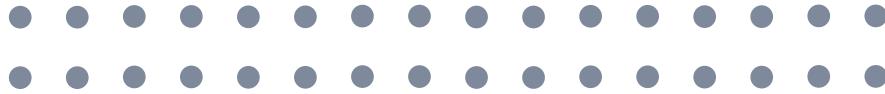


Final Project KASDD Resignation Intention (RES)



Dataset Resignation Intention

xx

Dataset Resignation Intention merupakan dataset dengan dimensi **30 kolom x 1470 baris** yang mendeskripsikan mengenai karyawan dan apakah suatu karyawan (employee) akan **resign** atau **tidak**. Adapun rincian dari kolom dataset resignation intention adalah sebagai berikut:

1. age: umur
2. resign: karyawan berhenti (1) atau tidak (0)
3. division: divisi karyawan
4. home_distance: jarak rumah karyawan ke kantor (km)
5. education: pendidikan terakhir karyawan
 - a. 1: Diploma 3
 - b. 2: Diploma 4
 - c. 3: Sarjana
 - d. 4: Magister
 - e. 5: Doktor
6. employee_id: ID karyawan

Dataset Resignation Intention

xx

- 7. score_environment: nilai kepuasan karyawan terhadap lingkungan kerja a. 1: sangat rendah
b. 2: rendah
c. 3: tinggi
d. 4: sangat tinggi
- 8. major: jurusan
- 9. gender: jenis kelamin
- 10. hourly_rate: bayaran per jam (USD)
- 11. score_contribution: nilai kepuasan kontribusi karyawan terhadap perusahaan a. 1: sangat rendah
b. 2: rendah
c. 3: tinggi
d. 4: sangat tinggi

Dataset Resignation Intention

**

12. job_rank: jabatan karyawan

- a. 1: Entry Level
- b. 2: Associate
- c. 3: Senior Associate
- d. 4: Manager
- e. 5: Senior Manager

13. role: peran karyawan dalam perusahaan

14. score_job_satisfaction: nilai kepuasan karyawan terhadap pekerjaannya

- a. 1: sangat rendah
- b. 2: rendah
- c. 3: tinggi d.
- 4: sangat tinggi

15. marriage_status: status pernikahan

16. monthly_income: gaji per bulan (USD)

17. companies_count: jumlah perusahaan yang pernah karyawan tersebut bekerja

18. underage: di bawah umur atau tidak

19. over_time: bekerja lembur atau tidak

Dataset Resignation Intention

**

21. rate_performance: nilai performa

- a. 1: sangat rendah
- b. 2: rendah
- c. 3: bagus
- d. 4: luar biasa

22. score_work_relationship: nilai kepuasan hubungan kerja karyawan terhadap rekan kerjanya

- a. 1: sangat rendah
- b. 2: rendah
- c. 3: tinggi
- d. 4: sangat tinggi

23. working_hours: jumlah jam kerja (jam)

24. time_total_working: jumlah waktu pengalaman kerja (tahun)

Dataset Resignation Intention

**

25. last_year_training_time: jumlah waktu untuk *training* tahun lalu (hari)
26. score_work_life_balance: nilai *work life balance*
 - a. 1: sangat buruk
 - b. 2: buruk
 - c. 3: bagus
 - d. 4: sangat bagus
27. time_current_company: jumlah waktu kerja di perusahaan saat ini (tahun)
28. time_current_role: jumlah waktu kerja di peran saat ini (tahun)
29. time_last_promotion: jumlah waktu semenjak promosi terakhir (tahun)
30. time_current_manager: jumlah waktu kerja dengan manajer saat ini (tahun)

Dataset Resignation Intention

xx

Overview dataset Resignation Intention

	age	resign	division	home_distance	education	employee_id	score_environment	major	gender	hourly_rate	...	rate_performance	score_work_rel
0	41	Yes	Marketing	1	2	FT1310001	2	Life Sciences	Female	94	3
1	49	No	Health and Technology	8	1	FT1010002	3	Life Sciences	Male	61	4
2	37	Yes	Health and Technology	2	2	FT1110004	4	Other	Male	92	3
3	33	No	Health and Technology	3	4	FT1110005	4	Life Sciences	Female	56	3
4	27	No	Health and Technology	2	1	FT1010007	1	Computer Science	Male	40	3

Dataset Resignation Intention

xx

Data columns (total 30 columns):			
#	Column	Non-Null Count	Dtype
0	age	1470	non-null
1	resign	1470	non-null
2	division	1470	non-null
3	home_distance	1470	non-null
4	education	1470	non-null
5	employee_id	1470	non-null
6	score_environment	1470	non-null
7	major	1470	non-null
8	gender	1470	non-null
9	hourly_rate	1470	non-null
10	score_contribution	1470	non-null
11	job_rank	1470	non-null
12	role	1470	non-null
13	score_job_satisfaction	1470	non-null
14	marriage_status	1470	non-null
15	monthly_income	1470	non-null
16	companies_count	1470	non-null
17	underage	1470	non-null
18	over_time	1470	non-null
19	salary_increment_percentage	1470	non-null
20	rate_performance	1470	non-null
21	score_work_relationship	1470	non-null
22	working_hours	1470	non-null
23	time_total_working	1470	non-null
24	last_year_training_time	1470	non-null
25	score_work_life_balance	1470	non-null
26	time_current_company	1470	non-null
27	time_current_role	1470	non-null
28	time_last_promotion	1470	non-null
29	time_current_manager	1470	non-null

Overview dataset Resignation Intention
(count null)

Dataset Resignation Intention

xx

Overview dataset Resignation Intention

	age	home_distance	education	score_environment	hourly_rate	score_contribution	job_rank	score_job_satisfication
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	9.192517	2.912925	2.721769	65.891156	2.729932	2.063946	2.728571
std	9.135373	8.106864	1.024165	1.093082	20.329428	0.711561	1.106940	1.102846
min	18.000000	1.000000	1.000000	1.000000	30.000000	1.000000	1.000000	1.000000
25%	30.000000	2.000000	2.000000	2.000000	48.000000	2.000000	1.000000	2.000000
50%	36.000000	7.000000	3.000000	3.000000	66.000000	3.000000	2.000000	3.000000
75%	43.000000	14.000000	4.000000	4.000000	83.750000	3.000000	3.000000	4.000000
max	60.000000	29.000000	5.000000	4.000000	100.000000	4.000000	5.000000	4.000000



00

Preprocessing

Memeriksa dan mengatasi nilai null, duplikasi
data, dan outlier



Preprocessing



Data preprocessing adalah teknik penambangan data yang melibatkan transformasi data mentah menjadi format yang dapat dimengerti. Data dunia nyata seringkali data tidak lengkap, tidak konsisten, kurang dalam perilaku atau tren tertentu, dan kemungkinan mengandung banyak kesalahan. Preprocessing adalah metode yang terbukti untuk menyelesaikan masalah tersebut.

Adapun untuk pengolahan dataset awal, kami melakukan beberapa tahap preprocessing, di antaranya adalah:

- **Cek null**
- **Cek duplikasi data**
- **Cek outlier**

Kemudian, kami juga melakukan **feature encoding**, **feature selection**, dan **dimensionality reduction** untuk masing-masing model yang akan lebih dijelaskan di bagian 2, yaitu eksplorasi model.

Preprocessing



Cek Null

```
# functions
def cek_null(df):
    col_na = df.isnull().sum().sort_values(ascending=True)
    percent = col_na / len(df)

    missing_data = pd.concat([col_na, percent], axis=1, keys=['Total', 'Percent'])

    if (missing_data[missing_data['Total'] > 0].shape[0] == 0):
        print("Tidak ditemukan missing value pada dataset")

    else:
        print(missing_data[missing_data['Total'] > 0])

# null
cek_null(df)

Tidak ditemukan missing value pada dataset
```

Dengan menjalankan fungsi **cek_null()** didapatkan bahwa **tidak ada** missing value pada dataset

Cek Duplikasi Data

```
# duplikasi
print("Jumlah duplikasi data : " + str(df.duplicated().sum()))

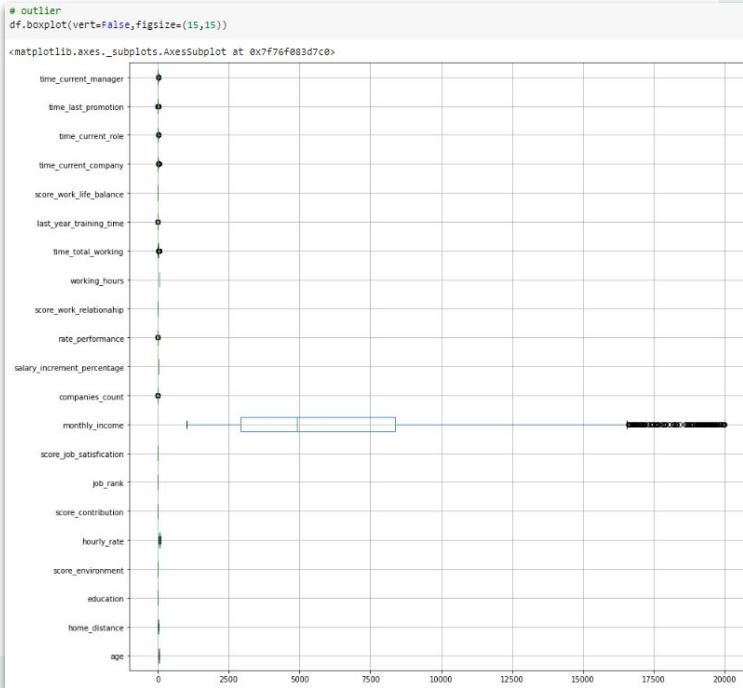
Jumlah duplikasi data : 0
```

Tidak ada duplikasi data pada dataset



Cek Outlier

Boxplot menunjukkan **adanya outlier** pada data



```
# menampilkan persentase outlier
jumlah_Outlier = []

Q3 = df_clean.quantile(0.75)
Q1 = df_clean.quantile(0.25)
IQR = Q3-Q1

for col in df_clean.select_dtypes(np.number).columns:
    outliers = ((df_clean[col] < (Q1[col] - 1.5 * IQR[col])) | (df_clean[col] > (Q3[col] + 1.5 * IQR[col])).sum()
    jumlah_Outlier.append([col, outliers, outliers/len(df_clean)])

indexOutput = list(range(0, len(jumlah_Outlier)))

pd.DataFrame(jumlah_Outlier, columns=['Column', 'Outlier','persentase'],index=indexOutput).sort_values(by=['persentase'], ascending = False,ignore_index=True)
```

	Column	Outlier	persentase
0	last_year_training_time	238	0.161905
1	rate_performance	226	0.153741
2	monthly_income	114	0.077551
3	time_last_promotion	107	0.072789
4	time_current_company	104	0.070748
5	time_total_working	63	0.042857
6	companies_count	52	0.035374
7	time_current_role	21	0.014286
8	time_current_manager	14	0.009524
9	score_work_relationship	0	0.000000
10	score_work_life_balance	0	0.000000
11	working_hours	0	0.000000
12	age	0	0.000000
13	home_distance	0	0.000000
14	score_job_satisfaction	0	0.000000
15	job_rank	0	0.000000
16	score_contribution	0	0.000000
17	hourly_rate	0	0.000000
18	score_environment	0	0.000000
19	education	0	0.000000
20	salary_increment_percentage	0	0.000000

Setelah melihat **persentase data outlier tersebut terhadap keseluruhan data.**

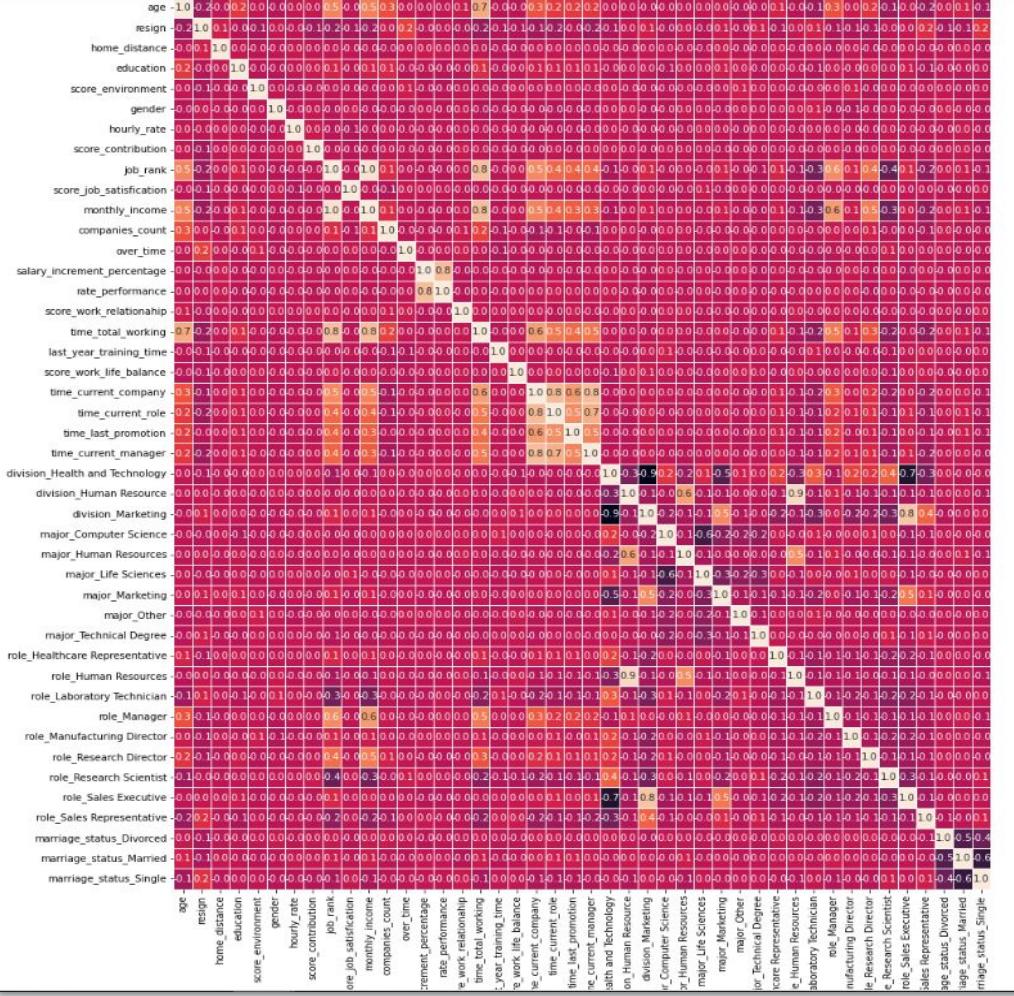
Outlier pada data **dibiarkan** karena dianggap akan **membantu** dalam penarikan kesimpulan.



01

Eksplorasi





Heatmap Korelasi

Untuk melihat korelasi antar atribut di dataset yang telah dilakukan preprocessing sebelumnya, kami menggunakan data visualization dengan heatmap.

Eksplorasi

1a. Visualisasikan karakteristik karyawan yang resign dari perusahaan tersebut!

Menurut kami, **berdasarkan** nilai **korelasi** antar atribut terhadap target (**resign**),

karakteristik karyawan yang resign dapat kita lihat dari **atribut-atribut berikut**:

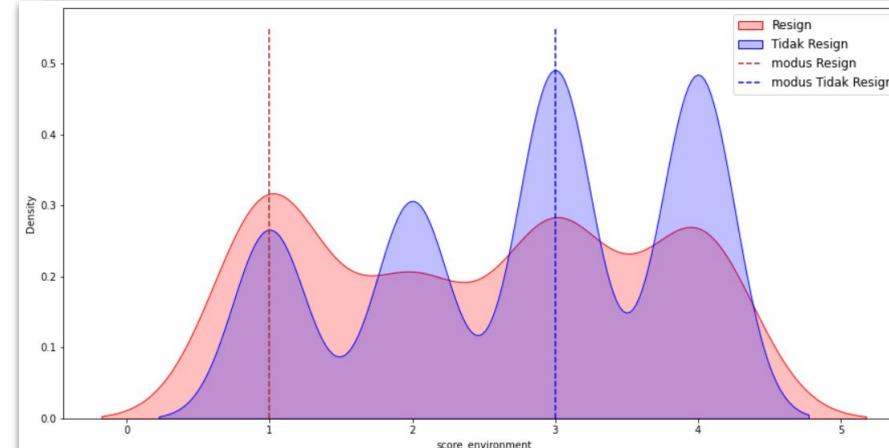
- score_environment,
- score_contribution,
- job_rank,
- score_job_satisfaction,
- over_time,
- score_work_relationship,
- time_total_working,
- time_current_role,
- time_last_promotion,
- time_current_manager,
- role_Sales Executive,
- marriage_status_Married.



score_environment

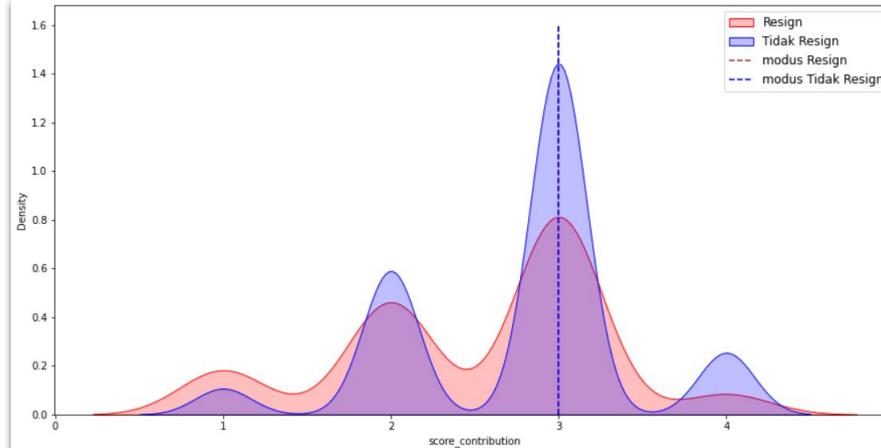
Persebaran data karyawan yang resign berdasarkan nilai kepuasan karyawan terhadap lingkungan kerjanya berada di kategori sangat rendah.

Sedangkan untuk persebaran data karyawan yang tidak resign berada di kategori tinggi.



score_contribution

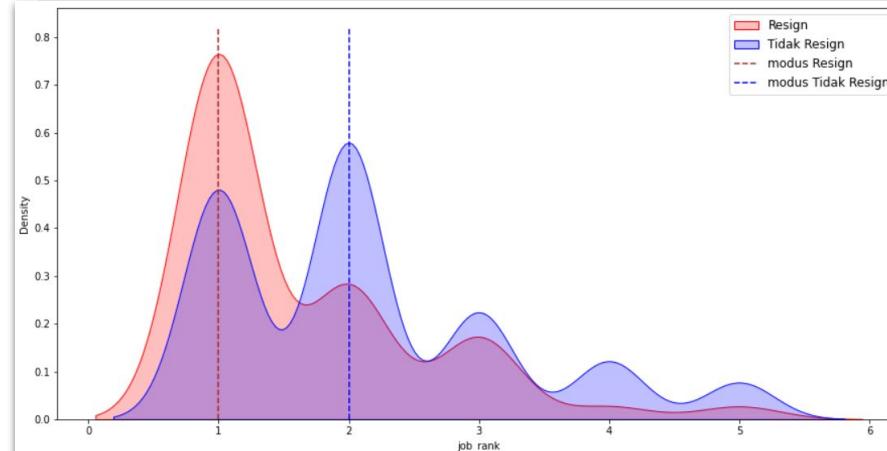
Persebaran data karyawan yang resign dan tidak resign berdasarkan nilai kepuasan kontribusi karyawan terhadap perusahaan keduanya berada di kategori tinggi.



job_rank

Persebaran data karyawan yang resign berdasarkan jabatannya berada di kategori Entry Level.

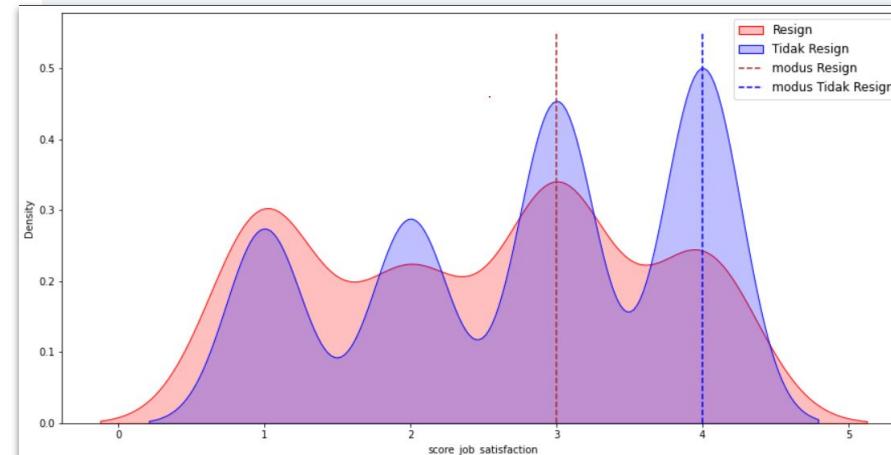
Sedangkan untuk persebaran data karyawan yang tidak resign berada di kategori Associate.



score_job_satisfaction

Persebaran data karyawan yang resign berdasarkan nilai kepuasan karyawan terhadap pekerjaannya berada di kategori tinggi.

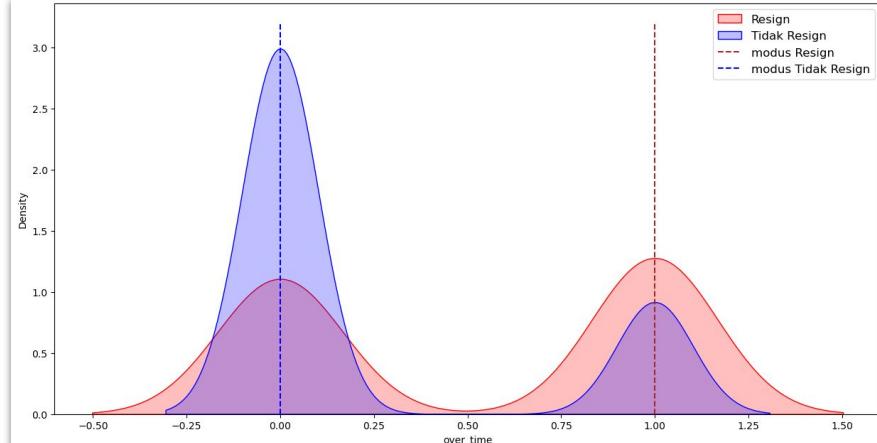
Sedangkan untuk persebaran data karyawan yang tidak resign berada di kategori sangat tinggi.



over_time

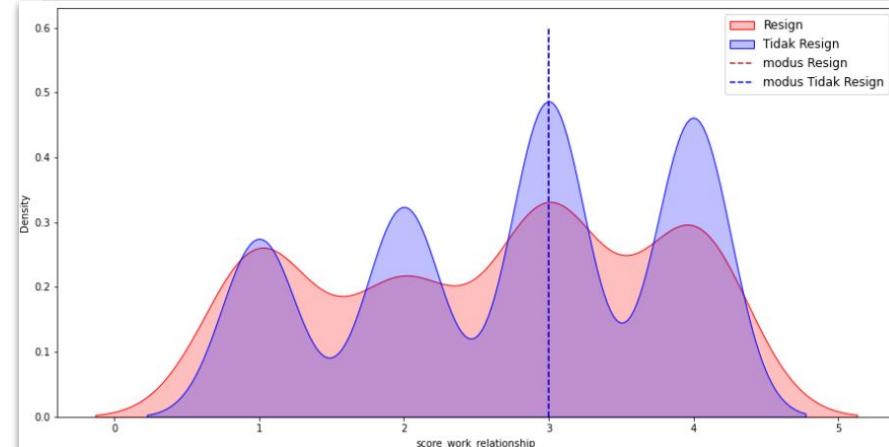
Persebaran data karyawan yang resign berdasarkan bekerja lembur atau tidaknya karyawan paling banyak bekerja lembur.

Sedangkan untuk persebaran data karyawan yang tidak resign sebagian besar tidak bekerja lembur



score_work_relationship

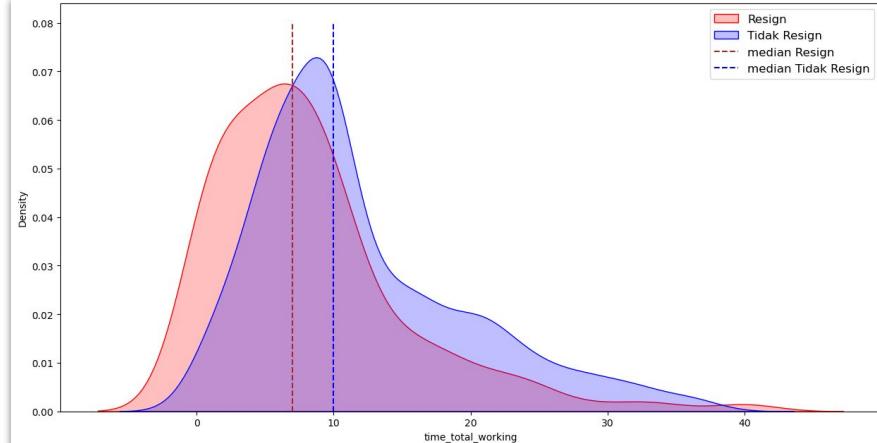
Persebaran data karyawan yang resign dan tidak resign berdasarkan kepuasan hubungan kerja karyawan terhadap rekan kerjanya keduanya berada di kategori tinggi.



time_total_working

Persebaran data karyawan yang resign berdasarkan pengalaman kerjanya berada di sekitar 7 tahun.

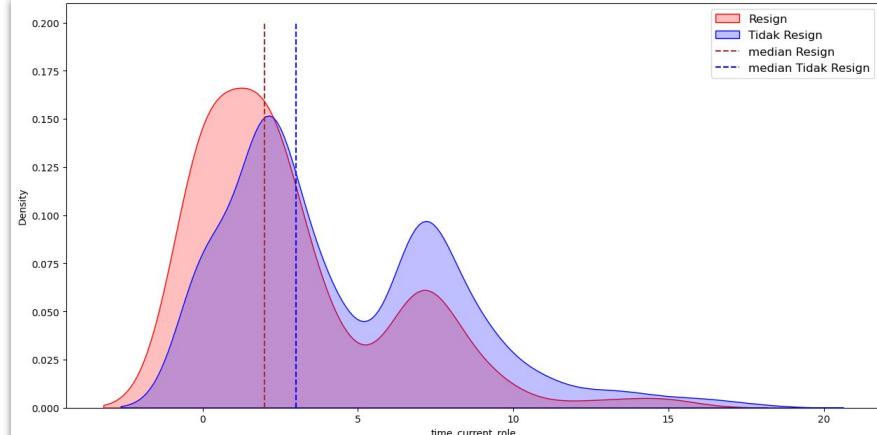
Sedangkan untuk persebaran data karyawan yang tidak resign berada di sekitar 10 tahun.



time_current_role

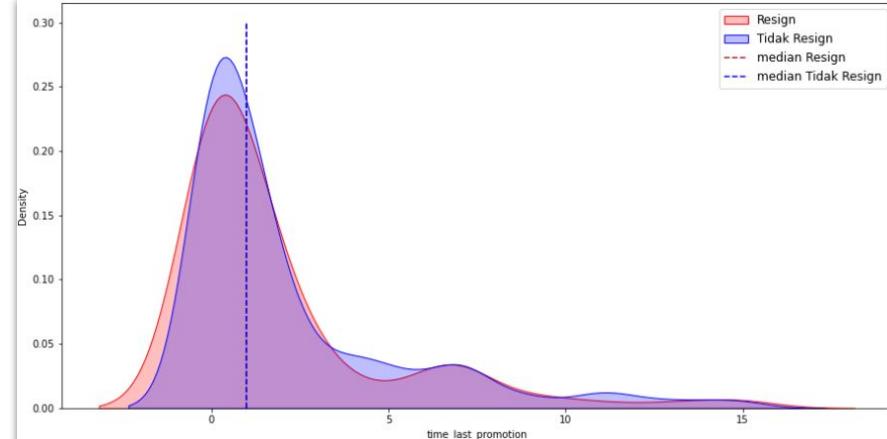
Persebaran data karyawan yang resign berdasarkan lama kerja di peran saat ini berada di sekitar 2 tahun.

Sedangkan untuk persebaran data karyawan yang tidak resign berada di sekitar 3 tahun.



time_last_promotion

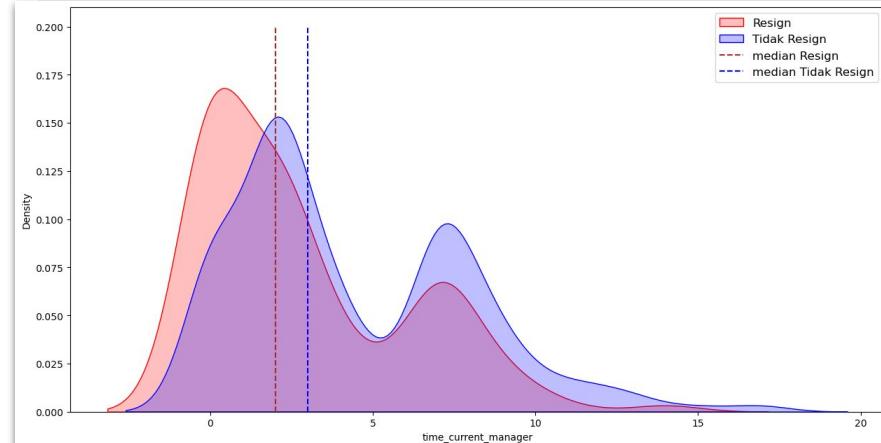
Persebaran data karyawan yang resign dan tidak resign berdasarkan waktu semenjak promosi terakhir keduanya berada di sekitar 1 tahun.



time_current_manager

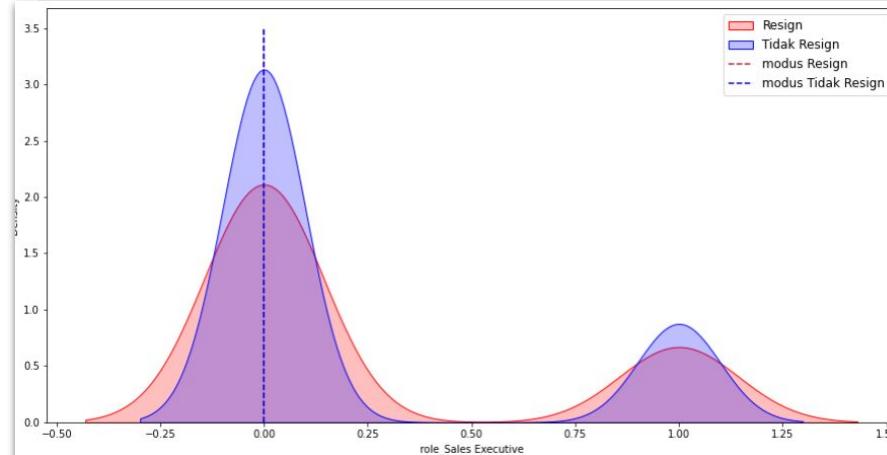
Persebaran data karyawan yang resign berdasarkan lama kerja dengan manajer saat ini berada di sekitar 2 tahun.

Sedangkan untuk persebaran data karyawan yang tidak resign berada di sekitar 3 tahun.



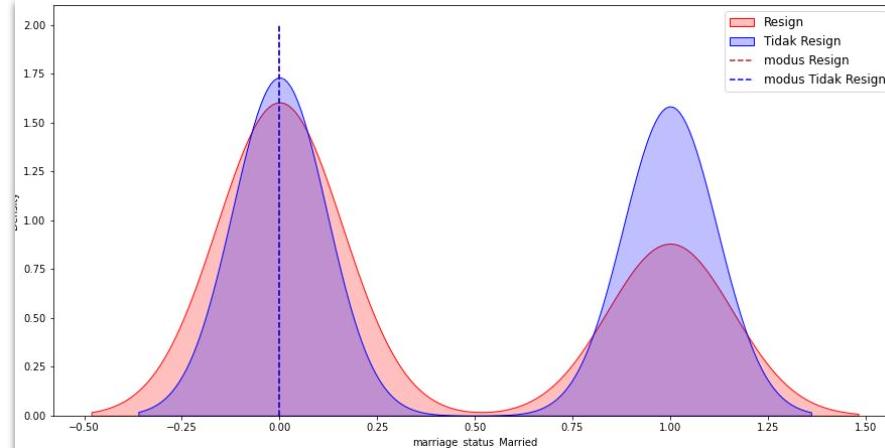
role_Sales Executive

Persebaran data karyawan yang resign dan tidak resign berdasarkan role Sales Executive keduanya paling banyak merupakan bukan role Sales Executive.



marriage_status_Married

Persebaran data karyawan yang resign dan tidak resign berdasarkan marriage status Married keduanya paling banyak memiliki status sedang tidak menikah.



Kesimpulan Karakteristik Karyawan ✘✘

Resign

Karyawan yang resign memiliki karakteristik sebagai berikut,

- Score_environment = 1 (Sangat rendah)
- Score_contribution = 3 (Tinggi)
- Job_rank = 1 (Entry Level)
- Score_job_satisfaction = 3 (Tinggi)
- Over_time = 1 (Bekerja lembur)
- Score_work_relationship = 3 (Tinggi)
- Time_total_working = 7 Tahun
- Time_current_role = 2 Tahun
- Time_last_promotion = 1 Tahun
- Time_current_manager = 2 Tahun
- Role_Sales Executive = Bukan Sales Executive
- Marriage_Status Married = Tidak Sedang Menikah

Tidak Resign

Karyawan yang tidak resign memiliki karakteristik sebagai berikut,

- Score_environment = 3 (Tinggi)
- Score_contribution = 3 (Tinggi)
- Job_rank = 2 (Associate)
- Score_job_satisfaction = 4 (Sangat tinggi)
- Over_time = 0 (Tidak Bekerja lembur)
- Score_work_relationship = 3 (Tinggi)
- Time_total_working = 10 Tahun
- Time_current_role = 3 Tahun
- Time_last_promotion = 1 Tahun
- Time_current_manager = 3 Tahun
- Role_Sales Executive = Bukan Sales Executive
- Marriage_Status Married = Tidak Sedang Menikah



Eksplorasi

1b. Apakah karyawan memilih untuk resign setelah mendapatkan promosi?

Untuk mencari apakah karyawan akan **resign setelah mendapatkan promosi**, kami mendefinisikan karyawan yang pernah dipromosikan dengan menggunakan atribut **job_rank**, **time_current_company**, dan **time_last_promotion**. Atribut tersebut kami pilih secara intuitif untuk menarik kesimpulan.



- **time_current_company**: jumlah waktu kerja di perusahaan saat ini (tahun)
- **job_rank**: jabatan karyawan
- **time_last_promotion**: jumlah waktu semenjak promosi terakhir (tahun)

Dari penjelasan ketiga atribut tersebut, secara intuitif kami memilih atribut tersebut.

Eksplorasi

1b. Apakah karyawan memilih untuk resign setelah mendapatkan promosi?

Adapun sebelumnya kami menggunakan **3 asumsi** yang tidak dijelaskan dalam dataset:

- **Asumsi 1:** Karyawan yang telah **dipromosi setidaknya** telah bekerja **di atas satu tahun** (`time_current_company`).
- **Asumsi 2:** Untuk edge case di mana seorang karyawan memiliki `job_rank > 1` dan `time_current_company = 0`, karyawan tersebut dari awal sudah berada di jabatannya tersebut sehingga bisa dibilang **belum** pernah **dipromosikan di current company**.
- **Asumsi 3:** Dengan mengambil `time_last_promotion <=1`, pertanyaan “**setelah mendapatkan promosi**” adalah **baru** saja mendapatkan promosi, sehingga kami mengambil karyawan yang **baru** saja **dipromosikan** dalam **0-1 tahun**.



Eksplorasi

1b. Apakah karyawan memilih untuk resign setelah mendapatkan promosi?

Dengan ketiga asumsi tersebut, kami menggunakan algoritma untuk melakukan filtering untuk mendapatkan kesimpulan apakah karyawan memilih untuk resign setelah mendapatkan promosi. Kemudian, catatan tambahan yang perlu diperhatikan adalah **job_rank > 1** artinya **pernah promosi**, atau **job_rank = Entry Level tidak pernah promosi** karena dianggap adalah karyawan dengan level paling rendah yang belum pernah dipromosikan.



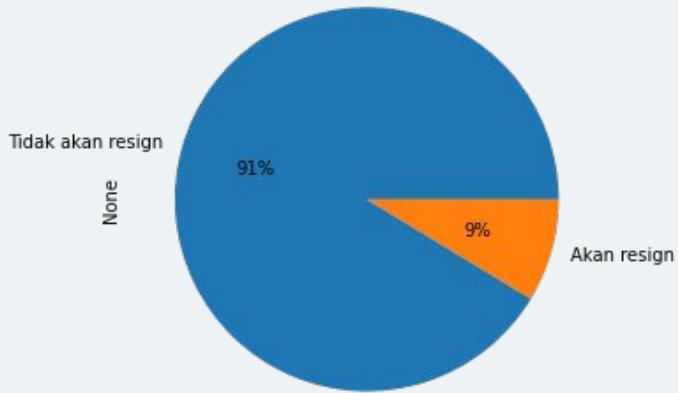
Visualisasi

Kode dan Pie Chart

```
[ ] df_norookie = df[(df['job_rank'] > 1)]
df_senior = df_norookie[(df_norookie['time_current_company'] > 0)]
df_last_promoted = df_senior[(df_senior['time_last_promotion'] <= 1)]
will_resign = len(df_last_promoted[(df_last_promoted['resign'] == "Yes")])
wont_resign = len(df_last_promoted[(df_last_promoted['resign'] == "No")])
print("Banyak karyawan akan resign setelah mendapat promosi:", will_resign)
print("Banyak karyawan tidak akan resign setelah mendapat promosi:", wont_resign)
```

Banyak karyawan akan resign setelah mendapat promosi: 45
Banyak karyawan tidak akan resign setelah mendapat promosi: 474

Berdasarkan percobaan kami, dari **519** karyawan yang mendapatkan **promosi**, sebanyak **45** karyawan (**9%**) **resign** dan sebanyak **474** karyawan (**91%**) **tidak resign**.

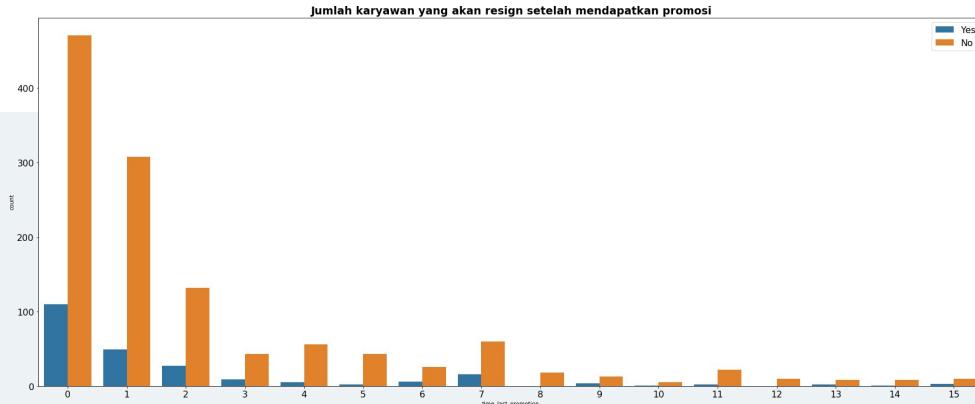


Karena **majoritas** karyawan **tidak akan resign** setelah mendapatkan promosi, dapat diambil **kesimpulan** bahwa karyawan cenderung **tidak akan resign** setelah mendapatkan promosi.

Eksplorasi

1b. Apakah karyawan memilih untuk resign setelah mendapatkan promosi?

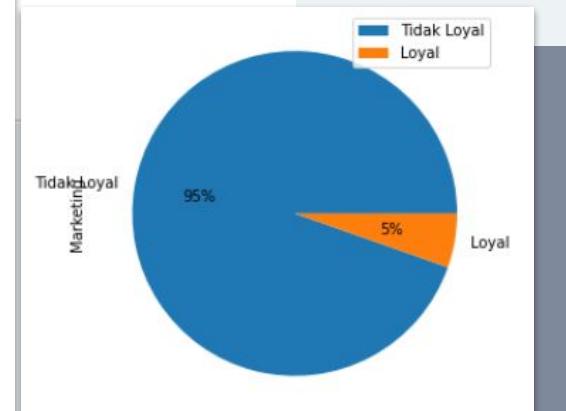
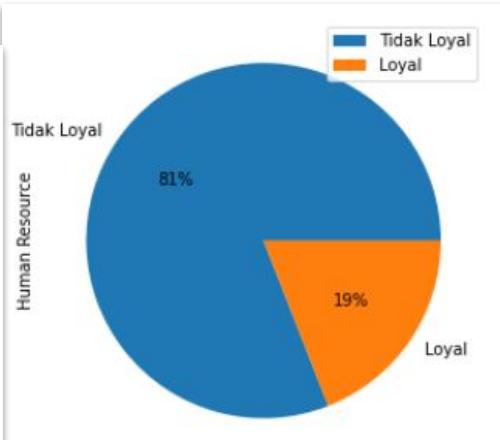
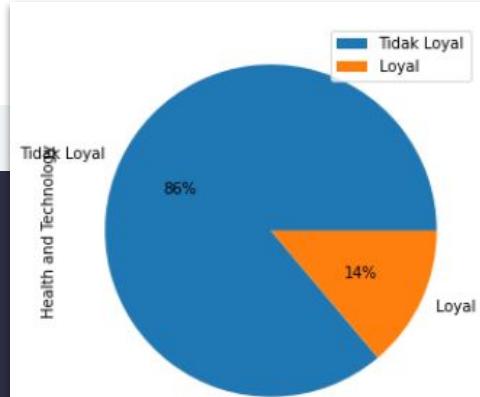
Namun, apabila dilihat dari grafik di bawah, berapa lama pun waktu seorang karyawan baru mendapatkan promosi (time_last_promotion), jumlah karyawan yang tidak akan resign selalu lebih banyak dibanding karyawan yang akan resign, sehingga informasi hubungan antara karyawan akan resign atau tidak dengan time_last_promotion **kurang dan tidak begitu berpengaruh**.



Eksplorasi

1c. Departemen manakah yang memiliki karyawan loyal paling banyak?

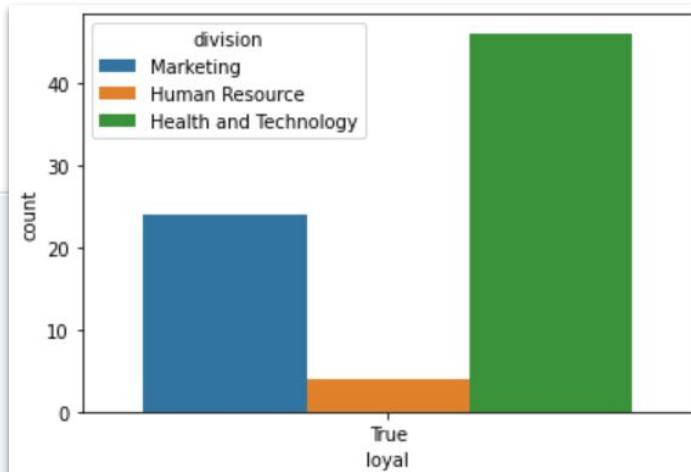
Akan tetapi, perlu dilihat bahwa divisi **Health and Technology** merupakan divisi dengan persentase karyawan tidak loyal yang banyak juga, hal ini disebabkan karena banyaknya jumlah karyawan pada divisi **Health and Technology**.



Eksplorasi

1c. Departemen manakah yang memiliki karyawan loyal paling banyak?

Kami mengambil karyawan dengan **companies_count** 0 beserta **mean** dari **time_current_company** tiap divisi untuk mendapatkan karyawan yang Loyal. Dari hasil tersebut, kami mendapatkan bahwa divisi **Health and Technology** merupakan divisi dengan karyawan loyal terbanyak



Eksplorasi

1d. Analisis Korelasi Atribut

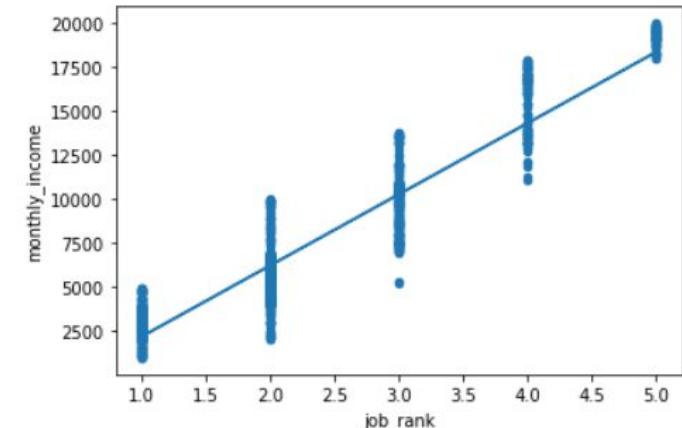
Atribut yang kami anggap berkorelasi adalah sebuah atribut yang memiliki korelasi lebih besar dari 0.7 atau lebih kecil dari -0.7, atribut dibawah ini adalah atribut yang kami anggap berkorelasi

job_rank	monthly_income	0.950300
job_rank	time_total_working	0.782208
monthly_income	time_total_working	0.772893
salary_increment_percentage	rate_performance	0.773550
time_current_company	time_current_role	0.758754
time_current_company	time_current_manager	0.769212
time_current_role	time_current_manager	0.714365
division_Health and Technology	division_Marketing	-0.906818
division_Health and Technology	role_Sales Executive	-0.733497
division_Human Resource	role_Human Resources	0.904983
division_Marketing	role_Sales Executive	0.808869



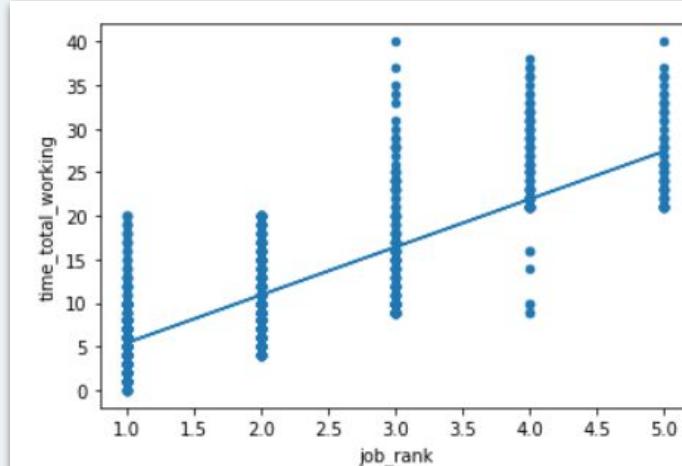
Job_rank, monthly_income

Dapat dilihat bahwa atribut **job_rank** berkorelasi positif dengan **monthly_income**. Artinya semakin tinggi **job_rank** nya akan semakin tinggi **monthly_income** nya. Akan tetapi dapat dilihat bahwa ada karyawan yang memiliki job rank berbeda namun memiliki pendapatan yang sama



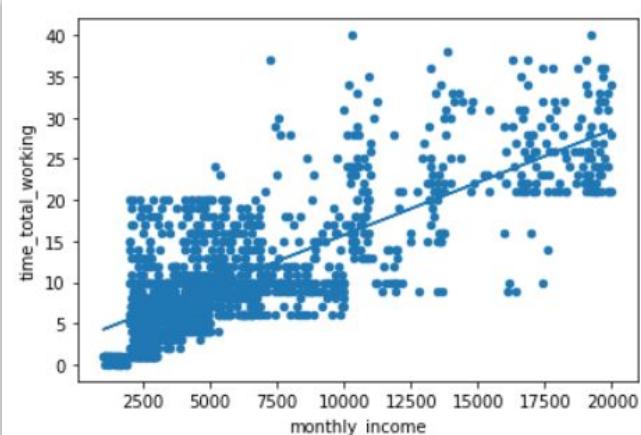
Job_rank, time_total_working

Dapat dilihat bahwa atribut **job_rank** berkorelasi positif dengan **time_total_working**. Artinya semakin tinggi **job_rank** nya akan semakin tinggi **time_total_working** nya. Tetapi, seperti sebelumnya, ada **job_rank** yang berbeda dengan **time_total_working** yang sama pada setiap **job_rank**



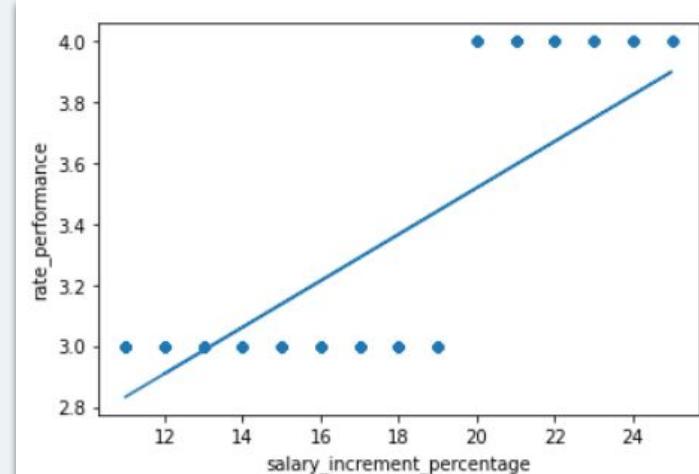
Monthly_income, time_total_working

Dapat dilihat bahwa atribut **monthly_income** berkorelasi positif dengan **time_total_working**. Artinya semakin tinggi **monthly_income** nya akan semakin tinggi **time_total_working** nya. Selain itu, meskipun kedua atribut ini memiliki korelasi positif, data yang dimiliki cukup tersebar.



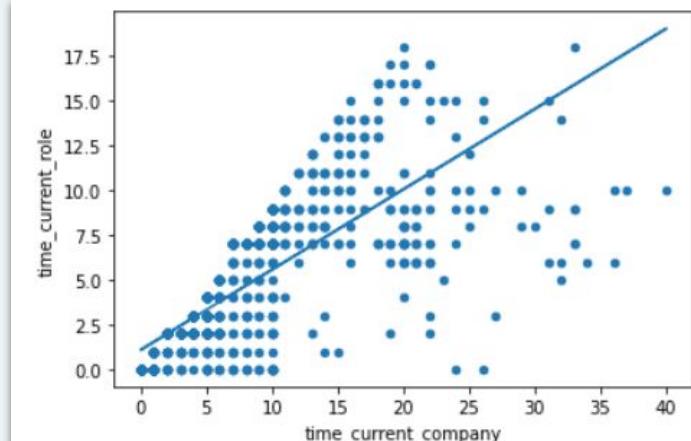
Salary_increment_percentage, rate_performance

Dapat dilihat bahwa atribut **salary_increment_percentage** berkorelasi positif dengan **rate_performance**. Artinya semakin tinggi **salary_increment_percentage** nya akan semakin tinggi **rate_performance** nya. Selain itu, dapat dilihat bahwa karyawan dengan rating performa 3 memiliki persentase kenaikan gaji dibawah 20%, sedangkan yang memiliki performa 4, memiliki persentase kenaikan gaji diatas 20%



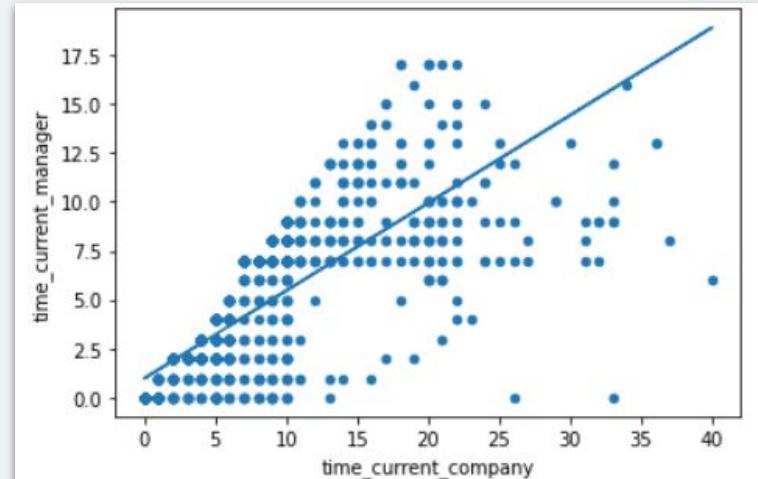
Time_current_company, time_current_role

Dapat dilihat bahwa atribut **time_current_company** berkorelasi positif dengan **time_current_role**. Artinya semakin tinggi **time_current_company** nya akan semakin tinggi **time_current_role** nya. Selain itu dapat dilihat bahwa rata-rata sedikit karyawan yang pindah role ketika sudah lebih dari 10 tahun di perusahaan tersebut. Kebanyakan karyawan pindah role pada sebelum 10 tahun mereka bekerja di perusahaan tersebut



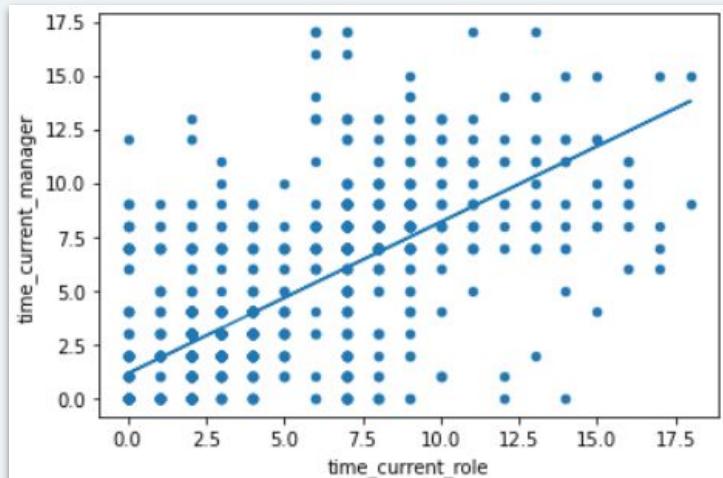
Time_current_company, time_current_manager

Dapat dilihat bahwa atribut **time_current_company** berkorelasi positif dengan **time_current_manager**. Artinya semakin tinggi **time_current_company** nya akan semakin tinggi **time_current_manager** nya. Sama seperti time_current role, kebanyakan karyawan yang berganti manager belum berada di perusahaan lebih dari 10 tahun



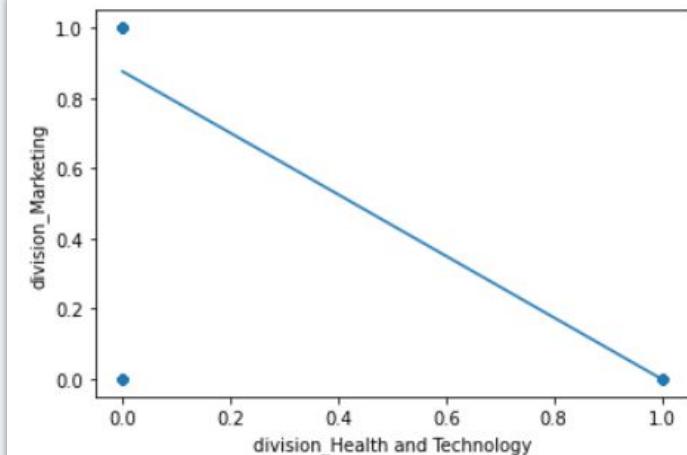
Time_current_role, time_current_manager

Dapat dilihat bahwa atribut **time_current_role** berkorelasi positif dengan **time_current_manager**. Artinya semakin tinggi **time_current_role** nya akan semakin tinggi **time_current_manager** nya. Akan tetapi, persebaran datanya cukup beragam, namun tetap berkorelasi positif karena data yang menunjukkan korelasi positif lebih banyak



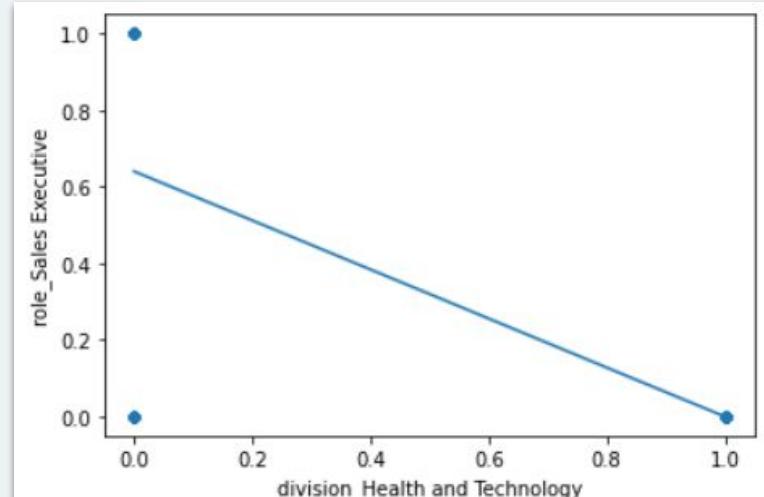
Division_Health and Technology, division_Marketing

Dapat dilihat bahwa atribut **division_Health and Technology** berkorelasi negatif dengan **division_Marketing**. Artinya semakin tinggi **division_Health and Technology** nya akan semakin rendah **division_Marketing** nya. Hal ini masuk akal karena pada umumnya seseorang tidak bisa berada pada dua divisi yang sama.



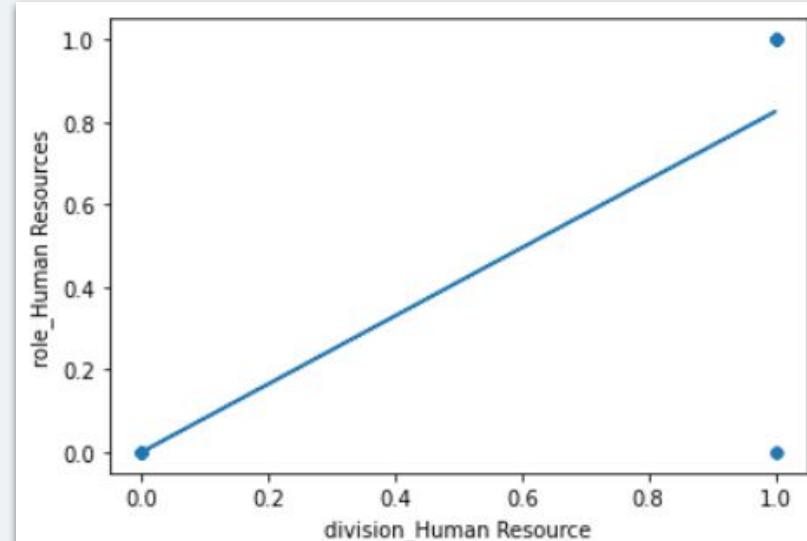
Division_Health and Technology, role_Sales Executive

Dapat dilihat bahwa atribut **division_Health and Technology** berkorelasi negatif dengan **role_Sales Executive**. Artinya semakin tinggi **division_Health and Technology** nya akan semakin rendah **role_Sales Executive** nya. Hal tersebut masuk akal karena pada divisi **Health dan Technology** tidak terdapat role **Sales Executive** yang merupakan role pada divisi **Marketing**



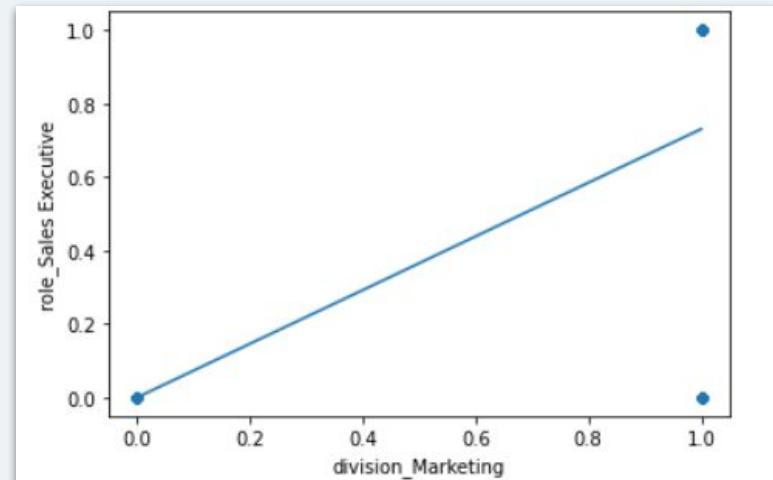
Division_Human Resource, role_Human Resources

Dapat dilihat bahwa atribut **division_Human Resource** berkorelasi positif dengan **role_Human Resources**. Artinya semakin tinggi **division_Human Resource** nya akan semakin tinggi **role_Human Resources** nya. Hal ini wajar karena role HR merupakan salah satu role pada divisi HR.



Division_Marketing, role_Sales Executive

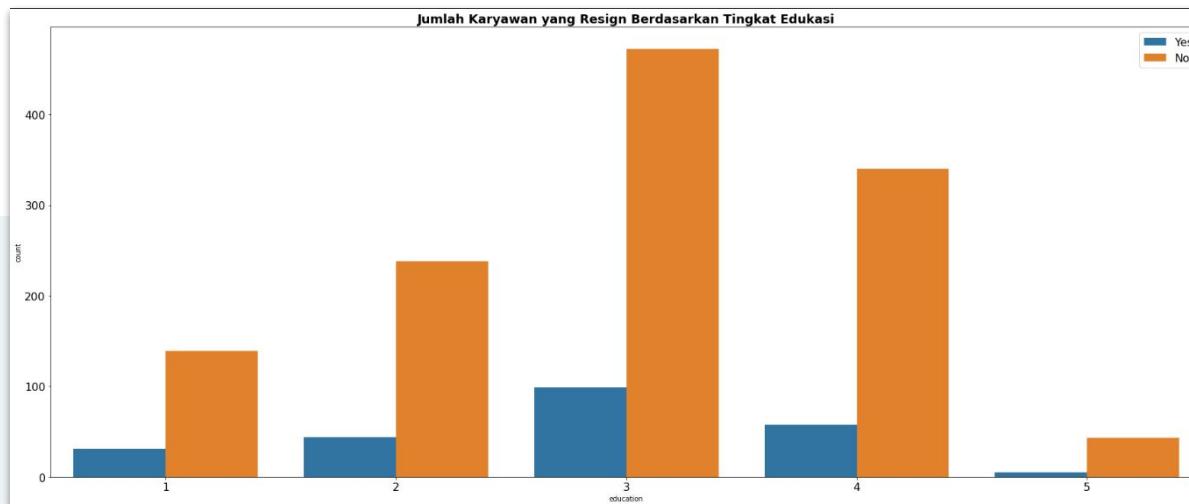
Dapat dilihat bahwa atribut **division_Marketing** berkorelasi positif dengan **role_Sales Executive**. Artinya semakin tinggi **division_Marketing** nya akan semakin tinggi **role_Sales Executive** nya. Hal ini wajar karena salah satu role pada divisi Marketing adalah Sales Executive



Eksplorasi

Apakah Karyawan dengan Tingkat Edukasi Tinggi Cenderung untuk Resign?

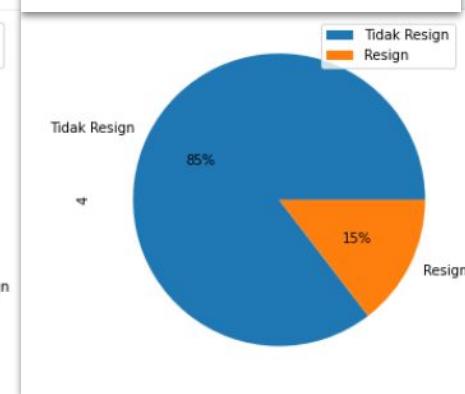
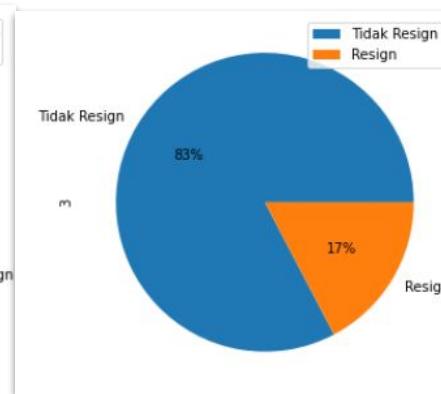
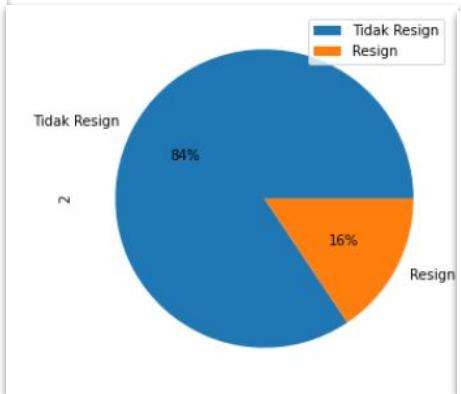
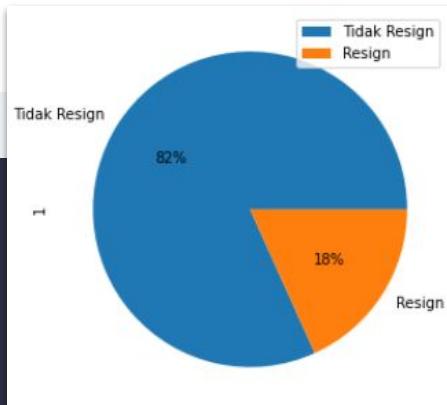
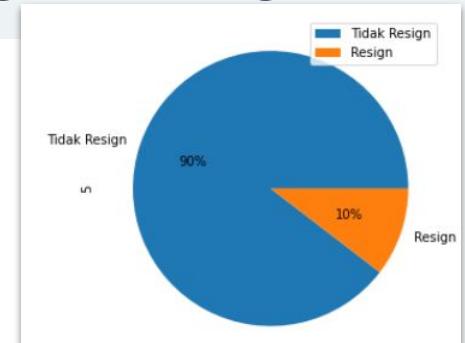
Dapat dilihat bahwa jumlah karyawan yang resign lebih banyak pada tingkat edukasi 3 yaitu karyawan dengan lulusan **Sarjana**. Sedangkan karyawan dengan tingkat edukasi tertinggi yaitu 5, atau bisa dibilang karyawan dengan lulusan **Doktor** lebih sedikit yang **Resign**



Eksplorasi

Apakah Karyawan dengan Tingkat Edukasi Tinggi Cenderung untuk Resign?

Jika melihat pada grafik, karyawan dengan tingkat edukasi tertinggi lah yang memiliki jumlah karyawan resign yang lebih sedikit dengan yang lain, dengan persentase 10% dari total karyawan pada tingkat edukasi tersebut



Kesimpulan

Apakah Karyawan dengan Tingkat Edukasi Tinggi Cenderung untuk Resign?

- Karyawan dengan Tingkat Edukasi Tinggi cenderung dengan tidak resign
- Karyawan yang memiliki kecenderungan untuk resign adalah karyawan dengan tingkat edukasi terendah



Eksplorasi

Apakah karyawan resign karena tidak senang bekerja di perusahaan tersebut?

Kesenangan karyawan terhadap perusahaan dapat kita lihat dari seberapa baik lingkungan kerjanya(score_environment), seberapa puas dia terhadap pekerjaannya(score_job_satisfaction) , dan seberapa baik hubungan dia dan rekan kerjanya(score_work_relationship). Kami menetapkan nilai 3 sebagai Threshold kesenangan karyawan terhadap perusahaan. Untuk nilai yang kurang dari 3 kami anggap kurang senang dengan perusahaan, dan untuk yang lebih dari sama dengan 3 kami anggap senang.

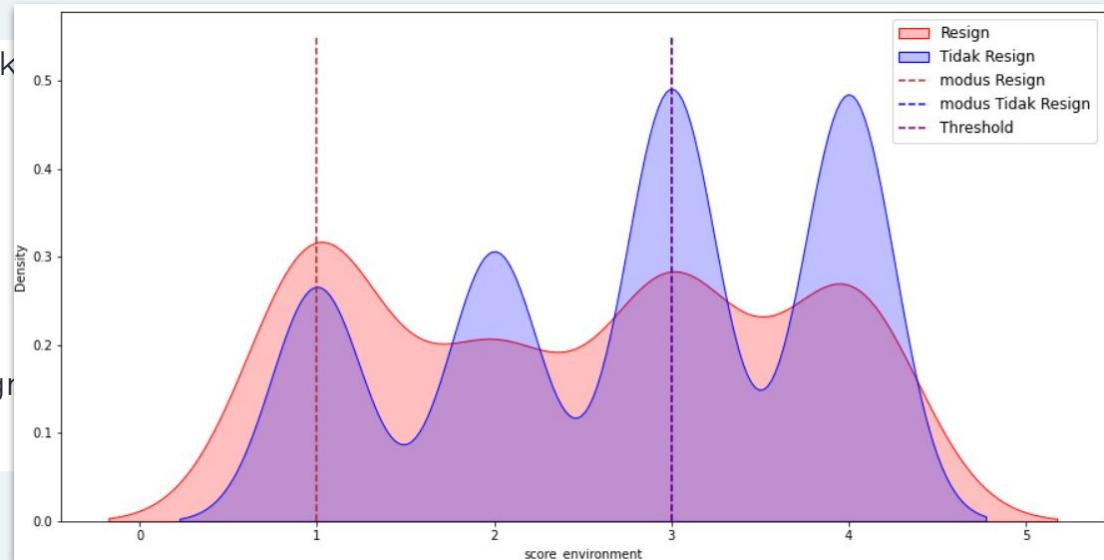


Eksplorasi

Apakah karyawan resign karena tidak senang bekerja di perusahaan tersebut?

Berdasarkan grafik di samping, nilai terbanyak untuk karyawan yang Resign berdasarkan score_environmentnya yaitu pada nilai 1 atau sangat rendah.

Sedangkan untuk karyawan yang Tidak Resign nilai 3 sebagai nilai terbanyak.

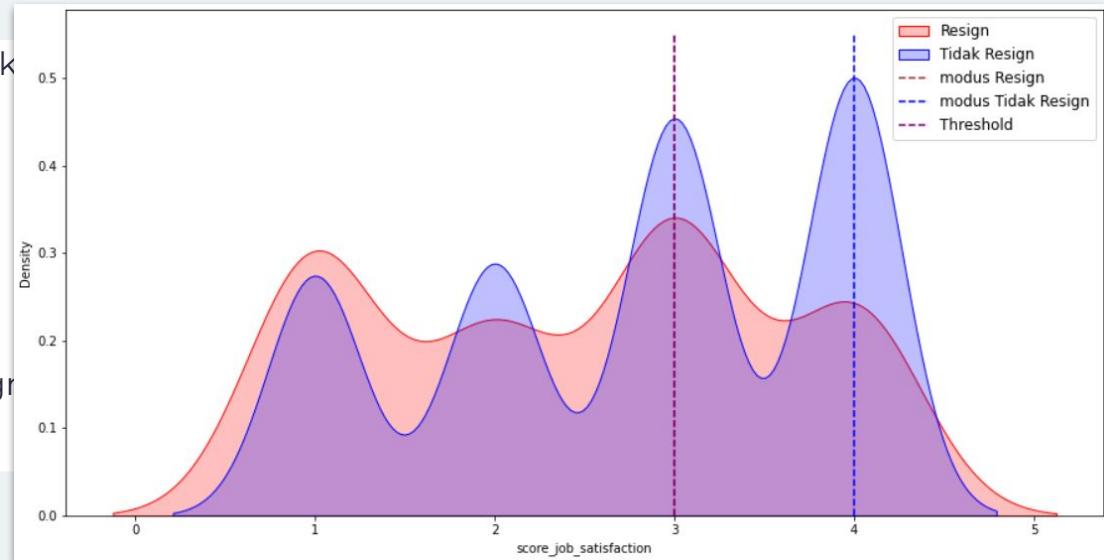


Eksplorasi

Apakah karyawan resign karena tidak senang bekerja di perusahaan tersebut?

Berdasarkan grafik di samping, nilai terbanyak untuk karyawan yang Resign berdasarkan score_job_satisfactionnya yaitu pada nilai 3 atau tinggi.

Sedangkan untuk karyawan yang Tidak Resign nilai 4 sebagai nilai terbanyak.

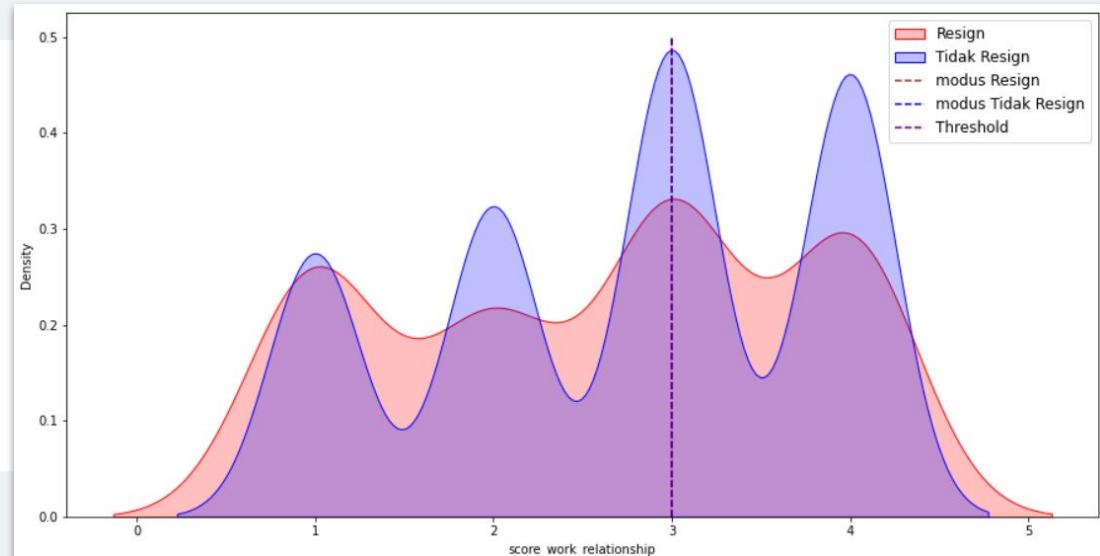


Eksplorasi

Apakah karyawan resign karena tidak senang bekerja di perusahaan tersebut?

Berdasarkan grafik di samping, nilai terbanyak untuk karyawan yang Resign berdasarkan score_work_relationshipnya yaitu pada nilai 3 atau tinggi.

Sedangkan untuk karyawan yang Tidak Resign nilai 3 sebagai nilai terbanyak.

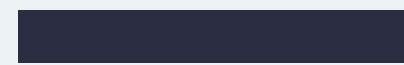


Eksplorasi

Apakah karyawan resign karena tidak senang bekerja di perusahaan tersebut?

Berdasarkan 3 grafik sebelumnya, resign atau tidaknya karyawan dipengaruhi oleh kepuasan karyawan terhadap lingkungannya dan kepuasan karyawan terhadap pekerjaannya, tetapi tidak dipengaruhi oleh kepuasan hubungan kerja karyawan dengan rekannya.





02

Model Training



Model Training Klasifikasi

2a. Prediksi karyawan akan **resign** atau tidak di perusahaan tersebut

- Metode yang digunakan adalah **classification** karena label bersifat **diskret** (**resign/tidak**).
- **Normalisasi** dengan **Standard Scaler** sebelum melakukan **data training/fitting**.
- Metode klasifikasi yang digunakan pada prediksi ini adalah **Logistic Regression** karena model di training terhadap nilai **resign=0** atau **resign=1**.
- Sebelum melakukan model training, data di **oversampling** terlebih dahulu karena jumlah **resign** tersedia dalam jumlah yang **sedikit** (sekitar **15%** dari keseluruhan data). Metode oversampling yang digunakan adalah **borderline SMOTE**.
- Dilakukan **eksplorasi metode lain** untuk mengatasi data yang **imbalanced**.



Model Training Klasifikasi

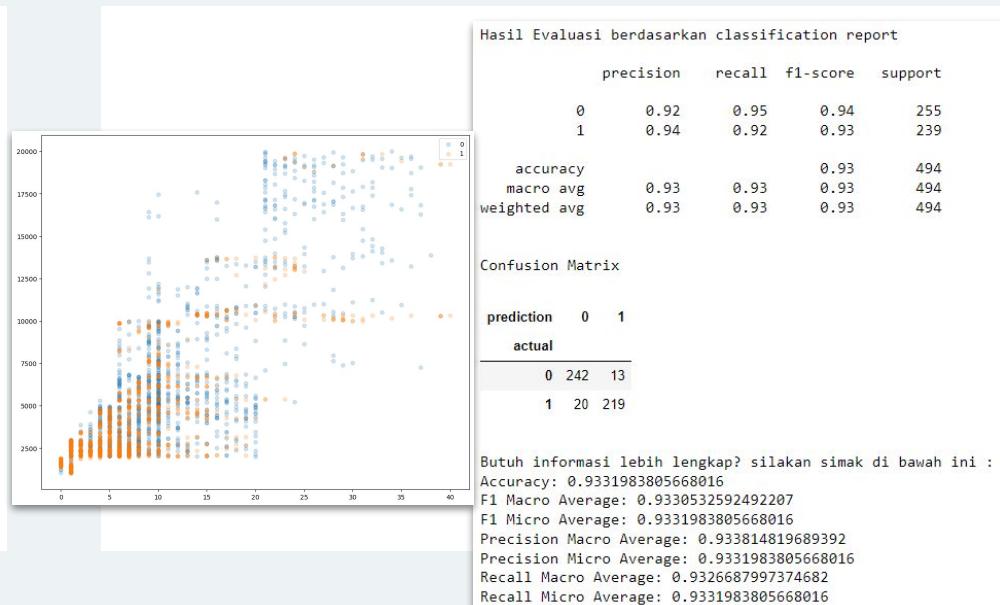
xx

2a. Prediksi karyawan akan **resign** atau tidak di perusahaan tersebut

Borderline SMOTE

Oversampling dengan **borderline SMOTE** menghasilkan model dengan **akurasi terbaik**

Akurasi yang dicapai oleh metode ini mencapai **93%**, dengan metrik yang **seimbang**.



Model Training Klasifikasi

Tanpa
Imbalanced
Method



Random Over
Sampler



SMOTE

Borderline
SMOTE

Metrik dari metode **imbalanced data handling** lainnya:

```
imbalanced method: none
classification model: LogisticRegression(max_iter=250, random_state=1234)
Counter({0: 1233, 1: 237})
Butuh informasi lebih lengkap? silakan simak di bawah ini :
Accuracy: 0.8673469387755102
F1 Macro Average: 0.7088527770859277
F1 Micro Average: 0.8673469387755102
Precision Macro Average: 0.7636363636363636
Precision Micro Average: 0.8673469387755102
Recall Macro Average: 0.6798606453086773
Recall Micro Average: 0.8673469387755102
```

```
imbalanced method: RandomOverSampler()
classification model: LogisticRegression(max_iter=250, random_state=1234)
Counter({1: 1233, 0: 1233})
Butuh informasi lebih lengkap? silakan simak di bawah ini :
Accuracy: 0.7611336032388664
F1 Macro Average: 0.7611179411523645
F1 Micro Average: 0.7611336032388665
Precision Macro Average: 0.7625400476464306
Precision Micro Average: 0.7611336032388664
Recall Macro Average: 0.7621954211018952
Recall Micro Average: 0.7611336032388664
```

```
imbalanced method: SMOTE()
classification model: LogisticRegression(max_iter=300, random_state=1234)
Counter({1: 1233, 0: 1233})
Butuh informasi lebih lengkap? silakan simak di bawah ini :
Accuracy: 0.9271255006728745
F1 Macro Average: 0.9269220769230877
F1 Micro Average: 0.9271255060728745
Precision Macro Average: 0.9282194941089231
Precision Micro Average: 0.9271255086728745
Recall Macro Average: 0.9263946491090551
Recall Micro Average: 0.9271255060728745
```

```
imbalanced method: BorderlineSMOTE()
classification model: LogisticRegression(max_iter=300, random_state=1234)
Counter({1: 1233, 0: 1233})
Butuh informasi lebih lengkap? silakan simak di bawah ini :
Accuracy: 0.9311740890688259
F1 Macro Average: 0.9308840715684823
F1 Micro Average: 0.9311740890688259
Precision Macro Average: 0.933695166548078
Precision Micro Average: 0.9311740890688259
Recall Macro Average: 0.930051685946345
Recall Micro Average: 0.9311740890688259
```

```
imbalanced method: SVMSMOTE()
classification model: LogisticRegression(max_iter=250, random_state=1234)
Counter({1: 1233, 0: 1233})
Butuh informasi lebih lengkap? silakan simak di bawah ini :
Accuracy: 0.9210526315789473
F1 Macro Average: 0.9208811245672609
F1 Micro Average: 0.9210526315789473
Precision Macro Average: 0.9216241116083179
Precision Micro Average: 0.9210526315789473
Recall Macro Average: 0.9205102961686766
Recall Micro Average: 0.9210526315789473
```

```
imbalanced method: ADASYN()
classification model: LogisticRegression(max_iter=250, random_state=1234)
Counter({0: 1233, 1: 1177})
Butuh informasi lebih lengkap? silakan simak di bawah ini :
Accuracy: 0.9273058921161826
F1 Macro Average: 0.9266276101129509
F1 Micro Average: 0.9273058921161826
Precision Macro Average: 0.9337882639068415
Precision Micro Average: 0.9273858921161826
Recall Macro Average: 0.9246163936689853
Recall Micro Average: 0.9273858921161826
```

```
imbalanced method: NearMiss()
classification model: LogisticRegression(max_iter=250, random_state=1234)
Counter({0: 237, 1: 237})
Butuh informasi lebih lengkap? silakan simak di bawah ini :
Accuracy: 0.7263157894736842
F1 Macro Average: 0.7248217468805784
F1 Micro Average: 0.7263157894736842
Precision Macro Average: 0.7412280701754386
Precision Micro Average: 0.7263157894736842
Recall Macro Average: 0.7322222222222222
Recall Micro Average: 0.7263157894736842
```

SVMSMOTE

ADASYN

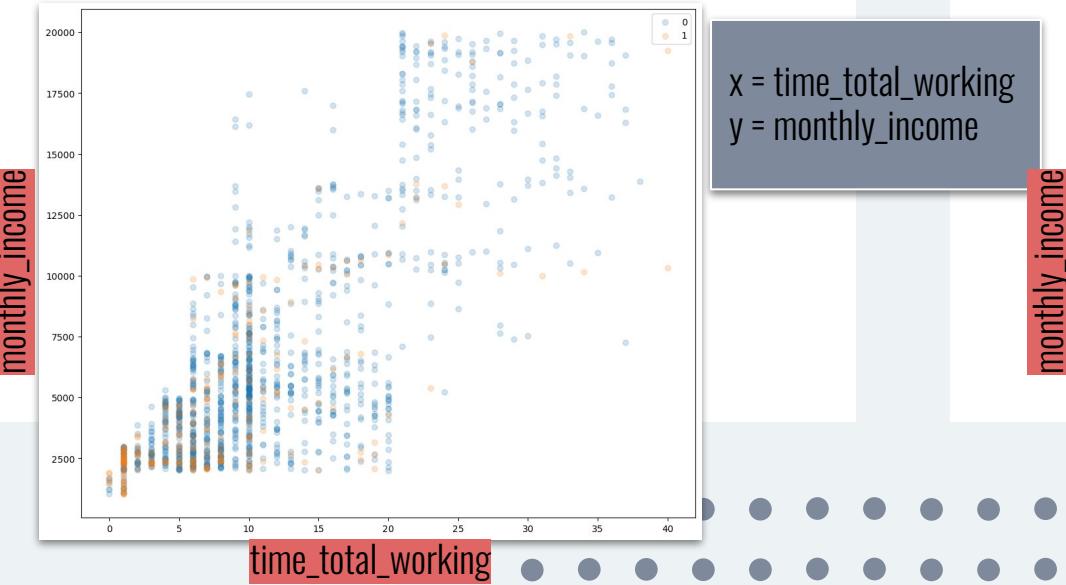
NearMiss

Model Training Klasifikasi

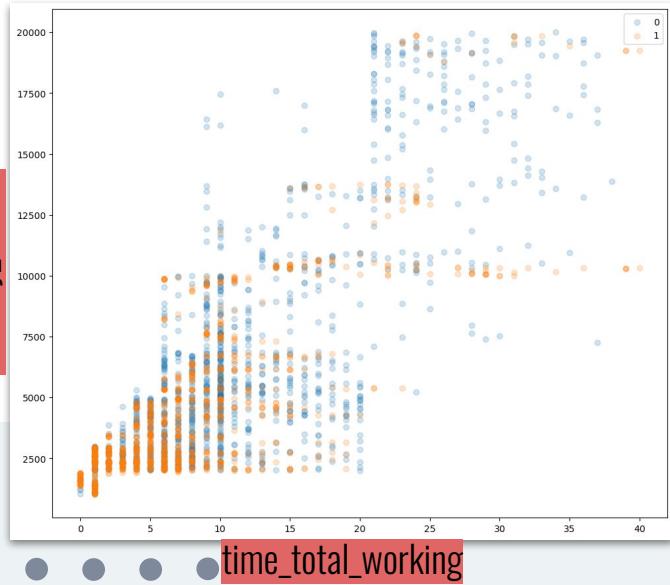
**

Dengan dan tanpa **Oversampling** Borderline SMOTE

Sebelum Borderline SMOTE



Setelah Borderline SMOTE



Model Training Klasifikasi

Bagaimana **hasil prediksi** anda dapat membantu perusahaan dalam mengambil **keputusan**?

Model **prediksi** yang dibuat memiliki tingkat **akurasi** yang **sangat tinggi**, mencapai **93%**. Hal tersebut berarti **model** dapat **memprediksi** orang yang akan **resign** dari perusahaan dengan sangat akurat.



Hasil prediksi tersebut dapat digunakan oleh **perusahaan** untuk mengambil keputusan untuk **memberhentikan karyawan** apabila memiliki **kinerja kurang baik** dan **cenderung** akan **resign**, atau mencoba **mempertahankan karyawan** yang **sangat eksepsional** dalam bekerja **namun** memiliki **kecenderungan** untuk **resign**.



Model Training Regresi

2b. Berapa lama seorang karyawan akan bertahan di perusahaan?

Pada masalah ini, kami menyatakan bahwa masalah ini adalah masalah regresi. Untuk mengetahui berapa lama seorang karyawan akan bertahan di perusahaan, kami mengambil data dari karyawan yang telah resign dan menjadikan **time_current_company** sebagai target dari regresi.



Model Training Regresi

Berapa lama seorang karyawan akan bertahan di perusahaan?

Model yang digunakan:

- Lasso Regression
- RandomForestRegression
- RandomForestRegression dengan Hyperparameter Tuning
- Linear Regression



Model Training Regresi

Berapa lama seorang karyawan akan bertahan di perusahaan?

Dari keseluruhan model yang telah dibuat, dapat dilihat bahwa pada model RandomForest yang telah di tuned hyperparameternya, memiliki nilai R2 yang lebih baik dan juga nilai error yang lebih kecil dibandingkan model lainnya.

```
LinearRegression
-----
MAE: 1.9990822040140142
MSE: 10.073958305278621
RMSE: 3.1739499531779987
R_squared: 0.792259774314747
```



Lasso Regression

```
-----
MAE: 1.9593944439559348
MSE: 12.581954570444973
RMSE: 3.5471050971806535
R_squared: 0.7405411058078081
```

RandomForest Regression

```
-----
MAE: 1.1870833333333335
MSE: 5.543479166666665
RMSE: 2.3544594213251298
R_squared: 0.8856850923671774
```

Tuned RandomForest Regression

```
-----
MAE: 1.2581258080106785
MSE: 4.5088906504299775
RMSE: 2.1234148559407737
R_squared: 0.9070198691591126
```

Model Training Regresi

Bagaimana hasil prediksi anda dapat membantu perusahaan dalam mengambil keputusan?

- Menyeleksi Karyawan yang akan diterima perusahaan
- Membantu dalam mengambil kebijakan seperti memberi overtime atau memberikan promosi

```
Fitur yang Mempengaruhi Regresi:  
score_environment  
score_contribution  
job_rank  
score_job_satisfaction  
over_time  
score_work_relationship  
time_total_working  
time_current_role  
time_last_promotion  
time_current_manager  
role_Sales Executive  
marriage_status_Married
```

Model Clustering

2c. Lakukan analisis cluster yang dapat terbentuk pada data karyawan. Deskripsikan karakteristik masing-masing cluster yang didapatkan!

Clustering merupakan unsupervised learning yang tidak memiliki predefined classes dan target variable. Pada dasarnya, tujuan clustering adalah untuk menemukan kelompok yang berbeda dalam elemen dalam data. Dalam analisis clustering ini, kami menggunakan dua metode, yaitu **K-Means Clustering** dan **Hierarchical Clustering** dengan algoritma **Agglomerative Clustering**.



Model Clustering

Preparation (Standard scaling)

Sebelum menerapkan metode clustering, kami melakukan preparation terhadap dataset. Dalam dataset ini, range data antar variabel berbeda jauh. Hal ini dapat memengaruhi hasil clustering, khususnya K-Means Clustering karena metode tersebut menggunakan Euclidean Distance untuk mengukur jaraknya. Oleh karena itu, kami melakukan **Standard Scaling**.



Model Clustering

Preparation (Standard scaling)

	age	resign	home_distance	education	score_environment	gender	hourly_rate	score_contribution
0	41	1	1	2		2	0	94
1	49	0	8	1		3	1	61
2	37	1	2	2		4	1	92
3	33	0	3	4		4	0	56
4	27	0	2	1		1	1	40



```
[191] kolom = ['age', 'home_distance', 'hourly_rate', 'monthly_income', 'companies_count', 'salary_increment_percentage', 'time_to_promotion', 'last_year_training_time', 'time_current_role', 'time_last_promotion', 'time_current_manager', 'time_current_company']
      std_scale = StandardScaler()
      df_norm[kolom] = std_scale.fit_transform(df_norm[kolom])
      df_norm.head()
```

	age	resign	home_distance	education	score_environment	gender	hourly_rate	score_contribution	job_rank	score_job
0	0.446350	1	-1.010909	2		2	0	1.383138	3	2
1	1.322365	0	-0.147150	1		3	1	-0.240677	2	2
2	0.008343	1	-0.887515	2		4	1	1.284725	2	1
3	-0.429664	0	-0.764121	4		4	0	-0.486709	3	1
4	-1.086676	0	-0.887515	1		1	1	-1.274014	3	1

Model Clustering

Preparation (Dimension Reduction)

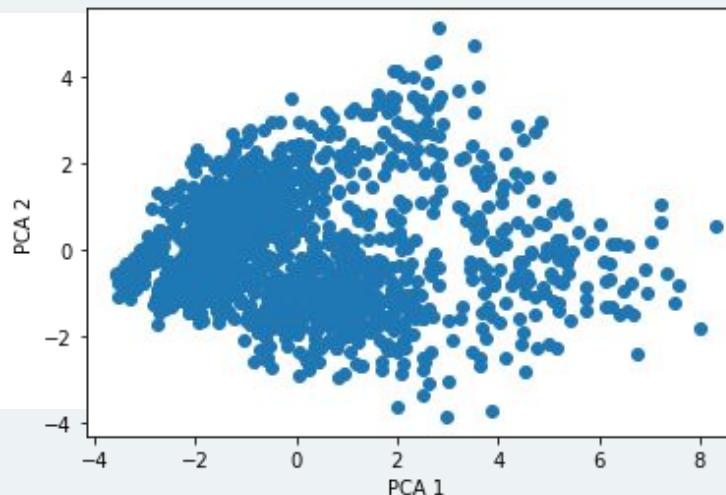
Principle Component Analysis (PCA) adalah teknik untuk mengubah data berdimensi tinggi menjadi berdimensi lebih rendah sambil menyimpan informasi sebanyak mungkin. Metode ini kami pilih untuk melakukan dimension reduction karena dataset ini merupakan dataset dengan high dimension dan memiliki banyak sekali fitur sehingga dengan dilakukan PCA, clustering akan lebih mudah dilakukan.



	P1	P2
0	-0.314834	1.124209
1	0.516758	-0.898667
2	-2.520416	1.199138
3	-0.958451	-0.789451
4	-1.973344	0.700560

Model Clustering

Visualisasi awal

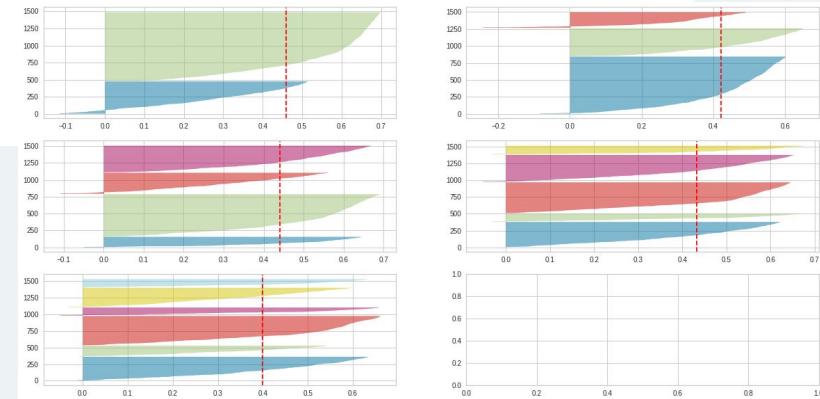
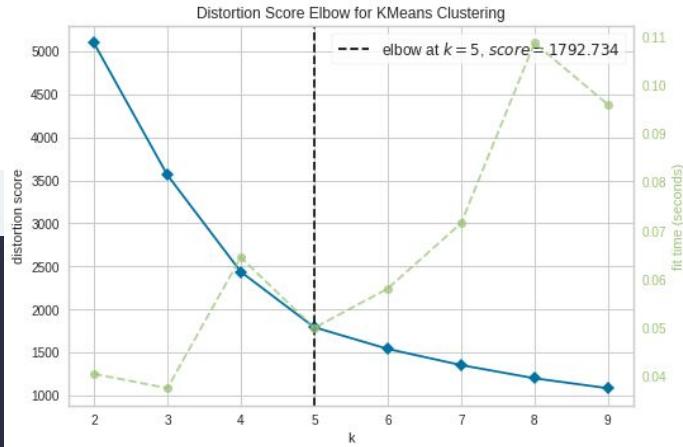


Visualisasi dataset awal setelah dilakukan PCA dan sebelum clustering.

Model Clustering

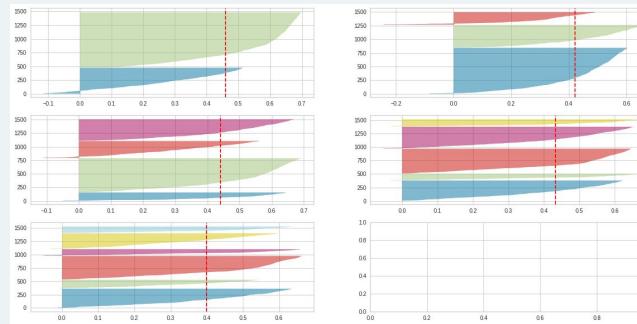
K-Means Clustering

Hal pertama yang kami lakukan adalah menentukan nilai K yang optimal untuk melakukan clustering. Dalam menentukan nilai K, kami menggunakan dua internal measures, yaitu **Elbow Method** dan **Sillhouette Coefficient**.



Model Clustering

K-Means Clustering



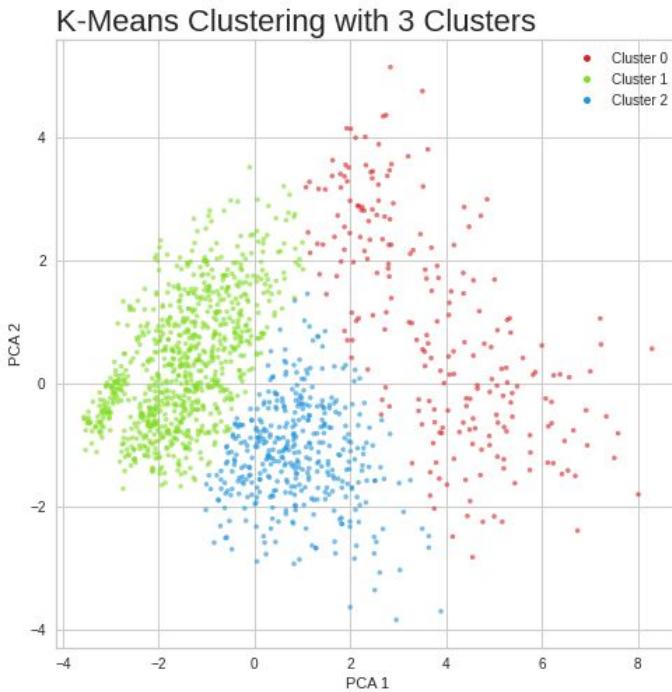
Dari grafik Elbow Method, ditunjukkan bahwa nilai $k = 5$. Akan tetapi, dari grafik tersebut, tidak terlihat jelas titik elbow, sehingga internal measure dengan elbow method kurang bisa diandalkan. Oleh karena itu, kami kemudian menggunakan Sillhouette Coefficient. Dari hasil Sillhouette Score yang dihasilkan di atas, $k = 2$ menunjukkan score yang paling besar. Akan tetapi, terdapat fluktuasi besar dalam ukuran cluster dalam $k = 2$ (ukuran cluster hijau jauh lebih lebar dibanding cluster biru). Lebar pada cluster mewakili jumlah titik data. Oleh karena itu, kami menggunakan **$k = 3$** karena ukuran cluster yang lebih uniform dibanding $k = 2$.



```
For n_clusters = 2 The average silhouette_coefficient is : 0.16295742375329347
For n_clusters = 3 The average silhouette_coefficient is : 0.09879640771181052
For n_clusters = 4 The average silhouette_coefficient is : 0.0725944895211867
For n_clusters = 5 The average silhouette_coefficient is : 0.07532102870034671
For n_clusters = 6 The average silhouette_coefficient is : 0.06322327934602581
```

Model Clustering

Visualisasi Clustering dengan K-Means



class_name	instance_count	rule_list
1	0	408 [0.8538135593220338] (time_current_company > -0.08297300623962656) and (time_total_working <= 1.185426652431488)
0	1	837 [0.9898862199747156] (time_current_company <= -0.08297300623962656) and (time_total_working <= 1.185426652431488)
2	2	225 [0.8852459016393442] (time_current_company <= -0.08297300623962656) and (time_total_working > 1.185426652431488) [0.9931506849315068] (time_current_company > -0.08297300623962656) and (time_total_working > 1.185426652431488)

Model Clustering

Karakteristik cluster dengan K-Means Clustering

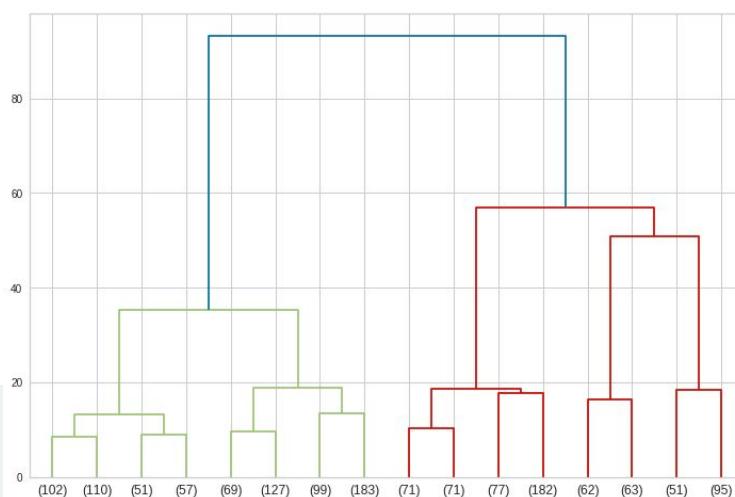
Dari hasil yang kami dapatkan, karakteristik cluster yang terbentuk adalah:

- **Cluster 0 (merah):** semua instance yang memenuhi kondisi $(\text{time_current_company} \leq -0.08297300623962656)$ dan $(\text{time_total_working} > 1.185426652431488)$, 88% dari instance tersebut berada di Cluster 0.
- **Cluster 1 (hijau):** semua instance yang memenuhi kondisi $(\text{time_current_company} \leq -0.08297300623962656)$ and $(\text{time_total_working} \leq 1.185426652431488)$, kurang lebih 99% dari instance tersebut berada di Cluster 1.
- **Cluster 2 (biru):** semua instance yang memenuhi kondisi $(\text{time_current_company} > -0.08297300623962656)$ and $(\text{time_total_working} \leq 1.185426652431488)$, 85% dari instance tersebut berada di Cluster 2.



Model Clustering

Visualisasi Agglomerative

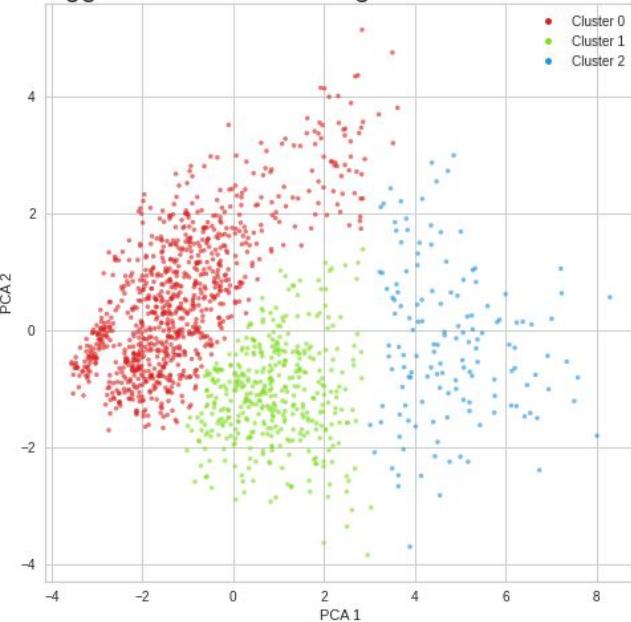


Visualisasi dendogram tanpa jumlah cluster

Model Clustering

Visualisasi Agglomerative

Agglomerative Clustering with 3 Clusters



A table titled "rule_list" showing the rules for the three clusters. The columns are "class_name", "instance_count", and "rule_list".

class_name	instance_count	rule_list
0	908	[0.9929577464788732] (time_current_company <= -0.08297300623962656)
1	416	[0.8601694915254238] (time_current_company > -0.08297300623962656) and (time_total_working <= 1.185426652431488)
2	146	[0.8835616438356164] (time_current_company > -0.08297300623962656) and (time_total_working > 1.185426652431488)



Model Clustering

Karakteristik cluster dengan Agglomerative Clustering

Dari hasil yang kami dapatkan, karakteristik cluster yang terbentuk adalah:

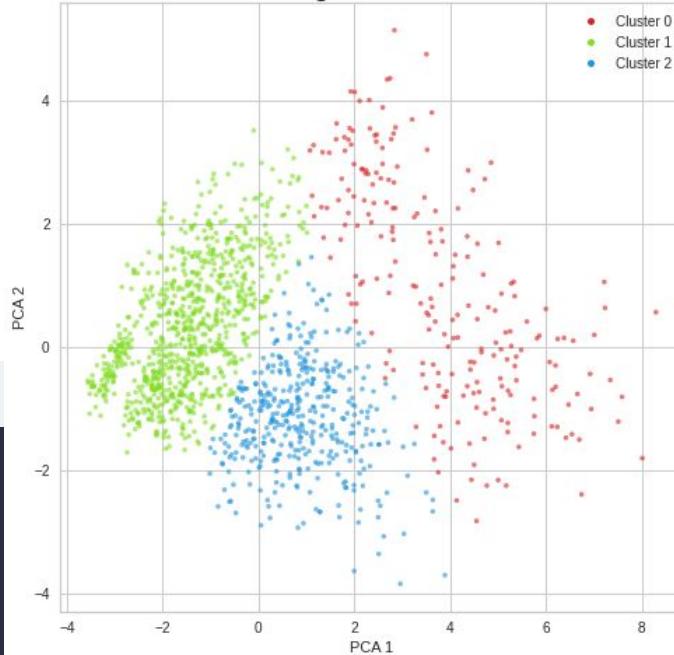
- **Cluster 0 (merah):** semua instance yang memenuhi kondisi $(\text{time_current_company} \leq -0.08297300623962656)$, 99% dari instance tersebut berada di Cluster 0.
- **Cluster 1 (hijau):** semua instance yang memenuhi kondisi $(\text{time_current_company} > -0.08297300623962656)$ and $(\text{time_total_working} \leq 1.185426652431488)$, 86% dari instance tersebut berada di Cluster 1.
- **Cluster 2 (biru):** semua instance yang memenuhi kondisi $(\text{time_current_company} > -0.08297300623962656)$ and $(\text{time_total_working} > 1.185426652431488)$, 88% dari instance tersebut berada di Cluster 2.



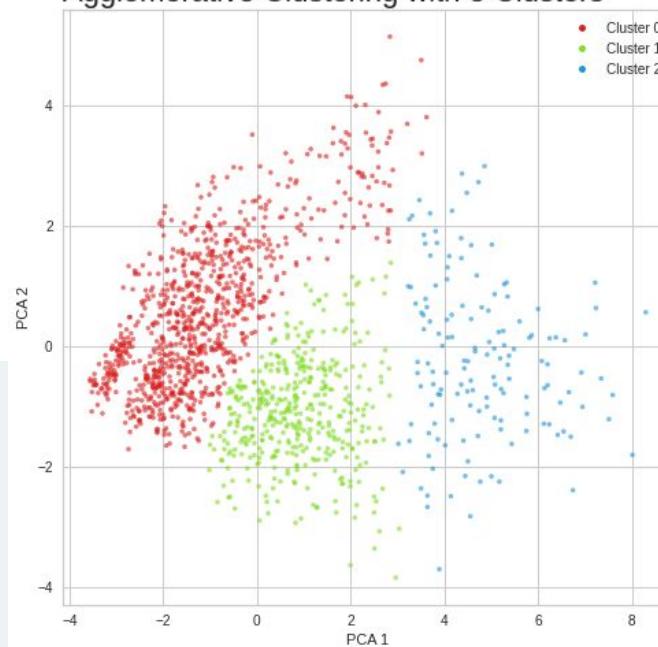
Model Clustering

Perbandingan Visualisasi K-Means Clustering dengan Agglomerative Clustering

K-Means Clustering with 3 Clusters



Agglomerative Clustering with 3 Clusters



X
X

X

Model Clustering

Perbandingan Karakteristik K-Means Clustering dengan Agglomerative Clustering

K-Means Clustering:

- **Cluster 0 (merah):** semua instance yang memenuhi kondisi $(\text{time_current_company} \leq -0.08297300623962656)$ dan $(\text{time_total_working} > 1.185426652431488)$, 88% dari instance tersebut berada di Cluster 0.
- **Cluster 1 (hijau):** semua instance yang memenuhi kondisi $(\text{time_current_company} \leq -0.08297300623962656)$ and $(\text{time_total_working} \leq 1.185426652431488)$, kurang lebih 99% dari instance tersebut berada di Cluster 1.
- **Cluster 2 (biru):** semua instance yang memenuhi kondisi $(\text{time_current_company} > -0.08297300623962656)$ and $(\text{time_total_working} \leq 1.185426652431488)$, 85% dari instance tersebut berada di Cluster 2.

Agglomerative Clustering:

- **Cluster 0 (merah):** semua instance yang memenuhi kondisi $(\text{time_current_company} \leq -0.08297300623962656)$, 99% dari instance tersebut berada di Cluster 0.
- **Cluster 1 (hijau):** semua instance yang memenuhi kondisi $(\text{time_current_company} > -0.08297300623962656)$ and $(\text{time_total_working} \leq 1.185426652431488)$, 86% dari instance tersebut berada di Cluster 1.
- **Cluster 2 (biru):** semua instance yang memenuhi kondisi $(\text{time_current_company} > -0.08297300623962656)$ and $(\text{time_total_working} > 1.185426652431488)$, 88% dari instance tersebut berada di Cluster 2.



Model Clustering

Kesimpulan

Dari hasil kedua model clustering, walaupun menghasilkan dua visualisasi yang berbeda, dapat disimpulkan karakteristik dari karyawan yang memiliki segmentasi atau grup yang sama bisa dilihat dari atribut **time_current_company** dan **time_total_working**. Adapun kemudian masing-masing cluster dan model memiliki threshold-nya masing-masing untuk menspesifikasikan kedalam cluster mana instance data tersebut masuk.



“Terima Kasih”



—Kelompok DiCARRY

