

Proyek Akhir KASDD: Resignation Intention

Kelompok YukBisaYuk

ANGGOTA KELOMPOK YUKBISAYUK

Link project:

ristek.link/DeepnoteYukBisaYuk

2006531951 - Andi Afifah Khairunnisa

2006532903 - Muhammad Damar Kusumo

2006533811 - Sultan Fahrezy Syahdwinata

2006596314 - Ekky Aliansyah

OUTLINE

1 Data Understanding

Pada bagian ini kami akan menjelaskan data yang kami gunakan. Penjelasan tiap kolomnya, jenis data dari setiap kolom, statistik data, dsb.

2 Data Preprocessing

Pada bagian ini kami akan melakukan penanganan terhadap banyak penemuam yang kami lakukan pada tahap sebelumnya agar model yang akan dibuat menjadi lebih baik

3 EDA

Pada bagian ini kami akan menampilkan data yang kami miliki dalam bentuk visualisasi agar lebih mudah dipahami oleh pembaca

4 Membuat Model

Pada bagian ini kami akan membuat model untuk melakukan klasifikasi, regresi, dan clustering

Co.

1

Data

Understanding



JUMLAH BARIS DAN KOLOM

```
total_rows, total_attributes = df.shape  
print()  
print('Jumlah data:', total_rows)  
print("Jumlah atribut:", total_attributes)
```

[51]

Jumlah data: 1470

Jumlah atribut: 30

Terdapat 1470 Rows Data dengan 30 Fitur

RINGKASAN DESKRIPSI STATISTIK DATASET

df.describe()

	age float64	home_distance fl...	education float64	score_environm...	hourly_rate float64	score_contributi...	job_rank float64
count	1470.0	1470.0	1470.0	1470.0	1470.0	1470.0	1470.0
mean	36.92380952380 9524	9.192517006802 72	2.912925170068 027	2.721768707482 993	65.89115646258 503	2.729931972789 1156	2.063945578231 2925
std	9.135373489136 ---	8.106864435666 ---	1.024164944597 ---	1.093082214635 ---	20.32942759399 ---	0.711561142963 ---	1.106939898935 ---

df.describe(include= 'object')

	resign object	division object	employee_id obj...	major object	gender object	role object	marriage_status o...
count	1470	1470	1470	1470	1470	1470	1470

RINGKASAN DESKRIPSI STATISTIK DATASET

- Rata-rata Usia karyawan adalah 37 tahun
- Karyawan mendapatkan promosi dalam 2-5 tahun
- Rata-rata karyawan sudah bekerja selama 7 tahun di perusahaan
- Dataset imbalanced, karena jumlah data dengan nilai resign = No sejumlah 1233 sementara resign = Yes (mengundurkan diri dari perusahaan) hanya 237 yang sebetulnya menunjukkan hal yang baik karena lebih banyak karyawan yang memilih untuk tetap bekerja di perusahaan dibandingkan resign.
- Perusahaan memiliki lebih banyak karyawan laki-laki dibanding perempuan
- Tidak ada karyawan underage yang bekerja di perusahaan
- Kebanyakan karyawan tidak bekerja over_time (1054/1470)

DATA DUPLIKASI

Tidak terdapat data duplikat pada dataset ini. Sehingga tidak ada penanganan khusus untuk duplikasi data.

```
stats = df.duplicated().to_frame("redundant")
red = stats[stats["redundant"] == True].count()
print("Jumlah data " + str(red))
stats[stats["redundant"] == True]
```



```
Jumlah data redundant    0
dtype: int64
```

Berikut adalah fitur dengan outlier pada dataset:

	Amount of Outli...
companies_count	52
last_year_training_time	238
monthly_income	114
rate_performance	226
time_current_company	104
time_current_manager	14
time_current_role	21
time_last_promotion	107

OUTLIER



2

Data

Preprocessing

OUTLIER HANDLING

Untuk tahap awal yang berlaku untuk semua EDA dan model kami hanya melakukan **penanganan terhadap outlier** dengan mengubah nilai outlier dengan **median** kecuali pada variabel **rate_performance** karena itu adalah data kategorikal. Untuk preprocessing lebih lanjut dilakukan pada tiap pengembangan model.

Berikut adalah fitur dengan outlier setelah diubah menjadi median:

	Column	object	Outlier	int64	percentase	float...
	rate_perfor...	11.1%	0 - 226		0.0 - 0.1537414965...	
	companies_...	11.1%				
	7 others	77.8%				
0	rate_performance			226	0.15374149659863	945
1	companies_count			0		0.0
2	last_year_training_time			0		0.0
3	monthly_income			0		0.0
4	time_current_company			0		0.0
5	time_current_manager			0		0.0
6	time_current_role			0		0.0
7	time_last_promotion			0		0.0
8	time_total_working			0		0.0

OUTLIER



Co.

3

Exploratory Data Analysis (EDA)

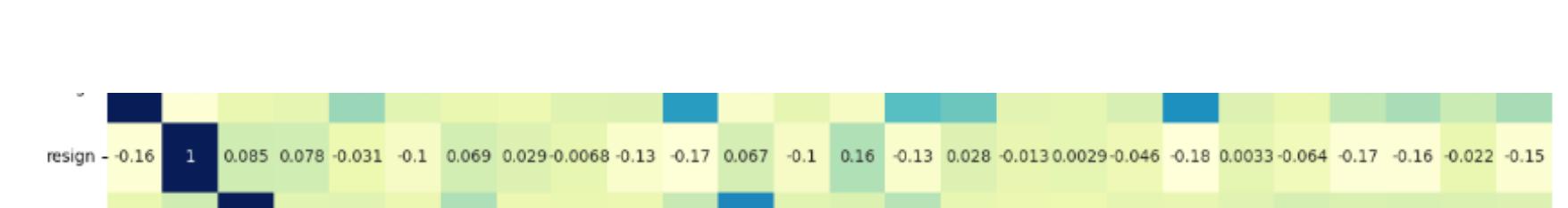
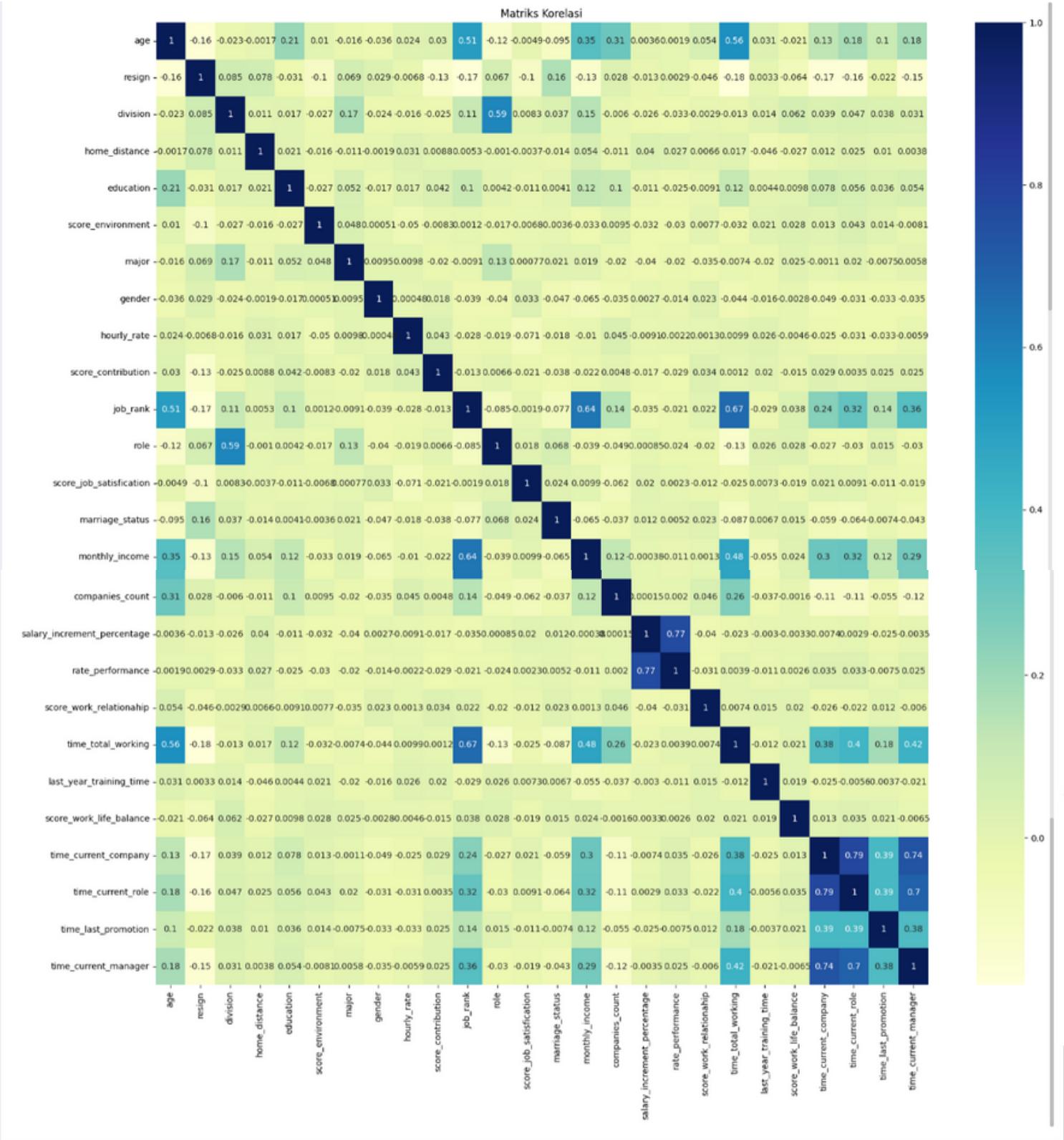
OUTLINE EDA

- A** Karakteristik karyawan yang resign?
- B** Karyawan resign setelah mendapat promosi?
- C** Departemen yang memiliki karyawan loyal terbanyak?
- D** Analisis korelasi antar atribut?
- E** Karyawan dengan edukasi lebih tinggi akan resign?
- F** Karyawan dengan tingkat edukasi yang mana yang akan resign?
- G** Karyawan yang mana yang sering berpindah perusahaan?
- H** Tingkat kepuasan dan work-life-balance karyawan dengan keputusan resign?



**VISUALISASIKAN KARAKTERISTIK
KARYAWAN YANG RESIGN DARI
PERUSAHAAN TERSEBUT!**

ANALISIS KORELASI



Pertama-tama kami melakukan pemetaan nilai korelasi antara fitur-fitur dengan target "resign". Berdasarkan pemetaan ini kami tidak menemukan adanya nilai korelasi yang signifikan (>0.5).

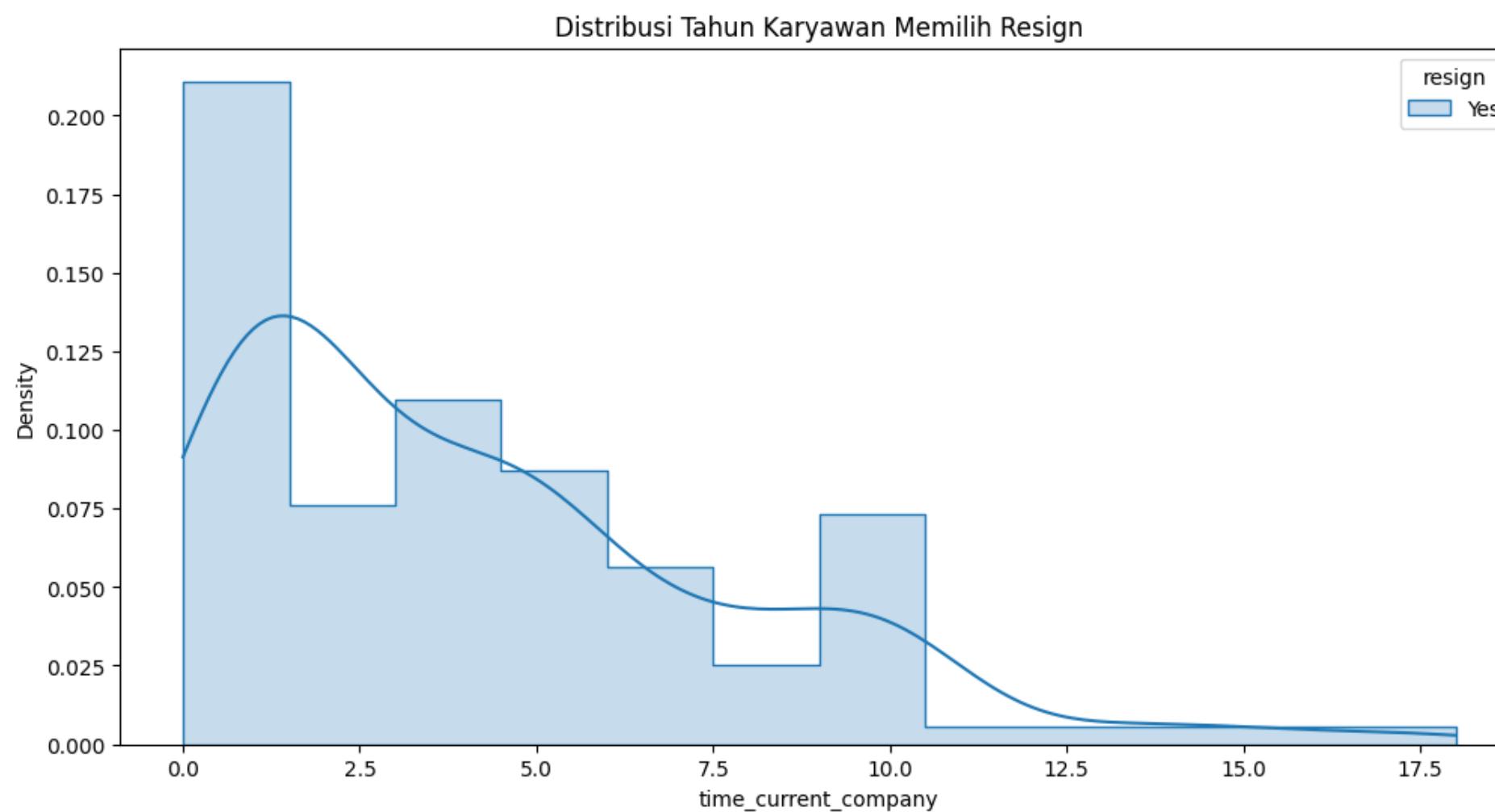
Oleh karena itu, eksplorasi karakteristik dilakukan dengan memilih fitur-fitur yang berdasarkan hipotesis kami akan berpengaruh terhadap resignation.

BERAPA BANYAK KARYAWAN RESIGN DI PERUSAHAAN?



Terdapat sekitar **16,12% karyawan perusahaan yang melakukan resign (237 karyawan)** sementara yang tidak resign berjumlah 1233 yang menunjukkan hal yang baik karena **Jumlah karyawan yang bertahan di perusahaan jauh lebih banyak.**

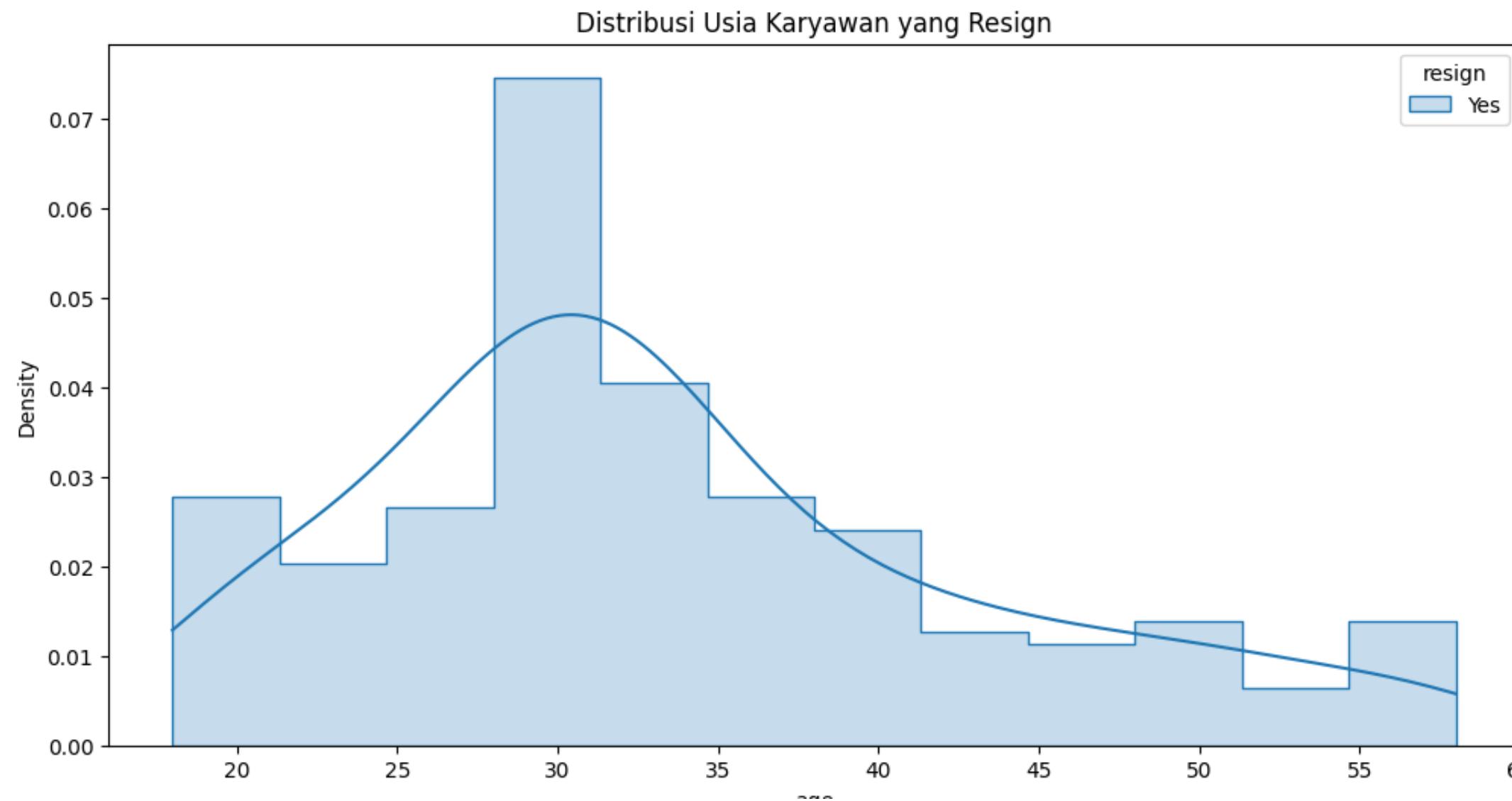
KAPAN KARYAWAN MEMUTUSKAN UNTUK RESIGN?



Terdapat **tiga tahun kritis** dimana kebanyakan karyawan memutuskan untuk resign:

- **1 - 2 tahun pertama**, pada masa ini karyawan paling banyak memutuskan untuk resign
- Tren resign kembali terjadi pada **tahun ke-5 dan ke-10** masa kerja karyawan di perusahaan, di mana pada tahun tersebut jumlah karyawan yang mengundurkan diri meningkat tajam setelah sebelumnya landai.

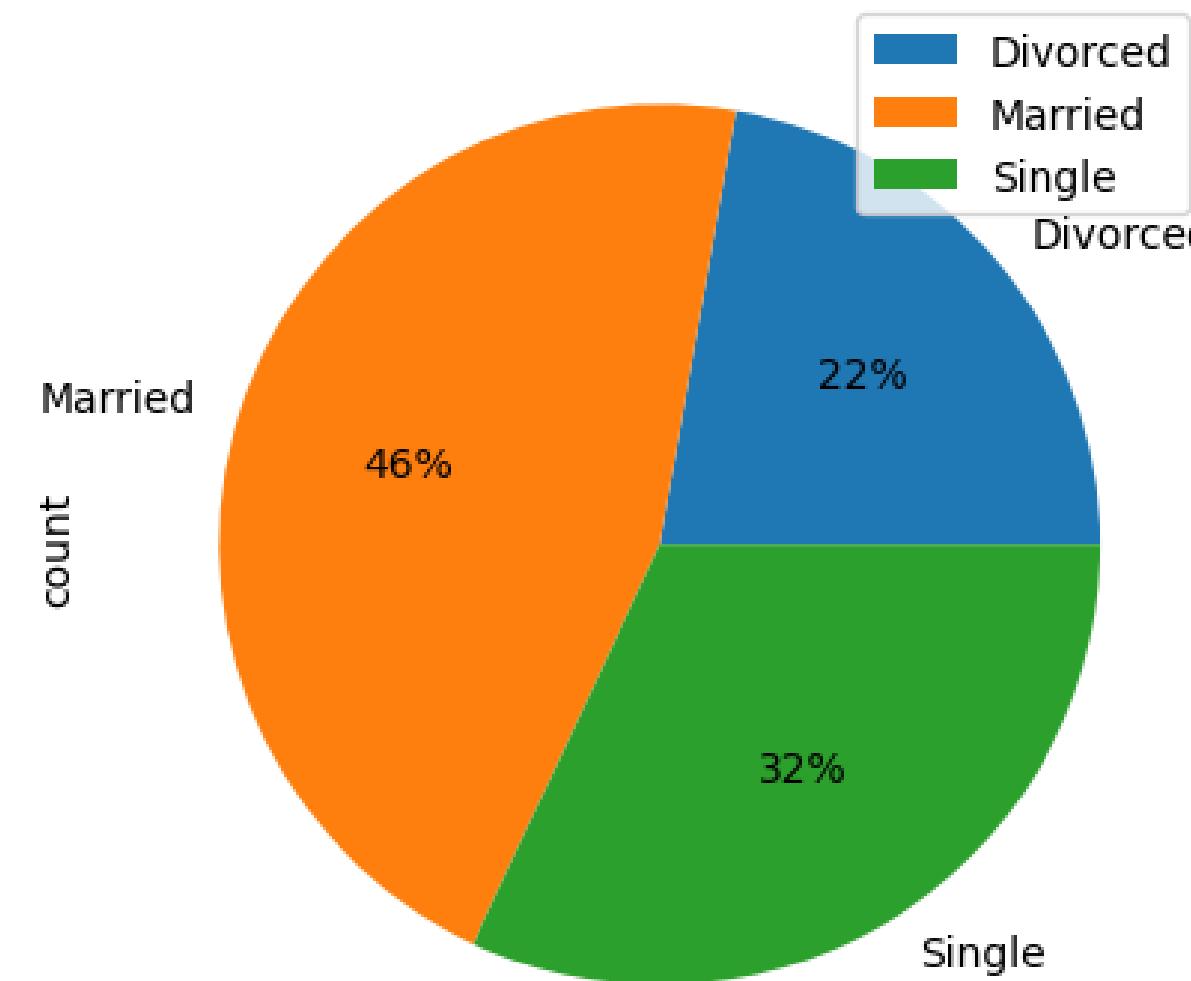
DISTRIBUSI USIA KARYAWAN RESIGN



Karyawan yang resign didominasi oleh karyawan yang berusia muda (sekitar 20 hingga pertengahan 30).

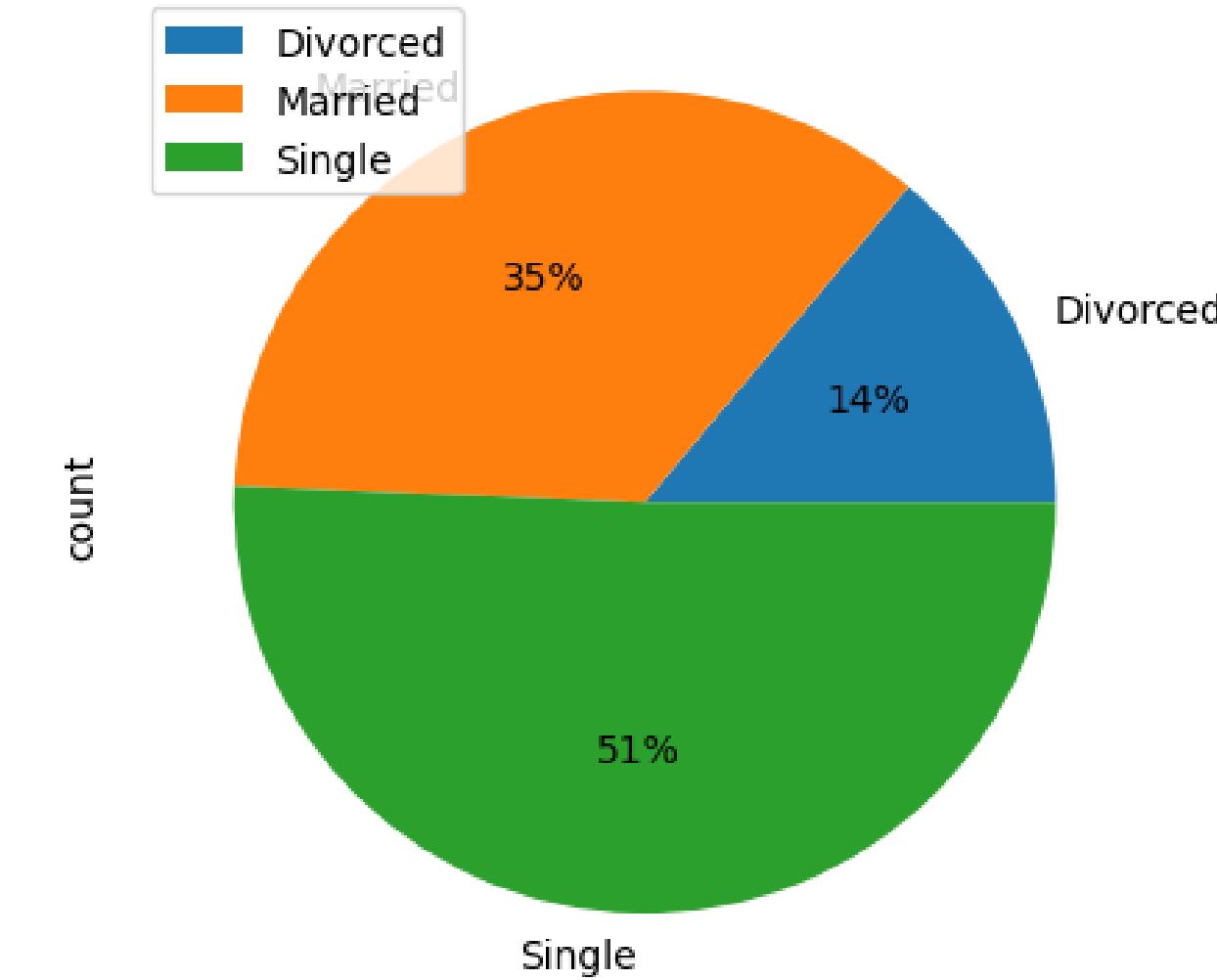
KETERKAITAN STATUS PERNIKAHAN TERHADAP RESIGNATION

Pie Chart Status Pernikahan Karyawan



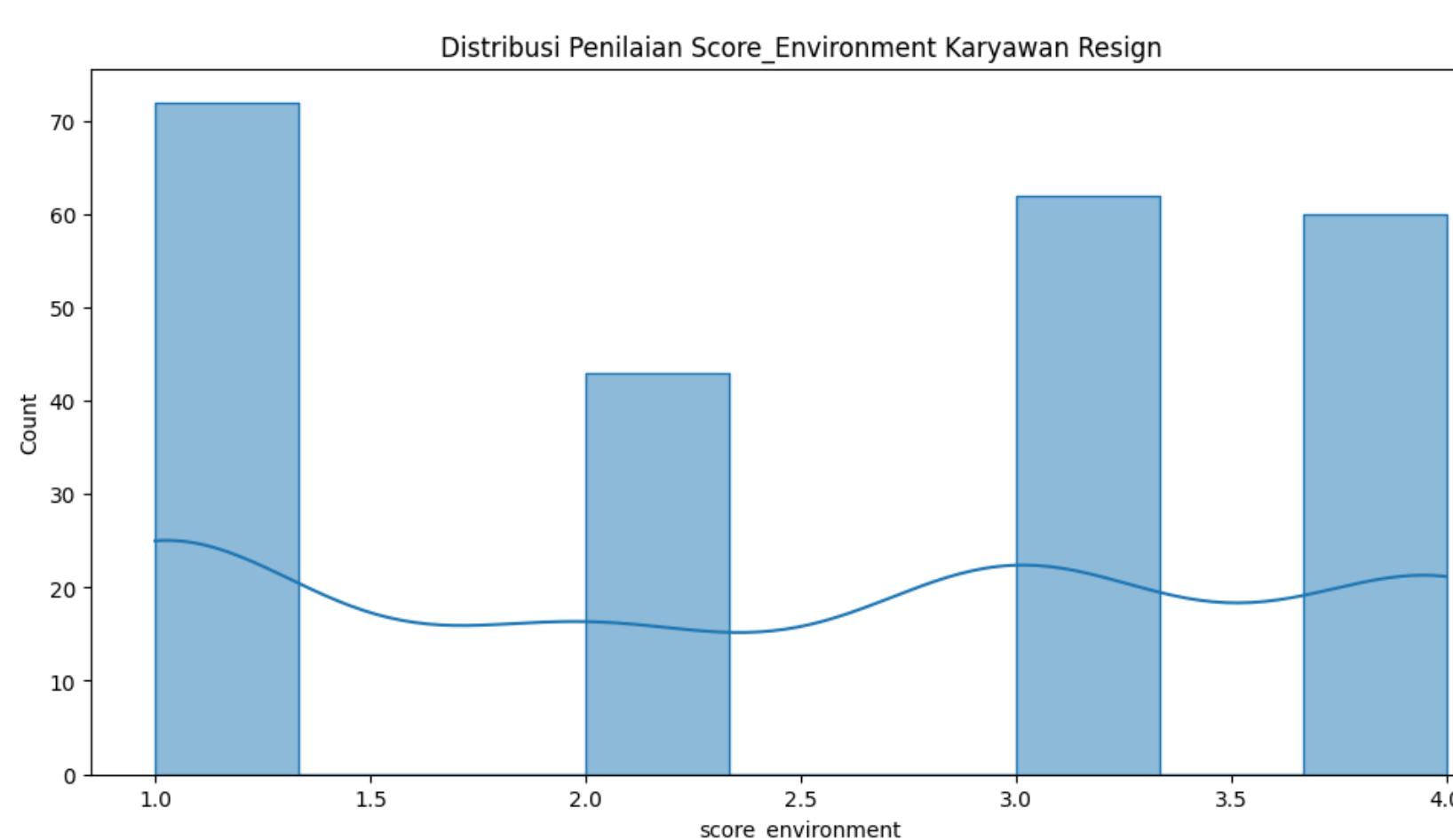
88% karyawan perusahaan sudah pernah menikah (46% masih berstatus menikah dan 22% berstatus divorce). Artinya mayoritas karyawan memiliki tanggung jawab finansial yang besar karena harus mencukupi kebutuhan finansial keluarga.

Pie Chart Status Pernikahan Karyawan yang Resign

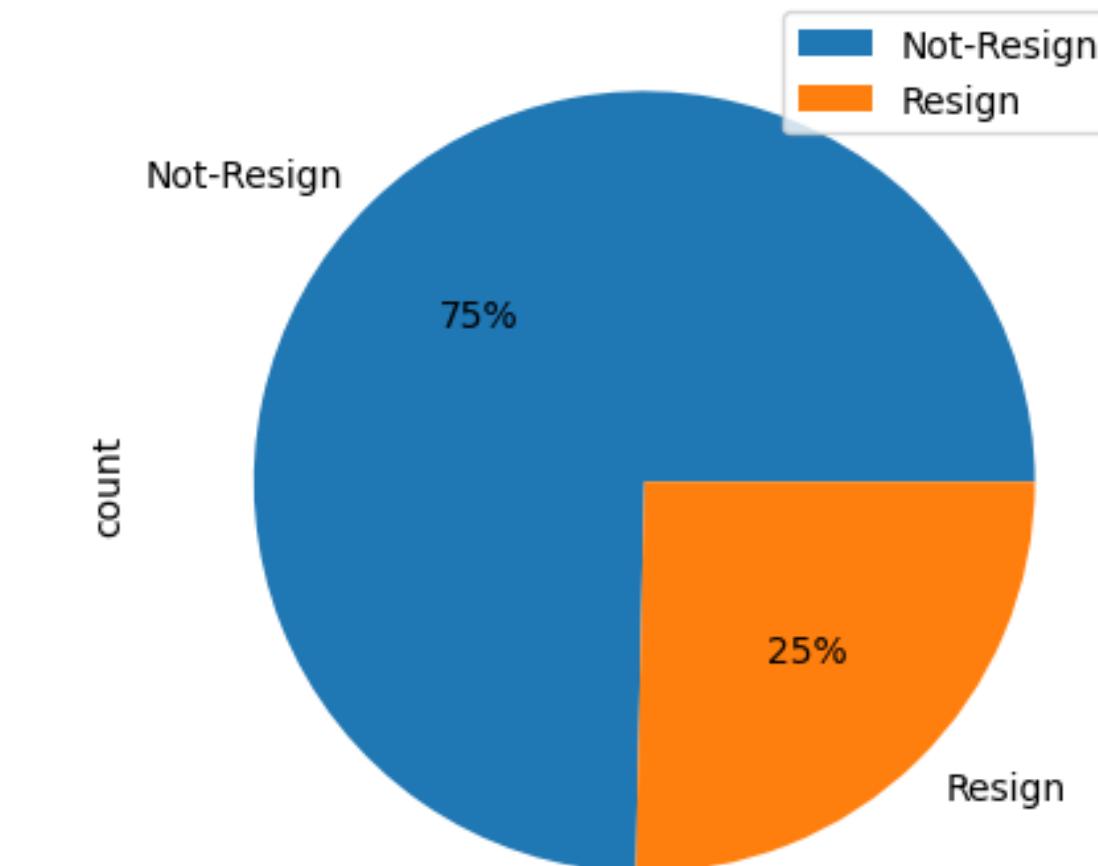


Mayoritas karyawan yang melakukan resign (51%) berstatus single, dimana kita ketahui karyawan yang berstatus single relatif memiliki kebutuhan finansial yang lebih rendah dibandingkan yang sudah menikah

KETERKAITAN LINGKUNGAN KERJA TERHADAP RESIGNATION



Pie Chart Distribusi Karyawan dengan score_environment terendah (1)

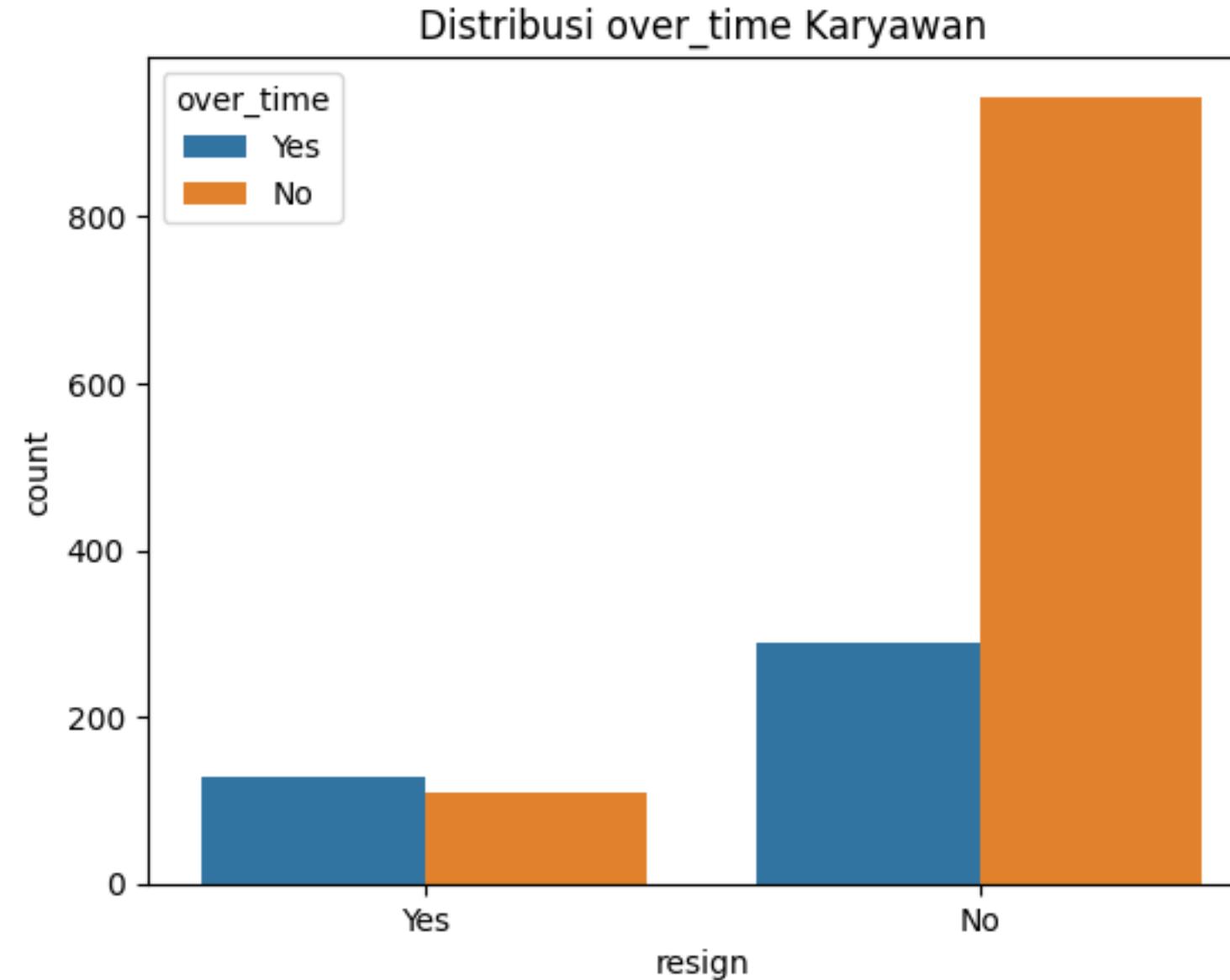


110 karyawan yang resign (hampir 50%) menilai lingkungan kerja dengan sangat baik (skala 3-4).

Di sisi lain, karyawan yang menilai lingkungan kerja buruk (nilai = 1) juga hanya 25% yang memilih untuk keluar sementara yang lainnya tetap di perusahaan.

Kesimpulan: Penilaian terhadap lingkungan kerja tidak menjadi karakteristik khusus yang dapat membedakan karyawan resign / tidak resign

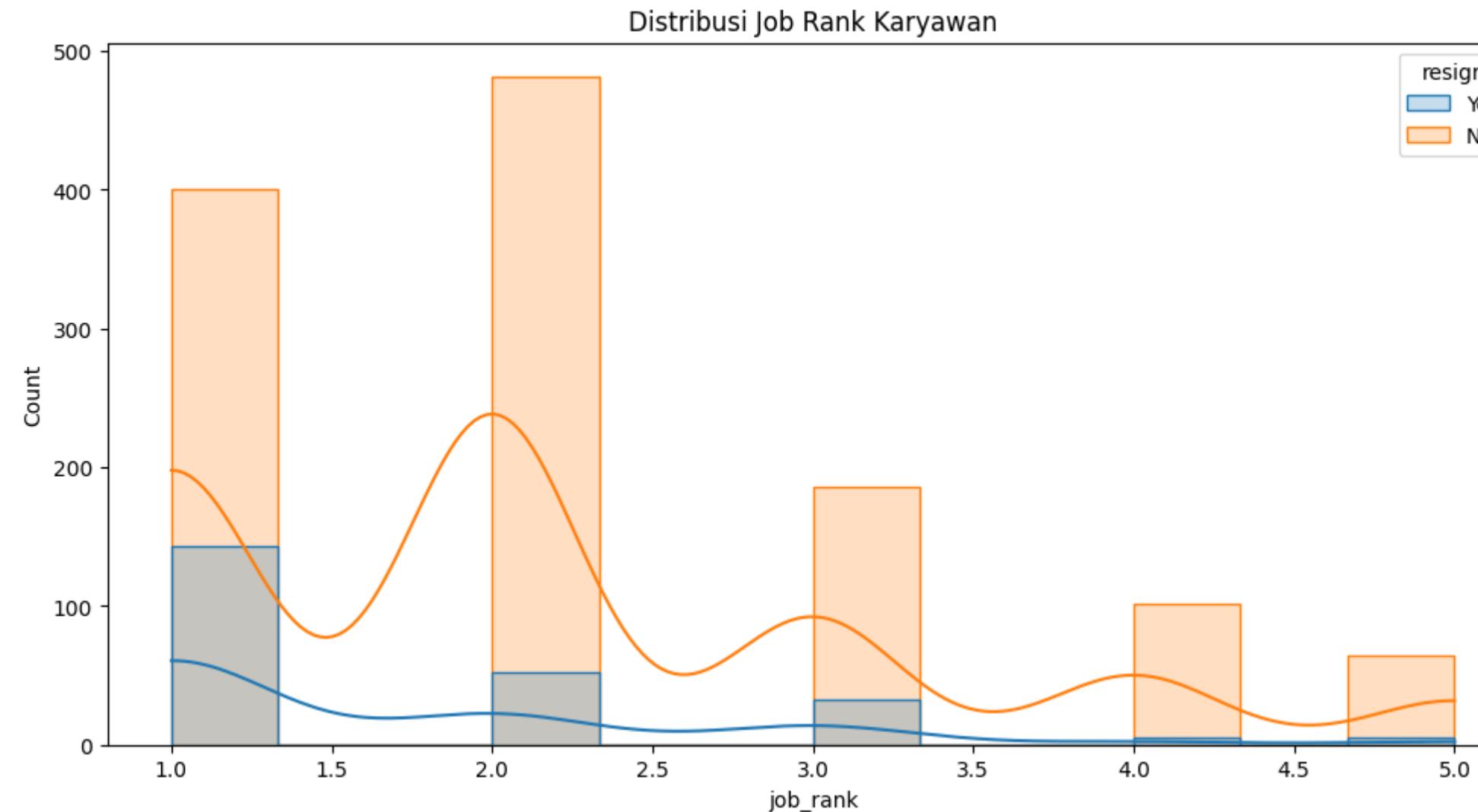
KETERKAITAN OVERTIME TERHADAP RESIGNATION



Meskipun pada karyawan yang resign lebih banyak yang bekerja overtime, namun selisih ini hanya sedikit dan **karyawan yang bekerja overtime lebih banyak yang tidak memilih resign.**

Oleh karena itu, kami menilai overtime tidak menjadi karakteristik pada karyawan yang resign.

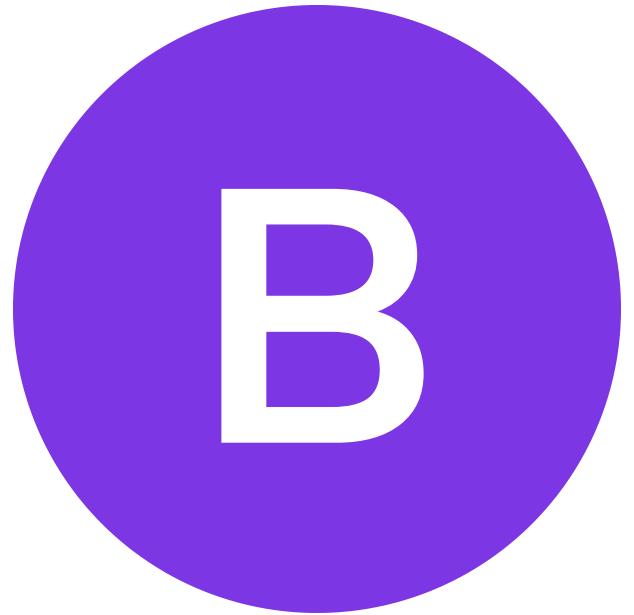
KETERKAITAN JOB RANK TERHADAP RESIGNATION



- Mayoritas karyawan memiliki job rank 1 atau 2
- **Karyawan yang resign paling banyak memiliki job rank 1**
- Persentase karyawan yang resign pada job rank yang tinggi (4 dan 5) relatif sangat kecil

KESIMPULAN

- Karyawan paling banyak resign pada tahun 1-2 bekerja di perusahaan, tren kemudian kembali meningkat pada tahun ke-5 dan ke-10
- Karyawan yang resign umumnya berusia muda (di bawah 20 tahun hingga pertengahan 30 tahun)
- Karyawan yang resign mayoritas memiliki status pernikahan single
- Menariknya, penilaian karyawan terhadap lingkungan kerja dan waktu bekerja yang overtime relatif tidak berkaitan dengan keputusan karyawan untuk mengundurkan diri.
- Kebanyakan karyawan resign memiliki job rank 1



**APAKAH KARYAWAN MEMILIH
UNTUK RESIGN SETELAH
MENDAPATKAN PROMOSI?**

APAKAH KARYAWAN MEMILIH UNTUK RESIGN SETELAH MENDAPATKAN PROMOSI?

Data yang ada utamanya yang menunjukan waktu menggunakan satuan tahun, sehingga ada beberapa asumsi yang kami tambahkan, yaitu:

- Saat masih entry level berarti karyawan belum pernah mendapatkan promosi
- Interpretasi dari kalimat "Setelah mendapatkan promosi" adalah promosi yang dilakukan 1 tahun ke belakang

TEMUAN

- Dari 1420 data, data non-entry level terdapat 972 data

```
use_for_b = df[["resign","job_rank","time_last_promotion"]]
```

```
use_for_b["promotion_as_new"] = np.where(use_for_b["job_rank"] == 1, 1, 0)
```

```
/tmp/ipykernel_88/936585172.py:1: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row_indexer,col_indexer] = value instead  
  
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus  
use_for_b["promotion_as_new"] = np.where(use_for_b["job_rank"] == 1, 1, 0)
```

```
filter_b = use_for_b[use_for_b["promotion_as_new"] == 0]  
filter_b = filter_b.drop(["job_rank", "promotion_as_new"], axis=1)  
filter_b.count()
```

resign	927
time_last_promotion	927
dtype:	int64

TEMUAN

- Dari 927 data non-entry level terdapat 321 data yang waktu promosi terakhirnya kurang dari 1 tahun

```
last_promotion_1_year = filter_b[filter_b["time_last_promotion"] < 1]
```

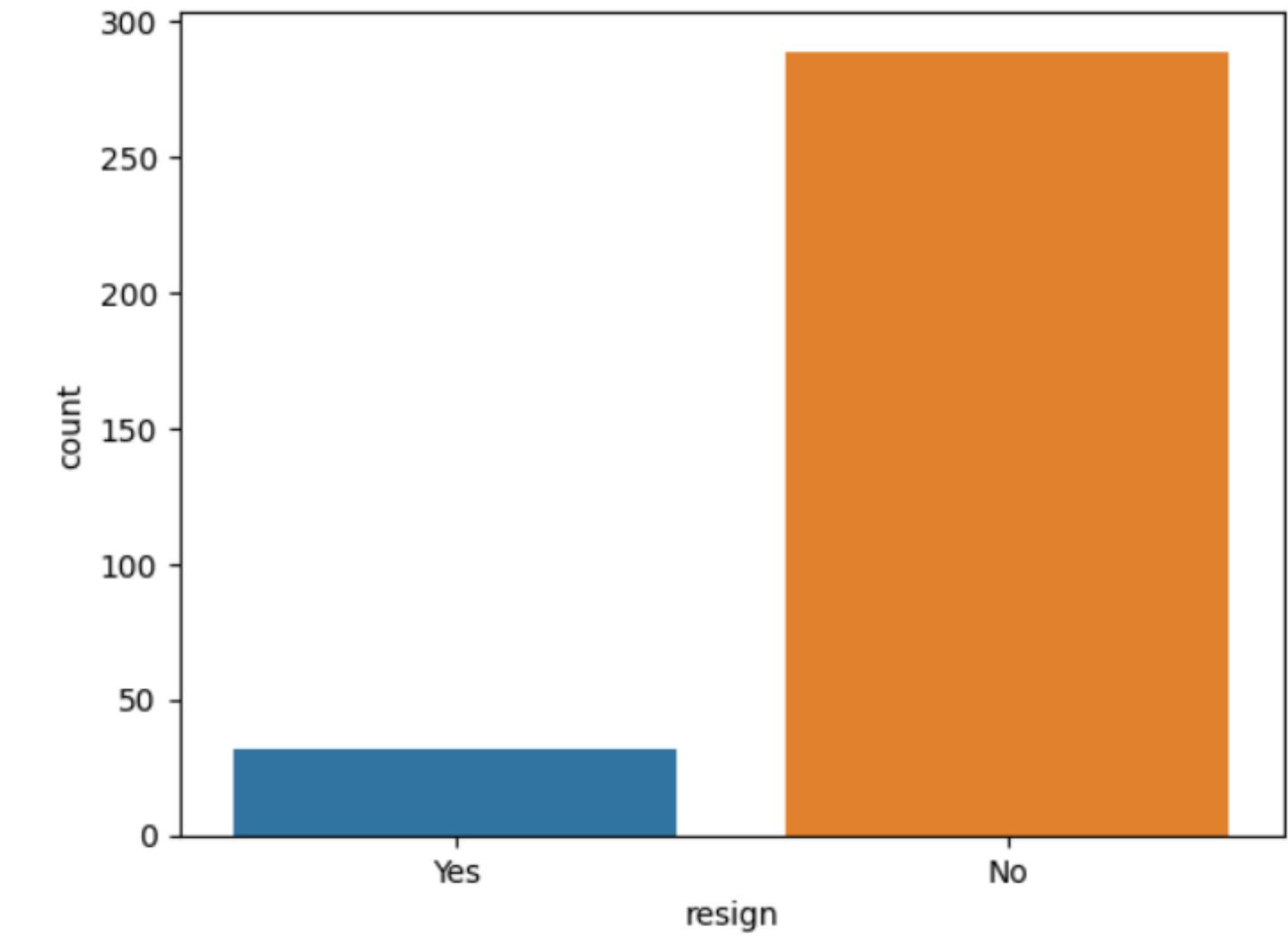
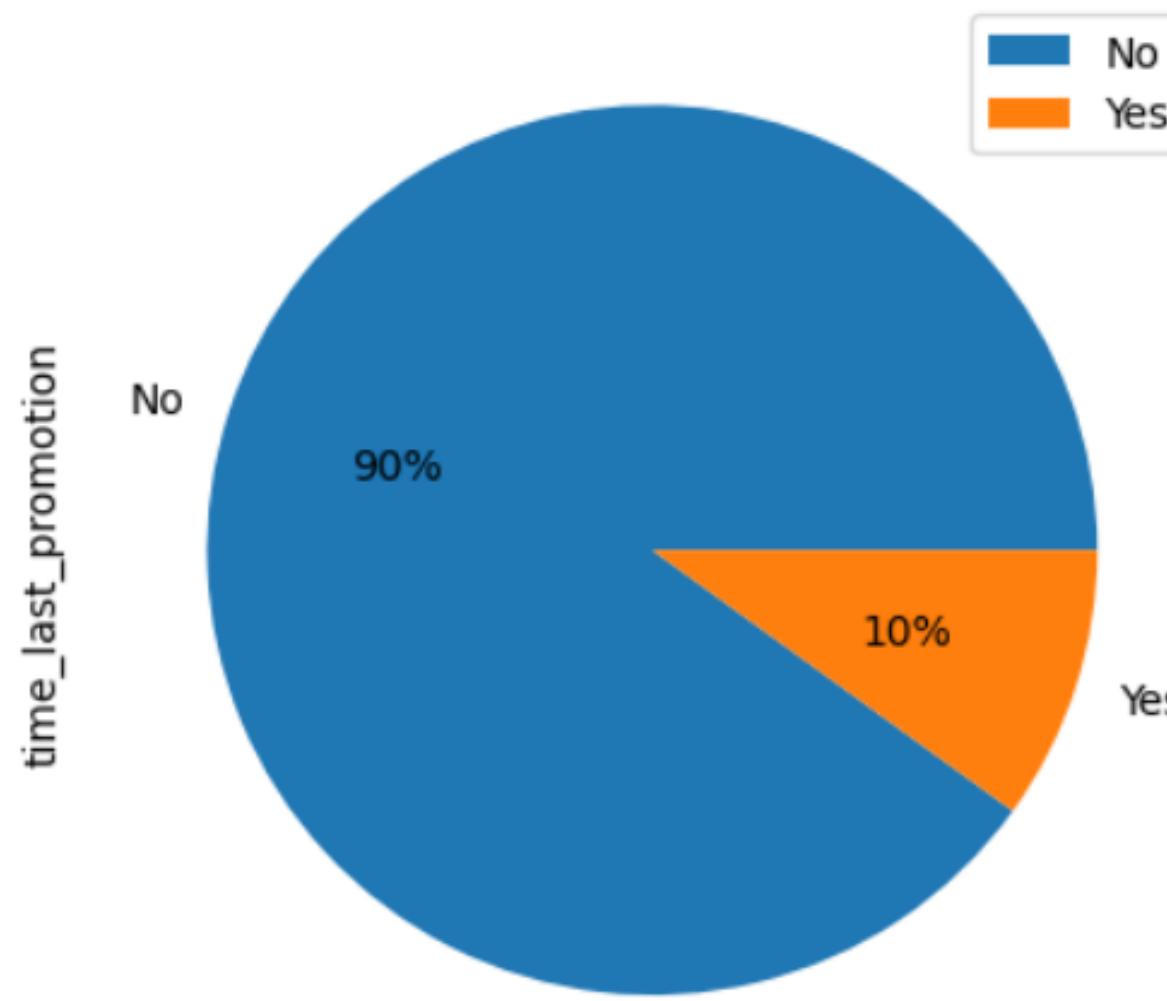


```
last_promotion_1_year.count()
```



```
resign          321  
time_last_promotion 321  
dtype: int64
```

**HASIL EKSPLORASI INI DITEMUKAN BAHWA DARI 321 YANG
MEMENUHI KRITERIA 289 DIANTARANYA MEMILIH RESIGN
DAN 32 SISANYA MEMILIH UNTUK TIDAK RESIGN**



ANALISIS DAN INTERPRETASI

- Jumlah resign jauh lebih kecil ketimbang karyawan yang tidak resign (1:9).
- Karyawan yang baru mendapatkan promosi cenderung untuk tidak resign.
- Promosi karyawan mempengaruhi intensi karyawan untuk keluar dari perusahaan.



**DEPARTEMEN MANAKAH YANG
MEMILIKI KARYAWAN LOYAL
PALING BANYAK?**

DEPARTEMEN MANAKAH YANG MEMILIKI KARYAWAN LOYAL PALING BANYAK?

Karyawan loyal yang dimaksud berdasarkan asumsi yang kami ambil adalah:

- Karyawan sudah bekerja pada perusahaan tersebut 10 tahun atau lebih
- Angka 10 tahun diambil dari grafik densitas yang sudah ada sebelumnya, pada tahun ke 10 menjadi daerah yang curam terakhir sebelum pada tahun berikutnya grafiknya menjadi landai. ([Page 17: Kapan karyawan memutuskan untuk resign?](#))

TEMUAN

- Dari 1420 data, terdapat 366 karyawan loyal pada perusahaan

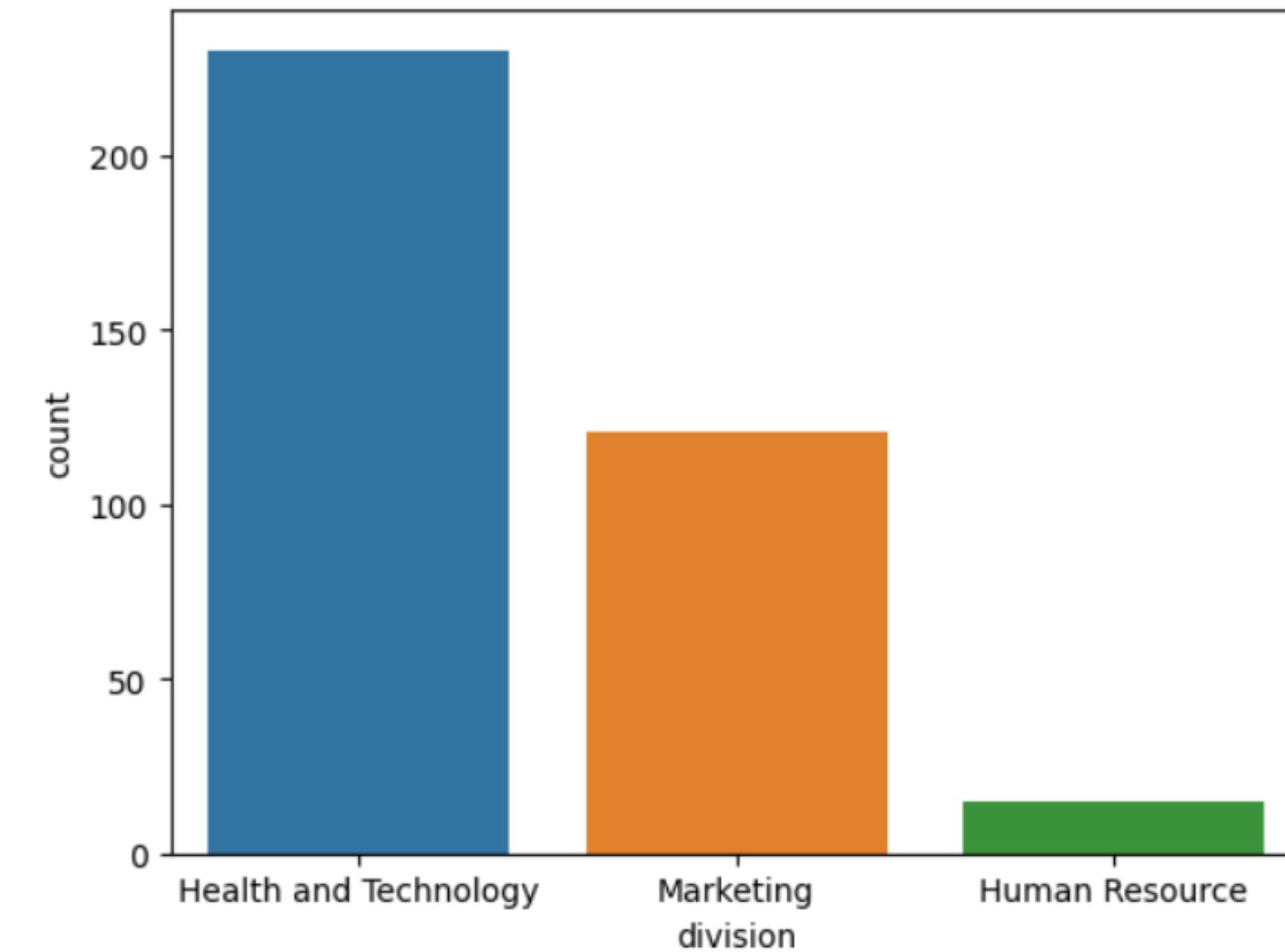
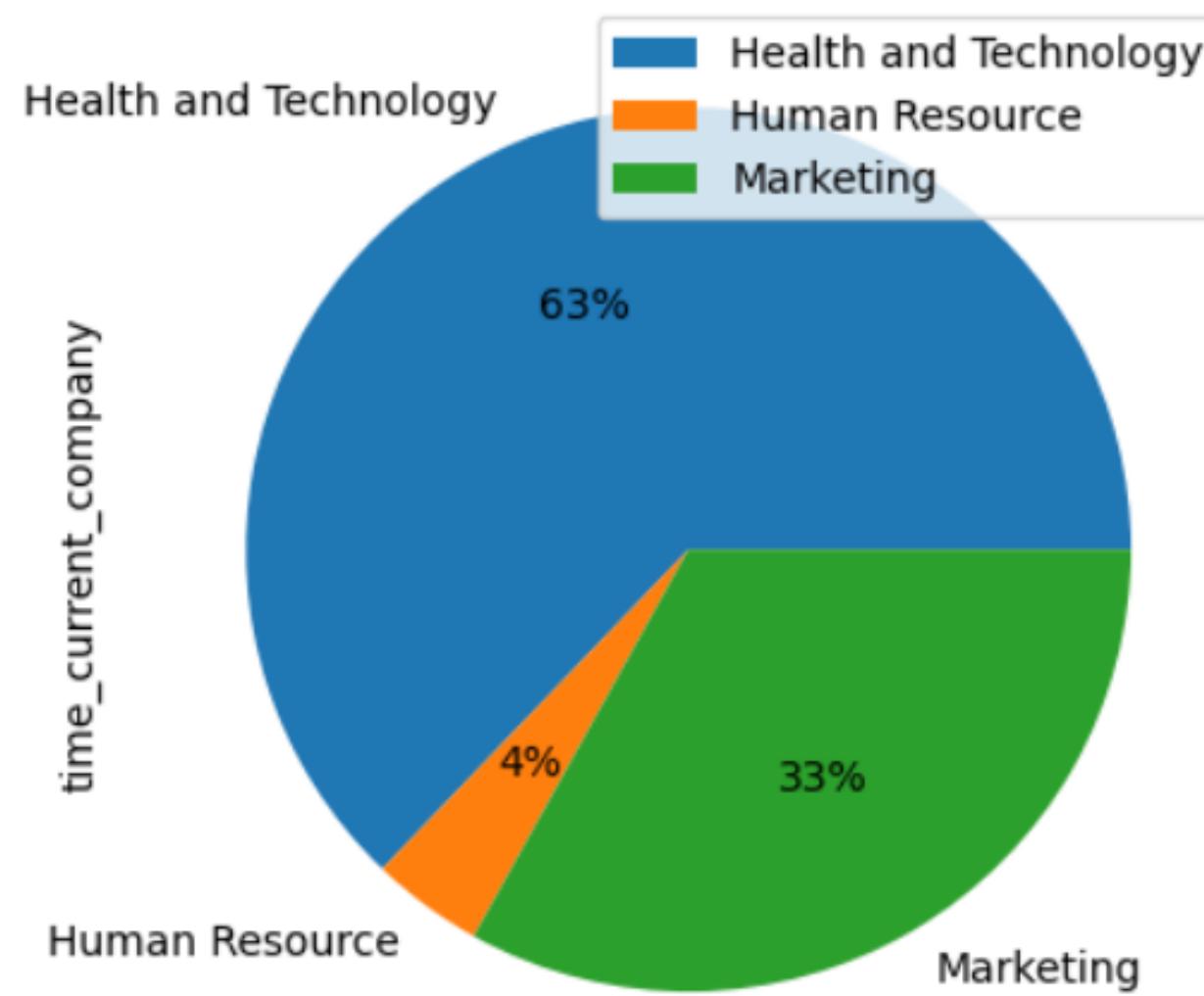
```
use_for_c = df[["division", "time_current_company"]]
```

```
filter_c = use_for_c[use_for_c["time_current_company"] >= 10]
```

```
filter_c.count()
```

```
division          366
time_current_company 366
dtype: int64
```

HASIL EKSPLORASI INI DITEMUKAN BAHWA TERDAPAT 366 KARYAWAN YANG BEKERJA SELAMA 10 TAHUN ATAU LEBIH, 230 DIANTARANYA ADALAH DIVISI HEALTH & TECHNOLOGY, 15 KARYAWAN HUMAN RESOURCE, DAN 121 KARYAWAN MARKETING.



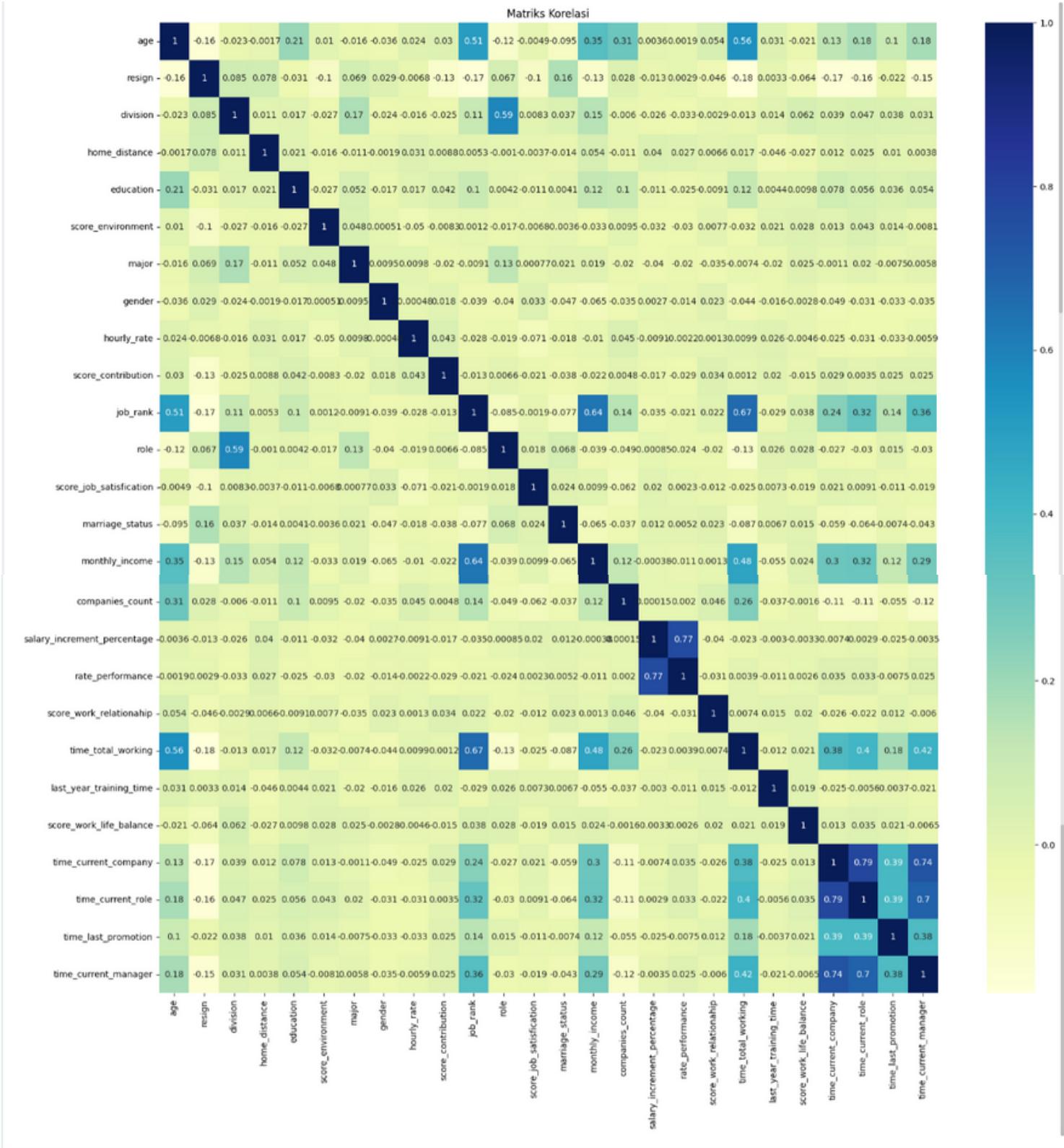
ANALISIS DAN INTERPRETASI

- Dengan asumsi yang kami berikan karyawan yang loyal pada perusahaan itu ada 366/1420 atau hanya sekitar 25% dari keseluruhan karyawan.
- Dari karyawan loyal tersebut didapatkan bahwa divisi yang paling banyak karyawan loyalnya adalah divisi Health and Technology, dengan 230 karyawan.



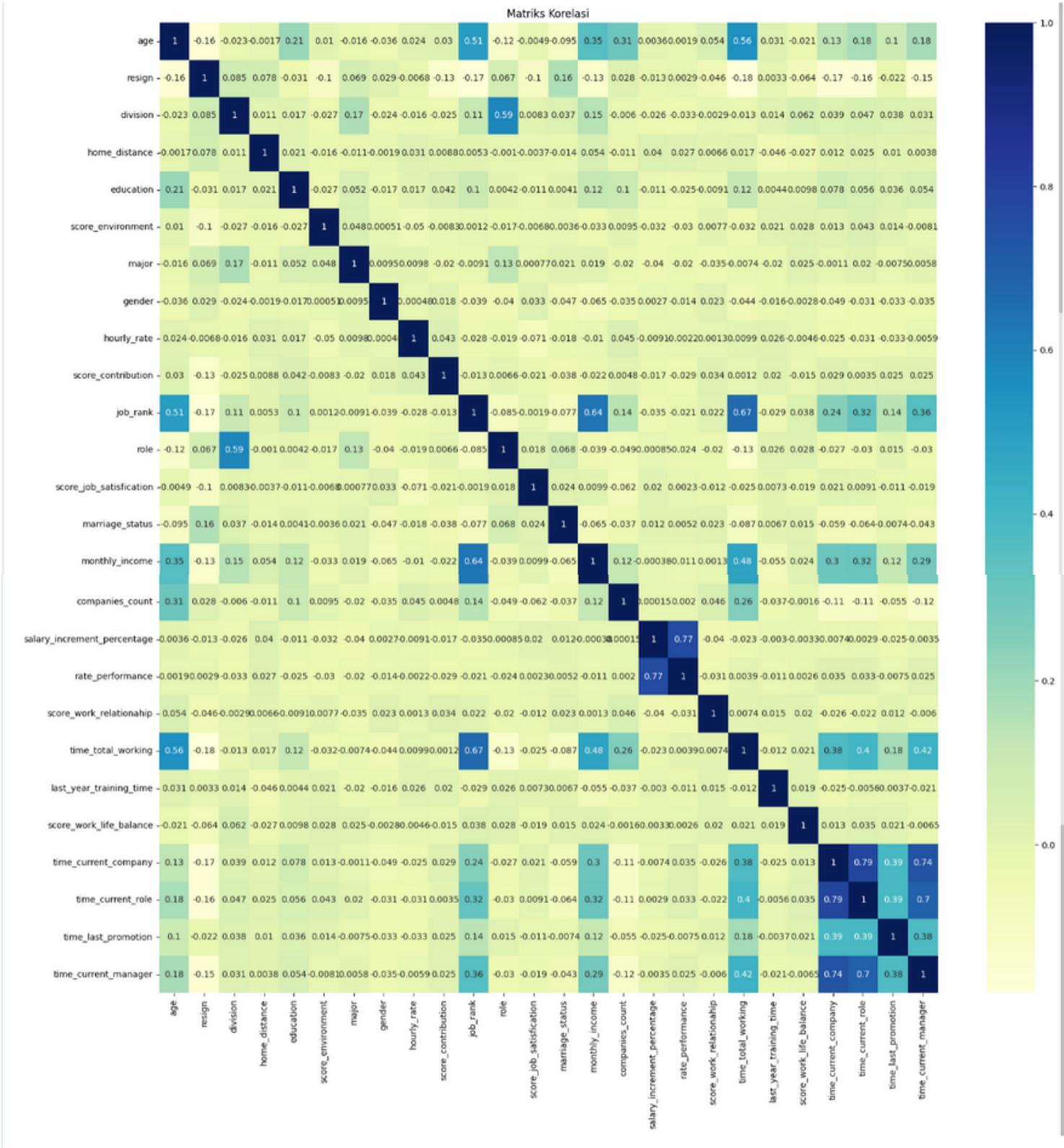
LAKUKAN ANALISIS KORELASI
ANTAR ATRIBUT, VISUALISASIKAN
ATRIBUT-ATRIBUT YANG MEMILIKI
KORELASI.

ANALISIS KORELASI ANTAR ATRIBUT



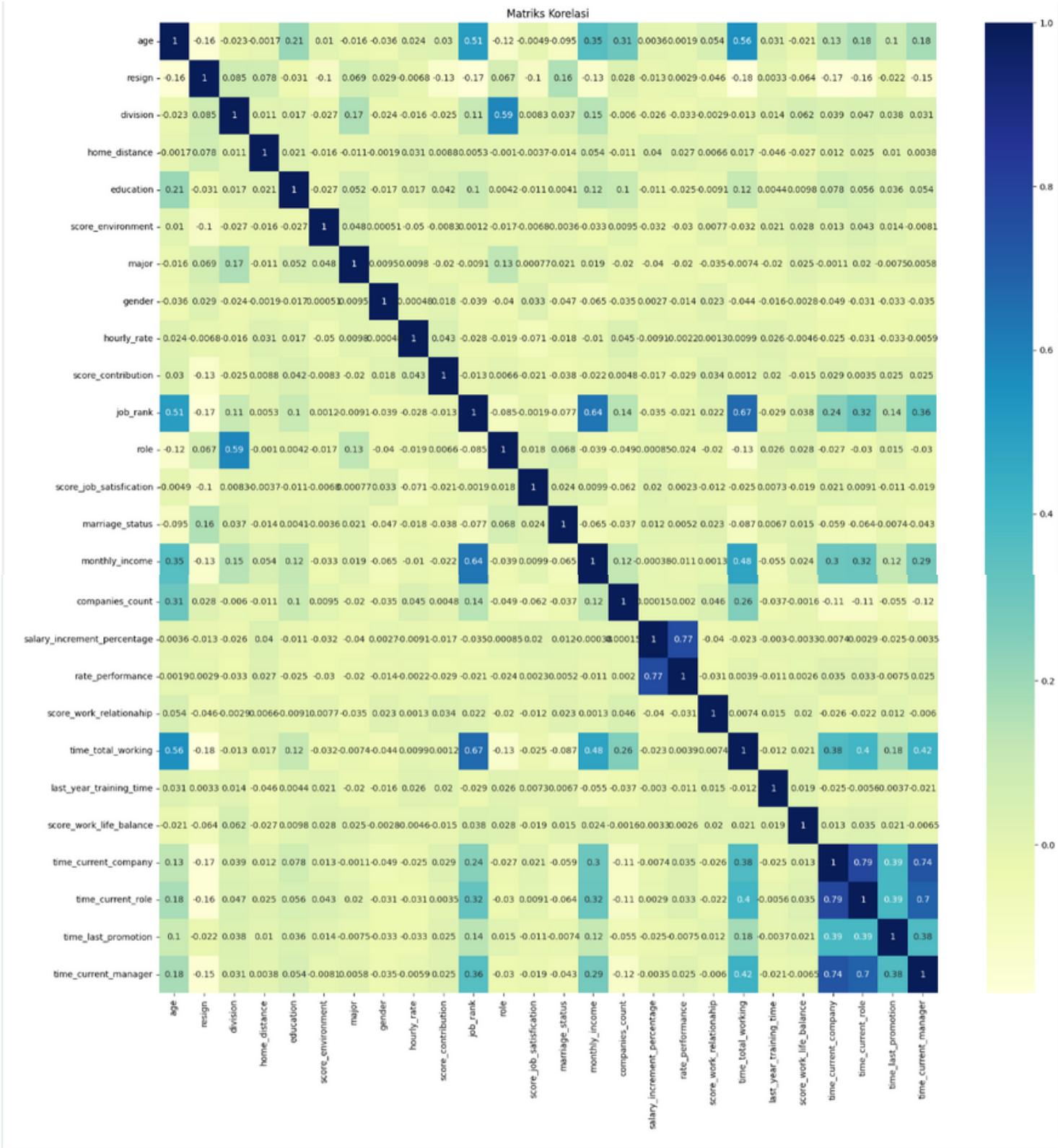
- **time_total_working** dan **job_rank** memiliki korelasi yang tinggi (0.67) hal ini sesuai dengan konteks nyata yaitu orang yang memiliki pengalaman bekerja banyak (dalam tahun) juga akan memiliki job_rank yang relatif tinggi.
- **job_rank** juga memiliki korelasi yang besar dengan **monthly_income** (0.64), dalam konteks di perusahaan ini maka kita dapat menilai apabila seseorang memiliki job_rank yang tinggi maka pendapatannya per bulan juga akan besar.

ANALISIS KORELASI ANTAR ATRIBUT

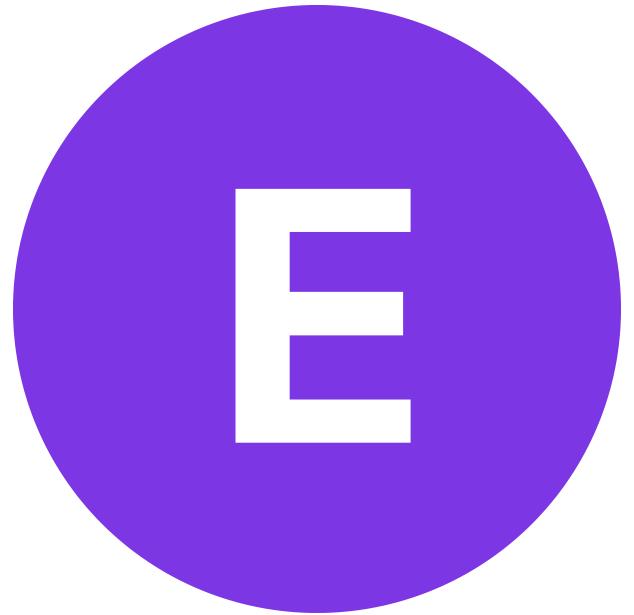


- **time_current_company**, **time_current_role**, dan **time_current_manager** memiliki nilai korelasi yang sama-sama besar antar atributnya : **time_current_company & timecurrent_role (0.79)**, **time_current_company & time_current_manager (0.74)**, **time_current_role & time_current_manager (0.7)**. Hal ini dapat menunjukkan salah satunya: orang-orang di dalam perusahaan yang sudah lama menekuni pekerjaan tersebut (role) relatif sudah lama bekerja juga di perusahaan, juga dapat menunjukkan bahwa orang-orang yang bekerja lama dengan manager saat ini relatif loyal karena sudah bekerja lama juga dengan perusahaan.

ANALISIS KORELASI ANTAR ATRIBUT



- **salary_increment_percentage** dan **rate_performance** juga memiliki nilai korelasi yang besar (**0.77**) yang dapat menunjukkan peningkatan persentase gaji akan sebanding dengan performance dalam pekerjaan.
- Nilai **hourly_rate** tidak berkorelasi dengan **monthly_income** (-**0.01**) yang menurut kelompok kami cukup membingungkan karena artinya terdapat sejumlah pendapatan lain yang didapat di luar dari pendapatan bekerja (per jam). Namun karena nilai **monthly_income** memiliki korelasi yang besar terhadap **job_rank** maka ke depannya kami akan memilih untuk menggunakan nilai **monthly_income** dibandingkan **hourly_rate**.



**APAKAH KARYAWAN DENGAN
EDUKASI LEBIH TINGGI LEBIH
MUNGKIN UNTUK RESIGN?**

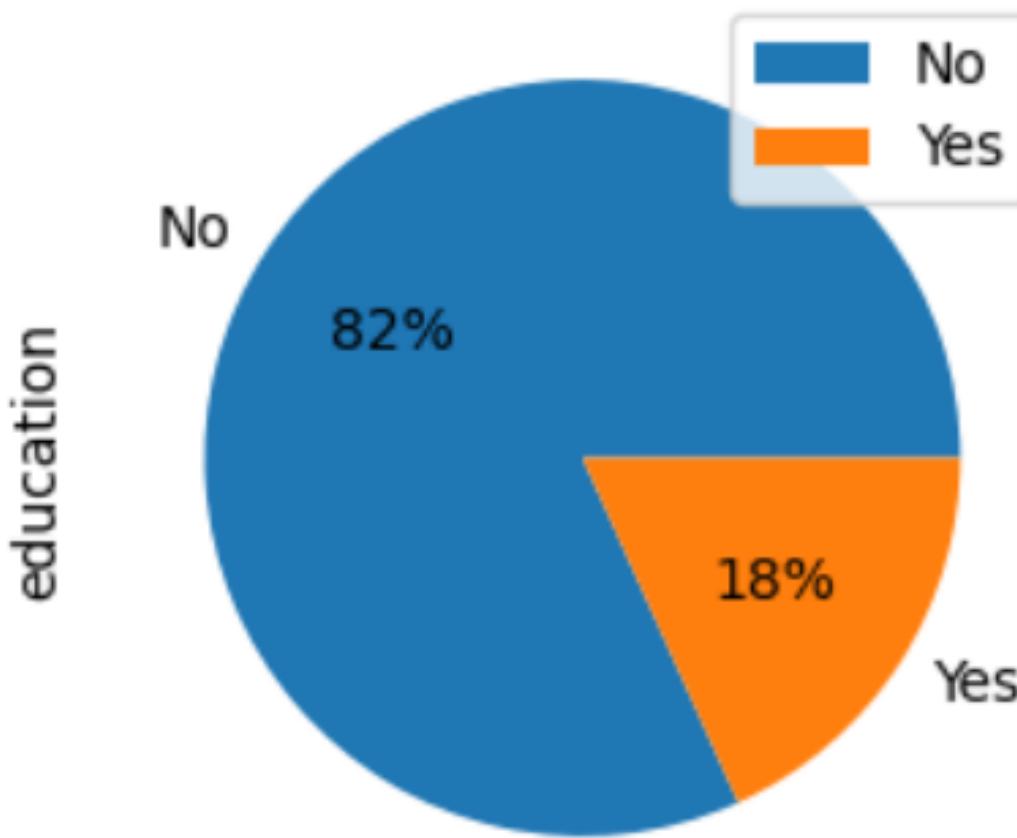
APAKAH KARYAWAN DENGAN EDUKASI LEBIH TINGGI LEBIH MUNGKIN UNTUK RESIGN?

Beberapa asumsi yang kami tambahkan dalam menjawab pertanyaan eksplorasi ini:

- Tingkat edukasi yang lebih tinggi disini berarti semakin tinggi pada angkanya ($5>1$)
- Atribut lain seperti major dan umur tidak berkontribusi apapun terhadap tingkat edukasi

TEMUAN

- Kami mencoba melihat probabilitas seorang untuk resign pada setiap tingkatan edukasi, berikut hasilnya:



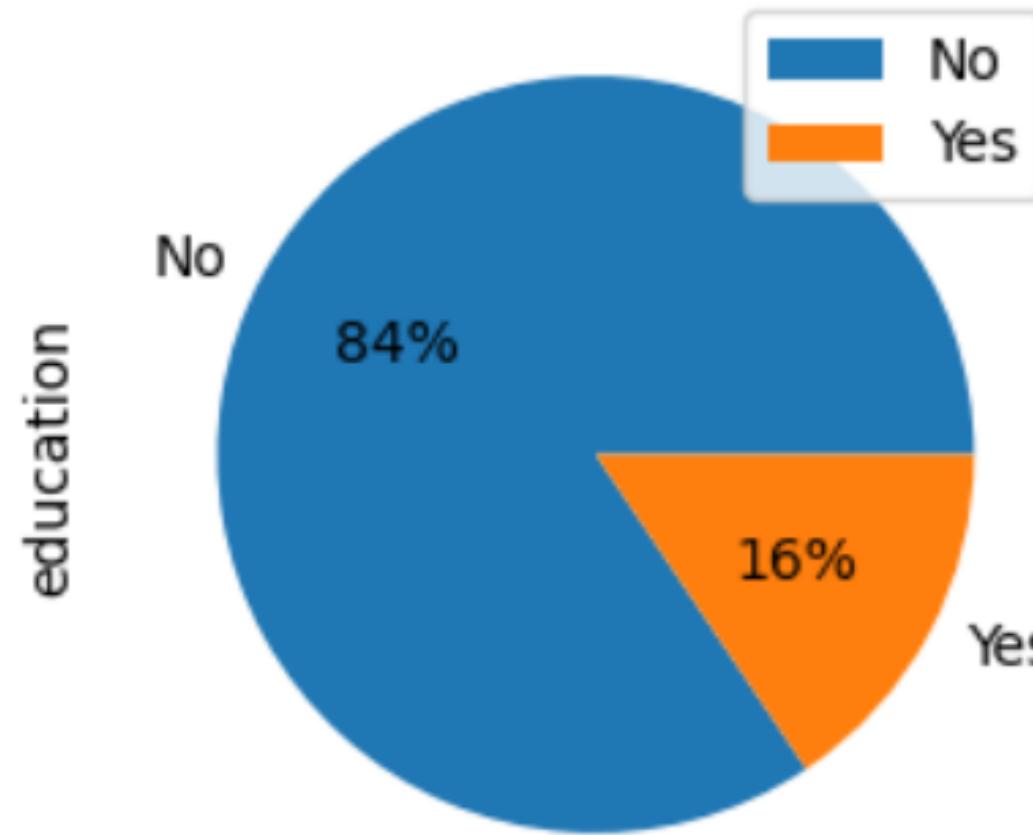
Tingkat Edukasi: 1

Mencakup karyawan dengan pendidikan Diploma 3

Memiliki probabilitas resign
sebesar **0.18**

TEMUAN

- Kami mencoba melihat probabilitas seorang untuk resign pada setiap tingkatan edukasi, berikut hasilnya:



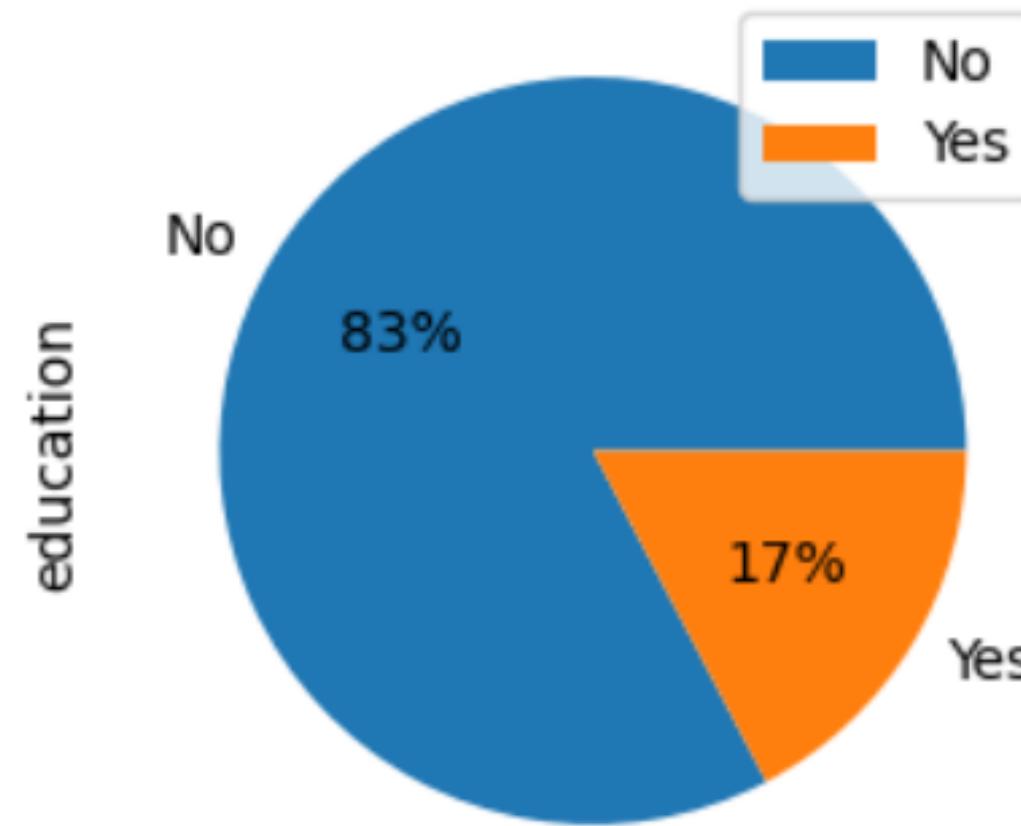
Tingkat Edukasi: 2

Mencakup karyawan dengan pendidikan Diploma 4

Memiliki probabilitas resign
sebesar 0.16

TEMUAN

- Kami mencoba melihat probabilitas seorang untuk resign pada setiap tingkatan edukasi, berikut hasilnya:



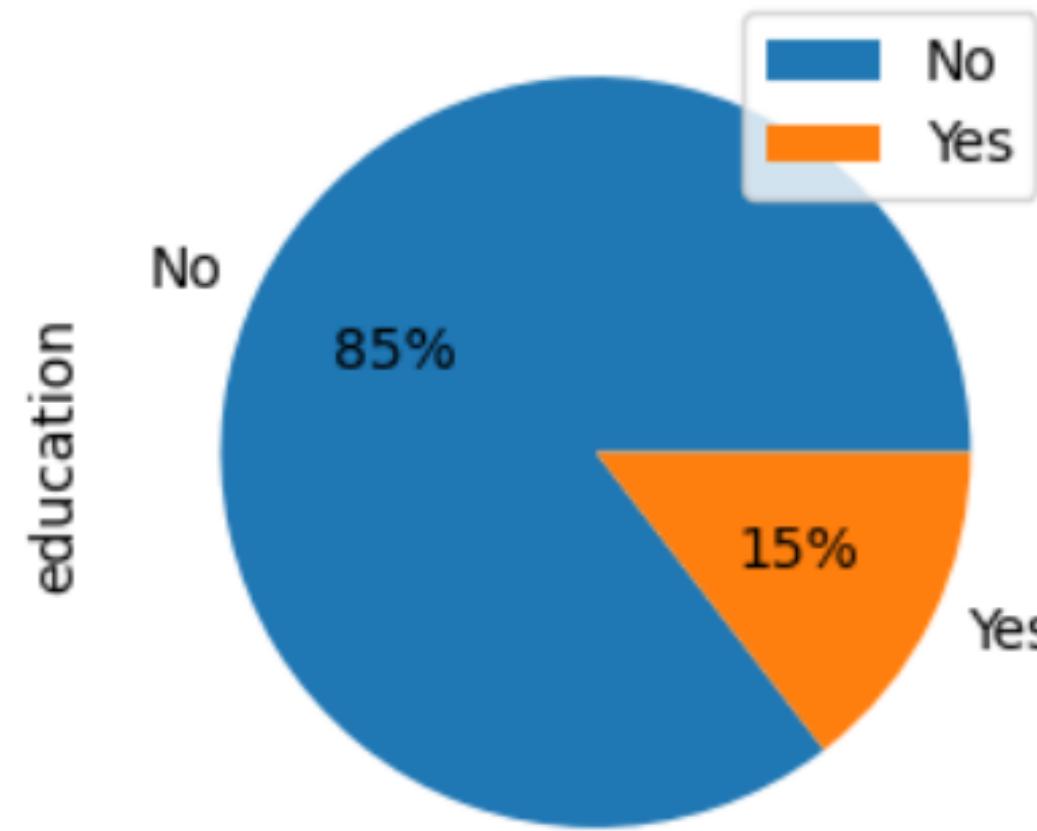
Tingkat Edukasi: 3

Mencakup karyawan dengan pendidikan Sarjana

Memiliki probabilitas resign sebesar 0.17

TEMUAN

- Kami mencoba melihat probabilitas seorang untuk resign pada setiap tingkatan edukasi, berikut hasilnya:



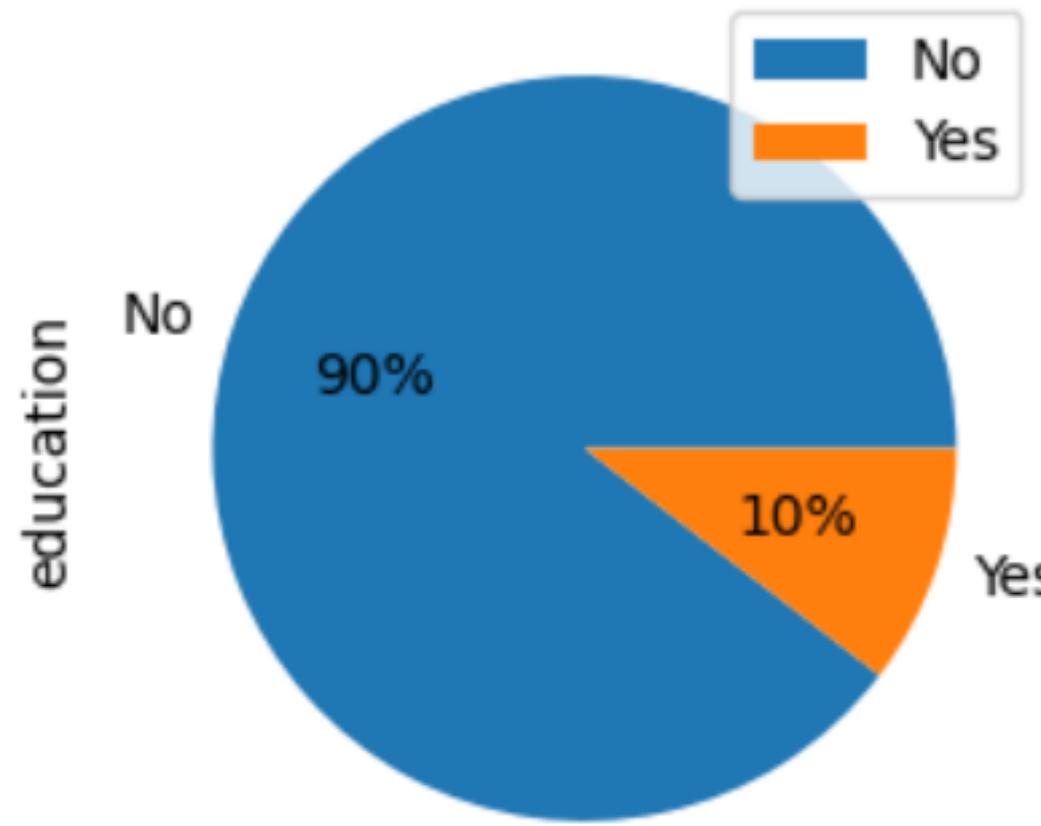
Tingkat Edukasi: 4

Mencakup karyawan dengan pendidikan Magister

Memiliki probabilitas resign sebesar 0.15

TEMUAN

- Kami mencoba melihat probabilitas seorang untuk resign pada setiap tingkatan edukasi, berikut hasilnya:

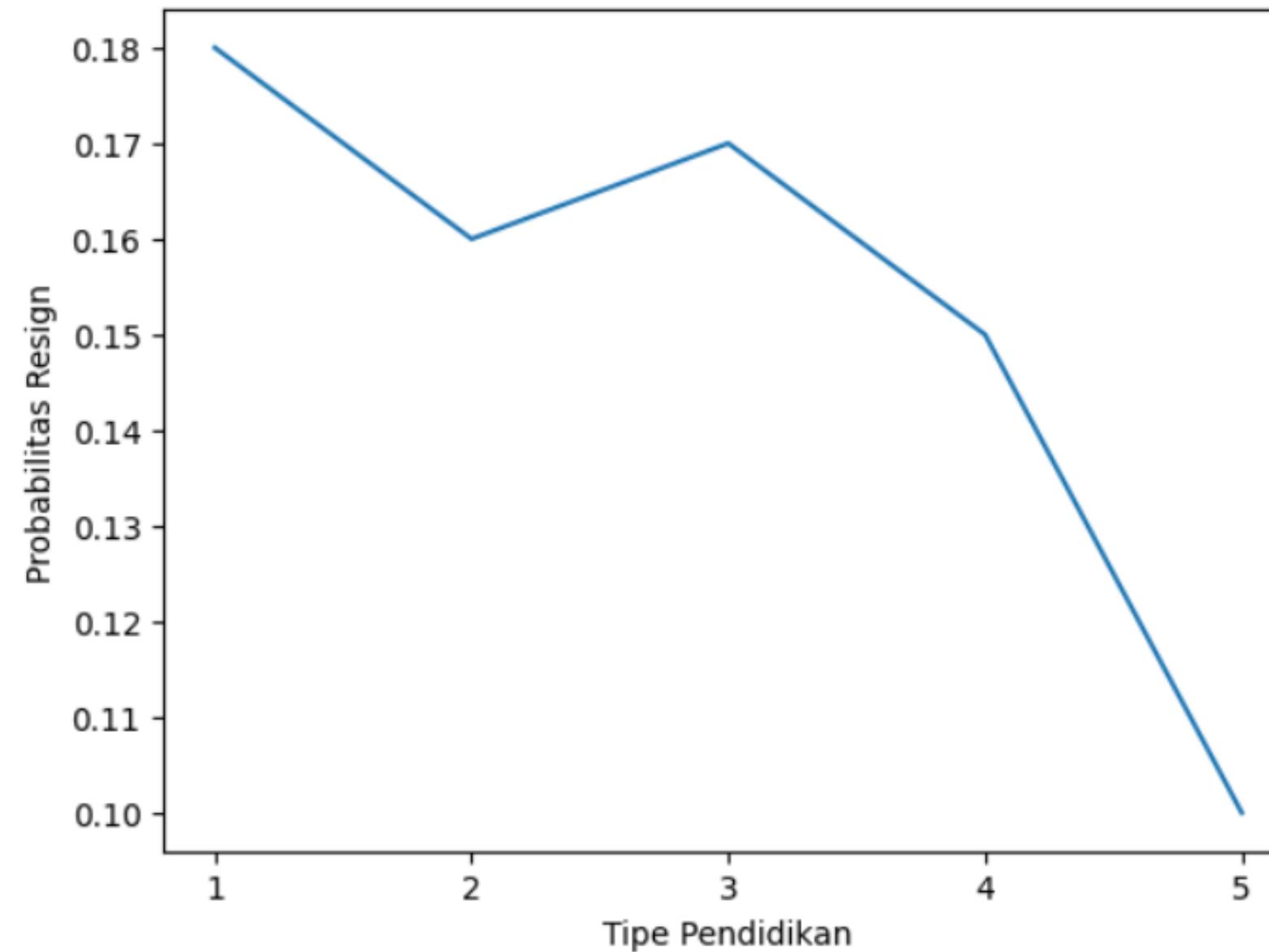


Tingkat Edukasi: 5

Mencakup karyawan dengan pendidikan Doktor

Memiliki probabilitas resign
sebesar 0.10

ANALISIS DAN INTERPRETASI



- Semakin tinggi tingkat pendidikan seseorang maka semakin kecil kemungkinan untuk resign, dengan pengecualian tingkat pendidikan 3
- Tingkat pendidikan 3 menjadi anomali, hal ini kami asumsikan karena tingkat pendidikan 3 merupakan jumlah terbanyak
- Tingkat pendidikan memiliki korelasi terhadap keinginan resign



KARYAWAN DENGAN TINGKATAN
EDUKASI MANAKAH YANG LEBIH
MUNGKIN UNTUK RESIGN?

KARYAWAN DENGAN TINGKATAN EDUKASI MANAKAH YANG LEBIH MUNGKIN UNTUK RESIGN?

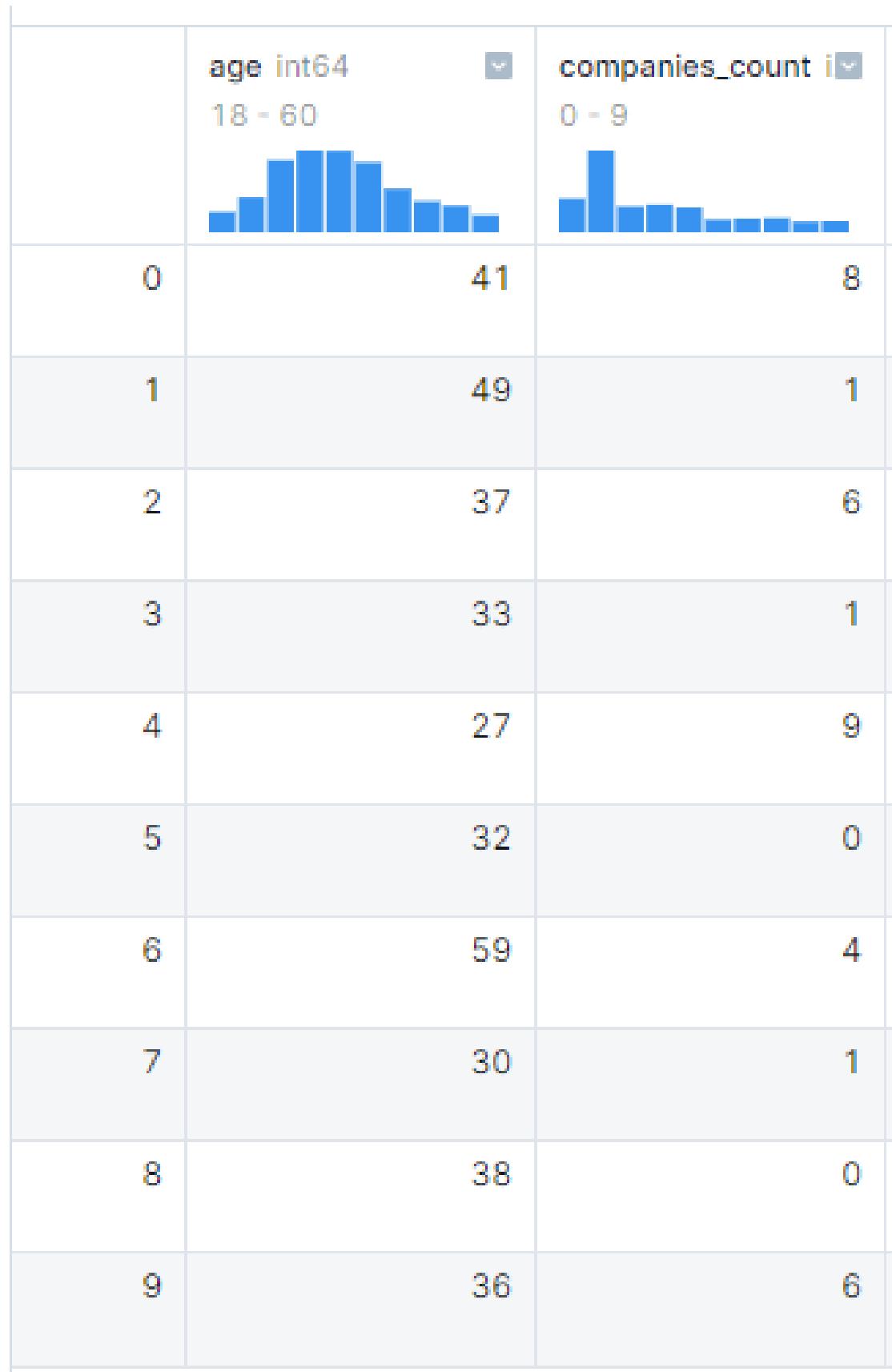
Karena eksploratori kami yang sebelumnya, pertanyaan ini dapat dijawab juga tanpa harus melakukan eksploratori lainnya

Tingkat edukasi yang paling mungkin untuk resign adalah tingkat edukasi 1



**BERDASARKAN UMUR, EMPLOYEE
SIAPA SAJA YANG SERING
BERPINDAH PERUSAHAAN?**

Pada dataset, kolom **age** memiliki nilai dari 18 tahun sampai 60 tahun. Agar saat divisualisasikan lebih mudah dipahami, kami membuat beberapa kelompok umur.

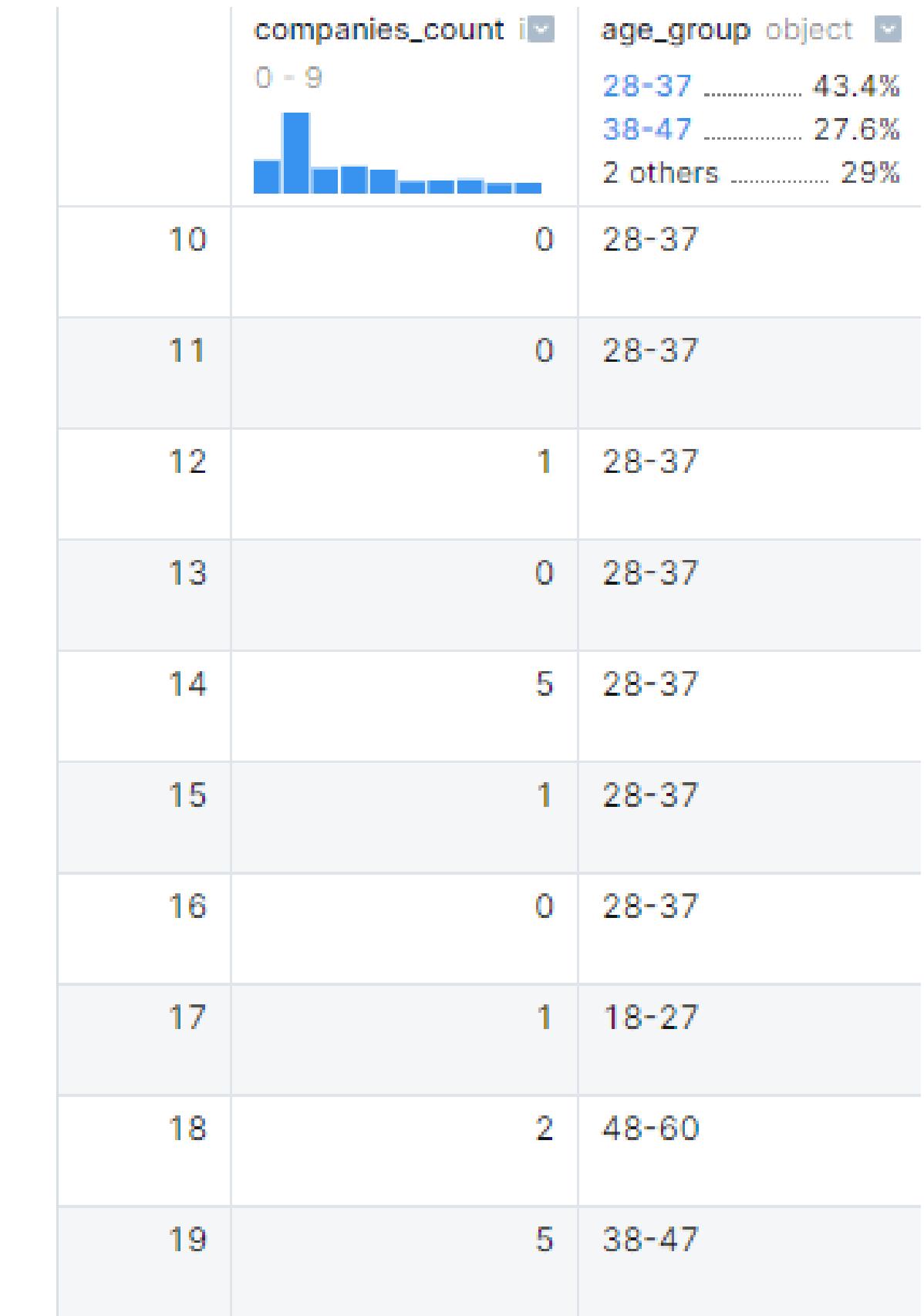


Berikut langkah yang kami lakukan agar data memiliki kolom kelompok umur

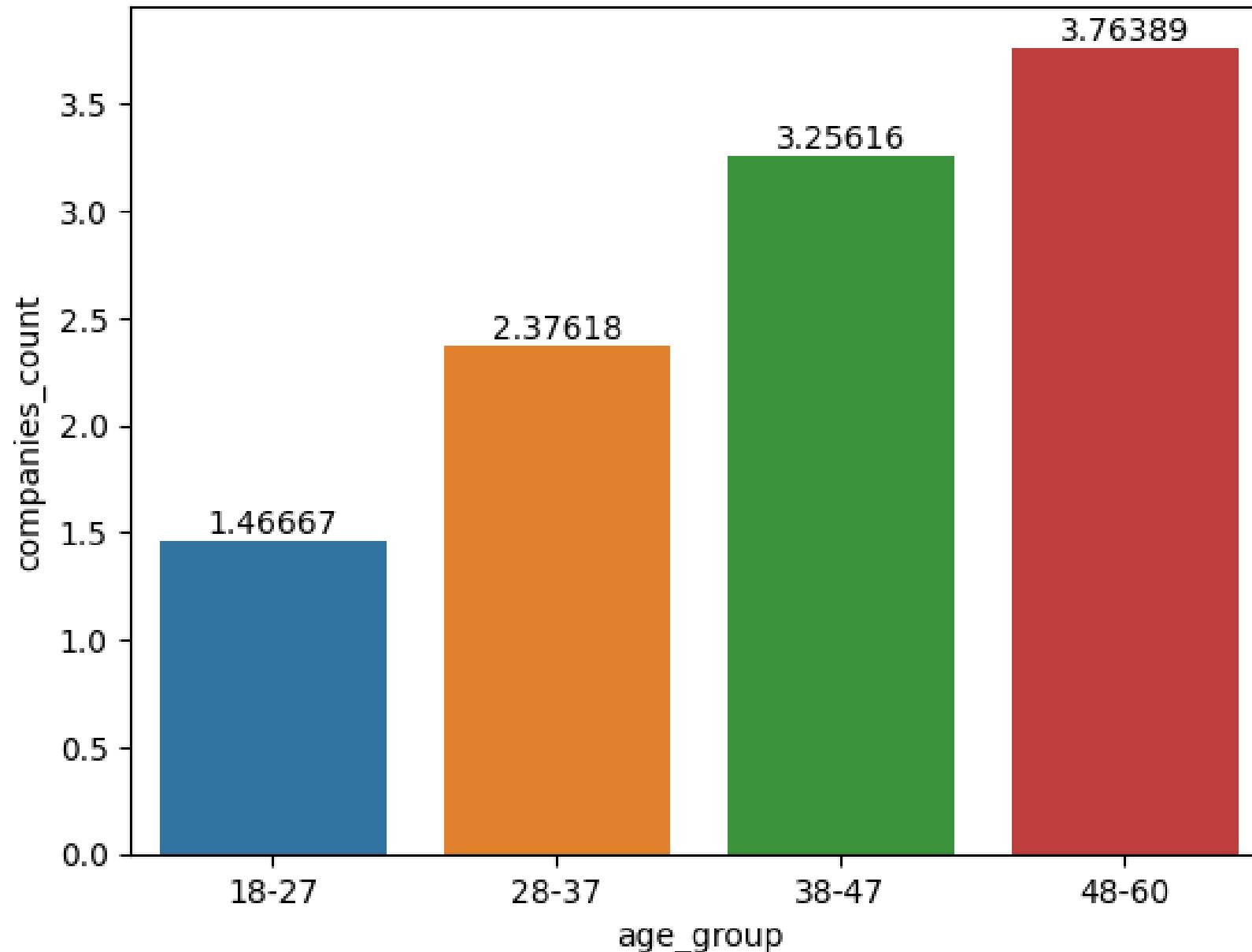
- Membuat kolom baru dengan label **age_group**
- Membuat kelompok umur seperti gambar di bawah, lalu menambahkannya ke dalam kolom baru
 - Employee berumur 18-27 tahun
 - Employee berumur 28-37 tahun
 - Employee berumur 38-47 tahun
 - Employee berumur 48-60 tahun
- Menghapus kolom **age**

Hasil data yang baru seperti ada gambar di kanan

```
df2.loc[df2['age'] <= 27, 'age_group'] = '18-27'  
df2.loc[df2['age'].between(28, 37), 'age_group'] = '28-37'  
df2.loc[df2['age'].between(38, 47), 'age_group'] = '38-47'  
df2.loc[df2['age'] > 47, 'age_group'] = '48-60'
```



Jumlah perusahaan employee pernah bekerja berdasarkan umur



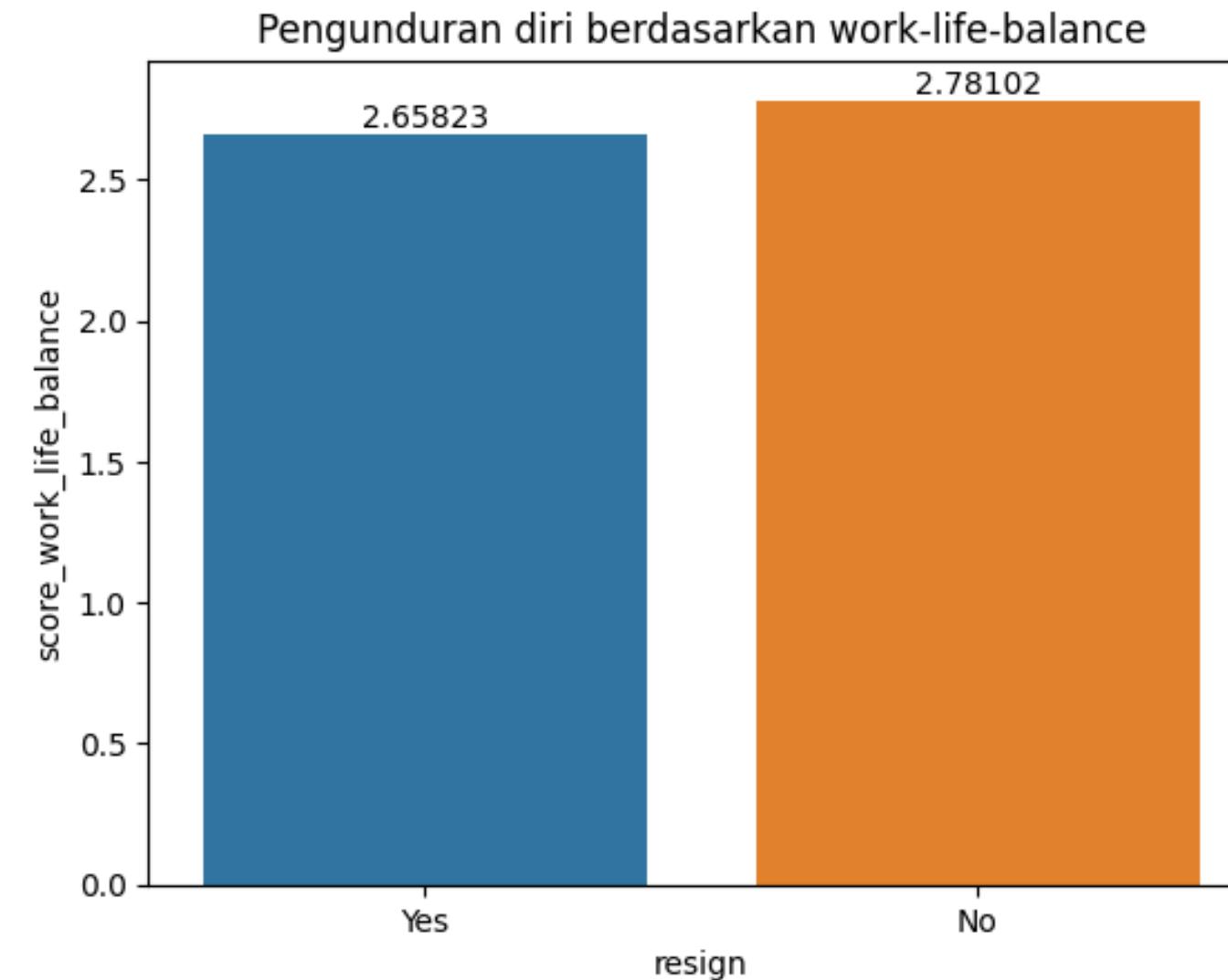
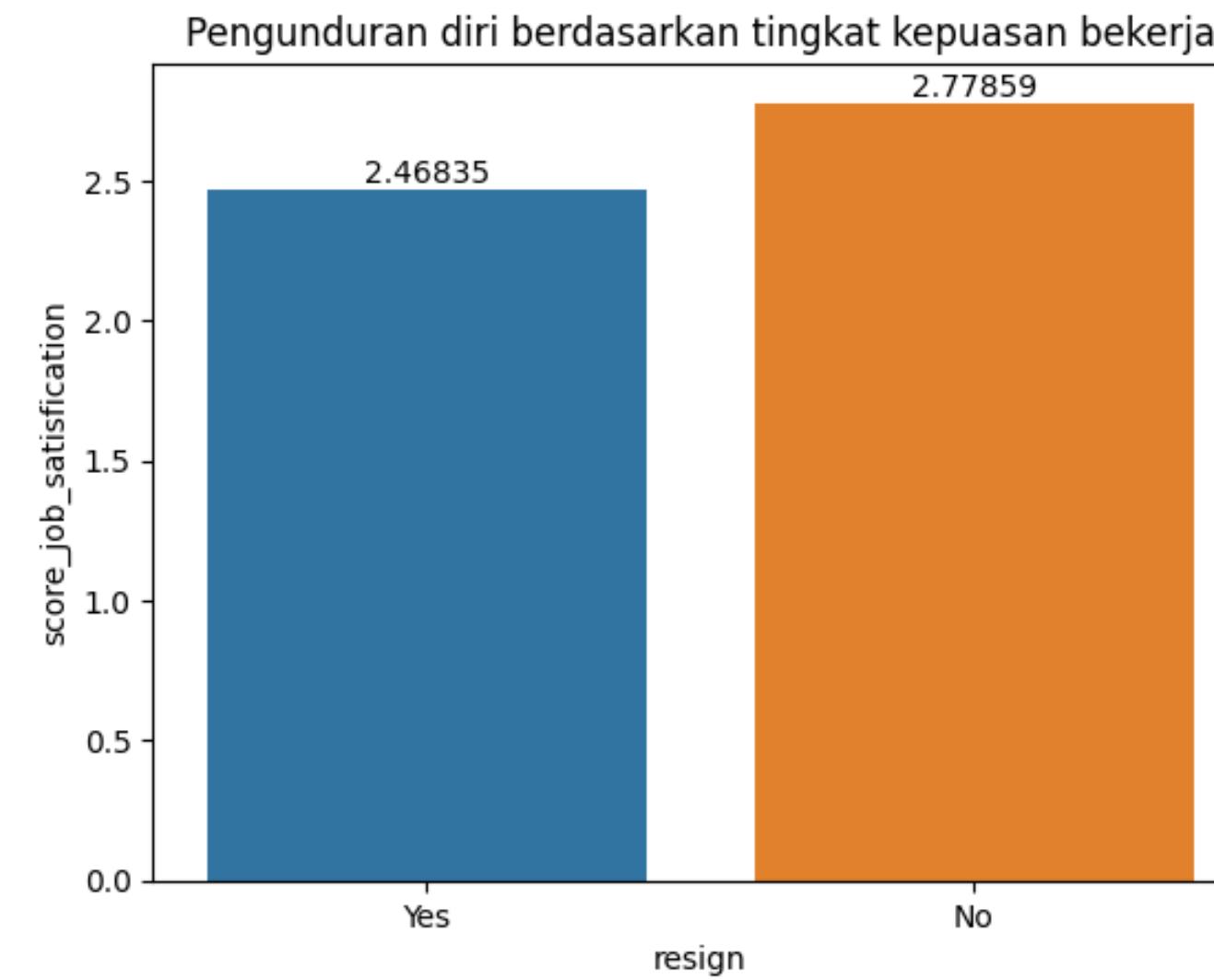
Grafik di atas adalah hasil plot dari data yang sudah diolah sebelumnya.

Berdasarkan visualisasi tersebut, **employee yang lebih berumur ternyata lebih sering berpindah perusahaan**. Employee yang berumur 48-60 tahun rata-rata sudah pernah bekerja di 3 perusahaan.



BAGAIMANAKAH TINGKAT KEPUASAN BEKERJA DAN WORK- LIFE-BALANCE EMPLOYEE TERHADAP RESIGNATION?

Untuk kasus ini, kami melakukan exploratory pada dua fitur, yaitu ***score_job_satisfaction*** dan ***score_work_life_balance*** terhadap ***resign***



Berdasarkan visualisasi tersebut, ternyata employee ***yang resign cenderung tidak puas terhadap pekerjaannya*** dan semua employee ***yang resign dan yang tidak resign cenderung dapat menyeimbangkan kehidupan sehari-hari mereka dengan pekerjaannya***.

Co.

4

Membuat Model



OUTLINE MEMBUAT MODEL

- A** Prediksi karyawan akan resign atau tidak resign
- B** Prediksi berapa lama karyawan akan bertahan di perusahaan
- C** Cluster yang terdapat pada dataset dan karakteristiknya



**PREDIKSI UNTUK MENGETAHUI APAKAH
KARYAWAN AKAN RESIGN ATAU TIDAK
DI PERUSAHAAN TERSEBUT.**

KLASIFIKASI

Permasalahan ini dapat diselesaikan dengan metode unsupervised learning yaitu **klasifikasi**. Terdapat dua tahapan yang dilakukan untuk klasifikasi yaitu :

- 1) Pre-Processing
- 2) Modeling



PRE-PROCESSING

1. Encoding

Terdapat 26 fitur input dengan **6 bersifat kategorikal**. Kami melakukan encoding fitur kategorikal dan output label menggunakan **LabelEncoder**.

Tidak dilakukan encoding / scaling pada fitur numerikal karena tidak diperlukan untuk modeling.

2. Feature Selection

Seleksi fitur digunakan untuk meminimalkan beban komputasi. Dengan menggunakan metode ANOVA kami mendapatkan 15 fitur sebagai berikut:

- age
- division
- home_distance
- score_environment
- major
- score_contribution
- job_rank
- score_job_satisfaction
- marriage_status
- monthly_income
- over_time
- time_total_working
- time_current_company
- time_current_role
- time_current_manager

MODELING

Dataset yang dimiliki **imbanced** dengan perbandingan 1233 : 237, sehingga memerlukan **oversampling / undersampling**

Selain itu, proses ini juga memerlukan **Hyper Parameter Tuning** dan **Cross Validation** untuk menemukan model optimal

Namun, terdapat setidaknya 2 kejadian yang memungkinkan **leakage** dari data testing ke data training di antara proses tersebut:

- Jika sampling dilakukan sebelum splitting data ke train dan test set
- Jika oversampling / undersampling dilakukan sebelum CV

Keduanya menyebabkan model overfit.

SOLUSI

Tahapan yang dilakukan adalah sebagai berikut:

1. Melakukan splitting data training dan testing
2. Melakukan oversampling / undersampling pada setiap fold dalam Cross Validation untuk mendapatkan hyperparameter yang optimal (Hyperparameter Tuning)
3. Melakukan training (model.fit) pada data training dengan Cross Validation
4. Melakukan prediksi terhadap data testing
5. Menampilkan evaluasi hasil prediksi

MODELING

1

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, random_state=42, stratify=y)
```

```
from sklearn.model_selection import StratifiedKFold  
kfold = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
```

Stratified KFold dipilih agar **distribusi tiap kelas** pada proses training dan testing **konsisten**

2

```
from sklearn.pipeline import Pipeline, make_pipeline  
from imblearn.pipeline import Pipeline, make_pipeline  
  
params = {  
    'n_estimators' : [150, 200],  
    'max_depth' : [None, 1, 3],  
    'min_samples_leaf': [1, 3],  
    'min_samples_split': [2, 7, 10]  
}  
  
imba_pipeline = make_pipeline(RandomOverSampler(sampling_strategy='minority', random_state=42),  
| | | | | | | RandomForestClassifier(n_estimators=100, random_state=42))  
cross_val_score(imba_pipeline, X_train, y_train, scoring='recall', cv=kfold)
```

Recall adalah metric utama karena tujuan klasifikasi utamanya untuk memprediksi **ground truth resign = yes** sebagai **yes**

3

```
new_params = {'randomforestclassifier__' + key: params[key] for key in params}  
grid_imba = GridSearchCV(imba_pipeline, param_grid=new_params, cv=kfold, scoring='recall',  
| | | | | | | return_train_score=True)  
grid_imba.fit(X_train, y_train);
```



4-5

```
y_test_predict = grid_imba.predict(X_test)  
evaluate_classifier_performance(y_test_predict, y_test)
```

EVALUASI MODEL

Model yang digunakan pada klasifikasi ini adalah **Random Forest Classifier** dan **Naive Bayes Classifier (Bernoulli)** hal ini karena keduanya dapat mengakomodasi klasifikasi untuk mixed input (nominal maupun kategorikal)

Metrics	Naive - RF	Random Oversampling - RF	SMOTE - RF	Random Undersampling - RF	Nearmiss2 - RF
Recall	0,6	0,74	0,73	0,72	0,51
Accuration	0,85	0,78	0,81	0,74	0,28
	0,725	0,76	0,77	0,73	0,395

CONFUSION MATRIX

prediction		0	1
actual	0	178	7
1	27	9	

prediction		0	1
actual	0	149	36
1	12	24	

prediction		0	1
actual	0	156	29
1	14	22	

prediction		0	1
actual	0	138	47
1	11	25	

prediction		0	1
actual	0	31	154
1	5	31	

EVALUASI MODEL

Metrics	Naive - NB	SMOTE - NB	Random Undersampling - NB
Recall	0,56	0,71	0,7
Accururation	0,81	0,81	0,71
	0,685	0,76	0,705

CONFUSION
MATRIX

prediction		0	1
actual	0	173	12
	1	30	6

prediction		0	1
actual	0	159	26
	1	16	20

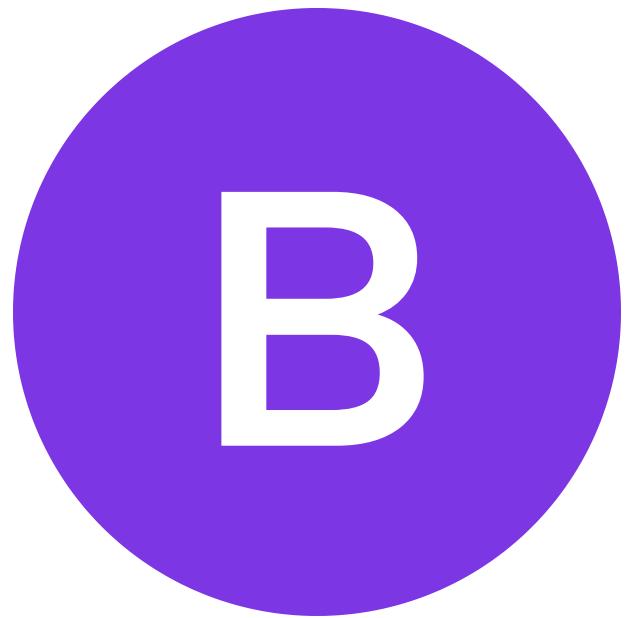
prediction		0	1
actual	0	131	54
	1	11	25

KESIMPULAN

Kami menilai model paling optimal adalah SMOTE Random Forest dengan akurasi 0.81 dan recall 0.73.

Model ini dapat digunakan oleh pihak HR untuk :

- memprediksi apakah karyawan akan melakukan resign atau tidak di tahun-tahun berikutnya
- memprediksikan ketersediaan sumber daya manusia, pengetahuan, dan kemampuan di perusahaan
- HR dapat lebih memperhatikan fitur-fitur (atau pada dunia nyata komponen) yang berperan penting dalam pengunduran diri karyawan sehingga dapat membuat keputusan yang bisa menurunkan tingkat pengunduran diri karyawan.



**MODEL PREDIKSI UNTUK MENGETAHUI
BERAPА LAMA SEORANG KARYAWAN
AKAN BERTAHAN DI PERUSAHAAN
TERSEBUT.**

REGRESI

Permasalahan ini kami tetapkan sebagai permasalahan supervised learning dengan model regresi.

- Target dari permasalahan ini adalah time_current_company
- Hanya mengambil data yang nilai resignnya = 1 (YES)
- Fitur yang digunakan berdasarkan seleksi fitur
- Seleksi fitur dilakukan dengan metode filter menggunakan korelasi dan metode embedded dengan menggunakan model lasso (untuk uji model)
- Menggunakan 3 model untuk komparasi yaitu normal linear model, lasso, dan random forest.

PREPROCESSING

- Mengambil data yang memiliki data resign = YES (237 data dengan 26 fitur / exclude resign)
- Membagi target dan fitur juga train dan test (0.85 & 0.15)
- Melakukan normalisasi untuk model lasso dan linear (min-max scaler)
- Melakukan encode untuk data data kategorikal pada model random forest (Label Encoder)
- Melakukan gridsearch cross validation untuk hyperparameter random forest

```
resign_data = cleaned[cleaned["resign"] >= 'Yes']
resign_data = resign_data.drop("resign",axis=1)

total_rows, total_attributes = resign_data.shape

print('Jumlah data resign:', total_rows)
print("Jumlah atribut:", total_attributes)
```

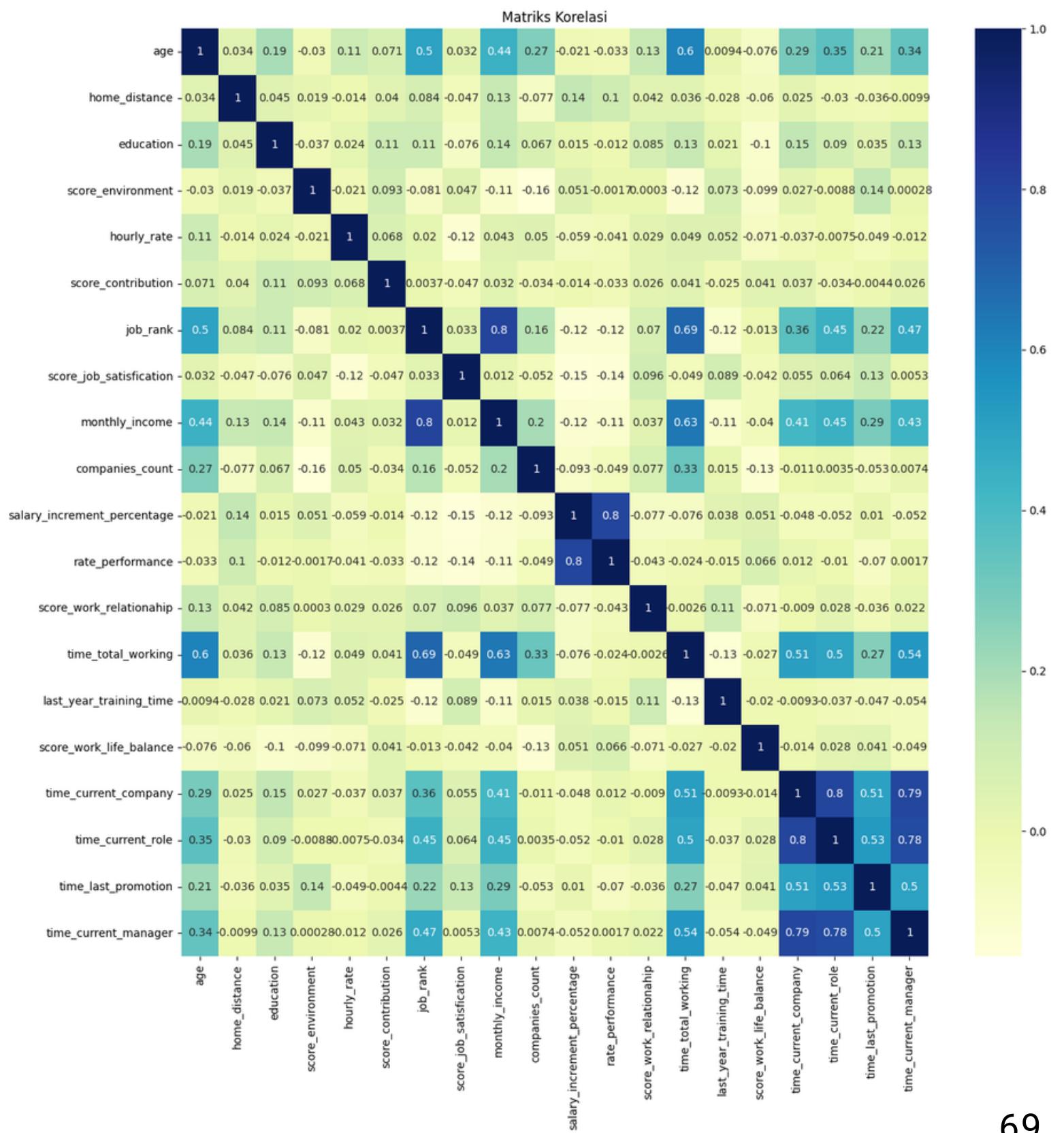


Jumlah data resign: 237
Jumlah atribut: 26

SELEKSI FITUR

Dari hasil uji korelasi, kami mengambil threshold 0,1 untuk mengaggap bahwa variabel tersebut memiliki korelasi dengan variabel target.

- age (0,29)
- education (0,15)
- job_rank (0,36)
- monthly_income (0,41)
- time_total_working (0,51)
- time_current_role (0,80)
- time_last_promotion (0,51)
- time_current_manager (0,79)



KOEFISIEN LASSO 8 FITUR (ALPHA = 0.1)

- age (0)
- education (0)
- job_rank (0)
- monthly_income (0)
- time_total_working (0, 06)
- time_current_role (6, 44)
- time_last_promotion (0, 53)
- time_current_manager (5, 70)

Variabel dengan koefisien = 0 kami tidak gunakan (sesuai dengan penggunaan lasso sebagai embedded feature selection)

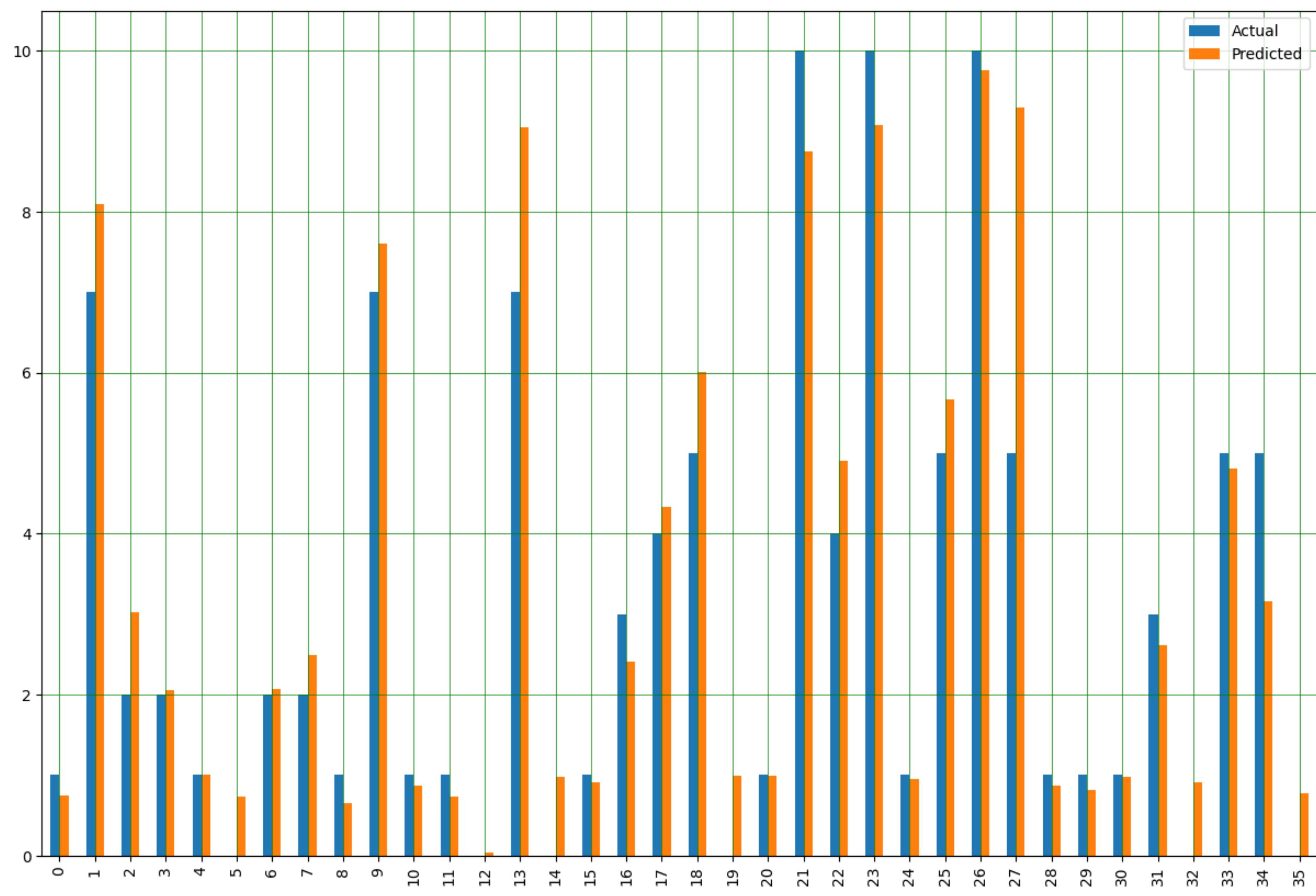
SELEKSI MODEL

TRAIN = 0.85 (201) & TEST = 0.15 (36)

Evaluasi	Lasso (8 Fitur)	Linear (4 Fitur)	Random Forest (4 Fitur)	Random Forest (8 Fitur)	Random Forest (25 Fitur)
MSE	2.07	1.80	1.38	1.05	0.94
RMSE	1.44	1.34	1.18	1.02	0.97
R_Squared	0.76	0.79	0.84	0.88	0.89

PREDIKSI DATA DARI MODEL FINAL (RANDOM FOREST 8 FITUR)

72





ANALISIS CLUSTER YANG DAPAT TERBENTUK PADA DATA KARYAWAN DAN MASING-MASING KARAKTERISTIKNYA

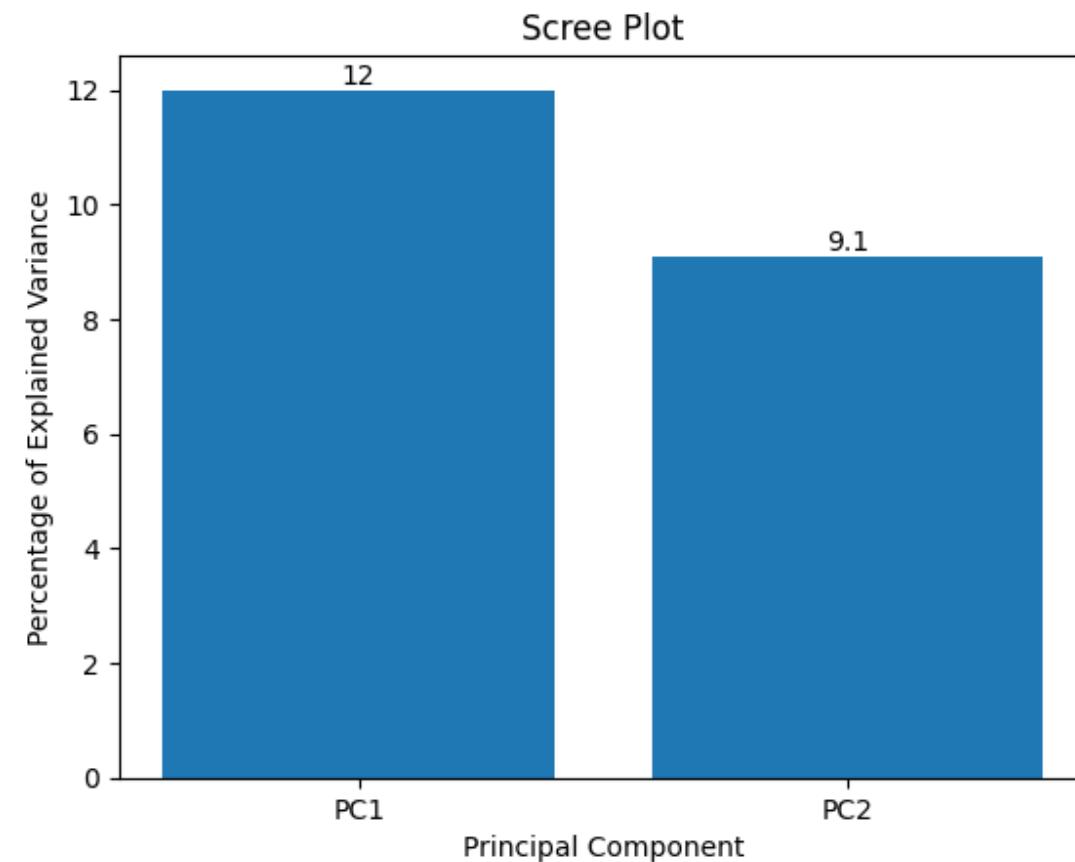
PREPROCESSING

Sebelum melakukan pemodelan clustering, ada beberapa langkah preprocessing yang harus dilakukan:

- **Mengubah nilai fitur *division*, *major*, dan *role* menjadi singkatan**
- **Melakukan One-Hot Encoding** untuk fitur-fitur yang kategorikal yang bernilai kata-kata, seperti *resign*, *division*, *major*, *gender*, *role*, *marriage_status*, dan *over_time*
- **Scaling data** yang sudah di-encode menggunakan *MinMaxScaler*
- Melakukan **feature selection** menggunakan metode PCA

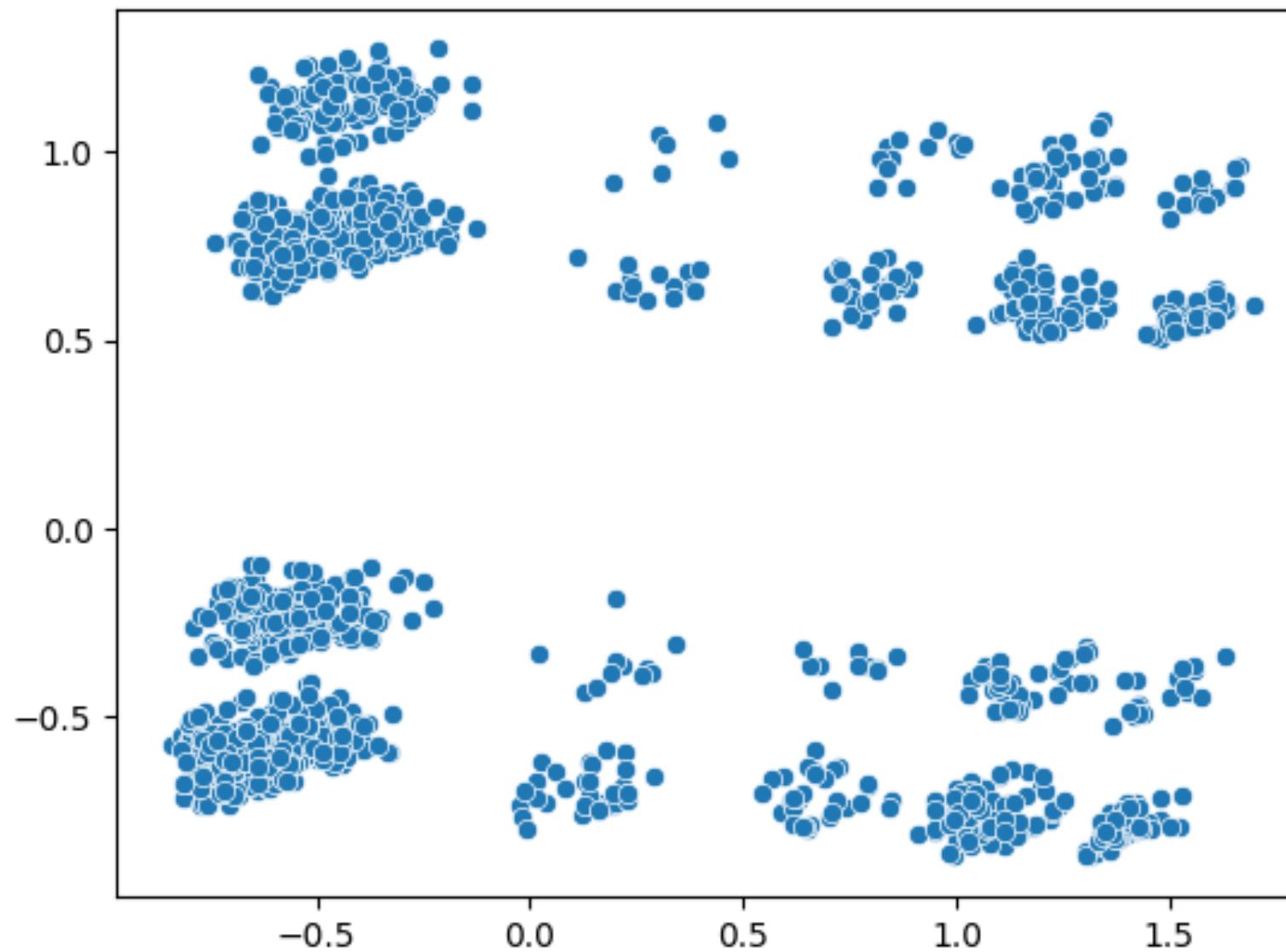
PRINCIPAL COMPONENT ANALYSIS (PCA)

Pada tahap ini, kami mengekstrak dua fitur baru dari 47 fitur yang ada pada preprocessed data untuk memudahkan visualisasi saat clustering.



Dari hasil PCA, berikut adalah explained variance pada principal component yang pertama dan kedua.

PRINCIPAL COMPONENT ANALYSIS (PCA)



Berikut adalah hasil scatterplot berdasarkan kedua principal component sebelumnya.

PEMODELAN CLUSTERING

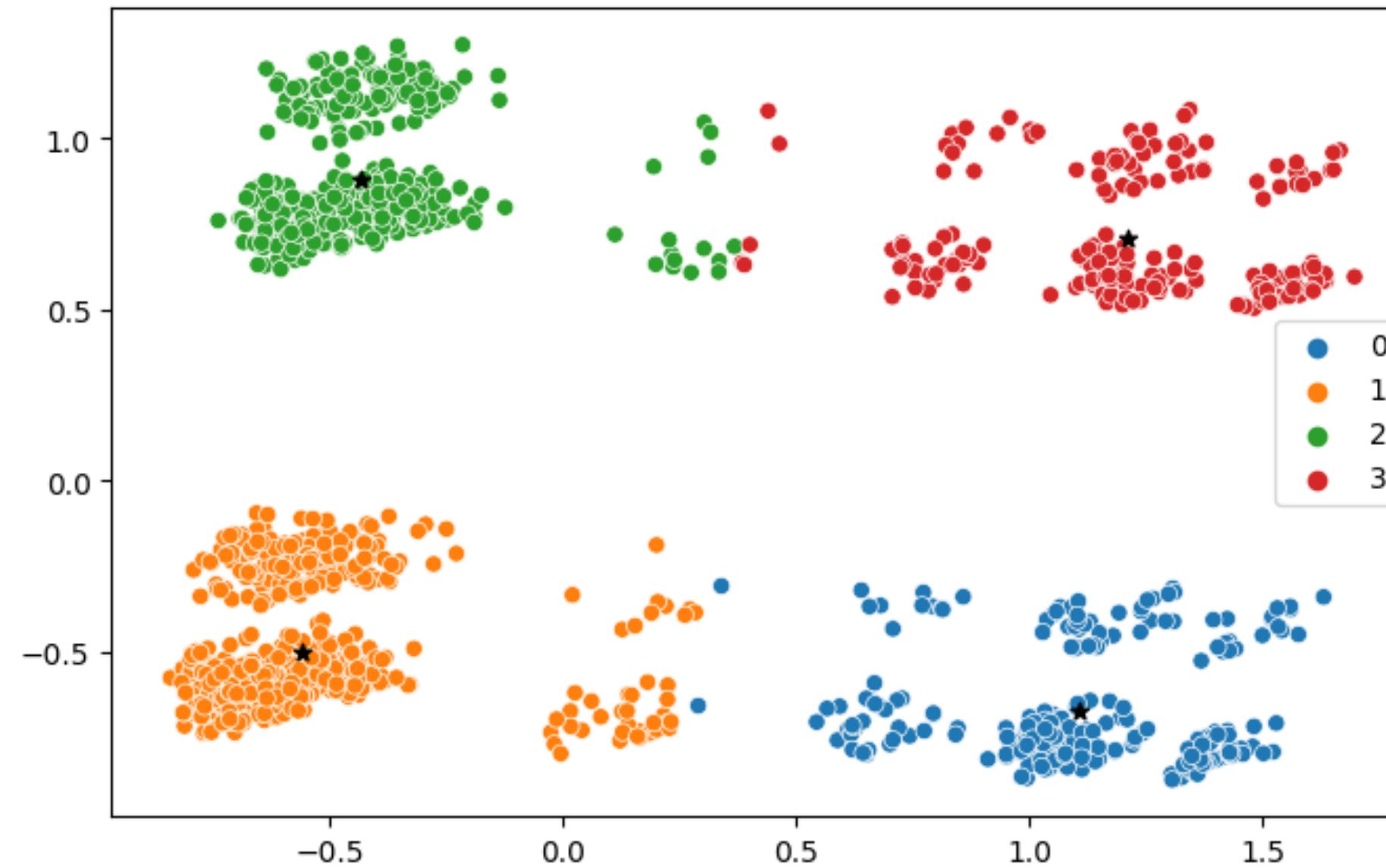
Model yang kami pilih untuk clustering adalah **K-Means**. Selain itu, agar hasil clustering optimal, kami **menguji silhouette score** berdasarkan k dari 2-10

```
For n_clusters = 2 , The average silhouette_score is: 0.547266118330054
For n_clusters = 3 , The average silhouette_score is: 0.6708857524997428
For n_clusters = 4 , The average silhouette_score is: 0.7430877424802874 ←
For n_clusters = 5 , The average silhouette_score is: 0.7113549574333344
For n_clusters = 6 , The average silhouette_score is: 0.6843430066318293
For n_clusters = 7 , The average silhouette_score is: 0.6130168882220658
For n_clusters = 8 , The average silhouette_score is: 0.5645905198765399
For n_clusters = 9 , The average silhouette_score is: 0.5392734240506308
For n_clusters = 10 , The average silhouette_score is: 0.5031974074630599
```

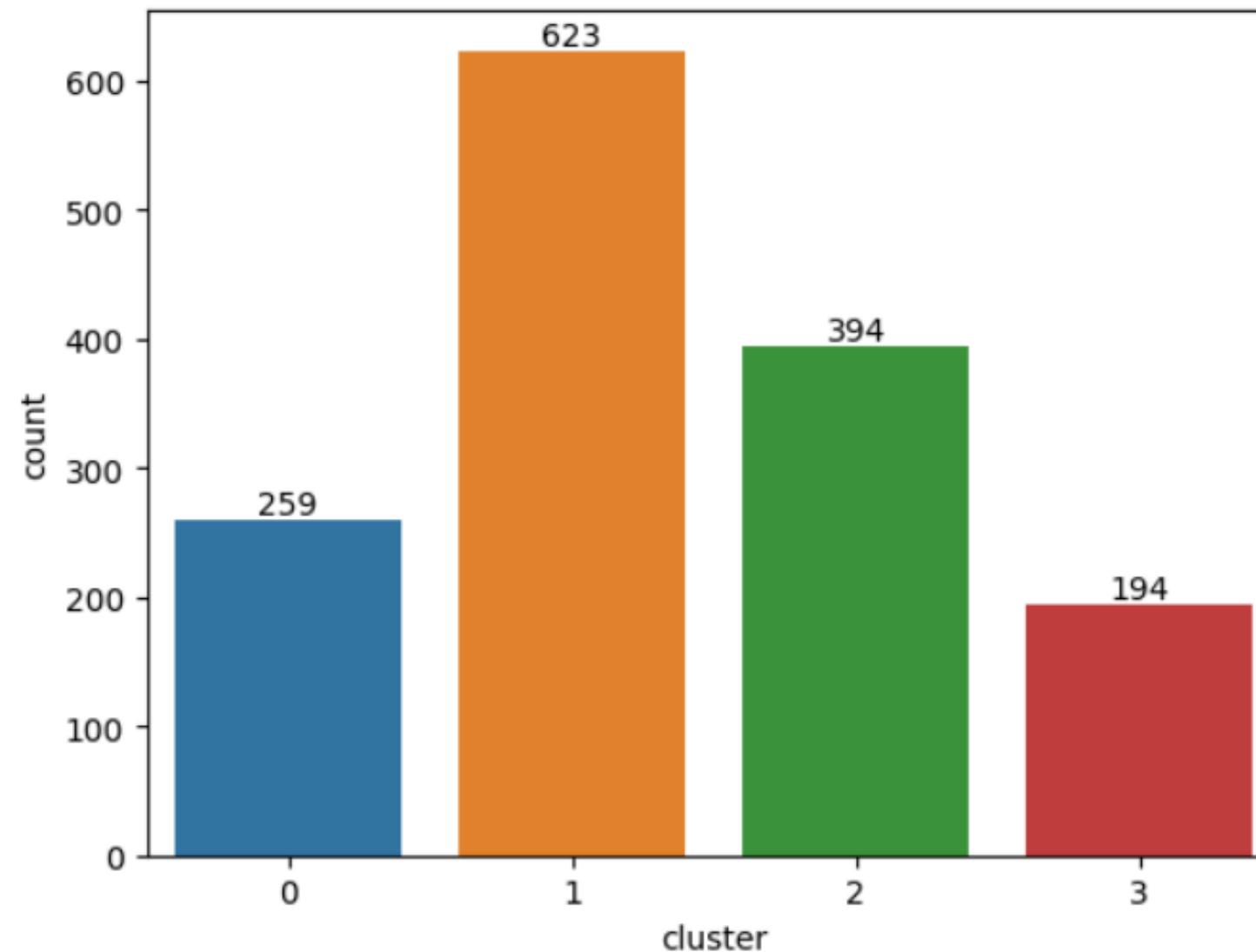
Dari hasil di samping,
 k bernilai 4 adalah
hasil yang terbaik

PEMODELAN CLUSTERING

Berikut adalah hasil clustering dengan k bernilai 4

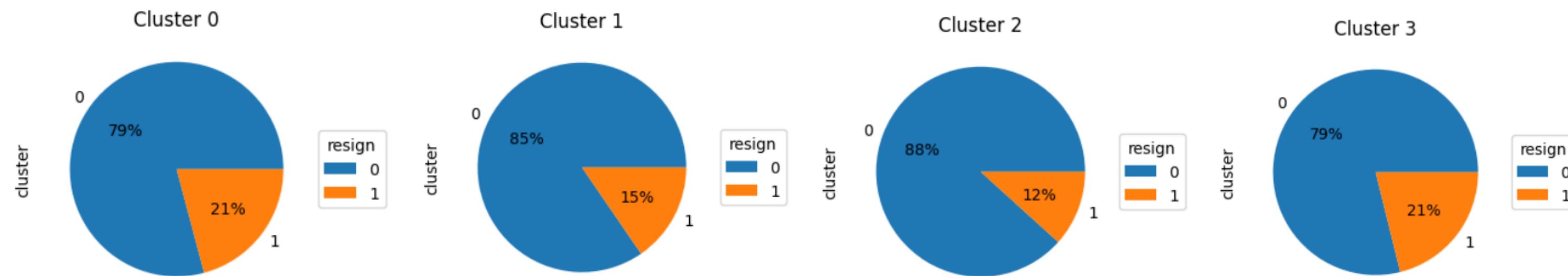


INTERPRETASI CLUSTERING : JUMLAH CLUSTER



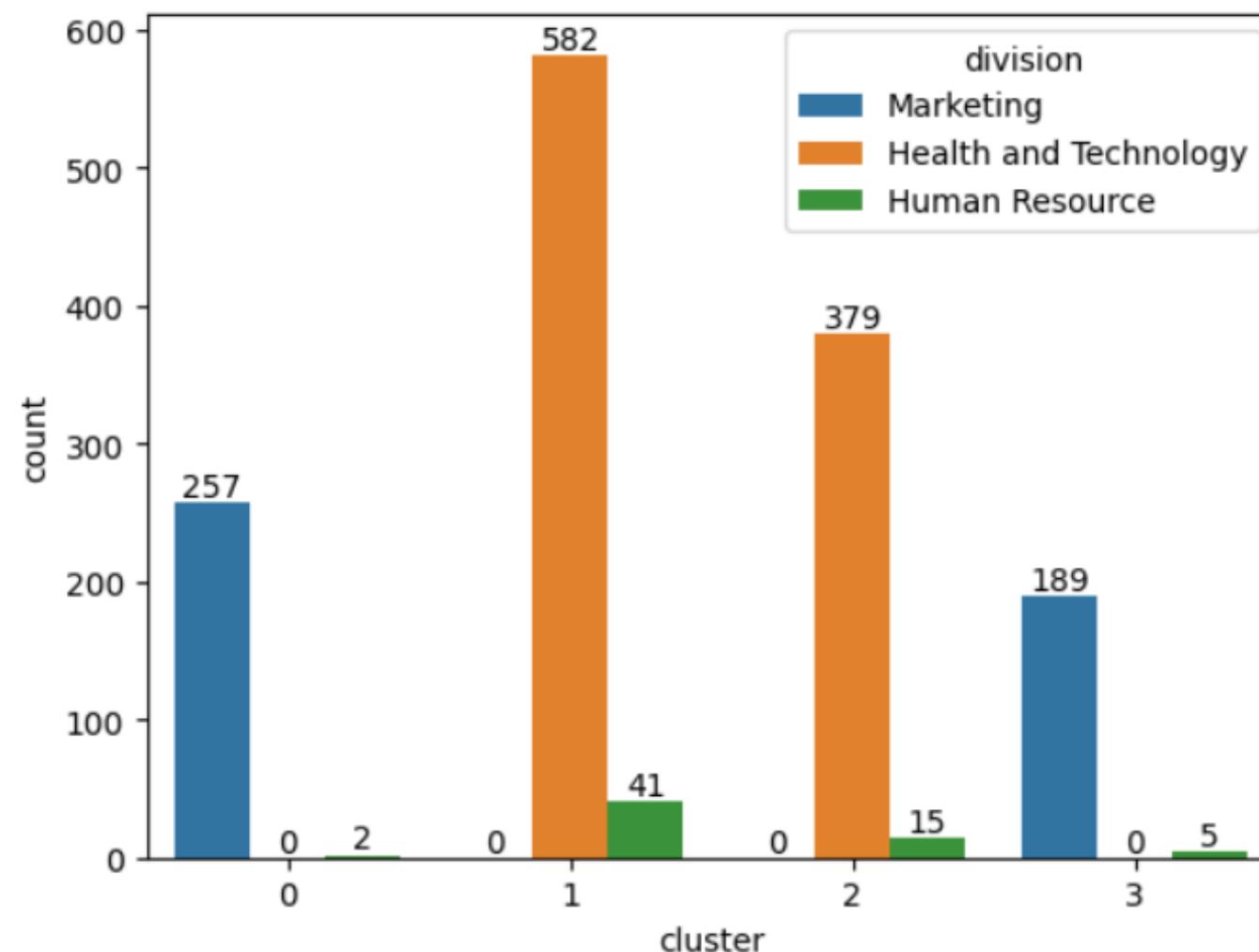
- Seperti yang ditunjukkan oleh visualisasi disamping **cluster 1** merupakan cluster dengan jumlah anggota paling **banyak**
- Sementara itu **cluster 3** merupakan cluster dengan jumlah anggota paling **sedikit**

INTERPRETASI CLUSTERING : PERSENTASE RESIGN



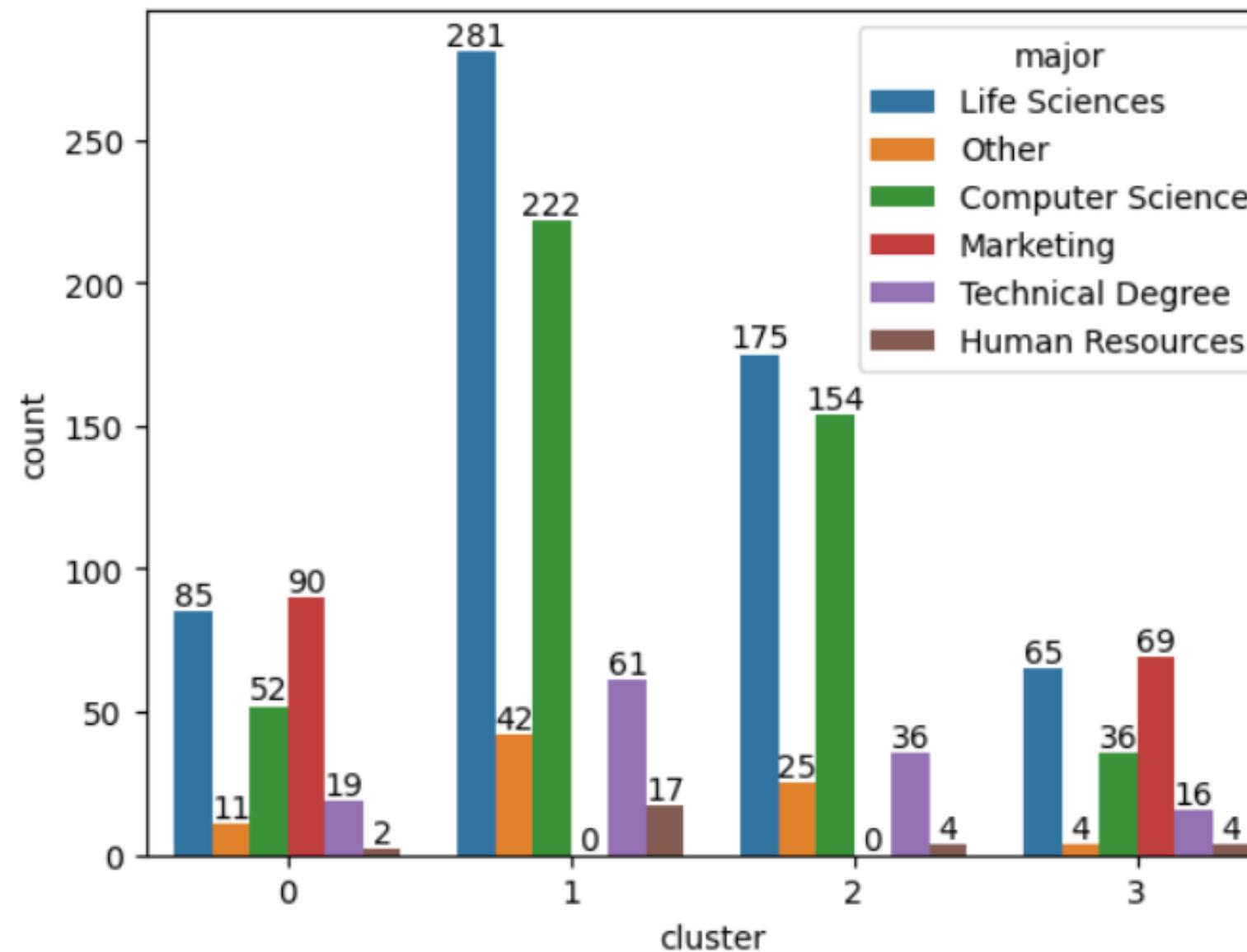
- Seperti yang dapat dilihat pada visualisasi diatas cluster dengan **persentase resign paling tinggi** adalah **cluster 0, cluster 3**
- Semetara itu cluster dengan **persentase resign paling rendah** adalah **cluster 1** dan diikuti oleh **cluster 2**

INTERPRETASI CLUSTERING : DIVISI BEKERJA



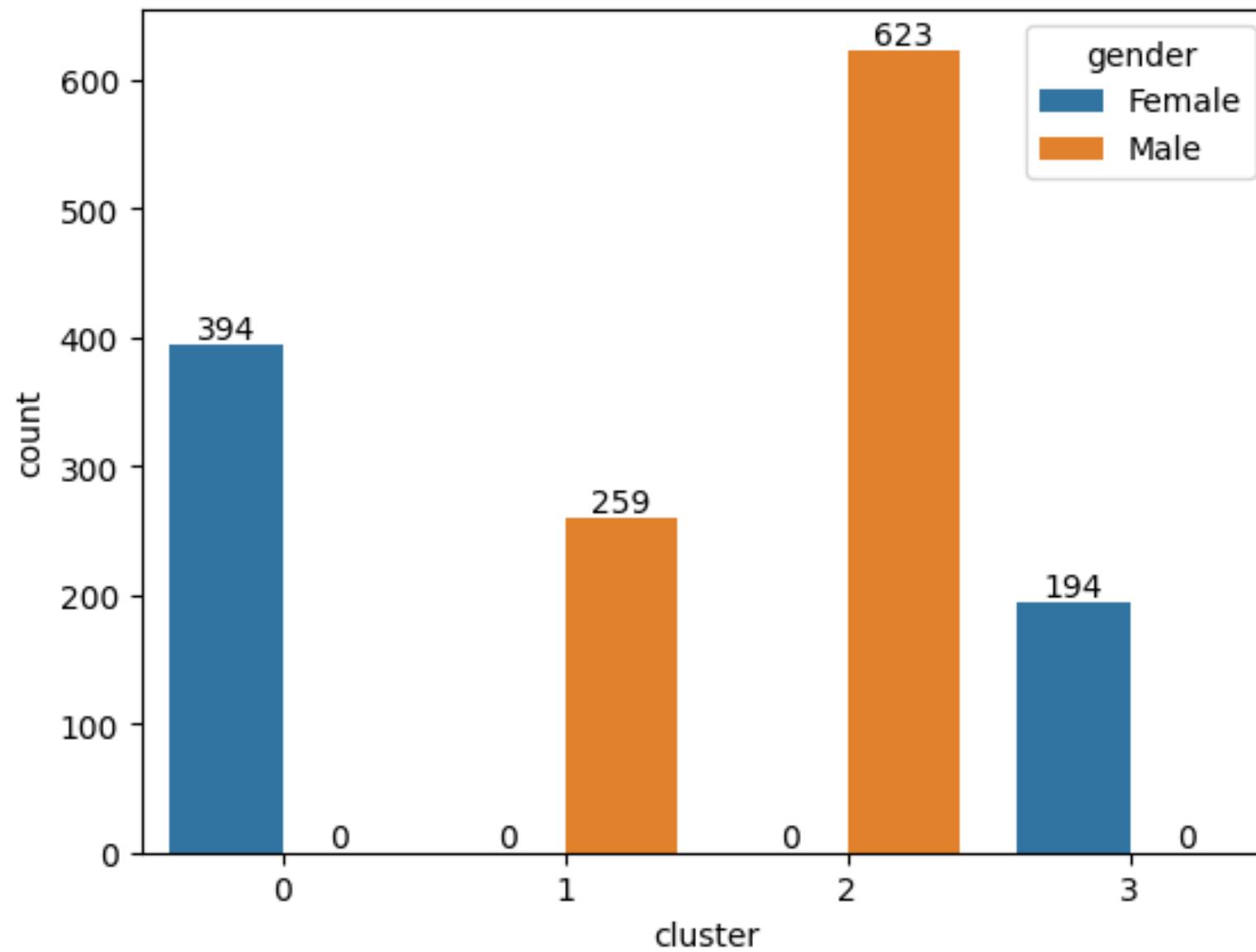
- Pada visualisasi disamping dapat dilihat **divisi Health and Technology mendominasi cluster 1 dan cluster 2**
- Sementara itu, **divisi Marketing mendominasi pada cluster 0 dan cluster 3**
- Perlu diketahui, **majoritas divisi Human Resource berada pada cluster 1**

INTERPRETASI CLUSTERING : JURUSAN LULUSAN



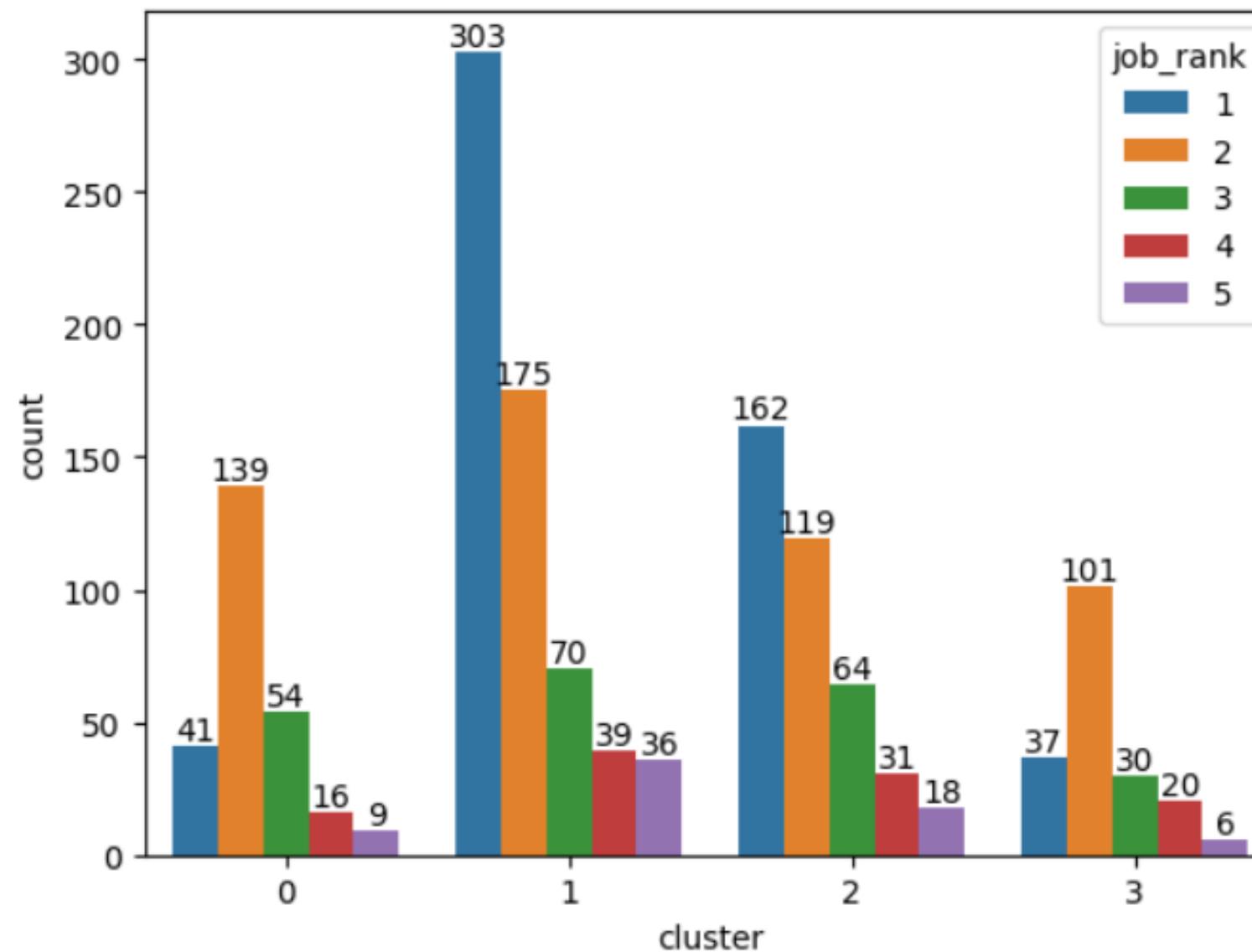
- **Cluster 0** didominasi oleh jurusan **Life Science** dan **Marketing**
- **Cluster 1** didominasi oleh jurusan **Life Science** dan **Computer Science**
- **Cluster 2** didominasi oleh jurusan **Life Science** dan **Computer Science**
- **Cluster 3** didominasi oleh jurusan **Life Science** dan **Marketing**

INTERPRETASI CLUSTERING : GENDER CLUSTER



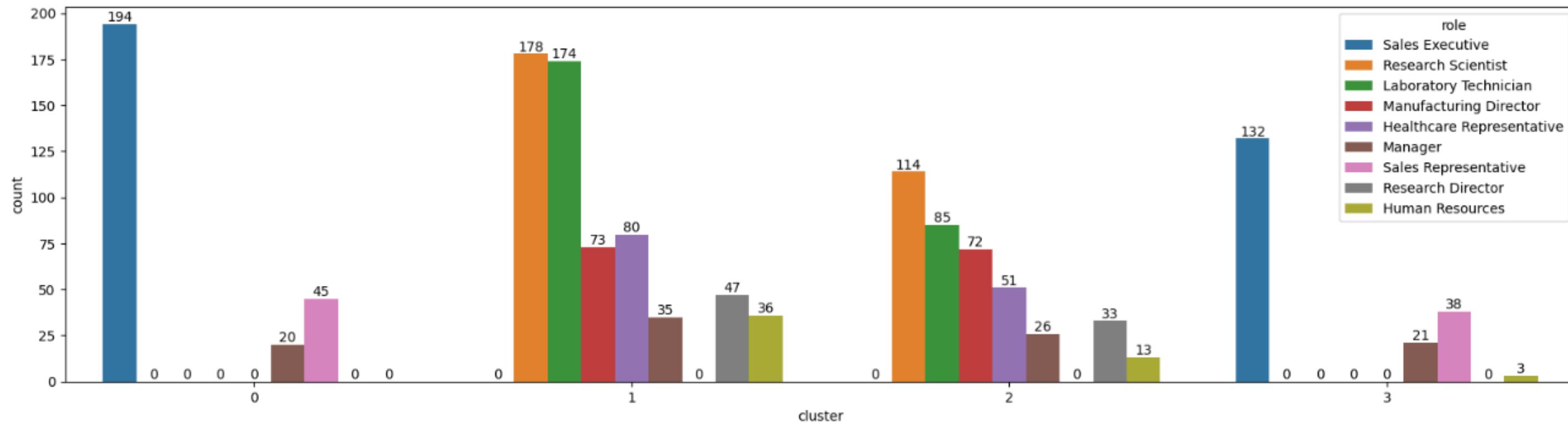
- Cluster 0 terdiri atas gender perempuan
- Cluster 1 terdiri atas gender laki-laki
- Cluster 2 terdiri atas gender laki-laki
- Cluster 3 terdiri atas gender perempuan

INTERPRETASI CLUSTERING : TINGKAT JABATAN



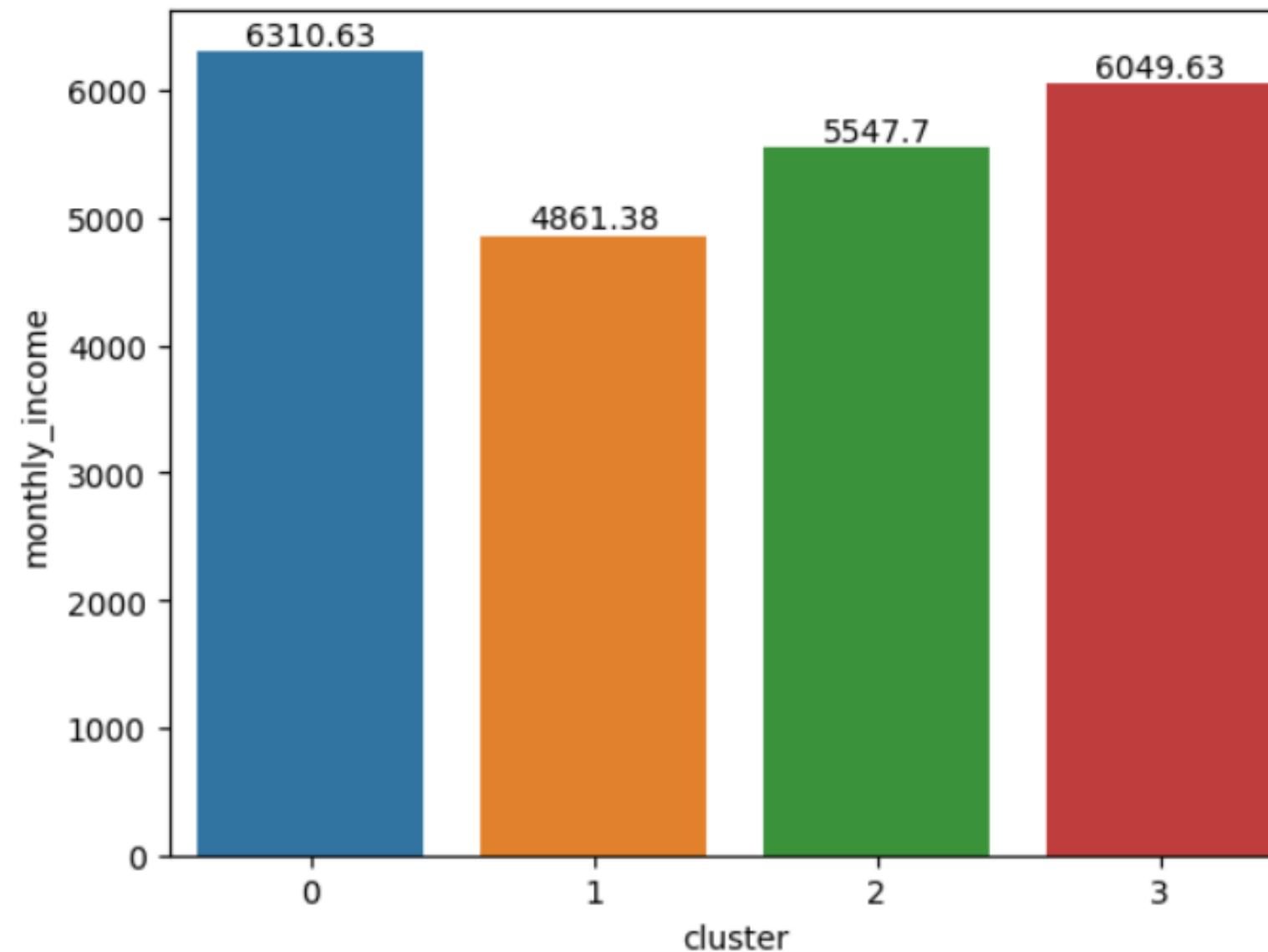
- **Cluster 0** didominasi oleh pekerja dengan **jabatan tingkat 2**
- **Cluster 1** didominasi oleh pekerja dengan **jabatan tingkat 1**
- **Cluster 2** didominasi oleh pekerja dengan **jabatan tingkat 1** dan diikuti **jabatan tingkat 2**
- **Cluster 3** didominasi oleh pekerja dengan **jabatan tingkat 2**

INTERPRETASI CLUSTERING : JABATAN



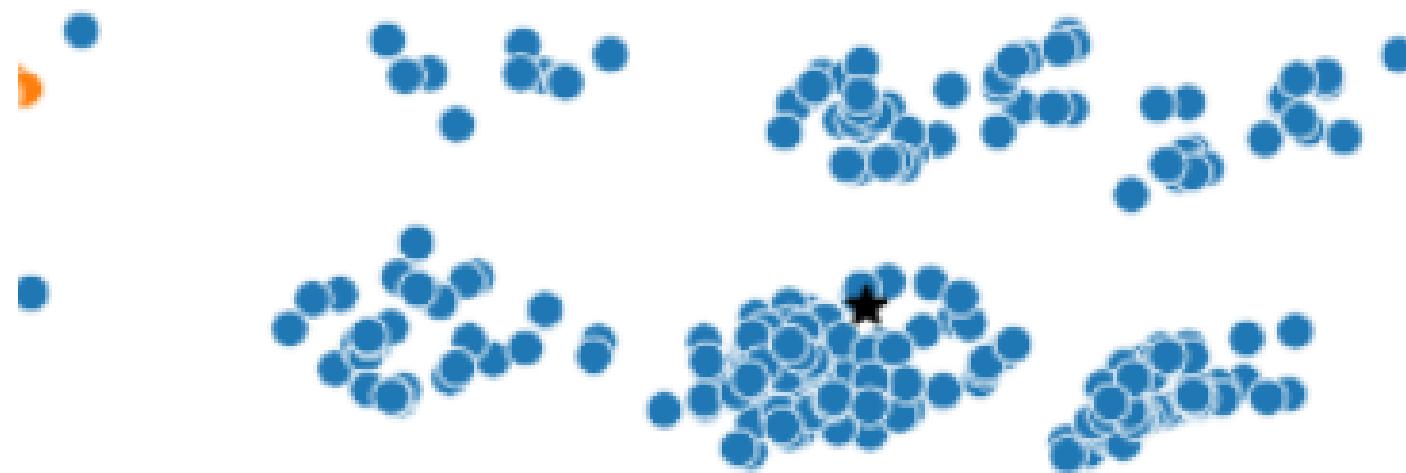
- **Cluster 0** didominasi oleh pekerja dengan **jabatan Sales Executive**
- **Cluster 1** didominasi oleh pekerja dengan **jabatan Research Scientist** dan **Laboratory Technician**
- **Cluster 2** didominasi oleh pekerja dengan **jabatan Research Scientist**
- **Cluster 3** didominasi oleh pekerja dengan **jabatan Sales Executive**

INTERPRETASI CLUSTERING : PENDAPATAN PER-BULAN



- **Cluster 0** memiliki jumlah **pendapatan** per-bulan **6310.63**
- **Cluster 1** memiliki jumlah **pendapatan** per-bulan **4861.38**
- **Cluster 2** memiliki jumlah **pendapatan** per-bulan **5547.7**
- **Cluster 3** memiliki jumlah **pendapatan** per-bulan **6049.63**

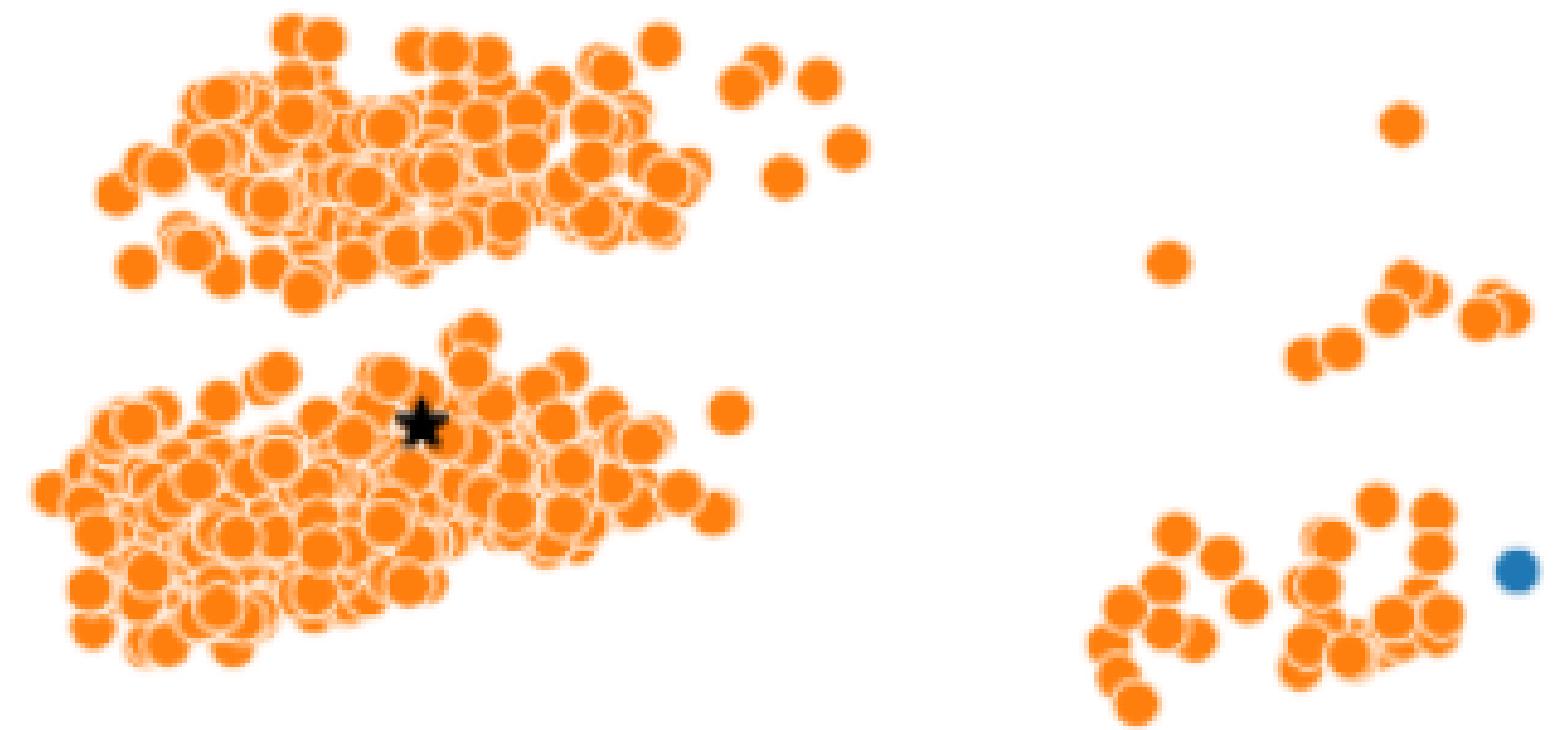
INTERPRETASI CLUSTER 0



Karakteristik Karyawan dalam Cluster 0 adalah sebagai berikut:

- **Persentase resign 21%**
- **Mayoritas divisi Marketing**
- **Didominasi lulusan Marketing dan Life Science**
- **Gender female**
- **Memiliki tingkat jabatan 2 paling banyak**
- **Didominasi jabatan sales executive**
- **Memiliki monthly income 6310.6 (paling besar)**

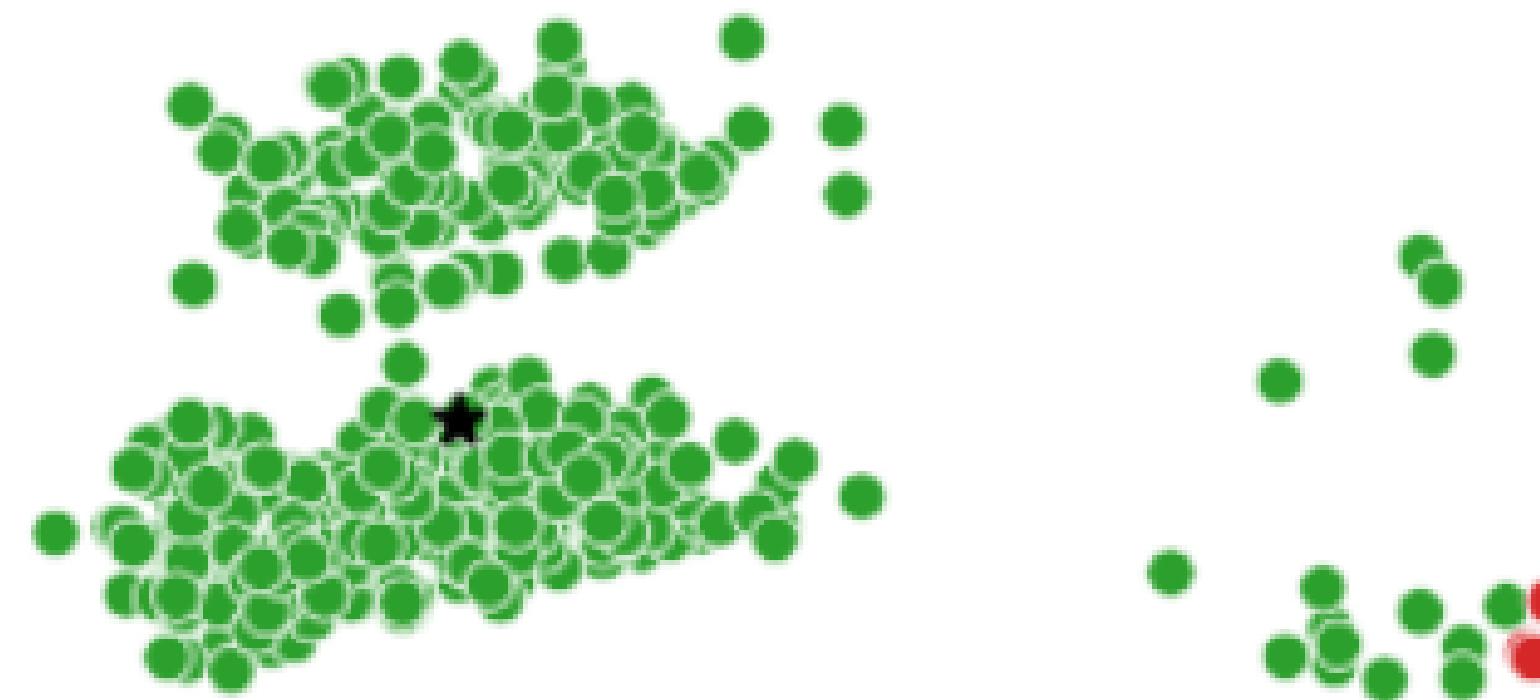
INTERPRETASI CLUSTER 1



Karakteristik Karyawan dalam Cluster 1 adalah sebagai berikut:

- **Persentase resign 15%**
- **Mayoritas divisi Health and Technology tapi memiliki jumlah HR yang besar**
- **Didominasi lulusan Life Science dan Computer Science (memiliki jumlah technical degree yang cukup banyak)**
- **Gender male**
- **Memiliki tingkat jabatan 1 paling banyak**
- **Didominasi jabatan research scientist dan laboratory technician (seimbang)**
- **Memiliki monthly income 4861.31 (paling kecil)**

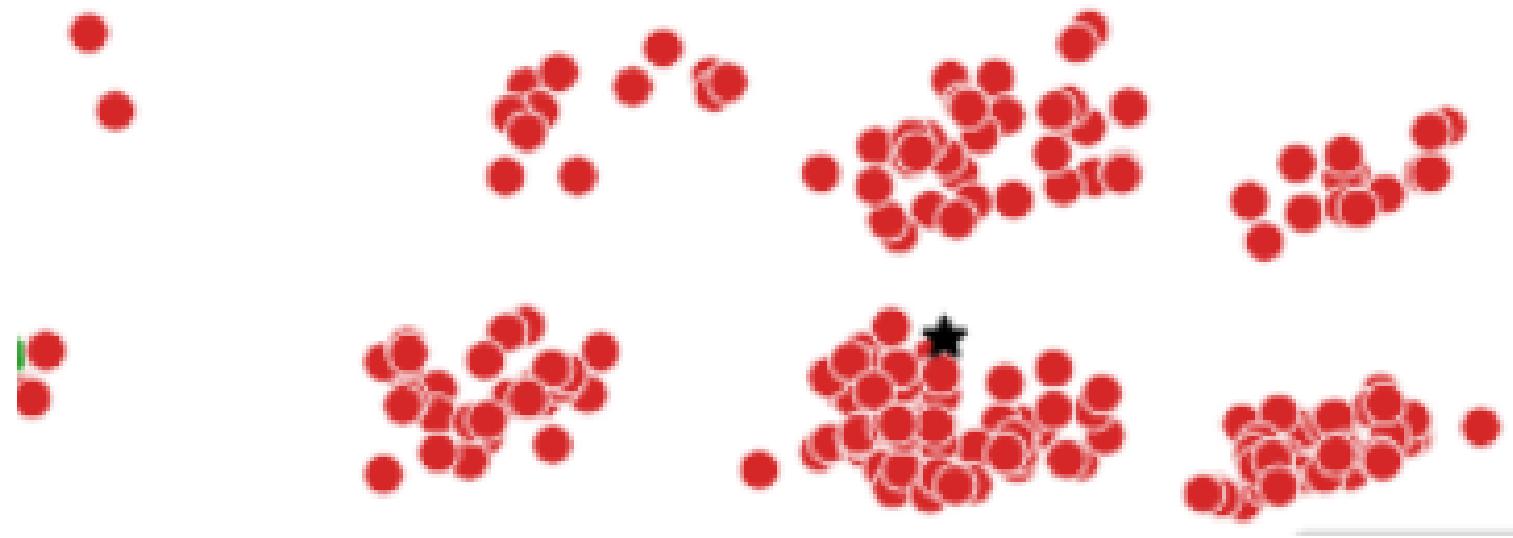
INTERPRETASI CLUSTER 2



Karakteristik Karyawan dalam Cluster 2 adalah sebagai berikut:

- Persentase resign 12%
- Mayoritas divisi Health & Technology
- Didominasi lulusan Life Sciences dan Computer Science
- Gender male
- Memiliki tingkat jabatan 1 paling banyak
- Didominasi jabatan research scientist
- Monthly income 5547.7

INTERPRETASI CLUSTER 3



Karakteristik Karyawan dalam Cluster 3 adalah sebagai berikut:

- **Persentase resign 21%**
- **Mayoritas divisi Marketing**
- **Didominasi oleh lulusan Marketing dan Life Science**
- **Gender female**
- **Memiliki tingkat jabatan 2 paling banyak**
- **Memiliki jabatan yang mirip cluster 1 dengan tambahan role human resource**
- **Memiliki monthly income 6049.63**

ANGGOTA KELOMPOK YUKBISAYUK

Link project:

ristek.link/DeepnoteYukBisaYuk

Bapak/Ibu dosen juga bisa memberikan comment
atau feedback di project kami :D

2006531951 - Andi Afifah Khairunnisa

2006532903 - Muhammad Damar Kusumo

2006533811 - Sultan Fahrezy Syahdwinata

2006596314 - Ekky Aliansyah

TERIMA KASIH