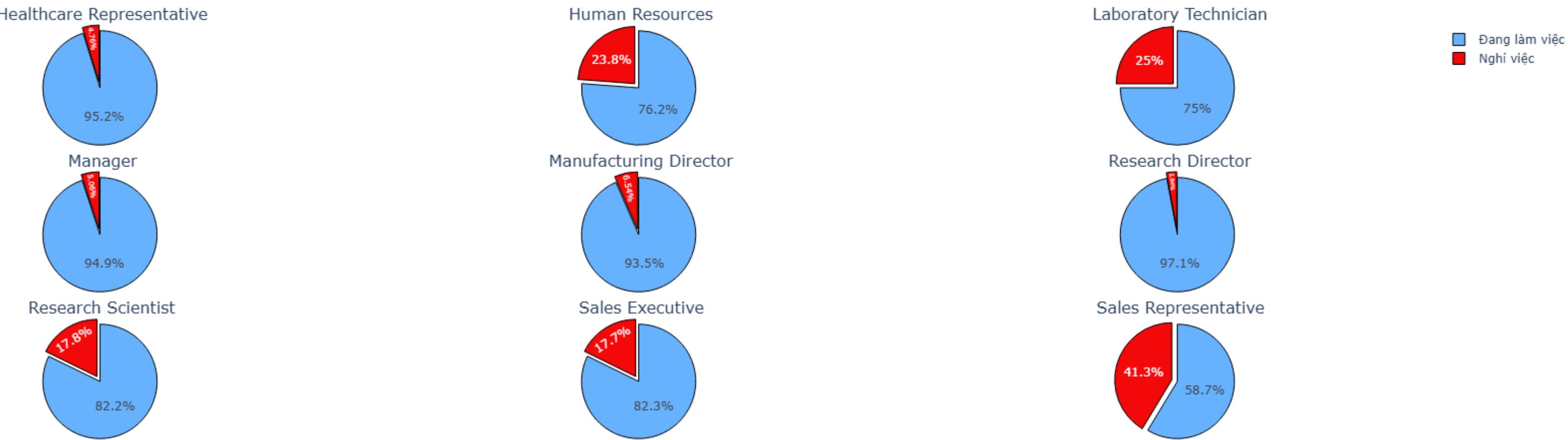
Tỷ lệ nhân viên nghỉ việc theo Lĩnh vực Giáo dục



- Nhận xét: 3 lĩnh vực Human Resources, Marketing và Technical Degree có **tỉ lệ nhân viên nghỉ việc cao nhất** trong tổng số 6 lĩnh vực giáo dục

4. PHÂN TÍCH VỀ THÔNG TIN CÔNG VIỆC CỦA NHÂN VIÊN

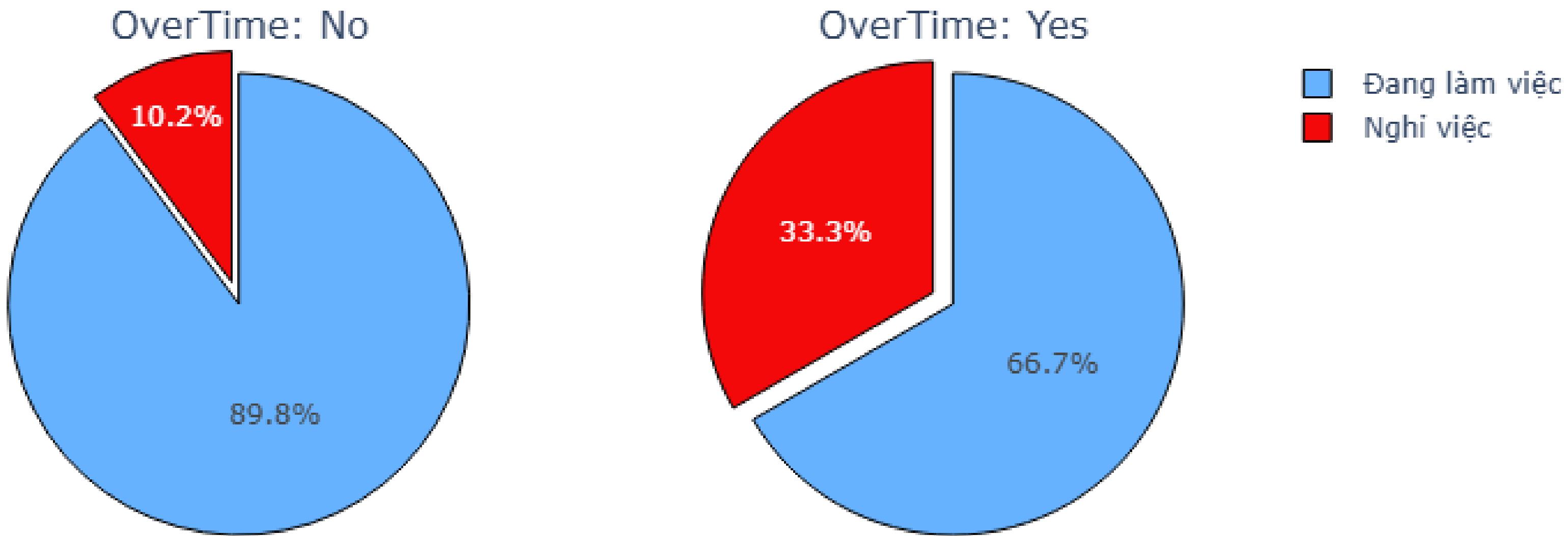
Tỷ lệ nhân viên nghỉ việc theo Vai trò công việc



- Nhận xét:

- Trong JobRole, có 4 trong tổng số 9 công việc có **tỷ lệ nghỉ việc** thấp hơn 7% bao gồm: Healthcare, Manager, Research Director và Manufacturing Director, đây là những vị trí có vai trò quan trọng trong công ty và thông thường những nhân viên ở vị trí này có sự ổn định và có thâm niên trong công ty nên **xu hướng nghỉ việc của họ thấp**.
- Các vai trò khác có tỷ lệ đều trên 15% trong đó ta có thể thấy ở vai trò Sales Executive và Sales Representative có **tỷ lệ nghỉ việc** khá cao (17.7% và 41.3%), điều này có thể giải thích vì sao ở Phòng ban này chiếm **tỷ lệ nhân viên nghỉ việc** cao nhất (đã chỉ ra ở trên)

4. PHÂN TÍCH VỀ THÔNG TIN CÔNG VIỆC CỦA NHÂN VIÊN



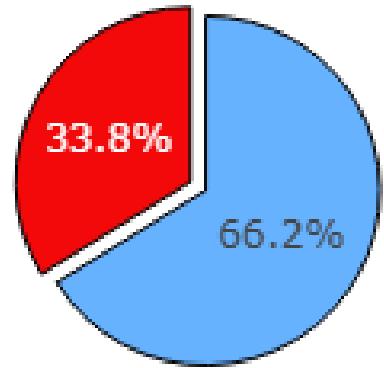
- Nhận xét:

- Ta nhận thấy đối với những nhân viên không làm thêm giờ tại công ty, **xu hướng họ nghỉ việc ít hơn**.
- Đối với những người làm **thêm giờ ở công ty**, ta nhận thấy **xu hướng họ nghỉ việc rất cao**, chiếm 1/4 trong tổng số nhân viên có làm thêm giờ. Nguyên nhân có thể đến từ áp lực từ công việc làm thêm, hay giá trị nhận được khi thực hiện làm thêm tại công ty không tương xứng với mức thưởng.

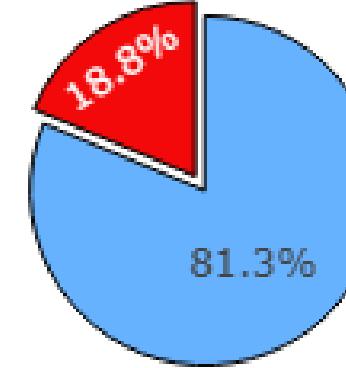
4. PHÂN TÍCH VỀ THÔNG TIN CÔNG VIỆC CỦA NHÂN VIÊN

Tỷ lệ nhân viên nghỉ việc theo Mức độ tham gia công việc

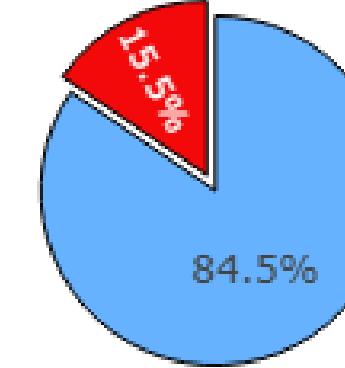
JobInvolvement: Low



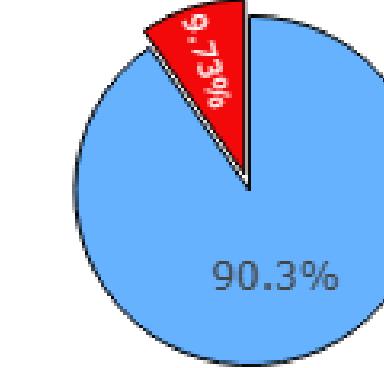
JobInvolvement: Medium



JobInvolvement: High



JobInvolvement: Very High



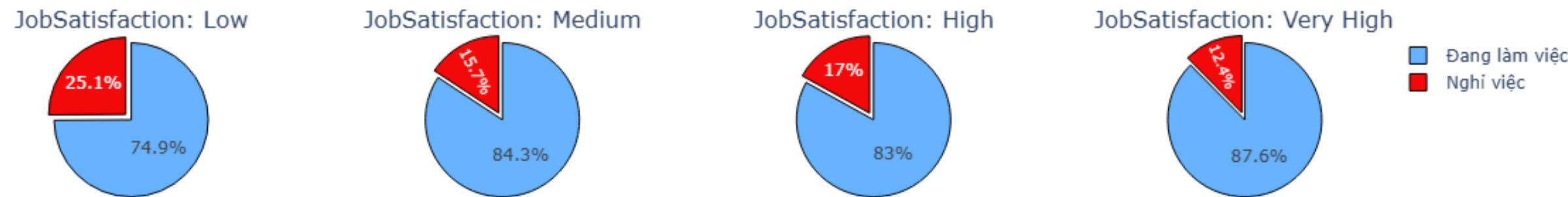
Đang làm việc
Nghi việc

- Nhận xét:

- Dựa vào biểu đồ ta có thể rút ra được nhận xét khi nhân viên có mức độ tham gia vào công việc của công ty càng nhiều thì **tỷ lệ nghỉ việc** của họ tại công ty càng giảm đi (33.8% đối với những người tham gia vào công việc ở mức độ Low và giảm dần còn dưới 10% cho vị trí Very High)

4. PHÂN TÍCH VỀ THÔNG TIN CÔNG VIỆC CỦA NHÂN VIÊN

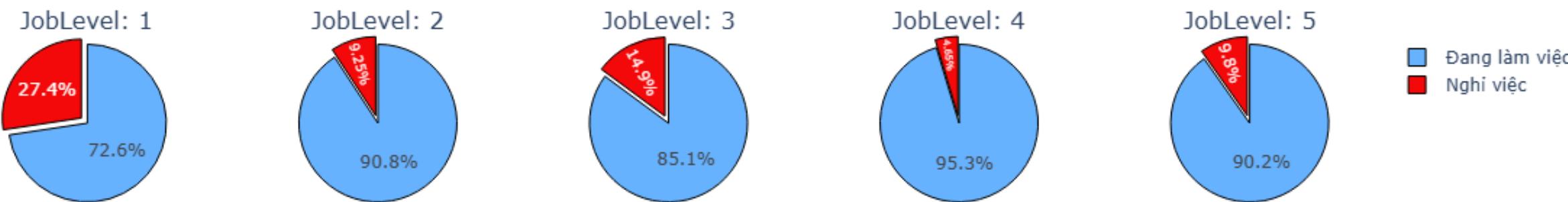
Tỷ lệ nhân viên nghỉ việc theo Mức độ hài lòng công việc



- Nhận xét:

- Tương tự như mức độ tham gia đóng góp vào công việc, mức độ hài lòng vào công việc của nhân viên có **tỷ lệ nhân viên nghỉ việc thấp** (khoảng 22.1%) và giảm dần khi mức độ hài lòng với công việc của họ rất cao (khoảng 12.4%), tuy nhiên những người có mức độ hài lòng công việc cao lại có **tỷ lệ nghỉ việc** nhỉnh hơn với những người có sự thỏa mãn với công việc ở mức trung bình (17% so với 12.7%) nhưng nhìn chung ở 2 cấp bậc này không chênh lệch nhiều.

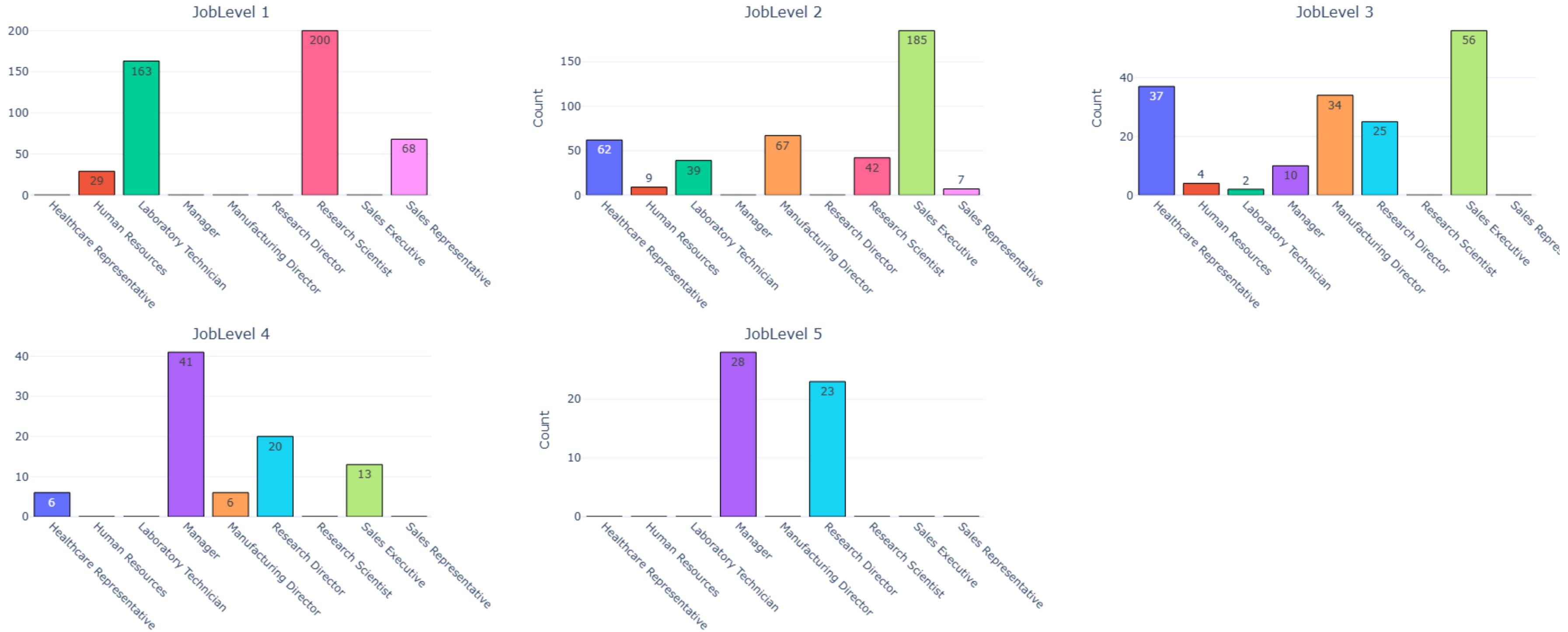
Tỷ lệ nhân viên nghỉ việc theo Cấp độ công việc



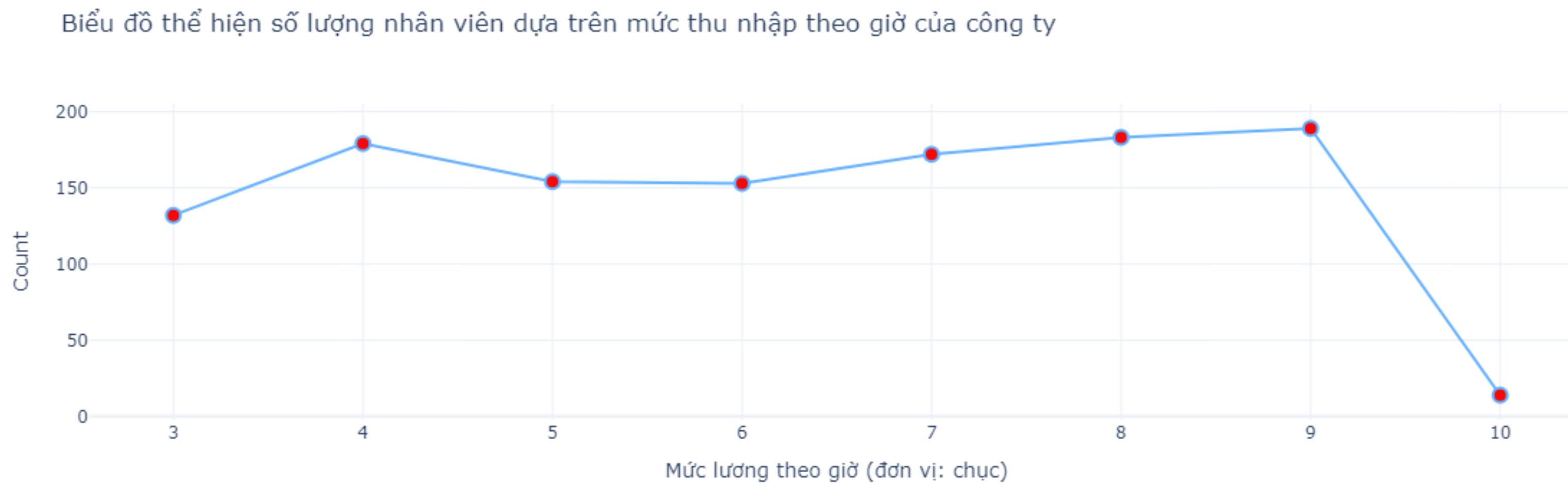
- Nhận xét: Quan sát theo từng cấp độ, ta nhận thấy cấp độ công việc có **tỉ lệ nghỉ việc cao nhất** là 1 (khoảng 27.4%) tuy nhiên nó không tuân theo một quy luật cụ thể nào. Vì thế ta sẽ làm rõ trong mỗi Cấp bậc công việc thì Vị trí công việc nào có tác động đến tỉ lệ nghỉ việc

4. PHÂN TÍCH VỀ THÔNG TIN CÔNG VIỆC CỦA NHÂN VIÊN

Phân bố Vai trò công việc theo Cấp độ công việc



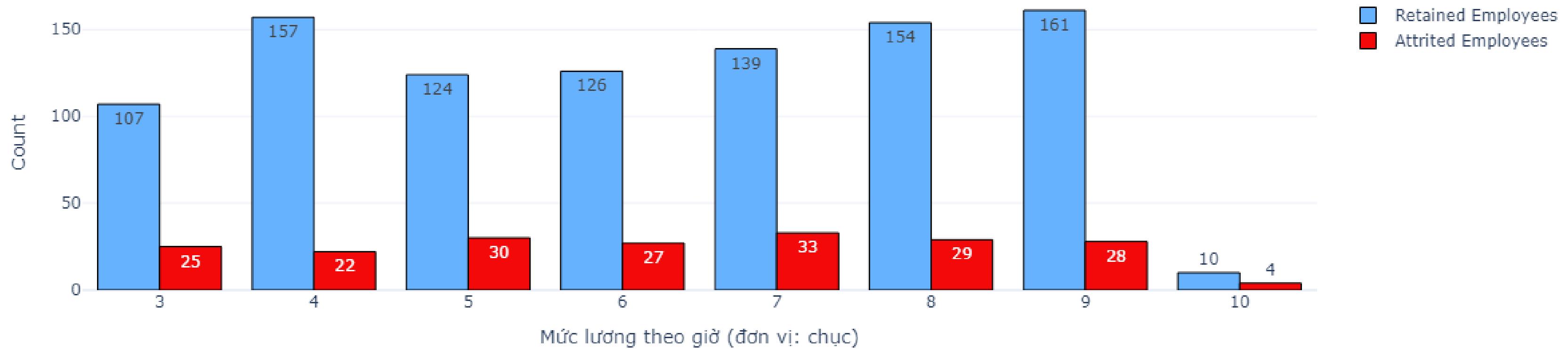
5. PHÂN TÍCH VỀ TÌNH HÌNH LƯƠNG THƯỞNG CỦA NHÂN VIÊN



Nếu tính lương theo giờ, số lượng nhân viên có **mức lương từ 30 - 90** không chênh lệch nhiều thể hiện qua đường nằm ngang của biểu đồ, tuy nhiên số lượng nhân viên có mức lương 90 chiếm rất ít

5. PHÂN TÍCH VỀ TÌNH HÌNH LƯƠNG THƯỞNG CỦA NHÂN VIÊN

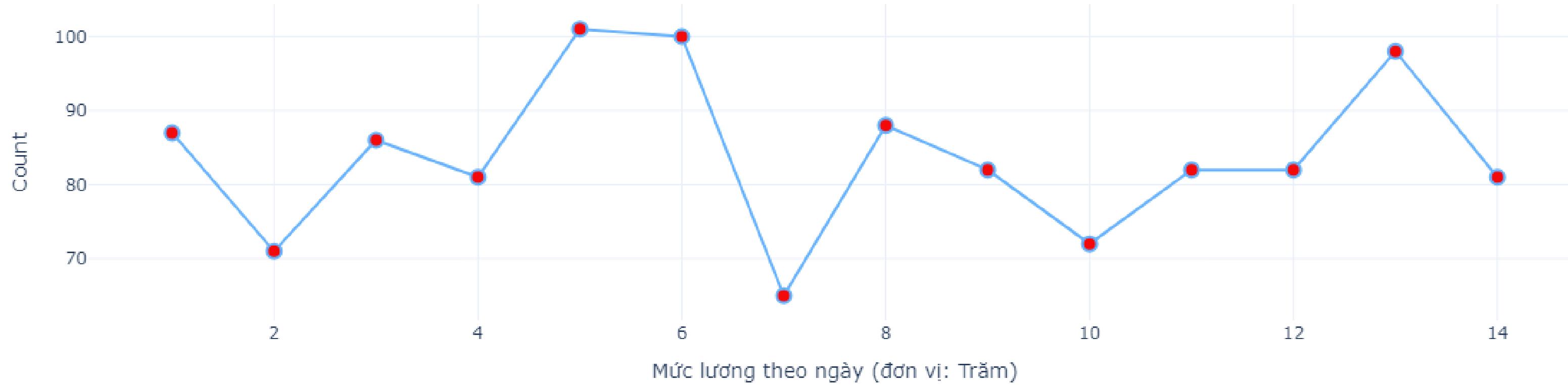
Biểu đồ thể hiện số lượng nhân viên đang làm việc và nghỉ việc trên mức thu nhập theo giờ của công ty



Tuy nhiên khi nhìn vào biểu đồ cột, ta lại thấy ở những người có mức lương 30 - 90, số lượng **người nghỉ việc** ở các nhóm lương theo giờ lại không chênh lệch quá nhiều (khoảng từ 22 - 35 người nghỉ việc trên từng mức lương khác nhau), và **số lượng người nghỉ việc** ở mức lương 100 chiếm ít nhất nhưng xét về tỉ lệ giữa người còn đi làm và người nghỉ việc thì lại chiếm tỉ lệ là 2:1

5. PHÂN TÍCH VỀ TÌNH HÌNH LƯƠNG THƯỞNG CỦA NHÂN VIÊN

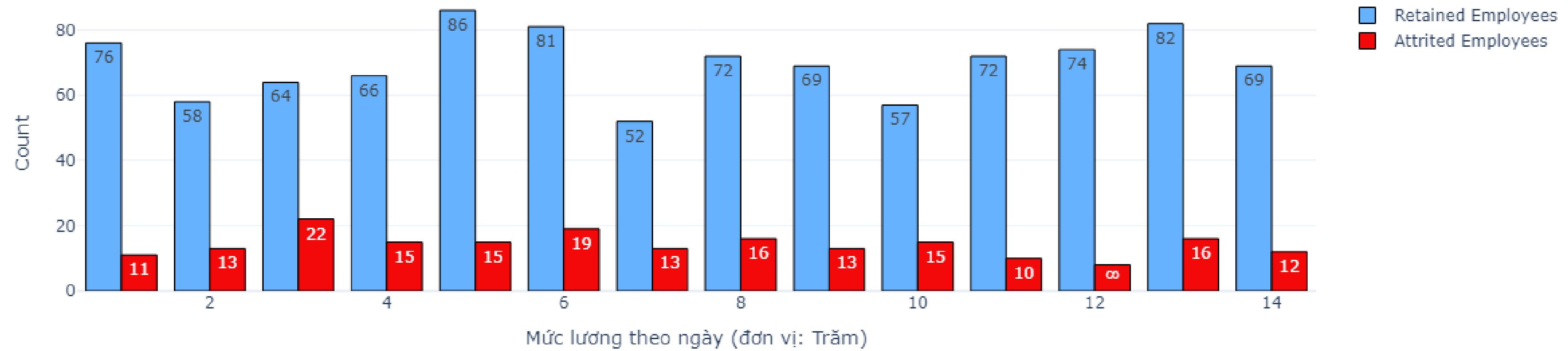
Biểu đồ thể hiện số lượng nhân viên dựa trên mức thu nhập theo ngày của công ty



Những người có mức lương tính trên ngày chiếm nhiều nhất là từ 500 - 600. Trong đó , ở mức từ 600 - 700 lại có số lượng thấp nhất (dưới 70 nhân viên)

5. PHÂN TÍCH VỀ TÌNH HÌNH LƯƠNG THƯỞNG CỦA NHÂN VIÊN

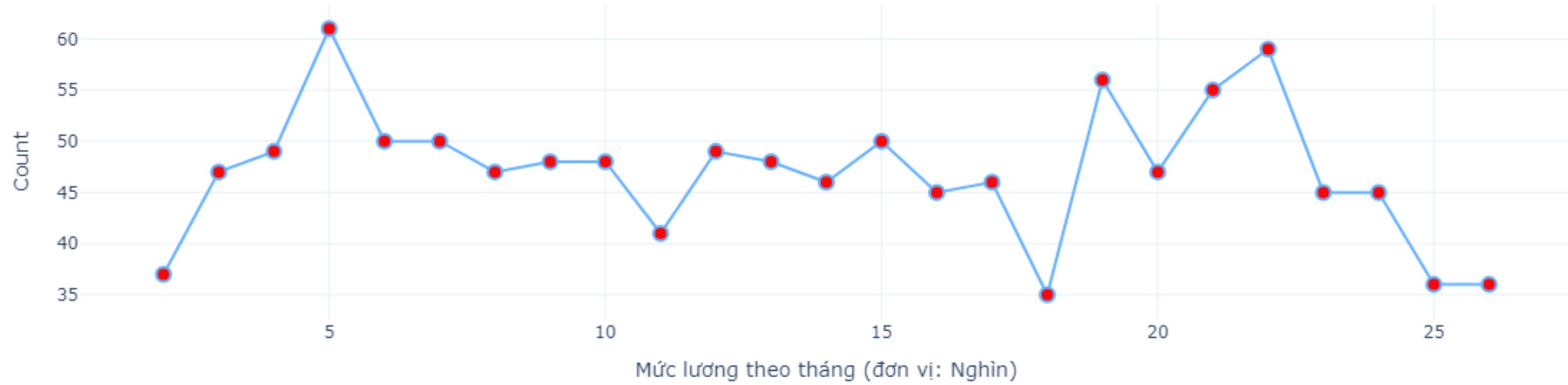
Biểu đồ thể hiện số lượng nhân viên đang làm việc và nghỉ việc trên mức thu nhập theo ngày của công ty



Tuy nhiên khi so sánh về **số lượng nhân viên nghỉ việc** ở mỗi mức lương ta thấy được sự chênh lệch không quá lớn (cao nhất có 22 nhân viên nghỉ việc ở mức lương 300/ngày và thấp nhất ở mức 1200/ngày chỉ có 8 nhân viên)

5. PHÂN TÍCH VỀ TÌNH HÌNH LƯƠNG THƯỞNG CỦA NHÂN VIÊN

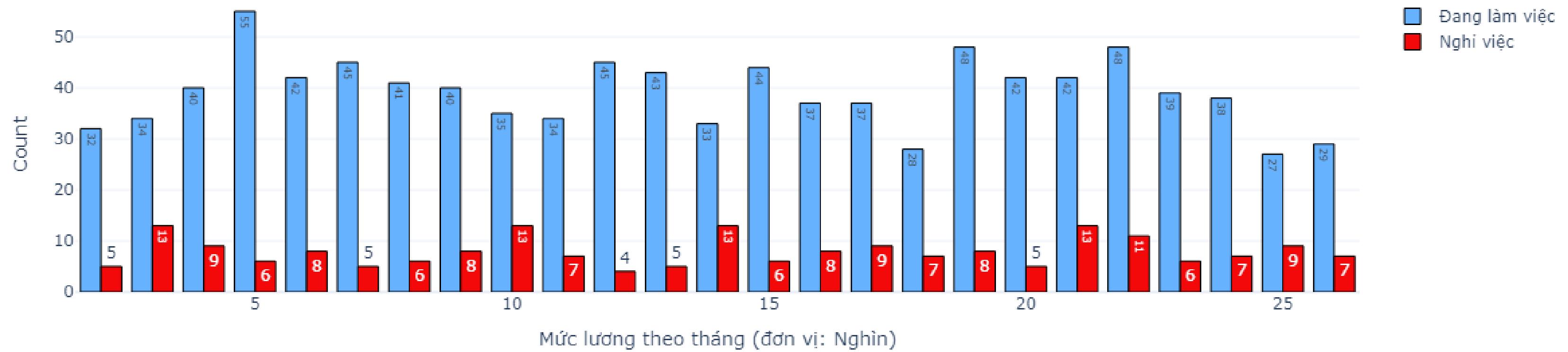
Biểu đồ thể hiện số lượng nhân viên trên mức thu nhập theo tháng của công ty



Ta thấy những người có mức lương từ 20000 - 22000 chiếm số lượng nhiều nhất

5. PHÂN TÍCH VỀ TÌNH HÌNH LƯƠNG THƯỞNG CỦA NHÂN VIÊN

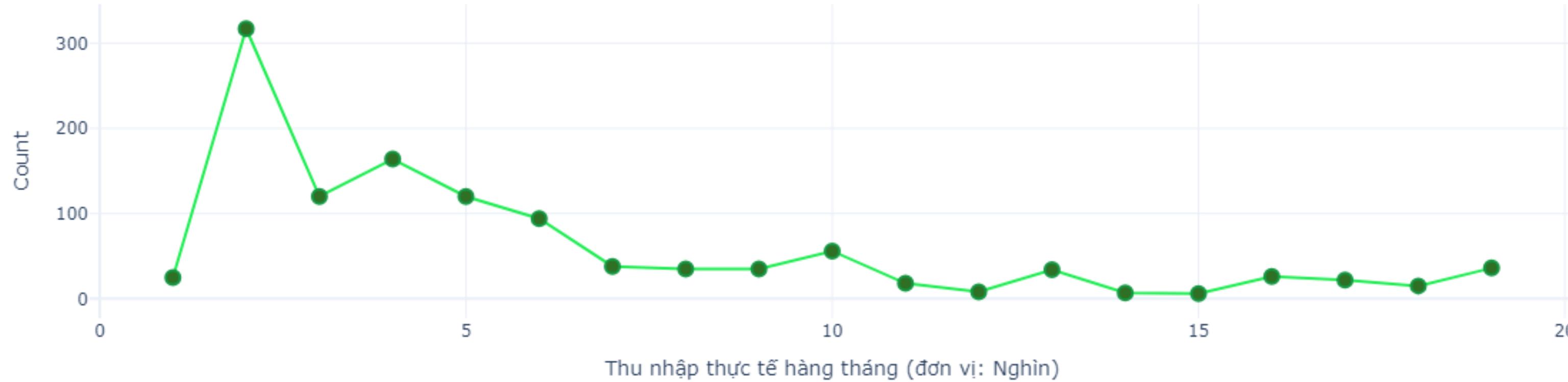
Biểu đồ thể hiện số lượng nhân viên đang làm việc và nghỉ việc trên mức thu nhập theo tháng của công ty



Số lượng **nhân viên nghỉ việc** ở từng mức lương không có sự chênh lệch lớn

5. PHÂN TÍCH VỀ TÌNH HÌNH LƯƠNG THƯỞNG CỦA NHÂN VIÊN

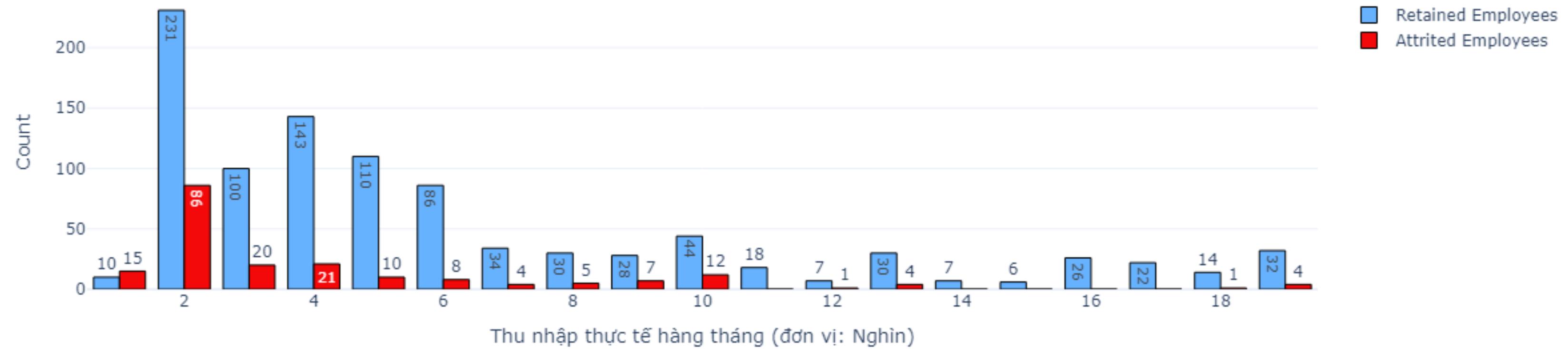
Biểu đồ thể hiện số lượng nhân viên của công ty dựa trên mức thu nhập thực tế



Nhân viên có mức lương từ 0 - 5000 chiếm số lượng đông nhất trong công ty và khi mức lương càng tăng cao, thì số lượng nhân viên có thu nhập cao càng giảm dần

3. PHÂN TÍCH VỀ TÌNH HÌNH TÀI CHÍNH CỦA CÔNG TY

Biểu đồ thể hiện số lượng nhân viên đang làm việc và nghỉ việc trên mức thu nhập thực tế của công ty



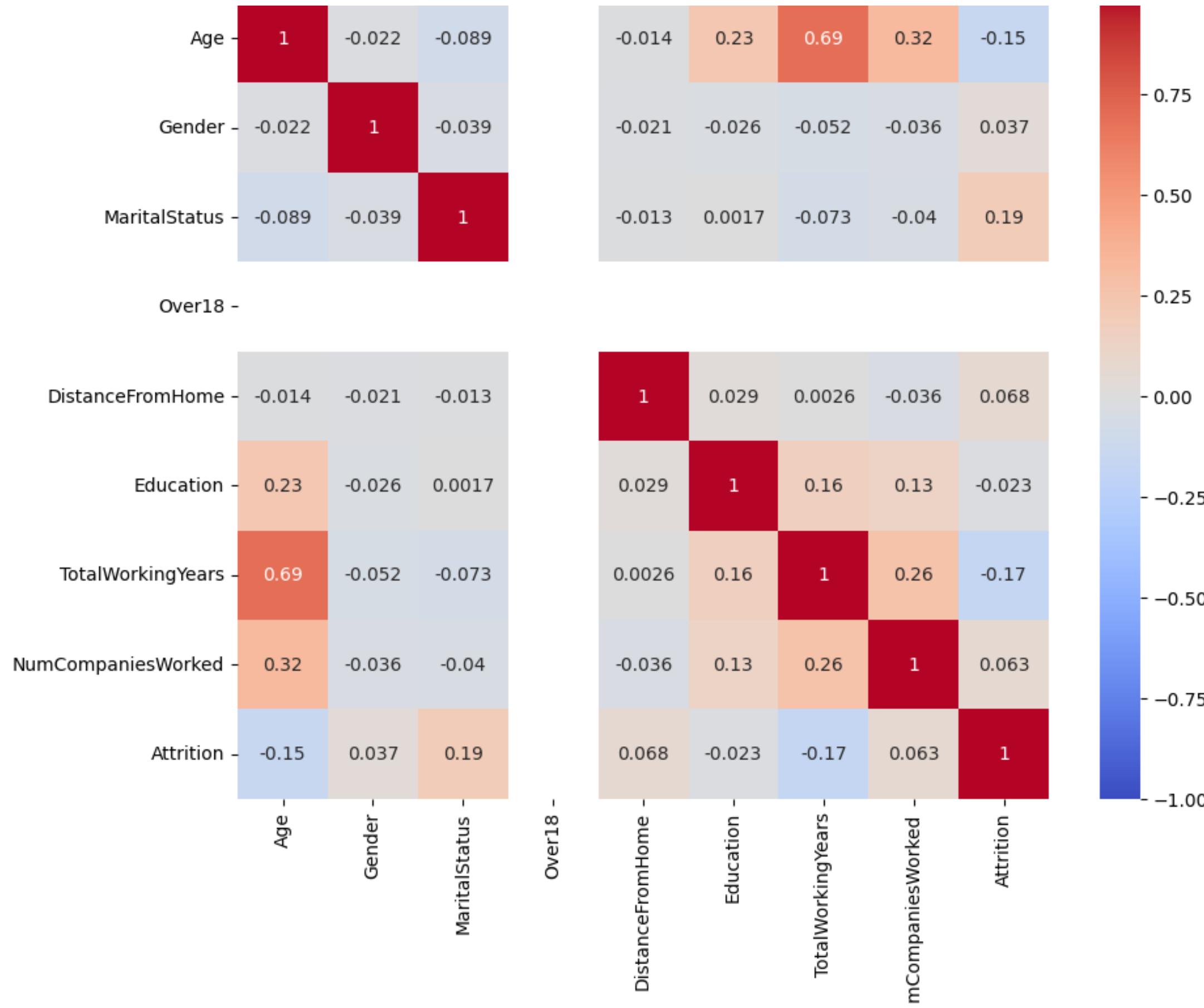
Đa số những nhân viên nghỉ việc sẽ rời vào nhưng người có thu nhập dưới 5000, có thể đến từ việc mức thu nhập thấp khiến họ phải rời khỏi công ty

TƯƠNG QUAN

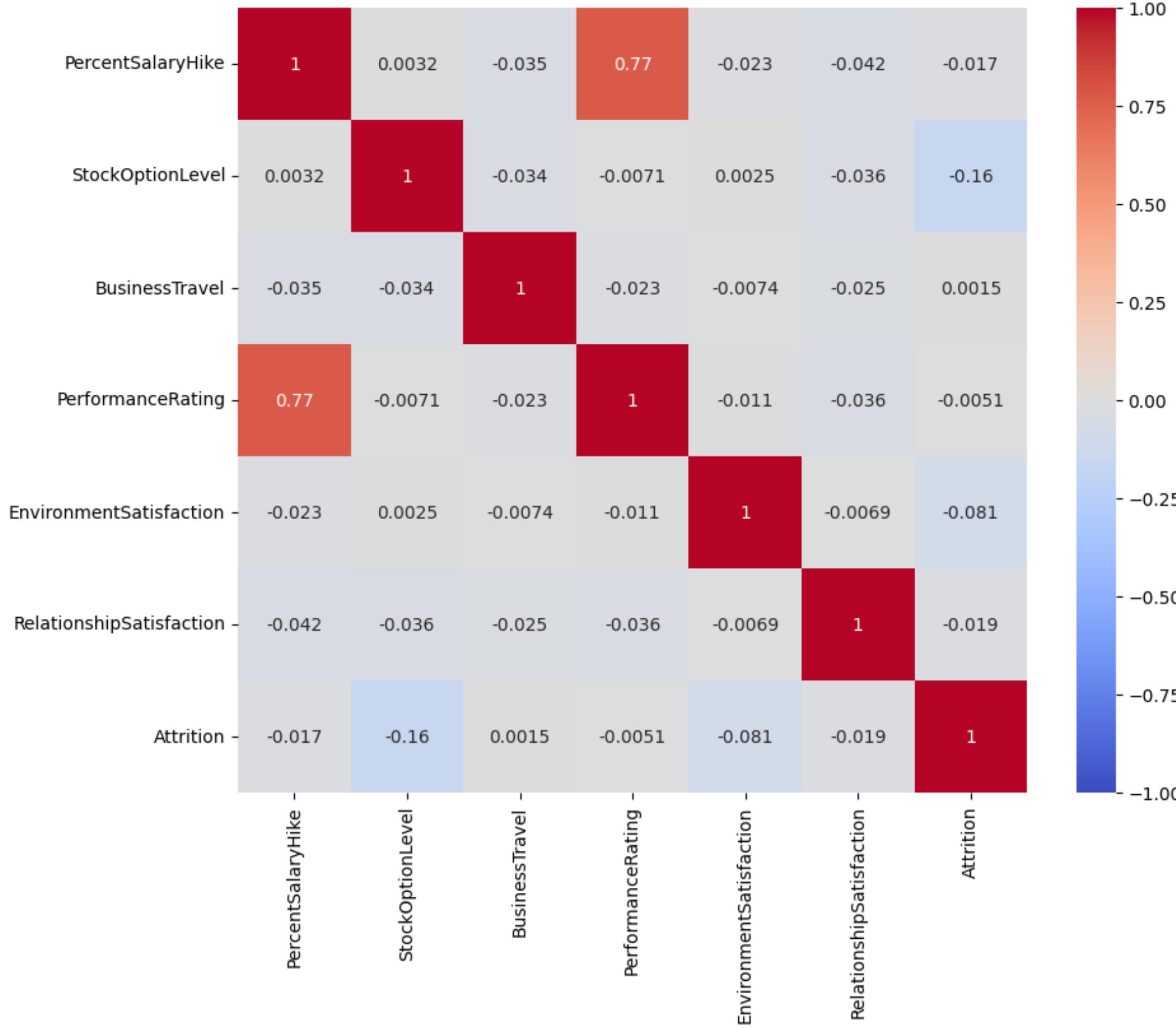
Dựa trên các trường dữ liệu, ta sẽ nhóm chúng thành 5 danh sách để phân tích tương quan với Attrition:

- **Features_info_employee** : Age, Gender, MaritalStatus, Over18, DistanceFromHome, Education, TotalWorkingYears, NumCompaniesWorked
- **Features_info_job**: EducationField, Department, JobLevel, JobRole, JobInvolvement, OverTime, JobSatisfaction
- **Features_info_emp_company**: YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager, YearsSinceLastPromotion, TrainingTimesLastYear, WorkLifeBalance
- **Features_info_company**: PercentSalaryHike, StockOptionLevel, BusinessTravel, PerformanceRating, EnvironmentSatisfaction, RelationshipSatisfaction
- **Features_salary**: MonthlyIncome, HourlyRate, DailyRate, MonthlyRate

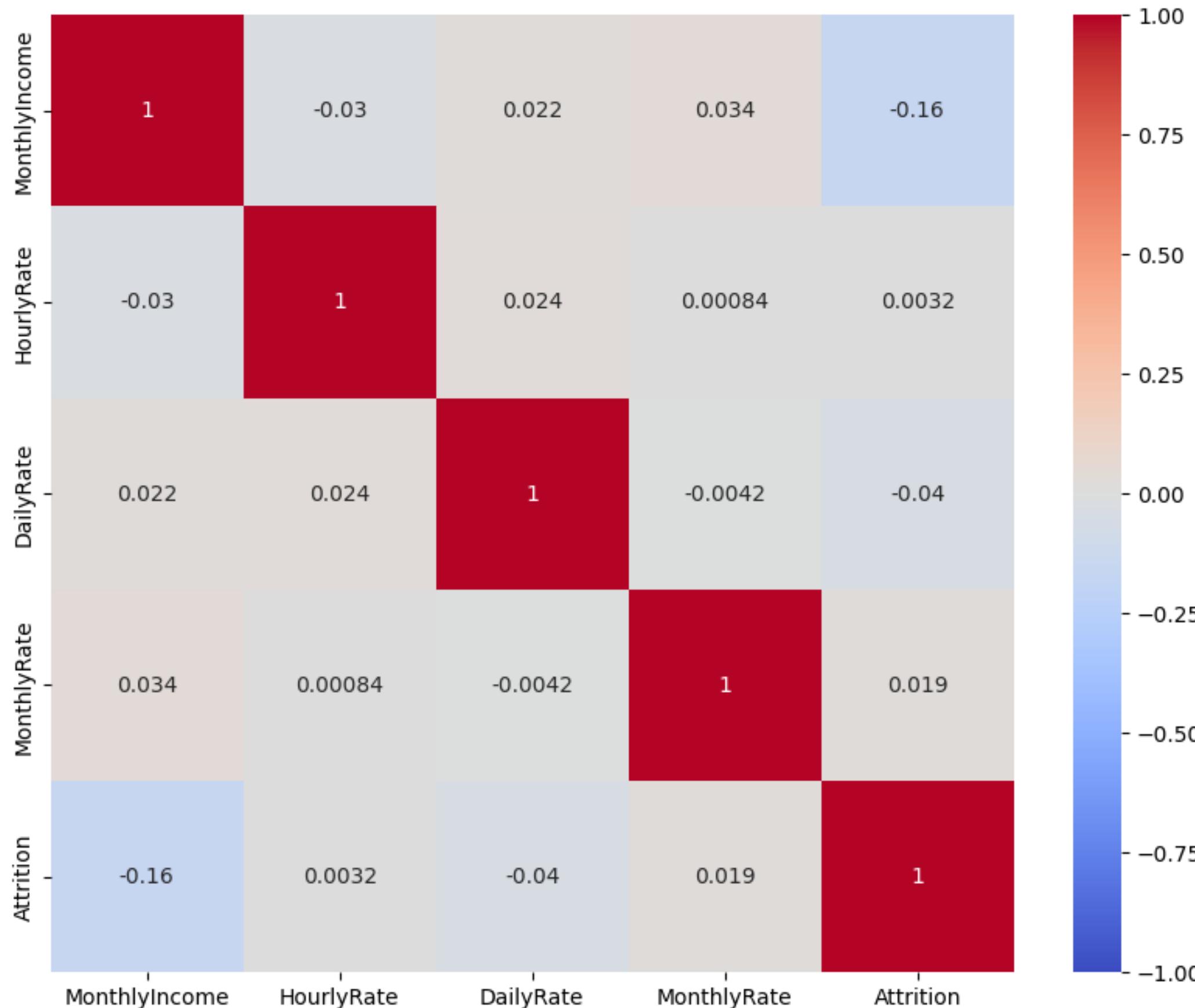
Ma trận tương quan của Attrition với nhóm Features_info_employee



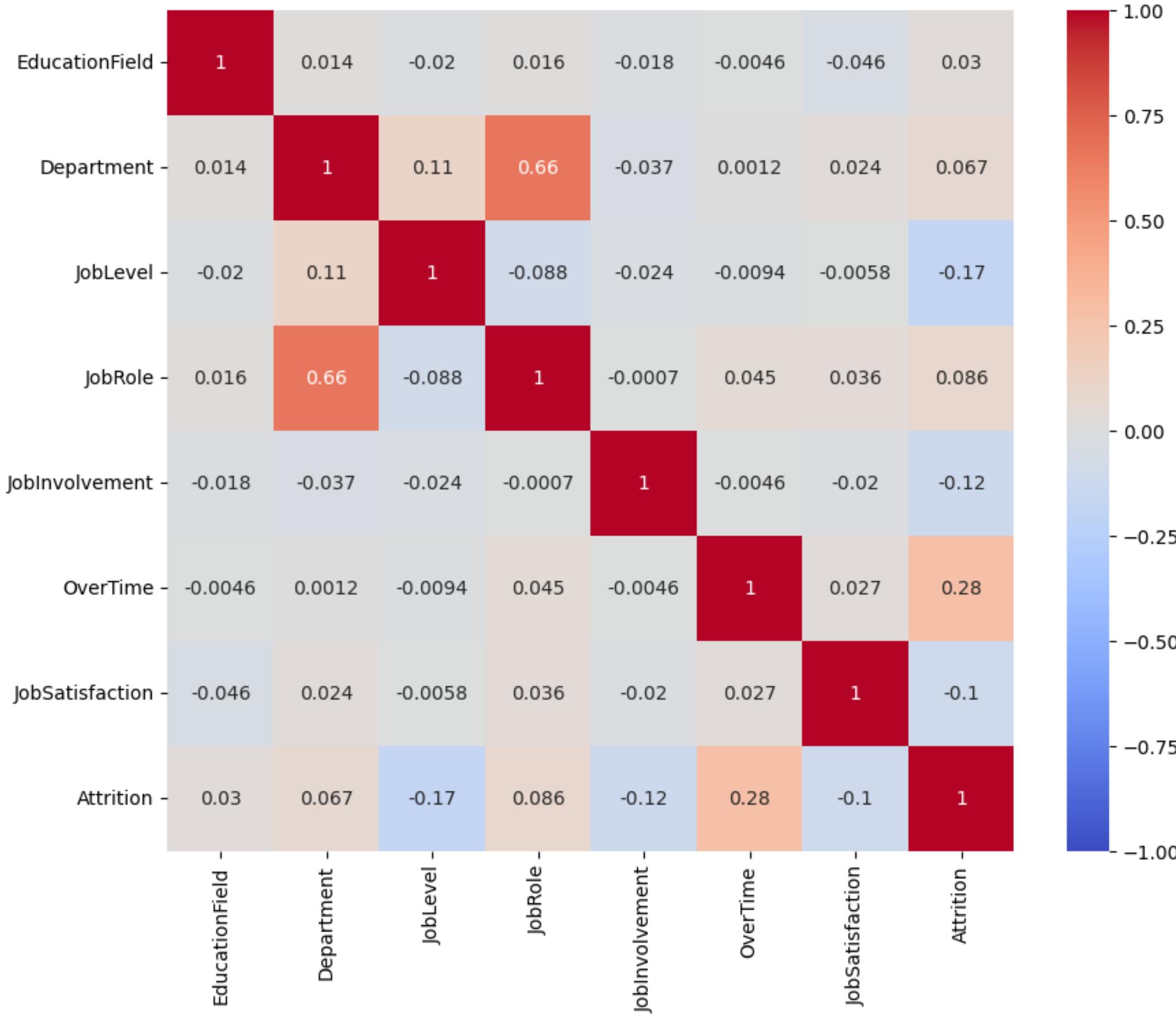
Ma trận tương quan của Attrition với nhóm Features_info_job



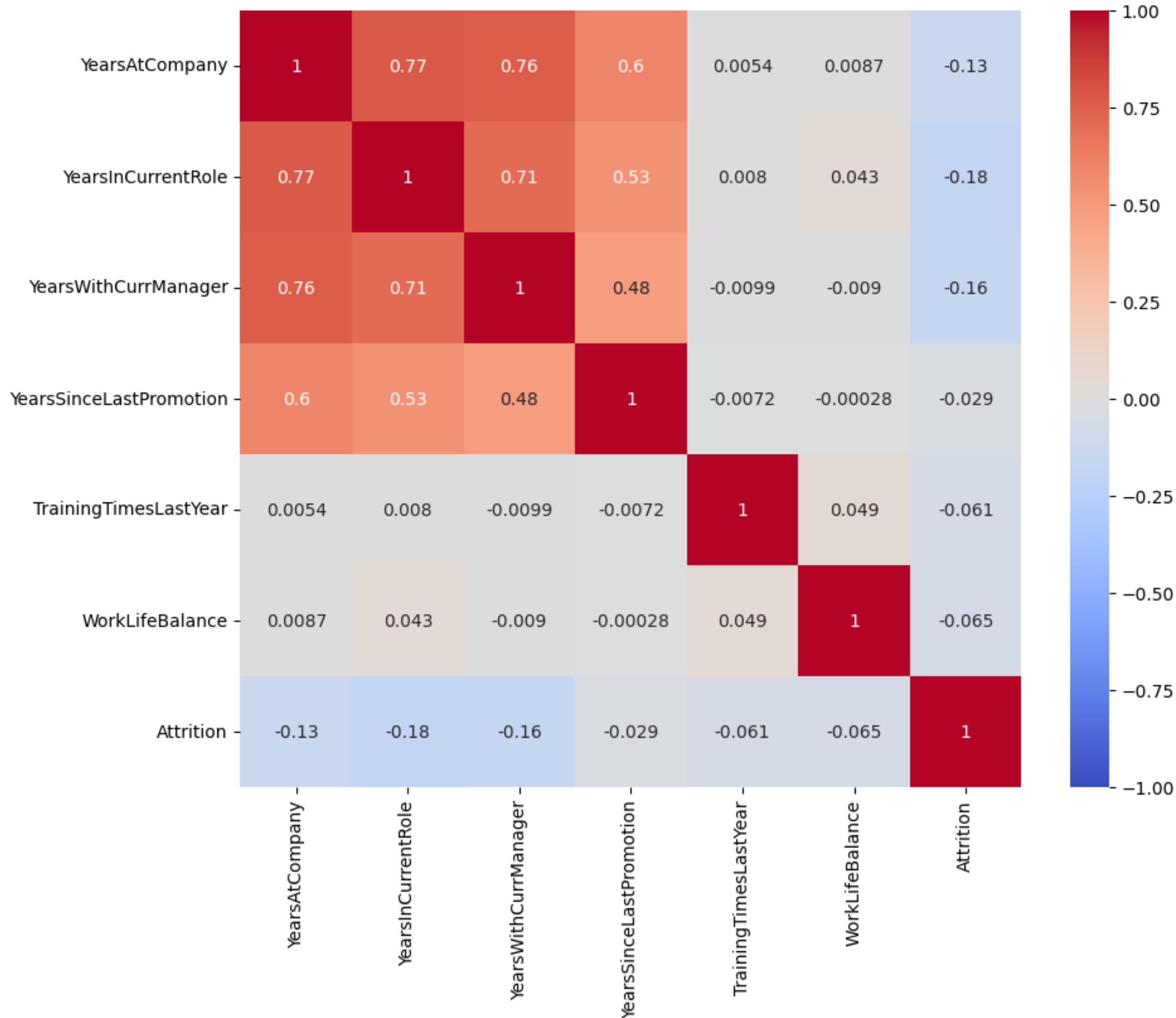
Ma trận tương quan của Attrition với nhóm Features_info_emp_company

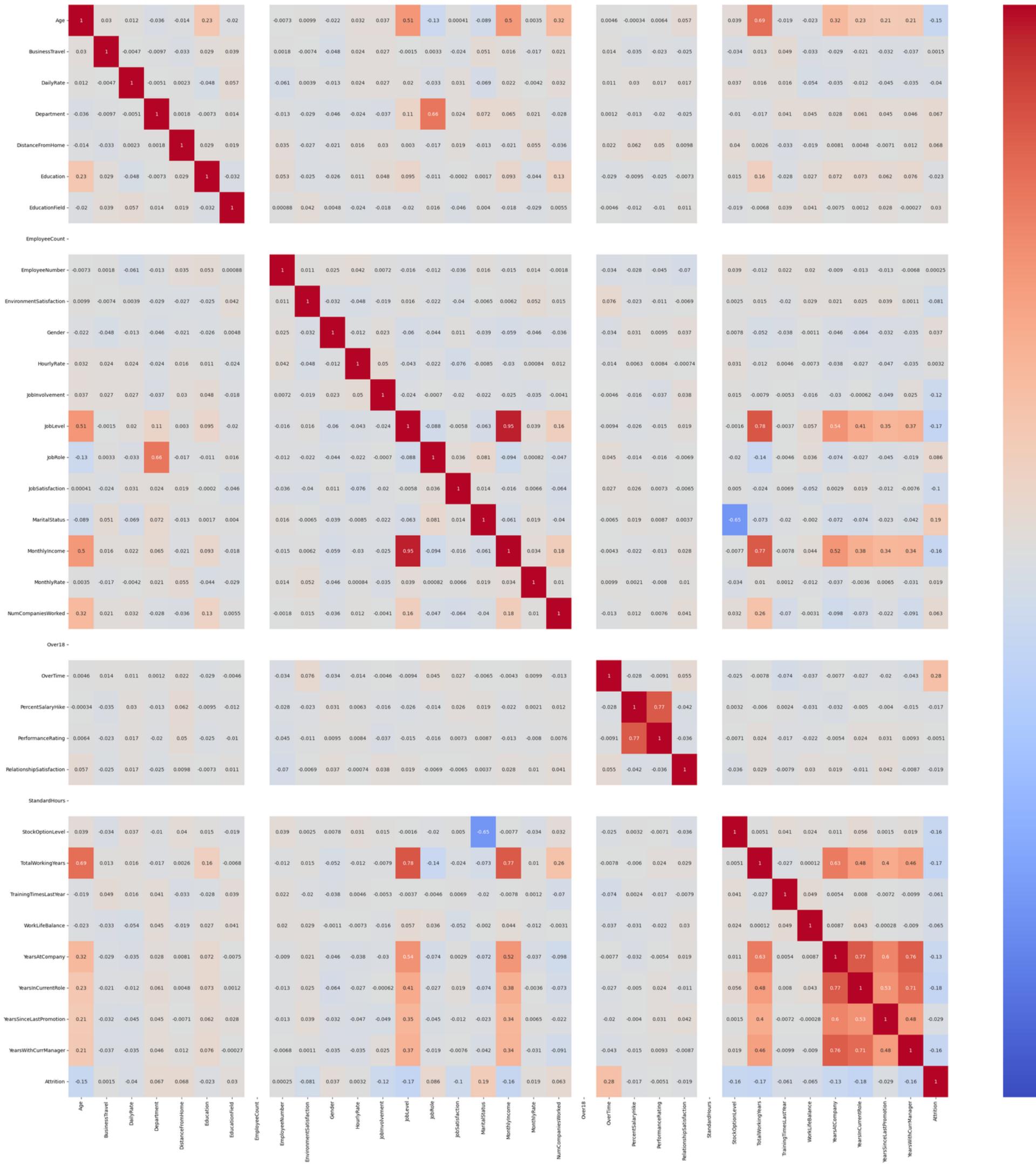


Ma trận tương quan của Attrition với nhóm Features_info_company



Ma trận tương quan của Attrition với nhóm Features_salary





XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH

CÁC BƯỚC TIẾN HÀNH XÂY DỰNG MODEL

Sử dụng tập train - data để thực hiện tìm kiếm hyperparams tốt nhất ở mỗi model



Sau khi có hyperparams phù hợp, tiến hành training lại model với hyperparams đó

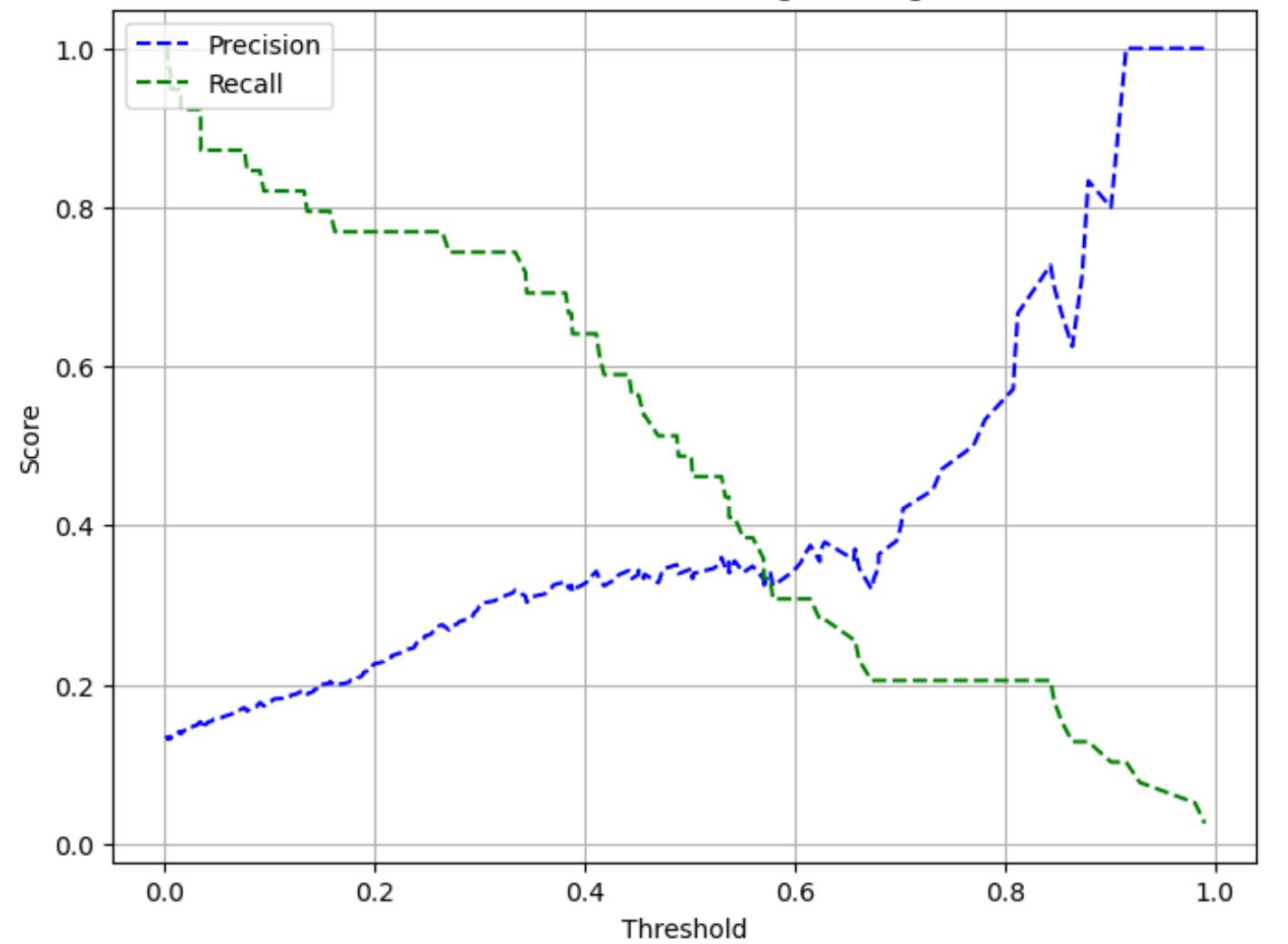


Sử dụng tập test-data để đưa ra đánh giá kết quả models

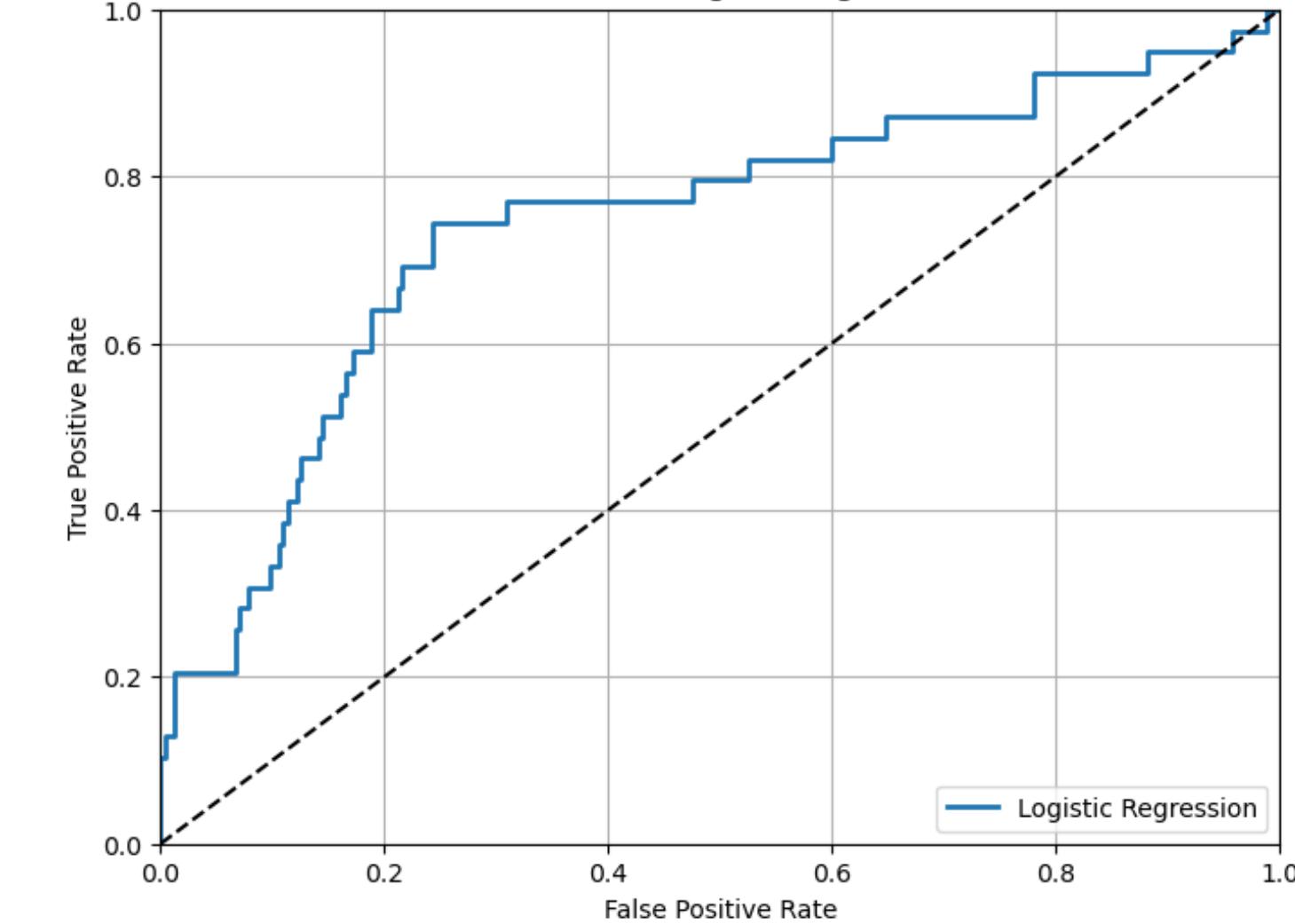


Sử dụng các metrics: Recall, Precision, F1 - scores, ROC Curve để đánh giá models

Precision/Recall Tradeoff - Logistic Regression



ROC Curve - Logistic Regression



Validation Logistic Regression model

Final evaluation on test set:

F1 Score: 0.4043

Recall: 0.4872

Precision: 0.3455

ROC AUC Score: 0.7432

Logistic Regression: Mô hình này thể hiện khả năng phân loại ổn định, nhưng hiệu quả dự đoán lớp nghỉ việc còn hạn chế do sự cân bằng giữa Precision và Recall chưa tối ưu. Phù hợp cho sàng lọc ban đầu, nhưng cần cải thiện để nhận diện sớm nhân viên có nguy cơ nghỉ việc với ít dự đoán sai hơn.



Validation Random Forest model

Final evaluation on test set:

F1 Score: 0.3514

Recall: 0.3333

Precision: 0.3714

ROC AUC Score: 0.7324

Random Forest: Dù được kỳ vọng cao, mô hình này không vượt trội, với khả năng phân loại kém và bỏ sót nhiều trường hợp nghỉ việc thực tế. Hiệu quả tổng thể thấp, khiến nó ít phù hợp cho bài toán này.



Validation Support Vector Machine model

Final evaluation on test set:

F1 Score: 0.2545

Recall: 0.1795

Precision: 0.4375

ROC AUC Score: 0.7169

SVM: Mô hình này kém hiệu quả nhất trong việc nhận diện các trường hợp nghỉ việc, bỏ sót nhiều trường hợp thực tế. Chỉ phù hợp khi ưu tiên độ chính xác cao, nhưng không lý tưởng cho mục tiêu phát hiện sớm.



Validation XGBoost model

Final evaluation on test set:

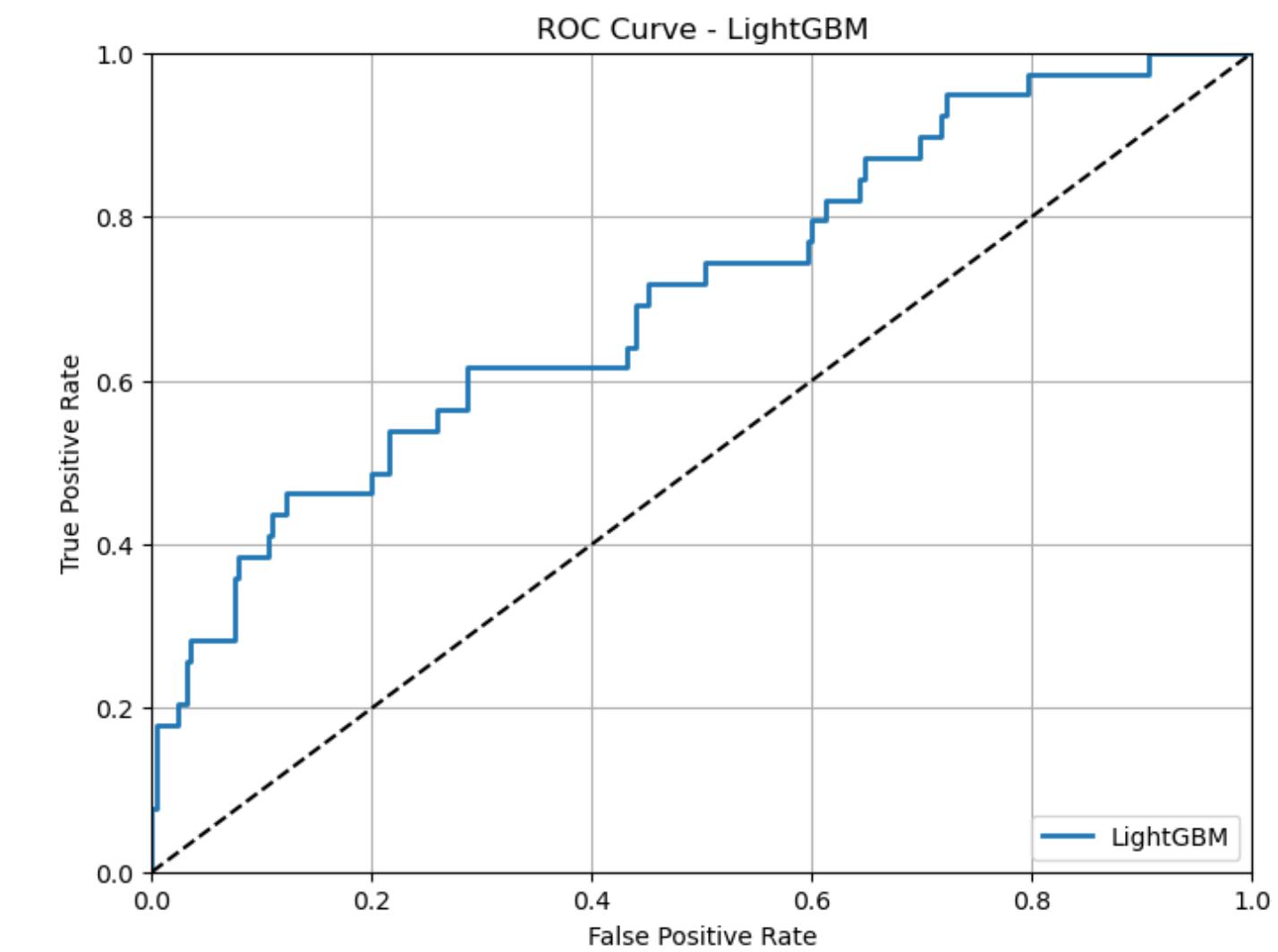
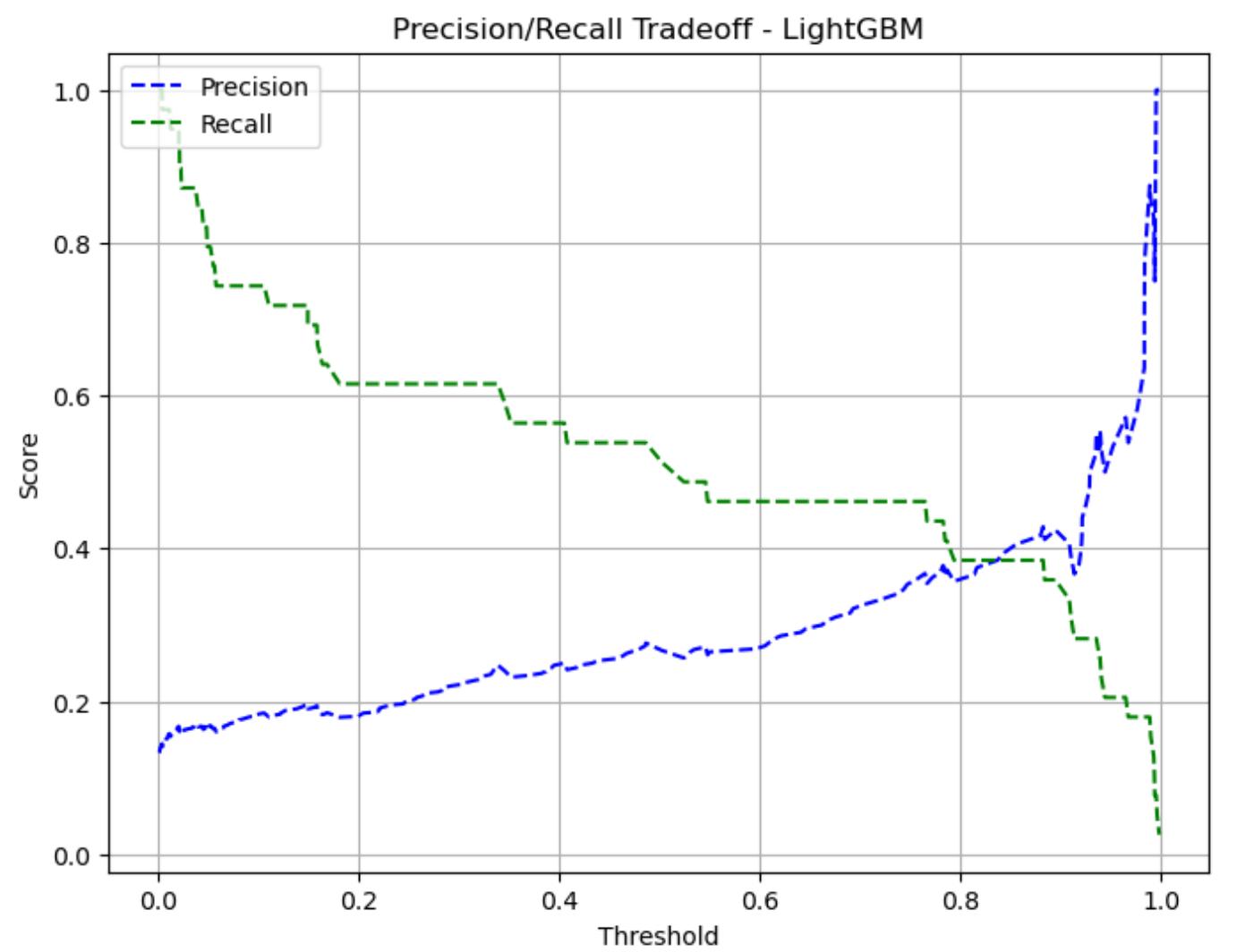
F1 Score: 0.3333

Recall: 0.6154

Precision: 0.2286

ROC AUC Score: 0.7192

XGBoost: Nổi bật với khả năng phân loại tốt, đặc biệt trong việc nhận diện các trường hợp nghỉ việc. Tuy nhiên, tỷ lệ dự đoán dương tính sai cao có thể gây lãng phí nguồn lực khi triển khai.



Validation LightGBM model

Final evaluation on test set:

F1 Score: 0.3509

Recall: 0.5128

Precision: 0.2667

ROC AUC Score: 0.7067

LightGBM: Mô hình này nhận diện tương đối tốt các trường hợp nghỉ việc, nhưng tỷ lệ dự đoán sai cao dẫn đến hiệu quả tổng thể kém hơn so với XGBoost và Logistic Regression.



Validation CatBoost model

Final evaluation on test set:

F1 Score: 0.3853

Recall: 0.5385

Precision: 0.3000

ROC AUC Score: 0.7191

CatBoost: Thể hiện hiệu suất vượt trội trong việc cân bằng Precision và Recall, rất phù hợp để phát hiện sớm nhân viên có nguy cơ nghỉ việc. Tuy nhiên, vẫn cần quản lý các dự đoán sai để tối ưu hóa hiệu quả.

Tổng thể: CatBoost là **lựa chọn tối ưu nhất** khi ưu tiên phát hiện sớm nhân viên có nguy cơ nghỉ việc, nhưng cần điều chỉnh để giảm thiểu dự đoán dương tính sai. XGBoost cũng là một **phương án khả thi** nếu cần tối ưu khả năng phân loại tổng thể. Công ty nên thử nghiệm thêm các kỹ thuật như điều chỉnh trọng số lớp hoặc kết hợp mô hình để cải thiện hiệu suất, đồng thời xây dựng quy trình ứng dụng kết quả dự đoán để can thiệp kịp thời, giảm tỷ lệ nghỉ việc hiệu quả.

=> Việc sử dụng các mô hình phù hợp giúp cho công ty đánh giá được nguồn lực nhân viên trong tương lai, từ đó có thể đưa ra những chính sách phù hợp trong việc tuyển dụng, cũng như việc tăng phúc lợi cho nhân viên, ... để tối ưu được chi phí và lợi nhuận cho công ty